Big data analytique

Data science
Master 2 ISiDIS
2015 / 2016

Le but est de mettre en place la chaîne de traitement qui permet de passer d'une base de données relationnelle existante à un base de données nosql où le big data analytique est possible.

1 Machine virtuelle cloudera

Cloudera est une société qui propose des logiciels et services basées sur hadoop. Elle propose une distribution de hadoop et de son environnement principal (MapReduce, HDFS, Hive, Avro, HBase, spark, etc.).

Q1.a Télécharger et lancer la machine virtuelle (gratuite) Cloudera quickStart VM CDH5 :
`http://www.cloudera.com/content/www/en-us/downloads/quickstart_vms/5-5.html`

Q1.b L'outil hue [1] est un interface web qui permet d'intéragir avec l'environnement hadoop. Lancer hue depuis un nagivateur web à l'adresse `localhost:8888` puisque hue communique avec le port 8888.

Q1.c Observer le système de fichier HDFS depuis le "file browser" ou "Navigateur de fichier" en français. Il est aussi possible d'observer les job MapReduce en cours d'exécution et bien d'autres choses comme lancer des commandes hive un peu comme on le ferait avec phpmysql.

2 De mySQL à HDFS avec scoop

Sqoop [2] est un outil pour échanger des données entre une base de données relationnelle et le système de fichier HDFS de hadoop. Il entre dans la catégorie des outils ETL (Extract Transform Load) comme les outils de Talend.

Scoop va lire la structure des tables mysql, importe les données et génère du code java avec les fonctions input et output pour MapReduce.

Q2.a Depuis un terminal, vérifier l'état du système de fichier HDFS avec la commande :
`hadoop fs -ls /user/hive`

Q2.b Effacer le répertoire warehouse avec la commande :
`hadoop fs -rm -f -r /user/hive/warehouse`

1. `http://gethue.com/`
2. `http://sqoop.apache.org/`

Q2.c La commande suivante permet d'importer une base de données mysql d'exemple de cloudera dans le système de fichier HDFS au format `parquet` [3] :

```
sqoop import-all-tables
    -m 4
    --connect jdbc:mysql://quickstart.cloudera:3306/retail_db
    --username=retail_dba
    --password=cloudera
    --as-parquetfile
    --warehouse-dir=/user/hive/warehouse > import.log
```

L'option `-m 5` correspond au nombre de mappers à exécuter simultanément.

Q2.d Vérifier que les fichiers ont bien été importé dans le système de fichier HDFS avec la commande : `hadoop fs -ls /user/hive/warehouse`

Vous trouverez beaucoup de documentation sur les commandes scoop comme par exemple ici : `http://cours.tokidev.fr/bigdata/cours/mbds_big_data_hadoop_cours_2.pdf`

3 De HDFS à Hive

Hive [4] est une surcouche analytique de hadoop : dans Hive, les données de HDFS sont abstraites sous forme tabulaire et relationnel. Un langage déclaratif proche de SQL (HiveQL) permet ensuite de manipuler ces données. Concrètement Hive transforme une requête HiveQL en des jobs MapReduce pour opérer sur les données sous un format particulier dans le système de fichier HDFS.

Vous pouvez directement taper les requêtes hive dans un interpréteur en lançant la commande `hive` dans un terminal, ou vous pouvez les requêtes hive dans l'interface web hue.

Q3.a Importer la table `customers` dans hive par la commande suivante :

```
CREATE EXTERNAL TABLE customers (
   customer_id int,
   customer_fname varchar(45),
   customer_lname varchar(45),
   customer_email varchar(45),
   customer_password varchar(45),
   customer_street varchar(255),
   customer_city varchar(45),
   customer_state varchar(45),
   customer_wipcode varchar(45)
)
STORED AS PARQUET
LOCATION 'hdfs:///user/hive/warehouse/customers';
```

3. https ://parquet.apache.org/
4. https ://hive.apache.org/

Q3.b Vérifier votre table en tapant la commande suivante qui vous donne le premier enregistrement :

```
SELECT * FROM customers limit 1;
```

Q3.c Maintenant, vous pouvez tester un requête simple et presque habituelle :

```
SELECT c.customer_fname, c.customer_lname as custname
FROM customers c
WHERE c.customer_fname LIKE 'A%'
AND c.customer_lname = 'Smith'
ORDER BY custname;
```

Ceci est un tout première exemple qui servira de base pour d'autres requêtes. N'hésitez à explorer les documents en ligne.