**Notion et outils du big data : Qu’est ce que le big data ?**

   La notion Big Data ou mégadonnées en français, est apparue pour la première fois dans les années 60, elle est utilisée pour décrire une collection de données complexes et de très grande taille, qui croît de façon exponentielle avec le temps. Ces données sont tellement volumineuses que les outils traditionnels de gestion des données ne pourront pas les stocker ou de les traiter. Ces données peuvent provenir par exemple des données utilisateurs, les historique des recherches, les études de marché… La gestion du big data nécessite des outils particuliers que nous allons découvrir dans la suite de la présentation. Le principe du Big data se décrit par cinq dimensions essentielles appelées les 5V et qui sont : le volume, la vitesse, la variété, la véracité, la valeur et la variabilité.

   Les résultats des analyses des « Big Data » peuvent être utilisés dans plusieurs domaines. Par exemple, dans le domaine du marketing, grâce aux données collectées, les entreprises peuvent étudier les comportements de consommation, identifier leurs besoins, et mesurer leur satisfaction par rapport à leurs produits et services. C’est ainsi que les entreprises utilisent les données collectées pour ajuster leur stratégie de marché et amélioration l’expérience client.

   Le Big data est être utilisé dans d’autres domaines d’applications comme  la recherche scientifique, le développement de produit et service, l’expérience client, le commerce, le machine learning, les prévision et la prise de décision, la gestion de risque ou encore la fraude et la conformité.

   Bien que le coût de la collecte, le stockage et l’étude des données restent onéreux et compliqué pour plusieurs entreprises, le “ big data “ est une révolution technologique qui permet aux entreprises d’avoir un regard sur l’ensemble de leur activité, d’établir des prévisions et d’améliorer la prise de décisions pour affiner la connaissance du risque.

Les outils du Big Data

Il existe de nombreux outils du Big Data existant, je vais donc vous en présenter quelques un, les principaux et vous expliquer leur utilité, leur force, faiblesse…

Bases No SQL :

* Mongo DB
* Apache HBase
* Redis
* Cassandra

Systèmes distribués :

* HortonWorks
* Cloudera
* Hadoop

Outils d’administration :

Outils de traitement des données :

* Spark

Outils d’accès aux données :

* Hive
* Hue
* Impala

**Outils du Big Data :**

   Pour convertir ces données massives et les rendre exploitables de nouveaux outils ont vu le jour. La gestion des « Big Data » requièrent donc des solutions plus avancées, vu le nombre considérable de données et leur complexité croissante. Je vais vous présentez les outils analytiques du Big data les plus utilisés.

* **Hadoop :** développé par Apache, Hadoop est la principale plateforme du Big Data. Il est utilisé pour le stockage et le traitement d’immenses volumes de données. C’est un framework logiciel open source. Ses différents composants sont utilisés par de nombreuses entreprises pour leurs projets Big Data. Hadoop repose sur des modèles de programmation simples pour assurer le traitement des données et les rendre disponibles sur des machines locales.
* **Hive :** représente le Data Warehouse de Apache Hadoop, c’est un logiciel qui a été développé par Facebook. Hive permet d’effectuer facilement et rapidement des requêtes. Ce logiciel permet d’extraire les données en provenance de Apache Hadoop. A la différence de Hadoop, Hive permet d’effectuer des requêtes SQL sans avoir besoin d’écrire en Java.
* **Cassandra :** Fait partie du projet Apache, Cassandra représente la base de données NoSQL qui est la plus utilisée pour le stockage et le traitement des larges volumes de données. C’est un système de base de données qui permet de faciliter le stockage et la gestion de données structurées sur différents serveurs à haute vélocité.
* **Spark :** C’est un moteur de traitement de données rapide qui fait aussi partie du projet Apache. Spark permet de traiter de larges volumes de données de manière distribuée (cluster computing). C’est un moteur de traitement parallèle de données open source permettant d’effectuer de très grande analyse à l’aide de machines en clusters.

Mongo DB :

**MongoDB est une base de données NoSQL orientée documents.**

Les données prennent la forme de documents enregistrés eux-mêmes dans des collections, une collection contenant un nombre quelconque de documents.

- Collections <-> tables (modèle relationnel)

- Documents <-> enregistrements (modèle relationnel)

MongoDB permet de manipuler des objets structurés au format BSON (JSON binaire).

MongoDB ne nécessite pas de schéma prédéfini des données. Contrairement aux BDDR, les champs d'un enregistrement sont libres et peuvent être différents d'un enregistrement à un autre au sein d'une même collection. Le seul champ commun et obligatoire est le champ de clé principale ("id").

* Format :

|  |  |
| --- | --- |
| **BSON = JSON binaire**  **JSON = JavaScript Object Notation**  **Tout est clé/valeur : “clé” : “valeur”**  **Champ “\_id“ obligatoire**  **On peut ajouter des champs à la volée**  **On peut imbriquer des documents, mettre des listes, …** | **C:\Users\u174319\Documents\database\Doc MongoDB - DEP - Confluence_fichiers\image2021-4-12_14-56-38.png** |

* NOSQL :

Le NoSQL (Not Only SQL) regroupe de nombreuses BDD, récentes pour la plupart, qui se caractérisent par une logique de représentation de données non relationnelle et qui n'offrent donc pas une interface de requêtes en SQL.

C’est une approche de la conception des bases et de leur administration particulièrement utile pour de très grands ensembles de données distribuées.

Enjeu : résoudre les problèmes de performances en matière d'évolutivité et de Big Data que les BDDR ne sont pas conçues pour affronter.

MongoDB ne permet ni les requêtes très complexes standardisées, ni les jointures, mais permet de programmer des requêtes spécifiques en JavaScript (langage natif).

Il est également possible de manipuler les données en C, C++, Java, .NET, Perl, PHP, Python, ...

* Haute disponibilité :

MongoDB utilise une architecture basée sur le principe « maître/esclave ».

Le serveur primaire reçoit toutes les requêtes (lecture/écriture), et s’occupe de gérer la cohérence des données. Les serveurs secondaires appliquent également les instructions reçues par le serveur primaire pour avoir les mêmes données.

Il faut au minimum 3 serveurs (1 primaire et 2 secondaires) => **ReplicaSet**

Il est possible de modifier la gestion du ReplicaSet pour router les requêtes de lectures vers les secondaires.

Pour gérer les pannes, on peut ajouter un arbitre qui vérifie l’état du réseau. Si le primaire tombe, un système de vote va permettre d’élire le secondaire qui remplacera le serveur tombé en panne. L’arbitre est toujours sur un serveur à part.

Il est également possible de répartir les données sur plusieurs serveurs => **Sharding**.

|  |  |
| --- | --- |
| MONGODB | POSTGRESQL |
| * Stocker les données et les lire * Cycle de vie très simple * Orienté développeur | * Transactionnel (commit/no commit) * Modèle relationnel (MCD) * Sauvegarde à chaud * Plus robuste |

Sources :

<https://dai.ly/x27wtco>

<https://www.octo.com/publications/19-hadoop-feuille-de-route/>

<https://mbamci.com/notion-du-big-data-outils-de-gestion/>