INF 438 BASE DE DONNEES AVANCEES
DEVOIR 2

Analyse Big Data avec le Jeu de Données CitiBike (Traitement Batch)

Objectif

L'objectif est d'utiliser Hadoop (déployé sur Docker) pour analyser le jeu de données CitiBike et développer plusieurs programmes MapReduce en Python via Hadoop Streaming.

Préparation du jeu de données

1. Téléchargez le jeu de données CitiBike depuis Kaggle : https://www.kaggle.com/datasets/sujan97/citibike-system-data
2. Nommer le fichier comme `citibike-data.csv`
3. Chargez le jeu de données dans le conteneur Docker
    a. Copiez le fichier de données dans le conteneur NameNode.
    ```
    docker cp citibike-data.csv namenode:/tmp/
    ```
    b. Connectez-vous au conteneur NameNode.
    ```
    docker exec -it namenode bash
    ```
    c. Créer un dossier dans HDFS
    ```
    hdfs dfs -mkdir -p /user/root/citibike/input
    ```
    d. Téléchargez le fichier de données vers HDFS.
    ```
    hdfs dfs -put /tmp/citibike-data.csv
    /user/root/citibike/input/
    ```
    e. Vérifiez que le fichier a bien été téléchargé.
    ```
    hdfs dfs -ls /user/root/citibike/input/
    hdfs dfs -cat /user/root/citibike/input/citibike-data.csv |
    head -n 5
    ```

Travaux à réaliser

Analyse 1 : Stations de départ les plus utilisées

Écrire un mapper et un reducer permettant de compter le nombre total de départs depuis chaque station. Produire le top 10.

Analyse 2 : Types d'utilisateurs

Comparer le nombre total de trajets réalisés par les types d'utilisateurs "Customer" et "Subscriber" et calculer la durée moyenne.

Analyse 3 : Analyse horaire

Déterminer les heures les plus actives de la journée en comptant le nombre de trajets par heure.

Livrables

Chaque étudiant.e doit rendre :

1. Les scripts mapper/reducer (sur un fichier *.ipynb)
2. Les résultats (captures ou fichiers texte)
3. Une courte interprétation analytique des résultats