YORK
U
UNIVERSITÉ
UNIVERSITY

EECS 4415 – Big Data Project
Airline Performance

In class presentation by:

Jaskaran, William, Gilbert, David

Motivation

- Did you know?
  - There were 3.8 billion air travelers in 2016. IATA predicts by 2035 there will be 7.2 billion passengers [1].
- Look for patterns in flight delays and make predictions to inform customers of what to expect from specific airline(s) & geographical regions

Applications

- Help Airline companies improve their services and increase revenue
- Maintenance of old aircrafts (causing delays)
- Customers (Passengers) can use analysis as a guide to plan and make reliable bookings from airlines

Dataset

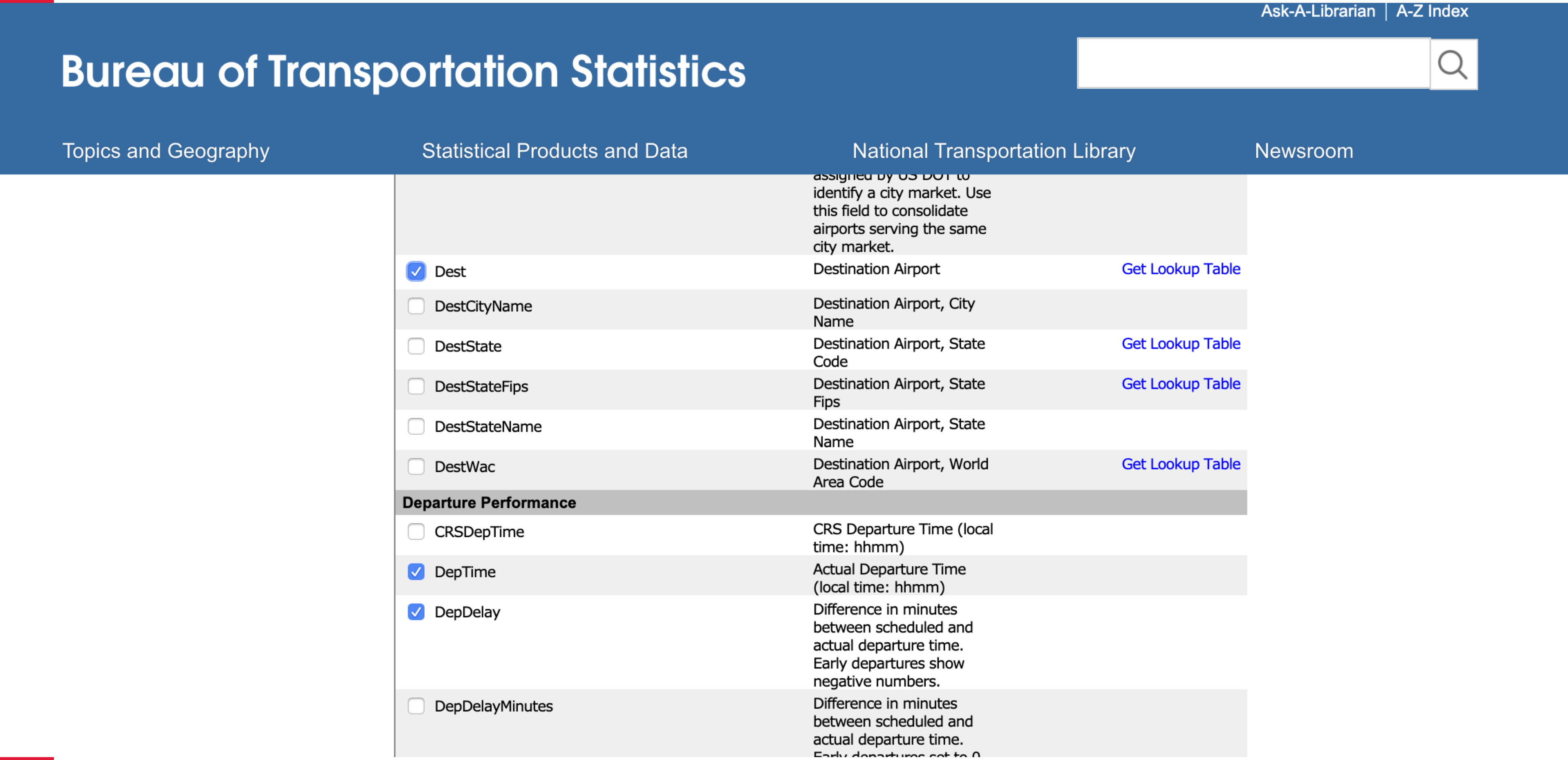
Flight Date	Origin	Dest	AirTime	Depart	Arrival	Delay	Tail #
2019-27-11	Miami - MIA	New York - JFK	240	1033	1549	0	N26232
…	…	…	…	…	…	…	…
2019-28-11	Ney York - JFK	Miami - MIA	267	620	1040	32	N26232

Bureau of Transportation Statistics (BTS)

- Aviation On-Time Performance 1987 to 2019
- We used data from 2016 – 2019

- Data Volume
  - 4 years ~ 5.5 GB (CSV file)

User Friendly Dataset, Minimal Data Cleaning


Ask-A-Librarian | A-Z Index
Bureau of Transportation Statistics
Topics and Geography
Statistical Products and Data
National Transportation Library
Newsroom

	assigned by US DOT to identify a city market. Use this field to consolidate airports serving the same city market.	
☑ Dest	Destination Airport	Get Lookup Table
□ DestCityName	Destination Airport, City Name	
□ DestState	Destination Airport, State Code	Get Lookup Table
□ DestStateFips	Destination Airport, State Fips	Get Lookup Table
□ DestStateName	Destination Airport, State Name	
□ DestWac	Destination Airport, World Area Code	Get Lookup Table
Departure Performance		
□ CRSDepTime	CRS Departure Time (local time: hhmm)	
☑ DepTime	Actual Departure Time (local time: hhmm)	
☑ DepDelay	Difference in minutes between scheduled and actual departure time. Early departures show negative numbers.	
□ DepDelayMinutes	Difference in minutes between scheduled and actual departure time. Early departures set to 0	

Analysis

Questions to answer

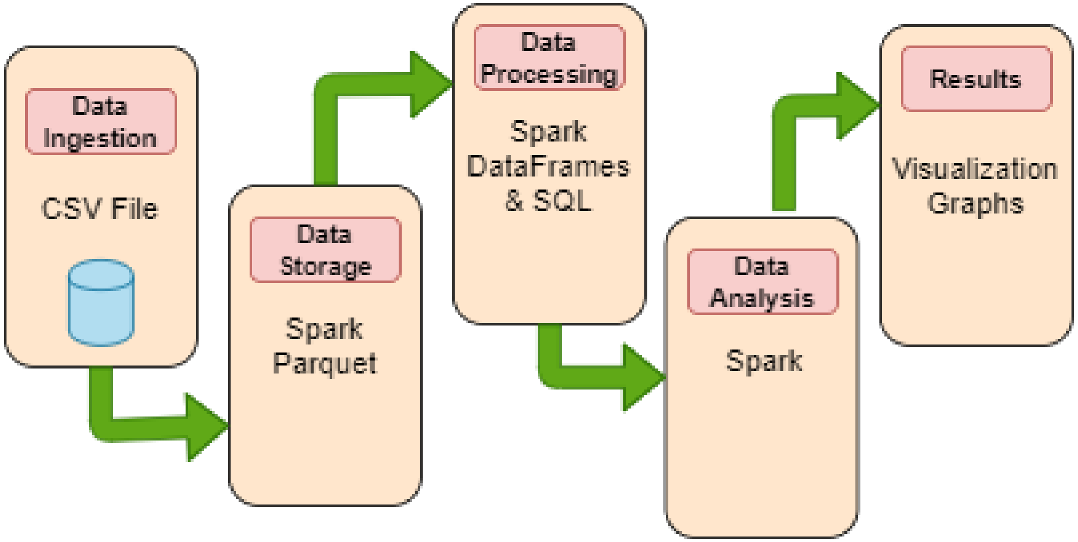
- Delays & Cancellation
  - most common cause
  - Which Airline companies have the most delays and cancellations
  - times of the year which have high probability of delays/cancellations

- Predict the best time of day/day of week/time of year to fly to minimise delays

- Fluctuations in Airline Industry
  - Increase/Decrease in # of flights in given time period (variance)
  - Ex. October 2019 Boeing 737 scandal.

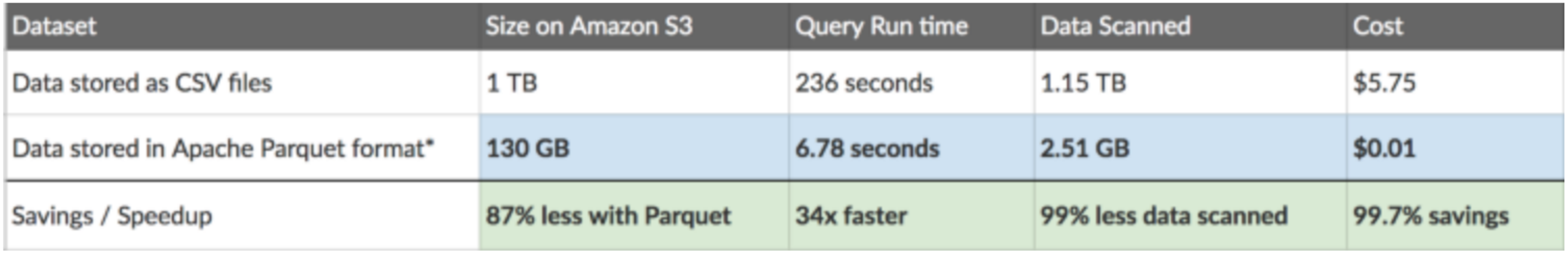
- Has there been an improvement or decline in the quality of service of an airline over a period?
  - Decrease in Delays and Increase in On-time performance

Architecture (Data pipeline)



Columnar Storage: CSV file → Apache Parquet file (Faster)

Parquet vs. CSV:

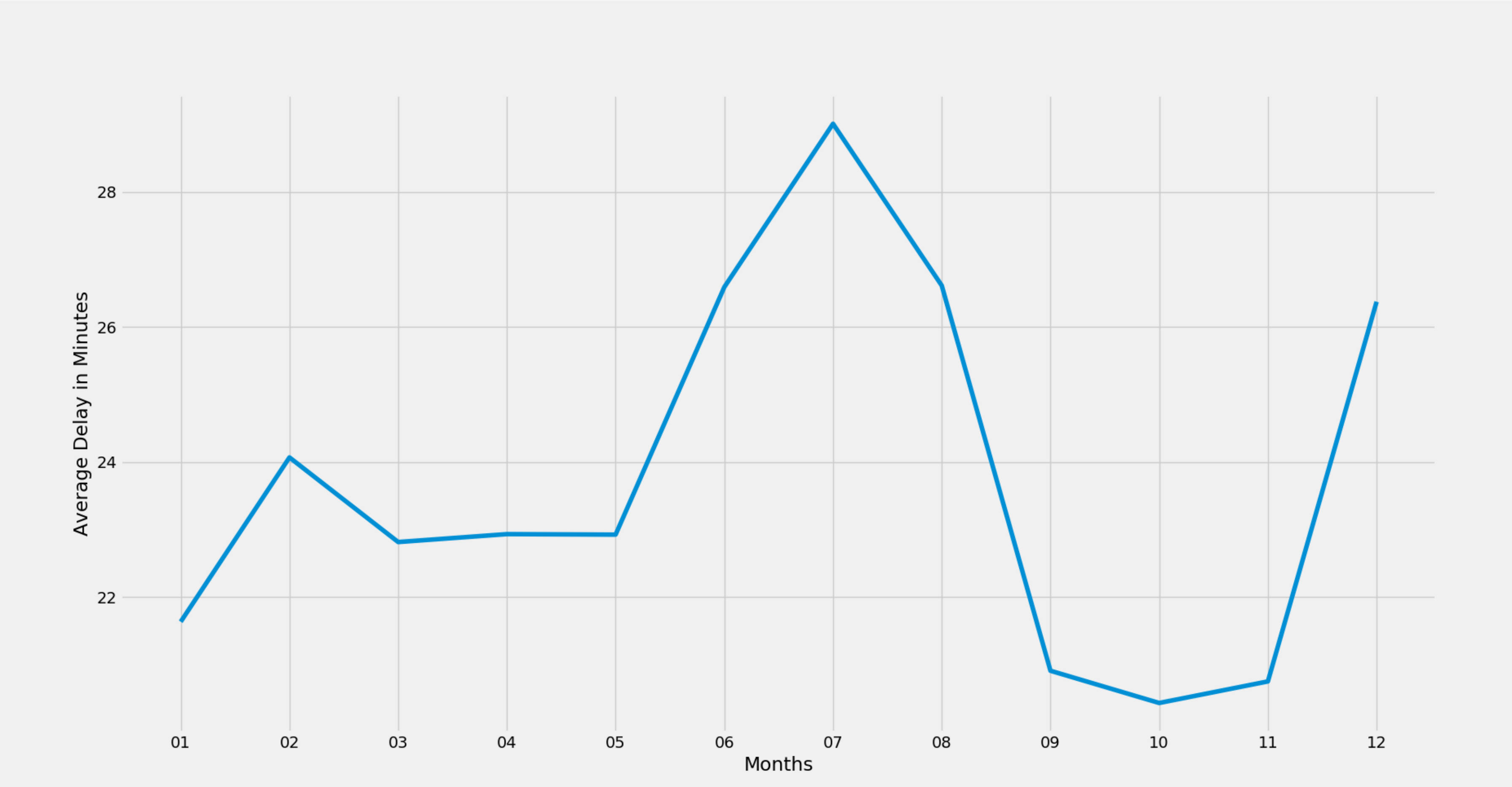


Dataset	Size on Amazon S3	Query Run time	Data Scanned	Cost
Data stored as CSV files	1 TB	236 seconds	1.15 TB	$5.75
Data stored in Apache Parquet format*	130 GB	6.78 seconds	2.51 GB	$0.01
Savings / Speedup	87% less with Parquet	34x faster	99% less data scanned	99.7% savings

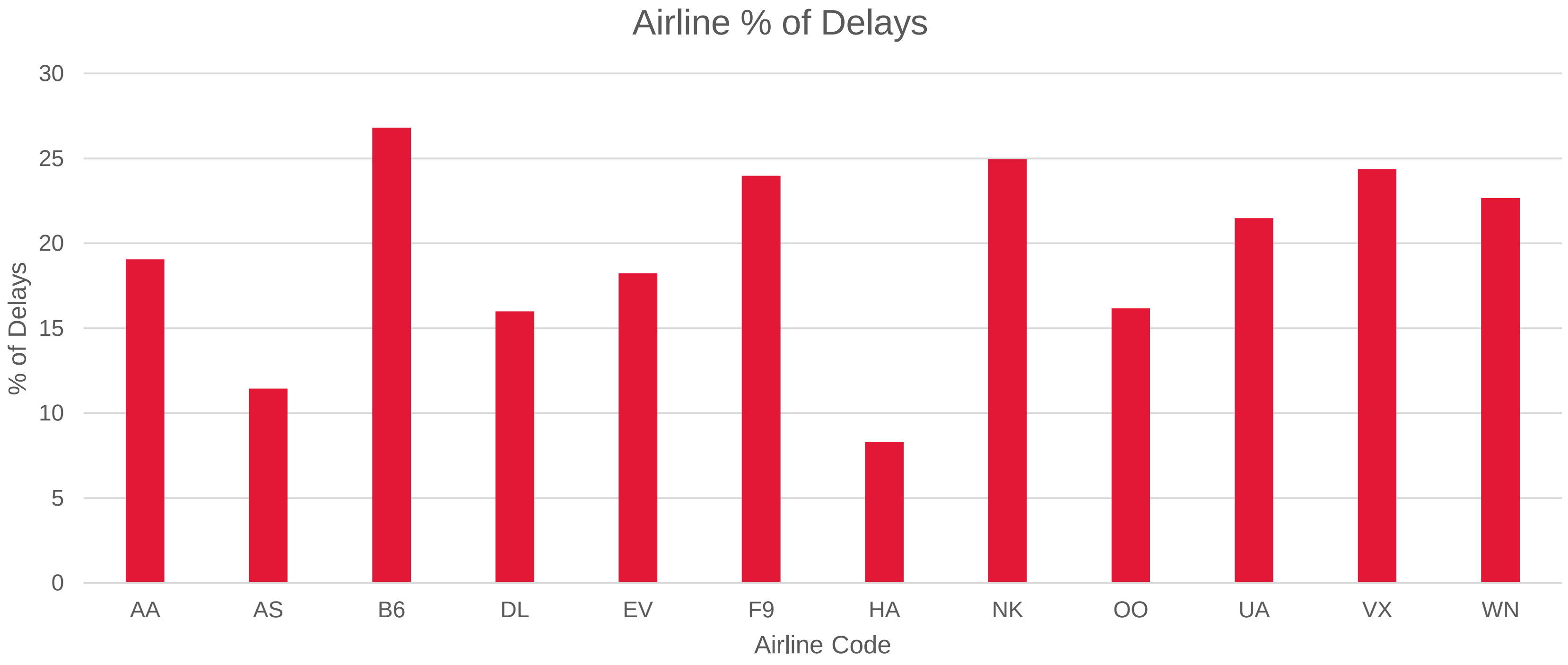
[2]

Results

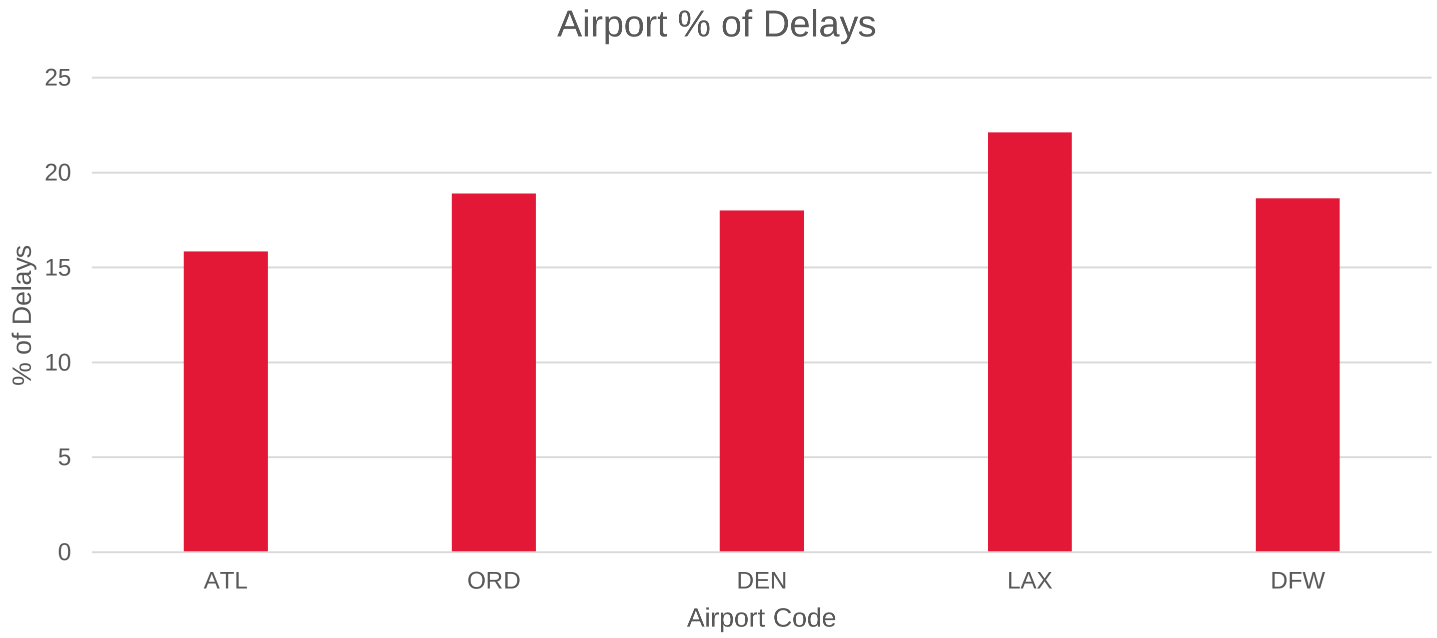
Monthly Delays 2016



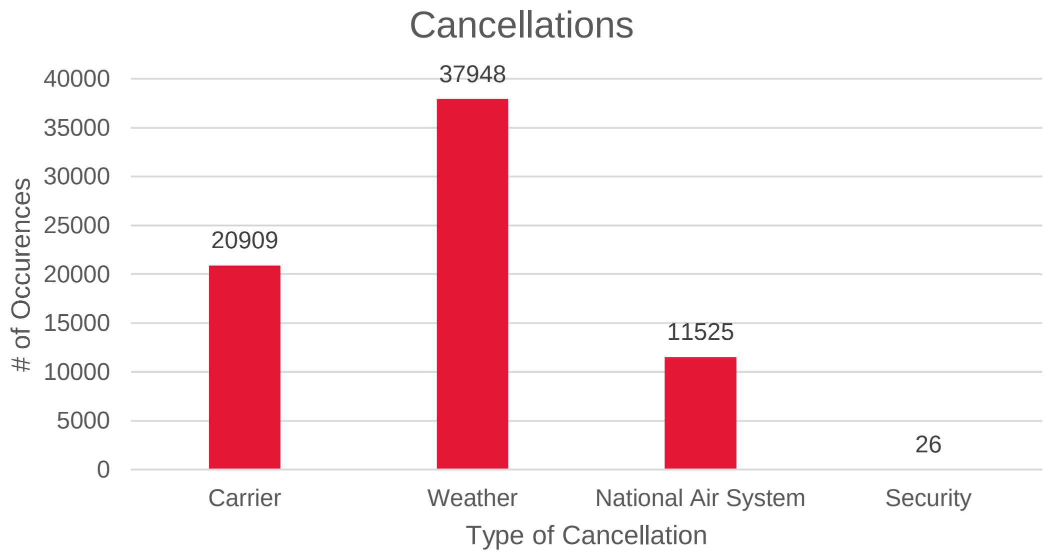
Results: Delays of Airlines 2016

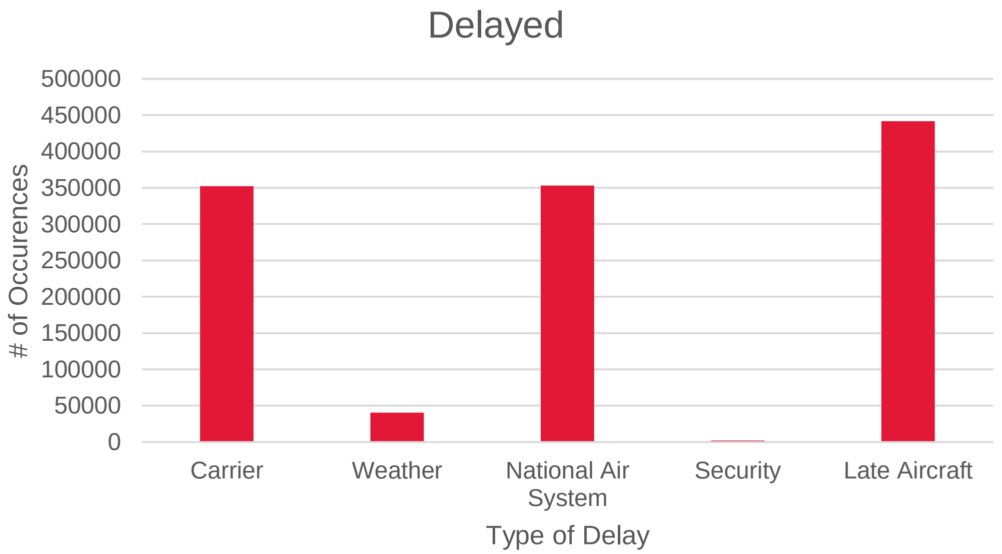


Results: Delays at Major Airports 2016

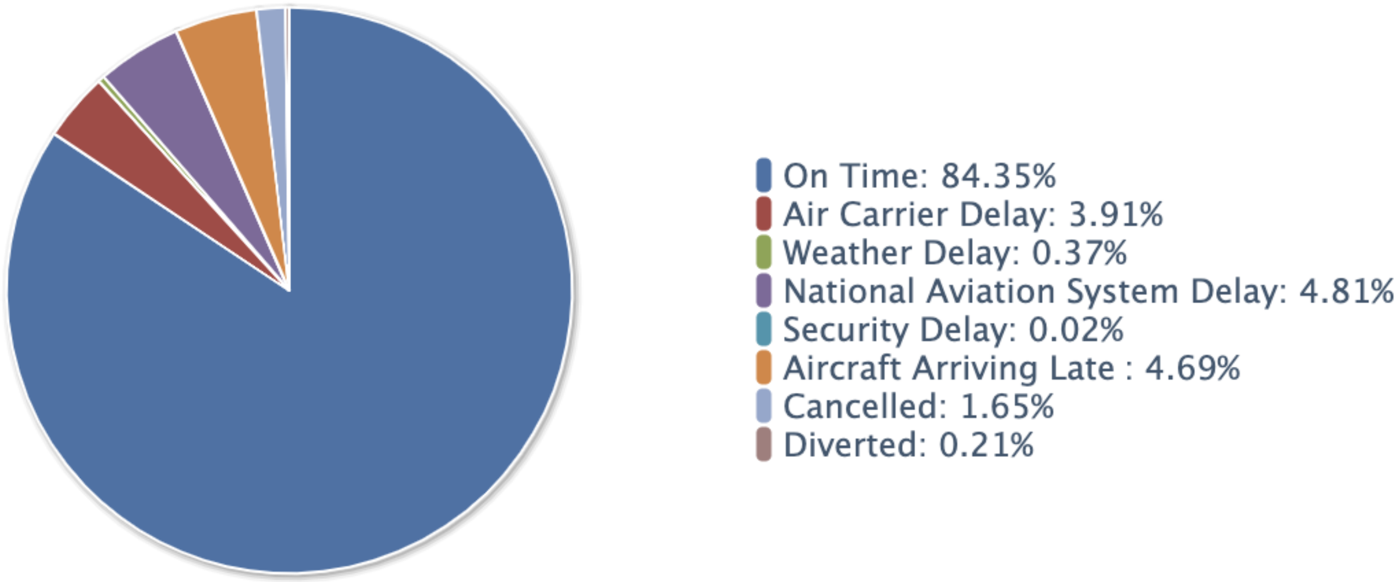


Results: Cancellations and Delays 2016





Overall Distribution



Limitations

Certain data types are unavailable for example:

- Aircraft types: manually correlate names, id, models, etc...

CSV files & Pandas didn’t work for “NULL” values

- Solved using parquet

Streaming Layer was hard to implement due to lack of real-time data from prominent sources

Future Work

Process Data from the past 20 years

- To see the historical improvement of delays/cancellations

Impact of external factors on air travel

- During Health crisis/epidemics
- Airline Fatalities affecting #of flights and airline

Implementing Page Ranking type mechanism, to recommend users with best flight airlines on a given week

Conclusion & Lessons Learned

- Fun project to implement
- Most of the delays were Weather specific or Airline specific
- Airlines delays cascaded – planes used for other flights are delayed on previous flights
    - Solving this issue could save money

References

[1] 2019. [Online]. Available: https://www.nationalgeographic.com/environment/urban-expeditions/transportation/air-travel-fuel-emissions-environment/. [Accessed: 27- Nov- 2019].

[2] "Apache Parquet vs. CSV Files - DZone Database", *dzone.com*, 2019. [Online]. Available: https://dzone.com/articles/how-to-be-a-hero-with-powerful-parquet-google-and. [Accessed: 27- Nov- 2019].