N° d'ordre :

REPUBLIQUE ALGERIENNE DEMOCRATIQUE & POPULAIRE

MINISTERE DE L'ENSEIGNEMENT SUPERIEUR & DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE DJILLALI LIABES - SIDI BEL ABBES

FACULTE DES SCIENCES EXACTES

DEPARTEMENT D'INFORMATIQUE

THESE DE DOCTORAT LMD

Présentée et soutenue par

Salheddine KABOU

La gestion de la confidentialité dans le Cloud Computing

Dirigée par :Pr. Sidi Mohamed BENSLIMANE
Professeur Ecole Supérieur en Informatique, SBA.

Soutenue le --/--/2017, devant le jury composé de :

Président:

Examinateur:

Examinateur:

Examinateur:

Examinateur:

Année universitaire : 2016-2017

« *L'âme relève de l'ordre de mon Seigneur et, en fait de science, vous n'avez reçu que peu de chose***».** Dieu honnête El Adhim.

DÉDICACE

Je dédie ce modeste travail à tous ceux qui me sont chers au cœur.

À ma défunte mère,

À mon père,

À tous ceux qui sentent participant dans ma réussite, et à toute personne qui reconnaîtra son empreinte sur ce modeste travail,

Je dédie ce travail.

Que la paix d'ALLAH soit avec tous…Que dieux nous réunisse dans son vaste paradis inch-ALLAH.

Remerciements

« Si j'ai pu voir aussi loin, c'est que j'étais debout sur des épaules de géants »

Issac Newton

Je suis loin d'avoir vu aussi loin que Newton, n'empêche que j'ai mes géants à qui je tiens à exprimer ma plus profonde reconnaissance.

Mes plus sincères remerciements vont :

À tous les membres du jury qui m'ont fait l'honneur de prendre mon modeste travail en considération et en suite de le juger.

Mon directeur de thèse, **Benslimane Sidi Mohamed** pour son soutien moral et scientifique efficace et constant, pour sa disponibilité et son écoute.

Je tiens à remercier tous ceux qui m'ont prêté main forte, et ceux qui ont contribué de près ou de loin à l'élaboration de ce travail.

RESUMÉ

Le Cloud Computing représente un nouveau modèle d'entreprise qui assure le partage de ressources informatiques contenant des informations personnelles à travers plusieurs bases de données distribuées et privées. La confidentialité et la sécurité sont les principaux obstacles qui empêchent l'adoption extensive de cette nouvelle technologie. La confidentialité de ces informations doit être préservée avant la publication dans le Cloud, c'est à dire aucune information sensible ne doit être divulguée. L'anonymisation des données est l'une des solutions qui peuvent être utilisées pour préserver la confidentialité des données tout en assurant leurs utilisations. La plus part des travaux utilisent le fameux modèle K-anonymat pour préserver la confidentialité des objets des données. Parmi les principaux inconvénients de ce modèle est qu'il ne préserve pas l'utilité de données. Dans ce travail de thèse, nous avons développé un nouveau protocole d'anonymisation distribué horizontalement pour la préservation de la confidentialité des données (individus), et la confidentialité des fournisseurs des données dans un environnement Cloud. Pour garantir la confidentialité des individus, nous avons utilisé le modèle *K-concealment* qui assure un haut niveau de sécurité et minimise la perte des données. Pour soutenir la confidentialité des fournisseurs de données, nous avons conçu un nouvel algorithme d'anonymisation distribué utilisant la structure de l'arbre- R* qui fournit une meilleur utilité de données. Les résultats des expérimentations, ont montré que notre approche fournit une meilleure utilité de données généralisées par rapport aux approches d'anonymisation centralisées.

Abstract

Cloud computing represents a new business model which enables diverse benefits such as the sharing of computing resources containing personal information across multiple distributed and private databases. However, privacy and security concerns are a significant obstacle that is preventing the extensive adoption of this new technology. The confidentiality of the personal information must be preserved before outsourcing to the commercial public cloud, i.e. any sensitive information should not be disclosed. Data anonymization is one of the solutions methods that can be used to preserve the privacy of data while still allowing the data to be used. Most of the existing works use a k-anonymity model for preserving privacy for data subject that offers lower utility. In this thesis, we have developed a new horizontally distributed anonymization protocol for privacy-preserving data subjects (individuals) and data providers in a cloud environment. For the privacy of data subjects, we have used a k-concealment model that offers a high level of security and the amount of information loss is minimized. For the privacy of data providers, we have adopted a new distributed anonymization algorithm that uses an R*-tree structure which provides better generalization. As demonstrated by our experiments, our approach provides better utility of generalized data compared to centralized anonymization approaches.

ملخص

تمثل الحوسبة السحابية نموذج جديد للعمل، حيث تمكن من فوائد متنوعة مثل تقاسم موارد الحوسبة والتي تحتوي على المعلومات الشخصية عبر قواعد بيانات موزعة، خاصة و متعددة. غير أن المخاوف المتعلقة بالخصوصية و الأمن هي العقبة الكبيرة التي تمنع من اعتماد واسع النطاق لهذه التكنولوجيا الجديدة. المحافظة على سرية المعلومات الشخصية، يجب أن تكون قبل الاستعانة بمصادر خارجية للسحابة العامة التجارية، بمعنى أن لا يجب الكشف عن أية معلومات حساسة. إخفاء هوية البيانات، هي واحدة من الحلول التي يمكن استخدامها للحفاظ على خصوصية البيانات مع ضمان استخدامها. معظم الأعمال الحالية تستخدم النموذج k-anonymity للحفاظ على خصوصية البيانات و التي من بين عوائقها الرئيسية عدم المحافظة على فائدة البيانات. في هذه الأطروحة، قمنا بتطوير بروتوكول موزع أفقيا لإخفاء الهوية و الحفاظ على خصوصية بيانات الأفراد و خصوصية مقدمي البيانات في بيئة عمل سحابية. لضمان خصوصية الأفراد، استخدمنا النموذج k-concealment الذي يضمن مستوى عال من الأمن و يقلل من فقدان البيانات. و لضمان خصوصية مقدمي البيانات، قمنا بتصميم خوارزمية موزعة لإخفاء الهوية باستخدام هيكل الشجرة R* و التي توفر أفضلية في التعميم. كما تبين من تجاربنا، فإن طريقتنا توفر أفضل فائدة للبيانات المخفية إذا ما قارناها بالطريقة المركزية.

TABLE DES MATIERES

Table des Figures

Liste des tableaux

Liste des algorithmes

Liste des abriviations

TI	Technologie d'Information
IPI	Informations Personnelles d'Individu
CHP	Le Calcul de Haute Performance
ASP	Application Service Provider
SaaS	Software as a Service
PaaS	Platform as a Service
IssS	Infrastructure as a Service
NIST	National Institute of Standards and Technology
SLA	Service Level Agreement
CRM	Customer Relationship Management
SOA	Service Oriented Architecture
PCA	Plan de Continuité d'Activité
SSL	Secure Sockets Layer
OECD	Organization for Economic Cooperation and Development
AICPA	American Institute of Certified Public Accountants
CICA	Canadian Institute of Chartered Accountants
DaaS	Database as a Service
PCDP	Préservation de la Confidentialité pour les Données Publiées
QID	Quasi-identifiant
EMD	Earth Mover's Distance

CMS Le calcul multipartite sécurisé

LM Loss Metric

Chapitre 1

Introduction Générale

1. Introduction

Ce chapitre présente l'introduction et la motivation de notre travail de recherche. Il décrit le domaine d'application, les problèmes rencontrés et introduit notre contribution. Il esquisse le cadre et la structure globale de la thèse.

2. Contexte

De nos jours, l'utilisation d'internet et des nouvelles technologies, pour satisfaire l'évolution des besoins de différents types d'utilisateurs (affaire, particulier), fait partie de la vie quotidienne. Toute information est disponible partout dans le monde à tout moment. Cela n'était pas possible il y a quelques années. Récemment, un nombre important de possibilités d'accès à l'information publique et privée sont apparues. Ainsi, nous avons un accès généralisé à un grand débit via Internet grâce au déploiement de dispositifs fixes, mobiles ou encore sans fil qui permettent la connexion à l'internet sans presque se soucier de la limitation géographique.

Aujourd'hui, différents types d'utilisateurs consultent leurs courriers en ligne via des Web-mail, rédigent des documents de collaboration en utilisant les navigateurs web, exécutent des applications et stockent des données dans des serveurs situés sur Internet et non dans leurs propres ordinateurs. De plus, ces services ainsi que d'autres sont utilisés d'une façon transparente pour l'utilisateur et sont donc perçus comme étant des services offerts par un nuage (Cloud) sans en connaître les détails. Cela signifie que de nombreux utilisateurs et organisations peuvent éviter l'installation de certaines applications sur leurs infrastructures ou peuvent avoir plus de puissance de calcul en utilisant les ressources de ce Cloud grâce à Internet. De plus, ces différents utilisateurs peuvent construire leurs propres Clouds privés et les administrer selon leurs propres politiques de gestion. Ainsi, la plupart des entreprises essayent de réduire leurs coûts d'exploitation et de traitement grâce à des techniques de virtualisation. Ces techniques et usages ont conduit à l'émergence d'un nouveau concept appelé Cloud Computing qui permet d'offrir plusieurs types de services avec une meilleure utilisation des ressources, des infrastructures et une réduction de leurs coûts d'exploitation.

Le Cloud Computing ou « informatique dans les nuages » promet en théorie de révolutionner le monde de l'TI (Technologie d'Information) en proposant des services accessibles à la demande au travers d'internet. À partir de 2010, le Cloud Computing a été

massivement médiatisé, les rachats stratégiques s'enchaînent, les offres se multiplient. En réalité, c'est une notion qui reste assez confuse, c'est un terme à la mode et souvent utilisé de manière abusive à des fins marketings. Mais la tendance est bel et bien réelle et présente, toutes les grandes entreprises du domaine informatique s'engagent très fortement dans cette nouvelle approche.

Aujourd'hui quasiment toutes les entreprises, de la petite PME à la grande multinationale, ont commencé à rendre compte les avantages en mettant leurs applications et leurs données dans le nuage, tel que Amazon EC2 et Microsoft Azure qui fournissent un support pour le calcul, la gestion des données et des services internet. Deux exemples de cas de réussite d'Amazon EC2 est *Pinterest*[1], qui gère une application sociale de haute performance et qui stocke plus de 8 milliards d'objets et de 400 téraoctets de données, et *SAP*[2], qui est l'une des plus grandes sociétés de logiciels d'entreprise dans le monde avec des bureaux dans plus de 130 pays, un réseau de partage de contenu social qui a partagé 430 millions d'articles dans 30.000 sites web. Jusqu'à maintenant la solution qui prédominait était de disposer de ressources locales avec un grand nombre de paramètres à gérer, maintenance, mise à jour, refroidissement, sécurité, sauvegardes, interface logicielle etc.

Le Cloud Computing reste une technologie relativement nouvelle. Par conséquent, la plupart des entreprises ne sont pas très confiantes lors de son adoption à cause de plusieurs défis qui restent à relever. Hélas, un tel partage de données est soumis à des contraintes imposées par la sécurité et la confidentialité des individus, un exemple est GoGrid qui a subi une violation de la sécurité, son équipe de sécurité a découvert qu'un tiers non autorisé peut avoir vu les informations de compte, y compris les données des cartes de paiement.

La confidentialité de ces données doit être préservée avant la publication dans le Cloud, c'est à dire aucune information sensible ne doit être divulgués. Selon une étude faite par 400 professionnelles dans le domaine de la technologie d'information, ils ont prouvé que la confidentialité est le facteur numéro *Un* pour la sécurité de Cloud [Chang et al, 2016]. Ce facteur assure que les données d'un client ne soient accessibles que par les entités autorisées. Les différentes solutions de Cloud Computing comportent plusieurs mécanismes de confidentialité telle que la gestion des identités et des accès, le cryptage, et l'anonymisation.

[1]https://aws.amazon.com/fr/solutions/case-studies/pinterest/
[2]https://aws.amazon.com/fr/solutions/case-studies/sap/

Anonymisation des données est l'une des techniques de la confidentialité qui se traduisent par la conservation de l'information, ce qui rend les données inutiles pour tout le monde sauf les propriétaires. Cette technique a été largement étudiée dans la littérature en proposant plusieurs modèles qui ont essayé de répondre aux problèmes causés au niveau de Cloud. La majorité des travaux qui s'articulent sur l'anonymisation touchent deux axes principaux :

- Préserver la confidentialité (anonymisation) des données et affiner la définition de l'anonymisation pour fournir des différentes garanties sur la sécurisation des données.
- Préserver l'utilité de données et minimiser la perte en ce qui concerne la qualité d'information après la publication tout en respectant la définition de l'anonymisation.

Dans cette étude, nous ciblons les deux axes à la fois. Atteindre les objectifs de la confidentialité tout en minimisant la généralisation des données (perte de la qualité d'information).

3. Problématique

Dans le monde numérique moderne, un partage efficace de l'information entre les individus et les organisations est devenu une exigence essentielle. Cela augmente la demande sur le partage des données et sur la confidentialité des individus. La présence d'informations personnelles d'individu (IPI) telles que les dossiers médicaux, les dossiers financiers et les dossiers scolaires a été identifiée comme un obstacle majeur au partage des données. Cela limite le partage des données pour des différentes raisons telles que la recherche universitaire ou commerciale, qui sont importantes pour soutenir les diverses activités dans la société telles que l'amélioration des soins de santé publics et l'élaboration des politiques.

Le problème de savoir comment partager efficacement ces données sans aucune révélation des IPI est encore un défi majeur. Un certain nombre d'approches, telles que l'anonymisation et le cryptage, sont apparues pour résoudre ce genre de problème, mais cela est réalisé avec une perte importante d'information. Il y a donc un problème de partage des micros données tout en protégeant la confidentialité des individus concernés. Le principal défi lors de la divulgation d'informations est de fournir autant d'informations que possible tout en garantissant la confidentialité d'un individu. Cela signifie que limiter la divulgation des

données partagées nécessite un examen minutieux entre l'utilité de données et la confidentialité d'individu.

Ce problème de recherche peut être représenté en posant les principales questions de recherche suivantes :

- Comment pouvons-nous assurer une confidentialité dans un environnement de Cloud tout en réduisant la perte d'information. ?
- Quelles approches peuvent être mises en place afin de réduire la quantité de perte d'information tout en s'efforçant la protection de la confidentialité de l'individu dans un environnement de Cloud?
- Comment concevoir, développer et mettre en œuvre des approches d'anonymisation afin d'améliorer la confidentialité et l'utilité de données dans un environnement distribué et particulièrement dans un mode Cloud?

4. Contribution

Le but de cette recherche est d'étudier le problème de l'anonymisation des données dans un environnement distribué horizontalement avec une perte d'information minimale ce qui rend les données utiles. La recherche vise à répondre à la question de savoir comment les éditeurs de données, tels que les hôpitaux, les collectivités locales, les organismes privés et gouvernementaux, peuvent publier leurs données, tout en préservant la confidentialité des individus. Pour répondre à cette question, notre recherche propose un algorithme d'anonymisation distribué pour la préservation de la confidentialité des données publiées à partir de plusieurs fournisseurs de données tout en réduisant la quantité de perte d'information dans un environnement de Cloud.

Nos travaux de recherche s'articulent sur deux points importants : La confidentialité des objets des données (individus), et la confidentialité des fournisseurs des données :

- Pour la confidentialité des individus, nous allons utiliser le modèle récent : *K-concealment* qui assure un haut niveau de sécurité et offre une plus grande utilité de données par rapport aux modèles suggérées dans les travaux précédents.
- Pour la confidentialité des Fournisseurs de données, nous allons concevoir un nouvel algorithme d'anonymisation distribuée qui focalise sur la structure de l'arbre- R* et qu'il donne une meilleur insertion des objets des données dans

l'espace de domaine de quasi-identifiant à travers un partitionnement (Split) adéquat, ce qui va nous donner par la suite une meilleur utilité de données.

5. Organisation de la thèse

En plus d'une introduction générale et d'une conclusion, notre thèse, comme l'indique *la Figure 1.1*, est composée de trois parties : Background, État de l'art et Contribution.

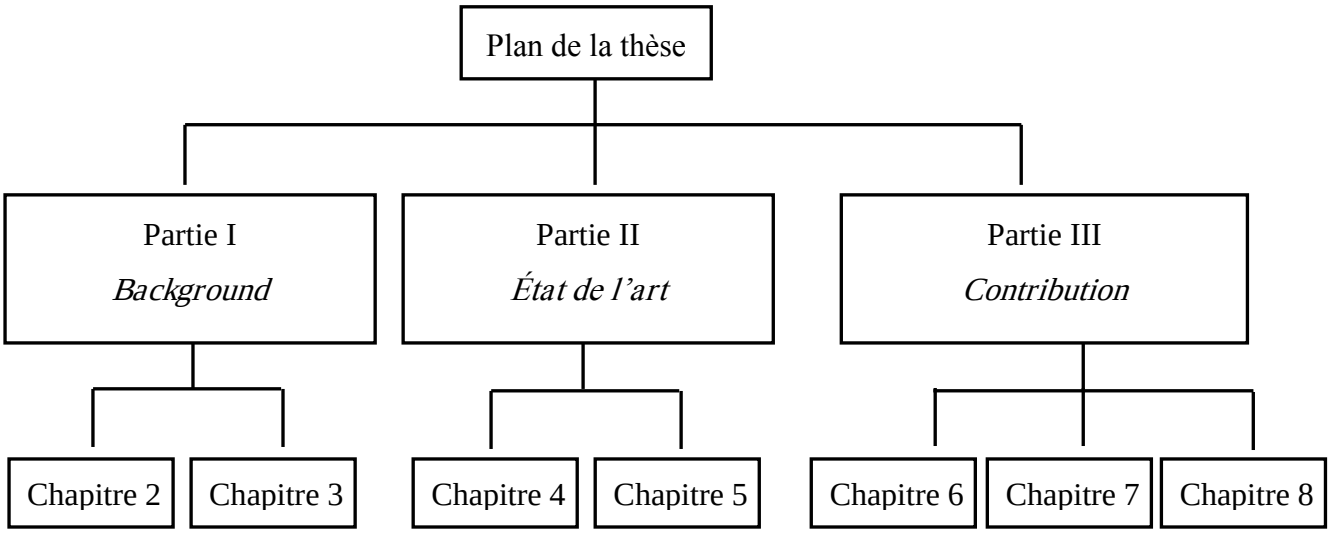


Figure 1.1 Synthèse du plan de la thèse

- La partie I : Background

La première partie propose une généralité sur les notions de base du Cloud Computing et la sécurité dans cet environnement. Cette partie comprend deux chapitres :

- *Chapitre 2* : Ce chapitre traite les définitions de tous ce qui concerne le Cloud. Il cherche à répondre aux questions : De quoi le Cloud Computing est-il constitué ? Quelles sont ses différentes architectures ? Quelles sont ses avantages ainsi que ses points faibles ?
- *Chapitre 3* : Le troisième chapitre se focalise sur l'aspect Sécurité (physique et logique) dans le Cloud, et plus particulièrement sur la confidentialité qui est considérée comme l'un des mécanismes de sécurité les plus importants dans cet environnement. Il présente un aperçu global sur les problèmes liés à ce mécanisme en classifiant les principales solutions proposées.

- La partie II : État de l'art

La deuxième partie présente les notions de base sur l'aspect Anonymisation ainsi qu'un État de l'art sur les travaux qui touchent la préservation de la confidentialité pour les bases des données centralisées et décentralisées. Cette partie se compose de deux chapitres :

- Chapitre 4 : Dans ce chapitre, nous détaillons les caractéristiques de base de la préservation de la confidentialité pour les données publiées. Nous donnons quelques définitions sur tous ce qui concerne l'approche d'anonymisation. Nous présentons les types d'anonymisation et les opérations sur lesquelles l'anonymisation se déroule.
- *Chapitre 5* : Dans le cinquième chapitre, nous introduisons les deux principaux axes liés à la préservation de la confidentialité. Nous abordons par la suite, les différents travaux connexes à la préservation de la confidentialité pour les bases de données centralisées ainsi que les différentes approches qui peuvent être considérées comme des solutions pour l'anonymisation des données dans les bases des données distribuées. Enfin, nous concluons par un tableau comparatif qui va permettre de mieux positionner notre contribution par rapport aux travaux de la littérature.

- La partie III : Contribution

La troisième partie décrit notre approche distribuée par une conception et une validation d'un protocole d'anonymisation distribué. Elle comprend trois chapitres :

- *Chapitre 6* : Ce chapitre présente le modèle K-concealment utilisé dans notre approche distribuée. Nous détaillons les différents types de k-anonymisations qui reposent sur la représentation graphiques des enregistrements. Ensuite, nous montrons l'insécurité de ces types qui n'offrent pas un haut niveau de sécurité par rapport au modèle k-concealment. Enfin, nous conclurons par une discussion sur l'algorithme d'anonymisation du modèle K-concealment.
- *Chapitre 7* : Dans ce chapitre nous introduisons notre protocole d'anonymisation distribué. Nous commençons par exposer l'architecture de notre approche distribuée en détaillant le nouvel algorithme d'anonymisation distribué. Une description de l'opération Split utilisée dans notre protocole est présentée. Cette opération fournit un partitionnement plus efficace que celui utilisé dans les

travaux précédents. Nous finissons par une présentation de la généralisation de l'arbre R* qui vise à insérer les données dans un chevauchement minimal.

- *Chapitre 8* : Nous présentons dans ce chapitre les expérimentations que nous avons mené pour valider notre contribution. Nous évaluons la qualité d'anonymisation à travers une comparaison de notre approche avec l'approche d'anonymisation centralisée en utilisant des métriques bien précises.

Chapitre 2

Généralité sur le Cloud Computing

1. Introduction

Dans ce chapitre nous allons commencer par quelques définitions de tout ce qui concerne le Cloud Computing. D'où est-on parti pour arriver à cette informatique dans les nuages ? De quoi le Cloud Computing est-il constitué ? Quelles sont ses différentes architectures ? Quelles sont ses avantages ainsi que ses points faibles ? C'est à ces questions que va répondre de manière globale ce chapitre.

2. Historique

Il n'y a pas de date-clé à laquelle nous puissions dire que le Cloud Computing est né. Pour comprendre ce que le Cloud Computing est, et n'est pas, il est important de comprendre comment ce modèle de l'informatique a évolué. Comme *Alvin Toffler* a noté dans son fameux livre *The ThirdWave* [Alvin Toffler, 1980], la civilisation a progressé dans des vagues (les sociétés agricoles, l'ère industrielle, et le troisième est l'ère de l'information). La notion de Cloud fait référence au l'ère de l'information, tel que nous l'avons l'habitude de l'utiliser dans des schémas techniques lorsque nous voulons représenter Internet. Un réseau comme Internet est constitué d'une multitude de systèmes fournissant des services et des informations. Le Cloud Computing est dans cette lignée : un ensemble de services et de données consommables [Grevet, 2009].

2.1. L'informatique utilitaire de *JOHN MCCARTHY*

Cette notion de consommation a été proposée en 1961, lors d'une conférence au MIT (Massachusetts Institute of Technology), par *John MCCarthy* aussi connu comme l'un des pionniers de l'intelligence artificielle (dont il proposa le nom en 1955) et pour avoir inventé du LISP en 1958.

Lors de ce discours, *John McCarthy* suggéra que la technologie informatique partagée (« time-sharing ») pouvait construire un bel avenir dans lequel la puissance de calcul et même les applications spécifiques pouvaient être vendues comme un service public.

> *Computation may someday be organized as a public utility.*
>
> *(Les ressources informatiques deviendront un jour d'utilité publique)*
>
> ***JOHN MCCARTHY***

Cette idée, très populaire dans les années 60, disparu au milieu des années 70 : à l'époque, les technologies matérielles, logicielles et réseaux n'étaient tout simplement pas prêtes.

Le Cloud Computing met en œuvre l'idée d'informatique utilitaire du type service public, proposée par *John McCarthy*. Il peut aussi être comparé au cluster de calcul dans lequel un groupe d'ordinateurs se relient pour former un ordinateur virtuel unique permettant le calcul de haute performance (CHP), mais aussi à l'informatique en grille (Grid Computing) où des ordinateurs reliés et répartis géographiquement permettent la résolution d'un problème commun.

2.2. Le service bureau

Au milieu des années soixante-dix, la notion de « service bureau » est inventée pour qualifier une entreprise louant des lignes téléphoniques, répondeurs, services informatiques etc. Généralement, les clients du « service bureau » n'ont ni l'ampleur ni l'expertise pour intégrer en interne ces services, c'est pourquoi ils passent par un prestataire. La combinaison de technologies, processus et expertise dans le domaine des entreprises est la valeur ajoutée du «service bureau », comme modèle économique basé sur leur capacité à produire des services et à les déployer en volume. IBM lui-même était un « service bureau » en proposant la notion de « on-demand ».

À l'époque, le coût d'achat et d'exploitation de mainframes IBM était hors de prix. C'est pourquoi, des solutions permettant aux entreprises de pouvoir exploiter ces technologies à moindre frais avec la notion de « paiement à la consommation » furent proposées.

2.3. Les Fournisseur de services applicatifs

Les ASP, « Application Service Provider » ont aussi leur part dans l'historique du Cloud Computing. Une ASP désigne une application fournie comme un service, c'est ce que l'on nomme actuellement SaaS pour « Software as a Service » dans la terminologie actuelle du Cloud Computing. Plutôt que d'installer le logiciel sur le poste client en ayant à assurer les phases d'installations et de maintenance sur chaque poste, les applications ASP sont hébergées et centralisées sur un serveur unique et accessible par les clients au travers de protocole standard. C'est par exemple le cas avec des applications Web accessibles par http : il n'y a alors plus de déploiement ou de maintenance à effectuer sur le poste utilisateur, celui-ci n'a alors besoin que d'un simple navigateur Internet. Le déploiement, la configuration, la maintenance, la sauvegarde, etc. sont désormais de la responsabilité du fournisseur du service, le client est alors consommateur.

2.4. La virtualisation

La virtualisation a été la première pierre vers l'ère du Cloud Computing. En effet, cette notion permet une gestion optimisée des ressources matérielles dans le but de pouvoir y exécuter plusieurs systèmes « virtuels » sur une seule ressource physique et fournir une couche supplémentaire d'abstraction du matériel. Les premiers travaux peuvent être attribués à IBM, qui dans les années soixante, travaillait déjà sur les mécanismes de virtualisation en développant dans les centres de recherche de Cambridge et de Grenoble, CMS (Conversation Monitor System), le tout premier hyperviseur [N. Grevet, 2009]**.**

C'est donc depuis presque cinquante ans que l'idée d'une informatique à la demande est présentée dans les esprits même si les technologies n'étaient jusqu'alors pas au rendez-vous pour pouvoir concrétiser cette idée.

Avec les différents progrès technologiques réalisés durant ces 50 dernières années, tant sur le plan matériel, logiciel et conceptuel, aux avancées des mécanismes de sécurité, à l'élaboration de réseaux complexes mais standardisés comme Internet, et à l'expérience dans l'édition et la gestion de logiciels, services, infrastructures et stockage de données, nous sommes maintenant prêts à entrer dans l'ère du Cloud Computing, telle que rêvait par John McCarthy en 1961.

3. Définition du Cloud Computing

La définition suivante du Cloud Computing est un extrait simplifié de la définition du *National Institute of Standards and Technology* (NIST, Etats-Unis) et du Groupe spécialisé de l'UIT.

«Le Cloud Computing *est un modèle qui offre aux utilisateurs du réseau un accès à la demande, à un ensemble de ressources informatiques partagées et configurables, et qui peuvent être rapidement mises à la disposition du client sans l'interaction direct avec le prestataire de service.»* [Peter et Tim, 2011]

4. Caractéristiques du Cloud Computing

Les Services *Cloud Computing* ont des caractéristiques qui les distinguent des autres technologies : [Peter et Tim, 2011]

- ***Libre-service à la demande*** : Le client peut consommer les services Cloud automatiquement selon son besoin sans aucune nécessitée d'une interaction humaine avec le fournisseur.

- ***Accès réseau (Ubiquité)*** : Les capacités sont disponibles sur le réseau et accessibles via des mécanismes standards qui favorisent l'utilisation de plates-formes.
- ***Mise en commun des ressources (pooling)*** : Les ressources et les services fournis au client sont souvent virtuels et partagés par plusieurs utilisateurs.
- ***Elasticité rapide*** : Les utilisateurs peuvent rapidement augmenter et diminuer leurs ressources en fonction des besoins, ainsi que de libérer les ressources pour d'autres utilisations quand ils ne sont plus nécessaires. (*Figure2.1)*
- ***Les services sont fournis selon le modèle pay-per-use*** : ou le modèle d'abonnement.

Ces spécificités font de la technologie *Cloud Computing* une nouvelle option qui offre à ses utilisateurs la possibilité d'accès à des logiciels et à des ressources informatiques avec la flexibilité et la modularité souhaitées et à des coûts très compétitifs.

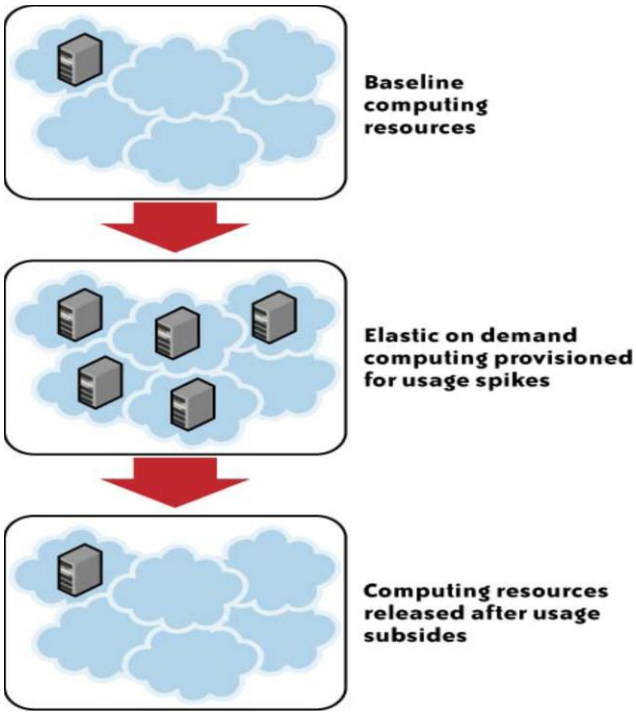


Figure 2.1. Attribut d'élasticité [Tim Mather, 2009]

5. Les différentes couches du Cloud Computing

La *Figure 2.2* ci-dessous représente les différentes couches du Cloud Computing de la couche la moins visible pour les utilisateurs finaux à la plus visible. L'infrastructure as a Service (Iaas) est plutôt gérée par les architectes réseaux, la couche PaaS est destinée aux

développeurs d'applications et finalement le logicielle comme un service (SaaS) est le « produit final » pour les utilisateurs.

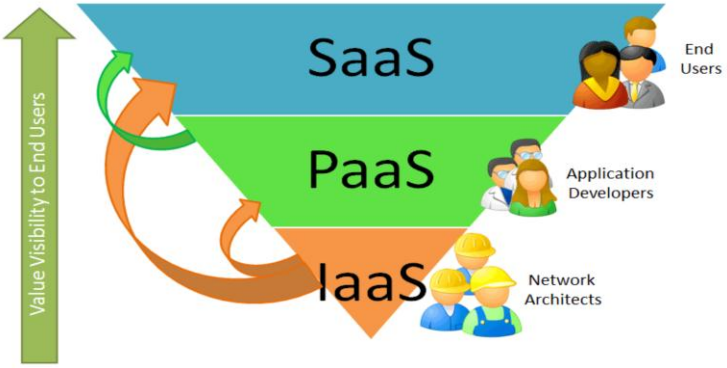


Figure 2.2 Les trois couches du Coud Computing [Warin, 2011]

5.1. SaaS (Software-as-a-Service)

Les méthodes traditionnelles de l'achat de logiciels impliquent le client de charger le logiciel sur son propre matériel. Dans le model SaaS (Le logiciel en tant que service) le client ne doit pas acheter un logiciel, mais plutôt louer pour une utilisation (utilisant le modèle pay-per-use). C'est en quelque sorte la partie « visible » du Coud Computing pour l'utilisateur final. Nous y trouvons différents types d'application, du CRM à la gestion des ressources humaines, comptabilité, messagerie etc.…

Les principaux avantages [Mather et al, 2009]:

- SaaS permet à l'organisation d'externaliser l'hébergement et la gestion des applications à un tiers (fournisseur de services), comme un moyen de réduire le coût des licences de logiciels d'application.
- SaaS permet aux éditeurs de logiciels pour contrôler et limiter l'utilisation, interdit la copie et la distribution, et facilite le contrôle de toutes les versions dérivées de leur logiciel.

5.2 PaaS (Platform-as-a-Service)

Dans un PaaS (modèle plate-forme-as-a-service), le fournisseur offre un environnement de développement pour les développeurs d'applications, qui développent des applications et offrent ces services par le biais de la plate-forme du fournisseur [Mather et al, 2009].

Les solutions PaaS sont des plates-formes de développement pour lesquels l'outil de développement lui-même est hébergé dans le nuage et accessible via un navigateur.

Avec PaaS, les développeurs peuvent souvent créer des applications Web sans avoir à installer des outils sur leurs ordinateurs, et peuvent ensuite déployer ces applications sans aucunes connaissances spécialisées du système d'administration. Parmi les solutions : *Windows Azure de Microsoft, App Engine de Google, Force.com de Salesforce*. Chaque fournisseur de PaaS propose des environnements de développement différents, Google App Engine se limite à Java et Python, tandis que Windows Azure permet de travailler avec les langages .NET, PHP, Python, Ruby et Java.

5.3 IaaS (Infrastructure-as-a-Service)

Dans le modèle traditionnel d'application hébergée, le vendeur fournit toute l'infrastructure pour un client pour exécuter ses applications. Souvent, cela implique un hébergement matériel dédié qui est acheté ou loué pour cette application spécifique. Le modèle IaaS fournit également l'infrastructure pour exécuter les applications, mais l'approche de Coud Computing permet d'offrir un modèle pay-per use et à l'échelle du service en fonction de la demande. Cette infrastructure fournit des capacités de calcul, de stockage et une bande passante suffisante et elle est mise à disposition de façon à gérer automatiquement la charge de travail requise par les applications. Il y a très peu de limitation pour le client si ce n'est la partie matérielle qui peut être contournée grâce aux systèmes de virtualisation. Les applications vont dès lors pouvoir être déployées sans être liées à un serveur spécifique. La virtualisation répond de manière dynamique là où les serveurs physiques fournissent un ensemble de ressources allouées selon les besoins, et où la relation entre les applications et les ressources de calcul, de stockage et de réseau pourront s'adapter de manière automatique pour répondre à la charge de travail et aux exigences demandées [Mather et al, 2009].

Du côté des fournisseurs, le marché est également en plein effervescence, de quelques acteurs en 2008, les annonces de services Cloud IaaS explosent, nous pouvons citer les plus connus : Microsoft avec Windows Azure, Amazon avec Amazon Web Services, VMware avec VMware vSphere.

En Cloud, la flexibilité est obtenue grâce à la virtualisation des systèmes d'exploitation. La plateforme est exécutée via des machines virtuelles et les ressources peuvent être allouées et délibérées à la demande. Ainsi, l'IaaS est considéré le service le plus flexible.

D'autre part, l'IaaS est le service le plus simple à mettre en place. *La figure 2.3* montre les trois couches du Cloud Computing en donnant un compromis flexibilité/simplicité.

Pour consolider ces différentes définitions, nous pouvons retenir qu'avec le SaaS, nous utilisons une application, avec le PaaS nous construisons ses applications et finalement l'IaaS permet d'héberger le tout.

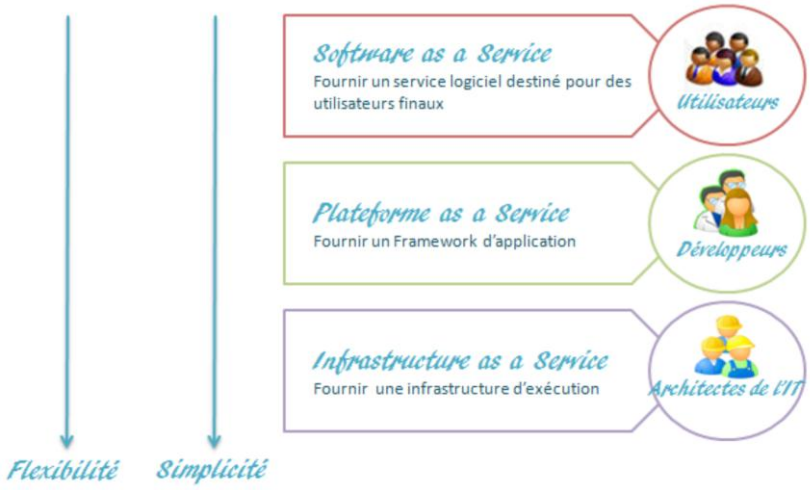


Figure 2.3 La flexibilité et la simplicité des trois couches du Cloud Computing

6. Types de Cloud

Comme nous avons vu, le Cloud Computing repose sur des ressources physiques. La question est « ***où sont ces ressources physiques*** ? » (Serveurs, commutateurs, routeur, solutions de stockage, etc.).

La réponse « dans le nuage » n'est pas vraiment acceptable. Du point de vue consommateur, l'abstraction est telle que nous pouvons déterminer sur quelles ressources physiques l'application est hébergée. De par son côté dynamique, les ressources physiques hébergeant une application et des données dans un Cloud ne sont jamais fixes et évoluent dans le temps.

6.1. Le Cloud publique

Pour certaines personnes, tel que *Werner Vogels* (Directeur des technologies d'Amazon), le Cloud ne peut être que public.

Un Cloud public est un service IaaS, PaaS ou SaaS proposé et hébergé par un tiers d'un ou plusieurs centres de données *Figure 2.4*. Amazon, Google et Microsoft propose un Cloud

public dans lequel n'importe quel particulier ou n'importe quelle entreprise peut y héberger ses applications, ses services ou ses données. Pour les consommateurs, il n'y a donc aucun investissement initial fixe et aucune limite de capacité.

Les fournisseurs de Cloud public sont les responsables pour la gestion de la sécurité et facturent à l'utilisation et garantissent une disponibilité de services au travers des contrats SLA (« Service Level Agreement » : document qui définit la qualité de service requise entre un prestataire et un client).

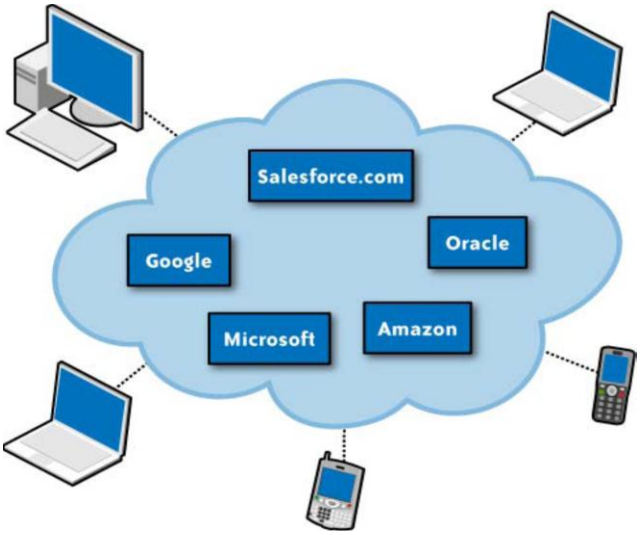


Figure 2.4 Un Cloud publique [Tim Mather, 2009]

6.2. Le Cloud privé

Ces ressources physiques peuvent être hébergées dans une infrastructure propre à l'entreprise et étant sous son contrôle, à sa charge donc de contrôler le déploiement des applications.

La fameuse question qui se pose à ce niveau est : « ***Si un Cloud privé est réellement un Cloud ?*** ». En effet, dans le sens où, comme nous avons dit plus haut, un Cloud ne doit pas imposer de dépenses en immobilisations, l'infrastructure physique dans un Cloud privé est à la charge de l'entreprise.

Le Cloud privé peut aussi désigner un Cloud déployé sur une infrastructure physique dédiée et mise à disposition d'un fournisseur de services.

Ainsi une entreprise peut louer à un fournisseur de services, un nombre conséquent de serveurs qui lui sont entièrement dédiés et sur lesquels une solution de Cloud sera déployée pour gérer dynamiquement l'application, la plate-forme ou l'infrastructure (virtuelle).

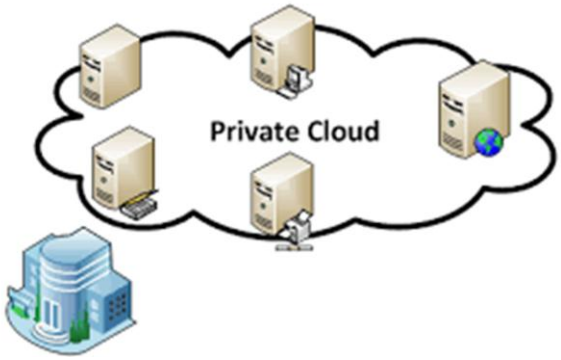


Figure 2.5 Cloud privé [Tim Mather, 2009]

6.3 Le Cloud hybride

Un Cloud Hybride est l'utilisation de plusieurs Clouds, publics ou privés. Nous pouvons ainsi déporter nos applications vers un Cloud public qui consommera des données stockées et exposées dans un Cloud privé *Figure 2.6*, ou bien faire communiquer deux applications hébergées dans deux Clouds privés distincts, ou encore consommer plusieurs services hébergés dans des Cloud publiques différents

Figure 2.6 Cloud hybride [Tim Mather, 2009]

7. Les principaux acteurs du Cloud

Dans cette section, nous allons parcourir les principaux fournisseurs de services dans le Cloud et observer les offres ainsi que les prix proposées.

7.1Amazon

Amazon a été la première société à proposer une plateforme du Cloud Computing avec Amazon Web Services[3] . Il s'agit d'un outil simple et facile à manipuler qui permet le développement des applications dans un environnement du Cloud Computing. Les services d'Amazon Web Services sont :

- *Amazon Elastic Compute Cloud (EC2)* : il fournit un stockage basé sur les services Web.
- *Amazon Simple Storage Service (S3)* : Amazon S3 offre une simple interface de services web à utiliser pour stocker et récupérer n'importe quelle quantité de données, à tout moment, de n'importe quel endroit sur le web.
- *Amazon Cloud Front* : il permet de distribuer des objets stockés sur S3 vers un emplacement proche de l'appelant.
- *Amazon Simple DB* : il permet aux développeurs d'exécuter des requêtes sur des données structurées.
- *AWS Management Console (AWS Console)* : c'est une interface Web pour gérer et surveiller les infrastructures Amazon, incluant Amazon EC2 et Amazon S3.

7.2 Google App Engine

Google App Engine est une plateforme de développement et d'hébergement d'applications Web. Il permet d'exécuter des applications Web sur l'infrastructure de Google[4] . Il est lancé en 2008 et est disponible seulement en Cloud publique sous la forme d'une offre gratuite. Sa grande force réside dans le fait qu'il soit purement gratuit.

Google Apps

Une suite bureautique (Gmail, Google Agenda, Google Documents, etc.) de la société Google qui est accessible par les particuliers, les entreprises, les établissements d'enseignement et les organisations[5]. Il existe différentes versions de Google Apps :

- *Google Apps for Groups* : pour tout le monde mais limité à 50 utilisateurs.
- *Google Apps for Education* : pour le primaire, le secondaire et l'université.
- *Google Apps for Governement* : pour les services gouvernementaux.

[3]Amazon web service. Http: //aws.amazon.com/fr/.
[4]Google app engine. Http: //code.google.com/Intel/fr-FR/appengine/.
[5]Google apps. Http: //www.google.com/apps/.

- *Google Apps for Business* : il s'agissait anciennement de Google Apps Premier Edition.

7.3 Salesforce

Salesforce.com est une société pionnière dans le domaine du SaaS, elle a été créée en 1999 déjà par *Marc Benioff*. Ces dix dernières années l'entreprise n'a cessé de croître et d'enchaîner les acquisitions. En 2004 la société est introduite à la bourse de New-York sous le symbole boursier CRM, apportant ainsi pas loin de 110 millions de dollars[6].

Basée sur la base de données *database.com* et une place de marché de logiciels, les solutions de *Salesforce.com* sont regroupées dans différentes grandes catégories : Sales Cloud, Service Cloud, Force.com et Chatter Collaboration Cloud.

- ***Sales Cloud*** : L'outil de CRM (Customer Relationship Management) par excellence disponible en plus de 25 langues et accessible depuis des appareils mobiles. Le produit fournit des outils de gestion des comptes et contacts clients, outils marketing, de ventes, plateforme de discussion.
- ***Service Cloud*** : Un service client de nouvelle génération permettant aux entreprises d'être plus sociable et collaborative. Service Cloud propose des services rapides et réactifs intégrant tous les canaux de communication ; du centre d'appel aux réseaux sociaux.
- ***Chatter Collaboration Cloud*** : Est une plateforme de collaboration en temps réel reprenant un peu la forme d'un réseau social.

Salesforce.com propose également une solution PaaS : ***Force.com*** qui permet de créer des applications au moyen de Visual force (un framework pour la création d'interfaces graphiques) et ***Apex***, un langage de programmation propriétaire qui reprend la syntaxe de Java mais qui est plus tourné vers la gestion des bases de données.

7.4 OVH

OVH est une société française indépendante, orienté grand public. Elle est le numéro un de l'hébergement en France et se positionne comme le deuxième hébergeur européen[7]. Elle propose des serveurs privés dédiés, des ressources de stockage, des ressources de bande

[6] http://investor.salesforce.com/about-us/investor/investor-news/investor-news-details/2015/Salesforce-Announces-Fiscal-2015-Fourth-Quarter-and-Full-Year-Results/default.aspx.
[7] Netcraft. Http: //news.netcraft.com/.

passante, des ressources, de CPU, de RAM, de SQL, du HTTP, des Emails etc. En 2010, OVH démarre une offre du Cloud Computing assez agressive ce qui fait son succès[8].

8. Cloud Computing et SOA

Les services Clouds ont quelques caractéristiques clés : élasticité, self-service, à la demande etc. Pour fournir de ces caractéristiques, les bases du service, de l'infrastructure Cloud doivent être bien conçus et bien architecturés. Afin de rendre cette approche possible, l'architecture du Cloud devrait être basée sur une approche modulaire. Une architecture modulaire permet la flexibilité et la réutilisation du service. L'approche cachée sous cette flexibilité n'est autre que SOA, une architecture orientée service. Nous allons donner un aperçu de SOA ainsi que la relation entre SOA et le Cloud **[**Hurwitz et al, 2010**].**

8.1 Définition de SOA

L'architecture orientée service (SOA) est une approche de conception et de construction de systèmes d'informations qui utilise des composants d'applications existantes. Les applications existantes exposent leurs fonctionnalités à travers des interfaces services qui peuvent être mis à disposition à l'intérieur même d'une entreprise ou par le biais d'internet. SOA facilite grandement la réutilisation de fonctionnalités **[**Hüsemann, 2009**].**

Cette approche permet de :

- Augmenter la flexibilité des applications
- Réduire les coûts de l'intégration
- Mettre à disposition des services réutilisables à travers différentes applications et/ou systèmes d'information **[**Hüsemann, 2009**].**

8.2 La relation entre le Cloud et SOA

Premièrement il est important de ne pas mélanger les deux termes. Selon *David Mitchell Smith*, vice-président de *Gartner*, la confusion est généralement faite à cause du terme « service » utilisé à la fois pour le Cloud et pour le SOA.

Dans le monde du SOA, nous parlons de services en faisant références aux logiciels, aux composants actifs et aux objets (éléments techniques), mais dans le monde réel quand nous

[8]Ovh. Https: //www.ovh.com/fr/index.xml

parlons de service, nous pensons résultats. *M. Smith* continue : « Avec le Cloud, vous payez pour des résultats, pas pour la technologie. ».

Cependant, les deux services sont liés, *le fait de bénéficier du SOA représente une bonne base pour le Cloud.*

Les deux concepts sont donc compatibles sans être basés sur la même idée : Le Cloud Computing est une architecture de déploiement et non une approche architecturale dans le but de construire le système IT de l'entreprise comme l'est SOA.

SOA est donc une approche au niveau de l'architecture afin de proposer un système comprenant des services réutilisables et interchangeables. Le Cloud Computing a pour but de délivrer une infrastructure et des services réutilisables en rapport avec les besoins de l'entreprise. Certains principes du Cloud peuvent être basés sur une architecture SOA mais le but en soi de ces deux concepts est différent.

9. La virtualisation

9.1 Définition

La virtualisation est l'ensemble de techniques et d'outils permettant de faire tourner plusieurs systèmes d'exploitation sur une même machine physique afin de délivrer une meilleure utilisation des ressources. Cette technologie vient pour répondre à certains problèmes tels que :

- *La sous exploitation des serveurs physiques* : il est estimé que dans un Datacenter privé, le taux d'utilisation moyen est de 10%. Celui-ci passe à 35% sur une architecture virtuelle [Devarieux, 2009].
- *La croissance du nombre de serveurs physiques* : les ressources physiques d'un serveur seront partagées entre différents serveurs virtuels ce qui permet de ne pas acheter plusieurs serveurs physiques.
- *La sécurité et la fiabilité* : isoler les services sur des serveurs différents.

Dans les systèmes de virtualisation, il faut noter les notions suivantes :

- ✓ **SE hôte** : le système d'exploitation installé sur la machine physique.
- ✓ **SE invité** : les systèmes d'exploitation des machines virtuelles.
- ✓ **Partage des ressources physiques** : les différentes machines virtuelles installées sur le serveur partagent ces ressources à savoir le processeur, les disques durs et d'autres périphériques.

- ✓ **Isolation** : les machines virtuelles sont considérées comme des ordinateurs physiques et donc possèdent chacune sa propre adresse IP.
- ✓ **Manipulation des machines virtuelles** : une machine virtuelle est un fichier situé sur un disque du serveur.

9.2 Techniques de virtualisation

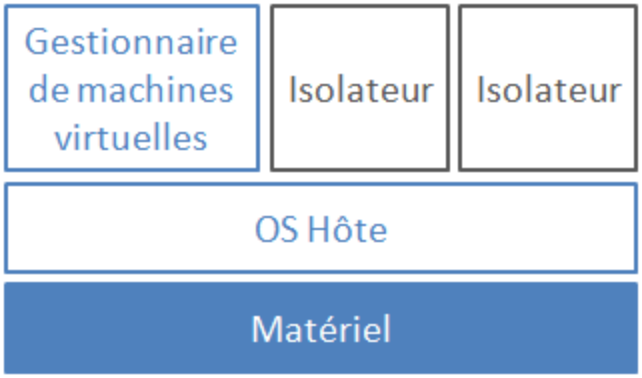
La virtualisation permet de cohabiter plusieurs systèmes d'exploitation complètement isolés dans un même hôte[9].

On distingue plusieurs techniques de virtualisation qui sont détaillées dans ce qui suit :

L'isolation

L'isolation permet de diviser un système d'exploitation en plusieurs espaces mémoires ou encore contextes. Chaque contexte est géré par le SE hôte. Cette isolation permet de faire tourner plusieurs fois la même application prévue pour ne tourner qu'une seule fois par machine.

Les programmes de chaque contexte ne sont capables de communiquer qu'avec les processus et les ressources associées à leur propre contexte. L'isolation est uniquement liée aux systèmes Linux.



La figure 2.7 Architecture de la technique d'isolation. [Mahjoub, 2011]

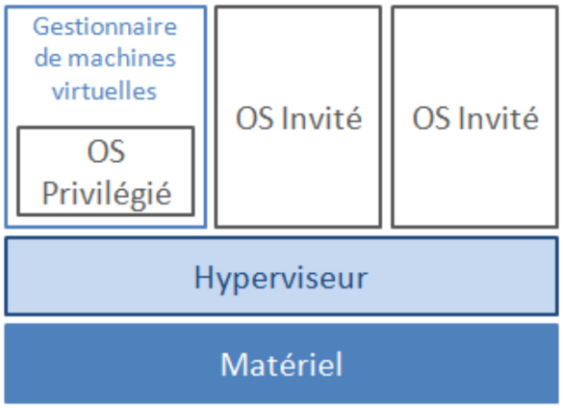
La para-virtualisation

La para-virtualisation est une technique de virtualisation qui présente à la machine invitée une interface logicielle similaire mais non identique au matériel réel. Ainsi, elle permet aux systèmes d'exploitation invités d'interagir directement avec le système d'exploitation hôte et donc ils seront conscients de la virtualisation.

[9]http : //doc.fedora-fr.org/wiki/Virtualisation

La virtualisation complète

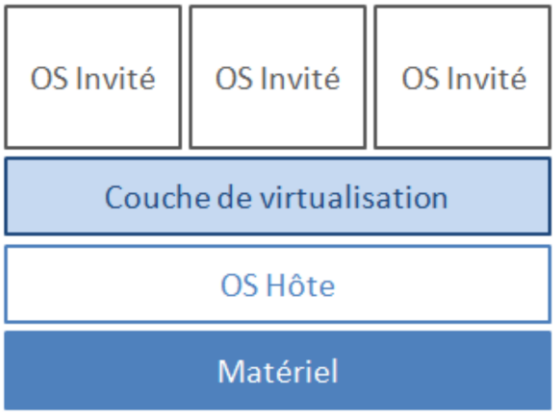
La virtualisation complète est une technique de virtualisation qui permet de créer un environnement virtuel complet. En utilisant cette technique, le système d'exploitation invité n'interagit pas directement avec le système d'exploitation hôte et donc il croit s'exécuter sur une véritable machine physique.



La figure 2.8 La para-virtualisation [Mahjoub, 2011]

Cette technique de virtualisation ne permet de virtualiser que des SE de même architecture matérielle que l'hôte**.**

Dans certaines architectures matérielles, le support de virtualisation est intégré avec le processeur. Les exemples les plus connus du marché sont : AMD-V et IntelVT.



La figure 2.9 La virtualisation complète [Mahjoub, 2011]

10. Les avantages de Cloud Computing

L'impact financier est important : N'importe qui peut bénéficier d'une infrastructure technique dernier cris, capable d'absorber n'importe quel type de charge et sans aucun investissement préalable. Il n'y a donc plus de dépense en immobilisation et en cas de sous-utilisassions, le Cloud Computing adaptera les ressources mises à disposition de manière proactive en fonction de la charge. *D'un point de vue financier, la facture sera relative à l'utilisation du service.*

En clair, le Cloud Computing permet de limiter les coûts d'exploitation en mutualisant les ressources physiques, garantit une haute disponibilité des services et des données et adapte l'infrastructure technique au volume d'activité de l'entreprise.

Pour finir, les développeurs, administrateurs, chef de projets, décideurs et les utilisateurs ne sont pas les seuls gagnants dans l'utilisation du Cloud Computing, la planète y trouve aussi une manière de réduire l'empreinte écologique de notre technologie actuelle et future.

11. Les freins au Cloud

Sur le papier, le Cloud Computing semble promis un très grand avenir. Pourtant plusieurs personnes ou entreprises vont à l'encontre de cette notion, comme le très célèbre *Richard Stallman* (fondateur de la « Free Software Foundation » et créateur du projet GNU) qui parle du Cloud Computing comme *un piège* [N. Grevet, 2009]. Deux des plus importants obstacles à l'adoption sont la sécurité et la confidentialité.

- *Sécurité* : Comment garantir la sécurité des informations stockées dans les nuages ? N'y a-t-il pas de risque d'intrusion, de perte ou de dégradation des données ? Le chapitre suivant sera dédié à cet obstacle dans la suite de ce mémoire.
- *Confidentialité* : La confidentialité reste comme l'une des plus grandes préoccupations en ce qui concerne le Cloud Computing. Ceci est largement dû au fait que les clients externalisent leurs tâches de données et de calcul sur les serveurs de nuages et qui sont contrôlées et gérées par les fournisseurs de Cloud. [Mather et al, 2009]

12. Conclusion

De l'informatique utilitaire des années soixante, au service bureau des années soixante-dix, tout en passant par l'émergence d'Internet et des avancées de virtualisation, le Cloud Computing comme les chiffres nous le confirme, est promis à un bel avenir. Dans ce chapitre,

nous avons présenté les principaux concepts qui concernent le Cloud, commençant par ses caractéristiques essentielles, arrivant à ses acteurs. Nous avons également définit l'aspect de virtualisation ainsi que ses différents types et nous avons conclu par citer les point forts et les points faibles qui touchent cette nouvelle technologie.

Il reste encore beaucoup à faire notamment concernant la sécurité et la confidentialité qui sont les deux les plus points faibles pour l'adoption de Cloud Computing. Dans le chapitre qui suit, nous allons décrire l'aspect confidentialité et sécurité dans le Cloud en essayant de répondre aux questions qui y affèrent.

Chapitre 3

La Sécurité et la Confidentialité dans le Cloud

1. Introduction

La sécurité est souvent citée comme le frein principal à l'adoption des services Cloud. L'accès aux données hébergées dans le Cloud présente en général un haut niveau de sécurité en raison des mécanismes d'authentification mis en place par les fournisseurs de service. Ces mécanismes peuvent d'ailleurs être renforcés par les solutions Corporate clients, de gestion d'identités, qui sont alors placées en amont d'un lien unique avec le fournisseur de solutions Cloud ; notons cependant que certains fournisseurs seulement acceptent une telle architecture.

Ce chapitre s'articule autour de deux parties principales :

- La première partie met l'accent sur l'aspect sécurité qui est cité comme le frein principal à l'adoption des services Cloud. Nous allons essayer dans cette partie de présenter les deux types de sécurité ; physique et logique.
- La deuxième partie a pour objectif de présenter l'aspect confidentialité qui est considéré comme l'un des mécanismes de la sécurité.

2. La sécurité dans le Cloud

2.1 Sécurité physique

Le Cloud Computing, par nature, est associé à une sorte de « dématérialisation » de l'hébergement (le nuage). En effet, le lieu d'hébergement du Cloud est généralement multiple, et réparti sur plusieurs data centres. Dans le cas du Cloud public, le client ne connaît donc pas avec précision le ou les lieux d'hébergement du Cloud.

Cette caractéristique, gage de disponibilité du Cloud, entraîne un changement important pour le client dans le mode de sélection de l'hébergeur.

Une visite de data centre ne suffit plus pour évaluer le niveau d'hébergement garanti par un fournisseur de Cloud. Ce dernier doit être en mesure d'apporter des garanties sur les conditions d'hébergement associées à son offre.

Un certain nombre de certifications et/ou de classifications existent à ce sujet, sont reconnues et adoptées par l'ensemble des hébergeurs.

2.1.1Accès physique

L'accès physique d'une seule personne mal intentionnée qui possède une excellente connaissance de l'implémentation physique du Cloud Computing et de ses points névralgiques peut suffire à mettre hors service le Cloud, provoquant une rupture dans la

continuité du service et empêchant tout accès externe au Cloud. Les conséquences d'une telle intrusion peuvent être désastreuses [Philippe Hedde, 2010]:

- Isolement complet ou partiel du service dans le cas de coupure des liaisons d'accès.
- Perte des données en production mais aussi des données sauvegardées, sans aucune possibilité de récupération si celles-ci sont détruites ou détériorées, hors des données déjà sauvegardées (externalisation et/ou réplications).
- Risque d'incendie élevé ou d'inondation, etc.

2.1.2. Contrôle et traçabilité des accès

Point critique de la sécurité physique, le contrôle des accès doit être maitrisé, que l'on soit dans un contexte de Cloud privé, public ou privé externalisé. Dans ces deux derniers cas, c'est au client final de s'assurer que les bonnes pratiques sont mises en œuvre chez son prestataire de service/opérateur de Cloud.

Concrètement, les va-et-vient du personnel interne et des prestataires externes (informatique et télécoms, société de maintenance, de nettoyage, etc.) sont nombreux dans une salle informatique ou dans les locaux techniques.

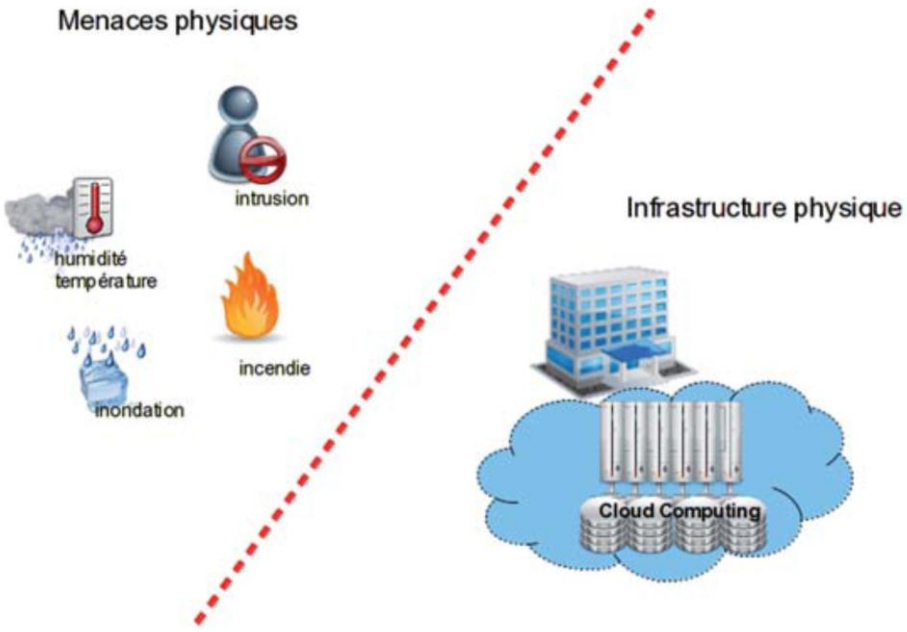
Ce flux représente une source potentielle de dysfonctionnements, volontaires ou non. Il convient de bien délimiter les zones les plus sensibles et de mettre en place des garde-fous suffisamment efficaces pour retrouver, le cas échéant, l'origine d'un incident. L'accès aux zones sensibles (serveurs, réseau, etc.) sera interdit et le passage dans les zones intermédiaires sera limité. Le personnel autorisé devra être informé du caractère sensible des zones dans lesquelles il est amené à intervenir. [Philippe Hedde, 2010]

2.1.3. Redondance matérielle

L'architecture Cloud Computing doit garantir un accès au service en très haute disponibilité avec des performances optimales. La seule défaillance d'un équipement matériel peut engendrer une dégradation ou une coupure du service voire une perte de données. Pour limiter les risques d'arrêt de service liés à la défaillance d'un équipement, il est nécessaire de le redonder.

Une réplication des configurations entre les équipements peut faciliter la bonne prise en charge de la redondance et ainsi augmenter la haute disponibilité du service. La mise en œuvre d'une redondance différentielle avec une sélection d'équipements de natures différentes

(ex : différents constructeurs, composants d'origines différentes, etc.) permet de se protéger d'un problème survenu à un équipement donné.



La figure 3.1 Sécurisation de l'environnement [Philippe Hedde, 2010]

De plus, une redondance des moyens de connexion, par la multiplication des liaisons, des opérateurs, et des chemins d'accès permet une accessibilité accrue au service en augmentant la tolérance aux pannes [Philippe Hedde, 2010].

2.1.4. Résilience

Une catastrophe d'origine humaine ou naturelle peut avoir des impacts radicaux sur le fonctionnement du Cloud Computing et amplifier une panne totale ou partielle du service. La perte totale de l'infrastructure du Cloud Computing pourrait entrainer une interruption de service d'une durée indéterminée et une perte de données irrémédiable sans possibilité de remise en service de l'infrastructure.

Une architecture de secours doit exister, sur un site géographiquement éloigné, avec des équipements redondants et permettant de réaliser un PCA (Plan de Continuité d'Activité) sans interruption de service.

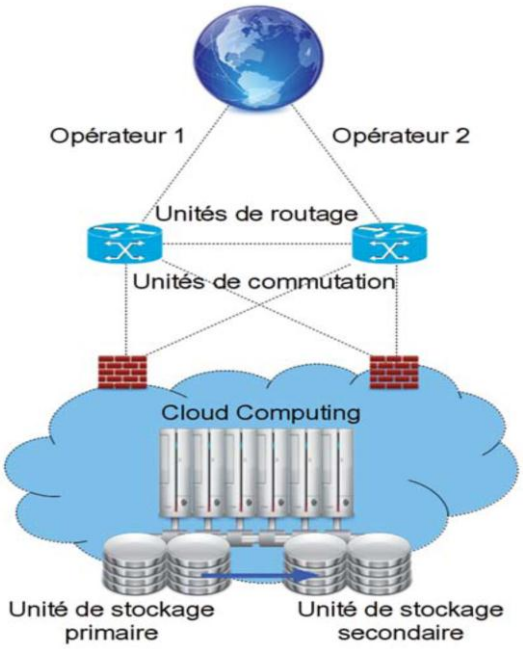


Figure 3.2 Présentation d'une architecture mono-data center [Philippe Hedde, 2010]

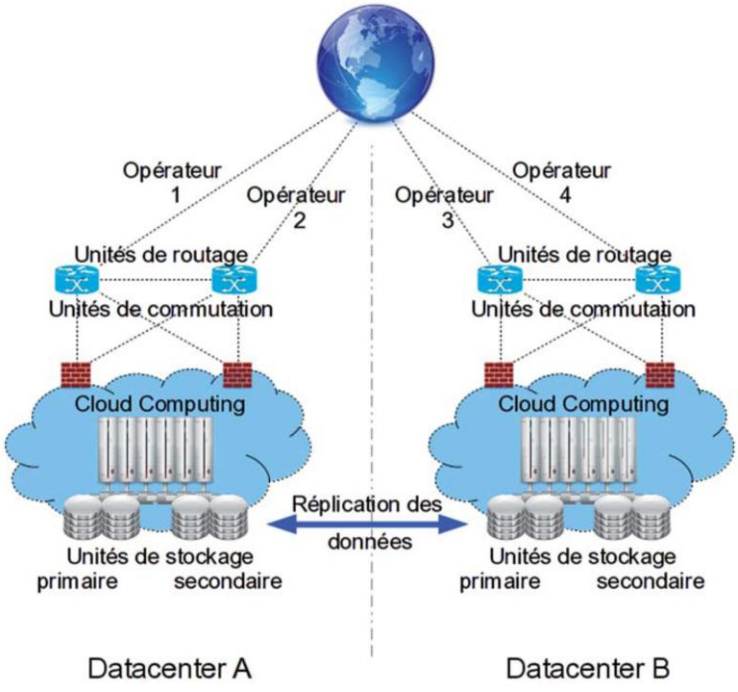


Figure 3.3 Présentation d'une architecture multi-data center [Philippe Hedde, 2010]

2.1.5. Bonnes pratiques de la sécurisation physique

- Découper les locaux informatiques en zones de sécurité concentriques, regrouper le matériel le plus sensible dans les zones les mieux protégées.
- Déporter à l'extérieur des locaux les accès de maintenance ordinaire (eau, électricité, ascenseurs, etc.)
- Eloigner les supports de sauvegarde (bandes, cassettes, CD, etc.) des locaux, si possible du bâtiment. La répartition du système de stockage sur plusieurs data centres permet de limiter les risques de perte totale du service en améliorant la tolérance de panne.
- Contrôler l'accès par des systèmes à clé, cartes, digicodes, etc. Faciles à utiliser et dont les listes d'accès sont actualisées en permanence.
- Installer des systèmes de surveillance extérieure permanente (caméras, détecteurs de présence, etc.)
- Enregistrer en vidéo les entrées et sorties (très dissuasif). L'enregistrement doit pouvoir durer entre le moment d'une intrusion et le moment de la constatation d'une malveillance, y compris pendant les congés tout en respectant les réglementations de protections liées à l'utilisation des données personnelles.
- Mettre en œuvre de bonnes pratiques pour accompagner durant leur venue les visiteurs. Nous les adapterons selon le niveau de criticité :
    - Etablir une politique d'accès générale comprenant toutes les exceptions.
    - Y soumettre tous les utilisateurs du site.
    - Identifier clairement les visiteurs par un badge spécial à durée limitée.
    - Installer des sanitaires/vestiaires pour les visiteurs.
    - Ne jamais laisser un visiteur seul se promener dans le bâtiment.
    - Tout visiteur doit avoir une autorisation d'accès délivrée par un responsable (maintenance, entretien, visites, réunions, etc.).
    - Concevoir le data center de façon à ce que la présence d'équipes de nettoyage ne soit pas nécessaire dans la salle des serveurs (mobilier antistatique, filtres à particules, etc.) [Philippe Hedde, 2010].

2.2. Sécurité logique

2.2.1. La confidentialité

La confidentialité assure que les données d'un client ne soient accessibles que par les entités autorisées. Les différentes solutions de Cloud Computing comportent des mécanismes de confidentialité comme la gestion des identités et des accès, le cryptage et l'anonymisation.

Les contrôles d'accès les plus sécurisés ne sont d'aucune protection contre un attaquant qui gagne l'accès à des informations d'identification ou des clés. Ainsi, les informations d'identification ou de gestion des clés sont des maillons essentiels dans la conception de la sécurité [Grevet, 2009].

La majorité des échangés internes ou externes au Cloud sont encapsulés en SSL (Secure Sockets Layer) et authentifiés avec un certificat généré par le client. Ce certificat n'est lié à aucune autorité de certification racine de confiance mais plutôt auto-signé par le client lui-même. Tant que ce dernier assure le contrôle de sa clé privée, le mécanisme permet un degré élevé d'assurance : seuls les clients autorisés et présentant cette clé peuvent accéder à des aspects spécifiques du service.

Le cryptage est à-priori séduisant, notamment la méthode classique à base de clé publique/clé privée : seul le destinataire de l'information peut déchiffrer la donnée qui lui est destinée avec sa clé privée, connue de lui seul, mais pas du fournisseur de la solution Cloud et moins encore d'un autre colocataire. C'est une méthode très sécurisée (selon la taille de la clé) et sélective car nous pouvons choisir de ne chiffrer que ce qui le nécessite.

Toutefois le chiffrement impose certaines réflexions quant à son implémentation. Notamment dans le cas où des traitements sont nécessaires (calcul, indexation, sauvegarde), pouvant obliger à la manipulation de données décryptées [Philippe Hedde, 2010].

Anonymisation des données est l'une des techniques de la confidentialité qui se traduisent par la conservation de l'information, ce qui rend les données inutiles pour tout le monde sauf les propriétaires.

2.2.2. L'intégrité

En plus de la confidentialité des données, les clients ont aussi besoin de se soucier de l'intégrité de leurs données. La confidentialité n'implique pas l'intégrité : les données peuvent être anonymes pour des raisons de confidentialité. L'anonymisation peut être suffisante pour la confidentialité, mais l'intégrité nécessite l'utilisation de codes d'authentification de message

(MAC). Un autre aspect de l'intégrité des données est important, surtout avec le stockage en utilisant IaaS. Une fois qu'un client a plusieurs giga-octets (ou plus) de ses données dans le Cloud, comment-il peut vérifier l'intégrité de ses données ? Il existe des coûts de transfert IaaS associés au déplacement des données dans et vers le Cloud ainsi que des considérations d'utilisation du réseau (bande passante) pour le réseau propre au client. Qu'est-ce qu'un client veut vraiment faire pour valider l'intégrité de ses données pendant que les données restent dans les Clouds sans avoir les charger. Cette tâche est encore plus difficile parce qu'elle doit être faite sans aucune connaissance explicite sur la totalité des données. Les clients généralement ne connaissent pas sur quelles machines physiques leurs données sont stockées. En outre, ces données se changent fréquemment et elles sont probablement dynamiques. Ces changements fréquents évitent l'efficacité des techniques d'assurance traditionnelles de l'intégrité [Mather et al, 2009].

2.2.3. La disponibilité

L'un des principaux avantages fournis par des plates-formes de Cloud Computing est la disponibilité robuste basée sur la redondance réalisée avec des technologies de virtualisation. Windows Azure par exemple offre de nombreux niveaux de redondance fournissant une disponibilité maximale des données et des applications. Les données sont répliquées au sein de Windows Azure sur trois nœuds distincts pour minimiser l'impact des pannes matérielles. Les clients peuvent exploiter la nature géographique de l'infrastructure Windows Azure en creusant un deuxième compte de stockage fournissant des capacités de basculement à chaud. Dans de tels scénarios, les clients peuvent créer des rôles personnalisés à répliquer et synchroniser les données entre les installations de Microsoft. Ils peuvent également créer des rôles personnalisés pour écrire des données de stockage pour des sauvegardes sur site privé.

Les agents tournant sur les machines virtuelles invitées surveillent la santé de ladite machine. Si l'agent ne répond plus, le contrôleur redémarre la machine virtuelle. Les clients pourront éventuellement choisir d'exécuter des processus de suivi de santé plus sophistiqués et adaptés à leur politique de continuité. En cas de défaillance du matériel, le contrôleur déplace l'instance du rôle vers un nouveau nœud et reprogramme la configuration réseau pour les instances de ce rôle afin de rétablir la disponibilité totale du service.

Les contrôleurs adhèrent au même principe de disponibilité grâce à la redondance et à un basculement automatique assurant la disponibilité continue des capacités de gestion des contrôleurs [Grevet, 2009]**.**

3. La Confidentialité dans le Cloud

«*Vous pouvez avoir la sécurité et ne pas avoir la confidentialité, mais vous ne pouvez jamais avoir une confidentialité sans sécurité* ». ***—Tim Mather***

Particulièrement dans les secteurs les moins réglementés, la responsabilité et l'imputabilité de la confidentialité est souvent assigné pour IT au lieu de l'unité d'affaires qui possède les données. Donc c'est quoi la confidentialité ?

3.1. Définition

Le concept de la confidentialité varie considérablement et parfois à l'intérieur des pays, cultures, et des juridictions. Il est façonné par *les expectations du public et les interprétations du droit.* Les droits de confidentialité ou les obligations sont liées à la collecte, l'utilisation, la divulgation, le stockage et la destruction des IPIs (Information personnelle d'un individu), de même, il n'y a pas de consensus universel sur ce qui constitue les données personnelles, Aux fins de cette discussion, nous allons utiliser la définition adoptée par OECD (Organization for Economic Cooperation and Development) l'Organisation de coopération et de développement économiques : « *Toute information relative à un individu identifié ou identifiable*». [Mather et al, 2009]

Une autre définition qui est fourni par AICPA (the American Institute of Certified Public Accountants) l'Institut américain des comptables publics certifiés, et par CICA (the Canadian Institute of Chartered Accountants) l'Institut Canadien des Comptables Agréés : « *Les droits et les obligations des individus et des organisations en ce qui concerne la collecte, l'utilisation, le stockage et la divulgation des informations personnelles* ».

3.2 Responsabilité juridique de la sécurité et de la confidentialité des données dans le Cloud

Le Client est juridiquement responsable de ses données et de leur utilisation, notamment de tout ce qui concerne leur conformité aux obligations juridiques.

Le prestataire est soumis à des obligations techniques et organisationnelles. Il s'engage à préserver l'intégrité et la confidentialité des données, notamment en empêchant tout accès ou utilisation fraudeuse et en prévenant toutes pertes, altérations et destructions. Sa responsabilité juridique peut être engagée dans le cas où il aurait transféré les données de son client sans l'en prévenir et sans s'assurer que les déclarations nécessaires ont été faites.

De façon générale, plus l'infrastructure est confiée au fournisseur Cloud, plus sa responsabilité est importante. Dans les cas de PaaS et de SaaS, le client ne contrôle que le contenu de ses données (et encore partiellement en SaaS où la responsabilité est partagée avec le fournisseur). En fonction du service fourni au client, le fournisseur peut être en charge (et donc responsable) de la sauvegarde, d'un niveau bien défini de disponibilité du service, et de la confidentialité des données. Même dans ce cas, le client final n'est pas affranchi de toutes responsabilités : il doit par exemple sécuriser les mots de passe ou les certificats qui lui servent à accéder à son environnement Cloud, ne pas laisser d'accès ouvert au service via son propre réseau. Les conséquences d'une négligence avérée du client final ne sauraient être imputées au fournisseur.

Le contrat de service doit aborder le domaine des responsabilités de chaque partie. Si le client doit exiger de la part son prestataire des engagements de confidentialité et les moyens de contrôle afin de surveiller que le prestataire respecte ses engagements, il est évident que ce dernier ne peut assumer pleinement la confidentialité des données confiées dans le système de stockage du Cloud. Il est nécessaire que les parties, chacune de leur côté, puissent maitriser leur propre sécurité, et d'autre part, contrôler le domaine tiers. Faute de contrats formalisant ces points de partage des responsabilités, et des outils de contrôle et d'enregistrement, des procédures en cas de litiges peuvent être longues et laborieuses, surtout dans un domaine où la jurisprudence est rare.

3.3. Données de Cloud accessibles aux autorités d'un autre pays

Tout pays a le droit légitime d'avoir accès, dans les conditions juridiques qui lui sont propres, aux données qui sont stockées sur son territoire, ou qui transitent par celui-ci. Ainsi, en France par exemple, dans le cadre d'une perquisition et conformément à l'article 97 du code de procédure pénale, l'hébergeur doit être en mesure d'extraire de son Cloud les éléments recherchés ou l'ensemble des informations concernant un client particulier, sans pour autant avoir à livrer l'ensemble des données des clients hébergés dans le Cloud.

La bonne pratique consiste à s'assurer contractuellement du (ou des) pays où seront physiquement installés les éléments d'infrastructures et de connaître, avant de s'engager dans un contrat de service Cloud avec un fournisseur, les juridictions compétentes. [Philippe Hedde, 2010]

3.4. Les axes de recherches liés à la confidentialité dans le Cloud

Cette partie donne un aperçu sur les problèmes liés à la confidentialité ainsi que quelques solutions proposées dans le Cloud. Dans notre classification des problèmes, nous nous sommes basés sur les travaux de [Chen et Zhao, 2012] et [Vimercati et al, 2012], qui ont analysé les problèmes de la confidentialité dans le Cloud selon le cycle de vie des données.

3.4.1. Le cycle de vie des données

Les informations personnelles d'un individu IPI doivent être gérées dans le cadre des données utilisées par l'organisation. *Elles doivent être gérées à partir du moment où l'information est conçue jusqu'à sa destruction finale*.

Le cycle de vie des données peut être décomposé en *sept* phases : génération, transfère, utilisation, partage, stockage, archivage et enfin destruction. Bien que ce cycle soit présenté de façon séquentielle, les données peuvent très bien le suivre de façon non-séquentielle. Cette approche peut aussi être utilisée dans le cas d'un système "classique", c'est-à-dire "non Cloud". Dans un tel contexte les contrôles de sécurité seront différents.

La protection des informations personnelles doit considérer l'impact du Cloud sur chacune des phases suivantes comme *la figure 3.4* indique :

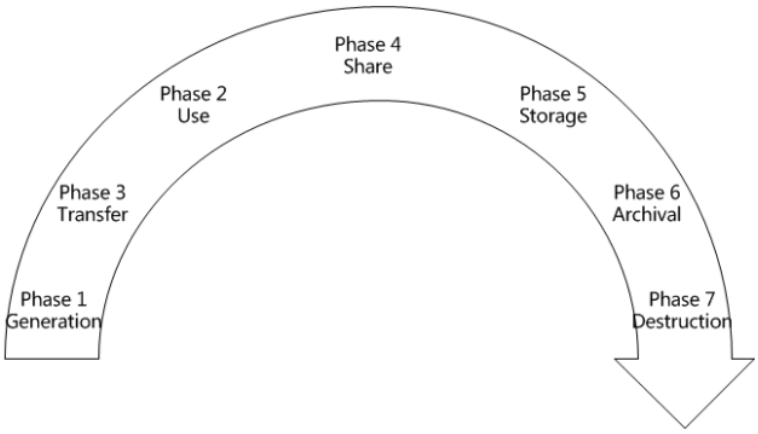


Figure 3.4 Le cycle de vie d'une donnée [Chen et Zhao, 2012]

La définition de chacune de ces phases est comme suit :

- **Génération de l'information**

Cette phase consiste en la création de nouvelles données ou de modifications significatives de données existantes. Les contrôles sécurité applicables à cette phase sont la classification (attribution d'un label, d'un niveau de sensibilité) et la définition des droits associés.

[Roy et al, 2010] ont appliqué un contrôle décentralisé des flux d'information où la technologie de protection de la confidentialité différentielle dans la génération des données est faite en mettant un système de protection de la confidentialité appelée *Airavat.* Ce système peut empêcher la fuite de la confidentialité sans l'autorisation du processus de calcul *Map-Reduce*. [Rabl et al, 2011] présentent un générateur de données pour des applications moyennes en Cloud. Leur architecture rend le générateur de données facile à étendre et à configurer. Le haut degré de parallélisme qui se considère comme l'élément clé, permet à un passage à l'échelle linéaire pour un nombre arbitraire de nœuds.

- **Le Transfer**

À l'intérieur des frontières de l'entreprise, la transmission des données ne nécessite généralement pas de chiffrement, ou tout simplement avoir une mesure simple de cryptage des données. Pour la transmission de données à travers les frontières de l'entreprise, la confidentialité et l'intégrité des données doivent être assurées afin d'empêcher que les données soient exploitées et falsifié par des utilisateurs non autorisés. La confidentialité et l'intégrité des transmissions de données doivent être assurées non seulement entre le stockage d'entreprise et le stockage de Cloud, mais aussi entre les différents services de stockage dans le Cloud.

[Porwal et al, 2012] ont proposé une approche pour sécuriser la transmission des données dans un Cloud privé**.** Dans leur méthode proposée, Au lieu d'utiliser SSL, les données transférées sont cryptées dans la couche supérieure au-dessus de la couche de transport. Le schéma pour l'amélioration de la performance peut être appliqué sans aucune modification de l'implémentation de la couche IP, et des communications efficaces et sécurisées sont réalisées par un prétraitement de cryptage dans la couche supérieure. La sécurité est appliquée sur les données du contexte en utilisant les algorithmes de cryptage. [Marconet al, 2011] mettent l'accent sur le problème des prix élevés de la bande passante facturés par les fournisseurs de Cloud pour le chargement de données.

- **Utilisation**

La phase d'utilisation regroupe les actions qu'un utilisateur réalise sur les données. La différence entre cette phase et la phase "*partage*" réside dans le fait qu'ici l'utilisateur est seul à manipuler les informations. Pour les contrôles, on retrouvera les grands classiques que sont la gestion des droits, la supervision des actions et les contrôles logiques au niveau applicatifs.

Les propriétaires de données privées doivent s'en concentrer et s'assurer que l'utilisation des informations personnelles est compatible avec les objectifs de la collecte d'informations et de savoir si ces informations sont partagées avec des tiers, par exemple, les fournisseurs de services de Cloud.

[Bowers et al, 2009] ont proposé un outil de gestion de confidentialité basé sur le client, ils ont fourni un modèle de confiance centré sur l'utilisateur pour l'aider à contrôler le stockage et l'utilisation de ses informations sensibles dans le Cloud. [Mulero et Nin, 2009] ont discuté les problèmes qui existent dans la technologie de la protection de confidentialité (comme les graphes d'anonymisation) quant appliqué à des données volumineuses.

- **Le partage**

Le partage est une phase qui inclue toutes les actions visant à partager les données avec d'autres utilisateurs, des clients ou des partenaires. Le partage de données élargit l'intervalle d'utilisation des données et rend les permissions de données plus complexe. Les propriétaires de données peuvent autoriser l'accès aux données pour une partie, et à son tour le parti peut partager aussi les données à un tiers sans l'accord des propriétaires des données. Par conséquent, au cours de l'échange de données, les propriétaires des données doivent se demander si la troisième partie garde le maintient des mesures de protection d'origine et les restrictions d'utilisation.

[Randike et al, 2011**]** ont proposé un Framework pour la protection de confidentialité basant sur les composants de la responsabilité d'information où l'agent peut identifier les utilisateurs selon le type d'information utilisé. Quand une mauvaise utilisation inappropriée est détectée, l'agent définit un ensemble de méthodes pour tenir les utilisateurs responsables d'abus.

- ## Le stockage

Le stockage est le fait d'enregistrer les données dans une structure ou un système (fichiers, base de données...). Du coté des contrôles, nous allons retrouver des moyens de contrôle d'accès, du chiffrement et enfin des techniques de découverte ou de surveillance des données.

[Hyun et al, 2012] proposent un schéma efficace et adaptable aux qualités saillantes. Ce schéma réalise la correction de stockage de données, permettant à l'utilisateur authentifié d'accéder aux données en localisant l'erreur de ces données (C.à.d. l'identification des serveurs mal conduite). [Wang et al, 2009] proposent un souple mécanisme distribué de stockage de l'intégrité de l'audit, en utilisant « the homomorphic token and distributed erasure coded data », Le modèle proposé permet aux utilisateurs d'auditer le stockage dans le Cloud avec une communication très léger et un faible coût de calcul. Il supporte aussi les opérations dynamiques, sécurisé et efficaces sur les données externalisées.

[Amar et al, 2012] proposent des différents schémas pour la fragmentation des données pour le stockage multiple dans le Cloud qui vise à fournir à chaque client une fiabilité, disponibilité et de meilleures décisions de stockage de données.

- ## Archivage

La phase d'archivage consiste à transférer les données vers un support de stockage pour qu'elles y soient conservées sur une durée plus ou moins longue. Les contrôles sont principalement le chiffrement et la gestion des supports. Si les fournisseurs de services de Cloud ne fournissent pas l'archivage *hors site*, la disponibilité des données sera menacée. Si la durée de stockage est compatible avec les exigences d'archivage ? Sinon Cela pourrait provoquer des menaces de la disponibilité ou la confidentialité**.**

- ## La Destruction

Lorsque les données ne sont plus nécessaires, est ce qu'elles sont complètement détruite ? En raison des caractéristiques physiques du support de stockage, les données supprimées peuvent encore exister et peuvent être restaurées. Cela peut conduire à divulguer par inadvertance des informations sensibles.

Le Ministère de la Défense américain présente deux méthodes approuvées pour la sécurité des données détruit et ils n'ont fourni aucunes exigences spécifiques pour la façon dont ces méthodes peuvent être atteintes **[**DoD, 2006**].** L'Institut National des Standards et de la

Technologie (NIST) donne des lignes directrices pour la désinfection des médias **[**Richard et al, 2006**].**

[Vimercati et al, 2012] ont caractérisé les différents aspects des problèmes de la confidentialité, et ils ont illustré les risques, et les différents problèmes ouverts qui ont une relation avec la confidentialité des utilisateurs accédant aux services dans le Cloud

3.4.2. Les risques de confidentialité dans les scénarios Cloud

- **Les communications anonymes**

Un utilisateur peut souhaiter envoyer un ou plusieurs messages à un destinataire donné, ou espérer recevoir des réponses, à ses messages. Les réponses peuvent arriver longtemps après l'envoi de ses messages ou de façon presque immédiate, et dans certains cas les réponses peuvent même être entrelacées avec les envois. La difficulté pour rester anonyme dans chacun de ces cas est croissante, et malheureusement dans le Cloud, le cas le plus courant est le dernier. Des protocoles de communication anonymes sont proposés pour régler cette situation [Ardagna et al, 2010]. Des solutions spécifiques prenant l'avantage des particularités des systèmes de Cloud sont également à l'étude **[**Jones et al, 2011**]**.

- **L'exécution collaborative des requêtes**

Pendant la distribution de l'information et la coopération du calcul des résultats des requêtes, la confidentialité des informations sensibles doit être prise en considération. La nécessité d'une partie de publier des informations et à coopérer avec un autre, peut caractériser plusieurs scénarios : allant des systèmes de bases de données distribuées traditionnelles (où l'administration centralisée d'une base de données est distribuée) aux systèmes de Cloud (où les différents serveurs de Cloud collaborent et intègrent leurs données et leurs services pour fournir aux utilisateurs des applications disponibles n'importe où et n'importe quand). Cette situation fait appel à des solutions innovantes supportant un partage sélectif des informations stockées sur plusieurs serveurs Cloud, même au-delà des frontières d'administratives et de l'entreprise [Benedikt et al, 2012] [Li, 2003].

3.4.3. Les risques de confidentialité pour l'utilisateur

Ils se concentrent sur les problèmes de la protection des identités des utilisateurs qui ont l'accès aux ressources dans le Cloud

- **Le contrôle d'accès basé sur Attribut**

Il permet aux utilisateurs d'interagir avec les serveurs de Cloud et de distribuer leurs informations sans divulguer leurs identités. Les approches traditionnelles pour l'accès aux ressources sont basées sur l'authentification des utilisateurs [Cimato et al, 2008] [Sandhu et Samarati, 1997] (ces approches ne sont pas adoptées dans le Cloud où les parties qui interagissent peuvent être inconnus les uns aux autres). Par la suite ils ont utilisé des solutions basées sur la spécification [Ardagna et al, 2011] [Lee et al, 2008] où les utilisateurs peuvent facilement accéder à toutes les ressources disponibles à partir (éventuellement inconnue) des serveurs sans avoir besoin de se souvenir de leurs mots de passe ou de gérer un compte spécifique pour chacun des serveurs sur lesquels ils accèdent.

- **Les préférences de confidentialité de l'utilisateur**

Elles précisent quelle information est préférable à divulguer pour accéder à un service en fonction de la sensibilité d'information (préférence entre une carte d'identité et un passeport pour accéder à un service qui exige l'un des deux). [Ardagna et al, 2010] et [Yao et al, 2008] se sont deux solutions qui ont touché cet aspect. [Chen et al, 2005] ont proposé un modèle qui permet à un utilisateur d'associer un coût différent avec chaque Accréditation dans son portefeuille représentant sa sensibilité (C.à.d. plus l'accréditation est sensible plus le coût est élevé) et de minimiser le coût total d'un processus de négociation. [Yao et al, 2008] proposent un modèle de gestion de la confiance basé sur des points où un utilisateur appelle son accréditation avec un score de confidentialité quantitative, et le serveur définit un crédit pour chaque accréditation que les utilisateurs peuvent posséder. Le serveur permet à un utilisateur d'accéder à une ressource si la somme des crédits d'accréditations publiées atteint un seuil fixe. [Karger et al, 2008] proposent un langage basé sur la logique pour la spécification des préférences de confidentialité qui déterminent un ordre partiel entre les propriétés des utilisateurs.

3.4.4. Les risques de confidentialité pour les données stockées

Depuis que la donnée est stockée dans les serveurs de Cloud (qui ne sont pas sous le contrôle du propriétaire de donnée) la confidentialité peut être mise en risque ! En plus, l'accès à la donnée doit être régulier (les différentes parties peuvent avoir des privilèges d'accès différents).

- **La confidentialité et l'intégrité**

Le problème de la préservation de confidentialité est adressé dans le scénario DaaS (Database as a Service), la solution proposée par [Samarati et al, 2010] consiste au cryptage des données avant leur stockage dans le serveur. Des fois, les données ne sont pas sensibles, mais ce qui est sensible est leurs associations avec d'autres informations. Donc le cryptage semble qu'une « OverKill » (Ex : La liste des noms des patients et la liste des maladies traités dans un hôpital peut être rendu publique alors que l'association du nom de chaque patient à une maladie spécifique doit être protégé). [Aggarwal et al-a-, 2005] [Ciriani et al, 2009] se sont des solutions proposés pour protéger les associations sensibles (elles sont basées sur la combinaison du Cryptage et de la Fragmentation).

- **Accès sélectif**

Dans un scénario de Cloud, ni le propriétaire des données, ni le serveur qui peuvent appliquer la politique de contrôle d'accès. Le serveur ne peut pas directement appliquer les restrictions de contrôle d'accès parce qu'il est non confiant de les appliquer, et aussi parce que la politique réglementant l'accès aux données peut dépendre du contenu des données (qui doit rester toujours confidentielle du serveur). Le propriétaire a besoin de faire pour chaque demande d'accès un filtrage, ce qui abandonne les avantages du Cloud. Donc il est nécessaire de concevoir un mécanisme tel que les données elles-mêmes appliquent des restrictions sur l'ensemble d'utilisateurs qui peuvent les accéder. [Ciriani et al, 2012] [Capitani. et al, 2007] se sont deux solutions basées sur « le cryptage sélective » qui consiste à crypter les portions différents des données en utilisant des clés différents, et dans la distribution des clés aux utilisateurs de telle sorte que chaque utilisateur peut décrypter tout et seulement les éléments d'information autorisés.

4. Conclusion

Le modèle Cloud Computing propose plus de choix, de flexibilité, d'efficacité opérationnelle et permet aux entreprises comme aux individus de réaliser d'avantage d'économies. Pour profiter pleinement de tous ces bénéfices, les utilisateurs doivent disposer de garanties fiables concernant la confidentialité et la sécurité de leurs données en ligne. Dans ce chapitre, nous avons abordé deux parties essentielles : La première partie qui prend en

considération l'aspect sécurité physique et logique, et la deuxième partie qui a pour objectif de présenter l'aspect confidentialité ainsi que quelques approches touchant cet aspect.

Dans le prochain chapitre, nous présenterons tous ce qui concerne l'aspect anonymisation qui est considéré comme l'une des techniques de la confidentialité.

Chapitre 4

La préservation de la confidentialité : Anonymisation

1. Introduction

La collecte de l'information numérique par les gouvernements, les entreprises et les particuliers a créé d'énormes possibilités pour la prise de décision basée sur la connaissance. Poussé par des avantages mutuels ou par des règlements qui exigent que certaines données soient publiées, une demande pour l'échange et une publication de données entre les différentes parties sont nécessaires. Les hôpitaux agréés en Californie à titre d'exemple sont tenus de présenter des données démographiques spécifiques sur chaque patient déchargé de ses installations [Carlisle et al, 2007]. En Juin 2004, le Comité consultatif sur les technologies de l'information a publié un rapport intitulé : La révolution de la santé via la technologie de l'information (Président de la comité consultatif sur les technologies de l'information 2004). L'un de ses points clés était d'établir un système national de dossiers médicaux électroniques qui encourage le partage des connaissances médicales par le soutien d'une décision clinique assistée par l'ordinateur. Les données détaillées qui spécifient un individu contiennent souvent des informations sensibles, ce qui fait que la publication de ces données risque de violer sa confidentialité. La pratique actuelle repose sur le principe des politiques et des lignes directrices qui restreignent les types de données publiables et ainsi sur les accords et le stockage des données sensibles. La limitation de cette approche nécessite un niveau de confiance plus élevé dans la plus part des scénarios de partage de données [Fung et al, 2010].

La tâche la plus importante est de développer des méthodes et des outils pour la publication de données de sorte que ces données doivent rester pratiquement utiles tout en préservant leurs confidentialités. Ce concept est appelé : PCPD, ***la Préservation de la Confidentialité pour les Données Publiées*** (Privacy preserving data publishing,). Au cours des dernières années, il existe plusieurs travaux qui ont essayé de répondre à ce défi et ont proposé de nombreuses approches.

Notre vision dans ce chapitre est de discuter et de donner quelques définitions sur ce qui concerne l'approche de l'anonymisation, les types et les opérations sur lesquelles se déroule parfaitement l'anonymisation et ainsi la différence qui existe entre le terme de cette anonymisation et celui du cryptage.

2. La préservation de la confidentialité pour les données publiées

La disponibilité d'un nombre important de bases de données qui contiennent une grande variété d'informations d'un individu, augmente le souci de la confidentialité dans le monde numérique moderne. Ceci permet de découvrir les informations d'un individu spécifique par une simple connexion d'un certain nombre de bases de données disponibles.

La confidentialité d'un individu peut être conservée dans de nombreux aspects qui se diffèrent dans la propriété et dans la limitation. Les principaux aspects de la confidentialité sont les suivants [Vallet, 2012]:

- **L'Anonymat**

Un individu ne doit pas être identifiable parmi un ensemble d'objets des données.

- **La non-chaînabilité (unlinkability)**

La non-chaînabilité représente l'impossibilité (pour d'autres utilisateurs) d'établir un lien entre différentes opérations réalisées par un même utilisateur ; par exemple, en interdisant la fusion ou le croisement d'informations à partir de différents fichiers ou bases de données.

- **Pseudonymat (pseudonymity)**

La pseudonymisation consiste à supprimer les champs directement identifiants des enregistrements, et à rajouter à chaque enregistrement un nouveau champ appelé *pseudonyme*, dont la caractéristique est de rendre impossible tout lien entre cette nouvelle valeur et la personne réelle. Un pseudonyme est une autre information généralement non personnelle qui doit être associée à une information personnelle, et c'est ainsi que quand le lien est découvert, le pseudonyme devient une information personnelle. Alors il est à noter qu'une faiblesse au niveau de ce lien surgisse entre la vraie identité et le pseudonyme, néanmoins une utilisation classique des pseudonymes peut se faire dans l'authentification de sujets sans que leur identité ne soit dévoilée. Un pseudonyme par exemple peut être une clé publique pour les signatures numériques.

- **La non-observabilité (Unobservability)**

La non-observabilité garantit qu'un utilisateur peut utiliser une ressource ou un service sans que d'autres utilisateurs soient capables de déterminer si une opération (l'utilisation d'une ressource ou d'un service) est en cours.

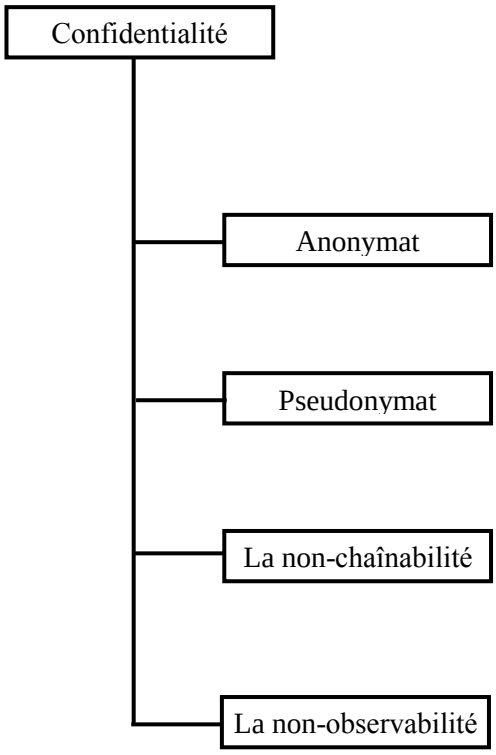


***Figure 4.1 Les aspects de la confidentialité** [Tinbo, 2013]*

La préservation de la confidentialité pour les données publiées est l'un des grands domaines de la confidentialité qui traite l'aspect de l'anonymisation des données. La PCPD a deux phases principales : La phase de la collecte des données et la phase de la publication des données comme est décrit dans *la figure 4.2*

Dans la phase de la collecte de données, l'éditeur de données (data publisher) collecte les données provenant des propriétaires de records (record owners). Dans la phase de la publication de données, l'éditeur de données publie les données collectées à un explorateur de données qui est considéré comme un destinataire (Data recipient) et qui effectuera par la suite une extraction sur ces données. Par exemple, un hôpital collecte les données des patients et les publie à un centre médical externe. Dans ce cas, l'hôpital est l'éditeur des données, les patients sont les propriétaires des records et le centre médical est le destinataire. L'exploration

de données effectuée au centre médical pourrait être un simple comptage du nombre d'hommes atteints de diabète.

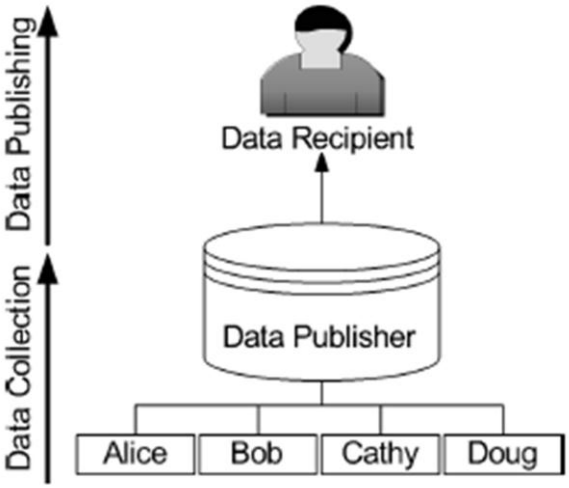


Figure 4.2. Les phases de la préservation de la confidentialité pour les données publiées

Il existe deux modèles d'éditeurs de données [Gehrke, 2006] :

- ***le modèle non confiant :*** l'éditeur de données n'est pas confiant et peut identifier les informations sensibles des propriétaires de records. Plusieurs solutions des cryptographies [Seethal, 2013], [Sakshi et Bamnote, 2015] et des communications anonymes [Pacheco et Puttini, 2012], sont proposées pour la collecte des records anonymes des propriétaires sans aucune révélation d'identités.
- ***Le modèle confiant*** : l'éditeur de données est confiant et les propriétaires de records sont prêts à lui fournir leurs informations sensibles.

Dans la pratique, chaque scénario de la publication de données a ses propres hypothèses et exigences pour le but de l'éditeur, le destinataire et la publication de données. Certaines propriétés et des hypothèses qui peuvent être traitées dans la publication de données pratiques sont les suivantes :

Éditeur de données non expert

L'éditeur de données ne doit pas avoir des connaissances pour effectuer l'exploration de données au nom du destinataire des données. Toutes les activités d'exploration de données doivent être effectuées par le destinataire des données après la réception des données de l'éditeur de données. Parfois, l'éditeur de données n'a aucune connaissance à propos de la

localisation des destinataires au moment de la publication. Par exemple, les hôpitaux de la Californie publient les records des patients sur le Web sans aucune connaissance de la localisation des destinataires et comment ils vont utiliser ces données [Carlisle et al, 2007]. Ils publient les records des patients dû aux réglementations ou au financement de la recherche médicale générale, et non pas au besoin du résultat de l'exploration de données. Par conséquent, et dans tel scénario de publication, l'éditeur de données ne fait pas plus qu'anonymiser les données.

Dans d'autres scénarios, l'éditeur de données est intéressé par le résultat de l'exploration de données, mais il manque de l'expertise interne pour effectuer l'analyse, donc il doit externaliser les activités d'exploration de données à certains explorateurs externes. Dans ce cas, la tâche d'exploration de données effectuée par le destinataire est connue au préalable.

Le destinataire des données peut être un attaquant

Parmi les scénarios que nous pouvons trouver dans le domaine de PCDP est qu'un destinataire, qui peut être considéré comme un attaquant. Par exemple, une société de recherche pharmaceutique est une entité confiante ; cependant, il est difficile de garantir que tout son personnel est confiant aussi. Cette hypothèse fait les problèmes et les solutions de PCDP différents par rapport à des approches de la cryptographie dans laquelle, seulement les destinataires autorisés et confiants qui ont le droit d'avoir la clé privée pour accéder au texte. Le défi majeur dans le PCDP est de faire simultanément :

- *Préserver la confidentialité des données anonymes.*
- *Préserver l'utilité d'informations personnelles.*

La publication des données, et non pas le résultat de l'exploration de données

PCDP met l'accent sur la publication des records des individus. Clairement, cette exigence est plus stricte que la publication des résultats de l'exploration de données, tels que les classifications, les règles d'association. Par exemple, dans le cas de la publication des données de ***Netflix***, les informations utiles peuvent être un certain type d'associations pour le classement des films. Cependant, Netflix a décidé de publier les records de données au lieu de telles associations parce que les participants, avec des records de données, ont une plus grande flexibilité dans l'exécution de l'analyse exigée et de l'exploration des données ; tels que les modes d'exploration dans une partition, et non aux d'autres partitions ; la visualisation des transactions contenant un motif spécifique [Barbaro et Zeller, 2006].

3. L'approche : Anonymisation

Les organisations privées publient leurs données sur le Cloud pour une recherche ou pour d'autres fins. La confidentialité de ces données doit être préservée. C'est à dire aucune information sensible ne doit être divulguée. Anonymisation des données [Cox, 1980] est l'une des techniques de la confidentialité qui se traduisent par la conservation de l'information, ce qui rend les données inutiles pour tout le monde sauf les propriétaires [Jeff, 2012], Les données ont également appelé micro données qui sont stockées dans une table sous la forme ***(Attributs explicites, QID, Attributs sensibles)*** et qui possèdent des records multiples. Ces records peuvent être classés comme :

- ***Identifiants explicites :*** les attributs contenant des informations qui identifient explicitement les propriétaires des records (individu), ex : Nom, prénom
- ***Quasi- Identifiants QID***: les attributs qui peuvent être liés à d'autres informations pour identifier un individu, ex : date de naissance.
- ***Identifiants sensibles*** : les attributs avec une valeur sensible, ex : salaire.

L'anonymisation [Cox, 1980] se réfère à l'approche de PCDP qui cherche à cacher l'identité et / ou les données sensibles des propriétaires. [Sweeney-C-, 2002] a montré à Monsieur *William Weld*, (ancien gouverneur de l'État du Massachusetts) une menace qui touche la confidentialité. Dans l'exemple de *Sweeney*, le nom d'individu qui est dans une liste électorale publique était lié dans une base de données médicale à travers une combinaison de code postal, date de naissance et son sexe *(figure 4.3).* Chaque attribut ne permet pas d'identifier de manière unique le propriétaire, mais la combinaison d'attributs (*quasi-Identifiants*) fait distinguer l'identité d'individu les uns par rapport aux autres. Selon [Sweeney-C-, 2002], 87% de la population des Etats-Unis avait déclaré des caractéristiques qui les rendaient probablement distinguer en se basent seulement sur ces quasi-Identifiants.

Dans l'exemple *(figure 4.3),* le propriétaire est ré -identifié par la liaison de son quasi-identifiant *(Record linkage (Liaison entre enregistrements)).* Pour effectuer de telles attaques de liaison, l'attaquant a besoin de connaître préalablement le record de la victime et son quasi-identifiant. Ces connaissances peuvent être obtenues par l'observation.

Par exemple, l'attaquant a remarqué que son patron a été hospitalisé, et donc il savait que le record médical de son patron apparaîtrait dans la base de données des patients. En outre, il n'a pas été difficile pour l'attaquant d'obtenir le code postal de son patron, date de naissance et le sexe pour faire une attaque de liaison.

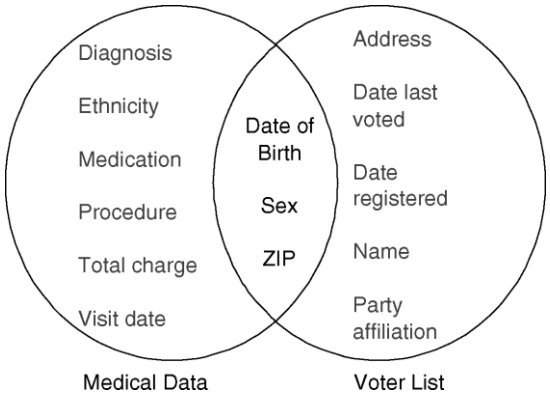


Figure 4.3. La ré-identification des propriétaires par la liaison [Fung et al, 2010]

Pour empêcher ce type d'attaque, l'éditeur de données doit fournir une table anonyme sous la forme : ***T' (QID', Attributs sensibles)***

QID' est une version anonyme de la QID originale obtenue en appliquant des opérations d'anonymisation aux attributs QID dans la table d'origine T. Les opérations d'anonymisation cachent quelques informations détaillées de sorte que plusieurs records viennent de se distinguer par rapport à QID'**.**

*Le problème de l'anonymisation est de produire une table anonyme T' qui satisfait une exigence de confidentialité donnée et déterminée par un modèle de confidentialité choisi et de conserver autant que possible l'utilité de données**.** Une métrique de l'information est utilisée pour mesurer l'utilité d'une table anonyme* [Fung et al, 2010]**.**

Dans le contexte de la confidentialité, les auteurs dans [Diaz et al, 2002] ont séparé deux types d'anonymisation : Anonymisation de connexion et Anonymisation de données *(figure 4.4).*

3.1. Anonymisation de connexion

Elle met l'accent sur la protection de l'identité de la partie source et de la partie destination durant une communication. Dans ce terrain, la plupart des travaux dans l'anonymisation de connexion touchent la notion d'authentification anonyme.

- **L'authentification anonyme**

L'authentification anonyme est une technique permettant aux utilisateurs de prouver qu'ils ont le privilège sans aucune divulgation des identités [Chai et al, 2006]. Les auteurs [Sushmita et al, 2012] proposent un nouveau schéma de contrôle d'accès décentralisé pour le stockage

sécurisé des données dans le Cloud, qui vérifie l'authenticité du service sans connaître l'identité de l'utilisateur avant le stockage des données, [Rajalekshmi et Lashma, 2015] réalisent une technique de contrôle d'accès décentralisé avec l'authentification anonyme des données qui sont stockées dans des serveurs de Cloud multiples.

- **Lien sémantique**

Il se base sur la relation sémantique entre le fournisseur et le client [Pacheco et Puttini, 2012].

3.2. Anonymisation des données

- **Anonymisation des données statiques**

La plupart des travaux d'anonymisation des données ont été effectuées sur des ensembles de données statiques.

Centralisé : Elle met l'accent à des bases de données centralisé

Décentralisé : Concerne tous les travaux qui touchent les bases des données décentralisé

- **Anonymisation des données dynamiques**

Généralement, une base de données est dite dynamique si et seulement si ses données sont différentes à des moments différents [Zhang et Bi, 2010]. Cette différence est due à deux types de mises à jour :

Mise à jour externe

Pour chaque entier *i* et *j* ($0 \leq i < j$), si l'enregistrement *t* ($t \neq \varphi$) satisfit l'une des conditions suivantes :

1) $t_i \in T_i$ et $t_j \notin T_j$
2) $t_i \notin T_i$ et $t_j \in T_j$

On dit que *t* est une mise à jour externe de T_j contrairement à *Ti.* En se basant sur cette définition, il y a deux types spécifiques d'une mise à jour externe : L'insertion et la suppression qui correspondent aux conditions 1 et 2 respectivement.

Mise à jour interne

Pour chaque entier *i* et *j* ($0 \leq i < j$), supposant pour un enregistrement *t* que $t_i \in T_i$ et $t_j \in T_j$. si t_i et t_j satisfont au moins l'une des conditions suivantes :

$$t_i[Q] \neq t_j[Q]$$
$$t_i[S] \neq t_j[S]$$

Alors, nous disons qu'il existe une mise à jour interne sur *t* dans la période *[i, j]*.

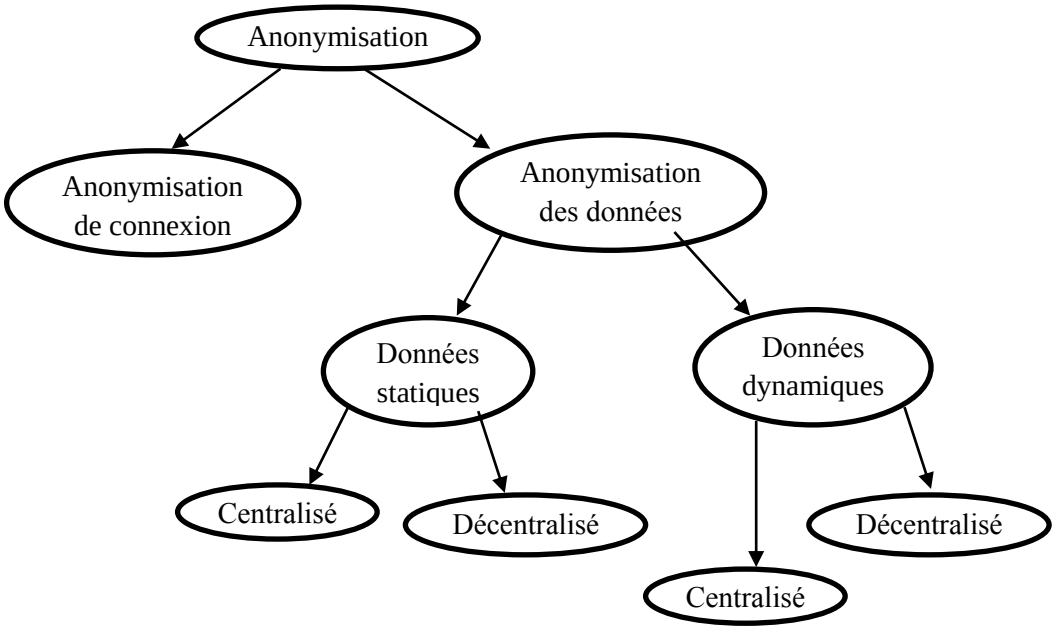


Figure 4.4. La classification des anonymisations

4. Les opérations d'anonymisation

Généralement, la table d'origine ne satisfait pas les exigences spécifiques de la confidentialité, donc le tableau doit être modifié avant d'être publié. Cette modification est faite par l'application des séquences d'opérations d'anonymisation. Ces opérations comprennent la généralisation, la suppression, la dissimulation et la permutation [Tinbo, 2013]**.**

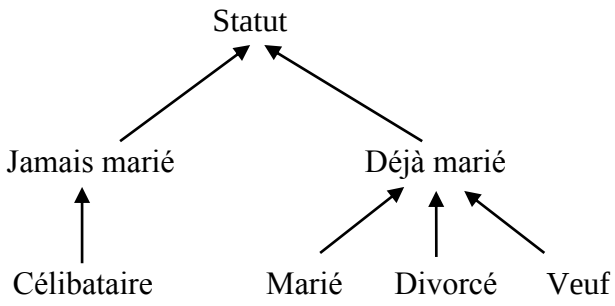


Figure 4.5. La hiérarchie de la généralisation de l'attribut : Marital Statut

4.1. Généralisation et suppression

La généralisation consiste à généraliser, ou diluer, les attributs des personnes concernées en modifiant leur échelle ou leur ordre de grandeur respectif dans la taxonomie des attributs. L'opération inverse de la généralisation est appelée la spécialisation. Le nœud père « Déjà marié » est plus général que les nœuds fils « Marié », « Divorcé » et « Veuf ». Pour l'attribut numérique, la valeur peut être remplacée par un intervalle qui le couvre. Le nœud racine « Statut » représente la valeur la plus générale de l'attribut.

Le *Bottom-up (ascendante)* et *top-down (descendante)* sont les principales stratégies de recherche utilisées pour traverser le long des hiérarchies de la généralisation. Dans la stratégie *ascendante*, l'algorithme commence par la table originale où les valeurs d'attributs sont remplacées par d'autres valeurs qui sont plus générales jusqu'au le modèle d'anonymisation soit atteigne. Dans la deuxième stratégie, la table est spécialisée à partir de l'état la plus générale où toutes les valeurs d'attributs ont les valeurs les plus généralisées dans la taxonomie de la généralisation. Dans chaque étape, ces valeurs généralisées sont remplacées par des valeurs qui sont moins générales en effectuant des contrôles pour déterminer la violation de l'exigence d'anonymisation. Le processus de spécialisation se termine quand le modèle d'anonymisation atteint son besoin [Bayardo et Agrawal, 2005]**.**

Dans ce travail, nous avons utilisé l'approche *ascendante* pour une meilleure efficacité de généralisation (une valeur minimum de la perte d'information) par rapport à l'autre approche. Notre choix est dû par ce que la stratégie *descendante* commence à partir de la valeur la plus générale et se termine quand l'exigence de l'anonymisation est violée. Ce processus peut arrêter au point où les données sont à un niveau de généralisation plus élevé ce qui augmente la perte d'information. Contrairement à notre stratégie qui se termine au point où les données sont bien spécialisées tout en atteignant l'exigence du modèle d'anonymisation.

[Lefevre et al, 2005] [Sweeney-A-, 2002] [Samarati, 2001] ont utilisé le schéma de généralisation de full-domaine. Dans ce schéma, toutes les valeurs d'attribut sont généralisées au même niveau de la taxonomie de l'arbre, par exemple dans *la figure 4.5*, si *Marié, Divorcé* et *Veuf* sont généralisé *à Déjà marié,* alors il faut que *Célibataire* soit généralisé à *Jamais marié.*

Suppression

La suppression est une approche qui consiste à remplacer certaines valeurs par une valeur spéciale, indiquant que les valeurs remplacées ne sont pas divulgués. Le but de cette opération est de réduire les valeurs de l'attribut. L'opération inverse de la suppression est appelée la divulgation.

4.2. La dissimulation des données

La dissimulation supprime une valeur de donnée et la remplace par une valeur '0 ', par exemple dans une base de données d'un hôpital, l'âge d'un patient ne peut pas être nécessaire pour le traitement, alors il sera remplacé par un constant '0' [Jeff, 2012].

Les auteurs [Abbasy et Shanmugam, 2011] ont utilisé la technique de dissimulation des données pour le partage des ressources dans les environnements Cloud basant sur les séquences d'ADN. L'objectif principal est de proposer un algorithme permettant d'implémenter les données cachées dans des séquences d'ADN. Dans le travail de [Aravinth et al, 2013], les auteurs ont utilisé la dissimulation des données pour les images en utilisant Spread Spectrum dans le Cloud.

4.3. La permutation des données

La permutation peut être considérée comme une forme spéciale d'ajout de bruit. Dans une technique de bruit classique, les attributs sont modifiés au moyen de valeurs aléatoires. La production d'un bruit cohérent peut se révéler une tâche difficile et le simple fait de modifier légèrement les valeurs des attributs risque de ne pas garantir la confidentialité adéquate. Au lieu de quoi, les techniques de permutation altèrent les valeurs au sein de l'ensemble de données en les échangeant simplement d'un enregistrement à un autre. Cet échange garantira que la fourchette et la distribution des valeurs resteront les mêmes, mais non les corrélations entre les valeurs et les individus. Si deux ou plusieurs attributs sont liés par une relation logique ou une corrélation statistique et sont permutés indépendamment l'un de l'autre, ce lien sera détruit. Il peut donc être important de permuter un ensemble d'attributs de façon à ne pas briser la relation logique, faute de quoi un attaquant pourrait identifier les attributs permutés et inverser la permutation.

[Zhang et al, 2007] ont proposé une approche appelée k-permutation. Leur idée principale est de dissocier la relation entre le quasi-identifiant et l'attribut sensible numérique par le

partitionnement d'un ensemble d'enregistrements de données en groupes et traînant leurs valeurs sensibles à l'intérieur de chaque groupe.

5. La différence entre le cryptage et l'anonymisation

Bien que l'anonymisation et le cryptage sont des sujets connexes et ce sont des techniques utiles pour la sécurisation des données (confidentialité) en Cloud.

L'anonymisation des données est le processus de transformation des données afin qu'il puisse être traité d'une manière utile, tout en évitant que les données ne soient liées à des identités individuelles des personnes, des objets ou de l'organisation

Le cryptage consiste à transformer les données afin de le rendre illisible pour ceux qui n'ont pas la clé pour décrypter.

Le cryptage peut être un outil utile pour faire l'anonymisation, et plus particulièrement lorsque la dissimulation des informations d'identification dans un ensemble de données **[**Jeff, 2012**].**

6. Conclusion

Dans ce chapitre, nous avons détaillé les caractéristiques de la préservation de la confidentialité pour les données publiées. Nous avons commencé par expliquer des principales phases et les différentes hypothèses qui peuvent être traitées dans la publication de données. Ensuite, nous avons présenté l'approche d'anonymisation qui est considérée comme l'une des techniques de la confidentialité. Enfin, nous avons énuméré les principales opérations de l'anonymisation.

Dans le prochain chapitre, nous allons nous intéresser aux différents travaux liés à la préservation de la confidentialité pour les bases des données centralisées et décentralisées.

Chapitre 5

La préservation de la confidentialité des données : État de l'art

1. Introduction

La majorité des travaux d'anonymisation touchent deux axes principaux **[**Iwuchukwu et Naughton, 2007**]**:

1. *Le critère de modèle des attaques* : Comment préserver la confidentialité (anonymisation) des données et affiner la définition de l'anonymisation pour fournir différentes garanties sur la sécurisation des données (par exemple, renforcer la sécurité à travers le modèle *l-diversité*) ?
2. *Le critère d'utilité de données* : Comment préserver l'utilité de données et minimiser la perte en ce qui concerne la qualité d'information après la publication, tout en respectant la définition de l'anonymisation (par exemple le modèle *k-concealment*) ?

Dans notre travail de thèse, nous allons considérer les deux axes à la fois pour atteindre les objectifs de la confidentialité tout en minimisant la généralisation des données (perte de la qualité des données). Dans ce chapitre, nous allons citer tous les travaux qui touchent l'aspect d'anonymisation par rapport à ces deux axes.

2. La préservation de la confidentialité pour les bases de données centralisées

D'après [Dalenius, 1977], toute information concernant un individu qui peut être extraite à partir de données publiées pourrait être apprise sans avoir l'accès à ces données. La préservation de la confidentialité des données publiées pour les bases des données centralisées a été largement étudiée [Fung et al, 2010]. Plusieurs techniques telles que ***k-anonymat, l-diversité*** et ***k-concealment*** ont été suggérées pour trouver le bon équilibre entre la publication des données et la divulgation de données.

2.1. Anonymisation statique

- K-anonymat

Pour remédier aux problèmes d'attaque *Liaison entre enregistrements* présenté dans le chapitre précédant (*la figure 4.3),* [Samarati, 2001**]** et [Sweeney-B-, 2002] ont défini le modèle de ***k-anonymat*** comme suit :

Un partage de données est dit adhéré à k-anonymat si chaque enregistrement publié comporte au moins (k) autres enregistrements dont les valeurs sont indistinctes sur un ensemble spécial de domaines appelés quasi-identifiant.

Supposant que la base de données contient les deux attributs : âge et code postal. L'ensemble de données est dit k-anonymes si, pour tout un enregistrement particulier, il existe *k* autres enregistrements qui ont le même âge et le même code postal.

Ce modèle ne répond pas à certains cas qui peuvent apparaître dans l'approche d'anonymisation. Considérant l'exemple suivant *(Liaison entre attributs)* où l'attaquant peut déduire les valeurs de l'attribut sensible à partir de la table originale.

Ainsi, selon les données publiées dans *la figure 5.1*, si le fournisseur de service sait qu'un individu a 27 ans et un code postal 08009, le fournisseur peut conclure qu'elle appartient à la 1ère section de la table anonyme et qu'il a une maladie cardiaque.

	Code postale	Age	Maladie
1	08001	29	Maladie cardiaque
2	08011	22	Maladie cardiaque
3	08009	27	Maladie cardiaque
4	22001	43	Grippe
5	22009	52	Maladie cardiaque
6	22011	47	Cancer
7	31000	30	Maladie cardiaque
8	31010	36	Cancer
9	31018	32	Cancer

Table originale des patients

	Code postale	Age	Maladie
1	080**	2*	Maladie cardiaque
2	080**	2*	Maladie cardiaque
3	080**	2*	Maladie cardiaque
4	220**	≥ 40	Grippe
5	220**	≥ 40	Maladie cardiaque
6	220**	≥ 40	Cancer
7	310**	3*	Maladie cardiaque
8	310**	3*	Cancer
9	310**	3*	Cancer

La version 3-anonymat

Figure 5.1 Le modèle k-anonymat

- L-diversité

Plusieurs travaux ont été proposés pour améliorer le modèle *k-anonymat* et combler les failles de confidentialité laissées par ce modèle. [Machanavajjhala et al, 2006] ont introduit la notion de *l-Diversité* où ils ont montré qu'une donnée k-anonyme permet de fortes attaques de ré-identification en raison du manque de diversité dans les attributs sensibles. Par manque de diversité, les auteurs veulent exprimer le nombre (*l*) de valeurs différentes des informations sensibles dans chaque classe d'équivalence. Ils ont proposé plusieurs types de *l*-Diversité s'intégrant facilement aux travaux de k-anonymat pour assurer un seuil *l* de diversité de valeurs dans les attributs sensibles.

Contrairement au modèle *k-anonymat*, le modèle *l-diversité* protège contre l'attaque *Liaison entre attributs*, à travers une certaine diversité au niveau de l'attribut sensible.

Attaque de similitude (Similarity attack) : Lorsque les valeurs d'attributs sensibles dans une classe d'équivalence sont distinctes, mais sémantiquement similaire, un adversaire peut apprendre des informations importantes.

Considérant l'exemple suivant : Étant donné la table (a) comme table originale et la table (b) présente la version anonyme qui satisfait à *3-diversité*. Il y a deux attributs sensibles : *Salaire* et *Maladie.* Supposant que le fournisseur sait que l'enregistrement d'un tel individu correspond à la première classe d'équivalence et qu'il a un salaire dans l'intervalle [3K–5K]. Ainsi, il peut déduire que le salaire de cet individu est relativement bât. Cette attaque ne s'applique pas seulement sur les attributs numériques, mais aussi sur les attributs catégoriques comme « Maladie ». Sachant que cet enregistrement appartient à la première classe, il conclut que cet individu souffre d'un problème d'estomac par ce que les trois maladies ont une relation avec l'estomac.

	Code postale	**Age**	**Salaire**	**Maladie**
1	08001	29	3 K	Ulcère gastrique
2	08011	22	4 K	Gastrite
3	08009	27	5 K	Cancer de l'estomac
4	22001	43	6 K	Gastrite
5	22009	52	11 K	Grippe
6	22011	47	8 K	Bronchite
7	31000	30	7 K	Bronchite
8	31010	36	9 K	Pneumonie
9	31018	32	10 K	Cancer de l'estomac

(a). Table originale des patients

	Code postale	Age	Salaire	Maladie
1	080**	2*	3 K	Ulcère gastrique
2	080**	2*	4 K	Gastrite
3	080**	2*	5 K	Cancer de l'estomac
4	220**	≥ 40	6 K	Gastrite
5	220**	≥ 40	11 K	Grippe
6	220**	≥ 40	8 K	Bronchite
7	310**	3*	7 K	Bronchite
8	310**	3*	9 K	Pneumonie
9	310**	3*	10 K	Cancer de l'estomac

(b). La version 3-diversité

Figure 5.2 Le modèle l-diversité

Il faut noter que les modèles *k-anonymat* et *l-diversité* sont les modèles les plus utilisés dans le premier axe (décrit précédemment). Par la suite, [Liu et al, 2015] ont proposé un nouvel schéma pour fournir les services personnalisés et [Stammler et al, 2016] ont amélioré le besoin de garder la diversité dans l'anonymisation.

- T-proximité

[Ninghuiet al, 2007] ont proposé un nouveau modèle de confidentialité appelé *t-proximité* qui exige que la distribution de l'attribut sensible dans une classe équivalence est proche à la distribution des attributs dans la table (la distance entre les deux distributions ne doit pas dépasser un seuil ***t***).

EMD (Earth Mover's Distance) est basé sur la quantité minimale de travail nécessaire pour transformer une distribution à l'autre en déplaçant la masse de distribution entre l'une à l'autre. Une distribution est considérée comme une masse de terre propagée dans l'espace et l'autre comme une collection de trous dans le même espace. EMD mesure la moindre quantité de travail nécessaire pour remplir les trous avec la terre.

t-proximité utilise EMD pour calculer la distance entre les deux distributions qui est considérée comme une proximité sémantique entre les valeurs d'attributs. Ce modèle souffre de deux principaux inconvénients : **[**Mahajan et Ganar, 2012**]**

- Protège contre l'attaque : *Liaison entre attributs* mais ne protège pas contre l'attaque *Liaison entre enregistrements*.
- Les attributs sensibles multiples présentent un défi supplémentaire

- Le nouveau modèle d'anonymisation *k-concealment.*

Dans cette étude, nous supposons que l'adversaire connaît les données de tous les individus de la population, et qu'il sait le sous-ensemble exact de la population qui est représentée dans le tableau. Le modèle *k-anonymat* assure que chaque enregistrement de la table origine peut être lié au moins à k enregistrements de la table anonymes. Pour cette raison, il faut que la table anonyme soit composée de groupes ayant une taille au moins de k, où tous les enregistrements dans le même groupe ont les mêmes quasi-identifiants généralisés, Ainsi, il devient impossible de distinguer entre les enregistrements d'un groupe donné dans le sens de l'information-théorique, d'où, l'adversaire ne pourra jamais détecter l'enregistrement ciblé.

Le modèle k-anonymat conduit à une généralisation excessive des quasi-identifiants, ce qui conduit par la suite à une perte en qualité d'information. Le modèle *k-concealment* [Tassa et al, 2012] a été proposé comme un modèle alternative qui étend le modèle k-anonymat dans le sens où chaque table k-anonyme satisfait les conditions de k-concealment mais l'inverse n'est pas nécessairement réalisé. L'avantage offert par ce nouveau modèle est l'augmentation d'utilité de la donnée : *k-concealment peut être réalisé avec moins de généralisation par rapport à ce qui est requis par k-anonymat.*

Exemple :

Étant donnée une table origine (a) (*figure 5.3*) ayant les quasi-identifiants : *Age* et *Code postal* et l'attribut sensible *Maladie*. La table (b) qui correspond au modèle 2-anonymat, est constituée de deux groupes d'enregistrements qui ont les mêmes quasi-identifiants généralisées (les trois premiers enregistrements et les deux derniers). Il est impossible de distinguer entre les enregistrements du même groupe, parce qu'ils sont identiques. La table (c) est 2-concealment de la table (a). Un adversaire qui connaît les quasi-identifiants de tous les enregistrements dans la table (a) ne peut pas lier un tel enregistrement de moins de deux enregistrements généralisés dans la table (c). Par exemple, en se basant sur les deux quasi-identifiants âge et Code postal, l'adversaire ne peut pas déterminer si l'enregistrement de Fatima dans la table (c) est le premier ou le troisième. Ces deux enregistrements ne sont pas identiques.

Nom	*Age*	*Code postale*	*Maladie*
Fatima	30	08001	rougeole
Saliha	21	08011	Grippe
Amine	21	08002	Angine
Mohamed	55	22001	Grippe
Salim	47	22005	Diabète

(a). Table originale des patients

Age	*Code postale*	*Maladie*
21-30	08***	Rougeole
21-30	08***	Grippe
21-30	08***	Angine
47-55	22***	Grippe
47-55	22***	Diabète

(b). La version 2-anonymat

Age	*Code postale*	*Maladie*
21-30	**08001**	Rougeole
21	08***	Grippe
21-30	08***	Angine
47-55	22***	Grippe
47-55	22***	Diabète

(c). La version 2-concealment

Figure 5.3 Le modèle k-concealment [Tassa et al, 2012].

Dans ce qui suit, nous allons présenter certaines solutions qui touchent le deuxième axe où la conservation de l'information est faite de telle sorte que les données publiées restent pratiquement utiles.

[Meyerson et Williams, 2004**]** ont étudié le problème de k-anonymat en utilisant la technique de suppression tout en se basant sur la minimisation de l'information perdue (information loss). [Aggarwal et al-b-, 2005**]** ont élargi le contexte de suppression en permettant des règles plus générales pour généraliser les entrées de données. [Gionis et Tassa, 2009**]** ont proposé trois fonctions pour mesurer la perte de l'information. Ces mesures sont plus générales que celles suggérées dans les deux travaux cités ci-dessus. [Jacob et Tassa, 2009**]** ont proposé une nouvelle mesure qui vise à maximiser la corrélation entre les données publiques généralisées et les données privées. Ils ont montré que cette mesure est beaucoup plus appropriée aux objectifs de l'exploration de données en visant à trouver des règles d'association pour la prédiction des données privées à partir des données publiques. [Sunyong et al, 2012**]** présentent une nouvelle méthode dans le cadre de l'information mutuelle. Cette méthode non seulement minimise la perte d'information, mais réalise également la diversité de l'attribut sensible, ce qui facilite l'utilisation des données et empêche les divers attaques.

[*Tinbo, 2013*] a suggéré un nouvel algorithme d'anonymisation nommé *kl-redInfo* qui assure l'anonymisation en utilisant les deux modèles k-anonymat et l-diversité. La perte d'information est réduite par une utilisation des nouvelles approches proposées. Le travail de [Kohlmayer et al, 2015] est considéré comme la première étude systématique qui décrit comment mettre en œuvre la généralisation et la suppression lorsque les données biomédicales doivent être anonymes avec des modèles de confidentialité syntaxiques communes et des mesures d'utilité. La mise en œuvre de ce modèle de transformation conduit à une forte augmentation de l'utilité par rapport aux données anonymes qu'avec la généralisation.

2.2. Anonymisation dynamique

Les données réelles sont dynamiques (surtout dans l'environnement Cloud). Les ensembles de données dynamiques sont complexes à travers l'utilisation des mises à jour de données (la suppression, l'ajout et la modification). La plupart des travaux dans le PCDP s'articulent sur l'anonymisation statique où ils assurent la protection à un certain niveau.

- M-invariance

[Xiaoet Tao, 2007] ont identifié les attaques qui peuvent exister dans une publication des données *(association des valeurs)* et ils ont proposé une solution pour empêcher ces attaques. Cependant, cette solution supporte que l'insertion des données (elle ne supporte pas la suppression et le chargement des données).

Pour comprendre ce concept, l'exemple suivant montre les données publiées par un hôpital (*figure 5.4*). Considérant qu'un adversaire sait que *Samir* a un enregistrement dans les deux publications (a) et (b), basant sur la dernière publication, l'adversaire peut conclure que *Samir* ayant, soit « *Grippe* ou *Gastrite* ». Alors par la combinaison de ces deux publications avec leurs versions anonymes, l'adversaire conclut, que *Samir* ayant une maladie (attribut sensible) *Grippe* et non *Gastrite (figure 5.4)* [Pei et al, 2007].

Pour éviter cette attaque, [Xiao et Tao, 2007] proposent le modèle M-invariance. Leur idée principale est l'utilisation de la généralisation des contrefais. Dans chaque instant, un enregistrement particulier peut être placé dans les partitions avec un ensemble fixe de valeurs sensibles, nommées *Signature*. Comme la publication *(c)* nous montre, *Amine* est classé dans une classe ayant comme attribut sensible, les maladies « *Infection virale, Gastrite* » ; ceci bloque l'attaque d'insertion. (C.à.d. si l'adversaire compare la publication *(a)* et *(e)*, il ne peut

pas déduire la maladie de *Amine*, par ce que les classes comprennent les mêmes ensembles de valeurs d'attributs sensibles.

Nom	***Sexe***	***Age***	***Code postale***	***Maladie***
Amine	Male	35	08001	Infection virale
Salim	Male	37	08002	Problème cardiaque
Mohamed	Male	39	08003	Grippe
Fatima	Femelle	54	08010	Grippe
Saliha	Femelle	58	08011	Problème cardiaque
Hamid	Male	54	08014	Infection virale
Samir	Male	41	08007	Grippe
Madjid	Male	46	08008	Grippe
Samira	Femelle	44	08009	Grippe

(a). 1er publication des données

Sexe	***Age***	***Code postale***	***Maladie***
*	< 40	0800*	Infection virale
*	< 40	0800*	Problème cardiaque
*	< 40	0800*	Grippe
*	5*	0801*	Grippe
*	5*	0801*	Problème cardiaque
*	5*	0801*	Infection virale
*	> 40	0800*	Grippe
*	> 40	0800*	Grippe
*	> 40	0800*	Grippe

(b). La version 3-anonymat

Nom	***Sexe***	***Age***	***Code postale***	***Maladie***
Amine	Male	35	08001	Infection virale
Sakina	Femelle	36	08005	Gastrite
Mohamed	Male	39	08003	Grippe
Fatima	Femelle	54	08010	Grippe
Saliha	Femelle	58	08011	Problème cardiaque
Asma	Femelle	53	08019	Gastrite
Samir	Male	41	08007	Grippe
Madjid	Male	46	08008	Grippe
Imad	Male	42	08021	Gastrite

(c). 2ème publication des données

Sexe	***Age***	***Code postale***	***Maladie***
*	3*	08001	Infection virale
*	3*	08005	Gastrite
*	3*	08003	Grippe
*	> 50	08010	Grippe
*	> 50	08011	Problème cardiaque
*	> 50	08019	Gastrite
*	> 40	080017	Grippe
*	> 40	08008	Grippe
*	> 40	08021	Gastrite

(d). La version 3-anonymat

Nom	***Sexe***	***Age***	***Code postale***	***Maladie***
Amine	Male	35	08001	Infection virale
F1	Male	37	08006	Problème cardiaque
Sakina	Femelle	36	08005	Gastrite
Mohamed	Male	39	08003	Grippe
Fatima	Femelle	54	08010	Grippe
Saliha	Femelle	58	08011	Problème cardiaque
F2	Male	54	08018	Infection virale
Asma	Femelle	53	08019	Gastrite
Samir	Male	41	08007	Grippe
Madjid	Male	46	08008	Grippe
F3	Femelle	44	08020	Grippe
Imad	Male	42	08021	Gastrite

(e). Republication des données avec des records contrefaits

Figure 5.4 Exemple de l'attaque Association des valeurs [Mahajan et Ganar, 2012]

Nom	*Salaire*	*Age*	*Maladie*
Karim	14K	20	Dyspepsie
Jamal	16K	23	Pneumonie
Tamer	24K	32	Pneumonie
Houcine	26K	35	Gastrite
Leila	29K	17	Glaucome
Bachir	31K	19	Grippe

(a). 1er publication des données

Nom	*Salaire*	*Age*	*Maladie*
Karim	14K	20	Dyspepsie
Jamal	18K	31	Cancer du poumon
Tamer	15K	27	Pneumonie
Houcine	23K	32	Dyspepsie
Leila	12K	17	Glaucome
Bachir	26K	35	Pneumonie

(b). 2èm publication des données

Nom	*Salaire*	*Age*	*Maladie*
Karim Tamer	[14K, 15K] [14K, 15K]	[19, 27] [19, 27]	Dyspepsie Pneumonie
Djamel Houcine	[18K, 23K] [18K, 23K]	[31, 32] [31, 32]	Cancer du poumon Dyspepsie
Leila C1	[10K, 12K] [10K, 12K]	[16, 17] [16, 17]	Glaucome Pneumonie
Bachir C2	[26K, 27k] [26K, 27k]	[35, 37] [35, 37]	Pneumonie Cataracte

(c). Republication des données avec des records contrefaits

Figure 5.5 La limite de M-invariance [Feing et Zhou, 2008]

La principale limite de M-invariance est qu'elle supporte seulement la mise à jour externe et ne supporte pas la mise à jour interne. Supposant, dans la 1ère publication (a) de *la figure 5.5*, *Djamal* est dans une classe dont l'ensemble des valeurs sensibles est {*la dyspepsie, la pneumonie*}. Ensuite et dans la publication *(b)*, la maladie est dégradée vers « *Cancer du poumon* ». Dans la publication (c), nous remarquons que l'ensemble des valeurs sensibles a changé, ce qui nous montre que M-Invariance ne satisfait pas le chargement interne.

- **M-Distinct**

Suite aux limites de M-invariance dans les mises à jours internes, [Feng et Zhou, 2008] ont proposé un nouveau principe de généralisation contrefait, appelé M-Distinct pour anonymiser de manière efficace les données avec des mises à jours externes et internes. Ils ont développé un algorithme de généralisation des ensembles de données pour arriver à satisfaire le modèle M-Distinct. Leur idée est de maintenir une indistinction persistante des valeurs sensibles de

chaque classe, même dans les cas des mises à jour internes, ou bien dans le cas où l'adversaire exploite la corrélation entre les différentes publications.

Dans chaque publication, ils ont partitionné chaque enregistrement d'individu en classes d'équivalences ce qui ne va pas conduire à une exclusion de ses valeurs sensibles possibles.

Quelques solutions y compris [He et al, 2011] et [Hoang et al, 2014] présentent l'extension de *m-invariance*, mais aucun d'entre elles ne se base sur le traitement des mises à jour arbitraires présenté par [Adeel et Guillaume, 2013**]**.

3. La préservation de la confidentialité des données publiées pour les bases des données décentralisées

Avant de présenter les architectures distribuées qui existent dans l'anonymisation, il faut noter que les modèles de l'anonymisation centralisée sont déjà utilisés dans les systèmes distribués. Il existe des approches qui peuvent être considérées comme des solutions pour l'anonymisation des données dans les bases des données distribuées.

3.1. Intégration/Anonymisation

C'est une simple approche qui assume une existence de tiers à qui fait confiance chacun des propriétaires de données. Les propriétaires de données envoient leurs données à un tiers où l'intégration et l'anonymisation des données sont effectuées. Les clients peuvent par la suite interroger la base de données centralisée.

Cette approche n'est pas réalisable pour de nombreux scénarios **[**Kohlmayer et al, 2013**]**. Trouver un tiers confiant n'est pas toujours possible et une fois le serveur est soumis à une violation par des hackers, il pourra conduire à une perte de confidentialité pour toutes les parties participantes. [Mohammed et Fung, 2010] ont utilisé cette solution en proposant un algorithme pour atteindre la confidentialité des données à la fois pour les scénarios centralisés et décentralisés. Cette solution a été réalisée pour préserver les informations des patients entre le Service de transfusion de la Croix-Rouge de Hong Kong et l'hôpital publique.

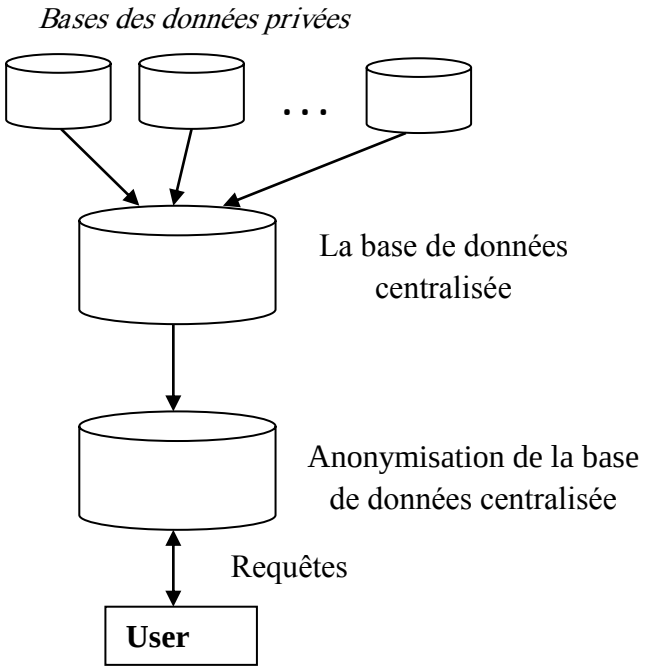


Figure 5.6 La solution : intégration / anonymisation [Jurczyk et Xiong, 2009]

3.2. Anonymisation /Intégration

L'idée principale de cette approche est que chaque partie réalise indépendamment une anonymisation locale des données qui seront par la suite intégrées dans un ensemble global.

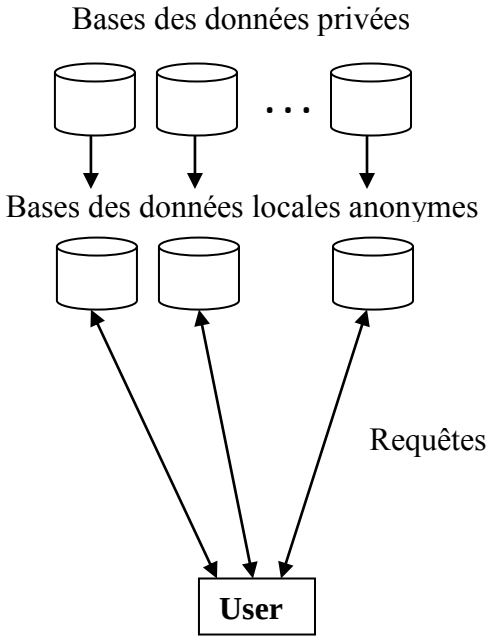


Figure 5.7 La solution : Anonymisation / Intégration [Jurczyk et Xiong, 2009]

Dans le cas de la distribution horizontale, l'approche de [Mohammed et Fung, 2010] a un impact négatif sur la qualité des données. Par exemple, l'application de k-anonymat local peut faire souffrir l'utilité de données plus que l'application de k-anonymat globale. Si les données sont distribuées verticalement [Jiang et Clifton, 2006], l'intégration des données localement anonymes peut provenir des données non anonymes au niveau global (Si, par exemple, des nouvelles colonnes sont ajoutées à un ensemble de données k-anonyme, le résultat dans la plupart des cas sera non k-anonyme).

3.3. La solution virtuelle

Dans *l'anonymisation virtuelle* [Jurczyk et Xiong, 2009] [Ding et al, 2013], les fournisseurs de données participent dans les protocoles distribués pour produire une base de données intégrée, virtuelle et anonyme. Il est important de noter que les données anonymes sont toujours restées dans les bases de données individuelles, et l'intégration et l'anonymisation des données sont réalisées par des protocoles distribués et sécurisés.

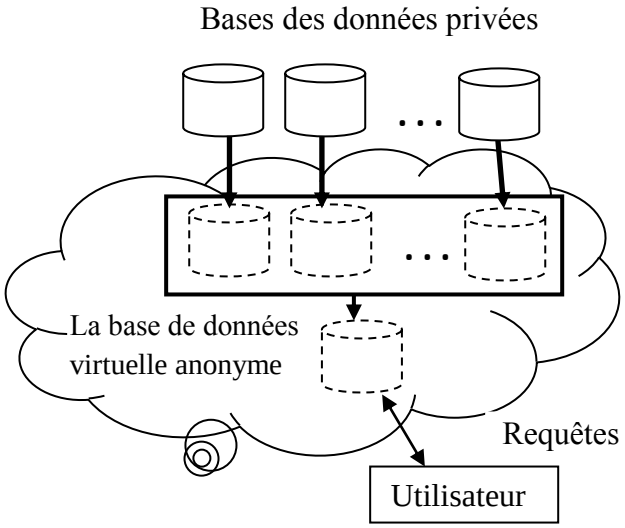


Figure 5.8 La solution : Anonymisation virtuelle [Jurczyk et Xiong, 2009]

Il existe différentes méthodes pour l'anonymisation de données dans un environnement distribué. Dans ce qui suit, nous allons mettre l'accent sur les approches qui utilisent l'approche d'anonymisation virtuelle.

Les auteurs dans [Jurczyk et Xiong, 2009] proposent une adaptation de l'algorithme Mondrian qui est basé sur la généralisation pour produire une version k-anonyme de l'union des ensembles de données en utilisant un ensemble de protocoles de calcul multipartis sécurisé [Goldreich, 2001] (Secure Computing Sum, Secure Median Protocols). La séquence

d'exécution est sous le contrôle d'un seul site *leader* qui suppose que les participants sont entièrement disponibles.

[Ding et al, 2013] ont présenté un algorithme d'anonymisation distribué en utilisant la structure d'index de l'arbre R [Guttman, 1984]. Leur algorithme utilise une approche pour insérer de manière récursive les objets de données dans l'espace de domaine de quasi-identifiant, et quand un débordement se produit, cet espace sera divisé en deux parties. Il choisit récursivement la branche pour insérer les objets de données les plus proches. La construction de la table de généralisation se faite lorsque tous les tûples de données sont insérés dans le arbre-R.

Récemment, certains travaux publiés pour l'anonymisation distribuée des données traitent seulement le critère de modèle des attaques, alors que la quantité d'information perdue n'est pas prise en compte. [Kohlmayer et al, 2013] ont construit une vue cryptée de l'ensemble de données. Ils ont implémenté les modèles *k-anonymat*, *l-diversité* et *t-priximité* avec une méthode d'identification globalement optimale dans les bases des données distribuées horizontalement et verticalement. [Lue et al, 2013] ont proposé un système d'anonymisation distribué pour les systèmes de recommandation préservant la confidentialité. Le système proposé permet aux utilisateurs d'anonymiser individuellement leurs propres données sans avoir l'accès les uns aux autres. [Grinshpoun et Tassa, 2014] ont introduit un nouvel algorithme pour préserver la confidentialité des agents participant au cours du processus de résolution. [Raymond et al, 2015] ont étudié comment l'anonymisation d'un profil clinique se fait dans plusieurs centres médicaux en utilisant le modèle k-anonymat.

Dans notre travail, nous allons nous baser sur la solution de l'anonymisation virtuelle.

4. Synthèse

Afin de concrétiser cette expérience et d'arriver au but de nos objectifs, cette synthèse présente l'aboutissement de ce travail. Ainsi la comparaison entre les différents modèles proposés permet d'éclairer l'objectif final de notre travail et mettre en valeur notre état de l'art.

Concernant la préservation de la confidentialité pour les bases de données centralisées, une étude analytique a été présentée et résumée dans *la table 5.1*.

Approches	***Attaques***	***Type d'anonymisation***	***Algorithme utilisé***	***Critères de recherche***	***Utilité de données***
Samarati et al, 2002	Liaison entre enregistrements	Statique	MinGen	Modèle d'attaque	Inférieure
Machanavajjhala et al, 2006	Liaison entre enregistrements & entre attributs	Statique	l-diversité	Modèle d'attaque	Inférieure
Li et al, 2007	Liaison entre attributs	Statique	t-proximité	Modèle d'attaque	Inférieure
Xiao et al, 2007	Association des valeurs (mise à jour externe)	Dynamique	m-invariance	Modèle d'attaque	Inférieure
Feng et al, 2008	Association des valeurs (mise à jour externe & interne)	Dynamique	m-distinct	Modèle d'attaque	Inférieure
Tassa et al, 2012	Liaison entre enregistrements & entre attributs	Statique	k-concealment	Modèle d'attaque & utilité de données	Supérieure

Table 5.1 Tableau récapitulatif pour la préservation de la confidentialité pour les bases de données centralisées

Alors que k-anonymat offre une protection contre *Liaison entre enregistrements*, elle est insuffisante pour empêcher l'attaque *Liaison entre attributs* due à certaines limitations qui sont déjà discutés. Contrairement au modèle l-diversité qui est capable de protéger contre l'attaque *Liaison entre attributs.*

D'autres parts, nous avons remarqué que tous les modèles présentés au-dessus ne prennent pas en considération le critère d'utilité de données où la conservation de l'information est faite de telle sorte que les données publiées restent pratiquement utiles. C'est pour cette raison, un nouveau modèle de confidentialité nommé ***K-concealment*** sera présenté dans le prochain chapitre comme un modèle alternative qui assure le même niveau de sécurité et offre une plus grande utilité par rapport aux autres modèles utilisés dans les travaux précédents. Ces techniques (modèles) peuvent être utilisées comme une meilleure approche pour sécuriser les données dans un environnement distribué.

Concernant la préservation de la confidentialité pour les bases des données décentralisées, les travaux ci-dessus (*Table.5.2*) ont utilisé les modèles *k-anonymat, l-diversité* et *t-proximité* pour la confidentialité des objets des données.

Approches	Modèles utilisés	Opération utilisée	Types d'anonymisation	Critères	Avantages	Inconvénients
Jiang et Clifton, 2006	k-anonymat	Généralisation (*descendante*)	Statique	Modèle d'attaque	Distribution verticale	Critère d'utilité
Jurczyk et Xiong, 2009	k-anonymat& l-diversité	Généralisation (*descendante*)	Statique	Modèle d'attaque & utilité de données	l'algorithme Mondrian au niveau de fournisseur de données	Mauvaise efficacité sur la qualité des données.
Mohammed et fung, 2010	k-anonymat& l-diversité	Spécialisation (*descendante*)	Statique	Modèle d'attaque & utilité de données	Distribution horizontale qui préserve les informations des patients par l'intermédiaire d'un tiers.	Inexistence du tiers confiant
Ding et al, 2013	k-anonymat& l-diversité	Généralisation (*ascendante*)	Statique	Modèle d'attaque & utilité de données	Algorithme de l'arbre *R* au niveau de fournisseur de données	la mauvaise insertion et le non adéquation de Split.
Kohlmayer et al, 2013	k-anonymat& l-diversité & t-proximité	Généralisation	Statique	Modèle d'attaque	Anonymisation flexible pour les données distribuées verticalement et horizontalement	Critère d'utilité de données
Raymond et al, 2016	k-anonymat	Généralisation	Statique	Modèle d'attaque	l'anonymisation d'un profil clinique dans plusieurs centres médicaux	Critère d'utilité de données
Notre approche	k-concealment	Généralisation (*ascendante*)	Statique	Modèle d'attaque & utilité de données	*Au niveau d'objet de données* : Utilisation du modèle k-concealment qui assure une grande utilité par rapport aux modèle utilisés précédemment. *Au niveau de fournisseur de données* : Utilisation d'un nouvel algorithme qui utilise la structure de l'arbre R* qui assure une insertion avec un chevauchement minimale	Anonymisation dynamique

Table 5.2 Tableau récapitulatif pour la préservation de la confidentialité pour les bases de données décentralisées

Notre travail vise à externaliser les données privées des fournisseurs des données vers des serveurs de Cloud pour la publication des données. En outre, le protocole d'anonymisation que nous présenterons dans *le chapitre 7*, vise à atteindre l'anonymisation à la fois des :

- *Objet des données* : en utilisant le modèle *k-concealment* qui assure le même niveau de sécurité et offre une plus grande utilité de données que les modèles proposés dans les travaux connexes (*Table 5.2*).
- *Fournisseurs de données* : par la conception d'un nouvel algorithme d'anonymisation distribué qui utilise la structure de l'arbre- R* **[**Beckmann et al, 1990**]** et qui donne une meilleur insertion des objets des données dans l'espace de domaine de quasi-identifiant en trouvant le partitionnement adéquat de cet espace.

5. Conclusion

Dans ce chapitre, nous avons montré les deux principaux axes liés à la préservation de la confidentialité. Nous avons précisé que nous allons nous baser sur les objectifs du premier axe qui touche l'aspect de sécurité tout en minimisant la généralisation des données qui est l'objectif principale du deuxième axe. Par la suite, nous avons détaillé les différents travaux connexes à la préservation de la confidentialité pour les bases de données centralisées et nous avons conclu que le modèle *k-concealment* que nous allons utiliser dans notre approche prend en considération les deux axes suscités.

Concernant la préservation de la confidentialité pour les bases de données décentralisées, nous avons présenté les différentes approches qui peuvent être considérées comme des solutions pour l'anonymisation des données dans les bases des données distribuées, tout en marquant les points fort et les points faibles pour chacune. Enfin, nous avons conclu par un tableau comparatif qui nous a permis de mieux positionner notre approche par rapport à la littérature.

Le prochain chapitre, a pour objectif de présenter le modèle k-concealment que nous allons utiliser dans notre approche d'anonymisation distribuée.

Chapitre 6

Le modèle k-Concealment

1. Introduction

Ce chapitre vise à mettre l'accent sur le modèle k-concealment que nous allons utiliser dans notre approche d'anonymisation distribuée. Durant ce chapitre, nous allons essayer de :

- Donner quelques définitions concernant le concept de cohérence entre les enregistrements ainsi que présenter les notions de k-types anonymisation sur lesquels reposent ce concept.
- Discuter de l'insécurité des trois notions de base d'anonymisations, en montrant qu'ils n'offrent pas le même niveau de sécurité que le modèle k-anonymat. Contrairement à la version k-concealment d'anonymisation qui assure une haute sécurité avec une plus grande utilité.
- Présenter les différents algorithmes des types d'anonymisations ainsi que le modèle k-concealment.

Préliminaires

Considérant une base de données qui contient des informations sur les individus dans une population $U = \{u_1, ..., u_n\}$. Chaque individu est décrit par un ensemble d'attributs $\boldsymbol{r}$, $A_1, ...,$ Ar et l'attribut sensible A_{r+1}. Nous notons D comme une projection de la base de données sur l'ensemble d'attributs $\boldsymbol{r}$ et les enregistrements de D sont notés comme R_i, $1 \leq i \leq n$.

Définition.1

Étant donné, l'ensemble finit A_j, $1 \leq j \leq r$, et son collection $\bar{A}$. L'enregistrement $\acute{R} \in \bar{A}_1 .... \bar{A}_r$ est la généralisation de l'enregistrement $R \in A_1 .....A_r$ (Autrement dit ; $\acute{R}$ est cohérent avec R), si $R\ (j \in \acute{R}\ (j)$ pour tous $1 \leq j \leq r$.

Si $D = \{R_1, . . . , R_n\}$ est la table d'enregistrements dans $A_1 .....A_r$. Alors $g(D) = \{\acute{R}_1, \ldots, \acute{R}_n\}$ est la généralisation de D pour tous $1 \leq i \leq r$ [Tassa et al, 2012].

Définition.2

Étant donné $S \subseteq A_1 \times .... \times A_r$ un ensemble d'enregistrements. Le coût minimal d'une généralisation est indiqué comme le minimum enregistrement généralisé dans $\bar{A}_1 \times ... \times \bar{A}_r$ qui est cohérent avec tous les autres enregistrements dans S. Ce cout de généralisation est définit comme $d(S) = c(S')$. (C est la métrique qui calcule l'information perdu d'un enregistrement généralisé). [Tassa et al, 2012]

2. Les k-types d'anonymisations

Nous allons maintenant présenter les notions de k-types anonymisations qui reposent sur le concept de cohérence définit précédemment dans *la définition1*.

Définition.3

Étant donné une table $D = \{R_1, \ldots, R_n\}$; $g(D) = \{\acute{R}_1, \ldots, \acute{R}_n\}$ représente la généralisation correspondante. Alors :

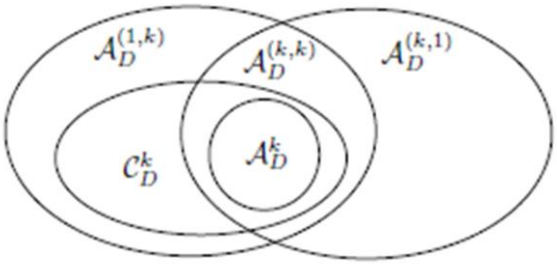
- *g (D)* est appelé (1, k) -anonymisation de *D* si chaque enregistrement de *D* est cohérent au moins avec k enregistrements de *g (D).*
- *g (D)* est appelé (k, 1) -anonymisation de *D* si chaque enregistrement de *g (D)* est cohérent au moins avec k enregistrements de *D.*
- *g (D)* est appelé (k, k) -anonymisation de *D* si elle est à la fois un (1, k) - et un (k, 1) -anonymisation de *D.*

D'une façon correspondante, nous définissons respectivement A_D^k, $A_D^{(1.K)}$, $A_D^{(k,1)}$, $A_D^{(k,k)}$ comme étant des collections de k, (1, k), (k, 1) et (k, k) –anonymisations de *D*.

Afin de comprendre la motivation derrière ces définitions, il faut que nous mettions l'accent sur la perspective de l'adversaire. Une attaque typique vise à révéler les informations sensibles sur une cible individuelle spécifique. Dans une telle attaque, l'adversaire sait l'enregistrement $R \in D$ et il essaie de trouver l'enregistrement généralisé correspondant $\acute{R} \in g(D)$. Alternativement, l'adversaire pourrait être intéressé par la re-identification de toute entité dans les données publiées où l'attaque fonctionne dans le sens opposé. En se basant sur l'enregistrement généralisé $\acute{R} \in g(D)$, l'adversaire tente de déduire sa pré-image correcte de $R \in D$. La notion de (1, k) - anonymisation vise à protéger contre la première attaque. La notion de (k, 1) - anonymisation vise à protéger contre la seconde. Enfin, (k, k) – anonymisation prend en considération les deux attaques.

La notion de (k, 1) - anonymisation a été déjà définie dans [Vinterbo, 2004] sous le nom de k-ambiguïté. Une notion de sécurité similaire est apparue dans [Sweeney, 2002]: une table anonyme adhère au modèle k-map si chaque enregistrement est cohérent avec au moins k entités de la population sous-jacente (et pas seulement dans la table d'origine, comme dans k-ambiguïté et son équivalent (k, 1) - l'anonymat).

Pour une table donnée *D*, la relation entre les collections A_D^k, $A_D^{(1.K)}$, $A_D^{(k,1)}$, $A_D^{(k,k)}$ qui sont définit dans *la définition 3*, est représentée dans *la figure.6.1*



$$A_D^k \subset A_D^{(k,k)} \subset A_D^{(1,k)}, A_D^{(k,1)},$$
$$A_D^{(1,k)} \setminus A_D^{(k,1)} \neq \emptyset, A_D^{(k,1)} \setminus A_D^{(1,k)} \neq \emptyset.$$

Figure.6.1 Relation entre les cinq classes d'anonymisation [Tassa et al, 2012]

Ces définitions d'anonymisation peuvent également être comprises via une terminologie graphique. Soit $D = \{R_1, \ldots, R_n\}$ une table et $g(D) = \{\acute{R}_1, \ldots, \acute{R}_n\}$ est la généralisation correspondante. Cette paire de tables définit un graphe biparti $V_{D, g(D)}$ sur l'ensemble des nœuds $D \cup g(D)$ où une *arête* relie $R_i \in D$ avec $\acute{R}_j \in g(D)$ si et seulement si les deux enregistrements sont cohérents. Avec cette formulation, $A_D^{(1.\ K)}$ ($A_D^{(k,\ 1)}$, ou $A_D^{(k,k)}$, respectivement) est la collecte de toutes les généralisations $g(D)$ pour lequel chaque nœud dans D ($g(D)$, ou $D \cup g(D)$, respectivement) dans le graphe biparti $V_{D, g(D)}$ a au moins un k degré.

Définition.4

Étant donnée la table D et sa généralisation $g(D)$, et $V_{D, g(D)}$ est le graphe biparti correspondant. L'enregistrement $\acute{R} \in g(D)$ est nommé un couple de $R \in D$ si $(R,\acute{R})$ est une arête dans $V_{D, g(D)}$ et il existe un couplage parfait dans $V_{D, g(D)}$ qui inclut cette arête. Si tous les enregistrements $R \in D$ ont au moins un k couples en $g(D)$. Alors $g(D)$ est appelé un k-concealment de D. [Tassa et al, 2012]

Nommément, k-concealment est une version améliorée de (1, k) –anonymat : Bien que (1, k) –anonymat exige que chaque enregistrement $R \in D$ a au moins un k arêtes adjacent dans $V_{D, g(D)}$. k-concealment exige que chaque enregistrement $R \in D$ aura au moins un k couples adjacent.

La relation entre la nouvelle classe d'anonymisation notée C^k_D et les classes précédentes est présentée dans *la figure 6.1.*

3. L'insécurité des k-types d'anonymisation

Dans cette section, nous allons discuter de l'insécurité des trois notions de base de (1, k), (k, 1) et (k, k) anonymisations, et nous montrerons qu'ils n'offrent pas le même niveau de sécurité que le modèle k-anonymat. Le but de cette discussion est de motiver la définition de k-concealment. Comme nous allons le présenter plus tard, ce modèle fournit le même niveau de sécurité que celui de k-anonymat (avec une plus grande utilité).

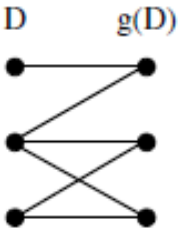


Figure 6.2 (k-1) anonymat [Tassa et al, 2012]

Considérant une base de données *D* et son (k, 1) anonymisation correspondante où chaque enregistrement de *g(D)* est cohérent au moins avec un k enregistrements de *D*. Cependant, il est possible que certains enregistrements $R \in D$ soient cohérents uniquement avec un seul enregistrement de *g(D)*, tel qu'il est illustré sur la *figure 6.2* (Il décrit un (2, 1) - anonymisation, mais le premier enregistrement dans *D* est cohérent uniquement avec un seul enregistrement de *g(D)*). Si l'adversaire arrive à cibler un individu dont son enregistrement dans *D* est cohérent uniquement avec l'enregistrement correspondant dans *g(D)*, alors il peut déduire, avec certitude, l'enregistrement généralisé qui correspond à son objectif individuel et, par conséquent, trouver l'attribut privée (sensible) correspondant de cet enregistrement.

Étant donné *g(D)* est (1, k) anonymisation de *D*, où les enregistrements de *D* sont cohérents avec au moins un k enregistrements de *g (D)*. Cette anonymisation semble satisfaire l'objectif de la confidentialité, en présentant l'exemple suivant qui montre le point faible de cette notion. Nous assumons que $D = \{R_1, \ldots, R_n\}$ et que $\acute{R}^*$ est l'enregistrement cohérent avec tous les enregistrements de *D*. Nous supposons que tous les entrées $\acute{R}^*$ sont supprimés ; Considérant la généralisation suivante : $g(D) = \{\acute{R}_1, \ldots, \acute{R}_n\}$, où $\acute{R}_i = R_i$ pour tous $1 \leq i \leq n-k$ et $\acute{R}_i = \acute{R}^*$ pour tous $n-k+1 \leq i \leq n$ (le graphie bipartie correspondant pour $n = 5$ et $k = 2$ est présenté dans *la figure 6.3*). Nous remarquons que $g(D) \in A_D^{(1.K)}$ et que la plus part des enregistrements de *g(D)* ne sont pas généralisés et par la suite, la perte de l'information est minimale. Cependant, une telle généralisation est totalement inacceptable. L'information privée de la plupart des individus représentés dans *D* est complètement révélé.

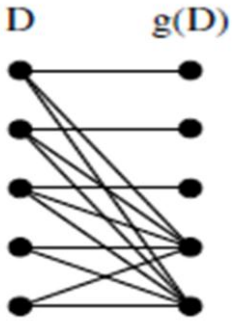


Figure 6.3 (1-k) anonymat [Tassa et al, 2012]

La notion de (k, k) -anonymat combine les deux notions précédentes et il semble qu'elle supporte les défauts de ces deux notions mentionnées ci-dessus. Toutefois, nous avons supposé que l'adversaire connaît le sous-ensemble exact de la population qui est représentée dans la table, et qu'il connaît les données publiques de chacune d'eux, il peut construire la table publique initiale *D*. Puisque la table anonyme *g(D)* est publiée, l'adversaire peut déduire le graphe biparti $V_{D,\ g(D)}$ où une arête relie l'enregistrement $R_i \in D$ à son enregistrement généralisé $\acute{R}_j \in g(D)$ si et seulement si $\acute{R}_j$ généralise R_i. La notion (k, k) –anonymat garantie que chaque nœud dans le graphe $V_{D,\ g(D)}$ ayant au moins un degré de K. Dans le cas où l'adversaire sait qu'une partie de *D*, cette borne inférieure du degré reste suffisante. Contrairement pour le cas où l'adversaire connait le graphe, il peut tester chaque arête et vérifier si elle peut être complétée dans un couplage parfait. Si elle ne peut pas, alors cette arête ne représente pas un lien possible entre l'enregistrement original et l'enregistrement généralisé.

En testant les arêtes du graphe et à travers la suppression des arêtes qui ne forment pas un couplage (*Définition 4*), l'adversaire peut obtenir les informations sensibles des nœuds de *D* qui ont moins de K couples. Cette discussion motive notre définition de k-concealment, comme un modèle sécurisé contre l'adversaire considéré.

4. La sécurité de k-concealment

Dans le modèle K-concealment, chaque $R_i \in D$ a au moins un k couples en *g(D).* Également comme le modèle k-anonymat, et en considérant l'adversaire décrit précédemment qui vise l'enregistrement particulier $R_i \in D$, il ne peut pas réduire le nombre d'enregistrements généralisés de *g(D)* à moins de k. Dans le modèle k-anonymat, les k enregistrements de *g(D)* sont identiques, où chacun d'eux est également susceptible d'être la véritable généralisation de

R_i, contrairement pour le modèle k-concealment où les enregistrements généralisés ne doivent pas être forcément identique lors de la généralisation. L'adversaire ne peut pas distinguer entre ces enregistrements. Pour qu'il puisse le faire, il doit résoudre au moins les problèmes k #P-complete ou bien utiliser les algorithmes d'approximation qui sont difficilement réalisables. Ce qui nous amène à conclure que le modèle k-concealment garantit le même degré de sécurité que le modèle k-anonymat [Tassa et al, 2012]**.**

5. Algorithmes

5.1. (k, k)- Anonymisation

Dans cette section, nous allons décrire l'algorithme de (k, k) -anonymisation. Premièrement, nous allons présenter une description de l'algorithme (k, 1) –anonymisation, puis, nous allons expliquer la façon de transformer cet algorithme pour qu'il devienne une version (k, k)- anonymisation.

5.1.1. (k, 1)- Anonymisation

Étant donnée une base des données $D = \{R_1, \ldots, R_n\}$, *l'Algorithme 1* réalise la version (k, 1) –anonymisation de D en trouvant pour chaque enregistrement R_i, $1 \leq i \leq n$, le meilleur sous-ensemble de k - 1 enregistrements supplémentaires tels que l'ensemble de coût de généralisation de ces enregistrements et de R_i, doit être minimal.

Algorithme 1 : (k, 1)-Anonymisation

Entrées : Table D, K : intègre
Sorties : Une table $g(D)$ qui satisfait la contrainte de (k, 1)-anonymat.
Pour tous $1 \leq i \leq n$ faire
  Pour toutes les sélections $\binom{n-1}{k-1}$ de k-1 enregistrements, $R_{i1}, \ldots R_{ik-1}$,
  Calculer le coût de généralisation $d(\{R_i, R_{i1}, \ldots, R_{ik-1}\})$.
    Soit $\{R_i, \ldots, R_{ik-1}\}$ La sélection qui représente le coût minimum de la généralisation minimum dans les étapes précédentes.
    Définir $\acute{R}_i$ comme étant l'enregistrement cohérent de $\{R_i, \ldots, R_{i\,k-1}\}$
  Fin pour.
Fin pour.

Algorithme 1. (k, 1)-Anonymisation [Tassa et al, 2012]

5.1.2. Transformation de (k, 1)-anonymisation à (k, k)- anonymisation

Soit $D = \{R_1, \ldots, R_n\}$ une base des données et $g(D) = \{\acute{R}_1, \ldots, \acute{R}_n\}$ son généralisation. Il existe une telle généralisation qui ne peut pas satisfaire le type (1, k)- anonymisation comme nous avons déjà présenté dans la section insécurité du type (k, 1)-anonymisation.

L'Algorithme 2 généralise encore plus les enregistrements de g*(D)* jusqu'ils deviennent un (1, k)-anonymisation. Spécifiquement, si un enregistrement donné R_i est cohérent avec seulement $\ell < k$ enregistrements dans *g(D)*, l'algorithme cherche d'autres enregistrements $(k - \ell)$ généralisés dans *g(D)* qui pourraient être généralisés encore plus pour devenir cohérents à R_i avec un coût minimum. Pour simplifier les notations, pour tout $R_i \in D$ et $\acute{R}_j \in g(D)$, nous notons $(R_i + \acute{R}_j)$ l'enregistrement minimal généralisé qui généralise R_i et $\acute{R}_j$.

En appliquant cet algorithme à la généralisation qui est déjà (k, 1)-anonyme, nous aboutissons à une table généralisée *g(D)* qui respecte (k, k) -anonymat.

Le temps d'exécution de *l'Algorithme 2* est $O(n^2)$: La boucle est constituée de *n* étapes, pour chaque enregistrement, nous vérifions le nombre de *n* enregistrements généralisés en *g(D)* avec lequel, cet enregistrement est cohérent. Si ce nombre est inférieur à *k*, il faut trouver les $(k - \ell)$ meilleurs enregistrements (opérations *O (kn)*) et les remplacer par *k* enregistrements généralisés. Par conséquent, le temps d'exécution global est *O (kn2)* (Le couplage de ces deux algorithmes est un : (k, k)- anonymisation).

Algorithme 2 : (1, k)-Anonymisation

Entrées *: Table D = {R1,…,Rn}, Table généralisé* g(D) *= {*Ŕ1*,…,*Ŕn*}, K : entier*

Sorties *: Une table g'(D) qui généralise* g(D) *et satisfait (1, k)-anonymat.*

Pour tous 1 ≤ i ≤ n faire

Soit ℓ est le nombre d'enregistrements Ŕj *qui sont cohérents avec Ri*

Si ℓ< k Alors

Scanner tous les enregistrements Rj qui ne sont pas cohérents

*avec Ri et trouver les k-ℓ qui minimisent c(Ri +*Ŕj*) - c(*Ŕj*).*

Remplacer chacun de ces enregistrements k-ℓ, Ŕj*, par Ri+* Ŕj

Fin si

Fin pour.

Algorithme 2. (1, k)-Anonymisation [Tassa et al, 2012]

5.2. Algorithme de K-concealment

Dans cette section, nous allons décrire *l'Algorithme 4* qui transforme (k, k)-anonymisation de la généralisation *g(D)* d'une base de données donnée *D* à une table k-concealment. Afin de comprendre l'idée principale de cet algorithme, nous commençons par trouver l'ensemble de tous les couples dans le graphe *(voir définition.4).*

5.2.1. Trouver tous les matches dans un graphe biparti

Soit $G = (U, V, E)$ un graphe biparti où $U = \{u_1,\ldots,u_n\}$, $V = \{v_1,\ldots,v_n\}$ et $E \subseteq U \times V$. Supposons également que G ait au moins un couplage parfait $\{(u_1,v_1),\ldots,(u_n,v_n)\}$. Selon *la Définition 4*, une arête dans E est appelée un couple si elle peut être étendue à un couplage parfait dans G.

Définition.5

Un ensemble d'arêtes $\ell \geq 1$ dans le graphe G est appelé un Bicycle (en respectant le couplage parfait assumé) s'il existe des indices ℓ, $1 \leq i_1,\ldots i_\ell \leq n$, telle que les ℓ arêtes sont :

$$(u_{i1},v_{i2}),(u_{i2},v_{i3}),\ldots\ldots,(u_{1\ell-1},v_{1\ell}),\ (u_{1\ell},v_{i1})\ldots\ldots\ldots \textbf{(1)}$$

Il est important de noter qu'un bicycle n'est pas un cycle. Par exemple, chacune des arêtes (u_i, v_i), $1 \leq i \leq n$, est un bicycle de longueur $\ell = 1$, ce qui n'est évidemment pas un cycle. Chaque bicycle de longueur $\ell \geq 1$ correspond à un cycle de longueur 2ℓ. En effet, si nous augmentons le bicycle ***(1)*** avec ℓ arêtes horizontales (u_{ij}, v_{ij}), $1 \leq j \leq \ell$, nous obtenons un cycle. L'inverse n'est pas vrai. Considérant l'exemple du graphe biparti présenté dans *la figure 6.4*, les quatre arêtes non horizontales forment un cycle de longueur 4 et qui ne correspondent à aucun bicycle. [Tassa et al, 2012]

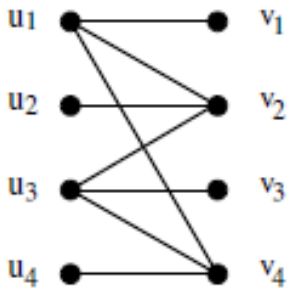


Figure 6.4. Exemple d'un graphe biparti [Tassa et al, 2012]

Définition.6

Soit $G = (U, V, E)$ un graphe biparti où $U = \{u_1,\ldots,u_n\}$, $V = \{v_1,\ldots,v_n\}$ et $E \subseteq U \times V$, et nous assumons que $\{(u_1, v_1),\ldots,(u_n, v_n)\} \subset E$. Alors l'arête $e \in E$ est un couple si et seulement si elle fait partie d'un bicycle. [Tassa et al, 2012]

Soit $G = (U, V, E)$ un graphe biparti qui correspond à une certaine anonymisation $g(D)$. $U = \{R_1, \ldots, R_n\}$ contient tous les enregistrements originaux et $V = \{\acute{R}_1, \ldots, \acute{R}_n\}$ contient tous les enregistrements généralisés. Puisque $\acute{R}_i$ est la généralisation de R_i, il existe une arête qui relie $u_i = R_i$ à $v_i = \acute{R}_i$ pour tous $1 \leq i \leq n$, et toutes ces arêtes sont des couples. Pour trouver toutes les autres arêtes de E qui sont des couples, nous procédons comme suit :

- Nous définissons le graphe orienté $H = (U, F)$ qui est induit par le graphe biparti $G = (U, V, E)$.
- Dans le graph orienté H, L'ensemble des nœuds est $U = \{u_1,\ldots,u_n\}$, et $(u_i, u_j) \in F$ si et seulement si $i \neq j$ et $(u_i, u_j) \in E$.

Nous remarquons dans *la définition.6*, une arête $(u_i, v_j) \in E$ est un couple dans G si et seulement si $i = j$ ou l'arête $(u_i, u_j) \in F$ fait partie d'un cycle dans H. Par conséquent, le problème de trouver tous les couples dans G se réduit au problème de trouver toutes les arêtes dans le graphe orienté H. Cela est réalisé comme suit :

- Trouver toutes les composantes fortement connexes de H (Un graphe orienté est fortement connexe s'il existe un chemin de chaque nœud dans le graphe vers chaque autre nœud).
- Si chaque composante fortement connexe est contractée à un seul nœud, le graphe résultant est un graphe orienté acyclique.
- Par conséquent, une arête donnée dans H est une partie du cycle si et seulement s'elle relie deux nœuds dans la même composante fortement connexes.

Algorithme 3 : Recherche de tous les couples dans un graphe biparti

Entrées: *Un graphe biparti G = (U, V, E), où U = {u1,…, un}, V= {v1, . . ., vn}, E ⊆ U × V, et pour tous 1 ≤ i ≤ n, (u1,….,v1)∈ E.*

Sorties: *Marquer toutes les arêtes de E, soit par OUI ou NON pour indiquer Si elles sont des couples dans G.*

```
Construire le graphe orienté H = (U, F) qui correspond à G.
Trouver toutes les composantes fortement connexes de H.
Pour toutes les arêtes (ui,uj) ∈ F faire
    Si ui et uj appartiennent à la même composante fortement connexe dans H,
marquer l'arête (ui,uj) ∈ E comme OUI,
```

```
      Sinon marquer NON
    Fin Si
Marquer toutes les arêtes (ui,uj) ∈ E, 1 ≤  i ≤ n, comme OUI.
Fin pour.
```

Algorithme 3. Recherche de tous les couples dans un graphe biparti [Tassa et al, 2012]

5.2.2. Algorithme

L'algorithme de k-concealment fonctionne comme suit : Pour chaque $R_i \in D$, il calcule le sous ensemble P de son ensemble des voisins Q, composé de tous les couples de R_i. Tant que $g(D)$ est (k, k)-anonymisation de D, alors $|Q| \geq K$, bien que $|P|$ doit être inférieur à K. Afin d'obtenir k-concealment, nous augmentons $|P|$ jusqu'elle devient au moins k. Si $|P| < k$**,** nous sélectionnons les voisins non couples $\acute{R}_{jh}$ de R_i qui minimisent la quantité $d_h = c(R_{jh} + \acute{R}_i) - c(\acute{R}_i)$. Ensuite, nous re-généralisons l'enregistrement $\acute{R}_i$ jusqu'il soit cohérent avec $R_{jh.}$ La discussion dans *la section 5.2.1* implique que cette mise à jour de $\acute{R}_i$ *améliore* $\acute{R}_{jh}$ d'un simple voisin de R_i à un couple avec R_i (puisque l'arête (R_i, R_{jh}) devient une partie d'un bicycle de longueur 2). Cette amélioration de l'arête (R_i, R_{jh}) en étant un couple peut avoir un effet similaire sur les autres arêtes. Par conséquent, nous recalculons l'ensemble des couples en répétant la procédure jusqu'à ce que $|P|$ devienne au moins k.

Algorithme 4 : Transformation de (k, k) -anonymat en k-concealment

```
Entrées : Table D={R1,…, Rn}, Table généralisé g(D)={Ŕ1,…,Ŕn} qui satisfait
au modèle (k, k)-anonymat, K : entier
Sorties : Une généralisation de g(D) qui satisfait au modèle k-concealment.
Trouver tous les couples dans le graphe VD, g(D) (Algorithme.3).
Pour tous 1 ≤ i ≤ n faire
      Q = {Ŕj1…, Ŕjq} est l'ensemble de q ≥ k voisins de Ri
      Extraire P, (Ensemble de Q qui consiste de tous les couples de Ri).
      Si |P| < K alors
      Pour tous 1 ≤ h ≤ q tel que Ŕjh ∈ Q.
Calculer dh = c (Rjh + Ŕi) – c(Ŕi).
       Sélectionner l'index 1 ≤ h ≤ q où Ŕjh∈ Q pour lequel h est minimal.
       Mettre Ŕi= Rjh+Ŕi
       Recalculer l'ensemble de tous les couples dans le graphe VD, g(D)
      Fin pour.
Fin pour.
```

Algorithme 4. Transformation de (k, k) -anonymat en k-concealment [Tassa et al, 2012]

Algorithme 4 invoque *l'Algorithme 3* une fois avant le lancement de la boucle principale. Ensuite, après chaque ajout d'une nouvelle arête dans le graphe, un recalcule des couples est nécessaire, vu que l'ajout d'une nouvelle arête pourrait améliorer plus d'une arête à un couple. Le calcule est fait par la conservation de la partition du graphe orienté $H = (U, F)$ aux composantes fortement connexes tout en gardant $H' = (U', F')$. (Le graphe orienté pour lequel $U' = \{C_1, \ldots, C_m\}$, est l'ensemble des composantes fortement connexes dans H, et F' ayant une arête de C_a à C_b si F possède une arête d'un nœud dans C_a à un nœud dans C_b). L'instruction $Ŕi = Rjh + Ŕi$ est équivalente à l'ajout d'une arête (u_{jh}, u_i) à H. Puisque ces deux nœuds appartiennent à deux composantes fortement connexes distinctes dans H ($u_{jh} \in C_1$ et $u_i \in C_2$). Elle se traduit également par l'ajout d'une nouvelle arête (C_1, C_2) à H'. En effet, ce que nous devons faire est d'appliquer l'algorithme de *Tarjan* sur le plus petit graphe H' ainsi que trouver dedans les composantes fortement connexes. Ces nouvelles composantes dans H' nous permettent d'éclairer comment mettre à jour la séparation de H dans les composants fortement connexes dont les arêtes dans H s'améliorent à être des couples *(Voir Exemple.1).*

Le temps d'exécution de *l'algorithme.4* est de l'ordre de $O(kn2)$.

Exemple.1

Supposons que H possède 5 composants fortement connexes. Par l'application de l'instruction Ŕi= Rjh+Ŕi, l'arête (u_{jh}, u_i) est ajoutée, où $u_{jh} \in C_1$ et $u_i \in C_2$. Ensuite, l'arête (C_1, C_2) est ajoutée à H', et on y exécute l'algorithme *Tarjan*. Supposons que suite à cette addition, les composantes fortement connexes dans H', sont $\{C_1, C_2, C_3\}$, $\{C_4\}$, $\{C_5\}$. Ainsi, nous concluons que H n'a pour l'instant que 3 composantes fortement connexes $C_1 \cup C_2 \cup C_3$, C_4, C_5. En outre, toutes les arêtes de H qui relient un nœud de C_i vers un nœud dans C_j, où $1 \leq i \neq j \leq 3$, seront des couples. Notons que le moins possible qui pourrait se produire suite à l'application de l'instruction Ŕi= Rjh+Ŕi, est l'unification de C_1 et C_2 (où $u_{jh} \in C_1$ et $u_i \in C_2$). En effet, puisque H possède déjà l'arête dans la direction opposée, (u_i, u_{jh}), (H' possède l'arête (C_2, C_1)), alors la nouvelle arête (C_1, C_2) rend $\{C_1, C_2\}$ fortement connexes dans H'. Cependant, comme il est possible que l'ajout de cette unique arête va créer une large composante fortement connexe, nous devons appliquer l'algorithme de Tarjan sur H'.

6. Conclusion

Dans ce chapitre, nous avons présenté le modèle de k-Concealment comme un modèle alternative de k-anonymat. Nous avons montré que k-Concealment offre essentiellement le

même niveau de sécurité que k-anonymat. Alors que ce dernier assure que chaque enregistrement de la table origine peut être lié à au moins k enregistrements de la table anonymes, K-concealment garantit une anonymisation des enregistrements sans la nécessité de généraliser tout l'ensemble de classe d'équivalence. L'avantage offert par ce modèle est l'augmentation d'utilité de données. Ainsi, k-concealment peut être réalisé avec moins de généralisation par rapport à ce qui est requis par k-anonymat.

Chapitre 7

Vers un protocole d'anonymisation distribué

1. Introduction

Dans le chapitre précédant, nous avons décrit le modèle d'anonymisation k-concealment utilisé dans notre solution distribuée. Ce chapitre vise à présenter notre approche distribuée qui a pour objectif de satisfaire les besoins des fournisseurs des données par une amélioration de la qualité de données. Dans ce chapitre nous allons :

- Clarifier les objectifs de confidentialité que nous visons dans notre travail.
- Présenter l'architecture de notre protocole d'anonymisation distribué basé sur le modèle k-concealment qui n'a pas encore été utilisé dans les environnements Cloud.
- Présenter l'algorithme d'anonymisation distribué pour les nœuds *Esclave*s et *Maître*.
- Expliquer le processus de Split utilisé dans notre algorithme, menant à un partitionnement plus efficace que celui utilisé dans les travaux précédents.
- Présenter le principe de la généralisation de l'arbre R*.

2. Les objectifs de la confidentialité

Dans cette section, nous allons présenter les deux principaux objectifs de confidentialité que nous nous basons dans notre travail.

- La confidentialité des individus, ou des objets des données : La base de données virtuelle et les résultats de la requête ne doivent pas contenir des informations d'identifications des individus. Pour atteindre cet objectif, nous avons adopté le modèle K-concealment présenté dans le chapitre précédant, qui est un modèle alternative de k-anonymat et qui offre une plus grande utilité.
- La confidentialité des fournisseurs de données : Les bases de données individuelles ne doivent pas révéler leurs données ou les propriétés de leurs données en dehors de la base de données anonyme virtuelle.

2.1. La confidentialité des objets des données sur la base de l'anonymisation

La préservation de la confidentialité des données publiées (PCDP) est un domaine de recherche évolutif qui vise à développer des techniques pour permettre la publication de données de telle sorte que la confidentialité doit être préservée et la généralisation de données doit être minimale. L'un des modèles les plus connus en PCDP est le modèle K-anonymat. Il

nécessite que chaque enregistrement publié comporte au moins (k) autres enregistrements dont les valeurs sont indistinctes. Une métrique est utilisée pour mesurer la quantité d'informations utiles lors de la modification des données, par la réduction de la quantité d'information perdue, ce qui augmente l'utilité de la table publié. Ainsi, l'objectif du modèle k-concealment est de généraliser les enregistrements de telle sorte que la table devient k-anonyme et la perte d'informations est minimale [Tassa et al, 2012].

2.2 La confidentialité entre les fournisseurs de données

Notre second but dans ce travail est d'assurer la confidentialité entre les fournisseurs des données à travers l'utilisation des protocoles du Calcul Multipartite Sécurisé (CMS) [Goldreich, 2001; Clifton et al, 2003; Vaidya et Clifton, 2004; Du et Atallah, 2001]. Le problème de CMS consiste à permettre à plusieurs parties de réaliser des calculs basés sur leurs données privées, de telle sorte que toutes les parties connaissent le résultat final du calcul, mais aucune d'entre elle ne puisse déduire les données privées des autres.

Dans notre travail, nous adoptons un modèle semi-honnête qui est utilisé dans le paradigme de CMS, où toutes les parties suivent les règles du protocole. Ce modèle est pratique pour notre scénario, où plusieurs organisations collaborent entre elles pour le partage des données et suivent le protocole pour obtenir le résultat précis.

3. Le protocole d'anonymisation distribué

3.1. Scénario de motivation

Nous supposons que les micros données sont distribuées entre S $(S > 2)$ sites (nœuds), et chaque nœud possède une base de données locale d_i $(1 \leq i \leq s)$. L'union de toutes les bases de données locales forme une vue globale de toutes les données ($D = \cup d_i$). En particulier, les caractéristiques des quasi-identifiants de chaque enregistrement sont identiques.

Pour préserver la confidentialité des données, notre travail se base sur le principe de k-concealment pour réaliser l'anonymisation des objets des données. Suivant le protocole d'anonymisation distribué, chaque site produit une base de données anonyme locale $\boldsymbol{a_i}$, qui satisfait son principe de confidentialité et leurs union forme une base de données virtuelle qui doit garantir le principe de k-concealment. Comme illustré dans *la figure 7.1*, les données sont distribuées horizontalement entre trois nœuds. Les ensembles de données anonymes a_1, a_2, et a_3 des bases de données d_1, d_2, et d_3 respectivement satisfont au modèle k-concealment.

Lorsque les utilisateurs font leurs requêtes sur la base de données virtuelle, chaque base de données d_i exécute la requête sur les données anonymes $a_{i,}$ et participe dans le protocole pour composer les résultats.

Dans l'exemple suivant, l'âge apparaît dans la table dans plusieurs enregistrements, il peut être laissé inchangeable dans certains cas, ou bien généralisé dans un intervalle *(Figure 7.1)*

Table origine

Age	Code postal	Maladie
18	08009	Grippe
22	08021	Angine

Table Concealment

Age	Code postal	Maladie
18-21	08005-08015	Grippe
22-25	08021	Angine

Nœud 1

Age	Code postal	Maladie
29	22004	Cancer
24	22015	Angine

Age	Code postal	Maladie
26-30	22001-22010	Cancer
22-25	22011-22021	Angine

Nœud 2

Age	Code postal	Maladie
20	31002	Cancer
34	31013	Angine

Age	Code postal	Maladie
18-21	31002	Cancer
34	31011-31019	Angine

Nœud 3

Figure 7.1 Les différentes tables K-concealment dans les trois nœuds

Age Zipcode Disease
18 10043 Flu
22 10028 Angina

3.2 Modèle de confidentialité

Il existe un grand nombre d'algorithmes proposés pour atteindre le modèle k-anonymat. Ces algorithmes ont été parfaitement utilisés pour améliorer l'opération d'anonymisation dans les environnements distribués. D'après les recherches que nous avons menées, le modèle k-concealment n'est pas encore utilisé dans les environnements distribués et plus particulièrement dans le Cloud. Il existe généralement deux types d'approches pour calculer un ensemble de données anonymes : *l'approche ascendante* et *l'approche descendante.* Dans

ce travail, nous avons choisi *l'approche ascendante* pour atteindre le modèle k-concealment pour une meilleure efficacité de généralisation *(voir chapitre 4).*

Selon notre modèle de confidentialité, nous introduisons un nouvel algorithme d'anonymisation qui contrôle les serveurs de Cloud. Étant donné une version centralisée de l'algorithme d'anonymisation, nous pouvons le décomposer et utiliser des protocoles distribués sécurisés pour la communication afin de générer une anonymisation distribués entre les serveurs de Cloud. *La figure 7.2* illustre l'architecture du protocole d'anonymisation distribué proposé.

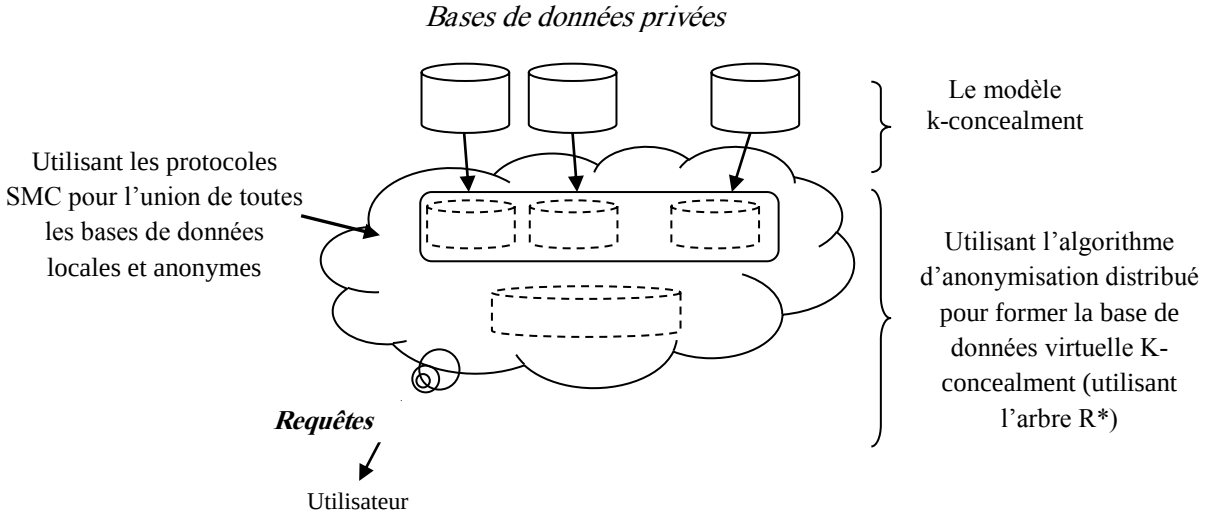


Figure 7.2 Architecture de notre protocole d'anonymisation distribué

3.3 Algorithmes

L'idée principale du protocole d'anonymisation distribué est d'utiliser les protocoles CMSs pour réaliser la méthode de l'Arbre-R* pour l'environnement Cloud, de sorte que chaque base de données locale produit un ensemble de données anonyme pour répondre aux besoins des fournisseurs, et leur union forme une base de données anonyme et virtuelle qui doit satisfaire le modèle k-concealment *(figure 7.2).*

Algorithme 5 : Anonymisation distribué pour le nœud-Maître

Entrées *: Classe d'équivalence de la table initiale (locale)*
Sorties *: Le rectangle B contenus dans R**
/***Phase 1:***Lecture des données (Esclave) */*
Lire la donnée (***B, num***) de chaque nœud esclave dans l'ensemble ***r***
/***Phase 2:***Insertion dans l'arbre R* (Généralisation) */*
Pour chaque rectangle *B* ∈ *r Faire*
    Trouver dans chaque niveau de l'arbre, le sous arbre le plus approprié Pour accueillir la nouvelle entrée (*le processus de la généralisation de l'arbre R**).
    Si le Split est possible, Partitionner*(Le processus de Split*).
Fin pour
/* **Phase 3:***Modification des bases des données locales k-concealment */*
Pour chaque classe d'équivalence E_i de la table *K-concealment* Faire
    Obtenir l'ensemble des rectangles ***P*** contenus dans ***R**** de E_i
    Pour chaque rectangle ***B*** ∈***P*** Faire
        Envoyer ***B*** et ***R**** au nœud esclave
    Fin pour
Fin pour.

Algorithme 5. Anonymisation distribué pour le nœud-Maître

Algorithme 6 : Anonymisation distribué pour le nœud-Esclave i (I>0)

Entrées *: Le rectangle B contenu dans R**
Sorties *: E_j : Classe d'équivalence élargit*
/* **Phase 1:***Envoi les données au nœud Maître **
Pour chaque classe d'équivalence E_i de la table *K-concealment* Faire
    Envoyer l'information de E_j au nœud Maître dans la forme (***B, num***)
Fin pour
/* **Phase 2:***La réception de la modification du maître **
Lire les données ***B*** et ***R**** envoyées par le nœud Maître
Pour chaque classe d'équivalence E_i de la table *K-concealment* Faire
    Si le rectangle de E_j égale à ***B***
        Élargir le rectangle de E_j à ***R****
    Fin si
Fin pour

Algorithme 6. Anonymisation distribué pour le nœud-Esclave

Nous supposons que le nœud *Maître* est sélectionné à partir du serveur de Cloud, et toutes les bases de données locales sont considérées comme des nœuds *Esclaves*. Les protocoles pour le nœud *Maître* et les autres nœuds *Esclaves* sont présentés dans *les algorithmes 5* et *6*.

La procédure réalisée au niveau du nœud *Maître* est similaire à la méthode de généralisation centralisée **[**Beckmann et al, 1990**]**. D'abord, le nœud *Esclave* envoie une donnée qui est une information abstraite de chaque classe d'équivalence pour le nœud *Maître* sous forme (*B, num*), où *B* désigne un rectangle de dimension *d* qui est un cadre de sélection des valeurs spatiales *QI* des classes d'équivalences $[l_1,u_1]$, ...,$[l_d,u_d]$, et *num* est le nombre totale des objets des données dans une classe d'équivalence.

Dans la première phase, le nœud *Maître* lit les données (*B, num*) de chaque classe équivalence envoyée à partir de chaque nœud *Esclave* dans un ensemble ***r.***

Dans la deuxième phase, l'algorithme insert chaque rectangle *B* dans un arbre *R** ; dans chaque niveau de l'arbre, il trouve le sous arbre le plus approprié pour accueillir la nouvelle entrée (minimum de chevauchement), et si le split est possible, il le fait.

Dans la troisième phase, le nœud *Maître* modifie toutes les bases de données initiales (k-concealment) en traversant chaque classe d'équivalence E_i de la table k-concealment pour trouver l'ensemble des rectangles ***P.*** Pour chaque rectangle $B \in P$ qui contient dans le cadre de sélection *R** de E_i, nous envoyons les données *B* et *R** vers le nœud *Esclave* correspondant. Quand le nœud *Maître* reçois les données *B* et *R**, le nœud *Esclave* découvre la classe d'équivalence dont le rectangle est égal à *B* et il élargit son rectangle à *R**.

ID	Age	Code postale
1	21-30	10000-11000
2	21	10000-11000
3	21-30	10000-11000

Nœud.1

ID	Age	Code postale
4	18-25	20000-22000
5	18-25	20000-22000
6	30-35	13000-17000
7	30-35	13000-17000

Nœud.2

ID	Age	Code postale
8	40-50	18000-20000
9	40-50	18000-20000
10	40-50	18000

Nœud.3

Figure 7.3 Les tables k_i-concealments initiales

ID	Age	Code postale
1	18-30	10000-22000
2	18-30	10000-22000
3	18	10000-22000

Nœud.1

ID	Age	Code postale
4	18-30	10000-22000
5	18-30	10000-22000
6	30-50	13000-20000
7	30-50	13000-20000

Nœud.2

ID	Age	Code postale
8	30	13000-20000
9	30-50	13000-20000
10	30-50	13000-20000

Nœud.3

Figure 7.4 Les tables k_i-concealments modifiées

Nous illustrons le protocole avec un exemple de scénario comme le montre les *figure 7.3* et *la figure 7.4.* Dans ce scénario, trois fournisseurs de données publient leurs bases de données privées à travers trois nœuds *Esclave*s avec leurs contraintes de confidentialités personnalisées : $k_1 = 3$, $k_2 = 2$, $k_3 = 2$ et $k_4=3$. Avant la réalisation de l'algorithme d'anonymisation distribué, trois tables initiales k-concealment ont été construites comme le montre *la figure 7.3*. Le premier nœud satisfait la contrainte de 3-concealment, le deuxième : 2-concealment et le troisième : 3-concealment.

L'algorithme modifie les rectangles de classe d'équivalence du premier nœud du [21-30][10000-11000] à [18-30][10000-22000], le rectangle des deux classes d'équivalences du deuxième nœud du [18-25][20000-22000] à [18-30][10000-22000], [30-35][13000-17000] à [30-50][13000-20000] respectivement, et le troisième nœud du [40-50][18000-20000] à [30-50][13000-20000].

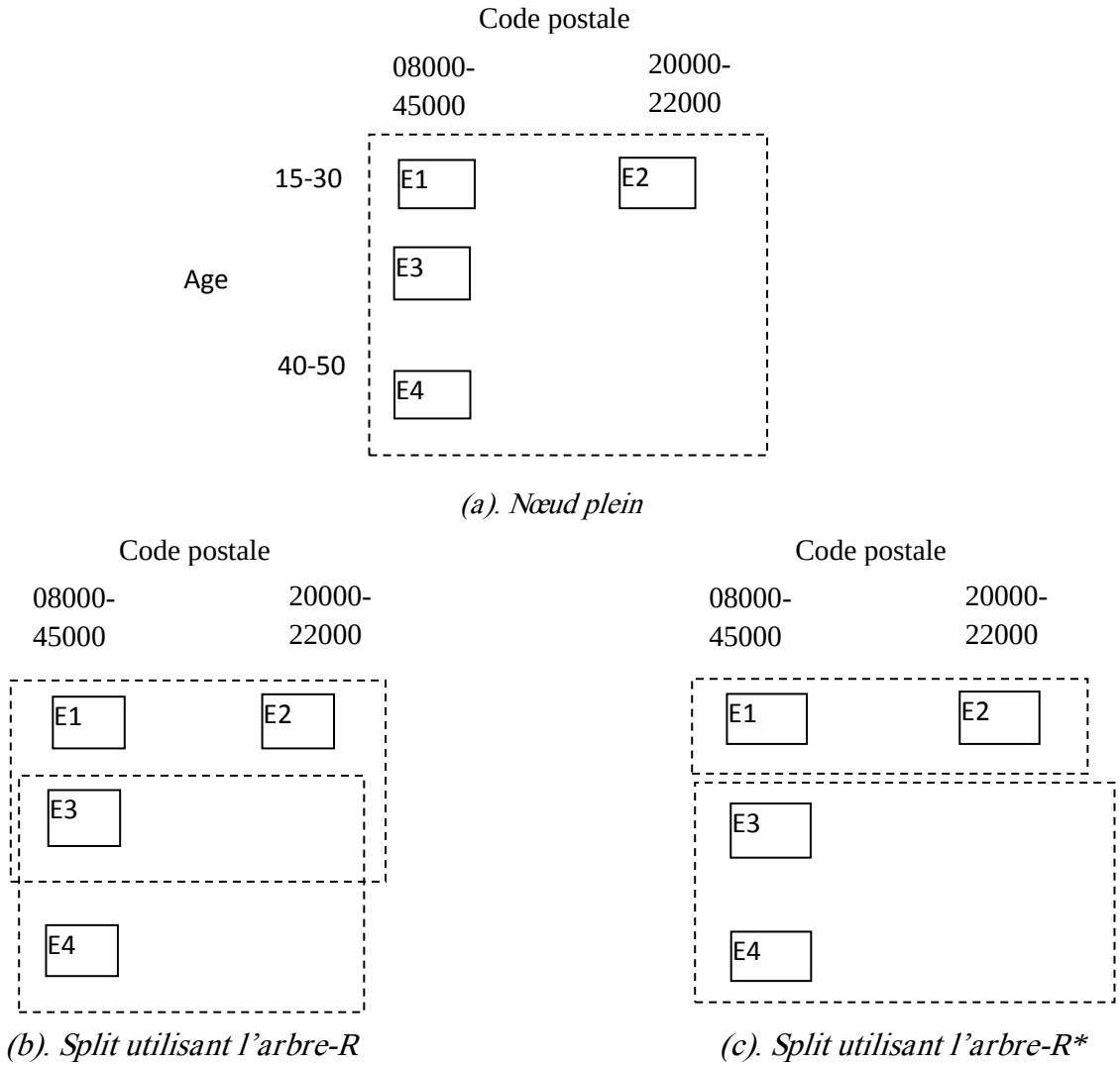
7. Le processus de Split

Le processus de split effectué pour notre algorithme est similaire au split utilisé dans l'arbre-*R* *, qui implique les rectangles de classes d'équivalence mises dans les bases de données locales k-concealment. L'objectif est de deviser les données autant que possible tout en respectant les contraintes de la confidentialité afin de maximiser l'utilité de données anonymes. L'exemple de la *figure 7.5* présente la stratégie de l'arbre *R** (c) **[**Beckmann et al, 1990] qui mène à un partitionnement plus efficace que celui de l'arbre *R* (b) [Guttman, 1984]. Pour trouver un meilleur partitionnement, l'arbre *R** utilise les étapes suivantes :

- Dans la première étape, l'axe de split doit être choisi. Au long de chaque axe, les entrées sont triées selon la valeur inférieure, et puis triées selon la valeur supérieure de

leurs rectangles (rectangles des classes équivalences). Pour chaque sorte *M-2m + 2* distributions de *M + 1*, les entrées en deux groupes sont déterminées (***M*** est le nombre maximal d'entrées qui s'adaptent dans un seul nœud, et ***m*** est le nombre minimum d'entrées). La k-th distribution (k=1, …., M-2m+2) est déterminée comme suit : Le premier groupe contient le premier (m-1) + k entrées, le second groupe contient les entrées restantes. Pour chaque distributions M-2m + 2, la valeur de marge est déterminée en résumant la longueur de la marge des deux zones de sélections minimales (rectangles des classes équivalences) des deux distributions. Enfin, L'axe qui renvoie la valeur minimale de marge est sélectionné comme un axe de split.

- Dans l'étape suivante, un partitionnement adéquat des entrées au long de l'axe de split est déterminé. La valeur du chevauchement de chaque distributions *2(M–2m+2)* est considérée où la valeur du chevauchement désigne la taille de la surface de chevauchement entre les deux cadres de sélection minimum de deux partitions. À ce niveau, les protocoles sécurisés de somme **[**Ottcher et Obermeier, 2008**]** peuvent être utilisés pour calculer la valeur de surface minimum.



(a). Nœud plein

(b). Split utilisant l'arbre-R

*(c). Split utilisant l'arbre-R**

Figure 7.5 Le processus de Split dans l'arbre-R et l'arbre-R*

8. La généralisation de l'arbre-R*

Similairement à l'algorithme multidimensionnel de Mondiran **[**Jurczyk et Xiong, 2009**]** qui généralise les données en utilisant l'arbre-*Kd* **[**Friedman et al, 1977**]**, et à l'algorithme **[**Ding et al, 2013**]** qui utilise l'arbre-*R* **[**Guttman, 1984**]**, notre algorithme utilise l'arbre- *R** qui généralise les données en l'insérant dans le chevauchement minimum de l'espace de domaine de quasi-identifiant. Quand un débordement se produit, l'espace de domaine de quasi-identifiant sera divisé en deux parties en choisissant le meilleur axe. Il choisit de manière récursive la meilleure branche à insérer les objets de données pour recueillir les objets de données les plus proches.

Tous les nœuds feuilles dans l'arbre-*R** seront récupérés et rassemblés pour la création de la table k-concealment. Un nœud non-feuille contient des entrées sous forme *(cp, B)* où *cp* est l'adresse du nœud fils et *B* est le rectangle de sélection minimum de tous les rectangles qui sont entrés dans ce nœud fils. Un nœud feuille contient des entrées sous forme *(Oid, B)* où *Oid* est un enregistrement dans la base de données, et $B= (B_1, B_2, \ldots, B_d)$ est un rectangle de dimension ***d*** qui est un cadre de sélection des valeurs spatiales *QI* (***d*** est le nombre de dimensions et ***B*** est un intervalle borné [x, y] qui décrit la valeur de *QI* au long de cette dimension).

La *figure 7.6* montre deux arbres- *R** dimensionnels où les tuples sont regroupés dans 7 nœuds feuilles $R_1, \ldots, R_7$, et en fonction de leur proximité spatiale, ils sont regroupés de manière récursive dans les racines *D1, D2*.

La construction de l'arbre-*R** est basée sur l'algorithme de l'insertion qui se base sur :

- *Le choix de la sous-arborescence* : À partir de la racine, jusqu'à la feuille, il trouve dans chaque niveaux le sous arbre le plus approprié pour accueillir le nouvel objet.
- *Le split* : Si la première étape se termine par un nœud rempli avec un nombre maximum d'objets *M*, le split doit distribuer *M + 1* rectangles en deux nœuds de la manière la plus appropriée.

Nous créons l'arbre-*R** en insérant chaque objet dans la structure d'indexation. Étant donné un nouveau objet, les auteurs dans **[**Ding et al, 2013**]** n'ont considéré que le critère de la zone. Dans notre algorithme, nous testons les paramètres de la zone et du chevauchement dans des différentes combinaisons. Le chevauchement d'une entrée est défini comme suit.

Soit $E_1, \ldots, E_p$ des entrées :

$$Overlap\ (E_k) = \sum_{i=1,\ i\neq k}^{p} area\ (E_i \cap E_k), 1 < k < p$$

Pour le choix d'un meilleur nœud non-feuille, nous pouvons utiliser la méthode proposée par [Guttman, 1984]. Pour les nœuds feuilles, la minimisation du chevauchement donne des résultats beaucoup plus efficaces. Supposons qu'il y a une insertion d'un objet ***x*** *(La figure 7.6).* Au niveau de la racine (non-feuille), l'algorithme choisit l'entrée dont le rectangle réalise l'élargissement dans une surface minimale pour couvrir *x* (*D1* est choisi). Pour le nœud feuille, l'algorithme choisit l'entrée dont le rectangle réalise le chevauchement minimal pour

couvrir *x* (*R3* est choisi). L'opération du Split est effectuée lorsqu'un nœud est rempli avec un nombre maximum d'objets de données.

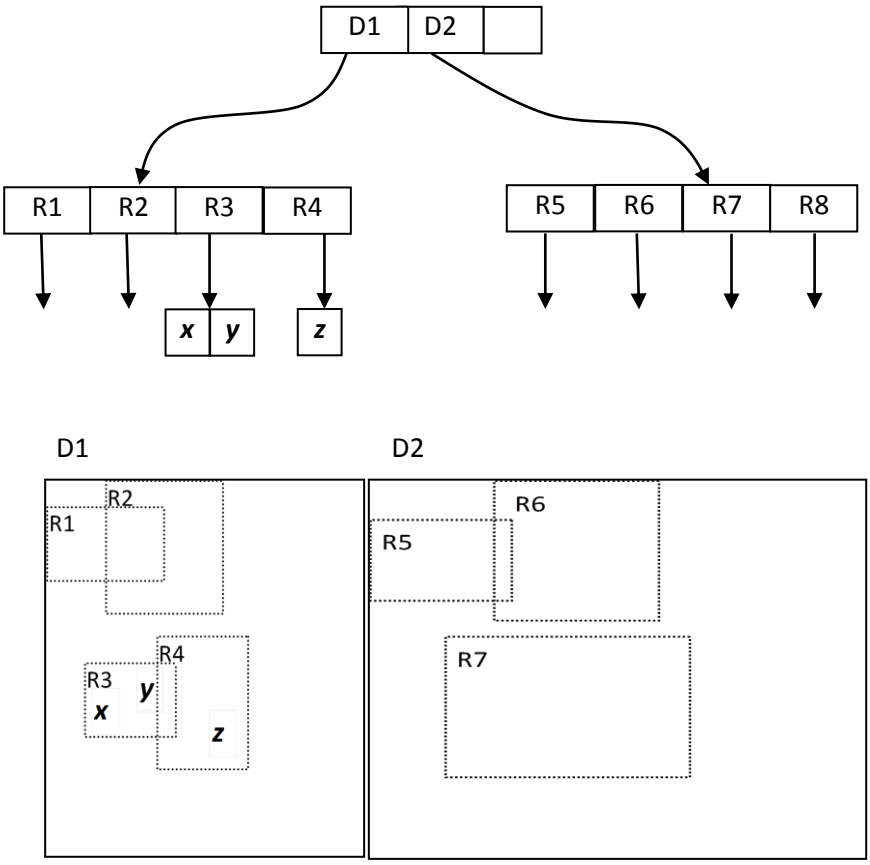


Figure 7.6. La structure de l'arbre-R*

9. Conclusion

Dans ce chapitre, nous avons présenté notre protocole d'anonymisation distribué pour la préservation de la confidentialité des données publiées à partir de plusieurs fournisseurs de données dans un environnement de Cloud Computing. Au début, nous avons commencé par schématiser l'architecture de l'approche distribuée et proposer le nouvel algorithme d'anonymisation distribué pour les nœuds Maître et Esclaves. Une explication du processus de Split utilisé dans notre protocole est donnée. Ce processus vise à fournir un partitionnement plus efficace que celui utilisé dans les travaux précédents. Nous finissons par présenter la généralisation de l'arbre R* qui insère les données dans un chevauchement

minimum. Cette opération de généralisation, choisit de manière récursive la meilleure branche à insérer les objets de données pour recueillir les objets de données les plus proches.

Afin de valider notre contribution, nous allons détailler, dans le chapitre suivant, les expérimentations menées pour évaluer la qualité d'anonymisation fournit par le protocole d'anonymisation distribué proposé.

Chapitre 8

Implémentation et Évaluation

1. Introduction

Après avoir présenté l'algorithme et la stratégie de partitionnement, nous passons dans cette partie à détailler la procédure d'expérimentation que nous allons adopter pour valider nos travaux. Nous exposons les déférentes étapes et scénarios de l'implémentation de notre approche en évaluant la qualité d'anonymisation fournit par notre protocole. Les étapes de notre implémentation sont comme suit :

- Nous montrons l'apport de notre protocole d'anonymisation (qualité des données anonymes) à travers une comparaison de plusieurs paramètres avec le système centralisé.
- Nous évaluons la qualité de généralisation du protocole en comparant notre algorithme avec celui utilisant l'arbre R.
- Pour la performance de notre protocole, nous comparons les deux scenarios (L'arbre R* est utilisé pour la satisfaction de la contrainte de confidentialité k-concealment et la satisfaction de la contrainte de confidentialité k-anonymat).

- **Ensemble de données**

Nous avons testé notre algorithme sur la base de données du Bureau du recensement des États-Unis ADULT qui contient les caractéristiques démographiques d'un petit échantillon de la population américaine. ADULT contient 48842 enregistrements avec 14 attributs publics (Age, WorkCLass, etc.). L'information privée signifie que les gains individuels sont plus ou moins de 50 mille dollars par an. Nous avons utilisé cet ensemble de données vu que la majorité des travaux qui touchent l'aspect anonymisation l'utilise [84],[100],[98]. Ceci Va nous permettre de mieux comparer nos résultats avec les autres travaux.

2. L'effet du protocole d'anonymisation distribué

Dans cette partie nous présentons la partie expérimentation de notre protocole distribué par l'évaluation de la qualité d'anonymisation des données en comparant ces résultats avec celles obtenus par l'exécution de l'approche d'anonymisation centralisée.

2.1. Les métriques

L'objectif de l'éditeur de données n'est pas seulement d'assurer une publication sécurisée des données, mais aussi une garantie de leurs utilités. La métrique est utilisée pour mesurer la

quantité d'informations utiles des données après une telle publication. L'éditeur de données a besoin de ces mesures pour l'évaluation de l'utilité de différentes publications ainsi que le destinataire qui en a besoin pour mesurer l'utilité d'une analyse.

Lors de la publication d'un ensemble de données, certaines informations sensibles sont invariablement supprimées. Cependant, la conséquence non voulue est que certaines informations agrégées ou statistiques sont également perdues. C'est pour cette raison que l'utilité d'un ensemble de données est mesurée par une mesure qui préserve les informations agrégées et statistiques.

Plusieurs mesures d'utilité ont été proposées dans la littérature et qui prennent diverses formes, [Gionis et al, 2008] ont présenté un aperçu général sur ces métriques. Dans notre expérimentation, nous allons utiliser la métrique LM (Loss Metric) qui semble être la mesure la plus précise par rapport aux autres mesures [Tassa et al, 2012]. Cette métrique est définie en termes de perte normalisée pour chaque attribut de chaque enregistrement.

Étant donnée la table d'enregistrements D= $\{r_1,..,r_n\}$ dans $A_1,...,A_z$, $g(D)$ représente la généralisation correspondante de D. Étant donné |A| qui représente la taille du domaine de l'attribut A, et |r| représente le nombre de valeurs dans ce domaine qui peuvent être généralisées. La métrique LM pour $r(A)$ est représentée par la formule :

$$r(A) = \frac{(|r| - 1)}{(|A| - 1)}$$

La perte d'information globale est la somme des pertes d'information pour chaque attribut. Pour l'évaluation, nous nous basons sur les scénarios suivants :

- Nous exécutons l'algorithme de l'arbre R* en localisant les données dans une base de données centralisée (approche centralisée).
- Les données sont distribuées sur quatre nœuds et nous utilisons l'approche d'anonymisation distribuée présentée dans le chapitre précédant.

Nous effectuons une comparaison de ces deux scénarios pour des valeurs différentes de *k* et *d*. *Les figures 8.1* et *8.2* présentent la perte d'information par rapport à différentes valeurs de *k* et *d* respectivement. Nous observons que notre protocole d'anonymisation distribué produit une perte d'information moindre par rapport à l'approche centralisée qui souffre de l'utilité de données particulièrement lorsque nous considérons *l-diversité* comme une mesure de sécurité supplémentaire (*figure 8.3*).

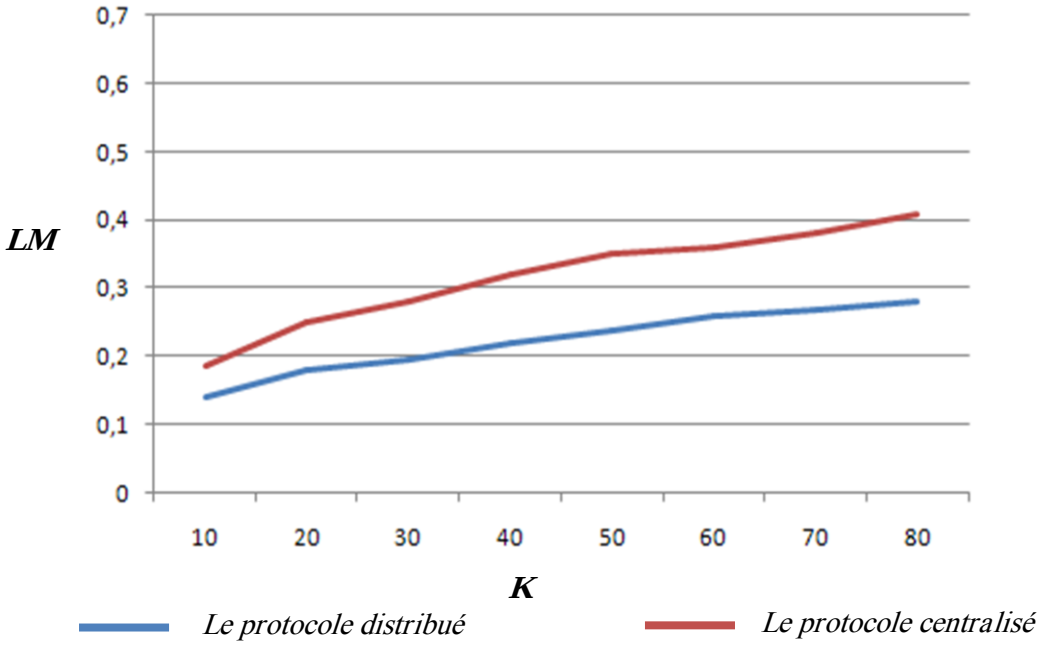


Figure 8.1 L'information perdue vs K

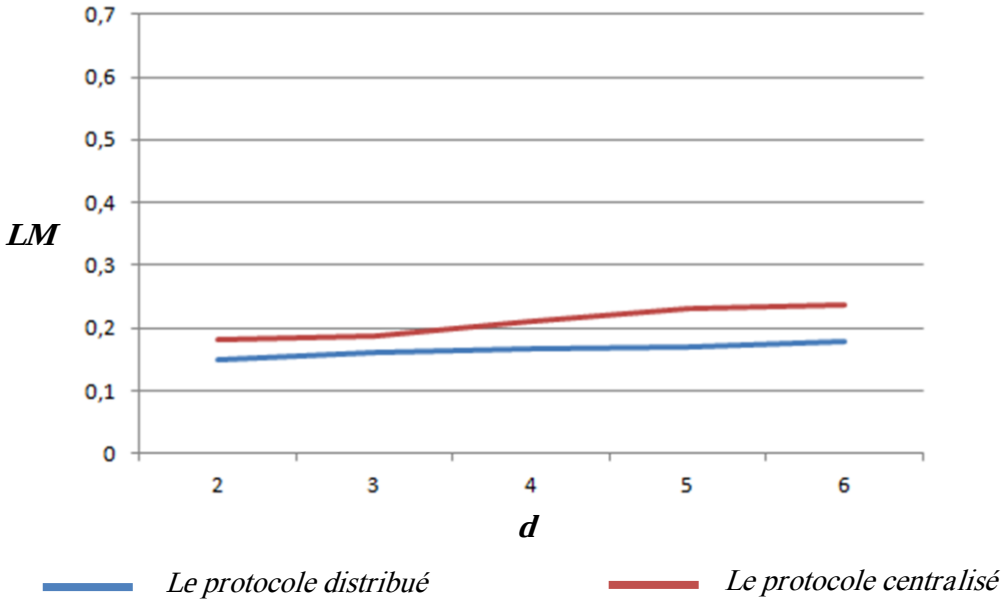


Figure 8.2 L'information perdue* vs *d

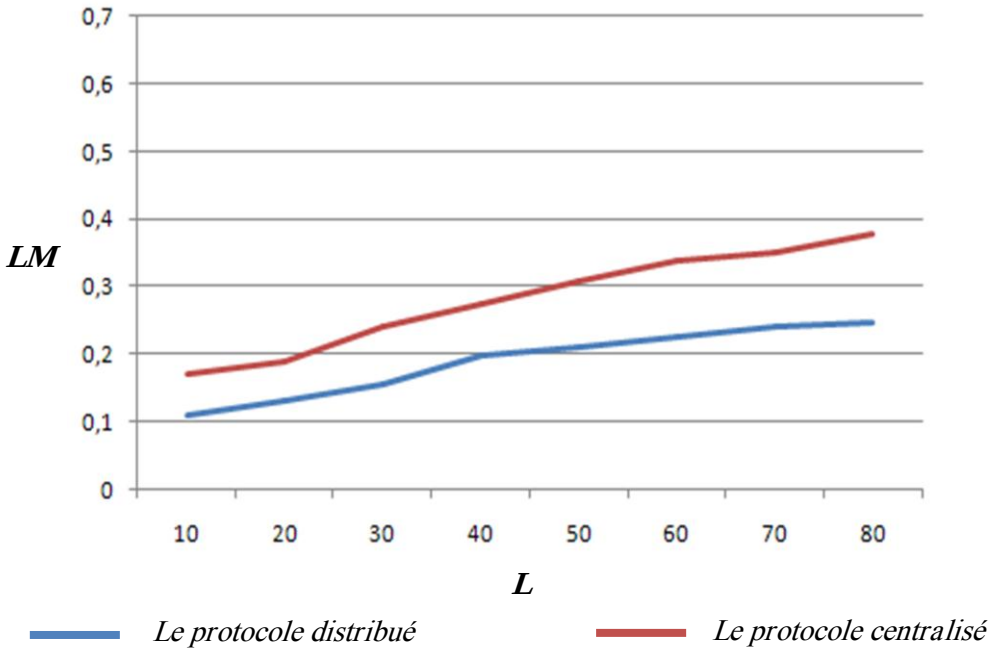


Figure 8.3 L'information perdue* vs *mesure de diversité L

3. La performance pour la satisfaction de l'anonymisation pour les fournisseurs de données.

Dans la deuxième partie de notre implémentation, nous nous intéressons aux performances de notre protocole par rapport à celui qui utilise l'arbre R. Les deux méthodes sont utilisées dans chaque base de données pour la généralisation des données, et l'union de ces données est envoyé au nœud *Maître* pour former une base de données globale qui satisfait les contraintes de confidentialité.

Étant donné une requête, pour la généralisation des valeurs d'attribut, la possibilité de faire l'opération inverse de l'anonymisation (transformer les enregistrements à partir de la table anonyme) peut faire un chevauchement avec le prédicat de sélection. Ce qui va produire des plus grands résultats par rapport à l'évaluation du prédicat de la table d'origine.

Pour cet ensemble d'expériences, nous calculons *l'erreur correspondante* qui est définie comme suit :

$$\boldsymbol{Erreur\ correspondante} = \frac{|\boldsymbol{Actual} - \boldsymbol{Estimate}|}{|\boldsymbol{Actual}|}$$

Où *actual* est la réponse correcte de la requête et *estimate* est la réponse calculée de la table anonyme.

Pour chacun des paramètres testés, nous envoyons 100 requêtes générées et nous rapportons une erreur moyenne correspondante sur les mêmes paramètres. Pour le calcul de l'erreur :

- Nous fixons le nombre de dimension à 3 et nous faisons varier k.
- Le nombre de dimension est varié entre 2 à 5, et nous fixons k à 10.

Les résultats sont présentés dans les *figures 8.4 et 8.5.*

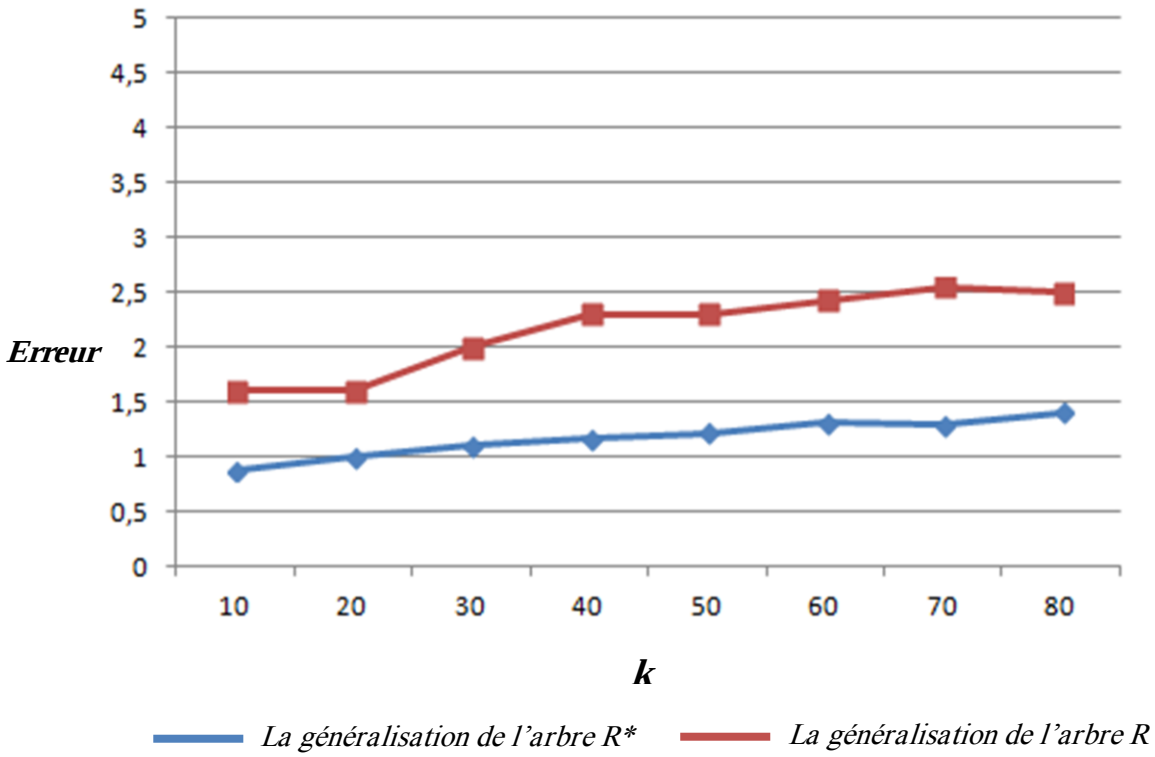


Figure 8.4 L'erreur correspondante* vs *K

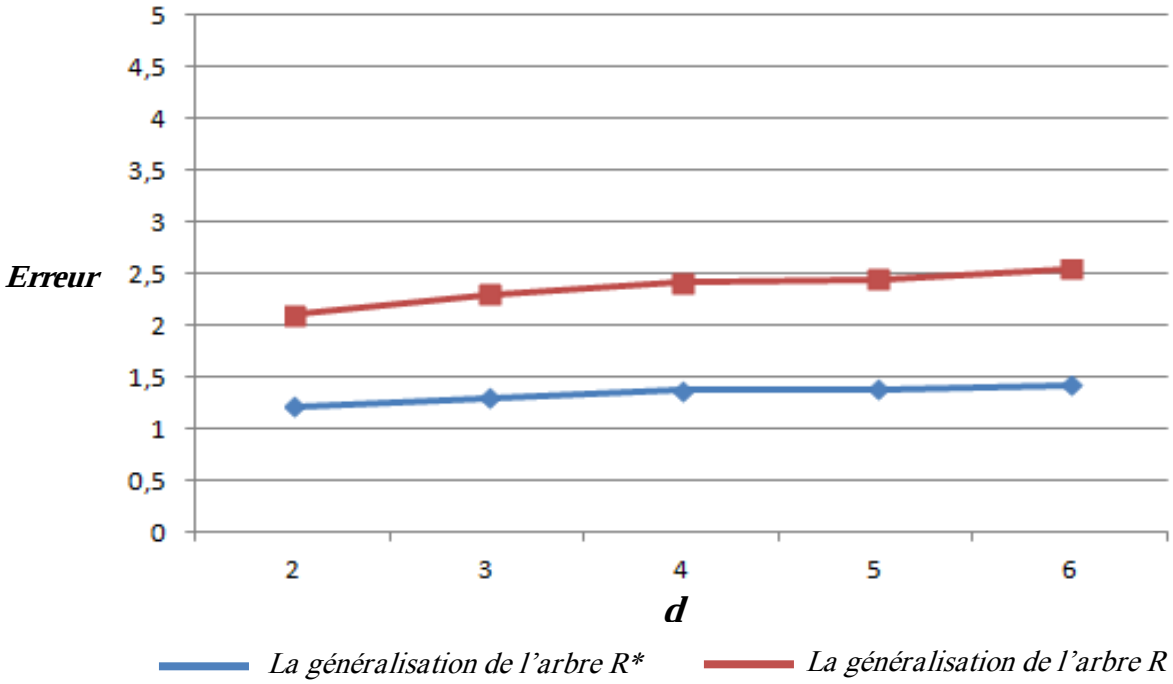


Figure 8.5 L'erreur correspondante* vs *d

Une autre expérimentation qui va nous permettre de tester la performance du protocole proposé est comme suit : Nous effectuons un premier test où la méthode R* est utilisée pour satisfaire la contrainte K-concealment, et un deuxième test où et la même méthode est utilisée pour tester la contrainte K-anonymat, (Voir *figure 8.6*).

Comme est illustré dans *les figures 8.4* et *8.5*, nous avons constaté que les requêtes fournit par notre protocole sont beaucoup plus précis que celles fournit par l'algorithme de généralisation de l'arbre R. Ceci est dû à l'utilisation de cet arbre, qui ne regroupe pas les objets de données proches dans les mêmes rectangles minimale. Ainsi *la figure 8.6* qui présente les informations perdues causées par le modèle *K-concealment* qui sont moins par rapport au modèle *K-anonymat*, ce qui nous indique une meilleure utilité de données généralisées.

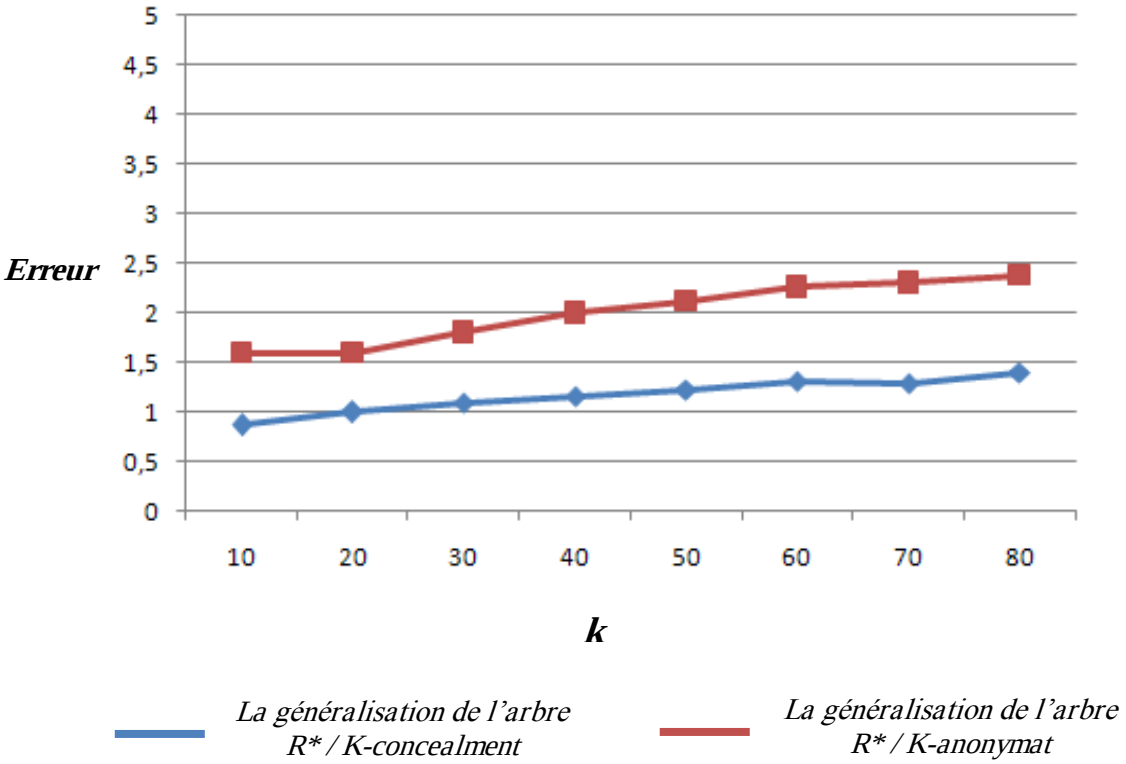


Figure 8.6 Arbre R*/K-concealment* vs *Arbre R*/K-anonymat

4. Synthèse

La comparaison entre nos résultats avec ceux existants dans la littérature, nous a permis de mettre en valeur nos contributions. Notre protocole d'anonymisation distribué, considéré comme une nouvelle technique pour préserver la confidentialité dans l'environnement Cloud, a permis d'améliorer deux points importants : la confidentialité des objets des données et la confidentialité des fournisseurs des données.

- Pour la confidentialité des objets de données (individus), nous avons utilisé le modèle k-concealment qui offre une meilleur utilité avec une généralisation minimum par rapport de ce qui est requis par le modèle k-anonymat utilisé dans [Jurczyk et Xiong, 2009] et [Ding et al, 2013].
- Pour la confidentialité des fournisseurs des données, nous avons adopté un algorithme qui utilise l'index de l'arbre-R* et qui donne une meilleure généralisation par rapport aux autres approches [Jurczyk et Xiong, 2009] et [Ding et al, 2013]. Nous avons également montré que la stratégie de l'arbre-R* conduit à un split et une insertion plus effective que celle utilisé dans l'arbre-R.

5. Conclusion

Dans cette partie, nous avons détaillé la procédure d'expérimentation pour la validation des résultats obtenus. Une exposition des déférentes étapes et scénarios de l'implémentation de notre approche a été présentée en évaluant la qualité d'anonymisation fournit par le protocole d'anonymisation distribué. Ce chapitre regroupe deux phases principales.

Durant la première étape, nous avons évalué la qualité d'anonymisation des données à travers des métriques tout en comparant ces résultats avec celles obtenus par l'exécution de l'approche d'anonymisation centralisée. Cette comparaison est faite selon des scénarios où l'exécution de l'algorithme de l'arbre R* est effectuée sur une base de données centralisée ensuite sur une autre distribuée sur des différents nœuds.

La deuxième étape consiste à mesurer la performance de notre protocole par rapport à celui qui se base sur l'arbre R par une évaluation de la qualité de généralisation. Une autre comparaison est faite suivant un scénario où l'arbre R* est utilisé pour satisfaire la contrainte de confidentialité K-concealment et l'utilisation de la même méthode pour satisfaire la contrainte de confidentialité k-anonymat. Nous avons conclu par une synthèse qui a résumé nos principales contributions.

Chapitre 9

Conclusion Générale

Conclusion générale

1. Bilan

Le modèle Cloud Computing propose plus de choix, de flexibilité, d'efficacité opérationnelle et permet aux entreprises comme aux individus de réaliser d'avantage d'économies. Pour profiter pleinement de tous ces bénéfices, les utilisateurs doivent disposer de garanties fiables concernant la confidentialité et la sécurité de leurs données.

L'utilisation et le partage des données collectées dans le Cloud sont limités en raison de la présence d'informations personnelles identifiables dont la confidentialité des individus peut être violée lors d'un tel partage. La difficulté dans le partage des données provient principalement du fait que la préservation de la confidentialité d'un individu résulte une perte d'information, ce qui rend les données moins utiles. Le défi d'assurer la confidentialité d'un individu tout en fournissant des informations utiles rend la préservation de la confidentialité pour les données publiées un domaine difficile.

Dans cette thèse, nos travaux de recherche ont porté sur la conception et l'implémentation d'un protocole d'anonymisation distribué horizontalement qui satisfait les besoins des fournisseurs des données dans un environnement Cloud avec une minimisation de perte d'informations.

Dans le second chapitre, nous avons présenté le paradigme du Cloud Computing, ses caractéristiques, ses modèles de services et de déploiement, les travaux de standardisation, ainsi que les outils d'implémentation et de simulation et les produits commerciaux relatifs aux services du Cloud. L'objectif était de s'assurer que le lecteur aurait une compréhension des concepts de base manipulés dans le reste du manuscrit.

Dans le troisième chapitre, nous avons présenté une généralité sur les deux concepts importants qui constituent un défi de recherche dans un environnement de Cloud Computing à savoir : la sécurité et la confidentialité.

Dans le quatrième chapitre, nous avons cité les principales techniques d'anonymisation, puis nous avons présenté les différents travaux liés à la préservation de la confidentialité selon les deux différents axes : *Le critère de modèle d'attaque* et *le critère d'utilité de données*.

Dans le cinquième chapitre, nous avons dressé un état de l'art sur les différentes approches pour la préservation de la confidentialité pour les bases des données. Ensuite, une présentation des différentes solutions pour la préservation de l'anonymisation pour les bases de données décentralisées est donnée. Enfin, une présentation d'un tableau comparatif est donnée en précisant les avantages et les inconvénients de chaque solution.

Dans le sixième chapitre, nous avons décrit le modèle k-concealment qui garantit une anonymisation des enregistrements sans la nécessité de généraliser toute l'ensemble de la classe d'équivalence. Nous avons montré que ce modèle peut être réalisé avec moins de généralisation par rapport à ce qui est requis par le modèle k-anonymat.

Dans le chapitre sept, nous avons présenté notre protocole d'anonymisation distribué pour la préservation de la confidentialité des données publiées à partir de plusieurs fournisseurs de données dans un environnement de Cloud Computing. Nous avons introduit le nouvel algorithme d'anonymisation distribué en montrant que le processus de Split utilisé dans notre protocole fournit un partitionnement plus efficace que celui utilisé dans les travaux précédents. Nous avons montré aussi, que la généralisation de l'arbre R* utilisée dans notre protocole, choisit de manière récursive la meilleure branche à insérer les objets de données pour recueillir les objets de données les plus proches.

Enfin pour valider nos contributions, nous avons illustré, à travers le chapitre huit, les expérimentations que nous avons entrepris pour évaluer la qualité d'anonymisation fournit par notre protocole. Dans un premier temps, nous avons réalisé une évaluation de l'information perdue causée par notre approche par rapport à l'approche centralisée en utilisation la métrique LM qui semble être la mesure la plus précise par rapport aux autres mesures. Dans un second temps, nous avons montré que les requêtes fournit par la stratégie de l'arbre-R* utilisée dans notre algorithme, sont beaucoup plus précis que celles fournit par la stratégie l'arbre-R.

2. Perspectives

Ce travail ouvre la voie à de nouvelles perspectives intéressantes. Nous soulignons dans la suite, certaines de ces perspectives qui vont contribuer à l'évolution des propositions que nous avons réalisées dans le cadre de cette thèse.

Ces perspectives s'organisent comme suit :

- Nous envisageons d'étendre notre algorithme avec d'autres modèles de confidentialité tel que *t-proximité* qui est un modèle plus sécurisé (assure une sécurité contres d'autres attaques) tout en gardant la minimisation de la perte d'information obtenu par le modèle k-concealment.
- Notre protocole est adapté à la distribution horizontale des bases des données. Nous envisageons de l'étendre pour qu'il prenne en charge aussi les bases des données distribuées verticalement.
- Nous nous intéresserons aussi à étendre notre protocole pour s'adapte avec la contrainte de confidentialité des objets des données pour les bases des données avec des publications dynamiques.
- Le calcul de la quantité d'informations perdue est toujours un problème difficile dans la préservation de la confidentialité. Chacune des métriques existantes utilise des aspects différents qui indiquent une réduction de la perte d'information. Développer une métrique qui prend en considération tous les aspects de la perte d'information simplifiera le processus de calcul de l'information perdu dans le domaine de la préservation de la confidentialité pour les données publiées.
- Enfin, nous planifions d'évaluer notre système avec des données réelles du Cloud.

3. Publications

- **Conférence internationales**

- Kabou salheddine and Sidi mohammed benSlimane, "*A new distributed anonymization protocol to satisfy multiple data providers privacy requirements*". The 3rd International Workshop on Advanced Information Systems for Enterprises (IWAISE'14), 10-12 November 2014, Tunis, Tunisia.
- Kabou salheddine and Sidi mohammed benSlimane, "*K-concealment Based Distributed Anonymization for the Cloud*." The International Conference on Advanced Aspects of Software Engineering (ICAASE'14). November 2 – 4, 2014, Constantine 2 University.
- Kabou salheddine and Sidi mohammed ben Slimane, " *Building virtual anonymized databases for the Cloud*" International conference on cloud computing technologies and applications, CLOUDTECH, Juin 2015, Marrakech, Morocco.

- **Journal international**
  - Kabou salheddine and Sidimohammed benSlimane, "*A New Distributed Anonymisation Protocol with Minimal Loss of Information*," International Journal of Organizational and Collective Intelligence (IJOCI), 7(1), IGI-Global, 2017.

BIBLIOGRAPHIE

[Ardagna et al, 2010] Ardagna. C, S. Jajodia, P. Samarati, and A. Stavrou, "*Providing mobile users' anonymity in hybrid networks,*" in *Proc. of ESORICS,* Athens, Greece, Sep. 2010.

[Amar et al, 2012] Amar.B,P.Raja,Sekhar, Reddy, « *Effective Data Distribution Techniques for Multi-Cloud Storage in cloud computing* », International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622, Vol. 2, Issue 5, September- October 2012, pp.1130-1134

[Ardagna et al, 2011] Ardagna. C, S. De Capitani di Vimercati, S. Paraboschi, E. Pedrini, P. Samarati, et M. Verdicchio, "*Expressive and deployable access control in open web service applications*," IEEE TSC, vol. 4, no. 2, pp. 96–109, Apr-Jun. 2011

[Ardagna et al, 2010] Ardagna. C, S. De Capitani di Vimercati, S. Foresti, S. Paraboschi, and P. Samarati, "*Minimizing disclosure of private information in credentialbased interactions: A graph-based approach,*" in Proc. of PASSAT Minneapolis, MN, USA, Aug. 2010

[Aggarwal et al-a-, 2005] Aggarwal. G, M. Bawa, P. Ganesan, H. Garcia-Molina, K. Kenthapadi, R. Motwani, U. Srivastava, D. Thomas, and Y. Xu, "*Two can keep a secret: A distributed architecture for secure database services*," in *Proc.of CIDR,* Asilomar, CA, USA, Jan. 2005

[Aggarwal et al-b-, 2005] Aggarwal, G., Feder,T,. Kenthapadi, K,.Motwani, R., Panigrahy, R., Thomas, D., and Zhu, A, '*Approximation algorithms for k-anonymity*' Journal of Privacy Technology, 2005.

[Alvin Toffler, 1980] Alvin Toffler, « *The Third Wave* », a book published in 1980

[Abbasy et Shanmugam, 2011] Abbasy.M, B.Shanmugam, « *Enabling Data Hiding for Resource Sharing in Cloud Computing Environments Based on DNA Sequences* », In Proc of IEEE World Congress on Services, 2011

[Aravinth et al, 2013] Aravinth.S, B .Rajkumar, M.Ramkumar, M. Kavipriya, M. MohanaPriya, « *Data Hiding Images Using Spread Spectrum in Cloud*

Computing », In Proceedings of International Journal of Scientific & Engineering Research, Volume 4, Issue 8, August-2013

[Adeel et Guillaume, 2013] Adeel. A, R. Guillaume. "*Anonymizing Sequential Releases under Arbitrary Updates*."Proceedings of the Joint EDBT/ICDT Workshops. Pages 145-154. 2013 Genoa, Italy.

[Bowers et al, 2009] Bowers KD, Juels A, Oprea A. « *Proofs of retrievability: Theory and implementation* ». In: Sion R, ed. Proc. of the 2009 ACM Workshop on Cloud Computing Security, CCSW 2009, Co-Located with the 16th ACM Computer and Communications Security Conf., CCS 2009. New York: Association for Computing Machinery, 2009. 43.54.

[Benedikt et al, 2012] Benedikt. M, P. Bourhis, et C. Ley, "*Querying schemas with access restrictions*," Proc. of VLDB Endowment, vol. 5, no. 7, pp. 634–645, Mar. 2012.

[Barbaro et Zeller, 2006]. Barbaro, M. et Zeller, T. "*A face is exposed for AOL searcher no"*. 4417749. *New York Times* (Aug.9) 2006.

[Bayardo et Agrawal, 2005] Bayardo. R et Agrawal. R, "*Data privacy through optimal k-anonymization"*. In International Conference on Data Engineering (ICDE), pages 217–228, 2005.

[Beckmann et al, 1990] Beckmann, N., Kriegel, H., Schneider, R., Seeger, B, "*The R*-tree: An efficient and robust access method for points and rectangles*". In: Proceedings of the International Conference on Management of Data (SIGMOD'90), NJ pp. 322–331, 1990, Atlantic City

[Chen et Zhao, 2012] Chen.D H.Zhao, « *Data Security and Privacy Protection Issues in Cloud Computing* », in Proc of International Conference on Computer Science and Electronics Engineering, 2012

[Cimato et al, 2008] Cimato. S, M. Gamassi, V. Piuri, R. Sassi, and F. Scotti, "*Privacy-aware biometrics: Design and implementation of a multimodal verification system*," in *Proc. of ACSAC*, Anaheim, CA, USA, Dec. 2008.

[Chen et al, 2005] Chen. W, L. Clarke, J. Kurose, and D. Towsley, "*Optimizing cost sensitive trust-negotiation protocols*," in Proc. of INFOCOM, Miami, FL, USA, Mar. 2005.

[Ciriani et al, 2009] Ciriani. V, S. CapitaniForesti, "*Fragmentation design for efficient query execution over sensitive distributed databases*," in *Proc. of ICDCS*, Montreal, Quebec, Canada, Jun. 2009.

[Ciriani et al, 2012] Ciriani. V, S. Capitani S. Foresti S.Jajodia "*Support for write*

privileges on outsourced data," in Proc. of SEC, Heraklion, Crete, Greece, Jun. 2012.

[Capitani et al, 2007] Capitani. S, di Vimercati, S. Foresti, S. Jajodia, S. Paraboschi, et P. Samarati, "*Over-encryption: Management of access control evolution on outsourced data*," in Proc. of VLDB, Vienna, Austria, Sep. 2007

[Carlisle et al, 2007] Carlisle, D,Rodrian. M, et Diamond.C. "*California inpatient data reporting manual, medical information reporting for California"* (5th Ed), Tech. rep., Office of Statewide Health Planning and Development.2007

[Cox, 1980] Cox, L. H. 1980. "*Suppression methodology and statistical disclosure control"*. J. Am. Statistical Assoc. 75, 370, 377–385.

[Chai et al, 2006] Chai. Z, Z. Cao, and R. Lu, "*Efficient password-based authentication and key exchange scheme preserving user privacy*," in Wireless Algorithms, Systems, and Applications, ser. Lecture Notes in Computer Science, X. Cheng, W. Li, and T. Znati, Eds. Springer Berlin / Heidelberg, 2006, vol. 4138, pp. 467– 477

[Clifton et al, 2003] Clifton, C., Kantarcioglu, M., Vaidya, J, "*Tools for privacy preserving distributed data mining*". ACM SIGKDD Explorations, 2003

[Chang et al, 2016] Chang, V., Kuo, Y. H., & Ramachandran, M. "*Cloud computing adoption framework: A security framework for business clouds*." Future Generation Computer Systems, 57, 24-41. 2016

[Devarieux, 2009] Devarieux. A, *Virtualisation de serveurs Solutions open source*, 2009

[DoD, 2006] DoD, "*National Industrial Security Program Operating Manual*", 5220.22-M, February 28, 2006

[Diaz et al, 2002] Diaz. C, S. Seys, J. Claessens and B. Preneel, "*Towards measuring anonymity*," Proc. 2nd international conference on Privacy enhancing technologies, 2002.

[Dalenius, 1977] Dalenius, T, "*Towards a methodology for statistical disclosure control*". Statistik Tidskrift, 15:429– 444, 1977.

[Ding et al, 2013] Ding, X., Yu, Q., LI, J., Liu, J., Jin, H, (2013) "*Distributed Anonymization for Multiple Data Providers in a Cloud System*", In: W. Meng et al. (Eds.): DASFAA 2013, Part I, LNCS 7825, pp. 346–360.

[Du et Atallah, 2001] Du, W., Atallah, MJ, "*Secure multi-party computation problems and their applications*", a review and open problems. In: NSPW 2001: Proceedings of the 2001 workshop on new security paradigms, pp. 13–22. ACM, 2001, New York

[Friedman et al, 1977] Friedman, J., Bentley, J. and Finkel, R, (1977) "*An algorithm for finding best matches in logarithmic time*". ACM Trans. On Mathematical Software, 3(3), 1977.

[Feing et Zhou, 2008] Feing. L, et Zhou. S, "*Challenging More Updates: Towards Anonymous Re-publication of Fully Dynamic Datasets*," in CoRR. 2008

[Fung et al, 2010] Fung, B., Wang, K., Chen, R, " *Privacy-preserving data publishing*", A survey of recent developments. ACM Computing Surveys, CSUR, 2010.

[Grevet, 2009] Grevet. N, *Le Cloud Computing* : *Réelle révolution ou simple évolution,* Mémoire de recherche, 2009.

[Gehrke, 2006] Gehrke, J. "*Models and methods for privacy-preserving data publishing and analysis"*. Tutorial at the 12th ACM SIGKDD.2006.

[Gionis et Tassa, 2009] Gionis, A. et Tassa, T, (2009), "*k-Anonymization with minimal loss of information*". IEEE Transactions on Knowledge and Data Engineering, 21:206–219. 2009.

[Goldreich, 2001] Goldreich, O, "*Secure multi-party computation*", Working Draft, Version 1.3, 2001

[Guttman, 1984] Guttman, A, "R-*trees: a dynamic index structure for spatial searching*". In: SIGMOD.1984

[Grinshpoun et Tassa, 2014] Grinshpoun. T, Tassa. T, "*A privacy-preserving algorithm for distributed constraint optimization"*. Proceedings of the international conference on Autonomous agents and multi-agent systems. Pages 909-916.Paris, France, May 05 - 09, 2014

[Gionis et al, 2008] Gionis, A., Mazza, A. and Tassa, T, (2008), "*k-Anonymization revisited*". In International Conference on Data Engineering (ICDE), pages 744–753. 2008.

[Hurwitz et al., 2010] Hurwitz, J., Bloor, R., Kaufman, M., et Halper, *Cloud Computing for Dummies*. Wiley, 2010.

[Hüsemann, 2009] Hüsemann, *Systèmes d'information*. Fribourg ,2009

[Hyun et al, 2012] Hyun.Y E. Gelogo J. Kim, « *Secure Data Storage In Cloud Computing* », Journal of Security Engineering, 2012

[He et al, 2011] He. Y, S. Barman, J. Naughton. "*Preventing equivalence attacks in updated, anonymized data"*. ICDE '11 Proceedings of the IEEE 27th International Conference on Data Engineering. Pages 529-540. 2011

[Hoang et al, 2014] Hoang. A, N. Son, M. Tran. « *Detecting Traitors in Re-publishing Updated Datasets"***.** Springer- Verlag Berlin Heidelberg, 2014

[Iwuchukwu et Naughton, 2007] Iwuchukwu, T., Naughton, F, " *K-Anonymization as Spatial Indexing : Toward Scalable and Incremental Anonymization*". VLDB Endowment, ACM 978159593649, 2007, Vienna, Austria..

[Jones et al, 2011] Jones. N, M. Arye, J. Cesareo, et M. Freedman, "Hiding *amongst the clouds: A proposal for cloud-based onion routing*," in *Proc. of FOCI*, San Francisco, CA, USA, Aug. 2011.

[Jeff, 2012] Jeff sedayao, "*Enhancing cloud security using Data Anonymization*", Intel white paper, June 2012

[Jacob et Tassa, 2010] Jacob, J et Tassa, T, "*Efficient Anonymizations with Enhanced Utility"*. TRANSACTIONS ON DATA PRIVACY 3 (2010) 149–175

[Jiang et Clifton, 2006] Jiang, W., Clifton, C, "*A secure distributed framework for achieving k-anonymity*". VLDB Journal 15(4), 316–333, 2006

[Jurczyk et Xiong, 2009] Jurczyk, P., Xiong, L, "*Distributed Anonymization: Achieving Privacy for Both Data Subjects and Data Providers*". In: Gudes, E., Vaidya, J. (eds.) Data and Applications Security 2009. LNCS, vol. 5645, pp. 191–207. Springer, Heidelberg.2009

[Karger et al, 2008] Karger. P, D. Olmedilla, et W.-T. Balke, "*Exploiting preferences for minimal credential disclosure in policy-driven trust negotiation*s," in Proc. of SDM, Auckland, New Zealand, Aug. 2008.

[Kohlmayer et al, 2015] Kohlmayer. F, Prasser. F, Kuhn. A. "*The cost of quality: Implementing generalization and suppression for anonymizing biomedical data with minimal information loss"*. Journal of Biomedical Informatics 37–48. 58 (2015)

[Kohlmayer et al, 2013] Kohlmayer, F., Prasser, F., Eckert, C., Kuhn, A, "*A flexible approach to distributed data anonymization*", In Journal of Biomedical Informatics. 2013.

[Li, 2003] Li. C, "*Computing complete answers to queries in the presence of limited access patterns*," *VLDB Journal*, vol. 12, no. 3, pp. 211–227, Oct. 2003.

[Lee et al, 2008] Lee. A, M. Winslett, J. Basney, and V. Welch, "*The Traust authorization service*," ACM TISSEC, vol. 11, no. 1, pp. 1–3, Feb. 2008.

[Lefevre et al, 2005] Lefevre, K., Dewitt, D.J., etRamakrishnan, R. 2005. Incognito: "*Efficient full-domain k-anonymity"*. In Proceedings of ACM

SIGMOD. ACM, New York, 49–60

[Liu et al, 2015] Liu.X, Q. Xie, L. Wang. "*Personalized extended (α, k)-anonymity model for privacy-preserving data publishing"*. Third International Conference on Advanced Cloud and Big Data, 2015.

[Lue et al, 2013] Lue, Z., Chen, S., Li, Y, (2013) " *A distributed anonymization scheme for privacy preserving recommendation systems*", 978-1-4673-5000-6/13 IEEE, 2013.

[Mahjoub, 2011] Mahjoub. M, « *Étude et expérimentations du Cloud Computing pour le monitoring des applications orientées services* », Mémoire Master, Ecole national de Sfax, Tunisie, 2011.

[Mohammed et Fung, 2010] Mohammed, N., Fung, B, "*Centralized and distributed anonymization for high-dimensional healthcare data*". ACM Trans Knowl Discovery Data;4(4):1–33. 2010.

[Meyerson et Williams, 2004] Meyerson, A. and Williams, R, (2004), "*On the complexity of optimal k-anonymity*". In ACM-SIGMOD Symposium on Principles of Database Systems (PODS), pages 223–228.2004

[Mahajan et Ganar, 2012] Mahajan.B, Ganar.S « *Review Paper on Preserving Confidentiality of Data in Confidentiality of Data in Cloud Using Dynamic Anonymization* » in IJCSN, vol 1, October 2012.

[Machanavajjhala et al, 2006] Machanavajjhala. A, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "*l-diversity: Privacy beyond kanonymity*," in *ICDE*, 2006, p. 24.

[Marcon et al, 2011] Marcon. M N. Santos P. Gummadi, *« NetEx-Cost-effective Bulk Data Transfers for cloud computing* », 2011

[Mather et al, 2009] Mather.T, Subra.K, Latif.S, « *Cloud security and privacy* », Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.,2009

[Mulero et al, 2009] Mulero V, Nin J. « *Privacy and anonymization for very large datasets* ». In: Chen P, ed. Proc of the ACM 18th Int'l Conf. On Information and Knowledge Management, CIKM 2009. New York: Association for Computing Machinery, 2009. 2117.2118.

[Ninghui et al, 2007] Ninghui Li Tiancheng Li, Suresh Venkatasubramanian, "*t-Closeness: Privacy Beyond k-Anonymity and l-Diversity*", 2007

[Ottcher et Obermeier, 2008] Ottcher, B., Obermeier, S, (2008) "*Secure set union and bag union computation for guar-anteeing anonymity of distrustful participants*", JSW 3(1), 9–17.2008.

[Pacheco et Puttini, 2012] Pacheco. V et R. Puttini. "Defining *and Implementing Connection AnonymityforSaaS Web Services"*. IEEE Fifth International Conference on Cloud Computing, 2012

[Peter et Tim, 2011] Peter Mell, et Tim Grance, «*The NIST Definition of Cloud Computing*, » Version 15, 2011, http://www.wheresmyserver.co.nz/storage/media/faq-files/cloud-def-v15.pdf.

[Philippe Hedde, 2010] Philippe Hedde, Syntec numérique *Livre blan sécurité de Cloud Computing* Analyse des risques, réponses et bonnes pratiques, 2010.

[Porwal et al, 2012] Porwal. A, R. Maheshwari G. Kakhani « *An Approach for Secure Data Transmission in Private Cloud* », International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-1, March 2012

[Pei et al, 2007] Pei. J, Xu. J, Z. Wang, W. wang, and K. Wang, "*Maintaining k-anonymity against incremental updates*," in SSDBM. 2007.

[Roy et al, 2010] Roy I, Ramadan HE, Setty STV, Kilzer A, Shmatikov V, Witchel E, « *Airavat: Security and privacy for MapReduce* » In: Castro M, eds. Proc of the 7th UsenixSymp. on Networked Systems Design and Implementation. San Jose: USENIX Association, 2010. 297.312.

[Rabl et al, 2011] Rabl. T,M.Frank H. Sergieh H. Kosch « *A Data generator for cloud sclale Benchmarking* », LNCS 6417, pp. 41–56, 2011

[Randike et al, 2011] Randike. G, Renato Iannella, and Tony Sahama, « *Sharing with Care An Information Accountability Perspective* », Internet Computing, IEEE, vol. 15, pp. 31-38, July-Aug. 2011.

[Richard et al, 2006] Richard Kissel, Matthew Scholl, Steven Skolochenko, Xing Li, "*Guidelines for Media Sanitization*," NIST Special Publication 800-88, September 2006,

[Rajalekshmi et Lashma, 2015] Rajalekshmi. V et K. Lashma. " *Anonymous Authentication and Access Control of Data Stored in Multi-Clouds"*. International Journal of Innovative Research in Computer and Communication Engineering. Vol. 3, Issue 11, November 2015

[Raumond et al, 2015] Raumond. H, V. Luke, L. Peggy, A. Pacheco, P. Harris, C. Denny, A. Malin. "*A multi-institution evaluation of clinical profile*

anonymization". Journal of the American Medical Informatics Association, 2015

[Sandhu et Samarati, 1997] Sandhu. R, et P. Samarati, "*Authentication, access control and intrusion detection*," in CRC Handbook of Computer Science and Engineering, A. Tucker, Ed. CRC Press Inc., 1997, pp. 1929–1948

[Samarati et al, 2010] Samarati. P , S. De Capitani di Vimercati, "*Data protection in outsourcing scenarios: Issues and directions*," in Proc. of ASIACCS, China, Apr. 2010

[Seethal, 2013] Seethal K S1, Siddana Gowda**. "***A Secure and Efficient Way of Accessing Encrypted Cloud Databases Using Adaptive Encryption Scheme"*. International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064. 2013

[Sakshi et Bamnote, 2015] Sakshi. D et R.Bamnote. "*An Efficient Approach to Encrypted Cloud Database"*. International Journal of Innovative Research in Computer and Communication Engineering. Vol. 3, Issue 5, May 2015

[Sweeney-C-, 2002] Sweeney, "*Achieving k-anonymity privacy protection using generalization and suppression"*. Int. J. Uncertainty, Fuzziness, Knowl.-Based Syst. 10, 5, 571–588. 2002

[Sushmita et al, 2012] SushmitaRuj, Milos Stojmenovic, AmiyaNayak, "*Privacy Preserving Access Control with Authentication for Securing Data in Clouds*", 12th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, pp. 556 – 563, 2012.

[Samarati, 2001] Samarati. P, "*Protecting Respondent's Privacy in Microdata Release*," IEEE Trans. Knowledge and Data Eng., vol. 13, no. 6, pp. 1010-1027, Nov./Dec. 2001.

[Sweeney-A-, 2002] Sweeney. "*Achieving k-anonymity privacy protection using generalization and suppression"*. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), 2002; 571- 588.

[Stammler et al, 2016] Stammler.S, S. Katzenbeisser, H. Hamacher. "*Correcting Finite Sampling Issues in Entropy l-diversity"*. International Conference in Privacy in Statistical Databases. Croatia, September 14–16, 2016

[Sunyong et al, 2012] Sunyong. Y, Shin. M, D. Lee. "*An Approach to Reducing Information Loss and Achieving Diversity of Sensitive Attributes in k-anonymity Methods"*. Journal of medical research. Nov 2012

[Sweeney-B-, 2002] Sweeney-B-. "*K-Anonymity: A model for protecting privacy"*.

International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10(5):557–570, 2002.

[Tinbo et al, 2013] Tinbo. R, "*A Modified Anonymisation Algorithm Towards Reducing Information Loss"*. PhD thesis, School of Computing Dublin Institute of Technology, Ireland, Juin 2013

[Tassa et al, 2012] Tassa, T., Mazza, A., Gionis, A, (2012), "k-*Concealment: An Alternative Model of k-Type Anonymity*", Transactions on Data Privacy 5, pp189–222. 2012.

[Vimercati et al, 2012] Vimercati. V S. Foresti P. Samarati, « *Managing and Accessing Data in the Cloud : Privacy Risks and Approaches* », IEEE, 2012

[Vallet, 2012] Vallet. L, « *Contribution au renforcement de la protection de la vie privée : Application à l'édition collaborative et anonyme des documents* ». Thèse de Doctorat, Université d'Aix-Marseille. Septembre 2012

[Vinterbo, 2004] Vinterbo. S, "*Privacy: a machine learning view"*. IEEE Transactions on Knowledge and Data Engineering, 16:939–948, 2004.

[Vaidya et Clifton, 2004] Vaidya, J., Clifton, C, "*Privacy-preserving data mining*": Why, how, and when. IEEE Security & Privacy 2(6), 19–27. 2004

[Wang et al, 2009] Wang. C, Ren N. Cao W. Lou, « Towards *Secure and Dependable Storage services in cloud computing* », 17th IEEE International Workshop on Quality of Service (IWQoS'09), 2009.

[Warin, 2011] Warin, S. (2011, Février 07). *Un livre blanc sur le cloud computing.* Consulté le Novembre 01, 2011,

[Xiao et Tao, 2007] Xiao. X et Tao. Y. "M-invariance: *Towards privacy preserving republication of dynamic datasets*." In *Proc. of SIGMOD*, 2007

[Yao et al, 2008] Yao. D, K. Frikken, M. Atallah, and R. Tamassia, "*Private information: To reveal or not to reveal*," ACM TISSEC, vol. 12, no. 1, pp. 1–27, Oct. 2008

[Zhang et al, 2007] Zhang. O, N.Koudas, D.Srivastava, T.Yu « *Aggregate Query Answering on Anonymized Tables* », In Proceedings of the 23rd IEEE International Conference on Data Engineering, 2007

[Zhang et Bi, 2010] Zhang. X et Bi. H, "*Secure and Effective Anonymization against Re-publication of Dynamic Datasets*", IEEE 2010

GLOSSAIRE DES TERMES

- **Anonymisation** : Est le processus qui garantit que les informations d'un individu restent non identifiées dans l'ensemble des données.
- **Donnée** : Une donnée est le résultat direct d'une mesure. Elle peut être collectée par un outil de supervision, par une personne ou être déjà présentée dans une base de données. Une donnée seule ne permet pas de prendre une décision sur une action à lancer
- **Base des données** : Est une collection de données organisées de façon à être facilement accessibles, administrées et mises à jour.
- **Éditeur des données** : Est l'individu ou la personne morale qui fait l'opération du contrôle et il est le responsable de la conservation et de l'utilisation des informations personnels.
- **Le destinataire** : Tout utilisateur de données ou informations produites par l'éditeur de données.
- **L'objet des données** : Est un individu qui est considéré comme l'objet de la donnée, par exemple, un patient admis à l'hôpital.
- **Classe d'équivalence** : Le nombre d'enregistrements qui ont les mêmes attributs quasi-identiques.
- **Information** : Les informations sont des données qui ont été traitées ou analysées pour produire quelque chose d'utile.
- **Information perdue** : Est la perte d'information due au processus d'identification.
- **La confidentialité** : le droit de protéger Toute information relative à un individu identifié
- **La préservation de la confidentialité pour les données publiées PCDP** : Est l'un des domaines de la recherche pour la préservation de la confidentialité, qui s'articule sur la protection de l'identité d'un individu lors d'un tel partage
- **Quasi-identifiants** : Les attributs qui peuvent être liés à d'autres informations pour identifier un individu, ex : date de naissance

- **Attributs sensibles** : Les attributs qui représentent l'information sensible d'un individu, ex : Salaire, maladie.