TECHNO : SPARK, ELASTICSEARCH, KIBANA

DONNEES SOURCES : avocado.csv

NOTES : le redu devra se faire sous forme de fichier Powerpoint et d'un document PDF tout cela encapsulé dans un dossier portant le nom du groupe. Y ajouter un maximum de capture dans vos fichier Powerpoint et PDF

HDFS

- Créer les répertoires HDFS **'/raw_avocado', '/stagging_avocado' , '/cleaned_avocado'**
- Scinder le fichier avocado.csv en 5 fichiers (avocado_1.csv, avocado_2.csv, avocado_3.csv, avocado_4.csv, avocado_5.csv)-> utiliser l'outil ou la méthode de votre choix ou préférence
- Chargez le premier fichier 'avocado_1.csv' dans le répertoire HDFS '/raw_avocado'

SPARK (toutes les opérations doivent se faire dans le code PYSPARK) **:**

- Écrire un script PYSPARK qui récupère le fichier csv dans '/raw_avocado' et pour chaque ligne du fichier, rajoute les colonnes 'jour' et 'mois' qui seront des extractions du champ 'date'. Ensuite sauvegardez le nouveau résultat dans un fichier csv (avocado_cleaned_1.csv, ...) et le stocker dans '/stagging_avocado', ensuite supprimer le fichier ('avocado_1.csv', ...) du répertoire '/raw_avocado'
- Écrire un script qui récupère le fichier (avocado_cleaned_1.csv, ...) csv du répertoire '/stagging_avocado' et l'indexe dans ELASTICSEARCH dans un indexe nommé **'avocado'** ensuite le déplace le fichier csv dans le répertoire '/cleaned_avocado'

KIBANA :

Faite un Dashboard constitué de quatre graphes de votre choix permettant de décrire les données (plus les graphes sont pertinents, meilleure est la note). faire en sorte que les graphes se rafraichissent toutes les 5s

Refaire :

- Déplacer le fichier **'avocado_2.csv'** dans '/raw_avocado' et exécuter les codes PYSPARK et vérifier les évolutions dans KIBANA
- Faire le même processus pour les fichiers 'avocado_3.csv, avocado_4.csv, avocado_5.csv'

Bonne Chance