The City College of New York | Department of Computer Science

DSE I2450: Big Data and Scalable Computation

SPRING 2019

Homework 5 – Spatial Join with Apache Spark

Due: 5:00 PM, Apr 10, 2019

In this homework, we would like to generate spatial statistics for yellow taxi trips in NYC. We are interested to know for destinations in each borough of New York, i.e. Manhattan, Brooklyn, Queens, Bronx and Staten Island, the top 3 neighborhoods (e.g. West Village, Williamsburg, Flushing, etc.) that those trips originated from. For example, an answer could be that *Upper East Side*, *Midtown West*, and *Laguardia Airport* are the top 3 origin neighborhood for trips ending up in Manhattan. 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.

DATA SET:

yellow_tripdata_2011-05.csv.gz

Source: http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml
Description: one-month extract of the TLC Trip Record Data for May 2011. The file is available on HDFS at: hdfs:///data/share/bdm/yellow_tripdata_2011-05.csv.gz
(**including** with its header)
For testing purposes, you can use a smaller file (e.g. on your local machine):
hdfs:///data/share/bdm/yellow.csv.gz

neighborhoods.geojson

Source: http://catalog.opendata.city/dataset/pediacities-nyc-neighborhoods
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:///data/share/bdm/neighborhoods.geojson

boroughs.geojson

Source: https://data.cityofnewyork.us/City-Government/Borough-Boundaries/tqmj-j8zm
Description: extracted from NYC Department of City Planning, containing the geometries and name for 5 boroughs of NYC. This file is also available on HDFS at:
hdfs:///data/share/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.

OBJECTIVE: (10 pts)

Please write a Spark application that takes the above file **yellow_tripdata_2011-05.csv.gz** on HDFS as its input and produce the top 3 origin neighborhoods that delivered passengers to each of the five borough based on the number of trips served.

Your 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 cluster. Please provide this command when submitting your code. For sanity check, please also include the results in the body of your Blackboard submission.
Evaluation: your code will be tested to run with exactly 25 cores (5 executors and 5 cores per executor). In other words, your code will be run with the folowing command structure (in a single line):

```
spark-submit --num-executors 5 –executor-cores 5 --files \
  hdfs:///data/share/bdm/neighborhoods.geojson,hdfs:///data/share/bdm/boroughs.geojson \
  application.py hdfs:///data/share/bdm/yellow_tripdata_2011-05.csv.gz
```

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 is not application.py).