CSE 511 Project Phase 2 Report

Team Members (Group 33):
Janam Vaidya - jvaidya4@asu.edu
Vishal Krishna Bettadapur - vbettada@asu.edu
Khushank Goyal - kgoyal7@asu.edu
Shreeshiv Patel - spate137@asu.edu

Abstract:

This progress report has been written to define the problem mentioned in phase 2 of the course project and our attempt at developing a solution for it.

Problem Statement:

In phase 2, we need to complete two different hot spot analysis tasks: Hot Zone Analysis and Hot Cell Analysis. A rectangle dataset and a point dataset is used in Hot Zone Analysis. For each of the rectangles, the number of points located within the rectangle will be obtained. The more points a rectangle consists of, the hotter it is going to be with our task being calculating the hotness of all rectangles. The task while performing hot cell analysis will be to apply spatial statistics to spatio-temporal Big Data in order to identify statistically significant hot spots using the Apache Spark framework.

Implementation:

- **Hot Zone Analysis**

  The input will be pick up point data of NYC taxi trip datasets Zone data. On the rectangle dataset and the point dataset, we performed a range join operation. The number of points located within the rectangles for each of the rectangles will be obtained. The more points a rectangle will have, we will consider that rectangle more hotter, i.e., The number of points in the rectangle will determine the hotness of the rectangle. We use the function ST_Contains implemented in Phase 1 to determine whether the rectangle contains a point or not. Once we determine which points are contained within which rectangles, we count the total number of points within each rectangle; the data is grouped by the rectangle and also ordered by the rectangle. The data that is returned is that of a rectangle along with its hotness. The output will be all zones with their count, sorted by "rectangle" string in an ascending order.

- **Hot Cell Analysis**

  Here, the input point data is a monthly NYC taxi trip dataset (2009-2012) like "yellow_tripdata_2009- 01_point.csv". In hot cell analysis, we initially parse a csv of pickup location points along with their corresponding timestamps and then we convert that to a spark dataframe of x,y,z coordinates of space/time. To reduce noise, we filter

out those coordinates which are not within a predefined boundary. On the spark dataframe, we perform aggregate functions such as mean, count which is to be used as a value attribute, standard deviation; and a self-join is performed with the constraint that only coordinates within 1 coordinate of each other are paired together. The joined pairs in the data frame is used to calculate the total sum of the value attribute. This joined data frame is then grouped and then in the Getis-Ord equation shown below, we enter the sum of values, standard deviation, total number of coordinates and the average.

$$G_i^* = \frac{\sum_{j=1}^{n} w_{i,j}x_j - \bar{X}\sum_{j=1}^{n} w_{i,j}}{S\sqrt{\frac{\left[n\sum_{j=1}^{n} w_{i,j}^2 - \left(\sum_{j=1}^{n} w_{i,j}\right)^2\right]}{n-1}}}$$

where x_j is the attribute value for cell j, $w_{i,j}$ is the spatial weight between cell i and j, n is equal to the total number of cells, and:

$$\bar{X} = \frac{\sum_{j=1}^{n} x_j}{n}$$

$$S = \sqrt{\frac{\sum_{j=1}^{n} x_j^2}{n} - (\bar{X})^2}$$