Université Claude Bernard Lyon 1

Data Processing and Analytics

SPARK LAB REPORT

Realized By :
Mr.Saad BELAOUAD
Mr.Adnan El Mouttaki
Mr.Anouar ZAHRAN

Under the supervision of :
Pr.Mauri Andrea

23 octobre 2023

Taxi Trips Analysis in New York City

Objective:

To analyze taxi trips data in New York City to understand taxi operations across different boroughs and derive insights that can be used for optimizing taxi operations.

1. Introduction:

The taxi industry is a vital part of New York City's transportation system. Understanding the patterns of taxi trips can provide insights into the city's mobility needs and help in optimizing taxi operations. This project aims to analyze taxi trips data to understand the patterns of trips within and across boroughs and the time taken by taxis between consecutive trips.

2. Dataset Overview:

The dataset provides details about individual taxi rides in New York City. Each record captures:

- Unique ID for the car (license)
- Pick-up location (longitude and latitude)
- Pick-up time
- Drop-off location (longitude and latitude)
- Drop-off time
- To enrich the dataset with borough information, a separate GeoJSON file containing the geographical boundaries of NYC boroughs was used.

3. Data Preparation:

3.1. Data Enrichment:

Using the Shapely library, the dataset was enriched by:

Converting the longitude and latitude of pick-up and drop-off locations into Point objects.
Associating each Point with a borough by checking its inclusion within the borough's polygon.

3.2. Data Cleansing:

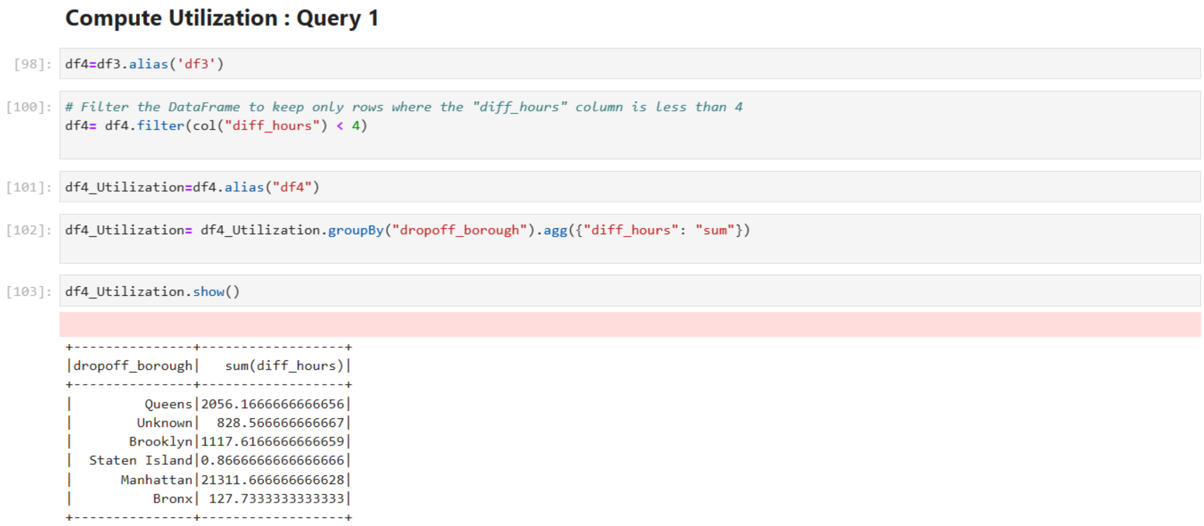
Records where the drop-off time was before the pick-up time were removed.

Trips with durations exceeding 4 hours were considered outliers and were excluded.

4. Analysis:

4.1. Taxi Utilization:

- Objective:
  Compute the utilization metric for taxis, which represents the fraction of time a taxi is occupied.
- Methodology:
  - Calculate the difference between consecutive trips for each taxi.
  - Filter out differences greater than 4 hours, considering them as new sessions.
  - Group by drop-off borough and sum the idle time.
- Implementation:



Compute Utilization : Query 1

```
[98]: df4=df3.alias('df3')

[100]: # Filter the DataFrame to keep only rows where the "diff_hours" column is less than 4
df4= df4.filter(col("diff_hours") < 4)

[101]: df4_Utilization=df4.alias("df4")

[102]: df4_Utilization= df4_Utilization.groupBy("dropoff_borough").agg({"diff_hours": "sum"})

[103]: df4_Utilization.show()
```

```
+---------------+-----------------+
|dropoff_borough|  sum(diff_hours)|
+---------------+-----------------+
|         Queens|2056.1666666666656|
|        Unknown|  828.566666666667|
|       Brooklyn|1117.6166666666659|
|  Staten Island|0.8666666666666666|
|      Manhattan|21311.666666666628|
|          Bronx| 127.7333333333333|
+---------------+-----------------+
```

- Results:
  Manhattan had the highest utilization, indicating that taxis in this borough have the least idle time. This suggests a high demand for taxis in Manhattan.

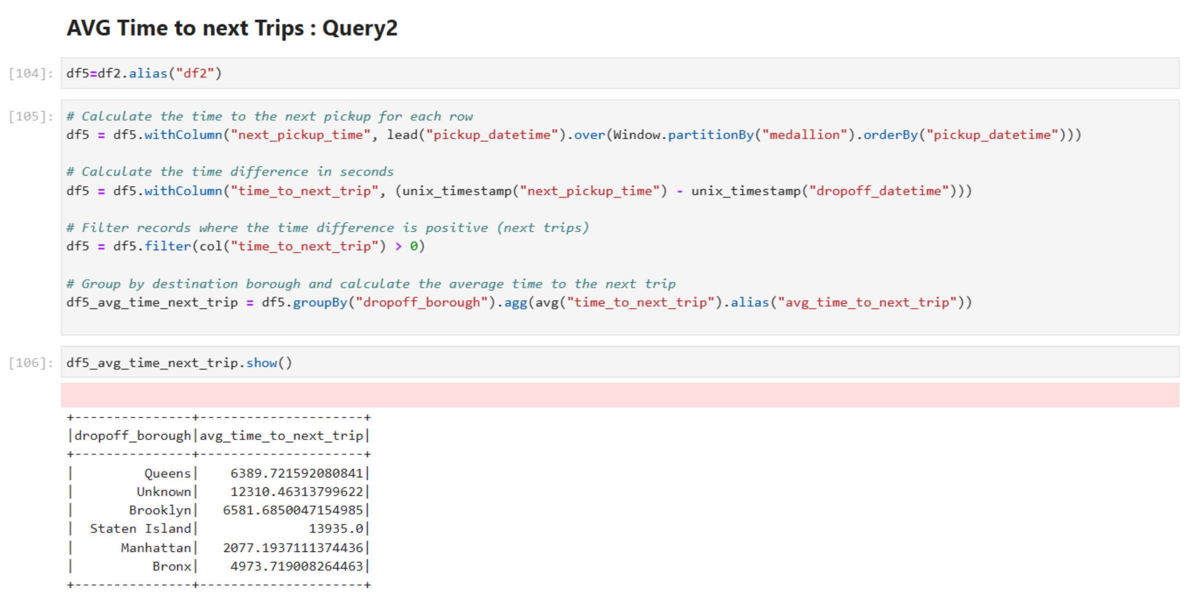
4.2. Average Time to Next Trip:

- Objective:

Determine the average time it takes for a taxi to find its next fare based on the drop-off borough.

- Methodology:
  - Calculate the time difference between the drop-off of one trip and the pick-up of the subsequent trip for the same taxi.

    - Group by drop-off borough and compute the average time difference.
- Implementation:



AVG Time to next Trips : Query2

```
[104]: df5=df2.alias("df2")
```

```
[105]: # Calculate the time to the next pickup for each row
df5 = df5.withColumn("next_pickup_time", lead("pickup_datetime").over(Window.partitionBy("medallion").orderBy("pickup_datetime")))

# Calculate the time difference in seconds
df5 = df5.withColumn("time_to_next_trip", (unix_timestamp("next_pickup_time") - unix_timestamp("dropoff_datetime")))

# Filter records where the time difference is positive (next trips)
df5 = df5.filter(col("time_to_next_trip") > 0)

# Group by destination borough and calculate the average time to the next trip
df5_avg_time_next_trip = df5.groupBy("dropoff_borough").agg(avg("time_to_next_trip").alias("avg_time_to_next_trip"))
```

```
[106]: df5_avg_time_next_trip.show()
```

```
+---------------+--------------------+
|dropoff_borough|avg_time_to_next_trip|
+---------------+--------------------+
|         Queens|   6389.721592080841|
|        Unknown|   12310.46313799622|
|       Brooklyn|  6581.6850047154985|
|  Staten Island|             13935.0|
|      Manhattan|  2077.1937111374436|
|          Bronx|   4973.719008264463|
+---------------+--------------------+
```

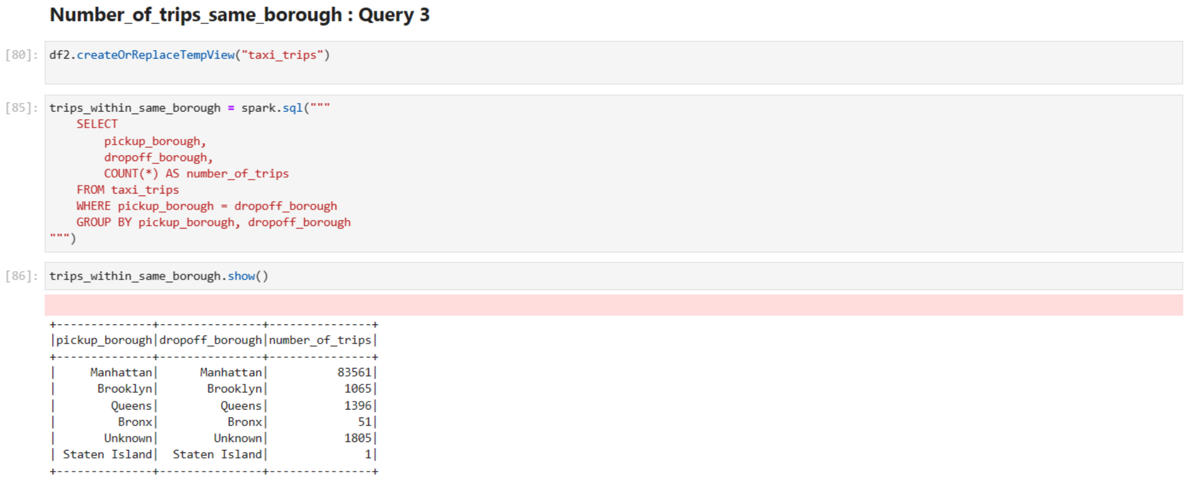
- Results:

Taxis dropping off in Manhattan have the shortest average time to their next trip, indicating a high demand and quick turnaround in this borough.

4.3. Same-Borough Trips:

- Objective:

  Identify the number of trips that both started and ended within the same borough.
- Methodology:
    - Filter records where the pick-up and drop-off boroughs are the same.
    - Group by pick-up borough and count the number of such trips.
- Implementation:



Number_of_trips_same_borough : Query 3

```
[80]: df2.createOrReplaceTempView("taxi_trips")
```

```
[85]: trips_within_same_borough = spark.sql("""
          SELECT
              pickup_borough,
              dropoff_borough,
              COUNT(*) AS number_of_trips
          FROM taxi_trips
          WHERE pickup_borough = dropoff_borough
          GROUP BY pickup_borough, dropoff_borough
      """)
```

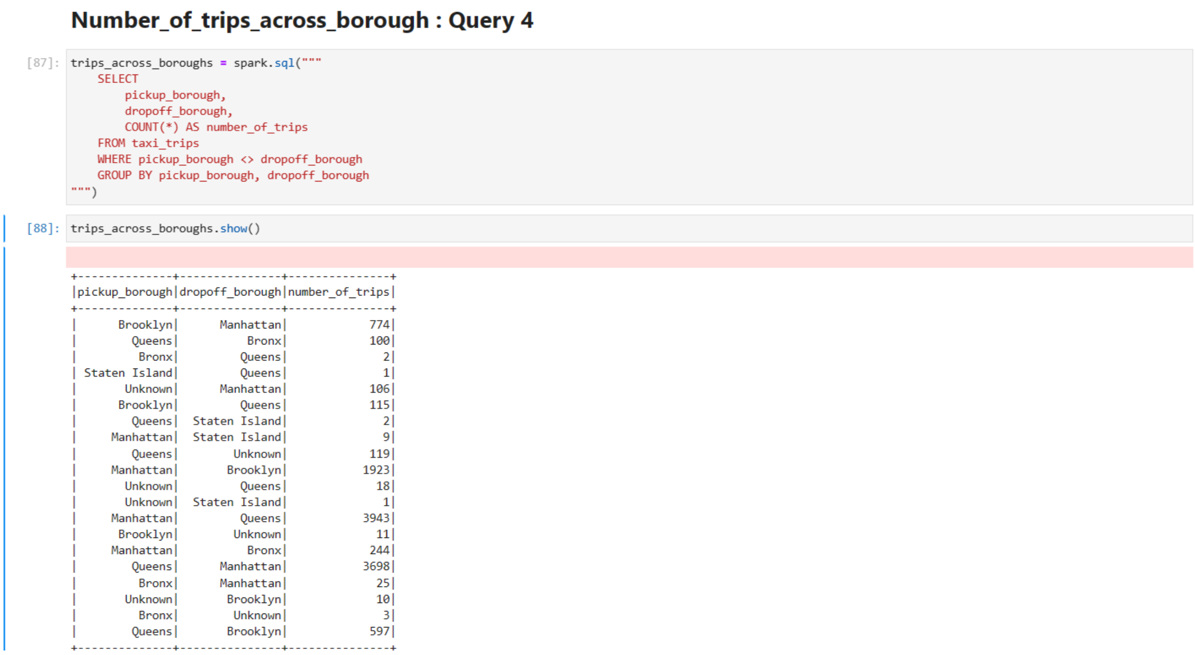
```
[86]: trips_within_same_borough.show()
```

```
+--------------+---------------+---------------+
|pickup_borough|dropoff_borough|number_of_trips|
+--------------+---------------+---------------+
|     Manhattan|      Manhattan|          83561|
|      Brooklyn|       Brooklyn|           1065|
|        Queens|         Queens|           1396|
|         Bronx|          Bronx|             51|
|       Unknown|        Unknown|           1805|
| Staten Island|  Staten Island|              1|
+--------------+---------------+---------------+
```

- <u>Results:</u>
  Manhattan: 83,561 trips
  Queens: 1,396 trips
  Brooklyn: 1,065 trips
  Bronx: 51 trips
  Staten Island: 1 trip
  Unknown: 1,805 trips
  Manhattan had the highest number of intra-borough trips, suggesting that most passengers prefer short-distance travels within the borough.

4.4. Across-Borough Trips:

- <u>Objective:</u>
  Determine the number of trips that started in one borough and ended in another.
- <u>Methodology:</u>
  - Filter records where the pick-up and drop-off boroughs are different.
  - Group by pick-up and drop-off boroughs and count the number of such trips.
- <u>Implementation:</u>



Number_of_trips_across_borough : Query 4

```
[87]: trips_across_boroughs = spark.sql("""
          SELECT
              pickup_borough,
              dropoff_borough,
              COUNT(*) AS number_of_trips
          FROM taxi_trips
          WHERE pickup_borough <> dropoff_borough
          GROUP BY pickup_borough, dropoff_borough
      """)
```

```
[88]: trips_across_boroughs.show()
```

```
+--------------+---------------+---------------+
|pickup_borough|dropoff_borough|number_of_trips|
+--------------+---------------+---------------+
|      Brooklyn|      Manhattan|            774|
|        Queens|          Bronx|            100|
|         Bronx|         Queens|              2|
| Staten Island|         Queens|              1|
|       Unknown|      Manhattan|            106|
|      Brooklyn|         Queens|            115|
|        Queens|  Staten Island|              2|
|     Manhattan|  Staten Island|              9|
|        Queens|        Unknown|            119|
|     Manhattan|       Brooklyn|           1923|
|       Unknown|         Queens|             18|
|       Unknown|  Staten Island|              1|
|     Manhattan|         Queens|           3943|
|      Brooklyn|        Unknown|             11|
|     Manhattan|          Bronx|            244|
|        Queens|      Manhattan|           3698|
|         Bronx|      Manhattan|             25|
|       Unknown|       Brooklyn|             10|
|         Bronx|        Unknown|              3|
|        Queens|       Brooklyn|            597|
+--------------+---------------+---------------+
```

- <u>Results:</u>
  The majority of inter-borough trips were between Manhattan and Brooklyn, indicating a strong connection between these two boroughs.

5. Optimizations:

5.1. Broadcasting:

Given the relatively small size of the GeoJSON data, it was broadcasted to different workers to speed up the enrichment process.

5.2. Data Ordering:

The GeoJSON data was sorted by polygon size in descending order. This ensured faster access since boroughs with larger areas are more frequent in the taxi ride dataset.

5.3. Windowing:

Spark's Window operator was used to efficiently calculate time differences between consecutive trips for the same taxi.

6. Conclusion:

The analysis of the NYC taxi dataset provided valuable insights into taxi operations and passenger preferences. The high utilization rates in Manhattan indicate the efficiency of taxi operations in the borough. The data also highlighted the importance of intra-borough trips, especially within Manhattan, and the strong connection between Manhattan and Brooklyn in terms of inter-borough trips. These insights can guide taxi operators in optimizing their operations and can also assist city planners in understanding mobility patterns.