RAPPORT PROJET INFRASTRUCTURE BIG DATA

Par les élèves :

- Huynh Ngoc Thai Duy
- MENGONG MEYANGA Simon Pierre

Table des matières

1. Description du projet

Partant du framework de base qui nous a été donné, nous avons construit dans un cluster de containers une chaine de traitement de données qui intègre les technologies : Kafka, Spark, Influx DB et Grafana.

Comme exemple d'application, nous avons utilisé les données de trafic routier provenant de plusieurs capteurs dans la ville de Belo Horizonte au Brésil. C'est un vaste dataset contenant des mesures des flux de trafic dans plusieurs endroits de la ville. Chaque capteur détecte périodiquement le type de véhicule circulant dans cette localisation sa vitesse et d'autres informations.

Dans ce scénario, Apache Kafka va être la couche intermédiaire entre les capteurs et l'application qui va consommer ces données.

2. Technologies

2.1. Apache Kafka

Nous avons dû utiliser la version 3.3.1 de Kafka pour des raisons de compatibilité avec Spark. Il suffisait alors de télécharger cette version là et de modifier le fichier **start-containers.sh**. Notre application compte le nombre de véhicules en circulation (peu importe l'endroit) toutes les 5 minutes. Etant donné que Spark lit les données sous forme de Batch pour éviter des complications au niveau du calcul et des regroupements, un petit pré-traitement a été fait au niveau de Kafka dès la réception des données.

2.2. Apache Spark

Nous avons utilisé la version 3.3.4 de Spark pour les mêmes raisons. Ici la difficulté a été de l'intégrer à Influx DB. Nous avons dû écrire une classe personnalisée pour pouvoir écrire les données dans la BD Influx.

2.3. Influx DB

Nous avons choisi d'utiliser Influx DB car c'est le SGBD le mieux adapté pour des séries temporelles pour le stockage en temps réel, de plus son intégration à Grafana est plus simple que les autres SGBD. Influx DB offre un container déjà personnalisé avec tout le nécessaire donc tout ce qu'il y a à faire c'est de modifier le fichier start-containers.sh pour l'ajouter. Bien sûr il faudra ajouter un dossier à la racine pour stocker les volumes de données.

2.4. Grafana

Pour l'affichage des données nous utilisons Grafana. Il faut juste modifier le fichier start-container.sh pour ajouter le container Grafana ensuite le connecter à Influx DB et il fera les requêtes tout seul.

3. Références

- A Fast Look at Spark Structured Streaming + Kafka. [Https://towardsdatascience.com/a-fast-look-at-spark-structured-streaming-kafka-f0ff64107325]
- Les données utilisées. [https://dados.gov.br/dataset/contagens-volumetricas-de-radares]