INGC 2 - InDIA

2025

Plateformes Big Data

Analyse de données de trajets vélo avec Apache Spark

NB: Rendre le Jupyter Notebook correspondant à la partie B du projet.

Installation de Spark et PySpark en local sous Windows

Pré-requis: Python et Jupyter NoteBook pré-installés

Suivez le tutoriel suivant pour installer Spark et PySpark sous Windows:
https://nonlineardata.com/install-spark-pyspark-on-windows

Description du jeu de données

Les données consistent en des détails des trajets des utilisateurs du service de vélos **Citibike** de New York City https://citibikenyc.com/homepage.
La description du jeu de donnés est fournie par l'url : https://citibikenyc.com/system-data.

Partie A :Téléchargement des fichiers de données

1. Téléchargez à partir du navigateur ou avec wget les **données de 2016**.
2. Décompressez l'archive.
3. Explorez les données pour comprendre la structure des fichiers et l'organisation des sous répertoires..

Partie B : Création d'un DataFrame et exécution de requêtes analytiques

Lancez un Jupyter NoteBook puis à partir du package **pyspark.sql** créez une SparkSession, puis répondez aux questions suivantes en utilisant soit **la méthode sql** de SparkSession, soit l'**API DataFrame**.

1. Créer un DataFrame **tripdataDF** à partir des fichiers csv téléchargés.
2. Opérez à partir d'un pré-traitement, au nettoyage des données et à la standardisation du format des données de type datetime.
3. Donnez pour chaque mois le nombre de trajets effectués. Tracez le diagramme en bâtons correspondant.
4. Calculez pour chaque jour (1 à 365) le nombre de trajets effectués. Représentez la courbe correspondante.
5. Donnez pour chaque station de départ et pour chaque mois le nombre d'utilisateurs.
6. Donnez le taux de fréquentation pour chaque station de départ. Représentez le résultat obtenu à l'aide d'un diagramme de votre choix.
7. Calculez et affichez pour chaque trajet identifié sa durée moyenne, minimale et maximale.