EMR Workshop Lab 2 – Hive, Pig & EMR Steps

Exercise 1: Process data interactively

* Create an S3 bucket with folders:
  + files
  + logs
  + input
  + output
* Choose S3 service from AWS console.
* Click Create bucket and choose a unique name
* Click Next twice and click Create. Leave all other options with no changes.
* Click on your bucket name and click Create folder.
* Create 4 folders as above leaving encryption settings as None.
* Get sample data from here (1.8MB file): <https://s3.amazonaws.com/aws-data-analytics-blog/emrimmersionday/tripdata.csv>
* Upload file to your "input" folder in your S3 bucket
  + Click on the input folder and click Upload.
  + Choose the tripdata.csv and click Upload.
* SSH to master node of your previously created cluster.
* Run “hive” and create external table following these steps:

[hadoop@ip-10-0-0-135 ~]$ hive;

* Copy and paste the following script, make sure that you don’t have invisible characters. Use vi on mac/Linux or Notepad on Windows. Alternatively, you can download it from [here](https://s3.amazonaws.com/aws-data-analytics-blog/emrimmersionday/ny-taxi-test.hql)

hive>

CREATE EXTERNAL TABLE ny\_taxi\_test (

vendor\_id int,

lpep\_pickup\_datetime string,

lpep\_dropoff\_datetime string,

store\_and\_fwd\_flag string,

rate\_code\_id smallint,

pu\_location\_id int,

do\_location\_id int,

passenger\_count int,

trip\_distance double,

fare\_amount double,

mta\_tax double,

tip\_amount double,

tolls\_amount double,

ehail\_fee double,

improvement\_surcharge double,

total\_amount double,

payment\_type smallint,

trip\_type smallint

)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

LINES TERMINATED BY '\n'

STORED AS TEXTFILE

LOCATION "s3://<YOUR-BUCKET>/input/";

• Run test query. This script will query the NY taxi data and show 5 different rate code ids.

hive> select distinct rate\_code\_id from ny\_taxi\_test;

Exercise 2: Processing data with EMR steps

After you’ve created the Hive table and queried your data, you can practice scheduling the job on the cluster using EMR steps.

# Hive Step

* You will have to create a ny-taxi.hql text file and upload it to your "files" folder.
* Copy and paste the following script into ny-taxi.hql, make sure that you don’t have invisible characters. Use vi on mac/Linux or Notepad on windows. Alternatively, you can download it from [here](https://s3.amazonaws.com/aws-data-analytics-blog/emrimmersionday/ny-taxi.hql).

CREATE EXTERNAL TABLE ny\_taxi (

vendor\_id int,

lpep\_pickup\_datetime string,

lpep\_dropoff\_datetime string,

store\_and\_fwd\_flag string,

rate\_code\_id smallint,

pu\_location\_id int,

do\_location\_id int,

passenger\_count int,

trip\_distance double,

fare\_amount double,

mta\_tax double,

tip\_amount double,

tolls\_amount double,

ehail\_fee double,

improvement\_surcharge double,

total\_amount double,

payment\_type smallint,

trip\_type smallint

)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

LINES TERMINATED BY '\n'

STORED AS TEXTFILE

LOCATION "${INPUT}";

INSERT OVERWRITE DIRECTORY "${OUTPUT}"

SELECT \* FROM ny\_taxi WHERE rate\_code\_id = 1;

This script will query the ny\_taxi table and extract trips where standard rate is used.

* Go to the EMR console and scroll down to the “Step”.
* Add step, choose Hive program in "Step type"
* You need to add 3 locations to this step.
  1. Script S3 location: The first is the location of the script you just uploaded to S3. The format is: s3://<YOUR-BUCKET>/files/ny-taxi.hql
  2. Input S3 location: Where is your data source  
     (Note that you don’t want to specific the file. Hive reads in folders, not files). The input location is: s3://<YOUR-BUCKET>/input/
  3. Output S3 location: Where to store your processed data. The output location is: <s3://<YOUR-BUCKET>/output/hive/
* After you’ve added the information necessary, click “Add”.
* Check "output/hive" in 3 minutes.

Pig Step

* Run PIG script to parse data in CSV format and transform into TSV format
* Create a ny-taxi.pig text file and upload it to the "files" folder.
* Copy and paste the following script into ny-taxi.pig, make sure that you don’t have invisible characters. Use vi on Mac/Linux or Notepad on windows. Alternatively, you can download it from [here](https://s3.amazonaws.com/aws-data-analytics-blog/emrimmersionday/ny-taxi.pig).

DEFINE CSVLoader org.apache.pig.piggybank.storage.CSVLoader();

NY\_TAXI = LOAD '$INPUT' USING CSVLoader(',') AS

(vendor\_id:int,

lpep\_pickup\_datetime:chararray,

lpep\_dropoff\_datetime:chararray,

store\_and\_fwd\_flag:chararray,

rate\_code\_id:int,

pu\_location\_id:int,

do\_location\_id:int,

passenger\_count:int,

trip\_distance:double,

fare\_amount:double,

mta\_tax:double,

tip\_amount:double,

tolls\_amount:double,

ehail\_fee:double,

improvement\_surcharge:double,

total\_amount:double,

payment\_type:int,

trip\_type:int);

STORE NY\_TAXI into '$OUTPUT' USING PigStorage('\t');

This script will parse data stored as CSV file on S3 and output data in tab delimited table format.

* Go to the EMR console and scroll down to the “Step”.
* Add step, choose Pig program in "Step type"
* You need to add 3 locations to this step.
  1. Script S3 location: The first is the location of the script you just uploaded. The format is: s3://<YOUR-BUCKET>/files/ny-taxi.pig
  2. InputS3 location: Your data source (unlike Hive, Pig needs file entry location). The input location is: s3://<YOUR-BUCKET>/input/tripdata.csv
  3. Output S3 location: Where to store your processed data. The output location is: s3://<YOUR-BUCKET>/output/pig/
* After you’ve added the information necessary, click “Add”.
* Check "output/pig" in 2 minutes.