ARIZONA STATE UNIVERSITY

IRA A. FULTON SCHOOLS OF ENGINEERING — SCHOOL OF COMPUTING AND AUGMENTED INTELLIGENCE

CSE 511 Data Processing at Scale: Individual Contribution Report

Group 22

Harsh Sanjaykumar Nagoriya

Instructor: Zhichao Cao

Tempe, Arizona April 26, 2022

Contents

1 Problem Statement

1.1 Project Phase 1

A leading peer-to-peer taxi cab company wants to use SparkSQL to create and perform many spatial queries on their massive database, which comprises geographic data as well as real-time customer location data. A spatial query is a form of query that geodatabases and spatial databases can support. Points, lines, and polygons can be used in the queries, which makes them different from standard SQL queries. The link between these geometries is also taken into account by the spatial queries. The project's purpose is to extract information from the database that will be used to make operational and strategic decisions.

1.2 Project Phase 2

1.2.1 Hot zone analysis

In this phase, a range join operation on a rectangle dataset and a point dataset is asked. The number of points within a rectangle will be determined for each rectangle. The larger the rectangle, the more points it contains. So this task is to determine the hotness of all the rectangles.

1.2.2 Hot cell analysis

The goal of this phase is to use Apache Spark to apply spatial statistics to spatio-temporal large data in order to discover statistically significant spatial hot spots by calculating the Getis-Ord statistic for NYC Taxi Trip datasets.

2 My roles and contributions

The team members agreed that each of them would try to implement the code on his or her own machine initially. Following that, we used to get together for team meetings and share our approaches. This allowed everyone to learn while also exposing us to a variety of different problems and how to handle them. Now, in the following sections, I'll go over the things that I accomplished explicitly during the project.

- The installation of the essential tools and environment required to run even very rudimentary code was the first big step I had to perform in phase 1. I needed to install Java 1.8, Apache Spark 2.4.7, and Hadoop 2.7 for this. I had to set up the path environment variables after successfully installing these programs. I was able to do these in a proper manner as I followed several tutorials available along with TAs session. My project environment was then ready to utilize.
- The next step was to learn the Scala programming language. I used some internet resources as well as the Scala language's official documentation to do this. I began working on the coding portion of phase 1 after this was completed.
- We mostly had to implement two functions in phase 1: ST_Within and ST_Contains. We divided these responsibilities into two sub-teams, Me and Savan in one and Krushali in another. I was

part of the subteam that was responsible for developing the ST_Conatins function. It was a simple function that examined if a point was inside or outside of a rectangle. If the point was inside the rectangle, True was returned; otherwise, False was returned. So, after writing the pseudocode for this function, I added it to the.scala file.

- The range query and range join query were run after the ST Contains function was implemented. The last half of phase 1 was done by the other members, and then we integrated the code and double-checked it for correctness.
- We had to complete two tasks in phase 2 of the project: Hotzone analysis and Hotcell analysis. We divided this phase into two sub-teams, just like phase 1. Krushali and I were designing the Hotzone analysis function. The hotness of a geographical area was evaluated by counting the number of taxi journeys that started from it in Hotzone Analysis. A heatmap was created using the given data sets of rectangles and taxi trips. The final outcomes were presented in ascending order.
- Savan finished the Hotcell Analysis after the Hotzone Analysis was completed. After then, the two codes were combined.
- We double-checked our codes with the provided testcases in both phases to ensure that no errors were present. We also compared our results to the test results provided. We had to present a project report at the conclusion of each phase, so at that time I worked on the conceptual half of the project, which corresponded to my coding part.

3 Conclusion

Being fascinated by the recent trends in the data technologies, I was able to study many new tools and technologies, such as Spark, Scala, Hadoop, Geospatial queries, and others. These tools and technologies benefited me learning cutting edge data processing technologies that would make a better path towards my career.

Acknowledgement

I'd like to express my gratitude to Savan Doshi and Krushali Shah for their unwavering support, co-operation, and encouragement throughout the project's teamwork journey, which proved crucial to the project's success. Without their support and co-operation this project would not have materialized. I would also like to thank Dr. Zhichao Cao (Instructor), Akkamahadevi Hanni (Teaching Assistant), Rithvik Chokkam (Grad Service Assistant) Arizona State University for their constant encouragement towards the realization of this work.