Master 1

FRAMEWORK LOGICIEL POUR LE BIG DATA

ATELIER HIVE

R.JAZIRI

Objectifs

- Se familiariser avec Hive et HQL.
- Manipuler, transformer, analyser les données avec Hive.

- **Enoncé:** on cherche à requêter les données des visiteurs de la maison blanche déjà intégrées dans HDFS

- **Données:** Le format des lignes du fichier est le suivant : lname,fname,time_of_arrival,appt_scheduled_time,meeting_location,info_comment

- **Questions**:

  - Créez une base de données Hive et spécifiez son chemin dans le HDFS
  - Associez une table externe aux données des visiteurs.
  - Lancez une requête pour vérifier l'état de la table en limitant à 20 lignes.
  - Affichez les lignes ayant la colonne lname qui commence par OME.
  - Comptez le nombre de lignes de la table.
  - Supprimez la table externe et affichez le fichier viste.txt dans le HDFS
  - Recréez la table externe et intégrez les données associées à cette table dans une table Hive finale partitionnée par l'année de visite.
  - Supprimez la table finale. Que remarquez-vous ?

- **Enoncé**: Nous allons nous intéresser aux visiteurs de la maison blanche en fonction de leurs dates d'arrivées.
- **Données**: Le format des lignes du fichier est le suivant : lname,fname,time_of_arrival,appt_scheduled_time,meeting_location,info_comment
- **Questions**:
  - Créez une requête qui récupère toutes les lignes de la table wh_visits_finale en éliminant les lignes ayant time_of_arrival vide Lancez une requête pour vérifier l'état de la table en limitant à 20 lignes.
  - Trouvez les 10 premières visites.
  - Trouvez la dernière visite.
  - Trouvez les commentaires les plus communs.
  - Vous constatez que le blanc est le commentaire le plus en commun, modifiez la requête afin qu'elle ignore les commentaires vides.
  - Trouvez les commentaires les moins fréquents.
  - Plusieurs variations de RECEPTION GÉNÉRALE se produisent dans le top 10.Trouvez le nombre le nombre de visites réelles impliquant une réception générale en essayant de nettoyer certaines de ces incohérences dans les données.
  - Trouvez les personnes qui ont le plus visité la maison blanche

- Trouvez les personnes qui ont le plus visité la maison blanche - utilisez la commande explain avant le select. Que constatez-vous?

- Affichez les 5 premiers enregistrement de la table - utilisez la commande explain avant le select. Que constatez vous?

- Expliquez pourquoi dans le select HIVE n'utilise pas le MapReduce.

- **Enoncé**: Nous allons nous intégrer des messages twitter avec des informations sur les utilisateurs dans une table Hive externe.

- **Données**: Le format du fichier JSON
- **Questions**:

  - Intégrez le fichier twitter_data dans le hdfs.
  - Créez un schéma Hive de telle sorte que vous pouvez afficher les tweet associés à l'utilisateur 'Amiee'
  - Créez une table externe associer au fichier twitter_data capable de supporter le format JSON.
  - Créez une requête qui récupère toutes les lignes de la table.
  - Créez une table finale (userlocation,id,name,screenname,geoenabled,tweetmessage,createddate,geolocation) partitionnée par l'année de création du tweet.
  - Créez un script d'insertion depuis la table externe vers la table finale.

Résultat

- Vous êtes en mesure de créer des scripts Hive de manipulation et de transformation d'une grande quantité de données.