Calgary traffic incidents data analysis

Design Document

Thi Uyen Thi Phan

2/13/24

Data Science Capstone Project

Table of Contents

1. Introduction

1.1 Purpose

The design document for the "Calgary traffic incidents analysis" project serves as a comprehensive blueprint outlining the structure, functionality, and implementation strategy of the project. Its purpose is to provide a clear understanding of the project's objectives, scope, requirements, architecture, and technical details to all stakeholders involved.

1.2 Scope

The Calgary traffic incidents analysis project investigates the dynamics of traffic incidents within Calgary, Alberta in the last five years from 2019 to 2023, considering various factors such as community characteristics and weather conditions. Focusing on patterns, trends, and influential factors, the project is essential for urban planning, public safety, and resource allocation, especially as urban populations expand, and transportation infrastructure evolves. Utilizing comprehensive datasets sourced from municipal authorities, this project employs advanced data analytics techniques to analyze traffic incidents, including collisions, road hazards, and related events.

Through exploratory data analysis, statistical modeling, and machine learning algorithms, this study aims to uncover spatial and temporal distributions of traffic incidents across neighborhoods, weather, and times of day. By identifying high-risk areas, contributing factors, and recurring patterns, the research informs targeted interventions, traffic management strategies, and infrastructure improvements to reduce incident frequency and severity.

Project objectives and audiences:

By pinpointing the audience as Calgary citizens and government, three objectives have been identified:

- Investigate trends and patterns of traffic incidents in Calgary.
- Raise awareness among the audience regarding high-risk factors when navigating traffic.
- Offer recommendations to the Calgary government for mitigating traffic incidents.

2. System Overview

2.1 System Architecture

- **Data Sourcing:** Datasets are collected from trustable sources including: https://data.calgary.ca/, https://climate.weather.gc.ca, https://www150.statcan.gc.ca.
- **Data Storage:** After preprocessing process, data will be stored in Azure SQL database named: **Traffic_Incidents.**
- **ETL Process:** Data will be loaded directly from websites via API or from local machine to Jupiter notebook to be cleaned here before being imported to SQL database.
- **Data Modelling:** The Star-schema will be used.
- **Visualization:** PowerBI will be used to analyze and create dashboards. The dashboard after creating will be published to PowerBI workplace.

2.2 System Components

- **User Interface:** The user interface component is developed PowerBI. It features an intuitive and responsive design that allows users to easily view visualizations, filter, search and interact with traffic incidents and other related data.
- **Backend Processing:** The backend processing component is implemented using Python through Jupyter notebook and Power Query in PowerBI. It handles collecting and cleaning data, import data to the database, retrieving and manipulates data from the database, executes data analysis algorithms, and generates relevant insights and reports to be displayed on the user interface.
- **Database:** The database component utilizes Serverless Azure SQL database as the relational database management system (DBMS). It consists of multiple tables, including Traffic_Incidents, WeatherData, Quadrants, Community, Street, and DateTable, which store detailed information about traffic incidents, daily weather, geographical locations, and calendar respectively. Foreign key constraints and indexes are applied to ensure data integrity and optimize query performance.

3. Data Management

3.1 Data Sourcing

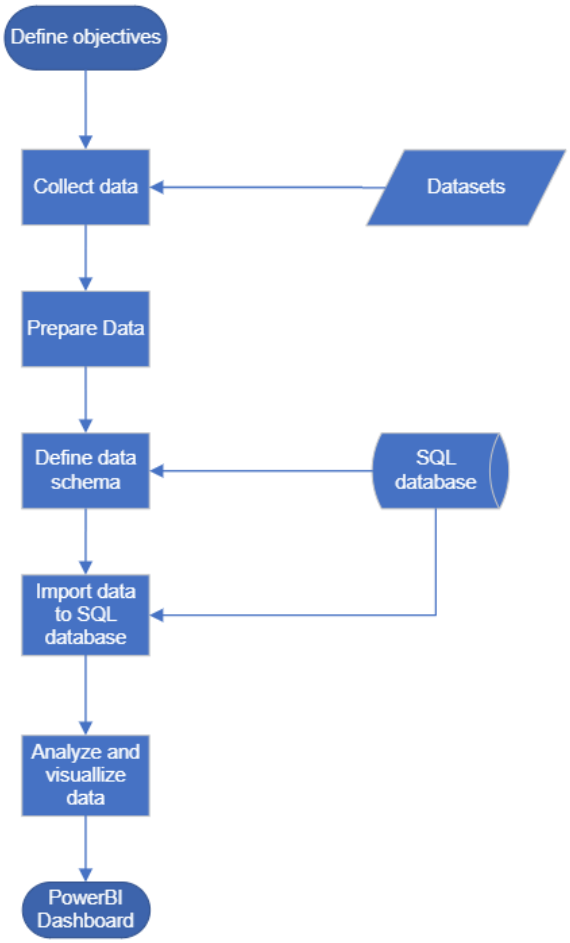
- **Calgary traffic incidents dataset:** This is a live dataset which contains information about all reported traffic incidents in Calgary from 2016, within the scope of this project, only data from 2019 to 2023 is selected.
  Data source: https://data.calgary.ca/Transportation-Transit/Traffic-Incidents/35ra-9556/data_preview.
- **Calgary community boundaries:** This is a geographic dataset which contains information about all communities in Calgary including Name, Class, Sector and Boundary.
  Data source: https://data.calgary.ca/Base-Maps/Community-Boundaries/ab7m-fwn6
- **Calgary street centreline:** This is a geographic dataset which contains information about all streets within Calgary, including Name, Full name, Type, Class, and Boundary. These streets are split into smaller lines at the intersection of the streets.
  Data source: https://data.calgary.ca/Transportation-Transit/Street-Centreline/4dx8-rtm5/data_preview
- **Daily weather in CALGARY INTL A station:** This dataset contains daily weather which was collected in INTL A Station (located at Calgary International Airport) including: Temperature, Snow, Rain, etc.
  Data source: https://climate.weather.gc.ca
- **Calgary four quadrants' boundaries:** This is a geographic dataset which contains Name and Boundary of four quadrants in Calgary (Southeast, Southwest, Northeast, Northwest).
  Data source: https://data.calgary.ca/Base-Maps/City-Quadrants/g2n2-qnvh/data_preview
- **Calgary downtown boundary:** This is a map which present boundary of Calgary Downtown. Since this is only an image, it is needed to convert it to Geospatial dataset.
  Data source: https://www150.statcan.gc.ca/n1/pub/91f0015m/2021001/m-c/m-c-calgary-eng.pdf

3.2 ETL Process

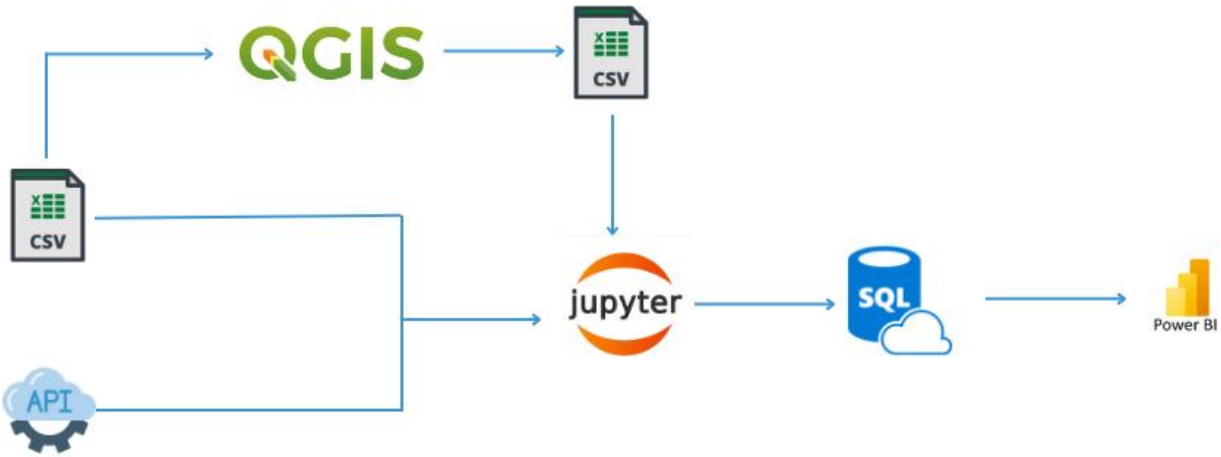
Step 1: Define objectives of the project.

Step 2: Research and select suitable datasets for the project. Identify formats of the data sources.
Step 3: Data preprocessing
- Utilize QGIS to handle Geospatial datasets that require advanced GIS techniques.
- Employ Python to transform data and handle missing values.

Step 4: Define data schema and create SQL database.
Step 5: Import data into the SQL database using Python.
Step 6: Analyze data and construct PowerBI dashboard.
ETL Process is described in the following chart:

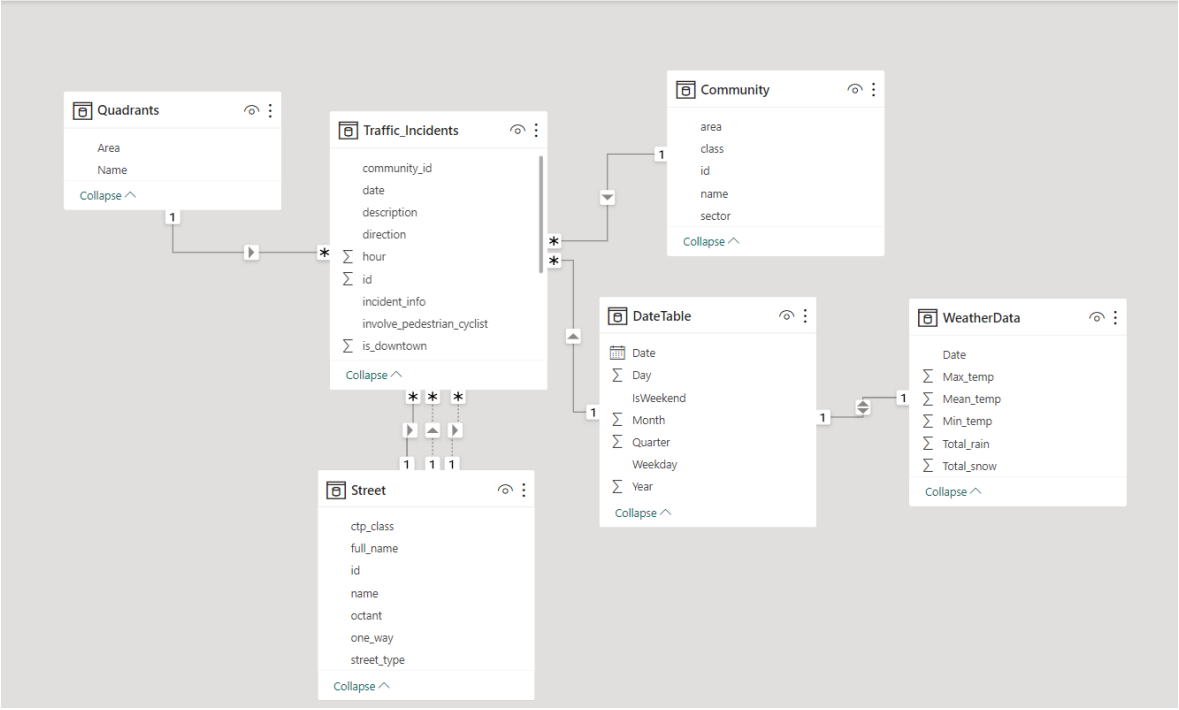


The dataflow of the project is described in the following flow chart:



4. Data Modelling

4.1 Entity-Relationship Diagram



4.2 Schema Design

This is a star-schema with one fact tables and four dimensional-tables:

Table Name	Column Name	Column Type	Columns description
Traffic_incidents (Fact table	Id	Int, Primary Key	Uniquely identify each row of table
	Date	Date	The date the incident happened
	Hour	Int	The hour the incident happened
	Incident_info	Text	Information about incident locations
	Description	Text	Short description about incidents
	Quadrant	Text (NE, NW, SE, SW, Unknown)	Name of the quadrant in which the incident happened. Unknown if there is no information about quadrants.

	Quadrant_excl_DT	Text, Foreign Key (NE, NW, SE, SW, Downtown Unknown)	Name of the quadrant in which the incident happened. “Downtown” if it happened in Downtown
	Latitude	Float	Location of the incidents
	Longitude	Float	
	Community_id	Int, Foreign Key	The id of the community in which the incident happened
	Is_downtown	Int	1: if it happened in downtown 0: it didn’t happen in downtown
	Direction	Text	The lanes that were blocked due to the incidents
	No_of_vehicle	Text (Single Vehicle/Two Vehicles/ Multiple Vehicles)	Number of vehicles attended in the incident.
	Street_id_1	Int, Foreign Key	The id of the street in which the incident happened. In some case, the incidents were in the intersection of 3 streets
	Street_id_2		
	Street_id_3		
Community Dimensional table	Id	Int, Primary Key	Uniquely identify each row of table
	Name	Text	Name of the community
	Sector	Text	Sector of the community
	Area	Float	Area of the community
Quadrants Dimensional table	Name	Text, Primary Key	Name of the quadrants
	Area	Text	Area of the quadrants
Street Dimensional table	Id	Int, Primary Key	Uniquely identify each row of table
	Full_name	Text	Full name of the street (usually includes Name of the Street and the quadrant)
	Name	Text	Name of the Street
	Street_type	Text	Type of the street
	Ctp_class	Text	Class of the street
WeatherData Dimensional table	Date	Date, Primary Key	
	Max_temp	Float	The maximum temperature of the day in Celsius degree
	Mean_temp	Float	The average temperature of the day in Celsius degree
	Min_temp	Float	The minimum temperature of the day in Celsius degree
	Total_rain	Float	The total rainfall of the day in mm
	Total_snow	Float	The total cm of snow of the day
DateTable Dimensional table	Date	Date, Primary key	
	Year	Int	
	Quarter	Int	
	Month	Int	
	Day	Int	
	Weekday	Text (From Monday to Sunday)	
	IsWeekend	Binary	

5. Visualization & User Interface

5.1 Layout

The dashboard architecture encompasses four distinct screens, each meticulously crafted to align with the specified factors. These screens serve as intuitive navigational hubs, catering to the diverse needs of users across various operational domains.

Each screen is thoughtfully tailored to encapsulate the essence of its associated factor, ensuring that users can seamlessly navigate through relevant insights and functionalities. From data analysis to strategic planning, each screen is purposefully designed to deliver a tailored experience that resonates with the unique requirements of its corresponding factor.

The UI/UX is carefully crafted to be inherently user-friendly, prioritizing simplicity, clarity, and efficiency in every interaction. Intuitive navigation pathways, contextual cues, and streamlined workflows empower users to effortlessly harness the full potential of the dashboard, irrespective of their level of expertise or familiarity with the platform.

5.2 Features & Functionalities

- **Interactive Elements:** All visuals are interactive visuals; audiences can interact within the visuals and cross visuals but can not make cross-pages interaction.
- **Filter & Search Option:** Date, Quadrants, Communities and Streets filter are visible in every page of the dashboard.
- **Navigation:** Page navigations are visible in every page of dashboard. There are some bookmark navigations and tooltip pages in the dashboard.

6. Software & Tools

- Jupiter notebook: to conduct data preprocessing.
- Azure SQL Server: to store data.
- PowerBI: to conduct data analyzing and visualization.
- QGIS: to prepare geo-spatial data.

7. Conclusion

This document provides a detailed roadmap for developing a Dashboard to analyze traffic data effectively. It outlines key components such as the user interface, backend processing, and database architecture, ensuring a structured approach to meet project objectives.

By defining clear requirements and design considerations, the document facilitates collaboration among stakeholders and guides the development team through implementation. The user interface offers an intuitive experience, while the backend processing handles data analysis and report generation efficiently.

With a focus on data integrity and scalability, the database architecture ensures reliable storage and management of traffic incident data. Overall, the design document sets the stage for delivering a valuable solution that enhances traffic management and safety in Calgary.