Bases de Données Documentaires et Distribuées

On distingue les types de données pour ne travailler que sur tout ce qui est lié à la notion de document => c'est une information UNIQUE (notion d'unicité).

Que ce soit un document structuré, non structuré ou semi structuré il faut récupérer une donnée qui est générée en temps réel que ce soit par un utilisateur, un capteur, un jeu en ligne….

On gère les données et on va rechercher des informations dans des documents textuels => collection massive de document (Documents, Collections => MongoDB).

On va maintenir de très grands volumes, on parle de stockage : distribution du stockage (c'est-à-dire la capacité à augmenter le nombre de disques, le nombre de machines (peta sur certaines machines).

Calcul distribué sur les données, on a des méthodes d'analyse sur de très grandes collections de données en adaptant les ressources allouées au système en fonction de la demande : élasticité.

Cycle de représentation de la gestion de données dans des bases NoSQL => on sauvegarde dans le cloud puis on accède à ces données via des moteurs de recherche).

Remarque : tout ceci est différent des bases de données Relationnelles qui sont structurées, régulières, normalisées qui respectent la notion de transaction (en représentation : Commit et Rollback) (en gestion, site sur le web, commerce en ligne, gestion de comptabilité).

En NoSQL on perd ces notions de Relationnel mais on gagne sur le volume de données et sur le traitement de données.

La notion d'information de document signifie que l'on travaille sur un document autonome qui ne dépend d'aucun autre document, il est stocké à un seul endroit et il ne fait référence à aucun autre document.

Exemple : on va stocker beaucoup d'informations dans un même document même si c'est complexe on va le gérer à partir du JSON, du XML

On effectue à la différence du Relationnel (jointure relationnelle) des recherches par similarité.

Techniquement : outils à mettre en œuvre

- Elasticité des systèmes pour accéder à de la volumétrie sans limite.
- Scalabilité pour des accès en temps réel
- Calcul : mapreduce, hadoop
- BD : MongoDB, CouchDB, Cassandra
- Moteur de Recherche : Elasticsearch

- Système distribué : MongoDB, Cassandra, Elasticsearch
- Traitement massif : Hadoop, Flink, Sark

Remarque : il existe bcp d'outils qui sont libres (Docker…..)