The City College of New York | Department of Computer Science

DSE I2450: Big Data and Scalable Computation

SPRING 2019

Lab 7 – Spatial-Temporal Processing

This is a comprehensive lab for processing spatio-temporal data sources in a streaming fashion. In particular, you are given two data sets:

- **citibike.csv** – 1 week worth of the citibike trip record from Feb-1 to Feb-8, 2015. This is the data set that we have been using extensively in class
- **yellow.csv.gz** – this is yellow taxi trip record for Feb-1 2015, in the GZip format. This is a compressed format similar to ZIP but has been used extensively in Unix system. This file format is also supported natively by Hadoop/Spark.

Problem Statement: we suspect that there were many people that used the yellow cab service to get to CitiBike stations for their daily commutes. As a first step to test this hypothesis, we would like to count how many CitiBike trips that could have been matched with a prior taxi trip around the "Greenwich Ave & 8 Ave" station on February 1st 2015. A CitiBike trip is defined as "matched" if it can be paired with a taxi trip given that:

- The taxi trip ended within 0.25 miles of the above station
- The Citibike trip starting at the above station must have happened after the taxi trip ended but cannot be more than 10 minutes after that.

Your Objecive: you are to write a complete .py script that can be executed on the NYU cluster. The output of your script is a single number indicating the number of potential connected CitiBike trips. Your .py file must run with the spark-submit command using the above two files, and output to the standard output (using 'print' is fine).

Note:

- Spark can read .gz file directly, e.g. sc.textFile("abc.csv.gz") as a regular text abc.csv text file.
- Since we only interested in potentially matched CitiBike trips, it is acceptable to match a taxi trip to multiple CitiBike trips. For example, if there is one taxi trip at 9:00am and two Citbike trips at 9:05am and 9:06am, that will count as **2**.
- Both input files are available on Blackboard under /data/share/bdm on HDFS.