République Algérienne Démocratique et polpulaire
Ministère de L'Enseignement Supérieur et de la Recherche Scientifique
Université de Larbi Tébessi- Tébessa-

Faculté des sciences Exactes et des sciences de la nature et de la vie
Département : Mathématique et Informatique

MEMOIRE DE MASTER

Domaine : Math et Informatique

Filière : Informatique

Option : Systéme d'information

Thème :

Une approche d'extraction des connaissances à partir des BIG DATA

Présenté par :

BEKKAI BESMA

ZIAD SAMIR

Devant le Jury :

M.Derdour	M.C.B	Université de Tébessa	Président
L.Bradji	M.C.A	Université de Tébessa	Encadreur
M.Amroune	M.A.A	Université de Tébessa	Examinateur

Date de soutenance : 25/05/2017

Note : .....

Mention : .....

Résumé

La notion de **Big Data** est un concept qui a pu être popularisé à partir 2012 et ce dans un but d'exprimer essentiellement le fait que les entreprises sont aujourd'hui devant un problème celui des volumes de données qu'il faut savoir traiter et avec une croissance rapide et considérable . Ces volumes de données massifs engendrent alors une évolution fulgurante de modèles technologiques possédant l'évolution nécessaires qui permet d'accéder à des nouvelles opportunités. De nouvelles techniques sont proposées dans un but d'améliorer le stockage et le traitement de ces données massives tel que le projet **Apache Hadoop** .Ces mêmes données ne peuvent être traitées qu'avec une gestion de base de données classiques c'est la raison pour laquelle ont été crées les bases de données **NOSQL(Not Only SQL)** .

Dans ce mémoire, on a établi un état de l'art impliquant l'extraction de connaissances à partir de données **(ECD)** définie comme un processus de découverte d'informations implicites, inconnues auparavant et potentiellement utiles à partir des données. Ce processus s'effectue en plusieurs étapes : préparation des données (recherche, nettoyage), fouille des données (recherche d'un modèle de connaissances), validation et interprétation du résultat et enfin intégration des connaissances apprises.

Enfin, nous avons travaillé sur la conception d'une architecture basée sur le nettoyage des données structurées et non structurées afin d'extraire les connaissances à partir des big data.

Mots clés : Big Data, Hadoop,NOSQL , ECD, fouille de données.

Abstract

The concept of Big Data is a concept that could be popularized from 2012, with the aim of expressing essentially the fact that companies are today face a problem of the volumes of data that must be treated and with Rapid and considerable growth. These massive volumes of data then lead to a rapid evolution of technological models with the necessary evolution that allows access to new opportunities. New techniques are proposed with the aim of improving the storage and processing of this massive data, such as the Apache Hadoop project. These same data can only be processed with a standard database management, which is why NOSQL (Not Only SQL) databases were created.

In this thesis, a state of the art involving knowledge extraction from data (ECD) has been established which is defined as a process of discovery of implicit information previously unknown and potentially useful from data. This process is carried out in several stages: data preparation (research, data cleaning and coding), data mining (search for a knowledge model), validation and interpretation of the result, and integration of the knowledge learned.

Finally, we build an architecture based on the cleaning of structured and unstructured data in order to extract the knowledge from the big data.

Keywords: Big Data, Hadoop,NOSQL , KDD, Datamining.

ملخـــــص

كلمة **Big Data** هو مفهوم شاع سنة 2012 و ذلك لغرض التعبير عن حقيقة أن الشركات تواجه اليوم مشكلة الكم الهائل من البيانات على نحو متزايد لذلك من الضروري معرفة كيفية التعامل مع هذه الزيادة الكبيـــرة و السريعة للبيانات . هذه الأحجام و الكميات الضخمة من البيانات نتج عنها تطور سريع للنماذج التكنـــولوجية الذي يسمح بالوصول إلى فرص جديدة. وقد إقترحت تقنيات جديدة لتحسين تخزين و معالجة هذه البيانات الضخمة مثل مشروع **Apache Hadoop**، لا يمكن أن تعامل هذه البيانات مع إدارة قواعد البيانات التقليدية لهذا تم إنشاء قواعد البيانـــات **NOSQL** (ليس فقط **SQL**). في هذه المذكرة ، قمنا بإنشاء حوصلة من الفن التي تنطـــوي على إستخراج المعرفة من خلال البيانات **(E C D)** الذي يعرف بأنه عملية إكتشاف المعلومات الضمنية التي كانت غير معروفة في السابق و التي يمكن أن تكون مفيدة ، و تنفذ هذه العملية في عدة مراحل : إعداد البيانات (البحث عن البيانات و تنظيفها) ، استخراج البيانات (البحث عن نموذج للمعرفة) و التحقق من صحة النتائج و تفسيرها و إدماج المعرفة المكتسبة .

أخيرا ، إننا عملنا على تصميم نموذج يوضح عملية تنظيف البيانات المنظمة و الغير المنظمة و ذلك بهدف إستخراج المعارف من البيانات الكبيرة.

Dédicace

Je dédie ce présent travail à :

Mon cher père, pour ses efforts, sa bienveillance et son sacrifice, qu'il n'a jamais cessé de consentir pour ma réussite, mon bonheur et ma joie dans la vie.

Quoi que je fasse, je ne pourrais jamais te récompenser cher papa pour les grands sacrifices que tu as consenti et pour tous les efforts que tu va encore fournir pour ta fille. Aucune dédicace ne pourrait exprimer ma grande admiration, considération et ma sincère affection pour toi.

Ma chère mère, à qui je dois tout, pour ton soutien, tes conseils, ton admiration pour ta fille chérie et sans doute la personne qui ma toujours réconfortée et encouragée dans les moments difficiles dans ma scolarité et également pour ton amour pour moi : sans toi, je ne serais peut être jamais ce que je suis maintenant.

Je suis arrivée aujourd'hui à ce niveau grâce à vous (mes parents), votre patience et vos innombrables sacrifices. Qu'Allah, le Tout Puissant, vous préserve et vous procure la santé et la longue vie afin que je puisse faire des prières salvatrices de reconnaissances et de remerciements.

Mon grand frère unique, Abdelmalek et son épouse Fatma Zohra et ses chers enfants Abdelhak, Abderrahim, Abdennour et Amine.

Mes sœurs, Meriem et son mari Khaled, Sonia et son mari Abdelhakim et son nourrisson Abdelmouhaimene.

Mes professeurs ;

Mes chères amies ;

Mes collègues ;

Sans oublier mon binôme, Ziad Samir.

Tous ceux qui de prés ou de loin m'ont aidé à réaliser ce mémoire.

BEKKAI BESMA

Dédicace

A mes parents

A ma femme

A mes enfants Abdelmalek et Mohamed

A mes beaux parents

A mes frères et sœurs et leurs enfants

A mes beaux frères et belles sœurs et leurs enfants

A tous mes amis

A ma binôme BEKKAI BESMA

ZIAD Samir

Remerciement

Avant toute chose, nous tenons à prosterner devant Allah pour le remercier de nous avoir donné le courage, la patience et la volonté de faire ce travail et de l'achever en temps opportun. Nous tenons à remercier très très fort notre encadreur ***Dr.BRADJI LOUARDI*** *pour ses précieux conseils, sa patience et ses nobles valeurs humaines.*

Nos remerciements vont également aux membres du jury pour nous avoir honorés par l'évaluation de notre travail.

Merci aussi pour tous ceux que nous avons omis de citer ici et qui de prés ou de loin ont contribué au bon déroulement de ce travail.

Merci à vous tous !

Liste des Figures

Liste des Tableaux

TABLE DES MATIERES

TABLE DES MATIERES

TABLE DES MATIERES

TABLE DES MATIERES

Chapitre 3 : Contribution

Introduction générale

Au cours de ces dernières années, il a été constaté une augmentation fulgurante de la capacité à recueillir des données issues d'un ensemble varié de capteurs, appareils et en différents formats à partir d'applications indépendantes ou même connectées. L'apport de données étant tellement fort et important qu'il a dépassé au large les veilles technologiques et ce sur plusieurs plans à savoir le traitement, l'analyse, le stockage et surtout leurs compréhensions. Prenons par exemple des données sur internet, les pages web indexées par google étaient en 1998 prés d'un million, elles atteindront un milliard en 2000 pour devenir et même dépasser un trillion en 2008. Cet essor fulgurant accéléré encore plus par l'utilisation des réseaux sociaux comme facebook, twitter, etc…, va permettre aux usagers un ajout inconditionnel d'informations (textes, images, vidéos, etc…) sur le web et ceci va encore accentuer sa montée en puissance. En conséquence, nous pouvons dire pour déduction que le phénomène Big Data a changé d'une manière radicale, la manière de gérer des données car il introduit de nouvelles problématiques concernant la volumétrie, la vitesse de transfert et le type de données.

On doit savoir que le développement et la croissance assez rapide des données collectées dans les bases de données avec la nécessité d'une réactivité efficace de la part des décideurs face à ces informations nouvelles surtout au cour de cette dernière décennie va avoir pour conséquence le développement rapide de l'ECD. L'ECD, ou Knowledge Discovery in Data base en anglais, processus non trivial d'identification de structures inconnues, valides et potentiellement utiles dans les bases de données. Son but est essentiellement de venir en aide à l'être humain pour extraire des informations utiles (connaissances) à partir de données très volumineuses avec une croissance très rapide. Les étapes de ce processus sont l'acquisition de données multiformes (textes, images, vidéos, etc…), la préparation de données (prétraitement), la fouille de données, et enfin la validation et mise en forme des connaissances.

La Fouille de Données va se situer dans le cadre de l'apprentissage inductif. Cette étape a recours alors à différentes techniques permettant la découverte de connaissances auparavant cachées dans les données et va permettre la possibilité de prendre une décision. L'ignorance des valeurs imparfaites (manquantes, imprécises.) va rendre la décision à prendre non représentative et donc être dangereuse.

Les principes de l'extraction de connaissances à partir de bases de données ont été introduites dans l'intention d'aider les décideurs dans l'analyse des informations reçues à partir des sources électroniques.

On dispose de différentes techniques automatiques proposées pour inférer de nouvelles connaissances, potentiellement utiles, à partir de gros volumes de données. Ces connaissances correspondent à des modèles ou des relations au départ inconnues mais existant de manière implicite dans les données. L'intérêt des connaissances extraites sera validé en tenant compte

du but de l'application. Seul l'utilisateur va avoir la possibilité de pouvoir déterminer la pertinence des résultats qu'il va obtenir par rapport à ses objectifs.

En effet, les données contribuent au succès de l'activité de toute organisation et toute entreprise. Leur qualité représente un enjeu très important. Le coût de la non-qualité peut s'avérer très élevé : prendre une décision à partir de mauvaises informations peut nuire à l'organisation ou à ses clients et partenaires. L'importance des données et de leur qualité est de plus en plus reconnue.

Notre contribution consiste à proposer une approche d'amélioration de la qualité de données structurées et non structurées.

Notre mémoire comprend trois grands chapitres :

Dans le premier chapitre nous allons définir les Big Data en abordant leurs principales caractéristiques et leurs intérêts en citant tout d'abord leurs avantages et leurs inconvénients ce qui va nous permettre de mentionner alors les différentes technologies des Big Data. Dans le deuxième chapitre, nous donnons avant tout la définition de l'extraction des connaissances à partir des données puis les étapes de son processus puis nous évoquons la fouille de données après avoir donné bien sûr sa définition de même que ses tâches, ses techniques et ses différents domaines d'application. Enfin pour ce qui est du dernier chapitre, nous proposons une approche d'amélioration de la qualité des données structurées et non structurées.

CHAPITRE 1
BIG DATA

PLAN DU CHAPITRE

1 .INTRODUCTION

Le système d'information connaissant des progrès et une évolution assez rapide que les entreprises sont dans l'obligation de traiter de plus en plus de données venant de sources variées. Selon IBM (International Business Machines), chaque heure peut générer jusqu'à 2,5 trillions d'octets de données soit 2,5 péta-octets. Des prévisions pour l'année 2020 tablent sur 35 Zetta-octets alors que seulement 1 Zetta-octects de données numériques ont pu être générées dans le monde dans le début de l'informatique. Un problème va se poser celui de stockage et de l'analyse des données. La capacité de stockage des disques durs augmente alors que le temps de lecture croit également. Les technologies associées comme **Apache Hadoop** ne forment pas un monde à part dans les systèmes d'information. L'adoption de ces solutions encore nouvelles va représenter un certain nombre d'enjeux aussi bien pour les gestionnaires d'entreprises que pour les utilisateurs métier.

Ainsi dans ce chapitre, nous présentons les différents concepts de base liées au phénomène du **BIG DATA** puis nous expliquons la technologie **HADOOP** et son système de fichiers distribué **HDFS** ; étant donné que les informations massives deviennent difficiles à gérer avec les outils classiques de gestion de base de données. A la fin, on mettra l'accent sur une certaine catégorie de système de gestion des bases de données volumineuses.il s'agit du célèbre **NOSQL**.

2. DEFINITION DU BIG DATA

Le terme « **Big Data** » a été la première fois présenté au monde de calcul par Roger Magoulas de media d'O'Reilly en 2005 afin de définir une grande quantité de données que les techniques traditionnelles de gestion des données ne peuvent pas contrôler et traiter en raison de la complexité et de la taille de ces données. Une étude sur l'évolution de grandes données comme recherche et sujet scientifique prouve que le terme « grandes données » était présent dans la recherche commençant depuis les années 1970 et des publications sont apparues en 2008[W1].

De nos jours le concept de « grandes données » est traité de différents points de vue couvrant ses implications dans beaucoup de domaines. Selon Mike 2,0 , la norme ouverte de source pour la gestion de l'information, de grandes données sont définies par sa taille, comportant une grande complexe et indépendante collecte des données des ensembles, chacune avec le potentiel d'agir l'un sur l'autre. En outre, un aspect important de grandes données est le fait qu'il ne peut pas être manipulé avec des techniques standards de gestion des données dûes à la contradiction et à l'imprévision des combinaisons possibles [W2].

Dans une définition plus simple nous considérons comme étant de grandes données une expression qui comporte différents ensembles de données très grands, fortement complexes, non structurés, organisés, stockés et traités utilisant des méthodes spécifiques et des techniques utilisées pour des processus d'affaires [1].

3. CARACTERISTIQUES DU BIG DATA

Afin de mieux cerner les caractéristiques du BIG DATA des spécialistes d'IBM ont proposé trois propriétés qui les caractérisent à des degrés divers. Il s'agit du volume, de variété et de la vélocité. On appelle communément les 3V. D'autres dimensions sont fréquemment rajoutées comme la Valeur et la véracité [2].

3.1. LE VOLUME

Le volume de données est très grand, il se situe maintenant entre quelques dizaines de téraoctets et plusieurs péta-octets pour former un seul jeu de données. Les entreprises et l'ensemble des secteurs d'activités doivent absolument utiliser des moyens efficaces et fiables pour la gestion de cette quantité de données qui ne cessent d'augmenter chaque jour. Par exemple, 90% des données actuelles ont été créées dans les deux dernières années seulement. Twitter, comme exemple arrive à générer 7 Téraoctets de données chaque jour [3].

3.2. LA VARIETE

Par variété, on entend l'origine variée des sources de données et leurs différents formats qui peuvent être gérés par les Big data .On distingue deux sortes de données : les données structurées et les données non structurées (musique, image, vidéo, métadonnées, capteurs..). En exploitant cette grande variété de sources de données, les entreprises ont la possibilité d'avoir une nouvelle source d'information, cela va les permettre de croiser des données alors que cette opération était très difficile à réaliser auparavant. Ici nous prenons un exemple : s'il s'agit d'une discussion ou un débat dans un centre d'appel, on peut les stocker sous une forme textuelle pour leur contenu comme on peut stocker l'enregistrement en entier afin d'interpréter le ton de voix du client [4] .

3.3. LA VELOCITE

La vélocité est l'action de décrire la fréquence à laquelle les données sont générées, capturées et partagées. Les entreprises peuvent générer des données dans des temps très courts et doivent appréhender la vélocité pas seulement en termes de création de données, mais également sur le plan de leur traitement, de leur analyse et de leur restitution à l'utilisateur avec le respect des exigences des applications en temps réel [3].

3.4. LA VALEUR

Le but principal de l'analyse du Big Data est la création de la valeur ajoutée pour l'entreprise, et ceci en trouvant des données assez pertinentes lors du processus de décision et de les rendre très accessibles au moment du processus de décision en particulier pour des informations non structurées [5].

3.5. LA VERACITE

L'intérêt ici se rapporte à la provenance des données pour savoir s'il s'agit de données fiables ou non. A ce niveau, on va accorder plus ou moins d'importance à la donnée dans les différentes chaînes de traitement. Il y a aussi des données de réseaux sociaux qui ne sont pas sûres il faut absolument s'en méfier parce que leur origine et leur objectivité ne sont pas faciles à évaluer. De plus même pour les données dont on connaît la provenance, la pondération n'est pas constante et assurée. On peut par exemple détenir des données incomplètes dont l'anonymisation a enlevé une partie de la valeur statistique ou encore, il peut s'agir de données dépassées par le temps (anciennes) [6].

4. L'ANALYSE : LE POINT CLE DU BIG DATA

Parmi les objectifs du Big Data, on peut trouver l'extraction d'informations basées sur des données stockées pour être ensuite analysées et enfin la restitution des résultats de l'analyse ou bien l'accroissement de l'interactivité entre utilisateurs et données. Ici, on peut donner l'exemple de Google et Facebook qui sont des « entreprises Big Data ». L'analyse est le point clé de l'usage du Big Data. Elle donne la possibilité de mieux connaître sa clientèle, d'optimiser son marketing, de détecter et prévenir des fraudes et enfin l'analyse de son image sur les réseaux sociaux et d'optimiser ses processus métiers [3].

5. INTERET DE BIG DATA

L'utilisation des Big Data a un impact fort sur le monde de l'entreprise, cette dernière pourra ainsi [7] :

- Disposer d'une solution qui permet la gestion et le traitement des données structurées et non structurées à la fois.
- Améliorer les capacités de traitement des données (data processing).
- Améliorer la prise de décision.
- Développer la réactivité et l'interactivité à l'égard des clients.

6. LES AVANTAGES ET LES INCONVENIENTS DU BIG DATA

Commençons tout d'abord par citer à la fois quelques avantages et inconvénients du BIG DATA :

6.1. LES AVANTAGES

Dans [8,9], les auteurs sont cités les avantages suivants :

- Le coût est petit pour les très gros volumes.
- Le temps de réponse est court.
- Les procédures Big Data transforment automatiquement les données non structurées sous une forme structurée, ce qui permet de réaliser des analyses quantitatives et qualitatives en temps réel et étend les bases des décisions de gestion.
- Avec Big Data, les entreprises peuvent réagir plus rapidement aux évolutions du marché, (par exemple en effectuant des analyses de médias sociaux et en évaluant les données sensibles des produits).
- Il permet de détecter des tendances jusqu'alors cachées et de nouvelles opportunités d'affaires.

6.2. LES INCONVENIENTS

Dans [8, W3], les auteurs sont cités les inconvénients suivants :

- Soucis de sécurité et de protection des données.
- Volume de données disponibles insuffisant.
- Être bien informé **ne suffit pas à prendre les bonnes décisions** : ne pas se noyer dans l'information et garder un regard stratégique global, voici ce à quoi le décideur avisé doit rester vigilant.
- Rapprocher des données entre elles peuvent parfois mener à **établir des liens sans causalité**. Par exemple, savoir que tel public est plus sensible à tel produit ou sujet à tel comportement, ne nous renseigne pas sur les causes de cette relation. C'est là où la capacité d'analyse prend toute son importance : au delà des chiffres, il faut savoir donner du sens.

7. TECHNOLOGIE BIG DATA

7.1. HADOOP

Un problème se pose alors quant au stockage et à l'analyse des données. La capacité de stockage des disques durs augmente mais le temps de lecture croît également. Il devient alors nécessaire de paralléliser les traitements en stockant sur plusieurs unités de disques durs. Toutefois, cela soulève forcément le problème de fiabilité des disques durs qui engendre la panne matérielle [W4].

7.1.1. DEFINITION

Le Hadoop est considéré comme la plateforme fondamentale du Big Data. Utilisé pour stocker et faire le traitement d'immenses volumes de données, ce Framework logiciel est utilisé par un grand nombre d'entreprises pour leurs projets Big Data**.** Hadoop est un Framework Open Source Apache développé sous java et géré par la fondation Apache.Sa conception répond aux besoins du Big Data. Il a la possibilité de stocker et de traiter efficacement un grand ensemble de données réparties entre des clusters à l'aide de modèles de programmation simples. Hadoop est donc un framework capable d'exécuter des applications sur des systèmes avec des milliers de nœuds et de téraoctets. Il assure la distribution du fichier entre les nœuds et permet également au système de poursuivre le travail en cas de défaillance d'un nœud. Cette approche réduit le risque de défaillance du système. Plus le nombre de nœuds est élevé moins le temps d'exécution des tâches est court [W5]. L'application Framework Hadoop fonctionne dans un environnement qui fournit le stockage et le calcul distribué entre les clusters [10].

Hadoop fournit aux développeurs et aux administrateurs un certain nombre de briques essentielles :

- Hadoop Distributed File System (HDFS)
- Hadoop MapReduce

7.1.2. HADOOP DISTRIBUTED FILE SYSTEM (HDFS)

C'est un système de fichiers distribué, extensible et portable développé par Hadoop et basé sur le principe MapReduce à partir du Google FS. Développé sous Java, il a été conçu pour stocker de très gros volumes de données sur un grand nombre de machines peu coûteuses équipées de disques durs banalisés. Il permet l'abstraction de l'architecture physique de stockage pour arriver à manipuler un système de fichiers distribué comme s'il s'agissait d'un disque dur unique [11].

L'enregistrement d'un fichier va nous obliger à le diviser en des blocs et stocker ces blocs avec un mécanisme de réplication. Ce mécanisme va permettre au HDFS d'être plus fiable, plus tolérant en cas de pannes. Pour accéder aux fichiers du HDFS et ce par une arborescence classique dossier/sous-dossier/fichier comme si on était sur un unique disque. Les données seront assez bien réparties sur différents nœuds. Le fait qu'il soit distribué va être imperceptible lors de son usage [W6].

a. LES COMPOSANTS D'HDFS

HDFS définit deux types de nœuds :

- **NAME NODE (NŒUD MAITRE)**

Un Name Node est un service central (généralement appelé aussi maitre) qui s'occupe de gérer l'état du système de fichiers. Il maintient l'arborescence du système de fichiers et les métadonnées de l'ensemble des fichiers et répertoires d'un système Hadoop. Le NameNode a une connaissance des Data Nodes (étudiés juste après) dans lesquels les blocs sont stockés. Ainsi, quand un client sollicite Hadoop pour récupérer un fichier, c'est via le NameNode que l'information est extraite. Ce NameNode va indiquer au client quels sont les Data Nodes qui contiennent les blocs. Il ne reste plus au client qu'à récupérer les blocs souhaités [W4].

- **DATA NODE (NŒUD DE DONNEES)**

Sa fonction est le stockage et la restitution des blocs de données. Il fonctionne de la manière suivante : une requête est envoyée au NameNode durant le processus de lecture d'un fichier pour localiser l'ensemble des blocs de données, ensuite il y a renvoi de l'adresse du DataNode le plus accessible, c'est-à-dire le DataNode qui dispose de la plus grande bande passante.la communication étant permanente entre les Data Nodes et le NameNode en ce qui concerne la liste des blocs de données qu'ils hébergent .Quand certains de ces blocs ne sont pas assez répliqués dans le cluster, l'écriture de ces blocs va se faire en cascade par copie sur d'autres [11].

- **SECONDARY NAMENODE**

Le NameNode d'après l'architecture Hadoop est le seul point de défaillance. Si ce service est arrêté, il n'y a pas moyen de pouvoir extraire les blocs d'un fichier donné. Pour donner une réponse à ce type de problème, on peut avoir recours à un NameNode secondaire appelé aussi **Secondary Namenode,** celui-ci a été mis en place dans l'architecture Hadoop. Son fonctionnement est assez simple puisque le NameNode secondaire a la possibilité de vérifier périodiquement l'état du NameNode. S'il est indisponible, le secondary Namenode prend alors sa place [W4].

b. ARCHITECTURE DE HDFS

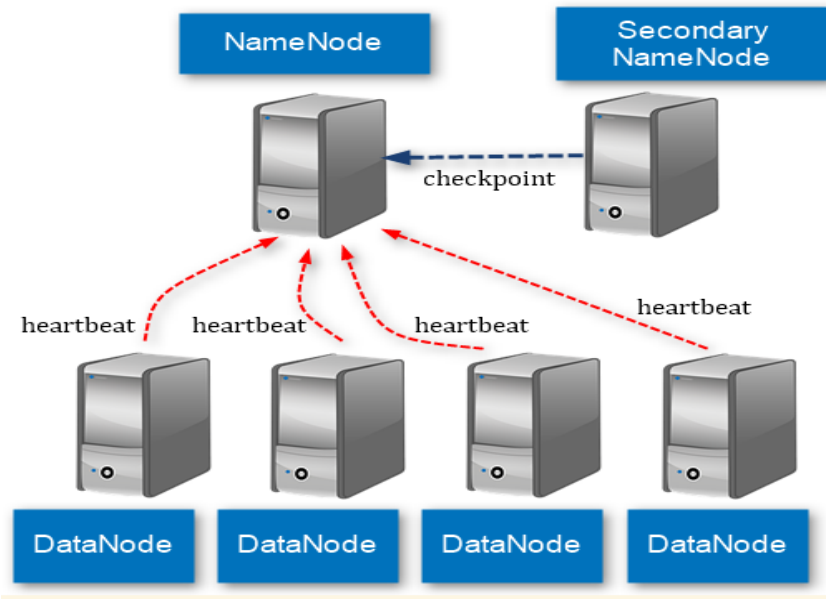

Figure 1.1 : Architecture de HDFS [W6].

7.1.3. HADOOP MAPREDUCE

MapReduce est le second composant assez important issu d'Hadooppour la gestion, la répartition et l'exécution des requêtes sur les données stockées par le Cluster. MapReduce est donc un logiciel de programmation développé au départ par Google en 2004. Il était conçu pour faciliter et simplifier la façon de traiter de grandes quantités de données en parallèle sur un nombre important de nœuds (machines) et ce d'une manière assez fiable et tolérant aux pannes.Un programme **MapReduce** peut se résumer à deux fonctions **Map ( ) et Reduce ( )** [12].

a. MAP

La fonction **MAP** va transformer les données d'entrées dans un but précis : en faire une série de couples clef/valeur. Elle va regrouper les données en les associant à des clefs, choisies de telle sorte que les couples clef/valeur contiennent un sens par rapport au problème rencontré pour le résoudre. Par ailleurs, cette opération doit absolument être parallélisable. On doit avoir la possibilité de faire des découpages dans les données d'entrées pour obtenir plusieurs fragments et faire l'exécution de l'opération **Map** à chaque machine du cluster sur un fragment distinct [13].

b. REDUCE

La fonction **REDUCE** sera chargée d'appliquer un traitement à toutes les valeurs de chacune des clefs différentes produites par l'opération **MAP.** Au terme de l'opération **REDUCE,** on aura un résultat pour chacune des clefs distinctes. Ici, on attribuera à chacune des machines du cluster une des clefs uniques produites par **MAP,** en lui donnant la liste des valeurs associées à la clef. Chacune des machines exécutera alors l'opération **REDUCE** pour cette clef [13].

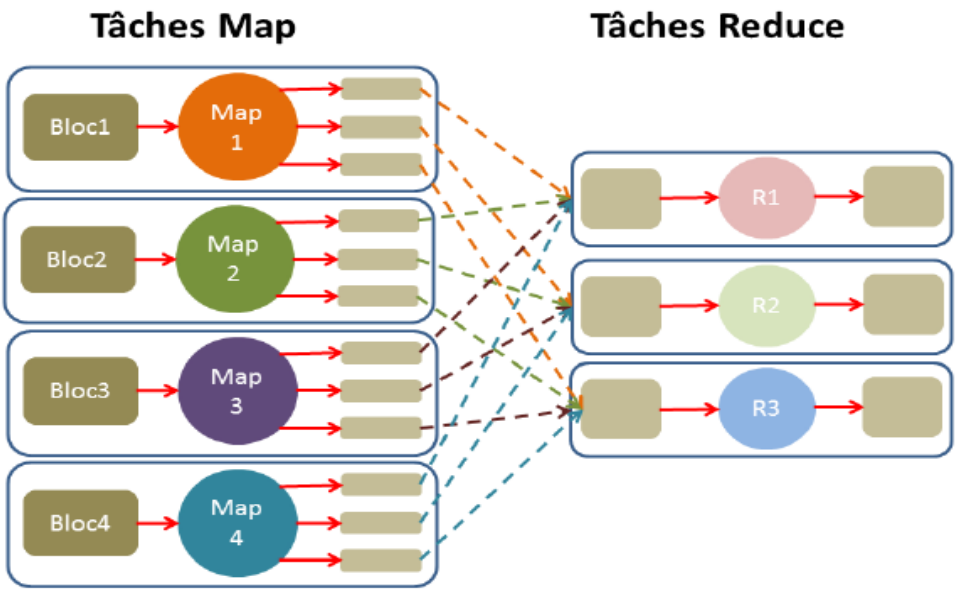

Figure 1.2: Exécution d'un Job Map Reduce [W7].

***EXEMPLE*:**

L'exemple classique est celui du **WordCount** qui permet de compter le nombre d'occurrences d'un mot dans un fichier. En entrée l'algorithme reçoit un fichier texte qui contient les mots suivants **voiture la le elle de elle la se la maison voiture.**

Dans notre exemple, la clé d'entrée correspond au numéro de ligne dans le fichier et tous les mots sont comptabilisés à l'exception du mot « se ». Le résultat de la fonction Map est donné ci-dessous :
(voiture, 1) /(la,1) /(le,1) / (elle,1) / (de,1) / (elle,1) / (la,1) / (la,1) / (maison,1) / (voiture,1).

Avant de présenter la fonction Reduce, deux opérations intermédiaires doivent être exécutées pour préparer la valeur de son paramètre d'entrée. La première opération appelée **shuffle** permet de grouper les valeurs dont la clé est commune. La seconde opération appelée **sort** permet de trier par clé. A la différence des fonctions Map et Reduce, shuffle et sort sont des fonctions fournies par le Framework Hadoop, donc, il n'a pas à les implémenter **[3]**.

Ainsi, après l'exécution des fonctions shuffle et sort le résultat de l'exemple est le suivant :

(de, [1]) / (elle, [1,1]) / (la, [1, 1,1]) / (le, [1]) / (maison, [1]) / (voiture, [1,1])

Suite à l'appel de la fonction Reduce, le résultat de l'exemple est le suivant :
(de, 1) / (elle, 2) / (la, 3) / (le, 1) / (maison, 1) / (voiture, 2).

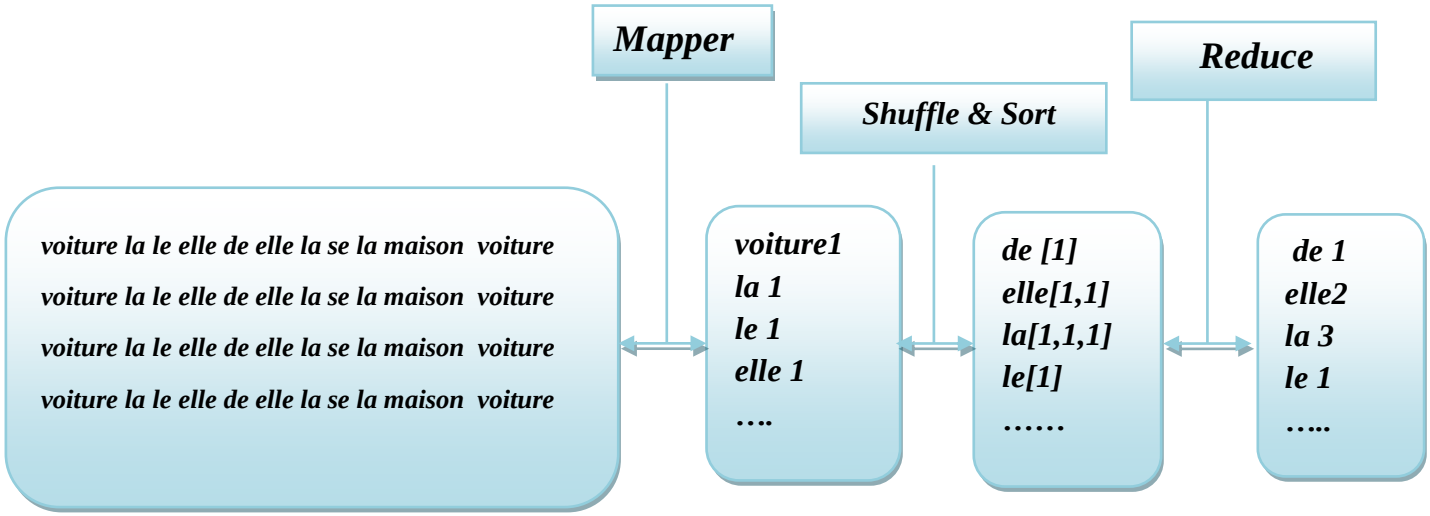

Figure 1.3: Exemple d'un programme MapReduce [3].

8. FOURNISSEURS DE DISTRIBUTION HADOOP

8.1. CLOUDERA

Cloudera représente une société de logiciels. Elle se présente comme la compagnie commerciale d'Hadoop. Elle a été fondée par des experts Hadoop en provenance de Facebook, Google, Oracle et Yahoo.

Elle a été conçue pour le déploiement, faire la configuration, la surveillance et le diagnostic des grappes d'ordinateurs assez facilement à partir d'une console de gestion Web centralisée. Sous licence libre, on peut la télécharger et l'utiliser. Cloudera offre une nouvelle version entreprise spécialement conçue pour faciliter le travail des entreprises : cette version contient aussi des nouvelles fonctionnalités à même de répondre aux exigences et aux besoins réels de ces entreprises [14].

8.2. HORTONWORKS

Elle est la seule plateforme 100% open source basée sur Apache Hadoop. La stratégie d'Horton works est de fonctionner en utilisant des versions stables et testées d'Apache Hadoop plutôt que sur les dernières versions. Leur solution de gestion du cluster, Ambari, n'est pas aussi mature que la concurrence : Cloudera Manager et HeatMap. HortonWorks a signé des partenariats importants avec IBM, Microsoft. Il y a eu un accord avec Microsoft pour assurer l'utilisation de leur plate forme [15].

8.3. MapR

Des collaborateurs de Google ont fondé en 2009 cette société Californienne appelée MapR. Elle propose une distribution de Hadoop qui se veut particulièrement facile d'utilisation. Elle participe à l'enrichissement du noyau Hadoop avec les solutions propriétaires et propres trois distributions : M3 qui est gratuite, M5 qui ajoute des fonctionnalités de haute disponibilité ainsi que M7 qui donne la possibilité d'intégrer une base de donnée haute disponibilité qui réimplémente l'API de HBase. Elle utilise un système de fichier propriétaire,MapR FS à la place du système HDFS de Apache dans le but d'augmenter les performances et propose également sa propre implémentation de MapR MapReduce[2].

9. LA MANIPULATION DES BIG DATA

Durant de longues années, les bases de données relationnelles furent la seule solution pour enregistrer des données ou alors la solution que beaucoup de personnes ont adopté de par le monde sans beaucoup réfléchir sur le sujet. Bien que certains personnes considèrent que le problème de stockage des données est pratiquement multiple et qu'il est plutôt convenable de se poser plusieurs questions:

- Est-ce que les données sont fortement structurées ou non ?
- Est-il acceptable de perdre un enregistrement sur un million ? Sur un milliard ?
- Est-ce que les données sont réparties sur plusieurs data-centres ?
- Est-ce que la taille des données peut être multipliée par 10 en l'espace d'un mois ?
- Etc…

Les bases de données relationnelles proposent des réponses à ces questions que l'on se pose ; Dans plusieurs cas, elles peuvent être acceptables et raisonnables mais ceci n'est pas une règle absolue, il y a quand même d'exceptions .Par exemple, les bases de données relationnelles s'adaptent très mal quand on tient à privilégier les performances au lieu de la garantie d'écriture des données. Alors, afin de donner une réponse adéquate à ces différents problèmes, un mouvement, NoSQL, vient pour proposer des outils différents et particulièrement conçus pour certains cas rencontrés.

Certaines bases de données NoSQL sont réalisées pour traiter les gros volumes de données et d'autres sont fait pour maximiser le nombre de requêtes par seconde qu'un serveur pourra traiter, etc.

Notons en particulier que la plupart des plus gros sites web ont quitté le monde relationnel (Google, Facebook, Twitter, Amazon), ce qui va nous conduire à trouver d'autres outils que les bases de données relationnelles [W8].

9.1. NoSQL

NoSQL signifie **Not Only SQL** , il regroupe de nombreuses bases de données récentes pour la plupart qui se caractérisent par une logique de représentation de données non relationnelles et qui n'offrent donc pas une interface de requêtes en SQL [16] .

Les bases de données **NoSQL** sont conçues pour gérer des volumes de données à très grande échelle. Elles ont généralement la propriété de pouvoir les répartir sur un grand nombre de serveurs qui les rendent extensibles (scalable).Ces bases ne sont pas relationnelles comme celles reposant sur le langage standardisé « SQL » permettant de manipuler des données dites « structurées ». Les bases NoSQL permettent de manipuler des données non structurées. En effet, NoSQL ne vient pas remplacer les BD relationnelles mais proposer une alternative ou compléter les fonctionnalités des SGBDR pour donner des solutions plus intéressantes dans certains contextes. Leurs principaux avantages sont leurs performances et leurs capacités à traiter de très grands volumes de données [17].

9.2. LES DIFFERENTS TYPES DE BASES NOSQL

Les différents types de bases NOSQL peuvent être classés en quatre catégories :

9.2.1. BASE DE DONNEES ORIENTEE « CLE/VALEUR »

Il s'agit de la catégorie de base de données la plus simple. Dans ce modèle chaque objet est identifié par une clé unique qui constitue la seule manière d'y accéder. Dans ce modèle on ne dispose généralement que de quatre opérations de base : Create, Retrieve, Update, Delete. Ces bases sont connues pour leur performance en lecture et en écriture. Elles ont une extensibilité (scalability) élevée. On les retrouve assez souvent comme système de stockage de cache ou de sessions distribuées, notamment là où l'intégrité relationnelle des données n'est pas significative. Les implémentations les plus répandues des bases « clé/valeur » sont :Riak, Redis,DynamoDB chez Amazon [18].

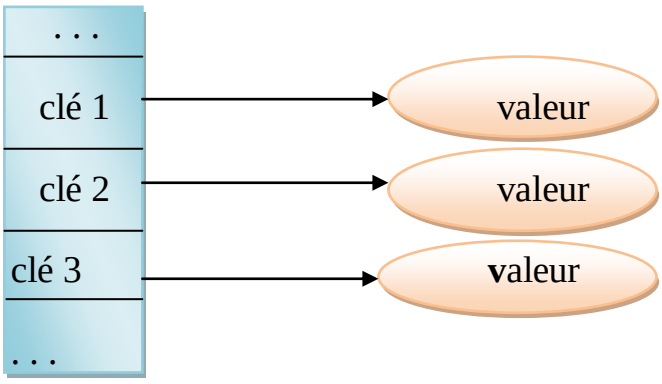

Figure 1.4 : Modèle de bases NoSQL type "clé / valeur" [18].

- ***EXEMPLE***

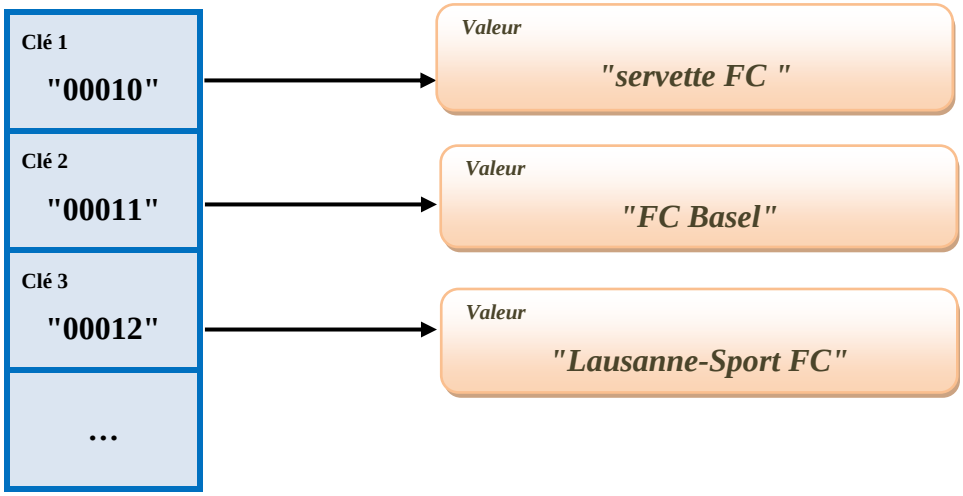

Figure 1.5 : Exemple d'une base NoSQL de type "clé / valeur [16].

9.2.2. BASE DE DONNEES ORIENTEE DOCUMENT

Avec ce modèle on va pouvoir ajouter au modèle Clé-valeur, l'association d'une valeur à structure non plane, c'est-à-dire qui nécessiterait un ensemble de jointures en logique relationnelle. Les bases de données documentaires sont formées de collections de documents. Un document est composé de champs et des valeurs associées, celles-ci pouvant être des fois requêtées. Par ailleurs, les valeurs peuvent être, soit d'un type simple (entier, chaine de caractère, date…), soit elles-mêmes composées de plusieurs couples clé-valeur. A ce niveau, on n'est pas obligé de définir au préalable les champs utilisés dans un document. Au sein même de la base les documents peuvent être très hétérogènes. Le stockage effectué des documents leur assure des fonctionnalités qui n'existent pas dans les bases clés valeurs simples et la plus évidente est la faculté de faire des requêtes sur le contenu des objets. Ce modèle se base sur le paradigme clé-valeur. La valeur, dans ce cas, est un document de type XML. L'avantage est de pouvoir récupérer, par une seule clé, un ensemble d'informations structurées de manière hiérarchique. Une même opération dans le monde relationnel va impliquer plusieurs jointures. Là, on peut retrouver principalement MongoDB et CouchBase comme solutions basées sur le concept de base documentaire [18].

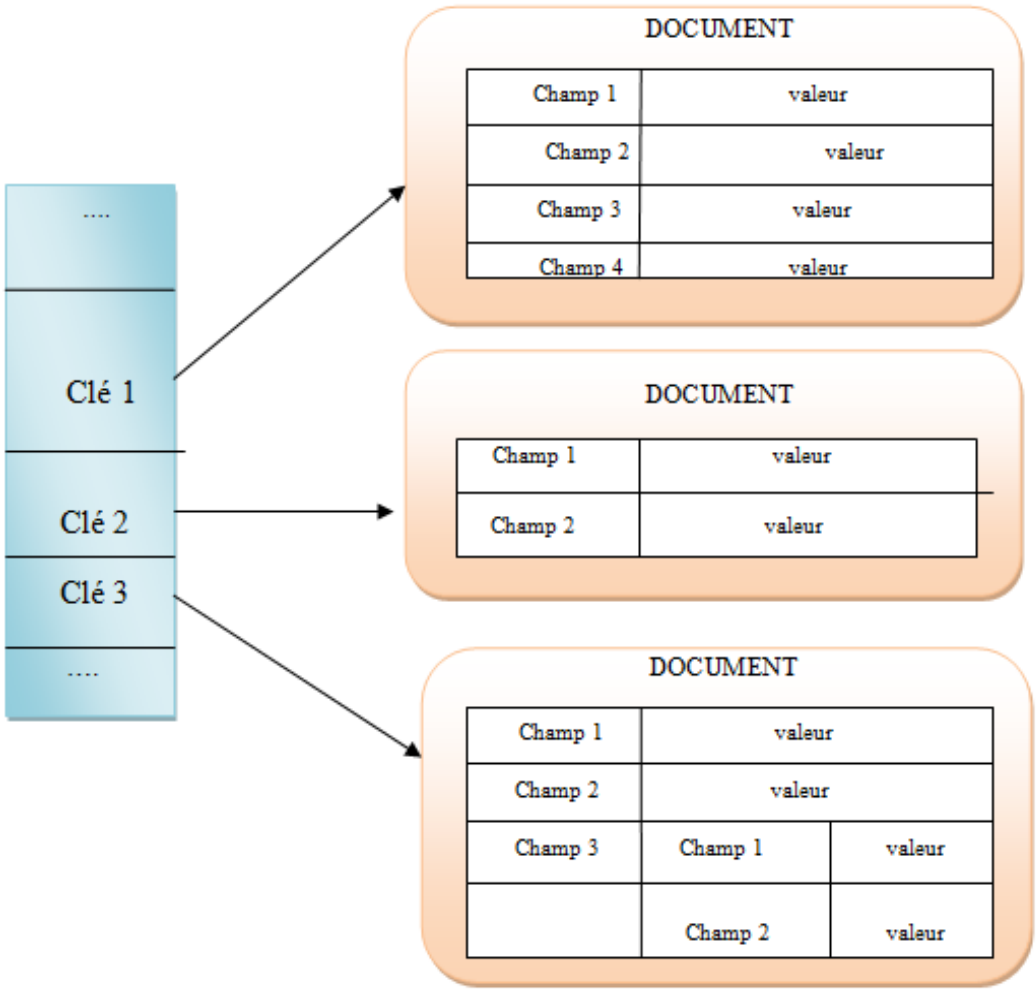

Figure 1.6 : Modèle de bases NoSQL type documentaires [18].

- ***Exemple***

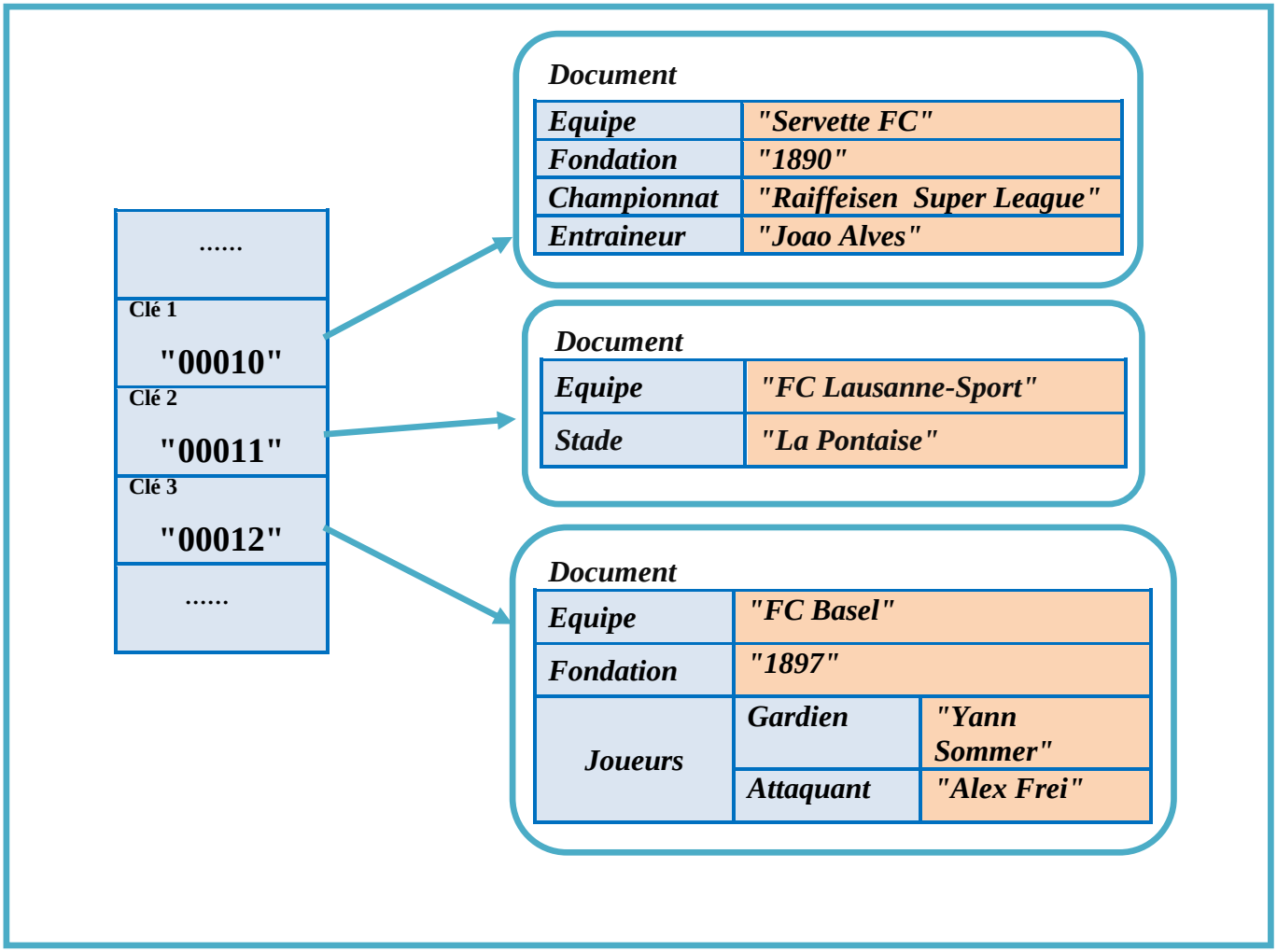

Figure 1.7 : Exemple d'une base NoSQL de type document [16].

9.2.3. BASE DE DONNEES ORIENTEE COLONNE

Ce modèle à priori peut parfois ressembler à une table dans un SGBDR à la différence qu'avec une BD NoSQL orientée colonne, le nombre de colonnes est dynamique. En effet, dans une table relationnelle, le nombre de colonnes est fixé à partir de la création du schéma de la table et ce nombre va rester le même pour tous les enregistrements dans cette table. Cependant, avec ce modèle, le nombre de colonnes peut varier d'un enregistrement à un autre ce qui évite de retrouver des colonnes ayant des valeurs NULL. Les bases les plus connues qui vont se baser sur ce concept sont HBase et Cassandra [W9].

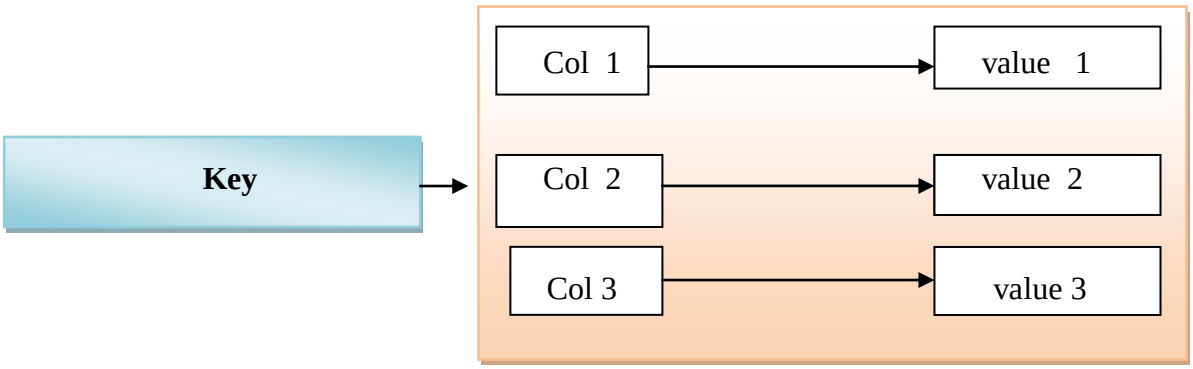

Figure 1.8 : Modèle de bases NoSQL type Colonne [19].

- ***EXEMPLE***

	A	B	C	D	E
1	Foo	Bar	Hello		
2		Tom			
3			Java	Scala	Cobol

1	A	Foo	B	Bar	C	Hello
2	B	Tom				
3	C	Java	D	Scala	E	Cobol

Figure 1.9 : Exemple d'une base NoSQL de type Colonne [19].

9.2.4. BASE DE DONNEES ORIENTEE GRAPHE

Dans de nombreuses situations comme les réseaux sociaux, il s'agit de relier des entités par différentes relations dotées d'attributs et ayant le sens de navigation. Telle personne « aime », «connait » ou « fait partie du cercle professionnel» de telle autre. Malgré leur nom, les SGBDR s'adaptent mal en parcourant rapidement un graphe.Un SGBDR ne s'adapte que pour décrire des relations simples comme celle qui lie un client à ses commandes par exemple. Par contre, dés que le nombre de liens à parcourir dépasse un ou deux, la nécessité de définir à chaque fois des clés étrangères et d'effectuer de nombreuses jointures va pénaliser les performances jusqu'à rendre cette approche impraticable [2].

Les bases de données orientées graphe s'appuient principalement sur deux concepts : D'une part l'utilisation d'un moteur de stockage pour les objets qui se présentent sous la forme d'une base documentaire, chaque entité de cette base étant nommée nœud. D'autre part, ce modèle, va s'ajouter un mécanisme qui permet de décrire les arcs (relations entre les objets), les arcs étant orientés, disposent de la possibilité d'avoir des propriétés (nom, date, …). Ici la principale solution est Neo4J: [20].

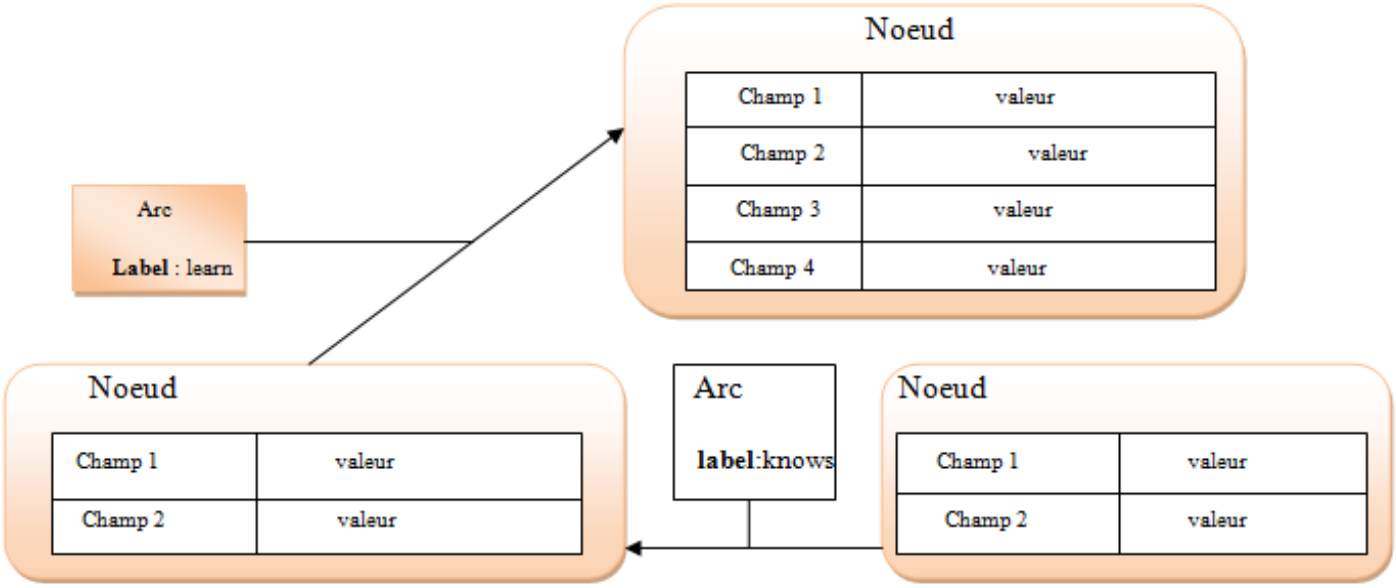

Figure 1.10 : Modèle de bases NoSQL orientées graphe [18].

- ***EXEMPLE***

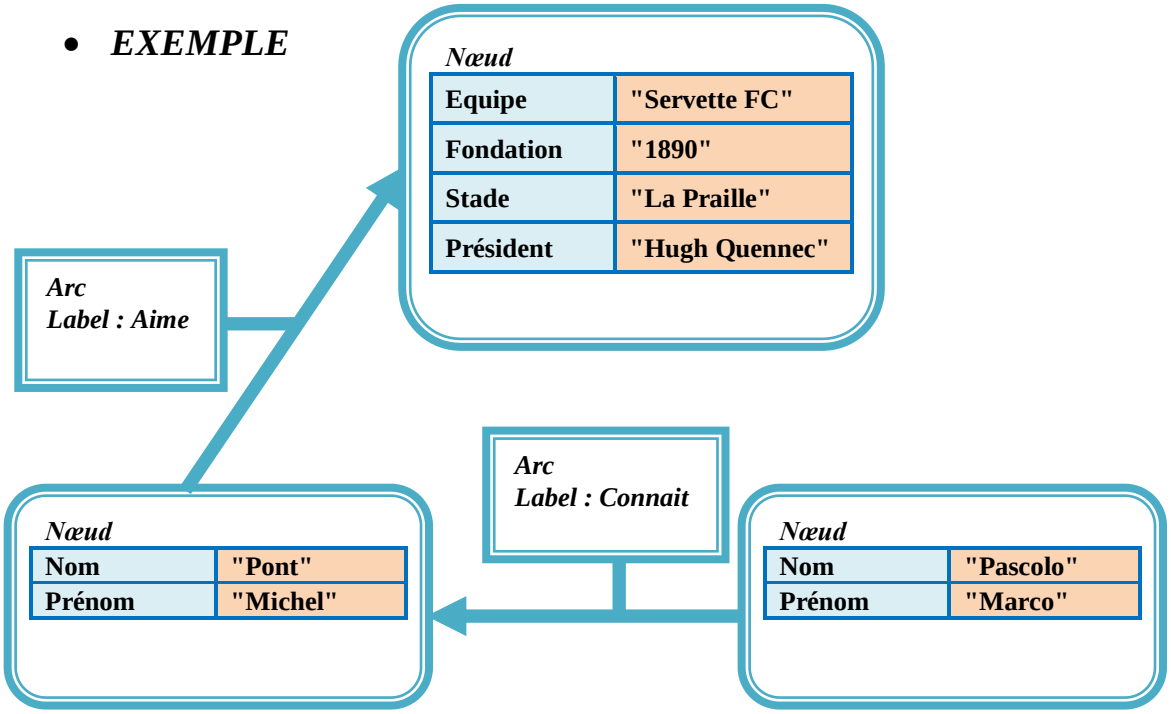

Figure 1.11 : Exemple d'une base NoSQL de type Graphe [16].

10. CONCLUSION

La quantité de données générées par des personnes, des appareils connectés à Internet et des entreprises, se développe à un rythme extraordinaire. Les différents organismes financiers, les entreprises, de même que les services de santé … génèrent de grandes quantités de données pendant leurs interactions avec les fournisseurs, les patients, les clients et les employés. Loin de ces mêmes interactions, de nombreuses données sont en cours de création et ce à travers les requêtes de recherches sur Internet. Le résultat obtenu est la constitution de la « révolution de données » ou ère du « Big Data ».

Le « Big Data » est un terme en général employé pour faire la description de l'augmentation fulgurante des volumes de données, à côté bien sûr de la croissance dans les capacités à transférer, stocker et analyser ces mêmes données. L'expression « Big Data » fait également référence aux technologies, processus et techniques qui permettent à une organisation de créer, manipuler et gérer des données à grande échelle et extraire de ces mêmes données de nouvelles connaissances afin d'en faire une valeur économique. Cependant stocker des quantités énormes de données est une chose, les traiter en est vraiment autre chose. Donc, dans ce chapitre, nous avons défini les principaux concepts liés au terme Big data, ensuite nous avons expliqué la technologie Hadoop et ses différents composants ainsi nous décrivons en détail le fonctionnement du mécanisme Mapreduce, nous terminons ce chapitre par une présentation du différents types des bases de données NOSQL.
Dans le chapitre qui suit, nous allons décrire les outils nécessaires pour l'extraction des connaissances à partir des Big data qui restent un domaine de recherche important et ouvert dans plusieurs filières.

CHAPITRE 2

EXTRACTION DES CONNAISSANCES A PARTIR DES DONNEES

PLAN DU CHAPITRE

7. Données et fouille de données

7.1. Les différents types de données

7.2. Distance et similarité

7.2.1. Notion de similarité et dissimilarité

7.2.2. Quelques mesures de similarité.

8. Les techniques de DataMining

8.1. Règles d'Association

8.1.1. Avantages et inconvénients

8.2. Les arbres de décision

8.2.1. Construction d'un arbre de décision

8.2.2. Avantages et inconvénients.

8.3. Clustering

8.3.1. Les algorithmes de clustering

9. Domaines d’application du Data Mining

9.1. Le secteur bancaire.

9.2. La détection de fraude.

9.3. Le secteur des assurances.

9.4. La médecine et la pharmacie.

10. Motivations du Data Mining

10.1. Explosion des données.

10.2. Améliorer la productivité.

10.3. Croissance en puissance/coût des machines capables.

11. Text Mining

12. Sound Mining

13. Image Mining

14. Vidéo Mining

15. Conclusion

1. INTRODUCTION

L'extraction de connaissances à partir des données est une discipline assez récente à l'intersection des domaines des bases de données, de l'intelligence artificielle, de la statistique, des interfaces homme / machine et de la visualisation. Par la collection de données faites par des experts, il s'agit de proposer des connaissances nouvelles pour enrichir les interprétations du champ d'application tout en donnant des méthodes automatiques permettant l'exploitation de ces informations.

Les principes de base de l'extraction des connaissances à partir des données vont être utilisés pour aider les décideurs dans l'analyse des informations issues des sources électroniques. Plusieurs techniques automatiques sont alors proposées pour inférer de nouvelles connaissances assez utiles à partir de données volumineuses. Ces connaissances vont correspondre à des modèles ou des relations inconnues au départ cependant existant de façon implicite dans les données. L'intérêt des connaissances extraites va être validé en tenant compte du but de l'application. Il y a seulement l'utilisateur averti qui puisse déterminer la pertinence des résultats obtenus par rapport à ses objectifs.

L'ECD va faire appel à des disciplines très diverses et assez variées comme les statistiques, l'intelligence artificielle, l'apprentissage automatique, la reconnaissance des formes, les bases de données et les techniques de visualisation. Son intention est alors d'automatiser ou d'aider l'extraction de nouvelles connaissances pertinentes et assez fiables à partir des grandes masses d'informations internes ou externes.

Ainsi dans ce chapitre, nous étudions l'extraction des connaissances à partir des données en commençant par une définition appropriée et présenter le processus d'ECD puis nous avons abordé la fouille de donnée comme une étape essentielle dans le processus ECD en donnant sa définition, ses tâches et ses techniques en finissant par donner quelques domaines d'applications de fouille de donnée .

2. DONNEE, INFORMATION ET CONNAISSANCE

Les données sont constituées par les faits, les observations, les éléments bruts, qui ne sont pas encore été interprétées, non traitées et non encore analysées. Prenons par exemple : 3000 nouveaux étudiants inscrits en première année , la température rectale du malade est de 37,5 °C [W10].

Lorsque les données sont interprétées, organisées, traitées, structurées ou présentées dans un contexte donné afin de le rendre utile, elles sont appelées **informations**. Par exemple : augmentation de 20% du nombre d'étudiants par rapport à l'année précédente [W11].

En effet, **la connaissance** est une information comprise c'est-à-dire assimilée et utilisée qui permet d'aboutir à des actions, les actions peuvent être une prise de décision ou la création de nouvelles informations. Prenons par exemple : vue l'augmentation des nombres d'étudiants inscris : un autre pôle universitaire sera établi [W12].

3. EXTRACTION DES CONNAISSANCES A PARTIR DE DONNEES

D'après certains experts, il est possible que les données peuvent doubler en partie et ceci dans les neufs mois [21].En se référant à des travaux scientifiques expérimentés, on s'aperçoit que ce sont des gigas octets de données qui sont en général collectées et stockées. Il y a également un nombre important de données collectées sur différents clients qui ne sont pas toutes exploitées dans certaines entreprises commerciales, par exemple les compagnies d'assurances ou même les banques.

Il faudrait donc savoir comment faire, comment agir pour répondre aux besoins de certaines entreprises et sociétés qui pourraient dans le cas échéant utiliser ces données en temps voulu tout en sachant que les requêtes traditionnelles de type SQL peuvent avoir des limites et ce au niveau de type d'information qu'elles sont censées obtenir d'une base de données : la question reste posée ? Tout cela fait partie d'un ensemble d'éléments de recherches et de développement appelés l'extraction des connaissances dans les données ou en anglais Knowledge Discovery in Database (KDD) [22].

C'est avec l'exploitation des techniques d'extraction des connaissances que les bases de données volumineuses sont devenues des sources riches et fiables pour la génération et la validation de connaissances. Data Mining en anglais ou fouille de donnée en français n'est qu'une étape du processus d'extraction des connaissances à partir des données et consistant alors à appliquer des algorithmes d'apprentissage sur les données pour obtenir des modèles (ou motifs). L'extraction des connaissances à partir des données se situe à l'intersection de nombreuses disciplines [23]. [W13] tels que l'apprentissage automatique, la reconnaissance de formes, les bases de données, les statistiques, la représentation des connaissances, l'intelligence artificielle, etc …

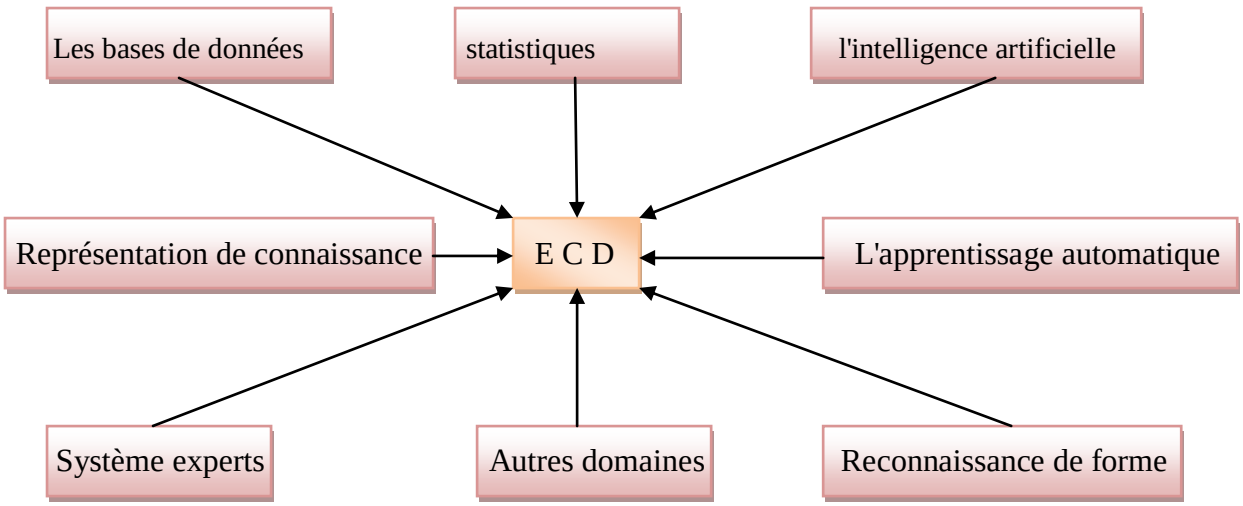

Figure 2.1 : Quelques domaines de l'ECD [23].

4. DEFINITIONS D'ECD

➢ DEFINITION 1

ECD est un processus non trivial qui consiste à identifier dans les données des modèles nouveaux, valides, potentiellement utiles et surtout compréhensibles et utilisables. Un expert du domaine relatif aux données, l'analyste est chargé de diriger l'extraction. Ces nouvelles connaissances viennent compléter le savoir de l'analyste sur le domaine. En fonction de ses objectifs, l'analyste va sélectionner les données et utiliser les outils de fouille de données pour construire des modèles du domaine expliquant les données. L'analyste peut ensuite sélectionner et exploiter les modèles qui représentent un point de vue « satisfaisant » [24].

➢ DEFINITION 2

ECD est un processus itératif et interactif d'analyse d'un grand ensemble de données brutes afin d'en extraire des connaissances exploitables par un utilisateur-analyste qui y joue un rôle central [25].

D'après ces deux définitions, L'utilisateur fait partie intégrante du processus: L'interactivité est liée aux différents choix que l'utilisateur est amené à effectuer. L'interactivité est liée au fait que l'utilisateur peut décider de revenir en arrière à tout moment si les résultats ne lui conviennent pas car l'extraction des connaissances est dirigée par l'objectif de l'utilisateur [26].

5. PRESENTATION DU PROCESSUS D'ECD

Les phases constituant le processus de l'ECD sont [27]:

- ✓ La compréhension du domaine d'application.
- ✓ Préparation des données.
- ✓ Fouille de données (Datamining).
- ✓ Interprétation et évaluation.

5.1. LA COMPREHENSION DU DOMAINE D'APPLICATION

Dans cette première étape, il est question d'exposer le problème et où l'on définit les objectifs, le résultat attendu de même que les moyens pour mesurer le succès du processus de l'ECD. Il y va de la compréhension du contexte de la recherche dans le but de donner un sens logique aux variables. Dans cette étape d'introduction, il va être intéressant de recueillir les intuitions de même que les connaissances des experts pour assurer l'orientation du processus de découverte ou tout simplement pour faire l'identification des variables les plus pertinentes capables d'expliquer les phénomènes analysés[28].

5.2. PREPARATION DES DONNEES

Pendant la préparation des données, le processus d'ECD s'intéresse à [27, 29,30]:

- L'acquisition des données où il procède à la sélection, le nettoyage et l'intégration des données cibles.
- Le prétraitement des données où l'ECD exécute des opérations complémentaires du nettoyage et sélection des données.
- La transformation des données où l'ECD réalise les opérations standards de normalisation, de standardisation et de conversion des données.

5.2.1. ACQUISITION DES DONNEES

- **Rechercher et bien sélectionner des données** est une opération importante et essentielle dans le processus ECD. Pour faire ce travail de sélection on va faire appel à des experts du domaine afin de déterminer au mieux les attributs à avoir pour décrire la problématique. Les mêmes experts peuvent trouver les variables à utiliser pour résoudre le problème. Dans cette étape, il est utile et préférable de connaitre les éléments du contexte pour construire une représentation élémentaire du problème. Cette étape a surtout pour but d'avoir assez de données potentielles pour en extraire des informations fiables, exploitables et pertinentes [31].

- **Le nettoyage des données**

C'est la phase de préparation d'un ensemble de données fiables nécessitant pour ce fait des techniques appropriées en ce qui concerne leur qualité pour en éliminer le cas échéant des données redondantes et/ ou des valeurs aberrantes pour ne retenir que des informations à même d'être utilisées en tenant compte du bruit , du choix de stratégie de traitement des valeurs manquantes pour enfin décider sur quelles bases de données faudrait-il avoir recours (types de données , le schéma à utiliser…). En effet, le nettoyage des données consiste à retravailler des données de mauvaises qualités soit en les supprimant, soit en les modifiant de manière à en tirer le meilleur profit [32].

EXEMPLE :

Soit l'exemple suivant qui présente une base de données d'un éditeur qui vend 5 sortes de magazines : sport, voiture, maison, musique et BD. Il souhaite mieux étudier ses clients pour découvrir de nouveaux marchés ou vendre plus de magazines à ses clients habituels [26].

Client	Nom	Adresse	Date d'abonnement	Magazine
23134	Bemol	Rue du moulin, Paris	7/10/96	Voiture
23134	Bemol	Rue du moulin, Paris	12/5/96	Musique
23134	Bemol	Rue du moulin, Paris	25/7/95	BD
31435	Bodinoz	Rue verte, Nancy	11/11/11	BD
43342	Airinaire	Rue de la source, Brest	30/5/95	Sport
25312	Talonion	Rue du marché, Paris	25/02/98	NULL
43241	Manvussa	NULL	14/04/96	Sport
23130	Bemolle	Rue du moulin, Paris	11/11/11	Maison

Tableau 2.1 : La base de données avant le nettoyage [26]

Intégrité de domaine : Dans notre exemple, la date d'abonnement des clients 23130, 31435 (11/11/11) semble plutôt correspondre à une erreur de saisie ou encore à une valeur par défaut en remplacement d'une valeur manquante. Pour les informations manquantes : Dans notre exemple, nous n'avons pas le type de magazine pour le client 25312. Il sera écarté de notre ensemble. L'enregistrement du client 43241 sera conservé bien que l'adresse ne soit pas connue. Après le nettoyage on aura la base de données suivante [26] :

Client	Nom	Adresse	Date d'abonnement	Magazine
23134	Bemol	Rue du moulin, Paris	7/10/96	Voiture
23134	Bemol	Rue du moulin, Paris	12/5/96	Musique
23134	Bemol	Rue du moulin, Paris	25/7/95	BD
31435	Bodinoz	Rue verte, Nancy	NULL	BD
43342	Airinaire	Rue de la source, Brest	30/5/95	Sport
43241	Manvussa	NULL	14/04/96	Sport
23130	Bemolle	Rue du moulin, Paris	NULL	Maison

Tableau 2.2 : La base de données après le nettoyage [26].

- L'intégration des données consiste à combiner des données éparses obtenues à partir de plusieurs sources (base de données, sources externes, etc …). Cette étape va permettre de résoudre le cas des données hétérogènes qui posent des problèmes, sachant que les données peuvent provenir de différentes sources et hétérogènes (bases de données relationnelles, fichiers XML, etc …). C'est également dans l'intention d'obtenir des entrepôts de données ou de magasins de données spécialisés contenant les données retravaillées pour faciliter leurs exploitations futures [30].

5.2.2. LE PRETRAITEMENT DES DONNEES

Il est possible que les bases de données contiennent à ce niveau un certain nombre de données incomplètes et/ou bruitées et c'est pour cela que ces données doivent absolument être traitées et surtout si elles ne l'ont pas été auparavant. Dans le cas contraire, les données sont stockées dans un entrepôt de données et/ou des magasins de données spécialisés pour faciliter leurs exploitations futures. Cette étape permet d'affiner les données. Si l'entrepôt de données est bien construit, le prétraitement de données permettra une amélioration qualitative des résultats lors de l'interrogation dans la phase de fouille de données [33].

- **EXEMPLE**

Soit la base de données nettoyée précédemment, le tableau 2.3 présente le résultat de prétraitement. Les clients qui ont des informations manquantes seront supprimés de la base [26].

Client	Nom	Adresse	Date d'abonnement	Magazine
23134	Bemol	Rue du moulin, Paris	7/10/96	Voiture
23134	Bemol	Rue du moulin, Paris	12/5/96	Musique
23134	Bemol	Rue du moulin, Paris	25/7/95	BD
43342	Airinaire	Rue de la source, Brest	30/5/95	Sport

Tableau 2.3 : La base de données après le prétraitement [26].

5.2.3. TRANSFORMATION DES DONNEES

Cette étape est d'une importance capitale pour la réussite du projet d'ECD. Il faut qu'elle soit adaptée en tenant compte de chaque base de données et des objectifs essentiels du projet. Pour cela et plus particulièrement lors de cette étape , il faut savoir utiliser les méthodes adéquates à même de réduire les dimensions pour avoir une représentation réduite de l'ensemble des données , volume plus réduit et capable de produire à peu prés les mêmes résultats analytiques [W14]. Cette opération sera réalisée sur les données sous forme tabulaire. L'objectif étant la réduction du nombre de données ou la sélection des lignes ou colonnes permettant un meilleur usage par l'utilisateur [34].

D'un autre côté, la transformation des données va consister à transformer un attribut A en une autre variable A' qui pourrait être plus appropriée pour les objectifs de l'étude. A ce niveau, plusieurs méthodes sont utilisées comme la discrétisation et la normalisation des données [34].

Une fois que toutes ces étapes seront terminées, les étapes suivantes seront liées à la partie de Data Mining, avec une orientation sur l'aspect algorithmique [35].

5.3. FOUILLE DE DONNEES

La fouille de données est au cœur du processus d'ECD, il se réfère à une série d'activités comme le choix du type de la tâche de fouille de données, la sélection de la technique de fouille de données. En premier lieu, il faut bien choisir le type de la tâche de fouille de données. A ce niveau, on peut distinguer dans l'ordre le clustering, la classification, la régression, l'analyse des associations, …. [36]

Le choix de la tâche étant fait, la technique de fouille est alors à son tour choisie et cela parmi plusieurs techniques en tenant compte des besoins des utilisateurs et des avantages et inconvénients des uns et des autres. L'étape suivante consistera au choix de l'algorithme retenu pour la technique de fouille choisie. Il faut pour cela penser à inclure une méthode de recherche de modèles dans les données. La décision de l'algorithme et des paramètres appropriés sera en principe guidé par le degré de précision et ce par rapport à la facilité d'interprétation des connaissances extraites. Les réseaux de neurones sont recommandés si l'on cherche la précision et pour plus de clarté on optera alors pour les arbres de décision [37].

Le problème de la fouille de données réside alors dans le strict choix de la méthode à même de résoudre tel ou tel problème. A ce niveau, il est possible de combiner plusieurs méthodes dans le but d'arriver à un meilleur résultat [38].

5.4. INTERPRETATION ET EVALUATION

Cette étape contient l'interprétation et l'évaluation des modèles extraits en les mesurant. Cette phase permet et donne la possibilité de faire un retour à une des étapes précédentes et assure une représentation visuelle de ces modèles, d'ôter systématiquement les modèles qui se répètent ou non représentatifs pour les transformer dans des termes assez clairs et compréhensibles pour celui qui les utilise (utilisateur) [35].

C'est un véritable expert, un spécialiste en la matière qui assure et mène l'opération d'interprétation et d'évaluation (expert, analyste,…). Ce travail de post traitement assez délicat à mener permet de faciliter la tâche du spécialiste pour lui fournir les critères de décision sous une forme de mesures de qualité ou d'intérêt des modèles. La conception de ces mesures obéit et combine deux méthodes de validation appelée aussi validation statistique et validation par expertise [39].

C'est par des méthodes de base de statistique descriptive que l'on peut assurer l'opération dite de validation statistique. Le but essentiel étant ici d'avoir des informations fiables à même de porter un jugement sur les résultats obtenus et d'en estimer leur qualité .On a la possibilité d'avoir cette validation par un calcul de moyenne et variances des attributs ou tout simplement d'en déterminer la classe majoritaire en ce qui concerne leur classification…etc. L'opération de validation ne peut être assurée que par un spécialiste, expert, pour juger de la pertinence des résultats obtenus. Exemple, l'opération de la ou les règle(s) d'association ne peut être faite que par un expert en la matière à même de juger convenablement leur pertinence [26].

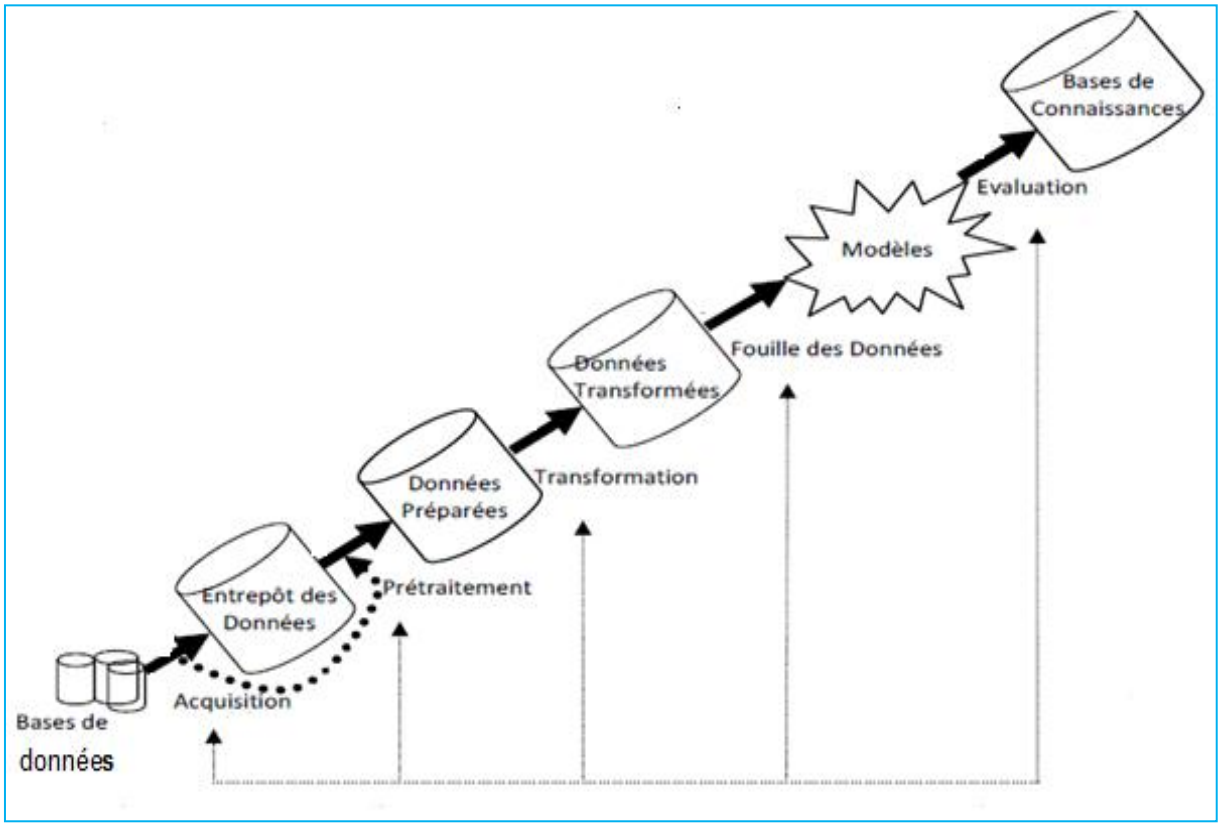

Figure 2.2 : Processus d'ECD [40].

6. FOUILLE DE DONNEES

6.1. DEFINITIONS DE LA FOUILLE DE DONNEES

Aujourd'hui, le Data Mining est un domaine très vague et qui « s'intéresse à la découverte d'informations utiles et nouvelles dans une quantité importante de données » [41].

Le terme de Data Mining (ou fouille de données) est souvent employé pour désigner l'ensemble des outils permettant à l'utilisateur d'accéder aux données de l'entreprise, de les analyser. Nous restreindrons ici le terme de Data Mining aux outils ayant pour objet de générer des informations riches à partir des données notamment des données stockées, de découvrir des modèles implicites dans les données. Ils peuvent permettre par exemple à un magasin de dégager des profils de client et des achats et de prévoir ainsi les ventes futures [42].

D'autres définitions existent. Nous avons sélectionné quelques-unes :

1. " La découverte de nouvelles corrélations, tendances et modèles par le tamisage d'un grand nombre de données"[43].
2. Le Data Mining est un processus non trivial qui consiste à identifier, dans des données, des schémas nouveaux, valides, potentiellement utiles et surtout compréhensibles et utilisables [44].
3. L'exploration et l'analyse, par des moyens automatiques ou semi automatiques, d'un large volume de données afin de découvrir des tendances ou des règles [45].

En bref, le data mining est l'art d'extraire des informations (ou même des connaissances) à partir des données [46].

6.2. TACHES DU DATA MINING

Nombreuses tâches peuvent être associées au Data Mining, parmi elles nous pouvons citer:

6.2.1. LA CLASSIFICATION

Pour faire une classification il suffit d'étudier les caractéristiques d'un nouvel objet afin de lui attribuer une classe prédéfinie. Les objets à classifier sont généralement des enregistrements d'une base de données, cette classification va permettre de mettre à jour chaque enregistrement en déterminant un champ de classe. La tâche de classification consiste à définir une classe assez précise et un ensemble d'exemples classés auparavant.Le but est la création d'un modèle qui peut être appliqué aux données non classifiées afin de les classifiées [47].

***Voici à ce propos quelques exemples de tâche de classification* :**

- Diagnostiquer la possibilité de l'existence d'une maladie.
- Déterminer si l'utilisation d'une carte de crédit est frauduleuse [48].

6.2.2. ESTIMATION

A partir des caractéristiques d'un objet, on peut arriver à faire l'estimation d'un champ en ce qui concerne sa valeur. Cette opération d'estimation peut se faire dans un but essentiel de classification. Il faut pour cela lui donner une classe particulière à un intervalle de valeurs d'un champ estimé. Par exemple : la classification peut concerner des événements discrets (le patient a été ou non hospitalisé). L'estimation peut alors se baser à des variables contenues à savoir la durée d'hospitalisation [49].

Voici ici quelques exemples sur l'utilisation des tâches d'estimation dans les domaines de recherche et commerce :

- Estimer le nombre d'enfants dans une famille [50].
- Estimer le montant d'argent qu'une famille de quatre membres choisis aléatoirement dépensera pour la rentrée scolaire [48].

6.2.3. LA PREDICTION

La classification et l'estimation tous deux ressemblent à la prédiction. Cependant, cela ce fait dans une échelle temporelle différente comme nous l'avons vu dans les tâches précédentes, tout s'appuie sur le passé et présent. Il y a seul le résultat qui appartient dans un futur à préciser. Parmi les techniques les plus appropriées à la prédiction sont [51] **:**

- L'analyse du panier de la ménagère (ou règles d'association)
- Le raisonnement basé sur la mémoire
- Les arbres de décision
- Les réseaux de neurones

Voici des exemples de tâche prédiction :

- Prédire au vu de leurs actions passées les départs de clients.
- Prévoir le champion de la coupe du monde en football en se basant sur la comparaison des statistiques des équipes [49].

6.2.4. REGLE D'ASSOCIATION

Grouper par similitude est une tache d'association. Cela va permettre de déterminer d'avance les attributs qui vont ensemble. Le data mining a pour fonction de donner un sens aux données, il faut pour cela en extraire les relations masquées et non triviales à utiliser la base de données. La technique la plus recommandée est celle qui consiste au regroupement par similitudes en faisant l'analyse de panier de la ménagère [52].

Voici un exemple de tâche d'association :

- Trouver dans un supermarché quels produits sont achetés ensemble et quels sont ceux qui ne s'achètent jamais ensemble.
- Déterminer la proportion des cas dans lesquels un nouveau médicament peut générer des effets dangereux [48].

6.2.5. LA SEGMENTATION

Il faut démontrer et arriver à trouver les observations qui s'associent sans pour cela privilégier aucune variable. On partage une certaine population hétérogène en sous groupes pour le rendre plus homogènes (clusters). A ce stade, les classes n'ont pas été définies. Ici la technique la plus appropriée à la segmentation est l'analyse des clusters [53].

6.2.6. DESCRIPTION

Data mining est des fois utilisé pour simplement décrire ce qu'il y a sur une base de donnée complexe pour expliquer les relations qui existent dans les données pour la bonne compréhension des individus , des produits et des processus existants sur cette base . Une bonne description d'un comportement implique souvent une bonne explication de celui-ci. Ici, la technique la plus appropriée à la description est l'analyse du panier de la ménagère [54].

7. DONNEES ET FOUILLE DES DONNEES

Une donnée est un enregistrement au sens des bases de données que l'on nomme aussi individu (terminologie issue des statistiques) ou instance (terminologie orientée objet en informatique) ou même tuple (terminologie base de données) et un point dans un espace euclidien ou un vecteur dans un espace vectoriel. Une donnée est caractérisée par un ensemble de champs de caractères ou encore d'attributs [55].

7.1. LES DIFFERENTS TYPES DE DONNEES

Un attribut peut être de nature **qualitative** ou **quantitative** en fonction de l'ensemble des valeurs qu'il peut prendre :

QUANTITATIVE	**CONTINU**	l'ensemble des valeurs qu'il peut prendre est réel ou un intervalle réel. Il s'agit donc d'un ensemble infini non dénombrable : on ne peut pas énumérer systématiquement l'ensemble de tous les points d'un intervalle réel. Par exemple, X peut être l'âge d'une personne prise au hasard, sa taille, son poids, …
	DISCRET	l'ensemble des valeurs qu'il peut prendre est un ensemble numérique ni comprenant un nombre ni d'éléments ou un ensemble infini dénombrable comprenant une infinité de nombres que l'on peut énumérer.
QUALITATIVE	**Nominal**	les valeurs sont justes des noms différents. par exemple : les codes postaux, les couleurs, le sexe, …
	Ordinal	les valeurs reflètent un ordre, rien de plus. Par exemple : bon, meilleur, mieux, ...

Tableau 2.4 : Les différents types de données [49].

7.2. DISTANCE ET SIMILARITE

7.2.1. NOTION DE SIMILARITE ET DISSIMILARITE

Gilles Bisson définit la similarité comme étant l'opérateur qui permet d'évaluer les ressemblances et les dissemblances qui existent au sein d'un ensemble de données. Il subdivise la similarité en deux grandes familles : les similarités numériques qui quantifient les ressemblances sous la forme d'une valeur dans l'intervalle [0,1] et les similarités symboliques qui permettent de caractériser les ressemblances [56].

Formellement, la similarité d(x, y) entre x et y est considéré comme une fonction à deux arguments satisfaisant les conditions suivantes [57] :

1. $\forall X_i, X_j \in X$; $dist(X_i, X_j) \geq 0$
2. $\forall X_i, X_j \in X$; $dist(X_i, X_j) = dist(X_i, X_j)$
3. $\forall X_i, X_j \in X$; $dist(X_i, X_j) = 0 \rightarrow X_{i=} X_j$
4. $\forall X_i, X_j, X_k \in X$; $dist(X_i, X_j) \leq dist(X_i, X_k) + dist(X_k, X_j)$

7.2.2. QUELQUES MESURES DE SIMILARITE

Il existe un grand nombre de mesures de similarité dans ce qui suit nous présentons quelques-unes des fonctions entre deux objets d(x1 ; x2) [49]:

a. DISTANCE EUCLIDIENNE

Le type de distance le plus couramment utilisé. Il s'agit d'une distance géométrique dans un espace multidimensionnel est définie comme suit :

D(X, Y) = $\sqrt{\sum_{i=1}(xi - yi)^2}$ tel que : **n** : nombre d'attributs

b. DISTANCE DE MANHATTAN :

Est définie comme suit :

D(x, y) = $\sum_{i=1}^{n}$**(xi − yi**) tel que **n** : le nombre d'attributs

c. DISTANCE DE MINKOWSKI

Cette distance est une généralisation de la distance euclidienne de la distance de Manhattan. Elle est calculée de la façon suivante :

D(x, y) = $\sqrt[p]{\sum_{i=1}^{n}(xi - yi)^{p}}$ **, p> 0**

Où p est un entier positif:
Pour **p= 1** : distance de Manhattan
Pour **p= 2** : distance euclidienne

8. LES TECHNIQUES DE DATAMINING

Nous distinguons différentes techniques de fouille de données. Ces techniques sont classées en deux grandes catégories: **les techniques non-supervisées** ou descriptives et les **techniques supervisées** ou prédictives [58].

Si l'on dispose seulement d'exemples non étiquetés et disponibles et si les classes et leur nombre ne sont pas connus on parle **d'apprentissage non supervisé**. Dans ce cas, l'apprentissage se ramène alors à cibler les groupes homogènes d'exemples existant dans les données, c'est-à-dire faire l'identification des groupes comme les exemples les plus similaires appartiennent au même groupe et que les exemples les plus différents sont alors séparés en différents groupes, la notion de similarité va être ramenée à une fonction de distance entre paires d'exemples. Autrement dit, il s'agit à ce niveau de rechercher la distribution sous-jacente des exemples dans leur espace de description.

Prenons ici un exemple celui de médecin qui peut être amené à s'intéresser particulièrement à une maladie donnée chez un ou des groupes de malades. Ces différents groupes de malades vont à leur tour refléter différentes causes possibles de la maladie [59].

Parmi les techniques d'apprentissage non supervisées les plus connues, on peut citer [60]:

- K-moyen
- K plus proches voisins
- Réseaux de Neurones avec cartes de kohonen
- Hierarchical Agglomerative Clustering
- Règles Associatives

Cependant si les classes possibles sont connues et si les exemples sont fournis avec l'étiquette de leur classe, on parle **d'apprentissage supervisé**. Dans ce cas, il s'agit alors d'utiliser les exemples fournis et déjà classés pour apprendre un modèle qui permette ensuite d'associer à tout nouvel exemple rencontré sa classe la plus adaptée. Prenons l'exemple d'application de l'apprentissage supervisé concernant la médecine : étant donnés les résultats d'analyse d'un malade et la connaissance de l'état d'autres malades pour lesquels les mêmes analyses ont été menées , il sera possible d'évaluer le risque de maladie de ce nouveau patient en fonction de la similarité de ses analyses avec celles des autres patients[59] .

Parmi les méthodes de classification supervisées les plus populaires, on peut citer [60]:

- Les réseaux de neurones
- les arbres de décision
- Les algorithmes génétiques
- Naïve Bayes
- Machines à vecteur supports (SVM)

Dans ce qui suit, nous allons présenter une description générale non exhaustive des algorithmes d'apprentissage.

8.1. REGLES D'ASSOCIATION

Les règles associatives sont des règles extraites d'une base de données transactionnelles et décrivant des associations entre certains éléments. Ces règles sont fréquemment utilisées dans le domaine de la distribution des produits où la principale application est l'analyse du panier de la ménagère dont le principe est l'extraction d'associations entre produits sur les tickets de caisse. L'objectif de cette méthode consiste à étudier la liste des produits achetés pour avoir des renseignements sur les clients et pourquoi ils font certains achats. La méthode consiste à trouver le genre des produits que les clients achètent en général ensemble.

Cette méthode peut être appliquée à tout secteur d'activité pour lequel il va s'intéresser dans la recherche des groupements potentiels de produits ou de services : services bancaires, services de télécommunications, par exemple. Elle peut être également utilisée dans le secteur médical afin de rechercher des complications causées à des associations de médicaments ou alors à la recherche de fraudes en recherchant des associations inhabituelles.

Une règle d'association est de la forme : Si **condition** alors **résultat**. Dans la pratique, nous nous limitons généralement à des règles où la condition se présente sous la forme d'une conjonction d'apparition d'articles et le résultat se constitue d'un seul article. Par exemple, une règle à trois articles sera de la forme : Si X et Y alors Z ; règle dont la sémantique peut être énoncée : Si les articles X et Y apparaissent simultanément dans un achat alors l'article Z apparaît. Parmi les algorithmes d'induction des règles associatives les plus connus sont : **APRIORI, FP-GROWTH, ECLAT**, … [61].

8.1.1. AVANTAGES ET INCONVENIENTS

Parmi les avantages et les inconvénients des règles d'association, nous citons :

a. LES AVANTAGES

- Résultats clairs : règles faciles à interpréter.
- Simplicité de la méthode et des calculs (calculs élémentaires des fréquences d'apparition).
- Aucune hypothèse préalable (Apprentissage non supervisé) [62].

b. LES INCONVENIENTS

- la méthode est coûteuse en temps de calcul.
- Qualité des règles : production d'un nombre important de règles triviales (des règles évidentes qui, par conséquent, n'apportent pas d'information) ou inutiles (des règles difficiles à interpréter provenant de particularités propres à la liste des achats ayant servi à l'apprentissage) [62].

8.2. LES ARBRES DE DECISION

Les arbres de décision constituent l'une des structures de données essentielles d'apprentissage statistique. Leur fonctionnement se base sur des heuristiques qui peuvent satisfaire l'intuition mais encore donnent des bons résultats en pratique notamment lorsqu'ils sont utilisés en « forêts aléatoires ». Leur structure arborescente va les rendre également lisibles par un être humain et ce contrairement à d'autres approches où le prédicteur construit, est une « boîte noire ».Un arbre de décision est donc tout simplement une structure qui permet de faire la déduction d'un résultat à partir de décisions successives afin de parcourir un arbre de décision et trouver ainsi une solution. Pour ce faire il faut partir de la racine [63].

Tout nœud est ou bien une feuille qui dénote une décision ou bien une branche qui spécifie un test sur la valeur d'un attribut. Le nombre de descendants de chaque nœud dépendra alors des résultats du test effectué à ce niveau. L'arbre de décision étant un arbre au sens informatique du terme. C'est sous une forme graphique qu'il est représenté, il est aussi représenté par une arborescence ou encore d'un diagramme illustrant des règles de décision. Un arbre décisionnel est un outil d'aide à la décision qui représente graphiquement un séquencement logique pour faire apparaître les différents résultats possibles et ce à partir des choix faits à chaque étape [64].

Un arbre de décision va aussi modéliser une hiérarchie de tests sur les valeurs d'un ensemble de variables qu'on appelle attributs. Les arbres de décision permettant de faire des classifications sur des données représentées par des ensembles d'attributs. On distingue deux catégories d'arbre de décision [65]:

- **Arbre de classification** : ici la variable expliquée est de type nominal (facteur). L'essentiel est de réduire à chaque étape du partage l'impureté totale des deux nœuds fils par rapport au nœud père.
- **Arbre de régression** : ici la variable expliquée est de type numérique et il s'agit de prédire une valeur très proche possible de la vraie valeur.

8.2.1. CONSTRUCTION D'UN ARBRE DE DECISION

Il existe une grande variété d'algorithmes pour construire des arbres de décision, nous citons quelques uns les plus répandus :

- ✓ **ALGORITHME ID3**

Il a été développé par Quinlan en 1986, ID3 est un algorithme de classification supervisé. Il se base essentiellement sur des exemples classés au préalable afin de générer des arbres de décision [66].

Partant du théorème de Shanon 1948, ID3 fait usage d'entropie[1] pour évaluer le désordre des données. Auparavant Shanon avait utilisé ce théorème pour calculer la taille minimale afin de coder un message [67].

Cet algorithme donne la possibilité de construire récursivement un arbre de décision. Il va se charger de calculer parmi les attributs restant celui qui va générer le plus d'informations qui permettront de classer les exemples d'un niveau quelconque de l'arbre de décision [65].

- ✓ **ALGORITHME C4.5 (J48)**

L'algorithme C4.5 est un algorithme de classification supervisé. Proposé par Quinlan en 1993, il se base principalement sur ID3 en lui apportant certaines améliorations qui permettront de résoudre les problèmes causés par ID3, comme par exemple [68] :

- Le traitement des attributs continus.
- Le traitement des valeurs nulles pour un attribut
- C4.5 permet de grouper l'ensemble des valeurs discrètes nominales, pour un attribut quelconque, afin de supporter des essais plus complexes.

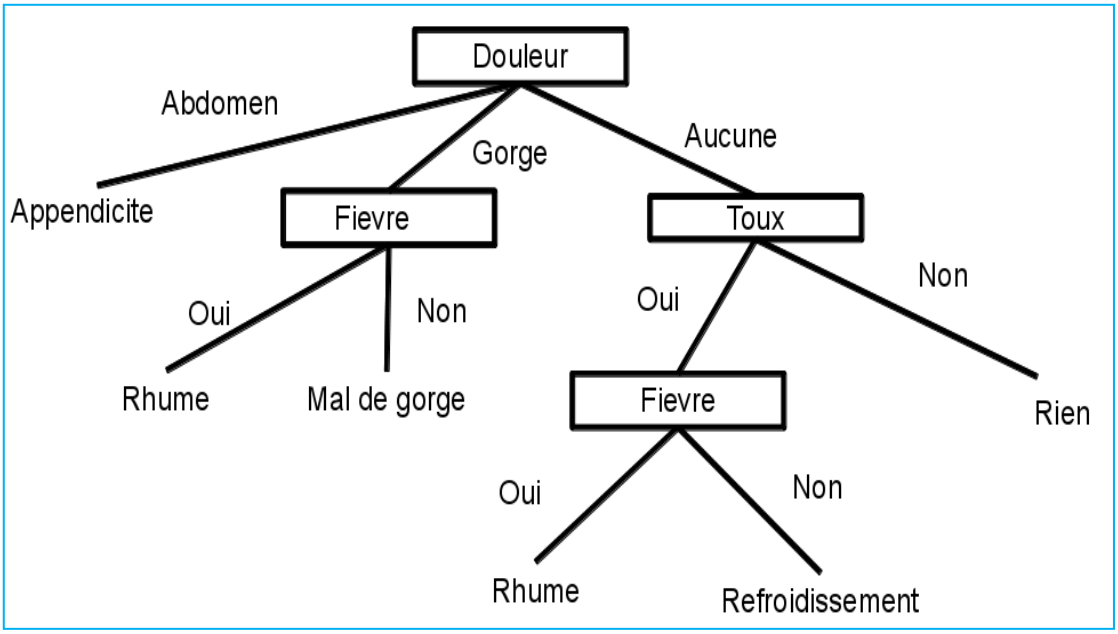

Figure 2.3: Exemple d'arbre de décision [69].

[1] **L'entropie de Shannon**, due à Claude Shannon, est une fonction mathématique qui intuitivement correspond à la quantité d'information contenue ou délivrée par une source d'information.

8.2.2. AVANTAGES ET INCONVENIENTS

Parmi les avantages et les inconvénients d'arbre de décision, nous citons :

a. LES AVANTAGES

- Les arbres de décision sont capables de produire des règles compréhensibles.
- Les arbres de décision effectuent la classification sans exiger beaucoup de calcul.
- Les arbres de décision sont en mesure de manipuler à la fois les variables continues et catégorielles [70].

b. LES INCONVENIENTS

- Manque de performance dans le cas de plusieurs classes; les arbres deviennent très complexes et ne sont pas nécessairement optimaux.
- Moins bonnes performances concernant les prédictions portant sur des valeurs numériques [71].

8.3. CLUSTERING

On définit le clustering comme une approche de classification avec des classes qui existent déjà au préalable. Au début, on a un ensemble d'objets non étiquetés (dont la classe est inconnue). En partant de ces objets, il faut arriver à détecter des objets similaires pour les regrouper dans des classes [72].

Selon [73], le clustering est une manière de diviser une population d'objets en sous ensembles d'objets appelés dans une classe pour les regrouper dans une même classe afin qu'ils soient similaires et séparer les objets de classes distinctes afin qu'ils soient dissimilaires. Parmi les buts que le clustering doit atteindre, il faut citer la réduction des données et la prédiction basée sur les groupes.

8.3.1. LES ALGORITHMES DE CLUSTERING

La majorité des algorithmes de clustering peuvent être répartis en trois grandes familles [74, 75,76]:

- **Les algorithmes de partitionnement** : Ils adoptent une recherche itérative des classes jusqu'à l'optimisation d'un critère d'arrêt qui peut être le nombre de classes désiré, le nombre minimum (ou maximum) d'objets dans chaque classe, le nombre d'itérations.

- **Les algorithmes hiérarchiques** : Ils sont descendants si, à partir d'une seule classe, ils cherchent à établir une partition par division ; ou ascendants s'ils cherchent à former des classes plus grandes par fusion de classes jusqu'à la satisfaction d'un critère d'arrêt.

- **Les algorithmes à base de densité** : Les classes sont formées selon le voisinage des objets et le niveau de densité de chaque objet.

a. CLUSTERING HIERARCHIQUE

Ces méthodes forment les classes graduellement sous forme hiérarchique, autrement dit, un arbre appelé un dendrogramme. Les algorithmes basés sur cette méthode vont essayer de créer une hiérarchie de clusters, les objets les plus similaires seront regroupés dans des clusters aux plus bas niveaux alors que les objets les moins similaires vont être regroupés dans les clusters aux plus hauts niveaux. A ce niveau, on peut distinguer deux sous types :

1) Agglomération (ascendant) : On commence par considérer chaque point comme une classe puis on essaie de faire la fusion de deux ou plusieurs classes appropriées et ce en fonction d'une similarité afin de former une nouvelle classe. Le processus est itéré jusqu'à ce que tous les points se trouvent dans une même classe.

2) Division (descendant) : En considérant tous les points comme une seule classe au départ, on divise alors successivement les classes en classes plus raffinées. Le processus sera réitéré pour que chaque classe contienne un seul point ou bien si le nombre de classe voulu est atteint.

Il y a de nombreux algorithmes hiérarchiques qui sont proposés dans la littérature, les plus connus sont : **BIRCH** pour le groupement agglomératif et **DIANA** pour division [77].

b. CLUSTERING PAR PARTITIONNEMENT

Le partitionnement de données a pour but de diviser un ensemble de données en différents clusters homogènes. Son principe est la subdivision de l'ensemble des individus en un certain nombre de classes et ce par l'emploi d'une stratégie d'optimisation itérative qui a pour principe général de générer une partition initiale et de chercher à l'améliorer en réattribuant des données d'une classe à l'autre. Contrairement aux algorithmes hiérarchiques qui donnent une structure de classes, les algorithmes de partitionnement produisent en ce qui les concerne une seule partition en recherchant alors des maxima locaux en optimisant ainsi une fonction objective traduisant le fait que les individus doivent être similaires au sein d'une même classe et dissimilaires d'une classe à une autre. Les algorithmes de partitionnement sont divisés en trois sous familles : **la méthode k-means, k-medoids, k-nn** [78].

c. CLUSTERING BASE SUR LA DENSITE

Ces algorithmes considèrent les classes comme étant des régions denses dans l'espace d'objets. Pour qu'un objet de l'espace soit dense il faut que le nombre de ses voisins dépasse un certain seuil fixé au préalable. Ils essayent alors de faire l'identification de classes en se basant sur la densité des objets dans une région. On groupe alors les objets non pas sur la base d'une distance mais sur la base de la densité de voisinage qui dépasse une certaine limite. Parmi les algorithmes les plus connus dans cette catégorie, nous citons DBSCAN et OPTICS [76].

9. DOMAINES D'APPLICATION DU DATA MINING

Le data mining est une spécialité transverse, elle regroupe un ensemble de théories et d'algorithmes ouverts à tout domaine susceptible de drainer une masse importante de données. Parmi ces domaines, on cite [77]:

9.1. LE SECTEUR BANCAIRE

- Identifier les clients "fidèles".
- Identifier les clients qui seront les plus réceptifs aux nouvelles offres de produits.
- Prédire les clients qui sont susceptibles de changer leurs cartes d'affiliation au cours du prochain trimestre.

9.2. LA DETECTION DE FRAUDE

La fouille de données est largement appliquée dans des processus de détection de fraude divers tel que :

- Détection de fraude de cartes de crédits.
- Détection de fausses demandes de remboursement médicale.

9.3. LE SECTEUR DES ASSURANCES

- Evaluation du risque d'un bien assuré prenant en compte les caractéristiques du bien et de son propriétaire.
- Formulation des modèles statistiques des risques d'assurance.

9.4. LA MEDECINE

- Prédiction de présence de maladies.
- Approvisionnement des médicaments les plus fréquemment prescrits [79].

10. MOTIVATIONS DU DATA MINING

10.1. EXPLOSION DES DONNEES

- **Volume des données : masse** importante de données (millions de milliards d'instances).
- données multi dimensionnelles (milliers d'attributs).
- Inexploitables par les méthodes d'analyse classiques.
- collecte de masses importantes de données (gbyte/heure).
- besoin de traitement en temps réel de ces données.

10.2. AMELIORER LA PRODUCTIVITE

- forte pression due à la concurrence du marché.
- besoin de prendre des décisions stratégiques efficaces.

✓ exploiter le vécu (données historiques) pour prédire le futur et anticiper le marché.

10.3. CROISSANCE EN PUISSANCE /COUT DES MACHINES CAPABLES

- de supporter de gros volumes de données.
- d'exécuter le processus intensif d'exploration.
- hétérogénéité des supports de stockage [80].

11. TEXT MINING

L'extraction de connaissances à partir de textes est un processus non trivial qui construit un modèle de connaissances valide, nouveau, potentiellement utile et au final compréhensible à partir de textes bruts. Donc l'objectif de la fouille de texte est le traitement de grandes quantités d'information qui sont disponibles sous une forme textuelle et non structurée [81].

12. SOUND MINING

Consiste à rechercher des éléments communs ou à classer des sons en fonction de leur contenu. Les applications potentielles sont l'indexation et la recherche des pièces musicales à partir des bases de données. Ces applications permettent aux utilisateurs de fouiller et recouvrer la musique, non seulement au moyen des requêtes textuelles (comme le titre, l'orchestre, le conducteur, la chanson de texte, le compositeur), mais aussi en fonction du contenu ou bien par une combinaison du son et du texte [82].

13. IMAGE MINING

Il s'agit de rechercher des relations entre les images ou des séquences d'images. L'image mining consiste à réaliser un système intelligent capable de fouiller dans les bases de données multimédias afin de fournir des images définies par l'utilisateur. Un des défis principaux dans la compréhension d'image est de développer un système flexible, adaptable, capable d'exécuter des tâches d'analyse d'image complexes et d'extraire des connaissances [83].

14. VIDEO MINING

Le Video Mining consiste à rechercher des éléments communs ou à classer des vidéos en fonction de leur contenu. Les applications potentielles sont l'indexation de bases de films ou l'optimisation des grilles de programmes des opérateurs de télévision. Actuellement, les systèmes de recherche de vidéo assurent une indexation par le contenu de documents vidéo. Ces systèmes sont basés sur une indexation automatique partielle dont les résultats peuvent être corrigés ou complétés par l'utilisateur en fonction de compromis coût-qualité recherché. Plusieurs briques de base de ces systèmes sont développés (segmentation en plans, recherche des mouvements de caméras, détection et suivi d'objets mobiles, caractérisation par vecteurs de caractéristiques, détection de personnages). Parallèlement, à ces traitements effectués sur la bande image, une indexation à partir de la transcription de la bande son ou des sous-titres lorsqu'ils sont disponibles sera intégrée et couplée à des systèmes d'indexation d'image [84].

15. CONCLUSION

Le processus d'ECD est géré par un analyste et un expert dans le domaine étudié. Ce processus va permettre de répondre aux buts de la veille stratégique. Il donne la possibilité d'offrir une vue synthétique et pertinente par la révélation qu'il fait des informations endogènes. Ces informations vont montrer les tendances, les signaux faibles, etc. d'un domaine donné.

L'objectif de l'ECD étant donc de permettre à l'expert de retrouver en fonction d'un corpus donné, des relations connues dans son domaine, de pouvoir les localiser de manière explicite, d'analyser les acteurs à partir d'une ou plusieurs de ces relations. L'ECD permet aussi de faire la découverte de nouvelles relations.

Dans ce chapitre, nous avons met en évidence la notion d'ECD et son processus. Nous avons également montré l'importance d'ECD pour aider un diagnostic des données et plus généralement pour l'aide à la décision.

Nous avons vu que le processus de KDD comporte une phase très importante qui est celle de Data Mining que nous avons présenté tout au long de ce chapitre ainsi ses tâches, ses techniques, etc.

Nous pouvons dire que, le Data Mining est la technologie idéale pour l'extraction de l'information des données et son importance est croissante dans les sociétés qui en détiennent en grande quantité. Il complète aussi d'autres modèles d'analyse en aidant les utilisateurs à naviguer et explorer un nombre important d'entrepôts de données pour faciliter et donner plus rapidement l'information cachée dans l'énorme volume de données. Le chapitre suivant est consacré à notre contribution qui a comme objectif ; l'amélioration de la qualité des données structurées et non structurées.

CHAPITRE 3
CONTRIBUTION

PLAN DU CHAPITRE

1. INTRODUCTION

La qualité des données n'a cessé de prendre une place de premier plan au sein des communautés de recherche en ECD .En effet, l'extraction des connaissances et la prise des décisions peuvent être réalisées sur des données de qualités médiocres (des données inexactes, incomplètes, ambigües, incohérentes et contenant des doublons). On peut alors s'interroger sur le sens à donner aux résultats de ces analyses et remettre en cause la qualité des connaissances ainsi les décisions prises ne seront pas efficaces. La mauvaise qualité des données est l'un des problèmes principaux rencontrés par les entreprises lorsqu'elles mènent des projets décisionnels. De ce fait, Le nettoyage des données fait partie des stratégies d'amélioration automatique de la qualité des données. Si nous ne mettons pas en place aucune gestion de la qualité des données, le système pourra rapidement être saturé de données manquantes ou incorrectes. Les problèmes de qualité des données se répandent de façon endémique à tous les types de données c'est-à-dire des données structurées ou données non structurées et dans tous les domaines d'applications. Les conséquences des données de mauvaises qualités sur les prises de décision et les coûts financiers qu'elles engendrent sont considérables. Dans ce chapitre, nous proposons une approche théorique basée sur le nettoyage des données structurées et non structurées.

2. BIG Data MINING: TRAVAUX DE RECHERCHE

Avec la croissance des données, l'utilisation des techniques de data mining et la découverte des informations précieuses cachées dans les BIG DATA est devenue de plus en plus importante.

Divers techniques existantes de data mining souvent par le biais de l'exploration des **Fréquent ItemSet** sont utilisés pour dériver des règles d'association et accéder aux connaissances pertinentes mais avec l'arrivée de l'ère de big data, les algorithmes traditionnelle de datamining ont été incapable de répondre aux besoins de l'analyse des big data.

1. **Le papier [85]** propose un algorithme parallèles basé sur MapReduce appelé **MRPrePost** sur la base de **PrePost** et décrit en détail la mise en œuvre de l'algorithme. **MRPrePost** est un algorithme parallèle basé sur la plateforme hadoop qui améliore PrePost par l'ajout d'un motif de préfixe, ce qui rend **MRPrePost** un algorithme adapté à l'exploitation des règles d'association associé au big data.

Les expériences montrent que l'algorithme MRPrePost est plus performant par rapport à PrePostet et la stabilité et l'évolutivité de l'algorithme MRPrePost est meilleur.

Le papier propose un algorithme parallèle basé sur **MRPrePost** appelé MRPrePost sur la base de PrePost et décrit en détail la mise en œuvre de l'algorithme.

Face à l'exploitation minière de grands ensembles de données, la parallélisassions est une bonne solution, les résultats expérimentaux.

Frequent itemset mining est un important sujet de recherche, car il est largement appliqué dans le monde réel il sert à trouver les motifs fréquents et les motifs derrière le comportement humain. Le processus **FIM** est gourmant en mémoire et en calcul. Comme les données croit de façon exponentielle chaque jour, l'achèvement et l'efficacité et le passage à l'échelle devient plus austère.

2. **Dans l'article [86]**, l'auteur propose un nouveau algorithme de la famille FIM ce dernier se distingue par sa possibilité à être implémenter sur la plateforme MapReduce. L'algorithme applique l'idée de l'ordre lexicographique pour construire un arbre appelé arbre de séquence lexicographique qui permet de trouver tous les motifs fréquents dans les bases de données de transaction sans recherche exhaustive.

En outre, breadth-wide support based pruning :est également un contributeur majeur dans l'efficacité et le passage à l'échelle de cet algorithme .

Pour tester les performances de son algorithme, l'auteur a mené de diverses expériences sur le cadre de map reduce avec les datasets de taille massive. Les résultats montrent l'impact breadth wide support based pruning et map reduce sur l'efficacité et le passage à l'échelle de l'algorithme.

3. **Lin et al [87]** ont proposés Single Pass Counting , Fixed Passes Combined Counting (FPC) et Dynamic Passes Counting (DPC) qui établit le comptage des étapes en paralléle par la distribution des dataset a travers les différents mappers .

4. **Hammoud** a proposé **MR Apriori [88]** une approche pour trouver des motifs fréquents par commutation entre la disposition verticale et horizontale itérative qui élimine le besoin de numération itérative des données. Il répète l'analyse d'autres données intermédiaires et elle est réduite avec chaque itération.

5. **MREclat [89]** est un algorithme Eclat modifié dans le cadre de mapreduce qui génère une liste de motifs fréquents, la listeest divisée en classes d'équivalence puis pour chaque classe d'équivalence frequent itemset sont calculés en utilisant le cadre de mapreduce.

6. **Parallel FP-Growth** (PFP) **[90]** a exploité itemsets tag à partir de laquelle la page webitemsets sont générés ce qui nécessite deux balayages sur la base de données. En utilisant mapreduce et son mécanisme de tolérance aux pannes, la tâche de big data mining est convertie en d'autres petites taches qui ne sont pas dépendantes les uns des autres.

7. La stratégie de regroupement des PFP a des problèmes de mémoire et de vitesse pour équilibrer les groupes de PFP **Zhou et al [91]** a proposé un algorithme pour une exécution plus rapide en utilisant des éléments simples qui n'est pas également une façon efficace.

8. **Moens et al [85]** ont proposés deux méthodes pour l'extraction des motifs fréquents pour big data sur mapreduce, première méthode Dist Eclat la version distribue de Eclat qui optimise la vitesse en répartissant l'espace de recherche de manière égale entre mappers, deuxième méthode BigFIM utilise à la fois la méthode basé Apriori et Eclat avec des bases de données projetées qui conviennent à la mémoire afin d'extraire frequent itemset.

3. NETTOYAGE DE DONNEES : TRAVAUX DE RECHERCHE

1. **J. R. Quinlan [92]** propose de traiter la valeur manquante comme une nouvelle valeur pour chaque attribut et donc comme toute autre valeur que peut prendre l'attribut. L'inconvénient de cette méthode vient du fait qu'elle se prête bien à l'analyse de valeurs manquantes catégorielles, mais plus difficilement à celle de valeurs manquantes au hasard. Il présente donc également une méthode, plus appropriée dans ce deuxième cas. Celle-ci est basée sur l'idée que les cas contenant des valeurs manquantes sont distribués de manière homogène dans l'ensemble d'apprentissage et attribue un statut différent â la valeur " inconnu ". Cependant, cette méthode traite spécifiquement chacune des valeurs manquantes et ne tient pas compte de la structure de l'ensemble de données, elle n'utilise donc pas l'intégralité de l'information disponible.

2. **S. G. Thompson et al [93]** utilise les informations disponibles (valeurs de l'attribut pour la classe, valeurs des autres attributs pour les cas de la même classe...) afin de déterminer les valeurs manquantes. Toutefois, il apparaît que cette technique n'est appropriée que pour une faible concentration de données incomplètes et un nombre limité d'attributs non-renseignés (explosion combinatoire).

3. **R. Pearson [94]** souligne notamment les problèmes liés à la détection des valeurs manquantes qui ne doivent pas être traitées de la même façon que des attributs volontairement non renseignés. A l'inverse dans certains cas, les valeurs inconnues, inapplicables ou encore non spécifiées sont encodées comme des valeurs valides.

4. **Ragel et al** [**94**] dans le cadre de la recherche de règles d'association**,** présentent un algorithme afin de traiter les valeurs manquantes. Celui-ci ne fait pas intervenir la logique floue mais divise la base de données en sous-ensembles complets. Par ailleurs, si la logique floue permet dans certains domaines de traiter les incomplétudes dans le cadre de la découverte de règles d'association et de motifs séquentiels, l'introduction de la logique floue dans les algorithmes d'extraction a permis de traiter un nouveau type d'attributs, les attributs quantitatifs.

La recherche bibliographique que nous avons réalisé nous a permis de constater que les travaux de recherche d'extraction des connaissance à partir des bases de données s'appuient sur les données structurées c'est-à-dire donnée relationnelle. Cependant sont rares les travaux qui traitent les big data. Ainsi, nous proposons une approche d'extraction des connaissances à partir des big data en concentrant sur le nettoyage des données ou autrement dit amélioration de la qualité des big data**.**

4. APPROCHE D'AMELIORATION DE LA QUALITE DES BIG DATA PROPOSEE

Dans cette section, nous présentons notre proposition :

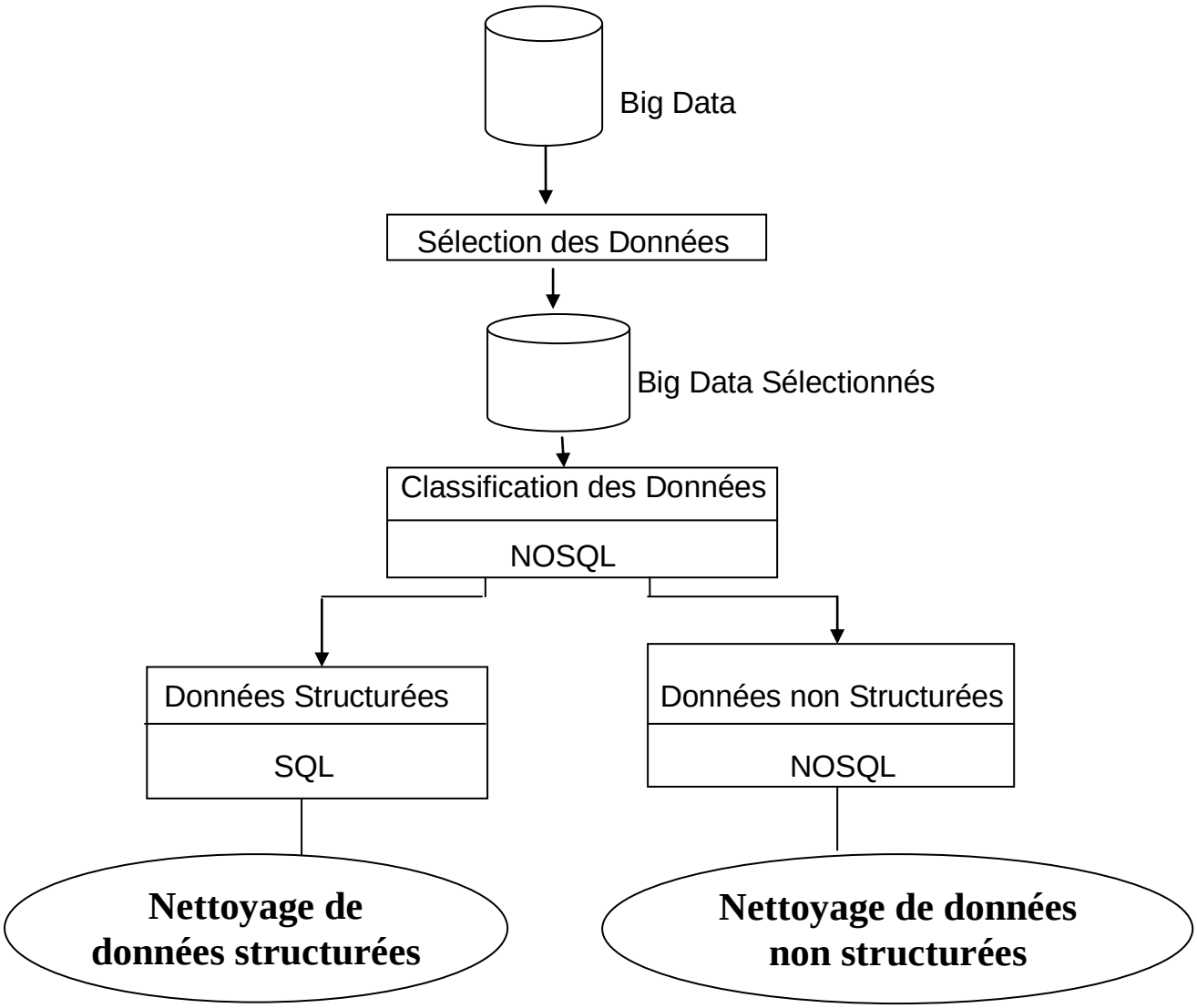

Figure 3.1 : Approche d'amélioration de la qualité proposée

La présente approche, après l'extraction des données et leur sélection sont stockées dans des bases de données temporaires (pour ne pas perturber les traitements transactionnels et aussi pour ne pas toucher aux valeurs originales des données). Le système proposé, dans une première phase et à l'aide des requêtes de type SELECT, selon les langages NOSQL divise ces données en deux classes selon leur type :

- Données structurées.
- Données non structurées.

Pour chaque type, nous proposons un traitement d'amélioration de qualité lui correspondant.

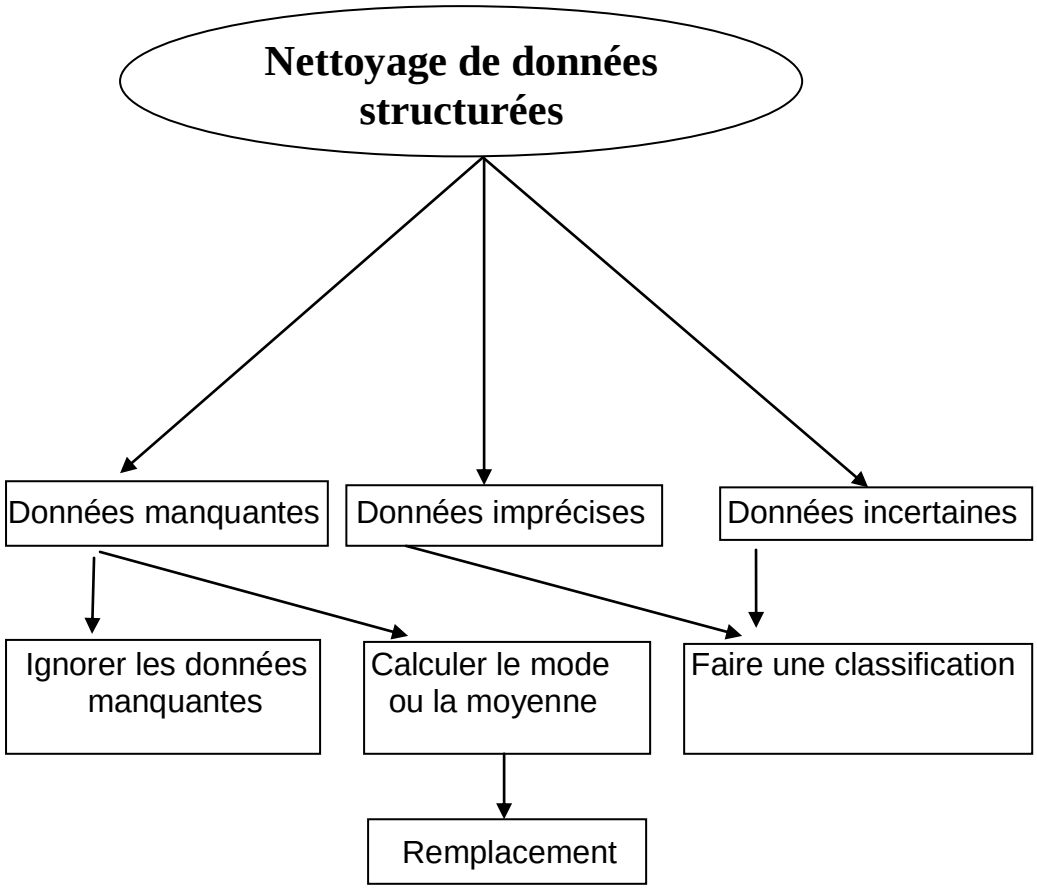

Figure 3.2 : Amélioration de la qualité des données structurées.

Pour réaliser le traitement et également l'amélioration de la qualité des données structurées, il faut d'abord faire le nettoyage de ces mêmes données qui sont de trois types : Données manquantes, données imprécises et données incertaines. Puisque ces dites données sont de mauvaises qualités, pouvant influer d'une manière assez négative sur la prise de décision. Là, on peut citer le problème des données manquantes étant donné que l'absence de connaissances précises ou l'inexistence de connaissances ont des effets très négatifs sur le résultat auquel on veut aboutir.

Afin de trouver une solution adéquate à ces différents problèmes et difficultés, on peut avoir recours à deux solutions : la première solution va consister à ignorer ces données manquantes. Il s'agira de modifier la dite technique de calcul par une similarité qui va en quelque sorte supprimer les attributs manquants.

Cependant, on peut remarquer que la solution adoptée n'est pas toujours efficace dans la mesure où le taux de manque est volumineux. Dans la seconde solution, nous proposons de remplacer les données manquantes par l'utilisation d'une des données techniques statistiques connues et assez maitrisées tels que le calcul de mode ou de la moyenne. Il y a également le problème des données imprécises et incertaines, dans n'importe quel domaine et contrairement aux attributs possédant obligatoirement des valeurs exactes, il doit y avoir des attributs descripteurs des cas pouvant faire l'objet d'imprécision et d'incertitude, pour mieux prendre en charge ce type d'attributs , on peut passer par des variables linguistiques qui faciliteront énormément la tâche dans le cas des valeurs numériques précisément et ce par la tolérance de quelques imprécisions dans le processus de mesurage.

Ces mêmes valeurs qui ne sont aucunement statistiques peuvent à leur tour changer d'un domaine à un autre. Notre suggestion quant à la solution pour faire le traitement des données imprécises et incertaines est la suivante : Etant donné que l'information imprécise en général peut avoir différentes valeurs possibles et peut même se présenter sous la forme d'intervalle, l'approche intuitive et logique va être l'utilisation des intervalles afin de faire face au problème des données imprécises surtout. Cette solution permettant d'éliminer efficacement l'effet des points sources d'aberrance et réduire en partie la possibilité infinie des valeurs de variables continues. La solution adéquate sera pour ce type de données est que pour chaque attribut découper l'espace des valeurs en intervalles classes par la transformation des valeurs continues.

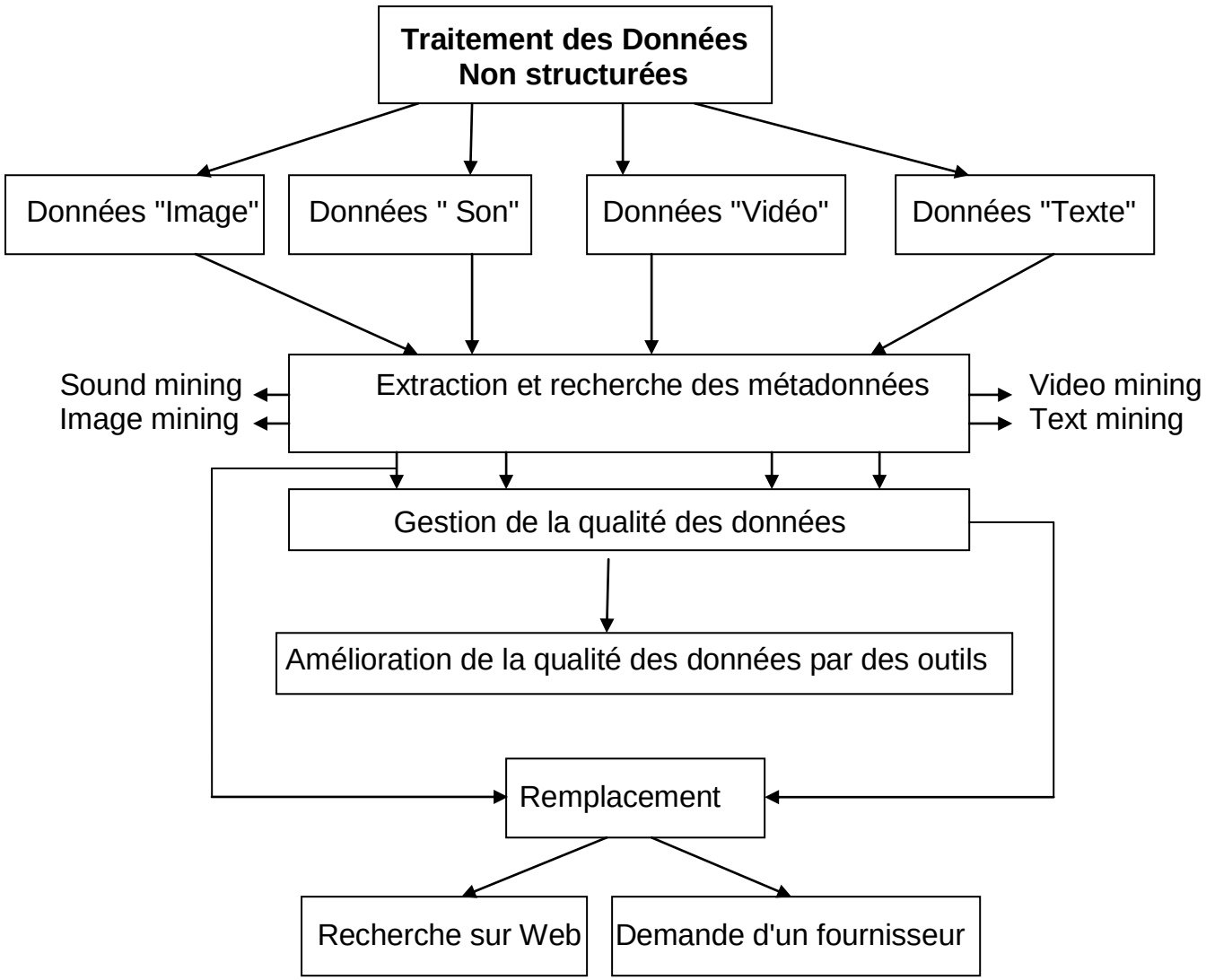

Figure 3.3 : Amélioration de la qualité des données non structurées.

Pour réaliser le traitement ainsi l'amélioration de la qualité des données non structurées qui se divisent en quatre types : données "image", données "son", données "vidéo" et données "texte" ; on se base sur deux phases :

La première phase consiste à retrouver et extraire les métadonnées à partir des Big data classifiées. Les métadonnées sont des données sur les données ; elles définies en général un ensemble d'informations descriptives. Par exemple parmi les métadonnées d'une image, on peut citer : Le type de l'image, le nom de l'auteur, la date de création de cette image, son titre et son contenu. Pour réaliser l'opération d'extraction des métadonnées, on peut utiliser les techniques : Sound Mining, Image Mining, Vidéo Mining et Text Mining.

La seconde étape consiste à gérer la qualité des données avec la possibilité d'améliorer davantage la qualité de ces données non structurées soit par l'usage d'outils tels que : Audacity, un logiciel libre de montage audio ; Wondershare Filmora, un bon logiciel de montage vidéo : Windows Movie Maker qui permet d'améliorer la qualité des vidéos mais également d'en produire, et aussi l'outil Pinnacle Studio, Movavi Photo Editor, GIMP permettant l'amélioration de la qualité des images et Blue Text Analytics pour améliorer la qualité d'un texte.

Une autre technique d'amélioration de qualité est le remplacement **;** pour cela on peut faire une recherche sur web afin d'avoir des meilleures données (gratuitement), sinon on va s'adresser à un fournisseur. Le remplacement se fait à l'aide des métadonnées extraites au préalable.

5. CONCLUSION

Dans ce chapitre, nous avons présenté quelques travaux de recherche de façon générale cependant nous n'avons pas constaté la présence de travaux de recherche qui traitent les big data autrement dit les travaux de recherche sur l'amélioration de la qualité des big data. Ainsi, nous avons présenté une approche permettant essentiellement l'amélioration de la qualité des données structurées et non structurées.

Conclusion générale

Les systèmes d'information des entreprises actuelles sont de plus en plus débordés par des données de tous types : structurées (bases de données, entrepôts de données) et non structurées (textes et multimédias). Pour cela de nouveaux défis ont été apparus, que ce soit pour les entreprises ou pour la communauté scientifique, parmi lesquels " comment comprendre et analyser de telles masses de données " afin d'en extraire des connaissances.

Par ailleurs, dans une organisation, un projet d'Extraction de Connaissances à partir de Données est le plus souvent mené par plusieurs experts (experts de domaine, experts d'ECD, experts de données…), chacun ayant ses préférences, son domaine de compétence, ses objectifs et sa propre vision des données et des méthodes de l'ECD.

Cependant, il se trouve que dans beaucoup de domaines, les données présentées sont incomplètes et/ou imprécises même de mauvaises qualités, ce qui entraine une exploitation très difficile et/ou impossible suite au taux élevé de la dimensionnalité des bases de données qui rend la tâche d'extraction de connaissances à partir de cette masse de données difficile ou complexe .

Dans ce travail, pour intervenir contre les données de mauvaises qualités, nous avons proposé une approche théorique basée sur deux phases : une phase pour les données structurées et l'autre pour les données non structurées.

Dans nos futurs travaux nous désirerons implémenter notre approche.

BIBLIOGRAPHIE

[1] Marinela Mircea, « Perspectives on Big Data and Big Data Analytics », Journal voll. III Data base Systems, 07 / 01/2013.

[2] Pirmin LEMBERGER, Marc BATTY, Médéric MOREL, Jean-Luc Raffaelli. Big Data et machine learning: Manuel du data scientist, Dunod, 2015.

[3] M.CORINUS, T.Derey, J.Marguerie, W.Techer, N.Vic, Rapport d'étude sur le Big Data, SRS Day, 54p, 2012.

[4] Maxime VIGIER, Mémoire professionnel présenté dans le cadre de la licence Marketing et Commerce sur Internet,Université d'Évry Val d'Essonne , 2014.

[5] Khadidjatou BAMB, Comprendre le BIG DATA. 2015.

[6] Christophe PARAGEAUD, Technologies Big Data : le livre blanc ,2016.

[7] Angeline KONE, Big Data (Rapports de stage), INSA LYON – Mastère spécialisé SI, 2013.

[8] Bernard Dousset, Le Big Data Mining enjeux et approches techniques, Institut de Recherche en Informatique de Toulouse UMR 5505, consulté le (11/01/2014).

[9] Christoph Spengler. Big Data – exploration de données temps réel - Direct Focus, Publié le: 7 février 2013.

[10] BENALLAL Zeyneb - TAHRAOUI Hayet, Etude comparative des bases de données NoSQL, mémoire de fin d'études pour l'obtention du diplôme de Master, Université Abou BakrBelkaid– Tlemcen, 2016.

[11] Olivier JOUANNOT .Présentation Générale Big Data, Guide Share France ,2013.

[12] CasyMcTaggart, « hadoopmapreduce », conférence présentation cadre orienté objet CSCI 5448 ,31/03/2011.

[13] Benjamin Renaut, « Hadoop/Big Data »,2016.

[14] Abdeljalil FELLAH, Abdelhamid BAACH, «Une Approche Scalable pour le Traitement de grande Quantité de Données », mémoire pour l'obtention du diplôme de Master en Informatique, université HassibaBenbouali de Chlef, 2016.

[15] Christophe PARAGEAUD, Big data : La jungle des différentes distributions open Source hadoop. Article Web, 2013.

[16] Matteo Di Maglie, Adoption d'une solution NoSQL dans l'entreprise, Haute École de Gestion de Genève (HEG-GE), Carouge, 12 septembre 2012 Haute École de Gestion de Genève (HEG- GE).

[17] Kouedi Emmanuel, Approche de migration d'une base de données relationnelle vers une base de données NoSQL orientée colonne, Mémoire présenté en vue de l'obtention du diplôme de MASTER II RECHERCHE, Option **:** S.I & G.L ; Université de YAOUNDE I, Mai 2012.

[18] FOUCRET Aurelien. NoSQL– Une nouvelle approche du stockage et de la manipulation des données. Smile. 55 p. [en ligne]. Disponible sur : http://www.smile.fr/Livres blancs/Culturedu-Web/ NoSQL (consulté le24/06/2015).

[19] Khaled Tannir, Introduction aux bases de données NoSQL, Montréal- 01 Décembre 2015.pdf

[20] BOUQUIN Sylvain, « Mise en œuvre d'un Système de Publication/Souscription basé sur les Flux D'Information de type RSS », mémoire d'ingénieur du CNAM, CONSERVATOIRE NATIONAL DES ARTS ET METIERS PARIS, 2016.

[21] Goetlas B, Zaki M.-J., "FIMI'03: Workshop on Frequent Itemset Mining plementation", FIMI'03 Workshop on frequent Itemset Mining Implementations, 2003.

[22] Piatsky-Schapiro G., Frawly W.J., "Knowledge Discovery in Databases", AAAI Presse, The MIT Press, Menlo Park, California, 1991.

[23] KODRATOFF Y, "techniques et outils de l'extraction de connaissances à partir des données", Signaux n°92 pp 38-43, Mars 1998.

[24] U. Fayyad, G. Piatetsky-Shapiro, et P. Smyth. "From data mining to knowledge discovery in databases". In AI Magazine, pp. 37-54,1996.

[25] Zighed D.A., kodratoff Y., Napoli A, "Extraction de connaissance à partir d'une base de donnée" Bulletin AFIA'01,2001.

[26] Alouane Basma, « Recherche de partitions floues optimales par segmentation floue pour la fouille de données quantitatives », Mémoire de magister, Université Mohamed BOUGARA de BOUMERDES.

[27] L. Brisson, « Intégration de connaissances expertes dans le processus de fouille de données pour l'extraction d'informations pertinentes », Thèse de Doctorat, Université de Nice – Sofia Antipolis- UFR Sciences, France, 2006.

[28] R.Lefébure et G.Venturi. " Data mining Gestion de la relation client Personalisation de sites web ", Editions Eyrolles.

[29] D. Cram, « Techniques d'extraction de connaissances pour la facilitation des taches à base de traces d'interaction », Deuxième partie du livrable T3.1: " États de l'art sur les traces", LIRIS, Février 2008.

[30] A. Vautier, M. O. Cordier, R. Quiniou, "Towards Data Mining Without Information on Knowledge Structure", In Proceedings of the 11th Européen Conférence on Principles and Practice of Knowledge Discovery in Databases (PKDD), Warsaw, Poland, September 17-21 2007, LNCS, Vol. 4702, pp. 300-311, Springer, 2007.

[31] RAMDANE CHAFIKA, « Le clustering des données : une nouvelle approche évolutionnaire quantique », Mémoire présenté en vue de l'obtention du diplôme de magistère, Université Mentouri de Constantine, 2006.

[32] El Moukhtar Zemmouri, « Représentation et gestion des connaissances dans un processus d'Extraction de Connaissances à partir de Données multipoints de vue »Apprentissage [Csl]. Thèse de doctorat, Université Moulay Ismaïl, Ecole Nationale Supérieure d'Arts et Métiers - Meknès, 2013. Français.

[33] BELGACEM, Brahimi,. « Extraction de connaissances à partir de données incomplètes et imprécises », Thèse de doctorat, Université Mohamed Boudiaf de Msila , 2011.

[34] D. A. Zighed et R. Rakotomalala, « Extraction des connaissances à partir des données (ECD) », Techniques de l'ingénieur, HA, 2002.

[35] OPREAN, Cristina, Etat de l'art sur les aspects méthodologiques et processus en Knowledge Discovery in Databases, 2010.

[36] L.Guo, « Applying Data Mining Techniques in Property/Casualty Insurance », In the Forum of the Casualty Actuarial Society, pp. 1 25,2003.

[37] RABAH, M. RAHMANI, « Découverte d'associations sémantiques dans les bases de données relationnelles par des méthodes de Data Mining ». Thèse de doctorat. Université Mouloud Maameri de Tizi Ouzou.

[38] M. Refaat, « Data Préparation for Data Mining Using SAS », Morgan Kaufmann Publishers, Elsevier, 2007.

[39] H. Briand, M. Sebag, R. Gras, F. Guillet, « Mesures de Qualité pour la Fouille de Données », Revue des Nouvelles Technologies de l'Information, RNTI-E-1, Cépaduès-Éditions, 2004.

[40] K. J. Cios, W. Pedrycz, R. W. Swiniarski, L. A. Kurgan, "Data Mining: A Knowledge Discovery Approach",1st edition, Springer, 2007.

[41] Nicolas BÉCHET, Extraction et regroupement de descripteurs morpho-syntaxiques pour des processus de Fouille de Textes, Thèse de doctorat, Univérsité Montpellier II(2009).

[42] Georges El Helu et Charbel Abou khalil , Data Mining , Techniques d'extraction des connaissances.

[43] GAINES B, Transforming rules and trees into comprehensible knowledge structure, Advances in Knowledge Discovery and Data Mining, 1996.

[44] Fayyad U, Piatetsky-Shapiro G., Smyth P, "From Data Mining to Knowledge Discovery in Databases", Advices in Knowledge Discovery and Data Mining, MIT Press, 1: pp 1-36, 1998.

[45] MICHIE D, Methodologies from machine learning in data analysis and software, The Computer Journal, 34, 6, 559-565, 1991.

[46] Wang W, "Data Mining: Concepts, Algorithms, and Applications", COMP pp 290-090, 2003.

[47] M. J. BERRY, G. S. LINOFF, Mastering Data Mining: The Art and Science of Customer Relationship Management, 2000.

[48] D.T. LAROSE, Discovering Knowledge In Data: An Introduction to Data Mining, Central Connecticut State University, 2005.

[49] BELGACEM, Brahimi. Extraction de connaissances à partir de données incomplètes et imprécises, Thèse de doctorat, Université Mohamed Boudiaf de Msila, 2011.

[50] M. J. BERRY, G. S. LINOFF, Data Mining Techniques For Marketing, Sales, and Customer Relationship, Management, Second Edition, 2004.

[51] Séraphin LOHAMBA OMATOKO, Analyse et détection de l'attrition dans une entreprise de Télécommunication.PDF.

[52] Fouille de données d'opinion des usagers de sites E-commerce. UKM Ouargla, Juin 2013.

[53] HOUMADI, Benamar, Étude exploratoire d'outils pour le Data Mining, Thèse de doctorat, Université du Québec à Trois-Rivières, 2007.

[54] B. AGARD, A. KUSIAK, Exploration Des Bases De Données Industrielles à L'aide Du Data Mining – Perspectives, 9ème Colloque National AIP PRIMECA, avril 2005.

[55] Alaoui Abdiya, « Application des techniques des métas heuristiques pour l'optimisation de la tâche de la classification de la fouille de données », mémoire pour l'obtention du diplôme de Magister, UNIVERSITE DES SCIENCES ET DE LA TECHNOLOGIE D'ORAN Mohamed Boudiaf, 2012.

[56] BISSON, Gilles. La similarité: une notion symbolique/numérique. Apprentissage symbolique-numérique, 2000.

[57] HAN, Jiawei, PEI, Jian, et KAMBER, Micheline, « Data mining: concepts and techniques ». Elsevier, 2011.

[58] MOHAMED EL HADI BENELHADJ, **«** Entrepôt de Données et Fouille de Données Un Modèle Binaire et Arborescent dans le processus de Génération des Règles d'Association », Thèse de doctorat en science, 2012.

[59] Laurent Candillier, Contextualisation, visualisation et évaluation en apprentissage non supervisé », thèse de doctorat, université de Charles de Gaulle-Lille3.

[60] Quang, C. T, Classification automatique des textes vietnamiens Hanoi, Institut de la Francophonie pour l'informatique, 2005.

[61] G. CALAS, Études des principaux algorithmes de data mining, Spécialisation Sciences Cognitives et Informatique Avancée, France.

[62] R.GILLERON, M. TOMMASI, Découverte de connaissances à partir de données, 2000.

[63] Leo Breiman, Jerome Friedman, Charles J. Stone, and R. A. Olshen. Classification and Regression Trees. Chapman and Hall/CRC, 1 edition, January 1984.

[64] Ron Kohavi and Ross Quinlan, Decision tree discovery. In Handbook of Data Mining and Knowledge Discovery, pages 267–276. University Press.

[65] J. Ross Quinlan, Induction of decision trees, Machine Learning, 1(1):81–106, 1986.

[66] Denoue, L, Classification supervisée de documents, 2003.

[67] Hai Anh, H, Usage des arbres de décision, Institut de la francophonie pour l'informatique, 2004

[68] Hanoune, M. and F. Benabbou,Modélisation informatique de Clients Douteux, en utilisant les Techniques de DATAMINING. Paris.

[69] Laurent Orseau, Induction d'arbres de décision, Agro Paris Tech.

[70] S. PRABHU, N. VENKATESAN, Data Mining and Warehousing, New Age International (P) Ltd., Publishers, New Delhi, 2007.

[71] B. LAVOIE, Arbres de décisions, Synthèse de lectures, Séminaire sur l'apprentissage automatique, Programme de Doctorat en Informatique Cognitive, Université du Québec à Montréal, 15 mars 2006.

[72] Nicola Beck. ''Application de méthodes de clustering traditionnelles et extension au cadre multicritère''. Mémoire d'ingénieur. Université Libre de Bruxelles, faculté des sciences appliquées, 2006

[73] Anh Tuan, Ifi Hanoi,''Réduction de base de données par la classification automatique''. Rapport de stage. Institut de la francophonie pour l'informatique (IFI).

[74] Jiawei Han, Micheline Kamber,''Data Mining, concepts and techniques''. Ouvrage.

Edition Morgan Kaufmann Publisher. 2000.

[75] Pierre Emmanuel Jouve, ‘‘Apprentissage Non Supervisé et Extraction de Connaissance à partir de Données’’. Thèse de Doctorat en Informatique. Université Lumière, Lyon 2. 2003

[76] Pavel Berkhin, “Survey of clustering data mining techniques”. Article en ligne. Accrue Software Inc. 2002, Consulté le 03.01.2010.

[77] MENOUER Tarek, DERMOUCHE Mohamed, Application de techniques de data mining pour la classification automatique des données et la recherche d’associations, Mémoire de fin d’études pour l’obtention du diplôme d’Ingénieur d’Etat.

[78] Hadj-Tayeb Karima, Approche de partitionnement pour un apprentissage non supervisé des Usagers du Web (Amélioration de l’approche k-means), Université des sciences et de la technologie d’Oran, Mohamed Boudiaf (USTO).

[79] CHAMI Djazia, Une plate forme orientée agent pour le data mining, Mémoire en vue de l’obtention du diplôme de Magister en informatique.

[80] Jiawei Han, Micheline Kamber, Jian Pei ; Data Mining: Concepts and Techniques, Third Edition (2011).

[81] Yannick Toussaint, Fouille de textes et organisation de documents : Extraction de connaissances à partir de textes structurés. LIRIS/INSA de Lyon, 2004.

[82] Najeh NAFFAKHI, Apprentissage supervisé pour la classification des images à l’aide de l’algèbre P-tree, Université de Tunis Institut Supérieur de Gestion de Tunis.

[83] LEFEBURE R , VENTURI G., Data Mining : Gestion de la relation client, Personalisation de sites web, Paris, Editions Eyrolles, Mars 2001.

[84] DOERMANN D, KOBLA V., FALOUTSOS C., Representing and visualizing structure in video sequences, Proceedings of Fifth ACM International Multimedia Conference, pages 335-346 November 2000.

[85] Sandy Moens,Emin Aksehirli and Bart Goethals. Frequent itemset mining for big data.in Big Data ,2013 IEEE International Conference on, page 111-118. IEEE ,2013.

[86] Yen-Hui Liang and Shiow-Yang Wu.Sequence-growth :A scalable and effective frequent itemset mining algorithm for big data based on mapreduce framework. In Big Data(Big Data Congress), IEEE International Congress on ,pages 393-400.IEE ,2015.

[87] Ming-Yen Lin,Pei-Yu Lee and Sue-Chen Hsueh, A priori-based frequent itemset mining algorithms on mapreduce. In Proceedings of the 6th international conference on ubiquitous information management and communication,page 76.ACM,2012.

[88] Suhel Hammoud, MapReduce network enable algorithms for classification based on association rules.PhD thesis,Brunel University School of Engineering and Design PhD Theses,2011.

[89] Zhigang Zhang,Genlin Ji and Mengmeng Tang . Mreclat : An algorithm for parallel mining frequent itemsets. In Advanced Cloud and Big Data(CBD),2013 International Conference on , pages 177-180. IEEE, 2013.

[90] Haoyuan Li,Yi Wang,Dong Zhang,Ming Zhang and Edward Y Chang.Pfp :parallel fp-growth for query recommendation . in Proccedings of the 2008 ACM conference on Recommender systems, pages 107-114. ACM, 2008.

[91] Le Zhou,Zhiyong ,Jin Chang,Junjie Li,Joshua Zhexue Huang and Shengzhon Feng . Balanced parallel fp-growth with mapreduce. In Information Computing and Telecommunications (YC-ICT),2010 IEEE Youth Conference on,pages 243-246. IEEE, 2010.

[92] J. R. Quinlan, Induction of decision trees, 1986.

[93] S. G. Thompson W. Z. Liu, A. P. White and M. A. Bramer. Techniques for dealing with missing values in classiffcation. In Computer Science, editor, Advances in Intelligent Data Analysis, Reasoning about Data, volume 1280, 1997.

[94] R. Pearson, The problem of disguised missing data. ACM SIGKDD Explorations Newsletter, 8(1) :83–92, 2006.

[**95**] B. Goethals T. Calders and M. Mampaey. Mining itemsets in the presence of missing Values. In ACM Symposium on Applied Computing (SAC'07), 2007.

WEBOGRAPHIE

[W1] H. Moed, The Evolution of Big Data as a Research and Scientific Topic: Overview of the Literature, ResearchTrends, http://www.researchtrends.com,2012.

[W2] MIKE 2.0, Big Data Definition, http://mike2.openmethodology.org/wiki/Big_Data_Definition

[W3] Qu'est-ce que le big data ?, http://www.espace-direct.com/qu'est-ce-que-le-big-data, 20 mars 2017.

[W4] Mickael BARON, partie1 – Généralités sur HDFS et Map Reduce, http://mbaron.developpez.com/tutoriels/bigdata/hadoop/introduction-hdfs-map-reduce/, 20 décembre 2014.

[W5] Bastien L, Hadoop – Tout savoir sur la principale plateforme Big Data http:// www.bigdataparis.com. 7 février 2017.

[W6] HDFS : explications, http://phamduc-a.com/tutoriels/Hadoop-et-Big-Data/Installation-et-HDFS/chapitre.html?chapitre=3, 2016.

[W7] François-Xavier Andreu, Déploiement d'une architecture Hadoop pour analyse de flux, françois-xavier.andreu@renater.fr , 2013.

[W8] Xavier Claude, Petit état des lieux du NoSQL, http://linuxfr.org/news/petit-etat-des-lieux-du-nosql, 2012.

[W9] Martin Laloux , Le NoSQL dans le domaine géospatial, approche préliminaire en Python avec SimpleGeo, http://www.portailsig.org/content/le nosql dans le domaine geospatial approche préliminaire en python avec simplegeo, Fevrier 2012.

[W10] Bruno Chaudet, Donnée, information, connaissance, http://www. Donnée, information, connaissance _ Logiques processuelles.htm, 30 mars 2009.

[W11] Data vs. Information, http://www.diffen.com/difference/ Data vs. Information ,2014.

[W12] Bruno Chaudet, donnée-information-connaissance, https://brunochaudet.wordpress.com, Publié le 30 mars 2009.

[W13] morgon.univ-lyon2.fr/Introduction_au_datamining_cours.htm.

[W14] J.Han et M. Kamber, « Data Mining: Concepts and Techniques »,Slides for Textbook, data Preprocessing, http://www.cs.sfu.ca ,Février1, 2006.