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

**“Phase 2”**

**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

September 15, 2024

Table of Contents

[Introduction 3](#_Toc177323115)

[DAP Design and Implementation 3](#_Toc177323116)

[Step 15: Data Protection 4](#_Toc177323117)

[Step 16: Data Governance 15](#_Toc177323118)

[Step 17: Data Monitoring 27](#_Toc177323119)

[DAP Architecture Analysis 39](#_Toc177323120)

[Operational Excellence 39](#_Toc177323121)

[Security 39](#_Toc177323122)

[Reliability 40](#_Toc177323123)

[Performance Efficiency 40](#_Toc177323124)

[Cost Optimization 41](#_Toc177323125)

[Sustainability 41](#_Toc177323126)

[Conclusion 41](#_Toc177323127)

# **Introduction**

# As mentioned today, cities need stable data platforms to store and analyze various data types, which are essential in the modern world. The City of Vancouver needs to process a massive amount of data concerning the urban infrastructure and public services to provide the people with the relevant data necessary for making efficient decisions. This project created an AWS Data Analytics Platform to meet these challenges by employing AWS's safe cloud services for storage, encryption, monitoring, and analysis of essential datasets while still making these datasets available and secure.

# The platform sets up a safe environment for the timely exercising of control, supervision, and management of data relevant to public services, ranging from operating permits to parking tickets and animal control, to name a few. AWS services such as Key Management Service (KMS), Amazon S3, CloudWatch, and CloudTrail are used to improve operation excellence, compliance, and cost control.

# Finally, this paper deals with the last steps, which consist of using key datasets for 2023 and 2024 to enable users to learn about changes in metrics. Such issues as data storage, processing, and security are solved through AWS, making the data secure and accessible. Thanks to the high level of hierarchy at this stage, AWS instrumentation helps the city to work with and analyze the necessary data. The paper further discusses the development and deployment of the platform with an emphasis on data security, access, and management, as well as monitoring the system's architecture for security, reliability, performance efficiency, cost optimization, and sustainability.

# **DAP Design and Implementation**

## **Step 15: Data Protection**

In the case of the city of Vancouver, we ensure the data Integrity, confidentiality, and replication of data with the help of services provided by AWS, namely Key Management Service (KMS), Encryption, Versioning, Replication, and Data backup to be retrieved from backup files, in case of accidental deletion or any cyber threat or server crash.

**Key Management Service (KMS)**

AWS KMS stands for AWS Key Management Service, an essential service that helps protect data under AWS through a centralized service to manage encryption keys. KMS is in charge of securely encrypting and decrypting data to allow only valid users and other systems access. In cases of many inquiries, KMS delivers high numbers of requests while ensuring the security of the encryption keys. In a way, it provides the confidentiality of the keys used in the data encryption and means that only authorized personnel or systems can access them. Customer-managed keys (CMKs) are where the user can create and manage keys with additional control over encryption and critical policies.

* This key is maintained within AWS KMS because this service involves data encryption and decryption.
* This helps in easy referencing and use throughout the AWS services by using the key alias.
* The key management (See role "LabRole”) lies in the hands of the key administrators to warrant the secure operations of the key.

**Figure 1**

*KMS for Operating permits for water systems*

KMS keys were created to gain access to necessary data with the confidence of its security.

*A screenshot of a computer

Description automatically generated*

**Encryption**

Amazon S3 provides server-side encryption (SSE) to protect data at rest automatically. Users can select from three main encryption options. SSE-S3 allows Amazon S3 to manage encryption keys using AES-256 without user intervention, making it ideal for those seeking a simple, low-maintenance solution. SSE-KMS uses AWS Key Management Service (KMS) to give users more control over key management, including key rotation, auditing, and access control, making it suitable for businesses needing compliance auditing or more granular control. For environments requiring the highest level of security, DSSE-KMS offers dual-layer encryption with two different keys, providing an extra layer of protection. Additionally, the Bucket Key option reduces costs by lowering the number of API calls made to KMS for decryption, which is useful for applications with frequent KMS requests. Users can choose the appropriate encryption method depending on organizational needs: SSE-S3 for simple encryption, SSE-KMS for greater control, or DSSE-KMS for maximum security (Protecting data with encryption - Amazon Simple Storage Service, n.d.).

In the below screenshot

1. The first thing I did here was enable the encryption of my primary bucket. The encryption type was SSE-KMS (Server-side encryption with AWS Key Management Service) keys. When the key was created in KMS, the role was selected as a lab role.
2. The second option was bucket versioning. This property was enabled to ensure that the multiple object variants are stored in the same bucket. This is good for integrity and security.

**Figure 5**

*Data Encryption for Operating permits for water systems*

A screenshot of a computer

Description automatically generated

**Backup**

AWS Backup is designed to automate the backup processes across various AWS services such as S3, EBS, RDS, and more. The service helps organizations manage their backups, ensuring their data is safe and recoverable during failure or disaster. The backup mechanism ensures that copies of data are regularly created and stored in separate locations, protecting against accidental deletion, corruption, or other data loss incidents. For higher inquiry volumes, AWS Backup scales to accommodate more frequent and more extensive backups, ensuring data availability and business continuity (Amazon S3 Backups - AWS Backup, n.d.).

**Figure 9**

*Backup for Operating permits for water systems*

A screenshot of a computer

Description automatically generated

The backup bucket for my primary bucket has been created. It is apparent from the naming convention. Like my primary bucket, versioning and encryption are enabled using the same key as the "lab role."

**Replication**

Replication, particularly across regions or availability zones, is critical to ensuring high availability and disaster recovery. AWS offers replication services such as S3 Cross-Region Replication, allowing organizations to duplicate their data to another AWS region automatically. This replication ensures that data can still be accessed from the replicated location even if a failure occurs in one area. For high inquiry volumes, replication services ensure data redundancy, maintaining consistency across multiple locations while minimizing downtime or data loss during disruptions (Replicating Objects Overview - Amazon Simple Storage Service, n.d.).

**Figure 13**

*Replication Rules for Operating permits for water systems*

A screenshot of a computer

Description automatically generated

I created a replication rule where the *cityvanc-opepermit-watersys* was my source bucket *and cityvanc-opepermit-watersys-backup* was my destination bucket. The assigned role was *labrole*. The scope was the entire bucket.

## **Step 16: Data Governance**

Data governance is a rigorous process of overseeing data management in a lifecycle manner to optimize the data's quality, security, availability, and compliance. Some of these are verifying the quality of the data and checking if the data is correct and has all the attributes needed for the intended use. Data governance also entails controlling sensitive information like PII through encryption, masking, or redacting privacy standards.

Besides the verification and certification of data, managers, and custodians of data apply the principles of data management in the area of data processing and sanitization in a way that helps to eliminate unwanted and redundant information in the systems, which are recognized as trusted. This includes sorting out information that may be useless and enhancing the quality of information that shall be processed or used.

Automation becomes a core aspect of modern data governance, where the defined workflows and jobs are executed periodically to perform data management tasks such as validation, transformation, and updating. In this way, by continuing such processes, organizations will guarantee their data's credibility, security, and compliance with standards and legal requirements.

In a nutshell, data governance is critical in exercising control over enterprise data to ensure the data available is sufficient, trustworthy, legitimate, and safeguarded from any illicit use.

Several Components used in designing the ETL process are: Several Components used in designing the ETL process are:

* Detect Sensitive Data: The process starts with identifying and masking Canadians' data restricted from being shared in different processes. From the feature, seven forms of sensitive data are masked, depicted by replacing the information with "\*\*\*\*\*. ".
* Evaluate Data Quality: A completeness check was carried out on the "Mechanical System Type" column of the data. This meets the validity of the data being gathered, hence readiness to take it to the next level of processing.
* Conditional Router: The valid data was filtered to pass through a router check to endorse only the sound data to proceed to the following stages.
* Change Schema: Some extra columns not used in the analysis were dropped; thus, the schema was changed.
* Trusted Zone: The final transformed data are written in the secure folder and stored under the primary S3 bucket.

**Figure 17**

*ETL For Operating permits for water systems*

A screenshot of a computer

Description automatically generated

The dataset is loaded from the raw zone, and the output is in the trusted zone. In this ETL job, we first implemented privacy using the Detect Sensitive Data feature from the transform. Here, seven types of Canadian sensitive data are redacted with \*\*\*\*\*. This was followed by other services like Evaluate Data Quality, which evaluated the "Mechanical System Type: column for completeness. Then, I used the Conditional Router to filter out the past data. Finally, the Change Schema is used to drop the unwanted columns, and finally, the data is stored in the trusted folder in my primary bucket.

This screenshot would show the output of the ETL job, displaying the cleaned and processed dataset ready for use. The output is stored in the trusted zone, confirming that the data transformation was successful and that any sensitive data is securely handled.

**Figure 18**

*ETL Output for Operating permits for water systems*

A screenshot of a computer

Description automatically generated

**Trusted Zone**

The Trusted Zone is where the validated and transformed data is stored after undergoing the ETL process. In this case, the operating permit data for water systems is now securely stored and available for further analysis or reporting.

**Figure 24**

*Trusted Zone for Operating permits for water systems*

Stored data in a trusted zone.

A screenshot of a computer

Description automatically generated

**Data Quality**

The Data Quality tab showcases the quality of the final dataset after the ETL process. In this instance, the data quality score is 100%, indicating that the data meets all the necessary quality standards and is complete, clean, and ready for use.

**Figure 27**

*Data Quality for Operating permits for water systems*

A screenshot of a computer

Description automatically generated

We can check the quality of our final dataset from the data quality tab. In my case, the quality was 100%.

**Workflows**

This screenshot shows the schedule set for the ETL job. In this case, the job is scheduled to run every Monday at 11:59 PM, ensuring the dataset is regularly updated with fresh data. Automation in scheduling makes the data pipeline more efficient and hands-off (Overview of Workflows in AWS Glue - AWS Glue, n.d.).

A schedule is created to run the job every Monday at 11:59 PM. The workflow screenshot would depict the overall flow of the ETL process, from data extraction through transformation to storage in the trusted zone. This workflow makes all defined tasks run automatically without interruption at the desired times. It gives you a big picture of the data pipeline process to ensure each node is appropriately set up and run in the correct sequence (Time-based Schedules for Jobs and Crawlers - AWS Glue, n.d.).

**Figure 31**

*Scheduling For Operating permits for water systems*

A screenshot of a computer

Description automatically generated

**Figure 32**

*Workflow For Operating permits for water systems*

A screenshot of a computer

Description automatically generated

The schedule is automated by a workflow that hosts the schedule and ensures that the scheduled job is run correctly at desired intervals and times.

**Step 17: Data Monitoring**

AWS data monitoring is possible through services like CloudWatch and CloudTrail to understand the performance, availability, and usage of resources. These services offer accurate time monitoring and control of actions through alarms. These services provide a safe and compliant process.

**Amazon CloudWatch**

CloudWatch provides services like automated and customized dashboards where the performance metrics can be monitored easily. The graphical view gives an easy way to analyze the metrics and can easily add metrics and alarms to it. The displays, such as line graphs and single number displays, provide an easy way to monitor across multiple accounts and regions (How Amazon CloudWatch Works - Amazon CloudWatch, n.d.).

**Figure 36**

*Data monitoring for Operating permits for water systems*

A screenshot of a computer

Description automatically generated

Dashboards for monitoring expenditures made it easier for us to track and trace the costs associated with each task completed during the project.

**CloudWatch Alarm**

CloudWatch Alarms are used to set thresholds for specific metrics. When a threshold is reached, alarms will trigger and send alert messages or even do auto-scaling. Alerts can be configured for various metrics such as CPU, memory, or any other application-level metrics so that users can act on them in case of a problem. Composite alarms include multiple conditions, creating more complicated monitoring schedules (Using Amazon CloudWatch Alarms - Amazon CloudWatch, n.d.).

In this screenshot, an alarm system is used to keep track of the health and efficiency of water system operating permits. This serves the purpose of alerting in real-time to check whether any deviation from the normal traffic flow is noted to alert. These alarms are usually configured by a threshold defined in AWS CloudWatch. If any of them exceed those thresholds, alert notifications will be generated to the concerned parties.

**Figure 40**

*Alarm for controlling Operating permits for water systems*

*A screenshot of a computer

Description automatically generated*

This figure depicts the actions for perpetual surveillance of the operating permits for the water systems. Data monitoring means observing its state and functionality and checking the conformity and reliability of data at a specific period. It also helps to check that any problems/ mistakes in the collected data are detected as they reflect on sample quality. Besides, it assists in monitoring whether the variables used comply with the company's operational standards and regulations. Alarms were designed to activate during unexpected activity or excessive spending.

Alarms were designed to trigger in the event of excessive expenditure or odd behavior.

**Data Monitoring and controlling**

Monitoring metrics like the status of data, performance, consistency, and accuracy help to identify issues and errors in data quality and can be modified and rectified quickly. Precise monitoring can help control the actions by triggering alarms and maintaining compliance with regulatory standards.

**Figure 44**

*Data Monitoring and controlling Operating permits for water systems*

A screenshot of a computer

Description automatically generated

**AWS Cloud Trail**

AWS CloudTrail works with CloudWatch to capture the events throughout the AWS service and intends to improve the API's performance. It logs who accessed the services, when, and from where, making it a preferred tool for security personnel, auditors, and lawyers. Regarding data utilization, CloudTrail can write logs to S3 for record-keeping analysis in the long term and also link with CloudWatch to issue alerts to possible unauthorized activities. It can also recognize contravention of API usage and even additional security breaches occurring that CloudTrail Insights can prevent.

AWS CloudTrail is a service that tracks all the API calls and user activities made in the AWS environment. These screenshots represent the CloudTrail logs, which contain all the activities related to the operating permits for water systems. CloudTrail allows tracking of user actions and is very useful for auditing, compliance, and security purposes. It ensures that all the changes and the accesses made to data are captured for future reference or in case of investigation (What Is AWS CloudTrail? - AWS CloudTrail, n.d.).

**Figure 48**

*Cloud Trail is for operating permits for water systems.*

A screenshot of a computer

Description automatically generated

**Figure 52**

*S3 Bucket cloud Trail output for Operating permits for water systems*

A screenshot of a computer

Description automatically generated

**DAP Architecture Analysis**

Cloud Architecture and AWS Framework designed in part 1 and part 2 from the Data Analytics platform for the city of Vancouver can be assessed in terms of six pillars: Operational Excellence, Security, Reliability, Performance Efficiency, Cost Optimization, and Sustainability.

## **Operational Excellence**

AWS Glue service is used to handle ETL operations and reduce the amount of human interaction to improve operational efficiency. The Platform is also leveraged with Amazon CloudWatch for performance monitoring and control with alarms by enabling the thresholds and sending an alert for safety standards.

However, we will need to create more automated processes to ensure we align well with the best operational excellence practices in Amazon Web Services. We might address this by implementing AWS Lambda for event management, such as automatically removing outdated information or triggering data validations.

Amazon CloudWatch is integrated into the Platform as a monitoring, logging, and alerting solution. This means that we can monitor the health and workload of AWS resources and address events or incidents on time.

We also use AWS Glue for ETL operations, a fully self-serverless service that enhances operational utilization. It is fully automated and does not require workers to perform tacit operations.

Some of the best practices identified for improving operational excellence are the automation of the production line and change intensification, known as SPCR – small, frequent, reversible changes. This is well supported by the architecture-level automation of Glue and CloudWatch in our architecture.

## **Security**

Security is one of the most vital concerns on our website. We have incorporated the AWS KMS and utilized it to manage the keys to encrypt all the data that flows in the Landing, Raw, and Trusted stages. Further, we have also embedded the quality and privacy assessment step, which can provide a second level of filtration to exclusively process information from trusted sources in the successive stages of the pipeline.

To strengthen security further, we'll refine our IAM policies to follow the principle of least privilege more closely, ensuring that each user or system only has access to the necessary resources. Adding VPC endpoints for S3 could also boost the security of our data transfers.

## **Reliability**

We've used Amazon S3 for data storage, which offers excellent durability and availability, and the serverless nature of AWS Glue enhances the Platform's resilience. Still, we want to improve on the reliability aspect by implementing automatic backups and versioning within S3, especially in the critical Trusted data stage.

Leveraging AWS Trusted Advisor will help us evaluate any reliability gaps and implement recommendations for multi-AZ deployments or adding redundancy where needed. I might also consider using Route 53 to manage failover critical failures effectively.

## **Performance Efficiency**

We're already capitalizing on serverless services like Glue and Athena, which scale effortlessly with demand. Using real-time ingestion and separating data into Landing, Raw, and Trusted stages ensures we process only the most relevant and clean data, which helps optimize performance.

To further optimize, we'll explore using Athena's caching capabilities to reduce query times or move towards Amazon Redshift for more complex, high-performance analytics if our query needs to grow. We might also leverage AWS Lambda to streamline specific processes, like automating quality checks as new data flows into the system.

## **Cost Optimization**

We're mindful of costs by using serverless components that only charge for what we use, like Glue, Athena, and CloudWatch. Amazon S3 is another cost-effective solution for storing data, and we can further reduce costs by using S3 lifecycle policies to move older data into lower-cost storage tiers, like S3 Glacier.

To maintain cost efficiency, we will create AWS Budgets to help us monitor expenses and explore S3 Intelligent-Tiering, automatically reducing costs for seldom accessed data.

## **Sustainability**

Sustainability is now a remarkable aspect of the design of a cloud solution, and our Platform utilizes the following AWS sustainable service. Engaging AWS Glue and Amazon Athena, serverless technologies, do not need an ever-running infrastructure and have less power consumption. AWS data centers are now more efficient than traditional data centers due to the increasing use of renewable energy sources.

Moreover, we can store large datasets in S3 Glacier with the help of S3 lifecycle policies, decreasing the consumptive factors that are the key drivers of environmental prices. Such optimizations lead to optimizing costs and, simultaneously, optimizing the Platform's carbon footprint.

In the future, there will be EC2 Spot Instances, which are also the best practices at AWS and do not require critical workloads. Thus, this approach helps save even more energy since excess capacity is oriented within the AWS regions. We will also need to use the AWS Customer Carbon Footprint Tool to monitor our carbon footprint and check how sustainable the architecture will be when it is expanded.

# **Conclusion**

In conclusion, we have ensured that the data analytics platform we have deployed is well-designed for operational excellence, security, and performance, thus meeting AWS's recommended standards. We focus on improving some areas, such as reliability and costs, by using more of the services provided by AWS, such as Trusted Advisor, automated failovers, and wise cost-tracking tools. These steps will assist in availably enhancing the Platform's capacity to expand the infrastructure and augment the safety and efficiency standards.

After working on this project, it was clear to see how helpful AWS can be if the proper tools are used and the objectives are clear; the potential of each of the tools allows users to generate information, storage, transformation, queries, predictions, and more with the certainty that the data is protected.

For the information management structure created, not only can information be analyzed and edited, but it also has the potential to grow over time, giving the Platform a broader dataset to generate analysis. If conscious of the costs associated with the services, Amazon Web Services should be an option for every organization looking for a high-quality, user-friendly, and secure cloud service to administer the data and information generated for any business in any industry.

**References**

Amazon S3 backups - AWS Backup. (n.d.). https://docs.aws.amazon.com/aws-backup/latest/devguide/s3-backups.html

How Amazon CloudWatch works - Amazon CloudWatch. (n.d.). <https://docs.aws.amazon.com/AmazonCloudWatch/latest/monitoring/cloudwatch_architecture.html>

Overview of workflows in AWS Glue - AWS Glue. (n.d.). https://docs.aws.amazon.com/glue/latest/dg/workflows\_overview.html

Protecting data with encryption - Amazon Simple Storage Service. (n.d.). https://docs.aws.amazon.com/AmazonS3/latest/userguide/UsingEncryption.html

Replicating objects overview - Amazon Simple Storage Service. (n.d.). https://docs.aws.amazon.com/AmazonS3/latest/userguide/replication.html

Time-based schedules for jobs and crawlers - AWS Glue. (n.d.). <https://docs.aws.amazon.com/glue/latest/dg/monitor-data-warehouse-schedule.html>

Using Amazon CloudWatch alarms - Amazon CloudWatch. (n.d.). https://docs.aws.amazon.com/AmazonCloudWatch/latest/monitoring/AlarmThatSendsEmail.html

What Is AWS CloudTrail? - AWS CloudTrail. (n.d.). https://docs.aws.amazon.com/awscloudtrail/latest/userguide/cloudtrail-user-guide.html