

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

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 average profit in that area, where we consider the sum of fares plus the sum tips (using the aggregation function *sum*) dividing by the number of trips (using the aggregation function *count*), in each area in the last 15 minutes.

To calculate the number of empty taxis, we selected, for each taxi, its highest *dropoff_datetime*, in the last 30 minutes, and for each area we count the taxis, using the aggregation function *count*.

For the presentation of results, we sorted by average profit, descending, selecting only the first five results.

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*, using the aggregation function *avg*.

For the presentation of results, we sorted by average trip time, descending, selecting only the first five results.

2.3 Question 5

Question 5 is intended to: *"Select the most pleasant taxi drivers"*.

For this query, the solution was quite simple. To select the most pleasant taxi driver, we grouped by taxi driver, adding (using the aggregation function *sum*) the tips received by the taxi driver on all trips that occurred in the time window. Then, sorting by the tip value in a descending way, just select the taxi driver that is at the top, using *"limit 1"*.

3 Detailed Query Presentation

In this section we present in more detail one of the answered questions, presenting the code and the rationale behind the decisions taken. The question we are going to present is Q4: *"Detect congested areas"*.

```
-----
@source(type='kafka',
        topic.list='productions',
        partition.no.list='0',
```

```

        threading.option='single.thread',
        group.id="group",
        bootstrap.servers='kafka:9092',
        @map(type='csv', delimiter=','))
define stream RawInputStream (medallion string, hack_license string,
    pickup_datetime string, dropoff_datetime string,
    trip_time_in_secs long, trip_distance string,
    pickup_longitude double, pickup_latitude double,
    dropoff_longitude double, dropoff_latitude double,
    payment_type string, fare_amount double,
    surcharge string, mta_tax string, tip_amount double,
    tolls_amount double, total_amount double);

define stream GridInputStream (medallion string, pick_lon_grid long,
    pick_lat_grid long, drop_lon_grid long,
    drop_lat_grid long, trip_time_in_secs long);

define stream FilteredGridInputStream (medallion string, pick_lon_grid long,
    pick_lat_grid long, drop_lon_grid long,
    drop_lat_grid long, trip_time_in_secs long);

@sink(type='log')
define stream CongestedAreasStream (pick_lon_grid long,
    pick_lat_grid long, avg_time double);
-----

```

For this question we use 4 streams to process the data, where `RawInputStream` is the stream that takes the data directly from the dataset's csv and `CongestedAreasStream` is the output stream where it gives us the answers of which are the most congested areas.

```

-----
@info(name='query2')
from RawInputStream
select medallion, math:round((pickup_longitude - (-74.916578)) /
(0.005986 / 2)) as pick_lon_grid,
math:round((41.47718278 - pickup_latitude) / (0.004491556 / 2)) as pick_lat_grid,
math:round((dropoff_longitude - (-74.916578)) / (0.005986 / 2)) as drop_lon_grid,
math:round((41.47718278 - dropoff_latitude) / (0.004491556 / 2)) as drop_lat_grid,
trip_time_in_secs
insert into GridInputStream;
-----

```

Using the formulas presented above, we transformed the longitude and latitude coordinates into grid cells, thus selecting the trip pickup and dropoff grid cells and also the medallion and the trip time (*trip_time_in_secs*), inserting the data in the `GridInputStream` stream.

```

-----
@info(name='query3')
from GridInputStream [ pick_lon_grid >= 0 and pick_lon_grid <= 600
    and pick_lat_grid >= 0 and pick_lat_grid <= 600 and

```

```

        drop_lon_grid >= 0 and drop_lon_grid <= 600 and
        drop_lat_grid >= 0 and drop_lat_grid <= 600 ]
select medallion, pick_lon_grid, pick_lat_grid,
        drop_lon_grid, drop_lat_grid, trip_time_in_secs
insert into FilteredGridInputStream;
-----

```

Then we filter only the events whose dropoffs and pickups occur with a latitude cell greater than zero and less than 600 and a longitude cell greater than zero and less than 600, inserting the same attributes of `GridInputStream` in `FilteredGridInputStream`.

```

-----
@info(name='query7')
from FilteredGridInputStream#window.timeBatch(10 sec)
select pick_lon_grid, pick_lat_grid, avg(trip_time_in_secs) as avg_time
       group by pick_lon_grid, pick_lat_grid
       order by avg_time desc
       limit 5
insert into CongestedAreasStream;
-----

```

Finally, in a 10 second tumbling window (`#window.timeBatch(...)` processes a tumbling window) we calculate the average travel time for each area where pickups occurred. In order to select the most congested areas, we order each area by the average travel time in a descending order (*order by avg_time desc*) and select the first five areas (*limit 5*), which represent the five with the highest average travel time and, therefore, are the most congested.

We insert the results into the *CongestedAreasStream*, which is the output stream that prints the results to the console (as it was defined with *sink(type='log')*)

4 Conclusion

conclusion