CSE 511
Individual Contribution Report

Name: Rajasree Chennupati
ASU ID: 1218519276
Group 20

Project Overview:

The main objective of this project is to develop and run multiple spatial queries on large database that contains geographic data as well as real-time location data and help a major taxi cab firm in Newyork in their operational (day-to-day) and strategic level (long term) decisions. Geospatial data or spatial data is information that has a geographic aspect to it. In other words, the records in this type of information set have features like coordinates, an address, city, postal code or zip code, included with them. Analysis of geospatial data is most often a complex problem due to high density i.e., it contains a very large number of data points. To support geodatabases and spatial databases we use spatial queries. Spatial 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.

The project has 2 phases.
Phase 1: Implement two user-defined functions, ST_Contains and ST_Within. Further these functions are called to run queries like Range query, Range join query, Distance query and the Distance join query.

Phase 2: Implement spatial hot spot analysis which is divided into two major parts, hot zone analysis and hot cell analysis.

Reflection:

In Phase 1, I worked specially on ST_Contains method. ST_Contains method takes 2 parameters: a point coordinate and rectangle coordinate and returns a boolean value indicating whether the given point lies within the rectangle. To check whether the given point lies within the rectangle, I checked whether the given point's x coordinate lies within the min and max of the rectangle's x coordinates and point's y coordinate lies within the min and max of the rectangle's y coordinates. I had also tested the function with different inputs and check whether everything was working fine.

In Phase2, I worked on Hot Zone analysis. The hot zone analysis uses a rectangle dataset and a point dataset and runs Range Join query on it. The result is then grouped on rectangle so that count of all the points within each rectangle can be obtained. I created the user defined function

to check if the point lies within the rectangle. I had also tested the function with different inputs and check whether everything was working fine.

In all team meetings, I actively interacted with all team members and gave my inputs for developing the project and suggestions for improving. I helped other team members in setting up environment when they are facing difficulty. I completed all the activities which are assigned to me and tested them thoroughly. I contributed my views and ideas for writing the Systems Documentation report.

Lessons Learned:

- Working in team built up my viable cooperation abilities. I figured out how to comprehend through conversation and clarification and to give and get input on execution
- Always install the compatible versions for the project. Sometimes, if we install a product with some version, it may not be compatible with some other product with different version. So, prior to any installation check the compatible product version for all the products so that everything runs fine.
- Whenever we have a complex task, try to break the task into small tasks. Also, instead of writing everything in a single function try to break into small functions which will allow us to debug it and make any modifications easily.
- Took feedback from someone for the code. This way I can know if there are any shortcomings in the code and provided me an opportunity to improve.
- Listening to everyone's ideas can sometimes give us simple solution. We may sometimes think wrongly about the idea to implement, listening to everyone's idea we can either confirm that our idea of implementing is correct or correct our idea if we are wrong.
- With proper planning and dividing of work in prior, we avoided repetitive tasks and completed the project faster.
- Having short demos after some part of implementation, can help everyone know about the functionality and also confirm whether the implementation is what is desired.
- Sharing any specific outcomes and learnings with the team helped in improving everyone's knowledge.

Assessment/Grading:

I spent almost a month to complete the project step by step with the help of TA and Professor. They both helped us and guided us in each stage as to what we need to improve and what is the right way forward. Initially, it took some time to get as I am new to Scala. But once I became familiar with language and the environment, I wrote the code. Overall by doing this project, I learned the different steps of a software development project. All the above learning learnings will without a doubt help me later on future endeavors. It was important when we begin any task, to get clear on the objectives directly in the beginning and afterward make a plan with milestones. We also typically broke down large tasks into smaller chunks, so that it is easier to distribute tasks and provides direction to start. Detailed planning was very important to ensure

pan important project goes smoothly. I completed my assigned tasks on time to complete before the deadline. Whenever required I consulted my team members for their feedback and suggestions. I think I played a good role as a team member and contributed towards the team by proactively participating and completing work on time.

Future Application:

Some of the skills that I learned in this course are:

- Prior to this I know working SQL queries and this course gave me working experience in spatial Queries using Spark SQL and Scala.
- Advantages and working of distributed database systems over centralized database systems.
- Perform queries (e.g., SQL) and analytics tasks in state-of-the-art database systems
- Apply leading-edge techniques to design distributed and parallel database systems
- Data management in Apache spark and Hadoop and detailed explanation of RDD data processing system in Apache Spark.
- Differentiating among major data models such as relational, spatial, and NoSQL and different NOSQL database systems.
- Performing different operations (map, filter, join) in cluster computing systems such as Hadoop/Spark and how spatial data is managed in Apache Spark.
- Advantages of cloud computing and different types of cloud services and models.
- How to program spatial queries with Scala and obtain the results.
- Introduction to AWS and about different products AWS offers.
- Learned benefits of Apache Spark and how with its in-memory processing feature we can increase the processing speed.

This project gave me a thorough understanding of various distributed frameworks such as Hadoop and Apache Spark. The GeoSpark framework based on Apache Spark integrals gave me a clear picture about processing of spatial data on a massive scale - the New York Yellow Taxi Trip Data. Based on these learnings, I want to explore further for implementing Spatial K nearest neighbors (KNN) query which finds the nearest taxi trip pickup points of New York Time Square. This is what the Uber uses in finding a nearest ride for a person in a location. I also want to investigate the most popular zone for taxi in New York. By knowing which areas have a great demand, the cab firm can have more taxis at that point which will help the cab firm in increasing their profit and also decrease the customer waiting time. This will benefit both the firm and the customers. I found Cloud computing very interesting during this course. This gave me a start and I want to explore more areas in cloud computing. With massive amount of data present everywhere it is advantageous for me to know how this data is managed with cluster computing systems like Hadoop/Spark or cloud computing environment like Amazon AWS. I am sure all these skills add a great advantage to me for my knowledge and also in my workplace.