**Big Data**

**Qu’est-ce que c’est ?**

Vous avez probablement entendu le terme Big Data - c'est l'un des termes les plus en vogue actuellement. Mais qu'est-ce que le Big Data exactement ?

Le Big Data est un ensemble de données très vaste, faiblement structuré, qui défie les méthodes de stockage traditionnelles.

**Les données générées**

Les données générées par l'homme sont des courriels, des documents, des photos et des tweets. Nous générons ces données plus rapidement que jamais. Imaginez simplement le nombre de vidéos téléchargées sur You Tube et de tweets qui tourbillonnent. Ces données peuvent aussi être des Big Data.

Les données générées par des machines constituent une nouvelle catégorie de données. Cette catégorie comprend les données de capteurs et les journaux générés par des "machines", tels que les journaux de courrier électronique, les journaux de flux de clics, etc. Les données générées par les machines sont beaucoup plus importantes que les données générées par l'homme.

Avant l'arrivée d'Hadoop, les données générées par les machines étaient généralement ignorées et non capturées. C'est parce que le traitement de ce volume n'était PAS possible, ou PAS rentable.

**Provenance des données ?**

À l'origine, le big data était constitué des données du web, c'est-à-dire de l'ensemble de l'Internet ! Rappelez-vous qu'Hadoop a été construit pour indexer le web. De nos jours, les Big data proviennent de sources multiples.

- Les données du Web - ce sont toujours des données volumineuses

- Données des médias sociaux : Les sites comme Facebook, Twitter, LinkedIn génèrent une grande quantité de données.

- Données de flux de clics : lorsque les utilisateurs naviguent sur un site Web, les clics sont enregistrés pour une analyse plus approfondie (comme les modèles de navigation). Les données relatives au flux de clics sont importantes pour la publicité en ligne et le commerce électronique.

- Données des capteurs : les capteurs intégrés dans les routes pour surveiller le trafic et d'autres applications génèrent un grand volume de données.

- Appareils connectés : Les téléphones intelligents en sont un excellent exemple. Par exemple, lorsque vous utilisez une application de navigation comme Google Maps ou Waze, votre téléphone renvoie des pings indiquant sa position et sa vitesse (ces informations sont utilisées pour calculer les points chauds du trafic). Imaginez des centaines de millions (voire des milliards) d'appareils consommant et générant des données.

**Exemple**

De combien de données parlons-nous ?

- Facebook : 40 PB de données et capture 100 TB / jour

- Yahoo : 60 PB de données

- Twitter : 8 To / jour

- EBay : 40 PB de données, capture 50 TB / jour

**Enjeux du Big Data**

*L'ampleur du Big Data*

Le Big Data est... eh bien... de grande taille ! La quantité de données qui constitue le Big Data n'est pas très claire. Ne nous laissons donc pas entraîner dans ce débat. Pour une petite entreprise qui a l'habitude de traiter des données en gigaoctets, 10 To de données, c'est GROS. En revanche, pour des entreprises comme Facebook et Yahoo, des pétaoctets, c'est énorme.

Rien que la taille des big data rend impossible (ou du moins prohibitif) leur stockage dans des systèmes traditionnels tels que les bases de données ou les classeurs conventionnels.

Nous parlons du coût du stockage de gigaoctets de données. L'utilisation de filers de stockage traditionnels peut coûter très cher pour stocker des Big Data.

*Les Big Data sont non structurées ou semi-structurées*

Une grande partie du Big Data n'est pas structurée. Par exemple, les données du journal du flux de clics peuvent ressembler à l'horodatage, à l'identifiant de l'utilisateur, à la page et à la page du référent.

Le manque de structure fait que les bases de données relationnelles ne sont pas bien adaptées au stockage des Big Data.

De plus, peu de bases de données sont capables de stocker des milliards de lignes de données.

*Il ne sert à rien de se contenter de stocker des données volumineuses si l'on ne peut pas les traiter.*

Le stockage des Big Data fait partie du jeu. Nous devons les traiter pour en extraire de l'intelligence. Les systèmes de stockage traditionnels sont plutôt "muets", c'est-à-dire qu'ils ne font que stocker des bits - ils n'offrent aucune puissance de traitement.

Dans le modèle traditionnel de traitement des données, les données sont stockées dans un "cluster de stockage", qui est copié sur un "cluster de calcul" pour être traité, et les résultats sont réécrits dans le cluster de stockage.

Ce modèle n'est toutefois pas adapté au Big Data, car copier autant de données sur un cluster de calcul peut prendre trop de temps ou être impossible. Quelle est donc la solution ?

Une solution consiste à traiter les Big Data "sur place", c'est-à-dire dans un cluster de stockage faisant office de cluster de calcul.

**Comme nous l'avons vu ci-dessus, le Big Data défie le stockage traditionnel. Alors comment gérer le Big Data ?**

Les bases de données relationnelles (SQL) atteignent actuellement leurs limites de performances et de stockage mais un nouveau modèle, le NoSQL (Not Only SQL) permet désormais de franchir ces limites.  
Ces bases sont beaucoup moins couteuses car elles peuvent fonctionner sur du matériel peu cher. Elles s’adaptent également facilement.

Cela vient du fait qu’il n’y a pas de limites sur le nombre de disques dur de la base de données, et que le système voit le système de stockage comme une seule entité.

Les disques sont donc faciles à ajouter ou à enlever, et n’importe quel disque peut être ajouté, notamment des disques peu couteux.

**Vers des outils du Big Data -> Hadoop**

**Hadoop est l’outil le plus populaire.**

Apache Hadoop est l'outil Big Data le plus populaire et le plus puissant, Hadoop fournit la couche de stockage la plus fiable au monde