Flight delay dataset Analysis using Hive

Naveen - (Founder & Trainer @ NPN Training) · Follow
2 min read · Feb 25, 2021

Listen Share

Hive is a data ware house infrastructure built on top of Hadoop ecosystem to query and analyze structured data. It gives SQl like semantics over Hadoop data(HDFS) by name Hive QL. Although now many SQl engine over hadoop like Impala,Drill,Presto has come but Hive was the original SQL engine on top of Hadoop.

Download the dataset
1. 2008.csv: Flight delay dataset from 2008.
2. airports.csv: Dataset linking airport codes to their full names.

There are 2 datasets in the repo.

a) The first dataset contains on-time flight performance data from 2008, originally released by Research and Innovative Technology Administration (RITA). The source of this dataset is http://stat-computing.org/dataexpo/2009/the-data.html. The dataset

b) The second dataset contains listing of various airport codes in continental US, Puerto Rico and US Virgin Islands. The source of this dataset is http://www.world-airport-codes.com/ The data was scraped from this website and then cleansed to be in its present CSV form.

Start all the services

```
[npntraining@centos8 ~] start-dfs.sh
[npntraining@centos8 ~] jps
NameNode
DataNode
SecondaryNameNode
[npntraining@centos8 ~] start-yarn.sh
```

```
[npntraining@centos8 ~] jps
NameNode
DataNode
SecondaryNameNode
ResourceManager
NodeManager
```

Hive Commands

Login to hive shell using the command shown below

```
[npntraining@centos8 ~] hive
hive>
```



```
hive> CREATE TABLE flight_data(
   year INT,
   month INT,
   day INT,
   day_of_week INT,
   dep_time INT,
   crs_dep_time INT,
   arr_time INT,
   crs_arr_time INT,
   unique_carrier STRING,
   flight_num INT,
   tail_num STRING,
   actual_elapsed_time INT,
   crs_elapsed_time INT,
   air_time INT,
   arr_delay INT,
   dep_delay INT,
   origin STRING,
   dest STRING,
   distance INT,
   taxi_in INT,
   taxi_out INT,
   cancelled INT,
   cancellation_code STRING,
   diverted INT,
   carrier_delay STRING,
   weather_delay STRING,
   nas_delay STRING,
   security_delay STRING,
   late_aircraft_delay STRING
)
```

```
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ',';
```

Load the data in flight_data

```
hive>LOAD DATA LOCAL INPATH '2008.csv' OVERWRITE INTO TABLE
flight_data;
```

Create airports table and load the data into the table

```
hive> CREATE TABLE airports(
   name STRING,
   country STRING,
   area_code INT,
   code STRING)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ',';

hive>LOAD DATA LOCAL INPATH 'airports.csv' OVERWRITE INTO TABLE
airports;
```

The rest of the blog can be read using this link: Flight delay dataset Analysis using Hive

Interested in Learning Big Data!

Join our 20 weekend program : Data Engineering Training

Big Data Hadoop Hive

Follow