CSE511- Data Processing at Scale

Project 2 Report

**Name** Yu Chang  
**Date** 10-09-2024

**Project Summary**

The primary goal of this project is to analyze New York City taxi data by completing two main tasks:

1. **Hot Zone Analysis:** Using Apache Spark to calculate the number of taxi pickups within specified rectangular spatial boundaries
2. **Hot Cell Analysis:** Using Apache Spark to process the data and compute the Getis-Ord statistic to identify the top 50 hotspot cells based on the pickup data

**First tasks: Hot Zone Analysis**

After loading and parsing the taxi pickup data and the rectangular boundary coordinates:

1. Use SQL queries to combine the pickup data and the rectangular boundary data with the ST\_Contains function. ST\_Contains determines whether each pickup point falls within the specified rectangular. Only save the points located inside the rectangles for further analysis.
2. Use SQL queries to calculate the number of pickup points and return the result.

**Second tasks: Hot Cell Analysis**

After loading and parsing the taxi pickup data**:**

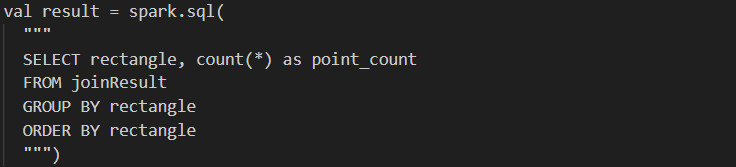
1. Use the CalculateCoordinate function with latitude, longitude, and time data to create space-time cells.
2. Use the SQL queries to calculate the number of pickups in each cell.
3. Use the SQL queries to calculate the mean and standard deviation.
4. Use the SQL queries to calculate neighbor sums and counts with the IsNeighbor function.
5. Use the SQL queries to calculate the Getis-Ord Statistic for each cell to determine the hotspot. Return the top 50 hotspot cells as the results.

**Experience Summary**

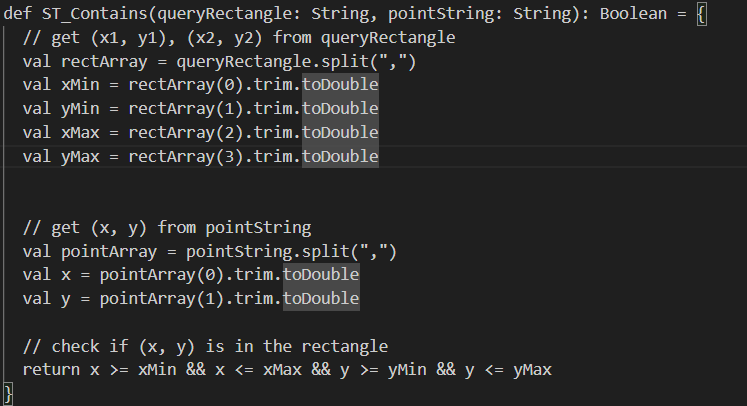
1. **Set up an environment:** I've experimented with various tools and versions to run Apache Spark, including Virtual Box with multiple Linux distributions and versions, Windows, and WSL with Ubuntu.
2. **Data Processing:** I gained experience in using Apache Spark to process datasets, learning how to write SQL queries, and implementing custom user-defined functions in Scala.

**Implementation Details**

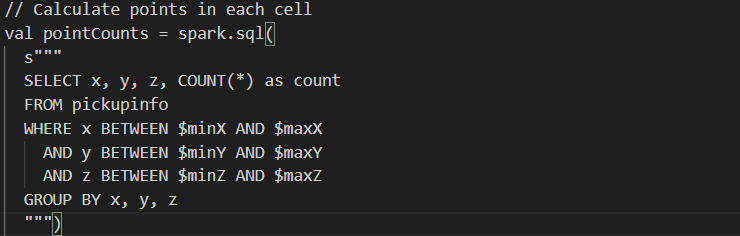
**HotcellAnalysis.scala**

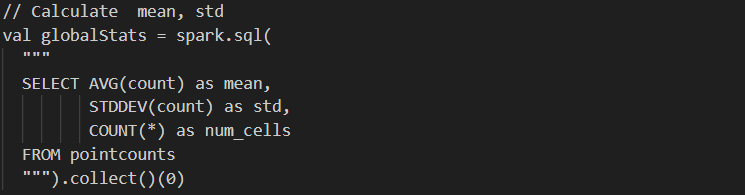
****

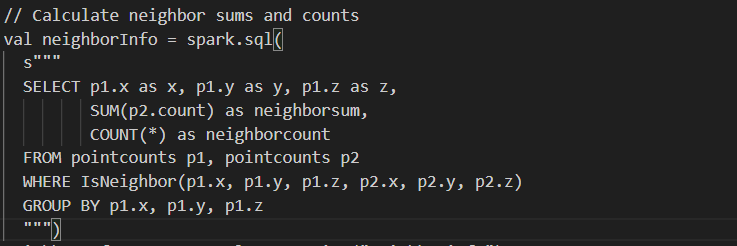
**HotcellUtils.scala**

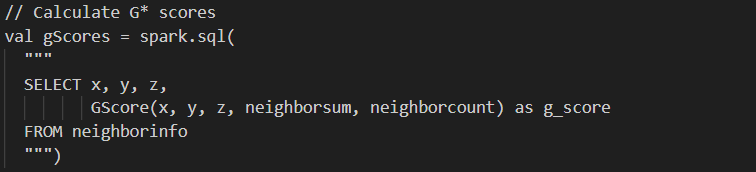
****

**HotzoneAnalysis.scala**

****

****

****

****

**HotzoneUtils.scala**

