K Mohan (2324600)

Group 2

AWS Data Analytic Platform for The City of Vancouver

Phase 2

Dataset – Public Trees

Assignment 2

**Data Analysis**

**Business Question 1**

A screenshot of a computer

AI-generated content may be incorrect.

The first Business Question 1 demonstrate that the dataset (Public Trees) was analyzed, refined, transformed, and structured to uncover the important information about the dataset with a purpose of supporting the decision making of the Vancouver Department. In First Business Question I used SQL Query by using Amazon Athena service. As I have multiples columns and rows, in Business Question 1 I figured out the minimum diameter of the various trees which was 10 Inches.

**Business Question 2**

A screenshot of a computer

AI-generated content may be incorrect.

In Business Question 2, I used maximum function to figure it out what the maximum length of trees diameter of the trees in Vancouver is. I found out that the maximum length of the diameter of various trees in Vancouver is 9 Inches.

**Business Question 3**A screenshot of a computer

AI-generated content may be incorrect.

In Business Question 3, I used average function to figure it out what the average length of the diameter of various trees in Vancouver is. I found out that the average length of the diameter of various trees in Vancouver is 5.0476 Inches.

Overall, the above screenshot helps us the analyze the public trees dataset. Data analysis methods can vary significantly and may include descriptive statistics, exploratory data analysis, inferential statistics, data visualization, and predictive modeling. However, I use Minimum, Maximum, and Average function to figure out the insights from the dataset.

In this step, I explored a dataset on Public Trees in Vancouver, which I obtained from the Open Data Vancouver website. I focused on answering three key business questions related to the dataset.

**Data Security**

A screenshot of a computer

AI-generated content may be incorrect.

The above screenshot demonstrates the keys, which I created to allowing the access to the 3 Different S3 Buckets for security of the data with the help of AWS Key Management.

A screenshot of a computer

AI-generated content may be incorrect.

Created a Buck-up Buckets for safeguarding any loss of the Data of original Buckets.

S3 – Raw Bucket – Default Encryption – Pairing Security Key To Controlling The Access.A screenshot of a computer

AI-generated content may be incorrect.

S3 – Raw Bucket – Bucket Versioning – Enabling To Store Data in Backup Bucket.

A screenshot of a computer

AI-generated content may be incorrect.

S3 – Raw Bucket – Management – Creating to Replication To Prevent Data Loss.

A screenshot of a computer

AI-generated content may be incorrect.

S3 – Transform Bucket – Default Encryption - Pairing Security Key To Controlling The Access.

A screenshot of a computer

AI-generated content may be incorrect.

S3 – Transform Bucket – Bucket Versioning - Enabling To Store Data in Backup Bucket.

A screenshot of a computer

AI-generated content may be incorrect.

S3 – Transform Bucket – Management - Creating to Replication To Prevent Data Loss.

A screenshot of a computer

AI-generated content may be incorrect.

S3 – Curated Bucket – Default Encryption - Pairing Security Key To Controlling The Access.

A screenshot of a computer

AI-generated content may be incorrect.

S3 – Curated Bucket – Bucket Versioning - Enabling To Store Data in Backup Bucket.

A screenshot of a computer

AI-generated content may be incorrect.

S3 – Curated Bucket – Management - Creating to Replication To Prevent Data Loss.

A screenshot of a computer

AI-generated content may be incorrect.

The Data Security in AWS is provide protection to store sensitive data from unwanted access, breaches, and data theft while ensuring compliance with industry regulations. By implementing data security service, it measures it ensures the confidentiality, integrity, accessibility, and availability of information. This typically involves deploying technologies, policies, and strategies to protect sensitive data.

In this step, I created the 3 AWS Keys for the Public Trees Dataset and implement security. I also created back-up S3 Buckets for the purpose of protecting the data and ensure data recovery in case of data loss. I created 3 encrypted keys to secure the data. By enabling versioning and edited the default encryption option, this way I enhanced data protection, by allowing new versions to be replicated into the backup backets. Finally, I paired each encryption key with its respective bucket and configured a replication rule of each 3 different buckets to automate the backup process.

**Data Governance**

Visual ETL Pipeline for filtering the quality of the dataset.

A screenshot of a computer

AI-generated content may be incorrect.

I have created the ETL Pipeline, to efficiently move and process data from various sources to a target destination, I created data quality to process the dataset, then I tested the dataset whether rows are failed or passed. The next step helped me to differentiate the dataset. One is for Passed, I paste all the dataset which are passed the threshold, and the failed one I paste in failed S3 Folder.

S3 Folders – Passed and Failed Folders

A screenshot of a computer

AI-generated content may be incorrect.

S3 Failed FileA screenshot of a computer

AI-generated content may be incorrect.

S3 Passed FilesA screenshot of a computer

AI-generated content may be incorrect.

Data Governance in AWS is designed to manage, protect, and control the access of the person who can access the dataset throughout the lifecycle of the dataset. It ensures that the data is secure and only the authorized person are accessing the dataset.

I have created the ETL Pipeline to process the dataset and figure out which of the rows from the dataset is failed the threshold rule and which of the dataset passed the threshold rule. Then I paste the passed data in Passed folder in S3 Bucket and the failed data in Failed folder of the S3 folder. Utilizing AWS Glue services, I successfully categorized the dataset into Passed and Failed, ensuring structured data governance and quality control.

**Data Monitoring**

Dashboard for the monitoring the usage of the services of the AWS.

A screenshot of a computer

AI-generated content may be incorrect.

Cloud Trail – Created the Dataset Cloud Trail to monitor the usage.

A screenshot of a computer

AI-generated content may be incorrect.

Cloud Trail Event History

A screenshot of a computer

AI-generated content may be incorrect.

Data Monitoring in AWS involves continuously tracking, analyzing, and managing data to ensure its accuracy, security, compliance, and performance. It helps organizations to maintain data integrity, detect anomalies, and optimize system performance.

In the above steps, I have created dashboard to visualize the service which I have used over the past 3 months for the purpose of created the project 1 and project 2. This dashboard provides information regarding how many times I have used the services of the AWS to create my projects. Moreover, this dashboard helped me to figure it out the usage of services into the utilization of raw data by AWS Glue and Amazon S3 services. Additionally, it displays an alarm system that tracks data and cost usage limits, enabling effective data and cost monitoring and management.

As the next step in my data monitoring process. I have created a folder in S3 which stores the data of the everything usage which I have using the services of the AWS, it shows the event history of the usage which includes an even history log, which records the various services utilized during this period to process the dataset and complete the tasks assigned to me by the City of Vancouver.