Uber Data Analysis Project

Technical Report

# 1. Project Overview

This project focuses on building a scalable, automated analytics pipeline for Uber trip data using ETL (Extract, Transform, Load) and a structured data warehouse, with the final output presented in an interactive Power BI dashboard.

# 2. Tools & Technologies Used

|  |  |  |
| --- | --- | --- |
| Component | Technology | Purpose / Justification |
| Programming | Python (Pandas) | For reading, cleaning, and transforming raw CSV files |
| Database | Microsoft SQL Server | Robust relational database with high performance for large datasets; seamless integration with Power BI |
| BI Tool | Power BI | Interactive dashboarding, KPI cards, filtering, and visual storytelling |
| IDE/Notebook | Jupyter Notebook | Easy step-by-step ETL implementation and data exploration |
| CSV | Uber\_Data.csv | Real-world structured Uber trip dataset for analysis |
| SQL Scripting | T-SQL | Schema creation, data loading, aggregation queries |

# 3. Architecture Diagram

Raw Uber CSV Files  
 ↓  
[ Python ETL Script ]  
 ↓  
Cleaned Uber Trip Dataset  
 ↓  
[ SQL Server Database ]  
 ↓  
Star Schema DW  
 ↓  
[ Power BI Report ]  
 ↓  
Interactive Dashboard

# 4. ETL Pipeline Overview

## Extraction

- Read CSV with Pandas  
- Inspect schema and data quality

## Transformation

- Handle missing values  
- Convert datetimes  
- Feature engineering: Trip Duration, Tip Rate, Revenue/Mile

## Load

- Export cleaned CSV  
- Use BULK INSERT to load into SQL Server  
- Populate fact and dimension tables

# 5. Data Warehouse Design (Star Schema)

FactTrips: Trip metrics (distance, fare, duration, tips, etc.)  
  
Dimension Tables:  
- Date\_Dim: Year, Month, Day, Weekday  
- Time\_Dim: Hour of day  
- Vendor\_Dim: Vendor name  
- Payment\_Dim: Payment method

# 6. Analytical Queries

- Revenue by distance band  
- Tip by vendor  
- Trips by hour  
- Fare by payment type  
- Surge vs non-surge revenue

# 7. Power BI Dashboard Design

- KPIs: Total Trips, Revenue, Avg Duration/Distance  
- Vendor/Payment comparisons  
- Drill-through to detailed trip data

# 8. Justification of Technology Choices

- Python: Fast, flexible ETL and EDA  
- SQL Server: Best for Power BI integration and indexing  
- Power BI: Interactive dashboards with drill-through and filters  
- Star Schema: Industry standard for analytics performance

# 9. Key Insights

- Peak revenue during evening hours (5–8 PM)  
- Vendor 2 outperforms in volume and revenue  
- Credit Card is most popular payment method  
- Higher tip rates in off-peak hours

# 10. Future Enhancements

- Add real-time streaming (Kafka/Azure)  
- ML models for trip duration/surge prediction  
- Map-based pickup/drop-off heatmaps  
- Cloud migration to Snowflake/BigQuery

# 11. Final Remarks

This full-stack project showcases ETL automation, structured warehousing, and business-focused visual insights using industry-standard tools.