Question 5 :

La première étape pour construire notre système est de construire la base du système qui est l'extraction des données :

- Pour cette étape, nous allons utiliser apache kafka et apache spark et combiner les deux pour le streaming et l'extraction des données à partir de la source en utilisant l'API par exemple.
- Nous avons utilisé ces deux solutions parce que :
    - Kafka est excellent pour l'ingestion durable et évolutive de flux d'événements provenant de nombreux producteurs vers de nombreux consommateurs.
    - D'autre part, Spark est excellent pour le traitement de grandes quantités de données, y compris les flux d'événements en temps réel et en temps quasi réel.

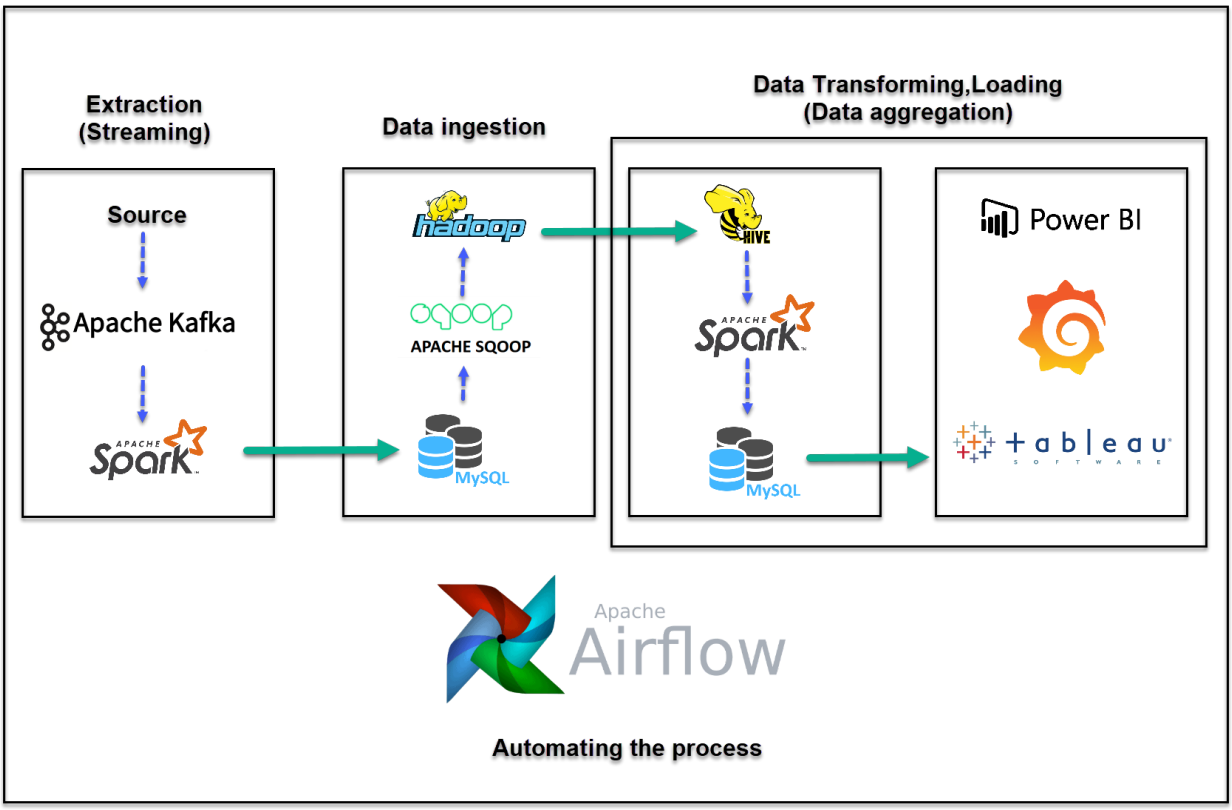
La deuxième étape concerne l'ingestion de données :

- C'est-à-dire que nous allons utiliser les données qui proviennent de spark dans l'étape précédente et les charger dans une base de données.
- Nous utilisons sqoop car c'est un outil conçu pour transférer des données entre Hadoop et des serveurs de bases de données relationnelles. Il est utilisé pour importer des données depuis des bases de données relationnelles telles que MySQL.
- Puis importer les données dans hadoop

la troisième étape est l'agrégation des données :

- La première étape consiste à télécharger les données dans Hive et à les manipuler depuis Hive Hive est une base de données open-source d'entreposage de données distribuées qui fonctionne sur le système de fichiers distribués Hadoop. Hive a été conçu pour l'interrogation et l'analyse des données volumineuses. Les données sont stockées sous forme de tables (tout comme les SGBDR).
- Chargez ensuite les données provenant de Hive dans spark pour effectuer différents types d'analyse, car spark peut fonctionner jusqu'à 100 fois plus vite en termes de mémoire et 10 fois plus vite en termes de vitesse de calcul sur disque que Hadoop.
- Charger ensuite les données analysées dans une base de données agrégée.
- On utilise les données agrégées et les présenter sous forme de graphiques à l'aide de différents tableaux de bord (Power BI, Grafana, Tableau).

La dernière étape consiste à utiliser Apache Airflow pour automatiser l'ensemble du processus.


Extraction
(Streaming)
Data ingestion
Data Transforming,Loading
(Data aggregation)
Source
Apache Kafka
Spark
hadoop
APACHE SQOOP
MySQL
HIVE
Spark
MySQL
Power BI
tableau
Apache
Airflow
Automating the process