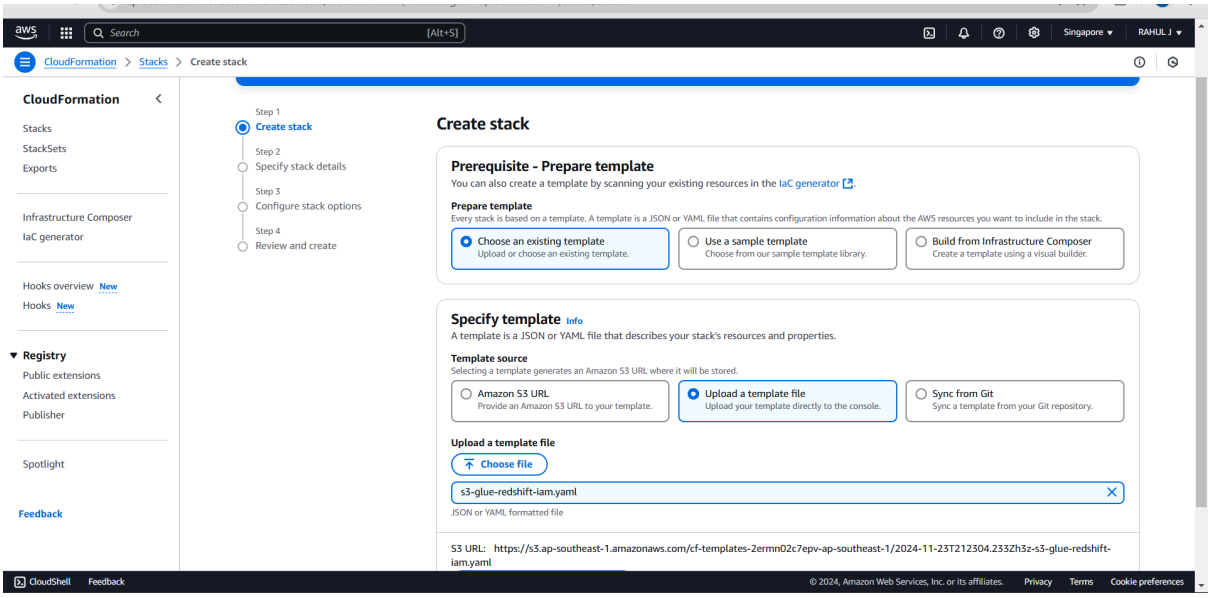
AWS Project - 2(Building End-To-End Datapipeline)

Summary

This end-to-end data processing and analysis on the AWS process involves creating a data pipeline using AWS services. It starts by setting up an S3 bucket and a Redshift cluster, followed by creating an AWS Glue database and crawler to catalogue data from the S3 bucket. An ETL job is configured using AWS Glue to transform and prepare the data for analysis. Finally, the Redshift Query Editor v2 is used to connect to the database and execute queries for data validation and insights.

Create the CloudFormation

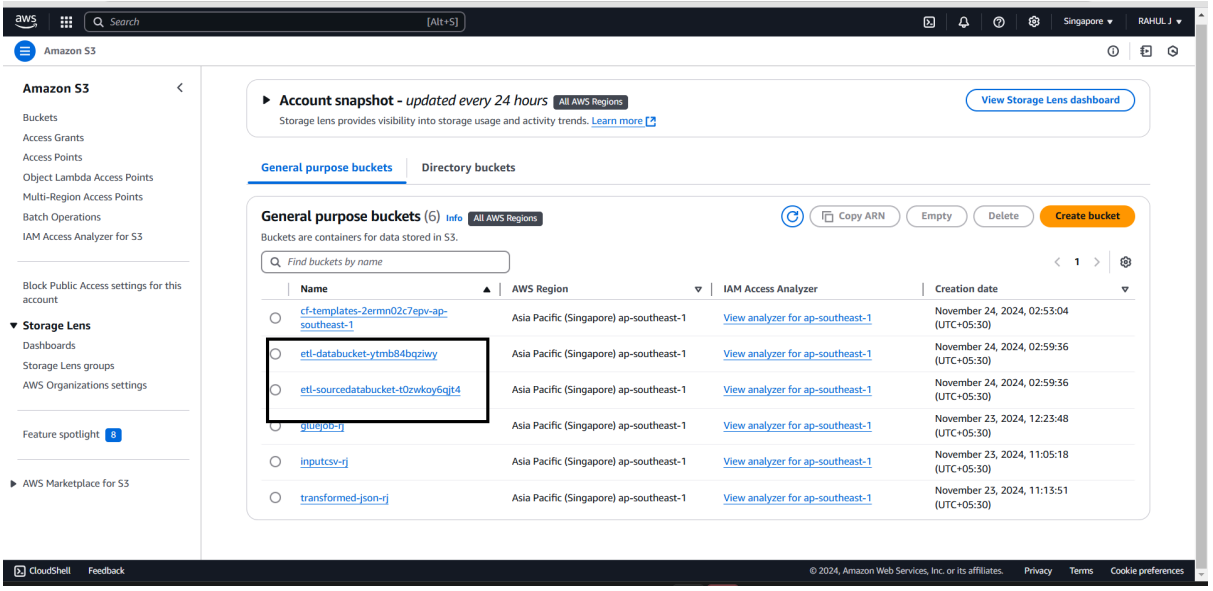
After login search for **CloudFormation** choose the Existing template select upload file choose the appropriate file to upload and click **next.**



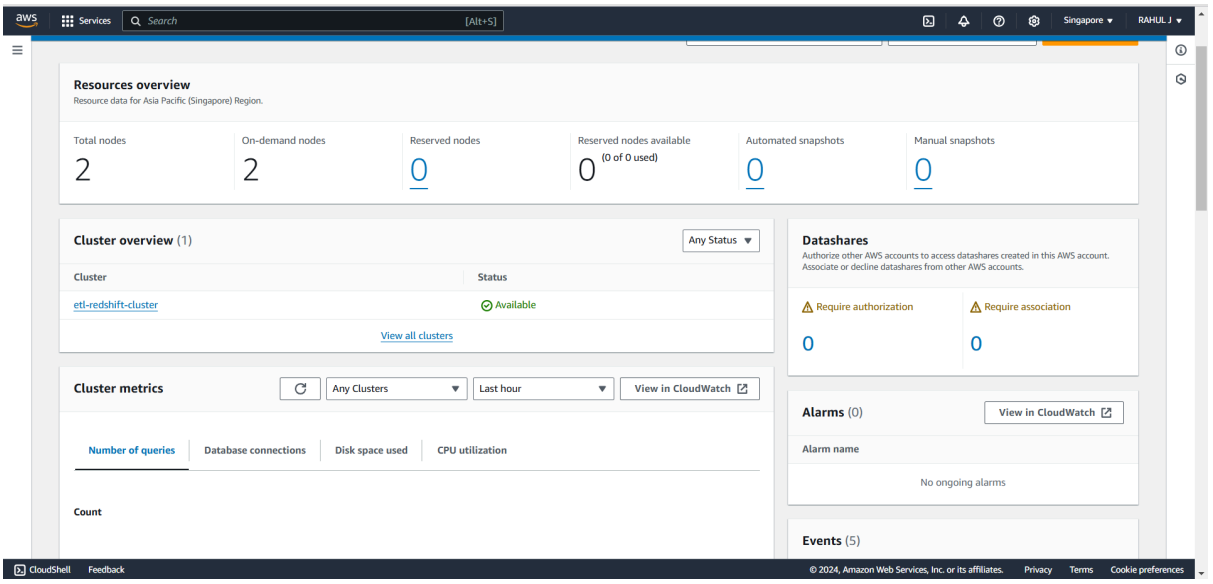
On the next page give the name of our **CloudFormation** and leave the rest of the configurations as the default and last tick the **acknowledgement.**

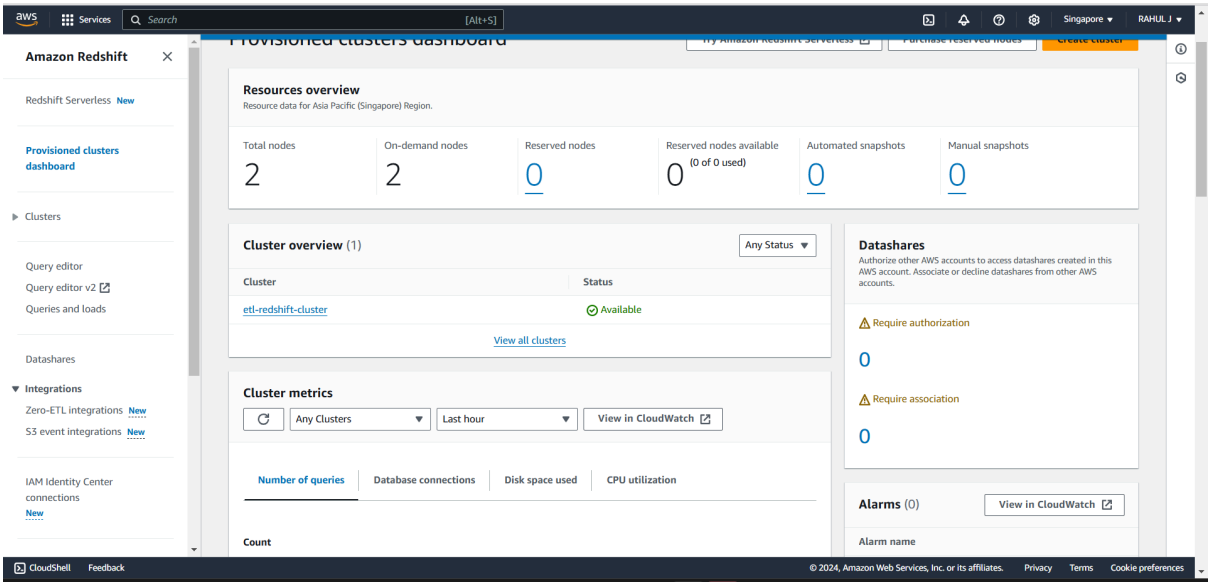
Now we need to check for the S3 bucket, Redshift cluster and its connection

First, we can start checking with S3 Bucket.



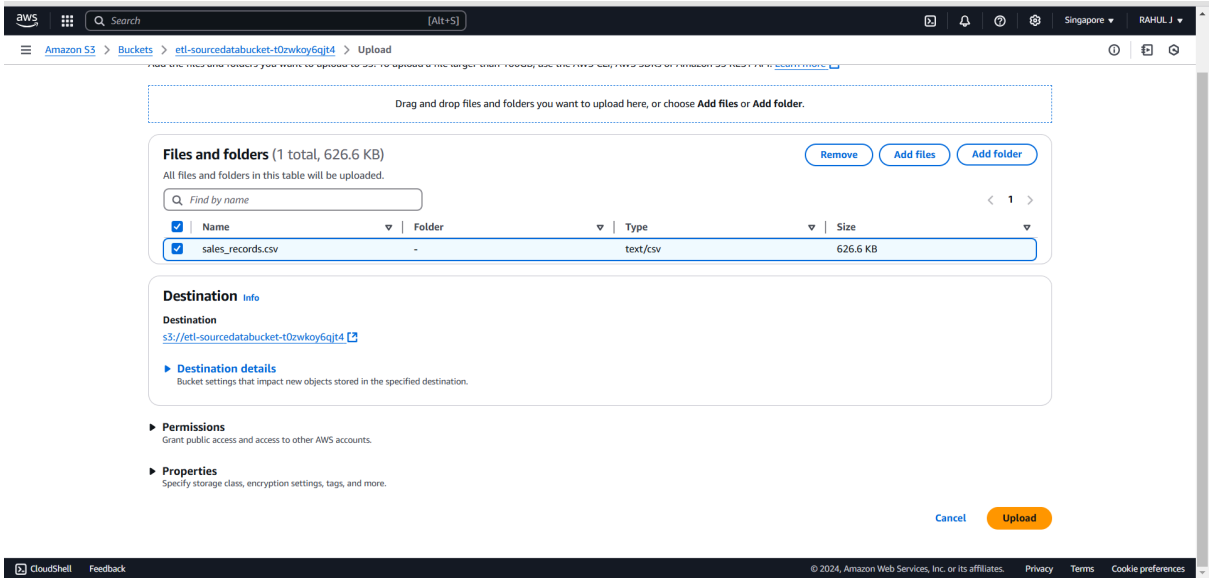
Now we can use **Redshift Cluster**





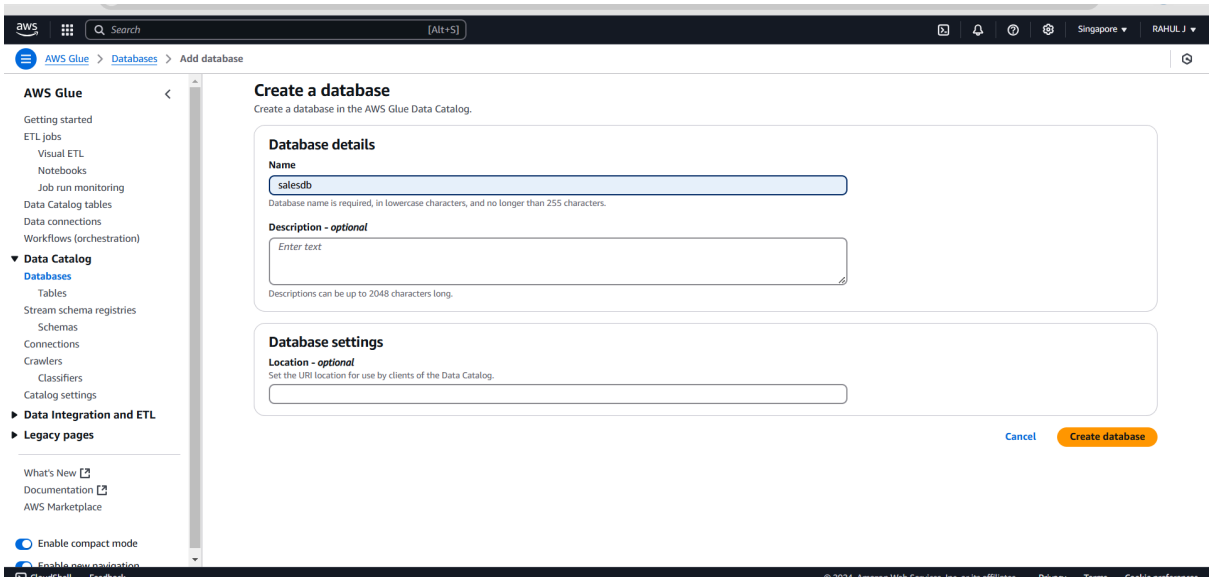
Now check for **AWS Glue** where the ETL Job is created or not

Upload file in S3

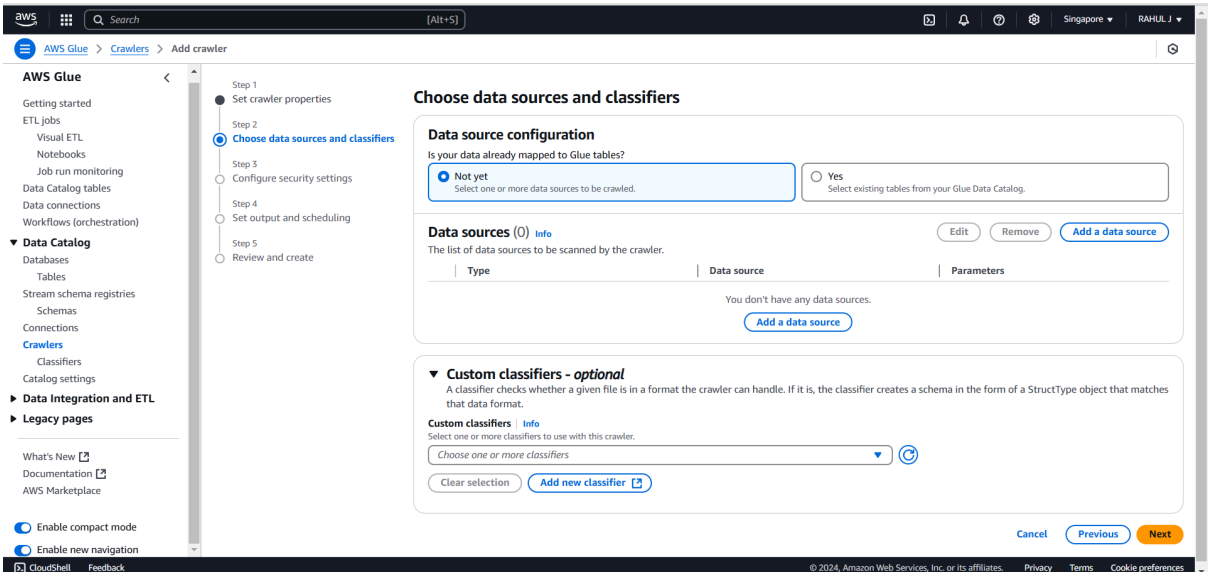


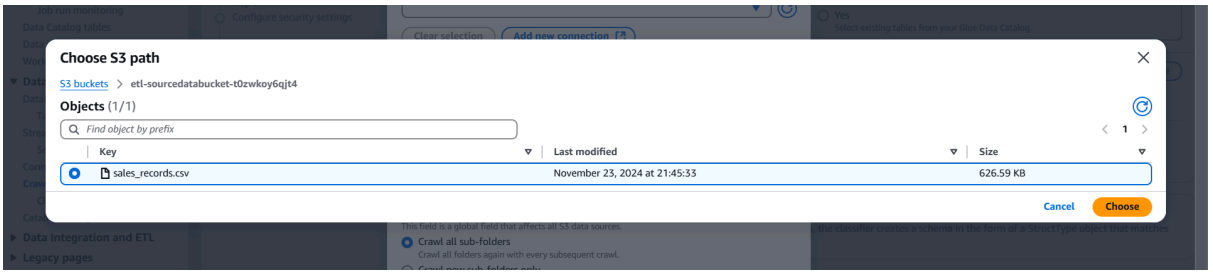
Create a Crawler in the AWS Database

Go to the **AWS Glue** in **Datacatalog** and click on **Add Database**
In that give the name of the database



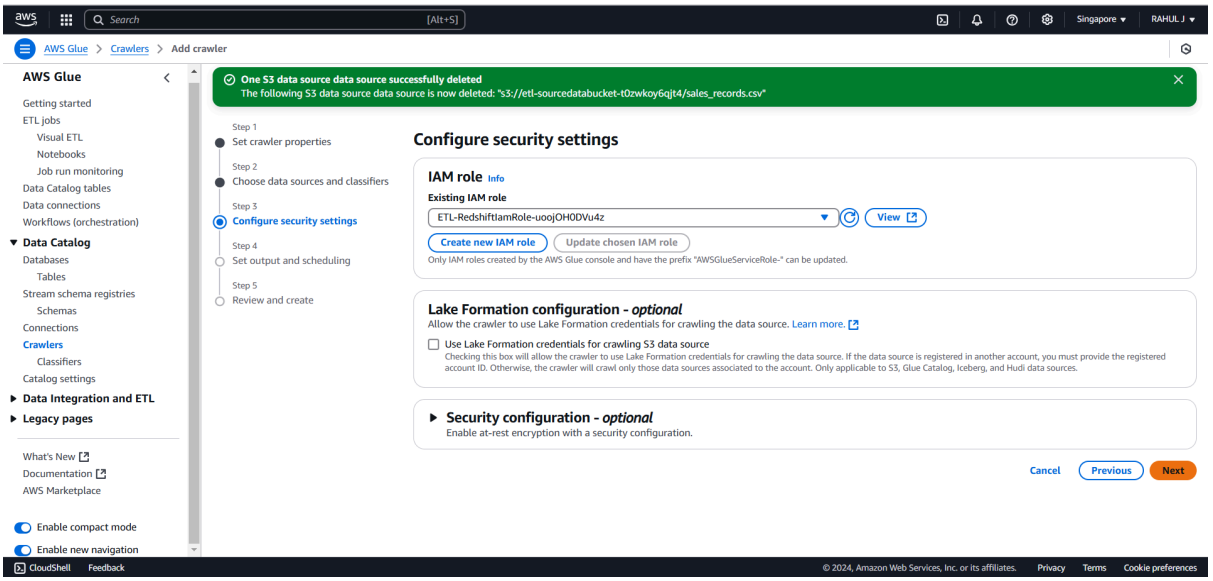
After clicking on **Create Database,** Now we need to choose our source file which is from our **S3 bucket** where we uploaded our file for that click on **Add a data source**



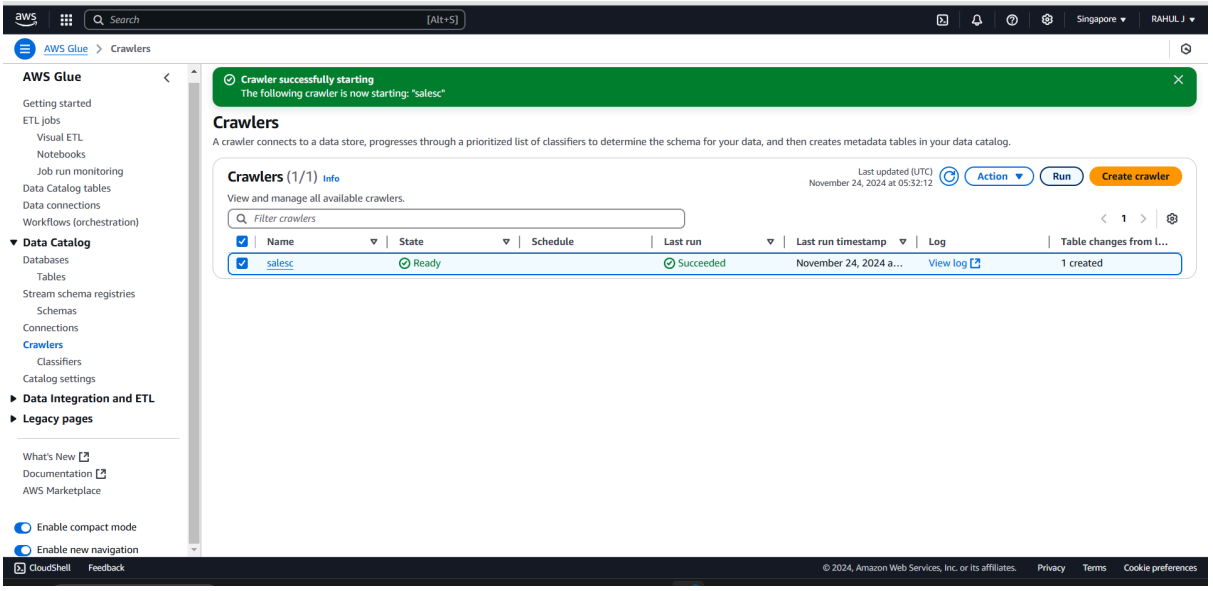


After **choosing** the file click on **next** then we need to choose the **IAM role** which we created before selecting the role

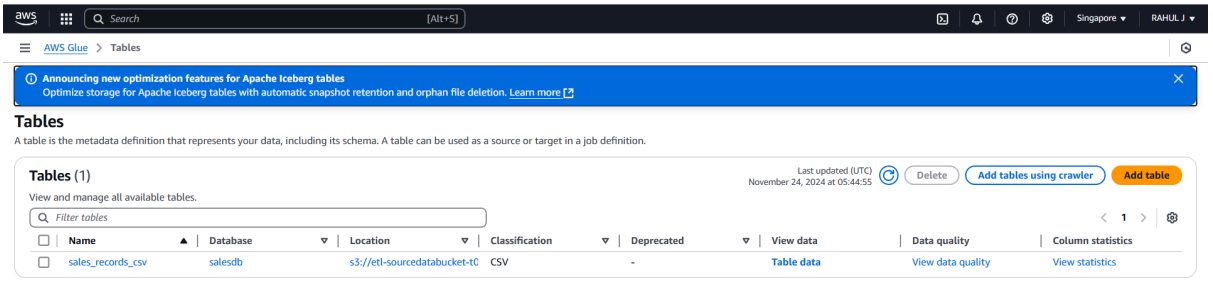
After choosing IAM Role click on next



Then click on Next for all and finally click on **Create Crawler.** After creating the crawler we need to run it

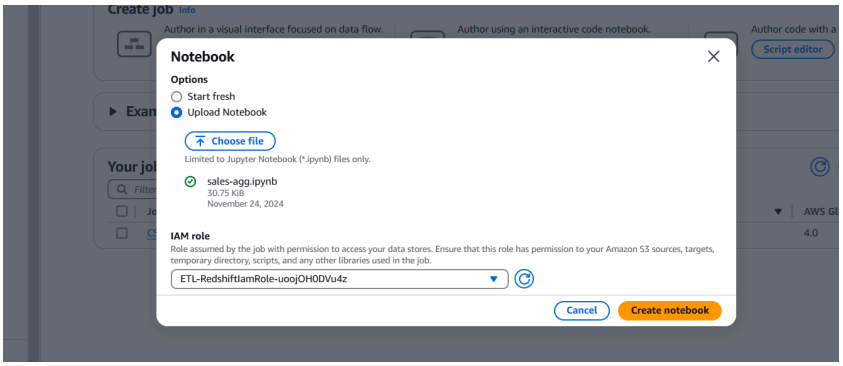


After **running** the crawler we can go and check for **Table** in **AWS Glue** itself

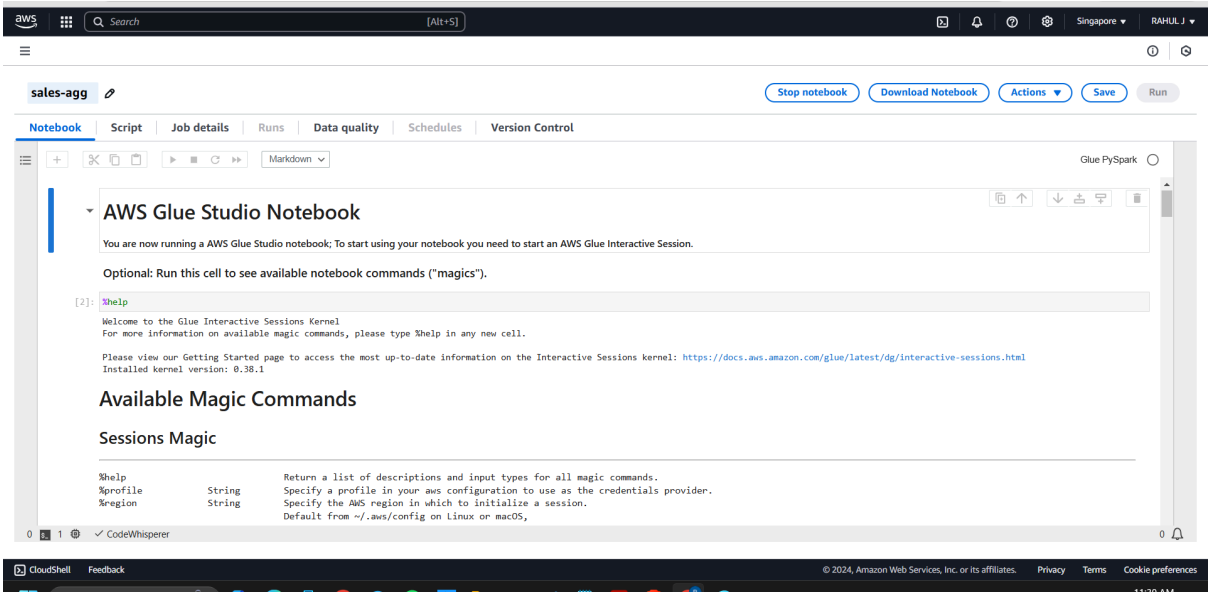


Creating an ETL Jobs

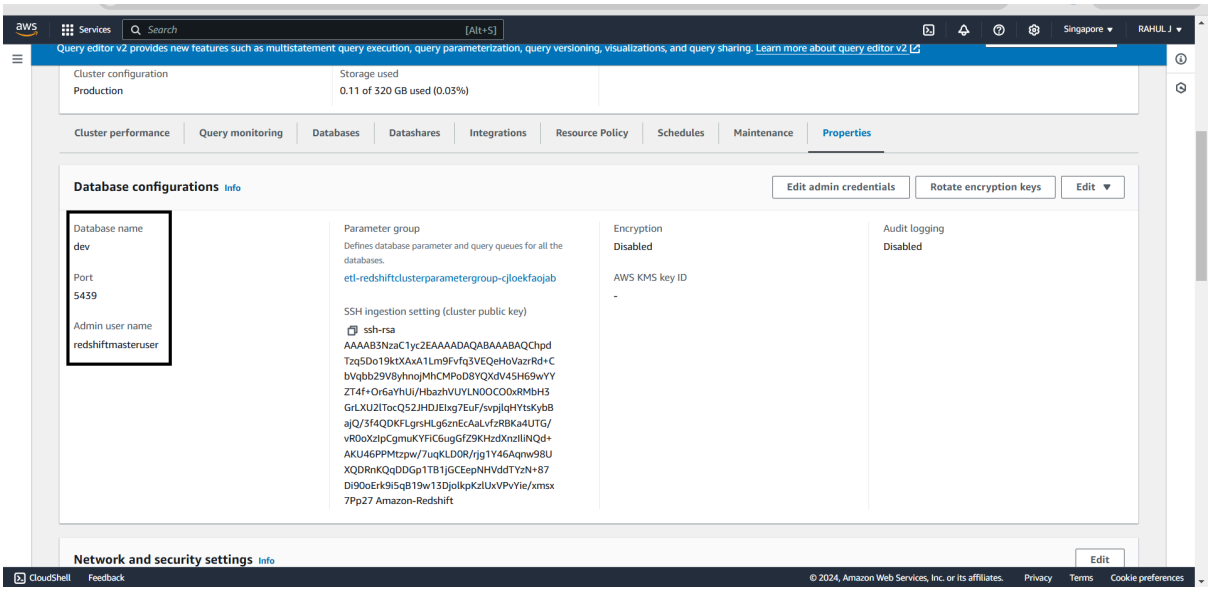
Now go to the **ETL Jobs** in choose **notebook** if we have an existing **IPYNB file** then we can upload it from our local machine and choose the **IAM Role** for the notebook



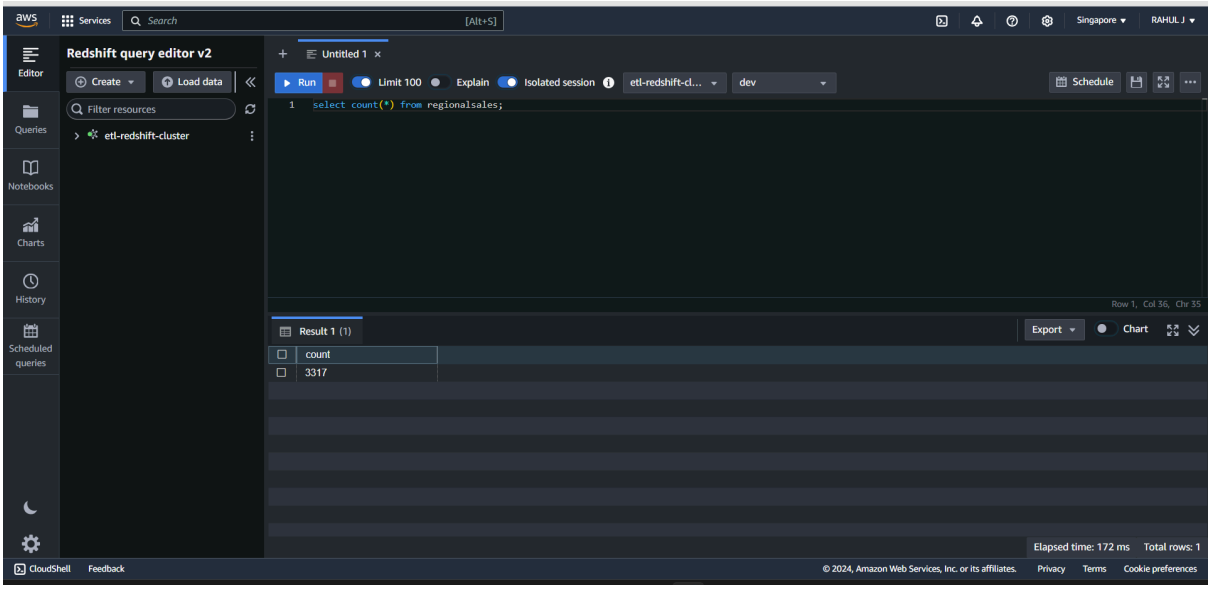
After **Creating notebook** we need to run our notebook



Now Go to the **redshift cluster** and open the **query editor v2**. If you are not connected to the database click on edit connection and connect your database. To connect to the database **go to your cluster** in the cluster and go to the properties section in that check for the **User and database name.**



In Query Editor v2 add the **database name** and **the username**. Now try to execute some query and check the output.


aws
Services
Search
[Alt+S]
Singapore
RAHUL J
Editor
Queries
Notebooks
Charts
History
Scheduled queries
Redshift query editor v2
Create
Load data
Filter resources
etl-redshift-cluster
Untitled 1
Run
Limit 100
Explain
Isolated session
etl-redshift-cl...
dev
Schedule
1 select count(*) from regionalsales;
Row 1, Col 36, Chr 35
Result 1 (1)
Export
Chart
count
3317
Elapsed time: 172 ms
Total rows: 1
CloudShell
Feedback
© 2024, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences