Hive – Hands-On

Objetivo: O objetivo deste hands-on é instalar e testar o software Hive.

Instalação do Hive

1. Faça download da versão 2.3.3 do Hive (https://hive.apache.org).

```
$ cd /vagrant
$ wget -c http://ftp.unicamp.br/pub/apache/hive/hive-2.3.3/apache-hive-2.3.3-bin.tar.gz
```

2. Descompacte o arquivo.

```
$ tar -zxvf apache-hive-2.3.3-bin.tar.gz
$ export HIVE_HOME=/vagrant/apache-hive-2.3.3-bin
```

3. Faça o download dos arquivos stations.csv e . Crie as pastas bikeshare/stations e bikeshare/trips no HDFS e faça upload dos arquivos para esta pasta.

```
$ wget -c
https://raw.githubusercontent.com/GerInfraBigDataPUCRS/CursoDataScience2018/master/07_H
ive/stations.csv
$ wget -c
https://raw.githubusercontent.com/GerInfraBigDataPUCRS/CursoDataScience2018/master/07_H
ive/trips.csv
$ $HADOOP_HOME/bin/hadoop fs -mkdir -p bikeshare/stations
$ $HADOOP_HOME/bin/hadoop fs -put stations.csv bikeshare/stations
$ $HADOOP_HOME/bin/hadoop fs -mkdir -p bikeshare/trips
$ $HADOOP_HOME/bin/hadoop fs -put trips.csv bikeshare/trips
```

4. Crie diretórios usados pelo Hive no HDFS e inicialize o Hive metastore usando o banco de dados Derby local

```
$ $HADOOP_HOME/bin/hadoop fs -mkdir /tmp
$ $HADOOP_HOME/bin/hadoop fs -mkdir -p /user/hive/warehouse
$ $HADOOP_HOME/bin/hadoop fs -chmod g+w /tmp
$ $HADOOP_HOME/bin/hadoop fs -chmod g+w /user/hive/warehouse
$ $HIVE_HOME/bin/schematool -dbType derby -initSchema
```

5. Abra uma sessão no HIVE CLI

```
$ $HIVE_HOME/bin/hive
```

6. Crie uma nova base de dados chamado bikeshare e defina ela para o contexto atual

```
CREATE DATABASE bikeshare;
SHOW DATABASES;
USE bikeshare;
```

7. Crie uma tabela para o dataset stations

```
CREATE EXTERNAL TABLE stations (
station_id INT,
name STRING,
lat DOUBLE,
```

```
long DOUBLE,
dockcount INT,
landmark STRING,
installation STRING
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
STORED AS TEXTFILE
LOCATION 'hdfs:///user/vagrant/bikeshare/stations';
```

8. Crie uma tabela para o dataset

```
CREATE EXTERNAL TABLE trips (
trip_id INT,
duration INT,
start_date STRING,
start_station STRING,
start_terminal INT,
end_date STRING,
end_station STRING,
end_terminal INT,
bike_num INT,
subscription_type STRING,
zip_code STRING
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
STORED AS TEXTFILE
LOCATION 'hdfs:///user/vagrant/bikeshare/trips';
```

9. Mostre as tabelas existentes

```
SHOW TABLES;
DESCRIBE stations;
DESCRIBE trips;
DESCRIBE FORMATTED stations;
DESCRIBE FORMATTED trips;
```

10. Execute uma consulta para verificar o número de viagens (trips) realizados por cada terminal.

```
SELECT start_terminal, start_station, COUNT(1) AS count
FROM trips
GROUP BY start_terminal, start_station
ORDER BY count
DESC LIMIT 10;
```

11. Execute uma consulta para realizar um join entre os datasets stations e trips

```
SELECT t.trip_id, t.duration, t.start_date, s.name, s.lat, s.long, s.landmark
FROM stations s
JOIN trips t ON s.station_id = t.start_terminal
LIMIT 10;
```

12. Saia do Hive CLI

```
EXIT;
```