Big Data Management

Project nr 2: DESB GRAND CHALLENGE 2015

Team members: Anna Maria Tammin, Maria Anett Kaha, Eidi Paas, Pirjo Vainjärv

Provided dataset

The source dataset file was provided by the course organizers in Moodle. Since the sorted_data.csv file is large, the project description advised us to use 1GB of the data. Therefore, we created a smaller file with the first 5 million rows. The code for the file creation is included in the notebook. In the csv file, each row represents one taxi ride information.

NB! ChatGPT was utilized for parts of the project.

Data Transformation

The raw dataset was filtered to remove invalid or malformed records:

- Trips with malformed medallion or hack_license (non-MD5 format);
- Missing timestamps;
- Trips with negative or zero distance;
- Non-positive fares.

Additionally, Grid-based indexing was applied based on the project requirements. The coordinates of pick-up and drop-off points were converted into cell coordinates formatted as X.Y . These coordinates were calculated based on the given starting point and grid size.

Queries

Query 1 - Frequent Routes

A grid system was established, a 500m × 500m grid system covering a 150km × 150km area centered at Barryville, NY (41.474937°N, -74.913585°W). This coarser resolution was chosen specifically for route frequency analysis because:

- It provides sufficient spatial granularity to distinguish meaningful routes while reducing computational complexity
- The larger cell size helps aggregate enough trips to identify statistically significant patterns
- It naturally filters out minor GPS variations that might artificially split what are essentially the same routes

We analyzed frequent taxi routes over 30-minute time windows, as shown below:

Top 10 Rows:

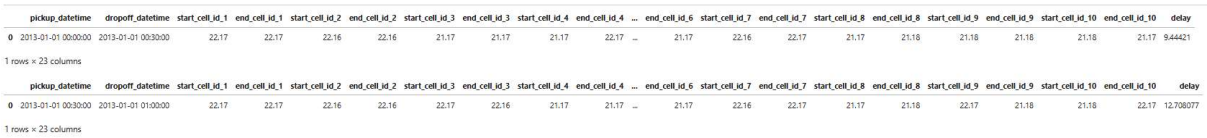
For each 30-minute window, the query identified the 10 most frequent routes, defined by start and end grid cells, and the number of rides between those cells.

Time Window: 2013-01-01 00:00:00 to 2013-01-01 00:30:00

	start_cell	end_cell	Number of Rides
0	22.17	22.17	1138
1	22.16	22.16	730
2	21.17	21.17	584
3	21.17	22.17	433
4	22.17	22.16	422
5	22.17	21.17	355
6	22.16	22.17	294
7	21.17	21.18	212
8	21.18	21.18	169
9	21.18	21.17	166

This result was computed by applying a 30-minute tumbling window to the data stream and aggregating the number of rides for each route (start_cell → end_cell). The most frequent routes show clear patterns of demand, particularly between grid cells 22.17 → 22.17, which represents a major route with 1138 rides in the first time window. As the time progresses, the frequency of rides on these routes fluctuates, which might be indicative of rush-hour patterns or other time-dependent factors.

We updated the query with the delay attribute which captures the time delay between reading the input event that triggered the output and the time when the output is produced. This query is updated only when the at least one of the top routes has changed since the last window.



	pickup_datetime	dropoff_datetime	start_cell_id_1	end_cell_id_1	start_cell_id_2	end_cell_id_2	start_cell_id_3	end_cell_id_3	start_cell_id_4	end_cell_id_4	...	end_cell_id_6	start_cell_id_7	end_cell_id_7	start_cell_id_8	end_cell_id_8	start_cell_id_9	end_cell_id_9	start_cell_id_10	end_cell_id_10	delay
0	2013-01-01 00:00:00	2013-01-01 00:30:00	22.17	22.17	22.16	22.16	21.17	21.17	21.17	22.17	...	21.17	22.16	22.17	21.17	21.18	21.18	21.18	21.18	21.17	9.44421

1 rows × 23 columns

	pickup_datetime	dropoff_datetime	start_cell_id_1	end_cell_id_1	start_cell_id_2	end_cell_id_2	start_cell_id_3	end_cell_id_3	start_cell_id_4	end_cell_id_4	...	end_cell_id_6	start_cell_id_7	end_cell_id_7	start_cell_id_8	end_cell_id_8	start_cell_id_9	end_cell_id_9	start_cell_id_10	end_cell_id_10	delay
0	2013-01-01 00:30:00	2013-01-01 01:00:00	22.17	22.17	22.16	22.16	22.17	22.16	21.17	21.17	...	21.17	22.16	22.17	21.17	21.18	22.17	21.18	21.18	22.17	12.708077

1 rows × 23 columns

Query 2 - Profitable Areas

For profitability calculations, a grid was implemented, a finer 250m × 250m grid covering the same geographic area. This higher-resolution approach was necessary because:

- Profitability metrics require more precise location data to accurately identify high-value zones
- The smaller cell size allows for better discrimination between adjacent areas with different profitability characteristics
- It provides taxi drivers with more actionable information about exactly where to position themselves

The same coordinate transformation logic was used, but with tighter spatial bins to calculate profitability. The profitability metric was defined as the median fare and tip amount divided by the number of available taxis in each cell. The finer grid allowed us to more precisely identify high-profit areas, providing actionable insights for taxi drivers on where to position themselves for maximizing earnings.

The results for one batch can be seen in the table below:

	pickup_datetime	dropoff_datetime	profitable_cell_id	empty_taxis_in_cell	median_profit_in_cell	profitability_of_cell
1661	2013-01-01 03:15:00	2013-01-01 04:00:00	30.40	1	145.00	145.00
4097	2013-01-01 03:45:00	2013-01-01 04:30:00	34.35	1	120.00	120.00
3110	2013-01-01 05:45:00	2013-01-01 06:30:00	54.39	1	110.00	110.00
4047	2013-01-02 11:15:00	2013-01-02 12:00:00	51.33	1	97.20	97.20
843	2013-01-02 01:15:00	2013-01-02 02:00:00	41.33	1	80.00	80.00
3773	2013-01-01 01:15:00	2013-01-01 02:00:00	35.30	1	70.00	70.00
626	2013-01-01 04:45:00	2013-01-01 05:30:00	43.28	1	68.00	68.00
366	2013-01-02 08:15:00	2013-01-02 09:00:00	50.32	3	197.25	65.75
441	2013-01-02 04:15:00	2013-01-02 05:00:00	37.31	1	65.00	65.00
3673	2013-01-02 05:15:00	2013-01-02 06:00:00	50.35	1	65.00	65.00

	pickup_datetime	dropoff_datetime	profitable_cell_id_1	empty_taxis_in_cell_1	median_profit_in_cell_1	profitability_of_cell_1	profitable_cell_id_2	empty_taxis_in_cell_2	median_profit_in_cell_2	profitability_of_cell_2	...	profitab
0	2013-01-01 03:15:00	2013-01-01 04:00:00	30.40	1	145.0	145.0	34.35	1	120.0	120.0	...	

1 rows × 43 columns

_of_cell_2	...	profitability_of_cell_8	profitable_cell_id_9	empty_taxis_in_cell_9	median_profit_in_cell_9	profitability_of_cell_9	profitable_cell_id_10	empty_taxis_in_cell_10	median_profit_in_cell_10	profitability_of_cell_10	delay
120.0	...	66.32	51.37	1	63.46	63.46	41.33	1	63.46	63.46	22.212939