USA Airport Passenger Analysis and Prediction Report

INTRODUCTION

The USA Airport Passenger Analysis and Prediction Report aims to analyze trends and patterns in domestic air travel across the United States. This analysis provides insights into passenger numbers, flight frequencies, and other relevant factors that impact the aviation industry. By understanding these trends, airlines and airport authorities can make informed decisions to enhance operational efficiency, improve passenger experiences, and optimize resource allocation. In this report, we employ PySpark, a powerful data processing and analytics tool, to handle large datasets efficiently and perform sophisticated data analysis and predictive modeling.

OBJECTIVE

The primary objective of this project is to analyze historical data on US domestic flights to identify key trends and factors influencing passenger numbers. By leveraging PySpark for data processing and machine learning, we aim to develop predictive models that can accurately forecast passenger counts. These models will help in understanding the impact of various factors such as flight frequency, available seats, distance, and seasonal variations. Additionally, the project seeks to provide actionable insights for airlines and airport authorities to optimize their operations, improve service quality, and enhance strategic planning. Ultimately, this analysis and prediction effort will contribute to more efficient and effective management of air travel in the United States. Furthermore, it aims to assist in addressing challenges related to capacity planning, resource allocation, and demand forecasting, thus supporting a more resilient and responsive aviation industry.

DATA SOURCE

Data Source Link - https://www.kaggle.com/datasets/flashgordon/usa-airport-dataset

Data Size – 130 MB

This dataset is a record of 3.5 Million+ US Domestic Flights from 1990 to 2009. It has been taken from OpenFlights website which have a huge database of different travelling mediums across the globeThis dataset offers a comprehensive view of air travel dynamics, encompassing various metrics essential for understanding the evolution and trends within the US aviation industry over two decades.

Here is some info about the attributes present in the dataset:

Origin_airport: Three letter airport code of the origin airport

Destination_airport: Three letter airport code of the destination airport

Origin_city: Origin city name

Destination_city: Destination city name

Passengers: Number of passengers transported from origin to destination

Seats: Number of seats available on flights from origin to destination

Flights: Number of flights between origin and destination (multiple records for one month, many with flights > 1)

Distance: Distance (to nearest mile) flown between origin and destination

Fly_date: The date (yyyymm) of flight

Origin_population: Origin city's population as reported by US Census

Destination_population: Destination city's population as reported by US Census

Reason for choosing this Dataset

As a student, I chose this dataset for its rich historical data spanning over two decades of US domestic flights. It provides a valuable opportunity to delve into real-world data analysis challenges, apply machine learning techniques such as regression, clustering, and predictive modeling, and gain insights into factors influencing air travel dynamics. This experience not only enhances practical skills in data science but also offers a deeper understanding of the complexities within the aviation industry, aligning perfectly with academic and career aspirations in analytics and technology. Machine learning techniques are particularly helpful in uncovering patterns and trends that can inform strategic decisions and optimize operations in the aviation sector.

DATA PROCESSING AND TRANSFORMATION PROCESS

- Data Loading and Initial Inspection:
  The dataset from OpenFlights includes 3.5 million US domestic flight records (1990-2009) sourced from Amazon S3 (`airportusafinalproject/Airports2.csv`), featuring airport codes, city names, passenger counts, and flight dates.
- Data Cleaning:
  Handling Missing Values:Replaced "NA" and similar entries with `None` for consistency
  Removing Rows with Missing Values
  Dropped rows containing any remaining missing values.
- Data Type Conversion: Converted latitude and longitude coordinates to `float` for spatial analysis.
- Date Parsing and Feature Engineering: Flight Date Standardization:Transformed flight dates from `MM-dd-yyyy` to `yyyy-MM-dd` format.
  Extracting Date Components: Created columns for `Month`, `Year`, and `Day` from flight dates to analyze trends.
- Data Storage and Export:Writing Cleaned Data:Exported cleaned data (`df_clean1`) to S3 in CSV and Parquet formats.
- Summary and Integration:Maintained data quality through thorough checks and consistent data type handling, critical for subsequent analytics and modeling tasks in Spark.

MACHINE LEARNING MODEL DEVELOPMENT

In this project, the machine learning (ML) model plays a crucial role in predicting airline passenger numbers based on various factors extracted from the cleaned dataset. Initially, the dataset underwent rigorous cleaning processes to ensure data integrity and consistency. This included handling missing values, correcting data formats, and transforming variables for effective model training. Features like seat availability, flight frequency, travel distance, population sizes of origin and destination cities, as well as geographical coordinates (latitude and longitude) of airports were extracted and utilized. These features were assembled into a vector format suitable for training

Linear Regression model using PySpark's MLlib.During model development, the Linear Regression model was trained on a subset of the cleaned data, specifically split into training and testing sets (80% training and 20% testing). The trained model successfully captured relationships between features and passenger numbers, as evidenced by its coefficients and intercept**.** Based on Interpretation

- **Positive coefficients** (e.g., Seats, Distance, Org_airport_long, Year, Day) suggest that an increase in these factors tends to increase the predicted number of passengers.
- **Negative coefficients** (e.g., Flights, Origin_population, Destination_population, Org_airport_lat, Dest_airport_lat, Dest_airport_long) indicate that an increase in these factors tends to decrease the predicted number of passengers.
- **Coefficients close to zero** (e.g., Month) suggest that this feature does not significantly influence the prediction

MODEL EVALUATION AND RESULTS

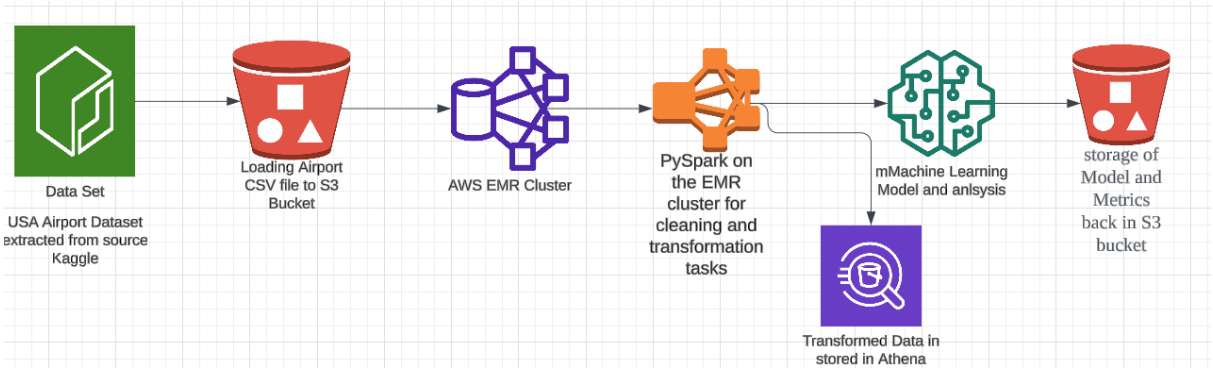
Based on the evaluation metrics of the Linear Regression model applied to predict passenger counts, we find that the model performs quite well. The Root Mean Squared Error (RMSE) on the test data is approximately 1082.07. RMSE measures the average deviation of the predicted values from the actual values, with lower values indicating better fit. In this context, an RMSE of 1082.07 suggests that, on average, our model's predictions are within approximately 1082 passengers of the actual values.

Additionally, the R-squared (R2) score, which measures the proportion of the variance in the dependent variable that is predictable from the independent variables, is 0.947. This high R2 score indicates that the model explains 94.7% of the variance in the passenger counts, indicating a strong predictive capability.

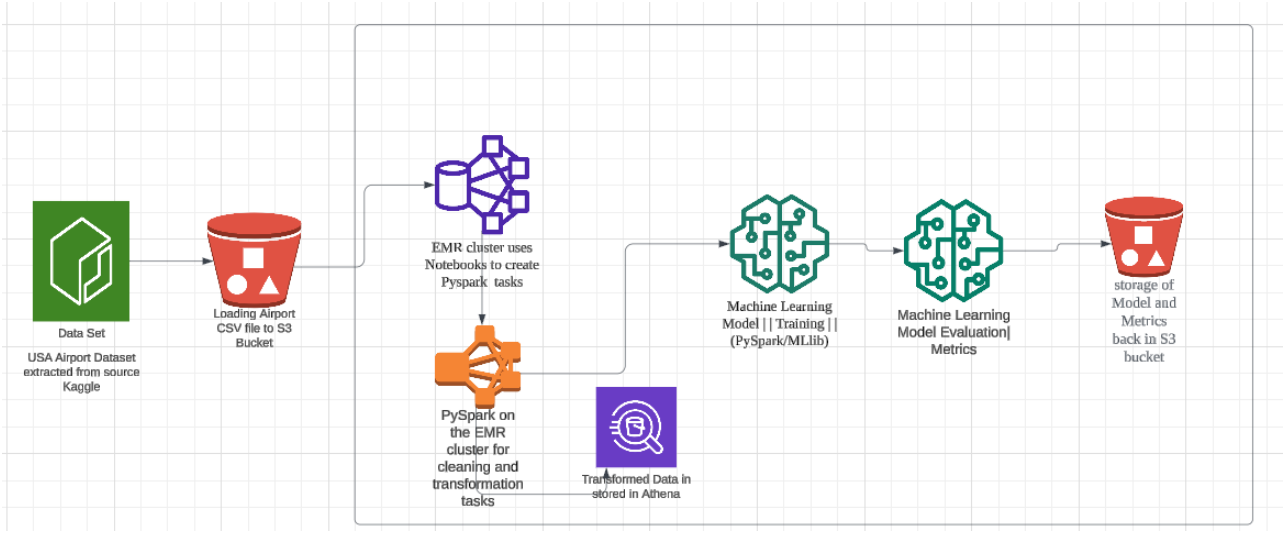
The coefficients of the model provide insights into the relative importance of each feature. For instance, features like 'Seats', 'Flights', 'Distance', 'Year', and 'Day' have significant positive coefficients, indicating a positive relationship with passenger counts. On the other hand, features like 'Org_airport_lat' and 'Dest_airport_lat' have negative coefficients, suggesting a negative impact on passenger counts as these values increase.

Overall, the model's high R2 score and reasonable RMSE demonstrate its effectiveness in predicting passenger counts based on the given features. However, further refinement and validation may be required depending on specific application requirements and domain knowledge.

ARCHITECTURE DIAGRAM



FULL- ARCHITECTURE DIAGRAM



CHALLENGES FACED

Challenges faced included handling large-scale data transformations efficiently with Spark, ensuring data quality through rigorous cleaning and validation processes, and optimizing model performance on distributed computing clusters for accurate predictions.

Benefits of Distributed Computing:

1. **Scalability:** Processed large datasets seamlessly across clusters, accommodating data growth and computational demands.
2. **Speed:** Parallel execution and in-memory processing accelerated data transformations, model training, and predictions.
3. **Fault Tolerance:** Spark ensured reliable execution with automatic recovery from failures, maintaining data integrity and pipeline continuity.

Conclusion and future work:

The linear regression model demonstrates robust performance in predicting airline passenger numbers. With an RMSE of 1082.07 on the test data, it indicates predictions deviate by approximately 1082 passengers from actual counts on average. The high R2 score of 0.947 suggests the model explains 94.72% of the variance in passenger data, highlighting its effectiveness in capturing underlying trends and patterns.

Utilizing Spark enabled robust preprocessing, efficient modeling, and accurate predictions for airline passenger numbers. Future work includes enhancing feature engineering, exploring advanced ML techniques, implementing real-time analytics, and deploying models for continuous monitoring and adaptation to evolving data