UNIT 5.2 GRADED ASSIGNMENT

Group members

Ifra Saleem (2303.khi.deg.003)
Umaima Siddiqui (2023.KHI.DEG.033)

UNIT 5.2 GRADED ASSIGNMENT

Task:

Using the earnings CSV as a base, prepare a new data file with employees' office locations. Make sure there are 5-6 distinct locations that are shared between employees.

Create a Glue job that aggregates the data based on the office location to calculate average salaries and raise percentages for these locations.

Solution:

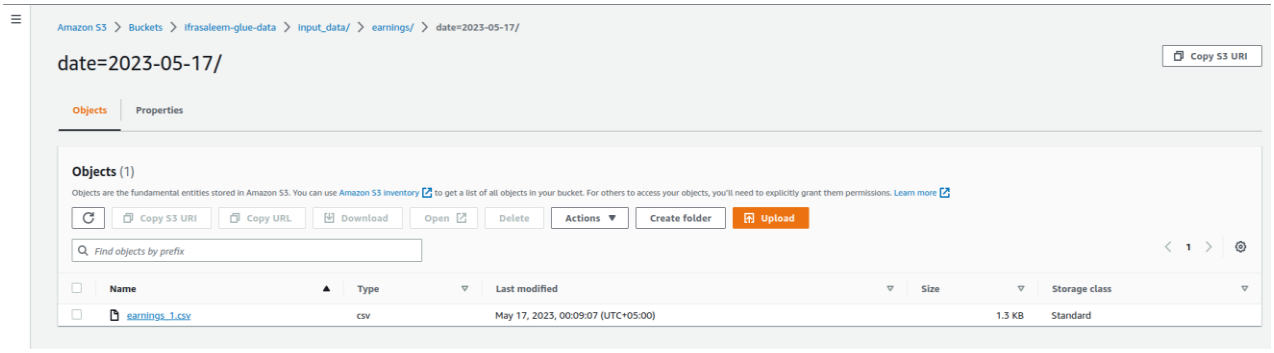
Input data:

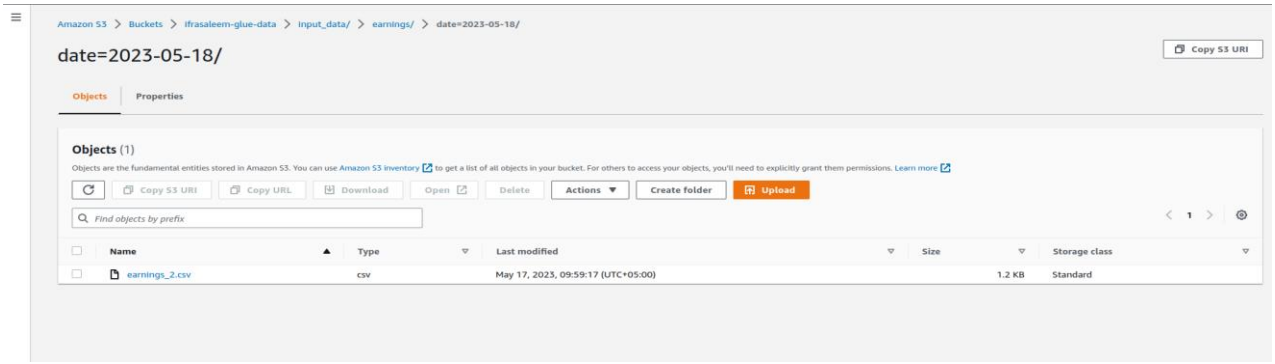
Here I upload three csv files to different folders of the bucket.

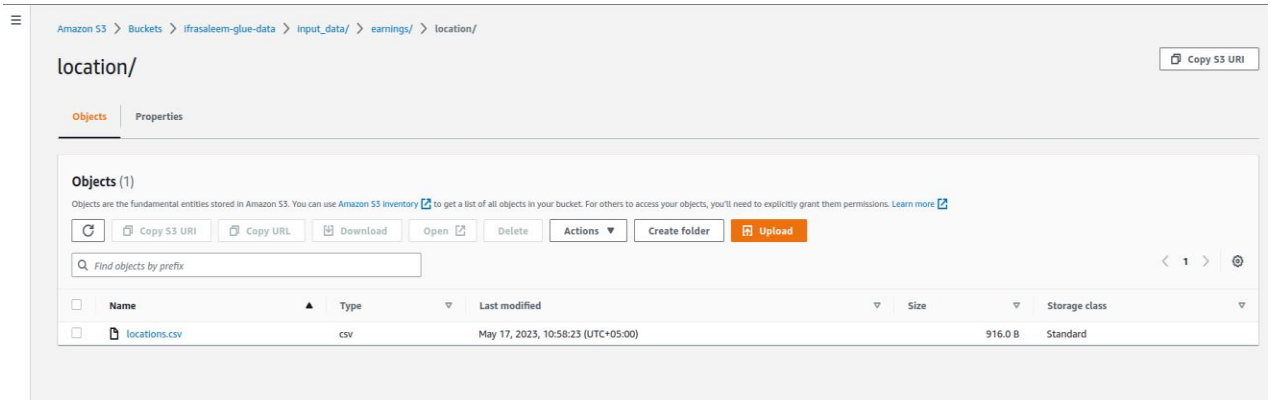
S3 bucket:

S3 bucket is used to retrieve or store large amounts of data in the form of objects within containers which is known as the bucket in AWS. It is used as backup or restore.

- So, in this task we created different folders inside that bucket and in the input data folder we created another folder earnings upload the csv files inside the folders inside the earning folder. So basically, we created the bucket to store our data and then we can use that data to create crawlers and jobs.



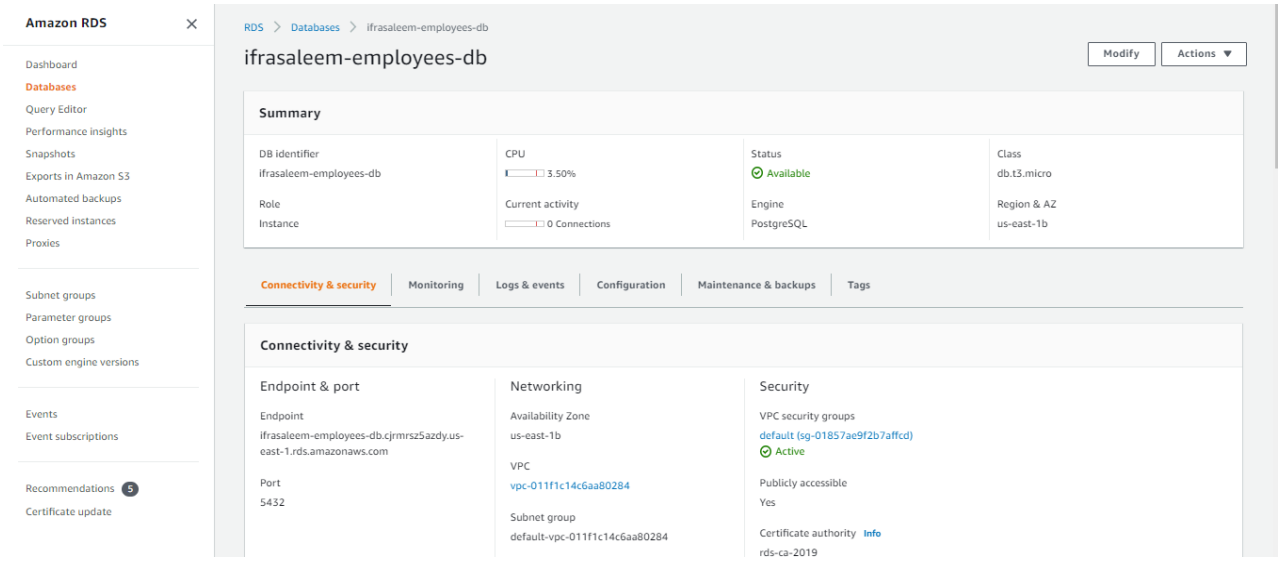




RDS:

RDS is a managed database service provided by AWS. Through RDS we can easily setup, operate and manage databases. It supports various database engines, including Amazon Aurora, PostgreSQL, MySQL, MariaDB, Oracle Database, and Microsoft SQL Server. It supports various database engines, including Amazon Aurora, PostgreSQL, MySQL, MariaDB, Oracle Database, and Microsoft SQL Server.

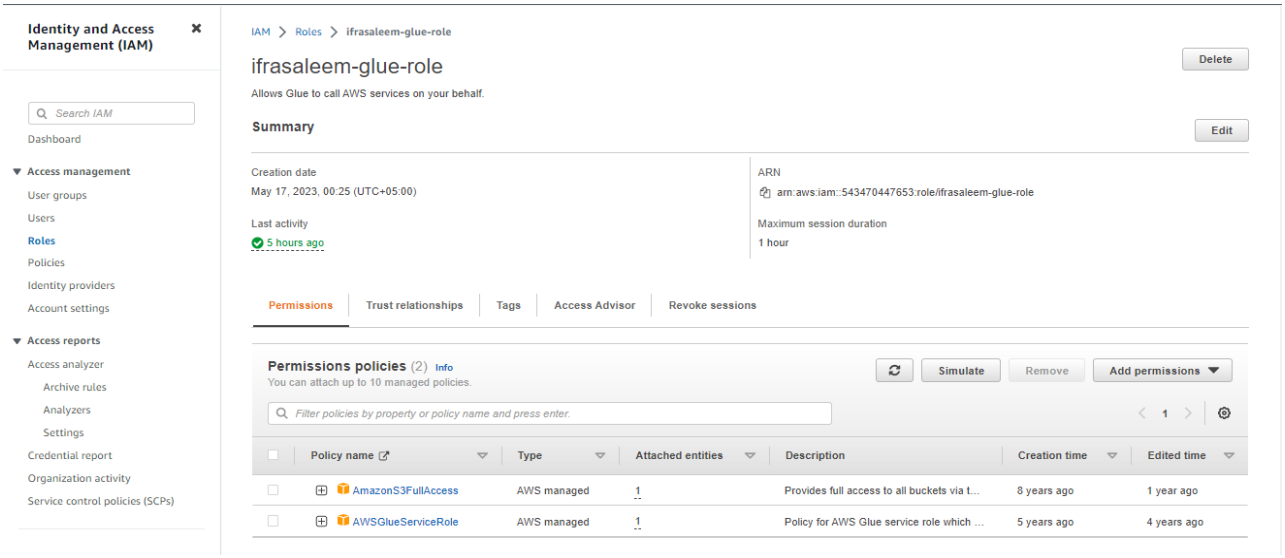
- So here we created a database, and we can use the credentials of this database to populate the database. We used the credentials in populate_db.py file and gave the values of username, password and endpoint URL.



IAM Roles:

IAM roles are used to give permission to users, services or applications.

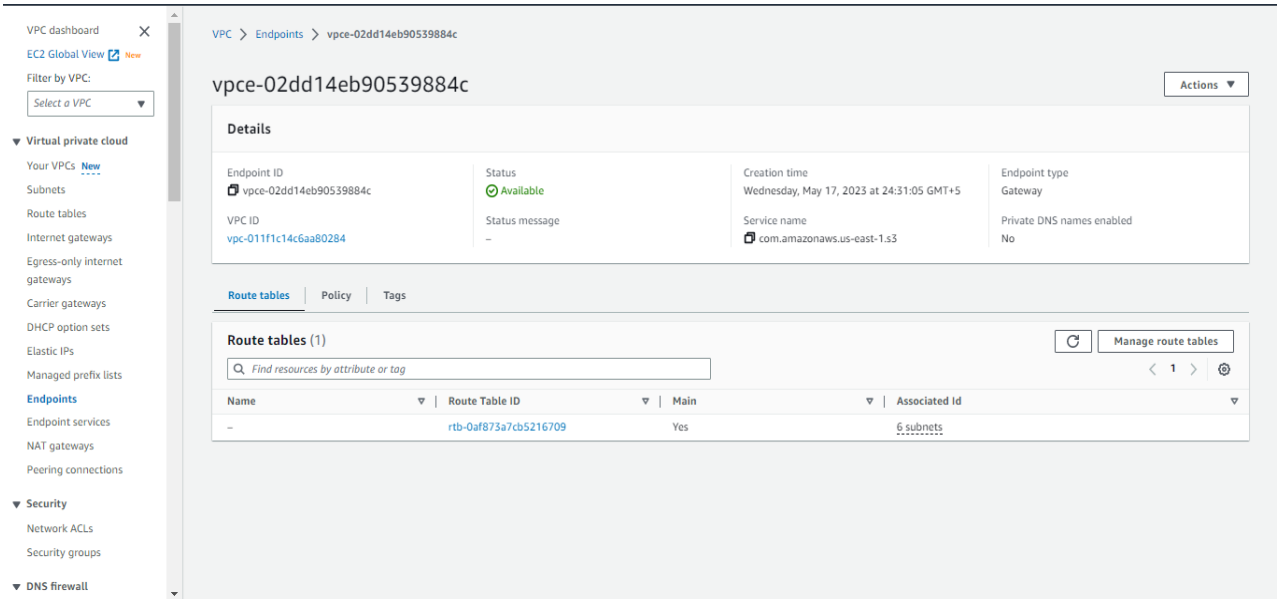
- We created an IAM role to use it while creating the glue job as it gives permission to services. And we used it while creating the crawlers and glue job.



VPC Endpoint:

VPC endpoints are used to privately connect virtual VPC to support AWS services and VPC endpoint services without the need for internet gateways. We can access AWS services by using private IP addresses.

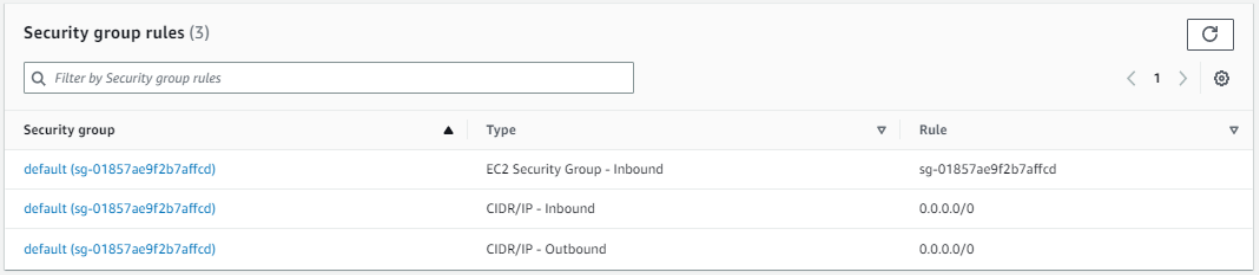
- We created VPC endpoint so that we can use it in our script file and populate the database. It is basically connecting the data present locally with the RDS database.

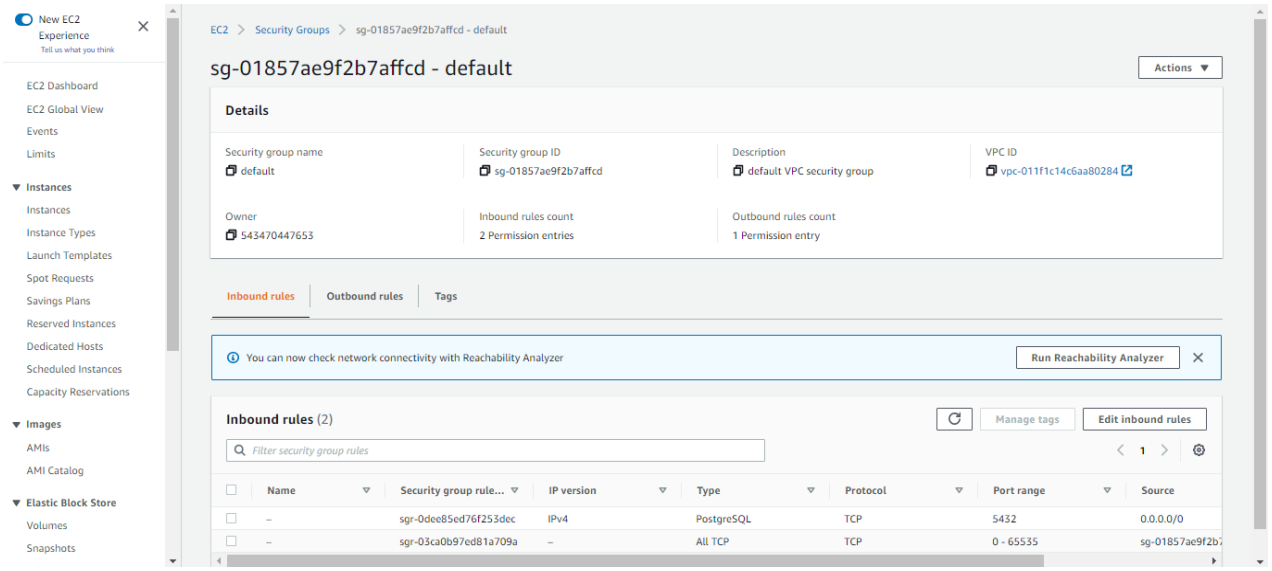


Security group rules:

It allows rules to define the types of traffic that are permitted to reach your instances. So, using these security group rules, we can allow or block any kind of traffic from any source.

- After creating the RDS database we define two inbound rules to give Glue Crawler access to RDS DB and to give access to RDS DB from our python script.

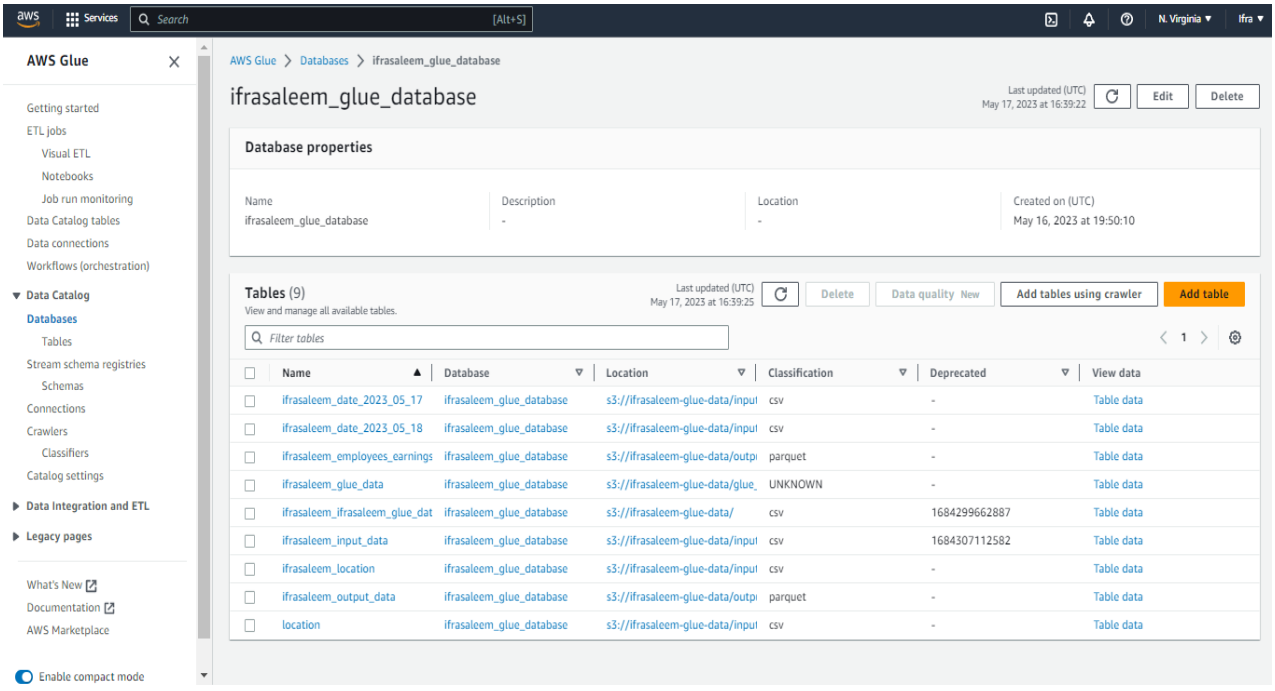




Data Catalog Database:

Data Catalog database stores metadata information of various data assets. It includes data source location, data formats, schema definitions etc.

- We created Data Catalog Database so that we can use it while creating S3 crawler as it will provide service to store metadata information. And we selected it as target database in S3 crawler as it will save all the necessary metadata.



Glue Crawlers:

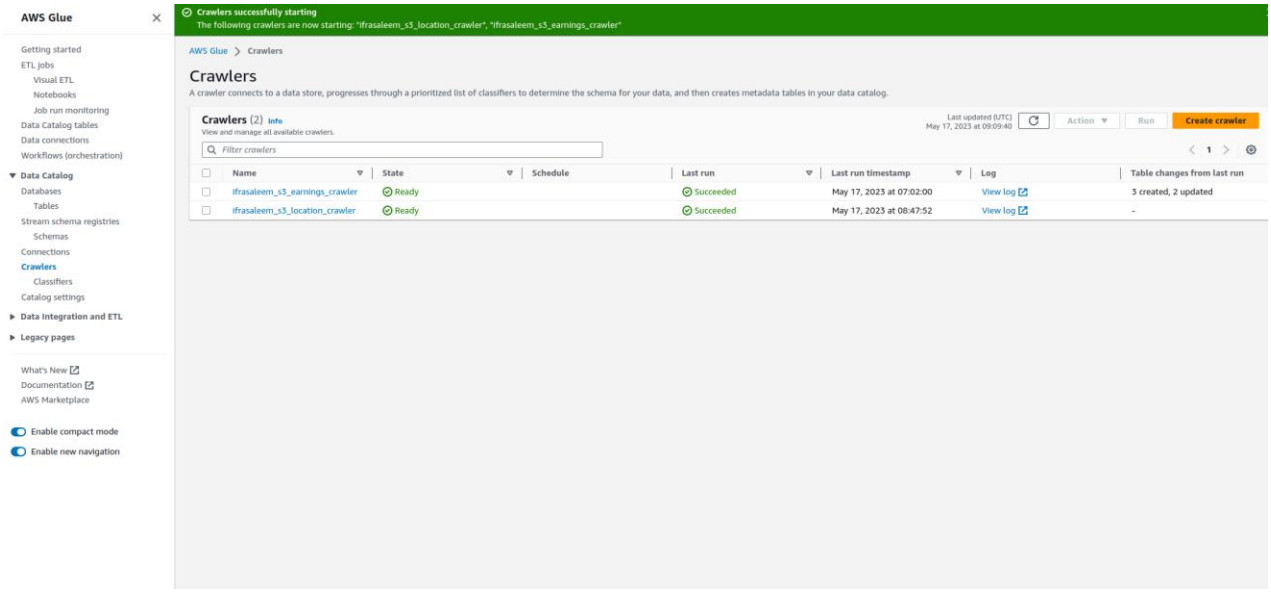
It is a fully managed ETL service. It automatically discovers and catalog metadata about data assets in various data stores. They analyze the data to infer its schema, structure, and format, and then create corresponding metadata tables in the AWS Glue Data Catalog.

- We created two glue crawlers. One crawler is used to analyze and manage earnings data and the other is used for locations data.

S3 Crawler:

AWS Glue Crawlers can be used to crawl data stored in Amazon S3 buckets. It analyzes and processes the data stored in s3 bucket using AWS glue and other AWS services.

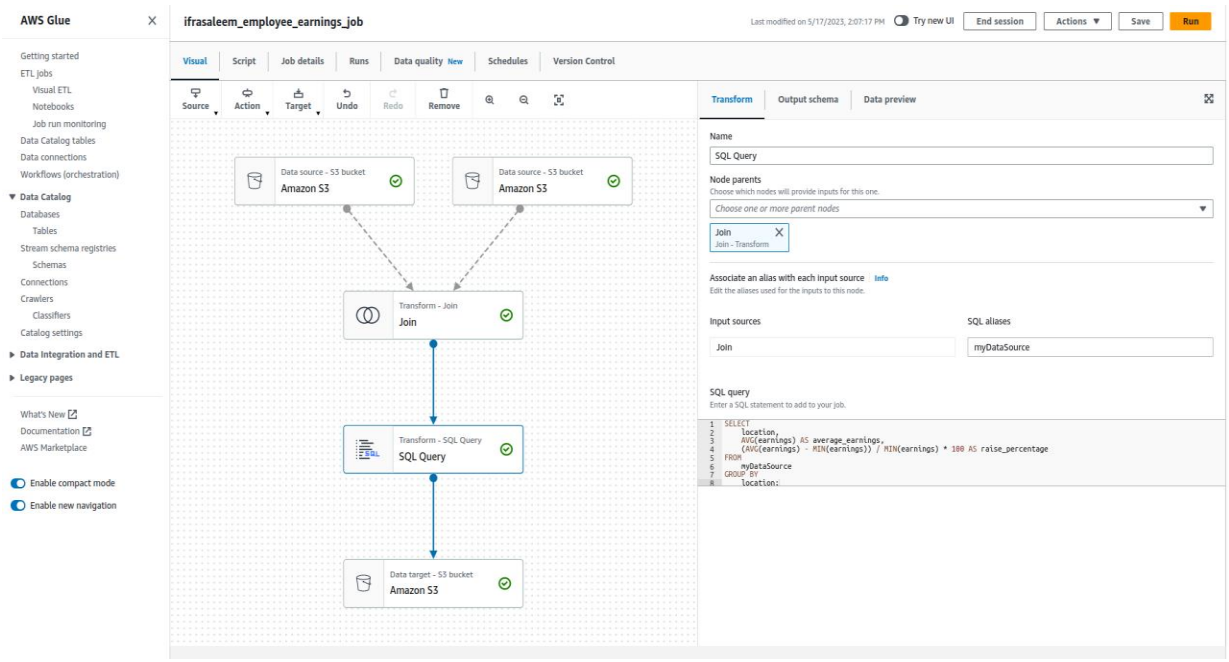
- We used S3 data source as it used to discover and catalog metadata about the data stored in your Amazon S3 buckets.



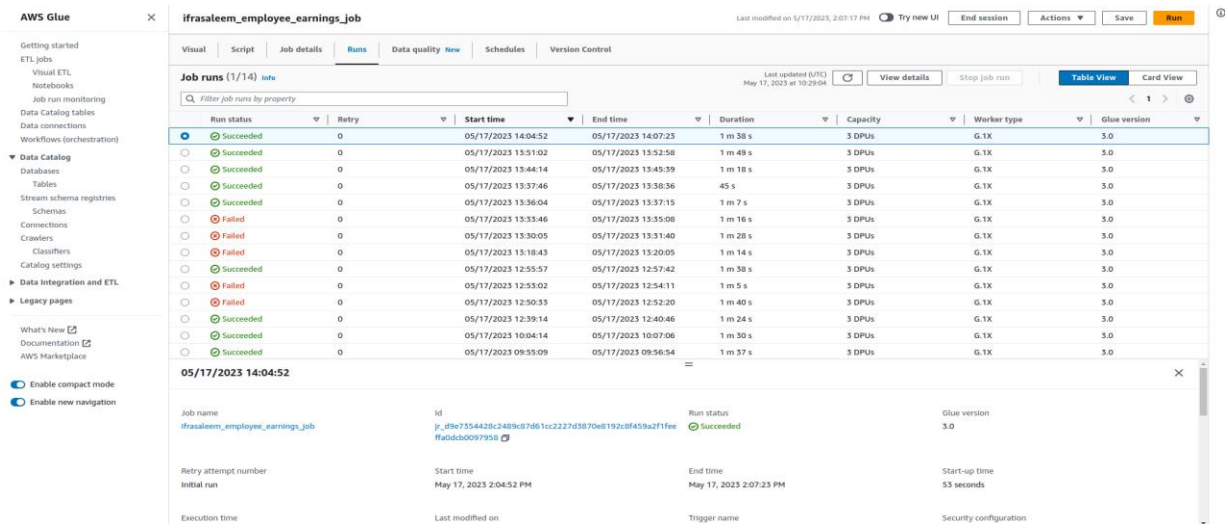
Glue Job:

It is also used to manage ETL service. It is used to create and run ETL workflows to process and transform data stored in various data sources, including amazon S3, relational database and data warehouses.

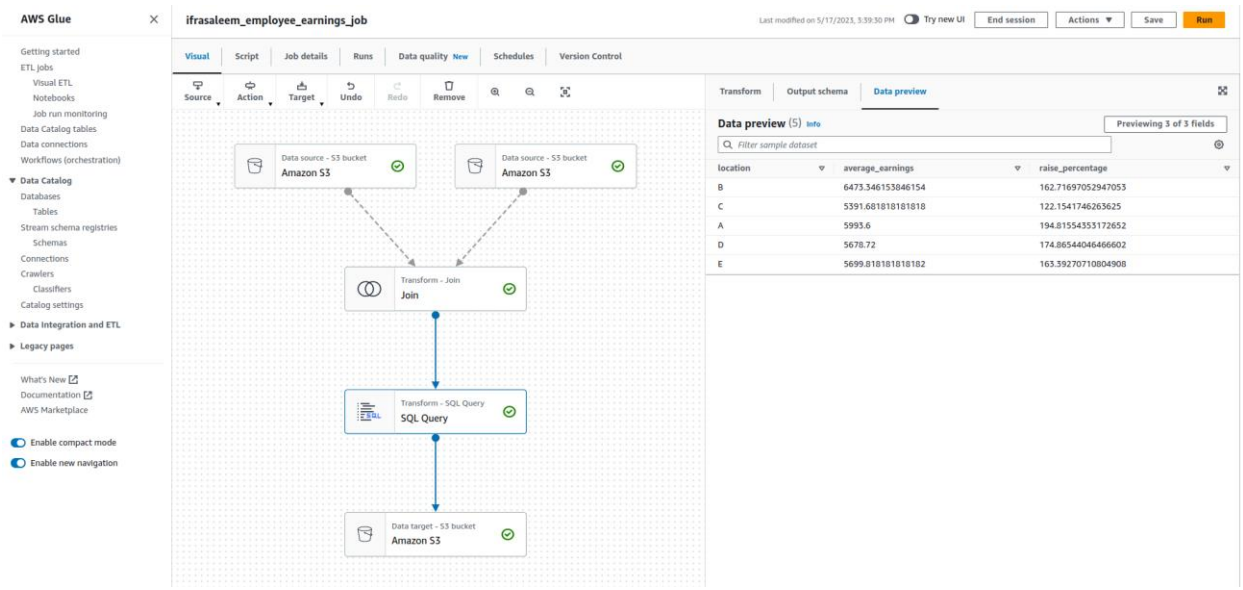
- We created a glue job to build a schema and to process the data source. We joined the earnings data and locations data and then wrote a SQL query to calculate the average earnings and raise percentage according to the locations.



Job runs:



Query output:



Output folder in the bucket: