**Analysis of Road Traffic Safety and Congestion Using Big Data**

Team Uncharted

DATA 228 -12: Project Plan

Version 2.0

Revision History

|  |  |  |  |
| --- | --- | --- | --- |
| **Date** | **Version** | **Description** | **Author** |
| 12/1/2023 | 1.0 | Setup the project plan document | Shashikumar K Mallikarjuna  Mahamaya Panda |
| 12/9/2023 | 2.0 | Worked on the content for project plan document | Shashikumar K Mallikarjuna  Mahamaya Panda  ReddySaketh R. Chappidi  Aradhya A. Rathnakar  Bhavan Kumar Basavaraju |

Table of Contents

1.Introduction 4

1.1 Purpose of this document 4

1.2 Intended Audience 4

1.3 Scope 4

1.4 Definitions and acronyms 6

1.4.1 Definitions 6

1.4.2 Acronyms and abbreviations 6

2. Background and Objectives 7

3. Architecture and High Level Design 8

4.Organization 9

4.1Project group 9

4.2 Customers 10

5.Development process 10

6.Deliverables 11

7.Project risks 12

8.Communication 12

8.1 Collaboration 12

8.2 Git 12

9.Project plan 13

9.1 Time schedule 13

9.2 Connection Test Plans 14

10.References 16

# Introduction

## Purpose of this document

In the face of growing urbanization, traffic congestion and safety issues pose significant challenges for huge cities like Chicago. The "Analysis of Road Traffic Safety and Congestion Using Big Data" project addresses these concerns by utilizing big data to monitor and analyze traffic safety and congestion in the city of Chicago. The project relies on data sourced from the City of Chicago's data portal, encompassing various datasets related to traffic crashes, vehicles involved in the crash, and the people involved in the crash. Analysis of traffic crashes helps to understand the reason for traffic crashes and improve the infrastructure to mitigate accidents caused due to poor infrastructure. This document serves as a resource for the team members and external stakeholders interested in collaboration to understand the project scope, technical aspects, and outcomes.

## Intended Audience

By tailoring the project's outputs to meet the diverse needs of these audiences, the project aims to serve as a valuable resource that fosters collaboration, informed decision-making, and a safer, more efficient road network in Chicago. A few such stakeholders include:

* City Planners and Urban Developers: can analyze the traffic pattern through the crash rate and plan future projects accordingly.
* Transportation Authorities: can understand the vehicle crash pattern and schedule more frequent buses to reduce the number of people using personal modes of transportation.
* The General Public: can use the application to see the crash rate in a particular route and plan their trips accordingly to avoid risky routes.
* Emergency Services: can add more patrols and ambulances dedicated to the areas where accidents happen more frequently to avoid any delays in terms of saving the lives of people involved in vehicle crashes.

## Scope

The scope of the project encompasses the following key elements:

**Project Objectives:**

- Define the overarching goals of the "Chicago Road Traffic Crash Analysis" project, emphasizing its role in improving road traffic safety, mitigating congestion, reducing vehicle crashes, and aiding in urban planning for the city of Chicago.

- Clearly articulate the desired outcomes and impact the project aims to achieve in traffic management and public safety.

**Project Team Organization:**

- Identify and define the roles and responsibilities of the project team members involved in the development and execution of the application.

- Establish a well-defined organizational structure, specifying the contributions of data engineers, cloud architects, data analysts, front-end developers, and other relevant roles.

**Development Process:**

- Specify the development methodology and processes that will be employed throughout the project.

- Detail the sequential stages of development, beginning with data acquisition from the City of Chicago’s data portal, processing using AWS cloud resources, and culminating in implementing an analytical report for user interaction to understand the vehicle crash statistics.

**Risk Assessment:**

- Identify potential risks associated with the project, drawing on considerations such as data integrity, cloud service reliability, and accuracy in traffic predictions.

- Develop comprehensive risk mitigation strategies and contingency plans to address identified risks, ensuring project resilience and success.

**Communication Plan:**

- Define the communication channels and protocols that will be utilized among project stakeholders, including city planners, transportation authorities, data analysts, and the public.

- Establish a systematic approach to regular reporting, status updates, and feedback mechanisms to maintain transparent and effective communication throughout the project's lifecycle.

**Project Plan:**

- Present a detailed schedule outlining the project's milestones, deadlines, and key deliverables, spanning from data acquisition to the deployment of the front-end application.

- Provide an activity plan that breaks down the project into manageable tasks, assigning responsibilities to team members and specifying timelines for each phase of development.

## Definitions and acronyms

### **Definitions**

| **Keyword** | **Definitions** |
| --- | --- |
| Analysis of Road Traffic Safety and Congestion using Big Data | The name of the project |
| Chicago Road Traffic Traffic Crash Analysis | Alternative name to reference the project which is a lot more specific |
| Project Manager | A person in charge of managing the project |
| Team Member | An active member of the team responsible for working on tasks that contribute to the project development |
| Milestone | A time in a project that marks the end of a project phase or the completion of an important deliverable. |
| Git | Version control system that will be used in this project |
| Scrum | An iterative and incremental agile software development method for managing software projects and product or application development |
| Scrum sprint | The basic unit of development in Scrum |
| Scrum master | Ensures the smooth working of the Scrum team and enforces Scrum practices |
| Team Leader | A person responsible to plan the project scope and handle the deliverables completed in each milestone |

### **1.4.2 Acronyms and abbreviations**

| **Acronym or**  **abbreviation** | **Definitions** |
| --- | --- |
| AWS | Amazon Web Services |
| ETL | Extract, Transform, and Load |
| API | Application Programming Interface |
| CSV | Comma Separated Values |

# Background and Objectives

**Background**:

The escalating population density in urban areas has led to a noticeable increase in road traffic and vehicle crashes. In response to this challenge, this project utilizes big data analytics to delve into road traffic safety and crash incidents. By thoroughly analyzing vehicle crash data, the project seeks to formulate data-driven solutions aimed at improving road safety and reducing the occurrence of vehicle crashes in urban settings.

**Objectives**:

**Collecting and Integrating Vehicle Crash Data**:

Aggregate, consolidate, and merge vehicle crash data sourced from the city of Chicago data portal.

Justification: Establish a comprehensive dataset that provides insights into traffic incidents and patterns for the city of Chicago.

**Implementing Intelligent Traffic Management Strategies**:

Utilize crash data-driven insights to develop intelligent traffic management strategies.

Justification: Enhance the efficiency of traffic flow, reduce congestion, and improve overall road safety through data-driven interventions.

**Optimizing Urban Infrastructure Planning**:

Leverage collected data to optimize urban infrastructure planning to meet growing traffic demands.

Justification: Support informed decision-making for infrastructure development, road expansions, and other urban planning initiatives.

**Enhancing Commuting Experience and Reducing Pollution**:

Improve the commuting experience for residents by minimizing crash-related delays.

Justification: Contribute to a safer and more efficient commuting environment while reducing pollution associated with vehicle delays and accidents.

# Architecture and High-Level Design

AWS will be used for the project implementation using various components like S3, Glue, Redshift, Glue crawlers, Athena, and Quicksight. The various components of the project architecture are mentioned below:

**Source Data**: Vehicle crash data obtained from government transportation-related websites in CSV format or through API calls.

**Data Storage**: Upload to Amazon S3 for secure and scalable storage. Amazon S3 provides scalable and cost-effective storage for the large volume of crash data.

**Data Processing**: AWS Glue will be used for ETL processes to prepare the data for analysis. AWS Glue Crawlers are used to generate the schema to read data from the underlying files in the S3 bucket. AWS Glue automates the ETL process, ensuring data integrity and consistency.

**Data Analysis**: AWS Athena and Redshift are utilized for querying and analyzing the processed data. AWS Athena and Redshift offer efficient querying and analytical capabilities for data-driven insights.

**Data Visualization**: Amazon QuickSight will be employed for creating visually appealing and interactive reports. Amazon QuickSight facilitates the creation of intuitive and actionable visualizations for stakeholders.

The figure below shows the various tools used in different phases of the project.

A screenshot of a computer

Description automatically generated

Fig 1. The above figure shows the various tools used in each phase of the project implementation.

The figure below shows the system architecture and data flow of the project.

A diagram of data flow

Description automatically generated

Fig 2. Project workflow design

# Organization

The project is split into five milestones and the responsibility of completing the different milestones is assigned to different members of the team.

## Project group

|  |  |  |
| --- | --- | --- |
| **Name** | **Initials** | **Responsibility (roles)** |
| ReddySaketh R. Chappidi | RC | Data Collection |
| Mahamaya Panda | MP | Exploratory Data Analysis |
| Aradhya A. Rathnakar | AR | Data Cleansing |
| Shashikumar K Mallikarjuna | SM | Data Transformation |
| Bhavan Kumar Basavaraju | BB | Data Visualization |

## 4.2 Customers

The target customers are listed below:

* City Planners and Urban Developers
* Transportation Authorities
* General Public
* Emergency Services

# Development process

Analysis of the road traffic safety and congestion project will be implemented using the waterfall development method since the project scope involves no changes for the current version of the implementation. The project is broken down into multiple phases for easier effort estimation and project delivery in phases. The table below shows the different phases of the project along with the tools used in each phase.

| **Phase** | **Description** | **Tools Used** |
| --- | --- | --- |
| Data Source Discovery | Explore various websites to get the traffic crash-related data for analysis. | Google search engine |
| Data Collection | Explore the Data API available in the source website and load it into a storage location for project usage. | AWS Glue |
| Dimensional Modeling | Understand the source data and the analytical requirement and design the data warehouse. | draw.io |
| Data Preprocessing / ETL | Cleanse the source data, transform it based on the business need, and load it into the data warehouse. | AWS Glue, AWS Redshift, and AWS Athena |
| Scalability and Performance Optimization | Add indexes and create table partitions to improve the performance of the analytical queries. | AWS Redshift |
| Data Verification/Testing | Verify that all the data in the dimension and fact tables are accurate to solve the business problem. | AWS Redshift |
| Data Analysis/ Visualization | Explore the trends and patterns in the data loaded into the data warehouse on the traffic crashes and make decisions based on the analysis. | AWS Quicksight |

# Deliverables

| **Sl.No.** | **Output** | **Planned Date** | **Promised Date** | **Late +/-** | **Delivered Date** | **Notes** |
| --- | --- | --- | --- | --- | --- | --- |
| 1 | Project Proposal | 8/28/2023 | 9/3/2023 | 0 | 9/3/2023 | On time |
| 2 | Project Design | 9/10/2023 | 9/15/2023 | -1 | 9/14/2023 | Delivered early |
| 3 | Project Implementation Code | 9/14/2023 | 11/2/2023 | -1 | 11/1/2023 | Delivered early |
| 4 | Data Verification | 10/16/2023 | 11/15/2023 | +2 | 11/17/2023 | People dataset didn’t meet the requirement and took long to fix |
| 5 | Visualizations | 10/20/2023 | 11/28/2023 | +1 | 11/29/2023 | Took longer due to change in report structure based on UAT |
| 6 | Project Documentation | 10/28/2023 | 12/10/2023 | 0 | 12/10/2023 | On time |

# Project risks

|  |  |  |
| --- | --- | --- |
| **Possibility** | **Risk** | **Preventive action** |
| Crashes data not uploaded in regular intervals | Up-to-date data availability | Find alternative websites and get licenses to get the most recent data regularly. |
| High AWS service cost | Cost | Set up budget alerts to inform the team when the AWS monthly usage cost surpasses the planned budget. Delete resources if not used. |

# Communication

The team will communicate with each other in terms of project planning, and task assignment over Zoom meetings which are scheduled for every two weeks. For day-to-day communication, WhatsApp will be used in case of task-related issues or questions. Canvas will be used to submit the project deliverables. The above-mentioned communication tools will be used for communication for four months.

## Collaboration

Team’s work collaboration is an important part of project development. Git is used as the version control tool to track the changes. After a particular feature is developed, the team member will raise a pull request for another team member to review and approve before pushing the code to production. This will help mitigate any possibility of causing issues in the production environment. Google Docs is used to collaborate on documentation of the project. When developing Proof of Concept to get the data from data sources, Google Colab is used so that the multiple team members can collaborate on getting data from multiple sources.

## Git

All source code and finished documentation will be uploaded to the GitHub repository.

RepositoryURL: <https://github.com/shashikumar1998/ChicagoVehicleCrashBigDataAnalaysis>

# Project plan

Project planning is an important part of project management. The project needs to be planned in a way such that it can be production-ready on time based on the promised delivery date. This project is planned to be completed in four months and have the system in production before 12/1/2023.

## Time schedule

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Id** | **Milestone** | **Responsible Dept./Initials** | **Delivery Date** | | | **Description** |
|  |  |  | **Forecast** | **+/-** | **Actual** |  |
| 1 | Project Proposal | RC | 9/3/2023 | 0 | 9/3/2023 | Decide and finalize a project topic |
| 2 | Project Design | RC | 9/15/2023 | -1 | 9/14/2023 | Design the big data analytics system to be scalable and plan the architecture. Perform dimensional modeling |
| 3 | Data Collection | MP | 9/25/2023 | -2 | 9/23/2023 | Source the data from the websites using the Data API option to automate the source data collection |
| 4 | Data Cleansing | AR | 10/18/2023 | +3 | 10/21/2023 | Cleanse the individual raw datasets to be merged for analysis |
| 5 | Data Transformation | SM | 10/29/2023 | +1 | 10/30/2023 | Transform the cleansed datasets based on the dimensional modeling |
| 6 | Load Data into Data Warehouse | SM | 11/3/2023 | 0 | 11/3/2023 | Load the transformed data into the datawarehouse for analysis |
| 7 | Data Verification | SM | 11/15/2023 | +2 | 11/17/2023 | Verify that the tables in the data warehouse meet the business requirement and test the accuracy of the data |
| 8 | Crash Analysis Report | BB | 11/28/2023 | +1 | 11/29/2023 | Create a report to analyze the crash dataset to understand the problem and formulate solutions |
| 9 | Project Documentation | MP | 12/10/2023 | 0 | 12/10/2023 | Document the project |

## Connection Test Plans

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Test No.** | 001 | **Phase:** | 1 | **Author Initial:** | SM | Date: 9/17/2023 |
| **Test Category:** | | **Data Collection from Glue** | | |  |  |
| **Software Product:** | | AWS Glue | | | |  |
| **Test Title:** | | Pull data from Chicago data portal using Data API from AWS Glue | | | | |
| **Test Purpose:** | | Test if AWS Glue can pull the crashes data from the source using API | | | | |
| **Test Setup:** | | Setup Glue notebook and import the dependent API libraries | | | | |
| **Prerequisites:** | | AWS account setup, create S3 bucket, create IAM for Glue to access S3 bucket | | | | |
| **Procedure:** | | Write Python code in AWS Glue notebook to use the Data API provided in the city of Chicago data portal to get the most recent data and load it into S3 bucket under a folder called raw | | | | |
| **Checks:** | | Check to see if there is a file created in the S3 bucket under raw folder | | | | |
| **Expected Results:** | | Data is stored in the S3 bucket under the raw folder in csv format and is not an empty file | | | | |
| **Result:** | | The python code written in Glue notebook successfully loaded the source data into the S3 bucket under the raw folder which is not empty | | | | |
| **Reason for Failure:** | | - | | | | |
| **Remarks:** | | Test successful | | | | |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Test No.** | 002 | **Phase:** | 1 | **Author:** | SM | Date:10/2/2023 |
| **Test Category:** | | **Load data to AWS Redshift using Glue** | | |  |  |
| **Software Product:** | | AWS Glue and Redshift | | | |  |
| **Test Title:** | | Load data into Redshift tables from S3 transformed folder | | | | |
| **Test Purpose:** | | Test connection from Glue to redshift to load the data into redshift tables | | | | |
| **Test Setup:** | | Setup Glue notebook and import the dependent S3 and redshift libraries | | | | |
| **Prerequisites:** | | Data in transformed folder of S3 bucket, Glue notebook, tables in redshift database, setup NAT gateway | | | | |
| **Procedure:** | | Write code in Glue notebook to load the CSV data from the transformed folder in S3 bucket. | | | | |
| **Checks:** | | Check to see if the data is populated in the redshift tables | | | | |
| **Expected Results:** | | The same number of rows in the S3 transformed folder are loaded into the redshift tables | | | | |
| **Result:** | | The same number of rows in the S3 transformed folder are loaded into the redshift tables | | | | |
| **Reason for Failure:** | | - | | | | |
| **Remarks:** | | Test Successful | | | | |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Test No.** | 003 | **Phase:** | 1 | **Author:** | SM | Date:10/20/2023 |
| **Test Category:** | | **Read data from redshift in Quicksight** | | |  |  |
| **Software Product:** | | AWS Redshift and Quicksight | | | |  |
| **Test Title:** | | Access data from redshift tables through Quicksight for reporting | | | | |
| **Test Purpose:** | | Test connection between Quicksight and Redshift | | | | |
| **Test Setup:** | | Setup a sample Quicksight report to test the connection | | | | |
| **Prerequisites:** | | data populated in redshift tables | | | | |
| **Procedure:** | | Create a Quicksight report and use redshift connector to connect to the redshift database which has the tables and create sample visuals to showcase the underlying data | | | | |
| **Checks:** | | Check to see if the data connection is successful | | | | |
| **Expected Results:** | | The data connection is successful and able to visualize the sample data in Quicksight based on the data in redshift tables | | | | |
| **Result:** | | The data connection is successful and able to visualize the sample data in Quicksight based on the data in redshift tables | | | | |
| **Reason for Failure:** | | - | | | | |
| **Remarks:** | | Successful | | | | |

# References

[1] W. Alajali, W. Zhou and S. Wen, "Traffic Flow Prediction for Road Intersection Safety," 2018 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI), Guangzhou, China, 2018, pp. 812-820, doi: 10.1109/SmartWorld.2018.00151.

[2] L. Zhu, F. R. Yu, Y. Wang, B. Ning, and T. Tang, “Big Data Analytics in Intelligent Transportation Systems: A survey,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 1, pp. 383–398, 2019. doi:10.1109/tits.2018.2815678

[3] H. Al Najada and I. Mahgoub, "Big vehicular traffic Data mining: Towards accident and congestion prevention," 2016 International Wireless Communications and Mobile Computing Conference (IWCMC), Paphos, Cyprus, 2016, pp. 256-261, doi: 10.1109/IWCMC.2016.7577067.

[4] H. A. Najada and I. Mahgoub, "Autonomous vehicles safe-optimal trajectory selection based on big data analysis and predefined user preferences," 2016 IEEE 7th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), New York, NY, USA, 2016, pp. 1-6, doi: 10.1109/UEMCON.2016.7777922.

[5] A. Saxena and S. A. Robila, "Analysis of the New York City’s Vehicle Crash Open Data," 2021 IEEE International Conference on Big Data (Big Data), Orlando, FL, USA, 2021, pp. 6017-6019, doi: 10.1109/BigData52589.2021.9672012.

[6] H. A. Najada and I. Mahgoub, "Anticipation and alert system of congestion and accidents in VANET using Big Data analysis for Intelligent Transportation Systems," 2016 IEEE Symposium Series on Computational Intelligence (SSCI), Athens, Greece, 2016, pp. 1-8, doi: 10.1109/SSCI.2016.7850097.