Flights Delay Analysis

Authors:

Dharmi Gala

Naman Mehta

Radhika Sharma

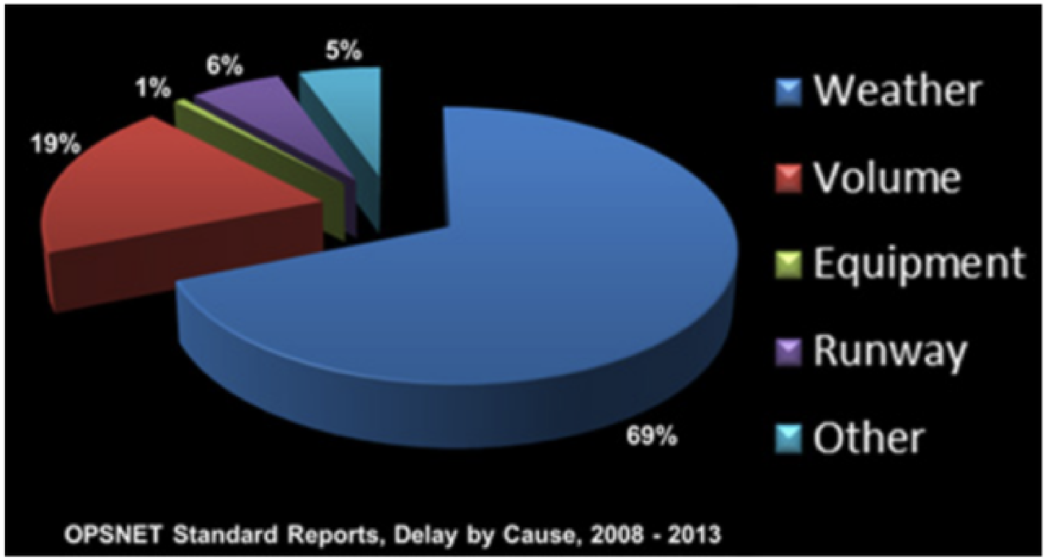
Agenda:

- **Introduction**
  - Business Problem
  - Solution Overview
- **Data Model**
  - Extract, Transform, Load (ETL)
  - Enhanced Entity Relationship (EER) Model
- **Data Analysis**
  - SQL
  - Tableau
- **Conclusion**
  - Challenges faced
  - Future Scope

Introduction

Business Problem

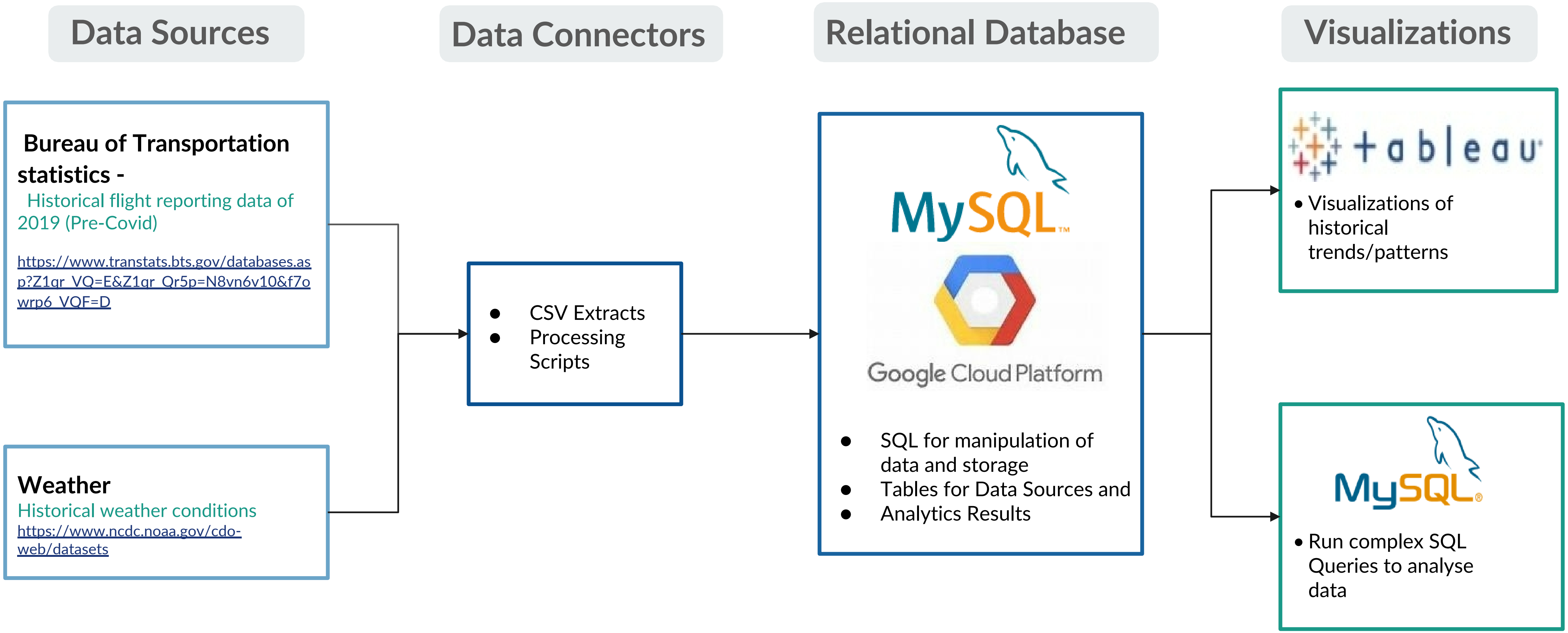
- Currently, the cost to the air carrier operators for an hour of delay ranges from about $1,400 to $4,500, depending on the class of aircraft and if the delay is on the ground or in the air.
- The largest cause of air traffic delay in the National Airspace System is weather.
- Weather caused around 70 percent of system impacting delays of greater than 15 minutes over the six years from 2008 to 2013, as recorded in the OPSNET standard "delay by cause" reports.
- With this in mind, we plan to analyze delays by airline and the effect of weather patterns on flight duration, cancellations and other aircraft operations.



Causes of air traffic delay in the National Airspace System.

Source: Federal Aviation administration (https://www.faa.gov/nextgen/programs/weather/faq#:~:text=Currently%2C%20the%20cost%20to%20the,ground%20or%20in%20the%20air.)

Solution Overview



Data Sources

Data Connectors

Relational Database

Visualizations

Bureau of Transportation statistics -
Historical flight reporting data of 2019 (Pre-Covid)

https://www.transtats.bts.gov/databases.asp?Z1qr_VQ=E&Z1qr_Qr5p=N8vn6v10&f7owrp6_VQF=D

Weather
Historical weather conditions
https://www.ncdc.noaa.gov/cdo-web/datasets

- CSV Extracts
- Processing Scripts

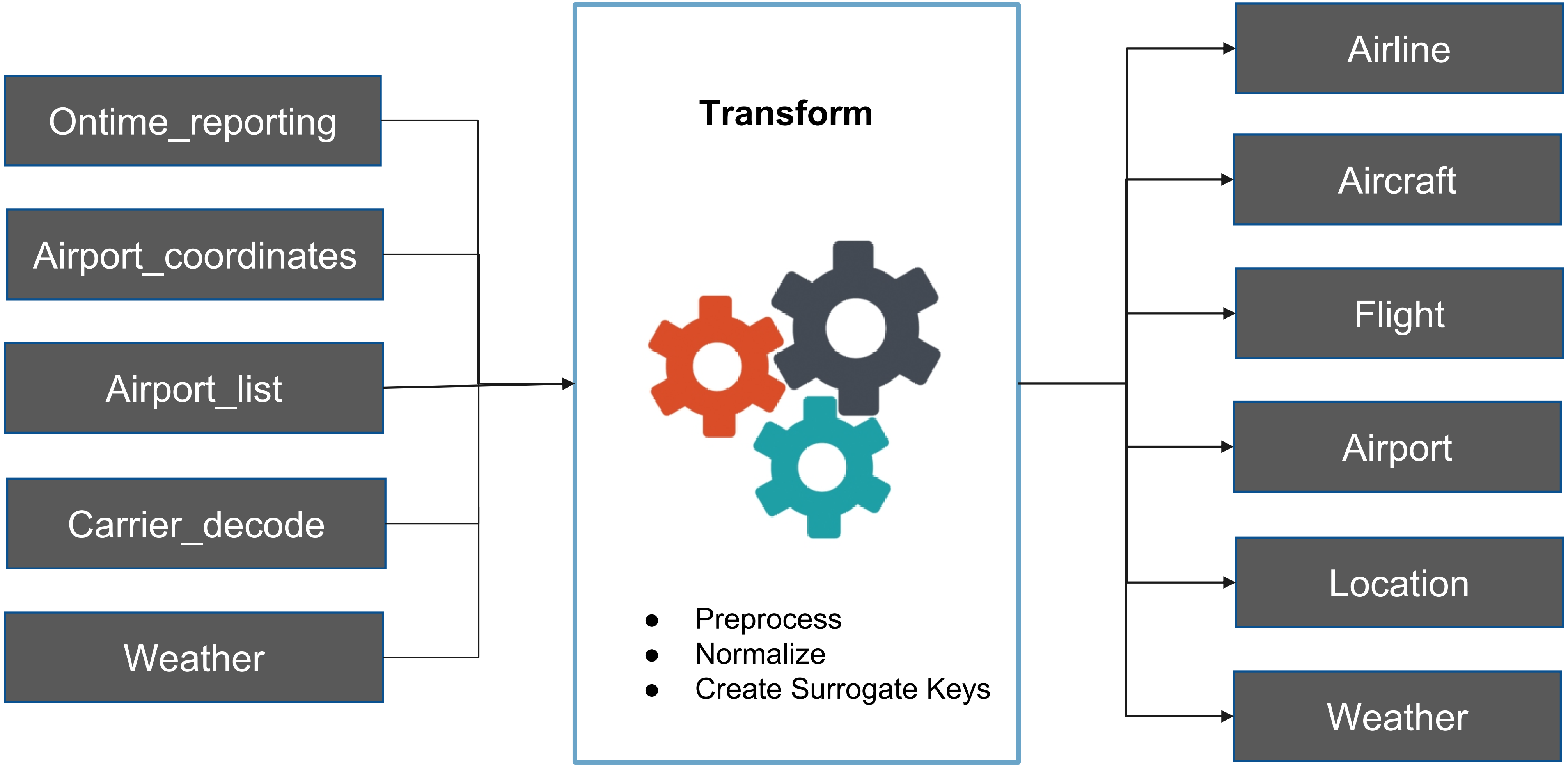
- SQL for manipulation of data and storage
- Tables for Data Sources and Analytics Results

- Visualizations of historical trends/patterns

- Run complex SQL Queries to analyse data

Data Model

Data Model: Extract, Transform



Ontime_reporting

Airport_coordinates

Airport_list

Carrier_decode

Weather

Transform

- Preprocess
- Normalize
- Create Surrogate Keys

Airline

Aircraft

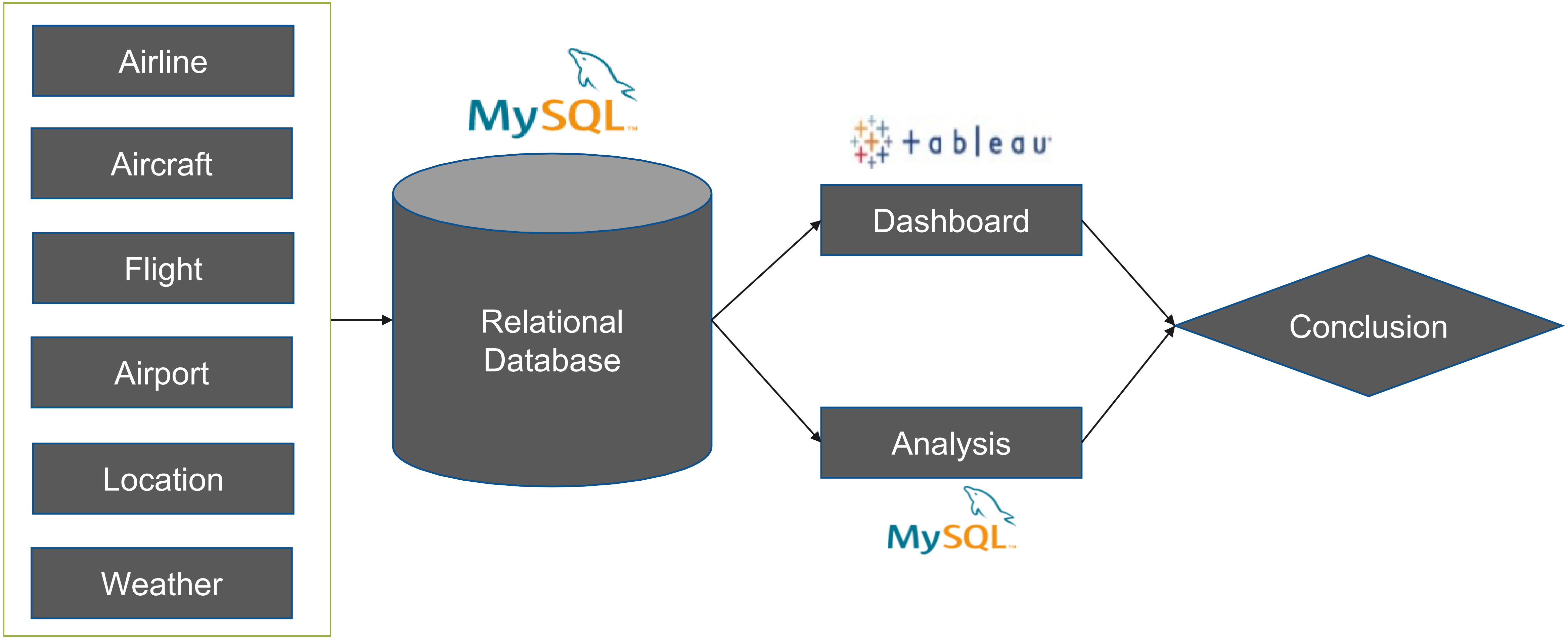
Flight

Airport

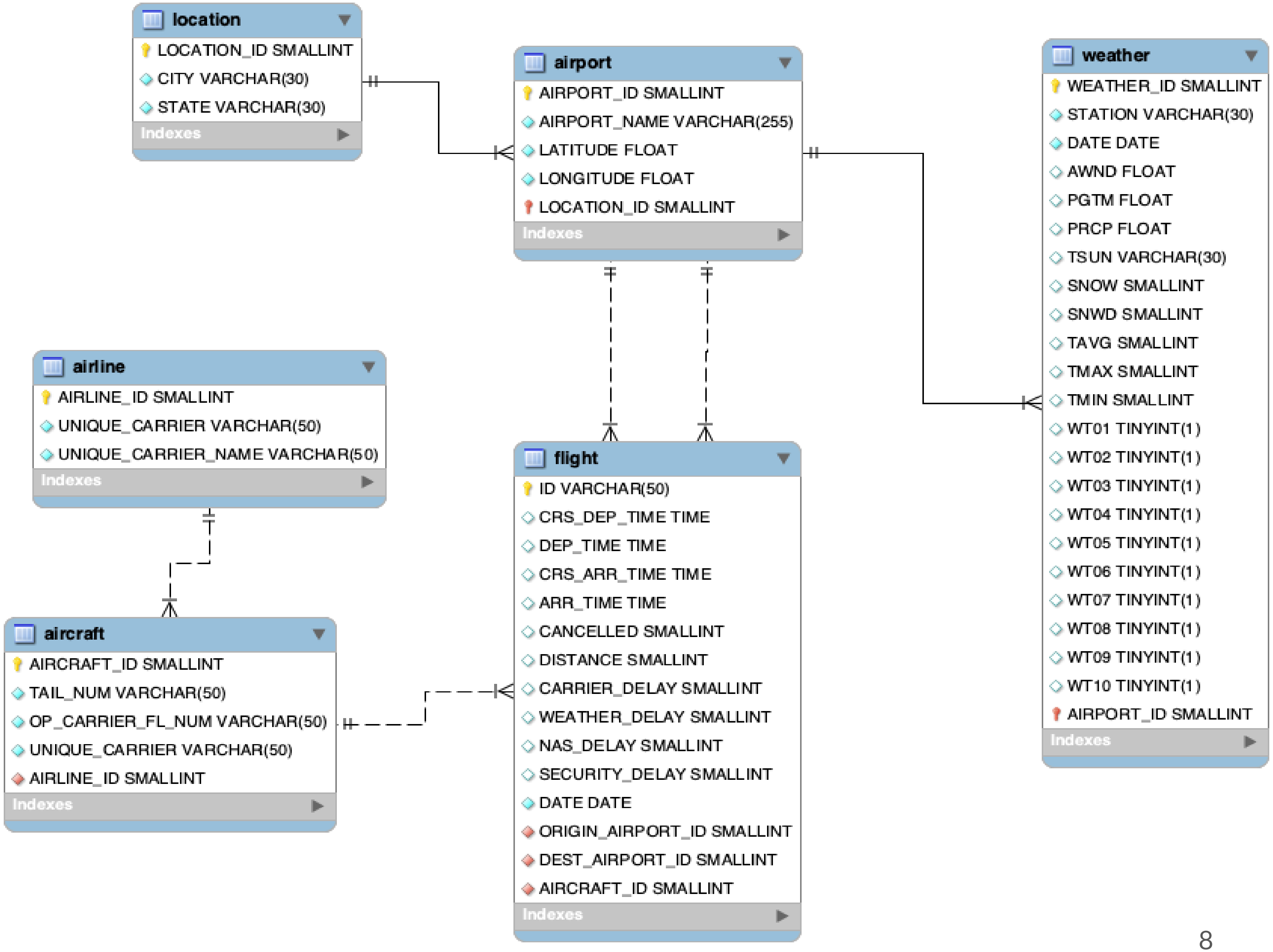
Location

Weather

Data Model: Load



EER Diagram:



location
LOCATION_ID SMALLINT
CITY VARCHAR(30)
STATE VARCHAR(30)
Indexes

airport
AIRPORT_ID SMALLINT
AIRPORT_NAME VARCHAR(255)
LATITUDE FLOAT
LONGITUDE FLOAT
LOCATION_ID SMALLINT
Indexes

weather
WEATHER_ID SMALLINT
STATION VARCHAR(30)
DATE DATE
AWND FLOAT
PGTM FLOAT
PRCP FLOAT
TSUN VARCHAR(30)
SNOW SMALLINT
SNWD SMALLINT
TAVG SMALLINT
TMAX SMALLINT
TMIN SMALLINT
WT01 TINYINT(1)
WT02 TINYINT(1)
WT03 TINYINT(1)
WT04 TINYINT(1)
WT05 TINYINT(1)
WT06 TINYINT(1)
WT07 TINYINT(1)
WT08 TINYINT(1)
WT09 TINYINT(1)
WT10 TINYINT(1)
AIRPORT_ID SMALLINT
Indexes

airline
AIRLINE_ID SMALLINT
UNIQUE_CARRIER VARCHAR(50)
UNIQUE_CARRIER_NAME VARCHAR(50)
Indexes

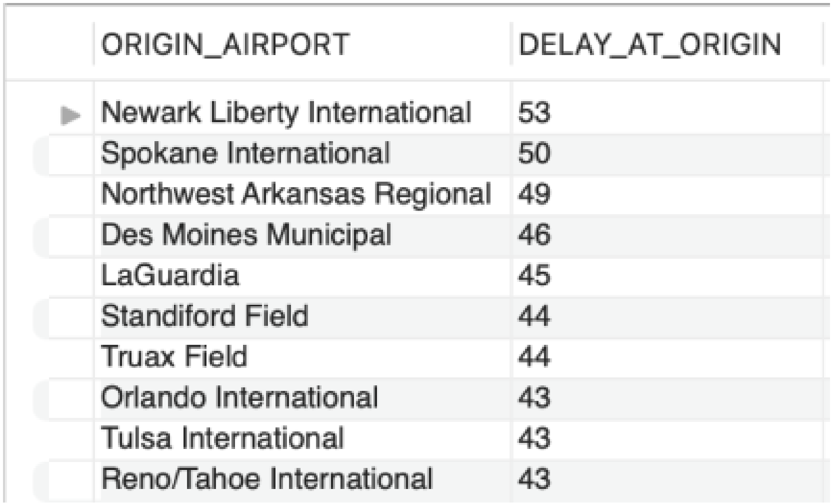
aircraft
AIRCRAFT_ID SMALLINT
TAIL_NUM VARCHAR(50)
OP_CARRIER_FL_NUM VARCHAR(50)
UNIQUE_CARRIER VARCHAR(50)
AIRLINE_ID SMALLINT
Indexes

flight
ID VARCHAR(50)
CRS_DEP_TIME TIME
DEP_TIME TIME
CRS_ARR_TIME TIME
ARR_TIME TIME
CANCELLED SMALLINT
DISTANCE SMALLINT
CARRIER_DELAY SMALLINT
WEATHER_DELAY SMALLINT
NAS_DELAY SMALLINT
SECURITY_DELAY SMALLINT
DATE DATE
ORIGIN_AIRPORT_ID SMALLINT
DEST_AIRPORT_ID SMALLINT
AIRCRAFT_ID SMALLINT
Indexes

Data Analysis: SQL

SQL Query 1: Average delay at Origin Airport

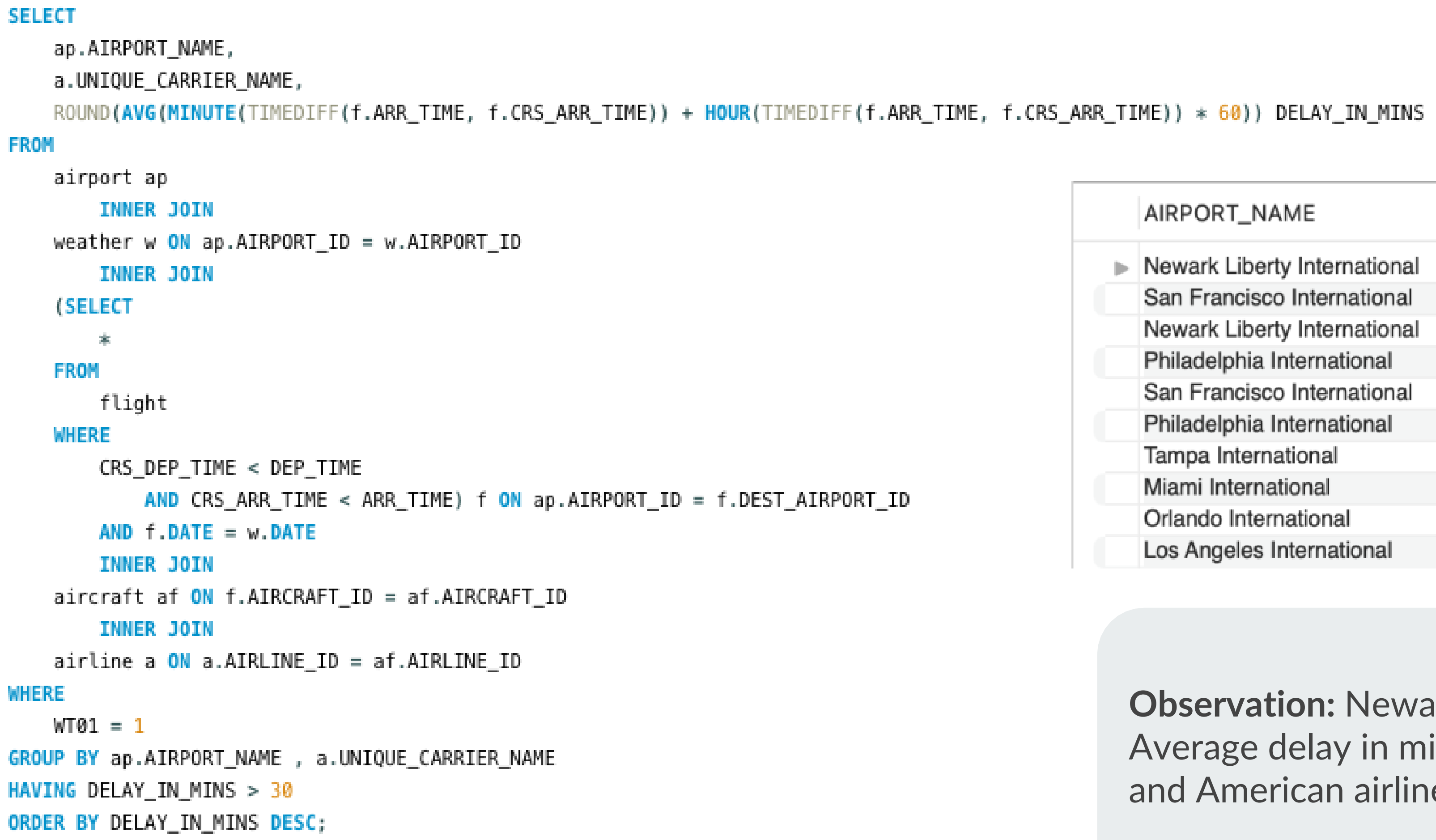
```sql
SELECT
    ap.AIRPORT_NAME as ORIGIN_AIRPORT,
    ROUND(AVG(MINUTE(TIMEDIFF(f.DEP_TIME, f.CRS_DEP_TIME)) + HOUR(TIMEDIFF(f.DEP_TIME, f.CRS_DEP_TIME)) * 60)) AS DELAY_AT_ORIGIN
FROM
    flight f
        INNER JOIN
    aircraft af ON f.AIRCRAFT_ID = af.AIRCRAFT_ID
        INNER JOIN
    airline a ON a.AIRLINE_ID = af.AIRLINE_ID
        INNER JOIN
    airport ap ON f.ORIGIN_AIRPORT_ID = ap.AIRPORT_ID
WHERE
    f.CRS_DEP_TIME < f.DEP_TIME
GROUP BY ap.AIRPORT_NAME
ORDER BY DELAY_AT_ORIGIN DESC;
```



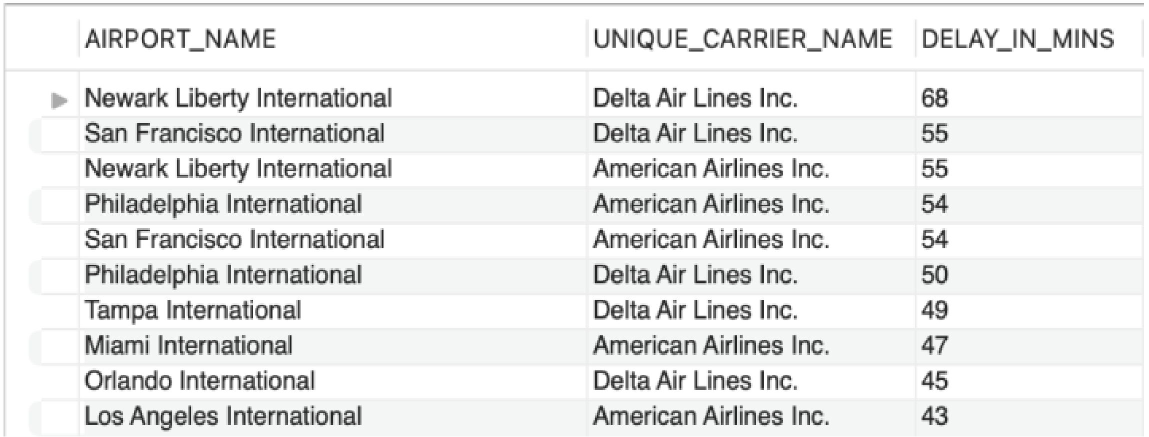
ORIGIN_AIRPORT	DELAY_AT_ORIGIN
Newark Liberty International	53
Spokane International	50
Northwest Arkansas Regional	49
Des Moines Municipal	46
LaGuardia	45
Standiford Field	44
Truax Field	44
Orlando International	43
Tulsa International	43
Reno/Tahoe International	43

Observation: Newark has the highest average delay in terms of minutes, followed by Spokane and Northwest Arkansas

SQL Query 2: Average Arrival delay in Mins due to fog at different airports for various airlines



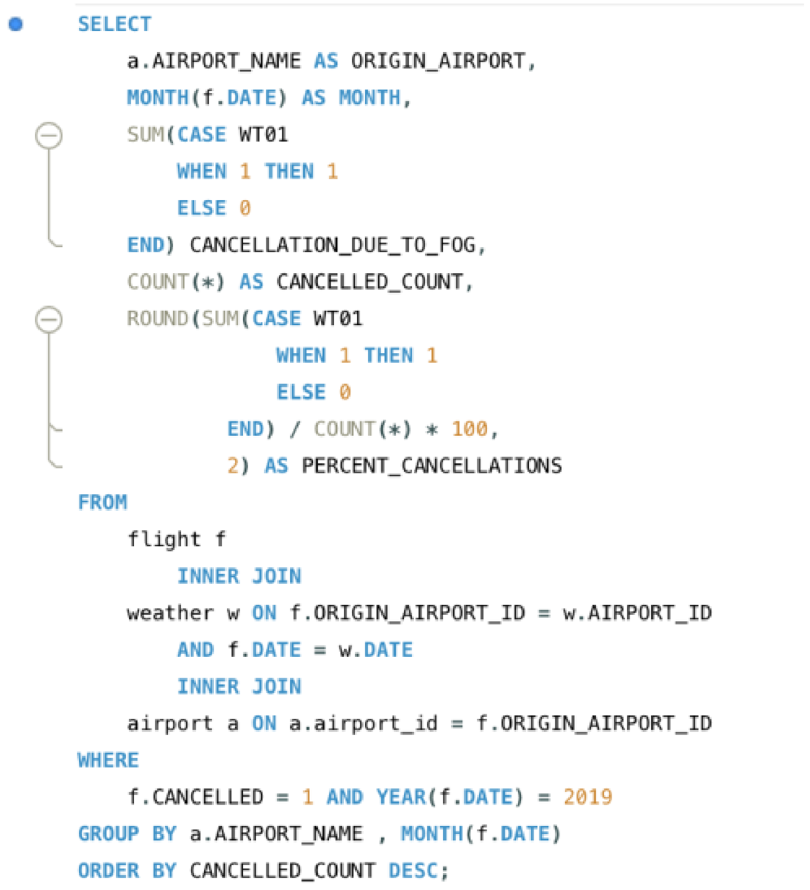
```
SELECT
    ap.AIRPORT_NAME,
    a.UNIQUE_CARRIER_NAME,
    ROUND(AVG(MINUTE(TIMEDIFF(f.ARR_TIME, f.CRS_ARR_TIME)) + HOUR(TIMEDIFF(f.ARR_TIME, f.CRS_ARR_TIME)) * 60)) DELAY_IN_MINS
FROM
    airport ap
        INNER JOIN
    weather w ON ap.AIRPORT_ID = w.AIRPORT_ID
        INNER JOIN
    (SELECT
        *
    FROM
        flight
    WHERE
        CRS_DEP_TIME < DEP_TIME
            AND CRS_ARR_TIME < ARR_TIME) f ON ap.AIRPORT_ID = f.DEST_AIRPORT_ID
        AND f.DATE = w.DATE
        INNER JOIN
    aircraft af ON f.AIRCRAFT_ID = af.AIRCRAFT_ID
        INNER JOIN
    airline a ON a.AIRLINE_ID = af.AIRLINE_ID
WHERE
    WT01 = 1
GROUP BY ap.AIRPORT_NAME , a.UNIQUE_CARRIER_NAME
HAVING DELAY_IN_MINS > 30
ORDER BY DELAY_IN_MINS DESC;
```



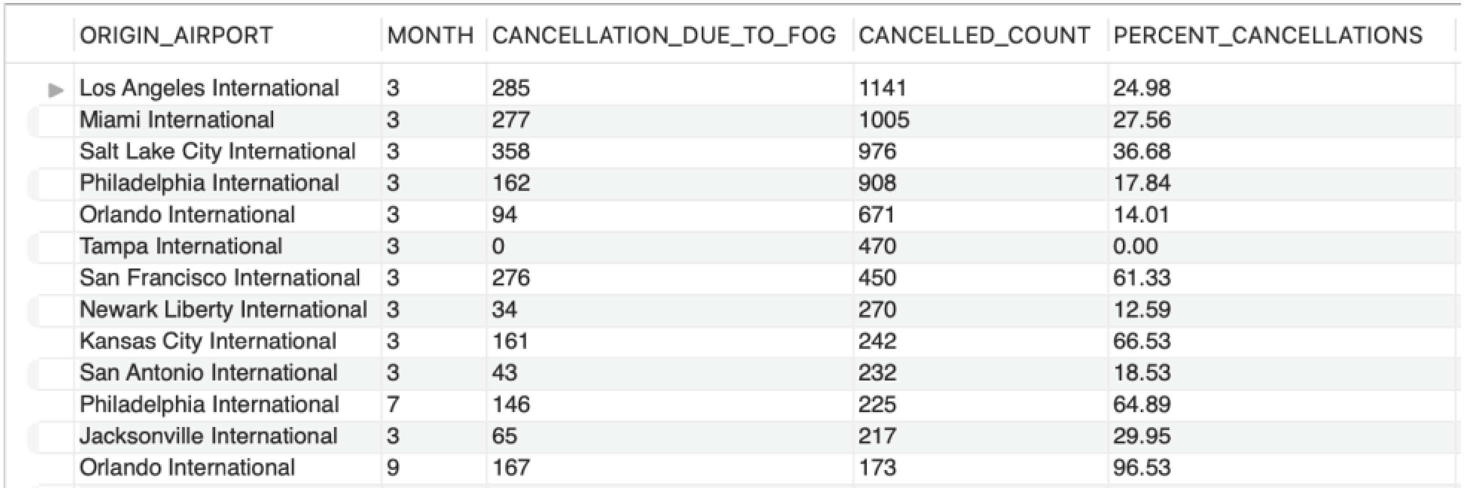
AIRPORT_NAME	UNIQUE_CARRIER_NAME	DELAY_IN_MINS
Newark Liberty International	Delta Air Lines Inc.	68
San Francisco International	Delta Air Lines Inc.	55
Newark Liberty International	American Airlines Inc.	55
Philadelphia International	American Airlines Inc.	54
San Francisco International	American Airlines Inc.	54
Philadelphia International	Delta Air Lines Inc.	50
Tampa International	Delta Air Lines Inc.	49
Miami International	American Airlines Inc.	47
Orlando International	Delta Air Lines Inc.	45
Los Angeles International	American Airlines Inc.	43

Observation: Newark Liberty Airport has the highest Average delay in mins due to heavy fog for both Delta and American airlines.

SQL Query 3: % Monthly cancellations due to fog by airports



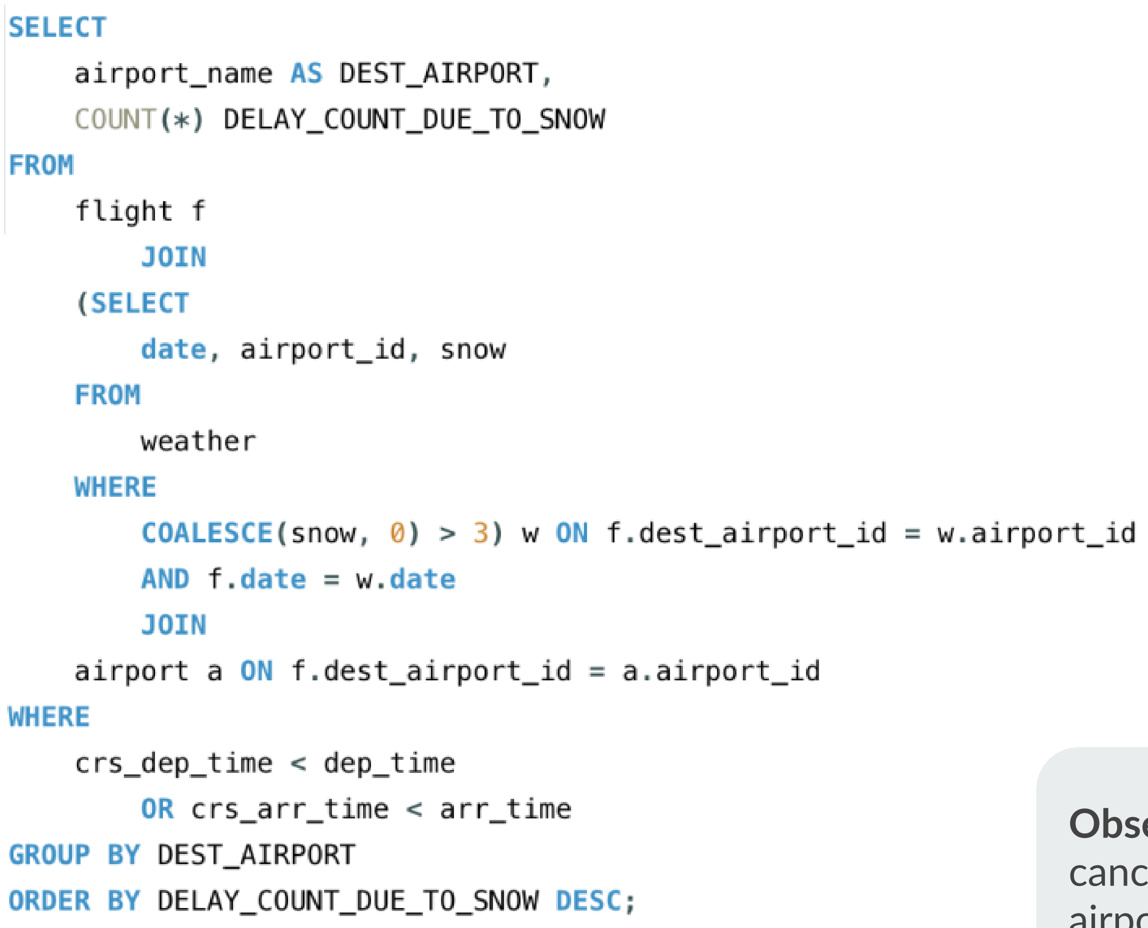
```
SELECT
    a.AIRPORT_NAME AS ORIGIN_AIRPORT,
    MONTH(f.DATE) AS MONTH,
    SUM(CASE WT01
        WHEN 1 THEN 1
        ELSE 0
    END) CANCELLATION_DUE_TO_FOG,
    COUNT(*) AS CANCELLED_COUNT,
    ROUND(SUM(CASE WT01
                WHEN 1 THEN 1
                ELSE 0
            END) / COUNT(*) * 100,
            2) AS PERCENT_CANCELLATIONS
FROM
    flight f
        INNER JOIN
    weather w ON f.ORIGIN_AIRPORT_ID = w.AIRPORT_ID
        AND f.DATE = w.DATE
        INNER JOIN
    airport a ON a.airport_id = f.ORIGIN_AIRPORT_ID
WHERE
    f.CANCELLED = 1 AND YEAR(f.DATE) = 2019
GROUP BY a.AIRPORT_NAME , MONTH(f.DATE)
ORDER BY CANCELLED_COUNT DESC;
```



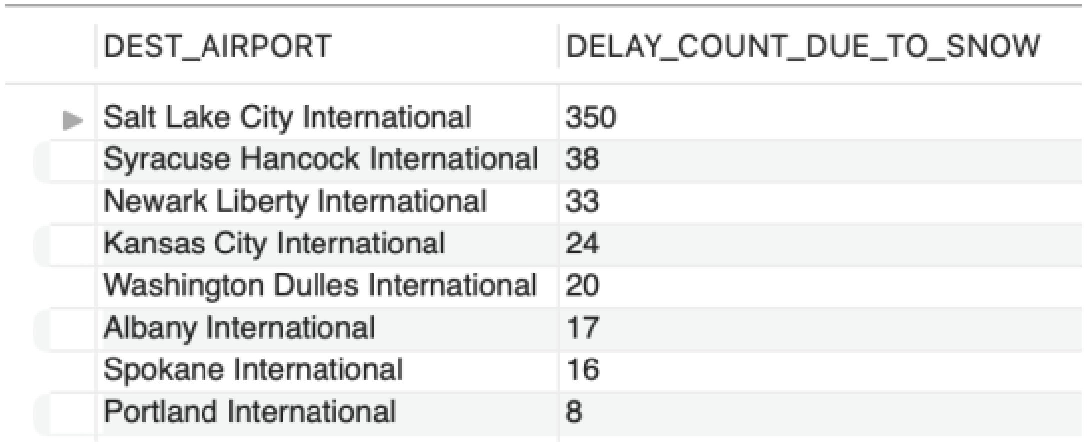
ORIGIN_AIRPORT	MONTH	CANCELLATION_DUE_TO_FOG	CANCELLED_COUNT	PERCENT_CANCELLATIONS
Los Angeles International	3	285	1141	24.98
Miami International	3	277	1005	27.56
Salt Lake City International	3	358	976	36.68
Philadelphia International	3	162	908	17.84
Orlando International	3	94	671	14.01
Tampa International	3	0	470	0.00
San Francisco International	3	276	450	61.33
Newark Liberty International	3	34	270	12.59
Kansas City International	3	161	242	66.53
San Antonio International	3	43	232	18.53
Philadelphia International	7	146	225	64.89
Jacksonville International	3	65	217	29.95
Orlando International	9	167	173	96.53

Observation: March has most flight cancellations for both the airlines.

SQL Query 4: Arrival Delay count due to heavy Snow



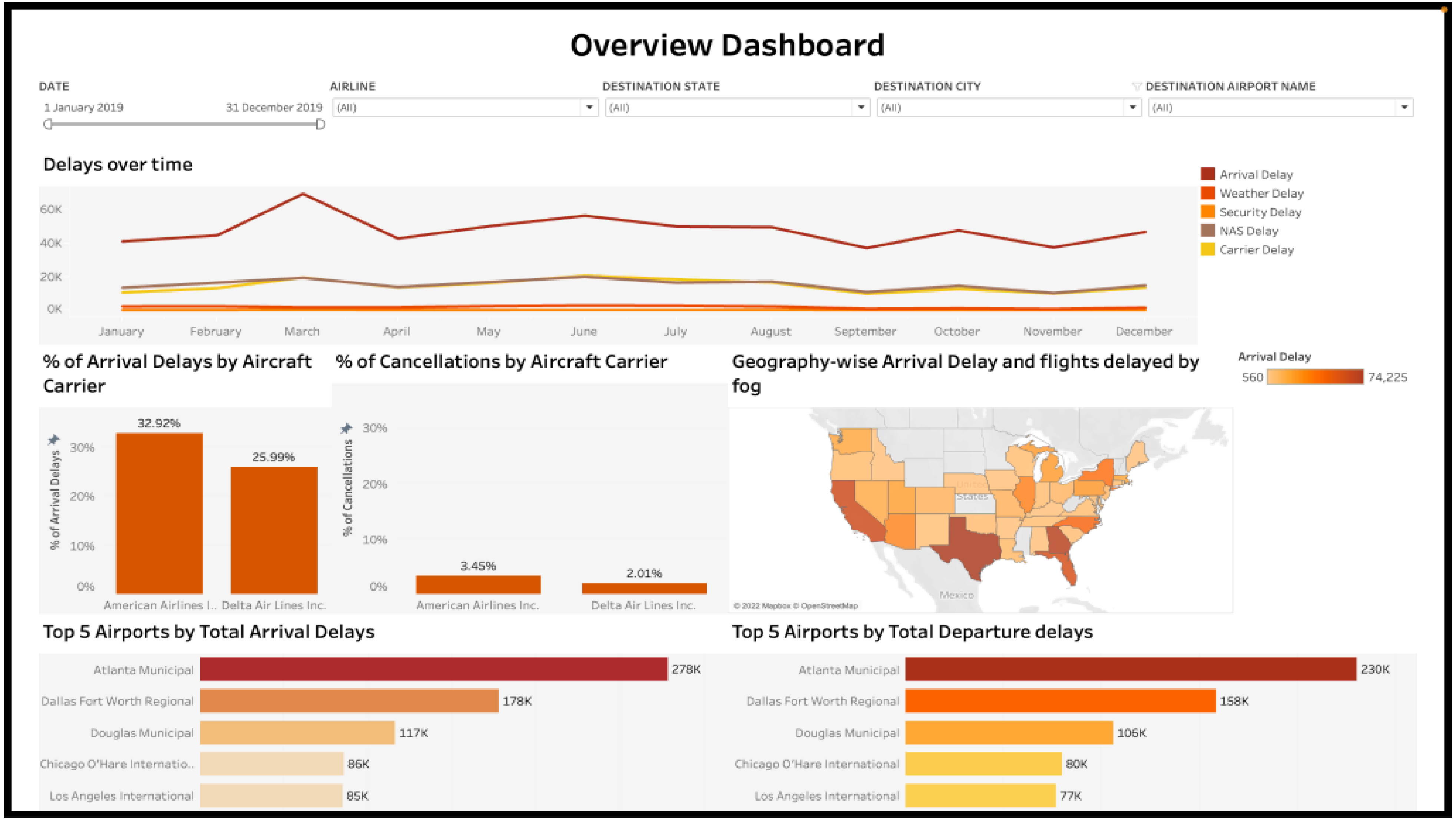
```sql
SELECT
    airport_name AS DEST_AIRPORT,
    COUNT(*) DELAY_COUNT_DUE_TO_SNOW
FROM
    flight f
        JOIN
    (SELECT
        date, airport_id, snow
    FROM
        weather
    WHERE
        COALESCE(snow, 0) > 3) w ON f.dest_airport_id = w.airport_id
        AND f.date = w.date
        JOIN
    airport a ON f.dest_airport_id = a.airport_id
WHERE
    crs_dep_time < dep_time
        OR crs_arr_time < arr_time
GROUP BY DEST_AIRPORT
ORDER BY DELAY_COUNT_DUE_TO_SNOW DESC;
```



DEST_AIRPORT	DELAY_COUNT_DUE_TO_SNOW
Salt Lake City International	350
Syracuse Hancock International	38
Newark Liberty International	33
Kansas City International	24
Washington Dulles International	20
Albany International	17
Spokane International	16
Portland International	8

Observation: Salt Lake city airport has significantly higher cancellations happening due to snow in comparison with other airports.

Data Analysis: Tableau



Overview Dashboard

DATE: 1 January 2019 – 31 December 2019

AIRLINE: (All)

DESTINATION STATE: (All)

DESTINATION CITY: (All)

DESTINATION AIRPORT NAME: (All)

Delays over time

% of Arrival Delays by Aircraft Carrier

% of Cancellations by Aircraft Carrier

Geography-wise Arrival Delay and flights delayed by fog

Top 5 Airports by Total Arrival Delays

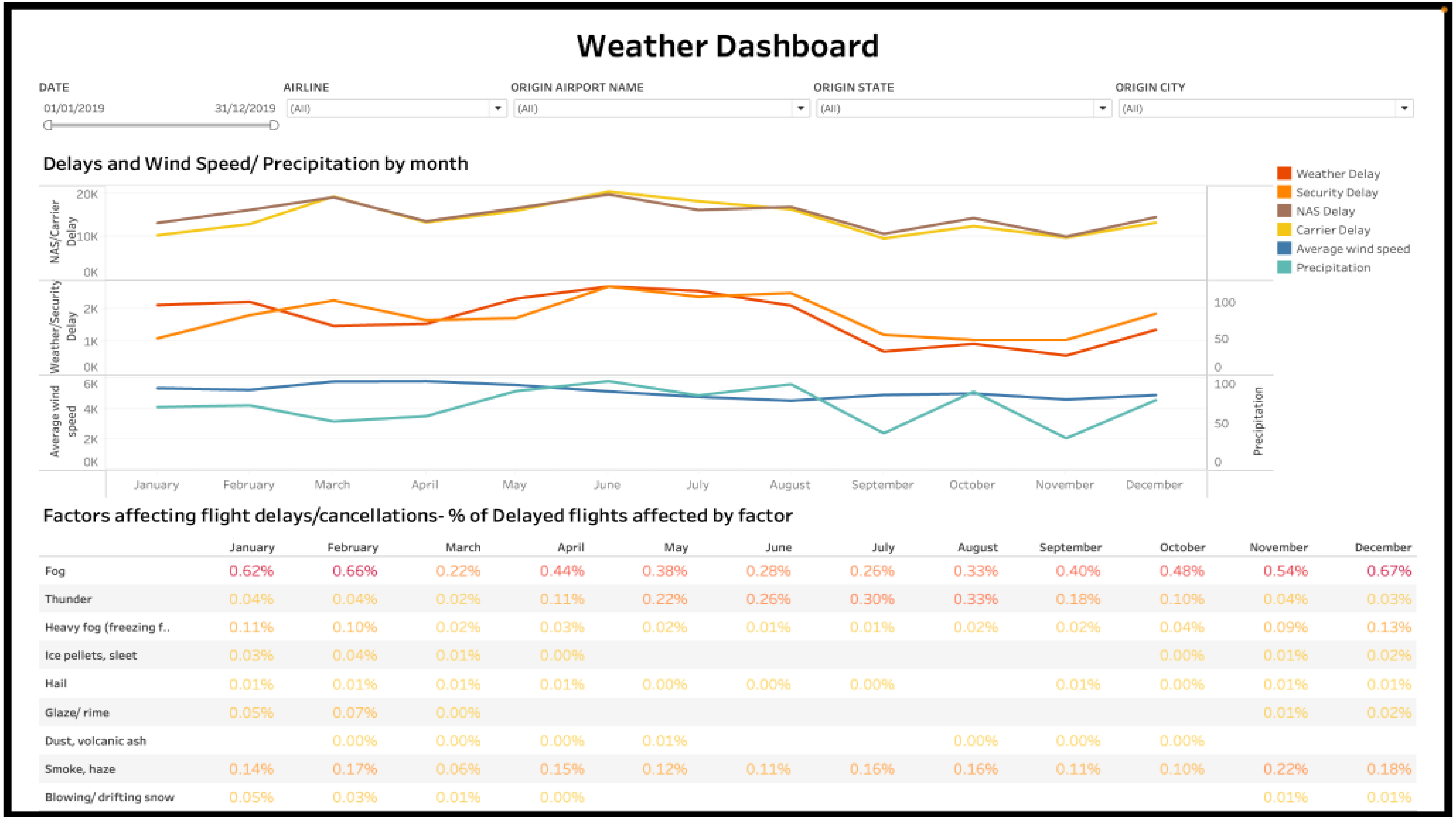
Atlanta Municipal
278K
Dallas Fort Worth Regional
178K
Douglas Municipal
117K
Chicago O'Hare Internatio..
86K
Los Angeles International
85K

Top 5 Airports by Total Departure delays

Atlanta Municipal
230K
Dallas Fort Worth Regional
158K
Douglas Municipal
106K
Chicago O'Hare International
80K
Los Angeles International
77K

Business insights:

- Maximum delays happen in March with NAS and Carrier delays acting as major reasons
- American Airlines has greater proportion of delays and cancellations in comparison to Delta
- Texas, Georgia, Florida and California show the highest arrival delays



Weather Dashboard

DATE: 01/01/2019 – 31/12/2019

AIRLINE: (All)

ORIGIN AIRPORT NAME: (All)

ORIGIN STATE: (All)

ORIGIN CITY: (All)

Delays and Wind Speed/ Precipitation by month

Factors affecting flight delays/cancellations- % of Delayed flights affected by factor

	January	February	March	April	May	June	July	August	September	October	November	December
Fog	0.62%	0.66%	0.22%	0.44%	0.38%	0.28%	0.26%	0.33%	0.40%	0.48%	0.54%	0.67%
Thunder	0.04%	0.04%	0.02%	0.11%	0.22%	0.26%	0.30%	0.33%	0.18%	0.10%	0.04%	0.03%
Heavy fog (freezing f..	0.11%	0.10%	0.02%	0.03%	0.02%	0.01%	0.01%	0.02%	0.02%	0.04%	0.09%	0.13%
Ice pellets, sleet	0.03%	0.04%	0.01%	0.00%						0.00%	0.01%	0.02%
Hail	0.01%	0.01%	0.01%	0.01%	0.00%	0.00%	0.00%		0.01%	0.00%	0.01%	0.01%
Glaze/ rime	0.05%	0.07%	0.00%								0.01%	0.02%
Dust, volcanic ash		0.00%	0.00%	0.00%	0.01%			0.00%	0.00%	0.00%		
Smoke, haze	0.14%	0.17%	0.06%	0.15%	0.12%	0.11%	0.16%	0.16%	0.11%	0.10%	0.22%	0.18%
Blowing/ drifting snow	0.05%	0.03%	0.01%	0.00%							0.01%	0.01%

Business insights:

- Precipitation contributes to weather delays for the month of June and August
- Out of the various weather conditions affecting flight schedules, Fog, Thunder and Smoke/Haze cause the most amount of delays.

Conclusion

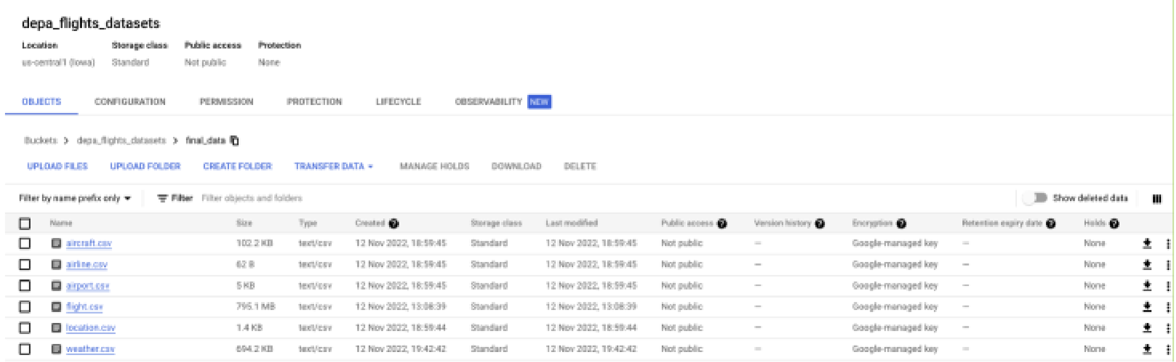
Challenges faced

Problem -

Due to the size of the dataset, inline function for inserting data for entity ‘flight’ was having time out issues.

Solution -

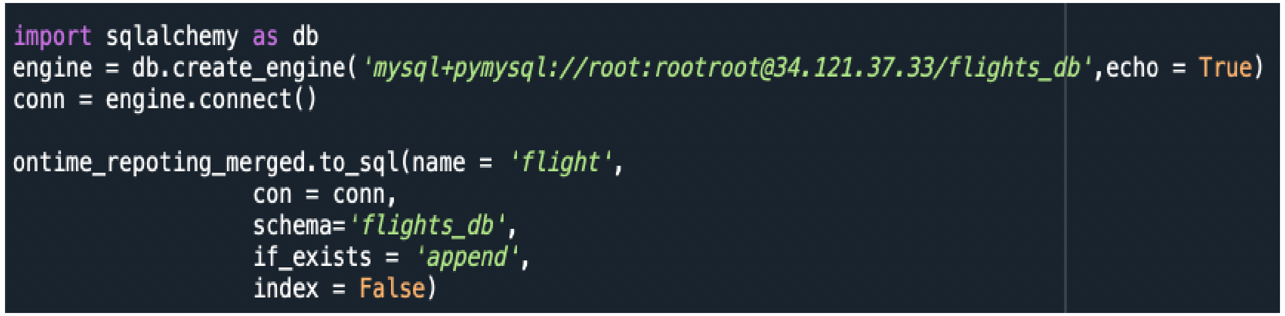
a) Filtered data to include only two airlines - American Airline and Delta Airline.
b) Transferred data to GCP Bucket and then used Data Import feature in Cloud MySQL to insert data into the database.
c) Created SQL pointer in Python to process and insert data.



depa_flights_datasets

Location	Storage class	Public access	Protection
us-central1 (Iowa)	Standard	Not public	None

□	Name	Size	Type	Created	Storage class	Last modified	Public access	Version history	Encryption	Retention expiry date	Holds
□	aircraft.csv	102.2 KB	text/csv	12 Nov 2022, 18:59:45	Standard	12 Nov 2022, 18:59:45	Not public	—	Google-managed key	—	None
□	airline.csv	62 B	text/csv	12 Nov 2022, 18:59:45	Standard	12 Nov 2022, 18:59:45	Not public	—	Google-managed key	—	None
□	airport.csv	5 KB	text/csv	12 Nov 2022, 18:59:45	Standard	12 Nov 2022, 18:59:45	Not public	—	Google-managed key	—	None
□	flight.csv	795.1 MB	text/csv	12 Nov 2022, 13:08:39	Standard	12 Nov 2022, 13:08:39	Not public	—	Google-managed key	—	None
□	location.csv	1.4 KB	text/csv	12 Nov 2022, 18:59:44	Standard	12 Nov 2022, 18:59:44	Not public	—	Google-managed key	—	None
□	weather.csv	694.2 KB	text/csv	12 Nov 2022, 19:42:42	Standard	12 Nov 2022, 19:42:42	Not public	—	Google-managed key	—	None



```
import sqlalchemy as db
engine = db.create_engine('mysql+pymysql://root:rootroot@34.121.37.33/flights_db',echo = True)
conn = engine.connect()

ontime_repoting_merged.to_sql(name = 'flight',
                    con = conn,
                    schema='flights_db',
                    if_exists = 'append',
                    index = False)
```

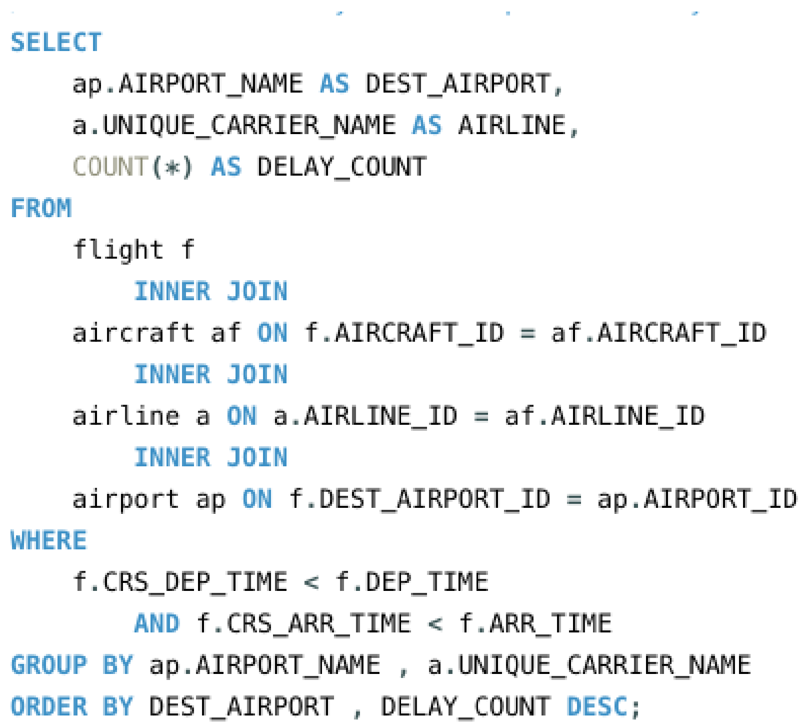
Recommendations/Future Scope

- Extend data to include more years of historical data
- Extend scope to include all airlines (for now, the analysis has been done on only American and Delta Airlines)
- Incorporating prediction model to forecast weather conditions based on previous delays due to weather such that costs could be minimized in similar situations
- Incorporating customer data and feedback, and flight revenue data to estimate implicit cost incurred due to the delays and cancellations by the airline

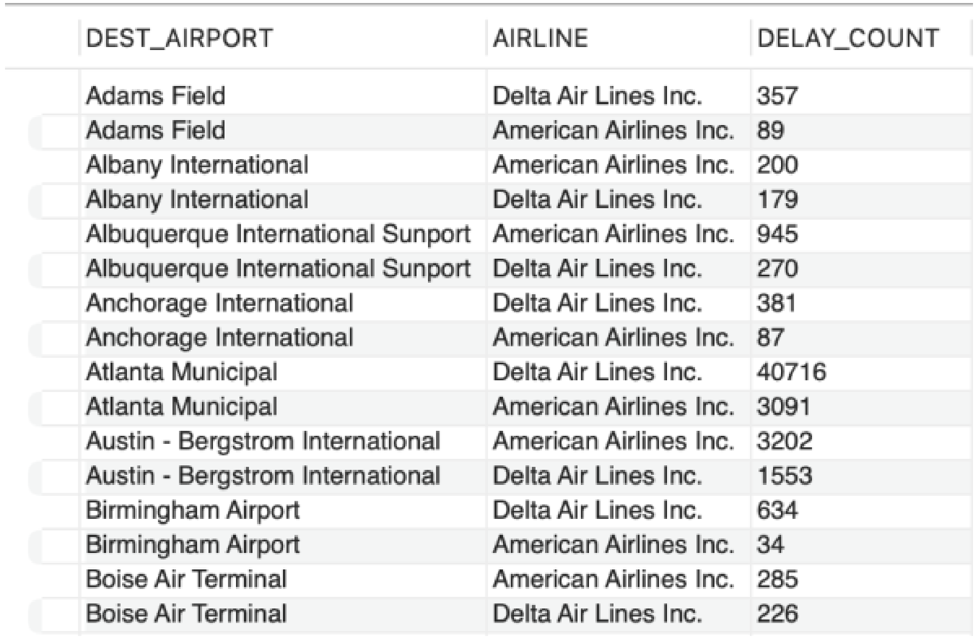
Thank You

Appendix

SQL Query 5: Arrival delay due to departure delay at Destination grouped by airline



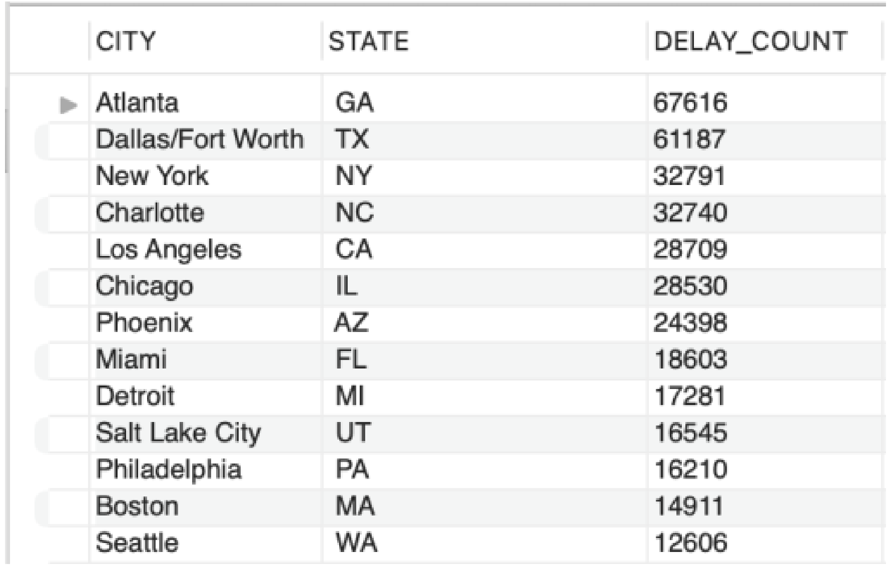
```
SELECT
    ap.AIRPORT_NAME AS DEST_AIRPORT,
    a.UNIQUE_CARRIER_NAME AS AIRLINE,
    COUNT(*) AS DELAY_COUNT
FROM
    flight f
        INNER JOIN
    aircraft af ON f.AIRCRAFT_ID = af.AIRCRAFT_ID
        INNER JOIN
    airline a ON a.AIRLINE_ID = af.AIRLINE_ID
        INNER JOIN
    airport ap ON f.DEST_AIRPORT_ID = ap.AIRPORT_ID
WHERE
    f.CRS_DEP_TIME < f.DEP_TIME
        AND f.CRS_ARR_TIME < f.ARR_TIME
GROUP BY ap.AIRPORT_NAME , a.UNIQUE_CARRIER_NAME
ORDER BY DEST_AIRPORT , DELAY_COUNT DESC;
```



DEST_AIRPORT	AIRLINE	DELAY_COUNT
Adams Field	Delta Air Lines Inc.	357
Adams Field	American Airlines Inc.	89
Albany International	American Airlines Inc.	200
Albany International	Delta Air Lines Inc.	179
Albuquerque International Sunport	American Airlines Inc.	945
Albuquerque International Sunport	Delta Air Lines Inc.	270
Anchorage International	Delta Air Lines Inc.	381
Anchorage International	American Airlines Inc.	87
Atlanta Municipal	Delta Air Lines Inc.	40716
Atlanta Municipal	American Airlines Inc.	3091
Austin - Bergstrom International	American Airlines Inc.	3202
Austin - Bergstrom International	Delta Air Lines Inc.	1553
Birmingham Airport	Delta Air Lines Inc.	634
Birmingham Airport	American Airlines Inc.	34
Boise Air Terminal	American Airlines Inc.	285
Boise Air Terminal	Delta Air Lines Inc.	226

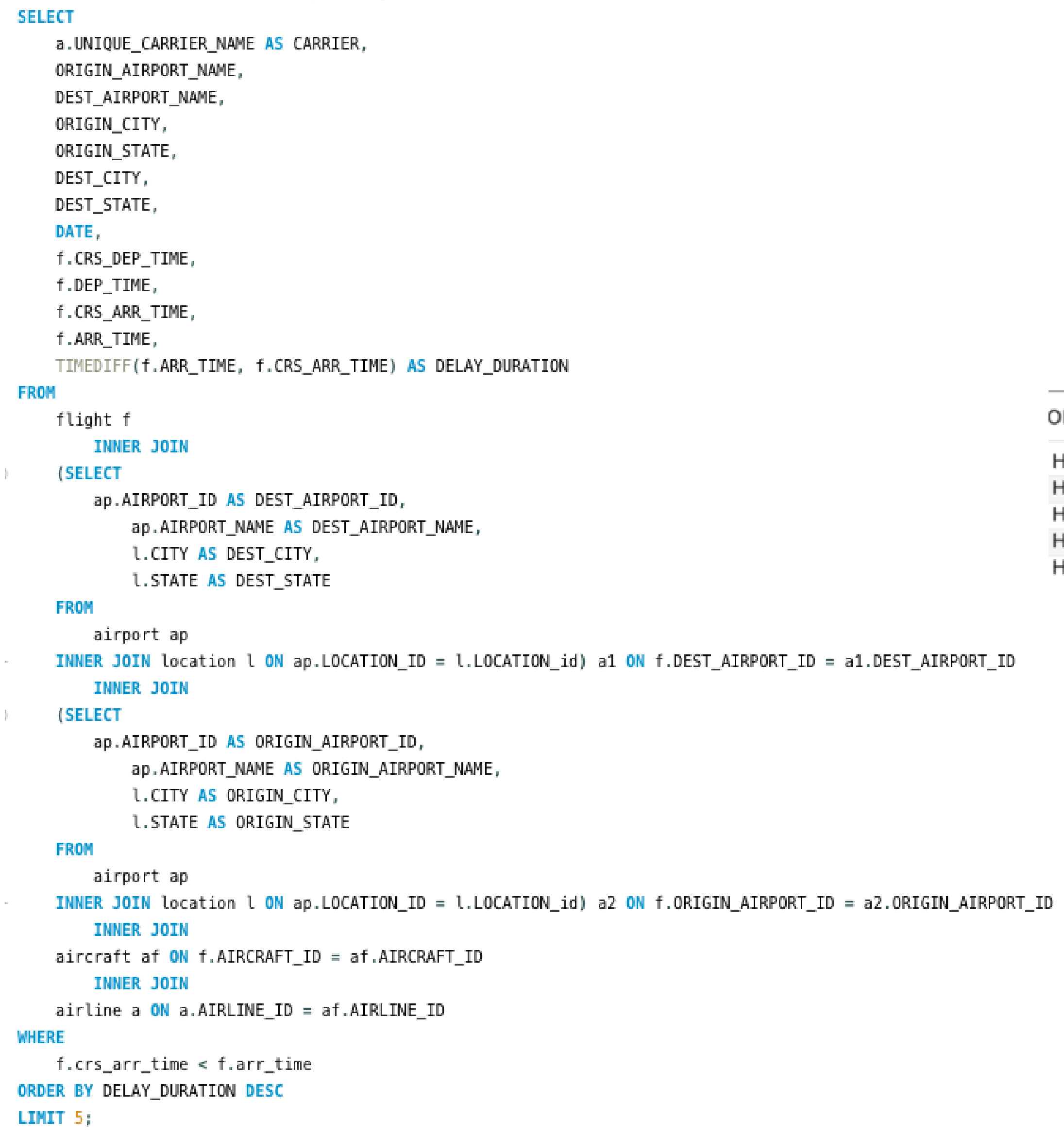
SQL Query 6: Arrival delay count by city and state

```
SELECT
    l.CITY, l.STATE, COUNT(*) AS DELAY_COUNT
FROM
    flight f
        INNER JOIN
    airport ap ON f.DEST_AIRPORT_ID = ap.AIRPORT_ID
        INNER JOIN
    location l ON ap.LOCATION_ID = l.LOCATION_ID
WHERE
    f.CRS_ARR_TIME < f.ARR_TIME
GROUP BY l.CITY , l.STATE
HAVING DELAY_COUNT > 999
ORDER BY DELAY_COUNT DESC;
```

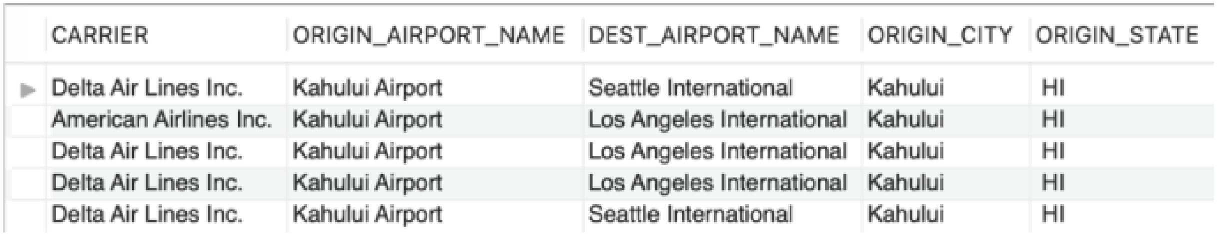


CITY	STATE	DELAY_COUNT
Atlanta	GA	67616
Dallas/Fort Worth	TX	61187
New York	NY	32791
Charlotte	NC	32740
Los Angeles	CA	28709
Chicago	IL	28530
Phoenix	AZ	24398
Miami	FL	18603
Detroit	MI	17281
Salt Lake City	UT	16545
Philadelphia	PA	16210
Boston	MA	14911
Seattle	WA	12606

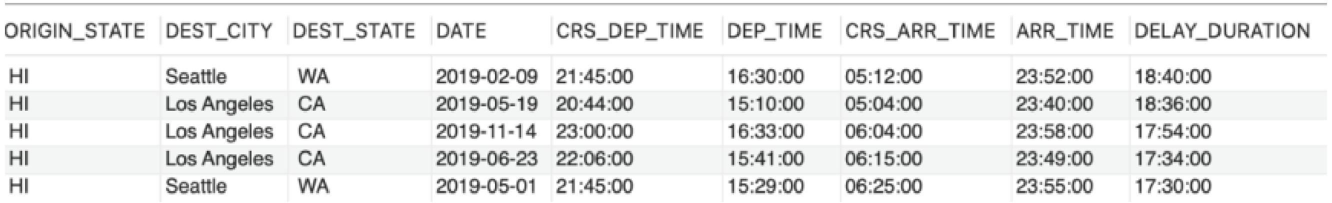
SQL Query 7: Top 5 delayed flights by minutes



```
SELECT
    a.UNIQUE_CARRIER_NAME AS CARRIER,
    ORIGIN_AIRPORT_NAME,
    DEST_AIRPORT_NAME,
    ORIGIN_CITY,
    ORIGIN_STATE,
    DEST_CITY,
    DEST_STATE,
    DATE,
    f.CRS_DEP_TIME,
    f.DEP_TIME,
    f.CRS_ARR_TIME,
    f.ARR_TIME,
    TIMEDIFF(f.ARR_TIME, f.CRS_ARR_TIME) AS DELAY_DURATION
FROM
    flight f
        INNER JOIN
    (SELECT
        ap.AIRPORT_ID AS DEST_AIRPORT_ID,
            ap.AIRPORT_NAME AS DEST_AIRPORT_NAME,
            l.CITY AS DEST_CITY,
            l.STATE AS DEST_STATE
    FROM
        airport ap
    INNER JOIN location l ON ap.LOCATION_ID = l.LOCATION_id) a1 ON f.DEST_AIRPORT_ID = a1.DEST_AIRPORT_ID
        INNER JOIN
    (SELECT
        ap.AIRPORT_ID AS ORIGIN_AIRPORT_ID,
            ap.AIRPORT_NAME AS ORIGIN_AIRPORT_NAME,
            l.CITY AS ORIGIN_CITY,
            l.STATE AS ORIGIN_STATE
    FROM
        airport ap
    INNER JOIN location l ON ap.LOCATION_ID = l.LOCATION_id) a2 ON f.ORIGIN_AIRPORT_ID = a2.ORIGIN_AIRPORT_ID
        INNER JOIN
    aircraft af ON f.AIRCRAFT_ID = af.AIRCRAFT_ID
        INNER JOIN
    airline a ON a.AIRLINE_ID = af.AIRLINE_ID
WHERE
    f.crs_arr_time < f.arr_time
ORDER BY DELAY_DURATION DESC
LIMIT 5;
```



	CARRIER	ORIGIN_AIRPORT_NAME	DEST_AIRPORT_NAME	ORIGIN_CITY	ORIGIN_STATE
▶	Delta Air Lines Inc.	Kahului Airport	Seattle International	Kahului	HI
	American Airlines Inc.	Kahului Airport	Los Angeles International	Kahului	HI
	Delta Air Lines Inc.	Kahului Airport	Los Angeles International	Kahului	HI
	Delta Air Lines Inc.	Kahului Airport	Los Angeles International	Kahului	HI
	Delta Air Lines Inc.	Kahului Airport	Seattle International	Kahului	HI



ORIGIN_STATE	DEST_CITY	DEST_STATE	DATE	CRS_DEP_TIME	DEP_TIME	CRS_ARR_TIME	ARR_TIME	DELAY_DURATION
HI	Seattle	WA	2019-02-09	21:45:00	16:30:00	05:12:00	23:52:00	18:40:00
HI	Los Angeles	CA	2019-05-19	20:44:00	15:10:00	05:04:00	23:40:00	18:36:00
HI	Los Angeles	CA	2019-11-14	23:00:00	16:33:00	06:04:00	23:58:00	17:54:00
HI	Los Angeles	CA	2019-06-23	22:06:00	15:41:00	06:15:00	23:49:00	17:34:00
HI	Seattle	WA	2019-05-01	21:45:00	15:29:00	06:25:00	23:55:00	17:30:00