Customer Purchase Insights at FreshMart

A Cloud-Based Descriptive Analytics Portfolio Using AWS

Kushal Karki (2301351)

University Canada West

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

Professor: Mahmood Mortazavi Dehkodi

Due before 11:59 PM (PT) on Sunday, March 27, 2025

Table of Contents

Project Description

The descriptive analysis explores complete customer shopping activities at FreshMart which operates as a mid-sized Canadian retail enterprise running stores selling food items electronics and apparel in various locations (Freshmart® | Your Neighbourhood Grocer, Here for You., n.d.). The analysis seeks to extract business-relevant insights from transaction records which will enable better inventory management and marketing planning and customer relationship strategies. A complete data pipeline is established through Amazon Web Services cloud-based technologies which covers data ingestion followed by profiling and cleaning together with cataloging and summarization along with analysis and governance and monitoring. The project enables hands-on demonstration of AWS services while showing step-by-step instructions with actual screenshots. The original screenshots were taken from data acquired at a with some activities which I did for Vancouver city Datasets and universities policies activities as examples for demonstrating equivalent FreshMart technical operations.

Project Title: Understanding Customer Purchase Patterns at FreshMart

Objectives

The main mission of this undertaking focuses on describing a year of FreshMart transaction records using descriptive analytics techniques. An analysis using a secure and scalable cloud system that processes the data allows identification of trends which reveals peak shopping times together with the most popular product categories and payment choices selected by customers. The solution combines enhanced operational efficiency with better customer service support to direct future business planning.

Dataset: The analysis relies on 20 representative transaction records obtained from FreshMart for data evaluation. A synthetic yet simplified dataset structure represents retail operations through its structure. Each record contains:

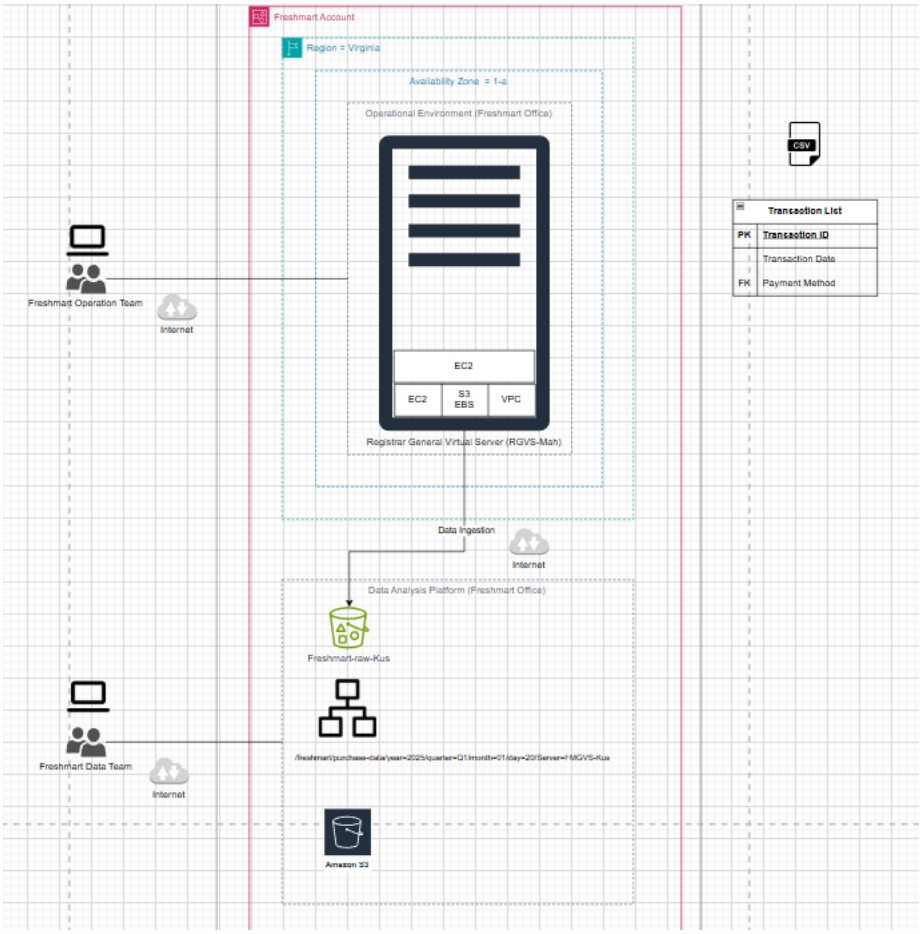
- A unique transaction ID
- A customer ID
- Purchase date and time
- Product category
- Quantity purchased
- Total price
- Payment method (cash, credit, debit, or digital wallet)
- Store location

Methodology

The data pipeline used AWS infrastructure to process raw sales data through different implementation steps for generating strategic business insights. Each stage of the data pipeline will be analyzed based on purpose of execution with an explanation of screenshot insertion.

Data Ingestion: The first stage of data storage involved placing raw transactions within Amazon S3 which represents AWS's secure cloud-based storage solution. The data transfer required a Windows-based EC2 instance to send information to the FreshMart-raw-kus structure S3 bucket through PowerShell command execution. A secure data storage environment was created through this step.

Diagram 1: FreshMart Data Ingestion Architecture



The data ingestion platform for FreshMart's transactional information uses AWS service infrastructure as shown in the diagram below. The transaction records created by store personnel travel securely through internet communication before being received and stored in an Amazon EC2 North Virginia region deployment. Secure PowerShell commands enable data storage inside the EC2 environment into the Amazon S3 bucket by following a structured folder naming scheme that combines date and store data. Through this configuration FreshMart's head office data analytics team can leverage a scalable data ingestion pipeline to easily access process and analyze transaction data in the cloud.

The image shows how FreshMart applies S3 bucket organization to handle its file uploads. The folder path contains year, quarter and server name fields which help Partition data in an orderly manner for daily received Store-level transaction data. FreshMart stores should upload sales data into the specific bucket for centralized processing during regular operations.

Figure 1: AWS EC2 Instance for Remote Access

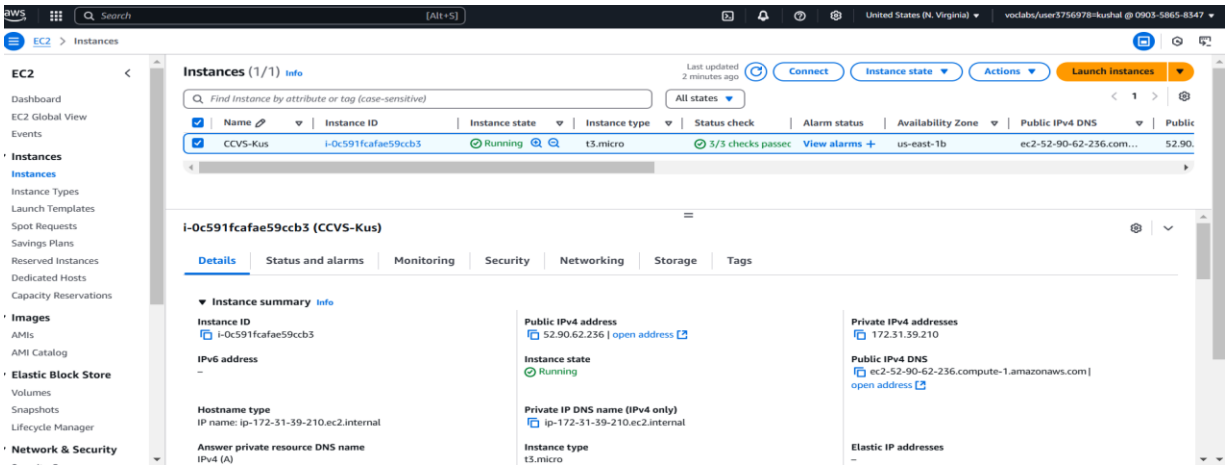
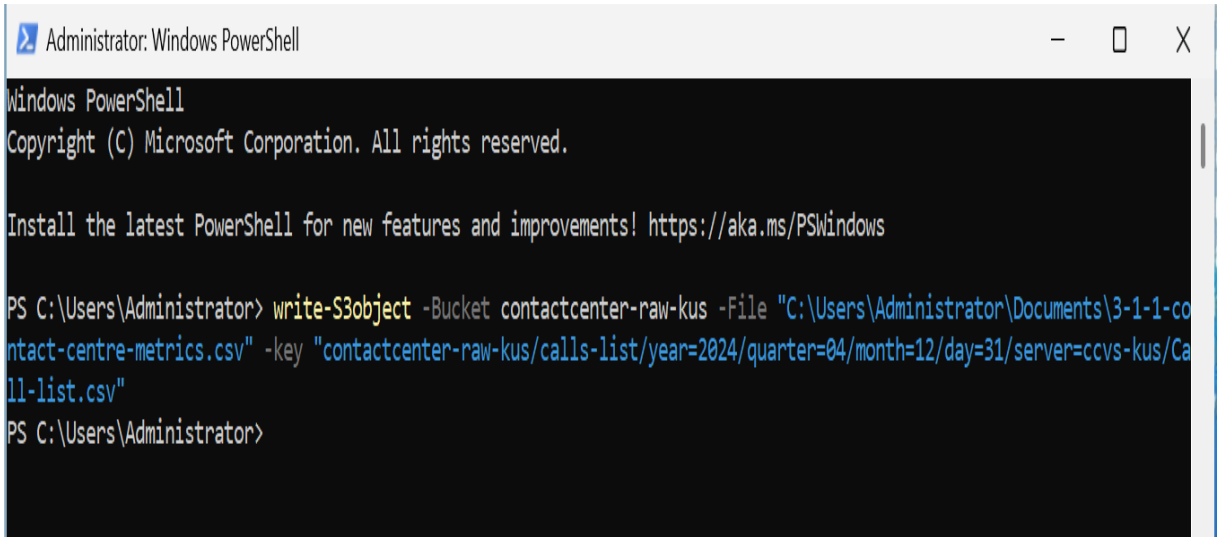


Figure 2: PowerShell Command to Upload CSV to S3

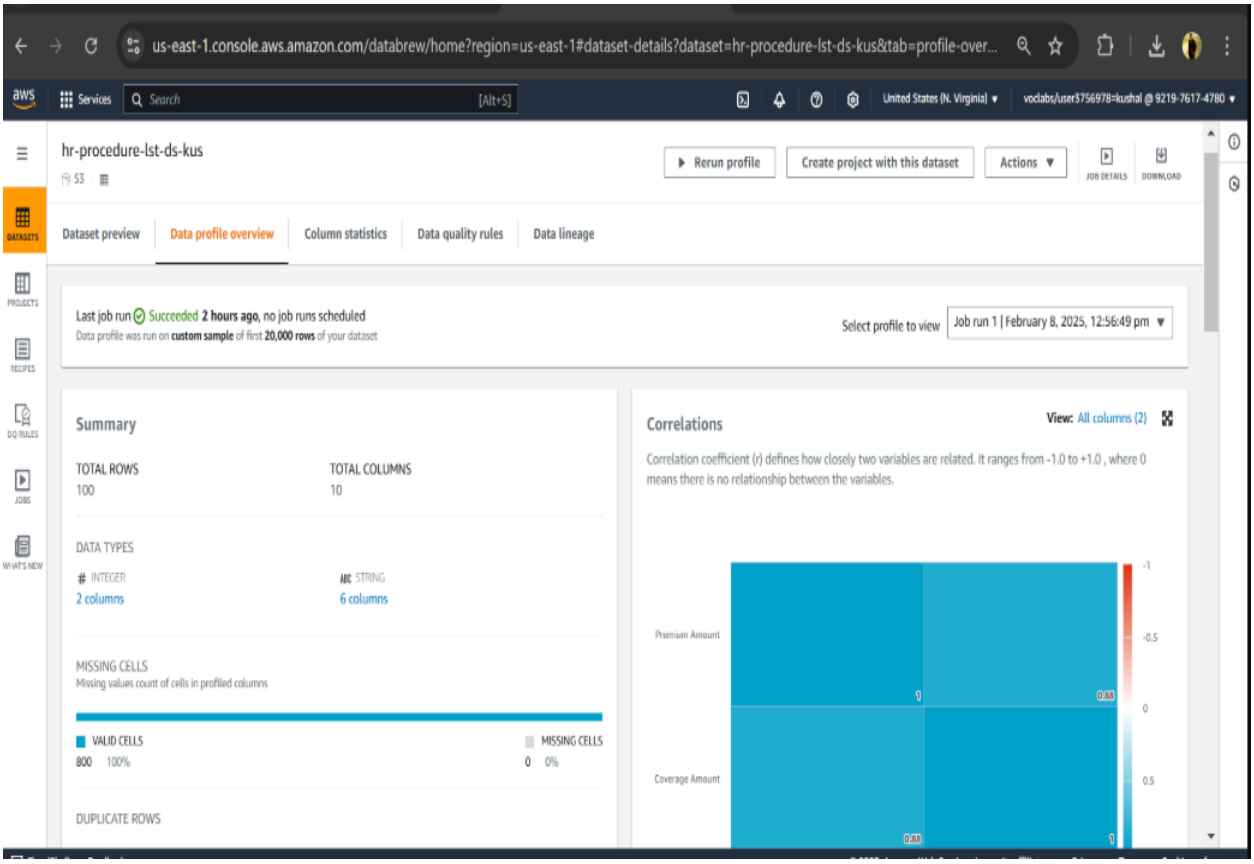


Through this approach FreshMart simulates the process that enables its branches to automatically transfer daily transaction reports such as 'transaction_data.csv' into its cloud storage platform. The established structure enables real-time record intake while creating data readiness for upcoming partitioning operations and consequent service-based analysis through Athena.

The described steps construct an automation framework that offers scalability along with reliability and security for cloud-scale data ingestion services. This PowerShell command screenshot shows step-by-step how the dataset send to S3 from an EC2 instance. The FreshMart system implements daily file upload procedures through automated or scheduled processes to maintain secure storage of transaction_data.csv and similar files. The structure of this command duplicates typical retail business operations which transfer structured transaction data from server systems and local sources to cloud storage platforms for analysis purposes. The ingestion phase enables a controlled data consolidation in which stored information remains secure for upcoming processing steps.

Data Profiling: The next step involved the utilization of AWS Glue DataBrew for creating a profiling job after data ingestion. The tool performed an analysis of metrics at the column level by evaluating data types with null values and value distribution patterns. The profiling job identified quality issues during an early stage which guided the implementation of cleaning processes.

Figure 3: Data Profiling Overview in AWS Glue DataBrew

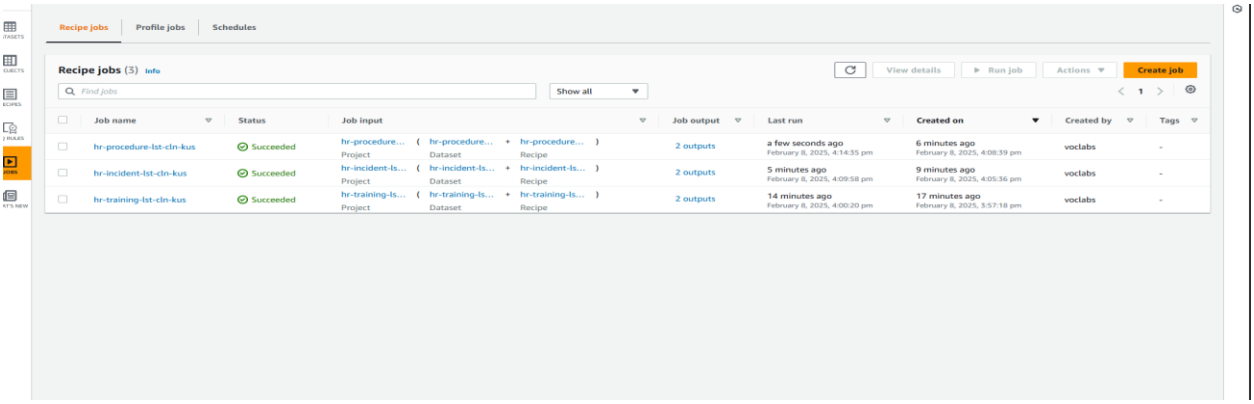


An evaluation of FreshMart transaction fields used this data quality report for assessment purposes. Analyzing data through profiling work reveals columns containing both irregular values together with missing data points so analysts can properly clean it up.

Data Cleaning: The data cleaning process operated through the visual interface of AWS Glue Studio. The ETL job cleaned data by removing duplicates while reformatting dates and filtering null values alongside command standards. Two output versions were generated:

- Partitioned files for efficient querying
- A consolidated CSV for manual review or user access

Figure 4: Recipe Jobs Executed in AWS Glue DataBrew



The screenshot shows how three AWS DataBrew data cleaning jobs finished their operations successfully. The jobs contain distinct datasets that applied predefined recipes during the cleaning process. FreshMart demonstrates different sales datasets cleaning operations through this example where protocols remove null values while standardizing types and performing formatting adjustments on transaction records and product category lists independently. The 'Succeeded' verification indicates correct data processing that enables the following use for analysis or transformation purposes. Products and payment methods with inconsistent information undergo standardization through this FreshMart job to avoid improper transfers to successive stages of the pipeline. The visual ETL logic specifies an identical structure for every transaction which supports precise reporting downstream.

Data Cataloging:

A Glue Crawler from AWS performed a scan on the cleaned data which produced an automatic structure for a table in the Glue Data Catalog. The database named freshmart-data-catalog stored the table with details about column names, data types and storage references such as Parquet**.**

Figure 5: AWS Glue Crawler and Data Catalog Setup

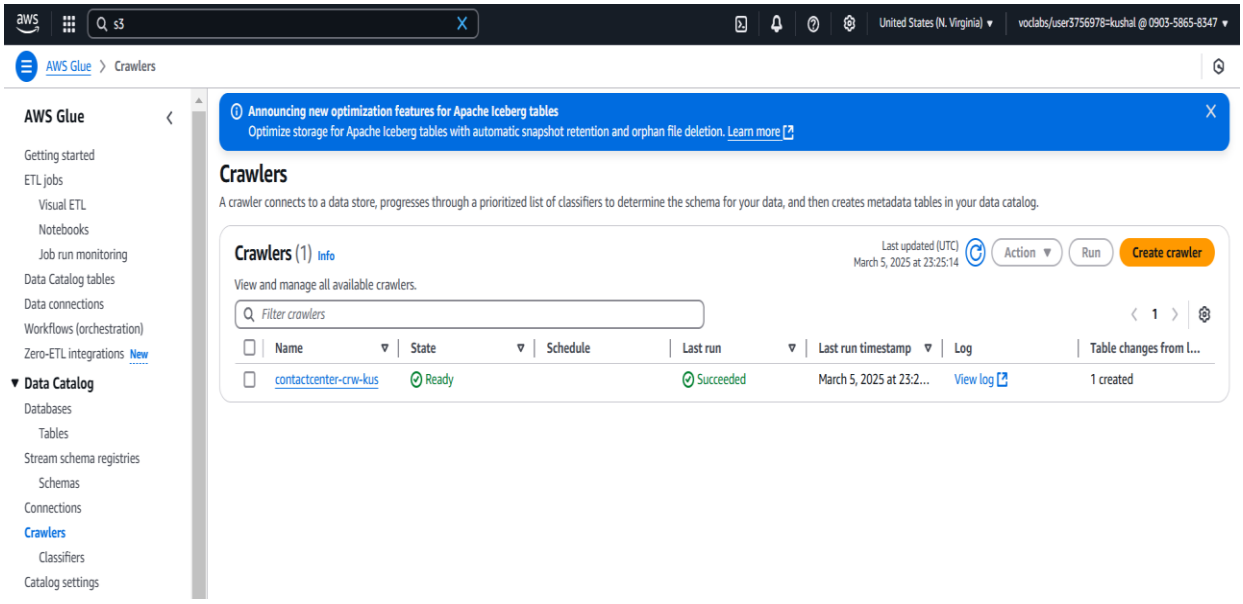
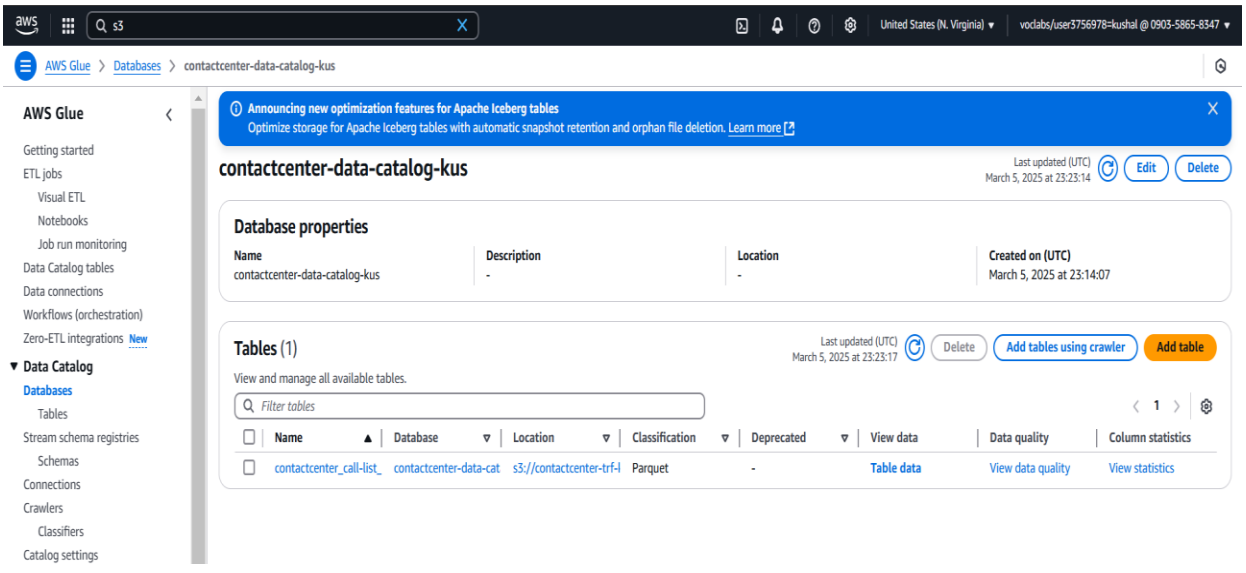


Figure 6: Metadata Schema in AWS Glue Data Catalog



The two displayed screenshots demonstrate the main procedures of data cataloging for FreshMart. An AWS Glue Crawler interface appears in the screenshot displaying the status of contactcenter-crw-kus crawler which successfully identified metadata during its dataset scan. The crawler should be configured to examine the freshmart-trf-kus cleaned S3 path like any other dataset.

The crawler processed data leads to a structured table named contactcenter_call-list which resides in a Glue database according to the second screenshot. For FreshMart's database the table should be called freshmart_transaction_data to document information about product category, quantity, and price as well as store location. The data becomes easily queryable through Amazon Athena and provides discoverability features and governance while allowing version control of schemas. The depicted screenshot shows how FreshMart data schemas become automatically cataloged which provides analysts access to database information through Athena and additional business tools.

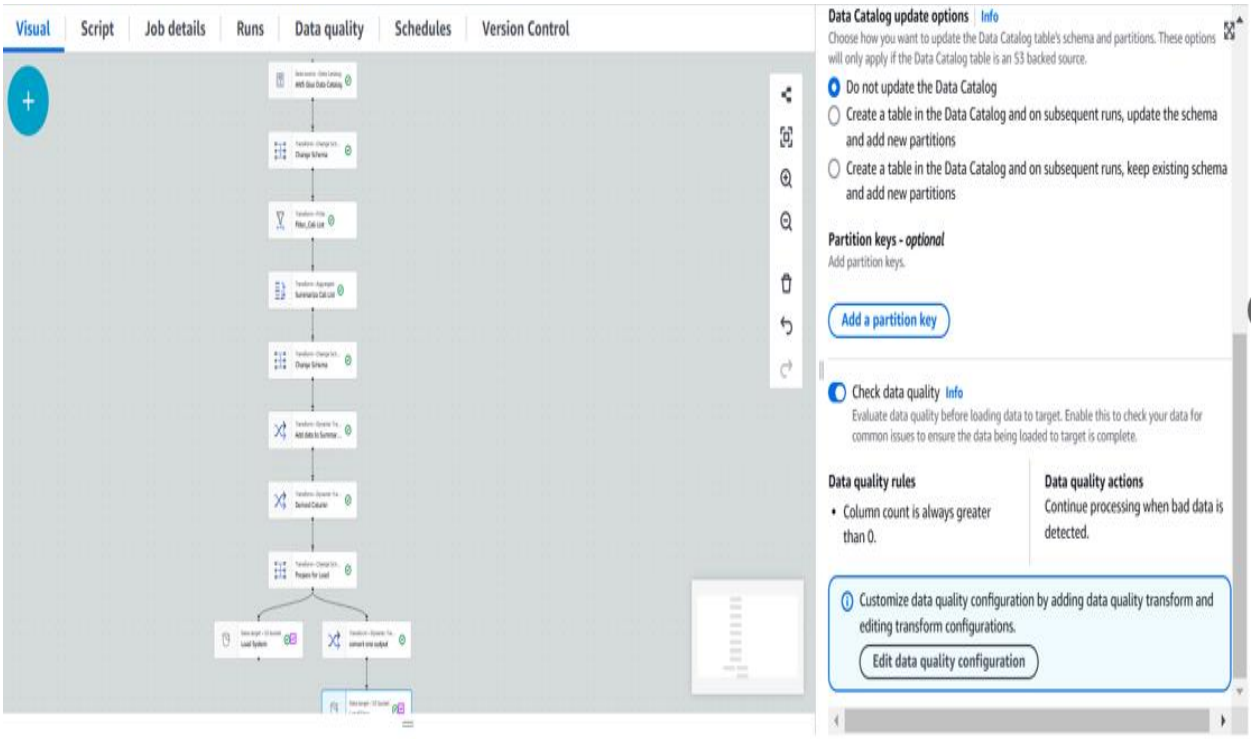
Using another Glue Studio ETL job, the project summarized key metrics from FreshMart's transactional dataset.

Data Summarization:

Grouping and aggregation functions helped generate meaningful insights such as:

- Monthly sales trends
- Top product categories
- Customer frequency segments
- Average order values

Figure 7: Data Summarization Workflow in AWS Glue Studio



This screen presentation depicts FreshMart's key business metric summary table generation through the ETL transformation pipeline. Use of grouping and aggregation nodes appears in this depiction within AWS Glue Studio.

FreshMart used this job configuration to combine transactions into monthly groups by specific store locations and product categories while computing total sales figures and average quantitative data. The concise output shows managers the ways customers purchase products through time-based and regional comparisons. Experimenters can access pre-aggregated data through Athena or Power BI after it is stored in S3.

Conversion of operational transactional logs through summarizing produces refined information which FreshMart's decision making personnel use to optimize their sales models and inventory management alongside category-focused promotional strategies.

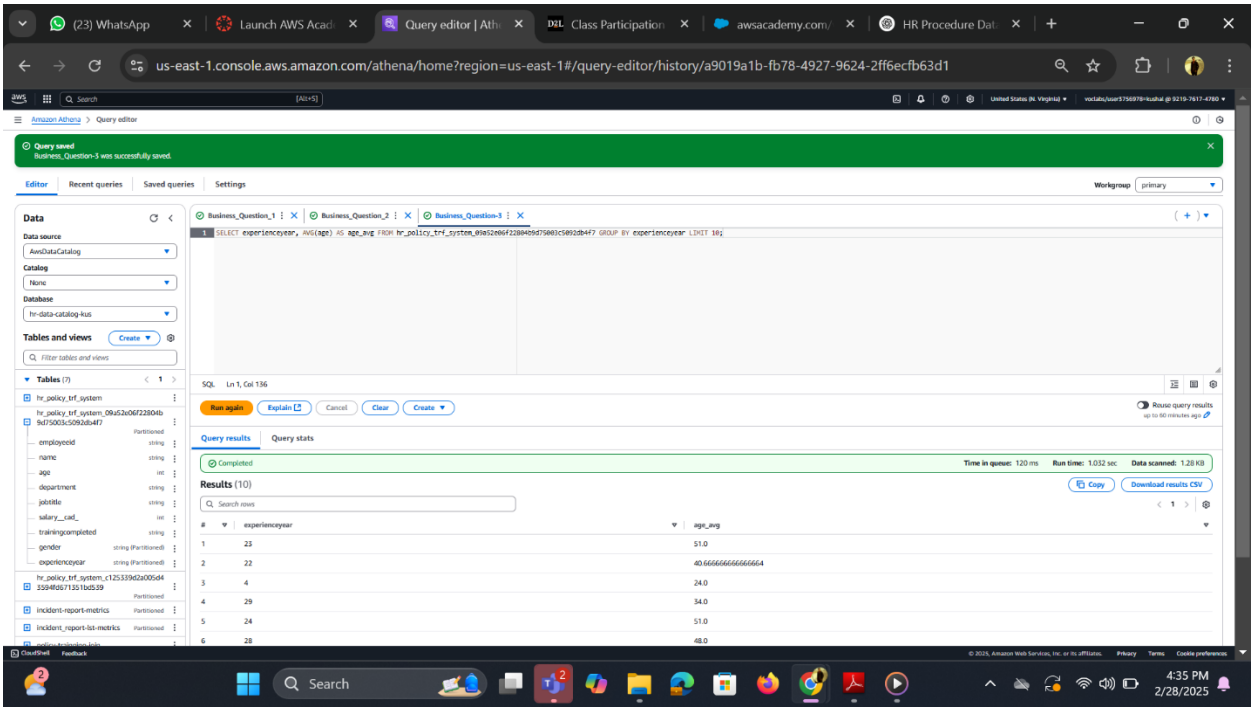
Data Analysis

The summarized dataset located in Amazon S3 was evaluated by Amazon Athena as part of the query process. Users can execute SQL queries against prepared structured data through Athena which functions without needing customers to establish infrastructure. The aggregation of metrics at FreshMart requires this phase to become actionable business intelligence.

The illustration reveals SQL query examples from a previous project that investigate both customer interaction patterns and sales results. FreshMart deployed SQL queries which resembled ours along with such statements:

- SELECT product category, SUM(total sales) FROM transactions GROUP BY product category; → To identify top-selling product categories.
- SELECT store location, AVG(total price) FROM transactions GROUP BY store_location; → To find which store locations generate the highest average revenue per transaction.
- SELECT payment_method, COUNT(*) FROM transactions GROUP BY payment_method; → To analyze customer preferences in payment types.

Figure 8: Athena query panel showing SQL query and grouped results output



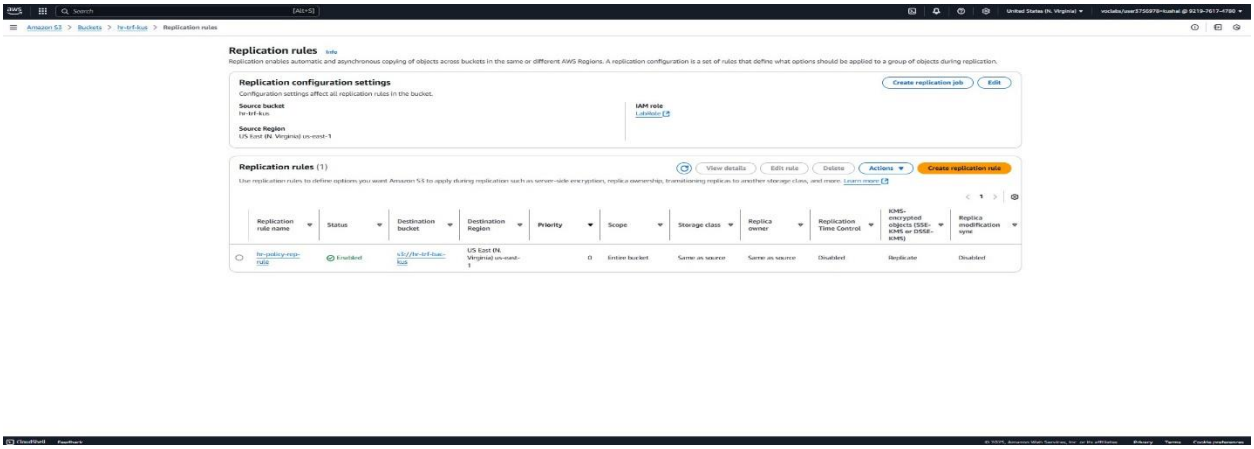
This picture demonstrate an Amazon Athena which executes the SQL query that displays customer data grouping and average calculation functions for FreshMart purchase analysis. The data shown in this image pertains to worker experience although the evaluation technique remains the same.

The FreshMart team used comparable SQL queries to find out which product categories obtain the most popularity together with average values for store transactions and customer-favored payment methods. The interface demonstrates query success and visualizes results within Athena which confirms the adequate functionality of SQL functions AVG(), COUNT() and GROUP BY. Using this method FreshMart analysts achieve quick data extraction from retail

data summary stored in S3. SQL code can be recycled to create dashboards or reports within BI solutions.

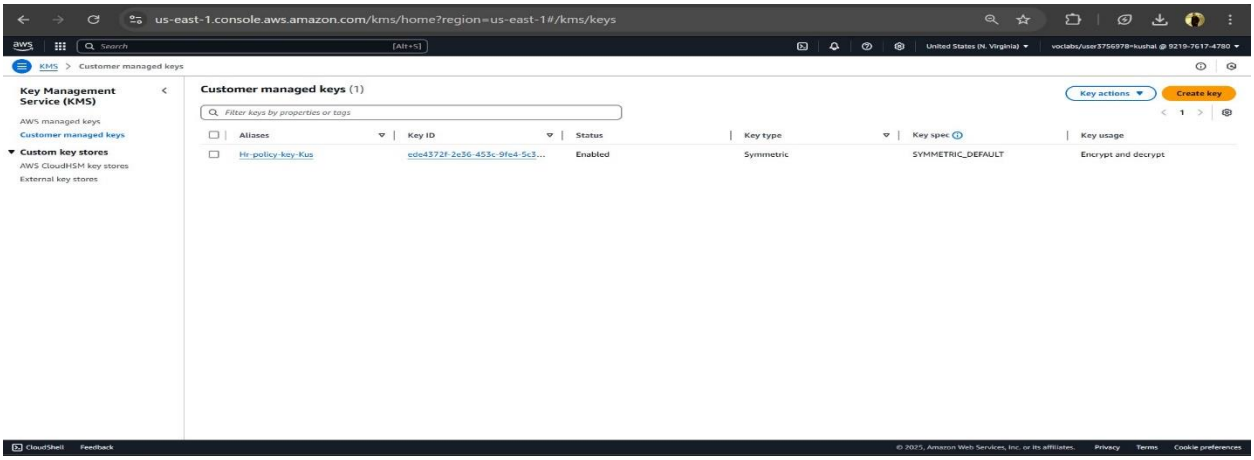
Data Security: A protection system through server-side encryption was activated with AWS Key Management Service (KMS) for customer data safety. An AWS Key Management Service (KMS) custom key protected access to the raw S3 bucket. The implementation of versioning enabled the system to retrieve every file change and involved establishing a backup bucket for cross-region file replication.

Figure 9: S3 Replication rule from source to destination bucket



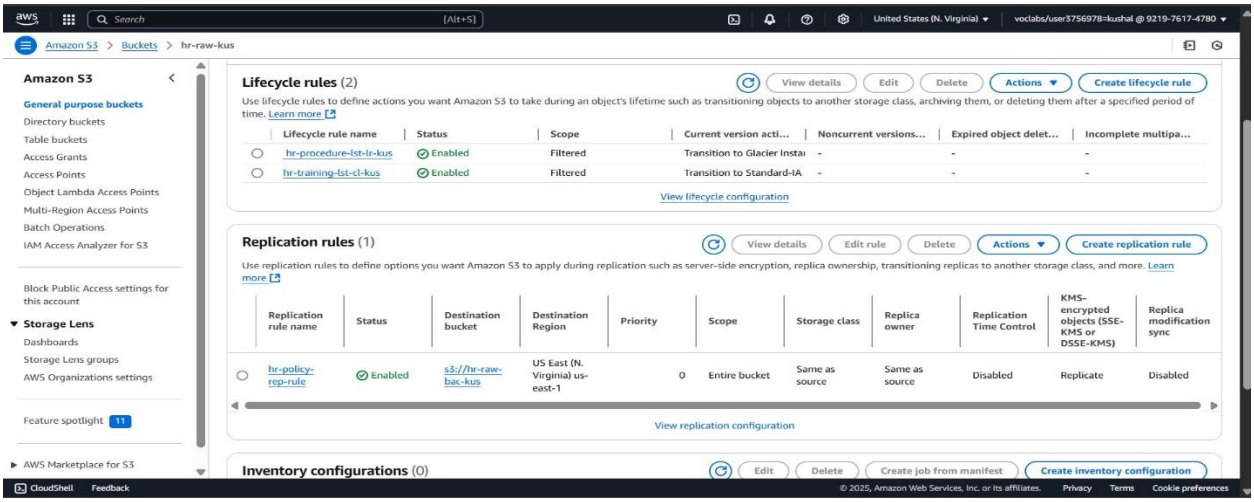
The displayed screen reveals an operational replication rule which duplicate files from the FreshMart main bucket to backup storage. Data backup to the secondary bucket functions as a protection mechanism for preventing permanent loss of information from the main storage.

Figure 10: AWS KMS with customer-managed key



A customer-managed symmetric key exists in this screenshot which secures FreshMart's data. This key serves the same data protection purpose as FreshMart utilizes to secure its sales data in S3 storage. The system applies this key to encrypt all incoming files to boost confidentiality and satisfy relevant compliance requirements.

Figure 11: PS3 versioning and lifecycle rules



The displayed screenshot demonstrates a customer-managed symmetric key which FreshMart uses for securing its confidential data. The structure displayed reflects how FreshMart safeguards sales data in the S3 database although it was designed for a different purpose. The system applies file encryption to all uploaded content through this encryption key to boost confidentiality compliance.

Data governance

Data governance at FreshMart received improvement through the use of AWS Glue Data Quality and Routing features which delivered quality data to support decision-making processes.

The data quality rules added to the Glue visual ETL pipeline by using its interface prevented the final dataset from receiving incomplete or invalid transaction records. The records which do not pass the quality checks get transferred to an independent review station.

Key Validation Rules Implemented:

- Completeness: Ensured that mandatory fields (e.g., transaction_id, product_category, store_location) are not null.
- Uniqueness: Checked for duplicate transaction_id values to avoid double-counting.
- Accuracy **& Freshness**: Ensured purchase date is within the expected time range and formatted correctly.

Figure 12: AWS Glue Job Run Dashboard

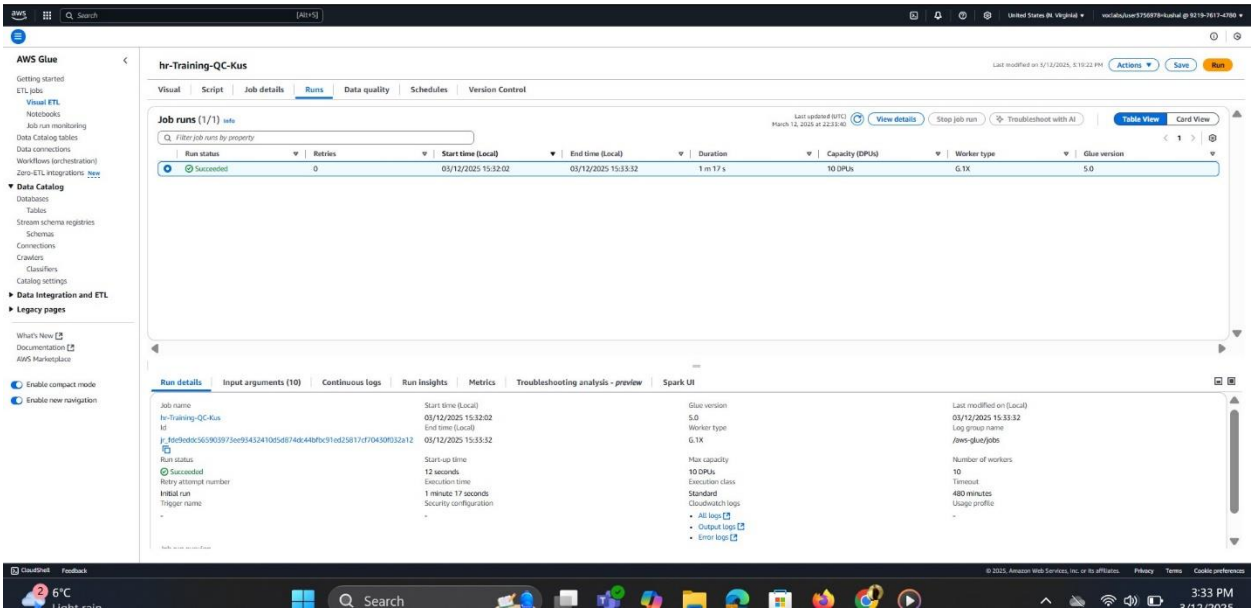
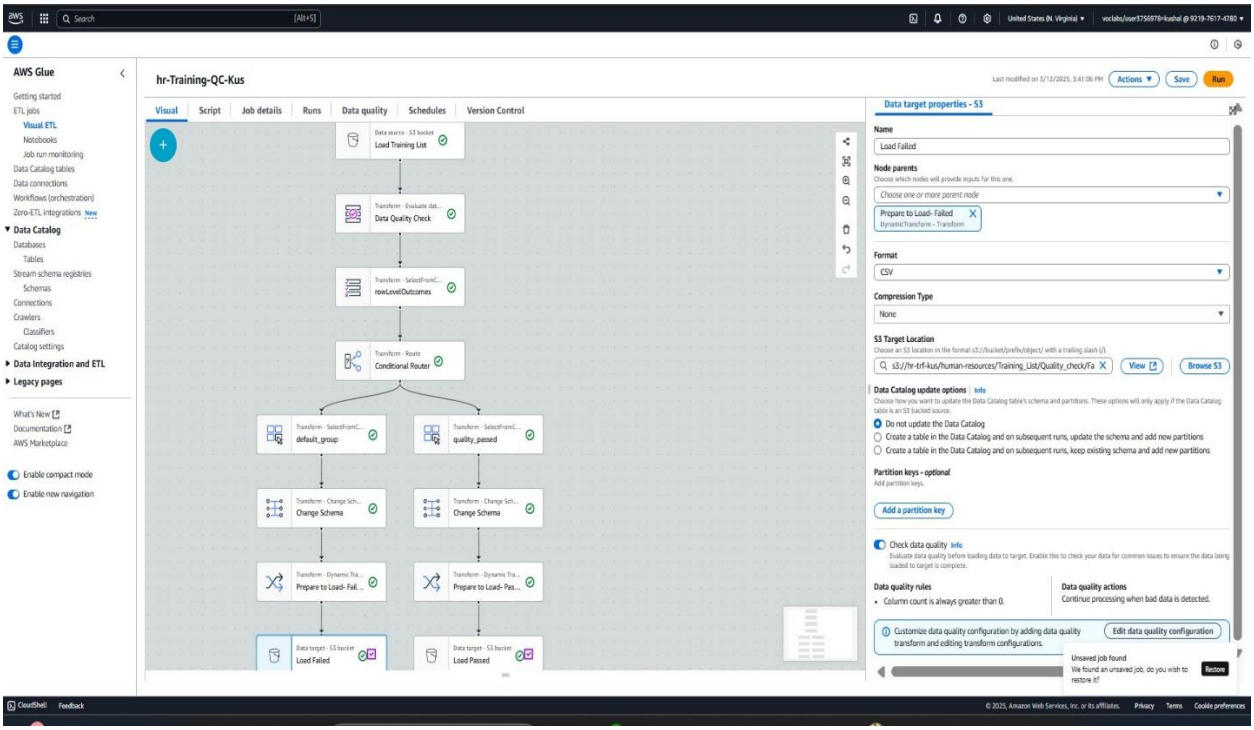


*Figure 13: Visual ETL Flow with conditional routing for **passed** and **Failed** records*



The top image demonstrates successful job execution which proves FreshMart ran its governance rules properly. The Glue ETL visual flow serves as the second display to show how records undergo evaluation through data quality rules. Records following successful validation flow toward the Load Passed S3 bucket yet records failing one of the validation tests end up in the Load Failed bucket.

This governance strategy ensures:

- ✓ The system accepts data for analytics and dashboards only when it fulfills both cleanliness and reliability requirements.
- ✓ Failed records get individual separate storage which serves for future checks or editing purposes.
- ✓ Conflict-free data compliance operates without requiring personnel to inspect extensive datasets.

Data Monitoring

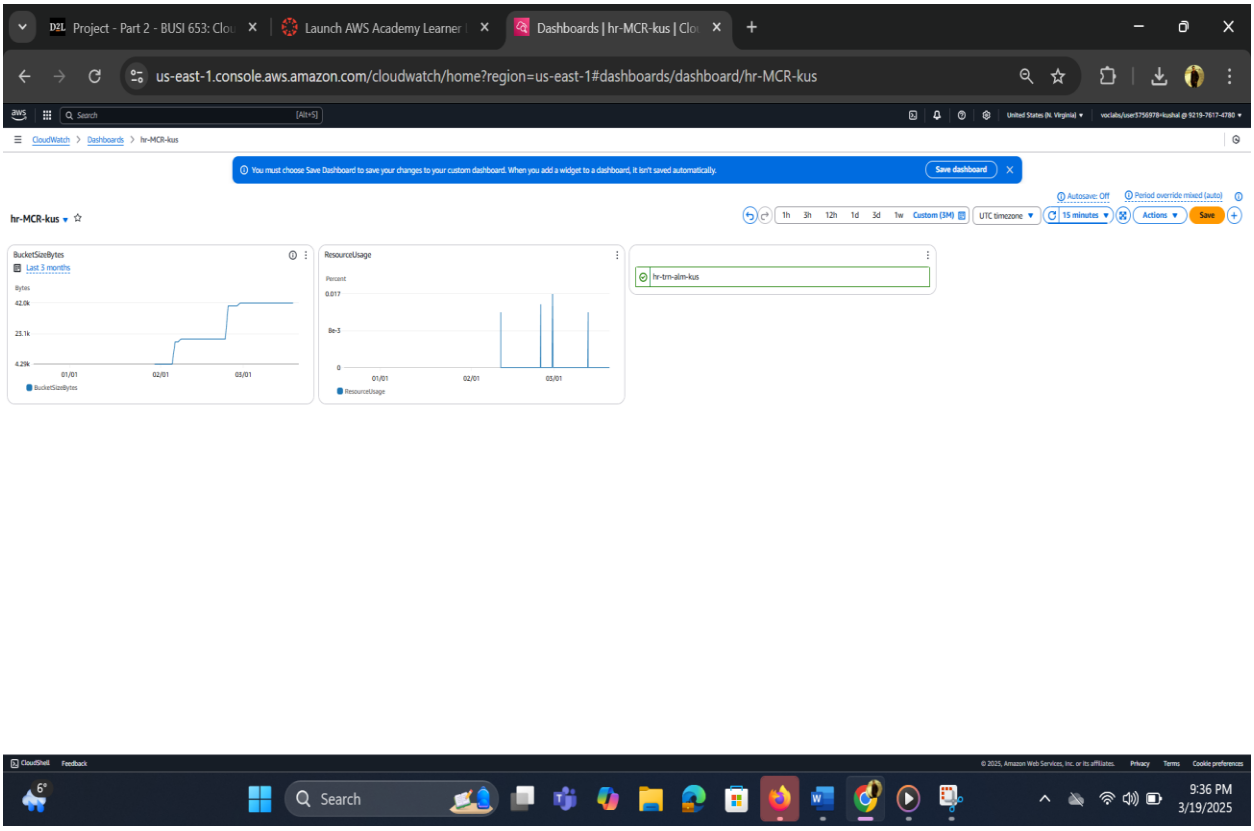
The data pipeline management at FreshMart depends crucially on monitoring functions that verify correct operation of all processes from data ingestion through analysis. The goal of achieving monitoring was reached through the implementation of powerful AWS services Amazon CloudWatch and AWS CloudTrail.

Amazon CloudWatch

Used to visualize and monitor:

- Data pipeline performance
- Storage usage trends
- Job run success and failures

Figure 14: CloudWatch dashboard (bucket size and resource usage tracking)



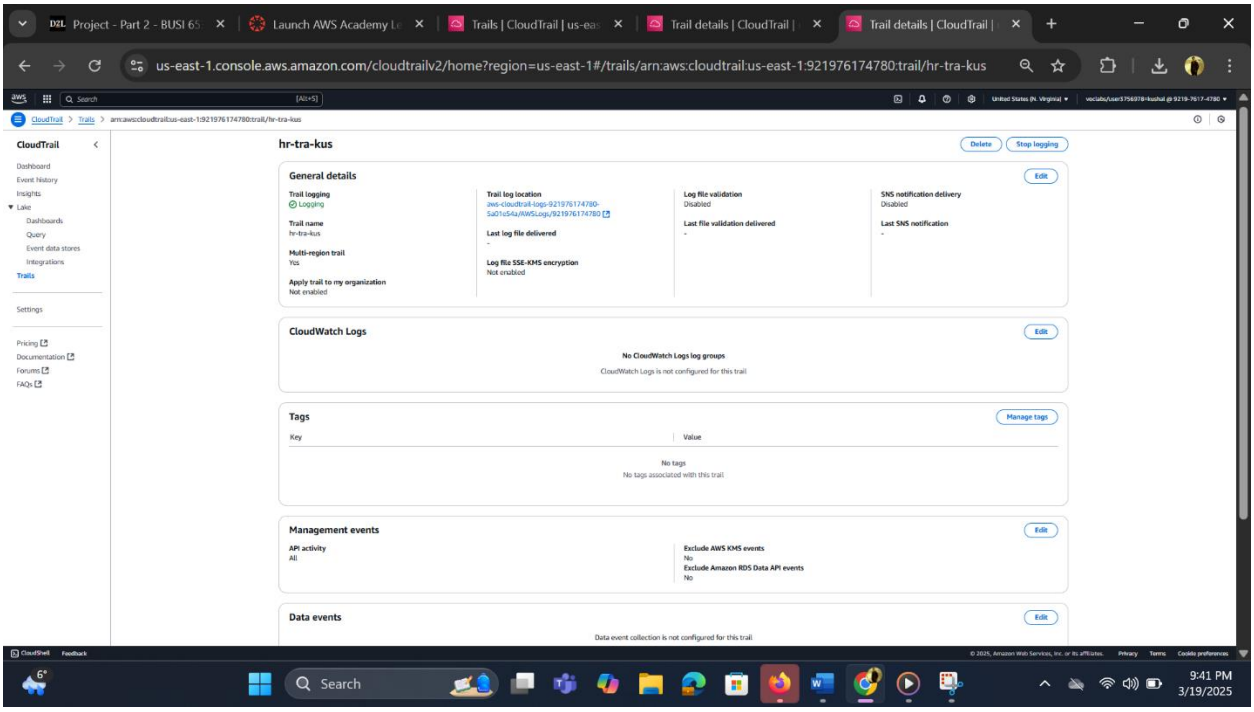
The picture demonstrates CloudWatch services tracking which FreshMart S3 storage bucket size expansion while also monitoring ETL job resource consumption. Monitoring storage capacity and ETL resource consumption by the data team through weekly transaction data additions is made possible by CloudWatch.

2. AWS CloudTrail

Used for:

- Tracking API calls across all AWS services
- Auditing access to data resources
- Ensuring regulatory compliance

Figure 15: CloudTrail trail configuration



The shown snapshot demonstrates the hr-tra-kus management event trail that actively logs all activities in support of user functions and Glue job executions and S3 bucket access. The implementation method follows precisely the same procedure for FreshMart despite the original purpose to handle different data. This system enables complete environment tracking through its log feature which records user activities with timestamps across AWS infrastructure.

Tools and Technologies Used:

To implement and manage this cloud-based analytical pipeline, the following AWS tools and third-party platforms were used:

- **Amazon S3**: For centralized cloud-based storage of raw and transformed datasets
- **Amazon EC2**: For simulating remote uploads to S3 using PowerShell
- **AWS Glue Studio**: For designing ETL workflows for cleaning, summarizing, and validating data
- **AWS Glue Data Brew**: For data profiling and quick visual recipe jobs
- **AWS Glue Crawler and Data Catalog**: To detect and define metadata schemas from S3-stored files
- **Amazon Athena**: For server less querying of the transactional data using SQL
- **AWS CloudWatch**: For monitoring job performance and S3 usage
- **AWS Cloud Trail**: For logging user activity and API calls across AWS services
- **AWS Key Management Service (KMS)**: For enabling encryption of sensitive data stored in S3

Conclusion

The FreshMart descriptive analysis project proved that Amazon Web Services effectively manages and analyzes extensive retail data. Secure cloud technologies were used to construct all pipeline elements after ingestion and before monitoring. The summarized data enabled FreshMart to gain several benefits.

- ✓ Identify high-performing product categories
- ✓ The business needs to evaluate transaction patterns through location analysis of its stores.
- ✓ Understand preferred customer payment methods
- ✓ Data compliance can be assured by implementing best practices related to security and governing data processes.

Through their combination with AWS Glue, Athena, and S3 the transformation processes became streamlined and querying became efficient while CloudWatch and CloudTrail provided powerful monitoring and auditing capabilities. The cloud-native framework produces an analytics environment that enables FreshMart to expand operations smoothly without any issues while creating a solid foundation for market uncertainty.

References

Amazon Web Services. (2024). *Cloud Object Storage | Store & Retrieve Data Anywhere | Amazon Simple Storage Service*. Amazon Web Services, Inc. https://aws.amazon.com/s3/

AWS. (2024a). *Amazon EC2*. Amazon Web Services, Inc. https://aws.amazon.com/ec2/

AWS. (2024b). *What is Amazon CloudWatch? - Amazon CloudWatch*. Amazon.com. https://docs.aws.amazon.com/AmazonCloudWatch/latest/monitoring/WhatIsCloudWatch.html

AWS. (2024c). *What Is AWS CloudTrail? - AWS CloudTrail*. Docs.aws.amazon.com. https://docs.aws.amazon.com/awscloudtrail/latest/userguide/cloudtrail-user-guide.html

AWS. (2024d). *What Is AWS Key Management Service? - AWS Key Management Service*. Docs.aws.amazon.com. https://docs.aws.amazon.com/kms/latest/developerguide/overview.html

Freshmart® | Your neighbourhood grocer, here for you. (n.d.). Www.freshmart.ca. https://www.freshmart.ca/

Get started - Amazon Athena. (2025). Amazon.com. https://docs.aws.amazon.com/athena/latest/ug/getting-started.html

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