AIRLINES DATA ANALYTICS IN AVAITION INDUSTRY

NALAIYA THIRAN PROJECT BASED LEARNING

On

PROFESSIONAL READINESS FOR INNOVATION, EMPLOYABILITY AND ENTREPRENEURSHIP

A PROJECT REPORT

TEAM ID: PNT2022TMID09956

DHIFANI STENIKSHA D	19104048
BARATH S	19104033
BHUVANESH S	19104036
GOKUL RAJ	19104059

BACHELOR OF ENGINEERING

IN

COMPUTER SCIENCE AND ENGINEERING

HINDUSTHAN COLLEGE OF ENGINEERING AND TECHOLOGY

Approved by AICTE, New Delhi, Accredited with 'A' Grade by NAAC

(An Autonomous Institution, Affiliated to Anna University, Chennai)

COIMBATORE – 641 032

November 2022

TABLE OF CONTENT

ABSTRACT

In the contemporary world, Data analysis is a challenge in the era of varied inters- disciplines though there is a specialization in the respective disciplines. In other words, effective data analytics helps in analyzing the data of any business system. But it is the big data which helps and axial rates the process of analysis of data paving way for a success of any business intelligence system. With the expansion of the industry, the data of the industry also expands. Then, it is increasingly difficult to handle huge amount of data that gets generated no matter what‘s the business is like, range of fields from social media to finance, flight data, environment and health.

An Airport has huge amount of data related to number of flights, data and time of arrival and dispatch, flight routes, No. of airports operating in each country, list of active airlines in each country. The problem they faced till now it‘s, they have ability to analyze limited data from databases.

How can it be gathered, stored, processed and analyzed it to turn the raw data information to support decision making. In this paper Big Data is depicted in a form of case study for Airline data.

1.INTRODUCTION

1.1. PROJECT OVERVIEW

Researchers working in the structured data face many challenges in analyzing the data. For in_stance the data created through social media, in blogs, in Facebook posts or Snap chat. These types of data have different structures and formats and are more difficult to store in a traditional business data base. The data in big data comes in all shapes and formats including structured. Working with big data means handling a variety of data formats and structures. Big data can be a data created from sensors which track the movement of objects or changes in the environment such as temperature fluctuations or astronomy data. In the world of the internet of things, where devices are connected and these wearables create huge volume of data. Thus big data approaches are used to manage and analyze this kind of data. Big Data include data from a whole range of fields such as flight data, population data, financial and health data such data brings as to another V, value which has been proposed by a number of researcher i.e., Veracity.

Most of the time social media is analyzed by advertisers and used to promote produces and events but big data has many other uses. It can also been used to assess risk in the insurance industry and to track reaction to products in real time. Big Data is also used to monitor things as diverse as wave movements, flight data, traffic data, financial transactions, health and crime. The challenge of Big Data is how to use it to create something that is value to the user. How to gather it, store it, process it and analyze it to turn the raw data information to support decision making.

An Airport has huge amount of data related to number of flights, data and time of arrival and dispatch, flight routes, No. of airports operating in each country, list of active airlines in each country. The problem they faced till now it‘s, they have ability to analyze limited data from databases. The Proposed model intension is to develop a model for the airline data to provide platform for new analytics based on the following queries.

1.2.Purpose

The main purpose of the project to explore detailed analysis on airline data sets such as listing airports operating in the India, list of airlines having zero stops, list of airlines operating with code share which country has highest airports and list of active airlines in united states. The main objective of project is the processing the big data sets using map reduce component of hadoop ecosystem in distributed environment.

2.LITERATURE SURVEY

1.Literature focusing on crew recovery

analytics Author : M. Selim Aktürk, Alper Atamtürk, Sinan Gürel

https://www.sciencedirect.com/science/article/pii/S0305054820302549#bb0020

The crew recovery problem (CRP) can be formulated as follows: given a flight schedule and a set of disruptions, re-assign to each (recovered) flight the necessary cabin and flight crew such that the disruption costs are minimized. For crew recovery, these disruption costs can include direct crew costs (e.g., remuneration or overtime compensation) and cost for deadheading crew. For studies that include flight cancellation as a recovery action, cancellation costs can be included in case a flight cannot be staffed. Alternatively, some authors opt to use minimizing the number of crew schedule changes as a proxy to the minimization of the crew recovery costs. The CRP is typically the second problem that is solved in the sequential solution approach. It is considered harder than the ARP since all regulations and restrictions dictated by government regulations, union agreements and airline-specific policies have to be taken into account.

2.Literature focusing on passenger recovery

Author : Bruno Aguiar, Jose Torres, António J M Castro

https://www.sciencedirect.com/science/article/pii/S0305054820302549#bb0010

Arguably, passenger recovery is the most relevant problem for airline disruption management since high passenger delay cost and continuous flight disruptions will lead to a potential loss of goodwill and long-term reputation damage. Passenger recovery can be formulated as follows: given a recovered flight and crew schedule and a set of disrupted passenger itineraries, re-assign to each disrupted itinerary the (recovered) flights necessary (given seat availability) to accommodate passengers from their current position to their destination while minimizing cost. These passenger recovery costs can include both hard and soft costs. Hard costs are directly incurred when a passenger cannot complete its scheduled itinerary (e.g., compensation for delay and cancellation as stipulated by government regulations). Soft costs are the potential losses of future revenue as a result of passenger inconvenience, possibly causing the passenger to switch to a different airline in the future. These costs are approximations made by the airline and can differ per passenger class or frequent flyer status. Alternatively, these passenger disruption costs are minimized by minimizing the total number of passenger delay minutes.

3.Literature focusing on integrated recovery

Industry Author : Khaled F. Abdelghany, Ahmed F. Abdelghany, Goutham Ekollu

https://www.sciencedirect.com/science/article/pii/S0305054820302549#bb0005

Both from a mathematical and computational perspective, the integration of all recovery stages (aircraft, crew, and passengers) is a difficult task. The purpose of this integration is to minimize the total disruption cost. This is achieved by weighing the disruption cost related to aircraft, crew, and passengers simultaneously to find the recovery solution that overall results in the lowest cost for the airline. To the best of the authors' knowledge, the first proposal of a truly integrated approach was the PhD Thesis of Lettovsky (1997), where the author formulated the 'Airline Integrated Recovery' problem which consists of aircraft routing, crew assignment, and passenger flow. The thesis presents a linear mixed-integer mathematical problem that captures the availability of the aforementioned resources. A decomposition scheme is presented where the 'Schedule Recovery Model' master problem controls the three sub-problems known as the ‘Aircraft recovery model’, ‘Crew recovery model’, and ‘Passenger flow model’. The solution is derived by applying Benders’ decomposition. A limitation is that the model only considers the cockpit crew and not cabin crew

2.1.EXISTING PROBLEM

Airline data analysis can provide a solution for businesses to collect and optimize large datasets, improve performance, improve their competitive advantage, and make faster and better decisions.

- By using airline data analysis, we can save time of users.
- The data could even be structured, semi-structured or unstructured.
- Cost savings
- Implementing new strategies
- Fraud can be detected the moment it happens

2.2. REFERENCES

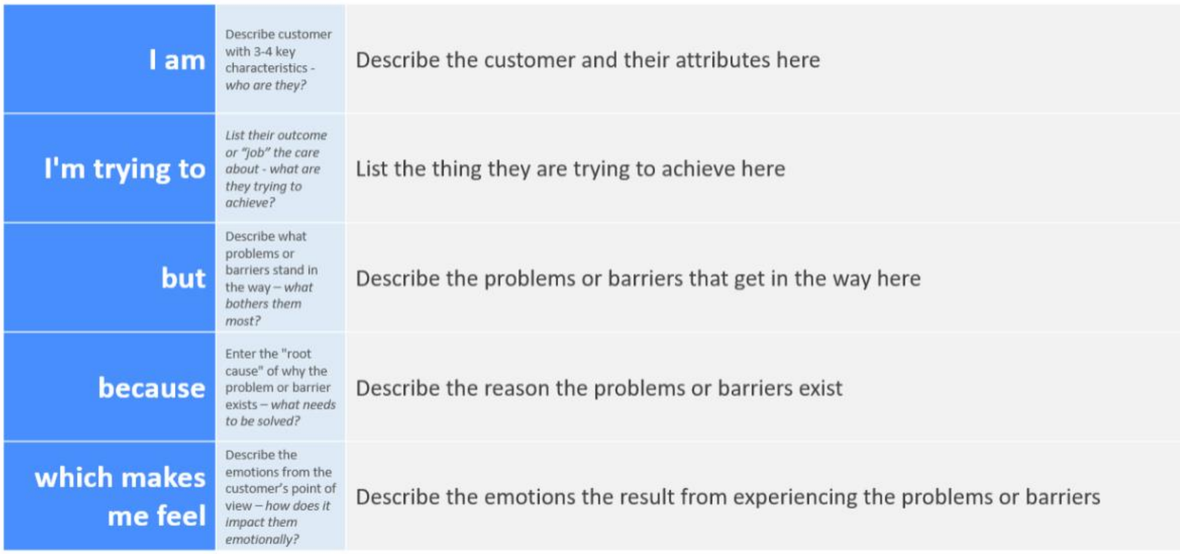
- https://www.iata.org/en/publications/store/world-air-transport-statistics/
- https://www.google.com/search?lei=cl9oY5byKqSvmgesiq-wDQ&q=data%20analytics%20in%20aviation%20industry&ved=2ahUKEwiW86_T9Zr7AhWkl-YKHSzFC9YQsKwBKAB6BAhDEAE
- https://www.google.com/search?lei=cl9oY5byKqSvmgesiq-wDQ&q=impact%20of%20covid-19%20on%20aviation%20industry%20research%20paper&ved=2ahUKEwiW86_T9Zr7AhWkl-YKHSzFC9YQsKwBKAJ6BAhDEAM
- https://dl.acm.org/doi/abs/10.1145/3469028
- https://www.nap.edu/read/21909/chapter/5

2.3. PROBLEM STATEMENT DEFINITION

Customer Problem Statement Template:

Create a problem statement to understand your customer's point of view. The Customer Problem Statement template helps you focus on what matters to create experiences people will love.

A well-articulated customer problem statement allows you and your team to find the ideal solution for the challenges your customers face. Throughout the process, you'll also be able to empathize with your customers, which helps you better understand how they perceive your product or service.



Reference: https://miro.com/templates/customer-problem-statement/

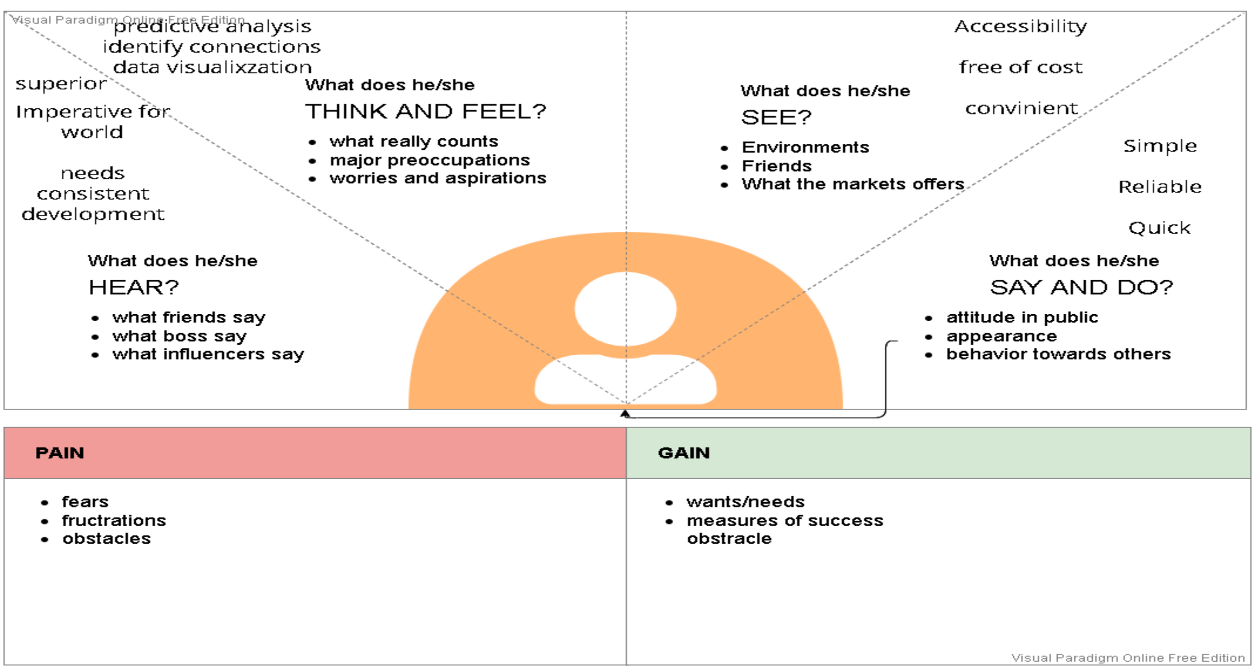
Example:

Problem Statement (PS)	I am (Customer)	I'm trying to	But	Because	Which makes me feel
PS-1	Facing flight delay	Evolve numerous techniques of improving the airlines transportation system	This has brought drastic change in airlines operations	Flights delay ocassinally cause inconvenience to the modern passengers	It hurts airports ,airlines,and affects a companys marketing strategies as companies rely on customer loyalty to

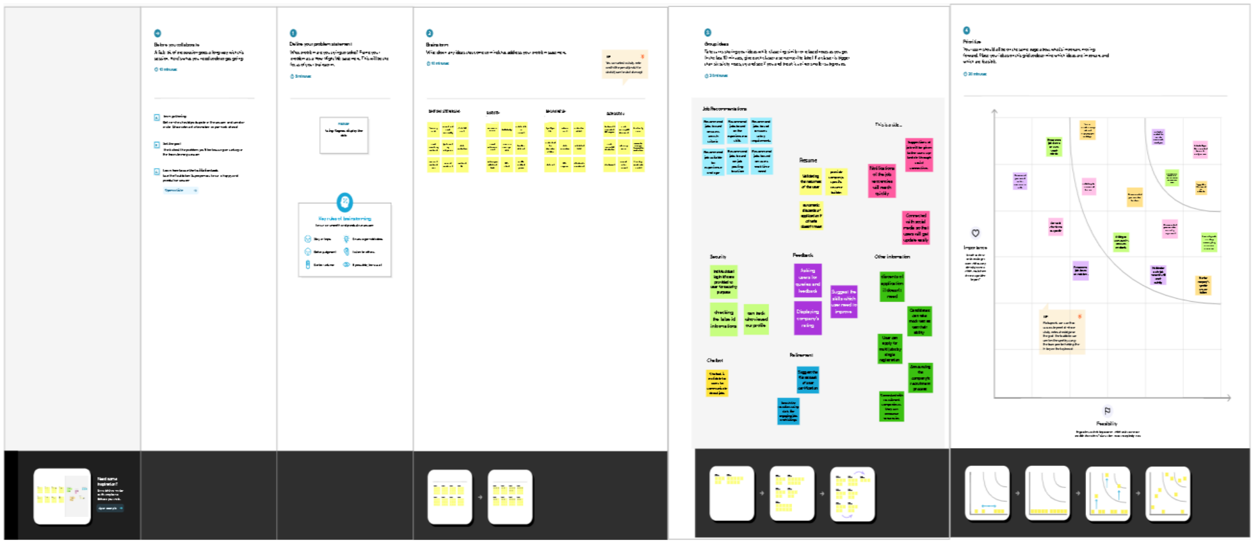
					support their frequent flying programs.
PS-2	Facing HAP emission from idling aircraft on ambient conditions	Improve our quantitative understanding of the largest aviation related HAPs emission source jet engines operating at low power	Potentially toxic emissions is hazardous air pollutant emissions and most important source of airport related HAPs compound at most commercia airports in idling jet engines	This will allow airport operators to utilize the latest scientific findings to construct HAPs emission estimates tailored to their specific airport	These estimates will be more defensible and better able to withstand litigation since they will be based on latest scientific findings regarding jet engine HAPs emissions and variabes that affect them.

3.IDEATION & PROPOSED SOLUTION

3.1. Empathy Map



3.2. Brainstroming



3.3 . Proposed Solution

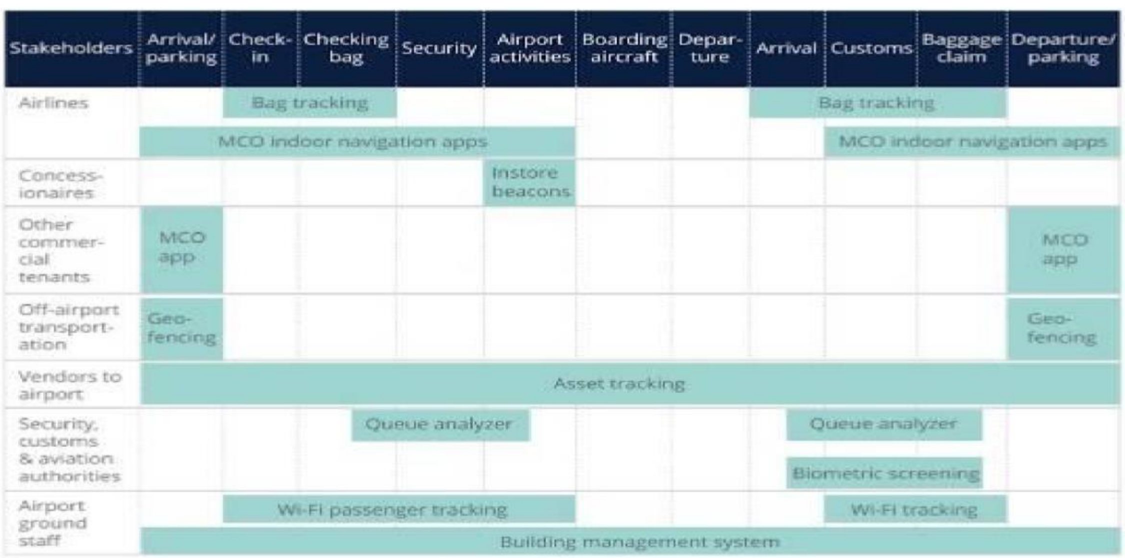
Proposed Solution Template:

Project team shall fill the following information in proposed solution template.

S.No.	Parameter	Description
1.	Problem Statement (Problem to be solved)	The airline industry has been keeping a tab on this information since long but it needs big data to help them analyse it and make it useful for the customer
2.	Idea / Solution description	The purpose of data analytics in aviation is to examine the vast amount of data generated daily and provide useful information to airlines, airports and other aviation stakeholders so that they can improve their operational planning and execution, as well as any related products and services
3.	Novelty / Uniqueness	Aforable. Easy to use. Services: Advance analysis, Easy to use and maintain, Actionnable report.
4.	Social Impact / Customer Satisfaction	The results of data analysis show that, in overall, full service airline customers are more satisfied than that of the low cost airline customers. Further, regression analysis on low cost airline data shows that the promptness and accuracy of service, employee attitudes, and price significantly influence customer satisfaction. While in full service airline physical evidence, the attitude of employees,

		and the price are significant predictors of customer satisfaction. This study underlines that the service quality especially the service employees' attitudes and price are factors that should be given more attention for developing customer satisfaction in both types of airlines, although their competitive strategy and target market are different.
5.	Business Model (Revenue Model)	There are two main business models in the airline industry: traditional Full-Service Carriers (FSCs) and Low-Cost Carriers (LCCs). The LCC business model was first pioneered by US-based Southwest Airlines. In a nutshell, low-cost airlines minimize operations costs to offer the cheapest tickets possible.
6.	Scalability of the Solution	This study illustrates how airlines successfully adopt big data technology. The paper also explores the opportunities and challenges of big data in the airline industry.

3.4. Problem Solution Fit



4. REQUIREMENT ANALYSIS

4.1. Functional Requirements:

Following are the functional requirements of the proposed solution.

FR No.	Functional Requirement (Epic)	Sub Requirement (Story / Sub-Task)
FR-1	User Registration	Can register through gmail/phone number
FR-2	User Confirmation	Receives Confirmation text via Email /OTP
FR-3	Visualization of data	Through IBM cognos analytics to know about delay of flights
FR-4	Generation of report	Users can know the timings and delay durations

4.2.Non-functional Requirements:

Following are the non-functional requirements of the proposed solution.

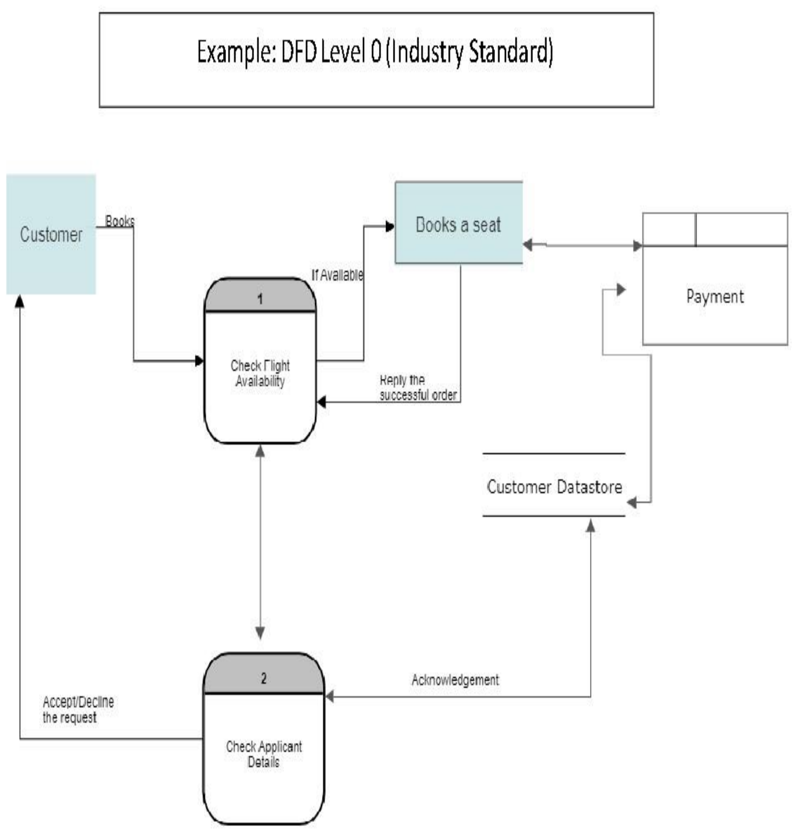
NFR-5	**Availability**	The application must be available to access at anytime , anydays
NFR-6	**Scalability**	Better scalability that large number of users could able to access at a time
FR No.	**Non-Functional Requirement**	**Description**
NFR-1	**Usability**	Users can access the application easily.Any functions can be performed with simple steps.
NFR-2	**Security**	Proper digital privacy system should be implemented to protect the details of user. Efficient login system should be made.
NFR-3	**Reliability**	When the system /server processing lowers,it should able to restain the saved datas of user

NFR-4	**Performance**	The system should have a efficient speed for browsing details

5. **PROJECT DESIGN**

5.1. Data Flow Diagrams

A Data Flow Diagram (DFD) is a traditional visual representation of the information flows within a system. A neat and clear DFD can depict the right amount of the system requirement graphically. It shows how data enters and leaves the system, what changes the information, and where data is stored.



User Type	Functional Requirement (Epic)	User Story Number	User Story / Task	Acceptance criteria	Priority	Release
Customer (Web user)	Registration	USN-1	As a user, I can register for the application by entering my email, password, and confirming my password.	I can access my account / dashboard	High	Sprint-1
		USN-2	As a user, I will receive confirmation email once I have registered for the application	I can receive confirmation email & click confirm	High	Sprint-1
		USN-3	As a user, I can register for the application through Gmail.		Medium	Sprint-1
	Login	USN-4	As a user, I can log into the application by entering email & password.	I can get to access my web portal	High	Sprint-1
	Dashboard	USN-5	As a user, I can get to know what my dashboard consists of.	I can my details of my registration.	Low	Sprint-2
Customer Care Executive	Organization	USN-6	The organization which owns this airplane analysis system will enable the option to customers to reach out the organization if • they have any problem with the organization's system of customer interaction or • airplane issues- delay, landing in a different location	The customer care workers will help out the customers in trouble.	High	Sprint-1
Administrator	Administration	USN-7	The organization takes in-charge of the administrative policies of different departments like: • registration • flight booking • delay visualization • generation of delay report	As an administrator, confirmation of user while registration is done.	High	Sprint-1

5.2. Solution & Technical Architecture

Technical Architecture:

The Deliverable shall include the architectural diagram as below and the information as per the table1 & table 2

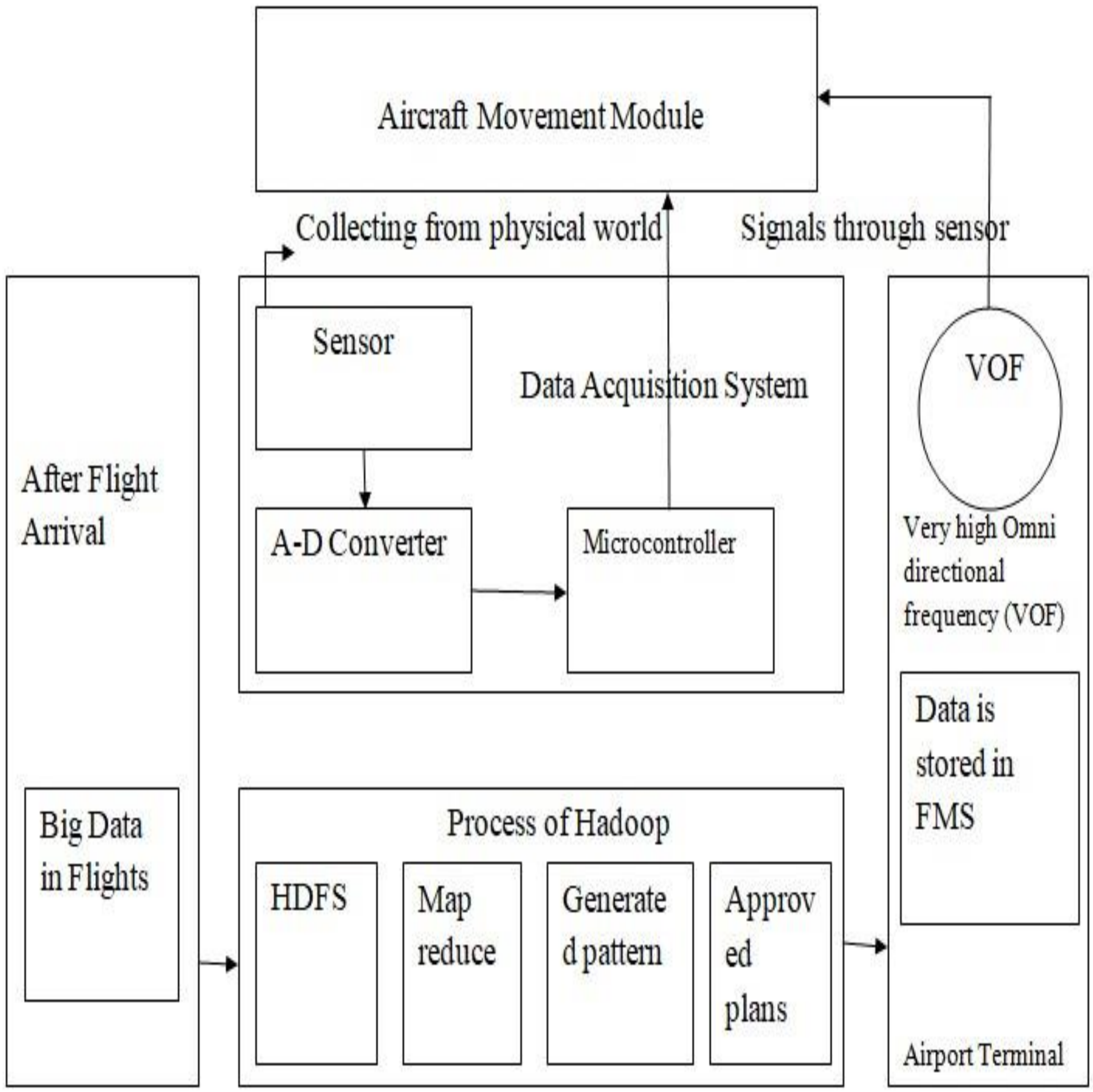

Aircraft Movement Module
Collecting from physical world
Signals through sensor
After Flight Arrival
Sensor
Data Acquisition System
A-D Converter
Microcontroller
VOF
Very high Omni directional frequency (VOF)
Data is stored in FMS
Big Data in Flights
Process of Hadoop
HDFS
Map reduce
Generated pattern
Approved plans
Airport Terminal

Table-1: Components & Technologies:

S.No	Characteristics	Description	Technology
1.	Open-Source Frameworks	List the open-source frameworks used	Technology of open-source framework
2.	Security Implementations	List all the security/access controls implemented, use of firewalls.	Example: SHA-256, Encryption, IAM Controls, OWASP

S.No	Components	Description	Technology
1.	User Interface	How user interacts with application. Example: Mobile App	HTML, CSS, Java Script, Excel
2.	Application Logic-1	Logic for a process in the application	IBM Watson STT service, Python
3.	Application Logic-2	Logic for a process in the application	IBM Watson Assistant
4.	Database	Data Type, Configurations	MySQL, NSQL
5.	Cloud Database	Database service on cloud	IBM DB2, IBM Cloudant
6.	File Storage	File Storage requirements	IBM Blocks Storage or other storage service or Local File system
7.	External API-1	Purpose of External API used in the application	IBM Weather API
8.	External API-1	Purpose of External API used in the application	Aadhar API

9.	Infrastructure (Server/Cloud)	Application Deployment on Local System/Cloud Local Server Configuration: Cloud Server Configuration	Local, Cloud Foundry

Table-2: Application Characteristics:

3.	Scalable Architecture	Justify the scalability of architecture	Cognos Used
4.	Availability	Justify the availability of application (e.g: use of load balancers, distributed servers)	AWS Used
5.	Performance	Design consideration for the performance of the application (number of requests per second, use of Cache, use of CDN's)	Dashboard,Reports,Stories

5.3. User Stories

Table-1: Components & Technologies:

Component	Description	Technology
User Interface	User can Interact with web Applications	HTML, CSS, JavaScript.
Data Preparation	Pre-processing of data should be done	Python
Feature Selection	Feature selection of the Dataset using the Correlation Feature Selection method.	Python
Data Analytics	Prediction of Flight delay using Decision Tree.	Python
Data Visualization	Data Type, Configurations etc.	Python
Data Storage	Database Service on Cloud	IBM DB2, IBM Cloudant etc.
User Interface	Dashboard showing the details of the flight delay	HTML, CSS, JavaScript.

Table 2: Application Characteristics:

Characteristics	Description	Technology
Security Implementations	The main security concern is for users' accounts hence proper login mechanisms should be used to avoid hacking.	e.g. SHA-256, Encryptions, IAM Controls, OWASP etc.
Availability	The system will be available 24 hours a day 7 days a week. Users can access it at any time.	

6.PROJECT PLANNING & SCHEDULING

6.1. Sprint Planning & Estimation

Product Backlog, Sprint Schedule, and Estimation (4 Marks)

Use the below template to create product backlog and sprint schedule

Sprint	Functional Requirement (Epic)	User Story Number	User Story / Task	Story Points	Priority	Team Members
Sprint1	Registration	USN-1	As a user, I can register for the application by entering my email, password, and confirming my password.	4	High	DHIFANI STENIKSHA D
Sprint1	Login	USN-2	As a user, I adapt to logging into the system with credentia	2	Low	BARATH
Sprint1	Designation of Region	USN-3	As a user, I can collect the dataset and select the region of interest to be monitored and analysed	2	Low	GOKUL RAJ
Sprint2	Exploration Of The Data	USN-4	As a developer,I will explore the given dataset through cognos.	3	Medium	BHUVANESH
Sprint2	Visualization Of The Dataset	USN-5	As a developer,I will visualize the given dataset into a dashboard using cognos	2	Low	DHIFANI STENIKSHA D
Sprint3	Customization Of The Dashboard	USN-6	As a user,I can customize the visualized dashboard.	2	Low	BARATH
Sprint3	Ease of Access	USN-7	As a user,I can easily access and	2	Low	GOKULRAJ
Sprint	**Functional Requirement (Epic)**	**User Story Number**	**User Story / Task**	**Story Points**	**Priority**	**Team Members**
			manipulate the dashboard.			

Sprint4	Report Generation	USN-8	As a user,I can view the detailed report of my visualization.	4	High	DHIFANI STENIKSHA D
Sprint4	Establishment of the Dashboard	USN-9	As a developer,I established the dashboard intoa website and submit the website.	3	Medium	BHUVANESH

Project Tracker, Velocity & Burndown Chart: (4 Marks)

Sprint	Total Story Points	Duration	Sprint Start Date	Sprint End Date (Planned)	Story Points Completed (as on Planned End Date)	Sprint Release Date (Actual)
Sprint1	20	6 Days	24 Oct 2022	29 Oct 2022	12	29 Oct 2022
Sprint2	20	6 Days	31 Oct 2022	05 Nov 2022	12	05 Nov 2022
Sprint3	20	6 Days	07 Nov 2022	12 Nov 2022	12	12 Nov 2022
Sprint4	20	6 Days	14 Nov 2022	19 Nov 2022	12	19 Nov 2022

Velocity:

Imagine we have a 10-day sprint duration, and the velocity of the team is 20 (points per sprint). Let's calculate the team's average velocity (AV) per iteration unit (story points per day)

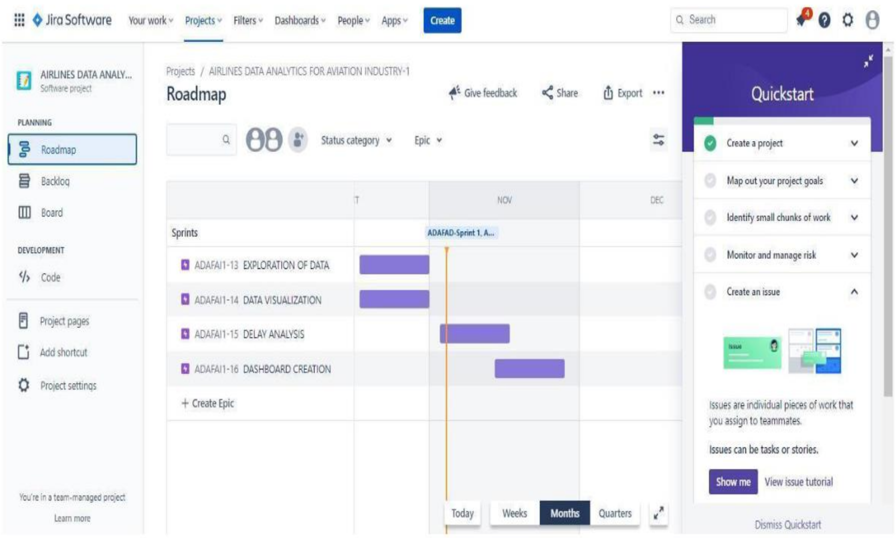
$$AV = \frac{sprint\ duration}{velocity} = \frac{20}{10} = 2$$

Average velocity=Sprint duration / velocity=12/6=2

6.2.Sprint Delivery Schedule

6.3.Report from JIRA

Project RoadMap



7.CODING AND SOLUTIONING

7.1.Feature 1

Top 5 Airports with Maximum Cancellations (decreasing order)

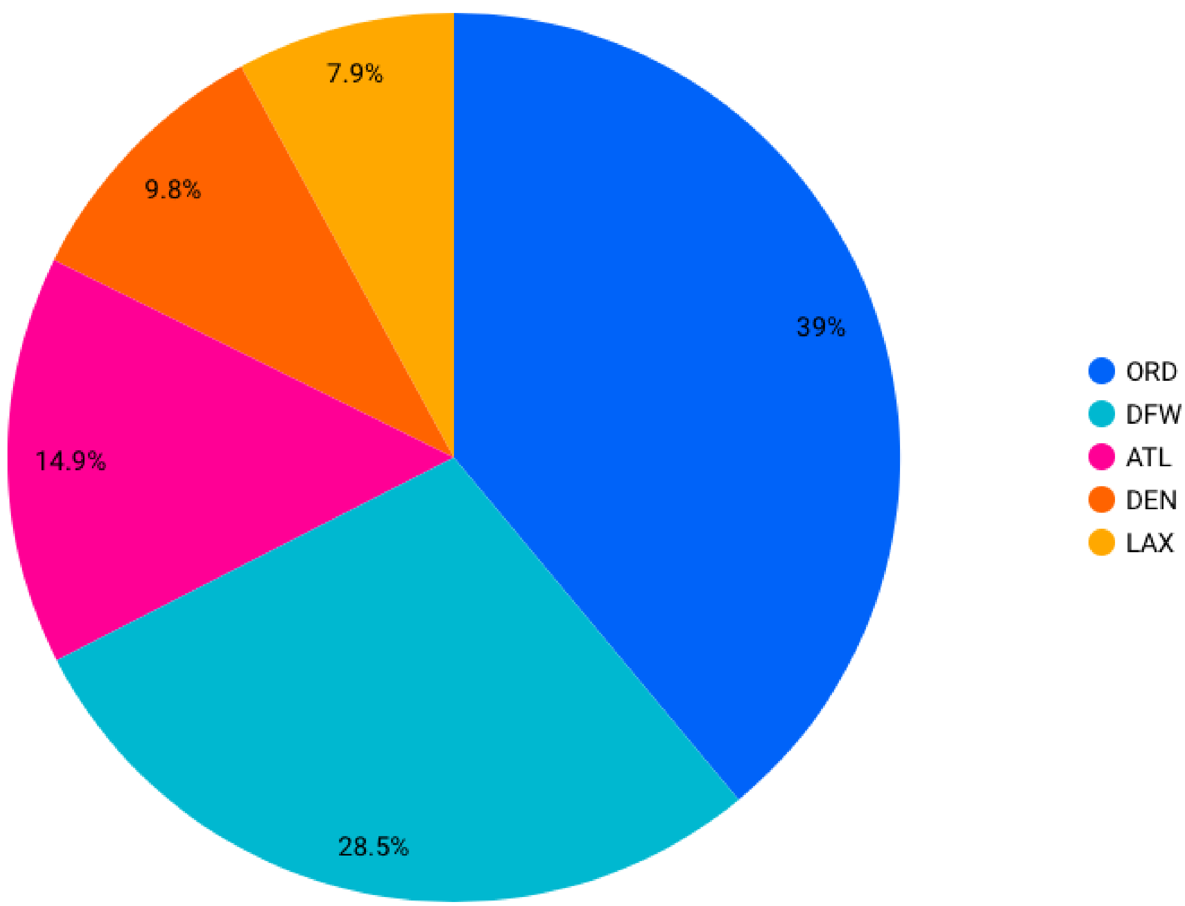
```
WITH
  top_5_airports AS (
  SELECT
    ORIGIN,
    COUNT(ORIGIN) AS count
  FROM
    `airline-delay-canc.airlines_data.delay_canc_data`
  GROUP BY
    1
  ORDER BY
    2 DESC
  LIMIT
    5 ),
  top_5_airlines AS (
  SELECT
```

```
    OP_CARRIER,
    COUNT(OP_CARRIER) AS count
  FROM
    `airline-delay-canc.airlines_data.delay_canc_data` main,
    top_5_airports top5
  WHERE
    top5.ORIGIN = main.ORIGIN
  GROUP BY
    1
  ORDER BY
    2 DESC
  LIMIT
    5),
  all_flights AS (
  SELECT
    main.ORIGIN AS Airport,
    main.OP_CARRIER AS Carrier,
    COUNT(*) AS all_cnt
  FROM
    `airline-delay-canc.airlines_data.delay_canc_data` main,
    top_5_airports top5_ap,
    top_5_airlines top_al
  WHERE
    top5_ap.ORIGIN = main.ORIGIN
    AND top_al.OP_CARRIER = main.OP_CARRIER
  GROUP BY
    1,
    2 ),
  cancelled_flights AS (
  SELECT
    main.ORIGIN AS Airport,
    main.OP_CARRIER AS Carrier,
    COUNT(*) AS cancelled_cnt
  FROM
    `airline-delay-canc.airlines_data.delay_canc_data` main,
    top_5_airports top5_ap,
    top_5_airlines top_al
  WHERE
    top5_ap.ORIGIN = main.ORIGIN
    AND top_al.OP_CARRIER = main.OP_CARRIER
    AND cancelled = 1
  GROUP BY
    1,
    2 )
SELECT
  af.Airport,
  af.Carrier,
  af.all_cnt - cf.cancelled_cnt AS all_cnt,
  cf.cancelled_cnt
```

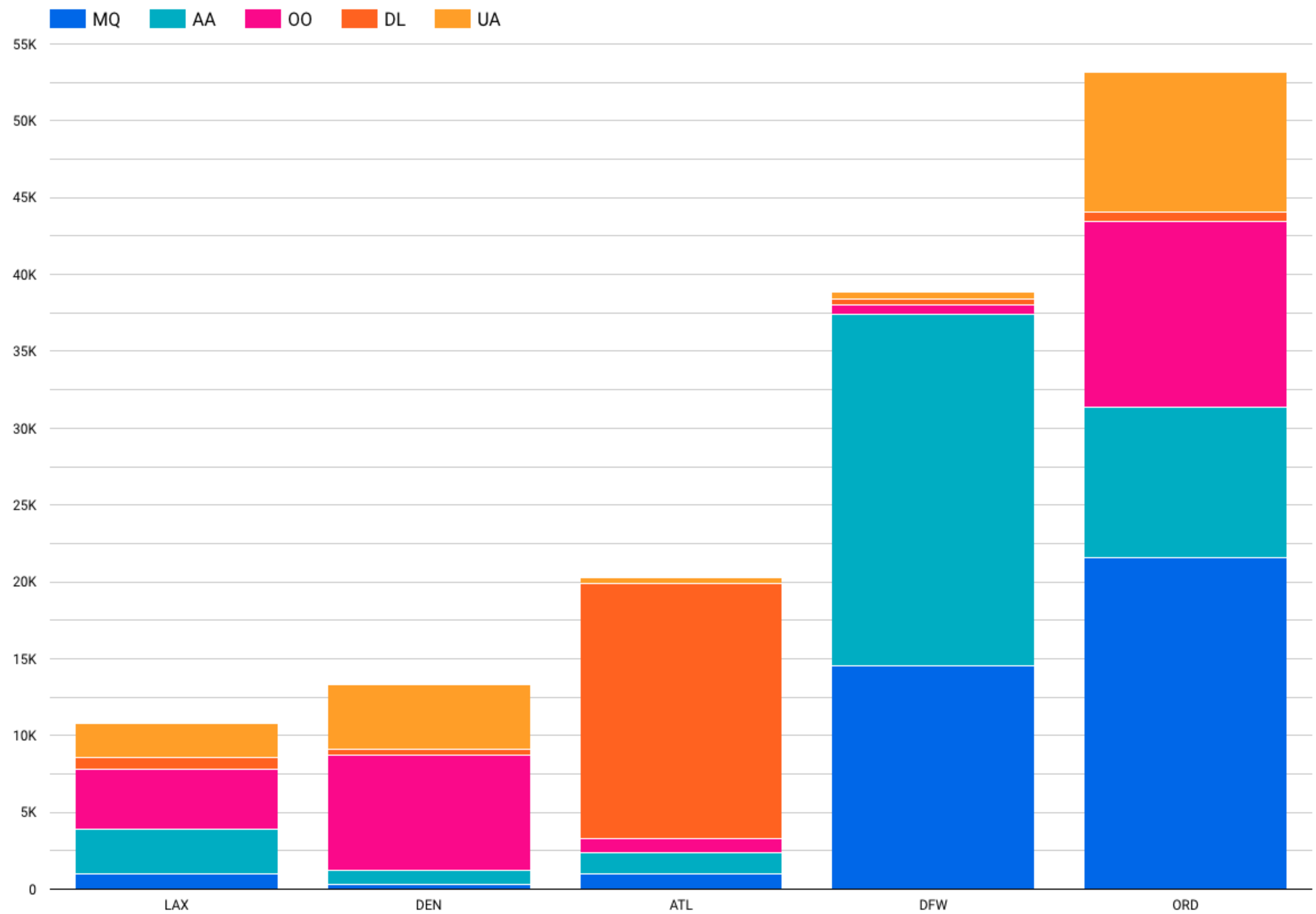
```
FROM
  all_flights af,
  cancelled_flights cf
WHERE
  af.Airport = cf.Airport
  AND af.Carrier = cf.Carrier
```

Results

| S No.| Airport Code | Airport Name | Cancellation (in %) | | - | - | - | - | - | 1. | **ORD** | (O'Hare International Airport) | 39| | 2. | **DFW** | (Dallas/Fort Worth International Airport) | 28.5| | 3. | **ATL** | (Hartsfield-Jackson Atlanta International Airport) | 14.9| | 4. | **DEN** | (Denver International Airport) | 9.8| | 5. | **LAX** | (Los Angeles International Airport) | 7.9|



Airline-wise Cancellation Bifurcation



Top Cancellation Reasons for Top 5 Busiest Airports

Query - JS UDF Function

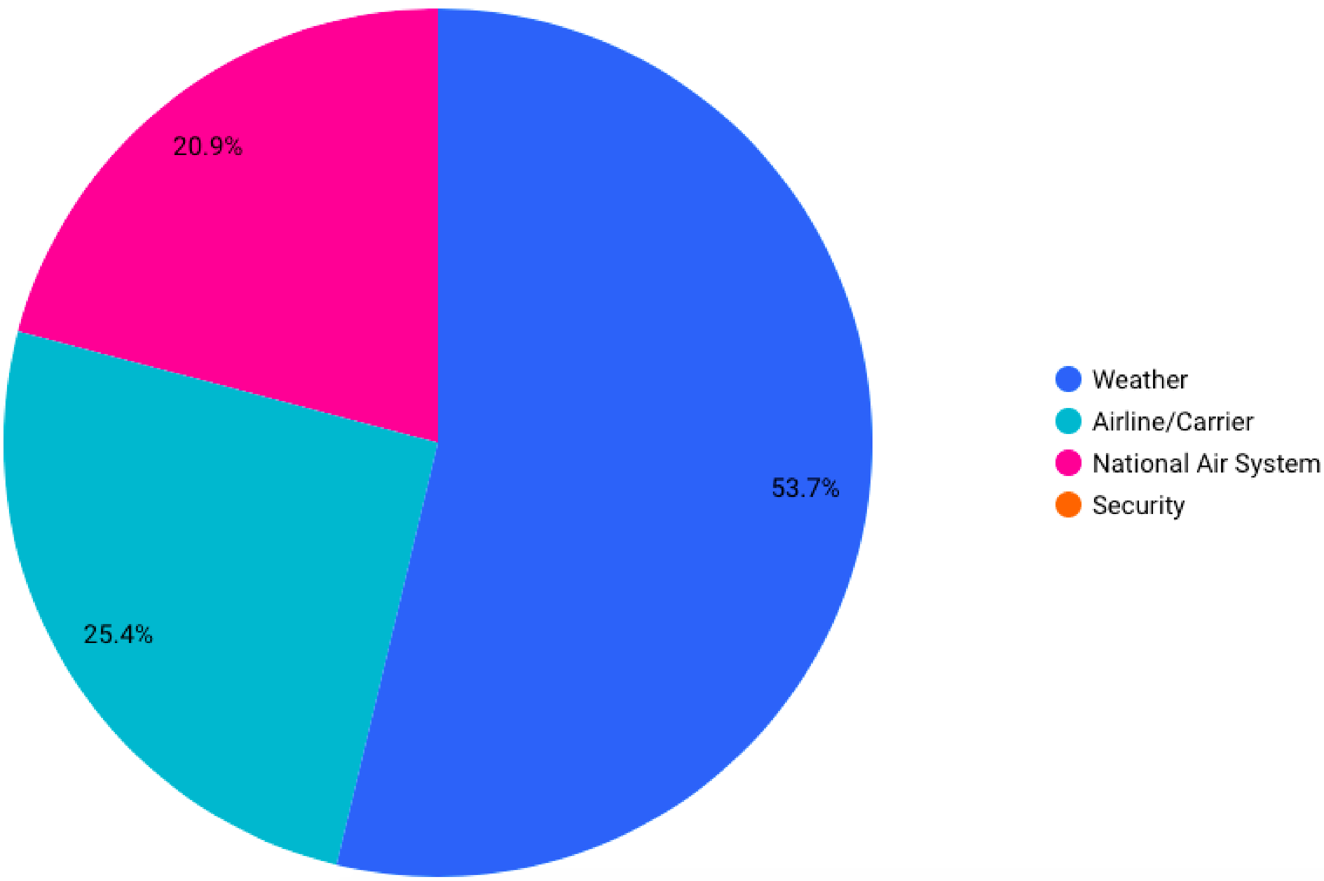
```
CREATE TEMP FUNCTION
  cancellation_reason(code string)
  RETURNS string
  LANGUAGE js AS """
    switch(code) {
        case "A":
          return "Airline/Carrier";
        break;
```

```
        case "B":
          return "Weather";
        break;
        case "C":
          return "National Air System";
        break;
        case "D":
          return "Security";
        break;
        default:
          return "Others";
        break;
    }
""";
WITH
  top_5_airports AS (
  SELECT
    ORIGIN,
    COUNT(ORIGIN) AS count
  FROM
    `airline-delay-canc.airlines_data.delay_canc_data`
  GROUP BY
    1
  HAVING
    count > 100000
  ORDER BY
    2 DESC
  LIMIT
    5 )
SELECT
  top5.ORIGIN,
  cancellation_reason(main.CANCELLATION_CODE) AS reason,
  COUNT(main.CANCELLATION_CODE) AS count
FROM
  `airline-delay-canc.airlines_data.delay_canc_data` main,
  top_5_airports top5
WHERE
  CANCELLED = 1
  AND EXTRACT(year
  FROM
    FL_DATE) = 2018
  AND top5.ORIGIN = main.ORIGIN
GROUP BY
  1,
  2
```

```
ORDER BY
  1,
  2
```

Result

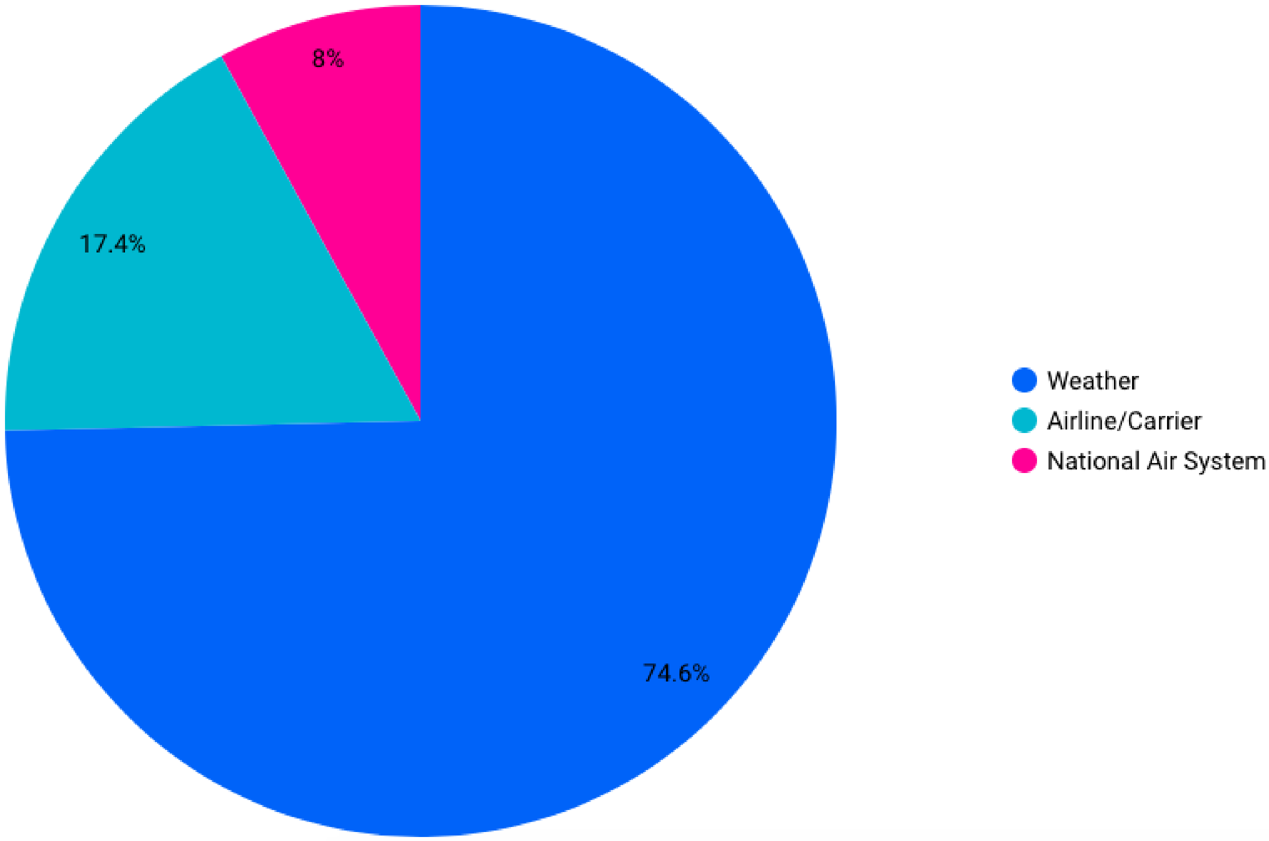
| S No.| Reason | Cancellation (in %) | | - | - | - | - | 1. | **Weather** | 53.7| | 2. | **Airline/Carrier Delays** | 25.4| | 3. | **National Air System** | 20.9| | 4. | **Airport Secutiy** | 0.01 (~ 0)|



Top Cancellation Reasons at the Most Busiest Airport in practice (Atlanta)

- Atlanta is one of the largest inter-connect point (airport) for domestic and international flights in USA.

|S No.| Reason | Cancellation (in %) | | - | - | - | - | 1. | **Weather** | 74.6| | 2. | **Airline/Carrier Delays** | 17.4| | 3. | **National Air System** | 8|



7.2.Feature 2

Overall Delays at Top 5 Airports for top 5 airlines

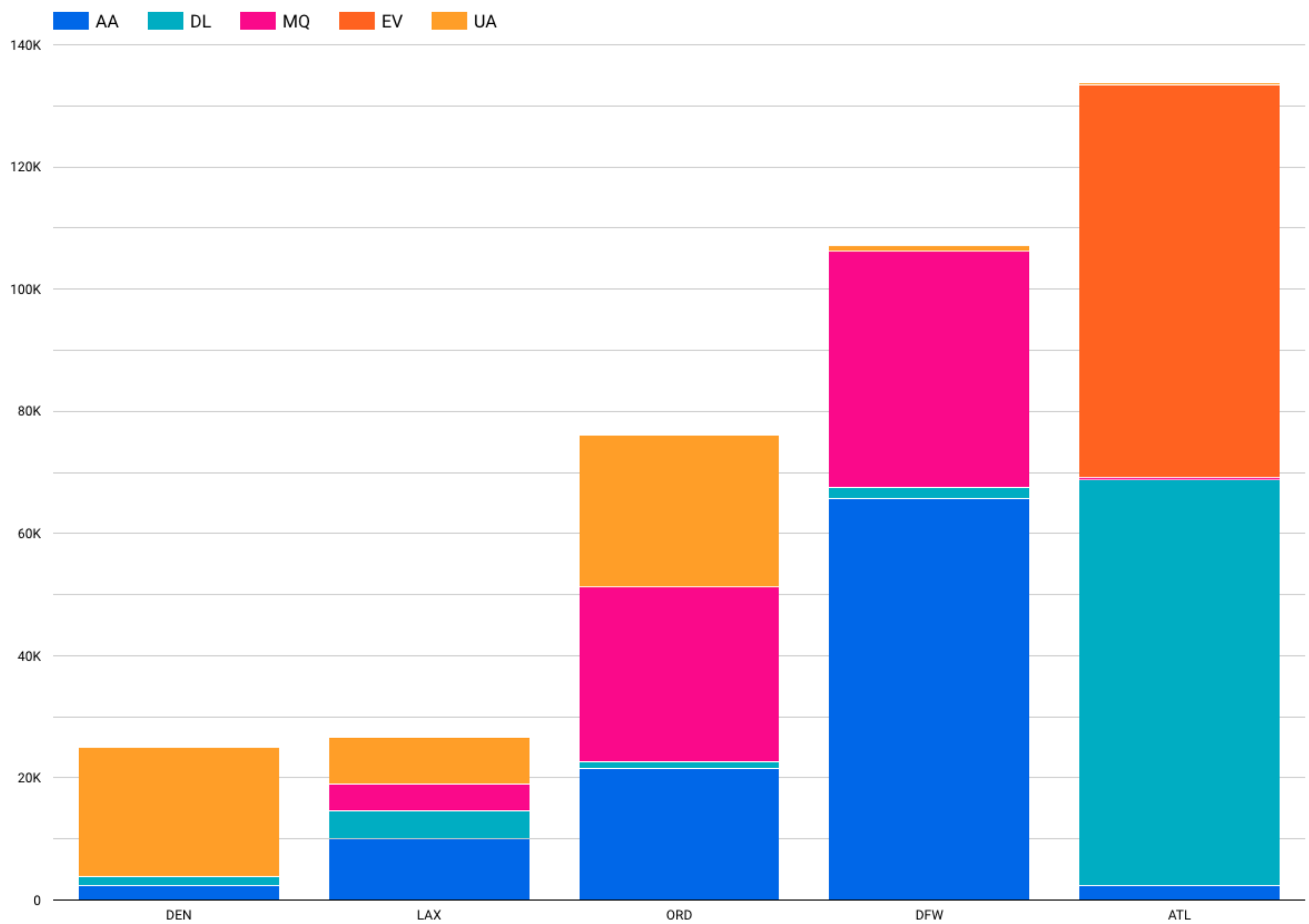
Query

```
WITH
  top_5_airports AS (
  SELECT
    ORIGIN,
    COUNT(ORIGIN) AS count
  FROM
    `airline-delay-canc.airlines_data.delay_canc_data`
  GROUP BY
    1
```

```
  ORDER BY
    2 DESC
  LIMIT
    5 ),
  top_5_airlines AS (
  SELECT
    OP_CARRIER,
    COUNT(OP_CARRIER) AS count
  FROM
    `airline-delay-canc.airlines_data.delay_canc_data` main,
    top_5_airports top5
  WHERE
    top5.ORIGIN = main.ORIGIN
  GROUP BY
    1
  ORDER BY
    2 DESC
  LIMIT
    5),
  all_flights AS (
  SELECT
    main.ORIGIN AS Airport,
    main.OP_CARRIER AS Carrier,
    COUNT(*) AS all_cnt
  FROM
    `airline-delay-canc.airlines_data.delay_canc_data` main,
    top_5_airports top5_ap,
    top_5_airlines top_al
  WHERE
    top5_ap.ORIGIN = main.ORIGIN
    AND top_al.OP_CARRIER = main.OP_CARRIER
  GROUP BY
    1,
    2 ),
  delayed_flights AS (
  SELECT
    main.ORIGIN AS Airport,
    main.OP_CARRIER AS Carrier,
    COUNT(*) AS delayed_cnt
  FROM
    `airline-delay-canc.airlines_data.delay_canc_data` main,
    top_5_airports top5_ap,
    top_5_airlines top_al
  WHERE
    top5_ap.ORIGIN = main.ORIGIN
    AND top_al.OP_CARRIER = main.OP_CARRIER
    AND (CARRIER_DELAY IS NOT NULL
      AND CARRIER_DELAY > 0
      OR ARR_DELAY IS NOT NULL
      AND ARR_DELAY > 0)
  GROUP BY
    1,
    2 )
SELECT
  af.Airport,
```

```
  af.Carrier,
  af.all_cnt all_with_del,
  df.delayed_cnt,
  af.all_cnt - df.delayed_cnt AS all_without_del
FROM
  all_flights af,
  delayed_flights df
WHERE
  af.Airport = df.Airport
  AND af.Carrier = df.Carrier
```

Overall Delays at Top 5 Airports with top 5 airlines



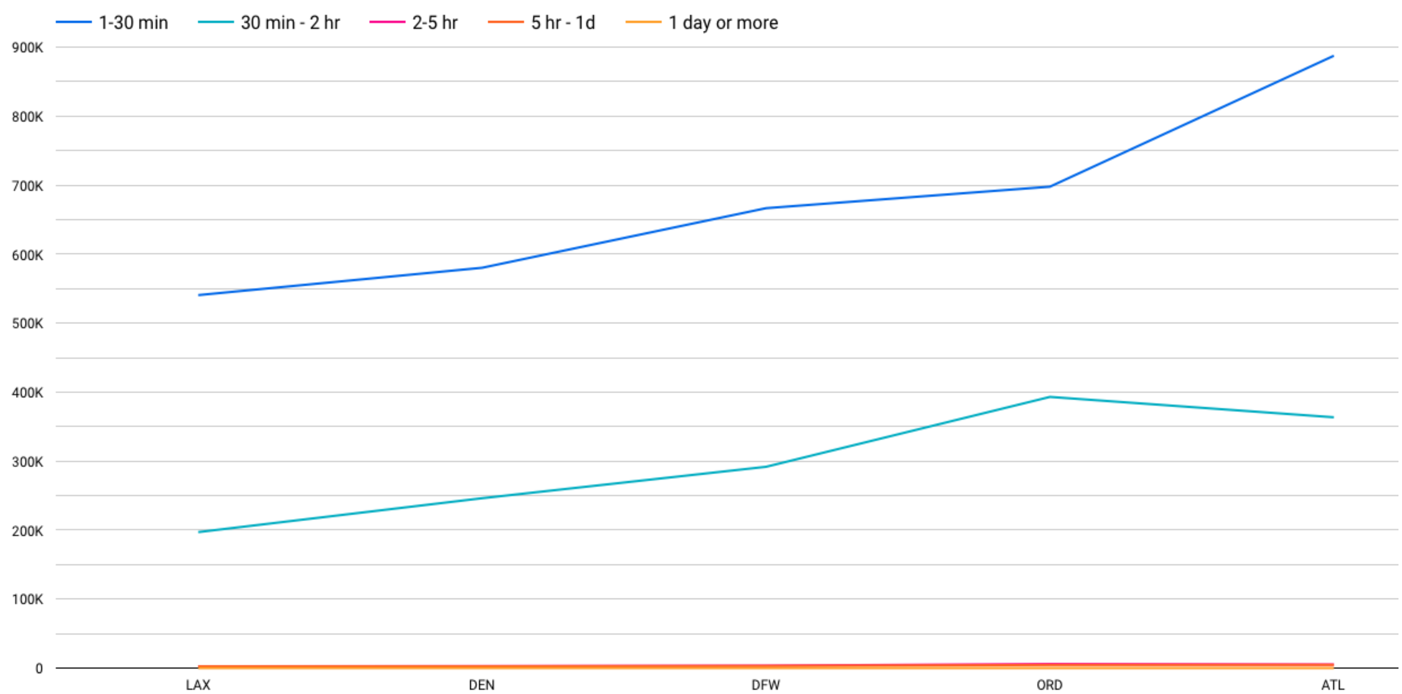
Overall Delay Time Frequency with Top 5 Airports

Query

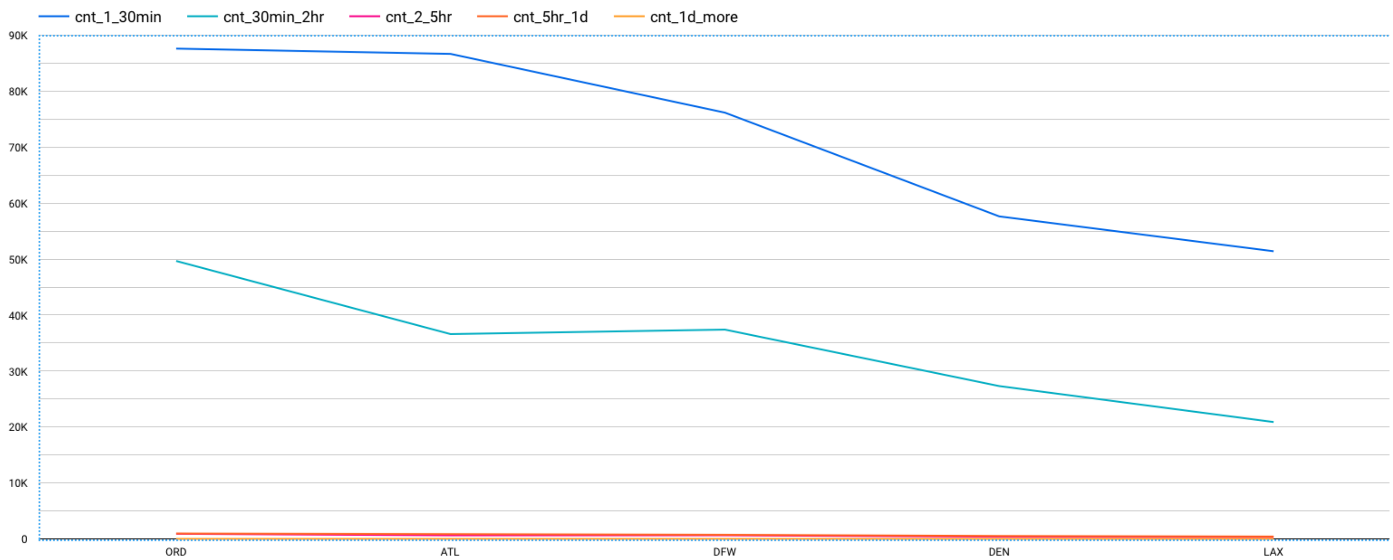
```
CREATE TEMP FUNCTION delay_bifurcation(slot_cnt ARRAY<STRUCT<slot int64,count
int64>>)
   RETURNS STRUCT<cnt_1_30 float64, cnt_30_2 float64, cnt_2_5 float64, cnt_5_24
float64, cnt_24 float64>
  LANGUAGE js AS """
  let response = {"cnt_1_30": 0.0, "cnt_30_2": 0.0, "cnt_2_5": 0.0, "cnt_5_24": 0.0,
"cnt_24": 0.0}
  for(let i = 0 ; i < slot_cnt.length; i++){
      let slotCntObj = slot_cnt[i];
      let result =   slotCntObj.count;
      switch(parseInt(slotCntObj.slot)){
        case 1:
          response["cnt_1_30"] =  result;
          break;
        case 2:
          response["cnt_30_2"] = result;
          break;
        case 3:
          response["cnt_2_5"] = result;
          break;
        case 4:
          response["cnt_5_24"] = result;
          break;
        case 5:
          response["cnt_24"] = result;
          break;
        default:
          response["cnt_1_30"] = 0.0;
          response["cnt_30_2"] = 0.0;
          response["cnt_2_5"] = 0.0;
          response["cnt_5_24"] = 0.0;
          response["cnt_24"] = 0.0;
          break;
      }
    }
    return response
""";

WITH top_5_airports as (
      SELECT ORIGIN, count(ORIGIN) as count
      FROM `airline-delay-canc.airlines_data.delay_canc_data`
      Group by 1
      having count > 100000
      order by 2 desc
      limit 5
      ),
    delay_bifurcation as (
      select ORIGIN,
          (case when ARR_DELAY > 1440 then 5
             when ARR_DELAY > 300 then 4
             when ARR_DELAY > 240 then 3
             when ARR_DELAY > 30 then 2
        else 1 end) as slot

  from `airline-delay-canc.airlines_data.delay_canc_data`
```

```
  where ARR_DELAY is not null and ARR_DELAY > 0
--   and EXTRACT(year FROM FL_DATE) = 2018
  ),

  airport_timeslots as(
  select db.ORIGIN, db.slot, count(db.slot) as count
  from delay_bifurcation db,top_5_airports top5
  where top5.ORIGIN = db.ORIGIN
  group by 1,2),

  airport_struct as(
      select origin, struct(slot,count) as slot_cnt from  airport_timeslots
  ),
  udf_result as (select origin, delay_bifurcation(ARRAY_AGG(slot_cnt)) as slot_struct
  from airport_struct
  group by 1
  )
  select origin, slot_struct.cnt_1_30 as cnt_1_30min,
      slot_struct.cnt_30_2 as cnt_30min_2hr,
      slot_struct.cnt_2_5 as cnt_2_5hr,
      slot_struct.cnt_5_24 as cnt_5hr_1d,
      slot_struct.cnt_24 as cnt_1d_more
  from udf_result
```

Overall Delay Time Frequency with Top 5 Airports (UDF Function)



Overall Delay Frequency (Year with max delays and cancellations)



7.3.Database Schema

Flight count from Top 5 Airlines at Top 5 Airports

```
WITH top_5_airports AS (
	SELECT ORIGIN, COUNT(ORIGIN) AS count
	FROM
		airline-delay-canc.airlines_data.delay_canc_data
	GROUP BY
		1
	HAVING
		count > 100000
	ORDER BY
		2 DESC
	LIMIT 5
),
top_5_airlines AS (
	SELECT
		OP_CARRIER,
		COUNT(OP_CARRIER) AS count
	FROM
		airline-delay-canc.airlines_data.delay_canc_data main,
		top_5_airports top5
	WHERE
		top5.ORIGIN = main.ORIGIN
	GROUP BY
```

```
        1
    ORDER BY
        2 DESC
    LIMIT 5
),
airportwise_carrier_cnt AS (
    SELECT
        main.ORIGIN AS Airport,
        main.OP_CARRIER AS Carrier,
        COUNT(*) AS count
    FROM
        airline-delay-canc.airlines_data.delay_canc_data main,
        top_5_airports top5_ap,
        top_5_airlines top_al
    WHERE
        top5_ap.ORIGIN = main.ORIGIN
        AND top_al.OP_CARRIER = main.OP_CARRIER
    GROUP BY
        1,
        2
),
resut_cte AS (
    SELECT
        Airport,
        Carrier,
        count,
        RANK() OVER (PARTITION BY Airport ORDER BY count) AS rank
    FROM
        airportwise_carrier_cnt
)
SELECT
    Airport,
    Carrier,
    count
FROM
    resut_cte
WHERE
    rank < 6
```

Results

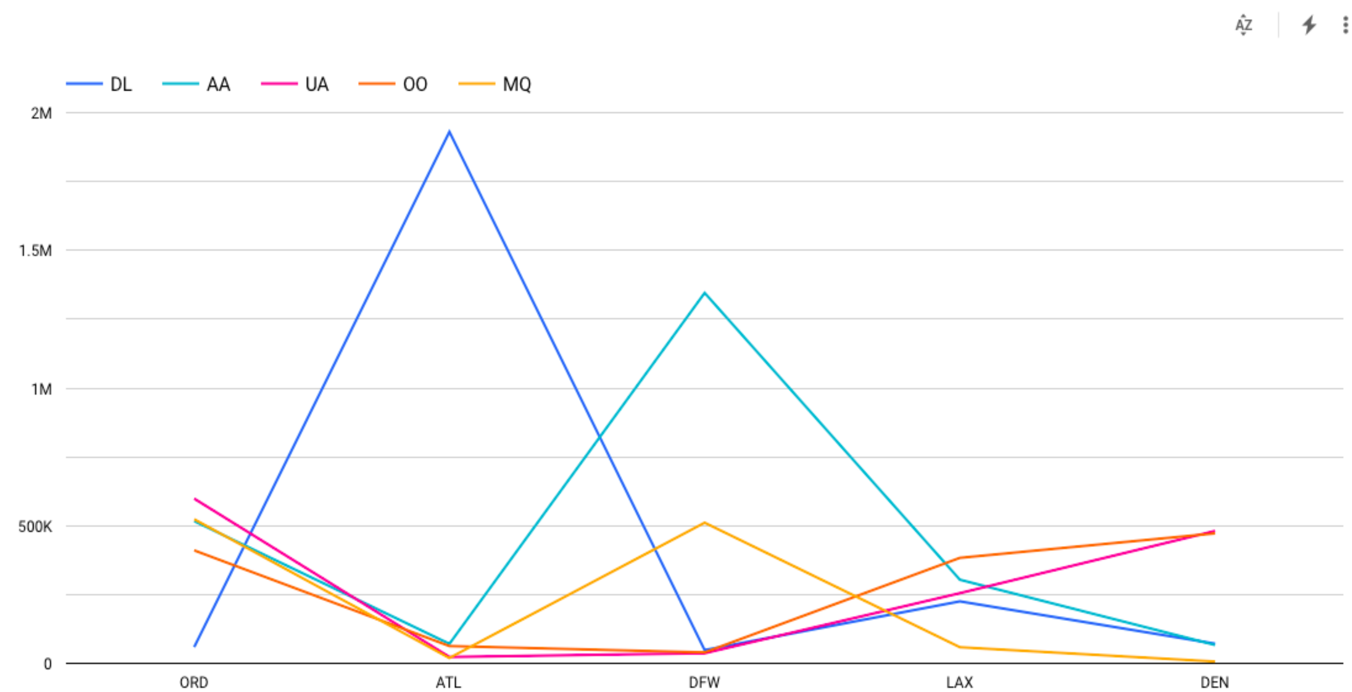
Top 5 Airports with maximum flight count:

1. **ORD** (O'Hare International Airport)
2. **ATL** (Hartsfield-Jackson Atlanta International Airport)

3. **DFW** (Dallas/Fort Worth International Airport)
4. **LAX** (Los Angeles International Airport)
5. **DEN** (Denver International Airport)

Top 5 Airlines with maximum flight count:

1. **DL** (Delta Air Lines)
2. **AA** (American Airlines)
3. **UA** (United Airlines)
4. **OO** (SkyWest Airlines)
5. **MQ** (American Eagle Airlines)



,

- From the above, it is realized that on **Delta Airlines** has the highest flight frequence on the **Atlanta** airport.

8.TESTING

8.1. Test Cases

- **Verify user is able to see home page without any content being hidden**
- **Verify user is able to enter data**
- **Verify user is able to navigate to the result**
- **Verify smooth usability**

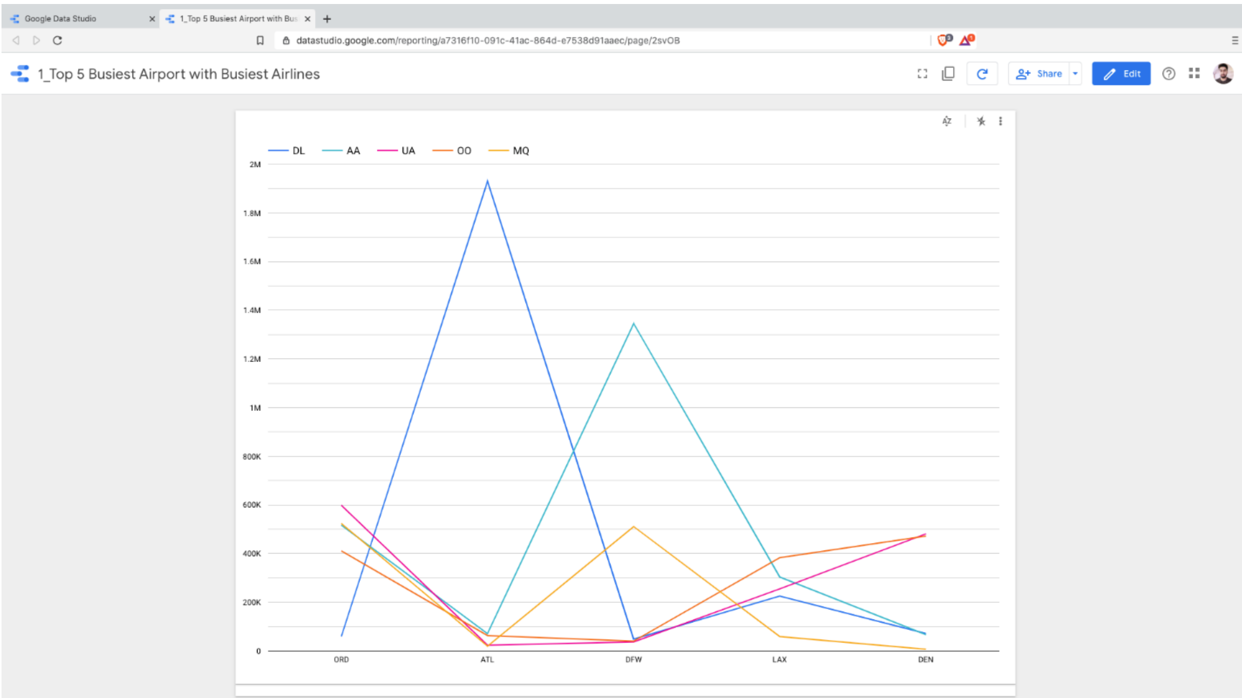
Resolution	Severity 1	Severity 2	Severity 3	Severity 4	Subtotal
By Design	10	4	2	3	20
Duplicate	1	0	3	0	4
External	2	3	0	1	6
Fixed	11	2	4	20	37
Not Reproduced	0	0	1	0	1
Skipped	0	0	1	1	2
Won't Fix	0	5	2	1	8
Totals	24	14	13	26	77

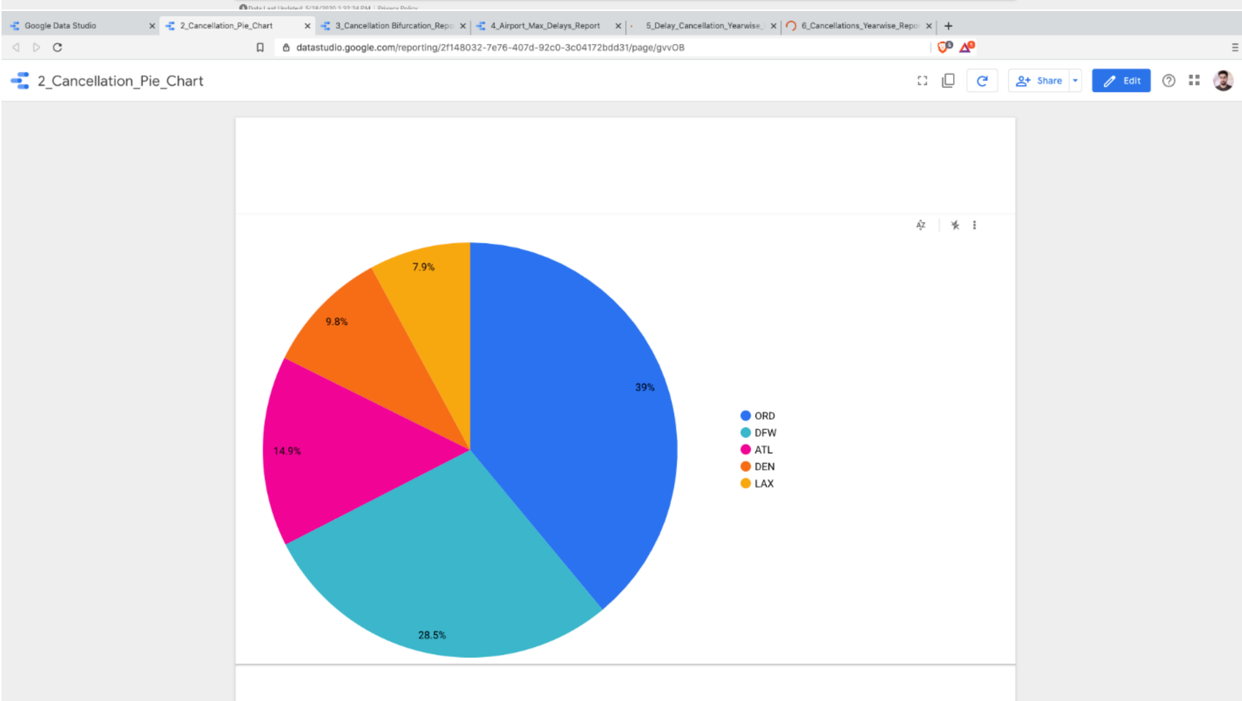
8.2.User Acceptance Testing

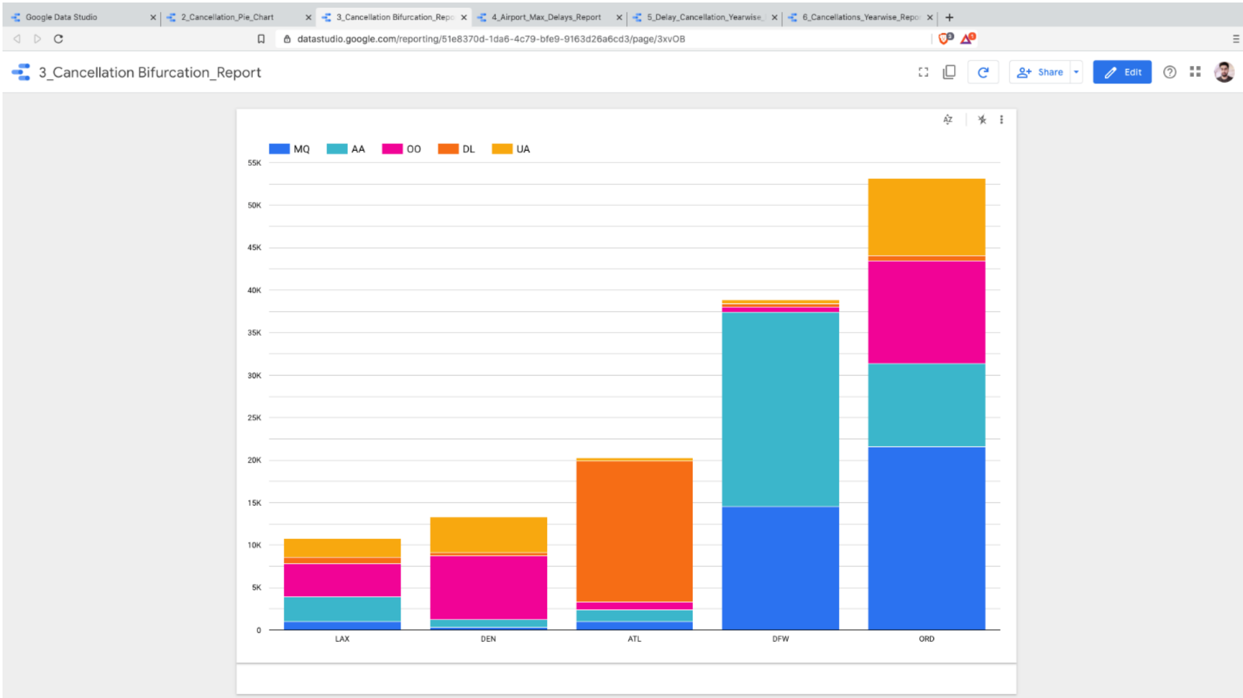
Section	Total Cases	Not Tested	Fail	Pass
Print Engine	7	0	0	7
Client Application	51	0	0	51
Security	2	0	0	2
Outsource Shipping	3	0	0	3
Exception Reporting	9	0	0	9
Final Report Output	4	0	0	4
Version Control	2	0	0	2

9.RESULTS

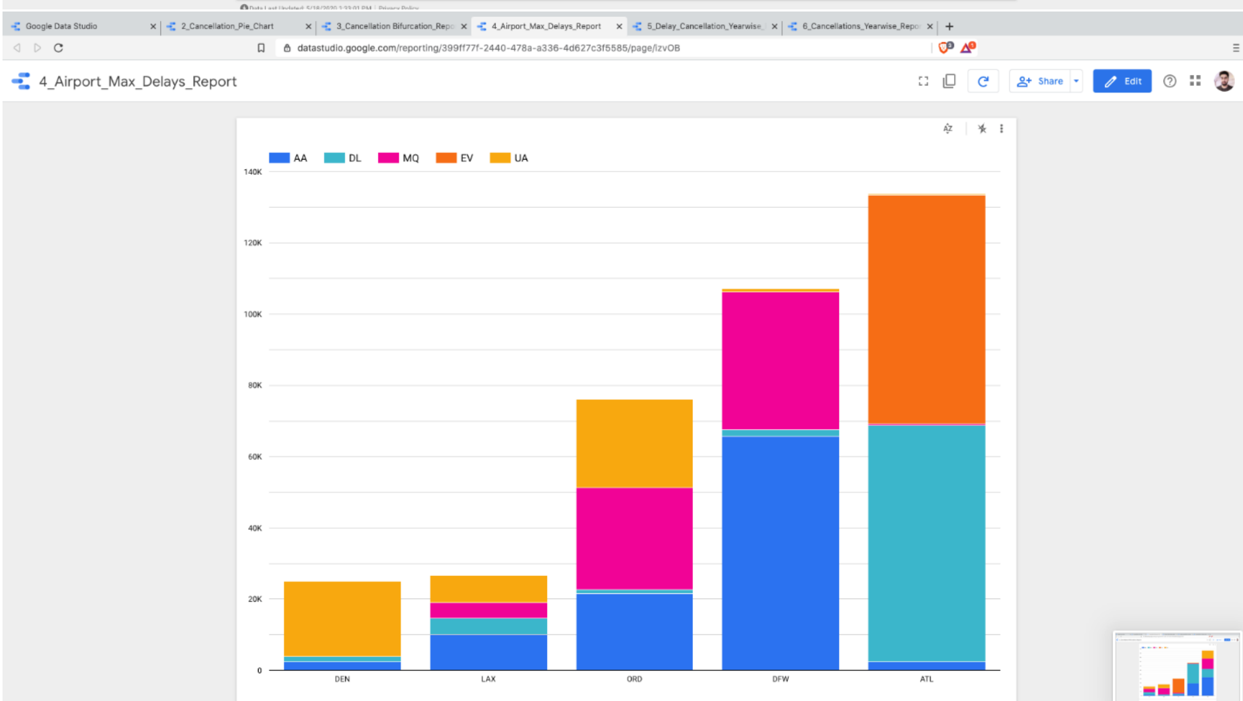
9.1.Performance Metrics



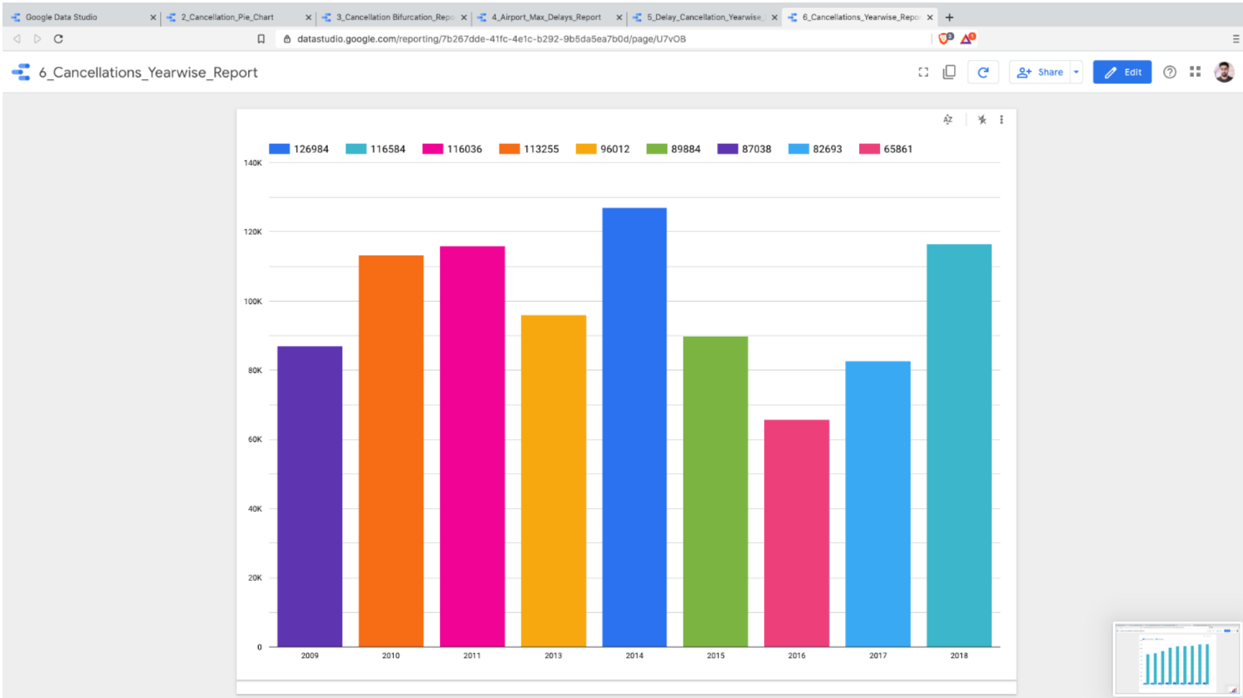




3_Cancellation Bifurcation_Report
datastudio.google.com/reporting/51e8370d-1da6-4c79-bfe9-9163d26a6cd3/page/3xvOB
MQ
AA
OO
DL
UA
55K
50K
45K
40K
35K
30K
25K
20K
15K
10K
5K
0
LAX
DEN
ATL
DFW
ORD



4_Airport_Max_Delays_Report
datastudio.google.com/reporting/399ff77f-2440-478a-a336-4d627c3f5585/page/lzvOB
AA
DL
MQ
EV
UA
140K
120K
100K
80K
60K
40K
20K
0
DEN
LAX
ORD
DFW
ATL



10.ADVANTAGES AND DISADVANTAGES

Advantages:

It can be used **to predict future glitches, prevent them from happening, and make the maintenance procedures more accurate and thorough**. As a result, it is possible to lower costs related to maintaining an aircraft. One of the companies using big data analytics this way is Boeing.

Disadvantages:

Airlines provide a vital service, but factors including the **continuing existence of loss-making carriers, bloated cost structure, vulnerability to exogenous events and a reputation for poor service** combine to present a huge impediment to profitability.

8.CONCLUSION

It can be used **to predict future glitches, prevent them from happening, and make the maintenance procedures more accurate and thorough**.

After analyzing the data, a lot of insights have been generated. Most of the delays and cancellations are due to three major reasons:

- Weather
- Airline/Carrier Issues
- National Air System

12.FUTURE SCOPE

Data Analytics eliminates guesswork and manual tasks. Be it choosing the right content, planning marketing campaigns, or developing products. Organizations can use the insights they gain from data analytics to make informed decisions. Thus, leading to better outcomes and customer satisfaction.

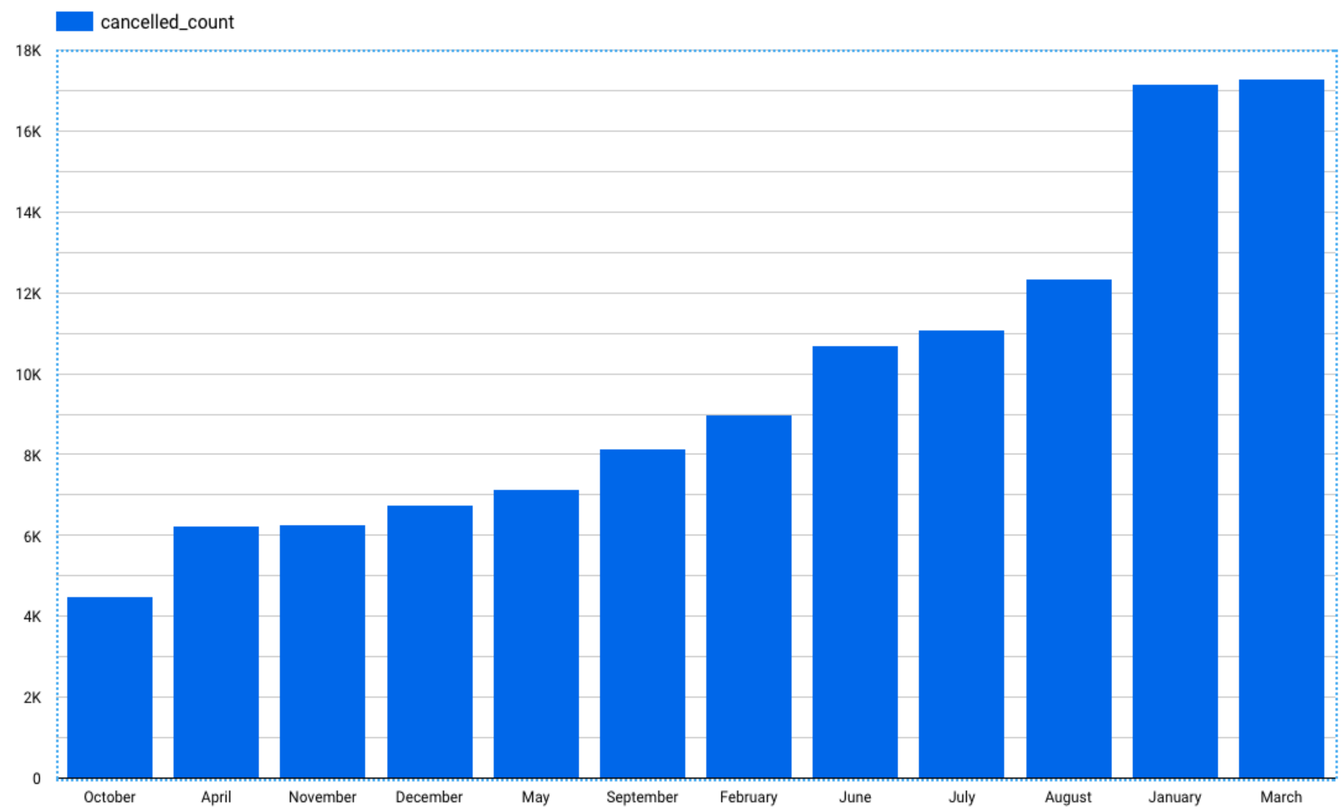
13.APPENDIX

SOURCE CODE

Most unreliable month (Cancellations in ascending order)

Query

```
WITH
  cancelled_count_cte AS (
  SELECT
    *,
    ROW_NUMBER() OVER (ORDER BY cancelled_count) AS RANK
  FROM (
    SELECT
      FORMAT_DATE('%B', FL_DATE) AS month,
      SUM(CANCELLED) AS cancelled_count
    FROM
      `airline-delay-canc.airlines_data.delay_canc_data`
    WHERE
      EXTRACT(year
      FROM
        FL_DATE) = 2018
    GROUP BY
      1) )
SELECT
  month,
  cancelled_count
FROM
  cancelled_count_cte
ORDER BY
  rank DESC
```



GITHUB LINK

https://github.com/IBM-EPBL/IBM-Project-45235-1660728963