Laboratórios de Hive

Atividade 01 - Obtenção de dados e inserção de dados no HDFS

Você pode executar este laboratório na sandbox HDP 2.6.4 no Hive ou em qualquer infraestrutura Hadoop que tenha o HDFS, MAPREDUCE, YARN, TEZ e HIVE instalados e executando. Se você for executar em outro ambiente, faça os ajustes nos caminhos dos arquivos no sistema local e no sistema de arquivos HDFS.

Faça download dos arquivos que deverão ser inseridos no HDFS:

```
$ git clone https://github.com/leonardoamorim/arquiteturadebigdata.git
$ cd arquiteturadebigdata
$ ls
```

Os arquivos que serão copiados para o HDFS:

- flight_delays1.csv
- flight_delays2.csv
- flight_delays3.csv
- sfo_weather.csv

```
$ hdfs dfs -mkdir /user/maria_dev/flightdelays
$ hdfs dfs -mkdir /user/maria_dev/sfo_weather
$ hdfs dfs -put flight_delays* /user/maria_dev/flightdelays
$ hdfs dfs -put sfo_weather.csv /user/maria_dev/sfo_weather
```

Agora, execute o hive via linha de comando:

```
$ hive
```

```
VERTICES: 01/01  [==========================>>] 100%  ELAPSED TIME: 4.79 s
--------------------------------------------------------------------------------
Loading data to table default.weather_partitioned partition (year=2008, month=1)
Partition default.weather_partitioned{year=2008, month=1} stats: [numFiles=1, numRows=31, totalSize=806, rawDataSize=4371]
OK
Time taken: 8.399 seconds
hive> SELECT * FROM weather_partitioned;
OK
SAN FRANCISCO INTERNATIONAL AIRPORT CA US       1       0       122     39      2008    1
SAN FRANCISCO INTERNATIONAL AIRPORT CA US       2       0       117     39      2008    1
SAN FRANCISCO INTERNATIONAL AIRPORT CA US       3       43      150     94      2008    1
SAN FRANCISCO INTERNATIONAL AIRPORT CA US       4       533     150     100     2008    1
SAN FRANCISCO INTERNATIONAL AIRPORT CA US       5       196     122     78      2008    1
SAN FRANCISCO INTERNATIONAL AIRPORT CA US       6       15      106     50      2008    1
SAN FRANCISCO INTERNATIONAL AIRPORT CA US       7       0       111     67      2008    1
SAN FRANCISCO INTERNATIONAL AIRPORT CA US       8       20      128     61      2008    1
SAN FRANCISCO INTERNATIONAL AIRPORT CA US       9       3       106     67      2008    1
SAN FRANCISCO INTERNATIONAL AIRPORT CA US       10      25      100     89      2008    1
SAN FRANCISCO INTERNATIONAL AIRPORT CA US       11      0       117     89      2008    1
SAN FRANCISCO INTERNATIONAL AIRPORT CA US       12      0       133     83      2008    1
SAN FRANCISCO INTERNATIONAL AIRPORT CA US       13      0       144     67      2008    1
SAN FRANCISCO INTERNATIONAL AIRPORT CA US       14      0       133     56      2008    1
SAN FRANCISCO INTERNATIONAL AIRPORT CA US       15      0       144     61      2008    1
SAN FRANCISCO INTERNATIONAL AIRPORT CA US       16      0       133     44      2008    1
SAN FRANCISCO INTERNATIONAL AIRPORT CA US       17      0       139     61      2008    1
SAN FRANCISCO INTERNATIONAL AIRPORT CA US       18      0       150     33      2008    1
SAN FRANCISCO INTERNATIONAL AIRPORT CA US       19      0       122     39      2008    1
SAN FRANCISCO INTERNATIONAL AIRPORT CA US       20      0       111     72      2008    1
SAN FRANCISCO INTERNATIONAL AIRPORT CA US       21      152     83      61      2008    1
SAN FRANCISCO INTERNATIONAL AIRPORT CA US       22      25      89      44      2008    1
SAN FRANCISCO INTERNATIONAL AIRPORT CA US       23      15      83      61      2008    1
SAN FRANCISCO INTERNATIONAL AIRPORT CA US       24      76      78      50      2008    1
SAN FRANCISCO INTERNATIONAL AIRPORT CA US       25      645     117     72      2008    1
SAN FRANCISCO INTERNATIONAL AIRPORT CA US       26      58      144     94      2008    1
SAN FRANCISCO INTERNATIONAL AIRPORT CA US       27      81      133     72      2008    1
SAN FRANCISCO INTERNATIONAL AIRPORT CA US       28      38      100     50      2008    1
SAN FRANCISCO INTERNATIONAL AIRPORT CA US       29      20      100     33      2008    1
SAN FRANCISCO INTERNATIONAL AIRPORT CA US       30      28      117     50      2008    1
SAN FRANCISCO INTERNATIONAL AIRPORT CA US       31      13      117     72      2008    1
Time taken: 0.331 seconds, Fetched: 31 row(s)
```

Atividade 02 - Criação de tabela externa no Hive

```
DROP TABLE IF EXISTS flightdelays;

CREATE EXTERNAL TABLE flightdelays (
    Year INT,
    Month INT,
    DayofMonth INT,
    DayOfWeek INT,
    DepTime INT,
    CRSDepTime INT,
    ArrTime INT,
    CRSArrTime INT,
    UniqueCarrier STRING,
    FlightNum INT,
    TailNum STRING,
    ActualElapsedTime INT,
    CRSElapsedTime INT,
    AirTime INT,
    ArrDelay INT,
    DepDelay INT,
    Origin STRING,
    Dest STRING,
    Distance INT,
    TaxiIn INT,
    TaxiOut INT,
    Cancelled INT,
    CancellationCode STRING,
    Diverted INT,
    CarrierDelay INT,
    WeatherDelay INT,
    NASDelay INT,
    SecurityDelay INT,
    LateAircraftDelay INT
)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE
LOCATION '/user/maria_dev/flightdelays/';
```

Atividade 03 - Analisando dados com o Hive

Escreva uma consulta do Hive para cada uma das tarefas abaixo.

1. Calcular o atraso médio dos voos que aterram em Denver (destino igual a "DEN")
2. Calcular o atraso médio dos voos em que a origem é LAX e o destino é SFO
3. Determinar qual aeroporto de destino teve o maior atraso médio

```
SELECT AVG(arrdelay) FROM flightdelays WHERE dest = 'DEN';
SELECT AVG(arrdelay) FROM flightdelays WHERE origin = 'LAX' AND dest = 'SFO';
SELECT AVG(arrdelay) AS delay, dest FROM flightdelays GROUP BY dest ORDER BY
delay DESC LIMIT 1;
```

Atividade 04 - Definir e preencher uma tabela ORCFile

Defina uma tabela Hive chamada sfo_weather que satisfaça todos os critérios a seguir:

1. Uma tabela gerenciada pelo Hive;
2. Os dados são armazenados no formato ORCFile;
3. A tabela é preenchida com registros no arquivo arquiteturadebigdata/sfo_weather.csv na máquina cliente;
4. O esquema corresponde às colunas em sfo_weather.csv - a primeira coluna é uma cadeia de caracteres denominada nome_da_estação, seguida de inteiros para Year, Month, DayOfMonth, precipitation, temperature_max e temperature_min

```
DROP TABLE IF EXISTS sfo_weather_txt;

CREATE TABLE sfo_weather_txt(station_name STRING,
          Year INT, Month INT, DayOfMonth INT, precipitation INT, temperature_max
INT, temperature_min INT)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
STORED AS TEXTFILE;

LOAD DATA LOCAL INPATH '/home/maria_dev/arquiteturadebigdata/sfo_weather.csv'
OVERWRITE INTO TABLE sfo_weather_txt;

DROP TABLE IF EXISTS sfo_weather;
CREATE TABLE sfo_weather(station_name STRING,
          Year INT, Month INT, DayOfMonth INT, precipitation INT, temperature_max
INT, temperature_min INT)
STORED AS ORC;

INSERT INTO TABLE sfo_weather SELECT * FROM sfo_weather_txt;

SELECT * FROM sfo_weather;
```

Atividade 05 - Junção de dados do Hive

Escreva uma consulta Hive em um arquivo chamado atividade03.hive que satisfaça os seguintes critérios:

1. Utilizar o Tez como motor de execução;
2. O resultado da consulta está numa nova tabela gerida pelo Hive, denominada flights_weather, armazenada como um ficheiro de texto;

3. Junta a tabela flightdelays com a tabela sfo_weather onde dest ou origin é igual a "SFO" em flightdelays, e o Year, Month e DayOfMonth são iguais nas duas tabelas;
4. Selecionar todas as colunas da tabela flightdelays, e a tabela tempera;

```
SET hive.execution.engine= tez;

DROP TABLE IF EXISTS flights_weather;
CREATE TABLE flights_weather STORED AS TEXTFILE AS SELECT fd.*,
sw.temperature_max, sw.temperature_min FROM flightdelays fd JOIN sfo_weather sw
ON fd.year = sw.year AND fd.month = sw.month AND fd.dayofmonth = sw.dayofmonth
WHERE fd.origin = 'SFO' OR fd.dest = 'SFO';

SELECT * FROM flights_weather;
```

Atividade 06 - Tabelas particionadas do Hive

Escreva uma consulta Hive que satisfaça os seguintes critérios:

1. Defina uma nova tabela gerenciada pelo Hive chamada weather_partitioned que tenha o mesmo esquema que a tabela sfo_weather;
2. A tabela é particionada nas colunas Year e Month;
3. Os dados são armazenados no formato ORCFile;
4. Inserir os registros de janeiro de 2008 da tabela sfo_weather na partição apropriada de weather_partitioned

```
DROP TABLE IF EXISTS weather_partitioned;

CREATE TABLE weather_partitioned(
station_name string,
  dayofmonth int,
  precipitation int,
  temperature_max int,
  temperature_min int)
PARTITIONED BY (year int, month int)
STORED AS ORC;

INSERT INTO TABLE  weather_partitioned PARTITION(year=2008, month=1) SELECT
station_name, dayofmonth, precipitation, temperature_max, temperature_min FROM
sfo_weather WHERE year = 2008 AND month = 1;

SELECT * FROM weather_partitioned;
```