BIG DATA

Razvan BIZOI

[razvan@bizoi.fr](mailto:razvan@bizoi.fr)

https://discord.gg/a3hCV5b2

# Introduction

Def dela BDD

Big data === distribuer du traitement entre les machine

Il existe deux typologie de BDD relationnel :

* La BDD transactionnel, qui fait beaucoup d’update et des inserts
* La BDD décisionnel, qui est un duplicata de la BDD de production et qui n’est dédier que pour l’analyse.

Ces deux types ont des fonctionnements complètement différents, elle se structure différents. Du fait des insert et des petites updates, il faut normaliser et éclater les informations sur une multitude de table pour réduire le temps d’insert.

Cependant, une fois que l’informations est éclaté sur une multitude de table, les requête pour retrouver l’information vont devenir très compliquer et parcourront de multiple table *(10-20)*. Un moteur relationnel fait un traitement itératif pour réunir la première table avec la deuxième, la troisième avec la quatrième.

Ainsi, une copie tel quel de la BDD n’est pas possible, il va falloir exécuter un traitement qui sera définie par la méthode en étoile. Il n’y a pas de manière unique pour résoudre ce problème, il faut s’intéresser au métier qui produit la donnée.

BDD SQL et noSQL

Dans les BDD no SQL il y a une grande sécurité sur les accès concurrent du au traitement d’objet lourd comme le JSON. Les Base sont alors « monomaniaque » qui savent très bien travailler un type de données.

# Niveau d’abstraction

## Conceptuel

* Abstraction des aspects techniques
* Analyse des besoins des décideurs

Dans les BDD décisionnel il y a l’idée de hiérarchie est liée au BDD dimensionnel où les information son regrouper par group By. A l’inverse dans le modèle Data Lak il n’y a pas de hiérarchie toutes les informations sont classes de manière horizontale.

### Dimension

C’est une manière de représenter un métier ou l’aspect d’un métier dans une BDD. Il faut tout d’abord identifier la dimension : temporel *(jour / semaine/ mois / années)*, géographique, produits… La dimension se construit dans un dialogue avec le cœur du métier.

Il faut bien séparer les métrique et les produits, le poids du produit s’il est utilisé pour son transport est une métrique.

### Hiérarchie

Les hiérarchies peuvent avoir plusieurs chemins pour arriver à la même donnée. Pour accéder à la donnée du jours, je peux soit passer par l’années, le trimestre le jour ou par l’années, la semaine, le jour.

### Modélisation

Il existe 2 schémas le modèle en étoile et …

Schéma en étoile :

* Le fait au centre et les dimensions autour
* Les dimensions n’ont pas de liens entre elle

Avantage de l’étoile :

* Facilité de navigation
* Nombre de jointure limité

Désavantage de l’étoile :

* Redondance dans les dimensions
* Toutes les dimensions ne sont

Grâce à l’architecture en étoile il est possible de supprimer les relations, puisqu’une application dédier remplit les informations de cette base à partir de la BDD transactionnel. Il est donc pertinent d’utiliser les BDD no-SQL pour le décisionnel. En effet, les BDD relationnel ont un défaut connu depuis les années 80, le temps de réponse augmente de manière linéaire en fonction de la qualité de données jusqu’à 1T, passer se seuille le temps de réponse augmente de manière exponentielle.

ETL = Extract Transforme and Load

Le modèle ETL qui consiste à faire le traitement la nuit pour mettre à jours le BDD décisionnel, est possible lorsque le volume de traitement n’est pas trop important.

### Map reduce

C’est quoi la combinaison map / reduce & en quoi elle est importante pour le big data ?

### Théorème du cap

Ceci n’est pas un théorème mais plutôt un constat.

* Disponibilité + tolérance aux partitions : No SQL *(n’existe pas encore mais des recherches se font)*

## Deux types de pointeur

Fonction et variable sont exactement égale au niveau de la mémoire, ainsi un pointeur peut pointer vers une fonction.

## Kafka

Traitement asynchrone des informations. Permet d’avoir simultanément des traitements synchrones et asynchrone.

## Spark

Spark est un Framework développer uniquement pour UNIX. Le package pyspark comprend tout l’environnement pour utiliser spark, à l’exception de la gestion de disque local.