**BIG DATA**

Table des matières

[Chapitre 1 : INTRODUCTION : 2](#_Toc96624983)

[Chapitre 2 : LES CARACTERISTIQUES DU BIG DATA : 2](#_Toc96624984)

[Chapitre 3 : PRINCIPES ET METHODES DU BIG-DATA : 3](#_Toc96624985)

[Chapitre 4 : LES OUTILS DU BIG DATA : 5](#_Toc96624986)

[Chapitre 5 : CONCLUSION : 5](#_Toc96624987)

# 

# : INTRODUCTION :

La BI c’est étudier les phénomènes anciens pour comprendre le futur, pour ce faire

Adhoc : pouvoir se poser n’importe quelle question (faire des requêtes dans la base de données) sur les données à n’importe heure.

Il y a aussi un grand flux de données généré au quotidien (par plusieurs sources, objets connectés, open sources des gouvernements, données météorologiques, géologiques, santé et pharmaceutique (scientifiques), politiques et géopolitiques, objets connectés, etc..), et l’apparition des réseaux sociaux n’a fait qu’exploser la volumétrie de ses données d’une façon exponentielle. (Données massives) hexa octets de données capturées au fil du temps).

Le souci c’est que les projets BI avec les requêtes ad hoc ne permettent pas tout le temp car c’est trop long et ce n’est pas trop efficace car il ne permet de prendre en compte les autres données générées par d’autres sources que celles de l’entreprise (données non structurées et très volumineuse)

Donc les entreprises essayaient de réfléchir à exploiter l’ensemble de ses données pour améliorer leurs opérations, fournir un meilleur service à la clientèle, créer des campagnes de marketing personnalisées et surtout augmenter leurs rentabilités et enfin garder un avantage concurrentiel potentiel par rapport à la concurrence en prenant des décisions commerciales plus rapides, plus éclairées et plus efficaces. Cette méthode alors est ce qu’on appelle le big-data

**Définition du big-data :** Le big data est l’ensemble de concepts, paradigmes, technologies, architectures et procédures permettant à une organisation de très rapidement et efficacement capter, stocker, traiter et analyser de gros volumes de données hétérogènes et changeantes et d’en extraire des connaissances pouvant améliorer leurs positions sur le marché.

# : LES CARACTERISTIQUES DU BIG DATA :

Les caractéristiques du big data furent ses contraintes à une époque récente. Au jour d’aujourd’hui le Bi data est caractérisé par les 6 V.

**Volume :** Le volume des données est la caractéristique la plus souvent citée du bi data. Cette volumétrie provient de différentes sources :

En outre des sources de données traditionnelles (propre aux entreprises) : les likes les commentaires sur les réseaux sociaux, les vidéos, les recherches sur les moteurs de recherche internet,

Objets connectés (IOT), open data des gouvernements et des organisations (locales nationales ou internationales), données scientifiques (médecines, biologiques, pharmaceutiques), géopolitiques (conflits, accords diplomatiques, guerres…), économiques, monétaires, naturelles (météorologiques, géologiques…).

Ces volumes sont en constances augmentation dans le temps d’une façon exponentielle, les statistiques disent que 80 % des données sont produites au cours des deux dernières années. (Depuis 10 ans). De l’ordre de Zettaoctet= 1021octets

**Variété :** Le big data traite différentes informations qui proviennent de différentes sources donc sont des données très variées : on en trouve des données structurées issues des bases de données relationnelles (langage sql), non-structurées (fichiers textes donc no sql), et semi-structurées (les journaux des serveurs web)

**Vélocité :** fait référence à la vitesse avec laquelle les données sont générées et doivent être traitées et analysées en temps réel (mises à jour en temps réel) au lieu des mises à jour quotidiennes, hebdomadaires ou mensuelles appliquées dans des BD traditionnelles. Les applications d’analyse données volumineuses ingèrent mettent en corrélation et analysent les données en temps réel puis fournissent une réponse.

**Véracité :** fait référence au degré de certitudes des ensembles des données, données brutes recueillies auprès de multiples sources tel que les plateformes des réseaux sociaux et les pages web peuvent entrainer de graves problèmes sur la qualité de l’information qui peuvent être difficiles à comprendre et à cerner. Mauvaises données alors conduisent à des analyses inexactes (fausses) et peuvent induire en erreur les analyses commerciales. D’où l’intérêt de s’assurer de la véracité de ces données. (Danger de ne pas faire confiance à la data qu’on utilise en tant que chef d’entreprise)

**Valeur :** Toutes les données collectées n’ont pas une réelle valeur commerciale et l’utilisation par exemple l’utilisation des valeurs inexactes peut affaiblir les informations fournies par des applications d’analyse. Il est essentiel que les entreprises emploient des pratiques telle que le nettoyage des données et confirment que les données sont liées à des questions commerciales pertinentes avant de les utiliser dans un projet d’analyse.

**Variabilité :** les données volumineuses sont moins cohérentes que les données de transaction classiques et formées de différentes manières d’une source à l’autre.

**Volatilité :**  le recueil des données doit s’adapter aux changements technologiques et réglementaires (RGPD, chatbot, blockchain…) pour rester pertinent.

Certaines personnes attribuent encore plus de V au big data (entre 7 et 10 V).

# : PRINCIPES ET METHODES DU BIG-DATA :

**Définition de la scalabilité :** La scalabilité en général est la capacité à chaque instant T d’un dispositif informatique (ordinateur, application, base de données, infrastructures …) de s’adapter à la demande. C’est-à-dire principalement d’augmenter sa capacité quand le besoin augmente et diminue sa capacité quand le besoin est en baisse.

Il existe deux types de scalabilité :

1. **Scalabilité horizontale** ( **scale out** évolutivité horizontale) : consiste à rajouter des composants matériels des composants matériels (machines) à des machines déjà existantes pour satisfaire le besoin par exemple de s’équiper de plus de serveurs au lieu d’un seul. Ça coute moins cher et la qualité moins bonne.
2. **Scalabilité verticale : scale up** elle consiste à ajouter des composants à une machine déjà existante pour satisfaire le besoin. Par exemple de rajouter de la mémoire au serveur ou acheter un serveur de plus grande mémoire pour satisfaire le besoin ça coute plus cher et la qualité meilleure.

Comme en Big data on a la problématique de volumétrie de données (données de volumétrie variées) donc la scalabilité de ses outils doit être à la hauteur des défis de l’analyse de ces données.

On prend un exemple : de google : en novembre 2016 google a annoncé que leur moteur de recherche a référencé plus 130 milliards sites web. Et a répondu à plus 5.48 milliards de requêtes par jour. Si on reste toujours sur les mêmes dispositifs et serveurs on risque de ne pas répondre aux besoins à partir d’un instant T surtout qu’on sait les données (les sites web) se multiplient d’un jour à l’autre. D’où la place primordiale de la notion de scalabilité.

De la même façon les traitements se multiplient d’une seconde à l’autre à chaque requête ce qui fait que la scalabilité doit être horizontale.

Aussi la scalabilité il faut qu’elle assure en même temps que les machines quand elles tombent en panne (au cours d’un traitement) la donnée n’est pas perdue et le calcul n’est pas perdu.

C’est deux paramètres font naître ce qu’on appelle les deux principes concepts de la big data :

Le calcul distribué et le stockage distribué.

1 Nœud = 1 machine 🡺 l’ensemble des machines connectées sur lesquelles on installe les services de la plateforme hadoop forment un cluster.

Le big data se base sur une plateforme hadoop qui se charge de réaliser le stockage distribué et le calcul distribué sur plusieurs machines (scalabilité horizontale) dans le but de s’adapter aux volumes de données de calculs et leur intégrité.

On dit stockage distribué car un seul fichier est subdivisé en plusieurs blocs, et chaque bloc est stocké dans une machine différente. A chaque fois qu’une machine reçoit un bloc le réplique sur plusieurs autres machines. Comme elle peut recevoir d’autres répliques des autres blocs d’un fichier.

Dans la plateforme hadoop il existe 4 services principaux :

* **HDFS (Hadoop Distributed File System)** : Le système (service) qui gère le stockage contenant des données très volumineuses et de haute performance. Tolérant aux pannes grâce au découpage (grâce aux réplications sur des nœuds). Grace à des capacités de stockage importantes d’une façon distribué le HDFS est devenu le cœur du big data. **HDFS remédie aux problèmes des précédents systèmes de gestion de données qui ne pouvaient prendre en charge les flux de données et les analyser en temps réel.** Comme il est impossible d’accéder à ces fichiers et à ces données comme dans un datawarehouse donc le seul moyen de retourner une valeur sur une donnée est le map reduce
* **MAP-REDUCE :** (Framework) est un système de calculs distribués. Le but de map reduce est de réaliser des calculs sur les différents nœuds. (sa méthode c’est répondre à la question qu’est ce que je fais quand je rencontre une ligne)
* **YARN :** s’occupe de la gestion des ressources du cluster. (des différents nœuds), quand on lance un job sur une machine si la machine n’pas assez de ressources ou tombe en panne ( CPU, RAM, disque dur) yarn décide de déplacer le job sur une autre machine plus adaptée. C’est lui qui alloue les ressources du cluster.
* **ZOO KEEPER :** permet de coordonner le travail des trois services (hdfs, mapreduce et yarn)

# : LES OUTILS DU BIG DATA :

En outre de la plateforme HADOOP la liste des outils du big data est non exhaustive est en constance expansion dans le monde du big data, on peut citer quelques outils :

En plus des bases de données traditionnelles (relationnelles) le big data utilise aussi les bases de données non relationnelles donc NoSql (Not Only Sql), dans ces bases de données on en trouve :

* Rapidminer : pour les données de type vidéo
* Mongo db : Elle est très scalable, NoSql. Langages de programmation comme JavaScript, Ruby et Python.
* Open refine : pour les bases de données désordonnées, permet de nettoyer rapidement les données et de les transformer en données exploitables.
* Hbase, Cassandra : Est une base de données nosql, hautement scalable.
* Sqoop : Permet de relancer les jobs différentiels au moment de migrer d’une base de données relationnel vers un datalake. Différentiel = la différence de temps au lieu que le job dépose les nouvelles données dans un data warehouse il les déposera dans le nouveau datawarehouse (datalake)
* Kafka, spark
* nifistreameset

# : CONCLUSION :

Le big data permet, au jour d’aujourd’hui grâce à ses avancées et ses résultats très optimisés tant au niveau économique que social et sociétal, de faire une liste indéfnie et indéterminée de services et d’avancées et les recherches sont toujours en cours :

* Commercial : Analyse complète (360°) du consommateur. (Comportement et ses exigences) ce qui permet aux entreprises d’optimiser leurs compagnes de marketing par exemple et les anticiper (quelqu’un qui navigue sur des page web de téléphonie on lui propose des pub sur les navigateurs de différentes page web qui vendent les mêmes produits) .
* Militaire : Très utilisé, on peut citer l’exemple de Londres pendant JO 2016 pour l’antiterrorisme

Les prouesses du big data sont le résultat de plusieurs facteurs, on peut en citer quelques un :

* Scalabilité horizontale : En big data les machines sont toutes de même valeur et elles exécutent toutes les mêmes tâches, la dépendance aux ressources est quasi nulle.
* Tolérance aux pannes : grâce au stockage et au traitement distribué le big data est considéré comme une révolution car même s’il des machines qui tombent en panne cela n’influencera pas l’exactitude des résultats de calculs ou de stockage.
* Facilité d’usage pour les développeurs : deux à 4 lignes de codes suffiront pour atteindre des résultats souhaités.
* Capacité de traiter des données très variées et très variables : la base de traitement dans le big data est le fichier ce qui permet de traiter tous les phénomènes existants.
* Schéma à la lecture : si par exemple on a une table composée de 40 colonnes on peut s’intéresser qu’à deux colonnes de la table selon le besoin.
* Pas de mise à jour : dans l’ancien système (base de données relationnelles) le plus chiant à faire est les mises à jour car cela prend beaucoup de temps alors qu’en big data cette action est totalement absente . (un fichier hdfs non soumis à des mises à jours)