REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université Mohamed Khider – BISKRA
Faculté des Sciences Exactes, des Sciences de la Nature et de la Vie
Département d'informatique

N° d'ordre : 03/M2/2019

Mémoire

Présenté pour obtenir le diplôme de master académique en

Informatique

Parcours : **RTIC**

Conception et réalisation d'un système de protection et d'assurance de vie privée sur les Bigdata

---

Par :

Leksouri Maria

Soutenu le 06 juillet 2019, devant le jury composé de :

Benameur Sabrina	DR	Président
Merizig Abdelhak	DR	Rapporteur
Saouli Hamza	DR	Examinateur

Remerciements

Au terme de ce travail, je tiens à remercier Dieu le tout puissant de m'avoir donné le courage, la volonté et la patience pour achever ce travail.

Que nos chers parents et familles, trouvent ici l'expression de nos remerciements les plus sincères et les plus profonds en reconnaissance de leurs sacrifices, aides, soutien et encouragement.

J'ai l'honneur et le plaisir de présenter ma profonde gratitude et mes sincères remerciements à mon encadreur Dr. SAOULI Hamza, pour ses précieuses aides, ces orientations et le temps qu'il m'a accordé pour mon encadrement.

Je remercie profondément tous les enseignants qui m'ont encouragé et soutenu pendant mon cursus.

Je remercie également les jurys Dr. Benameur Sabrina et Dr. Merizig Abdelhak qui ont accepté de juger notre travail.

Je remercie aussi tous ceux qui ont contribué de prêt ou de loin à la réalisation de mon mémoire

Dédicace

Je dédie ce modeste travail à :

A ma très chère mère Aicha, aucun hommage ne pourrait être à la hauteur de l'amour Dont ils ne cessent de me combler. Que Dieu lui donne une bonne santé et longue vie.

A mon cher père Ridha qui est mort et j'espère que Dieu me rassemblera au paradis.

A mon frère Mehdi Qui m'a soutenu tout au long de cette période et mes sœurs Cherifa, serine et Saida, sans oublié ma grand-mère Cherifa et ma grand-père Amara, toute ma famille et mes amis.

En fin à mon futur mari Abderraouf Qui m'a beaucoup aidé et m'a encouragé.

Je vous dis merci.

Résumé

Le Big Data se définit par les technologies et méthodes utilisées pour récolter, stocker et Analyser un grand volume de données issues de multiples ressources. Ces données peuvent être les informations que les internautes laissent sur le Web ou les objets connectés, aussi les données internes à l'entreprise ou encore des informations générales, L'objectif du Big Data est de réussir à corréler ces données entre elles, en temps réel, pour en tirer des conclusions d'analyse et prendre les décisions adéquates. D'une part, c'est un atout important pour les organisations professionnelles et les gouvernements pour la prise de décision d'autre part L'analyse de ces données Exige la confidentialité et l'anonymisation des données pour protéger la vie privée des l'informations sensibles et pour assuré la sécurité contre les hackers.

Ce travail vise à protégé et assuré la vie privée du big data. Pour cela nous avons proposé un nouveau système qui comprend divers composants, en prenant en compte les différents critères de sécurité et les caractéristiques des Big Data.

Afin de montrer la faisabilité de système proposée, nous avons développé un prototype qui pourra résoudre les problèmes mentionnée ci-dessus.

Mots clés : La protection de la vie privée, Big Data, confidentialité, anonymisation

Abstract

Big Data is defined by the technologies and methods used to harvest, store and analyze a large volume of data from multiple resources. This data can be the information that Internet users leave on the Web or the connected objects, as well as the data internal to the company or general information. The goal of Big Data is to successfully correlate these data with each other, in time. Real, to draw analytical conclusions and make the right decisions. On the one hand, it is an important asset for professional organizations and governments for decision-making on the other hand Analysis of these data Demands confidentiality and anonymization of data to protect the privacy of individual sensitive information and for assured security against hackers.

This work aims to protect and ensure the privacy of big data. For this we proposed a new system that includes various components, taking into account the different security criteria and characteristics of Big Data.

In order to show the proposed system feasibility, we have developed a prototype that will solve the problems mentioned above.

Keywords: Privacy, Big Data, Privacy, Anonymization.

Table des matières

Chapitre 3 Approches et travaux Connexes

Chapitre 4 : Conception et modélisation

Chapitre 5 : Implémentation du système

Table des figures

Chapitre 1 : Introduction générale

Chapitre 2 : Big Data & vie privée

Chapitre 3 Approches et travaux Connexes

Chapitre 4 : Conception et modélisation

Chapitre 5 : Implémentation du système

Chapitre 1

Introduction Générale

1.1 Contexte du travail

Nous vivons aujourd'hui dans une révolution numérique globale, alimentée par l'essor d'exploitation des Big Data. Ce sont nos données personnelles, le surgissement d'informations qui peuvent devenir le carburant de nouvelles révolutions industrielles, qui va profondément modifier nos modes de vie et qui a déjà commencé. Le traitement massif des données personnelles est utilisé dans tous les secteurs tels que le sport, le commerce, la politique afin de mettre en équation nos goûts, nos comportements et même nos désirs. « 90 % de l'ensemble des données du monde ont été crées ces deux dernières années» [Ralph Jacobson, 2013]. Donc l'enjeu de Big Data c'est de savoir collecter les données, les stockées, les analysées et ensuite les visualisées.

Les organisations sont aujourd'hui à un tournant dans la gestion des données. Nous sommes passés de l'ère où la technologie était conçue pour répondre à un besoin métier spécifique, comme la détermination du nombre d'articles vendus à combien de clients, à un moment où les entreprises disposent de plus de données, le traitement de toutes ces données est en train de changer d'échelle. Succès des médias sociaux, développement des objets connectés et des capteurs intelligents, dématérialisation de plus en plus poussée des échanges: tous ces phénomènes multiplient les sources de données potentiellement exploitables, générant, dans certains cas, des données à haute vélocité, c'est-à-dire qui se renouvellent très rapidement.

Face à cette masse de données une multitude de défis sont mis enjeux pour les Big Data, le plus complexe est celui concernant la vie privée des utilisateurs. Lorsque des données sensibles personnelles sont publiées et / ou analysées, une question importante à considérer est de savoir si cela peut violer le droit de la vie privée des individus. Les données humaines peuvent potentiellement révéler de nombreuses facettes de la vie privée d'une personne, mais un niveau de danger plus élevé est atteint si les différentes formes de données peuvent être reliées entre elles. Il est évident que le maintien du contrôle sur les données personnelles garantissant la protection de la vie privée est de plus en plus difficile et ne peut pas simplement être accompli.

1.2 Problématique et objectifs

Avec l'explosion en volume et en variété de données, Big Data est devenu un sujet brûlant, en revanche les risques de violation de la vie privée croissent en corrélation avec la quantité massive des données. De ce fait la préservation de la vie privée est l'une des plus grandes préoccupations et comment trouver une solution pour ce problème devenu un défi de taille.

Afin de protéger la vie privée des individus, il est nécessaire que les données doivent être correctement anonymisées avant la publication. Un tel anonymat doit non seulement satisfaire aux exigences de la vie privée, mais également préserver l'utilité des données. Sinon, il serait difficile d'extraire des informations utiles des données anonymisées.

Notre objectif est de :

- ✓ Etudier les concepts généraux de Big Data et la vie privée
- ✓ Présenter un état de l'art sur les approches et les travaux connexes.
- ✓ Construire un tableau comparatif entre les approches.
- ✓ Proposer un système de protection et d'assurance de la vie privée sur les Big Data.

1.3 Structure du mémoire

Hormis l'introduction et la conclusion générale qui sont le premier et le dernier chapitre Respectivement, ce mémoire est composé de quatre autres chapitres organisés comme suit :

Le deuxième chapitre « Big Data & Vie Privée» : Ce chapitre est consacré à l'étude des caractéristiques des Big Data pour mieux comprendre les concepts de base de cette technologie. Il comporte plusieurs notions fondamentales, nous présentons dans la première partie les notions des Big Data, les méthodes de traitement, les domaines d'application et les défis et enjeux. Dans la deuxième partie, nous aborderons les notions de vie privée.
Le troisième chapitre « Approches et travaux connexes» : Cette partie est attribuée à L'étude de plusieurs travaux qui proposent des solutions pour la protection et l'assurance de la vie privée Sur les Big Data, à partir desquels nous avons construit une table Comparative.
Le quatrième chapitre « Conception du système» : Ce chapitre présente une conception De notre système. On présentera notre architecture pour la protection de la vie privée sur les Big Data.
Le cinquième chapitre « Implémentation du système» : Cette partie consiste à présenter L'environnement logiciel sur lequel le système sera réalisé et validé, et ainsi que les détails D'implémentation de notre application. On donnera par la suite une description textuelle et Graphique de quelques interfaces du système réalisé.

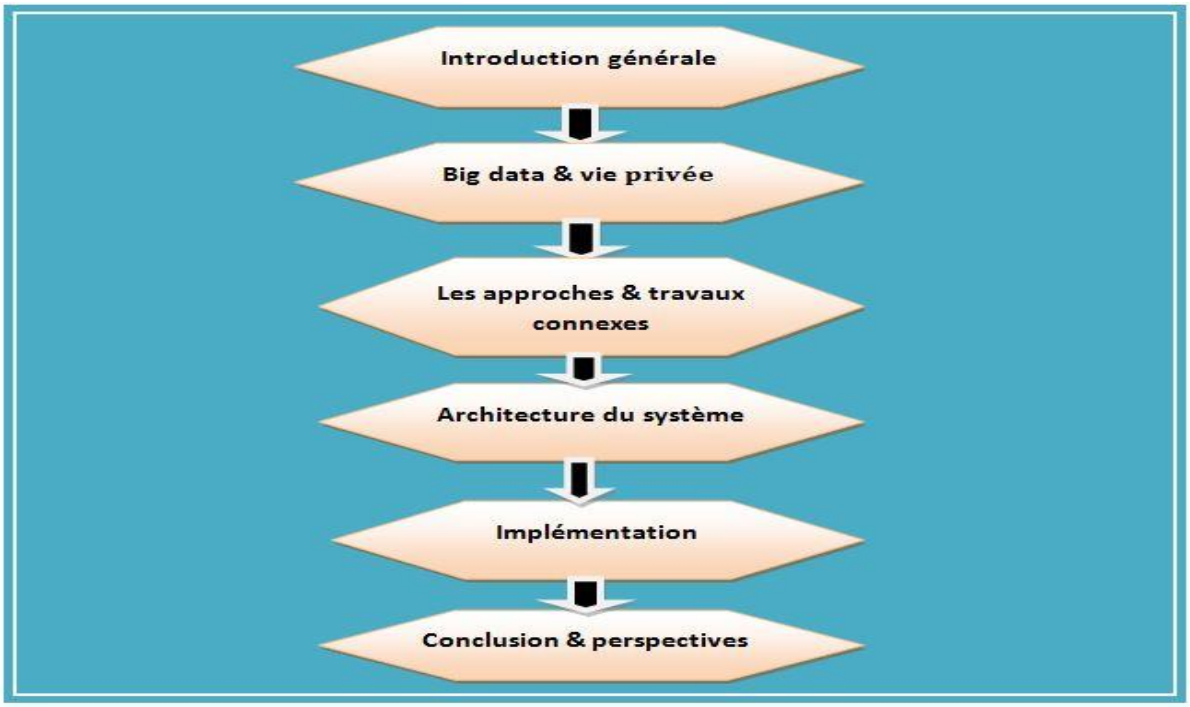


Figure 1.1 : structure du mémoire

Chapitre 2

Big Data & Vie Privée

2.1 Introduction

Le monde qui nous entoure devient de plus en plus numérique, des données sont maintenant générées dans tous les domaines de notre vie. La vitesse à laquelle nous produisons des données augmente régulièrement, créant ainsi des flux encore plus importants de données en constante évolution.

Ce chapitre fournit une compréhension fondamentale du Big Data, nous allons définir les termes, le vocabulaire et quelques propriétés du Big Data, nécessaires à la bonne compréhension des chapitres suivants, ensuite nous allons répondre aux questions qui se posent autour de la notion vie privée.

2.2 Big data

2.2.1 Émergence de Big data

Big data est considéré comme le dernier cri en matière de High Tech, il a résulté de la révolution de la technologie de l'information. Il a certainement une grande influence sur le présent et l'avenir à cause de son importance et du large spectre d'application dans le domaine de la communication et de traitement de données et la recherche scientifique.

La quantité d'information traitée par le micro-processeur se voit doublé chaque 18 mois environ selon la loi de *Moore.*

Il convient également de signaler que la capacité de stockage des données sur le disque dur évolue plus rapidement que la capacité de traitement de données sur le micro-processeur (loi de *Kryder).*

Le volume de données stocké est en croissance exponentielle. La capture, l'intégration et l'exposition de l'information était assez difficile. Les données non structurées prennent une importance considérable, cela est dû à l'évolution des réseaux sociaux et à leurs exigences.

Exemple : chaque appel téléphonique exige une masse de données ayant rapport à le lieu de l'abonné, son numéro, le temps de communication et le tarif unitaire. Un opérateur de télécommunications typique générera quelques téraoctets de données détaillées d'appels chaque mois. Nous constatons également une amélioration considérable dans le traitement du son (flac), de la vidéo (mkv) qui exigent plus de mémoire de stockage et de vitesse de traitement.

Cette masse énorme de données doit être stockée, et le cas échéant analysée et exploitée correctement pour répondre aux besoins des différents opérateurs industriels, économiques ou sociaux. En bref c'est l'ère de l'information.

Le Big data est une technologie révolutionnaire qui peut provoquer des perturbations aux niveaux économique, scientifique et culturelle. Cela est dû à l'importance des changements et des améliorations qu'il impose dans ces différents domaines et qui exigent une nouvelle réadaptation. [1]

2.2.2 Définition de Big data

Nombreuses définitions ont été proposées pour décrire le Big Data mais toutes ont pratiquement le même sens. En mai 2011 MGI (The McKinsey Global Institute) définit le Big Data comme «l'ensemble de données dont la taille dépasse la capacité des logiciels de base de données traditionnelles à capturer, stocker, gérer et analyser» [2]. Dans la même année un rapport a été annoncé par IDC (International Data Corporation) qui définit le Big Data comme suit «Les technologies qui décrivent une nouvelle génération de technologies et d'architectures, conçues pour extraire économiquement la valeur à partir de très grands volumes et d'une grande variété de données, en permettant une très grande vitesse de capture, une découverte et/ou une analyse» [3].

Le terme «Big data» se réfère à des ensembles de données numériques structurées ou non structurées qui dépassent la capacité des outils traditionnels de traitement et d'analyse des données pour les manipuler.

Big data est défini par les trois V :

- ✓ Haut Volume.
- ✓ Haute Variété.
- ✓ Haute Vitesse.

Big data exigent des formes innovantes et rentables de traitement de l'information pour une meilleure compréhension et une prise de décision correcte.

2.2.3 Modèle 5V

Big data peut être décrit en utilisant le modèle 5V illustré sur la figure 1. Ce modèle est une extension du modèle 3V précédemment cité, et comprend :

- ***Volume*** : il est d'une grandeur exceptionnelle par rapport aux normes connues. Les données produites sont de l'ordre de Zettabytes, et elles sont en croissance d'environ 40% chaque année.

- ***Vitesse*** **:** (the era of streaming data) la collecte et l'analyse des données doivent être rapides et en temps opportun afin de maximiser l'utilisation de la valeur commerciale du Big data.
- ***Variété*** : Les données sont sous plusieurs formats et types, elles comprennent des données semi-structurées et non structurées telles que l'audio, la vidéo, texte ainsi que les données structurées traditionnelles. La plupart des données existantes sont non-structurées ou semi-structurées.
- ***Valeur*** : les données sont devenues une marchandise qu'on peut vendre à des tiers pour une exploitation de nature commerciale, économique ou sociale. La maitrise et l'analyse correcte de ces données permettent une prise de décisions adéquates.
- ***Véracité*** : pour assurer l'exactitude et l'efficacité de cette masse de données, il est indispensable et même vital de procéder à son nettoyage de tout bruit qui peut générer des erreurs qui influent négativement sur les prises de décision. Des techniques sont utilisées pour atteindre cet objectif. [4]

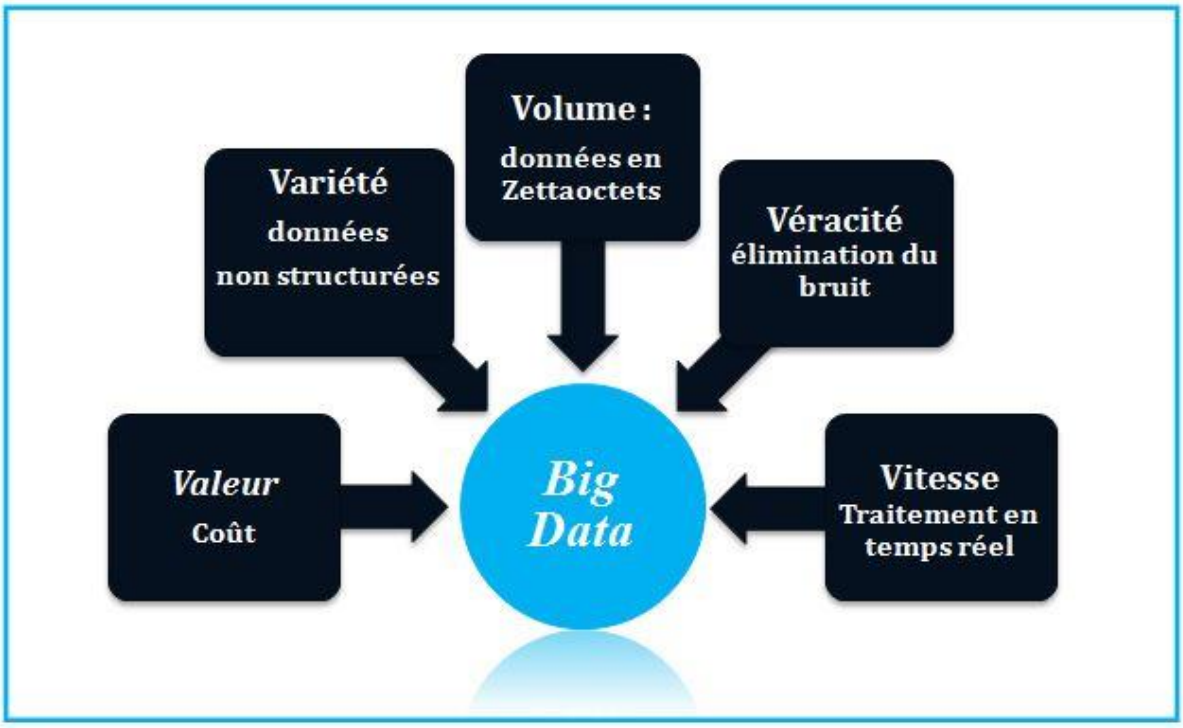


Figure 2.1 : le modèle 5V qui définit Big data.

2.2.4 Concepts de Big data

Dans cette section, nous allons examiner de plus près les concepts techniques communs et modèles généralement utilisés dans la plupart des principaux outils et plateformes Big data

2.2.4.1Cluster de Big data

Les concepts des Clusters Big data peuvent être divisés en deux grandes catégories :

1) Configuration et topologie des clusters.
2) Déploiement des clusters.

La première porte sur le modèle logique de la façon dont un cluster Big data est divisé selon les différents types de nœuds.

La seconde traite du déploiement réel de ces nœuds dans l'infrastructure matérielle physique.

2.2.4.1.1 Cluster Configuration et Topologie

Un cluster Big data est logiquement divisé en deux types de machines / nœuds, à savoir les nœuds de données et les nœuds de gestion.

Les nœuds de données servent deux objectifs fondamentaux d'une part, le stockage des données de manière distribuée et d'autre part le traitement secondaire pour la transformation et l'accès. Les nœuds de gestion servent de façade pour les applications clientes pour l'exécution des cas d'utilisation.

2.2.4.1.2 Déploiements des Clusters

L'utilisation du Big data tourne en grande partie autour de l'accès / stockage d'un très grand volume de données. L'accès / stockage de données depuis / vers le disque est le processus le plus lent dans l'exécution d'une tâche dans une machine ou un cluster.

2.2.4.2 Concept de stockage de Big data

Le stockage de données constitue une importance capitale et se situe au cœur des outils et des plateformes Big data. Cette masse immense de données doit être mémorisée d'une manière appropriée pour permettre son analyse et son exploitation efficacement.

Le concept de stockage se résume comme suit :

2.2.4.2.1 Modèles de données

Il existe plusieurs types de modelés de données, tels que le modèle relationnel, NoSQL.

2.2.4.2.2 Partitionnement de données

Le partitionnement des données doit se faire sur plusieurs nœuds de données afin de pouvoir traiter ces données en simultanée par toutes ces machines.

2.2.4.2.3La réplication de données

La réplication des données permet la protection de ces données en cas de défaillance d'un serveur et une plus large utilisation.

2.2.4.2.4 La compression de données

La compression nous permette de réduire l'espace de stockage et limite la bande passante nécessaire sur le réseau de transport. Toutefois elle présente un inconvénient en matière de temps de traitement de ces données suite à la compression et à la décompression qui s'imposent lors de l'écriture et de la lecture de ces données vers et à partir du disque de stockage.

2.2.4.2.5Indexation de données

Les données sont reparties sur plusieurs blocs à travers différents nœuds de données du cluster Big data. L'indexation sert à identifier les enregistrements réalisés dans ces blocs et déterminera leurs positions.

2.2.4.3 Concept de récupération informatique du Big data

Il nous permet le traitement d'un grand nombre de données en temps réel ainsi que l'accès à ces données d'une manière aléatoire pour la lecture et l'écriture.

2.2.4.3.1 Moteur de traitement distribué

Les moteurs de traitement distribué répondent à la nécessité pour la gestion, le traitement, le filtrage, l'interrogation, la modélisation, l'exportation, ainsi que l'archivage de grand volume de données dans l'infrastructure Big data.

2.2.4.3.2.Sécurité des données

La sécurité présente une dimension vitale pour tout ensemble de données informatiques. Différentes méthodes sont appliquées pour réaliser cet objectif tel que le chiffrement, l'authentification et le cryptage.

2.2.4.4 La gestion des ressources

La gestion des ressources présente une grande importance pour technologie Big data. Ces ressources informatiques (CPU, Mémoires et Disques) de tous les nœuds et du réseau les reliant doivent être distribuées de manière appropriée. [5]

2.2.5 Résistance de Big data (Souplesse et maniabilité)

Le tremblement de terre qui frappa le Japon et qui fut suivi d'un Tsunami d'une ampleur sans précèdent puis de l'explosion nucléaire qui a eu lieu à Fukushima n ° 1 et des évènements qui suivirent (exode de la population, destruction d'une masse importante

d'archives gouvernementale…etc.) a démontré l'utilité de la technologie Big data et le rôle que cette dernière pourrait jouer en sécurisant les données et en offrant des renseignements sur l'état des routes, l'intensité du trafic etc. Qui pourraient sauver des vies humaines et limiter relativement les dégâts. De nouvelles techniques permettant la préservation de données même en cas de catastrophe naturelles de grandes ampleurs sont apparues de nos jours telle que la technologie Cloud. [6]

2.2.6 Domaine d'application de Big data

Dans cette section nous présentons quelques principaux domaines d'applications du Big data

2.2.6.1Agriculture

D'ici 2050 on prévoie le dépassement de 9 milliards d'êtres humains sur le globe, ce qui rend l'agriculture un domaine prioritaire pour gérer les besoins alimentaires de la population mondiale. Le Big data représente un atout considérable pour l'organisation de l'agriculture à travers le monde, notamment pour la gestion de l'irrigation (l'eau potable étant une ressources de plus en plus rare), où nous avons besoin de gérer de gigantesque masses de données qui concernent les prédictions météorologique et la sécheresse du sol.

2.2.6.2 Assurance

L'assurance représente l'un des domaines directe d'application de Big data, vue qu'on est amené à effectuer des statistiques et des analyses sur les risques liés au comportement de millions d'individus.

La possibilité de récolter de gigantesques masses d'informations qui concernent la vie des individus permet de concevoir un modèle de vie pour chacun d'eux : hygiène de vie, conduite de voiture, amende, consommation électrique, relation professionnelle ….Etc. Ces modèles de vie permettent aux agences d'assurances d'améliorer leurs offres, d'optimiser leurs méthodes, et même de mener des enquêtes plus précises.

2.2.6.3 Marketing

Avec le marketing on est amené à gérer de gigantesques masses d'informations qui proviennent de divers sites et réseaux sociaux que des clients potentiels peuvent visiter. Mais ce qui révolutionne vraiment le marketing de nos jours c'est l'omniprésence de capteurs publics sur les centres commerciaux, métros, aéroports et universités, et qui sont destinés à capter le comportement des consommateurs, ce qu'il achète, ce à quoi ils s'intéressent, et les produits qu'ils ne trouvent pas aux marchés, ce qui permet d'analyser et d'étudier leurs

besoins en temps réel afin de produire des solutions et des méthodes de marketing plus efficaces.

L'utilisation des capteurs permet de capter des données de diverses formes : images de visages pour analyse émotionnelle, vidéos pour description comportementale, données textuelles pour décrire la nature des produits achetés, données numériques et statistiques. Cette diversité qui nécessite un traitement en temps réel ne peut être résolue qu'avec des méthodes de stockage et de traitement d'informations issues de Big data.

2.2.6.4 Au-delà du marketing

Le Big data a permis de refaçonner le monde du marketing en offrant les techniques et les stratégies nécessaires pour bénéficier des données que publient les consommateurs et fournisseurs en utilisant les : Réseau sociaux, applications mobiles, magasins, TV, catalogues, blog, presse, radios, etc. Sans la techniques Big data il sera tout simplement impossible de traiter les gigantesques masses d'informations que produisent ces moyens de publication. L'émergence de Big data a permis l'apparition de nouvelles notions telle que le pré-marketing et re-marketing qui représentant une nouvelle vision d'atteindre et de convaincre les consommateurs finaux.

2.2.6.5 Achat programmatique

L'achat programmatique est devenu la technique la plus utilisée pour l'achat/vente sur Internet, vue que cette technique permet d'utiliser un logiciel ou une plateforme intermédiaire entre les clients et les fournisseurs pour effectuer des opérations de : publicité, choix du meilleur prix, et payement électronique. L'achat programmatique permet d'alléger les tâches qui correspondent au processus d'achat/vente en s'occupant automatiquement du processus de négociation entre client et fournisseur ainsi que de toute opération manuelle traditionnellement demandée par le fournisseur. Cependant, l'achat programmatique impose la manipulation en temps réel de gigantesques masses d'informations qui sont échangées entre clients et fournisseurs en compétition pour trouver et acheter les meilleurs espaces publicitaires sur le Net. Les techniques de gestion des données issues du domaine Big data représentent un atout considérables et une alternative prometteuse pour la gestion des plateformes d'achat/vente intermédiaire.

2.2.6.6 Compétitivité et Innovation de produit

La possibilité de traiter de gigantesques masses d'informations en temps réel permet aux entreprises d'analyser les besoins de leurs clients afin de pouvoir optimiser et améliorer leurs propres produits et augmenter leur compétitivité sur le marché. C'est ainsi, que les

services qu'offrent les fournisseurs de téléphonie mobile permettent aux touristiques de localiser, en temps réel, leurs clients habituels afin de leurs envoyer des offres d'excursions, les lieux et la nature des évènements touristiques, et les réductions hôtelières et les billets d'avion par exemple. Les techniques d'analyse en temps réel de gigantesques masses d'informations, issues de Big data, permettent également aux entreprises de contrôler et d'être à jours par rapport aux produits des entreprises concurrentes ce qui garantit l'innovation et la compétitivité des produits.

2.2.6.7 Gestion de catastrophes naturelles

L'une des applications les plus intéressantes de Big data, est la possibilité d'analyser des données météorologiques en temps réel, ce traitement permet de suivre et de visualiser le déplacement des ouragans et de prédire les endroits géographiques où ces derniers vont frapper. C'est ainsi que les gouvernements locaux et les organisations internationales d'assistance humanitaire peuvent préparer les ressources nécessaires (couverture, alimentations, médicaments) ainsi que les moyens de transport et d'intervention rapide pour aider la population en détresse.

2.2.6.8 Contrôle d'épidémies

Le Big data peut contribuer à contrôler la propagation d'épidémies à travers le monde en surveillant par exemple la migration des insectes porteurs de maladies à travers le globe. Le big data est également utilisé pour traquer la population des rats dans les grandes villes telles que New-York ou Chicago où la police locale utilise un système Big data pour la surveillance visuelle et l'analyse des itinéraires des rats, afin de contrôler leurs croissances.

2.2.6.9 Prévention d'attaques cybernétiques

De nos jours, les techniques d'analyse de données qu'offre le Big data sont devenues incontournables pour pouvoir détecter les intrusions, les failles sécuritaires ainsi que les attaques cybernétiques, vue que le volume de données transportées sur le Net est devenu gigantesque, diversifier, et nécessitant un traitement en temps réel. Avec les techniques de traitement de données Big data on arrive à tracer le schéma relationnel entre les données et effectuer des calculs statistiques qui permettent de surveiller et d'intervenir, en temps réel, sur les menaces et les attaques cybernétiques à l'échelle mondiale. [7]

Figure 2.2 : Domaine d'application de big data.

2.2.7 Les méthodes de traitement des Big Data

Il existe de nombreuses méthodes différentes pour le traitement des grandes données on mentionne que la plupart de ces méthodes sont interconnectées et utilisées simultanément pendant le traitement, ce qui favorise l'utilisation du système, dans cette section nous allons évoquer les principales méthodes de traitement de données.

- **Méthode d'optimisation :** Ce sont les outils mathématiques qui s'appuient sur l'analyse numérique axée sur la résolution de problèmes dans divers défis du Big Data (volume, vitesse, variété et véracité), afin d'améliorer les performances du système par la recherche de l'ensemble optimal des actions nécessaires. Elle utilise certaines techniques d'analyse comme : la programmation génétique et évolutive.
- **Méthode statistique :** Elle permet de collecter, organiser et interpréter les données pour décrire les interconnexions entre les objectifs réalisés. Cette méthode contient des techniques d'analyse des clusters, de fouille des données.
- **Fouille des données :** La fouille des données comprend des techniques d'analyse de clusters, de classification, de régression et de règles d'association. Cette méthode vise à identifier et extraire des informations utiles à partir de données ou de jeux de données étendues.
- **L'apprentissage automatique :** Il vise à améliorer les comportements des ordinateurs sur les grandes données pour diminuer la variance et augmenter la précision. En outre c'est un domaine très important de l'informatique qui est encore très active à l'heure actuelle.
- **Méthode de visualisation :** C'est une méthode qui devient de plus en plus nécessaire car elle sert à rendre les grosses données visibles à l'aide de représentations graphique

(diagrammes, tableaux et images) ; chose qui est plus intéressante par rapport aux informations textuelles non structurées [9].

2.2.8 Défis et enjeux

La grande progression de données constitue un énorme défi en matière d'acquisition, de stockage, de gestion et d'analyse. Les systèmes de gestion et d'analyse de données traditionnels relationnelles (SGBDR) utilisant un équipement onéreux et ne peuvent traiter des masses énormes de données hétérogènes. Cela à amener les chercheurs à proposer de nouvelles technologies telles que le Cloud Computing et les bases de données NoSQL.

Les principaux défis sont énumérés comme suit :

- ✓ *La réduction de la redondance et la compression des données :* réduire au maximum la redondance des données et procéder à leur compression pour limiter le coût de l'ensemble du système sans pour autant influer négativement sur la valeur de ces données.
- ✓ *La gestion du cycle de vie des données :* vu la masse énorme de données affluente, il convient de ne garder que les données utiles et mises à jour et supprimer tout ce qui est superflue afin d'éviter la saturation des systèmes.
- ✓ *Mécanisme analytique :* Traiter un gros volume de données hétérogène dans un temps limité.
- ✓ *La confidentialité des données : La* capacité limitée des fournisseurs ou propriétaires de ces volumes de données, ils n'ont pas les moyens de procéder à un traitement et à une analyse efficace. Ils ont recours à des professionnels ou à d'autres outils pour réaliser ces taches. Cela présente un risque de sécurité potentiel.
- ✓ *Gestion de l'énergie :* La consommation d'énergie au niveau des systèmes de stockage et d'analyse sont en nette progression. Il convient de contrôler cette consommation et l'optimiser dans la mesure du possible.
- ✓ É*volutivité :* le système d'analyse Big data doit prendre en charge les ensembles de données actuelles et futures. Les algorithmes doivent être en mesure de traiter des ensembles de données en expansion permanente. [8]

2.3 vie privé

2.3.1 vie privée : survol général

La notion de « vie privée » diffère d'un pays à l'autre conformément à la culture et à la législation en cours. Une définition a été adoptée par l'Organisation de la Coopération et du Développement Economique (OCDE). Il s'agit de toute information concernant une personne physique identifiée. La vie privée des personnes est associée à la collecte, l'analyse, le stockage et la destruction de données personnelles ayant rapport à un individu donné. L'utilisation intensive des réseaux sociaux, des Smartphones et du marketing et d'autres services en ligne impose le control d'accès et la confidentialité des données. La confidentialité vise à protéger l'information qui est considérée comme privée et empêcher qu'elle soit partagée sans le consentement éclairé de son propriétaire.

Les recherches de ces dernières années ont abouti à un développement des techniques de protection de la vie privée. Parmi ces techniques, le chiffrement qui joue un rôle essentiel en matière de protection. L'anonymisation des données et d'autres techniques qui rendent l'accès à l'information par des personnes indésirables difficiles. [10]

2.3.2 Type de vie privée

Il est difficile de définir le concept exact de « vie privée ». Bien que la notion de « protection de la vie privée » soit liée au principe de protection de données, les deux défis sont loin d'être identiques.

Nous avons sept types de « vie privée » :

1) **La vie privée de la personne :** englobe le droit de conserver les fonctions du corps et ses caractéristiques (codes génétiques et biométriques).
2) **La vie privée du comportement et de l'action :** il s'agit des habitudes des personnes, de leurs activités politiques ou sociales.
3) **La vie privée de la communication :** éviter toute interception des communications ou des courriers ou l'usage d'outils d'espionnage tels que les microphones.
4) **La vie privée des données et des images :** empêcher que ces données soient à la portée d'autres individus ou organismes non autorisés.
5) **La vie privée des pensées et des sentiments :** protéger les pensées et les sentiments des individus de la divulgation sans leur autorisation.

6) **La vie privée du déplacement et de l'espace :** liberté de se déplacer dans les lieux publics, semi publics et chez soi sans être surveillé ou suivi.
7) **La vie privée de l'association :** la liberté de s'associer à des groupes sociaux sans être dérangé. [11]

2.3.3. Défis et enjeux de la vie privée

La vie privée est l'un des problèmes critiques du Big Data, qu'il faut régler de manière appropriée avant de profiter des applications du Big Data.

Il y a beaucoup problèmes et défis à venir en termes d'étude de la vie privée dans les Big Data. Nous résumons ici les principaux :

Mesure de la vie privée :

La protection de la vie privée étant un concept subjectif, elle varie d'une personne à l'autre, de temps en temps, même pour la même personne.

Il est difficile de définir et par conséquent, difficile à mesurer.

C'est un problème. Cela nécessite un effort technique et une perspective psychologique.

Cadre théorique de la vie privée :

Il ya des méthodes pour regrouper des données et le cadre de vie privée différentiel pour la confidentialité des données.

Mais, nous voyons les limites des différentes méthodes de regroupement des données et la nécessité d'adapter la confidentialité différentielle dans la pratique.

Donc il faut avoir des nouvelles bases théoriques pour l'étude de la protection de vie privée dans L'ère des données.

Evolutivité des algorithmes de la vie privée :

Nous avons des mécanismes et des stratégies pour gérer les grandes bases de données, et la stratégie principale est la division.

Mais les big data est bien plus grande qu'une base de données.

Par conséquent, il est difficile de concevoir des algorithmes évolutifs pour les algorithmes de vie privée**.**

Hétérogénéité de la source de données :

Les algorithmes de la vie privée disponibles concernent presque tous des sources de données homogènes, comme les enregistrements dans les bases de données.

Les sources des données du Big Data à venir sont hétérogènes donc Il est difficile de gérer ces sources.

Efficacité des algorithmes de la vie privée :

Étant donné le volume de données grandes, l'efficacité devient un élément très important des algorithmes de la vie privée dans l'environnement Big Data. [14]

MapReduce permet la manipulation d'une masse volumineuse de données en les distribuant dans un cluster de machines pour être traitées. Ce modèle connaît un succès incontestable auprès d'organisme possédant d'importants centres de traitement de données. Dans cette section, nous allons citer quelques défis concernant la sécurité et la vie privée pour les calculs *Mapreduce*.

- ✓ Le grand volume de données et son stockage pris en charge par Mapreduce présentent un défi de sécurisation des calculs et de confidentialité. Dans ses calculs Mapreduce fractionne les masses de données et les distribue sur plusieurs nœuds. Chaque fraction doit être transférée d'une manière sécurise et confidentielle.
- ✓ Haute distribution : Mapreduce exige un grand nombre de clusters de nœuds. Les nœuds reçoivent en parallèle les données qui vont être traitées et stockées. La mise en œuvre d'un grand nombre de nœuds pour l'analyse et le stockage de données est indispensable. La distribution des données sur plusieurs clusters augmente le risque d'attaque d'où la nécessité de rendre plus sûr ce mode de traitement. On doit également assurer la sécurité des clusters mis en jeu ce qui revient à dire qu'il faut sécuriser un nombre important de machines participant à cette opération. En conclusion le mode de traitement distribué présente un défi en matière de sécurité qu'il faut relever.
- ✓ L'accès aux données non fiables : Mapreduce est d'une grande flexibilité, il permet la réalisation des calculs définis par l'utilisateur. Cette souplesse exige par contre une grande prudence par les utilisateurs. Le risque de créer des codes perturbant le système existent. Des algorithmes de sécurité doivent être mis au point pour faire face aux codes endommagés.
- ✓ la protection de la vie privée des données des fournisseurs de Cloud : les utilisateurs ont la possibilité de stocker leurs données privées dans les Cloud publics. Les fournisseurs de ces services peuvent contrôler et accéder aux données et au code MapReduce et même procéder à des modifications. Dans ces conditions, il est impossible d'assurer la vie privée des clients en présence d'un fournisseur Cloud.

- ✓ Des multi-utilisateurs sur un seul nuage public : les fournisseurs de données et les fournisseurs de Cloud public doivent permettre à plusieurs utilisateurs d'accéder à leurs données simultanément sans que la vie privée de chacun ne soit menacée. On doit assurer à chaque utilisateur d'accéder à la totalité de ses données sans entraves tout en protégeant les données des autres utilisateurs. [12][13]

2.3.4 Sécurité Via vie privé

La sécurité et la vie privée sont des sujets très débattus depuis longtemps, car il est toujours délicat de cerner avec précision une définition à ces concepts, les informations fournies sont liées à des entités telles que les individus ou les entreprises et sont souvent requises, dès que la protection de la confidentialité s'impose, l'une des solutions c'est l'anonymisation des données elle sert à supprimer tous les informations directement et indirectement identifiable afin que la ré-identification soit impossible pour les concernés en question. Avec l'arrivée des Big Data d'énormes types de données sont collectées, donc la protection de vie privée devient l'un des grands défis pour l'humanité. [15]

La vie privée se définit comme un ensemble de renseignements de nature confidentielle se rattachant à un individu donné qu'il convient de protéger contre toute divulgation ou un usage non autorisé et à la volonté de rester hors de la vie publique.

La donnée personnelle concerne toute information relative à une personne physique identifiée ou qui peut être identifiée, directement ou indirectement, par référence un numéro d'identification ou à un ou plusieurs éléments qui lui sont propres.

La sécurité est un ensemble des moyens techniques, organisationnels, juridiques et humains nécessaires et mis en place pour conserver, rétablir, et garantir la protection de l'information contre les menaces accidentelles ou délibérées.

2.3.5 Gestion de la confiance

La sécurité et la vie privée sont étroitement liées à la confiance, cette dernière a été étudiée par diverses disciplines. La confiance est une entité qui se comporte de manière attendue, malgré le manque de capacité à contrôler l'environnement dans lequel elle opère. Il

est important de considérer deux grands types de confiance dans un environnement cloud la confiance dure ou les plates-formes de service sont fiables si l'existence de primitives de sécurité nécessaires est prouvable. La confiance douce implique des aspects tels que les émotions humaines intrinsèques, les perceptions, les expériences d'interaction les commentaires des utilisateurs. On peut tirer profit du TPM (Trusted Platform Module) qui contient une clé privée qui identifie de manière unique le TPM et aussi l'hôte physique et certaines fonctions cryptographiques, afin de rendre la sécurité plus résiliente a ces problèmes [16]

Selon différentes techniques de gestion de confiance adoptées dans la littérature,[Noor et al] ont classé ces modèles de confiance en quatre catégories différentes : politique,

Recommandation, réputation et prévision.

La gestion de la confiance joue un rôle important et notre ère rime avec l'utilisation de plus

De sources de données et la confiance dans différentes grandes étapes du cycle de vie des

Données, devrait recevoir plus d'attention et doit être étudiée de manière approfondie.

2.3.6 Infrastructure critique et Big data

L'infrastructure critique se définit différemment d'un pays à un autre. Généralement il englobe les installations et les organisations qui fournissent des services (santé, transport, sureté ...Etc.) ou des produits (production agricole, industrielle...Etc.) au pays. En bref c'est actif vital pour le fonctionnement d'un état ou d'une entreprise.
L'infrastructure critique a pris une grande importance aux yeux des états qui ont pris des mesures explicites pour assurer leur protection.

L'infrastructure critique de l'information présente un aspect stratégique et constitue une partie non négligeable de l'infrastructure critique d'une société donnée.

Big data à des ramifications qui touchent tous les secteurs vitaux de la société tels que la défense, la santé, l'économie et la recherche scientifique. En conséquence la protection de l'infrastructure critique du Big data est une priorité majeure dans la politique des états. Il est essentiel de signaler qu'en matière d'infrastructure critique le défi n'est pas la confidentialité mais la disponibilité c'est-à-dire l'efficacité. Dans le passé l'infrastructure critique d'information (CII) n'était pas connectée à internet mais aujourd'hui et pour diverses raison cela a changé. Le volume de données grandit au fur et à mesure et le traitement local de

l'information n'est plus possible, d'où le recours à un stockage et un traitement centralisés ou en cluster. [10]

2.4 Terminologie du domaine de la vie privée

2.4.1 Anonymat

Pour permettre l'anonymat d'un sujet, il doit toujours exister un ensemble approprié des sujets ayant potentiellement les mêmes attributs.

-L'anonymat est l'état d'être non identifiable dans un ensemble de sujets, l'ensemble d'anonymat.

-L'ensemble d'anonymat est l'ensemble de tous les sujets possibles.

-En ce qui concerne les acteurs, l'ensemble d'anonymat comprend les sujets susceptibles de provoquer une action.

-En ce qui concerne les destinataires, l'ensemble d'anonymat comprend les sujets pouvant être abordés.

- un expéditeur ne peut être anonyme que dans un ensemble d'expéditeurs potentiels.

-un destinataire peut être anonyme au sein d'un ensemble de destinataires potentiels.

-Les deux ensembles d'anonymat peuvent être disjoints, identiques ou se chevaucher.

-Les ensembles d'anonymat peuvent varier dans le temps.

-Un sujet est une entité potentiellement active telle que, par exemple, un être humain (c'est-à-dire une personne physique), une personne morale ou un ordinateur.

- (Une organisation n'agissant pas en tant que personne morale ne constitue ni un sujet ni une entité, mais un ensemble (éventuellement structuré) de sujets ou d'entités. Sinon, la distinction entre «sujets» et «ensembles de sujets» serait complètement floue.

-De [ISO]: "[Anonymat] garantit qu'un utilisateur peut utiliser une ressource ou un service sans divulguer l'identité de cet utilisateur.

-Les exigences en matière d'anonymat assurent la protection de l'identité de l'utilisateur.

-L'anonymat n'est pas destiné à protéger l'identité du sujet. L'anonymat exige que Les autres utilisateurs ou sujets sont incapables de déterminer l'identité d'un utilisateur lié à un sujet ou à une opération. Comparée à cette explication, cette définition est plus générale car elle ne se limite pas à identifier les utilisateurs, mais tous les sujets. C'est-à-dire les «suspects habituels». L'ensemble des sujets possibles dépend de la connaissance de l'attaquant.

Toutes choses étant égales par ailleurs, l'anonymat est d'autant plus fort que l'ensemble de l'anonymat est grand et que l'envoi ou la réception, respectivement, des sujets de cet ensemble sont répartis plus équitablement.

Afin de quantifier l'anonymat dans des situations concrètes, il faudrait décrire le système avec suffisamment de détails, ce qui n'est pratiquement pas (toujours) possible pour les grands systèmes ouverts (mais peut-être pour certaines petites bases de données, par exemple).

Nous pourrions utiliser la qualité d'anonymat comme un terme comprenant à la fois la quantité et la robustesse de l'anonymat. [17]

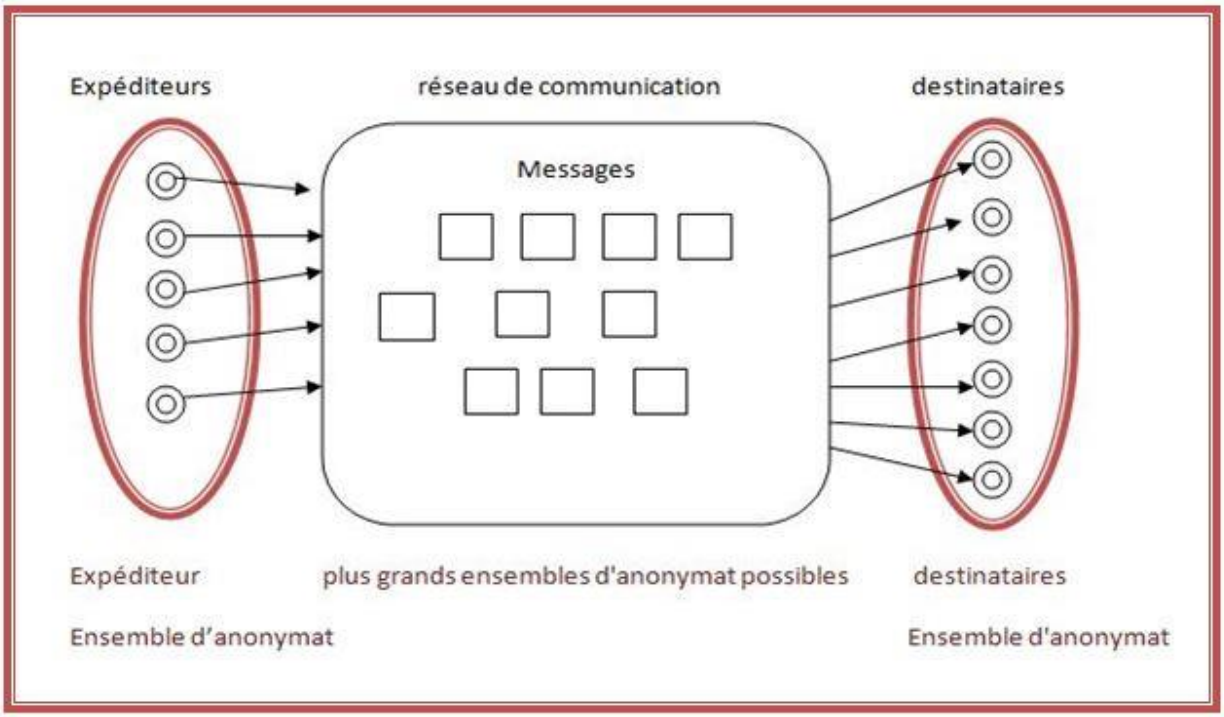


Figure 2.3 : l'anonymisation

2.4.2 Intraçabilité (Unlinkability)

Unlinkability n'a de sens que lorsque le système dans lequel nous voulons décrire l'anonymat ou les propriétés du pseudonyme ont été définis et que les entités intéressées par la liaison (l'attaquant) ont été caractérisées.

Ensuite: impossibilité de lier deux ou plusieurs éléments d'intérêt (par exemple, sujets, messages, événements, actions, ...) signifie que, dans le système (comprenant ces éléments et éventuellement d'autres éléments), ces éléments présentent ne sont ni plus ni moins liés après son observation que ne le sont ses connaissances concernant sa connaissance a priori.

Cela signifie que la probabilité que ces éléments soient liés du point de vue de l'attaquant reste la même avant et après l'observation de l'attaquant.

De [ISO]: “[Unlinkability] garantit qu'un utilisateur peut utiliser plusieurs ressources ou services sans que d'autres soient en mesure de les lier.

-Unlinkability nécessite que les utilisateurs et / ou les sujets ne puissent pas déterminer si le même utilisateur a causé certaines opérations spécifiques dans le système. ”

- Contrairement à cette définition, la signification de la non-liaison est moins centrée sur l'utilisateur, mais traite de la non-liaison des «éléments» et constitue donc une approche générale.

- les scientifiques peuvent distinguer entre «impossible de lier absolu» (comme dans [ISO]; c’est-à-dire «pas de détermination d’un lien entre les utilisations») et «relativement de dissociabilité» connaissances sur le lien entre les utilisations »).

-Dans le cas particulier où l'on sait auparavant que certains éléments sont liés, la probabilité que ces éléments soient liés reste bien sûr la même. [17]

2.4.3 Inobservabilité

L’inobservabilité est l’état des objets d’intérêt (IOI) ne pouvant pas être distingués de tout IOI (du même type)

-Cela signifie que les messages, par exemple, ne sont pas discernables. «Bruit aléatoire».

-Comme ils avaient des groupes de sujets liés à l'anonymat en ce qui concerne l'anonymat, ils ont des groupes de sujets non observables en ce qui concerne l'inobservabilité.

- Inobservabilité de l'expéditeur signifie alors qu'il n'est pas visible si un expéditeur faisant partie de l'ensemble des inobservances envoie.

- L'inobservabilité du destinataire signifie alors qu'il n'est pas visible si un destinataire appartenant à l'ensemble d'inobservabilité reçoit.

Inobservabilité signifie alors qu'il n'est pas visible si quelque chose est envoyé sur un ensemble d'expéditeurs potentiels à un ensemble de destinataires potentiels.

En d'autres termes, il n'est pas perceptible de savoir si un message est échangé dans une relation quelconque dans l'ensemble non observable de la relation de toutes les paires émetteur-destinataire possibles.

L'inobservabilité peut être considérée comme une propriété possible et souhaitable des systèmes stéganographiques.

Par conséquent, il correspond à la terminologie de masquage d'informations.

De [ISO]: “[inobservabilité] garantit qu'un utilisateur peut utiliser une ressource ou un service sans que d'autres, en particulier des tiers, puissent observer que la ressource ou le service est utilisé.

Pour être inobservable, les utilisateurs et / ou les sujets ne peuvent pas déterminer si une opération est en cours d'exécution. »Comme ils ont vu précédemment, leurs approche est moins centrée sur l'utilisateur et plus générale. [17]

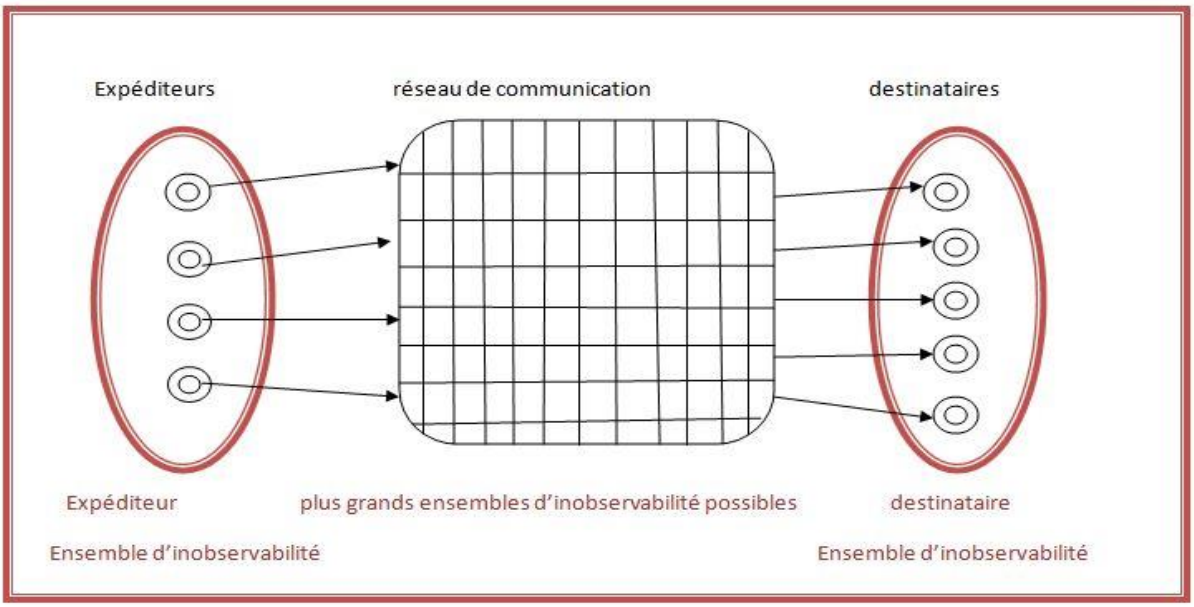


Figure 2.4 : l'inobservabilité

2.4.4 Pseudonymité

Le sujet auquel le pseudonyme fait référence est le titulaire du pseudonyme. Être pseudonyme, c'est utiliser un pseudonyme comme identifiant.

Dans le contexte habituel, les scientifiques supposent que chaque pseudonyme désigne exactement un titulaire, invariant dans le temps, non transféré à d'autres sujets.

Des types spécifiques de pseudonymes peuvent étendre ce paramètre: Un pseudonyme de groupe fait référence à un ensemble de détenteurs, c'est-à-dire qu'il peut faire référence à plusieurs détenteurs; un pseudonyme transférable peut être transféré d'un titulaire à un autre sujet devenant son titulaire.

Un tel pseudonyme de groupe peut induire un ensemble d'anonymat: à l'aide des informations fournies par le pseudonyme uniquement, un attaquant ne peut pas décider si une action a été effectuée par une personne spécifique de l'ensemble.

«Pseudonyme» vient du grec «pseudonumon» qui signifie «faux nom» (pseudo: faux; onuma: nom). Ainsi, cela signifie un nom autre que le «vrai nom».

Le «vrai nom» (écrit dans les papiers d'identité délivrés par l'État) étant quelque peu arbitraire (il peut même être changé de son vivant), ils étendrent le terme «pseudonyme» à tous les identifiants, y compris tous les noms ou autres chaînes de bits.

Sur un plan fondamental, les pseudonymes ne sont rien d'autre qu'un autre type d'attribut.

Mais alors que dans la construction de systèmes informatiques, son concepteur peut garder les pseudonymes sous son contrôle et / ou celui de l'utilisateur. Par conséquent, il est utile de donner à ce type d'attribut contrôlé par le système un nom distinct: pseudonyme.

Ils préfèrent le terme «titulaire» à «propriétaire» d'un pseudonyme, car il ne semble pas logique de «posséder» des ID, par exemple des chaînes de bits. En outre, le terme "titulaire" semble plus neutre que le terme "propriétaire"

Veuillez noter que, bien que les termes «anonyme» et «pseudonyme» partagent la plupart de leurs lettres, leur sémantique est très différente: Anonyme dit quelque chose sur l'état d'un sujet en ce qui concerne l'identifiabilité, pseudonyme ne dit que sur l'utilisation d'un mécanisme, à savoir , en utilisant des pseudonymes.

Veuillez noter que le simple fait qu'un pseudonyme ait plusieurs titulaires ne donne pas un pseudonyme de groupe: par exemple, créer le même pseudonyme peut arriver par hasard et même sans que les titulaires en soient conscients, en particulier s'ils choisissent les pseudonymes et préfèrent les pseudonymes qui sont faciles à retenir. [17]

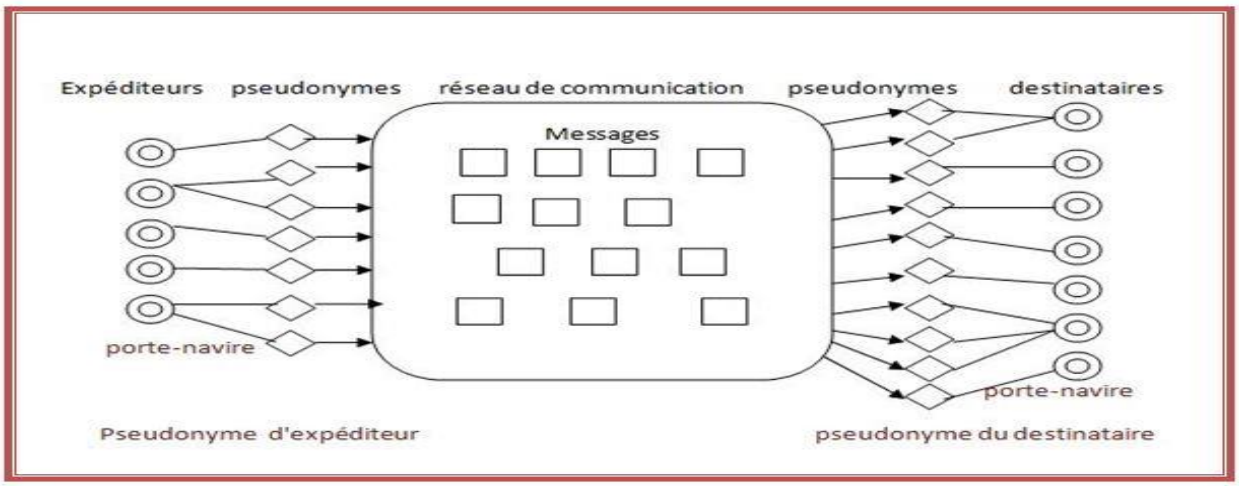


Figure 2.5 : Pseudonymité

2.4.5 Gestion d'identité

2.4.5.1 Réglage

Pour traiter correctement la gestion des identités renforçant la confidentialité, les scientifiques devient élargir leurs cadre :

• Il n'est pas réaliste de supposer qu'un attaquant pourrait ne pas obtenir d'informations sur l'expéditeur ou le destinataire des messages à partir du contenu du message et / ou du contexte d'envoi ou de réception (heure, informations de localisation, etc.) du message.

Nous devons considérer que l'attaquant est capable d'utiliser ces propriétés pour lier des messages et, par conséquent, les pseudonymes utilisés avec eux.

• De plus, ce ne sont pas seulement des êtres humains, des personnes morales ou simplement des ordinateurs qui envoient des messages et utilisent des pseudonymes à leur discrétion, mais utilisent des programmes d'application qui influencent fortement l'envoi et la réception de messages et peuvent même détermine fortement l'utilisation de pseudonymes.

2.4.5.2 Identité et identifiabilité

L’identité peut s’expliquer par une perception exclusive de la vie, une intégration dans un groupe social et une continuité, liée à un corps et façonnée par la société.

Ce concept d'identité distingue «moi»: c'est l'instance accessible uniquement par le moi individuel, perçue comme un exemple de liberté et d'initiative.

D'un point de vue structurel, l'identité peut être liée à n'importe quel sujet, qu'il s'agisse d'un être humain, d'une personne morale ou même d'un ordinateur.

-L'identifiabilité est l'état d'être identifiable au sein d'un ensemble de sujets, l'ensemble identifiable.

Figure2. 6 : anonymisation & identifiabilité

- l'identifiabilité est d'autant plus forte que l'ensemble d'identifiabilité correspondant est grand. Inversement, l'anonymat restant est d'autant plus fort que l'ensemble d'identifiabilité respectif est petit.

-Une identité est un sous-ensemble d'attributs d'un individu qui l'identifie au sein d'un ensemble d'individus.

2.4.5.3 Termes liés à l'identité

Il ya 3 terme lié a l'identité :

Identité partielle

Chaque identité d'une personne comprend de nombreuses identités partielles dont chacune représente la personne dans un contexte ou un rôle spécifique.

- Une identité partielle est un sous-ensemble d'attributs d'une identité complète, où une identité complète est l'union de tous les attributs de toutes les identités de cette personne.

- Sur le plan technique, ces attributs sont des données.

- Bien entendu, les valeurs d'attributs ou même les attributs eux-mêmes d'une identité partielle peuvent évoluer dans le temps.

Identité numérique

L'identité numérique désigne l'attribution de propriétés à une personne, qui sont immédiatement accessibles d'un point de vue opérationnel par des moyens techniques. Plus

précisément, l'identifiant d'une identité partielle numérique peut être une simple adresse électronique dans un groupe de discussion ou une liste de diffusion.

-Son propriétaire va atteindre une certaine réputation.

Identité virtuelle

L'identité virtuelle est parfois utilisée dans le même sens qu'identité numérique ou identité partielle numérique, mais en raison de la connotation avec «irréel, inexistant, semblant», le terme s'applique principalement aux personnages d'un MUD (Multi User Dungeon), d'un MMORPG (Jeux de rôle en ligne massivement multi-joueurs) ou d'avatars. [17]

2.5. Les techniques de protection de la vié privé en big data

2.5.1 L'identification

L'identification est une ancienne technique populaire de protection de la vie privée de données. Elle peut faire la préservation de la confidentialité dans l'analyse de données volumineuses.

Les intrus peuvent obtenir plus d'informations lors de l'identification sur la plateforme du Big Data. Donc cette méthode ne suffit pas pour préserver la confidentialité des données volumineuses.

Pour cela, nous devons améliorer cette méthode avec l'aide des méthodes préservant la confidentialité telles que l'anonymat, la diversité et la proximité.

1) anonymat:

Dans cette méthode, la base de données est conservée sous forme de tableau, comme les lignes et les colonnes.

Ce système construit et évalue des algorithmes pour les informations de version et expose les propriétés des entités à protéger.

Mais cette méthode a une limitation de l'homogénéité-attaque, la connaissance de base.

2) diversité:

Cette méthode est une extension du modèle d'anonymat qui résout également certaines faiblesses du modèle d'anonymat.

Le modèle de diversité préservant la confidentialité des ensembles des données et réduisant la granularité des données.

Dans cette méthode, chaque attribut sensible est représenté par des valeurs bien représentées.

Le problème de cette méthode est qu'elle repose sur les données sensibles.

3) proximité:

Cette méthode constitue un progrès supplémentaire de la diversité qui aide à protéger la confidentialité des ensembles de données et réduit également la granularité d'une représentation de données.

La proximité est la distance entre l'attribut sensible dans la classe et la distribution de l'attribut dans la table entière est inférieure au seuil.

L'avantage de la proximité est sa capacité à éviter les annonces d'attributs.

Le problème dans cette méthode est que si la taille et la variété des données augmentent, les chances de ré-identification du processus augmentent également. [18]

2.5.2 Confidentialité différentielle

La technologie de la confidentialité différentielle permettant aux analystes de bases de données d'obtenir les informations nécessaires à partir de la base de données contenant les informations personnelles des utilisateurs, mais sans conçu l'identité personnelle des individus. Les informations personnelles de l'utilisateur stockées dans la base de données.

L'analyste du Confidentialité différentielle souhaite accéder à ces moyens d'information à l'aide de logiciels pouvant le faire. Cette méthode restreint l'accès direct aux données de la base de données, car le logiciel intermédiaire développé et placé entre la base de données et l'analyste.

Ce logiciel aidé à préserver la confidentialité des données.

Étape 1: l'analyste de données envoie la demande à la base de données à l'aide d'un logiciel intermédiaire.

Étape 2: Le logiciel reçoit cette demande et traite cette requête.

Étape 3: Après cela, le logiciel obtient le résultat de la base de données.

Étape 4: Ajouter une modification dans ce résultat dépend du risque de confidentialité évalué et enfin du transfert à l'analyste.

La quantité d'emballage ajoutée aux informations d'origine est proportionnelle au risque d'atteinte à la vie privée. Donc risque pour la vie privée faible signifie que l'emballage est ajouté. [18]

Conclusion

Dans le chapitre 2 nous avons vu une généralité sur les big data puis on étudier la vie privée on big data et ces Défis et enjeux, Terminologie du domaine de la vie privée, Les techniques de protection du vié privée en bigdata.

On conclure qu'il existe plusieurs méthodes et propositions pour augmenter la sécurité de la vie privée dans les big data. Malheureusement il n'existe pas une standardisation mais que des propositions. Ces propositions ne sécurisent pas la totalité des big data mais elles essayent de fournir plus d'intimité et plus de protection aux vies et aux informations des utilisateurs contre les attaques probables par les voleurs d'identité.

Dans le chapitre 3 on va étudier les approches utilisé pour assurer la confidentialité des big data et chaque approche a son principe et ces méthodes.

Chapitre 3

Approches et travaux Connexes

3.1 Introduction

Avec l'avènement du Big Data, un volume de plus en plus important de données des appareils mobiles, de réseaux de capteurs, et de l'Internet des objets ont été générés et collecté. L'analyse des données est un processus essentiel dans les big data

Le Plateformes Cloud où la plupart des Big Data sont stockées utilisant une infrastructure Cloud est devenue de plus en plus importants pour la prise en charge du Big Data. De nombreuses personnes et entreprises choisissent de télécharger leurs données sur le Cloud, ceux-ci pouvant prendre en charge un service de stockage de données considérable, aussi une capacité de traitement de données efficace. Bien que les nuages comportent de nombreuses caractéristiques essentielles telles que le cout et l'évolutivité élevée, la vie privée est l'un des obstacles majeurs sur les big data du Cloud public.

Il y a plusieurs moyens pour assurer la vie privée sur les big data. Les nouvelles approches de la protection de la vie privée s'accordent sur la nécessité de déplacer la responsabilité des individus concernant leurs données personnelles vers les organisations qui les utilisent.

Dans ce chapitre on étudier plusieurs approches utilisé pour assurer la vie privée et chaque approche a son principe et ces méthodes et chaque approche a ses avantages et ses inconvénients.

3.2 Anonymisation multi dimensionnel

Il y a plusieurs moyens pour assurer la confidentialité comme l'anonymat.

Il existe de nombreuses méthodes d'anonymat, parmi ces méthodes, nous étudierons le schéma d'anonymisation multidimensionnel qui est un schéma de recodage global et qui établit un bon équilibre entre la distorsion des données et leur facilité d'utilisation.

Un algorithme de partitionnement récursif appelé Mondrian a été proposé pour mettre en œuvre le schéma sous k-anonymat. L'approche est nommée MR Mondrian pour référence, après la désignation des algorithmes de Mondrian. L'idée est de partitionner de manière récursive un jeu de big data en partitions plus petites à l'aide de MapReduce jusqu'à ce que toutes les partitions puissent tenir dans la mémoire de chaque nœud informatique. Ensuite, les algorithmes de Mondrian traditionnels peuvent être exécutés sur chaque nœud de manière parallèle. Des expériences approfondies sur des données réelles montrent que cette approche peut considérablement améliorer l'extensibilité de l'anonymisation multidimensionnelle.

Cette approche d'anonymisation multidimensionnelle basée sur MapReduce et évolutive, nommée MR Mondrian, contient 5 principale section.

La section 1 : (Pilote MR Mondrian)

Elle explore la mise en œuvre itérative de Mondrian et présente le pilote MR Mondrian. Selon la méthode de Mondrian, une approche simple basée sur MapReduce consiste à invoquer des tâches MapReduce de manière récursive. Les scientifiques peuvent exécuter une opération MapReduce pour calculer les statistiques de comptage pour l'heuristique de fractionnement d'un ensemble de données donné. Ensuite, ils exécutent une autre opération

MapReduce pour diviser les données en fonction de l'attribut de fractionnement et des valeurs de domaine.

Ces deux opérations MapReduce sont exécutées de manière récursive sur chaque partition. Cela consomme trop de nœuds de calcul et entraîne des frais généraux élevés pour l'initialisation et la planification des tâches.

La section 2 : Arbre d'indexation d'ID de partition :

La structure de données de base de l'arborescence d'indexation ID de partition (arborescence PID) est formulée à la section 2.

L'arborescence d'indexation ID de partition (arborescence PID) joue un rôle central dans l'algorithme itératif de Mondrian avec MapReduce. Il est essentiel que les travaux MapReduce identifient la partition à laquelle un enregistrement de données appartient au cours du processus de calcul des méthodes heuristiques de fractionnement ou des ensembles de données.

C'est-à-dire qu'avec un enregistrement r, les scientifiques essayent de connaître sa partition actuelle. Cela permet de traiter simultanément plusieurs partitions de données en profondeur de manière discontinue avec MapReduce.

La section 3 : Recherche de médiane évolutive à l'aide de MapReduce :

La médiane de l'anonymisation multidimensionnelle n'a pour but que de trouver un point de fractionnement approprié pour un attribut numérique permettant d'obtenir une bonne occupation des partitions. Le processus de base pour trouver la médiane des médianes se compose de trois étapes.

Tout d'abord, une série de groupes de données est construite, avec une taille de groupe N. Ensuite, la médiane de chaque groupe peut être identifiée rapidement car N est généralement défini comme un petit nombre. Enfin, les médianes des médianes de la dernière étape sont déterminées.

La section 4 : Calcul de CV et partitionnement des données :

Elle décrit les travaux MapReduce pour calculer le coefficient de variation (CV) de chaque attribut afin de sélectionner l'attribut de division et de la partition de données concrète.

La section 5 : Extension aux modèles de confidentialité différentiels :

Les solutions d'évolutivité peuvent également être étendues à d'autres modèles de confidentialité, y compris la confidentialité différentielle.

Donc l'approche de MR Mondrian pour l'anonymisation multidimensionnelle sur big data basée sur le paradigme MapReduce. l'idée de base consistant à diviser les ensembles de données en petites partitions de données pour les insérer dans la mémoire principale d'un seul nœud, puis proposé de procéder au partitionnement itératif des données de manière parallèle. Le partitionnement concret des données est effectué lorsque toutes les partitions de données peuvent être insérées dans la mémoire principale. Ensuite, chaque partition de données est divisée de manière récursive par la méthode série classique de Mondrian sur un seul nœud.

Pour prendre en charge le partitionnement itératif des données, une structure arborescente appelée PID-tree a été proposée pour indexer les partitions de données afin de rechercher les ID de partition.

Le coefficient de variation a été utilisé pour sélectionner les attributs de division d'un attribut catégorique, tandis qu'une méthode évolutive basée sur l'idée de la médiane des médianes et de la technique de l'histogramme a été proposée pour trouver la médiane d'un attribut numérique.

Les résultats expérimentaux d'ensembles de données du monde réel ont montré que cette approche peut considérablement améliorer l'évolutivité et la rentabilité du schéma multidimensionnel par rapport aux approches existantes.

À l'avenir, Il est prévu que cette approche à des plates-formes d'exploration de données évolutives afin de parvenir à une extraction de données Big Data évolutive préservant la confidentialité.

Les inconvénients du cette approche :

- La préservation de la confidentialité des données rencontre un problème lorsque les données évoluent dans un contexte de bigdata
- consomme trop de nœuds de calcul et entraîne des frais généraux élevés pour l'initialisation et la planification des tâches.
- difficile à mesurer la vie privée parce qu'elle subjectif.
- Hétérogénéité de la source de données.

3.3 Anonymisation par proximité avec MapReduce

Parmi les méthodes d'anonymat il existe une autre méthode appelée l'anonymisation par proximité, l'idée est trouver une solution pour l'anonymisation par recodage local du Big Data. Le schéma de recodage local, également connu sous le nom de généralisation de cellule, regroupe des ensembles des données dans un ensemble des cellules au niveau de l'enregistrement des données et anonyme chaque cellule individuellement.

Comme il est prouvé que le problème de satisfiabilité du modèle de confidentialité de proximité est difficile, il est intéressant et pratique de le modéliser comme un problème de regroupement consistant à minimiser à la fois la distorsion des données et la proximité entre des valeurs sensibles dans un groupe, pour trouver une solution satisfaisant rigoureusement le modèle de confidentialité. Le problème de regroupement sensible à la proximité fait référence au problème de regroupement Proximity-Aware à objectif unique SPAC (the Single-objective Proximity Aware Clustering).Pour résoudre le problème de SPAC dans les scénarios Big Data, ils proposent une approche de clustering en deux phases comprenant les algorithmes de classification par ancêtres T (la méthode de classification par affectation de points) et de classification par agglomération sensible à la proximité. Les deux phases basées sur MapReduce.

Esquisse d'un clustering en deux phases :

Afin de choisir les méthodes de clustering appropriées pour le problème SPAC Certaines observations relatives aux problèmes de clustering pour l'anonymisation des données doivent être prises en compte.

- Premièrement, le paramètre k dans le modèle de confidentialité k-anonymity est relativement petit par rapport à l'échelle d'un ensemble de données dans des scénarios Big Data. Dans la mesure où la limite supérieure de la taille d'un cluster pour l'anonymisation par recodage local est de 2k-1, la taille des clusters est également relativement petite. En conséquence, le nombre de clusters sera assez important.

-Deuxièmement, à condition que la taille d'un cluster ne soit pas inférieure à, plus un cluster est petit, plus il est préférable. La raison en est que cela entraîne généralement moins de distorsion des données. Idéalement, la taille de tous les groupes est exactement la même.

-Troisièmement, l'architecture de clustering intrinsèque dans un ensemble de données est utile pour l'anonymisation par recodage local, mais la construction d'une telle architecture n'est pas l'objectif final.

Mise en cluster d'ancêtres t pour la partition des données :

L'un des problèmes principaux de la méthode d'affectation de points est de savoir comment représenter un cluster. Un ancêtre d'un cluster désigne un enregistrement de données dont la valeur d'attribut de chaque quasi-identificateur qualitatif est l'ancêtre commun le plus bas des valeurs d'origine du cluster. Chaque quasi-identificateur numérique d'un enregistrement ancêtre est la médiane des valeurs d'origine dans le cluster. Pour faciliter le clustering des ancêtres t, les auteurs prennent en considération les attributs quasi-identifiants, mais aussi les attributs sensibles. Initialement, les ancêtres de la première série d'attribution de points sont des enregistrements qui sont dédiés à la graine.

En général, la sélection de tels enregistrements t influence dans une certaine mesure la qualité de la classification. Pour obtenir un bon ensemble de semences, les auteurs sélectionnent des enregistrements de données aussi éloignés que possible les uns des autres.

Concrètement, ils sélectionnent les semences via un travail MapReduce, SeedSelection, qui produit un ensemble de semences: s = {r1, ... rt}. Un seul réducteur est utilisé pour la sélection des semences en raison de la nature en série

Clustering agglomérant sensible à la proximité :

Dans la méthode de classification par agglomération, chaque enregistrement de données est considéré initialement comme une grappe, puis deux grappes sont sélectionnées pour être fusionnées à chaque tour d'itération jusqu'à ce que certains critères d'arrêt soient satisfaits.

En général, deux groupes dont la distance est la plus courte sont fusionnés. Ainsi, l'un des principaux problèmes de la méthode de regroupement par agglomération est de savoir comment définir la distance entre deux groupes.

Pour coïncider avec l'objectif du problème SPAC, ils exploitent la distance de liaison complète dans notre algorithme de classification agglomérée, c'est-à-dire que la distance entre deux grappes est égale à la distance pondérée la plus éloignée entre ces deux enregistrements (un dans chaque grappe). Loin les uns des autres.

Après la fusion de ces deux groupes, la distance qui les sépare correspond au diamètre du nouveau groupe.

Un travail MapReduce nommé Agglomertive Clustering. Un réducteur de MapReduce a collecté tous les enregistrements de données d'un cluster, il est exécuté pour générer des grappes finales (groupes QI). La fonction Map est relativement simple, elle n'émet qu'un record et son cluster correspondant.

Les inconvenants :

- le nombre de clusters sera assez important.
- difficile de Mesuré la vie privée.
- Dans un environnement Cloud, la préservation de la confidentialité pour l'analyse, le partage et l'exploration de données constitue un problème de recherche complexe en raison du volume de plus en plus important de jeux des données, ce qui nécessite des investigations approfondies.

3.4 Stockage multi partagé

L'utilisation de manière triviale des mécanismes de cryptage traditionnels, ne peut pas empêcher la divulgation de certaines informations sensibles au serveur Cloud mais également au public. Les systèmes de cryptage traditionnels ne tiennent pas compte de l'anonymat d'un expéditeur / récepteur de texte chiffré. En conséquence, une personne pouvant être quelqu'un ayant la capacité d'accéder au texte chiffré de l'enregistrement peut savoir sous quelle clé publique le texte chiffré est cryptée, à savoir qui est le propriétaire du texte crypté, de sorte que la personne associée au texte crypté puisse être facilement identifiée.

Pour résoudre ce problème et préserver l'anonymat, certains mécanismes de chiffrement bien connus sont proposés dans la littérature, tels que le BIE anonyme, le FBA anonyme. Avec l'utilisation de ces primitives, la source et la destination des données peuvent être protégées de manière privée. Mais les primitives ne peuvent pas supporter la mise à jour du récepteur de texte chiffré.

Proxy Re-Encryption (PRE) est proposé par Mambo et Okamoto pour résoudre le dilemme du partage de données. Il permet à une partie semi-digne de confiance appelée proxy de transformer un texte chiffré destiné à un utilisateur en un texte chiffré du même texte en clair destiné à un autre utilisateur sans que les connaissances des clés de déchiffrement ou du texte

en clair ne soient divulguées. La charge de travail du propriétaire des données est transférée vers le proxy et l'exigence de «en ligne tout le temps» pour le propriétaire est inutile.

Ce travail se concentre sur le paramétrage cryptographique basé sur l'identité.

Pour utiliser PRE dans le paramètre IBE, définissez la notion de rechiffrement de proxy basé sur l'identité (IBPRE), qui offre une solution pratique pour le contrôle d'accès dans le stockage de fichiers en réseau et la messagerie sécurisée avec IBE.

Pour capturer simultanément la propriété préservant la confidentialité et la mise à jour du destinataire du texte chiffré, a proposé un système IBPRE anonyme, qui constitue la sécurité CCA dans le modèle oracle aléatoire.

Un schéma de rechiffrement de proxy conditionnel basé sur une identité multi-sauts unidirectionnels (MH-IBCPRE) comprend plusieurs algorithmes.

Pour le rechiffrement, les auteurs utilisent une technique de chiffrement à clé symétrique à usage unique pour «encapsuler» le résultat du rechiffrement et le texte chiffré de niveau précédent (rechiffré), afin que ces éléments ne puissent pas être réutilisés par l'adversaire lors du tour suivant. Rechiffrement. Donc la nouvelle notion, le rechiffrement de proxy conditionnel anonyme basé sur une identité multi-sauts anonyme, afin de préserver l'anonymat pour l'expéditeur / le récepteur de texte chiffré, le partage de données conditionnel et la mise à jour de plusieurs destinataires.

Les inconvénients :

- difficile à mesurer la vie privée parce qu'elle varie d'une personne à l'autre.
- Étant donné le volume de données grandes, les algorithmes de chiffrements peut être moins efficace dans l'environnement Big Data.
- les primitives ne peuvent pas supporter la mise à jour du récepteur de texte chiffré.

3.5 Protection par Détection de compression

La solution idéale de préservation de la vie privée consiste à maximiser la préservation de la vie privée ou des données sensibles, sans toutefois affecter l'environnement, la préservation de la vie privée et l'utilité des données pour parvenir à un équilibre.

La détection comprimée (CS) a été appliquée pour la première fois dans le domaine du traitement du signal numérique. En 2006, Candice et Tao, les auteurs de CS, ont fait valoir que si la matrice de capteurs était utilisée comme clé, la théorie de CS était un schéma de cryptage et les applications de cryptage de données CS commençaient à être concernées.

L'accent est mis sur la préservation de la confidentialité des données volumineuses basée sur la théorie de la CS, qui comprend principalement les contributions suivantes:

1- Conception de l'architecture et du modèle théorique préservant la confidentialité des méga données.

2- obtenir un algorithme préservant la confidentialité des données volumineuses afin de promouvoir la protection de la confidentialité des données volumineuses de la théorie à la pratique, qui a une certaine valeur de référence.

3- les résultats permettent non seulement de préserver la confidentialité dans le Big Data, mais également d'assurer l'utilisation normale des données, ce qui constitue une étape dans l'équilibre entre la préservation de la confidentialité des Big Data et les applications de données.

La conception d'une solution de conservation de la confidentialité de big data sur la base de la détection compressée utilise plusieurs étape. Au début Trois problèmes doivent être résolus:

- Premièrement, comment gérer et préserver le big data set, de manière à éviter toute perte de confidentialité, mais aussi une réduction efficace des données brutes.
- Deuxièmement, comment améliorer l'intensité de la confidentialité en préservant l'utilisation de l'analyse de corrélation, des connaissances de base et d'autres attaques contre les données personnelles.
- Troisièmement, comment ne pas affecter l'utilité normale du Big Data.

Parmi les caractéristiques du Big Data, la confidentialité du Big Data se présente principalement sous la forme de trois types:

- le premier concerne les données de confidentialité essentielles et les données sont stockées en masse.
- la seconde catégorie regroupe les données de confidentialité contenues dans les méga données distribuées.
- la troisième catégorie concerne les données de confidentialité sous forme de flux.

Pour la première catégorie, en ajoutant des données redondantes pour compléter des ensembles de données de confidentialité, en obtenant des données de confidentialité fragmentées. Et pour la deuxième catégorie, par rapport aux autres données, les données privées sont déjà rares. Troisième est sous la forme de flux de données, similaire au signal continu, à travers une certaine transformation peut rendre les données deviennent rares.

Par conséquent, les données volumineuses relatives à la confidentialité peuvent répondre à la condition préalable selon laquelle «le signal peut être une représentation fragmentée». Ils peuvent donc utiliser la théorie de la détection compressive pour préserver la confidentialité des données volumineuses. Afin de résister aux attaques des connaissances de base, ils combinent l'anonymisation et le cryptage basé sur la détection compressive pour assurer la protection de la confidentialité des données volumineuses.

L'idée de base est:

1) Pour éviter le traitement de chiffrement et de déchiffrement de blocs de données volumineux, Car la pratique a prouvé que la surcharge de traitement de chiffrement et de déchiffrement de blocs de données volumineux est très importante.

2) Essayez de transformer le Big Data en données relativement petites grâce à la transformation clairsemée, grâce à l'algorithme de détection compressif.

Ils peuvent obtenir des données de confidentialité cryptées, afin d'éviter les attaques de connaissances en arrière-plan.

3) Les données compressées et cryptées peuvent être reconstruites, n'affecte pas l'utilisation normale pour les utilisateurs autorisés.

Donc cette méthode peut réaliser la compression, le chiffrement et la reconstruction de données volumineuses, ce qui peut éviter une surcharge de calcul liée au chiffrement direct des données volumineuses brutes, réaliser efficacement la préservation de la confidentialité des données et reconstruire les données en utilisant une petite quantité de données. Mais n'affecte pas l'utilisation normale des données.

Les inconvénients :

- Les sources de données du Big Data sont hétérogènes donc Il est difficile de gérer ces sources.

- certain information peut perdre pendant l'opération de compression

3.6 Protection par enregistrement local

Les scientifiques ont suggéré un autre mécanisme d'anonymisation. Il est nommé le mécanisme de gestion des enregistrements locaux (LRDM) contenant la métrique de confidentialité et un cadre permettant d'optimiser la méthode de conservation privée de l'utilisateur. La méthodologie LRDM utilise la confidentialité différentielle pour optimiser la méthode de préservation de la confidentialité de l'utilisateur, et propose un cadre pour atteindre le mécanisme optimal de protection de la confidentialité.

Dans LRDM, chaque utilisateur (individu ou organisation) a un ensemble de données R constitué de divers enregistrements de données R = {r1, r2, …, rn}, où ri signifie que le premier enregistrement a s'est produite et le nombre total d'enregistrements est N. On note l'événement réel ea (r) = <u, r, t>, ce qui signifie que l'utilisateur (individu ou organisation) a un enregistrement réel avec l'horodatage t, et l'événement observé eo (r) = <u, r ', t > signifie le résultat observé par l'adversaire (ou fournisseur de services, Cloud, utilisateur illégal).

Les scientifiques considèrent les enregistrements locaux de l'utilisateur et la connaissance de l'adversaire, et supposent les informations suivantes:

• Dans LRDM, les utilisateurs s'attachent à protéger ses informations sensibles contenant une série d'enregistrements. Utilise ensuite une fonction de préservation de la confidentialité pour transformer l'enregistrement réel r $\in$ R en enregistrement observé r'.

• L'adversaire qui connaît le profil de l'utilisateur cherche à trouver le véritable enregistrement de l'utilisateur r, puis l'adversaire observe le résultat de LRDM r' et connaît la fonction de protection de la vie privée.

En même temps, l'adversaire a la capacité d'analyser et de déduire le véritable enregistrement de l'utilisateur, et il découvre l'estimation r', puis l'événement inférant peut être décrit dans

ei (r') = <u, r', t>.

•De plus, il existe un écart entre l'enregistrement réel de l'utilisateur r et l'enregistrement estimé de l'adversaire r', défini comme la fonction de dissimilarité d (r, r').

Dans LRDM, ils préfèrent utiliser la distance de Hamming pour formuler la fonction de dissimilarité. Ils peuvent également utiliser d'autres méthodes pour la quantifier, qui dépend du contenu spécifique de l'enregistrement. La fonction de dissimilarité quantifie la perte en cas d'intimité découlant de l'attaque par inférence. Donc la méthodologie permettant de formaliser l'objectif d'utilisateur, qui contient une série de contraintes linéaires.

Tout d'abord, ils capturent la confidentialité différentielle pour concevoir leur LRDM.

Deuxièmement, ils adoptent l'inférence bayésienne pour calculer la distribution a posteriori, l'adversaire observant l'enregistrement de sortie r' de LRDM, et il connaît la fonction LRDM s (r' | r) ainsi que le profil d'accès de l'enregistrement π (r).

Du point de vue de l'utilisateur, l'adversaire vise à sélectionner un enregistrement r' afin de minimiser la confidentialité conditionnelle attendue de l'utilisateur.et d (r', r) est une fonction de dissimilarité entre l'utilisateur et l'adversaire.

Pour plus de commodité, ils utilisent la distance de Hamming pour quantifier la dissimilarité entre l'enregistrement réel r et l'enregistrement estimé r'.

Ensuite, il est évident que l'adversaire est dédié à l'estimation d'un enregistrement afin de minimiser l'équation de confidentialité conditionnelle attendue de l'utilisateur.

Enfin, l'utilisateur cherche à maximiser la confidentialité attendue, en même temps, la fonction LRDM s satisfait la confidentialité différentielle et aide les utilisateurs à trouver un schéma approprié pour protéger leur vie privée.

Inconvénient :

Les scientifiques doivent connaître l'adversaire pour appliquer le mécanisme de gestion des enregistrements locaux (LRDM).

3.7 Vie privée différentiel

L'importance de la confidentialité des big data dans le Cloud réside dans le fait que ces données sont généralement traitées ou utilisées par des tiers. Par conséquent, dans ce cas, le problème réside dans la sélection des méthodes de confidentialité pouvant assurer une sécurité élevée pour les données qui seront traitées ou utilisées par des tiers. Pour atteindre cet objectif, la présente étude propose de combiner plusieurs méthodes. Les auteurs présentent un aperçu des méthodes de confidentialité (K-anonymat, Différentiel-confidentialité) et ils

expliquent comment combiner ces deux méthodes pour former un nouvel algorithme appelé Diff-Anonym ".

k-anonymity est particulièrement apte à protéger contre la divulgation d'identité, mais il n'est pas entièrement sécurisé contre la divulgation d'attributs.

Si l'attaquant a connaissance des personnes et que les données sont similaires sur plusieurs lignes avec le même attribut, k-anonymity n'est souvent pas en mesure de fournir une confidentialité aux personnes ciblées.

De plus, pour résoudre les inconvénients de k-anonymity, les auteurs proposent la solution avec un modèle de confidentialité différentiel, car elle pourrait garantir une confidentialité réelle pour les données et les individus.

D'autres stratégies, par exemple, l'anonymisation sont sujettes à différentes attaques.

Avec un modèle de confidentialité différentiel, presque toute attaque par quasi-identificateur pourrait être évitée car elle est basée sur un ensemble de données différent généré à chaque fois avec différentes couches de bruit dans les données d'origine.

Indépendamment du fait que l'attaquant ou l'analyste dispose d'informations supplémentaires sur l'ensemble de données cible, les couches de bruit devraient les empêcher de pouvoir interpréter les données.

De cette façon, la confidentialité des données est garantie. La violation de la confidentialité des méga données survient en raison de limitations, du taux de garantie de la confidentialité ou de la diffusion de données précises pouvant être obtenues dans l'ensemble de données.

De plus, ils ont proposé dans un nouvel algorithme Diff-Anonym combinant des modèles k-anonymat et différentiel afin de fournir un anonymat des données avec la garantie de maintenir un équilibre entre l'ambiguïté des données privées et la clarté des données générales.

Algorithme Diff-Anonym
Entrée: ensemble de données à partir de n'importe quelle taille de données pour inclure la confidentialité. Sortie: ensemble de données avec modèles de confidentialité (k-anonymous - différentiel). Étape 1: Téléchargez les données dans le cadre. Étape 2: Sélectionnez les champs d'attributs à organiser dans des nouvelles tables temporaires. Étape 3: Détectez le quasi-identifiant dans les tables temporaires. Étape 4: Divisez les tables en mini-tables. Étape 5: Appliquez k-anonymity aux mini-tables temporaires. Étape 6: Détectez des attributs égaux dans les résultats. Étape 7: Diffusez les résultats de l'application de k-anonymity. Étape 8: Ajoutez du bruit aux données qui ont déjà des attributs identiques dans les résultats. Étape 9: Recombinez les résultats dans un ensemble de données volumineuses.

Figue3.1 : l'algorithme Diff-Anonym

Dans les travaux futurs, les auteurs vont implémenter leur proposition avec deux modèles avec un algorithme amélioré pour gérer la confidentialité des données volumineuses, ce qui garantira la sécurité des données contre les attaques d'attributs et garantira la prévention de la divulgation d'identité.

3.8 Appariement Cryptographique

Les big data dans le secteur de la santé désignent des ensembles des données de santé médicales électroniques volumineuses et complexes.il est difficile de gérer ces ensembles de données à l'aide de logiciels et / ou de matériel traditionnels.

La MBD dans le secteur des soins de santé inclut les données des patients dans les dossiers électroniques des patients (RPE); données cliniques des entrées informatisées d'ordonnances de médecins (EPEC); données générées par la machine / capteurs….

La télémédecine est l'un des domaines émergents de la recherche en cyber santé, dans ce domaine, les EMRs, y compris MBD, les images et les données médicales multimédia, sont transmis à la volée via des connexions Internet non sécurisées et cette opération n'assure pas la confidentialité si pour ça les auteurs utilise une nouvelle méthode appelé Appariement Cryptographique .

Le Cryptographie basée sur l'appariement Est une méthodologie présentée pour sécuriser la MBD des patients dans le nuage de soins de santé en utilisant la technique du leurre avec une installation de calcul par brouillard ,un DMBD est créé. Cette technique peut être considérée comme une technique d'illusion, car elle permet à l'attaquant de croire qu'il a accédé au MBD de l'utilisateur alors qu'il ne s'agit en fait que d'une galerie de leurres.

Une fois que l'utilisateur a accédé à son compte, le DMBD est affiché par défaut. Ainsi, les utilisateurs autorisés et non autorisés seront référés au DMBD en tant que première étape, tandis que les utilisateurs légitimes autorisés, en tant que deuxième étape, seront référés à l'OMBD après avoir été vérifiés.

Il est plus facile de vérifier que l'utilisateur est légitime que de détecter l'attaquant. C'est pourquoi les auteurs ont d'abord essayé de le traiter en proposant le DMBD.

Lorsque l'utilisateur accède à son compte, qu'il soit un utilisateur légitime ou un attaquant, sa première étape consiste à accéder au DMBD, situé dans la couche de calcul en brouillard, à côté du profilage de l'utilisateur.

Le profilage d'utilisateur est une technique familière qui peut être appliquée pour modéliser de quelle manière, à quelle heure et de quelle manière un nombre considérable d'utilisateurs accède à leurs informations dans le Cloud de la santé.

Cette méthode de sécurité basée sur le comportement est couramment utilisée dans les applications de détection de fraude.

Le DMBD contient de faux MBD, censés faire croire à un attaquant qu'il / elle a accédé aux photos / à l'image médicale de l'utilisateur, alors qu'il s'agit en réalité d'une galerie de leurres.

L'utilisateur légitime sait déjà que la galerie à laquelle il a accédé n'est pas la sienne, elle passerait donc à l'étape suivante.

Passant à l'étape suivante, l'utilisateur légitime peut accéder à son OMBD après avoir été vérifié en réussissant le défi de sécurité.

Le défi de la sécurité peut être une question de sécurité difficile, voire un code de vérification.

Ainsi, s'il réussit le défi de la sécurité, cela signifie qu'il est l'utilisateur légitime et qu'il pourra donc accéder à l'OMBD situé sur la couche de Cloud computing.

Si l'utilisateur n'accède qu'au DMBD, un SMS ou un courrier électronique sera envoyé à l'utilisateur légitime pour l'informer de l'accès à son compte.

Le message contiendra les informations de l'attaquant (par exemple, date et heure d'accès et adresse IP).

L'agent OMBD et le DMBD doivent communiquer dans différentes situations, par exemple, lorsque l'utilisateur télécharge une nouvelle photo / image, le OMBD est censé communiquer avec le DMBD pour l'informer de l'ajout d'une nouvelle photo leurre.

Ces communications entre trois parties (l'utilisateur, l'OMBD et la DMBD) doivent être sécurisées. En conséquence, cette méthodologie garantit la sécurité à 100% des utilisateurs MBD et raccourcit le processus.

Inconvénient :

- la méthode d'Appariement Cryptographique à un mot de passe si l'utilisateur Oublie ce mot de passe Le système le considère comme un attaquant et ne peut pas entré à l'OMBD.

3.9 Préservation du vié privé dans le Cloud

Le Cloud computing est l'endroit où les big data sont stockées à l'heure actuelle, exercent une influence significative sur le secteur informatique actuel et les communautés de recherche.
Le Cloud computing fournit une puissance de calcul et une capacité de stockage considérables qui permettent aux utilisateurs de déployer des applications sans investissement en infrastructure. Les scientifiques ont suggéré le cadre MapReduce. Il a été largement adopté par un grand nombre d'entreprises et d'organisations pour traiter d'énormes volumes de données. Contrairement à la solution traditionnelle, MapReduce intégré au cloud computing devient plus flexible, évolutif et économique. Comme le cadre MapReduce est généralement adopté pour gérer les données dans des scénarios tels que le cloud hybride ou les nuages multiples, des solutions aux problèmes de confidentialité du cadre sont urgentes.

Récemment, la recherche sur les questions de confidentialité dans le cadre MapReduce sur le cloud a commencé. Des mécanismes tels que le cryptage, le contrôle d'accès, la confidentialité différentielle et l'audit sont exploités pour protéger la confidentialité des données dans la structure MapReduce.

Il Ya un cadre souple, évolutif, dynamique et rentable, préservant la confidentialité, basé sur MapReduce sur le cloud.

La structure est construite sur le dessus de MapReduce et sert de filtre pour préserver la confidentialité des ensembles des données avant que ces ensembles des données ne soient utilisés et traités par MapReduce.

Plus précisément, le cadre fournit des interfaces aux détenteurs des données afin de spécifier diverses exigences de confidentialité en fonction de différents modèles de confidentialité.

Une fois les exigences de confidentialité spécifiées, la structure lance des algorithmes d'anonymisation de la version de MapReduce afin d'anonymiser efficacement les ensembles des données pour les tâches MapReduce suivantes.

Les ensembles de données anonymes sont conservés et réutilisés pour éviter les coûts de calcul. Ainsi, le cadre préservant la confidentialité gère la mise à jour dynamique des ensembles de données afin de maintenir les exigences de confidentialité de ces ensembles de données.

Outre l'anonymisation, le cadre intègre également des techniques de cryptage afin de garantir de manière rentable la confidentialité de plusieurs ensembles de données anonymisés indépendamment en termes d'exigences de confidentialité différentes.

Le cadre préservant la confidentialité peut anonymiser des ensembles de données à grande échelle et gérer les ensembles de données anonymes de manière très flexible, évolutive, efficace et rentable.

Pour répondre aux quatre exigences du système, il Ya quatre modules pour le cadre préservant la confidentialité :

- **l'interface PSI (Privacy Specification Interface) :** Les exigences de confidentialité spécifiées par un propriétaire de données sont définies en tant que spécification de confidentialité.
  Une spécification de confidentialité est formellement représentée par un vecteur des paramètres.
- **l'anonymisation des données (DA) :** Le module d'anonymisation des données (DA) consiste en une série d'algorithmes anonymisés de la version MapReduce.
  Fondamentalement, chaque algorithme anonymisé a un programme de pilote MapReduce et plusieurs paires de programmes Map et Reduce.
- **la mise à jour des données (DU)** : les ensembles de données dans les applications sur le Cloud sont dynamiques et augmentent considérablement au fil du temps, ce qui aboutit au Big Data.

Par conséquent, ils deviennent mettre à jour les ensembles des données d'origine et anonymisés. Trois opérations de base sont fournies dans le module DU, à savoir la mise à jour, la généralisation et la spécialisation.

- **la gestion des données anonymes (ADM) :** les ensembles de données anonymes sont conservés pour le partage, l'exploration et l'analyse de données.

3.10 Tableau comparative

Approches & critères	Confidentialité	Intégrité	Disponibilité	Division	Vélocité	Contrôle d'accès	Cloud fiabilité	Agrégation	L'anonymisation	Protection de la vie privée	L'utilité
Anonymisation multi dimensionnel	(-)	(-)	(-)	oui	Oui	x	x	x	k-anonymat	élevé	faible
Anonymisation par proximité avec MapReduce	x	x	x	x	x	x	x	Oui	k-anonymat	moyen	faible
Stockage multi partagé	x	(-)	x	x	x	oui	x	x	Proxy Re-Encryption (PRE)	moyen	moyen
Protection par Détection de compression	(-)	(-)	x	x	oui	oui	(-)	x	Cryptage avec Compression	élevé	moyen
Protection par enregistrement local	x	x	x	x	oui	x	(-)	x	confidentialité différentielle	moyen	faible
Vie privée différentiel	x	oui	oui	oui	x	(-)	(-)	oui	Diff-Anonym	élevé	faible
Appariement Cryptographique	oui	oui	oui	x	(-)	x	oui	x	Technique de DMBD	élevé	Moyen
Préservation du vié privé dans le Cloud	oui	(-)	(-)	x	oui	oui	oui	x	MapReduce	élevé	moyen

Figure3.2 : Comparaison des approches

3.11 Synthèse des travaux existants

Après avoir étudié chaque approche et établi le tableau comparatif ci-dessus on constate

Les limites suivantes :

- La majorité des approches ne prennent pas en compte ni l'intégrité ni la confidentialité

Des données.

- La disponibilité est ignorée par la plupart des travaux.

- L'utilité des données est vraiment faible ou moyenne pour chaque travail

3.12 Conclusion

Ce chapitre a été consacré à la présentation de différentes approches concernant la protection de la vie privée sur les big data et d'effectuer une étude comparative de ces derniers. Dans le chapitre suivant, nous présenterons notre architecture en prenant en considération les limites déduites à partir des travaux étudiés.

Chapitre 4

Conception et modélisation

4.1 Introduction

Après avoir effectuer un survol sur les divers travaux des chercheurs pour préserver la vie privé du big data dans le chapitre précédant.

On va s'intéresser a présent dans ce chapitre à l'architecture globale de notre système pour assuré la vie privée du big data, ainsi qu'à l'architecture détaillée de chaque composant, puis on développera une modélisation détaillée avec 'UML' dans laquelle la structure globale du système est fixée.

4.2 Considérations générales

4.2.1 Cible de protection

- Protégé les informations importantes et sensibles telles que les informations des patients au but de L'attaquant n'a pas pu voir ces données sensibles.
- Stocké les informations pour assurer qu'elles restent pendant le temps dont nous avons besoin.
- .L'attaquant est incapable d'accéder aux informations.
- L'incapacité de l'attaquant à falsifié ou saboté les informations du client et éliminé
  Les dommages qui ont causés la destruction du client et de son travail, tels que les informations d'entreprise.

4.2.2 Sources d'attaques possibles

Les opérations de hacking d'acquisition de données ont, elles, augmenté de 49%. Tous les deux ans, le cabinet de conseil PriceWaterHouseCoopers (PwC) réalise une étude sur les menaces informatiques pesant sur les entreprises.

La dernière a révélé que les cyberattaques contre les sociétés ont augmenté de 38% dans le monde et de 51% en France, soit l'équivalent de 21 incidents par jour.

Il existe de nombreuses sources d'attaques telles que :

- les attaquants dans le domaine de travaille, telle que les entreprises se font concurrence dans le même domaine.
- Les pirates qui ont un passe-temps de subversion sans argent.

- Les hackers qui sont leur métier de voler des informations confidentielles en faveur des autres parties.
- les Virus qui entrent dans l'emplacement de stockage du big data et les détruisent.

4.2.3 Hypothèses

On Suppose que ce système résolve le problème de la confidentialité et maintiendra la vie privée sur les big data et l'incapacité de l'attaquant à s'approcher de l'information, quelles que soient ses tentatives.

4.2.4 Objective

La sécurité de la vie privée d'un système informatique a pour mission la protection des informations et des ressources contre toute d´évaluation, modification ou destruction.
Les objectifs pris en confédération dans la sécurité sont les suivant :

1- La confidentialité :
Permet de garder les informations secrètes de tous sauf des personnes autorisées à les consulter.
2- L'authentification :
Permet la confirmation de l'identité d'une entité avant de lui donner l'accès `à une ressource.
3- L'intégrité des informations :
Permet d'assurer que les informations n'ont pas été altérées par des personnes qui ne sont pas autorisées.
4- La disponibilité :
Permet de garantir l'accès `à un service ou une donnée.
5- Non répudiation :
Permet d'empêcher le démenti d'engagement ou de l'action précédente.

4.3 Conception générale du système proposé

4.3.1 Architecture globale

Le rôle principal de cette section est de concevoir une architecture globale pour la réalisation d'un système de protection et d'assurance de vie privée sur les Bigdata.

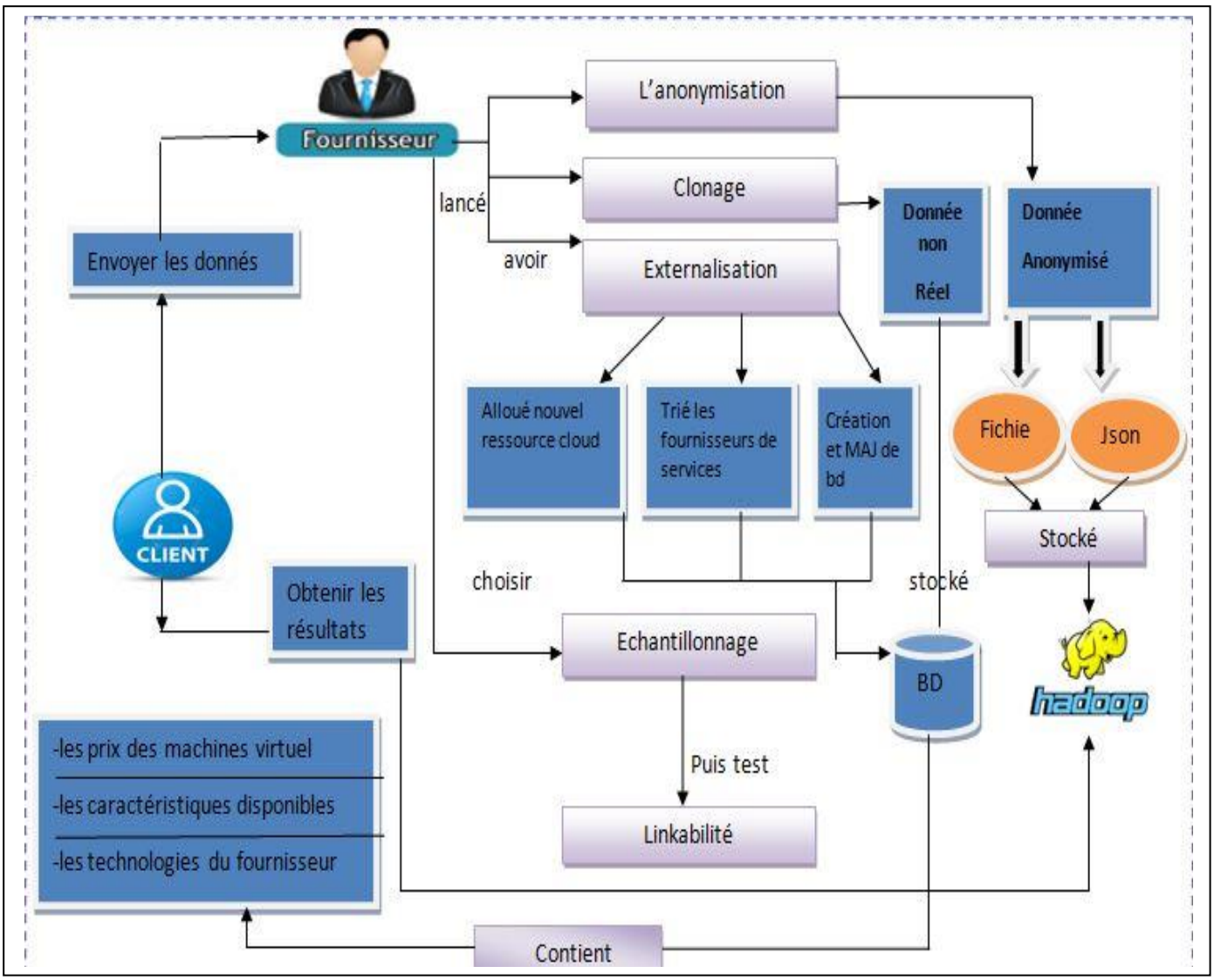


Figure 4.1 : L'architecture globale du système proposé

Description du système proposé :

Notre architecture est composée d'une collection de composants afin de trouver un bon système protégé et assuré la vie privée du big data, à cet égard on a utilisé un ensemble de composants : l'anonymisation, l'externalisation, L'échantillonnage, l'inkabilité, le clonage.

Tous ces composants travaillent en collaboration pour assurer la protection de La vie privée

Des données.

4.3.2 Architecture détaillée

4.3.2.1 Le composant anonymisation

-L'architecture du composant anonymisation

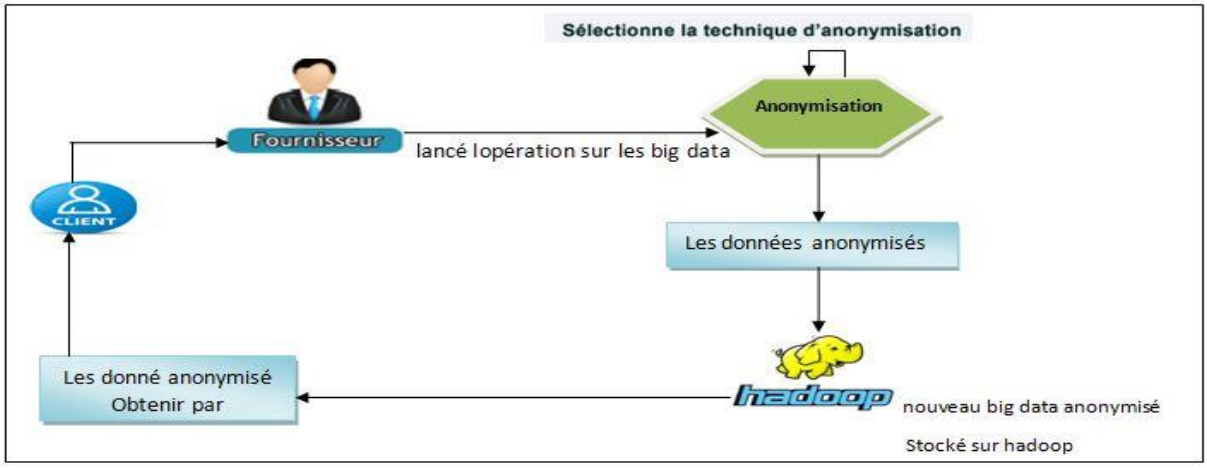


Figure 4.2 : L'architecture du composant anonymisation

- Le rôle du composant d'anonymisation :

- La réception des caractéristiques de Big Data d'après le fournisseur des données.
- Lancer l'opération d'anonymisation sur les données reçues.
- Déterminer le niveau d'anonymisation souhaité.
- Exporter les données anonymiser (sous forme de fichier texte ou JSON par exemple) et le stocké sur Hadoop.
- Demander plus de ressources au composant Externalisation en cas de besoin.

4.3.2.2 Le composant Externalisation

-L'architecture du composant Externalisation

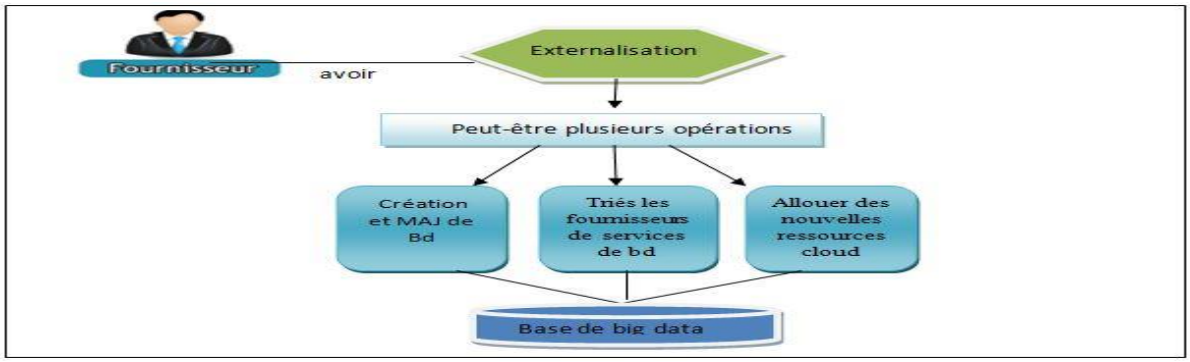


Figure 4.3 : L'architecture du composant externalisation

- Le rôle du composant d'externalisation

- La création et la mise à jours d'une base données pour stocké les caractéristique des fournisseurs cloud de stockage et de traitement comme : les prix des machines virtuel, les capacités des stockages et traitement fournis gratuitement, les caractéristique disponible et les technologies accessible sur chaque fournisseur, … etc.
- Triés les fournisseurs de services sur la base données selon leurs caractéristiques.
- Allouer des nouvelles ressources cloud à la demande des composants d'anonymisation, du composant d'échantillonnage, de linkabilité, et de clonages.

4.3.2.3 Le composant Echantillonnage & linkabilité

-L'architecture du composant Echantillonnage & linkabilité

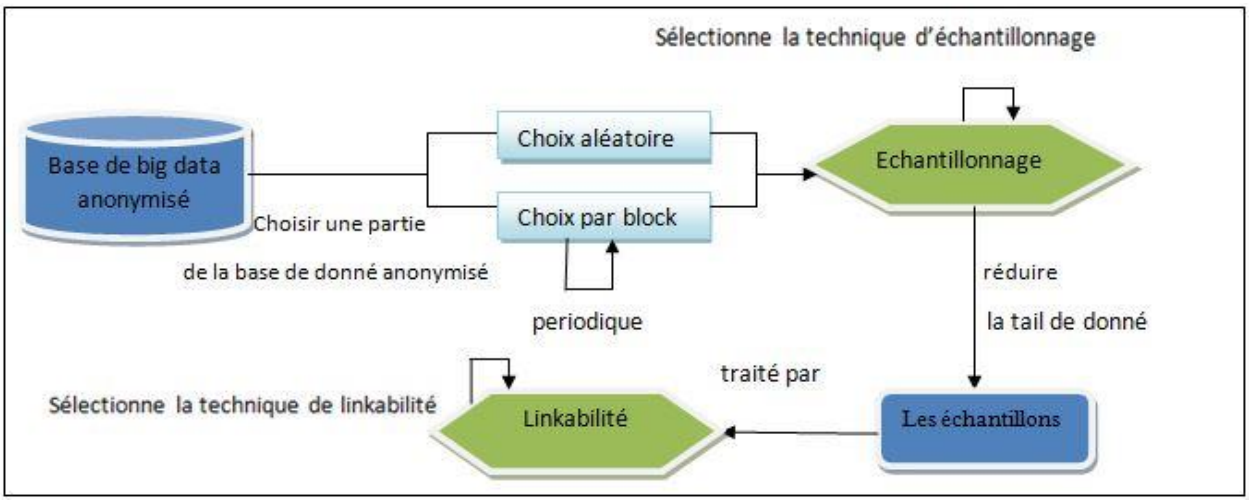


Figure 4.4 : L'architecture du composants Echantillonnage & linkabilité

- Le rôle du composant d'échantillonnage & linkabilité

- Sélectionner et fournir périodiquement des échantillons, qui représente des parties de la base de données à anonymisé… l'objectif est de réduire la taille de données (réduire l'effet du volume des bigdata) que le composant linkabilité doit traiter.
- Le choix des échantillons se fait de façon aléatoire ou par block périodiquement.
- Le composant échantillonnage fournit les échantillons au composant de linkabilité à la demande de ce dernier. L'opération de linkabilité faire la comparaison entre les donnés anonymisés et les donnés réel, Si la proportion de similarité est supérieure à 50 % donc l'anonymisation est faible si non l'anonymisation est fort.

4.3.2.4 Le composant Clonage

-L'architecture du composant Clonage

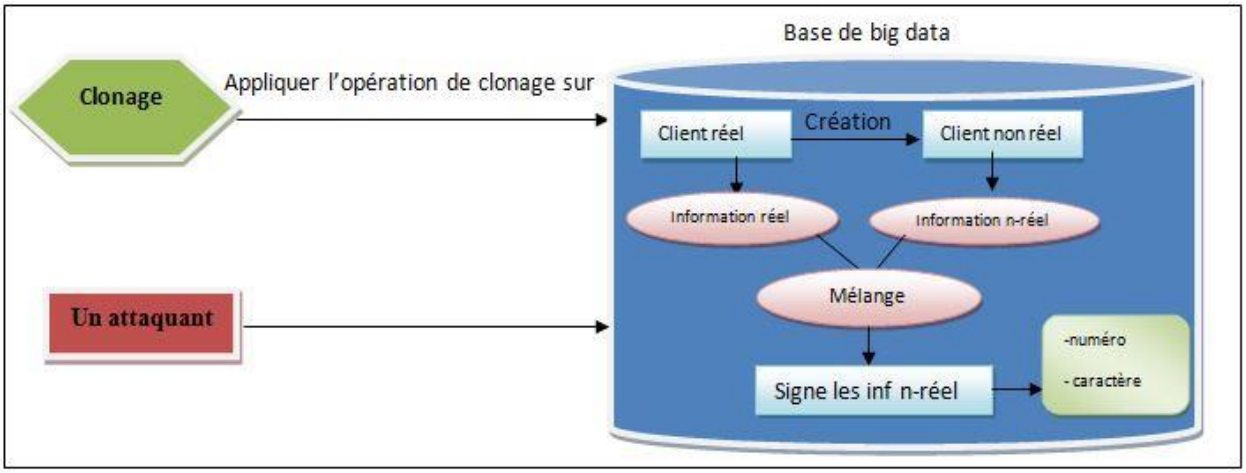


Figure 4.5 : L'architecture du composant clonage

- Le rôle du composant Clonage

- Création de nouveaux clients non-réels dans la base de données.
- Les informations des clients non réels devront être synthétisées à partir des informations des clients réels
- Marquer les clients non réels avec un signe (un numéro de série, un caractère spécial spécifique ou toute autre signe) pour les distingues des clients réels.
- Dispersé les lignes qui représente les informations des clients non réels dans la base de données pour augmenter l'efficacité de l'opération de camouflage
- L'objectif du clonage est que même si un attaquant arrive à pénétrer la base de données pour chercher des informations sur un client bien spécifique… on augmente les chance qu'il tombe sur un faux clients au lieu du clients réel ce qui permet au moins de ralentir l'attaquant afin d'avoir le temps nécessaires pour lancer des contre-mesures.

4.4 Conception et modélisation détaillée avec UML

4.4.1 Les Diagrammes de Cas d'utilisations

- Diagramme de Cas d'utilisation d'anonymisation

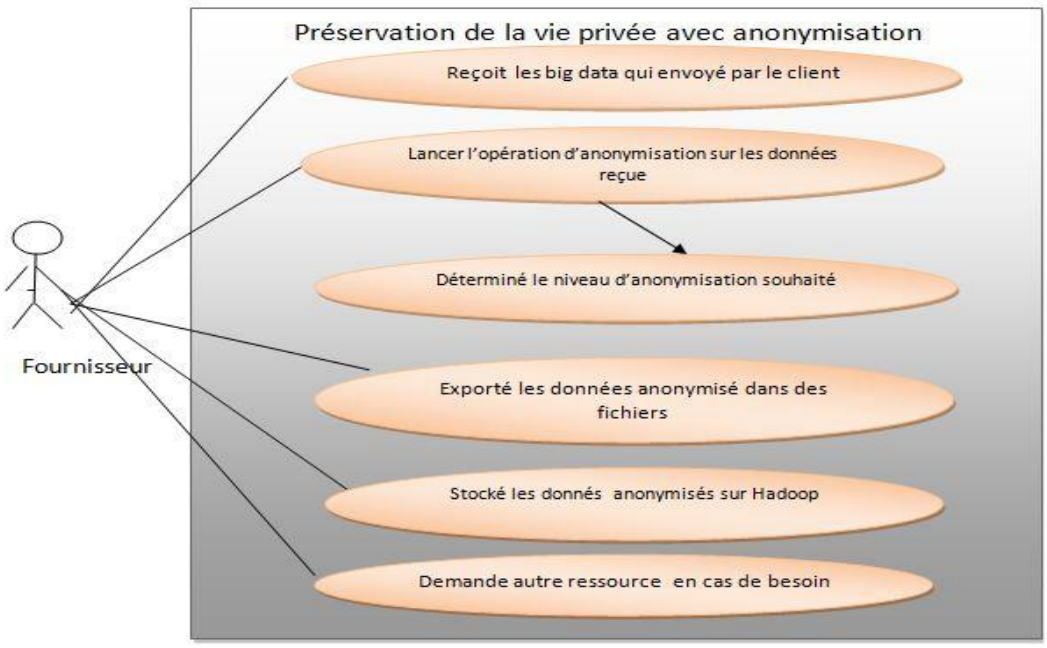


Figure 4.6: diagramme de cas d'utilisation d'anonymisation

- Diagramme de Cas d'utilisation d'Externalisation

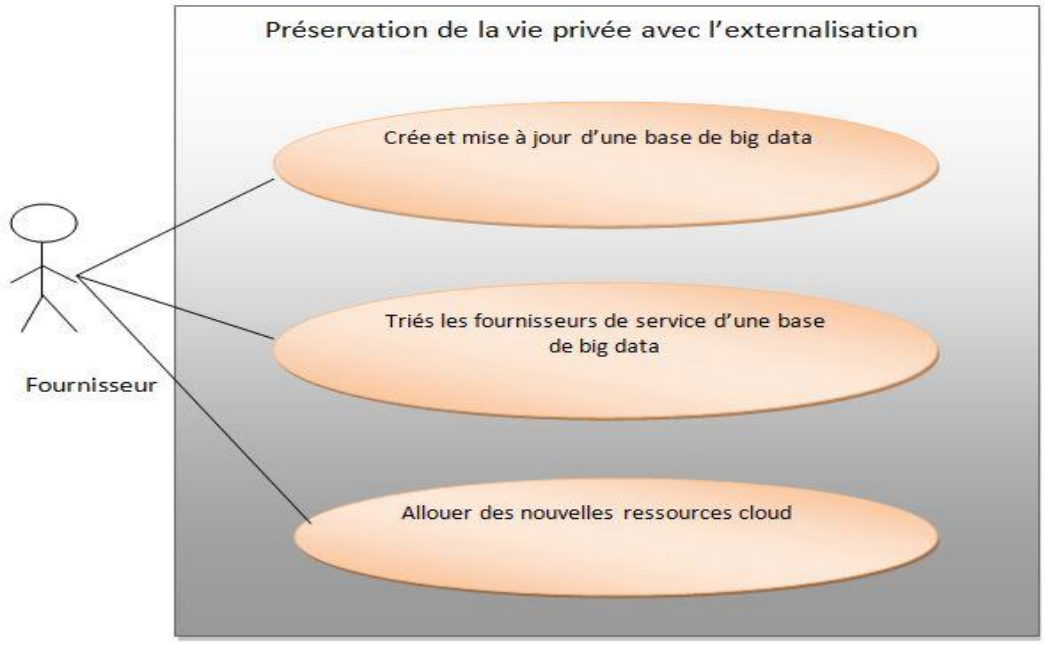


Figure 4.7 : diagramme de cas d'utilisation d'externalisation

- Diagramme de Cas d'utilisation d'Echantillonnage & linkabilité

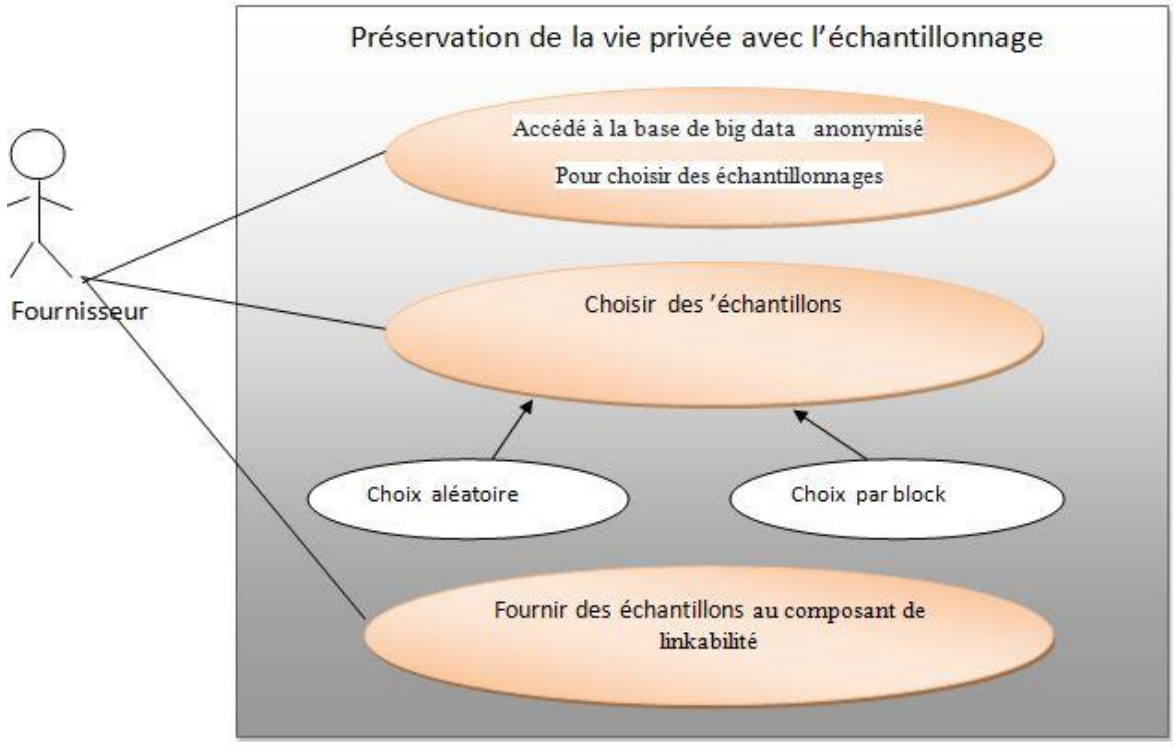


Figure 4. 8: diagramme de cas d'utilisation d'Echantillonnage & linkabilité

- Diagramme de Cas d'utilisation de clonage

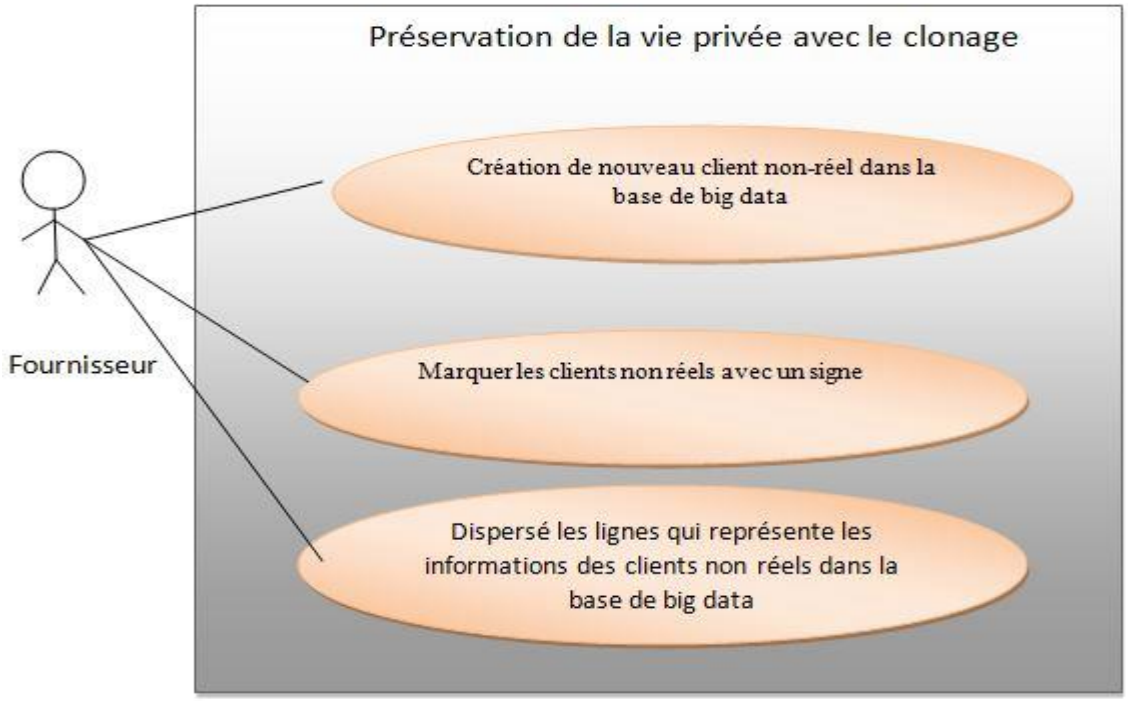


Figure 4.9 : diagramme de cas d'utilisation de clonage

4.4.2 Scénario temporelle d'exécution globale avec le diagramme de séquence

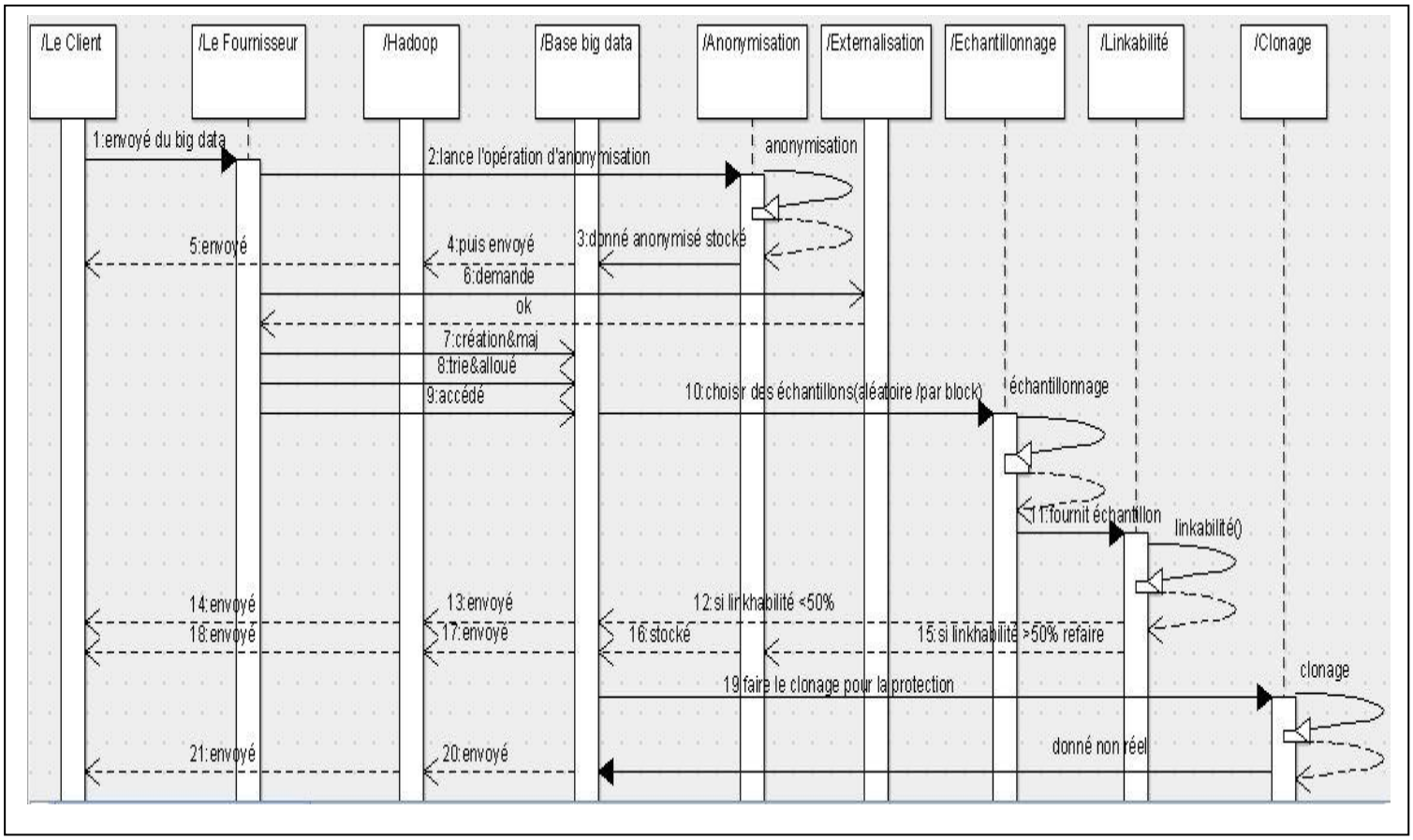


Figure 4.10: diagramme de séquence d'un Scénario temporelle d'exécution global

Description : ce diagramme de séquence permet de représente les échanges des messages entre des instances.

-le client envoie ces big data au fournisseur, ce dernier reçoit les données et faire quelque opération pour préservé la vie privée de ces donnés.

- le fournisseur Lance l'opération d'anonymisation sur les données reçus.

- Exporter les données anonymiser sous forme de fichier texte ou JSON par exemple et le stocké sur Hadoop.

- il Demande autre ressource comme la création et la mise à jour d'une base données pour stocker les caractéristiques des fournisseurs cloud de stockage et de traitement, triés les fournisseurs de services sur la base données selon leurs caractéristiques, Allouer des nouvelles ressources cloud à la demande des composant d'anonymisation, du composants d'échantillonnage, de linkabilité, et de clonages

-pour faire l'opération d'échantillonnage il accède à la base de big data anonymisés, Puis choisir des échantillons (choix aléatoire ou par block périodiquement), qui représente des parties de la base de données anonymisé, l'objectif est de réduire la taille des données.

- Le composant échantillonnage fournit les échantillons au composant de linkabilité.

-le principe d'opération clonage est la création de nouveaux clients non-réels dans la base de big data, Le but est de distraire l'attaquant et augmente les chances qu'il tombe sur un faux client.

4.4.3 Architecture détaillée avec les diagrammes d'activité

-Diagramme d'activité du composant anonymisation

Ce composant, a pour but de Masquer les données afin de protéger la vie privée des utilisateurs.

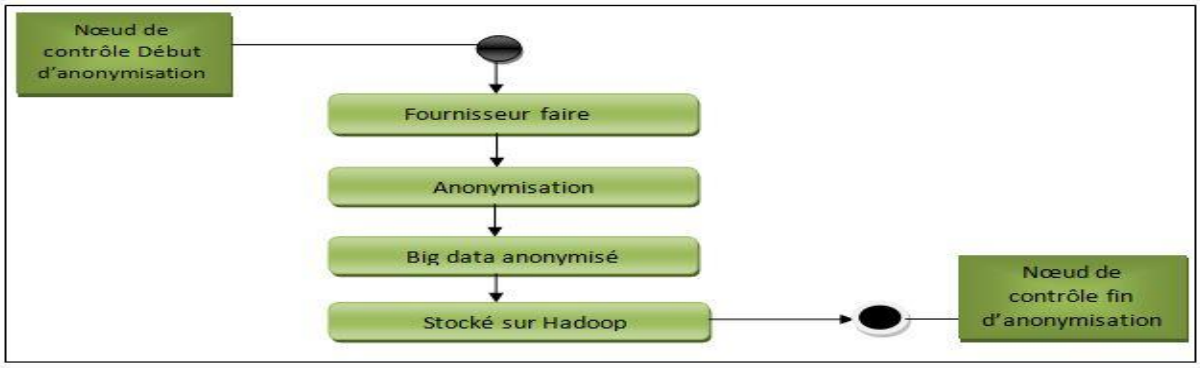


Figure 4. 11: diagramme d'activité du composant anonymisation

-Diagramme d'activité du composant d'externalisation

Le fournisseur peut faire plusieurs opérations comme la création et la mise a jour d'une base de donnés et aussi le trie des fournisseurs par apport a la propriété.

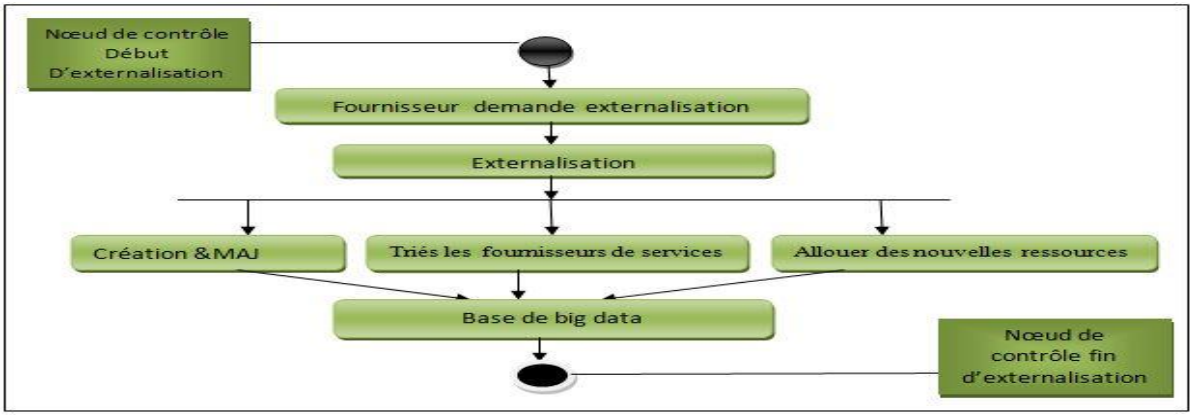


Figure 4.12 : diagramme d'activité du composant externalisation

- Diagramme d'activité du composant d'échantillonnage & linkabilité

Le but du composant d'échantillonnage pour réduire la taille du big data, le composant de linkabilité pour voir si l'anonymisation fort ou faible.

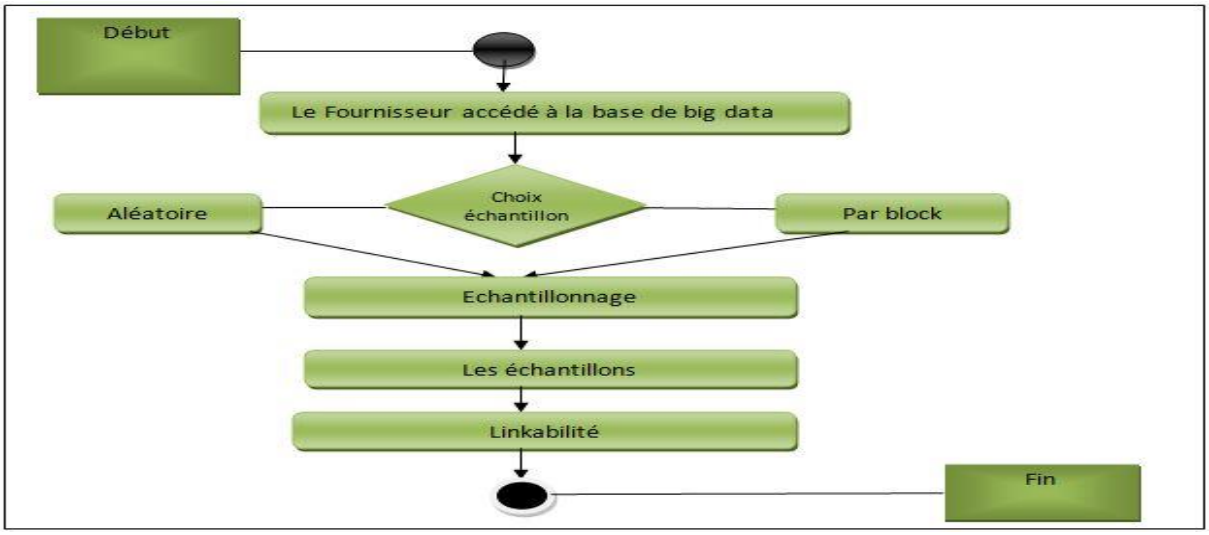


Figure 4.13 : diagramme d'activité du composant d'échantillonnage & linkabilité

- Diagramme d'activité du composant clonage

Le but de clonage pour ajouté un bruit au données afin de protéger la vie privée des utilisateurs

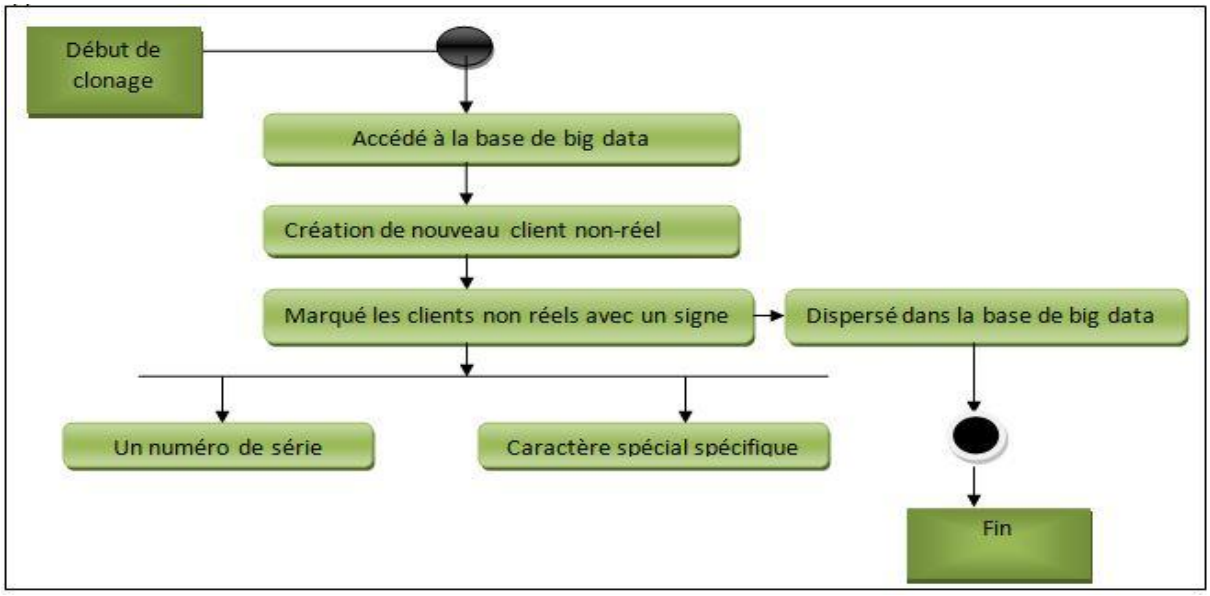


Figure 4.14 : diagramme d'activité du composant clonage

4.5 Projection sur Hadoop

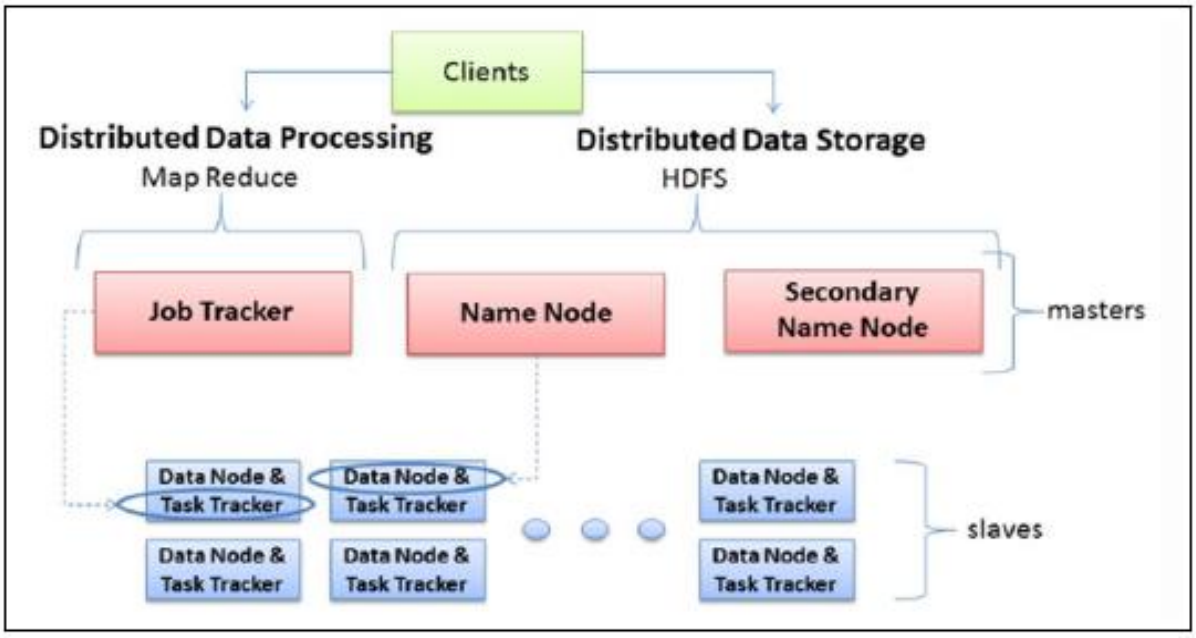


Figure 4.15 : Composants du noyau Hadoop

Les trois principales catégories de la machine dans un déploiement Hadoop sont les machines client, les nœuds maîtres 'Masters' et les nœuds esclaves 'Slave'. Les nœuds maîtres supervisent les deux pièces fonctionnelles clés qui composent 'Hadoop' : stockant beaucoup de données 'HDFS' et exécutant des calculs parallèles sur toutes ces données 'MapReduce'.
Le Name Node supervise et coordonne la fonction de stockage de données 'HDFS', tandis que 'JobTracker' supervise et coordonne le traitement parallèle des données à l'aide de'MapReduce'.
Les nœuds esclaves constituent la grande majorité des machines et font le stockage des données et exécutent les calculs. Chaque slave tourne à la fois un 'DataNode' et 'TaskTracker' qui communique et reçoit des instructions de leurs nœuds maître.

4.5.1 NameNode

Le 'NameNode' dans 'Hadoop' est le nœud où 'Hadoop' stocke toutes les informations de localisation des fichiers dans 'HDFS'.

4.5.2 Secondary NameNode

Le 'secondary Namenode' est responsable de l'exécution des fonctions d'entretien périodiques pour le 'NameNode'. Il ne créer que des points de contrôle du système des fichiers présents dans le 'NameNode'.

4.5.3 DataNode

Le 'DataNode' est chargé de stocker les fichiers dans 'HDFS'. Il gère les blocs de fichiers dans le nœud. Il envoie des informations au 'NameNode' sur les fichiers et les blocs stockés dans ce nœud et répond au 'NameNode' pour toutes les opérations du système de fichiers.

4.5.4 JobTracker

'JobTracker' est chargé de prendre des demandes d'un client et l'attribution des 'TaskTrackers' avec les 'Task' à effectuer. Le JobTracker tente d'assigner des tâches à 'TaskTracker' sur le 'DataNode' où les données sont présentés localement (Data Localité). Si cela est impossible, il va au moins essayer d'assigner des 'Task' à 'TaskTrackers' dans le même 'rack'. Si, pour une raison quelconque, le nœud échoue au 'JobTracker' affecte la tâche à 'TaskTracker' où la réplique des données existe depuis les blocs de données sont reproduits

à travers les 'DataNodes'. Cela garantit que le travail ne manque pas même si un nœud échoue au sein du cluster.

4.5.5 TaskTracker

TaskTracker est un démon qui accepte Task (Map, Reduce and Shuffle) de la JobTracker. Le TaskTracker continue à envoyer un message de heartbeat à un JobTracker de notifier qu'elle est vivante. Avec le rythme cardiaque il envoie aussi les emplacements libres disponibles à l'intérieur pour traiter des tâches. TaskTracker démarre et surveille le Map & Reduce Tasks et envoie progrès / informations d'état vers le Job Tracker. Le système externe travaille avec le HDFS (NameNode, DataNode), et le système interne : l'agent scanner travaille avec Secondary NameNode et Acces level agent travaille avec MapReduce (JobTracker, TaskTracker).

4.6 Comment le système proposé répond aux inconvénients des travaux connexes ?

Nous avons présenté notre système de titre « la Conception et la réalisation d'un système de protection et d'assurance de vie privée sur les Bigdata ».

Ce système proposé répond aux inconvénients des travaux connexes de chapitre Précédent comme la confidentialité peut traiter avec l'anonymisation, Le grand volume de donnée et son stockage peut traiter avec l'échantillonnage, Le composant échantillonnage fournit les échantillons au composant de linkabilité a la demande de ce dernier qui faire la comparaison entre les donnés anonymisés et les donnés réel , Si la proportion de similarité est supérieure à 50 % donc l'anonymisation est faible si non l'anonymisation et fort.

Pour l'opération de camouflage d'un attaquant peut utiliser le clonage ….

4.7. Conclusion

Dans ce chapitre nous avons présenté notre système de la préservation de la vie privée sur les big data. Cette étude conceptuelle présente l'architecture générale de notre travail. Dans le prochain chapitre nous allons présenter les techniques utilisées pour implémenter l'application. Ainsi qu'une étude de cas sur un exemple concret.

Chapitre 5

Implémentation du système

5.1 Introduction

Après avoir présenté en détails notre système dans le chapitre précédent, ce chapitre sera Consacré à la phase d'implémentation. Nous aborderons l'aspect pratique de notre application, il s'agit ici d'expliquer l'environnement matériel sur lequel notre système a été développé, les langages de programmation et les outils/technologies utilisés. Pour terminer, nous allons présenter les interfaces graphiques en décrivant les différentes fonctionnalités de notre application et nous présenterons aussi un exemple réel sur des données de patients que nous avons généré.

5.2 Environnement de développement

Avant de commencer l'implémentation de notre application, nous allons tout d'abord spécifier les langages de programmation et les outils utilisés qui nous ont semblé être un bon choix vu les avantages qu'ils offrent.

5.2.1 Environnement matériel et logiciel

Pour réaliser notre système, nous avons un PC I3 doté de Windows 10 (64 bits) qui est décrit avec la Figure suivante :

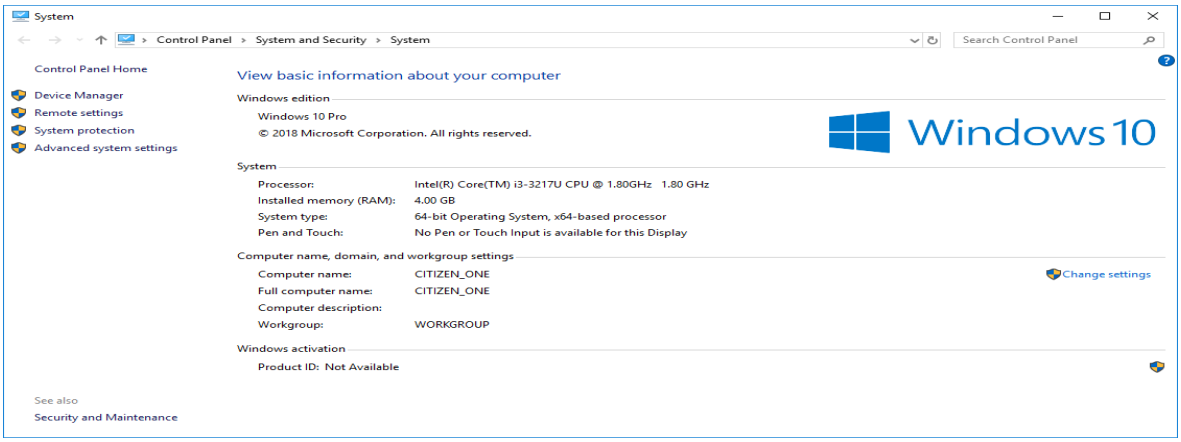


Figure 5.1 : Environnement matériel et logiciel

5.2.2 Outils et langages de programmation utilisés

Pour l'élaboration du système conçu, nous avons utilisé un ensemble de langages de programmation, et quelques environnements de développement. Nous les décrivons brièvement ci-dessous.

5.2.2.1 Langages de programmation

Le langage JAVA : Le langage Java est un langage de programmation et une plate-forme informatique évoluée et orientée objet qui est criée par James Gosling et Patrick Naughton, employés de Sun Microsystems, avec le soutien de Bill Joy (cofondateur de Sun Microsystems en 1982), présenté officiellement le 23 mai 1995 au SunWorld. La société Sun

a été ensuite rachetée en 2009 par la société Oracle qui détient et maintient désormais Java. Aujourd'hui, Java rassemble derrière lui une large communauté d'acteurs informatiques majeurs tels que HP, IBM, Oracle, Borland [Java]. Il est rapide, sécurisé et fiable. En outre, beaucoup d'applications et de sites Web ne fonctionnent pas si Java n'est pas installé et leur nombre ne cesse de croître chaque jour. À cause de sa simplicité, sa robustesse, sa portabilité ainsi que sa performance lui ont permis d'être le choix préféré pour le développement de notre application. [19]

5.2.2.2 Outils et technologies

Netbeans IDE : Nous avons écrit notre application en Netbeans version 8.2, le choix de Netbeans était fondamental puisqu'il est un logiciel permettant principalement le développement en Java. Mais aussi il permet également de supporter différents autres langages, comme C, CSS, XML et HTML. Il comprend toutes les caractéristiques d'un IDE moderne (éditeur en couleur, projets multi-langage, refactoring, éditeur graphique d'interfaces et de pages Web). Il fournit un environnement standard de développement pour créer des interfaces très puissantes.

Netbeans est un environnement de développement intégré (IDE) basé sur des normes, en une plate-forme d'application cliente riche, qui peut être utilisée comme structure générique pour créer n'importe quel type d'application avec une plus grande assurance de robustesse et de concevoir des applications qui résisteront à l'épreuve du temps. Il est placé en open source par Sun en juin 2000 sous licence CDDL (common development and Distribution license).).
Netbeans est disponible sous Windows, Linux et d'autres systèmes d'exploitation. Le projet de Netbeans IDE consiste en un EDI Open Source complet écrit dans le langage de programmation Java. [20]

Figure 5.2 : Logo de Netbeans

Hadoop : Hadoop est un Framework libre et open source écrit en Java destiné à faciliter la Création d'applications distribuées (au niveau du stockage des données et de leur traitement) et échelonnables (scalables) permettant aux applications de travailler avec des milliers de nœuds et des pétaoctets de données. Ainsi chaque nœud est constitué de machines standards regroupées en grappe. Tous les modules de Hadoop sont conçus dans l'idée fondamentale que les pannes matérielles sont fréquentes et qu'en conséquence elles doivent être gérées automatiquement par le Framework.
Hadoop a été inspiré par la publication de MapReduce, Google FS et BigTable de Google.

Hadoop a été créé par Doug Cutting et fait partie des projets de la fondation logicielle Apache Depuis 2009. Le noyau d'Hadoop est constitué d'une partie de stockage : HDFS 1, et d'une partie de traitement appelée MapReduce. Hadoop fractionne les fichiers en gros blocs et les distribue à travers les nœuds du cluster. [21]

Figure 5.3: Logo de Hadoop

MySQL & phpMyAdmin :

- MySQL : MySQL est un système de gestion de bases de données relationnelles (SGBDR). Il fait partie des logiciels de gestion de base de données les plus utilisés au monde. MySQL fait référence au Structured Query Language, le langage de requête utilisé.
- phpMyAdmin : PhpMyAdmin est une interface d'administration pour le SGBD MySQL. Il est écrit en langage PHP et s'appuie sur le serveur HTTP Apache.

Java FX : Java FX est un ensemble de graphiques et de packages multimédia permettant aux développeurs de concevoir, créer, tester, déboguer et déployer des applications client riches fonctionnant de manière cohérente sur diverses plates-formes.

Figure 5.4 : logo de java fx

5.3 Présentation des interfaces graphiques

5.3.1 Les interfaces de connexion et inscription

Afin de bénéficier de l'ensemble des services fournis par notre système, L'utilisateur doit fournir l'ensemble des informations requises et le mode d'utilisation (fournisseur/client)

Il est nécessaire de créer un compte pour le client Pour visualiser et télécharger les bases de données anonymisé.

Il ya 4 fournisseur qui peuvent entre au système et faire tous les opérations possibles .Toute inscription incomplète ne sera pas validée.

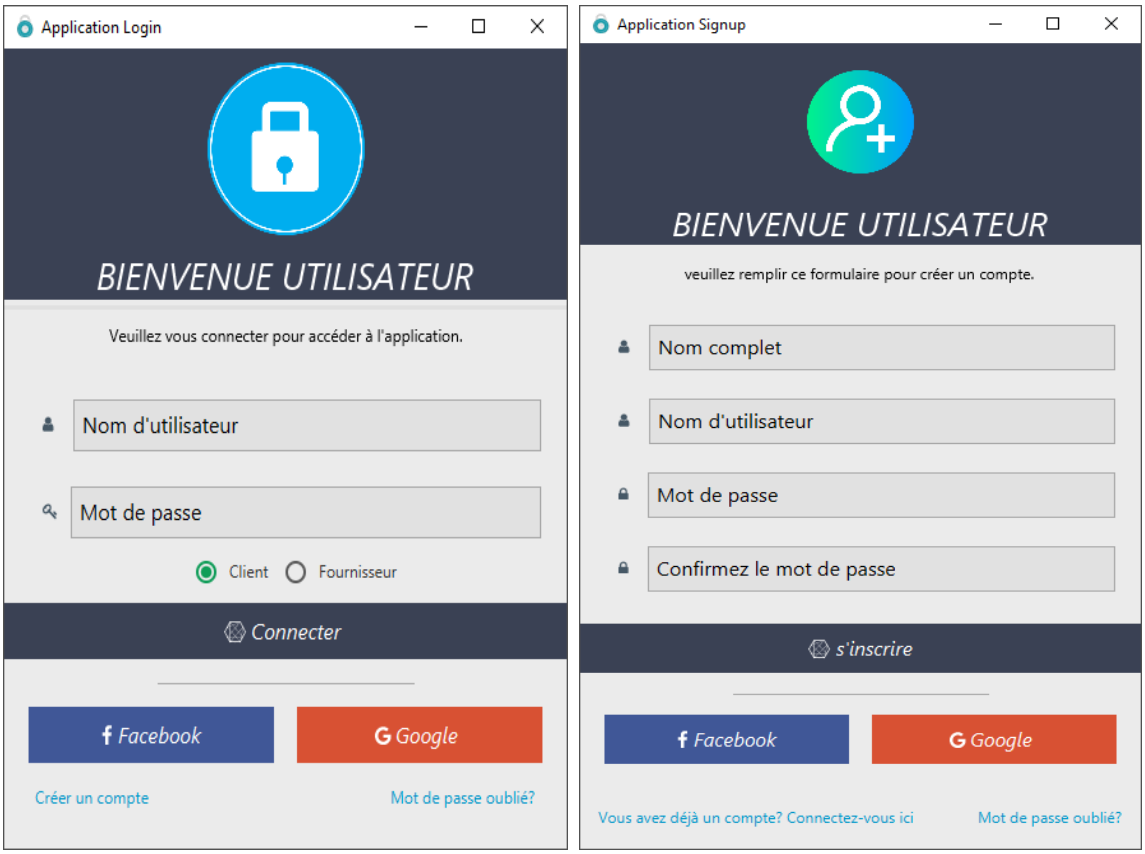


Figure 5.5 : Interface de connexion et inscription

5.3.2 Interface principale du fournisseur

L'interface principale du fournisseur est représentée sur la Figure (5.6). Il lui permet d'accéder rapidement aux services fournis par le système, il peut faire toutes les opérations possible, ces derniers sont discutés ci-dessous.

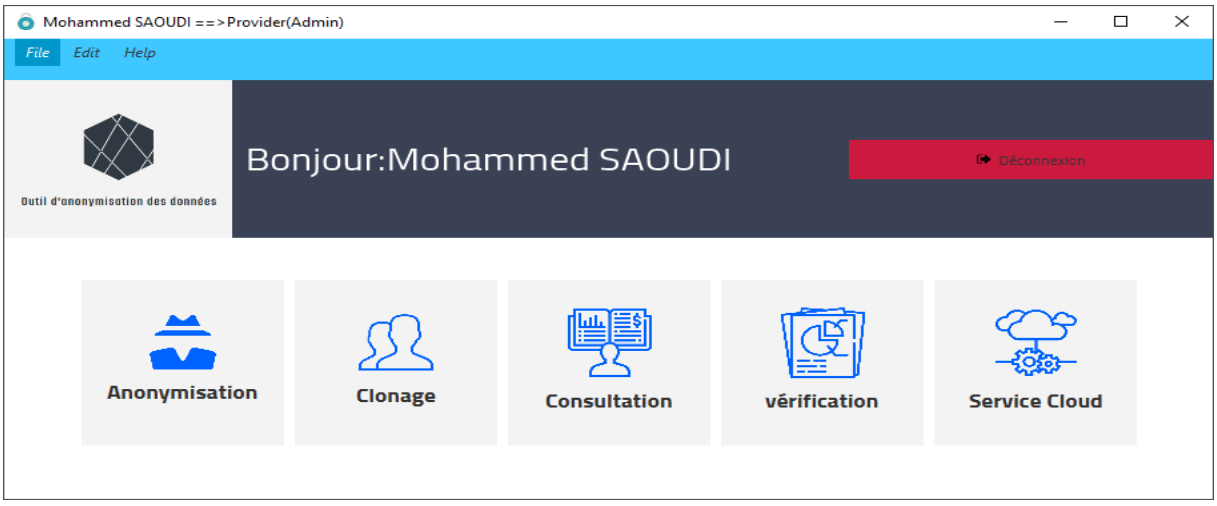


Figure 5.6 : Interface principale du fournisseur

5.3.3 Service “anonymisation”

Dans cette interface il ya les données avant faire l’anonymisation.

Il ya 2 méthodes d’anonymisation (k-anonymity et k-anonymity-généralisation).

Le fournisseur doit choisir une seule méthode et cliqué sur le bouton anonymisé

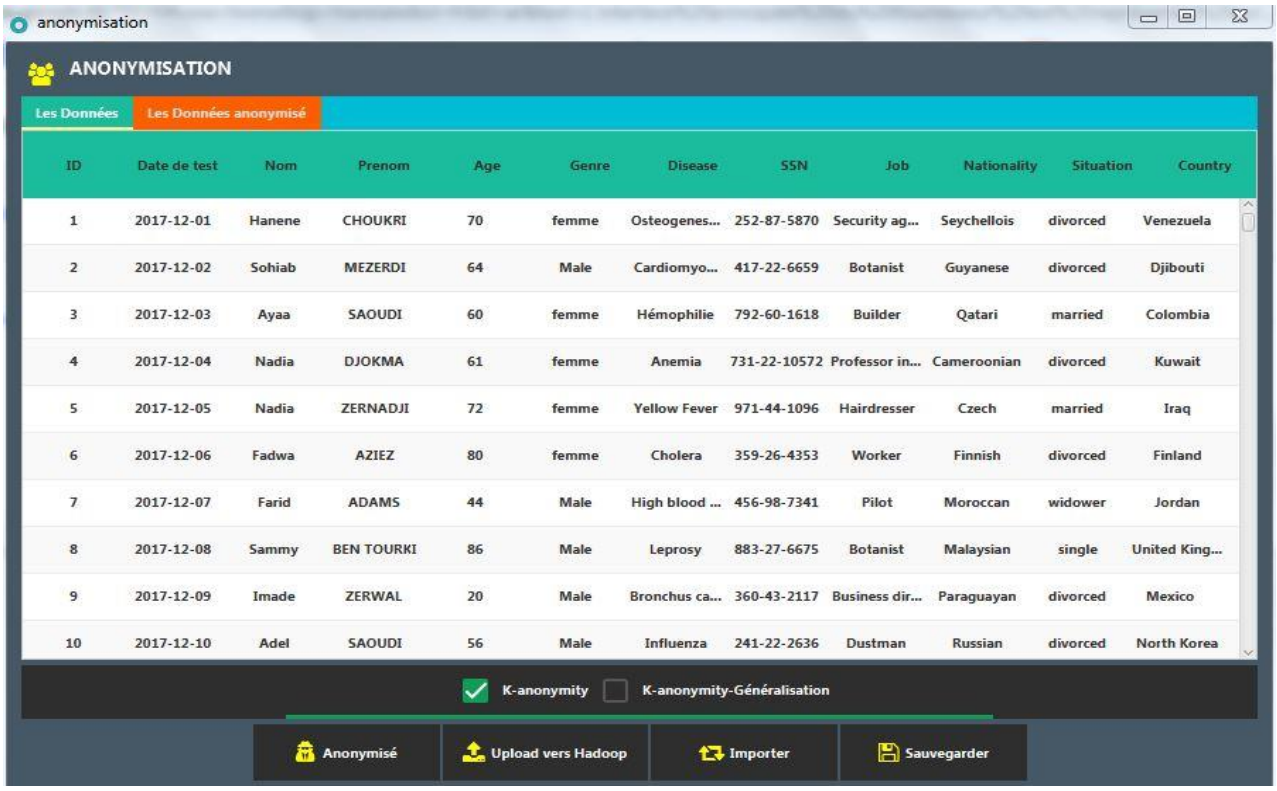


Figure 5.7 : interface d’anonymisation

Après avoir appuyé sur le bouton d’anonymisation il faut faire l’évaluation pour spécifier

Quels sont les attributs identifiant, quasi-identifiant et les attributs insensibles. Puis

Confirmer pour voir les donnés anonymisé. L’interface d’évaluation représentée a la figure (5.8)

Figure 5.8 : interface évaluation

Après avoir fini cette étape, notre système produit un nouveau “Big Data” anonymisé selon

L’évaluation fournis par le fournisseur comme indiqué sur la Figure (5.9).

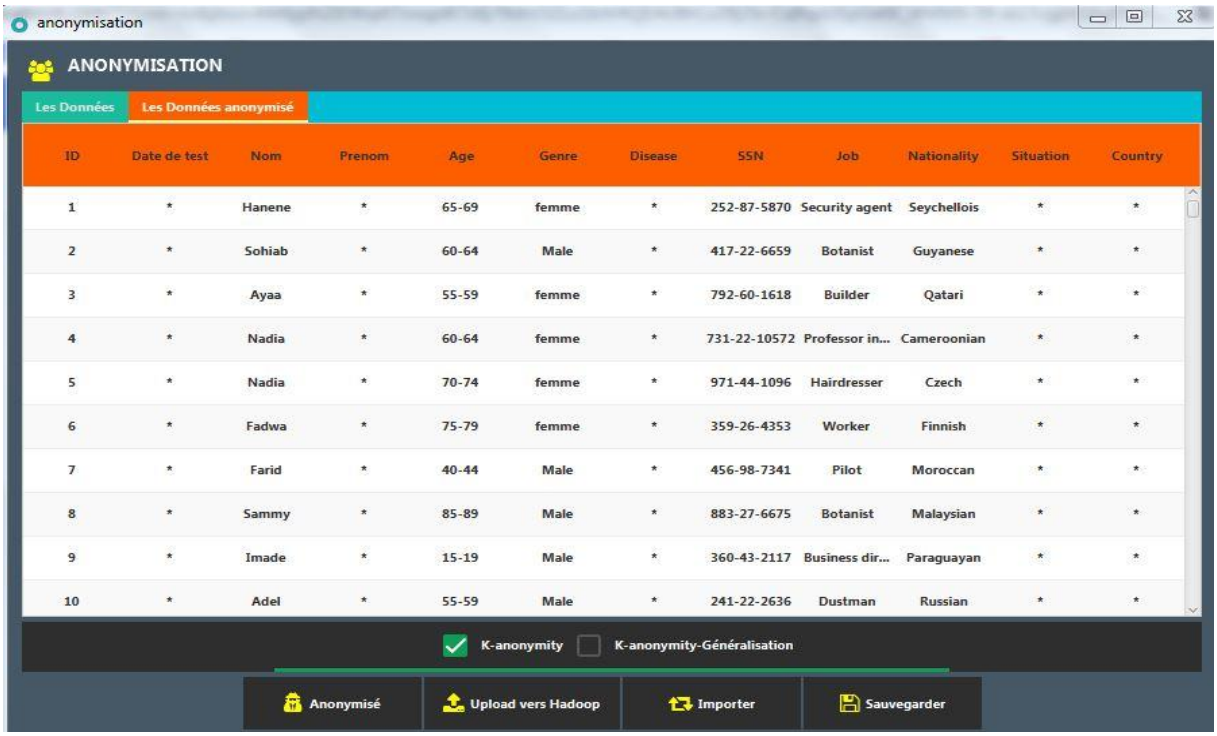


Figure 5.9 : les données anonymisés

Le fournisseur peut sauvegarder les informations internes anonymisé et aussi il peut importer des fichiers externes pour faire l’anonymisation.

Il peut upload les informations anonymisés vers Hadoop pour stocké.

5.3.4 Service “clonage”

Dans cette interface il ya les données avant faire le clonage.

Le fournisseur doit mètre le numéro de clone puis clique sur le bouton clonage pour faire

L’opération de clonage.

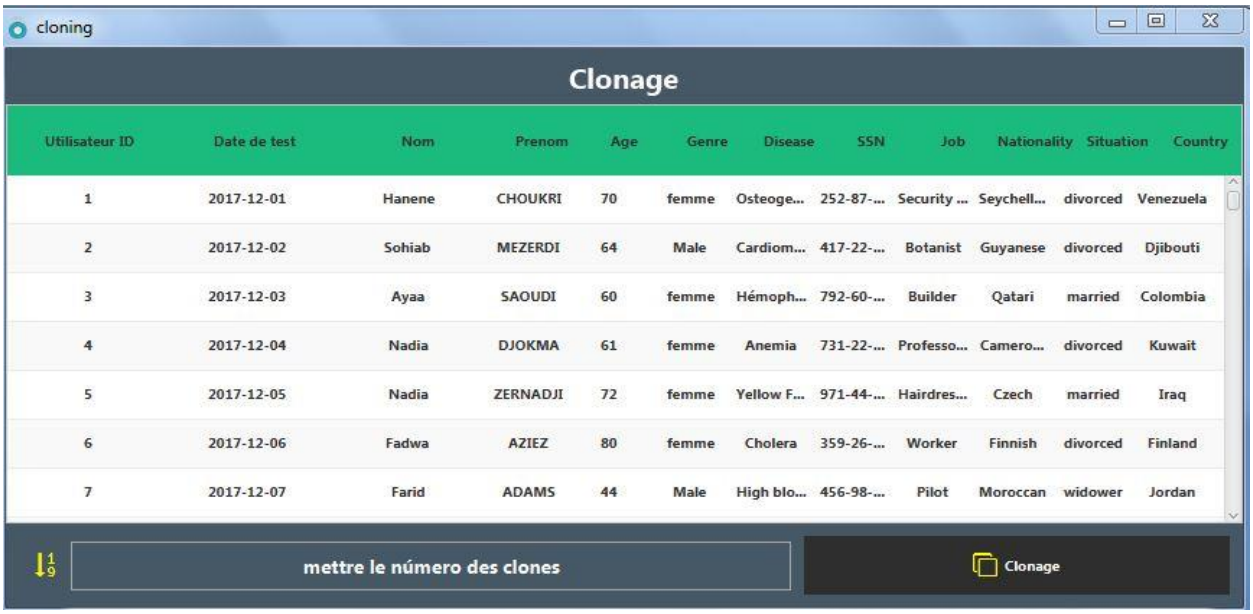


Figure 5.10 : interface clonage

La figure (5.11) représente les utilisateurs normaux avant faire le clonage.

Figure 5.11 : utilisateurs normales

La figure (5.12) représente les utilisateurs avec clonage. Ils ont des informations différentes sur leurs informations d'origine.

Figure 5.12 : utilisateurs avec clonage

Le fournisseur peut choisir des échantillons (Le choix est aléatoire) pour faire la comparaison entre les utilisateurs normaux et les utilisateurs avec clonage (la comparaison avec le bouton test linkabilité).

Si la proportion de similarité est supérieure à 50% alors le clonage est faible si non le clonage est fort. Comme représente la figure (5.13).

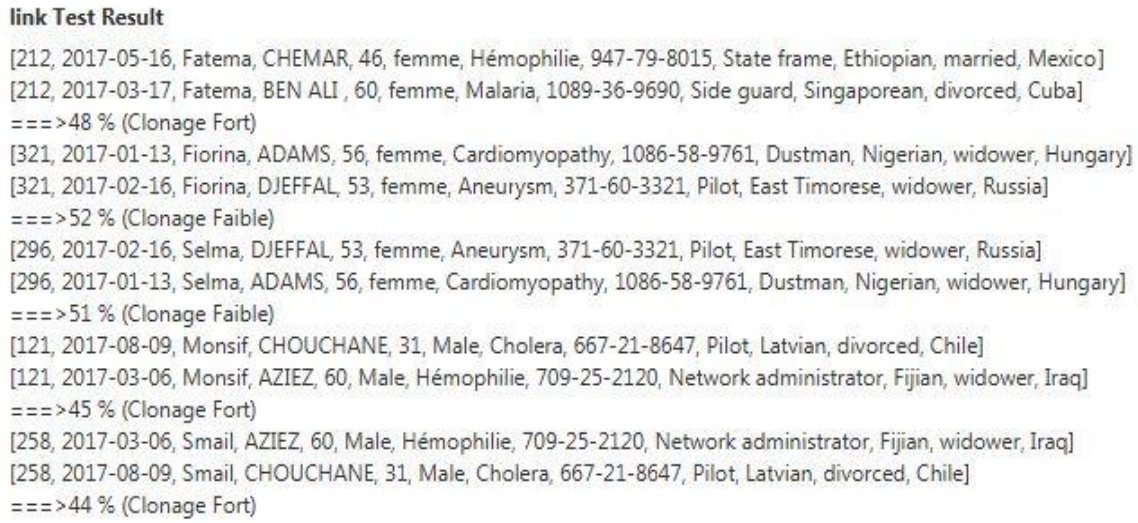


Figure 5.13 : test linkabilité de clonage

5.3.5 Service “cloud”

L’interface illustrée sur la Figure (5.14) permet de lister les meilleurs fournisseurs cloud. Le Meilleur fournisseur qui a le prix le plus bas, Le plus grand nombre possible de vm et d’espace de stockage.

Figure 5.14 : service cloud

5.3.6 Service “vérification”

Dans l’interface vérification le fournisseur peut choisir des échantillons et faire le test linkabilité entre les utilisateurs normal et les utilisateurs anonymisé

Figure 5.15 : interface vérification

Dans Le test linkabilité Si la proportion de similarité est supérieure à 50% alors l'anonymisation est faible si non l'anonymisation est forte. Comme représente la figure (5.16).

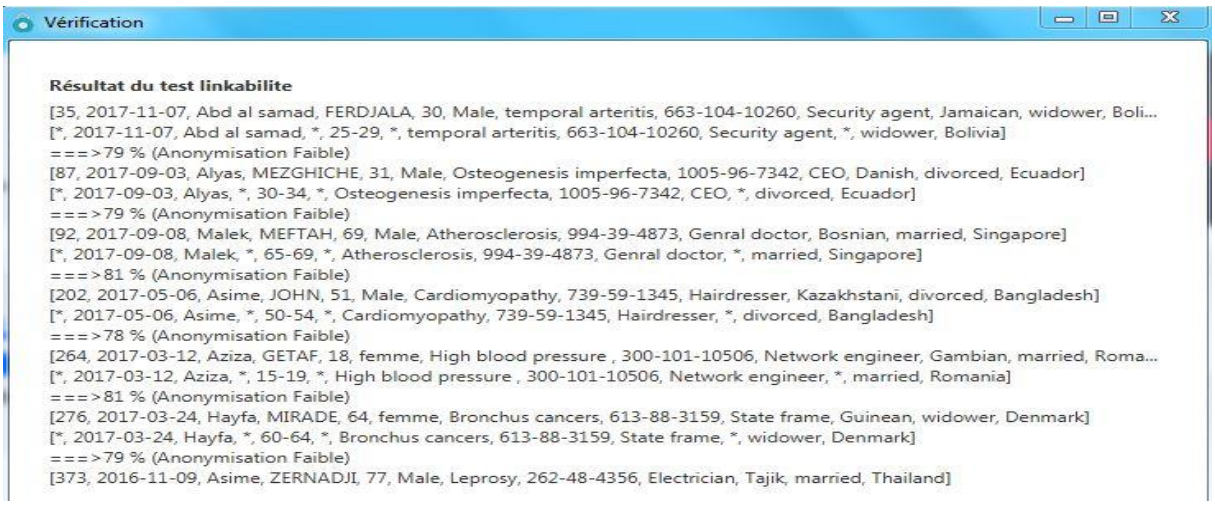


Figure 5.16 : test linkabilité de l'anonymisation

5.3.7 Interface principale du client

L'interface principale du client est représentée sur la Figure (5.17). Elle lui permet d'afficher tous les Big Data anonymisés disponibles, tous les fournisseurs et les branches médicales, le client peut télécharger n'importe quelle base de donné anonymisé.

Figure 5.17 : interface client

5.4 Hadoop et les principaux codes source

5.4.1Hadoop

Après l'installation, la configuration et le lancement du Hadoop (V-3.1.2) sur notre système D'exploitation, quatre fenêtres de CMD s'ouvrent qui sont les démons qui fonctionnent pour Hadoop d'une façon permanente afin d'assurer le bon fonctionnement DataNode, NameNode, NodeManager, RecourceManager (voir les Figures [5.18, 5.19, 5.20, 5.21]). Chacune d'eux à son rôle qui a déjà été expliqué dans le chapitre précédent. (Voir la Section [4.5]).

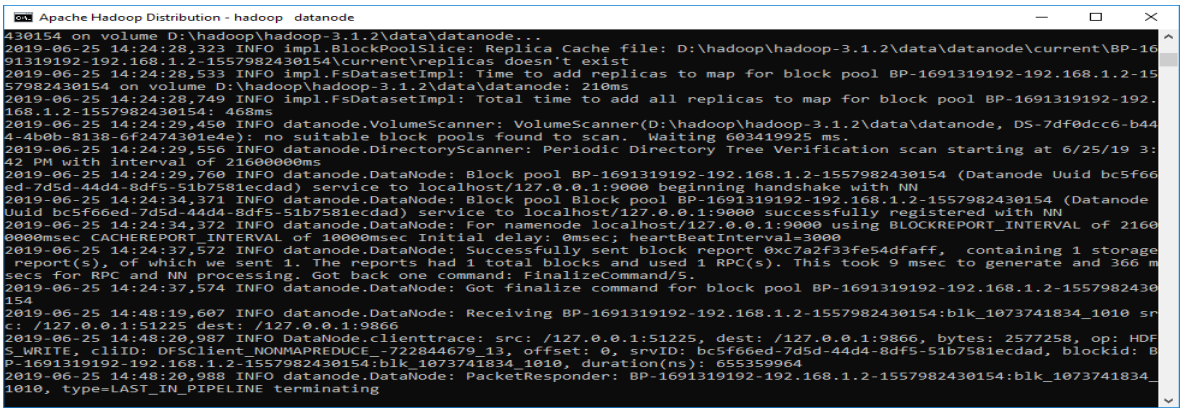


Figure 5.18: Interface CMD du DataNode

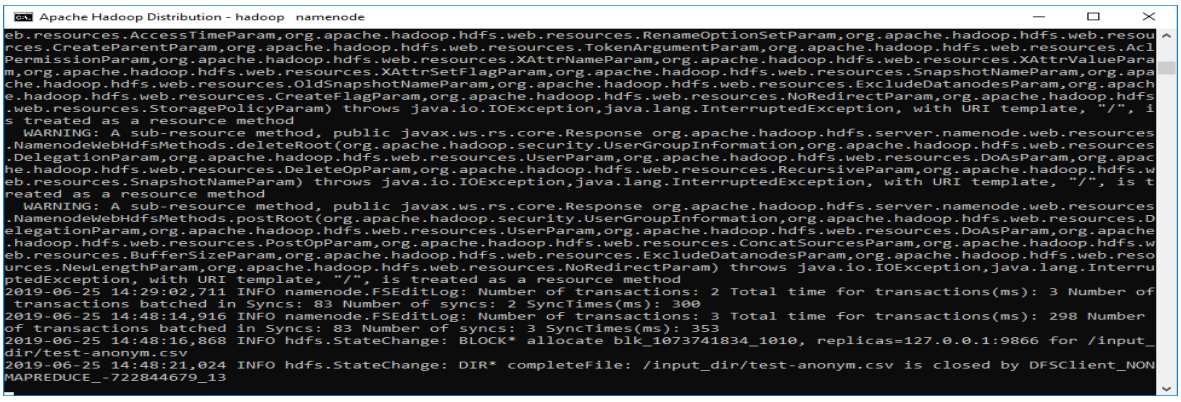


Figure 5.19: Interface CMD du NameNode

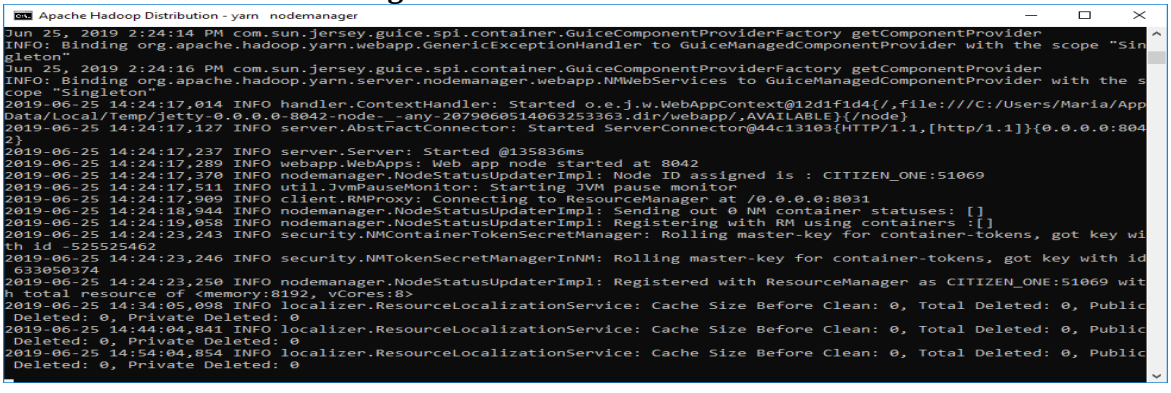


Figure 5.20: Interface CMD du NodeManager

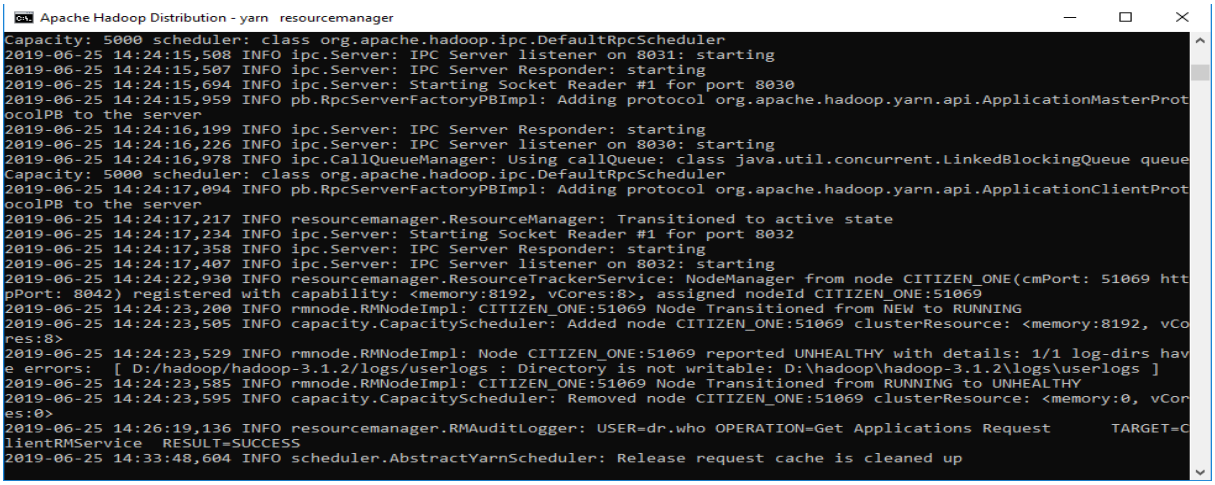


Figure 5.21: Interface CMD du RecourceManager

L'accès à la plateforme Hadoop est sur localhost : avec le numéro de port : 9870 voir la Figure (5.22). Le système de gestion de fichiers de Hadoop, HDFS, fournit un stockage de Données évolutif, tolérant aux pannes, écrit et lit les fichiers par blocs de 64 Mo par défaut. La Figure (5.23) montre le dossier qui a les fichiers transférés par le fournisseur. La figure (5.24) représente les fichiers transférés par le fournisseur.

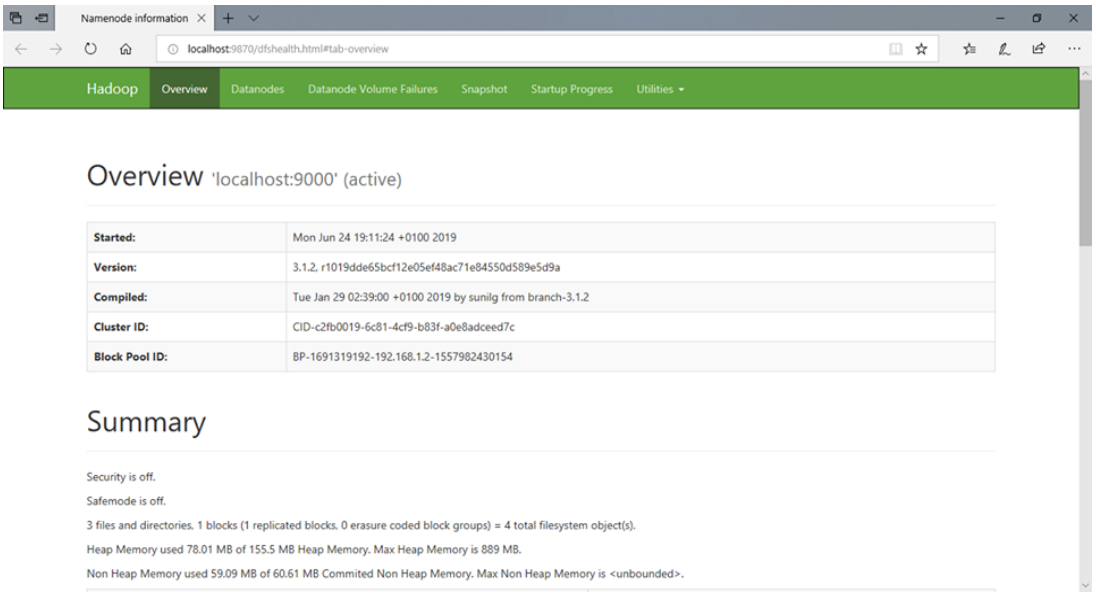


Figure 5.22 : Interface de la plateforme Hadoop « 1 »

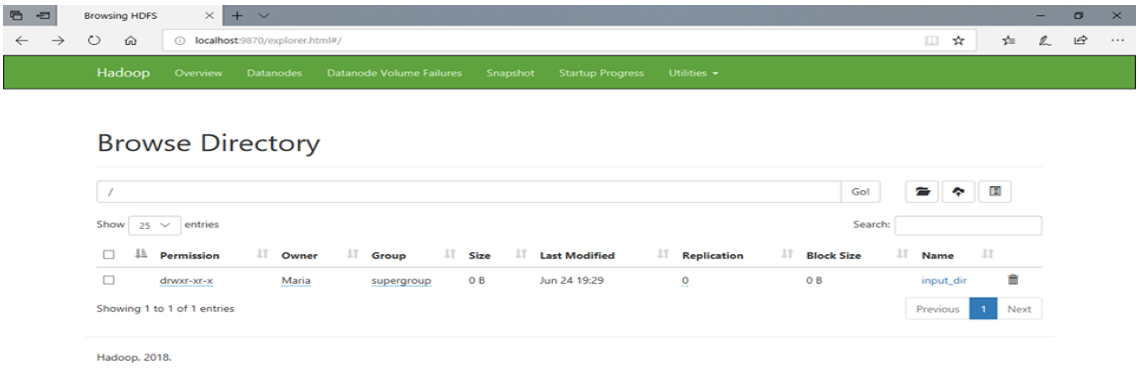


Figure 5.23 : Interface de la plateforme Hadoop « 2 »

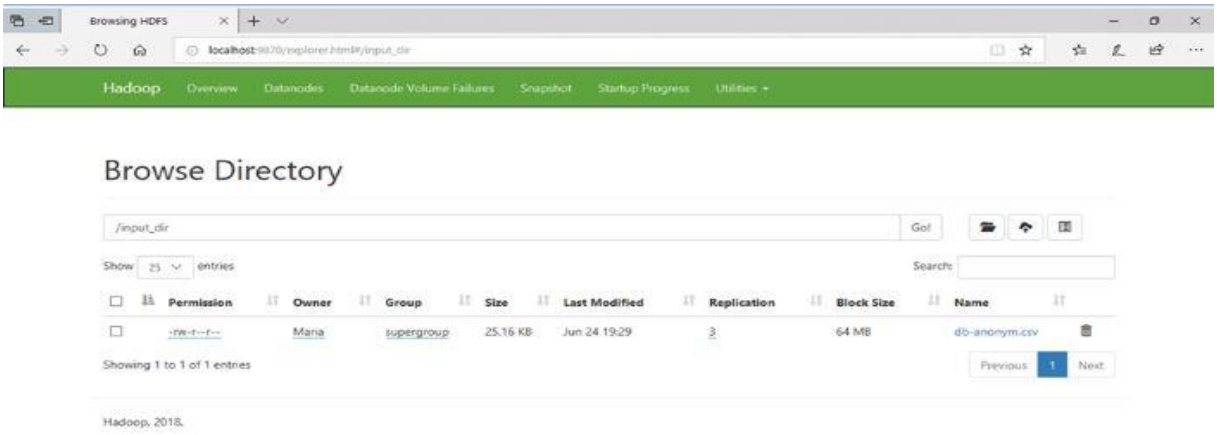


Figure 5.24 : Interface de la plateforme Hadoop « 3 »

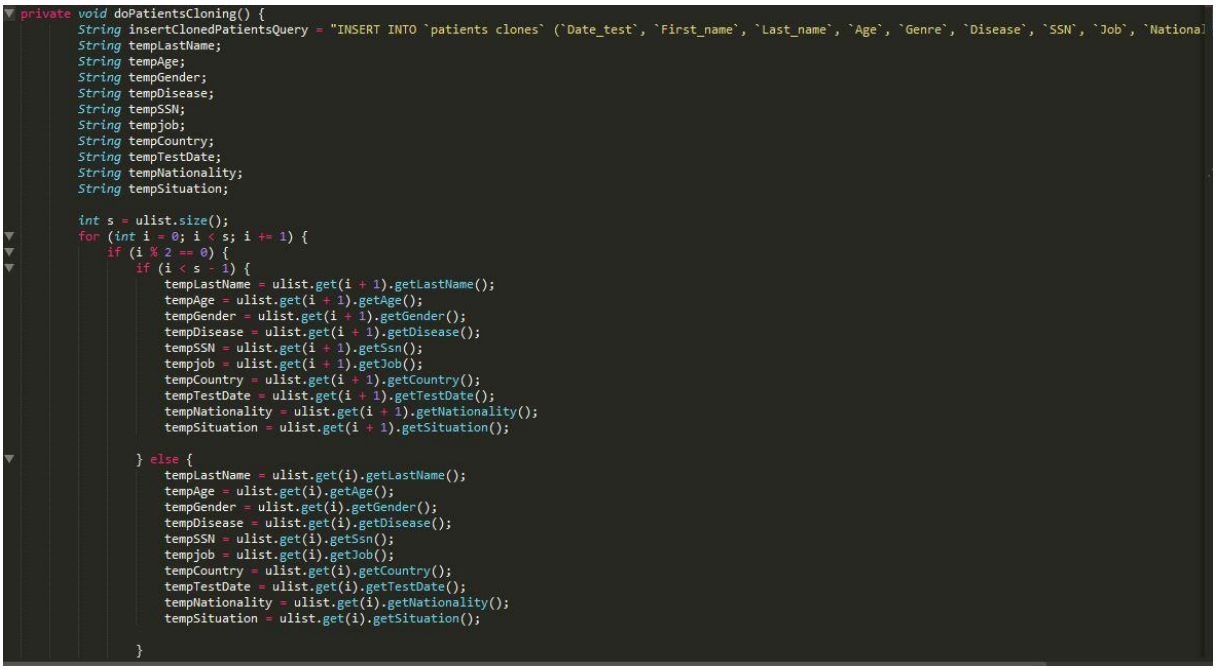
5.4.2 Les principaux codes source

```
private void performTheAnonymization() {
    anonymUsersListTV.getItems().clear();

    if (LOADED_DATA_TYPE.equals(INTERNAL_DATA)) {
        try {
            ObservableList<Patient> uls = Util.loadUsersDatabase(Util.PATIENT_INFO_SELECT_QUERY);
            loadDataEvalGUI(dataAttributes, "Evaluer les données");
            if (DataEvaluationController.attributesDefinitionList != null) {

                data = Data.create();
                data.add("id-patient", "date-test", "first-name", "last-name", "age", "genre",
                        "disease", "ssn", "job", "nationality", "situation", "country");

                uls.forEach((p) -> {
                    data.add(String.valueOf(p.getUserId()), p.getTestDate(), p.getFirstName(), p.getLastName(), p.getAge(),
                            p.getGender(), p.getDisease(), p.getSsn(), p.getJob(),
                            p.getNationality(), p.getSituation(), p.getCountry());
                });
                internalDataDefinition();
                // Create an instance of the anonymizer
                anonymizer = new ARXAnonymizer();
                // Execute the algorithm
                config = ARXConfiguration.create();
                config.addPrivacyModel(new KAnonymity(K_ANONYMISATION));
                config.setSuppressionLimit(0.02d);
                result = anonymizer.anonymize(data, config);

                for (int i = 0; i < result.getOutput().getNumRows(); i++) {
                    anonymUsersListTV.getItems().add(
                            FXCollections.observableArrayList(
                                    Arrays.asList(
```

Figure 5.25 : anonymisation part 1

```
                                    Arrays.asList(
                                            result.getOutput().getValue(i, 0),
                                            result.getOutput().getValue(i, 1),
                                            result.getOutput().getValue(i, 2),
                                            result.getOutput().getValue(i, 3),
                                            result.getOutput().getValue(i, 4),
                                            result.getOutput().getValue(i, 5),
                                            result.getOutput().getValue(i, 6),
                                            result.getOutput().getValue(i, 7),
                                            result.getOutput().getValue(i, 8),
                                            result.getOutput().getValue(i, 9),
                                            result.getOutput().getValue(i, 10),
                                            result.getOutput().getValue(i, 11)
                                    )
                            )
                    );
                }
                AlertMaker.showMaterialDialog(root, containerPane, Arrays.asList(new JFXButton("Okay!")),
                        "Résultat du lanonymisation",
                        "Données anonymisées avec succès");
            } else {
                AlertMaker.showErrorMessage("Aucune définition de données n'a été fournie",
                        "Erreur est survenue", (Stage) root.getScene().getWindow());

            }

        } catch (Exception e) {
            AlertMaker.showErrorMessage(e.getMessage(),
                    "Erreur est survenue", (Stage) root.getScene().getWindow());
            //e.printStackTrace();
        }

    } else if (LOADED_DATA_TYPE.equals(EXTERNAL_DATA)) {

        if (csvFile != null) {

            externalDataAnonymizationMethod(csvFile);
        } else {
            AlertMaker.showErrorMessage("Veuillez d'abord sélectionner un fichier",
                    "Erreur est survenue", (Stage) root.getScene().getWindow());

        }
```

Figure 5.26 : anonymisation part 2



```
private void doPatientsCloning() {
        String insertClonedPatientsQuery = "INSERT INTO `patients clones` (`Date_test`, `First_name`, `Last_name`, `Age`, `Genre`, `Disease`, `SSN`, `Job`, `National
        String tempLastName;
        String tempAge;
        String tempGender;
        String tempDisease;
        String tempSSN;
        String tempjob;
        String tempCountry;
        String tempTestDate;
        String tempNationality;
        String tempSituation;

        int s = ulist.size();
        for (int i = 0; i < s; i += 1) {
            if (i % 2 == 0) {
                if (i < s - 1) {
                    tempLastName = ulist.get(i + 1).getLastName();
                    tempAge = ulist.get(i + 1).getAge();
                    tempGender = ulist.get(i + 1).getGender();
                    tempDisease = ulist.get(i + 1).getDisease();
                    tempSSN = ulist.get(i + 1).getSsn();
                    tempjob = ulist.get(i + 1).getJob();
                    tempCountry = ulist.get(i + 1).getCountry();
                    tempTestDate = ulist.get(i + 1).getTestDate();
                    tempNationality = ulist.get(i + 1).getNationality();
                    tempSituation = ulist.get(i + 1).getSituation();

                } else {
                    tempLastName = ulist.get(i).getLastName();
                    tempAge = ulist.get(i).getAge();
                    tempGender = ulist.get(i).getGender();
                    tempDisease = ulist.get(i).getDisease();
                    tempSSN = ulist.get(i).getSsn();
                    tempjob = ulist.get(i).getJob();
                    tempCountry = ulist.get(i).getCountry();
                    tempTestDate = ulist.get(i).getTestDate();
                    tempNationality = ulist.get(i).getNationality();
                    tempSituation = ulist.get(i).getSituation();

                }
```

Figure 5.27: clonage part 1

```
        } else {
            tempLastName = ulist.get(i - 1).getLastName();
            tempAge = ulist.get(i - 1).getAge();
            tempGender = ulist.get(i - 1).getGender();
            tempDisease = ulist.get(i - 1).getDisease();
            tempSSN = ulist.get(i - 1).getSsn();
            tempjob = ulist.get(i - 1).getJob();
            tempCountry = ulist.get(i - 1).getCountry();
            tempTestDate = ulist.get(i - 1).getTestDate();
            tempNationality = ulist.get(i - 1).getNationality();
            tempSituation = ulist.get(i - 1).getSituation();

        }
        anonymUsersListTV.getItems().add(
                FXCollections.observableArrayList(
                        Arrays.asList(
                                String.valueOf(ulist.get(i).getUserId()),
                                tempTestDate,
                                ulist.get(i).getFirstName(),
                                tempLastName,
                                tempAge,
                                tempGender,
                                tempDisease,
                                tempSSN,
                                tempjob,
                                tempNationality,
                                tempSituation,
                                //ulist.get(i).getSituation(),
                                tempCountry
                        )
                )
        );
        DatabaseHndlr.getInstance().execAction(insertClonedPatientsQuery + "('" + tempTestDate + "', '" + ulist.get(i).getFirstName() + "', '" + tempLastName + "
                + tempAge + "', '" + tempGender + "', '" + tempDisease + "', '" + tempSSN + "', '" + tempjob + "', '" + tempNationality + "', '" + ulist.get(i).g

    }


}
```

Figure 5.28: clonage part 2

```
private void linkTestButtonAction() {
      String finalResult = "";
      System.out.println("linkTestButtonAction");

      if (samplingOn && randomSelectedGroup.length > 0) {
          Arrays.sort(randomSelectedGroup);

          int s = randomSelectedGroup.length;
          NormalizedLevenshtein l = new NormalizedLevenshtein();

          for (int i = 0; i < s; i++) {
              int k = randomSelectedGroup[i];
              String str1 = ulist.get(k).toString();
              String str2 = anonymUsersListTV.getItems().get(k).toString();
              double res = l.distance(str1, str2);
              int percentageInt = (int) ((1 - res) * 100);
              String percentage = percentageInt+ " %";

              if (percentageInt > 50) {
                  finalResult += str1 + "\n" + str2 + "\n===>" + percentage + " (Clonage Faible)\n";
              } else {
                  finalResult += str1 + "\n" + str2 + "\n===>" + percentage + " (Clonage Fort)\n";

              }
          }
          AlertMaker.showMaterialDialog(root, containerPane, Arrays.asList(new JFXButton("Okay!")),
                  "link Test Result",
                  finalResult);
          samplingOn = false;

      } else {
          usersListTV.getSelectionModel().clearSelection();
          anonymUsersListTV.getSelectionModel().clearSelection();
          samplesNumber.setText("00");
          AlertMaker.showErrorMessage("s'il vous plaît prendre un échantillon", "Erreur est survenue", (Stage) samplesNumber.getScene().getWindow());

      }

  }
```

Figure 5.29: test linkabilité

```
private void createAndUploadToHadoopHDFS() throws Exception {

        File localFile = fileChooser.showOpenDialog(root.getScene().getWindow());

        if (localFile != null) {
            String filePath = localFile.getPath();
            System.out.println(filePath);
            String hdfsDirUri = "hdfs://localhost:9000/input_dir";
            String uri = "hdfs://localhost:9000/";
            Configuration conf = new Configuration();

            FileSystem fs = FileSystem.get(URI.create(uri), conf);
            fs.mkdirs(new Path(uri));
            fs.copyFromLocalFile(new Path(filePath), new Path(hdfsDirUri));

            AlertMaker.showMaterialDialog(root, containerPane, new ArrayList<>(),
                    "Félicitations",
                    "La base de données anonymisée a été téléchargée sur hadoop avec succès.");

        } else {

            AlertMaker.showMaterialDialog(root, containerPane, new ArrayList<>(),
                    "Erreur de fichier",
                    "Sélectionnez un autre fichier s'il vous plaît.");

        }

}
```

Figure 5.30: upload to Hadoop

```
private void downloadButtonAction(ActionEvent event) {
    System.out.println("downloadButtonAction");
    if (this.newValue != null) {

        System.out.println("Downloading: " + this.newValue);
        String filePath = "C:\\Users\\" + System.getProperty("user.name") + "\\Downloads\\" + this.newValue;

        downloadFile(this.newValue);
        try {
            if (new File(filePath).exists()) {
                Runtime.getRuntime().exec("explorer.exe /select,\"" + filePath + "\"");
            } else {
                System.out.println("fichier n'existe pas");
            }

        } catch (IOException e) {
            e.printStackTrace();
        }
        closeStage();

    } else {
        AlertMaker.showErrorMessage("veuillez d'abord sélectionner un fichier", "Erreur est survenue", (Stage) root.getScene().getWindow());

    }

}
```

Figure 5.31: download Hadoop file

5.5 Les interfaces de la base de données

Le tableau suivant c'est le tableau des fournisseurs qui peuvent entré au système et faire tous les opérations possibles.

+ Options

	id	fullname	job	gender	birth_date	email	username	password
Modifier Copier Effacer	1	Mohammed SAOUDI	Informatique	mâle	1955-07-02	Mohammed@gmail.com	mohammed	mohammed
Modifier Copier Effacer	2	Fadwa DJOKMA	Docteur	femelle	1980-10-22	Fadwa@gmail.com	fadwa	fadwa123
Modifier Copier Effacer	3	Farid AZIEZ	Docteur	mâle	1982-02-28	Farid@yahoo.com	farid	farid1962
Modifier Copier Effacer	4	Ayaa ADAMS	Docteur	femelle	1993-05-01	Ayaa@gmail.com	ayaa	ayaa15

Figure 5.32: Interface de la table des fournisseurs

Le tableau suivant représente une partie de la base de donnés des patients.

ptions

	ID_patient	Date_test	First_name	Last_name	Age	Genre	Disease	SSN	Job	Nationality	Situation	Country
Modifier Copier Effacer	1	2017-12-01	Hanene	CHOUKRI	70	femme	Osteogenesis imperfecta	252-87-5870	Security agent	Seychellois	divorced	Venezuela
Modifier Copier Effacer	2	2017-12-02	Sohiab	MEZERDI	64	Male	Cardiomyopathy	417-22-6659	Botanist	Guyanese	divorced	Djibouti
Modifier Copier Effacer	3	2017-12-03	Ayaa	SAOUDI	60	femme	Hémophilie	792-60-1618	Builder	Qatari	married	Colombia
Modifier Copier Effacer	4	2017-12-04	Nadia	DJOKMA	61	femme	Anemia	731-22-10572	Professor in microbiology	Cameroonian	divorced	Kuwait
Modifier Copier Effacer	5	2017-12-05	Nadia	ZERNADJI	72	femme	Yellow Fever	971-44-1096	Hairdresser	Czech	married	Iraq
Modifier Copier Effacer	6	2017-12-06	Fadwa	AZIEZ	80	femme	Cholera	359-26-4353	Worker	Finnish	divorced	Finland
Modifier Copier Effacer	7	2017-12-07	Farid	ADAMS	44	Male	High blood pressure	456-98-7341	Pilot	Moroccan	widower	Jordan
Modifier Copier Effacer	8	2017-12-08	Sammy	BEN TOURKI	86	Male	Leprosy	883-27-6675	Botanist	Malaysian	single	United Kingdom
Modifier Copier Effacer	9	2017-12-09	Imade	ZERWAL	20	Male	Bronchus cancers	360-43-2117	Business director	Paraguayan	divorced	Mexico
Modifier Copier Effacer	10	2017-12-10	Adel	SAOUDI	56	Male	Influenza	241-22-2636	Dustman	Russian	divorced	North Korea
Modifier Copier Effacer	11	2017-12-11	Alyas	FIGHOULI	34	Male	Cardiomyopathy	356-27-3905	Pilot	Bruneian	divorced	United States of America
Modifier Copier Effacer	12	2017-12-12	Johiana	JACKSON	86	femme	Cholera	1087-59-9575	Dustman	Egyptian	single	Malaysia
Modifier Copier Effacer	13	2017-12-13	Jameela	GREEN	42	femme	Epilepsy	817-60-9115	CEO	Saint Lucian	widower	Netherlands
Modifier Copier Effacer	14	2017-12-14	Hanene	ABDALI	30	femme	Anemia	535-103-10095	Anesthetist	Chilean	widower	Guinea

Figure 5.33: Interface de la table des patientes

Ce tableau représente les clients inscrites qui peuvent voir les données anonymisés

+ Options

	id	fullname	username	password
Modifier Copier Effacer	1	Salah Eddine	salah	test
Modifier Copier Effacer	2	MacKensie B. Trevino	marco	test
Modifier Copier Effacer	3	Hayley A. Dyer	rizzy158	ifqt31
Modifier Copier Effacer	4	Ezra D. Carson	bupai007	yvhs98
Modifier Copier Effacer	5	Jameson P. Hale	myyua467	gwxn65
Modifier Copier Effacer	6	Camden Z. Cantrell	vuzks687	xfke46
Modifier Copier Effacer	7	Mason D. Olsen	mxzgm029	qbyf96
Modifier Copier Effacer	8	Josiah H. Ellison	sstgz432	gzlb86
Modifier Copier Effacer	9	Lacy V. Austin	odwa499	bkki59
Modifier Copier Effacer	10	Luke Z. Browning	sxvde794	ovur52

Figure 5.34: Interface de la table des clients

La dernière table pour les services cloud

	id	provider	vm_number	price	storage_space
Modifier Copier Effacer	2	kamatera	24	100	210
Modifier Copier Effacer	6	Adobe	32	234	111
Modifier Copier Effacer	8	IBM Cloud	18	325	456
Modifier Copier Effacer	10	Red Hat	22	325	452
Modifier Copier Effacer	13	SAP	12	254	458
Modifier Copier Effacer	16	Dropbox	23	223	669
Modifier Copier Effacer	18	OneDrive	247	123	110
Modifier Copier Effacer	19	kkkkkk	120	300	500
Modifier Copier Effacer	20	test	20	100	452
Modifier Copier Effacer	21	fcf	24	100	500

Figure 5.35: Interface de la table des services cloud

5.6 Conclusion

Dans ce dernier chapitre nous avons proposé un nouveau système pour résoudre les problèmes de la protection et l'assurance de la vie privée, nous avons montré l'implémentation de notre système, et décrit les outils utilisés pour cette implémentation.
Nous avons illustré les interfaces graphiques avec une description textuelle, la plateforme Hadoop et les principaux codes source et aussi présenté un exemple illustrant les différents services offerts par notre application.

Chapitre 6

Conclusion et perspectives

6.1 Conclusion

Aujourd'hui, près de la moitié de la population mondiale interagit avec les services en Ligne. Les données sont générées à une échelle sans précédent à partir d'un large éventail de Sources. Grâce à des technologies nouvelles de stockage et surtout d'analyse Big Data permet De collecter, de stocker, et d'analyser toutes ces données à des coûts raisonnables. Ces données permettent aux organisations de comprendre le fonctionnement de leurs utilisateurs et de prédire leurs besoins. Les Big Data devraient désormais ouvrir de nouvelles opportunités de revenus pour les entreprises et, en même temps, faciliter la vie de tous les jours, mais cette opportunité ne pourra pas être saisie sauf si le respect de la vie privée des clients est assuré et protégé. Cette protection se fait par l'anonymisation des données. Les Big data se heurtent au problème suivant : comment assurer la protection des vies privées des utilisateurs ?

6.2 Contribution

Dans ce travail nous avons essayé de résoudre cette problématique en :

- ✓ Etudiant quelques approches et travaux connexes concernant la vie privée pour pouvoir faire une étude comparative entre toutes ces approches.

- ✓ Déceler quelques inconvénients pour ces approches.

- ✓ Nous avons proposé un nouveau système afin de trouver une solution pour la vie privée dans les Big Data.

  Pour cela nous avons utilisé la plateforme Hadoop qui est la plateforme Big Data utilisée avec Les différents projets pour manipuler ces données, Nous avons utilisé cette plateforme aussi Pour manipuler les différents types de fichiers vers les systèmes de fichier distribué de Hadoop (HDFS).

6.3 Perspectives

Comme perspectives, nous pouvons envisager les points suivants :

- Ajouté des autres composant pour enrichir le système.
- Ajouté des autre méthodes pour le composant d'anonymisation.
- Ajouter d'autres fonctionnalités et lancer la version commerciale.

Bibliographie

[1] Evan Stibbs , "Big Data,Big Innovation", 2014.

[2] Manyika, J., Chui, M., Bughin, J., Dobbs, R., Roxburgh, C., and Byers, “Big data: The next frontier for innovation, competition, and productivity. Mckinsey Global institute”,A. H. (2011). pages 1–20.

[3] Gantz, J. and Reinsel, D. “Extracting value from chaos state of the universe: An executive summary. IDC iView”,(2011), pages 1–12.

[4] Lisbeth.R , Aaron. C , Jose. Luis, Jair. C, Jorge luis. G, Giner. A, "A general perspective of Big Data: application, tools, challenges and trends", New York 2015.

[5] Shui yu . Song Guo , "Big Data Concepts, Theories, and Applications", pages 31_50.

[6] Koichiro. Hayashi, "Social Issues of Big Data and Cloud: Privacy, Confidentiality, and Public Utility", International Conference on Availability, Reliability and Security, 2013.

[7] Saouli.H, Kazar.O, Kassimi.D, "Applications et enjeux des Big Data dans le contexte des défis mondiaux", Laboratoire LINFI.

[8] Min Chen · Shiwen Mao · Yunhao Liu, "Big Data: A Survey", 2014, pp 175_176.

[9] Olshannikova, E., Ometov, A., Koucheryavy, Y., Borko, T. O., and Flavio, V. Big Data Technologies and Applications, chapter Visualizing Big Data,(2016) pages 100–131. Springer International.

[10] Sithu . Sudarsan, Raoul . Jetley , Srini Ramaswamy, "Security and Privacy of Big Data".

[11] J.Camenisch, S.fischer, M.Hansen, "Privacy and Identity Management for the Future Internet in the Age of Globalisation", IFIP – The International Federation for Information Processing, pp 42_43.

[12] Leslie P. Francis "Introduction: Technology and New Challenges for Privacy".

[13] P.Derbeko, S.Dolev , E . Gudes , S . Sharma, "Security and privacy aspects in MapReduce on clouds: A survey", Israel, 2015.

[14] Shui Yu School of Information Technology, Deakin University, Victoria, Australia syu@deakin.edu.au , " Big Privacy: Challenges and Opportunities of Privacy Study in the Age of Big Data",(2016),page10.

[15] Sudarsan, S. D., Jetley, R. P., and Ramaswamy, S. "Security and privacy of big data. In Big Data", (2015), pages 121–136. Springer.

[16] Derbeko, P., Dolev, S., Gudes, E., and Sharma, S. "Security and privacy aspects in mapreduce on clouds : A survey. Computer science review, 20 ", (2016). Pages 1–28.

[17] Andreas Pfitzmann Marit Hansen, Anonymity, Unlinkability, Unobservability, Pseudonymity, and Identity Management – A Consolidated Proposal for Terminology, (Version v0.25 Dec. 6, 2005).

[18] Saravanan.k , Hemavathi.D," A Journey on Privacy protection strategies in big data", International Conference on Intelligent Computing and Control Systems ICICCS ,2017.

[19] Oracle,C.(2018).Java technology oracle corporation. https://go.java/ index.html?intcmp=gojava-banner-java-com. Accessed : 2018-06-01.

[20] IDE, N. (2018). NetBeans IDE netbeans community. https://netbeans. org/index_fr.html. Accessed : 2018-06-01.

[21] Hadoop (2018). Hadoop apache hadoop community. http://hadoop. apache.org. Accessed : 2018-05-01.