BUSINESS DATA MANAGEMENT FINAL PROJECT

Professor Behnaz Bojd

Chia Jo Chen (Tina), Tzu Ying Yu (Winnie),
Biying Han (Candice), Yu Hsin Lee (Kathy)

PROJECT OVERVIEW

01 STAGE I

Business Background
Description of Entities & Attributes
ERD

02 STAGE II

Relational Schema
As-is Dependency Diagram
3rd Normal Form

03 STAGE III

SQL Queries
Analysis

04 CONCLUSION

Key Takeaways

01 STAGE I

Business Background
Description of Entities & Attributes
ERD

BUSINESS BACKGROUND

We want to know more about the flight schedules from different airlines. Most people took a long vacation during Christmas and New Year, we want to see how holiday influences the flights. In this project, we will only analyze data in December 2015.

The dataset was collected from Kaggle.

DESCRIPTION OF ENTITIES & ATTRIBUTES

ENTITIES

- AIRLINE
- AIRCRAFT
- AIRPORT
- FLIGHT SCHEDULE - Scheduled flights
- ACTUAL FLIGHT - Actual record of flights
- DELAY （weak entity)
- CANCEL (weak entity)

ER DIAGRAM

AIRLINE
Airline Code
Name
Country
Number of Employees

AIRCRAFT
Tail Number
Capacity
Seat Class
Model

AIRPORT
Airport Code
Name
Location
Category

FLIGHT SCHEDULE
Flight ID (Flight Number, Date)
Scheduled Departure Time
Scheduled Arrival Time
Distance
Origin Airport
Destination Airport
Air Time

ACTUAL FLIGHT
Actual Flight ID (Flight Number, Date)
Actual Departure Time
Actual Arrival Time
Origin Airport
Destination Airport
Status

DELAY
Delay ID (Flight Number, Date)
Actual Departure Time
Departure Delay Time
Actual Arrival Time
Arrival Delay Time
{Delay Reason}

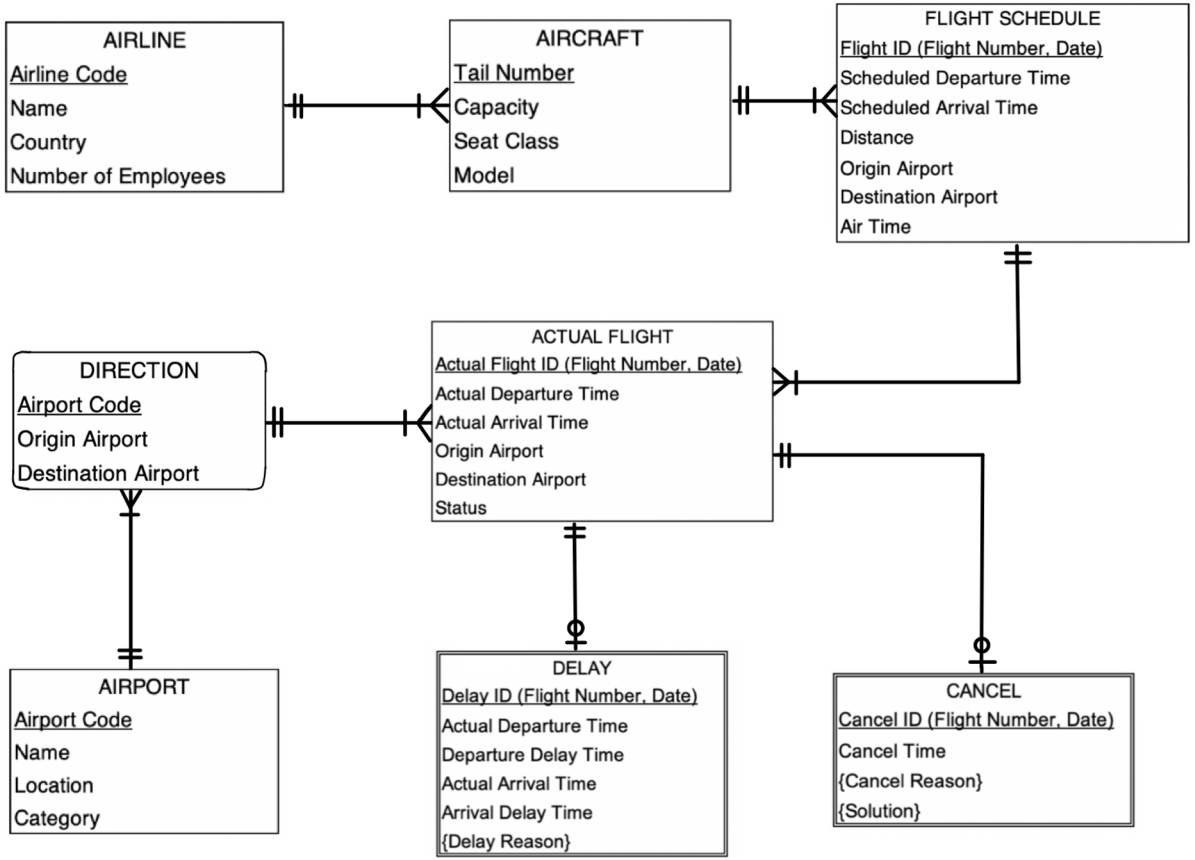
CANCEL
Cancel ID (Flight Number, Date)
Cancel Time
{Cancel Reason}
{Solution}

02

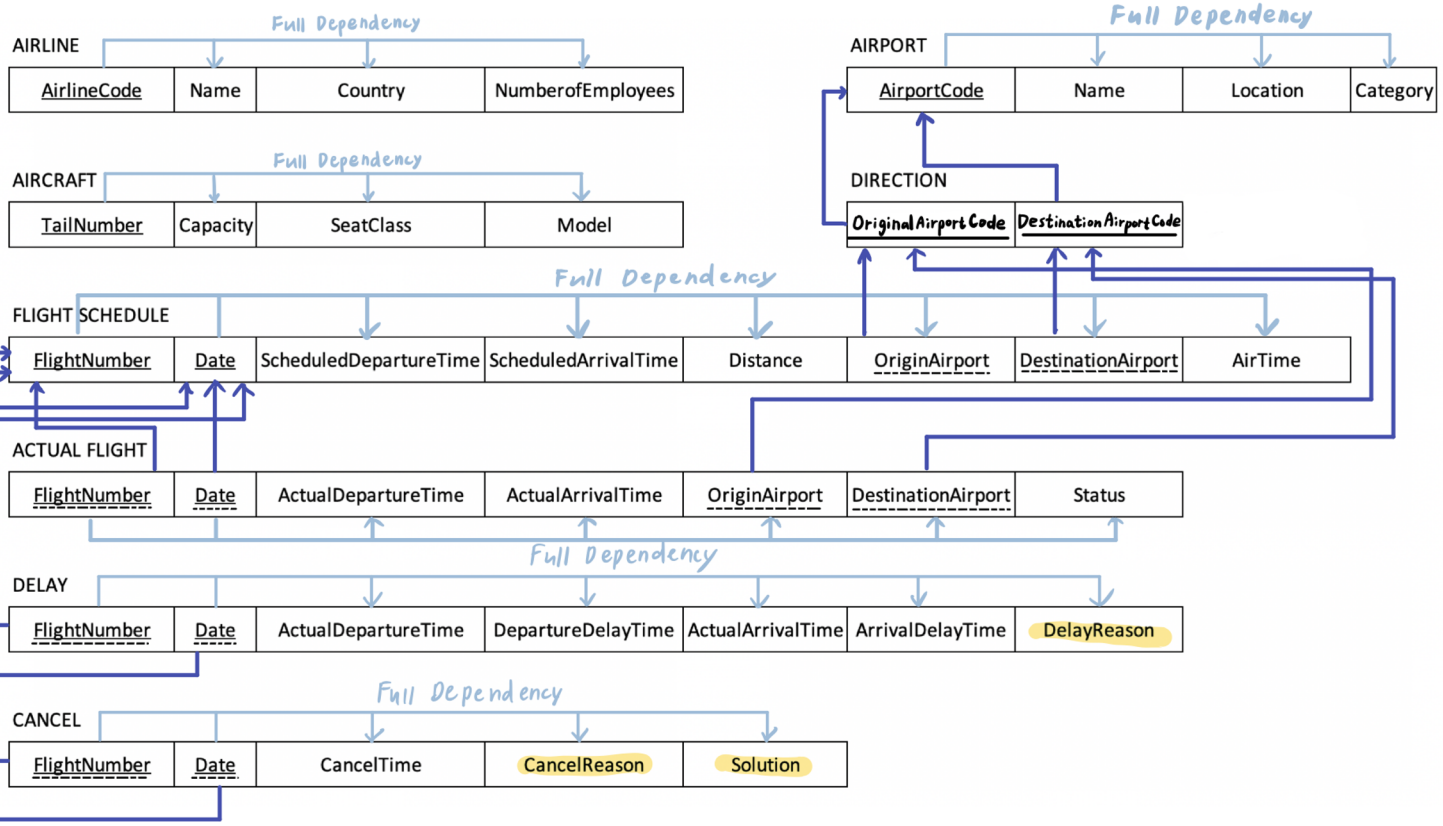
STAGE II

Relational Schema
As-is Dependency Diagram
3rd Normal Form

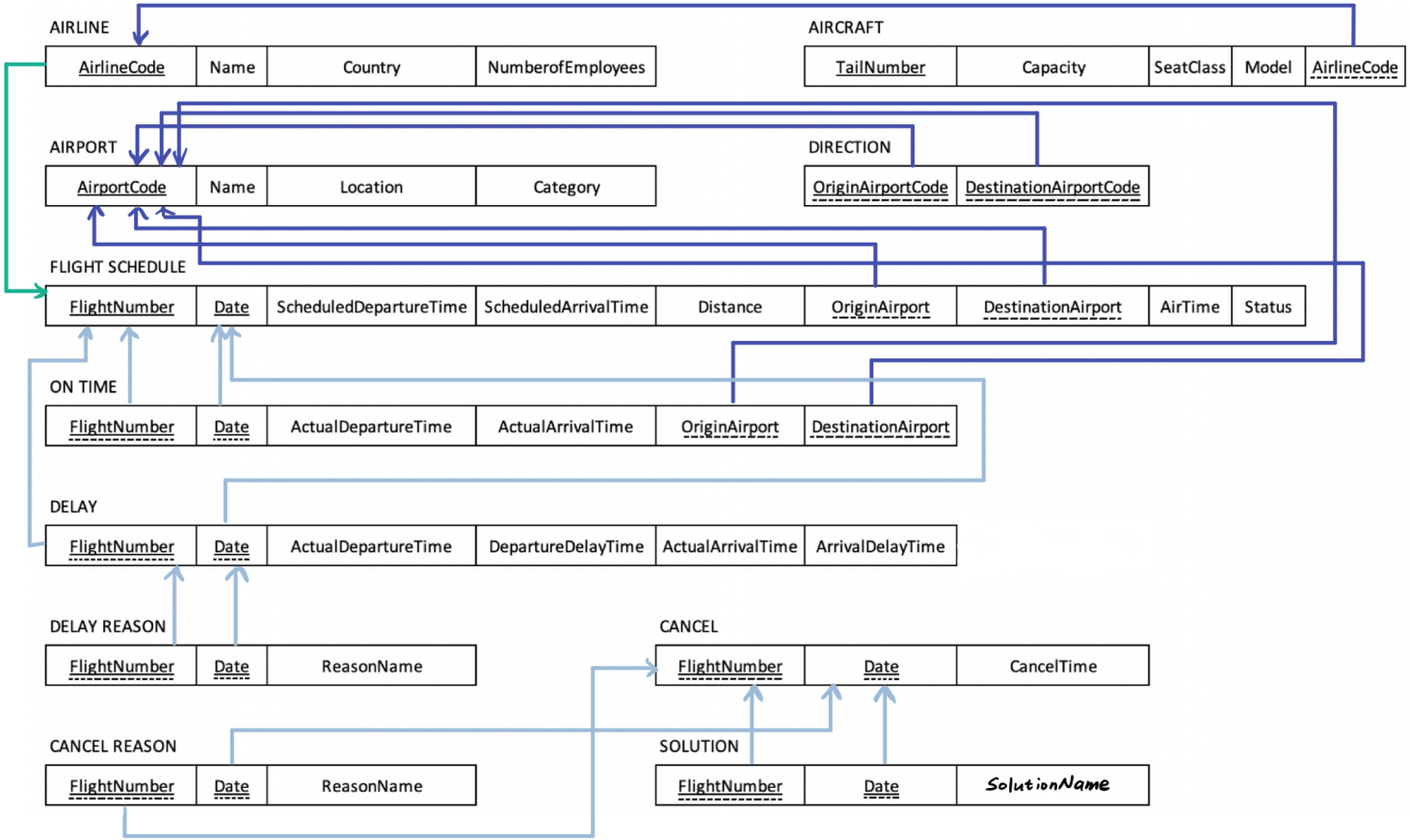
RELATIONAL SCHEMA



AS-IS DEPENDENCY DIAGRAM



3RD NORMAL FORM



03 STAGE III

SQL Queries
Analysis and Findings

SQL QUERIES

How many airports are in each state?

```
SELECT STATE, COUNT(STATE) AS StateCount
FROM `data-mgnt-project.flights_data.flights_dec` JOIN `data-mgnt-project.flights_data.airports`
ON `data-mgnt-project.flights_data.flights_dec`.ORIGIN_AIRPORT=`data-mgnt-project.flights_data.airports`.IATA_CODE
GROUP BY STATE
ORDER BY COUNT(STATE) DESC;
```

STATE	StateCount
CA	58083
TX	55292
FL	40832
IL	34028
GA	32121

STATE	StateCount
NY	21039
CO	20035
AZ	15028
NC	13312
NV	13232

SQL QUERIES

Which day in December has more flights?

```
SELECT DAY, COUNT(FLIGHT_NUMBER) AS FlightCount
FROM `data-mgnt-project.flights_data.flights_dec`
GROUP BY DAY
ORDER BY COUNT(FLIGHT_NUMBER) DESC;
```

DAY	FlightCount
18	16710
17	16579
27	16509
23	16481
28	16312
21	16306
3	16278
11	16263
22	16261
30	16260
10	16259
4	16215
29	16199
16	16100
14	16056

SQL QUERIES

How many flights are cancelled in December?

```
# how many flights are cancelled in December
SELECT COUNT(CANCELLED) AS CountofCancelled
FROM `data-mgnt-project.flights_data.flights_dec`
WHERE CANCELLED=1;
```

Row	CountofCancelled
1	8063

SQL QUERIES

Most popular cancel reason

```
SELECT CANCELLATION_REASON, COUNT(CANCELLATION_REASON) AS Counts
FROM `data-mgnt-project.flights_data.flights_dec`
WHERE CANCELLED=1
GROUP BY CANCELLATION_REASON
ORDER BY COUNT(CANCELLATION_REASON) DESC
LIMIT 1;
```

CANCELLATION_REASON	Counts
B Weather	5613

SQL QUERIES

Which day of the week have highest amount of cancellations?

```
SELECT DAY_OF_WEEK, COUNT(DAY_OF_WEEK) AS CancelledCount
FROM `data-mgnt-project.flights_data.flights_dec`
WHERE CANCELLED=1
GROUP BY DAY_OF_WEEK
ORDER BY COUNT(DAY_OF_WEEK) DESC;
```

DAY_OF_WEEK	CancelledCount
1	2743
2	1496
7	1444
3	744
6	665
4	603
5	368

SQL QUERIES

Number of delays by Airline

```sql
SELECT AIRLINE, COUNT(AIRLINE) AS Counts
FROM `data-mgnt-project.flights_data.flights_dec`
WHERE ARRIVAL_DELAY > 0 OR DEPARTURE_DELAY > 0
GROUP BY AIRLINE
ORDER BY Counts DESC;
```

AIRLINE	Counts
WN	60033
AA	36017
DL	30867
OO	23515
UA	23086
EV	18186
B6	13635
MQ	7056
AS	6162
NK	6103
F9	3953
VX	3290
HA	216[illegible]

SQL QUERIES

Number of delays by air route

```
SELECT ORIGIN_AIRPORT, DESTINATION_AIRPORT, COUNT(ORIGIN_AIRPORT) AS Counts
FROM `data-mgnt-project.flights_data.flights_dec`
WHERE ARRIVAL_DELAY > 0 OR DEPARTURE_DELAY > 0
GROUP BY ORIGIN_AIRPORT, DESTINATION_AIRPORT
ORDER BY Counts DESC;
```

ORIGIN_AIRPORT	DESTINATION_AIRF	Counts
SFO	LAX	827
LAX	SFO	822
LAX	JFK	569
LAS	LAX	561
JFK	LAX	546
LAX	LAS	504
ORD	SFO	497
LAS	SFO	456
ORD	LAX	451
ATL	MCO	429
SEA	LAX	416
LAX	SEA	407
LAX	PHX	406
ATL	FLL	406
SFO	SEA	404
SFO	ORD	403
DEN	PHX	402
SFO	LAS	

SQL QUERIES

Which airline has longer average delay time than total average delay time?

```
SELECT string_field_1 AS Airline, AVG(ARRIVAL_DELAY) AS AvgDelayTime
FROM `data-mgnt-project.flights_data.flights_dec` JOIN `data-mgnt-project.flights_data.airlines`
ON `data-mgnt-project.flights_data.flights_dec`.AIRLINE=`data-mgnt-project.flights_data.airlines`.string_field_0
WHERE ARRIVAL_DELAY > (SELECT AVG(ARRIVAL_DELAY)
FROM `data-mgnt-project.flights_data.flights_dec`)
GROUP BY AIRLINE
ORDER BY AVG(ARRIVAL_DELAY) DESC;
```

Airline	AvgDelayTime
United Air Lines Inc.	63.20179847
Frontier Airlines Inc.	60.24369748
Virgin America	53.20336538
JetBlue Airways	52.44429394
Delta Air Lines Inc.	52.05596921
Atlantic Southeast Airlines	51.61707655

Airline	AvgDelayTime
Skywest Airlines Inc.	50.78573733
American Eagle Airlines Inc.	50.75460695
Spirit Air Lines	45.69037859
American Airlines Inc.	43.80877547
Southwest Airlines Co.	40.1271813
Alaska Airlines Inc.	33.37859327
Hawaiian Airlines Inc.	25.311758

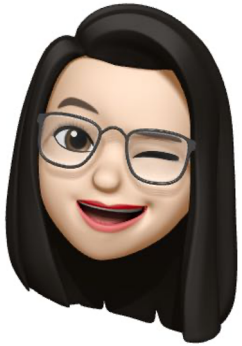
04

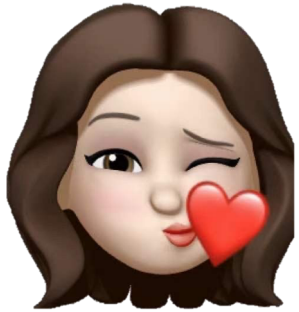
CONCLUSION

Key Takeaways

KEY TAKEAWAY

- Dates before (18, 17, 23) and after Christmas (27, 28) have more flights in a day.
- More well-developed states have more airports. For example, California and Texas have more high-tech companies, such as Google, Apple, Tesla.
- Only 1.7% of flights in December are cancelled and 70% of cancellations are due to weather.
- According to the average delay time, Alaska Airlines and Hawaii Airlines have shortest average delay time, while the average delay time of United Airlines and Frontier Airlines is over 1 hour.
- Choose the day around the end of a week (Thu, Fri, Sat) to lower cancel probability.
- We can find out which airline may be a good/ bad choice if we want to avoid delays






THANK YOU !