**AWS Data Analytic Platform for The City of Vancouver**

**“Phase 1”**

**Group 7**

Akash Haryani - 2237217

Emtiaz Ahmed Emon - 2221313

Martin Brunner - 2307307

Perumalla Vijaya - 2233471

BUSI 653: Section 05: Cloud Computing Technologies

Prof. Mahmood Mortazavi Dehkordi

August 27, 2024

[Introduction (Teamwork) 3](#_Toc379653748)

[DAP Design and Implementation (Individual work) 4](#_Toc1887761586)

[Step 1: Data Analytical Question Formulation 4](#_Toc1635678385)

[Step 2: Data Discovery 4](#_Toc1615259722)

[Step 3: Data Storage Design 8](#_Toc1700180782)

[Step 4: Dataset Preparation 10](#_Toc235666149)

[Steo 5: Data Ingestion 16](#_Toc1529270637)

[Step 6: Data Storage 19](#_Toc848114813)

[Step 7: Data Pipeline Design 23](#_Toc1777204349)

[Step 8: Data Cleaning 23](#_Toc715190994)

[Step 9: Data Structuring 29](#_Toc366715064)

[Step 10: Data Pipeline Implementation 32](#_Toc1211431193)

[Step 11: Data Analysis 35](#_Toc1148269958)

[Step 12: Data Visualization 35](#_Toc322022985)

[Step 13: Data Publishing 40](#_Toc627709591)

[DAP Estimated Cost (Teamwork) 49](#_Toc1750497960)

# **Introduction**

The project focusses on developing Data Analytic Platform (DAP) for the City of Vancouver to analyze various datasets like Operating permits for Water Treatment, Parking Tickets, Lost and Found Pets, 3-1-1 Inquiry Volume from 2023 and 2024. DAP is implemented using Amazon Web Services which is an online platform allowing users to store and interact with information, generating value from each action. Taking advantage of the wide variety of services offered by AWS, we were able to design, store, edit, and visualize our information for future analysis using different AWS available tools.

S3, Glue Data Brew, Athena, Glue, and EC2 were the services used. Buckets were generated to store information using S3, then transformed and evaluated using Glue Data Brew and edited using Glue services to finally publish and visualize our information using EC2. All these services allowed us not only to understand the platform better but also to improve the way we treat our information.

In the following pages, a step-by-step guide on how each of the datasets and services were used will be visible, with a brief explanation of each of the steps that will guide the reader to a comprehensive in-depth analysis on the potential of AWS toolkit for storing, managing, analyzing, evaluating, visualizing information and a lot more.

# **DAP Design and Implementation**

## **Step 1: Data Analytical Question Formulation**

Open data portal provides us with public data of City of Vancouver through tables, maps and charts. To analyse the critical issues for the city of Vancouver, our team exported the data for Operating permits for water systems, the evolution of parking tickets, Animal Control Inventory: Lost and Found animal control inventory and inquiry volume of 311 services for the years of 2023 and 2024. As the city of Vancouver needs to implement some transformations on these issues, our team analyses the below aspects for each department and these records will be evaluated and maybe merged with additional information to create a more complete picture of Vancouver trends in each case.

* What are the trends in Legionella pneumophila concentrations in mechanical system type over the past two years?

## **Step 2: Data Discovery**

From opendata.vancouver.ca, Data team identified a robust dataset that could help us answer our initial doubts; the data from the website ensures open Government license, data quality, consistency, and governance, providing flexibility to leverage data effectively.

**Figure 1**

*Data Discovery for Operating permits for water systems*

A screenshot of a computer

Description automatically generated

Our team exported the data for Operating permits for water systems for the years of 2023 and 2024 covering our two-year research span and categorizing the data for mechanical system type to analyze the trends for Legionella pneumophila concentrations.

## **Step 3: Data Storage Design**

In data storage phase of AWS Data Analytic Platform, the main objective is to set up the strong foundation and architecture to store the operational as well as the analytical data securely and optimally. This is made possible through Amazon Simple Storage Service (S3) that provides highly accessible, scalable as well as durable object storage. S3 is well suited for storage of any kind of data, within a broad range, from raw operational data and through analytical data structures to processed datasets, all stored and managed with no compromises on integrity and availability.

Using Amazon S3 we have created three buckets for storing the excel files. S3 buckets are designed in a way that each bucket can store the data for two years 2023, 2024 and in each year every operational bucket has 3 folders Landing, Raw and Curated. Operational dataset is loaded into landing zone in excel format and the analytical, structured, cleaned data is loaded into raw folder with high quality and the analytical data with summarized results are stored in curated folder in csv format.

**Figure 5**

*Data Storage Design for Operating permits for water systems*

A screenshot of a computer

Description automatically generated

## **Step 4: Dataset Preparation**

The main aim of dataset preparation is to assimilate and categorise the data by ensuring quality and functionality of all steps for analysis purpose. This encompasses data ingestion, cleaning, normalization, and structuring. AWS Glue DataBrew is used for data categorization and allows working with different data sources in different platforms. Data preparation takes the ingested data and processes to a format data (What Is AWS Glue DataBrew? - AWS Glue DataBrew, n.d.).

In AWS Glue DataBrew, we have created 2 separate projects one for each datasets for 2023 and 2024. In Cleaning the datasets, we changed the column names and their datatypes and delete the unwanted columns for precise results. Created 2 jobs to prepare for two CSV file dataset

**Figure 9**

*Data Discovery for Operating permits for water systems* A screenshot of a computer

Description automatically generated

For both the datasets the following steps were taken,

1. **Delete column** Permit details link
2. **Change type**ofTurbidityto**String**
3. **Change type**ofTemperatureto**String**
4. **Change type**ofSystem report dateto**Date**
5. **Create column**System report date\_YEAR using dateTime function YEAR
6. **Create column**System report date\_MONTH using dateTime function MONTH
7. **Delete column** System report date

## **Step 5: Data Ingestion**

AWS DataBrew is used for ETL job generation which automates and scales the data ingestion process, applying defined transformations to data consistently. The goal of data ingestion is to clean and store data in an accessible and consistent central repository to prepare it for use within the organization. It manages the execution, scheduling, and output storage, ensuring efficient and reliable data preparation.

Here, Our Data Team created 2 jobs to prepare for two CSV file datasets.

**Figure 13**

*Data Discovery for Operating permits for water systems*

A screenshot of a computer

Description automatically generated

## **Step 6: Data Storage**

The analytical datasets obtained from Datasets are stored in the raw folders of S3 buckets for each year respectively. The datasets in raw folder are of high quality with the structured csv format.

**Figure 17**

*Data Discovery for Operating permits for water systems*

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

## **Step 7: Data Pipeline Design**

Data pipeline design in AWS Glue involves creating a workflow that automates data ingestion, transformation (ETL), and cataloging. In Extract it retrieve data from raw folder from S3 Bucket and in Transform step, join, and aggregate are used to normalize and reformat the extracted data to ensure consistency and prepare it for analysis and the output is loaded into Curated folder of S3 buckets (What Is AWS Glue? - AWS Glue, n.d.).

The pipeline is designed with S3 bucket first to get the data from raw folder of S3 and then transformed the data from functions and the two data tables are then joined to get the consolidated data and the final output is loaded into storage of S3 buckets.

**Figure 21**

*Data Discovery for Operating permits for water systems*

A screenshot of a computer

Description automatically generated

ETL Pipeline is created to create a comparison table for the presence of Legionella pneumophila in Colling Tower machinal system type for the first seven months of 2023 and 2024

## **Step 8: Data Cleaning**

In AWS Glue, the filter function is used to refine data during the transformation process by selecting only the rows that meet the criteria. This helps in narrowing down the dataset to include only the relevant information, improving the efficiency and focus of the analysis.

**Figure 25**

*Data Discovery for Operating permits for water systems*

A screenshot of a computer

Description automatically generated

## **Step 9: Data Structuring**

Change Schema in AWS Glue, involves modifying the structure of the data as it's being processed, such as adding, removing, or altering columns and column names.

**Figure 28**

*Data Discovery for Operating permits for water systems*

A screenshot of a computer

Description automatically generated

## **Step 10: Data Pipeline Implementation**

The job created in Glue is responsible to load the output in the curated folder of storage of S3 bucket. The dataset we obtain is analytical dataset with high quality and in the structure format with the summarized results in CSV format. But the data loaded is not in a readable format. So, we need to proceed to next steps.

**Figure 32**

*Data Discovery for Operating permits for water systems*

A screenshot of a computer

Description automatically generated

## **Step 11: Data Analysis**

## The most important goal in this stage is to analyze processed datasets in order to gain meaningful insights and information. Amazon Athena is an essential tool intended for this task, it helps to perform an interactive query on data which is stored in Amazon S3 using SQL.

## Athena is a serverless analytical query service offered by AWS that let you analyze data in S3 using SQL. They support wide data formats, and here they have aws glue integration for partitioning the schema. Athena was specifically designed to cost only for the amount of data scanned, which makes it easy on the pocket and very accommodating to data scientists. The use case is especially suited for ad hoc querying, log analysis and querying data held in a data lake (What Is Amazon Athena? - Amazon Athena, n.d.).

**Figure 36**

*Data Discovery for Operating permits for water systems*

A screenshot of a computer

Description automatically generated

## **Step 12: Data Visualization**

The data can be downloaded in CSV Format and is available in the readable format, with the available data the Excel file generates the recommended charts to be presentable and can be easily analysed. The developed results are stored in pdf format so keep the format un disturbed.

**Figure 40**

*Data Discovery for Operating permits for water systems*

A screenshot of a computer

Description automatically generated

## **Step 13: Data Publishing**

Amazon EC2 (Elastic Compute Cloud) provides scalable computing capacity in the AWS cloud, allowing you to run virtual servers, known as instances, with various configurations based on your needs. EC2 instances, including t2 instances, can be used for a wide range of purposes, such as running general-purpose servers and web servers. T2 instances are designed to provide a baseline level of CPU performance with the ability to burst to higher levels when needed, making them suitable for applications with variable workloads.

A general-purpose EC2 server is used for a variety of tasks like providing access to files to remote hosts with security.

**Figure 44**

*Data Discovery for Operating permits for water systems*

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

An EC2 web server is specifically configured to host websites or web applications. It handles HTTP requests from clients and serves web pages or web applications in response. By installing IIS (Internet Information Services) platform from server manager for hosting websites, web applications, and services. Making the data accessible from the wwwroot as this is the key directory for the context of web servers. With the assigned IP address from EC2 instance the data is published and accessed through Internet.

**Figure 48**

*Data Discovery for Operating permits for water systems*

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

**Figure 49**

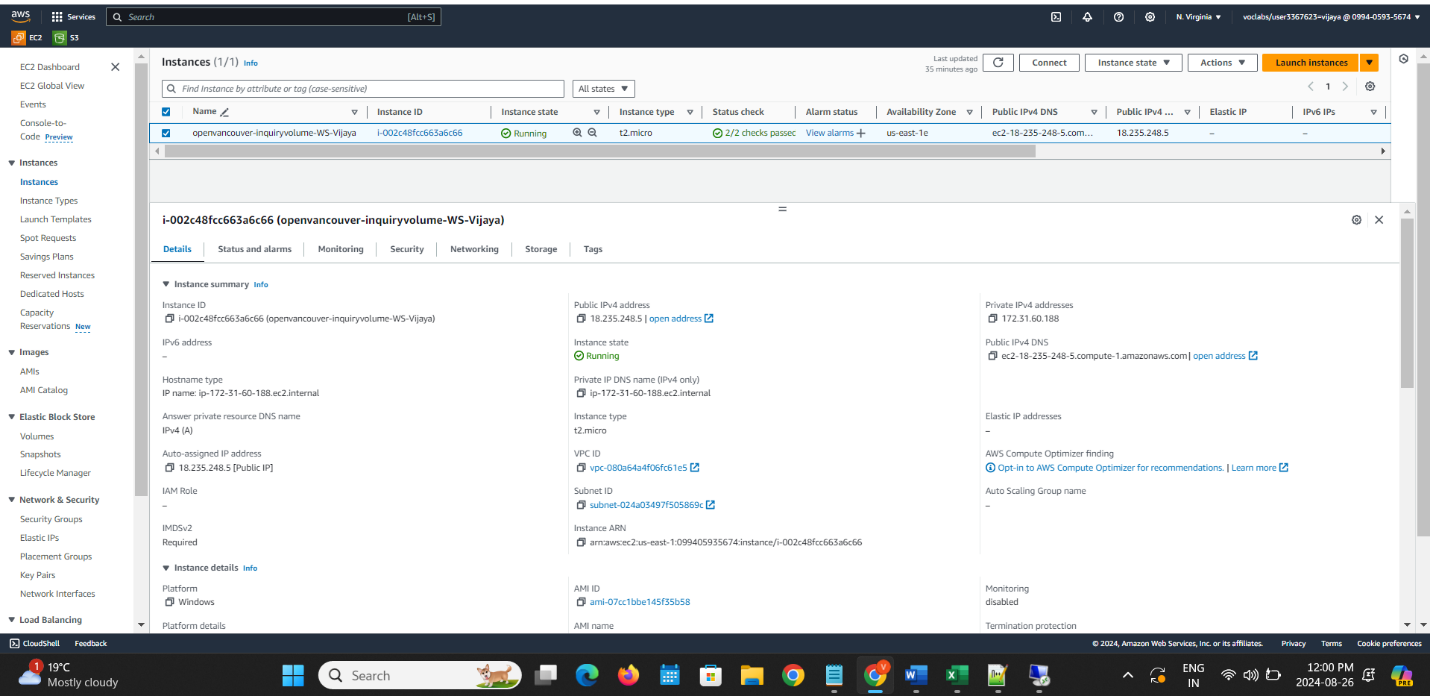
*Data Discovery for Operating permits for water systems*

A computer screen shot of a computer

Description automatically generated

**Figure 50**

*Data Discovery for Operating permits for water systems*

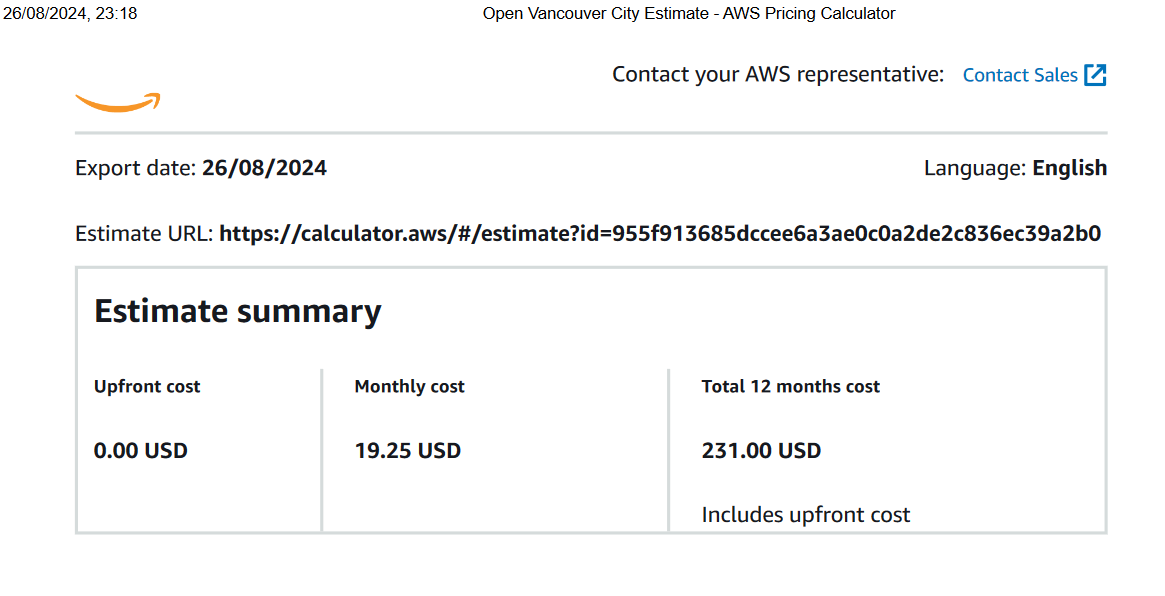


# **DAP Estimated Cost**

We used the AWS pricing calculator to estimate the costs associated with our platform, including all of our datasets, and considering the services used, we came up with a total cost of 19.25 USD per month or 231 USD per year. This includes S3, AWS Glue, Athena, and EC2 services. We also expanded the storage capacity in case the datasets continue to be fed (assuming the DAP continues to be used in future months) to mitigate storage issues in the future (What Is AWS Pricing Calculator? - AWS Pricing Calculator, n.d.).

**Figure 52**

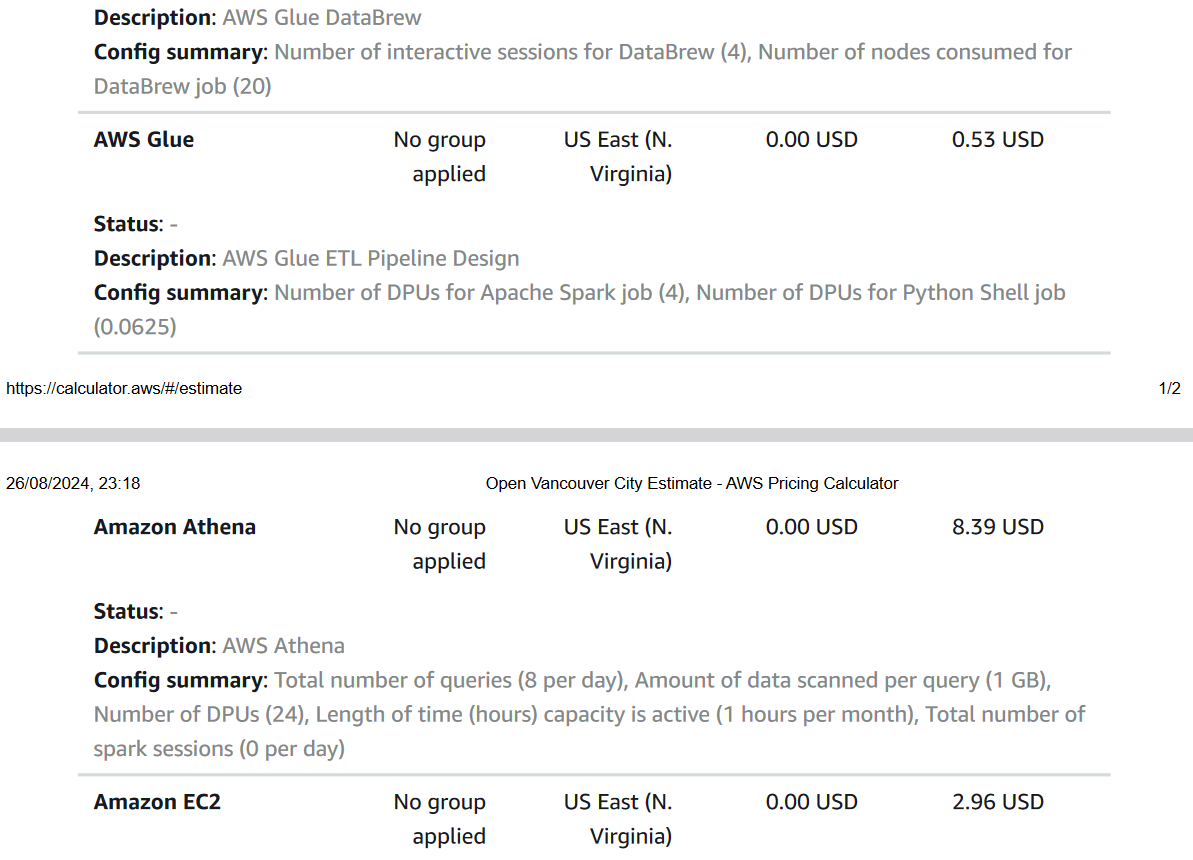
*Data Discovery for Operating permits for water systems*

**

Also, on observing the Storage system in AWS is very affordable, AWS Storage is preferable for integrity purposes. The Glue and Athena are expensive as they are main part of migration however, this gives a proper structure and giving in a readable format and Ec2 is not very expensive but when adding extra features for confidentiality, integrity and availability, it would be expensive.

**Figure 53**

*Data Discovery for Operating permits for water systems*

****

These estimations allowed us to understand the potential costs that using AWS services would have to cover and realize how convenient it is to have complete administrative service that helps you perform different activities with your information and guarantees a safe and secure environment for the datasets.

**References**

*3-1-1 inquiry volume*. (2024, August 1). <https://opendata.vancouver.ca/explore/dataset/3-1-1-inquiry-volume/information/?disjunctive.department&disjunctive.type&disjunctive.channel>

*Demo: Getting started with Amazon S3 (7:37)*. (n.d.). [Video]. Amazon Web Services, Inc. https://aws.amazon.com/s3/getting-started/#:~:text=Amazon%20Simple%20Storage%20Service%20(Amazon,at%20any%20time%2C%20from%20anywhere.

*Parking tickets*. (2024, July 15). https://opendata.vancouver.ca/explore/dataset/parking-tickets/information/

*What is AWS Glue DataBrew? - AWS Glue DataBrew*. (n.d.). <https://docs.aws.amazon.com/databrew/latest/dg/what-is.html#:~:text=AWS%20Glue%20DataBrew%20is%20a,data%20without%20writing%20any%20code>.

*What is AWS Glue? - AWS Glue*. (n.d.-b). https://docs.aws.amazon.com/glue/latest/dg/what-is-glue.html

*What is AWS Pricing Calculator? - AWS Pricing Calculator*. (n.d.). https://docs.aws.amazon.com/pricing-calculator/latest/userguide/what-is-pricing-calculator.html

*What is Amazon Athena? - Amazon Athena*. (n.d.). https://docs.aws.amazon.com/athena/latest/ug/what-is.html#:~:text=Amazon%20Athena%20is%20an%20interactive,Amazon%20S3)%20using%20standard%20SQL.