Rapport de la séance sur le Big Data

Au cours de la séance sur le Big Data, après avoir passé en revue les différentes notions : Big Data, Framework, Plateformes, Intelligence Artificielle, nous avons vu ensemble comment traiter des données en utilisant Snowflake et DBT (Data Build Tools).

Pami les notions abordées aujourd'hui, on peut retenir que :

- Hadoop est un framework de stockag et de calcul : le stockage est fait avec HDFS et l'agrégation avec Map Reduce.
- L'unité de stockage par défaut de Hadoop est de 128Mb
- Map Reduce est un framework de calcul distribué avec des unités de calculs sont très importants
- Toute transformation implique un mappeur et toute agrégation un reduceur
- Il existe une fonction qui permet qui permet de faire l'agrégation par clé : on parle de Reduce by Key
- Spark est un framework de développement qui est plus rapide que Hadoop en termes de temps de temps de traitement.

Afin d'appliquer toutes les notions abordées, nous avons réalisé un TP dont l'objectif était d'opérer des transformations sur une base de données à l'aide de DBT et de SQL.

Après avoir chargé notre jeu de donnée sur Snowflake, nous nous sommes connectés depuis un terminal pour opérer les transformations sur la base de données.

Après avoir créé différents fichiers sql dans 3 différents dossiers :

- Dossier : models/src
    - src_hosts.sql
    - src_listings.sql
    - src_reviews.sql
- Dossier : models/dim
    - dim_hosts.sql
    - dim_listings.sql
- Dossier models/fct
    - fct_reviews.sql

Nous avons fait usage des commandes ci-après :

- dbt init : pour nous initier le projet sur notre machine locale
- dbt debug : pour nous assurer que nous arrivions bien à nous connecter à la base de données distante et qu'il n'y avait pas d'erreur
- dbt run : pour exécuter le projet afin que les fichiers SQL précédemment crés ont bien été chargés sans erreur
- dbt test : pour exécuter le fichier schema.yml afin de tester la bonne exécution de notre projet et la prise en compte de nos transformations.
- dbt seed : pour synchroniser les données en local avec les données sur le serveur.
- dbt docs generate : pour générer une documentation du projet
- dbt docs serve : pour afficher la documentation