Aviation Data Analysis

Kusumitha Korrapati
MEng in Computer Engineering
University of Guelph

Reesa Susan Sabu
MEng in Computer Engineering
University of Guelph

ABSTRACT
Data analysis in today's interconnected world presents challenges due to the increasing interdisciplinarity of data and the need for specialized expertise in various fields. This paper presents a case study on leveraging big data for airline data analysis using Spark tools, within the context of the airline industry. The study explores the challenges and opportunities of analysing massive amounts of data generated from diverse sources such as social media, finance, flight data, environment, and health. The utilization of big data for purposes such as risk assessment, real-time product tracking, and monitoring various parameters is discussed. It highlights the importance of effective data analytics in extracting meaningful insights from big data to support decision-making. The findings of the case study contribute to the growing body of knowledge on leveraging big data for airline data analysis and provide valuable insights for researchers and practitioners in the field of data analytics.

KEYWORDS
Data Analytics; Data Visualization; Big Data; Tableau and PySpark.

1. INTRODUCTION

Big Data is a modern approach used to analyze extensive and intricate data sets, with no universally agreed-upon definition. One commonly utilized framework is the "4 V's" concept. The first "V" represents Volume, referring to the vast amount of data that poses challenges for traditional data analytics. Big Data aids in processing and analyzing such data. The second "V" is Velocity, signifying the high speed at which data is generated, processed, and analyzed. The third "V" is Variety, enabling analysis of data from diverse sources like social media, encompassing text messages, attachments, images, and more. The fourth "V" is Veracity, which pertains to the accuracy and reliability of the data, given the massive volume of data being processed.

Structured data researchers face challenges in analyzing data from sources like social media, blogs, Facebook posts, and Snapchat, as these data have different structures and formats that are not easily stored in traditional databases. Big Data encompasses data in different shapes and formats, including structured data. It involves working with a variety of data formats and structures, such as data from sensors tracking object movement or environmental changes. The Internet of Things (IoT) generates vast volumes of wearable data, which require big data approaches for management and analysis. Big Data includes data from diverse fields like flight data, population data, financial and health data, which bring value to users.

Big Data has a wide range of applications beyond just advertising on social media. It is utilized in industries such as insurance for risk assessment, real-time tracking of product reactions, monitoring of areas such as wave movements, flight data, traffic data, financial transactions, health, and crime. The key challenge lies in harnessing the potential of Big Data to create value for users by effectively gathering, storing, processing, and analyzing raw data to support decision-making.

Apache Spark is a powerful distributed computing framework for big data processing that provides various data processing capabilities, including batch processing, real-time processing, machine learning, and graph processing. One common use case is to load data from a Pandas Data Frame into a Spark cluster for distributed processing. In this report, we will explore how to load a Pandas Data Frame into a Spark cluster using the Spark Session class's, create Data Frame () method, and how to register it as a temporary table for further data manipulation using them. createTempView () or. createOrReplaceTempView () methods. By following this approach, we can leverage Spark's distributed computing capabilities for big data processing while utilizing SQL-like queries for data manipulation. However, it's important to note that the temporary table is only accessible within the specific Spark Session and should be used accordingly.

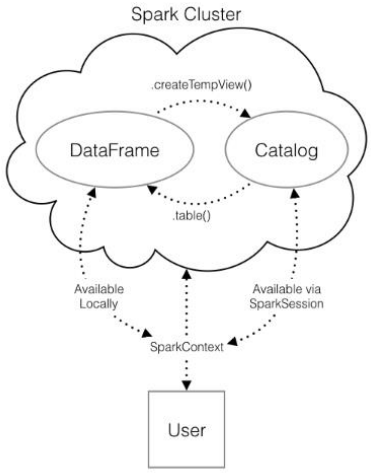


Figure 1: Spark Data Structures and Interactions

This project includes the creation of a logistic regression model using PySpark's Logistic Regression estimator, which will be tuned using k-fold cross validation. This involves splitting the training data into multiple partitions (in this case, three), fitting the model to some partitions, and evaluating its performance on the held-out partition. This process is repeated for each partition, and the average error is calculated, providing an estimate of the model's performance on unseen data. To compare different models, a binary classification evaluator is used with the metric "Area Under ROC" to evaluate the performance of each model. The report can mention the use of cross validation and the binary classification evaluator as a method for selecting the best hyperparameters for the logistic regression model, which can help improve the model's performance and generalization to unseen data.

Tableau can be a valuable tool in flight data analysis, as it allows for the visualization and analysis of large volumes of data related to flights, such as flight routes, schedules, delays, cancellations, and passenger information. With Tableau's intuitive interface and powerful visualizations, flight data analysts can easily explore, analyze, and identify patterns and trends in the data. For example, Tableau enables the creation of dynamic and interactive dashboards that offer real-time updates on flight statuses, visualizations of airline and airport performance, and identification of key factors contributing to delays or cancellations. These valuable insights empower stakeholders in the aviation industry, including airlines and airports, to make informed decisions based on data, leading to enhanced operational efficiency, improved customer experience, and optimized resource management. Tableau's ability to handle large and complex datasets, combined with its powerful visualizations and analytics capabilities, make it a valuable tool in flight data analysis for gaining actionable insights from the vast amount of data generated by the aviation industry.

2. ASSESSMENT OF DATASETS

The "airports" table contains information about airports in the USA, including the IATA code, airport name, city, state, country, latitude, and longitude. The columns provide details about various airports in the country. This table is a valuable resource for retrieving data about airports for analysis or reporting purposes in a database system.

The "flights5" table, on the other hand, contains information about flights. This table is useful for tracking flight information and analyzing flight patterns in a database system. More details of these two tables are given below.

The airlines table contains data regarding the IATA code and the name of the airline.

Table 1. Airport Table

Column Name	***Description***
IATA_CODE	IATA code of the airport
AIRPORT	Name of the airport
CITY	City where the airport is located
STATE	State where the airport is located
COUNTRY	Country where the airport is located
LATITUDE	Latitude of the airport
LONGITUDE	Longitude of the airport

Table 2. Flights5 Table

Column Name	***Description***
AIRLINE	Airline code
FLIGHT_NUMBER	Flight number
TAIL_NUMBER	Tail number of the aircraft
ORIGIN_AIRPORT IATA	Code of the origin airport
DESTINATION_AIRPORT IATA	Code of the destination airport
SCHEDULED_DEPARTURE	Scheduled departure time in HHmm format

Table 3. Airlines Table

Column Name	***Description***
IATA_CODE	IATA code of the airport
AIRLINE	The full name of the airline

3. METHODOLOGY

The methodology for analyzing the Flights Data containing information on flight schedules, departure and arrival times, airline details, and other relevant variables, was collected from a reliable source. The data was then loaded into a PySpark DataFrame, a distributed collection of data, for efficient processing. Data preprocessing techniques, such as data cleaning, data transformation, and data aggregation, were applied using PySpark's rich set of data processing capabilities. Advanced machine learning tasks, such as regression, were performed using PySpark's built-in MLlib library.

Visualization tools may have been used to communicate the findings effectively. The methodology may have been iterative, involving multiple rounds of data preprocessing, analysis, and visualization to arrive at meaningful insights and conclusions.

i. Data Collection: The dataset used in this analysis was sourced from two CSV files i.e. flights5 and airport. These tables are likely used for data loading and data analysis tasks. The "airports" table may be used to gather information about the airports involved in the flights data, such as their geographic locations, which can be used for visualizations or calculations. The "flights5" table may be used for analyzing flight-related information, such as airline performance, flight delays, or route analysis, using PySpark's data processing capabilities.

Finally, the airline's dataset includes two columns: IATA_CODE and AIRLINE. The IATA_CODE column contains the airline's IATA code, and the AIRLINE column contains the name of the airline. The data in this table appears to represent a list of airlines with their corresponding IATA codes and names.

ii. Data Loading: The data loading process begins by importing the necessary libraries, including PySpark's SQL module, which provides functions for reading and processing structured data. Then make use of the read method from the SQL module to create a DataFrame, which is a distributed collection of data organized into named columns. The read method takes various parameters, such as the file format, file path, and options for handling headers, delimiters, and encoding.

iii. Data Analysis: Once the data is loaded into a DataFrame, then it demonstrates how to perform basic data exploration tasks, such as displaying the first few rows of data using the show method, checking the schema using the print Schema method, and obtaining summary statistics using the describe method. The author also discusses how to handle missing data, including using PySpark's built-in functions for imputation and filtering out rows with missing values. After loading the data, it performs basic data exploration tasks using PySpark's built-in functions. This can include inspecting the data schema to understand the structure of the data, checking for missing values, and understanding the statistical properties of the data.

iv. Data cleaning: This is a crucial step in the data analysis process, and the notebook may illustrate how to clean the flights data using PySpark's data manipulation functions. This can involve handling missing values, filtering out irrelevant data, and transforming the data into a more suitable format for analysis.

Incorporating the logistic regression estimator, the cleaned data can be used to build a predictive model. The notebook may demonstrate how to use PySpark's Logistic Regression estimator to create a logistic regression model and tune it using k-fold cross validation. This can involve selecting optimal hyperparameters, such as elasticNetParam and regParam, to improve the performance of the model.

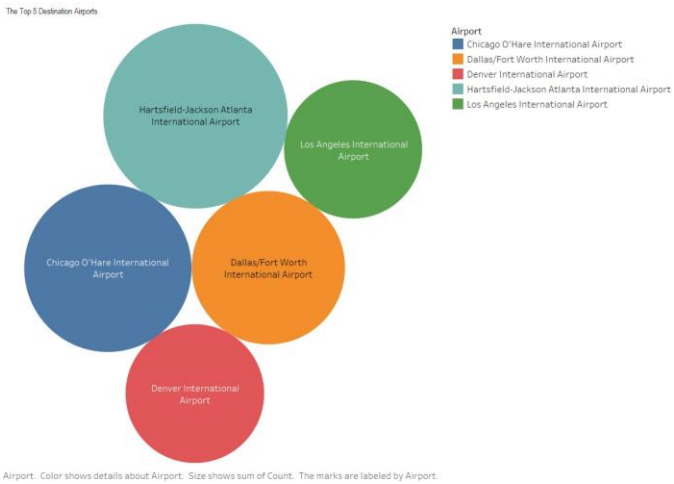
v. Data Visualization: Once the data has been processed and analyzed using Spark, the results can be seamlessly integrated into Tableau for visualizing and analyzing the data in a more user-friendly and interactive manner. Tableau, a powerful data visualization tool, allows users to create visually appealing and insightful visualizations from the output data obtained from Spark. The data can be imported into Tableau directly through a Spark connection.

Therefore, the combination of PySpark's robust data processing capabilities and Tableau's powerful visualization tools enables effective data analysis and communication of insights from the Flights Data, leading to informed decision-making and actionable recommendations.

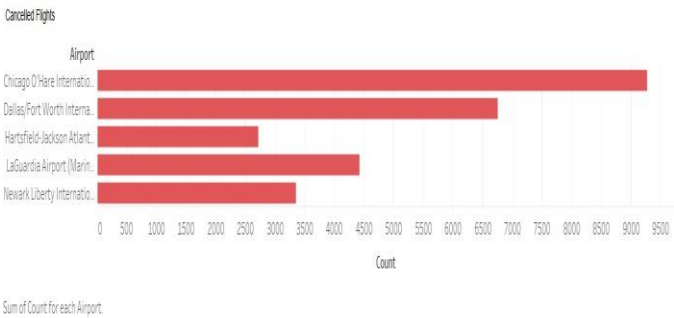
4. RESULTS AND DISCUSSIONS

Tableau is a powerful tool for visualizing the results of data analysis performed using PySpark. It allows us to create a wide range of interactive and visually appealing visualizations to explore and communicate the insights gained from the data analysis process. Data visualization is a valuable tool for enhancing the visual representation of datasets. It involves leveraging questions and their corresponding graphical results to provide a more compelling and informative view of the data.

1. Which airports were ranked in the top 5 for both origin and destination flights with the highest total number of flights in the year 2015?



2. What were the top 5 origin airports with the highest number of cancelled flights in the year 2015?



3. Which origin airports were ranked in the top 5 for the highest percentage of cancelled flights in 2015, considering the percentage in relation to the total flights departing from each respective airport?

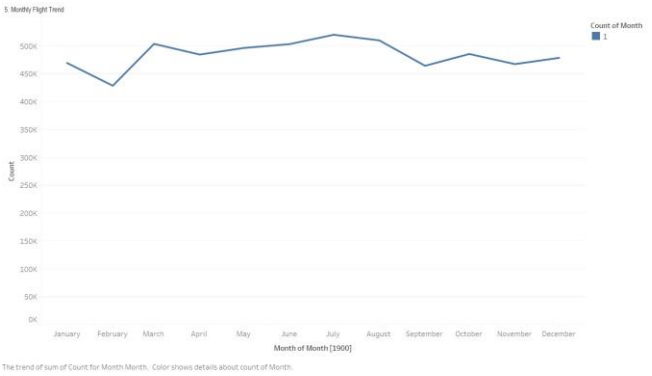
Highest Cancellation in Percentage %

Airport	Round(Percentage, 4)
Adak Airport	7.292
Devils Lake Regional Airport	7.795
Ithaca Tompkins Regional Airport	11.429
Mammoth Yosemite Airport	10.256
Muskegon County Airport	7.196

Round(Percentage, 4): 7.196 – 11.429

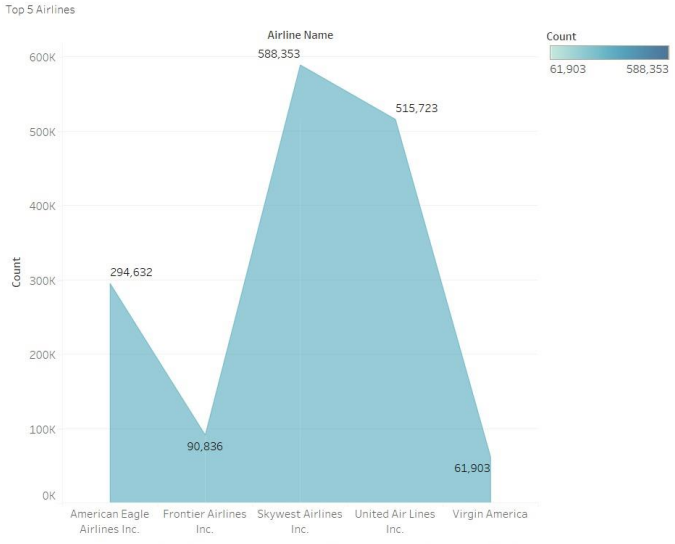
Sum of Round(Percentage, 4) broken down by Airport. Color shows sum of Round(Percentage, 4). The marks are labeled by sum of Round(Percentage, 4).

4. How many flights were recorded for each month listed in the Flights data frame in 2015?



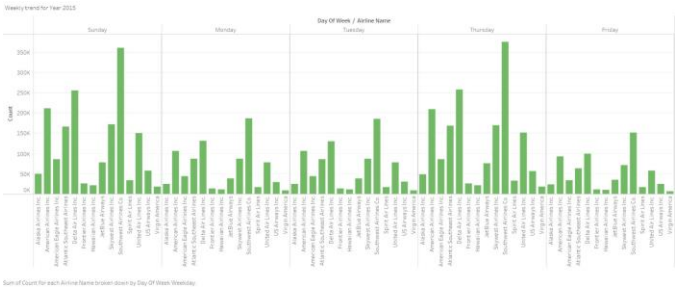
The trend of sum of Count for Month Month. Color shows details about count of Month.

5. Which airlines were ranked in the top 5 based on the number of flights operated in the year 2015?



Sum of Count for each Airline Name. Color shows sum of Count. The marks are labeled by sum of Count.

6. What was the total number of flights for each day of the week during the year 2015?

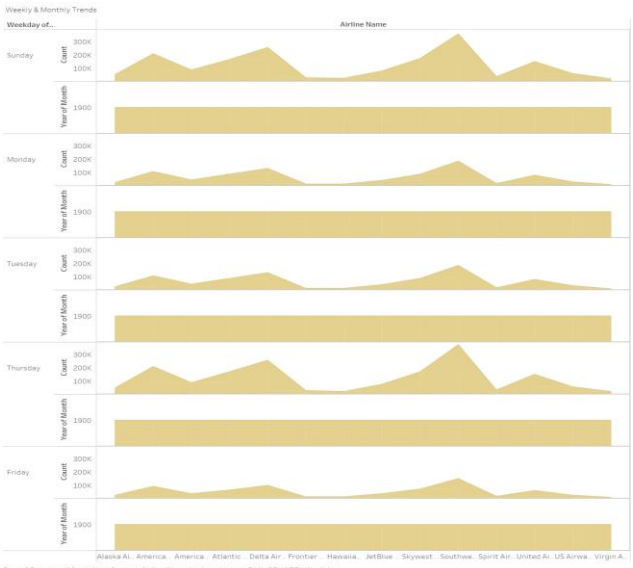


Sum of Count for each Airline Name broken down by Day Of Week Weekday.

7. How many flights were recorded for each airline in each month of 2015?

Sum of Count for each Month Month broken down by Airline Name.

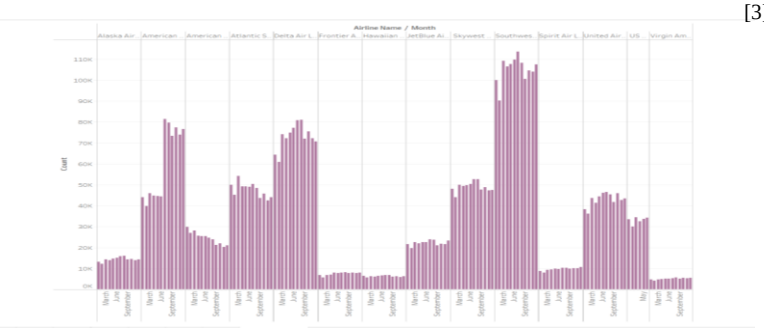
8. How many flights were recorded for each day of the week during each month of 2015?



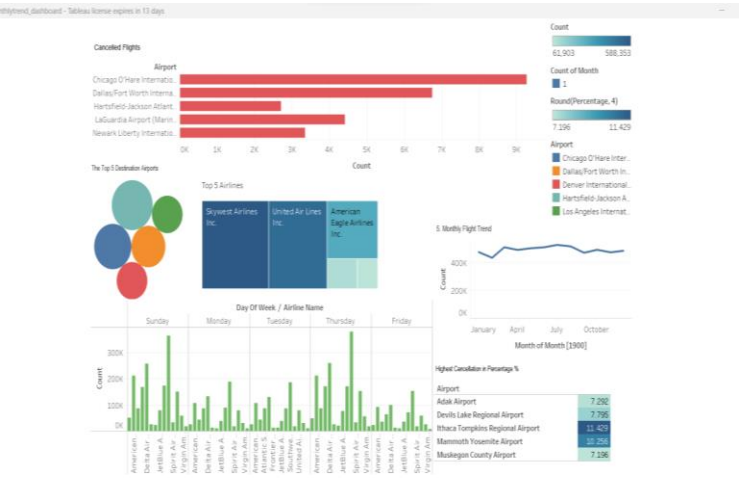
Sum of Count and Month Year for each Airline Name broken down by DAY_OF_WEEK Weekday.

Tableau Dashboard Visualization

We created 2 Dashboards to extract key insights from the data that has been explained above –



Complete Dashboard View -



5. CONCLUSIONS

In conclusion, this project involves data loading, data exploration, and data cleaning using PySpark's built-in functions. It demonstrates how to inspect data schema, handle missing values, filter out irrelevant data, and transform data for analysis. The project also includes the usage of PySpark's Logistic Regression estimator for building a logistic regression model to predict if there is any delay in flights. Overall, the Flight project provides a comprehensive introduction to PySpark and its application in analyzing flights data for insights and decision-making in the airline industry.

One key finding from the results of data analysis using Spark and visualized with Tableau is the importance of data visualization in enhancing the understanding of datasets. The top 5 origin and destination airports for total flights in 2015, the airports with the highest number of cancelled flights, and the airlines with the most flights operated in 2015 were among the notable insights gained. These findings highlight the significance of leveraging data visualization techniques to uncover meaningful insights and communicate them effectively.

6. REFERENCES

[1] "Introduction to PySpark using FlightsData" by Asrsaiteja, Kaggle, [online]. Available: https://www.kaggle.com/code/asrsaiteja/introduction-to-pyspark-using-flightsdata/notebook. Accessed: April 11, 2023.

[2] "Flight Delay Data Analysis" by Mouneim, Kaggle, [online]. Available: https://www.kaggle.com/code/mouneim/flight-delay-data-analysis/notebook. Accessed: April 11, 2023.

[3] "Flight Analysis" by LifeIsAFile, [online]. Available: http://www.lifeisafile.com/flight-analysis/. Accessed: April 11, 2023.