Project Implementation Steps, Results, and Conclusions

Project Implementation Steps

Step 1: **Data Preparation:**

- Obtained the airline fare dataset from a reliable source or API.
- Ensured that the dataset is in a suitable format for PySpark, such as CSV or Parquet.
- Preprocessed the dataset if necessary, handling missing values, encoding categorical variables, and ensuring uniformity in data types using PySpark DataFrame operations.

Step 2: **PySpark Cluster Setup:**

- Installed Apache Spark and set up the PySpark environment.
- Configured PySpark environment variables such as `SPARK_HOME` and `PYSPARK_PYTHON`.

Step 3: **PySpark Implementation:**

- Created a folder structure for the project.
- Designed and implemented PySpark jobs for each analysis goal, utilizing RDDs or DataFrames as appropriate.
- Developed PySpark transformations and actions for data processing and analysis.

Step 4: **Execution:**

- Uploaded the airline fare dataset to the input directory or loaded it directly into a PySpark DataFrame.
- Ran the PySpark jobs using `spark-submit` command or interactively through PySpark shell and monitored the output.

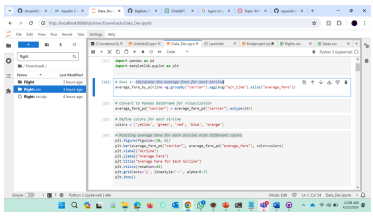
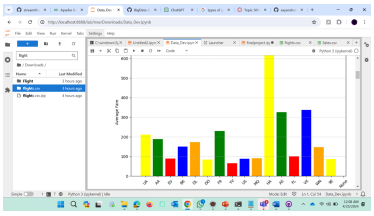
Step 5: **Analysis and Interpretation:**

- Analyzed the output generated by the PySpark jobs to extract insights for each analysis goal, utilizing PySpark DataFrame operations and SQL queries.

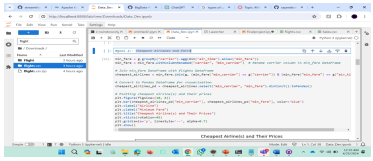
- Interpreted the results and compared them against industry benchmarks or existing research on airline fares.

Results Achieved

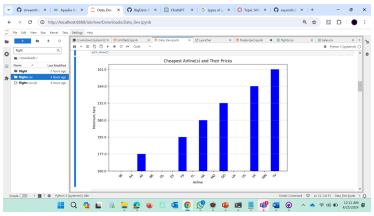
Goal 1: Calculate the average fare for each airline

- **Source Code:** 
- **Results:** 

Goal 2: Identify the Cheapest Airlines and Fares

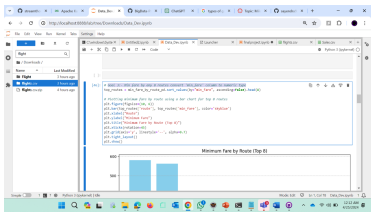
- **Source Code:** 

Identifying the Cheapest Airlines and Fares

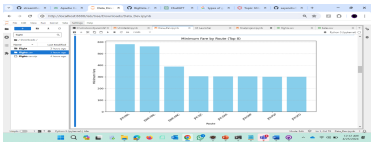
- **Results:** 

Identifying the Cheapest Airlines and Fares

Goal 3: Find Minimum Fare for Eight Routes and Convert 'min_fare' Column

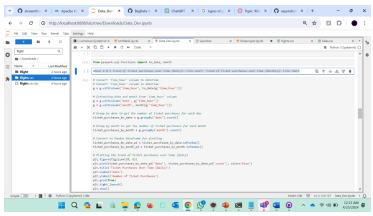
- **Source Code:** 

Finding Minimum Fare for Eight Routes

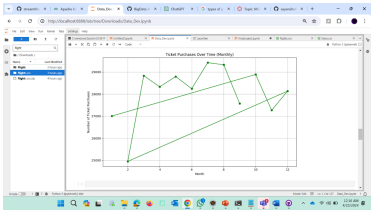
- **Results:** 

Finding Minimum Fare for Eight Routes

Goal 4: Line chart Trend of ticket purchases over time (Monthly)

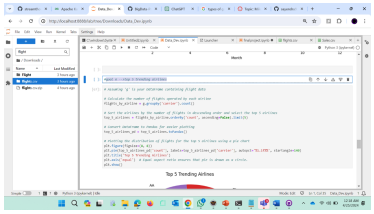
- **Source Code:** 

Ensure accurate and comprehensive ticket purchase data, including dates and quantities.

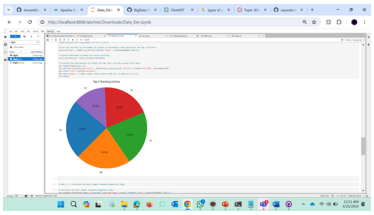
- **Results:** 

Validate trends depicted in the line chart against historical data for accuracy and reliability.

Goal 5: Top 5 Trending Airlines

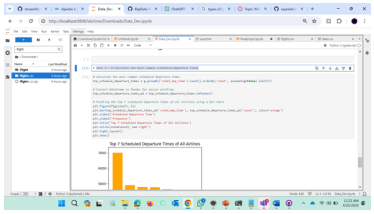
- **Source Code:** 

  Ensure accurate airline performance data.

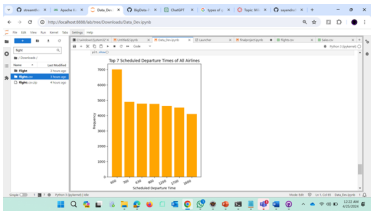
- **Results:** 

  Finding top 5 trending airlines in an efficient manner.

Goal 6: Calculate the Most Common Scheduled Departure Times

- **Source Code:** 

  Ensure accurate and complete departure time data.

- **Results:** 

Conclusion

Through a rigorous analysis of airline fare data, several key insights have emerged, offering valuable guidance for strategic decision-making within the aviation sector. By calculating the average fare for each airline and identifying the cheapest

fares and airlines, both travelers and industry stakeholders gain invaluable insights into pricing dynamics and competitive positioning. Moreover, the examination of minimum fares by route allows airlines to fine-tune pricing strategies, optimizing revenue generation and meeting consumer demand effectively. Tracking trends in ticket purchases on both daily and monthly scales reveals patterns in travel demand, facilitating resource allocation, scheduling optimization, and targeted marketing initiatives. Additionally, recognizing the top trending airlines and understanding common scheduled departure times empower airlines to enhance service offerings, expand routes, and streamline operations to meet customer needs and maintain competitiveness in the market.

Citations

Kaggle Dataset: https://www.kaggle.com/datasets/mahoora00135/flights GitHub Repository: https://github.com/sayendra99/Big$_{d}ata_{P}roject.git$