Data Warehousing & Integration
IE6750
Spring 2025

Customer Personality Analysis

Milestone 6 (starts on pg. 4)
Group 5

Anushka Jami
Barath Keshav Sriram Kumaran

jami.a@northeastern.edu
sriramkumaran.b@northeastern.edu

4/13/2025

Problem Statement:

Customer Personality Analysis is a critical process that enables businesses to gain a deep understanding of their customer segments. By analyzing customer data, businesses can separate their customer base into distinct groups based on various characteristics, needs, behaviors, and preferences. The goal of this analysis is to provide businesses with insights into how different types of customers interact with products and services, ultimately allowing companies to customize their offerings to meet the unique demands of each customer segment.

The key objective of customer personality analysis is to optimize marketing efforts and product development by identifying and targeting specific customer segments that are more likely to engage with and purchase particular products. Rather than investing resources in marketing a product to every customer within a business's database, customer personality analysis helps identify which customer segment is most likely to respond to a given product and marketing effort. By focusing on these high-potential segments, businesses can enhance marketing efficiency, improve customer satisfaction, and ultimately drive higher sales.

By leveraging this customer personality analysis, businesses can make more informed decisions about product development, promotional strategies, and resource allocation. This approach allows for more personalized marketing campaigns and a deeper understanding of customer behavior, leading to increased customer sales and better business outcomes.

Datasets and Dimensions:

1. https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis
2. Datasets extracted from: https://www.bls.gov/news.release/pdf/cesan.pdf
    a. Tables

i.

Expenditure_ID	Year	Category	Subcategory	Amount (USD)
1	2021	Food	Food at Home	5,259
2	2021	Food	Food Away from Home	3,030
3	2021	Food	Total (Food)	8,289
4	2022	Food	Food at Home	5,703
5	2022	Food	Food Away from Home	3,639
6	2022	Food	Total (Food)	9,343
7	2023	Food	Food at Home	6,053
8	2023	Food	Food Away from Home	3,933
9	2023	Food	Total (Food)	9,985
10	2021	Housing	Owned Dwellings	7,591
11	2021	Housing	Rented Dwellings	4,684
12	2021	Housing	Other Lodging	983
13	2021	Housing	Total (Housing)	22,624
14	2022	Housing	Owned Dwellings	8,230
15	2022	Housing	Rented Dwellings	4,990

ii.

	A	B	C	D	E	F	G
1	Item	All Consumer Units	Married Couple Only	Married Couple with Children	Other Married Couple Consumer Units	One Parent, at Least One Child Under 18	Single Perso and Other Consumer Un
2	Housing	32.9	31.2	30.6	30.7	37.3	3
3	Transportation	17	16.9	17.4	19.3	19	1
4	Food	12.9	12.2	13.6	14.5	14.3	1
5	Personal Insurance & Pensions	12.4	11.9	15.1	13.3	9	1
6	Healthcare	8	10	6.9	7.8	4.9	
7	Entertainment	4.7	5	4.9	4.1	3.8	
	Other	12.1	12.9	11.6	10.3	11.7	1

iii.

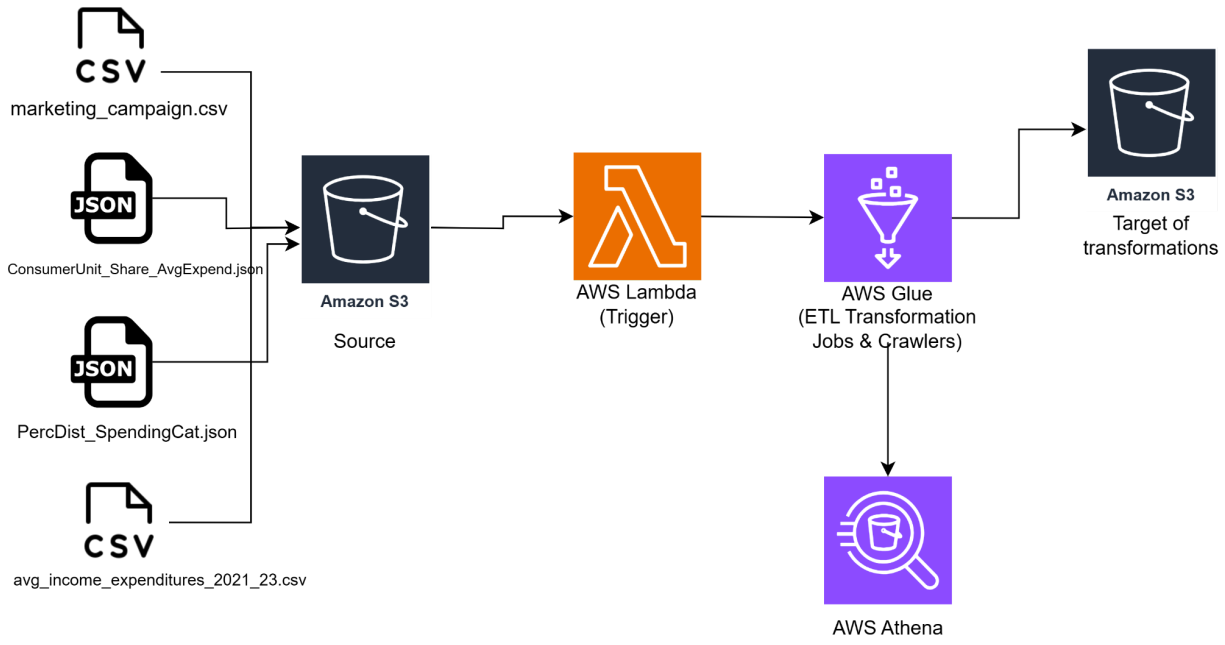
Spending Category	2020	2021	2022	2023
Food	11.9	12.4	12.8	12.9
Alcoholic Beverages	0.8	0.8	0.8	0.8
Housing	34.9	33.8	33.3	32.9
Apparel and Services	2.3	2.6	2.7	2.6
Transportation	16	16.4	16.8	17
Healthcare	8.4	8.1	8	8
Entertainment	4.7	5.3	4.7	4.7
Personal Care Products and Services	1.1	1.2	1.2	1.2
Reading	0.2	0.2	0.2	0.2
Education	2.1	1.8	1.8	2.1
Tobacco Products and Smoking Supplies	0.5	0.5	0.5	0.5

1. Education
2. Marital status
3. Children present
4. Income
5. Product Type
6. Method of Purchase

Datasets in analysis: Two csv files and two JSON files

1. marketing_campaign.csv
2. avg_income_expenditures_2021_23.csv
3. ConsumerUnit_Share_AvgExpend.json
4. PercDist_SpendingCat.json

Architecture Diagram:



Ingesting Raw Data Files

The architecture diagram above describes a cloud-based ETL pipeline that uses AWS services and components to prepare a target datasource ready for analysis. This pipeline processes raw data from both CSV and JSON files using Glue jobs to transform the raw data into a format to easily integrate multiple sources. Once the data is ready, it is loaded into a target S3 bucket to store the final analysis-ready result and is saved in a Glue Catalog data table to be queried using Athena. Below is a more detailed explanation of the architecture:

1. Ingesting the data files (CSV and JSON)

The raw data files, two CSV files and two JSON files, are ingested into an Amazon S3 bucket to be stored in their own folders within a cohesive project folder.

2. Amazon S3 source file storage

This S3 bucket serves as a place for raw data storage as well as intermediate dataset storage for data not fully processed. Files uploaded here are checked for filetype using a trigger.

3. AWS Lambda
    a. Trigger

A Lambda function is triggered when a file is uploaded to the source S3 bucket, warning when the file is not a CSV or JSON as expected.

4. AWS Glue
    a. ETL Transformation Jobs

Data was processed using ETL Notebook jobs written in Python. These were used to clean the data by dropping unnecessary columns and renaming column names for standardization and interpretability. In order to merge data from the four tables, three were merged on the “Spending Category” / ”Item” / ”Category” columns, the result of which was then merged to the fourth using a mapping column to map conditions of the dataset to values in the initial merged dataset.

    b. Crawlers

A crawler was created for each step in the transformation process, used to catalog the data into a Glue Database. These were then viewable and queryable using AWS Athena.

5. AWS Athena

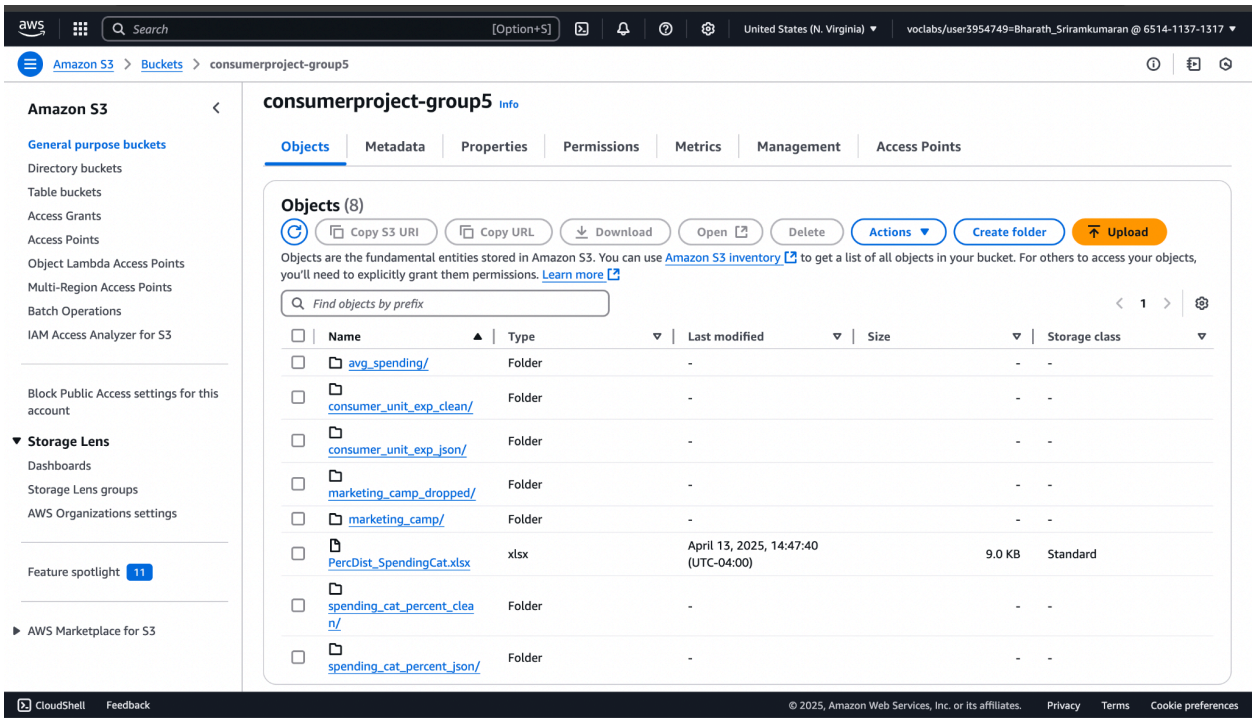
Athena was used to query the data at different steps throughout the transformation process and additionally to query the final resulting data table.

6. Amazon S3 target file storage

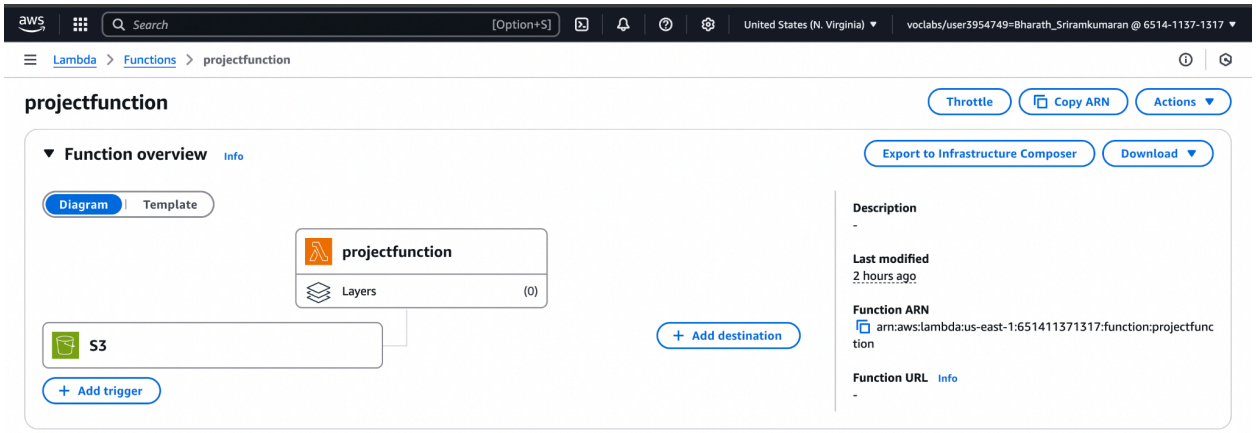
This S3 bucket serves as a storage space for the output of the final Glue ETL transformation. This dataset is a clean and analysis-ready dataset that can be downloaded in a CSV file.

Data Flow screenshots

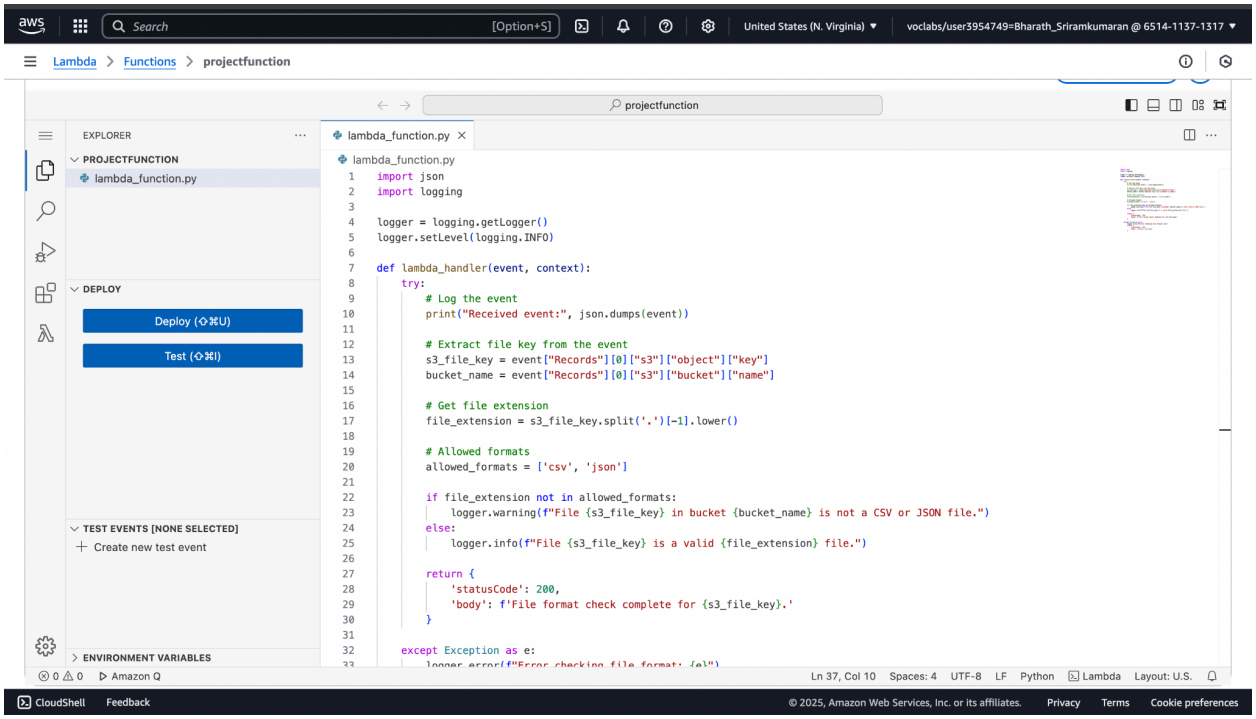
The Source Bucket is created with a folder for each file to be able to retrieve its schema.



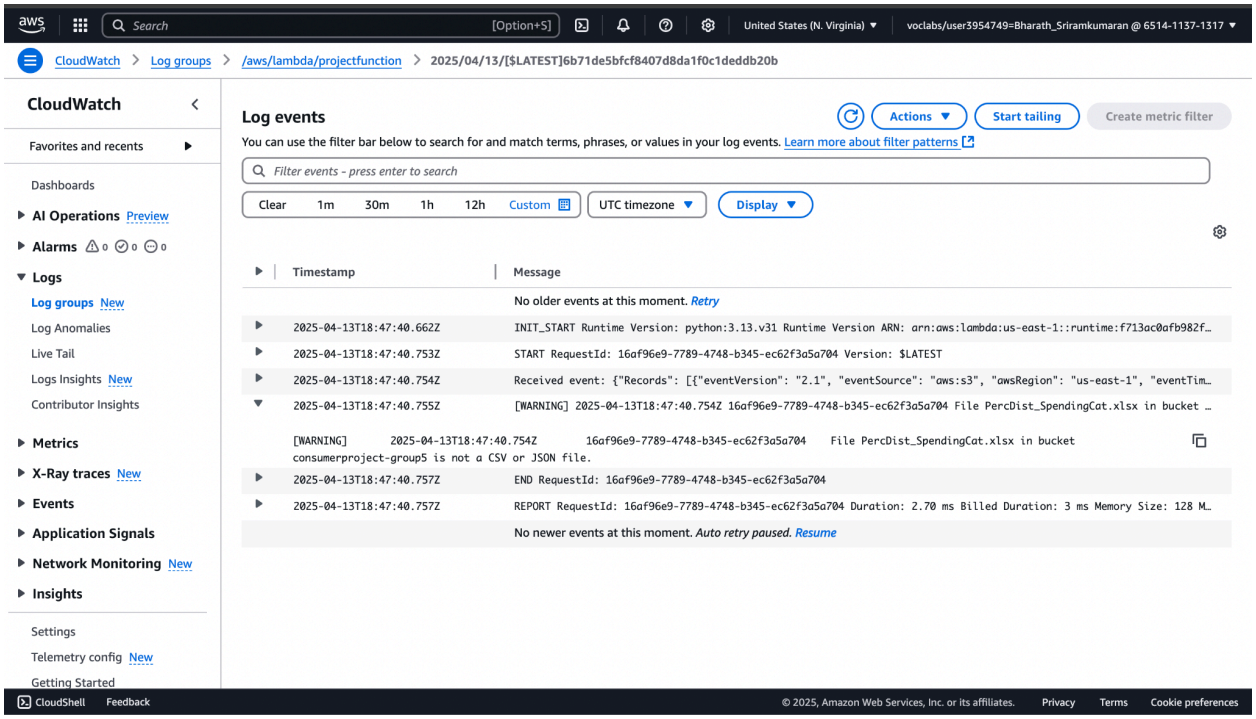
Creating a Lambda function and adding a trigger that warns when a file is added, which is not in .csv or .json format.



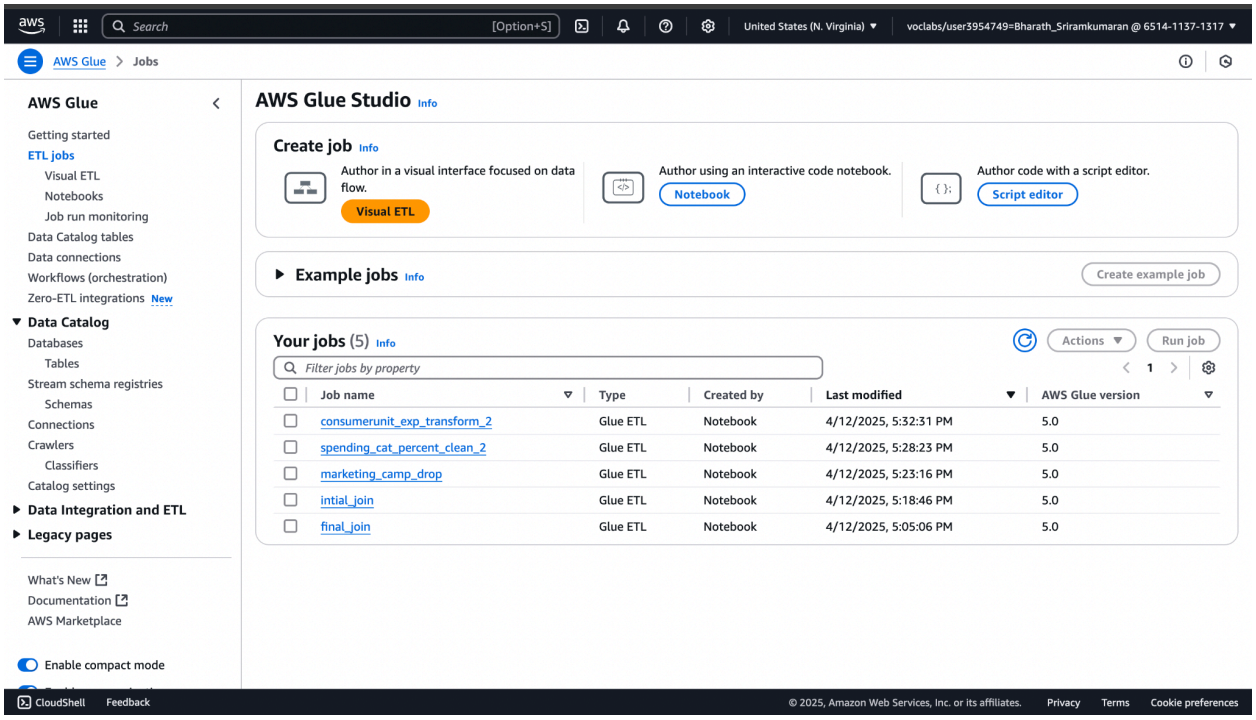
Code source for the Trigger.



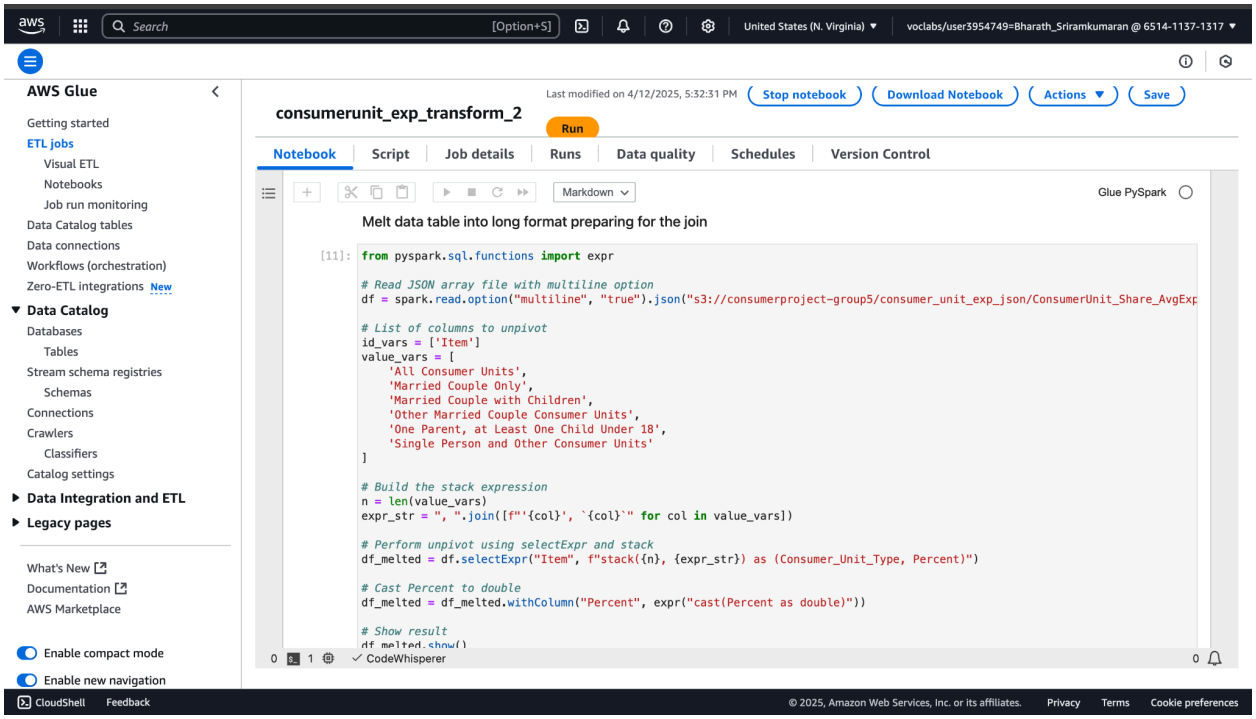
Log Event to show that the trigger has given a warning as there was a file added that was not in a .csv or .json format.



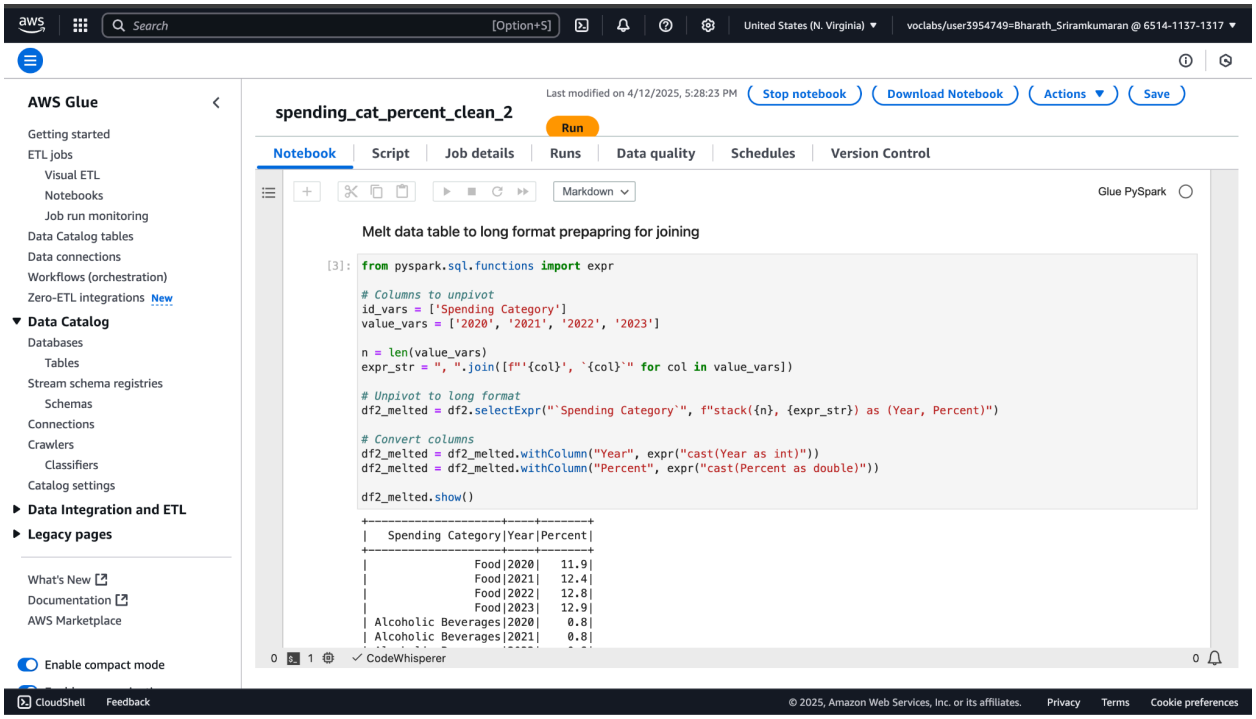
Using AWS Glue for doing the cleaning, transformations, and the joining of all the tables.



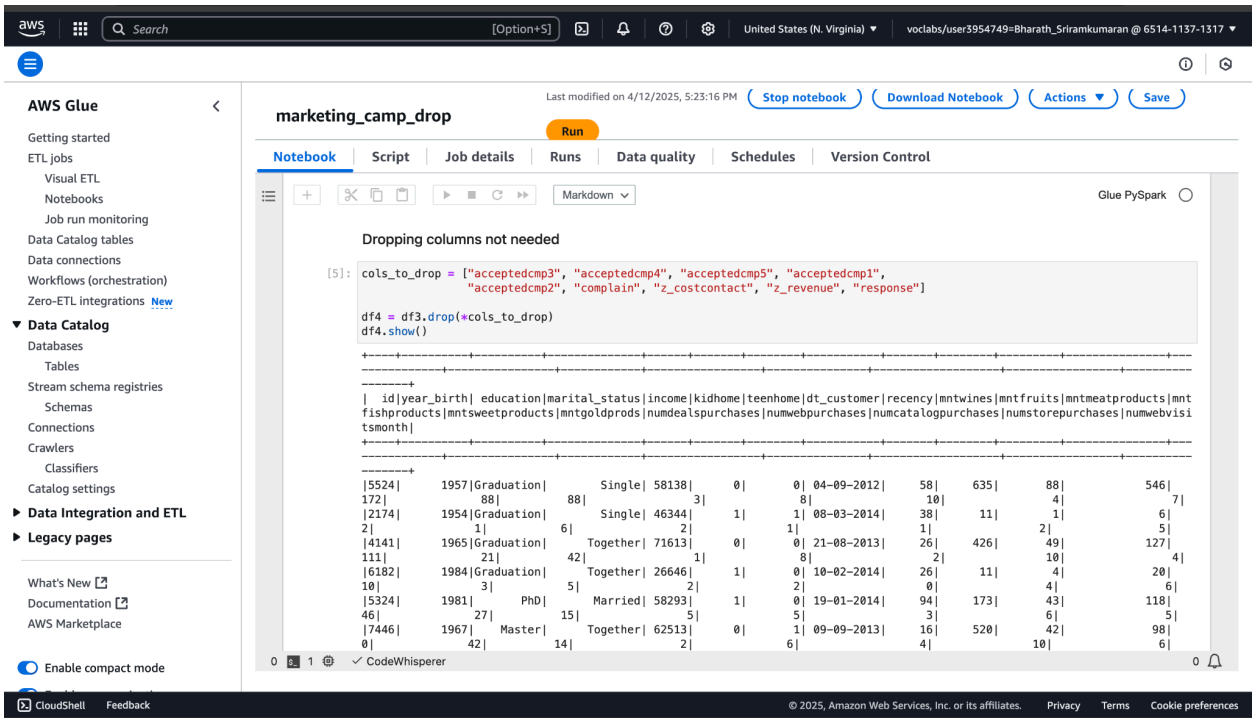
Cleaning of the table Consumer Units Average Expenditure.



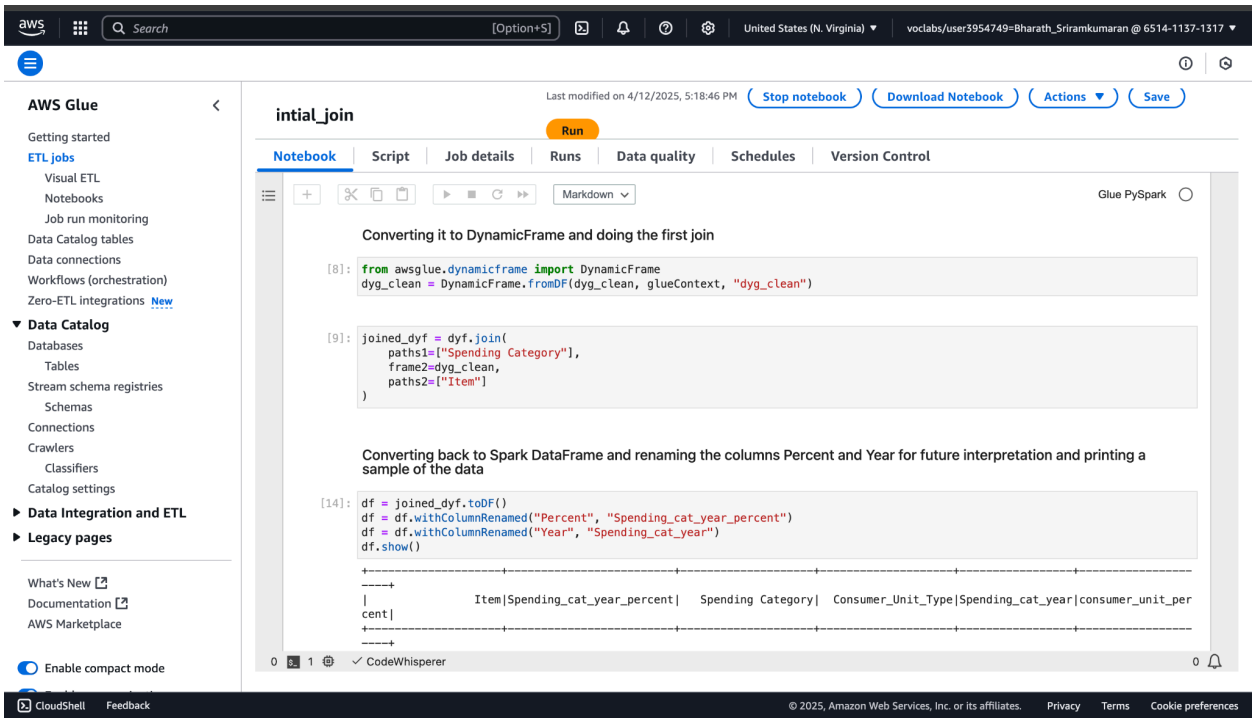
Cleaning of the table Spending Category Percentage.

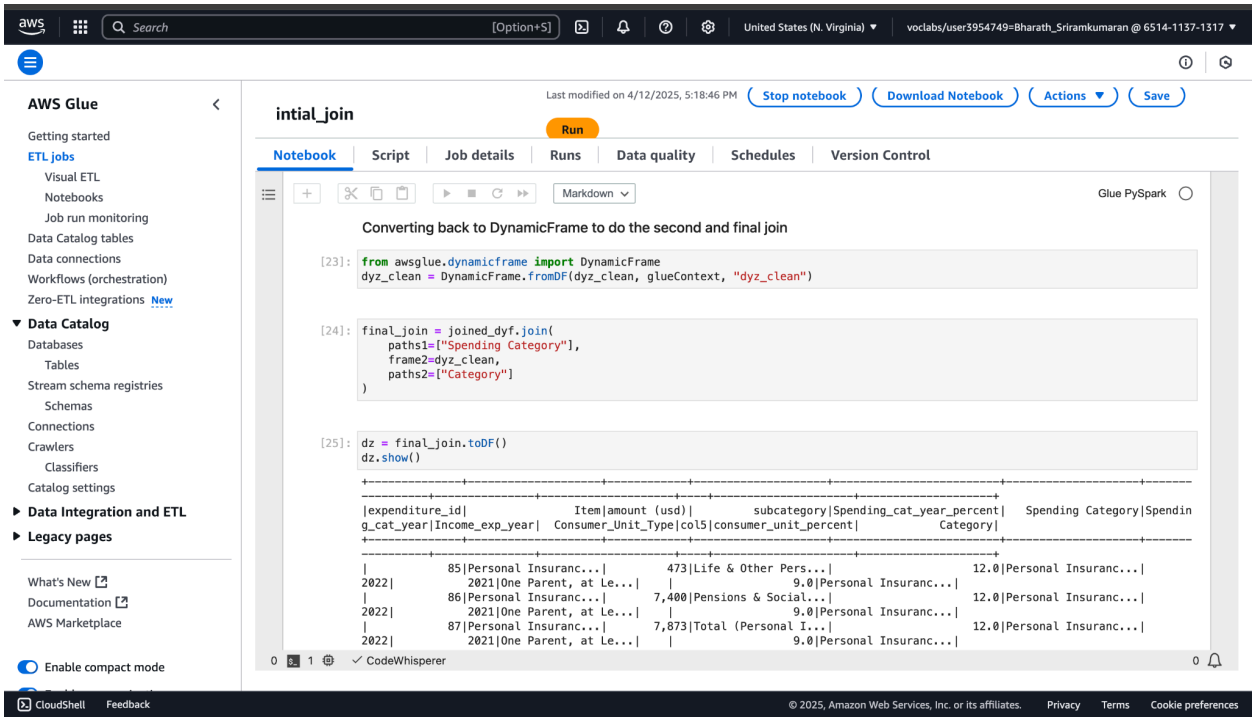


Dropping unused columns from the Marketing Campaign table.

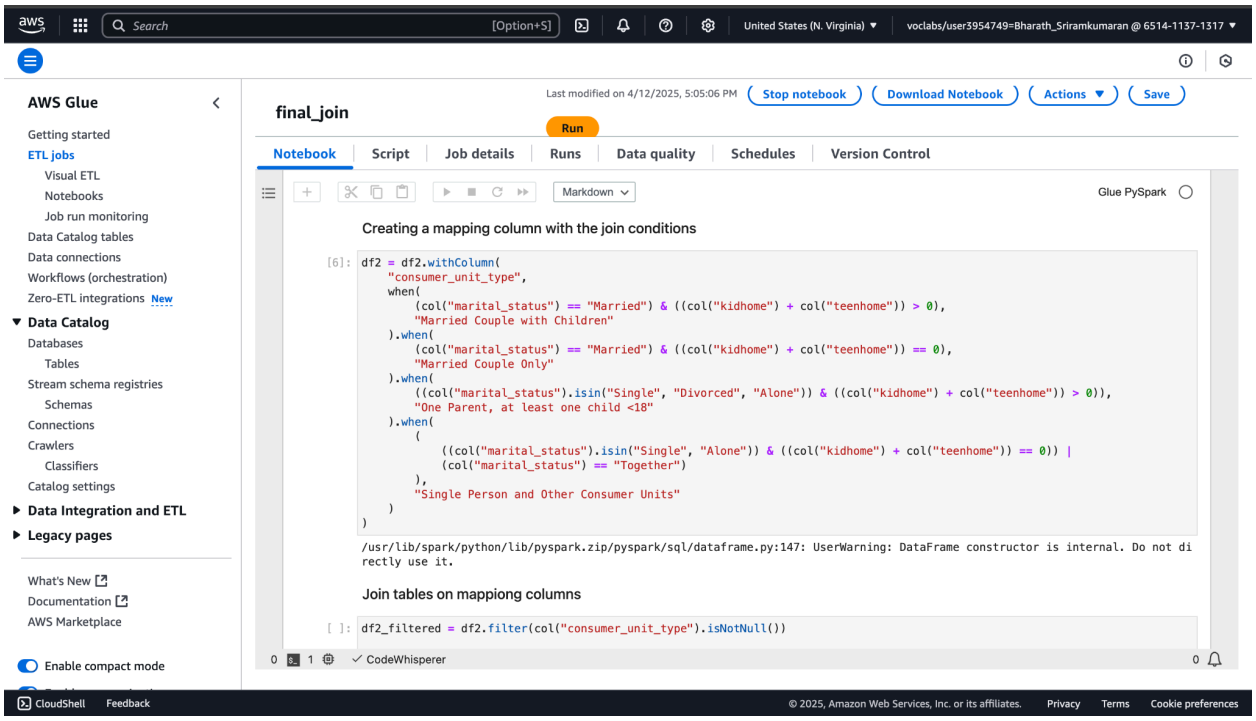


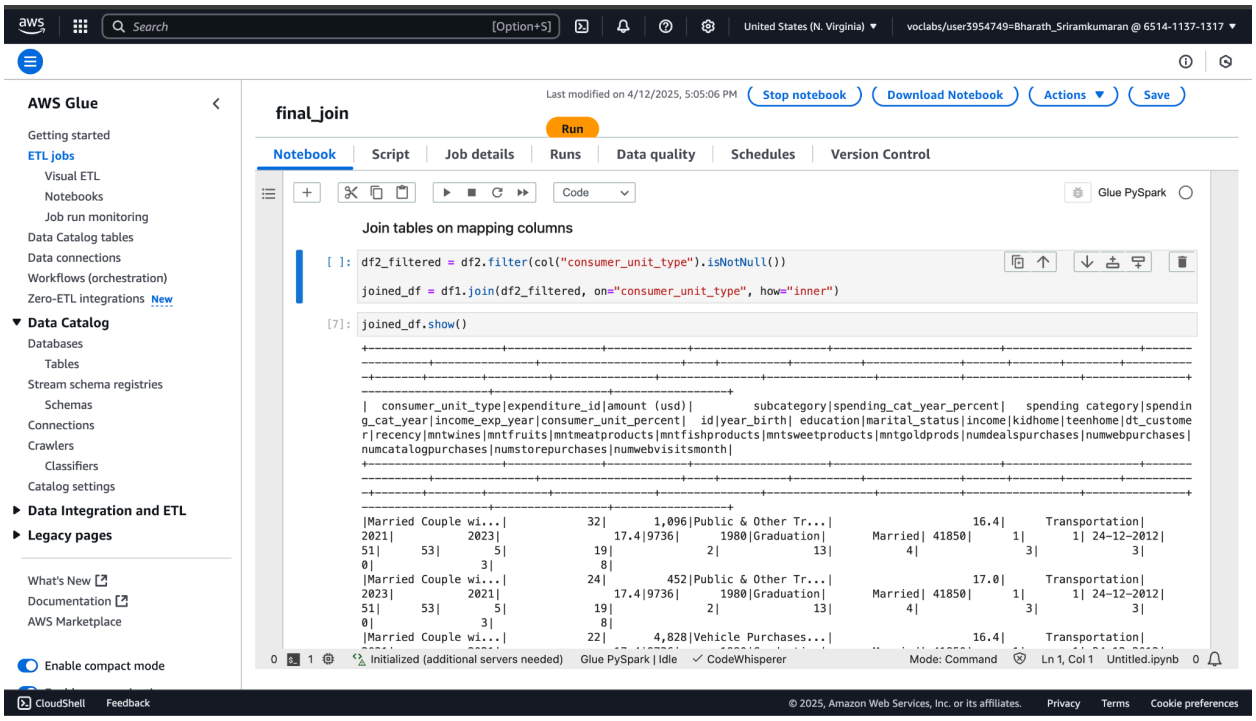
Doing the Initial join of the tables avg_income_expenditures_2021_23 , ConsumerUnit_Share_AvgExpend, and PercDist_SpendingCat, each of which has been cleaned and transformed.



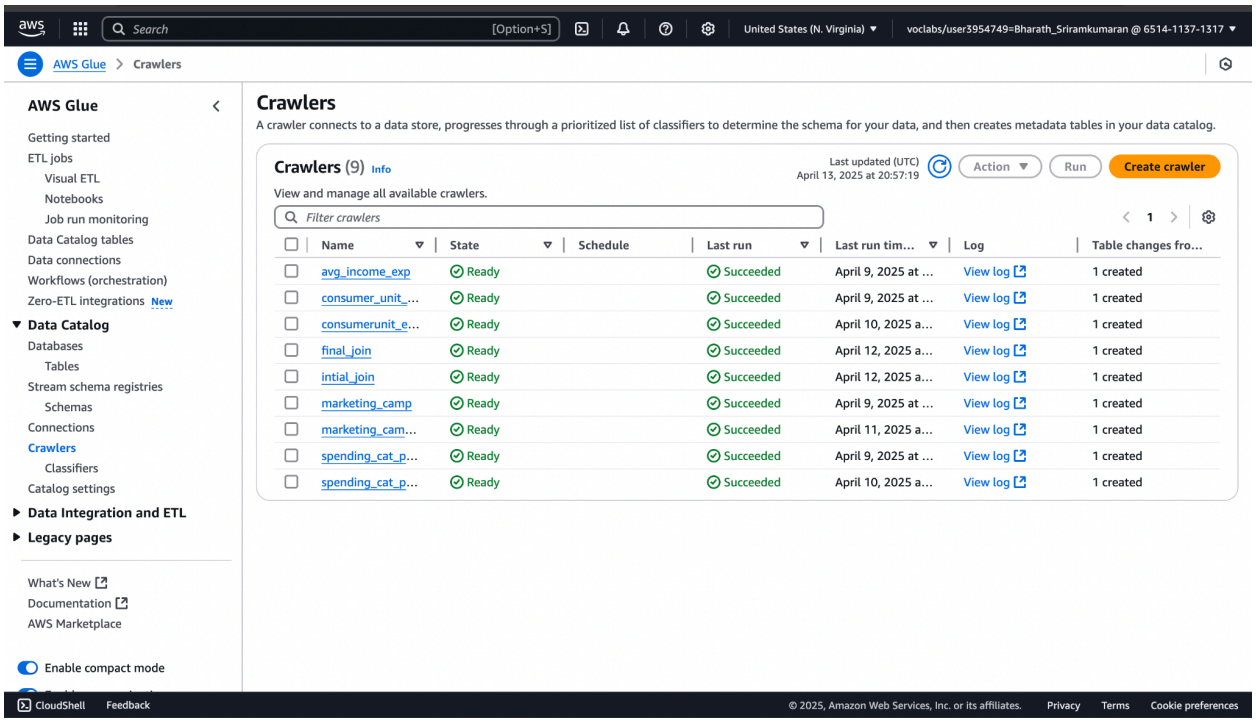


Doing the Final join, in which the initial join and the marketing campaign table are joined together. The mapping of the join is shown.

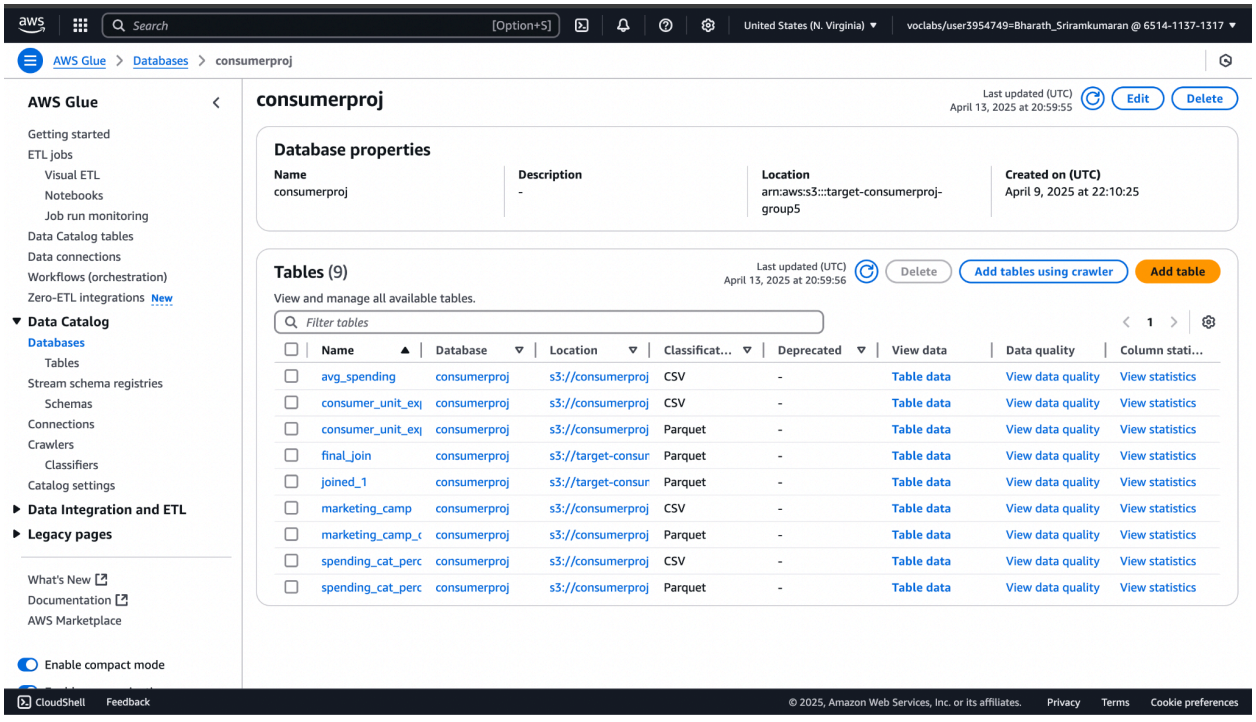




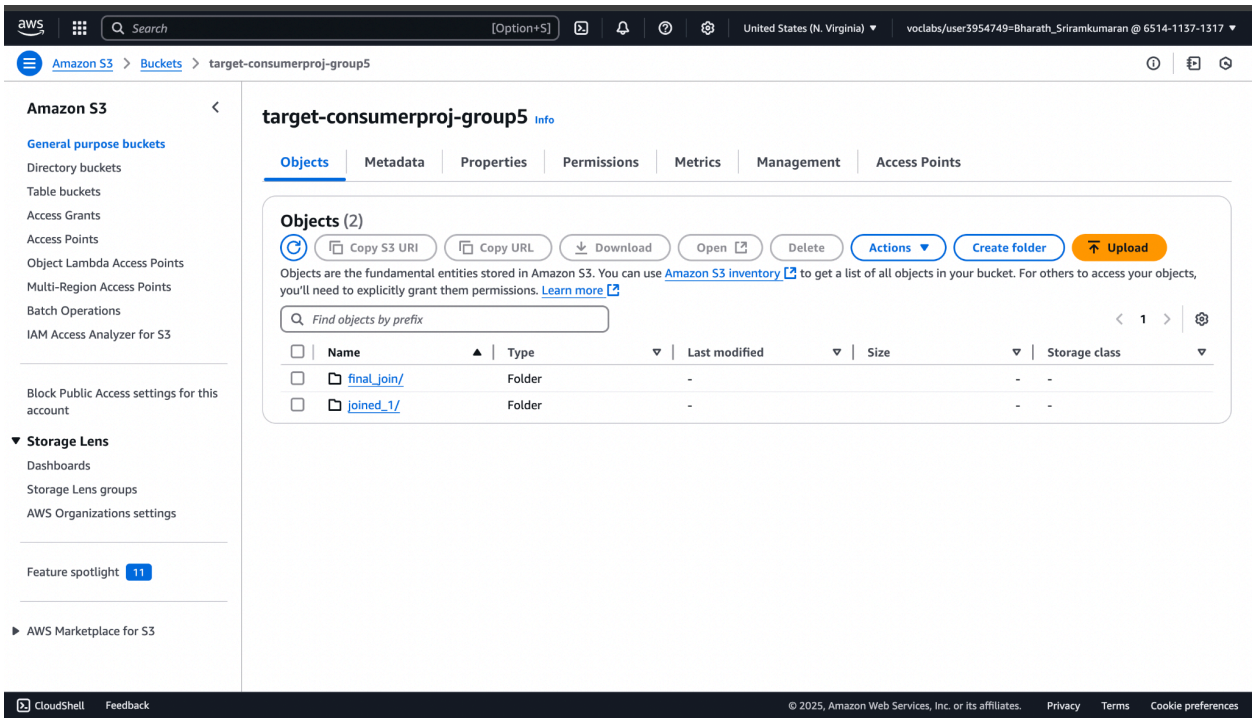
AWS crawlers are used to load into a Glue Database to visualize the tables using Athena.



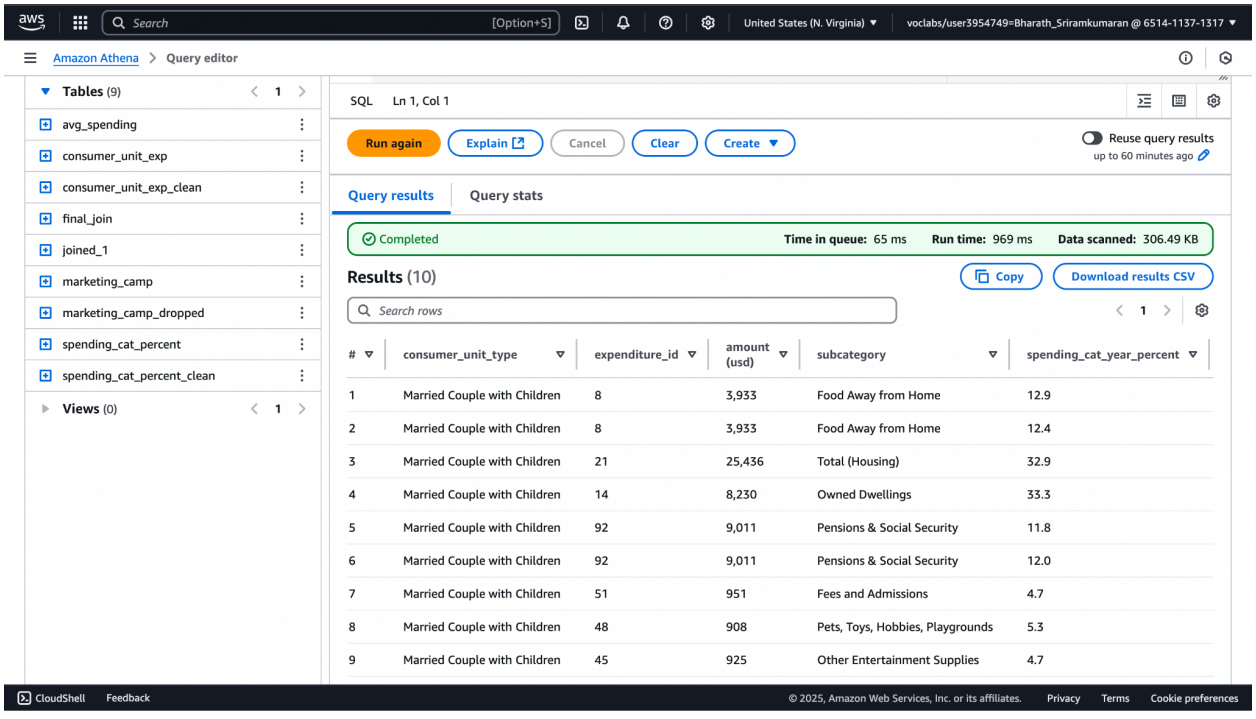
All the tables are in a Glue database called consumerproj.



The final Target Bucket where the joined tables are stored in with a folder for each to be able to retrieve the schema.



Visualizing the final table using Athena.



#	consumer_unit_type	expenditure_id	amount (usd)	subcategory	spending_cat_year_percent
1	Married Couple with Children	8	3,933	Food Away from Home	12.9
2	Married Couple with Children	8	3,933	Food Away from Home	12.4
3	Married Couple with Children	21	25,436	Total (Housing)	32.9
4	Married Couple with Children	14	8,230	Owned Dwellings	33.3
5	Married Couple with Children	92	9,011	Pensions & Social Security	11.8
6	Married Couple with Children	92	9,011	Pensions & Social Security	12.0
7	Married Couple with Children	51	951	Fees and Admissions	4.7
8	Married Couple with Children	48	908	Pets, Toys, Hobbies, Playgrounds	5.3
9	Married Couple with Children	45	925	Other Entertainment Supplies	4.7