Project Phase 1

Due   Apr 10 by 11:59pm   **Points**   0   **Submitting**   a file upload

File Types   zip, tar, tar.gz, gz, tgz, and jar

Available   until Apr 11 at 2:59am

This assignment was locked Apr 11 at 2:59am.

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.

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.

The full instruction is here: **Project Milestone 2: Introduction to Course Project**

The coding template is here: **https://github.com/jiayuasu/CSE512-Project-Phase2-Template (https://github.com/jiayuasu/CSE512-Project-Phase2-Template)**

Submission Guidance:

You will submit two files as follows:

1. One zip file. Please compress your Scala project to a single ZIP file and submit it in this Phase. Note that: you need to make sure your project can compile by entering "sbt assembly" in the terminal.

2. One jar file. The jar file should be able to run using "spark submit".