IE 7374 - Cloud Project: Aviation Accident Analysis

Group 10 - Subhasree Vemprala Sathyanarayanan and Stuti Dhebar

Problem Definition

Over the years, the aviation industry has witnessed tragic accidents, claiming hundreds of lives due to various errors. These are primarily caused by factors such as pilot error, weather conditions, mechanical failure, fuel mismanagement, and more. As a result, analyzing air accident data becomes crucial in addressing important questions like how safety measures can be improved, what technological advancements should airline companies consider while designing aircrafts, and which areas and situations can be categorized as high-risk. Most importantly, this analysis plays a pivotal role in maintaining public trust in air travel.

Objective

The aim of this project is to reveal insights into the main causes of air crashes and other aviation accidents, with a focus on understanding how these can be prevented in future. Here, we plan to perform analysis on various attributes within the data files by implementing a data pipeline in AWS. This pipeline will merge data files and transform all necessary rows and columns for final analysis. Some of the columns we will focus on are accident location, number of injuries sustained, weather conditions, and aircraft design details like type, make, model, among other relevant factors.

Some key analytical points we will explore are:

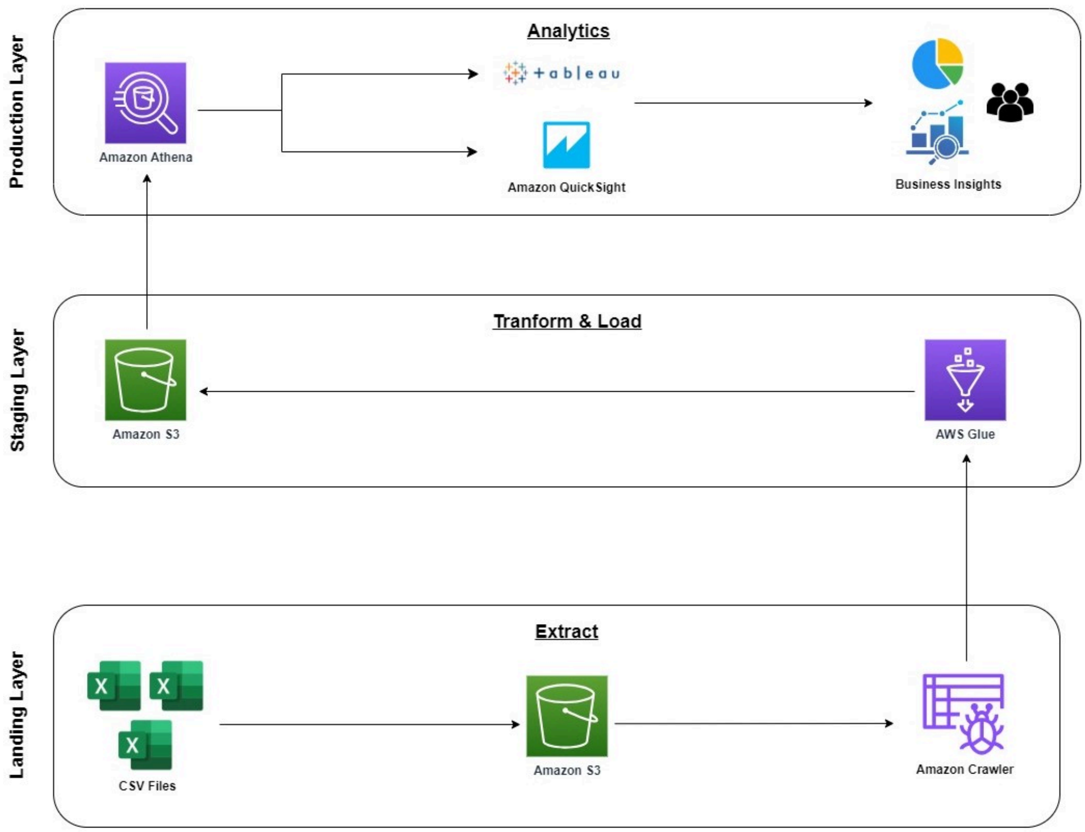
- What areas and weather conditions are more prone to airline accidents?
- Which type of accidents and incidents are common and why?
- What are common technical faults that lead to such tragedies?
- With advancements in aviation technology, have the number of accidents reduced over the years?

By diving into these attributes, we aim to discover patterns and identify correlations which have the potential to inform safety and mitigation strategies for similar incidents in the future.

Data Source

For this project, we will use Kaggle's [Aviation Accident Database](). This dataset has more than 30 columns and ~80,000 records providing air accident information over the past 40 years. The attributes are both continuous and categorical in nature and since there are many missing values, we will be using different techniques to transform the data before performing analysis. Finally, the dataset will be divided into 2 or more files before loading in the pipeline.

Data Pipeline Architecture



For this project, we divided our data file of aviation accidents data into 3 files - accident investigation, accident location, and US states codes. A 3-layer pipeline was created to implement the ETL process using Amazon S3, AWS Crawler, AWS Glue, and Amazon Athena. Before implementing the pipeline, we performed data cleaning for a few columns in Python to impute null values. The most frequently occurring value (mode) was used for imputation. In some other cases implemented in the ETL Workflow, we used ’Unknown’ as enough details were not available to impute null values.

```
[ ] # Replacing NONE values with None
    df.replace('NONE', 'None', inplace = True)

    df['Weather_Condition'].replace('Unk', 'UNK')

[ ] # Splitting Injury Severity to remove unnecessary details
    df[['Injury_Severity', 'Num']] = df['Injury_Severity'].str.split('(', expand = True)
    df.drop(['Num'], axis = 1, inplace = True)

    # Imputing missing values in columns with the most frequently occuring values

    df['Purpose_of_Flight'].value_counts()
    df['Purpose_of_Flight'].fillna('Personal', inplace = True)
```

Fig 1. Sample of Python Data Cleaning Script

Landing Layer

In the first stage of the data pipeline, we created source and target S3 buckets to manage data prior to and post ETL, respectively. After creating S3 buckets, the CSV files were uploaded to specific folders within the source bucket. Next, to perform ETL and query the transformed data for analysis, we created source and target AWS Glue Databases using Amazon Athena. The data was loaded in the database with the help of AWS Crawlers which automatically detects the schema of the files and creates tables to store data in the database. The screenshots below show the steps described here.

General purpose buckets (3) Info

Buckets are containers for data stored in S3. Learn more

Copy ARN | Empty | Delete | Create bucket

Find buckets by name

Name	AWS Region	Access	Creation date
aviation-accident	US East (N. Virginia) us-east-1	Bucket and objects not public	November 23, 2023, 11:00:08 (UTC-05:00)
aviation-accidents-output	US East (N. Virginia) us-east-1	Bucket and objects not public	November 28, 2023, 13:46:28 (UTC-05:00)

Fig 2. Creation of Source and Target S3 Buckets

Databases (3)

A database is a set of associated table definitions, organized into a logical group.

Last updated (UTC) December 1, 2023 at 22:48:01 | Edit | Delete | Add database

Filter databases

Name	Description	Location URI	Created on (UTC)
aviation_accidents	-	-	November 28, 2023 at 20:37:41
aviation_accidents_etl	-	-	December 1, 2023 at 21:47:41
default	-	-	November 23, 2023 at 16:55:28

Fig 3. Creation of Source and Target AWS Glue Databases

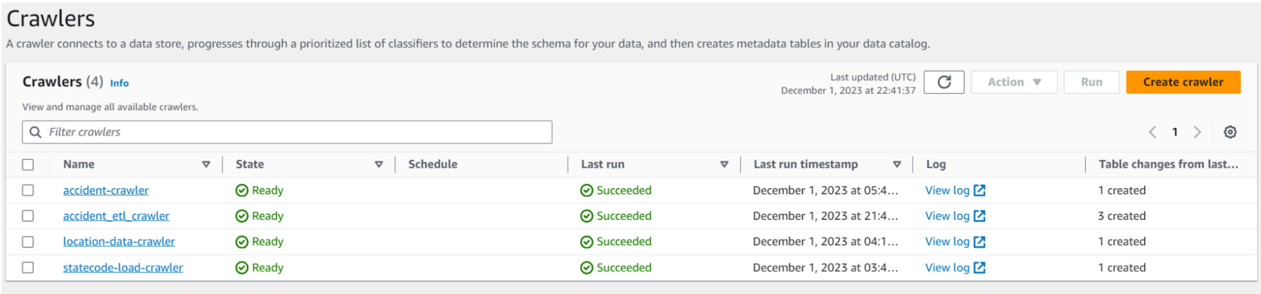


Fig 4. Creation of Crawlers to Load Data into Source Database

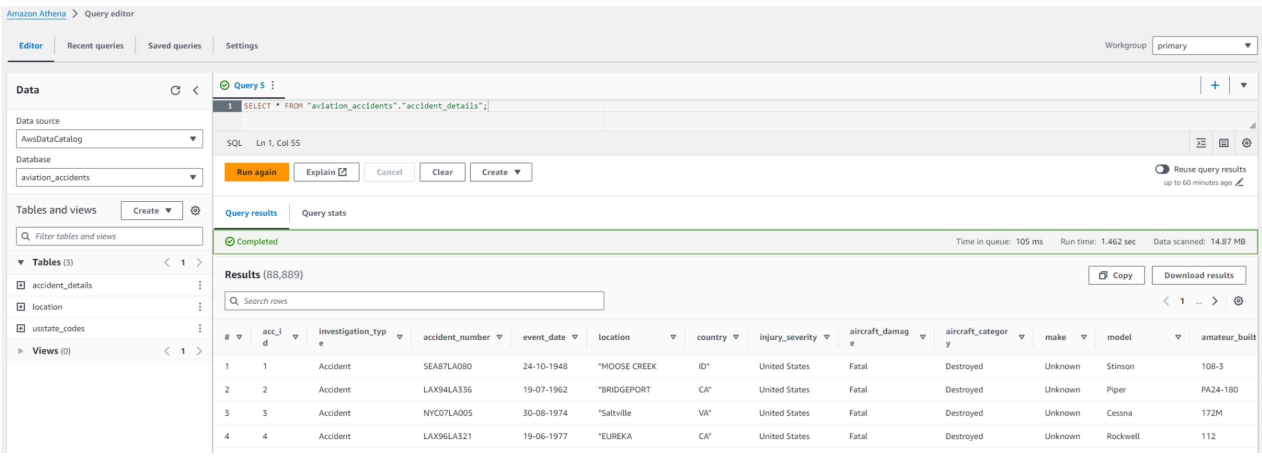


Fig 5. Preview of Accident Details Table

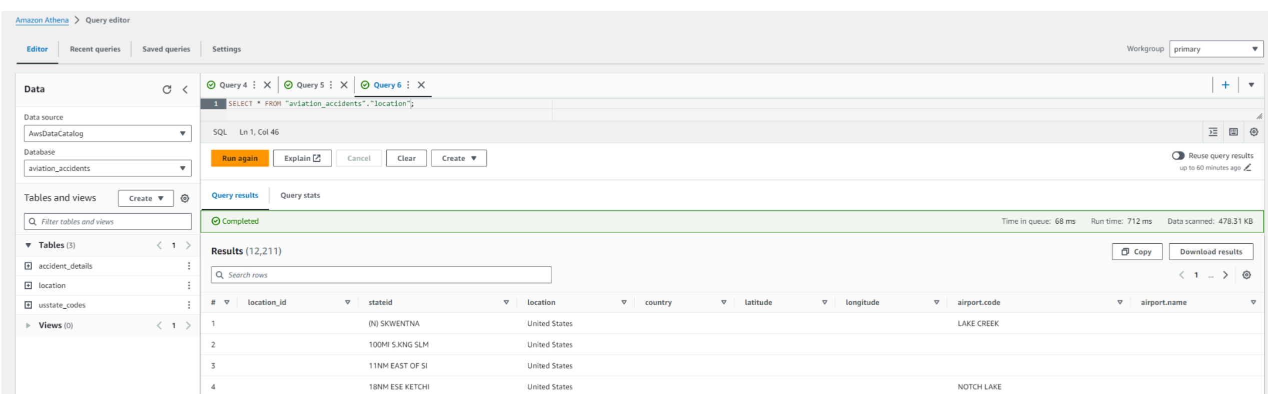


Fig 6. Preview of Accident Location Table

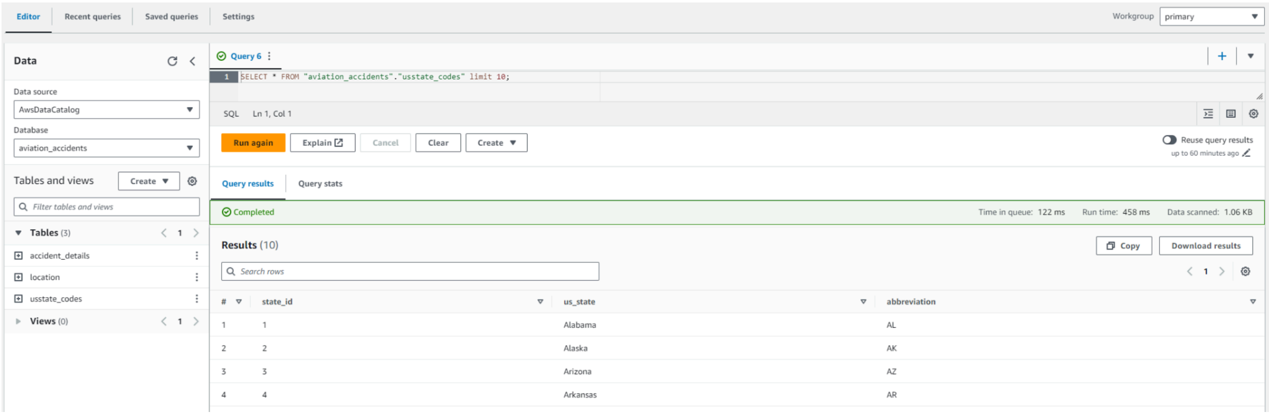


Fig 7. Preview of US States Code Table

Staging Layer

In the second stage of the pipeline, ETL transformations were performed on the data. Three jobs were created to transform the three tables. For the first table i.e accident investigation details, the data was loaded from AWS Glue Data Catalog and the following transformations were performed:

- Split Transform for splitting date column into day, month, and year arrays to perform analysis at different levels of time granularity. These arrays were converted using Array to Columns Transform.
- Split Transform for location column to store the exact location (eg. city) of accident and the state of accident in different arrays. Again, the Array to Columns Transform was used.
- Derived Column Transform was used to create a new column that stores a bucket value for injuries in each accident based on conditions in existing column. For ex, ‘0-100 injuries’.
- Finally, schema for some columns was changed and final data was stored in target S3 bucket.

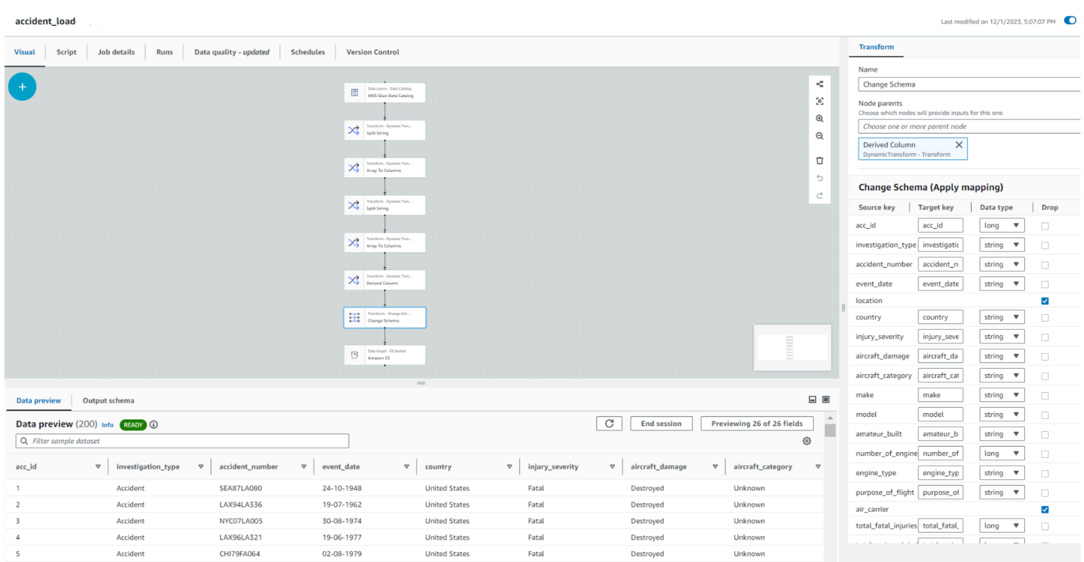


Fig 8. Accident Details - ETL Workflow

Next, for the accident location table, the following transformations were performed before loading data into target S3 bucket:

- Change of Schema to drop latitude and longitude columns
- Derived Column Transform to impute missing values with ‘Unknown’ for airport name and code columns.

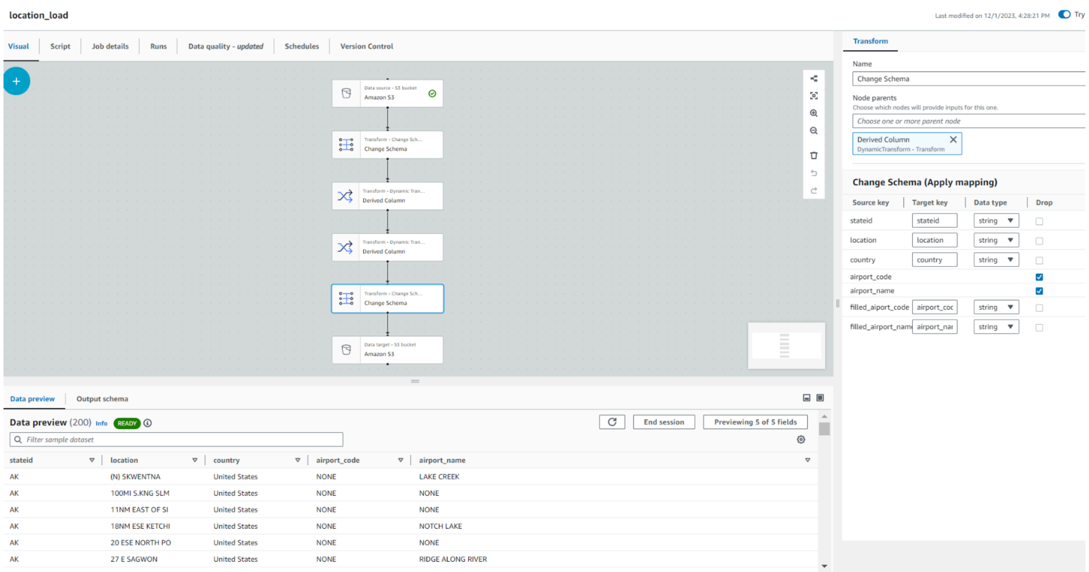


Fig 9. Accident Location - ETL Workflow

Finally, for the US state codes table, as it has very few columns and no transformations were needed, the data was directly loaded into target S3 bucket.

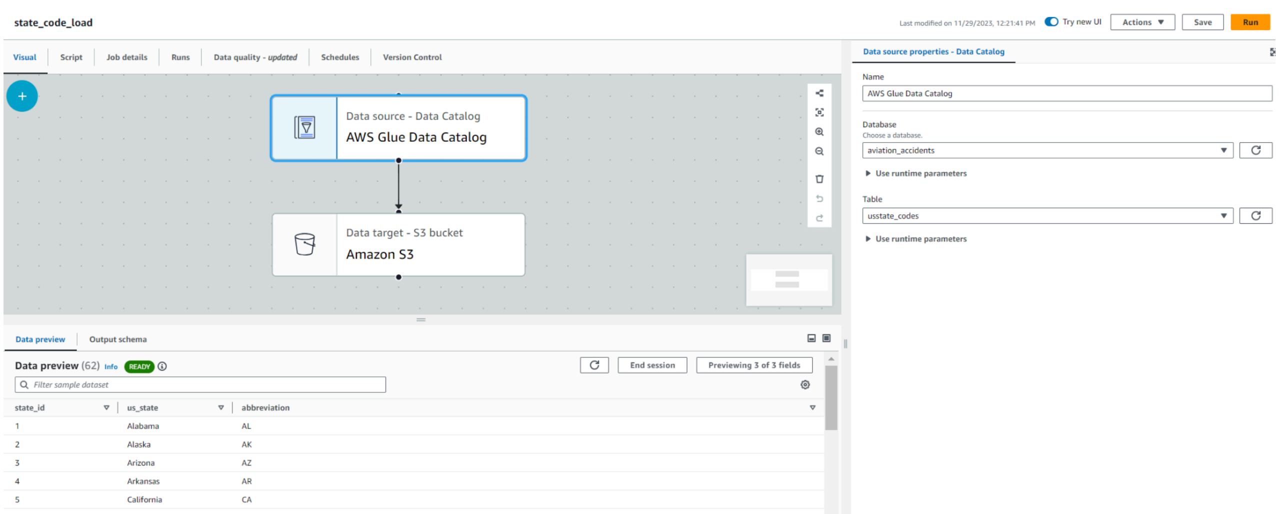


Fig 10. US States Code - ETL Workflow

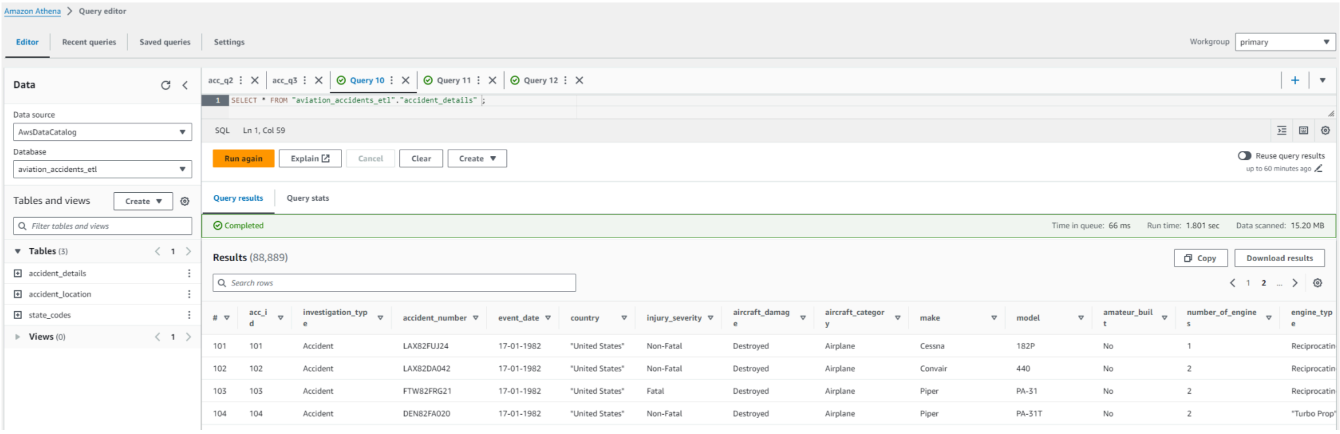
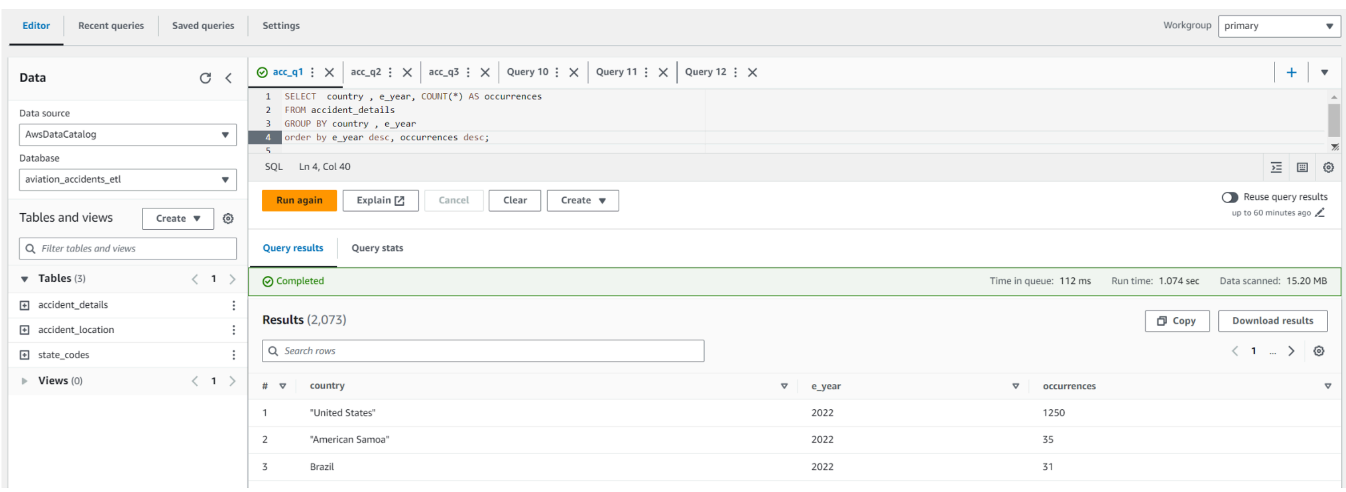


Fig 11. Preview of Accidents Table in Target S3

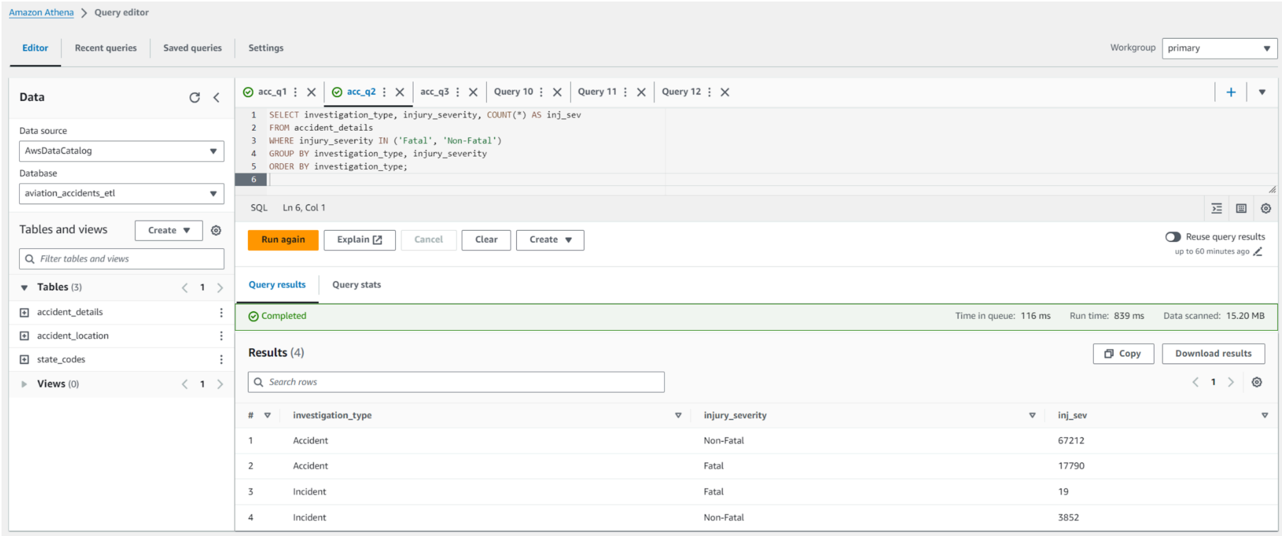
<u>Production Layer</u>

In this layer, the transformed data was used to perform analysis in Amazon Athena.

1) Countries with the highest number of accidents or incidents by year - This query gives an idea about the countries with the most number of aviation accidents. Also, note that our data is limited to a time frame and does not represent all time aviation accidents.



2) Number of non-fatal and fatal injuries across different accident types - This query gives the total number of injuries that could possibly lead to death indicating the severity of accidents.



3) Top 10 makes of aircrafts that were destroyed during aviation accidents - Analyzing the makes of aircrafts can be useful in identifying if a certain type of make/model is leading to such accidents over time.

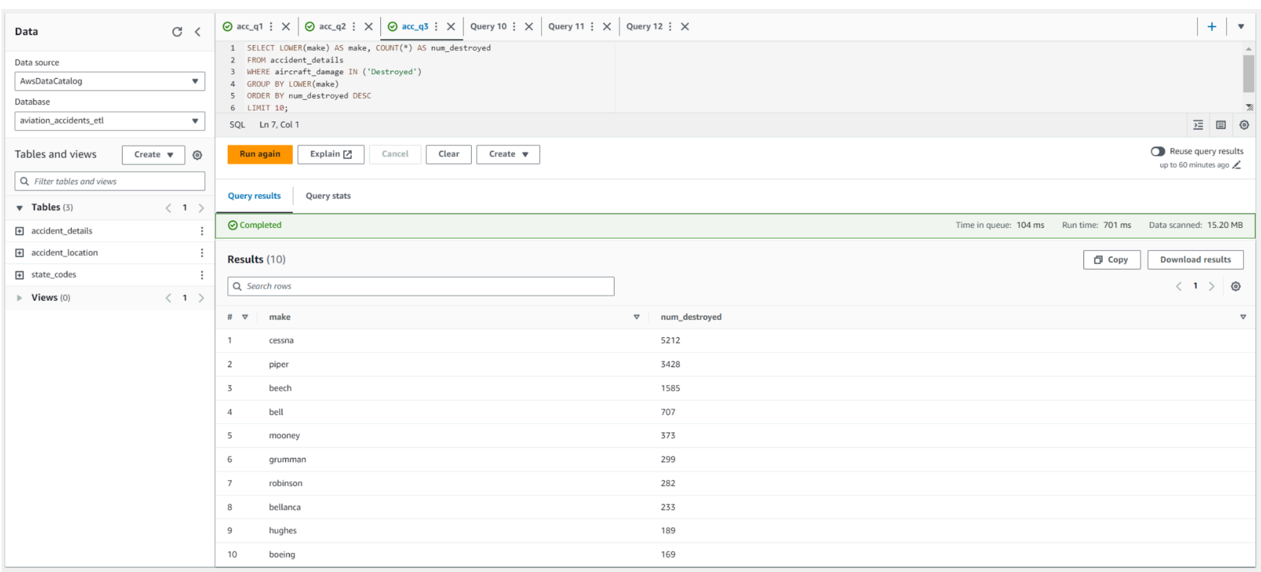
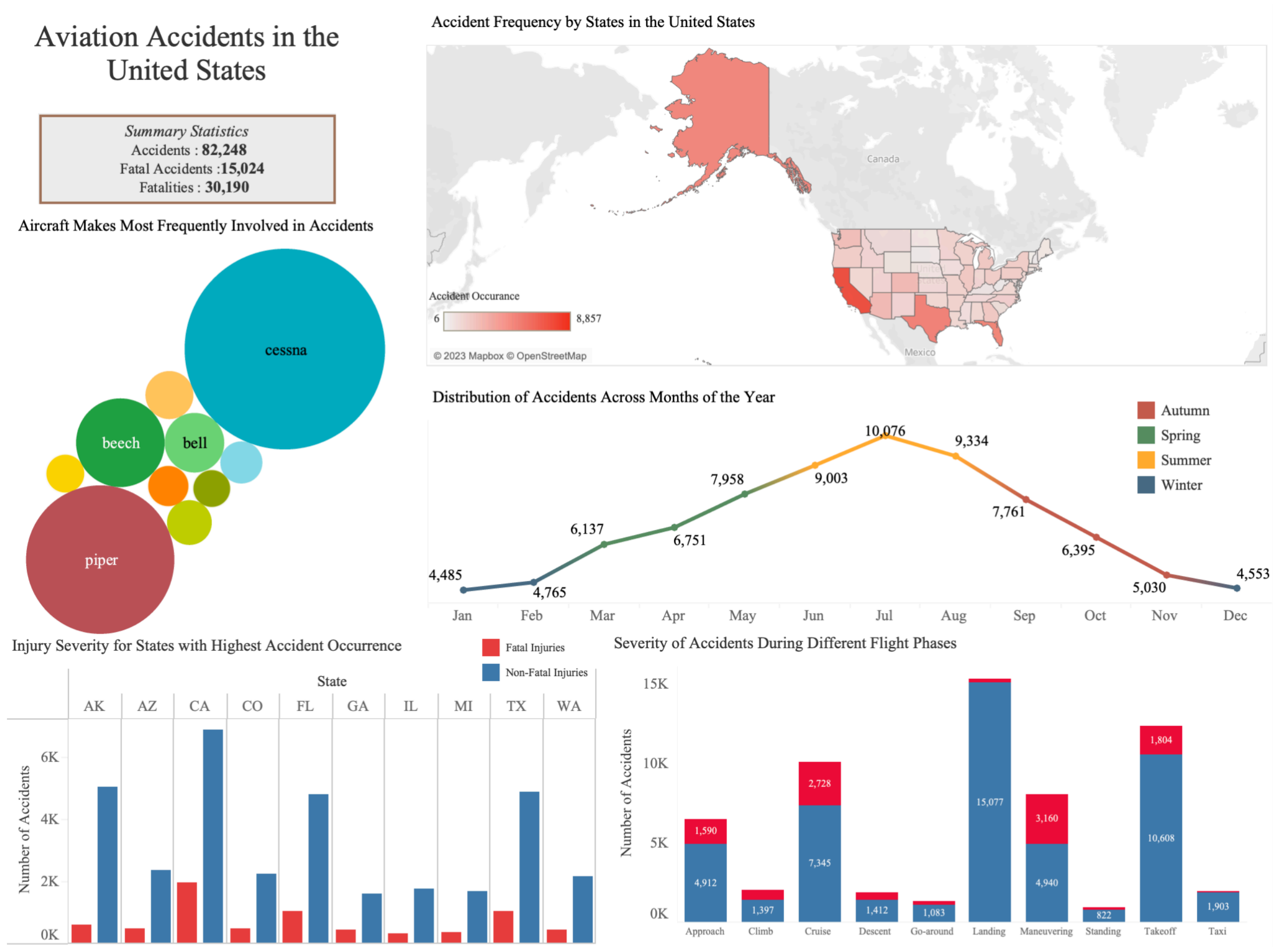


Tableau Dashboard:



The Tableau dashboard on U.S. aviation accidents provides key insights into prevalent aircraft makes, accident frequencies by states, seasonal patterns, injury severity in high-incident states, and severity across flight phases. These insights could aid stakeholders in identifying potential safety improvements, directing attention to specific aircraft models, geographical areas, and times of the year prone to higher accident rates. Understanding injury severity by state informs emergency response planning, while insights into accidents during different flight phases contribute to targeted safety measures during critical stages of air travel. Overall, the dashboard helps equip decision-makers with actionable information to enhance aviation safety protocols and mitigate risks effectively.