US Flight Delay Trend and Prediction

Department of Applied Data Science, San Jose State University

DATA 228: Big Data Technologies & Applications

Professor: Andrew H. Bond

May 22, 2022

By,
Team Data WatchDogs

Purva Badve - 016007797
Radha Krishna Boddu - 015946853
Manyam Soumith Reddy - 016000062
Varun Jaiswal - 016023397

Copyright ©2022

[Purva Badve, Radha Krishna Boddu, Manyam Soumith Reddy, Varun Jaiswal]

ALL RIGHTS RESERVED

Table of Contents

Abstract —------ 4

1. Introduction
1.1 Project Background —--- 5

1.2 Project Goals and Motivation —----- 5

1.3 Project Application and Impact —----- 5

1.4 Expected Deliverables —------ 6

2. Background and Related Work —------ 6

2.1 Literature Review —------ 6

3. Data Preparation —-------- 6

3.1 Data Description —--------- 6

3.2 Data Cleaning and Exploration —------- 7

4. Architecture —-------- 10

4.1 AWS S3 —-------- 10

4.2 AWS Glue —-------- 11

4.3 AWS Redshift —---------11

4.4 Tableau —-------- 11

4.5 Flask —-------- 11

4.6 EC2 —-------- 12

5. Security Management Services —---------12

5.1 Identity And Access Management —-------- 12

5.2 Virtual Private Cloud —-------- 12

5.3 AWS Secret Manager —-------- 12

5.4 Endpoint —-------- 13

6. Implementation —-------- 13

6.1 Web Application —---------19

7. Data Visualization —-------- 19

8 Conclusion —-------- 24

9 Future work —-------- 25

10 References —-------- 25

Abstract

Flight delays are an inevitable aspect of flying. Your plane may arrive late, there may be a line for departure, or the weather may impose hour-long delays (or even cancellation); whatever the reason for the delay, it is a major inconvenience for passengers. Common issues arise as a result of flight cancellations and delays. It has an impact on daily air traffic operations, as well as time delays, disgruntled passengers, and financial losses.According to the United States Bureau of Transportation Statistics (BTS), more than 20% of domestic flights were delayed in 2018. These flight delays have a significant economic impact in the United States, amounting to $40.7 billion each year. Passengers lose time, miss business opportunities, and airlines' attempts to compensate for delays result in increased fuel usage and a greater negative environmental impact. As a result, given the large amount of data on flight journeys (16+ million flights annually in the United States alone), we may acquire significant insights from this data to better understand flight delays. Furthermore, given the availability of data, machine learning models can be trained to predict these delays, which might be extremely useful for both individual travelers and corporations.

Keywords: US Flight Delay, Airline, NAS

1.Introduction

1.1 Project Background

We know how crucial transportation is in globalization and the growth of civilization, and airways have made great contributions in this regard.

Flight cancellations and delays are frequent occurrences. It has an impact on daily air traffic operations, as well as time losses, disgruntled passengers, and financial losses. We will focus on parameters that cause flight cancellation and delays in this study, which will be limited to passenger aircraft. The dataset was chosen from kaggle and is approximately 1.3 GB in size, with detailed information on airlines, weather, and airports for the year 2019.Many fields are included in the database, including flight delays by month, weekday, number of seats in the delayed aircraft, and carrier name.

Aviation can be divided into two categories.

1 Passenger Aircraft

2 Cargo Aircraft

In this project we are focusing particularly on passenger aircrafts

1.2 Project Goals and Motivation

a) With this project we analyzed and visualized the following:-

   1) Cities with highest number of flight delays

   2) Trend in the number of arrival delays over the year

   3) Which airports are best and worst for on time departures.

b) We also built web application which will predict the probability of flight delays flying from Airport A to Airport B

1.3 Project Application and Impact

This project is targeted to any person or organization who is looking forward to research on the delay parameters affecting flight delays. Flight delays are unpredictable as weather conditions

have been a contributing factor to flight delays. Our project takes historical data into consideration and based on it we predict the delay and we have also visualized the trend in the flight delays.

1.4 Expected Deliverables

The project's deliverables include the creation of an AWS pipeline for processing large amounts of flight data.The Data is preprocessed before ingesting into the pipeline the data is fed into cloud services, the pipeline is run, and the visualizations provide instantaneous insight. Furthermore, a machine learning model is created and a web app is created to predict delay. A detailed report with screenshots that details the entire pipeline.

2. Background and Related Work

2.1 Literature Review

Researchers and analysts are mostly interested in forecasting the reasons of aircraft delays, thus they've concentrated their efforts on obtaining flight and meteorological data. At Orlando International Airport, Mohamed et al. [2] looked into the pattern of non-stop domestic flight arrival delays. They focused on the weather at that specific airport as well as the cyclic fluctuations in air travel demand. [3] Shervin et al. claim that this can be done without compromising or impacting the budget. Adrian and colleagues [4] developed a data mining methodology that allows for flight delays. Their research goal is to offer a way for improving operational performance while staying within budgetary restrictions. Adrian et al. [4] created a data mining program that can anticipate flight delays depending on weather variables. They constructed

3. Data Preparation

3.1 Data Description

For this project, we have used two data sources for our analysis. It consists of data of fifteen months. The first data source is “2019 Airline Delays w/Weather and Airport Detail” on the Kaggle website. The Dataset name is full_data_flightdelay.csv. It has data of the US flights from January 2018 to March 2019 pre pandemic, this dataset has a total of 30 columns and 700 thousand records. The fields of the dataset are provided below.

MONTH:Month
DAY_OF_MONTH:Day of the month (1-31)
DAY_OF_WEEK:Day of the week
OP_UNIQUE_CARRIER:Carrier code
TAIL_NUM:Unique tail number,
OP_CARRIER_FL_NUM:Flight number
ORIGIN_AIRPORT_ID:Airport ID,
ORIGIN:origin airport abbreviation
ORIGIN_CITY_NAME:Origin city name
DEST_AIRPORT_ID:Destination airport ID
DEST:Destination airport abbreviation
DEST_CITY_NAME:Destination city name
CRS_DEP_TIME:Planned departure time
DEP_TIME:Actual departure time
DEP_DELAY_NEW:Departure delay in minutes
DEP_DEL15:TARGET VARIABLE Binary if delayed over 15 min, 1 is yes
DEP_TIME_BLK:Departure time block
CRS_ARR_TIME:Planned arrival time
ARR_TIME:Actual arrival time
ARR_DELAY_NEW:Arrival delay in minutes
ARR_TIME_BLK:Arrival time block
CANCELED:Flag if flight was canceled
CANCELLATION_CODE:Cancellation Code
CRS_ELAPSED_TIME:Flight planned elapsed time
ACTUAL_ELAPSED_TIME:Flight actual elapsed time
DISTANCE:Flight Distance in miles
DISTANCE_GROUP:Flight distance group
CARRIER_DELAY:Flag for a carrier delay
WEATHER_DELAY:Flag for a weather delay
NAS_DELAY:Flag for a NAS delay
SECURITY_DELAY:Flag for a security delay
LATE_AIRCRAFT_DELAY:Flag for a late aircraft delay.

3.2 Data Cleaning and Exploration

Kaggle.com and USCensus.com provided the data for this project. For feature extraction and data quality checks, datasets are rigorously inspected. The raw datasets needed a lot of cleaning and preparation because they were unclean and unprepared. Data preparation techniques include categorizing linear data into discrete categories and keeping categorical variables. The Flight Delays dataset came from kaggle.com, while the airline and weather information came from the US Census. There are many missing entries, null values, and unnecessary fields in the Flight Delays dataset. Because the datasets comprise textual, numerical, null, and zero-valued information, they will need to be modified and cleaned. Tableau Prep Builder, JupyterLab, and AWS Glue were used for data cleansing and preparation. In the first stage, Tableau Prep Builder was used to clean, transform, split, and merge the datasets. Tableau is a software program that makes data preparation simple. Tableau Prep Builder is a tool for preparing data for analysis by aggregating, organizing, and classifying it. At this point, We first separated the origin and destination fields because they contained the city name and state. We then cleaned the flight delay dataset, converted the datatypes of the fields, and assigned geographical responsibilities to the city and state fields. Using the column to row transformation, the weather and carrier datasets were pivoted, and geographical duties were assigned to the city and state fields, as well as city and state/region. We used the Jupyter Notebook for exploratory data analysis and data preparation in the second stage. We've done data analysis with histograms and scatter plots, removed unneeded and repetitive columns, and substituted null values with mean, median, and standard deviation at this point. In four columns, there are no values. Because the number of missing values is small, they may be in the same rows. To make sure, we used the missingno module in Python to show missing values.

FIGURE-1

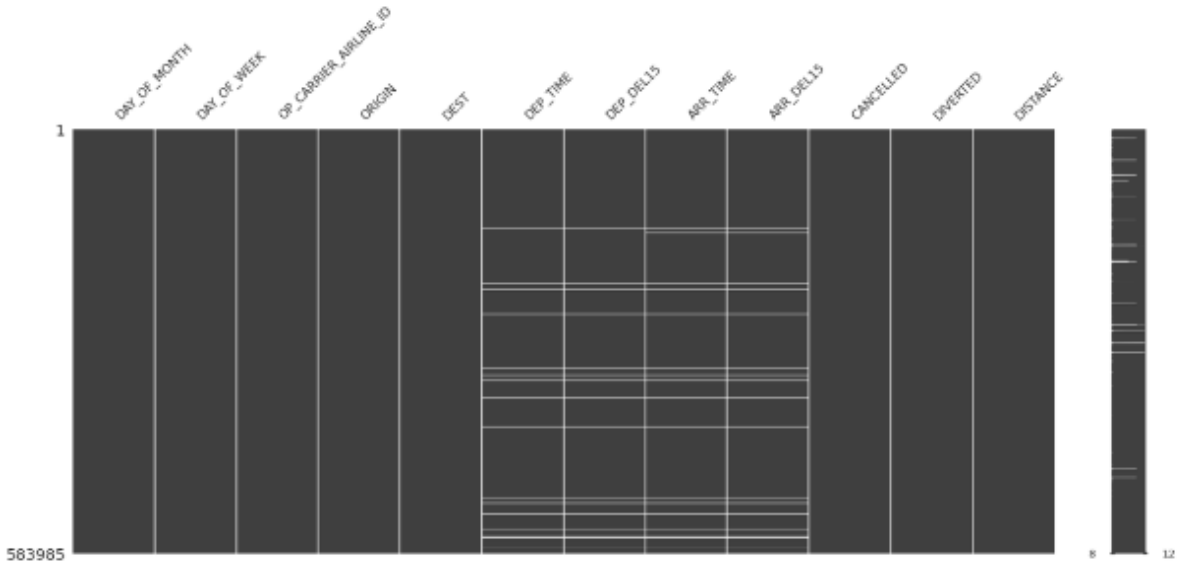


FIGURE-2

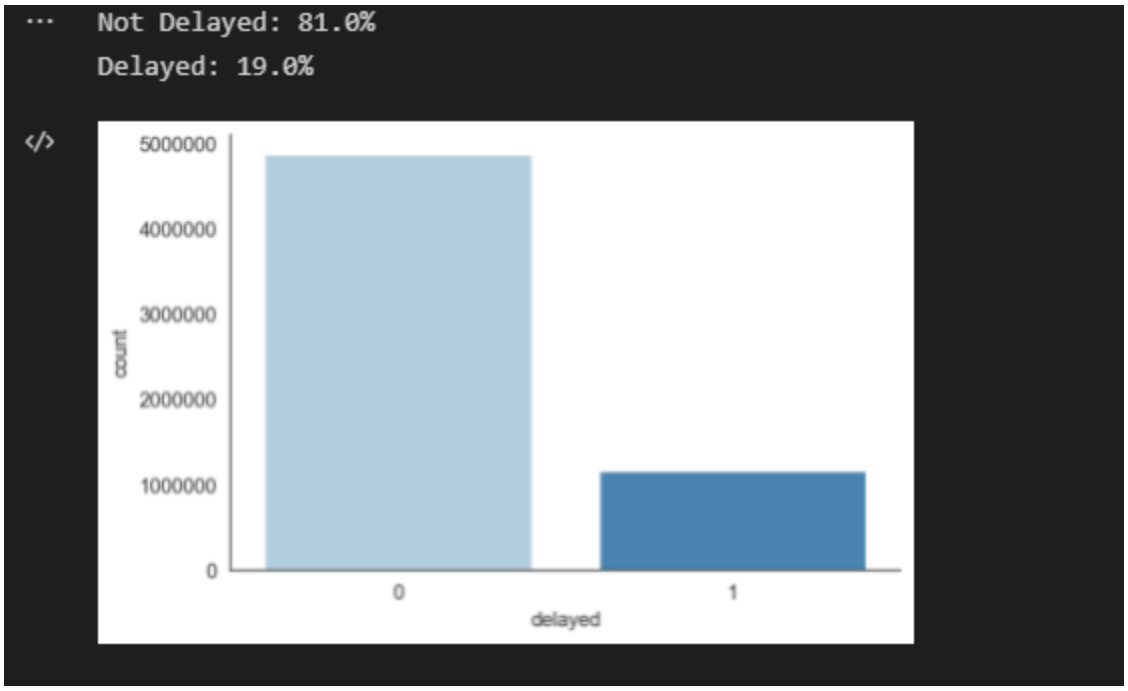
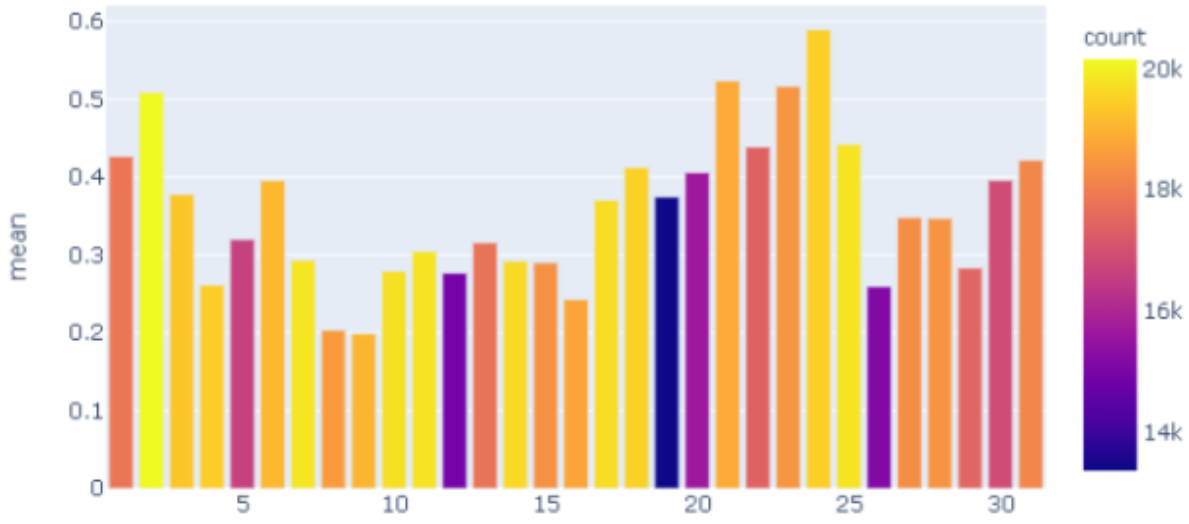


FIGURE-3

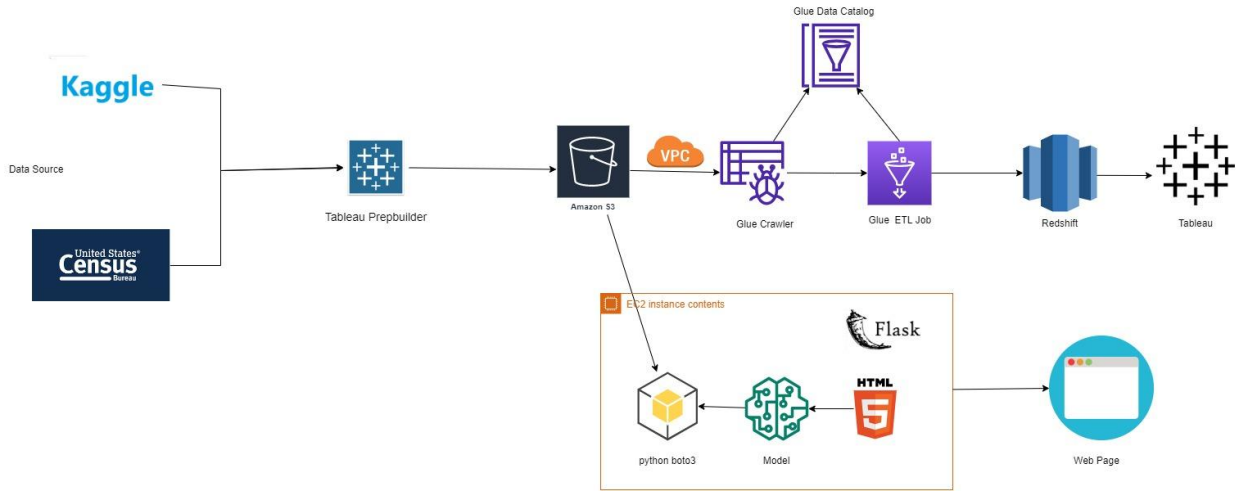


Based on the count scale on the right side, the color of the bar shows the number of flights every day. The delay ratio is represented by the height of the bar. In terms of days, there is no clear trend, but days with more flights have greater delay percentages.

4. Architecture

Our project's data flow diagram is shown in figu re 4. It comprises the many stages of the project, beginning with data collecting and ending with visualization. We acquired data from kaggle.com and USCensus.com in the first stage and used Tableau Prep Builder to normalize and clean the data sets. The data is then uploaded to an Amazon Web Services S3 bucket. Then, the required IAM roles are formed and then ETL Glue procedures, Crawlers are used to delete null values, update, and convert the data. The data is then uploaded to Redshift via AWS Glue, where the data is integrated in the desired manner. Tableau (Data Visualization), which is connected to Redshift, is used for the majority of the data analysis. We used the Randomforest classifier to train a model that predicts the likelihood of a flight delay, and we used Flask to develop a website to deploy our model over the internet. Finally, we used an EC2 container to host this website.

FIGURE-4



4.1 AWS S3

S3 storage is a cloud storage service from Amazon Web Services that allows you to access data from any location. When generating deployment buckets, this path is chosen. AWS provides 'pay as you go' services, with prices based on the amount of data kept and the network's data transit capability. Each item in an S3 bucket has its own set of characteristics and permissions. The information in the S3 buckets will be used for future research. An item for backup and disaster recovery storage is also included in these buckets. To avoid any setbacks, both the production and backup data would be subjected to the same security restrictions, such as encryptions, permissions, and audit requirements. Applications and software are likewise hosted and delivered via S3 buckets.

4.2 AWS Glue

For analytics, machine learning, and application development, AWS Glue is a serverless data integration tool. All of the data integration tools you'll need to get started analyzing data are included in AWS Glue. When new data enters, AWS Glue may execute ETL procedures. As soon as fresh data in Amazon S3 becomes available, the ETL operations are run using an AWS Lambda function. As part of the ETL process, a new dataset is registered in the AWS Glue Data Catalog. It analyzes all datasets, identifies data types, and suggests storage options. Glue automatically develops the application required to perform data processing and loading tasks. It makes running and managing ETL procedures, as well as combining and duplicating data across many data storage, a breeze.

4.3 AWS Redshift

Amazon Redshift combines AWS technologies to give the most cost-effective pricing at all scales by examining structured and semi-structured data in data warehouses, operational databases, and data lakes using SQL. Amazon Redshift, the most popular cloud data warehouse, is well positioned to benefit from all three innovations and can be a great place to start monetizing or extracting important insights from data.

4.4 Tableau

Tableau is a data visualization application that is commonly used for Business Intelligence, but it has a wide range of features. It aids in the construction of dynamic graphs and charts for market intelligence in dashboards and spreadsheets. Tableau can analyze data from a range of heterogeneous data sources, such as databases, spreadsheets, big data, and cloud data, in real time. It can retrieve information from any site. may access information from a simple database

4.5 Flask

Flask is a microweb framework written in Python. It is classified as a microframework because it does not require any specific tools or libraries. It lacks a database abstraction layer, form validation, and any other components that may be found in third-party libraries. Extensions, on the other hand, can be used to extend the functionality of an application as if it were built into Flask itself. Extensions are available for object-relational mappers, form validation, upload handling, several open authentication protocols, and other framework-related features.

4.6 AWS Elastic Cloud Computing

Amazon Elastic Compute Cloud (EC2) is a component of Amazon.com's AWS cloud computing platform that allows users to rent virtual processors to run their own applications. Customers can use Amazon's EC2 web service to boot an Amazon Machine Image (AMI) and configure a virtual machine, referred to as a "instance" by Amazon, with any software they choose. The ability of a user to build, launch, and terminate server instances as needed is referred to as "elastic." The second charges are applied to active servers. Users can specify the physical location of their instances on EC2, allowing for latency optimization and high redundancy.

5 Security Management Services

AWS provides services to protect data, identities, and applications against cyberattacks. The following are some of the services that are used:

5.1 Identity and Access Management

The AWS Identity and Access Management (IAM) service allows for seamless access management throughout the whole AWS infrastructure. IAM enables you to control who has access to which resources and services, and under what circumstances. You utilize IAM policies and processes to control permissions to ensure the least amount of access to your personnel and systems. The data and services are protected because there is no overhead for tracking authorized and unauthorized users when using this service.

5.2 Virtual Private Cloud

In the virtual networking environment, Amazon VPC allows you complete control over resource placement, connection, and security. It's a logical container that isolates our resources from those created by other Amazon cloud customers. To interface with other AWS services, different services with essential IAM roles are placed within each container. However, by creating a barrier between users, the use of cloud services is hidden from one customer to the next. With Amazon VPC, you can have even more exact control over your cloud network, ensuring that your operations and data are protected. A user can manage network gateways and subnets as well as configure network parameters such IP address ranges, route tables, and network gateways.

5.3 AWS Secret Manager

Customers and other users can't see sensitive or non-exposable credentials thanks to the secret manager. Because anyone may tamper with the database data, which would have a huge impact on our conclusions, the database passwords are extremely sensitive credentials. The values are stored as a key-value pair in the secret manager service, which is encrypted and kept private.

These values are retrieved through an API that decrypts the information and links it to the appropriate service. In this project, we encrypt the database name, username, password, port number, and server where the database is stored. Other types of secrets, such as API keys and OAuth tokens, can also be managed by the service.

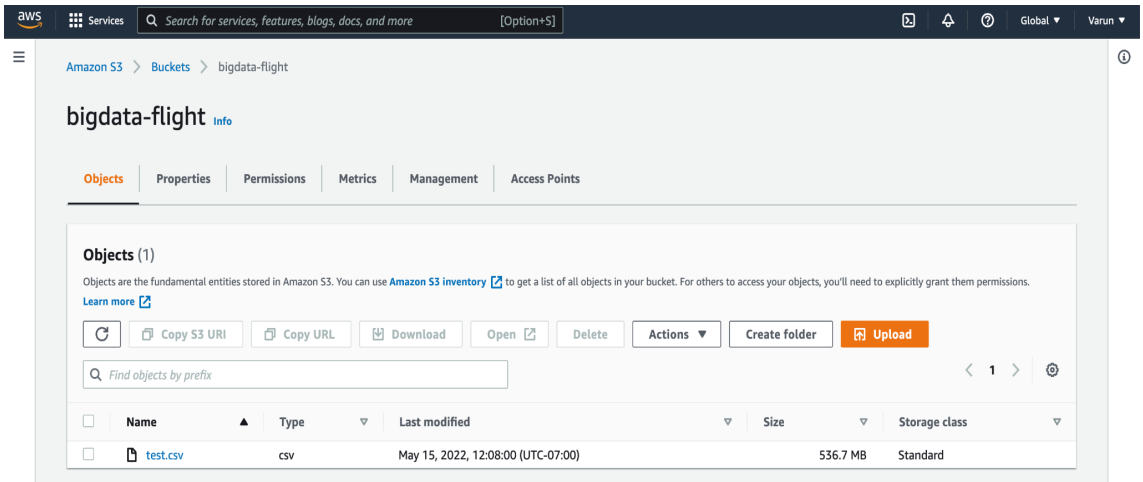
5.4 Endpoint

Endpoints were utilized to communicate amongst the VPC's services. They enable VPC time and goods to communicate while preventing errors. A gateway load balancer distributes users among a variety of digital devices. Auditing, adherence, cover tampering, and other community objects can all be protected using these devices. When you create an interface VPC endpoint, an Elastic Network Interface with a private IP address is deployed in your subnet. An Amazon EC2 instance in the VPC can connect to an Amazon S3 bucket via the ENI and AWS network. The interface endpoint allows apps on your on-premises network architecture to access S3 buckets using AWS Direct Connect or Site-to-Site VPN. The interface supports an ever-expanding list of AWS services.

6 Implementation

To begin with, Tableau Prep Builder is used to combine the data sets taken from kaggle.com and USCensus.com. The fifteen datasets are combined here using similar parameters such as departure_delay15, arrival city, departure city.. The.csv file with the name 'Flight_cleaned_final.csv' has been created. Following that, the data is transferred to Amazon S3 buckets for additional analysis. As demonstrated, this S3 bucket is named 'bigdata-flight' as shown in the figure 5.

FIGURE-5



A Virtual Private Cloud with several endpoints is then built to enable access to the Amazon buckets. A security group was built to connect the AWS glue to JDBC. We'll connect all of the

services using the default VPC, and we'll set inbound rules for connecting to JDBC for Redshift. Create an IAM role after establishing a glue crawler, selecting a source type, and adding a data storage. Create a database to hold the records before running the crawler. Before validating the connection, create a cluster in Amazon Redshift and assign the role that was established in the previous stage. When you connect to Redshift, create a new crawler in the data store, choose a connection, describe the role, and choose the destination.

FIGURE-6

Crawler created

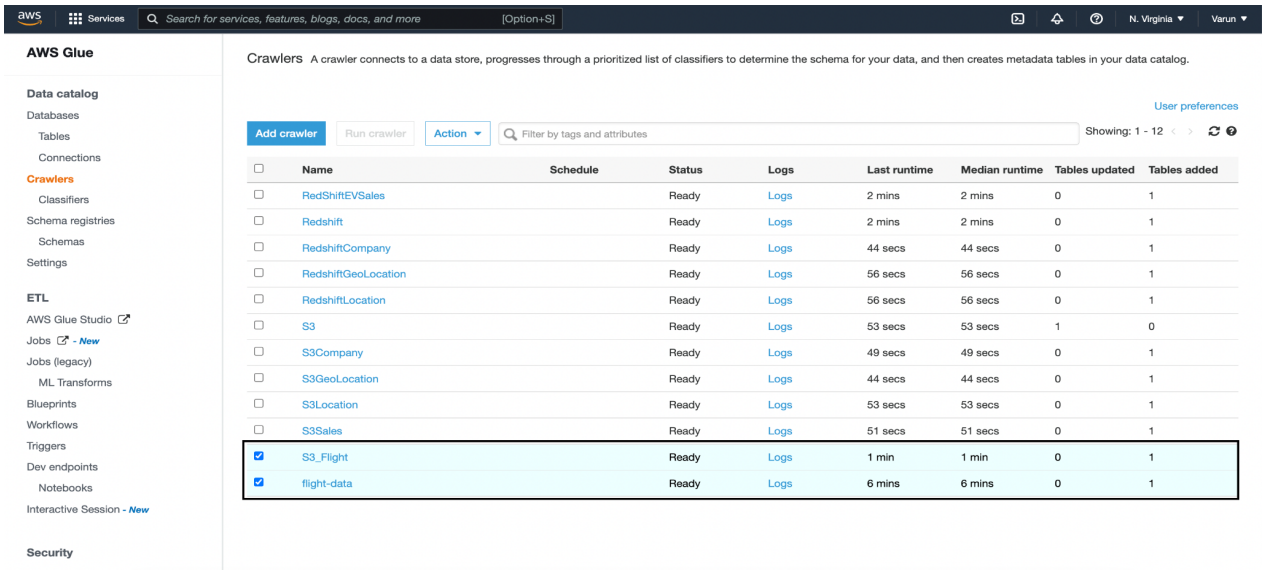


FIGURE-7

Configurations of the AWS crawler

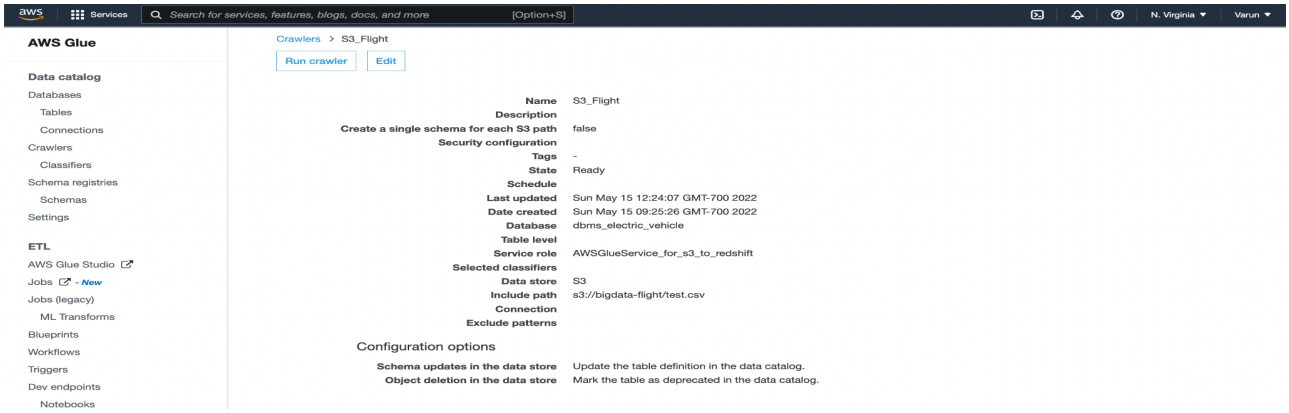
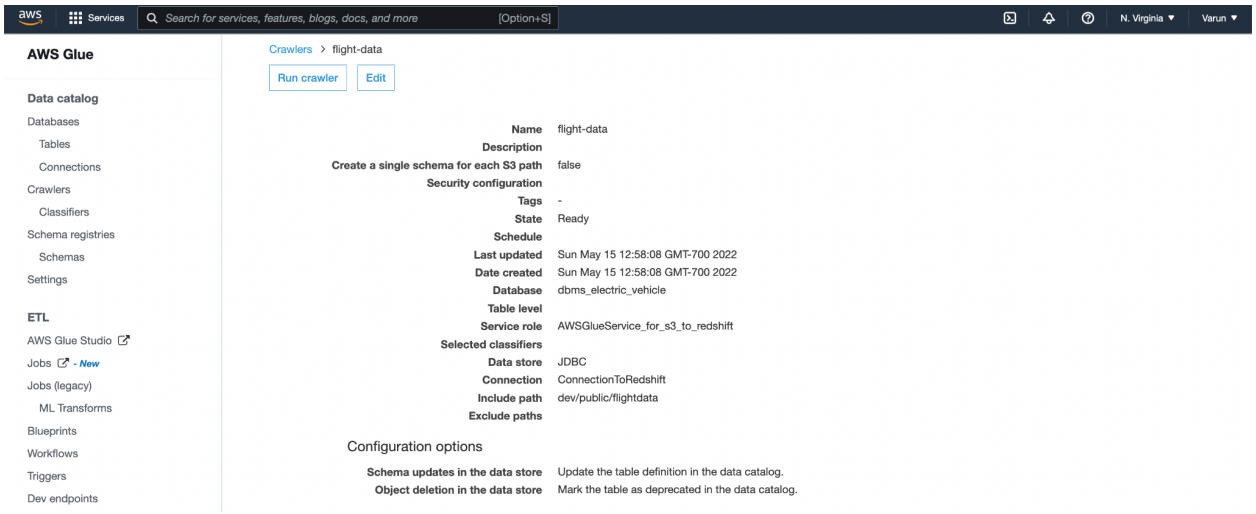


FIGURE-8



A crawler has also been constructed to create the table structure for the data in the S3 bucket. Use either the source or destination connection to create an instance job. For connecting the AWS glue to S3 and Redshift, two new connections, project-s3-glue-connection and project-s3-glue-redshift-job-connection, have been established. The metadata for AWS Glue tables is automatically built after the connection is established and the crawler is launched.

FIGURE-9

AWS Glue connections to the S3 and Glue

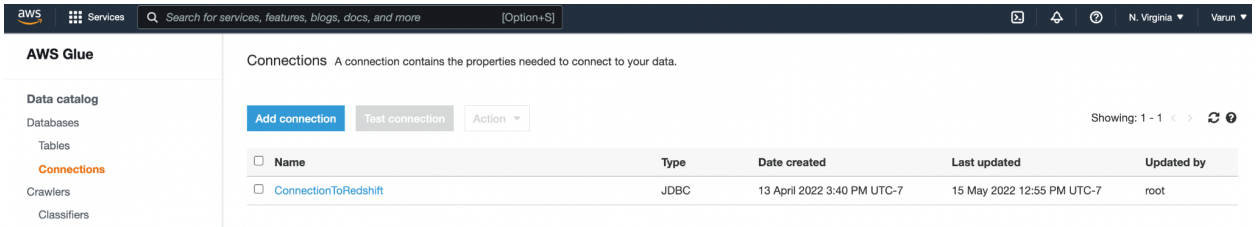


FIGURE-10

AWS Glue Tables

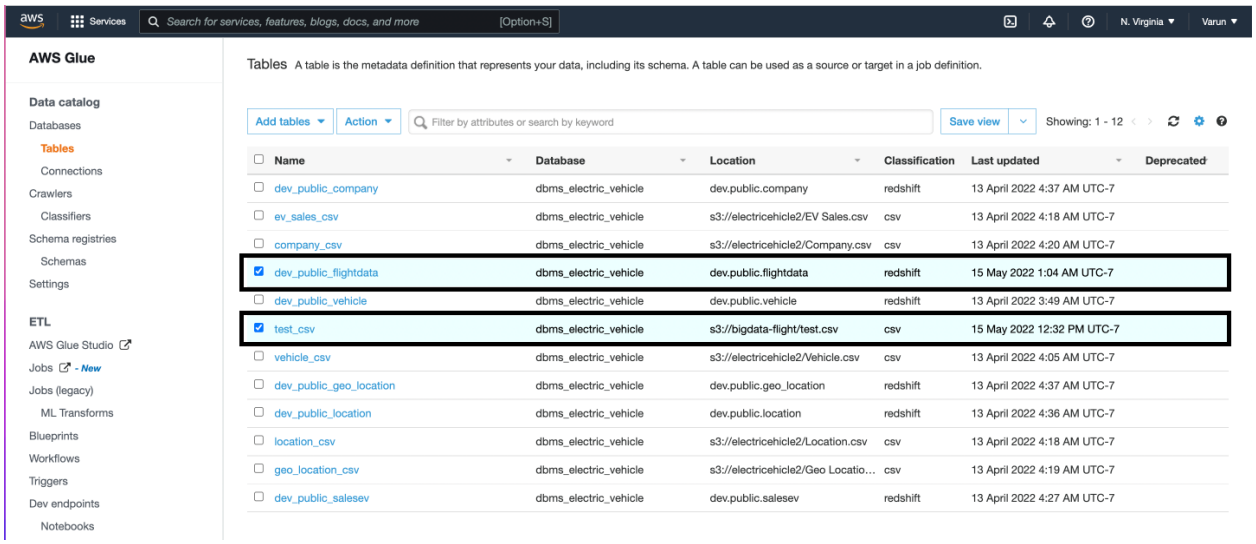
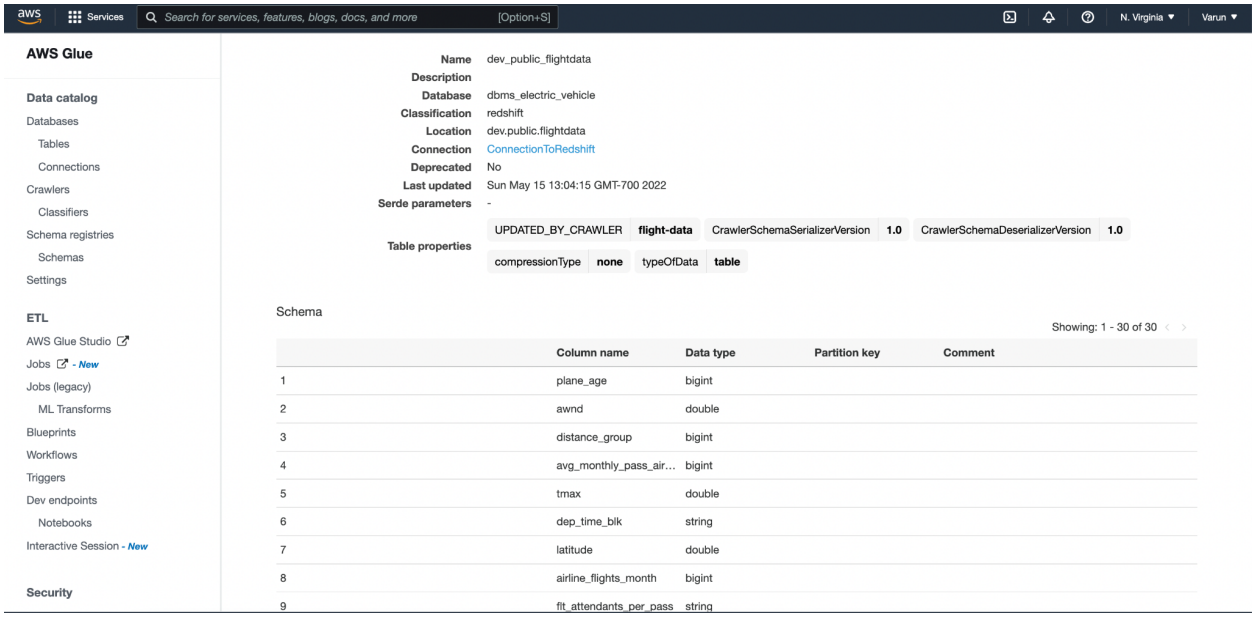


FIGURE-11

Table Metadata



Run the AWS job to move data from the S3 bucket to Redshift now. This will fill the table with the appropriate data. To clean and change the data, a pyspark script is used.

FIGURE-12

AWS Glue Job to put data from S3 to Redshift

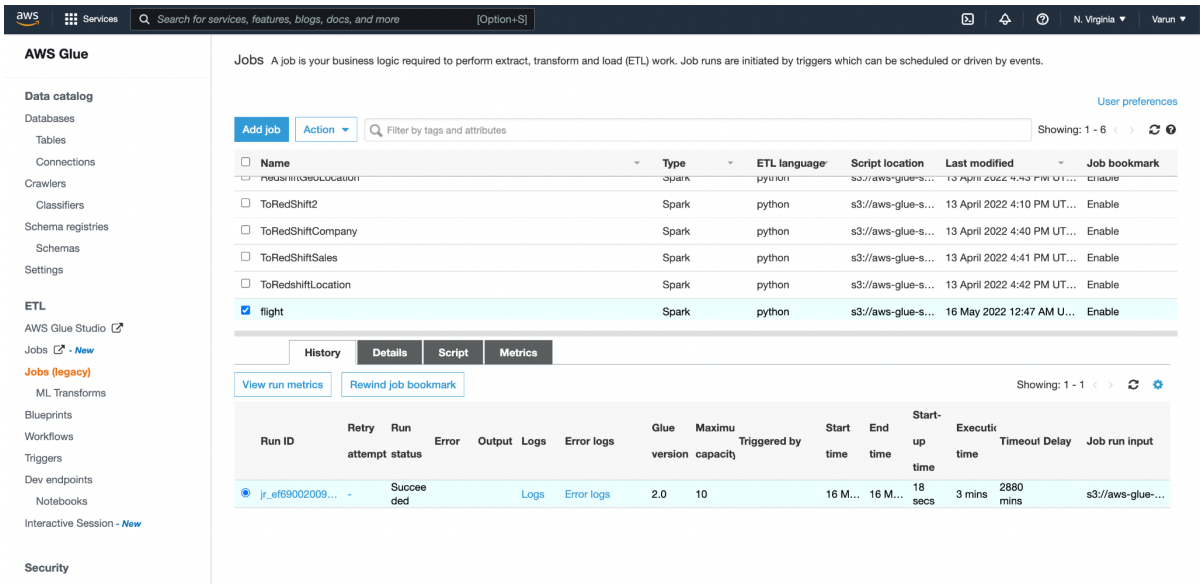
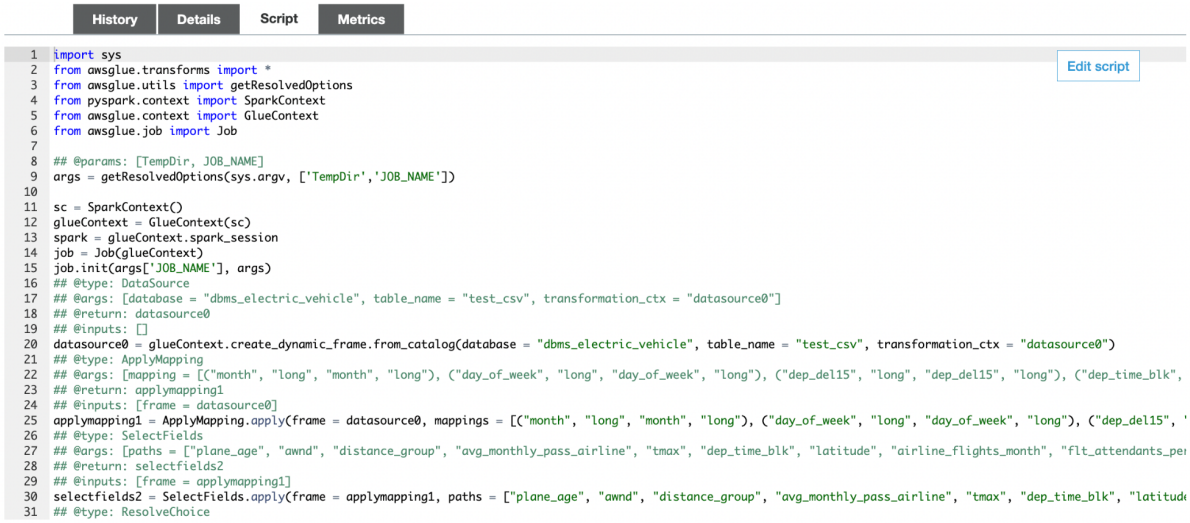


FIGURE-13

Pyspark Script



The 'project-cluster' cluster is created using the AWS Redshift service. AWS Redshift is a data warehouse with a cluster unit that consists of two components: compute node and leader node. The compute node has its own processor, memory, and storage. The data is sent to the 'dev' database using the account 'awsuser' and password once the process is completed.

FIGURE-14

Redshift cluster

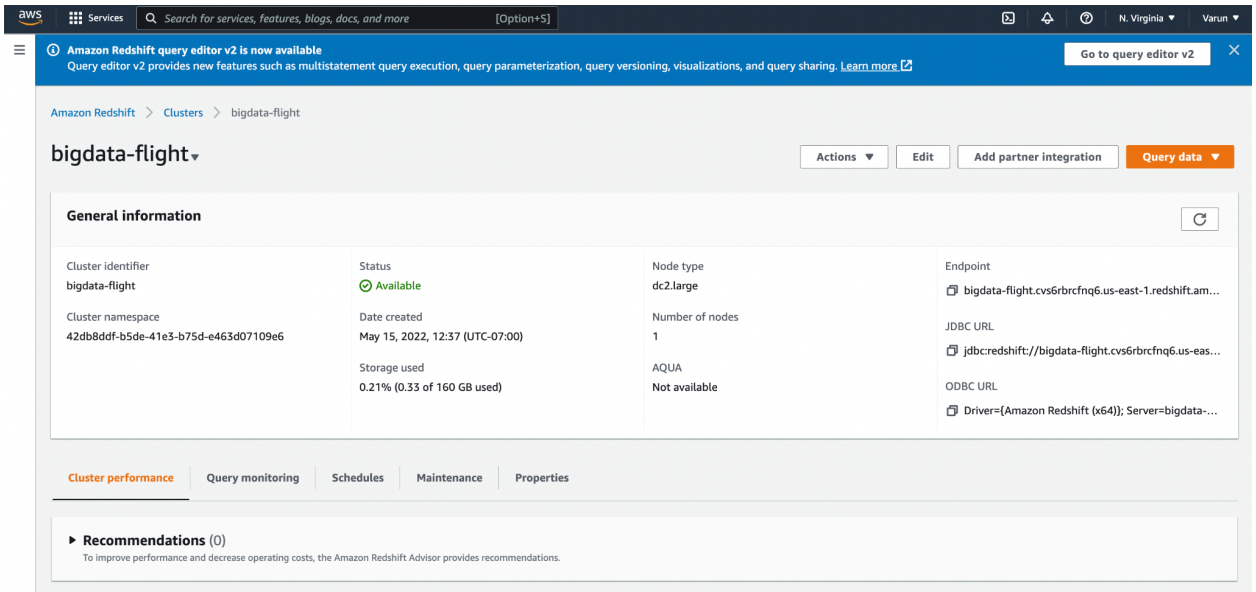
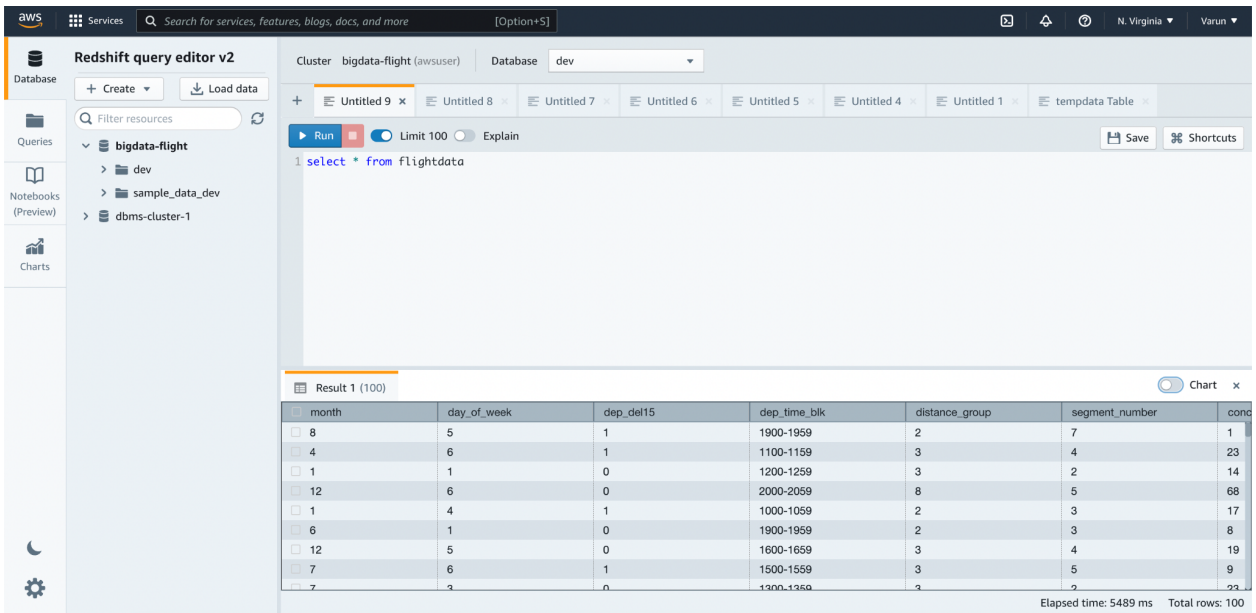


FIGURE-15



Finally, we constructed a visualization in Tableau by using the database name, server name, port numbers, login, and password to connect Tableau to Redshift.

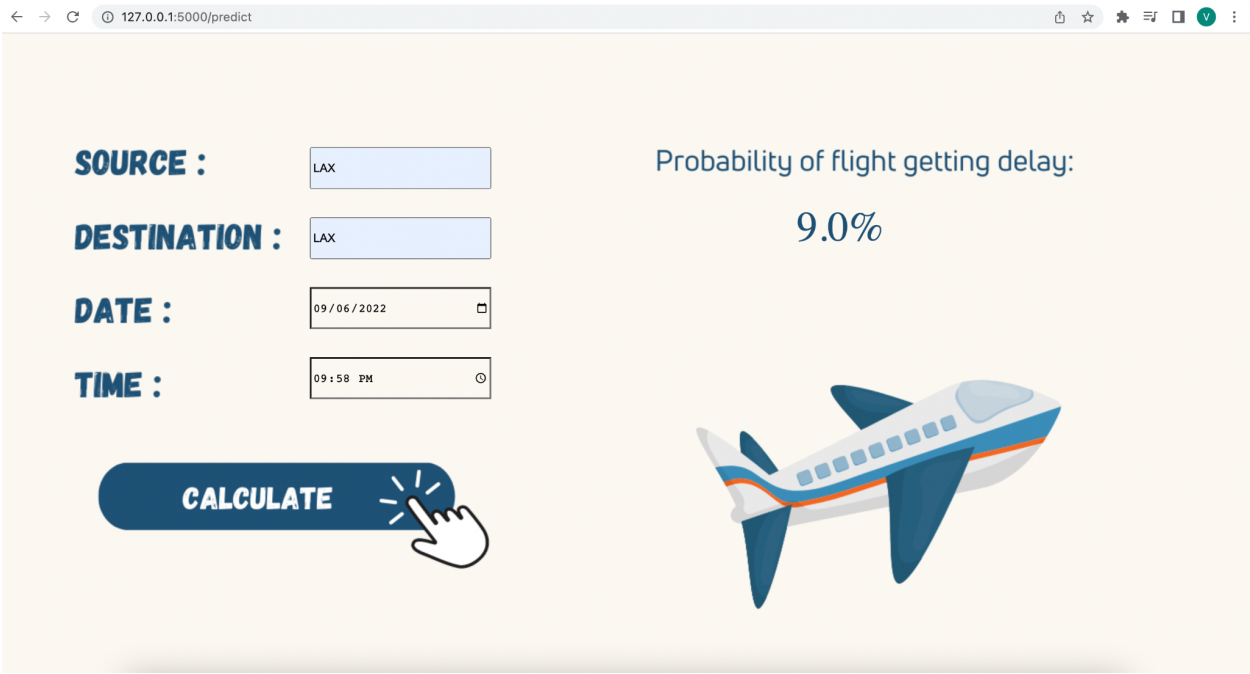
Web Application

Flask is a Python-based web framework. This means flask gives us the tools, frameworks, and technologies we need to create a web app. This web application might be as simple as a few web pages, a blog, or a wiki, or it can be as complex as a web-based calendar or a machine learning webapp.

Steps in creating web application

- Import Flask
- Create a Flask project
- Define a route
- Define the web resource for that route which in this case, is to render the HTML page and use one of the images that have been uploaded to the drive as a part of the site

FIGURE-16
The Web Application interface looks like this



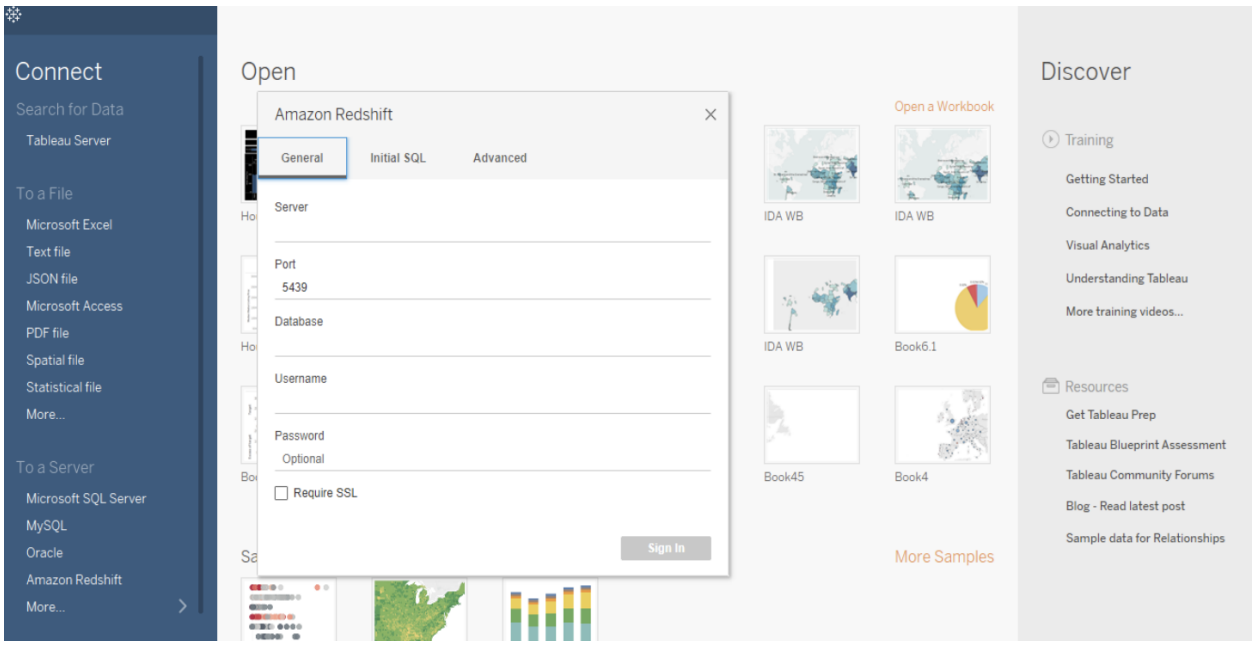
7. Data Visualization

At this point, we've used our dataset to construct some visualizations that show our goals in a graphical format. It's a method of condensing information and presenting it in a way that makes

it easier to analyze and spot outliers, trends, and patterns. Therefore, we used Tableau Desktop to visualize our findings. It's a program for displaying data. With the help of the Redshift Endpoint URL, we were able to connect the tableau desktop to a cleaned and processed dataset that had already been stored in AWS Redshift. To access the datasets available in Redshift, it requires the database name, port number, and password to assure the dataset's security and integrity. We've provided all of the details and used Redshift to retrieve the dataset in Tableau. After that, we double-checked that all fields' formats and data types were correct. Finally, we've generated a number of charts and graphs to help you get the most out of the data, which are available below.

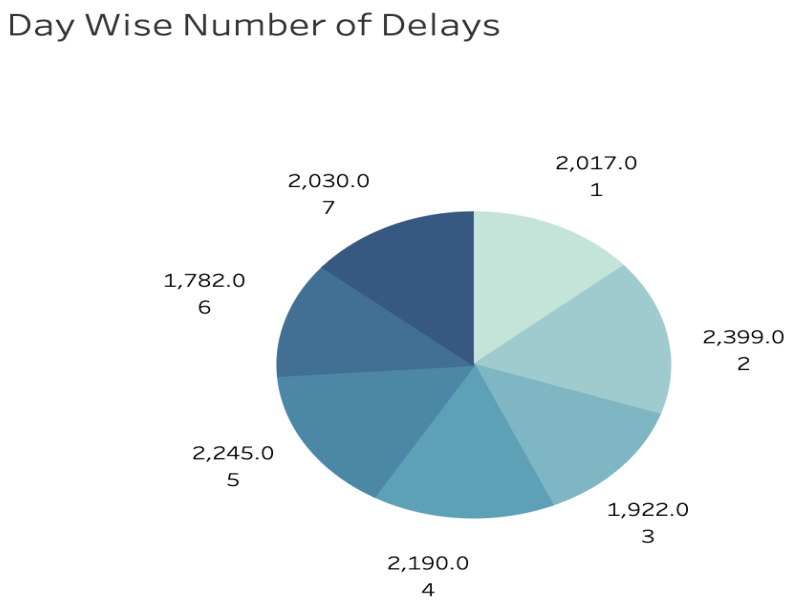
FIGURE-17

Connecting a Redshift dataset to Tableau Desktop



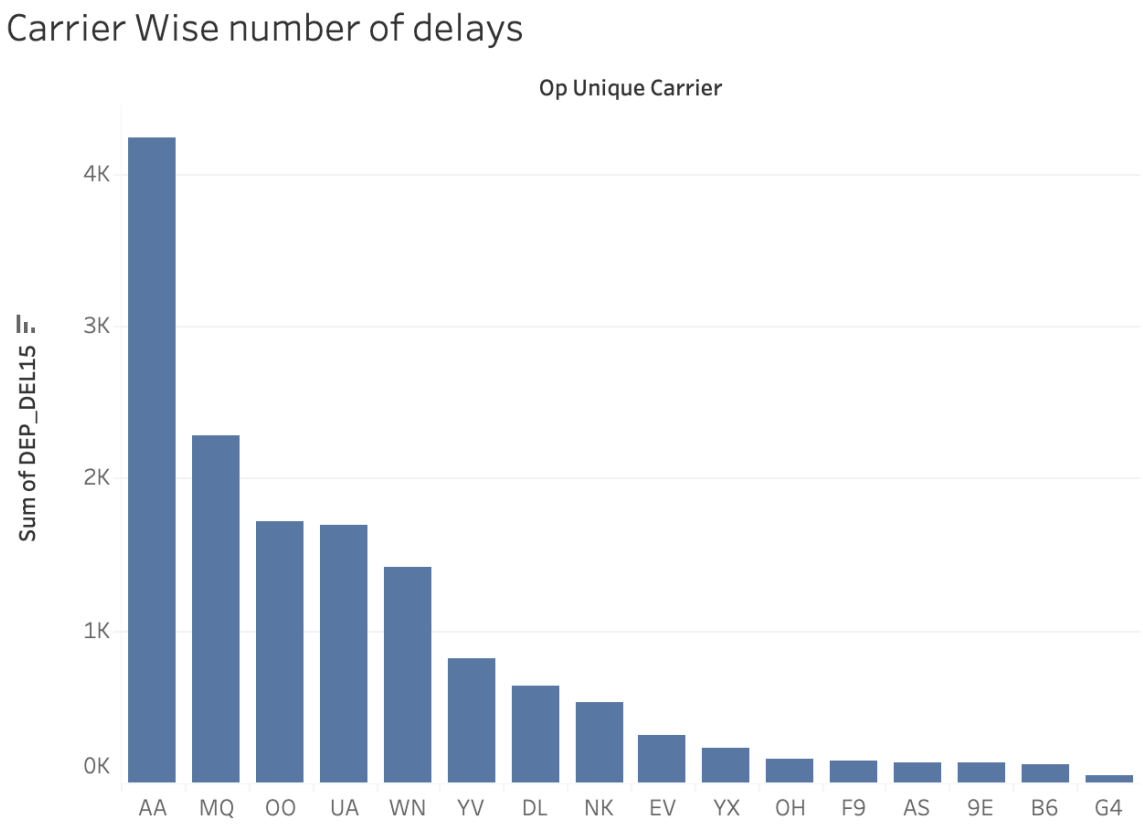
Some of the visualizations done for our project are:

FIGURE-18



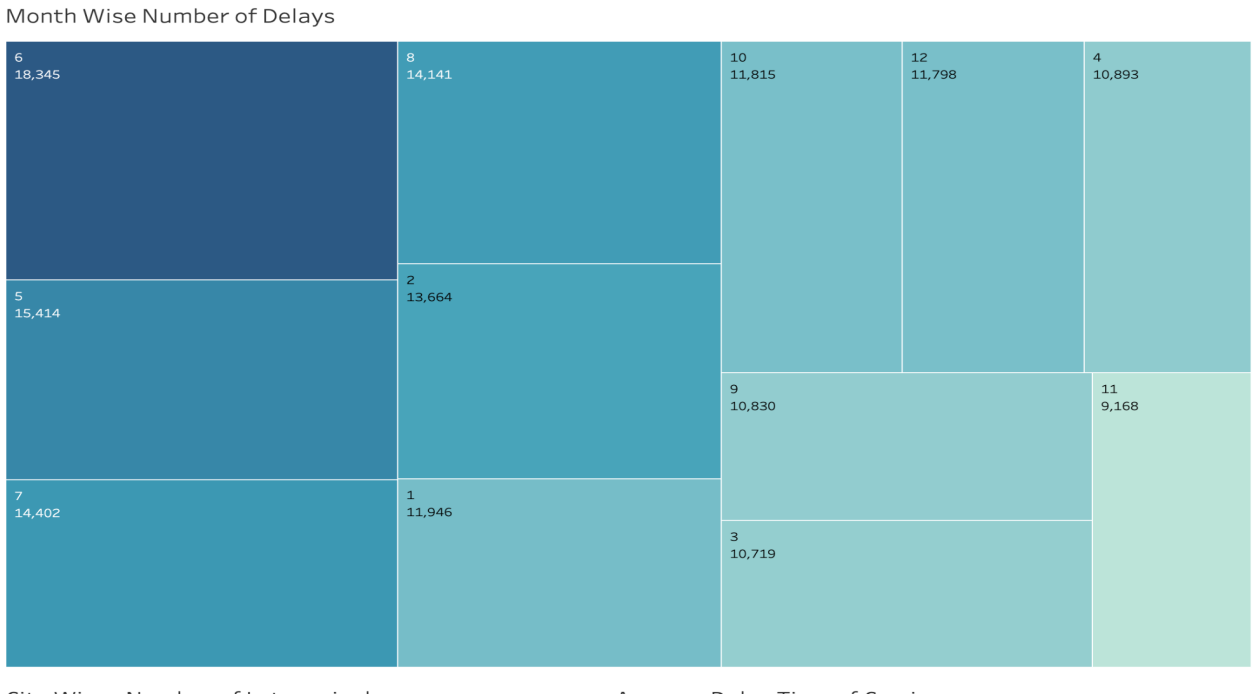
The above chart depicts the day wise number of flight arrival delays monday to sunday, 1 being monday to 7 being sunday. We have created global filters based on 12 months and flight carriers. So we can choose a particular month and we will get total delays of that month broken into day wise delays.

FIGURE-19



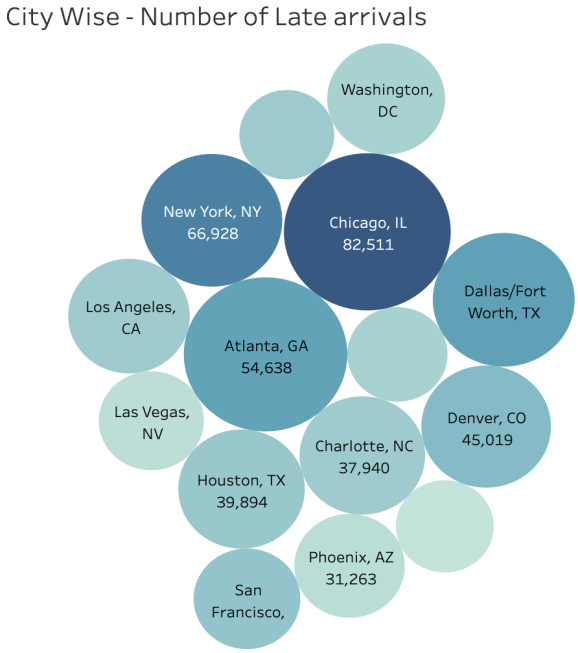
This chart shows the number of arrival delays based on carrier and as we have a global filter on month we can observe month wise delay count of each flight.

FIGURE-20



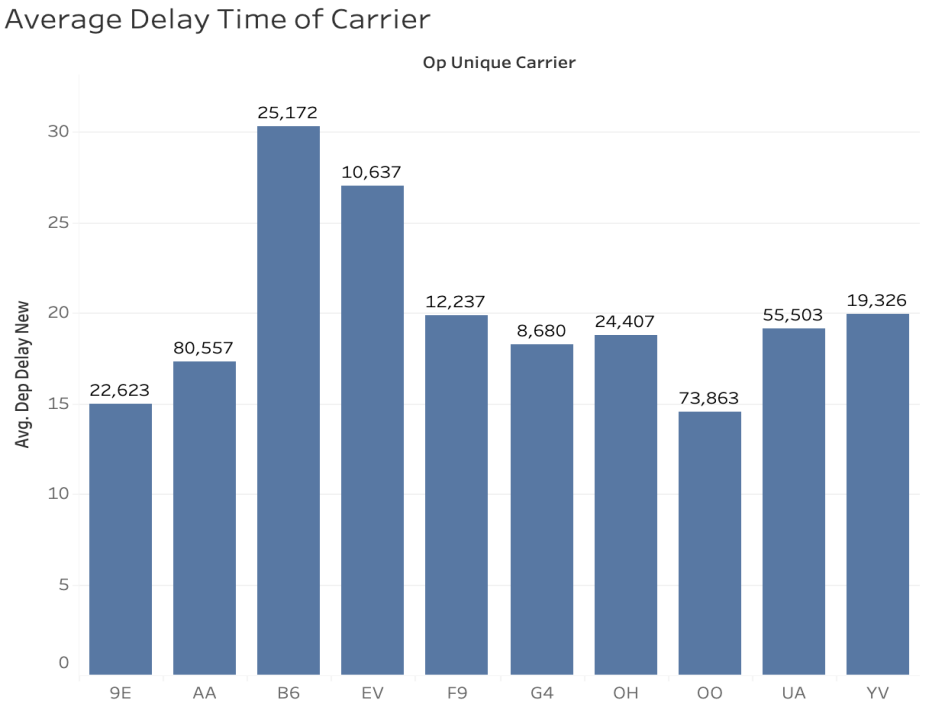
This chart represents month wise count of arrival delays and we could observe that the summer had the highest number of arrival delays starting from may to august.

FIGURE-21



This is a city wise count of the number of arrival delays and Chicago, Atlanta and New York has the highest number of arrival delays.

FIGURE-22



This chart represents the average delay time of each carrier and we observed a range of 15 minutes to 30 minutes with an average around 23 minutes.

Conclusion:-

By constructing an information investigation pipeline, we sought to evaluate and estimate US Flight Trends in order to investigate and inspect flight delays for the year 2019 and the first three months of 2020. For intake, readiness, stockpiling, and perceptions, several AWS administrations were used. Attempts were made to grasp the clump information in order to determine whether the aircraft will be delayed. Some of the observations were that the airports with the most arrival delays were Chicago and New York, with an average delay time of about 15-30 minutes, etc. Any novice may use our estimates to determine the likelihood of airline delays and plan their journey appropriately.

Future Work:-

The project may be improved to handle streaming data more easily and to give faster insight into real-time data. In order to estimate delay more precisely, the model should be trained with additional data points utilizing HPC. We also want to enable Interactive Chatbots in the web app to communicate delay specifications. Change from in-memory ML libraries to GUP-based libraries such as TensorFlow, PyTorch, and others for better processing. We would like to expand our initiative from the national to the global level, taking into account other variables that affect flight delays such as weather, air traffic, and so on.

References -

[1] A. B. Guy, "Flight delays cost $32.9 billion, passengers foot half the bill". [Online] Available : https://news.berkeley.edu/2010/10/18/flight_delays/3/. [Accessed on June 2017].

[2] M. Abdel-Aty, C. Lee, Y. Bai, X. Li and M. Michalak, "Detecting periodic patterns of arrival delay", Journal of Air Transport Management,, Volume 13(6), pp. 355– 361, November, 2007.

[3] S. AhmadBeigi, A. Cohn and M. Lapp, "Decreasing Airline Delay Propagation By Re-Allocating Scheduled Slack", Annual Conference, Boston, 2008.

[4] A. A. Simmons, "Flight Delay Forecast due to Weather Using Data Mining", M.S. Dissertation, University of the Basque Country, Department of Computer Science, 2015.