TECHNO : MYSQL, NIFI, HDFS, SPARK, HIVE

DONNEES SOURCES : avocado.csv

Délai : 2 semaines

NOTES : le redu devra se faire sous forme de présentation (date à définir) et de fichiers (Powerpoint et un document PDF) qui sera 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',** , **'/refine_avocado'**

MYSQL

- Créer une bd MYSQL nommé '**TP_MASTER'**
- Créer une table MySQL nommé **TP_MASTER.'avocado'** respectant la structure du fichier '**avocado.csv'**
- 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 la table MYSQL **TP_MASTER.'avocado'** -> utiliser l'outil ou la méthode de votre choix ou préférence

NIFI :

Avec NIFI, créer un process permettant d'ingérer les données de cette table MYSQL, et de stocker le résultat dans le répertoire HDFS **'/raw_avocado'** dans un fichier csv dont le nom sera au format **'avocado_YYYYMMJJHHmm.csv'**

HIVE partie 1 :

- Créer une table interne HIVE nommé **TP_MASTER.'avocado_volume_tracking'** avec 2 champs '**filename'** (STR) et '**somme_volume'** (FLOAT)
    - '**filename'**: nom du fichier ingéré (peut être avocado_1.csv, avocado_2.csv, avocado_3.csv, avocado_4.csv, avocado_5.csv)
    - '**somme_volume'**: somme des valeurs du champ **'volume'** pour le fichier enregistré dans **'filename'**
    - **ps**: cette table sera alimenté par **PYSPARK** dont les instructions seront définies ci-dessous
    - Exemple contenu de la table **'avocado_volume_tracking'**:

filename	**somme_volume**
avocado_1.csv	56000.5
avocado_2.csv	40400.5

SPARK partie 1(toutes les opérations doivent se faire dans le code **PYSPARK**) **:**

- Écrire un script permettant de récupérer le fichier déposé dans **'/raw_avocado'** et faire les opérations nécessaires pour alimenter la table HIVE **'avocado_volume_tracking'**. Ensuite, déplacer le fichier csv de **'/raw_avocado'** vers **'/stagging_avocado'**

SPARK Partie 2 :

Écrire un script PYSPARK qui récupère le fichier csv dans **'/stagging_avocado'** et pour chaque ligne du fichier, rajoute les colonnes **'jour'** et **'mois'** qui seront des extractions du champ **'date'**. Sauvegardez le nouveau résultat dans un fichier csv (**avocado_cleaned_YYYYMMJJHHmm.csv**, ...) et le stocker dans **'/refine_avocado'**, ensuite supprimez le fichier (**'avocado_YYYYMMJJHHmm.csv'**, ...) du répertoire **'/stagging_avocado'**

HIVE partie 2:

- Créer une bd HIVE TP_MASTER
- Créer une table externe HIVE qui pointe sur le répertoire HDFS **/refine_avocado**.
- Créer une vue HIVE qui retourne **la somme du champ 'Volume' et par mois, ce peu importe l'année**

Refaire :

- Rajoutez les données du fichier **'avocado_2.csv'** dans la table MYSQL **TP_MASTER.'avocado'**
- Exécuter le process **NIFI** et les script **PYSPARK**, ensuite observez les résultats
- Faire le même processus pour les fichiers **'avocado_3.csv, avocado_4.csv, avocado_5.csv'**

Cron :

- Scheduler votre process NIFI pour qu'il tourne toute les 5 mins
- Cronner vos jobs PYSPARK pour qu'il s'exécute tous les 10 et 7 mins
- Faite tourner tout votre WORKFLOW (NIFI et SPARKs)

Bonne Chance