BDM-1024: DATA TECHNOLOGY SOLUTIONS

PROJECT REPORT

SUBMITTED TO

BHAVIK GANDHI

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

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

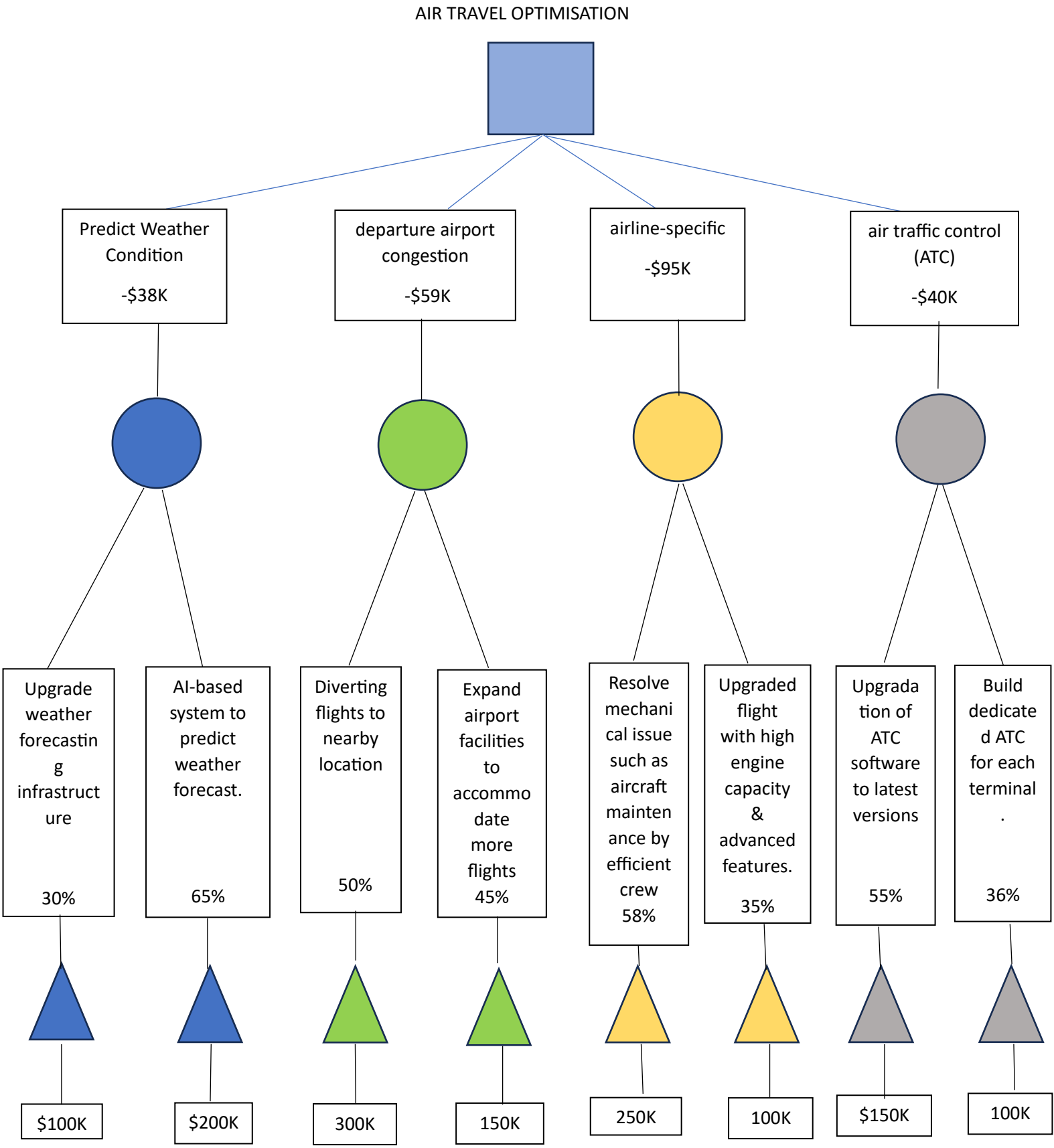
3.Flight

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.

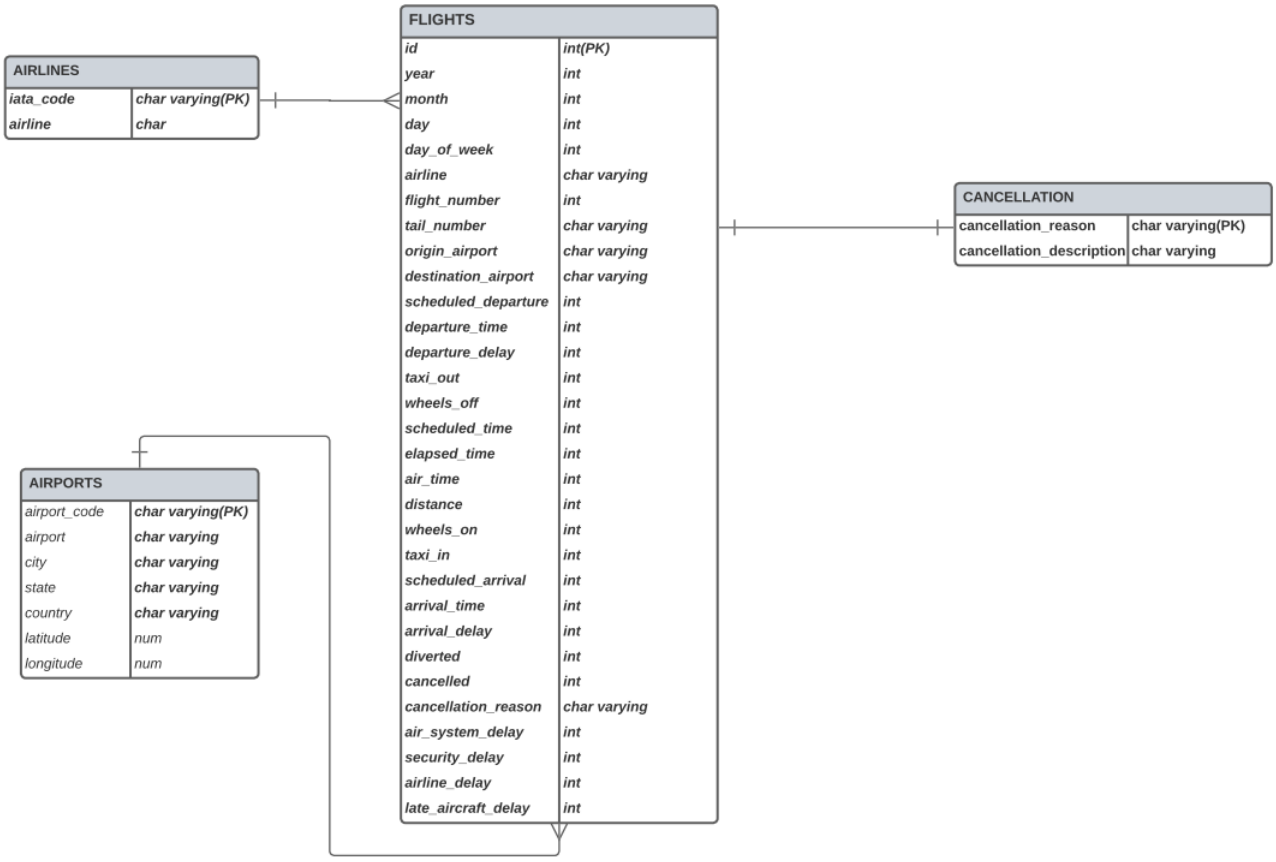
DECISION TREE



EXPECTED VALUE

100×0.3+200×0.65−38 = **$112K**	300×0.50+150×0.45−59 =$**158.5K**	250×0.58+100×0.35−95 =**$85K**	150x0.5+100x0.36−40= **$78.5K**

DATA MODEL DIAGRAM



SQL CODES & INSIGHTS

1. TIME VARIATION OF ALL FLIGHTS AND STATUS OF ARRIVAL

```sql
--TIME VARIATION OF ALL FLIGHTS AND STATUS OF ARRIVAL
select
ap.city as FROM_CITY,
ap.airport as from_airport,
fl.tail_number,
aps.city as destination_city,
aps.airport as destination_airport,
aline.airline,
fl.day_of_week,
(fl.departure_delay+fl.arrival_delay) TIME_VARIATION,
fl.arrival_delay,
fl.departure_delay,
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'
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
inner join airlines aline on aline.iata_code = fl.airline
order by day_of_week;
```

Data Output | Messages | Notifications

	from_city character varying	from_airport character varying	tail_number character varying	destination_city character varying	destination_airport character varying	airline character varying	day_of_week integer	time_variation integer	arrival_delay integer	departure_delay integer	satus_completed text
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	N87512	Denver	Denver International Airport	United Air Lines Inc.	1	-12	-12	0	EARLY ARRIVAL
8	Los Angeles	Los Angeles International Airport	N34131	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

```sql
-- LIST OF ALL FLIGHTS WITH CANCELLATION
select
ap.city as from_city,
ap.airport as from_airport,
fl.tail_number,
aps.city as destination_city,
aps.airport as destination_airport,
aline.airline as Airline,
fl.day_of_week as Day_OF_WEEK,
fl.month as MONTH,
c.cancellation_description AS CANCELLATION_DESCRIPTION
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
inner join airlines aline on aline.iata_code = fl.airline
inner join cancellation c on  c.CANCELLATION_REASON = fl.CANCELLATION_REASON
where
fl.CANCELLED = 1
order by fl.day_of_week,fl.month;
```

Data Output | Messages | Notifications

	from_city character varying	from_airport character varying	tail_number character varying	destination_city character varying	destination_airport character varying	airline character varying	day_of_week integer	month integer	cancell charac
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

```sql
--ANALYSIS OF DELAY ON DAILY BASIS

SELECT
     AP.AIRPORT,
     aline.airline as AIRLINE,
     fl.DAY_OF_WEEK AS day_of_week,
     fl.DAY AS day,
     fl.MONTH AS month,
     COUNT(*) AS num_delays,
     ( case when fl.day_of_Week <=5  then'Weekday' else 'Weekend' end)as weektype

FROM
  flight fl
  inner join  airports ap on ap.airport_code = fl.origin_airport
  inner join airlines aline on aline.iata_code = fl.airline
WHERE
  departure_delay > 0
GROUP BY
    AP.AIRPORT,
     aline.airline,
     fl.DAY_OF_WEEK,
     fl.DAY,
     fl.MONTH
ORDER BY
  num_delays DESC;
```

Data Output | Messages | Notifications

	airport character varying	airline character varying	day_of_week integer	day integer	month integer	num_delays bigint	weektype text
1	Hartsfield-Jackson Atlanta International Airpo...	Delta Air Lines Inc.	7	23	8	577	Weekend
2	Hartsfield-Jackson Atlanta International Airpo...	Delta Air Lines Inc.	3	30	12	574	Weekday
3	Hartsfield-Jackson Atlanta International Airpo...	Delta Air Lines Inc.	5	10	4	552	Weekday
4	Hartsfield-Jackson Atlanta International Airpo...	Delta Air Lines Inc.	2	24	2	488	Weekday
5	Hartsfield-Jackson Atlanta International Airpo...	Delta Air Lines Inc.	4	5	3	484	Weekday
6	Hartsfield-Jackson Atlanta International Airpo...	Delta Air Lines Inc.	2	18	8	472	Weekday

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

```sql
-- ANALYSIS by AIRLINE ON ARIVAL-DEPARTURE DELAY

select
aline.airline AIRLINE,
count(*) Number_of_Flights,
SUM(CASE WHEN fl.departure_delay> 0 THEN 1 ELSE 0 END) Number_Of_Delays_At_Departure,
SUM(CASE WHEN fl.arrival_delay > 0 THEN 1 ELSE 0 END) Number_Of_Delays_At_Arrival,
SUM(CASE WHEN fl.departure_delay< 0 THEN 1 ELSE 0 END) Number_Of_Early_Departure,
SUM(CASE WHEN fl.arrival_delay < 0 THEN 1 ELSE 0 END) Number_Of_Early_Arrival,
SUM(CASE WHEN fl.departure_delay= 0 and fl.arrival_delay = 0  THEN 1 ELSE 0 END) On_time_flights

from
 airlines aline
 inner join flight fl on aline.iata_code = fl.airline

group by aline.airline
```

Data Output | Messages | Notifications

	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
1	Alaska Airlines Inc.	172521	43566	56953	121253	110162	306
2	American Airlines Inc.	725984	245904	252191	436039	446655	916
3	American Eagle Airlines Inc.	294632	93726	103505	171512	170102	522
4	Atlantic Southeast Airlines	571977	169897	213217	365593	328301	924
5	Delta Air Lines Inc.	875881	282463	250840	522581	601859	1697
6	Frontier Airlines Inc.	90836	34893	41232	52144	46873	114
7	Hawaiian Airlines Inc.	76272	20146	30179	52841	42545	272
8	JetBlue Airways	267048	102061	101998	147160	155170	403
9	Skywest Airlines Inc.	588353	171572	222435	380370	339967	1094
10	Southwest Airlines Co.	1261855	566807	470767	587783	742332	3093
11	Spirit Air Lines	117379	52089	56887	58663	55893	154
12	United Air Lines Inc.	515723	256550	186227	224854	312159	644
13	US Airways Inc.	198715	62565	76285	123032	113399	314

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

```
--ANALYSIS BY AIRLINE BY DELAY CAUSE

select
aline.airline AIRLINE,
count(*) Number_of_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
 airlines aline
 inner join flight fl on aline.iata_code = fl.airline
group by aline.airline
```

Data Output | Messages | Notifications

	airline character varying	number_of_flights bigint	number_of_delay_by_air_system_delay bigint	number_of_delays_by_security_delay bigint	number_of_delays_by_airline_delay bigint	number_of_delays_by_late_aircraft_delay bigint	number_of_delays_by_weather_delay bigint
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 Air Lines	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

Total rows: 14 of 14 Query complete 00:00:01.616 Ln 89, Col 1

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 | Query History

```
-- ANALYSIS by AIRLINE ON CANCELLATION and RESAONS
select
aline.airline AIRLINE,
(SELECT Count(*) from Flight f where aline.iata_code = f.airline group by f.airline) Total_number_of_Flights,
SUM(CASE WHEN fl.CANCELLED = 1 THEN 1 else 0 END) total_Flights_Cancelled,
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
 airlines aline
 inner join flight fl on aline.iata_code = fl.airline
 inner join cancellation c on  c.CANCELLATION_REASON = fl.CANCELLATION_REASON
group by aline.airline,aline.iata_code;
```

Data Output | Messages | Notifications

	airline character varying	total_number_of_flights bigint	total_flights_cancelled bigint	cancelled_by_issue_airlineorcarrier bigint	cancelled_by_issue_weather bigint	cancelled_by_issue_national_air_system bigint	cancelled_by_issue_security bigint
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 Air Lines	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

Total rows: 14 of 14 Query complete 00:00:34.304 Ln 104, Col 1

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

```sql
--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 text	no_of_delayed_flights bigint	number_of_d bigint	number_of_d bigint	number_of_de bigint	number_of_delays_by_late_aircraft_delay bigint	number_of_delays_by_weather_delay bigint
1	Los Angeles International Airport to San Francisco International Airport	6004	1780	6	1367	1895	119
2	San Francisco International Airport to Los Angeles International Airport	5552	1889	8	1353	2111	88
3	Los Angeles International Airport to John F. Kennedy International Airport (New York International Airp...	4497	1026	9	988	718	80
4	Chicago O'Hare International Airport to Los Angeles International Airport	4415	985	10	1320	899	274
5	Chicago O'Hare International Airport to San Francisco International Airport	4282	937	2	1106	808	228
6	Los Angeles International Airport to McCarran International Airport	4058	955	4	1150	1501	54
7	Chicago O'Hare International Airport to LaGuardia Airport (Marine Air Terminal)	4048	1312	3	912	916	265
8	McCarran International Airport to Los Angeles International Airport	4002	1529	3	1189	1469	76
9	John F. Kennedy International Airport (New York International Airport) to Los Angeles International Air...	3943	887	20	1027	467	228
10	McCarran International Airport to San Francisco International Airport	3660	977	2	735	1081	127

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

```sql
--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 text	no_of_delayed_flights bigint	number_of_d bigint	number_of_d bigint	number_of_de bigint	number_of_delays_by_late_aircraft_delay bigint	number_of_delays_by_weather_delay bigint
1	Los Angeles International Airport to San Francisco International Airport	6004	1780	6	1367	1895	119
2	San Francisco International Airport to Los Angeles International Airport	5552	1889	8	1353	2111	88
3	Los Angeles International Airport to John F. Kennedy International Airport (New York International Airp...	4497	1026	9	988	718	80
4	Chicago O'Hare International Airport to Los Angeles International Airport	4415	985	10	1320	899	274
5	Chicago O'Hare International Airport to San Francisco International Airport	4282	937	2	1106	808	228
6	Los Angeles International Airport to McCarran International Airport	4058	955	4	1150	1501	54
7	Chicago O'Hare International Airport to LaGuardia Airport (Marine Air Terminal)	4048	1312	3	912	916	265
8	McCarran International Airport to Los Angeles International Airport	4002	1529	3	1189	1469	76
9	John F. Kennedy International Airport (New York International Airport) to Los Angeles International Air...	3943	887	20	1027	467	228
10	McCarran International Airport to San Francisco International Airport	3660	977	2	735	1081	127

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

```
--Top 10 diverted route

SELECT
ap.airport || ' to ' || aps.airport AS route,
COUNT(*) AS no_of_diverted_flights
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 diverted = 1
GROUP BY ap.airport, aps.airport
ORDER BY no_of_diverted_flights DESC
LIMIT 10;
```

Data Output | Messages | Notifications

	route text	no_of_diverted_flights bigint
1	Miami International Airport to LaGuardia Airport (Marine Air Terminal)	76
2	Salt Lake City International Airport to Friedman Memorial Airport	72
3	Chicago O'Hare International Airport to LaGuardia Airport (Marine Air Terminal)	66
4	Gen. Edward Lawrence Logan International Airport to Los Angeles International Airport	63
5	Gen. Edward Lawrence Logan International Airport to San Francisco International Airport	62
6	Los Angeles International Airport to Aspen-Pitkin County Airport	54
7	Cyril E. King Airport to Philadelphia International Airport	52
8	Hartsfield-Jackson Atlanta International Airport to LaGuardia Airport (Marine Air Termi...	48
9	Dallas/Fort Worth International Airport to LaGuardia Airport (Marine Air Terminal)	47
10	Chicago O'Hare International Airport to Dallas/Fort Worth International Airport	47

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 fl
inner join airports ap on ap.airport_code = fl.origin_airport
inner join airports aps on  aps.airport_code = fl.destination_airport
GROUP BY ap.airport, aps.airport
ORDER BY no_of_flights DESC
LIMIT 10
```

Data Output | Messages | Notifications

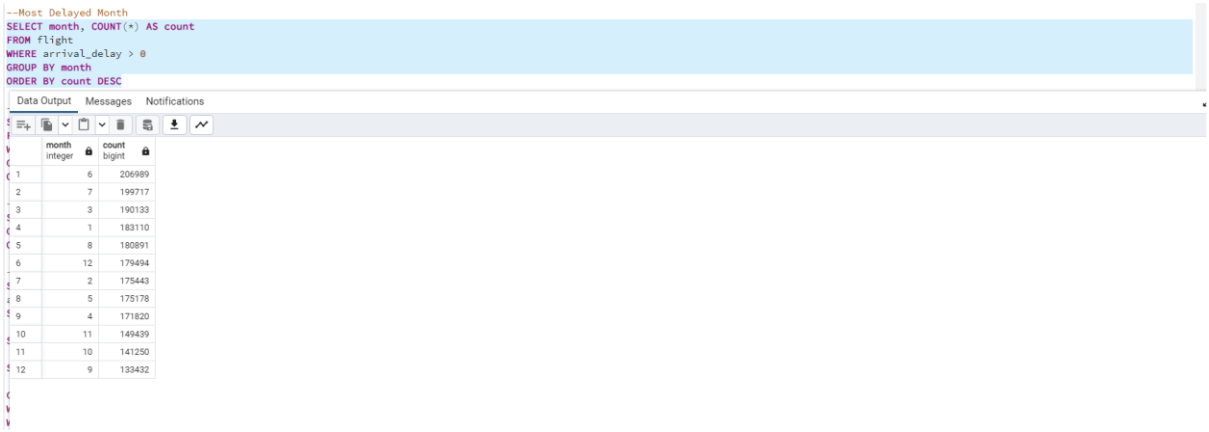
	route text	no_of_flights bigint
1	San Francisco International Airport to Los Angeles International Airport	13744
2	Los Angeles International Airport to San Francisco International Airport	13457
3	John F. Kennedy International Airport (New York International Airport) to Los Angeles International Airport	12016
4	Los Angeles International Airport to John F. Kennedy International Airport (New York International Airport)	12015
5	McCarran International Airport to Los Angeles International Airport	9715
6	LaGuardia Airport (Marine Air Terminal) to Chicago O'Hare International Airport	9639
7	Los Angeles International Airport to McCarran International Airport	9594
8	Chicago O'Hare International Airport to LaGuardia Airport (Marine Air Terminal)	9575
9	San Francisco International Airport to John F. Kennedy International Airport (New York International Airp...	8440
10	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 rote 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
```

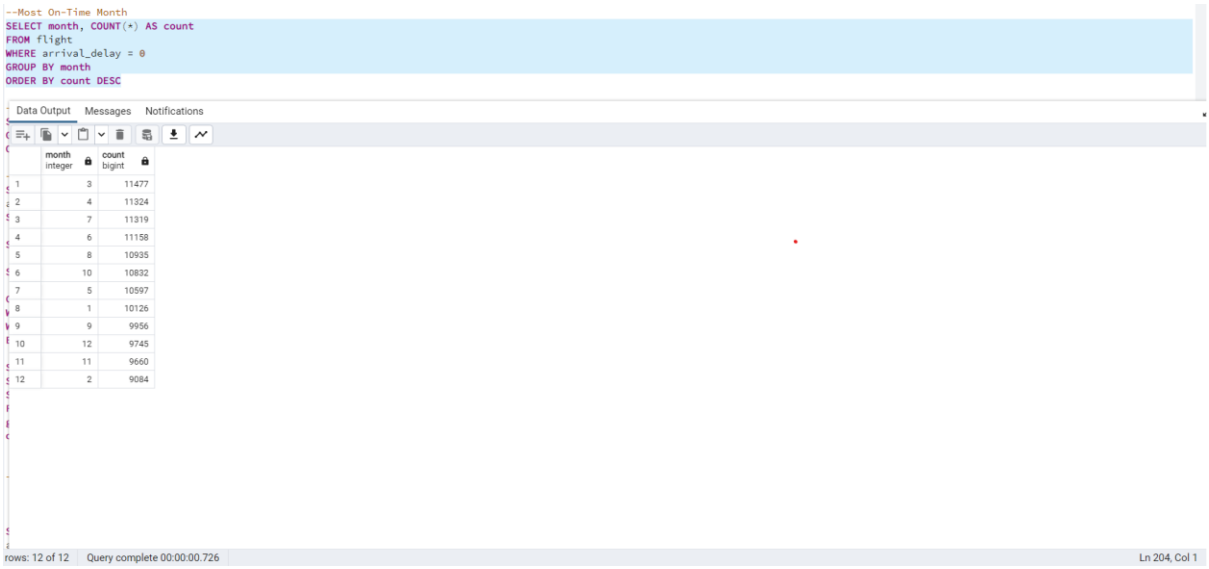
Data Output | Messages | Notifications

	month integer	count bigint
1	6	206989
2	7	199717
3	3	190133
4	1	183110
5	8	180891
6	12	179494
7	2	175443
8	5	175178
9	4	171820
10	11	149439
11	10	141250
12	9	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
```

Data Output | Messages | Notifications

	month integer	count bigint
1	3	11477
2	4	11324
3	7	11319
4	6	11158
5	8	10935
6	10	10832
7	5	10597
8	1	10126
9	9	9956
10	12	9745
11	11	9660
12	2	9084

rows: 12 of 12 | Query complete 00:00:00.726 | Ln 204, Col 1

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 TRAVELLED DAYS IN A WEEK

```
-- Most Travelled Days in a Week in US
SELECT day_of_week, Sum(distance) as Total_distance FROM flight
GROUP BY day_of_week
ORDER BY Total_distance DESC;
```

Data Output | Messages | Notifications

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

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 the significantly low in number.

15. TOP 10 DELAYED ROUTES BY AIRCRAFT ISSUES

```
SELECT
ap.airport || ' to ' || aps.airport AS route,
al.airline,
COUNT(*) AS no_of_delayed_flights,
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
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
inner join airlines al on  al.iata_code = fl.airline
WHERE departure_delay > 0 and fl.AIRLINE_DELAY >0 and fl.AIRLINE_DELAY >0
GROUP BY ap.airport, aps.airport,al.airline
ORDER BY no_of_delayed_flights DESC
LIMIT 10
```

Data Output | Messages | Notifications

	route text	airline character varying	no_of_delayed_flights bigint	number_of_delays_by_airline_delay bigint	number_of_delays_by_late_aircraft_delay bigint
1	Kahului Airport to Honolulu International Airport	Hawaiian Airlines Inc.	886	886	806
2	Dallas Love Field to William P. Hobby Airport	Southwest Airlines Co.	869	869	545
3	George Bush Intercontinental Airport to Los Angeles International Airport	United Air Lines Inc.	820	820	399
4	William P. Hobby Airport to Dallas Love Field	Southwest Airlines Co.	804	804	502
5	Dallas/Fort Worth International Airport to Los Angeles International Airp...	American Airlines Inc.	773	773	320
6	Chicago O'Hare International Airport to San Francisco International Airp...	United Air Lines Inc.	749	749	366
7	Chicago O'Hare International Airport to Los Angeles International Airport	United Air Lines Inc.	723	723	362
8	Newark Liberty International Airport to San Francisco International Airport	United Air Lines Inc.	721	721	350
9	McCarran International Airport to Los Angeles International Airport	Southwest Airlines Co.	714	714	484
10	San Francisco International Airport to Chicago O'Hare International Airp...	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

```
-- Actual number of flights flawn by each airlines on each month
select
aline.airline AIRLINE,
count(*) Number_of_Flights,
SUM(CASE WHEN fl.cancelled = 1 THEN 1 ELSE 0 END) no_of_cancelled,
(count(*) - SUM(CASE WHEN fl.cancelled = 1 THEN 1 ELSE 0 END)) ACTUAL_FLIGHTS
from
 airlines aline
 inner join flight fl on aline.iata_code = fl.airline
group by aline.airline;
```

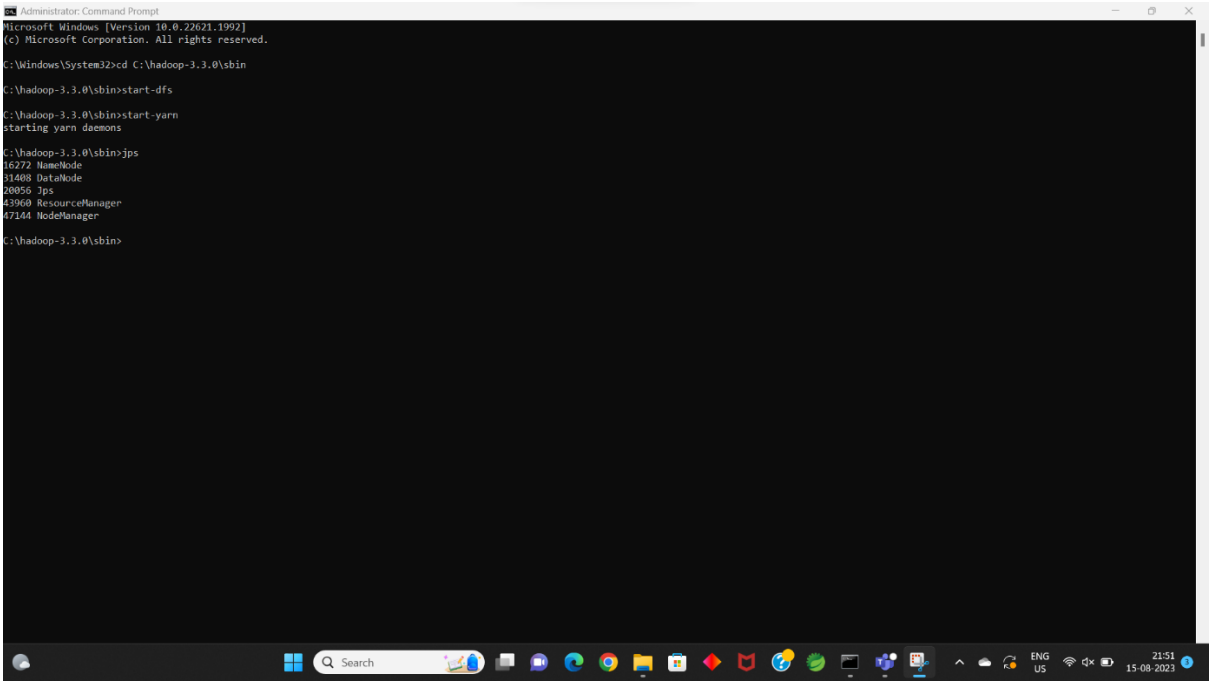
Data Output | Messages | Notifications

	airline character varying	number_of_flights bigint	no_of_cancelled bigint	actual_flights bigint
1	Alaska Airlines Inc.	172521	669	171852
2	American Airlines Inc.	725984	10919	715065
3	American Eagle Airlines Inc.	294632	15025	279607
4	Atlantic Southeast Airlines	571977	15231	556746
5	Delta Air Lines Inc.	875881	3824	872057
6	Frontier Airlines Inc.	90836	588	90248
7	Hawaiian Airlines Inc.	76272	171	76101
8	JetBlue Airways	267048	4276	262772
9	Skywest Airlines Inc.	588353	9960	578393
10	Southwest Airlines Co.	1261855	16043	1245812
11	Spirit Air Lines	117379	2004	115375
12	United Air Lines Inc.	515723	6573	509150
13	US Airways Inc.	198715	4067	194648
14	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.

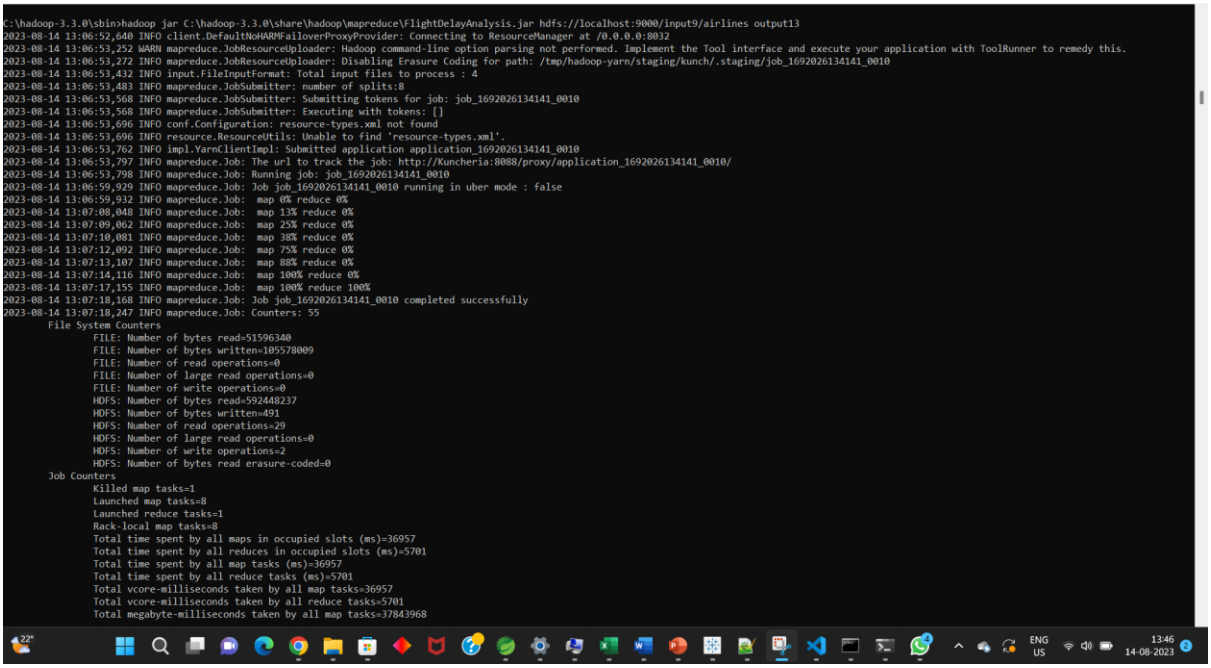
MAPREDUCE JOBS

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



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



OUTPUT:

```
C:\hadoop-3.3.0\sbin>hadoop fs -cat output13/part-r-00000
American Airlines       8.900856346719806
Alaska Airlines 1.7858007096736666
JetBlue Airways 11.5143526744102
Delta Airlines  7.36925417661782
ExpressJet Airlines     8.71593449776958
Frontier Airlines       13.350858345331709
Hawaiian Airlines       0.4857131596570407
Mesa Airlines   10.125188203309524
Spirit Airlines 15.944765880783688
SkyWest Airlines        7.801103808415331
United Airlines 14.435441010805953
US Airways Inc. 6.141136917746696
Virgin America  9.022595096521952
Southwest Airlines      10.581986295158847

C:\hadoop-3.3.0\sbin>
```

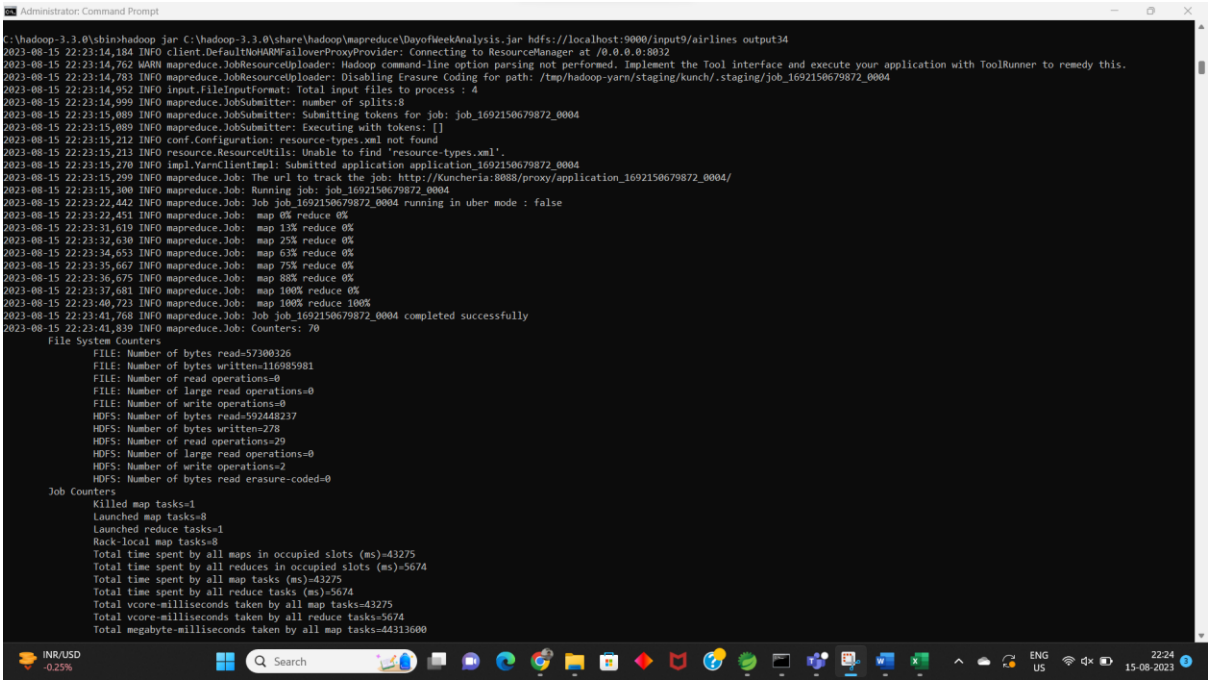
2. The number of flights and distance travelled based on category of HAUL (flight distance)

```
C:\hadoop-3.3.0\sbin>hadoop jar C:\hadoop-3.3.0\share\hadoop\mapreduce\FlightDistanceAnalysis.jar hdfs://localhost:9000/input9/airlines output27
2023-08-14 16:15:20,942 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2023-08-14 16:15:21,509 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 16:15:21,520 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/kunch/.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 tokens 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 16:15:22,370 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2023-08-14 16:15:22,425 INFO impl.YarnClientImpl: Submitted application application_1692026134141_0025
2023-08-14 16:15:22,453 INFO mapreduce.Job: The url to track the job: http://Kuncheria:8088/proxy/application_1692026134141_0025/
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 16:15:29,573 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 16:15:46,794 INFO mapreduce.Job: Job job_1692026134141_0025 completed successfully
2023-08-14 16:15:46,861 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=592423545
                HDFS: Number of bytes written=310
                HDFS: Number of read operations=20
                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=5
                Launched reduce tasks=1
                Rack-local map tasks=5
                Total time spent by all maps in occupied slots (ms)=25664
                Total time spent by all reduces 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=26279936
                Total megabyte-milliseconds taken by all reduce tasks=5966848
        Map-Reduce Framework
```

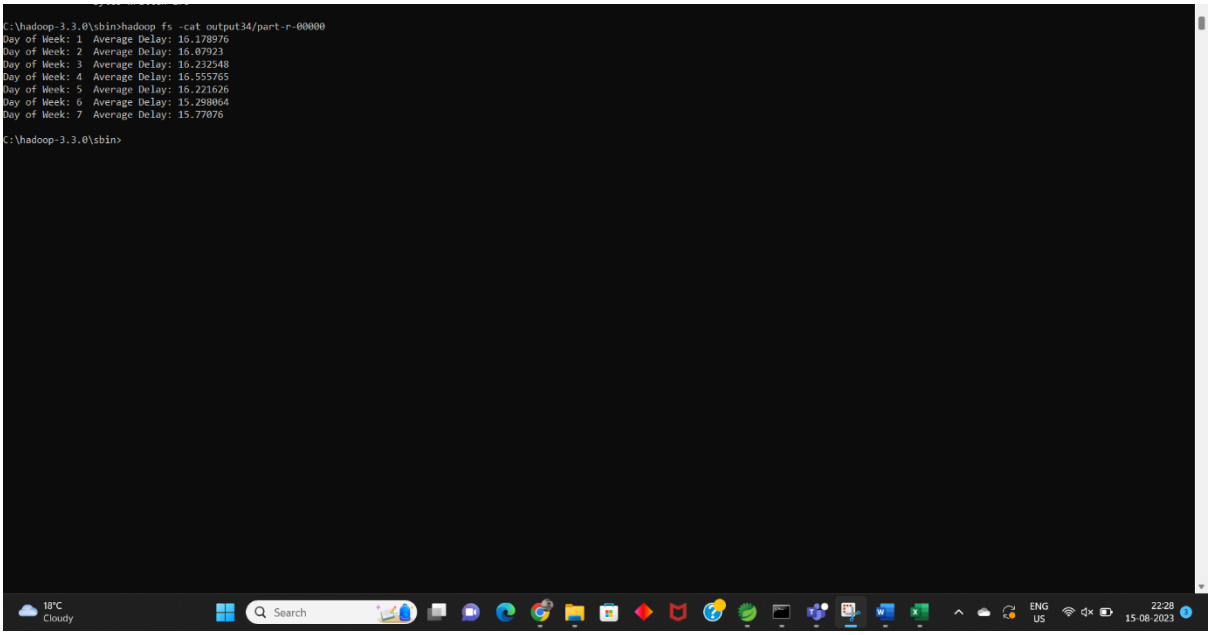
OUTPUT:

```
Administrator: Command Prompt

C:\hadoop-3.3.0\sbin>hadoop fs -cat output27/part-r-00000
Long Haul       Total Flights: 786370, Total Distance: 1617874289, Max Distance: 4983, Avg Distance: 2057.40
Medium Haul     Total Flights: 2912554, Total Distance: 2524139631, Max Distance: 1500, Avg Distance: 866.64
Short Haul      Total Flights: 2120155, Total Distance: 643343489, Max Distance: 497, Avg Distance: 303.44

C:\hadoop-3.3.0\sbin>
```

3. Average delay on different days of a week



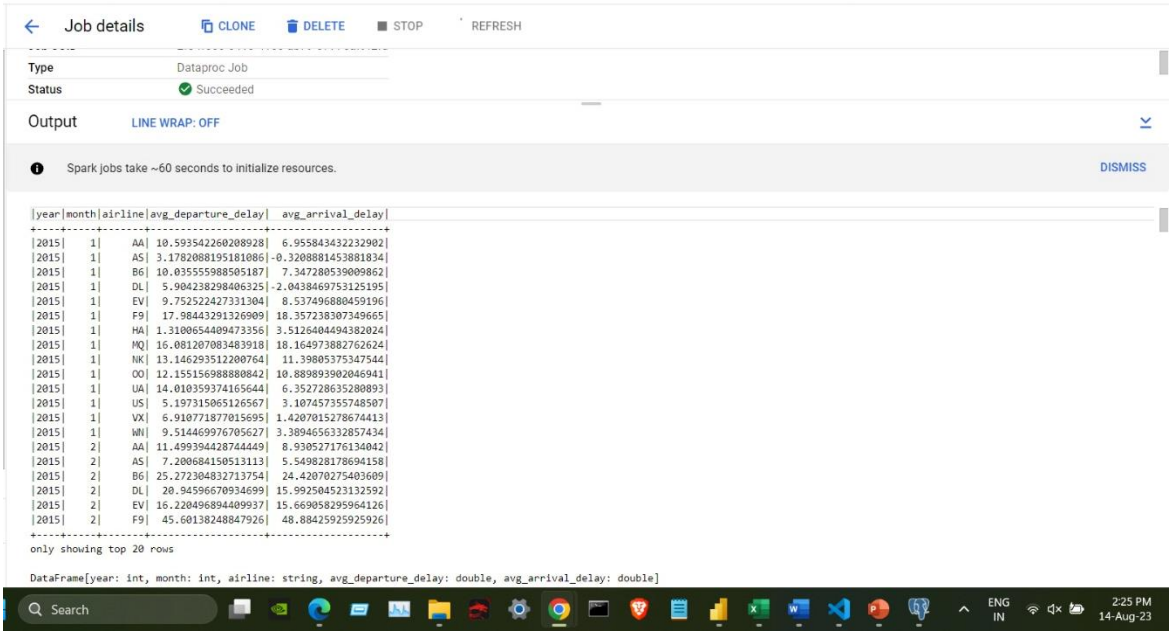
OUTPUT:



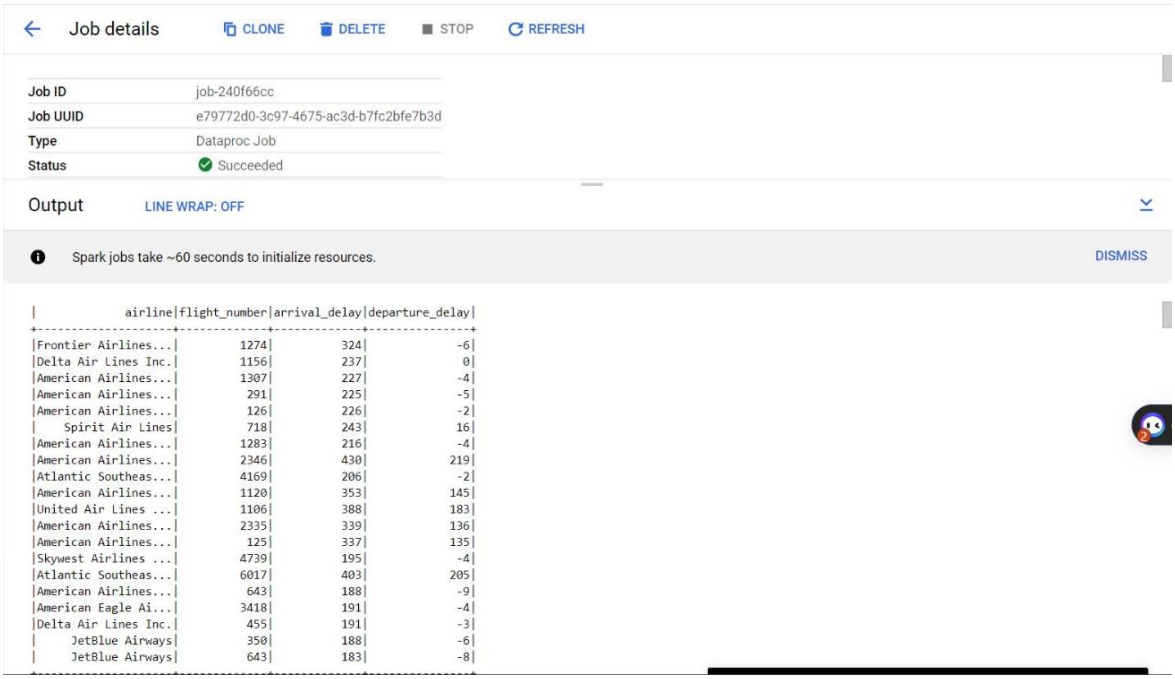
SPARK JOBS

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

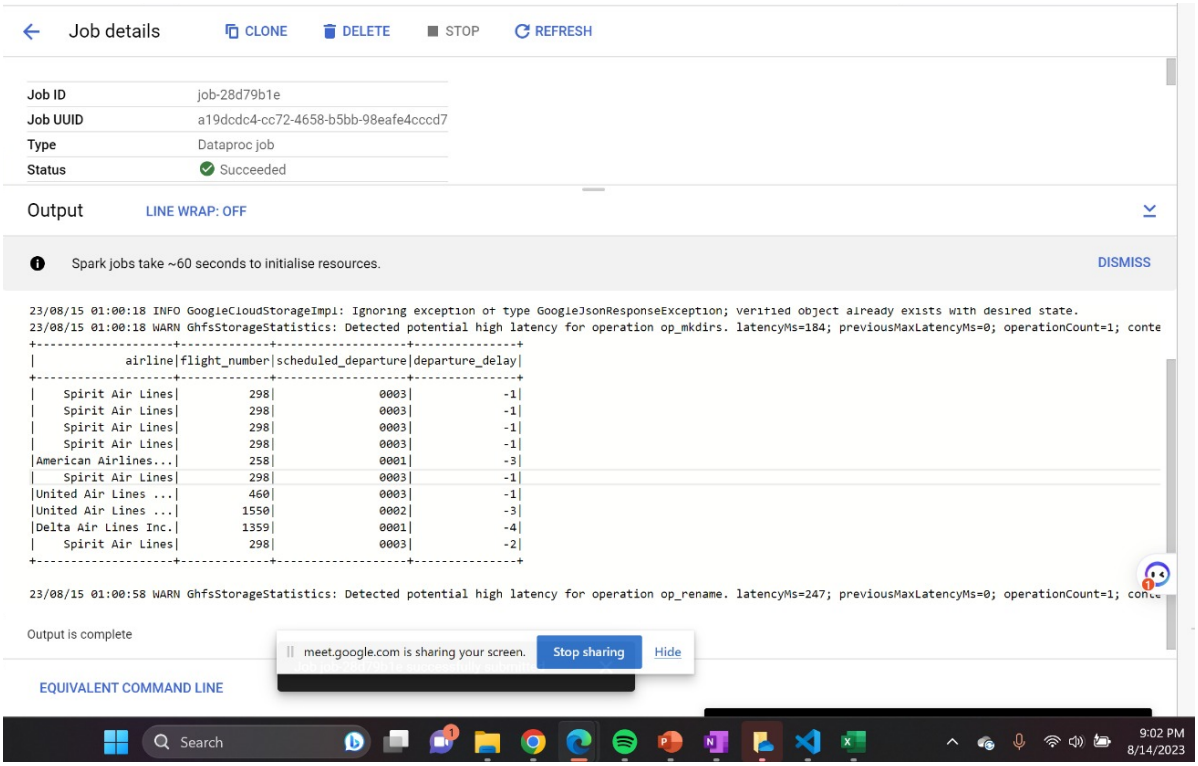
1. Average departure and arrival delays for each airline based on year and month.



2. Flights with highest arrival delay



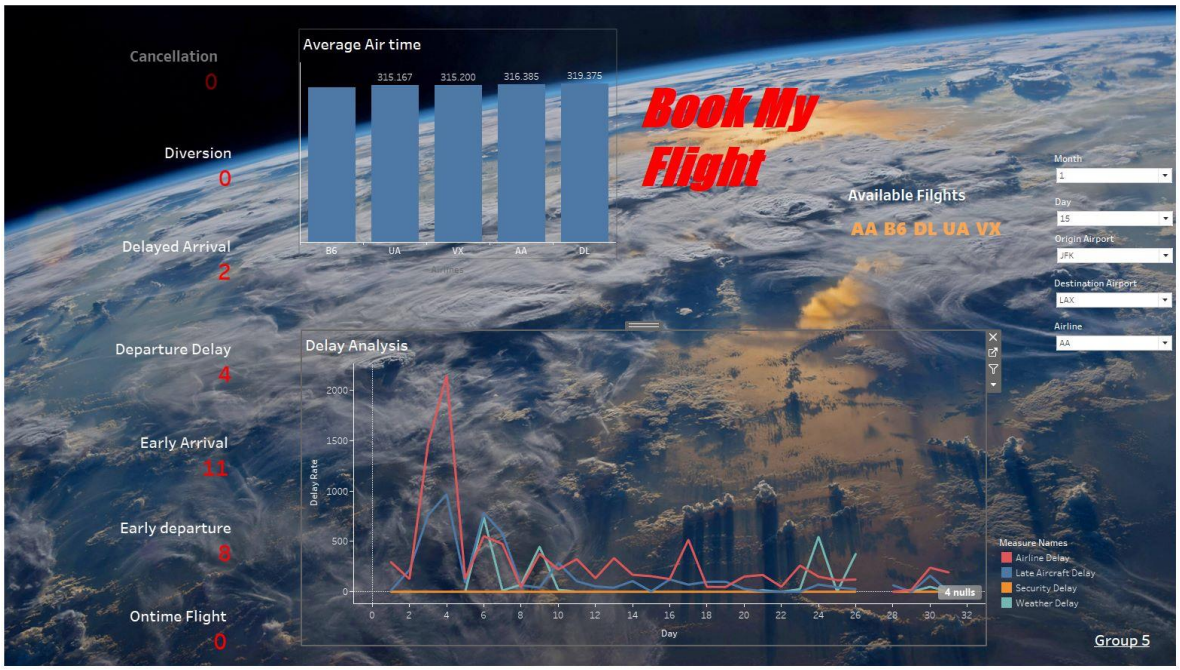
3. Flights that departed earlier than their scheduled departure time



VISUALIZATION

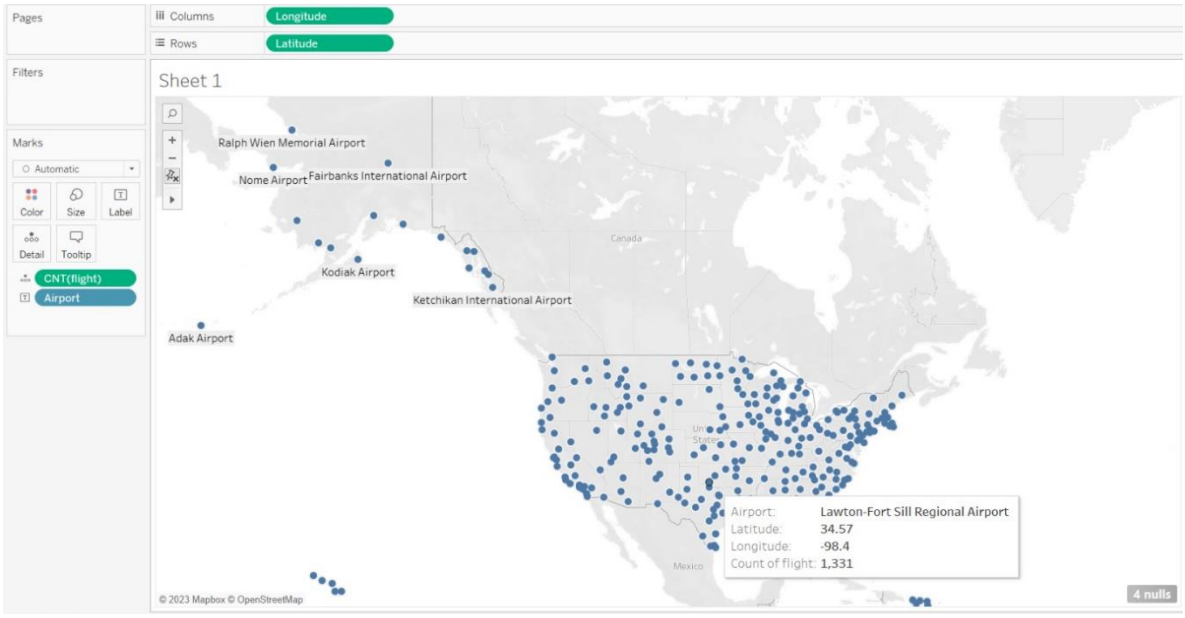
Visualization was done using tableau. A dynamic dashboard for airline and flight analysis is done. In which the users will get to know real-time information about any flight airtime, delays, and cancellations due to specific reasons.

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



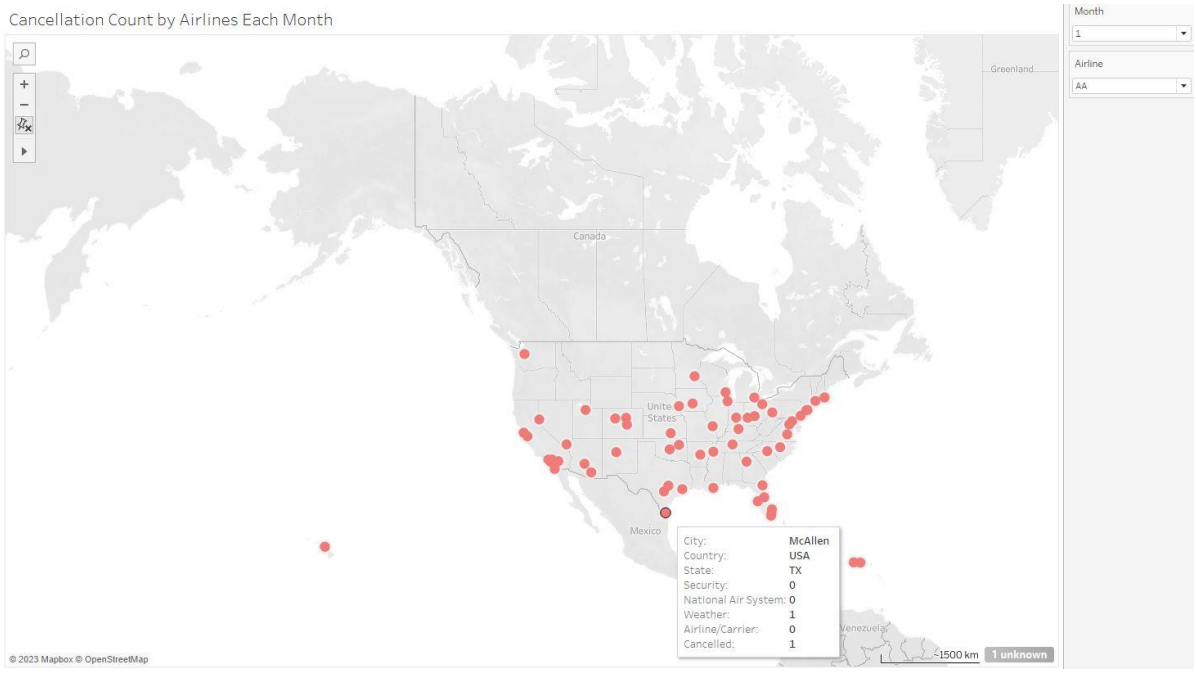
The dashboard is dynamic and provides results regarding a flight delay due to airline-delay, late aircraft delay, security delay, weather delay. Average airtime for each aircraft based on the user mentioned airport is obtained. The dashboard also provides better understanding of how many cancellations 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



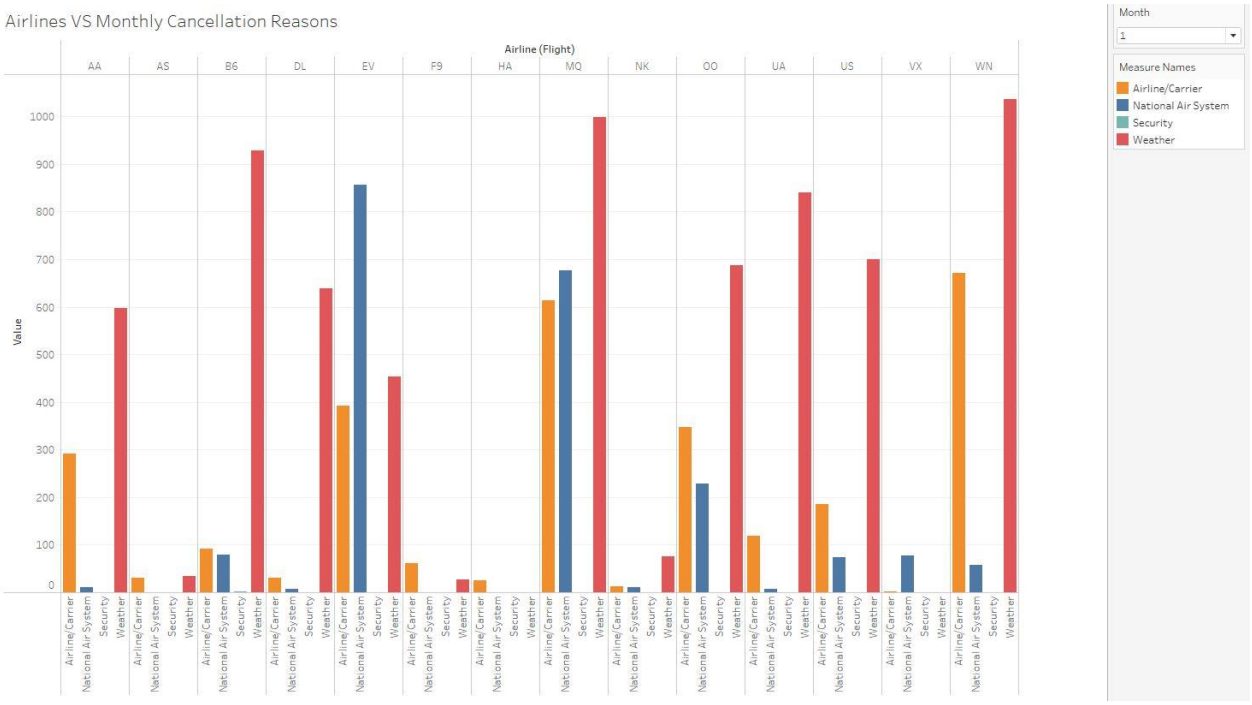
This provides understanding about the number of flights flown from the origin airports.

2. Cancellation count by airlines for each month



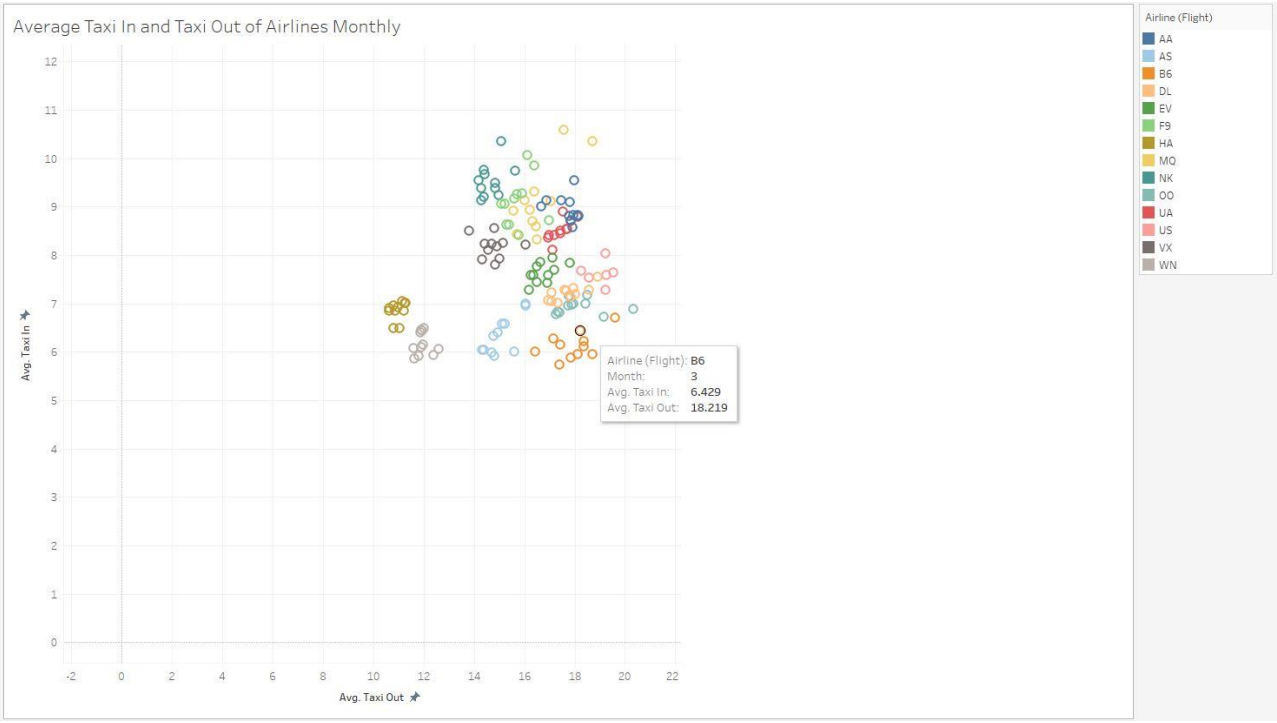
It depicts the count of airlines which was cancelled due to some specific reasons on a specific month.

3. Based on monthly updates on cancellation reasons for airline



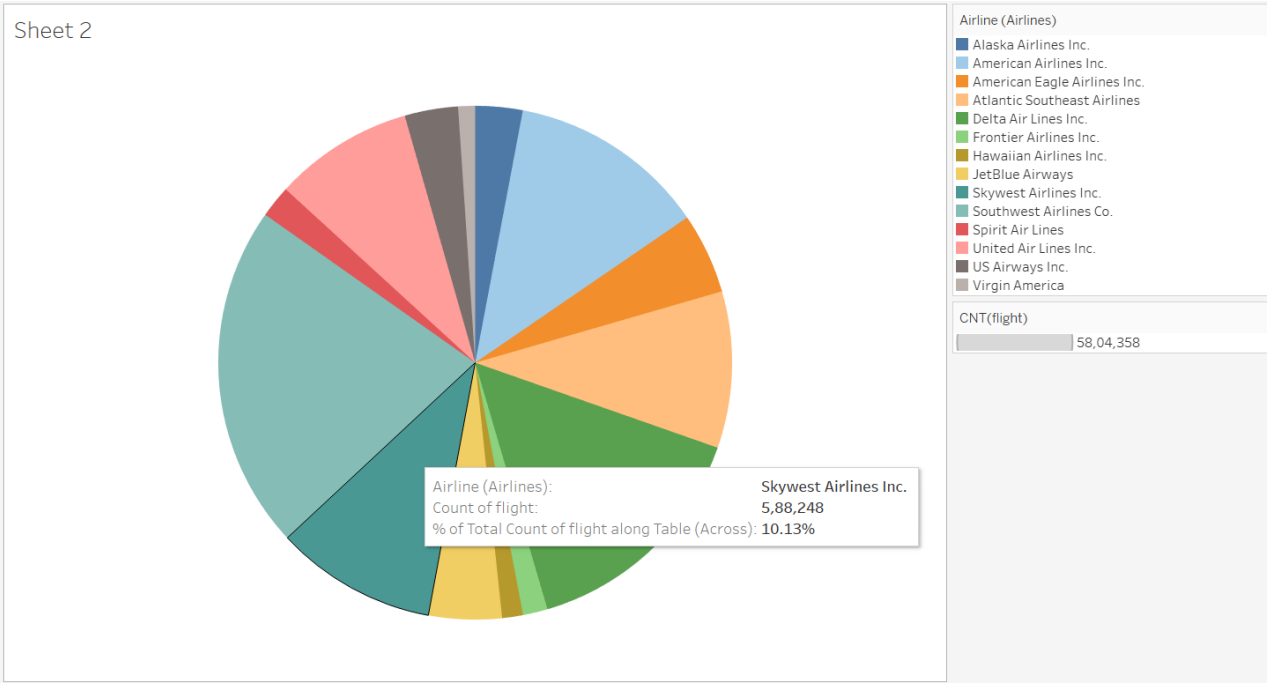
This chart helps to understand the cancellation rate for airlines(flight) which were cancelled.

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 minimalize 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