K Mohan (2324600)

Group 2

AWS Data Analytic Platform for The City of Vancouver

Phase 1

Dataset – Public Trees

**Role and Importance of Developing a Data Analytic Platform for the City of Vancouver**

Creating Data Analytics Platform (DAP) for the City of Vancouver is an important part for fostering data-driven decision-making, improving public services, and utilising the resource properly. By establishing an information pipeline in AWS, the Vancouver city can utilize the power of cloud computing to efficiently analyze large datasets. This platform will support descriptive analytics, providing valuable information across different areas such as traffic management, environmental monitoring, public safety, and citizen engagement. Furthermore, AWS ensures scalability, security, and cost-effectiveness, adaptable to evolve the platform to remain sustainable for future needs. The dataset used will specifically focus on public trees, enabling more targeted analysis in areas like urban planning and environmental sustainability.

**Screenshot of Draw.IO**

**A screenshot of a computer

AI-generated content may be incorrect.A screenshot of a computer

AI-generated content may be incorrect.**

**A screenshot of a computer

AI-generated content may be incorrect.**

The diagram of **Draw.IO** demonstrate that the data analytics platform for the City of Vancouver, built on AWS, which acquire the data from multiple sources, including public trees – species name reports and user logs. The raw data is stored in S3 buckets, with the **public-trees-project1-kmohan**, name bucket receiving unprocessed data.

AWS Glue plays a key role in creating a data catalog and performing ETL Pipeline (Extract, Transform, Load) tasks, particularly for summarizing **public trees** data. EC2 instances within a VPC in the Canada (Central) Region provide the necessary compute power for data processing and analysis. The City of Vancouver Operations Team, along with other users, interacts with the platform via the internet.

Overall, the diagram shows the flow of data from ingestion and storage to processing and analysis, with a focus on key components and data systems of Vancouver's analytics platform on AWS. This dataset will specifically focus on public trees in Vancouver, offering insights for urban planning, environmental sustainability, and other related areas.

Screenshot of Public Trees S3 Buckets (Raw/ Transform/ curated)

**Step 1 - Data Ingestion**

A screenshot of a computer

AI-generated content may be incorrect.

Screenshot of Raw Bucket Folders and Raw DataSet – Public Trees

**Raw Data Ingestion**

A screenshot of a computer

AI-generated content may be incorrect.

Step 1: In this part I have store the dataset CSV file within an S3 bucket. The "public-trees-project1-kmohan" bucket functions as a storage location for raw data related to the City of Vancouver, specifically including public trees of Vancouver. The file’s CSV format indicates that the data is organized in a structured, tabular form, which is ideal for analysis.

Key Points:

1. Raw Data Storage: The S3 bucket’s name and the presence of a CSV file suggest that this is where the City of Vancouver stores its raw, unprocessed data.
2. Data Organization: S3 offers a scalable and reliable solution for storing such data.
3. Data Accessibility: The AWS Management Console provides an easy interface for users to access and manage the stored data.

Screenshot of Transform Bucket Folders – System and User

**Step 2 – Data Profilling**

A screenshot of a computer

AI-generated content may be incorrect.

This screenshot demonstrate the folders of System and User in pulic-trees-project1-trf-kmohan S3 Bucket.

Screenshot of Transform Bucket Folders – System

**Data Profiling**

A screenshot of a computer

AI-generated content may be incorrect.

Screenshot of Transform Bucket Folders – User

**Data Profilling**

A screenshot of a computer

AI-generated content may be incorrect.

Screenshot of DataBrew Dataset

**Data Profilling**

**A screenshot of a computer

AI-generated content may be incorrect.**

**A screenshot of a computer

AI-generated content may be incorrect.**

Step 2: This section focuses on profiling the dataset "**publictrees-ds-kmohan.csv**" using AWS Glue service which is DataBrew. The dataset was imported into DataBrew for analysis, revealing 470 rows and a single column with a BIGINT data type. Interestingly, 100% of the cells contained valid data, meaning there are no missing values. However, the presence of only one column suggests a potential formatting issue within the CSV file. I have separated the columns by using Separate Column Function which makes 15 columns within the dataset. I have renamed the column which we can see the history on receipt section. This section shows how many times I have changed my dataset.

This data profiling step is crucial for gaining insights into the structure and characteristics of the dataset. The findings will help identify any data quality issues, guide the necessary data transformation steps, and determine the most suitable analytical methods. Fixing the formatting issues is essential for ensuring that the analysis of the dataset—particularly the data on Vancouver’s public trees—can provide accurate and meaningful insights.

Top of Form

Bottom of Form

Screenshot of DataBrew Project

**Step 3 - Data Cleaning**

**A screenshot of a computer

AI-generated content may be incorrect.**

My Dataset – Public Trees, has 470 rows which consist in one single such column. My dataset has right information however I have to separate my dataset into column so I use delimitation fuction as each of the column are seperated by comma, so once I have seperated the column it showed my total 15 columns.

A screenshot of a computer

AI-generated content may be incorrect.

Screenshot of DataBrew Jobs

**Data Cleaning**

A screenshot of a computer

AI-generated content may be incorrect.

The data cleaning process applied within AWS Glue DataBrew to the **City of Vancouver's public trees dataset** involved several important modifications to the "**publictrees-project1-kmohan.csv"** file.

Initially, column names were updated to be more descriptive, enhancing the dataset's clarity. Duplicate rows were removed ensuring the uniqueness of the data. All the columns were renamed and separated. Moreover, all the columns were standardized by converting all characters to uppercase, improving consistency across the dataset. Lastly, missing values in the "Geom" and “Geo\_point\_2d” columns were replaced with "null" to explicitly indicate the absence of data. So I have deleted these columns so that my dataset will be cleaned and enhance.

These data cleaning actions significantly improve the dataset’s quality and consistency. The data is now better structured for further analysis, reporting, and integration with other systems.

Top of Form

Bottom of Form

Screenshot of AWS DataBrew

**Step 4 - Data Catalog Folders and Curated Bucket**

**A screenshot of a computer

AI-generated content may be incorrect.**

This section demonstrates how the City of Vancouver utilizes the AWS Glue Data Catalog to organize and manage **public trees dataset**. A dedicated database named “pub-trees-trfststem” has been set up within the catalog to serve as a centralized repository.

Within this database, two important tables are stored: "**public-trees-matrics**" and "**public-trees-metrics2**" Both tables are stored in Parquet format, which suggests a focus on efficient storage and fast querying. The "**public-trees-matrics**" table likely contains transformed data, while the " **public-trees-metrics2**" table is expected to hold key metrics derived from the water quality data.

By utilizing the AWS Glue Data Catalog, the City of Vancouver benefits from several advantages. Data discovery and access are simplified, enabling users to easily locate and understand the available water quality data. Moreover, the catalog integrates seamlessly with other AWS services, such as Athena and EMR, which support data analysis and reporting.

Top of Form

Bottom of Form

Screenshot of ELT Pipeline 1

**Step 5 - Data Summarization**

**A screenshot of a computer

AI-generated content may be incorrect.**

Screenshot of ELT Pipeline 2

**Data Summarization**

**A screenshot of a computer

AI-generated content may be incorrect.**

This step has demonstrate that to further summarize the dataset I have did the ELT Pipeline and Data Summarization technique to refined the dataset. In this part, I have used 2 Pipeline to refined the dataset. This 1st screenshot illustrate the Public tree height range and the other is based on the public tree diameter based.

Screenshot of S3 bucket – Curated Bucket

System Folder 1

**Data Summarization**

A screenshot of a computer

AI-generated content may be incorrect.

Both jobs follow a similar sequence of steps: schema transformation, filtering, aggregation, and dynamic transformations such as adding timestamps and adjusting data to the local time zone. After processing, the data is stored in S3 buckets for subsequent analysis and reporting.

Screenshot of S3 bucket – Curated Bucket

System Folder 2

**Data Summarization**

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

Screenshot of S3 bucket – Curated Bucket

User Folder 1

**Data Summarization**

A screenshot of a computer

AI-generated content may be incorrect.

This screenshot demonstrates the multiple dataset of the public dataset after summarizing the dataset based on the trees height range, species name according to the Max calculation function.

Screenshot of S3 bucket – Curated Bucket

User Folder 2

**Data Summarization**

A screenshot of a computer

AI-generated content may be incorrect.

This screenshot demonstrates the summarized dataset of the public trees based on the diameter of the public trees in Vancouver city. This dataset is in standard format.

This automated workflow enables the City of Vancouver to efficiently derive insights from their water data, facilitating informed decision-making and better resource management. The consistent process and seamless integration with other AWS services demonstrate a well-structured and scalable data analytics pipeline.