
🚦 Theme: Smart City Traffic Monitoring System

📁 Dataset: `traffic_logs.csv`

```
LogID,VehicleID,EntryPoint,ExitPoint,EntryTime,ExitTime,VehicleType,SpeedKMH,TollPaid
L001,V001,GateA,GateC,2024-05-01 08:01,2024-05-01 08:20,Car,60,50
L002,V002,GateB,GateC,2024-05-01 08:10,2024-05-01 08:45,Truck,45,100
L003,V003,GateA,GateD,2024-05-01 09:00,2024-05-01 09:18,Bike,55,30
L004,V004,GateC,GateD,2024-05-01 09:15,2024-05-01 09:35,Car,80,50
L005,V005,GateB,GateA,2024-05-01 10:05,2024-05-01 10:40,Bus,40,70
```

1️⃣ Data Ingestion & Schema Analysis

- Load CSV using PySpark with schema inference
- Manually define schema and compare
- Ensure EntryTime/ExitTime are `timestamp`

2️⃣ Derived Column Creation

- Calculate `TripDurationMinutes = ExitTime - EntryTime`
- Add `IsOverspeed = SpeedKMH > 60`

3️⃣ Vehicle Behavior Aggregations

- Average speed per `VehicleType`
- Total toll collected per gate (EntryPoint)
- Most used ExitPoint

4️⃣ Window Functions

- Rank vehicles by speed within `VehicleType`
- Find last exit time for each vehicle using `lag()`

5️⃣ Session Segmentation

- Group by VehicleID to simulate route sessions
- Find duration between subsequent trips (idle time)

6️⃣ Anomaly Detection

- Identify vehicles with speed > 70 and TripDuration < 10 minutes
- Vehicles that paid less toll for longer trips
- Suspicious backtracking (ExitPoint earlier than EntryPoint)

7️⃣ Join with Metadata

Prepare this small `vehicle_registry.csv`:

```
VehicleID,OwnerName,Model,RegisteredCity
V001,Anil,Hyundai i20,Delhi
V002,Rakesh,Tata Truck,Chennai
```

```
V003,Sana,Yamaha R15,Mumbai
V004,Neha,Honda City,Bangalore
V005,Zoya,Volvo Bus,Pune
```

- Join and group trips by `RegisteredCity`

8️⃣ Delta Lake Features

- Save traffic_logs as Delta Table
- Apply `MERGE INTO` to update toll rates for all Bikes
- Delete trips longer than 60 minutes
- Use `DESCRIBE HISTORY` and `VERSION AS OF`

9️⃣ Advanced Conditions

- `when/otherwise`: Tag trip type as:
  - "Short" <15min
  - "Medium" 15-30min
  - "Long" >30min
- Flag vehicles with more than 3 trips in a day

🔟 Export & Reporting

- Write final enriched DataFrame to:
  - Parquet partitioned by VehicleType
  - CSV for dashboards
- Create summary SQL View: total toll by VehicleType + ExitPoint

---