Executive Summary Document

This project is all about digging into data from Denton to get some insights on crime and traffic incidents. We worked with two different datasets—one with real-time crime data and the other with records of traffic cases from September 2024. Our goal was to understand what's happening in Denton using the power of data analysis and make it easier for city officials to act on that information.

Overview of the Use Case and Data Lifecycle

The key focus of this project was to make sense of both real-time and historical data. We wanted to give the city a clearer picture of the patterns behind these incidents. To do this, we took the data through several phases:

1. Data Collection and Storage:
    - For the crime data, we pulled it in batches using the City of Denton's public API. Instead of grabbing all the data at once (which would be too much), we fetched it in chunks to keep things efficient.
    - For the traffic cases dataset, we first cleaned it up using OpenRefine, then uploaded it to Google Cloud for easy access and future processing.
2. Data Handling and Processing:
    - We used a method called pagination to make sure we didn’t overload our system. This involved fetching a manageable number of records at a time. We also cleaned up the column names so they could be used in BigQuery without any issues.
    - We filtered the crime data by date, which helped us focus on specific days for more detailed insights. This filtered data was also uploaded to BigQuery for further analysis.
3. Analyzing the Data:
    - To make sense of the data, we used BigQuery, Hive, and Spark. These tools helped us answer key questions, like where incidents were most common, what types of violations happened most frequently, and which areas had the most activity.
    - With Hive and Spark, we could process the data stored in our Hadoop system, while BigQuery let us run more in-depth queries to create reports.
4. Making the Data Available:
    - We saved a filtered version of the data as a CSV file in Google Cloud Storage. This makes the data easy to share and work with further, either for analysis or archiving.

Tools We Used

We used a bunch of tools to make all this happen:

- Google Cloud Platform (GCP): To store the data and set up virtual machines.
- Hadoop: For processing the static dataset in a way that could scale if needed.
- BigQuery: This was our go-to for analyzing large datasets, allowing us to easily query both crime and traffic records.
- Hive and Spark: These were used for processing data stored in our Hadoop setup, giving us the ability to dig deeper into the numbers.

Key Takeaways

From the crime data, we found that Gregory Creek Drive had the highest number of incidents. Not surprisingly, Denton itself topped the list for overall violations. When we looked at the traffic data, speeding was the biggest issue, and zip codes 76201 and 76209 had the highest number of incidents.

References:

1. Daniel Gutierrez(November 30, 2015), An overview of spark SQL, Inside AI news, https://insideainews.com/2015/11/30/an-overview-of-spark-sql/
2. Petrova-Antonova, D., & Tancheva, R. (2020, September). Data cleaning: a case study with OpenRefine and trifacta wranglerLinks to an external site. In International Conference on the Quality of Information and Communications Technology (pp. 32-40). Springer, Cham, https://www.researchgate.net/profile/Dessislava-Petrova-Antonova/publication/343983146_Data_Cleaning_A_Case_Study_with_OpenRefine_and_Trifacta_Wrangler/links/5fc6d32ba6fdcc697bd33a0c/Data-Cleaning-A-Case-Study-with-OpenRefine-and-Trifacta-Wrangler.pdf
3. George Lawton (17 Feb 2022), Hadoop vs. Spark: An in-depth big data framework comparison, Tech Target, https://www.techtarget.com/searchdatamanagement/feature/Hadoop-vs-Spark-Comparing-the-two-big-data-frameworks
4. Tim Stobierski (02 Feb 2021), 8 steps in Data life cycle, Harvard Business School, https://online.hbs.edu/blog/post/data-life-cycle