Réalisation d’une architecture pour la préservation de l’utilité des Big Data.

Mezghiche Dia eddine

3 juillet 2018

. *Version de 18/06/2018*

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 Pr. BOUCHANA Belkacem et Dr. BOUREKACHE Samir, 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 Amani Souad, pour toi très respectueux père Taher. Aucun hommage ne pourrait être à la hauteur de l'amour Dont ils ne cessent de me combler. Que dieu leur procure bonne santé et longue vie.

A mon frère Oussama et mes sœurs Asma et Chaima, sans oublié ma grand-mère Massouda.
A Dr. Widad, toute ma famille et mes amis.

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, mais 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 ouvre la porte à la confidentialité et de nombreuses informations sont colligées avec le consentement des intéressés sans pour autant que ceux-ci aient véritablement conscience de la « richesse » de la donnée offerte ni du spectre d'application qui en découle. Cependant protéger l'information nous fait diminuer son utilité, autrement dit les techniques d'anonymisation des données influent sur l'utilité de ces dernières et donc sur des Big Data.

Ce travail vise à maintenir l'utilité de la donnée tout en l'a protégeant. pour cela nous avons proposé une nouvelle architecture 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 l'architecture proposée, nous avons développé un prototype qui pourra résoudre les problèmes mentionnée ci-dessus..

Mots clés : La protection de la vie privée, Big Data, utilité des données, confidentialité

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 users leave on the Web or connected objects, but also the internal data to the company or general information, The purpose of Big Data is to successfully correlate these data with each other, in real time, to draw analytical conclusions and make the right decisions. On the one hand, it is an important asset for the professional organizations and the governments for the decision-making on the other hand the analysis of these data opens the door to the confidentiality and a lot of information is collected with the consent of the interested parties without these being truly aware of the "richness" of the data offered or the spectrum of application that results. However, protecting information reduces its usefulness, ie data anonymization techniques influence the usefulness of the data and therefore Big Data.

This work aims to maintain the usefulness of the data while protecting it. for this we proposed a new architecture that includes various components, taking into account the different security criteria and characteristics of Big Data. In order to show the feasibility of the proposed architecture, we have developed a prototype that can solve the problems mentioned above.

Keywords : Privacy , Big Data,Utility, Confidentiality.

Table des matières

Table des figures

Liste des tableaux

Chapitre 1

Introduction générale

‘It’s not who has the best algorithm that wins,
It’s who has the most data’.
Prof.Andrew NG

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'enjeux 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 un compromis entre cette dernière et l'utilité des Big Data est 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 une nouvelle architecture qui crée un compromis entre la protection de la vie privée et la préservation d'utilité dans 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 autre chapitres organisés comme suit :

Le deuxième chapitre « Sécurité des *Big Data* » : 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 sécurité et 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 de la vie privée et la préservation d'utilité dans 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 tout en préservant l'utilité des 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.

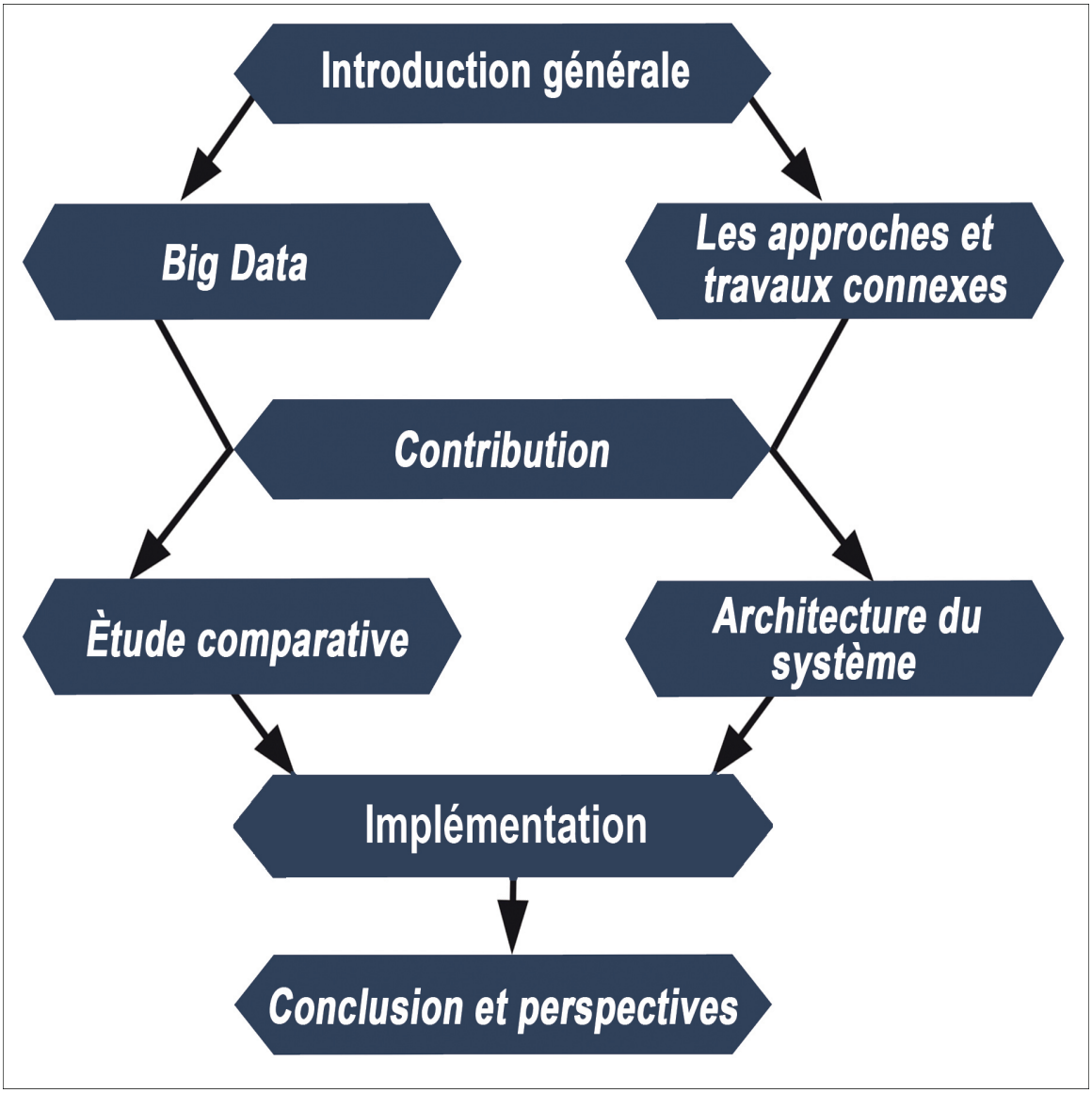


FIGURE 1.1 – Structure du mémoire.

Chapitre 2

Sécurité des *Big Data*

‘You can’t talk about Big data without talking
about things like privacy and ownership’.
Mr.Rick Smolan

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 sécurité et la vie privée.

2.2 Big Data

2.2.1 Définition

Nombreuses définitions ont été proposées pour décrire le Big Data mais toutes ont pratiquement le même sens. En mai 2011 MGI [1] 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»*[Manyika et al., 2011]*. Dans la même année un rapport a été annoncé par IDC [2] 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» *[Gantz and Reinsel, 2011]*. Une autre définition de Big Data a été annoncé par *TechAmerica Foundation* «Big Data est un terme qui décrit l'immense volume des données à haute vitesse, complexes et variables qui nécessitent des techniques et des technologies avancées pour permettre la capture, le stockage, la distribution, la gestion et l'analyse de l'information»*[Commission, 2012]*.
Donc, le Big Data est un volume massif de données doté de caractéristiques spécifiques lesquels pourraient être décrit par le modèle 5V, que nous allons présenter dans la section suivante.

2.2.2 Modèle 5V

On peut décrire le Big Data à l'aide du modèle 5V, ce dernier est composé de 5 éléments clés (volume, variété, vélocité, véracité, valeur). Illustrés dans la figure [2.1]. Le modèle 5V a marqué un progrès de la règle de 3V qui fait référence à (volume, variété, vélocité) ; puis avec la règle 4V on rajoute la véracité[Rodríguez-Mazahua et al., 2016].

1. The McKinsey Global Institute
2. International Data Corporation

FIGURE 2.1 – 5 éléments clés du modèle 5V.

2.2.3 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 *[Olshannikova et al., 2016]*.

2.2.4 Qualité des Big Data

Lorsqu'on évoque un projet Big Data, volume et variétés de données sont rapidement mentionnés, mais la qualité de ces données vient moins spontanément dans la conversation. L'un des plus grands challenges du Big Data ne tient plus dans la récolte des données mais bien plus dans l'assurance de leurs qualités.

"Potholes" pour la qualité de l'information

La qualité de l'information elle-même tombe sur des *"potholes"* qui affectent ses diverses dimensions par exemple :

- Une panoplie de sources de la même information produisant différentes valeurs touchent directement la cohérence et la crédibilité de l'information.
- Le manque de ressources informatiques suffisantes atteints l'accessibilité et la valeur de l'information.
- Les systèmes hétérogènes distribués liés à des définitions, des formats et des valeurs incohérentes affecte la ponctualité et la représentation concise de l'information.
- Un accès facile à l'information provoque des conflits avec les exigences en matière de sécurité et de confidentialité ce qui affecte directement la sécurité, l'accessibilité et la valeur de l'information.
- Les erreurs systémiques dans la production de l'information entraînent la perte d'informations qui affecte son intégralité et sa correction.*[Sivarajah et al., 2017]*.

La qualité de Big Data avec le modèle 3C

On considère que les principales caractéristiques de la qualité des données pour évaluer le niveau de qualité dans l'utilisation d'ensembles de données hétérogènes est "cohérence", il faut mentionner qu'il y a différents types de cohérences : (la cohérence contextuelle, la cohérence temporelle et la cohérence opérationnelle). Le niveau de qualité de l'utilisation des ensembles de données globaux doivent être mesuré en fonction des 3C en combinant les caractéristiques de qualité des données *ISO 25012*. parmi ses caractéristiques :

- La précision, l'intégralité, sont obtenu grâce à la cohérence opérationnelle et contextuelle.
- La crédibilité est mesurée grâce à la cohérence contextuelle.
- La cohérence temporelle nous évalue la disponibilité, la Compréhensibilité et la conformité.
- La cohérence opérationnelle estime les caractéristiques probabilité, traçabilité, accessibilité.*[Caballero et al., 2014]*

Impact de 3V sur 3C

Les 3V (volume, variété, vélocité), affectent directement la conception et la mise en œuvre des méthodes de mesure. Chaque V affecte de façon spécifique les modèles de mesures 3C comme suit :

- **Volume**
  - Affecte la cohérence contextuelle en l'intégralité et la crédibilité des données.
  - Touche la cohérence temporelle en sa disponibilité
  - Atteints la cohérence opérationnelle en la recouvrabilité, l'efficacité et la disponibilité.
- **Variété**
  - Affecte la cohérence contextuelle en la compréhensibilité et la précision.
  - Touche la cohérence temporelle en l'actualité et la conformité des données.
  - Atteints la cohérence opérationnelle en disponibilité, traçabilité et l'efficacité des données.
- **Vélocité**
  - Affecte la cohérence contextuelle en la crédibilité et la confidentialité.
  - Touche la cohérence temporelle en actualité et la disponibilité des données.
  - Atteints la cohérence opérationnelle en accessibilité, intégralité, efficacité et la traçabilité. *[Caballero et al., 2014]*

2.2.5 Framework d'implémentation.

Avec l'avènement de la technologie Big Data qui est un secteur en pleine effervescence, les entreprises doivent créer les conditions nécessaires pour traiter ce volume de données important. Les données liées à l'entreprise passent par quatre phases de traitement qui sont : la capture, l'organisation, l'analyse et la valorisation qui sont discuter ci-dessous ainsi que les différents outils disponibles. Il est nécessaire de signaler que ces quatre phases ne sont pas obligatoires pour une organisation afin de s'adapter aux solutions des Big Data, tout dépend de son activité, ses besoins et de son budget.

- **La capture de données :** Cette phase conçue pour collecter les données variées provenant à partir de différentes sources et à toute vitesse. Selon la variété des données, plusieurs outils tels que RDBMS [3], HDFS [4] et non seulement SQL, mais aussi NoSQL, peuvent être utilisés pour stocker des données.

- **L'organisation des données :** Une fois les données sont capturées et stockées, il est important d'affiner ces dernières afin d'éliminer tous les redondances. Elles peuvent être organisées dans des bases de données relationnelles ou des entrepôts de données. Nombreux outils peuvent être utilisés comme Hive, Cloudera qui sont intégrés sur Hadoop.

- **L'analyse de données :** Le but de cette phase est d'utiliser des outils d'analyse conçus pour extraire la valeur des données organisées et qui sont intégrées dans une base de données relationnelle ou des entrepôts de données. Certains SGBD relationnels bien connus tel que DB2 d'IBM, SQL Server de Microsoft et Oracle peuvent être utilisés.

- **La valorisation de données :** Cette phase est basée sur l'analyse de données qui a été faite dans la phase précédente afin de les valorisées et de prendre des décisions. On peut prévoir des résultats qui comprennent des rapports d'entreprise, des graphiques multidimensionnels, des tableaux de bord,...Etc.*[Sanyal et al., 2016]*

2.2.6 Domaines d'application

Dans cette section nous présentons les principaux domaines d'application du Big Data.

- **Agriculture :** 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. 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

3. Relational Database Management System
4. Hadoop Distributed File System

étant une ressource de plus en plus rare), où nous avons besoin de gérer de gigantesques masses de données qui concernent les prédictions météorologiques et la sécheresse du sol.

- **Assurance :** L'assurance représente l'un des domaines directs d'application des Big Data, vu 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 d'importante 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.

- **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 les visiter. Mais ce qui révolutionne vraiment le marketing de nous 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'ils achètent, à quoi ils s'intéressent, et les produits qu'ils ne se trouvent pas aux marchés, ce qui permet d'analyser et d'étudier leurs besoins en temps réel afin de produire des solutions et méthodes de marketing plus efficaces. L'utilisation des capteurs permet de capter des données de diverses formes : image de visage pour analyse émotionnelle, vidéos pour description comportementale, données textuelles pour décrire la nature des produits achetés, 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 des Big Data.

- **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 taches 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 grandes 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 considérable et une alternative prometteuse pour la gestion des plate-formes d'achat/vente intermédiaire.

- **Compétitivité et innovation de produit :** La possibilité de traiter d'importantes masses d'informations en temps réel permet aux entreprises d'analyser les besoins de ses clients afin de pouvoir optimiser et améliorer ses propres produits et augmenter sa compétitivité sur le marché. C'est ainsi que les services qu'offrent les fournisseurs de téléphonie mobile permettent aux agences 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 en billet d'avion par exemple. Les techniques d'analyse en temps réel de grandes masses d'informations, issues des 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.

- **Gestion de catastrophes naturelles :** L'une des applications les plus intéressantes des 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. Les organisations internationales d'assistance humanitaire peuvent préparer les ressources nécessaires (couverture, alimentations, médicaments) ainsi que les moyens de transport et d'interventions rapides pour aider la population en détresse.

- **Contrôle d'épidémies :** Les Big Data peuvent 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. Les Big Data sont également utilisés pour traquer la population des rats sur les grandes villes telle 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.

- **Prévention d'attaques cybernétiques :** De nos jours, les techniques d'analyses de données qu'offrent les Big Data sont devenues incontournables pour pouvoir détecter les intrusions, les failles sécuritaires ainsi que les attaques cybernétiques, vue que le volume de données transporté sur le Net est devenu gigantesque, diversifier, et nécessitant un traitement en temps réel. Avec les techniques de traitement de données Bigdata 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. La figure[2.2] résume les principaux domaines d'application des Big Data*[Saouli et al., 2016]*.

FIGURE 2.2 – Domaines d'application des Big Data.

2.2.7 Défis et enjeux

2,5 quintillions octets, tel est la taille de données produite chaque jour dans le monde. Le Big Data affronte plusieurs lourds défis. Dans cette partie nous allons citer les plus importants :

- **Représentation des données :** Vu l'immense niveaux d'hétérogénéité de nombreuses données une bonne représentation des données est nécessaire ; structure, classe et type de données. Ce défi est très important il permet de rendre la donnée plus significatif et facile à l'analyse informatique et l'interprétation des utilisateurs.
- **La réduction de la redondance et la compression des données :** La réduction de la redondance et la compression des données sont indispensables pour réduire le coût de l'ensemble du système ; sans nuire à la valeur potentielle des données.
- **La gestion du cycle de vie des données :** Les systèmes de stockages actuels et leurs progrès lents ne sont plus aptes à supporter les données massives générées par l'informatique à des vitesses et des échelles sans précédent ; ce qui nous pousse à réaliser une analyse précise de ces données pour décider quelle donnée doit être stockée et laquelle doit être rejetée.
- **Mécanisme analytique :** Les SGBDR traditionnels sont conçus avec un manque d'évoluti-

vité et d'extensibilité, qui ne répond pas aux exigences de performance. Quant aux bases de données non relationnelle, ces dernières ont prouvé leurs avantages dans le traitement des données non structurées malgré cela, il existe encore des problèmes dans leur performance et leurs applications particulières pour remédier à cet obstacle il faudrait miser pour le duo SGBDR/ bases de données non relationnelles.

- **Confidentialité des données :** La plupart des grands fournisseurs de services de données ou les propriétaires présents pourraient conserver et analyser de manière rigoureuse ces énormes ensembles de données en raison de leur capacité limitée. Ils doivent s'appuyer sur des professionnels ou des outils pour analyser ces données, ce qui augmente les risques potentiels pour la sécurité.
- **Gestion de l'énergie :** Les différentes opérations appliquées aux données (traitement, entreposage, transmission) consomment de plus en plus d'énergie électrique ; ce qui nécessite l'établissement d'un mécanisme de contrôle et de gestion de la consommation d'énergie au niveau des systèmes.
- **Évolutivité :** Dans ce défi le rôle des algorithmes analytique est très important car il doit pouvoir traiter des jeux de données de plus en plus complexes.
- **Coopération :** L'analyse des grandes données est une recherche interdisciplinaire ; qui permet aux ingénieurs de différents domaines d'utiliser pleinement leur expertise, afin de coopérer. Une architecture globale complète du réseau de données est nécessaire*[Chen et al., 2014]*.

2.3 Sécurité et vie privée

2.3.1 Sécurité : survole général

Une solution complète de la sécurité des données doit répondre à trois exigences : confidentialité, l'intégrité et la disponibilité ; ces derniers font partie de tous les environnements d'application. La protection des données est assurée par différents composants des systèmes de gestion des données, tels que les SGBD [5] qui fournissent des techniques de protection comme, les mécanismes de contrôle d'accès et le contrôle de la reprise et de la simultanéité. La mise en œuvre de ces techniques devient un grand défi en ce qui concerne les Big Data*[Bertino, 2015]*. À l'heure où nombreuses plates-formes et outils des Big Data sont émergés pour gérer ces données massives, les mesures de sécurité imposent des problèmes non négligeables car la plupart des grandes plates-formes s'appuient sur des pares-feux traditionnels ou des implémentations sur la couche d'application afin de limiter l'accès aux données, mais ces derniers n'assurent pas une protection adéquate *[Li et al., 2016]*.

5. systèmes de gestion de bases de données

Actuellement, la plateforme *Hadoop* est largement répandue dans les secteurs industriels et les recherches scientifiques, elle prend en charge certaines fonctionnalités de sécurité grâce à la mise en œuvre actuelle de *Kerberos*, tel que les autorisations *HDFS*[6] de base. Cependant il est nécessaire d'être attentif que malgré les efforts qui ont été fait pour améliorer la sécurité de la plateforme Hadoop, la situation reste compliquée et elle ne répond pas adéquatement aux exigences des entreprises pour une protection robuste, car *Hadoop* n'est pas une technologie unique, mais tout un écosystème d'applications, y compris *Hive*, *HBase*, *Zookeeper* et Oozie*[Bertino, 2015]*.

2.3.2 Les six éléments de sécurité

La sécurité de l'information inclut à la fois six éléments de sécurité essentiels qui garantissent la protection de tous les propriétaires de l'informations, l'absence de l'un de ces éléments a occasionné une sécurité insuffisante et certains problèmes. Dans cette section nous allons les évoquer et déterminer l'impact de chaqu'un.

1. La disponibilité : La disponibilité de l'information assure que les données sont accessibles et disponibles en tout temps sans interruption ni dégradation. Une perte de cette dernière due à la suppression, la destruction ou l'altération, par des actions volontaire ou involontaire sur les données ou par des phénomènes techniques rendent l'information ou une partie d'elle inaccessible ou indisponible. L'opération de récupération consiste à recouvrer les données perdues, mais elle n'est jamais parfaite. C'est pour cette raison qu'il est nécessaire de prendre tous les mesures préventives et de considérer ce défi en tant qu'un but crucial de la sécurité de l'information

2. L'utilité : Nombreuses organisations considèrent le cryptage et l'anonymisation comme un remède absolu afin de protéger la confidentialité des données sensibles et personnelles, cependant une gestion précise et des mécanismes de protection puissants doivent être adopté afin de les protégée. L'utilité des données est très importante car la perte de la clé de chiffrement rend le décryptage de l'information très difficile et parfois impossible, dans cette situation l'information est disponible, mais avec une perte d'utilité, ou une grossière anonymisation qui rend l'information sous une forme inutile.

3. L'intégrité : Comprend des mesures qui protègent les données contre les modifications illégales, y compris la création, la suppression et la modification non autorisées de données, l'intégrité garantit la cohérence, l'exactitude et la fiabilité des données pendant toute leur durée de vie. Plusieurs contrôles peuvent fournir une protection de l'intégrité des données, tels

6. Hadoop Distrbuted File System

que l'utilisation et la vérification des numéros de séquence, l'authentification et le contrôle d'accès.

4. L'authenticité : Les informations authentiques garantissent qu'elles sont créées à partir d'une source spécifiée, cette authenticité assure que l'origine de données peuvent être prouvée sans aucun doute. Vu que l'authentification assure l'intégrité et la confidentialité de l'information il est nécessaire de prendre toutes les mesures pour éviter de douter de l'origine correcte de ces données et permettre aux utilisateurs d'être correctement authentifiés.

5. La confidentialité : La confidentialité signifie que les données ne doivent être mises à la disposition qu'a des personnes autorisées, non seulement les données elles-mêmes mais aussi les systèmes. Une attaque contre la confidentialité est la collecte non autorisée d'informations (par exemple en espionnant les données de connexion par une personne non autorisée). La protection de la confidentialité exige que des mesures de sécurité soient prises pour empêcher l'accès non autorisé aux données stockées et transmises.

6. La possession : Il est nécessaire de garantir la protection de la possession d'informations afin d'éviter tout vol et duplication des données. La gravité de la perte de possession varie selon la nature de l'infraction. Une perte de confidentialité peut accompagner la perte de possession, mais nous devons traiter la confidentialité et la possession séparément pour déterminer les actions illégales et quels contrôles nous devons appliquer pour les prévenir *[Bosworth et al., 2016]*.

2.3.3 Sécurisation : SI via Big Data

Changement de paradigme à la sécurité centrale des données

Lors du déplacement de données d'un système à l'autre ou entre des domaines, nombreux problèmes sont engendrés autour de la sécurité et la confidentialité, un changement de paradigme de sécurité devient plus nécessaire lorsque les données passent par un certain nombre de transformation et deviennes distribuées davantage entre les domaines traditionnels de la sécurité. Pour favoriser ce changement on trouve plusieurs facteurs supplémentaires qui fondent de nouveaux défis :

- La virtualisation offre un niveau d'efficacité supplémentaire pour la sécurité de l'environnement de traitement des données, mais les problèmes de sécurité de données au repos sont encore irrésolubles.
- La mobilité des différentes composantes de l'infrastructure de données.
- La structure complexe de Big Data nécessite des politiques différentes pour contrôler et gérer l'accès.

La confiance aux solutions des virtualisation

Ces années, la virtualisation est devenue de plus en plus importante pour les entreprises et les utilisateurs qui doivent stocker et traiter leurs données. Ces dernières sont associées avec un environnement de stockage sous le contrôle des fournisseurs. La relation client-fournisseur et sa pérennité doit se baser sur la protection de la vie privée afin d'augmenter la confiance.

La propriété des données

La propriété des données est une notion importante et largement discutée dans la protection et la gouvernance des données, elle a une relation et un effet sur l'identité de contrôle d'accès centrée et la délégation, comme elle est liée à la propriété individuelle ou organisationnelle. Les modèles de sécurité doivent adopter le concept de propriété des données, car elle constitue un véritable droit de l'Homme numérique.

Renseignements personnels, vie privée et opacité

L'explosion du Big Dataet son impact majeur sur le respect et la protection de la vie privée des personnes, surtout dans le domaine des services modernes et les réseaux sociaux a motivé l'évolution des nouveaux modèles de protection afin de fournir une sécurité fiable de toutes ces données et faire face a l'opacité actuelle des droits de propriété et des possibilités de partage des données*[Demchenko et al., 2014]*.

2.3.4 Types de sécurité

Plusieurs définitions ont été discutées autour du concept " sécurité " selon des critère linguistiques, culturelles ou historiques ; c'est pour cette raison qu'il est très difficile de la définir. Dans cette partie nous allons identifier sept types généraux de contextes de sécurité et les mesures d'accompagnement pour sauvegarder et protéger ces contextes

1. **La sécurité physique :** S'appuie sur l'ensemble des mesures physiques visant à garantir les propriétés des systèmes, des espaces, des objets, des caractéristiques physiques et même des êtres humains.
2. **La sécurité politique :** Le but de celle-ci est la protection des institutions et les structures établies, des choix politiques reconnus et aussi des droits acquis.
3. **La sécurité socio-économique :** Ce sont des mesures destinées à sauvegarder le système économique, son développement et son impact sur les individus.
4. **La sécurité culturelle :** Repose sur la protection de la permanence des schémas traditionnels, de la langue, de la culture, des associations, de l'identité et des pratiques religieuses.
5. **La sécurité environnementale :** Des mesures qui ont pour fonction d'assurer la sécurité contre les dangers environnementaux liés à des processus naturels ou de l'accident mais

aussi d'autres facteurs rentrent également en jeux, tels que l'humain en raison de l'ignorance, de la mauvaise gestion ou de la conception intentionnelle.

6. **La sécurité d'incertitude radicale :** Inclut des mesures qui ont pour but d'assurer la sécurité contre la violence et des menaces exceptionnelles et rares, qui ne sont pas délibérément infligées par un agent externe ou interne, mais peuvent encore menacer de façon considérable et de dégrader la qualité de vie.
7. **Sécurité de l'information :** Des mesures qui permettent d'assurer la sécurité des systèmes d'information et des données contre l'accès, l'utilisation et l'enregistrement non autorisés, la divulgation, la modification et l'interruption*[Friedewald et al., 2014]*.

2.3.5 Défis de protection et sécurisation

L'augmentation des données importantes représente un défi fondamental pour les principes établis de protection des données, le modèle utilisé actuellement basé sur les objectifs, le consentement et la transparence ne fonctionnent plus en raison de la complexité de l'analyse, cependant la limitation dans un cadre de protection des données nous fait risquer de perdre leurs avantages. Par ailleurs il est préférable de tenir compte de la manière dont les données sont exploitées et non sur le contrôle, la collecte et la conservation des données personnelles pour protéger la vie privée. Il est difficile pour les organisations d'être utiles et novateurs pour dire aux gens ce qu'ils font avec leur données personnelles, et doivent continuer à examiner comment leur traitement des données personnelles influe sur la vie privée des personnes*[Derbeko et al., 2016, Sudarsan et al., 2015, Information Commissioner's Office, 2014]*.

L'anonymisation des ensembles de données individuelles s'avère insuffisante, et l'accès global à de grandes donnés n'est pas réellement assimilé. Les principaux défis de recherche sont :

- La compréhension des caractéristiques desBig Data et leur étendu de variété.
- Le contrôle d'accès lorsque la disponibilité est difficile à contrôler.
- L'anonymat des données y compris celles contenants les PII [7].
- Conception d'algorithmes cryptos et surveillance en temps réel avec des données de diffusion cryptées.
- Audit et vérification de la conformité des systèmes d'information*[Derbeko et al., 2016]*.

2.3.6 Vie privée : survole général

Dans le monde des donnée la protection de la vie privée reste un défi important a relever ; elle fait référence généralement a la PII. La protection de la vie privée vise à protéger l'infor-

7. Personally Identifiable Information

mation qui est considérée comme personnelle et non divulguée. Cependant l'accès aux données devient plus facile dans notre ère et l'anonymat n'est plus maintenu comme au passé.

- **Vie privée en ligne :** Toute donnée avec PII en ligne est à la fois un problème de sécurité et de confidentialité. Le type d'informations, (nom / prénom, le numéro de permis de conduire, la date de naissance) qui constitue une PII permet d'utiliser d'autres informations qui permettent suffisamment d'identifier un public cible spécifique.
- **Vie privée hors ligne :** Certes une donnée hors ligne ne pose aucun problème de confidentialité, mais à mesure que nous avançons dans un environnement intelligent (surveillance, maison intelligente.) Tôt ou tard, la confidentialité hors ligne deviendra une chose du passé.
- **Post archives :** Le moment où les données sont archivées elles sont accessibles et deviennent "actionables". Dans ce cas la question de la vie privée se développe. En analysant les données archivées, nous allons simplement prédire les comportements futurs*[Li et al., 2016, Sudarsan et al., 2015, Bertino, 2015]*.

2.3.7 Types de vie privée

La vie privée est un concept qui n'est pas facile à mesurer ni à définir. D'après le professeur Bernard Eignier « La vie privée n'est-elle que l'opposé de la vie publique ? Mais alors il faudrait savoir si la vie privée doit se définir par rapport à celle-ci ou bien s'il faut raisonner inversement »*[Eignier, 1997]*. Ou comme l'a déploré Daniel Solove « La vie privée est un concept en plein désarroi. Personne ne peut exprimer ce que cela signifie »*[Solove, 2008]*. Les sept types de vies privées que nous allons énumérer par la suite est une taxonomie développée par *[Finn et al., 2013]*.

1. **La vie privée de la personne :** S'appuie sur le droit de la protection des fonctions et des caractéristiques du corps tels que les codes génétiques et la biométrie, et encore plus les intrusions non physiques par exemple celles qui se produisent avec les scanners corporels de l'aéroport. Tous ces critères donnent à la personne humaine sa valeur infinie.
2. **La vie privée du comportement et de l'action :** Englobe toutes les questions cruciales comme les habitudes, les activités politiques et les pratiques religieuses. Il faut mentionner aussi que la vie privée du comportement personnel s'intéresse aux activités qui se déroulent dans l'espace public et l'espace privé.
3. **La vie privée de la communication :** Sert à éviter l'interception de la communication par téléphone sans fil ou l'enregistrement ainsi que l'utilisation de bogues, les microphones directionnels ou l'accès non autorisé aux messages électroniques.

4. **La vie privée des données et de l'image :** Comprend la protection des données de l'individu pour qu'elles soient inaccessible à d'autres personnes ou organisations.
5. **La vie privée des pensées et des sentiments :** Peut être distinguée de la vie privée de la personne, ce n'est que le droit que possède tout individu de ne pas partager ou révéler ses pensées ou ses sentiments.
6. **La vie privée de l'emplacement et de l'espace :** Fait référence a le droit de l'individu de se déplacer dans un espace public ou semi-public sans être suivi ou surveillé. Il nous faut ajouter aussi que cette dernière inclut le droit à la solitude et à la vie privée dans des espaces plus particuliers tels que la maison, la voiture ou le bureau.
7. **La vie privée de l'association :** Représente le droit des personnes à s'associer à ceux qu'elles souhaitent et sans être surveillées.

2.3.8 Défis et enjeux pour la vie privée

Les problèmes de la sécurité et la vie privée suscite énormément de débats et sont amplifiés par la vitesse, le volume et la variété de Big Data, tels que les infrastructures cloud à grande échelle, la diversité des sources de données, l'acquisition de données et la migration des grandes volume inter-cloud. Il est très important de garantir la transparence vis-à-vis de ces traitements d'énorme volume de données*[Francis, 2014]*. Dans cette partie nous allons mettre en évidence certains défis en matière de sécurité et vie privée pour les calculs MapReduce.

- **La taille des données d'entrée et son stockage :** La sécurisation des calculs de MapReduce et la protection de la vie privée envers ces énormes quantités de données impliquent un défi majeur. Dans les calculs MapReduce les données sont divisées en petites tailles et distribuées à plusieurs nœuds, Chaque division doit être transférée de manière sécurisée.
- **Nature hautement distribuée des calculs MapReduce :** Les capacités de traitement en parallèle et la répartition entre plusieurs nœuds de clusters participent à la rapidité de l'analyse en MapReduce, mais il nécessite à des mécanismes pour protéger ces nœuds et ces données car le traitement distribué sur les données répliquées à une probabilité plus élevée d'attaques par rapport à un système centralisé.
- **Flux de donnée / cloud hybride :** Les calculs MapReduce exigent un flux de données complexe entre différents nœuds de calculs et de stockage. A titre d'exemple EMR [8] utilise deux cloud différents : l'un consiste à exécuter les calculs MapReduce et l'autre pour le stockage des données. Cette structure à dual cloud nécessite un flux de données complexe entre eux. Ce dernier devient plus complexe lorsqu'on effectue des calculs MapReduce dans des cloud

8. Amazon Elastic MapReduce

hybrides ou publics. En outre, MapReduce a la capacité de fonctionner dans un seul cloud et cette caractéristique pose des défis supplémentaires pour soutenir un déploiement de cloud hybride qui fournit un traitement performant des données sensibles et non sensibles et une gestion efficace et économique des ressources.

- **Évolutivité, tolérance de panne et transparence :** L'intégration des mécanismes et l'implication des protocoles de sécurité et de confidentialité ne devrait pas réduire l'efficacité, l'évolutivité et la tolérance aux pannes des algorithmes MapReduce et doit être transparente pour les utilisateurs, sans réduction des fonctions.
- **Accès aux données non fiables :** MapReduce permet une grande souplesse dans la validation des calculs définis par l'utilisateur, ce qui implique un grand problème car les utilisateurs peuvent fournir des codes qui perturbent le bon fonctionnement des cluster MapReduce, cependant il est important de développer des algorithmes de sécurité pour faire face à ces codes corrompus.
- **Protection des données contre les fournisseurs de cloud :** Les utilisateurs peuvent stocker leurs données privées dans un cloud public, les fournisseurs de ces services ont la capacité d'espionner et de contrôler les données et les codes MapReduce et même de modifier ou supprimer ces derniers. Assurer la vie privée d'utilisateurs en présence d'un fournisseur cloud est un défi difficile à accomplir.
- **Des multi-utilisateurs sur un seul cloud public :** Il est nécessaire que les fournisseurs permettent à plusieurs utilisateurs d'accéder à leurs données ou à une partie spécifique. Des mécanismes d'authentification et d'autorisation précieuses suffisent pour résoudre les problèmes de confidentialité, mais la situation devient plus complexe quand les données sont fournies par un certain nombre de fournisseurs, chacune ayant des exigences de confidentialité différentes*[Derbeko et al., 2016]*.

2.3.9 Sécurité via vie privée

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*[Sudarsan et al., 2015]*.

2.3.10 Cryptage de données

Le cryptage est une technique cruciale et largement acceptée qui permet de protéger la confidentialité des données sensibles, cette technique s'effectue à l'aide des fonctions mathématiques qui génère une clef de chiffrement, cette dernière n'est autre qu'une suite de caractères. Il s'avère très difficile d'appliquer cette technique au sein des Big Data en raison de son volume important et sa grande diversité. Cela ne l'empêche pas de développer des nouvelles approches et des algorithmes efficaces afin de les sécurisées. Dans ce qui suit nous allons discuter les recherches menées dans la méthodologie de cryptage.

Le cryptage des recherches « SE » : Dernièrement le cryptage des recherches est apparu comme un problème important à l'intersection de la cryptographie de cloud et Big Data, cette technique cryptographique permet de rechercher des informations spécifiques dans un contenu crypté, elle est construite sur le modèle client / serveur où le serveur stocke des données cryptées et indexées par un ou plusieurs clients. À la demande des données un ou plusieurs clients sont générer des « drapdoors » pour le serveur qui effectue la recherche. Le cryptage des recherches est basé sur l'index "SE" qui est un tuple de six algorithmes : *"KeyGen ;BuildIndex ;Encryption ;Query ;Search ;Decryption"*, le tableau[2.1] résume les entrées et sorties de chaqu'un *[Dong et al., 2011]*.

Algorithme	**Prend à son entrée**	**Génère à sa sortie**
'KeyGen'	• Un paramètre λ	• Une clé secrète K
'BuildIndex'	• Ensemble des documents $D = \{D1, ..., Dn\}$	• Un Index ρ
'Encryption '	• Ensemble des documents $D = \{D1, ..., Dn\}$ • Un Index ρ • Une clé secrète K	• Ensemble des documents $C = \{C1, ..., Cn\}$ • Un Index ρs sécurisé
'Query'	• Un mot-clé ω • Une clé secrète K	• Une requête crypté $q\omega$
'Search'	• Un Index ρs sécurisé • Un $q\omega$ query	• Une collection de documents identificateurs dont le fichier de données contient le mot-clé ω
'Decryption'	• Un fichier crypté *Ci* • Une clé secrète K	• Un documents décrypté *Di*

TABLE 2.1 – Entrées et sorties de l'index « SE ».

Le cryptage des recherches « SE » : Deux types de cryptage des recherches ont été développés qui sont symétrique et asymétrique. La grande différence entre les deux, repose sur la clé du chiffrement, ceci est montré clairement dans la figure [2.3]

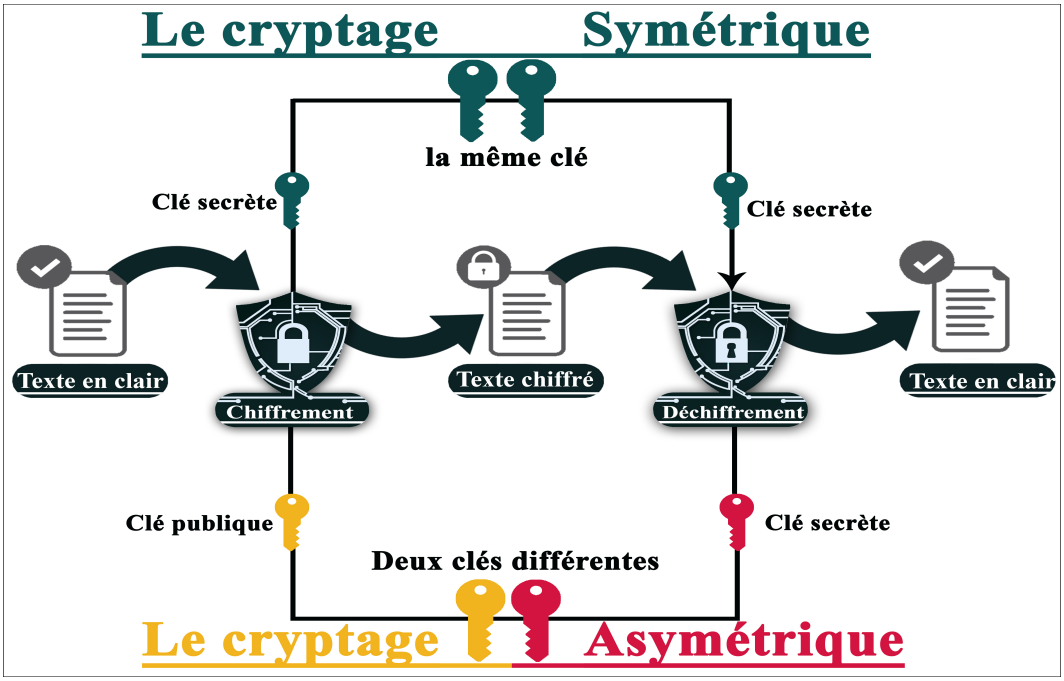


FIGURE 2.3 – Cryptage symétrique vs cryptage asymétrique.

Le cryptage de préservation de l'ordre « OPE » : 'OPE' c'est la méthode de cryptage où les chiffrages conservent l'ordre des fichiers en claires, elle était à l'origine étudiée de manière heuristique dans la communauté de base de données par *[Agrawal et al]*. Un grand nombre d'améliorations fonctionne sur les schémas OPE afin de minimiser les fuites d'informations et augmenter la sécurité comme : Modulaire OPE, Mutable OPE..., Etc*[Derbeko et al., 2016]*.

Le cryptage structuré : Les données structurées habituellement portent une information très concentrée et précieuse et par conséquent sa protection est indispensable. Le cryptage structuré est très nécessaire pour améliorer la sécurité et fournir une protection plus fiable pour ce type de données.*[Derbeko et al., 2016]*.

Cryptage homomorphique (HE) : Le problème du cryptage homomorphique était annoncer par [Ron Rivest, Michael Dertouzos et al.,1978], mais le premier algorithme d'homomorphisme était proposé en 2009 par Craig Gentry, cette forme de cryptage permet d'effectuer des opérations algébriques spécifiques (l'addition et la multiplication) sur les données chiffrées pour produire des résultats correctes sans devoir décrypter les données tout au long du processus*[Derbeko et al., 2016]*.

2.3.11 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 [9] 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 *[Derbeko et al., 2016]*.

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.

2.3.12 Définition et types de vulnérabilité

Cette section s'articule autour de deux éléments importants du concept vulnérabilité : sa définition et ses types.

Définition de vulnérabilité :

Ce terme touche plusieurs domaines, une vulnérabilité dans l'informatique définit toutes les failles ou les faiblesses d'un système ou un mécanisme de protection ce qui permet aux attaquants d'utiliser des cyber-attaques pour les exploiter dans le but de voler, endommager des informations confidentielles ou prendre des sites web ce qui rend l'information indisponible.

Type de vulnérabilité

- **La vulnérabilité des logiciels :** Est les failles dans un programme informatique lors de sa conception, c'est ce qui permet aux pirates de les exploiter afin d'obtenir un accès non autorisé au système, une fois l'accès obtenu, le système est à la disposition de l'attaquant et des cyber-attaques peuvent être lancées.Une faille dans les logiciels de base de données peut causer des dommages énormes aux grandes quantités d'informations stockées par les orga-

9. Trusted Platform Module

nisations et les rendant accessibles aux cyber-attaques. Donc une mise à jour est primordiale pour tout les systèmes et les logiciels afin de rectifier toutes les vulnérabilités.

- **La vulnérabilité personnelle :** Comprend des employés qui peuvent endommager l'information de leur organisation. Les employés mécontents ont plus d'avantage que les attaquants externes, puisqu'ils ont accès à des informations confidentielles et savent où les systèmes critiques sont situés, par contre des attaquants externes ont besoin de chercher des vulnérabilités afin de les exploiter. Les attaquants externes peuvent profiter de la naïveté des employés pour se leurs procurer ses informations confidentielles, ces méthodes sont utilisées dans l'ingénierie sociale.

- **Les vulnérabilités des protocoles réseau :** Plusieurs vulnérabilités ont été découvertes dans les protocoles de réseau leurs exploitation permet de perturber des site web ou l'inondation d'un réseau afin d'empêcher son fonctionnement, l'un de ses protocoles c'est HTTP [10] qui était exploité par DDoS attaques. DNS [11] est un autre protocole qui est vulnérable, l'exploitation de ce dernier permet de créer des sites web malveillants et les utilisés pour voler des informations confidentielles auprès des victimes. Il est nécessaire que les organisations prennent des mesures pour éviter l'exploitation de ces protocoles *[Tsegaye and Flowerday, 2014]*.

2.3.13 Infrastructure critique et Big Data

L'infrastructure d'information critique constitue une partie non négligeable de l'infrastructure critique. L'identification et la définition de cette dernière est variée selon les pays observés, elle est présentée dans plusieurs secteurs tels que les télécommunications, les services d'urgence, les banques et la finance, les services de santé, les systèmes de stockage et de transport du gaz et du pétrole et la distribution de l'eau…etc. Elle fait partie de tous les pays et sa protection assurent la continuité du gouvernement.

Durant les années qui se sont écoulées la protection de IIC a été plus importante car elle n'était pas connectée à Internet mais lorsqu'une telle connexion est devenue essentielle, plusieurs mécanismes ont été mis en place pour répondre aux exigences de sécurité. Ces dernières années, nous avons vu plusieurs secteurs principaux de la société se développer avec l'arrivée du Big Data pour couvrir ses besoins. la protection de l'infrastructure critique du Big Data font l'objet d'une attention particulière pour le bon fonctionnement de la société et de l'économie.*[Sudarsan et al., 2015]*

10. HyperText Transfer Protocol
11. Domain Name System

2.3.14 Contrôle de sécurité et protection d'infrastructure

Aujourd'hui, l'infrastructure est en risque de menace par groupes de criminels, des hackers ou des gouvernements étrangers hostiles , une stratégie de protection précieuse doit faire face à ces menaces qui mènent à des pannes atroces. Dans cette section, nous allons discuter les différentes principales catégories de contrôles de sécurité, en commençant par les contrôles préventifs se sont des catégories préliminaires qui empêchent les incidents de sécurité et qui sont meilleures par rapport aux deux contrôles qui suivent, les contrôles détectives qui détectent les incidents de sécurité qui ont échappés aux contrôles préventifs et en dernier les contrôles correctifs corrigent les incidents qui sont détectés.

2.3.14.1 Les contrôles préventifs

Comme nous venons de le mentionner plus haut le contrôle préventif est une catégorie préliminaire qui empêchent les incidents de sécurité, mais avant la mise en place de ce contrôle une stratégie de défense doit être sélectionnée afin d'empêcher l'exploitation des vulnérabilités, les politiques sont un exemple parfait.

2.3.14.2 Les contrôles détectifs

Le contrôle détectif considéré comme le deuxième axe de sécurité, ce dernier permet de détecter les menaces qui a pu s'injecter, une stratégie d'atténuation doit être sélectionnée afin de réduire l'impact causé par l'exploitation d'une vulnérabilité et garantit que les attaques sont détectées tôt.

2.3.14.3 Les contrôles correctifs

C'est le dernier axe de sécurité qui est utilisé une fois les cyber-attaques ont infiltrés les deux premiers axes de sécurité. Une stratégie d'atténuation doit être sélectionnée pour mettre en place des contrôles correctifs afin d'assurer les corrections des menaces infligées par l'attaquant le plus tôt possible.

2.4 Conclusion

Dans ce chapitre, nous avons présenté les principaux concepts liés aux sécurités des Big Data, nous avons commencé par une présentation du principe des Big Data, ses caractéristiques, son fonctionnement ainsi que les différents domaines d'application. Ensuite nous avons recensé les différents types de vulnérabilité, sécurité et la vie privée. Dans le chapitre suivant, nous aborderons les approches et les travaux connexes qui proposent des solutions pour le problème de la vie privée et la préservation d'utilité des Big Data.

Chapitre 3

Approches et travaux connexes

‘Privacy is not an option, and it shouldn’t be the price
we accept for just getting on the Internet’.

Mr.Gary Kovacs

3.1 Introduction

A l'heure de Big Data, la protection et la protection des données fait plus que jamais l'objet de nombreux débats. En quête de nouveaux principes, les spécialistes veulent sensibiliser les individus et responsabiliser les organisations.

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 nous faisons une revue des mécanismes de protection de la vie privée dans les Big Data. Dans la partie qui suit, nous convergeons vers les travaux des chercheurs qui se sont intéressés aux vie privée et utilités des données dans les Big Data. Nous établirons une évaluation critique avec le point fort et faible pour chaque travail pour au final en construire un modèle comparatif.

3.2 La protection de la vie privée

3.2.1 La vie privée dans les Big Data

Le cycle de vie des données se compose en générale de trois phases ; la génération des données, le stockage et l'exploration de données. Chaque phase a ses propres mécanismes qui préservent la vie privée, nous allons les discuter ci-dessous.

1- La phase de la génération des données : Pour minimiser le risque de violation de la vie privée pendant cette phase en restreignant l'accès ou le propriétaire des données doit adopter des méthodes de contrôle d'accès efficaces afin que les données ne puissent pas être volées par un tiers, dans certaines circonstances, il n'est pas possible d'empêcher l'accès aux données sensibles. Dans ce cas, les données peuvent être falsifiées à l'aide de certains outils pour que les données ne soient récupérées par un tiers.

2- La phase de stockage des données : Lorsque les données sont stockées sur le cloud, la sécurité des données a principalement trois dimensions, la confidentialité, l'intégrité et la disponibilité. Les deux premiers sont directement liés à la vie privée, les fonctionnalité et les limites des mécanismes qui préservent la confidentialité sont résumer dans le tableau [3.1].

En ce qui concerne l'intégrité le tableau [3.2], résume les fonctionnalité et les limites des mécanismes de vérification d'intégrité.

Schéma de chiffrement	Les fonctionnalité	Les limites
'ID-based encryption'	• Le contrôle d'accès est basé sur l'identité d'utilisateur • Accès complet aux ressources	• Mise à jour de texte chiffré n'est pas possible • Consommation de temps de calcul • Les données doivent être téléchargées et décryptées pour les traiter
'Attribute-based encryption'	• Le contrôle d'accès est basé sur l'attribut d'utilisateur • Plus sûr et flexible	• Mise à jour de texte chiffré n'est pas possible • Les données doivent être téléchargées et décryptées pour les traiter
'proxy re-encryption'	• Peut être déployé dans IBE ou ABE • possibilité de mise à jour de texte	• Frais généraux de calcul • Les données doivent être téléchargées et décryptées pour les traiter
'homomorphic encryption'	• Les calculs sont effectués sur les données cryptées • Très sécurisé	• Frais généraux de calcul très élevé

TABLE 3.1 – Mécanismes de protection de confidentialité .

Vérification d'intégrité	Les fonctionnalité	Les limites
'PDP' [1]	• Sécurisé pour la vérification des données à distance. • Fonctionne bien avec des données statiques et se base sur homomorphic.	• Non sécurisé dans un environnement dynamique en raison d'attaques de rejeu.
'POR' [2]	• Garantit la possession correcte des données. • *'ECC'*[3] sont utilisés pour récupérer des blocs corrompus	• Nombre limité de requêtes. • L'audit de données dynamiques est difficile en raison de *'ECC'*

1. Privacy and data protection
2. Proofs of retrievability
3. Error correcting codes

'Public auditing'	• L'audit est effectué par une tierce partie. • Utiliser les signatures BLS pour générer des valeurs d'authentification • Le système est prouvé pour être sécurisé	• Fuite de certaines informations lors du processus de vérification

TABLE 3.2 – Mécanismes de vérification d'integrité.

3- La phase l'exploration de données : Dans cette dernière phase, l'objectif est d'extraire des informations significatives à partir des données collectées qui peuvent contenir des informations sensibles, sans violer la vie privée*[Mehmood et al., 2016]*. Nous allons détailler les principes de protection de la vie privée pendant cette phase dans la section qui suit .

3.2.2 La vie privée et *Data-Maning*

Comme nous l'avons mentionné dans la section précédente, l'objectif est de générer de nouvelles informations à partir des données collectées et sans violer la vie privée. Les données collectées doivent être modifiées, de sorte que les informations sensibles ne puissent pas être trouvées et d'un autre coté assurer que l'utilité suffisante des données est conservée, ce processus appelé PPDP[4]. Il se base sur l"anonymisation des données, il existe plusieurs opérations afin de rendre les données anonymes.

- **Généralisation et suppression :** La généralisation signifie qu'une valeur de domaine parent est remplacée par ses valeurs de domaine enfant dans son arborescence de taxinomie de domaine, l'opération de suppression remplace certaines valeurs par une valeur spéciale, par exemple '*'.
- **Anatomisation et permutation :** Les deux opérations ne modifient pas les quasi-identifiants[5] ou les attributs sensibles, mais elles désassocient les relations entre eux.
- **Perturbation :** En perturbation, les valeurs de données d'origine sont remplacées par certaines valeurs de données synthétiques par exemple l'ajout de bruit*[Xu et al., 2014]*.

Plusieurs méthodes d'anonymisation ont été développés tels que : **k-anonymity**, signifie que le nombre d'enregistrements anonymes correspondant à un quasi-identifiant doit être supérieur à un seuil. Ce modèle est pour but d'empêcher les liaisons entre enregistrements. **l-diversity** exige que les valeurs sensibles correspondent à un quasi-identifiant n'est pas inférieur à un seuil. Ce modèle a pour but d'empêcher les liaisons entre attributs et les liaisons entre enre-

4. privacy preserving data publishing
5. les attributs qui peuvent être liés à des données externes pour ré-identifier des enregistrements individuels.

gistrements, il est plus robuste que k-anonymity. **t-closeeness**, nécessite que la distribution des valeurs sensibles corresponde à un quasi-identifiant pour être proche de celle des ensembles de données originaux. Ce modèle à pour but d'empêcher les attaques probabilistes et les liaisons entre attributs*[Wang et al., 2010]*.

3.3 L'utilité dans les Big Data

3.3.1 Récupération des volontés des utilisateurs

Les auteurs *[Gkiotsalitis and Stathopoulos, 2015]*, ont proposé un modèle de maximisation d'utilité pour capturer automatiquement la volonté des utilisateurs de parcourir une certaine distance pour participer à différents types d'activité dans une ville dense.

1. **La structure du modèle proposé :**

   Ce modèle est basé sur l'analyse des données générées par les utilisateurs à partir des réseaux sociaux afin de sélectionner les activités conjointes.La structure du modèle proposé se compose de trois modules :

   - **Module de reconnaissance (PRM) :** Ce module est proposé avec une double portée : traiter automatiquement des volumes massifs de données générées par les utilisateurs et capturer les schémas de mobilité des utilisateurs et relier les types d'activité aux emplacements géo-marqués sur la base de l'analyse spatio-temporelle des interactions des utilisateurs.
   - **Module de maximisation de l'utilité des données :** Les individus choisissent de participer à différents types d'activités en fonction d'un mécanisme de prise de décision qui cherche à maximiser leur niveau de satisfaction par une fonction d'utilité. L'indice de satisfaction pour la participation à différents types d'activités en fonction de leur distance par rapport à l'emplacement précédent est défini sous la forme d'une fonction d'utilité linéaire, ils ont appliqué l'algorithme d'optimisation BHHH [6] proposé par, *[Berndt et al]* pour estimer les coefficients par optimisation lorsqu'un modèle non-linéaire est ajusté aux données.
   - **Regroupement des utilisateurs en fonction de leurs modèle d'utilité :** Après la dérivation des modèles d'utilisateurs et de leurs modèles d'utilité, les profils d'utilisateurs sont regroupés en fonction de leur volonté. Dans cette phase de regroupement, les utilisateurs sont traités comme des entités sans informations personnelles via l'utilisation d'un ID aléatoire. Afin de calculer la distance entre deux utilisateurs et comparer la distance entre leurs modèles de maximisation d'utilité au fil du

6. Berndt–Hall–Hall–Hausman

temps. le regroupement basé sur la densité des applications avec bruit (DBSCAN) proposé par [Ester et al,1996] et modifié par [Gkiotsalitis et Alexandrou, 2014].

2. **Les inconvénients du modèle :**
   - Ils ont considéré l'utilité des données comme une caractéristique fondamentale.
   - L'utilisation d'un ID aléatoire n'atteint pas un niveau suffisant de protection de la vie privée des utilisateurs.

3.3.2 Exploitation sociale et vie privée

[Monreale et al., 2014] ont proposé une instanciation du paradigme de la protection de la vie privée dès la conception Pbd [7], introduit par *[Anne Cavoukian, 1990]*. Ils l'ont appliqué dans différents domaines afin de garantir la protection de la vie privée, tels que :

1. ***mobility data publishing :*** ils ont proposé une méthode pour la publication des données de mouvement basées sur le respect de la vie privée, qui permet une analyse de classification pour comprendre le comportement de la mobilité humaine dans des zones spécifiques. Les trajectoires sont rendues anonymes par la généralisation des trajectoires d'origines.
   - **Les inconvénients de cette méthode :**
   - Il est possible de déduire des trajectoires dans certains cas car l'opération de généralisation n'est pas suffisante pour l'anonymisation des données.
   - Cette méthode ne traite pas les problèmes d'intégrité et de confidentialité.
2. **Systèmes analytiques distribués :** ils ont proposé une méthode pour une analyse des données distribuée en fonction du respect de la vie privée, une station centrale recueille des statistiques agrégées et calculées par chaque nœud et les analyse, cette méthode fournie une protection de la vie privée au niveau individuel des nœuds par l'application du modèle *'Differential Privacy'*.
   - **Les inconvénients de cette méthode :**
   - L'inconvénient majeur de cette méthode réside dans la vraisemblance des données fictives.
   - Les risques de conflits liés à la complexité des tâches à effectuer par la méthode et une augmentation importante du temps de réponse .
   - *'Differential Privacy'* est trés difficile à implémenter d'une manière fiable, ainsi, elle perturbe les données des nœuds avant de les transmettre à la station centrale d'une manière grossière .

7. Privacy by design

3.3.3 Préservation de la vié privée dans le Cloud

[Zhang et al., 2014], ont proposé un modèle de préservation de la vie privée dans le cloud en fonction des quatre exigences la flexibilité, la scalabilité, le dynamisme et la rentabilité.

1. **Le modèle proposé :** Ce modèle se compose de quatre modules principaux
   - **L'anonymisation des données (DA) :**Dans ce module ils ont utilisé l'opération de généralisation pour l'anonymisation des données.
   - **La mise à jour des données (DU) :** Trois opérations de base sont prévues dans ce module : la mise à jour des données car dans le cloud les données sont dynamiques et augmentent considérablement au fil du temps, La généralisation est utilisée pour élever le niveau d'anonymisation tandis que la spécialisation consiste à abaisser le niveau d'anonymisation.
   - **La gestion des ensembles de données anonymes (ADM) :** Le dernier module ADM, utilise directement l'infrastructure de cloud pour accomplir les tâches.
   - **L'interface de spécification de la vie privée (PSI) :** Ce module est représenté par un vecteur de paramètres présenté dans le tableau [3.3]

Paramètres	Représentation
'PMN'	Le nom d'un modèle de protection de la vie privée, ils ont utilisé trois principes définit précédemment ; k-anonymity, l-diversity et t-closeeness.
'Thr'	Le seuil du modèle de spécification de la vie privée, c'est-à-dire, k, l et t dans les trois principes qui se précède.
'AT'	Indique le type d'application.
'Agl'	Ce paramètre indique les algorithmes d'anonymisation.
'Gra'	Ce paramètre représente la granularité de la spécification de la vie privée.
'Uti'	Ce paramètre représente l'utilité des données.

TABLE 3.3 – Les paramètres du module PSI.

2. **Les inconvénients du modèle :**

   L'utilisation de la généralisation comme modèle d'anonymisation est insuffisante pour avoir un niveau acceptable du protection de la vie privée, mais aussi les auteurs n'ont pas pris en considération le problème de la confidentialité.

3.3.4 Membership

[Li et al., 2013] ont étudié le problème de la divulgation de la vie privée et les incidents liées à celle-ci, par la suite ils ont proposé un Framework pour la protection de la vie privée des membres sous le nom *'Membership Privacy'*.

1. **Framework proposé :** Ce Framework comprend deux modules principales, PMP *'Positive Membership Privacy'* et NMP *'Negative Membership Privacy '*
   - **PMP *'Positive Membership Privacy'* :**
     Ce module, a pour but d'empêcher la divulgation positive des membres, Afin d'établir une connexion claire par le mécanisme d'anonymisation *differential privacy*.
   - **NMP *'Negative Membership Privacy'* :**
     Le rôle principale de ce module est d'empêcher un adversaire d'augmenter significativement la confiance que les données d'une entité particulière ne se produisent pas dans l'ensemble de données.
2. **Les inconvénients du Framework :** Lors de l'évaluation de ce modèle plusieurs inconvénients ont été découverts :
   - La vraisemblance des données fictives causée par l'utilisation de *differential privacy* comme mécanisme d'anonymisation.
   - Les auteurs n'ont pas pris en compte le problème d'utilité des données et ce Framework ne résolve pas les problèmes d'intégrité et volume des données.

3.3.5 Amélioration de la réplication des données

[Villasenor et al., 2014], ont développé MorphStore système de fichiers local qui améliore les performances lors de l'accès à des fichiers volumineux.

1. **Le système proposé :** MorphStore se compose de deux modules principaux :
   - **LAAS 'Load-adaptive access scheduling' :** C'est un composant clé de ce système ; qui décide si des requêtes sont émises pour exploiter le striping entre réplicas, ils ont utilisé une valeur seuil configurable qu'ils ont défini sur le nombre de périphériques de la matrice de stockage, Si la charge mesurée dépasse un seuil, MorphStore passe à une configuration de charge élevée dans laquelle les demandes sont envoyées à des disques spécifiques. En revanche, dans le cas d'une opération à faible charge les demandes sont réparties sur les répliques disponibles.
   - **Utility-driven Replication (UDR) :** Le deuxième composant de MorphStore est la réplication utilitaire qui vise à utiliser la capacité de réplication pour maximiser les performances. Le nombre de réplicas à générer pour chaque fichier est décidé en fonction de la notion d'utilité, qui est déterminée par le nombre d'accès en lecture et en écriture à ce fichier.
2. **Les inconvénients de MorphStore :**

   Ce système souffre de deux problème importants, le premier réside dans le traitement des

données d'une manière brut, ce qui augmente le risque de violation de la vie privée. Le second dans l'utilisation de la réplication qui élargie le volume de données.

3.4 Compromis utilité et vie privée hors Big Data

Le problème de la vie privée et de la fuite d'informations a été étudiées pendant plusieurs décennies par des communautés de recherche. Les approches du problème fondées sur la théorie de l'information sont rares et se sont principalement concentrées sur l'utilisation de mesures théoriques de l'information. Les stratégies respectueuses de la vie privée devraient trouver un compromis acceptable entre la vie privée et l'utilité des données et doit être l'objectif principal.

3.4.1 Approche informationnelle

[Sankar et al., 2013]., ont proposé une approche analytique qui garantie un compromis entre l'utilité et la vie privée, dans lequel certains attributs doivent rester privés alors que la source peut révéler une fonction de certains ou de tous les attributs K.

Ils ont noté Kr et Kc pour désigner des ensembles d'attributs privés (indice c pour cacher) et public (indice r pour révéler), tels que $Kr \cup Kc = K \equiv \{1, ..., K\}$, en outre les collections correspondantes d'attributs publics et privés par $XKr \equiv \{Xk\}k \in Kr$ et $XKc \equiv \{Xk\}k \in Kc$. Plus généralement, $XSc \equiv \{Xk : k \in Sc \subseteq Kc\}$, et $XSr \equiv \{Xk : k \in Sr \subseteq Kr\}$, pour désigner des sous-ensembles d'attributs privés et publics, respectivement. Ils ont considéré que cette notation permet à un attribut d'être à la fois public et privé ; ceci est dû au fait qu'une base de données peut avoir besoin de révéler une fonction d'un attribut tout en gardant l'attribut lui-même privé.

- **Les inconvénients :**
  - N'est pas applicable aux grandes bases de données, car la corrélation survient généralement entre les attributs et peut être ignorée entre les entrées en fonction de la taille de la base de données.
  - Définit la vie privée et l'utilité comme des caractéristiques fondamentales des sources de données qui peuvent être en conflit et peuvent être échangées.

3.4.2 Anonymisation sans regroupement

[Nagendrakumar et al., 2014], ont proposé, un modèle d'anonymat sans regroupement ; qui sert à donne un niveau basique d'anonymisation qui empêche la ré-identification d'un individu à partir de ses données publiées par l'application minimales des opérations de la généralisation et de la suppression afin de conserver des micro-données d'origine et obtenir une utilité optimale des données.

Il ont proposé un algorithme qui prend dans son entrée les données origines $D = \{d1, d2, \ldots\ldots, dn\}$, et produire à sa sortie des données anonyme $D' = \{d'1, d'2, \ldots..., d'n\}$. Ce algorithme [1] comprend 7 étapes :

Algorithm 1: Algorithme anonymisation sans regroupement

Input : Original micro data set $D = d1, d2, \ldots.dn$.

Output : Anonymized data set D* = $d1', d2', \ldots.dn'$.

Step 1 : Run Grubb's test.

Step 2 : **If** Grubb's p-value>0.05, **then** remove outliers.

Step 3 : **Else** p-value>0.05, **then** remove SuppressExplicit identifier S ← (DEi).

Step 4 : Move QI-attributes $Dqi = q1, q2, \ldots..qn$ to Table **T1** and rest to table **T2**.

Step 5 : **For** every Quasi Identifier attribute Dqi IN **T1**, initialize a generalization function Gf←Dqi

Step 6 : Join both tables **T1** and **T2**. D*←**T1** ***JOIN*** **T2**

Step 7 : Publish **D***

3.4.3 Approche basée sur un contrat

[Xu et al., 2015], ont proposé une nouvelle approche théorie contractuelle, cette approche est basée sur un contrat ente le collectionneur des données et le fournisseur afin d'équilibrer le compromis entre la protection de la vie privée et l'utilité des données. En effet, un haut niveau d'anonymisation implique la bonne protection de la vie privée des propriétaires, ainsi les fournisseurs sont prêts à fournir plus de données ou exiger moins de compensation. En ce sens, l'anonymisation est bénéfique pour le collectionneur. Cependant, un haut niveau d'anonymisation entraîne également une diminution importante de l'utilité des données, ce qui signifie que le collectionneur bénéficiera moins des données. La conception du contrat optimal, repose sur deux étapes. En premier lieu, la détermination de la nature des données et le niveau de protection. Ensuite, optimise le gain du collectionneur par rapport au niveau d'anonymisation.

3.5 Modèle de Comparaison

3.5.1 Table de comparaison

Approches & Critéres	Volume	Variété	Vélocité	Agrégation	Intégrité	Confidentialité	Disponibilité	Contrôle d'accès	Cloud Fiabilité	L'anonymisation	Protection de la vie privée	L'utilité
[Mehmood et al., 2016]	Aucune proposition n'a été trouvée											
[Xu et al., 2014]	Aucune proposition n'a été trouvée											
[Wang et al., 2010]	Aucune proposition n'a été trouvée											
[Gkiotsalitis and Stathopoulos, 2015]	✓	✓	(-)	✓	X	X	X	X	(-)	X	X	✓
[Monreale et al., 2014] *mobility data publishing*	✓	(-)	✓	X	X	X	(-)	(-)	X	Généralisation	Faible	Moyen
[Monreale et al., 2014] Systèmes analytiques distribués	✓	✓	(-)	✓	X	X	X	✓	✓	*Differential privacy*	Élevé	Faible
[Zhang et al., 2014]	✓	(-)	✓	(-)	X	X	X	✓	✓	*k-anonymity*	Moyen	Moyen
[Li et al., 2013]	X	X	X	X	X	✓	X	✓	✓	*Differential privacy*	Élevé	Faible
[Villasenor et al., 2014]	X	X	✓	X	✓	X	✓	✓	(-)	X	X	✓
[Sankar et al., 2013]	X	X	X	X	X	X	(-)	(-)	(-)	X	Faible	Faible
[Nagendrakumar et al., 2014]	X	X	X	(-)	(-)	(-)	(-)	(-)	(-)	Généralisation Suppression	Faible	Moyen
[Xu et al., 2015]	X	X	X	X	X	(-)	(-)	(-)	(-)	✓	✓	✓

TABLE 3.4 – Comparaison des approches.

3.5.2 Synthèse des travaux existants

Après avoir étudier chaque approche et établi le tableau comparative 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.
- La plupart des approches étudiées n'ont pas pris en considération la variété des données.
- L'utilité des données est vraiment faible ou moyenne pour chaque travail

3.6 Conclusion

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

Chapitre 4

Conception du système

‘Every company has Big Data in its future and every company will eventually be in the data business’.

Mr.Thomas Davenport

4.1 Introduction

L'un des problèmes fondamentaux dans les Big Data est de savoir comment faire le bon compromis entre la protection de la vie privée et la préservation de l'utilité des données. Après avoir effectuer un survol sur les divers travaux de chercheurs concernant ce sujet dans le chapitre précédant. On va s'intéresser a présent dans ce chapitre à l'architecture globale de notre système, 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 Conception générale du système proposé

4.2.1 Architecture globale

La visée principale de cette section est de concevoir une architecture générale afin de protéger la vie privée tout en préservant l'utilité des données.

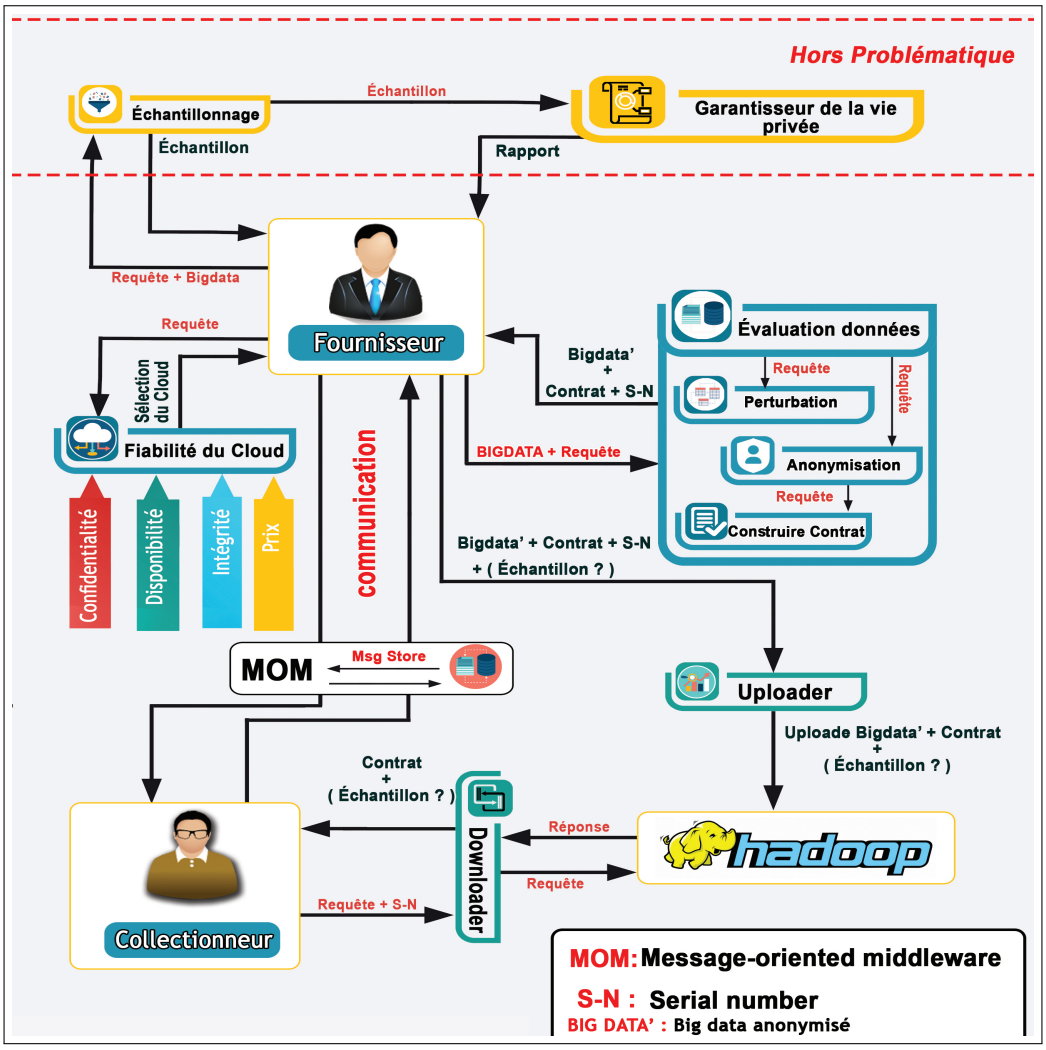


FIGURE 4.1 – L'architecture globale du système proposé

- **Description du système proposé :** Notre architecture est composée d'une collection de composants afin de trouver un bon compromis entre la protection de la vie privée et préservation d'utilité des données, à cet égard on a utilisé un ensemble de composants : le composant évaluation des données, ce dernier inclus trois composants : le composant perturbation, le composant d'anonymisation des données et le composant contrat.
  Et on a le composant fiabilité du cloud, et les deux derniers composants *Uploader* et *Downloader*. Tout ces composants travaillent en collaboration pour assurer la protection de la vie privée tout en gardant un niveau suffisant d'utilité des données.

4.2.2 Architecture détaillée

4.2.2.1 Le composant fiabilité du *cloud*

- **L'architecture du composant fiabilité du cloud :**

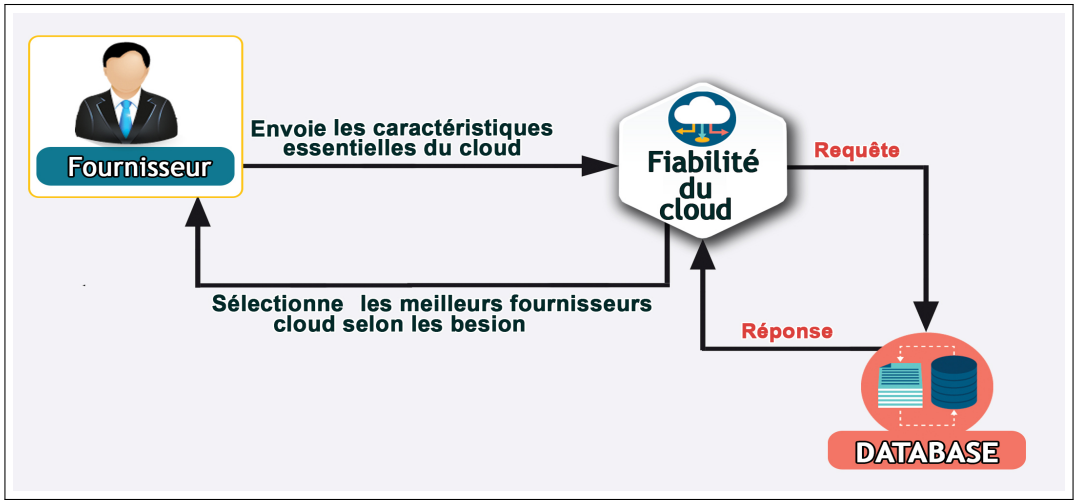


FIGURE 4.2 – L'architecture du composant fiabilité du cloud

- **Le rôle du composant fiabilité du cloud :**
  - La réception des caractéristiques essentielles du cloud d'après le fournisseur des données.
  - Citer les caractéristiques de chaque fournisseur cloud.
  - Proposer les meilleurs fournisseurs cloud selon les besoins du fournisseur des données .

4.2.2.2 Le composant évaluation des données

- **L'architecture du composant évaluation des données :**

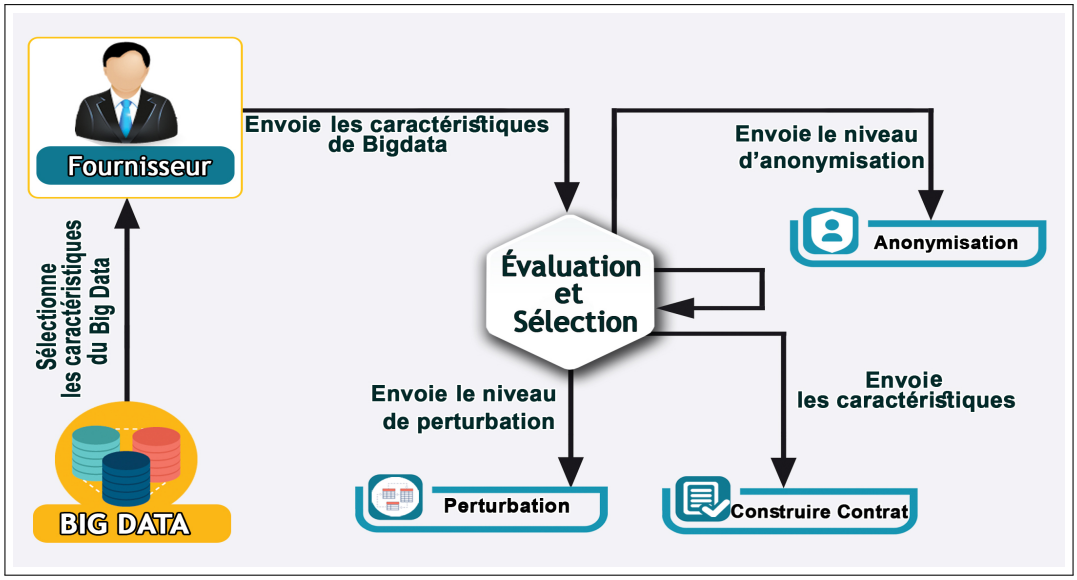


FIGURE 4.3 – L'architecture du composant évaluation des données

- **Le rôle du composant évaluation des données :**
  - La réception des caractéristiques de Big Data d'après le fournisseur des données.
  - L'évaluation des caractéristiques et la sélection du niveau de perturbation, d'anonymisation et les caractéristiques du contrat .
  - L'envoi du niveau de perturbation des données au composants perturbation.
  - L'envoi du niveau d'anonymisation des données au composant anonymisation.
  - L'envoi des caractéristiques principales au composants contrat.

4.2.2.3 Le composant d'anonymisation des données

- **L'architecture du composant d'anonymisation des données :**

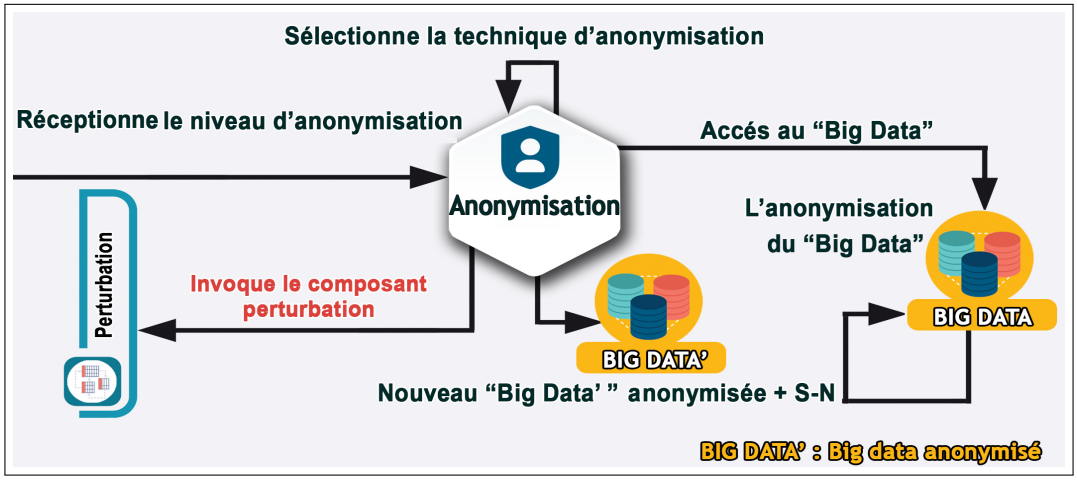


FIGURE 4.4 – L'architecture du composant d'anonymisation

- **Le rôle du composant d'anonymisation des données :**
  - La réception du niveau d'anonymisation d'après le composant évaluation des données.
  - Accès au Big Data, l'anonymise et génère un nouveau "Big Data ' " et un numéro de série.
  - Invoque le composant perturbation.

4.2.2.4 Le composant perturbation

- **L'architecture du composant perturbation :**

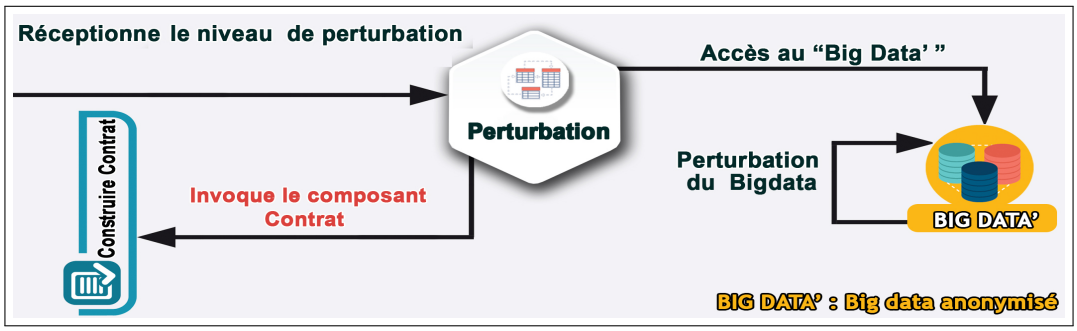


FIGURE 4.5 – L'architecture du composant perturbation

- **Le rôle du composant perturbation :**
  - La réception du niveau de perturbation d'après le composant évaluation des données.
  - Accès au Big Data' et assurer la perturbation.
  - Invoque le composant contrat.

4.2.2.5 Le composant contrat

Le contrat est avantageux pour le collectionneur et le fournisseur de données pour la préservation d'un niveau acceptable d'utilité des Big Data.

- **L'architecture de composant contrat :**

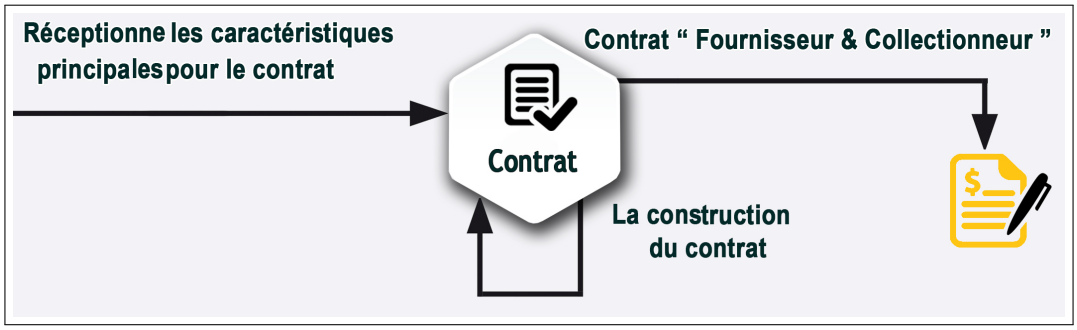


FIGURE 4.6 – L'architecture du composant contrat

- **Le rôle du composant contrat :**
  - La réception des caractéristiques principales.
  - La construction du contrat.

4.2.2.6 Les composant : Uploader & Downloader

- **L'architecture des : Uploader & Downloader :**

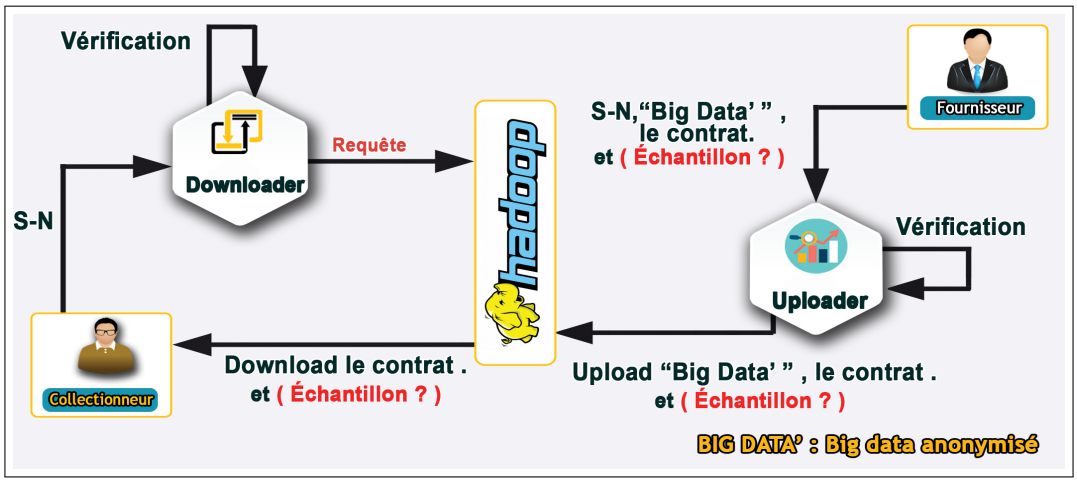


FIGURE 4.7 – Architecture des composants : Uploader & Downloader

- **Le rôle des composant Uploader & Downloader :**
  - Le composant *'uploader'* : Vérification du numéro de série, interroge *'Hadoop'* et crée un dictionnaire, transfère le "Big Data'" et le contrat vers le dictionnaire préalablement créé. Et transfère également l'échantillon dans le cas où il est disponible.
  - Le composant *'downloader'*,vérification du numéro de série, interroge *'Hadoop'* et a accès au dictionnaire, télécharge le "Big Data'" ou le contrat ou l'échantillon dans le cas où il est disponible.

4.3 Projection sur *Hadoop*

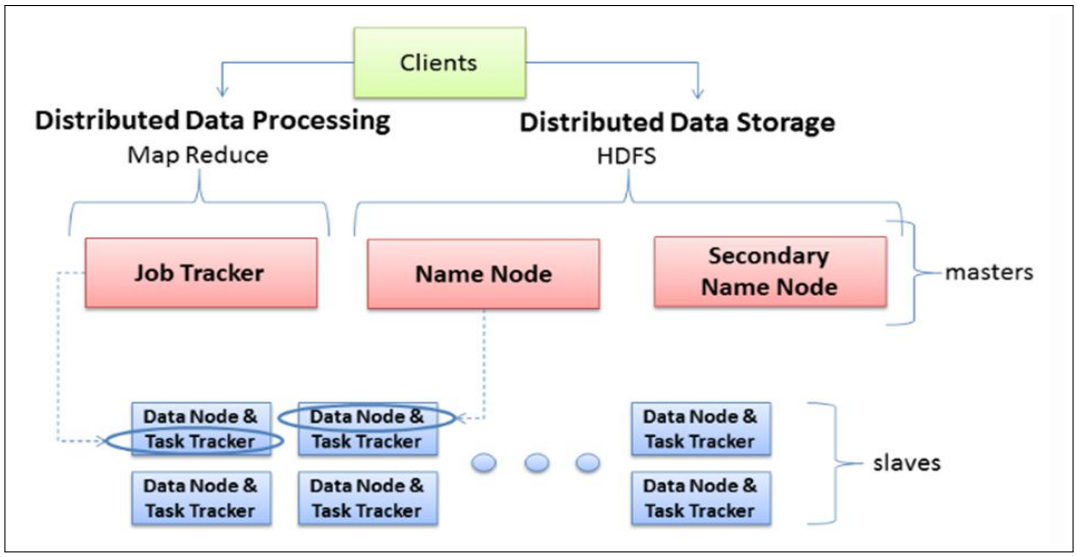


FIGURE 4.8 – Composants du noyau *Hadoop*

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 noeuds 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.

4.3.1 NameNode

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

4.3.2 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'*.

4.3.3 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.

4.3.4 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*.

4.3.5 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*).

4.4 Conception et modélisation détaillée avec *UML*

4.4.1 Diagramme de séquence

1- Diagramme de séquence du fournisseur

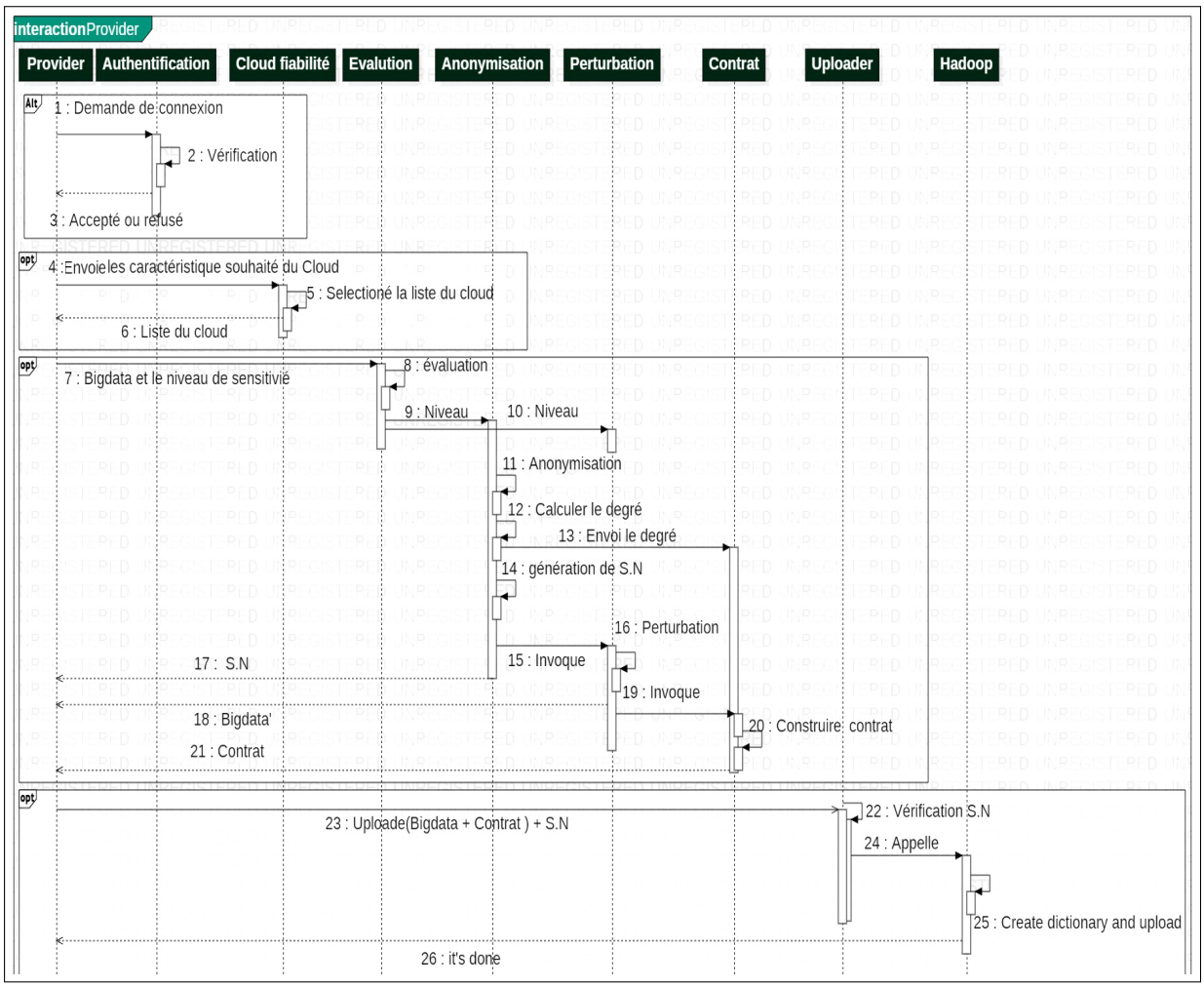


FIGURE 4.9 – Diagramme de séquence du fournisseur.

Description : L'utilisateur *'Provider'* doit être authentifier afin de bénéficier des services proposés :

- Le fournisseur choisit un cloud pour stocker ces données à l'aide du composant *cloud fiabilité* qui permet la classification des *clouds* appropriés pour le stockage des données, en tenant compte du prix introduit par le fournisseur des données et particulièrement le degré de sécurité.
- Le fournisseur choisit le niveau de sensibilité du Big Data et l'envoie au composant *évaluation* qui a son tour évalue ce niveau et envoie un niveau de perturbation et d'anonymisation au composant perturbation et anonymisation respectivement .

- Le composant anonymisation réceptionne le niveau d'anonymisation, génère un nouveau Big Data' anonymisé, calcule le degré d'anonymisation, et génère un numéro de série et invoque le composant *perturbation* qui s'occupe de la perturbation du Big Data' et invoque également le composant *contrat* afin de la construire un contrat .
- Finalement le fournisseur peut enregistrer le contrat et Big Data' dans la Framework *Hadoop* a l'aide du numéro série.

2-Diagramme de séquence du collectionneur

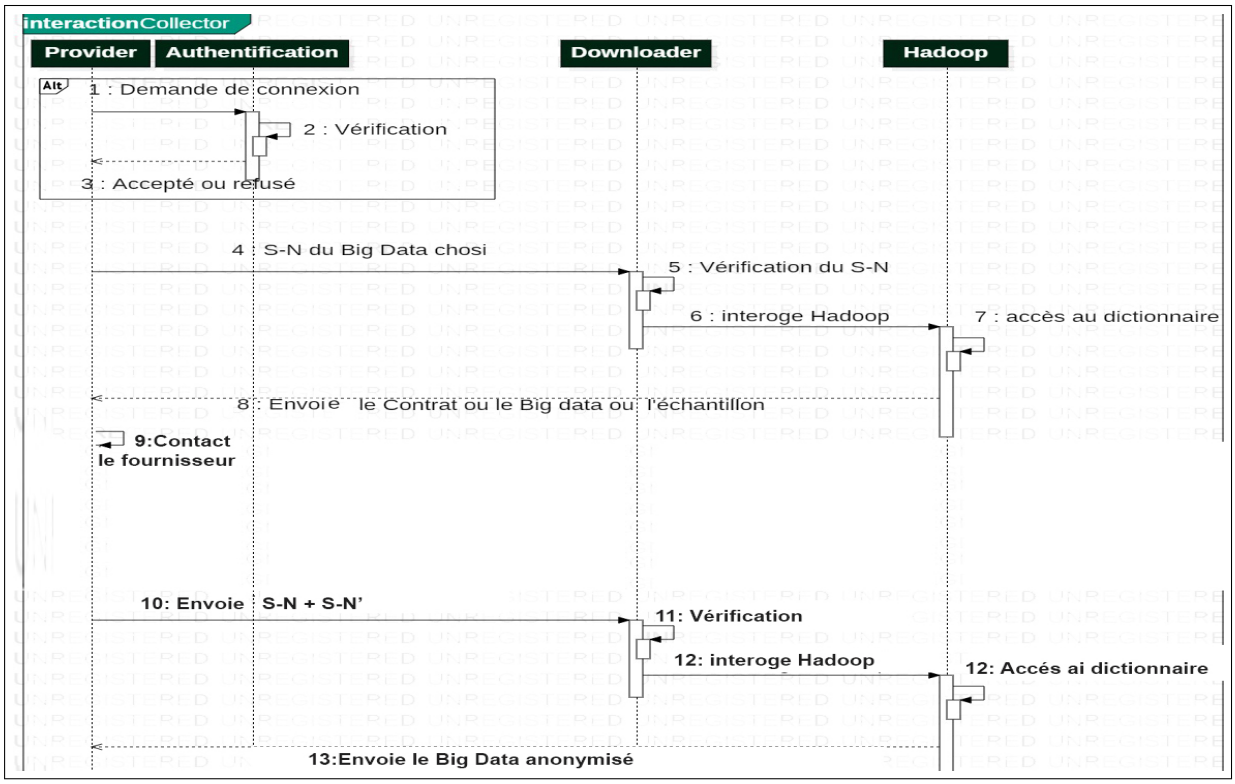


FIGURE 4.10 – Diagramme de séquence du collectionneur.

Description : L'utilisateur *'collectot'* doit être authentifier afin de bénéficier des services proposés :

- Le collectionneur choisit le type du Big data qu'il lui convient pour par la suite télécharger le contrat ou l'échantillonnage.
- Le collectionneur peut télécharger le contrat ou l'échantillonnage a partir du Framework *Hadoop* a l'aide de numéro série.
- Si le collectionneur veut télécharger le Big Data anonymisé, il doit contacter le propriétaire pour obtenir un autre numéro de série.

4.4.2 Diagramme d'activité

1- Diagramme d'activité du composant fiabilité *cloud*

Par le biais de ce composant, le fournisseur détermine ses exigences, ensuite une vérification sera lancée à l'aide des caractéristiques des fournisseurs *cloud* afin de choisir les meilleurs.

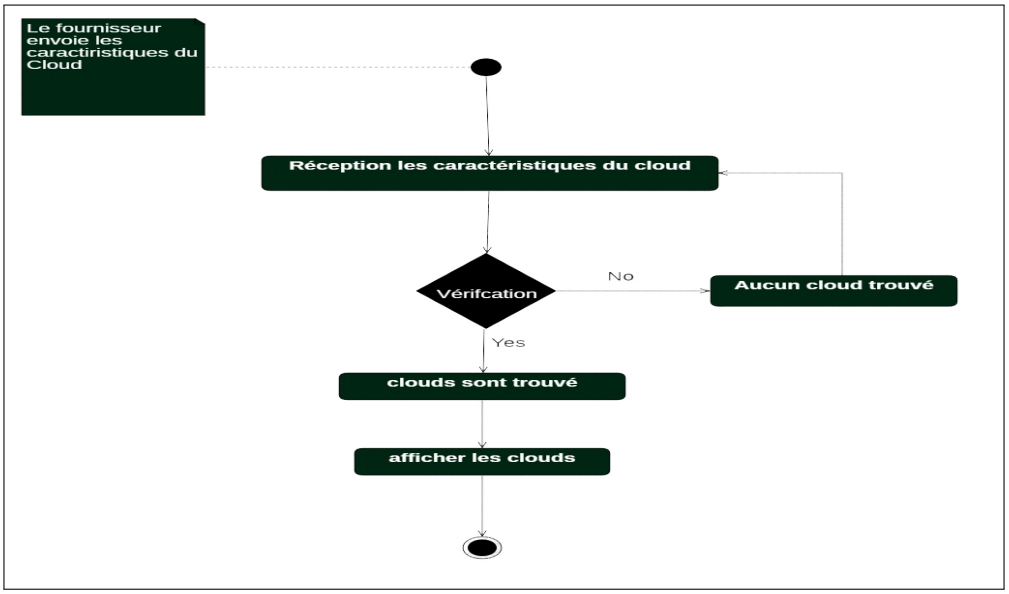


FIGURE 4.11 – Diagramme d'activité du composant fiabilité *cloud*

2- Diagramme d'activité du composant évaluation des données

Ce composant permet au fournisseur d'évaluer le niveau de sensibilité du Big Data.

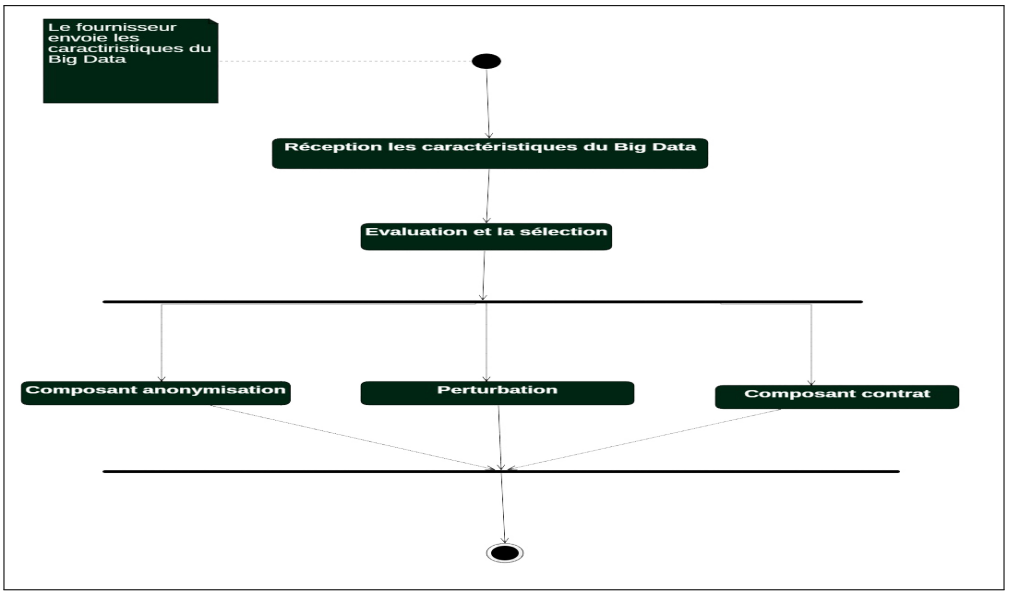


FIGURE 4.12 – Diagramme d'activité du composant évaluation des données.

3- Diagramme d'activité du composant anonymisation des données

Ce composant, a pour but d'ajouter un bruit au données afin de protéger la vie privée des utilisateurs

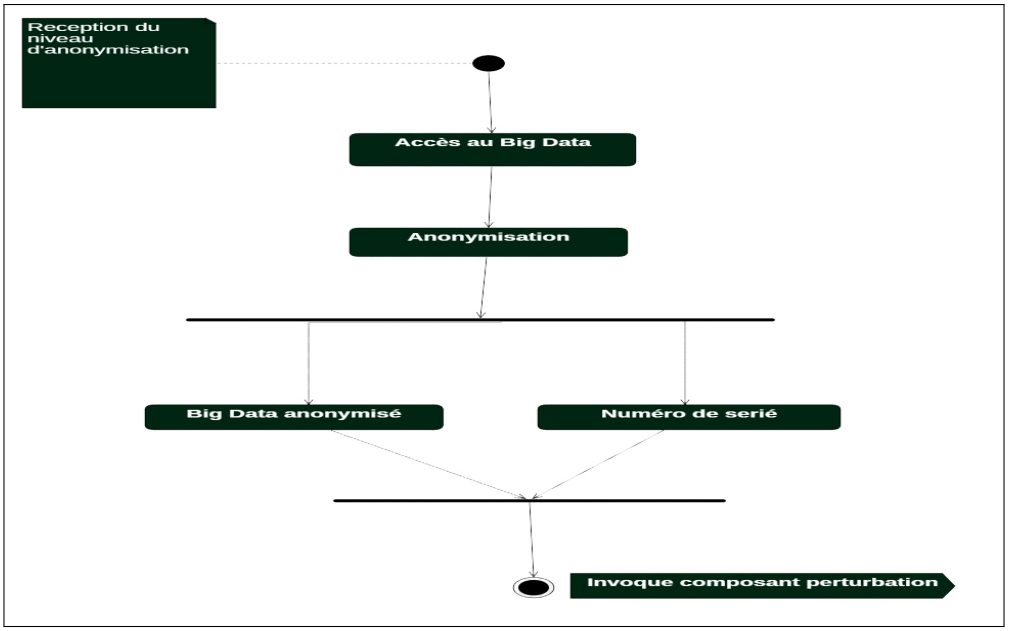


FIGURE 4.13 – Diagramme d'activité du composant anonymisation des données.

4- Diagramme d'activité du composant perturbation des données

Ce composant, a pour but d'ajouter des données synthétiques au Big Data d'origine.

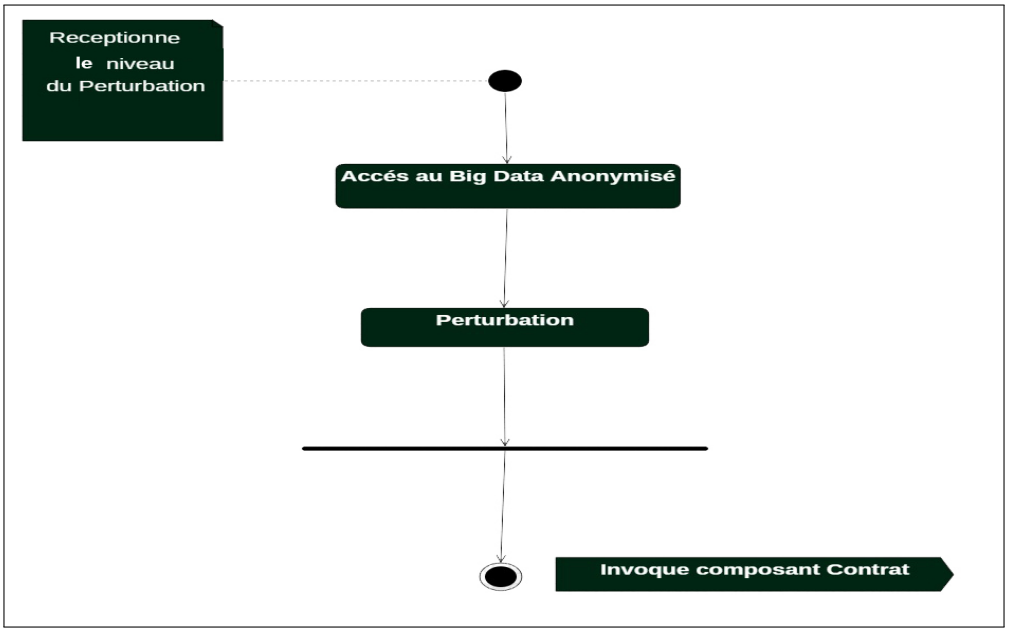


FIGURE 4.14 – Diagramme d'activité du composant perturbation des données.

5- Diagramme d'activité du composant contrat

Le but de ce composant est de construire un contrat pour les Big Data anonymisés.

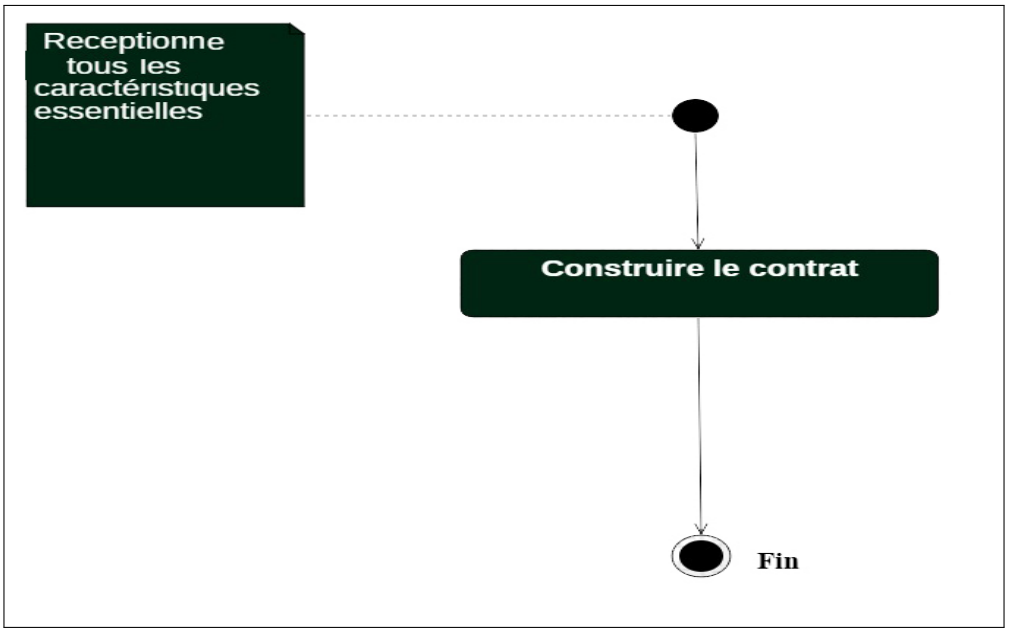


FIGURE 4.15 – Diagramme d'activité du composant contrat.

6- Diagramme d'activité des composants *Uploader* & *Downloader*

Le but de ces composants est d'interroger *Hadoop*, transférer ou télécharger des fichier a partir du *HDFS*.

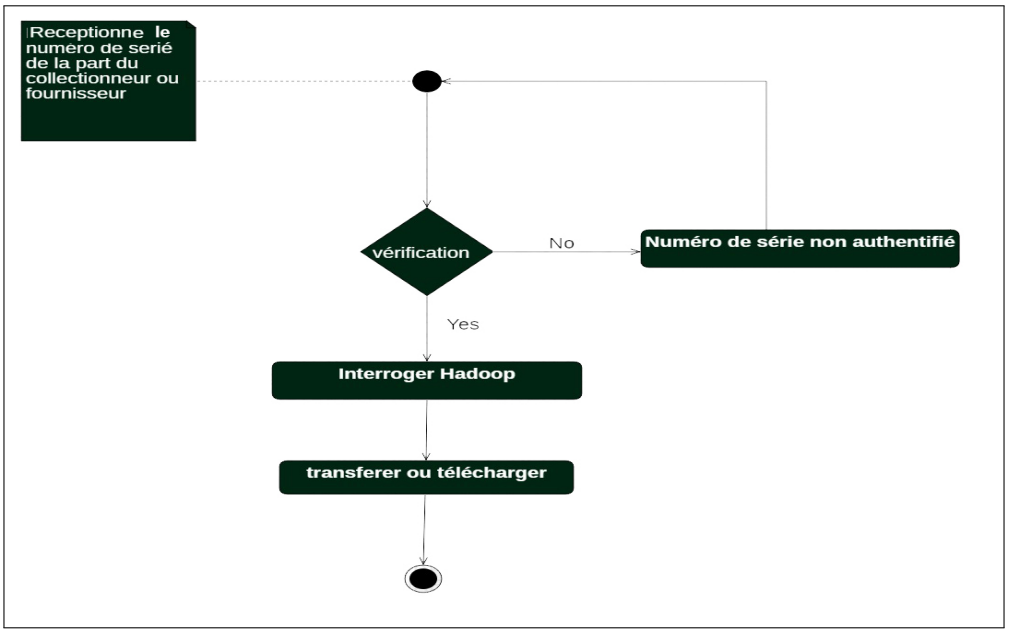


FIGURE 4.16 – Diagramme d'activité des composants *Uploader* & *Downloader*

4.5 Conclusion

Dans ce chapitre nous avons présenté notre système de la préservation de la vie privée en gardant un niveau suffisant d'utilité. 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

103 billion dollars, it will be Big data market size
revenue forecast worldwide by 2027.
according to the website
Reproduced from : www.statista.com

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 (64bits) qui est décrit avec la Figure suivante :

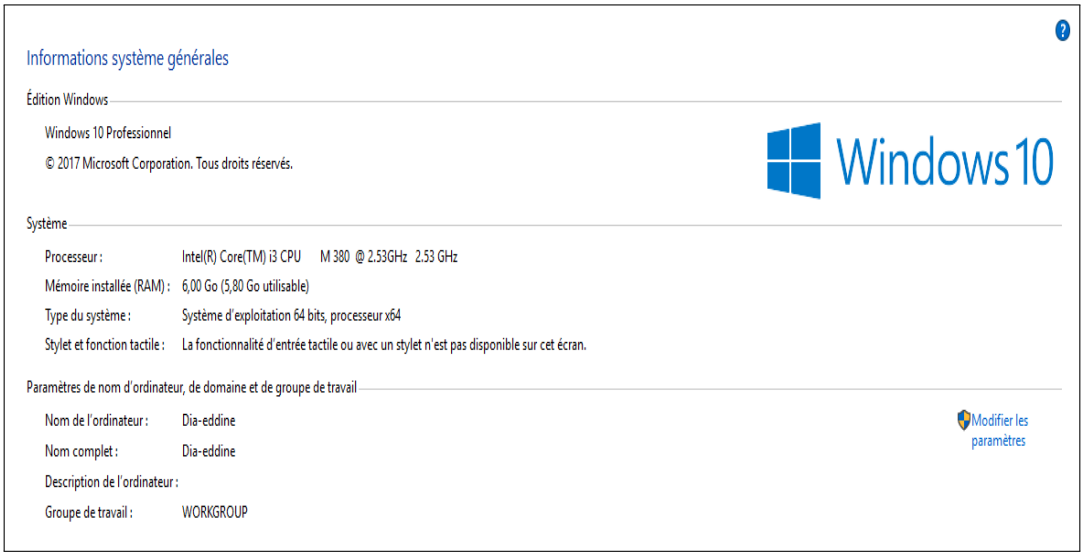


FIGURE 5.1 – Environnement matériel et logiciel

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.[Oracle, 2018]

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[IDE, 2018]

***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*, GoogleFS 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.[Hadoop, 2018]

***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.

5.2.3 Les données de test

Il existe plusieurs sites Web qui offrent ce service, malheureusement, aucune base de données massive (Data set) ne répond à nos besoins puisque la plupart des bases de données disponibles sont soit du texte (Sans relation) ou des bases qui ont des attributs qui ne sont pas suffisants pour le test.

Donc, nous avons choisi de générer notre propre Data set de manière aléatoire avec des scripts. Pour cela nous utilisons le *PHP MyAdmin* pour construire notre Data set. Ensuite, nous générons des données aléatoires pour créer un volume de données massif avec un (1) millions d'enregistrements, voici des parties [5.2,5.3] code de génération des données :

```
public static void main(String[] args) throws InterruptedException {
    for (int l = 1; l <= 6000; l++) {
        for (int y = 2017; y >= 2001; y--) {
            for (int m = 12; m >= 1; m--) {
                for (int i = 1; i <= 28; i++) {

                    new Privacy().conection(y + "-" + m + "-" + i);


                    System.out.println(y + "-" + m + "-" + i);
                    Thread.sleep(10);
                }
                System.out.println("mois " + m + "fin");
                Thread.sleep(10);

            }
            System.out.println("mois " + y + "fin");
            Thread.sleep(10);
        }
    }
}
```

FIGURE 5.2 – Code de génération des données "1"

1. Hadoop Distributed File System

```
    return age;
}
public String getssn(){
     Random random=new Random();
    String nbr=String.valueOf(random.nextInt(999) + 100);
    String nbr2=String.valueOf(random.nextInt(99) + 10);
    String nbr3=String.valueOf(random.nextInt(9999) + 1000);
    return nbr+"-"+nbr2+"-"+nbr3;
}

public String conection(String date) {

    try {
        Class.forName("com.mysql.jdbc.Driver");
        Connection con = DriverManager.getConnection("jdbc:mysql://localhost/privacy_utility", "root", "");
        PreparedStatement ps = con.prepareStatement("INSERT INTO `patient information` (`ID_patient`,`Date_test`, `First_name`, "
                + "`Last_name`, `Age`, `Genre`, `Disease`, `SSN`,`Job`,`Nationality`,`Situation`,`Country`) VALUES "
                + " (?,?,?,?,?,?,?,?,?,?,?,?)");

        int r1 = (int) (Math.random() * 80);
        int r2 = (int) (Math.random() * 80);
        int r3 = (int) (Math.random() * 40);
        int r4 = (int) (Math.random() * 40);
        int r5 =(int)   (Math.random()*180);
        int r6 =(int)   (Math.random()*118);
        int r7 =(int)   (Math.random()*4);
        int temp = (Math.random() <= 0.5) ? 1 : 2;
```

FIGURE 5.3 – Code de génération des données "2"

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 *Utivacy*, il est nécessaire de créer un compte qui sera utilisé pour s'authentifier. L'utilisateur doit fournir l'ensemble des informations requises et le mode d'utilisation (fournisseur/collectionneur) .Toute inscription incomplète ne sera pas validée.

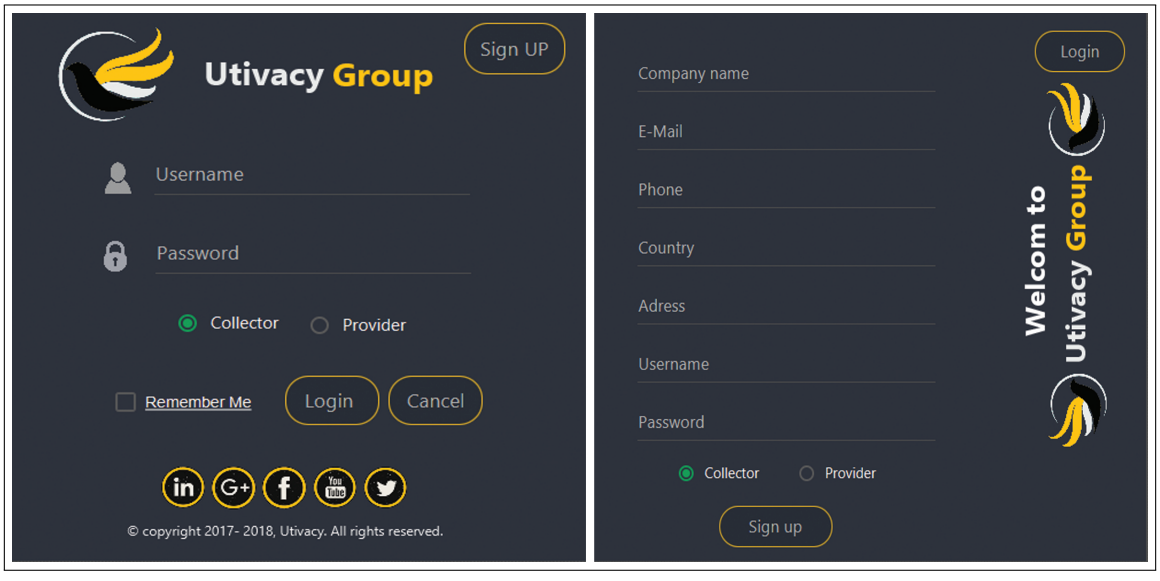


FIGURE 5.4 – Interface de connexion et inscription

5.3.2 Interface principale du fournisseur

L'interface principale du client fournisseur est représentée sur la Figure [5.5]. Elle lui permet d'accéder rapidement aux services fournis par *Utivacy*, ces derniers sont discutés ci-dessous.

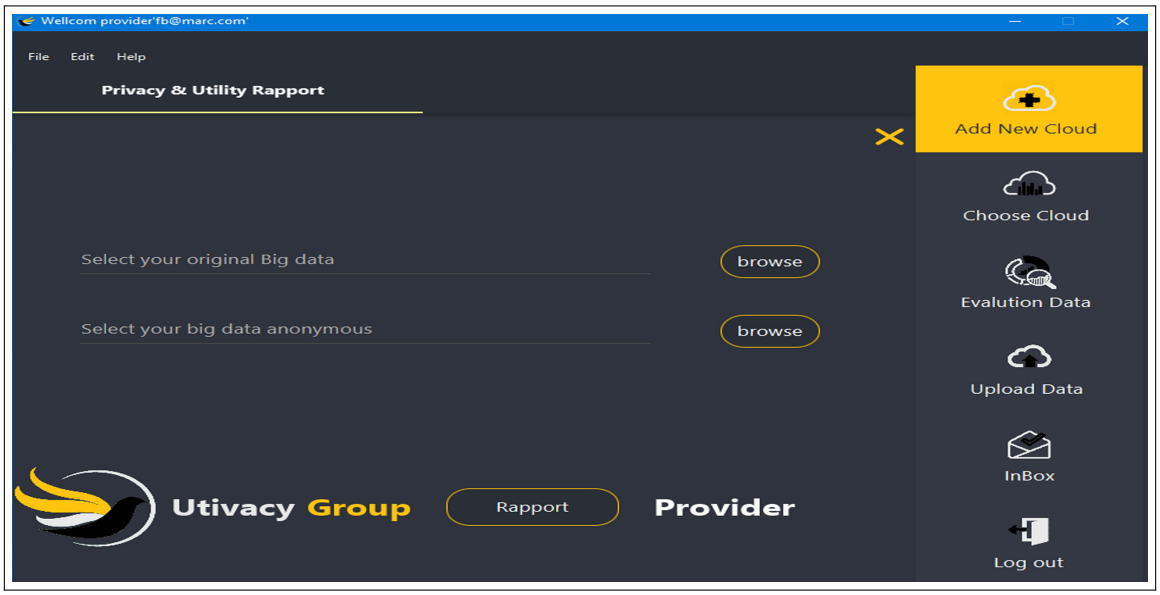


FIGURE 5.5 – Interface principale du client fournisseur

5.3.3 Service *"choose cloud"*

L'interface illustrée sur la Figure[5.6] permet de lister les meilleurs fournisseurs *cloud*. Il nécessaire d'évaluer les critères de la sécurité et le prix qui convient au besoin du collectionneur.

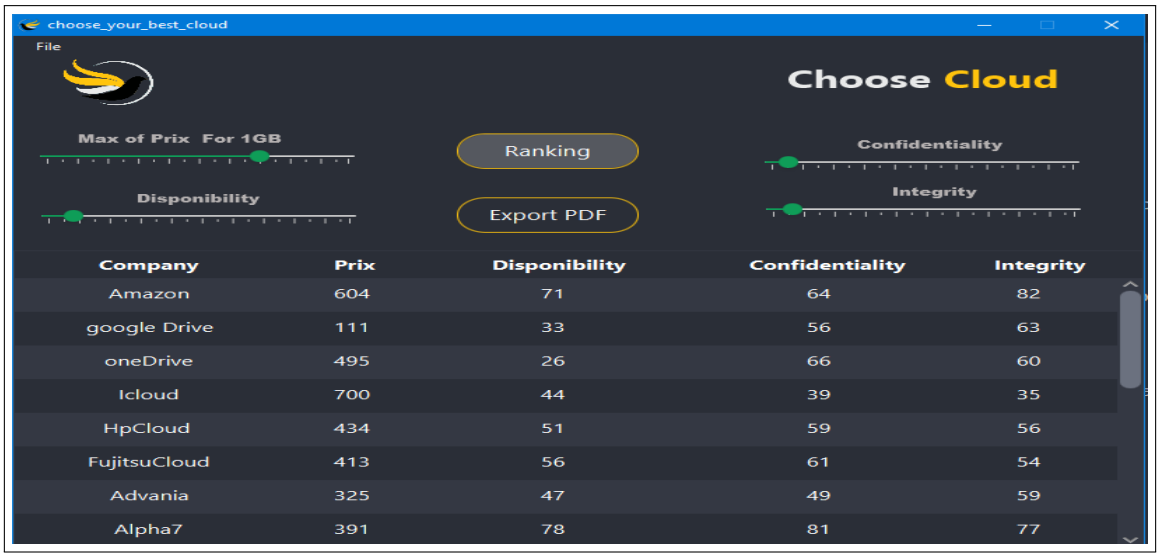


FIGURE 5.6 – Interface Service *"choose cloud"*

5.3.4 Service d'évaluation du Big Data

La Figure[5.7], illustre l'interface qui permet au fournisseur de sélectionner le fichier Big Data et son type, son prix et son niveau de sensibilité.

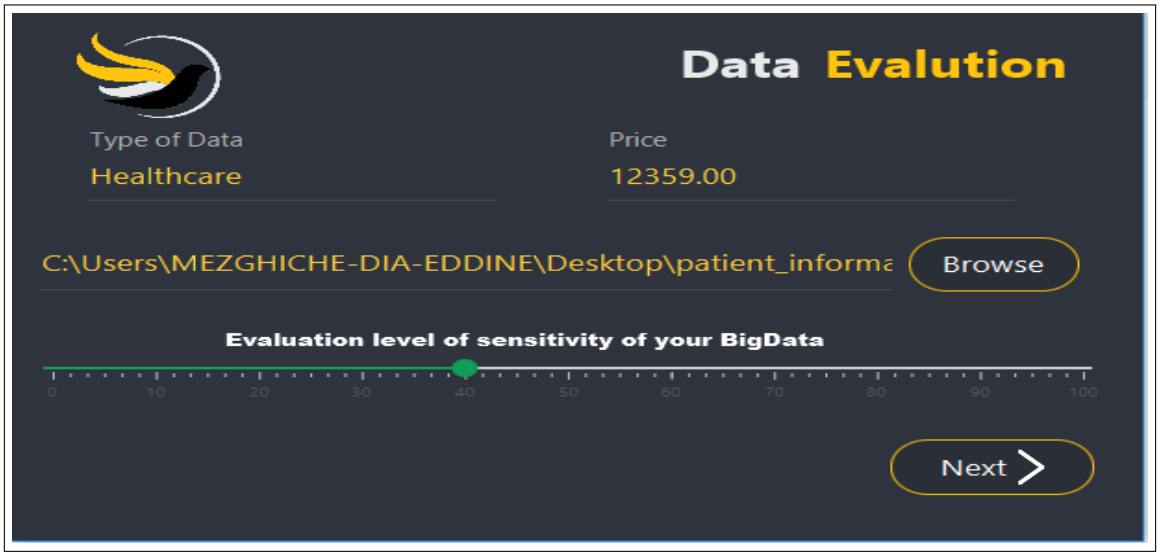


FIGURE 5.7 – Interface d'évaluation du "Big Data" "1"

La deuxième interface [5.8], lui permet de spécifier quels sont les attributs identifiant,quasi-identifiant et les attributs insensibles.

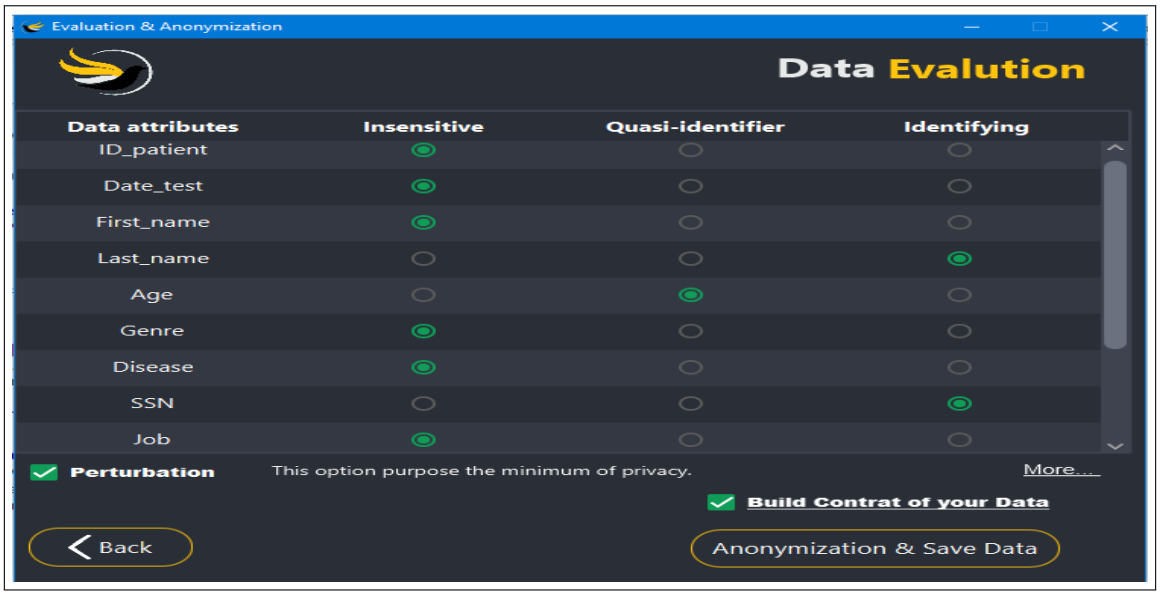
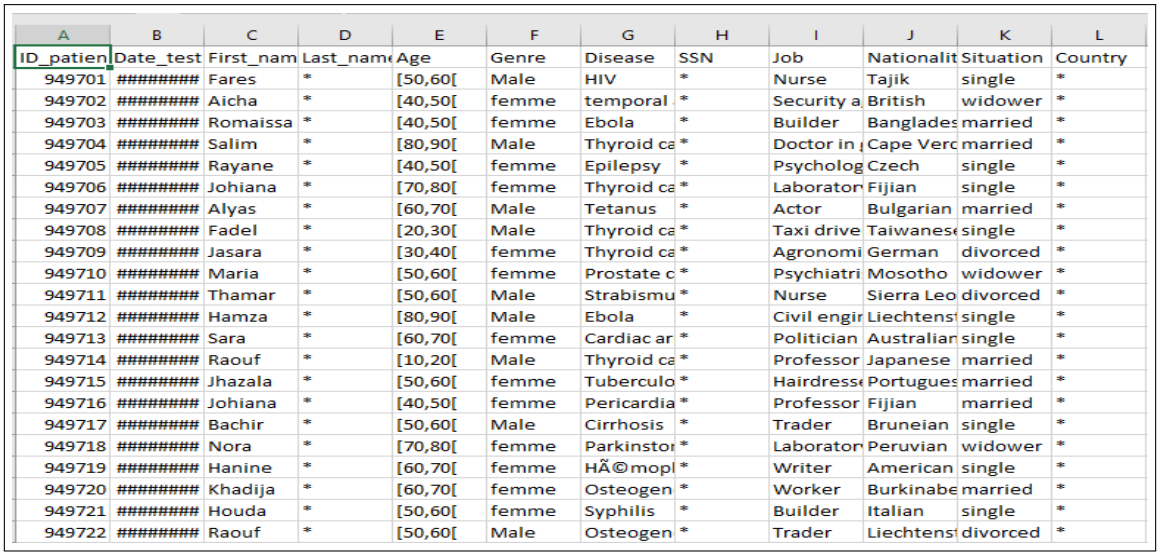


FIGURE 5.8 – Interface d'évaluation du "Big Data" "2"

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]. Un numéro de série se basant sur l'algorithme de chiffrement *AES64* [2], et un contrat (voir *Annexe A*).

2. Advanced Encryption Standard



A	B	C	D	E	F	G	H	I	J	K	L
ID_patien	Date_test	First_nam	Last_nam	Age	Genre	Disease	SSN	Job	Nationalit	Situation	Country
949701	########	Fares	*	[50,60[	Male	HIV	*	Nurse	Tajik	single	*
949702	########	Aicha	*	[40,50[	femme	temporal	*	Security a	British	widower	*
949703	########	Romaissa	*	[40,50[	femme	Ebola	*	Builder	Banglades	married	*
949704	########	Salim	*	[80,90[	Male	Thyroid ca	*	Doctor in	Cape Verc	married	*
949705	########	Rayane	*	[40,50[	femme	Epilepsy	*	Psycholog	Czech	single	*
949706	########	Johiana	*	[70,80[	femme	Thyroid ca	*	Laborator	Fijian	single	*
949707	########	Alyas	*	[60,70[	Male	Tetanus	*	Actor	Bulgarian	married	*
949708	########	Fadel	*	[20,30[	Male	Thyroid ca	*	Taxi drive	Taiwanese	single	*
949709	########	Jasara	*	[30,40[	femme	Thyroid ca	*	Agronomi	German	divorced	*
949710	########	Maria	*	[50,60[	femme	Prostate c	*	Psychiatri	Mosotho	widower	*
949711	########	Thamar	*	[50,60[	Male	Strabismu	*	Nurse	Sierra Leo	divorced	*
949712	########	Hamza	*	[80,90[	Male	Ebola	*	Civil engir	Liechtenst	single	*
949713	########	Sara	*	[60,70[	femme	Cardiac ar	*	Politician	Australian	single	*
949714	########	Raouf	*	[10,20[	Male	Thyroid ca	*	Professor	Japanese	married	*
949715	########	Jhazala	*	[50,60[	femme	Tuberculo	*	Hairdresse	Portugues	married	*
949716	########	Johiana	*	[40,50[	femme	Pericardia	*	Professor	Fijian	married	*
949717	########	Bachir	*	[50,60[	Male	Cirrhosis	*	Trader	Bruneian	single	*
949718	########	Nora	*	[70,80[	femme	Parkinsor	*	Laborator	Peruvian	widower	*
949719	########	Hanine	*	[60,70[	femme	HÃ©mopl	*	Writer	American	single	*
949720	########	Khadija	*	[60,70[	femme	Osteogen	*	Worker	Burkinabe	married	*
949721	########	Houda	*	[50,60[	femme	Syphilis	*	Builder	Italian	single	*
949722	########	Raouf	*	[50,60[	Male	Osteogen	*	Trader	Liechtenst	divorced	*

FIGURE 5.9 – Le nouveau “Big Data” anonymisé

5.3.5 Service *upload “Big Data”*

La Figure[5.10], présente l’interface *upload*. Le fournisseur doit spécifier le numéro de série du “Big Data” qui est généré dans la phase d’évaluation, il doit également sélectionner le contrat , le “Big Data” anonymisé et l’échantillon dans le cas ou il est disponible.Tous ces fichiers sont transférés au *HDFS*[3]. Une description détaillée de *Hadoop* sera indiqué dans la Section [5.4.1].

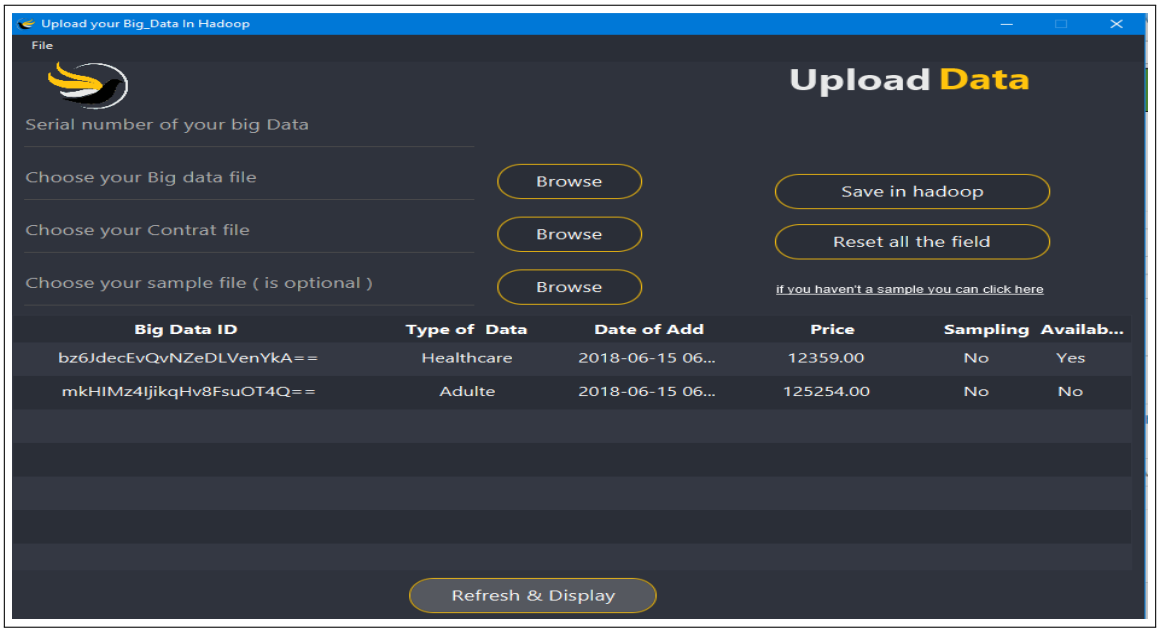


FIGURE 5.10 – Interface de transfert du fichier vers *Hadoop*

3. Hadoop Distributed File System

5.3.6 Service de communication

La Figure[5.11], montre l'interface qui permet au fournisseur de recevoir et répondre aux messages.

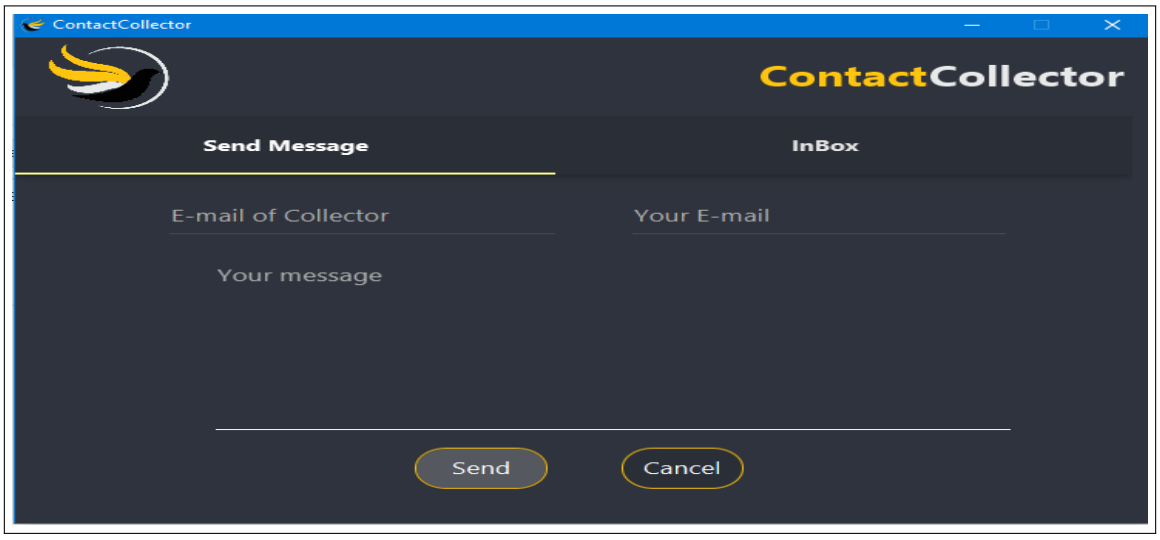


FIGURE 5.11 – Interface de communication

5.3.7 Interface principale du collectionneur

L'interface principale du client collectionneur est représentée sur la Figure [5.12]. Elle lui permet d'afficher tous les Big Data disponibles, tous les fournisseurs avec leurs *e-mails* et elle lui permet aussi d'accéder rapidement aux services fournis par *Utivacy*, ces derniers sont discutés par la suite.

FIGURE 5.12 – Interface principale du client collectionneur

5.3.8 Service de téléchargement contrat et échantillon

La Figure[5.13] montre les deux interfaces de téléchargement d'échantillon et du contrat a partir de la plateforme *Hadoop* . Le collectionneur doit indiquer le numéro de série de Big data choisi afin de l'identifier.

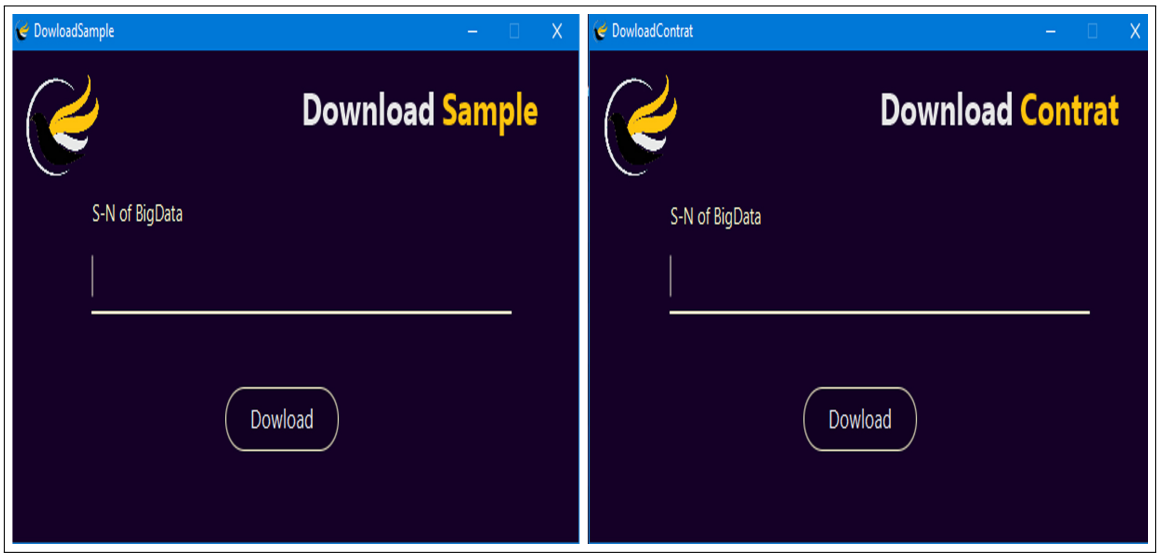


FIGURE 5.13 – Interface de téléchargement d'échantillon et du contrat

5.3.9 Service de téléchargement "Big Data"

La Figure[5.14], présente l'interface qui permet au collectionneur de télécharger le fichier Big Data, le nom de Big Data est requit. Pour l'avoir il est nécessaire de contacter le fournisseur par le service de communication *Utivacy*, Ou à travers les informations disponibles dans le contrat.

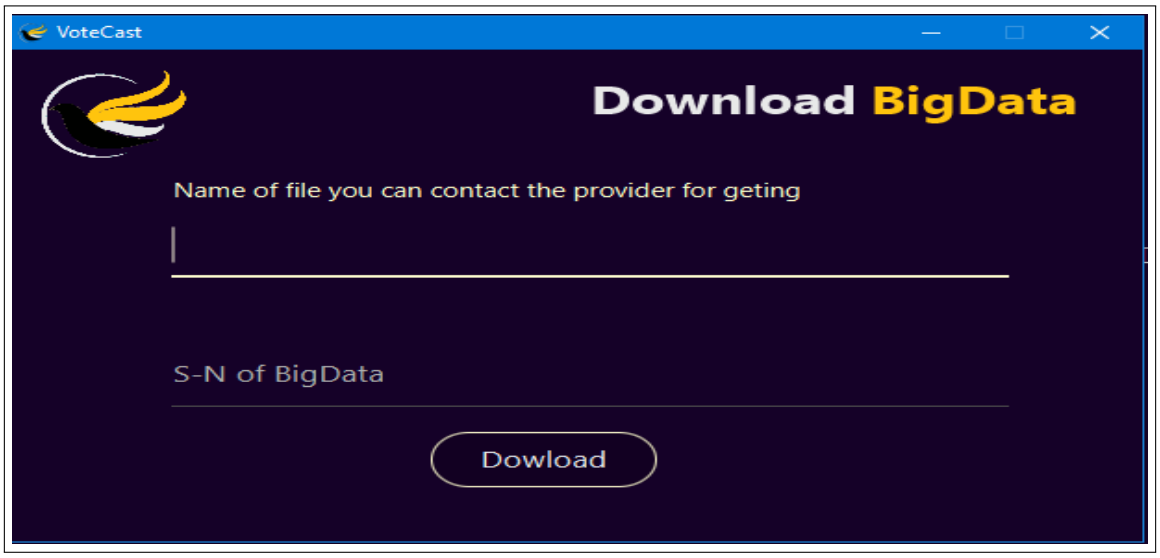


FIGURE 5.14 – Interface de téléchargement "Big Data"

5.3.10 Service de communication

Comme le montre la Figure[5.15], cette interface permet au collectionneur d'envoyer et de recevoir des messages afin de faciliter la communication avec les fournisseurs.

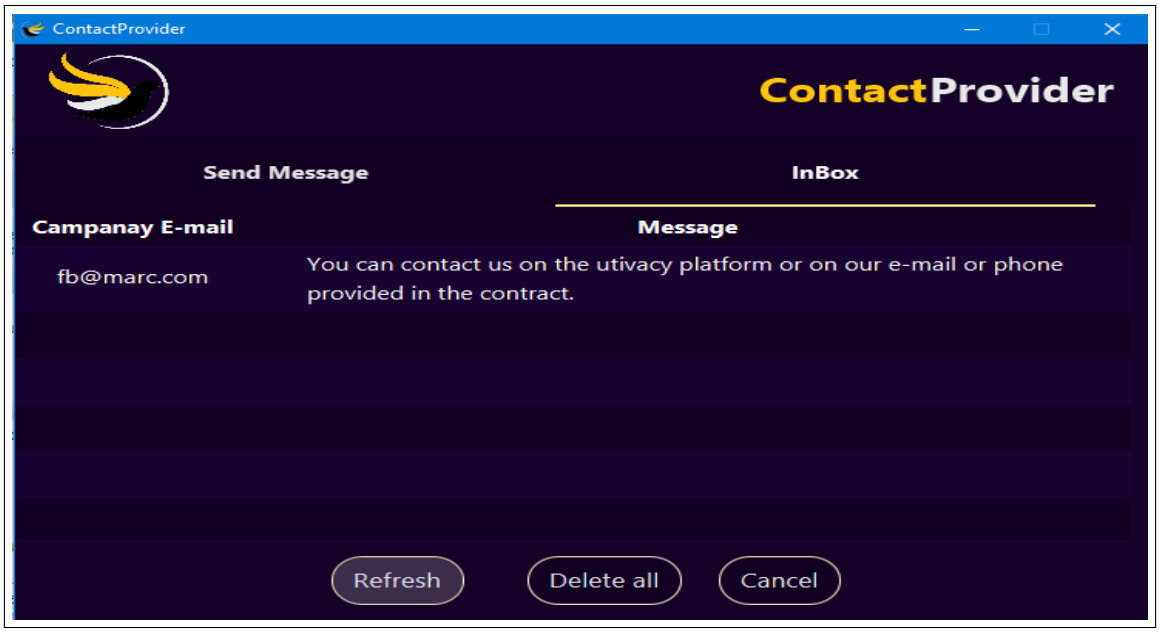


FIGURE 5.15 – Interface de communication

5.4 *Hadoop* et les principaux codes sources

5.4.1 *Hadoop*

Après l'installation, la configuration et le lancement du *Hadoop V-2.9.1*) 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 [5.16,5.17,5.18,5.19,]). Chacune d'eux à son rôle qui a déjà été expliqué dans le chapitre précédent.(voir la Section [4.3])

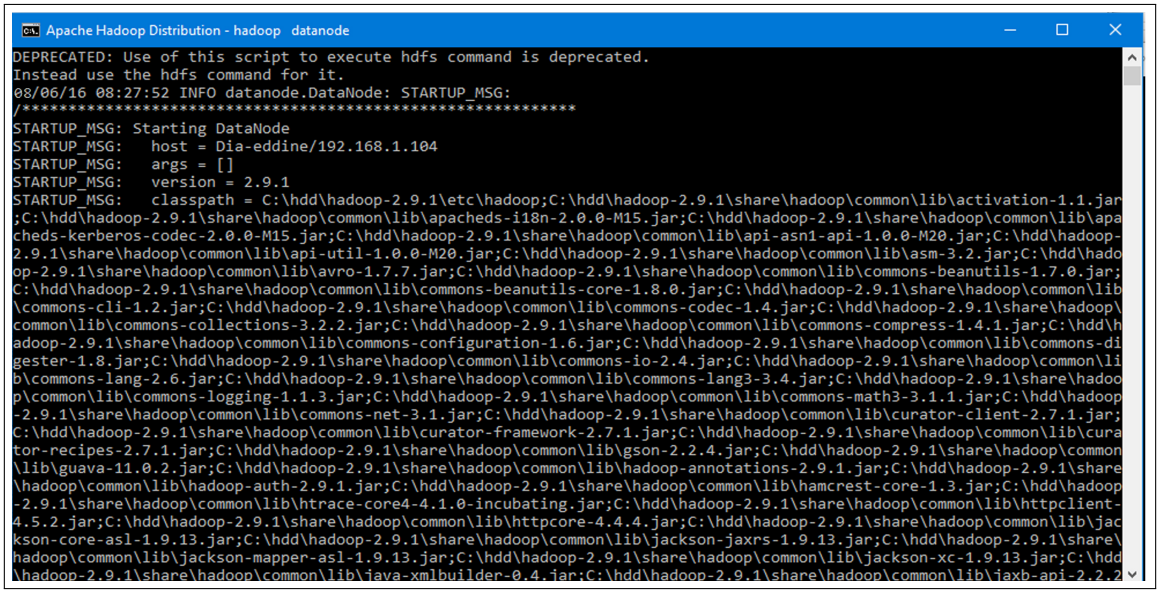


FIGURE 5.16 – Interface CMD du *DataNode*

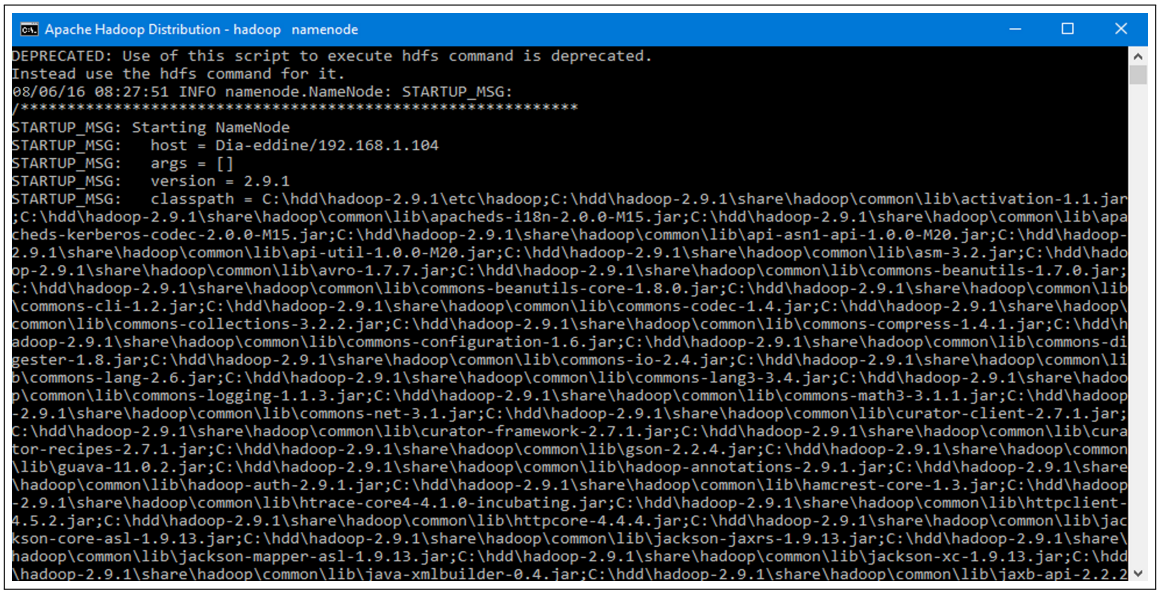


FIGURE 5.17 – Interface CMD du *NameNode*

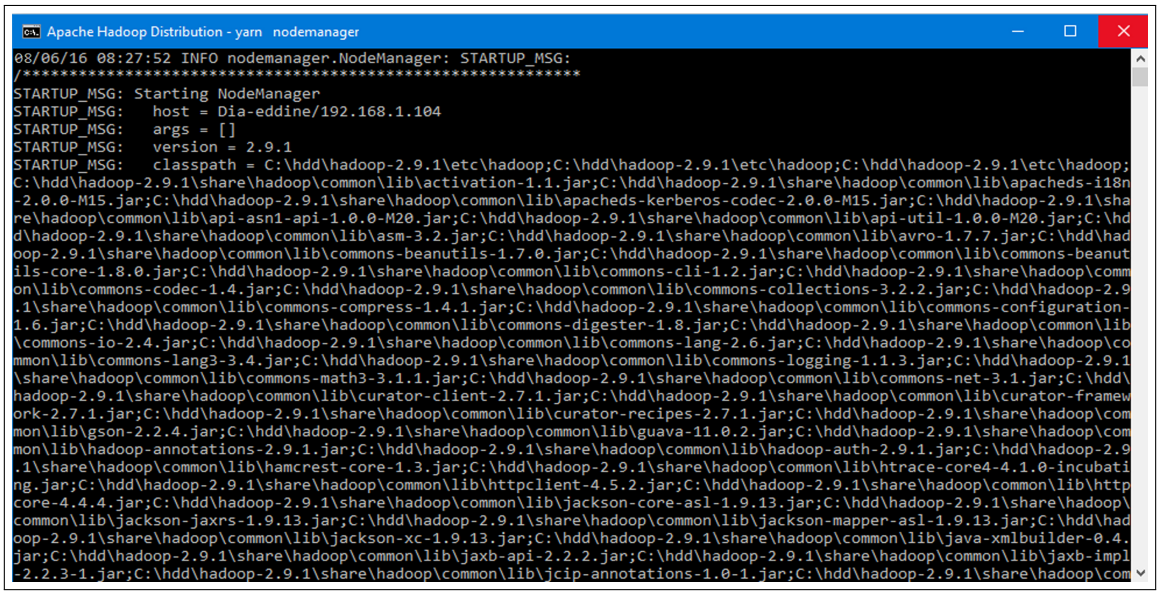


FIGURE 5.18 – Interface CMD du *NodeManager*

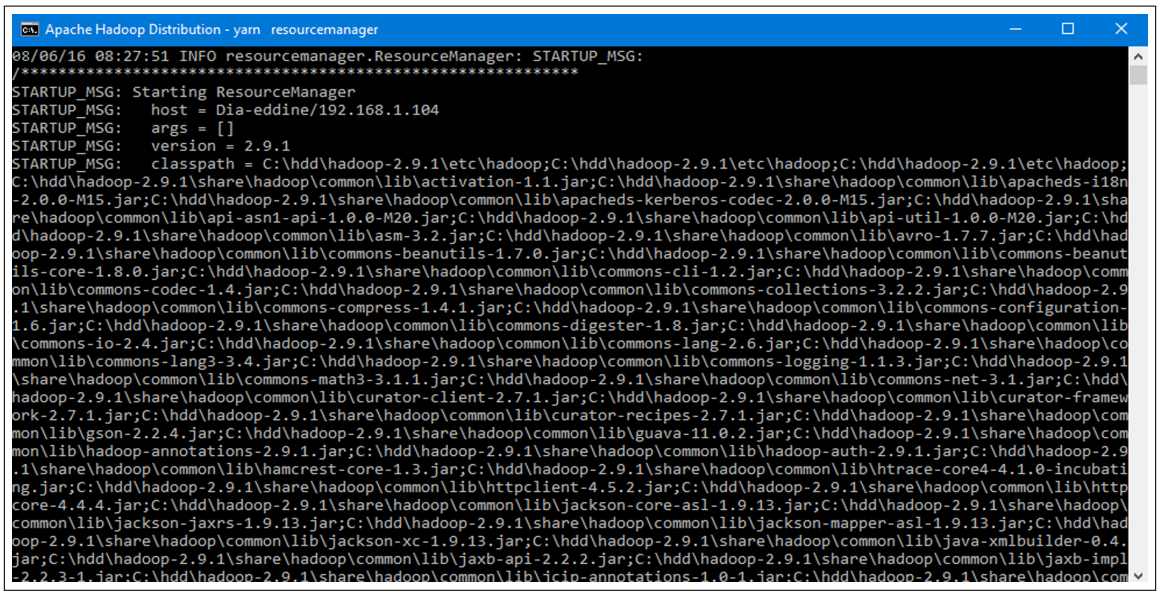


FIGURE 5.19 – Interface CMD du *RecourceManager*

L'accès à la plateforme hadoop est sur localhost : avec le numéro de port :50070 (voir la Figure[5.20]). 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[5.21] montre les fichiers transférés par le fournisseur.

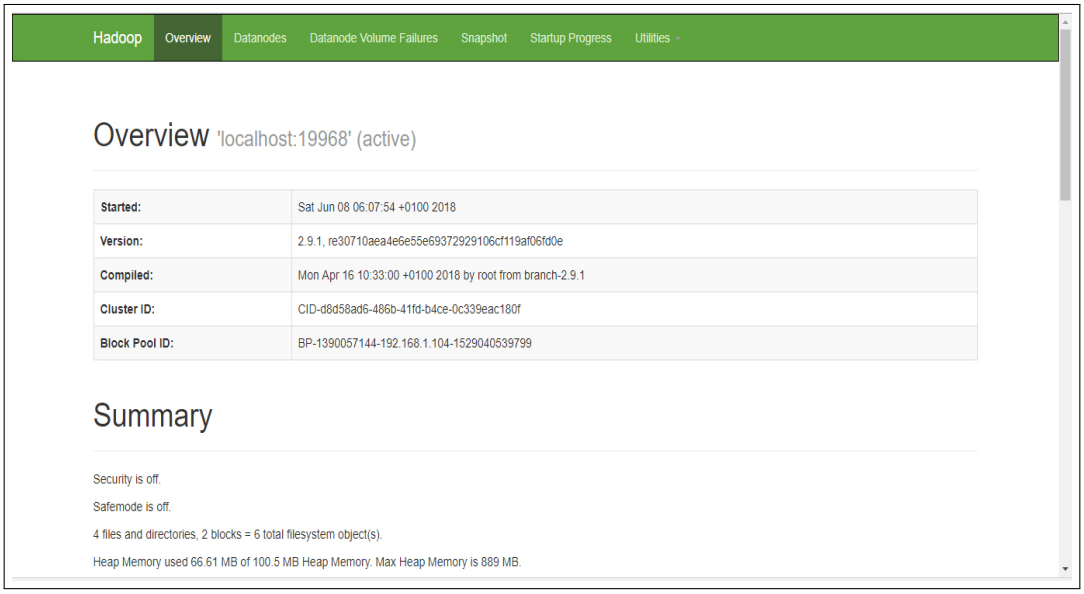


FIGURE 5.20 – Interface de la plateforme *Hadoop* '1'

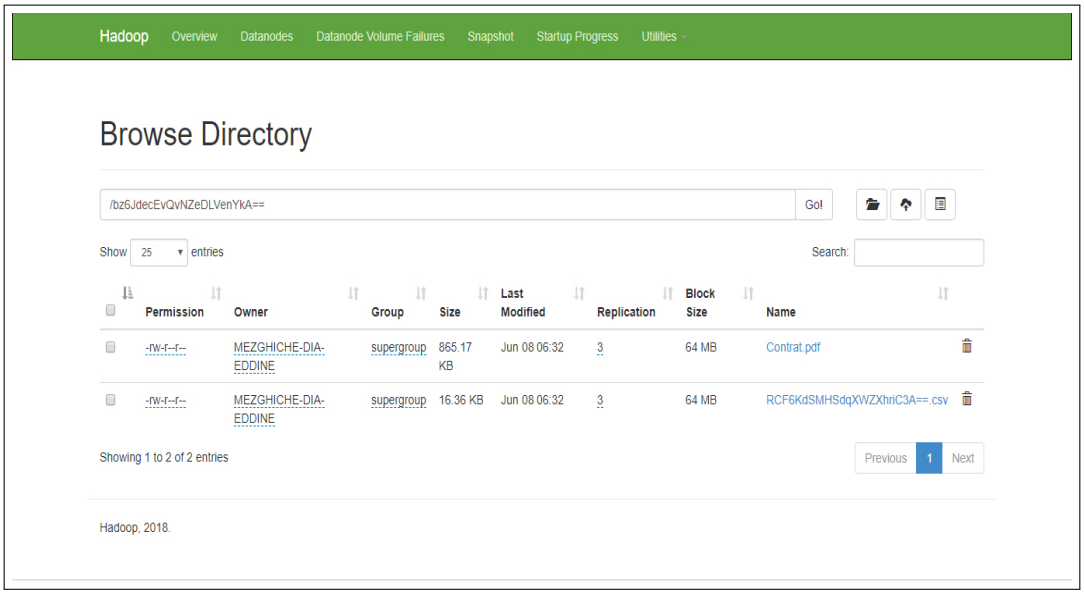


FIGURE 5.21 – Interface de la plateforme *Hadoop* '2'

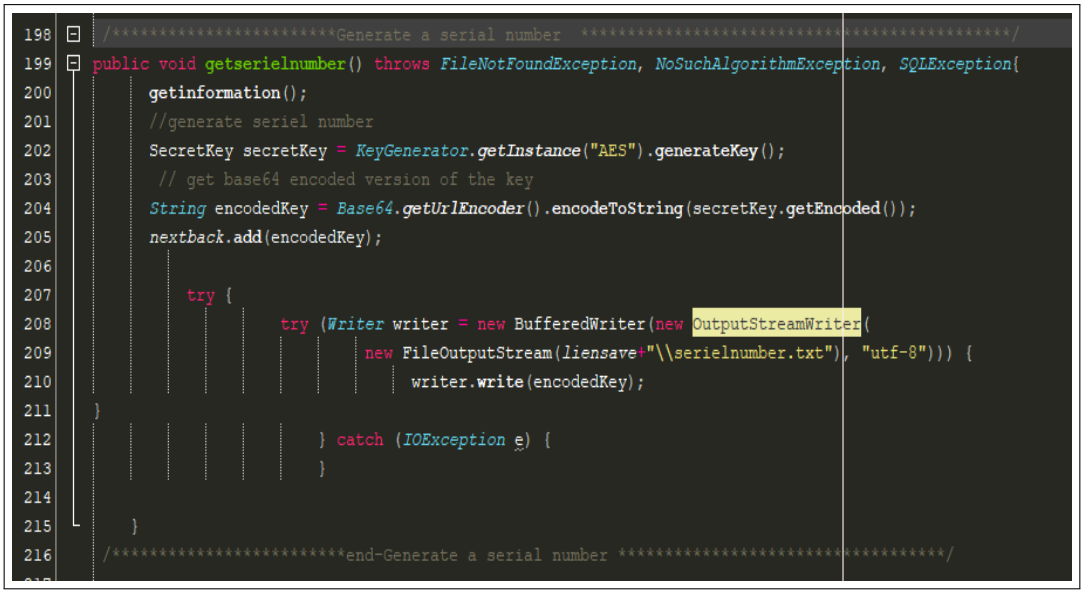
5.4.2 Les principaux codes sources

```
335
336  private  void uploaddatainhadoop() throws IOException, SQLException, SQLException{
337      String uri="hdfs://localhost:19968";
338      String distination="hdfs://localhost:19968/"+content;
339      Configuration conf= new Configuration();
340      FileSystem fs =FileSystem.get(URI.create(uri), conf);
341      fs.copyFromLocalFile(new Path(urlfile),new Path(distination));
342      System.out.println("your file  in hadoop");
343      fs.copyFromLocalFile(new Path(urlcontrat),new Path(distination));
344      System.out.println("your contrat  in hadoop ");
345      if ("".equals(urlsample.getText())) {
346          updatedatabase();
347
348     }else{
349          fs.copyFromLocalFile(new Path(urlsimple),new Path(distination));
350            updatedatabase();
351
352      }
353  }
```

FIGURE 5.22 – Code pour le transfert des fichiers du local à *HDFS*

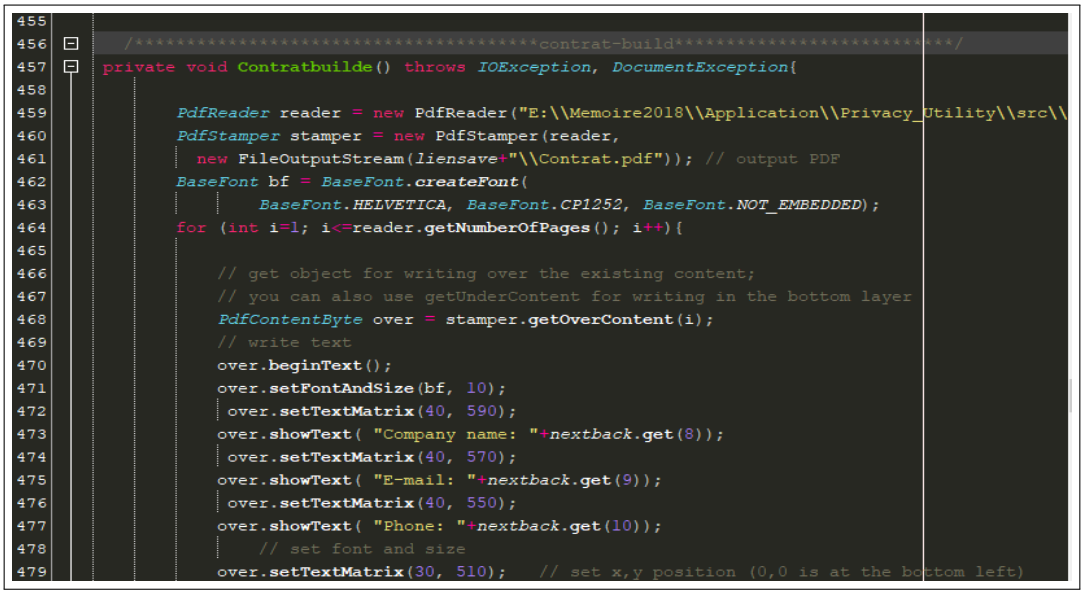
```
74
75      private void Dowloadcontrat() throws SQLException, IOException {
76          File file = dc.showDialog(null);
77        if (file != null) {
78
79        lienofsavedata=file.getPath();
80         System.out.println("lien"+lienofsavedata);
81         String localpath="hdfs://localhost:19968/";
82         String filepath="hdfs://localhost:19968/"+serialnubmring+"/"+nameofcontrat;
83          Configuration conf= new Configuration();
84          FileSystem fs =FileSystem.get(URI.create(localpath), conf);
85          fs.copyToLocalFile(new Path(filepath), new Path(lienofsavedata));
86          DowloadContat.getScene().getWindow().hide();
87
88         }else{
89          AlertMaker.showMaterialDialog(stackpanecontrat, AnchodowContrat, new
90           ArrayList<>(), "No Directory ", "please choose your Directory");
91
92          }
93      }
```

FIGURE 5.23 – Code pour le téléchargement des fichiers de *HDFS* au local



```
198  /**********************Generate a serial number  ***************************************/
199  public void getserielnumber() throws FileNotFoundException, NoSuchAlgorithmException, SQLException{
200      getinformation();
201      //generate seriel number
202      SecretKey secretKey = KeyGenerator.getInstance("AES").generateKey();
203       // get base64 encoded version of the key
204      String encodedKey = Base64.getUrlEncoder().encodeToString(secretKey.getEncoded());
205      nextback.add(encodedKey);
206
207          try {
208                  try (Writer writer = new BufferedWriter(new OutputStreamWriter(
209                          new FileOutputStream(liensave+"\\serielnumber.txt"), "utf-8"))) {
210                              writer.write(encodedKey);
211  }
212                      } catch (IOException e) {
213                      }
214
215      }
216  /************************end-Generate a serial number *********************************/
```

FIGURE 5.24 – Code de générateur du numéro de série



```
455
456    /**************************************contrat-build*************************/
457  private void Contratbuilde() throws IOException, DocumentException{
458
459          PdfReader reader = new PdfReader("E:\\Memoire2018\\Application\\Privacy_Utility\\src\\
460          PdfStamper stamper = new PdfStamper(reader,
461            new FileOutputStream(liensave+"\\Contrat.pdf")); // output PDF
462          BaseFont bf = BaseFont.createFont(
463                  BaseFont.HELVETICA, BaseFont.CP1252, BaseFont.NOT_EMBEDDED);
464          for (int i=1; i<=reader.getNumberOfPages(); i++){
465
466              // get object for writing over the existing content;
467              // you can also use getUnderContent for writing in the bottom layer
468              PdfContentByte over = stamper.getOverContent(i);
469              // write text
470              over.beginText();
471              over.setFontAndSize(bf, 10);
472               over.setTextMatrix(40, 590);
473              over.showText( "Company name: "+nextback.get(8));
474               over.setTextMatrix(40, 570);
475              over.showText( "E-mail: "+nextback.get(9));
476               over.setTextMatrix(40, 550);
477              over.showText( "Phone: "+nextback.get(10));
478                  // set font and size
479              over.setTextMatrix(30, 510);    // set x,y position (0,0 is at the bottom left)
```

FIGURE 5.25 – Code de construction du contrat

5.5 Conclusion

Dans ce dernier chapitre nous avons proposé un nouveau système pour résoudre les problèmes de la protection de la vie privée et la préservation d'utilité, 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

‘Small things are always teetering on the brink
of becoming big ones ’.
Adapted from : Mr.Max Lerner

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 utilisateur 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. Cette protection de données personnelles qui se fait par anonymisation risque d'engendrer des données fragmentées.

Les Big data se heurtent au problème suivant : comment assurer la protection des vies privées des utilisateurs mais sans nuire à l'utilité des Big data qui est de traiter des informations utiles.

Contribution

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

- ✓ Etudiant quelques approches et travaux connexes concernant la vie privée et l'utilité des données, pour pouvoir faire une étude comparative entre toutes ces approches.
- ✓ Déceler quelques inconvénients pour chaque approche.
- ✓ Nous avons proposé une nouvelle architecture afin de trouver une solution pour la vie privée et l'utilité, 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).

6.2 Perspectives

Comme perspectives, nous pouvons envisager les points suivants :

- ✓ Ajouter un composant pour l'apprentissage automatique dans le but d'aider les fournisseurs à prendre des décision dans la phase d'évaluation des Big Data d'une manière semi-automatique.
- ✓ Ajouter d'autres fonctionnalités et lancer la version commerciale.

Bibliographie

[Bertino, 2015] Bertino, E. (2015). Big data-security and privacy. *Proceedings - 2015 IEEE International Congress on Big Data, BigData Congress 2015*, pages 757–761.

[Bosworth et al., 2016] Bosworth, S., Kabay, M., and ERIC, W. (2016). *TOWARD A NEW FRAMEWORK FOR INFORMATION SECURITY*, volume 1, pages 109–131. John Wiley & Sons, 6 edition.

[Caballero et al., 2014] Caballero, I., Serrano, M., and Piattini, M. (2014). A data quality in use model for big data. *Advances in Conceptual Modeling*, pages 65–74.

[Chen et al., 2014] Chen, M., Mao, S., and Liu, Y. (2014). Big data : A survey. *Mobile Networks and Applications*, 19 :171–209.

[Commission, 2012] Commission, T. F. F. B. D. (2012). A practical guide to transforming the business of government. *TechAmerica*, pages 1–40.

[Damiani, 2015] Damiani, E. (2015). Toward big data risk analysis. In *Big Data (Big Data), 2015 IEEE International Conference on*, pages 1905–1909. IEEE.

[Demchenko et al., 2014] Demchenko, Y., Ngo, C., de Laat, C., Membrey, P., and Gordijenko, D. (2014). Big security for big data : Addressing security challenges for the big data infrastructure. In *Workshop on Secure Data Management*, volume 8425, pages 76–94. Springer.

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

[Dong et al., 2011] Dong, C., Russello, G., and Dulay, N. (2011). Shared and searchable encrypted data for untrusted servers. *Journal of Computer Security*, 19(3) :367–397.

[Eignier, 1997] Eignier, B. B. (1997). Vie privée et vie publique. 41 :163–180.

[Finn et al., 2013] Finn, R. L., Wright, D., and Friedewald, M. (2013). Seven types of privacy. In *European data protection : coming of age*, pages 3–32. Springer.

[Francis, 2014] Francis, L. P. (2014). Introduction : Technology and new challenges for privacy. *Journal of Social Philosophy*, 45(3) :291–303.

[Friedewald et al., 2014] Friedewald, M., van Lieshout, M., Rung, S., Ooms, M., and Ypma, J. (2014). Privacy and security perceptions of european citizens : A test of the trade-off model. In *IFIP International Summer School on Privacy and Identity Management*, pages 39–53. Springer.

[Gaddam, 2015] Gaddam, A. (2015). Securing your big data environment. *Black Hat USA 2015*.

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

[Gkiotsalitis and Stathopoulos, 2015] Gkiotsalitis, K. and Stathopoulos, A. (2015). A utility-maximization model for retrieving users' willingness to travel for participating in activities from big-data. *Transportation Research Part C : Emerging Technologies*, 58 :265–277.

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

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

[Information Commissioner's Office, 2014] Information Commissioner's Office (2014). Big Data and Data Protection. pages 1 – 51.

[Li et al., 2013] Li, N., Qardaji, W., Su, D., Wu, Y., and Yang, W. (2013). Membership privacy : a unifying framework for privacy definitions. In *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*, pages 889–900. ACM.

[Li et al., 2016] Li, S., Gao, Jerry/S, Y., and S, G. (2016). *Big Data Concepts, Theories, and Applications,*, chapter Security and Privacy for Big Data. Springer International.

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

[Mehmood et al., 2016] Mehmood, A., Natgunanathan, I., Xiang, Y., Hua, G., and Guo, S. (2016). Protection of big data privacy. *IEEE access*, 4 :1821–1834.

[Monreale et al., 2014] Monreale, A., Rinzivillo, S., Pratesi, F., Giannotti, F., and Pedreschi, D. (2014). Privacy-by-design in big data analytics and social mining. *EPJ Data Science*, 3(1) :10.

[Nagendrakumar et al., 2014] Nagendrakumar, S., Aparna, R., and Ramesh, S. (2014). A non-grouping anonymity model for preserving privacy in health data publishing. In *Science Engineering and Management Research (ICSEMR), 2014 International Conference on*, pages 1–6. IEEE.

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

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

[Rodríguez-Mazahua et al., 2016] Rodríguez-Mazahua, L., Rodríguez-Enríquez, C. A., Sánchez-Cervantes, J. L., Cervantes, J., García-Alcaraz, J. L., and Alor-Hernández, G. (2016). A general perspective of Big Data : applications, tools, challenges and trends. *Journal of Supercomputing*, 72(8) :3073–3113.

[Sankar et al., 2013] Sankar, L., Rajagopalan, S. R., and Poor, H. V. (2013). Utility-privacy tradeoffs in databases : An information-theoretic approach. *IEEE Transactions on Information Forensics and Security*, 8(6) :838–852.

[Sanyal et al., 2016] Sanyal, M. K., Bhadra, S. K., Das, S. C. S., Mandal, J. K., and Siba K. Udgata, V. B. (2016). *Information Systems Design and Intelligent Applications*, volume 1, chapter A Conceptual Framework for Big Data Implementation to Handle Large Volume of Complex Data, pages 455–465. Springer India.

[Saouli et al., 2016] Saouli, H., Kazar, O., and Kassimi, D. (2016). Applications et enjeux des big data dans le contexte des défis mondiaux. *journal*.

[Sivarajah et al., 2017] Sivarajah, U., Kamal, M. M., Irani, Z., and Weerakkody, V. (2017). Critical analysis of big data challenges and analytical methods. *Journal of Business Research*, 70 :263–286.

[Solove, 2008] Solove, D. J. (2008). I've got nothing to hide and other misunderstandings of privacy. *San Diego Law Rev.*, 44 :745–772.

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

[Tsegaye and Flowerday, 2014] Tsegaye, T. and Flowerday, S. (2014). Controls for protecting critical information infrastructure from cyberattacks. In *Internet Security (WorldCIS), 2014 World Congress on*, pages 24–29. IEEE.

[Villasenor et al., 2014] Villasenor, E., Pritchett, T., Dyaberi, J. M., Pai, V. S., and Thottethodi, M. (2014). Morphstore : A local file system for big data with utility-driven replication and load-adaptive access scheduling. In *Mass Storage Systems and Technologies (MSST), 2014 30th Symposium on*, pages 1–10. IEEE.

[Wang et al., 2010] Wang, K., Chen, R., Fung, B., and Yu, P. (2010). Privacy-preserving data publishing : A survey on recent developments. *ACM Computing Surveys*, pages 1–53.

[Xu et al., 2015] Xu, L., Jiang, C., Chen, Y., Ren, Y., and Liu, K. R. (2015). Privacy or utility in data collection ? a contract theoretic approach. *IEEE Journal of Selected Topics in Signal Processing*, 9(7) :1256–1269.

[Xu et al., 2014] Xu, L., Jiang, C., Wang, J., Yuan, J., and Ren, Y. (2014). Information security in big data : privacy and data mining. *IEEE Access*, 2 :1149–1176.

[Zhang et al., 2014] Zhang, X., Liu, C., Nepal, S., Yang, C., and Chen, J. (2014). Privacy preservation over big data in cloud systems. In *Security, Privacy and Trust in Cloud Systems*, pages 239–257. Springer.

Annexe A

Contrat

Company name: facebook

E-mail: fb@marc.com

Phone: +00077777777

the type of this big data is Healthcare ,their level of sensitivity is 40 % ,and their degree of anonymity is 25.0% ,their price is 12359.00 $.

Before you start using this big data, it is important that you understand that you need to consider the content of these big data and the level of their sensitivity to put the users in a high level of privacy protection.

The mission of the utivacy group, is to constantly improve the level of protection of user privacy and provide you with a high level of utility.

Company name (collector) : .....

Date : .....

Signature

.

+213-000-000-000
utivacy@utivacy.com
university of biskra-Algeria

This Big data are proved by Utivacy groupe 2017/2018

Erratum

- Concernant les chapitres [4 et 5] : On s'excuse du non netteté de certaines figures.
- On remercie d'avance toute personne qui nous signalera, les erreurs qu'il pourrait déceler, à l'adresse suivante : dia.mezghiche@gmail.com