Homework 2 (2023): Taxis data

Due date : 2023-05-24 23:55 (this is a hard deadline)

Lagarde et Michard

2023-04-12

New York taxis trips

This homework is about New York taxi trips. Here is something from Todd Schneider:

> The New York City Taxi & Limousine Commission has released a detailed historical dataset covering over 1 billion individual taxi trips in the city from January 2009 through January 2023. Taken as a whole, the detailed trip-level data is more than just a vast list of taxi pickup and drop off coordinates: it's a story of a City.

How bad is the rush hour traffic from Midtown to JFK? Where does the Bridge and Tunnel crowd hang out on Saturday nights? What time do investment bankers get to work? How has Uber changed the landscape for taxis? How did covid and lockdowns impact traffic and taxis?

The dataset addresses all of these questions and many more.

The NY taxi trips dataset has been plowed by series of distinguished data scientists.

The dataset is available from New York City Government:

> https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page

There is one parquet file for each NY taxi service (`yellow`, `green`, `fhv`) and each calendar month). Each file is moderately large (up to hundreds of megabytes). The full dataset is relatively large if it has to be handled on a laptop (several hundred gigabytes).

You will focus on the `yellow` and `fhv` taxi servicse and a pair of months, from year 2019, 2020, 2021 and 2022. Before those years, For Hire Vehicles services had taken off and carved a huge marketshare. During Spring 2020, NYC was hit by the covid-19 pandemic.

You will download the appropriate `parquet` files (this takes time, but this is routine).

Using `parquet` require decisions about bucketing, partitioning and so on. Such decisions influence performance. It is your call to reengineer the downloaded `parquet` files.

Many people have been working on this dataset, to cite but a few:

- 1 billion trips with a vengeance
- 1 billion trips with R and SQL
- 1 billion trips with redshift
- nyc-taxi
- https://flyte.org/blog/analyzing-covid-19-impact-on-nyc-taxis-with-duckdb
- https://medium.com/ibm-data-ai/analyzing-geospatial-data-in-apache-spark-f638601e405a

You **might** need the following stuff in order to work with GPS coordinates and to plot things easily.

```
!pip install geojson geopandas plotly
```

```
!pip install ipyleaflet
```

For this homework *we let you decide on the tools to use* (except that you should use `Spark`) and to *find out information all by yourself.*

Using data as `parquet` files

1. In the rest of your work, *you will only use the parquet files you created.*

Hint. Don't forget to ask `Spark` to use all the memory and ressources from your computer.

Hint. Don't forget that you should specify a partitioning column and a number of partitions when creating the `parquet` files.

Hint. When working on this, ask yourself and answer to the following questions:

1. What is the number of partitions of the dataframe?
2. Is it possible to tune this number at loading time?
3. Why would we want to modify the number of partitions when (re)-creating the `parquet` files?

Investigate (at least) one month of data in 2019

We shall visualize several features of taxi traffic during one calendar month in 2019 and the same calendar month in 2020 (pick months that correspond to hard lockdowns in 2020).

Hint. In order to build appealing graphics, you may stick to `matplotlib + seaborn`, you can use also `plotly`, which is used a lot to build interactive graphics, but you can use whatever you want (vega-altair).

The following longitudes and lattitudes encompass Newark and JFK airports, Northern Manhattan and Verazzano bridge.

```
long_min = -74.10
long_max = -73.70
lat_min = 40.58
lat_max = 40.90
```

1. Using these boundaries, *filter the 2019 data* (using pickup and dropoff longitude and latitude) and count the number of trips for each value of `passenger_count` and make a plot of that.

Trips with 0 or larger than 7 passengers are pretty rare. We suspect these to be outliers. We need to explore these trips further in order order to understand what might be wrong with them

1. What's special with trips with zero passengers?
2. What's special with trips with more than 6 passengers?
3. What is the largest distance travelled during this month? Is it the first taxi on the moon?
4. Plot the distribution of the `trip_distance` (using an histogram for instance) during year 2019. Focus on trips with non-zero trip distance and trip distance less than 30 miles.

Let's look at what Spark does for these computations

1. Use the `explain` method or have a look at the Spark UI to analyze the job. You should be able to assess
   - Parsed Logical Plan
   - Analyzed Logical Plan
   - Optimized Logical Plan
   - Physical Plan
2. Do the Analyzed Logical Plan and Optimized Logical Plan differ? Spot the differences if any. How would a RDBMS proceed with such a query?
3. How does the physical plan differ from the Optimized Logical Plan? What are the keywords you would not expects in a RDBMS? What is their meaning?

4. Inspect the stages on Spark UI. How many *stages* are necessary to complete the Spark job? What are the roles of `HashAggregate` and `Exchange hashpartitioning`?
5. Does the physical plan perform `shuffle` operations? If yes how many?
6. What are tasks with respect to stages (in Spark language)? How many tasks are your stages made of?

Now, compute the following and make relevant plots:

1. Break down the trip distance distribution for each day of week
2. Count the number of distinct pickup location
3. Compute and display tips and profits as a function of the pickup location

Investigate one month of trips data in 2019, 2020, 2021, 2022

Consider one month of trips data from `yellow` and `fhv` taxis for each year

1. Filter and cache/persist the result

Assessing seasonalities and looking at time series

Compute and plot the following time series indexed by day of the week and hour of day:

1. The number of pickups
2. The average fare
3. The average trip duration
4. Plot the average of ongoing trips

Rides to the airports

In order to find the longitude and lattitude of JFK and Newark airport as well as the longitude and magnitudes of Manhattan, you can use a service like geojson.io.

Plot the following time series, indexed by the day of the week and hour of the day

1. Median duration of taxi trip leaving Midtown (Southern Manhattan) headed for JFK Airport
2. Median taxi duration of trip leaving from JFK Airport to Midtown (Southern Manhattan)

Geographic information

For this, you will need to find tools to display maps and to build choropleth maps. We let you look and find relevant tools to do this.

1. Build a heatmap where color is a function of
    1. number of `pickups`
    2. number of `dropoffs`
    3. number of `pickups` with dropoff at some airport (JFK, LaGuardia, Newark)
2. Build a choropeth map where color is a function of
    1. number of pickups in the area
    2. ratio of number of payments by card/number of cash payments for pickups in the area
    3. ratio of total fare/trip duration for dropoff in the area
3. Build an interactive chorophet with a slider allowing the user to select an `hour of day` and where the color is a function of
    1. average number of dropoffs in the area during that hour the day
    2. average ratio of tip over total fare amount for pickups in the area at given hour of the day
4. Spot traffic imbalances. For each day and each hour, and each zone, compute the number of trips arriving and leaving the zone, compute the ratio between the two quantities and build a choropleth spotting possible imbalances.

Covid impact

Give a picture of the traffic changes induced by the pandemics.