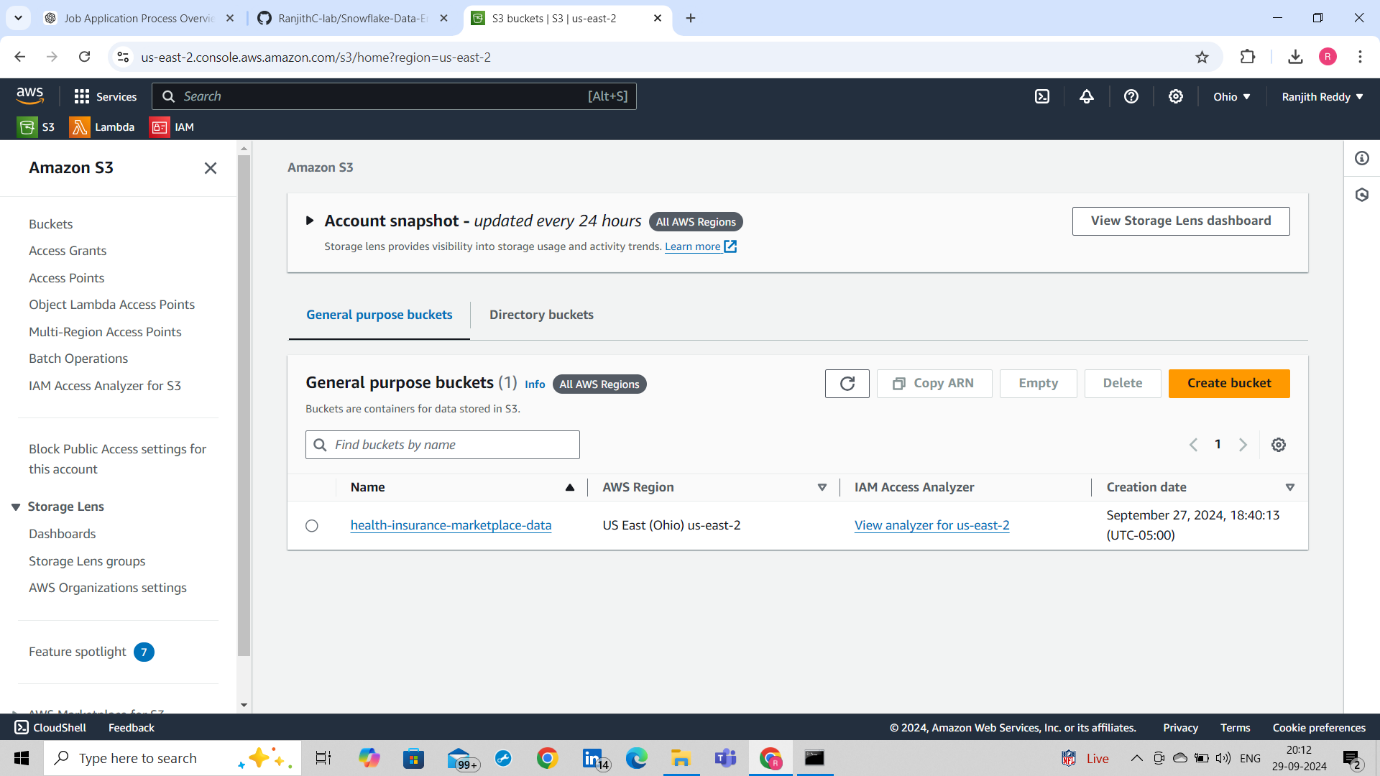
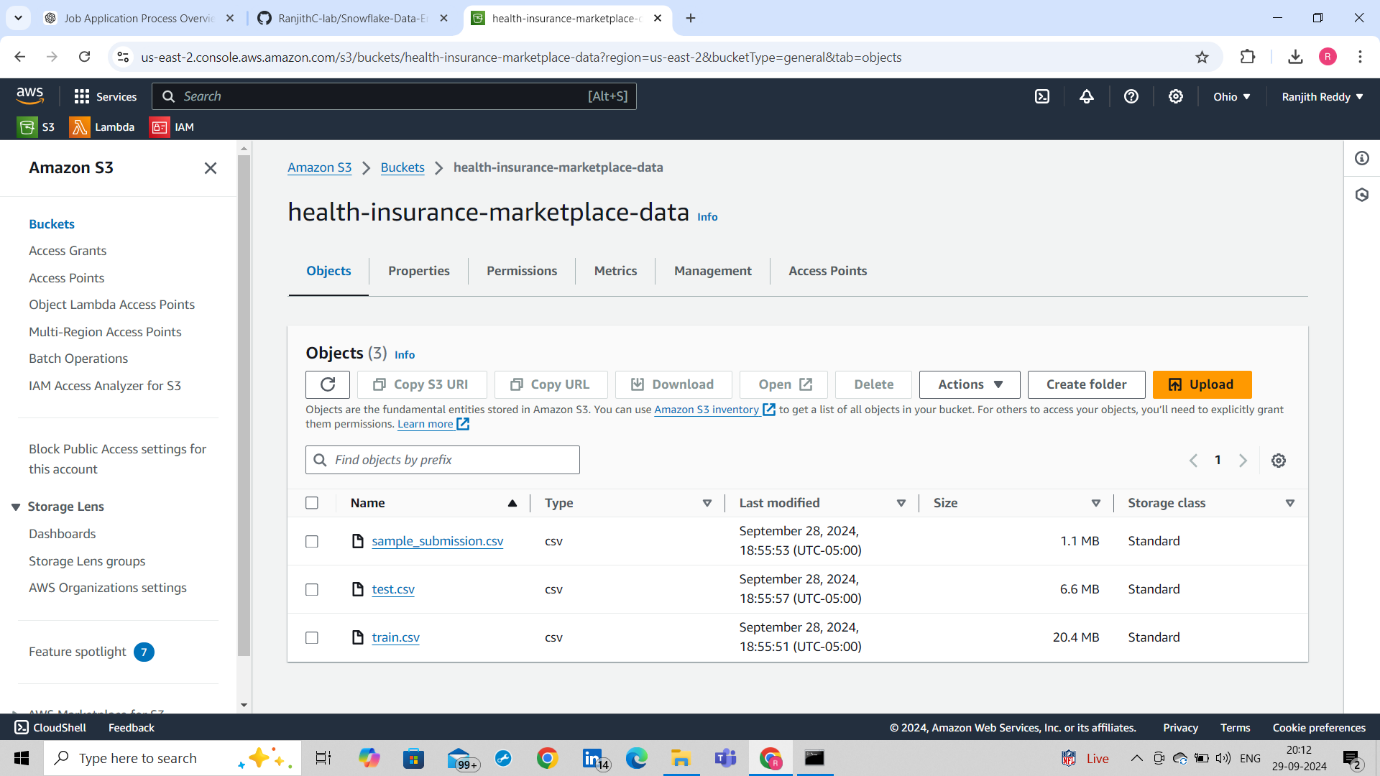
Snowflake Data Engineering with AWS S3 Integration

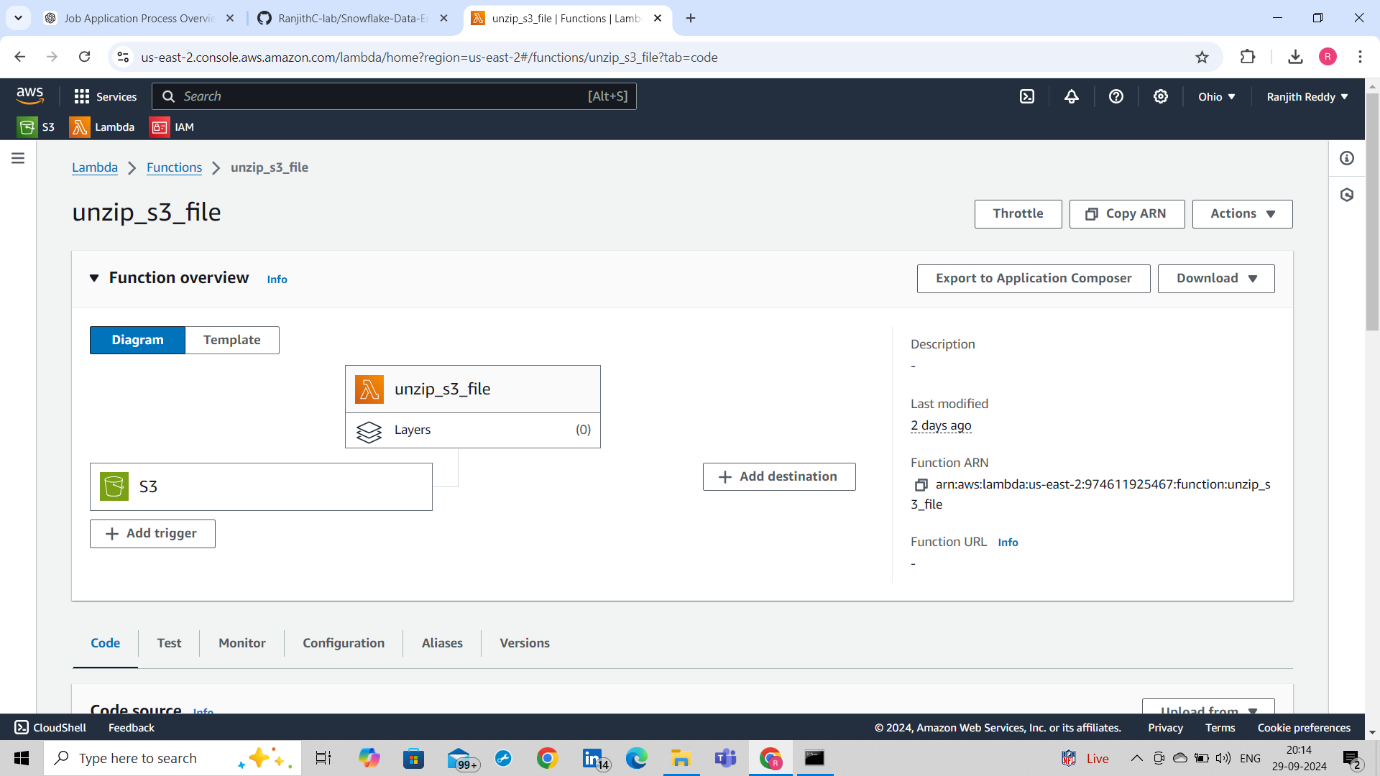


**Setting Up AWS S3 Bucket**

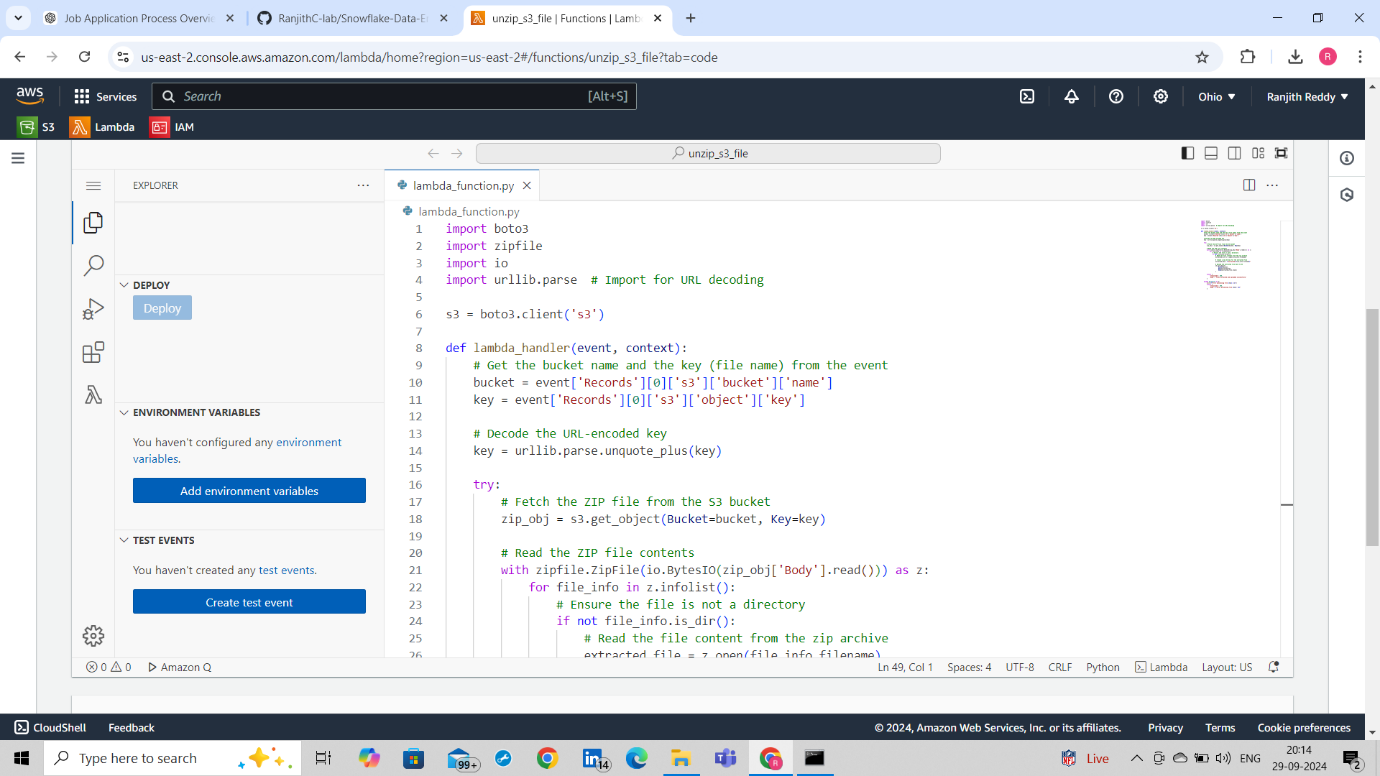
* Created an S3 bucket named health-insurance-marketplace-data.
* Uploaded train.csv, test.csv, and sample\_submission.csv into the S3 bucket.
* **Screenshot 1:** Displaying the uploaded files in S3.

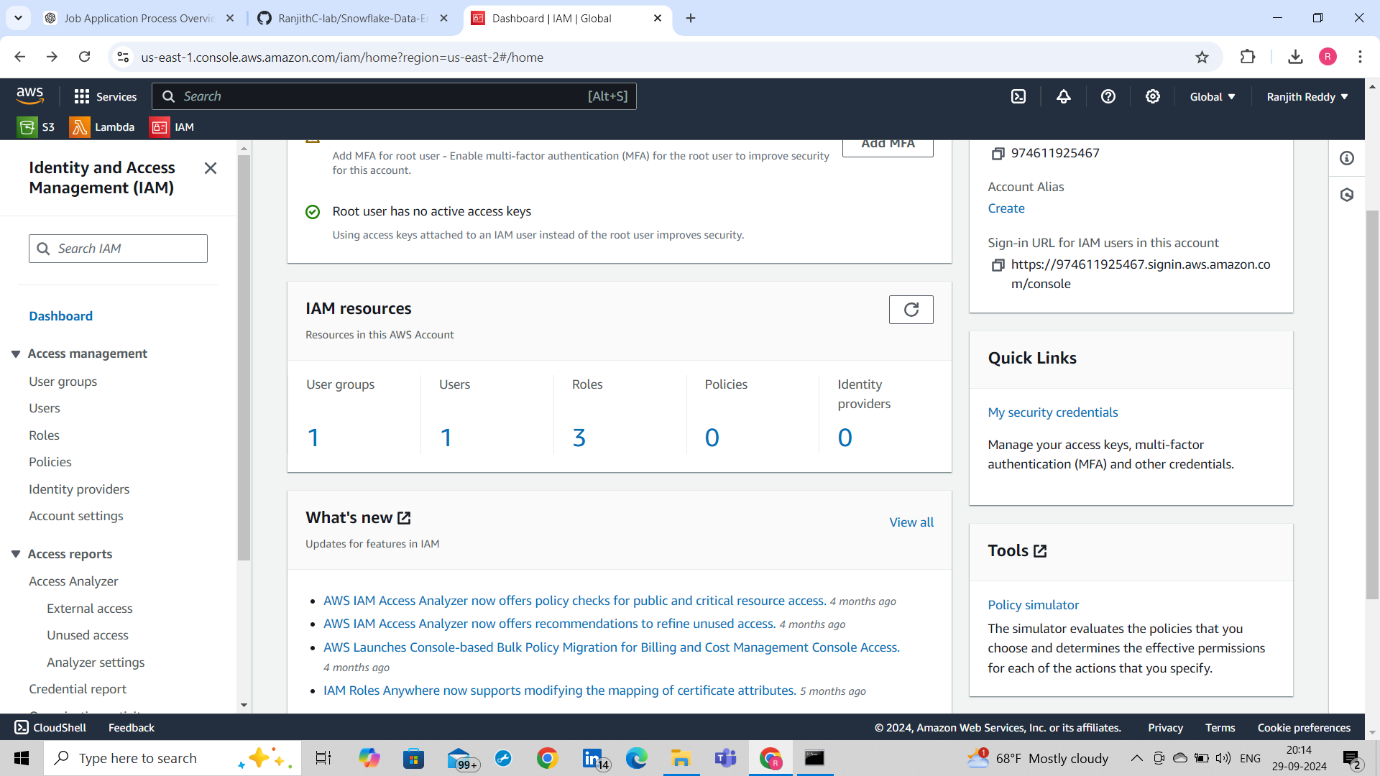


* **Data Description:**
  + train.csv: Used for training the model, contains customer demographic data.
  + test.csv: Data for testing, used for predictions.
  + sample\_submission.csv: Submission format for predictions.

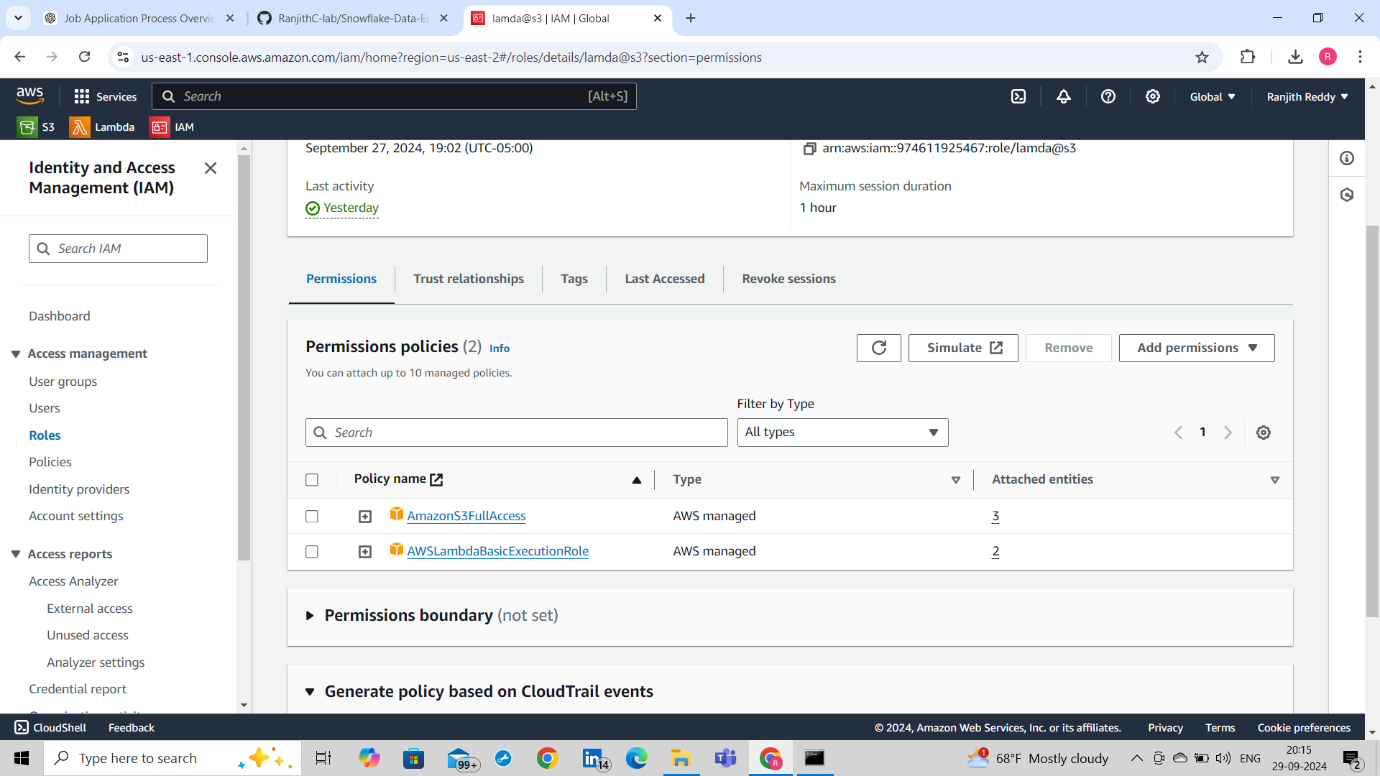


To automate the unzipping of files uploaded to the S3 bucket, an **AWS Lambda function** was triggered whenever a new file was uploaded to a specific S3 bucket. The Lambda function was designed to automatically unzip the uploaded file and store its contents back into the same bucket.

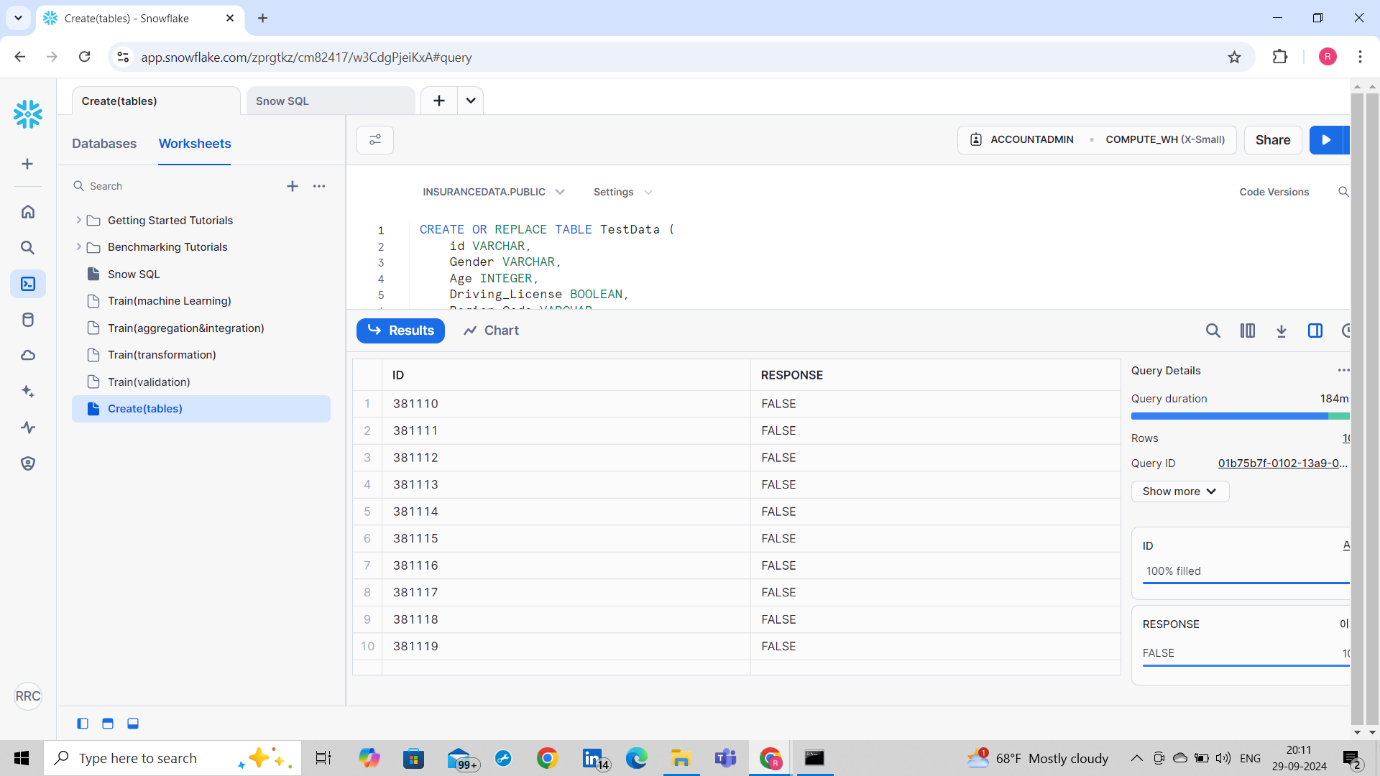
The **Lambda function code** is deployed to automatically retrieve the uploaded ZIP file from the S3 bucket, unzip its contents, and upload the extracted files back to the same bucket for further processing.



In this project, **IAM (Identity and Access Management)** was used to manage permissions and roles, ensuring that the Lambda function had the necessary access to interact with the S3 bucket and perform operations securely.



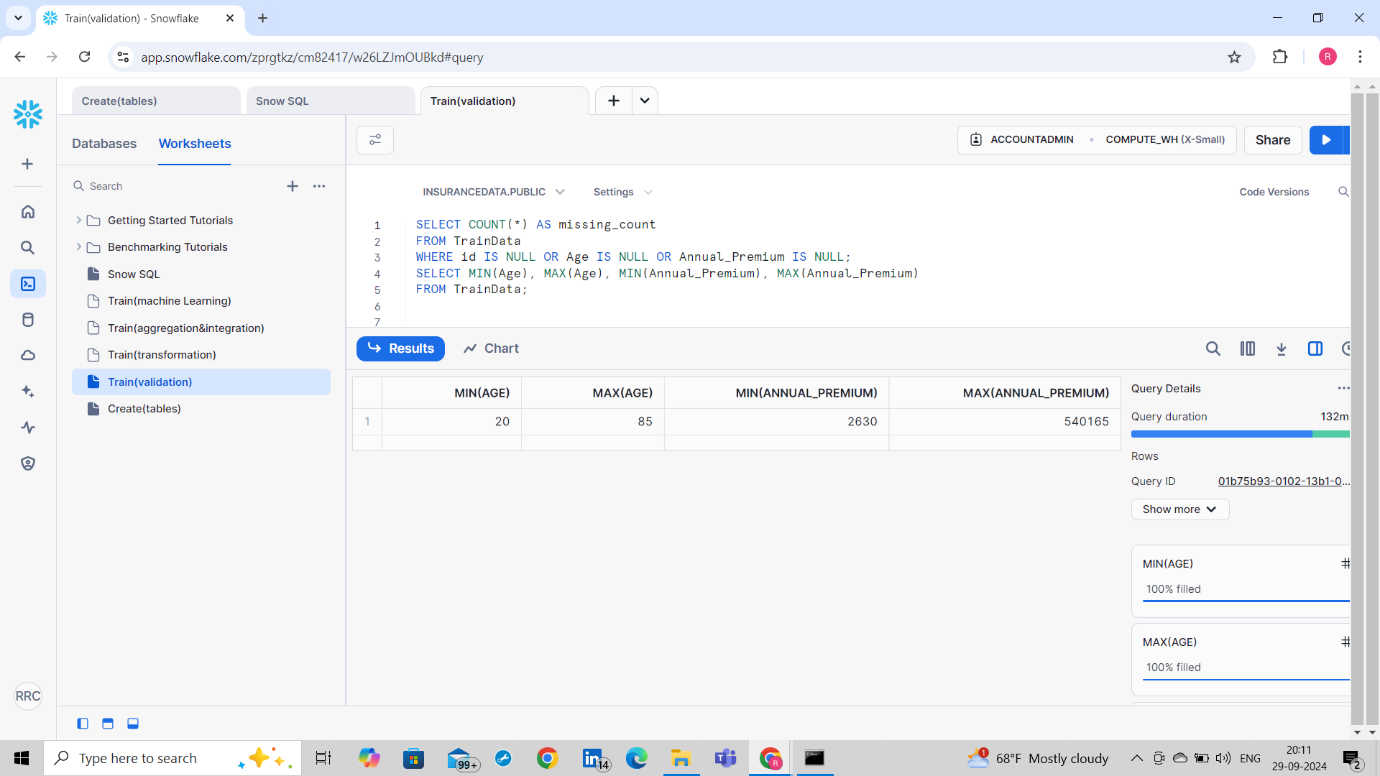
**AmazonS3FullAccess** was assigned to allow the Lambda function full access to the S3 bucket, enabling operations like reading, writing, and listing objects. Additionally, the **AWSLambdaBasicExecutionRole** was assigned to provide the necessary permissions for the Lambda function to execute and write logs to Amazon CloudWatch for monitoring and debugging purposes.



Tables like TrainData and TestData were created to store the datasets.

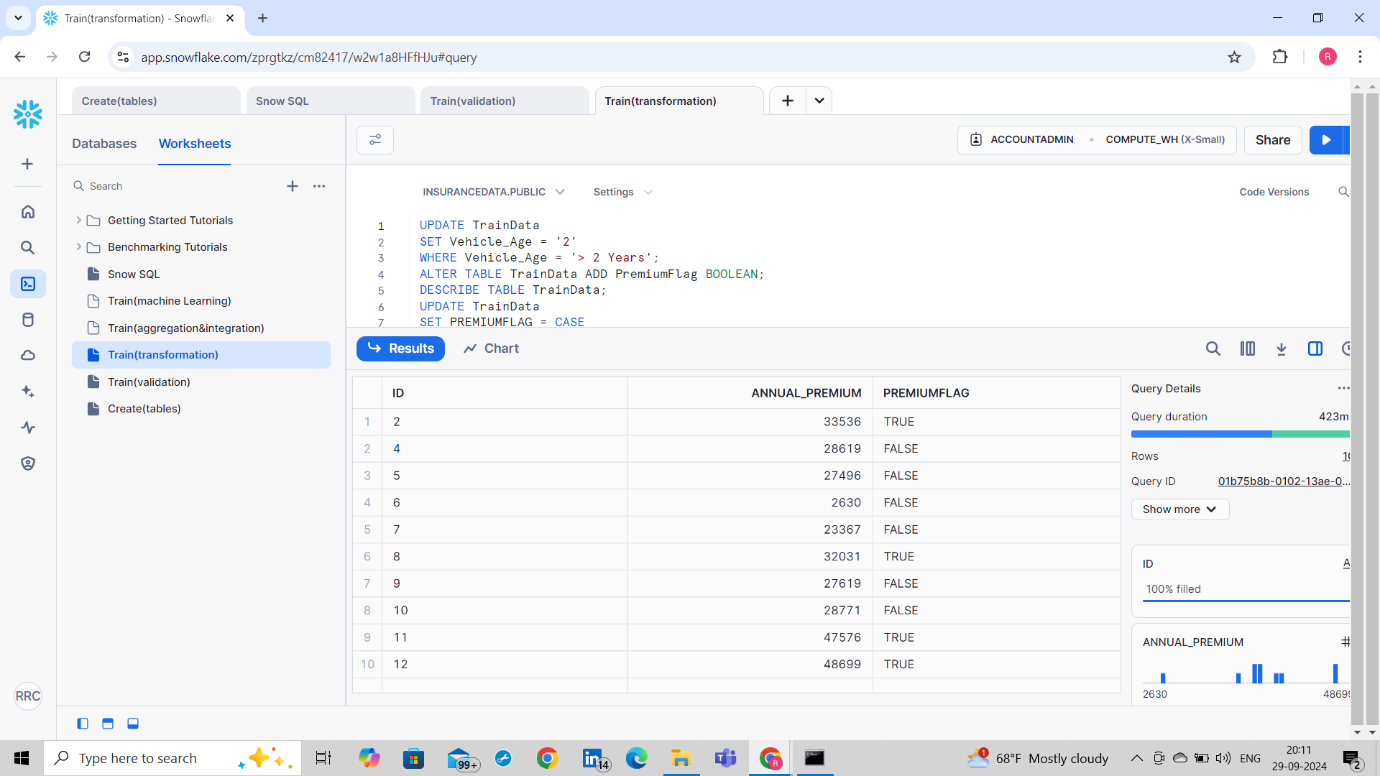
SQL scripts were used to define the table schema.

SQL script showing table creation for TrainData.



**Data Validation**

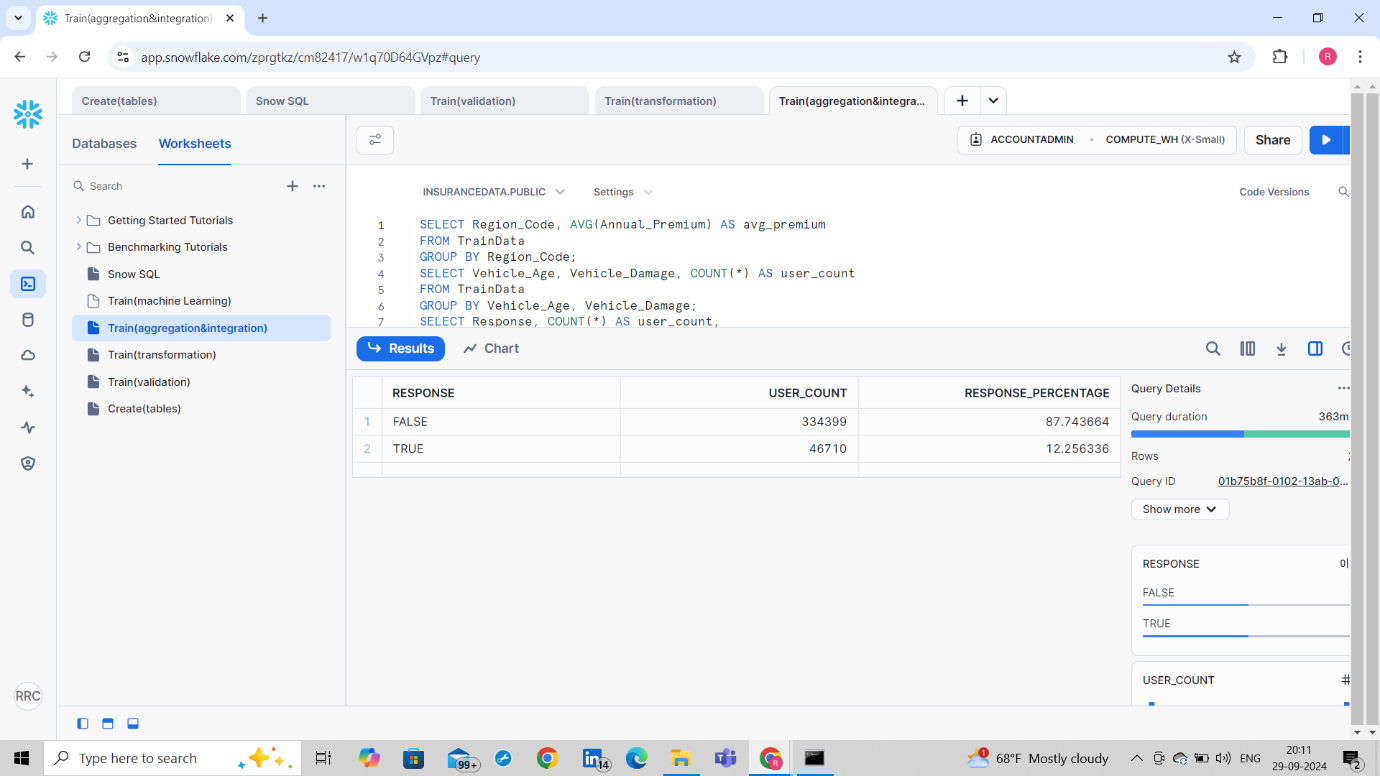
Performed data validation checks like missing values and minimum/maximum values.



**Transformation**

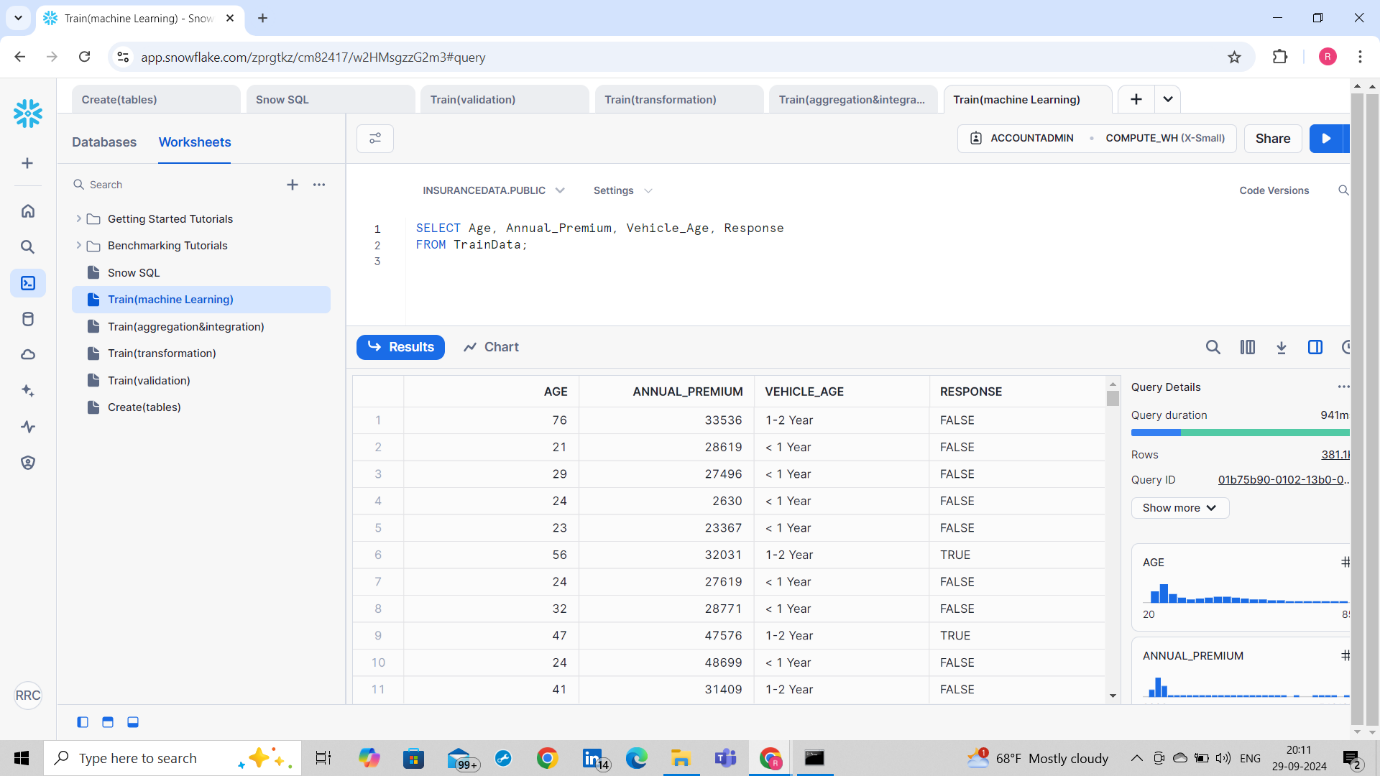
Applied data transformations like flagging premium customers.

* + Shows how missing values were handled and transformations applied.



**Data Analysis Queries**

* Used aggregation functions to analyze data trends.
* Example: Analysis of the average premium by region and vehicle damage by age.
  + Displaying the results of these aggregation queries in Snowflake.



In this project, **machine learning** can be applied to the dataset by building predictive models using the cleaned and transformed data. After data extraction and transformation in Snowflake, features like customer demographics and policy details can be used to train models, such as **classification models**, to predict customer behaviour, such as whether they are likely to buy additional insurance. Snowflake integrates well with machine learning workflows, enabling seamless analysis and model deployment.

**Challenges Faced**

* Encountered permissions issues with S3 which were resolved by updating IAM roles.
* Dealt with column mismatches and adjusted file format options for smooth data loading.

**Final Output**

The project successfully loaded and processed the data in Snowflake.