EMR Workshop Lab 2 – Hive, Pig & EMR Steps

(Updated 14-Nov-18)

This lab demonstrates submitting Hive/Pig work to an Amazon EMR cluster.

You can submit Hive work to your cluster interactively, or you can submit work as a cluster step using the console, CLI, or API. You can submit steps when the cluster is launched, or you can submit steps to a running cluster.

Exercise 1: Process data interactively

- Create an S3 bucket with folders:
  - files
  - logs
  - input
  - output
- 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
- 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 this script from here and edit it:

```
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 this script from here and edit it:

```
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 this script from here and edit it:

```
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.