Année universitaire : 2022-2023

Mini projet
Proof-of-concept for Kafka + Spark streaming from scratch
(Suite de Projet N5)

Mini Projet :

Twitter peut être considéré comme l'une des sources de données textuelles les plus importantes. En effet, il offre des informations massives en temps réel sur les opinions de ses utilisateurs et offre la possibilité d'effectuer des transformations sur ses données. L'objectif de ce projet est d'utiliser Twitter pour effectuer les tâches suivantes :

1. Collecter un flux de tweets à partir de Twitter
2. Extraire les tweets du cluster Kafka
3. Calculer le nombre de caractères et le nombre de mots pour chaque tweet
4. Enregistrer ces données dans une table Hive

Concernant la première tâche (la collecte d'un flux de tweets à partir de Twitter) et étant donné que l'API Twitter n'est pas gratuite, nous allons plutôt contourner ce problème de non-gratuité de l'API à travers une simulation des tweets en utilisant des fake Tweet. Le code python ***fake_tweet_stream.py***, disponible dans le dossier partagé, permet de générer des fake tweets que vous pouvez utiliser.

Les étapes à suivre pour la réalisation de ce projet :

1. **VM Setup**
   Pour commencer la réalisation de ce projet je vous recommande de configurer une machine virtuelle sur votre ordinateur. On va utiliser une machine virtuelle Lubuntu ubuntu-22.04.1 sur VirtualBox et lui ai donné 4 Go de mémoire et 15 Go de stockage, mais n'hésitez pas à utiliser votre distribution Linux préférée.
2. **Configuration de l'apache Kafka : Créer un flux de tweets qui sera envoyé à une file d'attente Kafka.**
3. **Configuration de Hadoop**
4. **Configuration de Hive**
5. **Configuration de Spark**
6. **Exécution de code**

Rendu :

- Un rapport comprenant des capteurs d'écran des étapes les plus importantes avec quelques explications.
- Une vidéo démonstrative individuelle qui présente les configurations de chaque étape avec les explications nécessaire.

Eléments nécessaire et des références pour la réalisation de ce mini projet :
https://drive.google.com/drive/folders/1Pe7-9DvHllQ9f2ggHYoy4F2Y64Nm5E8J?usp=sharing