# Title: Big Data Analysis for Ride-Sharing Service

**Problem Statement:** A ride-sharing company wants to analyze ride data to gain insights into ride patterns, driver performance, and customer preferences. They need to build a data pipeline that collects, processes, and analyzes data from various sources to generate meaningful insights for decision-making. As a data engineer, your task is to design and implement the data pipeline to support this analysis.

**Dataset:**

The dataset for this project will consist of the following sources:

* Ride Data: Contains information about each ride, including ride ID, driver ID, customer ID, start time, end time, pickup location, dropoff location, and fare.
* Driver Data: Contains details of drivers, such as driver ID, name, age, gender, and years of experience.
* Customer Data: Includes customer demographics, such as customer ID, name, age, gender, and location.

## Project Steps:

**Data Ingestion:**

- Create an ingestion process to receive and store the raw data from different sources.

- Store the data in a distributed file system (e.g., HDFS).

**Data Processing:**

- Design ETL (Extract, Transform, Load) processes to cleanse and transform the raw data.

- Implement data quality checks and filtering to ensure data integrity.

- Perform data enrichment by joining different datasets based on common keys.

- Utilise Apache Spark for distributed data processing and transformation.

- Apply data modelling techniques (e.g., data normalisation, denormalization) as per the analysis requirements.

**Data Storage:**

- Choose a suitable storage like HDFS for storing processed data.

- Create optimised tables and partitions for efficient querying and analysis.

- Ensure data security and privacy measures are in place.

**Data Analysis and Visualization:**

- Use SQL queries or Spark SQL to extract relevant insights from the processed data.

- Perform ride pattern analysis based on time, location, and fare.

- Generate reports, dashboards, and visualisations using tools like Apache Superset, Tableau, or Power BI.

**Overall architecture flow:**

****

**Task 1: Data Ingestion and Storage**

Outcome: Ingest the bike sharing data into Hadoop using Apache Sqoop, process it using Apache Spark, and store the results in Hive.

Deliverables:

- Sqoop command to ingest data from the CSV file into Hadoop.

- Spark code to process the ingested data.

- Hive commands to store the processed data.

**Task 2: Usage Pattern Analysis**

Outcome: Analyse the bike sharing data to identify usage patterns based on time, station, and user demographics using Apache Spark and SQL queries.

Deliverables:

- Spark code and SQL queries to analyse the bike sharing data.

- A report detailing the usage patterns and any interesting trends in the data.

**Task 3: Data Visualisation**

Outcome: Create visualisations of the usage patterns using a tool like Apache Superset, Tableau, or Power BI.

Deliverables:

- Visual representations (e.g., line charts, bar charts, heatmaps) of the usage patterns.

- A dashboard that displays the visualisations and allows users to interact with the data.

* Average trip duration per time of day
* Trend of taxi usage over time
* Top 10 busiest pickup and dropoff locations
* Correlation between trip distance and duration

For this assignment, you can use the [New York City Taxi Trip Duration dataset](https://www.kaggle.com/c/nyc-taxi-trip-duration/data) available on Kaggle for the ride data. You might need to create synthetic datasets for the driver and customer data based on the specifications provided.

**Tools required to achieve the end dashboard:**

1. **Hadoop ( hdfs , hive , spark )**
2. **Tableau/Power BI**