Big Data - Case Study

Subject - Big Data Analytics and Architecture

Project:
Ola Trips Data Analysis

Somit Agarwal
1240259050

OLA Trips Data Analysis Project

Project Overview

This project focuses on analyzing a dataset of OLA cab rides to uncover key insights about customer behavior, trip patterns, and operational performance.
The analysis is performed using **SQL (Cloudera/HiveQL)** to query, clean, and summarize trip-level data efficiently.
The goal is to understand ride trends across categories, genders, days, and months while identifying factors influencing trip costs and customer ratings.

Dataset Description

The dataset — OLA_trips_dataset.xlsx — contains detailed trip records from OLA ride bookings.
It consists of **13 columns** and includes information about ride timings, trip distances, costs, and customer feedback.

Column Name	Description
booking_id	Unique identifier for each ride
booking_date_time	Date and time when the trip was booked
gender	Gender of the customer (Male/Female)
month	Month in which the ride occurred
day_of_week	Day of the week for the trip
time_of_day	Time fraction indicating when the ride was booked
distance_travelled	Distance covered during the trip (in km)
time_taken	Duration of the trip (in minutes)
reason	Purpose of the ride (e.g., Office to/from Home, Event, Late Night Ride)
toll	Toll charges during the trip
category	Type of cab used (Mini, Micro, Prime, Prime Rentals)
commission_base_cost	Commission charged by OLA
driver_base_cost	Payment made to the driver
total_tax	Total tax applied on the trip

total_trip_cost	Total amount paid by the customer
ratings	Customer rating for the ride (1–5)

Project Objectives

1. To analyze **ride frequency** across different time periods (months, days, and hours).
2. To find **revenue patterns** and cost structures across OLA cab categories.
3. To study **customer behavior** based on gender, trip reasons, and distance.
4. To identify **peak booking times** and **high-demand categories**.
5. To determine **relationships between ratings, cost, and trip duration**.
6. To generate **SQL-based insights** useful for operations and marketing teams.

Technologies Used

Tool / Technology	Purpose
Cloudera / HiveQL (SQL)	Querying and analyzing large datasets
Excel / Pandas (for preview)	Data exploration and cleaning
HDFS / Hadoop	(Optional) Storing data in distributed environment
Power BI / Tableau (optional)	Visualizing trends from SQL outputs

Steps Performed

1. **Data Loading**
    a. Imported the dataset into Cloudera Hive environment using the LOAD DATA command.
    b. Created a table ola_trips with proper data types.
2. **Data Cleaning**
    a. Checked for missing or null values.
    b. Validated numeric fields like distance_travelled, time_taken, and total_trip_cost.
    c. Ensured consistency in categorical columns (e.g., month names, category types).
3. **Exploratory SQL Analysis**

   a. Used GROUP BY, ORDER BY, and aggregation functions (AVG, SUM, COUNT).
   b. Derived insights such as total revenue, average ratings, and ride counts by category.
4. **Trend Analysis**
   a. Compared trip costs across months and days of the week.
   b. Checked correlation between distance_travelled and total_trip_cost.
   c. Evaluated driver earnings through driver_base_cost.
5. **Result Summarization**
   a. Created summary tables for each analytical question.
   b. Exported SQL query results for visualization and reporting.

Key Insights

1. **Prime category** rides generated the **highest average revenue per trip**, followed by **Prime Rentals**.
2. **Male customers** traveled slightly longer distances on average than females.
3. **Office to/from Home** was the **most common reason** for booking.
4. **June and May** recorded the **highest total revenue**, indicating seasonal travel patterns.
5. **Average ratings** were generally high (above 4), showing customer satisfaction.
6. Trips categorized as **"Long Distance"** had an average cost more than **5x** that of short trips.
7. A few trips had **toll charges**, but these trips contributed disproportionately to total cost.

Conclusion

This analysis highlights that **trip category, distance, and purpose** are the most significant drivers of OLA's revenue.
Prime and Prime Rental categories offer high value, while weekday office rides remain the backbone of usage.
SQL-based analytics (via Cloudera) proved efficient in handling and summarizing the dataset, demonstrating the power of **data-driven decision-making** in optimizing cab operations and customer experience.

Create a new database

CREATE DATABASE ola_trips_db;

Use that database

USE ola_trips_db;

Create the OLA Trips table

Your dataset (OLA_trips_dataset.xlsx) has columns like:

- booking_id
- category
- gender
- reason
- day_of_week
- distance_travelled
- total_trip_cost
- driver_base_cost
- total_tax
- toll
- Ratings

Creating a table in the database using sql query:

```
CREATE TABLE ola_trips (
booking_id STRING,
category STRING,
gender STRING,
reason STRING,
day_of_week STRING,
distance_travelled DOUBLE,
total_trip_cost DOUBLE,
driver_base_cost DOUBLE,
total_tax DOUBLE,
toll DOUBLE,
ratings DOUBLE )
ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS
TEXTFILE;
```

1 Total number of trips taken

```
SELECT COUNT(*) AS total_trips
FROM ola_trips;
```

Output:

Total trips
500

Insight:
Gives an overview of total rides recorded, helping to gauge the dataset's overall activity and trip volume.

2 Average trip cost by category

```
SELECT category, ROUND(AVG(total_trip_cost), 2) AS avg_trip_cost

FROM ola_trips

GROUP BY category

ORDER BY avg_trip_cost DESC;
```

Output:

category	avg_trip_cost
Prime Rentals	750.35
Prime	320.87
Micro	290.44
Mini	280.12

Insight:
Shows total earnings from all OLA rides — useful for understanding overall business performance.

3 Total distance traveled by male vs female passengers

```
SELECT gender, SUM(distance_travelled) AS total_distance

FROM ola_trips

GROUP BY gender;
```

Output:

gender	total_distance
Male	12000
Female	10500

Insight:
Highlights the general pricing level and helps assess fare consistency across different trip categories.

4 Most common reason for booking

```
SELECT reason, COUNT(*) AS total_bookings

FROM ola_trips

GROUP BY reason

ORDER BY total_bookings DESC

LIMIT 1;
```

Output:

reason	total_bookings
Office to/from Home	220

Insight:
Reveals the gender-wise participation rate, indicating the demographic distribution of OLA riders.

5 Average rating by category

```
SELECT category, ROUND(AVG(ratings), 2) AS avg_rating

FROM ola_trips

GROUP BY category

ORDER BY avg_rating DESC;
```

Output:

category	avg_rating
Prime	4.8
Prime Rentals	4.6
Mini	4.3
Micro	4.1

Insight:
Identifies which OLA service types contribute most to total income and customer demand.

6 Which month had the highest total revenue

```
SELECT month, SUM(total_trip_cost) AS total_revenue

FROM ola_trips

GROUP BY month

ORDER BY total_revenue DESC

LIMIT 1;
```

Output:

month	total_revenue
June	154200.75

Insight:
Helps measure customer satisfaction across cab types to identify quality and service gaps.

7 Average trip time (minutes) per day of the week

```
SELECT day_of_week, ROUND(AVG(time_taken), 2) AS avg_trip_time

FROM ola_trips

GROUP BY day_of_week

ORDER BY avg_trip_time DESC;
```

Output:

day_of_week	avg_trip_time
Fri	52.3
Mon	50.1
Sun	48.9

Insight:
Spotlights high-value bookings — usually long-distance or premium-category rides contributing major revenue.

8 Top 3 highest earning drivers (based on driver_base_cost)

```
SELECT driver_base_cost, booking_id

FROM ola_trips

ORDER BY driver_base_cost DESC

LIMIT 3;
```

Output:

driver_base_cost	booking_id

960.54	1890061540
890.30	1672692603
875.20	1542148932

Insight:
Shows customer intent patterns like office, airport, or personal rides — valuable for market segmentation.

9 Trips with toll charges greater than 0

```
SELECT booking_id, category, toll, total_trip_cost

FROM ola_trips

WHERE toll > 0

ORDER BY toll DESC;
```

Output:

booking_id	gender	month	day_of_week	distance_travelled	time_taken	reason	toll	category	total_trip_cost	ratings
192560002 01	Female	June	Thu	15	49	Office to/from Home	35	Micro	312.00	5
177331627 2	Male	April	Mon	30	91	Office to/from Home	35	Micro	751.00	4
1644	M	Ma	Tue	62	98	Office	3	Mi	1079.0	4

393205	ale	rch				Event	5	ni	0	
194183159	Male	June	Mon	15	44	Office to/from Home	3 5	Mi ni	290.00	3
170346896	Female	April	Fri	23	81	Office to/from Home	3 5	Mic ro	470.79	3
169627563	Female	April	Wed	38	78	Office to/from Home	3 5	Mi ni	753.07	1
187042830	Male	May	Thu	31	76	Office to/from Home	3 5	Mi ni	562.00	3
167287950	Female	April	Wed	20	58	Office to/from Home	3 5	Mic ro	317.68	5
153768205	Female	February	Wed	31	72	Office Event	6 0	Mic ro	524.82	5
194204249	Female	June	Tue	35	76	Market/Site Visit	7 0	Pri me	679.00	5
197109207	Female	June	Mon	31	67	Office to/from Home	7 0	Mic ro	762.00	2
166980331 5	Male	April	Tue	30	54	Office to/from Home	7 0	Mic ro	500.24	4
140817330 3	Male	January	Wed	13	28	Office Event	6 0	Mic ro	269.00	5

197889316	Male	June	Wed	15	48	Office to/from Home	35	Mini	427.00	4
195201157	Female	June	Thu	31	68	Office to/from Home	70	Prime	795.00	4
191826521	Male	June	Tue	30	90	Office to/from Home	35	Micro	603.00	3
187214189	Male	May	Fri	20	46	(Null)	70	Micro	337.00	4
188659187	Female	May	Mon	16	57	Office to/from Home	35	Micro	414.00	5
183362062	Female	May	Mon	17	58	Office to/from Home	35	Micro	293.00	2
169006388	Female	April	Tue	14	30	Office to/from Home	35	Micro	228.75	4

Insight:
Highlights highway or intercity trips that included toll payments, indicating travel beyond city limits.

10 Correlation between distance and total cost (rough check using grouping)

```
SELECT
CASE
WHEN distance_travelled < 5
THEN 'Short'
WHEN distance_travelled BETWEEN 5 AND 15
THEN 'Medium'
ELSE 'Long' END AS trip_length, ROUND(AVG(total_trip_cost), 2) AS avg_cost
FROM ola_trips
GROUP BY CASE WHEN distance_travelled < 5
THEN 'Short'
WHEN distance_travelled BETWEEN 5 AND 15
THEN 'Medium' ELSE 'Long'
END;
```

Output:

trip_length	avg_cost
Short	120.45
Medium	275.34
Long	690.78

- **Insight:** Compares ride lengths across cab types — longer distances often belong to rental or premium services.

1️⃣1️⃣ Which category generates the highest profit margin per trip?

```
SELECT category, ROUND(AVG(total_trip_cost - (driver_base_cost + total_tax)), 2) AS avg_profit_margin

FROM ola_trips

GROUP BY category

ORDER BY avg_profit_margin DESC;
```

Output:

Category	Avg Profit Margin (₹)
Mini Rentals	235.44
SUV Rentals	229.48
Prime Rentals	229.33
Prime Play Rentals	204.77
Lux	120.20
Prime Play	73.32
Prime	48.29
Mini	47.60
Micro	44.86
Play	20.83
Sedan	14.55

📊 **Insight:**
Premium and rental-based categories (like *Mini Rentals* and *Prime Rentals*) generate much higher profit margins per trip.

1️⃣2️⃣ Find the top 5 longest trips (by distance) and their total costs

```
SELECT booking_id, distance_travelled, total_trip_cost, category, gender, reason

FROM ola_trips
```

ORDER BY distance_travelled DESC

LIMIT 5;

Ourput:

Booking ID	Distance (km)	Total Trip Cost (₹)	Category	Gender	Reason
1630132872	66	1512.70	Prime Rentals	Female	Office Event
1725296449	65	1232.00	Mini	Female	Office Event
1633384022	64	1828.12	Prime Play Rentals	Female	Office Event
1390385354	64	1714.00	Prime Play Rentals	Female	Customer/Partner Visit
1721890414	63	1174.00	Mini	Male	Office Event

📊 **Insight:**
Longest rides are mostly **corporate-related (Office Events)** and handled by **premium or rental categories**, often led by **female passengers**.

1️⃣3️⃣ Determine average cost per kilometer for each category

SELECT category, ROUND(AVG(total_trip_cost / distance_travelled), 2) AS avg_cost_per_km

FROM ola_trips

WHERE distance_travelled > 0

GROUP BY category

ORDER BY avg_cost_per_km DESC;

Output:

Category	Avg Cost per km (₹)
Lux	107.78
Prime Play Rentals	59.17
Prime Rentals	47.48
SUV Rentals	45.78
Prime	32.47
Mini	31.62
Prime Play	29.68
Mini Rentals	26.34
Micro	25.90
Play	24.37
Sedan	21.65

📊 **Insight:**
Lux is the most premium segment (₹107.78/km), while **Sedan** and **Play** are the most affordable.

1️⃣4️⃣ Average customer rating by gender and category

```
SELECT
  gender,
  category,
  ROUND(AVG(ratings), 2) AS avg_rating
FROM ola_trips
GROUP BY gender, category
```

ORDER BY avg_rating DESC;

Output:

Gender	Category	Avg Rating
Male	Prime Rentals	4.25
Female	Prime Play Rentals	4.14
Male	Play	4.13
Female	Lux	4.00
Male	Mini Rentals	4.00
Female	Prime Play	3.85
Male	Prime	3.84
Female	Mini	3.78
Male	Mini	3.76
Female	Sedan	3.75
Male	Micro	3.71
Female	Prime	3.70
Female	Micro	3.67
Female	Prime Rentals	3.62
Male	Prime Play	3.58
Female	Play	3.57
Male	Sedan	3.56
Male	Prime Play Rentals	3.40
Female	Mini Rentals	3.00
Female	SUV Rentals	1.00

📊 **Insight:**
Males using Prime Rentals rated highest (4.25), while *Females using SUV Rentals* gave the lowest (1.00). Premium categories generally receive better feedback.

1️⃣5️⃣ Which day of the week has the most expensive average trips?

```
SELECT day_of_week, ROUND(AVG(total_trip_cost), 2) AS avg_trip_cost
FROM ola_trips

GROUP BY day_of_week

ORDER BY avg_trip_cost DESC;
```

Output:

Day of Week	Avg Trip Cost (₹)
Friday	283.65
Monday	278.06
Thursday	271.05
Tuesday	256.39
Wednesday	244.51
Saturday	224.80
Sunday	204.72

📊 **Insight:**

Friday rides are most expensive on average — likely due to **weekend rush and surge pricing**.