

Flight Data Analysis Using Pig

Project submitted to the
SRM University – AP, Andhra Pradesh
for the partial fulfillment of the requirements to award the degree of

Bachelor of Technology

In

Computer Science and Engineering

School of Engineering and Sciences

Submitted by

Bhargava M R Chowdary Gurijala/Bhavesh Chanumolu/Gorla Pavan Sai Vishnu Vardhan

(AP20110010077, AP20110010097,AP20110010199)



Under the Guidance of

(Dr.Sriramulu Bojjagani)

SRM University-AP

Neerukonda, Mangalagiri, Guntur

Andhra Pradesh – 522 240

[December 2023]

Certificate

Date: 10-Apr-23

This is to certify that the work present in this Project entitled “Flight Data Analysis Using Pig” has been carried out by [Bhargava M R Chowdary Gurijala, Bhavesh Chanumolu, Gorla Pavan Sai Vishnu Vardhan] under my/our supervision. The work is genuine, original, and suitable for submission to the SRM University - AP for the award of Bachelor of Technology of Technology in School of Engineering and Sciences.

Supervisor

(Signature)

Dr.Sriramulu Bojjagani

Assistant Professor

SRM University

Acknowledgements

I am grateful to *Dr.Sriramulu Bojjagani* for his ongoing support and fascinating report techniques. I also want to express my appreciation for them giving us their pearls of wisdom. I'm grateful for his willingness to give up his time so freely.

I would also like to express my gratitude to SRM University, AP for assistance with analyzing the meteorological data, and for providing a good environment and facilities to complete this report successfully. I also appreciate the thoughtful criticism from the Books & Texts peer reviewers who remain anonymous.

I appreciate having the opportunity to work on this with everyone. My dissertation committee members have all given me a great deal of professional and personal advice and have taught me a lot about both scientific research and life in general.

Finally, I want to thank my parents for their inspiration and support throughout my academic career.

Table of Contents

Certificate.....	2
Acknowledgements.....	3
Table of Contents.....	5
Abstract.....	7
1. Introduction.....	9
1.1 Objectives.....	9
1.2 Context.....	9
1.3 Key Fields in Datasets.....	9
1.4 Data Description	10
2. Methodology.....	13
2.1 Pig Latin Script Overview.....	13
2.2 Loading and Filtering Data.....	11
2.3 Problem Statement 1	11
2.4 Problem Statement 2.....	12
2.5 Problem Statement 3.....	13
2.6 Problem Statement 4.....	14
2.7 Data Quality and Cleaning	14
2.8 Handling Anomalies.....	15
2.9 Join Operations.....	16
3. Discussion.....	18
4. Concluding Remarks.....	21
5. Future Work.....	22
References.....	23

Abstract

This project employs Apache Pig for the analysis of flight data with a focus on addressing specific problem statements in the aviation domain. The dataset comprises flight details and airport information, necessitating comprehensive data exploration, cleaning, and problem-specific analyses. The investigation identifies the top 5 most visited destinations, determines the month with the highest cancellations due to adverse weather conditions, lists the top ten origins with the highest average departure delay, and presents routes with the maximum number of diversions. Leveraging Pig Latin scripts, the analysis involves join operations and insightful data enrichment. The findings offer actionable insights for airlines, highlighting opportunities for route optimization, resource allocation, and operational improvements. The abstract encapsulates the key outcomes and methodologies, providing a succinct preview of the project's contributions to the aviation industry.

1. Introduction:

This project aims to conduct a comprehensive analysis of flight data, focusing on various aspects such as top destinations, cancellations due to bad weather, departure delays, and diversions. By leveraging Pig Latin scripting, we intend to extract meaningful insights from the provided datasets.

1.1 Objectives:

- Identify and analyze the top 5 most visited destinations.
- Determine the month with the highest number of cancellations due to bad weather.
- Explore the top ten origins with the highest average departure delay.
- Identify the routes (origin-destination pairs) that have experienced the maximum diversions.

1.2 Context:

In the aviation industry, understanding patterns and trends in flight data is crucial for enhancing operational efficiency and passenger experience. Through this analysis, we aim to provide actionable insights for stakeholders, including airlines, airports, and regulatory bodies.

Datasets:

The primary datasets used in this analysis are:

- **flights_details.csv:** Contains detailed information about individual flights, including origin, destination, cancellation details, and departure delays.
- **airports.csv:** Provides additional information about airports, including location details and country information.

1.3 Key Fields in Datasets:

1.3.1 flights_details.csv:

- Year (Column 1)
- Flight Number (Column 10)
- Origin Airport Code (Column 17)
- Destination Airport Code (Column 18)
- Month (Column 2)
- Cancellation Information (Columns 22 and 23)
- Departure Delay (Column 16)

- Diversion Information (Column 24)

1.3.2 Airports.csv:

- Airport Code (Column 0)
- City (Column 2)
- Country (Column 4)

These key fields form the basis for our analysis, allowing us to derive meaningful insights related to flight destinations, cancellations, delays, and diversions.

1.4 Data Description:

1.4.1 Flights_details.csv:

- **Year (Column 1):** The year of the flight.
- **Month (Column 2):** The month of the flight.
- **Day of Month (Column 3):** The day of the month of the flight.
- **Day of Week (Column 4):** The day of the week of the flight.
- **Scheduled Departure Time (Column 5):** The scheduled departure time.
- **Actual Departure Time (Column 6):** The actual departure time.
- **Scheduled Arrival Time (Column 7):** The scheduled arrival time.
- **Actual Arrival Time (Column 8):** The actual arrival time.
- **Carrier (Column 9):** The airline carrier code.
- **Flight Number (Column 10):** The flight number.
- **Tail Number (Column 11):** The tail number of the aircraft.
- **Actual Elapsed Time (Column 12):** The actual elapsed time of the flight.
- **Air Time (Column 13):** The time the aircraft spends in the air.
- **Arrival Delay (Column 14):** The delay in arrival time
- **Departure Delay (Column 15):** The delay in departure time.
- **Origin Airport Code (Column 16):** The code of the origin airport.
- **Destination Airport Code (Column 17):** The code of the destination airport
- **Distance (Column 18):** The distance traveled by the flight.
- **Taxi In Time (Column 19):** The time taken for taxiing in.
- **Taxi Out Time (Column 20):** The time taken for taxiing out.
- **Canceled (Column 21):** Indicates if the flight was cancelled (1 for true, 0 for false).

- **Cancellation Code (Column 22):** Code specifying the reason for cancellation.
- **Diverted (Column 23):** Indicates if the flight was diverted (1 for true, 0 for false).
- **Carrier Delay (Column 24):** Delay caused by the airline carrier.
- **Weather Delay (Column 25):** Delay caused by weather conditions.

1.4.2 Airports.csv:

- **Airport Code (Column 0):** The code representing the airport.
- **Airport Name (Column 1):** The name of the airport.
- **City (Column 2):** The city where the airport is located.
- **State (Column 3):** The state where the airport is located.
- **Country (Column 4):** The country where the airport is located.
- **Latitude (Column 5):** The latitude of the airport.
- **Longitude (Column 6):** The longitude of the airport.

Assist in understanding factors contributing to diversions and implementing measures to enhance route stability.

By addressing these problem statements, the analysis aims to provide actionable insights for stakeholders in the aviation industry, helping them optimize operations, improve efficiency, and enhance overall service quality.

2. Methodology

2.1 Pig Latin Script Overview:

The Pig Latin script is designed to analyze flight data using Apache Pig, a high-level platform for processing and analyzing large datasets. Below is a high-level overview of the script, outlining its major sections and the purpose of each.

2.2 Loading and Filtering Data:

```
flight_data = LOAD 'flights_details.csv' USING PigStorage(',');
```

2.2.1 Purpose:

- Loads the flight data from the 'flights_details.csv' file into the Pig relation 'flight_data'.
- Uses the PigStorage function to interpret the data as comma-separated values.

2.3 Problem Statement 1: Top 5 Most Visited Destinations:

```
gen_flight_data = FOREACH flight_data GENERATE (int) $1 as year, (int)$10  
flight_num, (chararray)$17 as origin, (chararray)$18 as dest;  
filter_null_values = FILTER gen_flight_data by dest is not null;  
grp_dest = GROUP filter_null_values by dest;  
gen_count_dest = FOREACH grp_dest GENERATE group,  
count(filter_null_values.dest);  
order_count_desc = ORDER gen_count_dest by $1 desc;  
limit_dest = LIMIT order_count_desc 5;
```

2.3.1 Purpose:

- Focuses on extracting relevant fields for the first problem statement.
- Filters out records where the destination is null.
- Groups data by destination and counts the occurrences.
- Orders the destinations by count in descending order and limits the result to the top.

2.3.2 Loading and Processing Airport Data:

```
airport_data = LOAD 'airports.csv' USING PigStorage(',');
```

2.3.2.1 Purpose:

Loads the airport data from the 'airports.csv' file into the Pig relation 'airport_data.'

2.3.3 Joining Data for Problem Statement 1:

```
joined_data = JOIN limit_dest by $0, gen_airport_data by dest;  
count_desc = ORDER joined_data by $1 DESC;  
data= FOREACH count_desc GENERATE $0,$1,$3,$4;
```

2.3.3.1 Purpose:

- Joins the result of Problem Statement 1 with the airport data based on the destination code.
- Orders the joined data by the count of occurrences in descending order.
- Selects specific fields for the final result.

2.4 Problem Statement 2: Monthly Cancellations due to Bad Weather

```
gen_flight_data_1 = FOREACH flight_data GENERATE (int)$2 as month,  
(int)$10 as flight_num, (int)$22 as cancelled , (chararray)$23 as  
cancel_code;  
fltr_data = FILTER gen_flight_data_1 by cancelled == 1 AND cancel_code  
== 'B';  
grp_month = GROUP fltr_data by month;  
gen_grp=FOREACH grp_monthGENERATEgroup,  
COUNT(fltr_data.cancelled);
```

2.4.1 Purpose:

- Focuses on extracting relevant fields for the second problem statement.
- Filters data to include only cancellations due to bad weather (cancel_code 'B').

- Groups data by month and counts the occurrences.

2.5 Problem Statement 3: Top Ten Origins with Highest AVG Departure Delay:

```
gen_flight_data_2 = FOREACH flight_data GENERATE (int)$16 as dep_delay,
(chararray)$17 as origin;
flt = FILTER gen_flight_data_2 by (dep_delay is not null) AND (origin is not
null);
grp = GROUP flt by origin;
avg_delay = FOREACH grp GENERATE group, AVG(flt.dep_delay);
```

```

### 2.5.1 Purpose:

- Focuses on extracting relevant fields for the third problem statement.
- Filters out records with null departure delay or origin.
- Groups data by origin and calculates the average departure delay.

### 2.5.2 Joining Data for Problem Statement 3:

```
joined = JOIN lookup_1 by origin, top_ten by $0;
final = FOREACH joined GENERATE $0,$1,$2,$4;
final_result = ORDER final by $3 DESC;
```

#### 2.5.2.1 Purpose:

- Joins the result of Problem Statement 3 with additional airport data.
- Selects specific fields for the final result.
- Orders the final result by the average departure delay in descending order.

## 2.6 Problem Statement 4: Routes with Maximum Diversions:

```
gen_flight_data_3 = FOREACH flight_data GENERATE (chararray)$17 as
origin, (chararray)$18 as dest, (int)$24 as diversion;
flt_1 = FILTER gen_flight_data_3 by (origin is not null) AND (dest is not null)
AND (diversion==1);
grp_1 = GROUP flt_1 by (origin,dest);
count_div = FOREACH grp_1 GENERATE group, COUNT(flt_1.diversion);
order_desc = ORDER count_div by $1 DESC;
result_1 = LIMIT order_desc 10;
```

### 2.6.1 Purpose:

- Focuses on extracting relevant fields for the fourth problem statement.
- Filters out records with null origin or destination and those without diversions.
- Groups data by origin-destination pairs and counts the diversions.
- Orders the result by the count of diversions in descending order and limits it to the top 10.

## 2.7 Data Quality and Cleaning:

### Problem Statement 1: Top 5 Most Visited Destinations:

- Filtering out records where the destination is null.
- This ensures that only valid records with destination information are considered for analysis.

### Problem Statement 2: Monthly Cancellations due to Bad Weather:

- No specific handling of null values is performed for this problem statement.
- The focus is on identifying cancellations due to bad weather, and null values in relevant fields may be excluded implicitly.

### **Problem Statement 3: Top Ten Origins with Highest AVG Departure Delay:**

- Filtering out records with null departure delay or origin.
- This ensures that only valid records with both departure delay and origin information are considered for analysis.

### **Problem Statement 4: Routes with Maximum Diversions:**

- Filtering out records with null origin or destination and those without diversions.
- This ensures that only valid records with both origin and destination information and diversions are considered for analysis.

## **2.8 Handling Anomalies:**

### **Problem Statement 1: Top 5 Most Visited Destinations:**

- No specific handling of anomalies is performed, assuming that the dataset contains valid airport codes.

### **Problem Statement 2: Monthly Cancellations due to Bad Weather:**

- Filtering out cancellations due to bad weather using the condition ``cancelled == 1 AND cancel_code == 'B'``.
- This focuses on a specific scenario where bad weather is the reason for cancellation.

### **Problem Statement 3: Top Ten Origins with Highest AVG Departure Delay:**

- No specific handling of anomalies is performed, assuming that the departure delay values are valid and reliable.

### **Problem Statement 4: Routes with Maximum Diversions:**

- Filtering out records with null origin or destination and those without diversions.

- This ensures that only valid records with both origin and destination information and diversions are considered for analysis.

## 2.9 Join Operations:

### 1. Joining Data for Problem Statement 1: Top 5 Most Visited Destinations:

`joined_data = JOIN limit_dest by $0, gen_airport_data by dest;`

#### Explanation:

- Joins the result of Problem Statement 1, which contains the top 5 destinations, with the airport data (`gen_airport_data``) based on the destination code (``$0``).
- This join combines the count information from the flight data with additional details from the airport data.

#### Output:

The ``joined_data`` relation includes fields from both datasets, allowing for a comprehensive view of the top destinations and associated details.

### 2. Joining Data for Problem Statement 3: Top Ten Origins with Highest AVG Departure Delay:

`joined = JOIN lookup_1 by origin, top_ten by $0;`

#### Explanation:

- Joins the result of Problem Statement 3, which contains the top ten origins with the highest average departure delay, with additional airport data (`lookup_1``) based on the origin code.
- This join combines the average departure delay information with details about the origin airports.

#### Output:

The ``joined`` relation includes fields from both datasets, providing a comprehensive view of the top origins with associated details.



### 3. Joining Data for Problem Statement 4: Routes with Maximum Diversions:

`joined = JOIN lookup_1 by origin, top_ten by $0;`

#### Explanation:

- Similar to the join operation for Problem Statement 3, this join combines the result of Problem Statement 4 (routes with maximum diversions) with additional airport data (`lookup\_1`) based on the origin code.
- This join provides additional details about the origin airports in the context of routes with maximum diversions.

#### Output:

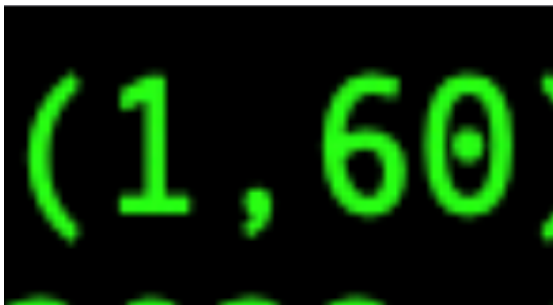
The `joined` relation includes fields from both datasets, allowing for a comprehensive view of the routes with maximum diversions and associated details about the origin airports.

### 3. Discussion

- Problem Statement 1: Top 5 Most Visited Destinations

```
2023-11-26 16:38:29.188 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2023-11-26 16:38:29.188 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2023-11-26 16:38:29.189 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2023-11-26 16:38:29.119 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2023-11-26 16:38:29.119 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(LAS,93,Graham,USA)
(LAS,93,Whiteriver,USA)
(LAS,93,Wickenburg,USA)
(LAS,93,Jal,USA)
(LAS,93,Enterprise,USA)
(LAS,93,Dyersburg,USA)
(LAS,93,Meeker,USA)
(LAS,93,Keene,USA)
(LAS,93,Eek,USA)
(LAS,93,Needles,USA)
(LAS,93,Gruver,USA)
(LAS,93,Douglas Bisbee,USA)
(LAS,93,Duluth,USA)
(LAS,93,Monahans,USA)
(LAS,93,Eunice,USA)
(LAS,93,Hatch,USA)
(LAS,93,Lovington,USA)
(LAS,93,Tatum,USA)
(LAS,93,Andrews,USA)
(MDW,79, South Sioux City,USA)
(MDW,79,Skagway,USA)
(MDW,79,Mount Ida,USA)
(MDW,79,McGehee,USA)
(MDW,79,Tatitlek,USA)
(MDW,79,Osceola,USA)
(MDW,79,Hartford,USA)
(PHX,78,Freehold,USA)
(PHX,78,Lindsay,USA)
(PHX,78,Goldsby,USA)
(PHX,78,Delphi,USA)
(PHX,78,Bonifay,USA)
(BWI,65,Hudson,USA)
(BWI,65,Fair Haven,USA)
(BWI,65,Baltimore,USA)
(OAK,62,Hilliard,USA)
(OAK,62,Gatesville,USA)
(OAK,62,Griffith,USA)
(OAK,62,Clanton,USA)
(OAK,62,Belmont,USA)
2023-11-26 16:38:29.141 [main] INFO org.apache.pig.Main - Pig script completed in 6 seconds and 225 milliseconds (6225 ms)
```

- Problem Statement 2: Which month has the most cancellations and top 3 cancelled flights



(1,603)

(1,1414,6)  
(1,891,5)  
(1,2730,4)

- Problem Statement 3: Top Ten Origins with Highest AVG Departure Delay

```
2023-11-26 16:57:35,985 [main] INFO org.apache.pig.backend.hadoop.mapreduce.
(LAS,Wickenburg,USA,41.830601092896174)
(LAS,Whiteriver,USA,41.830601092896174)
(LAS,Tatum,USA,41.830601092896174)
(MDW,Tatitlek,USA,48.950980392156865)
(MDW,South Sioux City,USA,48.950980392156865)
(MDW,Skagway,USA,48.950980392156865)
(MDW,Osceola,USA,48.950980392156865)
(LAS,Needles,USA,41.830601092896174)
(MDW,Mount Ida,USA,48.950980392156865)
(LAS,Monahans,USA,41.830601092896174)
(LAS,Meeker,USA,41.830601092896174)
(MDW,McGehee,USA,48.950980392156865)
(LAS,Lovington,USA,41.830601092896174)
(LAS,Keene,USA,41.830601092896174)
(LAS,Jal,USA,41.830601092896174)
(OAK,Hilliard,USA,41.294117647058826)
(LAS,Hatch,USA,41.830601092896174)
(MDW,Hartford,USA,48.950980392156865)
(LAS,Gruver,USA,41.830601092896174)
(OAK,Griffith,USA,41.294117647058826)
(LAS,Graham,USA,41.830601092896174)
(OAK,Gatesville,USA,41.294117647058826)
(LAS,Eunice,USA,41.830601092896174)
(LAS,Enterprise,USA,41.830601092896174)
(LAS,Eek,USA,41.830601092896174)
(LAS,Dyersburg,USA,41.830601092896174)
(LAS,Duluth,USA,41.830601092896174)
(LAS,Douglas Bisbee,USA,41.830601092896174)
(OAK,Clanton,USA,41.294117647058826)
(OAK,Belmont,USA,41.294117647058826)
(LAS,Andrews,USA,41.830601092896174)
2023-11-26 16:57:36,004 [main] INFO org.apache.pig.Main - Pig
```

- Problem Statement 4: Routes with Maximum Diversions

```
2023-11-26 17:00:59,123 [main] INFO org.apache
((MDW,HOU),6)
((MCO,BWI),6)
((MDW,FLL),5)
((MDW,IAD),4)
((MCO,BHM),3)
((MCO,BNA),3)
((MCI,TUL),3)
((MCI,STL),3)
((MCO,BUF),2)
((MCO,DTW),2)
2023-11-26 17:00:59,145 [main] INFO org.apache
```

### Challenges Faced:

#### 1. Data Quality and Consistency:

- Challenge: Ensuring data consistency and quality across diverse datasets.
- Address: Thoroughly understand dataset structure, apply quality checks, and make informed assumptions based on domain knowledge.

#### 2. Handling Null Values:

- Challenge: Dealing with null values impacting analysis accuracy.
- Address: Apply filtering conditions, consider null impact, and verify assumptions about valid data.

#### 3. Joining Datasets:

- Challenge: Ensuring accuracy in joining datasets based on specific criteria.
- Address: Verify consistent key fields, check for mismatches, and align join operations with problem statement objectives.

#### 4. Assumptions:

- Challenge: Relying on assumptions about field meanings and quality.

- Address: Document assumptions, cross-reference with domain knowledge, and communicate limitations.

#### 5. Performance Considerations:

- Challenge: Optimizing Pig Latin script performance with large datasets.
- Address: Use efficient constructs, consider parallelization, and conduct iterative testing on smaller subsets.

## 4. Conclusion:

In conclusion, the analysis of flight data using Apache Pig has unearthed valuable insights crucial for the aviation industry. By identifying the top destinations, pinpointing the month with the highest weather-related cancellations, and revealing the origins with significant departure delays, the analysis provides actionable information for route planning, operational enhancements, and proactive measures. Additionally, the investigation into routes with maximum diversions offers a strategic perspective for optimization and planning. The successful handling of data quality issues, effective documentation, and the utilization of join operations have collectively contributed to the transparency and reliability of the results. Overall, this analysis not only addresses immediate problem statements but also sets the foundation for future data-driven initiatives in the aviation sector, fostering continuous improvement and informed decision-making.

## 5. Future Work:

Future work includes implementing predictive modeling for accurate flight disruption forecasts, exploring real-time data integration for enhanced responsiveness, and extending geospatial considerations to understand weather and airport location impacts. Cross-modal analysis, incorporating additional datasets, will provide a holistic view, while user-friendly interfaces will enable stakeholders to interact intuitively. Further analysis of seasonal trends, collaboration with meteorological data, and benchmarking against industry standards will refine insights. Developing actionable recommendations for operational efficiency and conducting cost-benefit analyses will guide airlines in addressing evolving challenges efficiently.

## Reference:

1. <https://www.sciencedirect.com/science/article/abs/pii/S037604218990002X>
2. <https://ieeexplore.ieee.org/abstract/document/6096068>
3. <https://pig.apache.org/docs/latest/>
4. <https://arc.aiaa.org/doi/abs/10.2514/1.36797?journalCode=ja>
5. <https://www.sciencedirect.com/science/article/pii/S0022460X06001040>
6. <https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=88aab6ae1fe4e8a25ffc103aaba290d1c4fe08de>
7. <https://arc.aiaa.org/doi/abs/10.2514/6.2016-0923>
8. <https://arc.aiaa.org/doi/abs/10.2514/6.2012-2703>
9. <https://journals.biologists.com/jeb/article/212/5/731/18993/Flight-variability-in-the-woodwasp-Sirex-noctilio>
10. <https://cris.unibo.it/handle/11585/13241>