

AIRLINE MARKET ANALYSIS

ALY 6110: DATA MANAGEMENT AND BIG DATA CRN: 81984

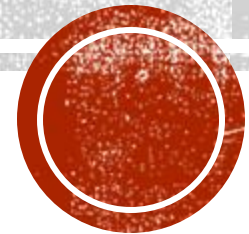
Submitted By: Group Beta

Ashwini Kumar Pathak [pathak.as@husky.neu.edu]

Ayush Jain [jain.ayush@husky.neu.edu]

Megha Ravi [ravi.m@husky.neu.edu]

Ashishkumar Bidap [bidap.a@husky.neu.edu]



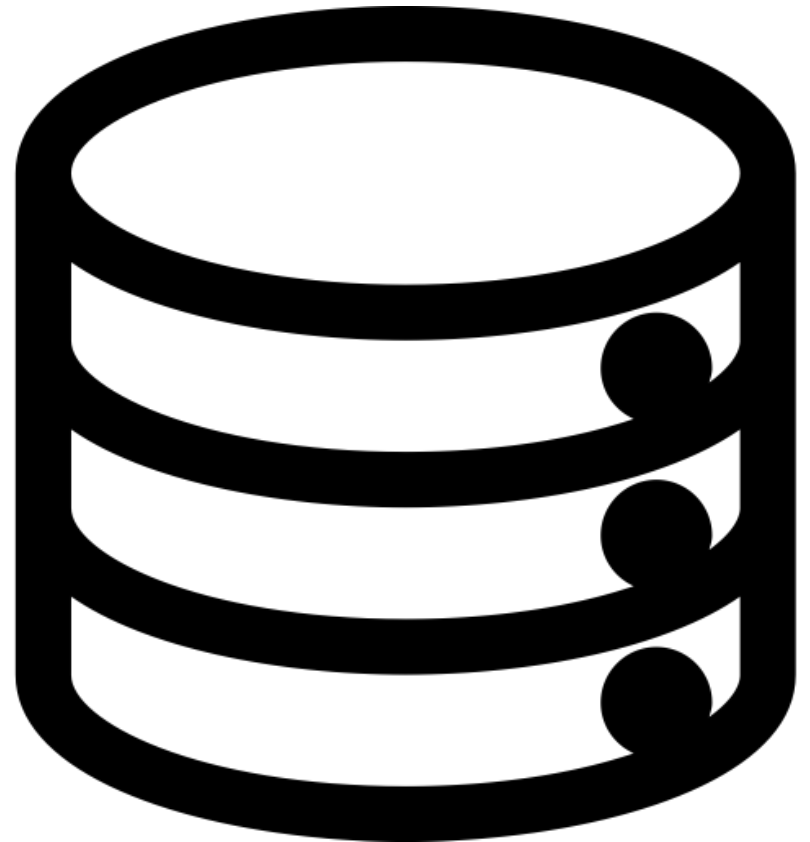
DATASET UTILIZED

- Airline Dataset
- Source: Bureau of Transport Statistics(BTS)
- Duration: Jan'2000 – Feb'2020
- Data size: 27 Gb
- Total files: 244 files
- Number of Rows: 128 Million
- Total Attributes: 54



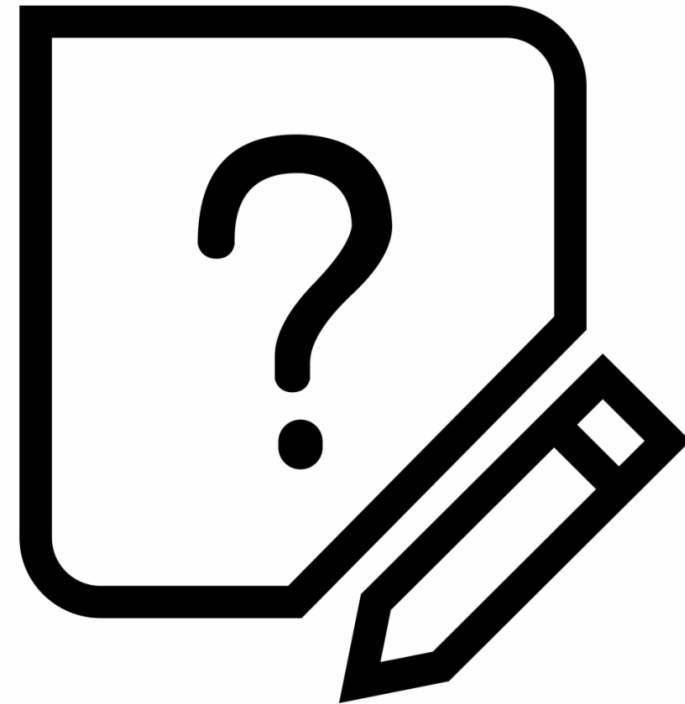
ABOUT DATASET

- Time period (Year, Month, DayofMonth, DayofWeek, FlightDate)
- Airline (Flight number, Unique Carrier number)
- Origin and Destination details, Departure and Arrival performances showcasing delays
- Cancellations, Cause of delays (Carrier delay, weather delay, National air system delay (in minutes), security delay, aircraft



PROBLEM STATEMENT:

- Exploratory Data Analysis
- Identify the Impact of global recession(2008)on US Flight industry.
- Fraud detection
- How Covid-19 impacted US Airline industry.

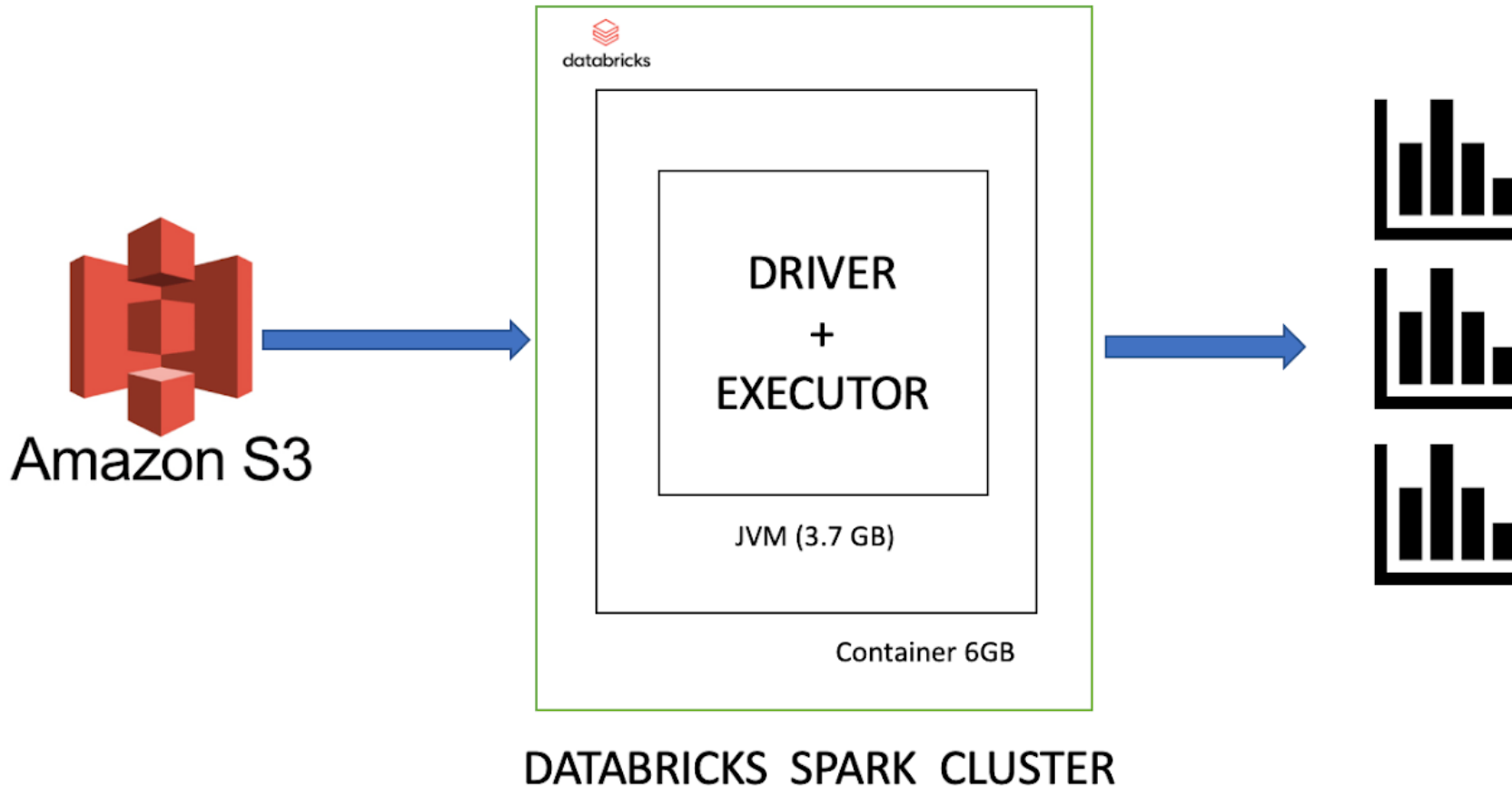


WHY BIG DATA?

- High Volume of data
- Fast retrieval
- Cost Saving
- Help in understanding market condition.
- Ability of tools to visualize insights.



OUR BIG DATA ARCHITECTURE



Databricks Workspace

Collaborative Notebooks, Production Jobs

Databricks Runtime



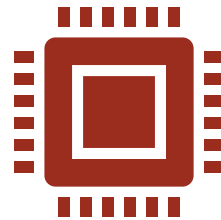
Databricks Cloud Service



CHALLENGES



Storage Space



CPU Compute
Power.



Multiple Node
Cluster



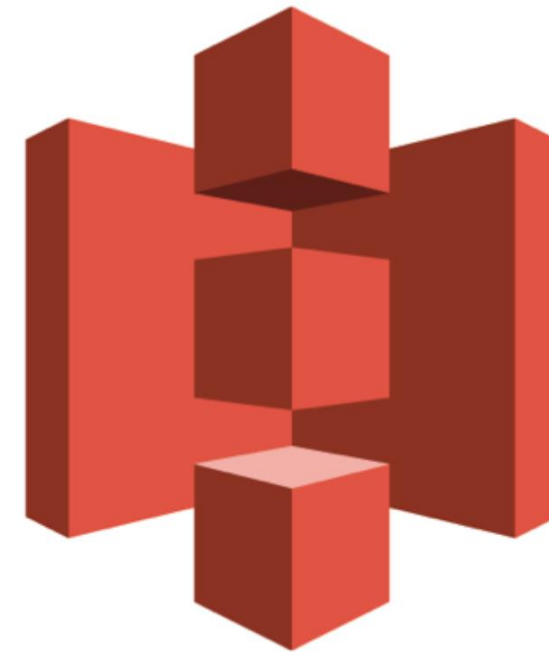
PERFORMANCE OPTIMIZATION OF APACHE SPARK QUERIES

- Repartitioning the data frame partitions
- Speeding up Shuffle.partitions
- Pulling data sets into a cluster-wide in-memory cache



AMAZON S3:

- Low Cost(0.023\$/Gb)
- Proven Availability of 99.99%
- Pay As you Go Model
- Owner & Bucket accessibility
- In & Out using Access key
- Scalable



Amazon S3



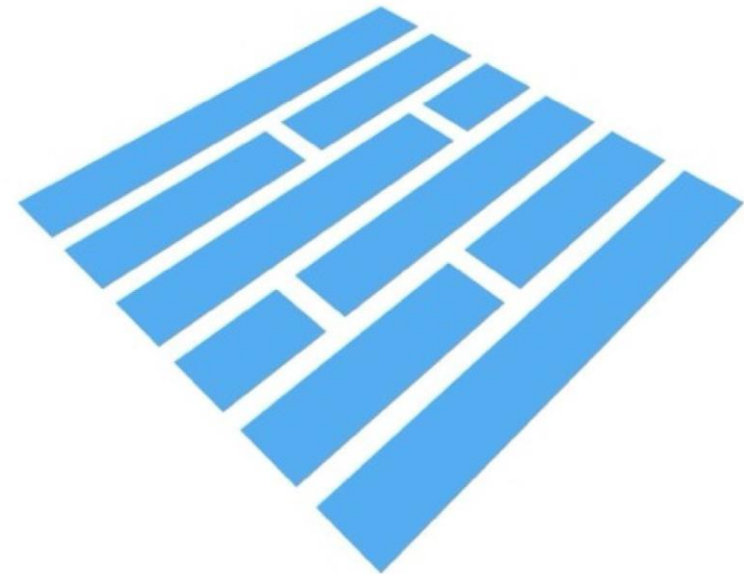
APACHE SPARK

- Data integration and ETL
- Interactive analytics
- Machine learning
- Advanced analytics
- Real-time data processing
- In-Memory computation
- Fault Tolerance



APACHE PARQUET

- Efficient columnar storage format compared to row-based storage files like that of csv or json.
- Flexible compression options
- Provides an efficient encoding system.
- Open source file format
- Available to any Hadoop file system.



DATABRICKS

- Free community edition
- Platform Connectivity
- Data Exploration
- Data Preparation
- Data Modeling
- Model Deployment
- Highly reliable
- Performant data pipelines



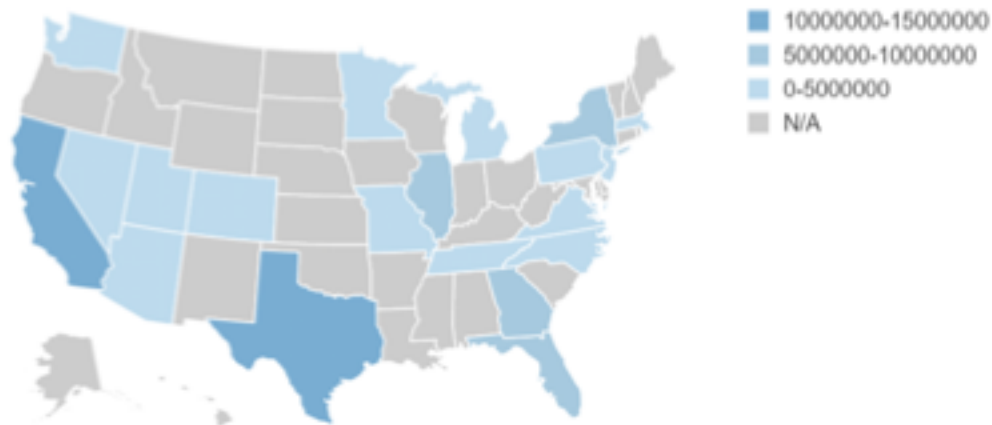
databricks



The background is a vibrant blue with various data-related icons. On the left, there's a red and white target with an arrow in the bullseye. Next to it is a computer monitor displaying a colorful donut chart. To the right of the monitor is a 3D pie chart with red, white, and grey segments. Further right is a 3D bar chart with three bars of increasing height. On the far right, there's a small bar chart with three bars in blue, green, and red. A large, complex network diagram with many nodes and connecting lines is positioned on the right side. The entire scene is filled with small white squares and dashed white circles.

EXPLORATORY DATA ANALYSIS





US Map displaying states mostly visited

- Highest bookings: California(15,425,336)
- Lowest bookings: Utah(2,413,319)

Monthly travel(Seasonal Influence)

- Most Travels: Summer [June and July]
- University move-in during September being next.



Factors influencing Delays

- Highest Delay: Aircraft delays with over 450Million
- Least Delay: Due to Security negligible records.

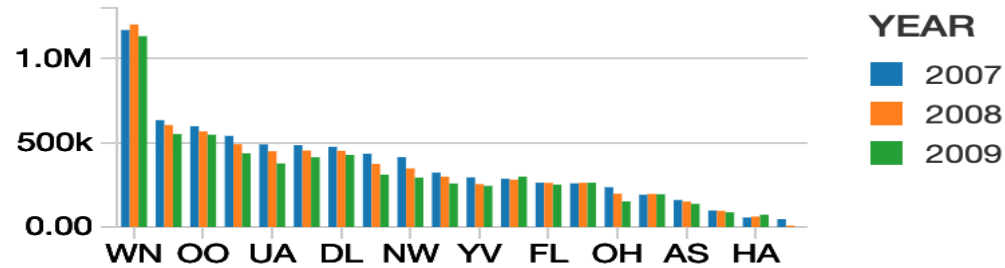




IMPACT OF GLOBAL RECESSION(2008)ON US FLIGHT INDUSTRY



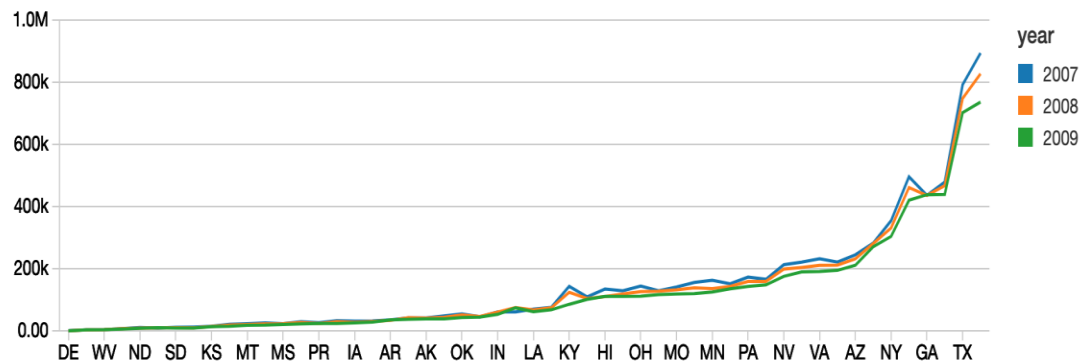
Count of Flights by Airline Carrier.



Month-wise analysis of the Global recession period.

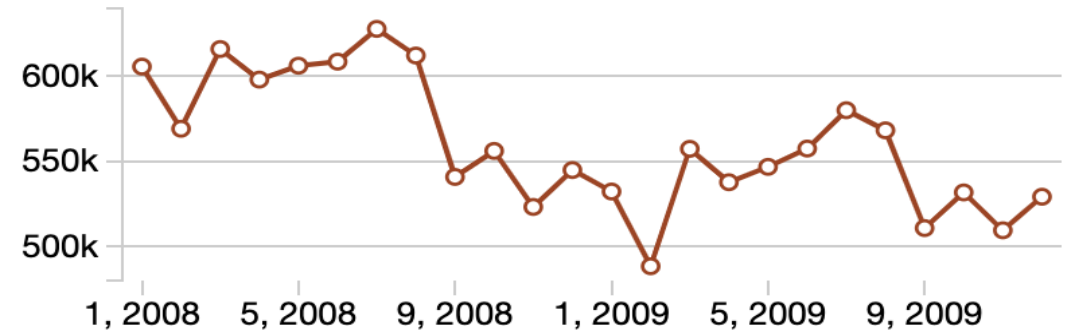
November 2008 and Jan 2009 observed a major fall in the number of flights.

State wise frequency of the flights journeys.(2007-2009)



Identifying the trends in the Frequency of Flights for the Airline Carriers from 2007-2009.


Dip in the flights in 2008-09



Fall in flights Journeys in the big states of US.

- 1.California
- 2.Texas
- 3.Illinois
- 4.Florida
- 5.New York



The background features a dark blue field with various geometric elements. There are several thick, orange, 3D-style rectangular bars of varying heights and orientations. A prominent teal-colored jagged line, resembling a stylized mountain range or a fluctuating data line, runs across the upper portion. A grey, textured banner with a grid-like pattern is positioned horizontally across the middle. The text is centered on this banner.

COVID-19 IMPACT ON US FLIGHTS





Jan-Feb 2019



Jan-Feb 2020

ANALYSIS

- Decrease in cancellation
- No significant changes
- Insufficient data

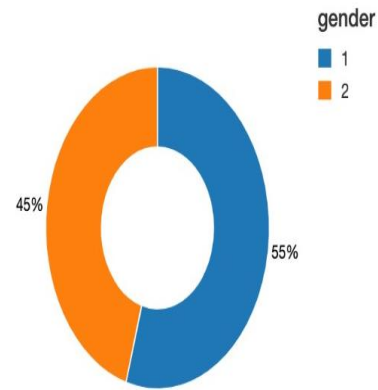




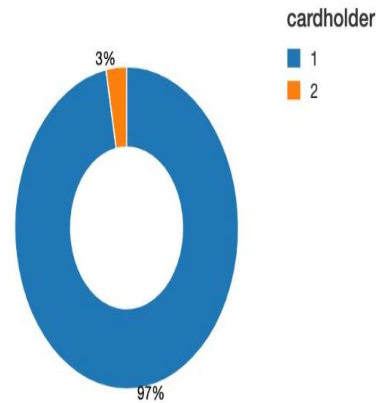
FRAUD DETECTION



Fraudulent Risk in 1.male and 2.female.



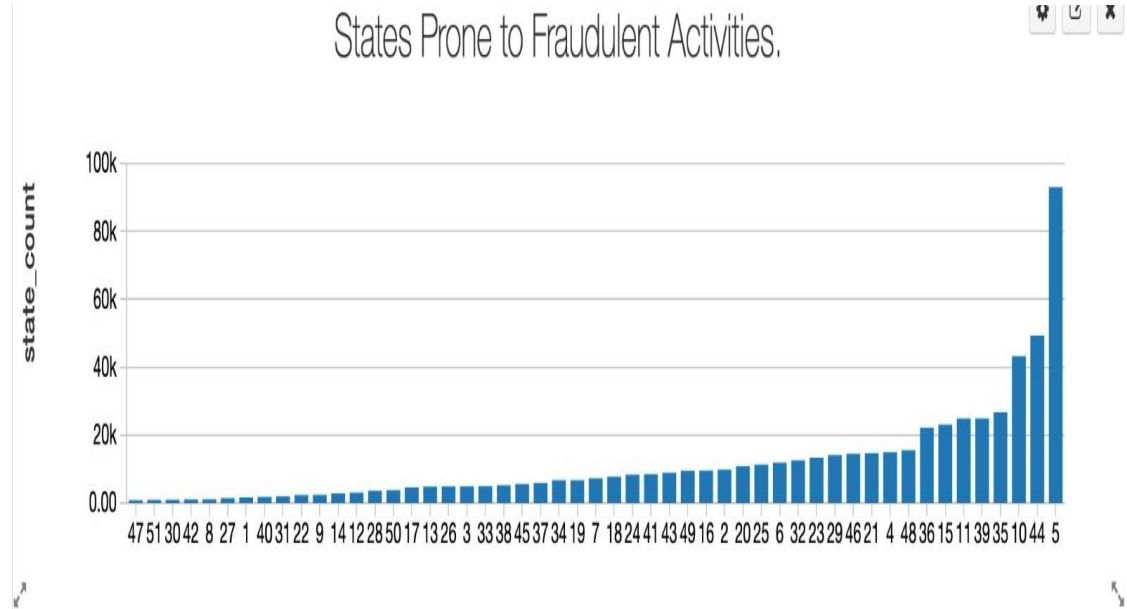
Cardholders with are fraudulent cases.



Identify user characteristics to anticipate chances of fraud

States of California, Texas, Illinois, New York and Arizona are the major states with maximum number of cases with risk of fraudulent activities

States Prone to Fraudulent Activities.



CONCLUSION

- AWS S3 provides a high-performance platform data lakes storage to handle 128 millions data entries
- AWS S3 not only has "Pay as You Go" model, but also has high security feature that which user can access its data
- Integration between Apache Spark and AWS S3 can be achieved using Databricks Community Edition by using Databricks File System, instead of HDFS
- Spark enables a user to optimize the existing queries using repartitioning



REFERENCES

EckersonFebruary, W. (n.d.). Which Big Data Platform Is Right For You? Retrieved from <https://tdwi.org/articles/2016/02/05/which-big-data-platform.aspx>

Top 5 Reasons for Choosing S3 over HDFS - The Databricks Blog. (2020, April 29). Retrieved from <https://databricks.com/blog/2017/05/31/top-5-reasons-for-choosing-s3-over-hdfs.html>

Understanding Amazon S3. (2019, May 22). Retrieved from <https://www.edureka.co/blog/understanding-amazon-s3/>

Author, V. (2020, May 14). 5 Benefits of using Amazon S3 vs your own server for hosting images/videos. Retrieved from <https://www.vizteck.com/post/5-benefits-of-using-amazon-s3-vs-your-own-server-for-hosting-imagesvideos>

Comparing Databricks to Apache Spark. (n.d.). Retrieved from <https://databricks.com/spark/comparing-databricks-to-apache-spark>

Big Data - Definition, Importance, Examples & Tools. (2019, September 5). Retrieved from <https://www.rd-alliance.org/group/big-data-ig-data-development-ig/wiki/big-data-definition-importance-examples-tools>

Unified Data Analytics. (n.d.). Retrieved from <https://databricks.com/>



Discussion Questions



THANK YOU

