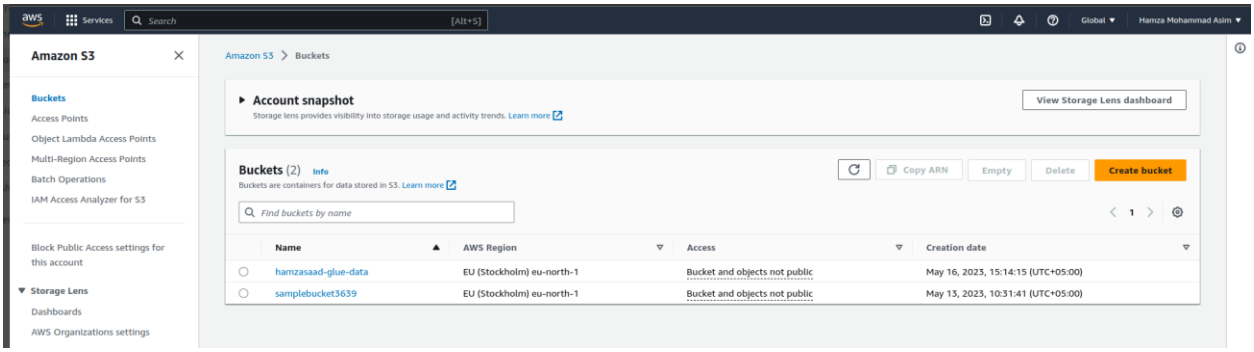
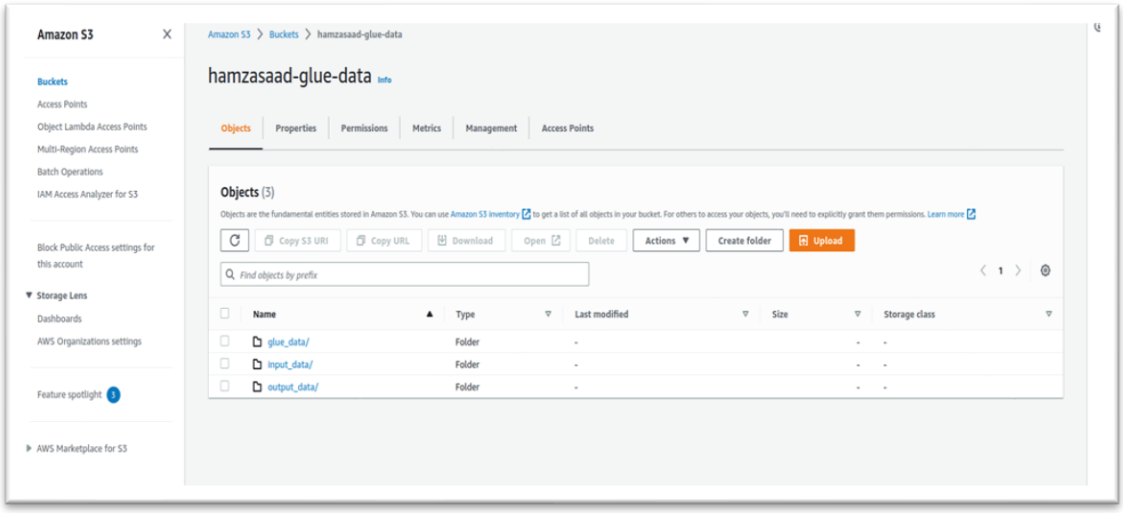
Graded Assignment 5.2

Name: Saad Sameer Khan
Employee#: 2303.KHI.DEG.034
Collaborated with: Mohammad Hamza Asim (2303.KHI.DEG.014)

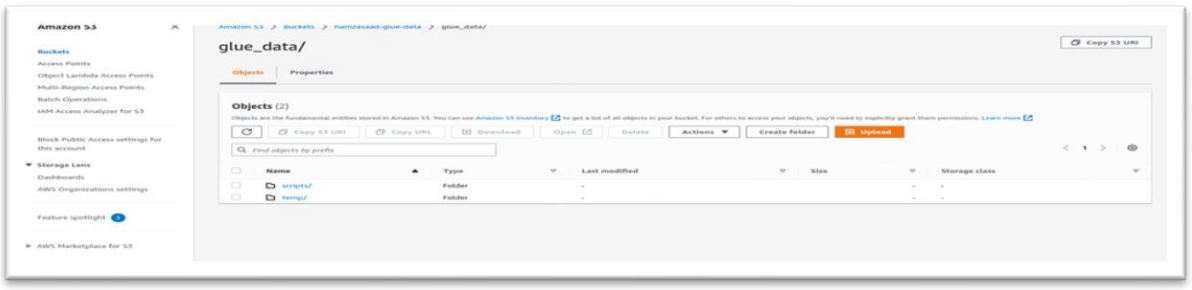
The following steps were taken for data access preparation:



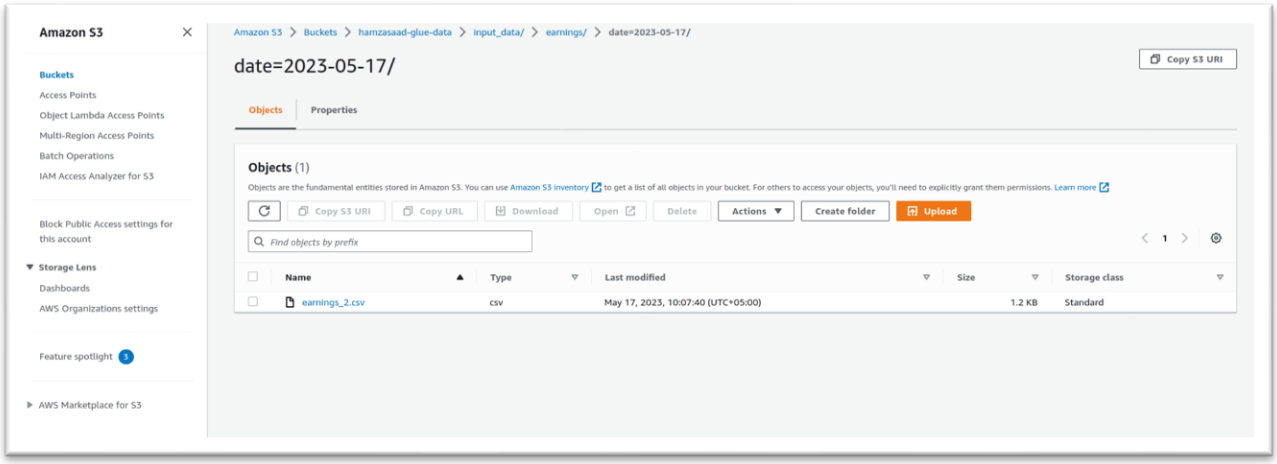
We created a bucket to dedicate a storage space in the cloud named ‘hamzasaad-glue-data’. This bucket includes objects stored in it which are directories consisting of folders and files, usually holding user metadata. Basically, everything is stored in a very organized way as in an inventory.



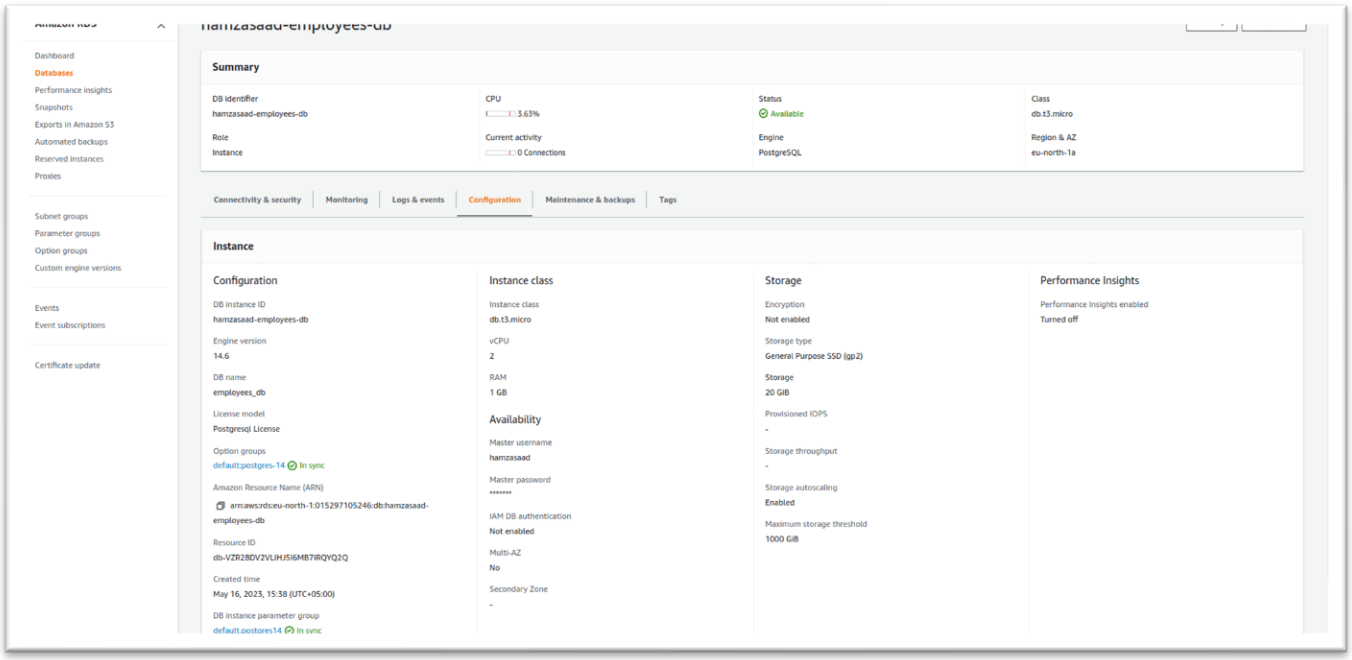
We created 3 Directories in it named glue_data, input_data, and output_data. Input data consists of data that will later be used as the source of data to perform the task. It has a location and earnings directory having locations.csv and earning_1.csv (and 2) file in it respectively. Output is where the target will be stored after performing the whole task.



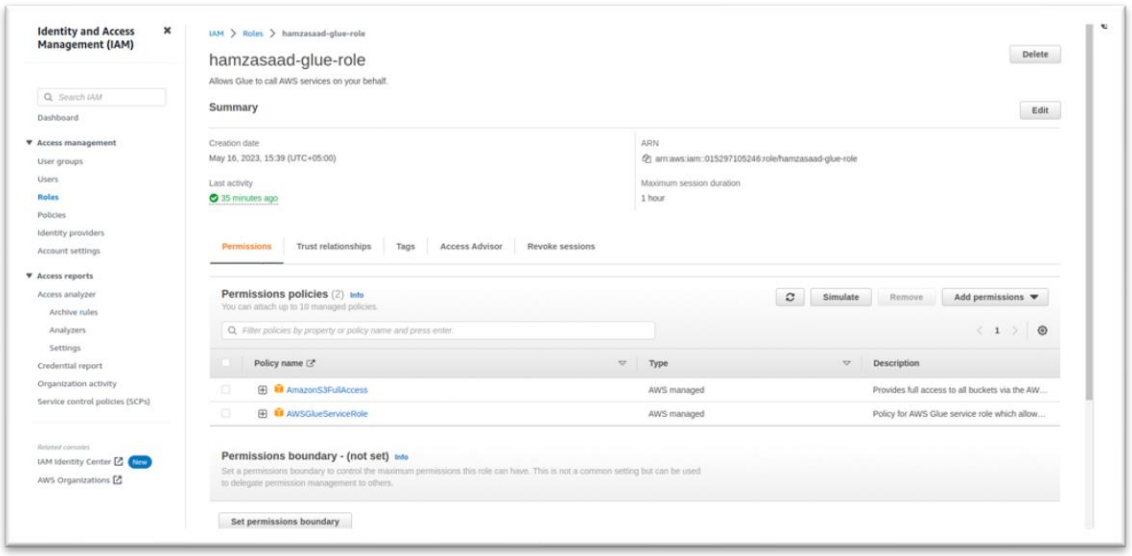
Glue data is a directory that has folders in it named script and temp. Script holds the Python file which is making all the logical groupings consisting of Python code.



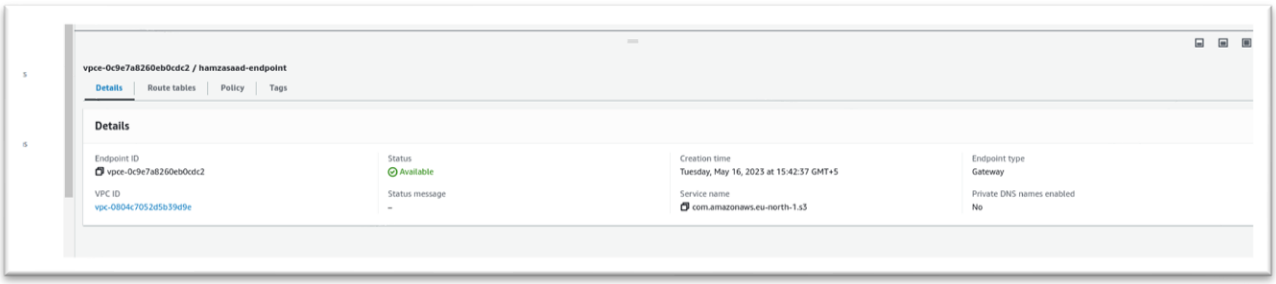
This date folder we have made is to store earnings_2.csv which will later be used as source file when working on job



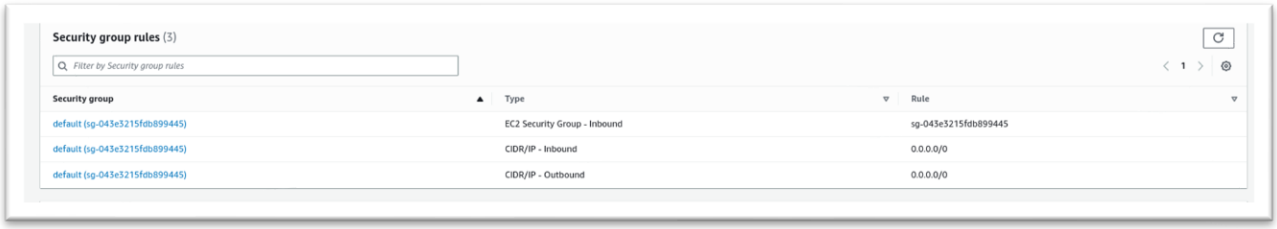
Configuration is mentioned in the image above.



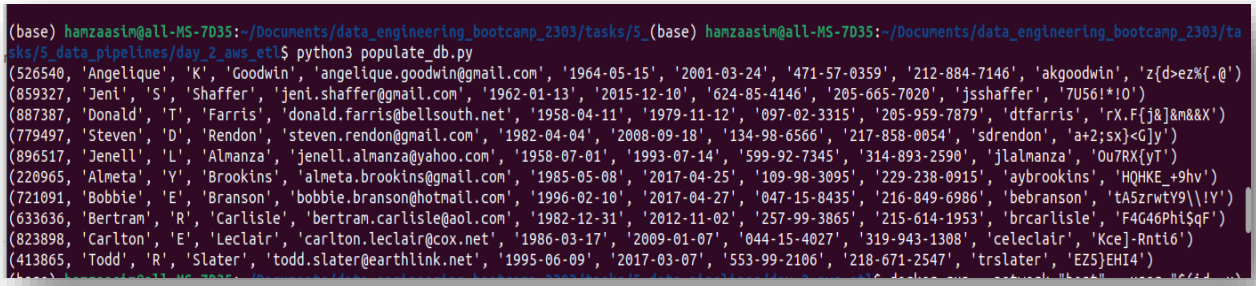
Here we are assigning a new IAM role which will allow glue or give authority and access to AWS services to perform required or desired tasks within the power handed over to them. By default, no IAM roles are assigned so they don't have any access and are restricted from performing any task. Access given here is 'AmazonS3FullAccess' which will allow glue to have complete access to S3. Second is 'AWSGlueServices' which will allow to have access to glue services.



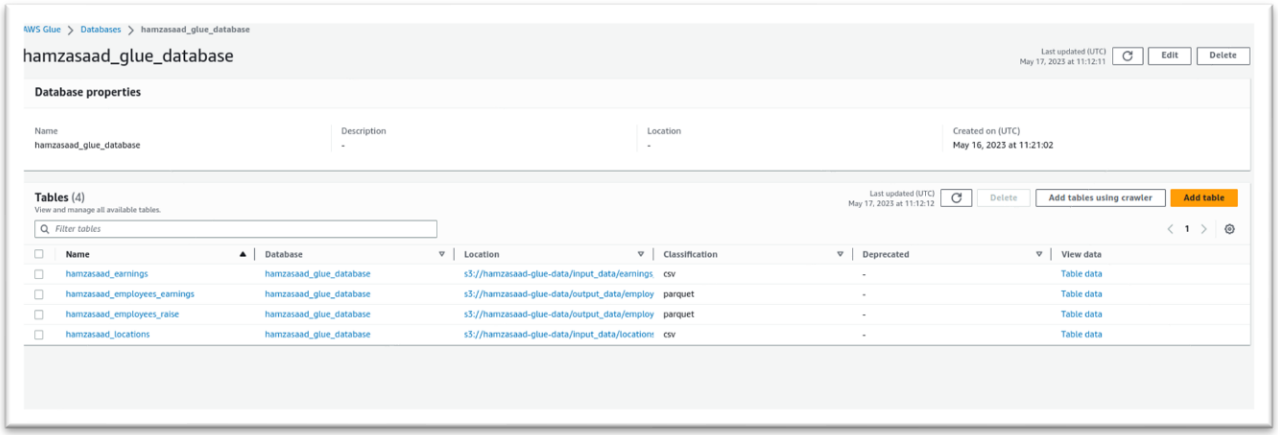
Vpc endpoints provide us with a private and direct connection to the virtual private cloud (VPC). Configurations are mentioned in the image.



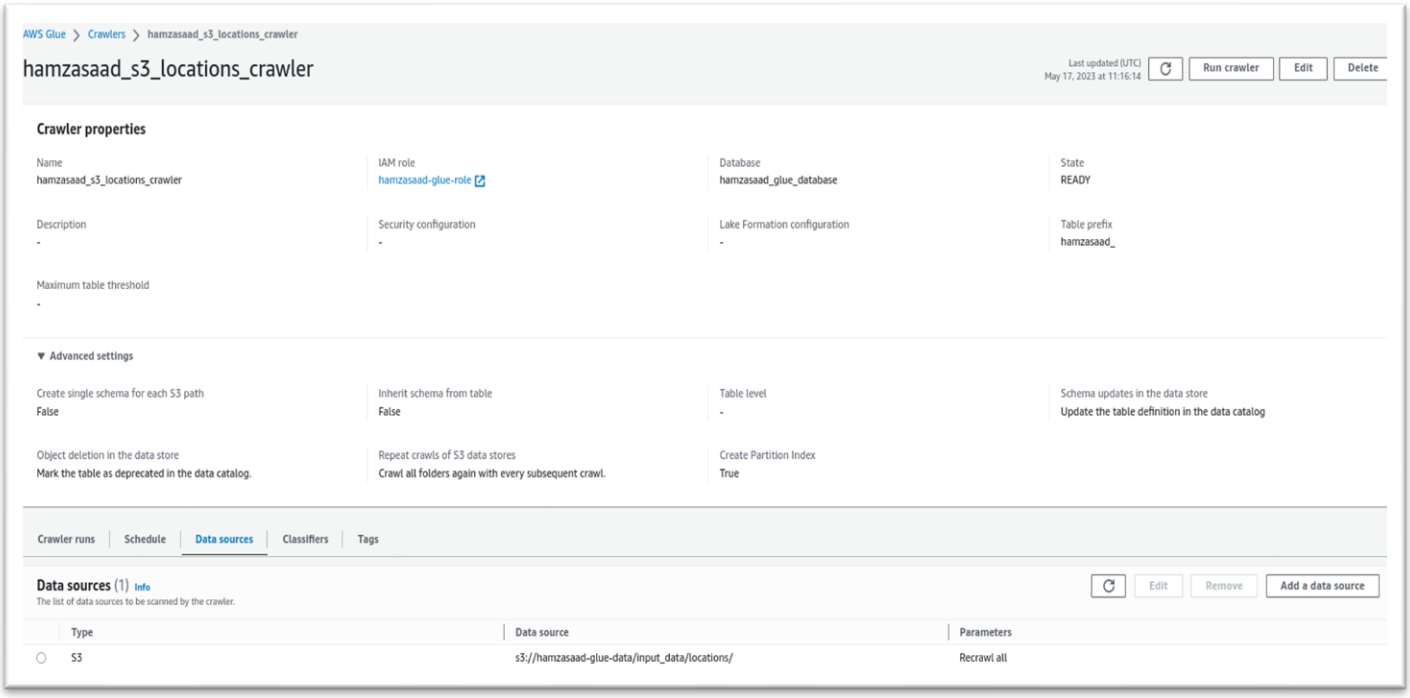
Security group rule allows us to control network traffic and it provides secure access to my resources in the cloud and database. It also helps maintain integrity. Two Inbound rules were added as seen in image above.

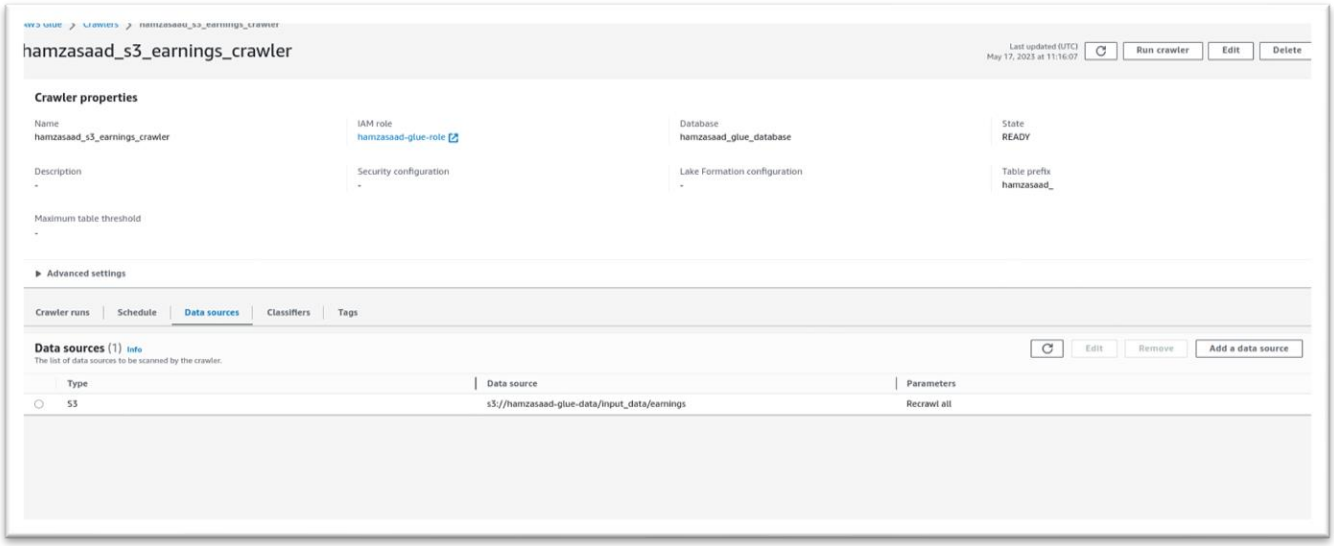


By providing my host endpoint URL from RDS database details, a connection was established between script & database. The script will be able to interact with the database

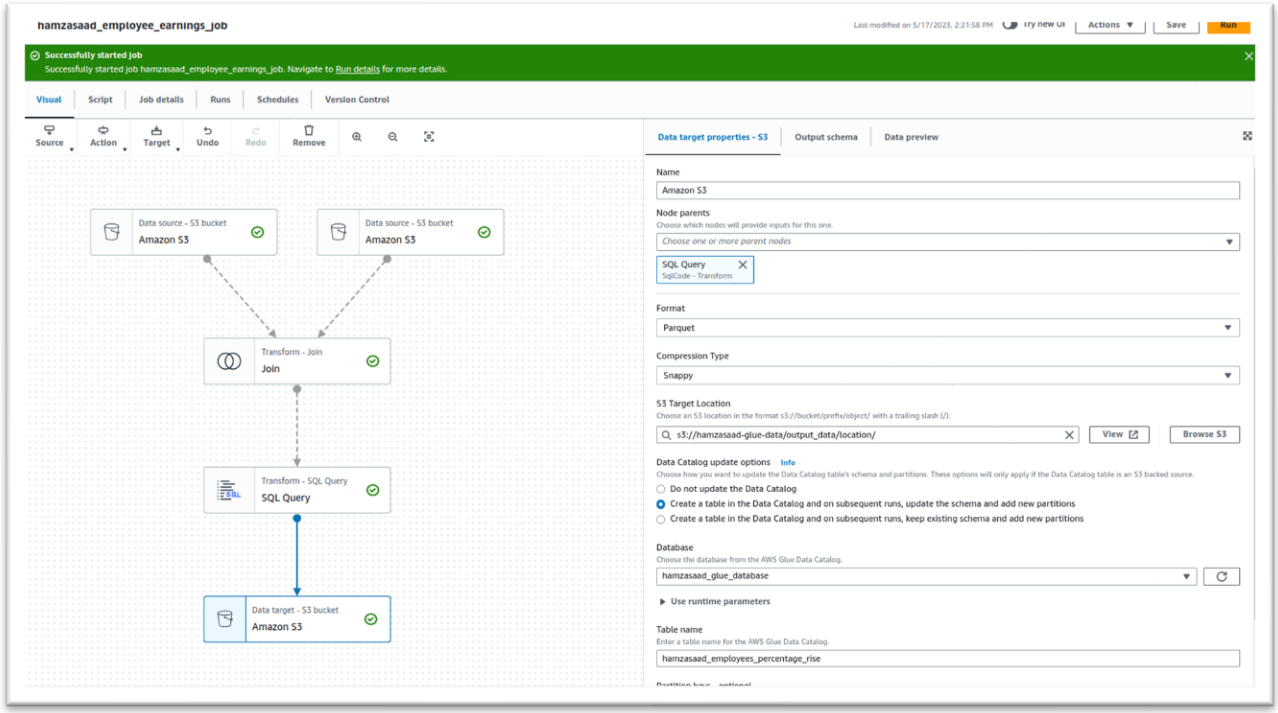


We created a Database here in AWS Glue. Glue, in simple words, allows us to catalog our data and perform ETL jobs to prepare data for analytics. Database here shown are the ones which two have been coming from the location of input_data which was located in S3. This recalls us of the csv files we added to act as the source to perform the task. The other two files in parquet format were generated after performing the assignment and got stored in the output_data folder.



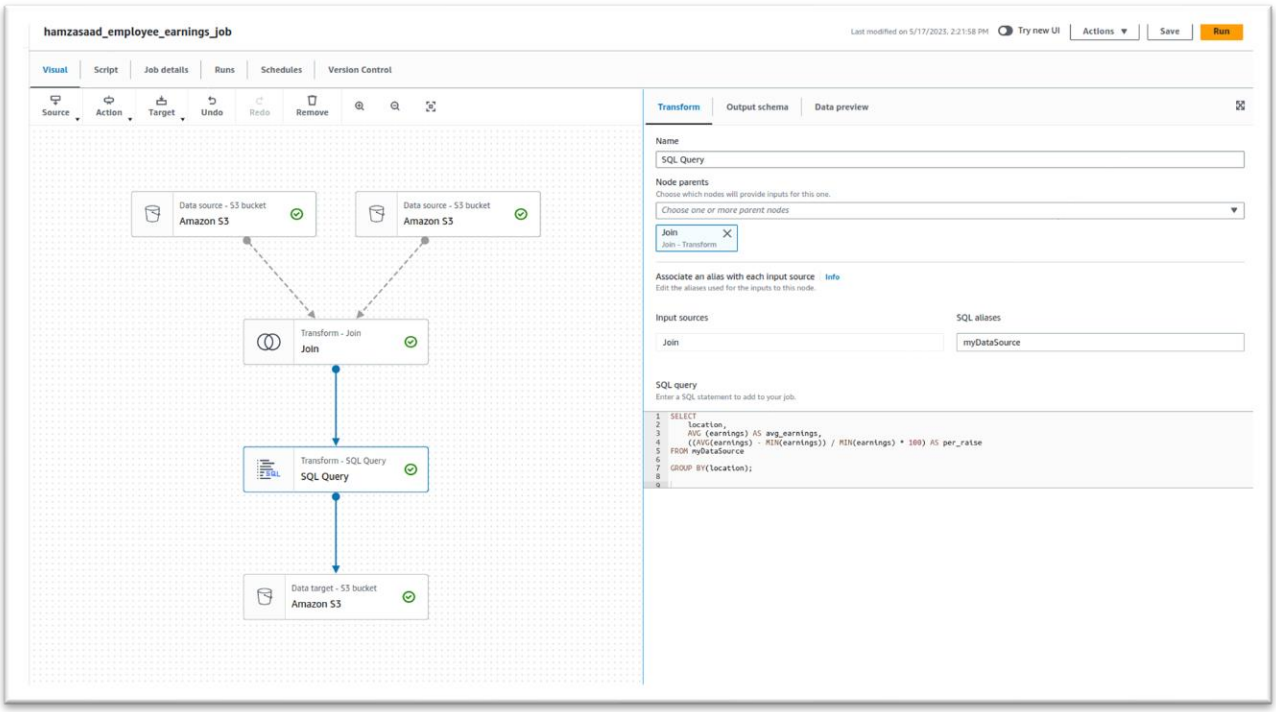


Glue crawler performs and analyzes the data, identifies patterns, and creates a summary of the data's format, schema, and relationships. Configurations are seen in the image above.

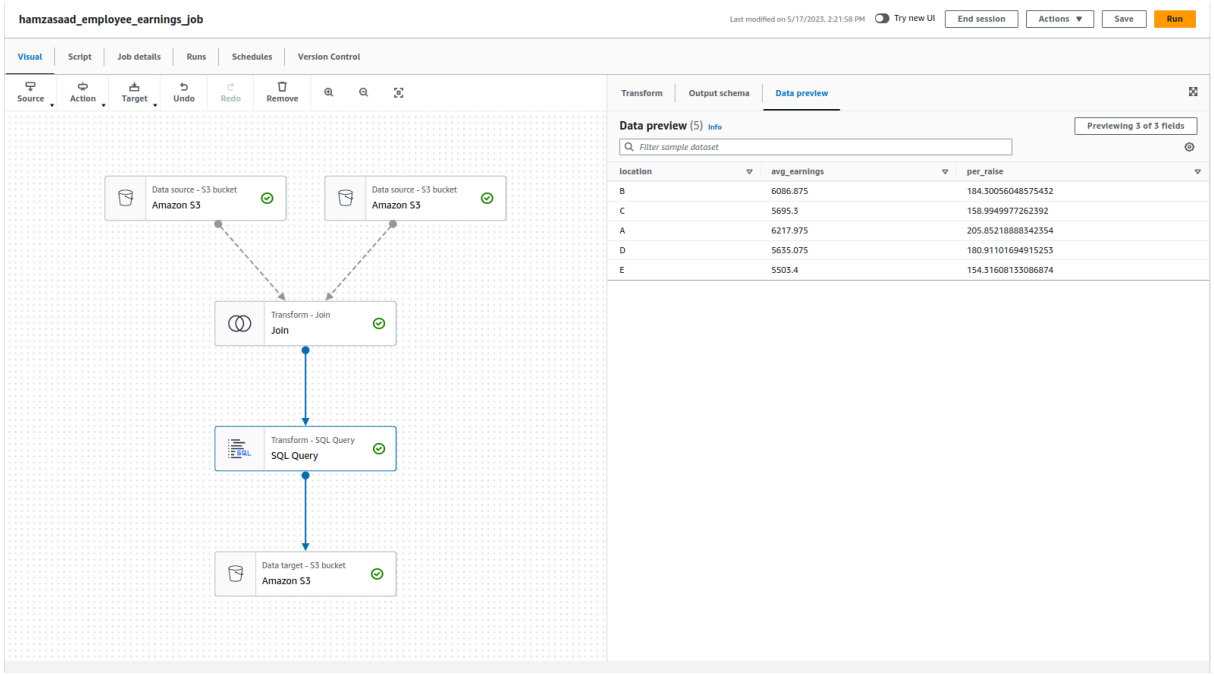


Here we have created a job. A job performs a specific data processing task. As you can see two Data sources are selected from the S3 bucket. One holds the data of earnings_2.csv and the other of locations.csv.

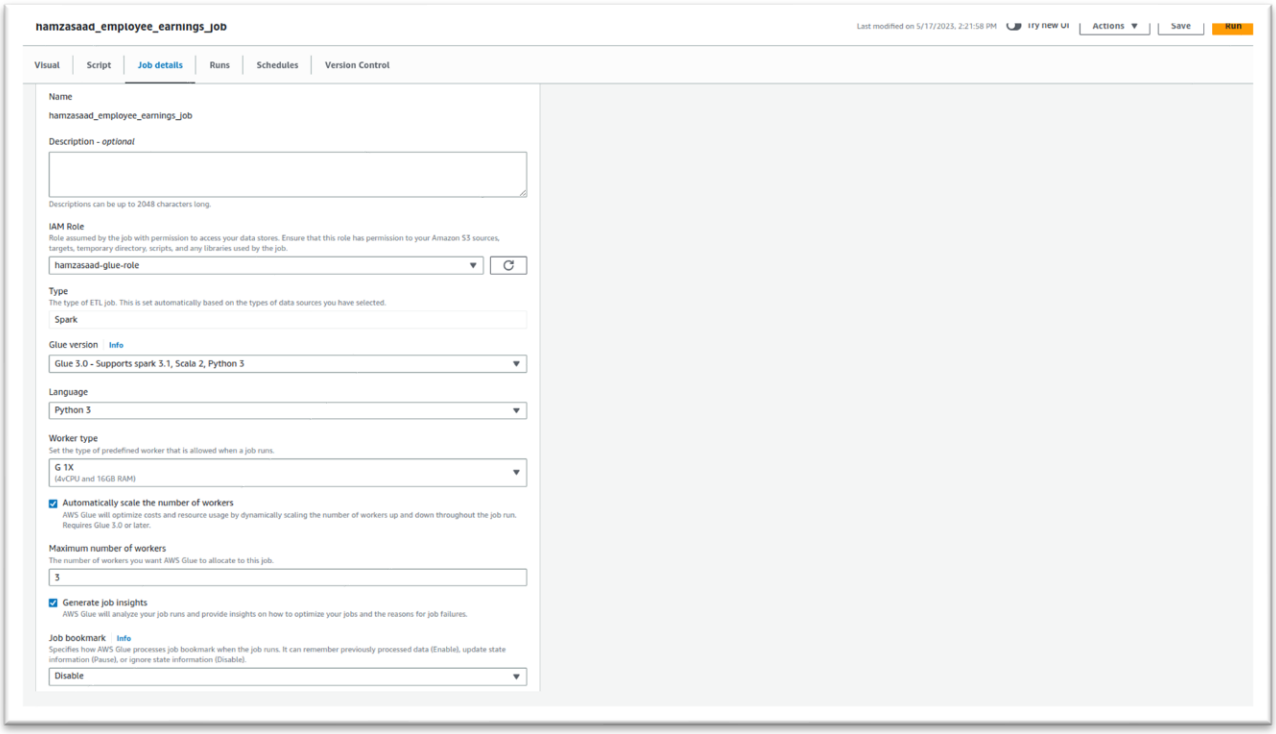
Inner join transform is made on the common field which was employee id. After this we performed SQL by adding ‘SQL Query‘. We had to group by all the locations and display the average earning on every location, and the earning rise. The data was sent to the data target which was also the S3 in the location of output_file which is stored in parquet format



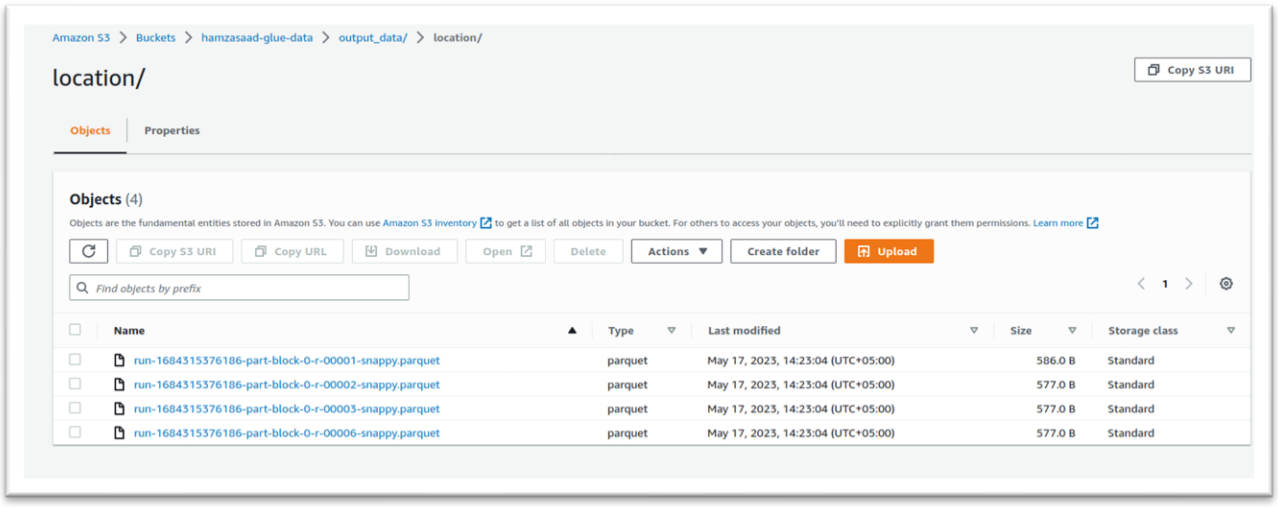
SQL query can be seen above



Result preview of the query



Final settings were made before saving and running the job.



After running the job we went back to S3 and check on the location directory we had in output_data. Parquet files were generated.