SAPIENZA
UNIVERSITÀ DI ROMA

Data Warehousing & Analysis

Data Management Project - a.y. 2024/2025

Daidone Giuseppe 2122594, Sheibani Davood 2126056

Dataset

The chosen dataset was downloaded from Kaggle and it contains the **US Airlines Domestic Departure Data in 2022**.

It consists of multiple .csv files:

- *CompleteData.csv*: Compiled dataset of US domestic flight data with added airport, aircraft and present weather data.
- *Stations.csv*: Information about airports.
- *Carriers.csv*: Information about operating carrier code and marketing name.
- *ActiveWeather.csv*: Information about weather status encoding.
- *Cancellation.csv*: Information about cancellation reason encoding.

ETL Operations(1)

To begin with, the focus of this project was on analyzing the reasons behind delayed and canceled flights. For this, we filtered the dataset to include only flights that were either delayed or canceled. Once the relevant data was isolated, we checked for duplicates in the dataset and removed them to maintain consistency.

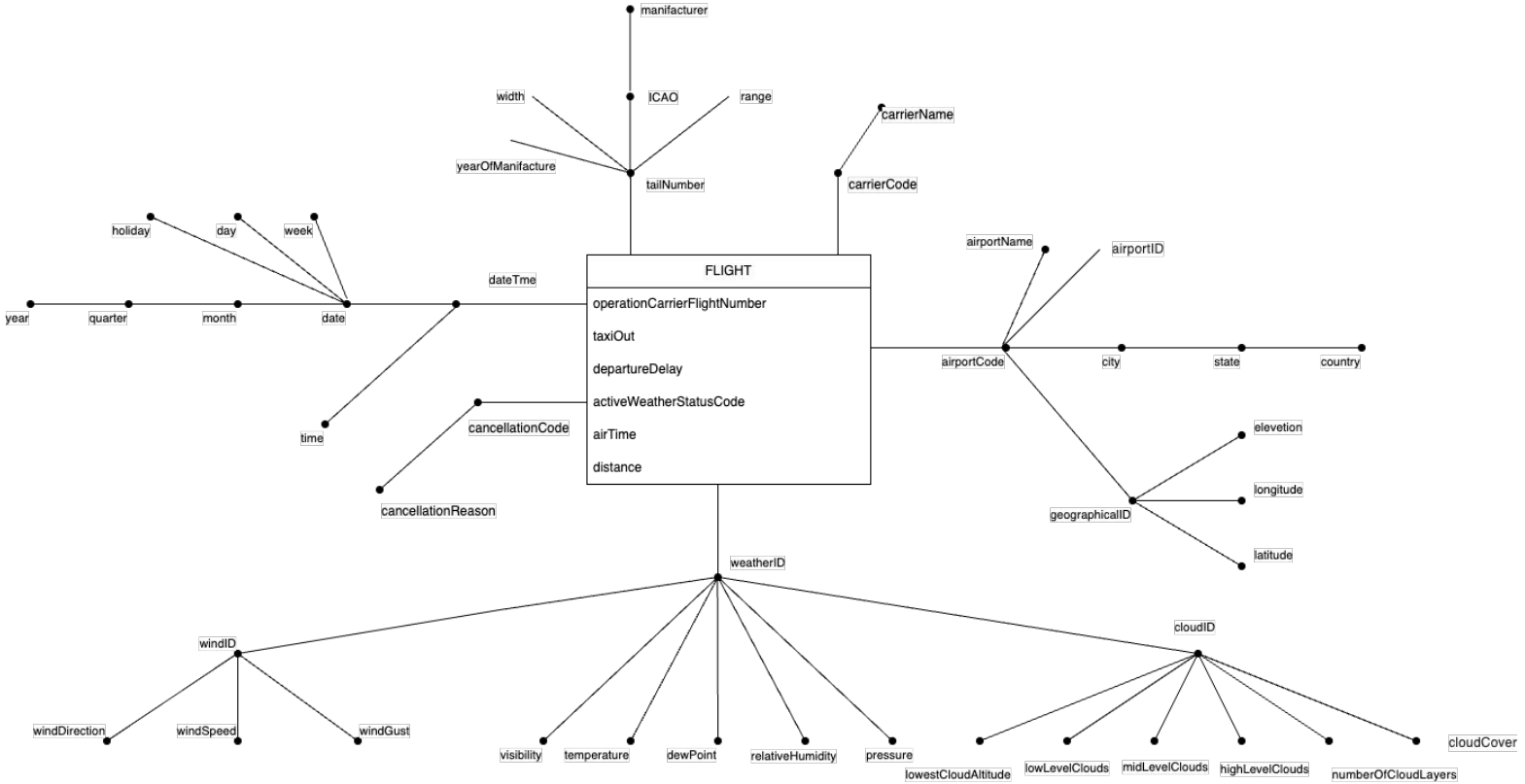
Next, we enriched the dataset by deriving new attributes. We incorporated holiday information using the US Federal Holiday Calendar for 2022, adding a column to indicate whether a flight occurred on a holiday. From the `DEP_TIME` attribute, we extracted several time-related features, such as the exact time, day, week, month, quarter, and year of the departure, which were then added to the main table.

ETL Operations(2)

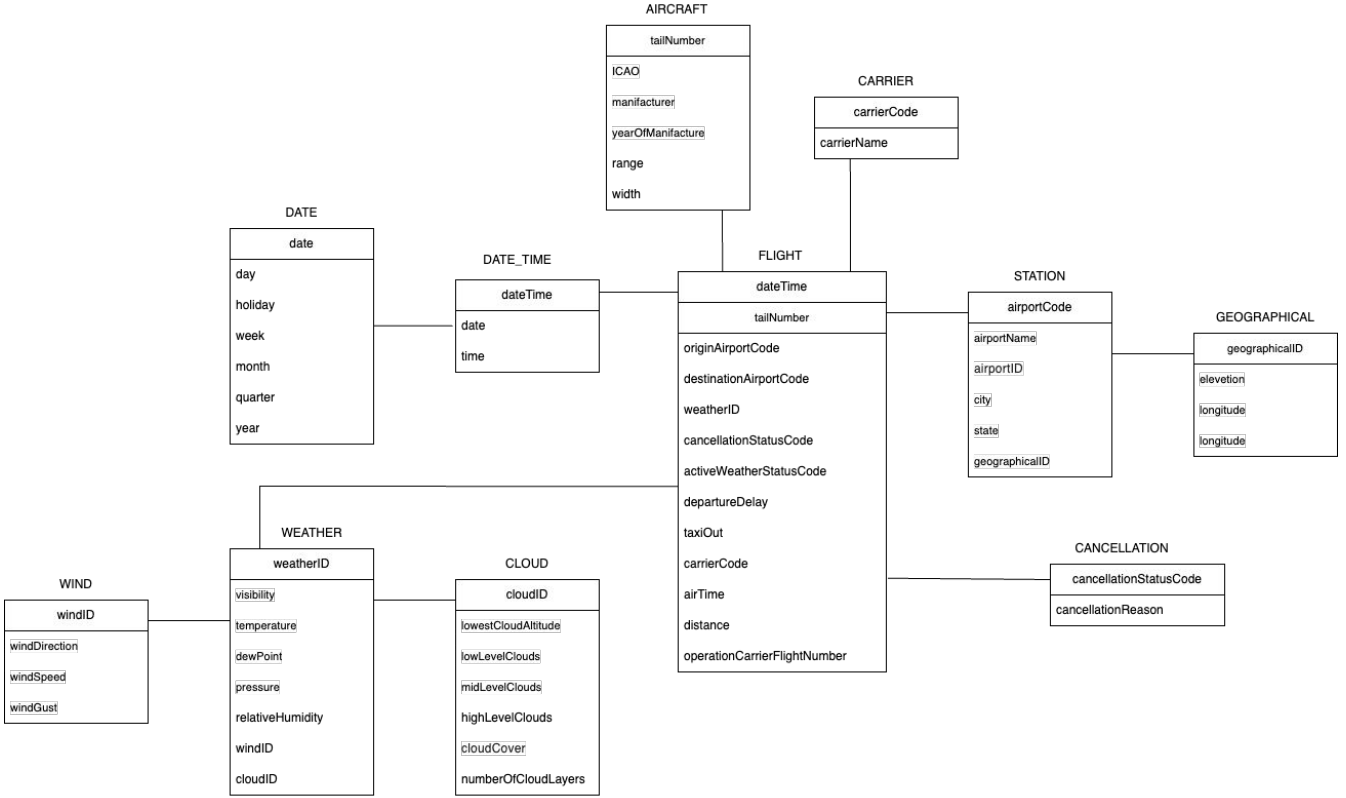
To normalize the data, we created foreign keys to link the main table to dimension tables. These included `weather_ID` for weather-related data, `cloud_ID` for cloud-related data, `wind_ID` for wind-related data, and `geographical_ID` for geographical data. Each foreign key referenced its respective dimension table, where the attributes for that category were stored separately.

Finally, we cleaned up the main table by dropping redundant and unnecessary attributes, leaving only relevant columns for analysis. The result is a cleaned **main table**, containing derived attributes and foreign keys, ready for in-depth analysis.

DFM Model


manifacturer
width
ICAO
range
yearOfManifacture
tailNumber
carrierName
carrierCode
holiday
day
week
dateTme
year
quarter
month
date
time
FLIGHT
operationCarrierFlightNumber
taxiOut
departureDelay
activeWeatherStatusCode
airTime
distance
cancellationCode
cancellationReason
airportName
airportID
airportCode
city
state
country
elevetion
longitude
geographicalID
latitude
weatherID
windID
windDirection
windSpeed
windGust
visibility
temperature
dewPoint
relativeHumidity
pressure
cloudID
lowestCloudAltitude
lowLevelClouds
midLevelClouds
highLevelClouds
numberOfCloudLayers
cloudCover

Snowflake Scheme

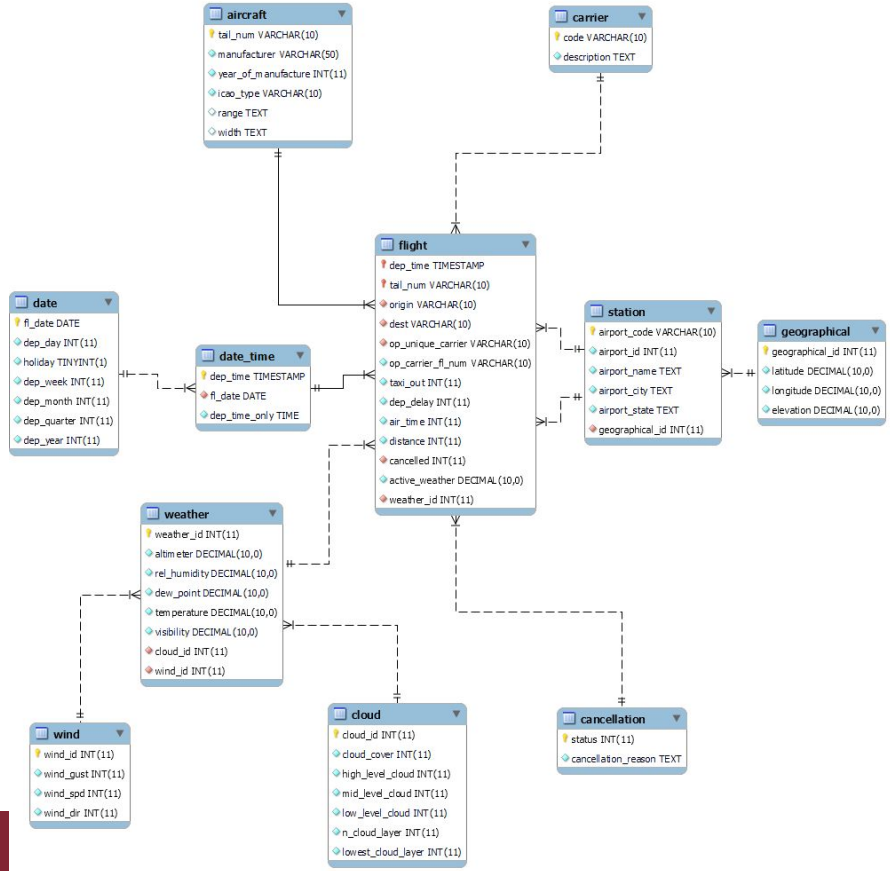


Constructing the DB

We used a python script to normalize and extract tables. We then created the database, the tables and the attributes in PostgreSQL.

After creating the database, we started executing the queries for analysis.

Database


aircraft
tail_num VARCHAR(10)
manufacturer VARCHAR(50)
year_of_manufacture INT(11)
icao_type VARCHAR(10)
range TEXT
width TEXT
carrier
code VARCHAR(10)
description TEXT
flight
dep_time TIMESTAMP
tail_num VARCHAR(10)
origin VARCHAR(10)
dest VARCHAR(10)
op_unique_carrier VARCHAR(10)
op_carrier_fl_num VARCHAR(10)
taxi_out INT(11)
dep_delay INT(11)
air_time INT(11)
distance INT(11)
cancelled INT(11)
active_weather DECIMAL(10,0)
weather_id INT(11)
date
fl_date DATE
dep_day INT(11)
holiday TINYINT(1)
dep_week INT(11)
dep_month INT(11)
dep_quarter INT(11)
dep_year INT(11)
date_time
dep_time TIMESTAMP
fl_date DATE
dep_time_only TIME
station
airport_code VARCHAR(10)
airport_id INT(11)
airport_name TEXT
airport_city TEXT
airport_state TEXT
geographical_id INT(11)
geographical
geographical_id INT(11)
latitude DECIMAL(10,0)
longitude DECIMAL(10,0)
elevation DECIMAL(10,0)
weather
weather_id INT(11)
altimeter DECIMAL(10,0)
rel_humidity DECIMAL(10,0)
dew_point DECIMAL(10,0)
temperature DECIMAL(10,0)
visibility DECIMAL(10,0)
cloud_id INT(11)
wind_id INT(11)
wind
wind_id INT(11)
wind_gust INT(11)
wind_spd INT(11)
wind_dir INT(11)
cloud
cloud_id INT(11)
cloud_cover INT(11)
high_level_cloud INT(11)
mid_level_cloud INT(11)
low_level_cloud INT(11)
n_cloud_layer INT(11)
lowest_cloud_layer INT(11)
cancellation
status INT(11)
cancellation_reason TEXT

Analysis Queries(1)

Total number of cancellations by reason.

```
SELECT c.CANCELLATION_REASON, COUNT(f.CANCELLED) AS total_cancellations
FROM Flight f
INNER JOIN Cancellation c ON f.CANCELLED = c.STATUS
GROUP BY c.CANCELLATION_REASON
ORDER BY total_cancellations DESC;
```

Cancellation Reason	Total Cancellations
Not Cancelled	2625802
Weather Cancellation	46544
Carrier Cancellation	33408
National Air System Cancellation	9055
Security Cancellation	81

Analysis Queries(2)

Weather situation with most delays.

```
SELECT w.TEMPERATURE, w.VISIBILITY, c.CLOUD_COVER, COUNT(f.DEP_DELAY) AS delay_count
FROM Flight f
JOIN Weather w ON f.weather_ID = w.weather_ID
JOIN Cloud c ON w.cloud_ID = c.cloud_ID
WHERE f.DEP_DELAY > 0
GROUP BY w.TEMPERATURE, w.VISIBILITY, c.CLOUD_COVER
ORDER BY delay_count DESC
LIMIT 10;
```

Analysis Queries(3)

Results for Weather Impact On Delays:

Temperature	Visibility	Cloud Cover	Count
26.11	10.0	3	19111
26.72	10.0	3	17341
27.22	10.0	3	17115
25.0	10.0	3	16986
25.61	10.0	3	16895
27.78	10.0	3	16853
28.28	10.0	3	16636
23.89	10.0	3	16395
22.78	10.0	3	15852

Analysis Queries(4)

Weather situation with most cancellations.

```
SELECT w.TEMPERATURE, w.VISIBILITY, c.CLOUD_COVER, COUNT(f.CANCELLED) AS cancellations
FROM Flight f
JOIN Weather w ON f.weather_ID = w.weather_ID
JOIN Cloud c ON w.cloud_ID = c.cloud_ID
WHERE f.CANCELLED = 2
GROUP BY w.TEMPERATURE, w.VISIBILITY, c.CLOUD_COVER
ORDER BY cancellations DESC
LIMIT 10;
```

Analysis Queries(5)

Results for Weather Impact On Cancellations:

Temperature	Visibility	Cloud Cover	Count
NaN	NaN	0	818
29.39	10.0	3	490
27.78	10.0	3	472
27.22	10.0	3	404
28.28	10.0	3	387
6.11	10.0	3	375
26.11	10.0	3	345
25.0	10.0	4	330
24.39	10.0	1	315
-4.39	10.0	4	314

Analysis Queries(6)

Find the airports ranked by number of delays with the most frequent weather conditions corresponding of delayed flights.

```
SELECT f.ORIGIN AS airport_code, s.AIRPORT_NAME, s.AIRPORT_CITY, w.TEMPERATURE, w.VISIBILITY, c.CLOUD_COVER,
COUNT(f.DEP_DELAY) AS delay_count
FROM Flight f
JOIN Weather w ON f.weather_ID = w.weather_ID
JOIN Cloud c ON w.cloud_ID = c.cloud_ID  -- Joining with Cloud table
JOIN Station s ON f.ORIGIN = s.AIRPORT_CODE
WHERE f.DEP_DELAY > 0 AND w.TEMPERATURE BETWEEN 22.0 AND 28.0 AND w.VISIBILITY = 10.0 AND c.CLOUD_COVER
IN (3, 4)  -- Example condition: cloud cover is 3 or 4
GROUP BY f.ORIGIN, s.AIRPORT_NAME, s.AIRPORT_CITY, w.TEMPERATURE, w.VISIBILITY, c.CLOUD_COVER
ORDER BY delay_count DESC
LIMIT 10;
```

Analysis Queries(7)

Delays by Airport with Specific Weather Conditions:

Airport Code	Airport Name	City	Temperature	Visibility	Cloud Cover	Delay Count
MIA	Miami International	Miami, FL	26.11	10.0	3	1448
MIA	Miami International	Miami, FL	26.72	10.0	3	1428
ATL	Hartsfield-Jackson Atlanta International	Atlanta, GA	27.78	10.0	3	1308
ATL	Hartsfield-Jackson Atlanta International	Atlanta, GA	26.11	10.0	3	1284
MIA	Miami International	Miami, FL	27.22	10.0	3	1279
ATL	Hartsfield-Jackson Atlanta International	Atlanta, GA	23.28	10.0	3	1276
ATL	Hartsfield-Jackson Atlanta International	Atlanta, GA	23.89	10.0	3	1251
MIA	Miami International	Miami, FL	27.78	10.0	3	1229
ATL	Hartsfield-Jackson Atlanta International	Atlanta, GA	27.22	10.0	3	1215
ATL	Hartsfield-Jackson Atlanta International	Atlanta, GA	26.72	10.0	3	1155

Analysis Queries(8)

Find the airports ranked by number of delays with the most frequent weather conditions corresponding of cancelled flights.

```
SELECT f.ORIGIN AS airport_code, s.AIRPORT_NAME, s.AIRPORT_CITY, w.TEMPERATURE, w.VISIBILITY, c.CLOUD_COVER, COUNT(f.CANCELLED) AS cancellation_count
FROM Flight f
JOIN Weather w ON f.weather_ID = w.weather_ID
JOIN Cloud c ON w.cloud_ID = c.cloud_ID  -- Joining with Cloud table
JOIN Station s ON f.ORIGIN = s.AIRPORT_CODE
WHERE f.CANCELLED = 2 AND w.TEMPERATURE BETWEEN 22.0 AND 28.0 AND w.VISIBILITY = 10.0 AND c.CLOUD_COVER IN (3, 4)  -- Example condition: cloud cover is 3 or 4
GROUP BY f.ORIGIN, s.AIRPORT_NAME, s.AIRPORT_CITY, w.TEMPERATURE, w.VISIBILITY, c.CLOUD_COVER
ORDER BY cancellation_count DESC
LIMIT 10;
```

Analysis Queries(9)

Cancellations by Airport with Specific Weather Conditions:

Airport Code	Airport Name	City	Temperature	Visibility	Cloud Cover	Cancellation Count
DFW	Dallas/Fort Worth InternationalDallas/	Fort Worth, TX	27.78	10.0	4	125
DFW	Dallas/Fort Worth InternationalDallas/	Fort Worth, TX	25.0	10.0	4	102
DFW	Dallas/Fort Worth InternationalDallas/	Fort Worth, TX	26.72	10.0	4	98
MCO	Orlando International	Orlando, FL	25.0	10.0	4	83
CLT	Charlotte Douglas International	Charlotte,NC	27.78	10.0	3	68
CLT	Charlotte Douglas International	Charlotte, NC	23.28	10.0	4	67
MIA	Miami International	Miami, FL	27.22	10.0	3	65
TPA	Tampa International	Tampa, FL	26.72	10.0	3	64
DEN	Denver International	Denver, CO	23.28	10.0	3	59
EWR	Newark Liberty International	Newark, NJ	23.89	10.0	3	56

Analysis Queries(10)

Month of the year with most delays.

```
SELECT EXTRACT(MONTH FROM f.DEP_TIME) AS month, w.TEMPERATURE, w.VISIBILITY, c.CLOUD_COVER,
COUNT(f.DEP_DELAY) AS delay_count
FROM Flight f
JOIN Weather w ON f.weather_ID = w.weather_ID
JOIN Cloud c ON w.cloud_ID = c.cloud_ID  -- Joining with Cloud table
WHERE f.DEP_DELAY > 0 AND w.TEMPERATURE BETWEEN 22.0 AND 28.0 AND w.VISIBILITY = 10.0 AND c.CLOUD_COVER
IN (3, 4)  -- Example condition: cloud cover is 3 or 4
GROUP BY month, w.TEMPERATURE, w.VISIBILITY, c.CLOUD_COVER
ORDER BY delay_count DESC
LIMIT 10;
```

Analysis Queries(11)

Delays by Date of the Year with Specific Weather Conditions:

Month	Temperature	Visibility	Cloud Cover	Delay Count
8	26.11	10.0	3	4144
8	27.78	10.0	3	3803
8	27.22	10.0	3	3724
8	26.72	10.0	3	3590
7	26.72	10.0	3	3503
7	27.78	10.0	3	3145
7	26.11	10.0	3	3096
7	25.61	10.0	3	3018
7	25.0	10.0	3	2970
8	24.39	10.0	3	2956

Analysis Queries(12)

Month of the year with most cancellations.

```
SELECT EXTRACT(MONTH FROM f.DEP_TIME) AS month, w.TEMPERATURE, w.VISIBILITY, c.CLOUD_COVER,
COUNT(f.CANCELLED) AS cancellation_count
FROM Flight f
JOIN Weather w ON f.weather_ID = w.weather_ID
JOIN Cloud c ON w.cloud_ID = c.cloud_ID  -- Joining with Cloud table
WHERE f.CANCELLED = 2 AND w.TEMPERATURE BETWEEN 22.0 AND 28.0 AND w.VISIBILITY = 10.0 AND c.CLOUD_COVER
IN (3, 4)  -- Example condition: cloud cover is 3 or 4
GROUP BY month, w.TEMPERATURE, w.VISIBILITY, c.CLOUD_COVER
ORDER BY cancellation_count DESC
LIMIT 10;
```

Analysis Queries(13)

Cancellations by Date of the Year with Specific Weather Conditions:

Month	Temperature	Visibility	Cloud Cover	Cancellation Count
8	27.78	10.0	3	158
5	25.0	10.0	4	150
8	26.72	10.0	4	138
8	27.22	10.0	3	136
8	26.11	10.0	3	126
6	27.78	10.0	3	120
6	27.22	10.0	3	114
9	25.0	10.0	4	103
6	23.28	10.0	4	97
8	25.0	10.0	3	96

Analysis Queries(14)

Delayed flights by day of the month.

```
SELECT d.DEP_DAY AS day_of_month, COUNT(f.DEP_DELAY) AS delayed_flights
FROM Flight f
JOIN Date_Time dt ON f.DEP_TIME = dt.DEP_TIME
JOIN Date d ON dt.FL_DATE = d.FL_DATE
WHERE f.DEP_DELAY > 0
GROUP BY d.DEP_DAY
ORDER BY delayed_flights DESC;
```

Analysis Queries(15)

Results for Delayed Flights By Day Of Month:

day_of_month	delayed_flights
18	95650
22	93709
17	93030
23	91604
13	91123
21	89879
19	89061
24	88371
11	88222
14	87909
25	87733
6	87528
5	87106
20	86520
15	86502
27	86328
28	86306
7	86272
16	86021
26	85740
4	85497
12	85484
2	83764
3	82819
1	82618
9	82015
10	81308
8	79307
30	74300
29	70587
31	46375

Analysis Queries(16)

Canceled flights holidays vs non holidays.

```
SELECT d.Holiday, COUNT(f.CANCELLED) AS canceled_flights
FROM Flight f
JOIN Date_Time dt ON f.DEP_TIME = dt.DEP_TIME
JOIN Date d ON dt.FL_DATE = d.FL_DATE
WHERE f.CANCELLED > 0
GROUP BY d.Holiday
ORDER BY canceled_flights DESC;
```

Analysis Queries(17)

Results for Canceled Flights Holidays Vs Non Holidays:

holiday	canceled_flights
False	80863
True	8225

Analysis Queries(18)

Delayed flights holidays vs non holidays.

```
SELECT d.Holiday, COUNT(f.DEP_DELAY) AS delayed_flights
FROM Flight f
JOIN Date_Time dt ON f.DEP_TIME = dt.DEP_TIME
JOIN Date d ON dt.FL_DATE = d.FL_DATE
WHERE f.DEP_DELAY > 0
GROUP BY d.Holiday
ORDER BY delayed_flights DESC;
```

Analysis Queries(19)

Results for Delayed Flights Holidays Vs Non Holidays:

holiday	delayed_flights
False	2481809
True	146879

Analysis Queries(20)

Peak days for delays.

```
SELECT d.FL_DATE AS flight_date, COUNT(f.DEP_DELAY) AS delayed_flights
FROM Flight f
JOIN Date_Time dt ON f.DEP_TIME = dt.DEP_TIME
JOIN Date d ON dt.FL_DATE = d.FL_DATE
WHERE f.DEP_DELAY > 0
GROUP BY d.FL_DATE
ORDER BY delayed_flights DESC
LIMIT 10;
```

Analysis Queries(21)

Results for Peak Days For Delays:

flight_date	delayed_flights
2022-01-02	11960
2022-12-22	11184
2022-01-03	11079
2022-06-17	10923
2022-12-23	10858
2022-06-12	10717
2022-02-18	10666
2022-02-25	10583
2022-07-24	10479
2022-03-13	10312

Analysis Queries(22)

Monthly delays and cancellations.

```
SELECT d.DEP_MONTH AS month,
COUNT(CASE WHEN f.DEP_DELAY > 0 THEN 1 END) AS delayed_flights,
COUNT(CASE WHEN f.CANCELLED > 0 THEN 1 END) AS canceled_flights
FROM Flight f
JOIN Date_Time dt ON f.DEP_TIME = dt.DEP_TIME
JOIN Date d ON dt.FL_DATE = d.FL_DATE
GROUP BY d.DEP_MONTH
ORDER BY month;
```

Analysis Queries(23)

Results for Monthly Delays And Cancellations:

month	delayed_flights	canceled_flights
1	183242	14801
2	183361	10008
3	227542	5131
4	225358	6978
5	233523	6477
6	250833	10466
7	249474	6474
8	237039	8774
9	190002	3937
10	197096	2360
11	202501	3423
12	248717	10259

Analysis Queries(24)

Quarter analysis of cancellations.

```
SELECT d.DEP_QUARTER AS quarter, COUNT(f.CANCELLED) AS canceled_flights
FROM Flight f
JOIN Date_Time dt ON f.DEP_TIME = dt.DEP_TIME
JOIN Date d ON dt.FL_DATE = d.FL_DATE
WHERE f.CANCELLED > 0
GROUP BY d.DEP_QUARTER
ORDER BY canceled_flights DESC;
```

```
quarter          | canceled_flights
-----------------------------------------------
1                | 29940
2                | 23921
3                | 19185
4                | 16042
```

Analysis Queries(25)

Quarter analysis of delays.

```
SELECT d.DEP_QUARTER AS quarter,COUNT(f.DEP_DELAY) AS delayed_flights
FROM Flight f
JOIN Date_Time dt ON f.DEP_TIME = dt.DEP_TIME
JOIN Date d ON dt.FL_DATE = d.FL_DATE
WHERE f.DEP_DELAY > 0
GROUP BY d.DEP_QUARTER
ORDER BY delayed_flights DESC;
```

Results for Quarterly Analysis Of Delayed Flights:

quarter	delayed_flights
2	709714
3	676515
4	648314
1	594145

Analysis Queries(26)

Hour distribution of delayed flights.

```
SELECT EXTRACT(HOUR FROM dt.DEP_TIME_ONLY) AS departure_hour, COUNT(f.DEP_DELAY) AS delayed_flights
FROM Flight f
JOIN Date_Time dt ON f.DEP_TIME = dt.DEP_TIME
WHERE f.DEP_DELAY > 0
GROUP BY departure_hour
ORDER BY delayed_flights DESC;
```

Analysis Queries(27)

Results for Hourly Distribution Of Delayed Flights:

departure_hour	delayed_flights
18	187586
17	184729
19	179764
16	177325
15	174777
20	167618
14	165196
13	164192
12	152860
11	151480
10	143118
21	135458
9	120460
8	115669
22	106180
7	95339
6	81804
23	58922
5	30239
0	23776
1	8414
2	2345
3	924
4	513

Analysis Queries(28)

Impact of holidays on delayed flights.

```
SELECT d.Holiday, AVG(f.DEP_DELAY) AS average_delay_minutes
FROM Flight f
JOIN Date_Time dt ON f.DEP_TIME = dt.DEP_TIME
JOIN Date d ON dt.FL_DATE = d.FL_DATE
WHERE f.DEP_DELAY > 0
GROUP BY d.Holiday
ORDER BY average_delay_minutes DESC;
```

holiday	average_delay_minutes
True	44.0670620034177793
False	39.7293732112342247

Analysis Queries(29)

Top 10 aircrafts with most delayed flights.

```
SELECT a.TAIL_NUM, a.MANUFACTURER, a.YEAR_OF_MANUFACTURE, c.DESCRIPTION AS carrier_name,
COUNT(f.DEP_DELAY) AS delayed_flights
FROM Flight f
JOIN Aircraft a ON f.TAIL_NUM = a.TAIL_NUM
JOIN Carrier c ON f.OP_UNIQUE_CARRIER = c.CODE
WHERE f.DEP_DELAY > 0
GROUP BY a.TAIL_NUM, a.MANUFACTURER, c.DESCRIPTION
ORDER BY delayed_flights DESC
LIMIT 10;
```

Analysis Queries(30)

Results for Top 10 Aircraft With Most Delayed Flights:

Tail Number	Manufacturer	Year of Manufacture,	Carrier Name	Count
N492HA	Boeing	2004	Hawaiian Airlines Inc.	1526
N475HA	Boeing	2001	Hawaiian Airlines Inc.	1467
N493HA	Boeing	2005	Hawaiian Airlines Inc.	1454
N476HA	Boeing	2001	Hawaiian Airlines Inc.	1444
N483HA	Boeing	2001	Hawaiian Airlines Inc.	1429
N489HA	Boeing	1998	Hawaiian Airlines Inc.	1425
N490HA	Boeing	2000	Hawaiian Airlines Inc.	1383
N478HA	Boeing	2001	Hawaiian Airlines Inc.	1381
N494HA	Boeing	2004	Hawaiian Airlines Inc.	1355
N484HA	Boeing	2001	Hawaiian Airlines Inc.	1347

Analysis Queries(31)

Top 10 aircrafts with most cancelled flights.

```
SELECT a.TAIL_NUM, a.MANUFACTURER, a.YEAR_OF_MANUFACTURE, c.DESCRIPTION AS carrier_name,
COUNT(f.CANCELLED) AS canceled_flights
FROM Flight f
JOIN Aircraft a ON f.TAIL_NUM = a.TAIL_NUM
JOIN Carrier c ON f.OP_UNIQUE_CARRIER = c.CODE
WHERE f.CANCELLED > 0
GROUP BY a.TAIL_NUM, a.MANUFACTURER, c.DESCRIPTION
ORDER BY canceled_flights DESC
LIMIT 10;
```

Analysis Queries(32)

Results for Top 10 Aircraft With Most Canceled Flights:

Tail Number	Manufacturer	Year of Manufacture	Carrier Name	Count
N329JB	Embraer	2011	JetBlue Airways	57
N284JB	Embraer	2008	JetBlue Airways	57
N733YX	Embraer	2016	Republic Airline	55
N258JB	Embraer	2006	JetBlue Airways	55
N727YX	Embraer	2015	Republic Airline	55
N375JB	Embraer	2013	JetBlue Airways	54
N192JB	Embraer	2005	JetBlue Airways	54
N746YX	Embraer	2017	Republic Airline	54
N324JB	Embraer	2010	JetBlue Airways	53
N731YX	Embraer	2015	Republic Airline	53

Analysis Queries(33)

Carrier that use in average the aircrafts with most delayed and cancelled flights, order by the average delay and average cancellation.

```
WITH Most_Delayed_Canceled_Aircraft AS (
SELECT f.TAIL_NUM,
COUNT(CASE WHEN f.DEP_DELAY > 0 THEN 1 END) AS total_delays,
COUNT(CASE WHEN f.CANCELLED > 0 THEN 1 END) AS
total_cancellations,
COUNT(*) AS total_flights
FROM Flight f
GROUP BY f.TAIL_NUM
ORDER BY COUNT(CASE WHEN f.DEP_DELAY > 0 THEN 1 END) +
COUNT(CASE WHEN f.CANCELLED > 0 THEN 1 END) DESC
 LIMIT 1000 -- We want to get top 1000 aircraft for calculation
 ),
 Carrier_Usage AS (
 SELECT c.DESCRIPTION AS carrier_name, f.TAIL_NUM,
  AVG(f.DEP_DELAY) AS avg_delay,
  AVG(f.CANCELLED::INT) AS avg_cancellation_rate
  FROM Most_Delayed_Canceled_Aircraft mdc
  JOIN Flight f ON mdc.TAIL_NUM = f.TAIL_NUM
  JOIN Carrier c ON f.OP_UNIQUE_CARRIER = c.CODE
  GROUP BY c.DESCRIPTION, f.TAIL_NUM
  ),
Carrier_Average AS (
 SELECT carrier_name,
 COUNT(DISTINCT TAIL_NUM) AS aircraft_count,
 AVG(avg_delay) AS avg_delay,
 AVG(avg_cancellation_rate) AS avg_cancellation_rate
 FROM Carrier_Usage
 GROUP BY carrier_name
)
SELECT carrier_name, aircraft_count, avg_delay, avg_cancellation_rate
FROM Carrier_Average
ORDER BY avg_delay DESC, avg_cancellation_rate DESC
LIMIT 10;
```

Analysis Queries(34)

Carriers Using Aircraft with Most Delayed and Canceled Flights (Top 10):

Carrier Name	Aircraft Count	Average Delay	Avg Cancellation Rate
SkyWest Airlines Inc.	9	60.2714652754177863	0.05013998435502001458
JetBlue Airways	43	53.2145644217918157	0.10767998267847953939
Mesa Airlines Inc.	1	51.6109215017064846	0.05631399317406143345
Frontier Airlines Inc.	93	50.6706681173784479	0.00257879490816387329
American Airlines Inc.	19	48.2014384300682322	0.08403699712228783032
PSA Airlines Inc.	12	43.5404850945158330	0.06592506261195212974
Allegiant Air	7	42.8226759893479071	0.00777529476366278443
Spirit Air Lines	37	40.9681206838788799	0.06039378323011508991
United Air Lines Inc.	1	39.6575809199318569	0.00511073253833049404
Delta Air Lines Inc.	41	34.4158949366381556	0.04367981876686748802

Analysis Queries(35)

Carrier with the most cancelled flights.

```
SELECT c.DESCRIPTION AS carrier_name, f.OP_UNIQUE_CARRIER AS carrier_code, COUNT(*) AS total_delays
FROM Flight f
INNER JOIN Carrier c ON f.OP_UNIQUE_CARRIER = c.CODE
WHERE f.DEP_DELAY > 0
GROUP BY c.DESCRIPTION, f.OP_UNIQUE_CARRIER
ORDER BY total_delays DESC
 LIMIT 10;
```

Analysis Queries(36)

Carriers with the Most Canceled Flights:

Carrier Name	Carrier Code	Total Cancellations
American Airlines Inc.	AA	18612
Southwest Airlines Co.	WN	17990
Delta Air Lines Inc.	DL	9254
SkyWest Airlines Inc.	OO	7192
Republic Airline	YX	7135
JetBlue Airways	B6	6441
Endeavor Air Inc.	9E	4223
Alaska Airlines Inc.	AS	3890
Spirit Air Lines	NK	3781
Envoy Air	MQ	2986

Analysis Queries(37)

Busiest airport.

```
SELECT s.AIRPORT_CODE, s.AIRPORT_NAME, s.AIRPORT_CITY, s.AIRPORT_STATE, COUNT(*) AS total_flights
FROM Flight f
JOIN Station s ON f.ORIGIN = s.AIRPORT_CODE OR f.DEST = s.AIRPORT_CODE
GROUP BY s.AIRPORT_CODE, s.AIRPORT_NAME, s.AIRPORT_CITY, s.AIRPORT_STATE
ORDER BY total_flights DESC
LIMIT 10;
```

Analysis Queries(38)

Results for Busiest Airports:

Airport Code	Airport Name	City	State	Total Flights
DEN	Denver International	Denver, CO	Colorado	257479
ATL	Hartsfield-Jackson Atlanta International	Atlanta, GA	Georgia	234542
DFW	Dallas/Fort Worth International	Dallas/Fort Worth, TX	Texas	209922
ORD	Chicago O'Hare International	Chicago, IL	Illinois	198984
LAS	Harry Reid International	Las Vegas, NV	Nevada	170667
LAX	Los Angeles International	Los Angeles, CA	California	146940
PHX	Phoenix Sky Harbor International	Phoenix, AZ	Arizona	141605
MCO	Orlando International	Orlando, FL	Florida	141278
CLT	Charlotte Douglas International	Charlotte, NC	North Carolina	140990
SEA	Seattle/Tacoma International	Seattle, WA	Washington	128398

Analysis Queries(39)

Airports with most cancelled flights.

```
SELECT s.AIRPORT_CODE, s.AIRPORT_NAME, s.AIRPORT_CITY,s.AIRPORT_STATE,
COUNT(f.CANCELLED) AS canceled_flights
FROM Flight f
JOIN Station s ON f.ORIGIN = s.AIRPORT_CODE
WHERE f.CANCELLED > 0
GROUP BY s.AIRPORT_CODE, s.AIRPORT_NAME, s.AIRPORT_CITY, s.AIRPORT_STATE
ORDER BY canceled_flights DESC
LIMIT 10;
```

Analysis Queries(40)

Results for Airports With Most Canceled Flights:

Airport Code	Airport Name	City	State	Canceled Flights
DFW	Dallas/Fort Worth International	Dallas/Fort Worth, TX	Texas	5043
LGA	LaGuardia	New York, NY	New York	4744
ATL	Hartsfield-Jackson Atlanta International	Atlanta, GA	Georgia	3350
CLT	Charlotte Douglas International	Charlotte, NC	North Carolina	3337
ORD	Chicago O'Hare International	Chicago, IL	Illinois	3135
JFK	John F. Kennedy International	New York, NY	New York	3112
DCA	Ronald Reagan Washington National	Washington, DC	Virginia	2952
DEN	Denver International	Denver, CO	Colorado	2948
BOS	Logan International	Boston, MA	Massachusetts	2486
MCO	Orlando International	Orlando, FL	Florida	2235

Analysis Queries(41)

Airports with most delayed flights.

```
SELECT s.AIRPORT_CODE, s.AIRPORT_NAME, s.AIRPORT_CITY, s.AIRPORT_STATE,
COUNT(f.DEP_DELAY) AS delayed_flights
FROM Flight f
JOIN Station s ON f.ORIGIN = s.AIRPORT_CODE
WHERE f.DEP_DELAY > 0
GROUP BY s.AIRPORT_CODE, s.AIRPORT_NAME, s.AIRPORT_CITY, s.AIRPORT_STATE
ORDER BY delayed_flights DESC
LIMIT 10;
```

Analysis Queries(42)

Results for Airports With Most Delayed Flights:

Airport Code	Airport Name	City	State	Delayed Flights
DEN	Denver International	Denver, CO	Colorado	146772
ATL	Hartsfield-Jackson Atlanta International	Atlanta, GA	Georgia	119059
DFW	Dallas/Fort Worth International	Dallas/Fort Worth, TX	Texas	103337
ORD	Chicago O'Hare International	Chicago, IL	Illinois	101182
LAS	Harry Reid International	Las Vegas, NV	Nevada	85863
PHX	Phoenix Sky Harbor International	Phoenix, AZ	Arizona	70664
CLT	Charlotte Douglas International	Charlotte, NC	North Carolina	70140
LAX	Los Angeles International	Los Angeles, CA	California	68471
MCO	Orlando International	Orlando, FL	Florida	68251
SEA	Seattle/Tacoma International	Seattle, WA	Washington	65351

Analysis Queries(43)

Geographical area with most delayed flights.

```
SELECT g.geographical_ID, s.AIRPORT_NAME, s.AIRPORT_CITY, s.AIRPORT_STATE,
COUNT(f.DEP_DELAY) AS delayed_flights
FROM Flight f
JOIN Station s ON f.ORIGIN = s.AIRPORT_CODE
JOIN Geographical g ON s.geographical_ID = g.geographical_ID
WHERE f.DEP_DELAY > 0
GROUP BY g.geographical_ID, s.AIRPORT_NAME, s.AIRPORT_CITY, s.AIRPORT_STATE
ORDER BY delayed_flights DESC
LIMIT 10;
```

Analysis Queries(44)

Results for Geographical Area With Most Delayed Flights:

Geographical ID	Airport Name	City	State	Count
63	Denver International	Denver, CO	Colorado	146772
95	Hartsfield-Jackson Atlanta International	Atlanta, GA	Georgia	119059
322	Dallas/Fort Worth International	Dallas/Fort Worth, TX	Texas	103337
118	Chicago O'Hare International	Chicago, IL	Illinois	101182
238	Harry Reid International	Las Vegas, NV	Nevada	85863
33	Phoenix Sky Harbor International	Phoenix, AZ	Arizona	70664
208	Charlotte Douglas International	Charlotte, NC	North Carolina	70140
54	Los Angeles International	Los Angeles, CA	California	68471
74	Orlando International	Orlando, FL	Florida	68251
352	Seattle/Tacoma International	Seattle, WA	Washington	65351

Analysis Queries(45)

Geographical area with most cancelled flights.

```
SELECT g.geographical_ID, s.AIRPORT_NAME, s.AIRPORT_CITY, s.AIRPORT_STATE,
COUNT(f.CANCELLED) AS canceled_flights
FROM Flight f
JOIN Station s ON f.ORIGIN = s.AIRPORT_CODE
JOIN Geographical g ON s.geographical_ID = g.geographical_ID
WHERE f.CANCELLED > 0
GROUP BY g.geographical_ID, s.AIRPORT_NAME, s.AIRPORT_CITY, s.AIRPORT_STATE
ORDER BY canceled_flights DESC
LIMIT 10;
```

Analysis Queries(46)

Results for Geographical Area With Most Canceled Flights:

Geographical ID	Airport Name	City	State	Count
322	Dallas/Fort Worth International	Dallas/Fort Worth, TX	Texas	5043
246	LaGuardia	New York, NY	New York	4744
95	Hartsfield-Jackson Atlanta International	Atlanta, GA	Georgia	3350
208	Charlotte Douglas International	Charlotte, NC	North Carolina	3337
118	Chicago O'Hare International	Chicago, IL	Illinois	3135
249	John F. Kennedy International	New York, NY	New York	3112
338	Ronald Reagan Washington National	Washington, DC	Virginia	2952
63	Denver International	Denver, CO	Colorado	2948
149	Logan International	Boston, MA	Massachusetts	2486
74	Orlando International	Orlando, FL	Florida	2235

Long story short…we could continue forever…

Summary and Information Inferred(1)

Brief summary and Information Inferred from this project

1. **Project Focus and Dataset**:
   - The project analyzed domestic flight delays and cancellations in the US during 2022.
   - Data was sourced from Kaggle, containing details about flights, stations, carriers, weather conditions, and cancellations.
2. **ETL Process Highlights**:
   - The dataset was filtered to include only delayed or canceled flights.
   - Derived attributes included time-related features (e.g., day, month, quarter) and holiday indicators.
   - Data normalization created foreign keys for linking dimension tables (e.g., weather, cloud, wind, and geographical data).
   - Unnecessary attributes were removed to simplify the main table for analysis.

Summary and Information Inferred(2)

3. **Database Design**:

- A Snowflake Schema was implemented, featuring normalized dimension tables for better query performance.
- PostgreSQL was used for database creation and querying.

4. **Key Analyses and Results**:

- **Delays**:
  - Peak delay days included January 2, December 22, and June 17.
  - Evening hours (16:00–20:00) had the highest number of delayed flights.
  - Specific weather conditions (e.g., temperatures between 22°C–28°C with cloud cover of 3) significantly contributed to delays.

Summary and Information Inferred(3)

- **Cancellations**:
  - January experienced the most cancellations, with weather being a leading reason.
  - Dallas/Fort Worth International Airport had the highest number of canceled flights.
- **Geographical Insights**:
  - Denver International Airport led in delays.
  - Dallas/Fort Worth International Airport had the most cancellations.
- **Carrier and Aircraft**:
  - Hawaiian Airlines' aircraft had the highest delay rates.
  - Carriers such as SkyWest and JetBlue were identified as using the most delayed and canceled aircraft.

Summary and Information Inferred(4)

5. **Conclusion**:

- The analysis highlighted critical trends in flight disruptions, providing insights into weather impacts, operational inefficiencies, and peak disruption periods. This information could guide decision-making in the aviation sector.

Thanks for the attention

/Davood-sh/Data-warehouse