N°d'ordre: .....

République Algerienne Démocratique et Populaire
Ministère de l'enseignement supérieur et de la Recherche scientifique

Université Djillali Liabès de Sidi Bel Abbès
Faculté des Sciences Exactes
Département d'informatique
Laboratoire EEDIS

Thèse de Doctorat en Sciences

Filière : Informatique
Spécialité : Informatique

Par

Mr DJAFRI Laouni

Analyse de données massives -Big Data- pour la prédiction

Soutenue le 24/06/2020. devant le jury :

Mr. BERRABAH Djamel	MCA	UDL de Sidi Bel Abbes	Président du jury
Mr. AMAR BENSABER Djamel	MCA	ESI de Sidi Bel Abbès	Directeur de thèse
Mr. ADJOUDJ Réda	MCA	UDL de Sidi Bel Abbès	Co-directeur de thèse
Mr. TOMOUH Adil	MCA	UDL de Sidi Bel Abbes	Examinateur
Mr. MALKI Mimoun	Pr	ESI de Sidi Bel Abbès	Examinateur
Mr. BENSLIMANE Sidi Mohamed	Pr	ESI de Sidi Bel Abbès	Examinateur

Année Universitaire : 2019. - 2020.

Je dédie cette thèse : A ma mère , A ma femme pour sa patience, son soutien et ses encouragements, et A mes enfants. . .

Remerciements

Nous remercions dieu tout-puissant de nous avoir donné le privilège d'étudier et de suivre le chemin de la science.

Nous tenons à adresser nos plus sincères remerciements au directeur de thèse Mr.«AMAR Bensaber Djamel» pour avoir encadré notre travail de thèse de doctorat. Nous le remercions pour toute la confiance qu'il m'a témoignée, pour son entière disponibilité, et pour tous ses conseils avisés.

Nous tenons également à remercier le co-directeur de thèse Mr. «ADJOUDJ Réda» d'avoir tout en nous faisons confiance d'accepter de nous suivre. Je le remercie pour les qualités scientifiques et pédagogiques de son encadrement et pour sa disponibilité.

Nous souhaiterions remercier les membres du jury, Mr.«BERRABAH Djamel», Mr.«TOMOUH Adil», Mr.«MALKI Mimoun» , et Mr.«BENSLIMANE Sidi Mohamed» qui nous avons fait l'honneur d'accepter d'être examinateurs et pour avoir bien voulu consacrer une partie de leur temps à examiner notre travail.

On remercie toutes les personnes avec lesquelles nous avons eu la joie de collaborer, de près ou de loin, à notre projet. Elles ont toutes su nous faire bénéficier de leurs connaissances et compétences.

Mes derniers remerciements et non les moindres, iront à nos proches qui m'ont toujours apportés leur soutien sans faille. Nous les remercions de toute l'affection et tout l'amour qu'ils m'ont témoignés.

Sidi Bel Abbès, le 30 juin 2020.

Table des matières

Liste des figures

Liste des tableaux

RÉSUMÉ

الملخص

منذ عدة سنوات ، شهدنا انفجارًا لمصادر جديدة للبيانات المتنوعة ذات الدقة العاليه والكمون المنخفض (المعروف باسم البيانات الكبيرة). حيث ظهرت مصادر جديدة واعدة للبيانات ، مثل الشبكة الاجتماعية أو البيانات المتعلقة بالإنترنت. تتكون البيانات الكبيرة من معالجة كميات كبيرة جدًا من البيانات المتنوعة للغاية في الوقت الفعلي وتحليلها. تهتم جميع الشركات ، خاصة تلك التي لديها كميات كبيرة من المعلومات وتريد غربلتها لتحسين معرفة العملاء وتحسين حملاتهم. البيانات الكبيرة هي محور رئيسي للتحول الرقمي للاقتصاد ورافعة مهمة للقدرة التنافسية للأعمال. حيث تساعد على فهم العملاء والموردين والشركاء بشكل أفضل من خلال تحليل هذه المعلومات المقدمة من العملاء والمستخدمين ، حيث نأمل في تحسين الخدمات التي تقدمها الشركات. في موضوعنا ، نهتم بتحليل البيانات الضخمة للتنبؤ بالاحتمالات المستقبلية بمستوى مقبول من الموثوقية ، وذلك لاتخاذ جميع التدابير اللازمة في المستقبل لتجنب الخسائر ، لتحسين الخدمات وكذا اتخاذ القرارات المقترحة والفعالة في أقصر وقت ممكن. ينصب التركيز الرئيسي لهذا الموضوع على الأساليب التحليلية المستخدمة في مجموعات البيانات الكبيرة استنادًا إلى خوارزميات التعلم الآلي. ومن ثم يمكن استخدام التحليلات التنبؤية للبيانات الضخمة لتوقع الصعوبات ، وتحسين خدمة العملاء ، وتوفير الخدمات الأكثر صلة.

الكلمات المفتاحية : البيانات الكبيرة ، التحليلات التنبؤية للبيانات الكبيرة ، تكنولوجيات البيانات الكبيرة ، أخذ العينات الإحصائية ، التعلم الآلي.

Résumé

Depuis plusieurs années, nous assistons à une explosion de nouvelles sources de données diverses à granularité fine et à faible latence (dites « Big Data »). De nouvelles sources de données prometteuses, telles que le web social ou le web des données liées sont apparues. Le «Big Data»

consiste à traiter, en temps réel, de très gros volumes de données extrêmement variées et à les analyser. Toutes les entreprises sont concernées, surtout celles qui possèdent de vastes gisements d'informations et souhaitent les passer au crible pour améliorer leur connaissance client et optimiser leurs campagnes. Le Big Data constitue un axe majeur de transformation numérique de l'économie et un levier important de compétitivité des entreprises. Il permet de mieux comprendre les clients, fournisseurs et partenaires. En analysant ces informations fournies par leurs clients et utilisateurs, on espère valoriser les services proposés par les entreprises. Dans notre thème, on s'intéresse à l'analyse du Big Data pour prévoir les tendances et les comportements futures de l'être humain ou des objets avec un niveau de fiabilité acceptable, ainsi pour prendre toutes les dispositions nécessaires à l'avenir afin d'éviter les pertes, d'améliorer les services proposés et la prise de décision efficace dans le plus bref délai. L'objectif principal de ce thème porte sur les méthodes d'analyse utilisées pour les grandes collections de données en se basant sur les algorithmes du Machine Learning. L'analyse prédictive de données massives peut être utilisée pour anticiper les difficultés, pour améliorer le service client et pour proposer les services les plus pertinents.

Mots clés : Données Massives, Analyse prédictive des Données Massives, les technologies du Big Data, échantillonnage statistique, Apprentissage Automatique.

Abstract

Several years ago, we witnessed an explosion of new sources of diverse data with high accuracy and low latency (known as "Big Data"). New promising data sources have emerged, such as social network or internet-related data. Big data consists of processing and analyzing very large amounts of very diverse data in real time. All companies are interested, especially those with large amounts of information and want to filter them to improve customer knowledge and improve their campaigns. Big data is a major focus of the digital transformation of the economy and an important lever for business competitiveness. It helps to better understand customers, suppliers and partners by analyzing this information from customers and users, where we hope to improve the services provided by companies. In our topic, we are interested in analyzing big data to predict future prospects with an acceptable level of reliability, to take all necessary measures in the future to avoid losses, to improve services as well as to make proposed and effective decisions in the shortest possible time. The main focus of this topic is on the analytical methods used in large datasets based on machine learning algorithms. Predictive analytics of big data can then be used to predict difficulties, improve customer service, and provide the most relevant services.

Key words : Big Data, Big Data predictive analytics, Big Data technologies, Statistical sampling, Machine Learning.

Introduction génerale

1

Sommaire

1.1 Introduction

Les données volumineuses garantissent de nouveaux niveaux de divulgation d'investigation et de qualité financière. Donc, qu'y a-t-il de nouveau dans le Big Data, et en quoi elles diffèrent des données traditionnelles à petite ou moyenne échelle ? Cette thèse décrit les portes ouvertes et les difficultés apportées par le Big Data, en mettant l'accent sur les éléments reconnus du Big Data et la technique mesurable et informatique, et aussi en enregistrant l'ingénierie pour les gérer. La gestion et l'analyse d'ensembles de données à grande échelle sont généralement associées au terme Big Data. Le Big Data est la pratique de la collecte et du traitement de grands ensembles de données en utilisant des systèmes et des algorithmes pour les analyser.

Nous entrons dans le temps du Big Data, un terme qui fait allusion à l'explosion de données maintenant accessibles. Avec l'analyse du Big Data, les entreprises peuvent prendre de meilleures décisions plus rapidement. En outre, cela aide également les clients à obtenir de précieuses informations sur ce qu'ils veulent et quels sont leurs besoins, ceci est connu sous le nom « l'analyse de données pour la prédiction ».

L'analyse des données est un ensemble de techniques permettant de découvrir une structure, éventuellement complexe, d'un tableau de données à plusieurs dimensions et de le traduire avec une structure plus simple et de le résumer autant que possible. Souvent, cette structure peut être représentée graphiquement. Les méthodes à utiliser pour l'analyse des données varient selon que si vous explorez une nouvelle idée dans votre recherche ou si vous souhaitez prouver une idéologie déjà présente. Ces méthodes visent également à fournir les liens pouvant exister entre différentes données et à en dessiner des informations statistiques permettant de décrire brièvement les informations clés contenues dans ces données. Nous pouvons également essayer de classer les données en différents sous-ensembles plus homogènes.

L'analyse des données recouvre principalement deux ensembles de techniques : les premières qui relèvent de la géométrie euclidienne et conduisent à l'extraction de valeurs et de vecteurs propres, sont appelées analyses factorielles ; les secondes, dites de apprentissage automatique sont caractérisées par le choix d'un indice de proximité et d'un algorithme d'agrégation ou de désagrégation qui permettent d'obtenir une partition ou arbre de classification. On s'intéressera surtout par l'apprentissage automatique qui suit la technique d'analyse des données qui est en charge de l'automatisation scientifique du modèle. Les algorithmes utilisés pour l'apprentissage automatique sont des outils très variés capables d'effectuer des prévisions tout en acquérant des connaissances issus de billions d'observations. L'apprentissage automatique est considéré comme une extension moderne de l'analyse prédictive. Donc, l'analyse de données prédictive repose sur les algorithmes d'apprentissage automatique.

En générale, l'analyse de données permet de dégager des tendances, des profils, de détecter des comportements ou de trouver des liens et des règles. Il existe trois types d'analyse de données les plus répandus (voir chapitre 3) : l'analyse descriptive, prédictive et prescriptive sont des solutions interdépendantes qui aident les entreprises à tirer le meilleur parti des données dont elles disposent. Parmi ces types d'analyse de données, nous nous intéresserons à l'analyse prédictive.

L'analyse prédictive aide les décideurs à faire des choix et à résoudre des problèmes qui ont des impacts durables. La base qui offre de nouvelles possibilités d'amélioration continue est issue des disciplines mathématiques ; de la collecte des données à l'analyse des données. L'analyse prédictive est donc fondamentalement dépendante des statistiques mathématiques et de l'informatique, ces deux branches s'imbriquent dans tous les aspects des efforts de recherche de données scientifiques. En fait, les rôles joués par les deux branches sont essentiels au processus de développement évolutif du Big Data, ces deux branches continuent de façonner notre compréhension collective de l'analyse prédictive. L'analyse prédictive a une vaste portée et une large application. L'analyse prédictive ne peut être réalisée de manière isolée, elle doit être considérée comme une approche systématique permettant aux systèmes de travailler en parallèle pour obtenir des résultats. À cette fin, l'analyse prédictive est réalisée par les algorithmes du Machine Learning, elle est également étayé par des techniques statistiques pour modéliser, analyser et prendre des décisions.

En outre, la plupart des entreprises ont adopté les plateformes pour l'analyse de données massives. De telles nouvelles et d'autres similaires motivent une réflexion sur les raisons pour lesquelles l'adoption du Big Data n'est pas encore une réalité pour beaucoup d'entreprises. D'une part, nous sommes en train de lutter contre un déluge de données qui requiert des solutions évolutives pour extraire la valeur d'un grand volume de données. D'autre part, les technologies Big Data sont généralement présentées comme la réponse clé à un tel besoin. Ainsi, l'industrie est devenue convaincue par les promesses du Big Data en comprenant des connaissances approfondies sur les avantages et les inconvénients des solutions disponibles pour les données massives.

A ce stade on peut dire que le Big Data est un écosystème large et complexe. Il nécessite la maitrise des technologies matérielles et logicielles diverses (stockage, parallélisation des traitements, virtualisation, ...). Le Big Data demande de la compétence et de l'expertise dans la maitrise et l'analyse des données. Les usages du Big Data sont très vastes qui touchent presque tous les secteurs d'activités (marketing, santé, sport, transport ...).

1.2 Positionnement du problème

Le domaine de l'analyse prédictive soit un sujet d'actualité pour les chercheurs dans les domaines des mathématiques de l'informatique. La modélisation et l'analyse prédictive peuvent revêtir une importance cruciale pour les organisations si elles sont correctement alignées sur leurs processus et leurs besoins métier. Elles peuvent également améliorer considérablement leurs performances et la qualité de leurs décisions augmentant leur valeur métier. Chaque organisation peut analyser statistiquement ses données et mieux connaître son environnement, mais le potentiel de profit le plus important réside parmi ceux qui sont en mesure de réaliser une modélisation et une analyse prédictive.

Étant donné que l'analyse prédictive joue un rôle important dans les organisations les plus rentables, elle représente également une tendance suivie par les organisations, elle doit être considérée comme un segment important de l'aide à la prise de décision. De plus, les recherches sur les données massives déjà effectuées dans des domaines associés à l'analyse prédictive sont trop nombreuses. Mais, il existe de nouvelles possibilités qui ouvrent la poursuite des recherches.

Pour soutenir l'initiative d'application de l'analyse prédictive, un certain nombre de problèmes liés à la gestion des données massives et aux ressources informatiques doivent être résolus. Du point de vue de l'infrastructure, quels composants et normes de réseau sont nécessaires pour prendre en charge l'analyse prédictive ? Du point de vue de l'architecture, quels logiciels et services sont nécessaires pour créer et prendre en charge l'analyse prédictive ? Et comment concevoir une architecture soutenue par les technologies du Big Data pour l'analyse prédictive de données massives ? Généralement comment on choisit la meilleure architecture de bout en bout pour résoudre notre problème de Big Data ? D'un point de vue conceptuel, quels sont les problèmes et les risques inhérents à l'analyse prédictive ? Comment pouvons-nous améliorer le résultat de l'analyse prédictive ? Quelles sont les méthodes de statistique à adopter pour le Big Data ? Et quels sont les algorithmes du Machine Learning très appropriés pour l'analyse prédictive de données massives ? En fin, la question cruciale qui inclut les questions de recherche ci-dessus auxquelles nous répondrons dans notre thèse est la suivante : Quelle est la méthode optimale pour traiter les données volumineuses (Volume), à condition : on garde la stabilité du résultat de l'analyse prédictive à un niveau acceptable (Véracité), en plus, on peut faire le traitement en temps réel (Vélocité) ? Ces questions et d'autres seront discutées dans notre thèse.

1.3 Objectif

L'analyse prédictive consiste à utiliser les données, les algorithmes statistiques et les techniques du Machine Learning permettant de prédire les probabilités du comportement humain, des catastrophes naturelles, des ventes, des résultats financiers des entreprises, etc., en se basant sur le passé. Compte tenu de l'augmentation du volume de données ; les outils d'analyse traditionnels ne sont pas assez puissants pour exploiter pleinement la valeur du Big Data. Le volume de données est trop large pour une analyse complète, et les corrélations et relations entre ces données sont extrêmement importantes pour que les analystes testent toutes les hypothèses afin de tirer une valeur de ces données. Pour analyser ces données, nous devons donc trouver des solutions plus efficaces que les méthodes traditionnelles en termes de précision et de rapidité.

Alors, l'objectif de notre travail consiste premièrement, à améliorer le résultat de prédiction (précision) dans le contexte du Big Data. Deuxièmement, notre objectif est d'accélérer le temps d'exécution (vitesse). Cela nécessite de concevoir de nouveaux outils d'analyse prédictive des données massives utilisant ces trois facteurs : (1) L'apprentissage automatique, (2) Les méthodes statistiques et (3) Le calcul parallèle et distribué.

Le premier facteur est utilisé pour suivre ou satisfaire seul le flux de données toujours croissant et en constante évolution et pour fournir des informations précieuses en constante évolution. Les algorithmes d'apprentissage automatique définissent les données entrantes et identifient les schémas pertinents, qui sont ensuite traduits en informations précieuses qui peuvent ensuite être implémentées dans les opérations commerciales. Après cela, les algorithmes automatisent également certains aspects du processus de prise de décision. Le cœur de l'analyse prédictive se fonde sur la capture des relations entre des données issues du passé et à les utiliser pour prédire les résultats futurs. Afin de pouvoir faire des prédictions sur un ensemble de données donné, une ou plusieurs variables prédictives sont utilisées pour prédire une variable de réponse. Sous sa forme la plus simple, l'analyse prédictive est un support permettant d'établir des prédictions pour la prise de décision. Pour des exigences plus complexes, des techniques d'analyse prédictive avancées (l'apprentissage supervisé et l'apprentissage non supervisé) sont utilisées afin d'orienter les processus stratégiques de l'entreprise. L'apprentissage supervisé est divisé en deux grandes catégories : la régression pour les réponses continues et la classification pour les réponses discrètes. L'apprentissage non supervisé permet de tirer des conclusions à partir d'ensembles de données composés de données en entrée sans réponses étiquetées. Dans notre travail, nous concentrons sur les algorithmes d'apprentissage supervisé, car nous allons prendre des décisions futures en fonction d'événements connus (les classes prédites sont connues au préalable) dans le passé. Il existe de nombreux algorithmes d'apprentissage supervisé utilisés pour prendre toutes les dispositions nécessaires à l'avenir pour éviter les pertes, pour améliorer les services proposés, pour la prise de déci-

sion, etc. On va détailler ces algorithmes au chapitre 3. Parmi ces algorithmes, nous avons ciblé l'algorithme Random Forests, car cet algorithme est l'un des meilleurs algorithmes d'apprentissage automatique supervisé [Fernandez-Delgado *et al.* 2014] pour résoudre les problèmes de classification et de régression, il est robuste face aux valeurs aberrantes et au bruit, il résout les problèmes de sur-apprentissage. Random Forests est un bon classifieur pour une implémentation parallèle.

Le deuxième facteur est utilisé pour réduire le volume du Big Dataset ; il représente l'ensemble des processus visant à réunir un certain nombre d'individus dans une population donnée. Pour que les résultats observés dans une étude puissent être généralisés à une population statistique, l'échantillon doit être représentatif de cette dernière, c-à-d qu'il doit refléter fidèlement sa composition et sa complexité. Il existe deux méthodes d'échantillonnage : l'échantillonnage probabiliste et l'échantillonnage non probabiliste. L'échantillonnage probabiliste implique la sélection d'unités dans une population selon le principe du choix aléatoire. Chaque unité de la population a une probabilité mesurable d'être choisie. Dans l'échantillonnage non probabiliste, une méthode subjective de sélection des unités est appliquée à la population. La distribution des caractéristiques au sein de la population devrait être égale. Dans notre travail, nous nous sommes appuyés sur l'échantillonnage probabiliste ; car ce type d'échantillonnage est souvent prise en charge à utiliser pour traitement de données à grande échelle [Mar *et al.* 2009] [Oliphant 2007] [Pedregosa *et al.* 2011]. Parmi les méthodes d'échantillonnage probabiliste, nous avons utilisé la méthode d'échantillonnage stratifié, car cette méthode donne toujours de meilleures performances que toutes les méthodes d'échantillonnage probabiliste [Okororie & Otuonye 2015] [Puech *et al.* 2014].

Le troisième facteur qui est le calcul parallèle et distribué est utilisé pour accélérer le temps d'exécution ; il aide à traiter des volumes importants en grande vitesse. Avec de tel calcul, il est possible d'améliorer les performances de traitement en utilisant simplement les solutions du Big Data, y compris Hadoop et son écosystème, les systèmes distribués qui supportent les algorithmes et les langages parallèles et distribués.

Les contributions de nos recherches ont donc concerné les trois points suivants :

1. **Apprentissage automatique (Algorithmes supervisés) :** Nous développons l'algorithme Random Forests (IRF) pour obtenir de meilleurs résultats que ceux obtenus par l'algorithme Random Forest traditionnel (CRF) en utilisant le principe du double élagage.

2. **Les méthodes statistiques (Échantillonnage statistique) :** Extraire une base d'apprentissage partagée par la méthode d'échantillonnage aléatoire stratifié inter-cluster, et extraire également les base d'ap-

prentissage partielles à partir du dataset original en utilisant la méthode MapReduce afin d'obtenir la base d'apprentissage représentative.

3. **Calcul parallèle et distribué :** Nous proposons une architecture distribuée portée par des nouvelles technologies qui permet d'extraire la base d'apprentissage partagée et l'échantillon représentatif. Cette architecture permet également d'accélérer le temps d'exécution au maximum (streaming) lors du traitement de grand volume de données.

1.4 Organisation de la thèse

Nous avons structuré notre thèse selon le plan décrit ci-dessous :

Chapitre 1 : Introduction générale

Le premier chapitre décrit en général notre sujet de la recherche, les problèmes auxquels nous sommes confrontés dans le domaine du Big Data, ainsi que les solutions qui seront abordées.

D'un point de vue organisationnel, le reste de la thèse s'articule autour de quatre chapitres.

Chapitre 2 : Big Data et Big Data Analytics.

Ce chapitre décrit les portes ouvertes et les difficultés engendrées par les données volumineuses, en soulignant les caractéristiques reconnues des données volumineuses et la manière de les gérer, à partir de la phase de collection de données, en passant par l'analyse de ces données, et comment les visualiser. Enfin, ce chapitre se termine par un point très important, à savoir la sécurité et l'intégrité de données massives.

Chapitre 3 : Statistique mathématique et Machine Learning pour Big Data analytics.

Dans ce chapitre, nous examinerons quelques concepts statistiques et nous présenterons les différentes méthodes d'échantillonnage statistiques, ainsi qu'une présentation complète des algorithmes d'apprentissage automatique.

Chapitre 4 : Calcul parallèle et distribué pour l'analyse du Big Data.

Dans ce chapitre, nous présentons divers systèmes et technologies qui nous permettent et qui nous aident à traiter les données volumineuses de manière plus efficace.Nous terminons ce chapitre par une synthèse de nos contributions, des limites et des perspectives ouvertes.

Chapitre 5 : Méthodologie

Dans ce chapitre, nous présenterons les travaux que nous avons proposés ainsi que les résultats obtenus, de sorte que, ces résultats sont comparés avec des travaux connexes dans le domaine de l'analyse du Big Data afin d'identifier la qualité de notre travail réalisé.

Finalement on terminera notre thèse de doctorat avec une conclusion générale et perspectives ouvertes par le travail présenté dans ce projet.

Big Data et Big Data Analytics

2

Sommaire

2.1 Introduction

Chaque seconde, nous voyons d'énormes quantités de données de croissance exponentielle, générées, acquises, stockées et analysées. Ces données sont communément appelées «Big Data» en raison de leur volume, de leur vélocité avec laquelle elles arrivent et de leur variété de formes qu'elles prennent. Le Big Data crée une nouvelle génération de gestion de données d'aide à la décision. Les entreprises reconnaissent la valeur potentielle de ces données, elles mettent en place les technologies, le personnel et les processus pour tirer parti des opportunités. L'utilisation de données analytiques est essentielle pour tirer parti des données massives. La collecte et le stockage de données volumineuses créent une grande valeur. Les décideurs de l'entreprise utilisent cette valeur pour une gestion optimale. La révolution dans la génération d'énormes quantités de données s'accompagne de l'utilisation d'Internet qui permet l'échange de données entre divers appareils électroniques et des humains. À cet égard, les domaines suivants : téléphones mobiles, médias sociaux, technologies d'imagerie permettant de déterminer un diagnostic médical, etc sont définis. Le volume de données disponibles continue de croître et augmente dans différents formats. Inversement, le prix du stockage de données continue de baisser, ce qui rend le stockage de données plus accessible. Bien que la création de stockage de données devienne moins chère et plus disponible, le volume croissant de données dans différents formats et de différentes sources crée de nouveaux problèmes en ce qui concerne le traitement des données, y compris dans l'analyse du Big Data dans les processus décisionnels de l'entreprise. Afin de stocker et de traiter le Big Data, les nouvelles technologies évoluent pour résoudre ces problèmes. Pour faire face à ces défis, une nouvelle approche, telle que la construction d'une architecture évolutive et élastique est nécessaire. Alors, quoi de neuf dans les données massives ? et en quoi elles diffèrent des petites et moyennes données traditionnelles ? Ce chapitre décrit les portes ouvertes et les difficultés engendrées par les données volumineuses, en soulignant les caractéristiques reconnues des données volumineuses et la manière de les gérer, à partir de la phase de collection de données, en passant par l'analyse de ces données, et comment les visualiser ? Enfin, ce chapitre se termine par un point très important, à savoir la sécurité et l'intégrité de données massives.

2.2 Définition du Big Data

Il existe plusieurs définitions du Big Data de différents points de vue. Par exemple, selon [Mills *et al.* 2012] le Big Data est un terme utilisé pour décrire des données volumineuses à grande vitesse et / ou à grande variété, il nécessite de nouvelles technologies et techniques pour les capturer, les stocker et les analyser ; il est également utilisé pour améliorer la prise de décision, fournir des informations et des découvertes, et soutenir et optimiser les processus. Selon [NIST 2015] le Big Data est un terme où le volume de données, la vitesse de traitement ou la représentation

des données déterminent la capacité à effectuer une analyse efficace à l'aide des approches traditionnelles, Big Data nécessite une mise à l'échelle significative (plus de nœuds) pour un traitement efficace. D'autre part [Barker & Ward 2013] définit les données massives comme un terme décrivant le stockage et l'analyse d'ensembles de données massives et / ou complexes à l'aide d'une série de techniques, notamment : NoSQL, MapReduce et Machine Learning.

En bref, le terme Big Data est un très grand ensemble de données, de sorte que les outils classiques (moteurs de gestion de données relationnelles ou analytiques) ou les outils de traitement de données (extraction et la transformation de données) ne peuvent pas traiter cette quantité de données dans les plus brefs délais (presque au rythme de traitement des données en temps réel). Tout simplement, l'ère du Big Data est en vigueur aujourd'hui parce que le monde vit une révolution numérique et informationnelle. Grâce à l'instrumentation, nous sommes en mesure de détecter plus de choses et, si nous pouvons le sentir, nous avons tendance à le stocker (au moins une partie). Grâce aux progrès de la technologie des communications, les personnes et les objets sont de plus en plus interconnectés, pas seulement de temps en temps, mais tout le temps, car ces données issues de notre monde numérique (Internet) sont produites en permanence (Internet ne dort jamais) à des vitesses, des volumes et des formats en augmentation constante et nette ; cette explosion quantitative des données numériques a forcé les principaux acteurs en ligne (Yahoo, Google, Amazon, Facebook, etc.)., dont les données sont des matières premières souvent utilisées pour l'analyse prédictive.

Les chercheurs dans le domaine Big Data restent toutefois perplexe quant à la manière d'utiliser efficacement toutes ces données. Ils cherchent de trouver un équilibre entre les deux équations pour l'analyse du Big Data ; la première équation, si le volume de données augmente, alors les algorithmes du Machine Learning donnent des résultats très précis, tandis que la deuxième équation, on espère que ces algorithmes sera capable de donner les résultats dans des délais acceptables. Peut-être à cause de ce conflit intrinsèque, de nombreux experts dans ce domaine considèrent que le Big Data non seulement apporte l'un des plus grands défis, mais aussi une opportunité des plus excitantes au cours des dix dernières années [Fan *et al.* 2012].

Le Big Data peut être classé en médias sociaux, données personnelles, données de capteurs, données transactionnelles [Japec *et al.* 2015]. Dans certains cas, les données d'enquête rapidement recueillies à l'aide d'outils techniques et en communicant un grand nombre d'unités peuvent être considérées comme des données volumineuses. D'un point de vue statistique, une quantité énorme de données pourrait être considérée comme un aspect positif pour les informations fournies par le biais de la collecte de données. Le rapport réalisé par [Dobre & Xhafa 2014] indique que le

monde produit environ un milliard de Gigaoctets et un grand pourcentage de ces données étant non organisées, également [Gantz & Reinsel 2012a] confirme que environ l'année 2020, plus de 40 Zettaoctets de données seront générés, imités et consommés. Avec cette grande quantité de données complexes et hétérogènes qui circulent de n'importe où, à tout moment et sur n'importe quel appareil, il est indéniable que l'ère du Big Data est un phénomène également connu sous le nom « déluge de données ».

Pour enrichir le terme Big Data, nous pouvons dire que les données massives vraiment contiennent une grande quantité de données, mais la qualité est une caractéristique à considérer avant de les utiliser, parce que l'utilisation efficace des données massives est très importante dans le monde dans lequel nous vivons. Aujourd'hui, nous vivons dans une société de l'information et nous évoluons vers une société fondée sur la connaissance. Afin d'obtenir de meilleures connaissances, nous avons besoin d'une plus grande quantité de données. La Société de l'information est une société où l'information joue un rôle majeur sur la scène économique, culturelle et politique. Donc le Big Data s'ajoute à la gamme de solutions que les sociétés ont mises en place pour le traitement, exploiter et diffuser des données pour les aider afin de prendre des décisions éclairées, que ce soit à des fins stratégiques ou opérationnelles, ainsi que pour atteindre à des informations précieuses soient extraites de toutes ces données afin d'améliorer la qualité de vie et de rendre notre monde meilleur.

2.3 Evolution historique du Big Data

À la fin du 17ème siècle, le monde entier a vu l'émergence de l'analyse de données statistiques, où « John Graunt » a mis au point un système d'alerte précoce pour la peste bubonique [1] qui ravage l'Europe en enregistrant des informations sur la mortalité. Au début du 19ème siècle, le monde entier a vu l'émergence des cartes perforées, où « Charles Babbage » a conçu une carte perforée, cette carte est une feuille de papier rigide contenant des informations numériques représentées par la présence ou non de trous dans des positions prédéfinies. Au milieu du 20ème siècle, le monde entier a vu l'invention de l'ordinateur électronique numérique, il a vu également l'apparition du Bande magnétique pour l'enregistrement de données, la majorité des données numériques sont stockées magnétiquement sur des disques durs des ordinateurs. En 1960, la technologie de système de fichiers a été introduite, de sorte que les données et les informations relatives sont stockées collectivement dans des formats de fichiers, le fichier est une série d'enregistrements stockés au format binaire [Marr 2015].

1. La peste bubonique : est une maladie grave (le plus souvent mortelle), hautement contagieuse.

Les études sur l'évolution du Big Data en tant que sujet de recherche et scientifique montrent que le terme "Big Data" était présent dans la recherche à partir des années 1970 [Boeth 1970] [Miller 1971].

Au début du 21ème siècle, exactement en 2001, le terme Big Data a été introduit pour la première fois dans le monde informatique par « Doug Laney » afin de définir une grande quantité de données, de sorte que les techniques traditionnelles de gestion des données ne peuvent pas gérer et traiter ces données en raison de sa taille et de sa complexité.

Pour mieux comprendre ce qu'est le Big Data et d'où il vient, il est essentiel de d'abord comprendre l'historique de stockage de données, les référentiels et les outils pour le gérer. Comme le montre la figure 2.1, le volume de données a considérablement augmenté au cours des trois dernières décennies.

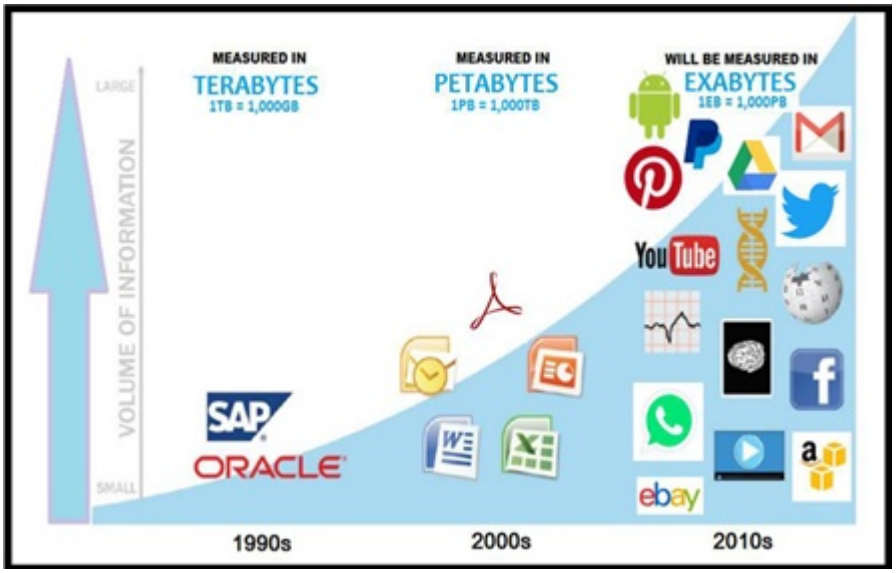

FIGURE 2.1 – *Présente l'évolution de données.*

Comme on peut le constater dans les années 90, le volume de données a été mesuré en téraoctets. Les bases de données relationnelles et les entrepôts de données représentant des données structurées en lignes et en colonnes étaient les technologies classiques pour stocker et gérer les informations d'entreprise.

La décennie suivante, la science de données a commencé à traiter différents types de sources de données basées sur la productivité et des outils de publication tels que les référentiels à gestion de contenu et les systèmes de stockage en réseau attachés. Par conséquent, le volume de données a commencé à être mesuré en pétaoctets.

À partir de l'an 2010 à ce jour, le volume de données augmente de façon exponentielle dicté par la variété et de nombreuses sources de données numérisées. Certaines applications générant beaucoup de données sont illustrées à la figure 2.1. En raison de la masse de données, il a été envisagé de commencer à mesurer les données en exaoctets, et en zéttaoctes à l'horizon 2020.

2.4 Caractéristiques du Big Data

Depuis l'apparition de l'Internet à ce jour, nous assistons à une croissance explosive du volume, de la vitesse et de la variété des données créées quotidiennement. Ces données proviennent de nombreuses sources, notamment les appareils mobiles, les capteurs, les archives individuelles, l'Internet des objets, les bases de données gouvernementales, les journaux de logiciels, les profils publics sur les réseaux sociaux, les ensembles de données commerciales, etc.

En 2001, Gartner a proposé une vue en trois dimensions (volume, variété et vélocité) concernant les défis et les opportunités liés à la croissance des données [Chen *et al.* 2014a]. En 2012, Gartner a mis à jour ce rapport comme suit : les données volumineuses sont des ressources d'information à grand volume, à grande vitesse et / ou très variées qui requièrent de nouvelles formes de traitement pour améliorer la prise de décision [Erl *et al.* 2016].

Les caractéristiques qui définissent le Big Data souvent appelées les trois V : Volume, Variété et Vélocité (comme illustré à la figure 2.2), de sorte que :

- **Volume :** Combien de données y a-t-il ?
- **Variété :** Quelle est la diversité des différents types de données ?
- **Vélocité :** À quelle vitesse les nouvelles données sont-elles générées ?

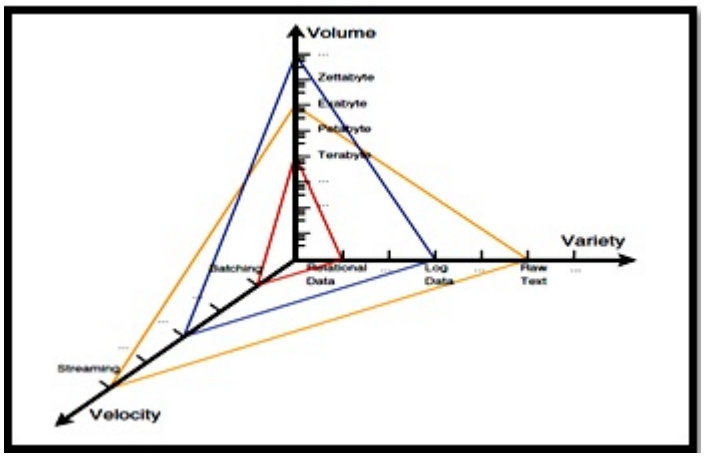

Figure 2.2 – *Les caractéristiques (3 Vs) du Big Data.*

- Volume : La première chose que tout le monde pense du Big Data est sa taille [Lyman *et al.* 2016][Eaton *et al.* 2011]. À l'ère de l'Internet, notamment les réseaux sociaux produisant des données en continu dont les volumes augmentent de manière exponentielle [Hota & Prabhu 2012] [DBTA 2013]. En 2000, huit-cent-mille (800.000) pétaoctets de données étaient stockés dans le monde [Eaton *et al.* 2012]. Nous nous attendons à ce que ce nombre atteigne trente à quarante (30-40) zettaoctets (Zo) à l'horizon 2020. Par exemple, sur Facebook, plus de cinq cent (500) téraoctets (To) de données sont créées chaque jour [Grinter 2013]. Twitter génère à lui seul plus de sept (7) téraoctets (To) de données chaque jour, et certaines entreprises génèrent des téraoctets de données chaque heure de chaque jour [Gantz & Reinsel 2012b].

- Variété : Auparavant, toutes les données nécessaires à une organisation pour exécuter ses opérations étaient des données structurées générées au sein de l'organisation, tels que les données de transaction des clients, etc. Aujourd'hui, les entreprises cherchent à exploiter beaucoup plus de données provenant d'une plus grande variété de sources, tant à l'intérieur qu'à l'extérieur de l'entreprise, tels que les documents, les contrats, les données machine, les données de capteurs, les médias sociaux, les dossiers médicaux, les courriels, etc. Mais, le problème c'est que beaucoup de ces données ne sont pas structurées ou ont une structure complexe difficile à représenter en lignes et en colonnes dans les bases de données structurées ou semi structurées [Pattnaik & Mishra 2016] [Chmidt 2012].

- Vélocité : Tout comme le volume et la variété des données que nous recueillons et stockons, la vélocité fait référence à la vitesse de génération des données et au temps nécessaire pour le traiter. Ou d'une autre manière, elle fait référence à la vitesse croissante de la génération de données, du traitement et de l'utilisation de ces données [Power 2014].

Souvent, ces caractéristiques sont complétées par un quatrième V, la véracité : dans quelle mesure les données sont-elles précises ?

- Véracité : fait référence au fait que les données doivent être crédibles, exactes, complètes et adaptées à la tâche. Étant donné que le Big Data provient de différentes sources indépendantes de la volonté d'organisations telles que les médias sociaux. La véracité est devenue un véritable problème. Les faux messages ou les spams sont très répandus, ils font de la confiance un défi majeur [Chan 2013][Roos *et al.* 2013].

Nous pouvons étendre ce modèle aux dimensions de Big Data sur dix Vs : volume, variété, vélocité, véracité, valeur, variabilité, validité, volatilité, viabilité et viscosité [Khan *et al.* 2018].

FIGURE 2.3 – *Les 10 Vs du Big Data.*

-Valeur : est un facteur majeur que toutes les organisations doivent prendre en compte lors de la mise en œuvre de données volumineuses, parce que les autres caractéristiques du Big Data n'ont pas de sens si vous ne tirez pas de valeur commerciale de ces données, Nous pouvons donc dire que la valeur aide les entreprises à mieux comprendre leurs clients [Chen *et al.* 2014b][Kayyali *et al.* 2013].

-Variabilité : La variabilité dans le contexte Big Data renvoie à différentes choses. L'un, est le nombre d'incohérences dans les données. Les données sont-elles cohérentes en termes de disponibilité ou d'intervalle de rapport ? Représente-t-il avec précision l'événement rapporté ? [Katal *et al.* 2013] Les données doivent être détectées par des méthodes de détection d'anomalies et de valeurs aberrantes afin de permettre toute analyse significative. Par conséquent, un traitement approprié de la propriété de variabilité augmente l'utilité des systèmes Big Data [Power 2014].

-Validité : Le terme fait référence à la validation des données. C-à-d, vérifier si les données utilisées sont correctes et exactes pour l'utilisation envisagée, de sorte que ces données sont donc utilisées pour évaluer la performance de la prévision [Ferguson 2013]. Nous prenons l'exemple des médias sociaux : contrairement aux sondages, les spécialistes du marché utilisent des méthodes correctes, mais n'ont en fait pas les mêmes concepts et théories. Par exemple, imaginez que la lune météorologique indique que la tempête a commencé dans un quelque part du monde, alors, comment cette tempête affecte-t-elle les individus ? Avec environ un demi-milliard d'utilisateurs, il est possible d'analyser les flux Twitter pour déterminer l'impact des tempêtes sur la population locale. Par conséquent, l'utilisation de Twitter avec des données d'un satellite météorologique pouvant aider les chercheurs à comprendre la validité des prévisions météorologiques. Cependant, l'analyse de données non structurelles issues d'un réseau social rend toujours difficile pour la prédiction fiable. Des données d'entrée correctes suivies d'un traitement de données approprié devraient

donner des résultats précis. Avec Big Data, vous devez être plus vigilant en matière de validité.

-Volatilité : La volatilité est la nature des changements brusques, instables, changés par inadvertance ou anonymement. La volatilité des données volumineuses (Big Data) fait référence à la durée de validité des données et à leur conservation [Ripon & Arif 2016]. Par exemple, certaines entreprises peuvent conserver les données et transactions les plus récentes de leurs clients, cela garantit une récupération rapide de ces informations en cas de besoin.

-Viabilité : La viabilité signifie que les données volumineuses doivent être actives très longtemps. Elle doit pouvoir de croître, d'évoluer et de produire davantage de données en cas de besoin. Nous pouvons identifier les caractéristiques et les facteurs les plus susceptibles pour prédire les résultats, de sorte que, le point le plus important pour les entreprises est de générer une valeur additive [Khan *et al.* 2018].

-Viscosité : La viscosité fait référence à la stabilité et à la résistance du flux de données volumineux. Le Big Data offre une perspective limitée en racontant une certaine narration. La viscosité mesure la résistance à l'écoulement dans le volume de données. Cette résistance peut provenir de différentes sources de données, des frictions résultant des taux d'intégration et du traitement requis pour transformer les données en informations. Les technologies permettant de traiter la viscosité incluent le traitement des événements complexes, l'intégration agile et la diffusion en continue [IBM 2014].

2.5 Structuration du Big Data

Les données volumineuses sont très diverses, car elles proviennent de différentes sources et de différents formats. Il existe de nombreuses façons de catégoriser les types de données, mais l'une des différences les plus fondamentales et les plus pertinentes réside entre les données structurées et non structurées.

Selon[Iafrate & Front 2015], environ 80 à 90% de la croissance future des données provient de types de données non structurés telles que des fichiers multimédias (vidéos, images et sons), des fichiers texte, des données géo-spatiales et financières, ce qui nécessite différentes techniques et outils pour stocker, traiter et analyser.La figure 2.4 présente quatre catégories de base.

La structuration du Big Data peut être assignée en quatre groupes :

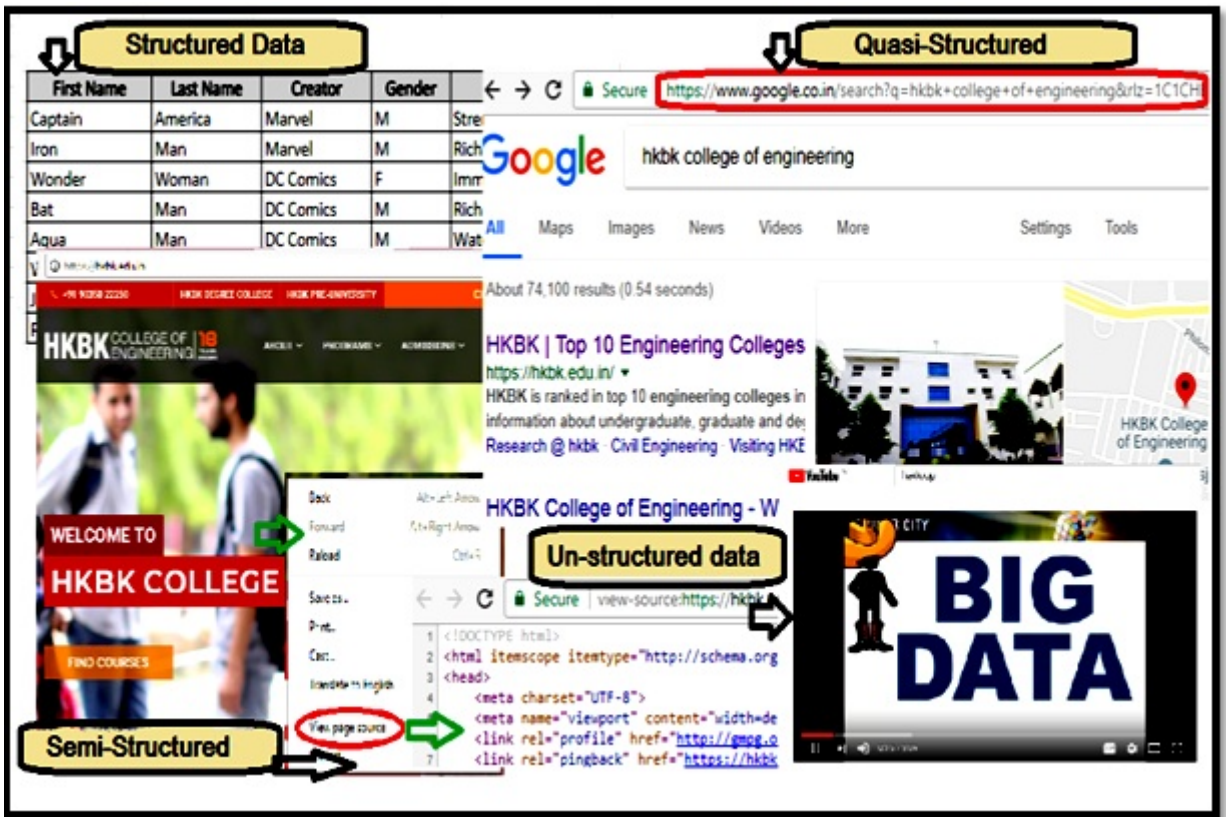

FIGURE 2.4 – *Structuration du Big Data.*

2.5.1 Données structurées

Les données structurées sont des données stockées dans une structure qui définit leur format [Hurwitz *et al.* 2013]. Les experts ont supposé que les données structurées sont comprises entre 5% et 10% de la quantité totale de données dans le monde [Kanimozhi & Venkatesan 2015]. Les données structurées sont divisées en deux catégories principales ; les données générées par la machine et les données générées par l'homme.

1. Les données structurées générées par machine incluent :

— *Les données de point de vente :* sont toutes les données transactionnelles générées chaque fois qu'un achat est effectué [Hurwitz *et al.* 2013].

— *Les données financières :* les données financières étaient principalement produites par des humains et certains d'entre eux sont encore produits ces données. La plupart des processus ont été automatisés et se déroulent sans aucune intervention humaine [Hurwitz *et al.* 2013]. Par exemple, la négociation d'actions qui contient des données structurées telles que le symbole de l'entreprise et la valeur en dollars ou euros, etc.

2. Les données structurées générées par l'homme incluent :

— *Données d'entrée :* Les humains entrent chaque jour différents types de données dans les ordinateurs et certaines de ces informations sont structurées, par exemple des noms, des âges et des adresses électroniques. Les données d'enquête qualitatives sont l'un des types de données les plus utiles. Elles peuvent être utilisées pour mieux comprendre les clients [Hurwitz *et al.* 2013].

— *Les données de flux de clics :* sont un autre type de données permettant de mieux comprendre le comportement des consommateurs. Chaque clic d'une personne en surfant sur le Web est enregistré et utilisé pour rechercher des motifs [Hurwitz *et al.* 2013].

— *Données relatives au jeu :* La même chose se produit avec chaque mouvement effectué par les clients dans un jeu vidéo. Avec la popularité croissante des jeux vidéo, ces données gagnent des volumes suffisamment importants pour être considérées comme un type séparé [Hurwitz *et al.* 2013].

En règle générale, Les données structurées peuvent être facilement stockées, traitées, analysées et interrogées à l'aide d'outils d'analyse et de logiciels traditionnels. Un exemple typique peut être considéré comme un système de gestion de base de données relationnelle (SGBDR), des données de transaction, des fichiers de données tels que des feuilles de calcul. Un autre exemple, Microsoft Excel est un excellent outil relativement simple pour travailler avec des données structurées. Des millions de lignes avec des numéros et des titres peuvent être manipulées facilement en connaissant les bonnes combinaisons de formules et de fonctions.

2.5.2 Données semi-structurées

Les données semi-structurées sont des données qui ne résident pas dans une base de données rationnelle, ou à d'autres formes de tables de données, elles peuvent contenir des balises ou d'autres marqueurs permettant de séparer les éléments sémantiques et d'appliquer des hiérarchies d'informations imbriquées, par conséquent, cette structure est également appelée structure auto-descriptive [2] .

Bien qu'il s'agisse d'un type de données semi-structurées, elles n'ont pas de structure de modèle stricte. Le meilleur exemple est celui des fichiers de données textuelles qui peuvent être analysées, tels que XML *-Extensible Markup Language-*. Ces fichiers de données se décrivent par un schéma XML. Le format JSON remplit également le même cahier des charges que le XML. XML et JSON sont utilisés pour gérer les données semi-structurées, et pour convertir les données semi-structurées en données structurées [Kanimozhi & Venkatesan 2015].

2. http ://en.wikipedia.org/wiki/Semi-structured_data (consulté le 17/10/2018)

2.5.3 Données quasi-structurées

Bien que cela ne soit pas couramment mentionné, ce groupe peut être ajouté à les trois catégories (données structurées, semi- structurées et non structurées), à savoir celui situé entre le semi-structuré et le non structuré. Ce type de données représente les données click-stream provenant de sites Web, des URL -*Uniform Resource Locators*-, des applications Web, etc. Ces données sont utilisées pour les contrats de niveau de service ou pour prévoir les atteintes à la sécurité [Hurwitz *et al.* 2013].

Les URL définit un flux de clics qui peut être analysé et exploité par les experts en données afin de découvrir les modèles d'utilisation et de mettre en évidence les relations entre les clics et les domaines d'intérêt des sites Web. Ensuite, les données quasi structurées peuvent être définies comme des données plutôt non structurées et dans un format erratique pouvant être manipulé avec des outils spéciaux. Les données de flot de données peuvent contenir des incohérences dans les valeurs et formats de données [EMCES 2015].

2.5.4 Données non structurées

Ce sont des données difficiles à classer qui en font le contraire des données structurées. Elles n'ont pas de structure inhérente. Elles représentent actuellement 90% de toutes les données dans le monde [Kanimozhi & Venkatesan 2015].

Les données non structurées peuvent également être divisées en deux catégories : Les données générées par la machine et Les données générées par l'homme [3].

1. Les données non structurées générées par la machine incluent :

— *Images satellitaires :* données météorologiques, formes de terrain, mouvements militaires.

— *Données scientifiques :* exploration pétrolière et gazière, exploration spatiale, imagerie sismique, données atmosphériques.

— *Surveillance numérique :* Photos et vidéos de surveillance.

— *Données du capteur :* capteurs de circulation, météo, océanographiques.

3. https ://www.datamation.com/big-data/structured-vs-unstructured-data.html (consulté le : 03/11/2018)

2. Les données non structurées générées par l'homme incluent :

— *Fichiers texte :* documents Word, tableurs, présentations, courriels, journaux.

— *Email :* grâce à ses métadonnées, le courrier électronique a une structure interne, Parfois, les scientifiques les appellent semi-structurés. Cependant, son champ de message est non structuré et les outils d'analyse traditionnels ne peuvent pas l'analyser.

— *Médias sociaux :* Facebook,YouTube, Twitter,Instagram LinkedIn...

— *Site Web :* Wikipedia,encyclopedia,google Map...

— *Données mobiles :* SMS,MMS...

— *Communications :* chat, messagerie instantanée, enregistrements téléphoniques, logiciel de collaboration.

— *Médias :* MP3, photos numériques, fichiers audio et vidéo.

— *Applications professionnelles :* documents MS Office, applications de productivité.

Une autre grande différence entre les données structurées et non structurées est que ces dernières ne peuvent pas être analysées via les outils et services traditionnels. Le premier obstacle est le volume de données. Il est impossible de stocker des quantités de données aussi importantes en utilisant les mêmes systèmes de stockage que pour les données structurées. Il est facile de comprendre la différence de taille : un document Excel comportant 1 000 lignes et 8 colonnes remplies d'informations à une taille 100 kilo-octets, tandis qu'une seule image au format JPEG peut facilement représenter environ 2 mégaoctets ou plus. Le problème suivant est le problème du format des données qui est très incertain, il passe d'un type de données non structuré à un autre. Excel fonctionne parfaitement avec les chiffres, mais il ne peut pas traiter les images, les vidéos, les profils Facebook et les textes longs. Parfois, il est nécessaire de traiter simultanément tous ces types de données pour obtenir un aperçu vraiment précieux.

2.6 Gestion des données massives

Dans ce chapitre, nous verrons comment générer un grand ensemble de données et comment le gérer ? Les données volumineuses sont générées à travers l'internet, tels que les réseaux sociaux, les sociétés scientifiques internationales, les organisations commerciales, ainsi que le contrôle à distance. Elles sont donc constamment échangées sur des

réseaux à grande échelle pour les traiter de manière efficace et rapide. Puisque les données volumineuses sont des données structurées ou non structurées énormes et complexes ; alors, elles sont difficiles à gérer avec des technologies traditionnelles telles que le système de gestion de base de données (SGBD).

La gestion des données volumineuses vise à garantir un niveau élevé de qualité et d'accessibilité des données pour les entreprises intelligentes, et les applications d'analyse du Big Data. Les entreprises, les agences gouvernementales et autres organisations utilisent des stratégies de gestion de données volumineuses pour les aider à faire face à la production rapide de gros volume de données. En plus la gestion efficace des données volumineuses aide les entreprises à localiser des informations précieuses dans de vastes ensembles de données provenant de diverses sources.

2.6.1 Collection de données

La collecte de données joue le rôle le plus important dans le cycle Big Data. Internet fournit des sources de données presque illimitées pour une variété de sujets. L'importance de la collecte de données dépend du type d'entreprise, mais les industries traditionnelles peuvent acquérir une source de données externes diversifiée et les combiner avec leurs données transactionnelles [4] [Benfield & Szlemko 2006].

La collecte de données est similaire à la collecte des ingrédients pour une recette. Si vous souhaitez créer un excellent plat, vous devez commencer par les bons ingrédients, ce qui signifie que vous devrez prendre une série de décisions à l'avance. Par exemple, si vous avez besoin de miel, voulez-vous un miel générique ou une variété spécifique telle que la fleur d'orange ? La marque de miel est-elle importante ? A-t-il besoin d'être cru ou biologique ? Préférez-vous obtenir le miel de votre agriculteur local ou d'un supermarché ? Et qui va avoir tous ces ingrédients ? Si vous n'avez pas le temps, êtes-vous prêt à payer quelqu'un pour l'obtenir pour vous, même si cela signifie que vous ne pourriez pas obtenir exactement ce que vous voulez ? Alors, il est préférable de collecter différents types de données, de différentes manières, et à différents endroits. Parfois, vous ne disposez pas de suffisamment de temps ni d'argent pour collecter vous-même les données dont vous avez besoin. Dans cette section, nous allons parler sur les méthodes de collecte de données pour nous aider à déterminer où et comment obtenir les meilleures informations que nous recherchons, et comment déterminer les informations dont nous avons réellement besoin.

4. La notion de donnée transactionnelle fait référence aux données que possède un commerçant sur ses clients et qui proviennent d'un comportement d'achat avéré.

2.6.1.1 Les méthodes de la collection des données

La manière de gérer le Big Data n'est pas tout à fait la même que de s'occuper d'informations classiques. Il existe diverses méthodes qui sont utilisées par les organisations, elles sont comme suit [Gill *et al.* 2008][Paradis *et al.* 2016][Ranney 2015] :

2.6.1.1.1 Observation

L'observation a pour but de recueillir des preuves de la réussite en observant la performance d'un apprenant pendant qu'il participe à une activité, mais sans nuire à son travail. L'activité peut être une situation réelle ou une situation simulée. L'observation vous permet de voir les connaissances mises en pratique, elle est mieux utilisée pour évaluer et mettre en évidence un apprentissage basé sur la citoyenneté active ou sur les compétences. Il existe deux catégories d'observations ; la première est l'observation directe, et la deuxième est l'observation indirecte. L'observation directe est effectuée personnellement par un enseignant, un moniteur ou un superviseur sur le lieu de travail. L'observation indirecte se produit lors de l'utilisation d'une technologie appropriée telle que l'enregistrement de vidéo [Mila 2018].

2.6.1.1.1.1 Observation directe

Par exemple, la fiche de suivi de l'enseignant est remplie lorsque l'étudiant commence l'activité (participation). L'enseignant enregistre ce que l'étudiant fait (les travaux à faires), et comment l'étudiant se comporte et interagit pendant la séance (assiduité). D'autre exemple, évaluation par les moniteurs, cela peut prendre la forme d'une discussion, d'une séance de questions-réponses ou de l'enregistrement d'informations sur une plateforme. Le moniteur sera un autre apprenant qui a pris part à l'activité aux côtés de l'apprenant évalué [Mila 2018]. Le moniteur notera ou fournira une rétroaction verbale sur ce que l'apprenant a fait pendant l'activité. Témoignage de témoin, il s'agit d'une déclaration d'un «tiers» qui a vu l'apprenant prendre part à l'activité. Le témoin peut être un travailleur communautaire, un superviseur de travaux ou un membre du public. Les commentaires peuvent être donnés sous forme verbale ou écrite [Kawulich 2005].

2.6.1.1.1.2 Observation indirecte

En observation indirecte, l'observateur n'a pas besoin d'être sur place et n'a même pas besoin de regarder un flux en temps réel de l'événement. Il peut être relayé par des moyens technologiques, tels qu'un événement enregistré qui est retransmis

à l'observateur à partir de transcriptions d'enregistrements audio, ou de comportements verbaux dans des environnements naturels [Anguera *et al.* 2018]. Par exemple, l'utilisation de médias sociaux, de blogs, de courriers électroniques ou d'autres supports d'archivage en ligne sont des méthodes d'observation indirecte[Johnson 2007].

2.6.1.1.2 Questionnaires

Le questionnaire fournit la technique la plus rapide et la plus simple pour rassembler des données sur des groupes d'individus dispersés dans un champ vaste et étendu. L'utilisation de questionnaires comme outil de mesure dépend du type et de la durée de l'activité. Les questions permettant de tester ou de mesurer l'apprentissage peuvent être présentées sous deux formats : interrogatoire verbal, par exemple une session de questions et réponses au début et à la fin d'une session, ou sous forme écrite, par exemple tests ou examens. Le format choisi doit être adapté à l'usage prévu selon que l'apprenant se trouve au début, au milieu ou à la fin de l'activité. Des questions peuvent être posées pour identifier les connaissances, l'expérience, les compétences et les réalisations [Kothari 2004]. Les questionnaires peuvent avoir des réponses ou non, s'il y avait des réponses, ce sont inévitablement des accords ou des désaccords. Ces réponses passent souvent par cinq niveaux allant de «fortement d'accord» à «fortement en désaccord». Rien n'est plus déroutant, frustrant et potentiellement embarrassant qu'un questionnaire mal conçu ou mal rédigé. Heureusement, avec réflexion et planification ces problèmes peuvent être facilement évités. La conception du questionnaire est un processus logique qui peut être divisé en étapes simples. Nous allons suivre les étapes suivantes pour nous aider à développer un outil valide, fiable et efficace [Robson 2002].

2.6.1.1.2.1 Déterminer les informations spécifiques nécessaires

La première étape de la conception du questionnaire consiste à examiner les objectifs, les sujets, les compétences ou les attitudes présentées dans le programme pour identifier les éléments de questionnaire potentiels. Il est parfois utile de développer ces informations sous forme de plan afin de pouvoir regrouper des questions ou des éléments connexes [Robson 2011].

2.6.1.1.2.2 Développer les questions

L'étape suivante consiste à élaborer des questions spécifiques en fonction du type de questions sélectionnées et des informations nécessaires. Les questions doivent être simples et directes afin d'éviter de dérouter les participants ou de les orienter vers la réponse souhaitée. Chaque question ne devrait porter que sur un seul pro-

blème. Si plusieurs problèmes doivent être résolus, alors nous divisons les questions en plusieurs parties ou nous développons des questions distinctes pour chaque problème, et nous évitons les termes ou expressions qui pourraient ne pas être familiers aux participants [Saunders 2012].

2.6.1.1.2.3 Tester les questions

Les questions proposées doivent être testées pour s'assurer qu'elles seront bien comprises. Idéalement, les questions devraient être testées sur un échantillon de groupes de participants. Si cela n'est pas réalisable, l'échantillon d'employés devrait se trouver à peu près au même niveau d'emploi que les participants. Nous sollicitons aussi les commentaires, les critiques et les suggestions de l'échantillon afin d'améliorer la conception du questionnaire avant qu'il ne soit administré aux participants. Nous devons faire en sorte que les questions reflètent les objectifs et le contenu du programme [Saunders 2012][Bell 2010]. Le questionnaire fournit des moyens rapides pour vérifier qu'un événement s'est produit. Il Peut être utilisé dans les domaines économiques, commerciaux, politiques et sociaux. Ainsi que, il aide à prendre les bonnes décisions immédiatement. Mais, il peut être difficile pour les personnes qui ont des difficultés de lecture et d'écriture. En plus, le modèle formel ne répond pas aux besoins des personnes utilisant des méthodes informelles.

2.6.1.1.2.4 Impliquer les parties prenantes dans le processus

Les parties prenantes telles que les clients, les sponsors, les supporters ou autres parties intéressées doivent participer autant que possible au processus de conception du questionnaire, de sorte que, nous demandons aux personnes les plus familiarisées de fournir des informations sur des points spécifiques. Souvent, les parties prenantes peuvent souhaiter contribuer à des problèmes ou à des éléments spécifiques.

2.6.1.1.2.5 Sélectionnez les types de questions

Un questionnaire peut contenir tout ou partie de ces types de questions :

— questions ouvertes permettent des réponses illimitées, les questions suivent donc un espace vide pour les réponses.
— Les listes de suivi offrent une liste d'éléments, et le participant est invité à vérifier ceux qui s'appliquent à la situation.
— Les questions à deux choix limitent les réponses à une paire de réponses alternatives, telles que oui et non.

— Les questions à choix multiples offrent plusieurs réponses possibles, et le participant est invité à sélectionner celle qui convient le mieux.
— Les échelles de classement obligent le participant à classer une liste d'éléments.

2.6.1.1.3 Interviews

L'interview est une méthode d'enquête directe. Il s'agit simplement d'un processus social dans lequel une personne connue sous le nom d'intervieweur, ce intervieweur pose des questions en face-à-face avec une ou plusieurs personnes dites interviewées pour découvrir des perspectives, des expériences, des sentiments et des idées sur un phénomène[Kothari 2004]. En outre, les interviews peuvent révéler des exemples de réussite pouvant être utilisés lors de la communication des résultats d'évaluation. Les participants peuvent hésiter à inclure leurs résultats dans un questionnaire, mais ils fourniront volontairement des informations et répondront aux questions. Nous pouvons maintenant dire que nous avons plus d'informations.

Les interviews se répartissent en deux catégories de base : les interviews structurées et les interviews non structurées. La première catégorie ressemble beaucoup à un questionnaire ; dans ce cas les intervieweurs posent des questions spécifiques avec peu de marge de manœuvre pour s'écarter des réponses souhaitées. L'un des principaux avantages des interviews structurées par rapport aux questionnaires est que ceux-ci garantissent que tous les participants répondent à toutes les questions et que l'intervieweur comprend les réponses du participant. Tandis que, la deuxième catégorie permet à l'intervieweur de rechercher des informations supplémentaires. Ce type d'interview utilise quelques questions générales qui peuvent conduire à des informations plus détaillées lorsque des données importantes sont découvertes ; dans ce cas, l'intervieweur doit être habile à poser des questions de suivi, ainsi qu'à rechercher des informations complémentaires si nécessaire [Dipboye 1994]. L'interviewé répond à ces questions, et l'intervieweur recueille diverses informations à partir de ces réponses par le biais d'une interaction sociale très saine et conviviale. Dans L'interview, nous basons sur certains points cruciaux sont les suivants :

— Nous préparons un formulaire d'entreview avec des questions correspondant à l'objectif de l'évaluation ;
— Nous utilisons des questions ouvertes et claires avec les invités ;
— Nous ne testons pas les connaissances, mais nous les explorer à travers des questions d'expérience et de description ;
— Nous ne dirigeons pas les répondants avec des questions biaisées et chargées d'hypothèses ;
— Nous enregistrons la conversation avec permission (si l'enregistrement sur bande n'est pas possible, alors nous prenons des notes abrégées).

Les principaux objectifs des interviews étaient de s'assurer que l'évaluateur identifiait toutes les personnes interrogées pertinentes pour l'évaluation, et que toutes les questions soient préparées de manière claire et concise. Certaines questions permettant d'évaluer si cet objectif a été atteint sont les suivantes :

— Avons-nous pu identifier toutes les personnes que nous devions interroger ?
— Y a-t-il des personnes qui ont été exclus des entretiens avec les personnes que nous devons inclure dans la prochaine itération ? Si oui qui ?
— Avons-nous reçu des informations précieuses lors des entretiens ?
— Y a t-il eu des personnes inclus dans les entretiens qui n'apportaient vraiment pas beaucoup de valeur au processus ? Peuvent-ils être exclus des futures interviews ? Si oui, pourquoi ?
— Nos questions étaient-elles claires et pertinentes ? Y a-t-il quelque chose que nous devons ajouter à nos questions ou changer afin que celles-ci soient plus claires ?

L'interview utilise une méthode de base pour la communication, il élimine les limitations et le caractère artificiel de la rédaction et du remplissage d'un questionnaire. Il permet de recueillir des données approfondies et détaillées. En plus, il est flexible, ouvert pour assurer le suivi. Mais, il nécessite beaucoup d'efforts et de temps.

2.6.1.1.4 Groupes de discussion

Les groupes de discussion sont utiles pour explorer les normes, les croyances, les attitudes, les pratiques et pour examiner comment les connaissances sociales sont produites, etc, de sorte que, l'animateur stimule la discussion afin d'examiner comment les connaissances et les idées se développent et fonctionnent dans un groupe donné. Le groupe est généralement composé de six à douze personnes [Nyumba *et al.* 2018].

2.6.1.1.4.1 Applications des groupes de discussion

Les groupes de discussion sont particulièrement utiles lorsque des informations qualitatives sont nécessaires sur le succès d'un programme. Par exemple, les groupes de discussion peuvent être utilisés dans les situations suivantes :

— Évaluer les réactions à des exercices, situations, simulations ou autres composants spécifiques du programme.
— Évaluer l'efficacité globale de la mise en œuvre du programme.
— Évaluer l'impact du programme par la suite.

Essentiellement, les groupes de discussion sont utiles lorsque les informations d'évaluation sont nécessaires, de sorte que, ces informations ne peuvent être collectées de manière adéquate à l'aide des questionnaires, des interviews ou des méthodes quantitatives. Par rapport aux questionnaires, aux entretiens, les groupes de discussion présentent de nombreux avantages :

— Les groupes de discussion sont très proche de formes de communication quotidienne.
— Les groupes de discussion peuvent être utilisés pour «explorer le terrain».
— Le chercheur obtient des informations sur un sujet particulier, il peut utiliser ces informations pour générer des idées et élaborer des méthodes plus structurées comme les questionnaires.

D'autre part, cette méthode présente également des inconvénients :

— Les groupes de discussion donnent des informations sur un groupe et non sur des individus, ils ne fournissent aucune information sur la fréquence ou la distribution des croyances dans la population.
— Beaucoup d'efforts et de temps sont nécessaires.

En outre, il existe d'autres méthodes utilisées pour collecter les données, nous citons certaines de ces méthodes :

2.6.1.1.5 Les documents

Il existe une quantité considérable de documents décrivant les informations des citoyens à divers niveaux d'organisation sociale. Les exemples incluent : Le registre nationale d'état civil (centre national à Alger), des registres tenus par la circulation des personnes (au niveau du wilaya DRAG), des registres qui sont trouvés au niveau des services administratifs ministériels pour suivre l'accession à la propriété, la migration, la production et la consommation d'énergie, l'emploi, la fiscalité, les hôpitaux et les écoles, des registres tenus par des organisations formelles et informelles, telles que des sociétés, des partis politiques.

2.6.1.1.6 Carte conceptuelle

La carte conceptuelle est un diagramme conçu pour clarifier la compréhension des relations entre les concepts contenus dans une zone particulière. Une liste de mots décrivant des aspects importants du sujet est compilée. Les mots de la hiérarchie sont classés du plus général au privé. Ils sont disposés de telle sorte que des termes similaires soient proches les uns des autres. Des liens sont ensuite créés entre les mots du concept et les instructions écrites pour décrire ou expliquer les liens [Faubert & Wheeldon 2009].

Il est également trouver dans ce contexte, les portefeuilles électroniques ; ce type permet de collecter de données sous forme électroniques (texte saisi, des fichiers électroniques tels que les fichiers Word, PDF, des images, des fichiers multimédias, des blogs et des liens Web, etc.) assemblées et gérées par un utilisateur, généralement en ligne. Il peut être géré de manière dynamique et en temps réel.

2.6.2 Préparation de données

La préparation des données (ou le pré-traitement des données) est la manipulation des données sous une forme appropriée pour une analyse et un traitement ultérieurs. Le but de cette phase est de fournir un jeu de données qui sera utilisé essentiellement dans l'analyse de données. Elle consiste en plusieurs tâches générales visant principalement le nettoyage et la transformation des données [Barapatre & Vijayalakshmi 2017].

2.6.2.1 Nettoyage de données

Le nettoyage des données est une étape de la préparation des données en vue de l'analyse en supprimant ou en modifiant des données incorrectes [5], incomplètes [6], non pertinentes ou répétitives, parce que ces données ne sont généralement ni nécessaires ni utiles pour l'analyse des données [Barapatre & Vijayalakshmi 2017]. Il existe plusieurs façons de nettoyer les données en fonction de la manière dont elles sont stockées avec les réponses requises.

La collecte et la saisie de données sont des processus sujets aux erreurs. Elles nécessitent souvent une intervention humaine et, comme ils ne sont que des êtres humains, ils font des fautes de frappe ou perdent leur concentration une seconde et introduisent des erreurs. Mais les données collectées par des machines ou des ordinateurs ne sont pas non plus exempt d'erreurs. Les erreurs peuvent provenir de la négligence humaine, alors que d'autres sont dues à une défaillance de la machine. Mais il faut corriger les erreurs dans le plus bref délai pour plusieurs raisons :

— Les décideurs peuvent commettre des erreurs coûteuses dans les informations en se basant sur des données incorrectes provenant d'applications qui ne corrigent pas les données erronées.
— Si les erreurs ne sont pas corrigées tôt dans le processus, le nettoyage devra être effectué pour chaque projet utilisant ces données.

5. Les données incorrectes : sont des données incohérentes qui peuvent conduire à de fausses conclusions et à de mauvaises prises de décision.

6. Les données incompletes : sont des données manquantes qui peuvent, introduire une quantité importante de biais, rendre le traitement et l'analyse des données plus laborieux, et réduire l'efficacité des méthodes statistiques.

— Les erreurs peuvent indiquer des équipements défectueux, tels que des lignes de transmission cassées et des capteurs défectueux.
— Les erreurs peuvent indiquer qu'un processus technique ne fonctionne pas comme prévu.

Les principales tâches de nettoyage des données [den Broeck *et al.* 2005] [Krause & Lipscomb 2016] comprennent :

— **Codage des valeurs manquantes :** Les données sont souvent manquantes dans certains cas. Peut-être que les données sur la famille sont collectées par la version électronique, mais les questions ne sont pas posées sur la version papier.

— **Standardisation :** La bonne pratique consiste à avoir toutes les données d'un type similaire dans un format similaire.

2.6.2.2 Transformation de données

La transformation de données est un processus permettant de convertir ou de fusionner des données d'un format ou d'une structure vers un autre format ou une autre structure [Kumar & Wu 2007]. Dans le domaine Big Data, généralement nous utilisons la technologie XML-*Extensible Markup Language*- pour convertir n'importe quel format de données en données semi-structurées, nous pouvons également les transformer en données structurées. À cette étape du processus de transformation, il devrait y avoir un bon exemple de transformation de données en format propre et bien formaté, ces données seront principalement utilisées lors de la phase d'analyse des données.

2.6.3 Analyse de données (massives)

Les données massives ne sont pas simplement massives, elles sont également rapides. Cette quantité de données est parfois créée par un grand nombre de flux constants qui envoient généralement les enregistrements de données simultanément et dans de petites tailles (ordre des kilo-octets). Le streaming de données[7] comprend une grande variété de des données telles que des données de flux de clics, des données de transaction financière, des fichiers journaux générés par des applications Web ou mobiles, des données de capteurs provenant de Internet des objets (IoT), l'activité de lecteur en jeu et la télémétrie à partir de périphériques connectés.

7. Streaming de données ou les données diffusées en continu sont des données générées en continu par des milliers de sources de données, qui envoient simultanément des enregistrements de données, et en petites quantités (quelques kilooctets).

L'analyse de données est l'outil scientifique et statistique d'analyse des données brutes permettant de rénover les informations nécessaires à l'acquisition de connaissances. L'analyse des données collabore avec les données pour formuler des décisions complexes à partir de différents points de vue afin de relever les défis du monde réel. Le rôle de l'analyse de données est d'assembler, de stocker, de traiter des données afin de mettre des méthodes empiriques dans le monde réel pour la prise de décision. Il est généralement classé en trois types : analyse descriptive, analyse prédictive et analyse prescriptive. L'analyses des données massives concernent désormais presque tous les aspects de la société moderne, tels que la fabrication, la vente en détail, les services financiers, etc.[Labrinidis & Jagadish 2012][Guerra *et al.* 2015].

2.6.3.1 Les types d'analyse de données

L'analyse de données ne peut pas être considérée comme une stratégie globale unique. En fait, ce qui distingue un scientifique ou un analyste de données de qualité supérieure des autres est sa capacité à identifier le type d'analyse pouvant être exploité au profit de l'entreprise de manière optimale. Les trois types d'analyse les plus répandus : l'analyse descriptive, prédictive et prescriptive sont des solutions interdépendantes qui aident les entreprises à tirer le meilleur parti des données dont elles disposent. Chacun de ces types d'analyse offre un aperçu différent. Dans cette partie, nous explorons les trois types d'analyses : analyse descriptive, analyse prédictive et analyse descriptive [Poornima & Pushpalatha 2016].

FIGURE 2.5 – *Les différents types de l'analyse de données.*

2.6.3.1.1 Analyse descriptive

Les techniques descriptives incluent souvent la construction de tableaux de quantiles[8], des mesures de dispersion telles que la variance ou l'écart-type, et des tableaux croisés[9] pouvant être utilisés pour examiner

8. Les quantiles sont des quantités qui peuvent être très utiles en statistique. Un quantile est obtenu à partir de la distribution de probabilité cumulée associée à une variable quantitative. On appelle généralement centile (ou percentile) le quantile ramené sur une échelle de 0 à 100.

9. Le tableau croisé est une technique courante pour étudier les relations entre les variables normales (catégoriques) ou ordinales.

de nombreuses hypothèses disparates. Ces hypothèses concernent souvent les différences observées entre les sous-groupes. Des techniques descriptives spécialisées sont utilisées pour mesurer la ségrégation [10], la discrimination et les inégalités. La discrimination est également mesurée à l'aide d'études d'audit ou de méthodes de décomposition. Une plus grande ségrégation par type ou par inégalité des résultats ne doit pas nécessairement être entièrement bonne ou mauvaise, mais elle est souvent considérée comme un marqueur de processus sociaux inéquitables. Une mesure précise des niveaux dans le temps et dans l'espace est une condition préalable à la compréhension de ces processus[Evans & Lindner 2012].

L'analyse descriptive fait exactement ce que le nom implique qu'elle «décrive», ou résume des données brutes et en fait quelque chose qui est interprétable par les humains. Elle est utilisée pour décrire les caractéristiques de base des données de l'étude. Elle limite la généralisation à un groupe particulier d'individus observés. Aucune conclusion ne s'étend au-delà de ce groupe et aucune similitude avec ceux en dehors du groupe ne peut être présumée. Les données décrivent un groupe et ce groupe seulement. Une recherche très simple implique une analyse descriptive et fournit des informations précieuses sur la nature du groupe particulier d'individus [Best & Kahn 2003].

L'analyse descriptive fournisse des résumés simples sur l'échantillon et les mesures, en parallèle à une simple analyse graphique, elle est constitue la base virtuelle de toute analyse quantitative de données. Avec l'analyse descriptive, on décrit simplement ce que les données montrent. La description des données est nécessaire pour déterminer la normalité de la distribution. La description des données est nécessaire, car la nature des techniques à appliquer pour les Statistiques inférentielles [11] des données dépend des caractéristiques des données [Bhaskar & Zulfiqar 2016]. L'analyse descriptive est utilisée pour résumer les données, telles que la moyenne, l'écart-type [12] pour les types de données continues (tels que l'âge), alors que la fréquence et le pourcentage sont utiles pour les données catégorielles (telles que le genre). L'analyse descriptive ne fournit pas de détails sur la raison pour laquelle certains événements sont survenus ni sur ce qu'il est possible de dire dans le futur [Davenport & Kim 2013].

2.6.3.1.2 L'analyse prédictive

L'analyse prédictive est un type d'analyse qui permet d'effectuer des prévisions concernant l'occurrence d'événements particuliers dans l'avenir en fonction des données du passé. L'ana-

10. La ségrégation désigne tout phénomène évolutif ou tout état de séparation de groupes ethniques ou sociaux, à l'échelle infra-urbaine, urbaine, régionale ou nationale.

11. Les statistiques inférentielles utilisent un échantillon aléatoire de données d'une population afin de décrire cette dernière et de faire des déductions à son sujet.

12. l'écart-type est une mesure de la dispersion des valeurs d'un échantillon statistique ou d'une distribution de probabilité. Il est défini comme la racine carrée de la variance ou, de manière équivalente, comme la moyenne quadratique des écarts par rapport à la moyenne.

lyse prédictive est largement intégrée dans les organisations décisionnelles[Davenport & Kim 2013][Brydon & Gemino 2008].

L'analyse prédictive est l'art de construire et d'utiliser des modèles qui font des prédictions basées sur des modèles extraits de données historiques. Cela inclut des modèles prédictifs empiriques (des modèles statistiques tels que les algorithmes d'exploration de données) qui prédisent des scénarios futurs et des méthodes d'évaluation permettant d'évaluer le pouvoir prédictif d'un modèle [Schmueli & Koppius 2011]. L'analyse prédictive identifie les relations entre les variables et ensuite, en fonction de ces relations, elle prédit la probabilité qu'un certain événement se produise. Bien que l'analyse prédictive repose sur des relations fortes entre les données, on peut s'attendre à des relations parfois mal définies[Evans & Lindner 2012][Davenport & Kim 2013]. Les applications de l'analyse prédictive de données incluent par exemple, la prédiction des prix comme les achats des billets en ligne. Des modèles d'analyse prédictive peuvent être formés pour prévoir des prix optimaux en fonction des historiques de vente. Les entreprises peuvent ensuite utiliser ces prévisions pour entrer dans leurs décisions en matière de stratégie de tarification[Etzioni *et al.* 2003].Elles incluent aussi la prédiction de dosage, c'est que les médecins décident souvent de la quantité d'un médicament ou d'un autre produit chimique à inclure dans un traitement. Les modèles d'analyse prédictive peuvent être utilisés pour faciliter cette prise de décision en prédisant les doses optimales en fonction des données sur les doses passées et les résultats associés [Nguyen *et al.* 2019][Coulet *et al.* 2018]. Les modèles d'analyse prédictive peuvent être utilisés pour prédire le risque associé aux décisions telles que l'octroi d'un prêt. Ces modèles sont formés à l'aide de données historiques. Les organisations peuvent utiliser les résultats des modèles de prédiction des risques pour mieux juger les risques [Rahul & Pravin 2016]. Dans le domaine commercial, la plupart des décisions seraient beaucoup plus faciles si nous pouvions prédire la probabilité, ou la propension de clients pour prendre des mesures différentes. Les applications réussies de la modélisation de la propension incluent la prévision de la probabilité que les clients quittent un opérateur de téléphonie mobile pour un autre, elles répondent à des efforts de marketing particuliers ou achètent différents produits[Davenport & Kim 2013]. Les modèles d'analyse prédictive peuvent aider les professionnels à établir de meilleurs diagnostics en exploitant de vastes collections d'exemples historiques à une échelle au-delà de tout ce qu'un individu pourrait voir au cours de sa carrière. Les diagnostics posés par les modèles d'analyse prédictive deviennent généralement un élément du processus de diagnostic existant du professionnel [Foster *et al.* 2014][Ghorbani & Ghousi 2019]. L'analyse prédictive des données peut être utilisée pour classer automatiquement les documents en différentes catégories. Les exemples incluent également le filtrage du courrier électronique, l'analyse de l'opinion, la redirection des plaintes des clients et la prise de décision médicale. En fait, la définition d'un document peut être étendue pour inclure des images, des sons et des vidéos qui peuvent tous être classés à l'aide de modèles d'analyse prédictive de données[Khan *et al.* 2014].

Tous ces exemples ont deux choses en commun. Premièrement,

dans chaque cas, un modèle est utilisé pour établir une prédiction afin d'aider une personne ou une organisation à prendre une décision. Dans l'analyse de données prédictive, nous utilisons une définition large du mot prédiction. Dans la pratique courante, le mot prédiction a un aspect temporel : nous prédisons ce qui se passera dans le futur. Cependant, dans l'analyse de données, une prédiction est l'affectation d'une valeur à une variable inconnue. Il peut s'agir de prédire le prix auquel un produit sera vendu à l'avenir ou, au contraire, de prévoir le type de document. Ainsi, dans certains cas, la prédiction a un aspect temporel mais pas du tout. La deuxième chose que les exemples énumérés ci-dessus ont en commun est qu'un modèle a été créée pour permettre des prédictions basées sur un ensemble d'exemples historiques. Dans notre travail, nous utilisons les algorithmes du Machine Learning pour former ces modèles.

2. 6.3.1.2 .1 Les six étapes clés de l'analyse prédictive

L'étude met en exergue un cycle de six étapes clés dans l'élaboration de solutions prédictives grâce au Big Data, ces étapes [Eckerson 2007][Kelleher *et al.* 2015][Rose *et al.* 2017] sont illustrées dans la figure ci-dessous :

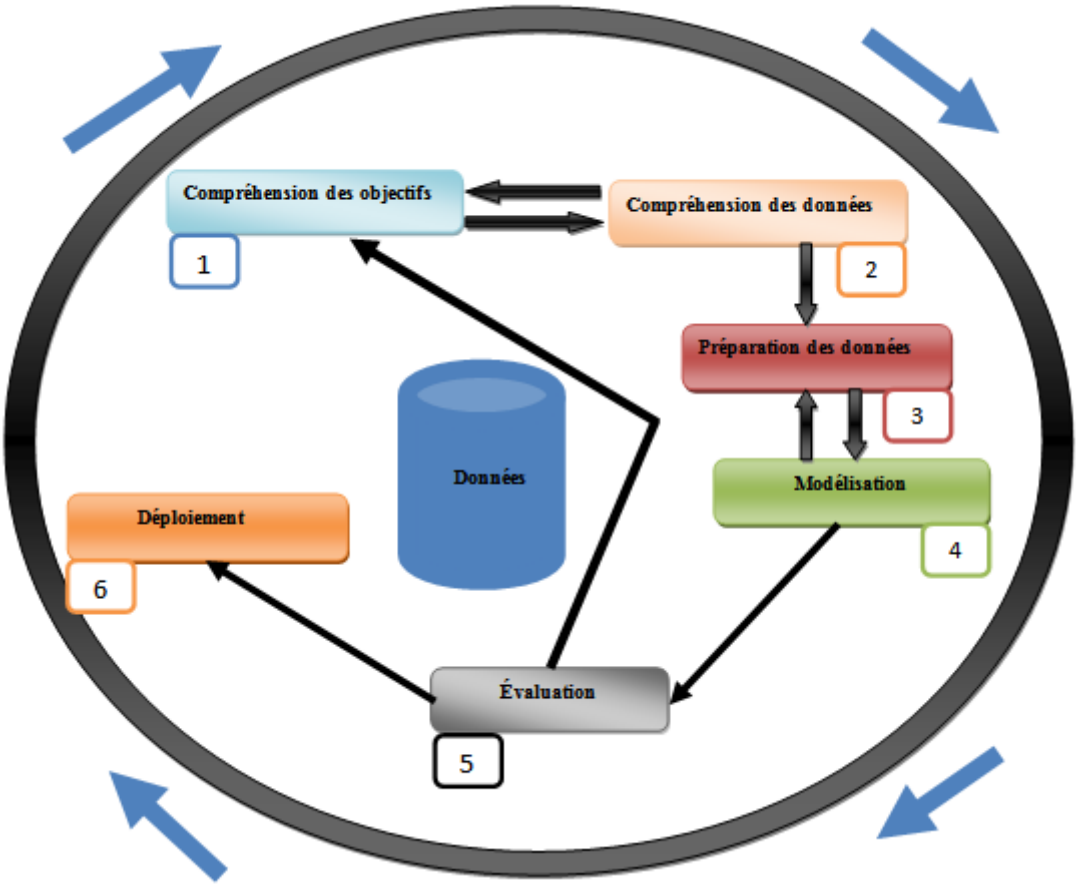

FIGURE 2.6 – *Les six étapes de l'analyse prédictive.*

1. **Compréhension des objectifs** ***(Définir le projet)*** **:** Tout d'abord, il est nécessaire de bien comprendre les objectifs de l'entreprise et de déterminer ses besoins. Ensuite, nous évaluons la situation actuelle en recherchant les ressources, hypothèses, contraintes et autres facteurs importants à prendre en compte. Ensuite, à partir des objectifs commerciaux et des situations actuelles, nous créons des objectifs d'exploration de données pour atteindre les objectifs commerciaux dans la situation actuelle. Enfin, un bon plan d'exploration de données doit être mis en place pour atteindre à la fois les objectifs commerciaux et l'exploration de données. Le plan devrait être aussi détaillé que possible.

2. **Compréhension des données** ***(Identifier les données utiles en évaluant diverses sources possibles)*** **:** Une fois que nous avons déterminé la manière dont les analyses de données prédictives seront utilisées pour résoudre le problème du travail, il est important que l'analyseur de données comprenne parfaitement les différentes sources de données disponibles au sein de l'organisation et les différents types de données contenues dans ces sources.

3. **Préparation des données** ***(Triturer les données, les agréger et les compléter)*** **:** Les modèles d'analyse de données prédictive nécessitent des types de données spécifiques. Cette phase inclut toutes les activités nécessaires à la conversion des différentes sources de données disponibles dans l'organisation, de sorte que ces données soient bien formées et permettent d'utiliser des modèles d'apprentissage automatique.

4. **Modélisation** ***(Construire un modèle prédictif à partir d'algorithmes statistiques et du Machine-Learning)*** **:** La modélisation se produit lorsque l'apprentissage est terminé. Différents algorithmes d'apprentissage automatique sont utilisés pour créer un ensemble de modèles prédictifs à partir desquels le meilleur modèle sera sélectionné pour le déploiement.

5. **Évaluation** ***(Evaluer l'efficacité et la précision du modèle prédictif)*** **:** Avant de pouvoir déployer des modèles destinés à être utilisés au sein d'une organisation, il est important qu'ils soient entièrement évalués et qu'ils se révèlent aptes à cet usage. Cette phase couvre toutes les tâches d'évaluation requises pour montrer qu'un modèle sera capable de faire des prédiction précises après son déploiement et qu'il ne souffre pas de sur-apprentissage ni de sous-apprentissage.

6. **Déploiement *(Utiliser le modèle prédictif pour orienter des décisions métiers)* :** Les modèles d'apprentissage automatique sont conçus pour servir une fonction au sein d'une organisation. Cette phase couvre tout le travail à effectuer pour intégrer avec succès un modèle d'apprentissage automatique dans les processus d'une organisation.

Nous pouvons ajouter une autre étape est la maintenance, cette étape assure l'efficacité du modèle prédictif, parce que la réalité n'est pas statique et donc les données aussi. Un modèle peut être valable pendant une certaine période alors que les conditions externes ne changent pas de manière significative. Il est judicieux de revoir périodiquement les modèles et de les tester avec de nouvelles données pour vous assurer qu'ils n'ont pas perdu de leur signification. Quand le modèle devient moins précis, donc il est nécessaire d'ajuster le modèle à nouveau, par exemple en ajustant les paramètres des algorithmes, ou trouver des données supplémentaires.

2. 6.3.1.3 L'analyse prescriptive

L'analyse prescriptive est l'un des grands mots à la mode de ces dernières années. Pouvoir prescrire automatiquement des actions en vue d'atteindre un objectif signifierait un énorme pas en avant dans l'aide ou la prise de décision automatique pour tous les domaines. Elle suggère la meilleure marche à suivre pour optimiser les résultats de votre entreprise. En règle générale, l'analyse prescriptive associe un modèle prédictif à des règles commerciales (telles que le refus d'une transaction si la probabilité de fraude est supérieure à un seuil donné). Cette technique est utilisée pour appuyer plus efficacement la prise de décision sur la base d'idées diverses émanant par les décideurs tels que les directeurs techniques et les directeurs généraux analysent et prédisent des situations complexes. Donc, elle est considérée comme nirvana dans l'analyse, elle est souvent utilisée par les organisations les plus sophistiquées sur le plan analytique [Gim *et al.* 2018].

L'analyse prescriptive, quant à elle, consiste à suggérer un certain nombre d'actions, elle inclut des méthodes de conception et d'optimisation expérimentales. Le plan expérimental explique les raisons pour lesquelles un phénomène se produit en effectuant des expériences au cours desquelles des variables indépendantes sont manipulées, des variables superflues sont contrôlées, et par conséquent, des conclusions sont tirées des actions qui doivent être prises par le décideur.

2.6.4 Visualisation de données

La visualisation des données explique l'importance des données en les plaçant en termes de contexte visuel [Olshannikova *et al.* 2015]. C'est une représentation visuelle des données. La visualisation des données permet à l'utilisateur d'acquérir plus de connaissances sur les données brutes collectées à partir de diverses sources.

2.6.4.1 La visualisation de données, dans le passé

La visualisation des données n'est pas nouvelle. La communication visuelle des données existe sous des formes diverses depuis des centaines, voire des milliers d'années.

Abo El-Rayhan Mohamed Ibn Ahmed El-Bayrouni, né à Khouarizm, est l'un des plus anciens chercheurs de représentation circulaire des phases de la lune. El-Bayrouni était connu dans le monde islamique et était l'un des érudits les plus acclamés de son temps. Al-Bayrouni était connu par son livre principal intitulé "Le canon Mas'udi", concernant à l'astronomie, à la géographie et à l'ingénierie, l'une de ces pages figurante dans la figure ci-dessous :

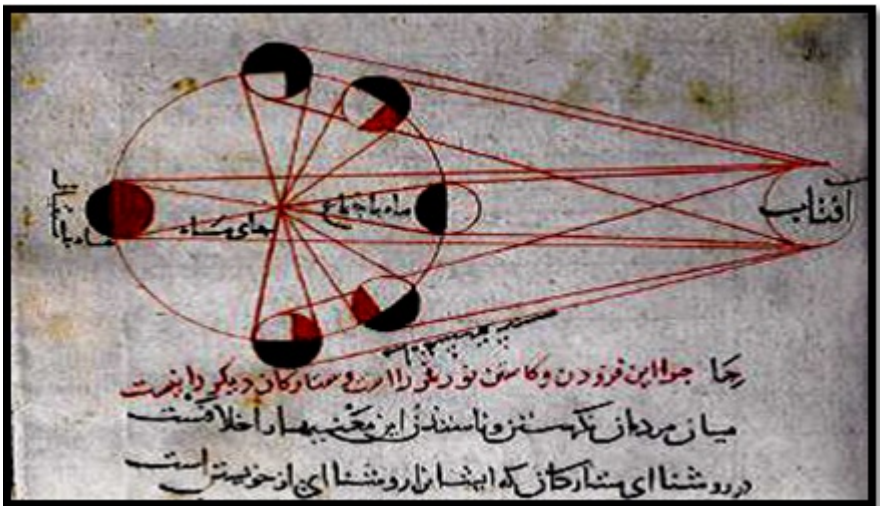

FIGURE 2.7 – *La visualisation de données dans le passé.*

Au dix-huitième siècle, les styles communs dominaient encore les conseils d'administration des entreprises dans tout le pays notamment les barres et diagrammes circulaires.

L'appétit pour la visualisation de données a considérablement augmenté ces dernières années. La visualisation de données est devenue un courant de conscience général au cours de la dernière décennie, elle est catalysée par de nouvelles capacités technologiques puissantes, ainsi que par un changement de culture vers une plus grande transparence et une

accessibilité accrue des données, le domaine a connu une croissance rapide de la participation enthousiaste. Alors que la pratique de la visualisation de données était autrefois l'apanage de statisticiens, d'ingénieurs et des experts spécialistes, le champ de la mondialisation qui existe aujourd'hui constitue une communauté très active, informée, inclusive et innovante de praticiens qui font avancer le métier dans des directions fascinantes.

2.6.4.2 La visualisation de données (massives) aujourd'hui

La visualisation de données massives permet de visualiser un grand ensemble de données dans un format facilement compréhensible par n'importe quel utilisateur. En plus la présentation des données est simple et directe, ainsi que les utilisateurs comprendront facilement le message[Alhadad 2018]. En termes simples, la visualisation des données est la représentation des informations sous une forme graphique comme illustrée dans la figure ci-dessous :

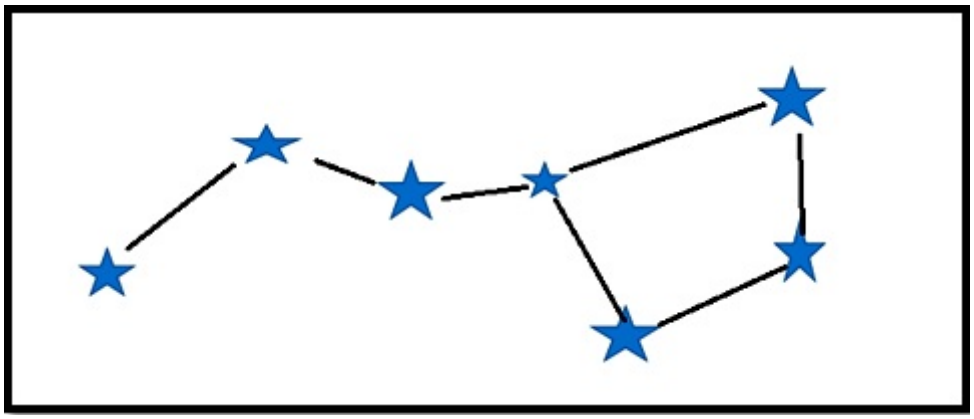

FIGURE 2.8 – *La visualisation de données aujourd'hui.*

Un exemple simple qui peut être utilisé pour définir la visualisation de données, on peut tracer des lignes entre des étoiles dans le ciel nocturne pour créer une image. Imaginez certaines étoiles comme des points de données qui vous intéressent (parmi les milliards d'autres étoiles visibles dans le ciel) et connectez-les dans un certain ordre pour créer une image permettant de visualiser la constellation. Actuellement, on dit dans l'industrie que de nombreuses disciplines considèrent la visualisation de données comme un équivalent moderne de la communication visuelle. Bon, alors quel est l'intérêt ou l'objectif principal de la communication visuelle ou de la visualisation de vos données ?

Lorsque vous utilisez la visualisation de données, l'objectif principal (bien qu'il y ait d'autres objectifs) est de faire en sorte que quelque chose compliquée paraisse simple. En regardant des tableaux et des graphiques sont généralement beaucoup plus facile pour les yeux. De plus, le fait est

que nous sommes, les humains, capables de traiter même de très grandes quantités de données beaucoup plus rapidement lorsque les données sont présentées sous forme graphique. Par conséquent, la visualisation de données est un moyen de transmettre des concepts de manière universelle permettant à votre public cible ou à votre cible d'obtenir rapidement votre point de vue.

Une visualisation bien conçue permet à nos yeux de discerner rapidement les schémas de données, ce qui nous permet de mieux comprendre les caractéristiques et les phénomènes sous-jacents de nos données. L'inspection visuelle nous amène à de nouvelles connaissances sur nos données en nous aidant à formuler des hypothèses sur les prochaines priorités. L'interaction nous permet d'approfondir ces hypothèses, soit en montrant d'autres parties des données, soit en montrant les mêmes données sous un angle différent. Ces fonctionnalités font de la visualisation des données un outil précieux pour l'exploration, l'analyse et la communication. La visualisation de données est largement utilisée dans les entreprises pour communiquer leurs données, et pour approfondir leur compréhension des données [AStephen 2015][Padilla *et al.* 2018].

La visualisation des données peut guider la prise de décision, elle devient un outil pour transmettre des informations essentielles à toute analyse de données [Kinkeldey *et al.* 2017][Kinkeldey *et al.* 2014]. Cependant, pour être réellement exploitables, la visualisation de données doit contenir la bonne quantité d'interactivité. Alors que le volume de données continue d'augmenter, d'une part, un nombre croissant de fournisseurs et de communautés développent des outils permettant de créer des graphiques clairs et percutants à utiliser dans les présentations et les applications. Et d'autre part, de nombreuses entreprises estiment que la compréhension de ces données nécessite l'utilisation d'une forme de visualisation des données [Padilla *et al.* 2018], parce qu'il est pratiquement impossible de visualiser un million de lignes de données et d'en comprendre le sens.

2.6.4.3 Les concepts traditionnels de la visualisation des données

Tout d'abord, nous clarifions ce que nous voulons dire quand nous disons le mot "traditionnel", c-à-d, nous nous référons aux idées et aux méthodes qui ont été utilisées avec un certain succès dans visualisation de données au fil du temps. En outre, la plupart des systèmes de visualisation traditionnels ne peuvent pas gérer la taille de nombreux jeux de données contemporains. Ils se limitent à traiter avec des ensembles de données de petite taille qui peuvent être facilement manipulés et analysés avec les techniques classiques de gestion de données. Bien qu'il semble que chaque jour, de nouvelles technologies et de nouvelles pratiques soient découvertes, développées et déployées offrant des options nouvelles et différentes. La visualisation des données en temps réel est toujours plus

ingénieuse, en plus la compréhension des concepts de base pour la visualisation des données reste également toujours indispensable [Bikakis 2018].

À ce point, il est essentiel de bien comprendre comment choisir la méthode de visualisation la plus appropriée ou la plus efficace. Pour faire ce choix, nous devons généralement répondre aux questions suivantes :

— Quel est le volume des données à visualiser ?
— Qu'est-ce que nous essayons de communiquer ?
— Quel est le point que nous souhaitons communiquer ?
— Qui est notre public ? Qui va consommer cette information ?
— Quel type de visualisation de données pourrait le mieux transmettre notre message à notre public ?

Nous avons également été réalistes sur le fait que parfois l'approche ou la méthode utilisée est uniquement basée sur votre temps et votre budget. Et vous connaissez probablement déjà ces méthodes de visualisation les plus courants incluent :

— Les tableaux.
— Les histogrammes.
— Les nuages de points.
— Les diagrammes à courbes, à barres, à secteurs, à aires, à flux et à bulles.
— Les séries de données ou les combinaisons de graphiques.
— Les chronologies.
— Les diagrammes de flux de données.

2.6.4.4 Visualisation interactive des données

La visualisation de données est une espèce intéressante. Les scientifiques savent souvent tout sur les données, mais les problèmes liés à la visualisation peuvent être difficiles pour eux. Les concepteurs d'interface utilisateur et les graphistes dominent les aspects visuels, mais le traitement des données n'est pas à leur portée. La visualisation de données permet aux concepteurs d'interface utilisateur de se familiariser avec les éléments inconnus. La visualisation de données est donc à la fois une science et un art [Dormehl 2014].

Les entreprises qui peuvent analyser les données en temps réel à partir de nombreuses formes de données présentent un grand avantage. Par exemple, elles peuvent connaitre les sentiments de leurs clients en consultant leurs habitudes achats. Les entreprises qui exploitent leurs données disponibles en les intégrant dans des visualisations de données interactives bénéficient des avantages suivants :

— Les utilisateurs peuvent manipuler les données pour trouver des informations spécifiques dont ils ont besoin.

— Les utilisateurs peuvent être alertés des situations nécessitant une attention immédiate.
— Lorsque tous les membres d'une équipe examinent les mêmes données, ils peuvent résoudre les problèmes plus facilement.
— Les utilisateurs ne présentent que les éléments clés qui leur permettent d'obtenir à la fois une vue d'ensemble et les détails dans une seule visualisation.
— Les utilisateurs peuvent obtenir d'importantes révélations sur les performances de l'entreprise à partir d'une bonne visualisation interactive.

La visualisation de données d'aujourd'hui est passée à des présentations interactives basées sur le Web. L'augmentation des données a également conduit à des présentations de plus en plus visuelles. Les interfaces visuelles font aujourd'hui face à des défis difficiles, car l'analyse doit être faite de manière visuelle. En conséquence, un nombre croissant d'interfaces utilisateur apparaîtra sous forme tableau de bord. Ces interfaces utilisateur ont un objectif visuel et les données doivent être ajustées à la volée [Koppal 2017].

2.6.4.5 La méthodologie de visualisation de données

L'adoption de cette méthodologie consiste à reconnaître les étapes clés, les considérations et les tactiques [Chittaro 2006] [Guanghui *et al.* 2015] [Blascheck & Ertl 2013] [Colin 2004] [Plaisant 2004] qui nous aiderons à naviguer sans heurts dans un projet de visualisation sont les suivantes :

— La première étape du processus de visualisation est appelée la cartographie . La cartographie signifie comment visualiser des informations ou comment coder des informations sous une forme visuelle. Dans la cartographie, les données ou les informations se transforment en une forme graphique sous l'hypothèse de caractéristiques visuelles. Une bonne cartographie produit une bonne représentation visuelle.

— La deuxième étape du processus de visualisation s'appelle la sélection . Rappelez-vous, la conception est rarement un processus net. En effet, certaines étapes peuvent parfois basculer en séquence et nécessiter des itérations. Il est naturel que de nouveaux facteurs apparaissent à tout moment sont influent sur les solutions alternatives. Il est donc important d'être ouvert d'esprit et flexible. Il peut être nécessaire de revoir les choses, inverser les décisions et modifier les directives. Ce que nous essayons de faire dans la mesure du possible est de trouver le meilleur chemin de choix de conception. Alors, la sélection signifie sélectionner des données parmi celles disponibles en fonction de la tâche ou du travail donné. La sélection des données dépend directement de

l'objectif consistant à obtenir des graphiques ou des représentations graphiques. Cette tâche dans le processus est la tâche la plus importante, car la sélection de données erronées induit l'utilisateur en erreur afin qu'il prenne des décisions cruciales qu'il subisse des pertes énormes (finances, temps, etc.,..). Pour cela, il doit éviter d'insérer des données inutiles.

— La troisième phase du processus est la présentation . Dans la visualisation, la présentation en perspective signifie comment gérer et comment organiser efficacement les informations dans l'espace disponible à l'écran. Après une cartographie intuitive et une sélection claire, il est vraiment important de présenter ces informations sous une forme plus compréhensible. Les défis de présentation liés à la visualisation des données sont principalement liés à la technologie ; cette étape nécessite généralement l'assistance d'une variété d'applications et de programmes.

— La quatrième étape du processus de visualisation s'appelle l'interactivité . L'interactivité signifie quelles sont les installations fournies pour organiser, explorer et réorganiser la visualisation. L'interactivité conviviale permet à l'utilisateur d'explorer, de comprendre et d'interpréter au mieux les données ou informations, ce qui améliore ses capacités d'exploration. Certains peuvent se sentir mal à l'aise de suivre un processus de l'interactivité, car il est itératif et créatif. Mais, nous voudrons dire que tout le monde devrait bénéficier de travailler de manière plus organisée et séquencée, en particulier si cela contribue à réduire l'inefficacité et le gaspillage de ressources.

— Le facteur humain est la cinquième étape de visualisation à prendre en compte. Le facteur humain implique deux grandes catégories, la facilité d'utilisation et les facteurs d'accessibilité. La visualisation est facile à utiliser pour n'importe quel utilisateur. La connaissance de la visualisation visuelle et les aspects cognitifs rendent très facile la conception d'une visualisation efficace. Ces facteurs sont la pratique courante de l'interaction homme-machine. En plus, une remarque plus importante, le contenu de la méthodologie doit éviter tout sens idéologique et dogmatique, nous préférons se concentrer uniquement sur des directives. Cette étape doit être adoptée de manière flexible en fonction de notre propre jugement et de notre discrétion.

— Après la création d'une interface de visualisation utilisable, la dernière étape consiste à évaluer le formulaire visuel créé. L'évaluation est également importante pour déterminer si la méthode de visualisation est efficace ou non et/ou si l'objectif est atteint ou non. Un autre point important, si la visualisation de données n'était pas déjà claire, donc elle n'est pas une science exacte. Dans ce cas, Il existe rarement, voire jamais, une bonne réponse ou une meilleure solution.

2.6.4.5.1 Conseils très importants en méthodologie de visualisation

— La visualisation de données devrait être attrayante. L'avènement d'outils de création visuelle plus sophistiqués et en grandes qualités ont par exemple placé au niveau des applications mobiles. Cela ne fera que s'accroître avec l'évolution de technologies.

— Nous garantissons que notre visualisation de données est construite sur un système évolutif et accessible pour la maintenance et les modifications futures. Parce que, si notre conception de la visualisation de données est réussie, d'autres voudront l'utiliser et l'exploiter.

— L'utilisateur doit obtenir les informations correctes. Parce que c'est un problème lorsque les utilisateurs se concentrent sur la visualisation ou sur une fonctionnalité particulière dont ils n'ont pas vraiment besoin.

— Avant de créer une visualisation, nous définissons exactement leur utilisation en self-service, en analyse approfondie ou en présentation générale.

— La visualisation de données doit être accessible et facile à utiliser, elle peut être modifiée facilement si nécessaire. De plus, les données doivent être accessibles sur n'importe quel appareil, à tout moment et n'importe où. Cette fonctionnalité est essentielle à l'adoption par l'utilisateur.

2.6.5 Sécurité et intégrité de données massives

Les entreprises sont toujours inquiètes par la sécurité et l'intégrité de leurs données, mais ne sont pas préparées aux complexités présentées par la gestion de données massives. La sécurité et l'intégrité sont des choses incontournables, car les programmes malveillants évoluent constamment. Cette stratégie de sécurité doit être associée à une stratégie d'intégrité. La combinaison de la sécurité et de l'intégrité assurera la responsabilité de toutes les parties impliquées dans le déploiement de notre gestion de données massives. La gestion de la sécurité et l'intégrité des données doit être considérée comme une responsabilité partagée dans l'ensemble de l'organisation. Vous pouvez mettre en œuvre les derniers contrôles techniques de sécurité et d'intégrité et faire face aux risques si vos utilisateurs finaux ne comprennent pas clairement leur rôle en matière de protection de toutes les données avec lesquelles ils travaillent. Voulez-vous que tout le monde ait accès à toutes vos données ? Vous devez connaître la nature de cette source de données. Les données ont-elles été vérifiées ? Est-elle sécurisée et contrôlée contre les intrusions ?

Les outils de sécurité des données massives nous permettent d'avoir un contrôle total sur ces données, elle est devenue un sujet en soi. Les chercheurs spécialisés dans ce domaine sont confrontés à de nombreux problèmes dans la sécurité et l'intégrité, mais ils invitent souvent d'autres experts dans d'autres domaines pour les aider[Moreno *et al.* 2016]. Par exemple, les sites de réseaux sociaux les plus réputés suivront de près les comportements malveillants des comptes utilisateurs ; s'ils existent alors ces comptes doivent être supprimés avant qu'ils puissent causer des dommages, cela nécessite un niveau sophistiqué pour la sécurité et l'intégrité des données massives[Ofori & Islam 2019], ce que beaucoup de sites ne peuvent pas le faire. Un autre exemple, si votre organisation a découvert un site merveilleux, mais ce site a été piraté et vous avez sélectionné ces données dans le cadre de votre plate-forme Big Data. Les conséquences peuvent être graves. Toutes les menaces à la sécurité ne sont pas délibérées. Vous ne souhaitez pas incorporer une source de données volumineuses contenant des informations sensibles personnellement identifiables qui pourraient nuire à la réputation de vos clients et à celle de votre entreprise.

Le Big Data devient critique pour les dirigeants d'entreprise qui essaient de comprendre l'orientation des nouveaux produits et les exigences des clients ou de comprendre de leur environnement global. Toutefois, si les données provenant de diverses sources introduisent des risques de sécurité et d'intégrité dans l'entreprise, des conséquences imprévues peuvent mettre l'entreprise en danger. Nous avons beaucoup de travail à faire pour comprendre la sécurité et l'intégrité, car ces deux cibles évoluent constamment. En plus, la formation des personnels est essentielle pour que chaque membre de l'organisation comprenne ses rôles et ses responsabilités en matière de sécurité, cette définition correspond à l'approche proposée par *Clark* et *Wilson* pour prévenir la fraude et les erreurs [Liu & Xu 2009]. Alors que, la sécurité et l'intégrité sont des problèmes qui concernent l'ensemble de l'entreprise ; quelques conseils et directives devraient être suivies. Par exemple, si vous collectez des données à partir de sources de données non structurées telles que des sites de médias sociaux, vous devez vous assurer que les virus ou les liens factices ne sont pas enfouis dans le contenu. Si vous prenez ces données et en faites une partie de votre système d'analyse, vous pourriez mettre votre entreprise en danger. De plus, gardez à l'esprit la source originale de ces données. Une source de données non structurée (contenant des commentaires intéressants sur le type de client que vous essayez de comprendre) peut inclure du bruit parasite. En plus, Certains experts estiment que différents types de données nécessitent différentes formes de protection. Dans certains cas, le chiffrement des données est obligatoire comme dans le cloud computing. Par exemple, vous pouvez chiffrer vos données sur votre disque dur, puis les envoyer et les stocker dans un cloud computing ; parce que le stockage des données volumineuses sur le cloud constitue une préoccupation majeure en matière de sécurité des données [Priyadharshini & Parvathi 2012]. Tout chiffrer de manière complète réduit votre exposition aux attaques ; Cependant, le cryptage constitue une perte de performances. Par exemple, de nombreux experts conseillent de gérer vos propres clés plutôt que de lais-

ser un fournisseur d'informatique en cloud le faire, ce qui peut devenir compliqué. Garder la trace de trop de clés peut être un cauchemar. Mais, le chiffrement peut créer d'autres problèmes. Par exemple, si vous essayez de chiffrer des données dans une base de données, vous devrez examiner les données au fur et à mesure de leur déplacement (chiffrement point à point [13]) et de leur stockage dans la base de données. Cette procédure peut être coûteuse et compliquée. De même, même si vous pensez que tout est crypté et que vous êtes en sécurité, cela n'est peut-être pas le cas. En revanche, Les algorithmes de chiffrement des données restent une solution prometteuse, et ils offrent des avantages plus précieux [Popa *et al.* 2011] [Poddar *et al.* 2016] [Arasu *et al.* 2013] [Kaashoek *et al.* 2013].

Conclusion

La disponibilité de données massives, de matériel de base peu coûteux et de nouveaux logiciels d'analyse de données ont créé un moment unique dans l'histoire de l'analyse des données massives. Ces données massives nécessitent le développement de techniques pouvant être utilisées pour faciliter leur analyse. La convergence de ces tendances signifie que nous disposons des capacités nécessaires pour analyser des ensembles de données étonnants rapidement et de manière rentable pour la première fois de l'histoire. Ces capacités ne sont ni théoriques ni triviales. Elles représentent un véritable bond en avant et une occasion claire de réaliser le succès à travers la meilleure prise de décision. Précédemment, avec de faibles volumes de données, les institutions décisionnelles n'ont pas eu beaucoup de succès. En conséquence, la prise de décision basée sur les données massives est devenue plus courante pour assurer un chemin raisonnable vers le succès. Cette situation est logique car il est facile de voir que les données ne diminuent pas mais augmentent. Les entreprises tirent maintenant parti de l'importance de la prise de décision en définissant les concepts de données volumineuses ainsi que leurs analyses. Il a été dit le plus souvent que l'utilisation de ces données massives avec les algorithmes d'apprentissage automatique donne de meilleurs résultats de prédiction. Il est également possible de trouver ces mêmes résultats si on prend des échantillons à partir du data-set original en utilisant des méthodes statistiques ; ceci sera détaillé dans le prochain chapitre, et sera également vérifié dans le dernier chapitre.

13. Le chiffrement point à point en anglais (End-to-end encryption ou E2EE) est un système de communication où seules les personnes qui communiquent peuvent lire les messages échangés.

Statistique mathématique et Machine Learning pour Big Data analytics

3

Sommaire

3.1 Introduction

Avec la grande quantité de données collectées dans différents domaines, les recherches récentes en statistique ont porté principalement sur le développement d'outils d'analyse de données volumineuses qui facilitent l'inférence statistique dans ce domaine. Des principes statistiques sont nécessaires pour justifier l'inférence de connaissances à partir de données. Cependant, il est difficile d'appliquer ces principes au Big Data. Mais, ces principes peuvent donner des résultats utiles. Dans toute discussion sur le Big Data et les inférences, il est essentiel de savoir qu'il est tout à fait possible de transformer des données en connaissances. Ensuite, nous commençons à creuser pour trouver les indices pouvant aider notre entreprise ou simplement nous donner plus de confiance pour prendre de meilleures décisions plus rapidement. Comme nous l'avons dit, nous sommes conscients que nous avons affaire à quelque chose d'immense, et que les analystes peuvent utiliser des techniques qui les aident à trouver des informations utiles. Ces techniques sont souvent des algorithmes d'apprentissage automatique et des modèles statistiques. En fait, dans un processus d'apprentissage automatique, si le volume de données augmente, alors la précision de l'apprentissage automatique augmente ; pour cette raison, il fonctionne bien avec le Big Data, sinon il ne peut pas fonctionner à son niveau optimal, car avec moins de données signifie que le programme a moins d'exemples ; ainsi, le résultat de la prédiction peut être influencé négativement. "Ce n'est pas grave", nous pouvons commencer à penser. En fait, la solution peut sembler évidente : «Nous utilisons l'apprentissage automatique pour l'analyse de données volumineuses, et c'est tout !». Dans ce chapitre, nous examinerons quelques concepts statistiques, et nous présenterons les différentes méthodes d'échantillonnage statistiques, ainsi qu'une présentation complète des algorithmes d'apprentissage automatique.

3.2 Statistique mathématique et Big Data analytics

Dans le monde réel, les données sont immenses, elles possèdent une structure complexe qui dépasse les capacités des systèmes traditionnels. Elles sont modifiées sous une forme compétente pour le traitement et l'analyse. Elles sont également générées en flux continu-*Streaming*- à partir d'innombrables sources dans notre monde numérique, elles sont gratuitement disponibles. Par conséquent, la croissance rapide des données numérisées offre de vastes possibilités d'analyse des données.

La statistique est la science qui consiste de collecter, d'analyser et de comprendre les données [John 1977]. Elle imprègne les sciences physiques, naturelles et sociales. Tandis que, le Big data reposent principalement sur la collecte d'un grand ensemble de données afin de les analyser dans le plus bref délai [1]. Les données massives sont particulièrement difficiles car certaines de ces données ne sont pas collectées pour répondre à une question scientifique particulière. Mais, les problèmes de données vo-

lumineuses nécessitent généralement des équipes multidisciplinaires, ces équipes généralement sont composés des experts dans le domaine informatique et statistique ; de sorte que, ces équipes garantirent l'extraction des informations significatives et précises à partir du Big Data. Le domaine de la statistique a développé des outils qui peuvent en principe résoudre les problèmes de l'analyse de données. Dans le contexte de données massives, il convient de veiller à ce que ces outils soient pris en charge pour deux raisons principales. Le premier, tous les outils statistiques reposent sur des hypothèses concernant les propriétés des données et comment elles sont échantillonnés, Ces hypothèses peuvent être violées lors du processus de collecte de grands ensembles de données. Le deuxième, les outils d'évaluation des erreurs de procédures et de diagnostic sont en eux-mêmes des procédures de calcul susceptibles ; ils peuvent être inutiles d'un point de vue du calcul dans le domaine Big data[Nongxa 2017].

3.3 Les différents types de données en statistique mathématique

Dans cette partie, nous présentons les différents types de données dans la statistique mathématique que nous devons connaître pour effectuer une analyse des données appropriée. Cette partie est l'une des parties les plus importantes pour développer et déployer un projet d'apprentissage automatique. Le type de données est très important dans la science des statistiques, ce qui doit être compris pour appliquer correctement les mesures statistiques à nos données afin de pouvoir déduire certaines hypothèses correctement. Généralement, les données en statistique sont divisées en deux classes : les données quantitatives et les données qualitatives [Deshpande *et al.* 2016]. Les données qualitatives sont également appelées données catégorielles. Malheureusement, cela devient un peu plus compliqué.

3.3.1 Les données quantitatives

Les données quantitatives traitent la quantité ou les nombres. Elles se réfèrent aux données qui calculent les valeurs, elles peuvent être exprimées en termes numériques[Bruce & Bruce 2017]. En statistique, la plupart des analyses sont effectuées à l'aide de ces données. Elles peuvent être utilisées dans le calcul et le test statistique, telles que la hauteur, le poids, le volume, la longueur, la taille, l'humidité, la vitesse, l'âge, etc. Ce type de données est également représenté sous forme des graphes, des tableaux, etc. En outre, ces données peuvent être classées comme des données discrètes ou continues [Mayya *et al.* 2017].

3.3.1.1 Les données discrètes

Les données discrètes est un nombre qui ne peut pas être précisé [Junhee & Kang 2015]. Typiquement, cela implique des entiers. Par exemple, le nombre d'enfants (ou d'adultes ou d'animaux domestiques) dans notre famille correspond à des données discrètes, car nous comptons des entités entières et indivisibles : nous ne pouvons pas avoir 2,5 enfants ou 1,3 animal domestique.

3.3.1.2 Les données continues

Les données continues peuvent prendre une infinité de valeurs dans un intervalle donné [Howard 2018]. En revanche, elles pourraient être divisées et réduites à des niveaux de plus en plus fins. Par exemple, nous pouvons mesurer la taille de nos enfants à des échelles de plus en plus précises : mètres, centimètres, millimètres, etc...

3.3.2 Données qualitatives

Les données qualitatives font référence aux données qui fournissent des informations et une compréhension sur un problème particulier. Les données qualitatives généralement sont des informations descriptives sur des caractéristiques difficiles à définir ou à mesurer ou impossibles à exprimer numériquement [Howard 2018]. On peut dire que ces données peuvent être approchées, mais ne peuvent pas être calculées. Par conséquent, le chercheur doit posséder une connaissance complète du type de caractéristique avant de collecte des données. La nature des données est descriptive, elle est donc un peu difficile de l'analyser. Ce type de données repose sur la base d'attributs physiques et les propriétés de l'objet. En outre, ce type de données peut être classé en deux catégories : les données nominales et les données ordinales [Campbell & Swinscow 2009].

3.3.2.1 Les données nominales

En statistique, les données nominales est un type de données utilisé pour étiqueter des variables sans fournir de valeur quantitative [Myles & Gin 2000]. C'est la forme la plus simple d'une échelle de mesure où les objets mesurés sont simplement classés en catégories uniques. Ces catégories sont mutuellement exclusives (aucune chose ne peut être placée dans plus d'une catégorie) et totalement inclusive (tout ce qui est classé doit être placé dans au moins une catégorie). Mathématiquement, la propriété d'être classifiable dans une et une seule catégorie peut être symbolisée par les symboles égal et non égal. Les catégories sur les données nominales ne sont en aucun cas ordonnées (par exemple, de petites à grandes), et les nombres ne sont utilisés que comme étiquettes pour les catégories. Aussi, les numéros de licence de voiture est un exemple des données nominales. Le nombre minimum de catégories sur des données nominales est de deux (par exemple, si une pièce de monnaie se pose pile ou face), il peut y avoir autant de catégories que nécessaire. D'autres

exemples de données nominales sont les suivants : type de poisson (par exemple, requin, plie, truite) ; présence ou absence de maladie ; et type de lésion professionnelle.

3.3.2.2 Les données Ordinales

En statistique, les données ordinales est un type de données dans lequel les valeurs suivent un ordre naturel [Myles & Gin 2000]. L'une des caractéristiques les plus remarquables des données ordinales est que les différences entre les valeurs des données ne peuvent pas être déterminées ou n'ont pas de sens. Plus clairement, la mesure du niveau ordinal conserve la propriété de niveau nominal consistant à classer les objets dans une et une seule catégorie (=, !=), mais les catégories sont maintenant ordonnées en fonction de la magnitude de la caractéristique mesurée. On peut maintenant dire que chaque catégorie est supérieure à ou inférieure à son voisin en fonction de la quantité de caractéristique qu'elle représente. Quelques exemples des données ordinales sont les suivants : classement de la taille d'un ensemble d'objets sur échelle à trois chiffres (I = petit, 2 = moyen, 3 = grand), classement de la qualité des films sur une échelle à cinq chiffres (de 1 = très mauvais à 5 = excellent).

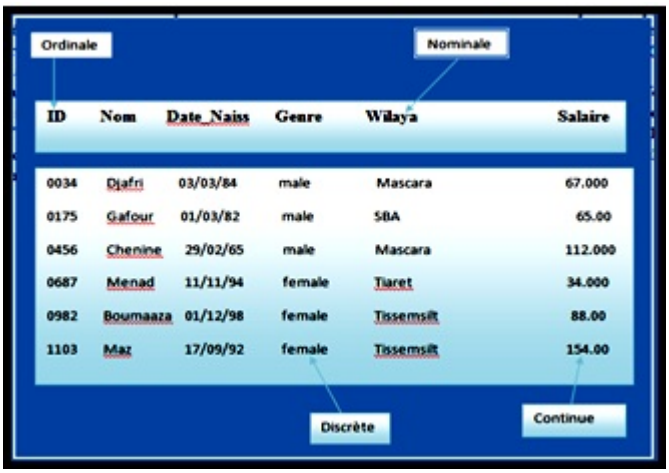

FIGURE 3.1 – *Exemple de jeu de données illustrant des différents types de données.*

Comme nous le verrons plus loin dans les algorithmes d'apprentissage automatique qui seront abordés dans ce chapitre, la présence de différents types d'attributs peut avoir un impact important sur le fonctionnement de ces algorithmes. Les attributs peuvent être de deux types : des attributs bruts et des attributs dérivés. Les attributs bruts sont des attributs provenant directement de sources de données brutes. Par exemple, l'âge du client, le sexe du client, le montant du prêt, etc., sont tous des éléments descriptifs que nous serions probablement en mesure de transférer directement d'une source de données brutes à des jeux de données (tables) pour l'analyse. Tandis que, les attributs dérivés n'existent dans aucune source de données brutes, ils doivent donc être construits à partir des données d'une ou de plusieurs sources de données brutes. Par exemple,

les achats mensuels moyens des clients, les intérêts bancaires, etc., sont autant de caractéristiques descriptives qui pourraient être utiles dans un jeu de données, mais qui devront probablement être dérivées de plusieurs sources de données brutes.

En outre, il existe d'autres termes concernant les types de données sont utilisés dans le domaine de l'apprentissage automatique [Chan & Stolfo 1998] [He & Garcia 2009b] [Kamal *et al.* 2017] [Qiu *et al.* 2016]. Ces termes sont décrits comme suit :

3.3.2.3 Données déséquilibrées

Des données déséquilibrées signifient que les classes ne sont pas équilibrées (c-à-d, pas informellement égales) [Akbani *et al.* 2004] [He & Garcia 2009a]. En termes simples, en utilisant un exemple simple à deux classes, nous pouvons dire que le nombre d'observations dans une classe est nettement inférieur à celui de l'autre classe. C'est un exemple de données déséquilibrées entre les classes. Dans ce cas, nous appelons la classe inférieure une classe minoritaire et l'autre classe une classe majoritaire [Kotipalli & Suthaharan 2014]. Dans certains ensembles de données, pour être significativement plus petite, la classe minoritaire ne doit avoir que 1% d'observations, tandis que la classe majoritaire en a 99% [He & Garcia 2009a], de sorte qu'elles ne sont pas égales en termes d'information.

3.3.2.4 Données inexactes

Des données inexactes signifient que les observations ne sont pas correctes [Biggio *et al.* 2011][Frenay & Verleysen 2014]. Dans les problèmes de classification, cela signifie que certaines des observations ne sont pas correctement étiquetées. Ce problème s'appelle le problème du bruit des étiquettes.

3.3.2.5 Données incomplètes

Data incomplètes désignent simplement les attributs incomplets. Il existe plusieurs modèles possibles comme indiqué dans [Little & Rubin 2002][Lakshminarayan *et al.* 1996].

3.4 Échantillonnage statistique et Big Data

On peut imaginer que les données ont été collectées (de manière réelle ou conceptuelle), nous sommes intéressés par l'exécution des calculs relativement coûteux sur les données. Pour surmonter ce défi, il est nécessaire d'échantillonner des ensembles de données massives, il peut

également être nécessaire d'utiliser plusieurs techniques d'échantillonnage pour éviter les biais de données [1] [Kandel *et al.* 2012]. Dans ce cas, l'échantillonnage est une technique statistique permettant de rechercher et de localiser des modèles dans le Big Data. L'échantillonnage permet aux scientifiques de données de travailler efficacement avec une quantité de données gérable. Les données échantillonnées peuvent être utilisées pour l'analyse prédictive qui repose sur les algorithmes du *Machine Learning* [Kandel *et al.* 2012][Lewis & Catlett 1994]. Les données peuvent être représentées avec précision lors qu'un grand échantillon de données est utilisé. Donc, l'échantillonnage est effectivement un procédé de réduction de dimensionnalité et/ou une technique de réduction de stockage permettant d'obtenir des solutions approximatives à des problèmes coûteux. Dans ce chapitre, nous présentons les bases de la méthodologie d'échantillonnage en définissant d'abord les composants d'une population en termes significatifs pour le prélèvement d'un échantillon. Une fois que ces propriétés de la population ont été définies et discutées, nous commençons l'élaboration de la méthodologie d'échantillonnage.

3.4.1 Population et échantillon

Si nous voulons mener une étude statistique sur la population en Algérie qui est composée de plus de 40 millions d'habitants, donc la population à laquelle nous aurions à demander est supérieure à 40 millions de personnes. Il est évident que faire une interview à plus de 40 millions de personnes nécessite de gros efforts dans de nombreux domaines. Tout d'abord, il ya un grand besoin de temps et d'argent, car nous embauchons beaucoup de personnes pour les entretiens, et nous payons leurs frais de déplacement pour les laisser se rendre dans tous les villages, etc. Par conséquent, il est difficile de toucher chaque algérien à travers des entretiens. Par exemple, il y aura des personnes dans des hôpitaux et d'autres qui se sont rendus dans un pays étranger, etc. Dans cette situation et pour des raisons économiques, il sera commode d'interviewer une certaine partie de la population, un échantillon étant sélectionné de manière appropriée afin que nous puissions obtenir des conclusions ultérieures pour l'ensemble de la population.

En vu des raisons que nous venons de mentionner, il est pratique dans de nombreux cas d'utiliser des échantillons. Mais si nous voulons en tirer de très bonnes conclusions, nous devons assurer que nous avons choisi la bonne option en utilisant des échantillons. Par exemple, dans le cas des citoyens satisfaits du pouvoir en Algérie, si nous choisissons 10 personnes sur les 40 millions d'habitants, cela n'est clairement insuffisant, il ne s'agit pas d'un échantillon représentatif. Je ne serai pas non plus représentatif si nous choisissons 100 personnes de Mascara ou tous nos

1. En statistique, un biais est une démarche ou un procédé qui engendre des erreurs dans les résultats d'une étude. Formellement, le biais de l'estimateur d'un paramètre est la différence entre la valeur de l'espérance de cet estimateur et la valeur qu'il est censé estimer.

amis et notre famille. Certains sujets doivent être clairement définis une fois que nous voulons échantillonner :

— La méthode de sélection des éléments de la population (méthode d'échantillonnage à utiliser).
— Taille de l'échantillon (voir les détails dans le chapitre V).
— Le degré de fiabilité des conclusions que nous pouvons obtenir, c'est une estimation de l'erreur que nous allons avoir (en termes de probabilité).

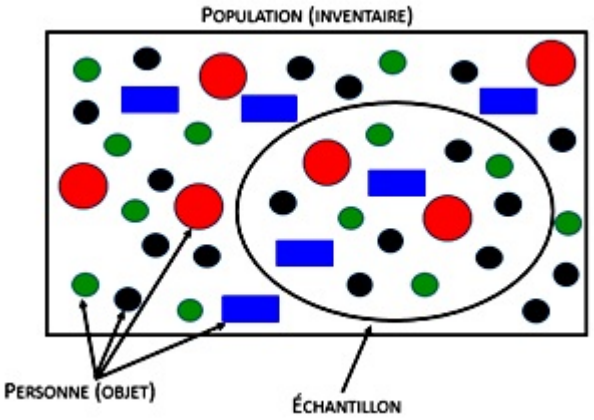

FIGURE 3.2 – *Population et echantillon en statistique mathematique.*

3.4.1.1 Population

En statistique, le terme "population" a une signification légèrement différente de celle utilisée dans la littérature. Il n'est pas nécessaire que cela se réfère à des personnes ou à des créatures animées. Les statisticiens parlent également d'une population d'objets, d'événements, de procédures ou d'observations, de choses, de cas, etc. Une population P est donc un ensemble des individus (n individus) auxquels les résultats de l'enquête doivent être extrapolés. Les individus de la population dont les caractéristiques doivent être mesurées sont appelés unités élémentaires ou éléments de la population $\{u_1, u_2, \ldots u_n\}$ sont identifiés par leurs étiquettes i = 1, 2,. . . n [Ewens 2004][Cochran 1977][Ardilly 2006]. Les populations d'étude peuvent être définies en fonction de l'emplacement géographique, de l'âge et du sexe, avec des définitions supplémentaires d'attributs et de variables telles que la profession, la religion et le groupe ethnique[Banerjee *et al.* 2007]. Par exemple, si nous voulons mener une enquête par sondage pour estimer le nombre de personnes habitant dans la wilaya de Mascara qui n'ont jamais rendu visite à un dentiste, la population comprend toutes les personnes habitant dans la wilaya de Mascara, et chaque personne habitant dans la wilaya de Mascara est une unité élémentaire ou élément de cette population. Un autre exemple, P peut être composée de tous les étudiants d'ESI-SBA (dans ce cas, n est connu), ou P peut être composée de tous les arbres en Sidi Bel Abbes (dans ce cas, n est inconnu). Au sens théorique du terme, il faut clairement savoir si quelque chose appartient à la population P ou non. Étonnamment, cette condition

apparemment simple peut être à l'origine de gros problèmes d'application (ex : qu'est-ce qu'un arbre ?). Donc, les différents types de population sont décrits ci-dessous :

3.4.1.1.1 Population finie

Lorsque le nombre d'éléments de la population est fixe, il permet ainsi de l'énumérer dans sa totalité. Nous considérons une population finie comme une collection de n unités identifiables, étiquetées s = 1,. . . , n, donc la population est dite finie [Lencina *et al.* 2004].

3.4.1.1.2 Population infinie

Lorsque le nombre d'unités d'une population est indénombrable et qu'il est donc impossible d'observer tous les éléments de l'univers [Kozak 2008].

3.4.1.2 Échantillon

Une population contient généralement trop d'individus pour pouvoir étudier convenablement, de sorte qu'une enquête est souvent limitée à un ou plusieurs échantillons prélevés. Un échantillon bien choisi contiendra la plupart des informations sur un paramètre de population particulier, mais la relation entre l'échantillon et la population doit effectivement permettre de déduire un ensemble à partir de cet échantillon, l'objectif est donc de généraliser les résultats de l'échantillon à partir de la population [den Broeck *et al.* 2013]. Par conséquent, Le premier point important de l'échantillon est que chaque membre de la population sélectionnée doit avoir une chance connue non nulle ; de sorte que, ces chances devraient être égales. Nous souhaitons que les choix soient faits de manière indépendante ; En d'autres termes, le choix d'un individu n'affectera pas le risque que d'autres individus soient choisis. Pour ce faire, nous choisissons par le biais d'un processus dans lequel seul le hasard agit, tel que faire tourner une pièce de monnaie ou plus, généralement on utilise un tableau de nombres aléatoires. Un échantillon ainsi choisi est appelé échantillon aléatoire. Le mot "aléatoire" ne décrit pas cet échantillon en tant que tel, mais la manière dont il est sélectionné [Bréchon 2015].

Comme nous venons de le dire, une sélection non appropriée des éléments de l'échantillon peut entraîner des erreurs supplémentaires dès lors que nous voulons estimer les paramètres correspondants dans la population. Mais nous pouvons trouver d'autres types d'erreurs, l'intervieweur peut être partiel, c-à-d, qu'il peut promouvoir certaines réponses plus que d'autres. Il peut également arriver que la personne que nous allons interroger ne veuille pas répondre à certaines questions (ou ne puisse pas répondre). Nous classons toutes ces erreurs possibles de la manière suivante :

1. **Erreur de sélection :** Si l'un des éléments de la population présente une probabilité plus élevée d'être sélectionnée que les autres. Cela se produit lorsque les répondants choisissent eux-mêmes leur participation à l'étude, mais seulement ceux qui sont intéressés à répondre [2]. Imaginons que nous voulions mesurer le degré de satisfaction des sportifs d'une salle de sport, nous en interrogerons donc certains de 08 à 12 heures du matin. Cela signifie que les sportifs qui iront à la salle de sport l'après-midi ne seront pas représentés et que l'échantillon ne sera pas représentatif de tous les sportifs. Un moyen d'éviter ce type d'erreur consiste à sélectionner l'échantillon de manière à ce que tous les sportifs aient la même probabilité d'être sélectionnés.

2. **Erreur de non-réponse :** Il est également possible que certains éléments de la population ne veuillent pas ou ne puissent pas répondre à certaines questions [Lessler & Kalsbeek 1992]. Ou il peut aussi arriver lorsque nous avons un questionnaire comprenant des questions personnelles auxquelles certains membres de la population ne répondent pas honnêtement. Ces erreurs sont généralement très compliquées à éviter, mais si nous voulons vérifier l'honnêteté des réponses, nous pouvons inclure des questions (questions filtres) pour détecter si les réponses sont honnêtes.

Théoriquement, les chercheurs en statistique veulent prélever des échantillons d'une population pour généraliser leurs résultats. Les populations accessibles sont des groupes d'unités de recherche que le chercheur peut réellement échantillonner. L'échantillon attendu est l'ensemble des unités de recherche choisies par le chercheur pour participer à la recherche. Généralement, on ne pourra pas obtenir de données de chacune des unités de recherche sélectionnées. Le chercheur peut ne pas être en mesure de trouver tous les participants prévus, certains peuvent choisir de ne pas participer, certains peuvent commencer mais ne pas terminer la recherche, certains peuvent donner de mauvaises données, etc. L'échantillon réel est le groupe d'unités de recherche à partir duquel nous pouvons réellement obtenir des données. Il convient de déterminer si l'échantillon réel est ou non représentatif de la population à laquelle on souhaite généraliser les résultats.

Pratiquement, il est impossible d'obtenir un échantillon représentatif de la population complètement similaire (100%),mais ce problème sera résolu (relativement) dans notre thèse. Par exemple, on ne peut pas être assez à l'aise pour obtenir des résultats similaires à un échantillon prélevé dans un pays et les comparer à un échantillon prélevé dans un autre pays similaire. Par exemple, les manifestations des pays arabes (Tunisie, Libye, Égypte, Syrie, Yémen) contre leurs systèmes, soi-disant le "printemps arabe" qui ont débuté d'une manière pacifique, mais ont terminé d'une manière tragique, par contre les manifestations pacifiques organisées en

2. https ://www.qualtrics.com/experience-management/research/sampling-errors/ (consulté le 24/03/2019)

Algérie qui ont terminé d'une manière merveilleuse, malgré de nombreux points communs entre ces pays telles que la religion, la langue et l'histoire.

3.4.2 Les méthodes d'échantillonnage

Une partie importante de la théorie statistique concerne les regles d' inférence [3] sur une population, c-à-d, sur un grand ensemble de données à partir desquels un échantillon a été prélevé, c'est ce qu'on appelle la " théorie de l'échantillonnage statistique ". Nous devons donc souligner l'importance d'un bon choix pour les éléments de l'échantillon afin de le rendre représentatif de notre population. Mais avant cela, comment pouvons-nous classer différentes manières de choisir un échantillon ? On peut dire qu'il y a deux types d'échantillonnage [Etikan & Bala 2017] :

3.4.2.1 Echantillonnage aléatoires ou probabilistes

Considérons un ensemble fini d'éléments identifiés par les entiers U = 1, 2, . . . , n. L'ensemble d'identificateurs, parfois appelés étiquettes, il peut être considéré comme formant une liste. L'existence d'une telle liste dans laquelle chaque élément est associé à un seul élément de la liste, elle constitue la pierre angulaire de l'échantillonnage probabiliste. La liste s'appelle également la base de sondage. En pratique, la base de sondage prend de nombreuses formes. Par exemple, il peut s'agir d'une liste au sens traditionnel, telles que la liste des employés d'une entreprise ou la liste des patients dans un hôpital. Le terme échantillons aléatoires ou échantillons de probabilité est utilisé pour les échantillons sélectionnés par des règles de probabilité. Ces règles servent à prédire les résultats de toute méthode d'échantillonnage, elles doivent être constantes si elles sont répétées plusieurs fois, car elles ont tendance à donner tous les échantillons possibles. Cette méthode est appelée méthode "aléatoire", ce qui signifie que l'échantillonnage aléatoire est associé à la procédure pour laquelle chaque échantillon a la même probabilité et chaque élément de la population a la même probabilité d'occurrence dans l'échantillon [4].

Dans l'échantillonnage probabiliste, étant donné que chaque élément a une chance connue d'être sélectionné, ou de d'autre façon, si le choix d'un individu de la population est indépendant de l'attribut qu'il assume, les résultats peuvent être ceux souhaités. Cependant, rien ne garantit que toute technique spéciale soit conforme à ces critères. L'échantillonnage aléatoire est caractérisé par des estimations non biaisées des paramètres de population qui sont des fonctions linéaires des observations (par exemple : des moyennes, des totaux et des proportions de population) peuvent être construites à partir des données de l'échantillon [5]. De plus, les erreurs standards de ces estimations peuvent être estimées ; à

3. les règles d'inférence sont les règles qui fondent le processus de déduction, de dérivation ou de démonstration.

4. https ://towardsdatascience.com/sampling-techniques-a4e34111d808 (consulté le :24/03/2019)

5. http ://www.fao.org/3/x6831e/X6831E12.htm (consulté le 24/03/2019)

condition que les probabilités d'inclusion de second ordre (c-à-d, la probabilité conjointe d'inclure deux unités d'énumération) soient connues. Mais qu'est ce qu'une unité d'énumération ? Dans le cadre d'enquêtes par sondage, il n'est souvent pas possible d'échantillonner directement les unités élémentaires. En effet, les listes d'unités élémentaires à partir desquelles l'échantillon peut être prélevé ne sont souvent pas facilement disponibles, elles ne peuvent être construites qu'à un coût considérable. Toutefois, les unités élémentaires peuvent souvent être associées à d'autres types d'unités pour lesquelles des listes sont disponibles, ou elles peuvent être facilement construites à des fins d'échantillonnage. Ces types d'unités sont appelés unités d'énumération ou unités de listage. Une unité d'énumération ou une unité de listage peut contenir une ou plusieurs unités élémentaires, elle peut être identifiée avant le tirage de l'échantillon [DESASO 1964].

3.4.2.1.1 L'échantillonnage aléatoire simple

L'échantillonnage aléatoire simple est une méthode de sélection d'un échantillon d'une taille donnée dans les unités d'échantillonnage, de sorte que chaque échantillon ait une chance égale d'être sélectionné, et que chaque unité ait une chance égale d'être incluse en tant qu'unité d'échantillonnage. Si l'unité d'échantillonnage est sélectionnée plusieurs fois, de sorte qu'elle est replacée dans la population avant de sélectionner l'unité suivante, on parle "échantillonnage aléatoire avec remise". Si l'unité d'échantillonnage est sélectionnée une seule fois, c.-à-d., n'est pas remplacée, on parle "échantillonnage aléatoire sans remise" [Antal & Tille 2011]. Dans l'échantillonnage avec remise, la variance estimée à partir d'un échantillon est plus simple par rapport à un échantillonnage sans remise. Ainsi, dans un plan d'échantillonnage complexe, un échantillonnage avec remise est utilisé. Normalement, le rapport[6] *-ratio-* entre les deux variables variant d'une unité à l'autre est la quantité estimée à partir d'un simple échantillonnage aléatoire. Cette estimation de rapport est utile lorsque l'intérêt de l'enquêteur est la moyenne de la population par individu [conrad taeuber 1961]. Dans l'échantillonnage aléatoire simple, on se procure une liste de toutes les unités statistiques de la population et on les numéroter de 1 à n, et en suite, on choisit au hasard « m» nombre déférents correspondant aux « n » unités statistiques font partie de l'échantillon.

3.4.2.1.2 L'échantillonnage stratifié

Dans l'échantillonnage stratifié, la population est divisée en sous-populations sans chevauchement appelées strates, cette division effectue selon certaines caractéristiques de sorte que, les unités d'une strate soient proches que possible [KSteven 2012]. Ensuite, même si une strate peut différer considérablement d'une autre, un échantillon stratifié avec le nombre souhaité d'unités de chaque strate de la population aura tendance à être «représentatif» de la population dans son ensemble. L'échantillonnage

6. En statistique, le rapport est une mesure statistique, souvent utilisée en épidémiologie, exprimant le degré de dépendance entre des variables aléatoires qualitatives.

stratifié est peu probable de choisir un échantillon absurde puisqu'on s'assure la présence proportionnelle de tous les divers sous-groupes composant la population [Etikan & Bala 2017]. L'échantillonnage stratifié est largement utilisé comme méthode puissante et flexible. Laissons-nous examiner quelques raisons de sa popularité :

— Supposons que des estimations de la précision spécifiée soient recherchées pour certaines sous-populations (domaines d'étude). Par exemple, conifères et feuillus, ou souris et éléphants. Chaque domaine d'étude peut être traité comme une strate séparée si l'appartenance au domaine est bien déterminée.

— Les aspects pratiques liés à la réponse, à la mesure et aux informations auxiliaires peuvent différer considérablement d'une sous-population à une autre, de sorte que l'on souhaite traiter chaque sous-population comme une strate séparée.

— Pour des raisons pratiques ou administratives, l'organisation d'enquête peut avoir divisé tout son territoire en plusieurs zones géographiques. Ici, il est naturel de considérer chaque zone comme une strate.

Le choix de la variable de stratification, les seuils de résultants et le nombre de strates sont des aspects importants [Jan *et al.* 2003]. En principe, une variable de stratification devrait être fortement corrélée à la variable de réponse d'intérêt principal. Chaque strate devrait alors être aussi homogène que possible et différer considérablement des autres, de sorte que la variance[7] entre les strates soit plus grande que la variance au sein des strates (encore une fois, pensons aux chats et aux chiens). Souvent, mais pas toujours, le choix du plan d'échantillonnage ainsi que celui de l'estimateur associé sont effectués de manière uniforme pour toutes les strates. Néanmoins, la diversité des objets rend souvent impossible l'obtention d'une conception optimale des échantillons stratifiés. La stratification présente trois avantages majeurs par rapport à un échantillonnage aléatoire simple :

1. Dans certaines conditions, la précision peut être accrue par rapport à un échantillonnage aléatoire simple (c-à-d, que la procédure d'estimation peut entraîner des erreurs types plus faibles).

2. Il est possible d'obtenir des estimations pour chacune des strates ayant une précision spécifiée.

3. - Il peut être aussi simple, pour des raisons politiques ou administratives, la collecte des informations pour un échantillon stratifié est le même cas pour un échantillon aléatoire simple. Si tel est le cas, peu d'éléments peuvent être perdus en prenant un échantillon stratifié,

7. En statistique et en théorie des probabilités, la variance est une mesure de la dispersion des valeurs d'un échantillon ou d'une distribution de probabilité.

car les erreurs standards (erreurs-types) vont rarement au-delà des simples erreurs d'échantillonnage aléatoire.

Lorsque les unités de population hétérogènes sont divisées en groupes homogènes non chevauchés, ces groupes sont appelés strates. Après avoir déterminé les strates, l'échantillon est extrait de chaque échantillon indépendamment par un échantillonnage aléatoire simple. Il est ensuite appelé échantillonnage aléatoire stratifié. L'échantillonnage ou sous-échantillonnage en deux étapes est une méthode dans laquelle l'unité n'est pas mesurée complètement mais est elle-même échantillonnée, c.-à-d. qu'un échantillon d'unités sélectionné pour la première fois est appelé unité primaire, puis qu'un échantillon de sous-unités est sélectionné pour chaque unité primaire choisie. L'échantillonnage aléatoire simple est un cas particulier d'échantillonnage stratifié et, par conséquent, la moyenne d'échantillonnage d'un échantillon aléatoire simple est normalement répartie dans la limite. On peut espérer un échantillon « représentatif » puisque l'échantillonnage aléatoire simple donne à chaque individu de la population une chance égale.

3.4.2.1.3 L'échantillonnage par grappes

L'échantillonnage par grappes est une technique dans laquelle des grappes de participants représentant la population sont identifiées et incluses dans l'échantillon [Jackson 2011]. La procédure d'échantillonnage par grappe consiste à sélectionner des groupes d'individus de la population (appelée : grappes) par échantillonnage aléatoire simple, puis à prélever un échantillon aléatoire simple de chaque grappe sélectionnée. Cette technique est couramment utilisée lorsque la population est répartie sur une vaste zone géographique[Gravetter & Forzano 2012]. Bien que l'échantillonnage en grappes ne nécessite aucune nouvelle théorie pour l'estimation, la conception des échantillons en grappes mérite une discussion pour deux raisons. Premièrement, il peut être possible de former des grappes de différentes tailles. Par exemple, si les grappes sont des grappes de ménages, il est possible de concevoir des grappes de tailles différentes en utilisant des mécanismes tels que les statistiques par blocs de recensement. Deuxièmement, la taille de grappe connue est une variable auxiliaire qui peut être utilisée aux étapes de la conception et de l'estimation. Toutefois, la taille de la grappe n'est pas toujours connue à la phase de la conception. Par exemple, un échantillon d'adresses de résidence fournit un échantillon en grappes de personnes, mais le nombre de personnes résident n'est souvent pas connu à la phase de la conception.

Les grappes doivent être formées pour que le collecteur de données trouve qu'il est relativement facile d'identifier correctement l'unité. Par exemple, si la zone de terrain est échantillonnée pour une enquête sur les pratiques agricoles, des unités basées sur la propriété de la terre ou des exploitants agricoles seraient pratiques, tandis que des unités avec des limites définies par la latitude et la longitude seraient moins souhaitables. Par ailleurs, un segment basé sur la latitude et la longitude serait utile pour l'étude des forêts étant donné que les unités cartographiques

peuvent être utilisées pour déterminer l'emplacement. Il existe des similitudes avec l'échantillonnage aléatoire stratifié, mais il diffère du fait que les unités d'échantillonnage (grappes) sont divisées en «sous-populations mutuellement exclusives et collectivement exhaustives» et ne sont pas traitées en tant qu'individus. Pour cela, nous pouvons dire que l'échantillonnage en grappes est souvent moins coûteux que l'échantillonnage stratifie ; par ce que, une collection d'unités dans une grappe est égale au nombre d'unités secondaires sélectionnées au hasard dans la population. Mais l'échantillonnage stratifié est plus avantageux que l'échantillonnage en grappes pour les raisons suivantes :

— L'échantillonnage en grappes est moins représentatif de la population que l'échantillonnage stratifié, étant donné que les individus d'une grappe ont tendance à présenter des caractéristiques similaires.

— Avec un échantillon en grappes, il est possible que le chercheur ait une grappe surreprésentée ou sous-représentée qui puisse biaiser les résultats de l'étude contrairement à l'échantillonnage stratifié.

— Dans l'échantillonnage stratifié, une fois que les strates sont créées, un échantillon aléatoire est tiré de chaque strate. Par contre, dans l'échantillonnage en grappes, les éléments ne sont pas sélectionnés dans chaque grappe.

— Dans l'échantillonnage stratifié, les sous-échantillons de chaque strate sont pris proportionnellement à la taille de l'échantillon de manière à ce que l'échantillon global corresponde à la population. mais dans l'échantillonnage en grappes, les sous-échantillons de chaque grappe ont tous la même taille, introduisant ainsi un biais.

— Bien que l'échantillonnage en grappes soit beaucoup plus facile à mettre en œuvre que l'échantillonnage stratifié, il engendre des erreurs d'échantillonnage plus importantes puisqu'il laisse une proportion importante de la population non échantillonnée.

— Il est plus difficile de calculer et d'interpréter l'analyse statistique inférentielle des données collectées par échantillonnage en grappes que l'analyse statistique inférentielle des données collectées par échantillonnage stratifié. Parce que s'il y a plus d'étapes dans un plan d'échantillonnage en grappe, alors il y a une augmentation significative de l'erreur d'échantillonnage.

La méthode d'échantillonnage en grappes ne nécessite pas une liste globale de la population puisque on compte seuls les individus inclus dans les grappes sont comptés. En plus, elle permet de limiter l'échantillon à des groupes compacts ce qui permet de réduire les couts de déplacement, de suivi et de supervision. En revanche, cette méthode peut entrainer des résultats imprécis (moins précis que les méthodes précé-

dentes) puisque les unités voisinés ont tendance se rassembler. Ainsi que, elle ne permet pas de contrôler la taille finale de l'échantillon.

3.4.2.1.4 L'échantillonnage systématique

L'échantillonnage systématique est fréquemment utilisé dans la pratique, car il est facile à appliquer, et il peut être facilement enseigné aux personnes peu familiarisées avec la méthodologie d'enquête. L'échantillonnage systématique est une méthode utilisée lorsque les éléments de population sont classés dans un ordre spécifique dans lequel un échantillon peut être tiré de manière systématique plutôt que de générer un échantillon aléatoire simple [Oaks 2012][Finney 1948]. D'une manière générale, il faut prélever un échantillon systématique en déterminant tout d'abord l'intervalle d'échantillonnage souhaité a, en choisissant un nombre aléatoire j compris entre 1 et a et en sélectionnant les éléments étiquetés j, j + a, j + 2 a, j + 3 a etc., Par exemple, si un échantillon de 1 à a unités est spécifié (où a est, par exemple, un nombre à deux chiffres), nous pouvons sélectionner un nombre aléatoire à deux chiffres compris entre 01 et a. Si le nombre sélectionné est j, les nombres à deux chiffres j, j + a, j + 2 a, etc., sont sélectionnés jusqu'à ce qu'un nombre à trois chiffres soit atteint, mais le problème est que les données peuvent être biaises à cause de la périodicité. Nous pouvons également considérer l'échantillonnage systématique comme un cas particulier d'échantillonnage stratifié avec n strates, chacune d'entre elles avec a éléments, de sorte que nous ne choisissions qu'un seul élément de chaque stratus. Dans l'échantillonnage stratifié, l'élément sélectionné est aléatoire, tandis que dans cette technique, nous choisissons de manière aléatoire le premier élément et les autres sont déterminés par a. En plus, l'échantillonnage systématique et l'échantillonnage en grappes semblent être opposés, le premier sépare les unités d'un échantillon et le deuxième les regrouper, mais réellement les deux modèles partagent la même structure. Parce que dans l'échantillonnage systématique, une seule unité principale est constituée d'unités secondaires réparties de manière systématique dans l'ensemble de la population. De la même manière dans l'échantillonnage en grappes, une unité principale consiste en une grappe d'unités secondaires, généralement très proches les unes des autres [Kalton 2017].

3.4.2.1.5 L'échantillonnage à plusieurs étapes (degrés)

Souvent, un plan d'échantillonnage particulier spécifie que l'échantillonnage doit être effectué en deux ou trois étapes ; cette conception s'appelle une conception d'échantillonnage à plusieurs étapes. Dans l'échantillonnage à plusieurs étapes, les unités de population sont d'abord divisées en groupes appelés unités primaires, puis plusieurs unités principales sont sélectionnées à l'aide de l'une des méthodes de sélection décrites ci-dessus. L'échantillonnage ultérieur est limité aux unités principales sélectionnées. Si des unités de population sont sélectionnées dans ces groupes, nous avons un échantillonnage en deux étapes : (1) premièrement, la sélection des grappes éligibles (unité primaire), (2) puis la sélec-

tion d'un échantillon d'individus appartenant à ces grappes (unité secondaire) [Gravetter & Forzano 2012]. Alternativement, des groupes d'unités de population peuvent être sélectionnés au sein d'unités primaires, ils sont alors définies comme des unités secondaires, et par conséquent, ils sont limités à un échantillonnage ultérieur. Dans ce cas, nous avons trois étapes ou plus dans la conception. Par exemple, une enquête auprès des ménages menée dans une grande ville (wilaya d'Oran) pourrait avoir un plan d'échantillonnage spécifiant qu'un échantillon à partir des daïras soit établi dans la wilaya d'Oran ; de sorte que dans chaque daïra sélectionnée dans l'échantillon, un échantillon de circonscriptions administratives de la daïra (les communes) soit constituées ; et que dans chaque commune, un échantillon de ménages soit constitué. Dans l'échantillonnage à plusieurs étapes, une base de sondage différente est utilisée à chaque phase de l'échantillonnage. Les unités répertoriées dans la base de sondage sont généralement appelées unités d'échantillonnage. Dans l'exemple mentionné ci-dessus, la base de sondage pour la première étape est la liste des daïras de la wilaya d'Oran, et chaque daïra est une unité d'échantillonnage pour cette étape. La liste des communes constitue la base de sondage de la deuxième étape et chaque commune constitue une unité d'échantillonnage pour cette étape. Enfin, la liste des ménages de chaque commune échantillonnée à la deuxième étape constitue la base de sondage de la troisième et dernière étape. Chaque ménage constitue une unité d'échantillonnage pour cette étape. Les unités d'échantillonnage pour la première étape sont généralement appelées unités d'échantillonnage primaires (UEP). Les unités d'échantillonnage pour la dernière étape d'un plan d'échantillonnage à plusieurs degrés sont appelées unités d'énumération ou unités de listage ; ceux-ci ont été discutés précédemment.

3.4.2.1.6 L'échantillonnage à probabilité inégale

Avec certaines procédures d'échantillonnage, différentes unités de la population ont des probabilités différentes, c.à.d. n'ont pas la même probabilité d'être incluses dans un échantillon, on parle d'échantillonnage à probabilité inégale. Lorsque l'échantillonnage à probabilités inégales est applicable, il donne généralement de meilleures estimations que l'échantillonnage à probabilités égales, parce que les probabilités d'inclusion différentes peuvent résulter d'une caractéristique inhérente de la procédure d'échantillonnage, ou elles peuvent être imposées délibérément pour obtenir ces meilleures estimations en incluant les unités plus importantes avec une probabilité plus élevée [Grafstrom 2010].

Contrairement à l'échantillonnage aléatoire simple, l'échantillonnage à probabilité inégale suppose des probabilités d'inclusion différentes pour chaque individus tel que : Si une zone d'étude est divisée en parcelles de tailles inégales, il peut être souhaitable d'attribuer des plus grandes probabilités d'inclusion à des plus grandes parcelles, cela peut être fait en sélectionnant des points dans la zone d'étude avec une probabilité égale, et en incluant un parcelle chaque fois qu'un point sélectionné y est situé. Plus généralement, des sélections de probabilités inégales peuvent être

effectuées en attribuant à chaque unité un intervalle de longueur égale à la probabilité désirée et en sélectionnant des nombres aléatoires dans la distribution uniforme. Une unité est incluse si le nombre aléatoire est dans son intervalle [KSteven 2012]. Avec l'échantillonnage à probabilité inégale, généralement la moyenne de l'échantillon ne permet généralement pas d'estimer la moyenne de la population, elle peut donc être totalement trompeuse, ceci est également vrai pour les échantillons aléatoires pour lesquels les probabilités d'inclusion sont inconnues ; par exemple, les entretiens avec des étudiants échantillonnés à la cafétéria, où les gros buveurs de café auront une probabilité beaucoup plus grande d'être échantillonnés.

3.4.2.2 Echantillonnage non probabiliste (non aléatoire)

L'échantillonnage non probabiliste est une technique d'échantillonnage dans laquelle l'échantillon statistique est choisi en fonction d'une estimation personnelle plutôt que d'une sélection aléatoire, ce type d'échantillonnage ne garantit pas l'égalité des chances pour chaque objet de la population cible [Gravetter & Forzano 2012][Moorley & Shorten 2014]. L'échantillonnage non probabiliste est fréquemment utilisé notamment dans les études de marché et les sondages d'opinion. Il est utilisé parce que l'échantillonnage probabiliste est souvent une procédure longue et coûteuse et, en fait, peut ne pas être réalisable dans de nombreuses situations. Il Indique que, même si les unités d'analyse de l'échantillonnage non probabiliste n'ont pas la même chance d'être incluses dans l'échantillon, c.à.d, tous les individus de la population n'ont pas une chance de participer à l'étude, contrairement à l'échantillonnage probabiliste où chaque individu de la population a une chance connue d'être sélectionné. Il est toujours utilisé fréquemment en raison de moins couteux, plus rapide et facile à réaliser. Cependant, il existe un certain nombre d'inconvénients. De telles méthodes peuvent être sujettes à un biais de sélection [Forster 2001].

Par exemple, en vu aux manifestations pacifiques qui a eu lieu dans notre pays du 22 février 2019, il incombait à l'autorité algérienne de constituer un nouveau gouvernement, de sorte que ce gouvernement est composé par exemple de vingt ministres, le premier ministre invite les représentants des partis politiques et les représentants de mouvement populaire comme suit : cinq représentants de FLN , cinq représentants de RND, cinq représentants de MHS, cinq représentants de TAJ, cinq représentants de MPA, cinq représentants de PT et vingt représentants de mouvement populaire, mais la sélection des représentants (individus) spécifiques dans chacune de ces catégories reste entre les mains du premier ministre . Il est fort probable qu'une telle méthode de sélection d'un échantillon puisse conduire à des estimations très biaisées. Par exemple, le premier ministre peut choisir les cinq représentants de FLN, les cinq représentants de RND, les cinq représentants de TAJ, deux représentants de HMS, deux représentants de MPA, un représentant de PT et aucun représentant de mouvement

populaire, par conséquent, ce nouveau gouvernement peut ne pas être représentatif de la communauté populaire dans son ensemble.

3.4.2.2.1 L'échantillonnage par quotas

Ce type de méthode d'échantillonnage est utilisé lorsque la population est hétérogène, dans ce cas le chercheur décide à l'avance de certaines caractéristiques clés qu'il utilisera pour stratifier l'échantillon. De sorte que, il prélève des échantillons d'un groupe d'éléments présentant des traits ou des caractéristiques spécifiques proportionnellement à leur taille dans la population, sachant que ces échantillons formés sont homogènes [8] [Alvi 2016]. Dans ce type d'échantillonnage, la base de sondage est divisée en autant de sous-ensembles qu'il y a d'attributs que nous voulons observer, et la proportion [9] de chaque sous-ensemble dans l'échantillon est la même que dans la base de sondage. La recherche ethnographique [10] sur un tel échantillon nécessite évidemment un investissement de temps considérable qui peut même durer plusieurs années. Par conséquent, il existe peu d'exemples d'échantillonnage par quota dans la littérature ethnographique. En plus, dans de nombreux pays, les données de recensement peuvent être médiocres ou inexistantes. Même les informations de recensement les plus fiables ne peuvent révéler toutes les caractéristiques pouvant affecter les opinions étudiées. Pour la plupart des populations, par exemple, on ne sait pas combien de personnes sont extravertis ou introvertis. Pourtant, ces caractéristiques peuvent être liées à des opinions sur certains sujets. L'échantillonnage par quota est un échantillonnage non-aléatoire, pour cela les statisticiens soulignent qu'il est impossible de donner à chaque membre de l'univers une chance connue d'être sélectionnée, et qu'il est donc impossible de calculer la marge d'erreur et le niveau de confiance des résultats [11] [12]. Mais, l'échantillonnage par quota nous aide pour la sélection des échantillons en fournissant des métadonnées sur le nombre d'échantillons à prélever pour chaque groupe cible de manière à ce qu'ils soient proportionnels à la population d'origine.

3.4.2.2.2 L'échantillonnage par jugement

L'échantillonnage par jugement est sélectionné sur la base des connaissances d'un expert dans le domaine étudié, c.-à-d, l'échantillon est pris en fonction de certains jugements sur la population globale. L'hy-

8. https ://humansofdata.atlan.com/2016/04/quota-sampling-when-to-use-how-to-do-correctly/ , (consulté le 28/03/2019)

9. La proportion est un rapport d'égalité entre deux quantités, traduit sous forme d'équivalence entre deux rapports (a/b = c/d) : Prix d'un paquet de chips = 10 DA, donc Prix de quatre paquet de chips = 40 DA, dans la mesure où 1/4 = 10/40.

10. L'ethnographie est le domaine des sciences sociales qui étudie sur le terrain la culture et le mode de vie de peuples ou milieux sociaux donnés.

11. https ://www.netquest.com/blog/en/quota-sampling, (consulté le 28/03/2019)

12. https ://humansofdata.atlan.com/2016/04/quota-sampling-when-to-use-how-to-do-correctly/ , consulté le 28/03/2019

pothèse sous-jacente est que l'investigateur choisira des unités caractéristiques de la population.

Il y a en fait deux raisons pour lesquelles nous pouvons faire un échantillonnage par jugement. Premièrement, ce serait le meilleur moyen de connaître le point de vue des personnes ayant des compétences spécifiques, tandis que l'autre raison pour laquelle nous pouvons utiliser un échantillonnage par jugement est la preuve de la validité d'une autre méthode d'échantillonnage que nous avons choisie.

Par exemple, supposons que nous effectuions un échantillonnage d'une population en sélectionnant les individus les plus typiques, mais nous craignons que les critères que nous avons utilisés pour définir les individus les plus typiques soient sujets à critique. Nous pouvons convoquer un groupe d'experts composé de personnes possédant une expérience reconnus et une connaissance approfondie de ce domaine ou de ce sujet, ensuite nous demandons leur avis pour examiner nos définitions modales et pour formuler des observations suivant leur pertinence et leur validité. L'avantage de cela est que nous ne tentons pas seul de défendre nos décisions. Nous avons le soutien d'experts renommés. Mais parfois, L'inconvénient est que même les experts peuvent avoir tort.

3.4.2.2.3 L'échantillonnage de boule de neige

On appelle l'échantillonnage en boule de neige, parce que si nous avons une boule de neige qui roule, elle ramasse plus de "neige" en cours de route, elle devient de plus en plus grande. En statistique, il permet de construire un échantillon en choisissant d'abords un petit groupe d'individus ayant les caractéristiques recherchées pour l'étude jusqu'à ce que l'échantillon compose le nombre d'individus voulu. Cette technique d'échantillonnage est souvent utilisée dans des populations cachées auxquelles les chercheurs ont difficilement accès, de sorte que, certaines personnes ne veulent pas être présentes [13]. Par exemple, si une étude portait sur la fraude lors d'examens, le vol à l'étalage, la consommation de drogue, ou tout autre comportement social «inacceptable», les participants potentiels se méfieront de présenter en raison des conséquences potentielles, dans ce cas, il est difficile d'établir une base de sondage. Cependant, d'autres participants à l'étude connaîtront probablement d'autres personnes dans la même situation qu'elles-mêmes, ils pourraient informer les autres des avantages de l'étude et les du rassurer secret absolu.

13. https ://en.wikipedia.org/wiki/Snowball_sampling#cite_note-1 , (consulté le : 28/03/2019)

3.4.2.2.4 L'échantillonnage de commodité (aveuglette)

Les chercheurs utilisent des techniques d'échantillonnage dans des situations où de grandes populations doivent être testées, car dans la plupart des cas, il est pratiquement impossible de tester l'ensemble de la population. Parmi ces techniques, l'échantillonnage de commodité. Il est utilisé pour créer un échantillon en termes de facilité d'accès, de disponibilité à faire partie de l'échantillon, de disponibilité à un créneau horaire donné ou de toute autre spécification pratique d'un élément particulier [Dörnyei 2007]. L'échantillonnage de commodité est utile à certaines fins, et il nécessite très peu de planification. Le chercheur choisit les membres uniquement sur la base de la proximité, il ne considère pas s'ils représentent la population entière ou non. En utilisant cette technique, on peut observer les habitudes, les opinions et les points de vue de la manière la plus simple possible.

En outre, l'échantillonnage de commodité est la technique d'échantillonnage la plus couramment utilisée, car il est extrêmement rapide, simple, économique et que les membres sont facilement accessibles pour faire partie de l'échantillon. Il est utilisé dans des situations où la recherche principale a été effectuée sans apport supplémentaire. Il n'y a pas de critères à prendre en compte dans cet échantillon, il est donc très facile d'inclure des éléments dans cet échantillon. Chaque élément de la population est éligible pour faire partie de cet échantillon, il dépend de la proximité du chercheur pour être inclus dans l'échantillon. Dans sa forme de base, l'échantillonnage de commodité peut être appliqué en posant des questions à des personnes aléatoires dans la rue. Par exemple, la campagne de mariage «future épouse» peut être considérée comme un exemple approprié de cette méthode d'échantillonnage. Par conséquent, tous les hommes de la communauté peuvent participer au concours sans aucune discrimination ou sélection [14].

3.5 Apprentissage automatique

L'analyse de données prédictive utilise l'apprentissage automatique pour créer des modèles qui capturent les relations dans des jeux de données volumineux entre des fonctionnalités descriptives et une fonctionnalité cible. L'apprentissage automatique est une méthode d'analyse de données qui automatise la création de modèles analytiques basés sur des expériences antérieures sans aucune assistance extérieure de la part de l'homme [Das & Behera 2017]. Aujourd'hui, un large assortiment de données est immergé qui manque de connaissances. Ces données abondantes sont classées sous forme structurée et non structurée. Les données structurées sont les données qui sont organisées dans des tableaux, tandis que les données non structurées soient des données représentées sous une forme irrégulière provenant de plates-formes de médias sociaux, d'applications

14. https ://research-methodology.net/sampling-in-primary-data-collection/convenience-sampling/ , (consulté le : 29/03/2019)

mobiles, de services de localisation et des technologies de l'Internet des objets [15]. Les données étant complexes et volumineuses, il est essentiel d'effectuer des calculs sur un environnement parallèle distribué à l'aide de techniques d'apprentissage automatique (pour plus de détails, voir le chapitre-4). En raison de cette croissance exponentielle des données, les techniques d'apprentissage automatique évoluent pour s'adapter à de nouvelles complexités. Dans une définition générale, l'apprentissage automatique est défini comme un processus automatisé de conception d'un bon moteur de recherche qui extrait des modèles à partir de données. Plus précisément, l'apprentissage automatique est la branche de l'informatique qui utilise l'expérience passée pour apprendre et utiliser ses connaissances pour prendre des décisions futures [Richardson *et al.* 2006].

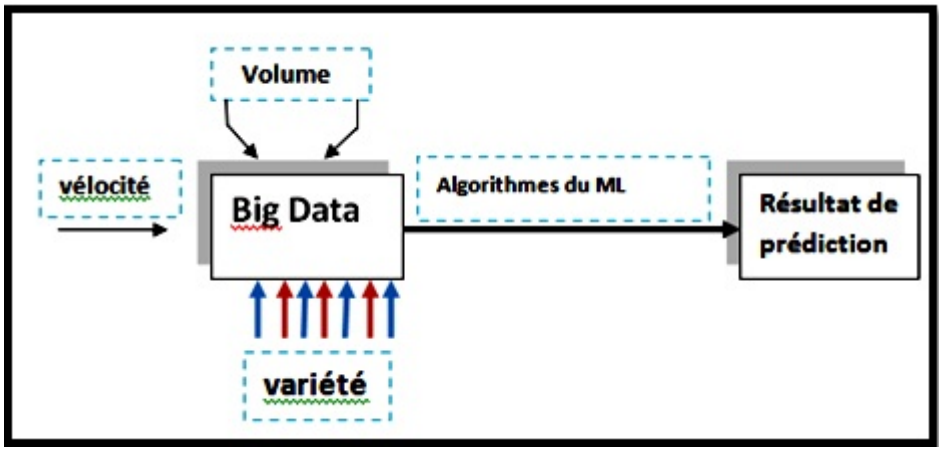

FIGURE 3.3 – *Illustration représentant le scenario de l'analyse prédictive des données massives.*

3.5.1 Objectifs et utilités de l'apprentissage automatique

L'apprentissage automatique-*Machine Learning*- a évolué à partir du vaste domaine de l'intelligence artificielle. Il fournit aux scientifiques un moyen d'explorer des modèles d'apprentissage et des algorithmes d'apprentissage pouvant aider les machines (ordinateurs, par exemple) à apprendre le système à partir de données ; donc l'objectif principal de l'apprentissage automatique est de doter les machines de l'intelligence de l'être humain ; de sorte qu' il est capable de fournir des prédictions basées sur une énorme quantité de données, ce qui est une tâche presque impossible à réaliser pour un être humain [Burhan *et al.* 2014]. En effet, l'apprentissage automatique a pour objectif de trouver le modèle prédictif [16] qui généralise le mieux. Afin de trouver ce meilleur modèle, un algorithme d'apprentissage automatique doit utiliser certains critères pour choisir parmi les modèles candidats qu'il considère lors de sa recherche.

15. https ://solutionsreview.com/data-management/key-differences-between-structured-and-unstructured-data/ ,(consulté le : 30/03/2019)

16. Au fur et à mesure que les algorithmes ingèrent les données d'apprentissage, il devient possible de créer des modèles prédictifs plus précis basés sur ces données. Donc, un modèle prédictif est le résultat qui est généré lorsque vous formez votre algorithme d'apprentissage automatique à l'aide de données.

3.5.2 Techniques et principe de fonctionnement de l'apprentissage automatique

L'apprentissage automatique consiste en de nombreux algorithmes puissants permettant d'apprendre des modèles, d'acquérir des connaissances et des prédictions. Plus précisément, ces algorithmes fonctionnent en recherchant dans un ensemble de modèles prédictifs possibles pour capturer la meilleure relation entre les caractéristiques descriptives et la fonctionnalité cible dans un jeu de données ; à partir duquel l'algorithme d'apprentissage effectue une sélection au cours de l'apprentissage. Un critère évident pour piloter cette sélection est de rechercher des modèles cohérents avec les données. Nous pouvons ensuite utiliser ce modèle pour établir des prédictions pour de nouvelles instances. Ces deux étapes distinctes sont illustrées à la figure 3.4

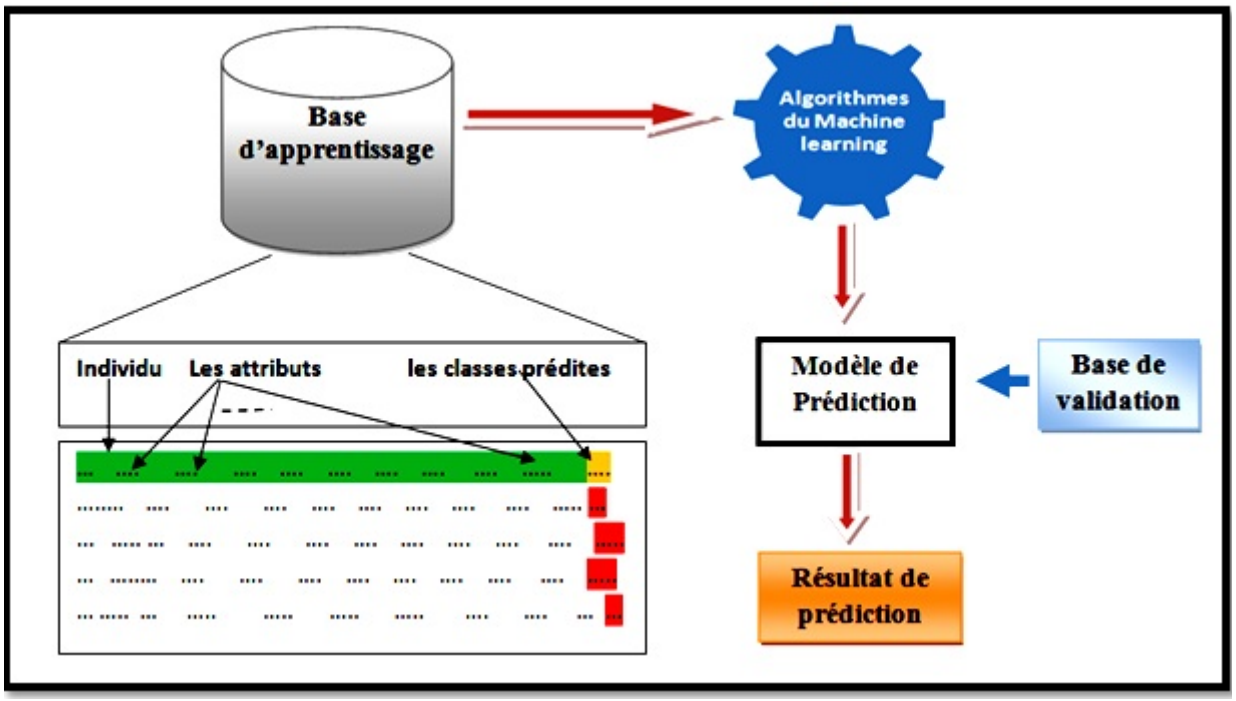

FIGURE 3.4 – *Procédure détaillée d'apprentissage automatique pour le traitement des jeux de données.*

Le développement et le déploiement de modèles d'apprentissage automatique impliquent une série d'étapes presque similaires au processus de l'analyse prédictive, afin de développer, valider et mettre en œuvre des modèles d'apprentissage automatique [17] [18]. Les étapes sont les suivantes :

17. https ://medium.com/dataseries/7-steps-to-machine-learning-how-to-prepare-for-an-automated-future-78c7918cb35d ,(consulté le : 31/03/2019)
18. https ://www.kdnuggets.com/2018/05/general-approaches-machine-learning-process.html, (consulté le : 31/03/2019)

1. **Collecte de données :** Les données pour l'apprentissage automatique sont collectées directement à partir de données source structurées, du Web, d'API, etc., l'apprentissage automatique pouvant fonctionner à la fois sur des données structurées et non structurées (voix, images et texte). Le résultat de cette étape est généralement une représentation des données que nous utiliserons pour l'apprentissage (soit étiquetées, ce qui signifie apprentissage supervisé, soit non étiquetées, signifiant apprentissage non supervisé).

2. **Préparation des données :** Les données doivent être formatées conformément à l'algorithme d'apprentissage automatique choisi ; nous éliminons les doublons, nous corrigeons les erreurs, nous gérons les valeurs manquantes, nous faisons la normalisation, nous faisons les conversions de types de données, etc.

3. **Choix du modèle :** Cette étape consiste à choisir l'un des nombreux modèles créés au fil des années par des chercheurs et des spécialistes des données. Nous faisons le bon choix pour faire le travail.

4. **Apprentissage :** Une fois les étapes précédentes terminées, nous passons ensuite à ce qui est souvent considéré comme l'essentiel de l'apprentissage automatique appelé : apprentissage, dans lequel les données sont utilisées pour améliorer progressivement la capacité du modèle prédictif.

5. **Évaluer le modèle prédictif :** Une fois que l'apprentissage est terminé, nous vérifions maintenant s'il est suffisant avec cette étape. Alors, l'évaluation permet de tester le modèle prédictif par rapport à des données qui n'ont jamais été vues et utilisées pour l'apprentissage, cela devrait être représentatif de la façon dont le modèle fonctionne dans le monde réel.

6. **réglage des paramètres :** Une fois l'évaluation terminée, il est possible d'améliorer notre apprentissage en ajustant les paramètres. Les paramètres de modèle simple peuvent inclure : le nombre d'étapes de l'apprentissage, le taux d'apprentissage, les valeurs d'initialisation et la distribution, etc. Une fois que nous avons terminé avec ces paramètres et que nous sommes satisfaits, nous pouvons passer à la dernière étape.

7. **Prédiction :** L'apprentissage automatique consiste essentiellement à utiliser des données pour répondre à des questions. C'est donc l'étape finale où nous devons répondre à quelques questions. Ici, nous pouvons enfin utiliser notre modèle pour prédire le résultat comme nous voulons.

En résumé, l'apprentissage automatique fonctionne en recherchant un ensemble de modèles potentiels pour trouver le modèle de prédiction qui se généralise le mieux au-delà du jeu de données.

3.5.3 Types d'algorithmes d'apprentissage automatique

L'apprentissage automatique peut être considéré comme un ensemble des algorithmes capables de reconnaître automatiquement les modèles de données, tandis que les modèles de données reconnus sont utilisés pour prédire les nouvelles valeurs observées. Les algorithmes d'apprentissage automatique sont classés généralement en trois catégories : apprentissage supervisé, non supervisé et par renforcement [Dasgupta & Nath 2016]. Chaque méthode d'apprentissage a sa signification et a sa dimensionnalité.

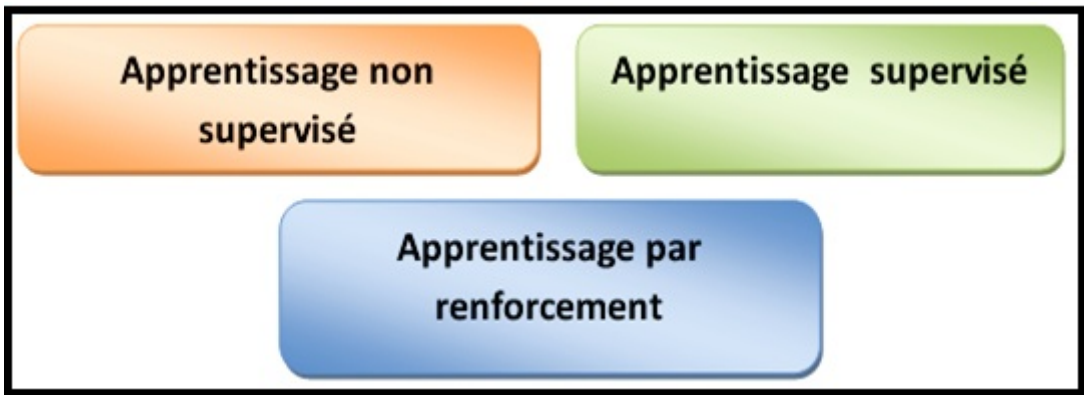

FIGURE 3.5 – *Types de méthodes d'apprentissage automatique.*

3.5.3.1 Apprentissage supervisé

Les algorithmes d'apprentissage supervisé varient d'une application à l'autre ; cependant, ils relèvent de trois catégories : phase d'apprentissage, phase de validation et phase de test[Kotsiantis 2007]. Les algorithmes développés pour un problème particulier dans ces catégories dépendent du concepteur et du développeur de l'algorithme. Les algorithmes d'apprentissage supervisé pour le Big Data sont plus complexes. Ils doivent prendre en compte l'opération physique. Le volume des données est ingérable, le nombre de types de classe est important et la vitesse requise pour traiter les données est élevée. Par conséquent, il a besoin du partage de fichiers distribué, de la technologie de traitement parallèle, des techniques d'apprentissage tout au long de la vie et des techniques d'apprentissage de la représentation sur plusieurs domaines[Chiliang & Wenting 2012]. Dans la méthode d'apprentissage supervisé, un modèle à partir de données d'apprentissage étiquetées est interprété pour permettre la prédiction de données de test[Caruana *et al.* 2008]. L'apprentissage supervisé fait référence à des étiquettes connues (les classes prédites sont connues au préalable) sous le

nom d'un ensemble d'échantillons pour obtenir le résultat souhaité. Il se décompose en trois phases : une phase d'apprentissage, une phase de validation et une phase de test.

1. *La Phase d'apprentissage* se compose d'un ensemble d'exemples avec lesquels l'ordinateur est formé [Abraham & Sathya 2013]. L'ensemble de données utilisé dans cette phase est appelé données d'apprentissage.

2. *La phase de validation* fournit un algorithme pour valider l'efficacité du modèle à l'aide d'un autre jeu de données qui n'a pas été utilisé lors de la phase d'apprentissage [19]. Dans ce cas, le jeu de données s'appelle le jeu de validation, il est également étiqueté. La phase de validation aide à montrer que les paramètres dérivés de la phase d'apprentissage fonctionnent sur la base d'une mesure quantitative. La mesure quantitative joue donc un rôle majeur dans ce processus de validation. Certaines de ces mesures sont l'entropie [20] et l'erreur quadratique moyenne [21]. Si les résultats ne sont pas satisfaisants, le modèle doit être entraîné à nouveau pour obtenir de meilleures valeurs de paramètre. C'est la phase où les problèmes (sélection d'un modèle incorrect ou utilisation d'un algorithme inefficace) peuvent être vus et corrigés. La phase de validation peut également aider à corriger le problème de l'over-fitting [22]. La technique principale utilisée à cet effet s'appelle la validation croisée [23] [Arlot & Celisse 2010].

3. *Phase de test*, cette phase est simple, elle fournit un algorithme pour vérifier si le modèle formé est validé de manière croisée, de sorte que, ce modèle fonctionne avec un autre jeu de données qui n'a pas été utilisé lors des phases d'apprentissage ou de validation [24]. Dans cet algorithme, le jeu de données étiqueté est utilisé uniquement pour comparer les résultats produits par le modèle final en termes d'exactitude de classification et de temps de calcul. À cet effet, plusieurs mesures sont disponibles, elles sont appelées mesure qualitative, car elles sont utilisées pour mesurer la performance du modèle.

19. https ://machinelearningmastery.com/difference-test-validation-datasets/ , (consulté le : 1/04/2019)

20. est une fonction mathématique qui, intuitivement, correspond à la quantité d'information contenue ou délivrée par une source d'information.

21. L'erreur quadratique moyenne (MSE :Mean Squared Error en anglais) est l'espérance du carré de l'erreur entre la vraie valeur et sa valeur estimée.

22. En statistique, over-fitting ou le surapprentissage, ou sur-ajustement, ou encore sur-interprétation est une analyse statistique qui correspond trop étroitement ou exactement à un ensemble particulier de données. Ainsi, cette analyse peut ne pas correspondre à des données supplémentaires ou ne pas prévoir de manière fiable les observations futures.

23. La validation croisée, en anglais cross-validation est, en apprentissage automatique, une méthode d'estimation de fiabilité d'un modèle fondé sur une technique d'échantillonnage.

24. https ://en.wikipedia.org/wiki/Training,_validation,_and_test_setscite_note-cann-faq-8, (consulté le : 1/04/2019)

L'apprentissage supervisé est divisé en deux grandes catégories [Gareth *et al.* 2013] : Classification et régression.

3.5.3.1.1 Classification

Les modèles de classification conviennent au système qui produit des réponses discrètes [Mohamed 2017]. Les algorithmes de classification s'appliquent généralement aux valeurs de réponse nominales. Autrement dit, les réponses sont des variables catégoriques, et les classes possibles sont vraies et fausses. Cependant, certains algorithmes peuvent accueillir des classes ordinales. Les algorithmes que nous discuterons dans notre thèse sont les suivants :

3.5.3.1.1.1 Arbre de décision

Le classificateur ou bien le classifieur arbre de décision est une technique significative de prédiction en considérant une fonctionnalité spéciale appelée interprétabilité. Ce modèle décompose les données pour permettre un processus de prise de décision approprié et pour répondre à une série de requêtes permettant d'interpréter les fonctionnalités. Le modèle d'arbre de décision apprend des caractéristiques du jeu d'apprentissage pour déduire les étiquettes de classe des échantillons [Raschka 2015]. L'algorithme arbre de décision commence à partir du nœud racine, il divise l'ensemble, ce qui génère un gain d'informations plus important. Il s'agit d'un processus itératif consistant à partitionner chaque nœud enfant jusqu'à ce que les feuilles soient pures [Kotsiantis 2007][Rokach & Maimon 2005]. Les échantillons de chaque nœud qui appartiennent parfois à la même classe entraînent un sur-apprentissage (over-fitting) .

3.5.3.1.1.2 Forêt Aléatoire

Une forêt aléatoire est une combinaison de divers classificateurs d'arbres de décision qui convient à un jeu de données en tant que méta-estimateur pour la précision prédictive et le contrôle de sur-apprentissage. Cette technique d'apprentissage est applicable aux modèles de classification et de régression [Liaw & Wiener 2002a]. Dans le cas de la classification, le nombre par défaut de variables choisies pour chaque division est $\sqrt{m}$, où m est le nombre total de variables. Cette méthode fonctionne sur de nombreux arbres de décision en tant qu'apprentissage, ce qui donne une classe d'arbres individuels pour former une sortie. Une forêt aléatoire considère l'importance des caractéristiques ; ces caractéristiques sont mesurées en tant qu'impureté moyenne. Les impuretés extraites des arbres de décision sont diminuées [Brownlee 2016]. Dans une forêt aléatoire, chaque nœud est divisé en utilisant le meilleur parmi un sous-ensemble de classifieurs choisis au hasard sur ce nœud. Ce clas-

sifieur prouve dans une certaine mesure qu'il est très puissant par rapport à de nombreux autres classifieurs, y compris l'analyse discriminante, les machines à vecteurs support et les réseaux de neurones. Les forêts aléatoires sont robustes contre le sur-apprentissage (over-fitting) [Breiman 2001]. (Voir le détail dans le 5ème chapitre). Les forêts aléatoires est une excellente technique pour traiter les problèmes de classification de données volumineuses en raison de sa structure parallélisée flexible qui répond aux exigences des technologies de données volumineuses modernes telles que Hadoop et Spark. L'utilisation de sous-espaces et l'amorçage pour créer des arbres peuvent résoudre les problèmes d'évolutivité associés aux applications Big Data [Li *et al.* 2012].

3.5.3.1.1.3 Machines à Vecteurs Support

Machines à Vecteurs Support (SVM) est un algorithme d'apprentissage automatique supervisé extrêmement populaire pour le traitement d'images. SVM implémente conceptuellement l'idée suivante : générer une limite de décision avec de grandes marges qui est la plus petite distance entre les points d'échantillon et la limite de séparation [Meyer 2015]. Au cours de la dernière décennie, SVM a été progressivement intégré dans le domaine Big Data. Il permet de résoudre les problèmes de classification des données volumineuses. En particulier, cela peut aider les applications multi domaines dans un environnement Big Data. Cependant, la machine à vecteurs de support est mathématiquement complexe et onéreuse en calcul [Demidova *et al.* 2016].

3.5.3.1.1.4 Régression logistique

La régression logistique est un modèle de classification dans lequel les résultats sont des classes discrètes plutôt que des valeurs continues. Par exemple, un client arrivera ou non, il achètera le produit ou non, etc. C'est un algorithme qui provient de la statistique qui est utilisé pour les problèmes de classification supervisés. Dans la régression logistique, nous cherchons le vecteur β de paramètres dans l'équation qui minimise la fonction de coût. En méthodologie statistique, il utilise la méthode du maximum de vraisemblance [25] - *likelihood*- pour calculer le paramètre de variables individuelles. En revanche, dans la méthodologie d'apprentissage automatique, la perte de log sera minimisée par rapport aux coefficients β (également appelés poids). La régression logistique a un biais élevé et une faible erreur de variance. La régression logistique applique une estimation du maximum de vraisemblance après avoir trans-

25. La vraisemblance est une fonction des paramètres d'un modèle statistique calculée à partir de données observées.

formé la variable dépendante en une variable logit [26] par rapport aux variables indépendantes. De cette manière, la régression logistique estime la probabilité qu'un certain événement se produise. La régression logistique fonctionne sur le principe de l'estimation du maximum de likelihood [27] ; en recherchant les valeurs de paramètres maximisant la probabilité d'effectuer les observations. Il s'agit donc de rechercher les paramètres maximisant la probabilité d'événement p à 1 et, la probabilité de non événement (1-p) à 0, comme vous le savez : probabilité (événement + non-événement) = 1 [Hyeoun-Ae 2013][Darlington 1990][Allison 2014]. Ce classifieur revêt également une importance considérable dans le monde d'apprentissage automatique et du Big Data [Prabhat & Khullar 2017].

3.5.3.1.1.5 K-plus proches voisins

K-Plus Proches Voisins (K-PPV ou K-NN : K-Nearest Neighbor) est un modèle d'apprentissage automatique non paramétrique qui stocke une observation d'instructions pour classer des données de test invisibles. K-NN pour la classification consiste à calculer le mode des valeurs de ses k plus proches voisins [28]. On peut aussi l'appeler apprentissage par instance. Il est souvent utile de prendre en compte plusieurs voisins de sorte que la technique est plus communément appelée classification k-plus proche voisin (k-NN) où k voisins les plus proches sont utilisés pour déterminer la classe. Ce modèle est également appelé apprentissage paresseux, car il n'apprend rien pendant la phase d'apprentissage comme la régression logistique, les forêts aléatoires, etc. Au lieu de cela, il commence à travailler uniquement pendant la phase de test / validation pour comparer les données de test avec les données d'apprentissage les plus proches, ce qui prendra un temps considérable pour comparer chaque point de données de test [Imandoust & Bolandraftar 2013][Cunningham & Delany 2007].
Par conséquent, cette technique n'est pas efficace pour le Big Data ; de plus, les performances se détériorent lorsque le nombre de variables est élevé en raison de la malédiction de la dimensionnalité.

3.5.3.1.1.6 Réseaux de neurones artificiels-ANN

Comme leur nom l'indique, les réseaux de neurones artificiels sont des réseaux informatiques qui tentent de simuler de manière générale les réseaux de cellules nerveuses (neurones) du système nerveux central biologique (humain ou animal). Cette simulation est une simulation de cellule par cellule (neurone par neurone, élé-

26. logit :logarithme népérien de la chance que la variable dépendante apparaisse ou non.

27. L'estimation du maximum de likelihood est une méthode d'estimation des paramètres d'un modèle à partir d'observations.

28. https ://mrmint.fr/introduction-k-nearest-neighbors, (consulté le : 02/04/2019)

ment par élément). Ils empruntent aux connaissances neurophysiologiques [29] sur les neurones biologiques et aux réseaux de tels neurones biologiques. Les ANN, comme l'être humain, apprennent par l'exemple. Un ANN est configuré pour une application spécifique, telle que la reconnaissance de formes ou la classification de données via un processus d'apprentissage. L'apprentissage dans les systèmes biologiques implique des ajustements aux connexions synaptiques qui existent entre les neurones. Néanmoins, nous soulignons que la simulation ordonnée par les réseaux de neurones est très grossière [Rabuñal & Dorado 2006]. Les réseaux de neurones artificiels constituent un critère réaliste dans le domaine du Big Data, la connaissance de ce domaine revêt donc une importance primordiale pour ceux qui souhaitent extraire des informations significatives des grandes bases de données disponibles à ce jour [Ossa 2017].

3.5.3.1.1.7 Bayésien naïf (Naïve Bayes)

Le classificateur Naïve Bayes est un algorithme d'apprentissage supervisé dans lequel les attributs utilisés pour prédire la classe sont considérés conditionnellement indépendants compte tenu de la classe. L'architecture sous-jacente de Naïve Bayes dépend de la probabilité conditionnelle. Cela crée des arbres en fonction de leur probabilité de se produire. Ces arbres sont également connus sous le nom de réseau bayésien. En termes simples, un classifieur Naive Bayes suppose que la présence ou l'absence d'un attribut particulier d'une classe n'est pas lié à la présence ou à l'absence d'un autre attribut [Kaur & Oberai 2014]. Une caractéristique très intéressante de l'algorithme Naïve Bayes est qu'il est extrêmement utile pour générer des données synthétiques lorsque les données réelles sont insuffisantes. Ce classifieurs peut également être utilisé dans le domaine Big Data. Cependant, il reste faible par rapport à d'autres classifieurs[Prabhat & Khullar 2017].

3.5.3.1.2 Régression

Les algorithmes qui développent un modèle basé sur des équations ou des opérations mathématiques sur les valeurs prises par les attributs d'entrée pour produire une valeur continue représentant la sortie sont appelés algorithmes de régression [Gareth *et al.* 2013]. L'entrée de ces algorithmes peut prendre des valeurs continues et discrètes en fonction de l'algorithme, alors que la sortie est une valeur continue. Nous allons décrire en détail les algorithmes de régression les plus couramment utilisés :

3.5.3.1.2.1 Régression linéaire

29. La neurophysiologie est l'étude des fonctions du système nerveux, reposant sur tous les niveaux de description, du niveau moléculaire jusqu'au niveau le plus intégré des réseaux neuronaux.

La régression linéaire est un algorithme d'apprentissage automatique supervisé dans lequel la sortie prévue est continue. Ce classifieur tente de modéliser la relation entre deux variables en ajustant une équation linéaire aux données observées, de sorte qu'une variable est considérée comme une variable indépendante et l'autre est considérée comme une variable dépendante. Cet algorithme permet de formuler l'équation de relation linéaire entre variable dépendante et variable indépendante. Dans ce cas, il donne les meilleurs résultats lorsqu'il existe une dépendance linéaire entre les données [30] [Güler & Uyanık 2013].

3.5.3.1.2.2 Perceptron multicouche

Le perceptron multicouche (MLP) est un réseau de neurones formé à l'aide de la rétro-propagation. Les MLP sont constitués de plusieurs couches d'unités de calcul connectées de manière anticipée pour former une connexion dirigée entre les unités inférieures et une unité de la couche suivante. La structure de base de MLP consiste une couche d'entrée, une ou plusieurs couches cachées et une couche de sortie. Les couches cachées sont dites cachées car leur sortie est utilisée uniquement sur le réseau et n'est pas visible en dehors du réseau. Les couches cachées et la couche de sortie sont basées sur des unités sigmoïdes [Balaji & Baskaran 2013][Kiyak & Timus 2015]. Une unité sigmoïde calcule une combinaison linéaire de son entrée, puis on applique la fonction sigmoïde au résultat. La fonction sigmoïde, pour l'entrée x est :

$$sigmoid(x) = \frac{1}{(1 + e^{-x})}$$

La sortie d'une unité sigmoïde, $sigmoid(x)$ est une fonction continue de son entrée x et se situe dans l'intervalle de 0 à 1.

En outre, il existe des algorithmes jouant les deux rôles : la classification et la régression, nous citons certains de ces algorithmes :

— Random Forests pour la régression fonctionne de la même manière que la classification, à ceci près que la taille du sous-ensemble aléatoire pour le fractionnement est maintenant $m/3$ et que le meilleur est choisi avec un critère différent [Liaw & Wiener 2002b].

— SVM peut également être utilisé comme méthode de régression, en conservant toutes les fonctionnalités principales qui caractérisent l'algorithme (marge maximale). La régression à vecteurs supports (SVR) utilise les mêmes principes que le SVM pour la classification avec seulement quelques différences mineures. Tout d'abord, étant donné que la sortie est un nombre réel, il devient très difficile de

30. https://www.vertica.com/docs/9.2.x/HTML/Content/Authoring/AnalyzingData/MachineLearning/LinearReg ,(consulté le : 02/04/2019)

prévoir les informations disponibles, ce qui offre des possibilités infinies, pace que le modèle produit par SVR ne dépend que d'un sous-ensemble des données d'apprentissage [Parrella 2007]. Cependant, l'idée principale est toujours la même : minimiser les erreurs, individualiser l'hyperplan qui maximise la marge en gardant à l'esprit qu'une partie de l'erreur est tolérée.

— K-NN pour la régression consiste à calculer la moyenne des valeurs de ses k plus proches voisins. Il peut être utile de pondérer les contributions des voisins, de sorte que les voisins les plus proches contribuent davantage à la moyenne que les plus éloignés [31] [Imandoust & Bolandraftar 2013].

— Le modèle bayésien naïf a été utilisé avec succès dans des problèmes de classification où la variable de classe est discrète. Tandis que ce modèle a également été appliqué à des problèmes de régression, de sorte que la distribution des variables de la classe est généralement gaussienne multivariée [Frank *et al.* 2000]. Le modèle bayésien naïf est plus puissant dans des problèmes de classification par rapport des problèmes de régression.

3.5.3.2 Apprentissage non supervisé

La méthode d'apprentissage non supervisé traite des données sans étiquette ou d'une structure inconnue. Dans ce type d'apprentissage, les algorithmes apprennent par eux-mêmes sans supervision ni aucune variable cible fournie. Il s'agit de trouver des modèles et des relations cachés dans les données fournies, de sorte que, l'apprenant reçoit exclusivement des données d'apprentissage non étiquetées, il fait des prédictions pour tous les points non vus. Le manque de direction de l'algorithme d'apprentissage dans l'apprentissage non supervisé peut parfois être avantageux, car il permet à l'algorithme de rechercher des modèles qui n'avaient pas encore été pris en compte [Kohonen & Simula 1996]. Il peut être difficile d'évaluer quantitativement les performances d'un apprenant.

En outre, le regroupement (clustering) et la réduction de la dimensionnalité sont des catégories de problèmes d'apprentissage non supervisés [Dayan 1999].

3.5.3.2.1 Clustering

Le clustering est l'une des techniques les plus largement utilisées pour l'analyse exploratoire de données. Dans toutes les disciplines, des sciences sociales à la biologie, en passant par l'informatique, les gens essaient de se faire une première idée de leurs données en identifiant des

31. https ://mrmint.fr/introduction-k-nearest-neighbors, (consulté le : 02/04/2019)

groupes significatifs parmi les points de données. Par exemple, les biologistes informatiques regroupent des gènes sur la base de similitudes dans leur expression dans différentes expériences, les détaillants regroupent les clients sur la base de leurs profils à des fins de marketing ciblé, et les astronomes regroupent les étoiles sur la base de leur proximité spatiale. Le point à clarifier est naturellement, qu'est-ce que le clustering ? Intuitivement, la mise en cluster consiste à grouper un ensemble d'éléments en sous-ensembles homogènes, de sorte que des éléments similaires se retrouvent dans le même groupe et que des éléments différents soient séparés en différents groupes [32]. Clairement, cette description est assez imprécise et peut-être ambiguë. De manière assez surprenante, il n'est pas du tout évident de parvenir à une définition plus rigoureuse. Sur le plan mathématique, la similarité (ou la proximité) n'est pas une relation transitive, tandis que le partage de clusters est une relation d'équivalence et, en particulier est une relation transitive. Plus concrètement, il se peut qu'il existe une longue séquence d'objets x_1 ,. . . , x_m tel que chaque x_i est très similaire à ses deux voisins, x_{i-1} et x_{i+1}, mais x_1 et x_m sont très différents. Si nous voulons assurer que lorsque deux éléments similaires partageant le même cluster ; alors nous devons placer tous les éléments de la séquence dans le même cluster. Cependant, dans ce cas, nous nous retrouvons avec des éléments dissemblables (x_1 et x_m) partageant un cluster, violant ainsi la seconde condition. On distingue quatre types d'algorithmes de clustering : les algorithmes hiérarchiques, les algorithmes de partitionnement, les algorithmes basés sur la densité et les algorithmes basés sur les graphes.

3.5.3.2.1.1 Les méthodes hiérarchiques

Dans les méthodes hiérarchiques les données sont regroupées hiérarchiquement sous la forme d'un arbre (ou dendrogramme), les nœuds de l'arbre issus d'un même parent forment des clusters, de sorte que, une arborescence ou une hiérarchie de clusters est construite de manière incrémentielle qui décrit la paire de clusters fusionnés à chaque itération. En supposant que les objets de données soient les feuilles de l'arborescence hiérarchique ; l'arborescence hiérarchique est construite en utilisant deux algorithmes : L'algorithme agglomératif (ascendante) ou l'algorithme divisif (descendante). Ces algorithmes sont décrits ci-dessous :

3.5.3.2.1.1.1 L'algorithme agglomératif

L'algorithme agglomératif fonctionne de manière ascendante. Autrement dit, chaque objet est initialement considéré comme un cluster à un seul élément (feuille). À chaque étape de l'algorithme, les deux groupes les plus similaires sont combinés dans un nouvel groupe plus grand (nœuds). Cette procédure est itérée jusqu'à ce que tous les points soient membres d'un seul grand cluster (racine) [33] [Omran *et al.* 2007].

32. https ://www.geeksforgeeks.org/clustering-in-machine-learning/, (consulté le : 03/04/2019)
33. https ://en.wikipedia.org/wiki/Hierarchical_clustering, (consulté le : 03/04/2019)

3.5.3.2.1.1.2 L'algorithme divisif

L'algorithme divisif construise la hiérarchie des groupes dans une approche descendante en commençant par le niveau supérieur, où tous les modèles sont enfermés dans un cluster unique, puis en séparant successivement des groupes jusqu'à ce que chaque modèle se trouve dans un cluster unique. Un algorithme divisif nécessite l'évaluation de toutes les $2^{C-1} - 1$ divisions possibles du cluster C en deux sous-clusters [34][Chavent *et al.* 2007].

3.5.3.2.1.2 Les algorithmes de partitionnement

Avec un jeu de données de n objets, il construit k partitions de données. Chaque objet doit appartenir à un seul groupe. Chaque groupe doit contenir au moins un objet. La technique de partition peut améliorer la technique de déplacement itératif en extrayant des objets d'un groupe à un autre. L'objectif principal de l'algorithme de partitionnement est de diviser les points de données en K partitions. Chaque partition reflètera un cluster. Il existe principalement quatre types d'algorithmes de partitionnement, à savoir l'algorithme K-Moyennes, l'algorithme K-Medoid ou PAM *(Portioning Around Medoids)*, l'algorithme CLARA *(Clustering LARge Applications)*, l'algorithme CLARANS *(Clustering LARge Applications upon RANdomized Search)*, et ainsi de suite [Khanali & Vaziri 2016].

3.5.3.2.1.2.1 K-Moyennes

K-moyennes *-K-means-* proposé par[MacQueen 1967], il est appelé aussi l'algorithme des centres-mobiles, qui a été généralisé en 1971 par Diday, Il a été nommé : la méthode des nuées dynamiques. K-means est un algorithme d'apprentissage automatique non supervisé le plus couramment utilisé pour la partition d'un ensemble de données donné en un ensemble de (k) clusters , où k représente le nombre de clusters prédéfinis par l'utilisateur. Il classe les objets en plusieurs clusters (groupes), de sorte que les objets d'une même groupe soient aussi proches que possible (c.-à-d., forte similarité entre les classes), tandis que les objets de différentes groupes sont aussi différents que possible (c.-à-d., faible similarité entre les classes). Dans le *clustering k-means*, chaque cluster est représenté par son centre *centroïde* qui correspond à la moyenne des points attribués au cluster [35] [Cardie & Wagstaff 2001]. La méthode *K-means* est sensible aux points de données anormaux et aux valeurs aberrantes [Gennari *et al.* 1989].

34. https ://en.wikipedia.org/wiki/Hierarchical_clustering, (consulté le : 03/04/2019)
35. https ://www.datascience.com/blog/k-means-clustering, (consulté le : 03/04/2019)

3.5.3.2.1.2.2 K-medoids (PAM)

PAM a été développé par [Kaufman & Rousseeuw 1990]. Cet algorithme est une approche de regroupement liée au regroupement de k-moyennes permettant de partitionner un ensemble de données en k clusters. Dans le clustering k-medoids, chaque cluster est représenté par l'un des points de données du cluster. Cet algorithme fonctionne efficacement pour les petits ensembles de données, mais ne s'adapte pas correctement pour les grands ensembles de données [Agrawal *et al.* 1998]. Il est comparativement plus robuste que K-Mean, en particulier dans le contexte du "bruit" ou du " valeur aberrante "; car il utilise des médoïdes [36] comme centre de cluster au lieu de moyens (utilisés dans les k-moyennes).

3.5.3.2.1.2.3 CLARA

CLARA été développé par [Kaufman & Rousseeuw 1990]. Cet algorithme de partitionnement est une extension des méthodes k-medoids permettant de traiter des données contenant un grand nombre d'objets (plus de plusieurs milliers d'observations) afin de réduire le temps de calcul. Cette technique sélectionne arbitrairement les données à l'aide de PAM, ceci est réalisé en utilisant une méthode d'échantillonnage. Au lieu de rechercher des médoïdes pour l'ensemble du jeu de données; CLARA considère un petit échantillon de données de taille fixe et en suite applique l'algorithme PAM pour générer un ensemble optimal de médoïdes pour l'échantillon. La qualité des médoïdes résultants est mesurée par la dissimilarité moyenne entre chaque objet de l'ensemble de données et le médoïde de son cluster. CLARA répète les processus d'échantillonnage, il rassemble en cluster un nombre de fois pré-spécifié afin de minimiser le biais d'échantillonnage [Frank *et al.* 2000].

3.5.3.2.1.2.4 CLARANS

CLARANS a été développé [Ng & Han 1994] . Il est introduit Pour surmonter les limites de l'algorithme K-Medoid. CLARANS n'est pas très efficace pour les grands ensembles de données. Mais, il est plus efficace et évolutif que PAM et CLARA [Pandya & Saket 2016].

3.5.3.2.1.3 Les algorithmes basés sur la densité

Contrairement aux méthodes hiérarchiques et de partitionnement, dans lesquelles des groupes reposent sur les distances des objets, ainsi que ces méthodes ne fonctionnent bien que pour des groupes bien séparés, elles sont également gravement affectées par la présence de

36. En statistiques, un médoïde est le représentant le plus central d'une classe. C.à.d. Objet ayant la plus faible dissimilitude aux autres objets d'un groupe.

bruit et des valeurs aberrantes. Les algorithmes de densité construisent les groupes sur la base d'un concept de densité. Cela signifie que les groupes sont situés dans des régions ayant des densités à peu près homogènes [Gennari *et al.* 1989]. La méthode la plus populaire basée sur la densité est DBSCAN (Density-Based Spatial Clustering and Application with Noise). L'algorithme DBSCAN a été introduit pour la première fois par[Ester *et al.* 1996], il repose sur une notion de grappes basée sur la densité. Les grappes sont identifiées en examinant la densité de points. Les régions à forte densité de points illustrent l'existence de grappes, tandis que les régions à faible densité de points indiquent les grappes de bruit ou les grappes de valeurs aberrantes. Cet algorithme est particulièrement adapté pour traiter de grands ensembles de données avec du bruit, il est capable d'identifier des grappes de différentes tailles et formes [Hinneburg & Keim 1998]. L'idée principale de l'algorithme DBSCAN est que pour chaque point d'un groupe ; le voisinage d'un rayon donné doit contenir au moins un nombre minimal de points, c-à-d., que la densité dans le voisinage doit dépasser un seuil prédéfini.

3.5.3.2.1.4 Les algorithmes basés sur les graphes

Une autre approche proposée par de nombreux auteurs est de considérer les relations de distance ou de similarité entre données comme un graphe [Matula 1977][Leahy & Wu 1993]. Cette représentation est indépendante du type des données (vecteurs, textes, images, etc.) puisque seule la similarité entre données est utilisée. Dans ce type de graphe, chaque nœud représente une donnée, et chaque poids associé à un arc représente la distance ou la similarité entre les deux données connectées [Hartuv & Shamir 1999]. Un cluster est alors défini comme un ensemble de données fortement connectées entre elles et faiblement connectées aux autres données. Une telle représentation permet l'utilisation d'outils issus de la théorie des graphes [Guha *et al.* 1998]. Nous présentons dans cette section trois algorithmes récents se basant sur une représentation sous forme de graphe qui présentent de bonnes performances : Chameleon [Karypis *et al.* 1999a], le Clustering Spectral [Ng *et al.* 2001][Malik & Shi 2000] et la propagation d'affinité [Frey & Dueck 2007].

3.5.3.2.1.4.1 Chameleon

Chameleon [Karypis *et al.* 1999a] est une méthode ascendante hiérarchique basée sur une représentation sous forme de graphe des k plus proches voisins. Il fonctionne sur un graphe fragmenté dans lequel les nœuds représentent des éléments de données, et les arêtes pondérées représentent des similitudes entre les éléments de données. Cette représentation graphique fragmentée du jeu de données permet à Chameleon de s'adapter à des jeux de données volumineux et de fonctionner avec succès sur des jeux de données disponibles ; uniquement dans l'espace de similarité et non

dans les espaces métriques [Ganti *et al.* 1999]. Chameleon trouve les grappes dans le jeu de données en utilisant un algorithme à deux phases. Au cours de la première phase, Chameleon utilise un algorithme de partitionnement graphique pour regrouper les éléments de données en un grand nombre de sous-groupes relativement petits. Au cours de la deuxième phase, il utilise un algorithme de classification hiérarchique agglomératif pour trouver les véritables grappes en combinant de manière répétée ces sous-groupes [Karypis *et al.* 1999b].

3.5.3.2.1.4.2 Clustering Spectral

L'algorithme de regroupement spectral peut être classé selon deux approches : le regroupement récursif bidirectionnel, et le regroupement direct k-voies. La première approche trouve le vecteur propre de Fiedler d'une matrice Laplacienne d'un graphe G et partitionne récursivement G jusqu'à ce qu'une partition à k-voies soit trouvée. Tandis que, la deuxième approche utilise les premiers vecteurs propres d$\geq$k , elle trouve directement une partition à l'aide de méthodes heuristiques. L'un des aspects positifs de ces méthodes est la possibilité de définir des limites supérieures ou inférieures pour la fonction objective des problèmes de partitionnement des graphes [Luxburg 2006][Sang *et al.* 2014].

3.5.3.2.1.4.3 Propagation d'affinité

La propagation d'affinité prend en entrée un ensemble de similitudes par paires entre les points de données, elle recherche les grappes sur la base de la maximisation de la similarité totale entre les points de données et leurs exemples. La similarité peut être simplement définie comme une distance euclidienne au carré négatif pour la compatibilité avec d'autres algorithmes, elle peut incorporer des modèles plus riches spécifiques à un domaine. Les exigences de calcul et de mémoire de la propagation d'affinité s'échelonnent linéairement avec le nombre de similitudes entrées ; pour les problèmes non épars où toutes les similitudes possibles sont calculées, ces exigences s'échelonnent quadratiquement avec le nombre de points de données. La propagation d'affinité est démontrée dans plusieurs applications de domaines tels que l'imagerie et la bioinformatique [Frey & Dueck 2007].

3.5.3.2.2 Réduction de dimensions

La réduction de dimensions consiste à transformer une représentation initiale d'éléments en une représentation de dimension inférieure tout en préservant certaines propriétés de la représentation initiale ; de sorte que, elle prendre des données dans un espace de grande dimension et à

les cartographier dans un nouvel espace dont la dimensionnalité est beaucoup plus petite. Ce processus est étroitement lié au concept de compression (avec perte) dans la théorie de l'information. Il existe trois manières différentes de réduire la dimensionnalité. La première est la sélection des fonctionnalités qui consiste généralement à parcourir les fonctionnalités disponibles et à déterminer si elles sont réellement utiles ; c-à-d., corrélées aux variables de sortie. Bien que de nombreuses personnes utilisent les algorithmes d'apprentissage supervisé parce qu'elles ne veulent pas «se salir les mains» et consulter les données elles-mêmes. La deuxième est la dérivation des nouvelles caractéristiques à partir des anciennes, généralement en appliquant des transformations au jeu de données qui changent simplement les axes (système de coordonnées) du graphe en les déplaçant et en les faisant pivoter. ce jeu de données peut simplement être écrit sous forme de matrice que nous appliquons aux données. La réduction de la dimensionnalité s'explique par le fait qu'elle nous permet de combiner des fonctions et d'identifier celles qui sont utiles et celles qui ne le sont pas. La troisième et la dernière consiste simplement à utiliser le clustering afin de regrouper des points de données similaires et de voir si cela permet d'utiliser moins de fonctionnalités.

Il existe plusieurs raisons pour réduire la dimensionnalité des données. Premièrement, les données de grande dimension posent des problèmes de calcul. De plus, dans certaines situations, une dimensionnalité élevée peut conduire à de faibles capacités de généralisation de l'algorithme d'apprentissage (par exemple, dans le classificateur K-PPV, la complexité de l'échantillon augmente de manière exponentielle avec la dimension). En plus, la réduction de dimensionnalité peut être utilisée pour l'interprétabilité des données, pour rechercher une structure significative des données et pour des fins d'illustration. Nous citons ci-dessous les méthodes les plus courantes de réduire les dimensions. Les techniques de réduction de dimension sont basées sur les méthodes linéaires et les méthodes non linéaires.

La réduction des dimensions est l'une des plus vieilles approches permettant d'apporter des éléments de réponse à ce problème. Son objectif est de sélectionner ou d'extraire un sous-ensemble optimal de caractéristiques pertinentes pour un critère fixé auparavant. La sélection de ce sous-ensemble de caractéristiques permet d'éliminer les informations non-pertinentes et redondantes selon le critère utilisé. Cette sélection/extraction permet donc de réduire la dimension de l'espace des exemples et de rendre l'ensemble des données plus représentatif du problème. En effet, les principaux objectifs de la réduction de la dimension sont :

- Faciliter la visualisation et la compréhension des données ;
- Réduire l'espace de stockage nécessaire ;
- Réduire le temps d'apprentissage et d'utilisation ;
- Identifier les facteurs pertinents.

Dans ce qui suit, nous allons présenter les principales méthodes linéaires et non linéaires utilisées pour la réduction de dimensions :

3.5.3.2.2.1 Méthodes linéaires de réduction des dimensions

Nous rappelons brièvement les principes de quatre méthodes classiques d'analyse de données, qui sont le fondement de plusieurs méthodes non linéaires plus récentes.

3.5.3.2.2.1.1 Analyse en composantes principales –ACP

L'analyse en composantes principales (ACP) est la technique de réduction de la dimensionnalité qui compte de nombreux utilitaires. Il réduit les dimensions d'un jeu de données en projetant les données sur un sous-espace de dimension inférieure [Jolliffe & Cadima 2016]. Parce que, il est difficile de comprendre la structure de données comportant des centaines de dimensions. Par conséquent, en réduisant les dimensions en 2D ou en 3D, les observations peuvent être facilement visualisées. Par exemple, un jeu de données à deux dimensions peut être réduit en projetant les points sur une ligne. Chaque instance du jeu de données serait alors représentée par une valeur unique plutôt que par une paire de valeurs. De manière similaire, un jeu de données à trois dimensions pourrait être réduit à deux dimensions en projetant des variables sur un plan. Donc, ACP permet d'atténuer le cours de la dimensionnalité et, de compresser les données tout en minimisant les informations perdues en même temps [Dharwat 2016].

La question posée est de savoir comment choisir les axes ? L'une des méthodes utilisées dans ce contexte est l'idée de la direction dans des données présentant la plus grande variation ; de sorte que, l'algorithme concentre d'abord sur les données en soustrayant la moyenne, puis on choisit la direction avec la plus grande variation, et on place un axe dans cette direction, ensuite on examine la variation restante, de sorte que, on trouve un autre axe orthogonal avec le premier axe. Il itère ensuite ceci jusqu'à ce qu'il soit à court d'axes possibles. Le résultat final est que toutes les variations sont situées sur les axes de l'ensemble de coordonnées, la matrice de covariance est donc diagonale : chaque nouvelle variable n'est pas corrélée avec toutes les variables sauf ses variables. Certains des derniers axes trouvés ont très peu de variation et peuvent donc être supprimés sans affecter la variabilité des données.

3.5.3.2.2.1.2 Analyse discriminante linéaire -ADL

L'analyse discriminante linéaire (ADL) est une méthode utilisée en statistique et en apprentissage automatique pour trouver la combinaison linéaire de caractéristiques qui séparent au mieux deux classes d'objet ou d'événement [Rusdiana 2016]. La combinaison résultante peut être utilisée pour la réduction de la dimensionnalité avant la classification ultérieure (pour assurer une classification plus efficace). L'analyse discriminante est similaire à la régression, sauf que la variable dépendante est catégorielle plutôt que continue. Dans l'analyse discriminante, l'intention est de prédire l'appartenance à une classe d'observations individuelles en fonction d'un ensemble de variables prédictives. ADL tente généralement de trouver des combinaisons linéaires de variables prédictives qui séparent le mieux les groupes d'observations [Kenobi *et al.* 2017]. L'ADL est également étroitement liée à l'analyse en composantes principales (ACP), et à l'analyse factorielle en ce sens qu'elles recherchent des combinaisons linéaires de variables qui expliquent le mieux les données. ADL tente explicitement de modéliser la différence entre les classes de données. En revanche, ACP ne prend en compte aucune différence de classe, ainsi que l'analyse factorielle construit les combinaisons d'entités en fonction des différences plutôt que des similitudes. L'analyse discriminante se distingue également de l'analyse factorielle en ce sens qu'il ne s'agit pas d'une technique d'interdépendance ; une distinction doit être faite entre les variables indépendantes et les variables dépendantes.

3.5.3.2.2.1.3 Analyse factorielle

L'analyse factorielle est une méthode statistique à plusieurs variables dont l'objectif principal est de simplifier les relations complexes et diverses existantes entre un ensemble de variables observées [Balasundaram 2009]. Cette technique aborde le problème de l'analyse de la structure des interrelations [37] (corrélations) entre un grand nombre de variables (scores de test, éléments de test, réponses au questionnaire, etc.) en définissant un ensemble de dimensions sous-jacentes communes, appelées facteurs ; de sorte que, les données observées peuvent s'expliquer par un plus petit nombre de facteurs non corrélés ou de variables latentes. Ainsi que ces données proviennent d'une source de données sous-jacente (ou d'un ensemble de sources de données) qui n'est pas directement connue. Mais, le problème de l'analyse factorielle est de trouver ces facteurs indépendants et le bruit inhérent aux mesures de chaque facteur. Plus de détails sur la méthode ainsi que sur des descriptions de l'algorithme peuvent être trouvées avec [Child 2006][Mulaik 2010].

37. En probabilités et en statistique, la corrélation entre plusieurs variables aléatoires ou statistiques est une notion de liaison qui contredit leur indépendance.

3.5.3.2.2.1.4 Analyse en Composantes Indépendantes -ACI

L'analyse en composantes indépendantes est une technique statistique et informatique permettant de révéler les facteurs cachés qui sous-tendent des ensembles de variables aléatoires, de mesures ou de signaux. ACI définit un modèle génératif pour les données multivariées qui est généralement donné sous la forme d'une grande base de données d'échantillons. L'objectif est de trouver des composants totalement indépendants et non gaussiens. Sa différence fondamentale avec les méthodes statistiques multi-variables classiques réside dans l'hypothèse de non gaussianité, ce qui permet l'identification de composants sous-jacents originaux contrairement aux méthodes classiques. L'ACI est superficiellement liée à l'analyse en composantes principales et à l'analyse factorielle. La technique ACI est cependant une technique beaucoup plus puissante, capable de trouver les sources ou les facteurs sous-jacents lorsque ces méthodes classiques échouent complètement. Les données analysées par l'ACI pourraient provenir de nombreux types de domaines d'application, notamment les images numériques, les bases de données de documents, les indicateurs économiques et les mesures psychométriques [38]. L'analyse en composantes indépendantes est une méthode statistique bien établie et fiable conduisant à la séparation des signaux. La séparation des signaux est un problème fréquemment rencontré, elle est au cœur du traitement statistique du signal qui a un large éventail d'applications dans de nombreux domaines technologiques ; allant du traitement d'images numériques, les bases de données de documents, les indicateurs économiques et aux mesures psychométriques [Hyvarinen *et al.* 2001].

3.5.3.2.2.2 Méthodes non linéaires de réduction de dimensions

Les méthodes non linéaires arrivent à modéliser la liaison entre les observations et les variables latentes d'une façon plus riche. Cependant leurs modèles engendrent plusieurs paramètres qui requièrent une grande masse de données pour son identification.

3.5.3.2.2.2.1 ACP Kernelisée- KPCA

L'ACP standard ne permet qu'une réduction de la dimensionnalité linéaire. L'un des problèmes de la ACP est qu'elle suppose que les directions de variation sont toutes des lignes droites, une ACP standard ne sera pas très utile. Pour résoudre ce problème, nous pouvons utiliser une très belle extension de la ACP qui utilise l'astuce du noyau (KPCA). KPCA pour *Kernel Principal Component Analysis* est la reformulation non linéaire de la technique linéaire

38. La psychométrie est la science qui étudie l'ensemble des techniques de mesures pratiquées en psychologie, ainsi que les techniques de validation et d'élaboration de ces mesures.

classique qu'est l'analyse en composantes principales en utilisant des fonctions à noyaux. Nous appliquons cette fonction à chaque point de données x qui transforme les données en espace noyau, puis nous effectuons une analyse ACP linéaire normale dans cet espace [Bernhard *et al.* 1999].

3.5.3.2.2.2.2 Cartes de Sammon

Il est souvent nécessaire de réduire la dimensionnalité d'un jeu de données afin de rendre l'analyse traitable, ou de faciliter la visualisation. Elle existe une autre méthode qui joue ce rôle, c'est les cartes de Sammon. Les cartes de Sammon est un algorithme non linéaire permettant de redimensionner des dimensions [Arren 1952]. Le but de cet algorithme est de projeter des données issues d'un espace à n dimension en deux dimensions. L'algorithme essaie de trouver les emplacements dans l'espace cible bidimensionnel de manière à maximiser la conservation de la structure d'origine des vecteurs de mesure dans un espace à n dimensions [Kolehmainen 2004]. Les cartes de Sammon est donc capable de représenter les distances relatives des vecteurs dans un espace de mesure [Kaski 1997].

3.5.3.2.2.2.3 Auto-encodeurs Multicouches

Un auto-encodeur est un réseau de neurones comportant trois couches : une couche d'entrée, une couche cachée (codage) et une couche de décodage. Ce réseau est formé pour reconstruire ses entrées, ce qui oblige la couche cachée à essayer d'obtenir de bonnes représentations des entrées [39]. Les auto-encodeurs appartiennent à la famille des réseaux de neurones, mais ils sont également considérés comme un algorithme d'apprentissage automatique non supervisé similaire à l'APC. Récemment, certaines des méthodes les plus puissantes de l'intelligence artificielle ont impliqué des auto-codeurs multicouches, en particulier dans l'apprentissage en profondeur *-Deep Learning-* . L'apprentissage en profondeur est une technique d'apprentissage automatique avancée dans laquelle plusieurs couches abstraites communiquent entre elles ; Chaque couche est profondément connectée à la couche précédente, elle prend ses décisions en fonction de la sortie alimentée par la couche précédente [Deng & Yu 2013].

3.5.3.2.2.2.4 Cartographie des Caractéristiques Isométriques- Isomap

La méthode de réduction de dimensionnalité Isomap a été introduite par [Langford *et al.* 2000]. Elle s'inspire de la méthode MDS *-Multi-Dimensional Scaling-* pour la connaissance d'une matrice

39. https ://codeburst.io/deep-learning-types-and-autoencoders-a40ee6754663, (consulté le : 05/04/2019)

de dissimilarité entre les paires d'individus, mais en lui donnant comme métrique la distance géodésique [40] (ou curviligne). Isomap est une méthode de réduction de dimensionnalité non linéaire basée sur la théorie spectrale, elle estime la distance géodésique [Saul & Weinberger 2004][Chen *et al.* 2006] de la façon suivante : Dans un premier temps, le voisinage de chacun des points est calculé. Une fois les voisinages connus, un graphe est construit en reliant tous les points voisins. Chaque arrêt du graphe est ensuite pondéré par la distance euclidienne [Saul & Weinberger 2006] entre les deux extrémités de l'arrête. Parce que, dans les variétés non linéaires, la métrique euclidienne de distance est valable si seulement si la structure du voisinage peut être approchée comme linéaire. Si le voisinage contient des trous ; les distances euclidiennes peuvent être très trompeuses. Contrairement à cela, si nous mesurons la distance entre deux points en suivant la variété, nous aurons une meilleure approximation de la distance ou de la proximité de deux points. Enfin, la distance géodésique entre deux points est estimée par la somme des longueurs des arêtes, puis nous choisissons le plus court chemin entre ces points. En pratique, le plus court chemin entre deux sommets du graphe est calculé par l'algorithme de Dijkstra [Javaid 2013].

3.5.3.3 Méthode d'apprentissage par renforcement

Apprentissage par renforcement est entre l'apprentissage supervisé et l'apprentissage non supervisé. Il est souvent considéré comme une branche de l'intelligence artificielle, il a été l'un des sujets centraux dans un large éventail de domaines scientifiques au cours des deux dernières décennies. Dans l'apprentissage par renforcement, les décisions séquentielles doivent être prises plutôt que par une prise de décision unique, ce qui rend par fois, la phase d'apprentissage un peu difficile ; de sorte que, l'algorithme est informé lorsque la réponse est fausse, mais ne dit pas comment le corriger. Il doit explorer et expérimenter différentes possibilités jusqu'à trouver la bonne réponse.

Apprentissage par renforcement vise à trouver une cartographie (mapping) appropriée des situations aux actions dans lesquelles une certaine récompense est maximisée. Il s'agit essentiellement de problèmes en boucle fermée, car les actions du système d'apprentissage influencent ses entrées ultérieures. Apprentissage par renforcement peut être défini comme une classe d'approches de résolution de problèmes dans lesquelles l'apprenant (l'agent) apprend à travers d'une série de recherches par d'essais (erreurs) et de récompenses différées (ultérieures) en considérant ces deux caractéristiques distinctives les plus importantes de l'apprentissage par renforcement. Le but est de maximiser non seulement la ré-

40. En géométrie, une géodésique désigne la généralisation d'une ligne droite sur une surface. En particulier, le chemin le plus court ou un des plus courts chemins, s'il en existe plusieurs, entre deux points d'un espace pourvu d'une métrique est une géodésique.

compense immédiate, mais aussi la récompense cumulée à long terme ; de sorte que, l'agent puisse apprendre à se rapprocher d'une stratégie comportementale optimale en interagissant en permanence avec l'environnement. Dans les cas les plus intéressants et les plus difficiles, les actions peuvent affecter non seulement la récompense immédiate, mais également la situation suivante, et par là même, toutes les récompenses ultérieures [Barto & Sutton 2014]. Le scénario général de l'apprentissage par renforcement est illustré à la figure 3.6. Contrairement au scénario d'apprentissage supervisé envisagé au dessus, l'apprenant ne reçoit pas passivement un jeu de données étiqueté. Au lieu de cela, il collecte des informations par le biais d'actions en interagissant avec l'environnement. En réponse à une *action*, l'apprenant ou l'agent reçoit deux types d'informations : son état (*state*) actuel dans l'environnement et une récompense (*reward*) réelle, spécifique à la tâche et à l'objectif correspondant [Richard 2001].

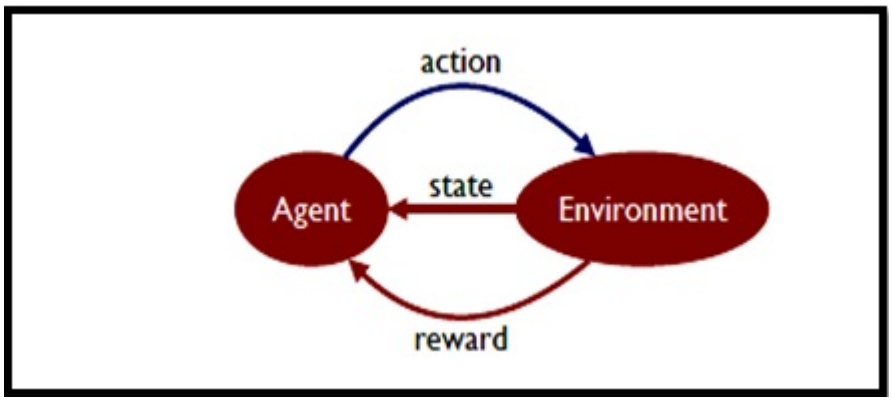

FIGURE 3.6 – *Processus d'apprentissage dans l'apprentissage par renforcement.*

Nous voyons qu'il existe plusieurs différences entre le scénario d'apprentissage par renforcement et le scénario d'apprentissage supervisé examiné précédemment. Contrairement à l'apprentissage supervisé, il n'y a pas de répartition fixée selon laquelle les instances sont dessinées ; c'est le choix d'une politique qui définit la distribution sur les observations. En fait, de légers changements à la politique peuvent avoir des effets dramatiques sur les récompenses reçues. De plus, l'environnement peut ne pas être corrigé, il peut également varier en fonction des actions sélectionnées par l'agent. Ainsi ce type d'apprentissage constituer un modèle plus réaliste pour certains problèmes d'apprentissage que l'apprentissage supervisé. Enfin, nous voyons aussi, contrairement à l'apprentissage supervisé, les phases d'apprentissage par renforcement (apprentissage, test) sont mélangées. Dans cette partie, nous aborderons diverses techniques utilisées dans l'apprentissage par renforcement à l'aide d'exemples pratiques. Les techniques abordées sont les suivantes :

3.5.3.3.1 Processus décisionnels de Markov-MDP

Dans l'apprentissage par renforcement, le MDP *-Markov Decision*

Process- est un cadre mathématique permettant de modéliser la prise de décision d'un agent dans des situations ou des environnements où les résultats sont en partie aléatoires et en partie sous contrôle. Dans ce modèle, l'environnement est modélisé comme un ensemble d'états et d'actions pouvant être effectuées par un agent pour contrôler l'état du système. L'objectif est donc de contrôler le système de manière à maximiser le rendement total de l'agent [Mannor *et al.* 2009][Even-Dar *et al.* 2009].

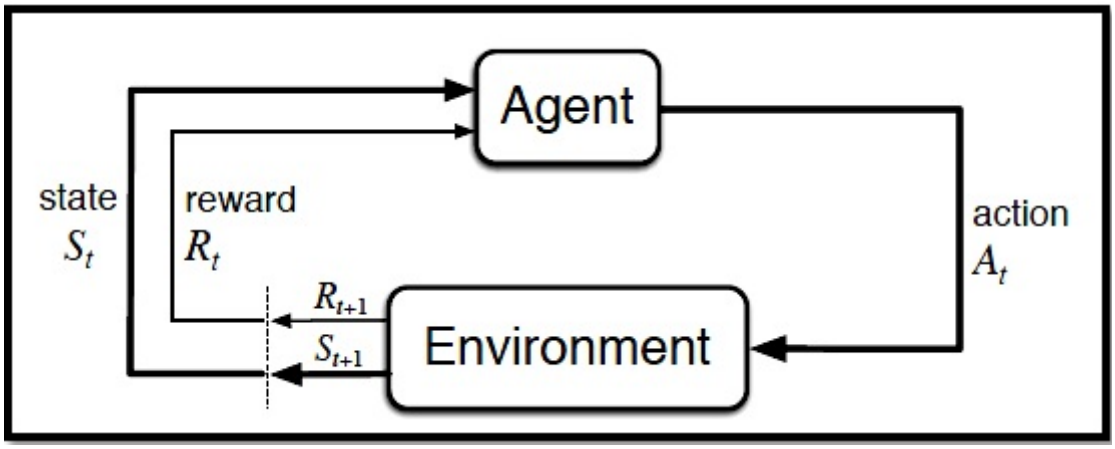

FIGURE 3.7 – *Interaction agent – environnement dans un processus décisionnel de Markov.*

Nous voyons que les deux figures 3.6 et 3.7 sont presque identiques. Pour cela, nous pouvons dire que le processus décisionnel de Markov décrit formellement un environnement d'apprentissage par renforcement. Où :

- L'environnement est entièrement observable ;
- L'état actuel caractérise complètement le processus (ce qui signifie que l'état futur dépend entièrement de l'état actuel plutôt que des états ou des valeurs historiques) ;
- Presque tous les problèmes d'apprentissage par renforcement peuvent être formalisés en tant que MDP (par exemple, le contrôle optimal concerne principalement des MDP continus).

MDP travaille sur la simple propriété markovienne d'un état. Par exemple, S_{t+1} dépend entièrement du dernier état S_t plutôt que des dépendances historiques. Nous invitons les chercheurs intéressés par les MDP de consulter l'un des nombreux excellents ouvrages sur ce sujet, tels que [Bertsekas & Shreve 1978] [Bertsekas 2007] [P.Bertsekas 2007] [Puterman 1994].

3.5.3.3.2 Équations de Bellman

Richard Bellman est un mathématicien américain qui a utilisé les équations pour la formulation mathématique du MDP pour obtenir les politiques optimales de l'environnement. Les équations de Bellman sont omniprésentes dans l'apprentissage par renforcement, elles sont nécessaires pour comprendre le fonctionnement des algorithmes de ce type

d'apprentissage. Les équations de Bellman est une condition nécessaire pour l'optimisation associée à une méthode d'amélioration mathématique appelée « programmation dynamique ». Les équations de Bellman sont des équations linéaires qui peuvent être résolues pour tout l'environnement. Cependant, la complexité en temps pour résoudre ces équations est $O(n^3)$; ce qui devient très coûteux en calcul lorsque le nombre d'états dans un environnement est grand ; et parfois, il n'est pas possible d'explorer tous les états, car l'environnement est très vaste [Mingyuan 2013][Huizhen *et al.* 2018]. L'importance des équations de Bellman est qu'elles nous permettent d'exprimer les valeurs des états en tant que valeurs des autres états, cela signifie que si nous connaissons la valeur de s_{t+1}, alors nous pouvons très facilement calculer la valeur de s_t. , cela ouvre beaucoup de portes aux approches itératives pour calculer la valeur de chaque état, car si nous connaissons la valeur de l'état suivant, nous pouvons connaître la valeur de l'état actuel. Enfin, avec les équations de Bellman en main, nous pouvons commencer à chercher comment calculer des politiques optimales et comment coder notre premier agent d'apprentissage par renforcement [41] [Philip 2018].

3.5.3.3.3 Programmation dynamique-DP

Le terme programmation dynamique *-Dynamic Programming-* désigne un ensemble d'algorithmes pouvant être utilisés pour calculer des politiques optimales à partir d'un modèle parfait de l'environnement sous la forme d'un processus décisionnel de Markov (MDP). La programmation dynamique permet de résoudre de gros problèmes complexes en les décomposant en sous-problèmes plus petits qui sont ensuite résolus indépendamment dont les résultats sont combinés pour former la solution au problème initial [Dreyfus 2002]. Parmi les méthodes de la programmation dynamique, nous citons la méthode *- Divide-and-Conquer-* , cette méthode est un paradigme qui résout de manière récursive des problèmes dont la solution peut être trouvée en termes d'instances plus petites d'elle-même. Avec *Divide-and-Conquer* , il est inévitable qu'un sous-problème soit résolu plus d'une fois (en fait, plusieurs fois) [42]. Alors, le but de la programmation dynamique est d'éliminer les redondances.

3.5.3.3.4 Les méthodes de Monte Carlo

Les méthodes de Monte Carlo sont des moyens de résoudre le problème de l'apprentissage par renforcement basé sur la moyenne des retours d'échantillon. Elles effectuent des sélections aléatoires à partir des échantillons sur la base d'un modèle supposé. Elles ne requièrent que des expériences à travers des séquences d'échantillons d'états, actions et récompenses tirées d'interactions réelles ou simulées avec un environne-

41. https ://joshgreaves.com/reinforcement-learning/understanding-rl-the-bellman-equations/, (consulté le : 05/04/2019)

42. http ://cgi.csc.liv.ac.uk/ ped/teachadmin/algor/dyprog.html, (consulté le : 05/04/2019)

ment [43][Nutini 2017]. En plus précisément, la méthode de Monte Carlo est un sous-ensemble d'algorithmes de calcul qui utilisent le processus d'échantillonnage aléatoire avec remise pour effectuer des estimations numériques de paramètres inconnus. Elles permettent de modéliser des situations complexes impliquant de nombreuses variables aléatoires, elles permettent également d'évaluer l'impact du risque. Les utilisations des méthodes de Monte Carlo sont extrêmement variées, elles ont conduit à un certain nombre de découvertes révolutionnaires dans les domaines de la finance pour déterminer quand lever une option sur un bien financier, de l' assurance pour évaluer le montant d'une prime, de la biologie pour étudier les dynamiques intra et intercellulaires, de la physique nucléaire pour connaître la probabilité qu'une particule traverse un écran, de la télécommunications pour déterminer la qualité de service, ou de façon générale pour déterminer la fiabilité d'un système, sa disponibilité ou son temps moyen d'atteinte de la défaillance [44][Paxton *et al.* 2001].

3.5.3.3.5 Apprentissage par différence temporelle-TD

S'il fallait identifier une idée comme principale et nouvelle dans l'apprentissage par renforcement, ce serait sans doute l'apprentissage par différence temporelle -*Temporal Difference*- . Il a été principalement utilisé pour résoudre le problème de l'apprentissage par renforcement. L'apprentissage par TD est une combinaison d'idées de Monte Carlo et d'idées de programmation dynamique. A l'instar des méthodes de Monte Carlo, la méthode TD peut tirer directement des informations de l'expérience brute sans modéliser la dynamique de l'environnement. C-à-d., il apprend en échantillonnant l'environnement selon certaines règles. TD est donc lié aux techniques de programmation dynamique, car il se rapproche de son estimation actuelle en fonction d'estimations acquises précédemment (un processus appelé bootstrapping) [Richard & Barto 1998][Richard 1988].

3.5.3.3.6 Algorithme Q-learning

Q-Learning est l'une des méthodes d'apprentissage par renforcement largement utilisées en robotique. Dans Q-learning, un agent sélectionne une action parmi toutes les actions possibles dans un état qui suit une politique, il provoque une interaction avec un environnement à un moment donné [Manju & Punithavalli 2011]. Une récompense est attribuée à l'action sélectionnée à partir de l'environnement en tant que valeur scalaire [Christopher 1989]. A ce moment, l'agent renouvelle la base de données en raison de la récompense donnée. En répétant ce processus, les valeurs d'action sont renouvelées et stockées dans chaque état. Après l'apprentissage, un mouvement optimal pour la tâche souhaitée peut être réalisé en sélectionnant simplement les actions avec la valeur d'action la

43. https ://www.sciencedirect.com/topics/computer-science/monte-carlo-method, (consulté le : 05/04/2019)
44. https ://towardsdatascience.com/an-overview-of-monte-carlo-methods-675384eb1694, (consulté le : 05/04/2019)

plus élevée dans chaque état. Dans Q-learning, la convergence vers la solution optimale est promise tant que la série de processus d'apprentissage suit le processus décisionnel de Markov (MDP). Q-learning utilise l'apprentissage par différence temporelle (TD) pour estimer la valeur de Q * (s, a), où a est une action et s est un état, Q * est la valeur attendue (récompense actualisée cumulative) de faire a dans l'état s et ensuite suivre la stratégie optimale [Drugan 2017].

3.6 Modélisation statistique Vs Machine Learning

Bien qu'il existe des similitudes inhérentes entre la modélisation statistique et les méthodologies d'apprentissage automatique *-Machine Learning-* , ce n'est parfois pas évident pour de nombreux praticiens. Nous expliquons brièvement les différences et les similitudes entre les deux concepts. Nous commençons par la modélisation statistique ; elle se caractérise par formalisation des relations entre variables sous forme d'équations mathématiques. Ainsi que, elle prendre la forme de la courbe du modèle avant de procéder à l'ajustement du modèle sur les données (par exemple, linéaire, polynomial, etc.). En plus, la modélisation statistique prédit la sortie avec une précision de 85% et une confiance de 95%, de sorte que les données sont réparties entre 30% et 70% afin de créer des données de l'apprentissage et de test respectivement. Mais, elle peut être développée sur un seul jeu de données appelé données d'apprentissage, car les diagnostics sont effectués à la fois avec une précision globale au niveau des variables individuelles. Principalement, la modélisation statistique est utilisée à des fins de recherche par des chercheurs statisticiens. Tandis que, Machine Learning est un ensemble d'algorithmes pouvant tirer des données sans recourir à la programmation basée sur des règles. Ainsi que, il n'est pas nécessaire d'assumer la forme sous-jacente, car les algorithmes d'apprentissage automatique peuvent apprendre automatiquement des modèles complexes en fonction des données fournies. En plus, l'apprentissage automatique ne fait que prédire le résultat avec une précision mois de 90%, de sorte que les données sont réparties entre 30% et 70% afin de créer des données de l'apprentissage et de test respectivement. Mais, en raison du manque de diagnostics sur les variables, les algorithmes d'apprentissage automatique doivent être formés sur deux jeux de données appelés données d'apprentissage et de validation. Généralement, Machine Learning est très apte à être implémenté dans un environnement de production par des informaticiens.

3.7 Applications de l'apprentissage automatique

Voici quelques applications de l'apprentissage automatique :

— **Marketing :** Comprendre le comportement du site en prévoyant le nombre possible d'utilisateurs visitant la page de l'entreprise pour l'analyse du produit.

— **Optimisation de la publicité :** Les techniques d'optimisation sont applicables à la publicité destinée à promouvoir l'entreprise.

— **Moteur de recommandation :** Techniques d'apprentissage automatique prévoient la préférence des clients en comprenant l'enquête analytique précédente.

— **Analyse du panier de marché :** Interprétation du comportement du client en fournissant des remises personnalisées proposées par l'entreprise pour augmenter la taille du panier.

— **Prévision de désabonnement des clients :** Des clients spécifiques reçoivent une attention supplémentaire de la part des équipes de service afin de les fidéliser.

— **Optimisation opérationnelle :** Les systèmes sont susceptibles d'échouer, par conséquent des contrôles réguliers sont effectués à titre préventif.

— **Maintenance préventive :** Veillez à protéger les opérations et les transactions à intervalles réguliers pour éviter les pertes de données.

— **Surveillance de la sécurité :** Examiner le comportement anormal des utilisateurs et rechercher le comportement malveillant parmi différents utilisateurs.

— **Évaluation des risques :** Évaluer le risque de certaines transactions sur la base de l'expérience passée d'une supervision et d'une analyse continues.

— **Détection de fraude :** Ces méthodes examinent minutieusement les caractéristiques anormales d'un type utilisateur pour détecter la fraude et la cybercriminalité.

— **Surveillance du réseau :** L'observation de la défaillance du réseau suit la force et la faiblesse d'un chemin de réseau en augmentant le filtrage.

Conclusion

L'analyse des données est l'application de méthodes mathématiques pour obtenir des informations à partir de données qui peuvent être utilisées pour améliorer les processus ou appuyer les décisions. L'analyse des données s'est donc principalement concentrée sur des concepts et des modèles statistiques en les transformant en algorithmes afin que les machines puissent apprendre par elles-mêmes ; c'est ce que nous appelons l'apprentissage automatique. L'apprentissage automatique est le dernier d'une longue série d'essais visant à résumer les connaissances humaines et la logique sous une forme adaptée à la construction de machines et à l'ingénierie de systèmes automatisés. L'apprentissage automatique est devenu de plus en plus omniprésent, ainsi l'utilisation de logiciels est devenue plus facile. Donc, nous pouvons maintenant dire que l'apprentissage automatique a eu de nombreux succès, car un logiciel est facilement disponible pour concevoir et construire des systèmes d'apprentissage automatique riches et flexibles. Dans le chapitre suivant, nous verrons comment traiter le Big Data en utilisant les nouvelles technologies utilisées dans ce domaine.

Calcul parallèle et distribué pour l'analyse du Big Data 4

Sommaire

4.1 Introduction

Le traitement de données volumineuses nécessite des ressources informatiques massivement parallèles et largement distribuées en raison de la quantité de données impliquée dans un calcul, ainssi que les résultats sont fournis dans un délai assez court, sinon, ce traitement peut perdre de la valeur avec le temps. Le calcul parallèle et distribué a émergé en tant que solution pour résoudre des problèmes complexes en utilisant d'abord plusieurs équipements de traitement, puis plusieurs nœuds de calcul dans un réseau. Le passage du traitement séquentiel au traitement parallèle et distribué offre des performances et une fiabilité élevées pour les applications. Mais, il introduit également de nouveaux défis en termes d'architectures matérielles, de technologies de communication interprocessus, d'algorithmes et de conception de systèmes. Dans ce chapitre, nous présentons divers systèmes et technologies qui nous permettent et qui nous aident à traiter les données volumineuses de manière plus efficace.

4.2 Calcul parallèle et distribué pour Big Data

Le monde d'aujourd'hui a une vision globale de la façon de gérer les problèmes du Big Data notamment le volume, la variété et la véracité. Mais cela devient très facile avec le développement terrible des machines et des logiciels. Premièrement, la puissance du matériel a augmenté, et son prix a diminué. En conséquence, beaucoup des logiciels sont apparus qui tirent parti de ce matériel en automatisant des processus tels que l'équilibrage de charge et l'optimisation sur un vaste cluster de nœuds. L'un des problèmes liés à la gestion de grandes quantités de données est l'impact de la latence qui touche tous les aspects de l'informatique notamment les communications, la gestion des données, les performances du système, etc. La possibilité de tirer parti des techniques de calcul distribué et de traitement parallèle a permis de réduire le temps de latence. Il peut s'avérer impossible de créer une application Big Data dans un environnement à latence élevée si des performances élevées sont nécessaires. Il est nécessaire de traiter, analyser et vérifier ces données en temps quasi réel. Dans le but de réduire le temps de latence, diverses techniques de calcul parallèle et distribué ont été proposées de temps à autre par des chercheurs et des praticiens. Le calcul parallèle utilise des nœuds de calcul ou des machines modernes qui contient des processeurs courants souvent multicœurs, multithread ou GPU comme des infrastructures matérielles, ou il utilise des plateformes ou des technologies sophistiquées, souvent Hadoop et son écosystème comme des infrastructures logicielles [CE 1998], cela permet aux technologies parallèles d'accroître rapidement la vitesse du processeur et l'efficacité énergétique. En plus, le traitement du Big Data est largement déterminé par les systèmes distribués utilisés pour permettre l'échange de données entre les nœuds de calcul [Pop *et al.* 2017]. Les ressources de traitement de connexion réseau peuvent également être des produits courants tels qu'Ethernet. Cependant, il est souvent utile de concevoir un réseau personnalisé, ou du moins d'utiliser une configuration personnalisée de commutateurs de base qui répondent aux exigences

de communication. Alors, le traitement du Big Data dans le sens parallèle et distribué unifie le calcul parallèle et distribué en même temps ; il permet d'exploiter les technologies du Big Data dans un ensemble de nœuds de calcul en réseau hétérogènes et les présenter comme une ressource unifiée.

4.2.1 Calcul parallèle

Le calcul parallèle est basé sur les systèmes parallèles qui sont composés de processeurs, d'une hiérarchie de mémoires et de réseaux d'interconnexion [Benjamin 2008][Inderpal 2013]. L'évolution de ces composants a été très importante, les systèmes parallèles sont encore améliorés, mais à des vitesses différentes. L'évolution de la technologie se poursuit et rend les processeurs toujours plus puissants. Même si les performances de la mémoire continuent d'augmenter, l'écart entre les processeurs et la mémoire s'accroît. Aujourd'hui, plusieurs niveaux de cache sont nécessaires pour le remplir. Enfin, la connexion entre les unités de calcul se fait par des bus si leur nombre est petit, ou par des réseaux d'interconnexion dans l'autre cas. Donc, dans le calcul parallèle, plusieurs processeurs coopèrent pour résoudre un problème, ce qui réduit le temps de calcul, car plusieurs opérations peuvent être effectuées simultanément. L'utilisation de plusieurs processeurs travaillant ensemble sur un même calcul illustre un nouveau paradigme en matière de résolution de problèmes informatiques, il est complètement différent du traitement séquentiel. Le calcul parallèle introduit des modèles et des architectures permettant d'effectuer plusieurs tâches au sein d'un seul nœud de calcul ou d'un ensemble de nœuds étroitement couplés avec un matériel homogène [Conti 2015]. Le parallélisme est obtenu en exploitant un matériel capable de traiter plusieurs instructions en parallèle. Différentes architectures exploitent le parallélisme pour augmenter les performances d'un système informatique selon que le parallélisme est réalisé sur des données, des instructions ou les deux. Le développement d'applications parallèles nécessite souvent des environnements et des compilateurs spécifiques offrant un accès transparent aux fonctionnalités avancées des architectures sous-jacentes. Du point de vue pratique, cela fournit une justification suffisante pour étudier le concept de traitement en parallèle et des questions connexes tels que les architectures parallèles, les algorithmes parallèles, les langages de programmation parallèles et l'analyse des performances qui sont étroitement liés.

4.2.1.1 Architectures et traitement parallèles

Le domaine de l'architecture des ordinateurs a connu une croissance explosive au cours des deux dernières décennies. Grâce à un flot continu de recherches expérimentales, d'efforts de création d'outils et d'études théoriques, la conception d'une architecture des ordinateurs est considérée autrefois comme un art. L'architecture des ordinateurs a été transformée en l'une des branches les plus quantitatives de la technologie informatique. Dans le même temps, une meilleure compréhension des

différentes formes de concurrence, du pipeline standard au parallélisme massif, et l'invention de structures architecturales pour prendre en charge un modèle de programmation relativement efficace et convivial pour de tels systèmes ont permis aux performances matérielles de poursuivre leur croissance exponentielle [Parhami 2002]. Avec cette croissance explosive, les performances continueront de croître de manière exponentielle avec chaque nouvelle génération de matériel, et que (contrairement au logiciel) le matériel informatique fonctionnera correctement dès qu'il sort de la chaîne de montage. Cela a entraîné une complexité matérielle sans précédent et des coûts de développement presque intolérables. Les défis que doivent relever les concepteurs des architectures des ordinateurs actuels consiste à instaurer la simplicité ; de sorte que les concepteurs utilisent les théories fondamentales développées dans ce domaine pour obtenir des avantages en termes de performances et de facilité d'utilisation grâce à des circuits plus simples, ainsi que de comprendre l'interaction entre les capacités et les limitations technologiques, d'une part, et les décisions de conception en fonction des besoins des utilisateurs et des applications, d'autre part.

Il est maintenant possible de construire de puissants systèmes multiprocesseurs [Atiquzzaman 1993][Baer 1976] et de les utiliser efficacement pour le traitement de données qui a connu une expansion explosive dans de nombreux domaines de l'informatique et de l'ingénierie. Pour répondre aux exigences de performances des applications, l'un des moyens consiste à utiliser le système multiprocesseur le plus puissant disponible à ce jour. Le concept du traitement parallèle diffère du traitement séquentiel. Dans le calcul séquentiel, un processeur est effectue une opération à la fois [Fischer & Plessow 2015]. D'autre part, dans le calcul parallèle, plusieurs processeurs coopèrent pour résoudre un problème, ce qui réduit le temps de calcul, car plusieurs opérations peuvent être effectuées simultanément [Fischer & Plessow 2015][Obiniyi *et al.* 2015]. L'utilisation de plusieurs processeurs travaillant ensemble sur un même calcul illustre un nouveau paradigme en matière de résolution de problèmes informatiques complètement différent du traitement séquentiel. Comme nous l'avons dit précédemment, le traitement parallèle implique l'utilisation de plusieurs facteurs tels que les architectures parallèles, les algorithmes parallèles, les langages de programmation parallèles et l'analyse des performances qui sont étroitement liés. Ces facteurs permettent d'améliorer les performances ou d'autres attributs (par exemple, la rentabilité, la fiabilité) des ordinateurs par le biais de diverses formes de simultanéité. Bien que les calculs aient été simultanés depuis les débuts de des ordinateurs, ils ont récemment été appliqués de manière à améliorer les performances ou la rentabilité par rapport aux supercalculateurs. Comme dans tout autre domaine de la science et de la technologie, l'étude des architectures et des algorithmes parallèles nécessite une motivation et une vue d'ensemble montrant les relations entre les problèmes et les différentes solutions pour les résoudre, ainsi que des modèles de comparaison, de mise en relation et d'évaluation de nouvelles idées.

Le parallélisme a introduit de nouveaux degrés de liberté dans les approches de conception d'architecture et d'algorithme [Schnabel 1985]. Pour une utilisation efficace des systèmes parallèles, il est essentiel d'obtenir une bonne adéquation entre les exigences de l'algorithme et les capacités de l'architecture. En général, l'exécution d'un problème de calcul parallèle comporte quatre étapes. La première étape consiste à fournir une architecture informatique parallèle. La deuxième étape concevoir un algorithme parallèle ou à mettre en parallèle l'algorithme séquentiel existant. La troisième étape consiste à mapper l'algorithme parallèle dans l'architecture informatique parallèle appropriée, et la dernière étape consiste à écrire le programme parallèle en utilisant une approche de programmation parallèle applicable.

4.2.1.1.1 Les architectures parallèles

L'architecture des ordinateurs est l'étude de l'organisation et de l'interconnexion de composants de systèmes informatiques. L'ordinateur peut être construit à partir des blocs de construction de base tels que les mémoires, les unités arithmétiques, les éléments de traitement et les bus [1]. Il est possible de construire à partir de ces blocs de construction n'importe quel type d'ordinateur, du plus petit (micro-ordinateur) au plus grand (super-ordinateur). Le comportement fonctionnel des composants de différents ordinateurs est similaire. Par exemple, le système de mémoire exécute les fonctions de stockage, l'unité de traitement centrale effectue les opérations, et les interfaces d'entrée et de sortie transfèrent les données d'un processeur aux dispositifs appropriés.

Les principales différences entre les ordinateurs est la façon dont les unités sont connectées les unes aux autres. Les caractéristiques de performance de ces unités est la manière dont elles contrôlent le système informatique au cours des opérations. Les deux principaux composants du système informatique traditionnel sont le processeur et la mémoire ; de sorte que, le processeur gère les données stockées en mémoire selon les instructions.

Le transfert de données dans un système est bidirectionnel, ce qui signifie que les données peuvent être lues ou écrites dans les modules de mémoire. La figure 4.1 représente l'interconnexion mémoire-processeur connue sous le nom de modèle de calcul de Von Neumann [Burks & Neumann 1946][Hepsa 2000]. Une autre extension naturelle du modèle de Von Neumann est un réseau d'ordinateurs [Asprey 1989][Asprey 1990], chaque nœud du réseau est un ordinateur autonome qui peut être considérablement complexe, il fonctionne de manière totalement autonome par rapport aux autres nœuds. Un réseau informatique peut être géographiquement distribué.

1. https ://computing.llnl.gov/tutorials/parallel_comp/, (consulté le :11/04/2019)

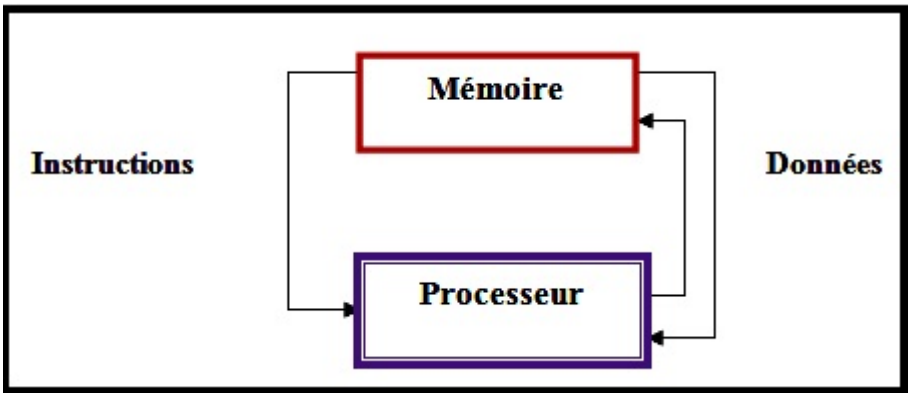

FIGURE 4.1 – *Interconnexion mémoire-processeur.*

En plus de cette extension naturelle du modèle de Von Neumann, il est possible d'adopter une approche plus fondamentale et de concevoir de nouveaux modèles de calcul exclusivement pour le traitement en parallèle. Ces modèles comprennent les processeurs multi-cœurs, les modèles de flux de données et les modèles de flux de contrôle [Najjara *et al.* 1999]. Sachant que, Les ordinateurs parallèles sont divisés en deux catégories principales : de flux de contrôle et de flux de données [Faraz *et al.* 2015][Dijkstra & Broy 1985]. Les ordinateurs parallèles à flux de contrôle reposent essentiellement sur les mêmes principes que l'ordinateur séquentiel ou de Von Neumann sauf que plusieurs instructions peuvent être exécutées à un moment donné [Brownbridge *et al.* 1982]. Les ordinateurs parallèles de flux de données, parfois appelés «non-Von Neumann», sont complètement différents en ce qu'ils n'ont pas de pointeur sur une ou plusieurs instructions actives ni sur un lieu de contrôle [Eassa & Zaki 1995][Panfilova & Salibekyan 2014]. Le contrôle est totalement distribué, la disponibilité d'opérandes déclenchant l'activation d'instructions. Dans ce qui suit, nous nous concentrerons exclusivement sur les ordinateurs parallèles à flux de contrôle.

4.2.1.1.1.1 Classification des architectures de machines

En 1966, M. J. Flynn a classé les architectures de machines selon diverses caractéristiques, notamment le nombre de processeurs, le nombre de programmes qu'elles peuvent exécuter et la structure de la mémoire. Cette classification révèle une bonne méthode pour la taxonomie des architectures de machines. La classification de Flynn est devenue une norme, il est largement utilisée. Flynn a inventé les abréviations SISD *-Single Instruction on Single Data-*, SIMD *-Single Instruction Multiple Data-*, MISD *-Multiple Instructions Single Data-* et MIMD *-Multiple Instructions on Multiple Data-* pour les quatre classes d'ordinateurs illustrées dans la Figure 4.9, en fonction du nombre de flux d'instructions (un ou plusieurs) et flux de données (un ou plusieurs) [Flynn 1966].

La classe SISD représente des machines qui utilisent des «mono-processeurs» dont les instructions ne fonctionnaient que sur des données uniques, c-à-d., ces machines ont une unité centrale qui exécute une instruction à la fois (flux d'instructions unique), cette unité extrait ou stocke un élément de données à la fois (flux de données unique). La figure 4.2 présente une structure générale de l'architecture SISD. Tous les ordinateurs SISD utilisent un seul registre appelé : compteur de programme [2], ce registre impose l'exécution en série des instructions [Flynn 1972][Flynn 1995]. Au fur et à mesure que chaque instruction est extraite de la mémoire, ce registre est mis à jour à l'adresse de l'instruction suivante à extraire et à exécuter, ce qui donne un ordre d'exécution en série, ce qui rend également cette exécution très lente.

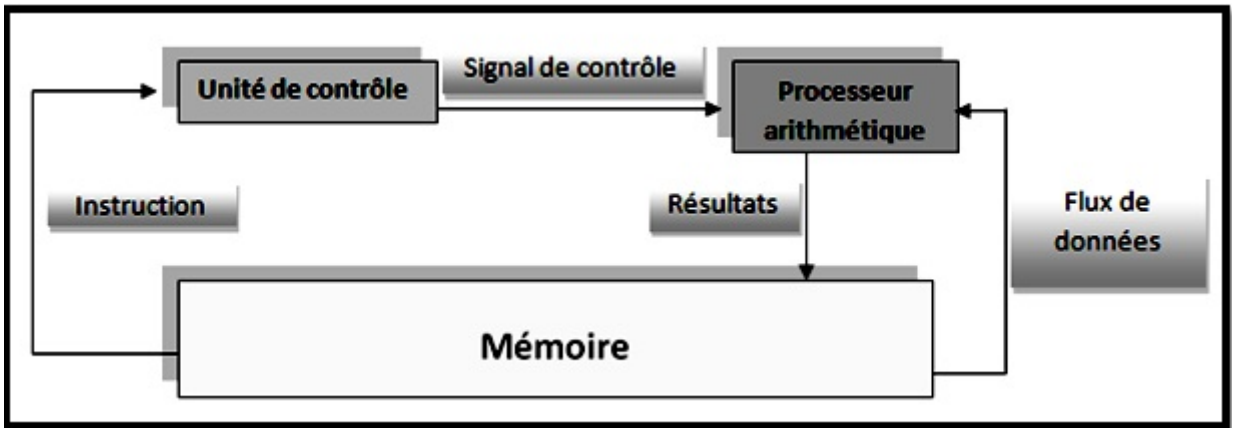

FIGURE 4.2 – *Modèle d'architecture SISD.*

Tandis que, les ordinateurs de la classe SIMD avec plusieurs processeurs dirigés par des instructions provenant d'une unité de contrôle centrale sont parfois appelés «processeurs multidisques». L'unité de contrôle génère les signaux de contrôle pour tous les éléments de traitement, lesquels exécutent la même opération sur différents éléments de données (donc de multiples flux de données), ce qui signifie qu'ils exécutent des programmes dans un mode pas à pas, chaque élément de traitement possède ses propres flux de données [Flynn 1995] [MinYou & Wei 1995]. En d'autres termes, de nombreux éléments de traitement distincts sont appelés par une seule unité de contrôle. La figure 4.3 présente une vue générale d'une architecture SIMD lorsqu'un seul élément de traitement est actif, il peut être une machine de Von Neumann.

2. Dans un processeur, le compteur de programme ou le compteur ordinal ou pointeur d'instruction est le registre (souvent nommé PC) qui contient l'adresse mémoire de l'instruction en cours d'exécution ou prochainement exécutée (cela dépend de l'architecture).

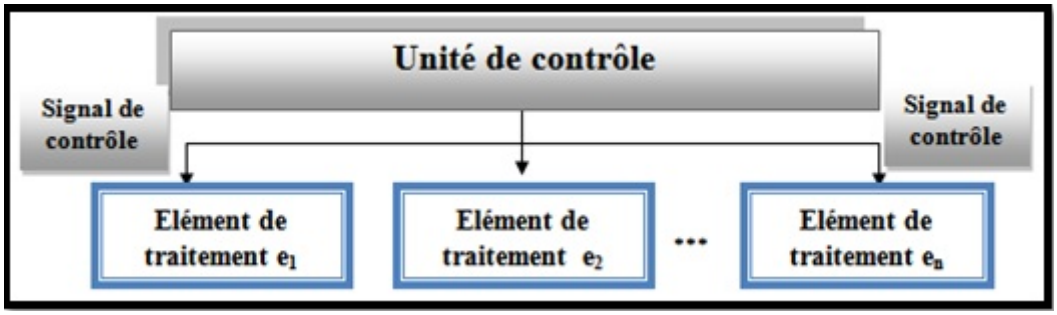

FIGURE 4.3 – *Modèle d'architecture SIMD.*

SIMD est largement utilisé pour le traitement d'image, de vidéo, et le traitement du signal numérique [Choi *et al.* 2016]. Ces ordinateurs sont principalement utilisés pour les problèmes de parallélisme à grain réduit. Parce que, ils effectuent le même calcul sur plusieurs points de données, ce qui entraîne un parallélisme au niveau des données et donc des gains de performance.

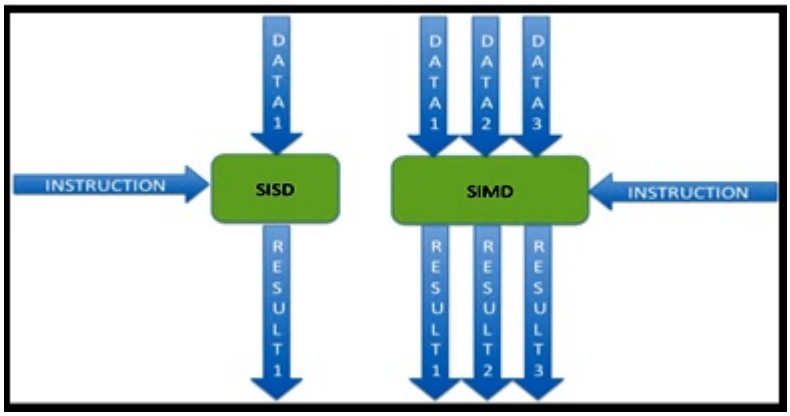

FIGURE 4.4 – *Différence entre SISD et SIMD.*

On peut voir sur la figure 4.4 que SIMD ne fournit pas le parallélisme au niveau instruction, mais seulement le parallélisme au niveau des données. SIMD peut traiter plusieurs vecteurs de données avec une seule instruction ; ceci est très utile pour certaines opérations en boucle. Par exemple, si vous avez deux listes d'octets et que vous souhaitez les ajouter à une liste, en supposant que la longueur des deux listes est de 1024, alors l'opération d'ajout prendra 1024 fois, mais si SIMD est pris en charge par l'ordinateur et le CPU est de 64 bits, cela ne prendra que 128 fois pour terminer le traitement.

Les machines de la catégorie MISD peuvent exécuter plusieurs programmes différents sur le même élément de données[3] [Popov 2017] ; cela implique que plusieurs instructions fonctionnent sur une seule donnée. La figure 4.5 représente la structure générale d'une architecture MISD.

3. https ://www.includehelp.com/cso/parallel-processing.aspx, (consulté le : 20/04/2019)

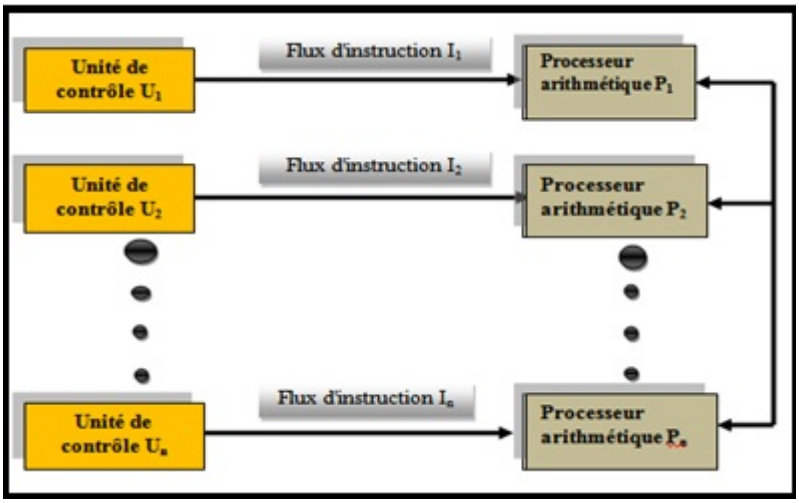

FIGURE 4.5 – *Modèle d'architecture MISD.*

MIMD est un autre type de parallélisme. Par rapport à la machine avec SIMD, les machines utilisant MIMD ont un certain nombre de processeurs qui fonctionnent de manière asynchrone et indépendante, chacun de ces processeurs peut exécuter un programme différent (flux d'instructions multiples) sur son propre élément de données (flux de données multiple). De plus, dans la plupart des systèmes MIMD, chaque processeur a accès à une mémoire globale, ce qui peut réduire les délais de communication. Ainsi que, chaque processeur possède une mémoire privée, ce qui aide à éviter les conflits de mémoire [4] [Azeem *et al.* 2015]. La plupart des architectures MIMD tirent parti du parallélisme à grain moyen et large. Dans les architectures parallèles MIMD actuelles, le nombre de processeurs est inférieur à celui des systèmes SIMD. La figure 4.6 représente la structure générale d'une architecture MIMD :

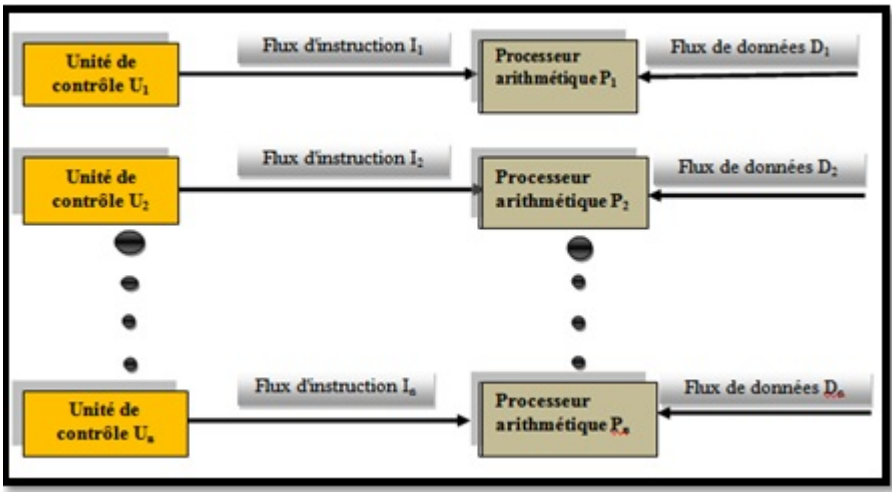

FIGURE 4.6 – *Modèle d'architecture MIMD.*

4. https ://www.techopedia.com/definition/3479/multiple-instruction-multiple-data-mimd, (consulté le : 20/04/2019)

Comme nous avons déjà parlé ci-dessus, un système MIMD est un système multiprocesseur ou multi-ordinateur dans lequel chaque processeur individuel possède sa propre unité de contrôle et exécute son propre programme. Dans un système MIMD, on peut accepter plusieurs instructions en même temps ; chaque instruction est indépendante des autres. Le système MIMD traite son propre flux de données. Il existe deux types de MIMD : MIMD à mémoire partagée (*SM- MIMD :Shared Memory- Multiple Instructions on Multiple Data-*) et MIMD à mémoire distribuée (*DM- MIMD :Distributed Memory- Multiple Instructions on Multiple Data-*) [Dongarra & van der Steen 2012]. Dans le modèle SM-MIMD, tous les processeurs partagent une mémoire centrale commune. La particularité des systèmes à mémoire partagée est que le nombre de blocs de mémoire utilisés n'a pas d'importance, mais nous sommes préoccupés par la manière dont ces blocs sont connectés aux processeurs, les espaces d'adresse de ces blocs de mémoire sont unifiés dans un espace d'adresse global parfaitement visible par tous les processeurs du système à mémoire partagée [Rajaraman & Murthy 2016] [Elnour *et al.* 2014].

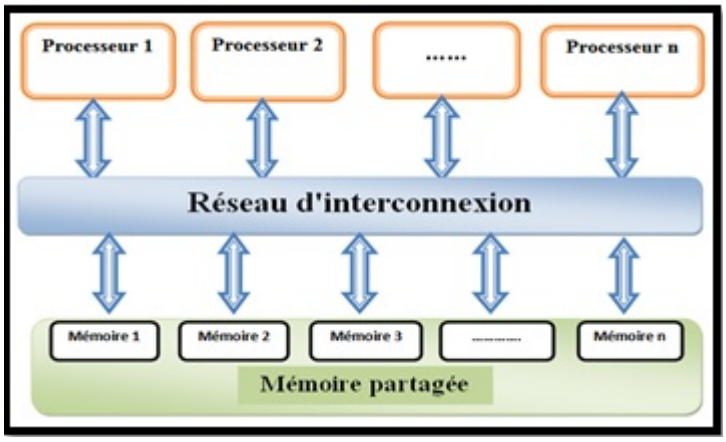

FIGURE 4.7 – *Mémoire partagée MIMD (SM- MIMD).*

La figure 4.7 représente SM-MIMD montrant les processeurs et les mémoires connectés par un réseau d'interconnexion, l'un des avantages du modèle de mémoire partagée est qu'il est facile à comprendre, un autre avantage est que la cohérence de la mémoire est gérée par le système d'exploitation. Il est donc facile pour le développeur de concevoir un programme parallèle dans un tel modèle. L'inconvénient est qu'il est difficile d'évoluer avec le modèle à mémoire partagée, et qu'il n'est pas aussi flexible que le modèle à mémoire distribuée (voir la figure 4.8 ci-dessous). Le modèle à mémoire partagée est couramment utilisé les GPUs.

Le marché des jeux vidéo est tellement lucratif que l'industrie a mis au point des cartes graphiques de plus en plus rapides afin de traiter des jeux vidéo de plus en plus détaillés. Il s'agit en réalité de périphériques de traitement parallèle. Au début du 21ème siècle, certains se sont demandé s'il serait possible de les utiliser pour le traitement parallèle d'applications non graphiques, cette programmation s'appelait GPGPU- *General-purpose processing on graphics processing units-*. Les GPGPUs sont

des co-processeurs fortement optimisés pour le traitement graphique, plus tard abrégée en GPU [Ohno *et al.* 2011].

A la fin des années nonante, NVIDIA présentait le premier GPU disponible sur le marché pour un ordinateur de bureau appelé GeForce 256. Il pouvait traiter 10 millions de polygones par seconde, ce qui lui permettait de décharger une quantité importante de traitement graphique de la CPU. Des technologies telles que CUDA permet désormais aux utilisateurs d'accéder directement à leurs calculs en utilisant du matériel graphique puissant et moderne. CUDA est une architecture logicielle développée par NVIDIA qui permet de programmer des GPU à l'aide de langages de programmation de haut niveau tels que C et C ++ [Anders & KAndrot 2011].

Comme nous avons indiqué précédemment, le matériel GPU et la programmation CUDA sont orientés sur les threads et la mémoire partagée. Jusqu'ici, tout va bien, mais il y a des problèmes dans certains des concepts. Nous commençons donc à nous préparer à un barrage de termes spécialisés : L'exécution sur le GPU (appelé le périphérique) est lancée sur le CPU (l'hôte). Le GPU est une unité de traitement graphique polyvalente capable d'implémenter des algorithmes parallèles [5] [Pratxa & Xing 2011]. Il est connecté à l'ordinateur hôte pour exécuter la partie d'une application qui nécessite beaucoup de temps de calculs et jeux de données volumineux. Ce périphérique est responsable de l'exécution de la partie parallèle de l'application. Tandis que, l'hôte est l'ordinateur qui s'interface avec l'utilisateur, il contrôle le périphérique utilisé pour exécuter la partie de l'application qui nécessite une grande quantité de données, mais il nécessite également des calculs parallèles. L'hôte est donc responsable de l'exécution de la partie série de l'application [6].

Nous voyons que le terme mémoire partagée dans ce contexte fait maintenant référence à la mémoire de la carte graphique et non à la mémoire utilisée par la CPU ; cette mémoire est appelée mémoire globale. Pour rendre les choses encore plus déroutantes, il existe en fait quelque chose appelé mémoire partagée, que nous verrons vraiment équivaloir à un cache. Lors du lancement, le programmeur configure également des structures spéciales, une grille et des blocs qui déterminent la manière dont les threads sont organisés.

En outre, la mémoire distribuée *-DM :Distributed Memory-* est un autre type de MIMD. Dans ce modèle, chaque processeur a son propre emplacement de mémoire. Chaque processeur n'a aucune connaissance directe de la mémoire des autres processeurs. Pour que les données soient partagées, elles doivent être transmises d'un processeur à un autre en tant que message [Elnour *et al.* 2014] [McClelland & Rumelhart 1985]. Puisqu'il n'y a pas de mémoire partagée, le conflit n'est pas un problème grave avec ces machines. DM-MIMD est la partie qui connaît la croissance

5. https ://www.scienceabc.com/innovation/what-is-a-gpu-how-exactly-does-it-help-in-running-high-graphic-games.html, (consulté le : 20/04/2019)

6. https ://www.scienceabc.com/innovation/what-is-a-gpu-how-exactly-does-it-help-in-running-high-graphic-games.html, (consulté le : 20/04/2019)

la plus rapide dans la famille des ordinateurs ou des serveurs hautes performances, car elle peut considérablement améliorer la bande passante en ajoutant plus de processeurs et de mémoires. La figure 4.8 ci-dessous illustre la structure de DM-MIMD.

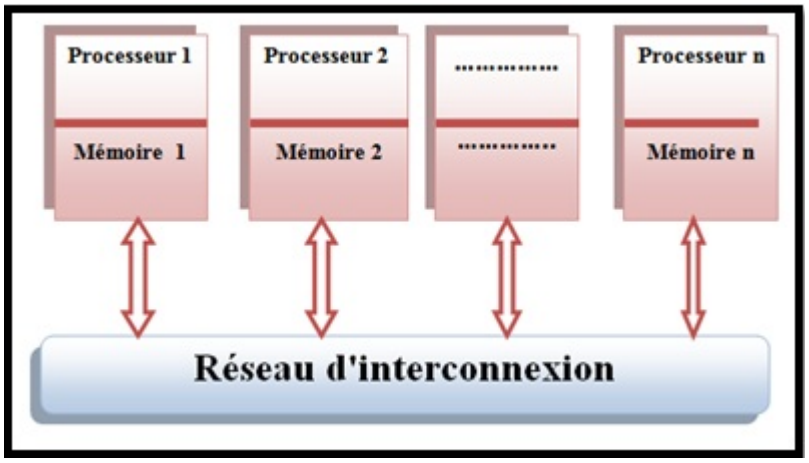

FIGURE 4.8 – *Mémoire distribuée MIMD (DM- MIMD).*

L'inconvénient de DM-MIMD est que les coûts de communication entre différents processeurs peuvent être très élevés et qu'il est difficile d'accéder aux données non locales qui se trouvent dans les mémoires d'autres processeurs. De nos jours, de nombreux systèmes ont été conçus pour réduire le temps et les difficultés entre les processeurs, tels que Hypercube et Mesh. MPP *-Massively Parallel Processors-* est l'un des exemples typiques de DM-MIMD et de nombreuses technologies Big Data réputées reposent sur MPP, comme BIG SQL (SQL sur Hadoop) d'IBM et Impala de Cloudera. En résumé, le MIMD est une tendance du développement actuel de l'architecture informatique, et la plupart des systèmes informatiques distribués sont basés sur de telles technologies.

En 1988, E. Johnson a proposé une classification supplémentaire de ces machines en fonction de leur structure de mémoire (globale ou distribuée) et du mécanisme utilisé pour la communication / synchronisation (variables partagées ou transmission de messages) [Johnson 1988]. Là encore, l'une des quatre catégories GMMP *-Global Memory Message Passing-* n'est pas largement utilisée. Les architectures GMMP ont des espaces d'adresse de processus isolés. Ils sont généralement virtuels dans la même mémoire [Johnson 1989][Johnson 1991]. L'architecture GMSV *-Global Memory Shared Variables-* est ce qu'on appelle vaguement un multiprocesseur à mémoire partagée. Elle est généralement adaptée au serveur d'applications [Hahn *et al.* 1997]. À l'opposé, l'architecture DMMP *-Distributed Memory Message Passing-* s'appelle multi-ordinateurs à mémoire distribuée, dans cette architecture tous les processus sont créés lorsqu'un programme est lancé et se terminent à leur fin, de sorte que chaque processus opère dans un espace d'adressage séparé, et les données sont distribuées aux processeurs. Ainsi que tous les processus exécutent le même programme, mais ils fonctionnent sur des ensembles de

données distincts. L'architecture DMMP est moins coûteuse, elle prend en charge de nombreuses applications commerciales avec un minimum de reprises [Anderson *et al.* 1995]. Enfin, l'architecture DMSV -*Distributed Memory Shared Variables*- gagne en popularité pour combiner la facilité d'implémentation de la mémoire distribuée avec la facilité de programmation du schéma à variable partagée. Elle est parfois appelée mémoire partagée distribuée. L'architecture DMSV présente encore des problèmes techniques pour surmonter le problème de la latence longue [Johnson 1988]. Lorsque tous les processeurs d'une machine de type MIMD exécutent le même programme, le résultat est parfois appelé données multiples à programme unique.

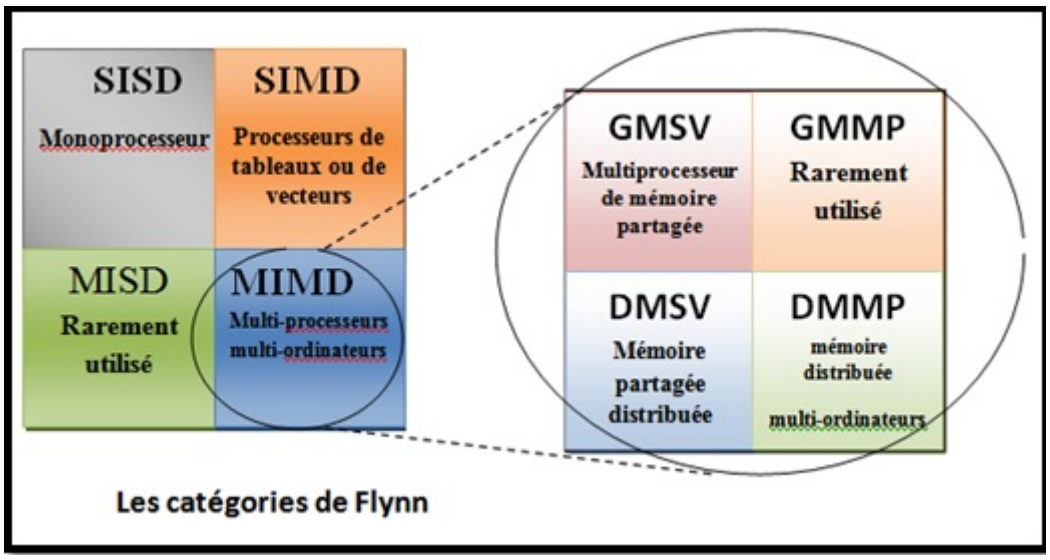

FIGURE 4.9 – *La classification des architectures parallèles selon Flynn-Johnson.*

Les concepteurs informatiques jouent un rôle clé dans la quête de la convivialité, de la compacité, de la simplicité, des performances élevées, du faible coût et de la consommation réduite. Autrefois, lorsque nous écrivons un code, nous n'avons pas besoin de connaître les détails sur caractéristiques du système informatique que ce soit software ou hardware, car le compilateur se chargera de ces détails. Cependant, nous pensons à une seule unité centrale (CPU) et à un traitement séquentiel lorsque nous commençons à écrire le code ou à déboguer la sortie. Maintenant, les processus de mise en œuvre d'algorithmes dans le matériel ou dans le logiciel pour des machines parallèles sont plus liés que nous pourrions le penser, parce que les monoprocesseurs hauts performances deviennent de plus en plus complexes, coûteux et gourmands en énergie. Il existe donc un compromis de base entre l'utilisation d'un ou de quelques-uns de ces processeurs complexes, à un extrême, et d'un nombre modéré à très grand de processeurs simples, à l'autre. Mais les moyens de communication entre le processeur sont logiquement simples, cette approche simplifie grandement le processus de conception. Cependant, deux obstacles majeurs ont jusqu'ici empêché l'adoption généralisée d'architectures modérément pa-

rallèles à massivement parallèles : le premier est le goulot d'étranglement [7] de la communication entre processeurs, le deuxième est le coût élevé du développement de la programmation parallèle (algorithmes et langages machine).

4.2.1.1.1.2 Programmation parallèle

Le concept de parallélisme est souvent utilisé pour désigner plusieurs événements se produisant côte à côte dans l'espace et dans le temps. En informatique, on l'utilise pour désigner le calcul simultané d'opérations sur plusieurs unités de traitement. Ainsi, la programmation parallèle est une spécification d'opérations dans un calcul qui peut être exécutée en parallèle sur différentes unités de traitement [Geist *et al.* 1994][Pacheco 2018]. En outre, les processus de mise en œuvre d'algorithmes dans une architecture informatique parallèle ou dans langage de programmation parallèle sont plus liés que nous pourrions le penser. Notamment, avec l'avènement des processeurs multi-cœurs, la programmation parallèle de la mémoire partagée est devenue courante. Cependant, la programmation parallèle est généralement difficile. Une solution de ce problème consiste à utiliser des bibliothèques qui masquent le parallélisme avec le programmeur. Cette pratique est pratiquée depuis longtemps dans le domaine de la technologie numérique en utilisant des bibliothèques très riches.Pour un domaine de programmation plus général, les modèles de conception sont considérés comme un bon moyen de permettre aux programmeurs de faire face aux difficultés de la programmation parallèle.

4.2.1.1.1.2.1 Les algorithmes parallèles

Un algorithme parallèle est un algorithme dans lequel les tâches sont exécutées simultanément sur de nombreuses unités de traitement différentes en raison de l'indépendance des données, ces tâches sont ensuite combinées pour obtenir le résultat final [8]. Les algorithmes parallèles sont très utiles pour traiter rapidement de gros volumes de données. Un exemple simple d'un tel algorithme purement parallèle est le serveur Web où chaque demande entrante peut être traitée indépendamment des autres demandes. Un autre exemple simple d'algorithmes parallèles est le multitâche dans les systèmes d'exploitation qui gère plusieurs applications, telles qu'un navigateur Web, un traitement de texte, etc. Les problèmes des algorithmes parallèles qui ne se posent pas dans les algorithmes séquentiels incluent la détermination du nombre de processeurs nécessaires pour calculer et allouer des données dans des mémoires. Un algorithme

7. Un goulot d'étranglement est un point d'un système limitant les performances globales d'un flux de production d'une entreprise.

8. https ://www.tutorialspoint.com/parallel_algorithm/parallel_algorithm_introduction.htm, (consulté le : 24/04/2019)

séquentiel est essentiellement une recette ou une séquence d'étapes de base permettant de résoudre un problème donné à l'aide d'un ordinateur série. De même, un algorithme parallèle est une recette qui nous dit comment résoudre un problème donné en utilisant plusieurs processeurs. De plus, un algorithme séquentiel est généralement déterminé par sa complexité temporelle et spatiale [9]. La complexité temporelle d'un algorithme renvoie à son temps d'exécution en fonction de la taille du problème. De même, la complexité de l'espace fait référence à la quantité de mémoire requise par l'algorithme en fonction de la taille du problème. La complexité temporelle est connue pour être la mesure la plus importante de la performance des algorithmes. Un algorithme dont la complexité temporelle est limitée par un polynôme est appelé algorithme pseudo-polynomial [10] [Fellows 1994]. Un algorithme est considéré comme efficace s'il s'exécute en temps polynomial. Les algorithmes inefficaces sont ceux qui nécessitent une recherche de tout l'espace énuméré et qui ont une complexité temporelle exponentielle. Tandis que les algorithmes parallèles, la complexité temporelle reste une mesure importante de la performance. De plus, le nombre de processeurs joue un rôle majeur dans la détermination de la complexité d'un algorithme parallèle. En général, nous disons que la performance d'un algorithme parallèle est exprimée en termes de rapidité et de nombre de ressources utilisées lors de son exécution. Ces critères peuvent être mesurés quantitativement comme suit :

— Le temps d'exécution est défini comme le temps passé lors de l'exécution de l'algorithme.
— Nombre de processeurs utilisés par l'algorithme pour résoudre un problème.
— Le coût de l'algorithme parallèle est le produit du temps d'exécution et du nombre de processeurs.

Le développement d'algorithmes est un élément essentiel de la résolution de problèmes à l'aide d'ordinateurs. Au moins, un algorithme parallèle a une dimension supplémentaire par rapport à la concurrence, le concepteur de l'algorithme doit identifier des ensembles d'étapes pouvant être exécutés simultanément. Ceci est essentiel pour obtenir des performances optimales grâce à l'utilisation d'un ordinateur parallèle.

Les algorithmes parallèles sont implémentés sur des structures parallèles, la relation entre l'espace algorithme et l'espace architecture doit être prise en compte. Il existe trois approches traitant de la conception

9. complexité temporelle : (ou en temps) : temps de calcul ; complexité spatiale : (ou en espace) : l'espace mémoire requis par le calcul.
- La complexité temporelle d'un algorithme est le nombre d'opérations élémentaires (affectations, comparaisons, opérations arithmétiques) effectuées par un algorithme. Ce nombre s'exprime en fonction de la taille n des données.
- La complexité spatiale est une mesure de l'espace utilisé par un algorithme, exprimé comme fonction de la taille de l'entrée. L'espace compte le nombre maximum de cases mémoire utilisées simultanément pendant un calcul.

10. https ://fr.wikipedia.org/wiki/Temps_de_calcul_pseudo-polynomial, (chapitre IV), (consulté le : 24/04/2019)

d'algorithmes parallèles. La première approche, nous mettons en parallèle les algorithmes séquentiels existants ou nous modifions un algorithme séquentiel existant en exploitant les parties de l'algorithme qui sont naturellement parallélisables. La deuxième approche, nous concevons un tout nouvel algorithme parallèle qui pourrait être adapté aux architectures parallèles. Et la dernière approche, nous concevons un nouvel algorithme parallèle à partir de l'algorithme parallèle existant. Cela indique que, d'un point de vue théorique, les algorithmes peuvent être transformés d'une forme séquentielle à une forme parallèle ou d'une forme parallèle à une autre. Une transformation d'algorithme n'est valide que si elle maintient l'équivalence d'algorithme. Ces trois approches doivent prendre en compte la structure de la machine parallèle. Par exemple, en concevant des algorithmes MIMD pour qu'ils s'exécutent sur un ordinateur parallèle, nous essayons de maximiser la quantité de simultanéité. Nous recherchons des opérations pouvant être effectuées simultanément et nous essayons de créer plusieurs processus pour gérer ces opérations, Ceci est possible dans les algorithmes MIMD, mais il contraste avec les algorithmes SIMD.

Dans ce qui suit, nous présentons les quatre étapes de la conception d'un programme parallèle, notamment : le partitionnement, la communication, les agglomérations et le mappage, Comme le montre la figure 4.10 ci-dessous :

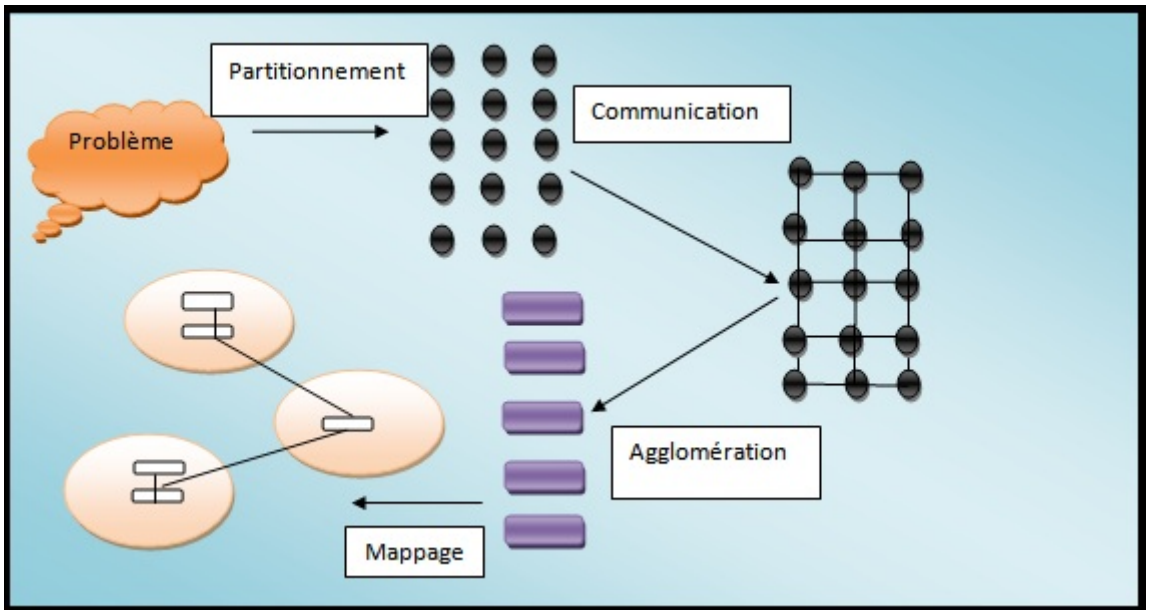

FIGURE 4.10 – *Différentes phases dans la conception d'un algorithme parallèle.*

Les deux premières phases (partitionnement et communication) du processus de conception peuvent permettre de diviser les calculs en un ensemble de tâches et de communications afin de fournir les données requises par ces tâches.

— **Partitionnement :** L'une des premières étapes de la conception d'un programme parallèle consiste à diviser le problème en "morceaux" distincts qui pouvant être répartis en plusieurs tâches, ceci

est appelé : partitionnement. Précisément, la phase de partitionnement est destinée à exposer les possibilités d'exécution parallèle en décomposant les calculs en petites tâches. Des problèmes tels que le nombre de processeurs dans l'architecture parallèle cible sont ignorés et l'attention est concentrée sur la définition d'un grand nombre de petites tâches afin d'obtenir une décomposition fine du problème [11] [BASU 2016]. En général, une bonne méthode de partitionnement divise à la fois le calcul et les données associées à un problème. Les méthodes de partitionnement reposent sur deux approches :

1. **La décomposition de domaine :** Elle concentre d'abord sur les données associées à un problème, puis on détermine un calcul approprié avec ces données (c-à-d., on décompose les données associées à un problème).
2. **La décomposition fonctionnelle :** Elle concentre d'abord sur la décomposition du calcul, puis on décompose les données associées (c-à-d., on divise le calcul en tâches disjointes).

Ces techniques peuvent être appliquées à différentes composantes d'un même problème pour obtenir d'autres algorithmes parallèles. Par conséquent, lors de la première phase d'un processus de conception, la réplication des calculs et des données est évitée, puis nous définissons des tâches qui partitionnent à la fois les calculs et les données en ensembles disjoints.

— **Communication :** La phase de communication est destinée à établir une structure de communication et des algorithmes appropriés pour coordonner l'exécution des tâches. Pour permettre le calcul, les données doivent être transférées entre les tâches, ce qui est le résultat de la phase de communication d'une conception [12] [BASU 2016]. Pour établir l'approche associée, nous nous concentrons sur deux points ; le premier point est la structure de canal, et le deuxième point est la structure de transmission de message. Dans la structure de canal, nous lions (directement ou indirectement) les tâches nécessitant des données à des tâches possédant ces données. Dans la structure de transmission de message, nous spécifions le message qui doit être envoyé et reçu sur ces canaux. En général, nous cherchons à optimiser les performances en répartissant les opérations de communication sur de nombreuses tâches et en organisant les opérations de communication de manière à permettre une exécution simultanée. Néanmoins, penser en termes de structures de canaux pourrait nous aider à évaluer les algorithmes du point de vue des coûts de communication.

11. https ://er.yuvayana.org/parallel-algorithm-design-approach-pcam/, (consulté le : 24/04/2019)
12. https ://er.yuvayana.org/parallel-algorithm-design-approach-pcam/, (consulté le : 25/04/2019)

— **Agglomération :** La phase d'agglomération est le processus consistant à regrouper des tâches en tâches plus grandes pour améliorer les performances. Elle est destinée à évaluer les tâches et la communication requise en matière de performance et de coûts de mise en œuvre [13] [BASU 2016]. Par exemple, il peut être inefficace de créer beaucoup plus de tâches qu'il n'y a de processeurs sur l'architecture parallèle cible. Le regroupement des tâches qui communiquent les unes avec les autres élimine le besoin de communication, ceci appelée augmentation du lieu. Il peut également nous permettre de combiner plusieurs communications en une seule. Ainsi, nous devons revoir les phases de partitionnement et de communication pour répondre aux problématiques de la phase d'agglomération ci-dessous. Alors comment :

  — Combiner ou agglomérer les tâches identifiées ?
  — Répliquer les données et les calculs ?
  — Maintenir la flexibilité en matière d'évolutivité ?
  — Réduire les coûts d'ingénierie logicielle ?
  — Définir le mappage des tâches sur les processeurs ?

  En général, dans la phase d'agglomération nous passons de l'abstrait au concret ; de sorte que, nous revisitons la deuxième phase du problème de processus de conception (communication) pour prendre les décisions appropriées en ce qui concerne les problèmes en jeu. Par exemple, le coût de la communication est un problème critique qui influence les performances parallèles. Les performances peuvent être améliorées en réduisant le temps passé à communiquer, ce qui signifie que nous envoyons moins de données. En plus des coûts de communication, nous pourrions être concernés par les coûts de création de tâches afin de réduire les exigences de communication ou le temps d'exécution.

— **Mappage :** Le mappage est le processus d'attribution de tâches agglomérées aux processeurs. Il est destinée à spécifier où chaque tâche doit être exécutée. Ici, nous pensons à une machine à mémoire distribuée, si nous choisissons le nombre de tâches agglomérées égal au nombre de processeurs, alors le mappage est déjà effectué, de sorte que chaque processeur reçoit une tâche agglomérée. En d'autres termes, chaque tâche est assignée à un processeur de manière à maximiser l'utilisation du processeur et à minimiser les coûts de communication. Le mappage peut être effectué de manière statique ou dynamique, c-à-d., avant le processus d'exécution ou à la phase d'exécution [14] [BASU 2016]. Le mappage ne concerne pas les monoprocesseurs ni les ordinateurs parallèles à mémoire partagée qui fournissent une planification des tâches, de sorte que le système d'exploitation ou des mécanismes matériels peuvent planifier des tâches exécutables sur des processeurs disponibles.

13. https ://er.yuvayana.org/parallel-algorithm-design-approach-pcam/, (24/04/2019)
14. https ://er.yuvayana.org/parallel-algorithm-design-approach-pcam/, (25/04/2019)

En général, il y a deux points importants impliqués dans la phase de mappage, le premier point est l'amélioration de la simultanéité, et le deuxième point est la localité croissante. L'amélioration de la simultanéité consiste à placer des tâches pouvant être exécutées simultanément sur différents processeurs ; tandis que, la localité croissante signifie placer des tâches susceptibles de communiquer fréquemment sur le même processeur. De manière générale, le processus de mappage est donc un problème difficile dans la conception d'algorithmes parallèles, il est connu sous le nom de problème NP-Complet [15], ce qui signifie qu'il n'existe pas d'algorithme de calcul général pour l'évaluation de cette méthode de mappage. Par conséquent, le résultat de ce processus de conception est un programme qui crée des tâches. Il effectue également le mappage des tâches sur des processeurs individuels. L'idée centrale dans le processus de conception est de cibler une conception avec des attributs généraux comme suit :

— **Parallélisme des données :** Le parallélisme des données utilise les données d'entrée pour certaines opérations comme moyen de partitionner en éléments plus petits. Les données sont réparties entre les processeurs disponibles afin de réaliser le parallélisme. Cette étape de partitionnement est souvent suivie de la réplication et de l'exécution d'opérations de programme essentiellement indépendantes sur ces partitions. En règle générale, la même opération est appliquée simultanément aux éléments de l'ensemble de données. Le parallélisme des données sera souvent soumise à la mise en œuvre de SIMD [Hillis & Steele 1986].

— **Parallélisme fonctionnel :** Il s'agit de décomposer l'algorithme en segments pouvant être affectés à différents processeurs. Ainsi, ce type de parallélisme repose sur différents blocs fonctionnels de notre application. L'idée est simple : l'application est divisée en unités de traitement distinctes qui communiquent avec un nombre fixe d'autres unités, de sorte que la sortie d'une partie serve d'entrée à une autre partie. Ainsi, nous pouvons visualiser un tel système comme un ensemble de nœuds reliés par des canaux dans lesquels les données ne circulent que dans une direction. Le parallélisme fonctionnel impliquera toujours l'exécution de MIMD [Foster 1995][Kumar *et al.* 1994][Kowalik 1995].

15. NP-complet (c-à-d un problème complet pour la classe NP) est un problème de décision vérifiant les propriétés suivantes :
-Il est possible de vérifier une solution efficacement (en temps polynomial) ; la classe des problèmes vérifiant cette propriété est notée NP ;
-Tous les problèmes de la classe NP se ramènent à celui-ci via une réduction polynomiale ; cela signifie que le problème est au moins aussi difficile que tous les autres problèmes de la classe NP(La classe NP est une classe très importante de la théorie de la complexité, l'abréviation NP signifie « Non déterministe Polynomial »).

— **Granularité de module :** La granularité de module peut être définie comme la taille moyenne d'une unité de calcul séquentiel dans le programme, donc elle s'agit d'une mesure du problème de synchronisation qui affectera : le choix des architectures SIMD par rapport à MIMD, l'attribution des processus aux processeurs, l'organisation de la mémoire, les exigences de communication et le temps d'exécution de l'algorithme. Les algorithmes caractérisés par une granulométrie fine [16] nécessiteront une synchronisation fréquente, ils conviennent souvent pour une exécution SIMD. Les algorithmes avec une granularité de gros élément composant la collection ont généralement moins besoin de communications efficaces, ils suggèrent donc souvent des opérations MIMD [Marshall 1980][Maheshwari 1996].

— **Granularité des données :** Ceci quantifie la taille des données à traiter. Granularité des données fournit une indication de la mesure nécessaire pour communiquer un seul élément de données. La granularité prend en compte le temps système de communication entre plusieurs processeurs ou éléments de traitement. Elle est généralement mesurée en termes de nombre d'instructions exécutées dans une tâche donnée. Alternativement, la granularité peut également être spécifiée en termes de temps d'exécution d'un programme combinant le temps de calcul et le temps de communication. Certains des problèmes liés à la granularité des données sont l'allocation de données, les exigences de communication, la capacité du processeur et les exigences de mémoire [17].

— **Degré de parallélisme :** Il est lié à la fois à la granularité des données et à celle des modules, il affecte le choix de la taille de la machine, il peut également atteindre l'accélération maximale. Degré de parallélisme est considéré comme une métrique qui indique le nombre d'opérations pouvant être simultanément par un ordinateur. Il est particulièrement utile pour décrire les performances de programmes parallèles et de systèmes multiprocesseurs. De plus, il est souvent lié au mode de fonctionnement et à l'organisation de la mémoire [18][Changtian *et al.* 2018].

— **Uniformité des opérations :** L'uniformité sera généralement associée au parallélisme des données. En général, Si les opérations à effectuer sont normalisées, alors le système ou le pipeline HMIS *-Hazardous Materials Information System-* peut être possible ; sinon, le traitement MIMD sera choisi. Il est possible de construire différents niveaux de normalisation en fonction de la granularité ou de la résolution à laquelle les opérations sont examinées [Kent & Williams 1992][Lin & Lee 1991].

16. La granulométrique est l'ensemble des opérations permettant de déterminer la distribution statistique des tailles des éléments composant la collection.
17. https ://computing.llnl.gov/tutorials/parallel_comp/, (consulté le : 25/04/2019)
18. https ://en.wikipedia.org/wiki/Degree_of_parallelism,(consulté le : 25/04/2019)

— **Synchronisation :** Le besoin de synchronisation apparaît chaque fois qu'il y a des processus simultanés dans un système, cela affectera l'affectation des processus aux processeurs. Ainsi que, elle affectera la planification de divers composants de l'algorithme. De plus, les contraintes de priorité sont des problèmes importants pour caractériser les exigences de synchronisation [Belzer *et al.* 1997].

— **Dépendance des données :** La dépendance des données entre les instructions est la capacité du compilateur à optimiser le code pour le parallélisme. Donc, pour tirer le meilleur parti du code parallèle, la dépendance des données doit être gérée avec soin. Elle est également permise de spécifier les modèles d'allocation de données, les caractéristiques de communication et l'organisation de la mémoire locale. Elle joue le rôle le plus important en dictant les modèles d'allocation de données et les caractéristiques de communication [Banerjee & Wolfe 1987] [Armstrong *et al.* 2014] [Ketterlin & Clauss 2012].

— **Génération et fin de processus :** Ils ont une incidence sur l'utilisation du processeur, la planification des sous-processus, le mode de traitement, l'organisation de la mémoire et les exigences de communication.

Des travaux récents sur les algorithmes parallèles se sont concentrés sur la résolution de problèmes relevant de domaines tels que l'appariement de modèles, les structures de données, le tri, la géométrie informatique, l'optimisation combinatoire, l'algèbre linéaire et la programmation linéaire. Les algorithmes sont également conçus spécifiquement pour les types d'ordinateurs parallèles disponibles aujourd'hui. Une attention particulière a été accordée aux machines avec une bande passante de communication limitée. L'industrie des machines parallèles a traversé une période de turbulences financières, plusieurs constructeurs ayant échoué ou interrompu la vente de machines parallèles. Cependant, ces dernières années un grand nombre de machines parallèles peu coûteuses ont été vendues. Ces machines sont généralement composées de 4 à 8 processeurs reliés par un bus à un système de mémoire partagée. Alors que, ces machines atteignent la taille imposée par l'architecture de bus, les fabricants ont réintroduit des machines parallèles basées sur la topologie maillée-2D notamment la topologie hypercube comme le montre la figure 4.11 ci-dessous.

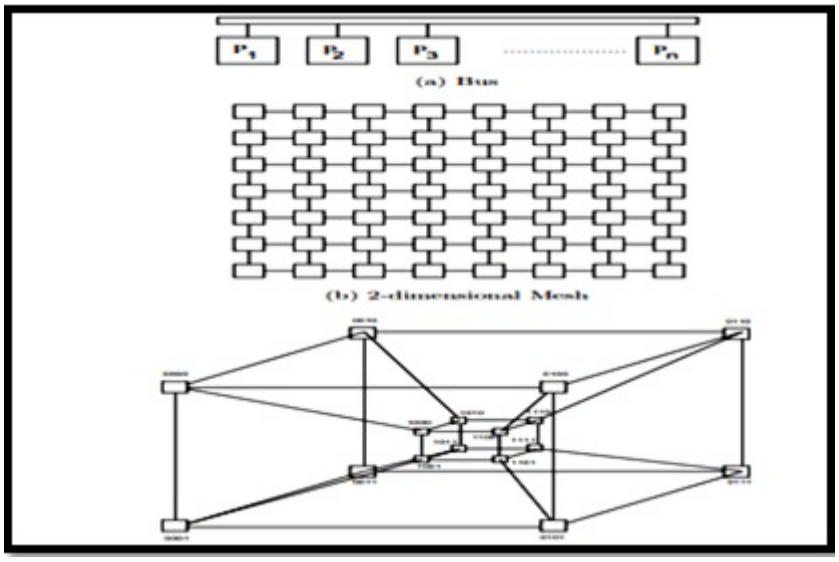

FIGURE 4.11 – *Les machines parallèles basées sur la topologie en bus, maillés et hypercube.*

Bien que ces topologies puissent conduire à des algorithmes parallèles très précis, elles présentent certains inconvénients. Premièrement, les algorithmes conçus pour un réseau ne peuvent pas fonctionner correctement sur d'autres réseaux. Par conséquent, pour résoudre un problème sur une nouvelle machine ; il peut être nécessaire de refaire une nouvelle conception de l'algorithme. Deuxièmement, les algorithmes qui exploitent un réseau particulier ont tendance à être plus compliqués que les algorithmes conçus pour des modèles plus abstraits comme les modèles PRAM *-Parallel Random Access Machine-*, car ils doivent intégrer certains détails du réseau. Néanmoins, certaines opérations qui sont souvent effectuées par une machine parallèle sont utiles pour concevoir un algorithme spécifique au réseau. Par exemple, l'algorithme qui achemine les messages ou les demandes d'accès à la mémoire via le réseau doit exploiter la topologie du réseau. D'autres exemples incluent des algorithmes pour la diffusion d'un message d'un processeur à de nombreux autres processeurs afin de collecter des résultats calculés dans plusieurs processeurs.

4.2.1.1.1.2.2 Les langages de programmation parallèles

La préparation des programmes pour une exécution parallèle revêt une immense importance pratique. Nous avons vu que la conception de structures parallèles et d'algorithmes parallèles sont pleines de problèmes complexes. En raison de la difficulté de ces problèmes, la principale raison de l'absence d'acceptation et de l'application limitée du traitement en parallèle n'est pas les équipements informatiques ou les algorithmes. mais plutôt les langages de programmation obscurs et fastidieux, nous mentionnons particulièrement les langages de programmation parallèles, car ils contiennent des modèles de données plus complexes que les langages de programmation séquentiels. Le modèle de données en langages

séquentielles est celui du modèle de machine à accès aléatoires (RAM) où il y a un emplacement de mémoire unique pouvant être lu et écrit par le processeur. L'analogique en langage parallèle est le modèle de mémoire partagée dans lequel tous les emplacements de mémoire sont situés dans un seul espace d'adressage, et tous les processeurs peuvent y accéder. Un modèle de données plus découplé est fourni par le modèle de mémoire distribuée, de sorte que chaque processeur a son propre espace d'adressage en mémoire auquel les autres processeurs ne peuvent pas accéder. Le choix du modèle de données détermine la manière dont les processeurs communiquent entre eux dans un modèle à mémoire partagée, c.à.d, ces processeurs communiquent en lisant et en écrivant des emplacements partagés, mais dans un modèle à mémoire distribuée, ils communiquent en envoyant et en recevant des messages.

Les langages de programmation parallèles sont des langages conçus pour programmer des algorithmes et des applications sur des ordinateurs parallèles. Le traitement parallèle est une excellente opportunité pour développer des systèmes hautes performances et pour résoudre de gros problèmes dans de nombreux domaines d'application [Feo 1992][Kubena *et al.* 1992]. Au cours des dernières années, les ordinateurs parallèles allant de dizaines à des milliers d'éléments informatiques sont devenus disponibles sur le marché. Ils continuent d'être reconnus comme de puissants outils de recherche scientifique, de gestion de l'information et d'applications d'ingénierie. Cette tendance est motivée par les langages de programmation parallèle et les outils qui contribuent à rendre les ordinateurs parallèles utiles pour prendre en charge un large éventail d'applications. De nombreux modèles et langages ont été conçus et mis en œuvre pour permettre la conception et le développement d'applications sur des ordinateurs parallèles. Les langages de programmation parallèles (également appelés langages concurrents) permettent de concevoir des algorithmes parallèles sous la forme d'un ensemble d'actions simultanées mappées sur différents éléments informatiques. La conception de langages de programmation parallèles est essentielle pour un traitement de données à grande échelle et une utilisation efficace de nouvelles architectures adaptées pour Big Data. Les langages de haut niveau réduisent les délais de conception et d'exécution des applications parallèles, ils facilitent l'approche des ordinateurs parallèles pour les nouveaux utilisateurs. Nous citons ci-dessous quelques langages de programmation parallèles utilisés de manière efficace dans le domaine Big Data :

— **Java :** Est un nouveau langage de programmation orienté objet qui possède une bibliothèque de classes prenant en charge la programmation avec des threads. La bibliothèque de threads est principalement destinée à l'écriture de programmes monoprocesseurs multithreads tels que des gestionnaires d'interface graphique (GUI). Le parallélisme en Java consiste en un certain nombre de threads qui s'exécutent dans un seul espace de nom d'objet global. Ces threads sont des instances de classes définies par l'utilisateur, ils sont généralement des sous-types de la classe « Thread » de la bi-

bliothèque Java, ils remplacent la méthode d'exécution de la classe Thread afin de définir les tâches à effectuer. Les threads sont des objets qui peuvent être nommés, transmis en tant que paramètres à des méthodes, renvoyés à partir de méthodes, etc. De plus, les méthodes héritées de la classe « Thread » permettent à un thread d'être suspendu, repris, mis en veille pour des intervalles spécifiés de temps, etc. Java prend également en charge la notion de groupes de threads. Les discussions d'un groupe peuvent être suspendues et reprises collectivement [Oaks & Wong 2004][Benmammar 2017].

— **Scala :** Est un langage à objet fonctionnel qui prend en charge pleinement la programmation fonctionnelle et orientée objet. Scala fonctionne sur la machine virtuelle Java, il offre une interopérabilité transparente avec Java, ce qui permet à Scala d'accéder à l'ensemble de l'écosystème proposé par Java. Contrairement aux langages fonctionnels traditionnels, Scala permet une approche progressive d'un style plus fonctionnel en raison de son paradigme multiple. Scala prend en charge les fonctions d'ordre supérieur, les fonctions imbriquées, les fermetures et une syntaxe concise pour définir les fonctions anonymes afin de faciliter la programmation fonctionnelle. Scala a des traits et des classes pour la programmation orientée objet. Les traits sont similaires aux interfaces en Java, mais ils permettent l'implémentation de méthodes par défaut. Scala s'appuie sur des méthodes de calcul parallèles, notamment la méthode des collections parallèles pour permettre des calculs parallèles. Scala utilise également Akka [19] pour tirer parti des traitements multi-cœurs [Odersky *et al.* 2016][Karim & Alla 2017].

— **Python :** En revanche, Python est l'un des langages les plus populaires et les plus utilisés pour le traitement des données en raison de sa simplicité et de sa facilité de maintenance. Il fournit un grand nombre de bibliothèques et de cadres facilitant le calcul haute performance. Dans la programmation parallèle, Python contient des modules intégrés et externes qui simplifient la mise en œuvre. Mais, il peut s'avérer assez délicat, car Python n'est pas vraiment multithread comme Java ou Scala [Pine 2019][Gowrishankar & Veena 2019].

19. Akka est un framework OpenSource soutenu par TypeSafe, disponible à la fois en Scala et en Java. Il permet de gérer efficacement des applications concurrentes et encourage la programmation réactive et événementielle.

4.2.2 Calcul distribué

Le paradigme distribué est apparu comme une alternative aux superordinateurs coûteux afin de répondre aux nouveaux besoins des utilisateurs et des applications. À l'inverse des superordinateurs, les systèmes informatiques distribués sont des réseaux composés d'un grand nombre de nœuds ou d'entités connectés via un réseau local rapide. Les nœuds (ordinateurs) d'un système distribué sont indépendants, ils ne partagent pas physiquement la mémoire ou les processeurs, ces nœuds apparaissent à ses utilisateurs comme un système cohérent unique [Andrew & Steen 2016]. Ils communiquent entre eux par des messages, des informations transférées d'un ordinateur à un autre sur un réseau. Les messages peuvent communiquer de nombreuses choses : les ordinateurs peuvent demander aux autres ordinateurs d'exécuter des procédures avec des arguments particuliers, ils peuvent envoyer et recevoir des paquets de données, ou ils peuvent envoyer des signaux aux autres ordinateurs d'un système distribué pouvant jouer différents rôles. Le rôle d'un ordinateur dépend de l'objectif du système et des propriétés matérielles et logicielles. Les systèmes distribués constituent un grand parapluie dans lequel plusieurs systèmes différents sont classés. Alors, la différence entre le calcul parallèle et le calcul distribué ; c'est que dans le calcul distribué plusieurs nœuds de calcul coopèrent pour résoudre un problème, mais ils ne peuvent pas réduire le temps de calcul, car plusieurs opérations ne peuvent pas être effectuées simultanément. La présence de plusieurs ordinateurs traitant les mêmes données signifie qu'un dysfonctionnement dans l'un des ordinateurs n'influence pas l'ensemble du processus. Le calcul distribué est désormais au cœur du Big Data. Les caractéristiques du Big Data exigent des systèmes informatiques distribués et des technologies de traitement de données innovantes et rentables qui permettent une meilleure compréhension, une prise de décision et une automatisation des processus.

4.2.2.1 Les systèmes informatiques distribués conçus pour Big Data

Les systèmes informatiques distribués divisent les gros problèmes ingérables liés au traitement, au stockage et à la communication en petites parties gérables et les résout efficacement de manière coordonnée. Ces systèmes sont constitués d'un nombre d'éléments de traitement interconnectés par un réseau informatique coopérant à l'exécution de certaines tâches assignées. Lorsque les données deviennent volumineuses, la base de données est distribuée sur différents sites. Les bases de données distribuées ont besoin un système distribué pour stocker, récupérer et mettre à jour les données de manière périodique. Dans cette partie, quelques systèmes distribués adaptés pour Big Data seront présentés.

4.2.2.1.1 Les réseaux Pair à Pair

Un réseau Pair à Pair en anglais -*Peer to Peer*- ou -*P2P*- est une topologie de réseau (filaire et/ou sans fil) dans laquelle chaque nœud appelé pair ou servent (client – serveur) à la fois. Tous les nœuds coopèrent dans la distribution des données dans le réseau, ils forment donc une structure sous forme de grille. Par conséquent, chaque nœud doit recevoir, envoyer et relayer les données. Un réseau P2P peut être conçu à l'aide d'une technique d'inondation ou de routage. Lors de l'utilisation d'une technique de routage, le message est propagé le long d'un chemin en sautant de nœud en nœud jusqu'à ce que la destination soit atteinte. Pour assurer la disponibilité de tous ses chemins, un réseau de routage doit permettre des connexions continues et une reconfiguration autour de chemins coupés ou bloqués à l'aide d'algorithmes à réparation automatique. Un réseau P2P auquel tous les nœuds sont connectés les uns aux autres est un réseau entièrement connecté [Diger 2001][Ciglaric *et al.* 2003][Djafri & Mekki 2012].

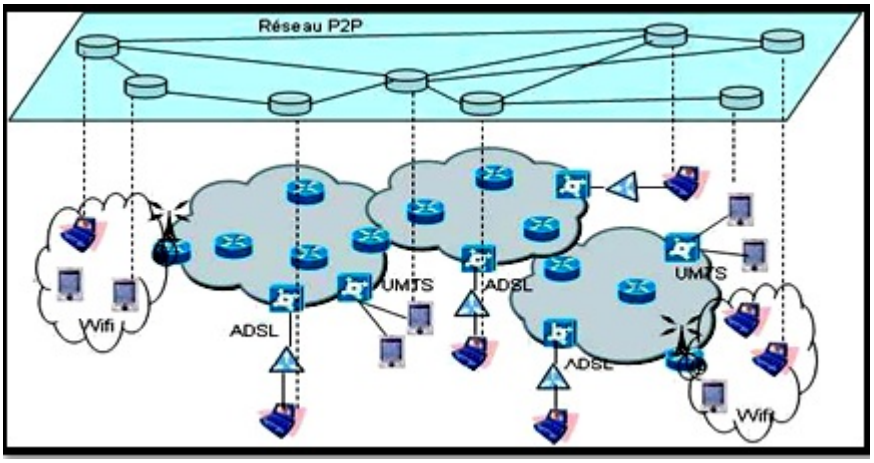

FIGURE 4.12 – *Topologie des réseaux P2P*

De manière générale, le terme pair-'a-pair fait référence à une classe de systèmes ou d'applications qui utilisent des ressources distribuées pour résoudre un problème de façon décentralisée. Ce problème peut être la recherche de fichiers, l'utilisation de ressources de stockage ou de calcul. Les systèmes P2P peuvent être caractérisés comme des systèmes distribués dans lequel tous les nœuds sont identiques, les capacités, les responsabilités et toutes les communications sont symétrique.

Jusqu'à présent, un réseau P2P est un nouveau paradigme appliqué à des ensembles de données volumineux pour les gérer et les traiter dans un délai acceptable. Le Big Data est très important pour comprendre le monde. Les données massives sont souvent capturées à partir de diverses sources, notamment les médias sociaux, les réseaux de capteurs, les applications scientifiques, la surveillance, les textes, documents Internet, les renseignements commerciaux et les journaux web. Il existe plusieurs caractéristiques typiques, notamment la grande taille, les structures hétérogènes, et le traitement complexe. En parallèle, il existe également

de nombreuses techniques importantes y compris un nouveau modèle de programmation, une nouvelle architecture de système, de nouveaux schémas de stockage de données et de schémas de partition de données, etc. Les technologies P2P démontrent une bonne évolutivité et un faible coût pour les services de streaming multimédia et de partage de fichiers. Les réseaux P2P peuvent être appliqués à la gestion des données volumineuses pour résoudre certains problèmes clés, notamment la manière d'organiser les données volumineuses pour l'indexation, la recherche, le traitement distribué et l'affectation de tâches de traitement de données dans un environnement distribué.

4.2.2.1.2 Les Clusters

Un cluster est un réseau homogène dans lequel les périphériques contiennent les mêmes composants matériels ainsi que le même système d'exploitation, ces composants sont étroitement connectés, ils fonctionnent ensemble, ce qui permet de nombreuses manières de les afficher en tant que système unique. Les composants d'un cluster sont généralement connectés les uns aux autres par le biais de réseaux locaux rapides (LAN), chaque nœud (ordinateur utilisé en tant que serveur) exécutant sa propre instance d'un système d'exploitation. Les clusters sont apparus à la suite de la convergence d'un certain nombre de tendances informatiques, notamment la disponibilité de microprocesseurs à faible coût, la disponibilité de réseaux à haut débit et la la disponibilité de logiciels pour le calcul distribué haut performance. Les clusters sont généralement déployés pour améliorer les performances et la disponibilité par rapport à un seul ordinateur, tout en étant généralement beaucoup plus économiques que des ordinateurs uniques d'une vitesse ou d'une disponibilité comparable [Mittal & Suri 2012][Khoussainov & Patel 2007].

FIGURE 4.13 – *Illustration du cluster construit par Aspen Systems, Inc.*

La quantité de données produites dans la communauté scientifique ou dans le monde commercial est en augmentation constante. Cette croissance induit des changements d'échelles dans les processus de collecte, de traitement et de stockage des données. Le domaine du Big Data a émergé pour faire face à ces nouveaux défis. Les outils de traitement disponibles dans la communauté Big Data sont faciles à utiliser, mais ils manquent de performances informatiques. L'un des principaux objectifs du domaine du calcul distribué est de fournir des infrastructures permettant aux ordinateurs de fonctionner le plus rapidement possible. Une infrastructure de calcul distribué (cluster) peut être utile pour le traitement rapide et fiable de données massives à l'aide de l'utilisation intégrée et collaborative des ressources autonomes géographiquement séparées. Plusieurs ressources informatiques sont reliées ensemble dans un cluster (y compris des processeurs et des périphériques de stockage) afin de constituer un ordinateur virtuel unique plus grand et plus puissant.

4.2.2.1.3 Les grilles de calcul

Une grille de calcul est un réseau hétérogène dans lequel les périphériques ont des composants matériels et des systèmes d'exploitation différents, ces composants sont connectés ensemble dans le même réseau. Une grille de calcul permet de faire du calcul distribué , elle exploite la puissance de calcul (processeurs, mémoires, ...) de milliers d'ordinateurs afin de donner l'illusion d'un ordinateur virtuel très puissant. Ce modèle permet de résoudre d'importants problèmes de calcul nécessitant des temps d'exécution très longs en environnement "classique". Une grille de calcul intègre également un intergiciel (terme anglais : -*middleware*-) qui permet à toutes les couches réseau et aux services logiciels de dialoguer entre les différents composants d'une application répartie. L'intergiciel masque la complexité des échanges inter-applications. D'une manière générale, la notion de grille de calcul est définie comme étant une infrastructure matérielle et logicielle fournissant un accès fiable -*dependable*-, cohérant -*consistent*-, à taux de pénétration élevé -*pervasive*- et bon marché -*inexpensive*- à des capacités de traitement et de calcul [Djafri & Mekki 2012] [Fedak *et al.* 2001].

La grille est physiquement constituée de nœuds qui sont des processeurs avec leurs disques, l'ensemble étant inter connecté via un réseau dont la qualité peut être primordiale suivant la technologie de grilles retenue. Suivant la voie technologique retenue, ces nœuds sont de serveurs plus ou moins puissants, voire des PC, ou des grappes de serveurs (clusters). Un logiciel d'interface est installé sur chaque nœud. Il assure les activités entre les nœuds, ce processus est supervisé par les systèmes d'exploitation de chaque serveur. Les outils de supervision et de management global de la grille sont logiquement uniques. Mais physiquement, ils sont distribués de nombreuses machines pour améliorer la fiabilité. L'ensemble des logiciels assurant la gestion de la grille est dénommé middleware de la

grille. Ce middleware gère toutes les ressources de la grille, il est constamment informé de son état : unités de traitement et de stockage qui constituent des nœuds, branches de réseau, librairies de programmes ou toute entité pouvant définie et administrée par le gestionnaire de ressources.

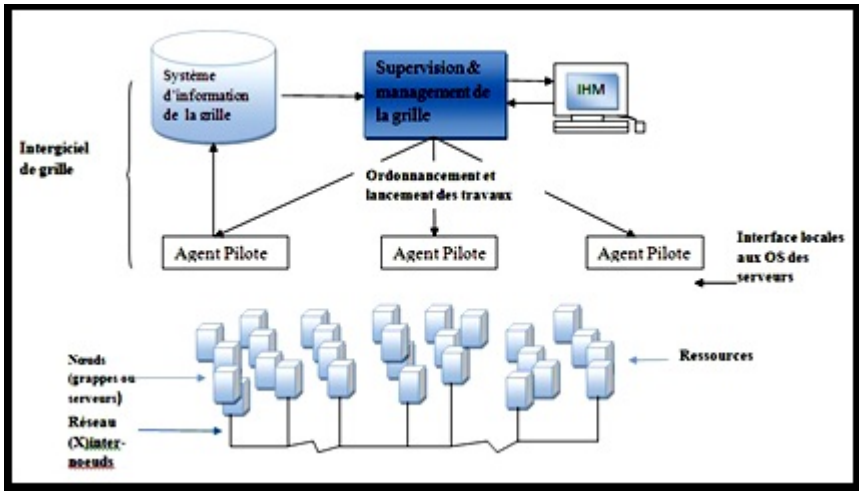

FIGURE 4.14 – *Architecture simplifiée d'une grille de calcul.*

La complexité de calcul s'est encore accrue avec la croissance des données. La grille de calcul est une solution potentielle aux défis informatiques dans le domaine du Big Data [Roberto 2014]. Elle se développe rapidement pour effectuer ce calcul complexe et pour développer de nouvelles applications. C'est une tendance prometteuse pour certaines raisons : sa capacité à produire une utilisation plus rentable de nombreuses ressources informatiques, puis comme moyen de résoudre des problèmes tres complexes, de sorte que la grille de calcul offre des capacités de stockage et de traitement de données volumineuses. Pour que la grille de calcul prenne en charge la gestion et le traitement de données volumineuses, certaines exigences basées sur le concept de données volumineuses doivent être prises en compte. Bien que la grille fournisse la technologie pour surmonter la limitation matérielle en termes d'espace de stockage, de capacité de mémoire, et de puissance de traitement. Mais, il est probablement qu'une grille de calcul peut nécessiter des techniques supplémentaires pour la gestion efficace des données massives.

4.2.2.1.4 Cloud Computing

Aujourd'hui, presque tout le monde est connecté à Internet, ce qui entraîne une croissance rapide des données. Le traitement des données massives est une tâche difficile et fastidieuse qui nécessite une vaste infrastructure informatique. Heureusement, le cloud computing offre une solution prometteuse pour traiter le Big Data de manière extrêmement efficace et économique. Le cloud computing est un type de système parallèle et distribué, il est constitué d'un ensemble d'ordinateurs interconnectés et virtualisés, telles que des ressources matérielles et logicielles, ces res-

sources sont provisionnés de manière dynamique, elles sont présentées également sous la forme d'une ou plusieurs ressources informatiques unifiées basées sur des accords établis entre le fournisseur de services et les consommateurs[Melekhova & Vinnikov 2015]. Le cloud computing repose sur le partage des ressources pour faire des économies d'échelle. Ces services sont divisés en trois catégories, la première catégorie : IaaS *-Infrastructure-as-a-Service-* est la possibilité de fournir des ressources de traitement, de stockage, de réseau et autres ressources informatiques fondamentales. IaaS offre au client la possibilité de déployer et d'exécuter des logiciels arbitraires tels que des systèmes d'exploitation et des applications. La deuxième catégorie :PaaS *-Platform-as-a-Service-* permet de déployer sur l'infrastructure de cloud des applications créées ou acquises par le client en utilisant de langages de programmation et d'outils pris en charge par le fournisseur. Et la dernière catégorie : SaaS *-Software-as-a-Service-* permet d'utiliser les applications du fournisseur fonctionnant sur une infrastructure cloud. Les applications sont accessibles à partir de divers périphériques clients via une interface de client léger tel qu'un navigateur Web (par exemple, un courrier électronique basé sur le Web) [Melekhova & Vinnikov 2015][Kulkarni *et al.* 2012][Aldossary & Allen 2016].

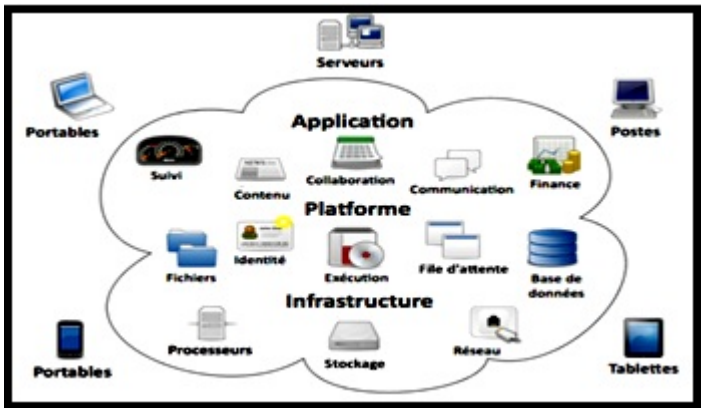

FIGURE 4.15 – *Illustration du cloud computing.*

La fonctionnalité du cloud computing permet d'obtenir un espace de stockage apparemment illimité et une puissance exceptionnelle de calcul. Le cloud computing est considéré comme une fonctionnalité très souhaitable dans chaque solution du Big Data. Les entreprises et les utilisateurs peuvent louer « en tant que service » des infrastructures hardware et software tels que les espaces de stockage, les processeurs, les logiciels, etc. De nombreuses organisations proposent un tel service, notamment Amazon [Amazon AWS], Microsoft [Microsoft Azure] et Google [Google Drive]. Malheureusement, ces systèmes publics ne suffisent pas pour effectuer des calculs complexes sur de gros volumes de données en raison de la faible bande passante ; idéalement, un système informatique en nuage pour Big Data devrait être dispersé géographiquement afin de réduire sa vulnérabilité en cas de catastrophe naturelle, mais devrait également bénéficier d'un niveau élevé d'interopérabilité et de mobilité des données. En fait, certains systèmes vont dans cette direction, tels que le projet OpenCirrus.

D'autre part, le cloud computing est utilisé comme un outil d'analyse et de stockage qui gère le paradigme du Big Data. Il s'agit d'un nouveau concept représentant l'hétérogénéité et la croissance des données qui doivent être gérées et traitées avec précision pour extraire les informations pertinentes nécessaires à la prise des meilleures décisions. Mais, malheureusement, les clouds d'aujourd'hui posent des problèmes de sécurité majeurs liés à l'intégrité, à la disponibilité et à la confidentialité des données et des applications externalisées vers le cloud [Aldossary & Allen 2016]. La multi-location et d'autres propriétés inhérentes au modèle du cloud computing ont introduit de nouvelles surfaces d'attaque et menaces. Les clouds et leurs clients sont également la cible de nouveaux types d'attaques menaçant leur fiabilité et leur durabilité économique. À moins que ces problèmes ne soient résolus, les clouds ne peuvent et ne doivent pas être utilisés dans des domaines sensibles telles que les données et les renseignements liés à la défense, les transactions financières ou les dossiers médicaux.

4.2.2.2 Les technologies (solutions) pour l'analyse de données massives

Le calcul parallèle et distribué revêt une importance primordiale en particulier pour atténuer les problèmes du Big Data tels que le volume et la vélocité en utilisant des solutions et des technologies appropriées. Les solutions ou les technologies du Big Data sont des plateformes qui font le calcul parallèle et distribué. Elles fournissent une puissance de calcul rentable et une efficacité inégalée dans le traitement de grand volume de données en batch, micro batch et en streaming (temps réel) [Akidau *et al.* 2015][da Silva *et al.* 2018], ceci est obtenu au moyen d'une couche d'abstraction réussie basée sur un paradigme de programmation approprié.

Pour gérer le Big Data, des technologies ont été créées, ces technologies sont capables d'utiliser la puissance de calcul et la capacité de stockage d'un cluster avec des performances accrues proportionnelle au nombre de machines existantes. En particulier, l'analyse du Big Data est un domaine prometteur pour la prochaine génération d'innovations dans le domaine de l'automatisation en raison de la nécessité croissante d'extraire de la valeur à partir de données provenant de multiples domaines d'application. Dans cet objectif, divers technologies ont évolué au cours des dix dernières années. Le plus utilisé de ces technologies est Hadoop et son écosystème [20].

20. https ://www.bmc.com/blogs/hadoop-ecosystem/, (consulté le : 02/05/2019)

4.2.2.2.1 Hadoop et son écosystème

Ecosystème Hadoop n'est pas un langage de programmation ; il est tout simplement une plateforme *-framework-* englobant un certain nombre de services (stockage, analyse et maintenance) permettant de résoudre les problèmes liés au Big Data. Il contient maintenant des dizaines de composants, allant de la recherche dans les bases de données, en passant par le stockage des données, le traitement des images, l'apprentissage en profondeur et le traitement du langage naturel. Avec l'avènement de Hadoop v.2, nombreux gestionnaires de ressources peuvent être l'utiliser pour offrir un niveau de sophistication et de contrôle encore plus élevé qu'auparavant.

Nous devons faire face à la question suivante : est-ce que Hadoop est mort ? Cela dépend des limites perçues de Hadoop lui-même. Pensons à Apache Spark qui fait partie de la famille Hadoop, car il utilise également HDFS. Donc le calcul distribué et parallèle est une cible mouvante, et les composants de Hadoop et de son écosystème ont changé de façon remarquable ces dernières années. Dans notre thèse, nous essayons de montrer certains des aspects divers et dynamiques de Hadoop et de son écosystème associé, et nous essayons de vous convaincre malgré les changements. Hadoop est toujours très vivant pour développer des logiciels spéciaux liés au traitement à grand échelle, en particulier dans d'analyse de données massives. Le schéma ci-dessous englobe les composants Hadoop, qui forment un ensemble d'écosystème Hadoop :

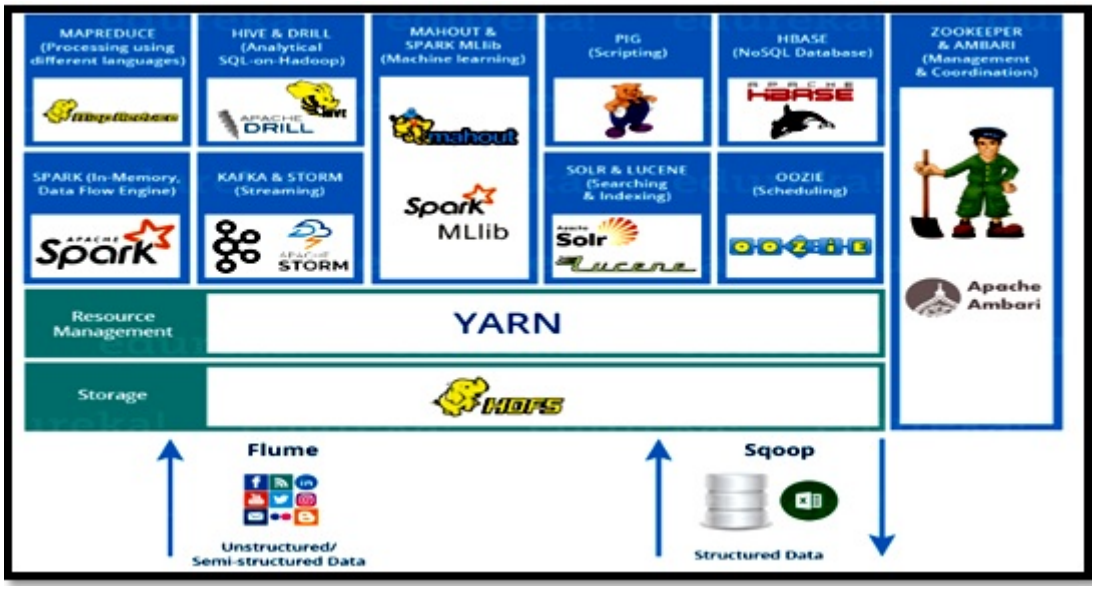

FIGURE 4.16 – *Ecosystème Hadoop.*

4.2.2.2.1.1 Hadoop

Hadoop *-High-availability distributed object-oriented platform-* est un ensemble de classes (Framework) écrites en java, open source de la fondation Apache permettant de répondre aux besoins du Big Data à savoir le volume (de l'ordre de plusieurs péta-octets), la variété et la vélocité des données. Il est composé de quatre composants de base [Mois 2016][Dhyani & Barthwal 2014] :

1. **Hadoop Distributed File System (HDFS) :** HDFS est un système de fichiers distribué, extensible et portable inspiré par le Google File System (GFS). Il est utilisé principalement pour le stockage persistant de données qui fournit un accès haut-débit aux données de l'application, il a été conçu pour stocker de très gros volumes de données sur un grand nombre de machines équipées de disques durs banalisés. Le HDFS est le composant principal de Hadoop Ecosystème. Il a été conçu pour stocker différents types de grands ensembles de données structurées, non structurées et semi-structurées. Il peut fonctionner dans des machines à faible coût avec une tolérance aux pannes élevée [Borthakur 2007]. HDFS définit deux types de nœuds [Borthakur 2007][Honnutagi 2014] :

   (a) **Le nœud principal *-NameNode-* :** Il est caractérisé par :
   - Responsable de la distribution et de la réplication des blocs ;
   - Stocker et gérer les données massive ;
   - Contient la liste des DataNodes pour chaque bloc (dans le cas de l'écriture) ;
   - Contient la liste des blocs pour chaque fichier (dans le cas de lecture).

   (b) **Le nœud de données *-DataNode-* :** Il est caractérisé par :
   - Stocker des blocs de données dans le système de fichier local ;
   - Maintenir des métadonnées sur les blocs possédés ;
   - Servir de bloc de données et de métadonnées pour le client HDFS.

   Le *-NameNode-* dans l'architecture Hadoop est un point unique de défaillance. Si ce service est arrêté, il est impossible de pouvoir extraire des blocs d'un fichier donné. Pour résoudre ce problème, un nœud secondaire appelé : *Secondary NameNode* a été mis en place dans l'architecture Hadoop.

2. **Hadoop YARN (Yet Another Resource Negotiator) :** Yarn est un framework pour la planification des tâches et la gestion des ressources de calcul en grappes. Il est considéré comme le cerveau de l'écosystème Hadoop. Il rend l'environnement Hadoop

mieux adapté aux applications opérationnelles, il effectue toutes les activités de traitement en allouant des ressources et en planifiant des tâches. Hadoop YARN comporte deux composants principaux : Le planificateur -*Scheduler*- et le gestionnaire d'applications -*Applications Manager*-. Le planificateur est responsable de l'affectation des ressources aux diverses applications en cours d'exécution ; tandis que, le gestionnaire d'applications est responsable d'accepter les soumissions de travaux et l'exécution d'Application Master [Kulkarni & Khandewal 2014].

3. **Hadoop Map-Reduce :** MapReduce est une structure de programme qui permet d'écrire des applications qui gèrent de grands ensembles de données à l'aide d'algorithmes distribués et parallèles dans l'environnement Hadoop. Un programme MapReduce peut être résumé en deux fonctions Map () et Reduce (). La fonction Map effectue plusieurs actions tels que le filtrage, le regroupement et le tri [Jeong & CHA 2019]. Tandis que, La fonction Reduce agrège le résultat produit par la fonction Map. Le résultat généré par la fonction Map est une paire de valeurs de clé (K, V) qui sert d'entrée à la fonction de réduction [Kulkarni & Khandewal 2014] [Jeong & CHA 2019].

4. **Hadoop Common :** Hadoop Common contient toutes les bibliothèques et les outils nécessaires pour que d'autres logiciels puissent se servir d'Hadoop [Mois 2016]. Plus concrètement, il permet de couvrir le stockage et la répartition des données, les traitements distribués, l'entrepôt de données, le workflow, la programmation, sans oublier la coordination de l'ensemble des composants.

4.2.2.2.1.2 Ecosystème Hadoop

Le terme Hadoop se réfère non seulement aux modules de base ci-dessus, mais aussi à son écosystème et à l'ensemble des logiciels qui viennent s'y connecter comme : Hbase, Pig, Hive, Flume, Oozie, Drill, Zookeeper, Sqoop, Solr, Lucene, Ampari, Mahout, Storm, Kafka, et Spark [21].

— **Apache HBase** est un système de gestion de base de données non relationnelle orienté colonne qui s'exécute sur le système de fichiers distribués Hadoop [22][Patel 2017].

— **Apache Pig** est développé par Yahoo, Pig est une plate-forme utilisée pour analyser de grands ensembles de données. Il est

21. https ://fr.wikipedia.org/wiki/Hadoop, (consulté le :03/05/2019)
22. Apache HBase Team (2015), Apache HBase TM reference guide (Version 2.0.0-SNAPSHOT ed.) Apache.

conçu pour fournir une abstraction sur MapReduce, cela réduit la complexité de l'écriture d'un programme MapReduce. Les opérations de manipulation de données sont effectuées facilement dans Hadoop avec Apache Pig. Pig se compose de deux parties : Pig Latin : langage procédural de haut niveau. Et Pig Engine : parse, optimise et exécute automatiquement les scripts PigLatin comme une série de jobs MapReduce au sein d'un cluster Hadoop [OMI 2012][Ansari & Swarna 2017].

— **Apache Hive** est à l'origine du projet Facebook, puis il a donné à la fondation Apache qui fait un lien entre SQL et Hadoop [23]. Apache Hive est un système d'entrepôt de données construit sur Hadoop, il est utilisé pour analyser des données structurées et semi-structurées. Hive fait abstraction de la complexité de Hadoop MapReduce. Fondamentalement, il fournit un mécanisme pour exécuter des requêtes écrites en HQL -*Hive Query Language*- similaires aux instructions SQL, ces requêtes ou HQL sont converties en mappage de tâches réduites par le compilateur Hive. Apache Hive prend en charge le langage DDL -*Data Definition Language*-, le langage DML -*Data Manipulation Language*- et les fonctions définies par l'utilisateur [Adamov 2018].

— **Apache Flume** est un outil d'ingestion de données dans HDFS. Il collecte, regroupe et transporte une grande quantité de données en continu tels que les fichiers journaux, les événements provenant de diverses sources tels que le trafic réseau, les médias sociaux, les courriers électroniques, etc. vers HDFS. Flume est donc un outil très fiable et distribué [Crinivasa *et al.* 2018].

— **Apache Oozie** est un système évolutif, fiable et extensible permettant de gérer les travaux Apache Hadoop, il est intégré au reste de la pile Hadoop. Il prend en charge de nombreux types de fonctions Hadoop prédéfinies tels que Pig, Hive, Sqoop et Distcp [24], il prend également des fonctions spécifiques aux systèmes tels que Programmes Java et scripts shell [25].

— **Apache Drill** est un framework open source qui fonctionne avec un environnement distribué pour analyser de grands ensembles de données. Il est utilisé pour explorer n'importe quel type de données tels que Parquet, JSON et CSV. Il supporte différents types de bases de données NoSQL et de systèmes de fichiers, ce qui est une fonctionnalité puissante de Drill [26] [27].

23. https ://fr.wikipedia.org/wiki/Apache_Hive, (consulté le :03/05/2019)
24. DistCp (distributed copy) est un outil utilisé pour la copie volumineuse inter / intra-cluster. Il utilise MapReduce pour effectuer sa distribution, sa gestion des erreurs, sa récupération et sa génération de rapports.
25. https ://fr.wikipedia.org/wiki/Apache_Oozie, (consulté le : 03/05/2019)
26. https ://www.oreilly.com/library/view/learning-apache-drill/9781492032786/, (consulté le : 03/05/2019)
27. https ://fr.wikipedia.org/wiki/Apache_Drill, (consulté le : 03/05/2019)

— **Apache Zookeeper** est le coordinateur de tout travail Hadoop qui comprend une combinaison de divers services dans un écosystème Hadoop. Il assure la coordination avec divers services dans un environnement distribué [28]. Avant Zookeeper, la coordination entre les différents services de l'écosystème Hadoop était très longue et difficile. Auparavant, les services avaient de nombreux problèmes d'interactions, telle que la configuration commune lors de la synchronisation des données. Même si les services sont configurés, des modifications apportées à leur configuration rendent le traitement complexe et difficile à gérer. Le regroupement et la dénomination prenaient également beaucoup de temps.

— **Apache Sqoop** est un outil conçu pour transférer efficacement des données entre Apache Hadoop et des data-stores structurés telles que des bases de données relationnelles. La principale différence entre Flume et Sqoop est que Flume ingère uniquement des données non structurées ou semi-structurées dans HDFS ; tandis que, Sqoop peut importer et exporter des données structurées à partir de SGBDR ou d'entrepôts de données d'entreprise vers HDFS ou vice versa [29].

— **Apache Solr et Lucene** sont les deux services utilisés pour la recherche et l'indexation dans Ecosystème Hadoop. Apache Lucene est une technologie basé sur Java qui aide également à la vérification orthographique, de mise en évidence des occurrences et d'analyse ou création de jetons avancée. Solr est un serveur de recherche hautes performances construit autour de Lucene [30].

— **Apache Ambari** est un projet de la fondation Apache Software qui vise à rendre l'écosystème Hadoop plus facile à gérer. Il comprend un logiciel de provisionnement, de gestion et de surveillance des clusters Apache Hadoop [31].

— **Apache Mahout** est une bibliothèque fournit un environnement permettant de créer des applications d'apprentissage automatique puissante et évolutive qui s'exécute sur Hadoop MapReduce [32]. L'apprentissage automatique est une discipline de l'intelligence artificielle qui permet aux systèmes d'apprendre en se basant uniquement sur des données, ce qui améliore continuellement les performances lors du traitement de grande quantité de données. L'apprentissage automatique est à la base de nombreuses technologies qui font partie de notre quotidien.

28. Tutorial : zookeeper, The Apache Software Foundation, disponible sur : https ://zookeeper.apache.org/doc/r3.1.2/zookeeperOver.pdf
29. https ://blogs.apache.org/sqoop/entry/apache_sqoop_overview, (consulté le : 03/05/2019)
30. https ://lucene.apache.org/, (consulté le : 03/05/2019)
31. https ://ambari.apache.org/,03/05/2019, (consulté le : 04/05/2019)
32. https ://fr.wikipedia.org/wiki/Apache_Mahout, consulté le : 04/05/2019

— **Apache Storm** est un logiciel open source géré par fondation Apache. Il a été déployé pour répondre aux besoins de traitement en temps réel de sociétés tels que Twitter, Yahoo. Apache Storm est donc devenu la plate-forme de choix des experts dans le domaine Big Data pour développer des plates-formes de traitement de données distribuées en temps réel. Il fournit un ensemble de primitives pouvant être utilisées pour développer des applications capables de traiter une très grande quantité de données en temps réel de manière hautement évolutive [33] [34].

— **Apache Kafka** Apache Kafka est un magasin dans lequel les messages sont stockés à partir de processus appelés producteurs, il est développé par l' *apache software foundation*, il écrit en Scala. Apache Kafka permet de résoudre les problèmes d'utilisation des données en temps réel à des fins de consommation. Il gère une charge de données considérable, il évite également les systèmes à contre-pression pour gérer les inondations. Il inspecte les volumes de données entrants, ce qui est très important pour la distribution et le partitionnement sur les nœuds du cluster [35][Javed 2017].

— **Apache Spark** Apache Spark est un framework d'analyse de données en temps réel dans un système distribué. Il exécute des calculs en mémoire pour augmenter la vitesse de traitement des données sur Map-Reduce. Il est plus rapide que Hadoop MapReduce pour le traitement de données à grande échelle en exploitant des calculs en mémoire et d'autres optimisations [Prakash & Atul 2016]. Spark présente plusieurs avantages par rapport aux autres technologies Big Data notamment MapReduce du Hadoop et Storm. D'abord, Spark propose un framework complet et unifié pour répondre aux besoins de traitements Big Data ainsi que par type de traitement (batch, micro batch, ou streaming). Ensuite, Spark permet à des applications sur Hadoop d'être exécutées jusqu'à 100 fois plus rapide en mémoire et 10 fois plus rapide sur disque [Bhattacharya & Bhatnagar 2016]. Il nous permet d'écrire rapidement des applications en Java, Scala, R ou Python, il inclut également un jeu de plus de 80 opérateurs haut-niveau [Dataflair 2018]. Il est composé de cinq composants de base [36] [37] [38] [39] [Bhattacharya & Bhatnagar 2016][Dataflair 2018][Haoyuan *et al.* 2013] [Meng *et al.* 2016][Gonzalez *et al.* 2014] :

33. https ://storm.apache.org/, (consulté le : 04/05/2019)
34. https ://www.journaldunet.com/web-tech/developpeur/1127535-storm-apache-integre-le-big-data-temps-reel-de-twitter/, (consulté le : 04/05/2019)
35. https ://fr.wikipedia.org/wiki/Apache_Kafka, (consulté le : 04/05/2019)
36. https ://spark.apache.org/docs/latest/sql-programming-guide.html, consulté le : 05/05/2019
37. https ://opensource.com/article/19/3/apache-spark-and-dataframes-tutorial, (consulté le : 05/05/2019)
38. https ://acadgild.com/blog/integrate-apache-flume-spark-streaming, (consulté le : 05/05/2019)
39. https ://data-flair.training/forums/topic/what-is-spark-core/, (consulté le : 06/05/2019)

1. **Spark SQL :** Spark SQL est un module Spark pour le traitement de données structurées [40]. Il fournit une abstraction de programmation appelée *-DataFrames-*, il peut également agir en tant que moteur de requête SQL distribué. Il permet aux requêtes Hadoop Hive non modifiées de s'exécuter jusqu'à 100 fois plus rapidement sur les déploiements et les données existants [Dataflair 2018]. Il fournit également une intégration puissante avec le reste de l'écosystème Spark (par exemple, en intégrant le traitement de requête SQL avec l'apprentissage automatique) [41].

2. **Spark Stearming :** Spark Streaming permet de puissantes applications interactives et analytiques pour Big Data en héritant des caractéristiques de facilité d'utilisation et de tolérance aux pannes de Spark. Il s'intègre facilement à une grande variété de sources de données populaires, notamment HDFS, Flume et Kafka. Spark Streaming peut être utilisé pour les traitements des données en temps-réel, il utilise le module DStream, c-à-d une série de RDD *-Resilient Distributed Dataset-* [42] [43] [Haoyuan *et al.* 2013].

3. **Spark MLlib :** MLlib est une bibliothèque d'apprentissage automatique *-Machine Learning-* permettant d'analyser le Big Data, il fournit plus de 55 algorithmes d'apprentissage automatique évolutifs qui bénéficient largement de la parallélisation des données [Meng *et al.* 2016]. Il fournit également à la fois des algorithmes de haute qualité (par exemple, plusieurs itérations pour augmenter la précision) et une vitesse fulgurante (jusqu'à 100 fois plus rapide que MapReduce) [Bhattacharya & Bhatnagar 2016]. MLlib du Spark est utilisé avec des applications écrites en Java, Scala, R et Python afin que nous puissions l'inclure dans le traitement du Big Data [Dataflair 2018].

40. https ://spark.apache.org/docs/latest/sql-programming-guide.html, (consulté le : 05/05/2019)
41. https ://opensource.com/article/19/3/apache-spark-and-dataframes-tutorial, (consulté le : 05/05/2019)
42. Un RDD est une collection de données calculée à partir d'une source et conservée en mémoire vive (tant que la capacité le permet).
-Résilient : tolérant aux pannes à l'aide du graphe de lignage RDD, et donc capable de recalculer les partitions manquantes ou endommagées en raison de défaillances de nœuds.
-Distributed :distribué avec des données résidant sur plusieurs nœuds d'un cluster.
-Dataset : est un ensemble de données partitionnées avec des valeurs primitives ou des valeurs de valeurs, par ex. n-uplets ou autres objets (qui représentent des enregistrements des données avec lesquelles vous travaillez).
43. https ://acadgild.com/blog/integrate-apache-flume-spark-streaming, (consulté le : 05/05/2019)

4. **Spark GraphX :** GraphX est un moteur de calcul graphique qui permet la composition de graphes à partir des données non structurées et des tableaux, il permet également d'afficher les mêmes données physiques sous forme de graphe, sans déplacement ni duplication de données [Gonzalez *et al.* 2014].

5. **Spark Core :** Spark Core est le moteur d'exécution général sous-jacent de la plateforme Spark sur lequel toutes les autres fonctionnalités sont créées. Spark Core utilise une structure de données fondamentale spécialisée : RDD, RDD est un ensemble logique de données partitionnées sur plusieurs ordinateurs[44]. Les RDD peuvent être créés de deux manières : la première consiste à référencer des ensembles de données dans des systèmes de stockage externes, et la seconde consiste à appliquer des transformations (par exemple, carte, filtre, réducteur, jointure) aux RDD existants.

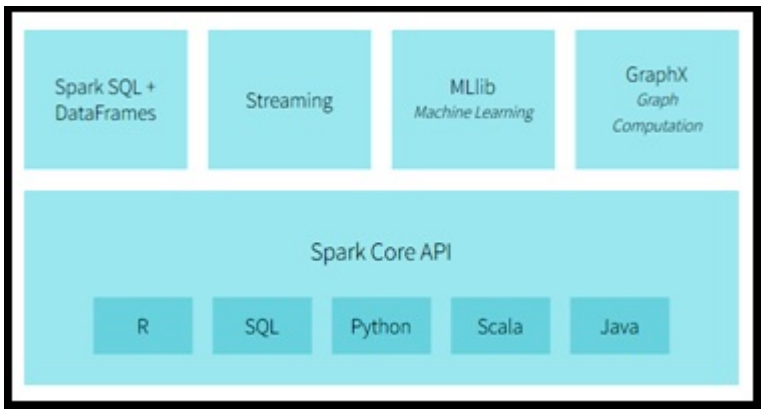

FIGURE 4.17 – *Ecosystème Apache Spark.*

4.3 Les technologies du Big Data et Machine Learning

Dans de nombreux domaines, les données arrivent plus rapidement que nous pouvons apprendre et comprendre. Pour éviter de gaspiller ces données, nous devons abandonner les méthodes traditionnelles d'apprentissage automatique, ces méthodes devraient plutôt être utilisées dans des systèmes capables d'analyser des données massives. Pour cela, nous présentons par la suite quelques solutions du Big Data qui peuvent être aidées les algorithmes du *Machine Learning* pour la prédiction à grandes échelles.

4.3.1 Machine Learning sous Spark

Spark est une infrastructure informatique open source essentiellement adaptée au traitement de données massives. Apache Spark a été

44. https ://data-flair.training/forums/topic/what-is-spark-core/, (consulté le : 06/05/2019)

conçu comme un moteur unifié pour le traitement distribué et parallèle. Spark a un modèle de programmation similaire à MapReduce qui s'étend à une abstraction de partage de données appelée :"RDD". Cette extension capture un grand ensemble de données pour les traiter qui nécessite SQL, l'apprentissage automatique et le traitement graphique [Matei & Mosharaf 2010]. Cette implémentation utilise les mêmes optimisations que le traitement par colonne et les mises à jour incrémentielles pour atteindre des performances similaires. Le traitement de données évolutif est essentiel pour la prochaine génération d'applications informatiques qui implique une séquence complexe d'étapes de traitement avec différents systèmes informatiques. Le projet Spark est une solution prometteuse pour les applications Big Data [Matei & Reynold 2016], c'est ce que nous allons démontrer dans le dernier chapitre .

Notre thèse comprenait de nombreux algorithmes d'apprentissage automatique et des outils d'analyse de données interactifs permettant de gérer le traitement des données. Spark est une nouvelle infrastructure qui prend en charge ces applications en conservant l'évolutivité et la tolérance aux pannes de Map-Reduce. Une abstraction (RDD) a été introduite en tant que collection en lecture seule d'objets partitionnés sur un ensemble de machines [Meng *et al.* 2015].

Apache Spark est l'un des solutions de traitement open source les plus utilisées pour les données massives avec des bibliothèques intégrées et des API plus riches. MLlib est une bibliothèque efficace pour apprendre la fonctionnalité des modèles (classifieurs), cette bibliothèque est composée de plusieurs primitives statistiques et d'optimisation.

4.3.2 Machine learning sous cloud computing

Le cloud computing est une technologie bien développée utilisée par de nombreuses entreprises réputées telles qu'Amazon, Microsoft et Google. Il semble difficile d'imaginer une tendance technologique qui perturberait le paysage actuel du cloud computing. L'intelligence artificielle présente toutefois des caractéristiques uniques qui peuvent certainement influer sur la prochaine génération de plates-formes de cloud computing. L'intelligence artificielle requiert une nouvelle infrastructure informatique pour prendre en charge de nouveaux modèles et paradigmes de traitement de données massives. L'infrastructure informatique traditionnelle n'a pas fourni suffisamment de puissance de calcul pour analyser les données volumineuses qui deviennent maintenant courantes. Le cloud computing constitue une alternative efficace pour améliorer la puissance de calcul. Les algorithmes d'apprentissage automatique sont des méthodes analytiques puissantes qui permettent aux machines de reconnaître des modèles, ils facilitent également l'apprentissage humain. Ainsi, la sélection robuste associant l'analyse des données en utilisant des algorithmes d'apprentissage automatique efficaces augmente la capacité des praticiens à utiliser largement le cloud computing [Buyya *et al.* 2016].

Conclusion

Nous avons parlé dans ce chapitre sur le calcul parallèle et distribué en présentant leurs systèmes. Ainsi que, nous avons présenté les technologies du Big Data et comment les utiliser pour le traitement à grand d'échelle. Le traitement du Big Data consiste à collecter, gérer, analyser et comprendre les données à des volumes et à des débits qui repoussent les frontières des technologies actuelles. Les enjeux du Big Data est de fournir des architectures matérielles, ainsi que des systèmes logiciels capables de traiter efficacement ces données volumineuses. Le traitement de données volumineuses nécessite des solutions différentes des solutions de traitement traditionnel. Ces solutions s'adaptent bien avec le calcul parallèle et distribue des données volumineuse. Elles nécessitent également un degré extrêmement élevé de tolérance aux pannes, de fiabilité et de disponibilité. De plus, les solutions du Big Data aident à donner des réponses relativement rapides. Dans le chapitre suivant, nous discuterons la méthodologie et les solutions proposées pour l'analyse prédictive des données massives.

Méthodologie 5

Sommaire

5.1 Introduction

Nous vivons maintenant à l'ère du Big Data. Mais l'analyse de données traditionnelle ne peut pas être en mesure de gérer de telles quantités de données. La question qui se pose maintenant est de savoir comment concevoir une architecture pour analyser efficacement ces données massives ? Et comment développer un algorithme de prédiction approprié pour l'analyse du Big Data ? C'est ce que nous discuterons en détail dans ce chapitre. Le présent chapitre commence par une présentation des travaux connexes, suivie de nos travaux proposés avec certaines questions importantes, et d'autres axes de recherche seront également présentés relatifs à l'analyse du Big Data. Enfin, Nous concluons ce chapitre en faisant une synthèse des résultats obtenus et en indiquant quelques pistes pour d'autres publications en lien avec notre sujet.

5.2 Etat de l'art

L'analyse du Big Data au cœur de l'apprentissage automatique *-Machine Learning-* qui fait partie de l'intelligence artificielle pour prédire un comportement futur face à une nouvelle donnée, et approximer une fonction de densité de probabilité [1]. Mais les algorithmes du Machine Learning pour l'analyse du Big Data prennent beaucoup de temps à s'exécuter. Ce problème exige de nouvelles technologies très gourmandes en ressources de calcul pour le traitement avec haute performance de grand volume de données. Certains algorithmes d'apprentissage automatique sont manipulés dans des architectures distribuées pour les calculs parallèles. Il existe plusieurs travaux dans ce contexte, par exemple [Mohan & Pramod 2017] présente l'algorithme parallèle de propagation d'étiquettes pour la détection de la communauté et l'algorithme de mesure de similarité parallèle pour la prédiction du lien. Un autre travail réalisé par [LakKang & Timothy 2017] qui permet de faire des calculs à la fois parallèles et distribués pour l'analyse du Big Data, ce travail est basé sur l'implémentation parallélisée de l'algorithme BFGS *-Broyden-Fletcher-Goldfarb-Shanno-* [2] sur la plate-forme HPCC-*High-Performance Computing Cluster-* [3]. On trouve également l'architecture Lambda qui fournit un modèle de traitement par lot *-Batch-* sur des volumes importants de données [Murphy *et al.* 2015]. Elle permet de conserver les principes du Big Data, telle que la scalabilité et la tolérance aux pannes. Par ailleurs l'architecture Kappa est plus simple que l'architecture Lambda. Elle permet de simplifier l'architecture Lambda en fusionnant les couches du temps réel et batch. Cependant, elle ne permet pas le stockage permanent des don-

1. La fonction de densité de probabilité permet de déterminer des zones de forte et de faible probabilité pour les valeurs d'une variable aléatoire.

2. BFGS : est une méthode permettant de résoudre un problème d'optimisation non linéaire sans contraintes.

3. HPCC, également connu sous le nom de DAS (Data Analytics Supercomputer) est une plate-forme de système informatique à source ouverte et à forte intensité de données développée par LexisNexis Risk Solutions.

nées [Bouazza 2017]. Il existe aussi l'architecture HDFS et l'architecture HBase [Bouazza 2017], tandis que les solutions techniques utilisées dans ces architectures sont basées sur la solution du Big Data Hadoop et son écosystème.

L'analyse du Big Data est influencée par d'autres techniques statistiques notamment l'échantillonnage. On trouve, dans ce contexte l'échantillonnage BLB-*Bag of Little Bootstrap*- pour Big Data [Kleiner *et al.* 2014], il est bien adapté aux architectures parallèles et distribuées. Il est également utilisé dans le cas où le traitement d'une population volumineuse est impossible. Les échantillons produits par les méthodes d'échantillonnage permettent de réduire la taille des données afin de diminuer le temps du calcul global avec la conservation de la précision des résultats obtenus à partir d'une population cible [Albattah 2016][Saleema *et al.* 2014].Ces échantillons ont une taille bien déterminée. Pour calculer la taille de l'échantillon [18] requise, il faut tenir en compte les facteurs suivants : (1) le niveau de confiance souhaité, (2) la marge d'erreur acceptable, et (3) l'effet du plan d'échantillonnage. Par exemple, une marge de précision de $\pm$ 10% nécessite une taille d'un échantillon d'environ 100 individus, s'augmentant à environ 400 individus pour une précision de $\pm$ 5%, et d'environ 10 000 individus pour une marge de précision de $\pm$ 1%. Dans le cas général, si la taille de la population est plus grande (des centaines de millions d'individus) ; alors la taille d'échantillon requise ne dépasse pas 10.000 individus pour une marge d'erreur $\pm$ 1% avec un niveau de confiance de 95%, elle ne dépasse pas 20.000 individus pour une marge d'erreur $\pm$ 1% avec un niveau de confiance de 99% [Etikan *et al.* 2016].

5.3 Présentation de notre travail réalisé

Notre travail est principalement axé sur l'étude de l'analyse de données massives pour la prédiction. Cette partie constitue la partie pratique de l'épreuve de notre projet. Notre projet est un ensemble des tests coordonnés et maîtrisés comportant des architectures et des solutions du Big Data supportées par les algorithmes du *Machines Learning*. Il implique donc la définition d'objectifs, la description des activités à réaliser et leur articulation, le choix des techniques à mettre en œuvre mais également des outils technologiques à mobiliser et enfin la maîtrise des contraintes temporelles et matérielles qui vont conditionner la réussite du projet. Nous pouvons être amenés à privilégier une dimension de la problématique, dans ce cas, notre choix devra être discuté. Nous allons donc être au cœur d'un travail dans lequel nous devrons identifier un besoin et proposer des solutions bien pensées pour répondre au problème posé.

5.3.1 Le modèle proposé pour l'analyse de données massives

Le développement des algorithmes du Machine Learning dans un contexte distribué est très complexe, mais avec les solutions du Big Data il devient simple. Le modèle proposé est basé sur : (1) une architecture distribuée portée par des solutions du Big Data tel que Hadoop, Spark et Zookeeper pour le calcul distribué et parallèle, (2) les méthodes de partitionnement des données Map-Reduce et l'échantillonnage aléatoire stratifié, et (3) la méthode d'apprentissage supervisé Random Forests pour la prédiction. Mais avant de réaliser ce modèle proposé, il faut répondre aux questions suivantes : (1) Comment organiser et gérer ces ressources informatiques dans une architecture adaptée avec Machine Learning distribués ? et comment rendre cette architecture plus efficaces pour l'analyse du Big Data ?, (2) Quelles sont les garanties du résultat et de performance si on divise la base d'apprentissage volumineuse sur différents nœuds de calcul lors de la prédiction ?, (3) Comment réaliser cette partition ? Et (4) A quelle vitesse devons-nous obtenir ces résultats ? La figure ci-dessous illustre notre modèle proposé.

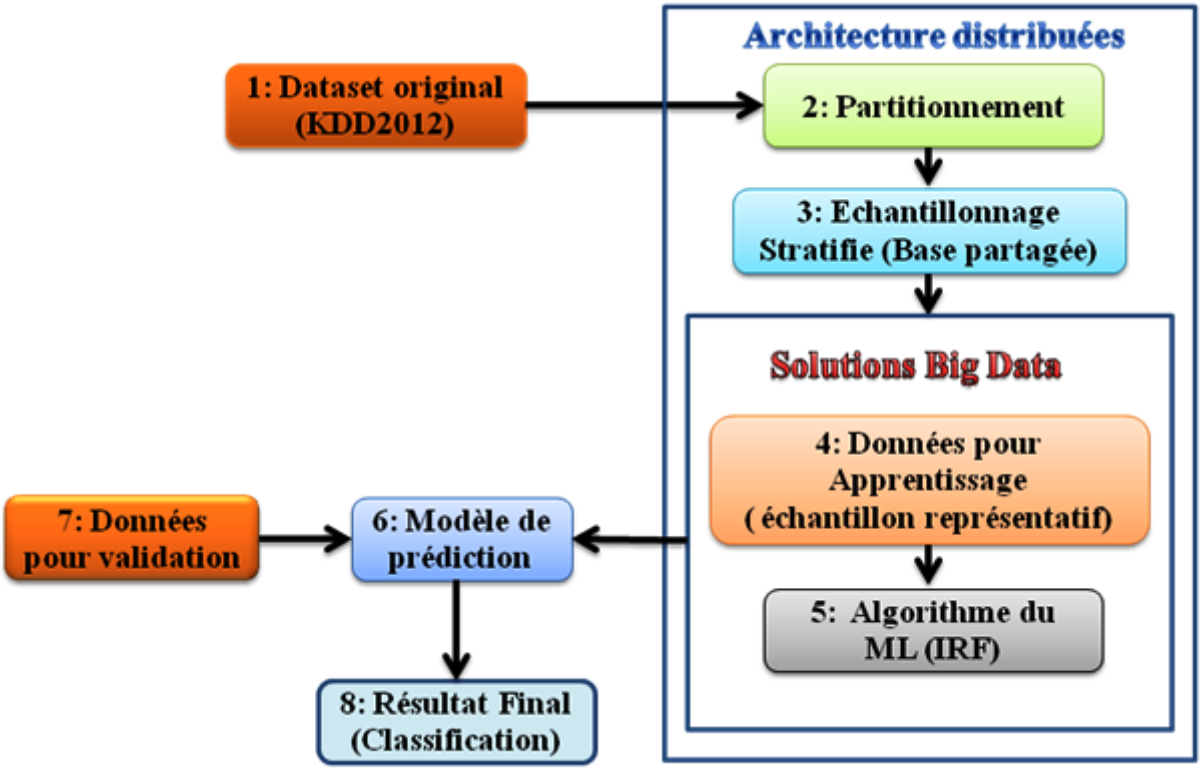

FIGURE 5.1 – *Le modèle proposé.*

Etape 1 : Mise en place le Big DataSet (KDD2012) dans notre système (architecture) distribué, exactement dans le nœud central, ce système proposé s'appuie sur des solutions du Big Data afin de traiter ce volume

de données. Le scénario de fonctionnement (organisation physique) de notre système distribué est bien détaillé dans la figure ci-dessous :

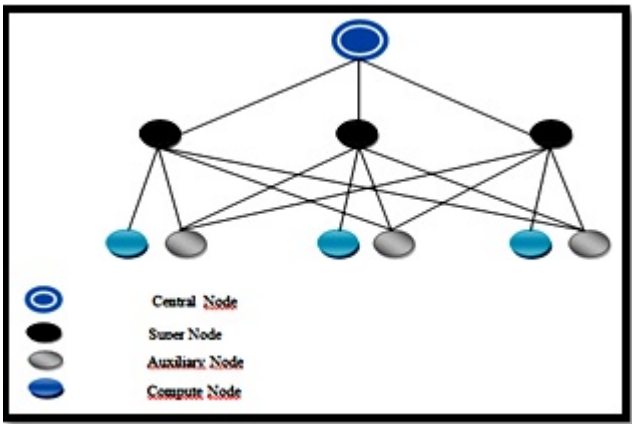

FIGURE 5.2 – *L'Architecture (organisation physique) proposée.*

Notre système (architecture) distribué est caractérisé par :

— L'installation du Apache Spark dans chaque nœud, de sorte que Spark Master soit pour les supers nœuds et Spark Worker soit pour les nœuds de calcul et les nœuds auxiliaires,

— L'implémentation d'un web service nommé « MRS-*Manager to Run Sample-* » est pour la gestion des échantillons partagés inter-clusters entre les supers nœuds et les nœuds auxiliaires,

— La méthode *Map-Reduce* est pour la partition de la base d'apprentissage globale en des bases d'apprentissage partielles distribuées aux différents supers nœuds, En outre, elle collecte les résultats de la classification à partir de ces nœuds et les envoyer au nœud central,

— La configuration de l'*Apache Zookeeper*, il permet la coordination entre le web service « MRS » et *Map-Reduce*, parce que le rôle d'*Apache Zookeeper* est de synchroniser le calcul parallèle entre les super-nœuds et les nœuds auxiliaires lors de l'analyse prédictive, car notre modèle proposé nécessite à la fois la base d'apprentissage partielle et la base d'apprentissage partagée.

Etape 2 : Partitionner le Big DataSet (KDD2012) entre les nœuds de calcul de manière simple, sachant que : la base d'apprentissage partielle = dataset original /nb nœuds de calcul.

Etape 3 : Extraire la base d'apprentissage partagée en utilisant l'échantillonnage stratifie, cet échantillon est constitué de toutes les bases partielles, sachant que la taille d'apprentissage partagée = 1/nb nœuds de calcul. Après cela, on supprime les individus dupliqués.

Etape 4 : Extraire la base d'apprentissage représentative en combinant la base d'apprentissage partielle avec l'échantillon partagé (voir le détail dans la section 5.3.2).

Etape 5 : La base d'apprentissage représentative est fournie à l'algorithme de classification Improved Random Forests (notre contribution, voir le détail dans la section 5.3.4). Ce classifieur (IRF) permet de déterminer les relations cachées entre divers éléments et résultats afin d'extraire le modèle de prédiction.

Etape 6 : En effet, l'algorithme Improved Random Forests va produire un modèle de prédiction, ce modèle se basera sur la base d'apprentissage représentative, il va capturer toutes propriétés et corrélations présentes dans la phase d'apprentissage. C'est pour cela que les données d'apprentissage doivent être assez représentatives du problème métier. Par conséquent, le modèle de prédiction calculé à la fin de la phase d'apprentissage sera plus généralisable et permettra des prédictions précises.

Etape 7 : Les nouveaux individus (données pour validation) sont transmis au modèle de prédiction pour prédire leurs classes. La sortie s'appelle également le résultat de classification.

Etape 8 : Le résultat de la prédiction est fourni avec des métriques de performance (Précision, RMSE, MAE,...) pour évaluer la performance de notre classifieur s'il prédit les classes de nouveaux individus correctement et avec une bonne précision.

Pour mieux comprendre les étapes de notre modèle proposé notamment les étapes : 2, 3 et 4 nous présentons un plan schématique cidessous qui explique en détail le scénario de fonctionnement (organisation logique) de ce modèle.

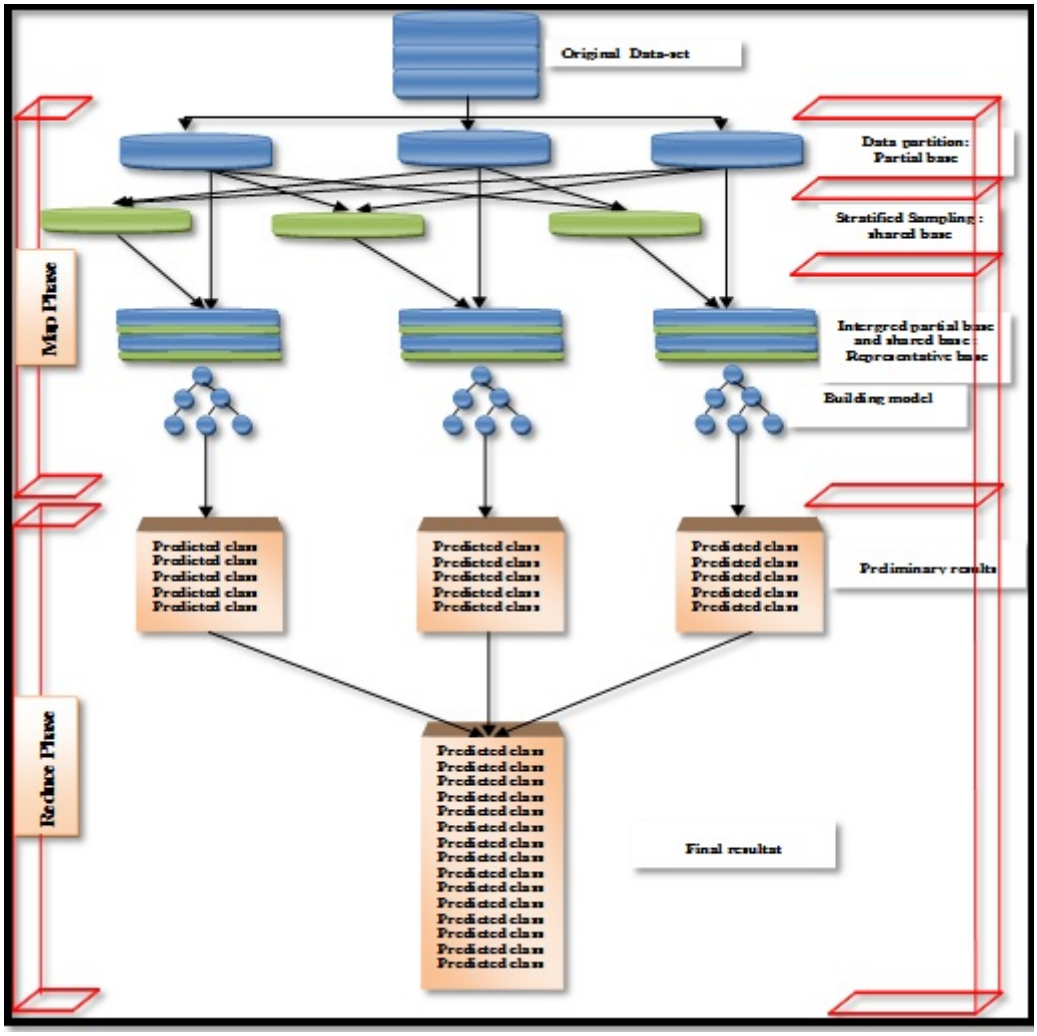

FIGURE 5.3 – *Le scénario de fonctionnement (organisation logique) du modèle proposé*

Les nœuds de notre système (architecture) distribué sont organisés (topologie physique et topologie logique) comme suit :

— **Le nœud central :** Ce nœud est connecté avec tous les supers nœuds, il contient la base d'apprentissage globale, il permet également d'afficher le résultat final de la prédiction.

— **Le super nœud :** (1) calcule l'échantillon e_i de sa partition de données d_i, il envoie ensuite cet échantillon vers tous les nœuds auxiliaires inter-clusters afin d'extraire la base d'apprentissage partagée d_s , tel que $d_{si} = \bigcup_{i=1}^{n} e_i\{d_i\}$, (2) cumule la base d'apprentissage partielle d_p avec la base d'apprentissage partagée d_s pour obtenir la base d'apprentissage représentative d_r, (3) envoie le résultat de prédiction préliminaire qui vient du nœud de calcul vers le nœud central.

— **Le nœud auxiliaire :** (1) chaque nœud auxiliaire est connecté avec tous les supers nœuds (intra et inter-cluster), il assure le calcul de la base d'apprentissage partagée, c-à-d, il cumule tous les échantillons qui viennent des supers nœuds inter-clusters), (2) ensuite il envoie cette base d'apprentissage vers le super nœud intra-cluster (local).

— **le nœud de calcul :** (1) exécute la méthode Random Forests à partir de la base d'apprentissage représentative, (2) ensuite, il envoie le résultat préliminaire vers le super nœud.

Notre modèle proposé est assez simple, mais il est capable de traiter rapidement (**Vélocité**) des données massives (**Volume**) et de fournir de bons résultats (**Véracité**) de l'analyse prédictive. On peut dire que notre modèle facilite à la fois la construction de la base d'apprentissage partagée et de la base d'apprentissage représentative.

5.3.2 Echantillonnage et Map-Reduce

L'échantillonnage est la phase qui consiste à sélectionner les individus que l'on souhaite interroger au sein de la population de base. Les résultats obtenus sur l'échantillon sont ensuite extrapolés à la population que l'on souhaite étudier [Nan *et al.* 2016].

Pour déterminer la taille de l'échantillon T_e , Il existe deux approches [Ataro 1967] :

1. A partir d'une proportion [Cochran 1977], on utilise la formule suivante : $T_e = \frac{n^2 * p * (1-p)}{m_e}$ T_e : taille de l'échantillon attendu, n : niveau de confiance selon la loi normale centrée réduite, p : proportion estimée de la population qui présente la caractéristique (lorsque inconnue, on utilise $n^2 = 95\%$, $p = 50\%$, $m_e = \frac{1}{4}\%$ or $m_e = m^2$ de sorte que $m = 5\%$),m_e :marge d'erreur tolérée.

2. A partir d'une moyenne [Jain *et al.* 2015] on utilise la formule suivante : $T_e = \frac{n^2 * \sigma^2}{m^2}$ T_e : taille de l'échantillon attendu, n : niveau de confiance déduit du taux de confiance, σ : écart type estimé de la moyenne du critère étudié, $m = \sqrt{m_e}$ =5%, m_e : marge d'erreur.

D'après[4] [MacInnis *et al.* 2018] [Espinosa *et al.* 2012] les méthodes d'échantillonnage probabiliste sont meilleures que les méthodes d'échantillonnage non probabiliste. D'après ce que nous avons présenté dans la première partie (section 3.3.2.1), et selon[5] [Espinosa *et al.* 2012][Fei 2015] [Peter 1976][Okororie & Otuonye 2015] [Puech *et al.* 2014], nous confirmons donc que la méthode d'échantillonnage stratifié est la meilleure

4. https ://en.wikipedia.org/wiki/Survey_sampling, (consulté le : 23/11/2017)
5. https ://saylordotorg.github.io/text_principles-of-sociological-inquiry-qualitative-and-quantitative-methods/s10-03-sampling-in-quantitative-resea.html,

parmi les autres méthodes d'échantillonnage probabiliste. Pour cela, dans notre travail nous avons appliqué un échantillonnage aléatoire stratifié afin d'extraire un échantillon représentatif e_i sans remise à partir de d_i de chaque super nœud N_s selon l'algorithme ci-dessous :

1. On sépare la base d'apprentissage partielle d_i de chaque N_s en différentes strates de manière à ce que chaque strate regroupe les individus ayant les mêmes classes prédites.

2. On compte le nombre des individus ayant les mêmes classes prédites de chaque strate $NB(I_{cs})$, ainsi que $NB(d_i)$: le nombre d'individus de la base d'apprentissage partielle d_i.

3. On détermine la proportion de chaque strate du d_i par la formule : $PR_i = \frac{NB(I_{cs})}{NB(d_i)}$.

4. On détermine le nombre d'individus qu'il faut dans chaque strate de l'échantillon représentatif e_i , de sorte que : $NB(e_i) = \sum_{i=1}^{n} PR_i * T_e$.

5. On sélectionne le nombre d'individus voulu dans chaque strate de la base d'apprentissage partielle d_i par échantillonnage aléatoire simple.

Map-Reduce est un paradigme de programmation qui permet de distribuer des traitements parallèles sur des volumétries de données dépassant typiquement 1To, dans un cluster composé de centaines, voire de milliers de nœuds.

Selon [Kaur *et al.* 2017] présente l'algorithme *Map-Reduce* en cinq étapes utilisant 100 nœuds de cloud computing. Il applique cet algorithme pour analyser les données numériques de la météorologie dans les plus brefs délais, pour cela, il traite 200 gigaoctet de données en moins de 200 secondes, ce qui est un peu de temps par rapport à la taille des données utilisées.

Algorithme MapReduce

Input : la base d'apprentissage globale D_g.
Processing :

- Calculer la base d'apprentissage partielle $d_p = \{d_{p1}, d_{p2}, ..., d_{pn}\}$ sachant que les d_{pi} ayant la même taille T_{d_p} (nombre d'instances), de sorte que $T_{d_p} = \frac{D_g(N_c)}{nbr\ N_s}$, N_s :super nœud, N_c : nœud central.
- Placer chaque d_{pi} dans N_{si}.
- Extraire e_i partir de N_{si}.
- Calculer d_s par N_a à l'aide du service web « MRS ».
- Calculer d_r , de sorte que : $d_{ri} = d_{pi} \bigcup d_{si}$.
- Placer d_{ri} dans N_{ki} à partir du N_s. N_k : nœud de calcul.
- Build a local random forest $\mathcal{L}_i$ du N_{ki}.
- Envoyer le résultat préliminaire du $\mathcal{L}_{N_{ki}}$ vers N_{si}.
- Intégrer tous les résultats (les classes prédites) de tous les forêts $\mathcal{L}_{N_k}$ into N_c , de sorte que $\mathcal{L}_{N_c} = \sum_{i=1}^{n} \mathcal{L}_{N_{ki}}$.

Output : Résultat final de la prédiction

5.3.3 Les Forêts

5.3.3.1 Les Forêts dans la nature

Une forêt est une vaste zone dominée par les arbres, en d'autres termes, la forêt est composée d'un grand nombre d'arbres sur une vaste zone. Nous avons tous besoin de forêts saines. Les forêts contribuent à maintenir notre climat stable en absorbant le dioxyde de carbone et en libérant de l'oxygène. Elles régulent notre approvisionnement en eau, et elles améliorent sa qualité. Elles abritent également plus de la moitié des espèces rencontrées sur la terre. Les forêts constituent une riche variété de vies qui permet à de nombreux systèmes naturels de fonctionner. Avec une meilleure protection, les forêts peuvent continuer à accueillir la faune et des ressources pour les populations autochtones et les communautés locales. Et, à l'échelle mondiale, Elles peuvent continuer à nous fournir tous les éléments essentiels tels que de l'air frais et de l'eau potable. Avec une meilleure gestion, nous pouvons augmenter la production de bois sans nuire à l'environnement local. Et avec une meilleure planification, nous pouvons produire suffisamment de nourriture pour les populations en croissance sans avoir à convertir les forêts en terres agricoles. Mais notre idée est inspirée de forêts composées d'arbres fruitiers, ce qui nous permet de voir la qualité des cultures après l'élagage des arbres.

FIGURE 5.4 – *Les forêts dans la nature (Les arbres improductifs et les arbres fruitiers).*

5.3.3.1.1 Elagage des arbres

L'élagage consiste à couper certaines parties d'un arbre ou d'une plante afin d'encourager la croissance et la fécondité [MORRIS 2010]. Dans notre thèse, nous parlerons plus particulièrement sur les arbres fruitiers. Nous devrions élaguer un tiers des branches de ces arbres chaque année. Sachant que l'élagage se déroule en deux phases, la première phase se déroule pendant l'hiver lorsque l'arbre est endormi et qu'il n'y a pas de feuilles. Cette phase est considérée comme une phase ordinaire [Palliotti *et al.* 2017]. Tandis que, la deuxième phase se déroule au printemps [6]. Nous connaissons que cela sonne beaucoup, mais il serait bénéfique pour l'arbre. Nous aiderions cet arbre à concentrer son énergie principalement sur la production de fruits plutôt que sur une croissance inutile, Et bien sûr, pour que l'arbre reste en bonne santé. De plus, les fruits d'un arbre régulièrement élagué seront plus gros, plus juteux, plus sucrés et plus sains.

5.2.3.1.1.1 Elagage d'hiver

La question posée : pourquoi devrions-nous élaguer les arbres en hiver ? L'élagage en hiver permet à l'arboriste de repérer plus facilement les défauts de l'arbre, ainsi que de visualiser toute la structure de l'arbre sans obstruction des feuilles. La gestion des défauts des arbres et le maintien d'une structure appropriée sont essentiels à la santé de l'arbre. Notez que, l'élagage d'arbre peut créer une plaie. Ces plaies commencent à se refermer quelques heures après l'élagage. En été, les insectes ont amplement le temps de visiter la plaie ouverte et de transmettre des maladies. Mais en hiver, les insectes et les maladies sont inactifs, il sera donc facile d'éliminer tout problème de transmission des parasites et des maladies [Palliotti *et al.* 2017] [Garcia *et al.* 2017].

6. https ://orchardpeople.com/when-to-prune-fruit-trees/, (consulté le 14/05/2019)

FIGURE 5.5 – *Une image montrant l'élagage d'hiver.*

Mais l'élagage en hiver se présente généralement dans les cas suivants : Si notre arbre a des branches mortes, alors nous les coupons à la base. Ainsi que, si notre arbre est malade, alors nous considérons l'élagage comme une période d'évaluation annuelle. C'est un bon moment pour examiner notre arbre de près. Si nous voyons des branches malades, alors nous les retirons. En plus, nous gardons notre arbre dans une taille confortable à gérer. Si nous ne voulons pas grimper sur l'échelle, alors nous devons couper les hautes branches de l'arbre, nous voulons également faire de notre mieux pour éliminer les branches croisés. Lorsque le vent souffle, les branches se frottent les unes contre les autres, elles peuvent donc endommager les fruits ou les branches elles-mêmes, ceci affaiblit l'arbre et le rend plus vulnérable aux parasites et aux maladies. De même, nous voulons éliminer les angles aigus, car ils affaiblissent les branches une fois qu'ils portent des fruits, ils risquent donc de se casser. Donc, si nous remarquons une branche en forme de « Y », nous élaguons une des branches supérieures. En fin, nous assurons d'élaguer les zones de ramifications pour favoriser une bonne circulation de l'air.

5.2.3.1.1.2 Elagage de printemps

L'élagage de printemps est effectué après que l'arbre commence à porter des fruits, mais parfois nous effectuons cet élagage en été suivant la nature des arbres tels que les arbres d'orangers, car leurs fruits mûrissent en hiver. L'élagage de printemps (d'été) consiste à couper les branches improductives et à éviter de couper les branches portant des fruits afin d'exploiter l'énergie uniquement dans les branches productives, il implique donc d'augmenter la taille (volume) des fruits, ceci est appelé produit amélioré [7] [8].

7. https ://orchardpeople.com/when-to-prune-fruit-trees/, (consulté le 14/05/2019)
8. https ://gardenerspath.com/how-to/pruning/basics-pruning/, (consulté le 14/05/2019)

FIGURE 5.6 – *Une image montrant l'élagage de printemps.*

L'élagage de printemps est un processus important tout aussi important que l'élagage d'hiver, où il est considéré comme un processus complémentaire de l'élagage d'hiver. Il sert à stimuler la croissance végétative pendant la croissance pour donner un produit de haute qualité tout en préservant les branches restantes de la maturité. Le processus de l'élagage de printemps requiert des compétences en performance, car toute erreur dans ce processus entraîne une diminution de la quantité et de la qualité du produit. L'élimination des branches inutiles est un processus important lors de l'élagage de printemps, cela se fait souvent dans les arbres fruitiers au début du printemps.

5.3.3.2 Les forêts en informatique

Les forêts aléatoires ou *Random Forests* en anglais, cas particulier de *bagging* [9], s'adapte particulièrement bien aux technologies du Big Data lorsqu'une grande taille de l'ensemble d'apprentissage fournit des échantillons indépendants. En outre, *Random Forests* semblent insensible au surapprentissage, cette méthode ne nécessite pas généralement de gros efforts d'optimisation de paramètres. Les forêts aléatoires évitent donc l'un des principaux écueils des approches Big Data en apprentissage automatique [Kuperman & Kazunaga Matsuki 2016] [Rio *et al.* 2014].

Dans notre travail, nous avons utilisé le classificateur *Random Forests* parmi les autres classificateurs d'apprentissage automatique. Il fonctionne bien notamment dans le traitement à grand d'échelle. Pour confirmer pourquoi nous avons choisi ce classifieur ? nous citons nombreux travaux récents [Kaur *et al.* 2017] [Rathore *et al.* 2016] [Dong *et al.* 2016] [Dong *et al.* 2013] [Bei *et al.* 2018] dans ce sens qui montrent la puissance de ce classificateur.

9. Bagging, appelé aussi Bootstrap Aggregation, est un méta-algorithme d'ensemble d'apprentissage automatique conçu pour améliorer la stabilité et la précision des algorithmes d'apprentissage automatique utilisés dans la classification et la régression.
- L'utilisation du bagging est adaptée aux algorithmes à fortes variance qui sont ainsi stabilisés (réseaux neuronaux, arbres de décision pour la classification ou la régression...), mais il peut également dégrader les qualités pour des algos plus stables (k plus proches voisins avec k grand, régression linéaire).

5.3.4 Double élagage (méthode proposée)

Cette méthode est inspirée de la nature par ce que nous appelons le « bio-mimétisme », où nous nous sommes appuyés sur l'élagage d'hiver et l'élagage de printemps. Le but de ce travail est d'améliorer la performance de la méthode supervisée CRF « les forêts aléatoires classiques » afin d'obtenir un bon résultat de l'analyse prédictive des données massives. Alors, CRF -*Classical Random Rorests*- deviendra IRF -*Improved Random Rorests*- , la méthode développée (IRF) est améliorée par une nouvelle méthode d'élagage différente de telle de l'élagage traditionnel, où cette nouvelle méthode proposée repose principalement sur le calcul des probabilités conditionnelles.

Pseudo code Random Forests amélioré

Input : A : Ensemble de données d'apprentissage, N : Nombre d'arbres, T :Arbre, v : Nœud d' arbre, f : Feuille, S : Echantillon bootstrap, $C_r = \{c_1, c_2, \ldots\ldots c_n\}$: Les classes prédites.

Output : E : ensemble d'arbres.

Processus :

Initialiser : $E \leftarrow \emptyset$

// 1er élagage traditionnel

For i=1..N **do**
- $T_i \leftarrow S$
- **If** $v > 1$ then
  - **For each** $v \in T_i$ do
    - -Calculer $\alpha_k = d(T_i(L), v)$
  - **End for**
  - **If** α_k is a minimum then
    - - Sélectionner le nœud v
    - - Remplacer le nœud v par la feuille f
    - **Repeat**
      - -Construire l'arbre T_i' de sorte que $T_i' = \{t_0, t_1, \ldots\ldots t_n\}$
    - **Until** T_i' soit l'arbre optimal
  - **End if**
- **End if**
- **Return :** T_i est remplacé par T_i'

// 2eme élagage (proposé) : le plus important

- **For each** $C_r \in T_i'$
  - -Calculer la probabilité $p(C_r)$, de sorte que : $p(C_r) = \frac{nbr(C_r, S(MC_\phi(T_i', v)))}{s'}$
  - /* $p(C_r)$: Probabilité de mauvaise classification de la classe C_r dans T_i' .
    $nbr(C_r, S(MC_\phi, (T_i', v)))$:le nombre d'individus mal classés par v dans l'arbre T_i' de S appartiennent à la classe C_r .
    $s' \in S$: Le nombre d'individus mal classés par v dans l'arbre T_i' de S.*/
  - **If** $\mu = \max(p(C_r))$
    - -Supprimer cette classe (feuille) de T_i' // élaguer les branches inutiles
  - **End if**
- **End for**
- **Return :** $E \leftarrow E \cup \{T_i'\}$

End for

Pour mieux comprendre l'algorithme : (if $\mu = \max(p(C_r))$), on applique la probabilité conditionnelle (Bayes). Pour cela, prenons l'exemple suivant :

Par exemple, un arbre est constitué de dix classes, il y a une classe a une mauvaise classification C_r parmi ces dix classes , en supposant que cette classe a la probabilité la plus élevée, nous devons alors élaguer cette classe E = "yes", sachant que $P(E/C_r)=1$, $P(\overline{E} / \overline{C_r}) = 1$

$$\underline{\text{Selon la formule de Bayes :}}\ P(C_r/\ E\ ) = \frac{P(C_r \cap E)}{P(E)} = \frac{P(E/C_r)P(C_r)}{P(E/C_r)P(C_r)+P(E/C_r)P(\overline{C_r})}$$

D'où:

$$P(C_r/\ E\ ) = \frac{0.1x1}{1x0.1+0x0.9} = 1 \text{ (élagage possible).}$$

$\underline{\text{Bien au contraire}}$: E = "no", $P(E/C_r)=0$, $P(\overline{E} / \overline{C_r}) = 0$

D'où:

$$P(C_r/\ E\ ) = \frac{0.1x0}{0x0.1+1x0.9} = 0 \text{ (élagage impossible).}$$

5.3.5 Expérimentation

Après la réalisation de l'architecture proposée qui est illustrée dans la figure 5.2, cette architecture est composée d'un nœud central, de trois supers nœuds, de trois nœuds auxiliaires, et de trois nœuds de calcul, de sorte que chaque nœud possède les caractéristiques suivantes : Intel Core i7-3.40 GHz processor, 8 GB RAM, 1TB Hard Disk et 1Gigabit Ehternet. Et comme software, nous avons besoin : Hadoop 2.6.2, Spark version 2.2.0, ZooKeeper version 3.4.10, Java version 1.8.0_101, et le système d'exploitation Ubuntu version 14.04.2 . On peut exécuter 1TB de données sur Spark dans chaque nœud de calcul (principe de RDD du Spark).

Dans notre travail, nous avons fait trois expérimentations pour l'évaluation de performance de la prédiction dans le contexte Big Data.

La première expérimentation concerne le développement des forêts aléatoires classiques.

Dans cette expérimentation, nous avons effectué un seul test (Test1).

Test 1 : On évalue l'algorithme Random Forests amélioré dans un nœud de calcul de notre model proposé. Puis, nous comparons le classifieur Random Forest traditionnel avec Random Forests amélioré (notre proposition), dont chaque Random Forest est composé de 1000 arbres en utilisant trois bases de données de l'UCI Machine Learning Repository [Bache & Lichman 2013], ces bases de données sont décrites dans le tableau 5.1 :

Data set	Instances	Features	Classes
HIGGS (2.1 Gigaoctet)	11,000,000	28	2
kdd2010 (algebra) (274 *Megaoctet*)	8,407,752	20,216,830	2
real-sim (33.6 *Megaoctet*)	72,309	20,958	2

TABLE 5.1 – *Un tableau descriptif des jeux de données utilisés pour évaluer la performances des forêts aléatoires.*

DataSet Higgs : Il s'agit d'un problème de classification qui fait la distinction entre un processus de signal (en avant-plan) qui produit des bosons de Higgs et un processus en arrière-plan qui ne produit pas des bosons de Higgs. Les données ont été produites à l'aide d'une simulation de Monte Carlo. Les 21 premiers attributs (colonnes 2 à 22) sont des propriétés cinétiques mesurées par des détecteurs de particules dans le métronome. Les sept derniers attributs sont des fonctions pour les 21 premiers attributs c.à.d. 21 attributs de bas niveau (quantité de particules) et sept attributs de haut niveau dérivées par physiciens des afin de permettre la distinction entre les deux classes (processus en avant-plan ou processus en arriere-plan) [10].

kdd2010 (algebra) : est une compétition permettant d'extraire des données pédagogiques pour prédire la performance d'un élève face à des problèmes mathématiques (algèbre), à partir d'informations relatives aux performances passées. À la fin de la compétition, le gagnant réel sera déterminé en fonction de ses performances. Cette tâche de prédiction présente non seulement un défi technique pour les chercheurs, mais également une importance pratique ,de sorte que des prédictions précises peuvent être utilisées pour mieux comprendre les performances des étudiants en mathématiques entre les deux classes : Bon ou faible, et pour améliorer les performances des étudiants [11].

real-sim : est un ensemble de données de classification des textes/documents. Il est souvent utilisé pour la classification binaire, où la tâche peut être définie comme la séparation de documents sur des sujets «réels» et «simulés» du corpus SRAA d'articles UseNet [12].

Les résultats du test (1) sont illustrés dans le tableau 5.2 ci-dessous :

10. https ://archive.ics.uci.edu/ml/datasets/HIGGS
11. https ://www.csie.ntu.edu.tw/ cjlin/libsvmtools/datasets/binary.html
12. https ://www.csie.ntu.edu.tw/ cjlin/libsvmtools/datasets/binary.html

Data Set	Classical Random Forest		Improved Random Forest	
HIGGS	**Precision**	0.861788617	**Precision**	0.881818181
	Kappa	0.8438	**Kappa**	0.9019
	MAE	0.0644	**MAE**	0.0432
	RMSE	0.2358	**RMSE**	0.1848
	Time(s)	238	**Time(s)**	221
kdd2010 (algebra)	**Precision**	0.940366972	**Precision**	0.971491228
	Kappa	0.8917	**Kappa**	0.9194
	MAE	0.0496	**MAE**	0.0355
	RMSE	0.2055	**RMSE**	0.1738
	Time(s)	47	**Time(s)**	43
real-sim	**Precision**	0.747685185	**Precision**	0.767220902
	Kappa	0.7128	**Kappa**	0.7327
	MAE	0.1155	**MAE**	0.108
	RMSE	0.3318	**RMSE**	0.3201
	Time(s)	7	**Time(s)**	6

TABLE 5.2 – *Comparaison des résultats de prévision entre CRF et IRF.*

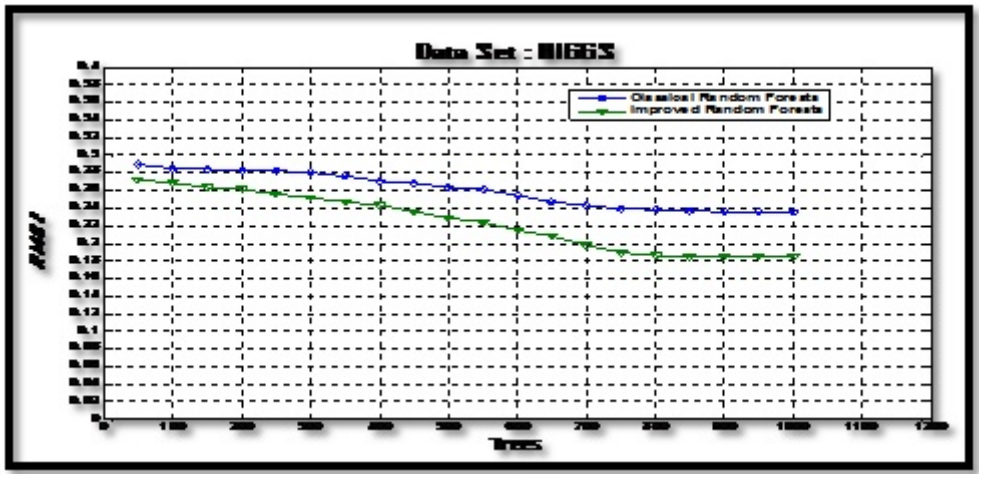

FIGURE 5.7 – *Comparaison du taux d'erreur entre le CRF et l'IRF du dataset :**HIGGS*
.

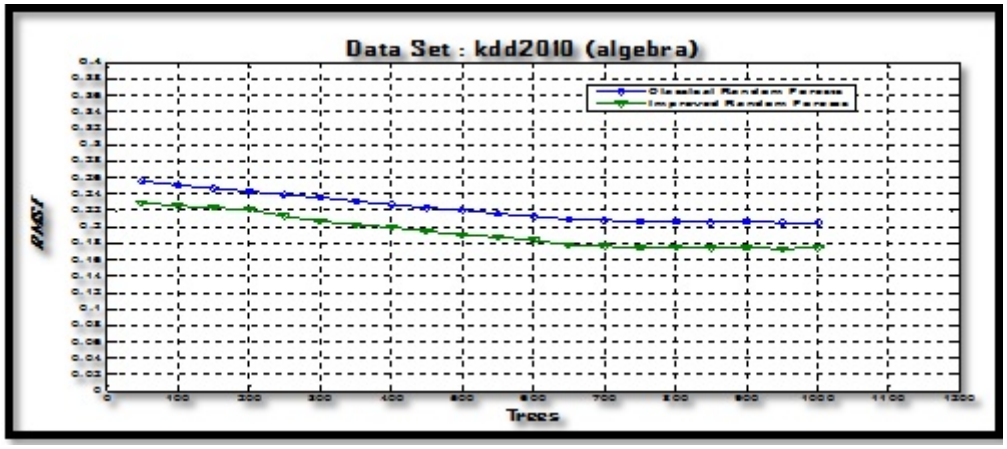

FIGURE 5.8 – *Comparaison du taux d'erreur entre le CRF et l'IRF du dataset :* ***kdd2010****.*

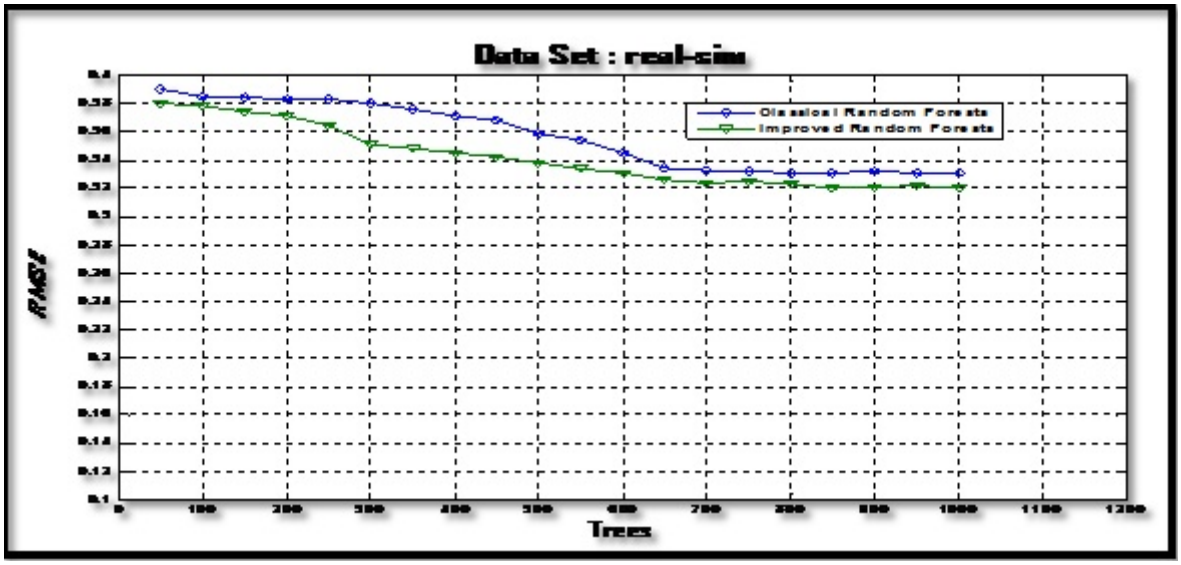

FIGURE 5.9 – *Comparaison du taux d'erreur entre le CRF et l'IRF du dataset :**real-sim*
.

La deuxième expérimentation concerne l'architecture proposée pour l'analyse prédictive de données massives (voir section 5.2.1).

La troisième expérimentation concerne l'abstraction de l'échantillon représentatif (la base d'apprentissage représentative). Dans cette expérimentation , nous avons effectué trois tests (Test2, Test3 et Test4).

Test 2 : présente l'analyse prédictive du Big Data sans l'utilisation de la base d'apprentissage partagée.

Test 3 : présente l'analyse prédictive du Big Data avec l'utilisation de la base d'apprentissage partagée, mais cette base est extraite de la base d'apprentissage globale qui se trouve dans le nœud central.

Test 4 : présente l'analyse prédictive du Big Data avec l'utilisation la base d'apprentissage représentative, de sorte que cette base d'apprentissage est extraite à partir de la base d'apprentissage partagée et la base d'apprentissage partielle, sachant que la base d'apprentissage représentative est sauvegardée sur les supers nœuds (notre modèle proposé).

Afin de présenter les trois tests (2,3 et 4), on utilise le data-set kdd2012 sauvegardé sous format LIBSVM [13], kdd2012 est décrit dans le tableau 5.3 comme suit :

Dataset	Instances training	Instances validation	Features	Classes
KDD2012 (1.95 Go)	119705032 (1.60 Go)	29934073 (458.26 Mo)	54686452	2

TABLE 5.3 – *Un tableau descriptif du dataset (KDD2012) utilisé pour l'analyse prédictive.*

13. http ://www.csie.ntu.edu.tw/ cjlin/libsvmtools/datasets/binary.htmlkdd2012

KDD Cup 2012 : est une base de données pour objectif de prédire le taux de clics sur des pages Web publicitaires contenant des informations sur la publicité, la requête et l'utilisateur, car le modèle économique derrière la publicité nécessite des informations introduite de la part des clients pour évaluer les annonces. Dans le fichier donné, chaque ligne indique un vecteur de caractéristiques sur les pages publicitaires et le nombre d'apparition avec click (consulter la page) ou non click (fermer la page). Avec sous ces valeurs de caractéristiques, nous définissons la classe prédite comme étant positive si le nombre de clics est non nul et négative dans le cas contraire.

Les résultats du test(2), test(3) et test(4) sont illustrés dans le tab-5.4 :

Data set	Clusters	Big Data predictive analytics: test 2		Big Data predictive analytics:test 3		Big Data predictive analytics:test 4	
Subset 1	Cluster 1	Correctly classified instances %	93.6599 %	Correctly classified instances %	95.053 %	Correctly classified instances %	96.9106 %
		Incorrectly classified instances %	6.3401 %	Incorrectly classified instances %	4.947 %	Incorrectly classified instances %	3.0894 %
		RMSE	0.2022	RMSE	0.1768	RMSE	0.1395
		MAE	0.0552	MAE	0.0446	MAE	0.0333
		Kappa	0.9047	Kappa	0.9256	Kappa	0.9537
		Precision	0.842857142857142	Precision	0.8727272727272727	Precision	0.9128440366972477
		Total time	88.16 s	Total time	88.11s	Total time	88.54s
Subset 2	Cluster 2	Correctly classified instances %	93.4884 %	Correctly classified instances %	95.1389 %	Correctly classified instances %	96.8543 %
		Incorrectly classified instances %	6.5116 %	Incorrectly classified instances %	4.8611 %	Incorrectly classified instances %	3.1457 %
		RMSE	0.2003	RMSE	0.1753	RMSE	0.1408
		MAE	0.0545	MAE	0.0439	MAE	0.0338
		Kappa	0.902	Kappa	0.9268	Kappa	0.9528
		Precision	0.8418079096045	Precision	0.8755555555555555	Precision	0.9120370370370371
		Total time	87.53s	Total time	88.08s	Total time	88.51s
Subset 3	Cluster 3	Correctly classified instances %	94.2623 %	Correctly classified instances %	95.2462 %	Correctly classified instances %	96.9937 %
		Incorrectly classified instances %	5.7377 %	Incorrectly classified instances %	4.7538 %	Incorrectly classified instances %	3.0063 %
		RMSE	0.1895	RMSE	0.1738	RMSE	0.1377
		MAE	0.0499	MAE	0.0434	MAE	0.0325
		Kappa	0.9136	Kappa	0.9282	Kappa	0.9549
		Precision	0.8486486486486	Precision	0.8782608695652174	Precision	0.915929203539823
		Total time	88.49s	Total time	88.53s	Total time	88.57s

TABLE 5.4 – *Un tableau descriptif du dataset (KDD2012) utilisé pour l'analyse prédictive.*

Sachant que chaque sous-ensemble (Subset) = (la base d'apprentissage représentative) + (la base de validation) = 1.114 Gigaoctet, de sorte que : la base d'apprentissage représentative = (la base d'apprentissage partielle (1/3 du data-set original) = 546,133 Mégaoctets) + (la base d'apprentissage partagée=137 mégaoctets)=683,133 Mégaoctets, ainsi que la base de validation=458,26 Mégaoctets.

Le tableau 5.5 ci-dessous présente les résultats de l'analyse prédictive de chaque subset (nous utilisons le jeu de données kdd2012) pour chaque cluster en utilisant la base d'apprentissage représentative, sachant que la base de validation est identique et répétée (nous utilisons la même base de validation = 458,26 Mégaoctets) dans tous les sous-ensembles (subset) afin de savoir que la base d'apprentissage représentative est similaire dans tous les clusters. Ainsi que, nous vérifions que cette base d'apprentissage donne des résultats égaux de la prédiction.

Data set	Clusters	Classical Random Forest:test 4		Improved Random Forest:test 4	
Subset 1	Cluster 1	Correctly classified instances %	94.8598 %	Correctly classified instances %	96.9106 %
		Incorrectly classified instances %	5.1402 %	Incorrectly classified instances %	3.0894 %
		RMSE	0.1938	RMSE	0.1395
		MAE	0.0493	MAE	0.0333
		Kappa	0.9211	Kappa	0.9537
		Precision	0.8823529411764706	Precision	0.9128440366972477
		Total time	92.47s	Total time	88.54s
Subset 2	Cluster 2	Correctly classified instances %	94.9458 %	Correctly classified instances %	96.8543 %
		Incorrectly classified instances %	5.0542 %	Incorrectly classified instances %	3.1457 %
		RMSE	0.1904	RMSE	0.1408
		MAE	0.0465	MAE	0.0338
		Kappa	0.9201	Kappa	0.9528
		Precision	0.8829787234042553	Precision	0.9120370370370371
		Total time	92.44s	Total time	88.51s
Subset 3	Cluster 3	Correctly classified instances %	94.9367 %	Correctly classified instances %	96.9937 %
		Incorrectly classified instances %	5.0633 %	Incorrectly classified instances %	3.0063 %
		RMSE	0.1875	RMSE	0.1377
		MAE	0.0442	MAE	0.0325
		Kappa	0.9012	Kappa	0.9549
		Precision	0.8828828828828829	Precision	0.915929203539823
		Total time	92.49s	Total time	88.57s

TABLE 5.5 – *Un tableau descriptif de l'ensemble de données (KDD2012) utilisé pour l'analyse prédictive.*

Nous concluons dans ce cas que nous pouvons diviser la base de validation (si elle est plus grandes) en sous bases de validation comme suit : *sous base de validation*=$\frac{\text{la base de validation}}{nbr\ du\ clusters}$, alors :

sous base de validation = $\frac{458.26}{3}$ =152.753 Megaoctet. Donc, la taille de tous les sous-ensembles de chaque cluster est : 683.133+152.753=835.886 Mégaoctets.

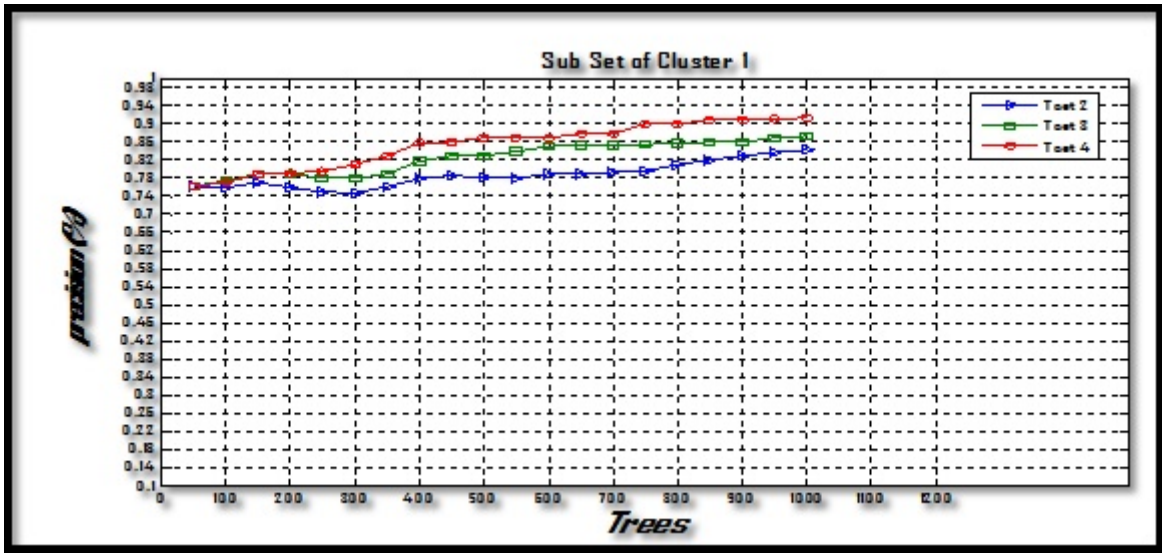

FIGURE 5.10 – *Prédiction préliminaire du cluster 1.*

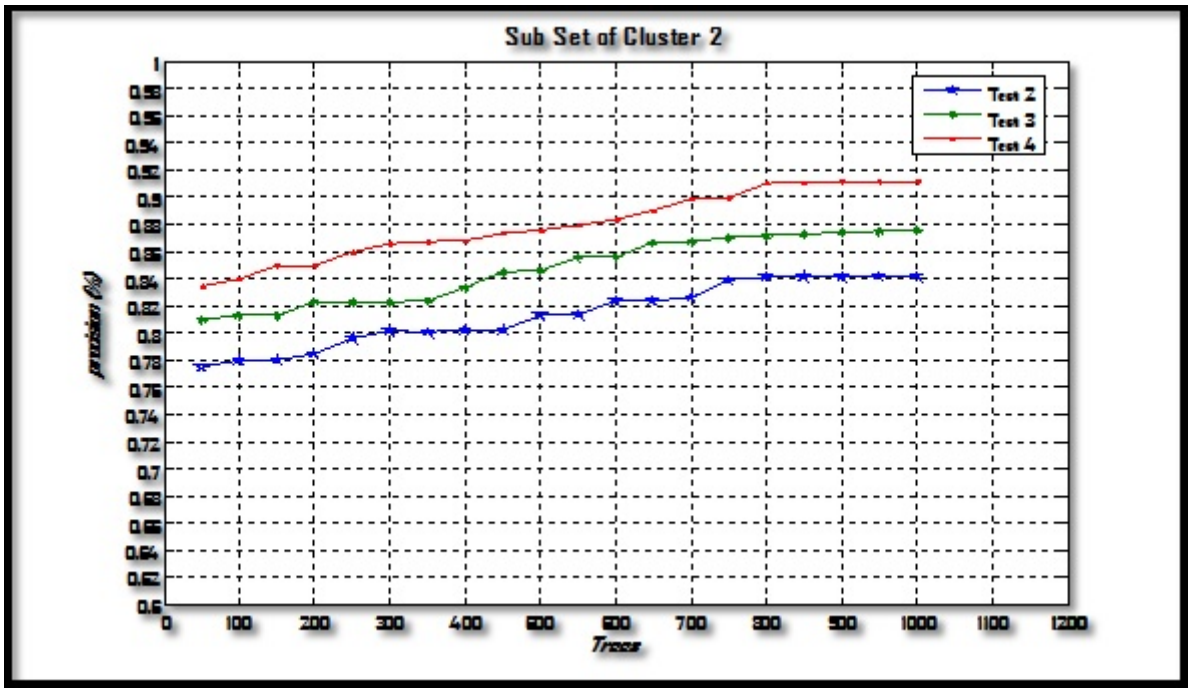

FIGURE 5.11 – *Prédiction préliminaire du cluster 2.*

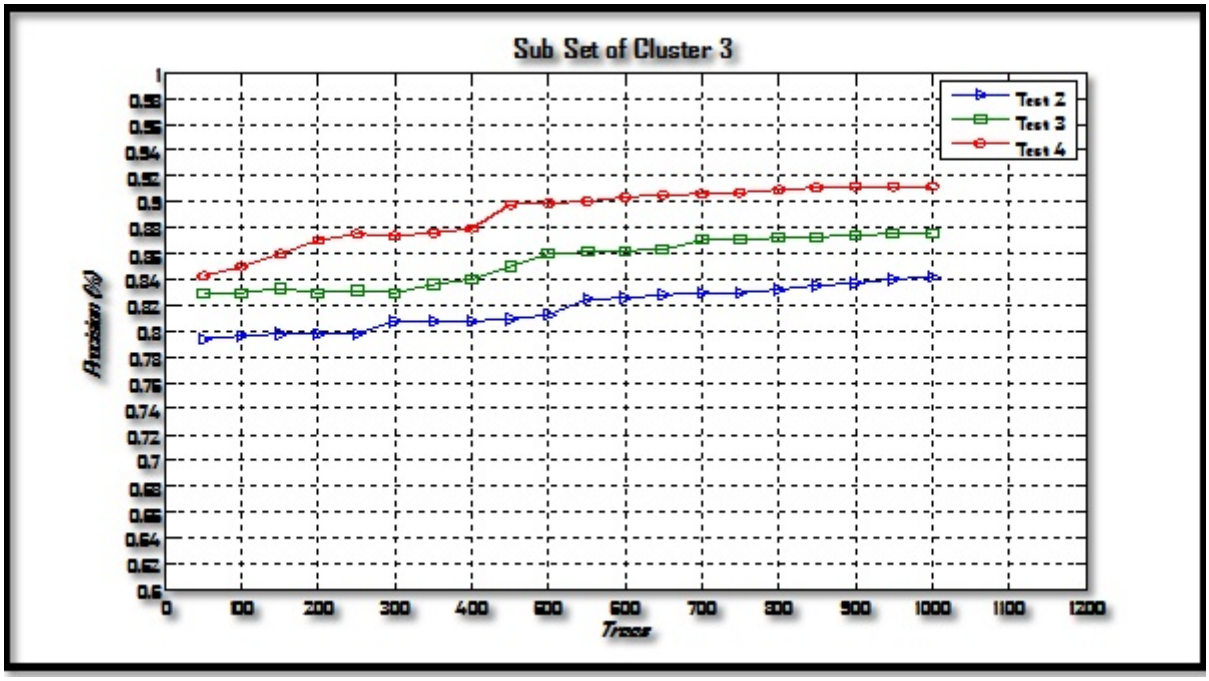

FIGURE 5.12 – *Prédiction préliminaire du cluster 3.*

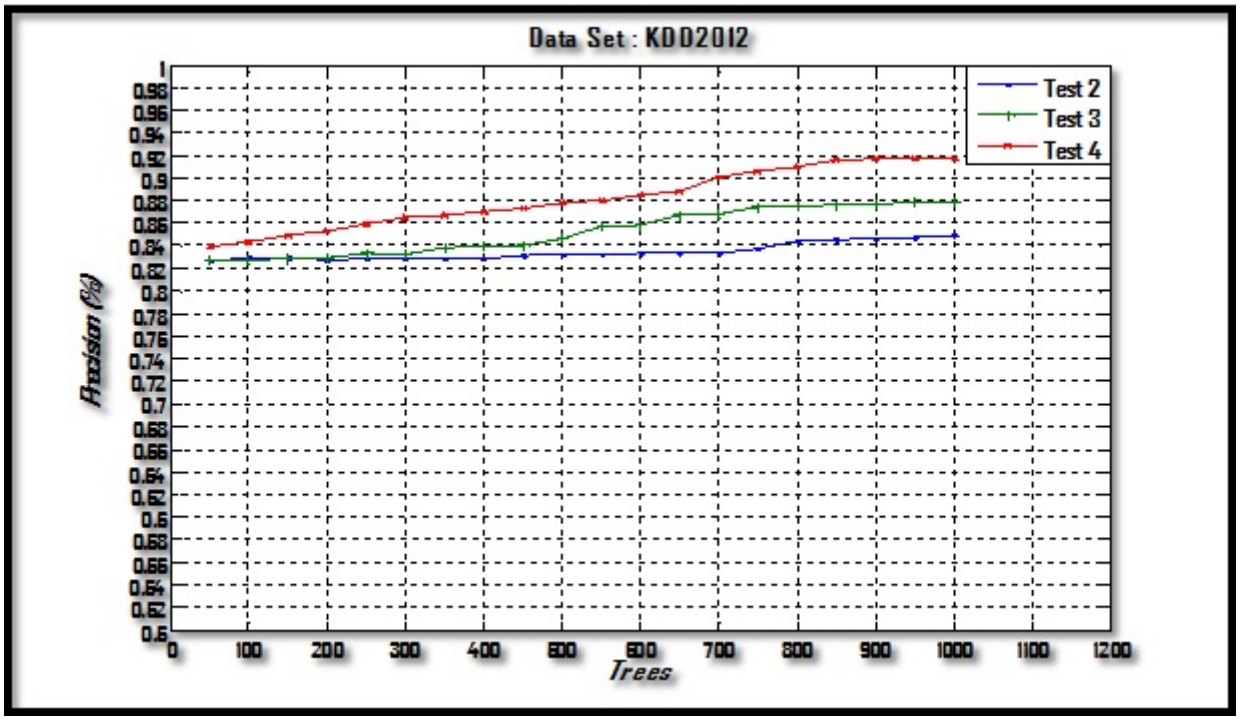

FIGURE 5.13 – *Prédiction final du Dataset orignal.*

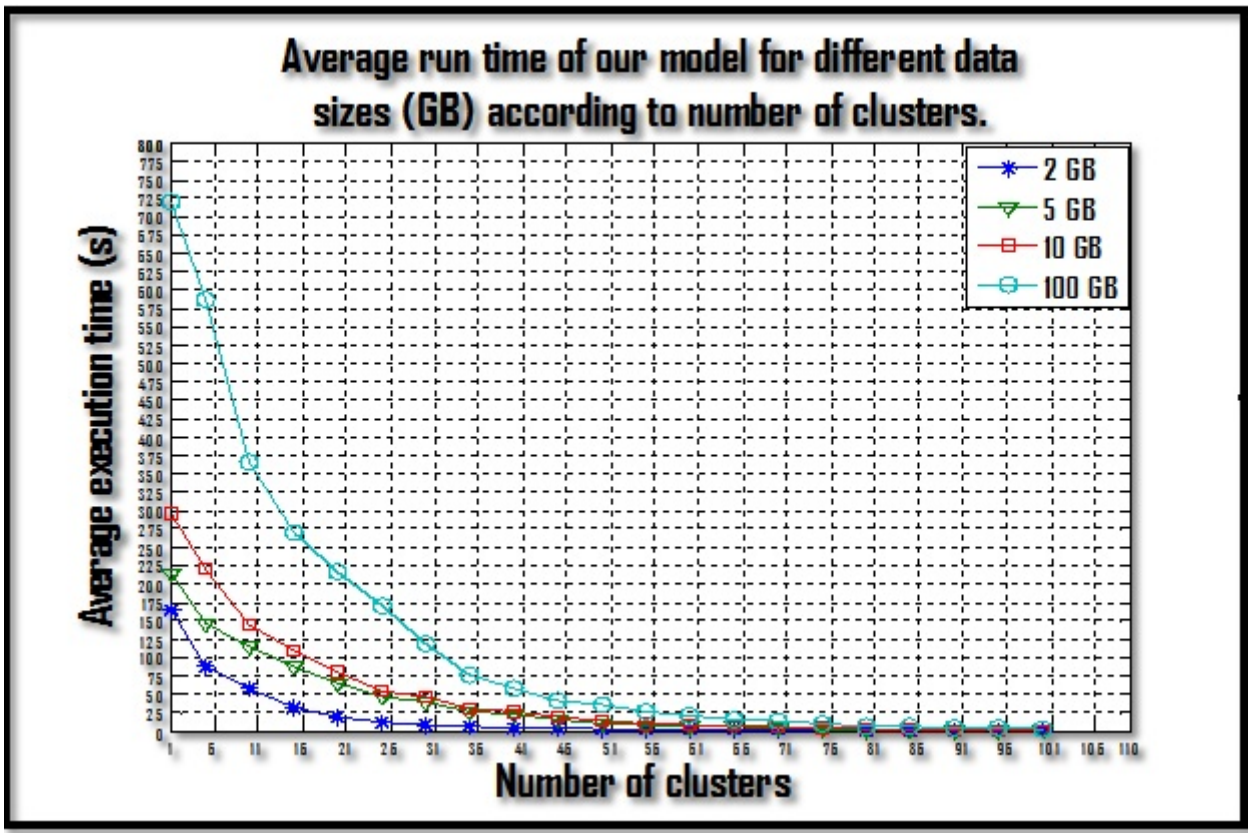

FIGURE 5.14 – *Durée moyenne d'exécution de notre modèle pour différentes tailles de données (Gigaoctet) en fonction du nombre de clusters.*

5.3.6 Discussion des résultats

Test 1 : La forêt aléatoire améliorée conserve son efficacité afin de fonctionner de manière stable quelle que soit la taille des données utilisées (voir tableau 5.2, tableau 5.5).

Test 2 : Ce travail est similaire au travail effectue par [Rio *et al.* 2014][Madhusudanan *et al.* 2017],mais d'après le premier test du tableau 4.4, nous avons obtenu des résultats insatisfaisants de l'analyse prédictive du Big Data en utilisant de la méthode *MapReduce*.

Test 3 : On trouve plusieurs travaux dans ce sens. Par exemple [Murata *et al.* 2015] utilise les données partagées extraites à partir de la base d'apprentissage globale *-original data-set-* et distribuées uniformément sur tous nœuds de calcul. Cette technique donne de bons résultats de l'analyse prédictive du Big Data. Mais, elle reste toujours un peu faible avec notre proposition.

Test 4 : La base d'apprentissage représentative obtenue par l'intégration de deux bases d'apprentissage partielle et partagée présente une excellente représentation du data-set original. Il donne donc de très bons résultats de l'analyse prédictive du Big Data.

Par ailleurs, nous n'oublions pas que ces résultats obtenus ont été soutenus par la méthode d'apprentissage supervisé Random Forests amélioré (IRF) qui a joué un rôle principal dans ce contexte.

Pour approfondir la compréhension, nous expliquons en détail les résultats obtenus comme suit :

Les résultats obtenus dans les figures 5.7, 5.8 et 5.9 montrent que le taux d'erreur dans différents ensembles de données s'est révélé être faible (entre 1% et 3%), cela montre que le double élagage du *Random Forests* est plus efficace pour donner de bons résultats de prévision.

Les résultats obtenus dans le tableau 5.4 et illustrés dans les figures 5.10, 5.11, 5.12 et 5.13 ci-dessus démontrent que la base d'apprentissage représentative (notre proposition) donne de très bons résultats de classification avec un temps d'exécution en mode micro-batch (voir figure 5.14). Mais il est possible d'améliorer ce mode de traitement de données en mode streaming en ajoutant simplement les clusters (voir figure 5.14), de sorte que chaque cluster soit composé d'un super-nœud, d'un nœud auxiliaire et d'un nœud de calcul.

A partir de figure 5.10, figure 5.11 et figure 5.12, nous concluons deux conclusions : la première conclusion est que le test 4 donne un meilleur résultat que les deux tests : test 2 et test 3 grâce à l'IRF (amélioration par double élagage), et la seconde conclusion, nous notons que les résultats de prédiction pour chaque sous-ensemble sont presque égaux grâce à l'échantillon représentatif. En outre, comme le montre la figure 5.13, le résultat final de la prédiction pour l'ensemble de données d'origine (KDD2012 d'une taille de 2 gigaoctet gérés dans un seul cluster) est représenté par le test 4 = ,917380683244 avec un temps de calcul égal à 171 s.

De plus, la moyenne de prédiction du test 4 obtenue à partir des trois sous-ensembles est la suivante : 0,9136034257 avec un temps d'exécution moyen = 88,54 s En effet, la différence de précision est très faible (0,0037772575) par rapport à la taille des données utilisée.

Les résultats obtenus dans le tableau 5.5 montrent que le taux d'erreur de l'analyse prédictive des données massives entre différents clusters est d'environ 1 <924> à 3 <924>. Cela montre que la base d'apprentissage représentative est similaire dans tous les clusters grâce à la technique proposée.

Maintenant, nous pouvons dire que notre travail proposé conserve le résultat de la prédiction dans les systèmes distribués ; c.à.d, ce résultat est également garanti dans le calcul distribué et parallèle (division de la base d'apprentissage entre différents nœuds de calcul lors de la prédiction) quel que soit le volume de données traité.

Conclusion

L'analyse de données massives pour la prédiction est un moyen pour prévoir les probabilités futures avec un niveau de fiabilité acceptable afin de prendre des décisions efficaces dans le plus bref délai, dans ce chapitre, nous avons présenté la méthode Random Forests amélioré par double élagage (IRF), cette technique permet d'augmenter la performance et la précision de classification de données massives, et nous avons également proposé une architecture distribuée supportées par les technologies du Big Data. Cette architecture nous aide à extraire la base d'apprentissage partagée, ainsi que l'échantillon représentatif en utilisant la méthode d'échantillonnage aléatoire stratifié. De plus, cette architecture permet de communiquer les résultats de l'analyse prédictive de données massives dans quelques seconds.

Mes contributions scienfifiques

— Djafri laouni, Amar bensaber Djamel and Adjoudj Reda, BIG DATA ANALYTICS FOR PREDICTION : parallel processing of the big learning base with the possibility of improving the final result of the prediction, Information discovery and delivery, ISSN : 2398-6247 vol 46,issue 3,2018.

Conclusion générale

Dans le domaine du Big Data, les données sont automatiquement enregistrées via un support de traitement numérique avec une propriété hétérogène créée à partir de différentes sources de données ou dans différents contextes. Les algorithmes du Machine Learning peuvent permettre d'extraire des modèles dans l'ensemble de données. Au fil des ans, ils ont eu un grand succès en raison de leur capacité à traiter les problèmes liés à l'analyse prédictive. Mais le problème, ils ne sont pas capables pour l'analyse de données volumineuses. Dans certains cas, , la simple logique est de savoir quand un modèle prédictif ne sera pas suffisant pour résoudre un problème particulier (dans le domain du Big Data), notamment si tous les cas sont traités rapidement et efficacement. Nous avons également montré que le domaine de l'analyse de données massives doit utiliser des algorithmes robustes d'apprentissage automatique et des solutions du Big Data pour développer de bons modèles prédictifs pour les utilisateurs.

Dans cette thèse, nous avons exploré comment utiliser des algorithmes d'apprentissage automatique dans des systèmes distribués pour l'analyse de données volumineuses. Nous n'oublions pas également les méthodes statistiques qui ont beaucoup contribué à cette évolution.

Parfois, les algorithmes d'apprentissage automatique ne peuvent pas être efficaces dans le traitement de données massives, car ils sont souvent confrontés au problème de sur-apprentissage, mais notre modèle proposé a prouvé son efficacité dans la classification des données massives. En outre, les techniques d'échantillonnage statistique constituent un élément important de toute recherche quantitative visant à généraliser les résultats de l'étude à une grande population.Il est nécessaire d'obtenir la taille d'échantillon requise et de sélectionner un échantillon représentatif à l'aide des techniques d'échantillonnage appropriées. Les expériences que nous avons menées dans cette thèse ont montré que l'échantillon représentatif (notre contribution) nous nous donnait de bons résultats de classification en utilisant le classifieur Improved Random Forests. En plus, n'oubliez pas que le système distribué (notre architecture) pris en charge par les solutions du Big Data a contribué pour accélérer le temps d'exécution.

En générale, nous pouvons dire maintenant que notre modèle proposé [Djafri *et al.* 2018] pour l'analyse prédictive des données massives contribue positivement pour résoudre les problèmes liés au Big Data notamment le volume, la véracité et la vélocité. Mais, notre modèle proposé est influencé négativement lorsque le nombre de clusters est très grand (des milliers de clusters), car chaque cluster représente une partition de

données, cela implique que la taille de la base d'apprentissage représentative devient un peu volumineuse. Dans ce cas, notre modèle fonctionne à pleine capacité, ce qui entraîne un déficit de traitement au niveau du nœud de calcul. Certaines personnes se posent la question suivante : pourquoi ne devrions-nous pas augmenter le nombre de nœuds de calcul dans chaque cluster ? La réponse est tout simplement que cette technique nous ramène à la manière classique. Ensuite, comme solution, nous extrayons un nouvel échantillon de taille n en utilisant les techniques de bootstrap [14] à partir de la base d'apprentissage globale, cette solution est similaire au travail effectué par [Kleiner *et al.* 2014], de sorte que il extrait cet échantillon une fois, puis il applique l'apprentissage. Tandis que, L'échantillon boostrap est extrait plusieurs fois afin de construire la base d'apprentissage représentative. L'échantillonnage bootstrap est appliqué afin de préserver parfaitement la précision du résultat global (éliminer la très petite différence de précision).

La survie du Big Data sera un grand défi à l'avenir, car nous vivons dans une ère numérique et, partout dans le monde, quelqu'un veux trouver une nouvelle technologie ou quelque chose pour rénover ce monde. Par conséquent, de nombreuses technologies ou de nombreuses tendances sont présentées chaque jour. Il est clair que le Big Data deviendra plus vitail dans les années à venir. L'analyse prédictive se passe en une analyse prescriptive qui devrait être plus courante et comprise par de nombreuses entreprises. L'obtention de la valeur métier à partir de données nécessite une action rapide sur les événements en temps réel, sinon cette valeur disparaîtra. De plus, prendre des mesures intelligentes rapidement exige plus qu'une simple prédiction ; il faut savoir exactement quoi faire et quand le faire. Grâce aux solutions Big Data et à l'infrastructure informatique distribuée omniprésente, nous prévoyons que la prochaine génération d'analyses du Big Data exploitera des fonctionnalités avancées telles que le cloud computing, les algorithmes d'apprentissage automatique avancés (Deep Learnig), ou utilisera d'autres solutions du Big Data telles que Storm et Solr.

Plusieurs pistes de recherche s'ouvrent et méritent d'être explorées, dont on peut citer :

— Le boom de l'apprentissage automatique va changer le jeu. L'apprentissage automatique qui joue un rôle important dans l'analyse de données massives devrait prospérer dans un proche avenir.

14. -En statistiques, les techniques de bootstrap sont des méthodes d'inférence statistique basées sur la réplication multiple des données à partir du jeu de données étudié, selon les techniques de rééchantillonnage.
-Précisément, et c'est le sens du terme « rééchantillonnage », un bootstrap consiste à créer des « nouveaux échantillons » statistiques, mais uniquement par tirage avec remise, à partir de l'échantillon initial.

— Self-service : La demande d'experts en données sera assez chère. À mesure que les volumes de données continueront de croître, l'écart entre les besoins et la disponibilité des experts augmentera considérablement. Dans ce cas, nous devons donc compter sur le Self-service.

— La confidentialité restera un sujet brûlant, car la croissance rapide des volumes de données crée des difficultés supplémentaires en matière de protection des intrusions et des cyber attaques, parce que les niveaux de protection des données ne suivent pas le rythme de croissance des données.

— Le terme Big Data pourrait être remplacé à l'avenir par deux nouveaux termes : Les données rapides et les données exécutables. Certains experts affirment que le Big Data est mort et obsolète, et que des données rapides le remplaceront bientôt.

Annexes

A

Sommaire

A.1 Quelques concepts en statistiques mathématiques

A.1.1 Échantillons

Lors de l'étude statistique d'ensembles d'informations, la façon de sélectionner l'échantillon est aussi importante que la manière de l'analyser. Il faut que l'échantillon soit représentatif de la population (nous ne faisons pas nécessairement référence à des populations humaines !). Pour cela, l'échantillonnage aléatoire est le meilleur moyen d'y parvenir.

Le statisticien part toujours de l'observation d'un ensemble fini d'éléments, que nous qualifions de "population". Les éléments observés, en nombre n, sont tous de même nature, mais cette nature peut être fort différente d'une population à l'autre.

A.1.1.1 Définitions

Nous sommes en présence d'un "caractère quantitatif" lorsque chaque élément observé fait explicitement l'objet d'une même mesure. A un caractère quantitatif donné, nous associons une "variable quantitative" continue ou discrète qui synthétise toutes les valeurs possibles que la mesure considérée est susceptible de prendre (ce type d'information étant représenté par des courbes de Gauss, de Bêta, de Poisson, etc.)

Nous sommes en présence d'un "caractère qualitatif" lorsque chaque élément observé fait explicitement l'objet d'un rattachement unique à une "modalité" choisie dans un ensemble de modalités exclusives (de type : homme | femme) permettant de classer tous les éléments de l'ensemble étudié selon un certain point de vue (ce type d'information étant représenté par des diagrammes à barre, fromages, diagrammes à bulles, etc.). L'ensemble des modalités d'un caractère peut être établi à priori avant l'enquête (une liste, une nomenclature, un code) ou après enquête. Une population étudiée peut être représentée par un caractère mixte, ou ensemble de modalités tel que genre, tranche salariale, tranche d'âge, nombre d'enfants, situation matrinomaile par exemple pour un individu.

Un "échantillon aléatoire" est un échantillon tiré au hasard dans lequel tous les individus d'une population ont la même chance, ou "équiprobabilité" (et nous insistons sur le fait que cette probabilité doit être égale), de se retrouver dans l'échantillon.

Dans le cas contraire d'un échantillon dont les éléments n'ont pas été pris au hasard, nous disons alors que l'échantillon est "biaisé" (dans le cas inverse nous disons qu'il est "non-biaisé").

Un petit échantillon représentatif est, de loin, préférable à un grand échantillon biaisé. Mais lorsque la taille des échantillons utilisés est petite, le hasard peut donner un résultat moins bon que celui qui est biaisé.

A.1.2 La Moyenne d'une série statistique

Une série statistique se représente souvent dans un tableau :

Valeurs	x_1	x_2	.....	x_p	Total
Effectifs	n_1	n_2	.....	n_p	N
fréquence	f_1	f_2	.....	f_p	1

- La **moyenne** d'une série statistique est le réel noté $\overline{x}$ tel que :

$$\overline{x} = \frac{n_1x_1 + n_2x_2 + \ldots\ldots\ldots\cdots + n_px_p}{N}$$

On note :

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{p} n_ix_i$$

Remarque :
On a aussi :

$$\overline{x} = f_1x_1 + f_2x_2 + \ldots\ldots\ldots\cdots + f_px_p = \sum_{i=1}^{p} f_ix_i$$

A.1.3 La variance

-La **variance** d'une série statistique est le nombre noté V tel que :

$$V = \frac{n_1(x_1 - \overline{x})^2 + n_2(x_2 - \overline{x})^2 + \ldots\ldots\ldots\cdots + n_p(x_p - \overline{x})^2}{N}$$

On note :

$$V = \frac{1}{N}\sum_{i=1}^{p} n_i(x_i - \overline{x})^2$$

Remarques :

1. La variance est un nombre positif.
2. On a aussi :

$$V = \sum_{i=1}^{p} f_i(x_i - \overline{x})^2$$

A.1.4 Ecart type

-L'**écart type** d'une série statistique est le réel, noté σ, tel que :

$$\sigma = \sqrt{V}$$

L'écart type est donc un paramètre de dispersion car comme la variance il mesure la dispersion des valeurs de la série autour de sa moyenne.

A.1.5 Espérance mathématique

-L'**espérance mathématique** d'une variable aléatoire continue X, de densité f sur l'intervalle I = $[a\,;\,b]$ est le réel défini par **E(X) =** $\int_a^b t \times f(t)dt$

A.1.6 Le biais

Le **biais** permet de détecter la présence éventuelle d'erreur systématique. Il correspond à la différence entre l'espérance mathématique de l'estimateur d'un paramètre et le paramètre lui-même.

$$Biais = E\left(\widehat{\theta}\right) - \theta$$

A.1.7 Erreur quadratique moyenne

Mean Squared Error(MSE) : On peut construire un indicateur de précision qui englobe les notions de biais et de variance, pour cela, il suffit de calculer la moyenne des carrés des écarts des estimateurs à la vraie valeur.

$$MSE = E(\widehat{\theta} - \theta)^2 = (biais)^2 + variance$$

A.1.8 Fonction de densité de probabilité

La **fonction de densité de probabilité** est donnée par la formule suivante :

$$f\,(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

A.1.9 Vraisemblance

La probabilité d'obtenir un ensemble (de taille n) de x-valeurs est :

$$L(\mu, \sigma) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \, ^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

A.1.10 Corrélation

La **corrélation** est donnée par la formule suivante : $r = \frac{cov(x;y)}{\sigma(x)\sigma(y)}$

A.1.11 Covariance

La **covariance** mesure la relation linéaire entre deux variables. La covariance est similaire à la corrélation entre deux variables, cependant elle est différente pour les raisons suivantes : Les coefficients de corrélation sont normalisés.

A.1.12 Précision

C'est le rapport (ratio) entre le nombre de vrais positifs et la somme des vrais positifs et des faux positifs.

A.1.13 Correctly Classified Instances

Le nombre d'individus bien classés, en valeur absolue, puis en pourcentage du nombre total d'instances.

A.1.14 Incorrectly Classified Instances

Sous le même format, le nombre d'instances mal classées.

A.1.15 Mean Absolute Error

Erreur absolue en moyenne : pour chaque exemple, on calcule la différence entre la probabilité (calculée parle classifieur) pour un exemple d'appartenir à sa véritable classe, et sa probabilité initiale d'appartenir 'à la classe qui lui a été fixée dans l'ensemble d'exemples (individus) (en général, cette probabilité vaut 1). On divise ensuitela somme de ces erreurs par le nombre d'instances dans l'ensemble d'exemples.

$$MAE = \frac{|p_1 - a_1| + |p_2 - a_2| + .. + |p_n - a_n|}{n}$$

A.1.16 Root Mean-Squared Error

Cette mesure d'erreur concerne principalement les prédicteurs. Racine carrée de l'erreur quadratique moyenne : avec les mêmes notations que ci-dessus, elle correspond à :

$$RMSE = \sqrt{\frac{(p_1 - a_1)^2 + \cdots + (p_1 - a_1)^2}{n}}$$

A.1.17 Relative Absolute Error

Cette mesure d'erreur concerne principalement les prédicteursErreur absolue relative : le nom para*i*t très mal choisi.

On compare l'erreur absolue avec l'erreur absolue d'un prédicteur très simple, qui retournerait toujours lavaleur moyenne des ai, soit

$$RAE = \frac{1}{n}\sum_i a_i \frac{|p_1 - a_1| + ... + |p_n - a_n|}{|a_1 - \ a_2 \ | + .. + |a_n - a_{n+1} \ |}$$

A.1.18 Root relative Squared Error

Cette mesure d'erreur concerne principalement les prédicteurs.Racine carrée de l'erreur quadratique relative : rapport entre l'erreur quadratique et ce que serait l'erreur quadratique d'un prédicteur qui retournerait toujours la valeur moyenne.

$$RSE = \sqrt{\frac{(p_1 - a_1)^2 + \cdots + (p_n - a_n)}{(a_1 - a)^2 + \cdots + (a_n - a)}}$$

A.1.19 Kappa

Le *coefficient Kappa* est censé mesure le degré de concordance de deux ou de plusieurs *juges*.

$$k = \frac{P_0 - P_e}{1 - P_e}$$

avec Po : La proportion de l'echantillon sur laquelle les deux juges sont d'accord (i.e. la diagonale principalede la matrice de confusion).

$$P_e = \frac{\sum_i P_i P_j}{n^2}$$

– pi : somme des éléments de la ligne i
– pj : somme des éléments de la colonne j
– n : taille de l'échantillon

Certains auteurs (Landis & Koch) ont proposé une échelle de degré d'accord selon la valeur du coefficient :

Accord	Kappa
Excellent	>0.81
Bon	0.80-0.61
Modéré	0.4-0.41
Médiocre	0.4-0.21
Mauvais	0.20-0.0
Très mauvais	<0

A.1.20 Les mesures d'exactitude par classe

Pour une classe donnée, un classifieur, et un exemple, quatre cas peuvent se présenter :

1. Le classifieur ne se trompe pas : c'est un **vrai positif.**

2. Le classifieur se trompe : c'est un **faux négatif.**

3. Le classifieur la lui attribue quand même : c'est **faux positif**.

4. Le classifieur ne le range pas non plus dans cette classe : c'est un **vrai négatif.**

A.1.20.1 TP Rate

Rapport (ratio) des vrais positifs. Il correspond à :

$$\frac{nbre\ de\ vrais\ positifs}{(nbre\ de\ vrais\ positfs + nbre\ de\ faux\ ngatifs)} = \frac{nbre\ de\ vrais\ positifs}{nbre\ d'exemples\ de\ cette\ classe}$$

C'est donc le rapport entre le nombre de bien classé et le nombre total d'éléments qui devraient être bien classes.

A.1.20.2 FP Rate

Rapport des faux positifs. Il correspond à :

$$\frac{nbre\ de\ faux\ positifs}{(nbredefaux\ positfs + nbre\ de\ vrais\ ngatifs)} = \frac{nbre\ de\ faux\ positifs}{nbre\ d'exemples\ n'etantpas\ de\ cette\ classe}$$

La donnée des taux TP Rate et FP Rate permet de reconstruire la matrice de confusion pour une classe donnée.

Bibliographie

[Abraham & Sathya 2013] Abraham et Sathya. *Comparison of Supervised and Unsupervised Learning Algorithms for Pattern Classification*. International Journal of Advanced Research in Artificial Intelligence, vol. 2, no. 2, 2013.

[Adamov 2018] Abzetdin Adamov. *large-scale data modelling in hive and distributed query processing using mapreduce and tez*. Research in the framework of Center for Data Analytics Research (CeDAR), 2018.

[Agrawal *et al.* 1998] Agrawal, Gunopulos Gehrke et Raghavan. *Automatic subspace clustering of high dimensional data for data mining applications*. 1998.

[Akbani *et al.* 2004] Akbani, Kwek et Japkowicz. *Applying support vector machines to imbalanced datasets*. ECML, Springer Berlin Heidelberg, page 39–50, 2004.

[Akidau *et al.* 2015] Tyler Akidau, Slava Chernyak Robert Bradshaw Craig Chambers et Rafael Fernandez Moctezuma. *The Dataflow Model : A Practical Approach to Balancing Correctness, Latency, and Cost in Massive-Scale, Unbounded, Out-of-Order Data Processing*. Proceedings of the VLDB Endowment, vol. 8, no. 12, 2015.

[Albattah 2016] Waleed Albattah. *The Role of Sampling in Big Data Analysis*. International Conference on Big Data and Advanced Wireless Technologies, vol. doi>10.1145/3010089.3010113, 2016.

[Aldossary & Allen 2016] Sultan Aldossary et William Allen. *Data Security, Privacy, Availability and Integrity in Cloud Computing : Issues and Current Solutions*. International Journal of Advanced Computer Science and Applications, vol. 7, no. 4, 2016.

[Alhadad 2018] Sakinah Alhadad. *Visualizing Data to Support Judgement, Inference, and Decision Making in Learning Analytics : Insights from Cognitive Psychology and Visualization Science*. The Journal of Learning Analytics, vol. 5, no. 2, pages 60–85, http ://dx.doi.org/10.18608/jla.2018.52.5, 2018.

[Allison 2014] Paul Allison. *Measures of Fit for Logistic Regression*. Statistical Horizons LLC and the University of Pennsylvania, Paper 1485, 2014.

[Alvi 2016] Mohsin Alvi. *A Manual for Selecting Sampling Techniques in Research*. University of Karachi, MPRA Paper No. 70218, vol. disponible sur : https ://mpra.ub.uni-muenchen.de/70218/, 2016.

[Anders & KAndrot 2011] JAson Anders et EdwArd KAndrot. *CUDA by Example : An Introduction to General-Purpose GPU programming*. NVIDIA Corporation, 2011.

[Anderson *et al.* 1995] Anderson, Culler et Patterson. *A case for NOW*. IEEEMicro, vol. 15, no. 1, pages 54–64, 1995.

[Andrew & Steen 2016] Andrew et Steen. *A brief introduction to distributed systems*. Computing, vol. 98, pages 967–1009, doi 10.1007/s00607–016–0508–7, 2016.

[Anguera *et al.* 2018] Anguera, Chacón-Moscoso Portell et Sanduvete-Chaves. *Indirect observation in everyday contexts : Concepts and methodological guidelines within a mixed methods framework*. Frontiers in Psychology, vol. 9, page Article ID 13. http ://dx.doi.org/10.3389/fpsyg.2018.00013, 2018.

[Ansari & Swarna 2017] Zahid Ansari et Swarna. *Apache Pig - A Data Flow Framework Based on Hadoop Map Reduce*. International Journal of Engineering Trends and Technology, vol. 50, no. 5, 2017.

[Antal & Tille 2011] Erika Antal et Yves Tille. *Simple random sampling with over-replacement*. Journal of Statistical Planning and Inference, vol. 141, no. 1, pages 597–601, 2011.

[Arasu *et al.* 2013] Arasu, Kossmann Ramamurthy Eguro Kaushik et Venkatesan. *A secure coprocessor for database applications*. In Proceedings of the 23rd International Conference on Field programmable Logic and Applications, pages 1–8, doi :10.1109/FPL.2013.6645524, 2013.

[Ardilly 2006] Ardilly. *Les techniques de sondage*. Edition TECHNIP, 2006.

[Arlot & Celisse 2010] Arlot et Celisse. *A survey of cross-validation procedures for model selection*. Statistics,surveys, vol. 4, page 40–79, 2010.

[Armstrong *et al.* 2014] Timothy Armstrong, Michael Wilde Justin Wozniak et Ian Foster. *Compiler Techniques for Massively Scalable Implicit Task Parallelism*. SC14, New Orleans, Louisiana, USA, 2014.

[Arren 1952] Torgerson Arren. *Multidimensional scaling : I. Theory and method*. Psychometrika, vol. 17, pages 401–419, 1952.

[Asprey 1989] William Asprey. *Von Neumann's contributions to computing and computer science*. Annals of the history of computing, vol. 11, no. 3, pages 189–195, 1989.

[Asprey 1990] William Asprey. *Von Neumann and the Origins of modern computing*. The MIT press. Cambridge, Mass, 1990.

[AStephen 2015] Thomas AStephen. *Data Visualization with JavaScript*. ed. s.l. :No Starch Press, 2015.

[Ataro 1967] Yamane Ataro. *Statistics, an introductory analysis*. 2nd ed. New York : Harper and Row, 1967.

[Atiquzzaman 1993] Atiquzzaman. *Performance modeling of multiprocessor systems for different data loading schemes*. Microprocessing and Microprogramming, vol. 36, no. 4, pages 167–178, https ://doi.org/10.1016/0165–6074(93)90241–C, 1993.

[Azeem *et al.* 2015] Muhammad Waqas Azeem, Arslan Tariq Farzan Javed Sheikh et Mirza Ahsan Ullah. *A Review on Multiple Instruction Multiple Data (MIMD) Architecture*. Proceedings of the 1st International Multi-Disciplinary Conference (IMDC), The University of Lahore , Gujrat Campus, PK, 23-24, 2015.

[Bache & Lichman 2013] Bache et Lichman. UCI machine learning repository, vol. http ://archive.ics.uci.edu/ml, 2013.

[Baer 1976] Baer. *Multiprocessing Systems*. IEEE Transactions on Computers, vol. 25, pages 1271–1277, doi : 10.1109/TC.1976.1674594, 1976.

[Balaji & Baskaran 2013] Arun Balaji et Baskaran. *Design and development of artificial neural networking (ann) system using sigmoid activation function to predict annual rice production in tamilnadu*. International Journal of Computer Science, Engineering and Information Technology, vol. 3, no. 1, 2013.

[Balasundaram 2009] Nimalathasan Balasundaram. *Factor Analysis : Nature, Mechanism and Uses in Social and Management Science*. Journal of Cost and Management Accountant, vol. 37, no. 2, pages 15–25, 2009.

[Banerjee & Wolfe 1987] Utpal Banerjee et Michael Wolfe. *Data dependence and its application to parallel processing*. International Journal of Parallel Programming, vol. 16, no. 2, page 137–178, 1987.

[Banerjee *et al.* 2007] Banerjee, Banerjee Mahato Chaudhury Singh et Haldar. *Statistics without tears - inputs for sample size calculations*. Indian Psychiatr Journal, vol. 16, pages 150–152, 2007.

[Barapatre & Vijayalakshmi 2017] Darshan Barapatre et Vijayalakshmi. *Data preparation on large datasets for data science*. Asian Journal of Pharmaceutical and clinical research (AJPCR), pages ISSN : 2455–3891 ,DOI https ://doi.org/10.22159/ajpcr.2017.v10s1.20526,, 2017.

[Barker & Ward 2013] Barker et Ward. *Undefined by data : a survey of big data definitions*. arXiv preprint arXiv :1309.5821, 2013.

[Barto & Sutton 2014] Andrew Barto et Sutton. *Reinforcement Learning : An Introduction*. Second edition, The MIT Press Cambridge, Massachusetts London, 2014.

[BASU 2016] De BASU. *Parallel and Distributed Computing : Architectures and algorithms*. pages 14–15, 2016.

[Bei *et al.* 2018] Zhendong Bei, Chuntao Jiang Chengzhong Xu Zhibin Yu Ni Luo et Shengzhong Feng. *Configuring*

in-memory cluster computing using random forest. Future Generation Computer Systems, vol. 79, pages 1–15, http ://dx.doi.org/10.1016/j.future.2017.08.011, 2018.

[Bell 2010] Bell. *Doing Your Research Project*. Maidenhead : Open University Press, vol. (5th ed), 2010.

[Belzer *et al.* 1997] De Jack Belzer, Albert Holzman et Allen Kent. *Encyclopedia of Computer Science and Technology*. computer selection to curriculum, vol. 6, pages 40–69, 1997.

[Benfield & Szlemko 2006] Benfield et Szlemko. *Internet-based data collection : Promises and realities*. Journal of Research Practice, vol. 2, no. 2, pages Article D1. Retrieved from, http ://jrp.icaap.org/index.php/jrp/article/view/30/51, 2006.

[Benjamin 2008] Wah Benjamin. *Interconnection networks for parallel computers*. Wiley Encyclopedia of Computer Science and Engineering, 2008.

[Benmammar 2017] Badr Benmammar. *Concurrent, Real-Time and Distributed Programming in Java : Threads, RTSJ and RMI*. FOCUS Series in Computer Engineering,Abu Bekr Belkaid University, Tlemcen, Algeria, 2017.

[Bernhard *et al.* 1999] Scholkopf Bernhard, Smola Alexander et Muller Klaus. *Kernel principal component analysis*. Advances in Kernel Methods – Support Vector Learning, page 327–352. MIT Press, 1999.

[Bertsekas & Shreve 1978] Bertsekas et Shreve. *Stochastic Optimal Control -The DiscreteTime Case-*. Academic Press, New York. 1, 1978.

[Bertsekas 2007] Bertsekas. *Dynamic Programming and Optimal Control*. Athena Scientific, Belmont, MA, 3 edition, vol. 1, a 2007.

[Best & Kahn 2003] Best et Kahn. *Research in Education*. 9th Edition, Prentice-Hall of India Private Limited, New Delhi., 2003.

[Bhaskar & Zulfiqar 2016] Bhaskar et Zulfiqar. *Basic statistical tools in research and data analysis*. Indian J Anaesth, vol. 60, no. 9, pages 662–669. doi : 10.4103/0019–5049.190623, 2016.

[Bhattacharya & Bhatnagar 2016] Abhishek Bhattacharya et Shefali Bhatnagar. *Big Data and Apache Spark : A Review*. International Journal of Engineering Research Science, vol. 2, no. 5, 2016.

[Biggio *et al.* 2011] Biggio, Nelson et Laskov. *Support vector machines under adversarial label noise*. Asian Conference on Machine Learning, JMLR : Workshop and Conference Proceedings, vol. 20, page 97–112, 2011.

[Bikakis 2018] Nikos Bikakis. *Big Data Visualization Tools*. ATHENA Research Center, Greece, vol. arXiv :1801.08336v2 [cs.DB], Springer, 2018.

[Blascheck & Ertl 2013] Tanja Blascheck et Thomas Ertl. Workshop on Visual and Spatial Cognition, vol. Techniques for Analyzing Empirical Visualization Experiments Through Visual Methods, 2013.

[Boeth 1970] Boeth. *The Assault on Privacy : Snoops, Bugs, Wiretaps, Dossiers, Data Bann Banks, and Specters of 1984*. Newsweek. Incorporated, 1970.

[Borthakur 2007] Dhruba Borthakur. *The Hadoop Distributed File System : Architecture and Design*. The Apache Software Foundation, 2007.

[Bouazza 2017] Naoufal Ben Bouazza. *Apprendre à bien choisir son architecture Big Data*. Tutorial, vol. disponible sur : https ://big-data.developpez.com/tutoriels/apprendre-faire-choix-architecture-big-data/, 2017.

[Breiman 2001] Leo Breiman. *Random forests*. Machine Learning, vol. 5, no. 2, page 5–32, 2001.

[Brownbridge *et al.* 1982] David Brownbridge, Philip Treleaven et Richard Hopkins. *Data-Driven and Demand-Driven Computer Architecture*. Computing Surveys, vol. 14, no. 1, 1982.

[Brownlee 2016] Jason Brownlee. *Machine learning – How it works*. pages 1–5, 2016.

[Bruce & Bruce 2017] Peter Bruce et Andrew Bruce. *Practical Statistics for Data Scientists*. Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472, 2017.

[Brydon & Gemino 2008] Michael Brydon et Andrew Gemino. *Classification trees and decision-analytic feedforward control : a case study from the video game industry*. Data Min Knowl Disc, vol. 17, pages 317–342 DOI 10.1007/s10618–007–0086–6, 2008.

[Bréchon 2015] Pierre Bréchon. *Random Sample, Quota Sample : The Teachings of the EVS 2008 Survey in France*. vol. 126, pages 67–83, DOI : 10.1177/0759106315572558, 2015.

[Burhan *et al.* 2014] Khan Burhan, Rashidah et Hunain. *Critical Insight for MapReduce Optimization in Hadoop*. International J of Computer Science and Control Engineering, vol. 2, no. 1, pages 1–7, 2014.

[Burks & Neumann 1946] Burks et Von Neumann. *Preliminary Discussion of the Logical Design of an Electronic Computing Instrument*. Princeton : Institute for Advanced Studies, 1946.

[Buyya *et al.* 2016] Rajkumar Buyya, Wu et Kotagiri Ramamohanarao. *Big Data Analytics = Machine Learning + Cloud Computing*. Published in ArXiv, vol. doi :10.1016/b978-0-12-805394-2.00001-5, 2016.

[Campbell & Swinscow 2009] Campbell et Swinscow. Statistics at Square One, 11th ed. Oxford : Wiley-Blackwell, 2009.

[Cardie & Wagstaff 2001] Claire Cardie et Kiri Wagstaff. *Constrained K-means Clustering with Background Knowledge*. In Proceedings of the Eighteenth International Conference on Machine Learning, page 577–584, 2001.

[Caruana *et al.* 2008] Rich Caruana, Nikos et Ainur. *An Empirical Evaluation of Supervised Learning in High Dimensions*. Conference on Machine Learning, ACM, 2008.

[CE 1998] Paz CE. *A survey of parallel genetic algorithms*. Calc Paralleles Reseaux et Syst Repar, vol. 10, no. 2, page 141–171, 1998.

[Chan & Stolfo 1998] Chan et Stolfo. *Toward Scalable Learning with Non-uniform Class and Cost Distributions : A Case Study in Credit Card Fraud Detection*. In Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining, page 164–168. AAAI Press, 1998.

[Chan 2013] Chan. *An architecture for big data analytics*. Communications of the IIMA, vol. 13, no. 2, page 1–13, 2013.

[Changtian *et al.* 2018] Ying Changtian, Ying Changyan et Ban Chen. *A performance optimization strategy based on degree of parallelism and allocation fitness*. Journal on Wireless Communications and Networking, vol. Article number : 240, 2018.

[Chavent *et al.* 2007] Marie Chavent, Yves Lechevallier et Olivier Briant. *DIVCLUS-T : A monothetic divisive hierarchical clustering method*. Comput. Statist. Data Anal, vol. doi : 10.1016/j.csda.2007.03.013, 2007.

[Chen *et al.* 2006] Chen, Jiang et Yoshihira. *Robust nonlinear dimensionality reduction for manifold learning*. In Proceeding of 18th International Conference on Pattern Recognition, page 447–450, 2006.

[Chen *et al.* 2014a] Chen, Mao et Liu. *Big Data : a survey*. mobile networks and application, vol. 19, no. 2, page 171–209, 2014.

[Chen *et al.* 2014b] Chen, Mao et Liu. *Big data : a survey*. Mob. Netw. Appl, vol. 19, page 171–209, 2014.

[Child 2006] Dennis Child. *The Essentials of Factor Analysis*. 3rd edition, continuum international publishing group, vol. The Tower Building, 11 York Road, SE1 7NX, London, 2006.

[Chiliang & Wenting 2012] Chiliang et Wenting. *Cross-domain representation-learning framework with combination of class-separate and domain-merge objectives*. Proceedings of the CDKD 2012 Conference, page 18–25, 2012.

[Chittaro 2006] Chittaro. *Visualizing Information on Mobile Devices*. ACM Computer, vol. 39, no. 3, pages 40–45, 2006.

[Chmidt 2012] Chmidt. *Data is exploding : the 3 versus of big data*. Bus Comput World, vol. 15, 2012.

[Choi *et al.* 2016] Seung-Hyun Choi, Yong-Min Tai Junguk Cho et Seong-Won Lee. *A parallel camera image signal processor for SIMD architecture*. EURASIP Journal on Image and Video Processing, vol. 29, pages doi 10.1186/s13640–016–0137–2, 2016.

[Christopher 1989] Watkins Christopher. *Learning from Delayed Rewards*. thesis, Cambridge University, Cambridge, England., 1989.

[Ciglaric *et al.* 2003] Mojca Ciglaric, Matjaz Pancur Matej Trampus et Tone Vidmar. *Message routing in pure peer-to-peer networks*. vol. Disponible sur : https ://pdfs.semanticscholar.org/c52a/666c1186b19ccc438260edbc921844608b57.pdf, 2003.

[Cochran 1977] Cochran. *Sampling techniques*. 3rded. New York : John Wiley and Sons,inc, page 75, 1977.

[Colin 2004] Ware Colin. *Information Visualization : Perception for Design*. Morgan Kaufmann, 2004.

[conrad taeuber 1961] Richard conrad taeuber. *On sampling with replacement : an axiomatic approach*. Institute of Statistics Mimeo Series, vol. 299, 1961.

[Conti 2015] Francesco Conti. *Heterogeneous Architectures for Parallel Acceleration*. Thèse de doctorat,University of Bologna, 2015.

[Coulet *et al.* 2018] Adrien Coulet, Mohammad Chawki Nicolas Jay Nigam Shah Maxime Wack et Michel Dumontier. *Predicting the need for a reduced drug dose, at first prescription*. Scientific Reports, vol. 8, page Article number : 15558, 2018.

[Crinivasa *et al.* 2018] Crinivasa, Siddesh et Srinidhi. *Network Data Analytics*. chapter 6, 1st edition, pages 95–105, 2018.

[Cunningham & Delany 2007] Padraig Cunningham et Sarah Jane Delany. *k-Nearest Neighbour Classifiers*. Technical Report UCD-CSI-2007-4, 2007.

[da Silva *et al.* 2018] Ticiana Coelho da Silva, Regis Magalh aes et Igo Brilhante. *Big Data Analytics Technologies and Platforms : a brief review*. LADaS, 2018.

[Darlington 1990] Darlington. *Regression and Linear Models*. Columbus, OH : McGraw-Hill Publishing Company., 1990.

[Das & Behera 2017] Kajaree Das et Rabi Narayan Behera. *A Survey on Machine Learning : Concept, Algorithms and Applications*. International Journal of Innovative Research in Computer and Communication Engineering, vol. 5, no. 2, pages ISSN : 2320–9798, DOI : 10.15680/IJIRCCE.2017. 0502001, 2017.

[Dasgupta & Nath 2016] Ariruna Dasgupta et Asoke Nath. *Classification of Machine Learning Algorithms*. International Journal of Innovative Research in Advanced Engineering (IJIRAE),ISSN : 2349-2763, vol. 3, no. 3, 2016.

[Dataflair 2018] Team Dataflair. *Spark Tutorial : Learn Spark Programming*. disponible sur : https ://data-flair.training/blogs/spark-sql-tutorial/, 2018.

[Davenport & Kim 2013] Davenport et Kim. *Keeping Up with the Quants*. Harvard Business Review Press, USA, 2013.

[Dayan 1999] Peter Dayan. *Unsupervised Learning*. In Wilson, RA Keil, F, editors. The MIT Encyclopedia of the Cognitive Sciences, 1999.

[DBTA 2013] DBTA. *Big Data Sourcebook*. Unisphere Media., 2013.

[Demidova *et al.* 2016] Demidova, Nikulchev et Sokolova. *Big Data Classification Using the SVM Classifiers with the Modified Particle Swarm Optimization and the SVM Ensembles*. International Journal of Advanced Computer Science and Applications, vol. 7, no. 5, 2016.

[den Broeck *et al.* 2005] Van den Broeck, Eeckels Argeseanu CunninghamS et Herbst. *Data cleaning : Detecting,diagnosing, and editing data abnormalities*. PLoS Med, vol. 2, no. 10, page e267, 2005.

[den Broeck *et al.* 2013] Van den Broeck, Sandøy et Brestoff. *The Recruitment, Sampling, and Enrollment Plan - Epidemiology : principles and practical guidelines*. Springer Netherlands, pages 171–196, 2013.

[Deng & Yu 2013] Li Deng et Dong Yu. *Deep Learning Methods and Applications, Foundations and Trends*. Signal Processing, vol. 7, no. 3-4, pages 197–387, doi : 10.1561/2000000039, 2013.

[DESASO 1964] Department of economic social affairs statistjcal office of the united nations DESASO. *Recommendations for the Preparation of Sample Survey Reports*. United nations new york, statistical papers, vol. 1, no. 2, 1964.

[Deshpande *et al.* 2016] Siddharth Deshpande, Nithya Gogtay et Urmila Thatte. *Data Types*. Journal of The Association of Physicians of India, vol. 64, 2016.

[Dharwat 2016] Alaa Dharwat. *Principal component analysis*. A tutorial, Frankfurt University of Applied Sciences, vol. doi : 10.1504/IJAPR.2016.079733, 2016.

[Dhyani & Barthwal 2014] Bijesh Dhyani et Anurag Barthwal. *Big Data Analytics using Hadoop*. International Journal of Computer Applications, vol. 108, no. 12, 2014.

[Diger 2001] Schollmeier Diger. *A Definition of Peer-to-Peer Networking for the Classification of Peer-toPeer Architectures and Applications*. Proceedings of the First International Conference on Peer-to-Peer Computing, page doi : 10.1109/P2P.2001.990434, 2001.

[Dijkstra & Broy 1985] Edsger Dijkstra et Manfred Broy. *Control Flow and Data Flow : Concepts of Distributed Programming*. International Summer School, first edition, 1985.

[Dipboye 1994] Dipboye. *Structured and unstructured selection interviews*. Research in Personnel and Human Resources Management, vol. 12, pages 79–123, ISBN : 1–55938–733–5, 1994.

[Djafri & Mekki 2012] Laouni Djafri et Rachida Mekki. *Monitoring and Resource Management in P2P Grid-Based Web Services*. Computer Engineering and Applications, vol. 1, no. 1, pages ISSN : 2252–5459, 2012.

[Djafri *et al.* 2018] Laouni Djafri, Djamel Amar bensaber et Reda Adjoudj. *BIG DATA ANALYTICS FOR PREDICTION : parallel processing of the big learning base with the possibility of improving the final result of the prediction*. Information discovery and delivery, vol. 46, no. 3, 2018.

[Dobre & Xhafa 2014] Dobre et Xhafa. *Intelligent services for big data science*. Future Generation Computer Systems, vol. 37, page 267–281, 2014.

[Dong *et al.* 2013] Long Jun Dong, Xi Bing Li et Kang Peng. *Prediction of rockburst classification using Random Forest*. Transactions of Nonferrous Metals Society of China, vol. 23, pages 472477, doi : 10.1016/S1003–6326(13)62487–5, 2013.

[Dong *et al.* 2016] Yanchao Dong, Jiguang Yue Yan Zhang et Zhencheng Hu. *Comparison of random forest, random ferns and support vector machine for eye state classification*. Multimedia Tools and Applications, vol. 75, pages 11763–11783, doi : 10.1007/s11042–015–2635–0, 2016.

[Dongarra & van der Steen 2012] Dongarra et van der Steen. *High-performance computing systems : Status and outlook*. Acta Numerica, pages 1–96, doi :10.1017/S09624929XXXXXXXX, 2012.

[Dormehl 2014] Dormehl. *The Five Best Libraries For Building Data Visualizations*. Fast Company, 2014.

[Dreyfus 2002] Dreyfus. *Richard Bellman on the birth of dynamic programming*. MLRG - Winter Term 2, vol. 50, no. 1, page 48–51, 2002.

[Drugan 2017] Madalina Drugan. *Reinforcement learning versus evolutionary computation : a survey on hybrid algorithms*. Technical University of Eindhoven, The Netherlands, 2017.

[Dörnyei 2007] Dörnyei. *Research methods in applied linguistics*. New York : Oxford University Press, 2007.

[Eassa & Zaki 1995] Fathy Eassa et Zaki. *A computational model for static data flow machines*. Computers Electrical Engineering, vol. 21, no. 6, pages 483–497, doi : https ://doi.org/10.1016/0045–7906(95)00018–P, 1995.

[Eaton *et al.* 2011] Eaton, Deutsch Zikopoulos DeRoos et Lapis. *Understanding Big Data*. McGraw-Hill, USA, 2011.

[Eaton *et al.* 2012] Eaton, Deutsch Deroos et Lapis. *Understanding Big Data : Analytics for Enterprise Class Hadoop and Streaming Data*. McGraw Hill Professional, McGraw Hill, New York, vol. ISBN : 978-0071790536, 2012.

[Eckerson 2007] Wayne Eckerson. *Predictive analytics : Extending the Value of Your Data Warehousing Investment*. tdwi best practices report, 2007.

[Elnour *et al.* 2014] Manhal Elfadil Eltayeeb Elnour, Muhammad Shafie Abd Latif et Ismail Fauzi Isnin. *Distributed Memory and Shared Distributed Memory Architecture for Implementing Local Sequences Alignment : A Survey*. International Journal of Computer Science and Telecommunications, vol. 5, no. 8, 2014.

[EMCES 2015] EMC Education Services EMCES. *Data Science and Big Data Analytics*. Indianapolis : John Wiley Sons, vol. 978-1-118-87613-8, 2015.

[Erl *et al.* 2016] Erl, Khattak et Buhler. *Big Data Fundamentals : Concepts*. Prentice Hall Press, Drivers Techniques, 2016.

[Espinosa *et al.* 2012] Mariano Martinez Espinosa, Isanete Bieski et Domingos Tabajara de Oliveira Martins. *Probability sampling design in ethnobotanical surveys of medicinal plants*. Revista Brasileira de Farmacognosia, vol. 22, no. 6, pages http ://dx.doi.org/10.1590/S0102–695X2012005000091, 2012.

[Ester *et al.* 1996] Martin Ester, Jörg Sander Hans-Peter Kriegel et Xiaowei Xu. *A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise*. In KDD'96 Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, pages 226–231, 1996.

[Etikan & Bala 2017] Etikan et Bala. *Sampling and Sampling Methods*. Biom Biostat Int J, vol. 5, no. 6, pages 138–149, DOI : 10.15406/bbij.2017.05.00149, 2017.

[Etikan *et al.* 2016] Ilker Etikan, Sulaiman Abubakar Musa et Rukayya Sunusi Alkassim. *Comparison of Convenience Sampling and Purposive Sampling*. American Journal of Theoretical and Applied Statistics, vol. doi : 10.11648/j.ajtas.20160501.11, 2016.

[Etzioni *et al.* 2003] Oren Etzioni, Rattapoom Tuchinda Craig Knoblock et Alexander Yates. SIGKDD, pages Washington, DC, USA, 2003.

[Evans & Lindner 2012] Evans et Lindner. *Business Analytics : The Next Frontier for Decision Sciences*. Decision Line, vol. 43, no. 2, pages 4–6, 2012.

[Even-Dar *et al.* 2009] Even-Dar, Kakade et Mansour. *Online markov decision processes*. Math. Oper. Res, vol. 34, no. 3, page 726–736, 2009.

[Ewens 2004] Ewens. *Mathematical Population Genetics 1 - Theoretical Introduction*. New York : Springer, 2004.

[Fan *et al.* 2012] Fan, Gondek Kalyanpur et Ferrucci. *knowledge extraction from documents*. IBM Journal of Research and Development, vol. 56 (3.4), no. 5, pages 1–10, 2012.

[Faraz *et al.* 2015] Ahmed Faraz, Faiz Ul Haque Zeya et Majid Kaleem. *A survey of paradigms for building and designing parallel computing machines*. Computer Science Engineering : An International Journal, vol. 5, no. 1, 2015.

[Faubert & Wheeldon 2009] Jacqueline Faubert et Johannes Wheeldon. *Framing Experience : Concept Maps, Mind Maps, and Data Collection in Qualitative Research*. International Journal of Qualitative Methods, vol. 8, no. 3, 2009.

[Fedak *et al.* 2001] Fedak, Néri Germain et Cappello. *XtremWeb : A generic global computing system*. Proceedings of the IEEE International Symposium on Cluster Computing and the Grid (CCGRID). IEEE Press, Piscataway, New Jersey, 2001.

[Fei 2015] Shi Fei. *Study on a Stratified Sampling Investigation Method for Resident Travel and the Sampling Rate*. Discrete Dynamics in Nature and Society, vol. Article ID 496179, http ://dx.doi.org/10.1155/2015/496179, 2015.

[Fellows 1994] Michael Fellows. *On Search, Decision, and the Efficiency of Polynomial-Time Algorithms*. Journal of computer and system sciences, vol. 49, 1994.

[Feo 1992] Feo. *A Comparative Study of Parallel Programming Languages : The Salishan Problems*. North-Holland, 1992.

[Ferguson 2013] Mike Ferguson. *Enterprise Information Protection- The Impact of Big Data*. IBM, 2013.

[Fernandez-Delgado *et al.* 2014] Fernandez-Delgado, Barro S Cernada E et Amorim D. *Do we need hundreds of classiffiers to solve real world classification problems.* J Mach Learn Res, vol. 15, 2014.

[Finney 1948] Finney. *random and systematic sampling in timber surveys*. International Journal of Forest Research, vol. 22, no. 1, pages 64–99, https ://doi.org/10.1093/oxfordjournals.forestry.a062953, 1948.

[Fischer & Plessow 2015] Rico Fischer et Franziska Plessow. *Efficient multitasking : parallel versus serial processing of multiple tasks*. Front Psychol, vol. 6, page doi : 10.3389/fpsyg.2015.01366, 2015.

[Flynn 1966] Flynn. *Very High–Speed Computing Systems*. Proceedings of the IEEE 54, 12,, page 1901–1909, 1966.

[Flynn 1972] Flynn. *Some Computer Organizations and Their Effectiveness*. IEEE Trans. Computers, vol. 21, no. 9, pages 948–960, 1972.

[Flynn 1995] Flynn. *Computer Architecture : Pipelined and Parallel Processor Design*. . Jones and Bartlett, Boston, 1995.

[Forster 2001] J.J. Forster. *Sample Surveys : Nonprobability Sampling*. International Encyclopedia of the Social Behavioral Sciences, 2001.

[Foster *et al.* 2014] Kenneth Foster, Robert Koprowski et Joseph Skufca. *Machine learning, medical diagnosis, and biomedical engineering research - commentary*. BioMedical Engineering OnLine, vol. 13, page Article number : 94, 2014.

[Foster 1995] Foster. *Designing and building parallel programs : Concepts and tools for parallel software engineering*. Addison-Wesley, 1995.

[Frank *et al.* 2000] Eibe Frank, Leonard Trigg et Geo rey Holmes. *Naive Bayes for Regression*. Machine Learning, vol. 41, no. 1, pages 5–25, 2000.

[Frenay & Verleysen 2014] Frenay et Verleysen. *Classification in the presence of label noise : a survey*. IEEE Transactions on Neural Networks and Learning Systems, vol. 25, no. 5, page 845–869, 2014.

[Frey & Dueck 2007] Frey et Dueck. *Clustering by passing messages between data points*. Science, vol. 315, page 972–976, 2007.

[Ganti *et al.* 1999] Ganti, Powell Ramakrishnan Gehrke et French. *Clustering large datasets in arbitrary metric spaces*. In Proceedings of the 15th Int'l Conf. on Data Eng, 1999.

[Gantz & Reinsel 2012a] Gantz et Reinsel. *The Digital Universe in 2020 : Big data, bigger digital shadows, and biggest growth in the Far East*. IDC – EMC Corporation, vol. disponible sur http ://www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf, 2012.

[Gantz & Reinsel 2012b] Gantz et Reinsel. *The digital universe in 2020 : Big data, bigger digital shadows, and biggest growth in the far east*. IDC : Analyze the Future, 2012.

[Garcia *et al.* 2017] Jesus Garcia, Pedro Balda Wei Zheng et Fernando Martinez. *Effects of late winter pruning at different phenological stages on vine yield components and berry composition in La Rioja, north-central Spain*. , Journal International des Sciences de la Vigne et du Vin, vol. doi : https ://doi.org/10.20870/oeno-one.2017.51.4.1863, 2017.

[Gareth *et al.* 2013] James Gareth, Trevor Hastie Daniela Witten et Robert Tibshirani. *An introduction to statistical learning*. Springer, vol. 112, 2013.

[Geist *et al.* 1994] Geist, Dongarra Sunderam et Manchek. *The PVM concurrent computing system : Evolution, experience, and trends*. Parallel Computing, vol. 20, pages 531–546, 1994.

[Gennari *et al.* 1989] Gennari, Langley et Fisher. *Models of incremental concept formation*. 1989.

[Ghorbani & Ghousi 2019] Ramin Ghorbani et Rouzbeh Ghousi. *Predictive data mining approaches in medical diagnosis : A review of some diseases prediction*. International Journal of Data and Network Science, vol. 3, pages 47–70, doi : 10.5267/j.ijdns.2019.1.003, 2019.

[Gill *et al.* 2008] Gill, Treasure Stewart et Chadwick. *Methods of data collection in qualitative research : interviews and focus groups*. BRITISH DENTAL JOURNAL, vol. 204, no. 6, page DOI : 10.1038/bdj.2008.192, 2008.

[Gim *et al.* 2018] Jangwon Gim, Sukhoon Lee et Wonkyun Joo. *A Study of Prescriptive Analysis Framework for Human Care Services Based On CKAN Cloud*. Hindawi Journal of Sensors, pages Article ID 6167385,https ://doi.org/10.1155/2018/6167385, 2018.

[Gonzalez *et al.* 2014] Joseph Gonzalez, Ankur Dave Reynold Xin et Daniel Crankshaw. *GraphX : Graph Processing in a Distributed Dataflow Framework*. Proceeding OSDI'14 Proceedings of the 11th USENIX conference on Operating Systems Design and Implementation, pages 599–613, 2014.

[Gowrishankar & Veena 2019] Gowrishankar et Veena. *Introduction to Python Programming*. 2019.

[Grafstrom 2010] Anton Grafstrom. *On Unequal Probability Sampling Designs*. Doctoral Dissertation Department of Mathematics and Mathematical Statistics Umea University, Sweden, vol. ISBN : 978-91-7264-999-6, 2010.

[Gravetter & Forzano 2012] Gravetter et Forzano. *Selecting Research Participants*. Res. Methods Behav. Sci., page 125–139, 2012.

[Grinter 2013] Grinter. *A big data confession*. Interactions, vol. 20, no. 4, page 10–11, 2013.

[Guanghui *et al.* 2015] Guanghui, Lidong et Cheryl Alexander. *Big Data and Visualization : Methods, Challenges and Technology Progress*. Digital Technologies, vol. 1, no. 1, pages 33–38, doi : 10.12691/dt–1–1–7, 2015.

[Guerra *et al.* 2015] Guerra, Simonini et Vincini. *Supporting image search with tag clouds : a preliminary approach*. Advances in Multimedia, pages 1–10, https ://doi.org/10.1155/2015/439020, 2015.

[Guha *et al.* 1998] Guha, Rastogi et Shim. *An efficient clustering algorithm for large databases*. Special Interest Group on Management of Data Record, vol. 27, no. 2, page 73–84, 1998.

[Güler & Uyanık 2013] Neşe Güler et Gülden Uyanık. *A Study on Multiple Linear Regression Analysis*. Procedia - Social and Behavioral Sciences, vol. 106, pages 234–240,doi :10.1016/j.sbspro.2013.12.027, 2013.

[Hahn *et al.* 1997] Woo-Jong Hahn, Kee-Wook Rim et Soo-Won Kim. *SPAX : A New Parallel Processing System for Commercial Applications*. Proceedings 11th International Parallel Processing Symposium, 1997.

[Haoyuan *et al.* 2013] Haoyuan, Zaharia, Scott Shenker Tathagata Das Timothy Hunter et Ion Stoica. *Discretized Streams : Fault-Tolerant Streaming Computation at Scale*. SOSP'13, Nov. 3–6, Farmington, Pennsylvania, USA. ACM, page doi :http ://dx.doi.org/10.1145/2517349.2522737, 2013.

[Hartuv & Shamir 1999] Hartuv et Shamir. *A clustering algorithm based on graph connectivity*. Information Processing Letters, vol. 76, page 175–181, 1999.

[He & Garcia 2009a] He et Garcia. *Learning from imbalanced data*. IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 9, page 1263–1284, 2009.

[He & Garcia 2009b] He et Garcia. *Learning from Imbalanced Data. Knowledge and Data Engineering*. IEEE Transactions, vol. 21, no. 9, page 1263–1284, 2009.

[Hepsa 2000] Silliman Hepsa. *Memorial Lectures*. The Computer and The Brain, 2nd edition- John von Neumann, 2000.

[Hillis & Steele 1986] Hillis et Steele. *Data parallel algorithms*. Communications of the ACM, vol. 29, no. 12, page 1170–1183, 1986.

[Hinneburg & Keim 1998] Hinneburg et Keim. *An efficient approach to clustering in large multimedia databases with noise*. 1998.

[Honnutagi 2014] Pooja Honnutagi. *The Hadoop distributed file system*. International Journal of Computer Science and Information Technologies, vol. 5, no. 5, pages 6238–6243, 2014.

[Hota & Prabhu 2012] Hota et Prabhu. *No problem with Big Data. What do you mean by Big ?* Journal of Informatics, page 30–32, 2012.

[Howard 2018] Seltman Howard. *Experimental Design and Analysis*. phdthesis - Carnegie Mellon University, 2018.

[Huizhen *et al.* 2018] Huizhen, Rupam Mahmood et Richard Sutton. *On Generalized Bellman Equations and Temporal-Difference Learning*. Journal of Machine Learning Research, vol. 19, pages 1–49, 2018.

[Hurwitz *et al.* 2013] Hurwitz, Halper Nugent et Kaufman. *Big Data For Dummies*. John Wiley Sons, Inc. Hoboken, vol. NJ 07030-5774, 2013.

[Hyeoun-Ae 2013] Park Hyeoun-Ae. *An Introduction to Logistic Regression : From Basic Concepts to Interpretation with Particular Attention to Nursing Domain*. J Korean Acad Nurs, vol. 43, no. 2, page http ://dx.doi.org/10.4040/jkan.2013.43.2.154, 2013.

[Hyvarinen *et al.* 2001] Aapo Hyvarinen, Juha Karhunen et Erkki Oja. *Independent Component Analysis*. A Wiley-Interscience Publication JOHN WILEY SONS, INC, 2001.

[Iafrate & Front 2015] Fernando Iafrate et Matter Front. *From Big Data to Smart Data*. John Wiley Sons., 2015.

[IBM 2014] IBM. *The top five ways to get started with big data*. 2014.

[Imandoust & Bolandraftar 2013] Sadegh Bafandeh Imandoust et Mohammad Bolandraftar. *Application of K-Nearest Neighbor (KNN) Approach for Predicting Economic Events : Theoretical Background*. Int. Journal of Engineering Research and Applications, vol. 3, no. 5, pages 605–610, 2013.

[Inderpal 2013] Singh Inderpal. *Review on Parallel and Distributed Computing*. Scholars Journal of Engineering and Technology, vol. 1, no. 4, pages 218–225, 2013.

[Jackson 2011] Jackson. *Research Methods and Statistics : A Critical Approach*. 4th edition, Cengage Learning, 2011.

[Jain *et al.* 2015] Jain, Gupta et Jain. *Estimation of sample size in dental research*. International Dental and Medical Journal of Advanced Research, vol. 1, page doi : 10.15713/ins.idmjar.9, 2015.

[Jan *et al.* 2003] Wretman Jan, Särndal Erik et Swensson Bengt. *Model Assisted Survey Sampling*. Springer Series in Statistics, 2003.

[Japec *et al.* 2015] Japec, Biemer Decker Lampe J.Lane C. O'Neil Kreuter Berg et Usher. *Big Data in survey research : AAPOR task force report*. Public Opin. Q., vol. 79, no. 4, page 839–880, 2015.

[Javaid 2013] Adeel Javaid. *Understanding Dijkstra Algorithm*. SSRN Electronic Journal, vol. doi : 10.2139/ssrn.2340905, 2013.

[Javed 2017] Ahmed Shaheen Javed. *Apache Kafka : Real Time Implementation with Kafka Architecture Review*. International Journal of Advanced Science and Technology, vol. 109, pages 35–42, http ://dx.doi.org/10.14257/ijast.2017.109.04, 2017.

[Jeong & CHA 2019] Hanjo Jeong et Kyung Jin CHA. *An Efficient MapReduce-Based Parallel Processing Framework for User-Based Collaborative Filtering*. symmetry, vol. 11, no. 6, page https ://doi.org/10.3390/sym11060748, 2019.

[John 1977] Tukey John. *Exploratory Data Analysis*. Pearson, London, 1977.

[Johnson 1988] Eric Johnson. *Completing an MIMD multiprocessor taxonomy*. ACM SIGARCH Computer Architecture News Homepage archive, vol. 16, no. 3, pages 44–47,doi>10.1145/48675.48682, 1988.

[Johnson 1989] Johnson. *GMMP Multiprocessor Architectures*. Proceedings, International Conference on Computing and Information, North-Holland, pages 187–191, 1989.

[Johnson 1991] Johnson. *A Global-Memory ,Message-Passing Multiprocessor*. Microprocessors and Microsystems, vol. 15, no. 8, pages 403–410, 1991.

[Johnson 2007] Johnson. *Technology, Indirect Observation Yield Insights*. Marketing News Journal, 2007.

[Jolliffe & Cadima 2016] Jolliffe et Cadima. *Principal component analysis : a review and recent developments*. Phil. Trans. R. Soc, vol. A 374 : 20150202, http ://dx.doi.org/10.1098/rsta.2015.0202, 2016.

[Junhee & Kang 2015] Seok Junhee et Yeong Seon Kang. *Mutual Information between Discrete Variables with Many Categories using Recursive Adaptive Partitioning*. Scientific Reports, vol. 5, page Article number : 10981, 2015.

[Kaashoek *et al.* 2013] Kaashoek, Madden Tu et Zeldovich. *Processing analytical queries over encrypted data*. Proc VLDB Endow, vol. 6, no. 5, page 289–300, 2013.

[Kalton 2017] Graham Kalton. *Systematic Sampling, Wiley StatsRef : Statistics Reference Online*. John Wiley Sons, Ltd, vol. DOI : 10.1002/9781118445112.stat03380.pub2, 2017.

[Kamal *et al.* 2017] Muhammad Kamal, Zahir Irani Sivarajah Uthayasankar et Vishanth Weerakkody. *Critical analysis of Big Data challenges and analytical methods*. Journal of Business Research, vol. 70, pages 263–286, https ://doi.org/10.1016/j.jbusres.2016.08.001, 2017.

[Kandel *et al.* 2012] Sean Kandel, Joseph Hellerstein Andreas Paepcke et Jeffrey Heer. *Enterprise data analysis and visualization : An interview study*. Visualization and Computer Graphics, IEEE Transactions, no. 2917–2926, 2012.

[Kanimozhi & Venkatesan 2015] Kanimozhi et Venkatesan. *Unstructured Data Analysis -A Survey*. International Journal of Advanced Research in Computer and Communication Engineering, vol. 4, no. 3, pages ISSN (Online) 2278–1021, 2015.

[Karim & Alla 2017] Rezaul Karim et Sridhar Alla. *Scala and Spark for Big Data Analytics : Explore the concepts of functional programming, data streaming, and machine learning*. first edition, 2017.

[Karypis *et al.* 1999a] Karypis, Han et Kumar. *Hierarchical clustering using dynamic modeling*. IEEE Computer, vol. 32, no. 8, page 68–75, 1999.

[Karypis *et al.* 1999b] George Karypis, Eui-Hong (Sam) Han et Vipin Kumar. *A Hierarchical Clustering Algorithm Using Dynamic Modeling*. Computer, vol. 32, no. 8, pages 68 –75, doi :10.1109/2.781637, 1999.

[Kaski 1997] Kaski. *Data exploration using self-organizing maps*. Thesis for the degree of Doctor of Technology, Helsinki University of Technology, Neural Networks Research Centre, 1997.

[Katal *et al.* 2013] Avita Katal, Mohammad Wazid et RH Goudar. *Big data : issues, challenges, tools and good practices*. In In Contemporary Computing (IC3) Sixth International Conference, page 404–409. IEEE, 2013.

[Kaufman & Rousseeuw 1990] Kaufman et Rousseeuw. *Finding groups in data : An introduction to cluster analysis*. New York : Wiley, 1990.

[Kaur & Oberai 2014] Gurneet Kaur et Neelam Oberai. *A review article on naive bayes classifier with various smoothing techniques*. International Journal of Computer Science and Mobile Computing, vol. 3, no. 10, pages 864–868, 2014.

[Kaur *et al.* 2017] Kaur, Chauhan et Chang. *Advancement and applicability of classifiers for variant exponential model to optimize the accuracy for deep learning*. Journal of Ambient Intelligence and Humanized Computing, vol. doi : 10.1007/s12652-017-0561-x, 2017.

[Kawulich 2005] Kawulich. *Participant Observation as a Data Collection Method*. FQS, vol. 6, no. 2, page Art. 43, 2005.

[Kayyali *et al.* 2013] Basel Kayyali, David Knott et Steve Van Kuiken. *The big-data revolution in us health care : Accelerating value and innovation*. Mc Kinsey Company, vol. 2, no. 8, page 1–13, 2013.

[Kelleher *et al.* 2015] John Kelleher, Brian Mac Namee et Aoife D'Arcy. *Fundamentals of machine learning for predictive data analytics Algorithms, Worked Examples, and Case Studies*. Massachusetts Institute of Technology, pages p49–50 and p.494–508, 2015.

[Kenobi *et al.* 2017] Kim Kenobi, Oorbessy Gaju Jayalath De Silva Jonathan Atkinson Darren Wells et John Foulkes. *Linear discriminant analysis reveals differences in root architecture in wheat seedlings related to nitrogen uptake efficiency*. Journal of Experimental Botany, vol. 68, no. 17, pages 4969–4981,https ://doi.org/10.1093/jxb/erx300, 2017.

[Kent & Williams 1992] Allen Kent et James Williams. *Encyclopedia of Computer Science and Technology*. vol. volume 26 - Supplement 11, 1992.

[Ketterlin & Clauss 2012] Alain Ketterlin et Philippe Clauss. *Profiling Data-Dependence to Assist Parallelization : Framework, Scope, and Optimization*. IEEE/ACM 45th Annual International Symposium on Microarchitecture, 2012.

[Khan *et al.* 2014] Muhammad Faheem Khan, Shahid Khan et Aurangzeb Khan. *Content based automatic classification of research articles*. Sci.Int. (Lahore), vol. 26, no. 5, pages 2495–2499, 2014.

[Khan *et al.* 2018] Nawsher Khan, Habib Shah, Gran Badsha, Aftab Ahmad Abbasi, Mohammed Alsaqer et Soulmaz Salehian. *10 Vs, Issues and Challenges of Big Data*. In International Conference on Big Data and Education ICBDE '18, pages 203–210. March 9–11, 2018, Honolulu, HI, USA, 2018.

[Khanali & Vaziri 2016] Hoda Khanali et Babak Vaziri. *A Survey on Clustering Algorithms for Partitioning Method*. International Journal of Computer Applications, vol. 155, no. 4, 2016.

[Khoussainov & Patel 2007] Rinat Khoussainov et Ahmed Patel. *Distributed Parallel Computing in Networks of Workstations A Survey Study*. vol. disponible sur :

https ://pdfs.semanticscholar.org/122a/fff7a1c6fc0383fb22c67e680affo8426e61.pdf, 2007.

[Kinkeldey *et al.* 2014] Kinkeldey, MacEachren et Schiewe. *How to assess visual communication of uncertainty ? A systematic review of geospatial uncertainty visualisation user studies.* The Cartographic Journal, vol. 51, no. 4, page 372–386. doi : 10.1179/1743277414Y.0000000099, 2014.

[Kinkeldey *et al.* 2017] Kinkeldey, Riveiro MacEachren et Schiewe. *Evaluating the effect of visually represented geodata uncertainty on decision-making : Systematic review, lessons learned, and recommendations*. Cartography and Geographic Information Science, vol. 44, no. 1, pages 1–2, doi : 10.1080/15230406.2015.1089792., 2017.

[Kiyak & Timus 2015] Erkan Kiyak et Oguz Timus. *Optimizing MLP Classifier and ECG Features for Sleep Apnea Detection*. Journal of Naval Science and Engineering, vol. 11, no. 1, pages 1–18, 2015.

[Kleiner *et al.* 2014] Ariel Kleiner, Purnamrita Sarkar Ameet Talwalkar et Michael I.Jordan. *A scalable bootstrap for massive data*. Journal of the Royal Statistical Society, vol. 76, no. 4, page 795–816, 2014.

[Kohonen & Simula 1996] Kohonen et Simula. *Engineering Applications of the SelfOrganizing Map*. In Proceeding of the IEEE, volume 84 of *10*, page 1354 – 1384, 1996.

[Kolehmainen 2004] Kolehmainen. *Data exploration with self-organizing maps in environmental informatics and bioinformatics*. Thesis for the degree of Doctor Science in of Technology, Helsinki University of Technology, Department of Computer Science and Engineering, Kuopio University Publications C. Natural and Environmental Sciences 167, 2004.

[Koppal 2017] Koppal. *Trends in Data Visualization*. vol. Available at : http ://www.labmanager.com/ask-the-expert/2017/05/trends-in-data-visualization.WXsTDISLSpo [Accessed 28 July 2017]., 2017.

[Kothari 2004] Kothari. *Research methodology : methods techniques*. New Delhi : New Age International (P) Ltd, 2004.

[Kotipalli & Suthaharan 2014] Kotipalli et Suthaharan. *Modeling of class imbalance using an empirical approach with spambase data set and random forest classification*. In Proceedings of the 3rd Annual Conference on Research in Information Technology, page 75–80. ACM, 2014.

[Kotsiantis 2007] Kotsiantis. *upervised machine learning : A review of classification techniques*. Informatica, vol. 31, page 249–268, 2007.

[Kowalik 1995] Kowalik. *Scalability of Parallel Systems : Efficiency Versus Execution Time*. Advances in Parallel Computing, vol. 10, pages 39–47, 1995.

[Kozak 2008] Marcin Kozak. *Finite and Infinite Populations in Biological Statistics : Should We Distinguish Them*. The Journal of American Science, vol. 4, no. 1, pages 59–62, ISSN : 1545–1003, 2008.

[Krause & Lipscomb 2016] Andy Krause et Clifford Lipscomb. *The Data Preparation Process in Real Estate : Guidance and Review*. Journal of Real Estate Practice and Education In Press, vol. 1, 2016.

[KSteven 2012] Thompson KSteven. *Sampling, Third Edition, chapter 6 : Unequal Probability Sampling*. John Wiley Sons, Inc., 2012.

[Kubena *et al.* 1992] Glenn Kubena, Kenneth Liao et Larry Roberts. *White Paper on Massively Parallel Programming Languages*. IBM, 1992.

[Kulkarni & Khandewal 2014] Amogh Pramod Kulkarni et Mahesh Khandewal. *Survey on Hadoop and Introduction to YARN*. International Journal of Emerging Technology and Advanced Engineering, vol. 4, no. 5, pages ISSN 2250–2459, 2014.

[Kulkarni *et al.* 2012] Gurudatt Kulkarni, Jayant Gambhir et Rajnikant Palwe. *Cloud Computing-Software as Service*. International Journal of Computer Science Information Technology Research Excellence, vol. 2, no. 1, 2012.

[Kumar & Wu 2007] Kumar et Wu. *Survey Paper on Top 10 Algorithms in Data Mining*. London : Springer-Verlag Limited, 2007.

[Kumar *et al.* 1994] Kumar, Gupta Grama et Karypis. *Designing and building parallel programs : Concepts and tools for parallel software engineering*. IEEE Computer Graphics and Applications, vol. 14, no. 4, pages 33–40, 1994.

[Kuperman & Kazunaga Matsuki 2016] Victor Kuperman et and Julie A Van Dyke Kazunaga Matsuki. *The Random Forests statistical technique : An examination of its value for the study of reading*. Scientific Studies of Reading, vol. 20, no. 1, page doi : 10.1080/10888438.2015.1107073, 2016.

[Labrinidis & Jagadish 2012] Labrinidis et Jagadish. *H.V : Challenges and opportunities with Big Data*. Proc. VLDB Endowment, vol. 5, no. 12, page 2032–2033, 2012.

[LakKang & Timothy 2017] Song LakKang et Sliwinski Timothy. *Applying Parallel Computing Techniques to Analyze Terabyte Atmospheric Boundary Layer Model Outputs*. Published by Elsevier Inc, vol. http ://dx.doi.org/10.1016/j.bdr.2017.01.001, 2017.

[Lakshminarayan *et al.* 1996] Lakshminarayan, Goldman Harp et Samad. *Imputation of missing data using machine learning techniques*. In KDD-96 Proceedings, AAAI, volume 4792, pages 140–145, Available at : http ://www.aaai.org/Papers/KDD/1996/KDD96–023.pdf, 1996.

[Langford *et al.* 2000] Langford, Tenenbaum et de Silva. *A global geometric framework for nonlinear dimensionality reduction*. Science, vol. 290, no. 5500, page 2319–2323, 2000.

[Leahy & Wu 1993] Leahy et Wu. *An approximation method of evaluating the joint likelihood for first-order GMRFs*. IEEE Transactions on Image Processing, vol. 2, no. 4, page 520–523, 1993.

[Lencina *et al.* 2004] Lencina, Singer et Stanek. *A unified approach to estimation and predition under simple random sampling*. Journal of Statistical Planning and Inference, vol. 121, page 325–338, 2004.

[Lessler & Kalsbeek 1992] Lessler et Kalsbeek. *Nonsampling Error in Surveys*. New York : Wiley Interscience, 1992.

[Lewis & Catlett 1994] David Lewis et Jason Catlett. *Heterogeneous uncertainty sampling for supervised learning*. In Proceedings of the 11th international conference on machine learning (ICML'94), pages 148–156. http ://doi.org/10.1016/B978–1– 55860–335–6.50026–X, 1994.

[Li *et al.* 2012] Li, Huang Chen et Feng. *Scalable random forests for massive data*. (Eds) : PAKDD 2012, Part I, LNAI 7301, page 135–146, 2012.

[Liaw & Wiener 2002a] Andy Liaw et Matthew Wiener. *Classification and Regression by randomForest*. vol. 2/3, pages ISSN 1609–3631, 2002.

[Liaw & Wiener 2002b] Andy Liaw et Matthew Wiener. *Classification and Regression by randomForest*. R News : ISSN 1609-3631, vol. 2/3, 2002.

[Lin & Lee 1991] Lin et Lee. *Fault-tolerant reconfigurable architecture for robot kinematics and dynamics computations*. IEEE Transactions on Systems, Man, and Cybernetics, vol. 21, no. 5, page doi : 10.1109/21.120051, 1991.

[Little & Rubin 2002] Little et Rubin. *Statistical analysis with missing data," Wiley Series in Probability and Statistics*. John Wiley and Sons, Inc. second edition, 2002.

[Liu & Xu 2009] Liu et Xu. *Configuring Clark-Wilson integrity model to enforce flexible protection*. In Proceedings of the International Conference on Computational Intelligence and Security (CIS '09), volume 2, page 15–20, 2009.

[Luxburg 2006] Von Luxburg. *A Tutorial on Spectral Clustering, Max Planck Institute for Biological Cybernetics*. 2006.

[Lyman *et al.* 2016] Lyman, Strygin Varian Dunn et Swearingen. *How much information ?* Counting-the-Numbers, vol. 6, no. 2, 2016.

[MacInnis *et al.* 2018] Bo MacInnis, Annabell S Ho Jon A Krosnick et Mu Jung Cho. *The Accuracy of Measurements with Probability and Nonprobability Survey Samples : Replication and Extension*. Public Opinion Quarterly, vol. 82, no. 4, pages 707–744, https ://doi.org/10.1093/poq/nfy038, 2018.

[MacQueen 1967] MacQueen. *Some methods for classication and analysis of multivariate observations*. In Proceedings of the fth Berkeley symposium on mathematical statistics and probability, Oakland, CA, USA, volume 1, pages 281–297, 1967.

[Madhusudanan *et al.* 2017] Madhusudanan, Gayathiri Subramanian Sivakumar et Ponnuramu. *Reducing the network traffic and handover the corresponding request using big data Hadoop*. International Journal for Scientific Research and Development, vol. 5, no. 1, 2017.

[Maheshwari 1996] Piyush Maheshwari. *Improving granularity and locality of data in multiprocessor execution of functional programs*. Parallel Computing, vol. 22, pages 1359– 1372, 1996.

[Malik & Shi 2000] Malik et Shi. *Normalized Cuts and Image Segmentation*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 8, page 888–905, 2000.

[Manju & Punithavalli 2011] Manju et Punithavalli. *An Analysis of Q-Learning Algorithms with Strategies of Reward Function*. International Journal on Computer Science and Engineering, vol. 3, no. 2, 2011.

[Mannor *et al.* 2009] Mannor, Yu et Shimkin. *Markov decision processes with arbitrary reward processes*. Math. Oper. Res, vol. 34, no. 3, page 737–757, 2009.

[Mar *et al.* 2009] Hall Mar, Frank Eibe, Reutemann Peter Holmes Geoffrey Pfahringer Bernhard et H. Witten Ian. *The WEKA Data Mining Software : An Update*. SIGKDD Explorations, vol. 11, no. 1, 2009.

[Marr 2015] Bernard Marr. *A brief history of big data everyone should read*. World Economic forum ,Business and data analytics, 2015.

[Marshall 1980] Yovits Marshall. *Advances in Computers*. academic press, vol. 19, page 15, 1980.

[Matei & Mosharaf 2010] Matei et Mosharaf. *Spark : cluster computing with working sets*. Proceeding HotCloud'10 Proceedings of the 2nd USENIX conference on Hot topics in cloud computing, pages 10–10, 2010.

[Matei & Reynold 2016] Matei et Reynold. *Apache Spark : A Unified Engine for Big Data Processing*. Communications of the ACM, vol. 59, no. 11, pages 56–65, 2016.

[Matula 1977] Matula. *Graph Theoretic Techniques for Cluster Analysis Algorithms*. In J. V. Ryzin, editor, Classification and Clustering, vol. Academic Press New York, 1977.

[Mayya *et al.* 2017] Shreemathi Mayya, Ashma Monteiro et Sachit Ganapathy. *Types of biological variables*. J Thorac Dis, vol. 9, no. 6, page 1730–1733. doi : 10.21037/jtd.2017.05.75, 2017.

[McClelland & Rumelhart 1985] James McClelland et David Rumelhart. *Distributed Memory and the Representation of General and Specific Information*. Journal of Experimental Psychology : General, vol. 114, no. 2, pages 159–188, 1985.

[Melekhova & Vinnikov 2015] Anna Melekhova et Vladimir Vinnikov. *Cloud and Grid Part I : Difference and Convergence*. Indian Journal of Science and Technology, vol. 8, no. 29, page doi : 10.17485/ijst/2015/v8i29/86860., 2015.

[Meng *et al.* 2015] Xiangrui Meng, Evan Sparks Shivaram Venkataraman Davies Liu Jeremy Freeman DB Tsai Manish Amde Sean Owen Doris Xin Reynold Xin Michael J.Franklin Reza Zadeh Matei Zaharia Joseph Bradley Burak Yavuz et Ameet Talwalkar. *MLlib : Machine Learning in Apache Spark*. Published in ArXiv, vol. doi :1505.06807, 2015.

[Meng *et al.* 2016] X Meng, Sparks Venkataraman Liu Freeman Tsai Amde Bradley Yuvaz et Owen. *Mllib : Machine learning in apache spark*. Journal of Machine Learning Research, vol. 17, no. 34, page 1–7, 2016.

[Meyer 2015] Meyer. *Support Vector Machines*. The Interface to libsvm in package e1071, 2015.

[Mila 2018] Steele Mila. *The SAGE Handbook of Qualitative Data Collection*. Book,ISBN 978-1-4739-5213-3, pages 314–322 and 587–589, 2018.

[Miller 1971] Miller. *The assault on privacy : computers, data banks, and dossiers*. University of Michigan Press, 1971.

[Mills *et al.* 2012] Mills, Lucas, Irakliotis, Ruppa, Carlson et Perlowitz. *Demystifying Big Data : A Practical Guide to Transforming the Business of Government*. Washington : TechAmerica Foundation, 2012.

[Mingyuan 2013] Zhong Mingyuan. *Value Function Approximation Methods for Linearly-solvable Markov Decision Process*. Thesis (Ph.D.)-University of Washington, 2013.

[MinYou & Wei 1995] Wu MinYou et Shu Wei. *Asynchronous Problems on SIMD Parallel Computers*. IEEE transactions on parallel and distributed systems, vol. 6, no. 7, page doi : 10.1109/71.395399, 1995.

[Mittal & Suri 2012] Sumit Mittal et Suri. *A Comparative Study of various Computing Processing Environments : A Review*. International Journal of Computer Science and Information Technologies, vol. 3, no. 5, page 5215 – 5218, 2012.

[Mohamed 2017] Amr Mohamed. *Comparative Study of Four Supervised Machine Learning Techniques for Classification*. International Journal of Applied Science and Technology, vol. 7, no. 2, 2017.

[Mohan & Pramod 2017] Anuraj Mohan et Venkatesanb Pramod. *A Scalable Method for Link Prediction in Large Real World Networks*. Journal of Parallel and Distributed Computing, vol. doi : http ://dx.doi.org/10.1016/j.jpdc.2017.05.009, 2017.

[Mois 2016] Martin Mois. *Apache Hadoop*. Tutorial, Copyright (c) Exelixis Media P.C, vol. disponible sur : http ://index-of.es/Varios-2/Apache-Hadoop-Tutorial.pdf, 2016.

[Moorley & Shorten 2014] Moorley et Shorten. *Selecting the sample*. Evidence-Based Nursing, vol. doi :10.1136/eb-2014- 101747, 2014.

[Moreno *et al.* 2016] Julio Moreno, Manuel Serrano et Eduardo Fernández-Medina. *Main Issues in Big Data Security*. Future Internet, vol. 8, no. 44, page doi :10.3390/fi8030044, 2016.

[MORRIS 2010] HUGH MORRIS. *Tree pruning : a modern approach*. IDS Yearbook, pages 217–225, 2010.

[Mulaik 2010] Stanley Mulaik. *Foundations Of Factor Analysis*. Second Edition by Taylor and Francis Group, LLC, 2010.

[Murata *et al.* 2015] Murata, T Yamashita Y Yamauchi Wakayama A Kimura et H Fujiyoshi. *Distributed forests for MapReduce-based machine learning*. 3rd IAPR Asian Conference on Pattern Recognition (ACPR), vol. doi :10.1109/ACPR.2015.7486509, page 276–280, 2015.

[Murphy *et al.* 2015] Mariam Kiran Peter Murphy, Inder Monga et Jon Dugan Sartaj Singh Baveja. *Lambda Architecture for Cost-effective Batch and Speed Big Data processing*. IEEE International Conference on Big Data, vol. doi : 10.1109/BigData.2015.7364082, 2015.

[Myles & Gin 2000] Myles et Gin. *Statistical methods for anaesthesia and intensive care*. Butterworth-Heinemann, Waltham, Massachusetts, 2000.

[Najjara *et al.* 1999] Walid Najjara, Edward Leeb et Guang Gao. *Advances in the dataflow computational model*. Parallel Computing, vol. 25, 1999.

[Nan *et al.* 2016] Feng Nan, Joseph Wang et Venkatesh Saligrama. *Pruning Random Forests for Prediction on a Budget*. arXiv[stat.ML], vol. 16, page doi>1606.05060v1, 2016.

[Ng & Han 1994] Ng et Han. *Efficient and effective clustering method for spatial data mining*. In Proceedings 1994 Int. Conf. Very Large Data Bases (VLDB'94),, pages 144–155, Santiago, Chile., 1994.

[Ng *et al.* 2001] Ng, Jordan et Weiss. *On spectral clustering : Analysis and an algorithm*. Advances in Neural Information Processing Systems, vol. 14, page 849–856. MIT Press., 2001.

[Nguyen *et al.* 2019] Dan Nguyen, Weiguo Lu Xuejun Gu Zohaib Iqbal Troy Long Xun Jia et Steve Jiang. *A feasibility study for predicting optimal radiation therapy dose distributions of prostate cancer patients from patient anatomy using deep learning*. Scientific Reports, vol. 9, page Article number : 1076, 2019.

[NIST 2015] NIST. *Definitions and Taxonomies Subgroup*. National Institute of Standards and Technology, vol. 1, http ://dx.doi.org/10.6028/NIST.SP.1500-1,, 2015.

[Nongxa 2017] Loyiso Nongxa. *Mathematical and statistical foundations and challenges of (big) data sciences*. South African Journal of Science, vol. 113, no. 3-4, pages ISSN 1996–7489,http ://dx.doi.org/10.17159/sajs.2017/a0200, 2017.

[Nutini 2017] Julie Nutini. *Monte Carlo Methods (Estimators, On-policy/Off-policy Learning)*. MLRG - Winter Term 2, 2017.

[Nyumba *et al.* 2018] Nyumba, Christina Derrick Kerrie Wilson et Nibedita Mukherjee. *The use of focus group discussion methodology : Insights from two decades of application in conservation*. Methods in Ecology and Evolution, vol. 9, pages 20–32, DOI : 10.1111/2041–210X.12860, 2018.

[Oaks & Wong 2004] Scott Oaks et Henry Wong. *Java Threads : Understanding and Mastering Concurrent Programming*. 3rd edition, 2004.

[Oaks 2012] Oaks. *P. Sampling, P. Guidelines, M.S. Choices*. CHAPTER 5, page 125–174, 2012.

[Obiniyi *et al.* 2015] Afolayan Obiniyi, Charles Saidu et Peter Ogedebe. *Overview of Trends Leading to Parallel Computing and Parallel Programming*. British Journal of Mathematics Computer Science, vol. 7, no. 1, pages 40–57, 2015.

[Odersky *et al.* 2016] Martin Odersky, Lex Spoon et Bill Venners. *Programming in Scala : A comprehensive step-by-step guide*. 3rd edition, 2016.

[Ofori & Islam 2019] Abel Yeboah Ofori et Shareeful Islam. *Cyber Security Threat Modeling for Supply Chain Organizational Environments*. Future Internet, vol. 11, no. 63, page doi :10.3390/fi11030063, 2019.

[Ohno *et al.* 2011] Kazuhiko Ohno, Takahiro Sasaki Dai Michiura Masaki Matsumoto et Toshio Kondo. *A GPGPU Programming Framework based on a Shared-Memory Model*. Proceeding (757) Parallel and Distributed Computing and Systems, pages doi : 10.2316/P.2011.757–097, 2011.

[Okororie & Otuonye 2015] Chigozie Okororie et Eric L Otuonye. *Efficiency of some sampling techniques*. Journal of Scientific Research and Studies, vol. 2, no. 3, pages 63–69, 2015.

[Oliphant 2007] Travis Oliphant. *SciPy : Open source scientific tools for Python*. Computing in Science and Engineering, vol. 9, no. 1, pages 10–20, 2007.

[Olshannikova *et al.* 2015] Ekaterina Olshannikova, Yevgeni Koucheryavy Aleksandr Ometov et Thomas Olsson. *Visualizing Big Data with Augmented and virtual reality : challenges and research agenda*. Journal of Big Data, vol. 2, no. 22, 2015.

[OMI 2012] Tom White OMI. *Hadoop : The definitive guide*. O'Reilly Media, Inc, 3rd Edition, 2012.

[Omran *et al.* 2007] Mahamed Omran, Andries Engelbrecht et Ayed Salman. *An overview of clustering methods*. Intelligent Data Analysis, vol. doi : 10.3233/IDA-2007-11602, 2007.

[Ossa 2017] Chinedu Ossa. *Integrated Big Data Analytics Technique for Real-Time Prognostics, Fault Detection and Identification for Complex Systems*. Infrastructures, vol. 2, no. 20, page doi :10.3390/infrastructures2040020, 2017.

[Pacheco 2018] Alexander Pacheco. *Parallel Programming Concepts*. tutorial, Research Computing, 2018.

[Padilla *et al.* 2018] Lace Padilla, Mary Hegarty Sarah Creem-Regehr et Jeanine Stefanucci. *Decision making with visualizations : a cognitive framework across disciplines*. Cogn Res Princ Implic, vol. doi : 10.1186/s41235-018-0120-9, 2018.

[Palliotti *et al.* 2017] Alberto Palliotti, Paolo Sabbatini Juan Guillermo Cruz-Castillo Tommaso Frioni Sergio Tombesi et Vania Lanari. *Double-Pruning Grapevines as a Management Tool to 3 Delay Berry Ripening and Control Yield*. American Journal of Enology and Viticulture, vol. doi : 10.5344/ajev.2017.17011, 2017.

[Pandya & Saket 2016] Sharnil Pandya et Swarndeep Saket. *Overview of Partitioning Algorithms in Clustering Techniques*. International Journal of Advanced Research in Computer Engineering and Technology, vol. 5, no. 6, 2016.

[Panfilova & Salibekyan 2014] Peter Panfilova et Sergey Salibekyan. *Dataflow Computing and its Impact on Automation Applications*. Procedia Engineering, vol. 69, pages 1286–1295, 2014.

[Paradis *et al.* 2016] Elise Paradis, Glen Bandiera Bridget O'Brien Laura Nimmon et Maria Athina. *Selection of Data Collection Methods*. J Grad Med Educ, vol. 8, no. 2, pages 263–264, doi : 10.4300/JGME–D–16–00098.1, 2016.

[Parhami 2002] Behrooz Parhami. *Introduction to Parallel Processing :Algorithms and Architectures*. Kluwer Academic Publishers, 2002.

[Parrella 2007] Francesco Parrella. *Online Support Vector Machines for Regression*. Thesis presented for the degree of Information Science, University of Genoa,Italy, 2007.

[Patel 2017] Hiren Patel. *HBase : A NoSQL Database*. Technical Report, page doi : 10.13140/RG.2.2.22974.28480, 2017.

[Pattnaik & Mishra 2016] Pattnaik et Mishra. *Introduction to big data analysis*. In : Techniques and Environments for Big Data Analysis, vol. Springer ,doi :10.1007/978-3-319-27520-8, page 1–20, 2016.

[Paxton *et al.* 2001] Pamela Paxton, Kenneth Bollen Patrick Curran et Jim Kirby. *Monte Carlo Experiments : Design and Implementation*. Structural equation modeling, vol. 8, no. 2, page 287–312, 2001.

[P.Bertsekas 2007] D. P.Bertsekas. *Dynamic Programming and Optimal Control*. Athena Scientific, Belmont, MA, 3 edition, vol. 2, b 2007.

[Pedregosa *et al.* 2011] Pedregosa, Varoquaux et Gramfort. *Scikit-learn : Machine Learning in Python*. Journal of Machine Learning Research, vol. 12, page 2825–2830, 2011.

[Peter 1976] Smith Peter. *The Foundations of Survey Sampling : a Review*. Journal of the Royal Statistical Society, vol. 139, no. 2, pages 183–204, 1976.

[Philip 2018] Thomas Philip. *Reinforcement Learning - Syllabus, Notes, and Assignments*. Fall 2018, vol. University of Massachusetts Amherst, 2018.

[Pine 2019] David Pine. *Introduction to Python for Science and Engineering*. 2019.

[Plaisant 2004] Plaisant. *The Challenge of Information Visualization Evaluation*. In Proceedings of AVI 2004 : 6th International Conference on Advanced Visual Interfaces, pages 109–116. ACM Press, New York, 2004.

[Poddar *et al.* 2016] Poddar, Boelter et Arx. *A Strongly Encrypted Database System : Cryptology*. ePrint Archive, Report 2016-591, 2016.

[Poornima & Pushpalatha 2016] Poornima et Pushpalatha. *A journey from big data towards prescriptive analytics*. arpn Journal of Engineering and Applied Sciences, vol. 11, no. 19, pages ISSN 1819–6608, 2016.

[Pop *et al.* 2017] Florin Pop, Ciprian Dobre et Alexandru Costan. *AutoCompBD : Autonomic Computing and Big Data platforms*. Soft Comput, vol. 21, pages 4497–4499 DOI 10.1007/s00500–017–2739–8, 2017.

[Popa *et al.* 2011] Popa, Zeldovich Redfield et Balakrishnan Cryptdb. *Protecting confidentiality with encrypted query processing*. In Proceedings of the Twenty-Third ACM Symposium on Operating Systems Principles. SOSP '11, page 85–100. ACM, 2011.

[Popov 2017] Aleksey Popov. *An Introduction to the MISD Technology*. Proceedings of the 50th Hawaii International Conference on System Sciences., 2017.

[Power 2014] Power. *Using 'Big Data' for analytics and decision support*. J. Decis. Syst, vol. 23, no. 2, page 222–228, 2014.

[Prabhat & Khullar 2017] Anjuman Prabhat et Vikas Khullar. *Sentiment classification on big data using Naïve bayes and logistic regression*. International Conference on Computer Communication and Informatics, page doi : 10.1109/ICCCI.2017.8117734, 2017.

[Prakash & Atul 2016] Verma Prakash et Patel Atul. *Comparison of MapReduce and Spark Programming Frameworks for Big Data Analytics on HDFS*. International Journal of Computer Science Communication, vol. 7, no. 2, pages 80–84, 2016.

[Pratxa & Xing 2011] Guillem Pratxa et Lei Xing. *GPU computing in medical physics : A review*. Med. Phys, vol. 38, no. 5, 2011.

[Priyadharshini & Parvathi 2012] Priyadharshini et Parvathi. *Data integrity in cloud storage*. In Proceedings of the 1st International Conference on Advances in Engineering, Science and Management (ICAESM '12), page 261–265, 2012.

[Puech *et al.* 2014] Pauline Lardin Puech, Herve Cardot et Camelia Goga. *Analysing large datasets of functional data : a survey sampling point of view*. Journal de la Société Française de Statistique, vol. 155, no. 4, 2014.

[Puterman 1994] Puterman. *Markov Decision Processes—Discrete Stochastic Dynamic Programming*. John Wiley Sons, Inc., New York, NY. 1, 1994.

[Qiu *et al.* 2016] Qiu, Xu Wu Ding et Feng. *A Survey of Machine Learning for Big Data Processing*. EURASIP Journal on Advances in Signal Processing, vol. 67, page 1–16, 2016.

[Rabuñal & Dorado 2006] Juan Rabuñal et Julián Dorado. *Artificial Neural Networks in Real-life Applications*. Idea Group Inc, 2006.

[Rahul & Pravin 2016] Vanve Rahul et Patil Pravin. *A survey on : Predictive Analytics For Credit Risk Assessment*. International Research Journal of Engineering and Technology (IRJET), vol. 3, pages e–ISSN : 2395 –0056, 2016.

[Rajaraman & Murthy 2016] Rajaraman et Siva Ram Murthy. *Parallel Computers Architecture and Programming*. 2nd Edition, pages 100–102, 2016.

[Ranney 2015] Ranney. *Interview-based Qualitative Research in Emergency Care Part II : Data collection, Analysis and Results Reporting*. Academic Emergency Medicine, vol. 22, pages 1103–1112, 2015.

[Raschka 2015] Sebastian Raschka. *Python Machine Learning'*. pages 32–41, 2015.

[Rathore *et al.* 2016] Mazhar Rathore, Awais Ahmad et Anand Paul. *Real time intrusion detection system for ultra-high-speed big data environments*. Journal of Supercomputing, vol. 72, pages 3489–3510, doi :10.1007/s11227–015–1615–5, 2016.

[Richard & Barto 1998] Sutton Richard et Andrew Barto. *Reinforcement Learning*. MIT Press, 1998.

[Richard 1988] Sutton Richard. *Learning to predict by the methods of temporal differences*. Machine Learning, vol. 3, pages 9–44, 1988.

[Richard 2001] Sutton Richard. *Reinforcement Learning Architectures*. GTE Laboratories Incorporated, Waltham, MA 02254, 2001.

[Richardson *et al.* 2006] Richardson, Prakash et Brill. *Efficient Selection of the Most Similar Image in a Database for Critical Structures Segmentation*. In Proceedings of the 15th international conference on World Wide Web, WWW (L. Carr, D. De Roure, A. Iyengar, C.A. Goble, and M. Dahlin, eds, page 707–715. ACM, 2006.

[Rio *et al.* 2014] Rio, Benitez Lopez et Herrera. *On the use of MapReduce for imbalanced big data using random forest*. Information Sciences - Journal – Elsevier, vol. http ://dx.doi.org/10.1016/j.ins.2014.03.043, 2014.

[Ripon & Arif 2016] Ripon et Arif. *Big Data : The V's of the Game Changer Paradigm*. In IEEE 18th International Conference on High Performance Computing and Communications ; IEEE 14th International Conference on Smart City ; IEEE 2nd International Conference on Data Science and Systems, volume DOI 10.1109/HPCC-SmartCity-DSS.2016.8. IEEE, 2016.

[Roberto 2014] Zicari Roberto. *Big data : Challenges and opportunities*. big data and cloud computing, 2014.

[Robson 2002] Robson. *Real World Research*. UK : Blackwell Publishing, vol. (2nd), 2002.

[Robson 2011] Robson. *Real World Research : A Resource for Users of Social Research Methods in Applied Settings*. Chichester : John Wiley, vol. (3rd edn), 2011.

[Rokach & Maimon 2005] Rokach et Maimon. *Top – Down Induction of Decision Trees Classifiers – A Survey*. IEEE Transactions on Systems, vol. 31, no. 4, 2005.

[Roos *et al.* 2013] Roos, Deutsch Corrigan Zikopoulos Parasuraman et Giles. *Harness the Power of Big Data : The IBM Big Data Platform*. New York : McGraw-Hill, 2013.

[Rose *et al.* 2017] Rose, Mathiason Berndtsson et Larsson. *The advanced analytics jumpstart : definition, process model, best practices*. JISTEM - Journal of Information Systems and Technology Management, vol. 14, no. 3, pages ISSN 1807–1775,http ://dx.doi.org/10.4301/s1807–17752017000300003, 2017.

[Rusdiana 2016] Siti Rusdiana. *Linear Discriminant Analysis to Clarifying Vacant Land Allocations of Baiturraman District in Banda Aceh City*. Proceedings, AIP Conference, vol. doi : 10.1063/1.4965191, 2016.

[Saleema *et al.* 2014] Saleema, deepa shenoy venugopal Bhagawathi monica et patnaik. *Cancer prognosis prediction using balanced stratified sampling, international journal on soft computing*. International Journal on Soft Computing, Artificial Intelligence and Applications, vol. 3, no. 1, 2014.

[Sang *et al.* 2014] Sang, Zeng et Tong. *Study on multicenter fuzzy C-means algorithm based on transitive closure and spectral clustering*. Applied Soft Computing, vol. 16, pages 89–101 , https ://doi.org/10.1016/j.asoc.2013.11.020, 2014.

[Saul & Weinberger 2004] Saul et Weinberger. *Unsupervised learning of image manifolds by semi definite programming*. In Proceeding of IEEE Conference on Computer Vision and Pattern Recognition, page 988–995, 2004.

[Saul & Weinberger 2006] Saul et Weinberger. *Unsupervised learning of image manifolds by semi definite programming*. International Journal of Computer Vision, vol. 70, no. 1, page 77–90, 2006.

[Saunders 2012] Saunders. *versus Mail : The Influence of Survey Distribution Mode on Employees' Response*. Field Methods, vol. 24, no. 1, pages 56–73, 2012.

[Schmueli & Koppius 2011] Schmueli et Koppius. *Predictive analytics in information systems research*. MIS Quarterly, vol. 35, pages 553–572, 2011.

[Schnabel 1985] Schnabel. *Parallel Computing in Optimization, Computational Mathematical Programming*. NATO ASI Series (Series F : Computer and Systems Sciences),Springer, Berlin, Heidelberg, vol. 15, pages doi : https ://doi.org/10.1007/978–3–642–82450–$o_1$3, 1985.

[W0093]minitocSome "*.mld" or "*.mlo" files are missing in your installation. Search for the I0050 and I0051 info messages in the \jobname.log file. The full list of the missing language files is given in the W0094 warning message. Please install the missing files from a recent distribution or from the CTAN archives[W0094]minitocMissing minitoc language file(s):