REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
MINISTERE DE L'ENSEIGNEMENT SUPERIEUR
ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE IBN KHALDOUN – TIARET

FACULTE DES MATHEMATIQUES ET DE L'INFORMATIQUE

DEPARTEMENT INFORMATIQUE

Mémoire

Présenter pour l'obtention du diplôme de

Master 2 en Informatique

THEME

Application des Big Data : Analyse des réseaux sociaux

Présenté par :
OUILEM Abdelhafidh
AISSAT Mohamed Abbes

Dirigé par : Mr. ZIOUAL Tahar

Année universitaire : 2016 – 2017

DEDICACE

A les deux qui s'ont toujours sacrifié pour me voir

Réussir, que Dieu le garde dans son vaste paradis , à mes Parents que j'adore. A ma grand mére.

A mes chers amis Qwider et Houssine Que dieu nous garde toujours unis .

A mon binôme Mohamed .

A tous mes collègues de la promontion Master 02 .

A toute personne ayant aidé, de prés ou loin, à la realisation de ce travail .

OUILEM ABDELHAFIDH

Dédicace

Je dédie ce mémoire

A mes chers parents ma mère et mon père

Pour leur patience, leur amour, leur soutien et leurs encouragements.

A mes soeurs et mes camarades.

A ma petie Nour el Houda spécialement .

Sans oublier tout les professeurs que ce soit du primaire, du moyen, du secondaire ou de l'enseignement supérieur.

Aissat Mohamed Abbas

Remerciement

Toute d'abord, ce travail ne serait pas aussi riche et n'aurait pas peu avoir le jour sans l'aide et l'encadrement de Mr ' ZIOUAL Tahar ', on lui remercie pour la qualité de son encadrement exceptionnel, pour sa patience, sa rigueur et sa disponibilité durant notre préparation de ce projet.

Nos remerciements s'adressent également à tous nos professeurs pour leurs générosités et la grande patience dont ils ont fait preuve malgré leurs charges académique et professionnelles

Tous ceux que nous avons omis de citer ici, et qui de près ou de loin ont contribué au bon déroulement de notre formation.

Résumé

L'utilisation croissante des réseaux sociaux à résulté dans de gigantesques masses de données générées rapidement et continuellement, l'analyse de ces données offre aux entreprises le moyen d'extraire des informations leur permettant d'avoir un avantage certain concernant les préférences et les intérêts du public. Les technologies Big Data se sont avérées particulièrement adaptées pour l'analyse des réseaux sociaux. Notre objectif est de réaliser un outil basé sur les technologies Big Data pour analyser les données générées par les utilisateurs du réseau social *Twitter.*

Mots clés: Big Data, réseaux sociaux, Big Data social, analyse des réseaux sociaux

Abstract

The increasing use of social networks has resulted in huge masses of data being generated rapidly and continuously, analyzing these data offers companies the means to extract information allowing them to have a definite advantage over preferences and public interests. Big Data technologies have proven to be particularly suitable for analyzing social networks. Our goal is to realize a tool based on Big Data technologies to analyze the data generated by users of social network Twitter.

Keywords: Big Data, social media, social Big Data, social media analysis

Sommaire

Liste des figures

Liste des tableaux

Introduction générale

De nos jours les réseaux sociaux sont devenus de plus en plus populaires, des sites et applications comme *Facebook* ou *Twitter* fournissent aux utilisateurs un espace communautaire pour interagir avec les autres, partager des intérêts communs ou échanger des informations.

L'utilisation intensive des réseaux sociaux à fait que d'énormes masses données portants sur les intérêts ou le profile des utilisateurs sont générés quotidiennement, ces données constituent une source d'information considérable aux entreprises qui veulent offrir des services en accord avec les intérêts des utilisateurs, des recommandations pertinentes ou bien mener des campagnes de marketing ciblées. De ce fait un besoin grandissant à vu le jour, portant sur l'analyse des données générées par les réseaux sociaux. Vu que les données sont générées à un très gros volume et à une grande vélocité, l'analyse des réseaux sociaux à recours aux solutions offertes par les technologies Big Data, ces technologies fournissent la possibilité d'analyser des grandes masses de données crées à une très grande vélocité et dont le format est souvent varié, ce qui les rend bien adaptées à l'analyse des réseaux sociaux.

Dans le cadre de notre projet de fin études nous proposons d'utiliser les technologies Big Data pour analyser les données portant sur un sujet d'intérêt commun publié sur le réseau social *Twitter*.

Pour décrire ce présent travail, nous avons organisé notre mémoire en quatre chapitres :

I. Introduction au big data, ce premier chapitre constitue une description des principaux concepts et technologies du domaine des Big Data.

II. **Analyse des réseaux sociaux avec les outils Big Data**, ce chapitre porte sur des définitions des réseaux sociaux et une description des données générées dans ce contexte, il illustre aussi un panorama des solutions offertes pour l'analyse des réseaux sociaux avec les big data.

III. Cas d'étude, ce chapitre est consacré à l'approche proposée dans le cadre de notre travail, une déscription détaillée de notre contribution est fournie.

IV. impléméntation, dans ce dernier chapitre, nous décrivons l'environnement de développement ainsi que l'outil réalisé à travers une série de captures d'écran.

Chapitre I : Introduction au Big Data.

I. Introduction

La notion de big data est un concept s'étant popularisé en 2012 pour traduire le fait que les entreprises sont confrontées à des volumes de données (data) à traiter de plus en plus considérables et présentant un fort enjeu commercial et marketing.

Le concept de big data n'est pas propre au commerce ou au marketing, mais le développement du commerce électronique et du marketing digital a joué un rôle important dans la mise en évidence de la problématique du big data. Ce sont en effet des secteurs qui par nature génèrent d'énormes volumes de données à traiter. [1]

Dans ce chapitre on va définir le Big Data et déterminer les différents concepts liés à ce grand sujet.

II. Concepts et technologies Big data

75% des dirigeants et managers d'entreprises interrogés par Opinion way se sont en tous cas déclarés incapables de donner une définition précise du big data. Pire : 86% avouent que la notion leur paraît « floue ». [2]

Pour enlever l'ambigüité du big data nous avons portée deux définitions intéressantes, mais avant de définir le Big Data, il est essentiel de dire ce qu'il n'est pas. « Big Data is NOT a bigger data warehouse » (Paul Doscher, LucidWorks). Autrement dit, exploiter le big data ne consiste pas à construire des data centres toujours plus gros pour stocker toujours plus de données. [2]

L'autre définition est que : Le Big data c'est la capacité de stocker et de traiter de très grandes quantités de données, de l'ordre du petaoctets. Facebook par exemple, gère une base de données de l'ordre de 100 petaoctets soit : 100 millions de milliards de données.[3]

Il ne faut pas considérer ces chiffres impressionnants uniquement en termes de capacité de stockage (Stock). La capacité de réactualisation de très grands volumes (flux) est tout aussi remarquable. Wal-Mart gère ainsi 1 million de transactions commerciales chaque heure. [3]

Enfin le Big Data c'est aussi la capacité de gérer de multiple formats : numériques, textes, images.

II.1. L'univers multidimensionnel du Big Data

En 2001, dans un rapport du Groupe META (futur Gartner Group), Doug Laney décrivit les nouveaux enjeux liés à la croissance de ce qu'on appellera par la suite le Big Data dans un modèle articulé autour de 3 attributs représentés par les **"3V" : Volume, Vitesse et Variété**.

Ce modèle toujours d'actualité fut complété par IBM qui ajouta un **4emeV**, la **Véracité.**

Plus récemment, quatre nouvelles dimensions pertinentes ont été ajoutées: la **Valeur**, la **Visibilité**, la **Data-visualisation** et les **Opportunités**.

Le modèle Big Data est aujourd'hui défini par 8 attributs :

II.1.1. Volume

Des Volumes importants de données stockées ou présentes à un instant donné et qu'il faut capturer (en terme de flux).Concrètement, pour une boutique en ligne, une entreprise de marketing ou un laboratoire de recherche, il est souvent nécessaire de disposer d'un grand nombre de données pour disposer d'un échantillon représentatif de la réalité et mener à bien les études ou les recherches.

Un commercial sera très heureux de pouvoir disposer de 10 ou 100 millions d'enregistrements afin d'identifier les clients potentiels. Un département se consacrant à l'étude des sciences de la terre ou du ciel exploite plusieurs centaines de millions voire des milliards d'enregistrements. [4]

II.1.2. Vélocité

La vitesse d'exécution d'enregistrement, d'analyse et de prise de décision. Le temps de capture des données, d'exécution des requêtes et des traitements doive correspondre au cycle de vie des données et aux besoins du métier, notamment aux attentes des clients qui souhaitent généralement une réponse ou une information en temps réel ou légèrement différée et même parfois proactive.

Ainsi, une agence de renseignements ou un bureau d'enquête analysera volontiers un fichier de 500 millions de transactions quotidiennement pour y déceler les fraudes éventuelles. Plusieurs fois par jour, une banque d'affaire doit pouvoir envoyer les centaines de milliers de rapports financiers à ses clients en moins d'une heure. En revanche, si un phénomène évolue rapidement dans le temps, un laboratoire de génie génétique ou de microbiologie doit disposer des résultats d'une analyse de millions d'échantillons en moins de 2 minutes ou les suivre en temps réel. [4]

II.1.3. Variété

La Variété ou l'hétérogénéité des données, des supports et des formats. Les données proviennent de sources internes ou externes dont les supports et les formats ne sont pas toujours contrôlés par l'entreprise. Le format des données stockées dans une base de données est différent de celui d'une feuille Excel, des pages html, des e-mails, des petites annonces, des SMS ou des vidéos. Les capteurs, les émissions radios, les photographies, les dessins et les films utilisent leurs propres structures de données. Ces données structurées, semi-structurées ou non structurées doivent être exploitées dans leur format natif.

La disponibilité des données étant un facteur essentiel pour les besoins du métier et donc la pérennité de l'entreprise, nous y reviendrons dans un instant. [4]

II.1.4. Véracité

La Véracité des données, c'est-à-dire leur fiabilité et leur qualité Proposée par IBM, cette dimension est essentielle car de plus en plus de données proviennent de sources extérieures, hors du périmètre de contrôle de l'entreprise. Il est stratégique pour sa pérennité et sa réputation que les données proviennent de sources fiables. Comme la rumeur propage de fausses informations, de fausses données génèrent de faux résultats. [4]

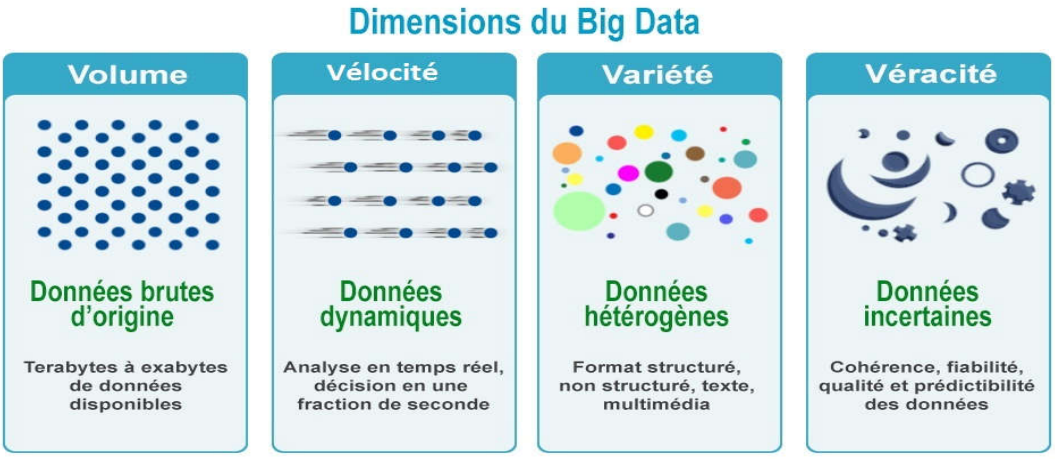


Figure 1 : Les 04 V proposées par IBM.

II.1.5. Valeur

Les Valeurs que représentent ces données apportent de nouvelles connaissances. L'analyse des données donne du sens aux informations collectées, à la fois individuellement en isolant l'épiphénomène et globalement.

Ainsi, la surveillance d'un réseau de milliers de caméras de surveillance permet de contrôler les zones suspectes qui, allié à un système de reconnaissance de forme permet d'identifier les suspects. Une gestion globale des tremblements de terre ou de l'activité volcanique à travers le monde grâce aux satellites radars et balises GPS permet d'évaluer avec précision l'évolution d'évènement imperceptibles depuis le sol ou étudiés manuellement et permet d'anticiper des cataclysmes et de prévenir les populations concernées. [4]

II.1.6. Visibilité

La Visibilité des données à travers des tableaux récapitulatifs ou Dashboard. Disposer de données mais ne pas pouvoir s'en servir ou les visualiser est une perte de temps et d'argent. Il est nécessaire que les informations soient factuelles, disponibles et visuellement présentables, c'est-à-dire accessibles rapidement afin de les surveiller, de les comprendre quel que soit le support utilisé. Les Dashboard sont généralement utilisés par les managers, chefs de projets et autres dirigeants des entreprises ainsi que par les équipes marketing qui ont peu de temps pour prendre une décision et de ce fait ont besoin d'une interface optimisée et intuitive (par ex. CaptainDash). [4]

II.1.7. Data-visualisation

La Data-visualisation concerne la représentation des données sous formes intelligentes, pratiques et interactives. Il n'est rien de plus pénible que de devoir lire un texte ou un tableau de nombres si on peut le résumer par un graphique. Le Big Data s'adapte parfaitement à l'expression "une image vaut mille mots". Si l'exploitation des données permet de leur donner du sens, ce sont les images, les diagrammes, les graphiques, les cartes et les infographies qui leur donnent un sens global, révèle les points forts et les détails. En éliminant les requêtes, en simplifiant les analyses, grâce à une interface interactive (par ex. Webmasters), les visuels apportent une réponse rapide, améliorent la qualité de l'aide à décision, accélèrent la prise de décision et offrent plus de liberté aux utilisateurs pour développer leur créativité. [4]

II.1.8. Opportunité

Les Opportunités offertes par la gestion de ces données en termes d'économie d'échelle et de métier. Tirer avantage des opportunités peut faire évoluer la stratégie de l'entreprise, augmenter sa compétitivité et améliorer sa réputation sur les plans technique et financier, les coûts d'installation et d'exploitation d'un réseau d'ordinateurs sont inférieurs à ceux d'une seule infrastructure équivalente, les traitements étant également répartis sur plusieurs "petites" machines. Point de vue métier, le client possède des données que les autres ne possèdent pas, lui offrant un avantage sur ses concurrents. Il peut également tirer parti où monnayer ses connaissances et ses services, les distribuer et améliorer la satisfaction de ses clients. [4]

II.2. Les enjeux du Big Data en entreprise

Pour les entreprises, le big data constitue une aide à la prise de décision au même titre que le business intelligence puisqu'il permet de mieux cerner les besoins des clients ou même d'anticiper leurs futures consommations**.** Les possibilités offertes sont vertigineuses et la plupart demeurent encore insoupçonnées pour l'heure. [5]

Toutefois, il existe de nombreuses applications concrètes, déjà accessibles ou qui devraient être proposées par les acteurs du marché à brève échéance :

- **Marketing** → La connaissance du client permise par le big data est tellement affinée qu'il devient envisageable pour l'entreprise de personnaliser la communication et de segmenter les offres promotionnelles avec une précision inégalée jusqu'ici. [5]

- **Production** → L'analyse des données peut être utilisée pour améliorer les processus, économiser les ressources énergétiques et naturelles, mieux gérer les stocks ou encore prévoir la maintenance des machines-outils. [5]

- **Commerce** → L'étude comportementale des clients basée sur le détail des commandes et la fréquence des achats peut servir à optimiser l'agencement des produits dans les rayons ou l'emplacement des points de vente et à fixer des prix de vente plus incitatifs. [5]

- **Recherche** → Le big data permet de traiter rapidement les masses d'informations issues des expérimentations, dans le but d'accélérer le prototypage et la mise sur le marché. [5]

- **Finance** → Les banques et les assureurs ont recours au big data pour déterminer le niveau de risque d'un client contractant un prêt ou une garantie et pour mettre en évidence un comportement suspect ou révéler des fraudes potentielles. [5]

- **Santé** → Le big data rend théoriquement possible un médecin préventive et personnalisée en lien avec des appareils connectés mesurant les données biométriques des patients et visant à leur proposer des conseils ou des traitements appropriés. [5]

- **Transport** → Le big data permet de modéliser les déplacements des usagers ou des salariés de l'entreprise afin d'optimiser les trajets et la fréquence des véhicules et donc d'accroître le taux de remplissage tout en faisant des économies de carburant. [5]

II.3. L'influence du Big data sur le processus décisionnel

La question qui se pose est Est-ce que les technologies Big Data transformer le processus décisionnel classique, et Comment ?

Réponse : Le décisionnel doit couvre deux aspects complémentaires mais distincts dans leur approche. On les développe par la suite :

a) L'aide à la décision pour le pilotage d'activité

Destinée aux managers-décideurs en charge d'une unité le Business Intelligence doit leurs apporter toute l'aide nécessaire pour conduire les activités dont ils ont la charge dans la bonne direction et selon les critères de performance attendus. Ils sont demandeurs d'instruments précis, délivrant une information rapide et bien ciblé pour réduire le risque inhérent à toutes prises de décision. [6]

b) L'analytique

Ce domaine est un peu en marge des processus de l'entreprise. C'était le domaine privilégié des statisticiens et des spécialistes du data mining. Ils sont aujourd'hui bousculés par les data scientistes ou scientifiques des données, les vrais spécialistes du « big data ». [6]

C'est un rôle bien plus complexe qui exige une maîtrise non seulement des techniques d'analyse de données, mais aussi de la technologie informatique et des métiers de l'entreprise.

Ils travaillent en effet avec les managers-décideurs, ils étudient des hypothèses de réflexion et bâtissent des modèles pour pousser plus avant la connaissance : clients, produits, processus…etc.

Le Big Data permet de bâtir des modèles bien plus complets qu'auparavant, Il améliore sensiblement la connaissance des thèmes habituellement prospectés et ouvre de nouveaux champs d'étude. Bien utilisé, il peut améliorer la connaissance des décideurs. [6]

II.4. Les sources des données

Les données représentent la matière première de l'entreprise mais avant de les exploiter il faut savoir où les trouver en fonction des besoins du métier.

On peut regrouper les sources de données en 6 grandes catégories :

II.4.1. Les outils professionnels

Les outils professionnels comprenant les logiciels et services des entreprises. Il s'agit des logiciels de gestion (Customer Relationship Management, Enterprise Resource Planning, Supply Chain Management, etc.), les outils de production de contenu et les suites bureautiques telles que MS-Office et autre suites Adobe.

Selon Microsoft, la moitié des données produites par la suite Office sont hors contrôle et ne sont donc pas valorisées. Le meilleur exemple est le courrier électronique avec ses plus de 100 millions d'e-mails envoyés chaque minute. [4]

II.4.2. Internet

Internet et ses volumes gigantesques de données stockées dans les sites d'actualités, scientifiques, gouvernementaux, de commerce en ligne, des entreprises, des organisations et des amateurs. Ces ensembles hétéroclites génèrent des interactions de plus en plus nombreuses, rendant les annuaires et moteurs de recherche indispensables, eux-mêmes étant une source de données grâce aux requêtes des internautes. [4]

II.4.3. Les réseaux sociaux

Bien qu'accessibles via Internet, ils constituent une source distincte de données car ils proposent aux internautes de nouveaux outils d'expression. Le Web 2.0 en particulier permet à chacun d'interagir et de collaborer sur le web dans le but de produire du contenu pour une communauté virtuelle. C'est valable pour les grands réseaux sociaux tels que Facebook ou YouTube mais également pour les sites de partage tels Flickr ou Instagram, les blogs, les flux RSS, les réseaux professionnels tels LinkedIn, Yammer, etc. Chaque seconde se sont des mégaoctets de données qui sont publiés sur ces sites depuis pratiquement tous les points du globe. [4]

II.4.4. Les systèmes connectés

Pour une entreprise analysant le Big Data, un Smartphone ou un ordinateur disposant d'une connexion Internet n'est pas un terminal mais une source de données, et la NSA le sait mieux que quiconque. En moyenne, un internaute établit 150 connexions par jour via son Smartphone pour consulter ses messages, se connecter aux réseaux sociaux et autres services en ligne. A l'avenir l'Internet des Objets permettra aux entreprises de recueillir et d'analyser directement des données opérationnelles pour améliorer les services ou en développer de nouveaux. [4]

II.4.5. Les données structurées

Il s'agit des données stockées par rangées et colonnes dans des bases de données et généralement exploitées par les entreprises (comptabilité, finance, gestion de stock, ressources humaines, recherche scientifiques, bases documentaire, etc.)[4]

II.4.6. Les données non structurées

Il s'agit des données produites par les outils les plus divers allant des logiciels bureautiques aux articles publiés sur les blogs ou les messages diffusés sur les forums. Ils ne sont pas organisés, les formats variés et leur nombre augmente de manière exponentielle et sans contrôle, principalement via les réseaux sociaux et autres communautés virtuelles. C'est leur développement qui est à l'origine du Big Data. [4]

Il faut noter toutefois que l'usage du big data en entreprise est encadré par la réglementation française en matière de protection des données personnelles**.** Dès lors, toute initiative doit être compatible avec les exigences de la CNIL : les personnes doivent être tenues informées de la collecte et de l'analyse de leurs données et ont la possibilité de s'y opposer à tout instant. Par ailleurs, seules les données nécessaires à un usage défini à l'avance peuvent être collectées. [5]

III. Volumétrie des données -Les Big Data en chiffres

Dans une étude publiée par la Commission Européenne en 2015 et reliée par l'IUCN, on apprend que dans le monde on a créé plus de données en 10 minutes qu'au cours de toute l'histoire de l'humanité jusqu'en 2003 ! Et pourtant les grandes bibliothèques comme les médiathèques numériques sont vastes !

Dans les années 1990, quand nous effectuions une recherche sur Internet, nous étions ravis quand nous trouvions une bribe d'information. Aujourd'hui nous sommes étonnés quand nous ne trouvons pas au moins une page d'hyperliens et déçus quand il n'y a pas la moindre image !

Dans un article publié en 2009 dans le Times et repris par Google, on apprenait que les internautes avaient effectué 100 milliards de recherches sur Internet, (ce qui engendrait par ailleurs 8400 tonnes d'émission de gaz à effet de serre chaque une année !). [4]

Début 2013, on a dépassé le nombre de 200 milliards de requêtes par mois, ce qui représente plus de 77000 requêtes par seconde !

En 2014, Google avait indexé plus de 30000 milliards de pages, ce qui représente 100 pétabytes de données selon Statistic Brain contre 462 milliards de pages pour Archive.

Mais toutes ces données ne représentent qu'une goutte d'eau quand on sait qu'en 2014 le web contenait 200 millions de fois plus d'information ! [4]

IV. Technologies du big data

Les traitements massivement parallèles, la gestion en temps réel des pannes systèmes ou la redondance systématique des données, c'est un peu tout cela le Big Data.

Bon pas seulement, après il y a les utilisations et là c'est une autre paire de manches. Mais intéressons déjà aux technos proprement dites.

IV.1. Map Reduce

Les premiers à être confronté à cette problématique de données numérique massive et à avoir trouvé une solution fut Google. En effet devant l'échec des approches classiques et des SGBD, le géant Américain Google à décider de développer son propre système distribué.

Un système distribuer prend en compte 2 critères majeurs qui sont :

- La distribution des données au sein du cluster
- La distribution des charges de traitements au sein du cluster

Google a donc mis en place son système de fichier distribuer appeler GFS (Google File System), dont le but principal est de répartir les données volumineuses au sein de plusieurs machines sur un cluster, tout en s'adaptant à l'évolution matériel et respectant les contraintes d'un système distribuer sur les problématiques de HA et de synchronisation des données, entre les machines au sein d'un cluster. [7]

Cependant une autre technologie fut mise en place pour le traitement sur ces fichiers distribuer au sein du cluster qui porte le nom de MapReduce.

MapReduce permet de lire les fichiers et d'effectuer 2 types d'actions bien distincts pour traiter les données qui sont :

- La phase Map qui lit les données concernées sur chaque machine du cluster
- La phase Reduce qui agrège les résultats pour produire une sortie.

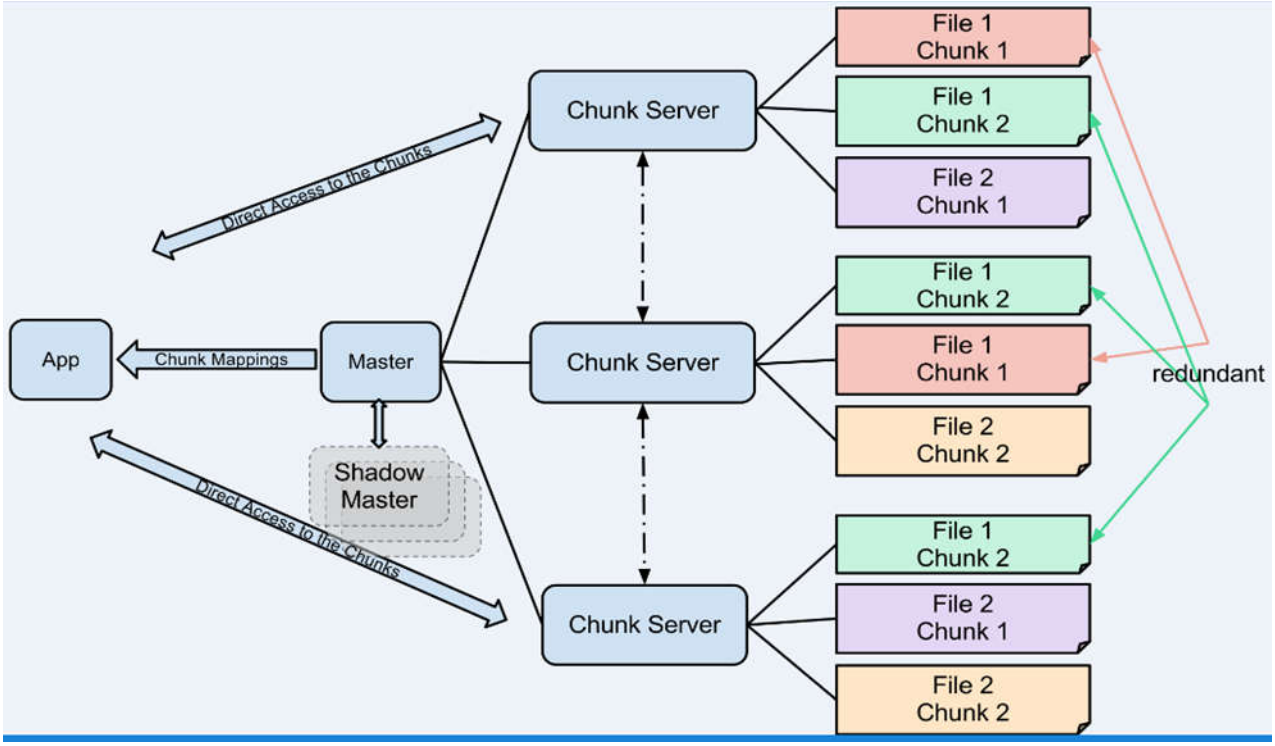


Figure 2 : Architecture de GFS (Google File System).

A ce jour encore GFS est toujours utilisé par Google et reste propriétaire pour l'entreprise.

IV.2. Hadoop

Devant la solution de Google qui reste propriétaire mais satisfaisante à cette problématique du BIG Data, un ingénieur du nom de Doug Cutting, actuellement Chief Architect de Cloudera, à l'origine du projet Apache Hadoop, s'est basé du manifeste de Google, comme cahier de charge pour réaliser son système distribuer.

Il a donc créé HDFS (Hadoop File System) qui devient le premier système de fichier distribué Open source au monde, avec une phase de traitement réalisé par la logique de développement MapReduce.

Suivant donc la définition du BIG DATA dite des **3V**, Hadoop apporta des solutions aux traitements des données de différents formats (Vidéo, Audio, XML, Json, Csv, Texte, Binaire…) avec des performances très apprécié par la communauté. [8]

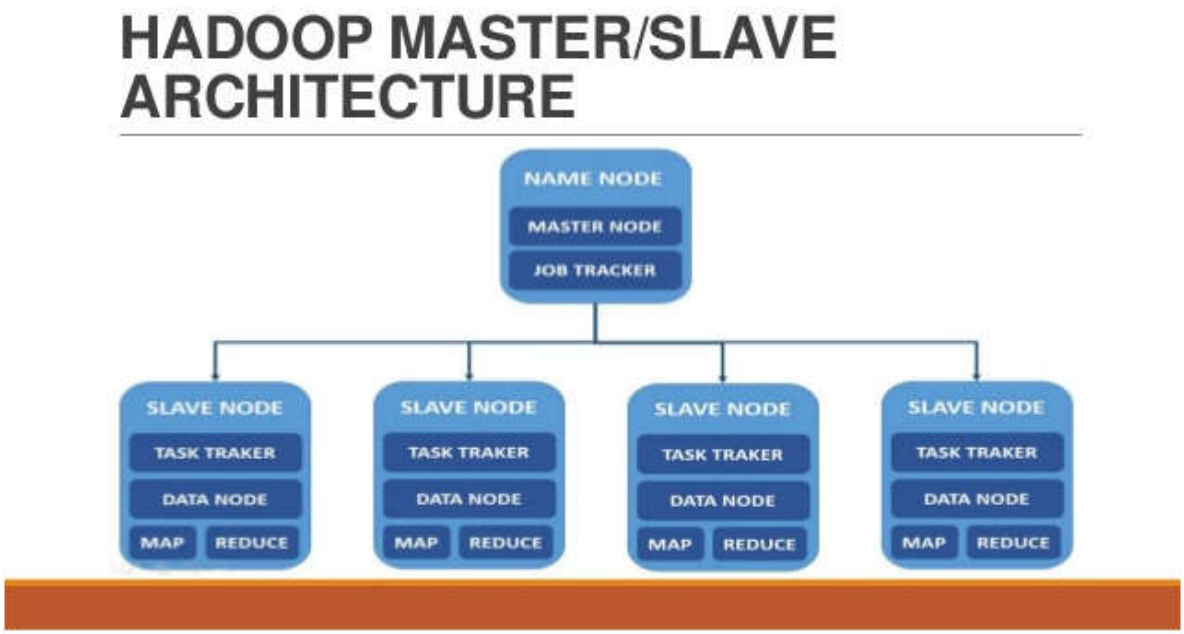


Figure 3 : Architecture HDFS (Hadoop Distributed File System).

Hadoop a été écrit en Java (natif), il est possible cependant, d'écrire des tâches de traitement MapReduce Hadoop en Node.js, Python, …

IV.3. Bases No SQL

Les bases de données relationnelles ont une philosophie d'organisation des données bien spécifiques, avec notamment le langage d'interrogation SQL, le principe d'intégrité des transactions (ACID), et les lois de normalisation. Bien utiles pour gérer les données qualifiées de l'entreprise, elles ne sont pas du tout adaptées au stockage de très grandes dimensions et au traitement ultra rapide. Les bases NoSQL autorisent la redondance pour mieux servir les besoins en matière de flexibilité, de tolérance aux pannes et d'évolutivité. [8]

IV.4. Stockage "In-Memory"

Pour des analyses encore plus rapides, les traitements directement en mémoire sont une solution. Une technologie bien qu'encore trop coûteuse il est vrai pour être généralisée. Le service est-il à la hauteur de l'investissement ? [8]

IV.5. Cloud Computing

Le Big Data exige une capacité matérielle hors du commun, que ce soit pour le stockage comme pour les ressources processeurs nécessaires au traitement. Nul besoin de s'équiper outre mesure, le "Cloud" est là pour cela. Encore faut-il avoir bien compris le concept pour différencier, le Cloud privé du Cloud public, l'interne de l'externe et les hybrides combinant plusieurs types de solutions. Ensuite il est aussi prudent de différencier les niveaux de services de chacune des solutions : IAAS, PAAS, SAAS... [8]

V. Applications du big data

Les usages du Big Data sont infinis, mais quelques domaines majeurs émergent.

V.1. Comprendre le client et personnaliser les services

C'est l'une des applications évidentes du Big Data. En captant et analysant un maximum de flux de données sur ses clients, l'entreprise peut non seulement dégager des profils génériques et concevoir des services spécifiques, mais aussi personnaliser ces services et les actions marketing qui y seront associées. Ces flux intègrent les données « classiques » déjà organisées via des systèmes de CRM, mais également les données non structurées issues des médias sociaux ou de capteurs intelligents capables d'analyser le comportement des clients sur le lieu d'achat.

L'objectif est de dégager des modèles susceptibles de prévoir les besoins des clients afin de leur fournir des services personnalisés en temps réel. On parle dès lors de segmentation attitudinale. Avec la quantité infinie de données qu'il collecte à notre sujet, Google est évidemment un acteur incontournable en la matière.

Ces modèles seront utilisés dans tous les secteurs d'activités, depuis les grandes enseignes commerciales pour améliorer et personnaliser les offres, notamment dans l'e-commerce, en passant par les assurances qui seront adaptées à chaque cas particulier ou encore au monde politique pour lequel la capacité à « interpréter » les souhaits des électeurs est depuis toujours un. [9]

V.2. Optimiser les processus business

Le Big Data va également impacter fortement les processus business. Des processus complexes tels que la Supply Chain Management (SCM) seront optimisés en temps réel en fonction de prévisions issues de l'analyse des données des médias sociaux, des tendances d'achats, de la circulation routière ou des stations météorologiques.

Un autre exemple concerne la gestion des ressources humaines, depuis le recrutement jusqu'à l'évaluation de la culture d'entreprise ou la mesure de l'engagement et des besoins du personnel. [9]

V.3. Améliorer la santé et optimiser les performances

Le Big Data va considérablement affecter les individus. Cela passe tout d'abord par le phénomène du « Quantified Self », c'est-à-dire la capture et l'analyse des données relatives à notre corps, notre santé ou nos activités, via le mobile, les « wear ables » (montres, bracelet, vêtements, lunettes, …) et plus généralement l'Internet des Objets. L'évolution des sites de rencontre passera également par l'utilisation d'algorithmes sophistiqués basés sur l'analyse de profils sociaux beaucoup plus riches et complexes.

Le Big Data va permettre des avancées considérables dans des domaines tels que le décodage de l'ADN ou la prédiction des épidémies ou la lutte contre des maladies encore incurables

comme le Sida. Avec les modélisations basées sur des quantités de données infinies, les essais cliniques ne seront plus limités par la taille des échantillons. Dans le domaine du sport, on peut citer l'exemple d'IBM a qui a développé SlamTracker pour le tennis. Grâce aux captures vidéo et à l'analyse des données liées, il est possible d'améliorer la préparation d'un match en analysant le jeu d'un adversaire sur base de paramètres inédits. Autre exemple remarquable, la victoire de l'Oracle Team USA lors de la fameuse compétition de l'America's Cup, ou comment 300 senseurs et 3000 variables ont permis un incroyable retournement de situation. [9]

V.4. Rendre les machines intelligentes

Le Big Data va rendre les machines et terminaux les plus divers plus intelligents et plus autonomes. Elles sont indispensables au développement de l'industrie 4.0. Avec la multiplication à l'infini des capteurs sur les équipements domestiques, professionnels et industriels, le Big Data appliqué au M2M (MachineTo Machine) va offrir de multiples opportunités pour les entreprises qui investiront ce marché. Les voitures intelligentes illustrent ce phénomène. Elles génèrent déjà d'énormes quantités de données qui peuvent être exploitées pour optimiser l'expérience de conduite ou les modèles de taxation. Les voitures intelligentes seront en mesure d'échanger entre elles des informations en temps réel et d'optimiser leur utilisation en fonction d'algorithmes spécifiques. Grâce aux capteurs équipant son matériel agricole, John Deere permet aux entreprises agricoles d'améliorer la gestion de leur flotte, de réduire les temps d'arrêt et d'économiser le carburant. Le système est basé sur le croisement des données en temps réel et historiques relatives à la météo, les conditions du sol, les caractéristiques des cultures, etc.

De même, les maisons intelligentes seront des contributeurs majeurs pour la croissance des données M2M. Les compteurs intelligents surveilleront les consommations énergétiques, mais seront surtout capables de proposer des comportements optimisés sur bases de modèles issus des analytiques.

Le Big Data est également indispensable au développement de la robotique. Les robots vont générer et utiliser des volumes considérables de données pour comprendre leur environnement et s'y insérer de manière intelligente. En utilisant des algorithmes d'auto-apprentissage basés sur l'analyse de ces données, les robots pourront améliorer leur comportement et effectuer des tâches toujours plus complexes, comme le pilotage d'un avion par exemple. Aux USA, des robots sont maintenant capables de percevoir les similarités ethniques grâce aux données issues du crowdsourcing. [9]

V.5. Développer les SmartCities

Le Big (Open) Data est indissociable du développement des villes et territoires intelligents. Un exemple classique concerne l'optimisation des flux de trafic sur base d'informations « crowdsourcées » en temps réels à partir des GPS, des capteurs, des mobiles ou des stations météorologiques.

Le Big Data va permettre aux villes, et singulièrement les mégalopoles de relier et faire interagir des secteurs fonctionnant jusque-là en silos : bâtiments privés et professionnels, infrastructures et systèmes de transport, production d'énergie et consommation des ressources, etc. Seules les modélisations issues du Big Data permettent d'intégrer et d'analyser les paramètres innombrables issus de ces différents secteurs d'activité. C'est également l'objectif de l'initiative SmarterCities d'IBM.

Dans le domaine de la sécurité, les autorités pourront utiliser la puissance des Big Data pour améliorer la surveillance et la gestion des événements mettant en péril notre sécurité ou pour prédire d'éventuelles activités criminelles, dans le monde physique (vols, accidents de la route, gestion des catastrophes, …) ou virtuel (transactions financières frauduleuses, espionnage électronique, …). [9]

VI. Les limites du Big Data

Si le terrain de jeu du Big Data est loin d'être restreint, il n'est pas sans limites. Elles tiennent, en premier lieu, à la nature des données et aux traitements envisagés, et lorsqu'il est question des données personnelles, la vigilance est nécessaire. Dans certains pays, le traitement des données à caractère personnel est régi par des dispositions particulières, ce qui n'est pas le cas dans la majorité des pays.

Nous sommes arrivés à un point où la protection des données personnelles, portée à la défense des libertés fondamentales de l'individu, est en train de devenir un argument économique. L'enjeu étant dorénavant, d'élaborer le cadre normatif le plus attractif pour le développement de l'économie numérique et des échanges de données, ceci nécessite d'être vigilent dans un contexte de forte concurrence entre les puissances économiques.

L'autre préoccupation provient de la sécurité : une faille minuscule peut menacer des quantités de données considérables. Si les utilisateurs perdent confiance dans l'utilisation de leurs informations, c'est donc tout l'édifice du big data qui risque de s'écrouler. Pour éviter cela, par exemple en Europe, la commission européenne a présenté, en début 2012, un règlement qui vise à protéger davantage les utilisateurs, il obligera les entreprises à demander le consentement explicite de l'utilisateur avant de collecter ses données. [10]

Ainsi un nouveau règlement européen sur la protection des données personnelles est paru au journal officiel de l'Union européenne le 4 mai 2016 et entrera en application en 2018. L'adoption de ce texte doit permettre à l'Europe de s'adapter aux nouvelles réalités du numérique. [11]

VII. Conclusion

En conclusion puisque le Big Data est un sujet assez vaste, nous avons introduit dans ce chapitres les concepts qui entourent le Big data et qui sont liés à notre sujet..

Ainsi nous avons conclu que la bonne gestion du Big Data peut apporter des solutions et offrir des opportunités aux entreprises pour gagner un avantage conséquent ce monde concurrentiel.

Chapitre II :

Analyse des réseaux sociaux avec les outils Big Data.

I. Introduction

L'Internet est actuellement le plus grand réseau informatique sur notre planète. L'Internet ne se limite plus aux universités, aux industries et aux gouvernements. Aujourd'hui tout le monde l'utilise, car chaque particulier peut maintenant se joindre au réseau. L'Internet permet d'échanger les informations en toute liberté. En même temps, on observe le développement dynamique des réseaux sociaux qui deviennent plus populaires et plus utilisés.

Toutes les entreprises et les sociétés cherchent à extraire une valeur et gagner des profits de l'internet via l'analyse des réseaux sociaux qui est n'est pas possible qu'à l'aide de Big Data et leurs technologies.

Dans ce chapitre nous allons définir et décrire les réseaux sociaux et plus particulièrement la relation entre les réseaux sociaux et les Big Data. Les sections suivantes constituent un état de l'art sur l'analyse des données via les Big Data et les différents outils du Big Data pour analyser les réseaux sociaux.

II. Les réseaux sociaux, c'est quoi ?

En tant que notion, un réseau social représente un groupement qui a un sens : la famille, les collègues, un groupe d'amis, une communauté, etc. Il s'agit d'un agencement de liens entre des individus et/ou des organisations.

Les réseaux sociaux sont des regroupements d'individus ou d'organisations qui discutent, parlent, échangent entre eux. Ils partagent des opinions, des idées ou encore du contenu. Sur le web, les réseaux sociaux sont grandement favorisés par l'avènement des plateformes comme Facebook, Twitter, YouTube ou LinkedIn [12].

Des réseaux sociaux peuvent être créés stratégiquement pour agrandir ou rendre plus efficient son propre réseau social (professionnel, amical).

Il existe des applications Internet aidant à se créer un cercle d'amis, à trouver des partenaires commerciaux, un emploi ou autres. Il s'agit de services de Réseautage social.

Figure 4 : les différents réseaux sociaux

II.1. Les types des réseaux sociaux

Il existe plusieurs « familles » de réseaux sociaux. Voici les principales.

✓ Les réseaux sociaux personnels

Ici, on parle des réseaux comme Facebook (avec plus de 1 milliard d'utilisateurs à ce jour, c'est le plus connu et le plus populaire). Ils permettent de retrouver des connaissances, membres de la famille ou des amis. On peut échanger des photos et des vidéos, discuter avec eux, organiser des évènements, etc.

✓ Les réseaux sociaux de divertissement

Ils permettent de partager et de diffuser de la musique ou des vidéos. Parmi les plus connus, on retrouve entre autres YouTube et aussi MySpace (qui vient tout juste de faire peau neuve). Ce sont des portails qui peuvent entre autres permettre aux artistes de se faire connaitre.

✓ Les réseaux sociaux d'affaires

LinkedIn est un bon exemple. Ce réseau permet de se connecter avec ses collègues de bureaux, ses fournisseurs, ses partenaires d'affaires ou encore des employeurs potentiels. En se créant un profil, on peut y inclure son curriculum vitae et ses réalisations. On peut aussi interagir en discutant de notre secteur d'activités. [12]

Figure 5: les différents types des réseaux sociaux

II.2. Le lien entre Les réseaux sociaux et Big Data

Les données sont la spécialité des réseaux sociaux. Grâce à la tendance des utilisateurs à partager de nombreuses informations, les réseaux sociaux collectent en effet de très nombreuses données. Cette large quantité de données reflète la façon dont les gens interagissent entre eux. Au centre de ces interactions, on retrouve des informations précieuses sur ce que les individus et la société considèrent important. Le volume de données, combiné à la vitesse rapide de flot des données qui caractérisent les réseaux sociaux, représentent l'essence du Big Data.

En utilisant les outils analytiques sur les données des réseaux sociaux, les entreprises peuvent aller au-delà des simples mécaniques d'interactions et permettent de voir la façon dont le contenu de ces interactions peut affecter les performances du business et son image de marque. Les analyses de contenu permettent aux entreprises de dégager des informations actionnables à partir des messagers postés par les utilisateurs. Par exemple, les outils analytiques peuvent être programmés pour pister les sentiments négatifs ou positifs au sujet d'une marque, lesquels peuvent représenter une menace pour la réputation ou le chiffre d'affaires de l'entreprise. [13]

III. L'Analyse des réseaux sociaux

III.1. Définition de l'analyse des réseaux sociaux

L'analyse des réseaux sociaux est avant tout une boîte à outils permettant de visualiser et modéliser les relations sociales comme des nœuds (les individus, les organisations…) et des liens (relations entre ces nœuds). De ce fait, l'analyse des réseaux sociaux repose sur des visualisations graphiques issues d'algorithmes permettant de calculer des degrés de force ou de densité entre les différents acteurs d'un réseau.

Ainsi, l'analyse des réseaux sociaux est fondée sur une approche structurale des relations entre membres d'un milieu social organisé. Elle s'attache à décrire les interdépendances entre acteurs et permet une simplification de leur représentation que Lazega (1998, p. 6) qualifie de « représentation simplifiée d'un système social complexe ». Cette simplification dans l'agencement des interdépendances est volontaire puisque l'analyse des réseaux sociaux se veut être une « technique d'exploration et de représentation » (Lazega, 1998, p.6).

Par ailleurs, l'analyse des réseaux sociaux est intégrée à une réflexion plus vaste autour de la sociologie des groupes. La compréhension de la structure des ensembles sociaux repose sur l'étude des relations entre membres d'un milieu social. Cette analyse dite structurale porte spécifiquement sur la description et l'analyse des différents modes de relation possibles : interdépendance des membres, réciprocité ou non des relations, place centrale de certains, absence de relations créant des « trous » relationnels au sein du réseau, fréquence des relations (liens forts versus liens faibles). La force de l'analyse structurale réside dans sa capacité à représenter de façon simplifiée la complexité et la diversité des relations entre acteurs.

Le système d'interdépendance est modélisé en prenant en compte l'imbrication progressive des acteurs au sein d'une « forme » structurale qui évolue, se contracte ou se dilate en fonction de l'activité de ses membres. La plasticité des réseaux est accentuée par leur absence définie de frontières. [14]

III.2. Les types d'analyse des données

Il existe quatre types d'analyse que les entreprises peuvent utiliser pour apprendre et mieux s'engager avec leurs clients. Et c'est à partir des deux autres types que l'on peut vraiment obtenir les connaissances dont vous avez besoin pour faire avancer votre entreprise.

III.2.1. Analyse descriptive

C'est ce que vous obtenez à partir de votre serveur Web via des outils comme Google analytiques, Omniture ou similaires. Vous pouvez rapidement comprendre ce qui s'est passé au cours d'une période donnée dans le passé et vérifier si une campagne a réussi ou si elle n'est pas basée sur des paramètres simples comme les pages vues. Environ 35% des entreprises interrogées disent qu'elles le font de façon cohérente.

III.2.2. Analyses diagnostiques

Si vous souhaitez approfondir les données que vous avez collectées auprès des utilisateurs afin de comprendre «Pourquoi certaines choses se sont produites», vous pouvez utiliser les outils de business intelligence pour obtenir des informations. Cependant, c'est un travail très laborieux qui a une capacité limitée à vous donner des idées utiles. Il s'agit essentiellement d'une très bonne compréhension d'une partie limitée du problème que vous voulez résoudre. Habituellement, moins de 10% des entreprises interrogées le font à l'occasion et moins de 5% le font de façon uniforme.

III.2.3. Analyse prédictive

Si vous pouvez collecter des données contextuelles et les corréler avec d'autres ensembles de données de comportement d'utilisateur, ainsi que d'élargir les données utilisateur au-delà de ce que vous pouvez obtenir à partir de vos serveurs Web, vous entrez dans un domaine entièrement nouveau où vous pouvez obtenir des informations réelles. Essentiellement, vous pouvez prédire ce qui se passera si vous gardez les choses telles qu'elles sont. Cependant, moins de 1% des entreprises interrogées ont déjà tenté de le faire. Ceux qui ont, ont trouvé des résultats incroyables qui ont déjà fait une grande différence dans leur entreprise.

III.2.4. Analyses prescriptives

Une fois que vous arrivez au point où vous 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. C'est à la pointe de l'innovation aujourd'hui, mais c'est réalisable !

Afin d'être en mesure de mettre en œuvre l'analyse prédictive et prescriptive, vous devez ajouter « cognition» à votre analyse à travers des algorithmes d'apprentissage machine. [15]

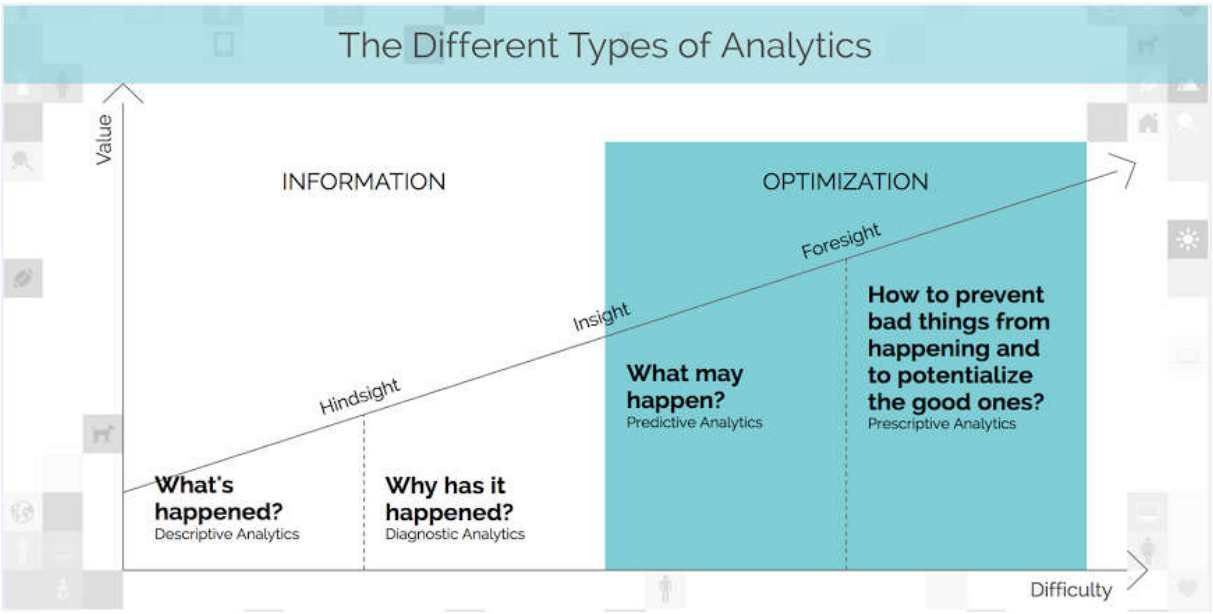


Figure 6 : les différents types d'analyse des données

III.3. L'Evolution de l'analyse de données

L'utilisation de données pour prendre des décisions n'est certainement pas une nouvelle, mais le domaine de l'analyse d'affaires est né au milieu des années 1950, avec l'avènement de la technologie qui pourrait générer et capter des informations et de détecter les modèles à partir de lui plus rapidement qu'un humain pourrait le faire manuellement sans l'aide de toute technologie (Davenport, 2013).

III.3.1. L'ère Analytiques 1.0

La première ère est également connue comme l'ère de la « Business Intelligence ». Analytique 1.0 fut un temps de réel progrès dans l'obtention d'une objective, compréhension approfondie des phénomènes commerciaux et donner aux gestionnaires la compréhension fondée sur les faits afin d'aller au-delà de l'intuition lors de la prise de décisions.
Pour la première fois, des données sur les processus de production, les ventes, les interactions avec les clients et plus ont été enregistrées, agrégés et analysés.
Les ensembles de données étaient assez petits en volume et assez statiques en vélocité pour être séparés dans des entrepôts pour les analyser. Toutefois, la préparation d'un ensemble de données dans un entrepôt était difficile. Les analystes ont passé une grande partie de leur temps à préparer des données pour l'analyse et relativement un peu de temps sur l'analyse elle-même - l'analyse était laborieuse et lente, prenant souvent des semaines ou des mois à exécuter (Davenport, 2013).

III.3.2. L'ère Analytiques 2.0

Aussi connu comme l'ère de «Big Data». L'ère d'analyse 1.0 a duré jusqu'au milieu des années 2000 et analytiques est entré dans la phase 2.0, la nécessité des nouveaux outils puissants et la possibilité de profiter de leur fournissant rapidement est apparu. Les entreprises se sont précipitées pour construire des nouvelles capabilités et acquise des nouveaux clients. La reconnaissance générale de l'avantage qu'un premier déménageur pourrait gagner a conduit à un battage, mais a poussée une accélération des nouvelles offres.

Exemple : LinkedIn, a créé des nombreux produits des données, y compris les gens que vous pouvez savoir, les emplois que vous peut être intéressé, les groupes que vous pouvez aimer, les entreprises que vous souhaitez mettre à suivre, les mises à jour des réseaux et des compétences et de l'expertise et pour ce faire, il a construit une solide infrastructure et a embauché des scientifiques intelligentes.
Des technologies novatrices de toutes sortes devaient être créées, acquises et maîtrisées à cette époque. Les grosses données ne pouvaient pas s'insérer ou être analysées assez rapidement sur un seul serveur, donc il a été traité avec Hadoop, Un Framework de logiciels open source pour le traitement rapide des données par lot sur des serveurs parallèles.
Pour négocier Avec des données relativement non structurées, les entreprises se tournent vers une nouvelle classe de bases de données appelée NoSQL.
Des nombreuses informations ont été stockées et analysées dans des environnements de Cloud Computing publics ou privés.
Parmi les autres technologies introduites durant cette période figurent les analyses «en mémoire» et «en base de données» pour le nombre croissant rapide. Méthodes Machin e-learning (développement de modèles semi-automatisés et testés) ont été utilisés pour générer rapidement des modèles à partir des données. Rapports en noir et blanc a donné lieu à des visuels complexes et colorés.
Les compétences / compétences requises pour l'analytiques 2.0 étaient très différentes de celles nécessaires pour 1,0. Les analystes quantitatifs de la prochaine génération ont été appelés scientifiques des données, et ils possédaient à la fois en termes de calcul et d'analyse (Davenport, 2013).

III.3.3. L'ère Analytiques 3.0

Comme les deux premières époques d'analytiques, celle-ci apporte de nouveaux défis et opportunités, pour les entreprises qui veulent rivaliser sur l'analyse et pour les fournisseurs qui fournissent les données et les outils avec qui les faire (Davenport, 2013).

III.3.3.1. Qu'est-ce que l'ère analytiques 3.0?

Analytiques 3.0 marque le stade de la maturité où les principales organisations réalisent des activités mesurables impact de la combinaison d'analyses traditionnelles et des données importantes. Des entreprises performantes Intégrer les analyses directement dans les processus décisionnels et opérationnels, et tirer profit Machin e-Learning et d'autres technologies pour générer des aperçus en millions par seconde plutôt qu'Un « aperçu d'une semaine ou d'un mois».
Les architectures de données (c.-à-d. Hadoop) approchent éliminant les barrières d'échelle. Analytiques devient véritablement le facteur de différenciation pour les entreprises qui capitalisent sur les possibilités de cette nouvelle ère (Institut international d'analytique, 2015). [16]

III.3.3.2. Scénario analytiques actuel et projections futures

Actuellement, 89% des chefs d'entreprise pensent que Big Data va révolutionner la façon dont les businesses sont fait de même façon que l'Internet, 83% d'entre eux ont pour suivies des projets de Big Data afin de gagner un avantage concurrentiel. Wiki bon - une communauté de praticiens et consultants sur la technologie des systèmes des entreprises.[16]

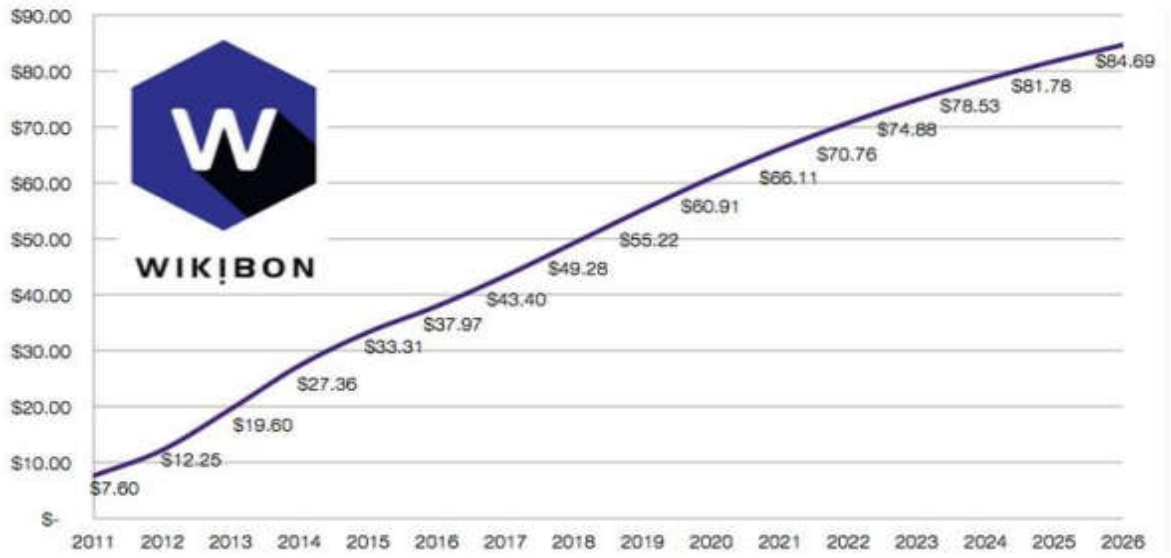


Figure 7 : la prévision du marché du Big Data , 2011- 2026 ($ US B) (Columbus, 2015)

III.3.3.3. Les leaders actuels du marché Big Data analytiques 3.0

IBM et SAS sont les leaders du marché de l'Analyse Prédictive du Big Data selon le dernier rapport Forrester Wave (Forrester est l'une des sociétés de recherche et de conseil influent dans le monde). Le dernier Forrester Wave est basé sur une analyse de 13 grandes fournisseurs d'analyse des données prédictives, y compris **Alpine Data Labs**, **Alteryx**, **AngossSoftware**, **Dell**, **FICO**, **IBM**, **KNIME.com**, **Microsoft**, **Oracle**, **Predixion Software**, **RapidMiner**, **SAP** et **SAS**.

Forrester spécifiquement appelé sur Microsoft Azure Learning est un impressionnant nouvel entrant qui montre le potentiel pour Microsoft d'être un acteur significatif sur ce marché (Columbus, 2015). [16]

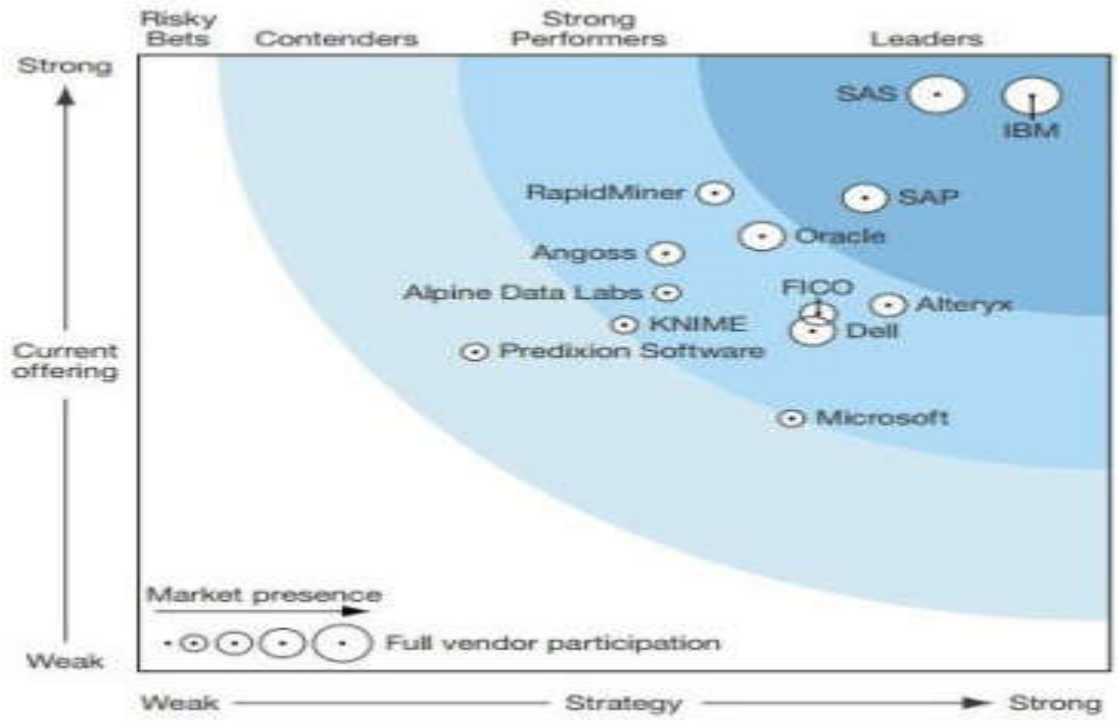


Figure 8 : Forrester Wave: Les dirigeants du marché de l'analyse prédictive du Big Data, Q2 2015 (Columbus 2015).

III.3.3.4. Tendances poussant les frontières de Big Data Analytiques 3.0

Les progrès actuels en matière de technologie ouvrent la voie à l'avenir de l'analyse.

1. Les clients sont à la recherche de matériel intégré et des logiciels pour analyser des charges de travail,
2. R- open source programmation Langages- pour les statistiques de calcul et de visualisation devient omniprésente,
3. Interfaces visuels font avancée Analytiques plus accessible aux utilisateurs professionnels,
4. La Visualisation des données devient une exigence de l'entreprise,
5. Les organisations insufflent l'analyse de données en toutes les activités commerciales,
6. Les entreprises se tournent vers la décision PMML- Predictive Model Markup Language - une norme pour les modèles statistiques et d'exploration de données (Olavsrud, 2014). [16]

III.3.3.5. Avenir de Big Data Analytiques 3.0

Il est prévu que les

i) Les volumes de données vont continuer à croître,
ii) SQL et Spark continueront pour améliorer la façon dont les données sont analysées,
iii) L'analyse prescriptives seront construits pour l'analyse d'affaires logicielles,
iv) En temps réel en continu un aperçu des données jouera un rôle majeur,
v) Les marchés Algorithme émergera,
vi) L'informatique et de l'analyse cognitive émergeront comme changeurs de jeu,
vii) Plus entreprises générer de la valeur et les revenus de leurs données,
viii) Les entreprises appliquant Analytiques témoin de 430 milliards de dollars en avantages de productivité par rapport à leurs concurrents n'utilisant pas l'analyse des données d'ici 2020,
ix) Des données rapides et utilisables remplaceront les données importantes (Mars, 2016). [16]

IV. Les outils du Big Data pour L'analyser des réseaux sociaux

Les analyses Big Data peuvent se révéler très utiles pour une entreprise, notamment pour booster les ventes, comprendre la clientèle et améliorer la gestion interne.

Cependant, pour convertir les données en informations exploitables, il est nécessaire de s'équiper de meilleurs outils analytiques. Voici une sélection de 7 outils Big Data très utilisées dans les entreprises.

- **Hadoop**

Créé par Apache, Hadoop est un Framework logiciel open source facilitant le traitement distribué de très larges ensembles de données au travers de centaines de serveurs opérant parallèlement. De nombreuses entreprises utilisent Hadoop depuis bien longtemps pour trier et analyser le Big Data. Ce Framework repose sur des modèles de programmation simples pour assurer le traitement des données et les rendre disponibles sur des machines locales.

- **Storm**

Storm est un autre produit développé par Apache. Il s'agit d'un système de traitement Big Data en temps réel open source. Il peut être utilisé aussi bien par les petites et les grandes entreprises. Storm est adapté à tous les langages de programmation, et permet de traiter des données même si un nœud connecté du cluster ne fonctionne plus ou si les messages sont perdus. Storm est également parfait pour le RPC distribué et le Machine Learning en ligne. Il s'agit d'un bon choix parmi les outils Big Data car il s'intègre aux technologies existantes.

❖ **Hadoop MapReduce**

Hadoop MapReduce est un modèle de programmation et un Framework logiciel permettant de créer des applications de traitement de données. Développé à l'origine par Google, MapReduce autorise le traitement rapide et parallèle de larges ensembles de données sur des clusters de nœuds.

Ce Framework a deux fonctions principales. D'abord, la fonction de mapping permettant de séparer les données à traiter. Deuxièmement, la fonction de réduction permettant d'analyser les données.

❖ **Cassandra**

Apache Cassandra est une base de données No SQL hautement scalable. Elle est capable de surveiller de larges ensembles de données répartis sur divers clusters de serveurs et sur le Cloud. Initialement développée par Facebook pour répondre à un besoin d'une base de données suffisamment puissante pour la fonction de recherche in box. Désormais, cet outil Big Data est utilisé par de nombreuses entreprises disposant de larges ensembles de données comme Netflix, eBay, Twitter et Reddit.

❖ **OpenRefine**

OpenRefine est un outil open source conçu pour les données désordonnées. Cet outil permet de nettoyer rapidement des ensembles de données et de les transformer dans un format exploitable. Même les utilisateurs sans compétences techniques peuvent se servir de cette solution. OpenRefine permet également de créer instantanément des liens entre les ensembles de données.

❖ **Rapidminer**

Rapidminer est un outil open source capable de prendre en charge des données non structurées, tels que des fichiers texte, des logs de trafic et des images. Concrètement, cet outil est une plateforme de science des données reposant sur la programmation visuelle pour les opérations.

Des fonctions comme la manipulation, l'analyse, la création de modèles, et l'intégration rapide dans les processus de business processus comptent parmi les avantages de Rapidminer.

❖ **Mongo DB**

Mongo DB est une base de données No SQL open source très utilisée pour ses hautes performances, sa disponibilité élevée et sa scalabilité. Elle est appropriée pour le traitement Big Data grâce à ses fonctionnalités et adaptée à des langages de programmation comme JavaScript, Ruby et Python. Mongo DB est facile à installer, à configurer, à maintenir et à utiliser. [17]

V. Conclusion

Grâce à les ouvertures de Big Data les entreprises sont dirigé vers un nouveau monde c'est le monde analytique qui est axé sur l'analyse des données et surtout les données des réseaux sociaux.

L'analyse des données des réseaux sociaux SNA (Social Network Analytics) joue un rôle très important pour les entreprises, car une entreprise peut stimuler ses ventes, augmenter son efficacité, améliorer ses opérations, ses services à la clientèle et sa gestion des risques avec une bonne plateforme de « Big Data Analytics ».

L'utilisation des outils Big Data tout seul ne suffit pas pour réaliser le succès donc il a besoin de scientifiques compétents et spécialisées dans l'analyse de données pour mettre en valeur les données collectées.

Chapitre III : Cas d’étude.

I. Introduction

L'intérêt que porte l'analyse des réseaux sociaux via les outils Big data pour ses utilisateurs attire l'attention de plusieurs secteurs. Pour satisfaire les besoins de ses clients, prévoir le succès ou l'échec des projets et ainsi augmenter les profits.

Pour concrétiser tout ce que nous avons étudié dans les chapitres précédents, nous avons pensé à prendre un cas réel et effectuer les tests et les analyses possibles qui permettent d'exploiter les technologies Big data pour l'analyse des réseaux sociaux nous avons choisi le cas le plus réel qui peut éclaircir le plus des résultats de l'analyse à effectuer.

Twitter.com est le site que nous avons choisi pour la collection et l'analyse de données. Considéré comme meilleur site qui facile la collection des données, notre choix a été effectuée selon plusieurs critères et raisons qu'on va détailler plus profondément dans le prochain chapitre.

Pour développer notre solution nous avons procédé à une étude des différentes étapes qui entrent en jeu lors de l'analyse des réseaux sociaux, de cette étude nous avons dégagé notre approche d'analyse des réseaux sociaux avec les solutions Big data.

Dans les prochaines sections on va décrire notre approche globale et ces principales étapes qui nous permettent d'analyser les données d'un réseau social concernant un sujet d'analyse spécifique.

II. Présentation

De nos jours il est difficilement envisageable d'ignorer la nécessité de l'analyse des réseaux sociaux et son rôle pertinent dans le développement des entreprises, il existe plusieurs façons de profiter de l'analyse des réseaux sociaux selon le domaine choisi, ci-dessous nous avons situé quelque exemples d'utilisation de cette analyse :

- L'analyse des réseaux sociaux pour voir ou prévoir l'opinion des publics sur un sujet d'intérêt public.
- L'analyse des réseaux sociaux pour concevoir des statistiques sur un produit ou un service à offrir notamment dans les campagnes de marketing.
- L'analyse des réseaux sociaux pour prévoir les résultats des élections présidentiels (exemple : les Etats Unis 2016, France 2017).

Il y a plusieurs façons d'utiliser un réseau social pour présenter et publier un sujet d'intérêt public afin de connaitre l'avis des internautes-parmi elles, on trouve deux méthodes, la méthode directe et la méthode indirecte :

A. **La méthode Direct :** la façon traditionnel d'utiliser les réseaux sociaux pour avoir les points des vues des internautes, cette technique consiste à publier le sujet et attendre la réaction des utilisateur, pour exploiter les réseaux sociaux avec cette méthode, l'analyseur à besoin des ressources, et faire des efforts comme la création des pages web et faire des publications du sujet sur les sites web, l'exemple le plus célèbre

d'utilisation des réseaux sociaux de cette façon est faire un sondage ou un questionnaire sur un thème ou événement

Figure 9 : Un sondage pour les élections présidentiel du France 2017 publie dans le site http://france-presidentielle.fr

Figure 10 : Exemple d'un questionnaire oriente vers les internautes.

B. **La méthode Indirecte :** le but centrale des réseaux sociaux est de connecter les internautes entre eux pour communiquer et publier leurs cultures, le sujet est déjà présenté dans les commentaires et les discussions des internautes dans les forums et les micros blogs de façon indirect, les ressources nécessaires pour utiliser les réseaux sociaux de cette façons sont juste le terme ou la phrase qui définir le sujet et une technique capable a faire la recherche

Figure 11 : Liste des commentaires concernant un film publie sur Youtube.

les internautes peuvent aussi exprimer leurs opinions de manières différentes, ils utilisent les commentaires, les images et les différentes émotions…etc. pour exprimer leurs points des vues selon le type du réseau et le sujet lui-même, la figure suivante porte une collection des tweets concernant un film sur twitter.

Figure 12: Collections des tweets concernant un film sur Twitter.

Le problème qui se pose dans l'analyse des réseaux sociaux et comment extraire les sentiments et identifier les points des vues des internautes.

Plusieurs travaux ont été menés par de grandes compagnies tel que IBM, GOOGLE et YAHOO pour résoudre ce problème, chacune de ces compagnies propose sa propre technologie pour analyses les réseaux sociaux.

III. L'approche globale

Dans cette section, nous allons présenter en détail les différentes étapes de notre approche d'analyse qui se base sur la solution proposée par Apache Software Foundation. Les différentes étapes que nous avons suivies sont illustrées dans la figure suivante :

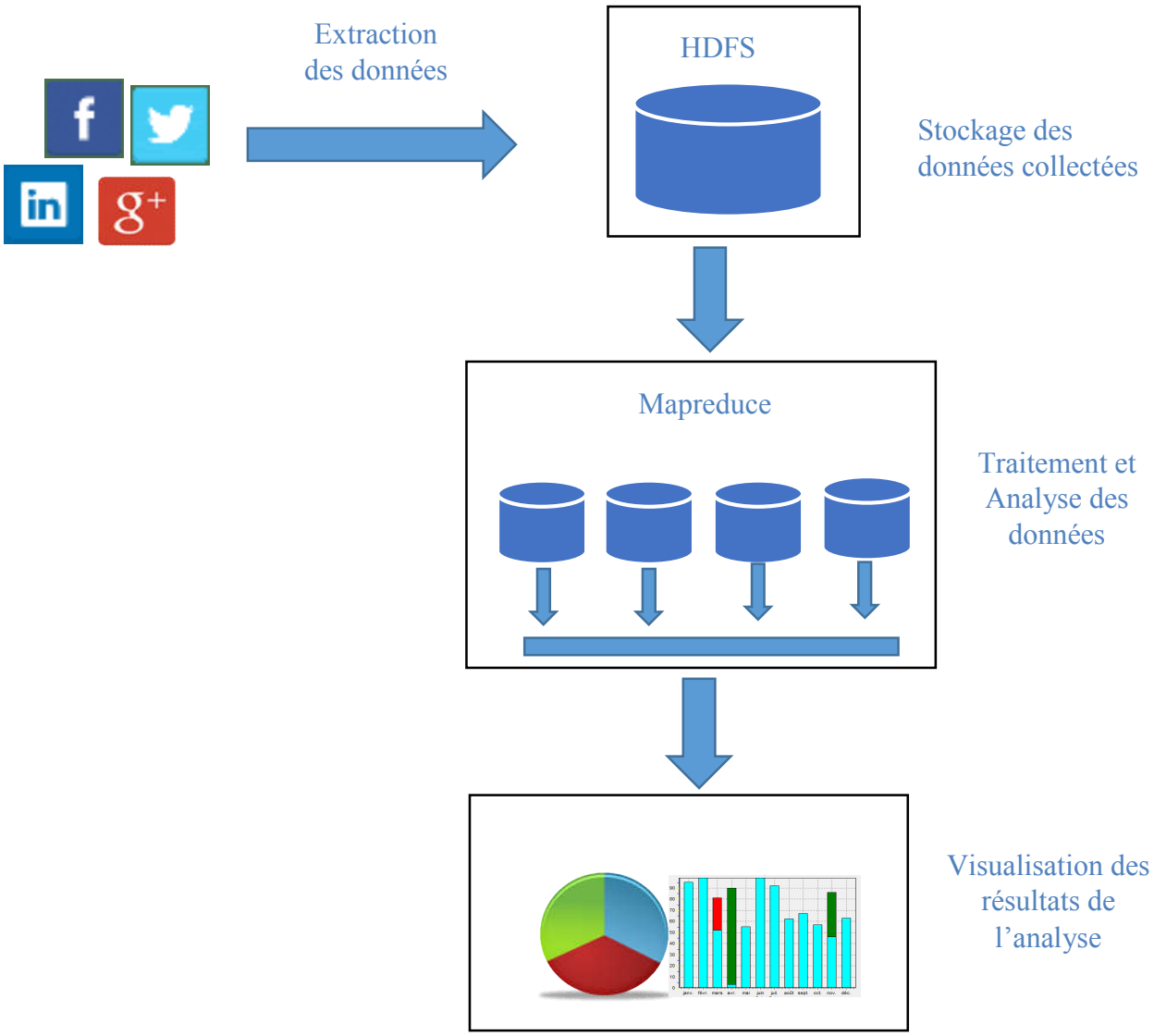


Figure 13 : Le schéma général de l'approche globale

III.1 Description de l'approche globale

L'approche que nous avons suivie passe par quatre étapes principales :

- L'extraction des données
- Le stockage des données
- Le traitement et l'analyse des données
- La visualisation des résultats de l'analyse

III.1.1. L'extraction des données

La première étape de l'approche sert à la collection et l'extraction des données des réseaux sociaux, ses données peuvent être des textes simples, des émotions, des images et des vidéos. La capacité de gérer, manipuler, partager, traiter et stocker ces données, varie d'un réseau social à un autre.

Plusieurs outils sont développés pour l'extraction et la collection des données et des réseaux sociaux, parmi Les outils très utilisé on trouve Facebook API, Twitter API, Apache Storm, Apache Flume, Apache Pig et d'autres.

Nous avons utilisé Apache Flume pour collecter les données de Twitter. Apache Flume utilise le streaming pour la collection des données, en d'autre façon il collecte les données en temps réel.

Apache Flume nous permet de créer un agent flume pour la recherche et la collection des données des twitter, il prend comme paramètre le nom du sujet à rechercher. Cet agent extraire tous les données fournis par twitter ou se trouve le nom du sujet, ces données peuvent être : noms d'utilisateurs, tweets, adresse web, titre des images, les métadonnées, etc. La figure 8 montre un exemple réel des données collectées en utilisant Apache Flume.

III.1.2. Le stockage des données collectées

Après le choix du réseau sociale et l'extraction de ses données, un nouveau besoin se produit c'est la capacité de stocker ces données localement pour les traiter. C'est le rôle du Big Data et ses technologies qui nous permet de stocker un grand volume de données el les manipuler rapidement (pour plus de détails sur l'importance du Big Data et ses technologies voire le chapitre 01).

Nous avons stocké et chargé les données collectées dans la première étape dans le HDFS (Hadoop Distributed File System) pour les traiter aux prochains étapes, la figure suivante montre le stockage des données dans le HDFS

Le HDFS est un système de fichiers distribué, extensible et portable développé par Hadoop à partir du Google FS. Écrit en Java, il a été conçu pour stocker de très gros volumes de données sur un grand nombre de machines équipées de disques durs banalisés. Il permet l'abstraction de l'architecture physique de stockage, afin de manipuler un système de fichiers distribué comme s'il s'agissait d'un disque dur unique. **[1]**

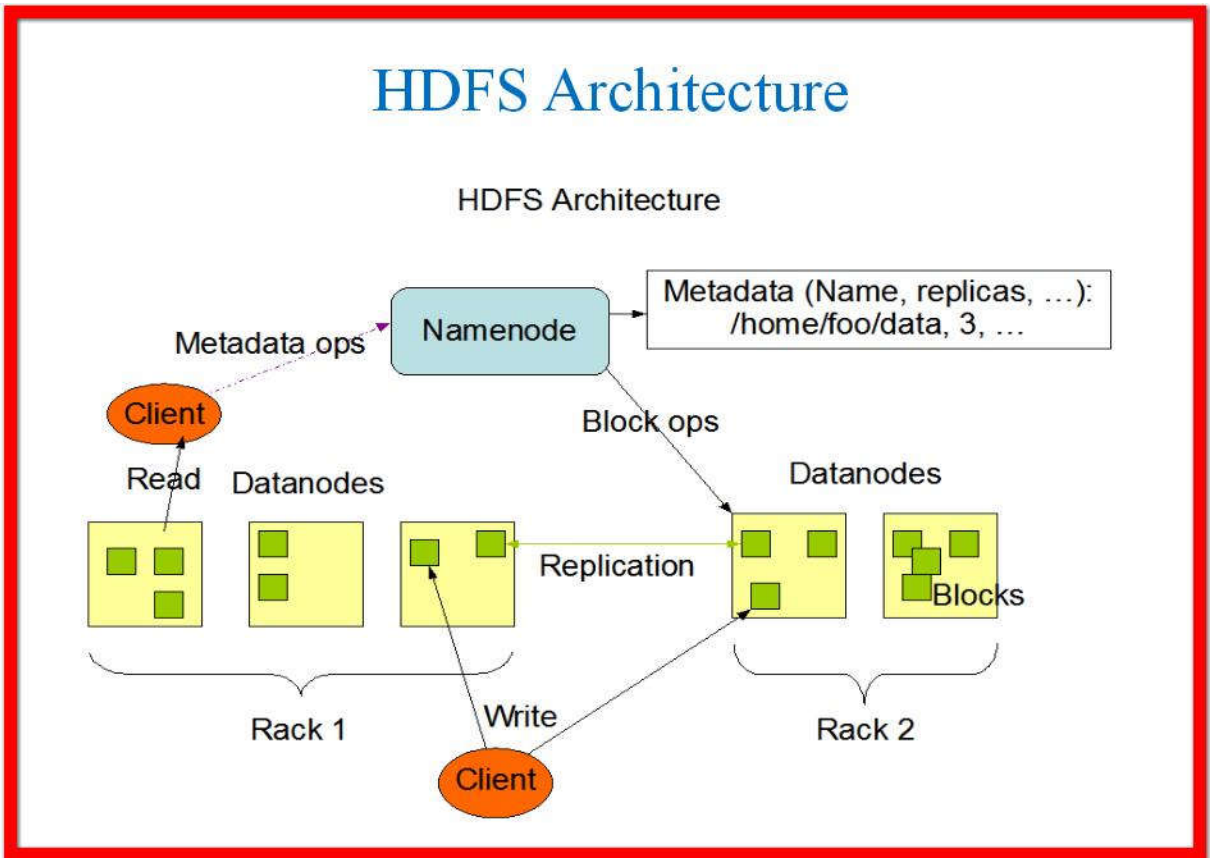


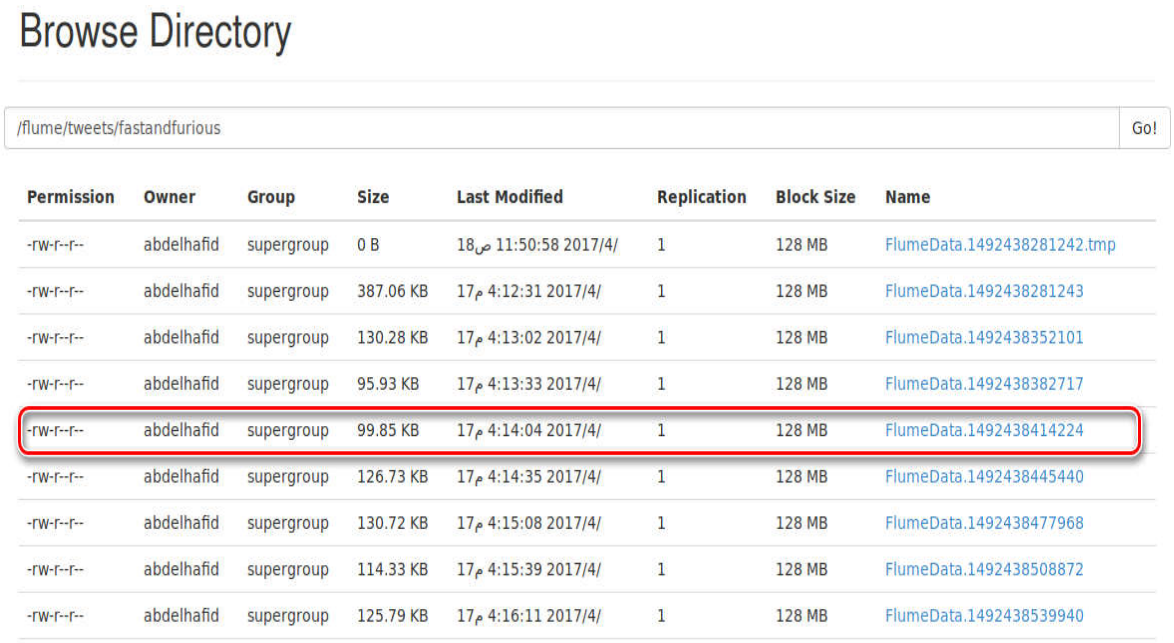
Figure 14 : l'architecture de Hadoop Distributed File Système.

III.1.3. Le traitement et l'analyse des données

C'est l'étape fondamentale de l'approche qui sert à traiter et analyser les données stockées dans le HDFS, le rôle de cette étape est reparti en deux sous étapes : pré traitement et traitement

a. Pré traitement des données

Puisque les données collectées sont des métadonnées des internautes, nous avons besoin de sélectionner juste les données pertinents par rapport à l'analyse ses données sont stockées avec un autre format dans le HDFS. Les trois figures suivantes montrent comment les données sont stockées dans le HDFS



Browse Directory

/flume/tweets/fastandfurious Go!

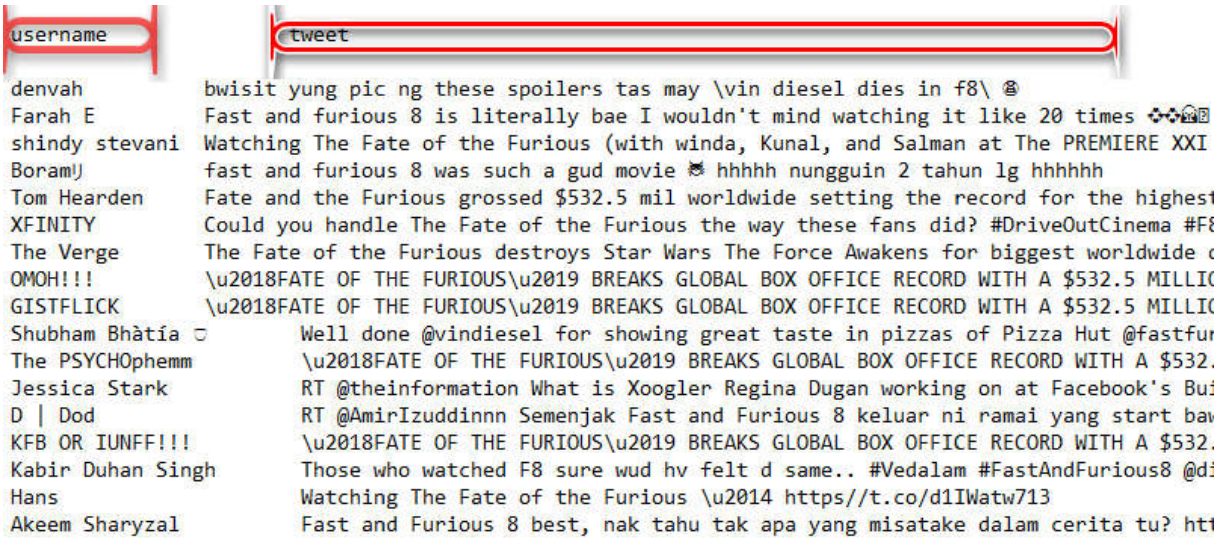
Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	abdelhafid	supergroup	0 B	18ص 11:50:58 2017/4/	1	128 MB	FlumeData.1492438281242.tmp
-rw-r--r--	abdelhafid	supergroup	387.06 KB	17م 4:12:31 2017/4/	1	128 MB	FlumeData.1492438281243
-rw-r--r--	abdelhafid	supergroup	130.28 KB	17م 4:13:02 2017/4/	1	128 MB	FlumeData.1492438352101
-rw-r--r--	abdelhafid	supergroup	95.93 KB	17م 4:13:33 2017/4/	1	128 MB	FlumeData.1492438382717
-rw-r--r--	abdelhafid	supergroup	99.85 KB	17م 4:14:04 2017/4/	1	128 MB	FlumeData.1492438414224
-rw-r--r--	abdelhafid	supergroup	126.73 KB	17م 4:14:35 2017/4/	1	128 MB	FlumeData.1492438445440
-rw-r--r--	abdelhafid	supergroup	130.72 KB	17م 4:15:08 2017/4/	1	128 MB	FlumeData.1492438477968
-rw-r--r--	abdelhafid	supergroup	114.33 KB	17م 4:15:39 2017/4/	1	128 MB	FlumeData.1492438508872
-rw-r--r--	abdelhafid	supergroup	125.79 KB	17م 4:16:11 2017/4/	1	128 MB	FlumeData.1492438539940

Figure 15 : Exemple des données collectées depuis Twitter.

"protected":false,"screen_name":"_farahhhe","id_str":"4904895059",
"profile_link_color":"1DA1F2","id":4904895059,"geo_enabled":false,
"profile_background_color":"F5F8FA","lang":"en","profile_sidebar_border_color":"C0DEED",
"profile_text_color":"333333","verified":false,
"profile_image_url":"http://pbs.twimg.com/profile_images/848176456081473536/xdU5_z9X_normal.jpg",
"time_zone":null,"url":null,"contributors_enabled":false,"profile_background_tile":false,
"profile_banner_url":"https://pbs.twimg.com/profile_banners/4904895059/1480187598",
"statuses_count":2473,"follow_request_sent":null,"followers_count":264,
"profile_use_background_image":true,"default_profile":true,"following":null,
name":"Farah E","location":"Doha, Qatar" "profile_sidebar_fill_color":"DDEEF6",
"notifications":null}}
{"extended_entities":{"media":[{"display_url":"pic.twitter.com/YpPZT8TAk6",
"source_user_id":350114400,"type":"photo","media_url":"http://pbs.twimg.com/media/C9bM_NKV0AA64C4.jpg",
"source_status_id":853094577753931776,"url":"https://t.co/YpPZT8TAk6","indices":[79,102],
"sizes":{"small":{"w":680,"h":383,"resize":"fit"},"large":{"w":1280,"h":720,"resize":"fit"},
"thumb":{"w":150,"h":150,"resize":"crop"},"medium":{"w":1200,"h":675,"resize":"fit"}},
text":"Fast and furious 8 is literally bae I wouldn't mind watching it like 20 times
"in_reply_to_status_id_str":null,"in_reply_to_status_id":null,"created_at":"Mon Apr 17 14:13:45 +0000 2017",
"in_reply_to_user_id_str":null,"source":"<a href=\"http://twitter.com/download/android\" rel=\"nofollow\">
|

Figure 16 : Une partie du contenue des données collectées.

Nous avons sélection le nom d'utilisateur et son tweet pour les traités par la suite, les données sélectionnées sont stockées avec un autre format dans le hdfs comme il montre la figure suivante :



username	tweet
denvah	bwisit yung pic ng these spoilers tas may \vin diesel dies in f8\
Farah E	Fast and furious 8 is literally bae I wouldn't mind watching it like 20 times
shindy stevani	Watching The Fate of the Furious (with winda, Kunal, and Salman at The PREMIERE XXI
Boram!J	fast and furious 8 was such a gud movie hhhhh nungguin 2 tahun lg hhhhhh
Tom Hearden	Fate and the Furious grossed $532.5 mil worldwide setting the record for the highest
XFINITY	Could you handle The Fate of the Furious the way these fans did? #DriveOutCinema #F8
The Verge	The Fate of the Furious destroys Star Wars The Force Awakens for biggest worldwide c
OMOH!!!	\u2018FATE OF THE FURIOUS\u2019 BREAKS GLOBAL BOX OFFICE RECORD WITH A $532.5 MILLIC
GISTFLICK	\u2018FATE OF THE FURIOUS\u2019 BREAKS GLOBAL BOX OFFICE RECORD WITH A $532.5 MILLIC
Shubham Bhàtía ⊃	Well done @vindiesel for showing great taste in pizzas of Pizza Hut @fastfur
The PSYCHOphemm	\u2018FATE OF THE FURIOUS\u2019 BREAKS GLOBAL BOX OFFICE RECORD WITH A $532.
Jessica Stark	RT @theinformation What is Xoogler Regina Dugan working on at Facebook's Bui
D | Dod	RT @AmirIzuddinnn Semenjak Fast and Furious 8 keluar ni ramai yang start bav
KFB OR IUNFF!!!	\u2018FATE OF THE FURIOUS\u2019 BREAKS GLOBAL BOX OFFICE RECORD WITH A $532.
Kabir Duhan Singh	Those who watched F8 sure wud hv felt d same.. #Vedalam #FastAndFurious8 @di
Hans	Watching The Fate of the Furious \u2014 https//t.co/d1IWatw713
Akeem Sharyzal	Fast and Furious 8 best, nak tahu tak apa yang misatake dalam cerita tu? htt

Figure 17 : Exemple d'une collection de données préparées pour les analyser.

b. Traitement des données

Le traitement des données consiste à déterminer comment utiliser les technologies du Big Data pour analyser, extraire et classer les points de vue des internautes.

Puisque les données collectées et préparées aux étapes précédentes sont volumineuses, le traitement de ses données va prendre un long temps et pour minimiser ce temps et optimiser le fonctionnement de cette approche nous avons utilisé Hadoop Mapreduce.

b.1. Rappel sur MapReduce

MapReduce est un patron d'architecture de développement informatique, inventé par Google, dans lequel sont effectués des calculs parallèles, et souvent distribués, de données potentiellement très volumineuses, typiquement supérieures en taille à 1 téraoctet.

Les termes « map » et « reduce », et les concepts sous-jacents, sont empruntés aux langages de programmation fonctionnelle utilisés pour leur construction (map et réduction de la programmation fonctionnelle et des langages de programmation tableau).

MapReduce possède les caractéristiques suivantes :

Le modèle de programmation du MapReduce est simple mais très expressif. Bien qu'il ne possède que deux fonctions, map() et reduce(), elles peuvent être utilisées pour de nombreux types de traitement des données, les fouilles de données, les graphes… Il est indépendant du système de stockage et peut manipuler de nombreux types de variable.

- Le système découpe automatiquement les données en entrée en bloc de données de même taille. Puis, il planifie l'exécution des tâches sur les nœuds disponibles.
- Il fournit une tolérance aux fautes à grain fin grâce à laquelle il peut redémarrer les nœuds ayant rencontré une erreur ou affecter la tâche à un autre nœud.
- La parallélisation est invisible à l'utilisateur afin de lui permettre de se concentrer sur le traitement des données.

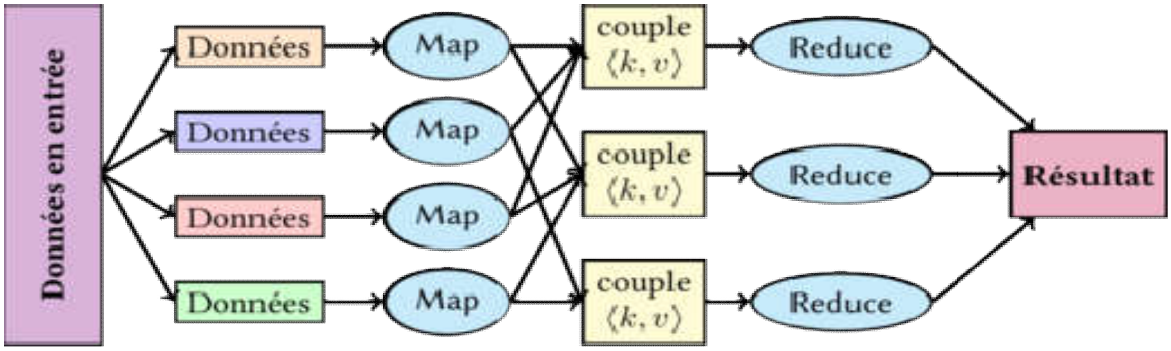


Figure 18 : Schéma de fonctionnement du MapReduce.

b.2. Modèle de programmation MapReduce

MapReduce est un modèle de programmation popularisé par Google. Il est principalement utilisé pour la manipulation et le traitement d'un nombre important de données au sein d'un cluster de nœuds.

Un cluster MapReduce utilise une architecture de type Maître-esclave où un nœud maître dirige tous les nœuds esclaves.

Le programme MapReduce utilisé dans notre approche est composé par deux fonctions map() et reduce().

- ✓ Dans l'étape Map le nœud analyse un problème et le découpe en sous- problèmes, et les délègue à d'autres nœuds (qui peuvent en faire de même récursivement). Les sous-problèmes sont ensuite traités par les différents nœuds à l'aide de la fonction Map.

Autrement dit, le nœud origine doit traiter un fichier texte et le partitionner en sous-fichiers, ces derniers sont affectés aux différents nœuds pour les traitées de façon indépendant et parallèlement à l'aide de la fonction Map.

Dans chaque nœud la fonction map reçoit comme entrée une partie d'un fichier texte et le décomposé en lignes, tel que chaque ligne est considéré comme une tweet (dans le cas d'analyse des données des Twitter), puis calculer la valeur sémantique de chaque tweet a l'aide de la fonction SentimentValue(), et en fin écrire le couple (tweet, valeur sémantique) dans un autre fichier.

```
// En pseudo code cela donnerait

Map (file Inputfile) {
foreach tweet in Inputfile {
// cles est la valeur du sentiment d'un tweet
    int cles = SentimentValue(tweet);
     List.add(tweet, cles);
       Outputfile.Write(tweet, cles);   }

       return List;  }
```

Figure 19 : La fonction Map().

Nous mettrons une explication bien détaillée sur le principe de fonctionnement de la méthode SentimentValue dans **la section (III.1.5. La Méthode SentimentValue).**

- ✓ Vient ensuite l'étape Reduce, où les nœuds les plus bas font remonter leurs résultats aux nœuds parents qui les avait sollicités. Celui-ci calcule un résultat partiel à l'aide de la fonction Reduce (réduction) qui associe toutes les valeurs correspondantes à la même clé à une unique paire (clé, valeur). Puis il remonte l'information à son tour.

En plus, la fonction reduce nous permet de calculer le nombre des éléments de chaque classe sémantique en reçois comme entrée le couple (valeur sémantique, liste des valeurs sémantique) puis calculer le nombre des éléments de cette classe sémantique.

```
// En pseudo code cela donnerait

      Reduce (int cles, List  values) {
      /* result est le nombre des element d'une classe
          (Positive/Négative/Neutre)  */
      int result = 0;
      foreach v in values
              {  if (v.getClass() ==cles) result ++;  }
      return result;        }
```

Figure 20 : La fonction reduce().

À la fin du processus, le nœud d'origine regroupe les différents résultats envoyés par les nœuds fils, ces résultats seront utilisés dans la prochaine étape de la visualisation.

III.1.4. La visualisation des résultats de l'analyse

La visualisation des résultats est une étape très importante dans toutes les approches et les algorithmes informatique, elle aide l'utilisateur à comprendre son problématique et faire des décisions logique.

La visualisation des résultats de l'analyse consiste à interpréter et visualiser les résultats obtenus sous forme des graphes et statistiques pour aider les utilisateurs à prendre leurs décisions et aussi les argumenter.

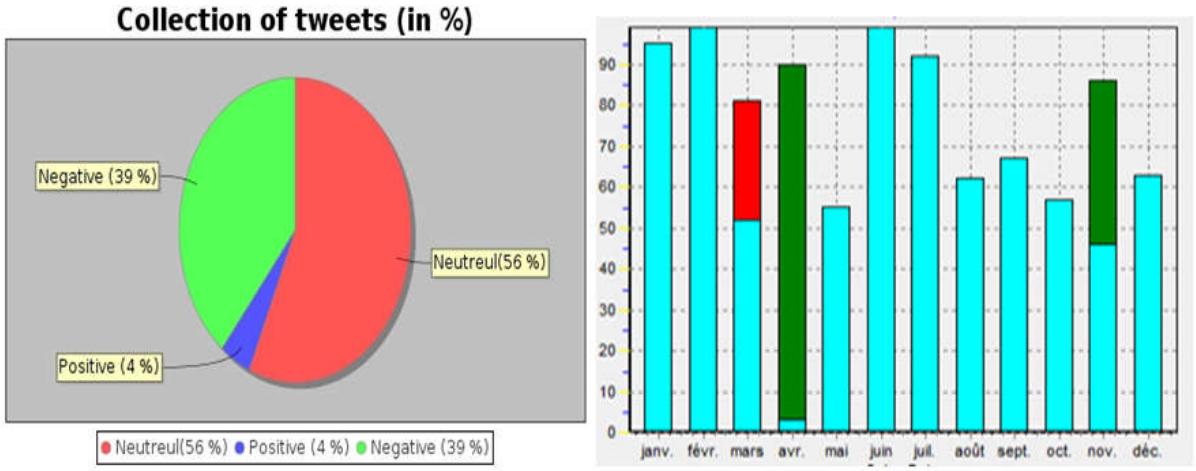


Figure 21: Exemple de visualisation des résultats d'analyse.

Pour notre approche la visualisation des résultats sert à représenter les résultats obtenus dans l'étape précédente sous forme d'un tableau contient les 03 champs suivant : le nom d'utilisateur, son commentaire (tweet) et la classe sémantique (positive, négative ou bien neutre) dans laquelle son commentaire a été affecté. En plus de ça dans cette étape on représente le nombre de chaque classe par une charte simple.

III.1.5. La Méthode SentimentValue

La méthode SentimentValue est le cœur de notre approche, elle nous permet d'extraire et classer les sentiments cachés dans un texte (une tweet). Cette méthode n'est pas notre propre méthode, mais elle est la réutilisation et l'implémentation de la fonction findSentiment de l'outil Stanford CoreNLP.

Stanford CoreNLP fournit un ensemble d'outils d'analyse du langage naturel. Il peut donner les formes de base des mots, leurs parties de la parole, qu'ils soient des noms d'entreprises, de personnes, etc., normaliser les dates, les heures et les quantités numériques, indiquent le sentiment, extraient les relations particulières entre les mentions d'entité, etc.

Le fonctionnement de la méthode SentimentValue est basé sur l'enchainement de 06 processus, comme il est illustré dans la figure suivante :

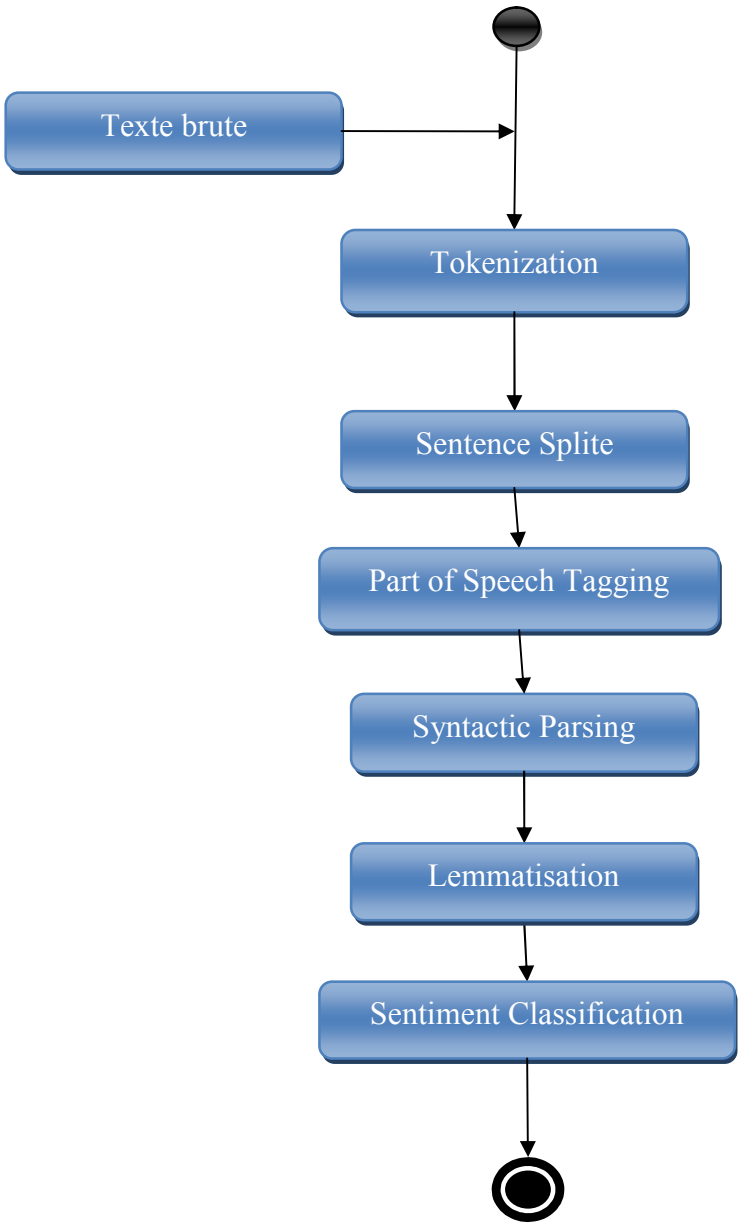


Figure 22 : Diagramme de fonctionnement de la fonction SentimentValue ().

III.1.5.1. Tokenization

Ce processus a pour but de découper le texte en plusieurs **tokens** qui sont les éléments porteurs de sens les plus simples. La tokenisation serait d'utiliser un simple découpage en mots graphiques, c'est-à-dire de séparer les mots en fonction des espaces présents entre eux.

Figure 23 : Exemple de tokenization d'un text.

III.1.5.2. Sentence Splite

Le rôle de ce processus est de découpe les tokens en plusieurs mots. La plus simple méthode pour découper un texte est d'utiliser les espaces blancs et les signes de ponctuations (point, virgules, trois points, etc.) comme il apparu dans l'exemple suivante :

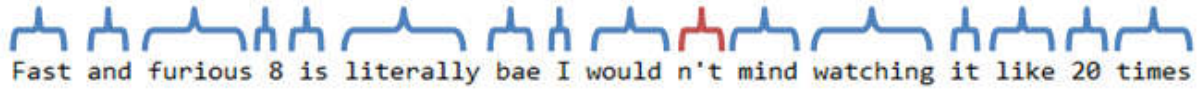


Figure 24 : Exemple d'une sentence splite d'un texte.

III.1.5.3. Part of Speech tagging

Il sert à déterminer le rôle ou la partie d'un mot dans la phrase (par exemple un verbe ou adjective).

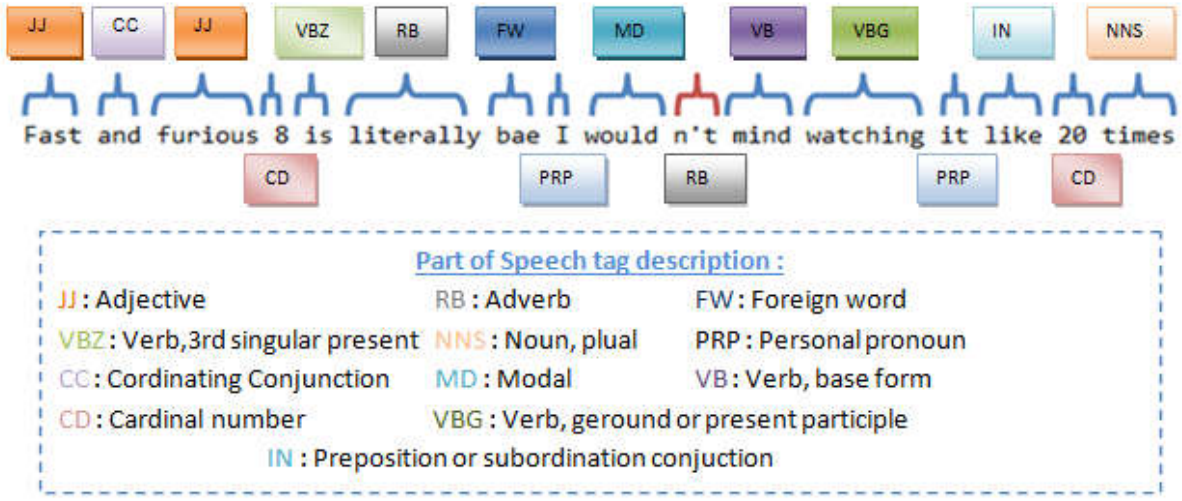


Figure 25 : Exemple de Part of Speech tagging d'un texte.

III.1.5.4. Syntactic Parsing

Son rôle est de dégager une représentation de la structure d'un texte, de manière à mettre en lumière les relations syntaxiques entre les mots. Son fonctionnement est basé sur un dictionnaire (le vocabulaire) et sur un ensemble des règles syntaxiques (la grammaire), pour déterminer les syntagmes, ou constituants, présents dans la phrase et les organiser selon leur hiérarchie dans la phrase.

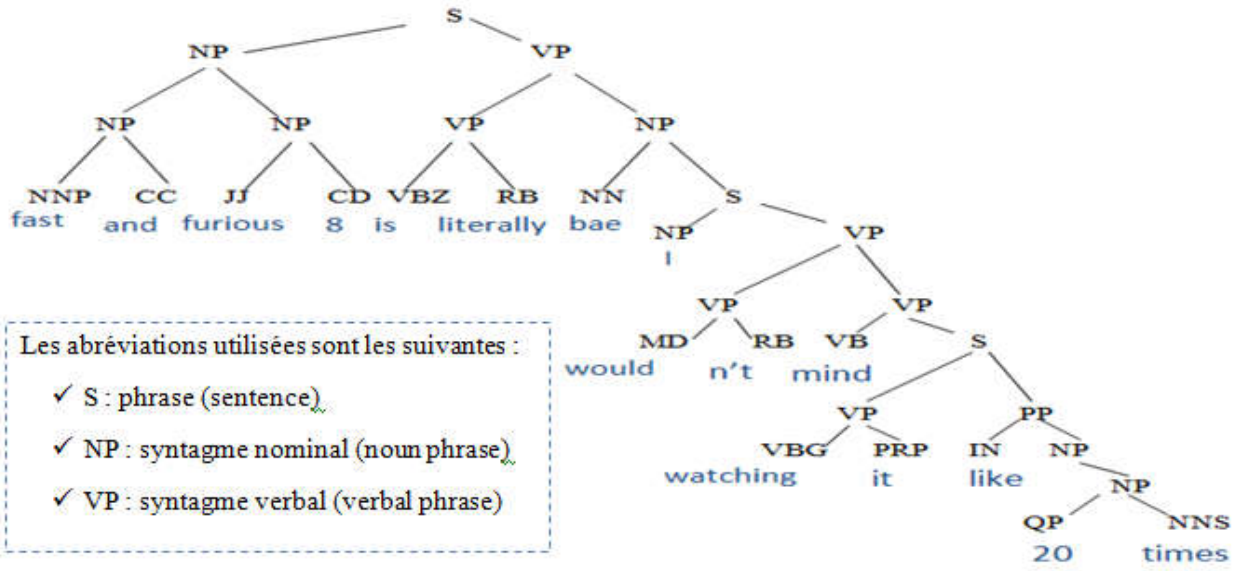


Figure 26 : Exemple d'un arbre syntaxique.

III.1.5.5. Lemmatisation

Chaque mot est considéré comme une composition de morphèmes (ou unités minimales de sens) : un ou des mots racines, ou **lemmes**, et des sous-mots, ou **flexions** (typiquement des préfixes ou suffixes apposés au lemme). Ce processus sert à trouver les mots racines en éliminant les préfixes ou suffixes apposés à ces mots.

La lemmatisation des mots est une étape très importante parce qu'il facilite la comparaison des mots avec les mots enregistrés dans un dictionnaire contient des mots et leurs classifications.

III.1.5.6. Sentiment Classification

La classification des sentiments joue un rôle très important, elle nous permet de classer le texte traité précédemment en 05 classe différentes, ces classes sont : very-negative, negative, neutrel, positive et very-positive. Cette affectation se fait par une comparaison de chaque mot avec un dictionnaire définie au préalable. Ce dernier a été produit par l'analyse d'un corpus de données contient 10,662 phrases (tweet) colletées depuis twitter. Ce dictionnaire a été préparé et valider par une groupe de grands chercheurs de l'université de Stanford.

Après l'affectation de chaque mot à son classe la fonction regroupe les mots un par un en respectant leurs hiérarchiques dans l'arbre syntaxique. La figure suivante montre un exemple complet d'une classification d'un texte (tweet posée sur twitter) avec la représentation de l'arbre syntaxique et la classe sémantique de chaque nœud :

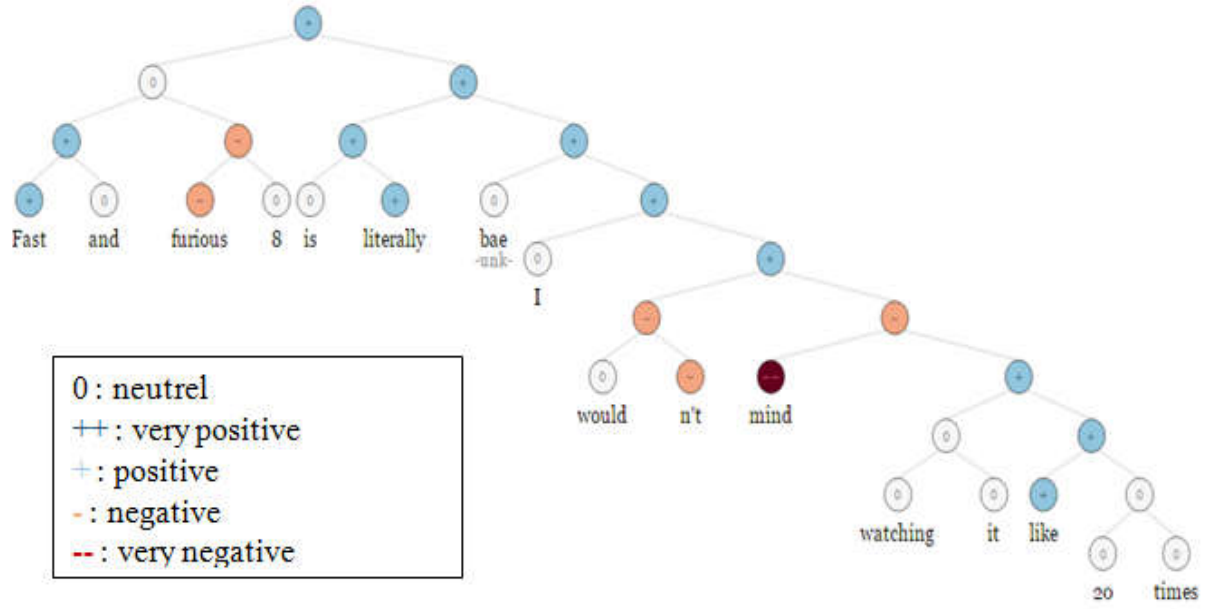


Figure 27 : Exemple d'une classification d'un texte.

IV. Conclusion

Ce chapitre a été consacré à la description de l'approche proposée pour l'analyse des réseaux sociaux via les outils du big data, nous avons cité et expliqué en détails les différentes étapes de notre approche.

Dans le prochain chapitre nous allons passer à l'étape d'implémentation et illustrer un exemple pratique de notre approche.

Chapitre IV : Implémentation.

I. Introduction

Toute approche à besoin d'un logiciel ou une application pour montrer un exemple pratique d'utilisation de cette approche, nous avons enrichie notre approche par une application développé sous java en utilisant les différentes outils des Big Data (Hadoop, Flume et Hive).

Dans le chapitre précédent, nous avons proposé et détaillé une approche pour l'analyse des réseaux sociaux, cependant ce chapitre a été crée afin de mettre la lumière sur l'exemple pratique de notre approche et la description des différentes outils et technologies utilises.

II. Le choix du réseau social

La première question qui peut être posée concernant notre choix est celle d'avoir choisi Twitter comme source de données. Il existe d'autres réseaux sociaux qui peuvent être aussi analysés. Facebook notamment est beaucoup plus populaire auprès d'une population davantage diversifiée. Le choix de Twitter repose alors sur un ensemble des critères:

- Le premier est d'ordre pratique. La grande majorité des messages postés sur Twitter étant publics, leur collecte et leur traitement automatique via l'API du service est facilitée. À l'inverse, les espaces d'échange sur Facebook sont davantage cloisonnés et l'accès conditionné par l'autorisation délivrée par tel ou tel membre à devenir son « ami ».
- Le deuxième raison qui motive notre choix est corrélative à la première. Twitter étant essentiellement une plateforme de diffusion publique. L'utilisation intensive de Twitter par le personnel politique, les journalistes, les blogueurs, les militants politiques et les activistes en témoigne.
- Sur Twitter, on parle de tout. Vraiment. Et même avant tous les autres réseaux sociaux. Souvent les informations partagées sur Facebook ont 2 jours de retard par rapport à Twitter, on parle ici sur la réaction en temps réel.
- Les tweet sont public et visible sur le fil d'actualité de tous les suiveurs (followers) d'un compte Twitter sans aucun algorithme.
- Vous pouvez Tweeter sur n'importe quel sujet. Twitter est un réseau social. Pour cette raison, de nombreux postes sont de nature sociale. Vous y trouverez le plus souvent des mises à jour quotidiennes. C'est La multiplicité des sujets.
- Le plus important est que Twitter fournit une API gratuit et compatibles avec la pluparts des langages des programmation (Java, Python,C++…etc.) pour la collection des données.
- Nous avons prend en considération le nombre d'utilisateur de Twitter, tel que Twitter comptait plus de 328 millions d'utilisateurs actifs par mois dans le monde (Avril 2017, selon une étude fait par le groupe investor de Twitter. https://investor.twitterinc.com) .

Bien sure, il existe d'avantages de critères qui permettent d'appuyer notre choix. Mais nous avons mentionné juste les plus pertinents.

III. Les outils utilisés

Nous avons utilisés un ensemble des outils pour faire notre application, chacune de ces outils joue un rôle très important et accomplie une tache spécifique, cet ensemble des outils est composé de :

III.1. Apache Hadoop

III.1.1. Hadoop

Hadoop est l'acronyme du ***High-Availability Distributed Object-Oriented Platform***. Il s'agit d'une plate-forme open source qui fournit aussi bien les capacités de stockage que de traitement. La communauté Apache (Yahoo) a repris le modèle MapReduce introduit par Google en 2004 et a proposé sa solution open-source Hadoop. Cette solution est construite autour de 2 concepts fondamentaux : HDFS et MapReduce.

Bien qu'il puisse aussi bien s'installer et fonctionner sur une seule machine, la vraie puissance de Hadoop ne sera visible qu'à partir d'un environnement composé de plusieurs ordinateurs (un cluster). Hadoop est donc la réponse à un constat simple : la multiplication de l'espace disque ne va pas avec l'accélération de la lecture des données. La solution serait alors de distribuer les données en plusieurs parties pour les stocker sur plusieurs machines. L'enjeu se situe alors dans le mécanisme qui gère l'accès partagé à cette ressource. [18]

III.1.2. HDFS

Le modèle MapReduce ne définit aucun système de fichier à utiliser. *Hadoop Distributed File System* (HDFS) est un système de fichiers distribué fournit par Hadoop. HDFS utilise le réseau pour manipuler et accéder aux données et utilise les ressources des autres ordinateurs clients présents sur ce réseau. De plus, HDFS est un système de fichier adapté pour travailler avec de très grands volumes des données tel que Google BigTable (1 GB et plus). Le grand plus de ce système est l'universalité. Il n'a pas besoin d'une machine très puissante pour fonctionner correctement tout comme Il a été conçu pour fonctionner sur des machines "ordinaires". [18]

Le système HDFS intègre un mécanisme de tolérances aux pannes, et peut ainsi garantir la disponibilité et l'intégrité des données malgré une défaillance système (plantage d'une machine). Il permet un traitement rapide des fichiers d'une grande taille, cependant HDFS est moins adapté aux fichiers de fichiers de petite taille.

Un block HDFS désigne une portion de données stockée dans le système de fichier. La taille d'un block est par défaut de 64 MB, ce qui est supérieure à celle des systèmes de fichiers standards. Lorsque la taille d'une donnée est inférieure à 64MB, elle n'occupe pas toute la taille du block, ce qui représente une autre différence par rapport aux systèmes de fichiers standard.

HDFS met en œuvre le pattern master-slaves, lequel est un modèle de traitement parallèle d'une ou de plusieurs opérations à travers plusieurs slaves et un *master* (un modèle similaire au modèle client/serveur). Dans HDFS, le master orchestre toutes les opérations et maintient la

cohérence des métadonnées, on l'appelle namenode. C'est lui qui réceptionne la demande du client et la passe directement à ses slaves appelés datanodes. Les datanodes stockent et récupèrent les blocks demandés. Lors de la création d'un fichier au sein de HDFS, l'utilisateur peut spécifier un nombre de réplica. Ainsi HDFS placera ces données sur plusieurs datanodes ce qui permet leurs disponibilités pour les slaves et une extraction plus rapide. [18]

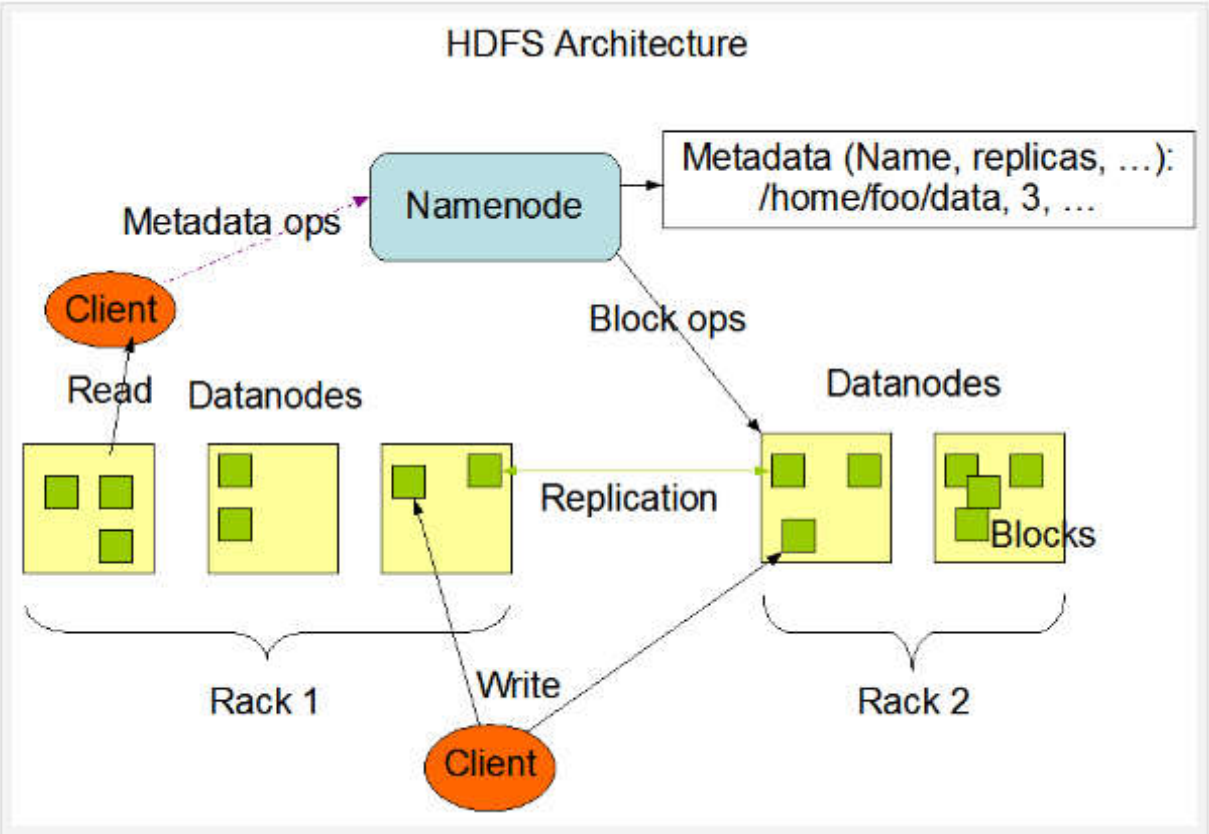


Figure 28 : L'architecture du HDFS.

III.1.3. MapReduce

Comme on a déjà mentionné, Hadoop est un système de traitement parallèle des données qui met en œuvre le modèle MapReduce lequel est un modèle de programmation parallèle permettant de traiter de grands volumes de données. Ce modèle se base sur 2 étapes principales:

Map(), ou étape de mapping (map tasks) : le développeur définit une fonction de mappage dont le but sera d'analyser les données brutes contenues dans les blocks de données stockés sur HDFS pour en sortir des données "correctement formatées". Les données sont considérées comme "correctement formatées" à partir du moment où elles respectent la forme key-value (clé-valeur). [18]

Reduce(), ou étape de réduction (reduce tasks) : cette tâche récupère les données construites dans l'étape du mappage et s'occupe de les trier et les regrouper dans le but de produire les données en sortie. [18]

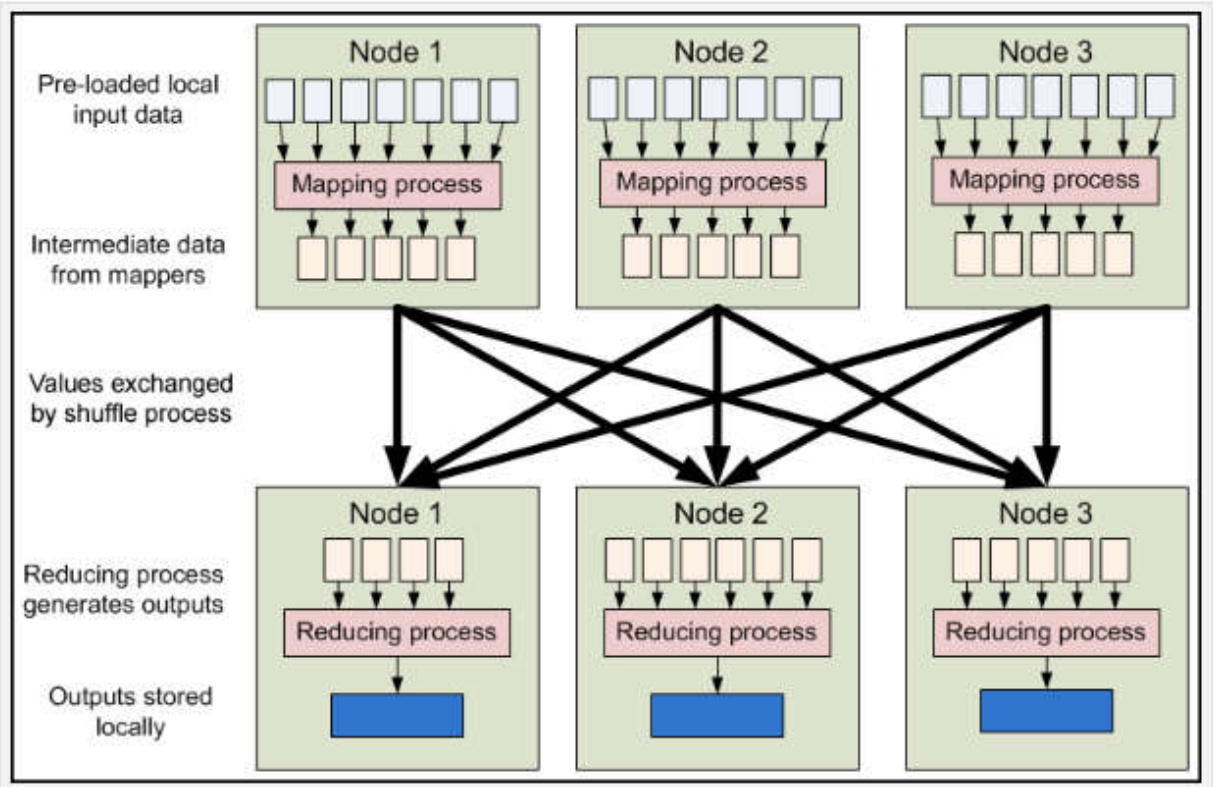


Figure 29 : L'architecture du MapReduce.

III.2. Apache Flume

Apache Flume est un mécanisme d'ingestion outil / services / données pour collecter l'agrégation et le transport de grandes quantités de données de streaming tels que les fichiers journaux, les événements…etc. provenant de diverses sources à une banque de données centralisée.

Flume est un outil très fiable, distribué et configurable. Il est principalement conçu pour copier les données en streaming (log des données) à partir de différents serveurs web à HDFS. [3]

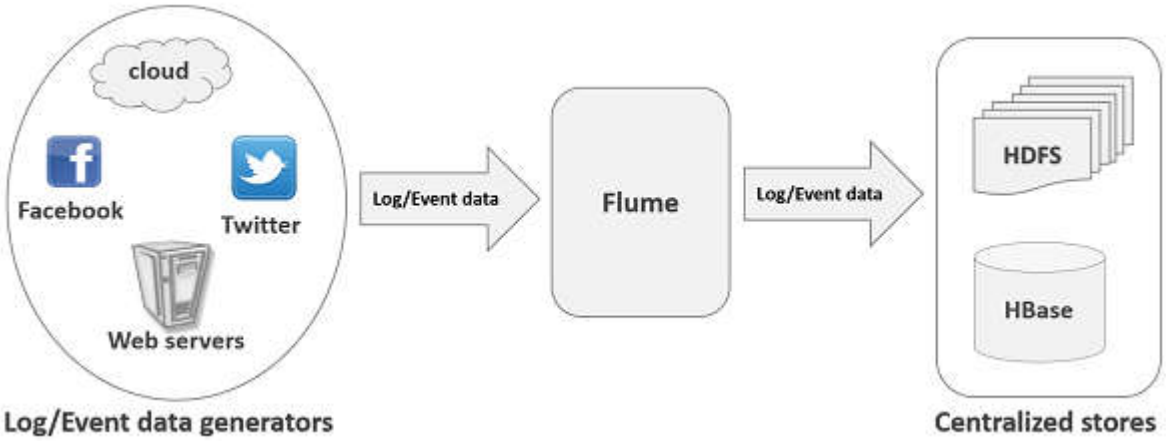


Figure 30 : L'architecture de l'Apache Flume.

III.2.1. Les caractéristiques de Flume

Voici les avantages de l'utilisation Flume

- Apache Flume nous permet de stocker les données dans des magasins centralisés (Hbase, HDFS).
- Lorsque le débit des données entrant dépasse la vitesse à laquelle les données peuvent être écrites sur la destination, Flume agit comme médiateur entre les producteurs de données et les magasins centralisés et fournit un flux continu de données entre eux.
- Flume fournit la fonctionnalité de routage contextuelle.
- Les transactions en Flume sont basées de canal où deux transactions (un émetteur et un récepteur) sont maintenues pour chaque message. Il garantit une livraison fiable des messages.
- Flume est fiable, tolérant aux pannes, évolutive, facile à gérer, et personnalisable.
- Utilisation de Flume, nous pouvons obtenir les données provenant de plusieurs serveurs immédiatement dans Hadoop.
- Avec les fichiers journaux, Flume est également utilisé pour importer d'énormes volumes de données d'événements produits par les sites de réseaux sociaux comme Facebook et Twitter, et les sites Web de commerce électronique comme Amazon et Flipkart.
- Flume prend en charge un grand nombre de sources et types de destinations.
- Flume prend en charge les flux multi-hop, ventilateur dans les flux de ventilateur-out, routage contextuel, etc. [19]

III.2.2. La configuration du Flume

Apache Flume fournit un ensemble d'agents (TwitterAgent, Avro…etc) pour collecter les données des réseaux sociaux et les sites web, le fonctionnement de ses agent est basée sur un fichier de configuration spécifique. [19]

Pour créer un agent Flume on fixe sa configuration de base (source, Channel, keywords, HdfsPath… etc.) dans un fichier portant l'extension .conf.

Dans l'exemple suivant nous avons crée un agent de type TwitterAgent pour collecter les données de Twitter.

```
TwitterAgent.sources.Twitter.type =
     com.cloudera.flume.source.TwitterSource
TwitterAgent.sources.Twitter.channels = MemChannel
TwitterAgent.sources.Twitter.consumerKey =
     7IWC71thW5hCkADhUmj926jkg
TwitterAgent.sources.Twitter.consumerSecret =
     riB8fZxZkF6YViuY5RvvGZEQGA7IXlHdQpfC3OPN7NKduAiDdH
TwitterAgent.sources.Twitter.accessToken =
     806775927350824964-LxVyevRvcS7EpWMV1xLzYkl5mUC8snI
TwitterAgent.sources.Twitter.accessTokenSecret =
     2ilZWtMXSJ6vWZBYkZm3KlkdTLn10trZ7Imnog0uxgIGV
TwitterAgent.sources.Twitter.keywords =    Fast and
     Furious 8, The fate of the Furious
TwitterAgent.sinks.HDFS.channel = MemChannel
TwitterAgent.sinks.HDFS.type = hdfs
TwitterAgent.sinks.HDFS.hdfs.path =
     hdfs://localhost:9000/flume/tweets/Fast8/
TwitterAgent.sinks.HDFS.hdfs.fileType = DataStream
TwitterAgent.sinks.HDFS.hdfs.writeFormat = Text
TwitterAgent.sinks.HDFS.hdfs.batchSize = 1000
TwitterAgent.sinks.HDFS.hdfs.rollSize = 0
TwitterAgent.sinks.HDFS.hdfs.rollCount = 10000
TwitterAgent.channels.MemChannel.type = memory
TwitterAgent.channels.MemChannel.capacity = 10000
TwitterAgent.channels.MemChannel.transactionCapacity =100
```

Figure 31 : La configuration de TwitterAgent pour collecter les données de Twitter.

III.3. Apache Hive

III.3.1. Hive

Apache Hive est un entrepôt données (Data Warehouse) open source pour Hadoop. C'est une abstraction sur Apache Hadoop, pour les familier avec SQL, Hive fournit un langage de haut niveau semblable a SQL, appelé HQL (Hive Query Language), pour interagir avec un cluster hadoop, dans le but de réaliser des analyses sur une masse importante de données. [20]

L'accès aux données se fait via des tables structurées. Il se matérialise par la création d'un plan d'exécution de la requête HQL qui se traduit par la création et l'exécution d'un job MapReduce. Hive aussi, il offre la possibilité, aux familiers du modèle parallèle MapReduce, de pouvoir utiliser des taches de type « mapper » et « reduce », dédiés aux traitements spécifiques de données non supportés par HQL. [20]

Hive vous permet de concevoir une structure sur des données largement non structurées. Une fois que vous avez défini la structure, vous pouvez utiliser HiveQL pour interroger les données sans connaître Java ou MapReduce. [21]

Le langage de requête HQL supporte les opérations de :

- ✓ Sélection

- ✓ Jointure
- ✓ Agrégation
- ✓ Union
- ✓ Ainsi que les sous-requêtes

HiveQL supporte le langage de définition de données (DDL) permettant ainsi la création de :

- ✓ Tables avec des formats de sérialisation spécifiques
- ✓ Partitions
- ✓ Buckets

Hive permet à travers le langage de manipulation de données (DML) le chargement des données dans les tables gérées par Hive. Cette manipulation par les commandes load et insert.

HiveQL est extensible, il intègre :

- ✓ Des scripts MapReduce écrit avec n'importe langage.
- ✓ Des fonctions spécifiques définissent par l'utilisateur
- ✓ Des Types spécifiques [20]

III.3.2. Hive et structure des données

Hive est capable de travailler avec des données structurées et semi-structurées. Par exemple, les fichiers texte et les fichiers xml/json dans lesquels les champs sont délimités par des caractères spécifiques (les espaces, les virgules, le saut de page (\n) et la tabulation (\t)) L'instruction HiveQL suivante crée une table à partir de données délimitées par la tabulation et le sauvegarder comme un fichier texte :

```
CREATE EXTERNAL TABLE employé ( Nom string, Prénom string, Service string, Id string, Salaire double ) ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t ' STORED AS TEXTFILE ;
```

Figure 32 : Exemple de création d'une table dans hive.

III.3.3. Tables internes et externes de Hive

Hive permet de créer deux types de tables :

Interne : les données sont stockées dans l'entrepôt de données Hive. L'entrepôt de données se trouve dans /hive/warehouse/ sur le stockage par défaut du cluster.

Les tables internes sont utilisées quand les données sont temporaires et quand vous voulez que Hive gère le cycle de vie de la table et des données.

Externe : les données sont stockées en dehors de l'entrepôt de données. Les données peuvent être stockées sur tout stockage accessible par le cluster.

Les tables externes sont utilisées quand :

Les données sont également utilisées en dehors de Hive. Par exemple, les fichiers de données sont mis à jour par un autre processus (qui ne verrouille pas les fichiers.)

Les données doivent rester dans l'emplacement sous-jacent, même après suppression de la table. Vous avez besoin d'un emplacement personnalisé, par exemple un compte de stockage non sélectionné par défaut. Un programme autre que Hive gère le format de données, l'emplacement, etc. [21]

III.4. Java

La technologie Java définit à la fois un langage de programmation et une plateforme informatique. Créée par l'entreprise Sun Microsystems (souvent juste appelée "Sun"), et reprise depuis par la société Oracle, la technologie Java est indissociable du domaine de l'informatique et du Web. On la retrouve donc sur les ordinateurs, mais aussi sur les téléphones mobiles, les consoles de jeux, etc.

le langage de programmation informatique orienté objet Java permet de développer des applications client-serveur. Les applications développées en Java peuvent fonctionner sur différents systèmes d'exploitation, comme Windows ou Linux ou Mac OS.

Puisque les technologies du Apache (Hadoop, Flume, Hive…etc) sont programmé sous java nous avons utilise java dans notre application pour éviter les problèmes d'incompatibilités

III.5. Versions des outils utilisées

Nous avons programmé notre application avec les outils et les APIs (Application Programming Interface) situé dans le tableau 1sous le système d'exploitation Linux Ubuntu 17.04 (Zesty Zapus).

Le tableau suivant montre tout 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	8 update 131
03	Jfree Charts	API	1.0.19
04	Stanford CoreNlp	API	3.7.0
05	Apache Hadoop	Plateforme	2.7.3
06	Apache Flume	API	1.7.0
07	Apache Hive	API	2.1.1

Tableau 1 : Versions des outils et APIs utilisés.

IV. Schéma d'exécution de l'application

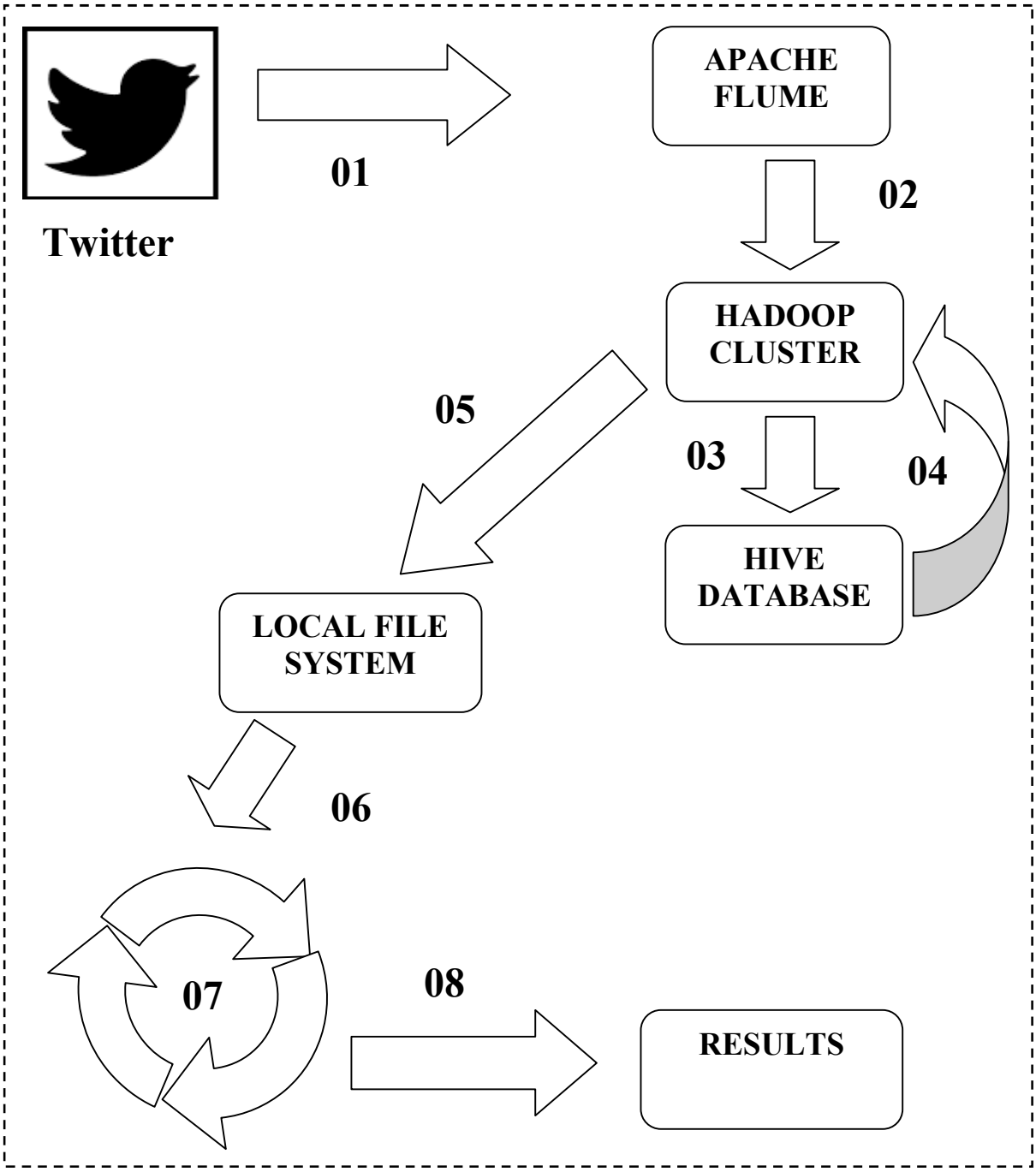


Figure 33 : Schéma d'exécution de notre application.

Le principe de fonctionnement de notre application a été devisé par les 08 étapes suivantes

Etape	Description
01	Extraction de donnés de Twitter
02	Stockage de données collecté dans le cluster
03	Filtrage et transformation des données avec Hive
04	Stockage de données filtrées dans le cluster
05	Transfert les données traitées précédemment dans System local
06	Chargement des données dans le programme Mapreduce
07	Analyse et Classification des données
08	Visualisation des résultats de l'analyse

Tableau 2 : Les 08 étapes de fonctionnement de notre application.

V. Description de l'application

V.1. Page d'accueil

La fenêtre principale de notre application.

Figure 34 : la page d'accueil.

V.2. La fenêtre de l'analyse

C'est la fenêtre principale de notre application, cette fenêtre nous permet d'analyser un fichier (extension .fda) des données collectées par Apache Flume et filtrées par Apache Hive, cette fenêtre aussi port un simple charte pour visualiser la résultat de l'analyse.

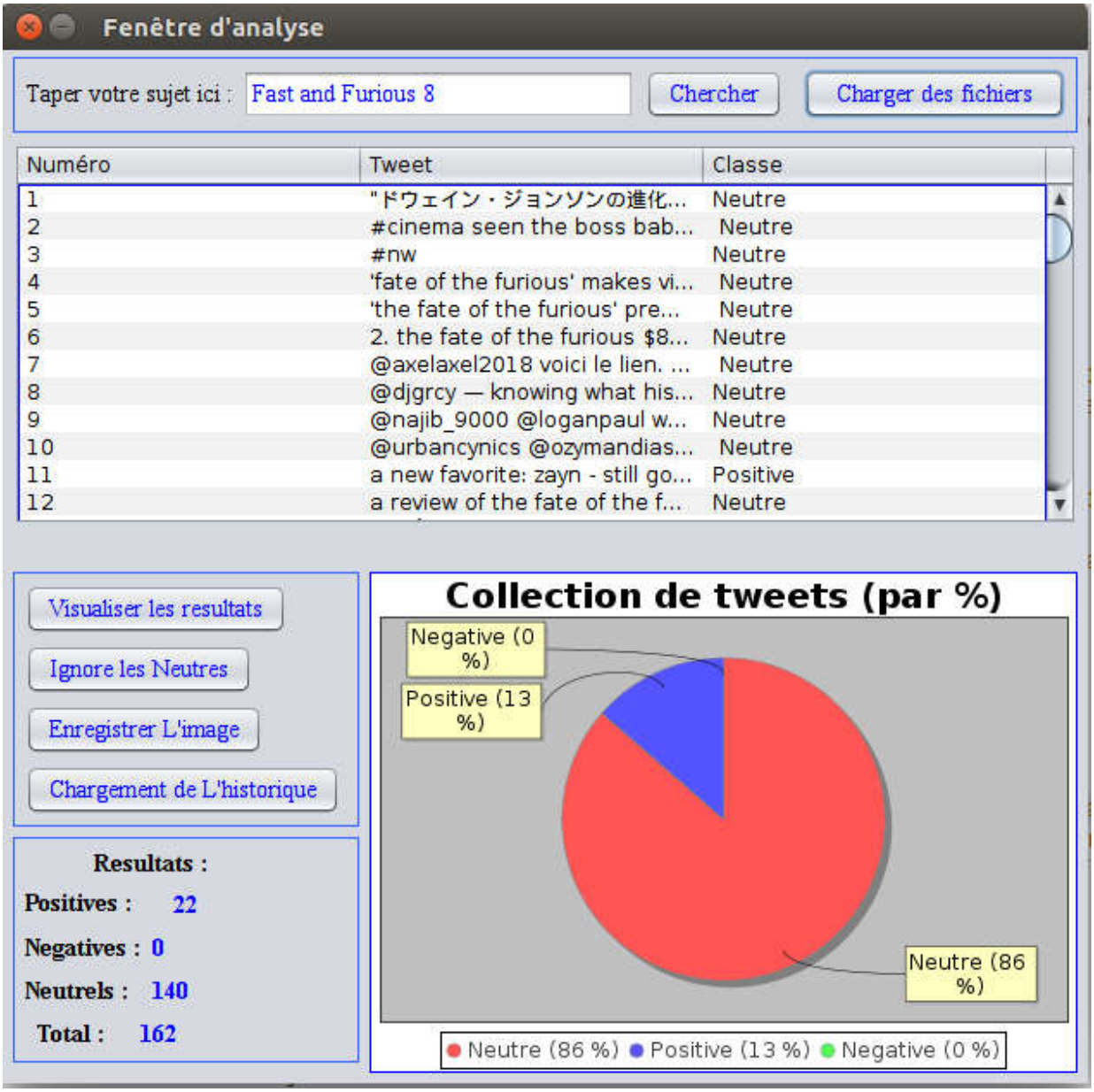


Figure 35 : La fenêtre de l'analyse des sentiments.

V.3. La Fenêtre de Chargement des fichiers

Cette fenêtre nous permet de sélectionner un fichier (extension .fda) pour l'analyser

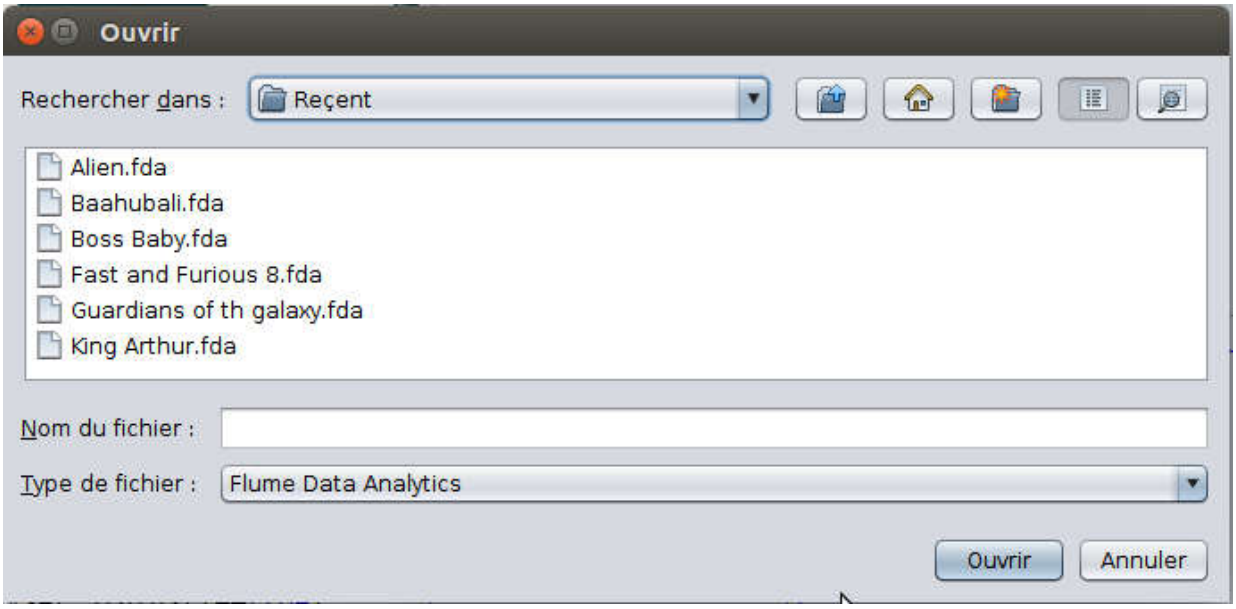


Figure 36 : La Fenêtre de Chargement des fichiers.

V.4. La Fenêtre du terminal

Pour faciliter et optimiser le fonctionnement de notre application nous avons programmé une méthode pour ouvrir automatiquement le terminal dans une fenêtre.

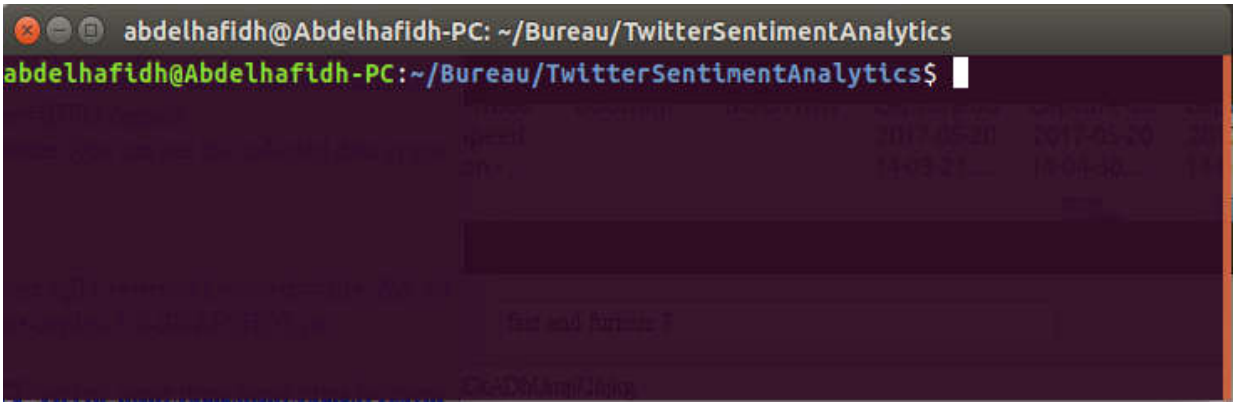


Figure 37 : La Fenêtre du terminal.

V.5. La Fenêtre de guide d'utilisation

Puisque nous ne pouvons pas lancer et faire fonctionner les différents outils du Big Data (Hadoop, Flume, Hive) depuis java, nous avons préparé une liste de commandes pour les exécuter dans le terminal.

L'utilisateur peut faire un copie/coller ces commandes afin de les exécuter dans le terminal

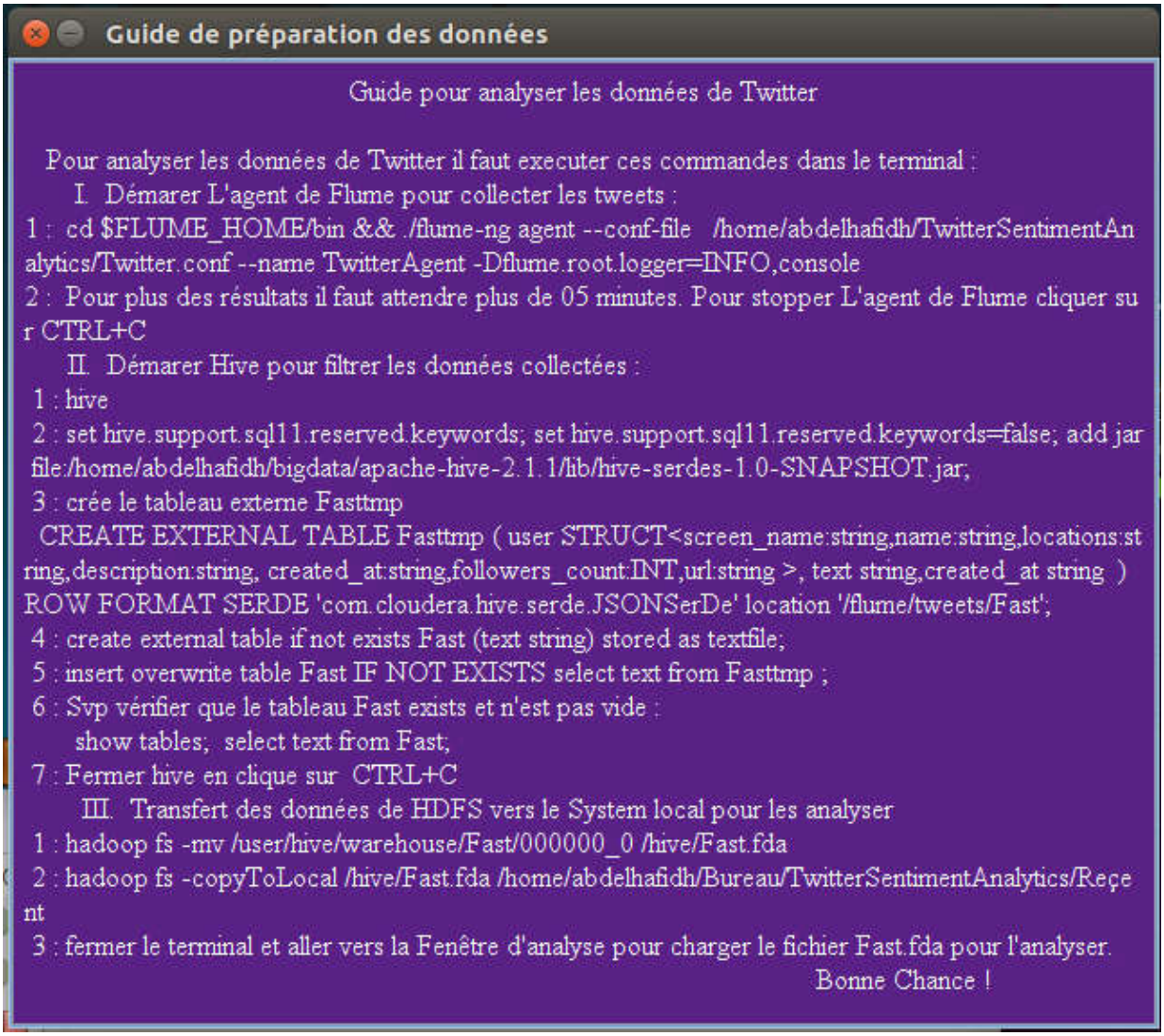


Figure 38 : La Fenêtre de guide d'utilisation.

V.6. La Fenêtre de configuration de TwitterAgent

Nous avons mentionné précédemment que pour collecter les données de Twitter il faut configurer et lancer un agent flume. La configuration de cette agent est spécifié dans un fichier (extension .conf),il faut remplir ce fichier par les différentes champs (Channel, Path,Type…) et les champs d'authentification et de sécurité de l'application crée dans Twitter (AccesToken, AccesTokenSecret, consumerSecret, consumerKey). Afin de remplir ce fichier de façon dynamique nous avons utilisé une fenêtre portant les différents champs nécessaires

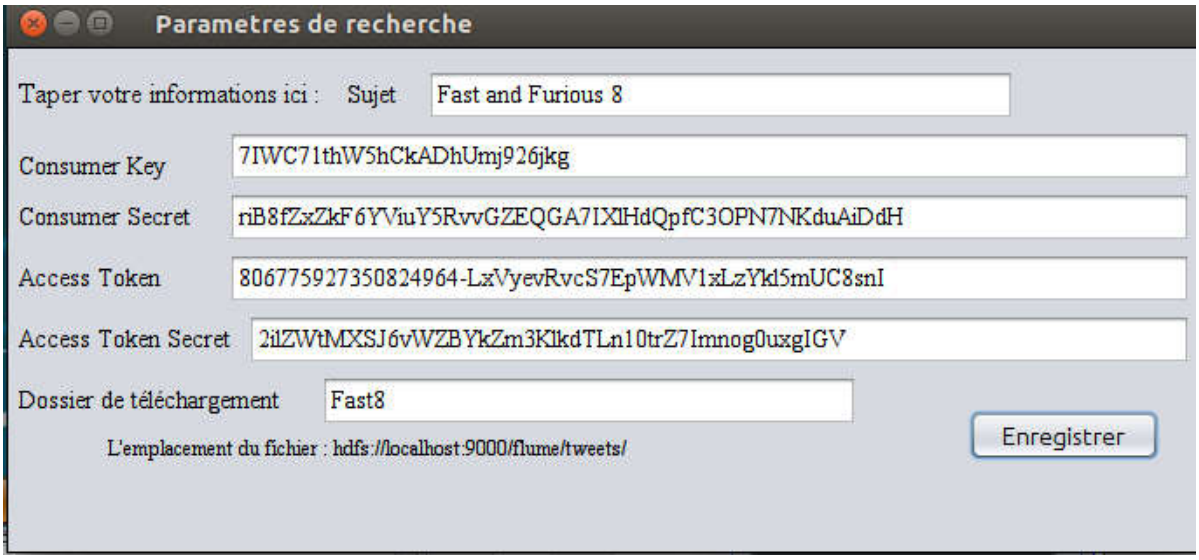


Figure 39 : La Fenêtre de configuration de TwitterAgent.

V.7. La Fenêtre de chargement de l'historique de l'analyse

Pour garder l'historique de l'exécution de notre application nous avons sauvegardé les résultats de l'analyse dans un fichier texte, afin de récupérer l'historique des résultats nous avons créé la fenêtre suivante

Figure 40 : La Fenêtre de chargement de l'historique de l'analyse.

VI. Conclusion

Ce chapitre a été mené pour faire une description détaillé de l'implémentation de notre application, afin de construire cette application nous avons utilisé un ensemble des outils des Big Data (Apache Hadoop, Apache Flume et Apache Hive), ces outils sont gratuit et facile a implémenter.

Dans ce chapitre nous avons cité les différents outils utilisés dans notre application.

Conclusion Générale

Ce projet de fin d'étude consiste à trouver une approche ou méthode pour analyser les réseaux sociaux via les outils de Big Data et de réaliser une application pour implémenter l'approche proposé.

Nous avons commencé par une introduction au Big Data dans laquelle nous avons défini les différents concepts liés au Big Data, ainsi que leurs technologies et leurs applications.

Par la suite nous avons mentionné un peu les réseaux sociaux, leurs types et le lien entre les réseaux sociaux et les Big Data.

Durant le développement de notre approche de l'analyse des réseaux sociaux avec les outils des Big Data nous avons choisi Twitter comme notre source de données, pour implémenter notre approche nous proposons une application pour analyser les sentiments portées par les tweet et les statues postées dans Twitter afin d'identifier L'opinion publique concernant un sujet donné.

En effet, ce travail étant une œuvre humaine, n'est pas un modèle unique et parfait, c'est pourquoi nous restons ouverts à toutes les critiques et nous sommes prêts à recevoir toutes les suggestions et les remarques tendant à améliorer d'avantage cette approche. Etant donné que tout travail informatique a été toujours l'œuvre d'une équipe

Tables de références

N°	Types	Adress / Titre	Consulté le :
1	Site web	http://www.definitions-marketing.com/definition/big-data/	17/01/2017
2	Site web	http://ideas.microsoft.fr/big-data-5-chiffres-secteur-expansion/	17/01/2017
3	Site web	Big Data pour les managers publie sur la portail www.piloter.org Lien direct : http://www.piloter.org/business-intelligence/big-data-quoi-pourquoi-comment.htm	17/01/2017
4	Site web	http://www.astrosurf.com/luxorion/Le Big Data et le data mining.htm	19/01/2017
5	Site web	http://ideas.microsoft.fr/Le big data, c'est quoi Quels enjeux pour votre entreprise - Microsoft pour les PME.htm	19/01/2017
6	Site web	http://www.piloter.org/business-intelligence/Le big data et le processus décisionnel.htm	19/01/2017
7	Site web	https://www.supinfo.com/La problématique du BIG DATA _ SUPINFO, École Supérieure d'Informatique.htm	19/01/2017
8	Site web	http://www.piloter.org/business-intelligence/technologie-big-data.htm	19/01/2017
9	Site web	https://www.digitalwallonia.be/Big Data. La révolution des données.htm	19/01/2017
10	Rapport de stage	rapport de stage écrit par Angeline KONE INSA LYON - Mastère spécialisé SI 2013 publie sur le site web Mémoire Online	25/01/2017
11	Site web	https://www.cnil.fr/fr/reglement-europeen-protection-donnees	25/01/2017
12	Site web	http://planmediassociaux.ca/les-reseaux-sociaux-cest-quoi/	26/02/2017
13	Site web	http://www.lebigdata.fr/reseaux-sociaux-big-data-0811	26/02/2017

14	Article	l'article de : ANALYSE DES RESEAUX SOCIAUX ET COMMUNAUTES EN LIGNE : QUELLES APPLICATIONS EN MARKETING ? **Maria Mercanti-Guérin**	26/02/2017
15	Site web	l'article publié http://www.ciandt.com/card/four-types-of-analytics-and-cognition	26/02/2017
16	Fichier pdf	" **The Evolution of Data Analytics: The Past, the Present and the Future** " Publié par M. **VarunNemmani** De l'université de Missouri-Kansas City Mars, 2016. Traduit en français	27/02/2017
17	Site web	http://www.lebigdata.fr/top-7-outils-big-data-0712	27/02/2017
18	Site web	http://blog.khaledtannir.net/2013/10/hadoop-et-mapreduce-introduction/#.WTVFLdWUfIU	02/03/2017
19	Site web	http://www.w3ii.com/fr/apache_flume/apache_flume-introduction.html	02/03/2017
20	Site web	http://www.linusnova.com/2013/09/hive-le-data-warehouse-de-hadoop/	02/03/2017
21	Site web	https://docs.microsoft.com/fr-fr/azure/hdinsight/hdinsight-use-hive	02/03/2017