Big Data Management
project-1

Students: Rene Puusepp, Swagata Datta, Shailaja Mahara, Muhammad Sohaib Anwar

Introduction

This report presents an analysis of the New York City Taxi dataset. The primary objectives of this project include:

- Measuring taxi utilization by computing idle time per taxi.
- Calculating the average time for a taxi to find its next fare per destination borough.
- Counting the number of trips that start and end within the same borough.
- Counting the number of trips that start in one borough and end in another.
- Applying optimizations to improve computational efficiency.

The dataset was preprocessed and optimized for performance, ensuring that computations were done efficiently despite the large data volume (approximately 28GB).

---

2. Query Results

Query 1: Utilization Per Taxi/Driver

Utilization was computed by determining the idle time per taxi and dividing the occupied time by the total operational time (occupied + idle time). The results show the utilization ratio per taxi.

Medallion (Taxi ID)	Total Occupied Time (sec)	Total Idle Time (sec)	Utilization (%)
0038EF45118925A510975FD0CCD67192	780,480	1,509,360	34.08%
00BD5D1AD3A96C997E49E0453A6C5DF1	810,120	1,080,300	42.85%
01A2F4366180AEB433600BAEA196BFC7	990,364	1,212,008	44.97%
01D13A056D9A26F84C328DFDD5534B55	629,460	826,080	43.25%
01F24976B8E3FF46A08187C86F1F9AB7	375,011	227,511	62.24%
02063AF23344CEA458E992EC448C5E73	638,880	929,160	40.74%
024E99A049B748C443A541B2F6F55E5F	343,440	429,240	44.45%
025B4E80E8A06FDB0FC0A05E319B0E60	829,481	1,133,337	42.26%
026B27179DE85CFDC57E5D97372C63F7	406,956	474,964	46.14%
02B196981B24858BCD38C205AA81D7D8	552,780	962,040	36.49%

Query 2: Average Time to Find Next Fare Per Destination Borough

The average time it takes for a taxi to find its next fare was computed by measuring the time difference between drop-off and the subsequent trip's pickup time, per borough.

Dropoff Borough	Avg Time to Next Fare (seconds)
Queens	2807.45 sec
Unknown	1235.57 sec
Brooklyn	1650.69 sec
Staten Island	3016.09 sec
Manhattan	728.07 sec
Bronx	2335.05 sec

As expected, taxis in Manhattan have the shortest wait time before their next fare, whereas Staten Island has the longest.

Query 3: Number of Trips Starting & Ending in the Same Borough

- **Count**: 13,108,532 trips

Query 4: Number of Trips Starting in One Borough & Ending in Another

- **Count**: 1,667,722 trips

---

3. Optimizations Applied

Given the large dataset size, several optimizations were implemented to enhance performance:

- **Column Reduction:** Removed unnecessary columns such as vendor_id, rate_code, passenger_count, and others to reduce processing time.
- **Parquet Format:** Converted CSV files to **Parquet** for optimized query execution.
- **Repartitioning:** Used repartition(50) before processing to balance data across workers and avoid skew.
- **Sampling Instead of Limit:** Used .sample(fraction=0.02) instead of .limit(500000) to quickly extract a subset without scanning all partitions.
- **Broadcasting:** GeoJSON borough data was **broadcasted** across executors to avoid repeated reads.
- **Indexing & Sorting:** Sorted borough polygons by size for **faster lookup** in geospatial queries.
- **Filtering Outliers:** Removed trips where duration was negative or exceeded 4 hours.

---

4. Conclusion

The analysis provided key insights into NYC taxi operations, revealing:

- **Utilization varies significantly per taxi**, with the highest utilization at ~62% and the lowest around 34%.

- **Manhattan taxis find their next fare the fastest (~728 sec)**, while **Staten Island taxis take the longest (~3016 sec)**.
- **Most trips occur within the same borough (~13.1 million trips)**, while **around 1.67 million involve cross-borough travel**.

These findings could be useful for optimizing taxi dispatching strategies and improving urban mobility planning.