**Data warehousing AWS Azure Assignment**

**25156**

**AWS Task –** username and region are visible in all screenshots in the top right

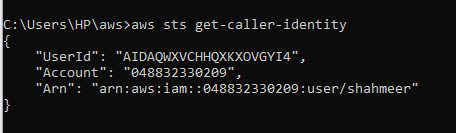
**Aws account creation**

Took some time due to payment verification



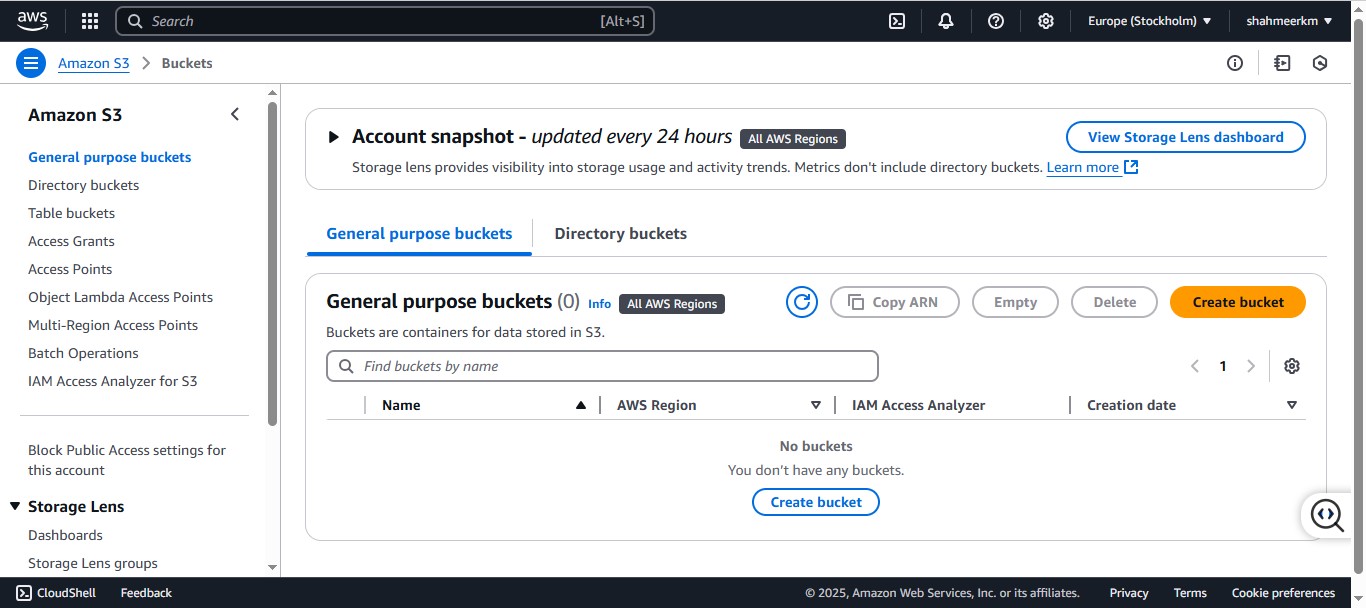
**Aws CLI configuration**

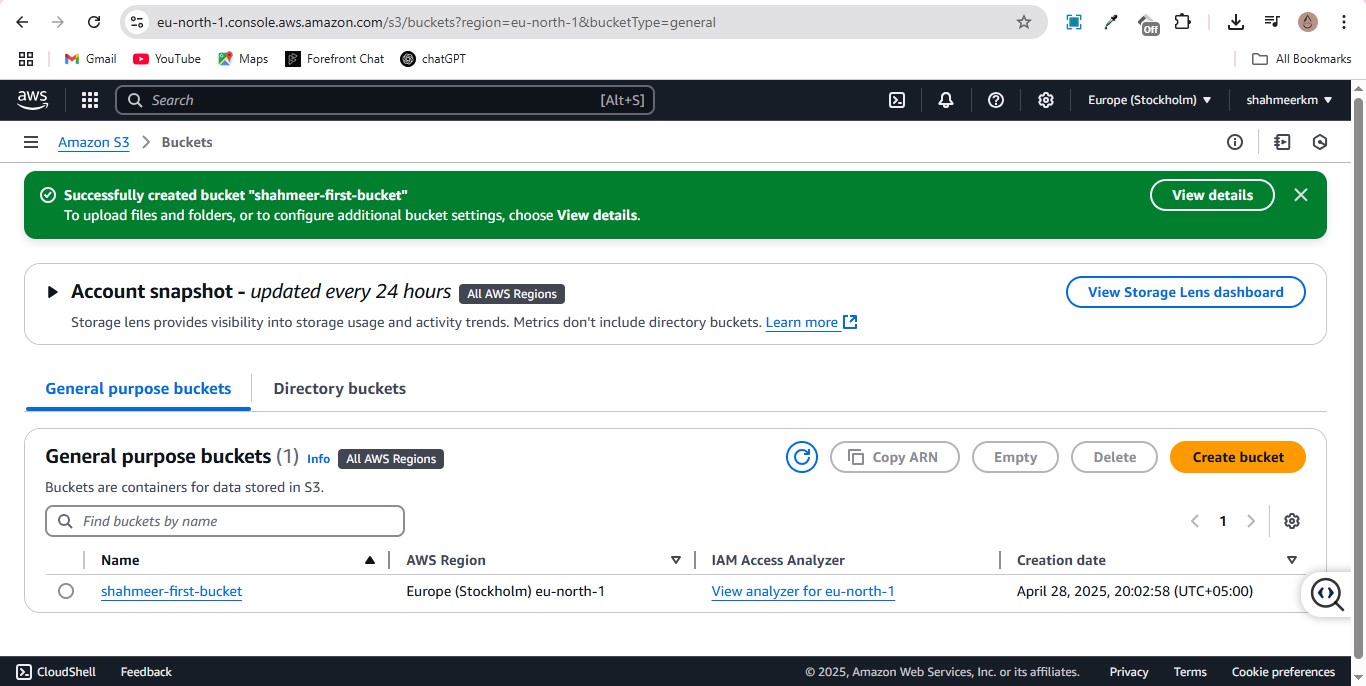
Later used for uploading json and csv files to aws



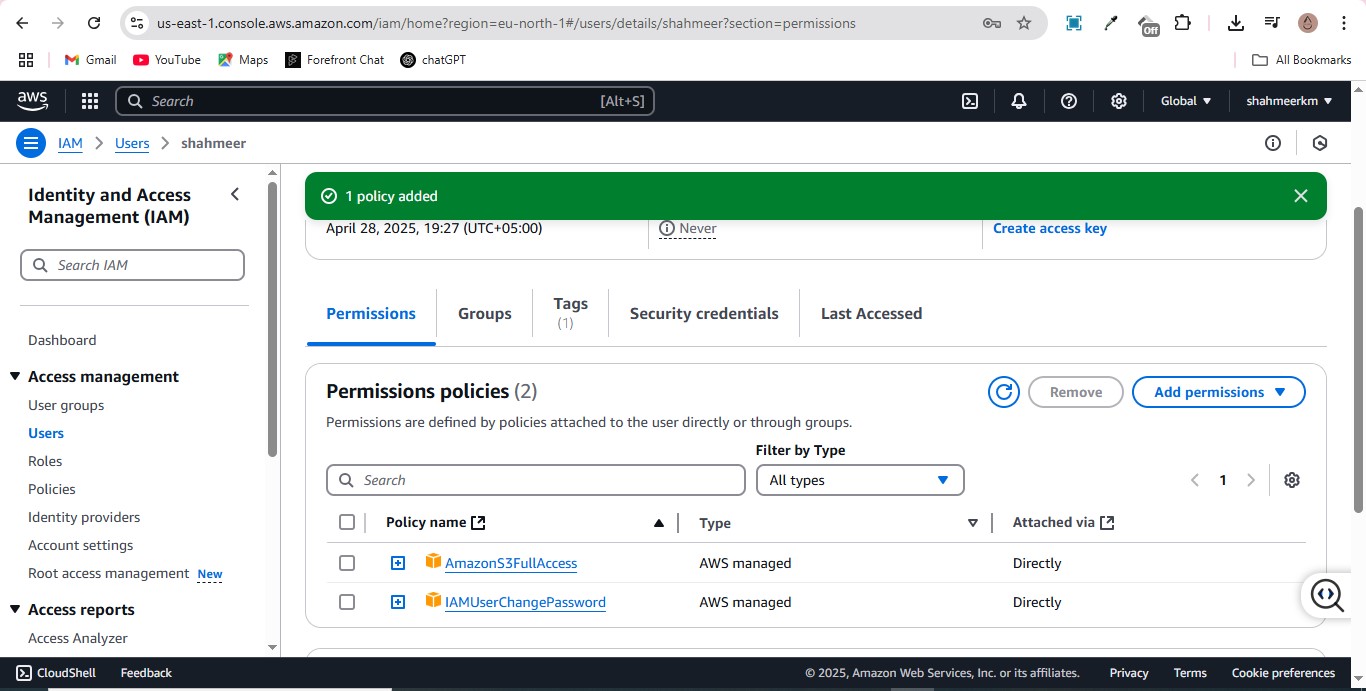
**Creating bucket**

All our relevant data has to reside in our buckets accordingly

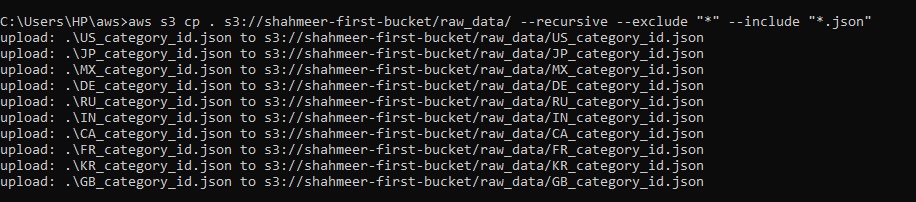
****

****

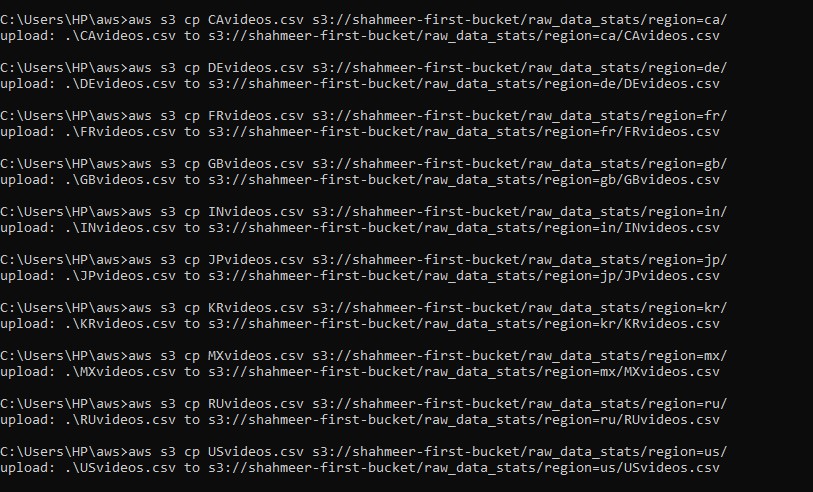
**Assigning the necessary policies to the root user account**



**Uploading files**

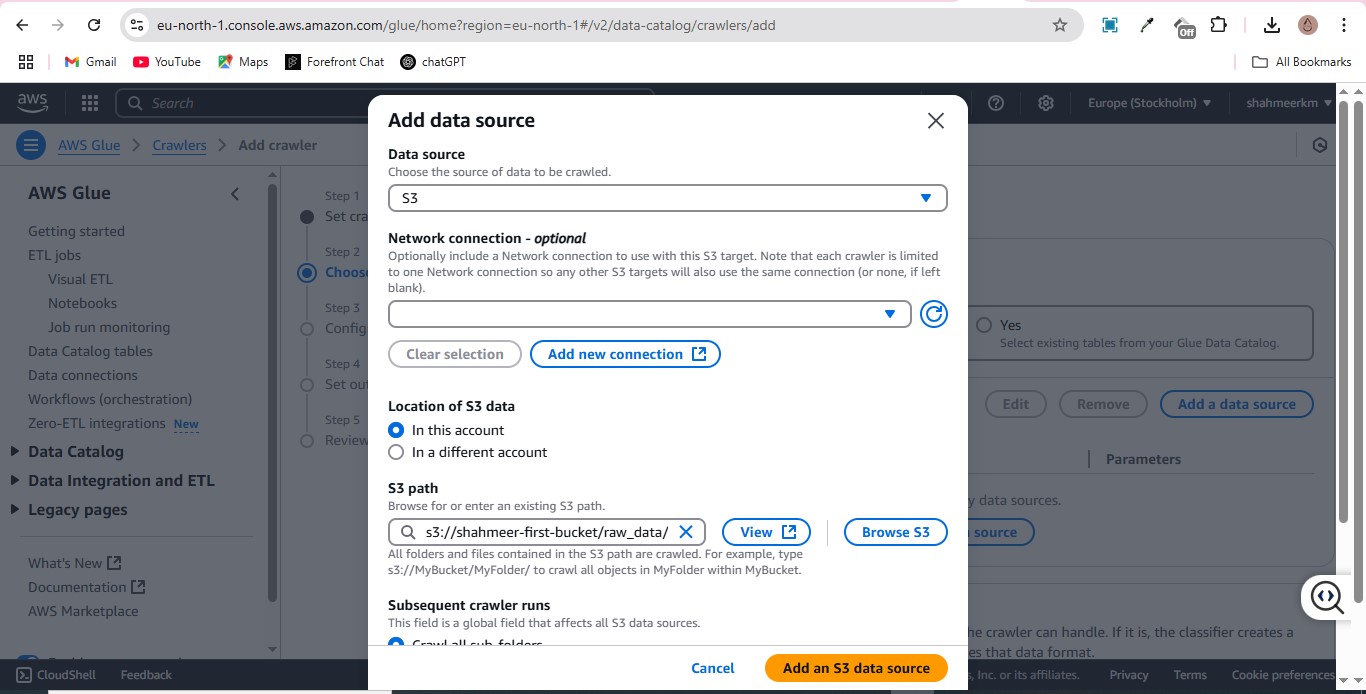
Uploading relevant json files to the AWS bucke

Uploading relevant csv files to the AWS bucket



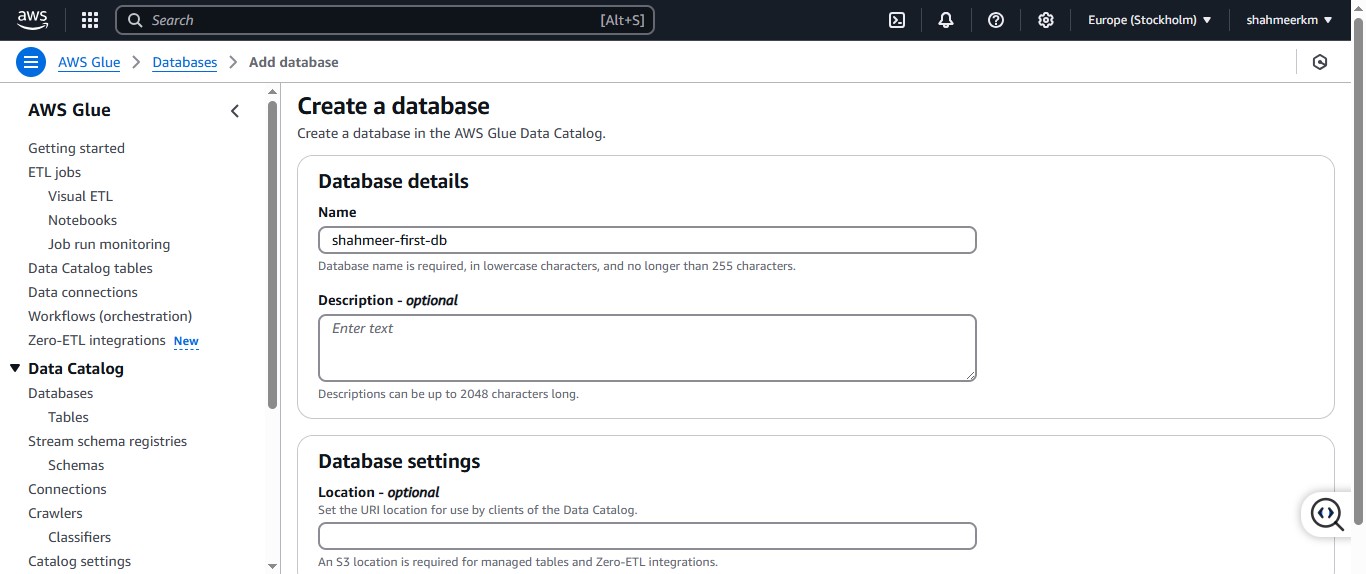
**Crawler creation and execution**

Crawlers are used to auto create tables and databases etc by accessing data sources, this one is configured to create a table for the raw\_data file

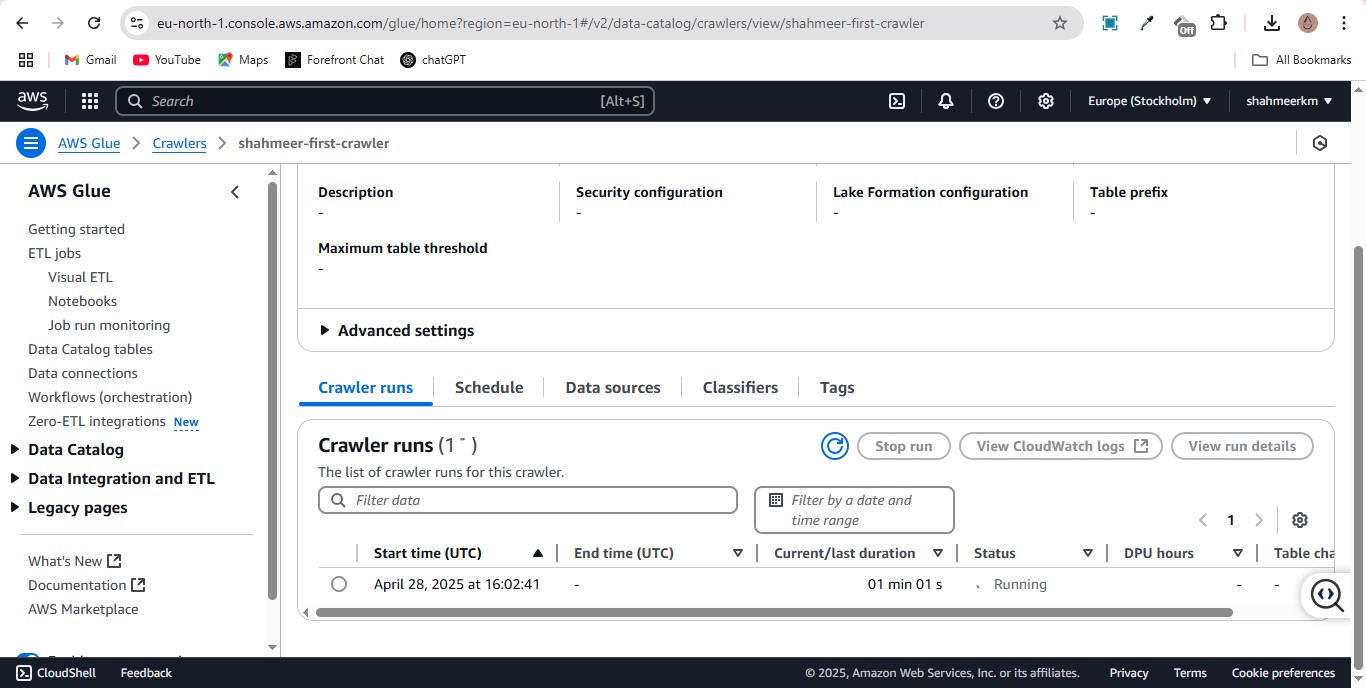


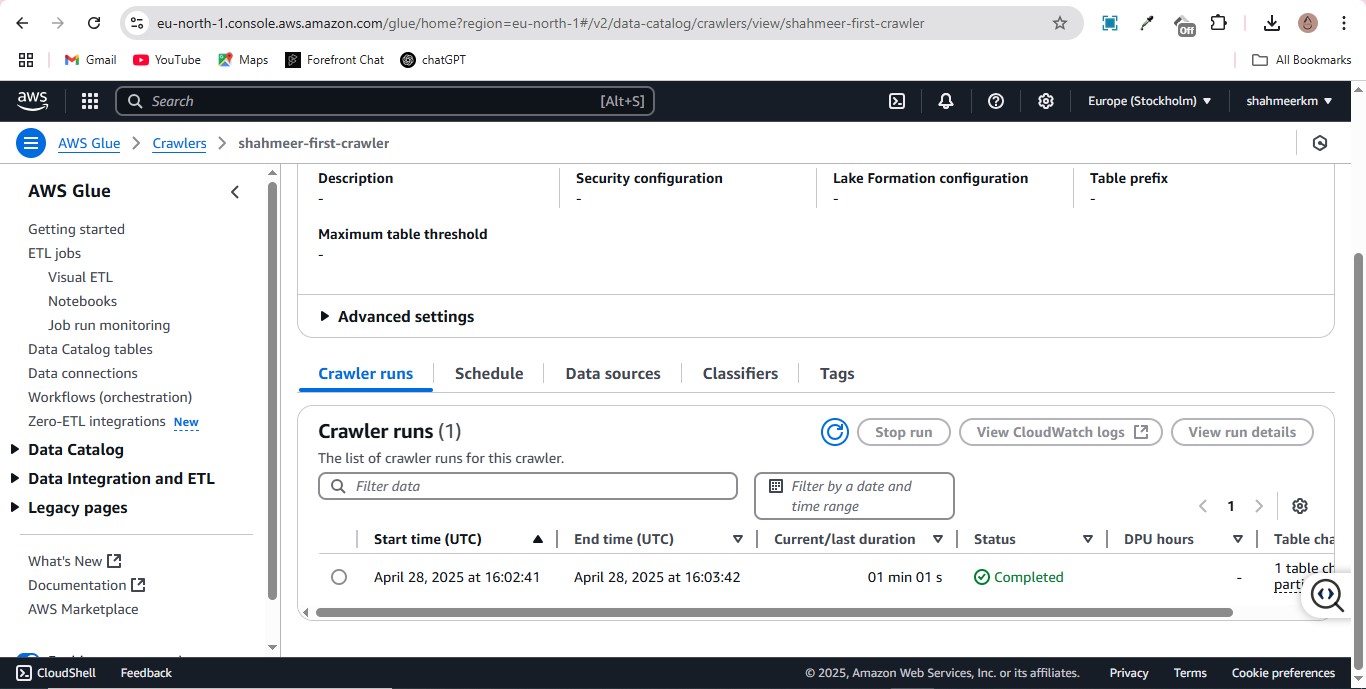
**DB creation**

DB creation is integral as it assigns a database for the crawler to write to



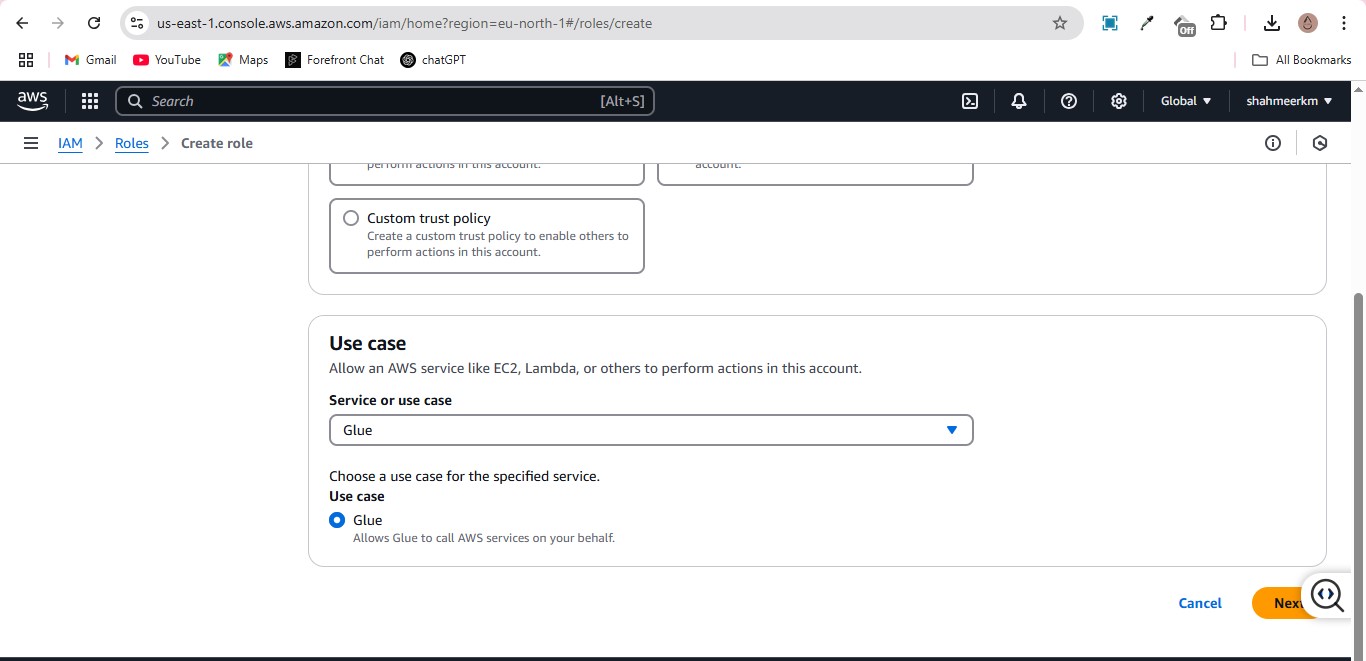
**Crawler execution**

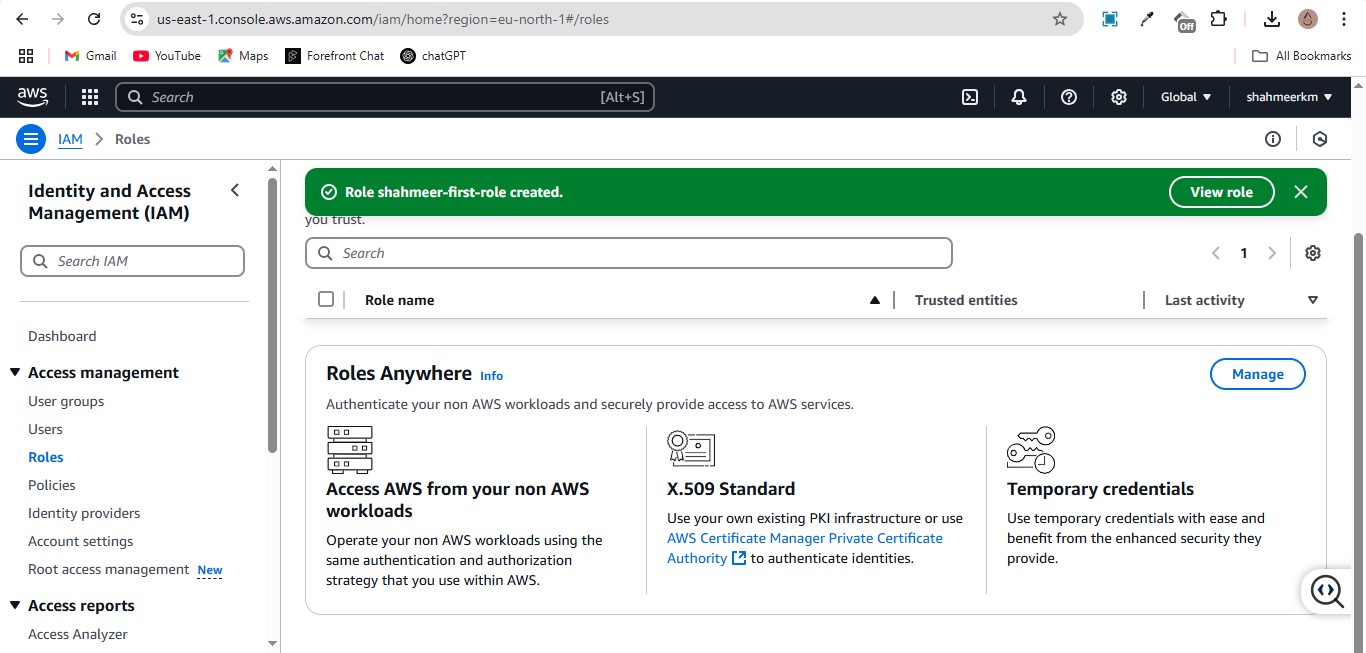




**Creation of the glue role**

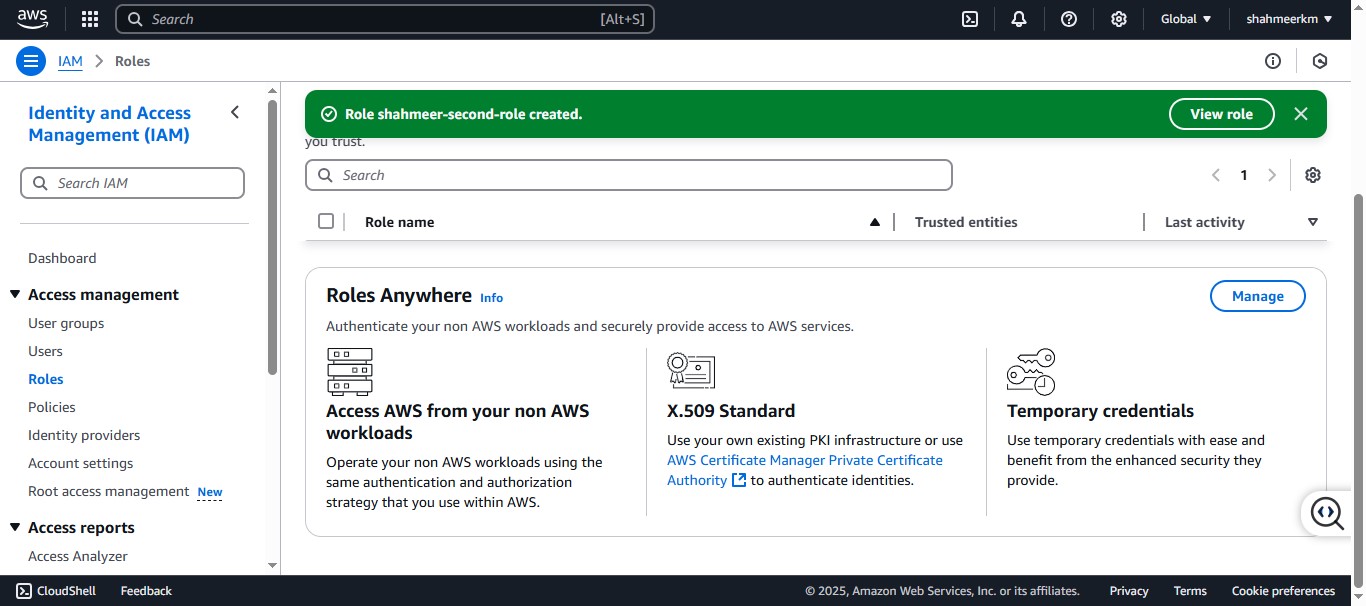
This role exists to handle all glue related tasks aka database and table management thus having glue access and s3 access



****

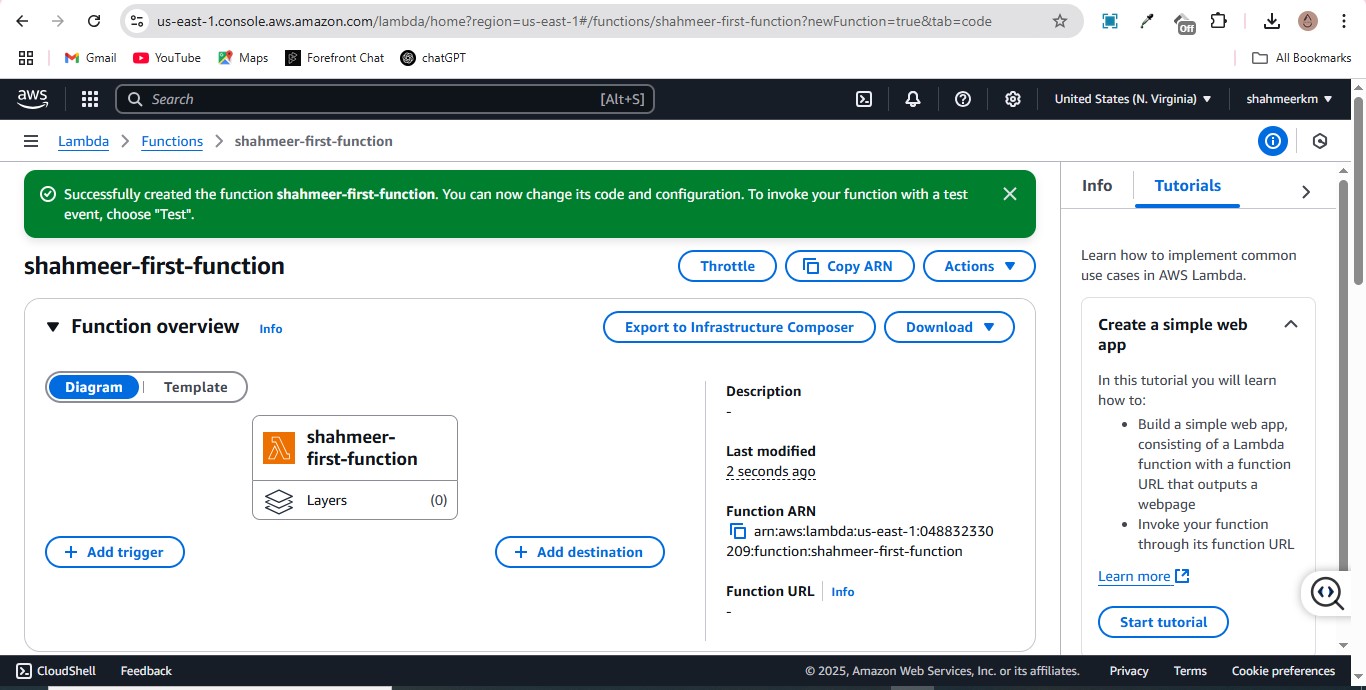
**Creation of the lambda role**

This role exists to manage lambda functions and thus has lambda access

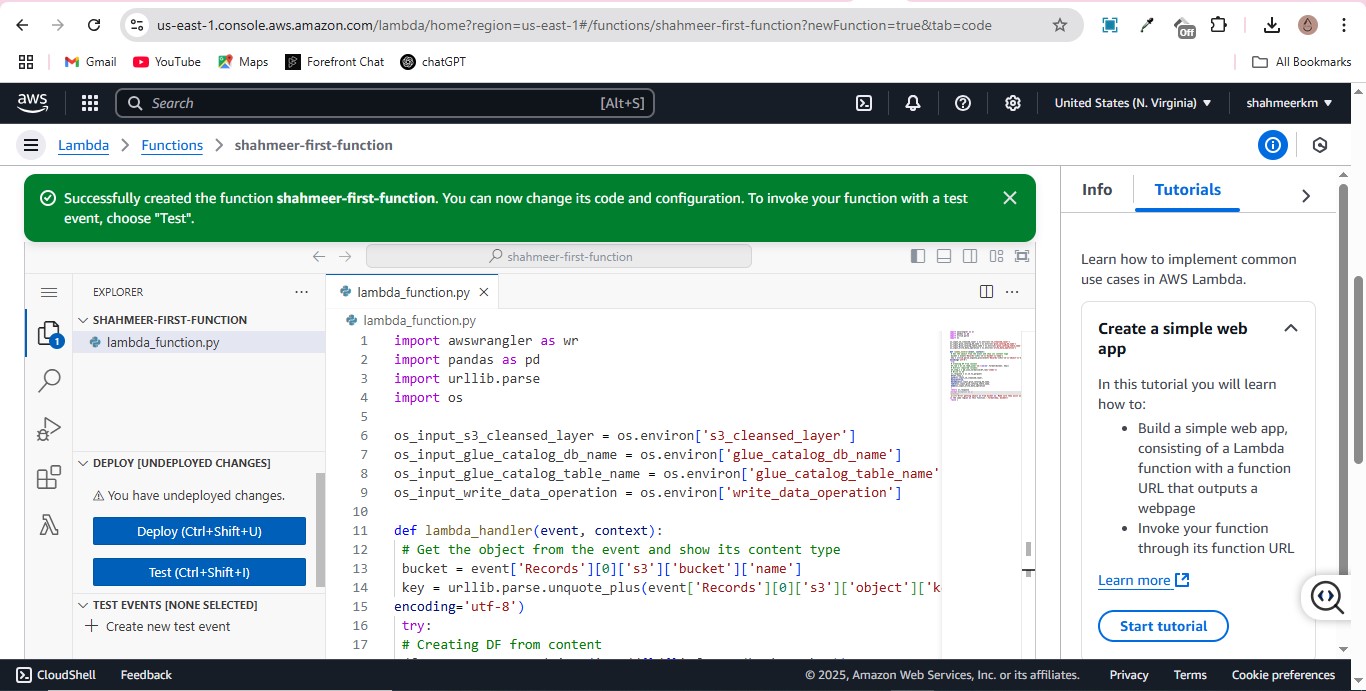


**Lambda function creation**

This function will be used to create the clean data table

****

Function code

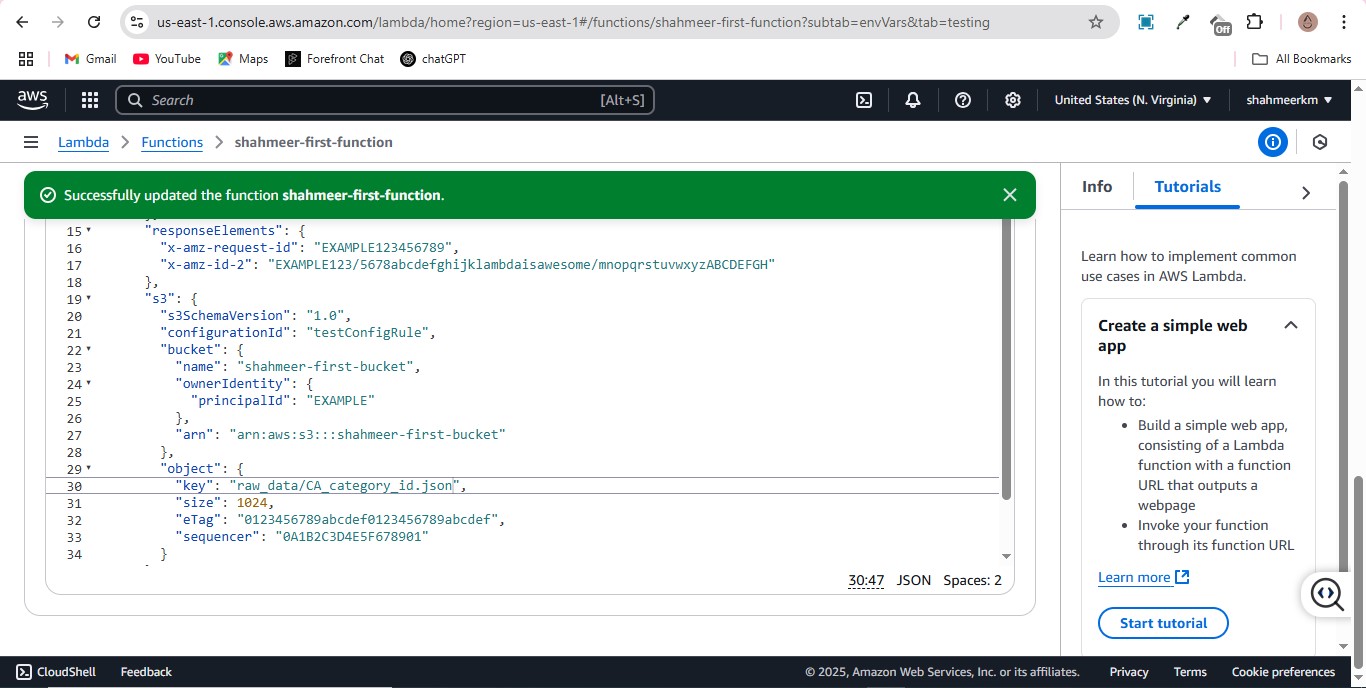


Assigning relevant layer to function

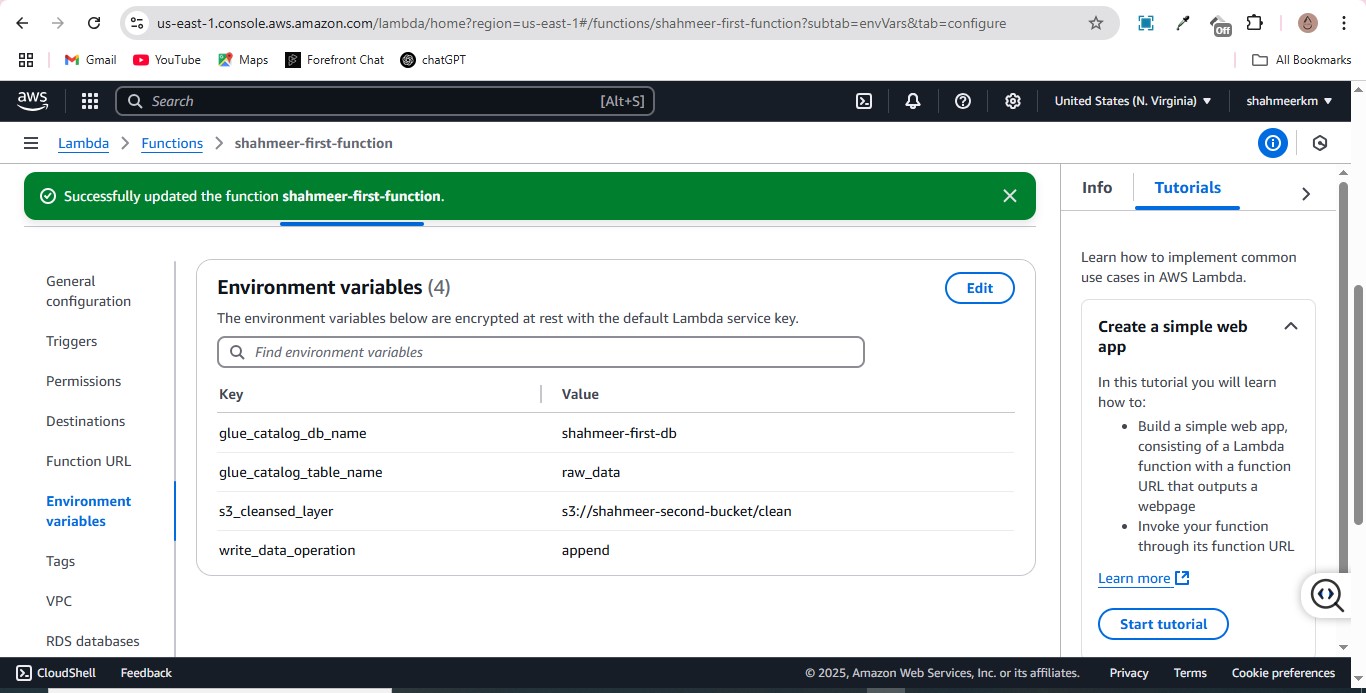


**Test function edits**

These are configured to access our bucket and the relevant Json file

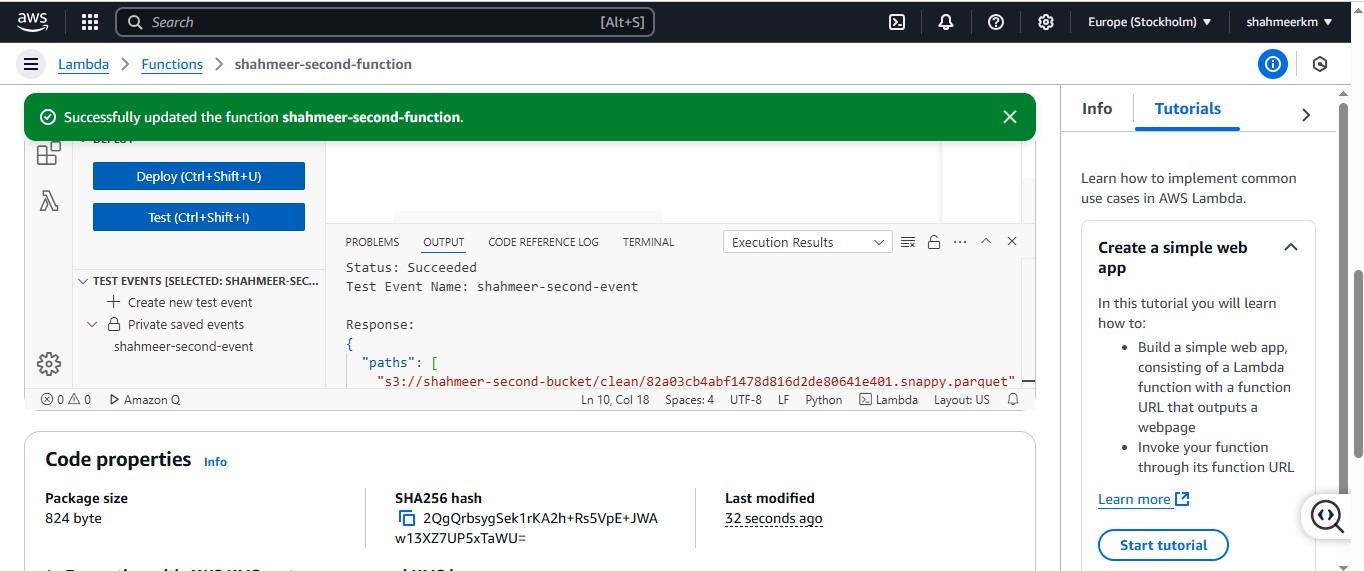


**Environment variables created**

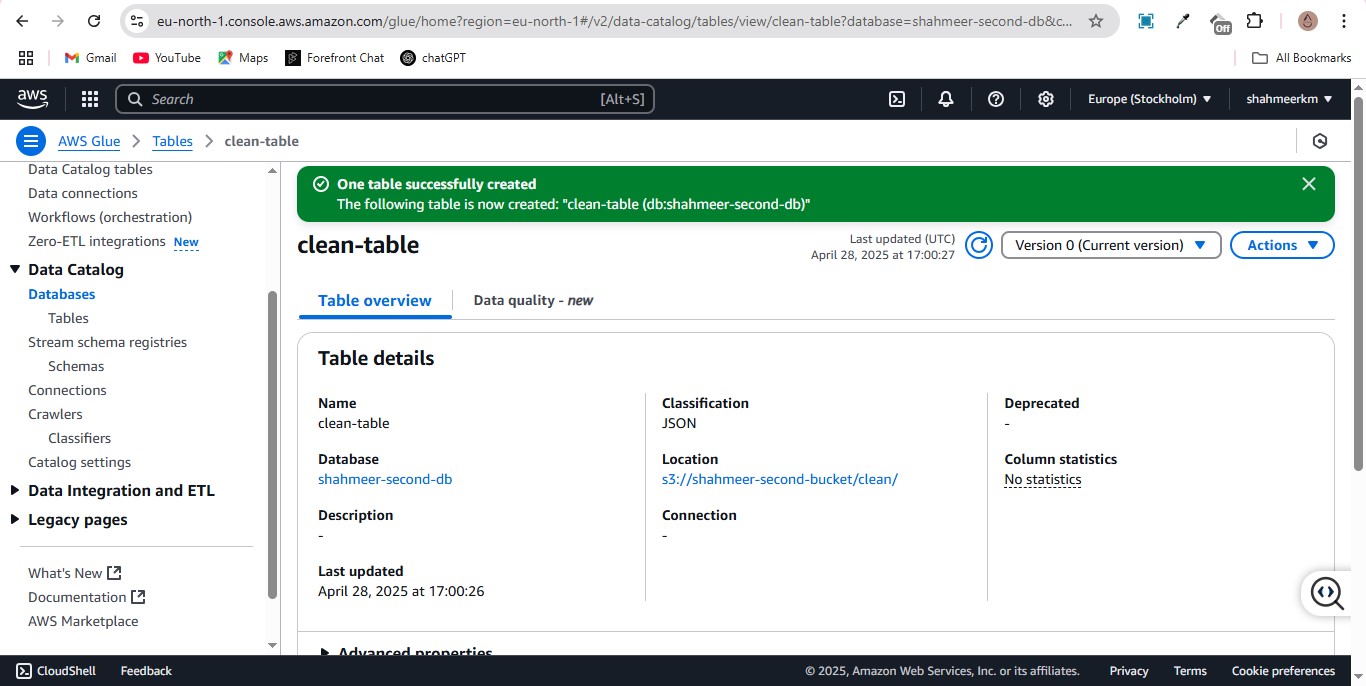


These ended up being changed, mainly the table name as initially I was creating a new table for the clean data and then running the function, this caused it to be an empty table as the function itself creates the table for you, I will refer to the issues faced at the end of the doc

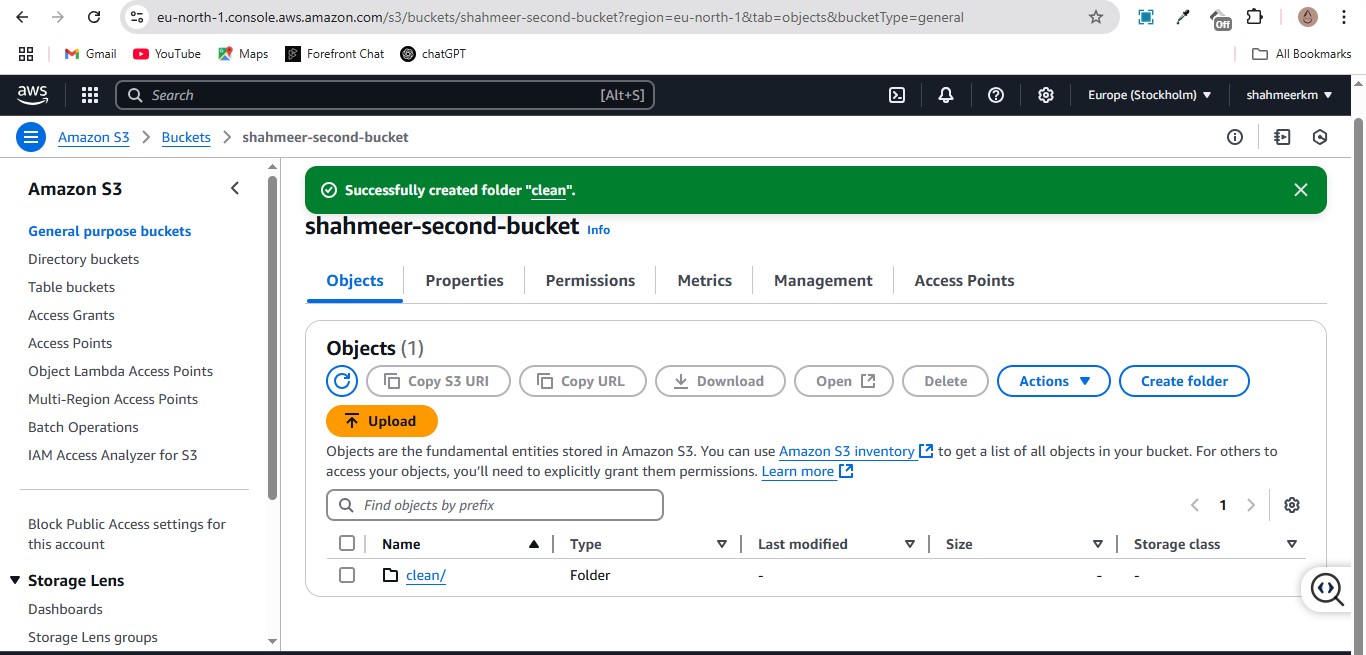
**Successful test run**



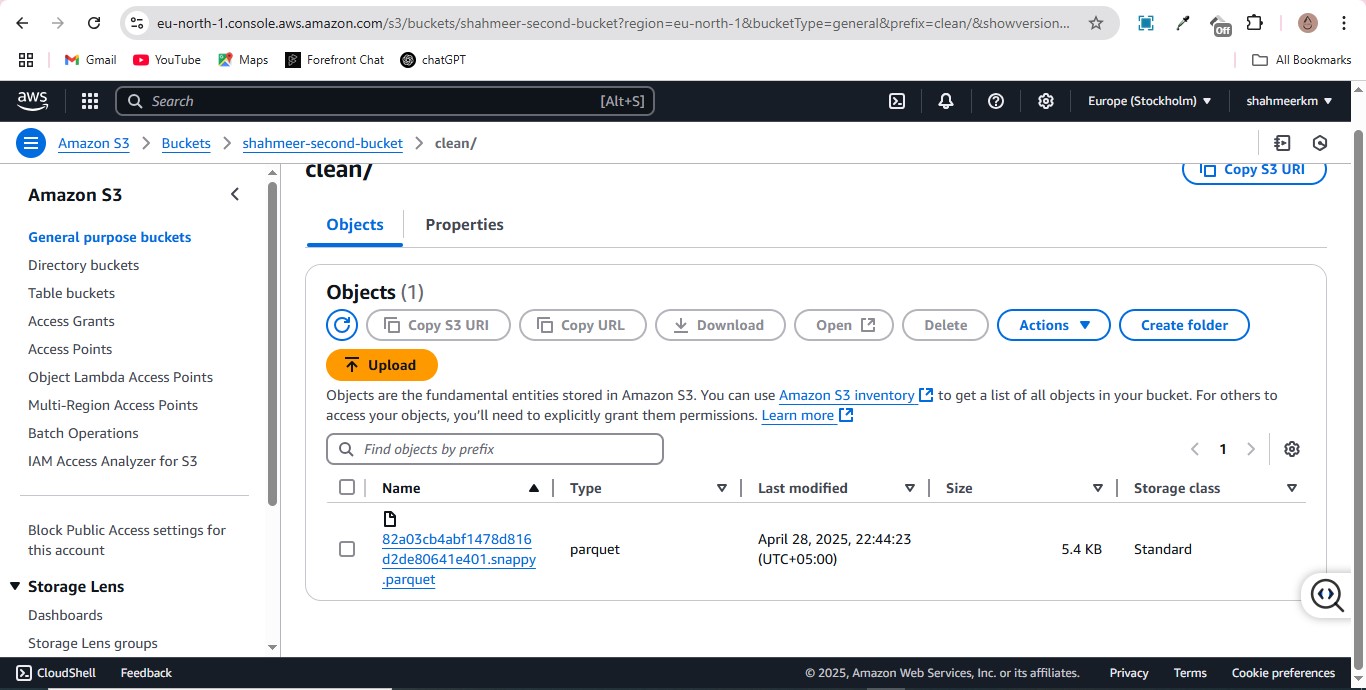
**Clean data table creation in db**



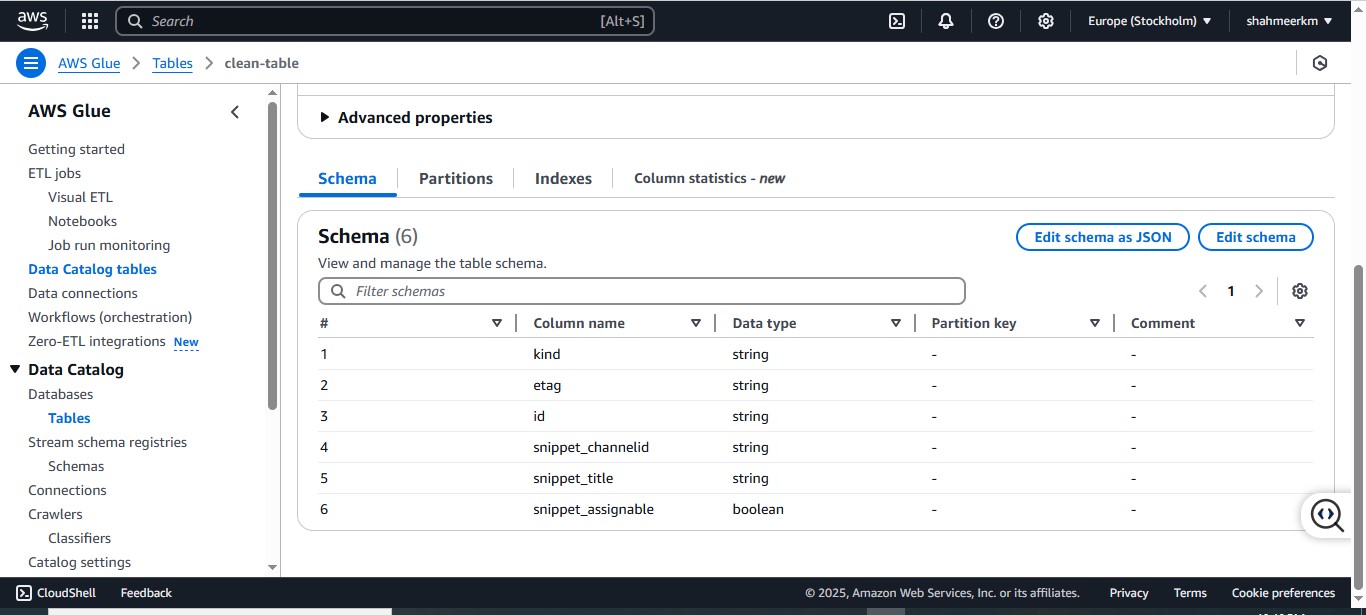
Created a new folder in bucket 2 for the new table



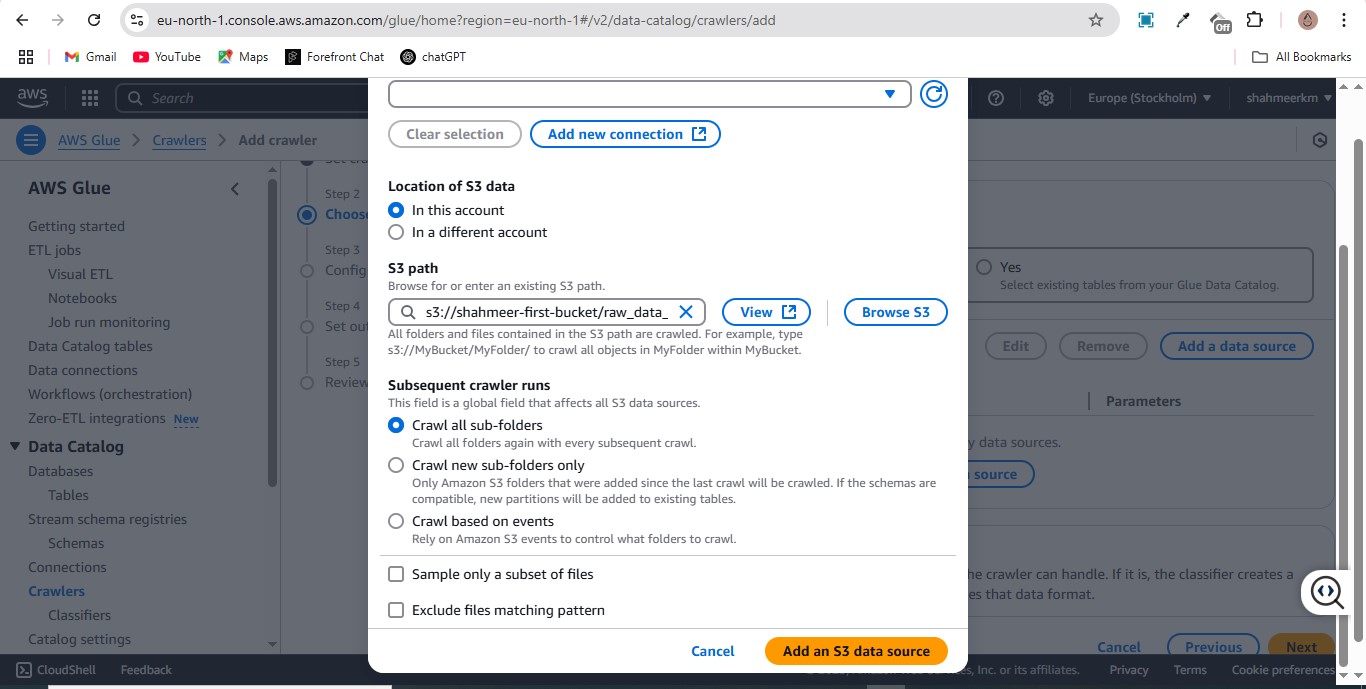
**Clean data parquet created in the secondary bucket**

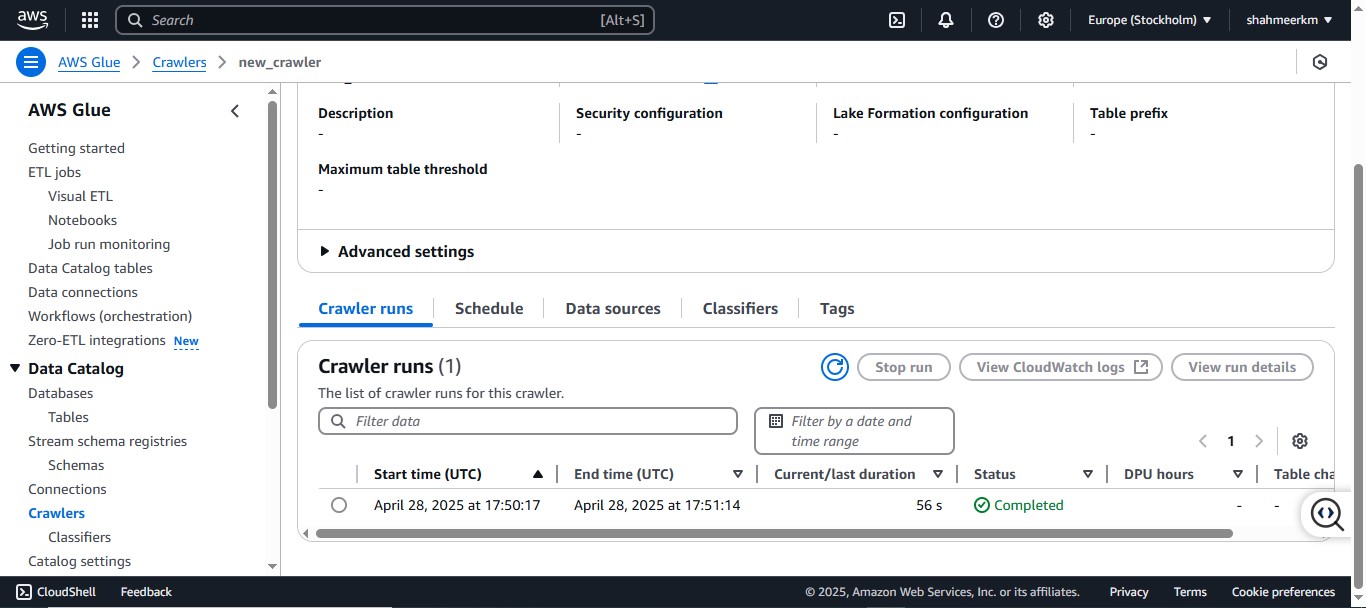


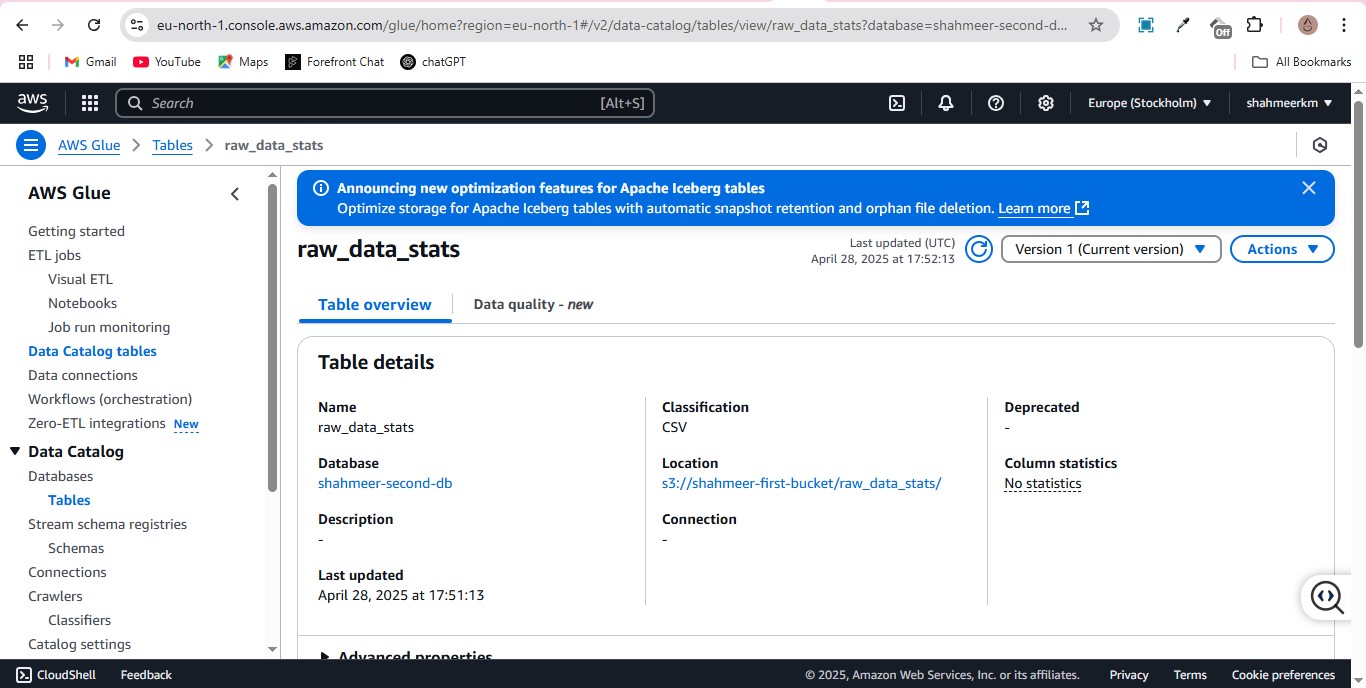
Schema for the new created clean data table



**2nd crawler created to create raw\_data\_stats table**

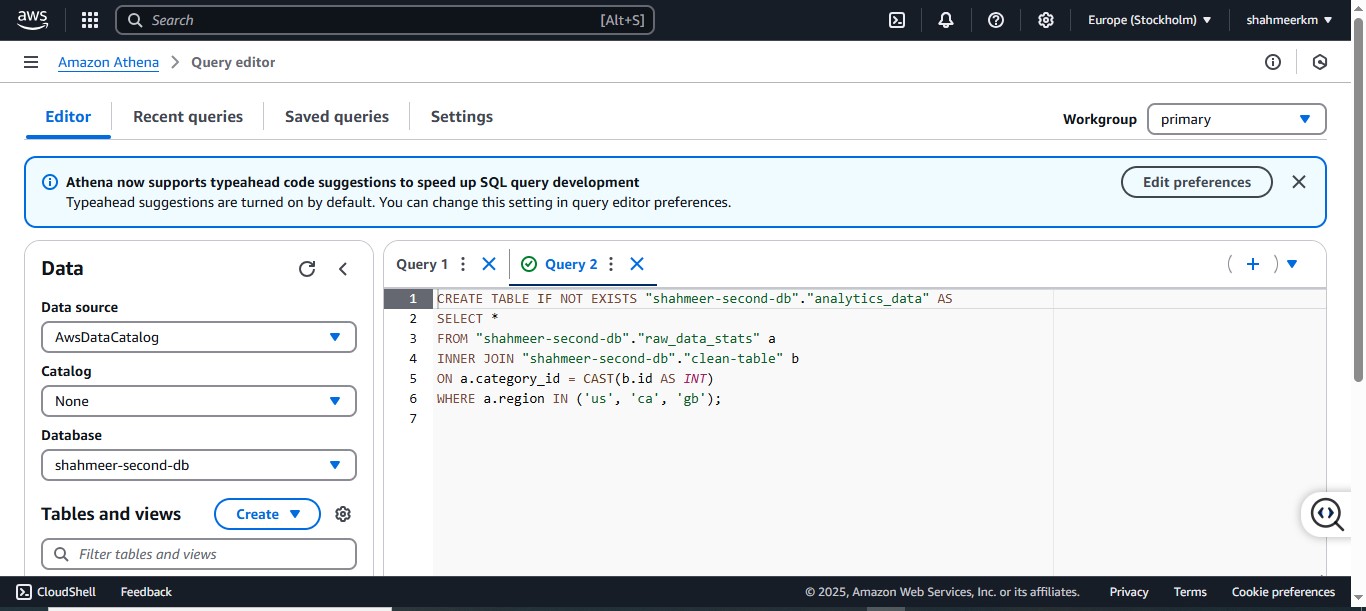
****

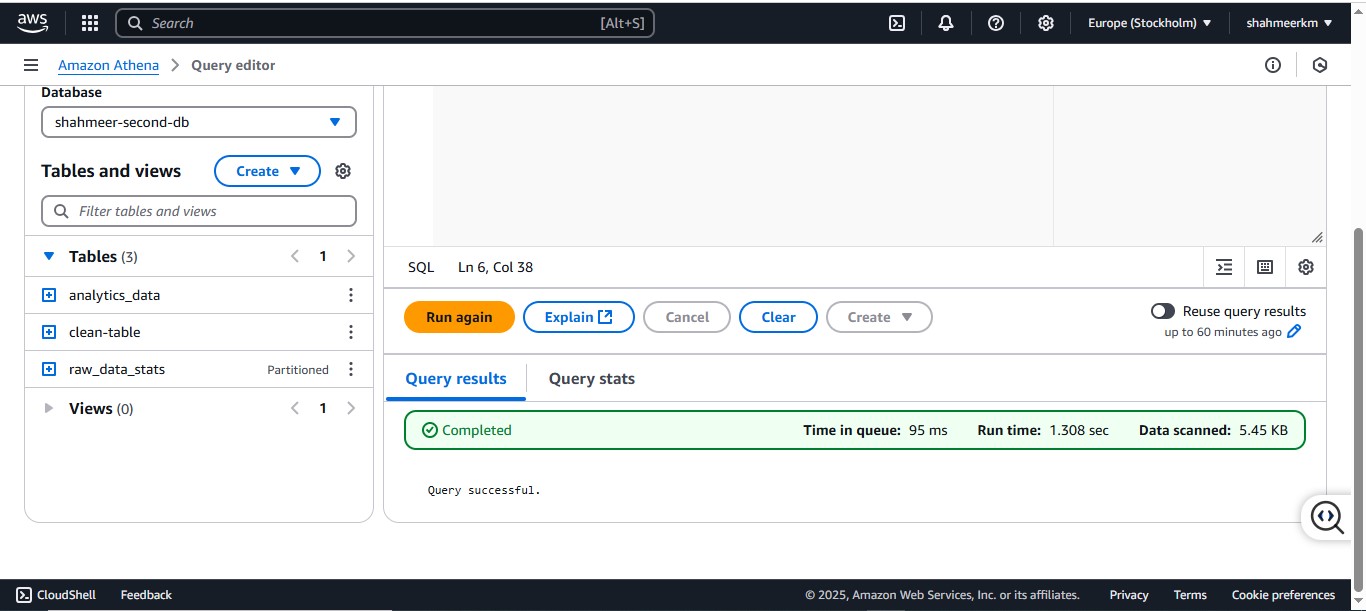
****



**Join query executed to create analytics table in Athena**

This query was a bit different later on mainly with reference to table name

****

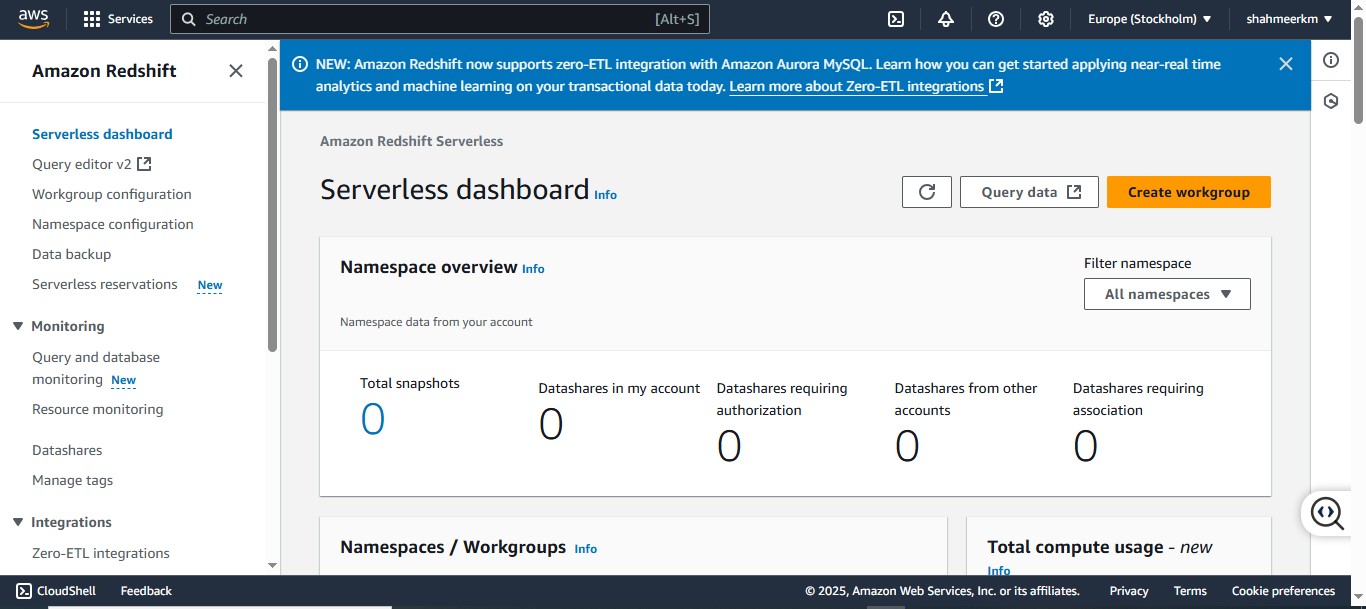


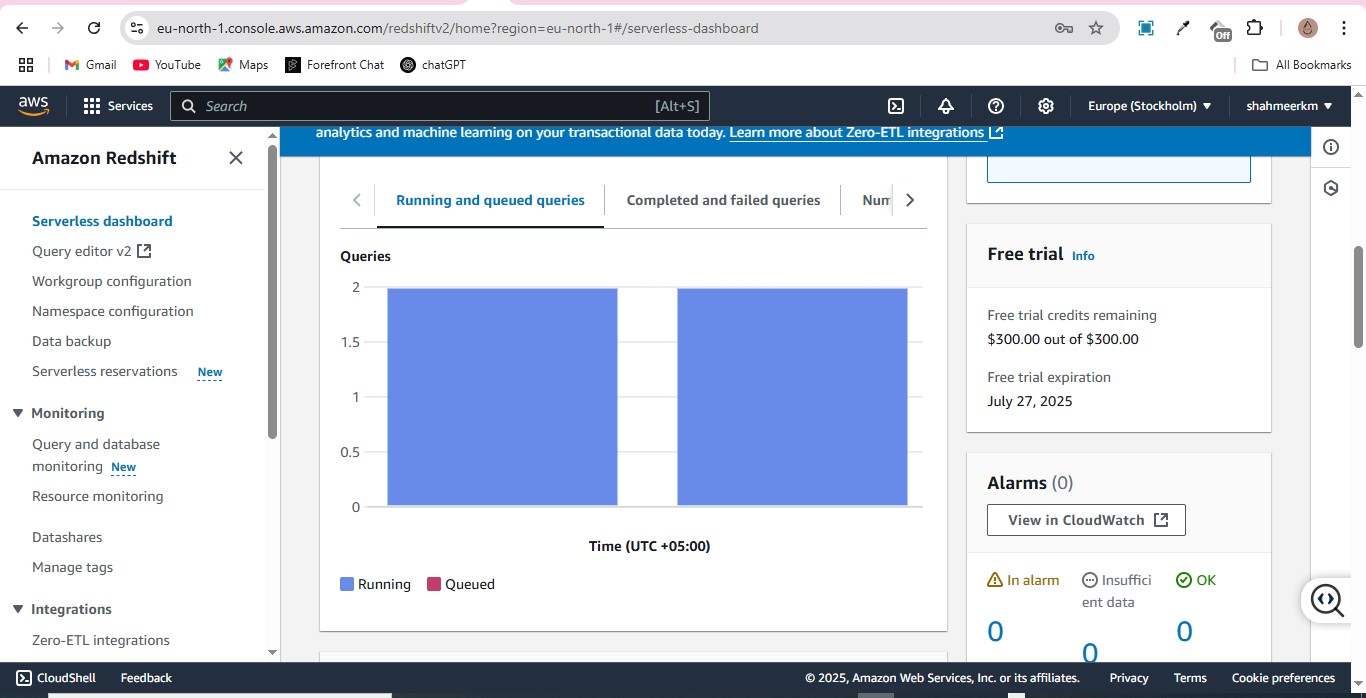
**Analytics table successfully created**

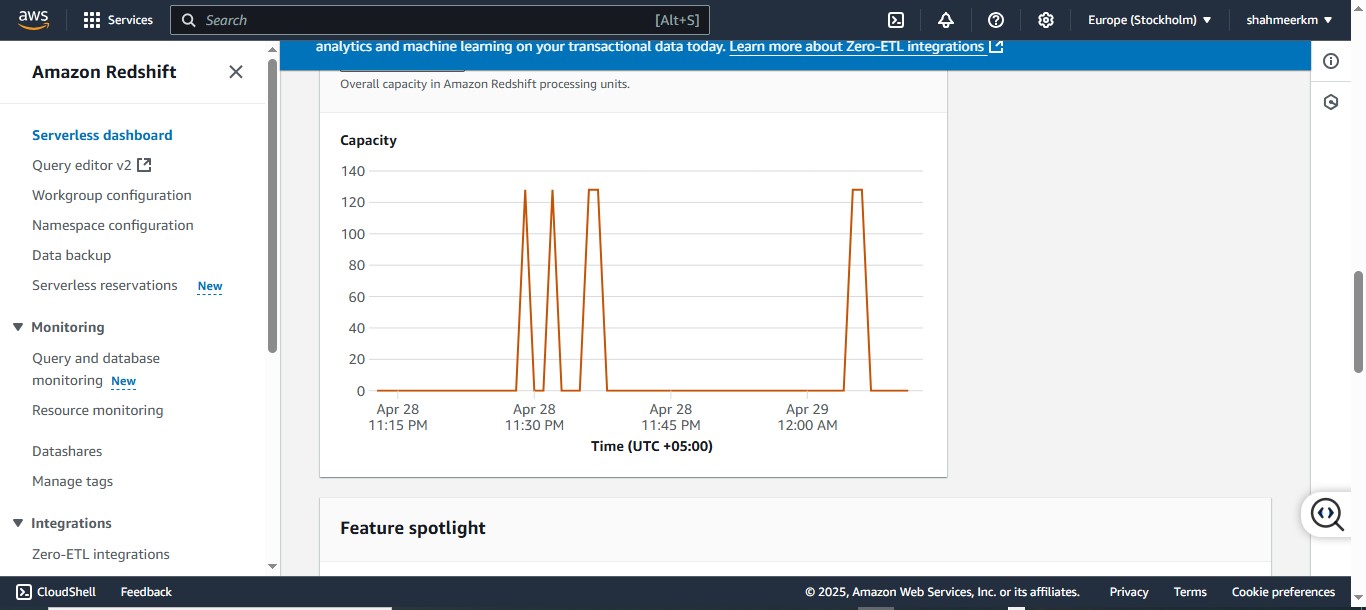


**Redshift task**

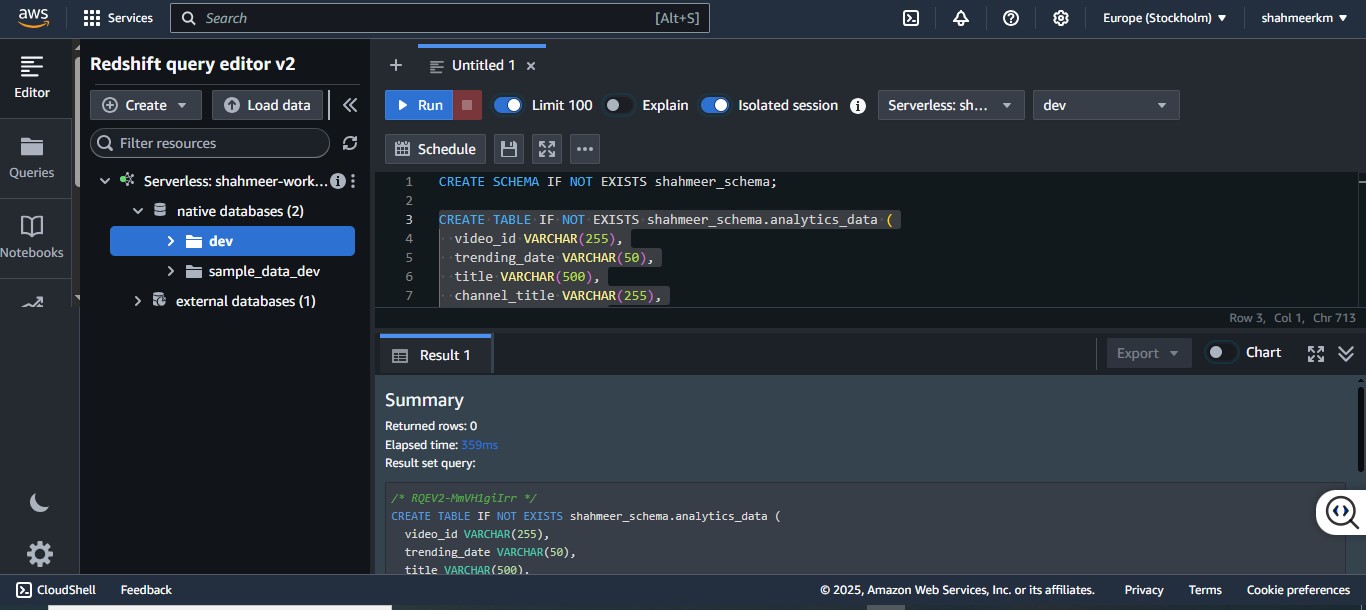
**Redshift account created**



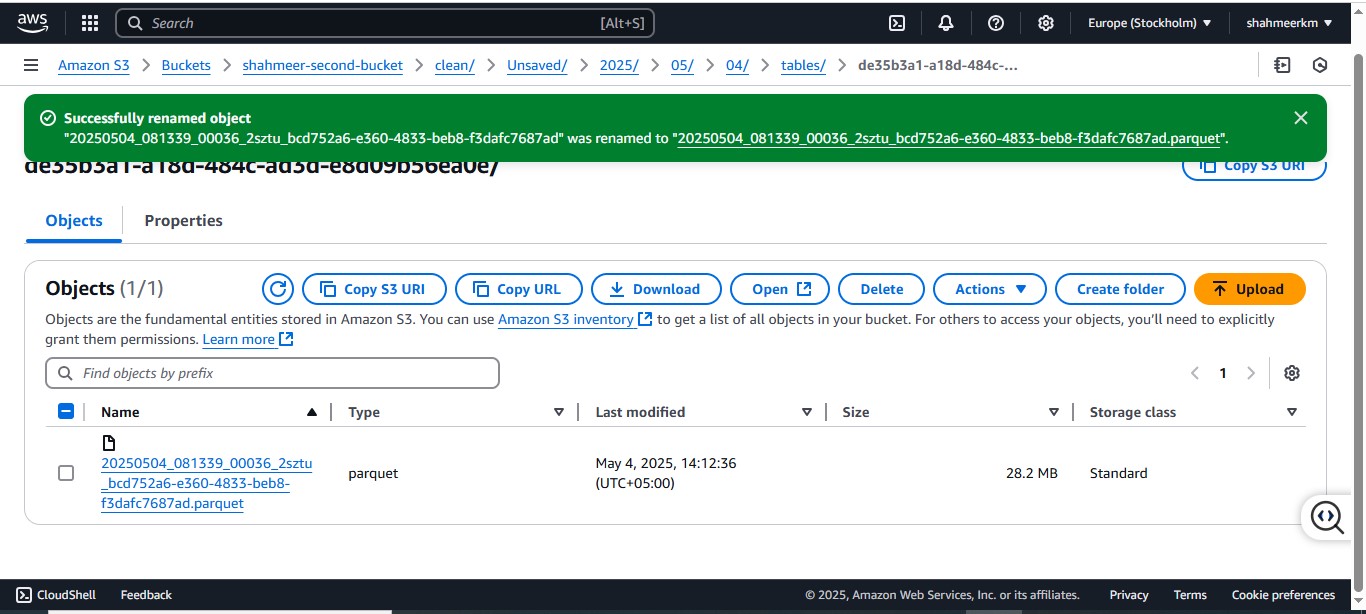




**Table schema created in redshift**



**The relevant file with the data had to be renamed to parquet format to be accessed**



**Loading relevant data into the table**

The IAM role here refers to the one with access to the buckets, the file accessed here is our analytics table

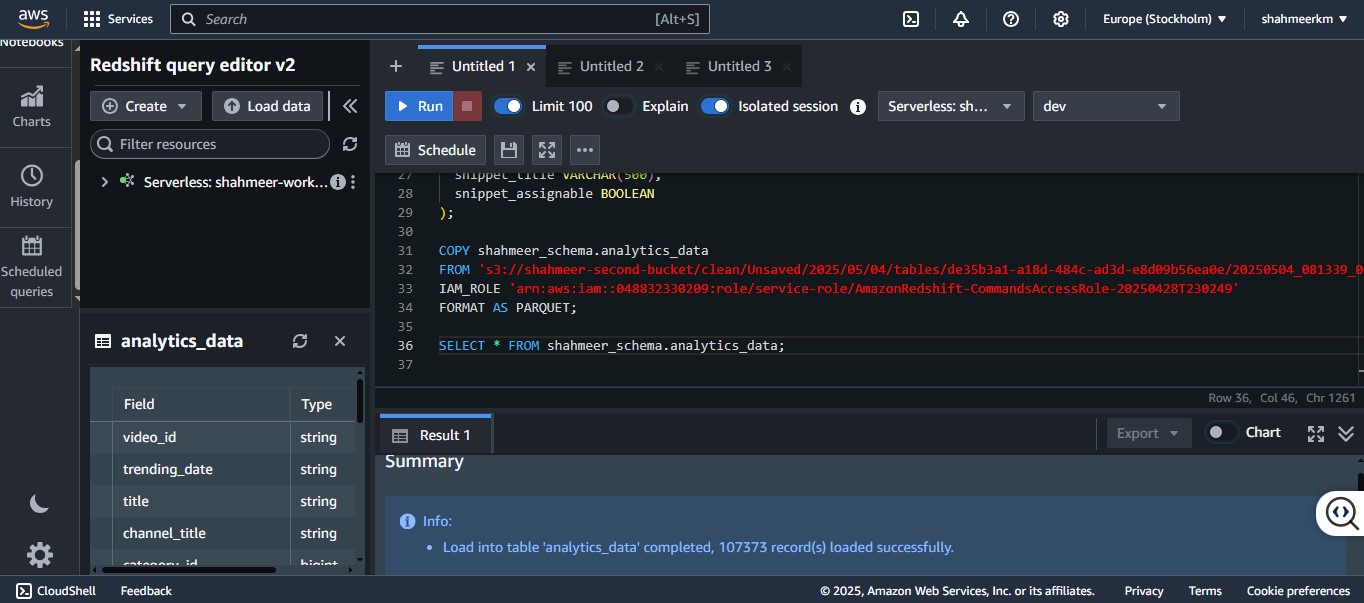
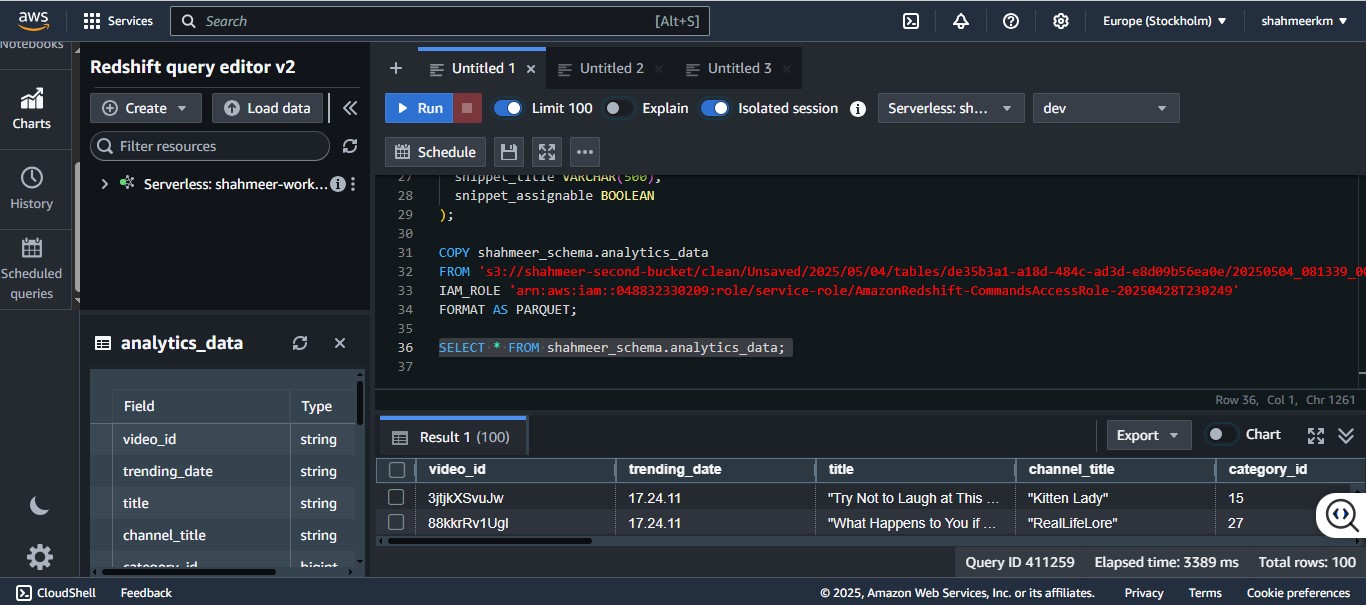


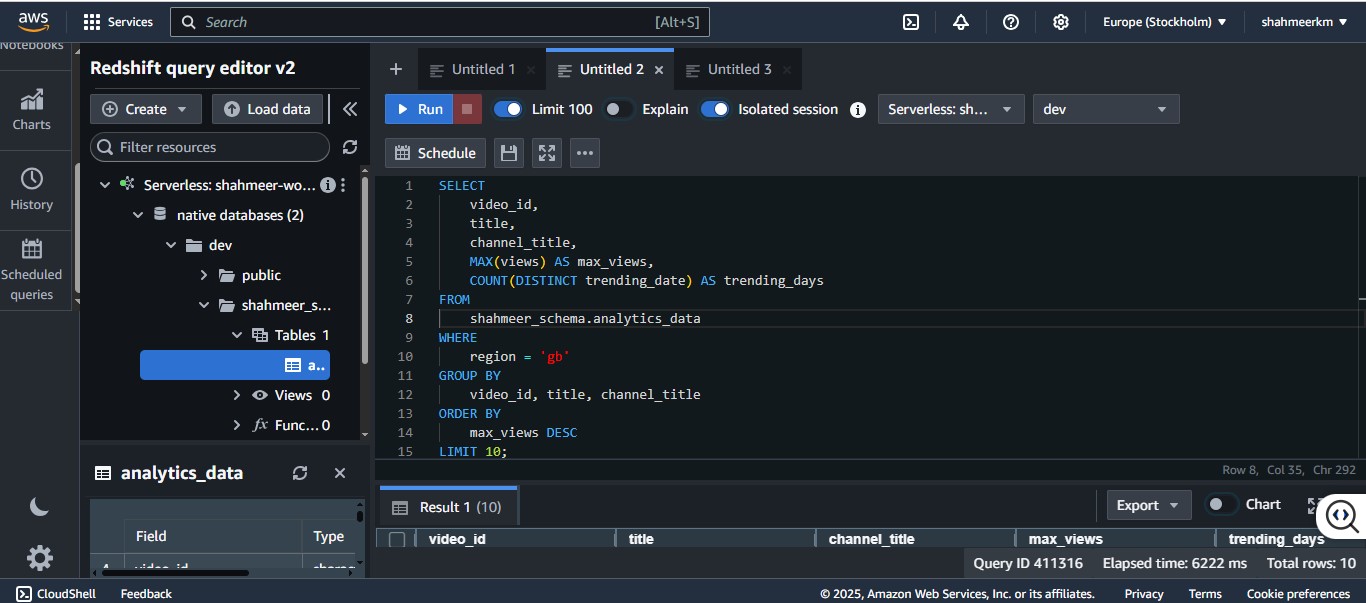
Table successfully loaded proven by a select statement



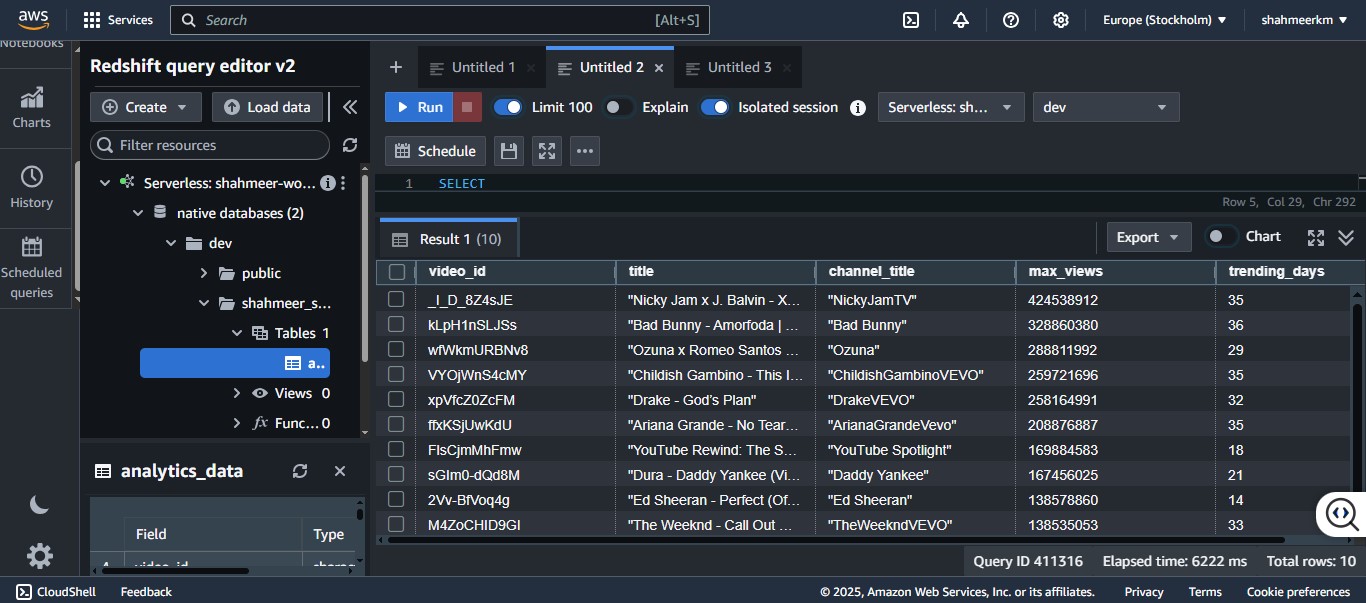
**Warehousing queries**

**Query 1**

Display the most viewd videos for the gb region and the days they remained on trending

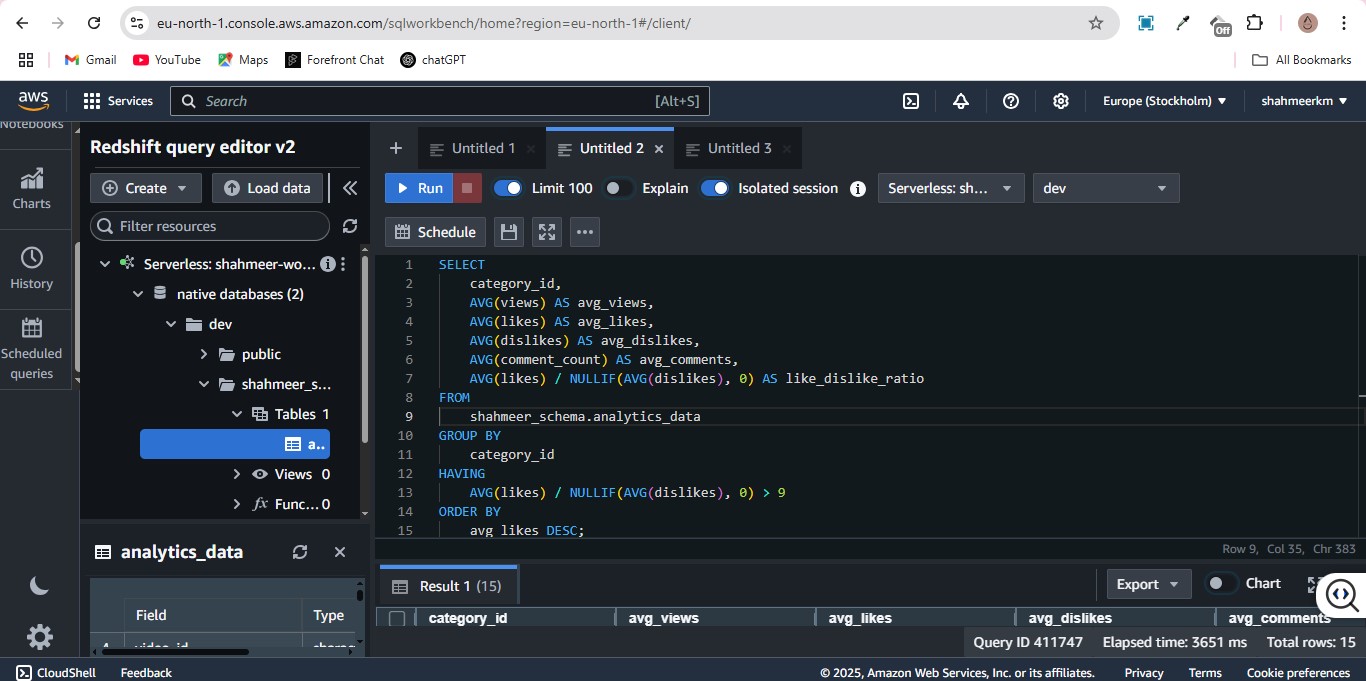


**Query 1 result**

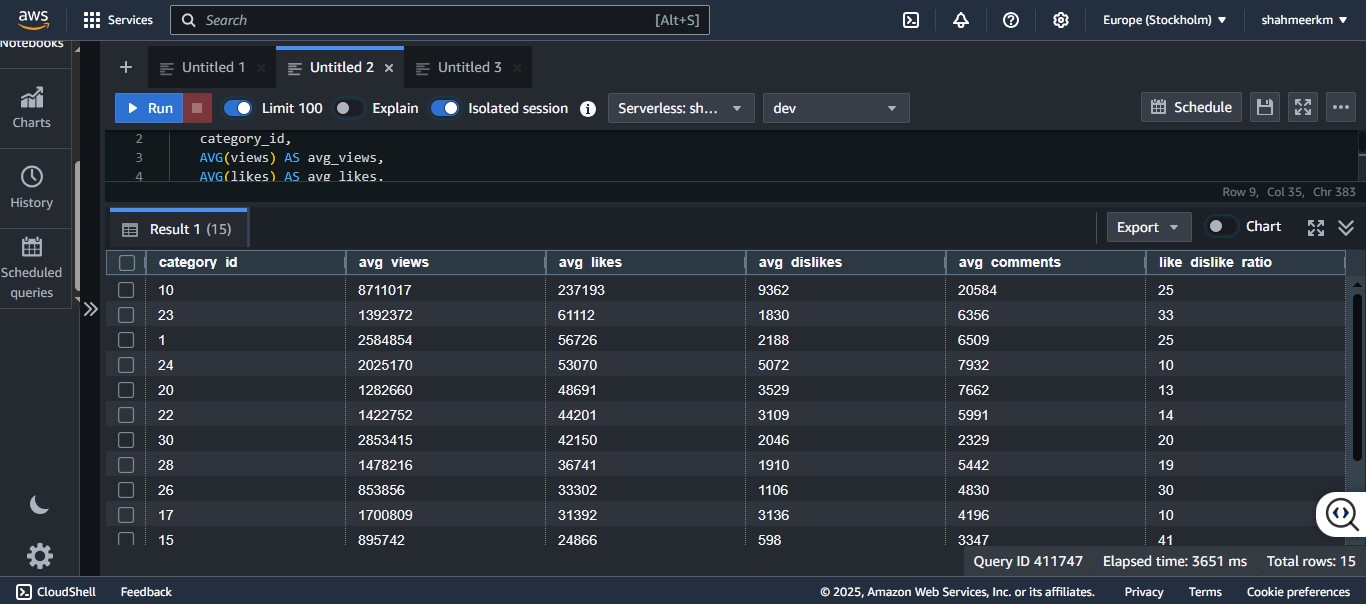
****

**Query 2**

Display video performance per category but only the categories where the like to dislike ration is at least 9:1



**Query 2 result**

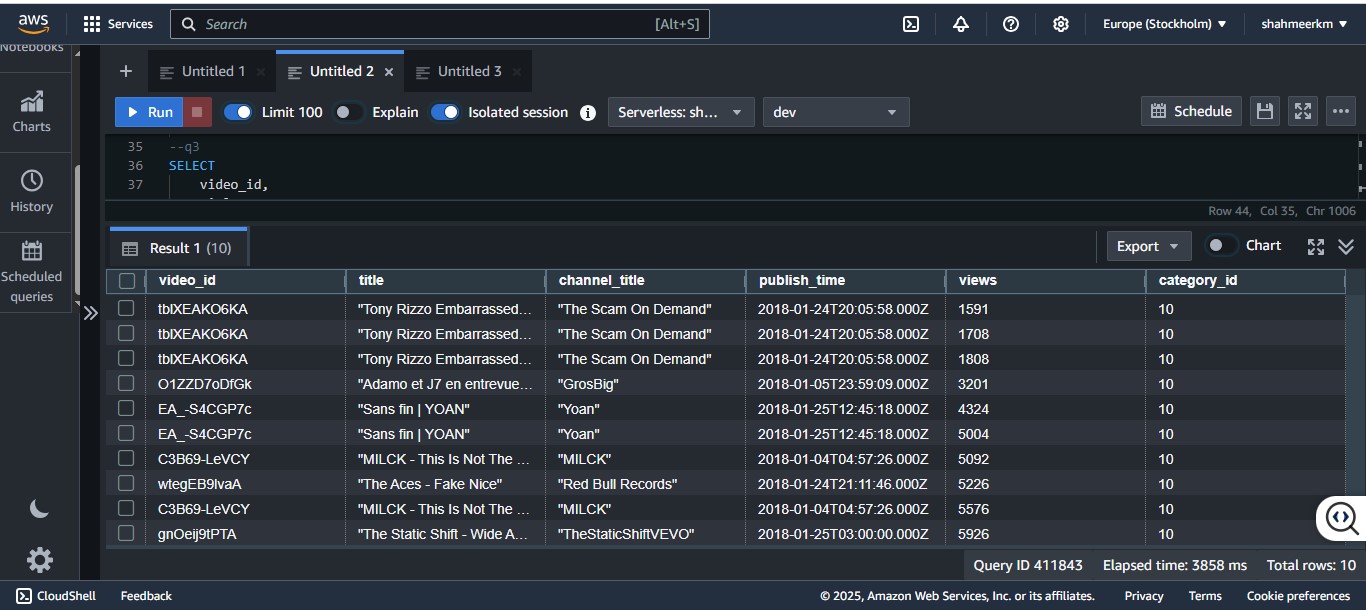


**Query 3**

Display least viewed music videos in January 2018

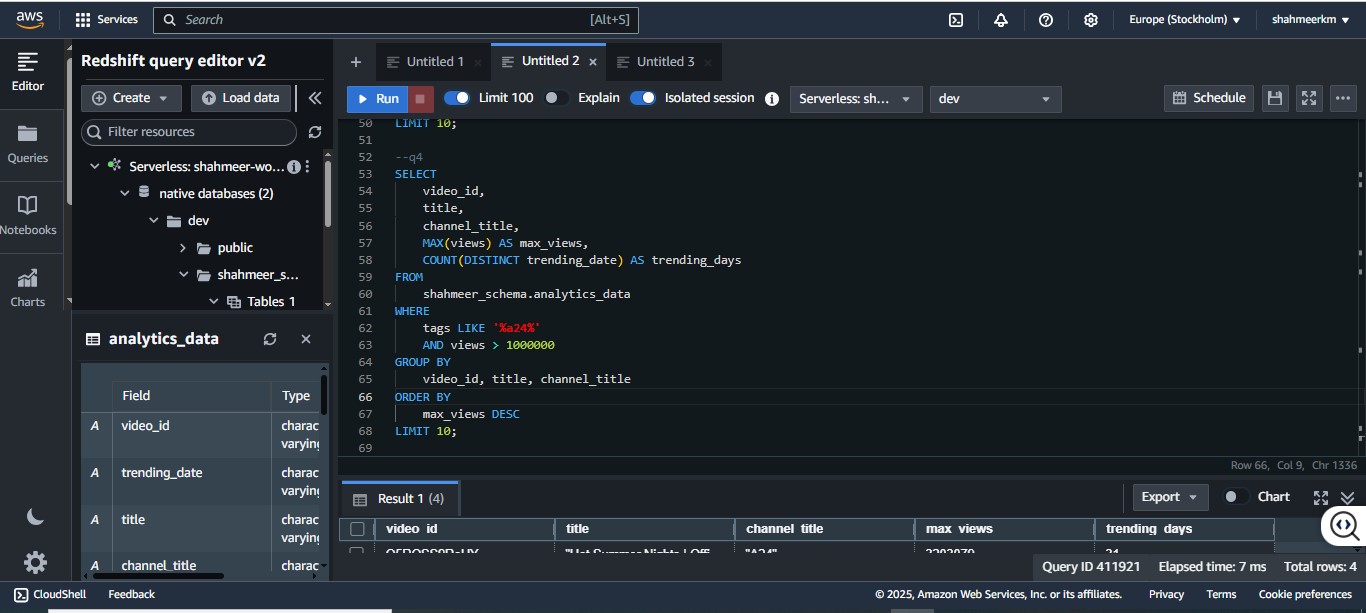


**Query 3 result**

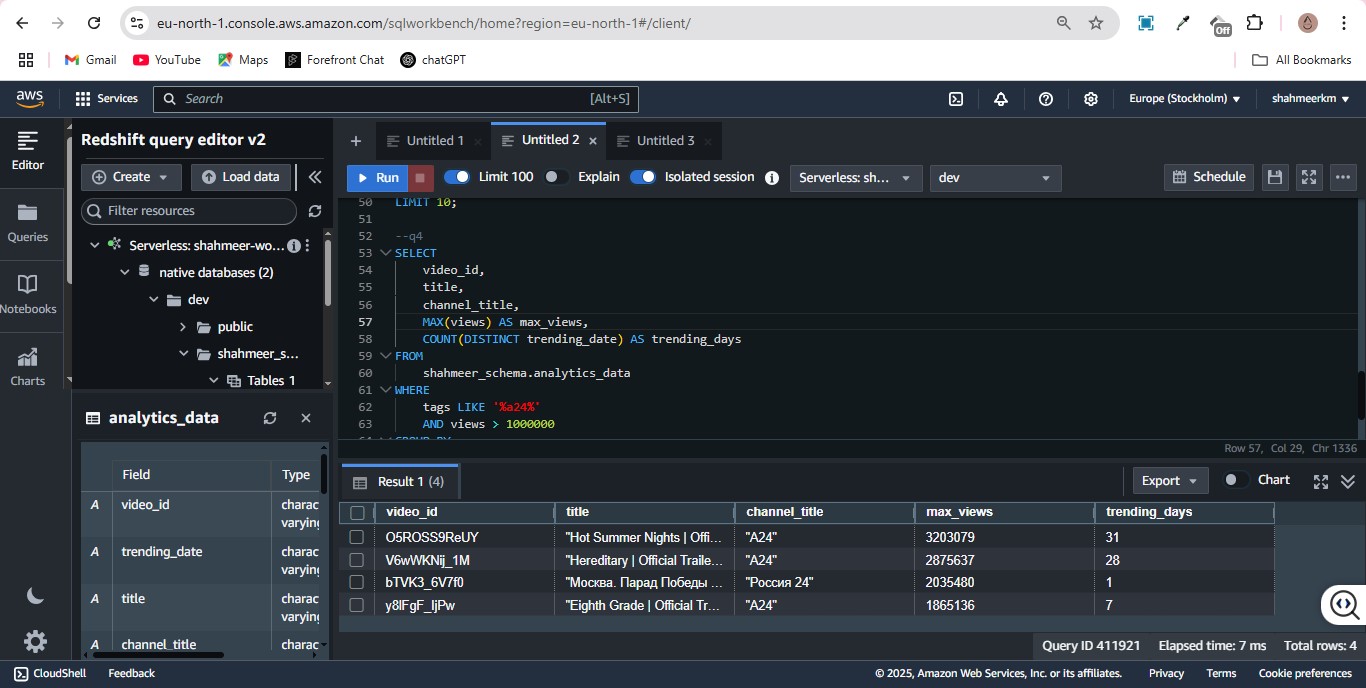
****

**Query 4**

Display videos with 1 million views or more that have a24 in their tags and also their trending duration

****

**Query 4 result**

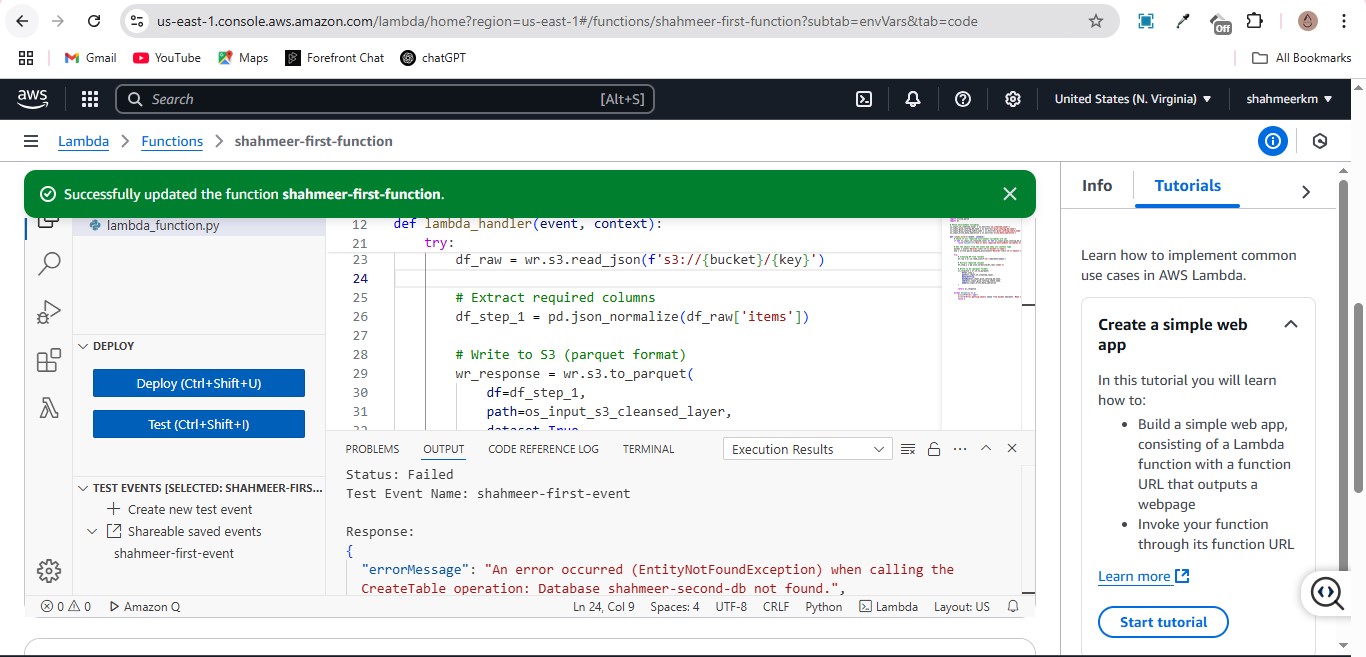


**Issues faced**

The table created wasn’t showing any rows at first, the reason was that I was creating a table for the clean data before running the function, this was causing overlap and thus an empty table so fixed that



The function wasn’t running as it was created in a different region by mistake, switched regions to have everything in the same region to fix this



**Conclusion**

The final run generated a table clean\_data which was then used for analytics data table, this was then loaded into redshift, a major issue was me creating an additional table that caused overlap issues, secondly the join function was for some reason not working and corrupting the clean data file as well, fixed that eventually to successfully generate analytics table, loading this caused issues as the file generated was hidden under several tables and wasn’t in the proper format, renaming it enable successful loading.

Overall query execution on both redshift and AWS Azure was pretty rapid and the whole process highlighted the benefits of cloud warehousing, though complex for beginners, once the learning curve is surpassed becomes a seamless experience for efficient data warehousing.