MILESTONE-2

IDEA-1

1. **(2 points) Title of the project – Be precise!**
   Airline Fare Prediction

2. **(5 points) Project idea – A detailed and concise description of what you plan to do in the project.**

   The Airline Fare Prediction dataset contains information about flights, including the airline, fare, and other variables such as the date and duration of the flight. The dataset includes both numerical and categorical data.

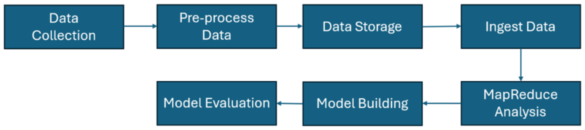
   For a big data project, this dataset can be used to analyze various factors that influence the price of a flight, such as the airline, the route, and the time of travel.

   By analyzing this data, you can gain insights into patterns and trends in airline pricing and make predictions about the fare for a given flight.

3. **(3 points) What tools and technologies you plan to use for the project?**

   Hadoop and Spark for distributed data processing, Hive and Impala for SQL-like queries, Kafka for real-time data streaming, and Apache Storm for complex event processing.

4. **(5 points) Sketch the high-level architecture or methodology of the project using a block diagram. In other words, draw the data flow diagram for your project.**



5. **(5 points) Explain the diagram in the above diagram in simple words using the bullet list.**

   - The data source for this project are the CSV file(dataset) that contain the airline fare information.
   - The "Data Preparation" block cleans and processes the data to ensure that it is in a suitable format for further analysis.
   - The "Data Storage" block uses SQL queries to perform various

operations such as aggregations, joins, and filtering.
- The file is ingested and loaded into the big data processing framework using the "Data Ingestion" block.
- The cleaned and processed data is then stored in a Parquet format for efficient querying and further analysis.
-MapReduce is a programming model and an associated implementation for processing and generating large data sets with a parallel, distributed algorithm on a cluster.
- Before building a model, it is essential to evaluate its performance using various metrics.
- After evaluating the model's performance, the next step is to build the model using appropriate algorithms and techniques.
- Finally, the data is analyzed using various techniques and tools to gain insights and make informed decisions.
- The results of the data analysis are then used to generate insights and make recommendations for the airline industry.

6. **(10 points) Formulate and write the goals your team wants to investigate.**

   The goals for the analysis in this project are as follows:

- To calculate the average fare for each airline.
- To determine the number of flights operated by each airline.
- To identify the maximum fare for each route.
- To determine the minimum fare for each route.
- To calculate the average fare for each route.
- To find the average fare for each unique route, sorted from highest to lowest.
- To identify the number of fares for each airline that are more than the median fare of that airline's flights.

These goals will be achieved by performing various operations and analyses on the dataset, using the tools and technologies mentioned.

Team Members:
1. Aashritha Dodda(s559361)
2. Volete,Sai Prashanth(s559234)
3. Punyam Anand, Maheshwar(s559173)
4. Patlolla,Venkateshwar Reddy(s555897)
5. Parvathaneni,Karthik(s559449)