


République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université AMO de Bouira

Faculté des Sciences et des Sciences Appliquées

Département d'Informatique

Mémoire de Master2

en Informatique

Spécialité : *ISIL+GSI*

Thème

La protection de la vie privée dans le Big Data

Encadré par

— Mr.BOUDJELABA Hakim

Réalisé par

— Mlle.BELKHAMSA Amel

— Mlle.YAHIAOUI Manel

2019/2020

Remerciements

Avant d'entamer ce mémoire de fin d'étude, nous tenons à exprimer notre sincère gratitude envers tous ceux qui nous ont aidé ou ont participé au bon déroulement de ce mémoire.

Tout d'abord, Nous exprimons nos profondes gratitudes et respectueuse reconnaissance à notre encadreur :**Mr. BOUDJELABA Hakim** pour sa bonne volonté d'accepter de nous encadrer, pour tout le temps qu'il nous a octroyé et pour tous les conseils qu'il nous a prodigué.

Aussi que **les membres de jury** trouvent ici nos remerciements les plus vifs pour avoir accepté d'honorer par leur jugement notre travail.

Nos vifs remerciements s'adressent également à **nos enseignants** et à nos amis, pour leur présence chaleureuse et leur encouragement.

Dédicaces

Avant toutes choses je tiens à remercier toutes personnes qui a contribuer de prêt ou de loin à la réalisation de ce travail.

Je dédie ce travail :

A mes chers parents Mes chers parents que nul dédicace ne peut exprimer mes sincères sentiments, je les remercie pour leur patience illimitée, leur encouragement continu ainsi que leur aide précieuse, en témoignage de mon profond amour et respect pour leur grands sacrifices.

A mes chers frères : Smail, karim, Rachid et Nacer pour leur grand amour et leur soutien qu'ils trouvent ici l'expression de ma haute gratitude.

A mes chères Sœurs : Kahina, Nadia et Malika pour leur présence à mes cotés dans les moments les plus difficiles lors de la réalisation de ce travail.

A ma chère binôme : Manel tu es été très collaboratrice et c'est un honneur pour moi d'avoir partagé ce travail avec toi.

Mes chers ami(e)s qui sans leur encouragements ce travail n'aura jamais vu le jour. Et en fin à toute ma famille et à tous ceux que j'aime.

BELKHAMSA Amel

Dédicaces

Je dédie ce travail :

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

A mon cher frère mohamed lamine et ma **chère sœur ;** Amina pour leurs encouragements permanents, et leur soutien moral,

A ma chère tante houda et mes autres tantes pour leur appui et leur encouragements

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

A ma chère binôme Amel, pour son soutien moral, sa patience et sa compréhension tout au long de ce mémoire.

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.

YAHIAOUI Manel.

Résumé

Le terme "big data" est utilisé pour désigner de très grands ensembles de données, plus variée et avec des structures plus complexes. Ses caractéristiques sont généralement liées à d'autres difficultés de stockage, d'analyse et d'application d'autres processus ou d'obtention de résultats.

Avec l'air du big data et la récolte massive de données et le développement des méthodes de machine learning, il est devenu difficile de protéger la vie privée des utilisateurs et de rester anonyme sur le web. L'objectif de notre travail consiste à protéger l'anonymat des utilisateurs dans le Big Data. Dans notre travail nous présentons plusieurs mécanismes de protection de l'anonymat des utilisateurs.

Mots clés : big data, l'anonymat, Vie privée, la généralisation ...

Abstract

The term "big data" is used to refer to very large datasets, more varied and with more complex structures. Its characteristics are usually related to other difficulties in storing, analyzing and applying other processes or obtaining results.

With the air of big data and massive data collection and the development of machine learning methods, it has become difficult to protect user privacy and remain anonymous on the web. The goal of our work is to protect the anonymity of users in the Big Data. In our work we present several mechanisms to protect user anonymity.

Key words : big data, anonymity, privacy, generalization ...

ملخص

يستخدم مصطلح ْالبيانات الضخمةْ للدلالة على مجموعات كبيرة جدًا من البيانات ، أكثر تنوعًا وبها هياكل أكثر تعقيدًا. خصائصه بشكل عام تتعلق بالصعوبات الأخرى في

تخزين وتحليل وتطبيق العمليات الأخرى أو للحصول على النتائج.

مع ظهور البيانات الضخمة والجمع الهائل لها وتطوير أساليب التعلم الآلي ، أصبح من الصعب حماية خصوصية المستخدمين بحيث أنها تظل مجهولة الهوية على الويب. الهدف من عملنا هو حماية سرية هوية المستخدمين في البيانات الضخمة. نقدم في عملنا عدة آليات حماية إخفاء هوية المستخدم.

الكلمات الرئيسية: البيانات الضخمة ، عدم الكشف عن الهوية ، الخصوصية، التعميم . . .

Table des matières

Table des figures

Liste des tableaux

Liste des abréviations

SGBD	système de gestion des base de données
LSST	Large Synoptic Survey Telescope
MOOC	Massive Open Online Courses
ACP	Analyse en Composante Principale
ACM	Analyse des Correspondances Multiples.
SVM	Support Vector Machine
KNN	K-plus proches voisins(k-nearest neighbors).
IoT	Internet of Things.
ABAC	Attribute Based Access Control
RBAC	Role Based Access Control
VPN	Virtual Private Network
TOR	The Onion Router
AMC	analyse multicritère.
GPL	General Public License.
PERL	practical Extraction and Report Language
PHP	Hypertext Preprocessor.
WWW	World Wide Web
QI	Quasi_Identifiants
DP	Protection Différentiel
GIC	Group Insurance Commission
AS	Attributs sensible
AQI	Attributs de Quasi_Identifiants

Introduction générale

Avec le développement technologique la quantité de données générée par internet, les réseaux sociaux, les sites, les applications de soins, de santé, ...augmentent de jour en jour ce qui a conduit à l'apparition du terme big data.

Il s'agit d'un concept permettant de stocker un nombre indicible d'informations sur une base numérique. Le big data maintenant est présent dans plusieurs domaines tels que la santé, le transport, l'économie, la recherche d'informations, ... En effet le big data est devenu une tendance pour beaucoup d'acteurs industriels du fait de l'apport qu'il offre en qualité de stockage, traitement et analyse de données.

Les mégadonnées peuvent avoir de nombreuses origines et des formes très différentes et avec le développement des méthodes de machine learning ces données sont vulnérables à la protection de la vie privée. L'objectif de notre travail est la protection de la vie privée dans le big data, en montrant les méthodes utilisées pour protéger la vie privée et l'anonymat des utilisateurs dans le big data en faisant une comparaison entre ses méthodes. Ensuite nous choisissant une technique pour l'anonymat ainsi un algorithme pour une implémentation qui sert à anonymiser des grandes quantités de données de plusieurs manières afin de les utiliser plus tard dans plusieurs domaines tels que la préservation de la vie privée dans l'apprentissage automatique.
Ce mémoire est organisé comme suit :

—**Le premier chapitre** est consacré à des définitions, des généralités sur le big data.

— **Le deuxième chapitre** On donnera des définitions sur la vie privée, l'anonymat et

on présentera les différentes méthodes utilisées pour protéger l'anonymat des utilisateurs dans le big data tout en expliquant le contexte utilisé par chaque technique. Ensuite on présentera une étude comparative entre les différentes méthodes (avantages et inconvenients).

— **Le dernier chapitre** est réservé à présenter la technique utilisée pour l'anonymat des données ainsi l'algorithme choisi pour notre implémentation.

Et nous terminerons par une conclusion générale et quelques perspectives.

Chapitre 1

Big Data

1.1 Introduction

Depuis longtemps, les données générées n'ont fait qu'augmenter : à l'heure actuelle, la quantité de données générée chaque année est très importante, estimée à près de 3 trillions ($3 * 10^{18}$) octets [1]. La croissance des données affecte tous les secteurs de la science et de l'économie, ainsi que le développement d'applications Web et de réseaux sociaux [2], ce qui a conduit à l'apparition du terme Big Data. Le mot anglo-saxon signifie littéralement «big data», et sa traduction officielle française recommandée est le big data, même quand on parle parfois de big data. Aujourd'hui, ces mégadonnées sont devenues le centre d'attention des participants dans tous les domaines d'activité.

L'objectif de ce chapitre est de définir le contexte et la classification du terme «big data», de citer les cas d'utilisation du big data, puis d'introduire les enjeux et les risques de sécurité dans le big data. Aussi nous présenterons les avantages et les inconvénients du big data.

1.2 Origine du Big Data

Le Big Data est un nouveau contexte et une grande quantité de données qui ne peuvent être traitées avec les technologies traditionnelles. Le premier projet Big Data concerne les participants qui effectuent la recherche d'informations sur les «moteurs de recherche» Internet (tels que Google et Yahoo). En fait, ces participants sont tous confrontés au problème de la mise à l'échelle du système et de la réponse aux demandes des utilisa-

teurs.

Le Big Data est devenu une tendance de base pour de nombreux acteurs de l'industrie car il contribue à la qualité du stockage, du traitement et de l'analyse des données[3].

1.3 La définition du big data

Littéralement : «Big data» est un ensemble de données de grandes quantités de données structuré ou non structuré. On parle également de grandes quantités de données à travers la similarité avec la biomasse.[4]

Conceptuellement : ce terme popularise non seulement la représentation de la quantité de données, mais vulgarise également l'infrastructure liée au traitement de ces données[5]. Il s'agit d'un très grand ensemble de données difficile à traiter avec des bases de données conventionnelles ou des outils de gestion de l'information [6]

1.4 Caractéristiques du big data

La caractérisation de ces big data se fait généralement selon 3 "V", à savoir le V de volume, de Variété et de Vélocité. D'autres "V" complémentaires peuvent s'ajouter, comme la valeur et la véracité /validité [7]

FIGURE 1.1 – Les Caractéristiques du Big data[8]

1.4.1 Le Volume

Fait référence à la grande quantité de données générées chaque seconde. Pensez simplement à tous les e-mails, tweets, photos, vidéos, données de capteurs que nous générons et partageons chaque seconde. Nous ne parlons plus en téraoctets, mais en zettabytes ou brontobytes.

Rien que sur Facebook, nous envoyons 10 millions de messages chaque jour, «j'aime» 4,5 millions de fois, et nous téléchargeons 350 millions de nouvelles photos chaque jour. Si nous extrayons toutes les données créées dans le monde dans la soirée de 2008, maintenant la même quantité de données sera générée chaque minute. Or, une telle quantité de données est trop grande pour être stockée ou analysée de manière «traditionnelle» (c'est-à-dire une base de données).avec le Big Data, nous pouvons utiliser des systèmes distribués pour stocker et utiliser ces ensembles de données, où différentes parties des données sont stockées à différents endroits mais collectées par logiciel[9]

1.4.2 La Vélocité

Désigne la vitesse à laquelle les nouvelles données sont générées et déplacées. Considérez simplement que les publications sur les réseaux sociaux se répandront en quelques secondes, les transactions bancaires frauduleuses peuvent être détectées en quelques minutes, ou un logiciel qui analyse les réseaux sociaux et saisit l'heure qui a déclenché l'achat. Doit être des millisecondes ! Désormais, le big data nous permet d'analyser le moment où les données sont générées au lieu d'avoir à les analyser dans la base de données[9]

1.4.3 La Variété

Désigne les différents types de données que nous pouvons utiliser. Dans le passé, nous nous appuyions principalement sur des données structurées. Nous pouvons présenter et organiser avec soin les types de données, tels que les transactions de vente par clients, régions, etc. Des données moins structurées, telles que des fichiers texte, des photos, du contenu vidéo, etc. Il a été largement ignoré. Aujourd'hui, nous pouvons utiliser et analyser toutes sortes de données, y compris le texte écrit, la voix et même les tonalités vocales ainsi que les données biométriques, les photos et le contenu vidéo[9]

1.4.4 La Véracité

Désigne la fiabilité des données. Avec autant de formes de mégadonnées, sa qualité et sa précision sont difficiles à vérifier (regardons les tweets avec des balises, des abréviations, des fautes de frappe, la familiarité, la fiabilité et l'exactitude du contenu). mais ! Le Big Data et l'analyse nous permettent désormais d'utiliser ces données pour la production. Le manque de qualité et de précision est généralement le résultat d'une production de masse[9]

1.4.5 La Valeur

C'est le dernier V à considérer quand on parle de big data. Avoir accès au big data c'est super, mais il faut quand même le convertir en valeur, sinon ce sera inutile ! Par conséquent, dans ce sens, on peut dire que la valeur V est très importante ! Il est également important pour les entreprises d'évaluer la rentabilité de la collecte de données. Sans bien comprendre et définir les avantages, nous pouvons facilement tomber dans le piège de la réalisation de projets Big Data. Combien nous coûtent-ils ?[9]

1.5 Classification des Big Data

Connaitre les caractéristiques des données est très important pour déduire ses modèles cachés. Selon le type de données, le format des données, la source de données, l'utilisateur de données, l'utilisation des données, l'analyse des données, le stockage des données et la fréquence des données, les mégadonnées sont divisées en dix catégories. L'émergence des données. Données, suggestions de traitement des données et méthodes de traitement des données [10]. (Comme illustrée par la Figure 1.2).

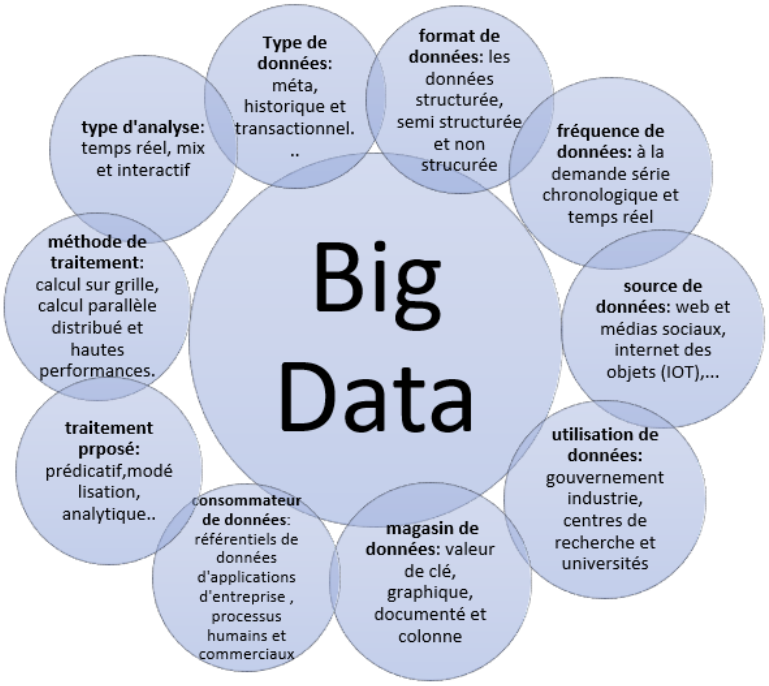


FIGURE 1.2 – Classification des Big Data[11]

1.6 Cas d'usage du Big Data

Aujourd'hui, les mégadonnées sont utilisées dans tous les domaines de la science, de la technologie et des activités socio-économiques. Donnons quelques exemples d'utilisation du big data dans différents domaines d'activité principaux :

FIGURE 1.3 – Les domaines d'usage du big data[12]

1.6.1 Domaine de la recherche scientifique

Dans le domaine de la science et de la technologie, les scientifiques et les ingénieurs sont confrontés au big data, en particulier au big data généré automatiquement par des capteurs ou des instruments de mesure.
Par exemple, dans le domaine de l'astronomie, le Sloan Digital Sky Survey a été un programme d'observation astronomique majeur pendant huit ans (2000-2008), enregistrant 140 To d'images (140.1012). Cependant, son successeur LSST (Large Weather Observation Telescope) peut obtenir ce volume en seulement cinq jours.
En physique, pour trouver le boson de Higgs, le Grand collisionneur de hadrons (LHC) a accumulé De son côté, il y a près de 15 pétaoctets de données (15.1015) chaque année, ce qui équivaut à plus de 3 millions de DVD[13]

1.6.2 Domaine de la santé

concernant la santé, McKinsey a estimé dans le rapport «Big Data Revolution in the Medical Field»[14] que la «Big Data Revolution in the Medical Field» coûterait entre 30 et 450 milliards de dollars.
Les «données» peuvent faire économiser au total 2,6 billions de dollars américains dans le système médical américain. Ces économies comprennent la prévention, le suivi des patients pour modifier leurs habitudes, le diagnostic et l'aide aux médecins pour choisir le traitement le plus approprié ; le personnel médical doit déterminer si le patient a besoin d'une infirmière, d'un médecin généraliste ou d'un spécialiste.
Maîtriser la fraude sur les dépenses en exécutant automatiquement des procédures de remboursement et en détectant les dépenses, et enfin l'innovation, la compréhension de la biologie et l'amélioration des méthodes de traitement grâce aux multiples apports du calcul haute performance. De même, grâce au big data, certaines maladies ou épidémies peuvent être mieux prévenues, ou le traitement des patients peut être amélioré. Par exemple, en analysant les recherches des internautes sur Google, une équipe a réussi à détecter plus rapidement l'arrivée d'une pandémie de grippe [15].
Dans un autre exemple, les chercheurs ont vérifié les données disponibles sur Facebook et ont constaté que les adolescents avaient des comportements dangereux afin de cibler les exercices préventifs[16].

1.6.3 Domaine socio-économique et politique

D'une manière générale, dans le domaine socio-économique, en écoutant mieux les opinions des utilisateurs et en comprenant comment les utilisateurs utilisent ces services[17], les mégadonnées peuvent être utilisées pour simplifier ou ajuster les services fournis. Par exemple, Google Analytics offre aux entreprises et aux administrations publiques la possibilité d'améliorer la conception de leurs sites Web en analysant les visites des internautes. Dans le domaine de l'éducation, à travers l'enseignement à distance (en particulier les cours publics en ligne à grande échelle-MOOC), le traitement des mégadonnées permet d'analyser les activités des étudiants (temps passé, méthode de suivi, temps d'arrêt) - regarder des vidéos éducatives sur Internet Recherche parallèle, etc.
L'analyse des mégadonnées permet également de mieux comprendre les sentiments ou les besoins des citoyens. . Par exemple, lors de la campagne de réélection de Barack Obama en 2012, des consultants ont analysé les messages Twitter en temps réel pour s'adapter à la diffusion en direct du président.

1.6.4 Domaine du transport et de l'énergie

dans le domaine de transport, les mouvements de population peuvent être simulés pour s'adapter aux infrastructures et aux services (horaires des trains, etc.). Pour cela, nous utilisons les données des tickets de transports en commun, du vélo et du covoiturage, ainsi que la localisation géographique des personnes ou des voitures (données cellulaires et systèmes de localisation par satellite). Dans le domaine de l'énergie et du développement durable, les systèmes de comptage intelligents (électricité, gaz naturel, eau) génèrent des mégadonnées pour rationaliser la consommation d'énergie [18].
En plus d'offrir aux citoyens la possibilité de mieux contrôler leur consommation, ces compteurs peuvent également déconnecter l'appareil à distance avec l'accord du client, évitant ainsi une surcharge du réseau. Dans le transport aérien, en combinant les données des capteurs installés sur l'avion avec les données météorologiques, le conduit d'air peut être changé pour économiser du carburant et améliorer la conception, la maintenance ou la sécurité de l'avion [19].

1.7 La sécurité dans le Big Data

La sécurité semble être l'un des principaux obstacles du «big data». Avec l'augmentation de la variabilité et la quantité de données échangées, ces barrières n'ont pas vraiment été éliminées. De nombreuses entreprises spécialisées dans la sécurité des données ont vu le jour, mais elles ne se concentrent que sur un ou plusieurs éléments. En revanche, le progrès technologique est plus rapide que le développement de ces solutions. Mais afin de minimiser le risque, vous devez au moins respecter les quatre règles de sécurité suivantes

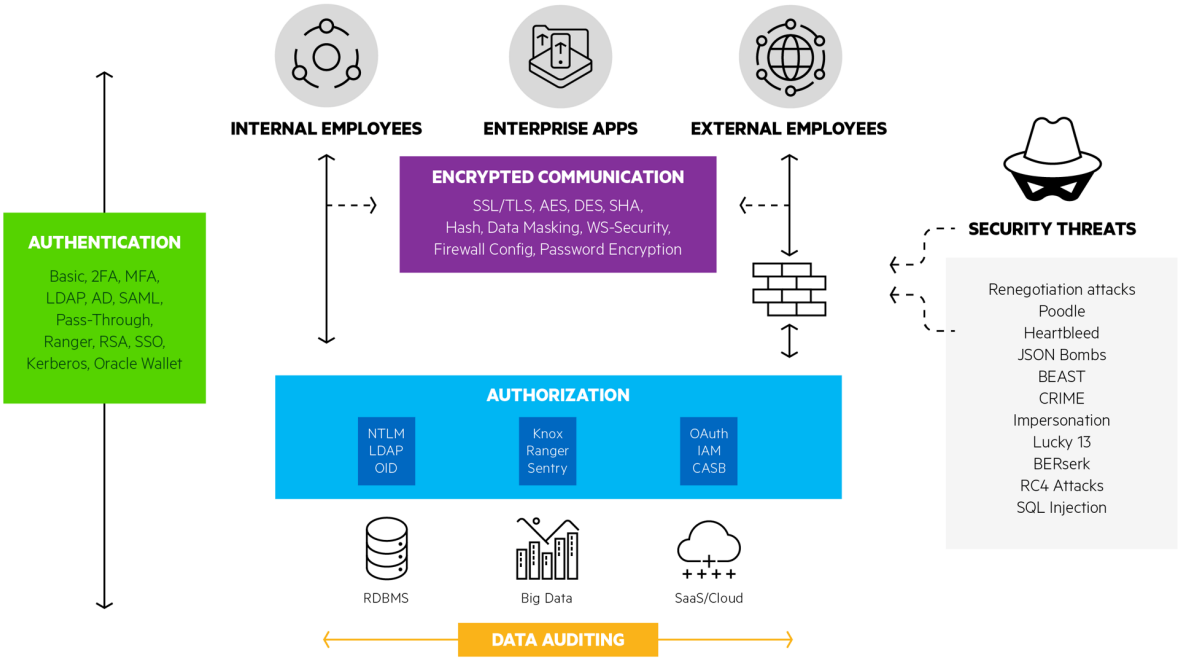


FIGURE 1.4 – Sécurité des données : authentification, autorisation,audit et cryptage[20]

1.7.1 Authentification

La vérification d'identité consiste à garantir l'identité de l'utilisateur, c'est-à-dire à s'assurer que le partenaire de chaque correspondant est bien l'identité qu'il revendique. Afin de protéger le Big Data, il est essentiel d'intégrer un outil permettant d'identifier les personnes ou les processus avant d'autoriser sa connexion. Il existe plusieurs techniques d'authentification qui peuvent être divisées en deux catégories :

— L'authentification de base est appelée authentification simple :

Qui suis-je : identifiant (login).

Preuve-le : Nom d'utilisateur (mot de passe).

— L'authentification forte :Renforcez la vérification d'identité simple Pour renforcer la vérification de L'authentification qui n'est pas suffisamment sécurisée, vous devez ajouter des "verrous" : empreinte biométrique,[21]

1.7.2 Autorisation

Une fois connecté, la personne n'est pas autorisée à effectuer des opérations ou à accéder à toutes les ressources. Plusieurs techniques peuvent être utilisées. Le plus ancien est l'utilisation d'ABAC (Attribute-Based Access Control) : contrôle d'accès basé sur les attributs. Une autre technologie la plus largement utilisée est RBAC (Role-Based Access Control), qui rend possible le contrôle d'accès basé sur les rôles [21].

1.7.3 Audit

Bien qu'une personne ait été identifiée, authentifiée et autorisée à accéder à certaines ressources, un système sûr et fiable doit suivre toutes les opérations de cette personne. Il s'agit de la traçabilité, qui est un processus essentiel [21].

1.7.4 Cryptage

Données Afin de répondre efficacement à l'évolution des menaces, une approche centrée sur les données est également nécessaire pour protéger les informations sensibles (telles que les informations de carte de crédit ou les dossiers confidentiels des patients). Vous pouvez également y parvenir en chiffrant certaines données, qui sont stockées en toute sécurité et deviennent difficiles à comprendre en raison de l'utilisation d'algorithmes de chiffrement pour les garder secrètes. Afin de rendre ce processus efficace, la clé de chiffrement doit être strictement protégée [21]. Les quatre règles décrites doivent former la base de votre stratégie Big Data. Si vous ne suivez pas ces règles, vos données seront menacées. C'est pourquoi vous devez suivre strictement ces règles [21].

1.8 Techniques d'analyse de données

Il existe trois principales méthodes d'analyse des données pour le Big Data :
-La méthode descriptive vise à mettre en évidence les informations présentes dans les

données. Mais il est masqué par une grande quantité de données [22].Certaines techniques et algorithmes utilisés dans l'analyse descriptive comprennent :

—- Analyse factorielle (PCA et ACM)

—- Méthode du centre mobile

—- Classification hiérarchique

—- Classification des neurones

—- Recherche d'association

-La méthode de prédiction vise à déduire de nouvelles informations à partir d'informations actuelles [23] Cette technologie utilise l'intelligence artificielle, les principales méthodes sont :

—- Arbre de décision

—- Les réseaux de neurones

—- Classification bayésienne

—- Support Vector Machine (SVM)

—- Voisin le plus proche (KNN)

-Les méthodes prescriptives visent à identifier et anticiper les actions / décisions les plus appropriées.

Le meilleur choix pour atteindre la situation souhaitée.[23]

1.9 Les avantages du Big data

L'architecture Big Data peut apporter certains avantages, tels que :

–**Extensibilité (scalabilité) :** le concept de Big Data fournit une architecture évolutive qui peut éviter la taille de l'infrastructure et l'espace disque requis.

–**Performance :** Grâce au traitement parallèle des données et à son système de fichiers distribué, le concept Big Data offre des performances élevées en réduisant la latence des requêtes.

–**Faible coût :** en raison du système de fichiers distribué, il n'est plus nécessaire de concentrer les données dans la matrice de stockage, ce qui est généralement trop coûteux, et le disque interne du serveur est suffisant.

–**Disponibilité :** les disques RAID, qui sont généralement chers, ne sont plus nécessaires. L'architecture Big Data fournit son propre mécanisme de haute disponibilité.[23]

1.10 Les inconvénients du big data

–**La première limitation** est la pertinence et l'exactitude des informations. S'il y a des erreurs dans les données utilisées, leur traitement automatique les reproduira, ce qui faussera l'interprétation et l'analyse qui en découlera. Par exemple, dans l'industrie, les erreurs de données peuvent causer de nombreuses difficultés. Par exemple, si le capteur n'est pas correctement réglé ou entretenu, les résultats obtenus seront erronés et toutes les prévisions qui peuvent être faites sont erronées. Par conséquent, il est nécessaire de contrôler ces «erreurs» approximatives.

–**La deuxième limite** est la sur-interprétation de la causalité. Le simple fait que l'algorithme trouve la corrélation entre différentes données ne signifie pas qu'il existe une relation causale : même si les statistiques montrent que les gens sont plus susceptibles de mourir dans des lits d'hôpitaux que dans leurs propres lits. Avoir son propre lit ne signifie pas que l'hôpital est le lieu de décès le plus dangereux.

–**La troisième limitation** est la répétabilité des résultats scientifiques. Par exemple, en astrophysique, avec de grandes quantités de données et des algorithmes très complexes, il est généralement impossible pour les chercheurs de copier des nombres dans des articles. La répétabilité est au cœur du processus scientifique. De nos jours, de plus en plus de scientifiques publient leurs résultats avec le code source utilisé pour analyser les données et les scripts utilisés pour traiter les données et générer des chiffres. Désormais, en raison des «données ouvertes», la disponibilité des données a également tendance à augmenter. Enfin, l'utilisation massive des données ne peut remplacer complètement les méthodes théoriques. Pouvons-nous découvrir les concepts de base de la physique uniquement à partir de données ? Même à son époque, Albert Einstein a répondu par la négative à cette question, et c'est encore un problème aujourd'hui [24].

1.11 Enjeux du Big data

La collecte et / ou le traitement de grandes quantités de données soulève les questions importantes suivantes pour les scientifiques et la société dans son ensemble :

–**La protection des données et les droits connexes** sont les principaux défis du Big Data. Les scientifiques doivent protéger les résultats de leurs recherches et les citoyens ont le droit de protéger leurs données personnelles. Le développement des objets connectés

(Internet des objets ou Internet des objets) transmettra de plus en plus de données, ce qui aggrave ce problème. À cette fin, le cadre juridique est progressivement renforcé. [25]
–**Superviser et gérer le «machine learning» pour éviter toute dérive.** Pour cette raison, la création d'un journal détaillant la phase de formation informatique est une idée qui facilitera les recours juridiques. [25]
–**L'impact écologique du big data** est le dernier problème, car le fonctionnement des serveurs, des supercalculateurs et des ressources de stockage et de communication (réseaux filaires ou sans fil) nécessite beaucoup d'énergie [26].
La gestion et l'utilisation efficace des mégadonnées restent un moyen d'accélérer l'acquisition de connaissances et donc de faire progresser la recherche scientifique. C'est aussi un moteur important pour la croissance de l'économie, des grands groupes ou de l'administration publique. L'Internet des objets et les quantités massives de données générées offrent d'innombrables possibilités pour améliorer la qualité de vie et la sécurité des citoyens [26].

1.12 Conclusion

L'explosion quantitative des données numériques a obligé les chercheurs à trouver de nouvelles façons de voir et d'analyser le monde. Il s'agit de découvrir de nouveaux ordres de grandeur concernant, la recherche, le partage, le stockage, l'analyse et la présentation des données ; le Big Data se présente comme une solution pour ces problèmes . le big data est utilisées dans plusieurs domaines tels que la santé , l'économie ,les recherches scientifiques .. d'un autre coté le Big data comporte des risques liées au respect de la vie privée et la confidentialité .dans le prochain chapitre nous allons aborder et détailler le domaine de l'anonymat dans le Big data ainsi les techniques et les méthodes utilisées.

Chapitre 2

L'anonymat dans le big data

2.1 Introduction

Le big data est la tendance du siècle, il devient de plus en plus utilisable dans le monde, cependant le big data contient plusieurs problèmes de sécurité et de protection de la vie privée. Dans ce chapitre nous allons définir la vie privée et l'anonymat et de tracer les frontières entre ces deux dernières. Nous allons ensuite présenter les différentes méthodes utilisées pour protéger la vie privée dans le big data et de mentionner le rôle de chacune, ensuite nous terminerons par une étude comparative entre ces différentes méthodes.

2.2 C'est quoi la vie privée ?

Que voulons nous dire par « la vie privée » ? la vie privée concerne toutes informations ou actes destinés à être individuel à une personne et à elle seule qui est privées au grand public, telle que l'identité de la personne, dossiers médicaux, conversations privées (mails, sms, etc.), photos et vidéos personnelles .... Afin de pouvoir préserver cette intimité et garder ces informations privées dans le monde du numérique, il existe plusieurs techniques telles que la cryptographie qui se base sur le chiffrement des données à l'aides des clés et pouvoir les échanger sur internet tout en les gardant privées. [28]

2.3 C'est quoi l'anonymat ?

L'anonymat signifie l'absence du nom d'une personne ou bien l'absence de son identité de manière générale. En informatique, un utilisateur anonyme est un utilisateur dont nous ignorons sont identité mais que nous savons ce qu'il est en train de faire. Par exemple, une personne peut se connecter à un service d'anonymat comme 'Tor' pour poster un message politique sous un nom d'utilisateur anonyme, dans ce cas, la personne peut publier un message public tout en gardant son identité anonyme[28]

Quelle est la différence entre l'anonymat et la vie privée ?

La frontière entre la vie privée et l'anonymat est très mince, ce sont deux définitions très proches dans leur sens mais qui peuvent être présentées dans des contextes très différents, il est donc facile de se confondre. Il est primordial de comprendre la différence, car les solutions varieront dans la façon dont elles garantissent la préservation de la vie privée et l'anonymat.[28]

2.4 Les relations entre la vie privée et l'anonymat ?

La vie privée et l'anonymat sont deux concepts proches mais différents. Ils sont tous les deux de plus en plus nécessaires pour remédier aux différents pièges et différents traçages présents de nos jours sur internet, que ce soit de manière légale ou non, il est important de comprendre pourquoi ils font partie intégrante de nos libertés civiles et pourquoi ils ne sont pas seulement bénéfiques pour l'individu, mais absolument critiques pour une société libre.

La vie privée est la capacité de garder ses informations personnelles pour soi-même, quel que soit leur impact sur la société. Par exemple, un patient dans un hôpital a le droit d'exiger que son dossier médical ne soit être consulter par aucune personne à part son médecin ni même pas par sa sécurité sociale même si cette dernière le suspecte de fraude[29].

La vie privée est donc un concept décrivant les activités que nous gardons entièrement à nous-même, ou à un groupe limité de personnes.

En revanche, l'anonymat à l'inverse du concept de la vie privée, nous souhaitons partager avec tout le monde (au public) une information mais à condition que notre identité ne soit pas révélée. L'anonymat est donc une sorte de préservation de vie privée car nous

souhaitons garder une information rien qu'à nous même qui est notre identité, nous pouvons donc dire que l'anonymat est inclus dans la vie privée.
Pour donner un exemple typique, lorsqu'une personne souhaite dénoncer un acte de criminalité mais qu'elle ne veut surtout pas que son identité soit divulguée pour des raisons de sécurité, la personne exige donc de garder son anonymat et c'est un choix privé qui rentre dans ses droits de préserver sa vie privée[29]

2.5 L'anonymisation de micro-données

Pour protéger la confidentialité, le but de l'anonymisation des données est d'empêcher :
1) Singulariser les individus dans un ensemble de données,
2) Le lien entre deux enregistrements (dont l'un correspond à des données individuelles) au sein d'un ensemble de données (ou entre deux ensembles de données séparés)
3) Dérivez les informations de l'ensemble de données. Par conséquent, la pseudonymisation couramment utilisée par les organisations (en d'autres termes, supprimer les identifiants clairs et les remplacer par des pseudonymes) ne peut garantir la ré-identification des individus, en particulier dans l'identification personnelle. Contexte de publication de données. Plusieurs situations spécifiques prouvent ce point. Citons, à titre d' exemple est celui de la société de location de films en ligne Netflix. En 2006, dans le cadre de la compétition, le «Netflix prize» visant à améliorer son système de recommandation, Netflix a annoncé un ensemble de données identifiées par des pseudonymes, qui contient plus de 100 millions de critiques de films par ses abonnés entre décembre 1999 et 2005.
Des chercheurs sont arrivés à démontré que la plupart des abonnés peuvent être identifiés par une connaissance limitée de jusqu'à 8 critiques de films faites par ces abonnés et leurs dates d' évaluation (Narayanan et Shmatikov 2006).Enfin, en 2002, Sweeney (Sweeney 2002b) a acheté des documents d'électeur et des fichiers de données de patients auprès d'une compagnie d'assurance médicale et a supprimé l'identifiant du patient pour prouver qu'il est possible de réidentifier les patients par simple appariement. Pour illustrer ce processus, les tableaux 2.1 et 2.2 montre un exemple de tableaux de patients et d'électeurs cités dans (Sweeney 2002a). Dans cet exemple, l'anonymisation est effectuée à l'aide de pseudonymes (C'est-à-dire supprimer l'identifiant explicite du nom et du numéro de sécurité sociale).[30]

SSN	**Name**	**Race**	**datedenaissance**	**sexe**	**Zip**	**état civil**	**maladie**
		asian	27/9/1964	F	10000	Divorcé	Cancer
		asian	30/9/1964	F	10000	Divorcé	Grippe
		asian	18/4/1964	M	10000	Mariée	TB
		asian	15/4/1964	M	10000	Mariée	Grippe
		black	13/3/1963	M	16000	Mariée	Cancer
		black	18/3 /1963	M	16000	Mariée	Rhume
		black	13/9/1964	F	12000	Mariée	Rhume
		black	7/9/1964	F	12000	Mariée	Grippe
		white	14/5/1961	M	16000	célibataire	TB
		white	8/5/1961	M	16000	célibataire	Grippe
		white	15/9/1961	F	15000	veuve	Rhume

TABLE 2.1 – Données médicales anonymisées

Name	**ville**	**zip**	**datedenaissance**	**sexe**	**état civil**	**maladie**
sue j.carlson	bouira	15000	15/9/1961	F	veuve	Rhume

TABLE 2.2 – la liste de votants

Ainsi, il est possible de révéler l'identité de la seule femme née le 15 septembre 1961 et résidant dans la zone 15000 en vérifiant les extraits de la fiche médicale anonyme et la liste électorale. Voici Sue J. Carlson. Cela vous permet de divulguer ses informations privées, comme le fait qu'elle souffre de difficultés respiratoires.

2.6 Modèles d'attaque de micro-données publiées

La littérature sur la protection de la vie privée dans le contexte de la divulgation de données souligne que la stratégie habituellement utilisée par les attaquants pour obtenir

des informations sensibles repose sur leur connaissance du contexte. Dans ces stratégies (également appelées modèles d'attaque), les adversaires déduisent des informations sensibles sur leurs victimes en établissant des liens et même en faisant des inférences probabilistes (voir Figure 2.1).

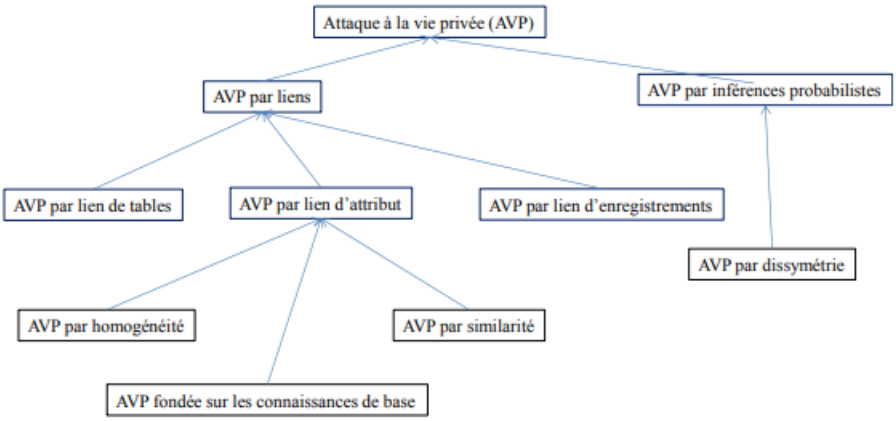


FIGURE 2.1 – Taxonomie des modèles d'attaque de la vie privée

Le modèle d'attaque par lien : est applicable à la situation suivante : l'attaquant connaît le quasi-identifiant de sa victime, c'est-à-dire la personne qu'il souhaite obtenir ses données sensibles (BCM Fung et al. 2010).

Dans l'attaque par lien de tables (« table linkage »), l'adversaire ne sait pas a priori s'il y a des données sur sa victime sur la table, mais il peut déduire des informations qu'il y a observées.Par exemple, une table publique externe E et une table T' anonymisée à partir de la table d'origine T. Soit l'individu l dont la valeur sensible doit être connue . Supposons que j'appartienne à un groupe de k individus dans T'et k'individus dans E. La probabilité de I individu dans T'est k / k '. Par exemple, si k = 4 et k '= 5, la probabilité de son existence est égale à $4/5 = 0{,}8$, ce qui est une probabilité élevée.

Le modèle d'attaque par «lien d'attributs» pour «attribute linkage», constitue une menace de réidentification lorsque l'adversaire connaît le quasi-identifiant de la victime et qu'il peut, par simple analyse des deux tables (celle dont il dispose et celle publiée), inférer des connaissances qui le mèneront vers des réidentifications sans pour autant savoir a priori si sa victime est présente ou absente de la table publiée.Par exemple, si deux tableaux T et E anonymes ont été publiés, et que l'un des deux tableaux (ex : E)

contient des données médicales (ex : maladies), et que l'attaquant connaît l'identifiant de sa victime, il pourra déterminer la victime Chacun des individus appartenant aux groupes T et E, et par cette identification, peut être déduit dans une certaine mesure que sa victime a une maladie du groupe E, parmi laquelle se trouve sa victime.

Dans le scénario «lien d'attributs» , Le risque de ré-identification existe toujours parce que l'attaquant connaît le QI de la victime et peut trouver des individus qui partagent le même QI. Ce suivi lui donnera éventuellement l'occasion de déduire des informations sensibles de la victime en fonction de la valeur de sensibilité associée à la population où la victime a été retrouvée. Si ces valeurs sont les mêmes pour les attributs sensibles (par exemple, tous les individus du même groupe souffrent par coïncidence de la même maladie), alors les données sensibles de la victime sont directement déduites.

On parle alors d'attaque par homogénéité. Par exemple, si nous supposons que le quasi-identifiant de la table «patient» comprend le sexe, l'âge et le code postal du patient, et dans ce tableau, tous les 52 ans vivant dans le district 20 (code postal 75020) ont Ulcères, afin que l'adversaire puisse vérifier si sa victime appartient au groupe et en déduire qu'il a la même maladie. Si toutes les femmes de 52 ans vivant dans le district 20 souffrent désormais d'ulcères d'estomac ou d'indigestion, l'adversaire pourra en déduire que leurs victimes ont des problèmes d'estomac. Ce type d'attaque repose sur l'analyse de la similitude sémantique des valeurs de données sensibles, **est nommé « attaque par similarité » (similarity attack).** Un autre exemple du scénario d'attaque «lien d'attribut» cité dans la littérature est «l'attaque de connaissances de base». Dans ce cas, l'attaquant a une connaissance suffisante des attributs sensibles et une fois qu'un groupe de victimes est découvert, il peut deviner les données sensibles de la victime. Par exemple, supposons que l'agresseur connaisse l'âge et le code postal de la victime et que le tableau «patients» montre qu'une patiente de 52 ans à Paris souffre d'une maladie cardiaque ou d'hypertension artérielle. Si l'agresseur sait que 80% des victimes souffrent encore d'hypertension artérielle, il peut conclure que la victime est susceptible de souffrir d'hypertension artérielle. Dans le scénario suivant que l'on nomme «lien d'enregistrements», Outre le fait que l'attaquant connaissait la quasi-identification de la victime, il savait également que les informations de la victime faisaient partie du formulaire publié. La menace est réelle lorsqu'il n'y a pratiquement aucun enregistrement dans le tableau publié qui soit identique à la valeur de QI de la victime. Enfin, dans le modèle d'attaque par inférence probabiliste,

l'attaquant n'établira pas de liens avec des tables, enregistrements ou attributs sensibles. Au lieu de cela, il s'est appuyé sur ses croyances probabilistes avant et après l'analyse de la distribution des valeurs d'attributs sensibles de la table publiée.
Le scénario d'attaque, fréquemment mentionné dans la littérature pour ce type de modèle, est l'attaque par dissymétrie (« skewness attack »). Dans ce cas, l'adversaire compare la distribution globale des valeurs d'attributs sensibles (croyances probabilistes avant l'analyse des données publiées) avec la distribution de la même valeur d'attribut sensible pour déduire la valeur de la donnée sensible de la victime. Dans un groupe de personnes ayant le même QI (croyance probabiliste après analyse des données publiées). Par exemple, si la proportion d'individus atteints d'un cancer gastrique dans un groupe donné est significativement plus élevée que celle de la population générale, on peut en déduire que les personnes de ce groupe de personnes atteintes d'un cancer gastrique ont une probabilité élevée de souffrir d'un cancer gastrique. pour faire face à ces scénarios d'attaque potentiels, des modèles de protection de la vie privée ont été proposés dans la littérature.
Compte tenu de l'utilisation ultérieure des données, ces modèles sont mis en œuvre grâce à la technologie d'anonymisation. Ces techniques sont instanciées par des algorithmes plus efficaces. Les sections suivantes présentent brièvement certains de ces modèles et les technologies impliquées[30].

2.7 Problématique

A cause de la croissance rapide des technologies de bases de données, de mise en réseau et d'informatique, une grande quantité de données personnelles peuvent être intégrées et analysées numériquement, ce qui conduit à une utilisation accrue des outils d'exploration de données pour déduire des tendances et des modèles.
Cette évolution a suscité des préoccupations universelles en matière de protection de la vie privée des individus. En effet, plusieurs techniques de croisement de données ont été développées pour pouvoir en déduire des informations critiques à travers des données publics. La problématique qui se pose dans ce contexte est :
Comment pouvons-nous remédier aux attaques par croisement de données à fin d'assurer l'anonymat de nos bases de données ?
Pour se faire, nous allons utiliser des méthodes d'anonymisation qui visent à rendre l'en-

registrement individuel indissociable d'un enregistrement d'un groupe en utilisant des techniques de généralisation et de suppression.

2.8 Les méthodes de sécurités dans le big data

2.8.1 La pseudonymisation

La pseudonymisation consiste à supprimer des champs directement identifiables de l'enregistrement et à ajouter un nouveau champ à chaque enregistrement, appelé pseudonyme, dont la caractéristique est de rendre impossible le lien entre la nouvelle valeur et la personne réelle. Pour créer ce pseudo, nous utilisons souvent une fonction de hachage appliquée à l'un des champs identifiant (par exemple, un numéro de sécurité sociale).

Cette fonction de hachage est une fonction spéciale qui rend (ou du moins très difficile) impossible de dériver la valeur initiale. Par conséquent, nous voyons que deux entités disposant d'informations sur la même personne identifiée par le numéro de sécurité sociale de la même personne peuvent partager ces données de manière anonyme en hachant l'identifiant.
Vous pouvez également utiliser une fonction aléatoire pour générer un identifiant unique pour chaque personne, mais nous verrons plus tard que cela ne résout pas tous les problèmes.
Le plus grand avantage des pseudonymes est qu'il n'y a aucune restriction sur le traitement ultérieur des données. Tant que nous ne pouvons pas identifier directement les champs que nous traitons, nous pouvons effectuer exactement les mêmes calculs que dans les bases de données non anonymes. Par exemple, la figure 2.1 montre un exemple de calcul de l'âge moyen pour une pathologie donnée. L'utilisation de données pseudonymes n'interférera pas avec ce calcul[31].

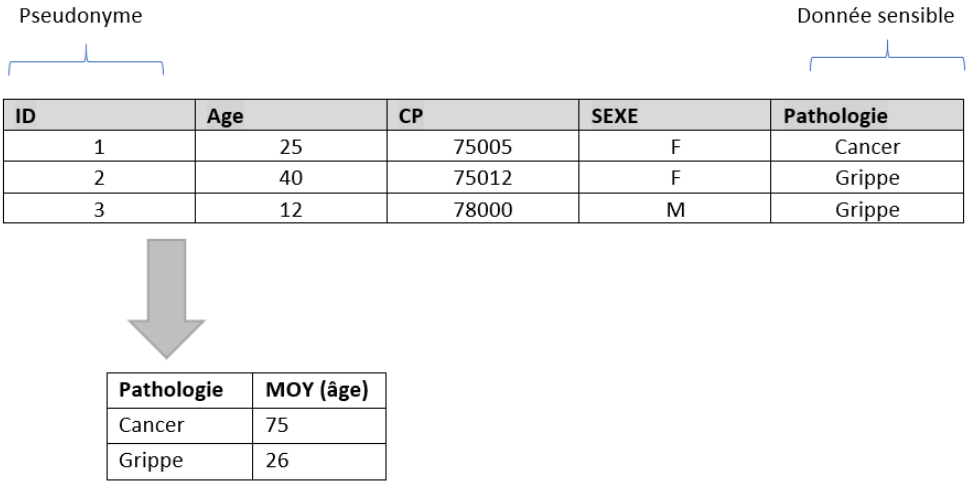


FIGURE 2.2 – Pseudonymisation et exemple de calcul

Cependant, les pseudonymes ne sont pas considérés comme un moyen d'anonymat car ils n'offrent pas un niveau de protection suffisamment élevé : la combinaison d'autres champs peut permettre de trouver des personnes pertinentes [31].

2.8.2 Le k-anonymat

Une publication de données est dite anonyme [32, 33] si les informations relatives à chaque personne contenue dans la publication ne peuvent être perçues par au moins k-1 personnes dont les informations figurent dans la publication. Dans le contexte des problèmes de k-anonymat, une base de données est une table qui se compose de n lignes et m colonnes, où chaque ligne de la table représente un enregistrement relatif à un individu particulier d'une population et les entrées dans les différentes lignes n'ont pas besoin d'être uniques. Les valeurs dans les différentes colonnes sont les valeurs des attributs liés aux membres de la population.

Le tableau 2.1 est une base de données non anonymisée comprenant les dossiers de patients d'un hôpital fictif à Hyderabad. Ces données comportent six attributs ainsi que dix enregistrements. Il existe deux techniques courantes pour réaliser le k-anonymat pour une certaine valeur de k.

Prénom	**Age**	**Sexe**	**Ville**	**Religion**	**maladie**
Amina	29	F	Bouira	musulman	cancer
Manel	24	F	Bejaia	musulman	Grippe
Kahina	28	F	Bouira	judaïsme	TB
Djamel	27	M	Alger	parsi	pas de maladie
Nadia	24	F	Bejaia	chretien	Rhume
Walid	23	M	Alger	buddist	TB
Adel	19	M	Bejaia	musulman	cancer
Morad	29	M	Alger	musulman	Rhume
Smail	17	M	Bejaia	chretien	Rhume
Karim	19	M	Bejaia	chretien	Grippe

TABLE 2.3 – Un ensemble de données non anonymisées comprenant les enregistrements des parties

1. Suppression Dans cette méthode, certaines valeurs des attributs sont supplantées par un astérisque "*". Toutes les valeurs d'une colonne ou certaines d'entre elles peuvent être remplacées par "*". Dans le a anonymisé le tableau 2.2, a remplacé toutes les valeurs de l'attribut "Name" et chacune des dans l'attribut "Religion" par un "*".

2. Généralisation Dans cette méthode, les valeurs individuelles des attributs sont remplacées par un catégorie plus large. Par exemple, la valeur "19" de l'attribut "age" peut être supplantée par "20", la valeur "23" par "20 ¡ age 30", etc.

Le tableau 2.2 est doublement anonyme en ce qui concerne les attributs "âge", "sexe" et "état de domicile", car pour toute combinaison de ces attributs dans une ligne du tableau, il y a toujours au moins deux lignes avec ces attributs exacts. Les attributs dont dispose un adversaire sont appelés "quasi-identifiants". Chaque tuple de "quasi-identifiant" se trouve dans au moins k enregistrements pour un ensemble de données avec k-anonymat. Les données k-anonymes peuvent toujours être impuissantes face à des attaques telles que l'attaque de correspondance non triée, l'attaque temporelle et l'attaque de libération complémentaire [34, 35].

Du côté positif, il présentera un algorithme d'approximation O(k log k) gourmand pour un

k-anonymat optimal via la suppression des entrées. La complexité de rendre les relations des enregistrements privés k-anonymes, tout en minimisant la quantité des informations qui ne sont pas divulguées et qui garantissent simultanément l'anonymat des personnes.

Prénom	**Age**	**Sexe**	**Ville**	**Religion**	**maladie**
*	29	F	Bouira	*	Cancer
*	24	F	Bejaia	*	Grippe
*	28	F	Bouira	*	TB
*	27	M	Alger	*	pas de maladie
*	24	F	Bejaia	*	Rhume
*	23	M	Alger	*	TB
*	19	M	Bejaia	*	Cancer
*	29	M	Alger	*	Rhume
*	17	M	Bejaia	*	Rhume
*	19	M	Bejaia	*	Grippe

TABLE 2.4 – 2-anonymat en ce qui concerne l'attribut "âge","sexe" et "Ville"

Jusqu'à un groupe de taille k, et de retenir un minimum d'informations pour atteindre ce niveau de confidentialité et ce problème d'optimisation est difficile à résoudre. En général, une autre restriction du problème où les attributs sont supprimés au lieu des entrées individuelles est également NP-difficile[37]. Nous nous dirigeons donc vers une stratégie de l- diversité d'anonymisation des données.

Limitation :

1. K-anonymat est insuffisant pour empêcher la divulgation des attributs.
2. Il peut souffrir d'une attaque d'homogénéité et d'une attaque de connaissance de fond.

2.8.3 l-diversité

On dit qu'une classe d'équivalence a une diversité L s'il y a au moins des valeurs "bien représentées" pour le sensible attribut. Si l'on sait que le salaire de Peter se situe entre 3 000 et 5 000 euros, on peut en conclure qu'il souffre d'une maladie de l'estomac. Ainsi, des fuites d'informations sensibles se produisent, ce qui donne lieu à une méthode plus

efficace appelée "t-closeness".[37]

Limitations :

1. La L-diversité est difficile à atteindre

2. La L-diversité est insuffisante pour empêcher la divulgation des attributs.

Age	**Salaire**	**maladie**
2*	3K	Gastric ulcer
2*	4K	Gastrics
2*	5K	Stomach cancer
>= 45	6K	Flu
>= 45	11K	Pneumonia
>= 45	8K	bronchitis

TABLE 2.5 – Données brutes

Age	**Salaire**	**maladie**
2*	3K	Gastric ulcer
2*	4K	Gastrics
2*	5K	Stomach cancer
>= 45	6K	Flu
>= 45	11K	Pneumonia
>= 45	8K	bronchitis

TABLE 2.6 – données anonymes et diverses

2.8.4 La t-proximité

Il s'agit d'une amélioration supplémentaire de l'anonymisation basée sur le groupe l-diversité qui est utilisée pour préserver la vie privée dans les ensembles de données en diminuant la granularité d'une représentation de données. Cette réduction est un compromis qui entraîne une certaine perte d'adéquation des algorithmes de gestion ou d'extraction des données afin de gagner un peu de vie privée. Le modèle de t-proximité (distance égale/hiérarchique) [31, 32] étend le modèle de l-diversité en traitant les valeurs d'un attribut de manière distincte en tenant compte de la distribution des valeurs des données pour cet attribut.

Une classe d'équivalence est dite proche de t si la distance entre le transport d'un attribut sensible dans cette classe et la distribution de l'attribut dans l'ensemble du tableau est inférieure à un seuil t. Un tableau est dit proche de t si toutes les classes d'équivalence sont proches de t. Le principal avantage de la t-proximité est qu'elle permet d'intercepter la divulgation d'un attribut. Le problème de la proximité t est que plus la taille et la variété des données augmentent, plus les chances de ré identification augmentent. L'approche de

force brute qui examine chaque partition possible du tableau pour trouver la solution optimale prend un temps de nO(n)mO(1) = 2O(nlogn)mO(1). Nous améliorons d'abord cette valeur liée à l'exponentielle simple dans n (Notez qu'elle ne peut être améliorée en polynôme à moins que P = NP) [38].

2.8.5 La confidentialité différentielle

La protection différenciée de la vie privée [40] est une technologie qui permet aux chercheurs et aux analystes de bases de données d'obtenir des informations utiles à partir des bases de données qui contiennent des données personnelles l'information des personnes sans révéler l'identité personnelle des individus.

Ce se fait en introduisant un minimum de distraction dans les informations fournies par le système de base de données. La distraction introduite est suffisamment importante pour protéger la vie privée et en même temps suffisamment petit pour que les informations fournies à l'analyste soient toujours utile. Auparavant, certaines techniques ont été utilisées pour protéger la vie privée, mais elles se sont avérées sans succès.

Au milieu des années 90, lorsque la Commonwealth of Massachusetts Group Insurance Commission (GIC) a rendu public le dossier médical anonyme de ses clients à des fins de recherche au profit de la société [39]. Le GIC cache certaines informations comme le nom, l'adresse, etc. afin de protéger leur vie privée.

Latanya Sweeney (alors doctorante au MIT) utilise le système de et de la base de données des électeurs publiée par le GIC, a réussi à identifier le dossier de santé en les comparant et en les mettant en relation. Ainsi, le fait de cacher certaines informations ne peut pas garantir la protection de l'identité individuelle. La protection différentielle de la vie privée (DP) vise à apporter une solution à ce problème, comme le montre la figure 2.2.

Dans la DP, les analystes ne bénéficient pas d'un accès direct à la base de données contenant les données personnelles. l'information. Un logiciel intermédiaire est introduit entre la base de données et l'analyste pour protéger la vie privée. Ce logiciel intermédiaire est également appelé "logiciel de protection de la vie privée garde".

Étape 1 L'analyste peut effectuer une recherche dans la base de données par cet intermédiaire de confidentialité garde.

Étape 2 Le gardien de la vie privée prend la requête de l'analyste et évalue cette requête et d'autres demandes antérieures concernant le risque pour la vie privée. Après évaluation

du risque pour la vie privée.

Étape 3 Le responsable de la protection de la vie privée obtient ensuite la réponse dans la base de données.

Étape 4 Ajoutez-y une distorsion en fonction du risque de violation de la vie privée évalué et transmettez-la enfin à l'analyste.

La quantité de distorsion ajoutée aux données pures est proportionnelle au risque pour la vie privée évalué. Si le risque pour la vie privée est faible, la distorsion ajoutée est suffisamment faible pour ne pas affecter la qualité de la réponse, mais suffisamment importante pour protéger la confidentialité individuelle de la base de données. Mais si le risque pour la vie privée est élevé, la distorsion ajoutée est plus importante.

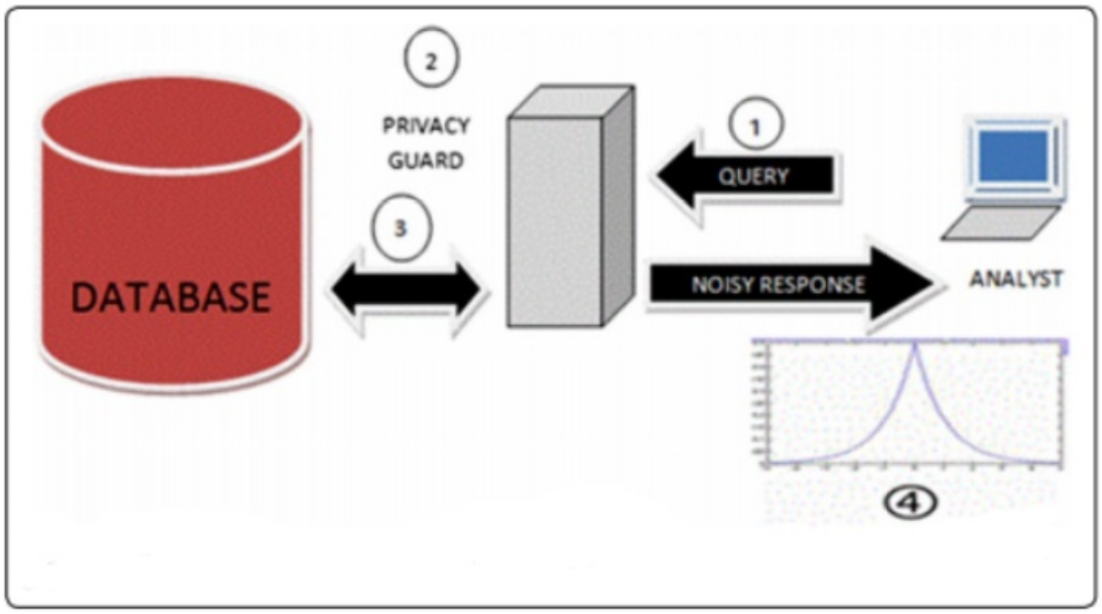


FIGURE 2.3 – Confidentialité Différentielle comme solution pour préserver la vie privée dans le big data

2.9 Comparaison entre méthodes

Les méthodes	**Avantages**	**Inconvénients**
k-anonymat	protéger contre les attaques de couplage d'enregistrements et d'attributs[41]	les attributs sensibles et l'attaque des connaissances de base[42]
La l-diversité	protéger contre les attaques par sub graphes et par liens d'attributs[42]	perte d'utilitaires de données importants et attaque d'asymétrie[42]
La confidentialité différentielle	garantit le respect de la vie privée même lorsque plusieurs parties accèdent à sa base de données[43]	les attaques par couplage d'enregistrements et le coût élevé des calculs
La t-proximité	éviter le problème de déduction d'informations (la l-diversité)[41]	dégradation de l'utilité des données [42]
La pseudonymisation	il n'y a aucune limite sur le traitement subséquent des données	dégradation de l'utilité des données

TABLE 2.7 – Comparaison entre méthodes

2.10 Les techniques d'anonymat

2.10.1 La généralisation

La généralisation consiste à remplacer la valeur réelle de l'attribut par une valeur moins spécifique, plus générale et fidèle à l'original [44]. Initialement, cette technique était utilisée pour les attributs catégoriels et employait des hiérarchies prédéfinies de domaine et de généralisation de valeur [44].

La généralisation a été étendue pour les attributs numériques en utilisant des hiérarchies prédéfinies [44] ou un modèle sans hiérarchie [46]. À chaque attribut catégoriel, une hiérarchie de généralisation de domaine est associée. Les valeurs des différents domaines de cette hiérarchie sont représentées dans un arbre appelé hiérarchie de généralisation des valeurs. Nous illustrons la hiérarchie de généralisation des domaines et des valeurs dans la figure 2.3 pour les attributs Age et Ville. Il existe plusieurs façons d'effectuer une généralisation.

La généralisation qui mappe toutes les valeurs d'un attribut catégoriel de quasi-identifiant de données initiale à un domaine plus général dans sa hiérarchie de généralisation de domaine est appelée « généralisation de domaine complet [45, 46] ». La généralisation peut également mapper les valeurs d'un attribut à différents domaines dans sa hiérarchie de généralisation de domaine, chaque valeur étant remplacée par la même valeur généralisée dans l'ensemble de données [45].
La généralisation la moins restrictive, appelée généralisation au niveau de la cellule [48], étend le modèle Iyengar [44] en permettant à la même valeur d'être mappée à différentes valeurs généralisées, dans des tuples distincts.

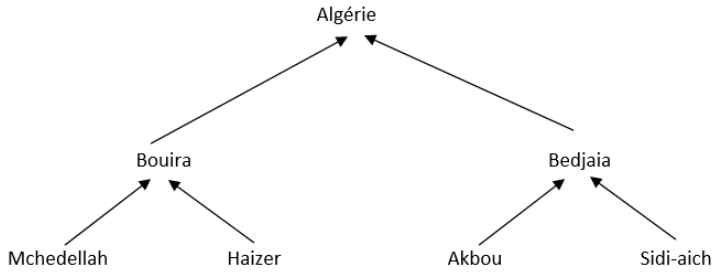


FIGURE 2.4 – Hiérarchie de généralisation de l'attribut Ville

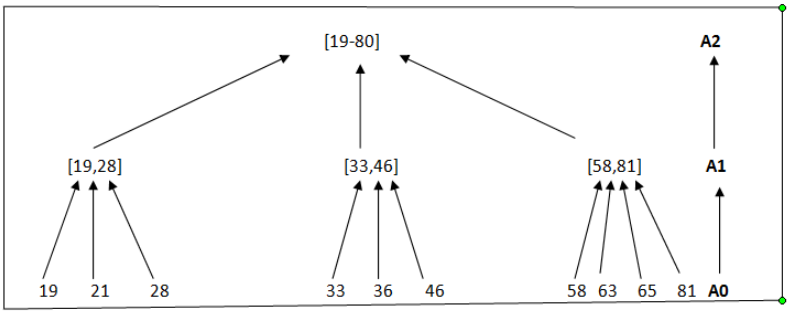


FIGURE 2.5 – Hiérarchie de généralisation de l'attribut Age

2.10.2 La Suppression

La suppression des tuples [47] est la seule autre méthode utilisée dans notre travail ,masquer les données initiales. En éliminant des tuples entiers, nous sommes en mesure de réduire la quantité de généralisation requise pour obtenir la propriété k-anonymity dans les tuples restants. Comme le modèle de k-anonymat contraint utilise des limites de

généralisation, pour de nombreux ensembles de données initiaux, la suppression doit être utilisée afin de générer des données masquées k-anonymes contraintes.[44]

2.11 Conclusion

Nous avons présenté dans ce chapitre des notions théoriques sur la protection de l'anonymat, En commençant par une définition avec les principes et les différents niveaux de protection dans la vie privée ainsi que quelques attaques et enfin nous avons terminé par les technologies et les méthodes permettant la protection de la vie privée, en finissant par les avantages et les inconvénients de chaque méthodes.

Chapitre 3

Proposition et implémentation

3.1 Introduction

Un énorme intérêt pour la confidentialité des données a été suscité récemment au sein du public et des médias [49], ainsi que dans la communauté des chercheurs. De nombreux efforts de recherche ont été dirigés vers la recherche de méthodes pour anonymiser les ensembles de données, Ces méthodes envisagent également de minimiser une ou plusieurs mesures de coût entre les micro-données initiales et publiées (un ensemble de données où chaque tuple correspond à une entité individuelle).

La publication de données sur des individus sans révéler d'informations sensibles à leur sujet est un problème important. Ces dernières années, une nouvelle définition de la vie privée appelée k-anonymat a gagné en popularité. Dans un ensemble de données k-anonymisé, chaque enregistrement est indiscernable d'au moins k 1 autres enregistrements en ce qui concerne certains attributs « d'identification ».

Dans ce chapitre, nous introduisons le modèle k-anonymat et nous montrons avec deux contraint que l'ensemble de données k-anonymisé présente des problèmes de confidentialité subtils mais graves.

Tout d'abord, nous montrons qu'un attaquant peut découvrir les valeurs d'attributs sensibles lorsqu'il y a peu de diversité dans ces attributs sensibles. Deuxièmement, les attaquants ont souvent des connaissances de base, et nous montrons que k-anonymat ne garantit pas la confidentialité contre les attaquants utilisant des connaissances de base. Nous donnons une analyse détaillée de ces deux attaques et nous proposons une nouvelle et puissante définition de la confidentialité appelée l-diversité. En plus nous montrons

dans une évaluation expérimentale que l-diversité est pratique et peut être mise en œuvre efficacement.

3.2 Le k-anonymat et les contraints

De nombreuses organisations publient de plus en plus de micro-données des tableaux qui contiennent des informations non agrégées sur les personnes. Ces tableaux peuvent comprendre des données médicales, des données d'inscription sur les listes électorales, des données de recensement et des données sur les clients. Les micro-données sont une source d'information précieuse pour l'allocation des fonds publics, la recherche médicale et l'analyse des tendances. Toutefois, si les micro-données permettent d'identifier des individus de manière unique, des informations privées les concernant (comme leur état de santé) sont alors divulguées, ce qui est inacceptable. Pour éviter l'identification des enregistrements dans les micro-données, les informations d'identification uniques comme les noms et les numéros de sécurité sociale sont supprimées du tableau. Cependant, cette première épuration ne garantit toujours pas la confidentialité des données des personnes concernées.

Les ensembles d'attributs (comme le sexe, la date de naissance et le code postal dans l'exemple ci-dessus) qui peuvent être liés à des données externes pour identifier de manière unique des individus dans la population sont appelés "quasi-identifiants". Pour contrer les attaques de couplage utilisant des quasi-identifiants, Samarati et Sweeney ont proposé une définition de la vie privée appelée anonymat [54, 55]. Une table satisfaite à l'anonymat si chaque enregistrement de la table est indiscernable d'au moins k-1 autres enregistrements en ce qui concerne tout ensemble d'attributs de quasi-identifiants ; une telle table est appelée k-Anonymous table.

Par conséquent, pour chaque combinaison de valeurs des quasi-identifiants dans la table k-anonyme, il existe au moins k enregistrements qui partagent ces valeurs. La table 3.1 montre les dossiers médicaux d'un hôpital psychiatrique situé dans le nord de l'État de New York. Notez que le tableau ne contient aucun attribut permettant d'identifier de manière unique le nom de famille, le numéro de sécurité sociale, etc. Dans cet exemple, nous divisons les attributs en deux groupes : les attributs sensibles (qui ne concernent que l'état de santé) et les attributs non sensibles (code postal, âge et nationalité).

Un attribut est marqué comme sensible si un adversaire ne doit pas être autorisé à dissimuler la valeur de cet attribut pour un individu de l'ensemble de données. Les attributs qui ne sont pas marqués comme sensibles ne sont pas sensibles. En outre, la collection d'attributs (code postal, âge, nationalité) doit être le quasi-identifiant de cet ensemble de données. La table 3.2 présente un tableau à 4 anonymes dérivé du tableau 3.2 (ici, "*" indique une valeur supprimée, de sorte que, par exemple, "code postal =1485*" signifie que le code postal se situe dans la plage [14850-14859] et "Age=3*" signifie que l'âge se situe dans la plage [3039]).

En raison de sa simplicité conceptuelle, le k-anonymat a été largement discuté comme une définition viable de la publication des données personnelles, et en raison des progrès algorithmiques dans la création de versions anonymes d'un ensemble de données [50], le k-anonymat a gagné en popularité. Cependant, l'anonymat garantit-il vraiment la vie privée ? Par la suite, nous montrerons que la réponse à cette question est curieusement non.[51] Nous donnons des exemples de deux attaques simples, mais subtiles, contre un ensemble de données anonymes qui permettent à un attaquant d'identifier des enregistrements individuels. Pour se défendre contre ces attaques, il faut renforcer la notion de vie privée, ce que nous appelons -diversité.[52]

Mais nous Montrons d'abord les deux attaques pour donner l'intuition derrière les problèmes liés à l'anonymat.

	non-sensitive			sensitive
	code	**Age**	**Nationalité**	**Condition**
1	10000	28	Algérienne	Cardiopathie
2	12000	29	Américain	infection virale
3	12000	21	Française	infection virale
4	10000	23	Américain	Cardiopathie
5	18000	50	Japonaise	Cancer
6	18000	55	Algérienne	Cardiopathie
7	19000	47	Américain	infection virale
8	19000	49	Américain	infection virale
9	06000	31	Américain	Cancer
10	06000	37	Japonaise	Cancer
11	15000	36	Française	Cancer
12	15000	35	Américain	Cancer

TABLE 3.1 – Microdonnées des patients hospitalisés

	non-sensitive			sensitive
	code	**Age**	**Nationalité**	**Condition**
1	100**	< 30	*	Cardiopathie
2	120**	< 30	*	infection virale
3	120**	< 30	*	infection virale
4	100**	< 30	*	Cardiopathie
5	1800*	>= 40	*	Cancer
6	1800*	>= 40	*	Cardiopathie
7	1900*	>= 40	*	infection virale
8	1900*	>= 40	*	infection virale
9	060**	3*	*	Cancer
10	060**	3*	*	Cancer
11	150**	3*	*	Cancer
12	150**	3*	*	Cancer

TABLE 3.2 – 4-anonymes sur les Microdonnées patients hospitalisés

3.2.1 Attaques sur k-anonymat

Nous présentons deux attaques, l'attaque d'homogénéité et l'attaque de connaissance d'arrière-plan, et nous montrons comment elles peuvent être utilisées pour compromettre un ensemble de données k-anonyme.

3.2.2 Attaque d'homogénéité

Manel et Ahmed sont des voisins antagonistes. Un jour, Ahmed tombe malade et est emmené en ambulance à l'hôpital. Après avoir vu l'ambulance, Manel part à la découverte de la maladie dont souffre Ahmed. Manel découvre le tableau anonyme des 4 dossiers d'hospitalisation actuels publié par l'hôpital (table 3.2), et elle sait donc que l'un des dossiers de ce tableau contient les données de Ahmed. Comme Manel est la voisine de Ahmed, elle sait que Ahmed est un Américain de 31 ans qui vit sous le code postal 13053. Manel sait donc que le numéro d'enregistrement de Ahmed est 9, 10, 11 ou 12, et que tous ces patients ont la même condition médicale (cancer), ce qui amène Manel à conclure que

Ahmed a un cancer.[53]

Observation 1

k-Anonymat peut créer des groupes qui fuient des informations en raison du manque de diversité dans l'attribut sensible .Notez qu'une telle situation n'est pas rare.
Supposons que nous disposions d'un ensemble de données contenant 60 000 tuples distincts et que l'attribut sensible puisse prendre trois valeurs distinctes et ne soit pas corrélé avec les attributs non sensibles. Une anonymisation de ce tableau aura environ 12 000 groupes et, en moyenne, 1 groupe sur 81 n'aura aucune diversité (les valeurs de l'attribut sensible seront toutes les mêmes). On devrait donc s'attendre à 148 groupes sans diversité.[54]
Par conséquent, les informations sur 740 personnes seraient compromises par une attaque d'homogénéité. Cela suggère qu'en plus d'un anonymat symbolique, le tableau analysé devrait également garantir la "diversité"[55,56] - tous les tuples qui partagent les mêmes valeurs de leurs quasi-identifiants devraient avoir des valeurs diverses pour leurs attributs sensibles. Notre prochaine observation est qu'un adversaire pourrait utiliser des connaissances "de base" pour découvrir des informations sensibles.
Attaque de connaissances en arrière-plan : Manel a un correspondant nommé Amel qui est admis dans les mêmes hôpitaux que AHMED, et dont les dossiers de patients apparaissent également dans le tableau de la table 3.2. Alice sait que Amel est une jeune Japonaise de 21 ans qui vit actuellement sous le code postal 13068. Sur la base de ces informations, Manel apprend que les renseignements concernant Amel figurent dans le dossier numéro 1, 2, 3 ou 4. Sans informations supplémentaires, Manel n'est pas sûre que Amel ait attrapé un virus ou souffre d'une maladie cardiaque. Cependant, il est bien connu que les Japonais ont une incidence extrêmement faible de maladies cardiaques. Manel conclut donc avec une quasi-certitude que Amel est atteint d'une infection virale.

Observation 2

k-Anonymat ne protège pas contre les attaques basées sur les connaissances de base

3.3 Matériels et méthodes

3.3.1 Définitions de base

A. Attributs sensibles

Peuvent également être nommés en tant qu'identificateurs clés en fonction de l'ensemble de données. Cela nous indique que les informations telles que le nom, le numéro SSN, l'adresse électronique, l'état de santé . . . etc.[57]

B. Attribut de quasi-identification

Un ensemble de quasi-identifiants est un ensemble minimal d'attributs dans la table T qui peut être associé à des informations externes pour désidentifier des enregistrements individuels.[57]

C. Attributs non sensibles

Ces types de données ne sont pas inclus avec d'autres attributs ne fuient pas ou n'indiquent pas d'informations sensibles comme de simples dossiers médicaux, une profession commune.[57]

3.3.2 Méthodologie

L'objectif de l'expérience est de préserver la vie privée des données . on utilisent l'algorithme de k_ anonymat ensuite on essaie de remédier les attaque du k anonymat avec un autre algorithme les l-diversité, en utilisant python en comparant les résultats anonymizé des deux algorithmes .
Les objectifs de cette étude sont visés par la méthodologie suivante en 6 étapes :
1. Importer dataset originale (adulte).
2. Déterminer les quas_identifiants et les attributs sensible
3. Choix d'un algorithme /technique(K-anonymat /l-diversité).
4. Déterminer les valeurs d'entrées .
5. Evaluation des résultats .
6. Exporter les données anonymiser .

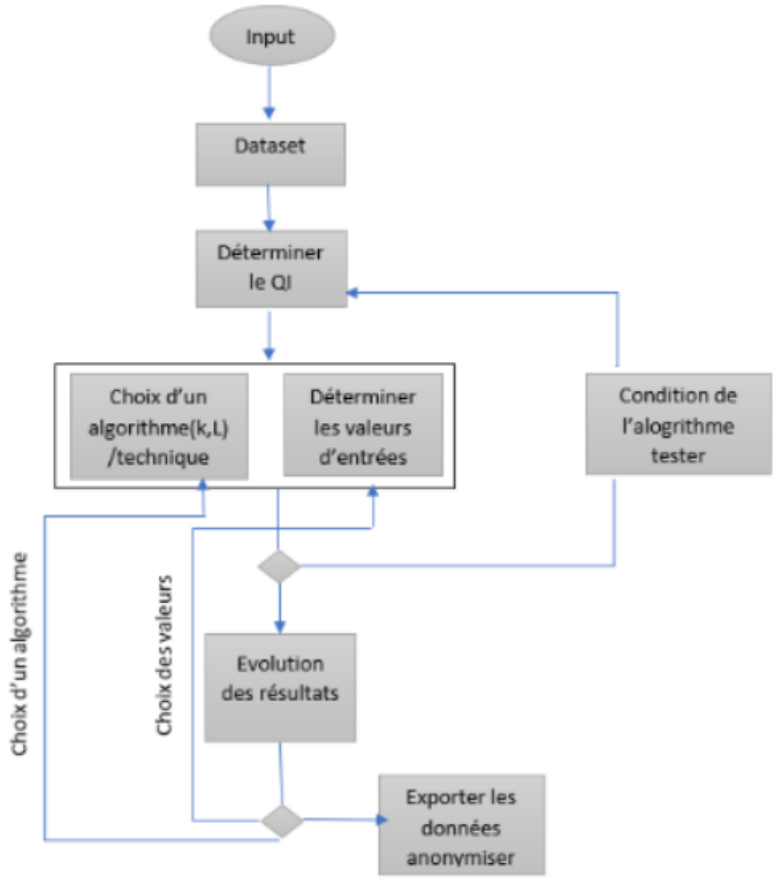


FIGURE 3.1 – L'architecture du système de base de l'algorithme généralisé

3.4 Environnement de travail et données

3.4.1 Environnement matériel

Pour la réalisation de ce projet on a disposé de :

Caractéristiques	**PC1 : DELL**	**PC2 :TOSHIBA**
Disque dur	128 GO SSD	500 GB
Type de processeur	Intel® core™ i5-7300U CPU @ 2.60GHz 2.71GHz	Intel® core™ i3-4005U CPU @ 1.70GHz 1.71GHz
Fréquence de processeur	2.71GHz	1.71 GHz
Mémoire mort S.E	Core i5 Windows 10	Core i3 Windows 8

TABLE 3.3 – Environnement matériel utilisées

3.4.2 Langage utilisé

Python

Python est un langage de programmation généraliste et très riche avec de nombreuses fonctionnalités. C'est l'un des principaux objectifs d'un style de programmation connu sous le nom de programmation orientée objet[58]. Il s'agit d'un langage de programmation interprété de haut niveau [59].

3.4.3 Plateforme et environnement de développement

Anaconda

Est une plateforme scientifique de données open source qui rassemble les meilleurs outils pour la science des données. Il s'agit d'une pile de données scientifiques qui comprend plus de 100 paquets populaires basés sur Python, Scala et R. Avec l'aide de son gestionnaire de paquets, conda, les utilisateurs peuvent travailler avec des centaines de paquets dans différents langages et effectuer facilement le prétraitement, la modélisation, le regroupement, la classification et la validation des données[59].

Jupyter Notebook Est un outil populaire pour écrire du code Python. Basées sur le Web, ce qui signifie que lorsque Jupyter s'ouvre, il le fait dans votre navigateur Web par défaut, qui peut être Google Chrome, Pale Moon, Edge ou Internet Explorer[60].

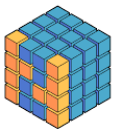
3.4.4 Bibliothèque Utilisés

Panda

Pandas est une bibliothèque Python open source pour l'analyse de données hautement spécialisées[61].

Matplotlib

Est la bibliothèque Python qui est actuellement la plus populaire pour produire des tracés et autres visualisations de données en deux dimensions. Comme l'analyse des données nécessite des outils de visualisation, elle est très répandue dans les milieux scientiques et techniques[61]



Numpy

Est une bibliothèque Python open source pour le calcul scientique. NumPy vous permet de travailler avec des tableaux et des matrices de manière naturelle. La bibliothèque contient une longue liste de fonctions mathématiques utiles, dont certaines pour l'algèbre linéaire, la transformation de Fourier et les routines de génération de numéros[63].

3.4.5 Présentation des ensembles de données

Le dataset adulte [64] utilisé dans notre projet. Il a une étiquette binomiale indiquant un salaire de $< 50K$ ou $> 50K$ USD. Dans cet ensemble de données, 76% des enregistrements ont une étiquette de classe de $< 50K$. Il y a un total de 14 attributs, dont 8 catégoriques et 6 continus. La classe d'emploi représente le type d'emploi tel qu'indépendant, bureau privé ou fédéral et la profession décrit le type d'emploi tel que l'agriculture, les travaux des champs, les commis ou dans la gestion.
Le niveau d'études mentionne le plus haut niveau d'éducation atteint, comme une licence, une maîtrise ou un doctorat. L'attribut relation a des catégories telles que célibataire ou mari et l'état civil a des catégories telles que marié ou séparé. Les autres attributs nominaux sont le pays de résidence, le sexe et la race. Les personnes ayant des caractéristiques démographiques similaires doivent avoir un poids similaire, etc.

3.5 Implémentation

3.5.1 La Mise en œuvre de k anonymat

la publication de données préservant la vie privée avec le k_anonymat :
L'algorithme procède alors comme suit pour partitionner les données en groups k-anonymes :
1. Initialiser l'ensemble fini de partitions en un ensemble vide p_fini=. P0=
2. Initialiser l'ensemble de travail des partitions en un ensemble contenant une partition avec l'ensemble complet des données p_travail = 1,2,. . . ,N. P=all
3. Tant qu'il y a des partitions dans l'ensemble de travail, faites-en sortir une et while p(all)
Calculez les portées relatives de toutes les colonnes de la partition.
— Triez les colonnes résultantes en fonction de leur portée (par ordre décroissant) et répétez l'opération sur ces colonnes. Pour chaque colonne,
—Essayez de diviser la partition le long de cette colonne en utilisant la médiane des valeurs de la colonne comme point de séparation.
***** Vérifiez si les partitions résultantes sont valables selon nos critères d'anonymat k (et éventuellement supplémentaires).
***** Si oui, ajoutez les deux nouvelles partitions à l'ensemble de travail et sortez de la boucle.
***** Si aucune colonne n'a produit une division valide, ajoutez la partition originale à l'ensemble des partitions finies.
4. Retournez l'ensemble des partitions terminées.

3.5.2 Les résultats

Dans cette partie, nous explorerons l'algorithme qu'on a défini ci-avant, qui utilise un algorithme de recherche gourmande pour partitionner les données d'origine en groupes de plus en plus petits.
En premier lieu, nous commençons par l'anonymisation
1– publient les données de data set dans un tableau comme suit :

I.2 - Affichage des données

```
df.head()
```

	age	classe-de-travail	poids-final	éducation	num-éducation	état civil	Occupation	relation	course	sexe	gain en capital	perte en capital	heures par semaine	pays d-origine	revenu
0	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50k
1	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	<=50k
2	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	<=50k
3	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50k
4	37	Private	284582	Masters	14	Married-civ-spouse	Exec-managerial	Wife	White	Female	0	0	40	United-States	<=50k

2– mettait en œuvre une fonction qui renvoie les intervalles (Max-Min pour les colonnes numériques, nombre de valeurs différentes pour les colonnes catégorielles) de toutes les colonnes pour une partition d'un data frame(trame de données).

```
{'age': 73,
 'classe-de-travail': 9,
 'poids-final': 1478115,
 'éducation': 16,
 'num-éducation': 15,
 'état civil': 7,
 'Occupation': 15,
 'relation': 6,
 'course': 5,
 'sexe': 2,
 'gain en capital': 99999,
 'perte en capital': 4356,
 'heures par semaine': 98,
 'pays d-origine': 42,
 'revenu': 2}
```

3– mettait en œuvre une fonction de division qui prend un tram de données, une partition et une colonne et renvoie deux partitions qui divisent la partition donnée de telle sorte que toutes les lignes dont les valeurs de la colonne sont inférieures à la médiane se trouvent dans une partition et toutes les lignes dont les valeurs sont supérieures ou égales à la médiane se trouvent dans l'autre Maintenant que toutes les fonctions d'assistance

sont en place, nous pouvons implémenter l'algorithme décrit ci-dessus :
On implémente l'algorithme de partitionnement décrit ci-dessus, en utilisant un critère k-anonyme pour les partitions créez. Ensuite nous essayons ceci sur notre jeu de données ! Pour simplifier les choses, nous sélectionnerons d'abord seulement deux colonnes de l'ensemble de données auquel nous appliquons le partitionnement aussi on affiche le nombre de partitionnements existe. Cela facilite la visualisation du résultat et accélère.

```
500
```

— Générer un ensemble de données k-Anonyme :bien sûr, pour utiliser les données, nous voulons produire un nouvel ensemble de données contenant une ligne pour chaque partition et la valeur de l'attribut sensible. Pour ce faire, nous devons agréger les colonnes de chaque partition.

```
Finished 1 partitions...
Finished 101 partitions...
Finished 201 partitions...
Finished 301 partitions...
Finished 401 partitions...
```

Nous trions les trames de données résultant en utilisant les caractéristiques de colonne et l'attribut sensible. Voici le résultat dans la figure

	age	num-éducation	revenu	count
469	17.0	3.000000	<=50k	3
615	17.0	4.000000	<=50k	5
110	17.0	5.000000	<=50k	36
111	17.0	6.000000	<=50k	198
0	17.0	7.200599	<=50k	334
...	...	...	...	...
746	90.0	9.000000	>50k	4
750	90.0	10.545455	<=50k	9
751	90.0	10.545455	>50k	2
818	90.0	14.000000	<=50k	2
819	90.0	14.000000	>50k	3

822 rows × 4 columns

Après avoir appliqué le fameux k_anonymat on remarque que les donnes sont anony-

misées et que l'analyse des données continue de fournir des résultats exacts. Cependant le k-anonymat, appliqué à une table, peut mener à la création de classes d'équivalences qui laissent échapper des informations sensibles en raison du manque de diversité dans les valeurs de l'attribut sensible.Pour pallier cet inconvénient (comme mentionné ci-avant) , on applique la méthode l-diversité.

3.5.3 Mettre en œuvre la l-diversité (la voie naïve)

Voyons maintenant comment nous pouvons mettre en œuvre la diversité afin de protéger encore mieux la vie privée des personnes concernées par l'ensemble des données. Pour mettre en œuvre la diversité, nous pouvons faire les choses suivantes :

A. Modifier notre fonction de k anonymat (is_valid) pour non seulement vérifier la taille d'une partition donnée, mais aussi s'assurer que les valeurs de l'attribut sensible de la partition est suffisamment diverses.

suivantes :

B. Modifier la fonction split pour produire des splits qui sont divers (si possible)

```
309
```

Implémenter une fonction de validation qui renvoie Vrai si une partition donnée contient au moins les valeurs différentes de l'attribut sensible, sinon Faux. encore une fois, nous construisons un ensemble de données anonymisé à partir des partitions l -Diverses

```
Finished 1 partitions...
Finished 101 partitions...
Finished 201 partitions...
Finished 301 partitions...
```

Voyons comment la diversité améliore l'anonymat de notre ensemble de données :

	age	num-éducation	revenu	count
0	17.706107	7.248092	<=50k	785
1	17.706107	7.248092	>50k	1
114	18.341463	3.365854	<=50k	40
115	18.341463	3.365854	>50k	1
4	19.320276	10.000000	<=50k	1301
...	...	...	...	...
587	89.727273	13.000000	>50k	2
574	90.000000	10.545455	<=50k	9
575	90.000000	10.545455	>50k	2
614	90.000000	14.000000	<=50k	2
615	90.000000	14.000000	>50k	3

618 rows × 4 columns

Dans l'ensemble, nous pensons que L-diversity est pratique, facile à comprendre, et qu'il répond aux lacunes du k-anonymat en ce qui concerne les connaissances de base et les attaques d'homogénéité.

3.6 Conclusion

dans ce chapitre, nous avons mentionné le langage de développement utilisé pour implémenter la solution proposée.Comme nous l'avons montré, qu'une base de données(dataset) k-anonymisée permet de fortes attaques en raison du manque de diversité des attributs sensibles.Nous avons proposé la diversité, et ce cadre offre une garantie plus forte pour la protection de la vie privée.

Conclusion générale et Perspectives

Le big data est apparu comme une réponse au problème du temps et du traitement des données. d'énormes quantités de données ont été réduites pour prendre peu de place lors de stockage . Ces données sont mises en relation et des calculs sont effectués grâce aux ordinateurs,(l'ENIAC et l'ABC se disputent la première place) et aux logiciels programmes.

Dans un premier temps, on a présenté une généralité sur la préservation de la confidentialité pour les données publiées. Au début, on a commencé par citer les définitions importantes qui concerne l'approche de l'anonymisation, les différents types, et les opérations sur les quelles, cette approche va se dérouler parfaitement.

Les mégadonnées peuvent avoir de nombreuses origines et des formes très différentes, et ces données sont vulnérables à la protection de la vie privée. Pour faire face à ces problèmes, il existe de nombreux algorithmes sont mis en œuvre en utilisant ces grands outils de données pour garantir la qualité, l'extensibilité et la sécurité. Les chercheurs ont déjà mis en œuvre divers algorithmes d'anonymisation sur différentes plateformes.
Dans Cette étude Nous avons considéré un ensemble de données « Adult » et nous avons exécuté l'algorithme K-Anonymat et L-Diversité avec des valeurs K=3, L¿=2 sur le ensemble de données afin de privatiser les données autant que possible avec la plus grande efficacité.

Il existe plusieurs pistes de travail pour l'avenir. Tout d'abord, nous voulons trouver la solution contre l'attaque de connaissances en arrière-plan et nous voulons développer des méthodes pour les attributs sensibles continus. Deuxièmement, bien que la vie privée et l'utilité soient deux choses distinctes, la vie privée a fait l'objet d'une attention beaucoup plus grande que l'utilité d'un tableau publié pour cela nous essayons de l'utilisé dans de nombreux domaines, par exemple l'utilisation de données anonymes pour l'apprentissage Automatique qui préserve la confidentialité.

Bibliographie

[1] [BRASSEUR (C.). – Enjeux et usages du big data. Technologies, méthodes et mises en oeuvre, Paris, Lavoisier, p. 30 (2013)] consulté :11 mars 2020

[2] [HELBING (D.) et POURNARAS (E.). – Build Digital Democracy : Open Sharing of Data that are Collected with Smart Devices would Empower Citizens and Create Jobs.Nature, Vol.527, Nov. 2015, Macmillan Publishers (2015)] consulté : 20 mars 2020

[3] https ://www.researchgate.net/publication/279848651-Rapport-sur-le-Big-Data consulté :23 mars 2020

[4] https ://2buseco.blogspot.com/2016/07/cest-quoi-le-big-data.html consulté :23 mars2020

[5] https ://www.oracle.com/ca-fr/big-data/what-is-big-data.html

[6] http ://www.gfii.fr/uploads/docs/BigDatasynthF.pdf consulté :23 mars 2020

[7] http ://big-data-iscomwiz.e-monsite.com/iscomwiz/le-big-data/les-5-v-du-bigdata-a-connaitre.html consulté :16 mars 2020

[8] https ://www.filfil.eu/2019/05/13/le-big-data/.consulté le 15/10/2020

[9] http ://big-data-iscomwiz.emonsite.com/iscomwiz/le-big-data/les-5-v-du-big-data-a-connaitre.html conculté 11mars 2020

[10] Seref SDJLURJOX Duygu Sinanc Terzi, Ramazan Terzi. A Survey on Security and Privacy Issues in Big Data, Gazi University, Computer Engineering Ankara, memoire master. 2015 consulté :23 mars 2020

[11] https ://statswiki.unece.org/display/bigdata/Classification+of/Types/of/Big+Data consulté le 26-10-2020.

[12] https ://data-flair.training/blogs/big-data-applications-various-domains/, consulté le 26-10-2020.

[13] H6040 V1 Introduction au Big Data - Opportunités, stockage et analyse des mégadonnées par Bernard ESPINASSE, Patrice BELLOT consulté :23 mars 2020

[14] KAUVALI (B.), KNOTT (D.) et VAN KUIKEN (S.). – The big data revolution in helthcare consulté :23 mars 2020

[15] GINSBERG (J.) et al. – Detecting influenza epidemics using search engine query data. Nature, n 457, pp. 1012-1014 (2009). consulté :20 mars 2020

[16] MORENO (M.) et al. – Associations between displayed alcohol references on Facebook and problem drinking among college students, Archives of Pediatrics Adolescent Medicine, 166 (2), pp. 157-163 (2012). consulté :10 mars 2020

[17] HAMEL (M.-P.) et MARGUERIT (D.). – Analyse des big data usages, quels défis ?Note d'analyse du Commissariat général à la stratégie et à la prospective, N 8, nov. 2013 (2013). consulté :23 mars 2020

[18] MOTHE (J.), PITARCH (Y.) et GAUSSIER (E.).– Big Data : Le cas des systèmes d'information,Revue Ingénierie des Systèmes d'Information, Hermès Editeur, Vol.´ 19/3 (2014). consulté :23 mars 2020

[19] JOUNIAUX (P.). – Big data au service de la sécurité du transport à erien : l'analyse des données de vol, Télécom, n 169, juillet (2013). consulté :6 mars 2020

[20] https ://blog.arcoptimizer.com/securite-des-donnees-authentification-autorisation-et-cryptage consulté :6 mars 2020

[21] Abdoul Seck. Sécurité et Big Data : 4 briques a mettre en place pour sécuriser votre projet, CCM Benchmark.consulté :6 mars 2020

[22] S. Tuffery, Cours de Data Mining , université de Rennes 1, 2014. consulté :15 février 2020

[23] D. Gaultier, Data Science Big Data – Etat de l'art , 2015. consulté :20 mars 2020

[24] Mr MATALLAH Houcine. Vers un nouveau modèle de stockage et d'accès aux données dans les Big Data et les Cloud Computing, UNIVERSITE ABOU-BEKR BELKAID - TLEMCEN. 2018 consulté :23 mars 2020

[25] https ://www.cea.fr/comprendre/pages/nouvelles-technologies/l-essentiel-sur-le-big-data.aspx Consulté le 8 mars 2020.

[26] https ://www.cea.fr/comprendre/pages/nouvelles-technologies/l-essentiel-sur-le-bigdata.aspx ? consulté 20 juin 2020

[27] https ://www.fifosys.com/blog/security/privacy-or-anonymity-which-is-moreimportant-in-the-digital-age/ consulté le 10/05/2020

[28] https ://www.fifosys.com/blog/security/privacy-or-anonymity-which-is-more-important-in-the-digital-age/ consulté le 10/05/2020

[29] https ://www.privateinternetaccess.com/blog/how-does-privacy-differ-from-anonymity-and-why-are-both-important/ : :text=Privacy Consulté le 05/05/2020

[30] https ://tel.archives-ouvertes.fr/tel-01783967/document consulté le 15/6/2020

[31] Techniques d'anonymisation,Benjamin NGUYEN Insa1 Centre Val de Loire et Inria2 Paris-Rocquencour/ss.pdf consulté le 10/05/2020

[32] Li N, et al. t-Closeness : privacy beyond k-anonymity and L-diversity. In : Data engineering (ICDE) IEEE 23rd international conference ; consulté le 10/05/2020.

[33] Ton A, Saravanan M. Ericsson research. [Online]. http ://www.ericsson.com/research-blog/data-knowledge/ big-data-privacy-preservation/2015. consulté le 14/05/2020.

[34] Samarati P, Sweeney L. Protecting privacy when disclosing information : k-anonymity and its enforcement through generalization and suppression. Technical Report SRI-CSL-98-04, SRI Computer Science Laboratory ; 1998. consulté le 14/05/2020.

[35] Sweeney L. K-anonymity : a model for protecting privacy. Int J Uncertain Fuzz. 2002 ;10(5) :557–70. consulté le 14/05/2020.

[36] Meyerson A, Williams R. On the complexity of optimal k-anonymity. In : Proc. of the ACM Symp. on principles of database systems. 2004.consulté le 14/05/2020.

[37] https ://www.ijedr.org/papers/IJEDR1702165.pdf consulté le 20/05/2020.

[38] Bredereck R, Nichterlein A, Niedermeier R, Philip G. The effect of homogeneity on the complexity of k-anonymity. In :FCT ; 2011. p. 53–64.consulté le 23/05/2020.

[39] Microsoft differential privacy for everyone, [online]. 2015. http ://download.microsoft.com/. . . /Differential$_{P}rivacy_{f}or_{E}veryone.pdfconsultéle$23/05/2020.

[40] Samarati P. Protecting respondent's privacy in microdata release. IEEE Trans Knowl Data Eng. 2001 ;13(6) : consulté le 23/05/2020.

[41] Xuyun Zhang,Chi Yang, Surya Nepal (2013) A MapReduce Based Approach of Sclable Multidimentional Anonymization for Big Data Privacy Preservation on Cloud ,IEEE Third International Conference on Cloud and Green Computing consulté 23/05/2020.

[42] Research India Publications : : : http ://www.ripublication.com,A Review of Big Data and Anonymization Algorithms consulté 23/05/2020

[43] Mohamed R.Fouad, Khaled Elbassioni, and Elisa Bertino, A Super Modularity Based Differential Privacy Preserving Algorithm for Data Anonymization IEEE Transactions on Knowledge and Data Engineering Vol.26, consulté 23/05/2020

[44] L. Sweeney, Achieving k-Anonymity Privacy Protection Using Generalization and Suppression, International Journal on Uncertainty, Fuzziness, and Knowledge-based Systems, Vol. 10, No. 5 (2002), pp. 571–588 consulté le 20/10/2020

[45] V. Iyengar, Transforming Data to Satisfy Privacy Constraints, in Proc. of the ACM SIGKDD (2002), pp. 279–288 .consulté le 20/10/2020

[46] K. LeFevre, D. DeWitt, and R. Ramakrishnan, Mondrian Multidimensional K-Anonymity, in Proc. of the IEEE ICDE (2006), pp. 25. consulté le 20/10/2020

[47] P. Samarati, Protecting Respondents Identities in Microdata Release, IEEE Transactions on Knowledge and Data Engineering, Vol. 13, No. 6 (2001), pp. 1010–1027. consulté le 20/10/2020

[48] M. Lunacek, D. Whitley, and I. Ray, A Crossover Operator for the k-Anonymity Problem, in Proc. of the GECCO (2006) pp. 1713–1720.consulté le 20/10/2020

[49] https ://www.cnil.fr/fr/lanonymisation-des-donnees-un-traitement-cle-pour-lopen-data consulté le 12/9/2020

[50] G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy, D. Thomas, and A. Zhu. k-anonymity : Algorithms and hardness. Technical report, Stanford University, 2004.consulté le 18/8/2020

[51] R. J. Bayardo and R. Agrawal. Data privacy through optimal kanonymization. In ICDE-2005, 2005.consulté le 19/8/2020

[52]] K. LeFevre, D. DeWitt, and R. Ramakrishnan. Incognito : Efficient fulldomain k-anonymity. In SIGMOD, 2005.consulté le 19/8/2020

[53]] A. Meyerson and R. Williams. On the complexity of optimal kanonymity. In PODS, 2004.consulté le 19/8/2020

[54] P. Samarati. Protecting respondents' identities in microdata release. In IEEE Transactions on Knowledge and Data Engineering, 2001 consulté le 19/8/2020

[55] L. Sweeney. k-anonymity : a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10(5) :557–570, 2002. consulté le 19/8/2020

[56] S. Zhong, Z. Yang, and R. N. Wright. Privacy-enhancing kanonymization of customer data. In PODS, 2005.consulté le 19/8/2020

[57] https ://core.ac.uk/download/pdf/211298232.pdf consulté le 19/6/2020

[58] https ://medium.com/datadriveninvestor/python-programming-language-ac762a3b5977 consulté le 01/6/2020

[59] https ://www.python.org/doc/essays/comparisons/ consulté le 18/6/2020

[60] https ://pandas.pydata.org/docs/pandas.pdf consulté le 18/6/2020

[61] https ://towardsdatascience.com/top-6-python-libraries-for-visualization-which-one-to-use-fe43381cd658 ?gi=42576256e867 consulté le 18/6/2020

[62] https ://towardsdatascience.com/numpy-the-king-of-scientific-computing-with-python-d1de680b811d consulté le 18/6/2020

[63] https ://www.kaggle.com/wenruliu/adult-income-dataset consulté le 12/4/2020