République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la

Recherche Scientifique

Université Abou Bekr Belkaid – Tlemcen

Faculté de Technologie

Département de Génie Electrique et Electronique

Filière du Génie Industriel

Projet de fin d'études

Intitulé :

Proposition d'une nouvelle architecture pour la préservation de la vie privée d'utilisateur dans le Big data

Présenté par :

- BEN DJEDDOU Smail
- CHAOUATI Rafik

Devant les jurys :

- Président : HADRI Abdelkader	MAA	UABB	Tlemcen
- Examinateur : HASSAM Ahmed	MCB	UABB	Tlemcen
- Examinateur : BESSENOUCI Hakim Nadhir	MAA	UABB	Tlemcen
- Encadreur : GUEZZEN Amine	MCB	UABB	Tlemcen

Année Universitaire: 2018/2019

REMERCIMENT

Avant tout, nous remercions ALLAH, le tout puissant qui nous a donné le courage, la volonté et la patience pour bien mener ce travail.

Nous avons l'honneur et le plaisir de présenter notre profonde gratitude et nos sincères remerciements à notre encadreur Dr. GUEZZAN Amine, pour ses précieuses aides, ces orientations et le temps qu'il m'a accordé pour notre encadrement.

Nous remercions aussi tous les enseignement de la filières Génie industriel qu'il nous ont encouragé et donner la force pendant la durée de notre étude.

Également nous remercions les membres de jury Dr HASSAM Ahmed , Dr BESSENOUSSI Nadhir et Dr HADRI Abdelkader qui ont accepté de juger notre travail.

Tous nos collègue du promotion génie industriel et les autre promotion de l'université de Tlemcen, et tous les personne qui nous donné l'aide pour compléter notre mémoire.

Dédicace,

A mes chers parents ***Madani*** *et* ***Zouina*** *, pour tous leurs sacrifices, leur amour, leur tendresse, leur soutien et leurs prières tout au long de mes études,*

A mes chères sœurs ***Samira, Dalila, Ouidad*** *et* ***Bahia*** *pour leurs encouragements permanents, et leur soutien moral,*

A mes chers frères, ***Fateh, Amine*** *et* ***Abdelhalim****, pour leur appui et leur encouragement,*

A toute ma famille pour leur soutien tout au long de mon parcours universitaire,

Que ce travail soit l'accomplissement de vos vœux tant allégués, et le fuit de votre soutien infaillible,

Merci d'être toujours là pour moi.

Rafik

Dédicace

Je dédie ce modeste travail :

À ma très cher mère ***Naziha*** *et à vous mon respectueux père* ***Laouni*** *pour qui n'ont jamais cessé de formuler des prières de me soutenir et de m'épauler pour que je puisse à atteindre mes objectif.*

À ma grand-mère qui je lui souhaite une bonne santé.

À mon unique chère sœur ***Nadjet*** *et son mari* ***Saifi,*** *à mes chère frères* ***Mohammed, Walid, Omar*** *et* ***Karim,*** *pour ses soutien morale et leur conseil tout au long mes études.*

À toute la famille et mes proches amis qui ont ma loyauté après le Dine.

Merci à tous d'être toujours à mes côtés.

Smail

Résumé: L'utilisation du Big Data a augmenté ces dernières années, ce qui a généré une grande quantité d'informations ,ses données peuvent produit dans l'internet et aussi à l'interne des entreprises, le but du big data est de corréler ses données entre elles en temps réels pour prendre des décision afin de gagner un profit ou un avantage concurrentielles. L'analyse de ses information ouvre la porte pour la question de confidentialité et la vie privé des utilisateurs et comment protéger ses données produisent chaque seconde.

Ce travail consiste à protéger la vie privé privée de l'utilisateur contre la violation et le non-respect de la confidentialité, pour cela nous avons proposé une nouvelle architecture qui garantit la protection des donnée et la vie privée afin de montrer l'efficacité de l'architecture. Nous avons développé un modèle qui pourra résoudre les problèmes précédentes.

Mots clés : Big data, Confidentialité , la protection de la vie privée, utilité des données.

Abstract : The use of Big Data has increased in recent years, which has generated a large amount of information ,its data can produce in the internet and also in-house companies, the purpose of big data is to correlate its data in real time to make decisions in order to gain a competitive profit or advantage. The analysis of its information opens the door to the issue of confidentiality and privacy of users and how to protect its data produce every second.

This work consists of protecting the user's privacy from breach and breach of confidentiality, for this we have proposed a new architecture that guarantees data protection and privacy in order to show the effectiveness of the architecture. We have developed a model that will solve the previous problems.

Keywords : Big data, Privacy, Confidentiality, data utility.

ملخص: لقد ازداد استخدام البيانات الضخمة هذه السنوات الأخيرة، وما أدى إلى كمية كبيرة من المعلومات، ويمكن أن تنتج هته البيانات على الإنترنت وأيضا في داخل الشركات، والهدف من البيانات الضخمة ربط المعلومات بينها في الوقت الحقيقي لاتخاذ قرارات لكسب ربح أو ميزة تنافسية. ويفتح تحليل المعلومات الباب أمام مسألة السرية، وكيف تنتج الحياة الخاصة لكل مستخدم، وكيف تحمي البيانات الخاصة به.

هذا العمل يتمثل في حماية الحياة الخاصة للمستخدم من الانتهاك وعدم احترام السرية، لقد اقترحنا هندسة جديدة تضمن حماية البيانات والحياة الخاصة، ولإظهار مدى نجاعة هته الهندسة، قمنا بتطوير نموذج من أجل حل المشاكل المذكورة سابقا .

الكلمات المفتاحية: حماية الخصوصية،البيانات الضخمة،البيانات المهمة.

SOMMAIRE

Table des Figures :

Liste des tableaux

Introduction Général

Nous vivons à l'ère de la technologie et de l'information, Avec l'avènement des médias sociaux les smartphones et les objets connectés (Internet of Things), la quantité des données produites chaque année augment d'une manière impressionnante, tout cela a conduit à l'émergence d'une nouvelle révolution technologique pour cela il vient le mot Big Data c'est-à-dire grand quantité des donné ou bien ''Méga-donnée'', ses données sont utilisé dans plusieurs domain de notre vie, parmi eux la médecine, le commerce, le marketing, la politique.. etc. afin de facilité notre vie et/ou pour gagner un profit dans le cas des organisation et des gouvernement. Donc l'enjeux de Big data et de savoir collecter les données , l'analysé et les visualiser sont perturbation et en temps réel .

Aujourd'hui le Big data présent un outil très important pour la prise de décision pour les entreprises moderne et il a aidé de comprendre les besoins du client, pour cela il est nécessaire pour la création de la valeur et acquisition d'un avantage compétitif et une long vie pour les entreprises .

Dans les derniers trois ans le Big data est devenu un sujet d'actualité grâce à l'explosion de l'utilisation de l'internet, les smartphones et le succès des réseaux sociaux, Cela a entraîné une augmentation des vols et du piratage via Internet, de ce fait, la préservation de la vie privé de l'utilisateur est devenu une nécessité et comment trouver un compromis entre cette dernière et l'utilité du Big data.

Afin de sécuriser la vie privée des individus, il est nécessaire que les données soit bien anonymisées de manière évidente pour satisfaire les individus et pour ne pas perdre ses valeurs.

Notre objective et de :

- faire une étude général sur les Big data et les notions de la vie privée.
- présenter un état de l'art sur les approches et travaux connexes.
- proposer une nouvelle architecture pour la préservation de l'utilité des Big data.
- appliquer notre architecture proposée.
- conclusion et perspective.

Notre mémoire commencé par une introduction générale suit d'un chapitre consacré pour étudier les fondements et les généralités sur le Big data. Dans le deuxième chapitre, nous avons présenté un état de l'art et les approches et des travaux connexes. Dans le troisième chapitre, nous avons proposé notre architecture pour la préservation de la vie privée des utilisateurs, dans le quatrième chapitre on a appliqué et implémenter notre architecture proposée, et on a terminé par une conclusion générale et quelques perspectives .

Chapitre I :
Généralités sur le Big data

I. 1 INTRODUCTION :

Dans ce chapitre, nous allons présenter un aperçu général sur le Big data, ses concepts, enjeux et domaines d'applications, on abordera ensuite la notion de la vie privée, ses types, ses défis ainsi que quelques terminologies de la vie privée. Ensuite, nous introduirons les techniques de protection de la vie privée dans les Big data et l'opportunité des Big data pour la protection de la vie privée. Enfin, nous nous intéresserons aussi au cloud computing dans le Big data en citant certaines technologies de cryptage qui préservent la confidentialité des données et de leurs propriétaires.

I. 2 BIGDATA :

1. Emergence de BigData :

Le BigData désigne le courant technologique dominant et le dernier cri en matière de High Tech que nous voyons émerger ces dernières années. Ce phénomène qui représente une révolution de la technologie de l'information et qui affecte les organisations de presque tous les secteurs, a une grande influence sur le présent et l'avenir vu son importance.

Gorden Moore a déduit que la puissance des ordinateurs allait croitre de manière exponentielle durant des années car la quantité d'information traitée par microprocesseur allée doubler chaque 18 mois (Loi de Moore), il est à noter aussi que la capacité de stockage de données a évoluée plus rapidement que la capacité de traitement de données (Loi de Kryder).

Les données augmentent de manière exponentielle en partie à cause de la loi de moore mais aussi suite à l'accroissement de la population ce qui nous pousse à penser aux réseaux sociaux et à leurs exigences. Ce qui rend difficiles la capture, l'intégration, le traitement et l'exposition des données.

Il est donc nécessaire de stocker cette énorme quantité de données numérique actuellement en circulation, les analyser, traiter et exploiter correctement pour répondre aux besoins des différents opérateurs industriels, économiques ou sociaux.

Le Big data qui porte sur porte sur la recherche, la capture, le stockage, le partage et la présentation de ces données : est une technologie révolutionnaire qui peut provoquer des perturbations aux niveaux économique, scientifique et culturelle. Cela est dû à l'importance des changements et des améliorations qu'il impose dans ces différents domaines et qui exigent une nouvelle réadaptation (Stubbs, 2014)

2. Définition de Big data

Le concept "Big data" ou "données massives" désigne un ensemble volumineux de données pouvant être structurées ou non structurées, leur manipulation dépasse les outils classiques de traitement et d'analyse.

On définit la problématique du big data par les trois V:

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

Le big data nécessite des approches innovantes et rentables de traitement de l'information pour une meilleure prise de décision. (Reinsel, 2011)

3. Modèle 5V :

Le modèle 5V comporte les 5 caractéristiques du big data comme illustré dans la figure 1.1. Il représente une extension du modèle 3V.

- **Volume**: Est d'une masse énorme de données de l'ordre de zettabytes, et test en croissance exponentiel.
- **Vitesse**: les données doivent être introduites et analysés le plus rapidement possible – traitement en temps réel- afin de maximiser le gain de production.
- **Variété** : Est caractérisé par la diversité des formats et type de données, comprend des données semi-structurées et non structuré comme l'audio, vidéos et textes. Et des données structurées comme chiffres, signes et mots.
- **Valeur**: concerne le degré d'importance attribué à une donnée, il est primordial d'analyser les données en prenant en compte ce critère dans le but de prendre des décisions adéquates.
- **Véracité**: il existe des techniques qui garantissent l'exactitude et fiabilité des données du big data afin d'éviter toute erreur ou modification non autorisé produite sur la donnée (Lisbeth.R, 2015)

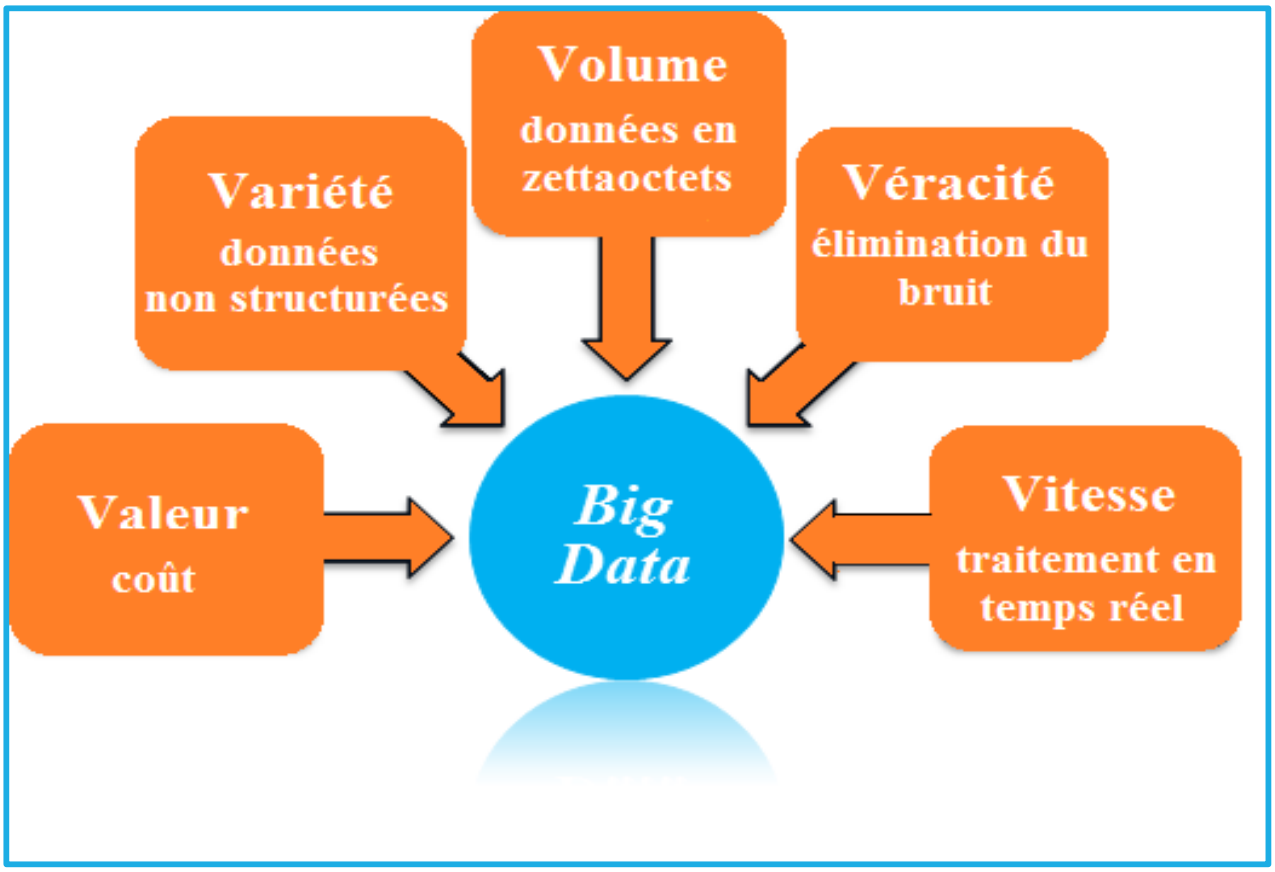

Figure 1-1: Modèle 5V

4. Concept de BigData :

Dans cette partie nous allons examiner attentivement les concepts techniques communs et modèles généralement utilisés dans la plupart des outils et plateformes BigData.

4.1 Cluster de BigData :

Un cluster est une grappe de serveurs (groupe) et d'autres ressources indépendantes fonctionnant comme un seul et même system, Les concepts clusters de BigData peuvent être divisés en deux catégories :

- Configuration et topologie du cluster :

On parlera ici du modèle logique du cluster, et des différents types de nœuds le constituant. Un cluster BigData est logiquement constitué de deux types de nœuds (machines) :

- Nœuds de données : Ces nœuds ont pour objectif d'une part le stockage des données d'une manière distribuée et d'autre part le traitement secondaire pour la transformation et l'accès.
- Nœuds de gestion : Sont une façade pour les applications client et exécute les différents cas d'utilisations.

- Déploiement des clusters :Traite le déploiement réel du cluster dans l'infrastructure physique.

4.2 Stockage des données en BigData :

Le stockage d'immense quantité de données est ce qui est primordiale dans n'importe quelle plateforme BigData. La quantité énorme de données doit être mémorisé d'une manière appropriée en BigData afin d'être exploiter et analyser d'une manière efficace.
Le concept de stockage se résume comme suit :

- Modèles de données :

Il existe plusieurs modèles de données, nous citons le modèle relationnel « NoSQL ».

- Partitionnement de données :

Dans cette étape les données seront partitionnées sur plusieurs nœuds de données du cluster dans le but de traiter toutes ces données simultanément par les machines.

- La réplication de données :

Signifie la protection des données dans le but d'assurer la continuité de service en cas de défaillance d'un des serveurs.

- La compression de données :

La compression permettre de réduire la taille physique de blocs d'information et de limiter la taille de la bande passante nécessaire dans le réseau de transport. Un décompresseur est donc nécessaire afin de reconstruire les données originelles. Cependant, cette étape présente un inconvénient majeur en temps de traitement suite à la lecture/écriture des données à partir/vers l'espace de stockage.

- Indexation de données :

L'indexation permet de déterminer la position des différents blocs positionnés dans les nœuds de données du cluster de BigData.

4.3 Gestion de ressource :

Les différentes ressources informatiques (RAM, CPU et Disque) des nœuds constituant le cluster BigData doivent être utilisées et partagées de manière appropriée.

5. Résistance de Big data (Souplesse et maniabilité) :

Le séisme et le tsunami du 11 mars qui ont ravagé le Japon ont induit à un nombre important de morts et disparus, suivi de l'accident de la centrale nucléaire à Fukushima et encore plus d'évènements qui ont démontré la nécessité de la technologie big data et le rôle qu'elle apporte sur la protection des données et la prévision des centres météorologiques et sismiques aux événements les plus dangereux. Et ainsi réduit d'une façon remarquable le nombre de victimes et dégâts. De nouvelles techniques permettant la préservation de données même en cas de catastrophe naturelles de grandes ampleurs sont apparues de nos jours, telle que la technologie Cloud (Hayashi, 2013)

6. Architecture lambda :

L'architecture lambda est une approche hybride dans la gestion des données massives en BigData, elle traite efficacement les données volumineuses et réponds aux tolérances aux pannes humaines et aux différents dommages causés par la tolérance aux pannes logicielles ou matérielles. Cette architecture comporte trois différentes couches :

- Couche batch (Batch layer) : Toutes donnée entrante est ajoutée à cette couche, les données présentes ici, sont immuables.
- Couche temps réels (Speed layer) : Cette couche traite uniquement les données récentes.
- Couche de service (Serving layer) : Cette couche analyse les données les plus récentes.

Afin de bien comprendre le fonctionnement de l'architecture Lambda, nous allons détailler le comportement de chaque couche ci-après :

6.A Couche batch (Batch layer) :

Cette couche est composée de deux grands composants très importants :

- Master Dataset : ou on va stocker les données massives récoltés

- Calculs distribués : qui va effectuer des calculs sur les données massives.

Le master Dataset est responsables du stockage de grandes quantités de données reçues, ces données sont toujours correcte et ne seront jamais supprimer ou modifier. Ainsi même si notre programme qui réalise des calculs et traitement de données comporte des erreurs, il sera toujours possible (à partir du master Dataset) de rectifier l'erreur suite aux données brutes sauvegardées dans le Dataset.

6.B Couche de service (Serving layer) :

Après les différents traitements par lot faits dans la première étape (Batch layer), les résultats seront transférer au Serving layer pour qu'il les stocks et expose aux clients lorsque ces derniers réalisent des requêtes.

6.C Couche temps réels (Speed layer) :

Le Speed layer est un composant qui va faire l'analyse des données les plus récentes, c'est-à-dire celles qui n'ont pas été encore ajouté à la Batch layer. Ce composant va permettre de faire un traitement rapide sur les données les plus récentes et d'exposer une vue pour que les clients puissent faire des requêtes sur les données les plus fraiches (celles qui n'ont pas encore été ajoutées au Serving layer). La couche temps réels ne contiendra non seulement pas des données immuables et source perpétuelle de vérité mais aussi, ne fournira pas un moteur de stockage permanent contrairement au Dataset.

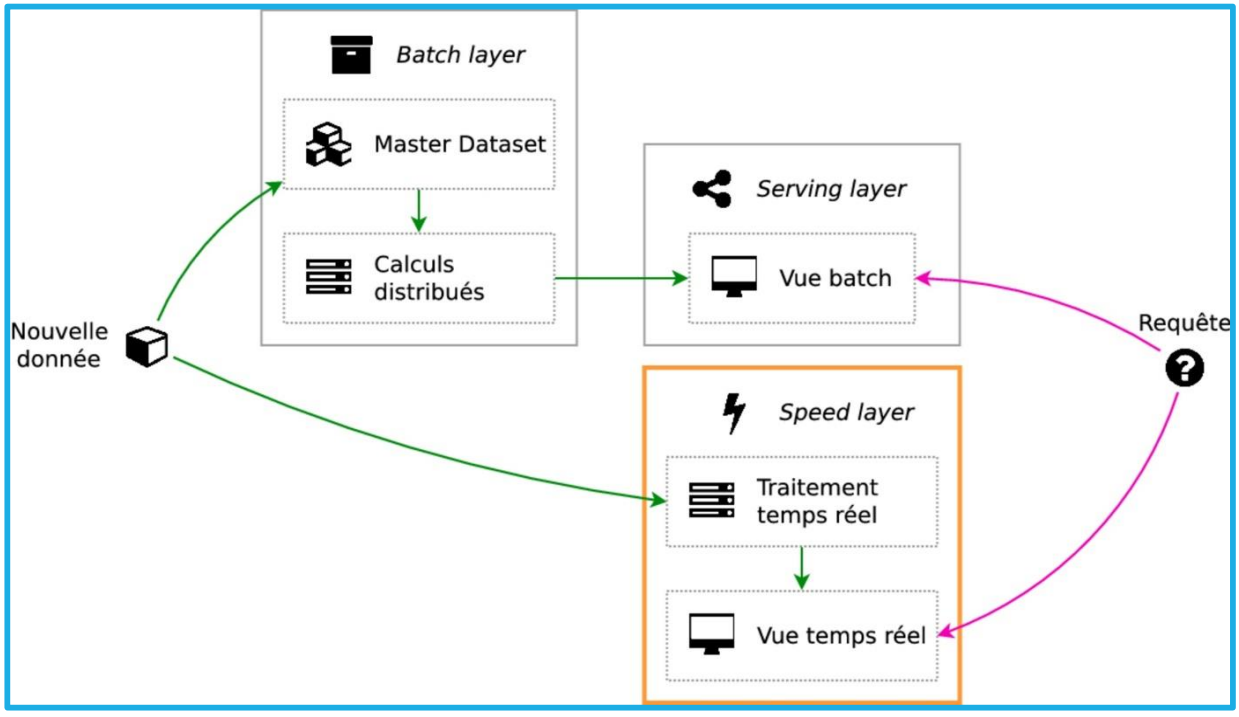

Figure 1-2: Figure illustrant l'architecture lambda (Behmo, 2019)

7. Domaine d'application de Big data :

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

7.A Agriculture :

La population mondiale ne cesse de se développer ce qui implique que le domaine de l'agriculture doit être parfaitement géré afin de garantir une alimentation saine et suffisante. Le big data représente aujourd'hui la meilleure façon d'atteindre ce but et de pouvoir organiser de gigantesque masses de données sur les prédictions météorologique et la sécheresse du sol.

7.B Assurance :

Le big data est utilisé dans le domaine de l'assurance afin d'effectuer des statistiques et analyses sur le comportement des assurés.

La récolte des informations additionnelles permet un meilleur profilage des assurés: hygiène, amende, conduite de voiture, relation professionnelle… Etc. Ainsi, les agences d'assurances garantissent un meilleur contrôle des coûts et une amélioration du service offert.

7.C Marketing **:**

Le domaine du marketing a toujours été ouvert à l'avancement technologique et informatique, ceci est dû au besoin d'un maximum d'information sur le consommateur, ceci en utilisant : les réseaux sociaux, applications mobiles, TV, catalogues, blog, presse, radios, etc. Ce qui permet de rendre les solutions marketing plus adéquates.

La révolution marketing est assurée aujourd'hui grâce à l'omniprésence des capteurs qui engendre la récolte d'une diversité de données : images de visages pour analyse émotionnelle, vidéos pour description comportementale, données textuelles, numérique ou statistique. La gestion de ces données n'est possible qu'avec les méthodes de traitement et stockage issues du big data.

7.D Achat programmatique :

L'achat programmatique facilite l'opération d'achat/vente traditionnelle, exploitable via un logiciel ou une plateforme intermédiaire entre le client et le fournisseur. Cette technique exige la gestion en temps réel de gigantesque masses d'informations échangées issus du big data qui représente un atout pour une meilleur rentabilité des plateformes achat/vente.

7.E Compétitivité et Innovation de produit :

La récolte et analyse en temps réel des données massives issus du big data a pour but pour les entreprises d'améliorer le contrôle de la production et de pouvoir la comparer avec celles des concourants ce qui assure l'innovation et la compétitivité des produits.

7.F Gestion de catastrophes naturelles :

La meilleure façon de gérer les catastrophes naturelles est la possibilité d'analyser les données météorologiques ce qui assure une prédiction des ouragans et sur quel endroit géographique vont-ils atteindre. Ainsi, en cas de dangers, des ressources nécessaires seront à la disposition de la population en détresse.

7.G Éradication des épidémies :

Le big data aussi utilisé dans le contrôle de diffusion d'épidémies, en surveillant la migration de la race porteuse de maladies à travers le monde et l'analyse des itinéraires pour garantir un contrôle de la propagation de l'épidémie.

7.H Prévention d'attaques cybernétiques :

Actuellement, le big data offre des outils d'analyse et de traitement de données permettant la détection des intrusions, les failles sécuritaires et aussi les attaques cybernétiques qui sont dû à l'immensité et diversité des données transmises sur internet. Avec les techniques de traitement de données Big data on arrive à tracer le schéma relationnel entre les données et effectuer des calculs statistiques qui permettent de surveiller et d'intervenir, en temps réel, sur les menaces et les attaques cybernétiques à l'échelle mondiale (Krishna, 2015)

Figure 1-3: Domain d'application de Big data

8. Défis et enjeux :

L'accroissement des données produites par les entreprises, particuliers, les smartphones ou réseaux sociaux constituent un énorme défi en matière d'acquisition, stockage et traitement. Cette masse volumineuse d'information est parfois difficile à gérer et traiter, donc il devient nécessaire d'utiliser les nouvelles technologies telles que le Cloud Computing et les base de données NoSQL.

Nous citons ci-dessous les différents défis auxquels est confrontée la technologie BigData :

- ✓ La réduction de la redondance et la compression des données : Compression des données après avoir réduit au maximum les redondances, ce qui permettra de réduire le cout sans influer négativement sur la valeur des données.

- ✓ La gestion du cycle de vie des données : Afin d'éviter la saturation du système, il est préférable de supprimer les données superflues et de ne garder que ce qui est important.

- ✓ Mécanisme analytique : traiter des données hétérogènes massives dans un temps limité.

- ✓ La confidentialité et la sécurité des données : puisque le BigData englobe de grandes quantité de données, il est très difficile pour les entreprise d'appliquer et assurer la sécurité de ces données, ils auront alors besoin de l'aide des professionnels dans le domaine ce qui présente aussi un risque de sécurité potentiel.

- ✓ Gestion de l'énergie : contrôler et optimiser la consommation de l'énergie, vu que cette dernière ne cesse de progresser au niveau des systèmes de stockage.

- ✓ Évolutivité : le système d'analyse Big data doit prendre en charge les ensembles de données actuelles et futures. Les algorithmes doivent être en mesure de traiter des ensembles de données en expansion permanente. (Liu, 2014)

I. 3 VIE PRIVEE :

1. Vie privé : survole générale :

Le concept de vie privée est relatif à chaque pays conformément à sa culture et à sa législation. Il représente toute information qui concerne une personne identifiée. On associe une atteinte à une vie privé toute collecte, analyse, stockage, divulgations, altération et destruction de données personnelles d'un individu donnée. La protection de l'information considérée comme privée est nécessaire afin d'empêcher qu'elle soit partagée sans le consentement de son propriétaire.

Les recherches de ces dernières années ont abouti à un développement des techniques de protection de la vie privée. Parmi ces techniques, le chiffrement qui joue un rôle essentiel en matière de protection. L'anonymisation des données et d'autres techniques qui rendent l'accès à l'information par des personnes indésirables difficiles (Sithu . Sudarsan, 2015)

2. Types de vie privée :

Le concept de « vie privée » diffère de « la protection de vie privée » qui est lié à la protection des données. Nous citons ci-dessous les différents types de vie privée :

- ✓ La vie privée de la personne : La vie privée d'une personne peut s'apparenter au droit de garder les caractéristiques du corps (ou fonctions) tel que les codes génétique et biométriques.
- ✓ La vie privée du comportement et de l'action : concerne les habitudes liés aux activités politiques (ou sociales) des individus.

- ✓ La vie privée de la communication : consiste à éviter toute tentative d'espionnage ou vis-à-vis des communications.
- ✓ La vie privée des données et des images : consiste à protéger ces données afin qu'elles ne soient pas à la portée de tout organisme non autorisés.
- ✓ La vie privée des pensées et des sentiments : concerne la protection des pensées et sentiments des individus contre la divulgation sans leurs autorisations.
- ✓ La vie privée du déplacement et de l'espace : concerne la protection des libertés des individus à se déplacer dans des lieux publics (ou semi public) et chez soi sans surveillance
- ✓ La vie privée de l'association : la liberté de s'associer à des groupes sociaux sans être dérangé. (J.Camenisch, 2014)

3. Défis et enjeux de la vie privée dans le MapReduce :

MapReduce est un modèle de programmation adapté au traitement de grande quantité de données, basé sur le parallélisme, il permet la distribution de ces données dans un cluster de machines pour être traitées.

Nous allons citer quelques défis qui concernent la sécurité et la vie privé pour les calculs mapreduce.

- Les masses de données fractionnées par le mapreduce et distribuées sur les nœuds du cluster doivent être transférées en toute sécurité.
- Haute distribution: Le mapreduce opère sur un nombre de clusters massifs, la distribution des données sur ces clusters peut alors causer un risque d'attaque, d'où le besoin de rendre plus sûr ce mode de fonctionnement. On doit alors assurer une sécurité robuste de ces clusters en procédant sur chaque machine participante au traitement et stockage de la donnée.
- L'accès aux donnée non autorisées: La flexibilité de mapreduce exige une grande prudence par les utilisateurs, car les opérations lecture/écriture indésirables représentent un grand risque, c'est pour cela qu'il faut mettre au œuvre des algorithmes de sécurité afin de contrôler l'accès aux données et faire face à ce genre de risque.
- La protection de la vie privée contre les fournisseurs du cloud: vu que les données privées des utilisateurs sont stockées au niveau du cloud sachant que les fournisseurs de ce service peuvent contrôler et accéder aux données et au code mapreduce et peuvent

même les modifier, ce qui implique l'impossibilité de garantir la protection de la vie privée en présence d'un fournisseur cloud.

- Des multi-utilisateurs sur un seul nuage public : les fournisseurs de données et les fournisseurs de Cloud public doivent permettre à plusieurs utilisateurs d'accéder à leurs données simultanément sans que la vie privée de chacun ne soit menacée. On doit assurer à chaque utilisateur d'accéder à la totalité de ses données sans entraves tout en protégeant les données des autres utilisateurs (P.Derbeko, 2015) (Francis, 2014)

4. Sécurité via vie privée :

Les scientifiques et législateurs en la matière cherchent toujours à protéger la vie privée des personnes ou à assurer la sécurité publique des citoyens ou des utilisateurs utilisant un certain service. La sécurité et la vie privée sont deux mots qui diffèrent en terme de définition. Avec l'apparition des différents réseaux sociaux et l'émergence de l'informatique, la vie privée est affectée ce qui a poussé les experts en sécurité informatique à développer des outils et méthode afin d'assurer la sécurité et la vie privée des utilisateurs.

La sécurité vise à réduire les risques concernant les informations qui sont reliées à des ressources, qui elles même sont reliées à des entités qui peuvent être des personnes ou des entreprises. Les craintes sur la vie privée sont en croissance continue. (Sithu . Sudarsan, 2015). La vie privée est ce qui n'appartient pas à la vie publique, c'est l'ensemble de renseignements lié à un individu, qui reste constamment cachés et qui doit être protégé contre toute divulgation.

5. Gestion de la confiance :

Il existe une forte relation entre la confiance, la sécurité et vie privée, en assurant la sécurité du système et la protection de vie privée on arrive à concrétiser la confiance.

Dans le cloud, deux aspects de confiance doivent être mis en considération, la confiance dur « hard trust » et la confiance douce « soft trust », la 1ere est établit si les plates-formes des services contiennent des primitives de sécurité nécessaires, le taux de confiance est tiré sur la base des techniques par exemple cryptage, des contrôles et des certificats. La 2eme plutôt subjective concerne toute opinion humaine telle que les émotions, les perceptions, les expériences et les commentaires.

Dans le big data il est important de gérer et renforcer la confiance par la disponibilité et fiabilité des services offerts, le renforcement de la sécurité des données utilisateurs, et la protection de leurs vie privées (Shuyu Li, 2016)

6. Infrastructure critique et Big data :

L'infrastructure critique (IC) est l'ensemble des biens et services qui ont une importance capitale pour la population (santé, transport, nourriture, énergie…etc.). La défaillance du IC peut avoir de grave conséquence ce qui pousse à assurer sa protection.

La protection de l'infrastructure critique (PIC) vise à neutraliser le risque de la discontinuité ou bien de la disponibilité du service ou de bien.

L'infrastructure critique de l'information présente une partie non négligeable de l'infrastructure critique d'une société donnée.

La protection de l'infrastructure critique BigData est devenue une priorité car le BigData est une révolution informatique et technologique qui touche plusieurs secteurs et plusieurs métiers tels que la défense, la santé, l'économie et la recherche scientifique. Le volume de données grandit au fur et à mesure et le traitement local de l'information n'est plus possible, d'où le recours à un stockage et un traitement centralisés ou en cluster (Sithu . Sudarsan, 2015).

I. 4 TERMINOLOGIE DU DOMAINE DE LA VIE PRIVEE

1. Anonymat :

L'anonymat se définit comme étant un état non identifiable au sein d'un ensemble de sujets, qui est l'ensemble d'anonymat (ensemble de tous les sujets possible). L'anonymat est fort si l'ensemble d'anonymat est grand et si les sujets de réceptions et d'envois sont répartis équitablement.

L'ensemble d'anonymat est constitué de deux ensemble qui peuvent être disjoints, identiques ou se chevaucher :

- Ensemble d'expéditeur (les entités agissantes) : qui sont les sujets susceptibles de provoquer une action (expéditeur).
- Ensemble de destinataire (les entités de destinations) : qui sont les sujets qui peuvent être abordés.

Il faut savoir qu'un expéditeur ne peut être anonyme que s'il appartient à un ensemble d'expéditeur potentiel, de même un destinataire ne peut être anonyme que s'il appartient à un ensemble de destinataire potentiel.

Il existe deux aspects qui décrivent parfaitement l'anonymat :

- Quantité d'anonymat : qui définit la quantité d'anonymat appartenant à un environnement ou ensemble particulier.
- Robustesse : La robustesse de l'anonymat caractérise la stabilité de la quantité d'anonymat par rapport aux modifications d'un certain paramètre par exemple.

Pour continuer, nous pourrions utiliser le terme « qualité d'anonymat » comme un terme comprenant à la fois la quantité et la robustesse de l'anonymat.

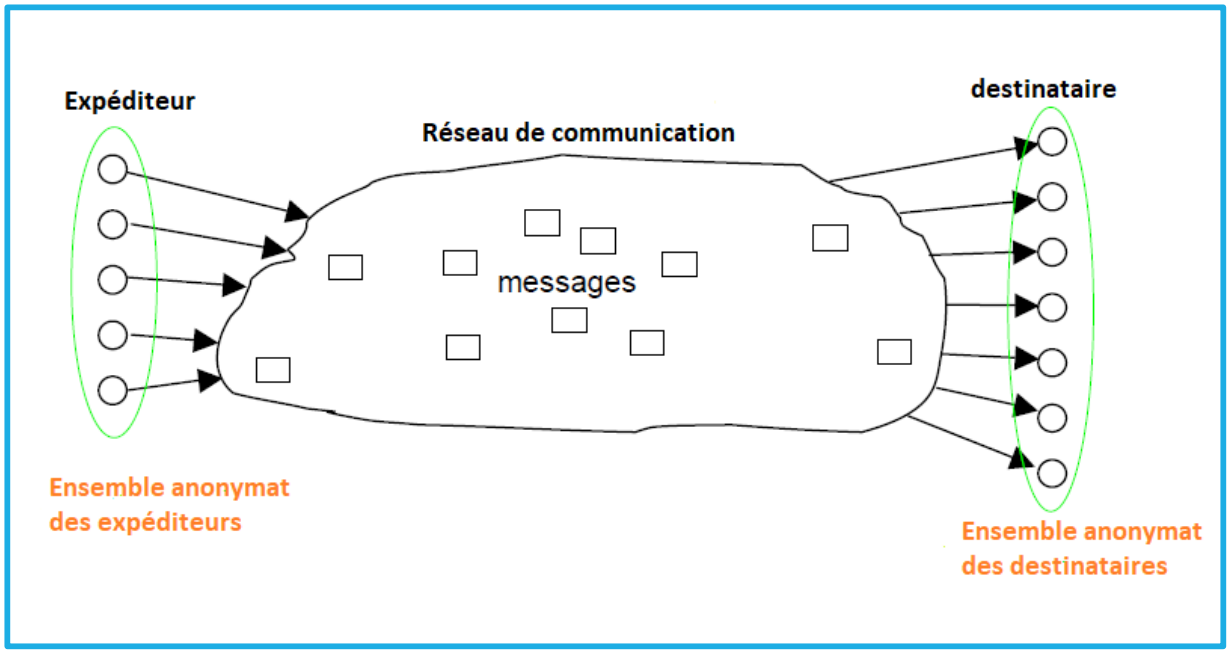

Figure 1.4: Schéma explicatif de la communication entre 2 ensembles d'anonymats

2. Intraçabilité (Unlinkability) :

Le concept de l'intraçabilité représente l'impossibilité de trouver une relation significative entre deux ou plusieurs entités d'intérêts (Items Of Interest or IOI) tels que les sujets, les messages, les évènements... etc. Cela signifie qu'au point de vue de l'attaquant, ces IOI ne sont ni moins ni plus liés avant et après l'observation et donc il est difficile, voire impossible de construire le schéma de liaison entre ces entités (Andreas Pfitzmann, 2010)

3. Inobservabilité (non-observabilité) :

L'inobservabilité est l'état d'un objet d'intérêt (sujet, message, action…) qui ne peut pas être distingué des autres objets d'intérêt (IOI). Comme nous avions un ensemble d'anonymat qui respecte l'anonymat, nous avons dans ce cas un ensemble d'inobservabilité qui respecte l'inobservabilité.

On peut distinguer 3 types dans l'ensemble d'inobservabilité :

- Sender unobservability : ou non-observabilité de l'expéditeur signifie qu'il n'est pas visible si un sujet appartenant à l'ensemble d'inobservabilité envoie un message.
- Recipient unobservability : ou non-observabilité du destinataire signifie qu'il n'est pas visible si un sujet appartenant à l'ensemble d'inobservabilité reçoit un message.
- Relationship unobservability : ou non-observabilité des relations signifie qu'il est impossible de savoir si un message est envoyé en dehors de l'ensemble d'expéditeurs potentiels et destinataires potentiels (Andreas Pfitzmann, 2010)

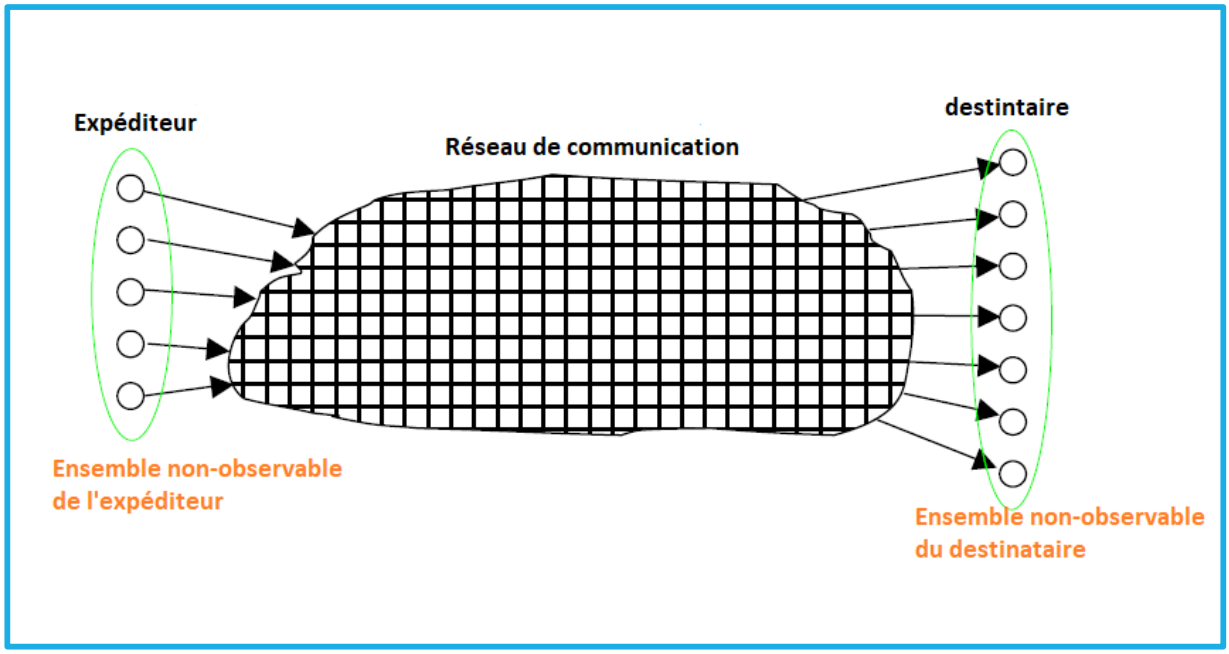

Figure 1.5: Schéma explicatif de la communication entre 2 ensembles non-observable

4. Pseudonymat :

Un pseudonyme est un identifiant cohérent attribué à un sujet, dans notre cas, peut être un expéditeur ou destinataire. Le pseudonymat garanti aux utilisateurs une communication

généralement anonyme, étant donné que personne ne connait la vraie identité de l'autre, cela contribue à assurer la confidentialité et à protéger la vie privée.

5. Gestion d'identité :

La gestion des identités est l'ensemble des processus et outils permettant d'identifier, authentifier et autoriser à un individu ou groupes de personnes l'accès à des ressources et applications de manière sécurisé.

5.A Identité et identifiabilité :

- Identité : Est un ensemble d'attributs qui caractérise un individu parmi les autres individus qui fait sa singularité.
- Identifiabilité : Est l'état d'être identifiable parmi un groupe d'individus, l'ensemble identifiable.

Plus l'identifiabilité est forte plus l'ensemble d'identifiabilité correspondant est grand, inversement, plus l'anonymat est fort, plus l'ensemble d'identifiabilité est petit. Voir Figure 1.6.

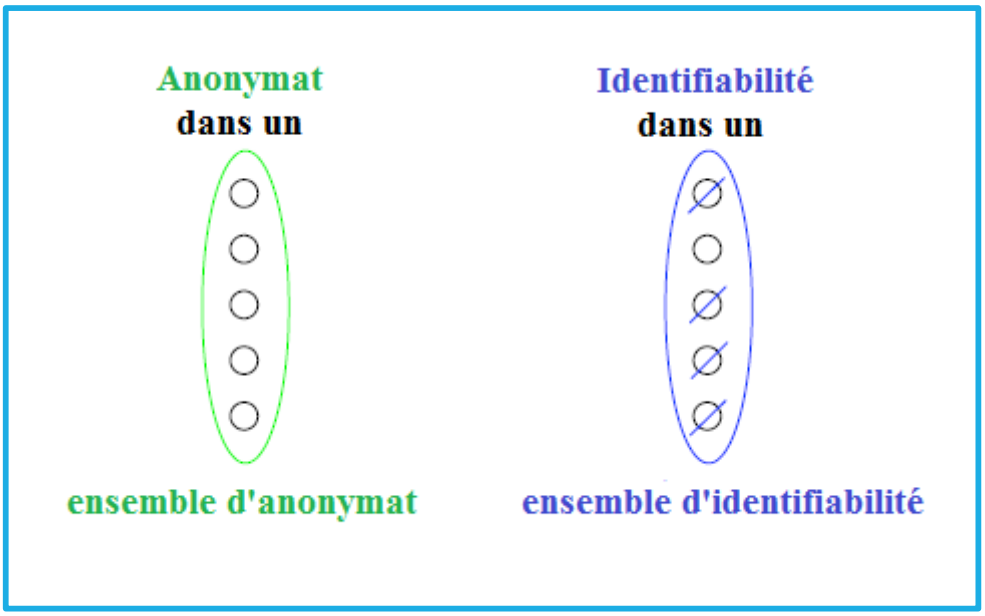

Figure 1.6: Schéma illustratif d'un ensemble d'anonymat et un ensemble d'identifiabilité

5.B Termes liés à l'identité

- Identité partielle : Est un sous-ensemble de caractéristiques d'une identité complète, où une identité complète est l'union de toutes les caractéristiques de toutes les identités partielles d'une personne donnée.
- Identité numérique : Désigne un ensemble de propriétés d'une personne qui peuvent être stockées et traitées par une technologie informatique.
- Identité virtuelle : Utilisée par un individu lors de ses activités sociales ou commerciales sur internet. Elle peut être compatible à son identité réelle, mais elle est souvent différente (souhaitant la protection de ses attributs personnels)

5.C Termes relatifs à la gestion d'identité :

- Gestion de l'identité : consiste à gérer plusieurs identités partielles qui sont généralement désignés par des pseudonymes d'un individu.
- Gestion des identités renforçant la confidentialité : la gestion d'identité est dite gestion des identités renforçant la confidentialité si elle ne permet pas plus de possibilité de lien entre les identités partielles.
- Gestion de l'identité renforçant la protection de la vie privée et permettant la conception d'applications : Une application est conçue de manière à permettre la gestion de l'identité renforçant la protection de la vie privée si ni le modèle d'envoi/réception ni les attributs donnés aux entités (personne, organisation, ordinateur…) n'impliquent plus de possibilité de liaison que ce qui est nécessaire pour atteindre les objectifs de l'application.
- Système de gestion d'identité (IMS) : fait référence à un ensemble de technologies (application, outils…etc.) qui assurent la gestion de l'identité au sein d'une entreprise ou inter-réseau.
- Système de gestion de l'identité renforçant la confidentialité (PE-IMS) : permet de donner à l'utilisateur un degré de contrôle en lui permettant de faire un choix de pseudonymes qui représentent ses identités partielles, il assiste aussi l'utilisateur dans la gestion des identités partielles (Andreas Pfitzmann, 2010).

6. Les techniques de protection de la vie privée en big data :

- De-identification : Est le processus qui empêche de retrouver les informations d'un individu à partir de son identité, ancienne technique appliquée dans le data mining, qui peut être lité dans le big data afin de garantir la confidentialité. Cette méthode seule ne suffit pas pour mettre en évidence la protection de la vie privée, c'est pour cela qu'elle a été enrichie par les techniques que nous allons résumer sur le tableau suivant : (Machanavajjhala A, 2006)

Techniques	Définitions	Faiblesses
K-anonymat	Une technique d'anonymisation se base sur la construction et l'évaluation des algorithmes et systèmes qui garantissent la protection des données de telle sorte à limiter ou à enlever un degré de précision à certains champs (voir exemple dessous la figure) (Nichterlein A, 2011).	Reste vulnérable aux attaques d'homogénéité.
L-diversité	Représente une extension du modèle K-anonymat, résout la majorité de ses faiblesses. Dans cette méthode, chaque attribut sensible ait au moins L valeurs distinctes. Ainsi, il est difficile de retrouver la valeur de la donnée exacte.	Repose sur la gamme de données sensibles.
T-proximité	Etend de la l-diversité, on dit qu'un ensemble d'enregistrement contenant les même données anonymisées a une proximité T, si la distance entre la distribution d'un attribut sensible de cet ensemble et la distribution de l'attribut dans la table entière n'est pas supérieure à un seuil T [(Microsoft, 2015)].	Si la taille des données augmente, la probabilité de ré-identification augmente aussi.

Tableau 1: Tableau comparatif des 3 techniques de de-indentification

Exemple : Si nous voulons identifier une personne et tout ce que nous disposons comme informations sont l'âge et la wilaya, il devrait au moins y avoir k nombre d'entrés répondant à cette exigence.

- Confidentialité différentielle : Est un mécanisme utilisé pour protéger la vie privé des individus, restreint l'accès direct aux données en élaborant un logiciel intermédiaire qui sera mis entre l'analyste et la base de données (Xu L, 2014).

  Le fonctionnement est définit comme suit :

Etape 1 : L'analyste envoie une requête à la base de données en passant par le logiciel intermédiaire.

Etape 2 : Le logiciel évalue l'impact de la requête sur la vie privée en utilisant un algorithme spécifique.

Etape 3 : Le logiciel envoie la requête à la base de données et récupère une réponse brute sans aucune modification.

Etape 4 : Ajoute ensuite la quantité appropriée de «bruit», selon le risque évalué sur la vie privée (étape 2), rendant la réponse incertaine afin de garantir la confidentialité des personnes dont les informations figurent dans la base de données, pour terminer, il renvoie la réponse modifiée à l'analyste.

On peut résumer les étapes précédentes dans la figure ci-dessous :

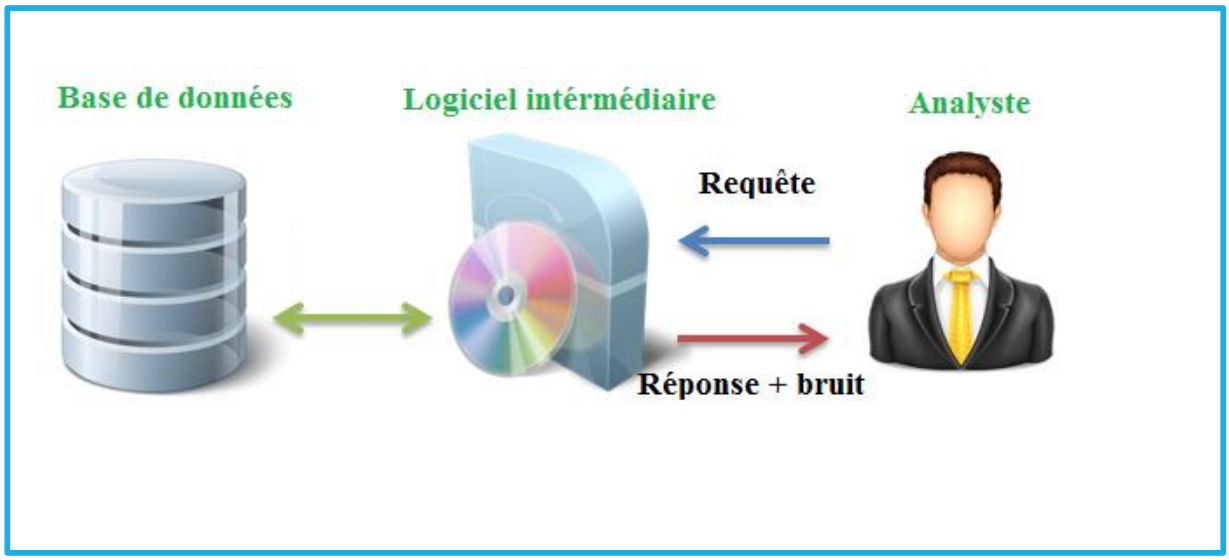

Figure 1.7: Schéma récapitulant la confidentialité différentielle

7. Opportunité des Big data pour la protection des vies privées :

Avec la croissance croissante de la protection des données personnelles dans les bigdata, nous citons ci-dessous quelques opportunités pour les entreprises utilisant cette technologie afin d'assurer la protection de ses données :

- Développement de nouveaux services et produits qui garantissent aux clients ou utilisateurs que leurs données sont stockées en toute sécurité.

- L'utilisation des données sans enfreindre les lois de confidentialité. Un exemple simple : l'utilisation de nouveaux guichets permettant de comprendre les clients d'une entreprise sans la collecte de données personnelles.
- L'innovation de nouvelles technologies afin de tirer parti de l'analyse « BigData » tout en maintenant la confidentialité.
- Les éditeurs de logiciels Intègre la protection de la vie privée dès la conception (TEETZ, 2018)

8. Les big data en cloud computing :

Plus la quantité de données est importante plus le besoin de nombreux systèmes de stockage augmente, Bien que le cloud computing fournit un traitement massif et une capacité de stockage vitale, le problème de confidentialité de données reste toujours posé, on présentera ci-dessous les défis concernant le mécanisme de traitement et de stockage sur le cloud (Li N, 2007)

- Externalisation : Quand les données sont externalisées cela veut dire que l'utilisateur n'a aucun contrôle sur ses données, d'où la principale cause de la perte de confidentialité des utilisateurs du système cloud.
- Multi-location : Est une architecture du cloud computing, qui permet à plusieurs clients, appelé locataire, de partager les mêmes ressources informatique dans l'environnement cloud. Le risque de rupture de données et de calcul est assez répandu dans cette architecture car n'importe quel utilisateur non lié au locataire pourrait accéder à ces données ce qui cause une atteinte à la confidentialité.
- Calcul massif : il est difficile d'assurer la confidentialité des données personnelles par les systèmes classiques, car ces derniers épuisent toutes leurs énergies pour la gestion d'une énorme quantité de données ainsi que leurs calculs.

8.A Confidentialité des Big Data en phase de génération de données:

Dans la phase de génération de données (introduction de données sur internet par exemple), les informations personnelles peuvent être transmises par des personnes non autorisées, afin d'éviter ce type de problème, les utilisateurs ont à leurs dispositions des outils qui masqueront l'identité du propriétaire de ces données ou même les données.

8.A Confidentialité des Big Data en phase de stockage de données :

Dans la technologie Big Data, les méthodes de stockage normales ne sont pas suffisantes. Bien que les données soient stockées sur le cloud, la confidentialité des données doit être respectée selon trois dimensions: confidentialité, intégrité et disponibilité. La confidentialité et l'intégrité sont pleinement liées à la confidentialité des données. Mais la disponibilité des données fait en sorte que seuls des personnes autorisées peuvent y accéder. Lors de la phase de stockage, certaines technologies de cryptage préservent la confidentialité du propriétaire des données comme (Mehmood A, 2016):

- Cryptage basé sur l'identité: le contrôle d'accès est basé sur l'identité de l'utilisateur.
- Cryptage basé sur les attributs: contrôle d'accès basé sur les attributs de l'utilisateur, plus sûr et plus flexible que celui basé sur l'identité. Il est très difficile de gérer utilisateurs de catégories différentes. Mise à jour du récepteur de texte crypté impossible.
- Rechiffrement proxy: la mise à jour du récepteur de texte chiffré est possible.
- Cryptage homomorphique: Calcul sur des données performantes et très sécurisées. Généralement exploité par les clients qui n'ont pas assez de moyens pour effectuer des calculs sur leurs données alors ils font appel au service de cloud computing avec cette propriété afin de garantir la confidentialité des données et résultats.

8.C Confidentialité des Big Data en phase de traitement de données :

Se divise en deux parties, la première concerne la protection des données contre les fuites non permises lors du traitement, la deuxième consiste à extraire les informations significatives à partir des données sans perdre la confidentialité.

I. 5 CONCLUSION :

Le BigData est une technologie révolutionnaire et naissante, qui prend de plus en plus d'importance pour un grand nombre d'entreprise vu l'énorme quantité de données stockées et traitées. Cependant, la majorité ne maitrise pas vraiment les concepts de base en matière de sécurité, ce qui poussent les pirates informatiques à récupérer les données du manière illicite, c'est pour cela que la sécurité du BigData est une étape cruciale à assurer afin d'éviter toute attaque qui peut mener à la divulgation ou perte des données sensibles du système tout entier.

Chapitre II :
Approches et travaux connexes

II.1 INTRODUCTION :

Dans ce chapitre, nous allons d'abord survoler les solutions suggérées qui consistent à renforcer la sécurité des données privées dans le big data afin de les analyser et d'en tirer les inconvénients pour chaque approche proposée, ensuite, construire une table comparative dans laquelle nous allons comparer entre les différentes solutions étudiées, enfin, on clôturera le chapitre par une synthèse de discussion.

II.2 ANONYMISATION MULTI DIMENSIONNELS :

1. Problème :

Avec l'apparition des Bigdata de grandes quantités de données sont récoltés à partir des téléphones mobiles (smartphones), internet des objets (IoT), capteurs réseaux et les médias sociaux. Les plateformes clouds dans lesquelles sont stockées cette quantité volumineuse de données et ses outils de traitement sont devenus de plus en plus connue et ciblés par les différents pirates informatiques pour la prise en charge des différentes applications de traitement et analyse des données et son extraction.

Bien que le cloud présente de nombreuses caractéristiques essentielles telles que la rentabilité et l'évolutivité, les problèmes de confidentialité restent l'un des obstacles majeurs à l'adoption des ressources de cloud public dans de nombreuses applications sensibles au respect de la vie privée dans des secteurs tels que la santé, la finance et la défense. Le risque de confidentialité est généralement causé par la redondance des informations provenant de diverses sources de données dans un ensemble volumineux de données (X. Wu, 2014) (Chaudhuri, 2012).

Diverses solutions et implémentations ont été mise en place afin de résoudre ce problème comme par exemple le schéma d'anonymisation multidimensionnel que nous allons traiter dans cette partie.

2. Contribution et implémentation :

Afin d'assurer la confidentialité des données contenu dans les Bigdata ou le cloud plusieurs techniques et méthodes ont été mises en place, ces méthodes sont classées en :

- Schéma de recodage global : qui partitionne les données par attribut.
- Schéma de recodage local : qui partitionne les données par instance.

Le schéma de recodage local est peu utilisé car ses données engendrées sont incohérente et qu'il a un problème d'exploration des données, malgré le fait qu'il engendre moins de distorsion des données et terme de confidentialité (B. Fung, 2010)

Plusieurs techniques qui suivent le schéma de recodage globale ont été mises en place, afin d'assurer la généralisation des données pour la préservation de la confidentialité des données dans des scénarios de publication ou de partage de données. La généralisation des données permet de masquer l'identité et/ou des données confidentielles en remplaçant des valeurs d'attributs détaillées par des valeurs plus générales, dans le but de préserver la confidentialité de l'individu.

Parmi les méthodes qui permettent d'assurer la généralisation des données nous citons :

- le schéma d'anonymisation multidimensionnel : est un schéma de recodage global qui établit un bon équilibre entre la distorsion des données et la facilité d'utilisation des données.
- le schéma de sous-arbre : Il partitionne les données en un seul attribut et entraîne une distorsion des données beaucoup plus grande que le schéma multidimensionnel.

Des problèmes d'extensibilité ont été remarqués dans le schéma multidimensionnel présentés ci-dessus et plusieurs solutions ont été proposées parmi elles :

- ✓ Un algorithme de partitionnement récursif appelé Mondrian
- ✓ approche basée sur un index R-tree pour obtenir une construction d'index efficace et une anonymisation en bloc.
- ✓ les algorithmes d'arbre de décision évolutif RainForest (J. Gehrke, 1998) et une technique d'échantillonnage qui consiste à diviser un ensemble de données en partitions de données plus petites pouvant tenir dans la mémoire.

3. Inconvénients :

-Les approches qui assurent l'évolutivité du système multidimensionnel sont essentiellement sérielles, même si elles sont extensibles à de grands ensembles de données. Ils doivent souvent analyser l'ensemble des données pour obtenir des statistiques de comptage lors du choix d'attributs de fractionnement ou de valeurs de domaine.

-Le temps qui lors de l'opération d'entrées et sorties sera énormément élevés pour les grands ensembles de données.

-L'utilisation de paradigmes parallèles distribués et parallèles directement pour le calcul récursif reste un défi, car ces paradigmes sont à l'origine conçus pour le traitement par lots ou l'informatique orientée flux.

-Il reste difficile d'arriver à un résultat exact et performant quand il s'agit de l'évolutivité et l'utilité des données pour le schéma multidimensionnel par rapport au Big Data.

II.3 ANONYMISATION PAR PROXIMITE AVEC MAPREDUCE :

1. Problème :

Le cloud computing fournit une infrastructure informatique évolutive prometteuse, capable de prendre en charge divers traitements de diverses applications Big Data dans des secteurs tels que la santé et les entreprises. Les ensembles de données tels que les dossiers médicaux électroniques dans de telles applications contiennent souvent des informations confidentielles, ce qui peut poser problème si les informations sont divulguées ou partagées avec des tiers dans le cloud.

Plusieurs approches ont été mises en place afin de palier au problème cité ci-dessus, Cependant, la plupart des approches existantes préservant la confidentialité comme le recodage local de l'anonymisation des données volumineuses contre les atteintes à la confidentialité à proximité, sont adaptées aux ensembles de données à petite échelle, font souvent défaut lorsqu'elles rencontrent des données volumineuses, en raison de leur insuffisance ou de leur faible évolutivité (L. Wang, 2012).

2. Contribution et implémentation :

Il est intéressant et pratique de modéliser le problème en tant que problème de regroupement visant à minimiser la distorsion des données et la proximité entre valeurs sensibles dans un cluster. Afin de résoudre le problème d'extensibilité (adaptation des solutions à grande échelle de données), quatre principales contributions ont été faites dans cette approche. Premièrement, la modélisation du problème du recodage local de grandes quantités de données contre les atteintes à la vie privée de proximité en tant que problème de clustering sensible à la proximité.

Deuxièmement, un aspect évolutif et efficace de clustering à deux phases est proposé pour paralléliser le recodage local sur plusieurs partitions de données. Troisièmement, plusieurs travaux MapReduce innovants sont conçus et coordonnés pour effectuer en toute confiance des calculs parallèles de données en vue de leur extensibilité (B.C.M. Fung, 2010).

- Modélisation du problème des clusters sensibles à la proximité :

Consiste à tenir en compte à la fois de la similitude des quasi-identifiants et de la proximité des valeurs sensibles. Intuitivement, on souhaite que les enregistrements tendent à être regroupés dans les mêmes clusters si leurs quasi-identifiants sont semblables, et que la proximité des valeurs sensibles entre eux est faible, c'est-à-dire que les valeurs sensibles sont dissemblables. Et ainsi, palier au problème du recodage local de grandes quantités de données contre les atteintes à la vie privée de proximité.

- Approche de groupement en deux phases :

La première phase divise un ensemble de données original en partitions qui contiennent des enregistrements de données similaires en termes de quasi-identifiants. Dans la deuxième phase, les partitions de données sont recodées localement par l'algorithme de clustering agglomératif de proximité en parallèle.

MapReduce, effectuer des calculs parallèles de manière créative et évolutive :

L'implémentation des algorithmes est faite avec MapReduce afin d'obtenir une extensibilité élevée en effectuant des calculs parallèles sur plusieurs nœuds de calcul dans le cloud. Ainsi, cette approche peut traiter des ensembles de données à grande échelle d'une manière linéaire, ce qui peut être accompli facilement dans des environnements cloud grâce à leur évolutivité.

3. Inconvénients :

La préservation de la vie privée pour l'analyse, le partage et l'exploration des données est une question de recherche difficile en raison du volume de plus en plus important d'ensembles de données, ce qui exige des recherches intensives.

La conception de tâches MapReduce appropriées pour des applications complexes reste un défi car MapReduce est un paradigme de programmation contraint. Habituellement, il est nécessaire de considérer les problèmes comme quelle partie d'une application peut être parallélisée par MapReduce, comment concevoir les fonctions Map et Reduce pour les rendre

évolutives, et comment réduire le trafic réseau entre les nœuds des travailleurs. Les réponses à ces questions varient souvent selon les applications. Par conséquent, des recherches approfondies sont encore nécessaires pour concevoir des tâches MapReduce pour une application spécifique (X. Zhang, 2013).

II.4 STOCKAGE MULTI PARTAGE :

1. Problème :

Certains mécanismes de sécurité comme le hashage et le cryptage existent aujourd'hui pour assurer la sécurité des données sur le Cloud, néanmoins ces mécanismes ne permettent pas de satisfaire tous les besoins essentiels pour la protection d'une grande quantité de données. Les systèmes de cryptage traditionnels ne tiennent pas compte de l'anonymat d'un expéditeur / récepteur de texte chiffré. En conséquence, n'importe quelle personne ayant la capacité d'accéder au texte chiffré peut savoir sous quelle clé publique le texte chiffré est chiffrée, et donc connaitre le propriétaire de ce texte chiffré. De manière similaire, le destinataire du texte chiffré peut être connu à partir du texte crypté sans aucune difficulté. Cela porte gravement atteinte à la vie privée du propriétaire d'un texte chiffré (ex : patient).

La mise à jour du destinataire du texte chiffré est souhaitable et nécessaire puisqu'un texte chiffré pourrait être partagé conditionnellement avec plusieurs personnes. Le propriétaire d'un texte chiffré (ex : le patient) a le droit de décider qui peut avoir accès à son contenu et quels types de données peuvent être consultés.

2. Contribution et implémentation :

Les chercheurs ont essayé de développer un mécanisme qui porte les propriétés suivantes :

- Anonymat : étant donné un texte chiffré, personne ne connaît l'expéditeur et le destinataire correspondants.
- Multiple receiver-update : étant donné un texte chiffré, le destinataire du texte chiffré peut être mis à jour en plusieurs fois.
- Partage conditionnel : un texte chiffré peut être finement partagé avec d'autres si les conditions pré-spécifiées sont remplies.

Pour arriver à model parfait les chercheurs ont passé par plusieurs modèles, que chacun d'eux a une propriété spécifique :

- Afin de préserver l'anonymat, certains mécanismes de chiffrement bien connus sont proposés dans la littérature, tels que le BIE anonyme (Waters X. B., 2006), le ABE anonyme (Y. Zhang, 2013). En utilisant ces primitives, la source et la destination des données peuvent être protégées de manière privée.

-L'une des approches mise en œuvre afin d'assurer la mise à jour du récepteur est le déchiffrement puis le chiffrement

-Proxy Re-Encryption (PRE) est proposé pour la 1ere fois par Mambo et Okamoto (Okamoto, 1997) afin résoudre le dilemme du partage des données. Il permet à une partie semi-digne de confiance appelée proxy de transformer un texte chiffré destiné à un utilisateur en un texte chiffré (du même texte en clair) destiné à un autre utilisateur sans qu'il y ait fuite de connaissance des clés de décryptage ou du texte en clair. Ce mode permet non seulement de tranférer la charge de travail du propriétaire des textes chiffré vers un proxy mais aussi de ne pas exiger la présence permanente du propriétaire du texte chiffré.

-Un modèle IBPRE anonyme qui est un système proposé par J. Shao (Shao, 2012) qui assure simultanément la confidentialité et la mise à jour du destinataire du texte chiffré.

- Le dernier modèle proposé afin d'assurer la mise à jour de plusieurs récepteurs est MH-IBPRE (MULTI HOP Identity-Based Proxy Re-Encryption) qui a été proposé par Chu et Tzeng (Tzeng, 2007) et qui connaitra plusieurs extensions parmi elles :

- MH-IBCPRE (Multi-Hop Identity-Based Conditional Proxy Re-Encryption)
- AMH-IBCPRE (Anonymous Multi-Hop Identity-Based Conditional Proxy Re-Encryption)

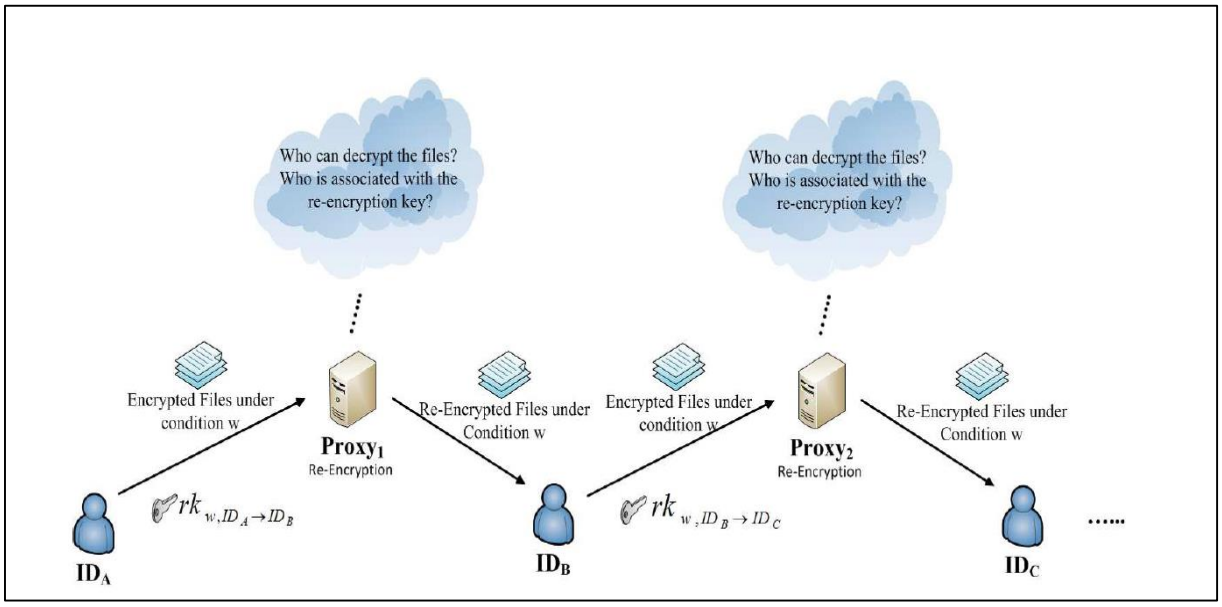

Figure 2.1 : Anonymous Multi-Hop Identity-Based Conditional Proxy Re-Encryption (Raju, 2016)

3. Inconvénients :

- Les primitives BIE et ABE mis en œuvre afin de préserver l'anonymat ne peuvent pas supporter la mise à jour du récepteur de cryptogramme.

- Dans le cas d'un déchiffrement et chiffrement par la suite pour assurer la mise à jour du recepteur, si les données cryptées sont volumineuses, le décryptage et le re-cryptage peuvent prendre du temps et être coûteux en calcul. De plus, ce mode de mise à jour souffre également d'une limitation selon laquelle le propriétaire des données doit être en ligne tout le temps.

- IBPRE ne prend en charge que la mise à jour d'un seul récepteur du texte chiffré et non de plusieurs.

II.5 PROTECTION PAR DETECTION DE COMPRESSION :

1. Problème :

L'exploration des bigdatas, leurs analyses et leurs partages devient de plus en plus fréquent, ce qui fait gonfler rapidement le volume de données, accélère le flux de données et améliore considérablement leur valeur. Parallèlement, les propriétaires de données perdent le contrôle absolu des données, et en raison du manque de supervision efficace, l'exploitation excessive des données, l'utilisation non autorisée, les transactions illégales et d'autres comportements auxquels est confrontée la sécurité des données.

Plusieurs méthodes de chiffrements ont été mis en place pour assurer la protection des données sensibles, cependant, les coûts généraux de ce genre de traitement est excessivement élevé, à savoir le temps, l'argent et l'énergie, vu la quantité énorme de données à prendre en charge.

2. Contribution et implémentation :

Une nouvelle méthode de protection de la vie privée basée sur la détection par compression est proposée afin de résoudre les problèmes auxquels sont confrontées les grandes entreprises de protection de la vie privée. Tout d'abord, rendre anonymes les grandes données et comprimer les grandes données anonymes en petites données à l'aide d'une technologie de détection par compression, tout en atteignant deux niveaux de cryptage des données. Ensuite, les utilisateurs des données peuvent appliquer la clé et l'algorithme de décryptage du propriétaire des données pour reconstruire avec précision les données originales et utiliser les données. Cette méthode évite le décryptage et le calcul directs de grosses données, réduit les frais généraux de calcul et peut reconstruire les données originales à partir de petites données sans affecter l'utilisation normale.

Deux principales contributions ont été faites dans cette méthode: D'abords concevoir l'architecture et le modèle théorique de préservation de la confidentialité des grandes données basées détection de compression (DC), Et ensuite réalisation d'un algorithme de préservation de la confidentialité des grandes données de la théorie à la pratique (Kargupta H, 2003) (Yingjie, 2015)

Nous présentons ci-dessous l'architecture de la méthode de protection de la vie privée par détection de compression.

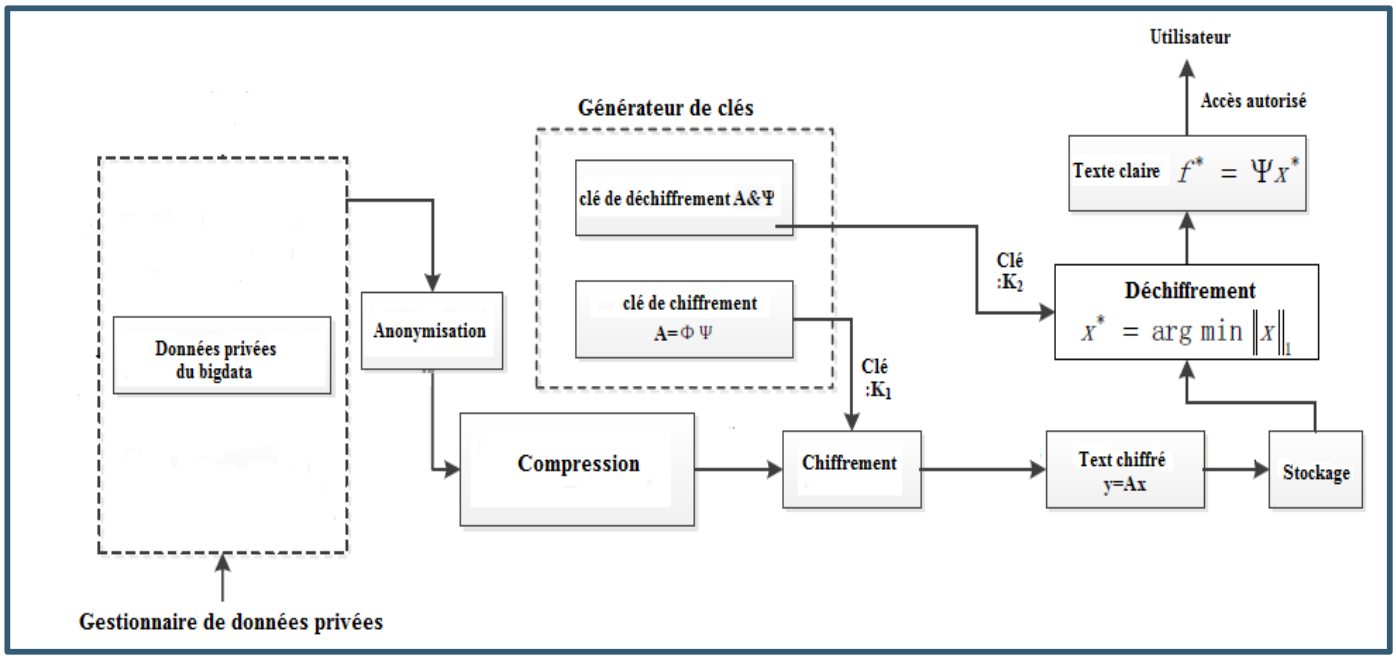

Figure 2.2: Architecture de protection de la vie privée par détection de compression

Principe de l'algorithme par détection de compression :

La solution se résume en 3 étapes :

1) Les données sont anonymisées en 1er lieu.
2) Les données subissent une compression afin d'éviter le cryptage de gros blocs de données et le traitement de décryptage.
3) Cryptage des données compressés, ces derniers peuvent être reconstruites, n'affecte pas l'utilisation normale pour les utilisateurs autorisés.

Grâce à l'algorithme de détection de compression, les frais généraux ont pu être réduits tout en garantissant une sécurité robuste des données sensibles dans le BigData.

3. Inconvénients :

Les gestionnaires des BigData représentent un facteur négatif, en raison de problèmes de gestion, il est facile d'atteindre à la vie privée et la divulgation de données sensibles en interne, corriger de nombreuses lacunes dans la plate-forme de gestion des données volumineuses reste encore un défi.

II.6 PROTECTION PAR ENREGISTREMENT LOCAL (LRDM) :

1. Problème :

Etant donné que le Big Data peut extraire de nouvelles connaissances pour la croissance économique et l'innovation technique, les chercheurs ont cherché à relever le défi pour capturer, stocker, gérer, partager, analyser et visualiser avec les outils de traitement les données existantes. L'analyse des ensembles de données donne un aperçu approfondi d'un certain nombre de secteurs clés de la société, les ensembles de données sont souvent partagés ou communiqués à des partenaires tiers ou au public (I. Stoica, 2001). Par conséquent, les informations privées des utilisateurs sont facilement capturées par le destinataire ou d'autres utilisateurs illégaux, il est donc souhaitable de concevoir des algorithmes efficaces et préservant la confidentialité pour le partage et le traitement des données volumineuses.

La confidentialité des BigData peut être préserver par trois approches, qui sont :

- approches basées sur le bruit.
- schémas basés sur le cryptage.
- techniques d'anonymisation.

Dans le scénario Bigdata, les approches actuelles de préservation de la vie privée se concentrent toujours sur la résolution des problèmes dans la phase de traitement et de stockage des données et ne tiennent pas compte de la phase de collecte des grandes données. La collecte de grandes données est une partie nécessaire, alors il ne fait aucun doute qu'il faut accorder plus d'attention à la phase de la collecte et aux informations locales.

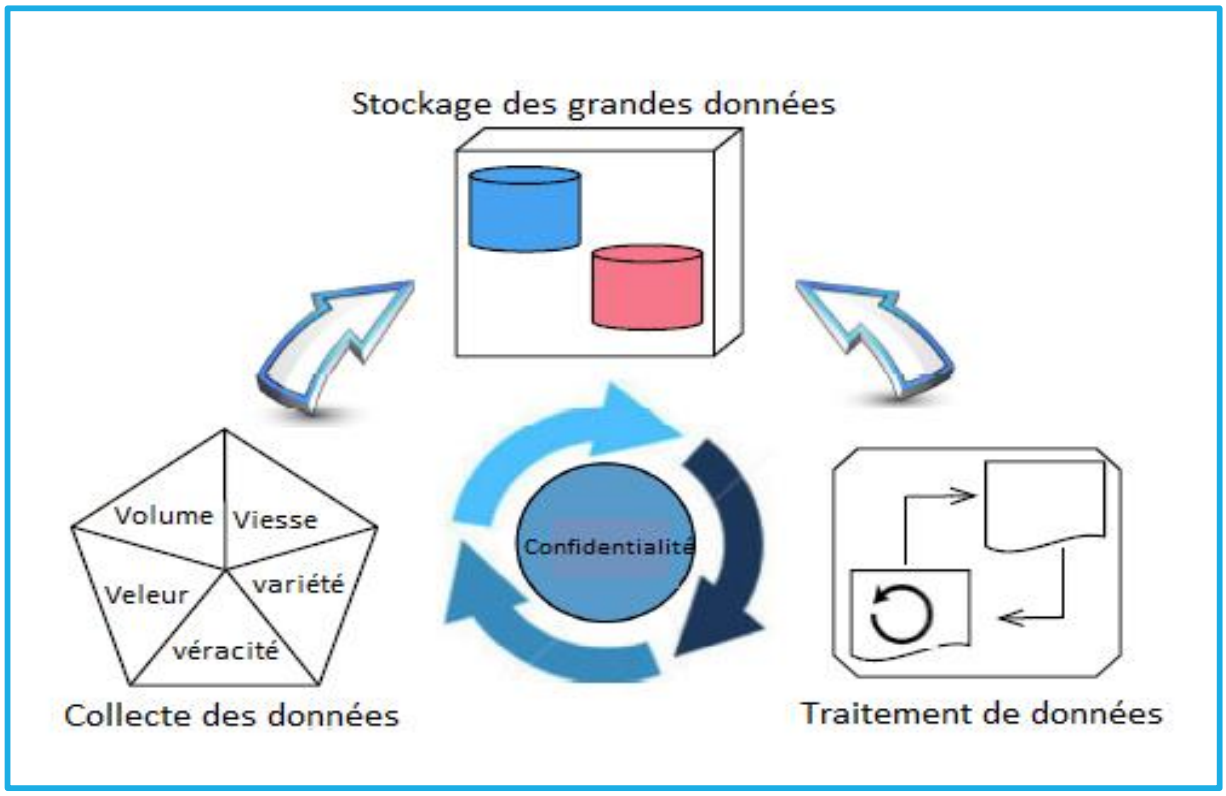

Figure 2.3 Architecture générale de la confidentialité des données volumineuses

2. Contribution et implémentation :

Un utilisateur individuel dispose de diverses données personnelles qui seront enregistré à l'aide des enregistrements locaux, ces données sont générées à partir de dispositifs électroniques intelligents personnels ou de dispositifs publics, tels qu'un smartphone, une tablette ou un PC.

Afin de protéger fondamentalement les informations sensibles personnelles des utilisateurs et de réduire les coûts de calcul et de communication, un mécanisme local de gestion des enregistrements a été proposé, qui consiste en une série de contraintes permettant à l'utilisateur de trouver la méthode optimale de préservation de la confidentialité qui est LRDM et utilise la protection différentielle de la vie privée pour optimiser la méthode de protection de la vie privée de l'utilisateur.

Le mécanisme LRDM passe par 3 étapes :

- Déclaration du problème : utilisation d'une fonction de dissimilarité qui permet de calculer la différence entre l'enregistrement réel de l'utilisateur et l'enregistrement estimé par l'adversaire (utilisateur illicite). (en LRDM on utilise la distance de Hamming pour formuler la fonction de dissimilarité).

- Mesure de la protection de la vie privée : La fonction de dissimilarité mentionnée précédemment permet de quantifier la perte de vie privée découlant de l'attaque d'inférence.
- Solution : Dans cette étape l'utilisateur cherchera à maximiser son intimité.

3. Inconvénients :

- L'utilisation permanente de la distance de Hamming pour quantifier la vie privée de l'utilisateur.

- Bien que le LRDM aide les utilisateurs à trouver un schéma approprié pour protéger leur vie privée mais il n'atteint pas la solution optimal.

II.7 VIE PRIVEE DIFFERENTIELLE :

1. Problème :

Le problème principal avec les bigdata dans le cloud est que les données traitées ou utilisées par un tiers. Il est très important pour les propriétaires de données ou les clients de faire confiance et d'avoir la garantie de confidentialité pour les données analysées ou stockées dans le cloud. Les modèles de protection de la vie privée étudiés dans le cadre de recherches antérieures ont montré que la violation de la vie privée pour les grandes données était due à la diffusion de données exactes qui peuvent être obtenues dans l'ensemble de données.

2. Contribution et implémentation :

Toutes les données ont les mêmes propriétés, ainsi que la question du respect de la vie privée, en particulier lorsque ces données sont stockées dans le cloud ou utilisées par des tiers. La présente étude propose de combiner plusieurs méthodes soit : anonymat k et confidentialité différentielle. La confidentialité différentielle implique la publication des résultats d'une requête avec un peu de bruit ajouté aux résultats de la requête. Dans ce cas, l'attaquant ne peut pas deviner les résultats de la requête car elle contient du bruit avec une garantie à 100%. Les résultats ont montré que l'anonymat k peut être utilisé pour améliorer la confidentialité différentielle et augmenter la garantie de protection des données. De plus, l'étude a évalué deux méthodes de protection de la vie privée (anonymat k et vie privée différentielle) en termes de

coût de calcul, de composabilité et de capacité de liaison lorsqu'il s'agit de grandes données. La recherche considère que l'anonymisation est le meilleur outil qui garantit la confidentialité des données volumineuses et réduit le risque de divulgation (Dwork, 2009)

Principe de l'algorithme diff-anonymat :

Entrée : Ensemble de données à partir de n'importe quelle taille de données.

Sortie : Ensemble de données avec modèles de confidentialité (k-anonyme - différentiel).

Etape 1 : Télécharger les données dans le framework.

Etape 2 : Sélectionner les champs d'attributs à disposer dans de nouvelles tables temporaires.

Etape 3 : Détecter le quasi identifiant dans les tables temporaires.

Étape 4 : Diviser les tables en mini tables.

Etape 5 : Appliquer k-anonymat aux minis tables temporaires.

Étape 6 : Détecter les attributs égaux dans les résultats.

Étape 7 : Diffuser les résultats de l'application de k-anonymat.

Etape 8 : Ajouter du bruit aux données qui ont déjà des attributs égaux dans les résultats.

Étape 9 : Recombiner les résultats en un grand ensemble de données.

Les avantages de cette proposition sont une garantie accrue par la confidentialité différentielle et la divulgation limitée de l'identité des personnes par la méthode K-anonymat. La combinaison de ces deux méthodes pourrait contribuer à assurer l'anonymat des données tout en garantissant l'équilibre entre l'ambiguïté des données privées et la clarté des données générales.

3. Inconvénients :

- L'obstacle majeur est que la protection différentielle de la vie privée ne donne pas d'assurance quant au couplage des ensembles de données et aux attributs des données.

- La méthode diff-anonymat est valide pour une quantité élevée mais limitée de données.

II.8 APPARIEMENT CRYPTOGRAPHIQUE :

1. Problème :

De plus en plus les données ne cessent pas d'augmenter et les espaces de stockage traditionnels ne sont plus efficaces pour de nombreuses organisations dans des différents domaines. Pour un accès efficace aux données, ces derniers doivent être sauvegardé dans un grand espace de stockage qu'on appelle BigData, mais malgré sa popularité, ce dernier est confronté à de nombreuses problème de sécurité parmi eux : questions juridiques et politiques, protection des données, protection de la vie privée, manque de transparence, problèmes de cyber sécurité, absence de normes de sécurité et de licences logicielles (Das Sargita, 2015). Alors comment peut-on sécuriser les données privés au niveau du BigData?

2. Contribution et implémentation :

Afin de sécuriser les données privées au niveau du BigData, une technique de leurre a été mise au point ou des fichiers de leurre seront appelés quand un attaquant voulant accéder au système est détecté. Lorsqu'un attaquant accède au nuage, un leurre lui est retourné afin que les données de l'utilisateur réel soient sécurisées.

Les documents leurres sont utilisés dans deux cas :

- Pour vérifier si l'accès aux données est autorisé ou non lorsqu'un accès anormal aux informations est détecté.
- Pour confondre l'attaquant en fournissant de faux documents.

En utilisant des installations informatiques de brouillard qui est une technique d'illusion et la technique de leurre, une fausse BigData est créé. Cela permet à l'attaquant de croire qu'il a accédé aux vraies données de l'utilisateur alors qu'il s'agit en réalité d'une galerie de leurres. Dans le système proposé dans la figure 2.4 les utilisateurs autorisés et non autorisés seront référés à la fausse BigData en tant que première étape. Les utilisateurs légitimes seront par la suite référés à la vraie BigData après avoir été vérifié en réussi le défi de sécurité.

Les deux BigDatas doivent communiquer entre eux en cas de mise à jour. Pour que le vrai système informe le faux de l'ajout d'une donnée par exemple. Cette communication doit être sécurisé c'est pour ça qu'un protocole a été mis au point. Ce dernier est basé sur la cryptographie

de couplage bilinèaire, qui permet de générer une clé de session parmi les participants. C'est ce qui permet d'accéder et stocker les données au niveau du BigData de manière sécurisée.

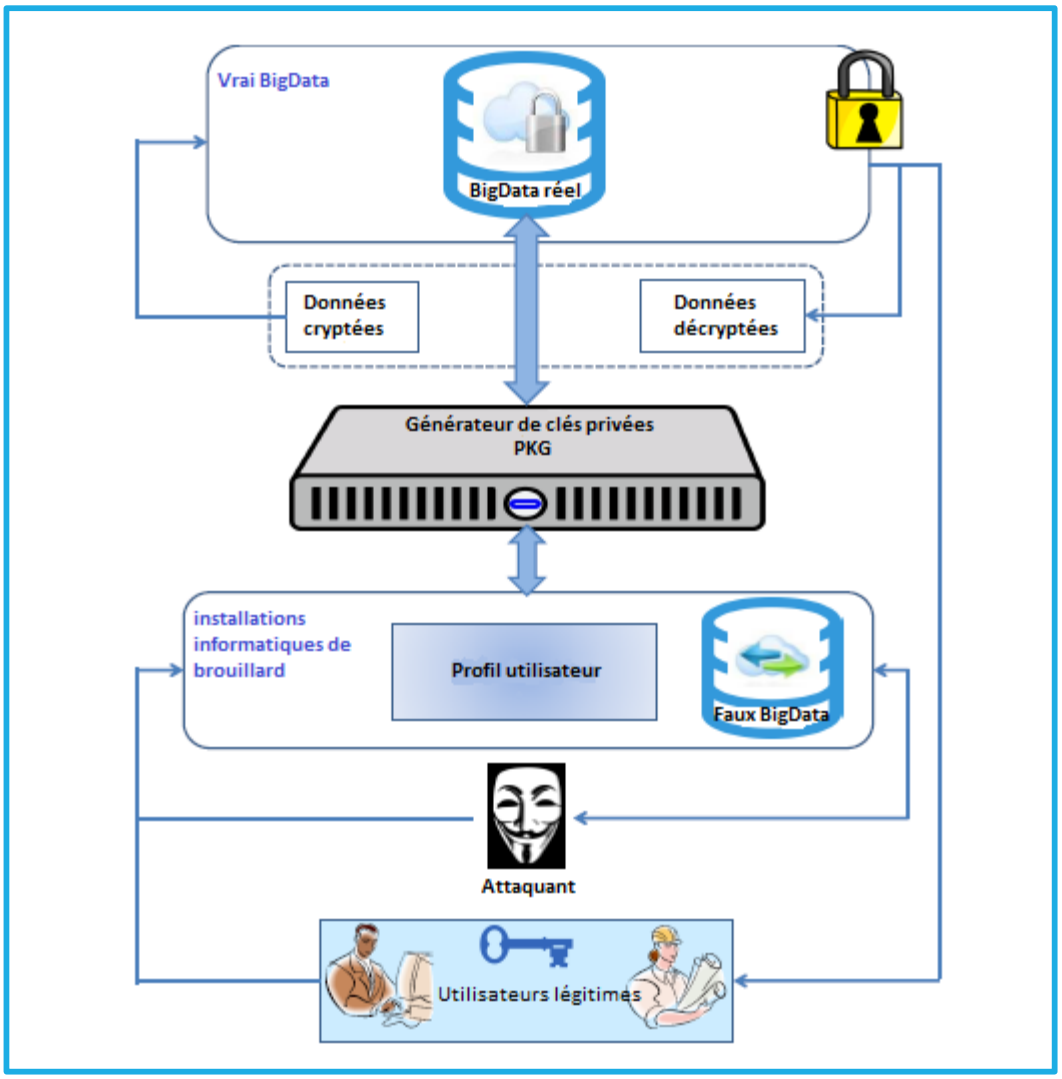

Figure 2.4: Architecture du système de communication (Hadeal Abdulaziz, 2016)

3. Inconvénients :

- En cas ou l'attaquant arrive à compromettre le faux système, ce dernier pourrait lui permettre d'accéder au vrai système et donc aux vrai données de l'utilisateur.

- l'attaquant pourrait interférer dans la communication entre les différentes parties et échanger les clés de sessions qui seront transmises. Ce qui implique une connaissance totale des informations échangées entre les parties.

- L'utilisation d'un défi afin de s'assurer de la légitimité de l'utilisateur reste un moyen peu sécurisé. Si le défi utilisé est un défi non complexe, l'attaquant peut exploiter cette vulnérabilité afin de se faire passer pour le vrai utilisateur.

II.9 PRESERVATION DE LA VIE PRIVEE DANS LE CLOUD :

1. Problème :

Les préoccupations relatives à la protection de la vie privée sont aggravées par le fait que l'information sensible dispersée dans divers ensembles de données peut être récupérée avec plus de facilité lorsque les données et la puissance de calcul sont considérablement abondantes. Bien que certaines questions relatives à la protection de la vie privée ne soient pas nouvelles, leur importance est amplifiée par le cloud.

Avec l'adoption généralisée des services en ligne dans le cloud et des services la prolifération des appareils mobiles, les préoccupations en matière de protection de la vie privée au sujet du traitement et du partage des renseignements personnels de nature délicate augmentent.

2. Contribution et implémentation :

Pour répondre à la problématique citée précédemment, un système de protection de vie privée exige quatre fonctionnalités importantes. La première est construite sur le dessus de MapReduce, et fonctionne comme un filtre pour préserver la confidentialité des ensembles de données avant que ces ensembles de données ne soient accédés et traités par MapReduce. La deuxième détermine les exigences de confidentialité spécifiques, le framework lance les algorithmes d'anonymisation de la version MapReduce pour anonymiser efficacement les ensembles de données pour les tâches MapReduce suivantes. Les ensembles de données anonymes sont conservés et réutilisés pour éviter les coûts de recalcule. La troisième constitue à gérer également la mise à jour dynamique des ensembles de données afin de préserver la confidentialité des données. Enfin, le cadre intègre également des techniques de chiffrement pour assurer de façon rentable la confidentialité de données multiples qui sont rendus anonymes de façon indépendante en fonction des différentes exigences en matière de protection de la vie privée (Zhang K, 2011).

Pour répondre aux quatre exigences du système, quatre modules ont été conçus pour le cadre de protection de la vie privée, soit :

- l'interface de spécification de confidentialité (ISP).
- l'anonymisation des données (AD).
- la mise à jour des données (DU).
- la gestion anonyme des données (ADM).

Enfin, un système prototype correspondant est développé à partir de notre environnement de cloud computing pour mettre en œuvre le framework avec les quatre fonctionnalités décrites ci-dessus.

Le schéma suivant représente la structure du système de préservation de vie privée :

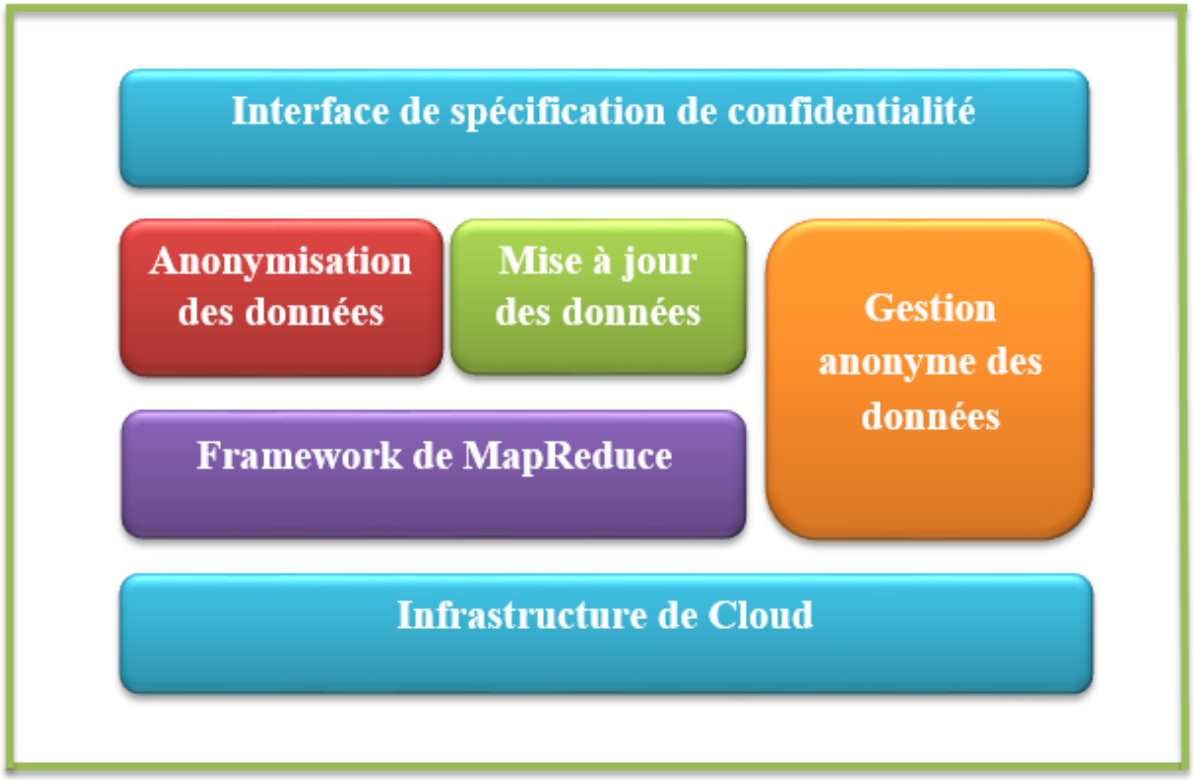

Figure 2.5: Structure du système de préservation de vie privée

3. Inconvénients :

- La protection de la vie privée des ensembles de données à grande échelle doit encore faire l'objet d'une enquête approfondie.

- La préservation de la confidentialité et l'utilité des données volumineuses dans le cadre MapReduce sur le cloud pour les applications d'exploitation ou d'analyse reste un défi.

II.10 TABLEAU COMPARATIF :

	Volume	*Vari-été*	*Vitesse*	*Véracité*	*Disponi-bilité*	*CA*	*Confidentialité*	*Anonymisation*	*Protection de vie privée*
Stockage multi-partagé [2007]	x	c	-	c	-	-	c	c	Moyen
Anonymisation par proximité avec MapReduce [2009]	x	c	c	c	c	-	c	Par proximité	Moyen
Anonymisation multi dimensionnels [2014]	c	c	x	x	-	-	c	Généralisation	Moyen
Protection par détection de compression [2015]	c	c	c	c	-	c	c	c	élevée
Protection par enregistrement local [2015]	-	c	x	-	-	c	-	différentielle	élevée
Vie privé différentiel [2009]	x	c	x	x	-	-	c	K-anonymat différentielle	elevée
Appariement Cryptographique[2015]	c	c	-	c	c	c	c	x	Faible
Préservation de la vie privée dans le Cloud[2011]	x	c	c	c	c	-	c	k-anonymity l-diversity	Moyenne

Tableau 2 : Tableau comparatif entre les approches étudiés

II.11 SYNTHESE DES TRAVAUX EXISTANTS :

Après une analyse faite sur le tableau de comparaison ci-dessus ainsi que les points forts et les points faibles de chaque approche présentée, nous réalisons certains des principaux inconvénients :

- La majorité des approches présentées n'assurent pas l'extensibilité lorsqu'il s'agit d'une quantité massive de données ce qui induit à une réduction de performance et réduction de la protection de la vie privée.

- Plus que la moitié des approches ne sont pas conçues pour assurer la disponibilité, ce qui peut causer un temps de latence assez important et une perte d'une partie ou la quasi-totalité des données.

- La plupart des solutions proposées ne prennent pas en considération ou n'assurent pas la vitesse de traitement des BigData, par conséquent atteinte de fiabilité.

- On constate qu'une partie importante des approches étudiées permettent la protection de la vie privée avec un niveau moyen ou faible.

- On peut dire qu'à l'état actuel de la technique, aucune approche de conception ne garantit à 100% une protection des données sensibles, ce qui peut nuire à la vie privée de l'utilisateur.

Bien que l'approche « Protection par détection de compression » semble être l'approche la plus complète qui ait traité la plupart des critères ; Cependant, certaines limitations ont également été identifiées. C'est pourquoi nous avons l'intention de développer un système de protection de vie privée optimisé.

II.12 CONCLUSION :

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

CHAPITRE III :
L'architecture proposée

III.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 effectué un survol sur les divers travaux de chercheurs concernant ce sujet dans le chapitre précédant. On va s’intéresser à présent dans ce chapitre à l’architecture globale de notre système, ainsi qu’à l’architecture détaillée de chaque composant.

III.2 **ARCHITECTURE GLOBALE :**

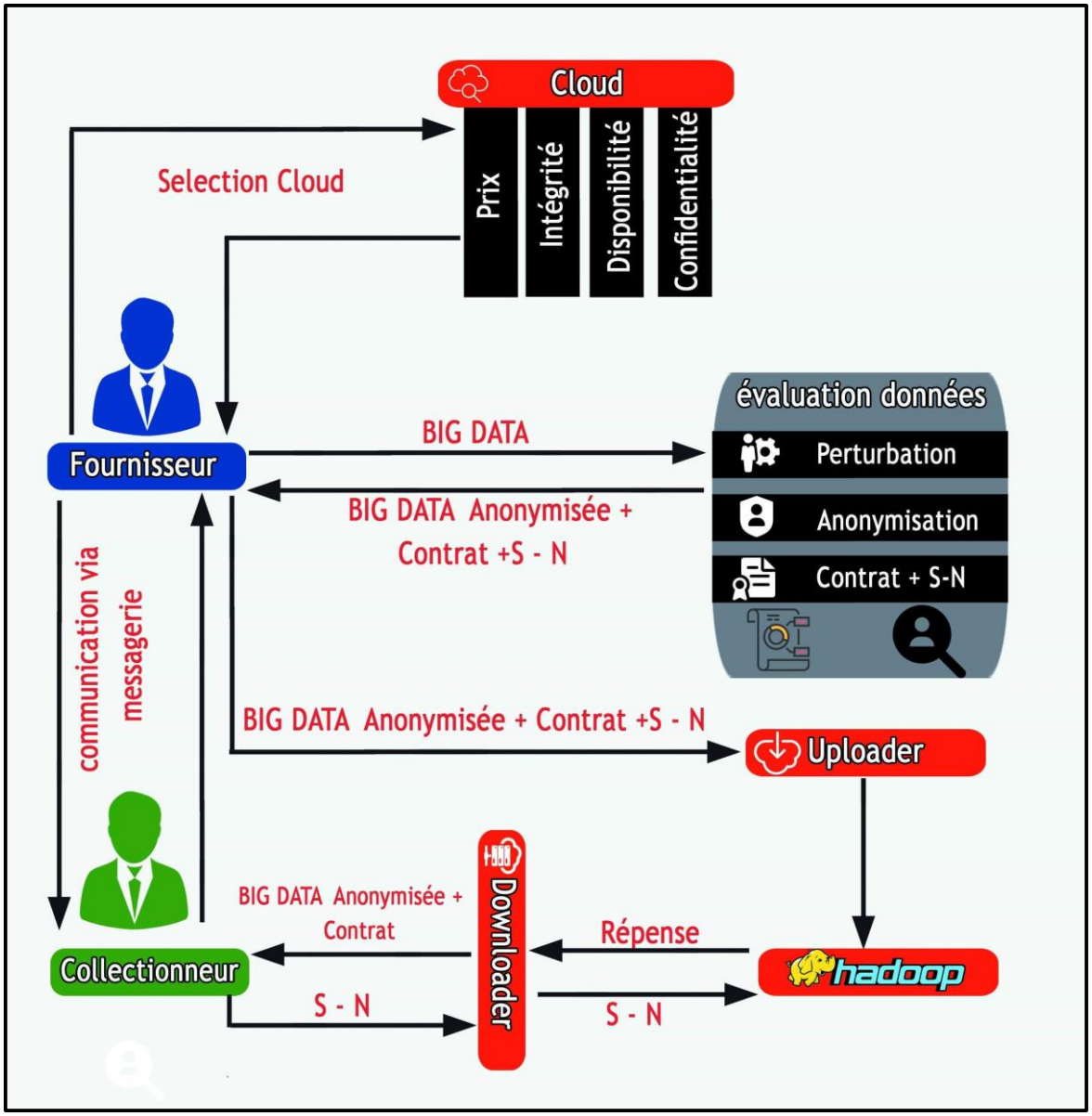

Figure 3.1: l'architecture proposé en Général

- **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. Tous ces composants travaillent en collaboration pour assurer la protection de la vie privée tout en gardant un niveau suffisant d’utilité des données.

III.3 ARCHITECTURE DETAILLEE :

III.3.1 Le composant fiabilité du cloud :

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

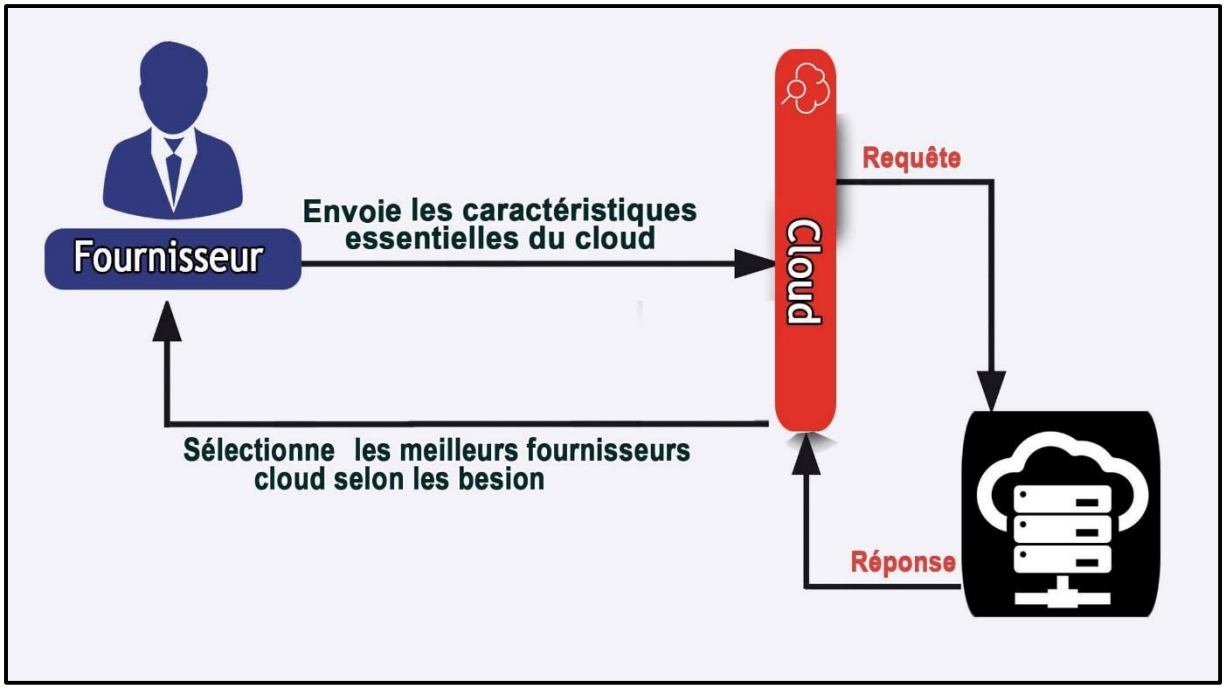

Figure 3.2 l'architecture du composant fiabilité du cloud

- **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

III.3.2 **Le composant évaluation des données :**

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

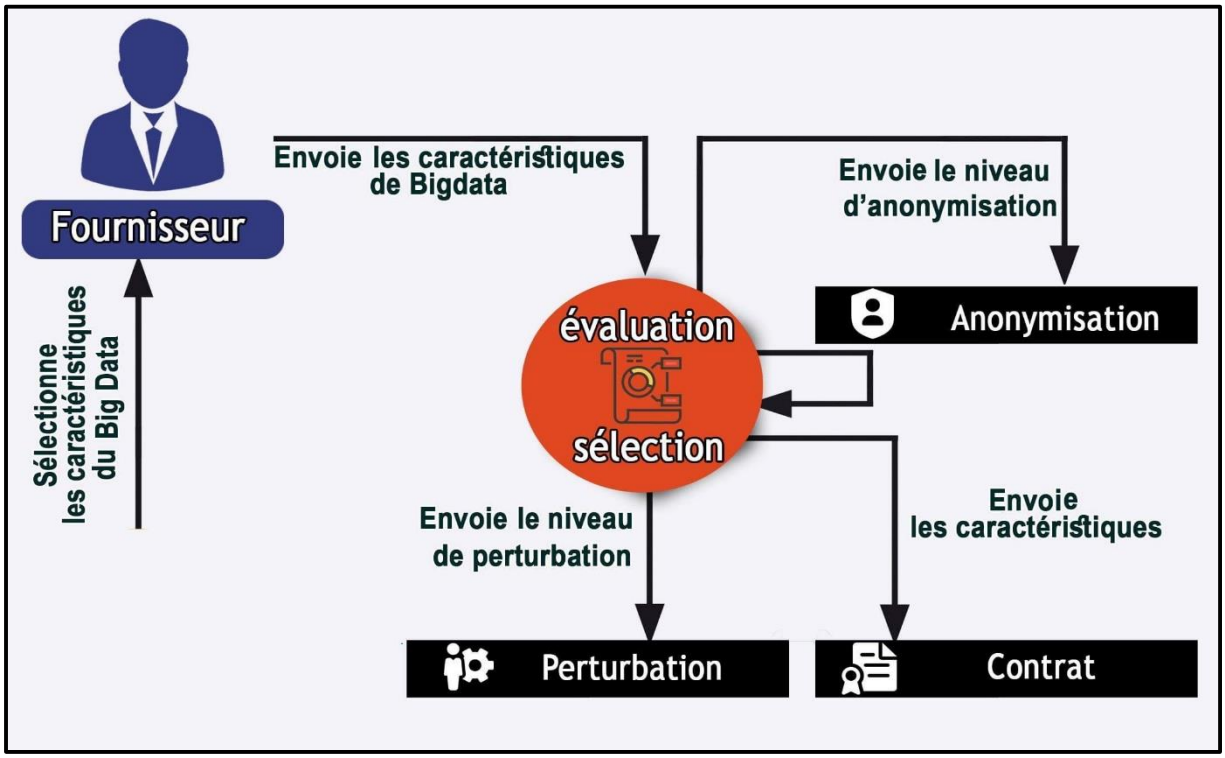

Figure 3.3: L’architecture du composant évaluation des données

- **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 à la composante perturbation.

✓ L’envoi du niveau d’anonymisation des données au composant anonymisation.

✓ L’envoi des caractéristiques principales au composant contrat.

III.3.3 Le composant d’anonymisation des données :

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

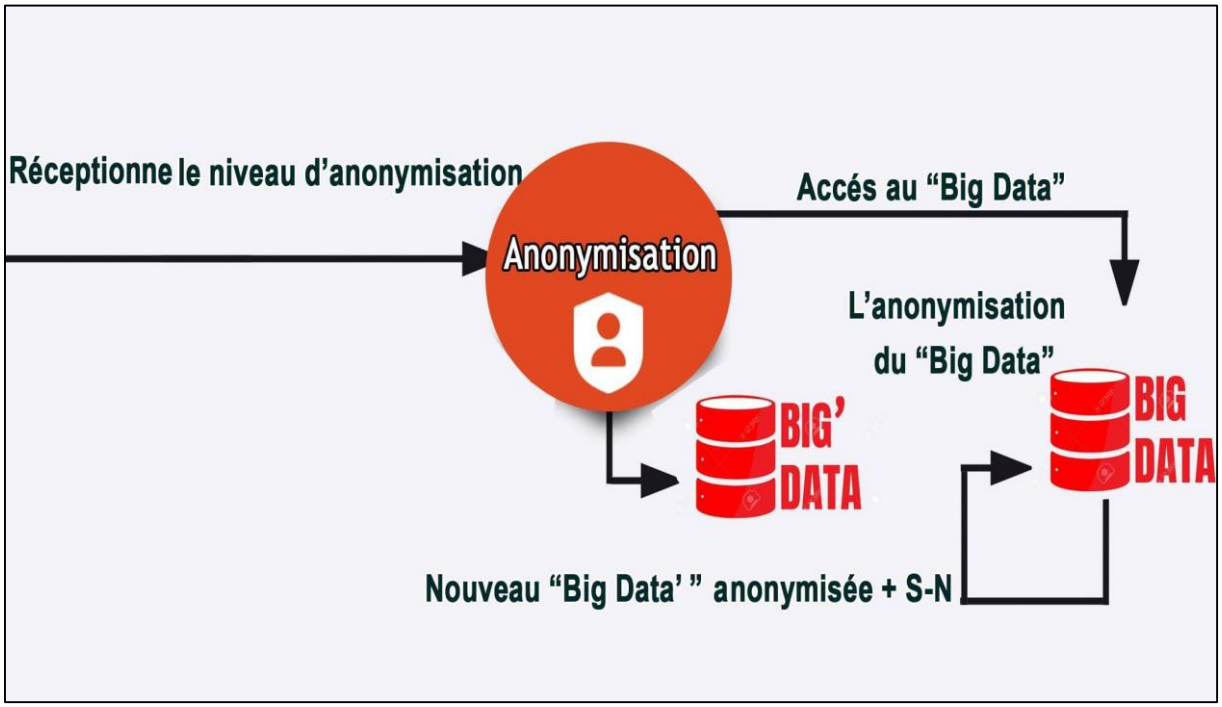

Figure 3.4: L’architecture du composant d’anonymisation des données

- **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.

III.3.4 **Le composant perturbation :**

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

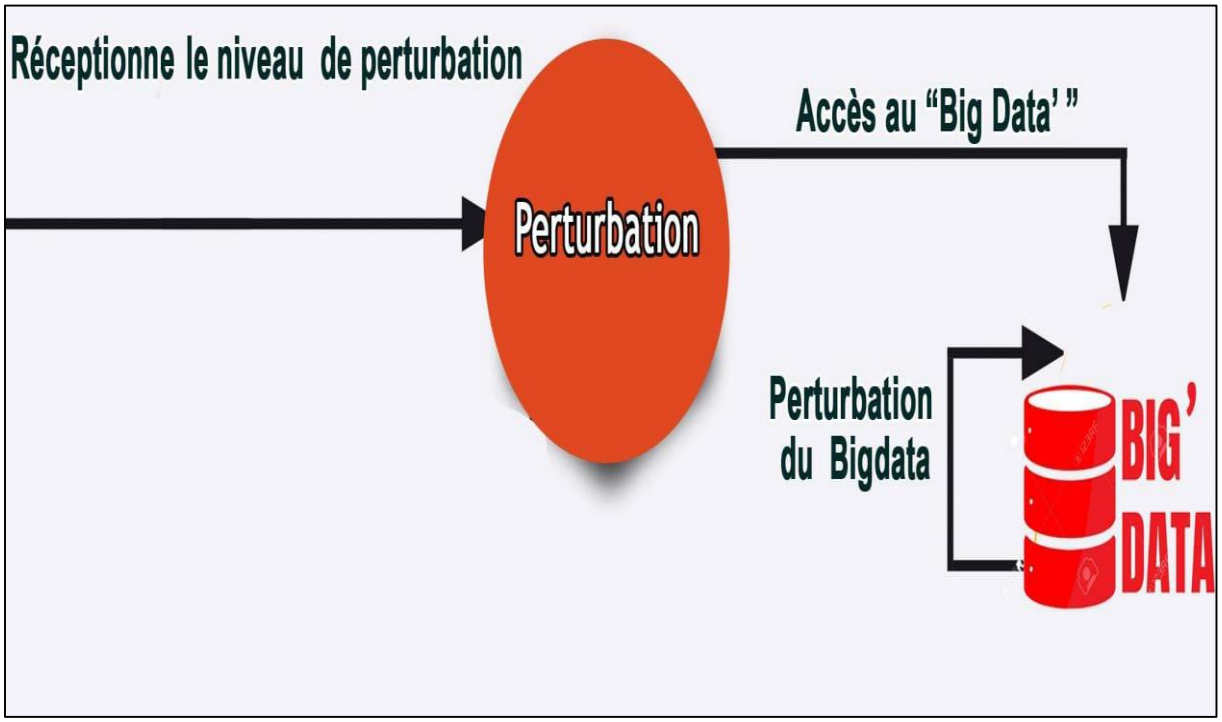

Figure 3.5: L’architecture du composant perturbation

- **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.

III.3.5 **Le composant contrat :**

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

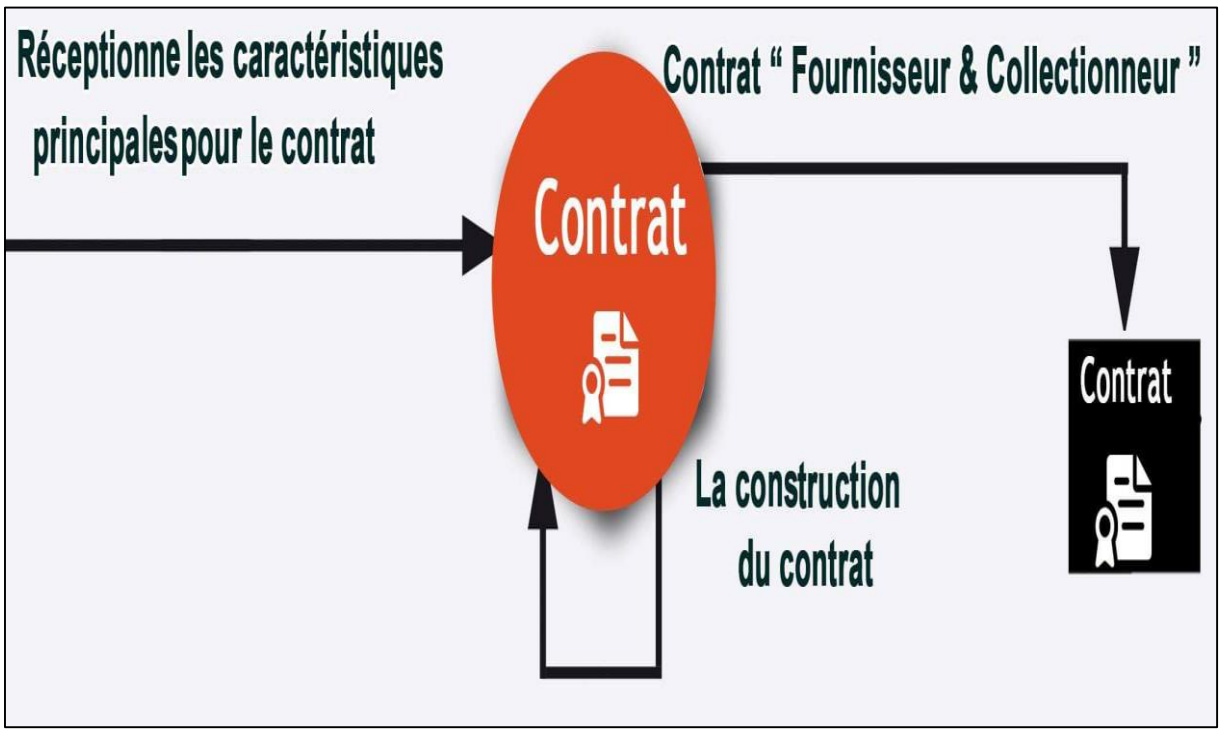

Figure 3.6: L’architecture de composant contrat

- **Le rôle du composant contrat :**

✓ La réception des caractéristiques principales.
✓ La construction du contrat.

III.3.6 Les composants Uploader et Downloader :

- **L’architecture des : Uploader et Downloader :**

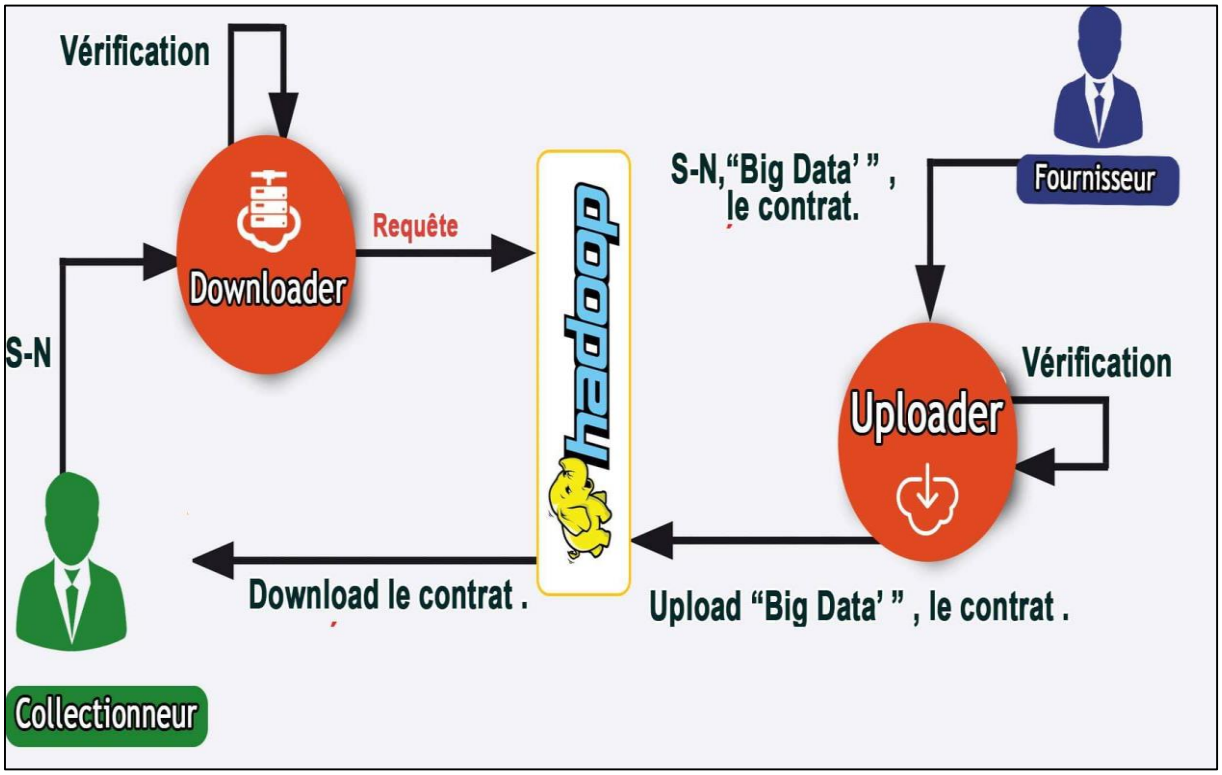

Figure 3.7: L’architecture des : Uploader et Downloader

- **Le rôle des composants Uploader et 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.

III.4 **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 IV :
Implémentation du système

IV.1 **INTRODUCTION :**

Dans ce chapitre nous allons décrire les détails d'implémentation du système, Nous commençons par la présentation des langages de programmation et les outils de développement utilisés pour la mise en œuvre du système conçu dans la section précédente, Nous donnons par la suite les principaux codes source.

IV.2 **OUTILS ET LANGAGES DE PROGRAMMATION UTILISES :**

Pour la réalisation du l'architecture proposée, nous avons utilisé un ensemble de langages de programmation, et quelques environnements de développement. Nous les décrivons brièvement dans les sous sections suivantes

IV.2.1 **Environnement de développement :**

On a utilisé le système d'exploitation Windows 10 avec les caractéristiques est décrites dans la figure ci-dessous :

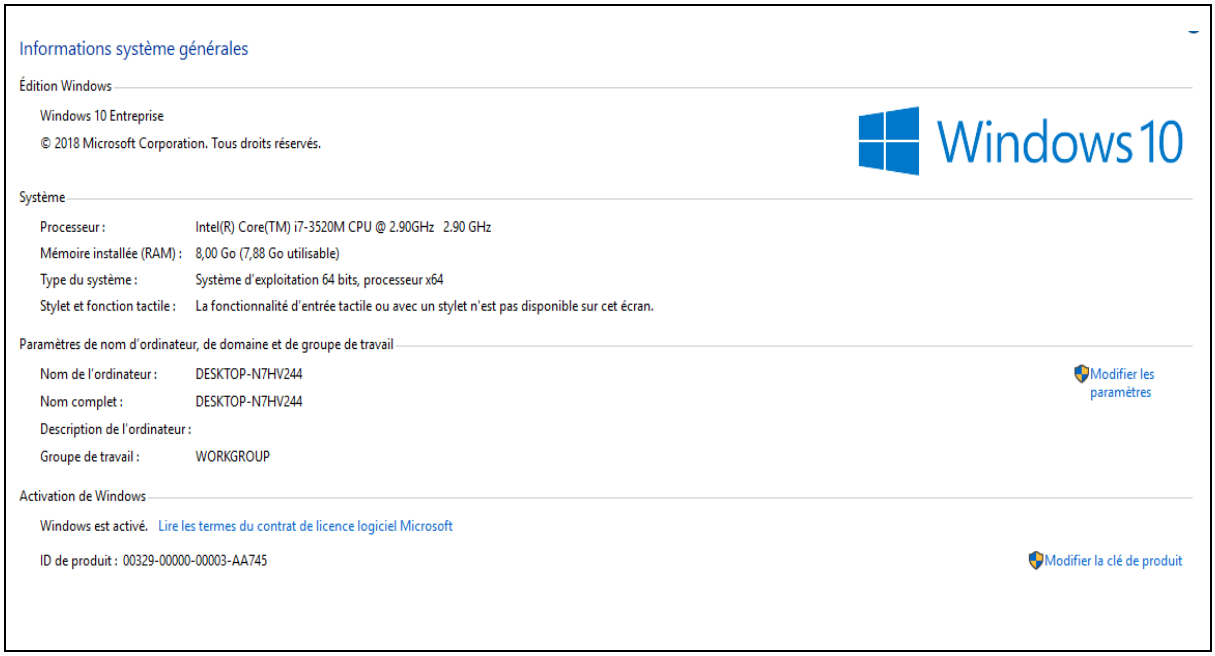

Figure 4.1: environnement de travail et logiciel

IV2.2 **Langages et outils de programmation utilisés :**

Langage Java : est un langage de programmation informatique orienté objet créé 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. La particularité et l'objectif central de Java est que les logiciels écrits dans ce langage doivent être très facilement portables sur plusieurs systèmes d'exploitation tels que UNIX, Windows, Mac OS ou GNU/Linux, avec peu ou pas de modifications. Pour cela, divers plateformes et Framework associés visent à guider, sinon garantir, cette portabilité des applications développées en Java. (Oracle, 2018)

Netbeans IDE : est un environnement de développement intégré (EDI), placé en open source par Sun en juin 2000 sous licence CDDL (Common Devlopement and Distribution License) et GPLv2. En plus de Java, NetBeans permet également de supporter différents autres langages, comme C, C++, JavaScript, XML, Groovy, PHP et HTML de façon native ainsi que bien d'autres (comme Python ou Ruby) par l'ajout de greffons. Il comprend toutes les caractéristiques d'un IDE moderne (éditeur en couleur, projets multi-langage, refactoring, éditeur graphique d'interfaces et de pages Web). (IDE, 2019)

MySQL : est un système de gestion de bases de données relationnelles (SGBDR). Il est distribué sous une double licence GPL et propriétaire. Il fait partie des logiciels de gestion de base de données les plus utilisés au monde3, autant par le grand public (applications web principalement) que par des professionnels, en concurrence avec Oracle, Informix et Microsoft SQL ServerArchitecteure de système. (MYSQL, 2019)

phpMyAdmin : est une application Web de gestion pour les systèmes de gestion de base de données MySQL réalisée en PHP et distribuée sous licence GNU GPL. Il s'agit de l'une des plus célèbres interfaces pour gérer une base de données MySQL sur un

serveur PHP. De nombreux hébergeurs, gratuits comme payants, le proposent ce qui évite à l'utilisateur d'avoir à l'installer.

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, 2019)

IV.3 DESCRIPTION DES INTERFACES GRAPHIQUES :

IV.3.1 Interface de connexion et inscription :

Tout d'abord, 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).

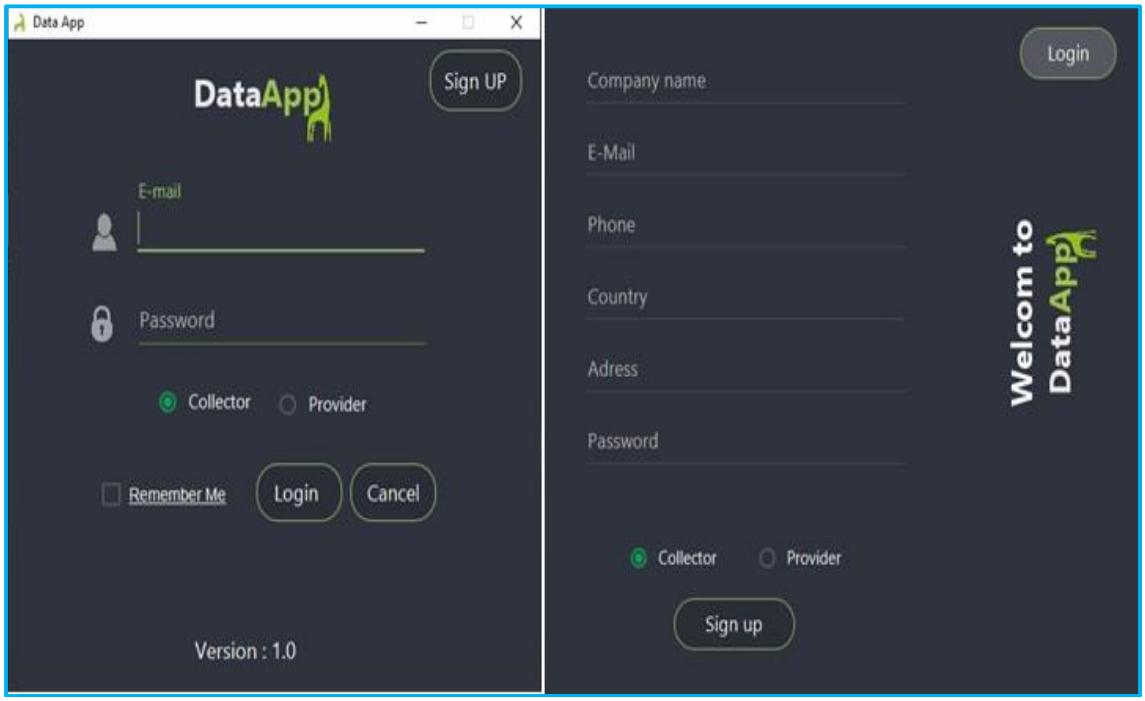

Figure 4.2-Interface de connexion et inscription

IV.3.2 **Interface principale du fournisseur :**

Elle permet à l'utilisateur d'accéder rapidement aux services fournis par cette application figure 4.3

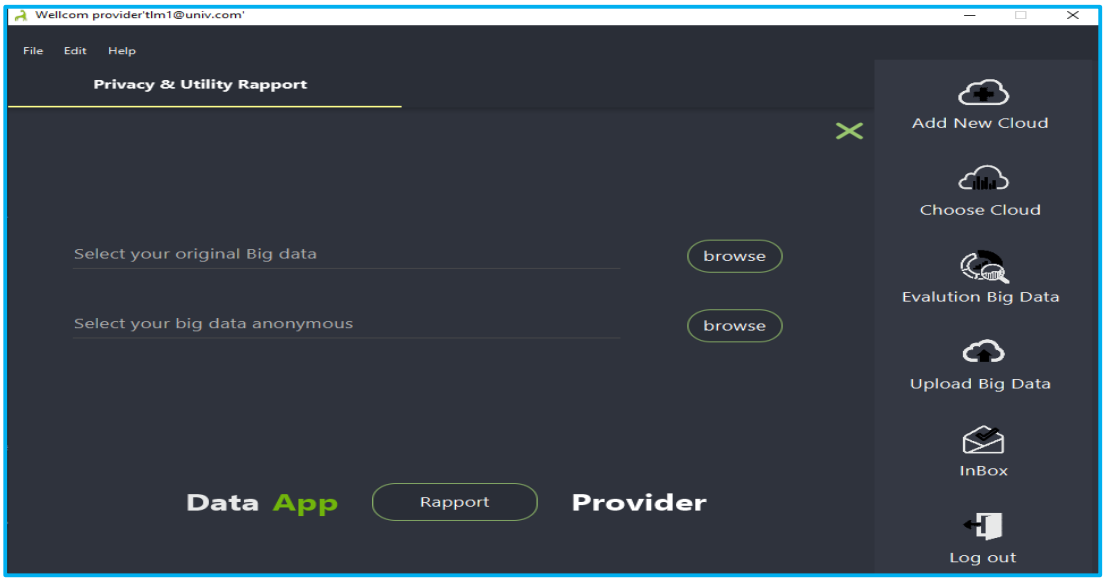

Figure 4.3-Interface Fournisseur

IV3.3 **Service choisir le Cloud :**

L'interface illustrée sur la Figure 4.4, permet de lister les meilleurs fournisseurs Cloud, il est nécessaire d'évaluer les critères de la sécurité et le prix qui convient aux besoins du collectionneur.

Company	Prix	Disponibility	Confidentiality	Integrity
Amazon	604	71	64	82
DELLCloud	731	73	64	50
Alpha7	391	78	81	77
Alibaba Cloud	394	62	64	61
OVH	51	99	99	99
OpenShift	808	73	64	64
sqddsq	247	63	54	46
googledr	500	70	50	30

Figure 4.4-Interface choisir Cloud

IV3.4 **Service d'évaluation du Big Data :**

La figure 4.5, illustre l'interface qui permet au fournisseur de sélectionner le fichier Big Data, son type, son prix et son niveau de sensibilité.

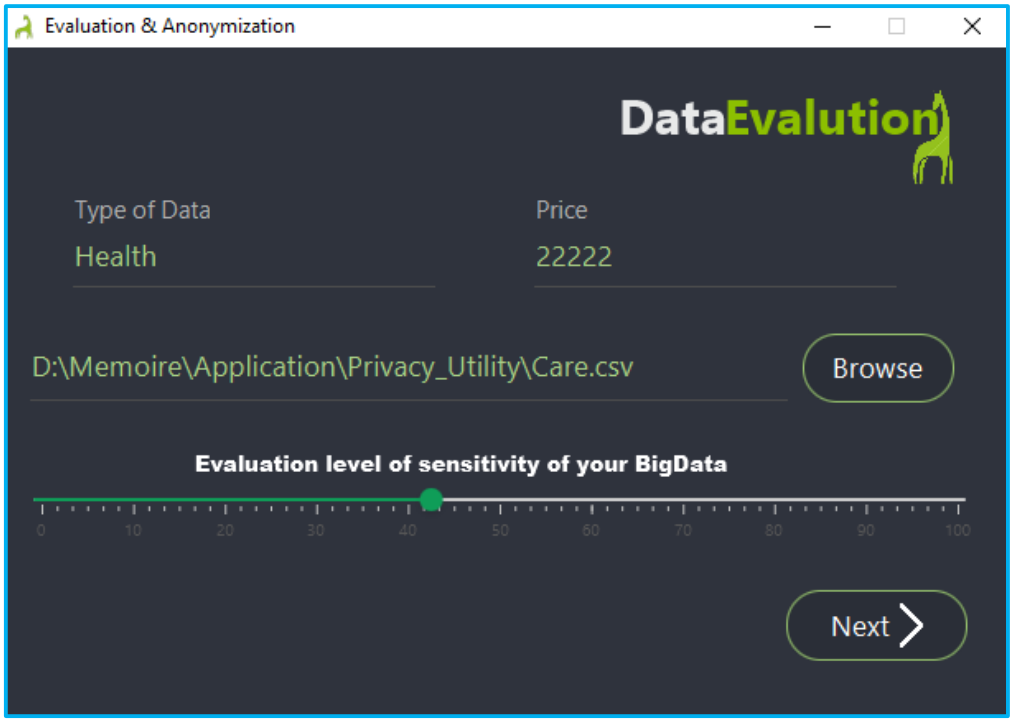

Figure 4.5-Interface évaluation des donnés

La deuxième interface 4.6, le fournisseur permet de spécifier quelles sont les attributs identifiant, quasi-identifiant et les attributs insensibles.

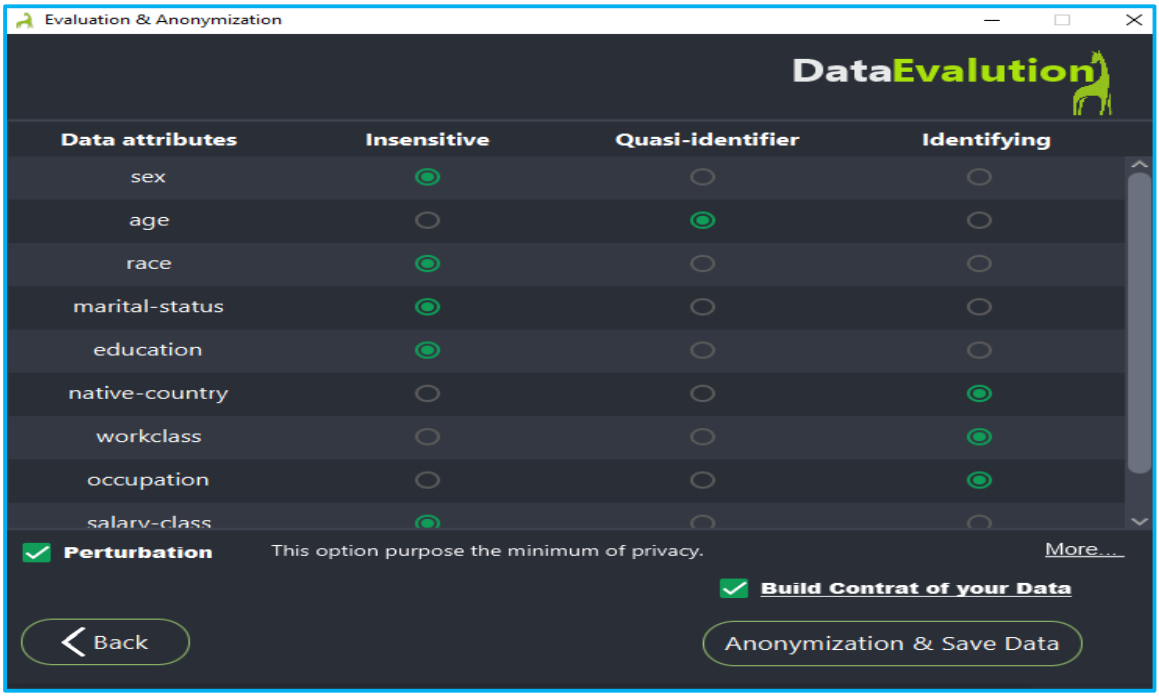

Figure 4.6-Interface évaluation des données'

IV.3.5 **Service Upload Big Data :**

La figure 4.7 présente l'interface UploadData, 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 où il est disponible. Tous ces fichiers sont transférés au HDFS.

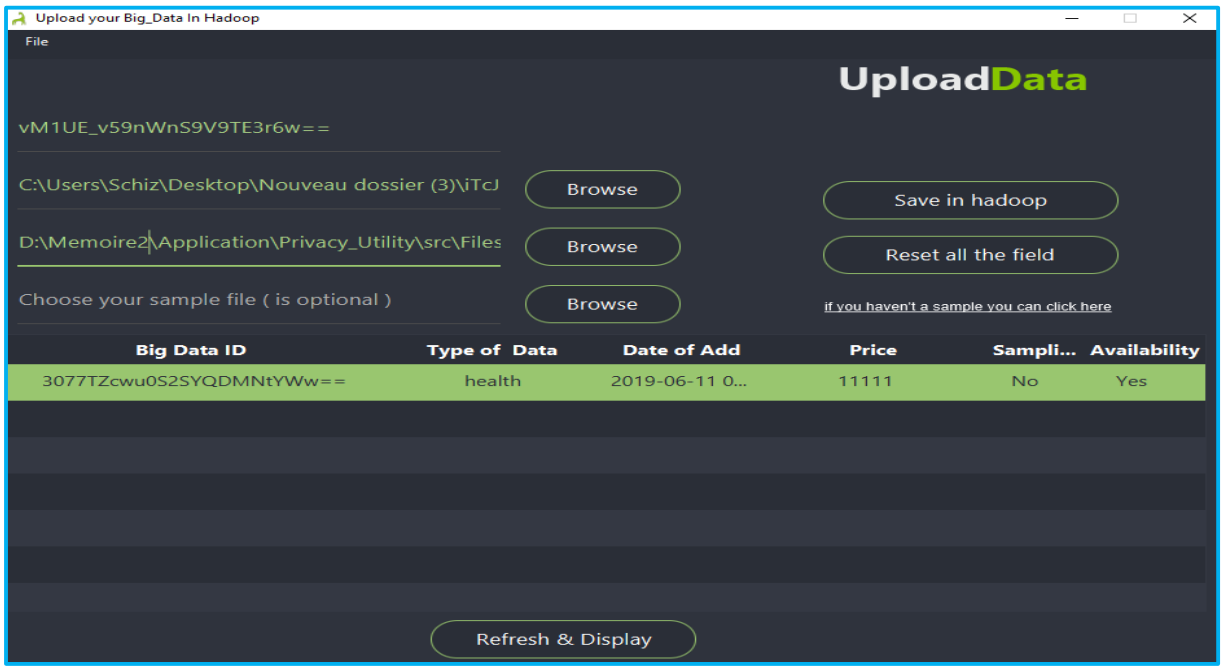

Figure 4.7-Interface Upload Big Data

IV.3.6 **Interface principale du collectionneur :**

L'interface principale du client collectionneur est représentée sur la figure 4.8, Elle lui permet d'afficher tous les Big Data disponibles, tous les fournisseurs avec leurs e-mails et elle lui permettent aussi d'accéder rapidement aux services fournis par l'Application.

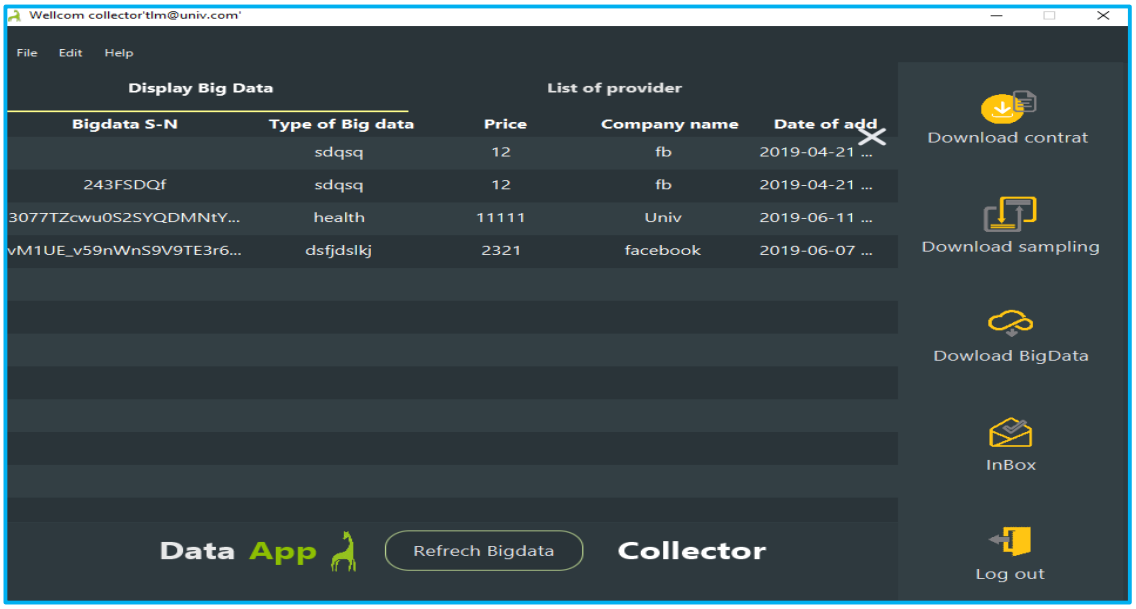

Figure 4.8-Interface Collectionneur

IV.3.7 **Service de téléchargement contrat et échantillon :**

Il suffit d'indiquer le numéro de série pour télécharger l'échantillon et le contrat comme la figure 4.9 montre.

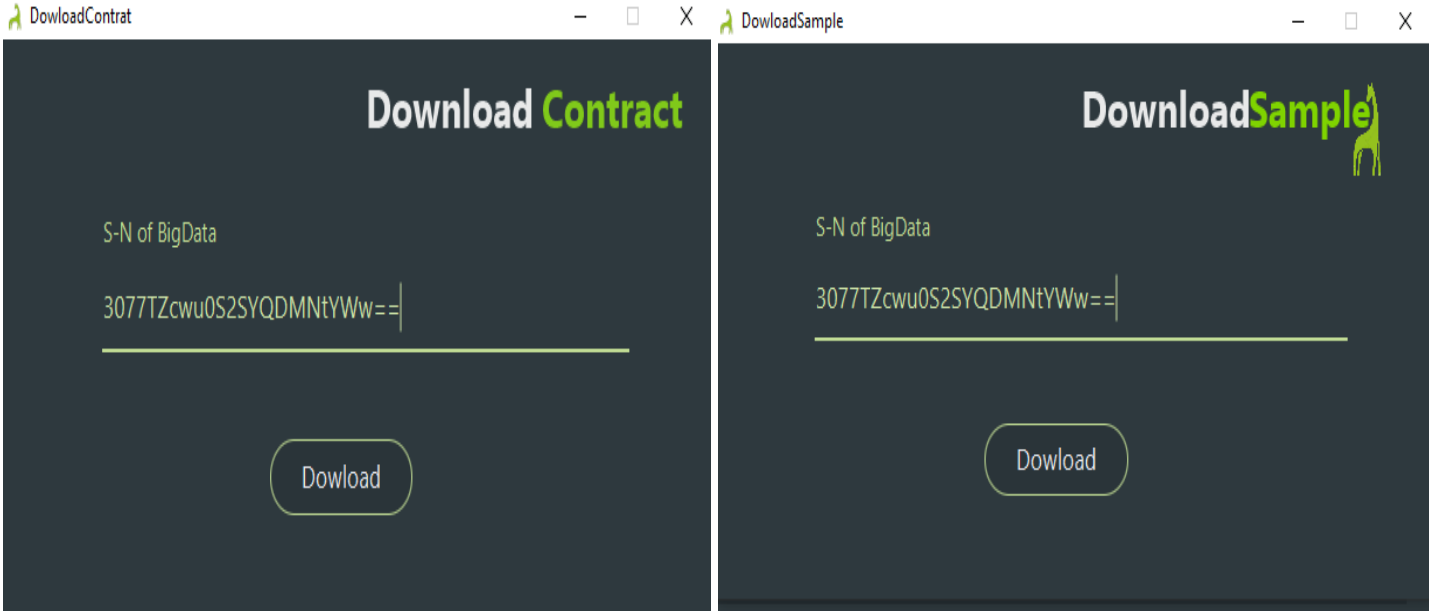

Figure 4.9-Interface de téléchargement contrat et échantillon

IV.3.8 **Service de téléchargement Big Data :**

La figure 4.10 présente l'interface qui permet au collectionneur de télécharger le fichier Big Data. Le nom de Big Data est requise. Pour l'avoir il est nécessaire de contacter le fournisseur par le service de communication, Ou à travers les informations disponibles dans le contrat.

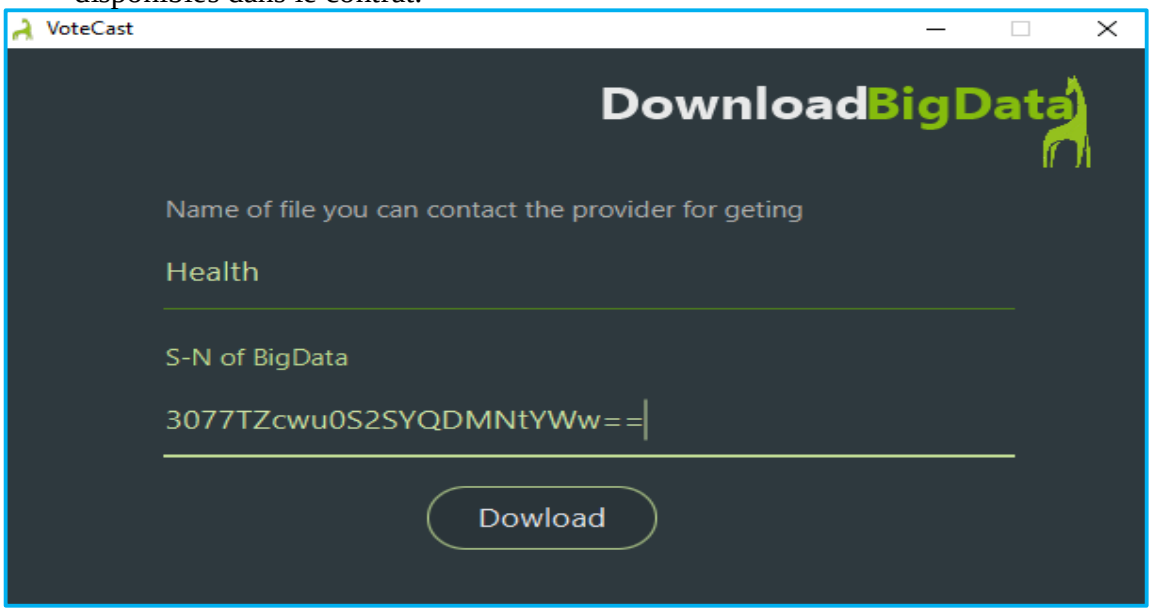

Figure 4.10-Interface télécharger Big Data

IV.3.9 **Service de communication :**

La figure 4.11 montre une interface qui permet au collectionneur d'envoyer et de recevoir des messages afin de faciliter la communication avec les fournisseurs.

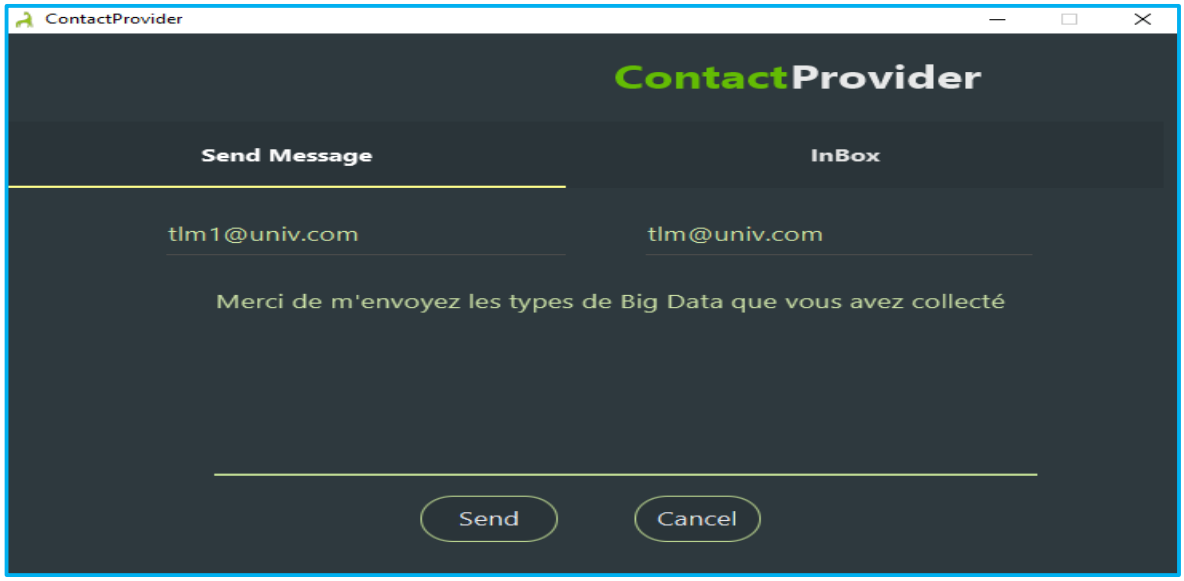

Figure 4.11-Interface contact

IV.4 **INTERFACES DE HADOOP :**

IV.4.1 Hadoop

Après l'installation de Hadoop sur le système d'exploitation, pour assurer le bon fonctionnement cinq fenêtres de CMD s'ouvrent et qui fonctionnement d'une façon permanente chaque une a son rôle.

Les figures [4.12, 4.13,4.14] montrent les principaux fenêtres Hadoop.

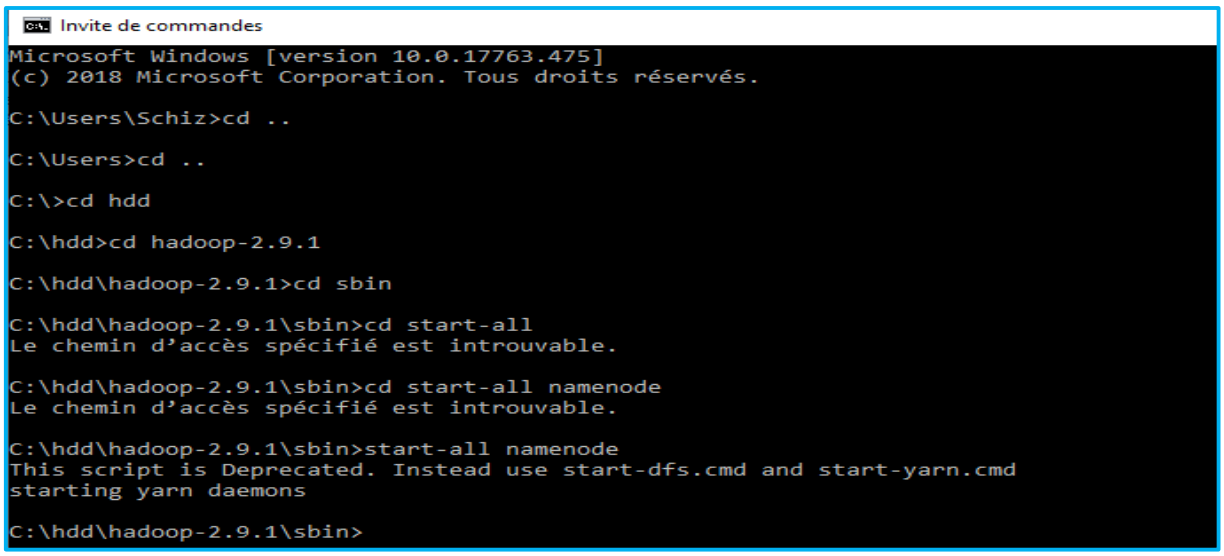

Figure 4.12-Les code de configuration Hadoop

La fenêtre de namenode, ce composant gère l'espace de noms, l'arborescence du système de fichiers et les métadonnées des fichiers et des répertoires.

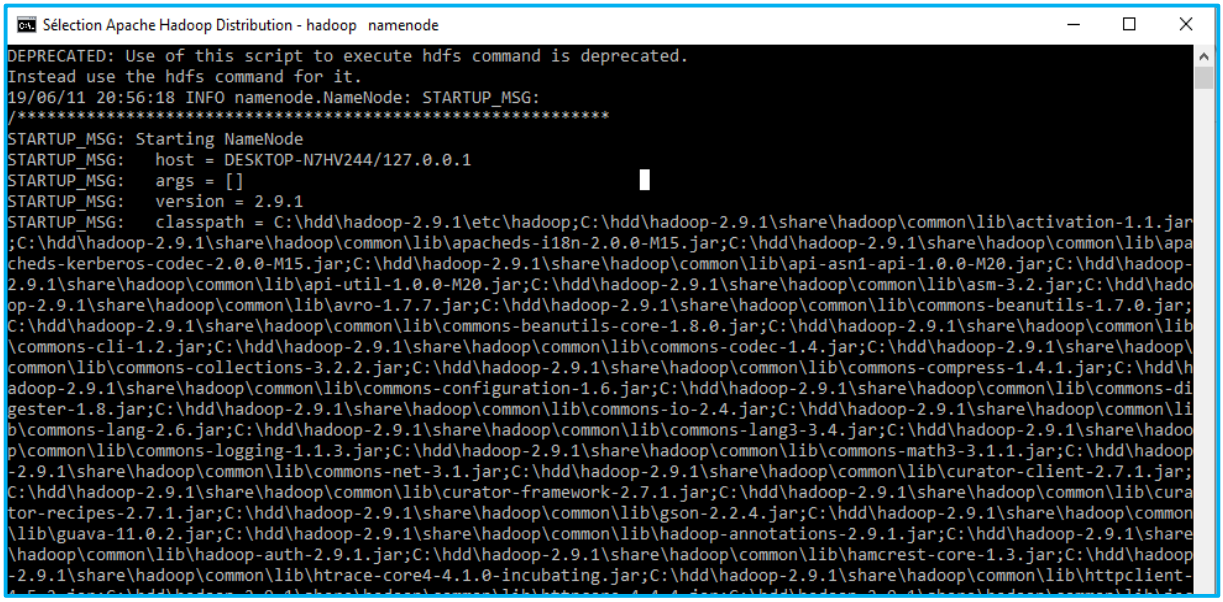

Figure 4.13-La fenêtre namenode

La fenêtre de datanode, ce composant stocke et restitue les blocs de données.

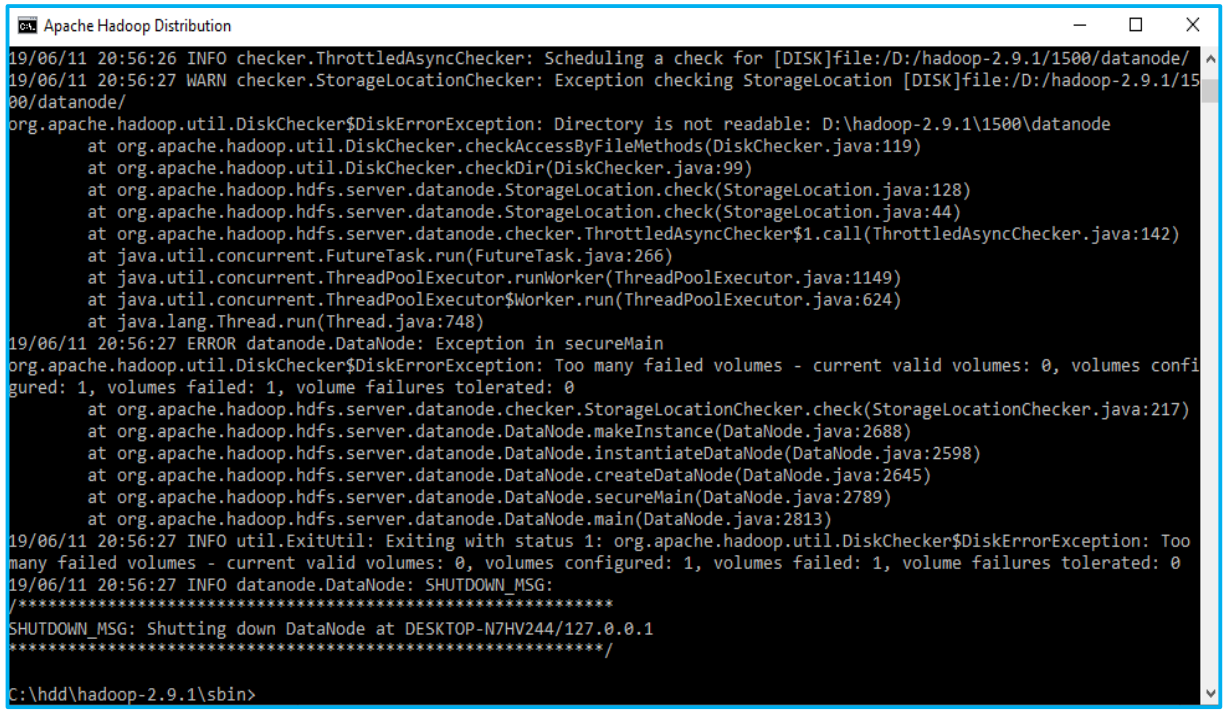

Figure 4.14-La fenêtre datanode

Voici la plateforme de Hadoop installée sur notre système d'exploitation, l'accès à cette plateforme est sur Local host avec le numéro de port : 50070

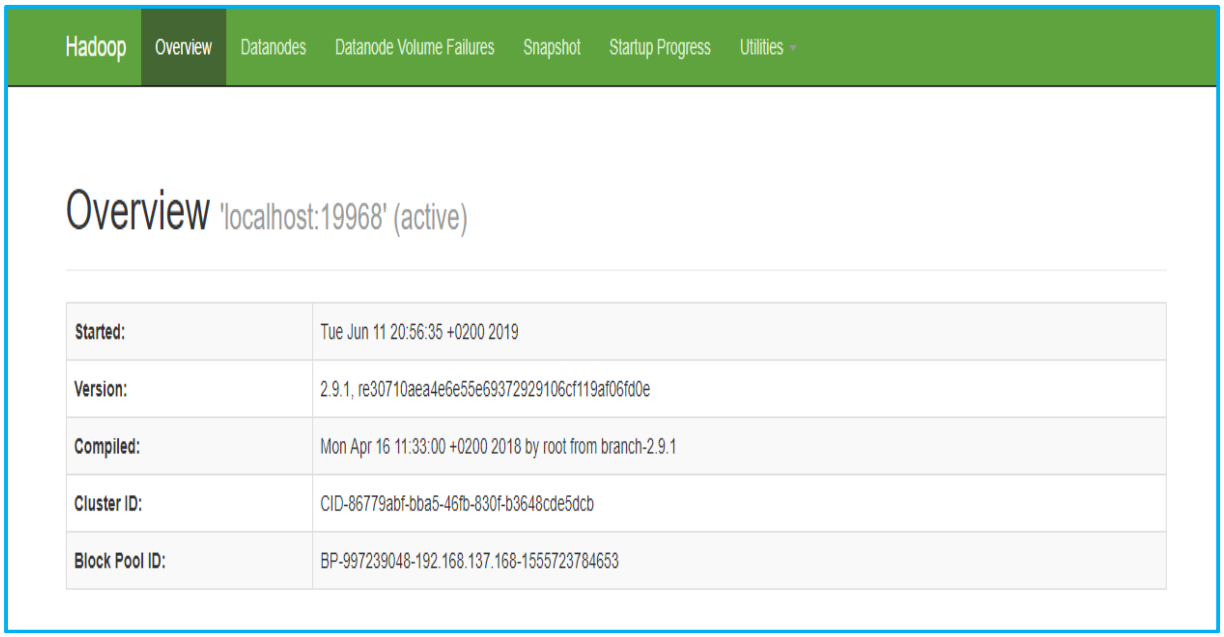

Figure 4.15-Vue générale sur Hadoop

Dans cette section nous trouvons tous les dossiers et les fichiers stockés dans HDFS (Hadoop Distributed File System).

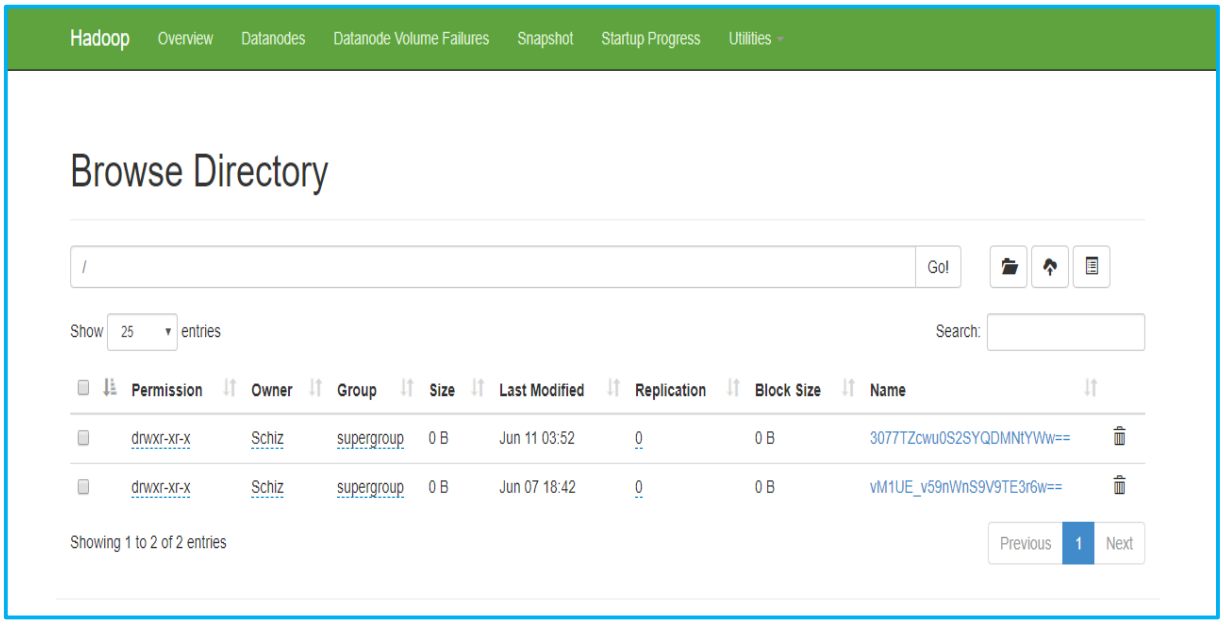

Figure 4.16- Fichiers stockés dans Hadoop

IV.4.2 **Principaux codes sources :**

```
(Collector.isSelected()){
    try {
        try {
            String insert = "INSERT INTO `collector`(`Id_collector`, `Companyc`, `Mailc`, `Phonec`, `C
            connection = handler.getConnection();
            pst=connection.prepareStatement(insert);
        } catch (SQLException e) {
            e.printStackTrace();
        }
        pst.setString(1, null);
        pst.setString(2, Company.getText());
        pst.setString(3, Mail.getText());
        pst.setString(4, Phon.getText());
        pst.setString(5, Countr.getText());
        pst.setString(6, Adres.getText());
        pst.setString(7, Pass.getText());
         boolean ver=virifiersingup();
       if (ver==true) {
            pst.executeUpdate();
            Company.setText("");
            Mail.setText("");
            Phon.setText("");
            Countr.setText("");
            Adres.setText("");
            Pass.setText("");
```

Figure 4.17-Code source d'inscription

```
void loginAction(ActionEvent event) throws IOException {
    if (Collector.isSelected()){

        try {
            String ql = "SELECT * FROM `collector` WHERE `Mailc`= ? AND `Passwordc`=?  ";

           connection = handler.getConnection();

            pst=connection.prepareStatement(ql);
            pst.setString(1, Username.getText());
            pst.setString(2, Password.getText());
            boolean verr=virifierlogin();
            if (verr==true) {
                ResultSet rs = pst.executeQuery();
                int count=0;
                while (rs.next()) {
                    count=count+1;
                }
                if(count+1==1){
                    AlertMaker.showMaterialDialog(stackpan,RootAnchor , new ArrayList<>(), "Try a gain pl

                }else{
        Login.getScene().getWindow().hide();
        Parent parent = FXMLLoader.load(getClass().getResource("/collector/Collector.fxml"));
        Stage stage = new Stage(StageStyle.DECORATED);
        stage.setTitle("Wellcom collector"+"'"+Username.getText()+"'");
```

Figure 4.18-Code source d'inscription

```
    for (String insensss:insensitivedata) {
   daata.getDefinition().setAttributeType(""+insensss, AttributeType.INSENSITIVE_ATTRIBUTE);
        }

  for (String identifis: identifdata) {
     daata.getDefinition().setAttributeType(""+identifis, AttributeType.IDENTIFYING_ATTRIBUTE);
         }

        // Create an instance of the anonymizer
        ARXAnonymizer anonymizer = new ARXAnonymizer();
        // Execute the algorithm
        ARXConfiguration config = ARXConfiguration.create();
        config.addPrivacyModel(new KAnonymity(2));
        config.setSuppressionLimit(0d);
        ARXResult result = anonymizer.anonymize(daata, config);



        // degeré d'anonimisation
        double degre = (insensitivedata.size()+quasidata.size()+identifdata.size())*100;
        double produit = ((quasidata.size() * 50)+(identifdata.size()*100)*100 );
        double degredanony=Math.floor(produit/degre);
        nextback.add(Double.toString(degredanony)+"%");
        //fint degré d'anonymisation
        getnamefile();
        result.getOutput(false).save(liensave+"\\"+namefile+".csv", ';');
```

Figure 4.19-Le code source d'Anonymisation

```
/************************seriel nebre ***********************************************/
   public void getserielnumber() throws FileNotFoundException, NoSuchAlgorithmException, SQLException{
     getinformation();
     //generate seryal number
     SecretKey secretKey = KeyGenerator.getInstance("AES").generateKey();
      // get base64 encoded version of the key
     String encodedKey = Base64.getUrlEncoder().encodeToString(secretKey.getEncoded());

     nextback.add(encodedKey);

         try {
                        try (Writer writer = new BufferedWriter(new OutputStreamWriter(


                         new FileOutputStream(liensave+"\\serielnumber.txt"), "utf-8"))) {
                              writer.write(encodedKey);
}


                       } catch (Exception e) {
                       }


   }
```

Figure 4.20-Le code source de génération de S-N

```
            pst.close();
            nameofofbig=null;
            serialnubmring=null;
            lienofsavedata=null;
            Nameofbigdata.clear();
            SNContrat1.clear();

 }else {

  AlertMaker.showMaterialDialog(StackpaneBig,AnchopaneBig , new ArrayList<>(), "Serial number or name of

        }
}
 private void dowloadBigdatainyourpath() throws IOException {
       File file = dc.showDialog(null);
     //int total =s+ins+id+quas;
      if (file != null) {

      lienofsavedata=file.getPath();
          System.out.println("lien"+lienofsavedata);
       String localpath="hdfs://localhost:19968/";
       String filepath="hdfs://localhost:19968/"+serialnubmring+"/"+nameofofbig;
        Configuration conf= new Configuration();
        FileSystem fs =FileSystem.get(URI.create(localpath), conf);
        fs.copyToLocalFile(new Path(filepath), new Path(lienofsavedata));
```

Figure 4.21-Code source de téléchargement des Big Data

IV.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, nous avons 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.

Conclusion Générale et perspectives

Dans les derniers années le Big data est devenu un opportunité pour les entreprises pour créer un avantage concurrentielle et gagner plus de profits, en même temps il facilite la vie quotidienne des individus ,grâce à des nouvelle technologie qui permettent de collecter, stocker et analyser les données produites en temps réel et avec un coût moins cher. Ces données collectées permettent aux entreprises de comprendre le comportement et les besoins de leur client , en revanche cette opportunité ne paurra pas être saisie sauf si le respect de la vie privé de l'utilisateur et ne pas jouer par ces données personnelles. Pour cela il est nécessaire d'anonymisé ces données partiellement pour éviter la manipulation et les fraudes sur internet .

Dans ce travail on a essyé de résoudre la problématique de la protection de la vie privée avec la conservation de l'utilité des big data , pour cela on a étudiér d'abord, les approches et les travaux connexes et on a essayé de comprendre les points faibles de chaque approche, ensuit, on a inspiré notre nouvelle architecture proposé afin de trouver une solution intégrée.

Finalement notre perspectives et d'intégrer l'intelligence artificiel et l'apprentissage automatique ''machine Learning'' a notre application, et de lancer la version commerciale.

Bibliographie

Andreas Pfitzmann, M. H. (2010). A terminology for talking about privacy by data minimization:.

B. Fung, K. W. (2010). Privacy-preserving data publishing: A survey of recent developments. *ACM Computing Surveys*, 14.

B.C.M. Fung, K. W. (2010). Privacy-Preserving Data Publishing: A Survey of Recent Developments. *ACM Computer Surveyv* , 1-53.

Behmo, R. (2019, 05 06). *Concevez des architectures Big Data*. Récupéré sur OpenClassrooms: https://openclassrooms.com/fr/courses/4467491-concevez-des-architectures-big-data/4896201-restez-a-jour-avec-la-speed-layer

C. Dwork, M. N. (381-390). On the complexity of differentially private data release: efficient algorithms and hardness results.

Chaudhuri, S. (2012, May). What next? : A half-dozen data management research goals for big data and the cloud. *PODS'12*, pp. 1-4.

Das Sargita, C. A. (2015). A Review on Issues and Challenges of Cloud Computing. 81-88.

Dwork, C. a. (2009, may 31). On the complexity of differentially private data release:efficient algorithms and hardness results. *STOC*, pp. 381-390.

Francis, L. P. (2014). Introduction: Technology and New Challenges for Privacy.

Hadeal Abdulaziz, S. M. (2016). A Security Model for Preserving the Privacy of Medical Big Data in a Healthcare Cloud Using a Fog Computing Facility With. *EEE Access*, 5964.

Hadoop. (2019). Récupéré sur Hadoop apache hadoop community: http://hadoop.apache.org/

Hayashi, K. (2013). Social Issues of Big Data and Cloud: Privacy, Confidentiality, and Public Utility. *International Conference on Availability, Reliability and Security*.

I. Stoica, R. M. (2001). A scalable peer-to-peer lookup service for internet applications. *SIGCOMM Comput*, 149-160.

IDE, N. (2019, 06 02). Récupéré sur NetBeans: https://netbeans.org/index_fr.html

J. Gehrke, R. R. (1998). Rainforest-a framework for fast decision tree construction of large datasets. 416–427.

J.Camenisch, S. M. (2014). *Privacy and Identity Management for the Future Internet in the Age of Globalisation.*

Kargupta H, D. S. (2003). On the Privacy Preserving Properties of Random Data Perturbation Techniques[C]. *nstitute of Electrical and Electronics Engineers Inc*, 99-106.

Krishna, P. (2015). Big data search and mining, serie studies in big data. 11-93.

L. Wang, J. Z. (2012). In Cloud, Can Scientific Communities Benefit from the Economies of Scale? *IEEE Transactions on Parallel and Distributed Systems*, 296-303.

Li N, e. a. (2007). t-Closeness: privacy beyond k-anonymity and Ldiversity. *Data engineering (ICDE) IEEE 23rdinternationalconference*.

Lisbeth.R, A. C. (2015). A general perspective of Big Data: application, tools, challenges and trends. new york.

Liu, M. C. (2014). Big Data: A Survey. 175_176.

Machanavajjhala A, G. J.-d. (2006). privacy beyond k-anonymity. *22nd international conference data engineering (ICDE)*, p. 24.

Mehmood A, N. I. (2016). Protection of big data privacy. *IEEE translations and content mining are permitted for academic research*.

Microsoft. (2015). Microsoft differential privacy for everyone.

MYSQL. (2019). Récupéré sur mysql: https://www.mysql.com/fr/why-mysql/presentations/

Nichterlein A, N. R. (2011). The effect of homogeneity on the complexity of k-anonymity. *FCT*, pp. 53-64.

Okamoto, M. M. (1997). Proxy cryptosystems: Delegation of the power to decrypt ciphertexts. *IEICE Transactions*, 54–63.

Oracle. (2018, 06 01). Récupéré sur Go Java: https://go.java/index.html?intcmp=gojava-banner-java-com

P.Derbeko, S. ,. (2015). Security and privacy aspects in MapReduce on clouds: A survey.

Raju, G. T. (2016). An Approach for Privacy Preserving and Multi-Sharing Control using Proxy Re-encryption in Big Data Storage. *ABHIYANTRIKI* , 20-25.

Reinsel, J. G. (2011). Extracting value from chaos. *IDC*, pp. 1-12.

Saouli.H, K. K. (s.d.). Applications et enjeux des Big Data dans le contexte des défis mondiaux. *Laboratoire LINFI*.

Shao, J. (2012). Anonymous id-based proxy re-encryption. 364–375.

Shuyu Li, J. G. (2016). Security and Privacy for Big Data. *International Publishing Switzerland*.

Sithu . Sudarsan, R. .. (2015). Security and Privacy of Big Data. *International Journal of computer application.*

Stubbs, E. (2014). Big Data,Big Innovation.

TEETZ, J. K. (2018). *#GDPR #DATA PROTECTION : CHALLENGES AND OPPORTUNITIES*. Récupéré sur HR OPEN: http://blog.hropenstandards.org/blog/gdpr-data-protection-challenges-and-opportunities

Tzeng, C.-K. C.-G. (2007). dentity-based proxy re-encryption without random oracles. *ICS*, 189-202.

Waters, X. B. (2006). Anonymous hierarchical identity-based encryption (without random oracles. *CRYPTO*, 290–307.

Waters, X. B. (2006). Anonymous hierarchical identity-based encryption (without random oracles. *Springer*, 290–307.

X. Wu, X. Z.-Q. (2014). *Data mining with big data.* IEEE TKDE.

X. Zhang, C. L. (2013). A Privacy Leakage Upper Bound Constraint-Based Approach for CostEffective Privacy Preserving of Intermediate Data Sets in Cloud. 1192-1202.

Xu L, J. C. (2014). Information security in big data. *IEEE*.

Y. Zhang, X. C. (2013). Anonymous attributebased encryption supporting efficient decryption test. *CCS*, 511-516.

Yingjie, W. (2015). Privacy protection data release: model and algorithm. *4inghua University Press*, 3-8.

Zhang K, Z. X. (2011). privacy-aware data intensive computing on hybrid clouds. pp. 515–526.