DAMG 7245 - Big Data Systems & Intelligence Analytics

# Experiments with Big data

## 

| Name | NUID |
| --- | --- |
| Sai Saketh Vootla | 001568984 |
| Shika Shyam | 002194543 |
| Sai Raghavendra Viravalli | 001568587 |

[GITHUB](https://github.com/shikashyam/BigDataSystemsCoursework)

# 

## Abstract

This tutorial attempts to analyze and visualize the SEVIR Storm Imagery dataset in two different aspects:

* First, we will be drilling down to one specific instance of a SEVIR event, and attempting to visualize the imagery from different sensors that have captured data associated with that event which are captured in HDF5 files stored on SEVIR’s AWS S3 cloud.
* Second, we look at SEVIR data on a holistic level by bringing the data catalog, Storm Details, Storm Fatalities and Storm locations files for the year 2019 onto the Google Cloud Platform. We will make use of GCP Storage Buckets to store this data and BigQuery to convert the csvs into a relational database form and run exploratory queries on the same. Additionally, we will be making use of Google Data Studio to create interactive and informative dashboards based on the results of our preliminary analysis on BigQuery.

## Dataset

The SEVIR Dataset is a Storm Event Imagery Dataset for Deep Learning Applications in Radar and Satellite Meteorology.

The SEVIR dataset consists of

– 107 files in HDF5 format containing imagery

• Files separated by image type and date range for easier access

• File sizes ranging between 1GB and 20 GB

– A catalog containing metadata of each image, including

• Unique ID assigned to each event

• Times of each image

• Lat / Lon bounds of slice

• Map projection and image extent in projection coordinates for exact georeferencing

• NCEI Storm Event ID and Episode ID (for non-random cases) SEVIR File Details

Total Size of SEVIR Dataset: ~900 GB

There is bulk data which contains between January 1950 to October 2021 at the NOAA’s National Weather Service(NWS). There are available in CSV format. Using this database we download the appropriate 2019 file to perform further analysis.

## 

# Jupyter Notebook



* SEVIR has made available a free to use tutorial which walks the user through how to access and georeference SEVIR sensor data (5 types).
* The Jupyter notebook created as a part of this recreates that tutorial with a reduced scope of looking at just one specific event in time, and accessing, visualizing and georeferencing the same.
* In order to access the h5 files present in the AWS buckets, we use the boto3 python package.
* Our analysis will be for the event ID=835047, once we get access to the AWS S3 buckets we search and retrieve the vis,ir069,ir107, vil and lght files for the specific Event ID that we are interested in.
* In order to access h5 files, we make use of the python h5py module.
* Once we have visualized the data from all the 5 different sensor types, we georeference this data with the catalog data to point to the coordinates of the nearest city affected by this event.
* We make use of the basemap module in Python in order to visualize these geo-coordinates more effectively by plotting state borders in addition to just the coordinates.

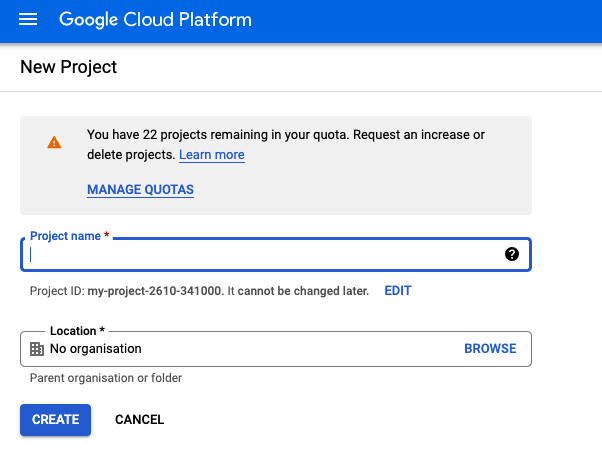
The Jupyter notebook can be found here. [Jupyter Notebook](https://colab.research.google.com/drive/1I1vOl2kPyQO5SnVXswd1NgNEKNouqCMj)

# Big Query

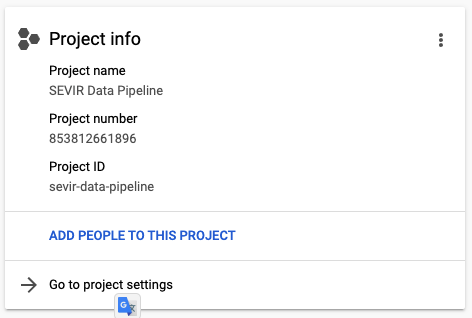


BigQuery is a fully managed enterprise data warehouse that helps you manage and analyze your data with built-in features like machine learning, geospatial analysis, and business intelligence.

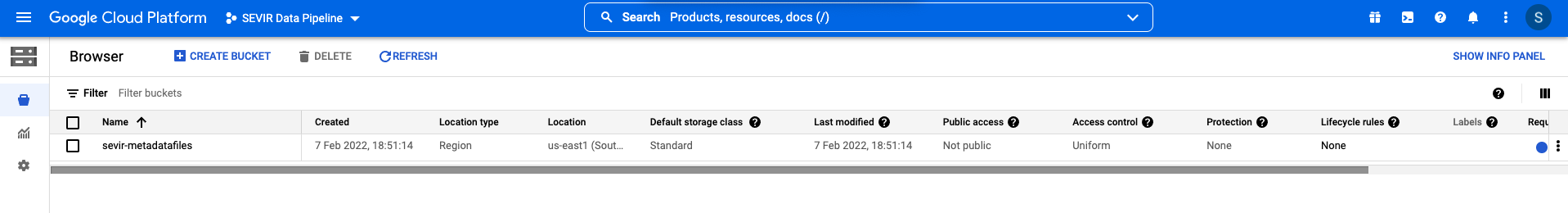
* Create a new project on GCP



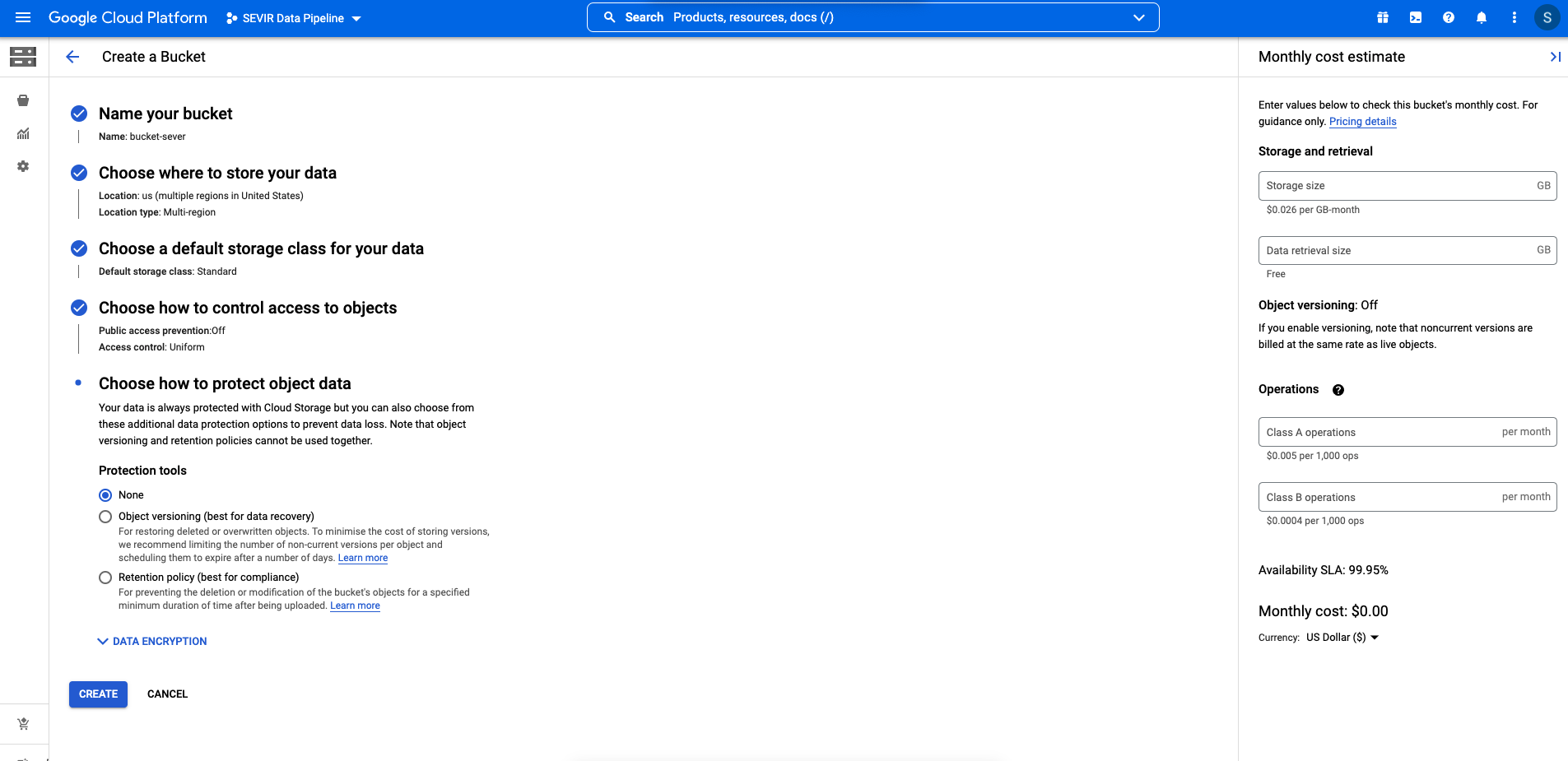
* Details of the project once created



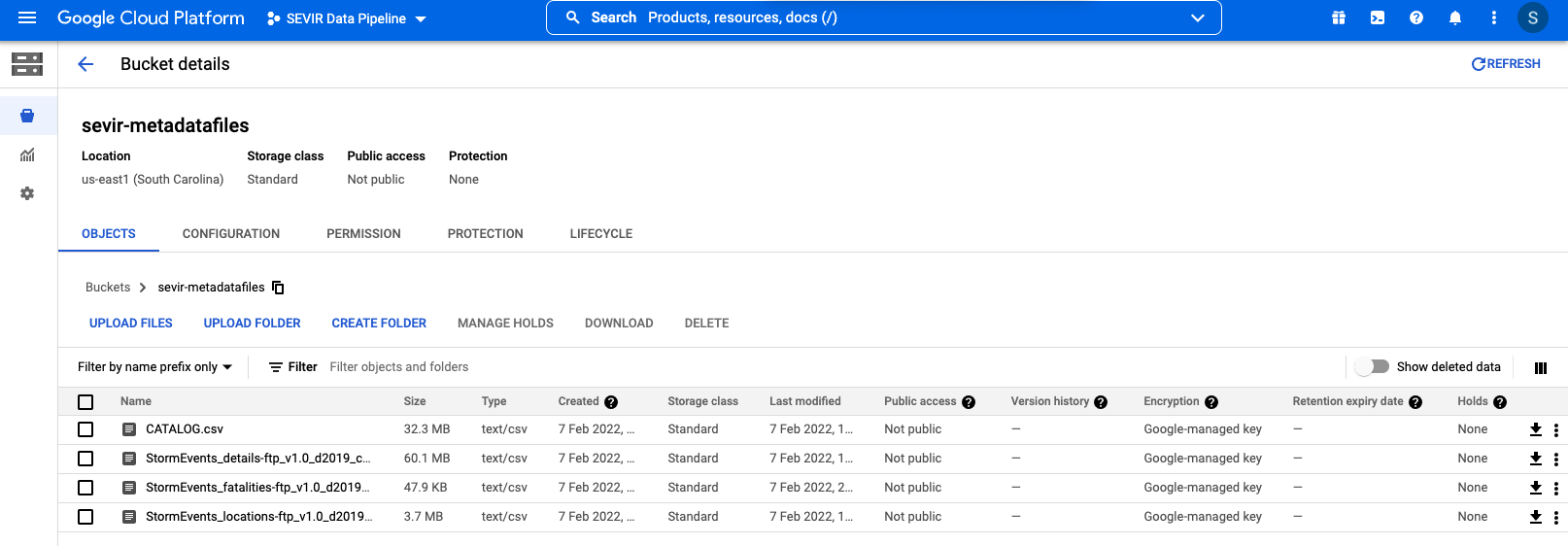
* Create a bucket to store our data files



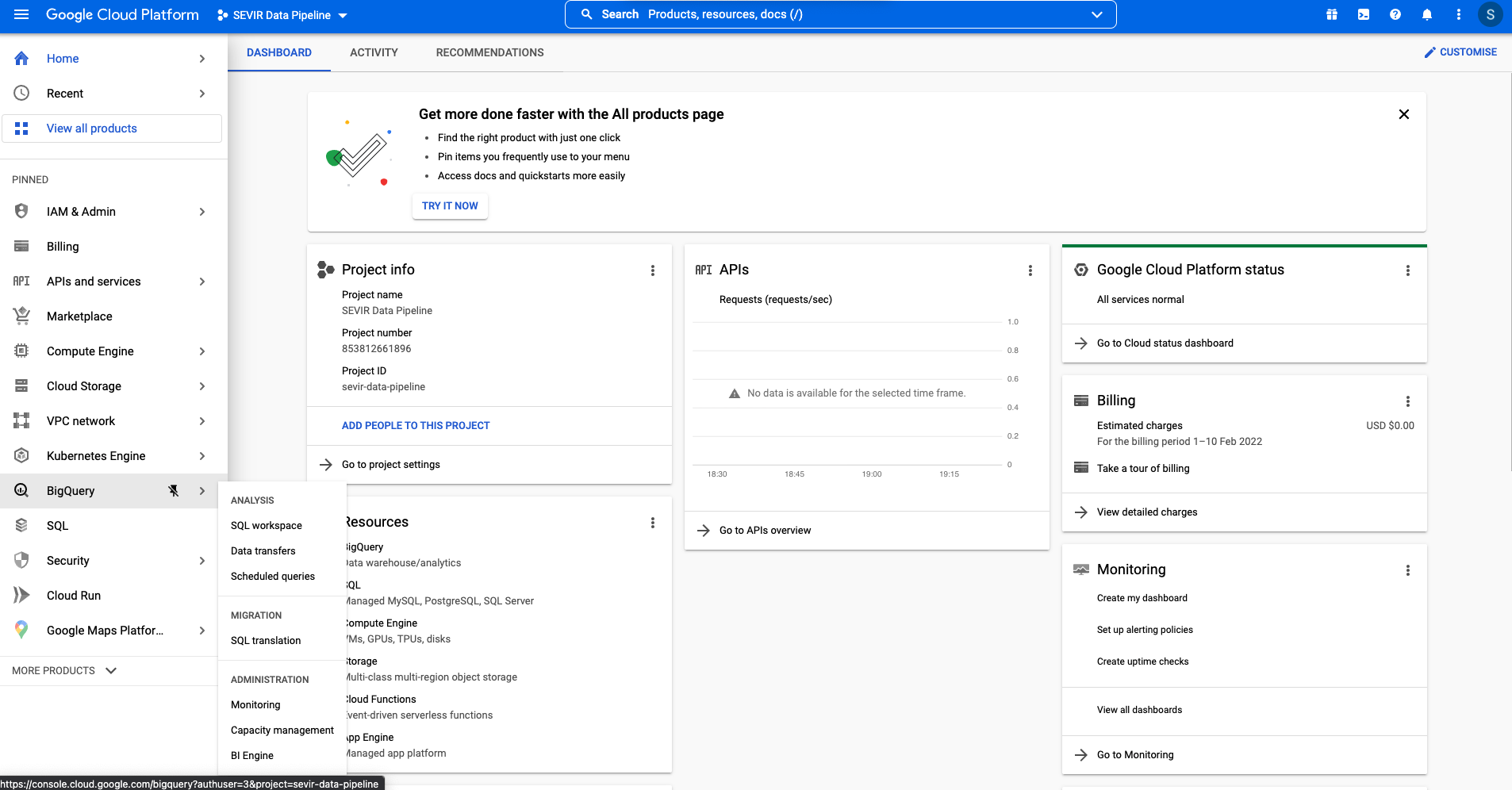
* Fill the necessary details for creating a bucket



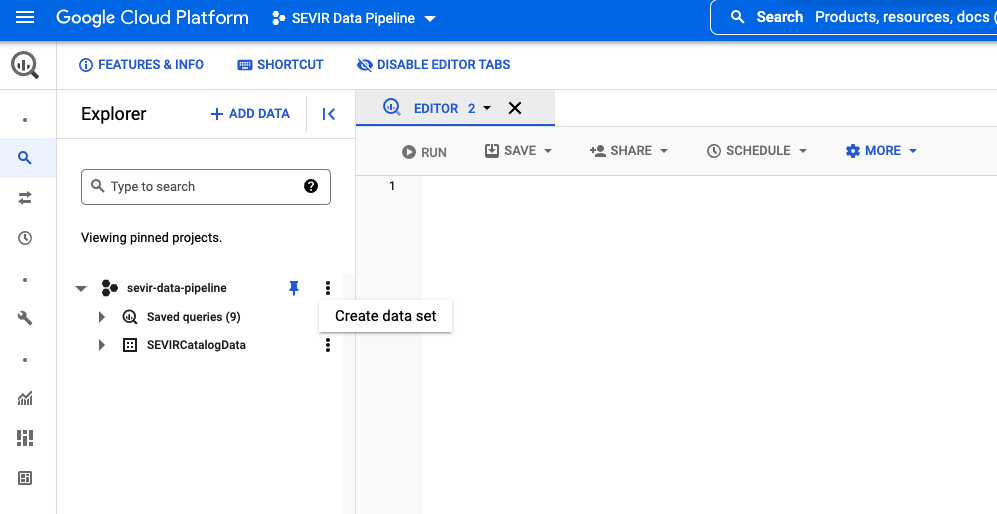
* Upload your datafiles onto the bucket once created



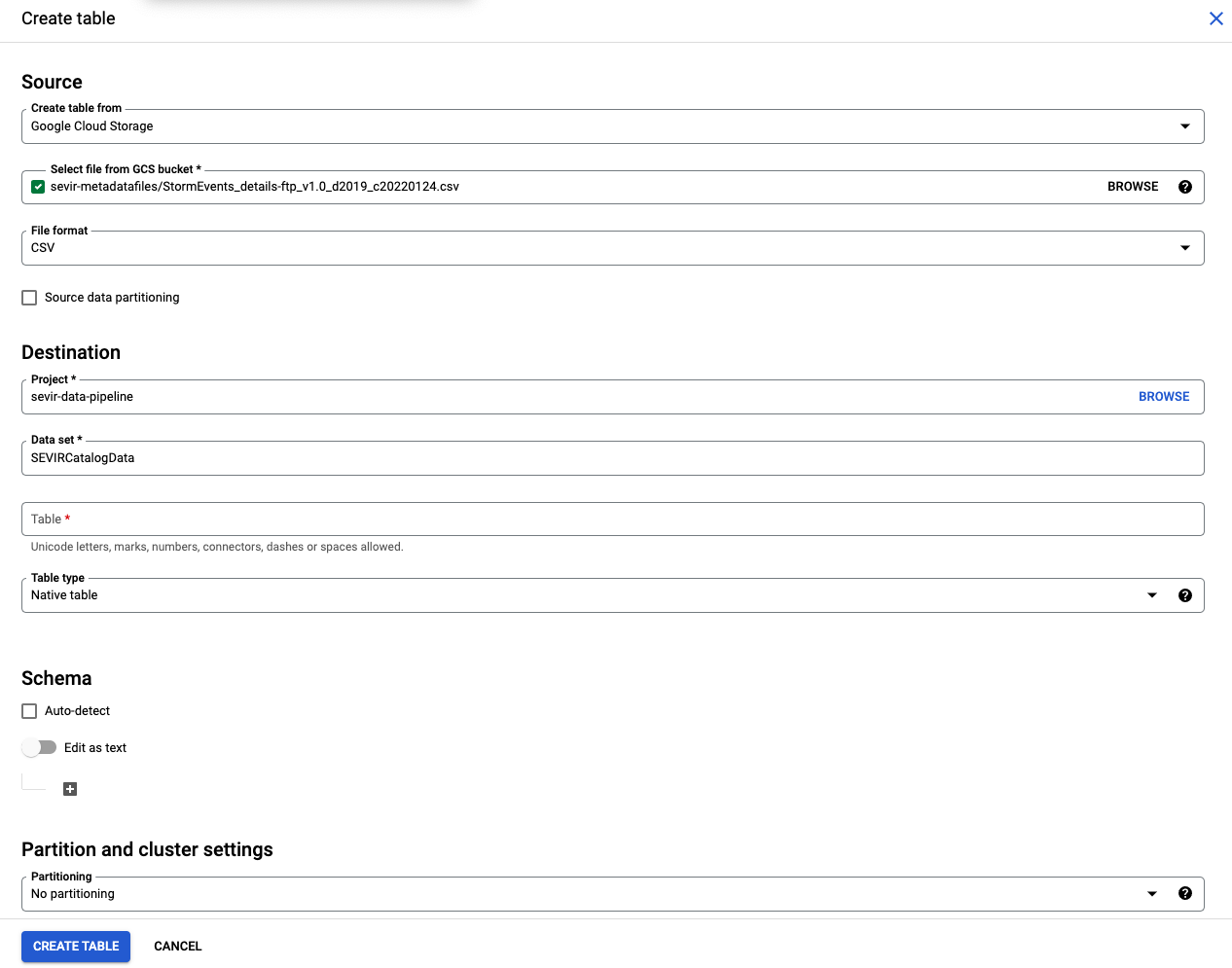
* After the datafiles are uploaded onto the bucket proceed to select BigQuery



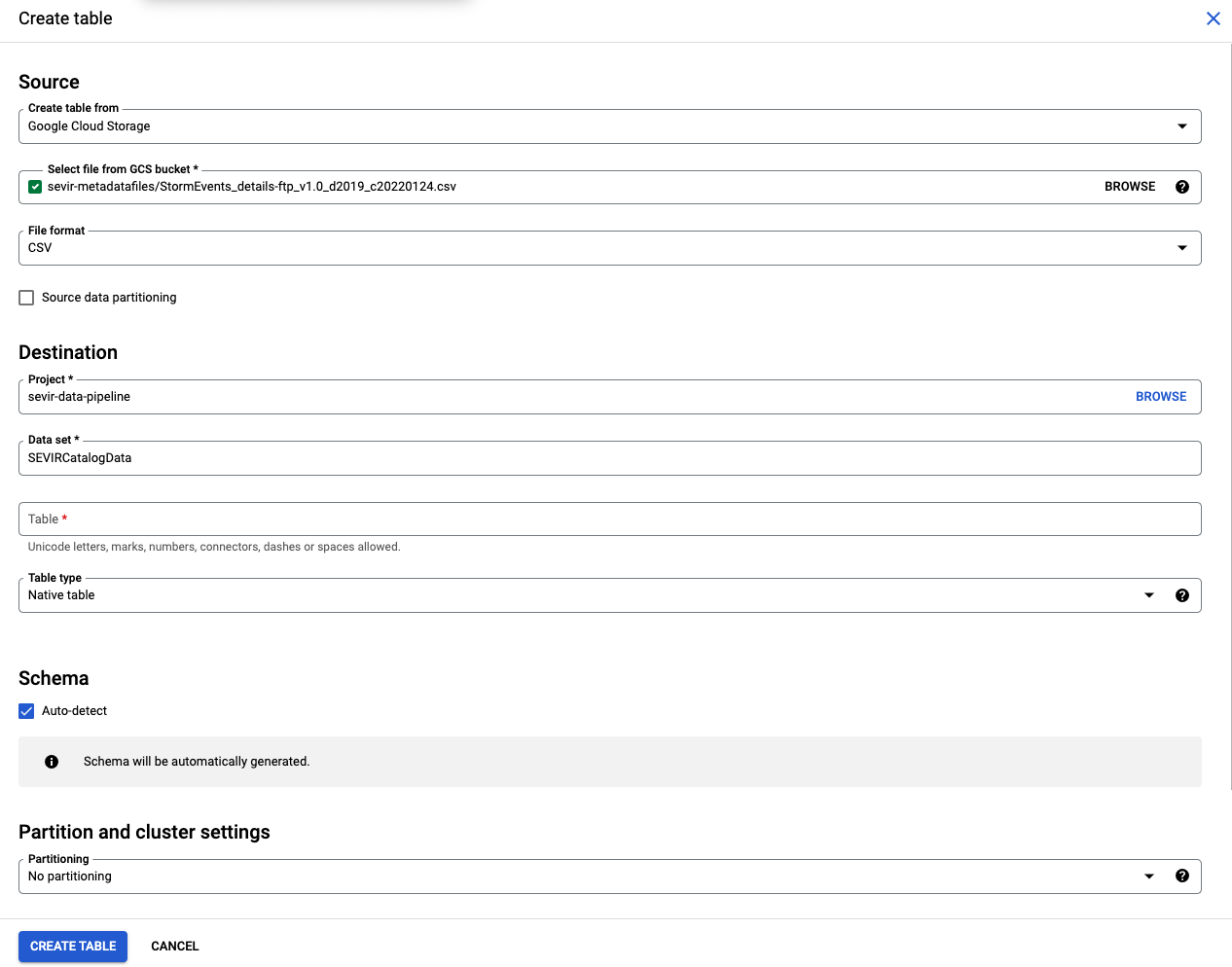
* Once BigQuery is selected under the GCP proceed to **Create dataset** under the project created

****

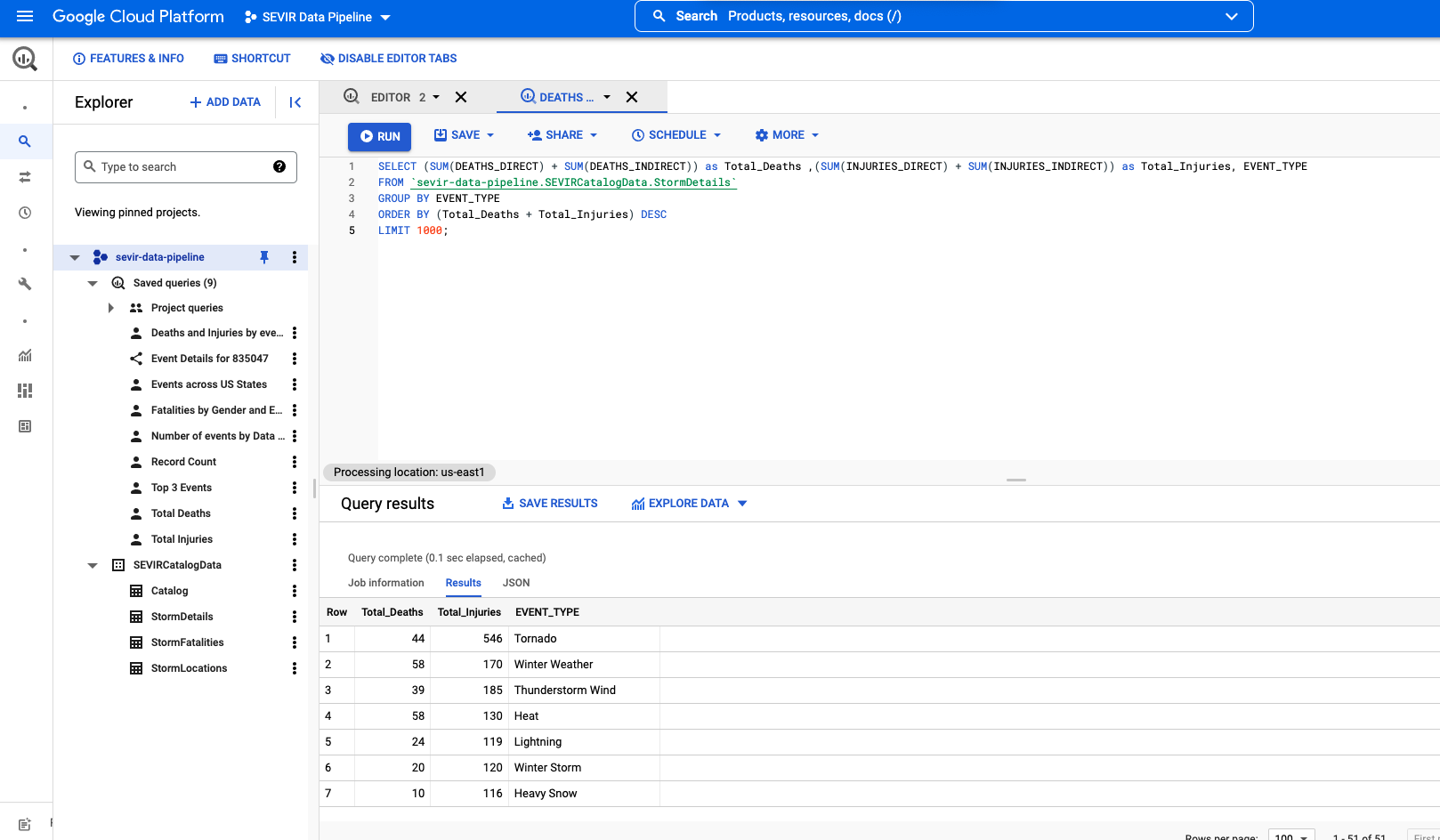
* **Creating the dataset**



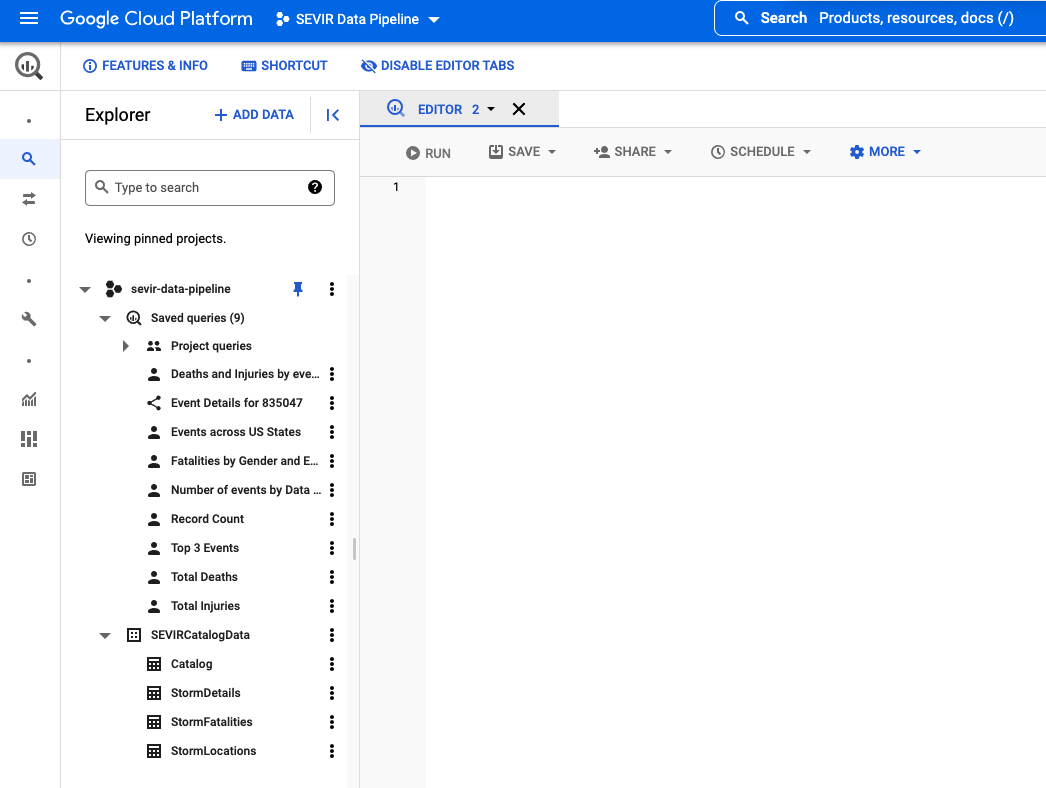
* After the necessary details have been filled to create a table and the bucket options make sure to toggle the auto detect checkbox



* After the data is added proceed to add the queries



* List of Queries and data under the project created can be viewed through the dropdown feature



# Data exploration in BigQuery

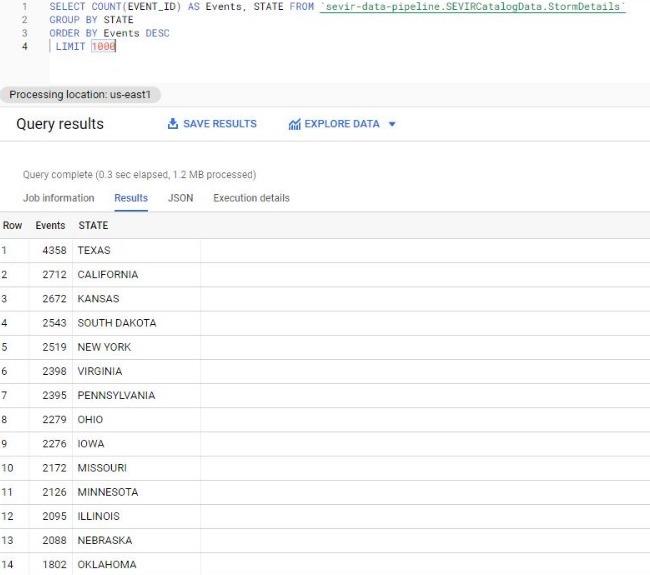
## Join the catalog, StormDetails and Stormlocations

## 

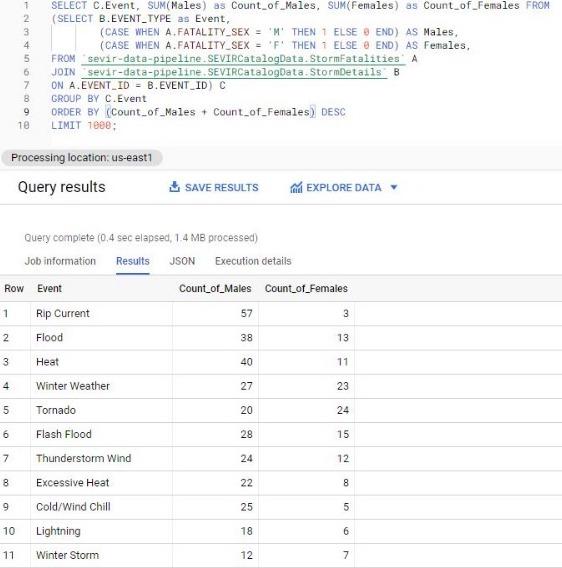
## No of instances per event type

## 

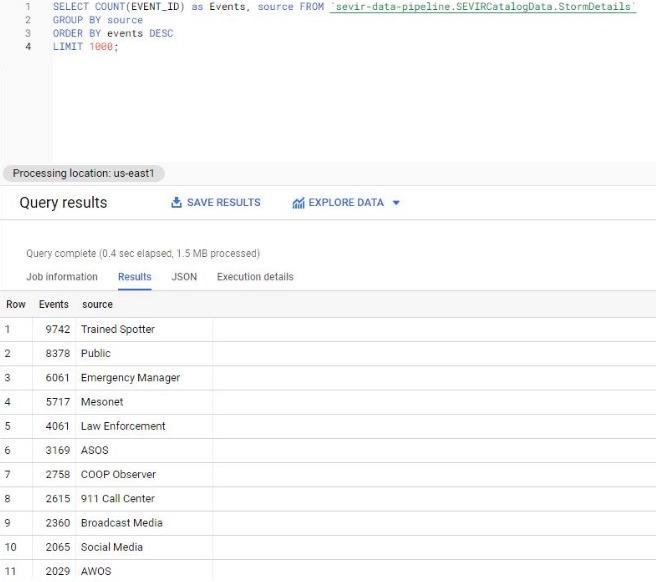
## No of events per State



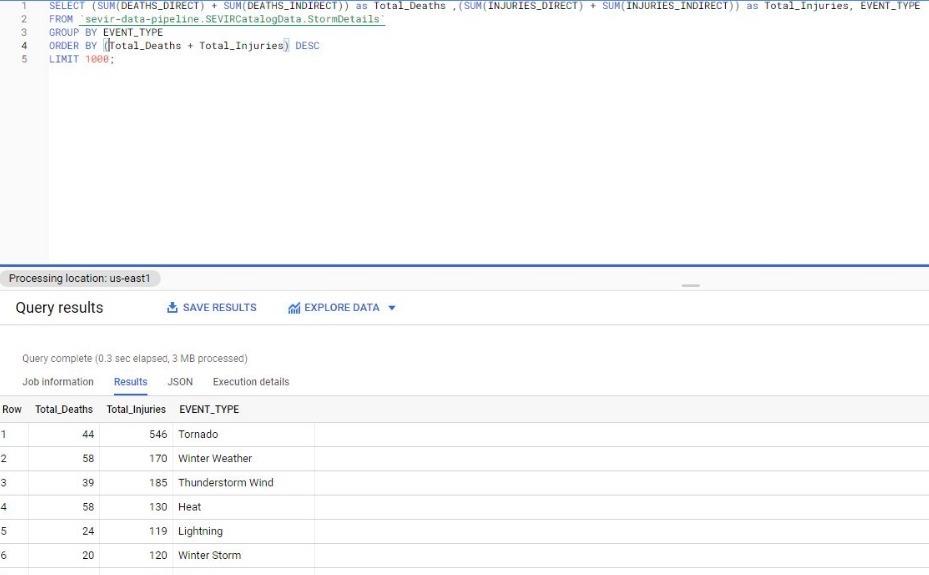
## Fatalities by Gender and event type



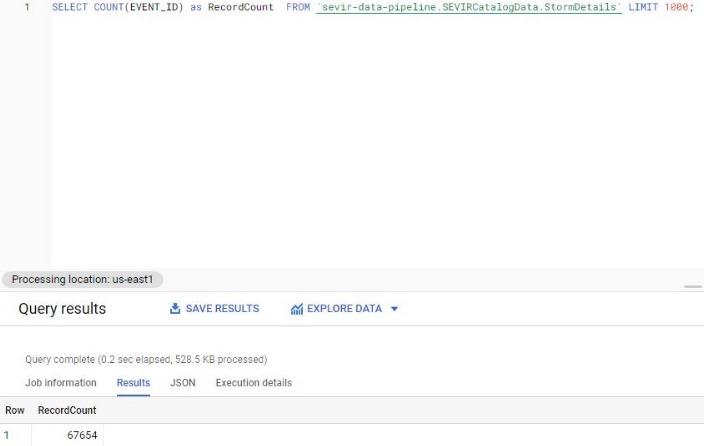
## No of events per data source



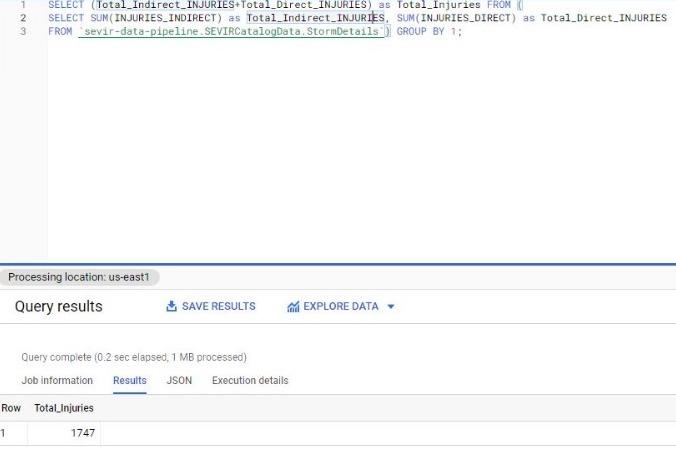
## Death and injury by event type



## Total no of events



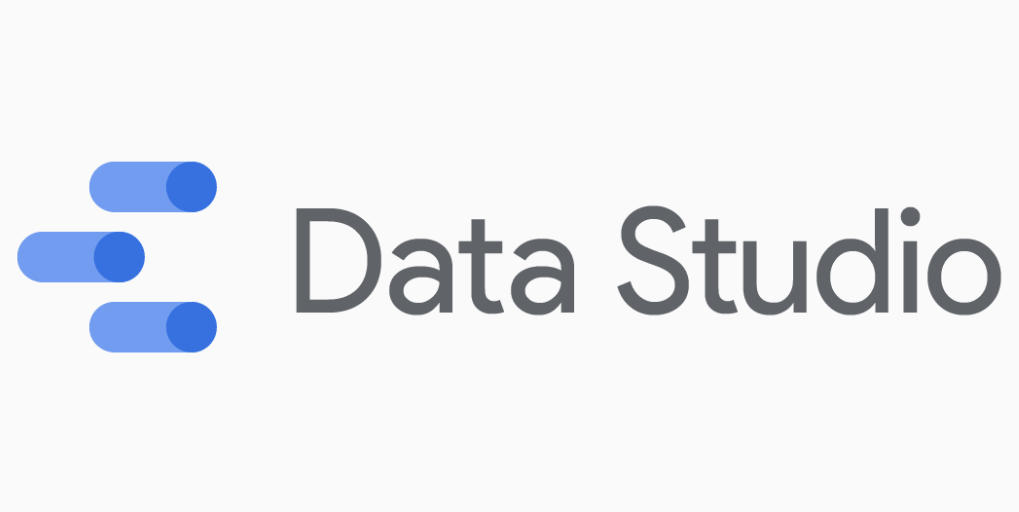
## Total injuries



## Total Deaths



# Data Studio



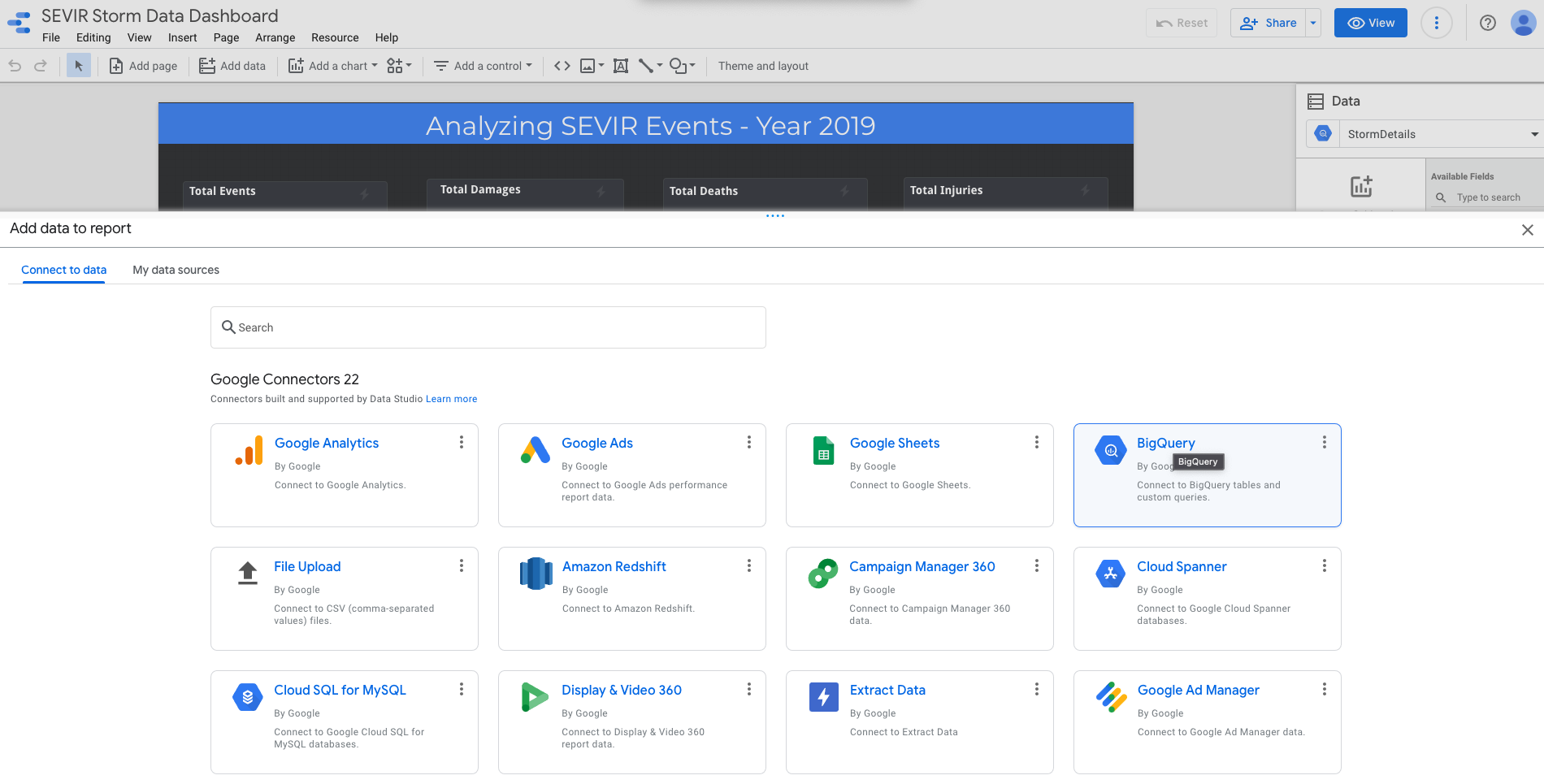
* [Google Data Studio](https://datastudio.google.com/) (GDS) is a visualization tool that lets you build interactive dashboards, and customized, beautiful reporting.
* A majority of Data Studio’s features are easy to use, and it allows for easy sharing and scheduling of reports.

**Benefits of using Data Studio**

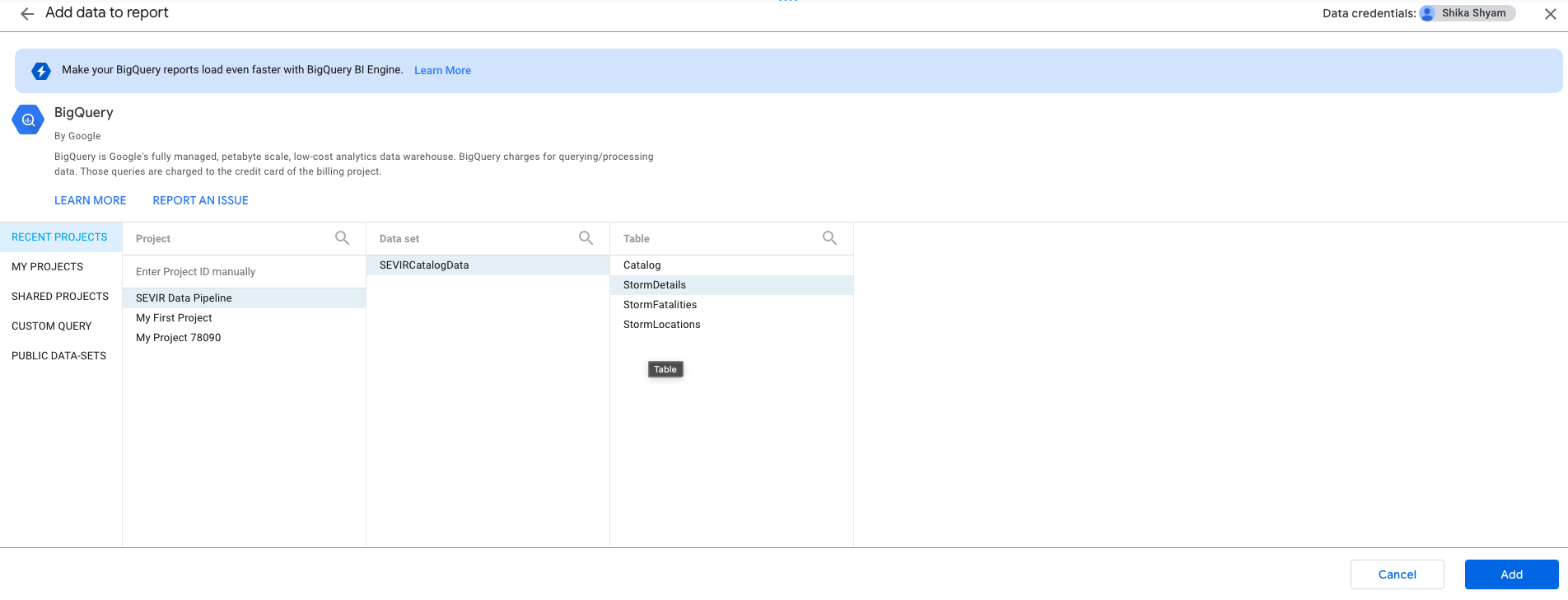
* Live data connection and access to over 220 connectors
* Full control over customization of reports and visuals, including adding as many pages or charts as needed
* Dynamic controls with report and page level filters
* Advanced formulas such as calculated metrics and calculated fields

Now we proceed to perform the operations to generate a report from our data source.

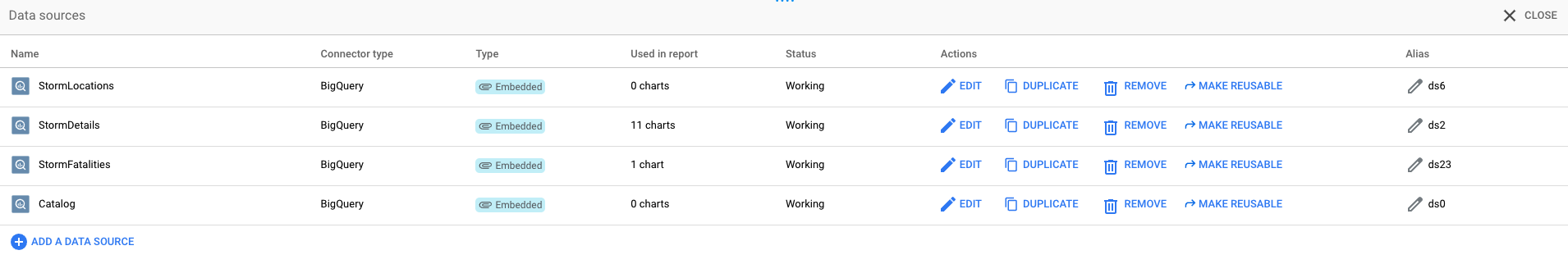
* Go to Add Data and select the BigQuery option to load our dataset



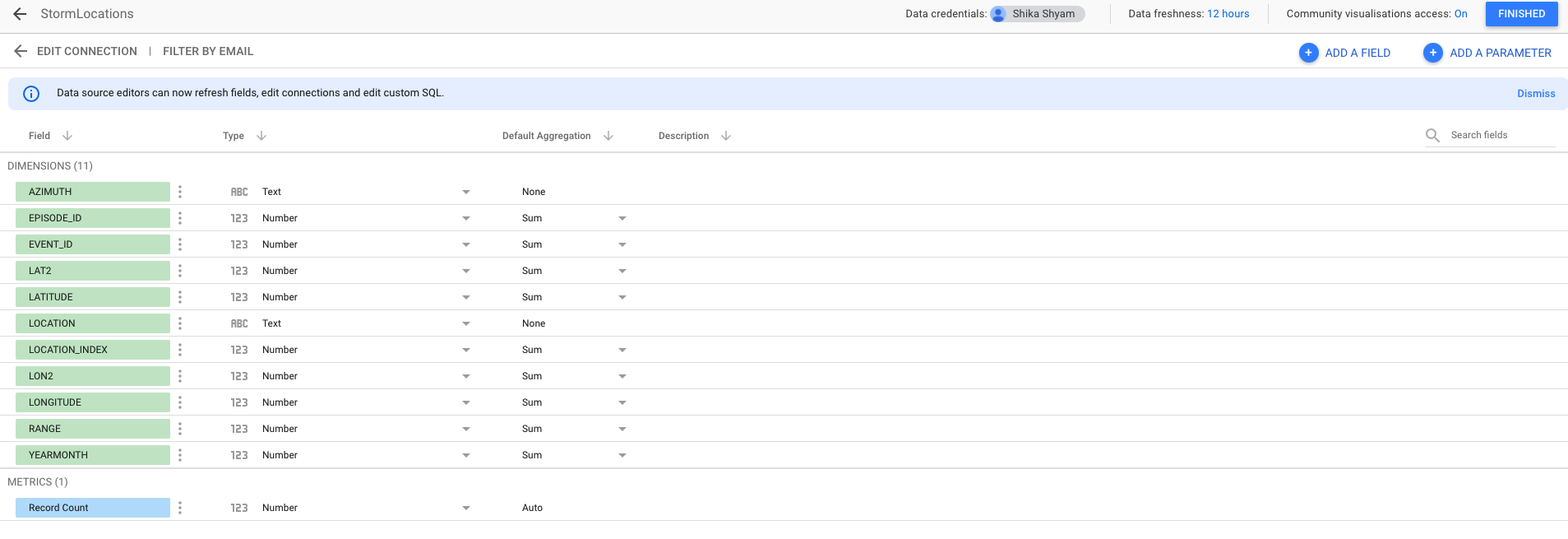
* List of data sources available



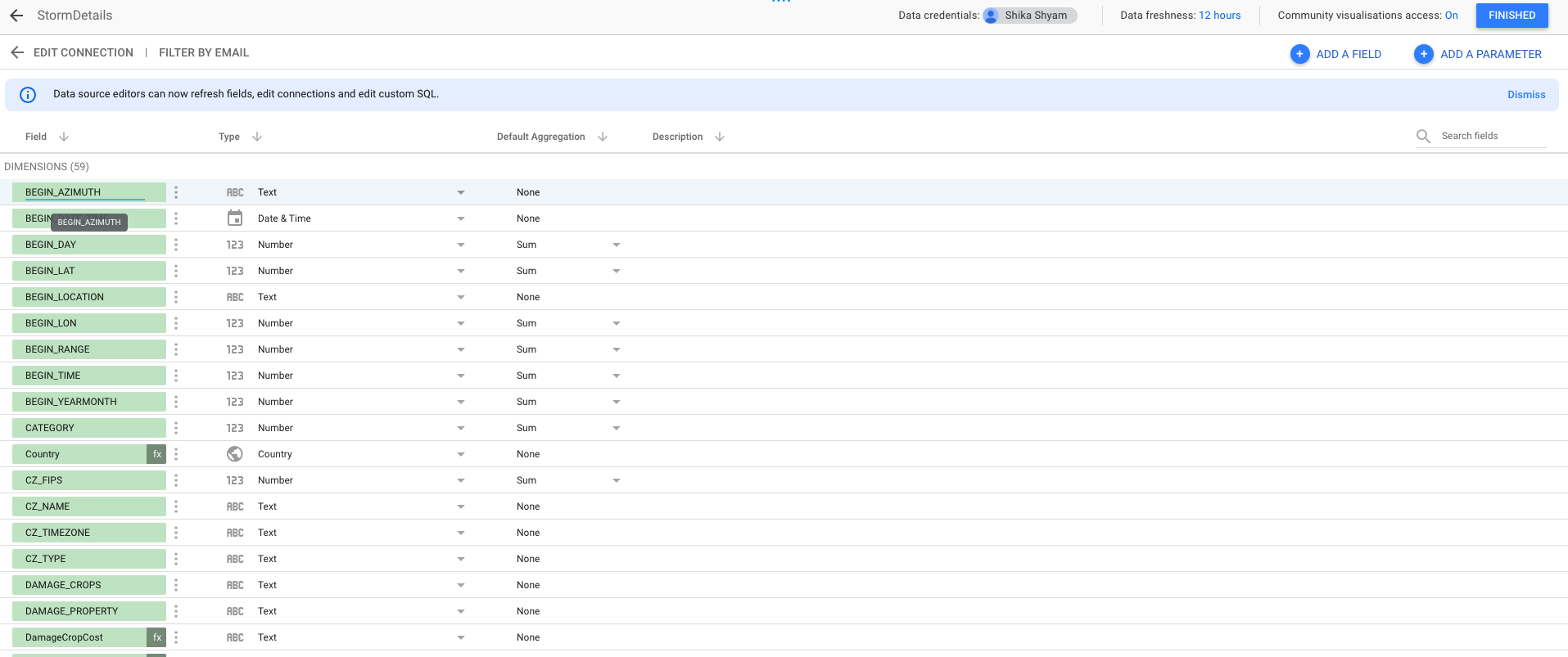
* Available data sources



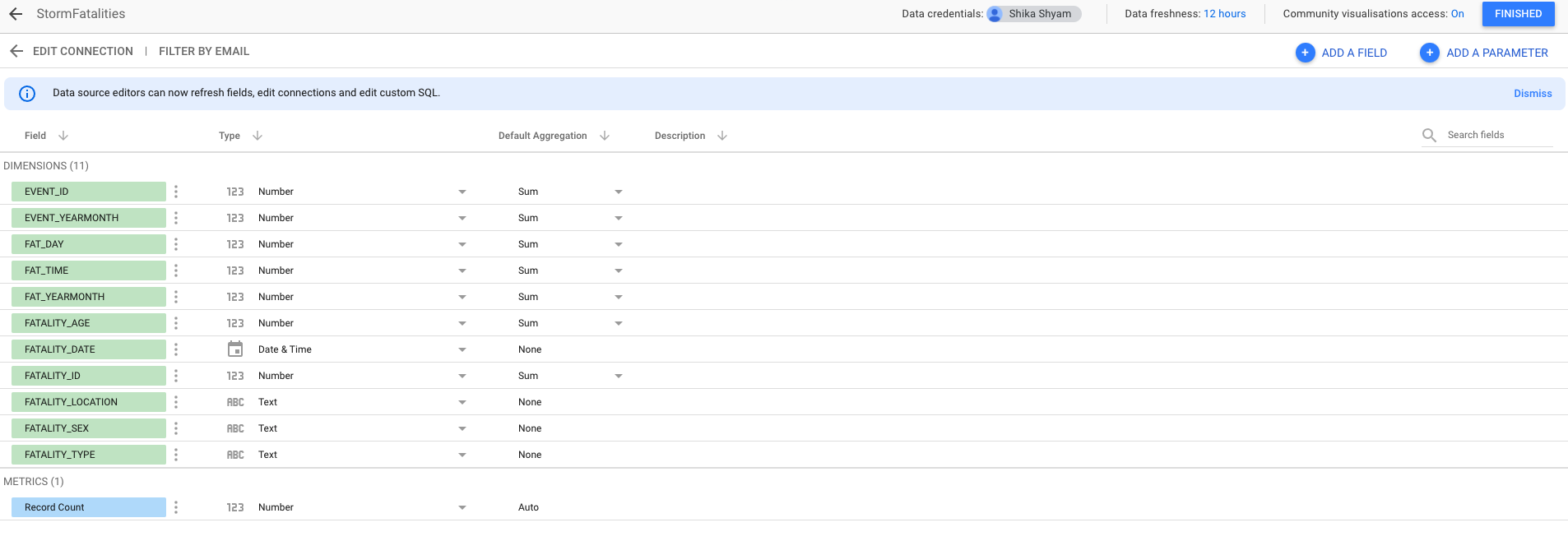
* Storm locations data source



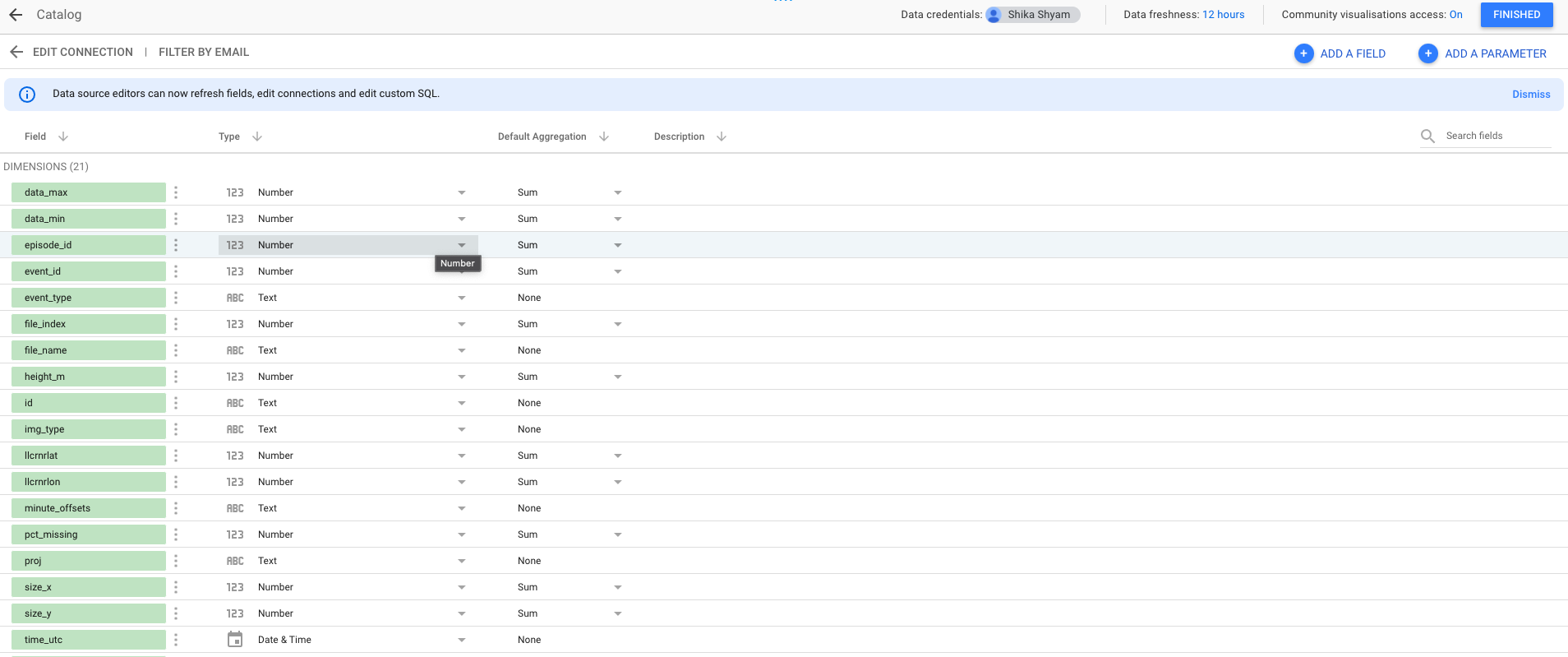
* Storm Details data source



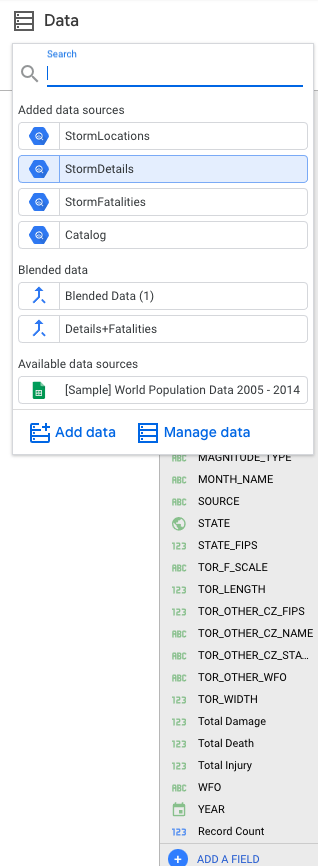
* Storm Fatalities data source



* Catalog data source

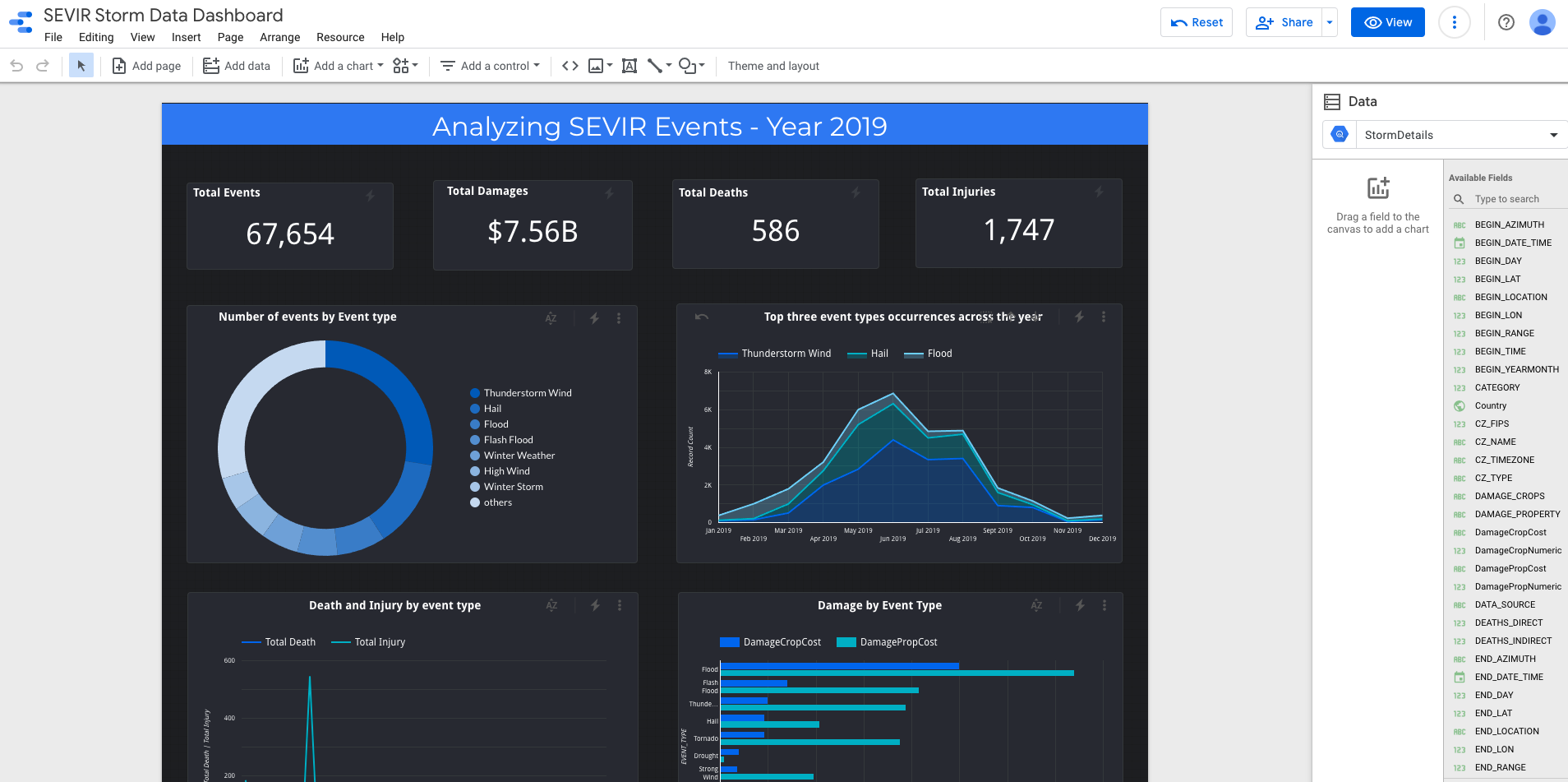


* Chose the appropriate data source



# 

* After selecting the data source drag ang drop the appropriate fields to visualize

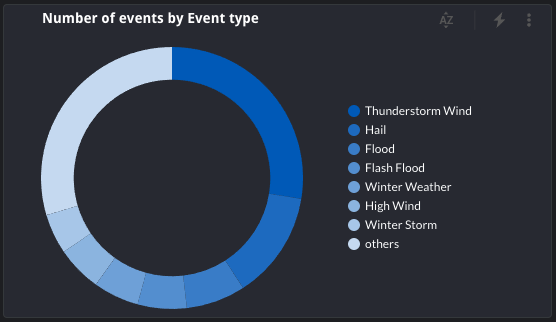


# 

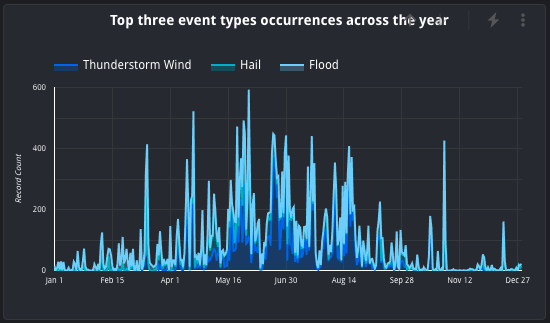
# Dashboards

We perform analysis for the following

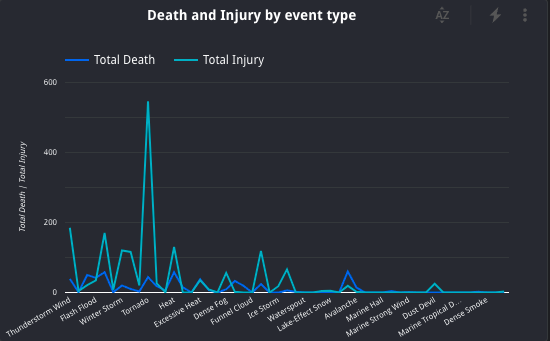
* Number of events by Event type



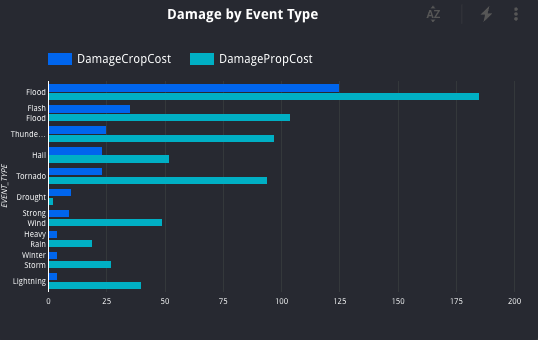
* Top three event types occurrences across the year



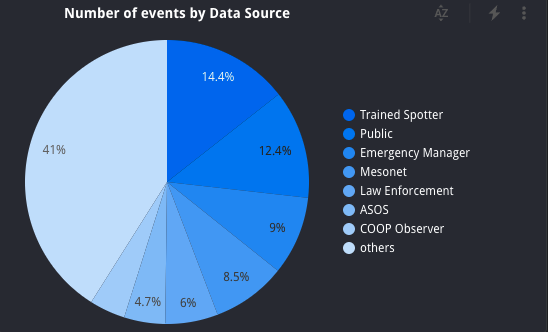
* Death and injury by event type



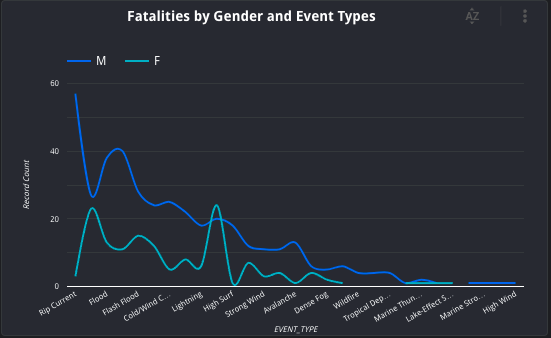
* Damage by event type



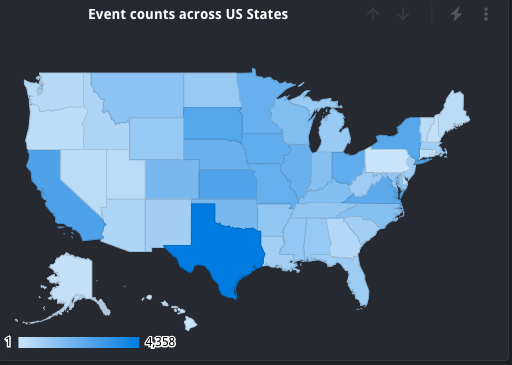
* Damage by event type



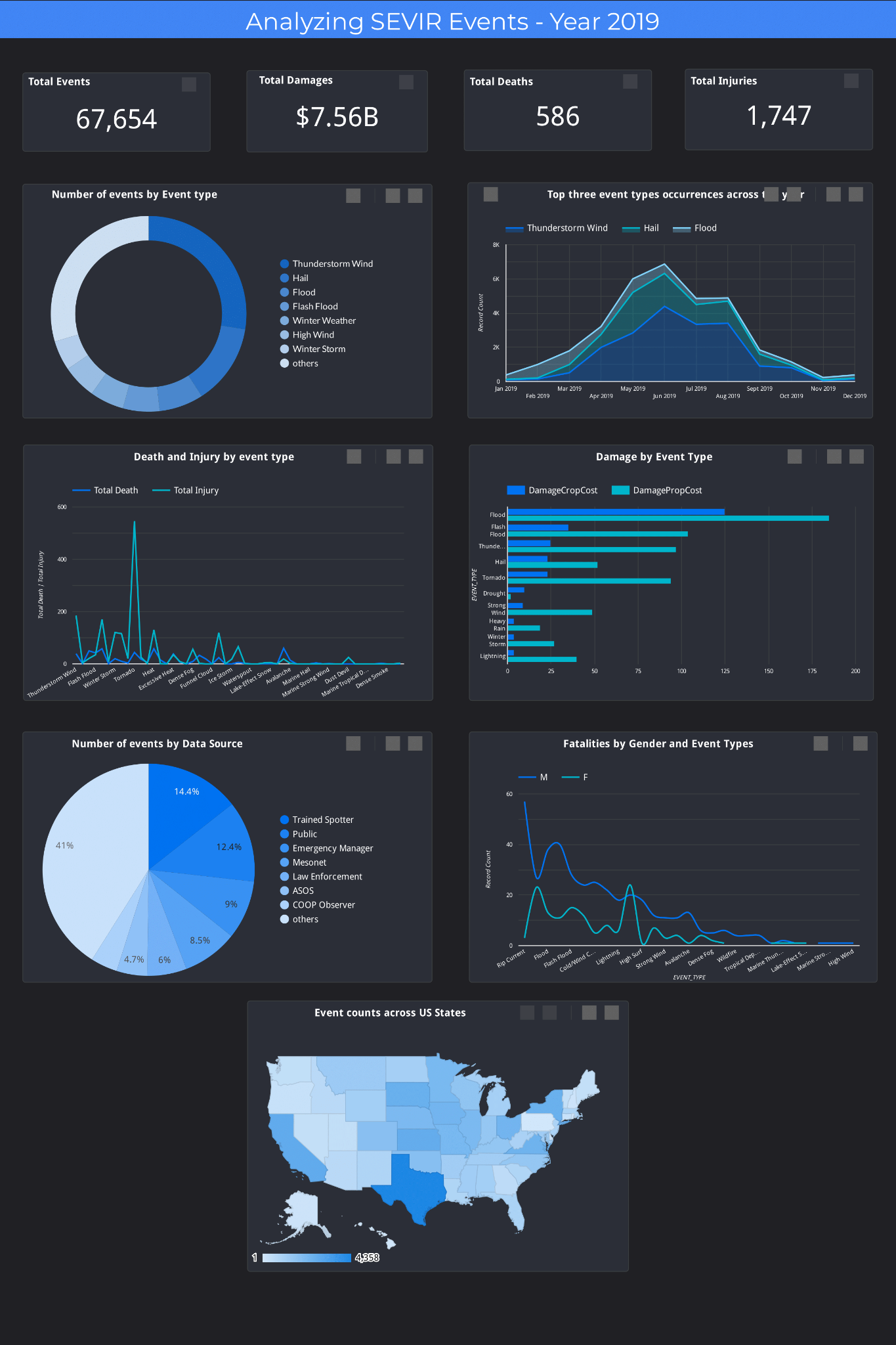
* Fatalities by gender and event types



* Events count across US states



Final Dashboard



# References

* [https://nbviewer.jupyter.org/github/MIT-AI-Accelerator/eie-sevir/blob/master/examples/SE VIR\_Tutorial.ipynb](https://nbviewer.jupyter.org/github/MIT-AI-Accelerator/eie-sevir/blob/master/examples/SE)
* [https://www.ncdc.noaa.gov/stormevents/ftp.jsp](https://nbviewer.jupyter.org/github/MIT-AI-Accelerator/eie-sevir/blob/master/examples/SE)
* [https://cloud.google.com/bigquery/docs/visualize-data-studio](https://nbviewer.jupyter.org/github/MIT-AI-Accelerator/eie-sevir/blob/master/examples/SE)
* <https://proceedings.neurips.cc/paper/2020/file/fa78a16157fed00d7a80515818432169-P>
* [aper.pdf](https://proceedings.neurips.cc/paper/2020/file/fa78a16157fed00d7a80515818432169-P)
* <https://sevir.mit.edu/sites/default/files/About_SEVIR.pdf>
* <https://medium.com/@zarinlo/publish-technical-tutorials-in-google-codelab-format-b07ef76972cd>
* <https://github.com/zarinlo/tools>

# 

# 

# 

# 

# 

# 

# 

# 

### 

# 