Travail Dirigé 3 - Spark Streaming

Nicolas Martin - Spark Streaming
Date de rendu : Vendredi 23 mai 23h59

Datasets :

- https://www.data.gouv.fr/fr/datasets/ficher-des-prenoms-de-1900-a-2018/ à télécharger et mettre dans src/resources/data/insee

Partie 1 - Prénoms

Lancer le websocket :

- Installer python3
- Lancer push_first_name.py
- Lancer read_socket.py pour vérifier que le soket envoie bien les données

Vous devrez lancer push_first_name.py avant chaque lancement de votre code Spark

Questions initiales
Ouvrez le fichier des des prénoms.

1. Combien de colonnes contient-il ? A quoi correspondent-elles ? Quels sont les types des données ?
2. Qualifiez les données. Il y a-t-il des données à éviter ?

Vous allez écrire votre code dans le fichier Streaming.scala. Vous me rendrez ce code avec les réponses aux questions.

Questions :

Découvrir le DataStream

1. Instanciez une SparkSession
2. Créez un DataStream qui collecte la donnée à partir du Socket
3. Ecrivez votre job, qui ne fait pour l'instant qu'écrire les données du DataStream dans la console (Utilisez les différents modes)
4. Lancez le script socket, puis le job Spark.
5. Attendez la fin du job, avec un timeout de 3 minutes. Décrivez le résultat.

Formater le DataStream

1. Créez une nouvelle colonne où les données sont splittées (Vous devriez avoir un Array)
2. Transformez votre DataStream pour obtenir une structure utilisable (Colonnes pour chaque donnée)

Utiliser le DataStream

1. Comptez le nombre de naissances par Sexe
2. Quel sont les prénoms les plus donnés ?
3. Quelles sont les années ayant le plus de naissance ?
4. Quel autre donnée pourrait-on sortir de ce stream ?

Partie 2 - IoT

Le fichier python push_iot_data.py simule un thermostat intelligent, ce thermostat envoie toutes les secondes une donnée dans un socket (localhost:9999).

Questions préliminaires

- Lancez le générateur et lisez ce qu'il envoie. Quel est le format ? Qualifiez la donnée.

Vous écrirez votre code dans le fichier StreamingIot.scala. Vous me rendrez ce code avec les réponses aux questions.

Questions

1. Donnez la température moyenne de la pièce toutes les minutes
2. Donnez la température moyenne sur 1 minute de la pièce toutes les 30 secondes
3. Quel temps avez vous utilisé pour créer vos fenêtres ?
4. Proposez d'autres calculs sur d'autres types de fenêtres.