

**BDM-1024: DATA TECHNOLOGY SOLUTIONS**

**PROJECT REPORT**

**SUBMITTED TO**

**BHAVIK GANDHI**



**Lambton**  
**College**

**SUBMITTED BY**

**GROUP E**

ABISHEAK DHANABAL(C0903766)  
ALWIN KANNYAKONIL SCARIA (C0894287)  
ASHNA VIJI ALEX(C0901082)  
JERIN THENGUMPALLIL THOMAS (C0896235)  
KUNCHERIA TOM(C0900973)  
PRINCE THOMAS (C0894907)  
MOHAMED AFTAB (C0891945)

# **CONTENT**

<b>INTRODUCTION</b>	<b>4</b>
<b>DATABASE</b>	<b>5</b>
<b>DECISION TREE</b>	<b>6</b>
<b>DATA MODEL DIAGRAM</b>	<b>7</b>
<b>SQL CODES AND INSIGHTS</b>	<b>8</b>
<b>MAPREDUCE</b>	<b>16</b>
<b>SPARK JOBS</b>	<b>19</b>
<b>VISUALIZATION</b>	<b>21</b>
<b>GITHUB REPOSITORY LINK</b>	<b>24</b>
<b>CONCLUSION</b>	<b>25</b>
<b>REFERENCES</b>	<b>26</b>

Type of activity	
Research Planning	<p><b>Kuncheria Tom:</b> Create a group in WhatsApp for group communication and compilation of the files needed for the presentation and report.</p> <p><b>Prince Thomas:</b> Identify the scope and objectives.</p> <p><b>Ashna Viji Alex:</b> Divide the topic into sub-topics that need to be covered.</p> <p><b>Jerin T Thomas:</b> Research specific sub-topics.</p> <p><b>Abhishek Dhanabal:</b> Set deadlines for each stage of the report process.</p>
Data Gathering	<p><b>Selected Flight analysis as main subject for the project together</b></p> <p><b>Prince Thomas:</b> Determine the specific areas of flight analysis to focus on.</p> <p><b>Ashna Viji Alex:</b> Gather data from various sources, such as official US Bureau of Transportation documentation, case studies, articles, etc.</p> <p><b>Alwin Kannyakonil Scaria:</b> Consolidate the gathered data via WhatsApp Group</p> <p><b>All group members:</b> Collaborate as a team and discuss the challenges encountered if there is any. The dataset is taken from US Bureau of Transportation website.</p>
ERD, DTD	<p><b>Ashna Viji Alex, Jerin T Thomas:</b> Developed Entity Relationship Diagram based on the dataset.</p> <p><b>Alwin Kannyakonil Scaria, Mohamed Aftab:</b> Developed Data Model Diagram</p>
Analysis	<p><b>Prince Thomas:</b> Analyze the gathered data and extract key insights features and use cases.</p> <p><b>All group members:</b> Compile the analysis and findings into a cohesive document.</p>
SQL Analysis	<b>Alwin Kannyakonil Scaria, Ashna Viji Alex, Jerin T Thomas:</b> Setup the dataset into PostgreSQL and analysed the dataset and created SQL queries, found out meaningful insights from the query outputs
MapReduce	<b>Jerin T Thomas, Kuncheria Tom:</b> Setup Hadoop cluster and ran MapReduce programs.
Spark Jobs	<b>Abhishek Dhanabal, Ashna Viji Alex, Mohamed Aftab:</b> Executed pyspark jobs in GCP dataproc to develop meaningful insights.
Visualization	<b>Prince Thomas, Kuncheria:</b> Obtain meaningful and dynamic dashboard and charts from tableau.
Report and Presentation PowerPoint Drafting	<p><b>All group members:</b> Collaborated on a shared platform to ensure consistency in style, formatting, and content.</p> <p><b>Kuncheria Tom:</b> Reviewed and provided feedback on each other's sections.</p>
Report Merging	<b>Ashna Viji Alex, Prince Thomas, Mohamed Aftab:</b> Consolidated and performed last editing.
Finalization	<b>Jerin T Thomas, Alwin Kannyakonil Scaria:</b> Overall check was done.
GitHub Repository	<b>Jerin T Thomas:</b> Created repository and updated all the files

## **INTRODUCTION**

In 2015, the Bureau of Transportation of federal agency in the United States, provided valuable data on domestic US airlines including various flights operators, flown flights, various airports, regarding the number of delays and cancellations. This data serves as a crucial resource for analysing the performance of airlines and identifying areas for optimization.

We look at the database as much as perspectives to identify the issues associated with existing flight schedule system, we primarily focused on resolving the root causes of delays, cancellations and diversions that cause immense operational cost, time and resource wastage rather than passenger analysis.

The analysis of delay and cancellation numbers can shed light on various factors affecting airline operations. By examining the data, airlines can identify common causes of delays such as weather conditions, air traffic congestion, airline issues. Understanding these patterns have a ripple nature, by predicting these causes we can make cost effective solution and improve the industry standards to a great extent.

## **DATABASE**

We found the data base from US Bureau of Transportation of size 658MB and has 4 tables as listed below.

### **1.Airports**

This Table holds the data of all airports having domestic terminals, airport code in addition to this it has the location details like city, state, and coordinates.

### **2. Airlines**

This table accommodates details like airline codes and airline carrier details.

### **3.Flight**

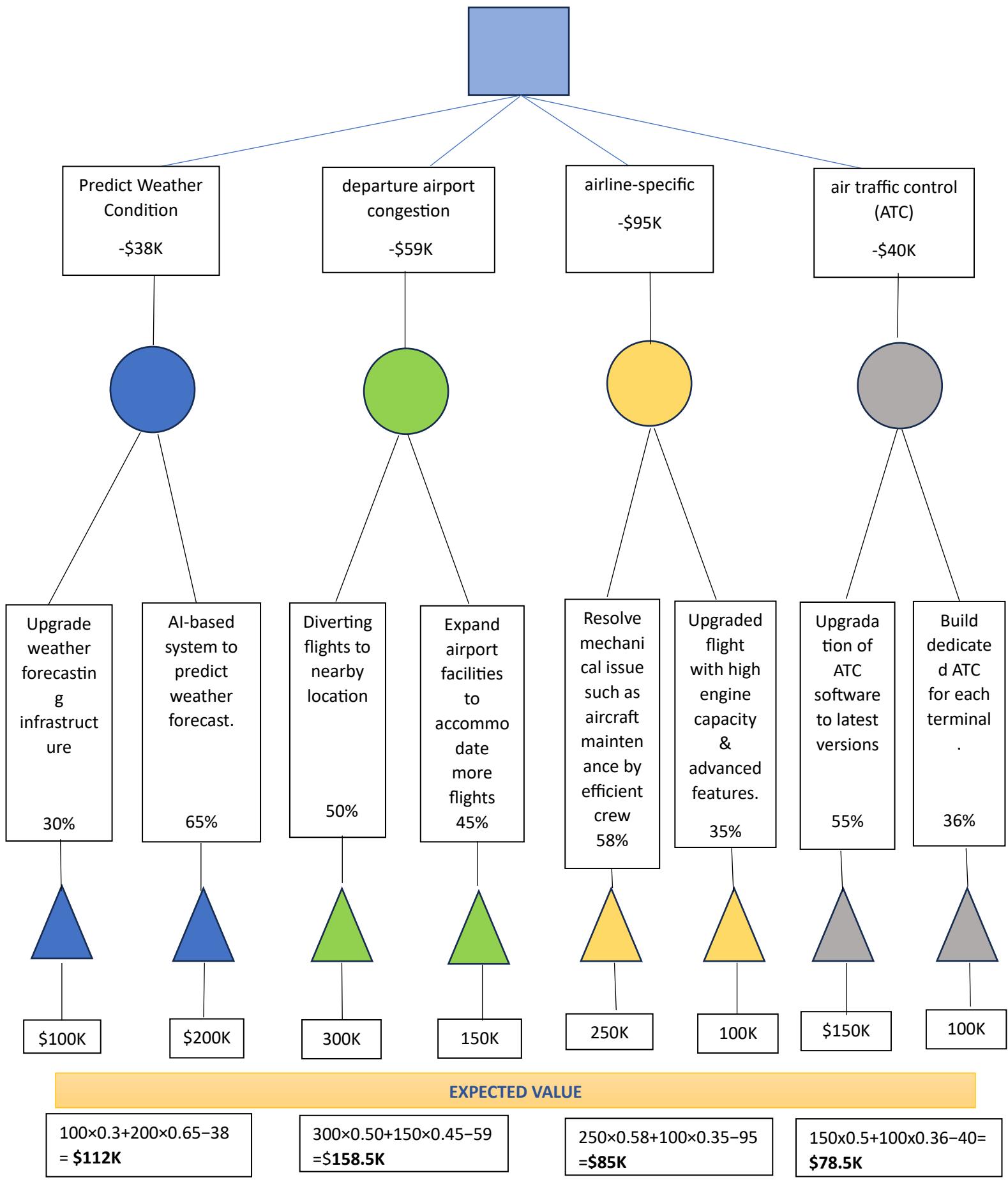
This is the main table that contains majority of data including the date of flight, source and destination airports, travel time, arrival and destination delays, cancellation code, diversion code, various factors of delay, taxi in, taxi out rates and many more.

### **4. Cancellation**

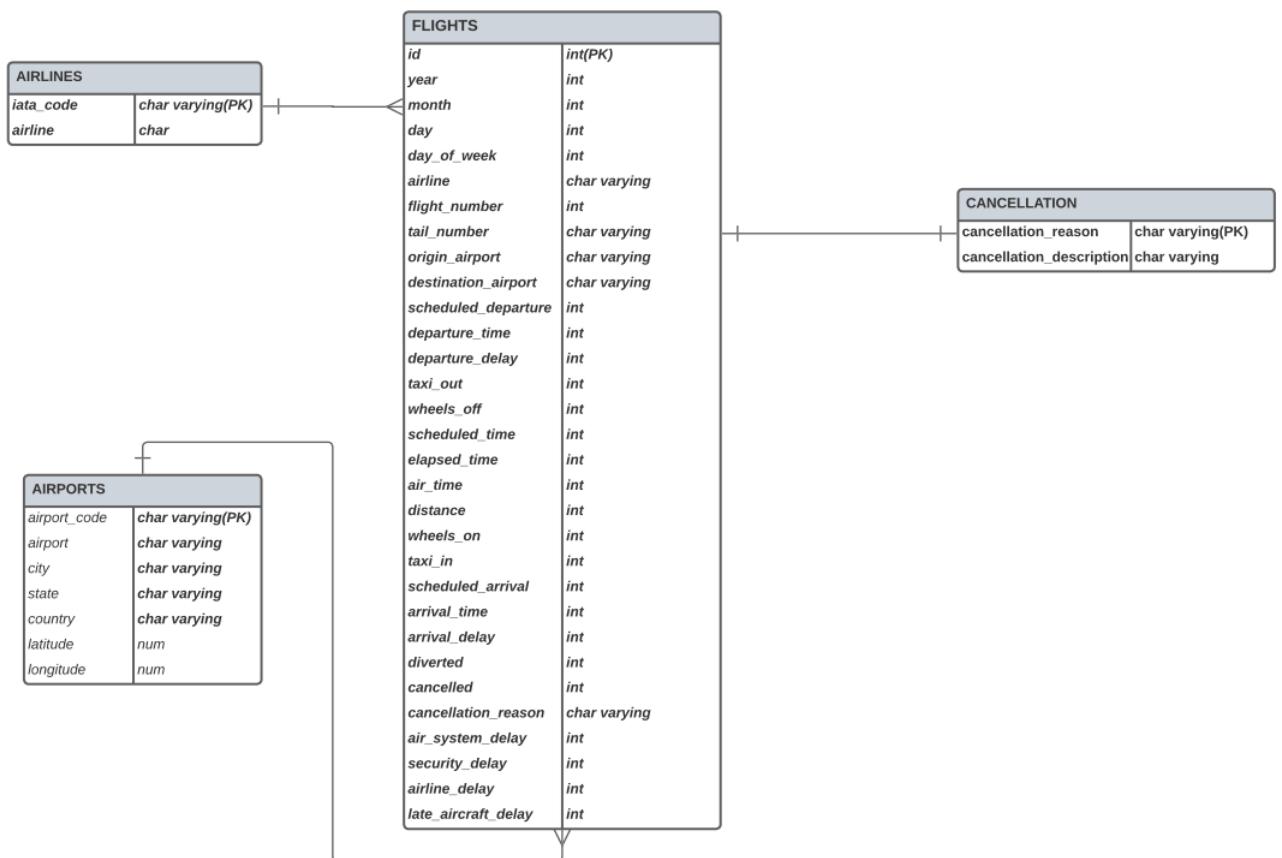
This table explains the details of each cancellation codes with description.

# DECISION TREE

## AIR TRAVEL OPTIMISATION



# DATA MODEL DIAGRAM



# SQL CODES & INSIGHTS

## 1. TIME VARIATION OF ALL FLIGHTS AND STATUS OF ARRIVAL

```

4 --TIME VARIATION OF ALL FLIGHTS AND STATUS OF ARRIVAL
5 select
6 ap.city as FROM_CITY,
7 ap.airport as from_airport,
8 fl.tail_number,
9 aps.city as destination_city,
10 aps.airport as destination_airport,
11 aline.airline,
12 fl.day_of_week,
13 (fl.departure_delay+fl.arrival_delay) TIME_VARIATION,
14 fl.arrival_delay,
15 fl.departure_delay,
16 CASE WHEN fl.arrival_delay<0 THEN 'DELAYED' WHEN fl.arrival_delay = 0 THEN 'ON TIME' WHEN fl.arrival_delay > 0 THEN 'EARLY ARRIVAL'ELSE 'NA'
17 from
18 flight fl
19 inner join airports ap on ap.airport_code = fl.origin_airport
20 inner join airports aps on aps.airport_code = fl.destination_airport
21 inner join airlines aline on aline.iata_code = fl.airline
22 order by day_of_week;
23

```

Data Output Messages Notifications

	from_city	from_airport	tail_number	destination_city	destination_airport	airline	day_of_week	time_variation	arrival_delay	departure_delay	status_completed
1	Detroit	Detroit Metropolitan Airport	N910EV	Grand Rapids	Gerald R. Ford International Airport	Skywest Airlines Inc.	1	-19	-14	-5	EARLY ARRIVAL
2	Pasco	Tri-Cities Airport	N616QX	Seattle	Seattle-Tacoma International Airport	Skywest Airlines Inc.	1	63	42	21	DELAYED
3	Las Vegas	McCarran International Airport	N689CA	Seattle	Seattle-Tacoma International Airport	Skywest Airlines Inc.	1	-27	-24	-3	EARLY ARRIVAL
4	Paducah	Barkley Regional Airport	N963SW	Chicago	Chicago O'Hare International Airport	Skywest Airlines Inc.	1	-13	-10	-3	EARLY ARRIVAL
5	Minneapolis	Minneapolis-Saint Paul International Airport	N151SY	Chicago	Chicago O'Hare International Airport	Skywest Airlines Inc.	1	27	26	1	DELAYED
6	San Francisco	San Francisco International Airport	N471UA	Las Vegas	McCarran International Airport	United Air Lines Inc.	1	-12	-11	-1	EARLY ARRIVAL
7	Chantilly	Washington Dulles International Airport	N8751Z	Denver	Denver International Airport	United Air Lines Inc.	1	-12	-12	0	EARLY ARRIVAL
8	Los Angeles	Los Angeles International Airport	N3413I	Newark	Newark Liberty International Airport	United Air Lines Inc.	1	-50	-41	-9	EARLY ARRIVAL
9	St Louis	St. Louis International Airport at Lambert Field	N833AW	Charlotte	Charlotte Douglas International Airport	American Airlines Inc.	1	-2	0	-2	ON TIME
10	Atlanta	Hartsfield-Jackson Atlanta International Airport	N593NW	Seattle	Seattle-Tacoma International Airport	Delta Air Lines Inc.	1	-53	-50	-3	EARLY ARRIVAL
11	Seattle	Seattle-Tacoma International Airport	N815DN	Salt Lake City	Salt Lake City International Airport	Delta Air Lines Inc.	1	-2	0	-2	ON TIME

Total rows: 1000 of 5332914 Query complete 00:00:19.932 Ln 4, Col 1

Analyse the general pattern of each flight on its delay by comparing the arrival & destination delays.

Insight: Majority of flights fail to departure/ arrive at scheduled time. Need to investigate more on this to identify the cause and need to look by airlines and airports statistics. Identified as serious issue in airline industry.

## 2. LIST OF ALL FLIGHTS WITH CANCELLATION

```

24 -- LIST OF ALL FLIGHTS WITH CANCELLATION
25 select
26 ap.city as from_city,
27 ap.airport as from_airport,
28 fl.tail_number,
29 aps.city as destination_city,
30 aps.airport as destination_airport,
31 aline.airline as Airline,
32 fl.day_of_week as Day_OF_WEEK,
33 fl.month as MONTH,
34 cancellation_description AS CANCELLATION_DESCRIPTION
35 from
36 flight fl
37 inner join airports ap on ap.airport_code = fl.origin_airport
38 inner join airports aps on aps.airport_code = fl.destination_airport
39 inner join airlines aline on aline.iata_code = fl.airline
40 inner join cancellation c on c.CANCELLATION_REASON = fl.CANCELLATION_REASON
41 where
42 fl.CANCELLED = 1
43 order by fl.day_of_week,fl.month;
44

```

Data Output Messages Notifications

	from_city	from_airport	tail_number	destination_city	destination_airport	airline	day_of_week	month	cancel char
1	San Francisco	San Francisco International Airport	[null]	Philadelphia	Philadelphia International Airport	United Air Lines Inc.	1	1	Weath
2	Seattle	Seattle-Tacoma International Airport	N371DA	New York	John F. Kennedy International Airport (New York International Airport)	Delta Air Lines Inc.	1	1	Weath
3	Phoenix	Phoenix Sky Harbor International Airport	N415WN	Philadelphia	Philadelphia International Airport	Southwest Airlines Co.	1	1	Weath
4	Philadelphia	Philadelphia International Airport	[null]	Las Vegas	McCarran International Airport	US Airways Inc.	1	1	Weath
5	Rochester	Greater Rochester International Airport	N516JB	New York	John F. Kennedy International Airport (New York International Airport)	JetBlue Airways	1	1	Weath
6	Covington	Cincinnati/Northern Kentucky International Airport	N665MQ	New York	John F. Kennedy International Airport (New York International Airport)	American Eagle Airlines Inc.	1	1	Weath
7	New York	LaGuardia Airport (Marine Air Terminal)	N371CA	Covington	Cincinnati/Northern Kentucky International Airport	Atlantic Southeast Airlines	1	1	Airline
8	San Francisco	San Francisco International Airport	N788AA	New York	John F. Kennedy International Airport (New York International Airport)	American Airlines Inc.	1	1	Weath
9	Grand Rapids	Gerald R. Ford International Airport	N607SW	Orlando	Orlando International Airport	Southwest Airlines Co.	1	1	Weath
10	Atlanta	Hartsfield-Jackson Atlanta International Airport	[null]	Philadelphia	Philadelphia International Airport	US Airways Inc.	1	1	Weath

Total rows: 1000 of 87430 Query complete 00:00:01.272 Ln 43, Col 34

Find the list of all flights cancelled and matched it with proper reasons to check the depth of the issue.

Insight: A sizable share of flights got cancelled due to weather and airline issues along with other reasons. Since it is a serious issue in Industry, need to find the patterns to cut down the numbers.

### 3. ANALYSIS OF DELAY ON DAILY BASIS

```

45 --ANALYSIS OF DELAY ON DAILY BASIS
46
47 SELECT
48   AP.AIRPORT,
49   aline.airline AS AIRLINE,
50   fl.DAY_OF_WEEK AS day_of_week,
51   fl.DAY AS day,
52   fl.MONTH AS month,
53   COUNT(*) AS num_delays,
54   ( case when fl.day_of_Week <=5 then'Weekday' else 'Weekend' end)as weektype
55
56 FROM
57 flight fl
58 inner join airports ap on ap.airport_code = fl.origin_airport
59 inner join airlines aline on aline.iata_code = fl.airline
60 WHERE
61   departure_delay > 0
62 GROUP BY
63   AP.AIRPORT,
64   aline.airline,
65   fl.DAY_OF_WEEK,
66   fl.DAY,
67   fl.MONTH
68 ORDER BY
69   num_delays DESC;
70
71 Data Output Messages Notifications
72
73
74 i | airport character varying | airline character varying | day_of_week integer | day integer | month integer | num_delays bigint | weektype text |
75 c | character varying | character varying | integer | integer | integer | bigint | text |
76 1 | Hartsfield-Jackson Atlanta International Airport.. Delta Air Lines Inc. | 7 | 23 | 8 | 577 | Weekend |
77 2 | Hartsfield-Jackson Atlanta International Airport.. Delta Air Lines Inc. | 3 | 30 | 12 | 574 | Weekday |
78 3 | Hartsfield-Jackson Atlanta International Airport.. Delta Air Lines Inc. | 5 | 10 | 4 | 552 | Weekday |
79 4 | Hartsfield-Jackson Atlanta International Airport.. Delta Air Lines Inc. | 2 | 24 | 2 | 488 | Weekday |
80 5 | Hartsfield-Jackson Atlanta International Airport.. Delta Air Lines Inc. | 4 | 5 | 3 | 484 | Weekday |
81 6 | Hartsfield-Jackson Atlanta International Airport.. Delta Air Lines Inc. | 2 | 18 | 8 | 472 | Weekday |
82
83
Total rows: 1000 of 248172 Query complete 00:00:06.113 Ln 45, Col 1

```

Attempt to identify the patterns in delay on weekday and weekends.

Insight: Major share of delays are shown in weekdays except few cases. Need to use larger carrier to accommodate more passengers and engine power on relevant days and reduce the number of smaller flights to reduce the traffic.

### 4. ANALYSIS by AIRLINE ON ARIVAL-DEPARTURE DELAY

```

71 -- ANALYSIS by AIRLINE ON ARIVAL-DEPARTURE DELAY
72
73 select
74   aline.airline AIRLINE,
75   COUNT(*) Number_of_Flights,
76   SUM(CASE WHEN fl.departure_delay > 0 THEN 1 ELSE 0 END) Number_Of_Delays_At_Departure,
77   SUM(CASE WHEN fl.arrival_delay > 0 THEN 1 ELSE 0 END) Number_Of_Delays_At_Arrival,
78   SUM(CASE WHEN fl.departure_delay < 0 THEN 1 ELSE 0 END) Number_of_Early_Departure,
79   SUM(CASE WHEN fl.arrival_delay < 0 THEN 1 ELSE 0 END) Number_of_Early_Arrival,
80   SUM(CASE WHEN fl.departure_delay = 0 and fl.arrival_delay = 0 THEN 1 ELSE 0 END) On_time_flights
81
82 from
83   airlines aline
84   inner join flight fl on aline.iata_code = fl.airline
85
86 group by aline.airline
87
88 Data Output Messages Notifications
89
90
91 i | airline character varying | number_of_flights bigint | number_of_delays_at_departure bigint | number_of_delays_at_arrival bigint | number_of_early_departure bigint | number_of_early_arrival bigint | on_time_flights bigint |
92 c | character varying | bigint | bigint | bigint | bigint | bigint | bigint |
93 1 | Alaska Airlines Inc. | 172521 | 43566 | 56953 | 121253 | 110162 | 306 |
94 2 | American Airlines Inc. | 725984 | 245904 | 252191 | 436039 | 446655 | 916 |
95 3 | American Eagle Airlines Inc. | 294632 | 93726 | 103505 | 171512 | 170102 | 522 |
96 4 | Atlantic Southeast Airlines | 571977 | 169897 | 213217 | 365593 | 328301 | 924 |
97 5 | Delta Air Lines Inc. | 875881 | 282463 | 250840 | 522581 | 601859 | 1697 |
98 6 | Frontier Airlines Inc. | 90836 | 34893 | 41232 | 52144 | 46673 | 114 |
99 7 | Hawaiian Airlines Inc. | 76272 | 20146 | 30179 | 52841 | 42545 | 272 |
100 8 | JetBlue Airways | 267048 | 102061 | 101998 | 147160 | 155170 | 403 |
101 9 | Skywest Airlines Inc. | 588353 | 171572 | 222435 | 380570 | 339967 | 1094 |
102 10 | Southwest Airlines Co. | 1261855 | 566807 | 470767 | 587783 | 742332 | 3093 |
103 11 | Spirit Air Lines | 117379 | 52089 | 56887 | 58663 | 55893 | 154 |
104 12 | United Air Lines Inc. | 515723 | 256550 | 186227 | 224854 | 312159 | 644 |
105 13 | US Airways Inc. | 198715 | 62565 | 76285 | 123032 | 113399 | 314 |
106 14 | Virgin America | 41000 | 20000 | 20100 | 20100 | 20000 | 100 |
107
108
109
Total rows: 14 of 14 Query complete 00:00:01.704 Ln 79, Col 36

```

Attempt to find time management of each airline in whole year.

Insight: Only a minority (less than 1%) could fly on scheduled time and rest either flew late or early. Checking further on next query.

## 5. ANALYSIS BY AIRLINE BY DELAY CAUSE

```

88 --ANALYSIS BY AIRLINE BY DELAY CAUSE
89
90 select
91 aline.airline AIRLINE,
92 count(*) Number_of_Flights,
93 SUM(CASE WHEN fl.AIR_SYSTEM_DELAY > 0 THEN 1 ELSE 0 END) Number_of_Delays_by_AIR_SYSTEM_DELAY,
94 SUM(CASE WHEN fl.SECURITY_DELAY > 0 THEN 1 ELSE 0 END) Number_of_Delays_by_SECURITY_DELAY,
95 SUM(CASE WHEN fl.AIRLINE_DELAY > 0 THEN 1 ELSE 0 END) Number_of_Delays_by_AIRLINE_DELAY,
96 SUM(CASE WHEN fl.LATE_AIRCRAFT_DELAY > 0 THEN 1 ELSE 0 END) Number_of_Delays_by_LATE_AIRCRAFT_DELAY,
97 SUM(CASE WHEN fl.WEATHER_DELAY > 0 THEN 1 ELSE 0 END) Number_of_Delays_by_WEATHER_DELAY
98 from
99 airlines aline
100 inner join flight fl on aline.iata_code = fl.airline
101 group by aline.airline
102
103 Data Output Messages Notifications
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
Total rows: 14 of 14 Query complete 00:00:01.616 Ln 89, Col 1

```

airline	number_of_flights	number_of_delays_by_air_system_delay	number_of_delays_by_security_delay	number_of_delays_by_airline_delay	number_of_delays_by_late_aircraft_delay	number_of_delays_by_weather_delay
1 Alaska Airlines Inc.	172521	14897	250	8363	7866	867
2 American Airlines Inc.	725984	70383	731	68656	57590	9967
3 American Eagle Airlines Inc.	294632	35806	308	27236	32125	9105
4 Atlantic Southeast Airlines	571977	63586	0	53943	54804	2852
5 Delta Air Lines Inc.	875881	64230	58	63128	50112	11838
6 Frontier Airlines Inc.	90836	18066	0	11517	11638	580
7 Hawaiian Airlines Inc.	76272	325	29	7395	4877	572
8 JetBlue Airways	267048	33913	433	38665	31684	2173
9 Skywest Airlines Inc.	588353	58509	345	39127	62432	4426
10 Southwest Airlines Co.	1261855	97772	554	144524	163188	10086
11 Spirit Airlines	117379	28378	299	17307	11722	925
12 United Air Lines Inc.	515723	49858	26	65999	49865	7531
13 US Airways Inc.	198715	23708	372	19717	13654	1722
14 Virgin America	61903	5395	79	4445	5396	2072

Checking on each airline on total number of delays and the share of various causes in total.

Insight: Least issues are reported by security issues and majority of airlines are affected by either flight issues or air traffic issues. Weather also has a share in delay.

## 6. ANALYSIS by AIRLINE ON CANCELLATION and RESAONS

```

Query History
103 -- ANALYSIS by AIRLINE ON CANCELLATION and RESAONS
104
105 select
106 (SELECT Count(*) from Flight f where aline.iata_code = f.airline group by f.airline) Total_number_of_Flights,
107 SUM(CASE WHEN fl.CANCELLATION_REASON = 'A' THEN 1 else 0 END) total_Flights_Cancelled,
108 SUM(CASE WHEN fl.CANCELLATION_REASON = 'B' THEN 1 else 0 END) CANCELLED_BY_ISSUE_AirlineOrCarrier,
109 SUM(CASE WHEN fl.CANCELLATION_REASON = 'C' THEN 1 else 0 END) CANCELLED_BY_ISSUE_Weather,
110 SUM(CASE WHEN fl.CANCELLATION_REASON = 'D' THEN 1 else 0 END) CANCELLED_BY_ISSUE_National_Air_System,
111 SUM(CASE WHEN fl.CANCELLATION_REASON = 'E' THEN 1 else 0 END) CANCELLED_BY_ISSUE_Security
112 from
113 airlines aline
114 inner join flight fl on aline.iata_code = fl.airline
115 inner join cancellation c on c.CANCELLATION_REASON = fl.CANCELLATION_REASON
116 group by aline.airline,aline.iata_code;
117
118 Data Output Messages Notifications
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
Total rows: 14 of 14 Query complete 00:00:34.304 Ln 104, Col 1

```

airline	total_number_of_flights	total_flights_cancelled	cancelled_by_issue_airlinecarrier	cancelled_by_issue_weather	cancelled_by_issue_national_air_system	cancelled_by_issue_security
1 American Airlines Inc.	725984	10919	2879	7306	730	4
2 Alaska Airlines Inc.	172521	669	334	317	18	0
3 JetBlue Airways	267048	4276	883	2464	928	1
4 Delta Air Lines Inc.	875881	3824	594	2973	257	0
5 Atlantic Southeast Airlines	571977	15231	3604	5082	6544	1
6 Frontier Airlines Inc.	90836	588	308	280	0	0
7 Hawaiian Airlines Inc.	76272	171	170	1	0	0
8 American Eagle Airlines Inc.	294632	15025	2475	9164	3385	1
9 Spirit Airlines	117379	2004	654	1068	279	3
10 Skywest Airlines Inc.	588353	9960	3205	5539	1216	0
11 United Air Lines Inc.	515723	6573	2870	3312	391	0
12 US Airways Inc.	198715	4067	1007	2490	570	0
13 Virgin America	61903	534	157	12	365	0
14 Southwest Airlines Co.	1261855	16043	6122	8843	1066	12

Checking on each airline on total number of cancellation and the share of various causes in total.

Insight: Airlines are primarily affected by either weather issues or carrier/traffic issues.

## 7. TOP 10 DELAYED ROUTES

Query History

```
--Top 10 delayed routes
SELECT
    ap.airport || ' to ' || aps.airport AS route,
    COUNT(*) AS no_of_delayed_flights,
    SUM(CASE WHEN fl.AIR_SYSTEM_DELAY > 0 THEN 1 ELSE 0 END) Number_of_Delay_by_AIR_SYSTEM_DELAY,
    SUM(CASE WHEN fl.SECURITY_DELAY > 0 THEN 1 ELSE 0 END) Number_of_Delays_by_SECURITY_DELAY,
    SUM(CASE WHEN fl.AIRLINE_DELAY > 0 THEN 1 ELSE 0 END) Number_of_Delays_by_AIRLINE_DELAY,
    SUM(CASE WHEN fl.LATE_AIRCRAFT_DELAY > 0 THEN 1 ELSE 0 END) Number_of_Delays_by_LATE_AIRCRAFT_DELAY,
    SUM(CASE WHEN fl.WEATHER_DELAY > 0 THEN 1 ELSE 0 END) Number_of_Delays_by_WEATHER_DELAY
FROM flight fl
INNER JOIN airports ap ON ap.airport_code = fl.origin_airport
INNER JOIN airports aps ON aps.airport_code = fl.destination_airport
WHERE departure_delay > 0
GROUP BY ap.airport, aps.airport
ORDER BY no_of_delayed_flights DESC
LIMIT 10;
```

Data Output Messages Notifications

route	no_of_delayed_flights	number_of_d	number_of_dk	number_of_delays_by_late_aircraft_delay	number_of_delays_by_weather_delay
text	bigrint	bigrint	bigrint	bigrint	bigrint
1 Los Angeles International Airport to San Francisco International Airport	6004	1780	6	1367	1895
2 San Francisco International Airport to Los Angeles International Airport	5552	1889	8	1353	2111
3 Los Angeles International Airport to John F. Kennedy International Airport (New York International Airp...	4497	1026	9	988	718
4 Chicago O'Hare International Airport to Los Angeles International Airport	4415	985	10	1320	899
5 Chicago O'Hare International Airport to San Francisco International Airport	4282	937	2	1106	808
6 Los Angeles International Airport to McCarran International Airport	4058	955	4	1150	1501
7 Chicago O'Hare International Airport to LaGuardia Airport (Marine Air Terminal)	4048	1312	3	912	916
8 McCarran International Airport to Los Angeles International Airport	4002	1529	3	1189	1469
9 John F. Kennedy International Airport (New York International Airport) to Los Angeles International Air...	3943	887	20	1027	467
10 McCarran International Airport to San Francisco International Airport	3660	977	2	735	1081

Total rows: 10 of 10 Query complete 00:00:05.072 Ln 147, Col 10

## Checking the highly delayed top 10 routes in U.S

Insight: Los Angeles to and from San Francisco is the highly delayed route, as well as routes involving airports like John F Kennedy, Chicago and McCarran are reported to be most delayed.

## 8. TOP 10 CANCELLED ROUTES

--Top 10 Cancelled routes

```
SELECT
    ap.airport || ' to ' || aps.airport AS route,
    COUNT(*) AS no_of_cancelled_flights,
    SUM(CASE WHEN fl.CANCELLATION_REASON = 'A' THEN 1 ELSE 0 END) CANCELLED_BY_ISSUE_AirlineORCarrier,
    SUM(CASE WHEN fl.CANCELLATION_REASON = 'B' THEN 1 ELSE 0 END) CANCELLED_BY_ISSUE_Weather,
    SUM(CASE WHEN fl.CANCELLATION_REASON = 'C' THEN 1 ELSE 0 END) CANCELLED_BY_ISSUE_National_Air_System,
    SUM(CASE WHEN fl.CANCELLATION_REASON = 'D' THEN 1 ELSE 0 END) CANCELLED_BY_ISSUE_Security
FROM flight fl
INNER JOIN airports ap ON ap.airport_code = fl.origin_airport
INNER JOIN airports aps ON aps.airport_code = fl.destination_airport
WHERE cancelled = 1
GROUP BY ap.airport, aps.airport
ORDER BY no_of_cancelled_flights DESC
LIMIT 10;
```

Data Output Messages Notifications

route	no_of_delayed_flights	number_of_d	number_of_dk	number_of_delays_by_late_aircraft_delay	number_of_delays_by_weather_delay
text	bigrint	bigrint	bigrint	bigrint	bigrint
1 Los Angeles International Airport to San Francisco International Airport	6004	1780	6	1367	1895
2 San Francisco International Airport to Los Angeles International Airport	5552	1889	8	1353	2111
3 Los Angeles International Airport to John F. Kennedy International Airport (New York International Airp...	4497	1026	9	988	718
4 Chicago O'Hare International Airport to Los Angeles International Airport	4415	985	10	1320	899
5 Chicago O'Hare International Airport to San Francisco International Airport	4282	937	2	1106	808
6 Los Angeles International Airport to McCarran International Airport	4058	955	4	1150	1501
7 Chicago O'Hare International Airport to LaGuardia Airport (Marine Air Terminal)	4048	1312	3	912	916
8 McCarran International Airport to Los Angeles International Airport	4002	1529	3	1189	1469
9 John F. Kennedy International Airport (New York International Airport) to Los Angeles International Air...	3943	887	20	1027	467
10 McCarran International Airport to San Francisco International Airport	3660	977	2	735	1081

## To identify the top 10 cancelled routes

Insight: The Los Angeles to San Francisco route has been reported as one of the routes with significant cancellation. Additionally, routes involving airports like John F Kennedy, Chicago, and McCarran have also been noted for experiencing a higher number of cancellations.

## 9. TOP 10 DIVERTED ROUTE

```
168 --Top 10 diverted route
169
170
171 ap.airport || ' to ' || aps.airport AS route,
172 COUNT(*) AS no_of_diverted_flights
173 FROM flight fl
174 inner join airports ap on ap.airport_code = fl.origin_airport
175 inner join airports aps on aps.airport_code = fl.destination_airport
176 WHERE diverted = 1
177 GROUP BY ap.airport, aps.airport
178 ORDER BY no_of_diverted_flights DESC
179 LIMIT 10;
180
181 Data Output Messages Notifications
182
183
184
185 route
186 text
187 no_of_diverted_flights
188 begin
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
d
e Total rows: 10 of 10 Query complete 00:00:00.780 Ln 170, Col 1
```

## Identifying the top 10 diverted routes

Insight: The highest number of flights are diverted on the route to LaGuardia Airport.

## 10. TOP 10 FLOWN ROUTES

```
--Top 10 flown routes

SELECT
ap.airport || ' to ' || aps.airport AS route,
COUNT(*) AS no_of_flights
FROM flight f
inner join airports ap on ap.airport_code = f.l.origin_airport
inner join airports aps on aps.airport_code = f.l.destination_airport
GROUP BY ap.airport, aps.airport
ORDER BY no_of_flights DESC
LIMIT 10
```

Data Output Messages Notifications

route	no_of_flights
San Francisco International Airport to Los Angeles International Airport	13744
Los Angeles International Airport to San Francisco International Airport	13457
John F. Kennedy International Airport (New York International Airport) to Los Angeles International Airport	12016
Los Angeles International Airport to John F. Kennedy International Airport (New York International Airport)	12015
McCarran International Airport to Los Angeles International Airport	9715
LaGuardia Airport (Marine Air Terminal) to Chicago O'Hare International Airport	9639
Los Angeles International Airport to McCarran International Airport	9594
Chicago O'Hare International Airport to LaGuardia Airport (Marine Air Terminal)	9575
San Francisco International Airport to John F. Kennedy International Airport (New York International Air...	8440
John F. Kennedy International Airport (New York International Airport) to San Francisco International Air...	8437

Rows: 10 of 10 Query complete 00:00:01.682 Ln 183, Col 1

This identifies the route with the highest number of flights.

Insight: San Francisco and Los Angeles Airport is having the highest number of flown flights and John F Kennedy and Los Angeles route is having the second highest number of flights.

## 11. MOST DELAYED MONTH

```
--Most Delayed Month
SELECT month, COUNT(<>) AS count
FROM flight
WHERE arrival_delay > 0
GROUP BY month
ORDER BY count DESC
```

month	count
1	206989
2	199717
3	190133
4	188310
5	180891
6	179494
7	175443
8	175178
9	171820
10	149439
11	141250
12	133432

To find the month which is having the highest number of delayed flights.

Insight: June and July months are having the highest number of delays.

## 12. MOST ON-TIME MONTH

```
--Most On-Time Month
SELECT month, COUNT(<>) AS count
FROM flight
WHERE arrival_delay = 0
GROUP BY month
ORDER BY count DESC
```

month	count
3	11477
4	11324
7	11319
6	11158
8	10955
10	10832
5	10597
1	10126
9	9956
12	9745
11	9660
2	9084

Identify the month having the highest number of flights arrived on time.

Insight: The month of march and April is having the highest number of flights arrived on time

## 13. MOST TRAVELED DAYS IN A WEEK

```

209
210 -- Most Travelled Days in a Week in US
211 SELECT day_of_week, Sum(distance) as Total_distance FROM flight
212 GROUP BY day_of_week
213 ORDER BY Total_distance DESC;
214

```

Data Output Messages Notifications

	day_of_week	total_distance
	integer	bigint
1	4	714260507
2	1	706583476
3	5	705225205
4	3	694803803
5	2	683362562
6	7	680372853
7	6	600749003

To find which day of the week is having the most distance travelled.

Insight: Thursday followed by Monday and then Friday is having the highest

## 14. NUMBER OF MORNINGS, EVENING ARRIVALS AND DELAYS ANALYSIS BY AIRPORT MONTH WISE

```

--number of morning, evening arrival and delay analysis by airport monthwise
SELECT
ap.airport,month,
SUM(CASE WHEN arrival_time < 1200 THEN 1 ELSE 0 END) AS morning_arrival,
SUM(CASE WHEN arrival_time >= 1200 THEN 1 ELSE 0 END) AS evening_arrival,
SUM(CASE WHEN arrival_time < 1200 THEN 1 ELSE 0 END) + SUM(CASE WHEN arrival_time >= 1200 THEN 1 ELSE 0 END) AS total_arrivals,
CASE
WHEN SUM(CASE WHEN arrival_time < 1200 THEN 1 ELSE 0 END) < SUM(CASE WHEN arrival_time >= 1200 THEN 1 ELSE 0 END) THEN 'Evening'
WHEN SUM(CASE WHEN arrival_time < 1200 THEN 1 ELSE 0 END) > SUM(CASE WHEN arrival_time >= 1200 THEN 1 ELSE 0 END) THEN 'Morning'
ELSE 'Equal' END AS higher_arrival_time,
SUM(CASE WHEN arrival_delay > 0 THEN 1 ELSE 0 END) AS delayed,
SUM(CASE WHEN arrival_delay = 0 THEN 1 ELSE 0 END) AS on_time,
SUM(CASE WHEN arrival_delay < 0 THEN 1 ELSE 0 END) AS early
FROM flight inner join airports ap on ap.airport_code = destination_airport
group by airport,month
order by airport,month ;

```

Data Output Messages Notifications

airport	month	morning_arrival	evening_arrival	total_arrivals	higher_arrival_time	delayed	on_time	early
	character varying	integer	bigint	bigint	text	bigint	bigint	bigint
Aberdeen Regional Airport	1	3	57	60	Evening	15	0	45
Aberdeen Regional Airport	2	0	55	55	Evening	11	1	41
Aberdeen Regional Airport	3	0	61	61	Evening	8	1	52
Aberdeen Regional Airport	4	1	58	59	Evening	6	1	52
Aberdeen Regional Airport	5	0	62	62	Evening	9	3	49
Aberdeen Regional Airport	6	1	58	59	Evening	18	1	40
Aberdeen Regional Airport	7	1	61	62	Evening	25	2	34
Aberdeen Regional Airport	8	0	61	61	Evening	17	3	41
Aberdeen Regional Airport	9	0	60	60	Evening	17	1	42
Aberdeen Regional Airport	10	4	58	62	Evening	23	2	36
Aberdeen Regional Airport	11	3	58	61	Evening	14	1	46
	12	...	...	...	...	...	...	...

rows: 1000 of 3417 Query complete 00:00:01.654

Ln 216, Col 1

Insight: The highest number of airplanes arriving at airports is consistently more in the second half of the day and the on-time arrivals are significantly low in number.

## 15. TOP 10 DELAYED ROUTES BY AIRCRAFT ISSUES

```

235
236 SELECT
237 ap.airport || ' to ' || aps.airport AS route,
238 al.airline,
239 COUNT(*) AS no_of_delayed_flights,
240 SUM(CASE WHEN fl.AIRLINE_DELAY > 0 THEN 1 ELSE 0 END) Number_Of_Delays_by_AIRLINE_DELAY,
241 SUM(CASE WHEN fl.LATE_AIRCRAFT_DELAY > 0 THEN 1 ELSE 0 END) Number_of_Delays_by_LATE_AIRCRAFT_DELAY
242 FROM flight fl
243 INNER JOIN airports ap ON ap.airport_code = fl.origin_airport
244 INNER JOIN airports aps ON aps.airport_code = fl.destination_airport
245 INNER JOIN airlines al ON al.iata_code = fl.airline
246 WHERE departure_delay > 0 AND fl.AIRLINE_DELAY > 0 AND fl.AIRLINE_DELAY > 0
247 GROUP BY ap.airport, aps.airport, al.airline
248 ORDER BY no_of_delayed_flights DESC
249 LIMIT 10
250

```

Data Output Messages Notifications

route	airline	no_of_delayed_flights	number_of_delays_by_airline_delay	number_of_delays_by_late_aircraft_delay
Kahului Airport to Honolulu International Airport	Hawaiian Airlines Inc.	886	886	806
Dallas Love Field to William P Hobby Airport	Southwest Airlines Co.	869	869	545
George Bush Intercontinental Airport to Los Angeles International Airport	United Air Lines Inc.	820	820	399
William P Hobby Airport to Dallas Love Field	Southwest Airlines Co.	804	804	502
Dallas/Fort Worth International Airport to Los Angeles International Airport	American Airlines Inc.	773	773	320
Chicago O'Hare International Airport to San Francisco International Airport	United Air Lines Inc.	749	749	366
Chicago O'Hare International Airport to Los Angeles International Airport	United Air Lines Inc.	723	723	362
Newark Liberty International Airport to San Francisco International Airport	United Air Lines Inc.	721	721	350
McCarran International Airport to Los Angeles International Airport	Southwest Airlines Co.	714	714	484
San Francisco International Airport to Chicago O'Hare International Airport	United Air Lines Inc.	708	708	396

Total rows: 10 of 10 Query complete 00:00:01.810

Ln 235, Col 1

Insight: Hawaiian Airlines is the aircraft causing the highest number of delays in the Kahului to Honolulu airport and United Air Lines is the aircraft having the highest number of delays

## 16. ACTUAL NUMBER OF FLIGHTS FLOWN BY EACH AIRLINE ON EACH MONTH

```

252 -- Actual number of flights flown by each airlines on each month
253 select
254 aline.airline AIRLINE,
255 count(*) Number_of_Flights,
256 SUM(CASE WHEN fl.cancelled = 1 THEN 1 ELSE 0 END) no_of_cancelled,
257 (count(*) - SUM(CASE WHEN fl.cancelled = 1 THEN 1 ELSE 0 END)) ACTUAL_FLIGHTS
258 from
259 airlines aline
260 inner join flight fl on aline.iata_code = fl.airline
261 group by aline.airline;

```

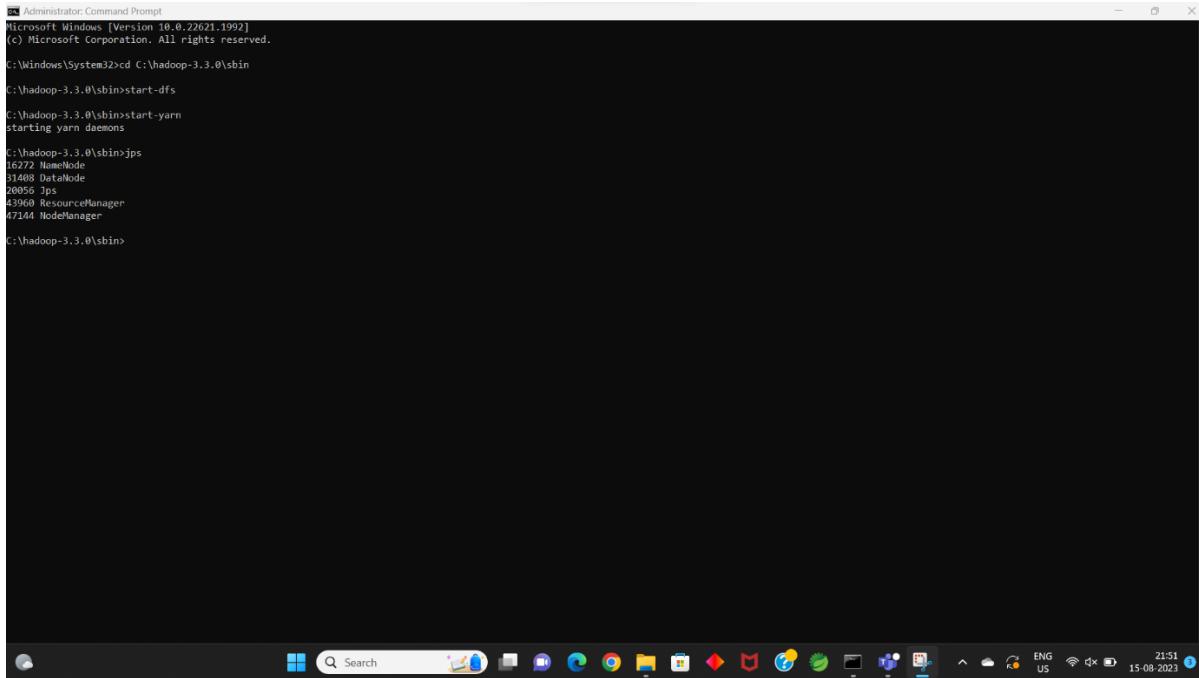
Data Output Messages Notifications

airline	number_of_flights	no_of_cancelled	actual_flights
Alaska Airlines Inc.	172521	669	171852
American Airlines Inc.	725984	10919	715065
American Eagle Airlines Inc.	294632	15025	279607
Atlantic Southeast Airlines	571977	15231	556746
Delta Air Lines Inc.	875881	3824	872057
Frontier Airlines Inc.	90836	588	90248
Hawaiian Airlines Inc.	76272	171	76101
JetBlue Airways	267048	4276	262772
Skywest Airlines Inc.	588353	9960	578393
Southwest Airlines Co.	1261855	16043	1245812
Spirit Airlines	117379	2004	115375
United Air Lines Inc.	515723	6573	509150
US Airways Inc.	198715	4067	194648
Virgin America	61903	534	61369

Insight: On observing the above data we could find that after reducing the cancelled flights from scheduled trips we can find a significant reduction in the actual flights flown.

# MAPREDUCE JOBS

Hadoop cluster set-up was done. All nodes and daemons are up and running.



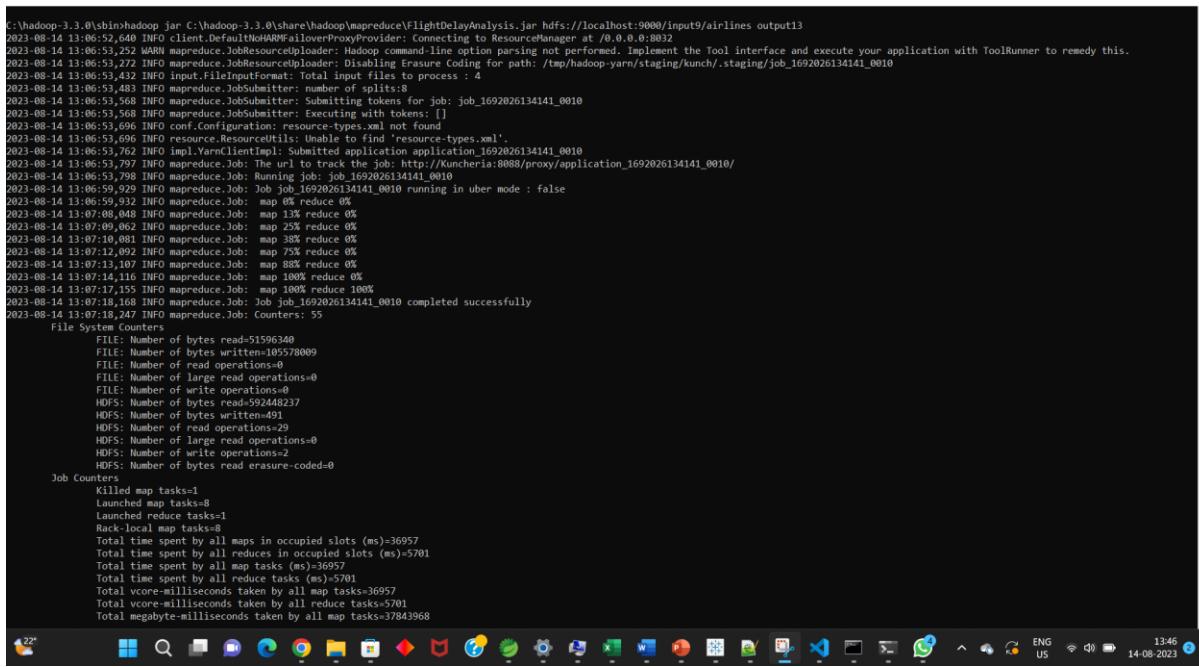
```
C:\Administrator: Command Prompt
Microsoft Windows [Version 10.0.22621.1992]
(c) Microsoft Corporation. All rights reserved.

C:\Windows\System32>cd C:\hadoop-3.3.0\sbin
C:\hadoop-3.3.0\sbin>start-yarn
starting yarn daemons

C:\hadoop-3.3.0\sbin>jps
16722 NameNode
31408 DataNode
20056 Jps
43969 ResourceManager
47344 NodeManager
C:\hadoop-3.3.0\sbin>
```

Few MapReduce jobs was done on the running Hadoop cluster and found some insights. Below are some of the insights got while running the MapReduce jobs:

## 1. Average delay by airline



```
C:\hadoop-3.3.0\sbin>hadoop jar C:\hadoop-3.3.0\share\hadoop\mapreduce\flightDelayAnalysis.jar hdfs://localhost:9000/input9/airlines output13
2023-08-14 13:06:52,640 INFO client.DefaultHDFSailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8082
2023-08-14 13:06:53,252 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
2023-08-14 13:06:53,252 INFO mapreduce.JobResourceUploader: Uploading Erasure Coding for path: /tmp/hadoop-yarn/staging/kunchi/.staging/job_1692026134141_0010
2023-08-14 13:06:53,432 INFO InputFormat: Total input files in process : 4
2023-08-14 13:06:53,483 INFO mapreduce.JobSubmitter: Number of splits:8
2023-08-14 13:06:53,568 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1692026134141_0010
2023-08-14 13:06:53,568 INFO mapreduce.JobSubmitter: Executing with tokens: []
2023-08-14 13:06:53,696 INFO conf.Configuration: resource-types.xml not found
2023-08-14 13:06:53,696 INFO resource.ResourceUtil: Unable to find 'resource-types.xml'.
2023-08-14 13:06:53,762 INFO impl.YarnClientImpl: Submitted application application_1692026134141_0010
2023-08-14 13:06:53,797 INFO mapreduce.Job: The url to track the job: http://Kuncheria:8088/proxy/application_1692026134141_0010/
2023-08-14 13:06:55,591 INFO mapreduce.Job: Running job: job_1692026134141_0010
2023-08-14 13:06:59,931 INFO mapreduce.Job: map 0% reduce 0%
2023-08-14 13:07:00,049 INFO mapreduce.Job: map 13% reduce 0%
2023-08-14 13:07:00,049 INFO mapreduce.Job: map 25% reduce 0%
2023-08-14 13:07:10,081 INFO mapreduce.Job: map 38% reduce 0%
2023-08-14 13:07:12,092 INFO mapreduce.Job: map 75% reduce 0%
2023-08-14 13:07:13,107 INFO mapreduce.Job: map 88% reduce 0%
2023-08-14 13:07:14,116 INFO mapreduce.Job: map 100% reduce 0%
2023-08-14 13:07:17,155 INFO mapreduce.Job: map 100% reduce 100%
2023-08-14 13:07:21,938 INFO mapreduce.Job: Job job_1692026134141_0010 completed successfully
2023-08-14 13:07:40,241 INFO mapreduce.Job: Counters
  File System Counters
    FILE: Number of bytes read=51596340
    FILE: Number of bytes written=105578009
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=592448237
    HDFS: Number of bytes written=491
    HDFS: Number of read operations=29
    HDFS: Number of large read operations=0
    HDFS: Number of write operations=2
    HDFS: Number of bytes read erasure-coded=0
  Job Counters
    Killed map tasks=1
    Launched map tasks=8
    Launched reduce tasks=1
    Rack-local map tasks=8
    Total time spent by all maps in occupied slots (ms)=36957
    Total time spent by all reducers in occupied slots (ms)=5701
    Total time spent by all map tasks (ms)=36957
    Total time spent by all reduce tasks (ms)=5701
    Total vcore-milliseconds taken by all map tasks=36957
    Total vcore-milliseconds taken by all reduce tasks=5701
    Total megabyte-milliseconds taken by all map tasks=37843968
```

## OUTPUT:

```
C:\hadoop-3.3.0\sbin>hadoop fs -cat output13/part-r-00000
American Airlines 8.900856346719886
Alaska Airlines 1.7858007096736666
JetBlue Airways 11.5143526744102
Delta Air Lines 1.3692541768134422
ExpressJet Airlines 8.1593449776958
Frontier Airlines 13.358958345331709
Hawaiian Airlines 0.48571315905796407
Delta Airlines 10.125188203309524
Spirit Airlines 15.944765880783688
SkyWest Airlines 7.801103880415331
United Airlines 14.435441010805953
US Airways Inc. 6.141136917746696
Virgin America 9.022595096521952
Southwest Airlines 10.581986295158847
C:\hadoop-3.3.0\sbin>
```

## 2. The number of flights and distance travelled based on category of HAUL (flight distance)

```
C:\hadoop-3.3.0\bin>hadoop jar C:\hadoop\mapreduce\FlightDistanceAnalysis.jar hdfs://localhost:9000/input9/airlines output27
2023-08-14 16:15:29 942 INFO Client:DefaultYARNFollowerProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2023-08-14 16:15:31,500 INFO mapreduce.JobResourceUploader: Uploading local file to resource manager via local file system
2023-08-14 16:15:23,520 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/kunchi-staging/job_1692026134141_0025
2023-08-14 16:15:21,694 INFO input.FileInputFormat: Total input files to process : 1
2023-08-14 16:15:22,159 INFO mapreduce.JobSubmitter: number of splits:5
2023-08-14 16:15:22,247 INFO mapreduce.JobSubmitter: Submitting tasks for job: job_1692026134141_0025
2023-08-14 16:15:22,247 INFO mapreduce.JobSubmitter: Executing with tokens: []
2023-08-14 16:15:22,370 INFO conf.Configuration: resource-types.xml not found
2023-08-14 16:15:22,429 INFO resource.ResourceTills: Unable to find 'resource-types.xml'.
2023-08-14 16:15:22,429 INFO impl.YarnClientImpl: Submitted application application_1692026134141_0025
2023-08-14 16:15:22,429 INFO mapreduce.Job: The application application_1692026134141_0025 is running on cluster
2023-08-14 16:15:22,454 INFO mapreduce.Job: Running job: job_1692026134141_0025
2023-08-14 16:15:29,571 INFO mapreduce.Job: Job job_1692026134141_0025 running in uber mode : false
2023-08-14 16:15:29,571 INFO mapreduce.Job: map 0% reduce 0%
2023-08-14 16:15:37,678 INFO mapreduce.Job: map 20% reduce 0%
2023-08-14 16:15:38,693 INFO mapreduce.Job: map 40% reduce 0%
2023-08-14 16:15:39,704 INFO mapreduce.Job: map 60% reduce 0%
2023-08-14 16:15:40,717 INFO mapreduce.Job: map 100% reduce 0%
2023-08-14 16:15:46,789 INFO mapreduce.Job: map 100% reduce 100%
2023-08-14 16:15:46,789 INFO mapreduce.Job: Job job_1692026134141_0025 completed successfully
2023-08-14 16:15:46,803 INFO mapreduce.Job: Counters: 55
File System Counters
FILE: Number of bytes read=101050533
FILE: Number of bytes written=203691425
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=59242345
HDFS: Number of bytes written=31000
HDFS: Number of read operations=20
HDFS: Number of large read operations=0
HDFS: Number of write operations=2
HDFS: Number of bytes read crasure-coded=0
Job Counters
Killed map tasks=1
Launched map tasks=5
Launched reduce tasks=1
Rack-local map tasks=3
Total time spent by all maps in occupied slots (ms)=25664
Total time spent by all reducers in occupied slots (ms)=5827
Total time spent by all map tasks (ms)=25664
Total time spent by all reduce tasks (ms)=5827
Total vcore-milliseconds taken by all map tasks=25664
Total vcore-milliseconds taken by all reduce tasks=5827
Total megabyte-milliseconds taken by all map tasks=2627936
Total megabyte-milliseconds taken by all reduce tasks=5966848
Map-Reduce Framework
2023-08-14 16:15:46,803 INFO mapreduce.Job: Counters: 55
File System Counters
FILE: Number of bytes read=101050533
FILE: Number of bytes written=203691425
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=59242345
HDFS: Number of bytes written=31000
HDFS: Number of read operations=20
HDFS: Number of large read operations=0
HDFS: Number of write operations=2
HDFS: Number of bytes read crasure-coded=0
Job Counters
Killed map tasks=1
Launched map tasks=5
Launched reduce tasks=1
Rack-local map tasks=3
Total time spent by all maps in occupied slots (ms)=25664
Total time spent by all reducers in occupied slots (ms)=5827
Total time spent by all map tasks (ms)=25664
Total time spent by all reduce tasks (ms)=5827
Total vcore-milliseconds taken by all map tasks=25664
Total vcore-milliseconds taken by all reduce tasks=5827
Total megabyte-milliseconds taken by all map tasks=2627936
Total megabyte-milliseconds taken by all reduce tasks=5966848
Map-Reduce Framework
2023-08-14 16:15:46,803 INFO mapreduce.Job: Counters: 55
File System Counters
FILE: Number of bytes read=101050533
FILE: Number of bytes written=203691425
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=59242345
HDFS: Number of bytes written=31000
HDFS: Number of read operations=20
HDFS: Number of large read operations=0
HDFS: Number of write operations=2
HDFS: Number of bytes read crasure-coded=0
Job Counters
Killed map tasks=1
Launched map tasks=5
Launched reduce tasks=1
Rack-local map tasks=3
Total time spent by all maps in occupied slots (ms)=25664
Total time spent by all reducers in occupied slots (ms)=5827
Total time spent by all map tasks (ms)=25664
Total time spent by all reduce tasks (ms)=5827
Total vcore-milliseconds taken by all map tasks=25664
Total vcore-milliseconds taken by all reduce tasks=5827
Total megabyte-milliseconds taken by all map tasks=2627936
Total megabyte-milliseconds taken by all reduce tasks=5966848
Map-Reduce Framework
```

## OUTPUT:

```
C:\hadoop-3.3.0\bin>Administrator: Command Prompt
C:\hadoop-3.3.0\sbin>hadoop fs -cat output27/part-r-00000
Long Haul   Total Flights: 786370, Total Distance: 1617872089, Max Distance: 4983, Avg Distance: 2857.44
Medium Haul Total Flights: 291254, Total Distance: 2524130631, Max Distance: 1590, Avg Distance: 865.64
Short Haul  Total Flights: 2126155, Total Distance: 043342489, Max Distance: 497, Avg Distance: 383.44
C:\hadoop-3.3.0\sbin>
```

### 3. Average delay on different days of a week

```
C:\> hadoop-3.3.0\bin\hadoop jar C:\hadoop-3.3.0\share\hadoop\mapreduce\DayofWeekAnalysis.jar hdfs://localhost:9000/input9/airlines output34
2023-08-15 22:23:14,184 INFO client.DefaultNWARNFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2023-08-15 22:23:14,762 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
2023-08-15 22:23:14,783 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/kunchi-staging/job_1692150679872_0004
2023-08-15 22:23:14,952 INFO input.FileInputFormat: Total input files to process : 4
2023-08-15 22:23:14,999 INFO mapreduce.JobSubmitter: number of splits:4
2023-08-15 22:23:15,001 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1692150679872_0004
2023-08-15 22:23:15,089 INFO mapreduce.JobSubmitter: Executing with tokens: []
2023-08-15 22:23:15,212 INFO configuration: resource-types.xml not found
2023-08-15 22:23:15,213 INFO resource.Resourcetools: Unable to find 'resource-types.xml'.
2023-08-15 22:23:15,270 INFO impl.YarnClientImpl: Submitted application application_1692150679872_0004
2023-08-15 22:23:15,299 INFO mapreduce.Job: The url to track the job: http://Kuncheria:8088/proxy/application_1692150679872_0004
2023-08-15 22:23:15,306 INFO mapreduce.Job: Running job: job_1692150679872_0004
2023-08-15 22:23:22,442 INFO mapreduce.Job: Job job_1692150679872_0004 running in uber mode : false
2023-08-15 22:23:22,451 INFO mapreduce.Job: map 0% reduce 0%
2023-08-15 22:23:22,452 INFO mapreduce.Job: map 1% reduce 0%
2023-08-15 22:23:22,630 INFO mapreduce.Job: map 10% reduce 0%
2023-08-15 22:23:22,631 INFO mapreduce.Job: map 63% reduce 0%
2023-08-15 22:23:34,657 INFO mapreduce.Job: map 75% reduce 0%
2023-08-15 22:23:36,675 INFO mapreduce.Job: map 88% reduce 0%
2023-08-15 22:23:37,681 INFO mapreduce.Job: map 100% reduce 0%
2023-08-15 22:23:37,723 INFO mapreduce.Job: map 100% reduce 100%
2023-08-15 22:23:41,768 INFO mapreduce.Job: Job job_1692150679872_0004 completed successfully
2023-08-15 22:23:41,839 INFO mapreduce.Job: Counters:
File System Counters
FILE: Number of bytes read=5730926
FILE: Number of bytes written=116985981
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=592448237
HDFS: Number of bytes written=278
HDFS: Number of read operations=29
HDFS: Number of large read operations=0
HDFS: Number of write operations=2
HDFS: Number of bytes read erasure-coded=0
Job Counters
Killed map tasks=1
Launched map tasks=8
Launched reduce tasks=1
Rack-local map tasks=8
Total time spent by all maps in occupied slots (ms)=43275
Total time spent by all reduces in occupied slots (ms)=5674
Total time spent by all map tasks (ms)=43275
Total time spent by all reduce tasks (ms)=5674
Total vcore-milliseconds taken by all map tasks=43275
Total vcore-milliseconds taken by all reduce tasks=5674
Total megabyte-milliseconds taken by all map tasks=44313600
22:24 15-08-2023
```

#### OUTPUT:

```
C:\> hadoop-3.3.0\bin\hadoop fs -cat output34/part-r-00000
Day of Week: 1 Average Delay: 16.178976
Day of Week: 2 Average Delay: 16.07923
Day of Week: 3 Average Delay: 16.232548
Day of Week: 4 Average Delay: 16.555765
Day of Week: 5 Average Delay: 16.721626
Day of Week: 6 Average Delay: 15.298064
Day of Week: 7 Average Delay: 15.77076
C:\> hadoop-3.3.0\bin\
```

# SPARK JOBS

Some spark jobs were run on GCP DATAPROC to obtain meaningful insights. Mentioned below are some of them:

## 1. Average departure and arrival delays for each airline based on year and month.

Job details    CLONE    DELETE    STOP    REFRESH

Type: Dataproc Job    Status: Succeeded

Output: LINE WRAP: OFF

Spark jobs take ~60 seconds to initialize resources. DISMISS

```
|year|month|airline|avg_departure_delay| avg_arrival_delay|
+---+---+---+---+---+
|2015| 1 | AA | 10.593542260208928 | 6.955843432232982 |
|2015| 1 | AS | 3.1782088195181086 | -0.3208881453881834 |
|2015| 1 | B6 | 10.035555988505187 | 7.347280539099862 |
|2015| 1 | DL | 5.984238298406325 | -2.0438467953125195 |
|2015| 1 | EV | 9.752522427331304 | 8.537496880459196 |
|2015| 1 | F9 | 17.98443291326909 | 18.357238307349665 |
|2015| 1 | HA | 1.210654409473356 | 3.5126404494382024 |
|2015| 1 | MQ | 16.081267083483918 | 18.164973882762624 |
|2015| 1 | NK | 13.146293512200764 | 11.39805375347544 |
|2015| 1 | OO | 12.155156988868642 | 10.889893962046941 |
|2015| 1 | UA | 14.01039374165644 | 6.35272035280893 |
|2015| 1 | US | 5.197315865126567 | 3.10745735548587 |
|2015| 1 | VX | 6.910771877015693 | 1.4207015278674413 |
|2015| 1 | WN | 9.514469976705627 | 3.3984656332857434 |
|2015| 2 | AA | 11.49939442874449 | 8.938527176134042 |
|2015| 2 | AS | 7.20868415851313 | 5.549828178694158 |
|2015| 2 | B6 | 25.272304832713754 | 24.42070275403609 |
|2015| 2 | DL | 28.94596670934699 | 15.992504523132592 |
|2015| 2 | EV | 16.220496894409937 | 15.669058295964126 |
|2015| 2 | F9 | 45.60138248847926 | 48.88425925925926 |
+---+---+---+---+---+
only showing top 20 rows
```

DataFrame[year: int, month: int, airline: string, avg\_departure\_delay: double, avg\_arrival\_delay: double]

Search    225 PM    14-Aug-23

## 2. Flights with highest arrival delay

Job details    CLONE    DELETE    STOP    REFRESH

Job ID: job-240f66cc    Job UUID: e79772d0-3c97-4675-ac3d-b7fc2bfe7b3d

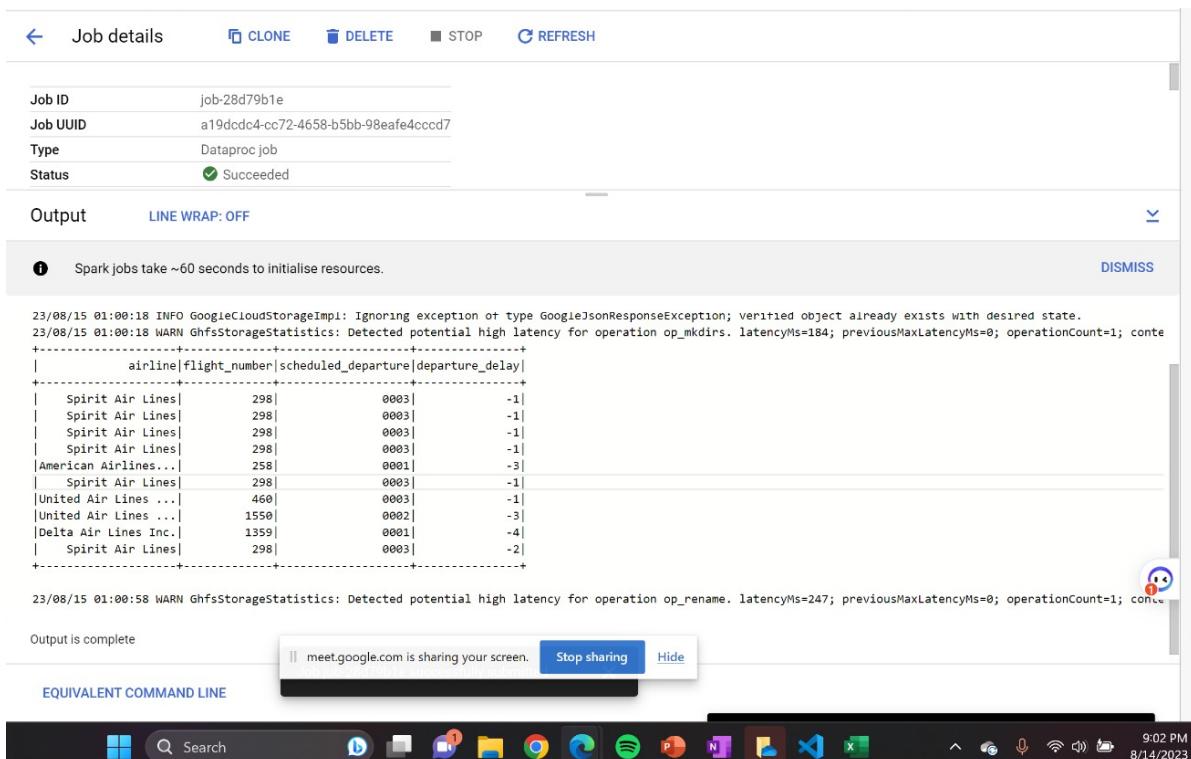
Type: Dataproc Job    Status: Succeeded

Output: LINE WRAP: OFF

Spark jobs take ~60 seconds to initialize resources. DISMISS

```
|airline|flight_number|arrival_delay|departure_delay|
+---+---+---+---+
|Frontier Airlines...| 1274 | 324 | -6 |
|Delta Air Lines Inc.| 1156 | 237 | 0 |
|American Airlines...| 1307 | 227 | -4 |
|American Airlines...| 291 | 225 | -5 |
|American Airlines...| 126 | 226 | -2 |
|Spirit Air Lines| 718 | 243 | 16 |
|American Airlines...| 1283 | 216 | -4 |
|American Airlines...| 2346 | 438 | 219 |
|Atlantic Southeas...| 4169 | 206 | -2 |
|American Airlines...| 1120 | 353 | 145 |
|United Air Lines ...| 1106 | 388 | 183 |
|American Airlines...| 2335 | 339 | 136 |
|American Airlines...| 125 | 337 | 135 |
|Skywest Airlines ...| 4739 | 195 | -4 |
|Atlantic Southeas...| 6017 | 403 | 205 |
|American Airlines...| 643 | 188 | -9 |
|American Eagle Ai...| 3418 | 191 | -4 |
|Delta Air Lines Inc.| 455 | 191 | -3 |
|JetBlue Airways| 350 | 188 | -6 |
|JetBlue Airways| 643 | 183 | -8 |
+---+---+---+---+
```

### 3. Flights that departed earlier than their scheduled departure time



The screenshot shows the 'Job details' page for a completed Dataproc job. The job ID is job-28d79b1e and the UUID is a19dcdc4-cc72-4658-b5bb-98eafe4cccd7. The status is Succeeded. The output log displays a table of flight information:

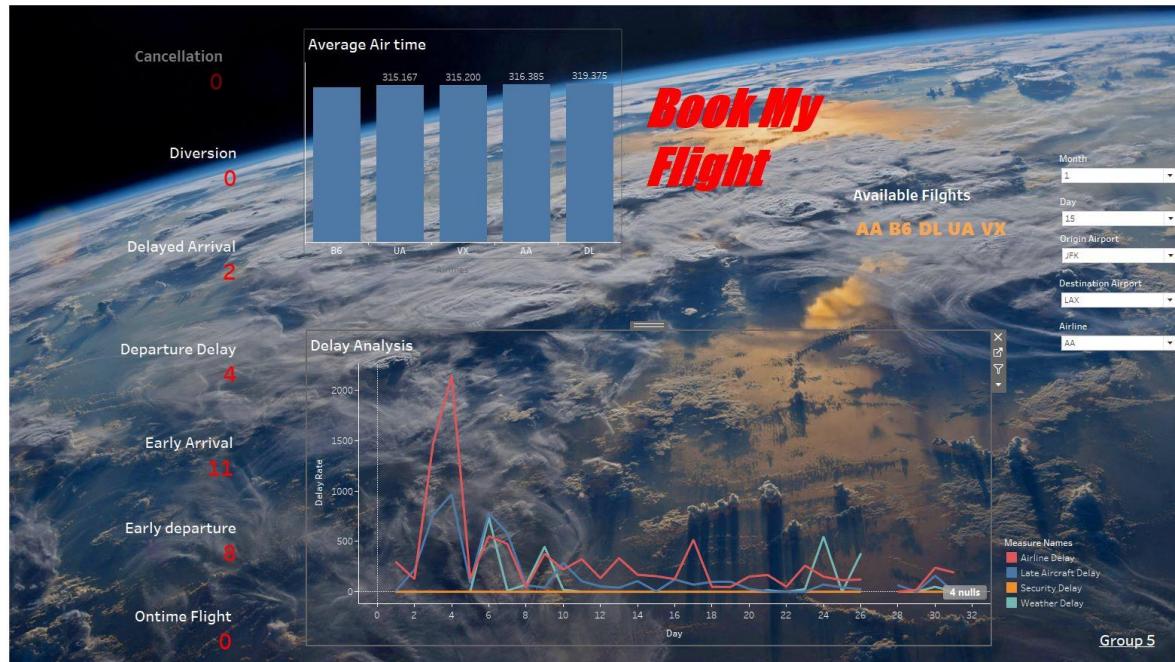
airline	flight_number	scheduled_departure	departure_delay
Spirit Air Lines	298	0003	-1
Spirit Air Lines	298	0003	-1
Spirit Air Lines	298	0003	-1
Spirit Air Lines	298	0003	-1
American Airlines...	258	0001	-3
Spirit Air Lines	298	0003	-1
United Air Lines ...	460	0003	-1
United Air Lines ...	1550	0002	-3
Delta Air Lines Inc.	1359	0001	-4
Spirit Air Lines	298	0003	-2

The log also includes several informational and warning messages from the Google Cloud Storage and HDFS storage systems.

# VISUALIZATION

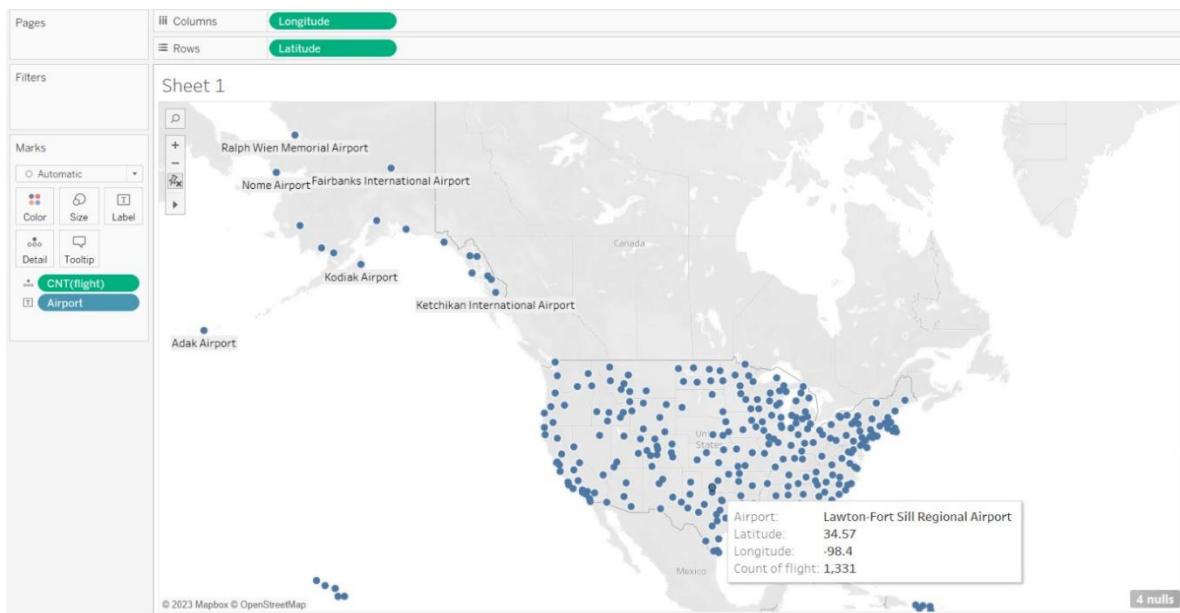
Visualization was done using tableau. A dynamic dashboard for airline and flight analysis is done. In which the users will get to know real-time information about any flight airtime, delays, and cancellations due to specific reasons.

## REAL-TIME DASHBOARD FOR FLIGHT BASED ON MONTH AND DAY.



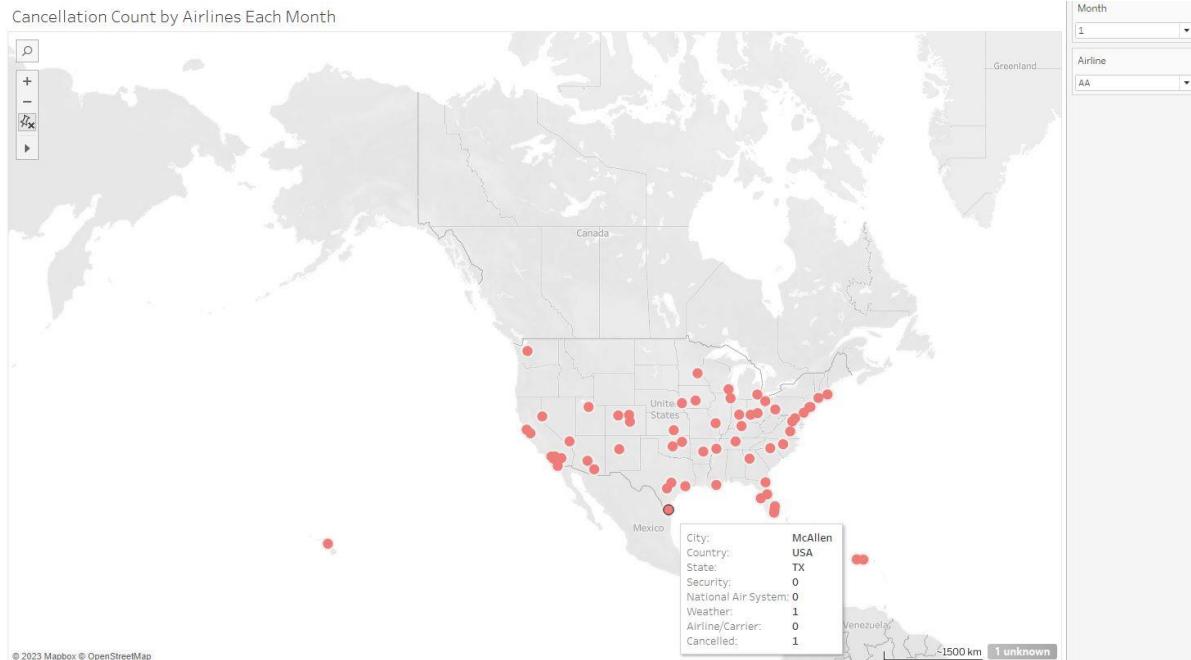
The dashboard is dynamic and provides results regarding a flight delay due to airline-delay, late aircraft delay, security delay, weather delay. Average airtime for each aircraft based on the user mentioned airport is obtained. The dashboard also provides better understanding of how many cancellations was happened to the airline, diversion taken, delayed arrival count, departure delay count, early arrival, early departure and on time flight is gained. Based on the inputs provided by the user, the dashboard updates on the available flights from the origin airport.

## 1. Total number of flights from airports



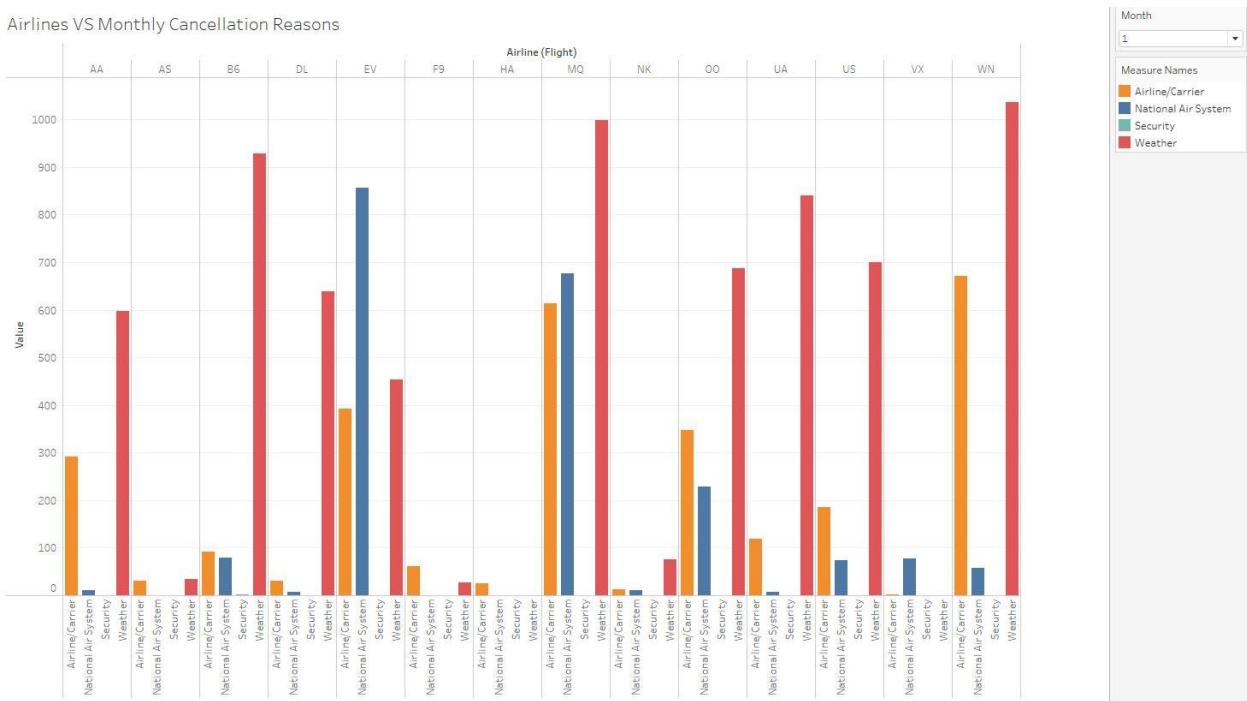
This provides understanding about the number of flights flown from the origin airports.

## 2. Cancellation count by airlines for each month



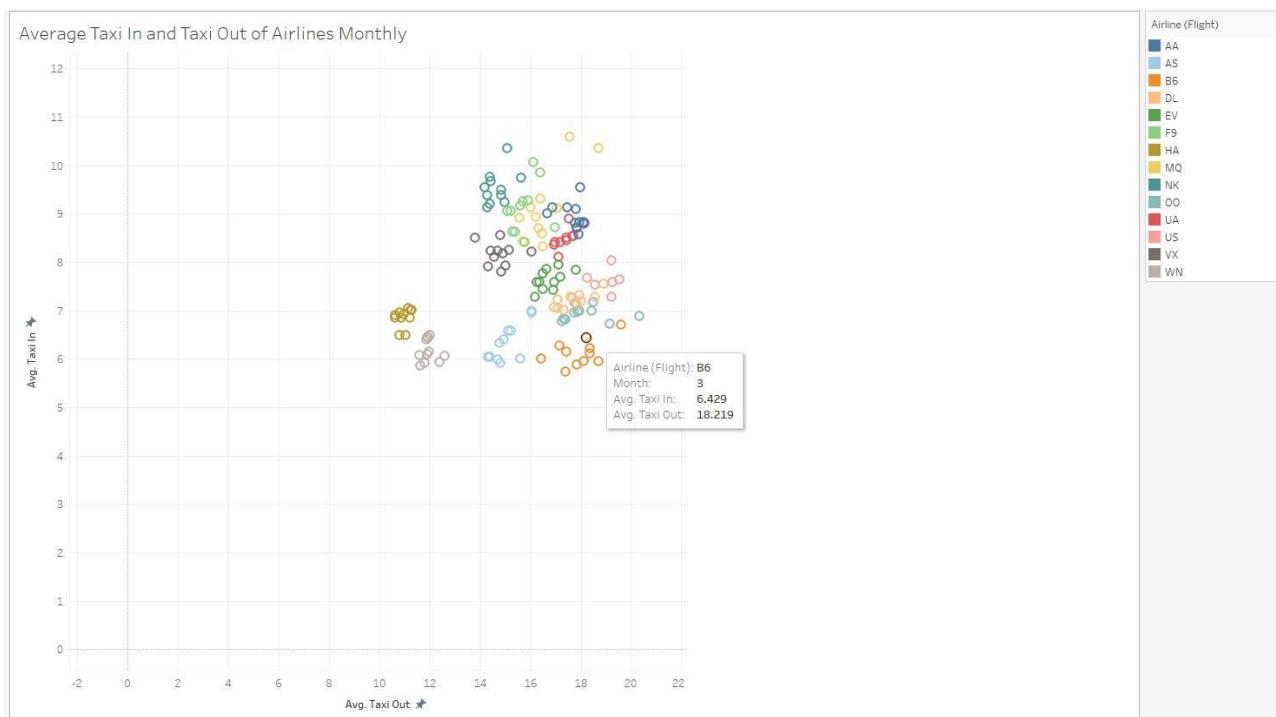
It depicts the count of airlines which was cancelled due to some specific reasons on a specific month.

### 3. Based on monthly updates on cancellation reasons for airline



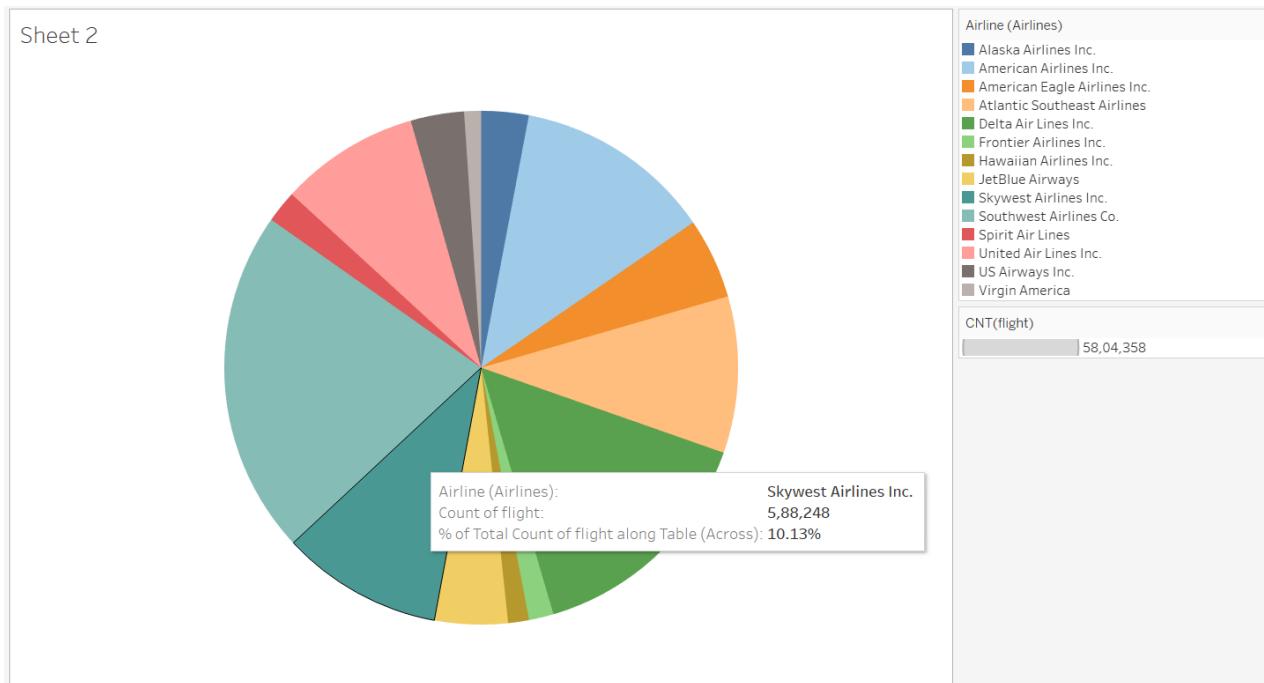
This chart helps to understand the cancellation rate for airlines(flight) which were cancelled.

### 4. Monthly average Taxi-in and Taxi-out for airlines



The chart shows the average rate of taxi-in and taxi-out for the mentioned airlines during the specific months.

## 5. Market share of each airline



## GITHUB REPOSITORY LINK

<https://github.com/Jerin-T/airline-analysis>

## **CONCLUSION**

On the analysis of the dataset, we could find that delay, cancellations, and diversion are the main challenges and cause serious resource wastages that leads to performance issue in the domestic airlines industry.

We could find a common pattern by airports, airlines and season that are usually delayed. Weather, airline issues and air traffic in each entity are the primary cause of this.

By addressing these issues by identifying the patterns from historic data we could minimize the root cause of the issues and enhance the industry performance up to a great extent.

## **REFERENCES**

1. <https://www.bts.gov/newsroom/2015-annual-and-december-us-airline-traffic-data>
2. <https://www.bts.gov/newsroom/2015-us-based-airline-traffic-data>