CUSP-GX-6002.001: Big Data Management & Analysis

SPRING 2020

Homework 4 – Spatial Join with Apache Spark

Due: 5:30 PM, Apr 28, 2020

In this homework, we would like to generate spatial statistics for yellow taxi trips in NYC. We are interested in knowing the top 3 destinations for trips starting from each five borough of New York, i.e. Manhattan, Brooklyn, Queens, Bronx and Staten Island. For example, for Manhattan, *Upper East Side*, *Midtown West*, and *Laguardia Airport* could be the top 3 destinations for trip starting from the borough. For the month of May 2011, there were more than 15 million trips, each with pick-up and drop-off location information (i.e. latitude and longitude) and the total size is over 3GB. You are asked to write a Spark application to compute such statistics for the taxi trip records in May 2011. You are also provided with the spatial boundaries for the NYC boroughs and neighborhoods.

INPUT DATA SET:

yellow_tripdata_2011-05.csv

Source: https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page

Description: one-month extract of the TLC Trip Record Data for May 2011. The file is available on HDFS at: `hdfs:///tmp/bdm/yellow_tripdata_2011-05.csv`
(**including** its header)

For testing purposes, you can use a smaller file (e.g. on your local machine):
`hdfs:///tmp/bdm/yellow.csv.gz`

neighborhoods.geojson

Source: https://data.beta.nyc/dataset/pediacities-nyc-neighborhoods/resource/35dd04fb-81b3-479b-a074-a27a37888ce7

Description: extracted from the Pediacities NYC Neighborhoods polygons and correlated data, containing only neighborhood geometries, their names and corresponding boroughs.

This file is available on HDFS at: `hdfs:///tmp/bdm/neighborhoods.geojson`

boroughs.geojson

Source: https://data.cityofnewyork.us/City-Government/Borough-Boundaries/tqmj-j8zm

Description: the geometries and name of 5 boroughs in NYC. This file is also available on HDFS at:
`hdfs:///tmp/bdm/boroughs.geojson`

Please note that the NYC borough boundaries could be derived from the neighborhoods file as borough information are also included. However, they are both made available for your convenience.

OUTPUT

Please write a Spark application that takes the above file **`yellow_tripdata_2011-05.csv`** on HDFS as its input and produce the top 3 destination neighborhoods along with it counts for trips starting from each borough. The output is expected to be in the CSV line format (fields separated by commas) without header. The output also needs to be sorted by Borough name. A sample output is provided below (not a correct anwer).

SAMPLE OUTPUT (header not to be included):

```
Borough,Top1_Name,Top1_Count,Top2_Name,Top2_Count,Top3_Name,Top3_Count
Brooklyn,Upper East Side,10000,Midtown West,9999,Laguardia Airport,9998
Bronx,Upper East Side,10000,Midtown West,9999,Laguardia Airport,9998
Manhattan,Upper East Side,10000,Midtown West,9999,Laguardia Airport,9998
Queens,Upper East Side,10000,Midtown West,9999,Laguardia Airport,9998
Staten Island,Upper East Side,10000,Midtown West,9999,Laguardia Airport,9998
```

SUBMISSION:

You can turn in one or more files including your application's main (Python) file and any dependencies that it may need. However, all of the submitted file(s) must be able to fit into a single **spark-submit** command running on NYU HPC Dumbo. Please provide this command when submitting your code. For sanity check, please also include the results in the body of your NYU Classes submission.

Evaluation: your code will be tested to run with exactly 25 cores (5 executors and 5 cores per executor), and 15GB of memory per executor (3GB per core). In other words, your code will be run with the folowing command structure (in a single line):

SAMPLE RUN:

```
spark-submit --num-executors 5 --executor-cores 5 --executor-memory 15g \
  --conf spark.executorEnv.LD_LIBRARY_PATH=$LD_LIBRARY_PATH \
  --files hdfs:///tmp/bdm/neighborhoods.geojson,hdfs:///tmp/bdm/boroughs.geojson \
  BDM_HW4.py hdfs:///tmp/bdm/yellow_tripdata_2011-05.csv output
```

Note: the above command is only an example to demonstrate how to specify the number of executors. It might not run if you just copy and paste into the console (since your file might not be BDM_HW4.py).