Stream Processing 2nd Project

João Martins 52422, Ricardo Ferreira 52915

Departamento de Informática, Faculdade de Ciências e Tecnologia, Universidade
Nova de Lisboa, 2829-516 Caparica, Portugal

Abstract.

1 Introduction

This project aims to present solutions (queries) on a dataset that contains information about taxi rides in New York City. Using Siddhi, which allows creating Apps where we can write Streaming processing and Complex Event Processing rules with Siddhi Streaming SQL, and WSO2 Streaming Processor which allows to integrate Siddhi Apps, we intend to present solutions for the following questions:

- **Q1:** Find the top 10 most frequent routes during the last 30 minutes.
- **Q2:** Identify areas that are currently most profitable for taxi drivers.
- **Q3:** Alert whenever the average idle time of taxis is greater than a given amount of time (say 10 minutes).
- **Q4:** Detect congested areas.
- **Q5:** Select the most pleasant taxi drivers.

To divide the map into an NxN grid, given the longitude and latitude values of the event, we use the following formulas to calculate the grid cell:

$$LONG\_CELL = round((longitude - MIN\_LON)/LON\_DELTA) \quad (1)$$

$$LAT\_CELL = round((MAX\_LAT - latitude)/LAT\_DELTA) \quad (2)$$

where $MIN\_LON = -74.916578$, $MAX\_LAT = 41.47718278$ and LON_DELTA and LAT_DELTA are deltas corresponding to the intended N for the NxN grid.

This report is structured as follows: in section 2 we present the rationale for solving each issue as well as our interpretation of the issue; in section 3, we present in greater detail our solution to one of the questions, as well as the results obtained; and, finally, in section 4, we present our conclusions about the solutions and results obtained.

2 Queries Presentation

In this section we individually present the interpretation and rationale for our solutions.

2.1 Question 2

Question 2 is intended to: *"Identify areas that are currently most profitable for taxi drivers"*.

For this, we start by dividing the map into a 600x600 grid. To identify the most profitable areas, we divided the profit made in that area by the number of empty taxis. For each area we calculate the profit in that area, where we consider the fares plus tips in each area in the last 15 minutes.

To calculate the number of empty taxis, we selected, for each taxi, its highest pickup_datetime and its highest dropoff_datetime, in the last 30 minutes, and for each area we count the taxis that have not had any pickups recently, that is, whose dropoff_datetime is highest than pickup_datetime.

2.2 Question 4

Question 4 is intended to: *"Detect congested areas"*.

For this query, as in query 2 we split the map into a 600x600 grid. To identify the most congested areas, we chose to identify the areas with the longest average travel time. For this, for each area where the pickup occurred, we calculated the average time of trips that started in that area, accessing the values in *trip_time_in_secs*.

3 Conclusion