**Cloud Computing and Data Analytic Portfolio Report: A Practical Approach Using AWS Services**

Chinazor Azubuike (2331035)

University Canada West.

BUSI 653 (HBD-WINTER25-06) Cloud Computing Technologies.

Mahmood Dehkodi

March 28, 2025

Table of Contents

[**Introduction** 2](#_Toc194023577)

[**Exploratory Data Analysis** 3](#_Toc194023578)

[**Title:** 3](#_Toc194023579)

[**Objective:** 3](#_Toc194023580)

[**Dataset**: 3](#_Toc194023581)

[**Methodology:** 3](#_Toc194023582)

[**Tools & Technologies:** 4](#_Toc194023583)

[**Timeline:** 7](#_Toc194023584)

[**Descriptive Analysis** 7](#_Toc194023585)

[**Title:** 7](#_Toc194023586)

[**Deliverables:** 8](#_Toc194023587)

[**Timeline:** 10](#_Toc194023588)

[**Diagnostic Analysis** 10](#_Toc194023589)

[**Title**: 10](#_Toc194023590)

[**Methodology:** 11](#_Toc194023591)

[**Timeline:** 13](#_Toc194023592)

[**Data Wrangling** 14](#_Toc194023593)

[**Title:** 14](#_Toc194023594)

[**Background Objective:** 14](#_Toc194023595)

[**Tools & Technologies:** 14](#_Toc194023596)

[**Deliverables:** 15](#_Toc194023597)

[**Timeline:** 17](#_Toc194023598)

[**Data Quality Control** 17](#_Toc194023599)

[**Title**: 18](#_Toc194023600)

[**Methodology:** 18](#_Toc194023601)

[**Tools & Technologies:** 18](#_Toc194023602)

[**Timeline**: 21](#_Toc194023603)

[**Conclusion** 22](#_Toc194023604)

**Introduction**

These files showcase my experience with cloudbased data management through utilizing Amazon Web Services (AWS) technologies to automatically implement powerful cloud based processes for handling, cleaning and validating large datasets. In the course I tackled various techniques on how to ensure data integrity, integrity, high quality standards and adopt best practices in data wrangling as well as data analysis.Screenshots and outputs included in this report are drawn from both my Project 1 and Project 2, as well as other selected weekly activities. Each phase Exploratory, Descriptive, Diagnostic, Data Wrangling, and Data Quality Control reflects my technical learning and ability to apply AWS services in real-life data situations. In this portfolio, projects and weekly activities are presented to show that AWS services like AWS S3, AWS Glue, AWS Athena, IAM and CloudTrail to handle, analyze, and check the quality of data. This work includes practical examples of how to ingest, cleanse, profile and validate data which were critical to building a robust pipeline.

Utilizing the cloud computing power, I was able to automate workflows, secure data, and have data quality checks to enable us to see what is changing and what action can be taken to help in making better decisions. The basis of this portfolio is that it encompasses both the technical skills and strategic thinking needed to work with cloud technologies in the context of data analytics.

**Exploratory Data Analysis**

**Title:** Exploratory Analysis of Employee Remuneration and Expenses (Over $75,000) in Engineering Services – Journeyman Mechanic**Project Description**: This initial review examines employee remuneration and expense patterns in the Engineering Services Department of the City of Vancouver. I applied the using exploratory data analysis (EDA) to the salaries and expenses for employees who earned over $75,000, with special attention to the Journeyman – Mechanic role. It helped me discover the patterns in salaries and spending across years and ensure the preparedness of the data for later analysis.

**Objective:** The projects and weekly activities in this portfolio make use of AWS S3, AWS Glue, AWS Athena, IAM and CloudTrail to handle, analyze, and quality check data. It's also practical examples of ingesting, cleansing, profiling and validating data that turned out to be critical in building a robust pipeline.

**Dataset**: The dataset comes from the City of Vancouver Open Data Portal and includes records of employees earning over $75,000. Specifically, I focused on the Journeyman – Mechanic category within the Engineering Services department.It includes fields such as:

* Title: Employee Remuneration and Expenses – Over $75,000
* Department: Engineering Services (City of Vancouver)
* Focus Role: Journeyman – Mechanic
* Fields: Name, Department, Title, Base Salary, Overtime, Expenses, Total Remuneration, Year

**Methodology:**

Data Preparation and Collection: I downloaded the dataset from the Open Data Portal, uploaded it to AWS S3 and prepared for it to be used to create the dataset.Profiling: I used AWS Glue DataBrew to review structure, duplicate records, and outliers, as well as to identify missing values.Data Cleaning: I filtered the data to only Journeyman – mechanic category, removed null entries and corrected format.Exploratory Charts: To understand change in salaries and expenses by year, I created informal tools like line charts and bar graphs that display salaries and expenses by year.Outlier Detection: I reviewed high salary and expenses values which seemed unusual or inconsistent.

**Tools & Technologies:**I used the below tools for this exploratory analysis:

Excel: This was used to clean, filter, and organize the datasetAWS S3: This was used to store both raw and cleaned data filesAWS Glue: This was used for profiling the data and organizing tables**Deliverables:**A cleaned CSV file of filtered Journeyman – Mechanic dataA summary Excel sheet showing salary and expense patterns by yearScreenshots of AWS Glue profiling resultsNotes on patterns and outliersExcel summaries and screenshots

**Figure 1: Raw Data Set from City Of Vancouver Open Portal**

A screenshot of a computer

AI-generated content may be incorrect.

Figure 1 displays the original unprocessed data as exported from the source. It contains various columns such as Year, Name, Department, Title, Remuneration, and Expenses. The dataset at this stage has blanks cells, inconsistent formatting, and duplicates. The raw form is the beginning before any cleaning, or transformation.

**Figure 2: Filtered Dataset by Year (2023)**A screenshot of a computer

AI-generated content may be incorrect.

Here in Figure 2 the data has been filtered to show only records for the year 2023. It helps in filtering out to focus on the most recent and relevant data entries. In particular, it is very useful when comparing trends or measurements of changes over time in remuneration or departmental performance.

**Figure 3: Uploaded Dataset to AWS S3**

A screenshot of a computer

AI-generated content may be incorrect.

This figure shows the successful upload of the raw employee remuneration dataset to an Amazon S3 bucket named *ren-raw-chi*. I first uploaded the dataset to S3, which gave me the ability to have the dataset being stored on the cloud and to access it. Storing the file in S3 allowed the dataset to easily work with other AWS services like AWS Glue and AWS DataBrew for more processing and analysis. CSV file format is used as structured tabular data and can be used with many Data Wrangling and Profiling tools. Also, this setup provides version control, scalable access and the security of large files during the phase of exploration of data.

**Timeline:**  
This phase was completed in about 2 days. I used Excel for exploration and filtering of the dataset and AWS S3 for raw and filtered files. This was part of the early work during Project 1.

The work of this EDA project laid the groundwork to perform further analysis. It helped in organizing the data for trend reviews, ensuring I had a reliable dataset to build the rest of the portfolio on.

**Descriptive Analysis**

**Title:** Descriptive Analysis of Salary and Expense Patterns for Journeyman Mechanics in Engineering Services**Project Description**: This phase builds on the EDA by summarizing the numerical patterns in the dataset. I quantified salary, expense patterns with descriptive statistics and there were differences by year and role. It shows the computing totals, averages, and distributions. The goal was to show what’s happening over time with salaries and expenses in the selected group.**Objective**: To describe trends in base salaries, overtime pay, and work-related expenses for the Journeyman – Mechanic employees earning over $75,000 from 2017 to 2023. This gave me a sense of patterns over the years and a comparison in metrics such as average remuneration. **Methodology:**Grouped data by year to compare average salaries and expensesCalculated trends over time using line chartsCompared total expenses to salary ratiosIdentified employees with the highest compensationSummarized salary brackets and distributions using Excel**Tools & Technologies**: The tools I used for the descriptive analysis are given below,

AWS S3: The cleaned dataset was stored in an structured format in AWS S3.Excel (with chart tools): It was used to generate summary tables and visualizations.AWS Glue: This allowed me to catalog and order the dataset by year and department.

## **Deliverables:**

Salary statistics by role and year represented in summary table. Salary and expense trend line and bar graphs. Screenshots from Excel with tables and charts.Yearly comparison of expense-to-salary ratios

**Figure 4: Total Remuneration Trend (2008–2023)**

A graph with a line and a line

AI-generated content may be incorrect.

This is a line chart that represents total employee remuneration from 2008 to 2023. The upward trend is clear, and there was a sharp increase between 2010 and 2013, which may be caused by rising salaries, or by the fact that the company started giving staff bonuses. The amount of remuneration stays fairly stable from 2014 to 2022 with a brief hiatus.

Nevertheless, the peak of 2023 salary level is an all time high, delineating a sharp change in salary structure or perhaps in workforce size. This trend helps highlight significant periods of financial growth and areas that may need further review.

**Figure 5: Salary and Expense Trends***A screenshot of a computer

AI-generated content may be incorrect.*

This chart compares the yearly average remuneration and expenses of employees. It’s divided into different colours to show how salaries and expenses changed from one year to another. They show which years had notable increase or decrease in their salary vs. expenses. It is a snapshot of pattern of overall compensation, and it sheds light on outliers or patterns (for e.g. sudden increase or decrease in salary).

**Figure 6: Top 5 Highest Paid Journeyman Mechanics – 2023**

**A graph with blue lines and black text

AI-generated content may be incorrect.**This bar chart displays the five employees who earned the highest remuneration in 2023. Their names are shown on the Y-axis, and their exact remuneration values are shown on the X-axis. It also helps identify the top earners in the dataset, which individuals earned significantly higher in 2023. This leads to further investigation as to what drove their pay, like their departments, their roles, or perhaps bonuses.

**Timeline:**  
This analysis took 3 days as part of Project 1. In Excel I made pivot tables, summarized earnings and expenses, and showed the results for the key trends in years.

The descriptive phase showed a consistent salary growth, which was also useful in estimating the financial behavior of employees. The result constituted a good starting point for diagnostic analysis.

**Diagnostic Analysis**

**Title**: Root Cause Analysis of Outliers in Journeyman Mechanic salary and expenses (Earnings Over $75,000)**Project Description:** Here I examined why some employees within the Journeyman – Mechanic group had much higher expense claims or overtime pay. This was done by querying the cleaned dataset using Athena on, I examined specific outliers in the data and comparing the outputs across different years.**Objective:** To learn more about specific cases when expenses or remuneration were out of the norm to understand potential reasons and confirm the accuracy of the data. To also find out what led to high overtime or abnormal expense ratios and whether it was due to specific years, job behavior, or reporting issues.

**Background Objective**: In descriptive analysis I identified some high expenses in some years and some employee salary spikes in other years. I used diagnostic techniques to investigate these.

**Methodology:**I used some Athena SQL queries to analyze salary and expense patterns. Calculated averages, variance, and standard deviation was used to detect anomalies.I focused on employees with higher expenses.I Made comparison with yearly values to see irregular spikes.I Created and monitored query access and activity using CloudTrail logs.**Tools & Technologies:**AWS Athena: To query my cleaned dataset stored in AWS S3, I used Athena. I used SQL Queries to find patterns and outliers for instance, the year with the highest salaries e.t.c.

AWS S3: This was used as a secure storage space for both the cleaned and raw data files. I uploaded my datasets here so they could be queried later through Athena.

AWS Glue: This helped to prepare and organize the dataset. I profiled my data here, removed the duplicates, cleaning the structure (like column names) before querying.

CloudTrail (to log access): This was used to track the data access. This recorded who queried the dataset and when, which was useful for spotting if someone got their hands on sensitive information or if any odd queries were made.  
**Deliverables:**Athena query resultsScreenshots and SQL logsOutlier explanation summary

**Figure 7: Athena query for Summary Statistics**

A screenshot of a computer

AI-generated content may be incorrect.

This figure shows an Athena query result summarizing total and average remuneration and expenses by year. It lists the max and min value, so it makes it very easy to spot an unusual spike or drop. For instance, 2023 had a high total salary payout, while 2016 had lower averages. It also flagged specific years for deeper checks and the possible cause like overtime, promotions or data issue.

**Figure 8: Athena Query for Year with the Highest Remuneration**

A screenshot of a computer

AI-generated content may be incorrect.

This Athena query was used to identify the year with the highest total remuneration from the dataset. The result showed 2018 as the top year, with the largest overall salary payout. This helped to quickly focus further analysis on that year to understand why it stood out—possibly due to increased overtime, promotions, or a higher number of employees. It was a useful step in the diagnostic phase.

**Timeline:**   
Completed in 3 days during Project 2. To check these outliers against historical averages, I ran SQL queries using Athena and identified them as outliers in salary and expenses.

This diagnostic step also indicates years when salaries have not kept pace with the increase in expense, which may support policy change or the possibility of misreporting. It gave it a certain depth to the understanding of financial trends.

**Data Wrangling**

**Title:** Cleaning and Preparing Remuneration Data (Over $75,000) for Journeyman Mechanics in Engineering Services**Project Description**: This phase was dedicated to cleaning and transforming the dataset. The raw file contained missing fields, duplicates and inconsistent format. Wrangling was needed to ensure accuracy and usability before analysis.**Objective:** To clean, filter, and format the employee compensation dataset to remove errors and inconsistencies for proper analysis in the AWS environment.

**Background Objective:** Some of the issues with the original dataset were missing values, column mismatches, and rows with irrelevant data. These needed to be fixed before any analysis

**Methodology:**I checked if the dataset has any missing values and formatting issues in AWS Glue DataBrew.Null fields were cleaned, and all column names and formats were standardized.I added a new calculated column which is Expense to Remuneration Ratio.Saved the cleaned dataset to AWS S3 with folder structure based on year. Made sure that currency fields appeared correctly formatted and consistent.

**Tools & Technologies:Excel**: It was used to manually inspect and clean the raw dataset. I removed unneeded records, ensured date, numeric fields, job titles and department names were standardized, an.

**AWS S3**: This Served as a cloud-based storage solution. Here I uploaded both raw and cleaned datasets as well and structured them into folders per year and job role for better organization.

**AWS Glue**: This was used to catalog and profile the datasets. Missing fields, inconsistent format, duplicate entries were all detected by glue. It also supported the schema organization, which made querying in Athena easier during later stages.

**Deliverables:**

Cleaned dataset.Screenshots of profiling dashboard.Documentation of cleaning steps.S3 folder structure by year and role.

**Figure 9: Data Profiling Summary**

A screenshot of a data analysis

AI-generated content may be incorrect.

This summary was generated using AWS Glue DataBrew. It gives a quick view of dataset – number of rows (508) and number of columns (6). The tool identified different data types: integer, double, and string. It also flagged missing cells 22 in total mostly in the expense’s column. This profiling step was important because it discovered early data quality issues such as missing or duplicate values. This insight directed me to plan out the cleaning process correctly, and to have an accurate dataset for further analysis.

**Figure 10: Handling Missing Values**

A screenshot of a computer

AI-generated content may be incorrect.

This figure shows how I handled missing values in AWS Glue DataBrew. The profiling had revealed that some rows in the Expenses column were empty. To fix this, I used a recipe step to delete any row with missing values in that column. It was necessary for data quality and to avoid errors during the analysis. By removing those incomplete rows, the dataset became more reliable and easier to work with in the later stages of the project.

## **Timeline:**

This phase was spread across a week. I cleaned the raw data in Excel, standardized the format, and uploaded the structured version to AWS S3 with folder separation.

This stage ensured the reliability of all analyses by creating a valid dataset. It formed the backbone of the cloud data project and was essential for downstream processing.

**Data Quality Control**

**Title**: Monitoring and Securing Remuneration Data for Journeyman Mechanics in Engineering Services**Project Description:** This part involves putting governance and quality control measures in place: making sure data is valid, secure, and well-monitored in the cloud environment.**Objective**: To validate all data that is uploaded into AWS is valid, secure, auditable and in a regular consistent manner.**Background Objective:** Insights as well as decision making are impaired by poor data quality. Once ingested, I used several AWS tools to monitor job success and validate incoming data as well as enforce access control.

**Methodology:**Validation Rules: I established limits for the maximum salaries and expenses to fall within historical ranges.Logging: I turned on CloudTrail to monitor access of dataset.Monitoring: To get job logs, I used CloudWatch, and enabled CloudTrail to monitor dataset access.Access Control: For data access and query control I assigned IAM roles.

**Tools & Technologies:**

**AWS CloudTrail:** This was used to track all data access activities. It logged detailed events of which data was accessed by whom, when, helping with accountability and tracing any unauthorized action.**AWS CloudWatch:** It was used to see real time logs and system metrics and helped to track user actions, detecting abnormal behavior, and checking that the validation processes ran successfully.**AWS KMS (Key Management Service):** This ensured encryption of sensitive data. It ensured that security of key creation and access controls enhanced compliance and databased storage or transmission while ensuring its protection.**AWS Athena**: This also validated the salary and expense patterns in the dataset. This was especially useful in helping to cross verify the records, catch anomalies and provide consistency in the final dataset.**AWS S3:** It was used as the main storage layer. Datasets were cleaned and validated and stored in well organized directory labeled as 'passed and failed' outputs to make data governance fuss free.**AWS IAM (for permission control):** It managed user roles and only allowed authorized individuals to access sensitive datasets. Maintaining data security and integrity.

**Deliverables**:IAM role showing restricted access to datasetCloudTrail access logs tracking dataset queriesSummary report of any data errors or issuesValidation checks for salary and expense limitsGlue job logs and execution statusSecurity policy summary for data protection

**Figure 11: Monitoring and Controlling on Alarm Dashboard – CloudWatch**

A screenshot of a computer

AI-generated content may be incorrect.

This figure was taken from my week 9 activity on using CloudWatch and cloud trail. It demonstrates how I used Amazon Cloud Watch to monitor my data environment and create alert if there is unusual activity. Metrics such as job runs, errors can be tracked and issues identified quickly through top dashboard. The CloudWatch Logs in the bottom section display all job activity including who accessed the data, when it happened, and what the action was. These tools allowed me to be sure that the dataset was being treated with security, that the dataset met quality expectations, and that I would be alerted if anything went wrong (e.g., job failures or unexpected spikes in the use of resources).

**Figure 12: AWS Glue Job Completion Confirmation**

A screenshot of a computer

AI-generated content may be incorrect.

This figure confirms that the job ran successfully. It displays the job status as “Completed” along with the run time details and job metrics. This step was vital in validating that all cleaning and transformation processes were executed correctly. The success status indicated that the job did not fail at any point, and the output was generated as expected. Through this final confirmation step, I was sure that my data pipeline worked, the dataset is ready for further analysis and everything fit quality control and monitoring standards.

**Timeline**:This was completed in 3 days during Week 10 of the course and as part of Project 2I then applied IAM policies, turned on CloudTrail for data access logging, and checked the values of some of the key fields in the dataset.

This step helped me have confidence in the quality and the security of my data pipeline. It made sure that analytics were reliable, and that data were both compliant and of integrity.

# **Conclusion**

This cloud computing portfolio represents a hands-on demonstration of my skills in using AWS services to perform data analysis, cleaning, profiling, transformation, and quality control. Drawing insights from both my Project 1, Project 2 and some of the weekly activities done during the course, I was able to analyze, organize, and validate data through various phases using tools like AWS S3, Glue, Athena, and CloudWatch. The purpose of each task, whether it was exploratory analysis, handling missing values, or checking diagnostics, was to build a clean and insightful dataset which helps in making sound decisions.

Outside of having the technical skills, this portfolio illustrates that I can manage real life data scenarios from start to finish, resolve issues independently, and follow best practices in cloud-based data governance and visualization. It also demonstrates how my knowledge of cloud platforms such as AWS can be instrumental to simplify and secure data workflows to be more reliable and scalable.

In general, this experience has helped me to develop my confidence in using cloud tools and further my passion in regard to data analytics. I am proud of what I was able to do and am excited to apply these skills to more of a real-world environment in the future.