Flight Delay Trends at Chicago O’Hare airport

Group 4: Daniela Matinho, Hanna Kerr, Yuling Gu, Wei Yin

Agenda

Flight delays trends in Chicago O'Hare airport

- Executive Summary & Business Use Case
- Data Source & Data preparation
- Modeling
- Visualization
- MongoDB & Neo4J Insights
- Summary & Future Work

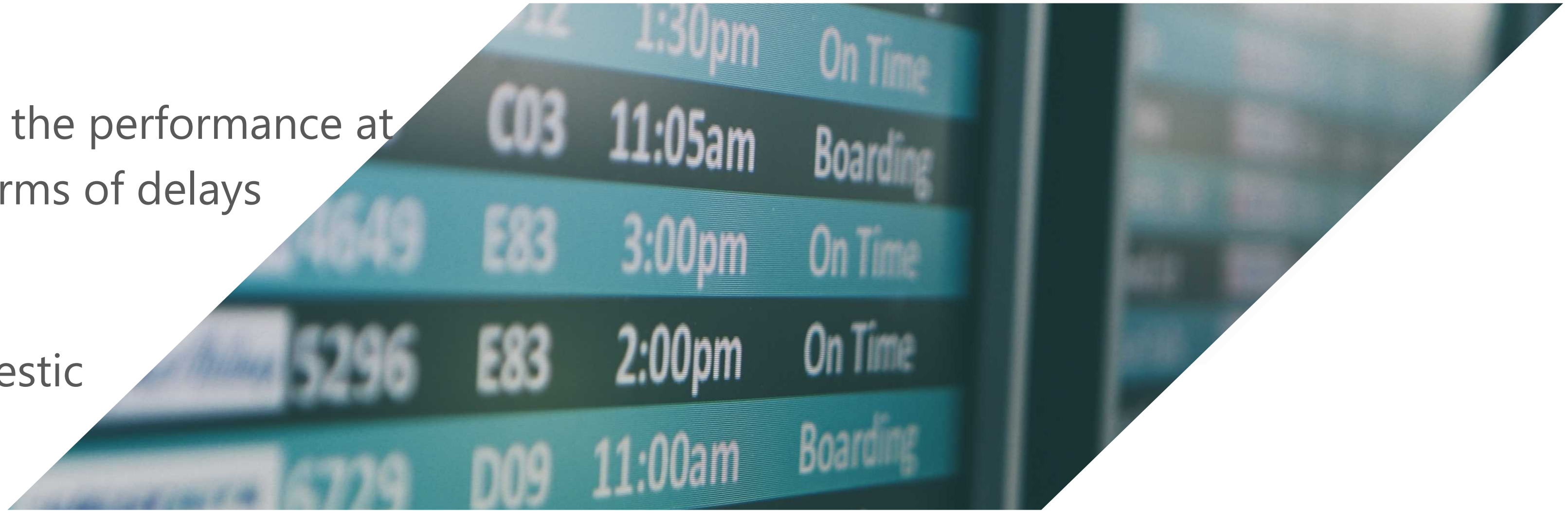
Executive Summary & Business Use Case

Overview

- ✓ Analyze flight disruptions that occur at the busiest airports in the US – Atlanta, LA, O'Hare, Denver, Dallas, JFK
- ✓ In depth analysis of the performance at O'Hare airport in terms of delays
- ✓ Data captures domestic flights from Jan to Dec 2018

7,213,446
2018
Total Number of Flights in U.S

1,304,214
2018
Total Departure Delays in U.S.



67,647 out of 332,953
2018
Total Departure Delays per Number of Flights at O'Hare airport

Business Use Case

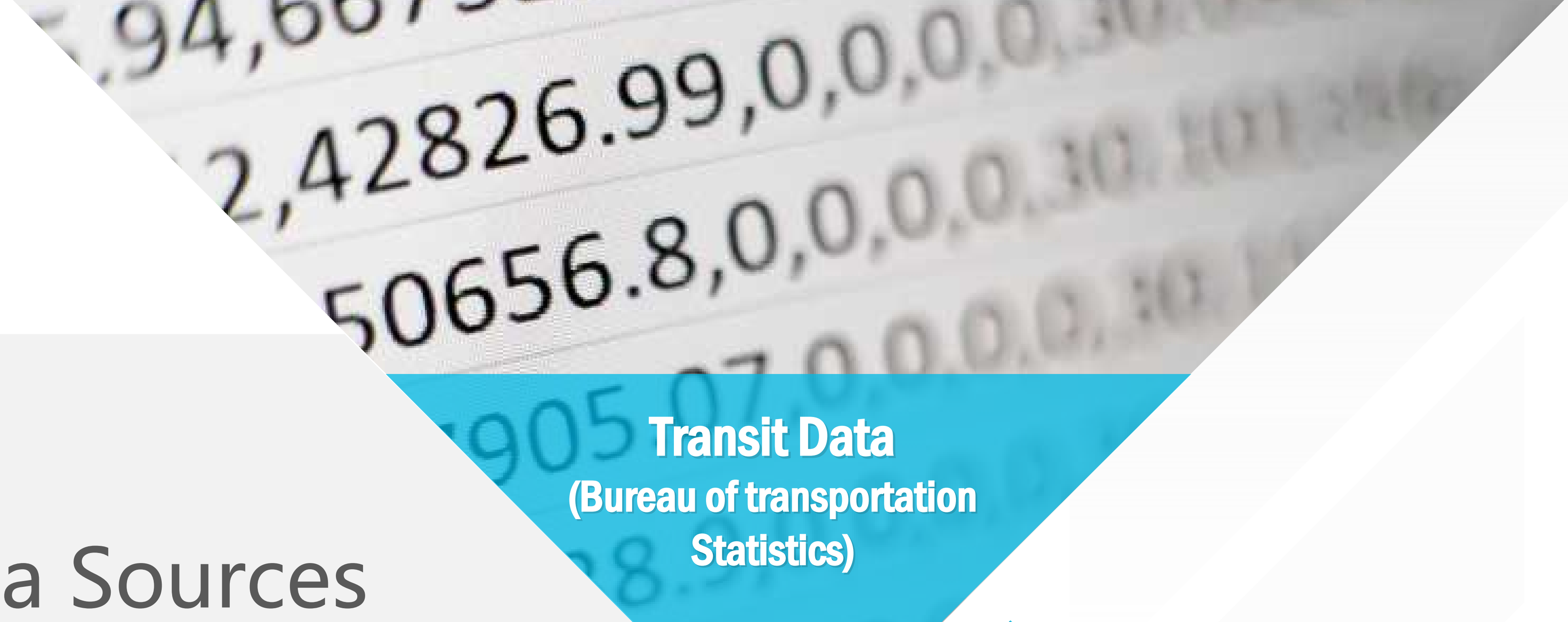
Can we define a pattern in the flight delays at O’Hare airport?

- Define critical periods of the year in terms of delays to prepare and optimize resources
- Identify the main causes of delays at O’Hare – air traffic, weather, etc
- Create an on-time arrival tool that describes airlines’ performance
- Pinpoint origins & destinations that are most likely to have delays

Data Source & Data preparation

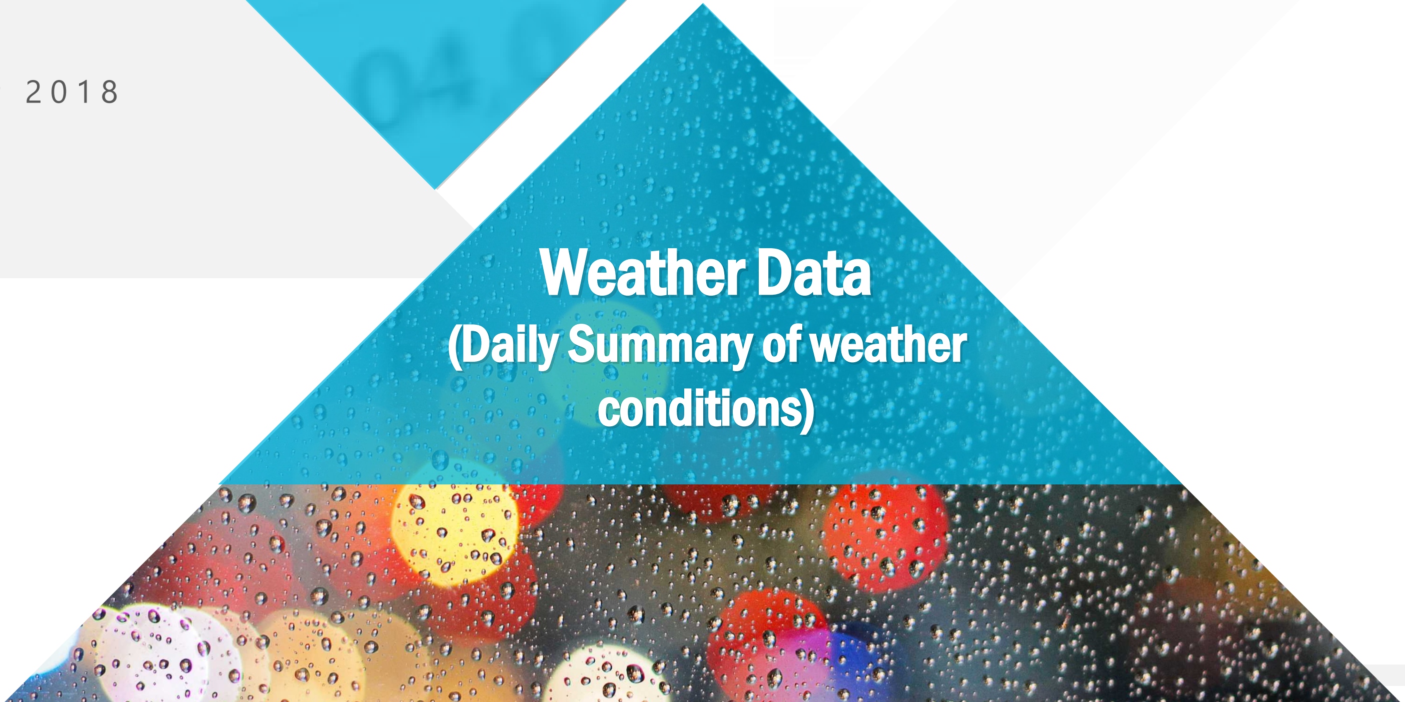
Data Sources

Data from January to December 2018



Transit Data

(Bureau of transportation Statistics)



Weather Data

(Daily Summary of weather conditions)

Platform considerations

01 Cleaning & Analysis

Python

- Data from different formats can easily be pulled in
- Automation of the cleaning process to save time for repeated tasks

02 Storing

MySQL & Google Cloud

- Security and control
- Cheap storage
- Group work options

03 Visualization

Tableau

- Connected to GCS
- Self-serving
- Easy to use

Data Preparation

Data Source

- Monthly files with flight data
- Daily files with weather information

Data Ingestion & Cleaning

- Format different data sources
- Clean columns

Data Store

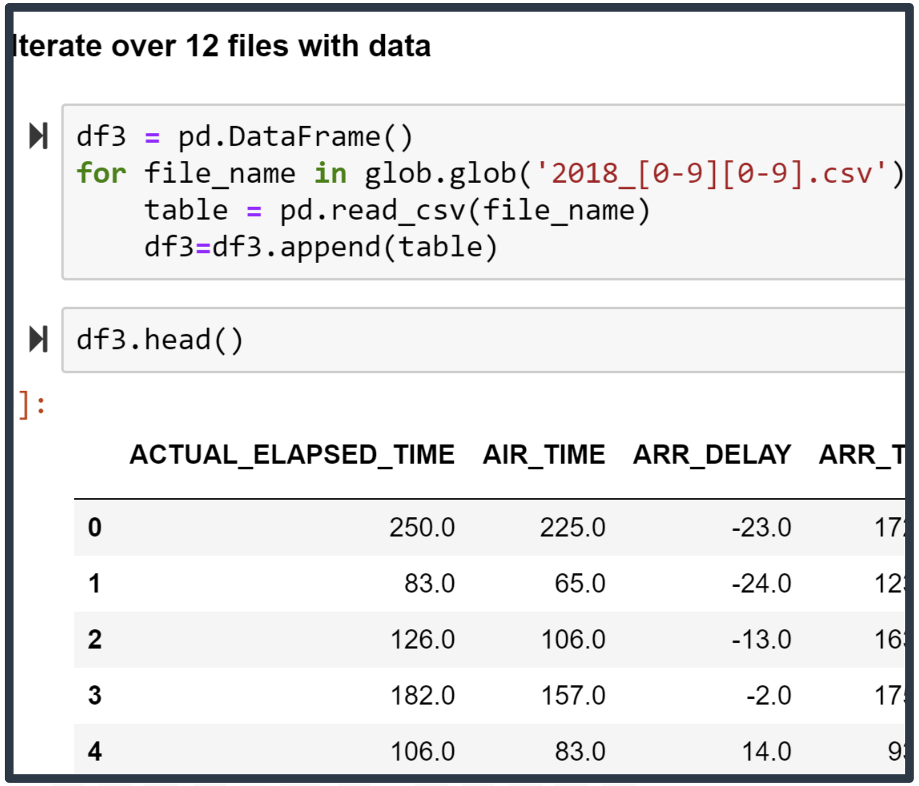
- Store data in google cloud using buckets

Dimension Schema

- Pull data into a dimensional model
- Connect Mysql with tableau to visualize data

Data Preparation – Step by Step

01. Python



Iterate over 12 files with data

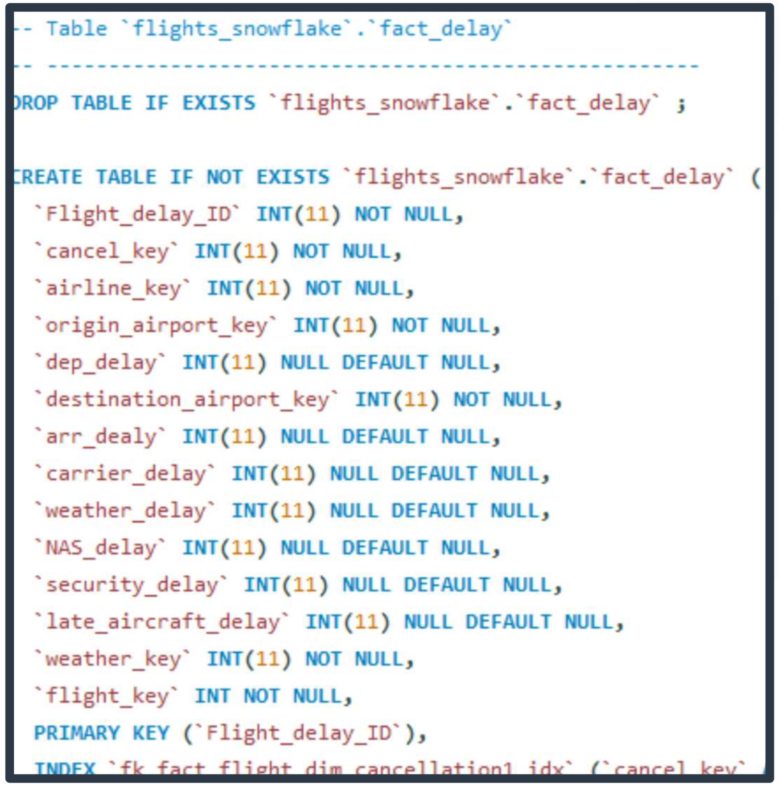
```
df3 = pd.DataFrame()
for file_name in glob.glob('2018_[0-9][0-9].csv')
    table = pd.read_csv(file_name)
    df3=df3.append(table)
```

```
df3.head()
```

]:

	ACTUAL_ELAPSED_TIME	AIR_TIME	ARR_DELAY	ARR_T
0	250.0	225.0	-23.0	17
1	83.0	65.0	-24.0	12
2	126.0	106.0	-13.0	16
3	182.0	157.0	-2.0	17
4	106.0	83.0	14.0	9

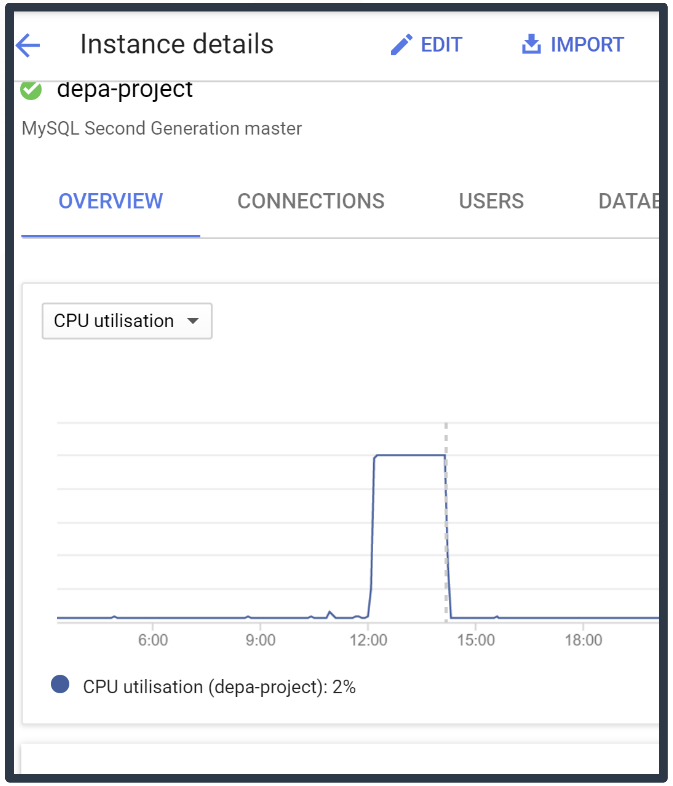
02. MySQL



```
-- Table `flights_snowflake`.`fact_delay`
-- -------------------------------------------------------
DROP TABLE IF EXISTS `flights_snowflake`.`fact_delay` ;

CREATE TABLE IF NOT EXISTS `flights_snowflake`.`fact_delay` (
  `Flight_delay_ID` INT(11) NOT NULL,
  `cancel_key` INT(11) NOT NULL,
  `airline_key` INT(11) NOT NULL,
  `origin_airport_key` INT(11) NOT NULL,
  `dep_delay` INT(11) NULL DEFAULT NULL,
  `destination_airport_key` INT(11) NOT NULL,
  `arr_dealy` INT(11) NULL DEFAULT NULL,
  `carrier_delay` INT(11) NULL DEFAULT NULL,
  `weather_delay` INT(11) NULL DEFAULT NULL,
  `NAS_delay` INT(11) NULL DEFAULT NULL,
  `security_delay` INT(11) NULL DEFAULT NULL,
  `late_aircraft_delay` INT(11) NULL DEFAULT NULL,
  `weather_key` INT(11) NOT NULL,
  `flight_key` INT NOT NULL,
  PRIMARY KEY (`Flight_delay_ID`),
  INDEX `fk_fact_flight_dim_cancellation1_idx` (`cancel_key`
```

03. Google Cloud



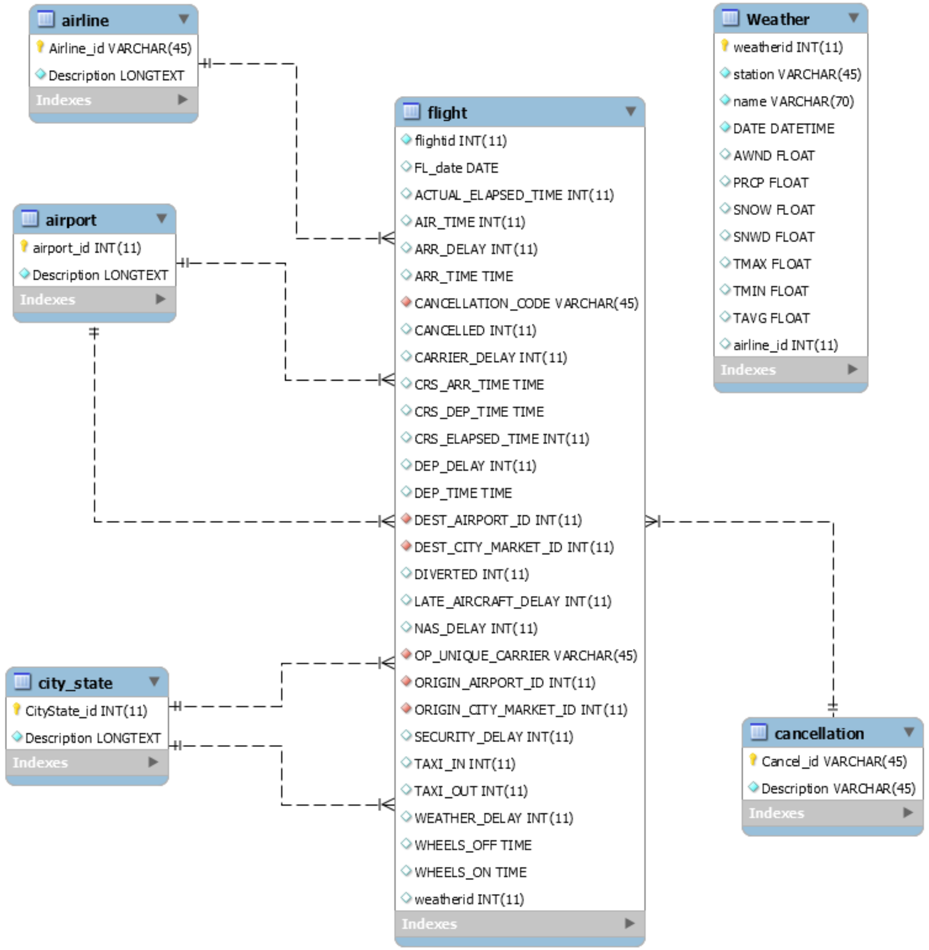
✓ depa-project

MySQL Second Generation master

● CPU utilisation (depa-project): 2%

Modeling

EER Diagram



airline
- Airline_id VARCHAR(45)
- Description LONGTEXT
- Indexes

airport
- airport_id INT(11)
- Description LONGTEXT
- Indexes

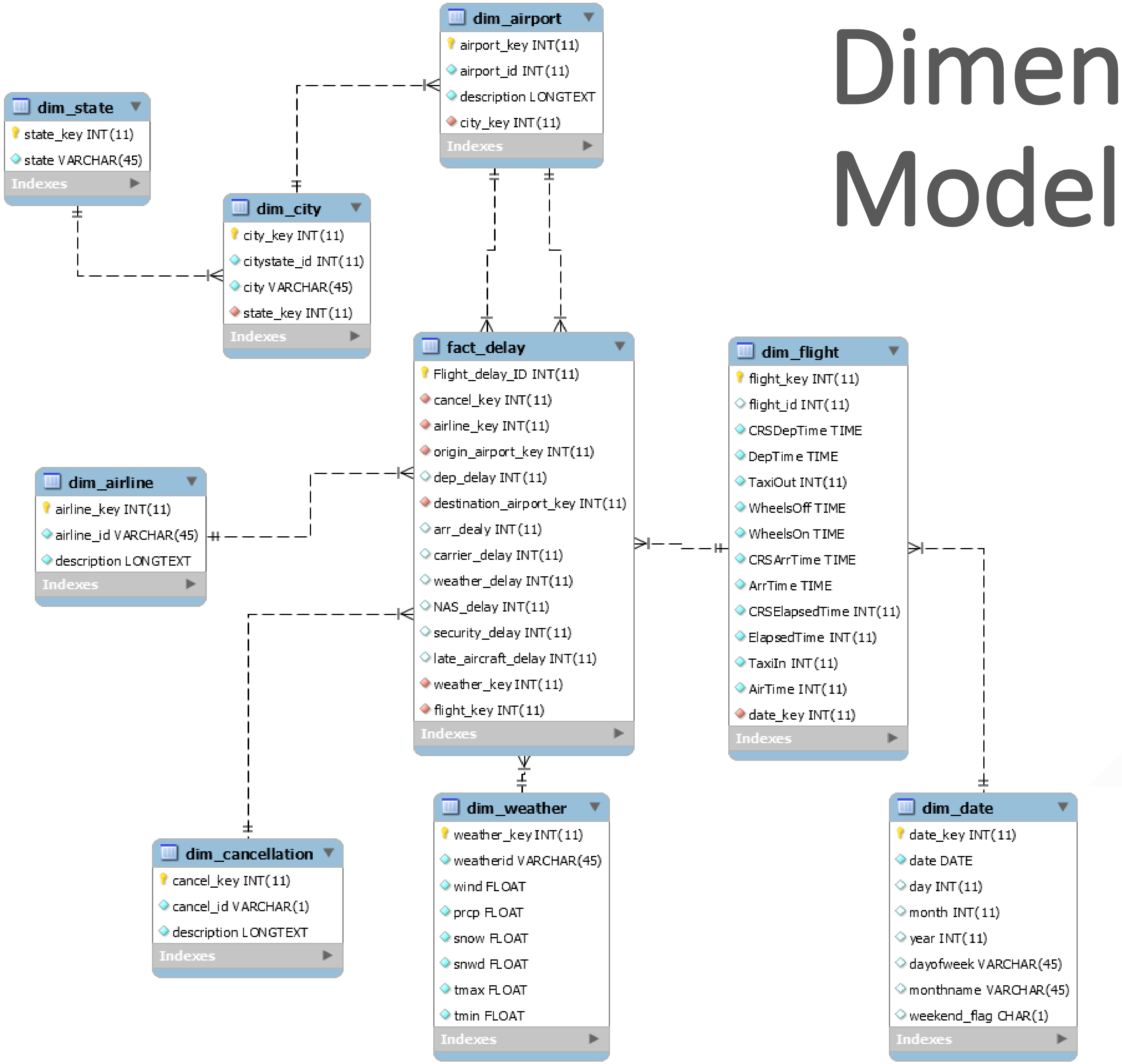
city_state
- CityState_id INT(11)
- Description LONGTEXT
- Indexes

flight
- flightid INT(11)
- FL_date DATE
- ACTUAL_ELAPSED_TIME INT(11)
- AIR_TIME INT(11)
- ARR_DELAY INT(11)
- ARR_TIME TIME
- CANCELLATION_CODE VARCHAR(45)
- CANCELLED INT(11)
- CARRIER_DELAY INT(11)
- CRS_ARR_TIME TIME
- CRS_DEP_TIME TIME
- CRS_ELAPSED_TIME INT(11)
- DEP_DELAY INT(11)
- DEP_TIME TIME
- DEST_AIRPORT_ID INT(11)
- DEST_CITY_MARKET_ID INT(11)
- DIVERTED INT(11)
- LATE_AIRCRAFT_DELAY INT(11)
- NAS_DELAY INT(11)
- OP_UNIQUE_CARRIER VARCHAR(45)
- ORIGIN_AIRPORT_ID INT(11)
- ORIGIN_CITY_MARKET_ID INT(11)
- SECURITY_DELAY INT(11)
- TAXI_IN INT(11)
- TAXI_OUT INT(11)
- WEATHER_DELAY INT(11)
- WHEELS_OFF TIME
- WHEELS_ON TIME
- weatherid INT(11)
- Indexes

Weather
- weatherid INT(11)
- station VARCHAR(45)
- name VARCHAR(70)
- DATE DATETIME
- AWND FLOAT
- PRCP FLOAT
- SNOW FLOAT
- SNWD FLOAT
- TMAX FLOAT
- TMIN FLOAT
- TAVG FLOAT
- airline_id INT(11)
- Indexes

cancellation
- Cancel_id VARCHAR(45)
- Description VARCHAR(45)
- Indexes

Dimensional Model



dim_airport
- airport_key INT(11)
- airport_id INT(11)
- description LONGTEXT
- city_key INT(11)
- Indexes

dim_state
- state_key INT(11)
- state VARCHAR(45)
- Indexes

dim_city
- city_key INT(11)
- citystate_id INT(11)
- city VARCHAR(45)
- state_key INT(11)
- Indexes

fact_delay
- Flight_delay_ID INT(11)
- cancel_key INT(11)
- airline_key INT(11)
- origin_airport_key INT(11)
- dep_delay INT(11)
- destination_airport_key INT(11)
- arr_dealy INT(11)
- carrier_delay INT(11)
- weather_delay INT(11)
- NAS_delay INT(11)
- security_delay INT(11)
- late_aircraft_delay INT(11)
- weather_key INT(11)
- flight_key INT(11)
- Indexes

dim_flight
- flight_key INT(11)
- flight_id INT(11)
- CRSDepTime TIME
- DepTime TIME
- TaxiOut INT(11)
- WheelsOff TIME
- WheelsOn TIME
- CRSArrTime TIME
- ArrTime TIME
- CRSElapsedTime INT(11)
- ElapsedTime INT(11)
- TaxiIn INT(11)
- AirTime INT(11)
- date_key INT(11)
- Indexes

dim_airline
- airline_key INT(11)
- airline_id VARCHAR(45)
- description LONGTEXT
- Indexes

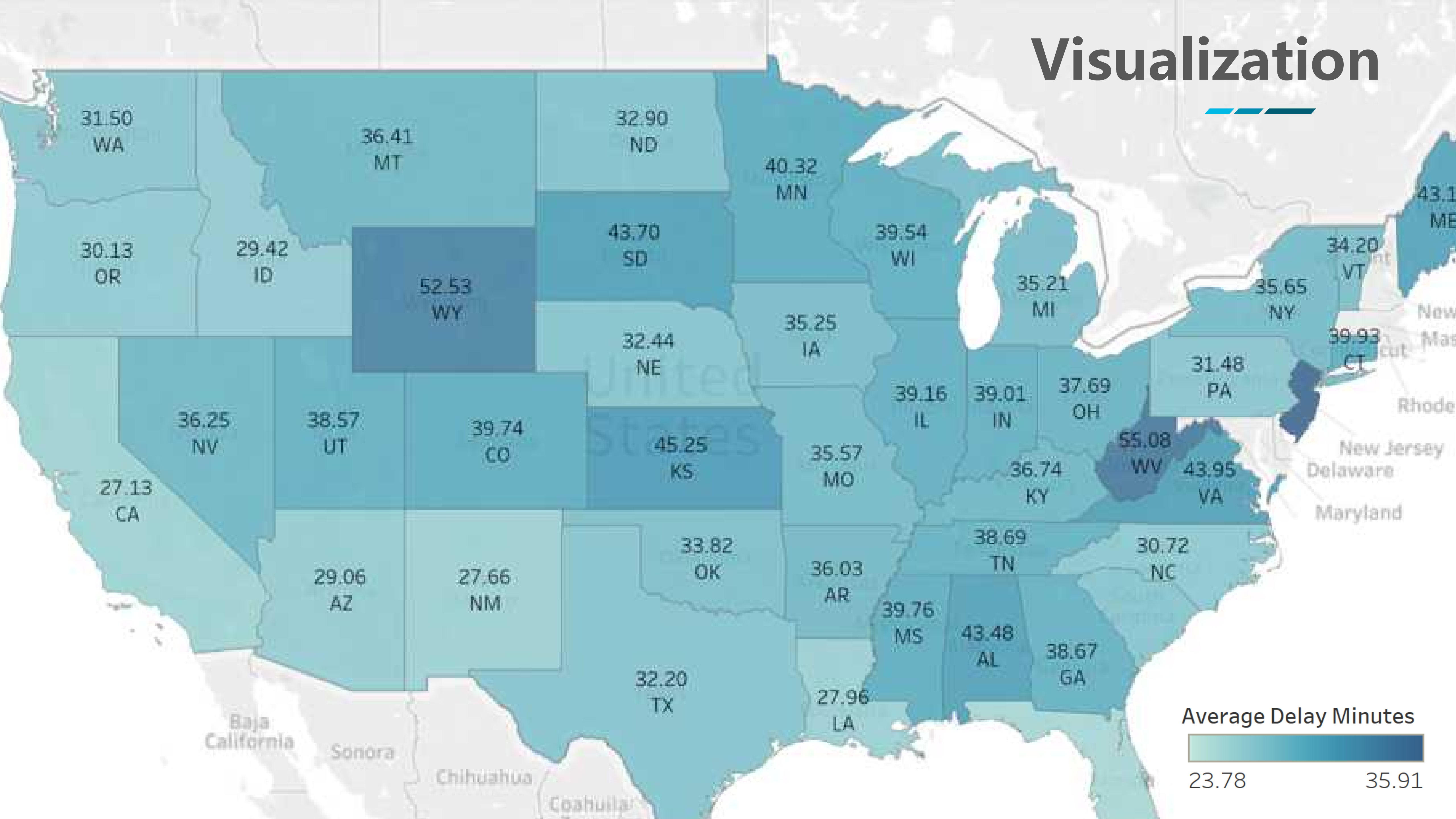
dim_cancellation
- cancel_key INT(11)
- cancel_id VARCHAR(1)
- description LONGTEXT
- Indexes

dim_weather
- weather_key INT(11)
- weatherid VARCHAR(45)
- wind FLOAT
- prcp FLOAT
- snow FLOAT
- snwd FLOAT
- tmax FLOAT
- tmin FLOAT
- Indexes

dim_date
- date_key INT(11)
- date DATE
- day INT(11)
- month INT(11)
- year INT(11)
- dayofweek VARCHAR(45)
- monthname VARCHAR(45)
- weekend_flag CHAR(1)
- Indexes

Visualization

Visualization



Analysis

About the 6 airport

- Atlanta is the busiest airport but is also the best in terms of delays
- Average departure delay for the 6 airports, range from 24 to 37 min (with JFK having the largest average)

O’Hare airport

- O’Hare has an average departure delay of 35 min
- The main type of delays are: Aircraft Delay & National Aircraft System & Carrier Delay
- Cancelation rate is higher during the severe weather periods – However rate decreases throughout the year
- Destinations that are more likely to have a flight delay: Wyoming & West Virgin
- O’Hare airport is performing better when the weather is better
- The cheapest airlines have the highest average departure delay minutes (Frontier: 54 min)

MongoDB & NEO 4J Insights

MongoDB

JSON Data Model

```
{
  "_id": "5cf8a6e1733d928192870a56",
  "OP_CARRIER_FL_NUM": 1,
  "ORIGIN_AIRPORT_ID": 12478,
  "ORIGIN": "JFK",
  "ORIGIN_CITY_NAME": "New York, NY",
  "ORIGIN_STATE_NM": "New York",
  "DEST_AIRPORT_ID": 11697,
  "DEST": "FLL",
  "DEST_CITY_NAME": "Fort Lauderdale, FL",
  "DEST_STATE_NM": "Florida",
  "DEP_TIME": 1000,
  "DEP_DELAY": 0,
  "ARR_TIME": 1319,
  "ARR_DELAY": 4
   + {... }
}
```

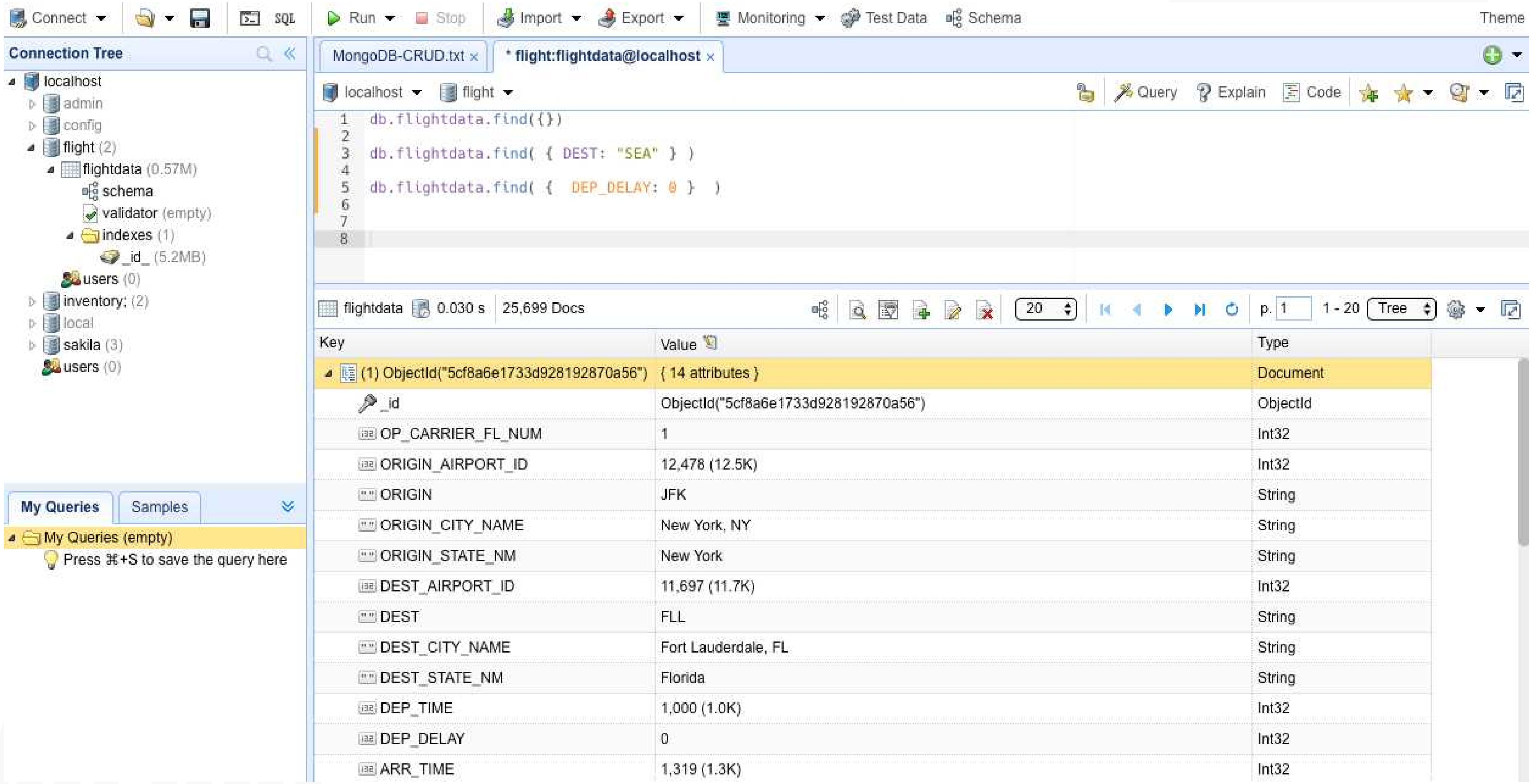
Pros

- Easy to create queries, optimize, & maintain
  Work with data in a natural, intuitive way
- Capacity to adapt & make changes quickly
- Great performance with less code
- Freedom to run anywhere

Cons

- High Memory Usage (due to no functionality of joins, there is data redundancy)
- Limited Data Size
- Limited Nesting (cannot perform nesting of documents for more than 100 levels)
- Less Secure

MongoDB – In action



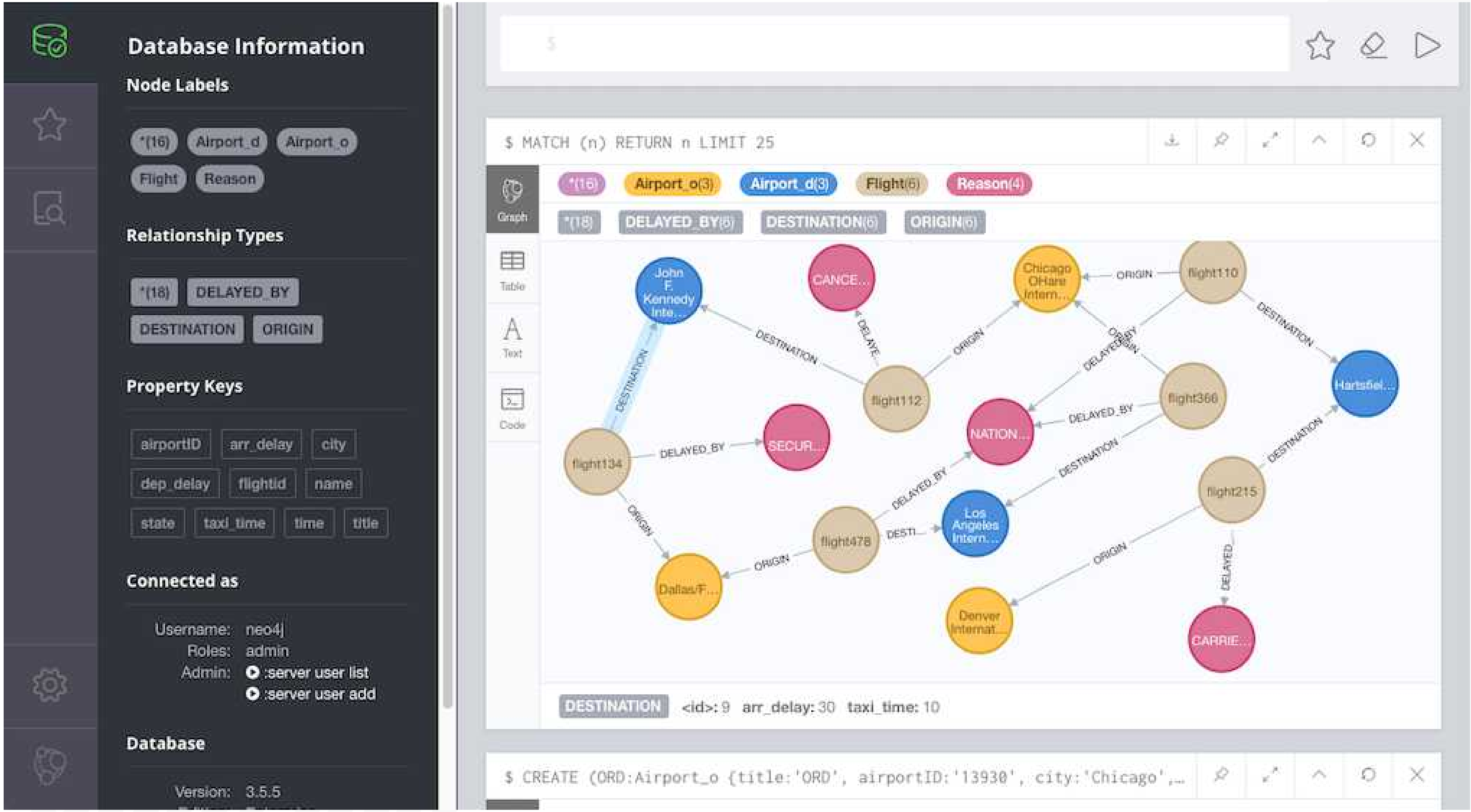
```
db.flightdata.find({})

db.flightdata.find( { DEST: "SEA" } )

db.flightdata.find( {  DEP_DELAY: 0 }  )
```

Key	Value	Type
(1) ObjectId("5cf8a6e1733d928192870a56")	{ 14 attributes }	Document
_id	ObjectId("5cf8a6e1733d928192870a56")	ObjectId
OP_CARRIER_FL_NUM	1	Int32
ORIGIN_AIRPORT_ID	12,478 (12.5K)	Int32
ORIGIN	JFK	String
ORIGIN_CITY_NAME	New York, NY	String
ORIGIN_STATE_NM	New York	String
DEST_AIRPORT_ID	11,697 (11.7K)	Int32
DEST	FLL	String
DEST_CITY_NAME	Fort Lauderdale, FL	String
DEST_STATE_NM	Florida	String
DEP_TIME	1,000 (1.0K)	Int32
DEP_DELAY	0	Int32
ARR_TIME	1,319 (1.3K)	Int32

Neo 4J



Database Information

Node Labels

*(16) Airport_d Airport_o Flight Reason

Relationship Types

*(18) DELAYED_BY DESTINATION ORIGIN

Property Keys

airportID arr_delay city dep_delay flightid name state taxi_time time title

Connected as

Username: neo4j
Roles: admin
Admin: :server user list
:server user add

Database

Version: 3.5.5

```
$ MATCH (n) RETURN n LIMIT 25
```

*(16) Airport_o(3) Airport_d(3) Flight(6) Reason(4)

*(18) DELAYED_BY(6) DESTINATION(6) ORIGIN(6)

DESTINATION <id>: 9 arr_delay: 30 taxi_time: 10

```
$ CREATE (ORD:Airport_o {title:'ORD', airportID:'13930', city:'Chicago',…
```

Summary & Future Work

Lessons learned

Data Preparation & Storing

- Cleaning the data is the most important step in the project – the better the quality of the data, the easier it is to treat, analyze & visualize data
- Keeping consistency in the cleaning process is crucial (e.g. format of variables)
- Creating indexes is necessary for querying large sets of data
- Data preparation & loading accounts for most of the overall process

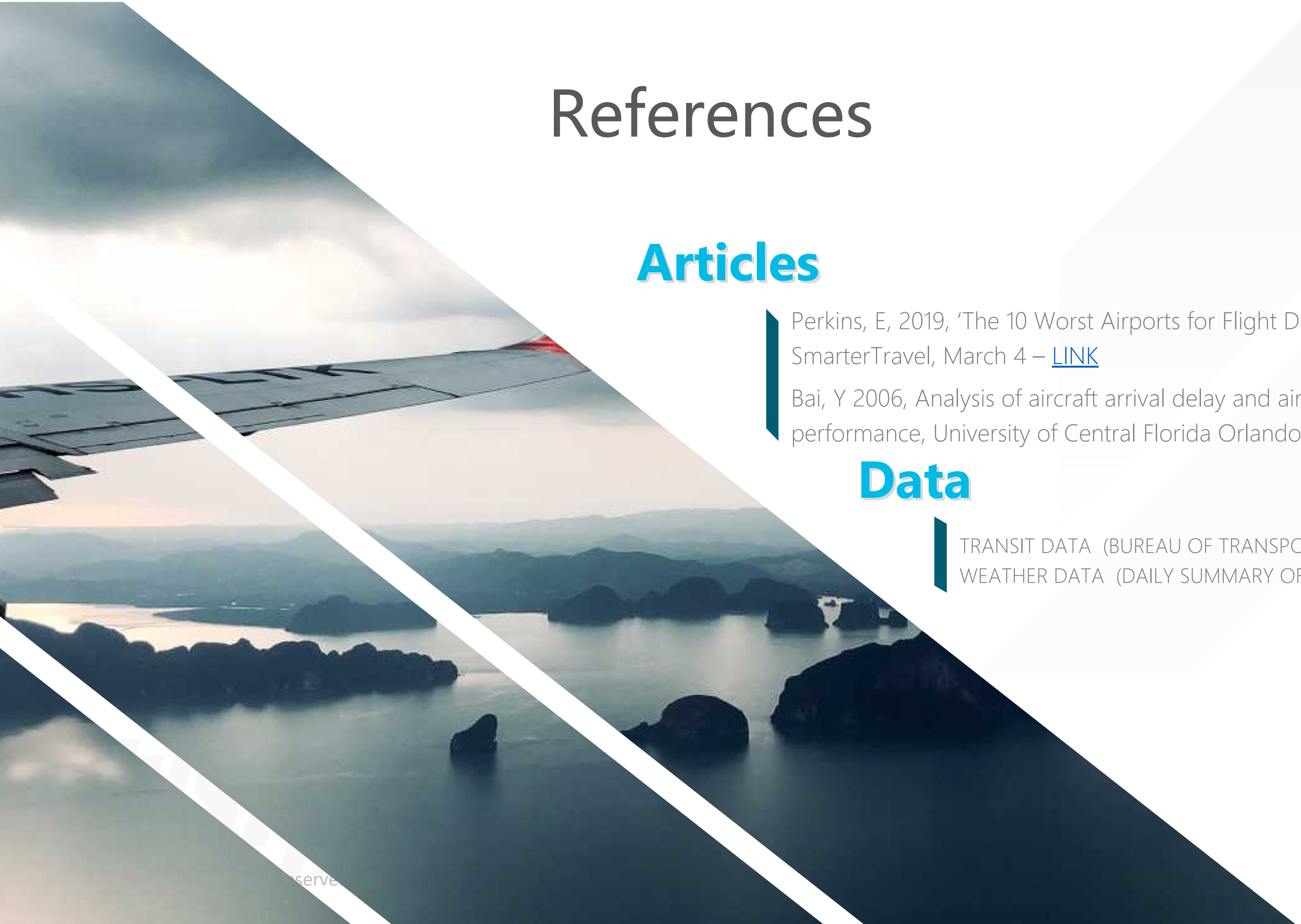
Visualization

- Keeping the business use case in mind while performing visualization

AND MUCH MORE… ▶

Future Work

- Build a **model that predicts the probability** of delay & length of delay for upcoming flight
- Compare **ticket prices** for airlines with rate of delay
- Consider data from **international flights**
- **Analyze data for several years** to get a better understanding on O’Hare performance
- Analyze **costs related** to delays & find ways to reduce them
- **Automation of the process,** to constantly update new data



References

Articles

Perkins, E, 2019, 'The 10 Worst Airports for Flight Delays, Ranked', SmarterTravel, March 4 – LINK

Bai, Y 2006, Analysis of aircraft arrival delay and airport on-time performance, University of Central Florida Orlando – LINK

Data

TRANSIT DATA (BUREAU OF TRANSPORTATION STATISTICS) – LINK

WEATHER DATA (DAILY SUMMARY OF WEATHER CONDITIONS) – LINK

THANK YOU

Q&A

Group 4: Daniela Matinho, Hanna Kerr, Yuling Gu, Wei Yin