Prise en main du framework Spark

1. Premiers pas sur Databricks

a) Commencez par une veille perso : qu'est-ce que c'est ? à quoi ça sert ? comment ça marche ?

b) Créez un compte et connectez-vous à la version *Community Edition*

c) Faites le tutoriel Quickstart disponible sur la plateforme

2. Manipulation de données avec pyspark

a) Téléchargez le fichier « transactions.csv » du projet dans Databricks

b) Créez un cluster pour l'exercice en prenant la dernier version stable

c) Dans un notebook, chargez le fichier CSV en PySpark dans un dataframe

d) Affichez le dataframe et affichez uniquement les 10 premiers éléments de la colonne Amount

e) Utilisez la méthode describe() et affichez les résultats

f) Calculez en utilisant les éléments du module pyspark.sql.functions :
- la moyenne, l'écart-type, la somme, le max et le min de la colonne Amount
- affichez les 5 plus importantes transactions
- la moyenne et la somme des transactions par pays émetteur
- la somme des transactions envoyés et reçues par "Laurent Alexandre" et le solde net

3. Du SQL depuis Databricks

a) À l'aide de la méthode df.createOrRempTempView, mettez le dataframe dans une table

b) Refaites les calculs de la question 2.f) en SQL

c) Créez une table avec les transactions émises depuis la France et l'enregistrez pour pouvoir y accéder ailleurs que dans ce notebook