Project Title: Real-Time Analytics of NYC Taxi Trip Data using Apache Spark

Query 0: Data Cleansing

The initial dataset was filtered to remove invalid or malformed data. Cleaning steps included:

- Dropping trips with missing or out-of-bound GPS coordinates.
- Filtering trips with non-positive duration or distance.
- Ensuring non-null values for key identifiers (medallion, hack_license).

Sample of Cleaned Data:

Query 0 - Sample of Cleaned Data:

+--------------------+--------------------+-------------------+-------------------+-----------------+-----------+---------------+--------------+----------------+---------------+------------+-----------+---------+-----------+----------+------------+

| medallion| hack_license| pickup_datetime| dropoff_datetime|trip_time_in_secs|trip_distance|pickup_longitude|pickup_latitude|dropoff_longitude|dropoff_latitude|payment_type|fare_amount|surcharge|mta_tax|tip_amount|tolls_amount|

+--------------------+--------------------+-------------------+-------------------+-----------------+-----------+---------------+--------------+----------------+---------------+------------+-----------+---------+-----------+----------+------------+

|945F1F65FAA293DA1...|D7421B620BD448E6B...|2013-02-17 16:13:00|2013-02-17 16:23:00| 600| 1.31| -73.974167| 40.753681| -73.990501| 40.751266| CRD| 7.5| 0.0| 0.5| 1.5| 0.0|

|94E10E3A3763877CE...|826FF187797D81521...|2013-02-17 16:01:00|2013-02-17 16:23:00| 1320| 3.67| -73.972206| 40.754059| -73.986916| 40.750572| CRD| 17.5| 0.0| 0.5| 3.5| 0.0|

|9B6A7942D02E1977A...|3E2D2C56FFAFFCAA7...|2013-02-17 15:59:00|2013-02-17 16:23:00| 1440| 4.17| -73.966209| 40.770863| -74.005684| 40.72559| CSH| 18.0| 0.0| 0.5| 0.0| 0.0|

|A1627FA9AB9437855...|4BB61985C755BF1BC...|2013-02-17 16:09:00|2013-02-17 16:23:00| 840| 2.6| -73.970741| 40.788445| -73.986275| 40.756046| CRD| 11.5| 0.0| 0.5| 2.3| 0.0|

|46BC499CF11522E49...|11BAF36F141401322...|2013-02-17 16:21:03|2013-02-17 16:23:11| 127| 0.4| -73.979561| 40.752693| -73.972618| 40.762199| CSH| 3.5| 0.0| 0.5| 0.0| 0.0|

+--------------------+--------------------+-------------------+-------------------+-----------------+-----------------+---------------+--------------+-----------+-----------+---------+-----------+------------+

only showing top 5 rows

Grid Cell Mapping

Grid cells were created for two resolutions:

- **500m x 500m** (Query 1)
- **250m x 250m** (Query 2)

Grid Cell Mapping Sample:

Grid Cell Sample:

+---------------+---------------+---------------+----------------+

|pickup_grid_500|dropoff_grid_500|pickup_grid_250|dropoff_grid_250|

+---------------+---------------+---------------+----------------+

| -160.157| -160.154| -320.313| -321.308|

| -160.157| -160.155| -320.314| -321.309|

| -156.158| -166.152| -312.316| -332.303|

| -152.157| -159.155| -304.314| -319.309|

| -160.156| -158.157| -320.311| -316.314|

+---------------+---------------+---------------+----------------+

only showing top 5 rows

Query 1: Frequent Routes

Part 1

Top 10 most frequent routes (start_cell, end_cell) within the last 30 minutes were identified:

```
+----------+--------+----------+
|start_cell|end_cell|ride_count|
+----------+--------+----------+
| -156.160|-159.154|         2|
| -154.155|-153.157|         2|
| -153.160|-157.158|         2|
| -149.158|-156.155|         1|
| -157.157|-153.156|         1|
| -170.151|-157.160|         1|
| -149.158|-148.159|         1|
| -156.175|-165.153|         1|
| -154.161|-162.156|         1|
| -157.159|-157.157|         1|
+----------+--------+----------+
only showing top 10 rows
```

Part 2

The query was extended to output a new result only when the top-10 changed.

```
+-------------------+-------------------+----------+--------+---------+
|   pickup_datetime|   dropoff_datetime|start_cell|end_cell|    delay|
+-------------------+-------------------+----------+--------+---------+
|2013-01-01 02:00:00|2013-01-01 02:30:00|  -152.157|-150.157|386371750|
```

|2013-01-01 02:00:00|2013-01-01 02:30:00| -157.158|-164.153|386371750|

|2013-01-01 02:00:00|2013-01-01 02:30:00| -144.161|-145.161|386371750|

|2013-01-01 02:00:00|2013-01-01 02:30:00| -151.157|-134.165|386371750|

|2013-01-01 02:00:00|2013-01-01 02:30:00| -151.160|-159.157|386371750|

|2013-01-01 02:00:00|2013-01-01 02:30:00| -155.161|-156.161|386371750|

|2013-01-01 02:00:00|2013-01-01 02:30:00| -156.154|-165.154|386371750|

|2013-01-01 02:00:00|2013-01-01 02:30:00| -156.158|-150.159|386371750|

|2013-01-01 02:00:00|2013-01-01 02:30:00| -156.158|-167.159|386371750|

|2013-01-01 02:00:00|2013-01-01 02:30:00| -157.155|-159.158|386371750|

+-------------------+-------------------+----------+--------+---------+

only showing top 10 rows

Query 2: Profitable Areas

Part 1

The profitability of a grid cell was calculated as:

profitability = median(fare + tip) / empty_taxis

Where:

- Trips were filtered to those ending in the last 15 minutes.
- Empty taxis were those with a recent dropoff and no subsequent pickup.

Top 10 Profitable Areas (Static):

Query 2 Part 1 - Top 10 Profitable Areas:

+--------+-----------+-------------+-------------+----+

| cell_id|empty_taxis|median_profit|profitability|rank|

+--------+-----------+-------------+-------------+----+

|-128.590| 0| 504.77| 504.77| 1|

|-386.228| 0| 300.0| 300.0| 2|

|-256.372| 0| 296.4| 296.4| 3|

|-173.480| 1| 288.0| 288.0| 4|

|-365.113| 1| 288.0| 288.0| 4|

| -387.75| 0| 273.0| 273.0| 6|

|-157.512| 0| 270.0| 270.0| 7|

|-332.283| 1| 260.0| 260.0| 8|

|-307.454| 0| 242.0| 242.0| 9|

|-186.456| 0| 241.2| 241.2| 10|

+--------+-----------+-------------+-------------+----+

Conclusion

All parts of Query 0, Query 1, and Query 2 Part 1 have been successfully implemented using PySpark. The dataset was streamed by chunking it into small files and processed using Spark Structured Streaming.

The system demonstrates real-time analytical capabilities over high-volume geospatial taxi trip data.