Project Milestone 2: Introduction to Course Project

A major peer-to-peer taxi cab firm has hired your team to develop and run multiple spatial queries on their large database that contains geographic data as well as real-time location data of their customers. A spatial query is a special type of query supported by geodatabases and spatial databases. The queries differ from traditional SQL queries in that they allow for the use of points, lines, and polygons. The spatial queries also consider the relationship between these geometries. Since the database is large and mostly unstructured, your client wants you to use a popular Big Data software application, SparkSQL. The goal of the project is to extract data from this database that will be used by your client for operational (day-to-day) and strategic level (long term) decisions.

The project has two phases. In each phase, you will be given data and a template code written in SparkSQL. In the first phase, you will write two user-defined functions ‘ST_Contains’ and ‘ST_Within’ in SparkSQL and use them to run the following four spatial queries. Here, a rectangle R represents a geographical boundary in a town or city, and a set of points P represents customers who request taxi cab service using your client firm’s app.

1. Range query: Given a query rectangle R and a set of points P, find all the points within R. You need to use the ‘ST_Contains’ function in this query.

2. Range join query: Given a set of rectangles R and a set of points P, find all (point, rectangle) pairs such that the point is within the rectangle.

3. Distance query: Given a fixed point location P and distance D (in kilometers), find all points that lie within a distance D from P. You need to use the ‘ST_Within’ function in this query.

4. Distance join query: Given two sets of points P1 and P2, and a distance D (in kilometers), find all (p1, p2) pairs such that p1 is within a distance D from p2 (i.e., p1 belongs to P1 and p2 belongs to P2). You need to use the ‘ST_Within’ function in this query.

In the second phase of the project, you will implement two major tasks using te
SparkSQL: ‘hot zone analysis’ and ‘hot cell analysis’. The hot zone analysis us
and a point dataset. For each rectangle, the number of points located within th
obtained. The more points a rectangle contains, the hotter (and more profitable
cell analysis’ applies spatial statistics to spatio-temporal Big Data in order to id
significant hot spots using Apache Spark.

To Get Started

Install Apache Spark and SparkSQL on Computer

You will be using Apache Spark and SparkSQL in this project. Apache Spark is
Data software application. Each team member needs to install Apache Spark a
his/her computer by carefully following the instructions on the page
(https://spark.apache.org/docs/latest/) **(https://spark.apache.org/docs/latest/)**
https://spark.apache.org/docs/latest/ **(https://spark.apache.org/docs/latest**

To get started, team members will need to do some research about Apache Sp
queries.

Required Resource:

(https://www.tutorialspoint.com/spark_sql/spark_sql_quick_guide.htm)
https://www.tutorialspoint.com/spark_sql/spark_sql_quick_guide.htm
(https://www.tutorialspoint.com/spark_sql/spark_sql_quick_guide.htm)

Required Templates

You will be using the following two templates. Keep in mind th assignments are not due until the last two weeks of the course requirement is in the README file of each template:

Project first phase (https://github.com/jiayuasu/CSE512-Project-

Project second phase (https://github.com/jiayuasu/CSE512-Proj Analysis-Template)