Atelier : Utiliser HIVE

L'objectif :

- Stocker et réaliser des calcules sur les statistique de baseball de 1871 à 2011.
    - Nous allons trouver le joueur avec le plus de « run » pour chaque année.

Emplacement du fichier : /hive/

Réalisation : Vous aller créer des tables Hive pour stocker les donner statistiques puis requêter ces tables pour obtenir les informations voulues

Chapitre correspondant : HIVE

1- Visualiser les données statistiques des batteurs :
   a- Ouvrir le terminal en ligne de commande de votre VM, et aller lire le fichier de Batting.cvs situé dans le répertoire /formation/ateliers/hive/
2- Copier ce fichier dans HDFS (soit via IHM soit via PIG)

3- Créer une table temporaire pour stocker les données:
   a- Se connecter à Hive
   a- Créer une table externe pour visualiser le fichier des batteurs
   b- Que s'est-il passé sous hdfs /user/hive/warehouse/
   c- Requêtez la table ? Qu'est ce que vous remarquez ?
   d- Déposer le fichier Batting.csv dans le répertoire hdfs `/hive/warehouse/`

4- Créer une table pour stocker les données:
   Créer une table nommée « batting » contenant l'identifiant des batteurs, l'année et le nombre de runs réalisés
5- Identifier pour chaque année le batteur qui a réalisé le plus de Run et insérer ces informations dans la table batting:

6- Récupérer le nom du fichier de stockage, l'id du joueur dont les runs sont supérieur à 150.
7- Supprimer la table temp_batting:
8- Vérifier le dossier /user/hive/warehouse/.
9- Que se passe si nous recréons la table ?
10- Créer une table « local » temp_master
11- Alimenter la table temp_master avec le fichier /user/cloudera/demo/Master.csv
12- Que se passe-t-il si nous supprimons la table temp master.csv ?

Partitions

13- Créer une table names qui contient une colonne id (entier) et une colonne name (text) et une colonne state (texte). Cette table sera partitionnée par la colonne state.

14- Charger les fichiers hivedata_<<state>>.txt dans la table names

15- Vérifier que toutes les données sont bien dans la table

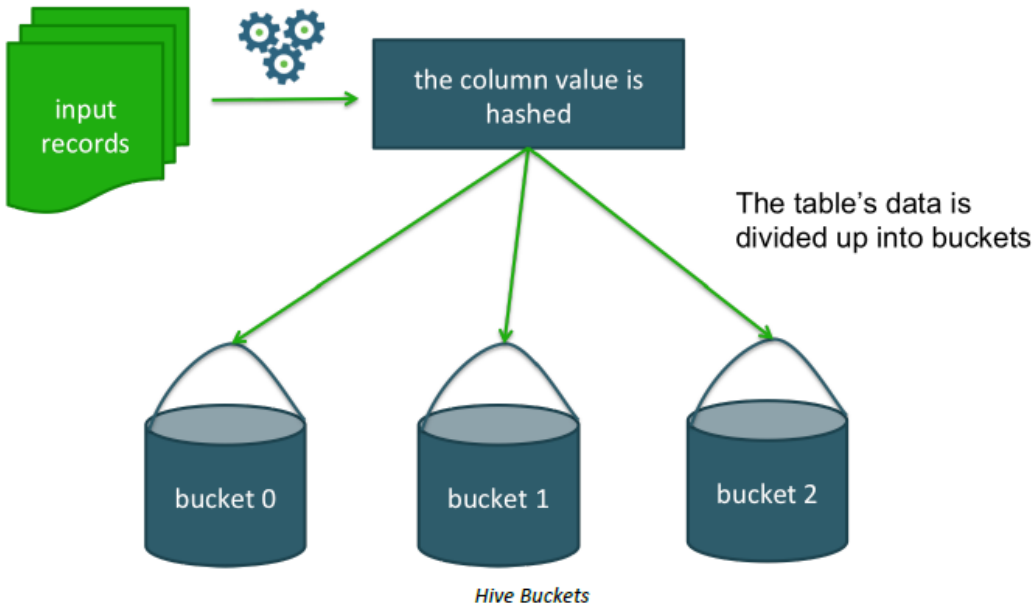
16- Vérifier les partitions avec la commande show partitions <table>

17- Comment se traduit les partitions sous hdfs :

18- Quand dans une requête on spécifie la partition, Hive est beaucoup plus performant et va lire seulement le répertoire correspondant

19- Remarque : il n'y a pas de job MapReduce qui s'est exécuté. Pourquoi ?

Bucket



20- Exécuter la requête suivante :

```
create table names_bucket (id int, name string, state string)
   clustered by (id) into 2 buckets;
```

```
load data local inpath
'/home/cloudera/formation/ateliers/hive/hivedata_ca.txt'
into table names_bucket ;
```

21- Insérer les données :

22- Faire un select * from names_bucket et aller voir dans HDFS

23- Insérer les données de names en écrasant les données de la table names_bucket , puis aller voir dans HDFS