Comprendre le BIG DATA

Document réalisé par Khadidjatou BAMBA

Sommaire

Avant Propos

Depuis plus de cinq (05) décennies l'informatique s'est implanté au cœur de nos entreprises, nos hôpitaux, nos ministères, nos foyers …..Etc. Cette forte utilisation de l'informatique à engendré de grands volumes de données qui ne sont pas gérable par les logiciels et matériels classique.

Prenons le cas d'entreprises de taille humaine comme Google et Microsoft, ces grandes filiales qui doivent avoir des milliards de données à conservés. Un autre exemple est celui des entreprises de téléphonie qui ont de grands volumes de données sur les clients et les prestations qui leurs sont offertes.

Cette perplexité dans la gestion de ces grands volumes de données a donné naissance au Big Data. Qu'est ce que le Big Data ? Au travers de cet article vous aurez l'historique du Big Data, La définitions et les Framework etc.., qui vous permettront à comprendre le concept du Big Data en quelque ligne.

Historique du Big Data

Selon Gill PRESS dans un article publié sur Forbes.com le 05 Mai 2013, l'explosion des données est en effet d'abord perçue comme une menace sur la vie privée. Côté technique l'espace de stockage grandit, mais les données s'étendent systématiquement jusqu'à le combler. Dans les années 70, la qualité des données est enfin mise en cause : tout est stocké, il n'est plus utile de faire le tri.

L'expression « Big data » fait finalement son apparition en octobre 1997 dans la bibliothèque numérique de l'ACM, au sein d'articles scientifiques qui pointent du doigt les défis technologiques à visualiser les « grands ensembles de données ». Le Big data est né, et avec lui ses nombreux défis.
Dans les années 2000, alors que l'exabytes (10^{18} bytes) entrent en jeu dans la quantification des données produites annuellement, la valeur du Big data est mise en avant, d'abord pour les bénéfices que peuvent en tirer la recherche dans les secteurs de la physique, de la biologie ou des sciences sociales.
La montée en puissance des sites Facebook, Linkedin, Amazon et Twitter dans les années 2000 et plus particulièrement à partir de 2005, révèle tout le potentiel des données publiques disponibles sur internet. Les succès économiques des grands du web commencent alors à nourrir deux idées principales :

1- Les données brutes accumulées ont une valeur intrinsèque de par les fameuses 3 composantes en V (pour volume essentiellement mais aussi variété et vitesse de leur production donc leur fraîcheur). Cette valeur est monétisable et Facebook en est l'illustration par excellence !

2- Ces données brutes ont une autre valeur liée aux analyses et corrélations auxquelles elles s'offrent et l'information qui en découle participe à la chaîne de valeur de l'entreprise (ex. : l'algorithme de recommandation d'Amazon).

Introduction

L'informatique est devenue indispensable à l'entreprise. Tous les processus et métiers sont touchés : services clients, finances, marketing, productions, logistiques... etc. De cet fait, on compare souvent le système d'information d'une entreprise à l'épine dorsale du corps humain : elle le construit, le soutient, et grandit avec lui. Mais ce corps humain l'entreprise, n'est rien sans des muscles, à savoir ces employés, et sans un flux sanguin continu : les données

Aujourd'hui il ya de plus en plus de données et informations à traiter. L'entreprise produit ses propres donnés, les échanges avec ses clients, ses fournisseurs, ses partenaires, ses actionnaires et en reproduit sans cesse de nouvelles. Cependant au fil du temps bon nombres de grandes entreprises se sont retrouvées avec des volumes de données ingérables. Face à ce problème, la question suivante concernant la gestion de ces volumes :

- Comment les exploiter et les analyser, pour mieux piloter l'activité de son entreprise ?

Pour répondre à cette question, un ensemble de réponse logicielles et matérielles étiquetées « Big Data » à vu le jour. Dans ce document nous vous expliquerons la technologie de Big Data.

Chapitre I : Présentation du Big Data

I. Généralités sur le Big Data

1. Qu'est ce que le Big data ?

Plusieurs définitions ont été donnés au Big Data nous retiendrons dans cet article, celle de Wikipedia qui est la suivante :

« «Big Data, littéralement les grosses données, est une expression anglophone utilisée pour designer des ensembles de données qui deviennent tellement volumineux qu'ils en deviennent difficiles à travailler avec des outils classiques de gestion de base de données. Dans ces nouveaux ordres de grandeur, la capture, le stockage, la recherche, le partage, l'analyse et la visualisation des données doivent être redéfinis. » Il s'agit donc d'un ensemble de technologies, d'architecture, d'outils et de quantités et contenu hétérogènes et changeants, et d'en extraire les informations pertinentes à un cout accessible.

Pour décrire le principe du Big Data, il est coutumier de résumer ses caractéristiques majeures en utilisant quatre lettres "V" :

- Le volume de données à collecter, stocker et traiter
- La vitesse à laquelle les données sont produites et évoluent dans le temps
- Variétés la nature des données
- Valeur qui représente l'importance de la donnée

2. Les caractéristiques du Big Data

1. Le volume

Le volume décrit la quantité de données générées par des entreprises ou des personnes. Le Big Data est généralement associé à cette caractéristique. Les entreprises, tous secteurs d'activité confondus, devront trouver des moyens pour gérer le volume de données en constante augmentation qui est crée quotidiennement.

Les catalogues de plus de 10 millions de produits sont devenus la règle plutôt que l'exception. Certains clients gérant non seulement des produits mais aussi leur propre clientèle peuvent aisément accumuler un volume dépassant le téraoctet de données.

2. *La vitesse*

La vitesse décrit la fréquence à laquelle les données sont générées, capturées et partagées. Du fait des évolutions technologiques récentes, les consommateurs mais aussi les entreprises génèrent plus de données dans des temps beaucoup plus courts.

A ce niveau de vitesse les entreprises ne peuvent capitaliser sur ces données que si elles sont collectées et partagées en temps réel. C'est précisément à ce stade que de nombreux systèmes d'analyse, de CRM, de personnalisation, de point de vente ou autres, échouent. Ils peuvent seulement traiter les données par lots toutes les quelques heures, dans le meilleur des cas. Or, ces données n'ont alors déjà plus aucune valeur puisque le cycle de génération de nouvelles données a déjà commencé.

3. *La variété*

La prolifération de types de données provenant de sources comme les médias sociaux, les interactions *Machine to Machine* et les terminaux mobiles, crée une très grande diversité au-delà des données transactionnelles. Les données ne s'inscrivent plus dans des structures nettes, faciles à consommer.

Les nouveaux types de données incluent contenus, données géo spatiales, points de données matériels, données de géo localisation, données de connexions, données générées par des machines, données de mesures, données mobiles, point de données physiques, processus, données RFID, données issues de recherches, données de confiance, données de flux, données issues des médias sociaux, données texte et données issues du Web.

4. *La valeur*

L'analyse Big Data a pour objectif de créer un avantage concurrentiel unique pour les entreprises, en leur permettant de mieux comprendre les préférences de leurs clients, de segmenter les clients de façon plus granulaire et de cibler des offres spécifiques au niveau de segments précis. Mais les entreprises du secteur public utilisent également Big Data pour éviter les fraudes et économiser l'argent des contribuables et offrir des meilleurs services aux citoyens, dans le domaine des soins de santé par exemple. Des cas d'utilisation du Big data font leur apparitions dans tous les secteurs d'activités.

Remarque : Pourquoi est-il important de comprendre tout cela ?

Le Big Data nous aide à obtenir une meilleure représentation de l'interaction des clients avec l'entreprise. Il permet une meilleure compréhension de ce que les clients aimeraient réaliser à chaque point de contact.

Il minimise ainsi le risque de perdre ces clients lors du passage d'un point de contact vers un autre et garantit la pertinence de l'information qui leur est délivrée, ainsi pour améliorer à la fois la qualité de service, aspect clé pour les clients, et le taux de transformation de ces clients, il est important pour l'entreprise de ne pas perdre de vue les 4 V du Big Data.

II. Les défis de stockage liés à Big Data au sein de l'entreprise

Les quatre composants du Big Data changent les besoins de l'entreprise en matière de protection de données, et relèvent les défis dans la gestion de données ci-dessus explication détaillée :

- Volume. L'augmentation des volumes de données est le défi le plus communément admis pour les responsables du stockage. Ils ont fort à faire avec la réduction des fenêtres de sauvegarde, tout en ayant des cycles de sauvegarde encore plus longs en raison des volumes plus importants à traiter. Ils éprouvent également des difficultés à respecter les exigences imposant des processus de restauration plus courts. Le Big Data accélère ces défis et soulève la question de la réorganisation de l'architecture des processus de sauvegarde ainsi que des questions sur la valeur des données et la question de savoir si toutes les données doivent être de façon égale.

- Variété .L'existence de différents types de données, qui ne sont pas toutes générées au sein de l'entreprise, soulève la question de la gouvernance de l'information. Comment protégez-vous les données qui ont été générées sur le Web social ? Comment pouvez-vous appliquer des politiques à des données qui vivent dans le Cloud, sont analysées dans le Cloud.etc.

- La vitesse. La performance est l'un des caractéristiques clé de Big Data. Et l'un de ses avantages est la réduction du temps de décision. Cela augmente les performances exigées dans l'infrastructure de stockage.

- La valeur. L'objectif de l'analyse du Big Data est de créer une valeur ajoutée pour l'entreprise. Un autre aspect de valeur est de trouver des

données pertinentes et de les rendre accessibles lors du processus de décision, en particulier des informations non structurées.

III. Quelques domaines d'utilisations du Big Data

Une grande partie des cas d'usage du Big Data existaient déjà avant son émergence. Les nouvelles techniques permettent cependant d'aller plus vite et de traiter plus de données. Car aujourd'hui, il existe beaucoup plus de données générées automatiquement (issues du web, des appareils mobiles et de capteurs divers).La plupart des contextes d'utilisations actuelles du Big Data se résume en quelques termes :

- Pressentir la naissance d'une tendance,
- Prédire l'évolution d'un phénomène,
- Repérer des corrélations pour optimiser une stratégie,
- Faire des contrôles pour découvrir une fraude,
- Organiser une communication virale,
- Mieux cibler,

En effet toutes les sociétés et tous les secteurs sont concernés par le Big Data, la vente, commerce, les administrations et le secteur public, les domaines scientifiques et médicaux, la banque/assurance, les télécoms, les usines de production. Ci-dessous quelques domaines applications du Big Data :

1. Marketing

Le Marketing est un client pour le Big Data que ce soit pour de l'analyse prédictive ou de l'analyse de sentiment, que l'on peut définir rapidement pour l'interprétation automatisée de l'opinion exprimée d'un individu. Ce jugement peu être caractérisé par une polarité (positive, neutre, un mélange des deux) et une intensité. Le Big Data est utilisé pour bon nombre de besoins notamment :

- L'e-réputation
- La fidélisation
- L'analyse de comportement
- L'optimisation des prix

2. Protection de la population et prévention

Depuis la fin des années 90, nous sommes entrés dans l'ère du renseignement. En effet de nombreux moyens ont mis en œuvre par les états au nom de la défense du territoire et de la protection des citoyens contre toute menace ou attaque ; de ce fait des milliards de données non structurées sont ainsi collectées sous forme d'images, d'enregistrement audio ou vidéo..Etc. , qu'l faut pouvoir stocker, trier en fonction de la pertinence et analyser afin d'en ressortir des informations critique.

L e Big Data aide à résoudre efficacement des enquêtes policières (analyser, des indices, trouver une corrélation entre plusieurs affaires), ou prévenir un attentat (suivre les déplacements d'un suspect, reconnaissance faciale sur des vidéos… etc.) Il permet donc de réduire le temps de résolution des affaires et d'en augmenter le taux de résolution.

Chapitre II : l'environnement du Big Data

De nombreuses technologies ont été développées pour intégrer, exploiter, gérer et analyser les Big Data, dans ce chapitre une présentation des solutions les plus utilisés sera faite.

I. Système de gestion de base de données NoSQL

NoSQL signifie "Not Only SQL" ' pas seulement SQL en francais" Ce terme désigne l'ensemble des bases de données qui s'opposent à la notion relationnelle des SGBDR.

Le premier besoin fondamentale auquel répond NoSQL est la performance. En effet ces dernières années, les géants du Web comme Google et Amazon ont vu leurs besoins en termes de charge et de volumétrie de données croitre de façon exponentielle. Et c'est pour répondre à ces besoins que ses solutions ont vu le jour. Les architectes de ces organisations ont procédé à des compromis sur le caractère ACID des SGBDR. Ces compromis sur la notion relationnelle ont permis de dégager les SGBDR de leur frein à la scalabilité.

En effet les solutions NoSQL existantes peuvent être regroupées en quatre grandes familles

- Clé/ valeur : Ce modèle peut être assimilé à une hashmap distribuée. Les données sont, représentées par un couple clé/valeur. La valeur peut être une simple chaine de caractères, un objet sérialisé…..Néanmoins, la communication avec la BD se résumera aux opérations PUT, GET et DELETE.

- Orienté colonne : Ce modèle ressemble à première vue à une table dans un SGBDR à la différence qu'avec une BD NoSQL orientée colonne, le nombre de colonnes est dynamique. En effet, dans une table relationnelle, le nombre de colonnes est fixé dés la création du schéma de la table et ce nombre reste le même pour tous les enregistrements dans cette table. Par contre, avec ce modèle, le nombre de colonnes peut varier d'un enregistrement à un autre ce qui évite de retrouver des colonnes ayant des valeurs NULL. Comme solutions, on retrouve principalement HBase (implémentation Open Source du modèle BigTable publié par Google) ainsi que Cassandra (projet Apache qui respecte l'architecture distribuée de Dynamo d'Amazon et le modèle BigTable de Google).
- Orienté document : Ce modèle se base sur le paradigme clé valeur. La valeur, dans ce cas, est un document de type JSON ou XML.

L'avantage est de pouvoir récupérer, via une seule clé, un ensemble d'informations structurées de manière hiérarchique. La même opération dans le monde relationnel impliquerait plusieurs jointures. Pour ce modèle, les implémentations les plus populaires sont CouchDB d'Apache, RavenDB (destiné aux plateformes .NET/Windows avec la possibilité d'interrogation via LINQ) et MongoDB.

- Orienté Graphe : Ce modèle de représentation des données se base sur la théorie des graphes. Il s'appuie sur la notion de noeuds, de relations et de propriétés qui leur sont rattachées. Ce modèle facilite la représentation du monde réel, ce qui le rend adapté au traitement des données des réseaux sociaux. La principale solution est Neo4

II. Les plateformes pour le Big Data

1. Apache Hadoop

Crée par Doug CUTTING 2009, Apache Hadoop est un Framework qui permet le traitement distribué de grands ensembles de données à travers des grappes d'ordinateurs utilisant des modèles simples de programmation. Il est conçu pour évoluer à partir de serveurs uniques à des milliers de machines, offrant à chaque calcul et le stockage local. Plutôt que de s'appuyer sur du matériel à fournir la haute disponibilité, la bibliothèque elle-même est conçu pour détecter et gérer les échecs à la couche d'application, afin de fournir un service hautement disponible sur un cluster d'ordinateurs, chacun d'eux pouvant être sujettes à des défaillances.

Hadoop met à la disposition des développeurs et des administrateurs un certain nombre de briques essentielles :

- Hadoop Distributed File System (HDFS) : Un système de fichiers distribué qui fournit un accès à haut débit aux données d'applications.
- Hadoop FILS : Un cadre pour la planification des taches et la gestion des ressources de cluster.
- Hadoop MapReduce : Un système basé FILS pour le traitement parallèle de grands ensembles de données.
- Hadoop commun : Les utilitaires communs qui prennent en charge les autres modules Hadoop.

Hadoop est écrit en java et soutenu par plusieurs startups américaines. Il est en outre devenu une sorte de standard de fait pour l'écriture d'application de traitement de données ralliant l'ensemble des acteurs majeurs du secteur.

2. Teradata

Teradata est une société informatique américaine qui vend des plateformes de données analytiques, les applications et les services connexes. Ses produits sont destinés à consolider les données provenant de différentes sources et de rendre les données disponibles pour l'analyse.

Les services proposés par Teradata pour le Big Data sont les suivants:

- Concentrer les données,
- Unifier vos données.

3. Netezza

Netezza est une Appliance d'entrepôt de données, qui à été conçue par IBM elle se caractérise par ca simplicité de déploiement, une optimisation immédiate, absence de réglages, une maintenance réduite au maximum. Vous disposez des performances et de la simplicité dont vous avez besoin pour explorer en profondeur les volumes croissants de données et tirer parti de ces dernières pour transformer l'information en action.
Les différents produits d'appliances d'entrepôt de données sont les suivants :

- IBM Netezza 100
- IBM Netezza 1000
- IBM Netezza High Capacity Appliance

III. Briques fonctionnelles en lien avec le Big Data

1. Pig

Pig est un outil de traitement de données qui fait partie de la suite Hadoop et qui permet l'écriture de scripts qui sont exécutés sur l'infrastructure Hadoop sans être obligé de passer par l'écriture de tache en Java via le Frame Work MapReduce. Il dispose en outre de fonctionnalités permettant le chargement de données depuis une source externe vers le cluster HDFS ou de fonctionnalités permettant l'export de données pour utilisation par des applications tierces.

Pig s'appuie sur son propre nommé Pig Latin. Il permet en outre d'accéder à la couche applicative Java. Ce langage est assez simple ce qui permet au développeur venant d'un autre monde que java de produits des scripts de traitement s'exécutant sur Hadoop beaucoup plus rapidement.

Dans la pratique, Pig est surtout utilisé pour charger des données externes vers des fichiers HDFS et transformer des fichiers afin de faciliter leur analyse surtout dans des cas ou plusieurs étapes sont nécessaires (du fait de

la nature procédurale du langage et de sa capacité à stocker des résultats temporaires).

2. Hive

Hive permet l'écriture de tache de traitement de données aux développeurs ne maîtrisant pas pas Java. Là ou Pig définit un langage procédural permettant d'exploiter le cluster, Hive permet de définir des tables structurées de type SQL et de les alimenter avec des données provenant soit du cluster, soit de sources externes.

Une fois le schéma des tables définies et les données insérées, il est possible d'utiliser le langage HiveQL pour requêter ces tables. HiveQL a une syntaxe proche de SQL et permet de réaliser l'essentiel des opérations de lecture permettant de produire des analyses classiques (sélection de champs, somme, agrégat, tri, jointure,…).

Son gros avantage est sa capacité à utiliser une compétence très répandue qui est la connaissance de SQL rendant les développeurs très rapidement opérationnels pour extraire les données.

3. Sqoop

Sqoop est un projet de la fondation Apache qui a pour objectif de permettre une meilleure cohabitation des systèmes traditionnels de type SGBDs avec la plateforme Hadoop.

Il est ainsi possible d'exporter des données depuis la base de données et de procéder aux traitements couteux en exploitant le cluster Hadoop. Les dispositifs de collecte basés sur une base de données sont à ce jour les plus répandus. Il est ainsi possible de procéder à la collecte de données au sein d'applications traditionnelles n'ayant pas la capacité de se connecter au cluster.

Inversement, il est possible d'exporter le résultat d'un traitement vers une base de données tierce afin qu'il soit exploité par une application à des fins de restitution par exemple.

4. HBase

HBase est un système de gestion de base de données non-relationnel distribuée, écrit en Java, disposant d'un stockage structuré pour les grandes tables. Il permet de distribuer les données en utilisant le système de fichiers distribué HDFS (Hadoop Distributed File System) d'Hadoop.

5. Cassandra

Développé par Facebook, Cassandra est une base de donnée orientée colonnes de type NoSQL. Elle supporte le traitement MapReduce et est particulièrement reconnue pour sa capacité à faciliter l'accessibilité des données, quelque soit le volume géré

Remarque :

On distingue deux types de solutions d'entrepôts de données pour le Big Data :

- Les solutions software d'entrepôts de données sont conçues pour simplifier et accélérer l'obtention d'informations synthétiques à partir de l'analyse métier. Elles incluent des dispositifs d'entrepôts de données qui intègrent une base de données, un serveur et un espace de stockage dans un système unique et facile à gérer qui ne nécessite un minimum de configuration et d'administration et permet une analyse plus rapide et plus cohérente.

- Les plateformes d'entrepôts de donnée et d' analyse préconfigurées, réintégrées et optimisées pour les charges de travail, cette offre est enrichies par la prise en charge des grands données de données (Big Data) et de nouveaux types de charge de travail d'analyse, comprenant l'analyse continue et rapide de volumes massifs de flux de données (Big Data) et de nouveaux types de charge de travail d'analyse,
- comprenant l'analyse continue et rapide de volumes massifs de flux de données.

Chapitre III : Les essentielles d'apache Hadoop

I. Hadoop Distribution Files System

HDFS est un système de fichiers distribué, extensible et portable développé par Hadoop à partir du GoogleFS. Conçu pour stocker de très gros volumes de données sur un grand nombre de machines équipées de disques durs banalisées ; 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. En effet une architecture de machine HDFS, aussi appelée cluster HDFS repose sur deux types de composants majeurs :

- NameNode : est un composant qui gère l'espace de noms, l'arborescence du système de fichiers et les métadonnées des fichiers et des répertoires. Il centralise la localisation des blocs de données répartis dans le cluster. Il est unique mais dispose d'une instance secondaire qui gère l'historique des modifications dans le système de fichiers. Ce NameNode secondaire permet la continuité du fonctionnement du cluster Hadoop en cas de panne NameNode d'origine.
- DataNode : est un composant qui stocke et restitue les blocs de données. Lors du processus de lecture d'un fichier, le NameNode est interrogé pour localiser l'ensemble des blocs de données. Pour chacun d'entre-eux , le NameNode renvoie l'adresse du DataNode le plus accessible, c'est-à-dire le DataNode qui dispose de la plus grande bande passante. Les DataNodes communiquent de manière périodique au NameNode la liste des blocs de données qu'ilss hébergent. Si certains de ces blocs ne sont pas assez répliqués dans le cluster, l'écriture de ces blocs s'effectue en cascade par copie sur d'autres.

Voici quelques-unes des principales caractéristiques qui pourraient être d'intérêt pour de nombreux utilisateurs.

- Hadoop, HDFS, y compris, est bien adapté pour le stockage distribué et le traitement distribué à l'aide du matériel de base. Il est tolérant aux pannes, évolutive et extrêmement simple à développer
- MapReduce, bien connu pour sa simplicité et son applicabilité pour grand ensemble d'applications distribuées, est une partie intégrante de Hadoop.
- HDFS est hautement configurable avec une configuration par défaut bien adapté pour de nombreuses installations. La plupart du temps, la configuration doit être réglée que pour de très grands groupes.

- Hadoop est écrit en Java et est pris en charge sur toutes les plateformes majeures.
- Hadoop prend en charge des commandes shell comme d'interagir avec HDFS directement.
- Le NameNode et DataNodes ont construit dans les serveurs Web qui le rend facile à vérifier l'état actuel de la grappe.
- Nouvelles fonctionnalités et améliorations sont régulièrement mises en œuvre dans HDFS. Ce qui suit est un sous-ensemble de fonctionnalités utiles dans HDFS:
    - Les autorisations de fichier et de l'authentification.
    - Rack sensibilisation: prendre l'emplacement physique d'un nœud en compte lors de la planification des tâches et l'allocation de stockage.
    - Safemode: un mode administratif de maintenance.
    - fsck : un utilitaire pour diagnostiquer la santé du système de fichiers, de trouver les fichiers manquants ou des blocs.
    - fetchdt : un utilitaire pour aller chercher DelegationToken et le stocker dans un fichier sur le système local.
    - Rééquilibreur: outil pour équilibrer le cluster lorsque les données sont inégalement répartis entre DataNodes.
    - Mise à niveau et à la restauration: après une mise à jour du logiciel, il est possible de rollback à l'état HDFS 'avant la mise à niveau en cas de problèmes inattendus.

II. MapReduce

MapReduce est un modèle de programmation massivement parallèle adapté au traitement de très grandes quantités de données. MapReduce est un produit Google Corp. Les programmes adoptant ce modèle sont automatiquement parallélisés et exécutés sur des clusters (grappes) d'ordinateurs.

Le principe de fonctionnement de principe MapReduce est le suivant : Le système de traitement temps réel assure le partitionnement et le plan d'exécution des programmes tout en gérant les inhérentes pannes informatiques et indisponibilités. Ainsi, une application typique MapReduce traite plusieurs tera-octets de données et exploite plusieurs milliers de machines. MapReduce est écrit en C++.Un cluster MapReduce utilise une architecture de type Maitre-esclave ou un nœud maitre dirige tous les nœuds esclaves .L'index de Google est généré avec MapReduce.

Ci-dessous quelque caractéristique de MapReduce :

- 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

Chapitre IV : cas d'utilisation du Big Data

Dans ce chapitre nous présenterons des cas concret d'utilisation du Big Data. Quelques exemples récents ou le Big Data a su apporter une valeur ajoutée et des innovations concrètes dans des domaines comme la santé, la physique ou la géopolitique.

1. Le Big Data dans la prédiction des conflits mondiaux

L'outil GDELT (Global Database of Events, Languages and Tones), développé par l'université de Georgetown et accessible de manière open source, compile toutes les actualités (communiqués de presse, articles, discours…) parues depuis 1979. Il applique ensuite des techniques d'analyse sémantique et des algorithmes auto-apprenants pour faciliter la compréhension des événements récents et des principes de cause à effet pour arriver à prédire les conflits mondiaux.

2. Big Data dans l'aide à la recherche contre le cancer

Project Data Sphère met à disposition de tous des données de tests cliniques passés pour permettre à chacun de conduire ses propres analyses, et, dans l'esprit du Crowd-Innovation, d'améliorer les méthodes ou de découvrir des corrélations encore inconnues.

3. Le Big Data nous aide à comprendre le monde

L'entreprise Kaggle, qui met à disposition sa communauté de 150 000 data-scientists pour aider les entreprises à résoudre des défis liés à l'analyse de données, vient de lancer un concours visant à définir un algorithme capable de comprendre les facteurs qui influencent la création d'un boson de Higgs lors de la collision de deux atomes. Le projet est mandaté par le CERN et a été élaboré par deux chercheurs du CNRS.

4. Le Big data permet de gérer les catastrophes naturelles

En utilisant des outils de tracking, d'analyse sémantique et de visualisation en temps réel, l'Organisation Mondiale de la Migration a pu assister les forces locales en dégageant les urgences sanitaires, la localisation des ressources clés et en optimisant l'allocation des ressources sur le terrain lors du typhon qui a frappé les Philippines en 2013.

5. Le Big Data aide à éradiquer les épidémies

Des scientifiques de l'université de Bringhma Youns essaient de simuler la localisation des mouches tsé-tsé dans le but d'aider à contrôler la propagation d'épidémies. De la même manière, la police de Chicago utilise le Big Data et la visualisation de données pour contrôler les populations de rats dans la ville.

Conclusion

Le Big Data, la gestion des grands volumes de données à un champ d'application très vaste et varié. Dans un futur proche le Big Data serait très utile dans la création de nouvelles entreprises, de l'amélioration de la satisfaction clients, la détection d'épidémie, la détection de foyer de tension …etc.

Selon un rapport publié par Gartner le Big data est la technologie qui va générée le plus d'emploi dans l'informatique dans les trois (03) années à venir.