CSCI 6444: Intro to Big Data and Analytics(Fall -2023)

System Design Document

Prof. Roozbeh Haghnazar

Team WeatherLink

Alper Cetinkaya

Aditya Gujral

Tharun Saravanan

Matthew

Winchester

Table of Contents

1. Introduction

1.1. Purpose

The purpose of this project is to analyse the correlation between adverse weather conditions and their impact on first responder response times. We aim to use historical data and predictive forecasting to visualise and predict how various weather conditions affect response times, enabling more effective resource allocation and preparedness for emergency situations.

1.2. Scope

The scope of this project includes the development of a big data system that integrates Google Cloud Platform (GCP) Big Data tools, Tableau for visualisation, Apache Spark for data processing, and Airbyte for data integration. It covers data collection, storage, analysis, visualisation, and forecasting.

1.3 Detailed Activity Phases Description

The architectural framework comprises three primary phases: Data Ingestion, Data Model Generation, and Data Visualization. The descriptions here follow along with the activity diagram flow charts for each phase.

1.3.1 Data Ingestion:

This flowchart explains the process of data ingestion within our system which encompasses the orchestrated transfer of diverse data streams from assorted sources to a storage medium where it can be accessed, used and analysed in further steps. The initial steps of this data pipeline entail the periodic acquisition of data from different sources, including Census data, weather data and accident data, as the trigger mechanism is activated. Subsequently, the acquired data is cleaned in a way that the unnecessary parts of the data are removed, null values are systematically addressed and the data is formatted to conform to our designated storage technology specifications. The resultant refined data is then loaded into the Master Google Big Query table, setting the stage for further analysis and utilisation.

1.3.2 Data Modelling:

This flow chart represents the process of data model generation activity. The process starts with the data generation job which is responsible for generating data to predict what will happen based on the previous data from prior years. This data is then retrieved from the master GBQ table. If the data has changed since the last training, the model trains on historical data. Once the model is ready, it generates predictions which are then stored in the master GBQ table. If there is no new data since the last training, the model generates predictions without training.

1.3.2 Data Visualization:

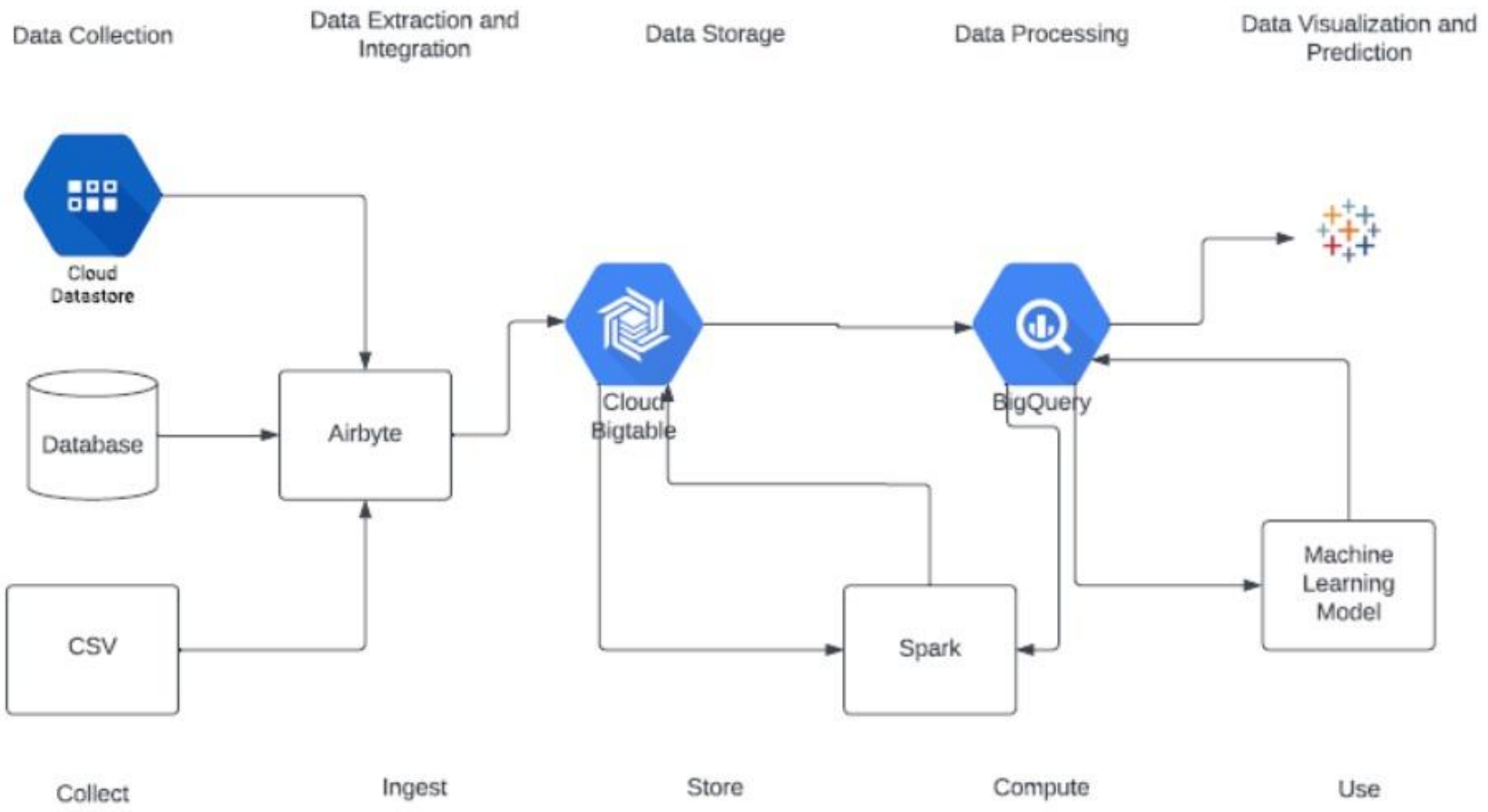
The data visualisation flow chart serves as a graphical representation of the process of data visualisation activity with the use of actions in Tableau platform. By default, data from a certain date and a predefined map location is selected and presented visually. In the later steps, the users are able to modify the date, altering the orientation of the map and specifying a particular data for visualisation. When users make adjustments to these defaults, the visualisation configuration is dynamically updated, triggering the retrieval of the corresponding data from the master Google Big Query Table. Subsequently, the revised configuration is visualised.

2. System Architecture

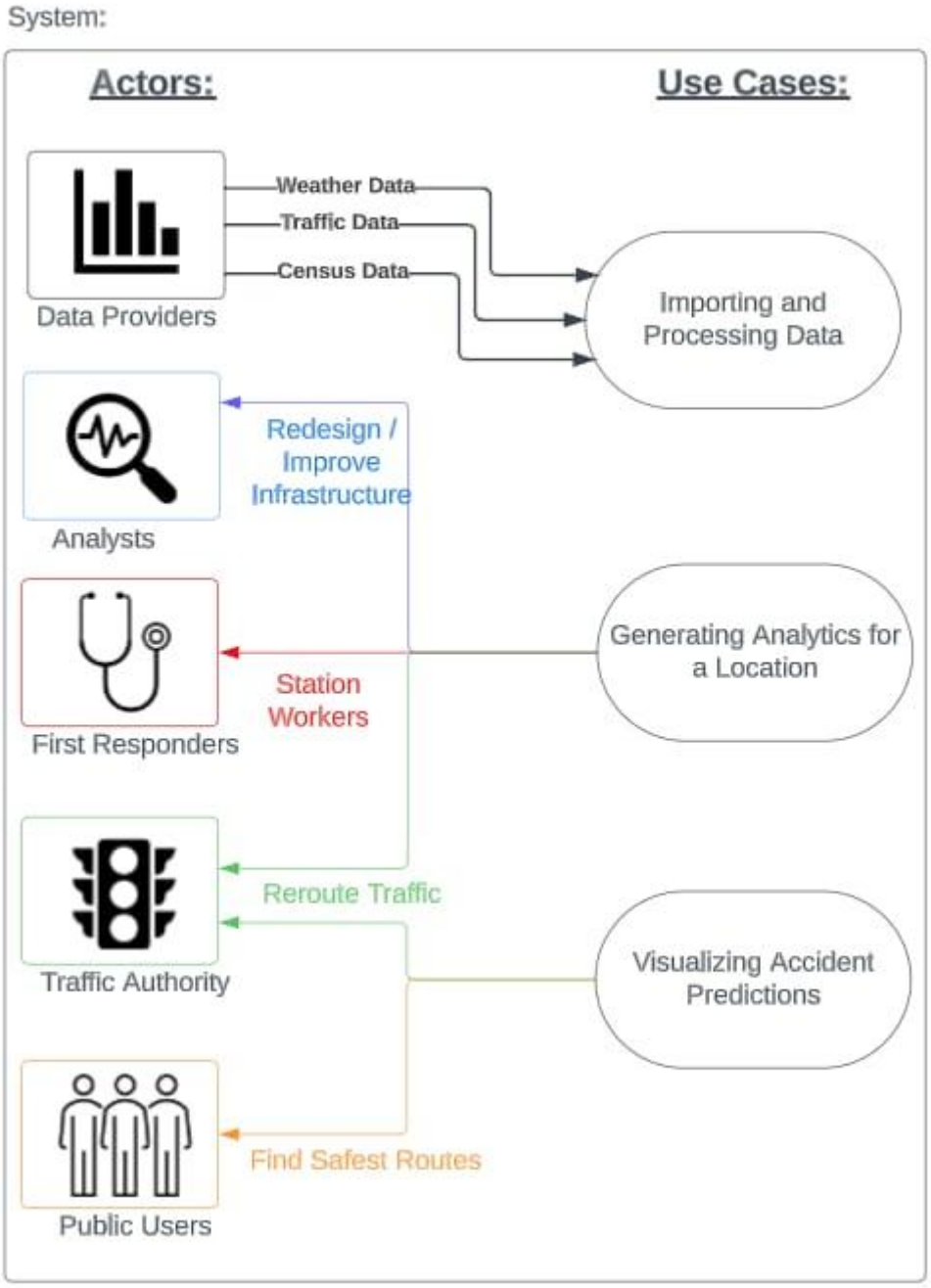
2.1 Architectural Design Choices

Design Aspect	**Key Technology**	**Purpose**
Data Sources	We will use data sources from Kaggle, Google Big Query public datasets, and NOAA Weather data	Required data for the project
Data Extraction and Integration	Airbyte Python Google Big Query SQL	Extraction of data in manual batches and on monthly intervals. (The NOAA data in particular will require custom Python tooling to read from their proprietary format)
Data Storage	GCP	Google Big Query Tables will store the cleaned data once it has been ingested.
Data Processing	GCP, BigQuery, Spark	Data will be cleaned as it is ingested before final storage in the master Big Query Table
Data Prediction	ML Model	We will train a model using historical data to produce predictions for future dates
Data Visualization	Tableau	Tableau will be used to create a dashboard to interface with the data and create visualisations

2.2 System Architecture Diagram

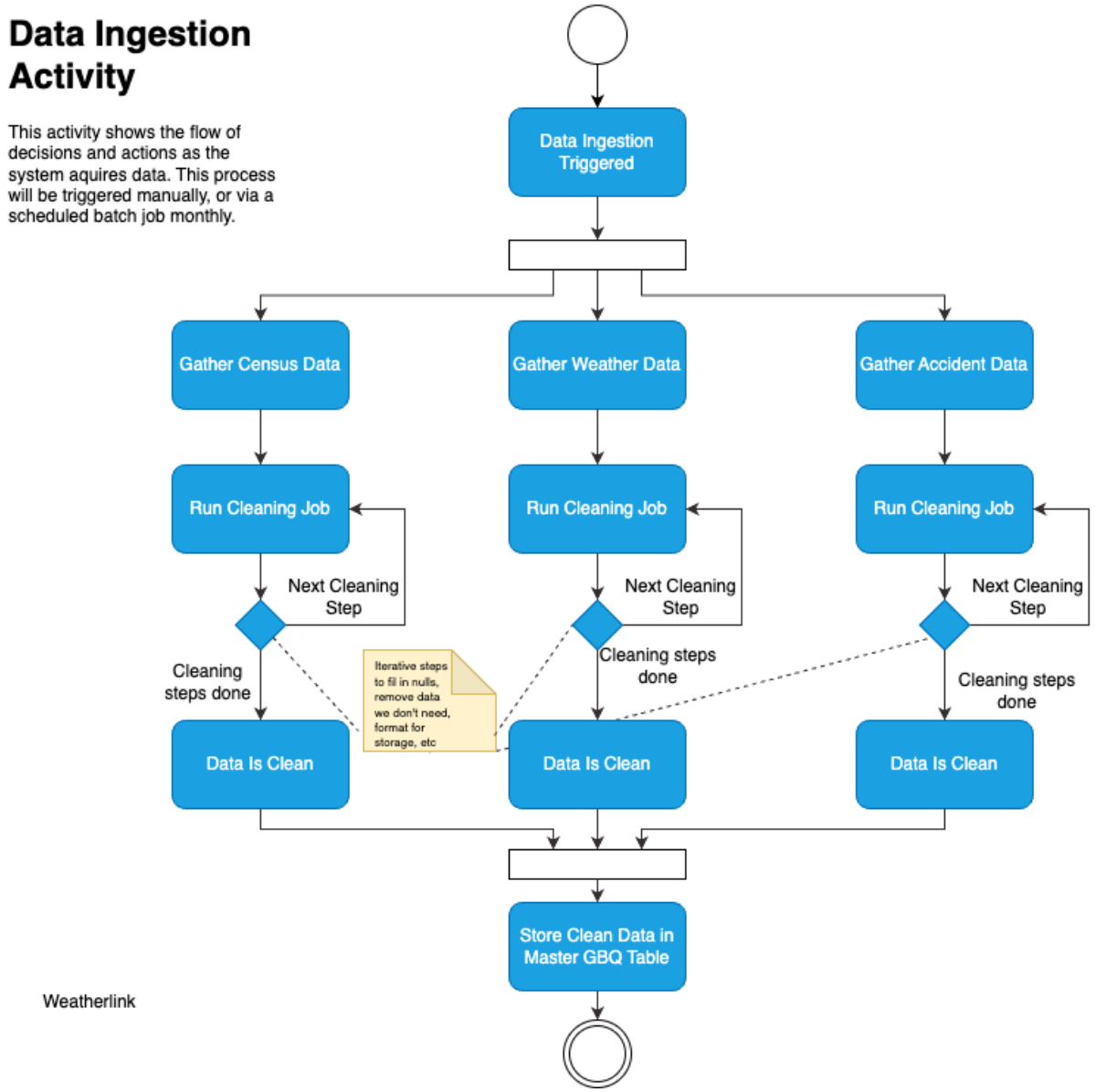


2.3 Use Case Diagram



2.4 Activity Diagrams

2.4.1 Data Ingestion

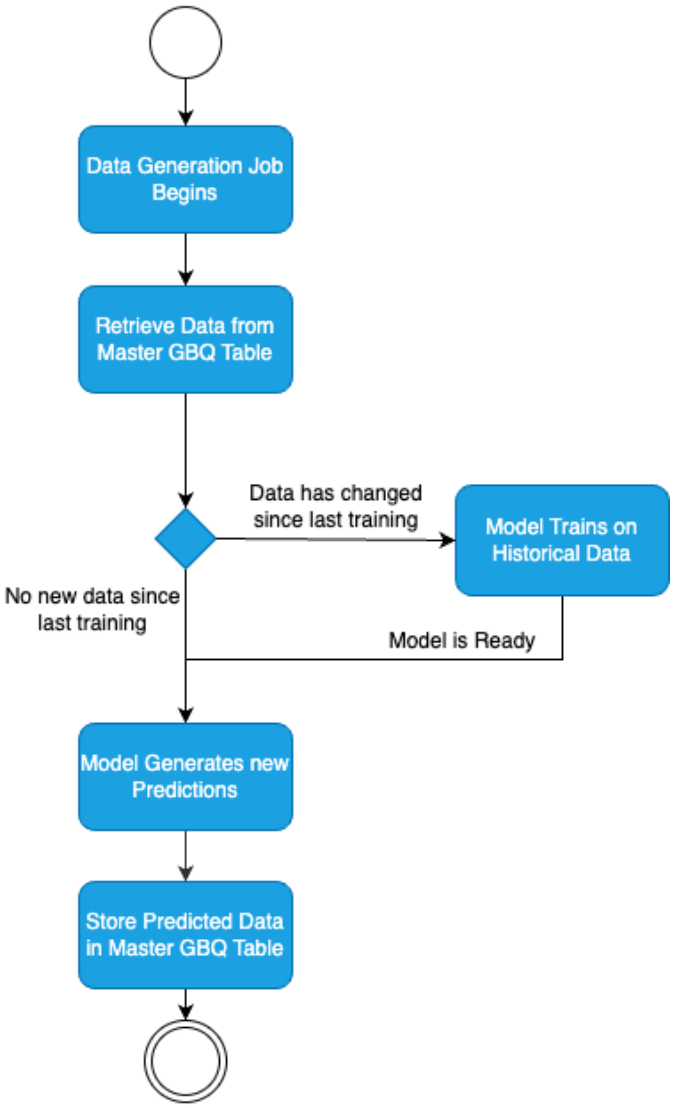


2.4.2 Data Prediction Generation

Data Model Generation Activity

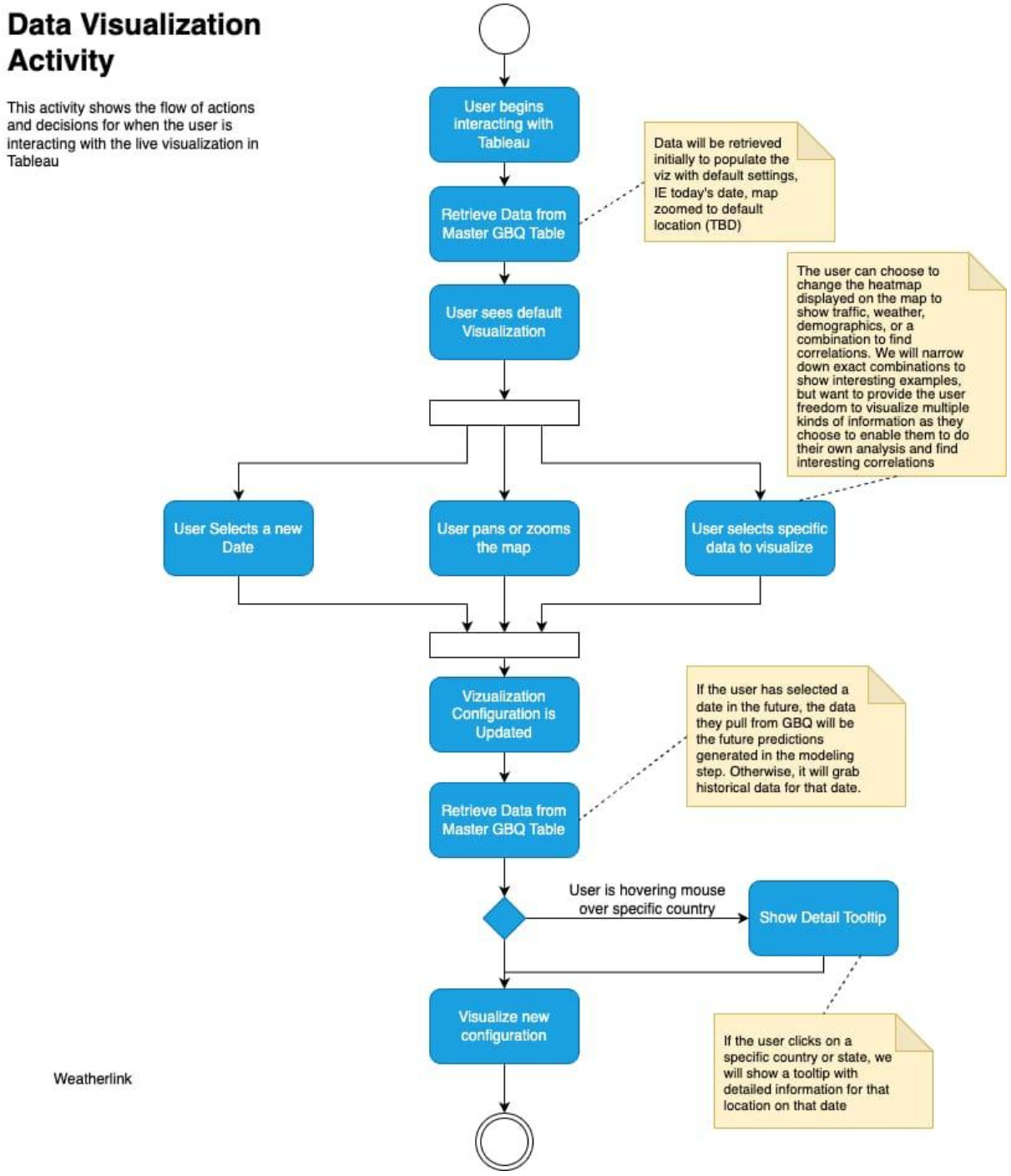
This activity shows how the system will generate data to visualize for future dates, predicting what will happen based off of previous data from prior years. This will allow the user to visualize historical data, as well as see a probable future scenario.

This will likely run monthly, to generate more future predictions as historical data replaces previous predictions.



Weatherlink

2.4.3 Data Visualization



3. Appendix

3.1. References

3.1.1. Data Sources we will use

Dataset	Usage
Weather Based Datasets:	
NOAA's: Daily Weather Dataset https://www.ncei.noaa.gov/access/crn/qcdatasets.html	Freely accessible weather data across North America
NOAA's lightning strike dataset: https://console.cloud.google.com/marketplace/product/noaa-public/lightning	Historical lightning strike data providing the origin of recorded lightning strikes on a daily basis starting from 1997
NOAA's Severe storm dataset: https://console.cloud.google.zom/marketplace/product/noaa-public/severe-sztorm-events	Historical major storm data including storm event's location, azimuth, distance, impact, and severity from 1950
Traffic Datasets:	
NHTSA Traffic Fatalities: https://console.cloud.google.com/marketplace/product/nhtsa-data/nhtsa-traffic-fatalities	Traffic Incident Data, including types of cars and roads, the manoeuvres that preceded the accident, and the involvement of pedestrians and cyclists.
US Accidents (2016-2023): https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents	Comprehensive dataset including geospatial location alongside key weather timestamps of traffic accidents
Miscellaneous Linkage Data:	
Census Bureau Data: https://console.cloud.google.com/marketplace/product/united-states-census-bureau/acs?q=search&referrer=search&project=mattbigquerytest1&pli=1	Miscellaneous data that can be used to link traffic data to weather based datasets

3.1.2. Papers studied in preparation of this project

1. Gutierrez-Osorio, C., & Pedraza, C. (2020). Modern data sources and techniques for analysis and forecast of road accidents: A review. Journal of Traffic and Transportation Engineering (English Edition), 7(4), 432–446. doi:10.1016/j.jtte.2020.05.002
2. Theofilatos, A. (2019). Utilizing Real-time Traffic and Weather Data to Explore Crash Frequency on Urban Motorways: A Cusp Catastrophe Approach. Transportation Research Procedia, 41, 471–479. doi:10.1016/j.trpro.2019.09.078
3. Alam, M. M., Torgo, L., & Bifet, A. (2022). A survey on spatio-temporal data analytics systems. ACM Computing Surveys, 54(10s), 1-38.
4. Mehdizadeh, A., Cai, M., Hu, Q., Alamdar Yazdi, M. A., Mohabbati-Kalejahi, N., Vinel, A., ... & Megahed, F. M. (2020). A review of data analytic applications in road traffic safety. Part 1: Descriptive and predictive modelling. Sensors, 20(4), 110
5. Data Integration: The Teenage Years Alon Halevy Google Inc. halevy@google.com Anand RajaramanKosmix Corp .anand@kosmix.com Joann Ordille Avaya Labs joann@avaya.com
6. A Framework for Scalable Real-Time Anomaly Detection
over Voluminous, Geospatial Data Streams
Walid Budgaga, Matthew Malensek, Sangmi Lee Pallickara, and Shrideep Pallickara
Computer Science Department, Colorado State University, Fort Collins, CO, USA
7. Jeffrey Dean, Sanjay Ghemawat. 2004. MapReduce: Simplified Data Processing on Large Clusters. Google, Inc.
https://static.googleusercontent.com/media/research.google.com/en//archive/mapreduce-osdi04.pdf
8. Konstantin Shvachko, Hairong Kuang, Sanjay Radia, Robert Chansler. 2010. The Hadoop Distributed Filesystem. Apache Foundation.
https://hadoop.apache.org/docs/r1.2.1/hdfs_design.pdf