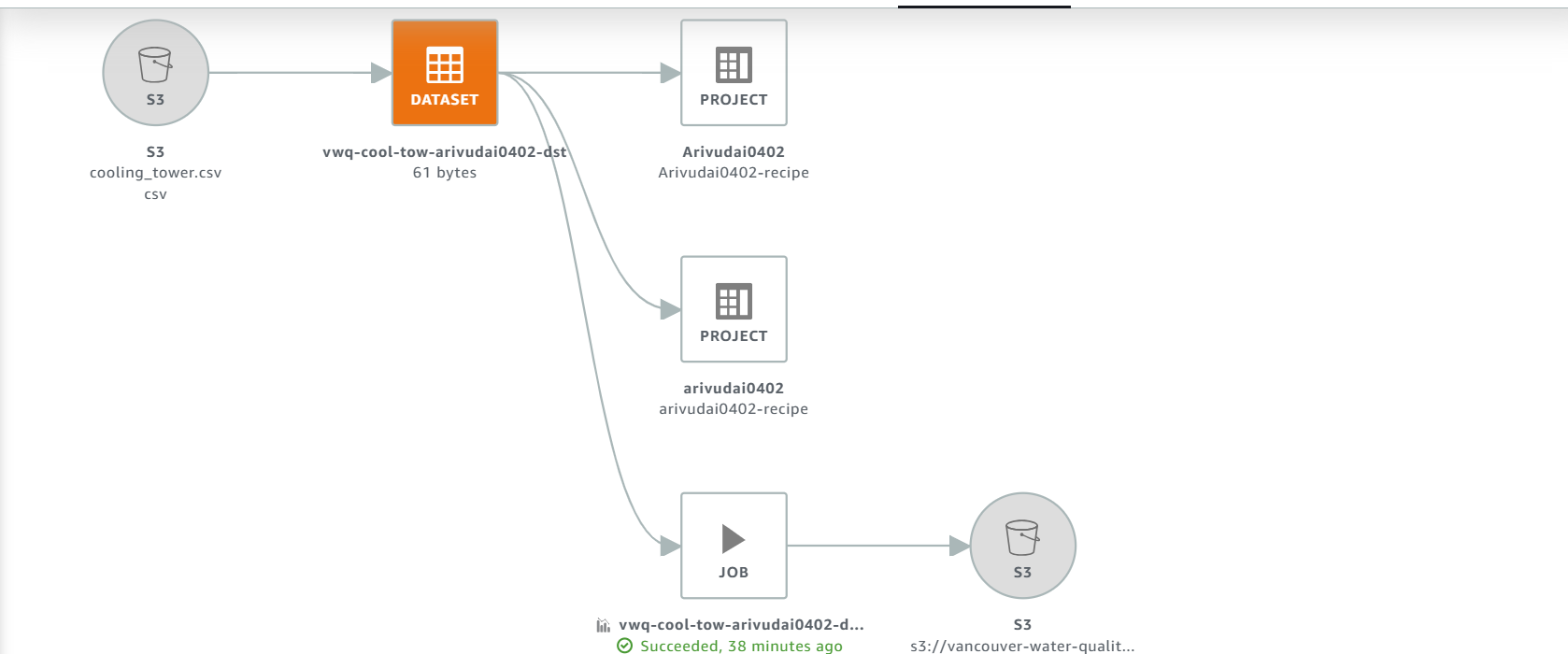
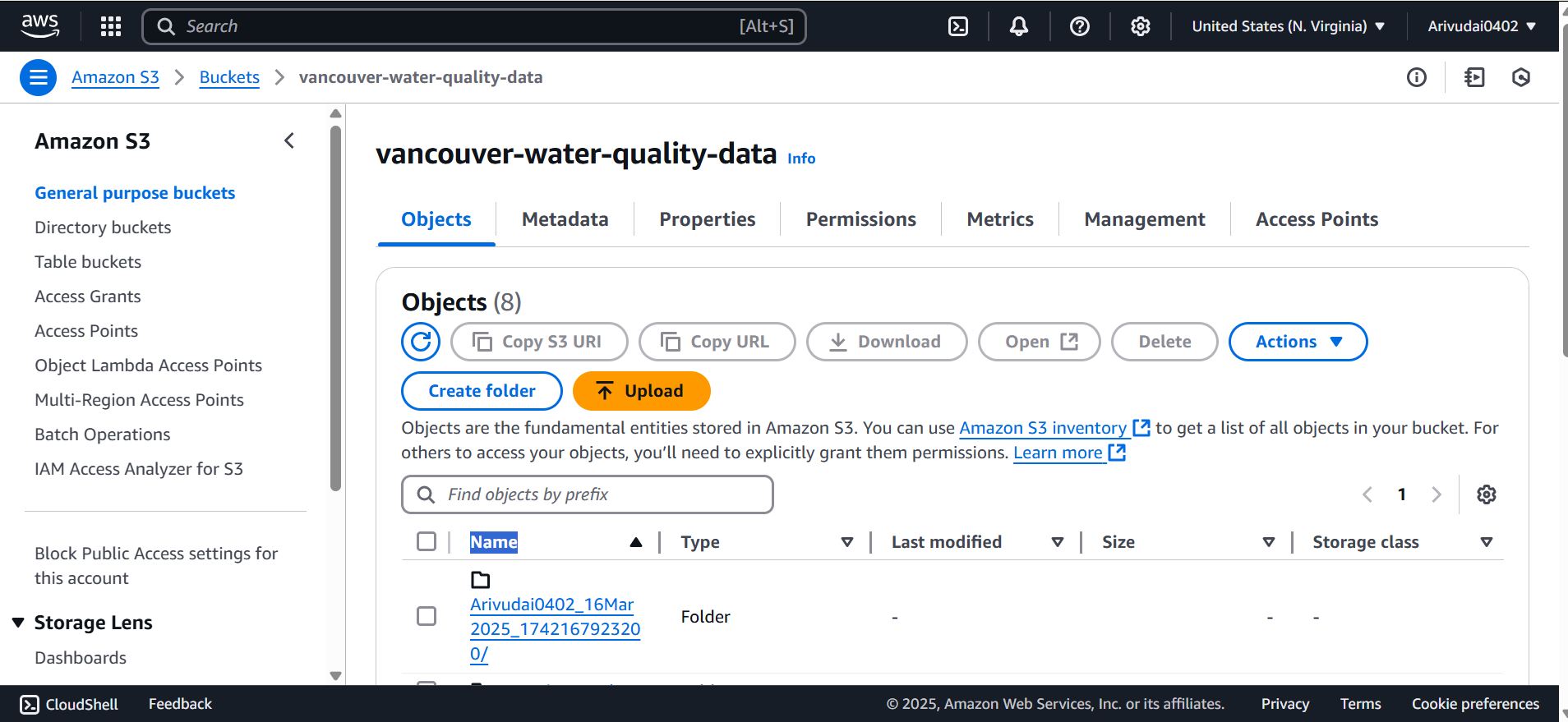
**Project Title: AWS Data Analytics Platform Migration (City of Vancouver)**

**Designing the data flow**



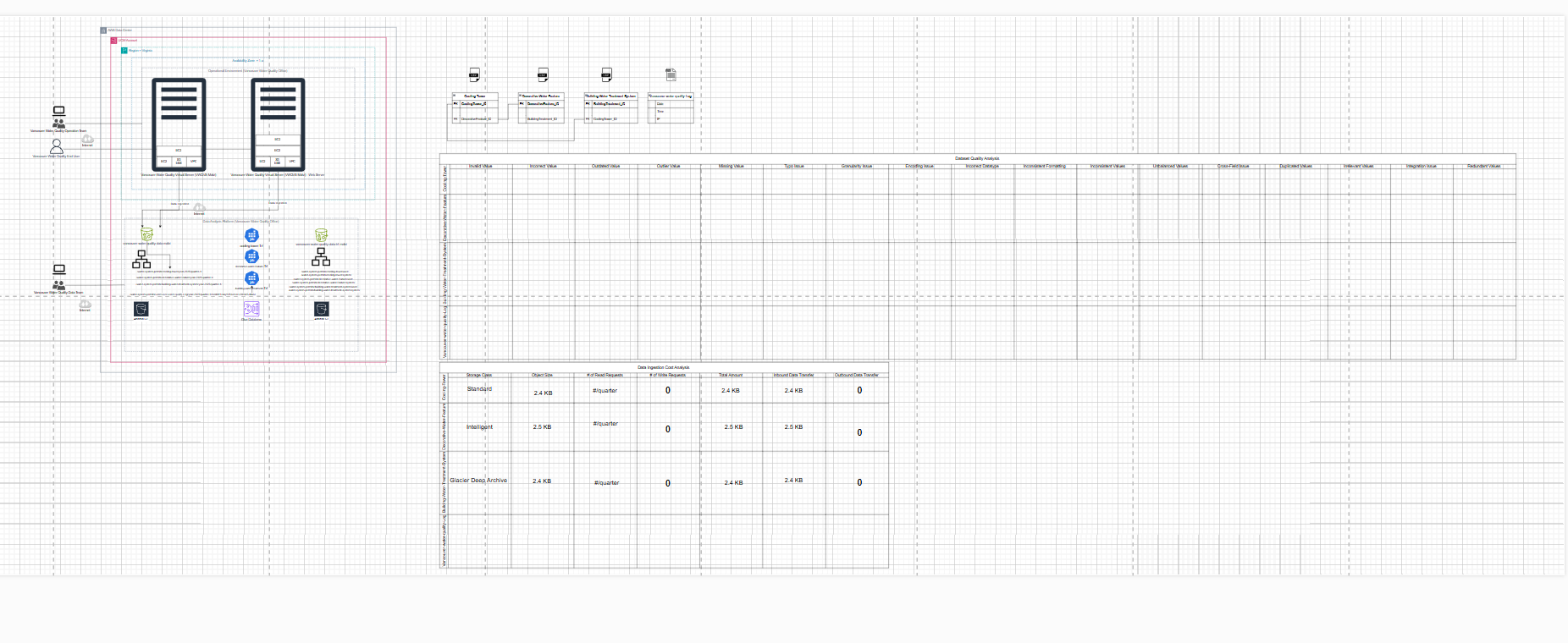
**Setting up the S3 Bucket**



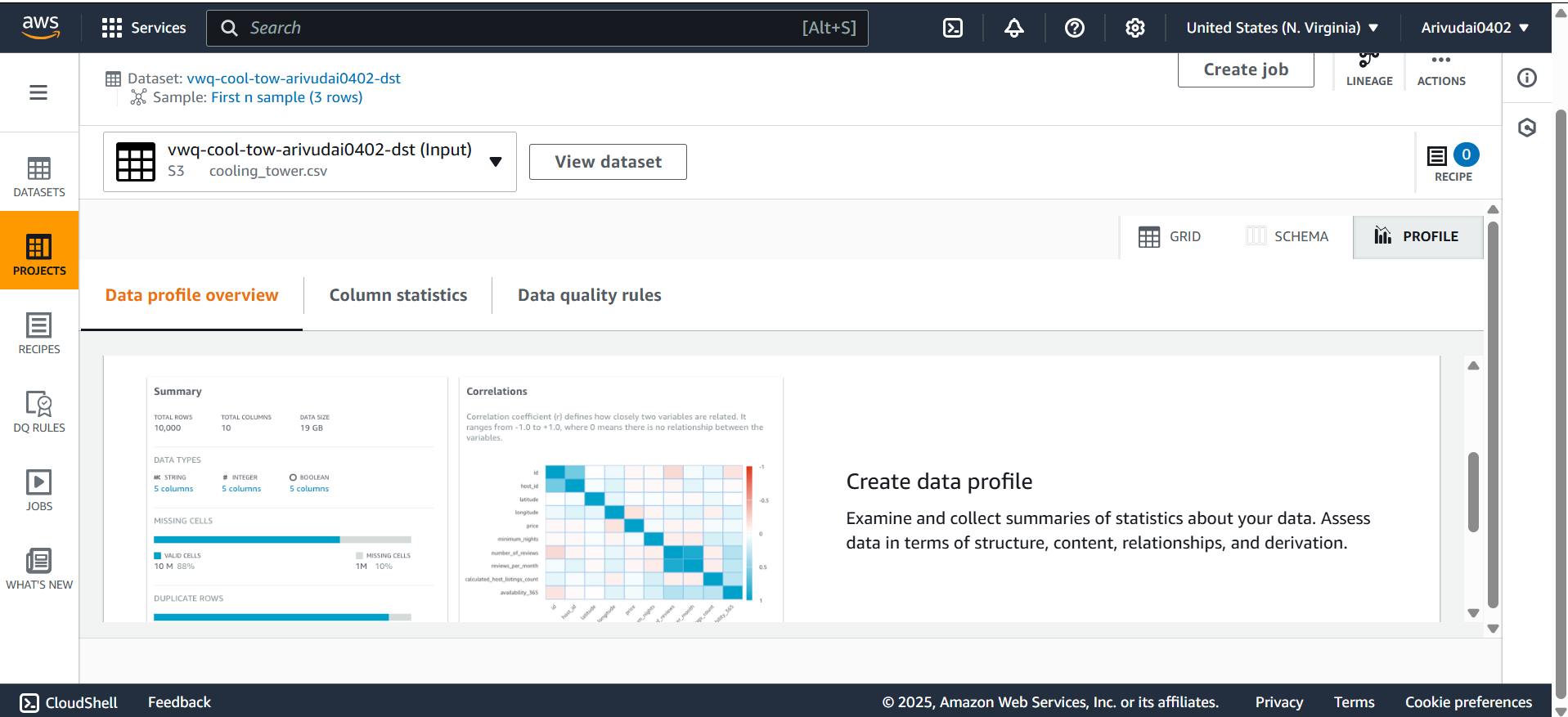
To serve as a central repository for all raw data files, I set up an Amazon S3 bucket.

I transferred data from my local system to Amazon S3 using PowerShell scripts. This enables data to be kept in the cloud for further processing and analysis. Data Verification for Ingestion: I checked the data within the AWS Management Console following the upload so that the files actually got stored correctly in the proper S3 bucket locations.

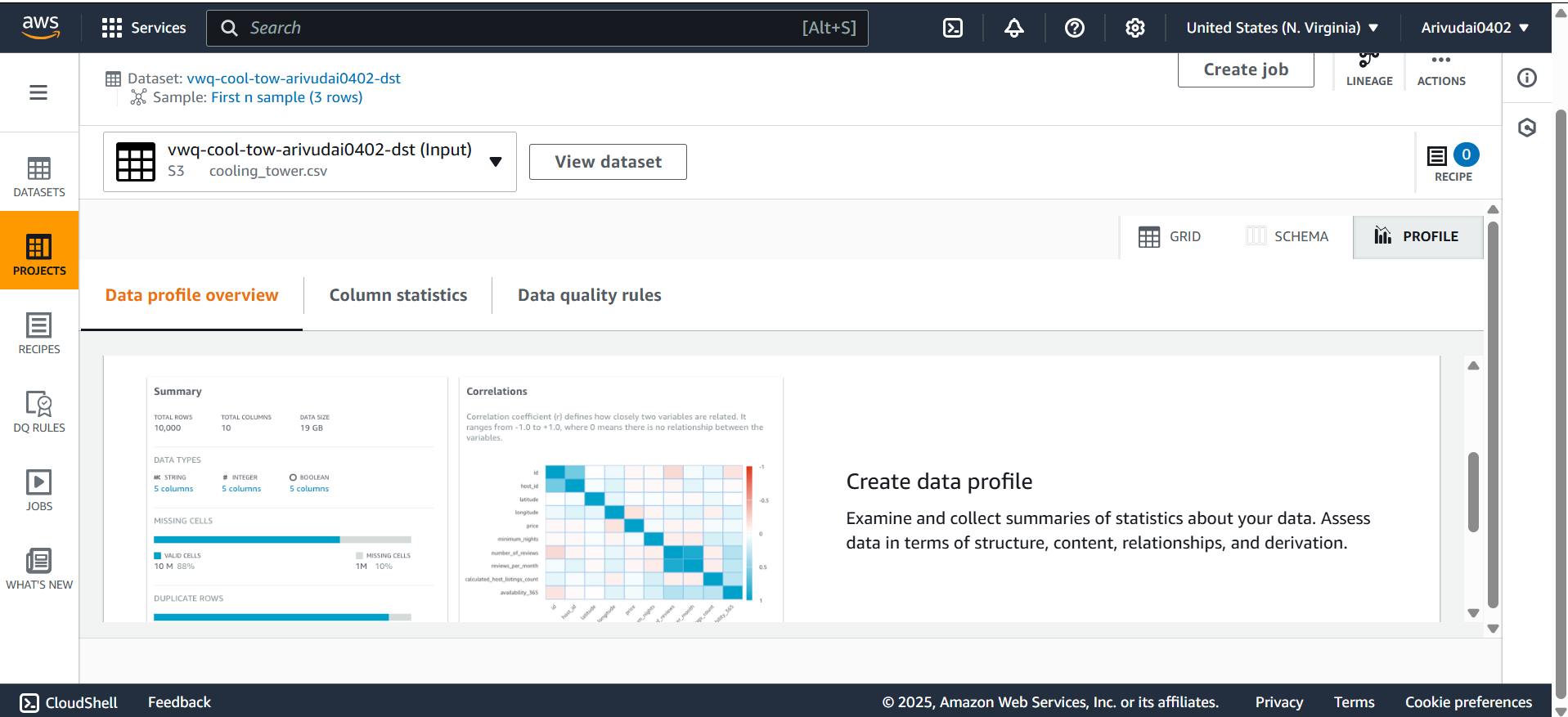
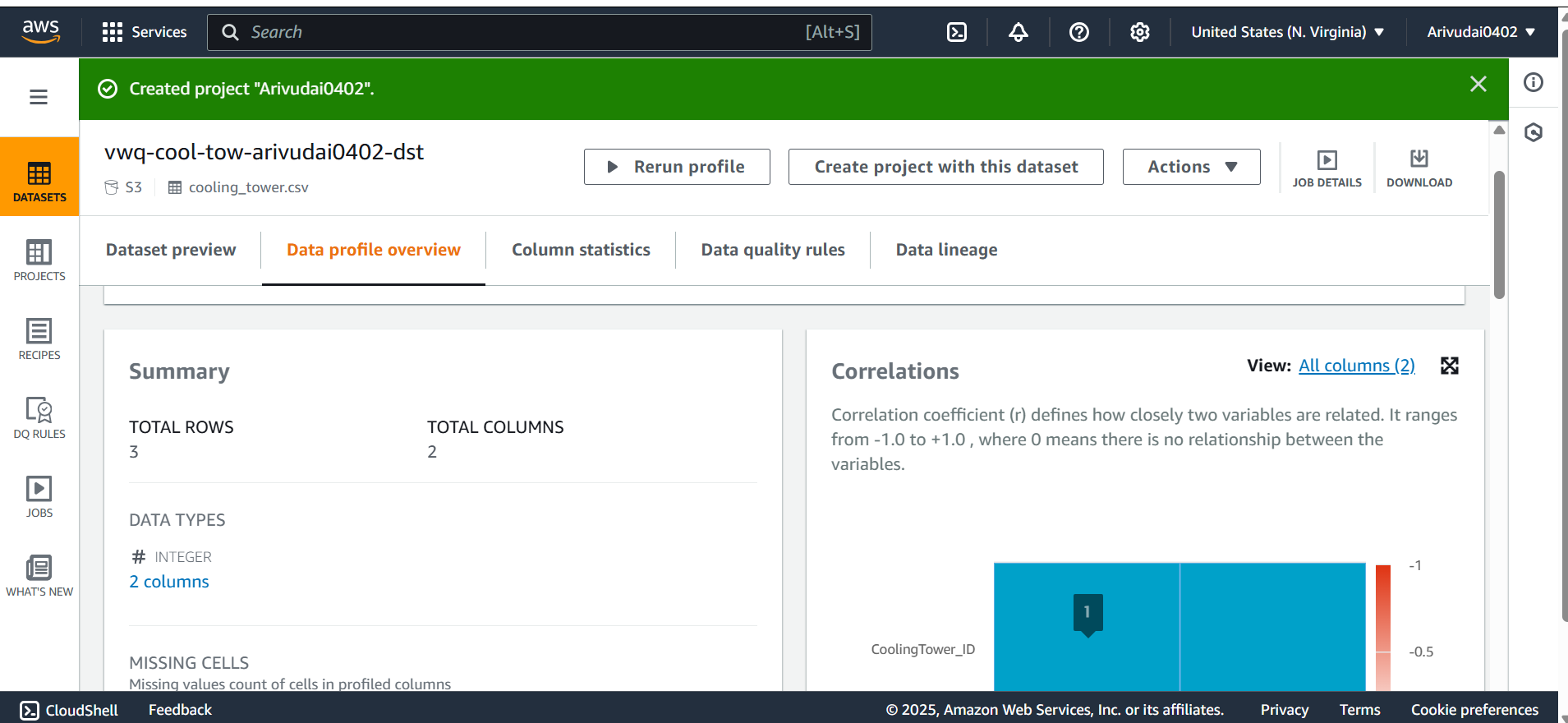
# Data profiling



I profiled and cleaned the water quality dataset using AWS Glue DataBrew to make it properly organized and ready for further analysis. The dataset, divided into sections such as cooling towers, decorative water features, and building water treatment facilities, was kept in an S3 bucket. An overview of data ingestion and processing pipeline is given by the Draw.io diagram that also describes storage, analysis, and transformation of various datasets.

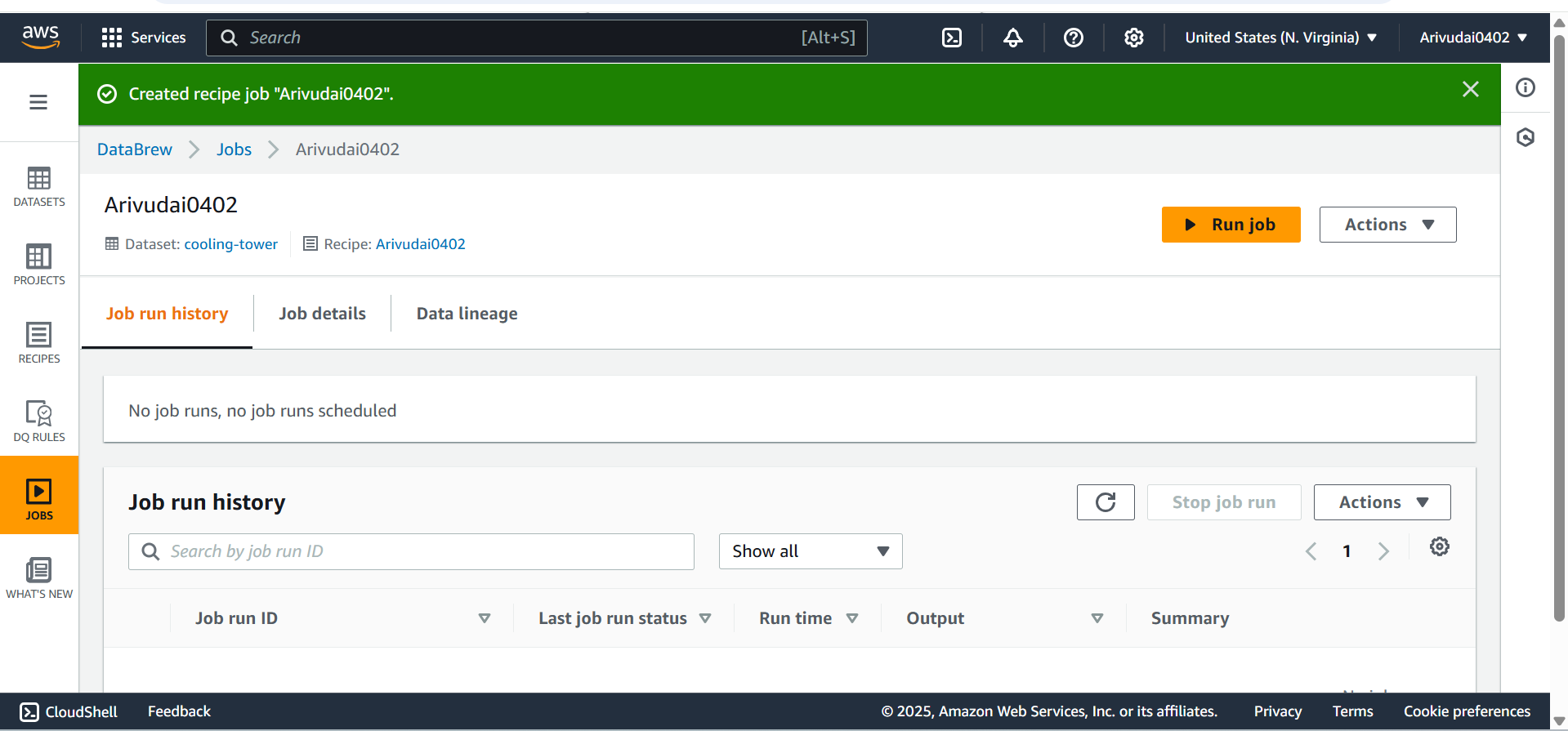


After the datasets were set up, I used DataBrew to build profile jobs that examined the data quality and structure. Profiling created insights into column data types, missing values, duplicate records, and variable correlation. Missing values in the Cooling Tower dataset's Legionella Level CFU, pH Level, and Chlorine Levels, for instance, were found. Inconsistencies in data types and formatting were also discovered while profiling the Building Water Treatment and Decorative Water Feature datasets.

**Data Cleaning**

I then conducted data cleaning to have the dataset properly organized, standardized, and ready for analysis once I had finished the data profiling process and discovered several data quality problems. I did this by creating a specific data cleaning action that fixed missing values, inconsistent data, improper delimiters, and storage format optimization.

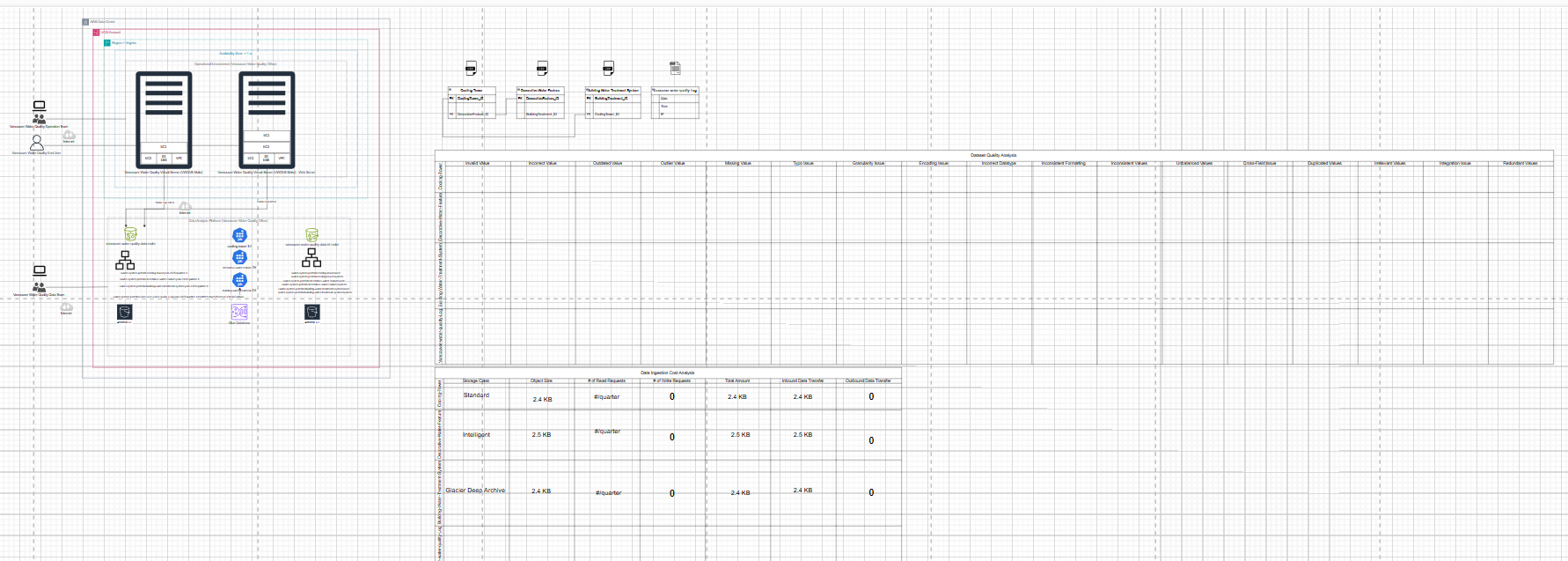
The "vwq-bwt-lst-cln-arivdai0402" data cleaning task was carried out using AWS Glue DataBrew. The dataset was refined using transformations in this task before being stored in an efficient manner.



numerical values.

**Formatting Output and Optimizing Storage**

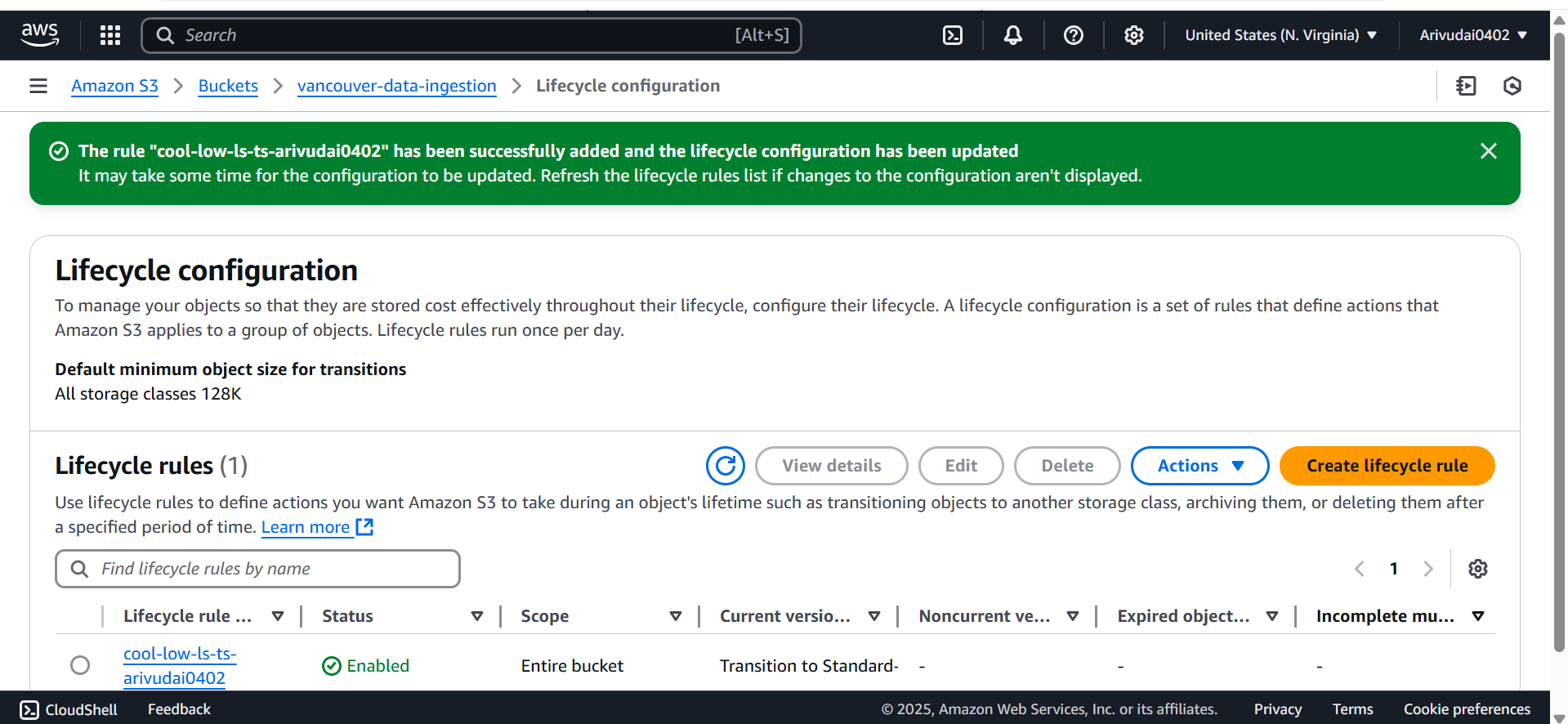
I saved the dataset in two optimal forms following the cleaning procedure:



**Optimization of Storage and Cost**

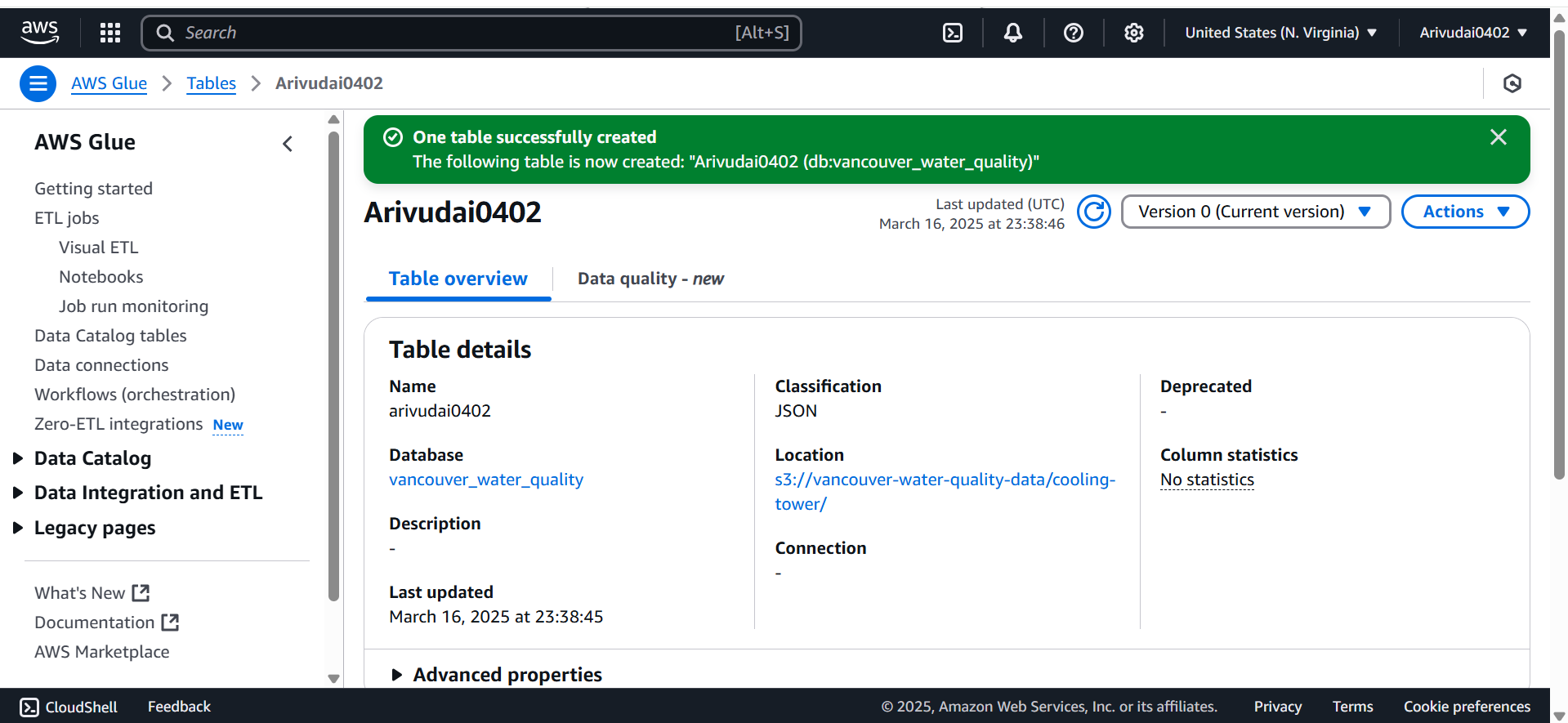
Once the data were cleaned, the converted datasets were placed in the "user" and "system" directories of the "vwq-bwt-lst-arivudai0402" S3 bucket.

Given that the dataset is refreshed frequently, I reduced storage costs by taking advantage of AWS S3 Lifecycle Rules. I maintained cost-effectiveness and quick access as needed by storing obsolete data in Amazon S3 Glacier Instant Retrieval. This helped me save a lot of money on long-term storage while not compromising data accessibility.



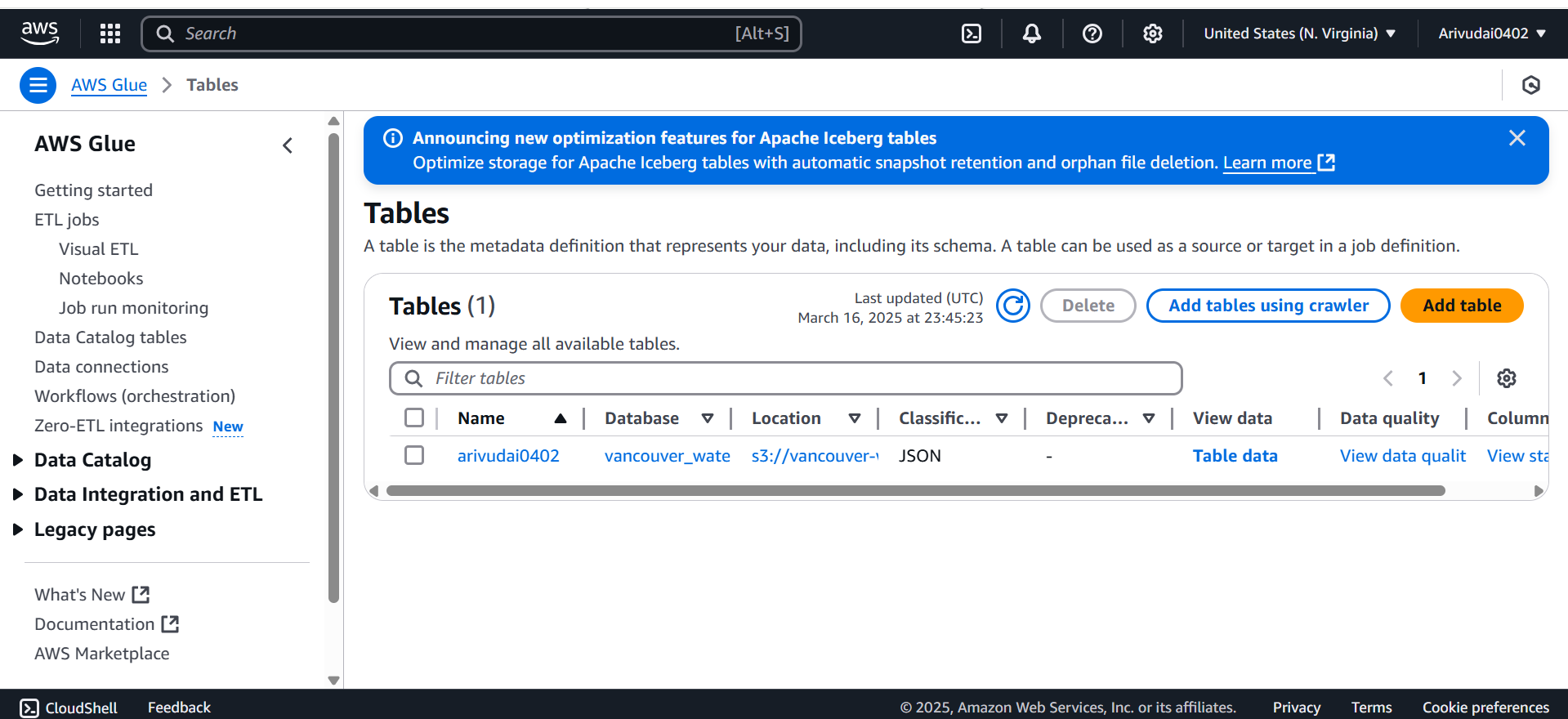
**Using the AWS Glue Crawler**

I initiated the data transformation and extraction process by running the AWS Glue Crawler after setting up the configurations. The tables created were examined in the AWS Glue Data Catalog after they were complete.



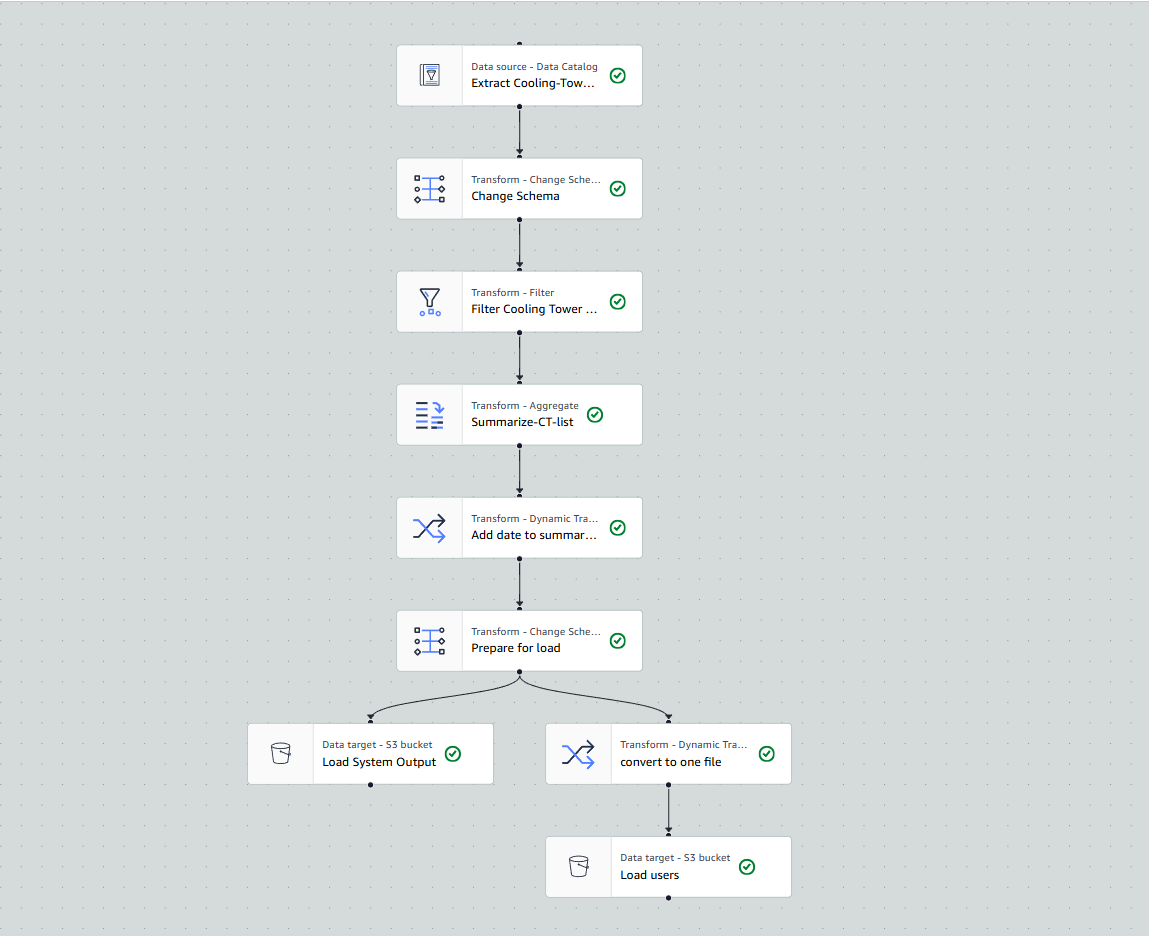
1. **Results of creating a data catalog**

The data was indexed successfully into structured tables that could be queried using Redshift, Amazon Athena, or other services. The data will remain well-structured and readily available for additional processing thanks to this orderly process.

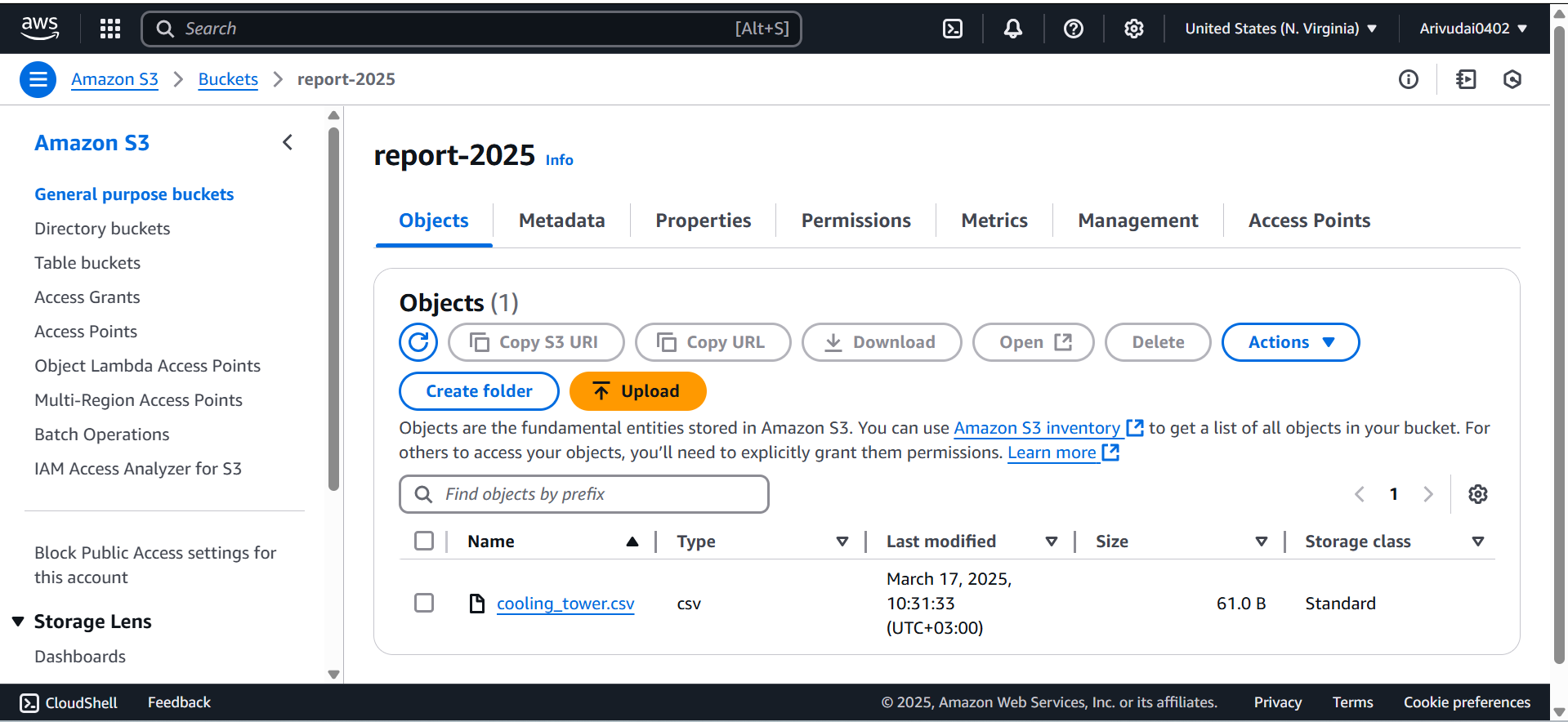


# **Data Summarization**

AWS Glue Visual ETL proved useful to enable such a change, enabling data transformation without too much coding. To make the pipeline of data processing set up for summarizing efficiently, a new Visual ETL job called "cooling-tower-list-Summarization" was established.

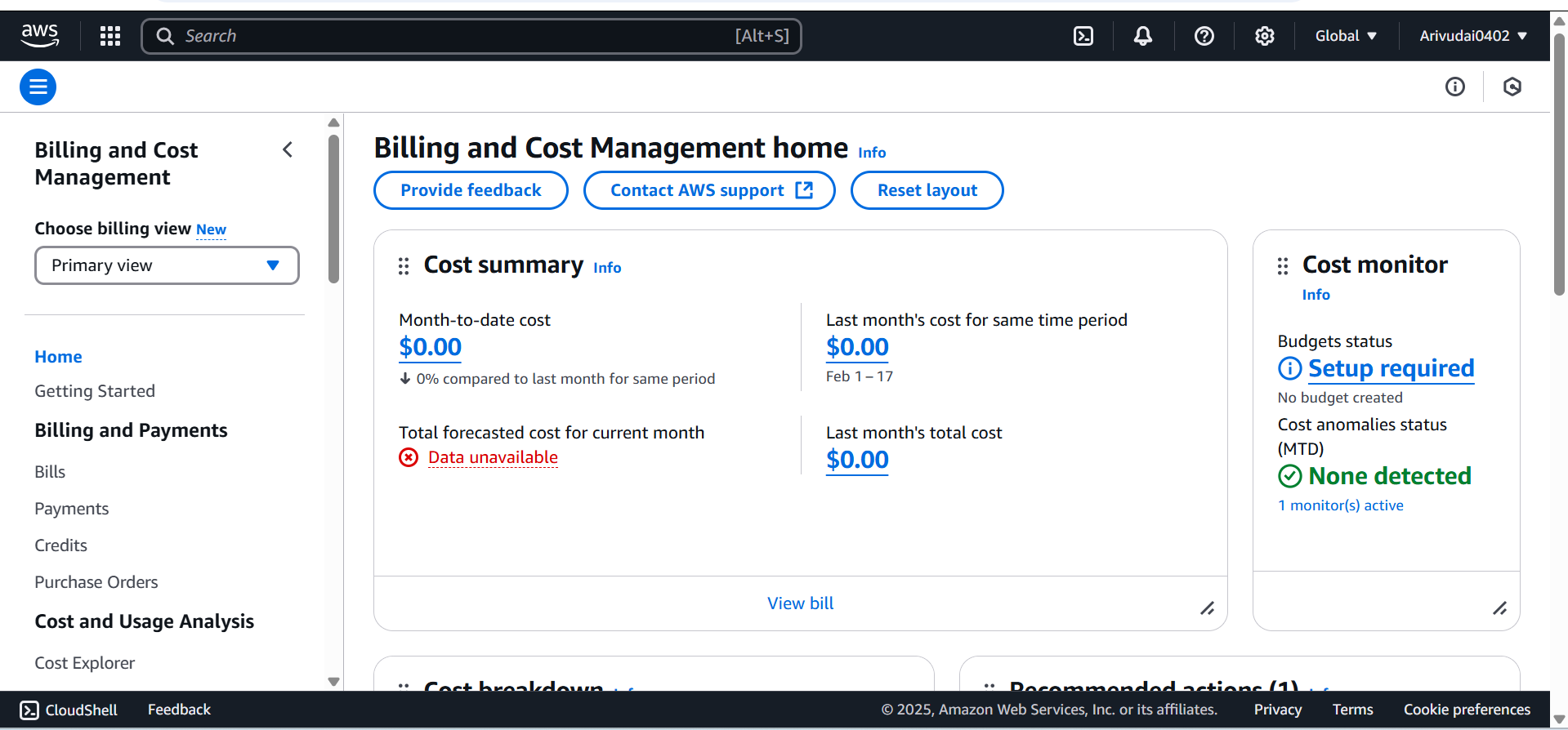


AWS Glue was provided with the required IAM permissions to allow access to the appropriate S3 buckets and tables prior to triggering the data transformation process. This made it possible for the ETL pipeline to load, extract, and transform data into the corresponding locations with ease. AWS Glue Data Catalog was chosen as the source of data for the ETL process, allowing structured data sets to be fetched with ease without the need for manual data extraction.



The information was well-structured and ready for additional analysis and reporting by adopting an ETL pipeline, providing easy access to summarized information for well-informed decision-making.

# **Monthly Cost**



The expense on the site is shown on the AWS Billing and Cost Management Dashboard. The overall estimated cost this month is $7.69, with the expense so far being $1.75. This is a 5,217%

**Recommendations:**

Improve operational efficiency through predictive insights.

**Tools & Technologies**

* AWS (S3, Redshift, Glue, Lambda, Athena, CloudTrail, GuardDuty)  
  Deliverables:
* AWS solution architecture and migration plan.

**References**

1. AWS Documentation. (n.d.). Amazon S3: Object Storage Service. Retrieved from <https://aws.amazon.com/s3/>
2. AWS Glue. (n.d.). AWS Glue DataBrew Overview. Retrieved from <https://aws.amazon.com/glue/databrew/>
3. City of Vancouver Open Data Portal. (n.d.). Water Quality Data. Retrieved from <https://opendata.vancouver.ca>
4. AWS Cost Management. (n.d.). Billing Dashboard Overview. Retrieved from <https://aws.amazon.com/aws-cost-management/>