Step 1: Data Ingestion

A screenshot of a computer

AI-generated content may be incorrect.

I uploaded raw data from Vancouver through the Amazon S3 platform for structured along with unstructured data storage. Secure data storage and management of the data took place through AWS S3 buckets before processing. The platform S3 provided durable storage features together with effortless system access that supported additional processing needs. The configuration provided easy combination possibilities with AWS services to conduct profiling and cleaning operations.

Step:2 Data Profiling

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

I consumed the data into S3 before using AWS Glue Data Brew to study data structure and detect anomalies and handle missing or inconsistent values. The analysis made it possible to understand the dataset before initiating the cleaning process. The automated profiling tools available in AWS simplified efficient data quality and structure evaluation procedures.

Step 3: Data Cleaning

A computer screen with a white screen

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

The data cleaning required an AWS EC2 instance to run scripts which cleaned missing entries and duplicate data before formatting it correctly. The consistency of the dataset was achieved through AWS Glue while preparing it for subsequent processing steps. The data cleaning operations and large dataset processing work efficiently on EC2 instances because they supply adequate computational capacity.

Sep:4 Data Cataloguing

A computer screen shot of a computer screen

AI-generated content may be incorrect.

A computer screen shot of a computer

AI-generated content may be incorrect.

A computer screen shot of a computer

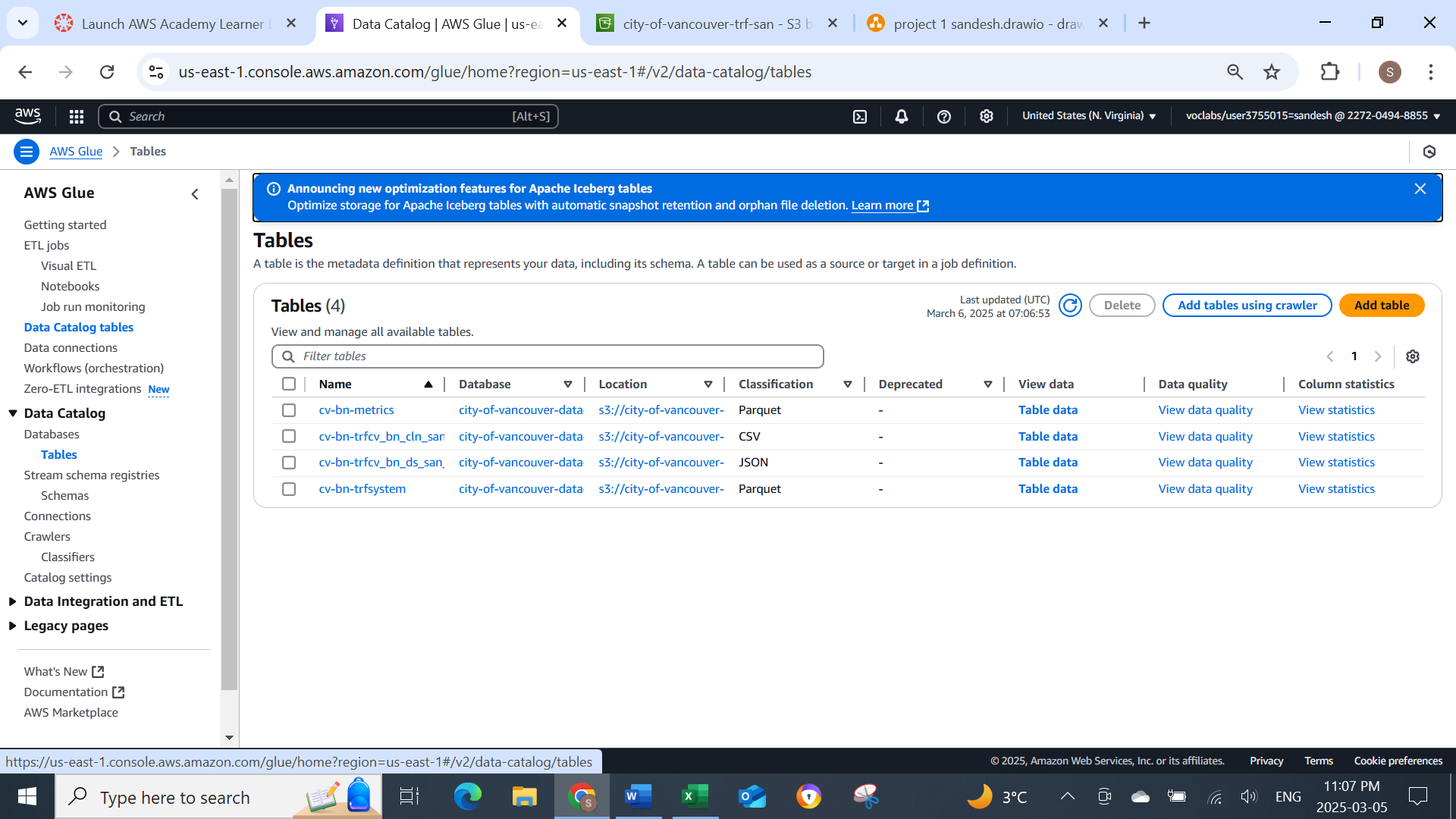
AI-generated content may be incorrect.

At this point I utilized AWS Glue Data Catalog(using crawler) to catalog cleaned data which provided searchable metadata within AWS services. By performing this step data analysts and both data scientists working with data mining models could effortlessly locate and implement the datasets. The catalog function acted as a metadata storage system for table schemas and S3 data locations in addition to other data elements.

Step:5 Data Summarizing

A screenshot of a computer

AI-generated content may be incorrect.



Finally, the processed data required summarization with AWS Glue( through ETL) for report and visualization generation. The EC2 instance served as an alternative execution environment to perform data queries together with analytical scripts that created statistical summaries. The cleaning of the City of Vancouver block number dataset enabled us to obtain analytical insights that support decision processes.