Découverte et mise en œuvre de Spark

A – Traitement de données GoBike avec Spark/Databricks

- Aller sur le site de lyft GoBike https://www.lyft.com/bikes/bay-wheels/system-data
  Examiner le schéma des données.
- Télécharger un jeu de données disponibles sur le site https://s3.amazonaws.com/fordgobike-data/index.html
- Charger les données
- Pour les types de données : utiliser les types des données importées SAUF pour l'attribut "Duration" (changer Tinyint en Int si possible).
  Les noms des colonnes doivent être `duration_sec`, `start_time`, `end_time`, ……

B - Travail à faire

- Les données sont prêtes à l'utilisation. Il s'agit de les exploiter dans un objectif informatif et/ou décisionnel. Pour un exemple de traitements, visiter le site : https://github.com/pavelk2/Bay-Area-Bike-Share
- En vous inspirant des tâches d'exploration et d'analyse décrites, proposer « vos tâches propres », en fonction de vos propres besoins/spécifications.
  a. Trace des analyses (requêtes, charts, etc.). Les traces seront restituées sous la forme de lignes de commande du **Notebook**.
  b. Certaines analyses pourront être faites avec des méthodes d'apprentissage (ML). Montrer à travers quelques exemples comment le ML peut aider à répondre à des besoins « métiers » : analyse du trafic, profilage/comportement utilisateurs, etc.
  c. Donner quelques critères de comparaison entre la situation actuelle (les données que vous avez choisies) et celles analysées par pavelk2 : augmentation du trafic, évolution de certains profils, etc.

C- Restitution du travail

- Sous la forme de Notebook déposé sur Ametice, d'un lien Git, etc.
  - Attention à bien noter et renseigner le lien vers vos données comme par exemple : https://s3.amazonaws.com/fordgobike-data/201801-fordgobike-tripdata.csv.zip

Ressources complémentaires :

- https://projectsbasedlearning.com/uncategorized/basics-about-databricks-notebook/
- https://projectsbasedlearning.com/uncategorized/provisioning-a-spark-cluster-or-creating-a-spark-cluster/

Attention

- Les ressources sont données à titre indicatif et pédagogique.
- Vous pouvez vous en inspirer mais éviter le « copier-coller » brut, sans valeur ajoutée, sans citation, etc.