République Algérienne Démocratique et Populaire
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 : Mathématiques et Informatique

MEMOIRE DE MASTER
Domaine : Mathématiques et informatique
Filière : Informatique
Option : Systèmes d'informations

Thème :

Un Modèle Basé Data Mining Pour L'exploitation Du Big Data Dans Le Cloud Computing

Présenté par :
SLIMI Hamda
MELLAB Nora

Devant le jury :

Zamar. A	MAA	Université de Tébessa	Président
Derdour. M	MCA	Université de Tébessa	Rapporteur
SOLI. A	MAA	Université de Tébessa	Examinateur

Date de soutenance : 29/05/2016

Note : ……. Mention : ……

ماخص

البيانات الكبيرة ولدت تطورا تكنولوجيا هائلا، وهو التطور الذي يوفر الوصول إلى فرص جديدة. تمثل البيانات الكبيرة بالتأكيد راحة كبيرة، ولكن أيضا تحديات محددة، وهذه التحديات تتمثل في حجم وتنوع البيانات.

في هذه المذكرة أنشأنا دراسة تنطوي على الغالبية العظمى من العمل على موضوع البحث " BIG DATA Mining "، وعرضت هذه المذكرة كيف أثرت البيانات الكبيرة على أساليب استخراج البيانات.

أخيرا عرضنا بنية الخدمة التي تستفيد من موارد السحابة المعلوماتية لتكون قادرة على إجراء عمليات حسابية مكثفة والتي هي ضرورية لاستخراج المعرفة البيانات الكبيرة.

كلمات البحث :البيانات الكبيرة، استخراج البياناتو تعدين البيانات الكبير، مابريديوس، كودا، ذاكرة رسم الخرائط، الحوسبة السحابية.

Résumé

Les volumes de données massifs (« Big Data ») engendrent une évolution considérable des modèles technologiques, une évolution qui permet d'accéder à de nouvelles opportunités.

Le Big Data représente certes une convenance significative, mais pose aussi des défis spécifiques, ces derniers incluent la volumétrie et la variété des données.

Dans ce mémoire on a établi un état de l'art impliquant la majorité des travaux relatifs au sujet de recherche « BIG DATA Mining » et on a montrée comment les big data ont influencé les méthodes de data mining.

Enfin nous avons réalisé une architecture d'un service qui profite des ressource cloud afin d'être en mesure d'effectuer des calculs intensifs qui sont nécessaire pour extraire des connaissances des big data.

Mots clés : Big Data, Data Mining, Big Data Mining, MapReduce, Cuda, Memory Mapping, Cloud Computing.

Abstract

Tremendous amount of data is getting explored through IOT (Internet of Things) from variety of sources such as sensor network, social media feed, internet applications, called as Big Data. Big Data cannot be handled by conventional tools and techniques.

The Big Data mining is essential in order to extract value from massive amounts of data which could give better insights using efficient techniques.

In this manuscript we established a state of art on the subject of big data mining where we talked about the different solutions that enable us to extract knowledge from big data using new approaches such as data mining algorithms based on Map Reduce framework then we compared the different platforms that allow big data mining based on related measures and in the end we developed an architecture for a cloud a service that would help the users in the process of choosing the right platform for their data mining tasks.

Keywords: Big Data, Big Data, Data Mining, Big Data Mining, Map Reduce, Cuda, Memory Mapping, Cloud Computing.

Remerciements

A Notre Encadreur Derdour Makhlouf
Nous avons eu le privilège de travailler parmi votre équipe et d'apprécier vos qualités et vos valeurs.
Votre sérieux, votre compétence et votre sens du devoir nous ont énormément marqués.
Veuillez trouver ici l'expression de notre respectueuse considération et notre profonde admiration pour toutes vos qualités scientifiques et humaines.
Ce travail est pour nous l'occasion de vous témoigner notre profonde gratitude.

Dédicace

A notre cher et dynamique encadreur
Derdour Makhlouf
Un remerciement particulier et sincère pour tous vos efforts
fournis. Vous avez toujours été présent.
Que ce travail soit un témoignage de ma gratitude et mon profond
respect.
A mon chère Père,
Que Dieu Bénisse son âme
A ma mère et ma petite sœur
En témoignage de l'attachement, de l'amour et de
L'affection que je porte pour vous.
Je vous remercie pour
votre affection si sincère.
A mon très cher oncle Madjed Maalem
et sa famille
Vous avez toujours été présents pour les bons conseils.
Votre affection et votre soutien m'ont été d'un grand secours au long de ma vie
professionnelle et personnelle.
Veuillez trouver dans ce modeste travail ma reconnaissance pour tous vos efforts.
A mes chères amies
Je ne peux trouver les mots justes et sincères pour vous
exprimer mon affection et mes pensées, vous êtes pour moi des
frères, des amis sur qui je peux compter.
En témoignage de l'amitié qui nous uni et des souvenirs de
tous les moments que nous avons passé ensemble, je vous dédie
ce travail et je vous souhaite une vie pleine de santé et de
bonheur.

SLIMI HAMDA

Dédicace

A notre cher et dynamique encadreur
Derdour Makhlouf
Un remerciement particulier et sincère pour tous vos efforts
Fournis. Vous avez toujours été présent.
Que ce travail soit un témoignage de ma gratitude
et mon profondrespect.
A Mes très Chers parents
A Mon marie
A Toute MAA FAMILLE.
A mes chères amies

MELLAB NORA

LISTE DES TABLEAUX

TABLE DES FIGURES

TABLE DES MATIÈRES

INTRODUCTION GÉNÉRALE

Ces dernières années ont vu une augmentation spectaculaire de la capacité à recueillir des données provenant d'une variété de capteurs, appareils, et en différents formats, à partir d'applications indépendantes ou connectées. Le déluge des données a dépassé largement la veille technologique sur plusieurs plans, à savoir, le traitement, l'analyse, le stockage et surtout leurs compréhensions. Considérez les données sur Internet. Les pages Web indexées par Google étaient autour d'un millions en 1998, mais ils ont rapidement atteint un milliard en 2000 et ils ont déjà dépassé un trillion en 2008. Cette expansion rapide est accélérée par l'augmentation surprenante de l'utilisation des réseaux sociaux, tels que Facebook, Twitter etc., qui permettent aux utilisateurs l'ajout inconditionnel des informations (textes, images, vidéos ?etc.) sur le Web, ce qui amplifie cette montée volumineuse. En outre, l'amélioration de la vie quotidienne est désormais possible via l'analyse des données récoltées en temps réel à partir des téléphones mobiles. De ces faits nous pouvons déduire que le phénomène Big Data change radicalement les modalités de gestion des données puisqu'il introduit de nouvelles problématiques concernant la volumétrie, la vitesse de transfert et le type de données. Alors comment surpasser les défis liés à l'extraction de connaissances à partir des Big data et comment adapter les anciennes techniques de data mining à ces nouvelles épreuves ? Dans ce mémoire nous avons tenté de répondre à cette question ; ce

dernier s'articule sur trois chapitres, dans le premier nous présentons les concepts de base relatifs à notre thème, dans le deuxième nous exposons les articles qui partagent la même problématique (Etat de l'art) et pour le dernier nous proposons une solution cloud pour le big data mining.

CHAPITRE I

BIG DATA & DATA MINING

1 Introduction

Le Big Data est un phénomène qui a vu le jour avec l'émergence de données volumineuses qu'on ne pouvait pas traiter avec des techniques traditionnelles. Les premiers projets de Big Data sont ceux des acteurs de la recherche d'information sur le web (moteurs de recherche) tel que Google et Yahoo. En effet, ces acteurs étaient confrontés aux problèmes de la scalabilité (passage à l'échelle) des systèmes et du temps de réponse aux requêtes utilisateurs.

Très rapidement, d'autres sociétés ont suivis le même chemin comme Amazon et Facebook. Le Big Data est devenu une tendance incontournable pour beaucoup d'acteurs industriels du fait de l'apport qu'il offre en qualité de stockage, traitement et d'analyse de données.

Big Data

2 Définition de Big Data

Comme l'expression l'indique, le Big Data se caractérise par la taille ou la volumétrie des informations. Mais d'autres attributs, notamment la vitesse et le type de données, sont aussi á considérer. En ce qui concerne le type, le Big Data est souvent rattaché á du contenu non structuré ou semi-structurée, ce qui peut représenter un défi pour les environnements classiques de stockage relationnel et de calcul. Les données non structurées et semi-structrées sont partout : contenu web, posts twitter ou commentaires client en format libre. Par vitesse on entend la rapidité avec laquelle les informations sont créées. Grâce á ces nouvelles technologies, il est maintenant possible d'analyser et d'utiliser l'importante masse de données fournie par les fichiers log des sites web, l'analyse d'opinions des réseaux sociaux, et même les vidéos en streaming et les capteurs environnementaux. Nous pouvons ainsi tirer parti d'une vision stratégique impossible jusqu'á ce jour .

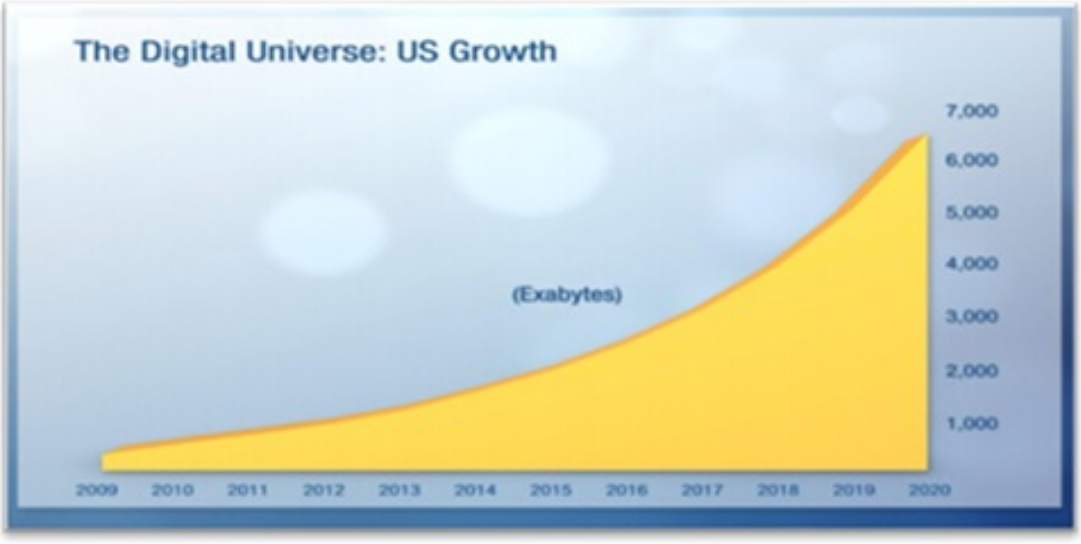

Figure I.1 – Accroissement du stockage de données entre 2009 et 2020

3 Historique des Big Data

C'est en 2008 que Gartner a utilisé pour la première fois l'expression (Big Data), en référence à l'explosion des données numériques. Certains ont parlé de déluge de données.

Quelle que soit la métaphore, le phénomène est réel. Son impact sur notre façon de travailler est présent et pressant. Le Big Data n'est pas un phénomène nouveau. Cette discipline a vu le jour avec l'émergence de données trop volumineuses.

4 L'origine des données du Big data

Les données traitées par le Big Data proviennent notamment :

Du Web : journaux d'accès, réseaux sociaux, e-commerce, indexation, stockage de documents, de photos, de vidéos, etc.

plus généralement, de l'internet et des objets communicants : réseaux de capteurs, journaux des appels en téléphonie ;

des sciences : génomique, astronomie, physique subatomique, climatologie (ex : le centre de recherche allemand sur le climat gère une base de données de 60 petaoctets), etc.

données commerciales (ex : historique des transactions dans une chaîne d'hypermarchés) ; données personnelles (ex : dossiers médicaux) ; Données publiques (open data) [1].

5 Contexte du Big Data

On peut parler de Big Data dés lors que :

- Les volumes à traiter atteignent des tailles (plus grandes) que les problèmes courants : Peta (web), Terra, Exa, Zettaoctets.

- Le problème ne peut pas être traité par les outils existants : SGBD relationnels.

6 Les caractéristiques du Big Data

Le Big Data (données massives) se caractérise par la problématique des 3V qui sont le Volume, la Variété et la Vélocité, certains auteurs ont rajoutés d'autres V comme la Valeur.

- ***Volume*** : désigne la masse de données collectées (giga-octets, téraoctets,...),

- ***Variété*** :Variété de Big Data révèle l'hétérogénéité des données par rapport à son type (structurée , semi- structurée, et non structurées), la représentation et l'interprétation sémantique .

- ***Vélocité*** : représente la génération et le traitement des données transitoires en vol dans le délai écoulé . La plupart des sources de données génèrent des données de haut flux continu qui se déplacent à une vitesse très élevée , ce qui rend les analyses plus complexe.

- ***Variabilité*** : le format et le sens des données peut varier au fil du temps.

- ***Valeur :*** Les entreprises doivent être en mesure de tirer parti de la valeur des informations stockées en masse grâce à une approche analytique des Big Data, c'est ce que l'on appelle les Big Analytics [2].

7 Domaines d'applicaion des Big Data

Les usages du Big Data sont très vastes qui touchent presque tous les secteurs d'activités, quelques exemples sont cités ci-dessous :

7.1 Transports

Exemple : Planification des voyages : mise à disposition du citoyen de données jusque là réservées aux administrations (gagner du temps / réduire le coût).

7.2 Santé

Exemple : Suivi des patients (dossier médical du patient).

7.3 Economie

Exemple : Accélération des temps d'analyse des données clients pour l'identification des comportements atypiques.

8 Les limites des SGBD Relationnels

Les organisations tels que Google, Facebook possèdent une très grande quantité des données c'est-à-dire plusieurs centaines de millions d'entrées dans leurs bases de données Par conséquent, une seule machine ne peut pas gérer la base de données, ces bases de données sont dupliquées pour que le service ne soit pas interrompu en cas de panne. La méthode consiste donc à rajouter des serveurs pour dupliquer les données et ainsi augmenter les performances et résister aux pannes. Seulement, dû aux propriétés fondamentales sur lesquelles une base de données relationnelle repose.

9 Le stockage des données du Big Data (Modèle NoSQL)

un certain nombre de limitations importantes sont apparues au fil des années. Les premiers acteurs à buter sur ces limites furent les fournisseurs de services en ligne, les plus populaires étant Yahoo, Google ou plus récemment les acteurs du web social comme Facebook, Twitter... Ces besoins de stockage et de traitement des quantités importantes de données ont poussé ces organisations à chercher de nouvelles techniques de stockage.

Les besoins majeurs identifiés par ces acteurs sont les suivants :

- Capacité à distribuer les traitements sur un nombre important de machines afin d'être en mesure d'absorber des charges très importantes. On parle de scaling des traitements.

- Capacité à répartir les données entre un nombre important de machines afin d'être en mesure de stocker de très grands volumes de données. On parlera plutôt de scaling des données.

- La distribution de données sur plusieurs Datacenter afin d'assurer une continuité de service en cas d'indisponibilité de service sur un Datacenter. Cette approche doit par ailleurs permettre une facile et rapide accessibilité aux données que des personnes éprouvent le besoin de les utiliser.

- Une architecture qui fonctionne sur du matériel peu spécialisé et donc facilement

remplaçable en cas de panne.

En effet, la plupart des SGBD (Système de Gestion de Bases de Données) relationnels ne permettent plus de répondre aux besoins de stockage et de traitement de ces grandes quantités. Ils existent de nouveaux SGBD qui sont regroupés sous l'appellation : les SGBD NoSQL (Not Only SQL). Ils représentent l'ensemble des technologies qui se distinguent par un relâchement des caractéristiques ACID (Atomicité, Consistance, Intégrité et Disponibilité) propres aux SGBDR (Système de Gestion de Bases de Données Relationnelles).

9.1 Une nouvelle approche de stockage et de manipulation de données"NoSQL"

9.1.1 Définition

- Comme son nom l'indique, ce n'est pas un remplaçant des SGBDR, mais il vient en complément de ces derniers. Le NoSQL est un type de magasin de données, c'est une manière de stocker et de récupérer des données de façon rapide, un peu comme une base de données relationnelle, sauf qu'il n'est pas basé sur une relation mathématique entre les tables comme une base de données relationnelle traditionnelle.

- Le NoSQL regroupe de nombreuses bases de données, récentes qui se différencient du modèle SQL par une logique de représentation de données non relationnelle, leurs principaux avantages sont leurs performances et leur capacité à traiter de très grands volumes de données.

La famille des SGBD NoSQL se compose de plusieurs catégories : orientés colonne, orientés document, orientés graphe et Clé/valeur.

- *Modèle clé / valeur :*

Dans ce modèle les données sont donc simplement représentées par un couple clé/valeur. La valeur peut être une simple chaîne de caractères, un objet sérialisé... Cette absence de

structure ou de typage ont un impact important sur le requêtage. La communication avec la BD se résumera aux opérations PUT, GET et DELETE.

- *Bases documentaires :*

Ce modèle se base sur le paradigme clé valeur. La valeur, dans ce cas, est un document de type JSON ou XML. L'avantage est de pouvoir récupérer, via une seule clé, un ensemble d'informations structurées de manière hiérarchique. La même opération dans le monde relationnel impliquerait plusieurs jointures.

- *Bases orientées colonnes :*

Ce modèle ressemble à première vue 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 dés la création du schéma de la table et ce nombre reste le même pour tous les enregistrements dans cette table. Par contre, 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. Comme solutions, on retrouve principalement HBase (implémentation Open Source du modèle BigTable publié par Google) ainsi que Cassandra (projet Apache qui respecte l'architecture distribuée de Dynamo d'Amazon et le modèle BigTable de Google).

- *Paradigme graphe :*

Le modèle de représentation des données se base sur la théorie des graphes. Il s'appuie sur la notion de noeuds, de relations et de propriétés qui leur sont rattachées. Ce modèle facilite la représentation du monde réel, ce qui le rend adapté au traitement des données des réseaux sociaux.

10 Les plateformes pour le Big Data

La mise en oeuvre d'un projet Big Data nécessite le choix d'une méthode de stockage,d'une technologie d'exploitation et des outils d'analyse de données. Pour optimiser

les temps de traitement sur des bases de données volumineuses, une panoplie de solutions existe, certains sont en open-source et d'autres sont propriétaires. Nous allons décrire certaines de ces solutions :

10.1 Hadoop

Hadoop est un projet open source de la fondation Apache qui est constitué de plusieurs composants(HDFS,MapReduce, Hive,...). Hadoop met à la disposition des développeurs et des administrateurs un certain nombre de briques essentielles :

- HDFS (Hadoop Distributed File System) système de fichiers distribués sur un ensemble de noeud (serveurs). C'est un système tolérant aux fautes (malgré les pannes il continue à fonctionner correctement).

- MapReduce, une technologie qui permet la parallélisation des traitements et qui s'effectue en deux phases, la phase Map consiste en la division des traitements en sous-ensembles et exécution en parallèle, la phase Reduce se charge de collecter les réponses des noeuds et les assemblent pour constitué le résultat.

- Hive, fournit un langage de haut niveau semblable à SQL, appelé HQL, pour interagir avec un cluster hadoop, dans le but réaliser des analyses sur une masse importante de données.

- Hbase, une base de données distribuée disposant d'un stockage structuré pour les grandes tables. Hbase est une base de données orientée colonnes, elle fait partie des BD NoSQL (Not only SQL).

- Pig, un système de traitement de gros volumes de données en utilisant la plateforme Hadoop MapReduce, Il fournit les opérations de filtrage, jointure et classement des données (conçu spécialement pour l'analyse de données)[2].

Data Mininig

11 Data mining

L'exploration de données, connue aussi sous l'expression de fouille de données, forage de données, prospection de données, data mining, ou encore extraction de connaissances à partir de données, a pour objet l'extraction d'un savoir ou d'une connaissance à partir de grandes quantités de données, par des méthodes automatiques ou semi-automatiques. Elle propose d'utiliser un ensemble d'algorithmes issus de disciplines scientifiques diverses telles que les statistiques, l'intelligence artificielle ou l'informatique, pour construire des modèles à partir des données, c'est-à-dire trouver des structures intéressantes ou des motifs selon des critères fixés au préalable, et d'en extraire un maximum de connaissances. L'utilisation industrielle ou opérationnelle de ce savoir dans le monde professionnel permet de résoudre des problèmes très divers, allant de la gestion de la relation client à la maintenance préventive, en passant par la détection de fraudes ou encore l'optimisation de sites web [3].

12 Les tâches du data mining

12.1 Règles d'association

Vise à divulguer les modèles (Pattern) qui apparaissent fréquemment dans l'ensemble des données.

Par exemple, une règle découverte dans les données de ventes dans un supermarché pourrait indiquer qu'un client achetant des oignons et des pommes de terre simultanément, serait susceptible d'acheter un hamburger. Une telle information peut être utilisée comme base pour prendre des décisions marketing telles que par exemple des promotions ou des emplacements bien choisis pour les produits associés. Les règles d'association sont employées aujourd'hui dans plusieurs domaines incluant celui de la fouille du web, de la détection d'intrusion et de la bio-informatique.

Une règle d'association peut être définie formellement de la façon suivante : Soit I = i1,

i2, ..., in un ensemble d'items. Soit T= t1, t2, ..., tm un ensemble de transactions telle que ti soit un ensemble de I (i.e. ti ' I). Une règle d'association s'exprime sous la forme [4] :

$$X \Rightarrow YouX, Y \subseteq IandX \cap Y = \%_0$$

12.1.1 Méthodes

12.1.1.1 Apriori

L'algorithme Apriori est un algorithme d'exploration de données conçu en 1994, par Rakesh Agrawal et Ramakrishnan Sikrant, dans le domaine de l'apprentissage des règles d'association. Il sert à reconnaitre des propriétés qui reviennent fréquemment dans un ensemble de données et d'en déduire une catégorisation[5].

12.1.1.2 FP-growth

FP-growth (fréquent pattern growth) utilise une structure d'arbre (FP-tree) pour stocker une forme compressée d'une base de données. FP-growth adopte une stratégie de découpage pour décomposer les taches d'exploration de données et les bases de données. Il utilise une méthode " pattern fragment growth " pour éviter le coûteux processus de génération et de test des candidats, utilisé par Apriori [4].

12.1.1.3 GUHA

GUHA (" General Unary Hypotheses Automaton ") est une méthode de génération automatique d'hypothèses à partir de données empiriques, c'est donc une méthode d'exploration de données. La procédure ASSOC est une méthode GUHA qui explore les données en vue de trouver rapidement des règles d'association généralisées en utilisant des structures de données en tableau (" Bit array ") [6].

12.1.1.4 OPUS

OPUS est un algorithme efficace pour la recherche de règles d'association, qui, par

opposition à d'autres, ne nécessite pas de contraintes anti-monotones et monotones tels que le support minimum. Initialement utilisé pour trouver des règles pour une conclusion donnée, il a par la suite été étendu pour trouver des règles avec n'importe quel item comme conclusion. Le moteur de recherche OPUS est la technologie centrale dans le populaire système de recherche d'association Magnum Opus[6].

12.2 Classification

La classification est un processus à deux étapes : une étape d'apprentissage (entrainement) et une étape de classification (utilisation).

Dans l'étape d'apprentissage, un classifier (une fonction, un ensemble de règles, ...) est construit en analysant (ou en apprenant de) une base de données d'exemples d'entrainement avec leurs classes respectives. Un exemple X = (x1, x2, ., xm) est représenté par un vecteur d'attributs de dimension m. Chaque exemple est supposé appartenir à une classe prédéfinie représentée dans un attribut particulier de la base de donnée appelé attribut de classe. Puisque la classe de chaque exemple est donnée, cette étape est aussi connue par l'apprentissage supervisé.

Dans l'étape de classification, le modèle construit dans la première étape est utilisé pour classer les nouvelles données. Mais avant de passer à l'utilisation, le modèle doit être testé pour s'assurer de sa capacité de généralisation sur les données non utilisées dans la phase d'entrainement. Le modèle obtenu peut être testé sur les données d'entrainement elles-mêmes, la précision (le taux de reconnaissance) est généralement élevée mais ne garantit pas automatiquement une bonne précision sur les nouvelles données. En effet, les données d'entrainement peuvent contenir des données bruitées ou erronées (outliers) qui ne représentent pas le cas général et qui tire le modèle vers leurs caractéristiques. Ce cas est appelé le sur-apprentissage ou en anglais "over fitting" et qui peut être évité en testant le modèle sur une base de données différente appelée base de test. La base de test est un Ensemble d'exemples ayant les mêmes caractéristiques que ceux de la base d'entrainement

et qui sont écartés au départ de l'entrainement pour effectuer les tests.

12.2.1 Méthode de classification

12.2.1.1 L'apprentissage par arbre de décision

Désigne une méthode basée sur l'utilisation d'un arbre de décision comme modèle prédictif permettant d'évaluer la valeur d'une caractéristique d'un système depuis l'observation d'autres caractéristiques du même système. On l'utilise notamment en fouille de données et en apprentissage automatique. Dans ces structures d'arbre, les feuilles représentent les valeurs de la variable-cible et les embranchements correspondent à des combinaisons de variables d'entrée qui mènent à ces valeurs. En analyse de décision, un arbre de décision peut être utilisé pour représenter de manière explicite les décisions réalisées et les processus qui les amènent (voir article Arbre de décision). En apprentissage et en fouille de données, un arbre de décision décrit les données mais pas les décisions elles-mêmes, l'arbre serait utilisé comme point de départ au processus de décision.

C'est une technique d'apprentissage supervisé : on utilise un ensemble de données pour lesquelles on connaît la valeur de la variable-cible afin de construire l'arbre (données dites étiquetées), puis on extrapole les résultats à l'ensemble des données de test.[7].

12.2.1.2 La prédiction

La prédiction ressemble à la classification et à l'estimation mais dans une échelle temporelle différente. Tout comme les tâches précédentes, elle s'appuie sur le passé et le présent mais son résultat se situe dans un futur généralement précisé. La seule méthode pour mesurer la qualité de la prédiction est d'attendre

Les techniques les plus appropriées à la prédiction sont Les arbres de décision les réseaux de neurones.

12.2.1.3 K plus proches voisins

En intelligence artificielle, la méthode des k plus proches voisins est une méthode d'apprentissage supervisé. En abrégé k-NN ou KNN, de l'anglais k-nearest neighbor.

Dans ce cadre, on dispose d'une base de données d'apprentissage constituée de N couples - entrée-sortie -. Pour estimer la sortie associée à une nouvelle entrée x, la méthode des k plus proches voisins consiste à prendre en compte (de façon identique) les k échantillons d'apprentissage dont l'entrée est la plus proche de la nouvelle entrée x, selon une distance à définir.

Par exemple, dans un problème de classification, on retiendra la classe la plus représentée parmi les k sorties associées aux k entrées les plus proches de la nouvelle entrée [8].

12.2.1.4 Clustering

Clustering est le processus de regroupement des données dans des classes ou des clusters, de sorte que les objets au sein d'un cluster ont une grande similitude l'un par rapport à l'autre, mais très dissemblables à des objets dans un autre cluster. [1]

Les approches de clustering sont :

— Partitionnement : partitionne les objets et évalue les partitions

— Algorithme des K-moyennes et variantes

— Clustering flou

— spectral

— Hiérarchique : décomposition hiérarchique d'ensembles d'objets

— Clustering hiérarchique ascendant (CHA) et variantes

— Clustering hiérarchique descendant

— Densité : basée sur une fonction de densité ou de connectivité

— Grille : basée sur une structure de granularité à plusieurs niveaux [référence cours]

13 D'ou pourrions-nous extraire les connaissances ?

Il existe une variété immense de dépôt de données, et généralement le DM est applicable à tous types de dépôt de données mais nous n'aborderons que quelque uns. En premier, le DM peut être appliqué sur les bases de données relationnelles afin d'en extraire des connaissances que le SQL est incapable de déterminer, tel que l'analyse des données des clients et la prédiction du risque de crédit d'un nouveau client en fonction de son âge, revenu et les informations sur son crédit précèdent. Les entrepôts de donnée représente eux aussi une source fréquente de connaissance malgré que ces derniers ont leur propre outils d'analyse de données comme OLAP, des outils de DM sont nécessaires pour permettre une analyse plus profonde et automatisée. Le web peut lui aussi être fouillée (Web mining) mais sa fouille est un vrai défi due à l'hétérogénéité des données résidents dans le web. Autre dépôts de données :

— BD transactionnelles

— BD objet et objet-relationnelles

— BD spatiales

— Séries temporelles

— BD Textes et multimédia

— BD Hétérogènes

14 Les applications du data mining

Le datamining est une spécialité transverse : elle regroupe un ensemble de théories et d'algorithmes ouverts à tout domaine métier susceptible de drainer une masse de données. La liste suivante illustre des applications courantes du datamining, mais elle reste loin de l'exhaustivité : Industrie

— Optimisation/fiabilisation d'une chaîne de montage

— Système expert de résolution de panne par la description des symptômes

— Prévision de pics de consommation d'un réseau (téléphone, énergie électrique...)

— Traitement d'images

— Reconnaissance de forme

— Reconnaissance de signatures biométriques Outils de collaboration

— Classification dynamique et contextuelle de documents non structurés

— Mise en relation de personnes par la création automatique de profil de centres d'intérêt.

15 Outils de data mining

Ces outils diffèrent l'un de l'autre selon l'approche adopte et la taille des données qu'ils peuvent traiter par exemple, Microsoft Analysis Service peut traiter jusqu'à 1 millions de lignes alors que Start Miner ne peut traiter qu'une dizaine de millier de lignes.

Produit	Spécialité	Editeur
StartMiner	Réseaux de neurones-arbre de décision	Grimmersoft
Alice	Arbre de décision	Isoft
KnowledgeSEEKER	Arbre de décision	Angoss
4Thoughts	Réseaux de neurones	Cognos
Intelligent Miner	Classification Relationnelles-Réseaux de neurones	IBM
Micrososft Analysis Service	Arbre de décision – clustering	Micrososft
KXEN	Théorie de l'apprentissage de Vapnik	KXEN

Tableau I.1 – Outils de Data Minning

16 Difficultés liées au data mining

16.1 La qualité des données

C'est-à-dire la pertinence et la complétude des données, est une nécessité pour l'exploration des données, mais ne suffit pas. Les erreurs de saisies, les enregistrements doublonnés,

les données non renseignées ou renseignées sans référence au temps affectent aussi la qualité des données. Les entreprises mettent en place des structures et des démarches d'assurance qualité des données pour pouvoir répondre efficacement aux nouvelles réglementations externes, aux audits internes, et augmenter la rentabilité de leurs données qu'elles considèrent comme faisant partie de leur patrimoine.

16.2 L'interopérabilité

D'un système est sa capacité à fonctionner avec d'autres systèmes, créés par des éditeurs différents. Les systèmes d'exploration de données doivent pouvoir travailler avec des données venant de plusieurs systèmes de gestion de bases de données, de type de fichier, de type de données et de capteurs différents. En outre, l'interopérabilité a besoin de la qualité des données. Malgré les efforts de l'industrie en matière d'interopérabilité, il semble que dans certains domaines ce ne soit pas la règle. Les données sont collectées dans le but de répondre à une question posée par le métier. Un risque de l'exploration de données est l'utilisation de ces données dans un autre but que celui assigné au départ. Le détournement des données est l'équivalent d'une citation.

16.3 La vie privée

Des personnes peut être menacée par des projets d'exploration de données, si aucune précaution n'est prise, notamment dans la fouille du web et l'utilisation des données personnelles collectées sur Internet où les habitudes d'achats, les préférences, et même la santé des personnes peuvent être dévoilées. Un autre exemple est fourni par l'Information Awareness Office et en particulier le programme Total Information Awareness (TIA) qui exploitait pleinement la technologie d'exploration de données et qui fut un des projets -post-11 septembre- que le Congrès des États-Unis avait commencé à financer, puis qu'il a abandonné à cause des menaces particulièrement importantes que ce programme faisait peser sur la vie privée des citoyens américains. Mais même sans être dévoilées, les données

des personnes recueillies par les entreprises, via les outils de CRM, les caisses enregistreuses, les DAB, les cartes santé, etc., peuvent conduire, avec les techniques de fouille de données, à classer les personnes en une hiérarchie de groupes, de bons à mauvais, prospects, clients, patients, ou n'importe quel rôle que l'on joue à un instant donné dans la vie sociale, selon des critères inconnus des personnes elles-mêmes. Dans cette optique, et pour corriger cet aspect négatif, Rakesh Agrawal et Ramakrishnan Sikrant s'interrogent sur la faisabilité d'une exploration de données qui préserverait la vie privée des personnes. Le stockage des données nécessaire à la fouille pose un autre problème dans la mesure où les données numériques peuvent être piratées. Et dans ce cas l'éclatement des données sur des bases de données distribuées87 et la cryptographie font partie des réponses techniques qui existent et qui peuvent être mises en place par les entreprises.

17 Conclusion

Dans ce chapitre nous avons définie tous les concepts liés aux deux domaines big data et data mining, dans le prochain chapitre nous allons faire le lien entre ces deux domaines et montrer les difficultés liées au big data mining.

CHAPITRE II

BIG DATA MINING

1 Introuduction

Le big data mining pose un vrai défis á travers différents aspects et afin d'avoir une idée générale sur ces difficultés, nous avons tenté de les définir dans une problématique bien incitée, après nous avons établi un état de l'art contenant les travaux réalisés par les diffèrents piliers du domaine pour aboutir á une résolution á cette obstacle, nous avons repartie ces travaux selon le type de solution en 3 catégories majeures, nous avons aussi établie une taxonomie des travaux liés á une de ces catégories.

2 Défis des Big Data Mining

Le monde informatique évolue á un rythme vertigineux, entrainant une révolution dans l'organisation des données. Ce qui a pour effet de laisser, de façon sans cesse croissante, des traces numériques (ou des données) á chacune de nos opérations et interventions.

Ces traces peuvent être minée et analysée á chaque fois que nous jugeons utile. Telle est l'idée de base derrière l'expression (Big-Data).

Mais le phénomène Big Data change radicalement les modalités de gestion des données puisqu'il introduit de nouvelles problématiques concernant la volumétrie, la vitesse de transfert, le type, et le traitement de ces données. Dans notre thèse nous nous focaliserons sur l'aspect traitement et analyse des big data autrement dit le BIG DATA Minig.

Big Data ne peut pas être manipulé par des outils et des techniques classiques et du fait que les réseaux sociaux sont de plus en plus dominant dans les communications sur Internet. Big Data Mining est essentiel afin d'en extraire des connaissances qui permettront de prédire les comportements et les futures tendances, permettant aux entreprises de prendre des décisions proactives, axées sur le savoir. Alors comment pourrions-nous adapter/proposer des méthodes classiques de data minig pour être en mésure d'extraire des connaissances á partir des big data ?

3 Travaux de recherche dans les Big Data Mining

3.1 Définitions des concepts

3.1.1 Map-Reduce (MR)

Est un framework de traitement distribué sur de gros volumes de données et un modèle de programmation parallèle conçu pour la scalabilité et la tolérance aux pannes, son principe est de répartir la charge sur un grand nombre de serveurs et gère la distribution de données.

Map-Réduce permet de :

✓Traiter de grands volumes de données.

✓Gérer de milliers de processeurs.

✓Paralléliser et distribuer des traitements.

✓Ordonnancement des entrées / sorties.

✓Gérer la tolérance aux pannes.

✓Surveiller des processus.

Principes de base de l'algorithme

Une tâche est divisée en deux ou plusieurs sous-tâche, Chaque sous-tâche est traitée indépendamment, puis leurs résultats sont combinés, 3 opérations majeures : Split, Compute et Join C'est le principe de (Diviser pour Reigner) avec une structure pseudo-hiérarchique.

3.1.2 CUDA

CUDA est une architecture de traitement parallèle développée par NVIDIA permettant de décupler les performances de calcul du système en exploitant la puissance des processeurs graphiques (GPU). Alors que des millions de GPU compatibles avec CUDA ont été vendus, des milliers de développeurs de logiciels, de scientifiques et de chercheurs utilisent CUDA dans une grande gamme de domaines, incluant notamment le traitement des images et des vidéos, la chimie et la biologie par modélisation numérique, la mécanique des fluides numérique, la reconstruction tomodensitométrique, l'analyse sismique, le ray tracing et

bien plus encore.

Traitement parallèle avec CUDA

Le calcul informatique a évolué en passant du traitement central exclusif des CPU vers les capacités de co-traitement offertes par l'association du CPU et du GPU. Pour permettre ce nouveau paradigme informatique, NVIDIA a conçu l'architecture de traitement parallèle CUDA, aujourd'hui incluse dans les GPU GeForce, ION Quadro, et Tesla en offrant ainsi une base matérielle significative aux développeurs d'applications. Du côté de la recherche scientifique, CUDA a été reçu avec enthousiasme. CUDA permet par exemple d'accélérer AMBER, un programme de simulation de dynamique moléculaire utilisé par plus de 60 000 chercheurs du public et du privé afin d'accélérer la découverte de nouveaux médicaments pour l'industrie pharmaceutique.

En matière de finance, Numerix et CompatibL ont annoncé leur support de CUDA pour une nouvelle application de recherche de risques de contrepartie, accélérant jusqu'à 18x les procédures de calcul existantes. La plateforme Numerix est utilisée par plus de 400 institutions financières.

Le parc existant de GPU Tesla, offrant d'importantes capacités en matière de calcul par le GPU, permet également de jauger le succès de CUDA. Plus de 700 clusters de GPU sont aujourd'hui actifs dans le monde entier. De nombreux groupes, allant de Schlumberger et Chevron (secteur énergétique) à BNP Paribas (secteur banquier) et figurant dans la liste des (500 entreprises les plus importantes au monde) publiée par Fortune, ont adopté CUDA.

Depuis l'avènement de Windows 7 de Microsoft et de Snow Leopard d'Apple, le calcul par le GPU est également une réalité pour le grand public. Dans ces nouveaux systèmes d'exploitation, le GPU ne tiendra pas seulement lieu de processeur graphique, il jouera également un rôle de processeur parallèle pour toutes les applications.

3.2 Travaux basés Map-Reduce

3.2.1 Algorithmes basés Map-Reduce

3.2.1.1 MRPrePost : MRPrePost-A parallel algorithm adapted for mining big data

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 Frequent ItemSet sont utilis és pour d ériver des règles d'association et acc éder aux connaissances pertinentes, mais avec l'arriv ée de l'ère de BIG DATA, les algorithmes traditionnelle de data mining ont ét é incapables de répondre aux besoins de l'analyse des BIG DATA.

Le papier [9] propose un algorithme parallèle 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 rapportà PrePost et le PFP, et la stabilité et l' évolutivité de l'algorithme MRPrePost est meilleur.

Le papier propose un algorithme parallèle basé sur MapReduce appel é MRPrePost sur la base de PrePost et d écrit en d étail la mise en œuvre de l'algorithme.

Face à l'exploitation minière de grands ensembles de donn ées, la parall élisation est une bonne solution, les r ésultats exp érimentaux

3.2.1.2 FP-Growth :Squence-Growth :A scalable and effective frequent itemset mining algorithm for big data based on mapreduce framework In big data(BigData Congress)

Frequent itemset mining (FIM) 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 de l'efficacité et le passage á l'échelle devient plus austère.

Dans cet article [10], 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 a l'échelle de cet algorithme.

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

3.2.2 Discussion des travaux basés MapReduce

On repésente dans le tableau (II-1) les algorithmes basés MapReduce :

Auteur	**Technique**	**Compatible MapReduce**	**Caractéristiques**	**Avantages**	**Limitations**
Agrawal et al [11]	Apriori	Non	Level Wise Search monotonicity Facile a implementer	Génère des règles d'association	évolutivité
Zaki et al [12]	Eclat	Non	Recherche en profondeur. Applicable sur les bases de donnée verticale	Améliore la localisation et requière seulement quelques scans	Dégradation de la performance avec de large nombre de transaction
Han et al[13]	FP-Growth	Non	Approche récursive. Emploie l'arbre FP comme structure de données	Recherche focalisée sur les bases de données de petite taille	Performance banale
Hammoud [14]	MRApriori	Oui	Un seul scan des données. Structures hybrides vertical et horizontal	Efficace avec Big Data. Performant	Une réduction insignifiante du temps de traitement
Li et al [15]	Parallel FP Growth	Oui	Version parallèle de FP Growth Independence entre la création de l'arbre FP et le regroupement des éléments	Evolutivité linéaire	Non efficace en matière de mémoire et vitesse
Moens et al [16]	BigFIM	Oui	Méthode hybride (Apriori+Ecalt)	évolutivité	Vitesse

Tableau II.1 – Analyse comparatives des différentes téchniques

Lin et al.[17] ont proposée Single Pass Counting(SPC), Fixed Passes Combined - Counting (FPC) et Dynamic Passes Counting (DPC) qui établit le comptage des étapes en parallèle par la distribution des dataset a travers les différents mappers .Hammoud a proposé MRApriori [14] un approche pour trouver des motifs fréquents par commutation entre la disposition verticale et horizontale itérative qui élimine le besoin de numérisation itérative des données. Il répète l'analyse d'autres données intermédiaires et elle est réduite avec chaque itération.

MREclat [18]est un algorithme Éclat modifié dans le cadre de MapReduce qui génère une liste de motifs fréquents, la liste est divisée en classes d'équivalence, puis pour chaque classe d'équivalence Frequent Itemset sont calculés en utilisant le cadre de MapReduce.

Parallel FP-Growth (PFP) [15] a exploité itemsets tag á partir de laquelle la page web itemsets sont générés, ce qui nécessite deux balayages sur la base de données. En utilisant MapReduce et son mécanisme de tolérance aux pannes, la tÃ¢che de Big Data Mining est convertie en d'autres petites tâches qui ne sont pas dépendants les uns des autres. La stratégie de regroupement des PFP a des problèmes de mémoire et de vitesse ; pour équilibrer les groupes de PFP Zhou et al.[19] a proposé un algorithme pour une exécution plus rapide en utilisant des éléments simples qui n'est pas également une façon efficace. Moens et al [9]. ont proposé deux méthodes pour l'extraction de motifs fréquents pour Big Data sur MapReduce , Première méthode Dist- Éclat la version distribue de Éclat, qui optimise la vitesse en répartissant l'espace de recherche de manière égale entre Mappers , deuxième méthode BigFIM utilise á la fois la méthode basée Apriori et Éclat avec des bases de données projetées qui conviennent á la mémoire afin d'extraire Frequent Itemset.

3.2.2.1 Contraintes exigées pour la parallélisassions d'algorithme en MapReduce

•Jugez si l'algorithme consiste d'étapes de traitement en parallèle : depuis une analyse théorique elle devrait considérer non seulement le parallélisme des étapes principales, mais aussi le parallélisme des étapes de raffinement des résultats.

•Jugez si l'algorithme répond aux conditions parallèles de MapReduce en théorie : si le traitement des données peuvent être divisées par l'algorithme et les résultats du traitement

partitionnées peuvent être fusionnés et le résultat final obtenu.

•Assurer le coût du temps de l'algorithme avec Map-Reduce les opérations ne peuvent pas être trop car l'algorithme ne tolère pas trop d'opérations d'itération, parce que le déclanchement de MapReduce.

•Prend trop de temps et ceci peut mettre en cause l'amélioration amené par MapReduce.

•Il est nécessaire de veiller á ce que l'efficacité de l'algorithme est promue après la parallélisassions MapReduce ; sinon á quoi bon d'utiliser MapReduce.

•Veiller á l'exactitude des résultats de l'algorithme après la parallelisaion a l'aide de MapReduce, pour assurer que les résultats sont similaires á celle obtenue en traitement en série.

3.3 Travaux basés CUDA

3.3.1 DBSCAN : Design and optimization of dbscan algorithm based on cuda

DBSCAN est un algorithme trés classique pour le CLUSTERING, ce dernier est largement utilisé dans de nombreux domaines. Cependant, avec l'explosion des volumes de données, l'algorithme de série traditionnel ne peut répondre á l'exigence de performance. Récemment, le calcul parallèle basé sur CUDA á gagner de l'ampleur et as débuté á tenir un territoire important en matière de traitement et analyse de BIG DATA. Le présent document est organisé de la façon suivante [20] : La premiére partie donne une brève introduction des concepts sur DBSCAN et résume les algorithmes proposēs auparavant. La deuxiéme partie décrit l'algorithme DBSCAN basé sur GPU ce dernier utilise la mémoire partagée d'une façon plus exhaustive que les autres algorithmes et cette algorithme á exhiber une trés bonne adaptation aux données volumineuse (Scalability). La troisiéme partie a donné les détails sur le procédé par lequel on peut améliorer l'algorithme basé sur CUDA. Dans la quatriéme partie, une analyse des résultats est établit et une conclusion est retenue qui vient comme suit le nouveau algorithme est 97 fois plus rapide que l'algorithme ordinaire.

3.3.2 CUDA : Parallel data mining techniques on graphics processing unit with compute unified device architecture (CUDA)

Le développement récent en unités de traitement graphique (GPU) a permis le calcul á haute performance pour les applications á usage général. Data mining est largement utilisée dans des applications importantes dans divers domaines. Cependant, les boites á outils de data mining actuelles ne peuvent pas répondre á les exigences d'applications avec des bases de données á grande échelle en termes de vitesse. Dans cet article [21], l'auteur propose trois techniques pour résoudre les problèmes fondamentaux des algorithmes de big data mining sur la plateforme CUDA :

1- un système évolutif d'ordonnancement des threads pour motif irrégulier.

2- Un système parallèle distribué top-k.

3- Un schéma parallèle de réduction de dimension.

Ils jouent un rôle clé dans la mise en œuvre des trois algorithmes de Data mining représentatives, CU-Apriori, CU-KNN et CU-K-moyens á base de CUDA. Ces implémentations parallèles surpassent les autres implémentations mentionnées dans l'état de l'art de manière significative sur un poste de travail xw8600 HP avec un GPU Tesla C1060 et un quad-Core Intel Xeon. Ses résultats ont montré que l'architecture parallèle GPU + CUDA est faisable et prometteuse pour les applications de Big Data Mining.

3.4 Travaux basés Memory Mapping

Mmap : Fast billion-scale graph computation on a pc via memory mapping

Des approches graphique de calcul, tels que GraphChi et TurboGraph ont récemment démontré qu'un seul PC peut effectuer des calculs efficaces sur des graphes contenant des milliards de nœuds. Pour Atteindre une haute performance et évolutivité, ils ont souvent besoin de structures de données sophistiquées et des stratégies de gestion de mémoire avancé. L'auteur propose une approche minimaliste qui renonce à ces exigences, en tirant parti de la capacité de Memory Mapping (MMap) disponible aux niveaux des systèmes d'exploitation. La contribution de [22] consiste en :

(1) une nouvelle notion que MMap est une technique viable pour la création d'algorithmes de graphes rapides et évolutifs qui dépasse les meilleures techniques ;

(2) la conception et la mise en œuvre d'algorithmes de graphes populaires pour des graphes de taille massive avec peu de code, grâce à Memory mapping ;

(3) des expérimentations abondante sur des graphes réelles, y compris le YahooWeb contenant 6,6 milliards de sommets, et montrer que cette nouvelle approche est brutalement plus rapide ou semblable aux méthodes hautement optimisées par exemple, il est 9.5 fois plus rapide que GraphChi pour calculer PageRank sur TwitterGraph (1.47 billions de nœuds) .

3.5 Synthèse des travaux basés (MR, CUDA et Memory Mapping)

	Application/Algorithme		
Support de l'algorithme	**Real Time Processing**	**Taille de donnée supportée**	**Tolérance aux Tâches itérative**
Hadoop/MapReduce	★	★★★★★	★★
GPU/CUDA	★★★★	★★	★★★★
MemoryMapping/MPI	★★★★★	★★	★★★★

Tableau II.2 – Les Algorithmes basés sur les différentes plateformes

•**Real Time Processing** :

Traitement en temps réel d'un système réside dans sa capacité à traiter les données et produire les résultats dans certaines contraintes de temps. Les réponses en temps réel sont souvent livrées en l'ordre de quelques millisecondes, et parfois microsecondes en fonction de l'application et les besoins des utilisateurs. Dans cette catégorie, CUDA et memory Mapping score 5 étoiles et surpassent les autres plates-formes. GPU avec ses milliers de cœurs de traitement et memory mapping avec sa haute bande passante de mémoire sont bien adapté pour traitement en temps réel des données. Bien que leur mémoire soit limitée, les GPU et memory mapping sont optimisées pour la vitesse et sont souvent utilisés pour le traitement en ligne.

•**Taille de donnée supportée** :

La taille des données supportée est la taille du DataSet que le système peut traiter et gérer efficacement. Dans cette catégorie, Hadoop/MapReduce ont 5 étoiles car ils peuvent évoluer jusquà' des dizaines de milliers de nœuds et de tels plateformes sont capables de traiter et de manipuler les Big Data. GPU et Memory Mapping ne sont pas bien adaptés pour le traitement de grands ensembles de données. Car ils ont une mémoire limitée à bord de l'ordre de plusieurs gigaoctets. Et donc ils obtiennent 2 étoiles.

•**Tolérance aux taches itérative** :

Ceci est la capacité d'un système de soutenir efficacement les tâches itératives. Comme beaucoup de tâches et algorithmes d'analyse de données sont de nature itérative, il est un indicateur important de comparaison des différentes plates-formes, en particulier dans le contexte de big data mining

GPU et Memory Mapping ont 4 étoiles dans cette catégorie et tous leur conviennent trèsp bien pour les algorithmes itératifs car ces plates-formes sont parfaitement adaptées pour les algorithmes itératifs, le résultat d'une itération peut être facilement utilisé dans la prochaine itération et tous les paramètres peuvent être stockés localement. Cependant, tous les algorithmes itératifs ne peuvent pas être facilement modifiés pour fonctionner sur chacune de ces plates-formes qui est la principale raison pour laquelle nous les avons donnés 4 étoiles au lieu de 5 étoiles. Hadoop/MapReduce n'est pas conçus pour gérer des tâches itératives et donc il ne reçoit que 2 étoiles. MapReduce n'est pas conçu pour le traitement itératif et les données doivent être écrites sur le disque après chaque itération, ainsi rendant E/S des données sur le disque un énorme goulot d'étranglement. Certains développements récents tels que HaLoop améliore les performances de MapReduce pour les tâches itératives dans une certaine mesure, qui est la raison pour laquelle on l'as pas donné 1 étoile.

4 Conclusion

Depuis la panoplie d'articles abordés dans l'état de l'art nous avons déduit que pour appliquer les techniques de data mining sur les big data il faut avoir des ressources matérielles et softwares qui ne sont pas á la disposition de tout le monde et l'astuce serait d'exploiter les ressources du cloud computing pour achever notre objectif, la solution serait abordée en détails dans le prochain chapitre.

CHAPITRE III

BIG DATA MINING BASÉ SUR LE CLOUD COMPUTING

1 Introduction

Après avoir établie l'état de l'art dans le chapitre précèdent nous avons figuré qu'il faut avoir des ressources matérielles et software pour être en mesure de traiter les big data, a partir de ce fait nous avons songé à tourner vers le cloud pour profiter de ces ressources.

Dans ce chapitre nous montrons les différentes solutions proposées en cloud pour soutenir le big data mining puis nous détaillons chaque aspect de notre approche.

2 Définition des concepts

2.1 Qu' est-ce que le Cloud Computing ?

Le cloud computing se traduit littéralement par "informatique dans les nuages", faisant référence aux technologies d'internet qui est souvent représenté schématiquement par un nuage..

C'est un ensemble de services qui permet à un utilisateur, de transférer ses stockages et ses traitements informatiques, depuis un ou plusieurs de ses serveurs locaux, vers d'autres serveurs extérieurs. . Mais à la demande d'un utilisateur, le Cloud Computing lui ouvre simplement l'accès à des ressources informatiques virtuelles, via internet.

Son but est de pousser les entreprises à externaliser les ressources numériques qu'lles stockent.

Ces ressources offrant des capacités de stockage et de calcul, des logiciels de gestion de messagerie, et d'autres services sont mis à disposition par des sociétés accessibles grâce à un système d'identification, via un PC et une connexionà Internet.

2.2 Les différents services du cloud computing

2.2.1 IaaS (Infrastructure as a Service)

Dans le modèle IaaS, seul le matériel (serveurs, baies de stockage, réseaux) qui constitue l'infrastructure est hébergé chez un prestataire ou un fournisseur.

2.2.2 Paas (Platform as a Service)

Le modéle PaaS se place sur le niveau supérieur du IaaS en offrant aux entreprises un environnement leur permettant de déployer leurs développements. Le PaaS fournit ainsi des langages de programmation, des bases de données et différents services pour faire fonctionner leurs applications. De plus, il automatise entièrement le déploiement (mises à jour, correctifs, etc)

2.2.3 SaaS (Software as a Service)

Le modèle SaaS fournit des applications à l'utilisateur sous la forme d'un service prêt à l'mploi qui ne nécessite aucune maintenance : les mises à jour étant régulièrement faites par l'éditeur. Il est donc perçu, à juste titre par les utilisateurs, comme un modèle de consommation des applications.

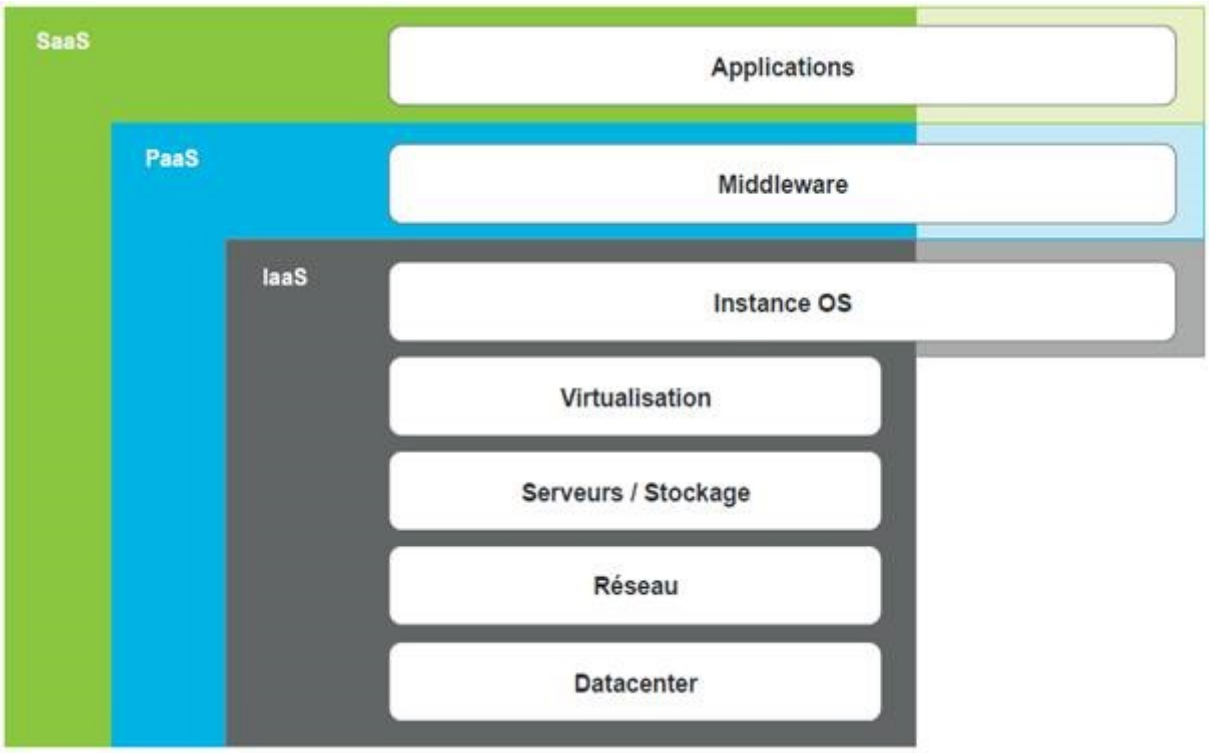

Figure III.1 – Les services du cloud computing

3 Types de Cloud Computing

Le concept de Cloud Computing est encore en évolution. On peut toutefois dénombrer trois types de Cloud Computing :

•Cloud privé (ou interne) : réseau informatique propriétaire ou un centre de données qui fournit des services hébergés pour un nombre limité d'utilisateurs.

•Le cloud public (ou externe) : prestataire de services qui propose des services de stockage et d'applications Web pour le grand public. Ces services peuvent être gratuits ou payants.

•Le cloud hybride (interne et externe) : un environnement composé de multiples prestataires internes et externes.

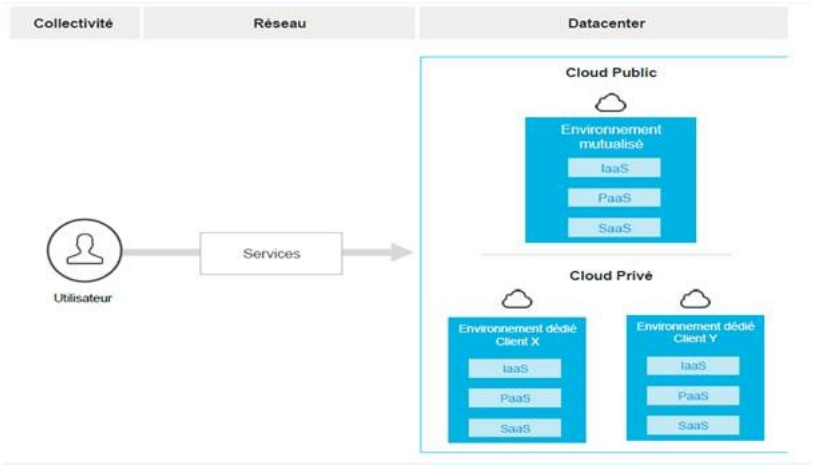

Figure III.2 – Types de cloud computing

4 Travaux connexes

–**Making knowledge discovery services scalable on clouds for big data mining** [23] Dans ce papier, introduit le sujet de "comment rendre la découverte de connaissance scalable pour le big data mining ", met l'accent sur les questions principaux de recherche, et présente une plateforme de data mining dans le cloud conçu pour développer et exécuter des applications distribuées d'analyse de données en tant que des services Workflow. Dans cet environnement, [23] utilise des DataSet, des outils d'analyse, des algorithmes de data mining et des modèles de connaissances qui sont mis en œuvre comme des services élémentaires qui peuvent être combinés par l'intermédiaire d'une interface de programmation visuelle dans des workflows distribuée prêt à être exécutée sur le cloud. Les principales caractéristiques de l'interface de programmation sont décrites et une évaluation des perforamances est établie.

–**Big Data Analytic Using Cloud Computing** [24]

Ce papier traite de divers problèmes liés au big data mining et propose des solutions en utilisant cloud computing. Le papier débute en premier par la définition des difficultés rencontrées dans le big data minig puis il montre ce qui a été proposé dans le domaine big

data dans le cloud et enfin il propose l'architecture Suivante :

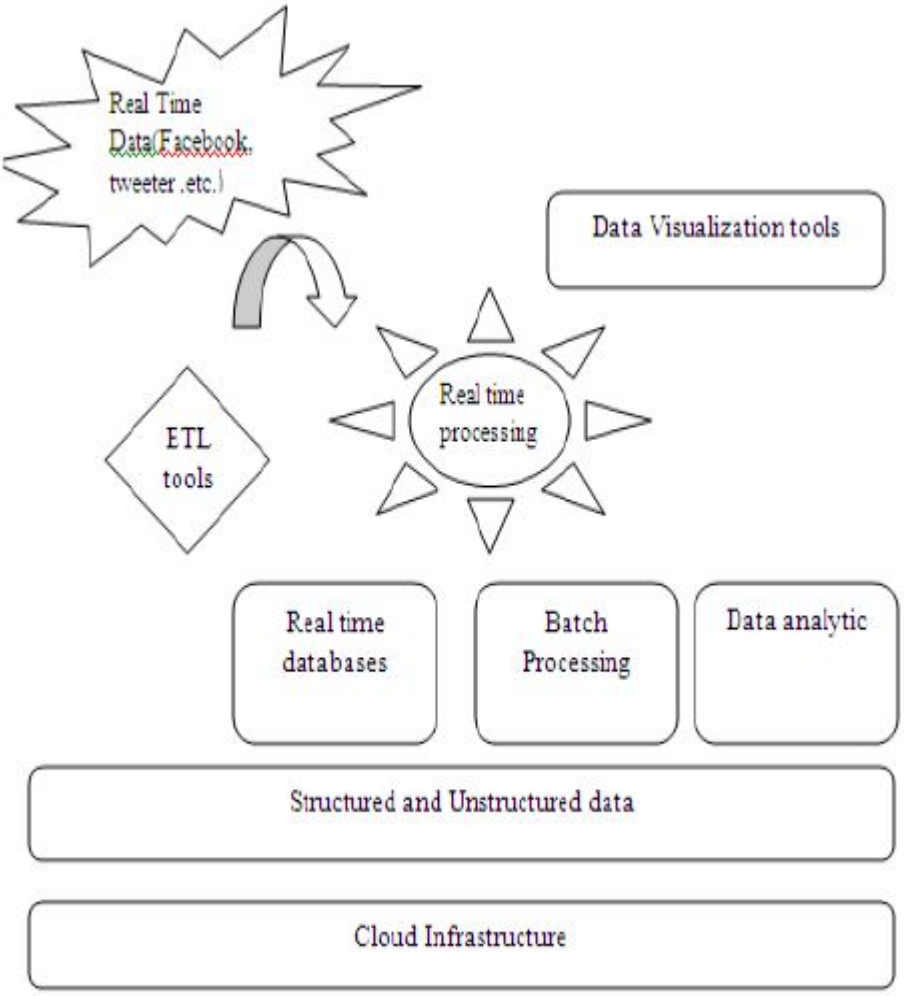

Figure III.3 – Architecture Big Data Mining

IBM Research a identifié AaaS comme étant un domaine qui peut avoir une valeur commercial [25]. Ceci est due au fait que AaaS peut permettre la transformation des données non structurée en opportunité lucratif. A partir de ce contexte, IBM a opté pour la création d'une Plateforme AaaS qui peut aider les utilisateurs finaux et les entreprises à soumettre leur donnée que ce soit en format structurée on non structurée afin de profiter de services d'analyse de données. Cette plateforme vise à réduire le fardeau financier des entreprises et les épargner la gestion d'analystes internes de données ; EMC partage le même point de vue [26]. IBM a identifiéles obstacles qui doivent être surmonté pour matérialiser AaaS :

•Définitions des accords de niveau de service.

•Qualité des méthodes de surveillance de service.

•Tarification.

•Gestion des données non structurées.

•processus d'affaires (modèles).

un autre contributeur majeure dans le déploiment des platforme AaaS est SAS [27]. Le service AaaS exhibé par SAS est basée sur les analyses prèdictive qui permetent aux clients d'acceder facilement à des solutions pour leur problèmes d'entreprise.

Sun et al. [28] propose une plateforme AaaS comme une extension de Software-as-a-Service (SaaS) qui permet aux entreprise d'accederà des analyses de données en tant que services . Tout en se concentrant sur mutualisation , les auteurs proposent une méthodologie de personnalisation des SLA qui prend en charge de multiples demandes d'analyse de capacité des locataires . Leur architecture détaille comment les serveurs virtuels sont conçus pour accepter l'ntrée des données des utilisateurs. Les serveurs qui exécutent les analyses réelles sont appelés scoring servers .

Deepak et al [29] . détaille aussi la conception architecturale de leur plate-forme AAAS qui est hébergé en cloud. C'est pratique pour les utilisateurs de télécharger des fichiers et des données sur une interface web qui est le moyen d'interaction avec le système. Les fichiers téléchargés sont poussés vers les moteurs d'analyse qui comprennent SPSS , R , SAS , Cluto et WEKA . Ce type de plate-forme AAAS nécessite des technologies d'autres fournisseurs aussi bien et le résultat analysé est stocké dans des systèmes de données tels que les systèmes Oracle, DB2.

5 Traitement et analyse des BIG Data

5.1 Architecture proposée

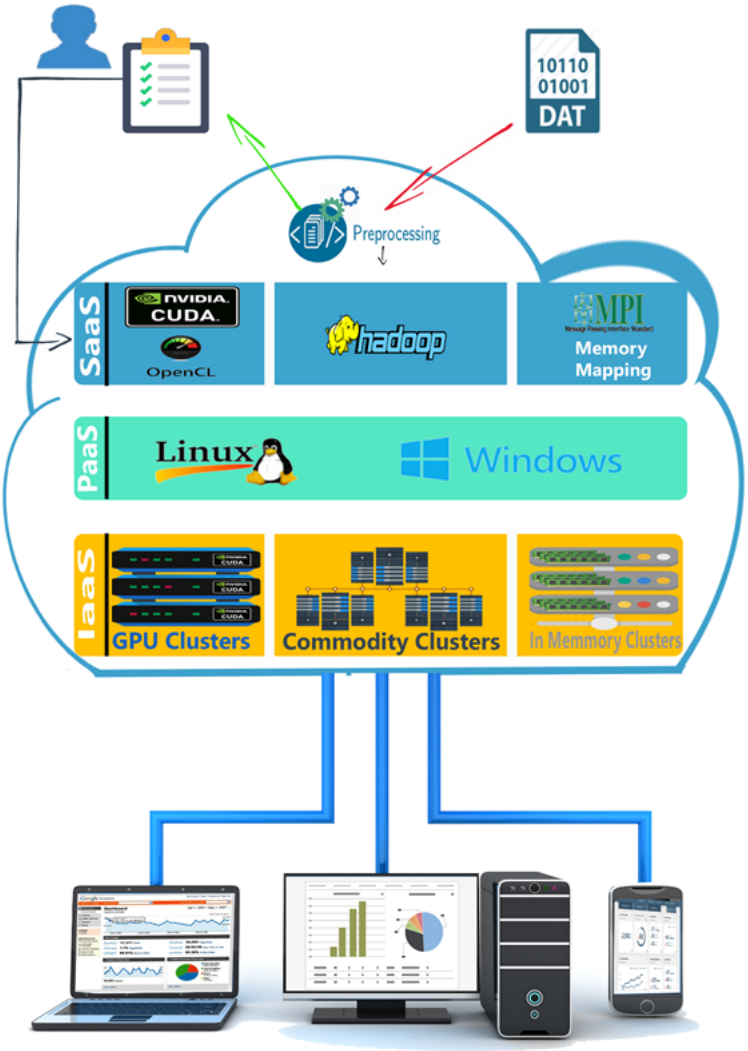

Figure III.4 – Architecture proposée

5.2 Présentation de l'architecture

SaaS Cette couche représente une livraison conjointe de moyens, de services et d'expertise qui permettent aux entreprises d'externaliser intégralement le traitement et l'analyse des BIG Data ce traitement est assuré par le biais de trois plateforme comme le montre la figure suivante :

Figure III.5 – Service SaaS

• MapReduce : une plateforme de traitement distribué sur de gros volumes de données, cette plateforme utilise des commodity clusters pour être en mesure de gérer les Big Data.

• Cuda : qui est une architecture de traitement parallèle développée par NVIDIA, permettant de décupler les performances de calcul du système en exploitant la puissance des processeurs graphiques, ces processeurs graphiques sont offerts par la couche infrastructure à travers un cluster GPU.

• Memory mapping : Ce dernier offrent un contrôle absolue sur la réservation de segment en mémoire centrale en établissant une projection en mémoire des fichiers. Pour être en mesure de projeter des fichiers de taille massive cette couche exploitera un cluster In memory qui offrent une mémoire de taille massive.

Figure III.6 – Service PaaS

Système d'exploitation : Cette couche contient les systèmes d'exploitation sur lesquels opérés les services offerts dans la couche de services. Ces systèmes sont windows 7 et linux Ubuntu. Hadoop/MapReduce opèrent sous linux, alors que le memory mapping et cuda opèrent sur les deux systèmes linux ubuntu et windows 7.

Infrastructure : L'infrastructure de cette architecture est composée de trois type de clusters chacun dédié a un des services offert.

•**Cluster de GPU** :

il représente une grappe d'ordinateurs dans lequel chaque nœud est équipé d'une unité de traitement graphique (GPU). En exploitant la puissance de calcul des GPU modernes via General-Purpose Computing sur des unités de traitement graphique (GPGPU) des calculs très rapides peuvent être effectuées.

•**Cluster in-memory computing** :

Il représente une grappe d'ordinateurs dans lequel chaque nœud est équipé d'une unité de traitement graphique (GPU). En exploitant la puissance de calcul des GPU modernes via General-Purpose Computing sur des unités de traitement graphique (GPGPU) des calculs très rapides peuvent être effectuées.

•**Commodity clusters** : Il utilise un type de logiciel qui permet de stocker des données dans la RAM, à travers un groupe d'ordinateurs, et les traiter en parallèle. Considérons les ensembles de données opérationnelles généralement stockées dans une base de données centralisée qui peuvent maintenant être stockée en RAM sans oublier que la RAM, est à peu près, 5000 fois plus rapide que le disque. Ajouter à tous sa le traitement parallèle, et les calculs deviennent vraiment rapide.

5.2.1 Définitions des critères d'évaluation

	Plateforme		
Support de l'algorithme	**Scalabilité**	**Performance d'E/S de donnée**	**Tolérance Aux Pannes**
MapReduce	★★★★★	★★	★★★★★
CUDA	★★	★★★★	★★★★
Memory Mapping/MPI	★★	★★★★★	★★★★

Tableau III.1 – Évaluation des plateformes

Le tableau ci-dessus représente une évaluation des trois plateformes offerte.

•**La scalabilité** : est définie comme la capacité d'un produit à s'adapter à un changement d'ordre de grandeur de la demande, en particulier sa capacité à maintenir ses fonctionnalités et ses performances en cas de forte demande. Dans notre cas, la scalabilité est considéré comme la possibilité d'ajouter plus de matériel pour améliorer la capacité et la performance

d'un système.

•**Performance d'E / S de données** : se réfère à la vitesse à laquelle les données sont transférées vers / à partir d'un périphérique. Dans le contexte du big data mining, cela peut être considéré comme le taux dans lesquels les données sont lues et écrites dans la mémoire (ou un disque) ou le débit de transfert de données entre les nœuds d'un cluster.

•**Tolérance aux pannes** : est la caractéristique d'un système qui lui permet de continuer à fonctionner correctement dans l'éventualité d'une défaillance d'un ou de plusieurs composants.

5.2.2 Évaluation des critéres

Cette discussion sera menée selon les critères mentionnés dans le tableau comparatif :

Scalabilité :

Pour le critère de scalabilité Hadoop/MapReduce auront 5 étoiles, du fait qu'il supporte vraiment le passage à l'échelle, et pour cette plateforme il est relativement facile d'ajouter plus de machine. 2 étoiles pour CUDA montre que la scalabilite n'est pas le point fort de GPU cluster .Il y a une limite sur le nombre de GPU qu'une seule machine peut avoir et l'ajout de plusieurs machines va créer un goulot d'étranglement pour le transfert de donnée sur le réseau. Même memory Mapping as 2 étoiles car le passage à l'échelle est vraiment pénible pour in memory clusters.

Performance d'E / S de données :

GPU reçoivent 4 étoiles, car il a une mémoire à haut débit et les opérations d'entrée /sortie des données sont extrêmement rapides. Les GPU de la génération actuelle sont disponibles avec une mémoire GDDR5 qui est beaucoup plus rapide que la mémoire système DDR3, Hadoop/MapReduce ont 2 étoiles, comme ils lisent principalement les données du disque qui est un peu lent. En outre, trop de communication sur le réseau dégrade la performance. Memory mapping as 5 étoile pour ce critère car il est par excellence le meilleur parmi les trois plateformes car toutes les données sont stockées en mémoire centrale (RAM) et celle-ci as une vitesse vraiment élevé de lecture et écriture 5000x de plus que le disque

Tolérance aux pannes :

Hadoop/MapReduce as 5 étoiles, car c'est le seul parmi ces plateforme qui offre un mécanisme de tolérance de pannes fiable et robuste. D'autre part, GPU et memory mapping obtiennent 4 étoiles. Bien qu'aucune de ces plateformes contient un mécanisme de tolérance aux panne, mais ils ont un matériel fiable et bien édifié qui rend la défaillance matérielle un événement extrêmement rare.

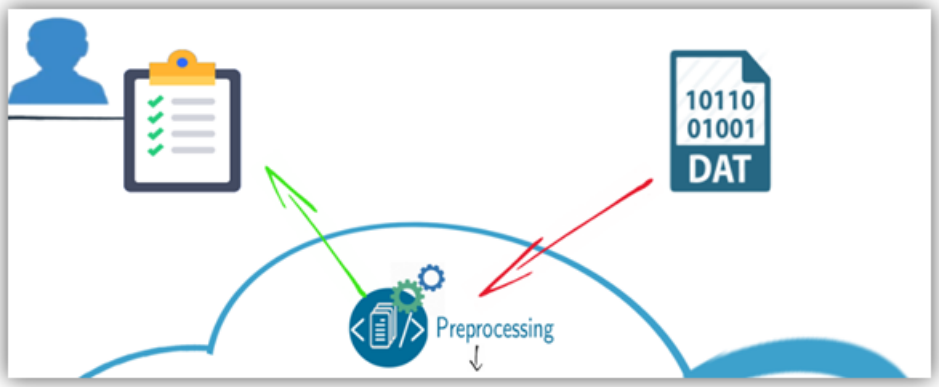

Figure III.7 – Prétraitement

Prétraitement :

Cette tâche est l'axe principal de notre service cloud, le prétraitement consiste à aider L'utilisateur à faire le bon choix de plateforme, ce choix est obtenu à travers un formulaire comme le montre la figure(III.8)

Ce formulaire est rempli par l'utilisateur, il nous permet de récolter des informations sur le type et la taille des données ainsi que les préférences de l'utilisateur.

Figure III.8 – Formulaire vide

Figure III.9 – Formulaire rempli

Aprés avoir rempli le formulaire l'opération de prétraitement commence en récoltons les préférences de l'utilisateur et les jugeant á travers l'évaluation effectué dans les 2 tableaux (II.2,III.1) Exemple : dans le cas du formulaire montré dans la figure remplit le prétraitement affichera la plateforme MapReduce comme étant la plus adéquate pour l'utilisateur car

il a précisé que la scalabilité et la tolérance aux pannes sont les caractéristiques les plus appropriés pour son application.

6 Conclusion

Depuis le travail achevé dans ce chapitre, nous avons déduit que le cloud sera une solution adéquate pour faciliter le big data mining du fait qu'il contient ressources adaptées au calcul intensif et ce qui manque est une service qui pourrait exploiter d'une maniére appropriée ces ressources tous en aidant l'utilisateur à prendre la bonne décision.

CONCLUSION GÉNÉRALE

Big Data et son effet de boule de neige a instauré une nouvelle discipline, celle du Big Data Mining, qui a provoqué une métamorphose inattendue en matière de gain scientifique.

Comme l'extraction de connaissances à partir des big data est un sujet d'actualité, le besoin était clair, celui de : comment adapter les méthodes classiques de data mining afin d'être en mesure de tirer profit des connaissances extraites à partir des big data?

Après avoir établi l'état de l'art nous avons pu dégagé trois types de solutions algorithmiques, la première basée MapReduce, la deuxième Cuda et la dernière est Memory Mapping.

Ensuite nous avons fait une comparaison entre chacune de ces solutions pour montrer leurs points forts et leurs vulnérabilités, après cette comparaison nous avons remarqué que l'implémentation de ces solutions nécessite des ressources matérielles et software, alors nous avons proposé une architecture qui se manifeste en un service cloud, qui a provoqué une naissance d'une symbiose entre toute entité désirant l'extraction des connaissances des big data et la bonne plateforme à adopté.

Dans nos futurs travaux nous désirons implémenter notre service cloud toute en améliorant la prise de décision offerte par ce dernier.

BIBLIOGRAPHIE

[1] Stefane Fermigier. Big data and open source une conférence inévitable, 2013.

[2] Amrane Abesselam. Big data concepts et cas d utilisations, 2015.

[3] Wikipédia. Exploration de données wikipédia, l'encyclopédie libre, 2016.

[4] Jiawei Han, Jian Pei, Yiwen Yin, and Runying Mao. Mining frequent patterns without candidate generation : A frequent-pattern tree approach. *Data mining and knowledge discovery*, 8(1) :53–87, 2004.

[5] Jiawei Han, Jian Pei, and Yiwen Yin. Mining frequent patterns without candidate generation. *SIGMOD Rec.*, 29(2) :1–12, May 2000.

[6] wikipédia. Régle d'association — wikipédia, l'encyclopédie libre, 2014.

[7] Wikipédia. Arbre de décision (apprentissage) — wikipédia, l'encyclopédie libre, 2016. [En ligne ; Page disponible le 22-mars-2016].

[8] Wikipédia. Méthode des k plus proches voisins — wikipédia, l'encyclopédie libre, 2015. [En ligne ; Page disponible le 22-mars-2016].

[9] Sandy Moens, Emin Aksehirli, and Bart Goethals. Frequent itemset mining for big data. In *Big Data, 2013 IEEE International Conference on*, pages 111–118. IEEE, 2013.

[10] 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 (BigData Congress), 2015 IEEE International Congress on, pages 393–400. IEEE, 2015.

[11] Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules. In *Proc. of 20th Intl. Conf. on VLDB*, pages 487–499, 1994.

[12] Mohammed J. Zaki, Srinivasan Parthasarathy, Mitsunori Ogihara, and Wei Li. Parallel algorithms for discovery of association rules. *Data Mining and Knowledge Discovery : An International Journal*, 1(4) :343–373, Dec 1997. Special issue on Scalable High-Performance Computing for KDD.

[13] J. Han, J. Pei, and Y. Yin. Mining frequent patterns without candidate generation. In *Proc. ACM SIGMOD Int'l Conf. on Management of Data*, pages 1–12, May 2000.

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

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

[16] Sandy Moens, Emin Aksehirli, and Bart Goethals. Frequent itemset mining for big data. In *Big Data, 2013 IEEE International Conference on*, pages 111–118. IEEE, 2013.

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

[18] 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.

[19] Le Zhou, Zhiyong Zhong, Jin Chang, Junjie Li, Joshua Zhexue Huang, and Shengzhong Feng. Balanced parallel fp-growth with mapreduce. In *Information Computing and*

Telecommunications (YC-ICT), 2010 IEEE Youth Conference on, pages 243–246. IEEE, 2010.

[20] Bingchen Wang, Chenglong Zhang, Lei Song, Lianhe Zhao, Yu Dou, and Zihao Yu. Design and optimization of dbscan algorithm based on cuda. *arXiv preprint arXiv :1506.02226*, 2015.

[21] Liheng Jian, Cheng Wang, Ying Liu, Shenshen Liang, Weidong Yi, and Yong Shi. Parallel data mining techniques on graphics processing unit with compute unified device architecture (cuda). *The Journal of Supercomputing*, 64(3) :942–967, 2013.

[22] Zhiyuan Lin, Minsuk Kahng, Kaeser Md Sabrin, Duen Horng Polo Chau, Ho Lee, and U Kang. Mmap : Fast billion-scale graph computation on a pc via memory mapping. In *Big Data (Big Data), 2014 IEEE International Conference on*, pages 159–164. IEEE, 2014.

[23] Domenico Talia. Making knowledge discovery services scalable on clouds for big data mining. In *Spatial Data Mining and Geographical Knowledge Services (ICSDM), 2015 2nd IEEE International Conference on*, pages 1–4. IEEE, 2015.

[24] Vinay Kumar Jain and Shishir Kumar. Big data analytic using cloud computing. In *Advances in Computing and Communication Engineering (ICACCE), 2015 Second International Conference on*, pages 667–672. IEEE, 2015.

[25] Richard K Lomotey and Ralph Deters. Analytics-as-a-service (aaas) tool for unstructured data mining. In *Cloud Engineering (IC2E), 2014 IEEE International Conference on*, pages 319–324. IEEE, 2014.

[26] Richard K Lomotey and Ralph Deters. Analytics-as-a-service framework for terms association mining in unstructured data. *International Journal of Business Process Integration and Management*, 7(1) :49–61, 2014.

[27] Richard K Lomotey and Ralph Deters. Analytics-as-a-service (aaas) tool for unstructured data mining. In *Cloud Engineering (IC2E), 2014 IEEE International Conference on*, pages 319–324. IEEE, 2014.

[28] Xi Sun, Bo Gao, Liya Fan, and Wenhao An. A cost-effective approach to delivering analytics as a service. In *Web services (icws), 2012 ieee 19th international conference on*, pages 512–519. IEEE, 2012.

[29] P Deepak, Prasad M Deshpande, and Karin Murthy. Configurable and extensible multi-flows for providing analytics as a service on the cloud. In *SRII Global Conference (SRII), 2012 Annual*, pages 1–10. IEEE, 2012.