Exploratory analysis of parking violation Project Plan

Version 1.0

Revision History

Date	Version	Description	Author

Table of Contents

1. Introduction

1.1 Purpose of this document

The purpose of this document is to provide a detailed project description of the application. Big data is the result of technological advancements that have resulted in the advent of massive amounts of data. Big data refers to datasets that are not only large in size but also contain a lot of different types of data. When properly analyzed, these data can assist industries in making important decisions in a variety of ways. Parking violations are a daily problem in today's fast-paced environment. Parking a vehicle illegally may result in an offense, resulting in a large number of traffic citations being issued. 'Parking Violation' data is one such data set for our exploratory investigation. Every day, millions of automobiles are parked in cities, and New York, as a major metropolis, is no exception, with most residents having parking issues. New York City itself collected approx $957 million in fine revenues In them more than 59% that is approx $565 million of the $957 million, come from parking tickets. The analytics and visualizations are performed using various AWS Services. We analyzed the dataset which consists of more than 50 million records from the years 2013-2018. We performed analytics using AWS services like S3 for data storage and Redshift for performing queries. The visualizations were performed using QuickSight.

1.2 Intended Audience

- Team members : Sonali,Sanjana,Shrivatson,Rishi

1.3 Scope

Our project aims to analyze the parking dataset for two different data sources , one with a span of 20 million data and second data source of 30 million parking records with various graphical representations to give an interpretation of these parking datasets.

1.4 Definitions and acronyms

1.4.1 Definitions

Keyword	Definitions
Project Name	Exploratory Analysis of Parking Violation
Project Supervisor	Sanjana Balagar
Project Leader	Sonali Gupta
Team Member	Sonali,Sanjana,Shrivatson,Rishi
Milestone	Aug 2021 - Dec 2021
Git	https://github.com/sanjanabalagar/Data228_Project
Scrum	An incremental and iterative agile software development method

	for managing software projects and product or application development
Kunagi	Web-based tool for integrated agile project management and collaboration based on Scrum
Scrum sprint	Weekly
Scrum master	Shrivatson
Product owner	Rishi

1.4.2 Acronyms and abbreviations

Acronym or abbreviation	Definitions

1.5 References for data source

2. https://data.cityofnewyork.us/City-Government/Parking-Violations-Issued-Fiscal-Year-2022/p_vqr-7yc4
3. https://www.kaggle.com/new-york-city/nyc-parking-tickets

4. Background and Objectives

Abstract:

In this project we have used the parking violation dataset from the open nyc and kaggle which contained more than 50 million records and performed various analytics to generate meaningful insights. The NYC Department of Finance collects data on every parking ticket issued in NYC and is responsible for collecting and processing payments of all tickets. Because of the huge number of cars and the limited geography, there are a lot of parking tickets. This prompted us to conduct an exploratory analysis on such data in order to gain insights such as when and where tickets are more likely to be issued, if there is a specific season for it, what types of vehicles are receiving tickets, comparing state data to determine which state has the most tickets issued on a monthly basis, and which vehicle body type receives the most tickets. This analysis is carried out utilizing different AWS services like the S3 for data storage, Redshift for performing queries and analytics , and QuickSight to perform dynamic visualizations with the goal of developing a graphical solution for real-time analytics using the parking dataset.

Objective

Our project aims to analyse the movies data for two different data sources , one with datalength of 20 million parking violation data and second data source of 20 million with various graphical representations to give an interpretation of these movie datasets.

5. Architecture & High Level Design

1. Using the Talend ETL tool we have cleaned the datasets. We were having files of four fiscal years from 2013 to 2017 , which were cleaned and transformed using ETL and then we combined all four transformed files and generated one CSV file containing around 50 million records and loaded it to S3 bucket to create tables in Redshift Clusters.
2. Created cluster on redshift service by providing access to users through an associated IAM role to perform analysis using its query editor.
3. We have created a ticket violation table including all required attributes with metadata in a database. Then, loaded data into it from a stored csv file in s3 bucket to perform queries and analysis.
4. For visualization we have connected S3 with QuickSight by creating and uploading manifest JSON to showcase dynamic visualization , we also performed visualization using matplotlib library in python.

6. Organization

6.1 Project group

Name	Initials	Responsibility (roles)
Shrivatson Ramaratnam Giridharan	SRG	Analysis and Development
Sonali Gupta	SG	Data sourcing/modeling and Development
Rishi Bamb	RB	Analysis and visualization
Sanjana Balagar	SB	Development, documentation

6.2 Customer

The target customers are listed below:

- Customers
- Financial Department
- Ticket Issuing Agencies

7. Development process

1. As we were having four fiscal years data from 2013 to 2017, we cleaned and normalized individual years files using Talend which provides feasible services to process and prepare data.
2. To have data in a single file makes an analysis and working on it easy hence we merged all four files into one using python code.
3. Created cluster on redshift service by providing access to users through an associated IAM role to perform analysis using its query editor.
4. Then we have created a ticket violation table including all required attributes with metadata in a database. Then, loaded data into it from a stored csv file in s3 bucket to perform queries and analysis.
5. Then we are showcasing visualizations on AWS Quicksight and python. Quicksight's analytic platform empowers any skill level target audience to work with data through actionable and insightful visualizations. Below are the screenshots of visual analysis done on Quicksight.We have done visualisations using quicksight where we loaded a JSON manifest file via S3 bucket into quicksight for performing visualisations.

8. Deliverables

To	Output	Planned week	Promised week	Late +/-	Delivered week	Notes
Data Extraction	**Data was downloaded from Kaggle and NYC Open Data and made it ready for loading and cleaning**	**Sept 3rd week**	**Sept 3rd week**	**no**	**Sept 3rd week**	
Data Normalization and cleaning	**Data was cleaned and normalized using Talend**	**Oct 1 week**	**Oct 1 week**	**no**	**Oct 1 week**	
Data Loading to S3	**Data was uploaded to s3 bucket**	**Oct 2 week**	**Oct 2 week**	**no**	**Oct 2 week**	
Creating Clusters in Redshift	**We created database, tables and uploaded the data from s3 bucket**	**Oct 3 week**	**Oct 3 week**	**no**	**Oct 3 week**	
Loading Data in Quicksight	**Loaded the data in QuickSight using JSON Manifest File**	**Oct 4 Week**	**Oct 4 Week**	**no**	**Oct 4 Week**	
Visualization and analytics	**Data was analysed using python and Quicksight**	**Nov 4 week**	**Nov 4 week**	**no**	**Nov 4 week**	

9. Project risks

Possibility	Risk	Preventive action
The use of S3 and Redshift	If accessed	When we were not using it we had to

could have costed us a lot of money	continuously it would charge us more.	pause the service.Once the project was done we deleted the service.
Initializing the cluster while giving public access	If it is given public access then there is a chance of vulnerability as other people can get access to our data.	Hence we have made it private

10. Communication

All the team members were connected through Zoom calls weekly and also met in person.

10.1 Collaboration

10.2 Git

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

Repository URL: https://github.com/sanjanabalagar/DATA228_Project

11. Project plan

11.1 Time schedule

Id	Milestone Description	Responsible Dept./Initials	Finished week				Metr.	Rem.	
			Plan	Forecast		Actual			
				Week	+/-				
1	Data Extraction	Rishi	Sept 3rd week	Sept 3rd week	No	Sep 3rd week			
2	Data Normalization and cleaning	Shrivatson	Oct 1 week	Oct 1 week	No	Oct 1 week			
3	Data Loading to S3	Sonali	Oct 2 week	Oct 2 week	No	Oct 2 week			
4	Creating Clusters in Redshift	Sanjana	Oct 3 week	Oct 3 week	No	Oct 3 week			
5	Loading Data in Quicksight	,Sonali	Oct 4 Week	Oct 4 Week	no	Oct 4 Week			
6	Visualization and analytics	Shrivatson	Nov 4 week	Nov 4 week	NO	Novl 4 week			

11.1.1 Remarks

Remark Id	Description

11.2 Test plan

Test No.	001	**Phase:**	1	**Author:**	Rishi	Date: Oct 2021
Test Category:		**Extract data and load into S3**				
Software Product:		Talend,AWS S3				
Test Title:		Similarity in the data count of source and destination				
Test Purpose:		The number of records should match at both ends				
Test Setup:		We queried in Redshift to find the number of records in destination and compared with the number of records in source which was extracted through talend.				
Prerequisites:		Queried in Redshift and source data in s3 bucket				
Procedure:		We queried in Redshift to find the number of records in destination and compared with the number of records in source which was extracted through talend.				
Checks:		The count of data				
Expected Results:		The count of data at the source and destination should be similar				
Result:		The count of data and the destination matched				
Reason for Failure:		No failure				
Remarks:						

Test No.	002	**Phase:**	1	**Author:**	Sonali	Date: Oct 2021
Test Category:		accurate data is being display on graph				
Software Product:		AWS Quicksight				
Test Title:		To check whether the data present on the graph is correct				
Test Purpose:		The data present in the graph must correlate with the actual dataset				

Test Setup:	We rechecked the data in the graph to validate whether the data present in the graph is matching with the actual data
Prerequisites:	data must be loaded in the S3 Bucket and connected to quicksight using JSON Manifest file
Procedure:	We checked the graph data to see the accuracy of data present in the graph along with the dataset present.
Checks:	Checked the validity of data present in the graph
Expected Results:	The graph visualisation must match with the loaded data
Result:	The graph visualisation must match with the loaded data
Reason for Failure:	There was no failure
Remarks:	

11.2.1 Testing Remarks

Remark Id	Description

12. References

- https://www.kaggle.com/new-york-city/nyc-parking-tickets
- https://data.cityofnewyork.us/City-Government/Parking-Violations-Issued-Fiscal-Year-2022/pvqr-7yc4
- https://www.talend.com
- Images - Google
- GitHub -https://github.com/sanjanabalagar/DATA228_Project/tree/main