US Traffic Accident Dataset

Cloud Analytics and Data Warehouse Implementation

Nayel Enikeev, Nassim Ali-Chaouche, Jason Djuandy, Quan Gu, Mohana Nukala (*Authors*)
DATA 225 Group 6

Simon Shim (*Professor*)
DATA 225 Prof. at SJSU

***Abstract*—"NHTSA (National Highway Traffic Safety Administration) projects that an estimated 42,915 people died in motor vehicle traffic crashes last year (2021), a 10.5% increase from the 38,824 fatalities in 2020."[1]. Utilizing cloud analytics and data warehouse implementation on US Traffic Accident Historical Dataset, meaningful visualizations and insights can be taken to improve these conditions.**

I. PROBLEM STATEMENT

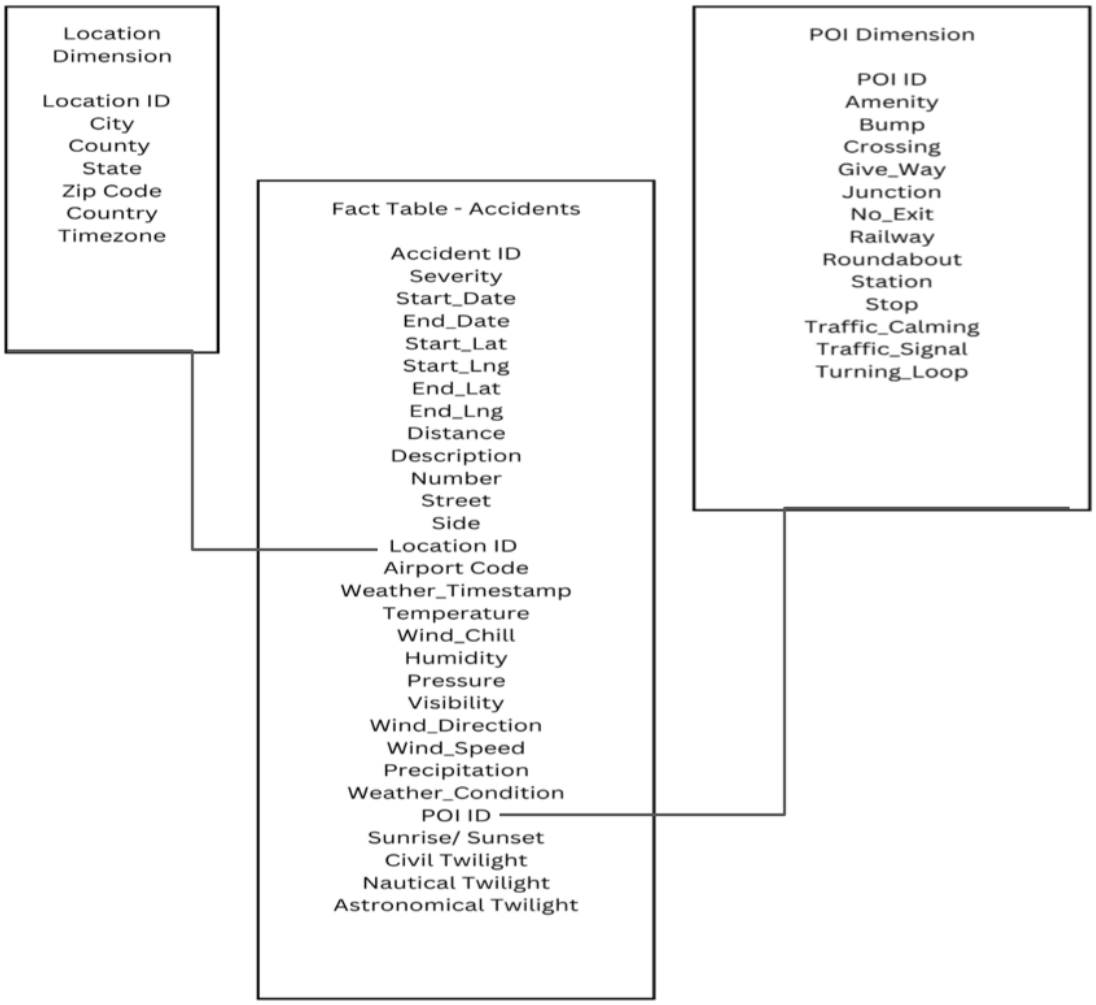
As mentioned above, the sudden increase of traffic accidents happening across the country needs to be addressed immediately. Although reducing traffic accidents is an important public safety challenge, it is a challenge that must be faced head on. Mitigating the frequency of accidents also means saving lives in the process.

The "US-Accidents: A Countrywide Accident Traffic Dataset"[2] is chosen to help tackle this issue. The dataset contains 2.8 million traffic accidents records covering 49 States in the USA collected from multiple sources including APIs spanning from 2016 to 2021. Summarizing the trends found in the dataset will hopefully provide a meaningful understanding and conclusions on the possible factors and/or conditions which increases the likelihood of accidents. This extracted information could then be used to be applied in the real world setting by coming up with actions and hopefully prevent/reduce the number of accidents which in turn reduces the number of lives lost every year due to traffic accidents.

II. SOLUTION

In order to create a solution, a Cloud SQL instance is created using the Google Cloud platform as an initial source of information configured to have a VPC (Virtual Private Cloud) network to be able to safely access various data sources. Using Apache Airflow which uses the Direct Acyclic Graph (DAG) script Architecture allowing the workflow to be designed and scheduled, ETL (Extract, Transform and Load) scripts are developed in Python and orchestrated using a Cloud Composer. This Cloud Composer is hosted on a container based cluster which enables auto scaling both horizontally and vertically, allowing for scalability.

The data initially stored in the Cloud SQL instance is extracted, transformed and loaded into the data warehouse with the help of Airflow. The Star Schema is used for the data model when storing data in the warehouse as it is very commonly used in OLAP (Online Analytical Processing) Apps which will be used in the Analytics and Business Intelligence. Analytics and Business Intelligence done on the data in the data warehouse will hopefully be able to analyze and extract trends resulting in meaningful insights as to why these traffic accidents happen by the Key Performance Indicators (KPI) extracted. The KPIs will be the key to implementing a solution to reduce traffic accidents.



III. CONCEPTUAL DATABASE DESIGN

The selected dataset will be stored in the data warehouse after the ETL process in the form of the star schema. The star schema stores data in the form of fact and dimension tables. Facts hold numeric values that represent a specific event which are the traffic accidents in this case. Dimensions contain values which describe these facts and attributes that are used to search, filter and classify facts. Location and POI is selected as Dimensions as from the 2.8 million records there are combinations of information below that can be used to describe the accident. More on the conceptual database design will be explained below:

- **Fact Table: Accidents** table contains the main specific event information of traffic accidents. This includes:
    - **Accident ID**, a unique identifier for each accident.
    - **Severity**, a number between 1 and 4 where 1 indicates the lease impact on traffic (short delay) and 4 indicates the significant impact on traffic (long delay).
    - **Start and End for Time, Latitude and Longitude** providing

information on the start time of the accident in local time, end time which is when the impact of accident on traffic flow is dismissed and the start and end point denoted by latitude and longitude.
    - **Distance**, which is the length of road extent affected by the traffic accident.
    - **Description**, gives a description on the traffic accident if any.
    - **Number, Street and Side**, provides information on the street number and side of the street of the accident.
    - **Location ID**, which corresponds to a unique identifier in the Location Dimension table.
    - Weather: **Airport Code** (which denotes the closest airport based weather station)**, Weather Timestamp** (the local timestamp of the weather observation)**, Temperature (F), Wind Chill (F), Humidity (%), Pressure (in), Visibility (in), Wind Direction, Wind Speed (mph), Precipitation (in) and Weather Condition** (rain, snow, thunder, etc)**.**
    - **POI ID** refers to a unique identifier in the POI Dimension table.
- **Dimension Table: Location** table contains values which describe the accident. Information here includes:
    - **Location ID** provides a unique identifier for each Location combination
    - **City, County, State, Zip Code, Country, Timezone**
- **Dimension Table: POI (Points of Interest)** table contains possible points of interest regarding the accident and are denoted in either True (meaning that it exists) and False (does not exist):
    - **POI ID** provides a unique identifier for each combination of the points of interest listed below.
    - Multiple points of interest, True if exist and False if does not exist: **Amenity, Bump, Crossing, Give Way, Junction, No Exit, Railway, Roundabout, Station, Stop, Traffic Calming, Traffic Signal, Turning Loop.**

IV. ANALYTICS

Having the dataset loaded on the data warehouse in a star schema data model. Analytics is done using SQL Queries where interesting insights and trends can be drawn from the traffic accident historical dataset. Below are the SQL Queries done and insights taken from each query using Google BigQuery:

- Top 10 States with the Most Accidents (Appendix 1)
    - The state with the most number of accidents is California, followed by Florida and Texas
- Number of Accidents based on Severity (Appendix 2)
    - The vast majority of accidents have a “Severity” level of 2 .
- Average Length of Road Affected Based on Severity (Appendix 3)
    - Accidents with a severity level of 4 affected the greatest extent of the road with an average distance of 1.45 miles. Severity levels 2 and 3 affected a similar distance.
- Most Common Weather Conditions at the Time of Accident (Appendix 4)
    - The most common weather condition at the time of the accident was “Fair”. (1,107,194). The next most common conditions were

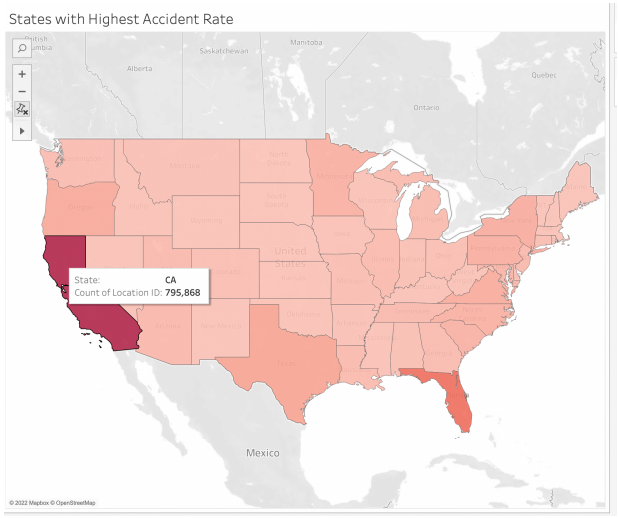
"Mostly Cloudy" (363959) and "Cloudy" (348,767). This suggests that adverse weather conditions are not a major contributing factor to car accidents.

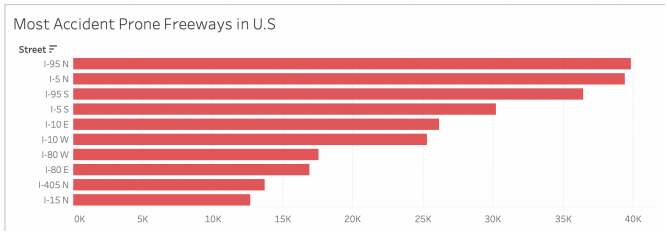
- Number of Accidents at Day/Night Time (Appendix 5)
    - The majority of accidents occur during the day (~63.7%)
- Number of Accidents Per Day of Week (Appendix 6)
    - Friday had the most number of accidents. Saturday and Sunday saw the fewest number of accidents by a significant margin compared to every weekday.
- Number of Accidents by Hour of Day (Appendix 7)
    - The highest number of accidents occurred between 5:00 PM and 6:00 PM, followed closely by between the time of 4:00 and 5:00 PM, and 3:00 PM to 4:00 PM.
- Number of Accidents by Month (Appendix 8)
    - December had the most accidents (473943), followed by November (360696) and October (299131).
- Number of Accidents by Year (Appendix 9)
    - 2021 had the largest number of accidents by a significant amount compared to other years at 1,511,745, followed by 2020 (625,864), and 2019 (258,615).
- Number of Accidents by Timezone (Appendix 10)
    - Most accidents occurred in the Eastern time zone, followed by Pacific, Central, and Mountain.
- Average Visibility Based on Severity (Appendix 11)
    - For each accident severity level, the average visibility was greater than 9 miles (with a maximum of 10). This suggests that visibility is not a major factor in the severity of a car accident.
- Top 10 Streets with the Most Accidents (Appendix 12)
    - I-95 N had the most number of accidents, followed by I-5 N and I-95 S.
- Top 10 Cities with the Most Accidents (Appendix 13)
    - The city with the most number of accidents is Miami with (4%)
- The Relative Side of the Street (Right/Left) in Address Field at the Time of Accident. (Appendix 14)
    - The relative side of the accident street is mostly Right (2353309) over Left (492032)
- The Average Temperature Based on Severity
  (Appendix 15)
    - The average temperature of lowest severity 1 is 71 Fahrenheit, average temperature for highest severity 5 is 58 Fahrenheit. Generally the severity goes up as temperature gets lower.
- Most Frequent Visibility at the Time of Accident
  (Appendix 16)

- The most frequent visibility during accidents is 10 miles with a total of 2230276 records.

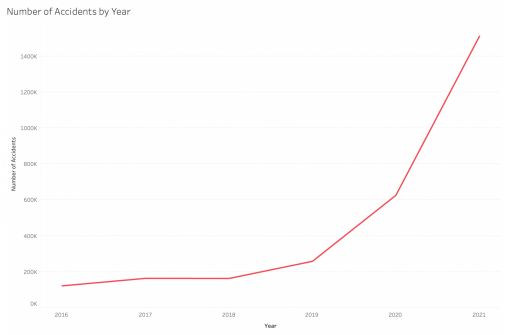
V. BUSINESS INTELLIGENCE

Business Intelligence is done using Tableau on the data in order to produce meaningful and interesting visualization which adds even more value on top of the analytics done beforehand. Various KPI (Key Performance Indicators) are extracted from the visualizations created. Below are several of the data visualizations created using Tableau:

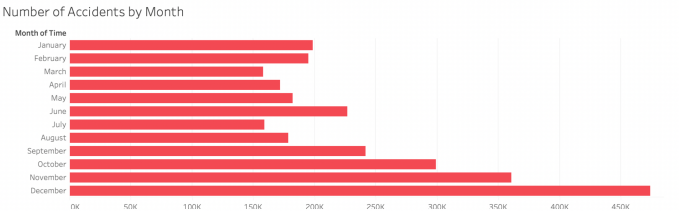




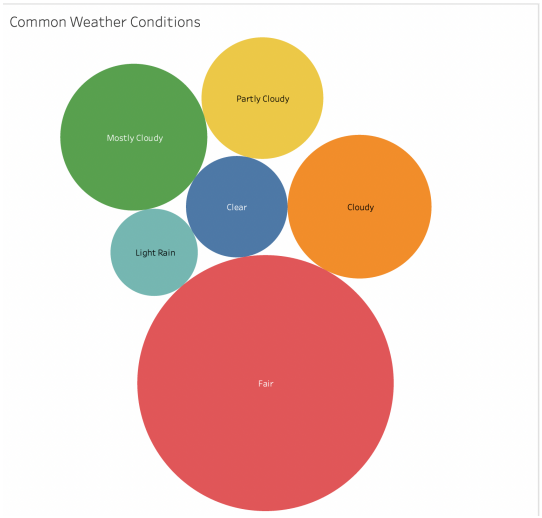
- Insurance companies can use the map and the freeway visual for pricing quotes for trucking companies that frequently travel through the high accident rate states/routes.



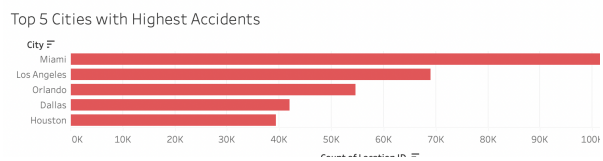
- Number of accidents drastically increased after 2020, which can be due to the lack of cars on the road in 2020 in the covid lockdown and drivers acclimated to higher speeds and rash driving on empty streets. These habits may be causing more accidents after 2020 when there are more drivers on the road.



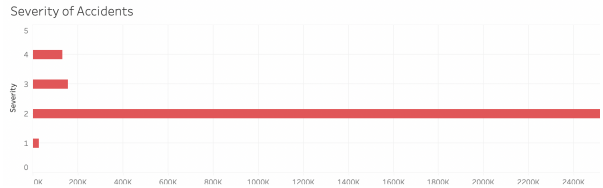
- From the above visual, December has the highest number of accidents. This makes sense as it is the holiday season, with a lot of road travel and transportation of goods. Also, the weather conditions are dangerous with low visibility and snow in a lot of areas which may cause more accidents.



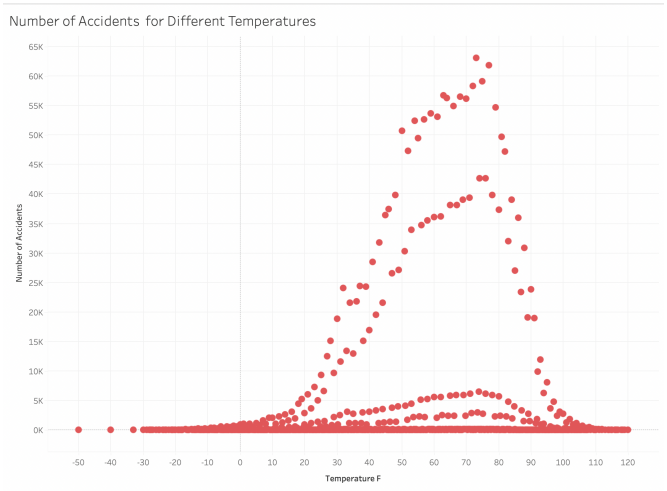
- Speaking of weather conditions, Fair weather seems to be the most prominent condition during the majority of the accidents, followed by Mostly Cloudy and Cloudy.



- Cities and state officials officials can use this accident data to improve the infrastructure like roads and bridges to reduce accidents in the area



- 2nd level of severity is the most common among the accidents, followed by 3 and 4.



- Scatter plot of different temperatures where the accidents occurred. Majority of accidents occurred between 70 and 80 degrees.

VI. CONCLUSION

Cloud analytics and data warehouse implementation on the dataset which contains traffic accidents enables analytics and business intelligence to be done on the data. For example, the queries and visualizations were able to show that there exists a significant gap on the traffic accidents happening in December compared to the other months. Several factors or conditions can be derived to attempt to explain this fact. The large number of accidents can be caused due to snow since December falls in the winter season and also due to December being a popular month for the holidays, larger traffic flow exists countrywide on people trying to get home to their families or basically people are just traveling for the holidays.

Possible conclusions drawn from this example such as snow and larger traffic flow will be helpful in determining possible solutions such as implementing lower speed limits on months where it usually snows and more highway patrols on months where there exists larger traffic flow such as the holidays. These examples are proof that analyzing trends derived from historical dataset can provide possible solutions which if implemented by the government such as the

NHTSA can help to reduce the frequency of accidents happening therefore saving lives.

VII. MOVING FORWARD

Moving forward, ETL scripts can be implemented to continuously extract, transform and load the data to the data warehouse and used for analytics will be real time data. All the analytics queries, dashboard and visualizations will be able to display the latest information gathered from various sources on traffic accidents. At the same time advanced machine learning can be implemented to extract even more meaningful trends which help create predictive models on accidents as well as where. These predictive models can also be used by the NHTSA to proactively look out and implement better safety precautions before the accident even happens.

REFERENCES

[1] NHTSA CITATION

https://www.nhtsa.gov/press-releases/early-estimate-2021-traffic-fatalities

[2] DATASET LINK

https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents?resource=download

[*] DATASET INFO

https://smoosavi.org/datasets/us_accidents

APPENDIX

[1]

```
SELECT State, count(*) AS Number_of_Accidents
FROM
`group-project-data-225.bidataset.bi_dataset_
fact` JOIN
`group-project-data-225.bidataset.bi_dataset_
dim_location` USING (Location_ID)
GROUP BY State
ORDER BY Number_of_Accidents DESC
LIMIT 10
```

[2]

```
SELECT Severity, count(*) AS
Number_of_Accidents FROM
`group-project-data-225.bidataset.bi_dataset_
fact` GROUP BY Severity ORDER BY
Number_of_Accidents DESC
```

[3]

```
SELECT Severity, AVG(Distance_mi) AS
Average_Distance_Affected FROM
`group-project-data-225.bidataset.bi_dataset_
fact` GROUP BY Severity ORDER BY
Average_Distance_Affected DESC
```

[4]

```
SELECT Weather_Condition, count(*) AS
Number_of_Accidents FROM
`group-project-data-225.bidataset.bi_dataset_
fact` WHERE Weather_Condition IS NOT NULL
GROUP BY Weather_Condition ORDER BY
Number_of_Accidents DESC
```

[5]

```
SELECT Sunrise_Sunset, count(*) AS
Number_of_Accidents FROM
`group-project-data-225.bidataset.bi_dataset_
fact` WHERE Sunrise_Sunset IS NOT NULL GROUP
BY Sunrise_Sunset ORDER BY
Number_of_Accidents DESC
```

[6]

```
SELECT FORMAT_DATETIME("%A", Start_Time) AS
Day_of_Week, count(*) AS Number_of_Accidents
FROM
`group-project-data-225.bidataset.bi_dataset_
fact` GROUP BY Day_Of_Week ORDER BY
Number_of_Accidents DESC
```

[7]

```
SELECT FORMAT_DATETIME("%H", Start_Time) AS
Hour_of_Day, count(*) AS Number_of_Accidents
FROM
`group-project-data-225.bidataset.bi_dataset_
fact` GROUP BY Hour_of_Day ORDER BY
Number_of_Accidents DESC
```

[8]

```
SELECT FORMAT_DATETIME("%B", Start_Time) AS
Month, count(*) AS Number_of_Accidents FROM
`group-project-data-225.bidataset.bi_dataset_
fact` GROUP BY Month ORDER BY
Number_of_Accidents DESC
```

[9]

```
SELECT FORMAT_DATETIME("%Y", Start_Time) AS
Year, count(*) AS Number_of_Accidents FROM
`group-project-data-225.bidataset.bi_dataset_
fact` GROUP BY Year ORDER BY
Number_of_Accidents DESC
```

[10]

```
SELECT Timezone, count(*) AS
Number_of_Accidents FROM
```

```
`group-project-data-225.bidataset.bi_dataset_
fact` JOIN
`group-project-data-225.bidataset.bi_dataset_
dim_location` USING (Location_ID)
WHERE Timezone IS NOT NULL
GROUP BY Timezone
ORDER BY Number_of_Accidents DESC
```

[11]

```
SELECT Severity, AVG(Visibility_mi) AS
Average_Visibility FROM
`group-project-data-225.bidataset.bi_dataset_
fact` GROUP BY Severity ORDER BY
Average_Visibility DESC
```

[12]

```
SELECT Street, count(*) AS
Number_of_Accidents FROM
`group-project-data-225.bidataset.bi_dataset_
fact` JOIN
`group-project-data-225.bidataset.bi_dataset_
dim_location` USING (Location_ID)
GROUP BY Street
ORDER BY Number_of_Accidents DESC
LIMIT 10
```

[13]

```
SELECT City, count(*) AS Number_of_Accidents
FROM
`group-project-data-225.bidataset.bi_dataset_
fact` JOIN
`group-project-data-225.bidataset.bi_dataset_
dim_location` USING (Location_ID)
GROUP BY City
ORDER BY Number_of_Accidents DESC
LIMIT 10
```

[14]

```
SELECT Side, count(*) AS Number_of_Accidents
FROM
`group-project-data-225.bidataset.bi_dataset_
fact` WHERE Side IS NOT NULL GROUP BY Side
ORDER BY Number_of_Accidents DESC
```

[15]

```
SELECT Severity, AVG(Temperature_F) AS
Average_Temperature_Affected FROM
`group-project-data-225.bidataset.bi_dataset_
fact` GROUP BY Severity ORDER BY
Average_Temperature_Affected DESC
```

[16]

```
SELECT Visibility_mi, count(*) AS
Number_of_Accidents FROM
`group-project-data-225.bidataset.bi_dataset_
fact` WHERE Visibility_mi IS NOT NULL GROUP
BY Visibility_mi ORDER BY Number_of_Accidents
DESC
```