Big Data – NoSQL

BIG DATA : C'est 3 composantes :

Calcul Scientifique (Modélisation, Langage C, Python)

Analyse de Données (statistiques : R langage)

BD : **NoSQL**

2015 : 2/3 des entreprises du numérique lançaient des projets Big Data :

-exploiter l'information,

- stocker la data,

- gestion temps réel

2017 : début de l'humanité => volume de data

Il est présent dans beaucoup de domaines :

- Stockage des données : Les réseaux sociaux, Web achat en ligne
- Temps réel : domotique, jeux online, lignes ferroviaires, Geo localisation => outils de stockage et de manipulation spécifique.

Le développement d'outils suscite un grand intérêt dans l'industrie => masse de données, stockage

Volume : Téraoctets ; Pétaoctets.

Concept des 5 V :

- Notion de Volume important de données : pétaoctets
- Notion de Variété des informations : sources de données multiples, hétérogènes, peu ou pas de structurées (MongoDB)
- Notion de Vélocité : traitement / fréquence de création et/ou de collecte => (outils de reporting) prise de décision
- Notion de Variabilité : s'adapter au format changeant des données qui doivent être gérées dans le SI (cela impose aux entreprises d'avoir un SI)
- Notion de Véracité : vérifier et intégrer les données dans un processus de traitement des données

Remarque : les objectifs pour les entreprises sont de disposer d'outils permettant de traiter, de gérer au mieux les masses de données dans leur SI.

Cela implique la mise en œuvre :

(1) De systèmes de stockage de masse,

(2) La capture de l'information à grande vitesse, en temps réel

(3) L'analyse de la data.

Les base de données NoSQL offrent des solutions de stockage.

NoSQL : Le terme NoSQL pour « Not Only SQL », regroupe des solution récentes (remarque : il existe différents types de solutions en NoSQL alors qu'en relationnel MySQL, Oracle permettent de faire pratiquement la même chose). NoSQL se différencie de la représentation relationnelle. C'est une logique de représentation des données différentes du SQL => performance en termes de temps de réponse, capacité à traiter de très grands volumes de données, stockage de données dont la structure varie.

Pour comparer avec le SQL et le Relationnel : BD cohérente, structurée, organisée…On fonctionne avec MySQL , Oracle, PG qui ont tous le même mode de fonctionnement…….

En NoSQL on va s'affranchir des contraintes ACID (Atomicité, Cohérence, Isolation, Durabilité), on va représenter les données via une architecture technique ou il suffit d'ajouter des capacités de stockage et de calcul pour gagner en performance (scalabilité et élasticité).

En relationnel nous modélisations une problématique via un MCD, un MLD…

Il existe 5 représentations de base de données NoSQL avec les 5 V du Big Data

Clé valeur : représentation simpliste (très utile dans la gestion de caches ou pour fournir un accès rapide aux informations (Redis)

Clé valeur avec contraintes d'intégrité : on ajoute des contraintes dans la représentation des données (Cassandra)

Document : on associe une valeur complexe au système de clé-valeur (CouchDB (plus ancien), MongoDB (langage d'interrogation de données natif))

Colonne : l'idée c'est de pouvoir disposer de plusieurs valeurs pour les stocker en relation de type one to many (Hbase, Cloudata)

Graphe : modélisation, stockage et manipulation de données complexes. (FlockDB, Neo4j)

Le cycle de vie, la chaine de traitement de la donnée est lié à :

EXTRACTION de données, STOCKAGE de données, AFFICHAGE de données, ANALYSE de données

L'extraction est liée à la récupération des données et/ou dans des bases données : ETL (Talend)

Le stockage de données, c'est constituer des base de données (BD décisionnelle : datawarehouse, datamining, datamart…..)

L’affichage de données que l’on trouve sur des portails : donner l’accès aux informations

L’analyse avec graphiques et tableaux de bord : reporting, datamining…..

Exemple : Facebook c’est 800 téraoctets de données par jour => machines, de la capacité de stockage et de traitement de la donnée.