REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

MINISTERE DE L'ENSEIGNEMENT SUPERIEURE ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE IBN KHALDOUN - TIARET

MEMOIRE

Présenté à :

FACULTÉ MATHEMATIQUES ET INFORMATIQUE

DÉPARTEMENT D'INFORMATIQUE

Pour l'obtention du diplôme de :

MASTER

Spécialité : Génie logiciel

Par :

M^lle^ SAIDI Chahrazed

M^lle^ DANOUNE Imane

Sur le thème

Vers une meilleure optimisation pour l'analyse prédictive de données massives

---

Soutenu publiquement le .. / .. / 2019 à Tiaret devant le jury composé de :

Mr TALBI Omar	Grade	Université MCB	Président
Mr DJAFRI Laouni	Grade	Université MCB	Encadreur
Mr OUARED Abdelkader	Grade	Université MCB	Examinateur

Remerciement

Tout d'abord, nos remerciements vont aux Allah qui nous a éclairé le chemin du savoir et de nous avoir donné le bon sens et la grande volonté pour réaliser ce modeste travail.

Nous tenons à remercier notre encadreur, Monsieur DJAFRI Laouni, qui a supervisé notre travail tout en nous laissant une grande marge de liberté, nous le remercions pour son encadrement, sa disponibilité et la pertinence de ses remarques tout au long de la réalisation de ce projet.

Nous remercions également le président du jury Mr. TALBI Omar, et l'examinateur Mr. OUARED Abdelkader d'avoir accepté d'évaluer ce travail.

Nous voulons, à cette occasion avec le plus grand honneur, remercier sincèrement tous nos enseignants.

Enfin, nous tenons à exprimer nos sincères gratitudes aux les personnes qui ont vraiment contribué à l'élaboration de la présente de cet mémoire.

Nous espérons que ce travail aura la valeur souhaitée.

Merci à tous !

Dédicaces

Je dédie ce mémoire à :

La personne la plus chère dans ma vie : Mon père. Merci puisque tu as su m'inculquer le sens de la responsabilité, de l'optimisme et de la confiance en soi face aux difficultés de la vie, je te dois ce que je suis aujourd'hui et ce que je serai demain et je ferai toujours de mon mieux pour rester ta fierté et ne jamais te décevoir, tu resteras toujours mon exemple éternel et l'homme de ma vie.

La personne qui je suis nulle sans elle : Ma mère. Tu n'es pas seulement une mère pour moi, tu es mon idole, ma sœur, mon amie, mon héroïne, tous mes remerciements ne sont pas suffisant devant ce que tu as fait pour moi. Merci puisque tu n'as comblé avec ta tendresse et affection tout au long de mon parcours. Tu n'as cessé de me soutenir et de m'encourager durant toutes les années de mes études.

Mon coup de cœur, ma raison de sourire, tous ce qui est sucré dans ma vie : Mon neveu JOUJOU.

Mes frères: Moncef et Ayoub et mes sœurs: Manel et Yousra et son homme Ghalem, merci pour vos mots, vos amours et vos encouragements.

Mon oncle Mehenni qu'était un deuxième père pour moi, que Dieu te garde pour nous.

Ma tante Sarah qu'était plus qu'une sœur, merci pour ton soutenir.

Une personne très spéciale dans ma vie : Karima, merci d'avoir être dans vie.

Une amie avec un gout d'une sœur : Hanane, merci pour tous ce qu'on a passé ensemble.

Mon amie d'enfance qui m'a trop manqué : Djihad.

Mon grand-père et mes deux grandes mères.

Mon uncle Mohamed, Sidahmed, Khaled, Hamida, Foufou, Houhou, Fatima, Saadia, ma tante Yamina, Zahia, Amina.

Mon cousin Abdenour, mes cousines Maroua et Safa.

Mes amis : Djamel, Sofiane, Ilyes, Medjeda, Toussa et rahil.

Ma partenaire dans ce travail : SAIDI Chahrazed.

DANOUNE Imane.

Dédicaces

Tous d'abord Je dédie ce modeste travail à :

L'homme de ma vie, mon exemple éternel, mon soutien moral et source de joie et de bonheur, celui qui s'est toujours sacrifié pour me voir réussir, que dieu te garde dans son vaste paradis, à toi mon père.

La lumière de mes jours, la source de mes efforts, la flamme de mon cœur, ma vie et mon bonheur ; maman que j'adore.

Mes chers et adorables frères : Ihab et Abdelillah.

Toute ma famille de loin ou de près.

Mes meilleurs amis : Djamel, Ilyes, Jimmi, Samah, Sara, Ismahan, Mokhtaria, Hanane.

Ma partenaire dans ce travail : DANOUNE Imane

Tous mes collègues de ma promotion.

Tous les enseignants qui ont contribué à ma formation durant mes études.

SAIDI Chahrazed

Résumé

Depuis plusieurs années, nous assistons à une explosion de nouvelles sources de données diverses à granularité fine et à faible latence (dites « Big Data »). Il consiste à traiter, en temps réel, de très gros volumes de données extrêmement variées et à les analyser. Toutes les entreprises sont concernées, surtout celles qui possèdent de vastes gisements d'informations et souhaitent les passer au crible pour améliorer leur connaissance du client et optimiser leurs campagnes. Tous les chercheurs dans le domaine de l'analyse des données volumineuses sont convaincus que ces dernières sont influencées par le volume, la variété et la vitesse, où l'augmentation et la complexité des données sont directement proportionnelles au taux d'erreur élevé et au temps d'exécution lent. Par conséquent, notre travail proposé vise principalement à résoudre ces problèmes en améliorant le résultat de la prévision à un niveau acceptable et dans les plus brefs délais. Dans cette partie, nous avons proposé une idée pour extraire une base d'apprentissage représentative par la méthode d'échantillonnage aléatoire stratifié au niveau des lignes et par la méthode de la sélection des variables importante au niveau des colonnes en utilisant l'algorithme « les forêts aléatoire ».

Mots-clés : Analyse prédictive des données massives, Apache Spark, échantillonnage, Les forêts aléatoires, variables importantes.

Abstract

For several years, we have witnessed an explosion of new sources of diverse data with fine granularity and low latency (called "Big Data"). It consists of processing, in real time, very large volumes of extremely varied data and analyzing them. All companies are concerned, especially those with large amounts of information and want to screen them to improve their knowledge of the customer and optimize their campaigns. All researchers in the field of large data analysis are convinced that they are influenced by volume, variety, and speed, where the increase and complexity of data are directly proportional to the high error rate and slow execution time. Therefore, our proposed work is primarily aimed at solving these problems by improving the forecast result to an acceptable level and in the shortest possible time. In this part, we proposed one idea for extracting a representative learning base from the stratified random sampling method at the row level and by the method of the selection of variables important at the level of the columns using the algorithm « the random forests ».

Keywords : Big Data Predictive Analytics, Apache Spark, Sampling, Random Forests, Importance features.

S o m m a i r e

Introduction générale

Introduction Générale

1. Contexte & Motivation

Big Data et l'intelligence artificielle sont deux technologies inextricablement liées, au point que l'on peut parler d'une Big Data Intelligence. L'intelligence artificielle est devenue omniprésente dans les entreprises de toutes les industries, le besoin en matière de décisions plus intelligentes et de gestion du Big Data sont les critères qui dirigent cette tendance. La convergence entre le Big Data et l'IA semble inévitable à l'heure actuelle où l'automatisation des prises de décisions intelligentes se présente comme la prochaine évolution du Big Data. Une agilité en hausse, des processus business plus intelligentes et une meilleure productivité sont les bénéfices les plus probables de cette convergence.

L'intelligence artificielle va être utilisée pour extraire du sens, déterminer de meilleurs résultats, et permettre des prises de décisions plus rapides à partir de sources Big Data. L'intelligence artificielle est forte, elle sans le moindre doute une technologie formidable, elle peut trouver des données inaccessibles pour l'être humain, et distiller du sens avec meilleure précision.

L'essor de l'intelligence Artificielle et du Machine Learning dépend fortement du Big Data, car les données massives permettent de développer des modèles prédictifs, et permettent de représenter des concepts à apprendre.

Ces dernières années, nous devrions voir apparaître davantage d'experts dans ce domaine, mais la demande devrait rester supérieure à l'offre. Le Machine Learning favorise l'adoption des solutions du Big Data, au même titre, nous avons trouvé que le cloud computing, Hadoop, et Spark et d'autres solutions du big data qui facilitent le déploiement de ces données massives, car l'analyse de ces données demeure très complexe, à ce moment les entreprises peinent à transférer leurs données depuis les systèmes opérationnels vers les systèmes analytiques, cette difficulté affecte directement la productivité.

2. Problématique

Quelle est la méthode optimale pour traiter les données volumineuse ?, sachant que : on garde la stabilité du résultat de l'analyse prédictive à un niveau acceptable, en plus on peut faire le traitement en temps réel.

3. Objectif

Notre objectif au terme de ce projet est : la sélection des variables importantes du data set obtenu par l'extraction d'un échantillon représentatif du data set original, et l'obtention du résultat de l'analyse prédictive des données massives (*Big Data Predictive Analytics*) en streaming.

4. Organisation de manuscrit

Notre mémoire se décline en cinq chapitres :

- Le premier chapitre rappel sur les notions fondamentales sur le Big data ;
- Le second chapitre retrace également, les différents techniques pour le traitement du Big data ;
- Troisième chapitre est consacré à l'analyse prédictive des données massives avec des statistiques mathématiques et machine learning ;
- Le quatrième chapitre représente les défirent architectures et technologies du Big data le dernier chapitre réservé à la réalisation et l'implémentation de notre application ;

En fin, nous cloutirons ce mémoire pour une conclusion qui décrit panoramiquement le travail réalisé et les résultats obtenus.

Chapitre I

Généralités et concepts sur le Big Data

I . Introduction

De nos jours, les quantités massives des données sont produites en raison de la croissance du Web, de l'essor des médias sociaux, de l'utilisation du mobile et de l'Internet des objets et par des personnes, des choses et de leurs interactions.

Le Big Data est devenu une vraie tendance de fond et sans conteste, le domaine de mode ultime de l'informatique, il se retrouve dans un état de développement très émouvant, et il continue à évoluer très activement. Le Big Data a décrit le problème d'explosion quantitative des données venant du n'importe quel support, tant que les outils classiques ne se trouvaient plus en mesure de suivre leur rythme.

Ce chapitre a pour but d'éclaircir les notions fondamentales liées au Big Data tels que les définitions, les caractéristiques, les architectures et les domaines d'application.

I.1. Big Data

I.1.1. Définition

Le terme "données volumineuses" désigne un grand ensemble de données, ces données sont complexes, et très difficiles à traiter avec des outils de traitement de données traditionnels [1].

Big Data est le stockage et l'analyse de données complexes et massives. Ces données ne peuvent pas être stockées sous la forme d'une base de données relationnelle, en raison de la plus grande quantité de données, ainsi que de la non-structuration de ces données diverses. De plus, elles doivent être analysées rapidement [2].

Les données volumineuses sont des actifs d'information volumineux, à grande vitesse et très variés qui nécessitent des formes de traitement de l'information novatrices et rentables pour une meilleure compréhension, une prise de décision et une automatisation des processus [3].

En général, la définition du terme Big Data est détaillée comme suit :

- **Littéralement**

Grosse données ou volume massif de données structurées ou non. On parle aussi de data masse par similitude avec la biomasse [4].

❖ **Conceptuellement**

Ce terme vulgarise à la fois la représentation du volume des données mais aussi les infrastructures liées au traitement de ces données [4].

Finalement, l'expression « Big Data » (traduite en français par « méga données » ou « données massives ») désigne la masse hétérogène des données numériques produites par les entreprises et les particuliers dont les caractéristiques (très grand volume, diversité de forme, vitesse de traitement) requièrent des outils d'analyse informatiques spécifiques [5].

Figure I.1. Big Data

I.1.2. Caractéristiques du Big Data

Depuis l'apparition de l'Internet, nous avons assisté à une croissance explosive du volume, de la vitesse et de la variété des données créées quotidiennement. Ces données proviennent de nombreuses sources, notamment les appareils mobiles, les capteurs, les archives individuelles, l'Internet des objets, les bases de données gouvernementales, les journaux de logiciels, les profils publics sur les réseaux sociaux, les ensembles de données commerciales, etc.

Le Big Data se définit principalement par les 3Vs : Volume, Variété et vélocité. Mais avec le développement des données dans ce domaine, on parle aujourd'hui de 10Vs au lieu de 3Vs. Voici une petite description de chacune de ces caractéristiques :

I.1.2.1. Le Volume

La première chose que tout le monde pense du Big Data est sa taille [6,7]. À l'ère de l'internet (Internet ne dort jamais), notamment les réseaux sociaux produisant des données en continu, dont les volumes augmentent de manière exponentielle [8,9].

Le volume correspond à la grande quantité de données à l'échelle de données générées à chaque instant. On estime que 90% des données disponibles dans le monde ont été générées lors des deux dernières années. Chaque seconde, 29 000 giga-octets d'informations sont publiés sur internet. Un volume de données qui croît de manière exponentielle pour atteindre 163 zetta-octets en 2025, selon le cabinet d'analystes *IDC(International Data Corporation).*Les systèmes d'information des entreprises traitent des données en téraoctets ou péta-octet [6].

I.1.2.2. La Variété

La Variété signifie le traitement de la grande diversité des données issues de connexions, de données mobiles, de géolocalisations, de machines, de capteurs de mesures, de flux, de l'utilisation des médias sociaux, de textes (journaux, SMS, mails…), de vidéos, de sons et de transactions sous fourme structuré, semi structuré et non structuré mais devant faire l'objet d'une analyse collective [10].

I.1.2.3. La Vélocité

La vélocité traduit la vitesse de l'élaboration et du déploiement des données. Les évolutions technologiques permettent aux entreprises et aux consommateurs de générer des données à des fréquences intenses. Pour les entreprises possédant la puissance de calcul et les outils d'analyse nécessaires, le traitement des données peut se faire de manière instantanée. En effet, la majorité des données perd de sa valeur à mesure que le temps passe. Il convient donc d'utiliser rapidement les données [10].

I.1.2.4. La véracité

La véracité concerne la qualité et la crédibilité de la source et la pertinence des données pour le public cible. Elle donne un caractère qualitatif au Big Data. Pour être en possession de contenus authentifiés, les spécialistes du Big Data devront mettre en place des techniques permettant la gestion des contenus des données [10]. Par exemple, les faux messages ou les spams sont très répandus et font de la confiance un défi majeur [11,12].

I.1.2.5. La Valeur

La valeur est un facteur majeur que toutes les organisations doivent prendre en compte lors de la mise en œuvre de données volumineuses, parce que les autres caractéristiques du Big Data n'ont pas de sens si vous ne tirez pas de valeur commerciale de ces données, Nous pouvons donc dire que la valeur aide les entreprises à mieux comprendre leurs clients [13,14].

I.1.2.6. La variabilité

La variabilité dans le contexte du Big Data fait référence à plusieurs choses différentes comme la vitesse incohérente et le nombre d'incohérences dans les données. Ces dernières doivent être trouvées par les méthodes de détection des anomalies et des valeurs aberrantes pour que toute analyse significative puisse se produire [15].

I.1.2.7. La visibilité

La visibilité proposée par les plates-formes du Big Data permet aux utilisateurs de bien utiliser les données et informations publiées en effectuant des recherches pertinentes [16].

I.1.2.8. La validité

La validité fait référence à la validation des données, c'est-à-dire : vérifier si les données utilisées sont correctes et exactes pour l'utilisation envisagée, de sorte que ces données sont donc utilisées pour évaluer la performance de la prévision [17].

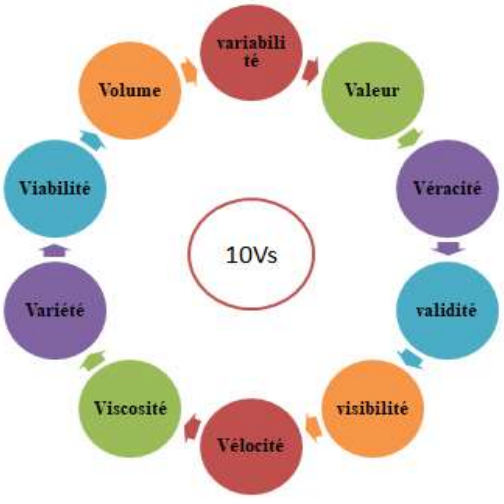


Figure I.2. Les 10 Vs du Big Data.

I.1.2. Architectures du Big Data

Le succès du fonctionnement du Big data dépend de leur infrastructure correcte et de leur utilité que l'on fait '' ***Data into Information into Value*** ''.

L'architecture du Big data est composée de cinq grandes parties *:* Intégration de données, Stockage de données, traitement de données, Sécurité de données et Opérations de données [2].

- **Intégration de données:** Consiste à charger le volume de données au sein du stockage.
- **Stockage de données:** C'est le stockage du volume de données.
- **Traitement de données :** Il s'agit de la manipulation et de traitement des données.
- **Sécurité de données :** Sert à l'autorisation, l'authentification et la protection des données.
- **Opérations de données :** Pour la gestion, le monitoring et les taches planifiés.

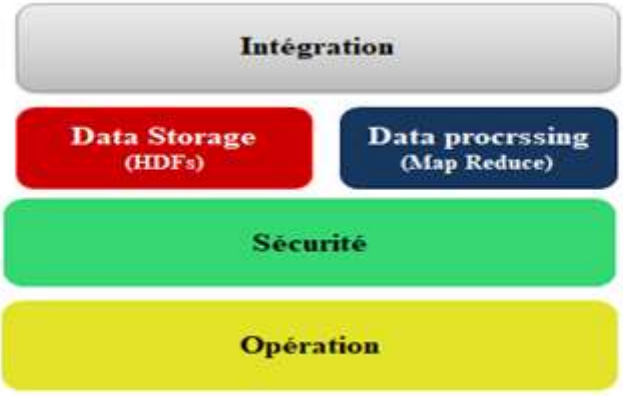


Figure I.3. Architecture du Big Data.

I.1.3.1. Avantages des architectures du Big Data

Plusieurs avantages peuvent être associés à une architecture du Big Data [18], nous citons par exemple :

- **Extensibilité (scalabilité) :** Le concept Big Data apporte une architecture scalable qui peut prévenir la taille d'infrastructure et l'espace disque nécessaire.
- **Performance :** Grâce au traitement parallèle des données et à son système de fichiers distribués, le concept Big Data est hautement performant en diminuant la latence des requêtes.
- **Coût faible :** Les baies de stockage de données centralisées ne seront pas nécessaires étant souvent coûteuses, grâce au système de fichiers distribué, nous ne serons satisfaits que des disques internes des serveurs.
- **Disponibilité :** On a plus besoin de disques, souvent coûteux. L'architecture Big Data apporte ses propres mécanismes de haute disponibilité.

I.1.4. Les domaines d'application du Big Data

Le Big Data trouvait sa place dans des nombreux domaines de notre quotidien .Parmi ces domaines on a :

❖ **La Santé** : La science a réalisé des avancées très importantes grâce au Big Data. Ce dernier permet des avancements impressionnantes dans l'analyse génomique, séquençage humain, l'analyse des dossiers médicaux, elle permet aussi de mieux comprendre le développement de pathologies, d'améliorer des traitements, d'améliorer les protocoles de soins et les mesures de prévention [16].

❖ **Les marchés financiers :** L'analyse des transactions pour la gestion des risques et la gestion des fraudes. Ainsi que pour l'analyse des clients.

❖ **Les services publics :** L'analyse des compteurs (gaz, électricité, etc…) et la gestion des équipements.

❖ **La Télécommunication** : l'analyse de l'état du réseau en temps réel.

❖ **Le Marketing :** C'est l'ensemble du secteur qui est déterminée et renouvelé à travers l'analyse prédictive. par exemple, l'analyse morale qui peut être rapidement déterminée par l'interprétation automatique de l'opinion exprimée de l'individu.

❖ **La Recherche** : Le Big Data autorise le traitement rapide des multitudes de données dont le but d'accélérer le prototypage et la mise en œuvre des décisions fiables et rapides sur le marché.

Figure I.4. Domaines d'applications Big Data.

I.1.5. Enjeux du Big Data [16]

Le Big Data apparaît comme le challenge technologique des années 2010-2020 dépassant les domaines techniques et informatiques, le Big Data suscite un vif intérêt auprès des politiciens, des scientifiques et des entreprises. Les enjeux du Big Data touchent plusieurs secteurs d'activités :

I.1.5.1. Enjeux techniques

Les enjeux techniques s'articulent autour de l'intégration, le stockage, l'analyse, l'archivage, l'organisation et la protection des données.

I.1.5.2. Enjeux économiques

Les entreprises collectent de plus en plus d'information en relation avec leurs activités (production, stockage, logistique, ventes, clients, fournisseurs, partenaires, etc.), toutes ces informations peuvent être stockées et exploitées pour stimuler leur croissance.

Les Big Data permettent :

- D'améliorer les stratégies marketing et commerciale ;
- D'améliorer les stratégies marketing et commerciale ;
- De fidéliser la clientèle ;
- De gagner de nouvelles parts de marché ;
- De réduire les coûts logistiques ;
- De favoriser la veille concurrentielle.

Le client est un acteur majeur dans ce contexte. Jusqu'à présent, la vente consistait à se demander « J'ai un produit, à qui vais-je pouvoir le vendre? ». A l'ère du Big Data, nous devons changer le paradigme pour dire « J'ai un client, de quoi a-t-il besoin aujourd'hui ? ». En connaissant mieux son public, à travers ses achats, ses activités sur Internet, son environnement, les commerçants peuvent améliorer l'expérience-client, exploiter la recommandation, imaginer le marketing prédictif (le marketing prédictif regroupe les techniques de traitement et de modélisation des comportements clients qui permettent d'anticiper leurs actions futures à partir du comportement présent).

I.1.5.3. Enjeux juridiques

Le principal enjeu juridique dans un contexte où les utilisateurs sont souvent des « produits »reste la protection de la vie privée.

I.1.6. Les intérêt du Big Data [19]

L'utilisation des Big Data pourrait impacter fortement le monde de l'entreprise et ce de façon méliorative, ainsi les entreprises pourront :

- Améliorer la prise de décision ;
- Réduire les coûts d'infrastructures informatiques via l'utilisation des serveurs standards et des logiciels open source ;
- Développer la réactivation et l'interactivité à l'égard des clients ;
- Améliorer les performances opérationnelles.

I.1.7. Les challenges du Big Data [20]

L'augmentation du déluge de données dans l'ère Big Data apporte d'énormes défis sur l'acquisition de données, le stockage, la gestion et l'analyse. La gestion traditionnelle de données et les systèmes d'analyse sont basés sur le système de gestion de base de données relationnelle(SGBDR). Cependant, de tels SGBDRs appliquent uniquement aux données structurées. En outre, ces SGBDRs utilisent de plus en plus de matériels coûteux. Il est clair que les SGBDRs traditionnels ne pouvaient pas gérer l'énorme volume et l'hétérogénéité des Big data.

La communauté de recherche a proposé certaines solutions à partir de différentes perspectives. Par exemple, le ***cloudcomputing*** est utilisé pour répondre aux exigences en matière d'infrastructures pour les Big data.

Pour les solutions de stockage permanent et de gestion d'ensembles de données désordonnés à grande échelle, les systèmes de fichiers distribués DFS et les bases de données NoSQL (***Not Only SQL)*** sont des bons choix.

Certains ouvrages discutent les obstacles dans le développement des applications des Big data. Les principaux défis sont énumérés comme suit :

I.1.7.1. La représentation des données

Des nombreux ensembles de données ont un certain niveau d'hétérogénéité dans le type, la structure, la sémantique, l'organisation, la granularité, et l'accessibilité. La représentation des données vise à rendre les données plus significatives pour l'analyse de l'ordinateur et de l'interprétation de l'utilisateur. Néanmoins, une mauvaise représentation des données réduira la valeur des données d'origine et peut même entraver l'analyse efficace des données.

I.1.7.2. La réduction des redondances et la compression de données

Généralement, il y a un niveau élevé de redondance dans les ensembles de données. La réduction des redondances et la compression de données est efficace pour réduire le coût indirect de l'ensemble du système sur la prémisse que les valeurs potentielles des données ne sont pas affectés. Par exemple, la plupart des données générées par les réseaux de capteurs sont très redondantes, ce qui peut être filtré et comprimé à plusieurs ordres de grandeur.

I.1.7.3. La gestion du cycle de vie des données

Par rapport aux progrès relativement lents des systèmes de stockage, la détection omniprésente et informatique produisent des données à des taux et des échelles sans précédent. Nous sommes confrontés à une foule de défis pressants, dont l'une est que le système de stockage actuel ne pouvait pas soutenir de telles données massives. Généralement parlant, les valeurs cachées dans le Big data dépendent de l'actualisation des données.

I.1.7.4. Les mécanismes analytiques

Le système d'analyse du Big data doit traiter des masses de données hétérogènes dans un temps limité. Cependant, les SGBDRs traditionnels sont strictement conçus avec un manque d'évolutivité et d'extensibilité, ce qui pourrait ne pas répondre aux exigences de performance.

I.1.7.5. La confidentialité des données

La plupart des fournisseurs ou propriétaire de services de Big data à l'heure actuelle ne pouvait pas maintenir et analyser efficacement ces gigantesques ensembles de données en raison de leurs capacités limitées. Ils doivent compter sur des professionnels ou des outils pour analyser ces données, ce qui augmente les risques de sécurité potentiels.

Par exemple, les ensembles de données transactionnels comprennent généralement un ensemble de données d'exploitation complète pour guider des processus métier. Ces données contiennent des détails de la plus faible granularité et certaines informations sensibles tels que les numéros de la carte de crédit.

I.1.7.6. Extensibilité et évolutivité

Le système analytique du Big data doit prendre en charge les actuels et futures ensembles de données. L'algorithme analytique doit être capable de traiter les ensembles de données les plus complexes.

I.1.7.7. La coopération

L'analyse du Big Data est une recherche interdisciplinaire, qui exige que des experts dans différents domaines coopèrent pour récolter le potentiel des Big data. Une architecture globale du réseau de Big data doit être mise en place pour aider les scientifiques et les ingénieurs dans divers domaines d'accéder à différents types de données et d'utiliser pleinement leurs compétences, afin de coopérer pour terminer les objectifs analytiques.

Figure I.4. Challenges du Big Data.

I.3.Conclusion

Le Big data est un domaine qui évolue très rapidement, nous avons également abordé dans ce premier chapitre les principes du Big Data, ces caractéristiques, son fonctionnement, ses domaines d'applications, ses enjeux, et ses challenges, ainsi que les différents dans lesquels ils sont utilisées. Tandis que, dans le prochain chapitre, nous allons aborder et détailler toutes les méthodes de construction et de traitement des données massives.

Chapitre II

Construction et traitement du Big data

II. Introduction

De plus en plus d'éléments de notre quotidienne sont mis en données. Ce phénomène est complété par celui de la digitalisation qui transversalise notre économie grâce au développement du self-service. La BI (*Business Intelligence*) repose actuellement en grande partie sur la centralisation du savoir-faire de l'IT (*information technology*) en matière d'extraction, de transformation et de chargement des données. Elle ne suffit plus pour répondre à l'écrasante demande des entreprises en termes de données.

C'est dans ce contexte qu'est apparue la préparation des données. Il s'agit d'une nouvelle entité du système d'information décisionnel permettant de démocratiser la transformation des données. Pour se faire, elle permet aux experts métiers, maîtres connaisseurs de leurs données, de raffiner eux-mêmes les données dont ils ont besoin.

Dans ce chapitre, on va parler sur la préparation des données, comment on la faire ? Pourquoi ? Et quelles sont ses étapes ? Aussi on va voir la modélisation, la visualisation, l'intégrité et la sécurité des données.

II.1. Préparation des données

II.1.1. Définition 1

La préparation des données consiste à rassembler, combiner, structurer et organiser les données afin de pouvoir les analyser dans le cadre de programmes d'informatique décisionnelle (*Business Intelligence*) et d'analytique métier (*Business Analytics*). Ce processus comprend la découverte, le profilage, le nettoyage, la validation et la transformation des données; Il implique souvent d'assembler des données provenant de différents systèmes internes et externes. Dans les applications du Big Data, la préparation des données est généralement une tâche automatisée [21].

L'un des principaux objectifs de la préparation des données consiste à assurer que les informations concernées sont exactes et cohérentes, afin que les applications BI et BA donnent des résultats pertinents. En effet, les données sont souvent créées avec des valeurs manquantes, des inexactitudes ou d'autres erreurs. De plus, les ensembles de données sont souvent stockés dans des fichiers ou bases de données sous des formats différents, qui doivent donc être harmonisés. Le processus de correction des erreurs et de jointure des ensembles de données représente une large part de la préparation des données [21].

II.1.1.2. Définition 2

Le terme préparation des données désigne les opérations de nettoyage et transformation qui doivent être appliqués aux données brutes avant leur traitement et analyse. Il s'agit d'une étape importante avant le traitement proprement dit, qui implique souvent de reformater et corriger les données et de combiner des datasets pour enrichir certaines données [22].

La préparation des données est généralement une opération de longue haleine pour les spécialistes des données ou les utilisateurs de l'entreprise, mais il est essentiel de mettre les données en contexte pour pouvoir les convertir en connaissances exploitables et éliminer les biais résultant d'une mauvaise qualité des données [22].

II.1.2. Étapes de la préparation des données

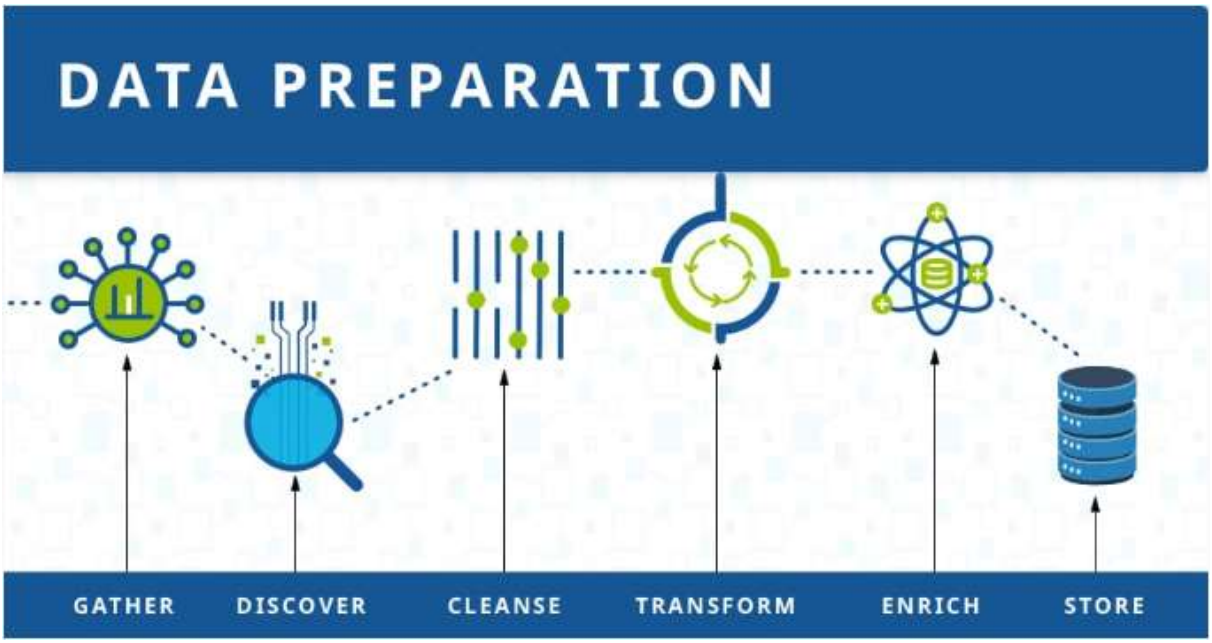


Figure II.1. Les étapes de la préparation des données [22].

II.1.2.1. Collecte de données

Le processus de préparation des données commence par la recherche des données les plus utiles. Ces données peuvent provenir d'un catalogue existant ou être ajoutées en mode ad hoc [22].

II.1.2.1.1. Définitions

II.1.2.1.1.1. Définition 1

La collecte de données est l'approche systématique permettant de collecter et de mesurer des informations provenant de diverses sources afin d'obtenir une image complète et précise d'un domaine d'intérêt. La collecte de données permet à une personne ou à une organisation de répondre à des questions pertinentes, d'évaluer les résultats et de prédire les probabilités et les tendances futures [23]. La composante de collecte de données de la recherche est commune à tous domaines d'études [24]. Une collecte de données précise est essentielle au maintien de l'intégrité de la recherche et à la prise de décisions [23].

II.1.2.1.1.2 Définition 2

La collecte de données est une compétence nécessaire pour tout individu. Les données sont utilisées dans diverses situations, telles que la rédaction de rapports de recherche à l'école, la recherche d'un élément spécifique ou l'obtention des informations nécessaires à un emploi. Quel que soit le motif des données, des outils similaires sont utilisés pour rechercher des informations et collecter des données [25].

II.1.2.1.2. Méthodes de collecte de données

Les méthodes principales de la collecte de données sont : Les enquêtes, les entretiens, les groupes de discussion, les observations et les livres. Aujourd'hui, avec l'aide d'outils Web et d'analyse, les organisations sont également en mesure de collecter des données à partir d'appareils mobiles, du trafic de sites Web, de l'activité des serveurs et d'autres sources pertinentes [23].

II.1.2.1.2.1. Enquête

Les enquêtes sont un excellent outil de collecte de données. Celles-ci sont utiles dans les entreprises, les études sur la santé mentale, les projets scolaire ou toute autre collecte de données nécessitant des informations d'un grand nombre de personnes. Les enquêtes posent des questions spécifiques qui sont remplies par des individus [25].

Une enquête est une méthode de recherche permettant de collecter des informations auprès d'un groupe de personnes appartenant à une population cible à l'aide de questionnaires standardisés ou d'interview [26].

Les enquêtes sont un bon moyen de collecter une grande quantité de données, offrant une perspective large. Les sondages peuvent être administrés par voie électronique, par téléphone, par courrier ou en face à face. Les enquêtes par courrier et par voie électronique ont une large portée, sont relativement peu coûteuses à administrer, les informations sont normalisées et la confidentialité peut être préservée [24].

II.1.2.1.2.2. Questionnaires

Le questionnaire est une méthode de collecte de données. Contrairement à l'entretien et à l'observation qui sont des méthodes individuelles ou collectives, le questionnaire est une méthode seulement collective. En effet, c'est la quantité d'éléments collectés qui confère au questionnaire sa validité et qui permet aux données d'être jugées authentique. La méthode du questionnaire repose sur une démarche mathématique purement rationnelle [27].

Les questionnaires présentent des avantages par rapport à d'autres types d'enquêtes car ils sont peu coûteux, ils ne nécessitent pas autant d'effort de la part du questionneur que les enquêtes verbales ou téléphoniques et ils ont souvent des réponses normalisées facilitant la compilation des données. En tant que type d'enquête, les questionnaires posent également nombre des problèmes liés à la construction de la question et à la formulation qui existent dans d'autres types de sondages d'opinion [24].

II.1.2.1.2.2.1. Les types du questionnaire

Il existe différents types de questionnaire pour récolter des informations claires et précises.

II.1.2.1.2.2.1.1. Le questionnaire ouvert

Dans ce type de questionnaire, l'ordre des questions et leur formulation sont fixés. Cependant, le participant peut s'exprimer aussi longtemps qu'il le souhaite. L'enquêté a la possibilité de le relancer [28].

II.1.2.1.2.2.1.2. Le questionnaire fermé

Dans ce type de questionnaire, les questions et la liste de propositions à soumettre au participant sont fixés à l'avance. Ceci afin de permettre au locuteur de faire le meilleur choix possible [28].

II.1.2.1.2.2.2. Avantages des questionnaires

Les avantages des questionnaires sont [24]:

- De grandes quantités d'informations peuvent être collectées auprès d'un grand nombre de personnes en peu de temps et de manière relativement rentable ;
- Peut être effectué par le chercheur ou par un nombre quelconque de personnes ayant une incidence limitée sur sa validité et sa fiabilité ;
- Les résultats des questionnaires peuvent généralement être rapidement et facilement quantifiés par un chercheur ;
- Peut être analysé de manière plus scientifique et objective que d'autres formes de recherche ;
- Lorsque les données ont été quantifiées, elles peuvent être utilisées pour comparer et contraster d'autres recherches et être utilisé pour mesurer le changement.

II.1.2.1.2.2.3. Inconvénients des questionnaires

Les inconvénients des questionnaires sont [24]:

- Il n'y a aucun moyen de dire à quel point un répondant est véridique.
- Le répondant peut être oublieux ou ne pas penser dans le contexte complet de la situation.
- Les gens peuvent lire différemment chaque question et donc répondre en fonction de leur propre interprétation de la question.

II.1.2.1.2.3. Entretien

L'entretien est une méthode de collecte de données utilisée pour obtenir des informations sur un sujet spécifique. Les entretiens sont généralement confiés à des experts dans un domaine spécifique. Les interviews sont souvent utilisées par les journalistes pour obtenir des informations de première main sur un article particulier [25].

Les entretiens consistent à poser des questions et à obtenir les réponses des participants à une étude. Les entretiens revêtent diverses formes: entretiens individuels, face à face et entretiens de groupe face à face. La question et la réponse aux questions peuvent être méditées par le téléphone ou d'autres appareils électroniques (ordinateurs, par exemple) [24].

II.1.2.1.2.3.1. Les type d'entretien

Il existe trois types d'entretien sont les suivants : Entretien directif, entretien semi directif et entretien non directif en vas détails chaque un d'eux :

II.1.2.1.2.3.1.1. Entretien directif

Ce type d'entretien se rapproche de la méthode du questionnaire. En effet, avant d'aller sur le terrain, le chercheur établit une série de questions précises qu'il va poser aux interviewés. Dans un souci de comparer scientifiquement les données, le chercheur va poser les mêmes questions à tous les interviewés. Certes, ce type d'entretien est sécurisant pour le chercheur. Ce dernier arrive avec une série de questions pré établies. Mais, il ne laisse qu'une petite large de manœuvre à l'enquêté. A cause des limites que lui pose l'enquêteur, l'enquêté n'aura pas une grande liberté pour s'exprime [29,30].

II.1.2.1.2.3.1.2. Entretien semi-directif

Il est préférable d'utiliser l'interview semi-directif si vous n'avez pas plus d'une chance d'interviewer quelqu'un et lorsque vous enverrez plusieurs intervieweurs sur le terrain pour collecter des données. Le guide d'interview semi-directif fournit un ensemble d'instructions claires aux intervieweurs et peut fournir des données qualitatives. Les entretiens semi-directifs sont souvent précédés d'observations, d'entretiens informels et non directif, afin de permettre aux chercheurs de bien comprendre le sujet d'intérêt nécessaire au développement de questions semi-structurées pertinentes et significatives. L'inclusion de questions ouvertes et la formation des enquêteurs aux sujets pertinents susceptibles de s'écarter du guide d'entretien permettent néanmoins d'identifier de nouvelles façons de voir et de comprendre le sujet à traiter [24,29].

II.1.2.2.1.3.1.3. Entretien non directif

L'entretien non directif est recommandé lorsque le chercheur a développé une compréhension suffisante du cadre et de son sujet d'intérêt pour avoir un ordre du jour clair pour la discussion avec l'informateur, mais reste ouvert à ce que sa compréhension du domaine d'enquête soit ouverte à la révision par les répondants. Parce que ces entretiens ne sont pas très structurés et que la compréhension du chercheur évolue toujours, il est utile d'anticiper le besoin de parler avec les informateurs à plusieurs reprises [24,30].

Il s'agit d'un type particulier d'entretien réalisé dans le but d'obtenir des informations dans un domaine particulier en rapport avec les forces et les capacités des participants. Bien que les entretiens servent généralement à obtenir des informations sur les candidats sans aucune influence de la part de l'enquêteur, les groupes de discussion permettent aux participants de partager leurs opinions, ce qui conduit parfois les gens à s'influencer et à mener des débats [31].

II.1.2.1.2.4. Groupes de discussion

Une discussion de groupe FGD (*Focus Group Discussion)* est une méthode de terrain approfondie qui rassemble un petit groupe homogène (généralement six à douze personnes) pour discuter de sujets inscrits à l'ordre du jour d'une étude. Le but de cette discussion est d'utiliser la dynamique sociale du groupe, avec l'aide d'un modérateur / animateur, pour inciter les participants à révéler leurs opinions, attitudes et motivations sous-jacentes. En bref, un groupe bien animé peut être utile pour déterminer le "comment" et le "pourquoi" du comportement humain [24].

II.1.2.1.2.5. Observation

Les observations sont souvent associées à la tâche d'enregistrer une communication individuelle entre des personnes. Ces processus nécessitent des observateurs bien formés et des instructions claires sur le processus d'observation, indiquant notamment qui doit observer et pendant combien de temps [24,31].

L'observation est un moyen fondamental de découvrir le monde qui nous entoure. En tant qu'êtres humains, nous sommes très bien équipés pour recueillir des informations détaillées sur notre environnement à travers nos sens. Cependant, en tant que méthode de collecte de données à des fins de recherche, l'observation est plus que juste regarder ou écouter. La recherche, définie simplement, est une «enquête systématique rendue publique» [24,31].

II.1.2.2. Découvrir et évaluer les données

Lorsque les données ont été collectées, il est important de découvrir les différents datasets. Cette étape permet de mieux connaître les données et de déterminer le traitement à leur appliquer avant qu'elles deviennent exploitables dans un contexte particulier [22].

L'étape de découverte est une tâche longue et complexe, mais la plate-forme *Talend* de préparation des données propose des outils de visualisation qui aident les utilisateurs à profiler et parcourir leurs données [22].

II.1.2.3. Nettoyer les données

Une fois les données collectées, nous avons généralement différentes sources de données avec des caractéristiques différentes. L'étape la plus urgente consiste à rendre ces sources de données homogènes et à développer davantage notre produit de données. Cependant, cela dépend du type de données. Nous devons nous demander s'il s'agit d'une homogénéité pratique des données [22].

En général, le nettoyage des données est l'étape la plus longue du processus de préparation des données, mais cette opération est cruciale pour éliminer les données erronées et combler d'éventuelles lacunes [22].

Lors du nettoyage, les tâches importantes sont notamment les suivantes [22] :

- Supprimer les données superflues et les valeurs aberrantes ;
- Ajouter les valeurs manquantes ;
- Adapter les données à une structure standard ;
- Masquer les données privées ou sensibles.

II.1.2.4. Transformer les données

Transformer les données consiste à mettre à jour les entrées de format ou de valeur de manière à obtenir un résultat clairement défini ou à rendre les données plus faciles à comprendre par un plus grand nombre d'employés. Ainsi que, il consiste à ajouter des données et à les relier à des données apparentées de manière à dégager des connaissances approfondies [22].

II.1.2.5. Stocker les données

Lorsque la préparation des données est terminée, celles-ci peuvent être stockées ou routées vers une application tierce avant leur traitement et analyse [22].

II.1.3. Pourquoi la préparation des données est-elle importante?

La préparation des données dicte les types d'analyses pouvant être effectuées à partir du système frontal de la solution d'analyse de données et la difficulté pour les utilisateurs finaux de répondre de manière simple aux questions de leur entreprise. En outre, une modélisation des données et des processus ETL *(Extract, Transform, Load)* efficaces pourraient avoir un impact majeur sur les performances globales de la solution de BI [32].

II. 1.4. Problèmes de la préparation des données

Quelques problèmes de la préparation des données [33] :

- 80 % du temps d'un processus (BI/Big Data) est dans la préparation des données ;
- Complexité de la construction d'un Data Warehouse ;
- Le développement prend un temps énorme (problème de la fraîcheur des données) ;
- Architecture Rigide (enrichissement, nouveau besoins) ;
- Intégration et alimentation complexe ;
- Faiblesse des outils de préparation classiques ;
- Pas de gouvernance de données ;
- Ne supportent pas le passage à l'échelle ;
- Découverte et transformation difficile.

II.1.5. Motivations et promesses de la préparation des données [33]

- L'analytique en self-service ;
- Gouvernance et collaboration ;
- Accélération de l'exploitation des données.

II.1.6. Avantages de la préparation des données

La préparation des données permet d'obtenir les résultats suivants [22] :

- Faciliter la détection des erreurs avant le traitement des données et les corriger rapidement **;**
- Obtenir des données de grande qualité **;**
- Prendre des décisions plus avisées, plus rapides, plus efficaces et de meilleure qualité.

Le mouvement de migration des données et processus vers le cloud n'épargne pas la préparation des données, qui en retire des avantages encore plus importants, en particulier [22] :

- Faciliter l'évolutivité **;**
- Pérenniser la solution **;**
- Accélérer l'utilisation des données et de la collaboration**.**

II.2. Modélisation de données

II.2.1. Définition

La modélisation de données fait référence à la formalisation et à la documentation de processus et d'événements qui se produisent au cours de la conception et du développement des applications. Les techniques et les outils de modélisation de données recueillent les conceptions de systèmes complexes et les traduisent en représentations simplifiées des processus et des flux de données de façon à créer un modèle pour la construction et la réingénierie [34].

Les modélisateurs utilisent souvent plusieurs modèles pour représenter les mêmes données et s'assurer que la totalité des processus, entités, relations et flux de données a été identifiée. La modélisation de données comprend plusieurs approches [34]**.**

II.2.2. Approches de modélisation de données

Voici les approches de modélisation de données :

II.2.2.1. La modélisation logique de données

Illustre spécifiquement des entités, attributs et relations impliqués dans une fonction métier. Elle sert de point de départ pour la création du modèle physique de données [35].

II.2.2.2. La modélisation physique de données

Représente l'implémentation propre à une application et à une base de données d'un modèle logique de données [35].

II.2.2.3. La modélisation de données d'entreprise

Est semblable à la modélisation conceptuelle de données, mais permet de répondre aux besoins spécifiques d'une entreprise [35].

II.2.2.4. La modélisation conceptuelle de données

Permet d'identifier le niveau de relations le plus élevé entre différentes entités [35].

II.3. Visualisation de données

II.3.1. Définitions

II.3.1.1 Définition 1

La visualisation des données est définie comme l'exploration visuelle et interactive de données de toutes volumétries, natures (structurées ou non structurées) et origines, et leur représentation graphique. Les visualisations aident à voir des choses qui n'étaient pas évidentes auparavant. Même quand le volume des données est très important, des tendances peuvent être perçues de façon rapide et simple. La visualisation facilite la transmission des informations de façon universelle et facilite le partage d'idées avec les autres [36].

C'est un moyen rapide d'obtenir des informations à travers l'exploration visuelle, des rapports fiables et un partage d'informations aisé. Toutes catégories d'utilisateurs peuvent ainsi donner un sens au nombre croissant de données de votre entreprise. Les Big data sont ainsi interprétés et utilisés rapidement par les utilisateurs métiers. La visualisation donne vie à vos données [36].

II.3.1.2. Définition 2

La visualisation de données est la présentation de données dans un format graphique. Il permet aux décideurs de voir les analyses présentées de manière visuelle, afin de pouvoir saisir des concepts difficiles ou d'identifier de nouveaux modèles. Avec la visualisation interactive, vous pouvez aller plus loin dans le concept en utilisant la technologie pour explorer en détail des graphiques, en modifiant de manière interactive les données que vous voyez et leur traitement [37].

La visualisation de données peut également [38]:

- Identifiez les domaines qui nécessitent une attention ou une amélioration;
- Clarifiez les facteurs qui influencent le comportement du client;
- Vous aider à comprendre quels produits placer où;
- Prédire les volumes de vente.

II.3.2. Pourquoi la visualisation des données est-elle importante?

La visualisation de données est un moyen efficace de partager universellement des concepts complexes [39] :

❖ **Les visuels sont plus efficaces que le texte :** Les graphiques et les photos transmettent des informations plus rapidement qu'un grand tableur ou un rapport dense.

❖ **Les métaphores visuelles sont un langage universel :** Les personnes de toutes les langues parlées ou écrites peuvent communiquer avec des métaphores visuelles.

❖ **Les données visualisées augmentent les connaissances :** Les données n'ont de valeur que lorsqu'elles sont traitées, analysées et mémorisées. La visualisation des données les rend encore plus utiles car elles sont plus faciles à consommer en tant qu'informations, qui deviennent à leur tour des connaissances.

II.3.3. Les approches de la visualisation des données

La visualisation est finalement un excellent moyen de laisser nos données parlent, voici les approches de la visualisation des données [39,40].

II.3.3.1. Comparer les données

Compare deux ou plusieurs valeurs.

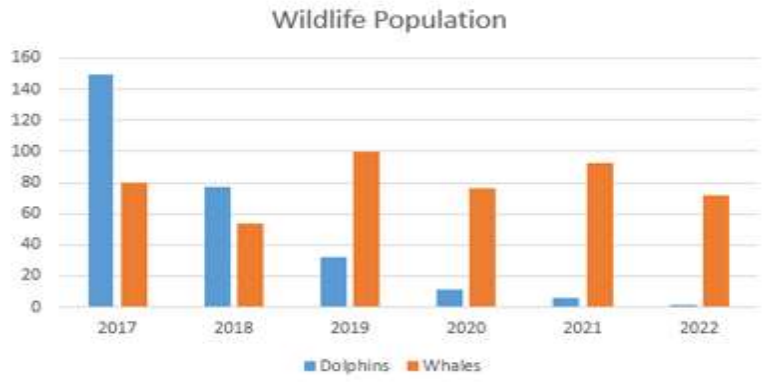


Figure II.2. Comparaison des données.

II.3.3.2. Explorer la composition et les relations partielles dans les données

Affiche les parties d'une seule unité.

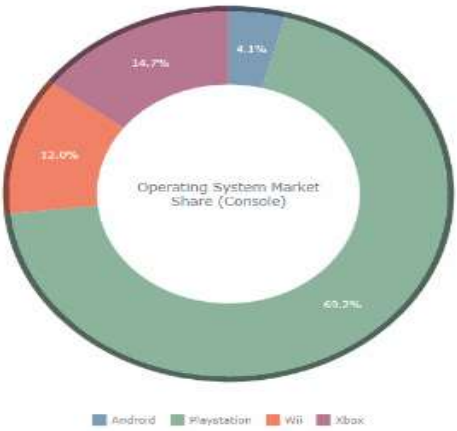


Figure II.3. Explorer la composition et les relations partielles dans les données.

II.3.3.3. Suivre les données au fil du temps

Montre comment quelque chose change avec le temps.

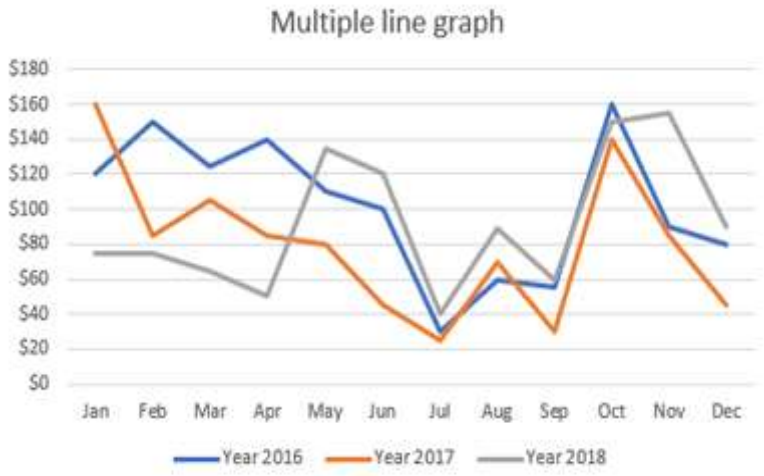


Figure II.4. Suivre les données au fil du temps.

II.3.3.4. Analyser la distribution des données

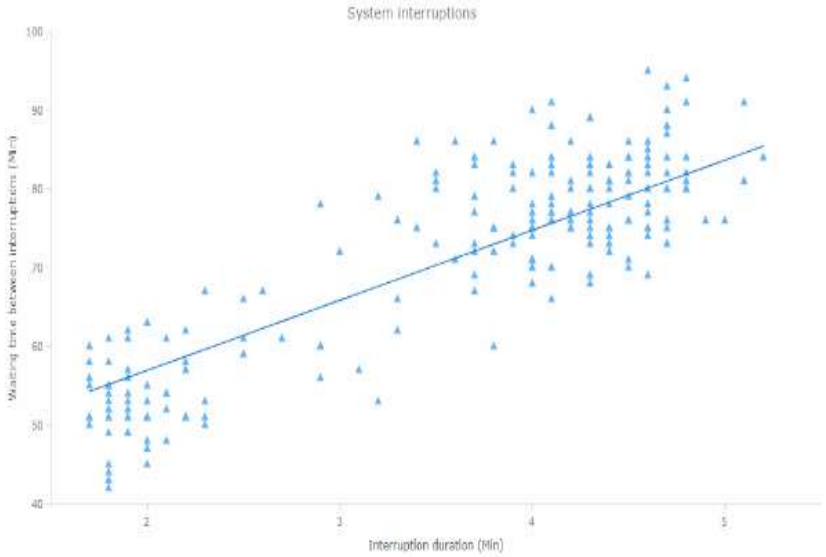


Figure II.5. Analyser la distribution des données.

II.3.3.5. Évaluer les données de performance actuelles

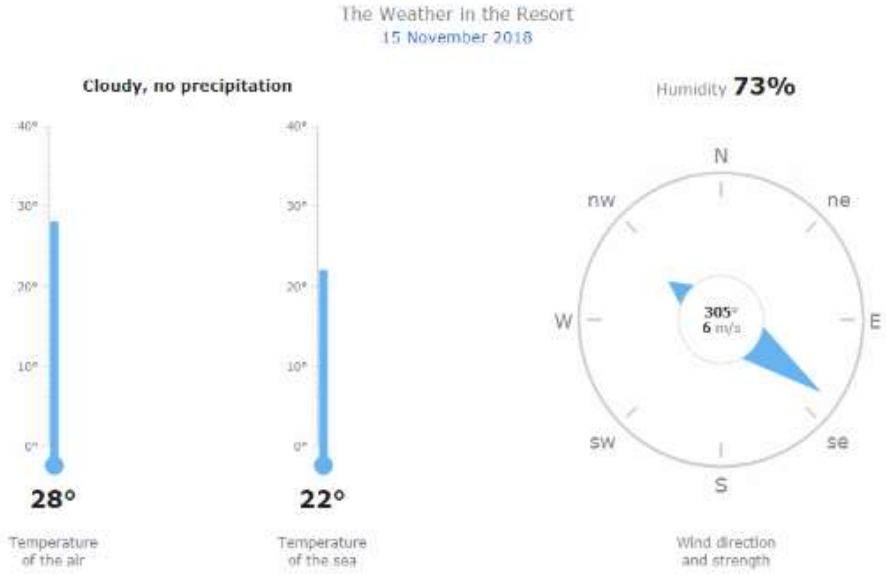


Figure II.6. Exemple d'évaluation les données de performance actuelles.

II.3.3.6. Examiner les données du projet

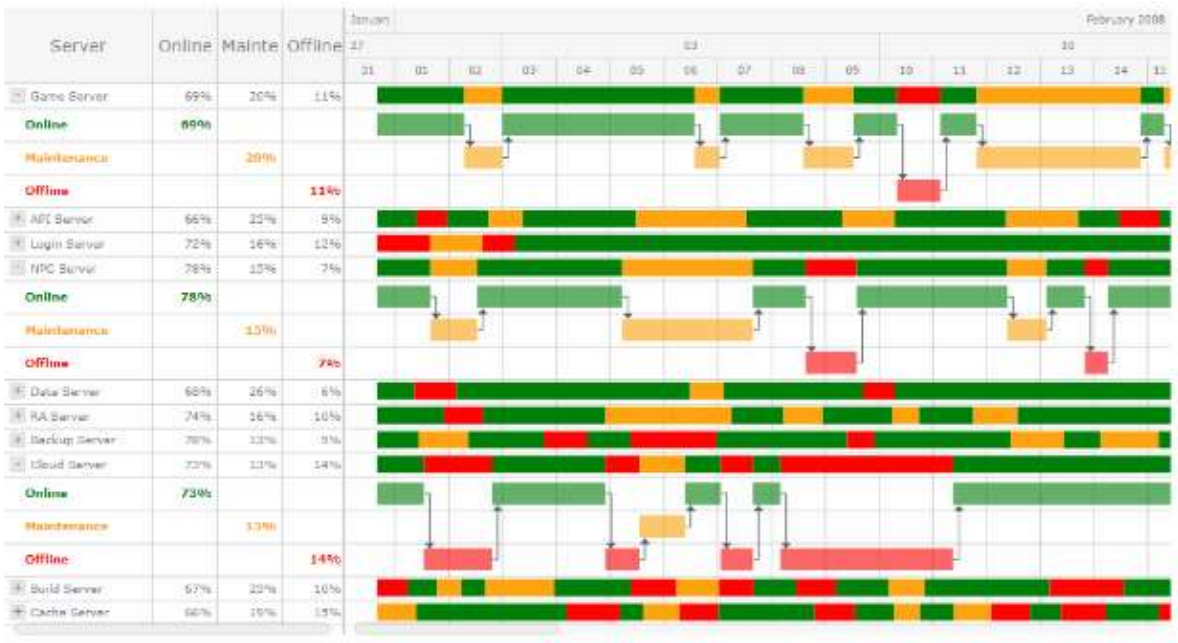


Figure II.7. Examination des données du projet.

II.3.3.7. Donner un sens aux données géographiques

Permet aux utilisateurs d'explorer des données, d'explorer et de trouver plus de détails dans un contexte géospatial.

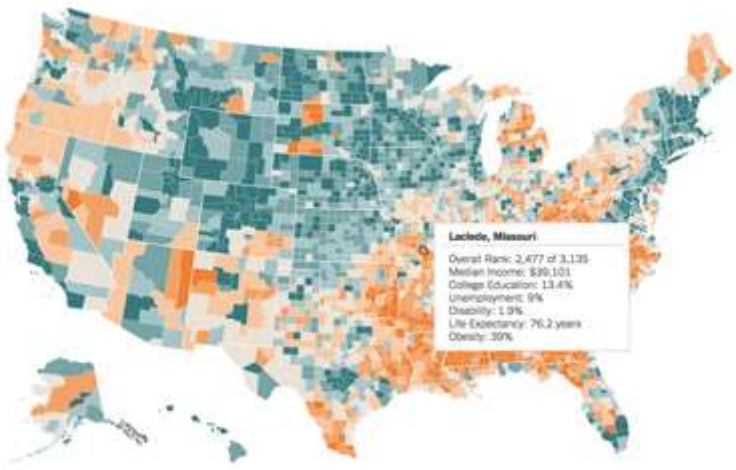


Figure II.8. Donner un sens aux données géographiques

II.3.4. Les avantages de la visualisation des données [36]

La visualisation des données permet de [36] :

- Prévoir les tendances du marché et de développer votre activité ;
- Connaître les dynamiques du marché comme jamais auparavant ;
- Cerner les exigences de chaque client et d'agir en conséquence ;
- Prendre les bonnes décisions au bon moment, et de partager l'information rapidement.

II.4. Sécurité et intégrité des données

II.4.1. Sécurité

II.4.1.1. Définition

La sécurité des données est un ensemble de normes et de technologies qui protègent les données contre la destruction, la modification ou la divulgation intentionnelle ou accidentelle. La sécurité des données peut être appliquée à l'aide d'un éventail de techniques et de technologies, notamment les contrôles administratifs, la sécurité physique, les contrôles logiques, les normes organisationnelles et d'autres techniques de protection qui limitent l'accès aux utilisateurs ou processus non autorisés ou malveillants [41].

II.4.1.2. Pourquoi la sécurité des données est-elle importante?

Aujourd'hui, toutes les entreprises traitent des données dans une certaine mesure. Des géants bancaires qui traitent d'énormes quantités de données personnelles et financières au personnel qui stocke les coordonnées de ses clients sur un téléphone mobile, les données sont en jeu dans les petites et grandes entreprises [41].

L'objectif principal de la sécurité des données est de protéger les données qu'une organisation collecte, stocke, crée, reçoit ou transmet. La conformité est également une considération majeure. Peu importe l'appareil, la technologie ou le processus utilisé pour gérer, stocker ou collecter des données, il doit être protégé. Les violations de données peuvent entraîner des litiges et de lourdes amendes, sans parler des dommages causés à la réputation d'une organisation. L'importance de protéger les données des menaces de sécurité est plus importante aujourd'hui qu'elle ne l'a jamais été [41].

II.4.1.3. Les technologies de sécurité des données

La technologie de sécurité des données se présente sous de nombreuses formes et protège les données d'un nombre croissant de menaces. Un grand nombre de ces menaces proviennent de sources externes, mais les organisations doivent également concentrer leurs efforts sur la sauvegarde de leurs données de l'intérieur [42].

Les moyens de sécuriser les données incluent [42] :

❖ **Cryptage des données :** le cryptage des données applique un code à chaque donnée et n'autorise pas l'accès aux données cryptées sans une clé autorisée.

❖ **Masquage des données :** le masquage de zones spécifiques de données peut empêcher leur divulgation à des sources malveillantes externes, mais également au personnel interne susceptible d'utiliser les données.

❖ **Effacement des données :** il arrive parfois que des données qui ne sont plus actives ou utilisées doivent être effacées de tous les systèmes.

❖ **Résilience des données :** en créant des copies de sauvegarde des données, les organisations peuvent récupérer des données si celles-ci sont effacées ou corrompues accidentellement, ou volées lors d'une violation de données.

II.4.1.4. les Avantages de la sécurité

La sécurité a une grande importance. Voici quelques avantages : [43]

❖ Protection contre les risques majeurs;

❖ Automatisation de l'ensemble des procédures de sauvegarde avec création de rapports journaliers ;

❖ Gestion simplifiée garantissant le bon déroulement des sauvegardes et leurs restitutions ;

❖ Garantir la sécurité informatique des entreprises dans le respect du cadre légal et juridique.

II.4.2. Intégrité des données

II.4.2.1. Définition

L'intégrité des données fait référence à la précision et à la cohérence (validité) des données tout au long de leur cycle de vie. Après tout, les données compromises sont peu utiles aux entreprises, sans parler des dangers présentés par la perte de données sensibles. Pour cette raison, le maintien de l'intégrité des données est au cœur des préoccupations de nombreuses solutions de sécurité d'entreprise [44].

L'intégrité des données peut être compromise de plusieurs manières. Chaque fois que des données sont répliquées ou transférées, elles doivent rester intactes et inchangées entre les mises à jour [44].

On se base généralement sur des méthodes de vérification des erreurs et des procédures de validation pour assurer l'intégrité des données transférées ou reproduites sans intention de les modifier [44].

L'intégrité des données signifie que les données doivent être fiables et précises tout au long de leur cycle de vie. L'intégrité des données et la sécurité des données vont de pair, même si ce sont des concepts distincts. Les données non corrompues (intégrité) sont considérées comme entières et restent ensuite inchangées par rapport à cet état complet [45].

II.4.2.2. Types d'intégrité des données

Il existe deux types d'intégrité des données: l'intégrité physique et l'intégrité logique. Les deux sont un ensemble de processus et de méthodes qui renforcent l'intégrité des données dans des bases de données hiérarchiques et relationnelles [46].

II.4.2.2.1. Intégrité physique

L'intégrité physique est la protection de la complétude et de la précision des données stockées et récupérées. L'intégrité physique est compromise lorsque des catastrophes naturelles se produisent, que des pannes se produisent ou que des pirates informatiques perturbent les fonctions de base de données. Les erreurs humaines, l'érosion du stockage et une foule d'autres problèmes peuvent également empêcher les responsables du traitement des données, les programmeurs système, les programmeurs d'applications et les auditeurs internes d'obtenir des données précises [46].

II.4.2.2.2. Intégrité logique

L'intégrité logique maintient les données inchangées, car elles sont utilisées de différentes manières dans une base de données relationnelle. L'intégrité logique protège les données contre les erreurs humaines et les pirates, mais de manière très différente de l'intégrité physique. Il existe quatre types d'intégrité logique [46].

II.4.3. Risques d'intégrité des données

Quelques risques d'intégrité des données [45] :

II.4.3.1. Erreur humaine

Lorsque des personnes saisissent des informations de manière incorrecte, dupliquent ou suppriment des données, ne respectent pas le protocole approprié ou ne commettent pas d'erreurs lors de la mise en œuvre de procédures destinées à protéger les informations, l'intégrité des données est mise en péril.

II.4.3.2. Erreurs de transfert

Une erreur de transfert s'est produite lorsque les données ne peuvent pas être transférées d'un emplacement de la base de données à un autre. Les erreurs de transfert se produisent lorsqu'un élément de données est présent dans la table de destination, mais pas dans la table source d'une base de données relationnelle.

II.4.3.3. Bogues et virus

Les logiciels espions, les logiciels malveillants et les virus sont des éléments de logiciel qui peuvent envahir un ordinateur et modifier, supprimer ou voler des données.

II.4.3.4. Matériel compromis

Des pannes soudaines d'ordinateur ou de serveur, et des problèmes de fonctionnement d'un ordinateur ou d'un autre périphérique, sont des exemples de pannes importantes qui peuvent indiquer que votre matériel est compromis. Le matériel compromis peut rendre les données de manière incorrecte ou incomplète, limiter ou éliminer l'accès aux données ou rendre les informations difficiles à utiliser.

II.4.3. Comment minimiser ou éliminer les risques d'intégrité des données ?

Les risques pour l'intégrité des données peuvent facilement être minimisés ou éliminés en procédant comme suit [46] :

- Limitation de l'accès aux données et modification des autorisations pour limiter les modifications d'informations par des tiers non autorisés ;
- Valider les données pour s'assurer qu'elles sont correctes à la fois quand elles sont rassemblées et utilisées ;
- Sauvegarde des données ;

- ❖ Utilisation de journaux pour garder une trace du moment où des données sont ajoutées, modifiées ou supprimées ;
- ❖ Réalisation d'audits internes réguliers ;
- ❖ Utilisation d'un logiciel de détection d'erreur.

II.4.5. Intégrité des données VS Sécurité des données

L'intégrité des données et la sécurité des données sont des termes liés, chacun jouant un rôle important dans la réussite de l'autre. La sécurité des données fait référence à la protection des données contre les accès non autorisés ou la corruption et est nécessaire pour assurer l'intégrité des données [44].

Cela dit, l'intégrité des données est un résultat souhaité de la sécurité des données, mais le terme intégrité des données ne fait référence qu'à la validité et à l'exactitude des données plutôt qu'à l'acte de protection des données. En d'autres termes, la sécurité des données est l'une des mesures pouvant être utilisées pour maintenir l'intégrité des données. Qu'il s'agisse d'une intention malveillante ou d'une compromission accidentelle, la sécurité des données joue un rôle important dans le maintien de l'intégrité des données [44].

Pour les entreprises modernes, l'intégrité des données est essentielle à la précision et à l'efficacité des processus métiers ainsi que du processus décisionnel. C'est également un élément central de nombreux programmes de sécurité des données. Grâce à diverses méthodes de protection des données, notamment la sauvegarde et la réplication, les contraintes d'intégrité de la base de données, les processus de validation, ainsi que d'autres systèmes et protocoles, l'intégrité des données est essentielle, mais gérable pour les entreprises [44].

II.5. Conclusion

Le taux des données ne cesse de croitre d'un jour à un autre. Avec cette croissance, nos bases de données traditionnelles sont limitées face à l'analyse et au traitement de ces données. Dans un souci de gain de temps, de nouvelle technologie sont venu pour soulager les entreprises génératrices d'un grand nombre de données. L'analyse de Big Data est sans aucun doute vouée à gagner une importance, certains parlent même de révolution. Et d'un autre côté, l'analyse des données massives comporte des risques liés au respect de la vie privée, à la confidentialité, au libre-arbitre, auxquels il convient de réfléchir dès maintenant.

Chapitre III

Big data analytics, la statistique mathématique et Machine Learning

III .Introduction

Il n'y a pas très longtemps, on ne pouvait pas traiter un tableau avec une taille énorme. L'apparition et le développement dans le domaine d'informatique a du coup levé cet obstacle de calcul, et a permis la conservation et l'exploitation des grandes masses de données. Cette amélioration continue de l'outil informatique a fortement contribué au développement et à la vulgarisation de nombreuses méthodes statistiques, devenues maintenant d'usage assez courant**.** Parmi ces méthodes on a l'analyse de donnée.

L'analyse de données est un ensemble de méthodes mathématiques qui permettent de traiter des "items" dans un tableau. Ces items sont décrits par un ensemble de variables.

L'analyse de données a pour l'objectif de traiter des informations du type : quels sont les items identiques ou dissemblables, quelles sont les relations entre les items et les variables associées.

Parfois, l'analyse de donnée devenue difficile devant le nombre imaginaire des données. Dans ce cas, la meilleure solution est de prendre un échantillon pour faciliter l'opération et ce qu'on l'appelle l'échantillonnage. Et pour faciliter l'échantillonnage beaucoup plus, on a permis au ordinateurs de faire ce l'être humain peut le faire.

Dans ce chapitre, on va voir l'analyse de données, l'échantillonnage et le Machine learning en détails.

III.1. Analyse de données

III.1.1. Définitions

III.1.1.1. Définition 1

On appelle l'ensemble de méthodes statistiques qui sont appliquées à un jeu de données dont son but est de détacher des informations pertinentes : l'analyse de données, cette extraction est la fouille de données ; Son but est de révéler des comportements et d'extraire des directions, de trouver des liens ou des règles [47].

III.1.1.2. Définition 2

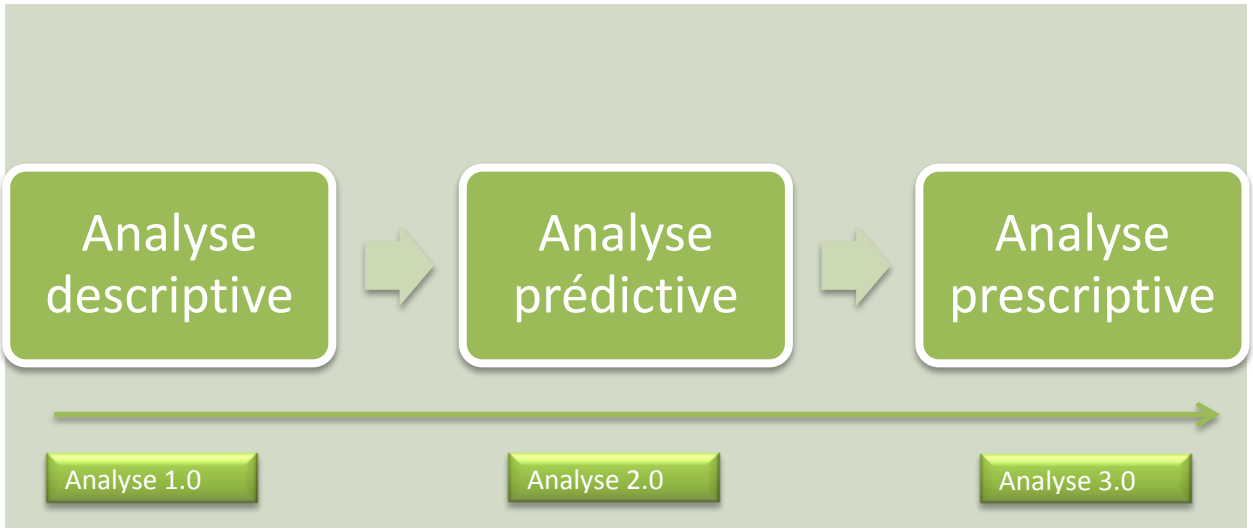
L'analyse de données est un domaine spécifique des statistiques mais communément employée comme expression pour décrire l'ensemble des techniques utilisées pour collecter, décrire, analyser, synthétiser, comprendre un ensemble de données, souvent immensément grand [47].

Depuis la création et la multiplication de bases de données, mais aussi la nécessité pour les entreprises de réaliser une veille permanente de la satisfaction de ses clients, de la satisfaction et la performance de ses salariés, de ses processus internes…etc, réaliser des enquêtes, des études et analyser ses données est devenu un enjeu majeur du décisionnel dans les entreprises à tous les niveaux [48].

III.1.2. Types d'analyse de données

Il existe trois grands types d'analyse de données : l'analyse descriptive, l'analyse prédictive et l'analyse perspective [48,49,50,51].

Ce que nous concerne c'est l'analyse prédictive.



FigureIII.1. Les types de l'analyse de données.

III.2.1. L'analyse descriptive

Est un ensemble des calculs mathématiques simples permettent de dégager des données une réelle tendance des résultats que ce soit positive ou négative [47].

L'objectif d'analyse descriptive est d'assigner une nouvelle représentation pour les données qu'elle a déjà les résumé, de synthétiser pour ressortir ce qui le volume a dissimulé. On peut classer les individus dans des catégories, trouver les individus les plus proches ou les plus éloignés entre eux ; mais aussi trouver les exceptions ou les cas atypiques. On peut également voir si des variables sont proches, expliquer une variable en fonction des autres ou encore repérer les variables les plus influentes [48]. Avant d'approfondir l'analyse dans les détails, il faut commencer par la description globale en basant sur les statistiques descriptives. Donc, les statistiques descriptives sont la base de toute analyse de données [47].

III.2.2. L'analyse prédictive

III.2.2.1. Définitions

III.2.2.1.1. Définition 1

L'analyse prédictive consiste à analyser les données courantes afin de faire des hypothèses sur des comportements futurs. On se sert des données que l'on possède déjà pour extrapoler et deviner le comportement de nouveaux individus mais également l'évolution des individus déjà présents [48].

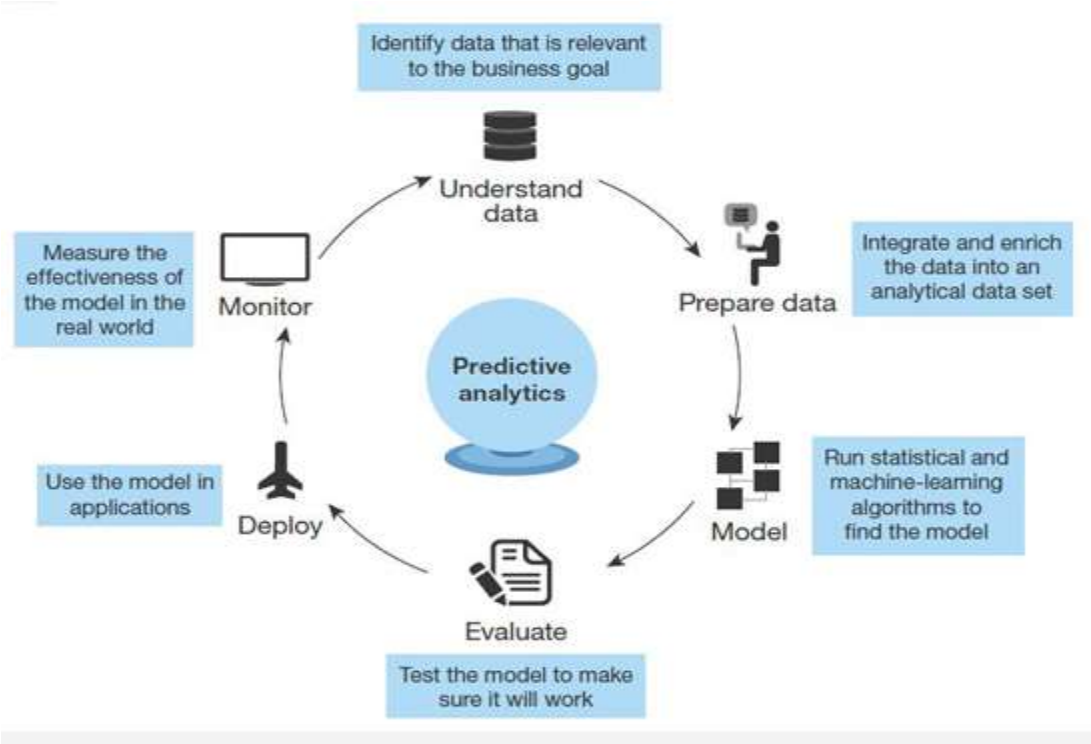
III.2.2.1.2. Définition 2

Les analyses prédictives sont le résultat pratique du Big Data et de la *Business Intelligence* (BI). Elles permettent d'exploiter les énormes quantités de données qui sont collectées par de nombreuses entreprises auprès de leurs clients, de leurs marchés, des réseaux sociaux, des applications en temps réel, ou encore du Cloud. En dégageant des informations tangibles, les analyses prédictives permettent de rester en tête de la compétition [49].

III.2.2.2. Les six étapes clés de l'analyse prédictive [50]

L'étude met en exergue un cycle de six étapes clés dans l'élaboration de solutions prédictives grâce au Big data :

- Identifier les données avantageuses en évaluant diverses sources possibles ;
- Triturer les *data*, les agréger, les compléter, etc ;
- Construire un modèle prédictif, à partir d'algorithmes statistiques et du *Machine Learning* ;
- Evaluer l'efficacité et la précision du modèle prédictif ;
- Utiliser le modèle prédictif pour orienter des décisions métiers ;
- Assurer un suivi de l'application et de l'efficacité du modèle prédictif.



FigureIII.2. Les six étapes clés de l'analyse prédictive [50].

III.2.3. L'analyse prescriptive

Une fois que vous arrivez au point ou pouvez constamment analyser vos données pour prédire ce qui va se passer, vous êtes très proche de pouvoir comprendre ce que vous devez faire afin de maximiser les bons résultats et également prévenir les mauvais résultats potentiels [51].

III.3.Echantillonnage

III.3.1. Définitions

III.3.1.1. Définition 1

L'échantillonnage est le procédé utilisé pour choisir un échantillon et qui est à la base de l'enquête par sondage [52].

III.3.1.2. Définition 2

L'échantillonnage est la phase qui consiste à sélectionner les individus que l'on souhaite interroger au sein de la population de base. Les résultats obtenus sur l'échantillon sont ensuite extrapolés à la population que l'on souhaite étudier. Le plus souvent l'échantillon est prélevé de manière aléatoire [53].

III.3.1.3. Définition

L'échantillonnage est le processus de sélection des unités d'une population d'intérêt afin qu'en étudiant l'échantillon, nous généralisions assez bien nos résultats à la population à partir de laquelle ils ont été choisis. De plus, l'échantillonnage peut être le meilleur choix si un modèle prédictif est besoin d'être construit, différent modèles peuvent être fabrique à partir de différents échantillon de grande données [54].

III.3.2. Les avantages de l'échantillonnage

La technique d'échantillonnage permet une réponse plus exacte avec [55]:

- Un moindre cout ;
- Une grande rapidité ;
- Moins complexe de prendre un échantillon ;
- Une précision des résultats.

III.3.3. Les méthodes l'échantillonnage

Les méthodes d'échantillonnage correspondent aux différentes manières de constituer un échantillon de la population étudiée [54].

III.3.3.1. Les méthodes probabilistes ou aléatoires

Voici les méthodes d'échantillonnage probabiliste les plus courantes :

III.3.3.1.1. l'échantillonnage aléatoire simple

L'échantillonnage aléatoire simple est la méthode d'échantillonnage la plus facile à appliquer et la plus communément utilisée [56].

III.3.3.1.1.1. Les étapes de l'échantillonnage aléatoire simple [52,56]

1- On se procure une liste de toutes les unités statistique de la population et on les numérote de 1 à N ;
2- 2- On choisit au hasard « n» nombre déférents correspondant aux « n » unité statistique faire partie de l'échantillon.

III.3.3.1.1.2. L'avantage

- On peut espérer un échantillon « représentatif » puisque la méthode donne à chaque individu de la population une chance égale [55].

III.3.3.1.1.3. l'inconvénient

❖ La méthode n'est applicable que lorsqu'il existe une liste exhaustive de toute la population [55].

III.3.3.1.2. l'échantillonnage systématique

Appelé aussi l'échantillonnage par intervalles ; Il signifie qu'il existe un écart, ou un intervalle, entre chaque unité sélectionnée qui est incluse dans l'échantillon [56].

III.3.3.1.2.1.Les étapes pour sélectionner un échantillon systématique

1. Numéroter de 1 à N les unités incluses dans votre base de sondage (où N est la taille de la population totale) [52];

2. Déterminer l'intervalle d'échantillonnage (K) en divisant le nombre d'unités incluses dans la population par la taille de l'échantillon que vous désirez obtenir ;

3. Sélectionner au hasard un nombre entre 1 et K. Ce nombre s'appelle l'origine choisie au hasard et serait le premier nombre inclus dans votre échantillon ;

4. Sélectionner chaque Ke unité après ce premier nombre.

III.3.3.1.2.2. L'avantage

❖ Parce qu'il y a un seul individu qui va être choisi au hasard, l'échantillon systématique est facile à sélectionner [55].

III.3.3.1.2.3. l'inconvénient

❖ À cause de la périodicité, les données peuvent être biaises [55].

III.3.3.1.3. l'échantillonnage avec probabilité proportionnelle à la taille

Pour l'échantillonnage probabiliste, il faut que chaque membre de la population observée ait une chance d'être inclus dans l'échantillon, mais il n'est pas nécessaire que cette chance soit la même pour tous. Si la base de sondage renferme de l'information sur la taille de chaque unité et si la taille de ces unités varie, on peut utiliser cette information dans le cadre de la sélection de l'échantillonnage afin d'en accroître l'efficacité. Cela s'appelle l'échantillonnage avec probabilité proportionnelle à la taille (PPT). Dans le cas de cette méthode, plus la taille de l'unité est grande, plus sa chance d'être incluse dans l'échantillon est élevée.

Il faut que la mesure de la taille soit exacte pour que cette méthode accroisse l'efficacité. C'est une méthode d'échantillonnage plus complexe dont nous ne traiterons pas ici davantage [56].

III.3.3.1.4. l'échantillonnage stratifié

Lorsqu'on utilise l'échantillonnage stratifié, on divise la population en groupes homogènes (appelés strates), qui sont mutuellement exclusifs, puis on sélectionne à partir de chaque strate des échantillons indépendants. On peut utiliser n'importe quelle des méthodes d'échantillonnage mentionnées dans la présente section pour sélectionner l'échantillon à l'intérieur de chaque strate [57].

La méthode d'échantillonnage peut varier d'une strate à une autre. Lorsqu'on utilise l'échantillonnage aléatoire simple pour sélectionner l'échantillon à l'intérieur de chaque strate, on appelle le plan d'échantillonnage un plan d'échantillonnage aléatoire simple stratifié. On peut stratifier avant l'échantillonnage une population au moyen de toute variable dont on dispose pour la totalité des unités incluses dans la base de sondage (comme l'âge, le sexe, la province de résidence, le revenu, etc.) [57].

III.3.3.1.4.1. Pourquoi doit-on créer des strates?

Pour bien des raisons, la principale étant que leur création peut rendre la stratégie d'échantillonnage plus efficace. Nous avons mentionné précédemment que vous aviez besoin d'un échantillon plus grand pour obtenir une estimation plus exacte d'une caractéristique qui varie beaucoup d'une unité à l'autre [56].

III.3.3.1.4.2. Les étapes de l'échantillonnage stratifié [58]

1- On se procure une liste de toutes les unités statistiques de la population ;

2- On sépare la population en déférents strates de manière à ce que chaque strate regroupe les individus de la population possédant une caractéristique commune qui pourrait avoir une influence sur les résultats de l'étude, on numérote ensuite les individu dans chaque strate ;

3- Déterminer la population de chaque strate dans la population ;

4- Déterminer la taille 'n' de l'échantillon voulu.

5- On détermine le nombre d'individu qu'il faudra dans chaque strate de l'échantillon ;

6- On sélectionne le nombre d'individus voulu dans chaque strate de la population par échantillonnage aléatoire simple.

III.3.3.1.4.3. Les avantages [55]

❖ Il est peu probable de choisir un échantillon absurde puisqu'on s'assure de la présence proportionnelle de tous les divers sous-groupes composant la population ;

❖ Cette méthode permet d'obtenir un échantillon représentatif car tous les individus d'un groupe ont la même probabilité de faire partie du sous-échantillon et l'échantillon obtenu est représentatif de la population en ce qui concerne la variable d'intérêt [59].

III.3.3.1.5. L'échantillonnage en grappes

Échantillonnage en grappes est parfois trop dispendieux de disséminer un échantillon dans l'ensemble de la population. Les coûts de déplacement risquent de devenir élevés lorsque les intervieweurs doivent sonder des gens d'un bout à l'autre du pays. Les statisticiens peuvent choisir la technique de l'échantillonnage en grappes pour réduire les coûts. La technique de l'échantillonnage en grappes entraîne la division de la population en groupes ou en grappes comme son nom l'indique. Suivant cette technique, on sélectionne au hasard un certain nombre de grappes pour représenter la population totale, puis on englobe dans l'échantillon toutes les unités incluses à l'intérieur des grappes sélectionnées. On n'inclut dans l'échantillon aucune unité de grappes non sélectionnées; ces unités sont représentées par celles tirées de grappes sélectionnées. La technique en question diffère de la technique d'échantillonnage stratifié, qui entraîne la sélection d'unités de chaque groupe [56].

III.3.3.1.5.1. Les étapes de l'échantillonnage par grappes [58]

1- On sépare la population en grappes hétérogène de taille semblable qu'on numérote ;

2- On détermine la taille « n » de l'échantillon voulu ;

3- On calcule le nombre de grappes qu'il faudra choisir pour constitue l'échantillon ;

nbr G = n /nombre moyenne d'individu par grappes

4- On choisit le nombre de grappes nécessaire par la méthode d'échantillonnage aléatoire simple.

III.3.3.1.5.1.2. Les avantages [59]

❖ La méthode ne nécessite pas une liste globale de la population puisque seuls les individus inclus dans les grappes comptent ;

❖ Elle permet de limiter l'échantillon à des groupes compacts ce qui permet de réduire les couts de déplacement, de suivi et de supervision.

III.3.3.1.5.1.3. Les inconvénients [59]

❖ La méthode peut entrainer des résultats imprécis (moins précis que les méthodes précédentes) puisque les unités voisinés ont tendance se rassembler ;

❖ Elle ne permet pas de contrôler la taille finale de l'échantillon.

III.3.3.1.6. L'échantillonnage à plusieurs degrés

La méthode d'échantillonnage à plusieurs degrés ressemble à la méthode d'échantillonnage en grappes, sauf qu'il faut dans son cas prélever un échantillon à l'intérieur de chaque grappe sélectionnée, plutôt que d'inclure toutes les unités dans la grappe. Ce type d'échantillonnage exige au moins deux degrés. On identifie et sélectionne au premier degré de grands groupes ou de grandes grappes. Ces grappes renferment plus d'unités de la population qu'il n'en faut pour l'échantillon final. Pour obtenir un échantillon final, on prélève au second degré des unités de la population à partir des grappes sélectionnées (à l'aide de l'une des méthodes d'échantillonnage probabiliste possibles). Si l'on utilise plus de deux degrés, le processus de sélection d'unités de la population à l'intérieur des grappes se poursuit jusqu'à l'obtention d'un échantillon final [52,56].

III.3.3.1.6.1. Les étapes de l'échantillonnage à plusieurs degrés [58]

L'échantillonnage à plusieurs degrés fonctionne comme suite :

1- On commence par construire des groupes d'individus qui soient disjoint, et dont la réunion soit la population toute entière ;

2- On consiste ainsi une partition de la population. On tire alors, par exemple par SAS ou toute autre méthode un certain nombre de groupes dans la base de sondage de groupes qui a été constituée.

Chaque group est donc un individu (à ce stade on parle d'unités primaires « UP»).

3- Ayant obtenu notre échantillon d'UP, et considérant ces UP les uns après les autres, on tire des individus dans chaque UP, par SAS par exemple (à ce stade les individus tirés au sien des UP sont appeler unités secondaires « US »), et ainsi de suite les tirages sont faites.

III.3.3.1.6.2. Les avantages [55]

- ❖ Pas de besoin de disposer de la liste de toutes les unités ;
- ❖ La méthode permet de contrôler la taille de l'échantillon notamment par stratification.

III.3.3.1.6.3. L'inconvénient [55]

- Précision des résultats est faible.

III.3.3.1.7. l'échantillonnage à plusieurs phases

Un échantillonnage à plusieurs phases entraîne la collecte de données de base auprès d'un échantillon d'unités de grande taille et ensuite, pour un sous-échantillon de ces unités, la collecte de données plus détaillées. La forme la plus courante d'échantillonnage à plusieurs phases est l'échantillonnage à deux phases (ou l'échantillonnage double), mais il est également possible d'effectuer un échantillonnage à trois phases ou plus. L'échantillonnage à plusieurs phases est assez différent de l'échantillonnage à plusieurs degrés, malgré les similarités entre eux sur le plan de leur appellation. Même si l'échantillonnage à plusieurs phases suppose aussi le prélèvement de deux échantillons ou plus, dans son cas, tous les échantillons sont tirés de la même base de sondage et les unités sont structurellement les mêmes à chaque phase. Comme dans le cas de l'échantillonnage à plusieurs degrés, plus l'on utilisera de phases, plus le plan d'échantillonnage et l'estimation deviendront complexes. L'échantillonnage à plusieurs phases est utile lorsqu'il manque à l'intérieur de la base de sondage des données auxiliaires qui pourraient servir à stratifier la population ou à rejeter à la sélection une partie de la population[56].

III.3.3.2. Échantillonnage non probabiliste (Non aléatoire)

L'échantillonnage non probabiliste repose sur un choix arbitraire des unités, c'est l'enquêteur qui choisit les unités et non le hasard. En ce sens, il serait donc aventureux de généraliser les résultats obtenus pour l'échantillon à toute la population. Malgré cela, ces méthodes sont souvent utilisées dans certaines disciplines [52].Voici les méthodes d'échantillonnage non probabiliste les plus courantes :

III.3.3.2.1. L'échantillonnage volontaire

Comme l'expression le laisse entendre, ce type d'échantillonnage intervient lorsque des gens offrent volontairement leurs services pour l'étude dont il est question. Souvent, à l'occasion des sondages d'opinion, seuls les gens qui se soucient assez fortement d'une façon ou d'une autre de la question étudiée ont tendance à y répondre. La majorité silencieuse n'y répond généralement pas, ce qui entraîne un important biais sur le plan de la sélection. Cette technique consiste à faire appel à des volontaires pour constituer l'échantillon [56].

III.3.3.2.2. L'échantillonnage par quotas

L'échantillonnage par quotas est l'une des formes les plus courantes d'échantillonnage non probabiliste. Il s'effectue jusqu'à ce qu'un nombre précis d'unités (de quotas) pour diverses sous populations ait été sélectionné. Puisqu'il n'existe aucune règle qui régirait la façon dont il faudrait s'y prendre pour remplir ces quotas, l'échantillonnage par quotas est réellement un moyen de satisfaire aux objectifs en matière de taille d'échantillon pour certaines sous-populations [52].

L'échantillonnage par quotas est un peu similaire à l'échantillonnage stratifié parce que dans son cas également les unités semblables sont regroupées. Toutefois, il en diffère, cependant, sur le plan du mode de sélection. Il camoufle toutefois des biais pouvant être significatifs [52].

III.3.3.2.2.1.Les étapes de l'échantillonnage par quotas

Les individus de l'échantillon ne sont pas choisir au hasard et comporte trois étapes[52]:

1- description la structure de la population selon des critères choisis au préalable appelées variable de contrôle ;
2- construction d'une maquette de la population à partir des mesures prises précédemment. (choix des variables de contrôle) ;
3- chaque enquêteur voit attribué des quotas qu'il doit réaliser, on connait donc le nombre de personne à interroger à l'aide d'une feuille de quotas.

III.3.3.2.2.2. Les avantages [56]

- L'échantillonnage par quotas est préférable à d'autres formes d'échantillonnage non probabiliste (comme l'échantillonnage au jugé), parce qu'il impose l'inclusion dans l'échantillon de membres de différentes sous-populations ;
- peu coûteux ;
- facile à administrer a la propriété souhaitable de respecter les proportions de la population.

III.3.3.2.3. L'échantillonnage de commodité (ou à l'aveuglette)

On appelle parfois l'échantillonnage de commodité l'échantillonnage à l'aveuglette ou accidentel. Cet échantillonnage n'est pas normalement représentatif de la population cible, parce qu'on ne sélectionne des unités d'échantillonnage dans son cas que si on peut y avoir facilement et commodément accès [52,56].

III.3.3.2.3.1. L'avantage

- Même si ses applications utiles sont limitées, la technique peut donner des résultats exacts lorsque la population est homogène [55].

III.3.3.2.3.2. L'inconvénient

- Malgré que cette méthode est facile à utiliser, mais la présence de biais annule énormément ce dernier [55].

III.3.3.2.4. Echantillonnage de convenance (de commodité)

Cas où les unités d'échantillonnage sont faciles à rejoindre, disponibles et généralement facile à convaincre. Le chercheur constitue son échantillon en choisissant les individus (appelés aussi unités statistiques) disponibles pour des raisons pratiques d'accessibilité et de coût et non de manière aléatoire. L'échantillon est obtenu sans méthode particulière [56].

III.3.3.2.5. Echantillonnage selon le jugement

On utilise la méthode d'échantillonnage au jugé lorsqu'on prélève un échantillon en se fondant sur certains jugements au sujet de l'ensemble de la population. Le chercheur juge que l'échantillon va lui permettre d'atteindre les objectifs de la recherche [56].

III.3.3.2.6. Echantillonnage boule de neige

Utile dans le cas de la rareté des unités d'échantillonnage ou de l'absence d'un cadre d'échantillonnage valide. On demande à un répondant de nous référer à un autre qui présente les mêmes caractéristiques que les siennes, et ainsi de suite [52].

On appelle parfois l'échantillonnage de commodité l'échantillonnage à l'aveuglette ou accidentel. Cet échantillonnage n'est pas normalement représentatif de la population cible, parce qu'on ne sélectionne des unités d'échantillonnage dans son cas que si on peut y avoir facilement et commodément accès [52].

III.4. Machine Learning

III.4. 1. Définitions

III.4.1.1. Définition 1

D'après [60], l'intelligence artificielle est un domaine très large, un de ses sous-domaines est L'apprentissage automatique (ou Machine Learning).

Machine Learning permet aux ordinateurs d'apprendre sans avoir été programmés explicitement à cet effet. Dont l'objectif de comprendre la structure des données et de les intégrer dans des modèles qui peuvent être compris et utilisés par les tout le monde [61].

III.4.1.2. Définition 2

L'apprentissage automatique est un domaine qui s'intéresse à comprendre et reproduire l'apprentissage humain à partir des systèmes artificiels. Il s'agit, très schématiquement, de concevoir des algorithmes et des méthodes permettant d'extraire l'infom1ation pertinente de données, ou d'apprendre des comportements à partir d'exemples. Son but essentiel est de déterminer la relation entre les objets et leurs catégories pour prédire et découvrir des connaissances [62].

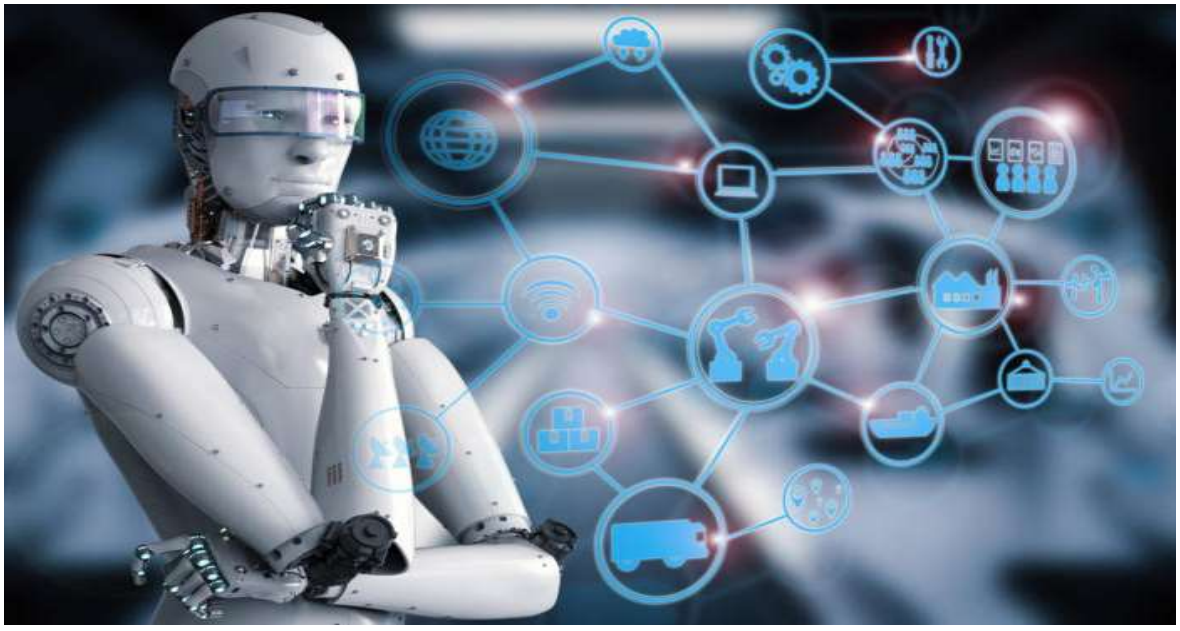


Figure III.3.L'apprentissage automatique [61].

III.4.2. La relation entre Machine Learning et Big Data

Selon **[**52**]**, on ne peut pas dire Big data sans mentionner l'apprentissage automatique car le Big Data est l'essence du Machine Learning, et c'est la technologie qui permet d'exploiter pleinement le potentiel du Big Data.

Le Machine Learning est idéal pour exploiter les opportunités cachées du Big Data. Cette technologie permet d'extraire de la valeur en provenance de sources de données massives et variées sans avoir besoin de compter sur un humain **[**60**].**

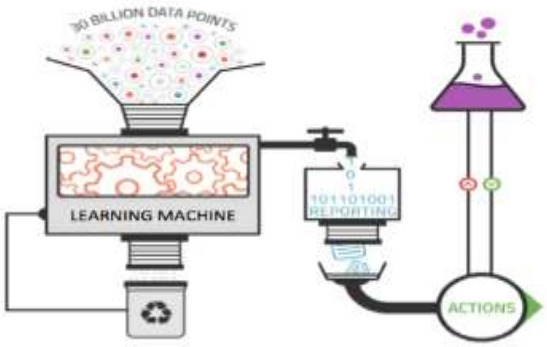


Figure III.4. Machine Learning et sa relation avec le Big data [60].

III.4.3. Méthodes du Machine Learning

Les tâches dans l'apprentissage automatique sont classées en grandes catégories qui sont basées sur la façon dont l'apprentissage est reçu ou comment le feedback sur l'apprentissage est donné au système développé [61].

L'apprentissage automatique a comme objectif de créer des programmes intelligents, au travers de processus d'apprentissage et d'évolution [61].

L'apprentissage supervisé**,** l'apprentissage non supervisé **et** l'apprentissage par renforcement sont les méthodes de la machine learning :

III.4.3.1. L'apprentissage supervisé

L'apprentissage supervisé forme des algorithmes basés sur des données d'entrée et de sortie étiquetées par l'homme ; C'est-à-dire dans l'apprentissage supervisé, l'ordinateur est fourni avec des exemples d'entrées qui sont étiquetés avec les sorties souhaitées. Le but de cette méthode est que l'algorithme puisse «apprendre» en comparant sa sortie réelle avec les sorties «enseignées» pour trouver des erreurs et

modifier le modèle en conséquence. L'apprentissage supervisé utilise donc des modèles pour prédire les valeurs d'étiquettes sur des données non étiquetées supplémentaires [61].

L'apprentissage supervisé a pour but d'établir des règles de comportement à partir d'une base de données contenant des exemples de cas déjà étiquetés. La base de données est en principe un ensemble de couples entrées / sorties {(X, Y)}. Son but est d'apprendre à prédire pour toute nouvelle entrée X, la sortie Y [62].

III.4.3.2. L'apprentissage non supervisé

L'apprentissage non supervisé ou le « clustering » ne fournit pas à l'algorithme des données étiquetées pour lui permettre de trouver une structure et de découvrir une logique dans données entrées [61].

III.4.3.3. L'apprentissage par renforcement

L'apprentissage par renforcement est un cadre formel qui modélise des problèmes décisionnels séquentiels. Il diffère fondamentalement des problèmes supervisés et non supervisés par ce côté interactif et itératif [63,64].

L'apprentissage par renforcement référence à une classe de problèmes d'apprentissage automatique dont le but est d'apprendre, à partir d'expériences successives, ce qu'il convient de faire de façon à trouver la meilleure solution [64].

III.4.4. Algorithmes du Machine Learning

Parmi les algorithmes du Machine Learning les plus courants, on a :

III.4.4.1. L'arbre de décision

III.4.4. 1.1. Définition 1

Un arbre de décision sert à classifier des observations futures étant donné un corpus d'observations déjà étiquetées. L'arbre commence par une racine puis une série de branches dont les intersections s'appellent des nœuds et termine par des feuilles qui correspondent chacune à une des classes à prédire. On parle de profondeur de l'arbre comme étant le nombre maximum de nœuds avant d'atteindre une feuille. Chaque nœud de l'arbre représente une règle. Parcourir l'arbre c'est donc vérifier une série de règles. L'arbre est construit de telle sorte que chaque nœud correspond à la règle (type de mesure et seuil) qui divisera le mieux l'ensemble d'observations de départ [65].

III.4.4.1.2. Définition 2

L'arbre de décision est un algorithme qui se base sur un modèle de graphe **(les arbres)** pour définir la décision finale. Chaque nœud comporte une condition, et les branchements sont en fonction de cette condition (Vrai ou Faux). Plus on descend dans l'arbre, plus on cumule les conditions [66].

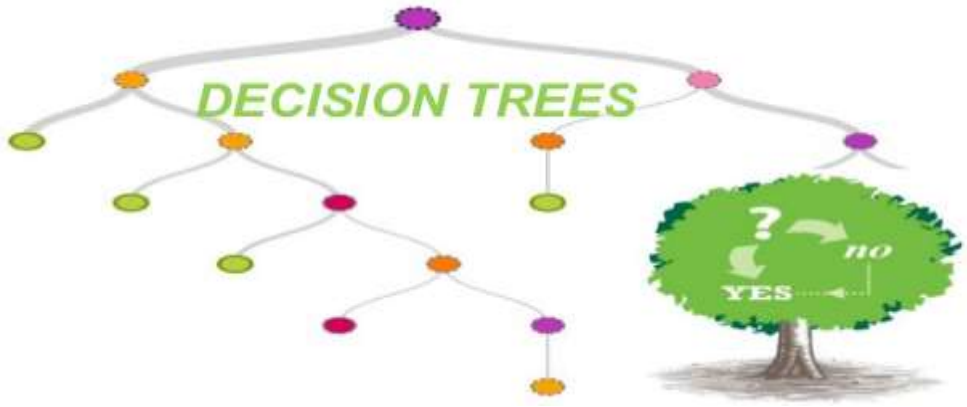


Figure III.5. L'arbre de décision.

III.4.4.2. Les forêts aléatoires

Comme son nom l'indique, l'algorithme des forêts aléatoires se fonde sur les arbres de décisions [65].

Voici les principales étapes :

1. On prend un nombre X d'observations du jeu de données de départ (avec remise) ;
2. On prend un nombre K des M variables disponibles (*features*) ;
3. On entraîne un arbre de décision sur ce jeu de données ;
4. On répète les étapes 1. à 4. N fois de sorte à obtenir N arbres.

Pour une nouvelle observation dont on cherche la classe on descend les N arbres. Chaque arbre propose une classe différente. La classe retenue est celle qui est la plus représentée parmi tous les arbres de la forêt. [66]

III.4.4.3. Le gradient boosting

La méthode du gradient boosting sert à renforcer un modèle qui produit des prédictions faibles, par exemple un arbre de décision [66].

Le principe du gradient boosting est que vous allez refaire un modèle sur l'écart entre la valeur prédite et la vraie valeur à prédire [65].

Vu son importance, on inclut l'algorithme Le gradient boosting dans cette liste bien qu'il ne soit pas "vraiment" un algorithme de machine Learning. En effet, Le gradient boosting (ou Gradient Descent) est un algorithme itératif de minimisation de fonction de coût. Cette minimisation servira à produire des modèles prédictifs comme la régression logistique et la régression linéaire [66].

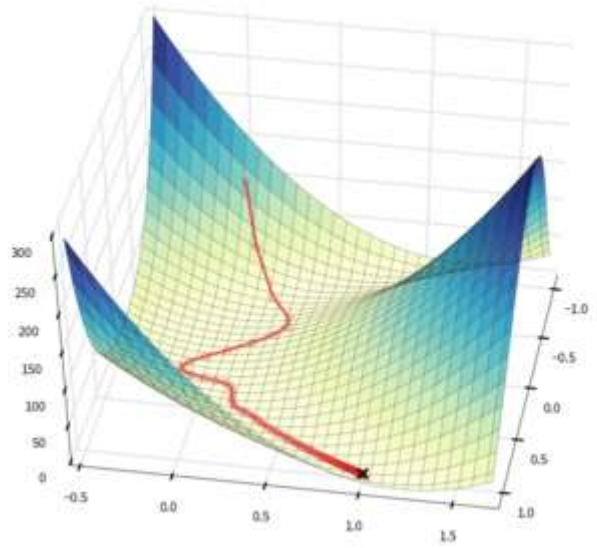


Figure III.6. Le gradient boosting [66]

III.4.4.4. Les machines à vecteurs de support

Aussi connu sous le nom de "SVM" (Support Victor Machine).Cet algorithme sert principalement à des problèmes de classification même si il a été étendu à des problèmes de régression [65].

Machine à Vecteurs de Support (SVM) est lui aussi un algorithme de classification binaire. Tout comme la régression logistique. Si on prend la figure dans la page suivante, nous avons deux classes. La régression Logistique pourra séparer ces deux classes en définissant le trait en rouge. Le SVM va opter à séparer les deux classes par le trait vert [66].

Sans entrer dans les détails, et pour des considérations mathématiques, le SVM choisira la séparation la plus nette possible entre les deux classes (comme le trait vert). C'est pour cela qu'on le nomme aussi *Large Margins classifier* (classifieur aux marges larges) [66].

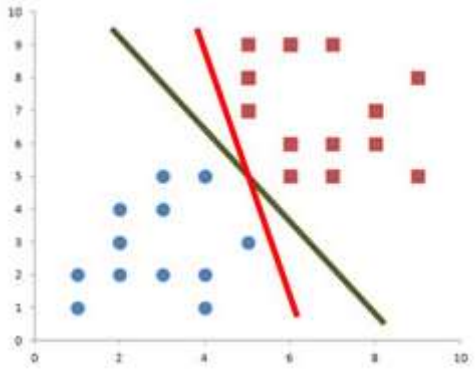


Figure III.7. Machine à Vecteurs de Support (SVM) [66].

III.4.4.5. Les K plus proches voisins (ou K-Means)

K-Means est un algorithme de clustering en Unsupervised Learning. On lui donne un ensemble d'éléments (des données), et un nombre de groupes K. K-means va segmenter en K groupes les éléments. Le groupement s'effectue en minimisant la distance euclidienne entre le centre du cluster et un élément donné [66].

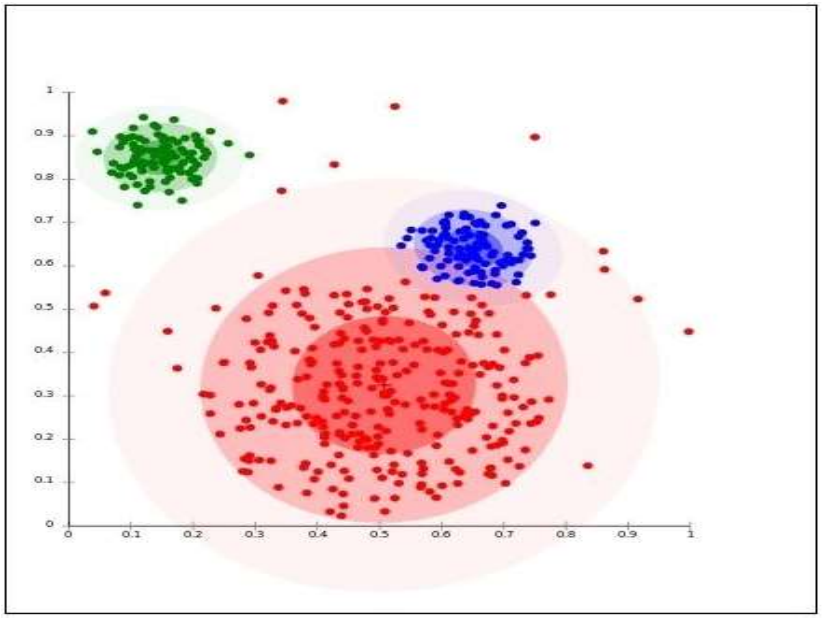


Figure III.8. Les K plus proches voisins [66].

III.4.4.7. La régression logistique

La régression logistique est une méthode statistique pour effectuer des classifications binaires. Elle prend en entrée des variables prédictives qualitatives et/ou ordinales et mesure la probabilité de la valeur de sortie en utilisant la fonction sigmoïd (représentée dans la figure9) [66].

On peut effectuer la classification multi-classes. En utilisant la régression logistique et la méthode un-contre-tous (*One-Versus-All classification*) [66].

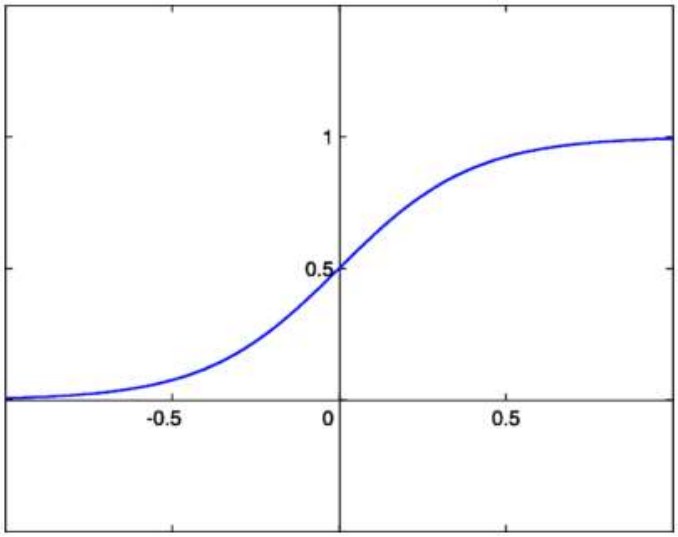


Figure III.9. La régression logistique en utilisant la fonction sigmoïd [66].

III.4.4. 8. Le clustering

Le regroupement consiste à diviser la population ou les points de données en un certain nombre de groupes, de sorte que les points de données du même groupe soient plus similaires aux autres points de données du même groupe et différents des points de données des autres groupes. C'est fondamentalement une collection d'objets basée sur la similitude et la dissimilarité entre eux [67].

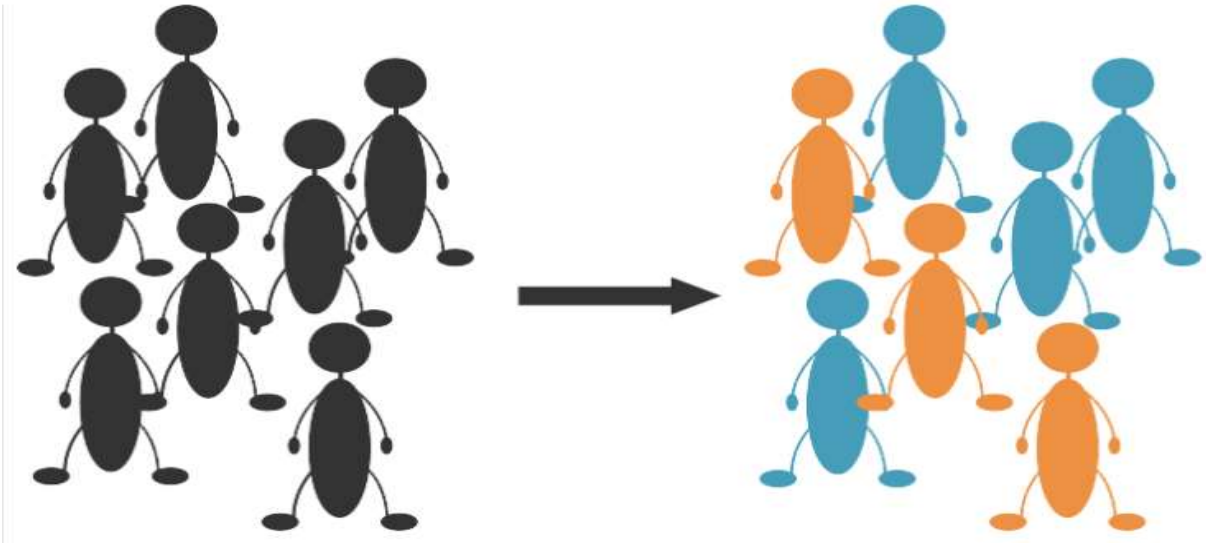


Figure III.9. Le Clustering [68]

III.4.4. 9. Régression linéaire

Les algorithmes de régression linéaire modélisent la relation entre des variables prédictives et une variable cible**.** La relation est modélisée par une fonction mathématique de prédiction. Le cas le plus simple est la régression linéaire univariée. Elle va trouver une fonction sous forme de droite pour estimer la relation. La régression linéaire multivariée intervient quand plusieurs variables explicatives interviennent dans la fonction de prédiction. Et finalement, la régression polynomiale permet de modéliser des relations complexes qui ne sont pas forcément linéaires [66].

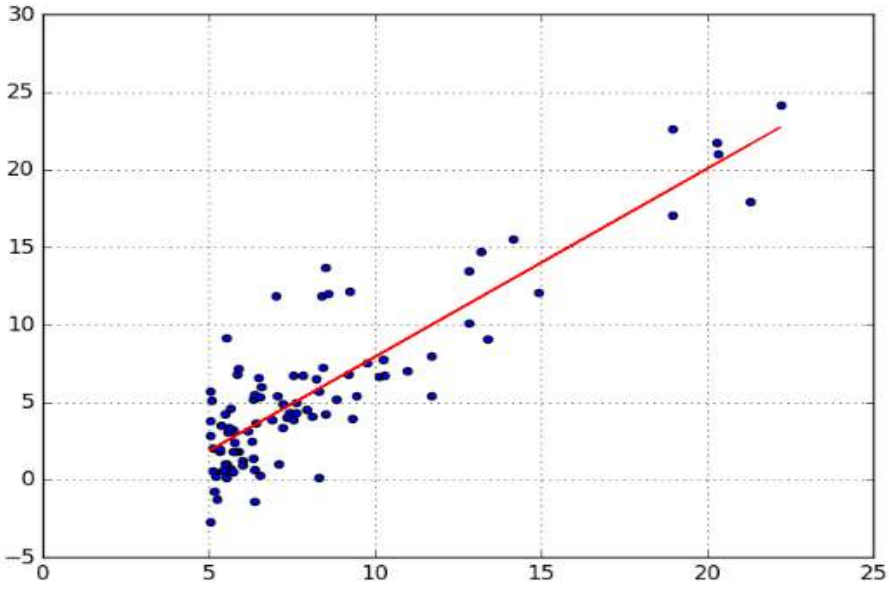


Figure III.10.Régression linéaire [66].

III.4.4. 10. Naïve Bayes

Naïve Bayes est un classifieur assez intuitif à comprendre. Il se base sur le théorème de Bayes des probabilités conditionnelles. Naïve Bayes assume une hypothèse forte (naïve). En effet, il suppose que les variables sont indépendantes entre elles. Cela permet de simplifier le calcul des probabilités [66].

Généralement, le Naïve Bayes est utilisé pour les classifications de texte en se basant sur le nombre d'occurrences de mots [66].

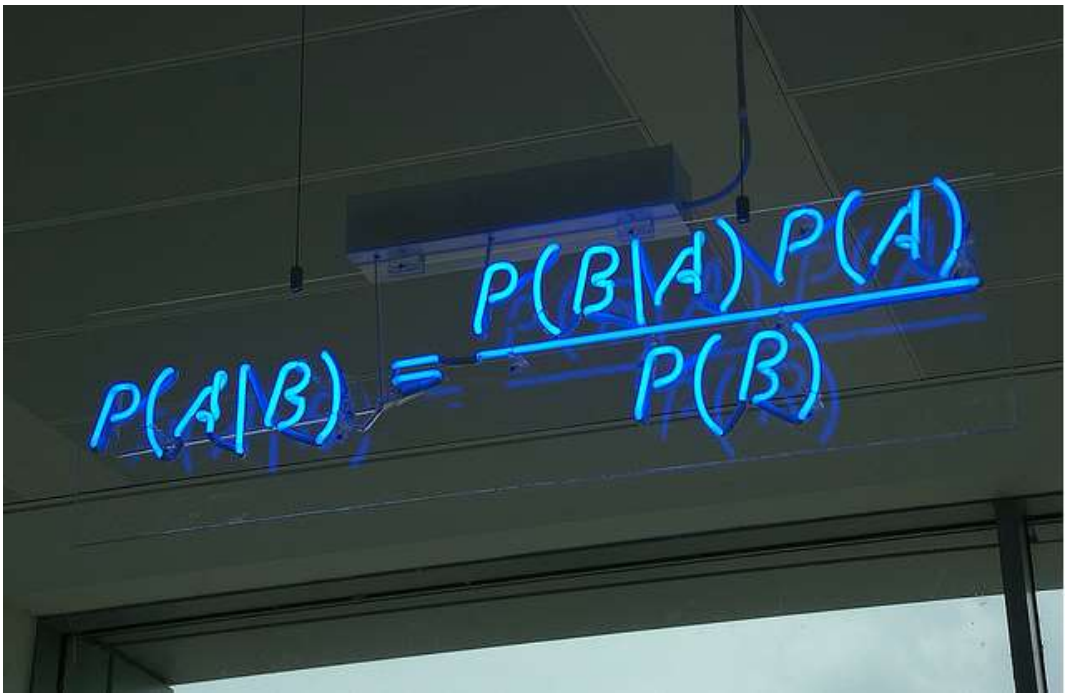


Figure III.11. La formule du théorème de Bayes [66].

III.4.4. 11. Détection d'une anomalie

Détection d'une anomalie est un algorithme de Machine Learning pour détecter des patterns anormaux. Il est très utile pour la détection de fraudes dans les transactions bancaires, et les détections d'intrusions.

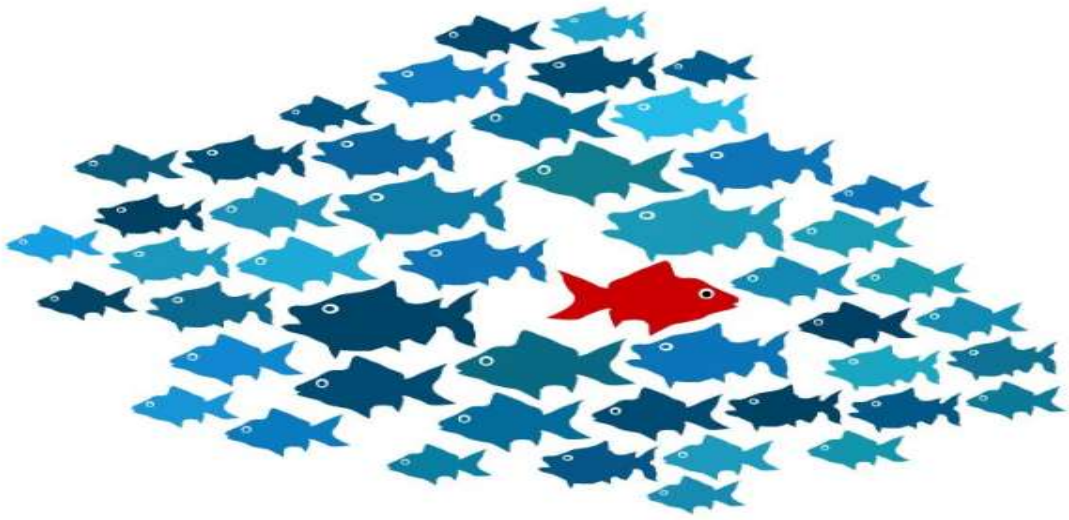


Figure III.12.Détection d'une anomalie [66].

III.4.4. 12. Les réseaux de neurones

Les réseaux de neurones sont inspirés des neurones du système nerveux humains. Ils permettent de trouver des patterns complexes dans les données. Ces réseaux de neurones apprennent une tâche spécifique en fonction des données d'entrainement.

Les réseaux de neurones se composent de nœuds (les cercles dans l'image). Dans ces réseaux, on retrouve le tiers d'entrée (*Input Layer*) qui va recevoir les données d'entrées. *L'Input Layer* va propager les données par la suite aux tiers cachés (*HiddenLayers*). Finalement le Tiers de sortie (le plus à droite) permet de produire le résultat de classification. Chaque tiers du réseau de neurones est un ensemble d'interconnexions des nœuds d'un tiers avec ceux des autres tiers [66].

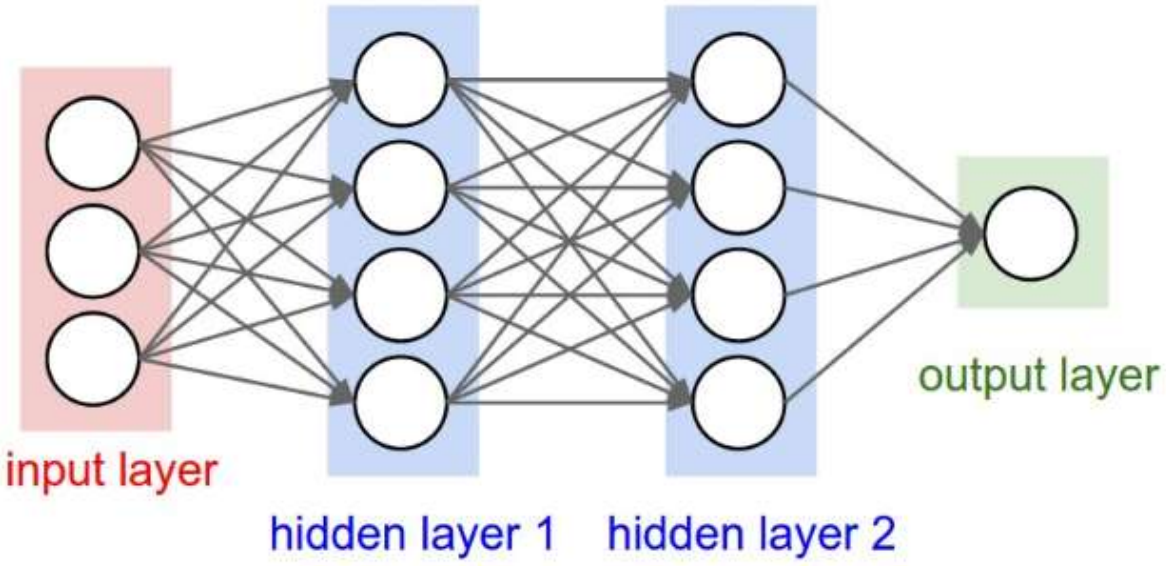


Figure III.13. Les réseaux de neurones [66].

III.5. Conclusion

Les données massives peuvent signifier un stockage important, en supposant que toutes les données doivent être stockées. Contrairement à l'archivage et à l'analyse de données traditionnels, nous pouvons effectuer des analyses du Big Data au moment de la collecte de données. Par conséquent, il se peut que nous devions gérer uniquement un sous-ensemble plus petit, comme des échantillons à l'aide du Machine learning.

Chapitre IV

Manipulation et traitement parallèle du Big data

IV. Introduction

La gestion des données évolutives et distribuées qui a été la vision de la communauté des chercheurs dans le domaine des bases de données depuis plus de trois derniers décennies, s'est accéléré remarquablement ces dernières années pour pallier aux rudes contraintes liées aux serveurs, aux réseaux et aux applications évoluant dans des environnements distribués imposés par les nouvelles tendances informatiques. Les dernières étapes de l'évolution informatique ont émergé de nouvelles technologies.

La mise en œuvre de la séparation du traitement des métadonnées et du transport des données n'est pas aisée, la recherche dans cet axe est toujours d'actualité. Les différentes approches d'amélioration de performance et d'extensibilité dans la matière seront très considérées dans les prochaines échéances. Ce chapitre construit une collection des architectures distribuées et des technologies relatives au Big Data.

IV.1. Approches de traitement des données

IV.1.1. Introduction

Les approches de traitement de données se diffèrent selon les besoins de l'entreprise et les objectifs attendus de l'application. Principalement, les applications actuelles utilisent une vision Batch en récupérant les données, les stocker et puis les requêter via des jobs ce que l'on nomme un traitement en Batch. Mais ce n'est pas le seul moyen, en effet le streaming ou le traitement en temps réel répond aussi assez bien à d'autres besoins des entreprises.

Il existe trois approches, on parle de : Batch, Micro Batch et streaming.

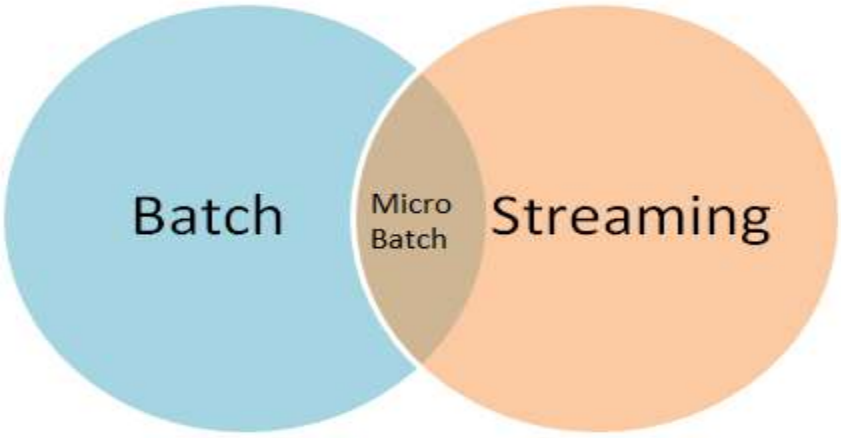


Figure IV.1. Approches de traitement des données.

IV.1.2.1. L'approche Batch

L'approche batch est un moyen efficace pour le traitement de gros volumes de données où on applique un ensemble de transactions dans une période de temps.

Les données sont collectées, entrées, traitées et ensuite, les résultats sont produits. Cette approche nécessite des programmes distincts entre les différentes étapes entre l'entrée, le traitement et la sortie [69,70].

IV.1.2.2. L'approche Micro-Batch

Micro-Batch est une approche qui permet à un processus ou une tâche de traiter un flux comme une séquence de batch ou un bloc de données. Pour les flux entrants, les évènements peuvent être partagés en petits Batch et ensuite délivrés à un système de batch globale pour le traitement [70,71].

IV.1.2.3. L'approche temps réel (Streaming)

L'approche Streaming ou traitement en temps réel implique une entrée continue des données, des traitements à l'arrivée et une sortie continue des résultats de traitement. Les données doivent être traitées dans une courte période ce qu'on appelle un temps presque réel [69].

IV.2. Les architectures du Big Data

IV.2.1. Les architectures avancées

IV.2.1.1 L'architecture Lambda

L'architecture lambda est une approche qui permet le traitement et le stockage des données massives, elle mixte entre les deux approches de traitement de données Batch et streaming (real-time) dans un même Framework. Donc elle admet les problématiques relatives à la vélocité et le volume des données. Elle est divisée en trois couches : Batch, service et temps réel comme montré dans le schéma suivant [72]:

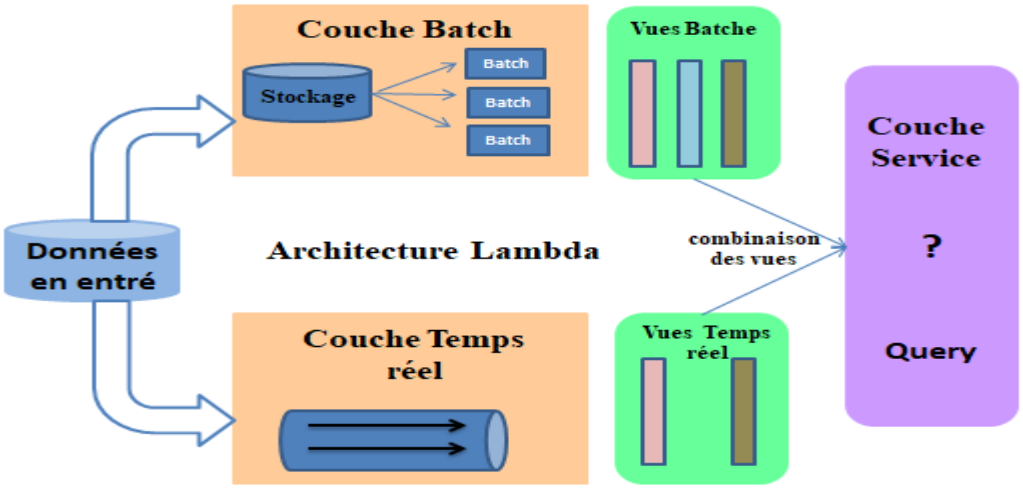


Figure IV.2. Architecture Lambda.

IV.2.1.1.1. La couche Batch

C'est la partie qui implémente la vue Batch et qui est définie par l'équation :

La vue Batch = Fonction (toutes les données) et La requête = Fonction (La vue Batch).

Le traitement en batch dans cette architecture est souvent prédominé par le Framework Hadoop avec son paradigme de traitement *MapReduce* ou bien Spark. L'avantage de Hadoop est sa capacité à écrire des programmes (scripts) qui peuvent être automatiquement propagés et distribués sur l'ensemble des nœuds constituant le cluster [73].

IV.2.1.1.2. La couche temps réel

Pour un système qui reçoit une grandes masses des données en temps réel où les besoins métiers sont très exigeants en matière de latence et de rapidité de rafraîchissement, la couche Batch toute seul ne suffit pas, des nouvelles données arrivent entre un calcul d'une vue batch à une autre, c'est cette différence qui est comprise par la couche temps réel. Le but d'une telle couche est bien de permettre la prise en compte de cette différence. On peut voir la couche temps réel comme étant une couche Batch à la différence qu'elle prend en considération que les données récentes.

On peut donc formaliser cette couche en utilisant l'équation suivante [70] :

La vue temps réel = Fonction (La vue temps réel, données récentes).

Donc une vue temps réel est mis à jour automatiquement à l'arrivée des nouvelles données. Ainsi, on peut résumer l'approche lambda par les trois équations suivantes :

La vue batch = Fonction (toutes les données)

et La vue real-time = Fonction (La vue réel time, données récentes)

donc :

Requête = Fonction (La vue réel time, La vue batch).

Un avantage crucial de la couche temps réel est bien l'isolation de la complexité, cela vaut dire que cette partie d'une part, elle élimine les anciennes vues une fois les données représentées sont chargées par la couche Batch vers la couche service, d'une autre part, et au cas de problème, il est nettement possible de désactiver la couche temps réel sans toucher les ensembles des données stockées physiquement dans la couche Batch, ce qui isole la complexité dans un seul endroit temporaire qui est la couche temps réel et qui ne garde pas ses résultats indéfiniment. Donc on obtient une faible latence avec une grande robustesse toute en gardant un niveau de complexité faible et gérable. C'est l'esprit de l'architecture lambda [73].

IV.2.1.1.3. La couche service

La couche Batch prépare ses fonctions qui sont ensuite classées et utilisées par la couche service pour qu'elles soient requêtées. Cette même couche est responsable sur le chargement de nouvelles vues et les vues déjà chargés qui ont connu une mise à jour. Il est remarquable que les deux couches ensembles puissent satisfaire la majorité des besoins d'un système du Big Data : une tolérance aux pannes, une scalabilitée assurée par la qualité distribuée de Hadoop, une flexibilité concrétisée par la possibilité de rajout des requêtes et une débugabilité avec des logs bien détaillés. La seule propriété qui manque est la faible latence des requêtes ce qui nous induit à introduire la nouvelle couche de temps réel [73].

IV.2.1.2. L'architecture Kappa

L'idée de l'architecture Kappa a été issue sur la base de l'architecture Lambda exposée dans un article présenté par Jay Kreps de LinkedIn. Elle permet donc de simplifier l'architecture Lambda en fusionnant les couches temps réel et batch [74].

L'idée peut se formuler comme suit : Contrairement à l'approche Lambda, Kappa utilise un système de traitement en temps réel pour tous le processus et évite de se coller à un autre système. En d'autres termes, pourquoi ne pas utiliser le mode temps réel pour le système global et même quand il s'agit de changement de code ou de fonctionnalités de notre système. Vu que déjà que le mode temps réel a cette notion de parallélisme, alors ça sera peut-être une meilleure approche en ce qui concerne les tâches qui devraient être en mode batch en retraitant les données en temps réel pour obtenir les mêmes résultats [70].

Voici l'architecture Kappa ayant pour but de bien gérer le retraitement des données :

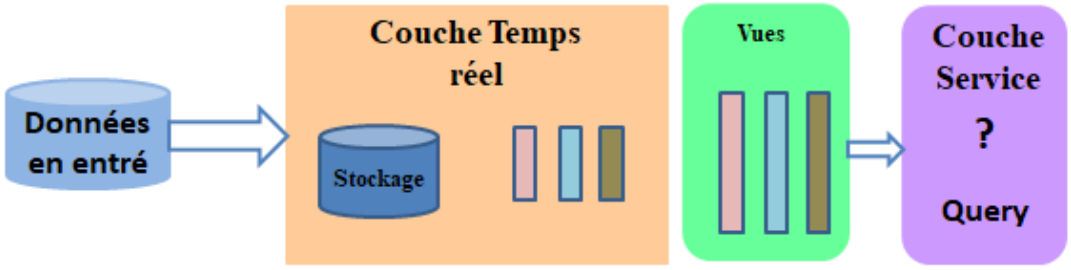


Figure IV.3. Architecture Kappa.

IV.2.1.3. L'architecture Zeta

L'architecture Zeta est une structure d'architecture d'entreprise de haut niveau, semblable à l'architecture Lambda, qui simplifie les processus d'entreprise et définit un moyen évolutif d'accroître la vitesse d'intégration des données dans l'entreprise [75].

Au fait, [76] a conçu cette architecture durant son travail dans une entreprise de publicité numérique. Il a pensé directement au traitement en temps réel. Au départ, il a mis en place un système qui répond au bout de 2 à 3 heures pour que le système change et reçoive le feedback de client. Il a continué à améliorer le système jusqu'à avoir comme délai de réponse de 5 à 10 minutes, mais il a trouvé que ce n'était toujours pas suffisant. Il a compris enfin que l'utilisateur ne sens pas le temps réel et du coup la notion de « *good enough* » n'étais jamais suffisante mais plutôt « *we need to be faster* ! » c'était le vrai besoin de l'utilisateur et c'est pourquoi, il a trouvé que c'est nécessaire de mettre en place une approche temps réel et qui répond aux exigences de l'utilisateur. Et c'est de là que de l'architecture Zeta est née inspirée principalement de modèle architectural de Google dans l'utilisation de Gmail [76].

Ci-dessous l'architecture proposé par [76] dans le cadre d'utilisation de l'écosystème Hadoop avec une abstraction théorique ou on présente le fonctionnement de chaque composant [75,76] :

❖ **Système de fichier distribué partagé :** les applications seront capables de lire et écrire dans un emplacement commun, qui permet la simplification de reste de l'architecture.

❖ **Stockage des données en temps réel :** Ce mécanisme pour supporter les exigences des applications métiers des grandes vitesses en permettant l'accès à des bases de données partagées en temps réel.

❖ **Système de gestion de récipient :** Ce système est utilisé pour avoir une approche standard pour déployer les logiciels il faut que chaque logiciel puisse être isolé et déployé selon cette manière standard.

❖ **La solution architecture :** Ce composant a été mise en place pour répondre à un besoin bien spécifique. Certaines solutions imposent l'utilisation de plus qu'une seule application ayant une grande interaction entre les algorithmes et les libraires de programmation. Avec l'architecture Zeta cette différenciation ne pose pas de grand problème car elle est bien gérée en séparant les solutions.

❖ **Moteur d'exécution :** Différents groupes de business requièrent des besoins et des exigences différentes, et pour bien répondre à toutes ces contraintes on a besoin d'un moteur qui peut supporter différents modèles.

❖ **Application d'entreprise :** dans le passé, c'était ce genre d'application qui guide les autres composants, mais avec la nouvelle approche c'est les autres composants qui préparent le terrain à ces applications pour atteindre les buts de business plus facilement.

❖ **Gestionnaire de ressources globale** : Permet l'allocation dynamique des ressources pour permettre à l'entreprise de s'adapter facilement à n'importe quelle tâche jugée plus importante dans un moment donné.

Figure IV.4. Architecture Zeta.

IV.2.1.4. L'architecture SMACK

L'architecture SMACK est différente que les architectures Lambda, Kappa et Zeta, puisqu'elle est composée de plusieurs solutions du Big Data (Spark, Mesos, Akka, Cassandra, Kafka) plutôt que sur des principes et pattern. La figure ci-dessous illustre bien le mode de fonctionnement de l'architecture SMACK [77].

IV.2.1.4.1. Spark

Apache Spark est un moteur rapide et général pour le traitement de données à grande échelle, il fournit une analyse des données en temps quasi réel (en vas voir le détailles ultérieurement).

IV.2.1.4.2. Mesos

Mesos est un noyau de systèmes distribués, repose sur les mêmes principes que le noyau Linux, mais à un niveau d'abstraction différent. Le noyau Mesos s'exécute sur toutes les machines et fournit aux applications (Hadoop, Spark, Kafka) des API pour la gestion des ressources et la planification dans des environnements de centre de données [78].

IV.2.1.4.3. Akka

Akka est une implémentation de modèle d'acteur, une boîte à outils et un environnement d'exécution permettant de créer des applications pilotées par message hautement concurrentes, distribuées et résilientes sur la machine virtuelle Java. Akka a été conçu pour permettre aux développeurs de créer facilement des applications réactives utilisant un niveau d'abstraction élevé. Il le fait de manière très naturelle et simple, sans avoir à traiter avec des concepts de bas niveau [78].

IV.2.1.4.4. Cassandra

Apache Cassandra est une base de données distribuée permettant de gérer de grandes quantités de données structurées sur de nombreux serveurs de base, tout en offrant un service hautement disponible et aucun point de défaillance unique. L'architecture de Cassandra est responsable de sa capacité à évoluer, à exécuter et à offrir une disponibilité continue. Plutôt que d'utiliser un ancien maître-esclave ou une architecture fragmentée manuelle et difficile à maintenir [78].

IV.2.1.4.5. Kafka

Apache Kafka est un journal de validation distribué, une alternative à la messagerie publication-abonnement. Il gère une charge de données considérable et évite les systèmes à contre-pression pour gérer les inondations. Il inspecte les volumes de données entrants, ce qui est très important pour la distribution et le partitionnement sur les nœuds du cluster [78].

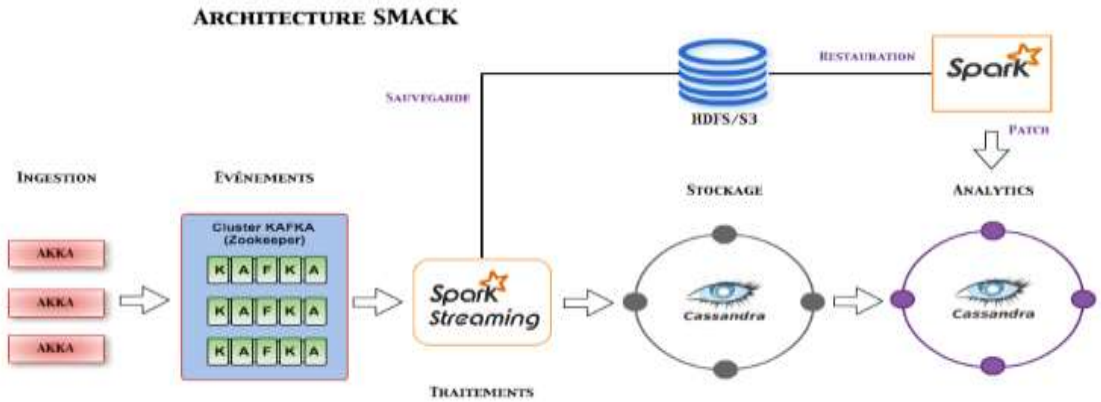


Figure IV.5. Architecture SMACK.

IV.2.2.Les architectures distribués

IV.2.2.1. Les nuages informatiques

IV.2.2.1.1. Définition

L'organisme NIST (*National Institute of Standards and Technology*) définit le Cloud Computing comme un modèle fournissant, à la demande et au travers d'un réseau, un ensemble partagé de ressources informatiques incluant des serveurs, des espaces de stockage, des applications, des traitements et des plates-formes de déploiement qui peuvent être rapidement mises en service avec un effort minimum de gestion et d'interaction avec le fournisseur de ce service [79].

Le Cloud Computing permet de rendre une infrastructure matérielle et logicielle dynamique et flexible en exposant les capacités des Data-Centers comme étant un "réseau de services virtuels". Dans cette infrastructure, les utilisateurs peuvent accéder et déployer des applications à partir d'Internet suivant leurs demandes et la qualité de service exigée [80].

IV.2.2.1.2 Modèles de services

Le Cloud Computing possède trois modèles de service [80,81] :

IV.2.2.1.2.1. SaaS (Software as a Service)

Dans ce modèle, le logiciel est offert sous la forme d'un service. Le fournisseur de Cloud de type SaaS gère entièrement sa plateforme matérielle et logicielle, et les clients du Cloud utilisent ainsi le logiciel fourni sans s'occuper de la pile en dessous (plateforme applicative, matériel, …etc) ni de l'installation du logiciel en question.

IV.2.2.1.2.2.PaaS (Plateforme as a Service)

Dans ce second modèle, le fournisseur offre une plateforme sous forme d'environnement complet sur laquelle des développeurs ou éditeurs de logiciels des clients peuvent déployer des applications. La pile en dessous de cette plateforme à savoir le socle applicatif, le système d'exploitation, le matériel et le réseau, sont gérés par le fournisseur de service. Notons que certaines offres PaaS exigent un langage de programmation spécifique.

IV.2.2.1.2.3. IaaS (Infrastructure as a Service)

Infrastructure en tant que Service (IaaS) est la proposition d'un ensemble de services informatiques comprenant le matériel, le réseautage et le stockage. Le fournisseur offre une plateforme sur laquelle les clients vont pouvoir déployer de ressources d'infrastructures dont la plus grande partie est localisée à distance dans des *Data Centers*. Le client acquiert une ressource et est facturé pour cette ressource en fonction de la quantité utilisée et de la durée d'utilisation. On trouve des versions publiques et privées d'IaaS.

Dans le modèle IaaS publique, l'utilisateur utilise une carte de crédit pour acquérir ces ressources, dès que l'utilisateur cesse de payer, la ressource disparaît. Dans un service IaaS privé, c'est généralement la structure informatique qui crée l'infrastructure conçue pour fournir des ressources à la demande pour les utilisateurs internes et parfois les partenaires commerciaux. Ce modèle considéré comme une évolution des Data Centers virtualisés, permet au client de faire abstraction du modèle physique (gestion des serveurs physiques, l'électricité, la climatisation, la sécurité physique des centres de données). Dans ce modèle, le fournisseur contrôle le matériel et la couche de virtualisation. Sur l'aspect de données, le contrôle est effectué au niveau de la machine virtuelle qui est stockée et sauvegardée par le fournisseur de Cloud IaaS.

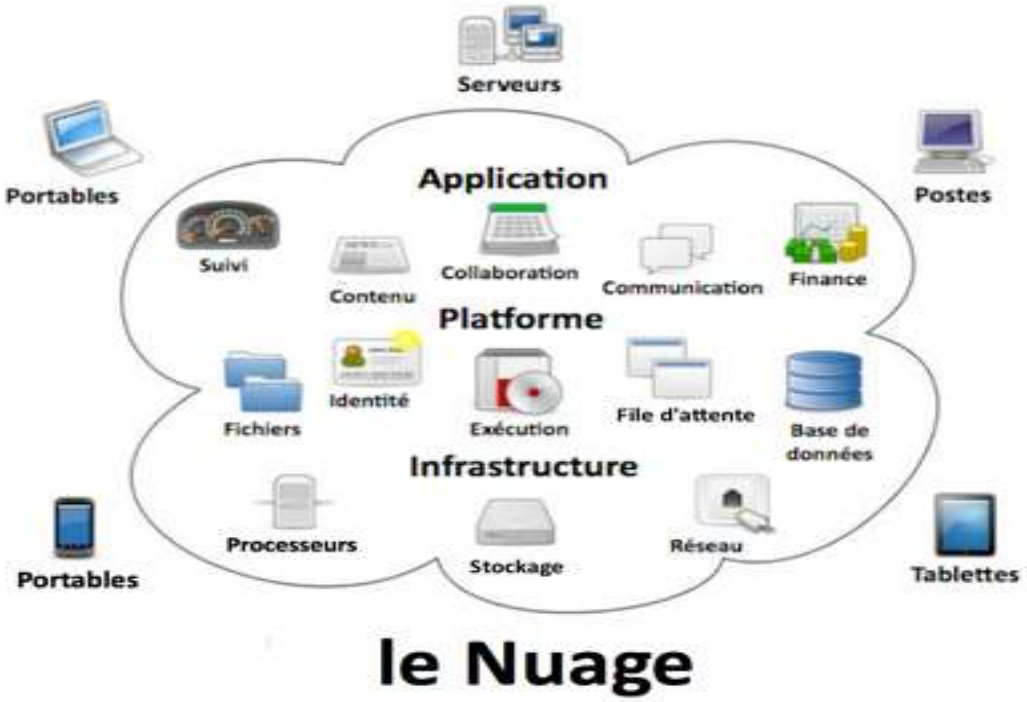


Figure IV.5. Le nuage informatique (Cloud Computing).

IV.2.2.2. Les grilles

Le terme grille désigne un ensemble beaucoup plus important de machines, hétérogènes, réparties sur différents domaines d'administration réseau. Les différentes entités composant une grille peuvent être réparties sur l'ensemble de la planète et communiquent entre elles en utilisant une grande diversité de réseaux allant de l'Internet à des réseaux privés très haut débit [81].

Une grille informatique est une infrastructure matérielle et logicielle qui fournit un accès consistant et peu onéreux à des ressources informatiques. Le but est ainsi de fédérer des ressources provenant de diverses organisations désirant collaborer en vue de faire bénéficier aux utilisateurs d'une capacité de calcul et de stockage qu'une seule machine ne peut fournir.

Cependant, tout système informatique distribué ne peut posséder l'appellation de grille. En effet, une grille est un système qui coordonne des ressources non soumises à un contrôle centralisé, qui utilise des protocoles et interfaces standards dans le but de délivrer une certaine qualité de service (en termes de temps de réponse ou bien de fiabilité par exemple).

Plusieurs types de grilles peuvent être distingués selon l'utilisation recherchée :

- **Grille d'information:** la ressource partagée est la connaissance. L'Internet en est le meilleur exemple: un grand nombre de machines hétérogènes réparties sur toute la surface du globe autorisant un accès transparent à l'information.
- **Grille de stockage:** l'objectif de ces grilles est de mettre à disposition un grand nombre de ressources de stockage d'information afin de réaliser l'équivalent d'un "super disque dur" de plusieurs pétabyte.
- **Grille de calcul:** l'objectif de ces grilles est clairement d'agréger la puissance de traitement de chaque nœud de la grille afin d'offrir une puissance de calcul "illimitée".

IV.2.2.3. Les grappes

Une grappe de machine (***Cluster***) désigne un ensemble de d'ordinateurs, appelés nœuds, tous interconnectés, dans le but de partager des ressources informatiques. Une grappe peut être constituée d'ordinateurs de bureaux, de "racks" de machines constituées de composants standards ou de "lames" également constituées de composants standards afin d'optimiser l'espace physique[81].

Une grappe est généralement composée de machines homogènes en termes d'architecture et de système d'exploitation. Elle ne regroupe que des machines appartenant au même domaine d'administration réseau et les nœuds communiquent entre eux en utilisant un réseau de communication rapide. Les différents nœuds d'une grappe possèdent souvent une configuration logicielle semblable [81].

IV.3. Les technologie du Big Data

Les données massives générées sont hétérogènes, ils exigent des technologies de stockage et de traitement avancés (non traditionnels). Plusieurs technologies ont été proposées dans ce contexte, nous citons principalement Hadoop et son écosystèmes.

IV.3.1. Ecosystèmes Hadoop

Ecosystème Hadoop n'est pas un langage de programmation ni un service, c'est un Framework englobant un certain nombre de services (stockage, analyse et maintenance) permettant de résoudre les problèmes liés au Big Data

Le schéma ci-dessous englobe les composants Hadoop, qui forment ensemble un écosystème Hadoop :

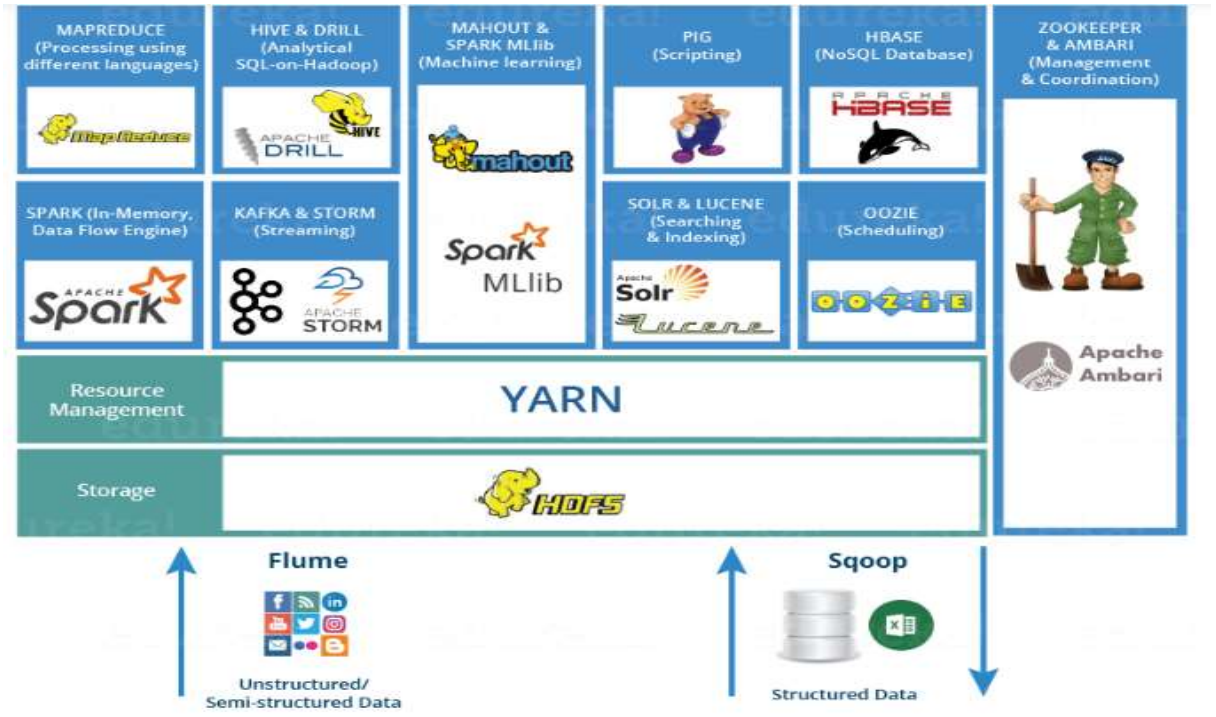


Figure IV.6. Ecosystème Hadoop.

IV.3.1.1. Hadoop [82]

Hadoop (*High-availability distributed object-oriented platform)*est un Framework open source d'apache permettant de stocker et de réaliser des traitements sur les volumes de données massives, de l'ordre de plusieurs pétaoctets.

Il est caractérisé par :

- **Robuste :** si un nœud de calcul tombe, ses tâches sont automatiquement réparties sur d'autres nœuds. Les blocs de données sont également répliqués.
- **Coût :** il optimise les coûts via une meilleure utilisation des ressources présentées.
- **Souple :** car il répond à la caractéristique de variété des données en étant capable de traiter différents types de données.
- **Virtualisation :** ne plus se reposer directement sur l'infrastructure physique (baie de stockage coûteuse), mais choisir la virtualisation de ses clusters Hadoop**.**

Hadoop est principalement constitué de quatre composants [83] :

- ***Hadoop Distributed File System (HDFS) :*** Un système de fichiers distribués qui fournit un accès haut-débit aux données de l'application.
- ***Hadoop YARN :*** Un Framework pour la planification des tâches et la gestion des ressources du cluster.
- ***Hadoop MapReduce :*** Un système basé sur YARN pour le traitement parallèle des gros volumes de données.
- ***Hadoop Common :*** Les utilitaires communs qui supportent les autres modules d'Hadoop. Plus concrètement l'écosystème Hadoop comprend de nombreux autres outils couvrant le stockage et la répartition des données, les traitements distribués, l'entrepôt de données, le workflow, la programmation, sans oublier la coordination de l'ensemble des composants. On parle des outils comme Hive, Pig, Hbase, Flume**.**

IV.3.1.1.1. Le système de fichier distribué Hadoop

Le HDFS (***Hadoop Distributed File System)*** est un système de fichier distribué, extensible et portable inspiré par le Google File System (GFS) [84].

Le HDFS est le composant principal de Hadoop Ecosystème. Il a été conçu pour stocker différents types de grands ensembles de données structurées, non structurées et semi-structurées. Il peut fonctionner à faible coût machines avec une tolérance aux pannes élevée [84].

IV.3.1.1.1.1.1 Les composantes d'HDFS

HDFS définit deux types de nœuds :

- **Le nœud principal (*NameNode*)**

Il est caractérisé par :

- ✓ Responsable de la distribution et de la réplication des blocs ;
- ✓ Serveur d'informations du HDFS pour le client HDFS;
- ✓ Stocke et gère les données massive ;
- ✓ Contient la liste des DataNodes pour chaque bloc (dans le cas de l'écriture) ;
- ✓ Comporte la liste des blocs pour chaque fichier (dans le cas de lecture).

- **Le nœud de données(*DataNode*)**

Il est caractérisé par :

- ✓ Stocke des blocs de données dans le système de fichier local ;
- ✓ Maintenir des métadonnées sur les blocs possédés ;
- ✓ Serveur de bloc de données et de métadonnées pour le client HDFS.

- **Secondary NameNode**

Le NameNode dans l'architecture Hadoop est un point unique de défaillance. Si ce service est arrêté, il n'y a pas moyen de pouvoir extraire les blocs d'un fichier donné. Pour résoudre ce problème, un NameNode secondaire appelé Secondary NameNode a été mis en place dans l'architecture Hadoop.

IV.3.1.1.2. Yarn [85]

Apache Hadoop YARN (***Yet Another Resource Negotiator***) est une technologie de gestion des ressources du cluster. Elle Considéré comme le cerveau de l'écosystème Hadoop.

Elle rend l'environnement Hadoop mieux adapté aux applications opérationnelles, elle effectue toutes les activités de traitement en allouant des ressources et en planifiant des tâches.

YARN comporte deux composants principaux Resource-Manager et Node-Manager.

- **Resource-Manager :** est un nœud principal du service de traitement. Il reçoit les requêtes de traitement, puis transmet les parties de requêtes aux gestionnaires de nœuds correspondants.

❖ **Node-Managers** : sont installés sur chaque DataNode. Il est responsable de l'exécution de la tâche sur chaque DataNode.

IV.3.1.1.3. MapReduce

C'est le composant de base du traitement dans un écosystème Hadoop, il fournit la logique de traitement. En d'autres termes, MapReduce est un framework logiciel qui aide à l'écriture d'applications qui traitent de grands ensembles de données en utilisant des algorithmes distribués et parallèles dans l'environnement Hadoop [85]**.**

Un programme MapReduce peut se résumer à deux fonctions Map () et Reduce ().

❖ La fonction **Map** : cette fonction effectue plusieurs actions telles que le filtrage, le regroupement et le tri.

❖ La fonction **Reduce** : agrège et résumer le résultat produit par la fonction Map.Le résultat généré par la fonction Map est une paire de valeurs de clé (K, V) qui sert d'entrée à la fonction de réduction.

IV.3.1.1.4. Hadoop Common

IV.3.1.1.4.1. HBase

HBase est un système de gestion de base de données non relationnel orienté colonne qui s'exécute sur le système de fichiers distribués Hadoop (HDFS) [85]. HBase fournit un moyen tolérant aux pannes de stocker des ensembles de données fragmentés, qui sont courants dans le cas d'utilisation de données volumineuses. Il est bien adapté au traitement de données en temps réel ou à un accès en lecture / écriture aléatoire à de gros volumes de données [86].

IV.3.1.1.4.2. Apache Pig

Apache Pig est une plate-forme utilisée pour analyser de grands ensembles de données les représentant sous forme de flux de données. Il est conçu pour fournir une abstraction sur MapReduce, réduisant ainsi la complexité de l'écriture d'un programme MapReduce. Les opérations de manipulation de données effectuent facilement dans Hadoop avec Apache Pig.PIG se compose de deux parties: Pig Latin, la langue et le temps d'exécution pour l'environnement d'exécution [85,87].

IV.3.1.1.4.3. Apache Hive

Hive est à l'origine un projet Facebook puis il a donné à la fondation Apache qui permet de faire le lien entre le monde SQL et Hadoop [85].

Apache Hive est un système d'entrepôt de données construit sur Hadoop et est utilisé pour analyser des données structurées et semi-structurées. Hive fait abstraction de la complexité de Hadoop MapReduce. Fondamentalement, il fournit un mécanisme pour projeter la structure sur les données et exécuter des requêtes écrites en HQL (*Hive Query Language*) similaires aux instructions SQL,ces requêtes ou HQL sont converties en mappage de tâches réduites par le compilateur Hive.Apache Hive prend en charge le langage DDL (*Data Definition Language*), le langage DML (*Data Manipulation Language*) et les fonctions définies par l'utilisateur (*UDF*) [88].

IV.3.1.1.4.4. Apache Flume

Apache Flume est un outil d'ingestion de données dans HDFS. Il collecte, regroupe et transporte une grande quantité de données en continu telles que les fichiers journaux, les événements provenant de diverses sources telles que le trafic réseau, les médias sociaux, les courriers électroniques, etc. vers HDFS. Flume est un outil très fiable et distribué [85,89].

IV.3.1.1.4.5. Apache Oozie

Oozie est un système évolutif, fiable et extensible de planification de flux de travail permettant de gérer les travaux Apache Hadoop.Il est intégré au reste de la pile Hadoop et prend en charge plusieurs types de travaux Hadoop prédéfinis (tels que Java Map -Reduce, Streaming Map-Reduce, Pig, Hive, Sqoop et Distcp), ainsi que des travaux spécifiques au système (tels que Programmes Java et scripts shell) [90].

IV.3.1.1.4.6. Apache Drill

Apache Drill est un Framework open source qui fonctionne avec un environnement distribué pour analyser de grands ensembles de données. Il est utilisé pour explorer n'importe quel type de données. Il supporte différents types de bases de données NoSQL et de systèmes de fichiers, ce qui est une fonctionnalité puissante de Drill [85].

IV.3.1.1.4.7. Apache Zookeeper

Apache Zookeeper est le coordinateur de tout travail Hadoop qui comprend une combinaison de divers services dans un écosystème Hadoop. Il assure la coordination avec divers services dans un environnement distribué [91].

Avant Zookeeper, la coordination entre les différents services de l'écosystème Hadoop était très longue et difficile. Auparavant, les services avaient de nombreux problèmes d'interactions, telles que la configuration commune lors de la synchronisation des données. Même si les services sont configurés, des modifications apportées à leur configuration rendent le traitement complexe et difficile à gérer. Le regroupement et la dénomination prenaient également beaucoup de temps [85].

IV.3.1.1.4.8. Apache SQOOP

Apache Sqoop est un outil conçu pour transférer efficacement des données entre Apache Hadoop et des datastores structurés tels que des bases de données relationnelles [92].

La différence principale entre Flume et Sqoop est que:

Flume ingère uniquement des données non structurées ou semi-structurées dans HDFS.

Sqoop peut importer et exporter des données structurées à partir de SGBDR ou d'entrepôts de données d'entreprise vers HDFS ou inversement [85].

IV.3.1.1.4.9. Apache Solr & Lucene

Apache Solr et Apache Lucene sont les deux services utilisés pour la recherche et l'indexation dans Ecosystème Hadoop [85].

Apache Lucene est une technologie basé sur Java, qui aide également à la vérification orthographique, de mise en évidence des occurrences et d'analyse / création de jetons avancée. Solr est un serveur de recherche hautes performances construit autour de Lucene [93].

IV.3.1.1.4.10. Apache AMBARI

Ambari est un projet de la fondation Apache Software qui vise à rendre l'écosystème Hadoop plus facile à gérer.Il comprend un logiciel de provisionnement, de gestion et de surveillance des clusters Apache Hadoop [85].

IV.3.1.1.4.11. Apache MAHOUT

Apache Mahout est une bibliothèque fournit un environnement permettant de créer des applications d'apprentissage automatique puissante et évolutive qui s'exécute sur Hadoop MapReduce. L'apprentissage automatique est une discipline de l'intelligence artificielle qui permet aux systèmes d'apprendre en se basant uniquement sur des données, améliorant continuellement les performances à mesure que davantage de données sont traitées [94].

IV.3.1.1.4.12. Apache SPARK

Apache Spark est un Framework d'analyse de données en temps réel dans un environnement informatique distribue. Il exécute des calculs en mémoire pour augmenter la vitesse de traitement des données sur Map-Reduce. Il est 100 fois plus rapide que Hadoop pour le traitement de données à grande échelle en exploitant des calculs en mémoire et d'autres optimisations. Par conséquent, il nécessite une puissance de traitement élevée par rapport à Map-Reduce [85].

Spark présente plusieurs avantages par rapport aux autres technologies Big Data et MapReduce comme Hadoop et Storm. D'abord, Spark propose un Framework complet et unifié pour répondre aux besoins de traitements Big Data pour divers jeux de données, divers par leur nature (texte, graphe) aussi bien que par le type de source (batch ou temps-réel).

Ensuite, Spark permet à des applications sur clusters Hadoop d'être exécutées jusqu'à 100 fois plus vite en mémoire, 10 fois plus vite sur disque. Il vous permet d'écrire rapidement des applications en Java, Scala, R ou Python et inclut un jeu de plus de 80 opérateurs haut-niveau [95].

IV.3.1.1.4.12.1. Spark SQL

Spark SQL est un module Spark pour le traitement de données structurées. Il fournit une abstraction de programmation appelée DataFrames et peut également agir en tant que moteur de requête SQL distribué. Il permet aux requêtes Hadoop Hive non modifiées de s'exécuter jusqu'à 100 fois plus rapidement sur les déploiements et les données existants. Il fournit également une intégration puissante avec le reste de l'écosystème Spark (par exemple, en intégrant le traitement de requête SQL avec l'apprentissage automatique).

IV.3.1.1.4.12.2. Spark Stearming

Spark Streaming permet de puissantes applications interactives et analytiques sur les données en continu et historiques, tout en héritant des caractéristiques de facilité d'utilisation et de tolérance aux pannes de Spark. Il s'intègre facilement à une grande variété de sources de données populaires, notamment HDFS, Flume et Kafka [96].

Spark Streaming peut être utilisé pour traitement temps-réel des données en flux. Il s'appuie sur un mode de traitement en "micro batch" et utilise pour les données temps-réel DStream, c'est-à-dire une série de RDD (*Resilient Distributed Dataset*). [97]

IV.3.1.1.4.12.3. Spark Mlib

MLlib est une bibliothèque d'apprentissage automatique (*Machine Learning*) Le évolutive qui fournit à la fois des algorithmes de haute qualité (par exemple, plusieurs itérations pour augmenter la précision) et une vitesse fulgurante (jusqu'à 100 fois plus rapide que MapReduce). La bibliothèque est utilisable dans Java, Scala, R et Python dans le cadre d'applications Spark, afin que vous puissiez l'inclure dans des flux de travaux complets [96].

IV.3.1.1.4.12.4. Spark GraphX

GraphX est un moteur de calcul graphique construit sur le dessus de Spark qui permet aux utilisateurs de créer, transformer et raisonner de manière interactive sur des données structurées graphiquement à l'échelle [96].

IV.3.1.1.4.12.5. Spark Core

Spark Core est le moteur d'exécution général sous-jacent de la plateforme Spark sur lequel toutes les autres fonctionnalités sont créées. Il offre des fonctionnalités informatiques en mémoire pour assurer la vitesse, un modèle d'exécution généralisé pour prendre en charge une grande variété d'applications, ainsi que des API Java, Scala, R et Python pour faciliter le développement [96].

Spark utilise une structure de données fondamentale spécialisée appelée RDD (*Resilient Distributed Datasets*), un ensemble logique de données partitionnées sur plusieurs ordinateurs. Les RDD peuvent être créés de deux manières: la première consiste à référencer des ensembles de données dans des systèmes de stockage externes et la seconde consiste à appliquer des transformations (par exemple, carte, filtre, réducteur, jointure) aux RDD existants [95].

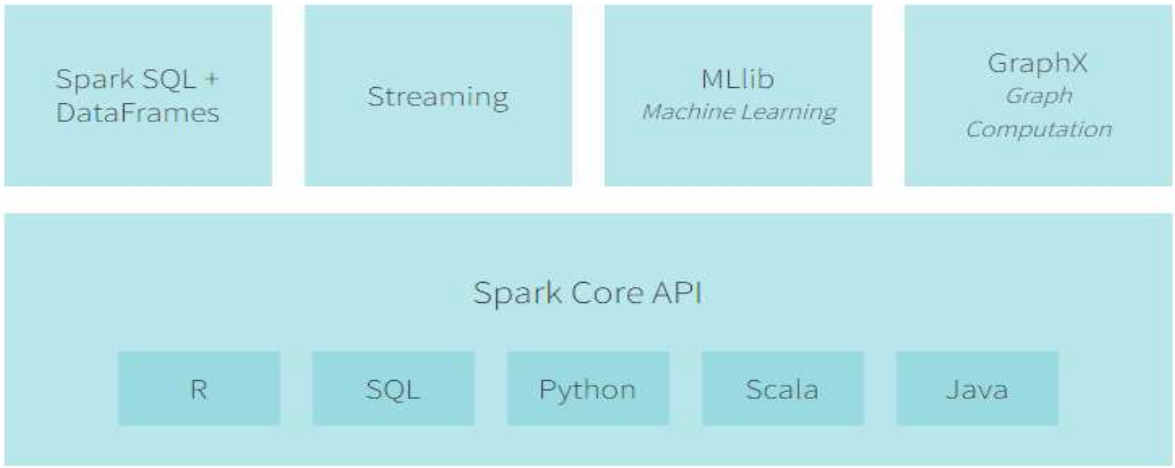


Figure IV. 7. Ecosystème Apache Spark

IV.4. Les Traitements parallèles

IV.4.1. Traitement Massivement Parallèle

C'est une stratégie pour traiter des données à grande échelle. En Anglais***(MPP)Massively Parallel Processing*** autorise chaque serveur ou nœud disposant de sa propre mémoire et de son propre disque, et ces nœuds peuvent partager la charge de travail.[98] Comparé au multitraitement symétrique (SMP), MPP permet également d'effectuer des recherches en parallèle dans de nombreuses bases de données et constitue actuellement la base de l'infrastructure de bon nombre des plus grands super calculateurs. Mais dans l'état actuel de la technologie, les systèmes MPP nécessitaient encore un coût élevé et une installation compliquée, ce que ne peuvent se permettre que les plus grandes entreprises et les organismes gouvernementaux [99].

IV.4.2. SISD (Simple Instruction Simple Data)

Un programme est un flot d'instructions sur un flot de données. Les machines monoprocesseur ou pipelines entrent également dans cette catégorie, en remarquant plus du traitement simultané que de l'exécution concurrente [100].

IV.4.3. SIMD (Simple Instruction Multiple Data)

Dans ce modèle un seul flot d'instructions et plusieurs flots de données. La même instruction est exécutée de façon concurrente par plusieurs processeurs utilisant différents flots de données. On ne dispose que d'une seule mémoire d'instructions pour tous les processeurs, mais chacun d'entre eux possède sa propre mémoire de données. Un seul processeur de contrôle gère les instructions. Les processeurs sont spécifiques à la tâche qu'ils exécutent [100].

IV.4.4.MISD (Multiple Instruction Simple Data)

Dans ce modèle plusieurs flots d'instructions et un seul flot de données correspond à des architectures composées de plusieurs unités fonctionnelles, qui travaillent chacune leur tour sur un même jeu de données, de façon synchrone [100].

IV.4.5. MIMD *(Multiple Instruction Multiple Data)*

Plusieurs flots d'instructions et plusieurs flots de données il s'agit du modèle le plus large pour exprimer le parallélisme. Chaque processeur exécute ses propres instructions et opère sur ses propres données de façon autonome [100].

IV.4.6. Unité de traitement graphique (GPU)

Les unités de traitement graphique (*GPU*) ont été initialement développées en tant que circuits électroniques spécialisés pour le traitement rapide des images et le rendu graphique. Les GPU sont aujourd'hui très utilisés pour toutes les applications générales nécessitant une puissance de calcul très performante, car leur structure hautement parallèle les rend plus efficaces que les CPU universelles pour les algorithmes dans lesquels le traitement de grands blocs de données est effectué en parallèle. Les GPU sont en train de devenir une alternative valable aux clusters de super ordinateurs classiques basés sur CPU, également pour le rapport consommation / performances optimisé et leur coût réduit [101].

IV.5. Conclusion

Le Big Data regroupe une famille des nouvelles technologies très gourmandes en ressources de calcul et des plateformes adéquates qui répondent à un triple problématique : un **V**olume de données important à traiter, une grande **V**ariété d'informations (en provenance de plusieurs sources, non-structurées, structurées) et un certain niveau de **V**élocité à atteindre - c'est-à-dire de fréquence de création, collecte, traitement/analyse et partage de ces données.

Chapitre V

Réalisation et implémentation

V. Introduction

Dans ce chapitre, nous allons aborder la partie pratique de notre projet qui consiste à décrire la conception technique de notre modèle proposé, et nous allons définir les différents outils utilisés dans notre application notamment la configuration du Apache Spark en utilisant la bibliothèque du Spark pour implémenter la méthode «les forets aléatoires » (*Random Forest*) sous Netbeans.

V.2. Versions des outils utilisés

Nous avons programmé notre application avec les outils et les APIs *(Application Programming Interface*) situé dans le tableau sous le système d'exploitation Linux Ubuntu 18.10. Le tableau suivant montre tous les outils et les APIs utilisé pour construire notre application :

Numéro	Outil	Type	Version
01	Netbeans IDE	IDE	8.2
02	Java Développement kit	Plateforme	1.8.0_111
03	Apache Spark	API	2.4.3
04	Scala	API	2.12.8

Tableau V.1. Versions des outils et APIs utilisés.

Nous avons utilisé la data-set originale connect-4 [102] avec les caractéristiques listées dans ce tableau ci-dessous :

Data-set	Type	Nombre d'individu (Taille)	Nombre de classes	Les attributs	Source
Connect-4	Classification	67557 (14.7 Mo)	3	218	UCI

Tableau V.2 Les caractéristiques de la base d'apprentissage connect-4.

Nous avons obtenu la nouvelle base Result qu'est le résultat de l'échantillonnage stratifier avec les caractéristique listées dans ce tableau ci-dessous.

Data-set	Type	Nombre d'individu (Taille)	Nombre de classes	Les attributs	Source
Result	Classification	22518 (4.9 Mo)	3	218	UCI

Tableau V.3. Les caractéristiques de la base d'apprentissage Result après échantillonnage.

Nous avons utilisé la nouvelle data-set Result qu'est le résultat de l'échantillonnage stratifie pour la sélection des variables importantes puis nous avons obtenu une nouvelle data-set nommé Result-1 avec les caractéristiques listées dans ce tableau ci-dessous.

Data-set	Type	Taille	Nombre de classes	Les attributs	Source
Result-1	Classification	22518 (4.1Mo)	3	181	UCI

Tableau V.4. Les caractéristiques de la base d'apprentissage Result-1 après la sélection des variables importante.

V.3. Le scenario de fonctionnement du système proposé

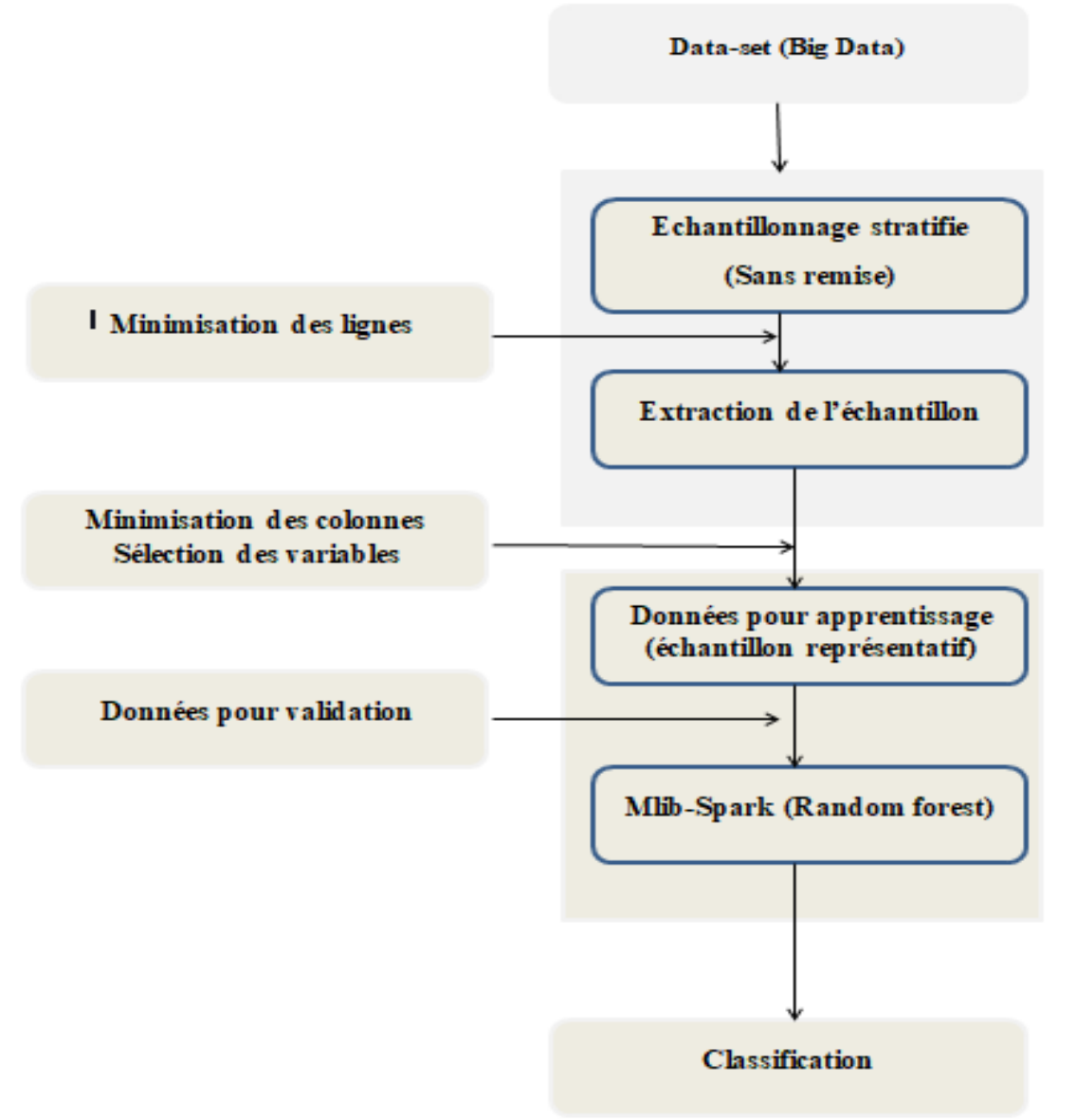


Figure V.1. Le scénario de fonctionnement du Système proposé.

Dans le troisième chapitre nous avons parlé sur Machine Learning et ses types (supervisée, non supervisée) et les méthodes de chaque type parmi les méthodes nous avons choisi la méthode (*Random Forest)* pour la multitude des réponses telles que bonnes performances en prédiction, pas de problème d'overfiting, l'évaluation de l'erreur intégrée.

D'après l'étude réalisée par [103], il a confirmé que (*Random Forest*) est le meilleur classificateur parmi les autres classifieurs, tel que (K-NN, Naïve Bayes).

Target class	Alive					Dead-prostaticca				
Accuracy measure	Spec	AUC	F1	Prec	Recall	Spec	AUC	F1	Prec	Recall
Naive Bayes	0.0338	0.0135	0.083	0.125	0.2025	0.98	0.9397	0.1286	0.9	0.0583
Random Forest	0.5896	1	1	0.7863	0.0738	1	0.9985	0.7789	0.7161	0.0692
K-NN	0.3783	0.6149	0.833	0.6968	0.3783	0.99	0.994	0.9692	0.9692	0.612
Target class	Dead-respiratory disease					Dead-cerebrovascular				
Accuracy measure	Spec	AUC	F1	Prec	Recall	Spec	AUC	F1	Prec	Recall
Naive Bayes	0.9856	0.9397	0.083	0.125	0.1625	1	0.9397	0.1176	0.6667	0.6450
Random Forest	1	0.9985	1	1	1	1	0.9985	0.9688	0.9394	1
K-NN	0.9956	0.994	0.879	0.8213	0.6896	1	0.994	0.4878	1	0.6926

Tableau.V.4. Représentation des classes cibles.

V.4. Présentation de l'application

V.4. 1. L'authentification

La figure ci-dessous présente l'authentification d'un utilisateur, elle demandera de l'utilisateur introduire leur nom et leur mot de passe afin de commencer à utiliser l'application.

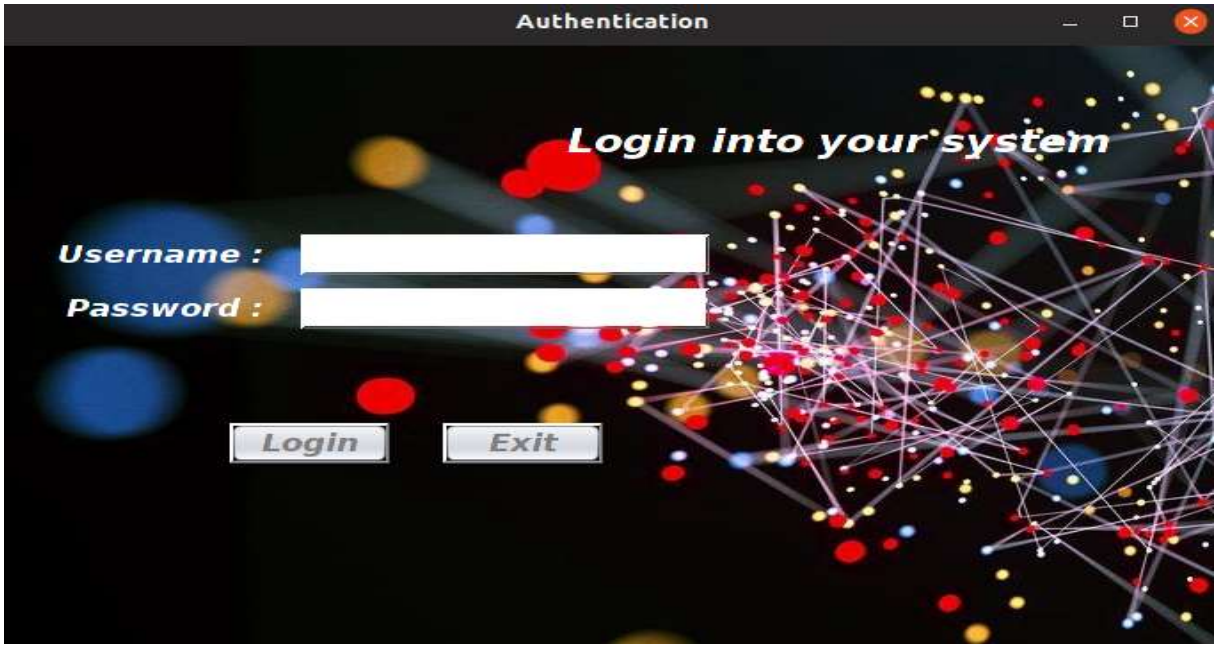


Figure V.2. Authentification.

V.4.2. L'interface principale

L'interface de notre system contient d'un menu principal, composé de sous menus dont chacun eux correspond à une interface.

La figure ci- dessous présente l'interface principale de notre system :

FigureV.3.L'nterface principale.

La figure suivante illustre l'exécution de la première étape de notre application :

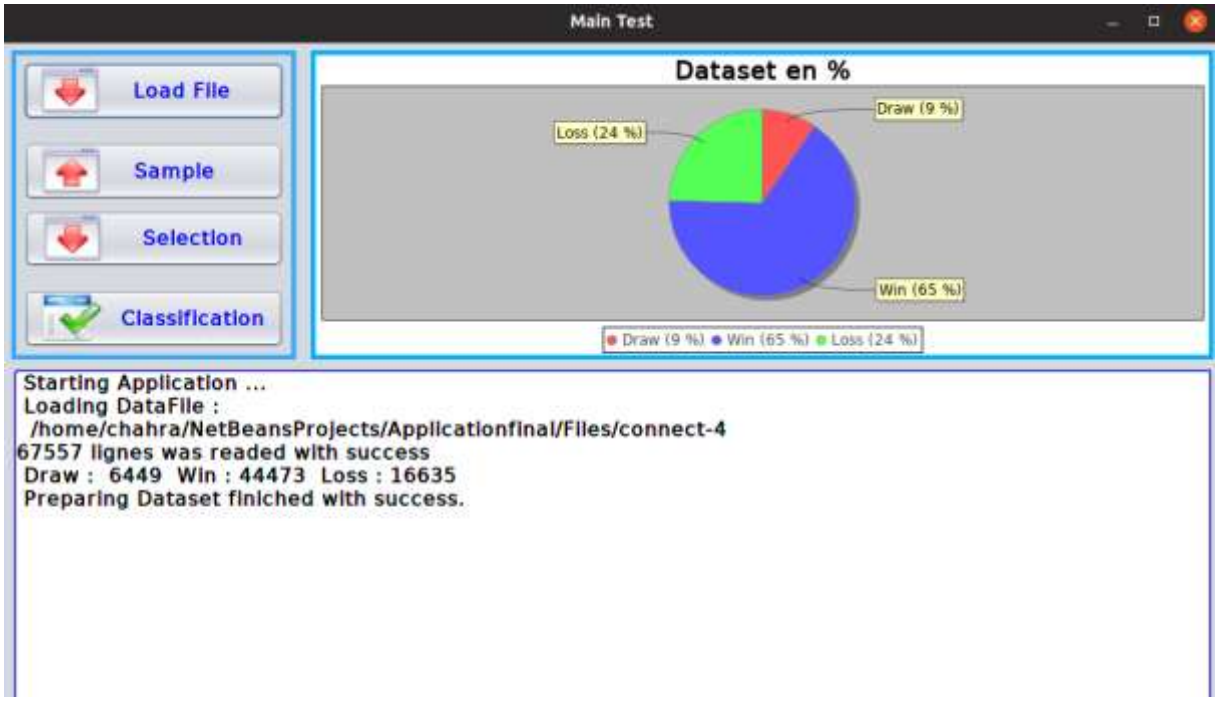


Figure.V.3 Chargement de data-set originale d'apprentissage (connect-4).

La figure suivante illustre l'exécution de la deuxième étape de notre application, qui est l'échantillonnage stratifie (Sampling).

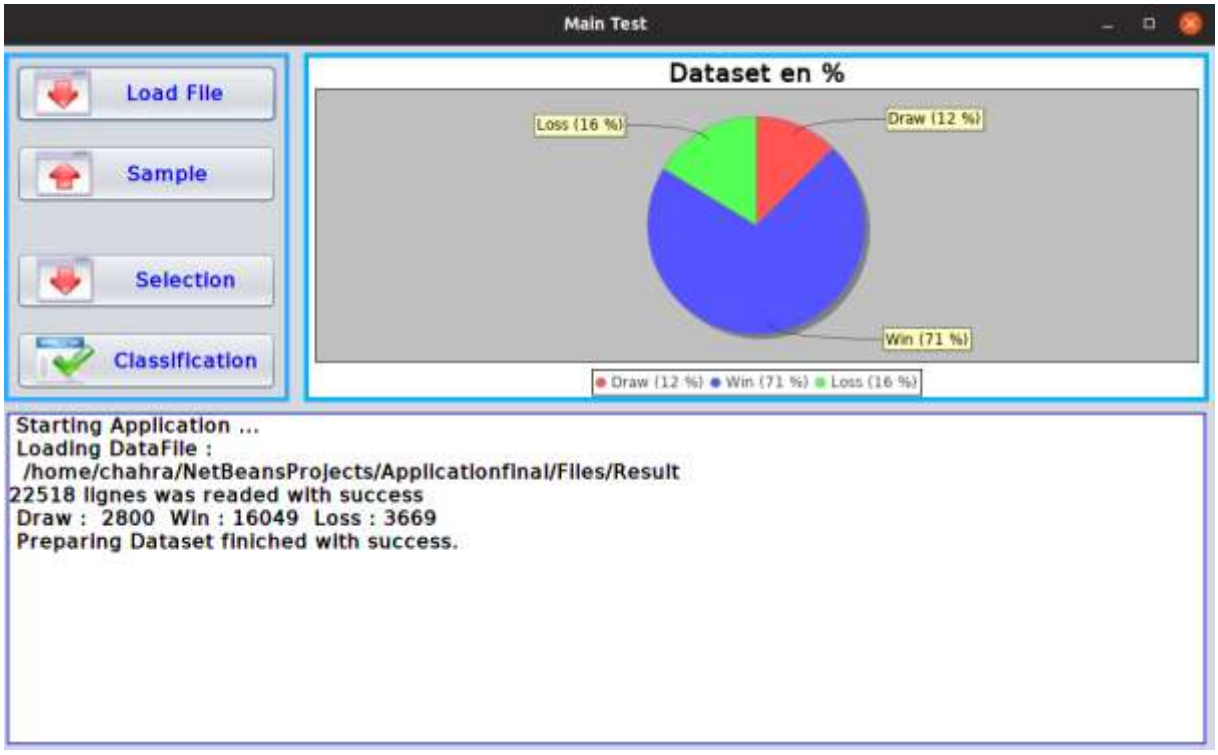


Figure V.4. Chargement de data-set d'échantillonnage stratifie (Result).

La figure suivante illustre l'exécution de la troisième étape de notre application, qu'est la sélection des variables importante (Selection).

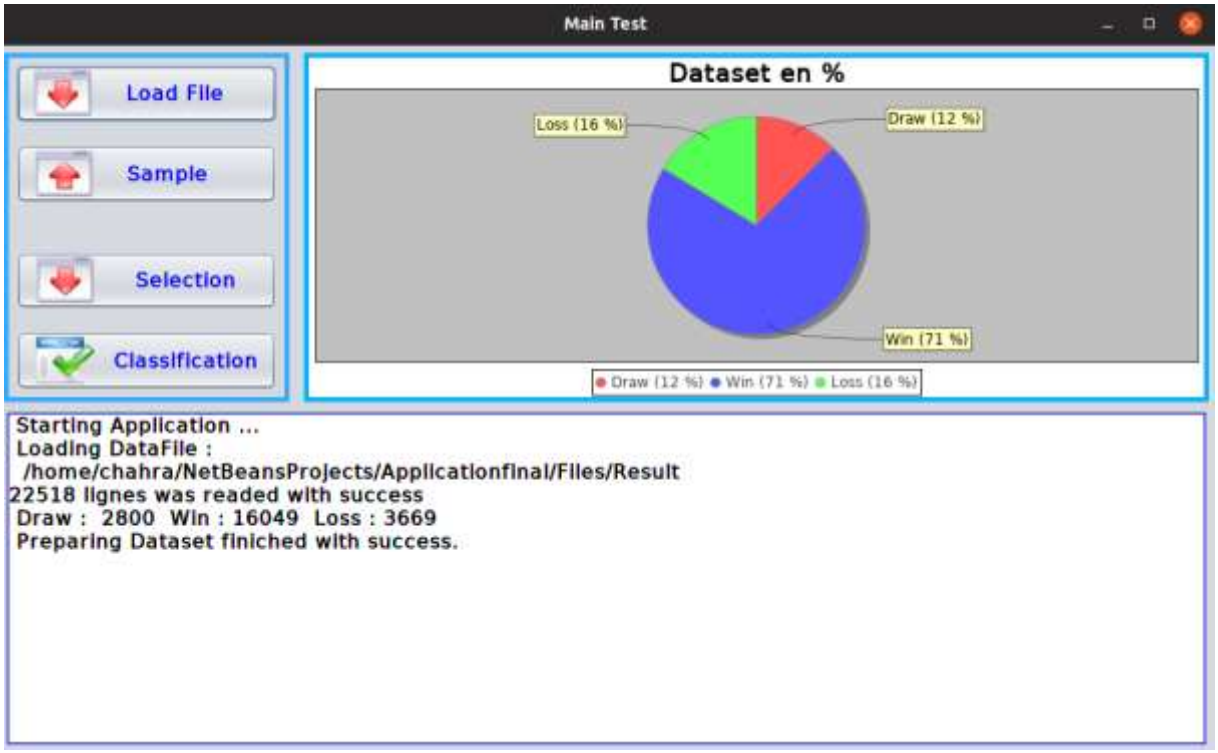
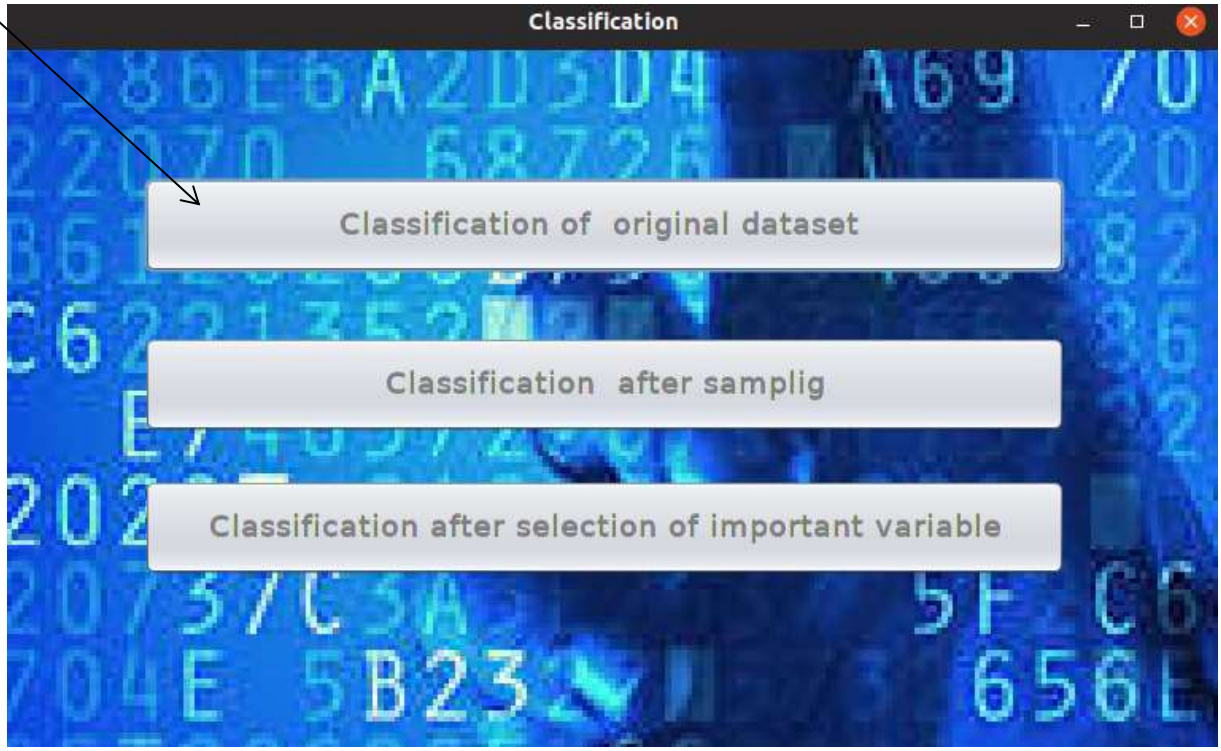


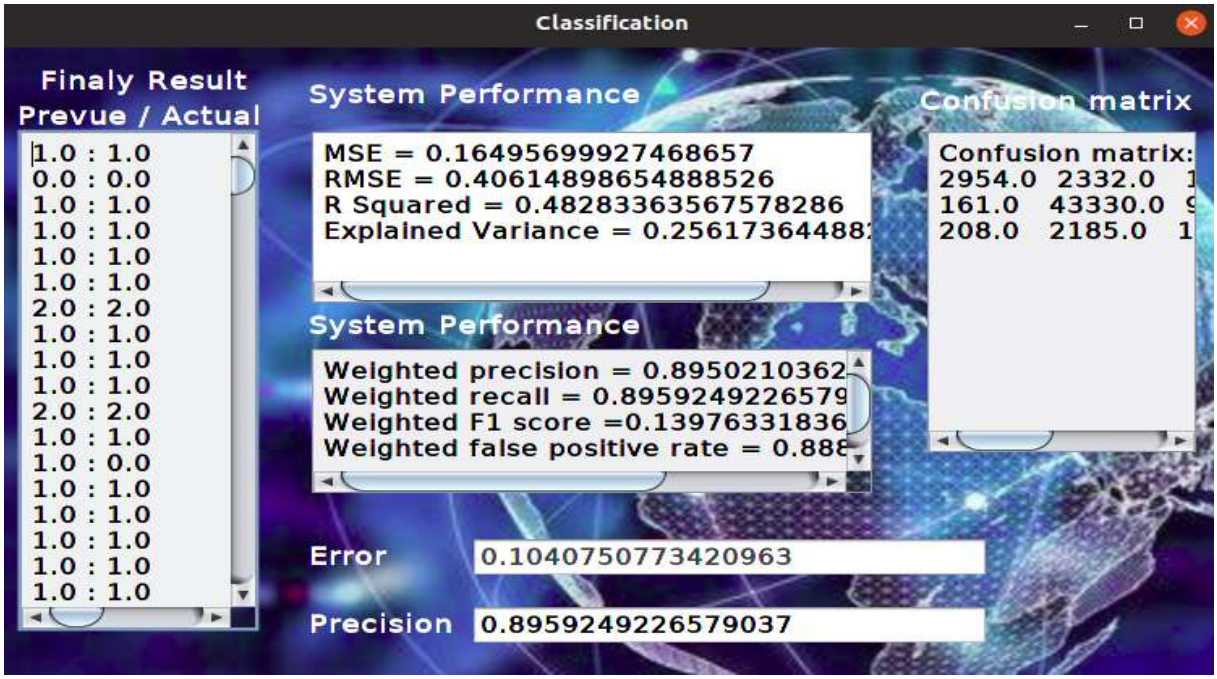
Figure.V.5 Chargement de data-set de sélection (Result-1).

Le bouton **classification** est pour choisir le data-set à classifier :



FigureV.6. Chois de data set pour classification.

Le résultat obtenu après le choix du data set figuré ci-dessous :



FigureV.7. Le résultat final de la prédiction.

V.5 Performance du système

V.5.1 Précision

C'est le rapport entre le nombre de vrais positifs et la somme des vrais positifs et des faux positifs.

V.5.2 Correctly Classified Instances

Le nombre d'individus bien classés, en valeur absolue, puis en pourcentage du nombre total d'instances.

V.5.3 Incorrectly Classified Instances

Sous le même format, le nombre d'instances mal classées.

V.5.4 Root mean-squared error

Cette mesure d'erreur concerne principalement les prédicteurs.

Racine carrée de l'erreur quadratique moyenne : avec les mêmes notations que ci-dessus, elle correspond à :

$$\sqrt{\frac{(p_1-a_1)^2+\cdots+(p_1-a_1)^2}{n}}$$

V.5.5 Mean absolute error

Erreur absolue en moyenne : pour chaque exemple, on calcule la différence entre la probabilité (calculée parle classifieur) pour un exemple d'appartenir à sa véritable classe, et sa probabilité initiale d'appartenir `à la classe qui lui a été fixée dans l'ensemble d'exemples (individus) (en général, cette probabilité vaut 1). On divise ensuite la somme de ces erreurs par le nombre d'instances dans l'ensemble d'exemples.

$$\text{mean absolute error} = \frac{|p_1-a_1|+|p_2-a_2|+..+|p_n-a_n|}{n}$$

V.5.6 Relative absolute error

Cette mesure d'erreur concerne principalement les prédicteursErreur absolue relative : le nom paraıt très mal choisi.

On compare l'erreur absolue avec l'erreur absolue d'un prédicteur très simple, qui retournerait toujours la valeur moyenne des ai, soit$\bar{a}=\frac{1}{n}\sum_i a_i$

$$\frac{|p_1-a_1|+\ldots+|p_n-a_n|}{|a_1-\bar{a}|+..+|a_n-\bar{a}|}$$

V.5.7. Root relative squared error

Cette mesure d'erreur concerne principalement les prédicteurs. Racine carrée de l'erreur quadratique relative : rapport entre l'erreur quadratique et ce que serait l'erreur quadratique d'un prédicteur qui retournerait toujours la valeur moyenne :

$$\sqrt{\frac{(p_1-a_1)^2+\cdots+(p_n-a_n)^2}{(a_1-a)^2+\cdots+(a_n-a)}}$$

V.5.8 Les mesures d'exactitude par classe

Pour une classe donnée, un classifieur, et un exemple, quatre cas peuvent se présenter :

1. L'exemple est de cette classe, et le classifieur ne se trompe pas : c'est un **vrai positif.**

2. L'exemple est de cette classe, mais le classifieur se trompe : c'est un **faux négatif.**

3. L'exemple n'est pas de cette classe, mais le classifieur la lui attribue quand même : c'est **faux positif**.

4. L'exemple n'est pas de cette classe, et le classifieur ne le range pas non plus dans cette classe : c'est un **vrai négatif.**

- **TP Rate**

Rapport des vrais positifs. Il correspond à :

$$\frac{Nbr\ de\ vrqi\ positifs}{(Nbr\ de\ vrais\ positif + Nbt\ de\ faux\ nègatifs} = \frac{Nbr\ de\ vrais\ positifs}{Nbr\ d'exemplede\ cette\ classe}$$

, entre le nombre de bien classé et le nombre total d'éléments qui devraient être bien classes.

- **FP Rate**

Rapport des faux positifs. Il correspond à :

$$\frac{Nbr\ de\ faux\ positifs}{(Nbr\ de\ faux\ positifs + Nbr\ de\ vrais\ nègatifs} = \frac{Nbr\ de\ faux\ positifs}{(Nbr\ d'exemplen'etantpas\ de\ cette\ classe)}$$

La donnée des taux TP Rate et FP Rate permet de reconstruire la matrice de confusion pour une classe donnée.

V.6. Les graphes

Il existe corrélation inverse entre la précision et l'erreur : si la précision est élevée, alors l'erreur est diminuée, cela implique que l'inexactitude est vraie.

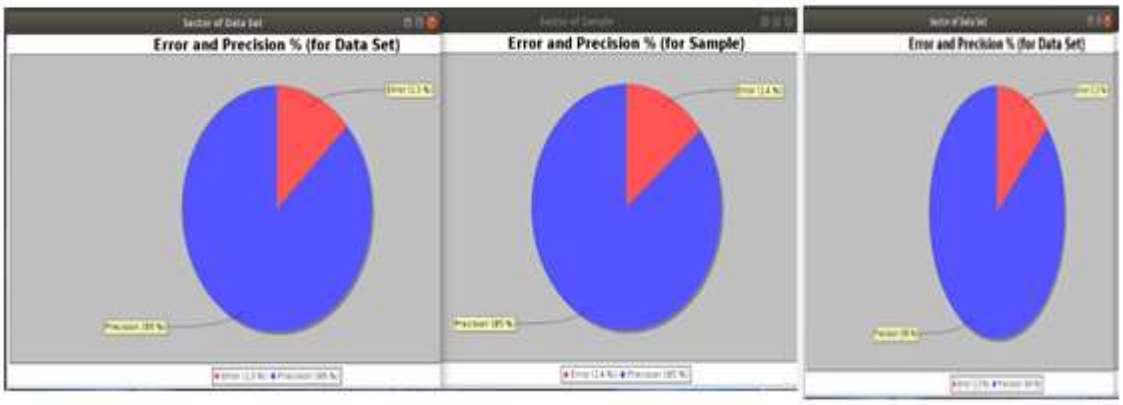


Figure V.6. Comparaison entre Data-Set original, l'échantillon représentatif et la sélection des variables importantes.

V.7. Synthèse

Nous voyons maintenant une comparaison concernant l'analyse prédictive du Big data notamment le temps d'exécution, pour cela on va tester notre application en utilisant deux outils : Weka and Spark, et les résultats obtenus sont récapitulés dans le tableau (05) ci-dessous :

Outils (Framework)	Nom & Taille du Data-Set (Mo)	Temps d'exécution (S)
Spark	Echantillon représentatif (4.9 Mo)	53 seconds
Weka	Titanic (4.89 Mo)	9 munit

Tableau V.5. La comparaison entre weka et spark.

V.8 Discussion

D'après les résultats obtenus dans la partie implémentation du troisième chapitre nous déduisons deux points crucials :

Le premier point, selon la figure (Figure 4) nous déduisons que l'échantillon extrait à partir du data-set original (la base d'apprentissage globale) est vraiment représentatif, car elle donne un résultat de prédiction presque égal que celui trouvé par data-set original (0.01 de déférence de précision) et selon la figure (Figure 5) nous déduisons que les variables importante sélectionnées a partir du det-set obtenu après l'échantillonnage stratifie (la base d'apprentissage result) est vraiment représentatif, car elle donne un résultat de prédiction presque égal que celui trouvé par data-set original (0.02 de déférence de précision les résultats obtenus sont récapitulés dans le tableau (06) ci-dessous :

Outils (Framework)	Taille (Mo)	Précision	Temps d'exécution (S)
Data-Set (base globale)	14.7 Mo	86 %	1 minute et 23 seconds
Apres l'échantillonnage	4.9 Mo	85 %	53 seconds
Après la sélection	4.4 Mo	83%	1 minute et 45 second

Tableau V.6. La comparaison entre Data set et Echantillon.

Dans le deuxième point, et suivant le tableau (05) nous conclurons que le Framework Spark donne des résultats de l'analyse prédictive (classification) dans le meilleur délai (53Second).

V.8. Conclusion

Nous avons présenté dans ce chapitre, le système proposé et leur scenario de fonctionnement, ce système basé sur le modèle prédictif d'apprentissage supervisé « la méthode des forets aléatoires », il est configuré dans la solution du Big Data Spark pour le traitement parallèle de données massives, avec l'utilisation du RandomForest pour l'analyse prédictive de données.

On peut dire maintenant que notre system proposer a résolues les problèmes suivant :

- Le premier problème du volume en utilisant la méthode de l'échantillonnage stratifie

Ainsi que la sélection des variables importante.

- Le deuxième problème de la véracité : est que notre system garde presque a même précision de la prédiction du data set original et finalement le troisième problème de vélocité en utilisant le system de calcul distribuer e parallèle « spark ».

Conclusion générale

Conclusion générale

L'analyse prédictive de données massives est un moyen pour prévoir les probabilités futures avec un niveau de fiabilité acceptable afin de prendre des décisions efficaces dans le plus bref délai.

Dans notre travail, nous avons présenté la méthode d'échantillonnage stratifie sans remise pour obtenir un meilleur résultat de classification à partir de l'échantillon représentatif, cet échantillon permet de faire le calcul parallèle.

Par ailleurs, nous avons utilisé les solutions du big data tel que le Spark afin de communiquer les résultats en streaming.

Finalement, à partir des résultats obtenus dans la partie expérimentation que ce soit la précision ou bien le temps d'exécution, on peut dire que notre travail proposé a résolu les problèmes de l'analyse prédictive Big Data notamment le volume et la vitesse.

Références Bibliographiques

Référence Bibliographique

[1] Jérémie, B, Adrien, R. (2016). Big Data, Hadoop, MapReduce – Introduction pour statisticien non-initié, Université de Toulouse.

[2] Jean, P. (2018).Généralité sur les données massives : BIG DATA disponible sur le site web de supinfo. Récupuré de : https://www.supinfo.com/articles/single/6676-generalite-donnees-massives-big-data.

[3] Gartner, IT Glossary. Récupéré de : http://www.gartner.com/it-glossary/big-data/.

[4] Olivier. J. (2013). Présentation Générale de Big Data Guide Share France.

[5] Emmanuel. F. (2015). Les particularités des projets liés au « Big Data ».

[6] Boukharta. M. (2018). Les enjeux du Big Data dans le développement durable des territoires touristiques (Master). Université Jean Jaures, Toulouse. Récupère de : http://www.isthia.fr/core/modules/download/download.php?memoires_id=669

[7] Firican, G. (2017). The 10 Vs of Big data.disponiblesur le site webhttps://tdwi.org/articles/2017/02/08/10-vs-of-big-data.aspx.

[8] HADJARI, I, BENBACHIR,M and all. (2017). Big DATA : Conceptions, architectures, Fonctionnements et applications (MASTER). Université Abou BakrBelkaid, Tlemcen.

[9] Nawsher, K, Habib, S, Gran, B, Aftab, A, Abbasi, M, Alsaqer, S, Salehian. (2018). The 10 Vs, Issues and Challenges of Big Data, International Conference on Big Data and Education ICBDE '18, Honolulu, HI, USA.

[10] Angeline, K. (2013). Big Data (rapport de stage) (Master). Insa Lyon. Recuperate de: https://www.memoireonline.com/05/14/8890/m_Big-data-rapport-de-stage.html

[11].Lyman, H. Varian, J. Dunn, A. Strygin, and K. Swearingen. (2000). How much information? Counting-the-Numbers, vol. 6, no. 2.

[12] Zikopoulos, C, Eaton, D. DeRoos, T. Deutsch, and G. Lapis. (2011). Understanding Big Data. McGraw-Hill, USA.

[13] Hota,K , and D. Madam Prabhu.(2012) No problem with Big Data. What do you mean by Big?, Journal of Informatics, pp. 30–32.

[14] DBTA. (2013), Big Data Source book, Unisphere Media.

[15]Chen, M, Mao, S, Liu, Y. (2014). Big Data: a survey. MOB. NEtw. Appl. 19 171-209.

[16] Basel, K, David, Kt, and Steve, V. (2013). The big-data revolution in us health care: Accelerating value and innovation. Mc Kinsey & Company, 2(8):1–13.

[17] Chan, J. (2013).An architecture for big data analytics, Communications of the IIMA, 13(2), 1–13.

[18] Zikopoulos, P. C, deRoos, D., Parasuraman, K., Deutsch, T., Corrigan, D., and Giles, J.(2013) . Harness the Power of Big Data: The IBM Big Data Platform. New York:

McGraw-Hill.

[19]Mike, F. (2013), Title: "Enterprise Information Protection- The Impact of Big Data",IBM.

[20] IBM (2014), Title: "The top five ways to get started with big data".

[24] Syed, M. (2018). Methode of data collection.

[25] Mayank, J. (2019). Tools used to collect Data. Rècupèré de : https://www.theclassroom.com/

[26] Margruet, R. (2016). Data Collection. Récupéré de : https://searchcio.techtarget.com/definition/data-collection

[27] Jean. Christophe, V. (2007). Méthodologie de l'enquête par questionnaire. Laboratoire Culture & Communication Université d'Avignon.

[28] Sincero, S. (2014). Surveys and Questionnaires-Guide. Rècupèré de : https://explorable.com/surveys-and-questionnaires

[30]Justin, D. (2018). Les différents types d'entretiens. Récupéré de : https://www.scribbr.fr/memoire/types-entretiens/

[31] T, T. (2017). Entretien ou questionnaire ? Quelle méthode de collecte de données pour son mémoire ? Récupéré de : https://arlap.hypotheses.org/8170

[33] Kheddouchi, I. (2017). Qu'est-ce que la 'Data préparation' et a quoi sert-elle ? Récupéré de : https://blogbi.asi.fr/2017/06/27/QUEST-CE-QUE-LA-DATA-PREPARATION-ET-A-QUOI-SERT-ELLE/

[35]Margarter, R. (2016). Datawarehouse, Bases de données, Modélisation de donnée. Récupéré de : https://www.lemagit.fr/definition/Modelisation-de-donnees.

[37] Margruet, R. (2016). Data visualisation. . Récupéré de : https://searchcio.techtarget.com/definition/data-collection

[44] Brook, C. (2019) : Data protection 101, whatis Data Integrity ? Definition, Best Practice & more. Recapture de: https://digitalguardian.com/blog/what-data-integrity-data-protection-101?fbclid=IwAR1phaAjrXv_QhhZVkDoFFoHA3jKuCmCG1bQxopZqzf16jptH9xwhlYHvg

[47] People Vox. (2019).Analyse Données, Analyse descriptive de données. Consulté le : 21 Mars 2019 de : http://www.analyse-donnees.fr/

[49] Bastien.L. (2016).Analyse prédictive, définition et secteurs d'application. Consulté le : 21 Mars 2019 de : https://www.lebigdata.fr/

[50] DKT Group. (2016). Big data, Les six étapes de l'analyse prédictive. Consulté le : 21 Mars 2019 de : https://dkt-group.cm

[52] KHERRI.A. (2014). Statistique de Gestion, Echantillonnage. Consulté le 15 Avril 2019 de : www.sg-ehec.jimdo.com

[53] BATHELOT.B. (2017). Définition : échantillonnage étude. Récupéré de : https://www.definitions-marketing.com/definition/echantillonnage-etude/

[55] HAMADI.F, MAZ.S. (2018). ANALYSE PREDICTIVE DU BIG DATA-Sampling and Streaming (Master 2). Université IBN Khaldoun, Tiaret.

[58] Saleema, J, Bhagawhi, N, Venugopal, K, and all. (2014) cancer prognosis prediction using balanced stratified sampling.Vol.3, No. 1.

[60] LEPINE.B. (2018). Data Analytics, Intelligence artificielle. Récupère de : https://www.lebigdata.fr/.

[61] Metom, J bertrand, R. (2017). Machine Learning : Introduction à l'apprentissage automatique. Consulté le : 19 février 2019 de : https://www.supinfo.com/fr/Default.aspx

[62] Ait Mohammed, F. (2018). Approches d'apprentissage automatique pour la détection du SPAM WEB : Exploration de diverses caractéristiques (Mémoire présenté comme exigence partielle de la maitrise en informatique). Université du Québec, Montréal.

[63] Zimmer, M. (2018). Intelligence artificielle, Apprentissage par renforcement développemental .Université de Lorraine, France.

[64] JULLIEN.S. (n.d). L'apprentissage par renforcement. Récupéré de : https://dataanalyticspost.com/ .

[65] Gaël.B. (2017). 8 Machine Learning Algorithmsexplained in Humanlanguage. Retrieved April 17, 2019, fromhttps://www.linkedin.com/pulse/8-algorithmes-de-machine-learning-expliqu%C3%A9s-en-humain-ga%C3%ABl

[66] Benzaki, Y. (2017). 9 Algorithmes de Machine Learning que chaque Data Scientist doit connaitre. Consulté le 18 Avril 2019 de : https://mrmint.fr/9-algorithmes-de-machine-learning-que-chaque-data-scientist-doit-connaitre

[69] Micheal, W. (2017). Batche vs Real Time Data Processing. Récupéré de : https://www..datasciencecentral.com/.

[70] Mehdi, R. (2018). Datalake, lambda, kappa, trois architectures Big Data incontournables. Récupéré de : http://www.cyres.fr/

[71] Nahtan, M. & James, w. (2015). Big Data: Principles and Best Practices of Scalable Real-time Data Systems, Manning. Récupéré de : http://index-of.co.uk/Big-Data-Technologies/

[72] Jim, S. (2015). Zeta Architecture: Hexagon is the new circle. Récupéré de :https://www.oreilly.com/ideas/zeta-architecture-hexagon-is-the-new-circle

[73] Kumar, C. (2016). What is SMACK (Spark, Mesos, Akka and Kafkaf) ?. Récupéré de : https://bigdata-madesimple.com/smackspark-mesos-akka-kafka/

[74] Jay, K. (2014). Questioning the Lambda Architecture - Kappa architecture. Récupéré de : https://www.oreilly.com/ideas/questioning-the-lambda-architecture

[75] Matallah, H. (2018). Vers un nouveau modelé de stockage et d'accès aux données dans les Big Data et les Cloud Computing (Doctorat en science). Université Abou-Baker Balkaid. Tlemcen.

[76] Thierry, M. (2010). Du cluster à la grille sous l'angle de la performance. Laboratoire d'Analyse et d'Architecture des Systèmes du CNRS. Récupéré de : http://tel.archives-ouvertes.fr/tel-00547021/document.

[77] Michael, B. (2014). Tutoriel d'introduction a apache Hadoop. Récupéré de : https://mbaron.developpez.com/tutoriels/bigdata/hadoop/introduction-hdfs-map-reduce/

[78] Shubham, S. (2019). Hadoop écosystème :Hadoop touls for crunching Bid Data. Recuperate de: https://www.edureka.co.

[79] Zoiner, T. (2017).Mastering Azure Analytics (1st, ed). O'Reilly Media, Inc. ISBN: 9781491956649.

[80] Peter, M, Timothy, G. (2011). The INST definition Of cloud computing (DRAFT) Recommendation of national ins tutu of standards an technologie

[81] Rachmi, J. (2017). What are the main componement of a hadoop application ? Récupéré de : https://www.quora.com/

[82] Saeed, S, Saeed, J. (2017). Beyound Batche processing :Towards Real-Time and Stearming Big Data. Récupéré de :https://arxiv.org/abs/

[83] Cristophe, P. (2016). Big Data : Panorama des solutions 2016 Récupéré de : https://blog.ippon.fr/2016/03/31/big-data-panorama-des-solutions-2016/

[84] Bastien, L. (2018). Hadot-tout savoir sur la principale platforme Big Data. Récupéré de : https://www.lebigdata.fr/HADOOP

[85] Khauchik, P. (2015). Zeta Architecture_ A Solution to Integrate Data into Business

[89]Shubham, S. (2019).Pig Tutorial : Apachepig Architecture & Twitter case study. Récupéré de : http://www.edureka.co/blog/pig-tutorial/#architecture.

[90]Ashich, B. (2019). Hive Tutorial-Hive Architecture and NASA case study. Récupéré de : http://www.edureka.co/blog/hive-tutorial/

[91] Shubham, S. (2019). Apache Flume Tutorial-Twiter Data Streaming. Récupéré de : http://www.edureka.co/blog/apache-flume-tutorial/

[97]Sirini, p. (2015). Traitements Big Data Avec Apache Spark-1er partie: Introduction. Récupéré de : https://www.infoq.com/fr/articles/apache-spark-introduction/

[98] Alexey, G. (2019) Distributed systems architecture. Récupéré de : https://0x0fff.com/hadoop-vs-mpp/

[99]Edmondson, G. (2012). "Massively parallel processing and the parallel data warehouse," Récupéré de: https://garrettedmondson.wordpress.com/]

[100] Ezio, B.GPU Computing & Architecture. Université of Tübingen. Récupéré de : http://www.eziobartocci.com/gpucomputing2015_tubingen.php

[101]Gabriel, N. (2013). Un environnement parallèle de développement haut niveau pour les accélérateurs graphiques :mise en œuvre à l'aide d'open (Doctorat). Université de Reims Champagne-Ardenne. Récupéré de : https://tel.archives-ouvertes.fr/tel-00822285/document

[103] Chauhan R, Kaur H. (2017) .Chang Advancement and applicability of classifiers for variant

exponential model to optimize the accuracy for deep learning, Received: 24 May 2017, Accepted: 2 August 2017, Springer-VerlagGmbH Germany .

Web graphique

[21] https://whatis.techtarget.com/fr/definition/preparation-des-donnees Consulté le 19 Mars 2019.

[22] https://fr.talend.com/resources/what-is-data-preparation/. Consulté le 19 Mars 2019.

[23] https://www.tbs-sct.gc.ca/cee/pubs/meth/pem-mep04-fra.asp . Consulté le 4 fèvrie 2019

[29] https://tpelesbigdata.wordpress.com/2016/01/15/partie-ii/. Consulté le 11 Mais 2019

[32] https://www.sisense.com/glossary/data-preparation/. Consulté le 06 avril 2019

[34] https://www.lemagit.fr/definition/Modelisation-de-donneesConsulté le 03Avril 2019.

[36] Sas https://sas-data-visualization-ebooklet-fr.pdfConsulté le 04 Avril 2019.

[38]https://www.sas.com/en_us/insights/big-data/datavisualization html?fbclid=IwAR0f8Z8V_guCr4WJhInibdVmo6_ffInr0sr6eL5yM3JtgFy18XlXyIO-rU. Consulté le 07 Avril 2019. Consulté le 12 Mai 2019

[39]https://www.omnisci.com/learn/resources/datavisualization?fbclid=IwAR2VXb8MHvuluGWZAGEI04WLS2AMz0sGvL0KeVS3lgXQslJXdCaaOPQqJcs. Consulté le 19 Avril 2019.

[40] https://hackernoon.com/what-is-data-visualization-definition-history-and-examples-e51ded6e444a?fbclid=IwAR2bN0HzvW2qJQ2bPxNQmhqUispuHpjT-7AmjVGdx0c3myyBEBgX84fxenI. Consulté le 21 Avril 2019.

[41] https://www.forcepoint.com/cyber-edu/data-security?fbclid=IwAR2vD_1y2EGAbEohglYJysFFuQML4UyOZUfaQO9eRz5YesRuRjohWa8ptf8. Consulté le 19 Avril 2019.

[42]https://www.forcepoint.com/cyberedu/datasecurity?fbclid=IwAR2vD_1y2EGAbEohglYJysFF uQML4UyOZUfaQO9eRz 5YesRuRjohWa8ptf8. Consulté le 05 Mai 2019.

[43] https://www.inovera.fr/solutions-dematerialisation/securisation-donnees-informatiques/ Consulté le 02 Mai 2019.

[45]https://www.talend.com/resources/what-is-data-integrity/?fbclid=IwAR3u6Zn_ xnRE55EncMCvYK_ZBypb2cbJZ4i69LsNKgShrIcdKXYCsNYLA-A8 . Consulté le 05 Mai 2019.

[46]https://www.talend.com/resources/what-is-dataintegrity/?fbclid=IwAR3u6Zn_xnRE 55EncMCvYKZBypb2cbJZ4i69LsNKgShrIcdKXYCsNYLA-A8.Consulté le 02 Mai 2019.

[48] https://www.aubay.com/wp-content/uploads/2016/07/Analyse-de-donnes-VF.pdf Consulté le 1 juin 2019

[51] https://halobi.com/2016/07/descriptive-predictive-and-prescriptive-analytics-explained/ consulté le : 22 février 2018.

[54] https://fr.khanacademy.org/math/statistics-probability/designing-studies/sampling-methods-stats/a/sampling-methods-review .consulté le : 2 Mai 2019

[56] Statistiques Canada- Méthodes d'échantillonnage. (2017). Consulté le : 03 Avril 2019 de : https://www150.statcan.gc.ca/n1/edu/power-pouvoir/ch13/5214895-fra.htm#tphp

[57] http://uis.unesco.org/fr/glossary-term/echantillonnage-stratifie consulté le : 19 Mai 2019.

[59] https://www.dissertationsenligne.com/Divers/La-collecte-des-donn%C3%A9es-et-l'analyse-univari%C3%A9e-des/23524.html consulté le : 21 Mai 2019.

[67] https://www.geeksforgeeks.org/clustering-in-machine-learning/ . Consulté le 16 Mai 2019.

[68] https://openclassrooms.com/fr/courses/4379436-explorez-vos-donnees-avec-des-algorithmes-non-supervises/4379551-decouvrez-l-interet-des-algorithmes-de-clustering. Consulté le 16 Mai 2019.

[86] https://oozie.apache.org/ . Consulté le 30 Mars 2019

[87] https://zookeeper.apache.org/ Consulté le 4 avril 2019

[88] https://www.ibm.com/analytics/hadoop/hbase . Consulté le 5 juin 2019

[92] https://sqoop.apache.org/ Consulté le 28 avril 2019

[93] https://lucene.apache.org/ . Consulté le 16 Mai 2019

[94] https://mapr.com/products/product-overview/apache-mahout/ Consulté le 30 mai 2019

[95] https://www.tutorialspoint.com/apache_spark/apache_spark_core_programming.htm https://zookeeper.apache.org/ Consulté le 22 juin 2019

[96] https://databricks.com/spark/about . Consulté le 06 Mars 2019

[102] Connecter-4 disponible sur http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/ . Consulté le 11 janvier 2019