

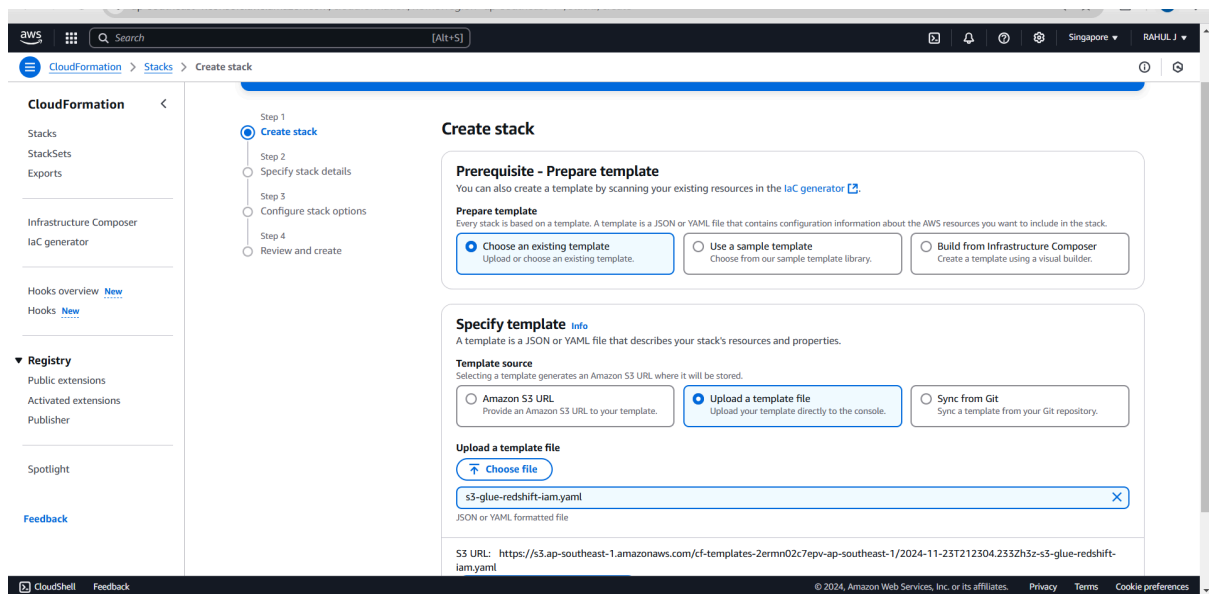
# AWS Project - 2(Building End-To-End Datapipeline)

## 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.

Amazon S3

Account snapshot - updated every 24 hours All AWS Regions [View Storage Lens dashboard](#)

Storage lens provides visibility into storage usage and activity trends. [Learn more](#)

General purpose buckets Directory buckets

General purpose buckets (5) All AWS Regions [Copy ARN](#) [Empty](#) [Delete](#) [Create bucket](#)

Buckets are containers for data stored in S3.

Find buckets by name

Name	AWS Region	IAM Access Analyzer	Creation date
<a href="#">cf-templates-2ermn02c7epv-ap-southeast-1</a>	Asia Pacific (Singapore) ap-southeast-1	<a href="#">View analyzer for ap-southeast-1</a>	November 24, 2024, 02:53:04 (UTC+05:30)
<a href="#">eti-databucket-ytmb84bqzwy</a>	Asia Pacific (Singapore) ap-southeast-1	<a href="#">View analyzer for ap-southeast-1</a>	November 24, 2024, 02:59:36 (UTC+05:30)
<a href="#">eti-sourcedatabucket-t0zrkoy6sqit4</a>	Asia Pacific (Singapore) ap-southeast-1	<a href="#">View analyzer for ap-southeast-1</a>	November 24, 2024, 02:59:36 (UTC+05:30)
<a href="#">gluup00-rj</a>	Asia Pacific (Singapore) ap-southeast-1	<a href="#">View analyzer for ap-southeast-1</a>	November 23, 2024, 12:23:48 (UTC+05:30)
<a href="#">inputcsv-rj</a>	Asia Pacific (Singapore) ap-southeast-1	<a href="#">View analyzer for ap-southeast-1</a>	November 23, 2024, 11:05:18 (UTC+05:30)
<a href="#">transformed-json-rj</a>	Asia Pacific (Singapore) ap-southeast-1	<a href="#">View analyzer for ap-southeast-1</a>	November 23, 2024, 11:13:51 (UTC+05:30)

## Now we can use Redshift Cluster

Resources overview

Resource data for Asia Pacific (Singapore) Region.

Total nodes	On-demand nodes	Reserved nodes	Reserved nodes available (0 of 0 used)	Automated snapshots	Manual snapshots
2	2	0	0	0	0

Cluster overview (1) Any Status

Cluster	Status
<a href="#">eti-redshift-cluster</a>	Available

[View all clusters](#)

Cluster metrics

[Any Clusters](#) [Last hour](#) [View in CloudWatch](#)

[Number of queries](#) Database connections Disk space used CPU utilization

Count

Datashares

Authorize other AWS accounts to access datashares created in this AWS account. Associate or decline datashares from other AWS accounts.

[Require authorization](#) [Require association](#)

0 0

Alarms (0) [View in CloudWatch](#)

Alarm name

No ongoing alarms

Events (5)

Amazon Redshift

Provisioned clusters dashboard

Resources overview

Resource data for Asia Pacific (Singapore) Region.

Total nodes	On-demand nodes	Reserved nodes	Reserved nodes available (0 of 0 used)	Automated snapshots	Manual snapshots
2	2	0	0	0	0

Cluster overview (1) Any Status

Cluster	Status
<a href="#">eti-redshift-cluster</a>	Available

[View all clusters](#)

Cluster metrics

[Any Clusters](#) [Last hour](#) [View in CloudWatch](#)

[Number of queries](#) Database connections Disk space used CPU utilization

Count

Datashares

Authorize other AWS accounts to access datashares created in this AWS account. Associate or decline datashares from other AWS accounts.

[Require authorization](#) [Require association](#)

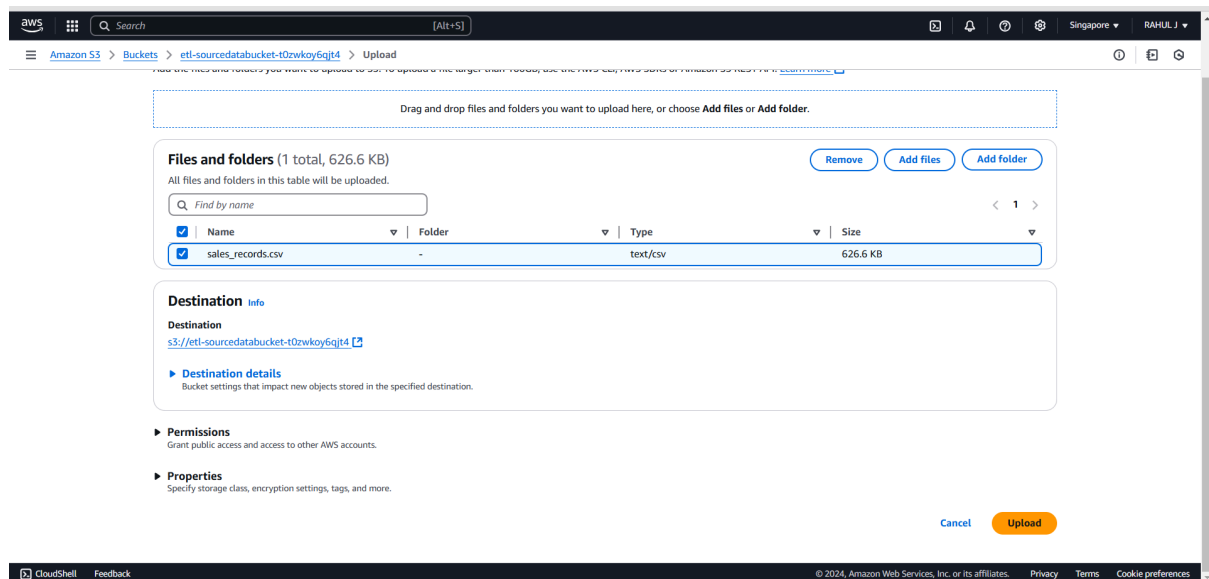
0 0

Alarms (0) [View in CloudWatch](#)

Alarm name

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

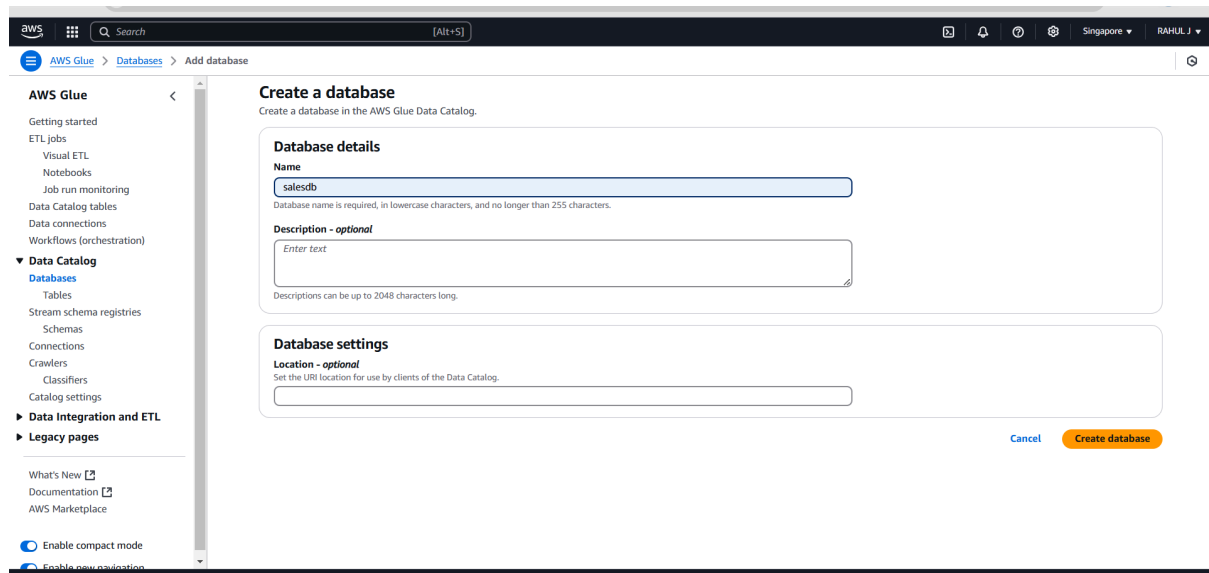
## Upload file in S3



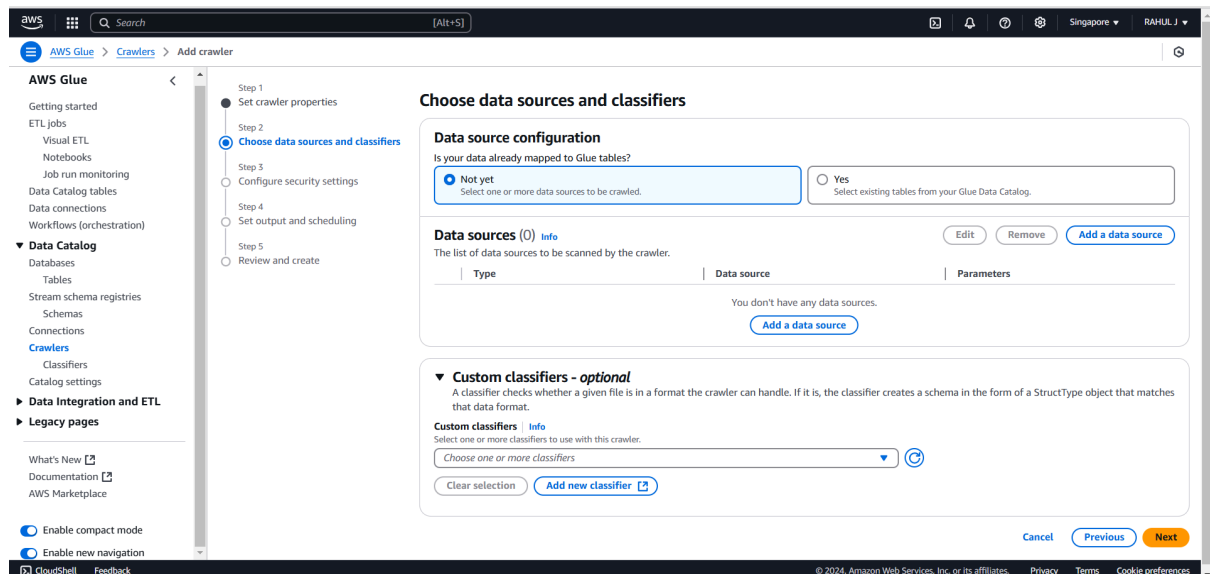
## Create a Crawler in the AWS Database

Go to the **AWS Glue** in **Datacatalog** and click on **Add Database**

In that give the name of the database

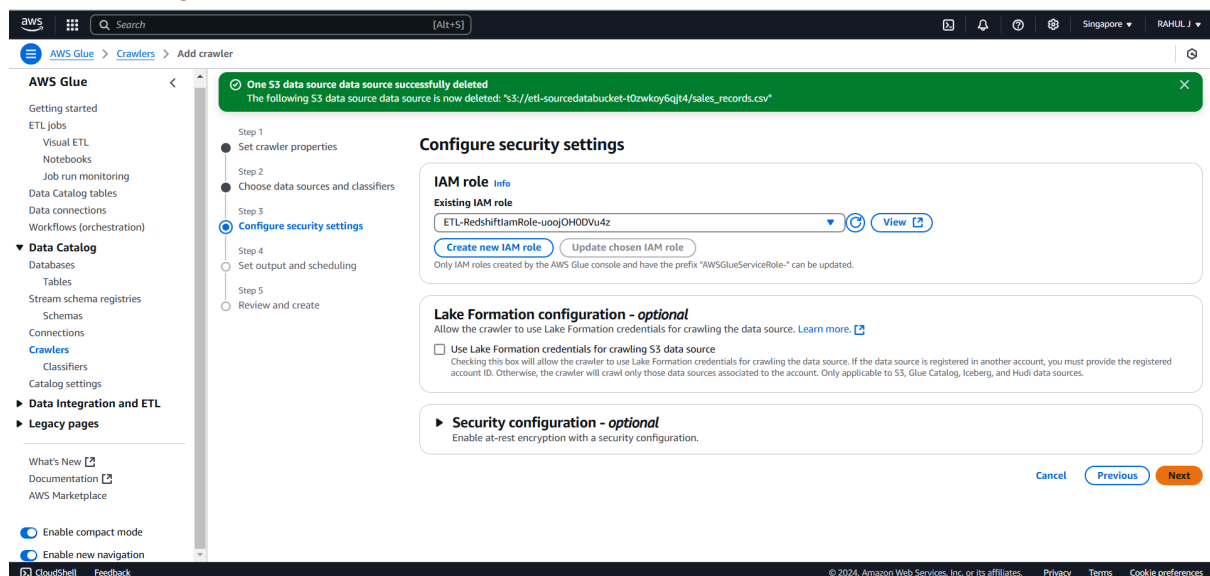


After clicking on **Create Database**, Now we need to choose our source file which is from our **S3 bucket** where we uploaded our file for that click on **Add a data source**

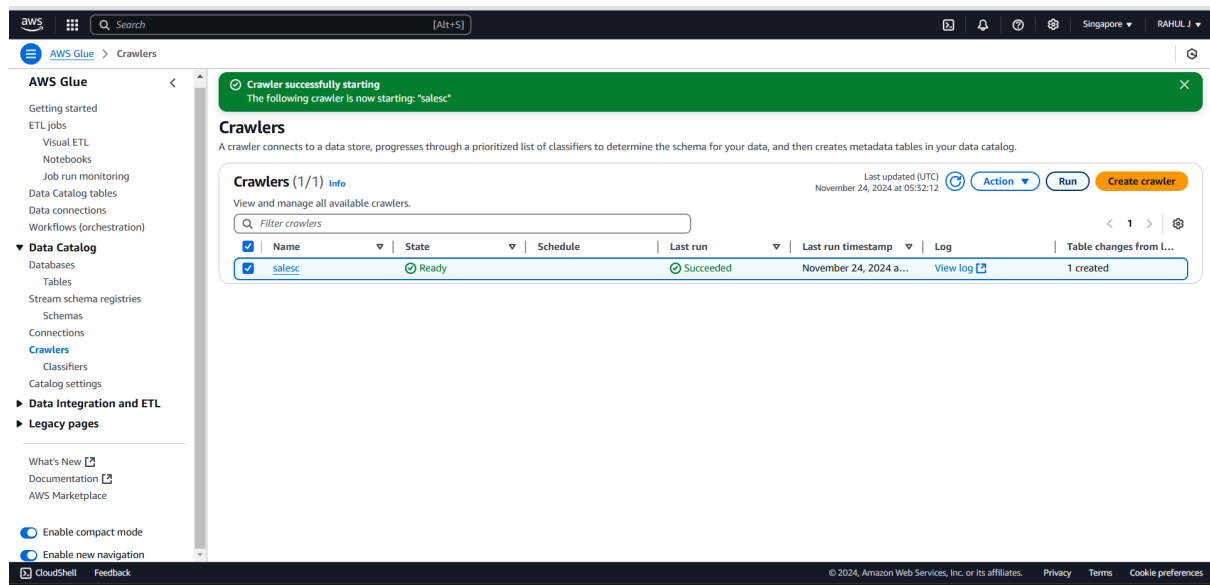


After **choosing** the file click on **next** then we need to choose the **IAM role** which we created before selecting the role

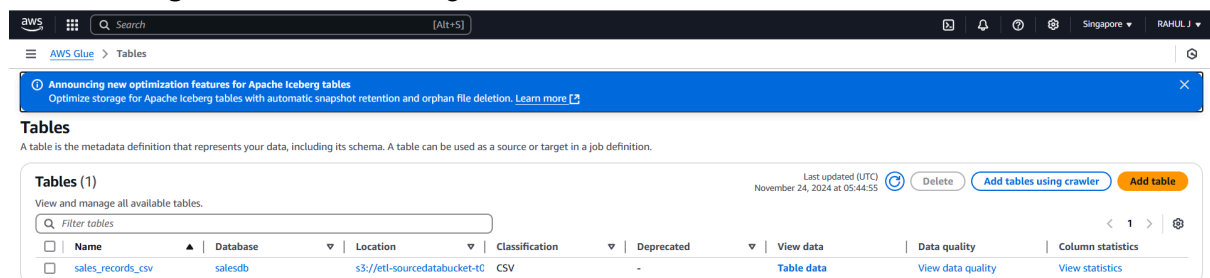
After choosing IAM Role click on next



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

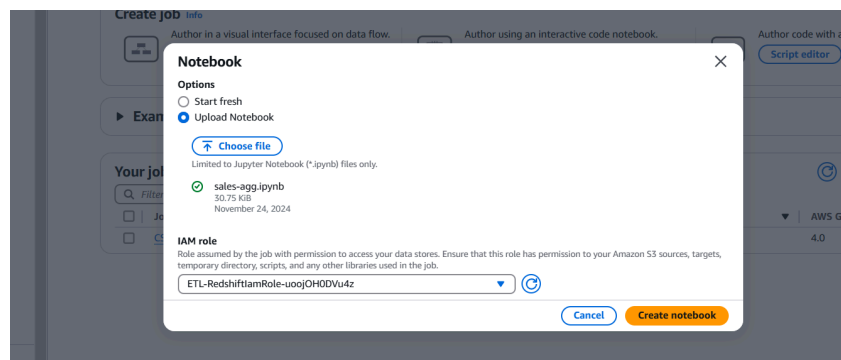


After running the crawler we can go and check for **Table** in **AWS Glue** itself

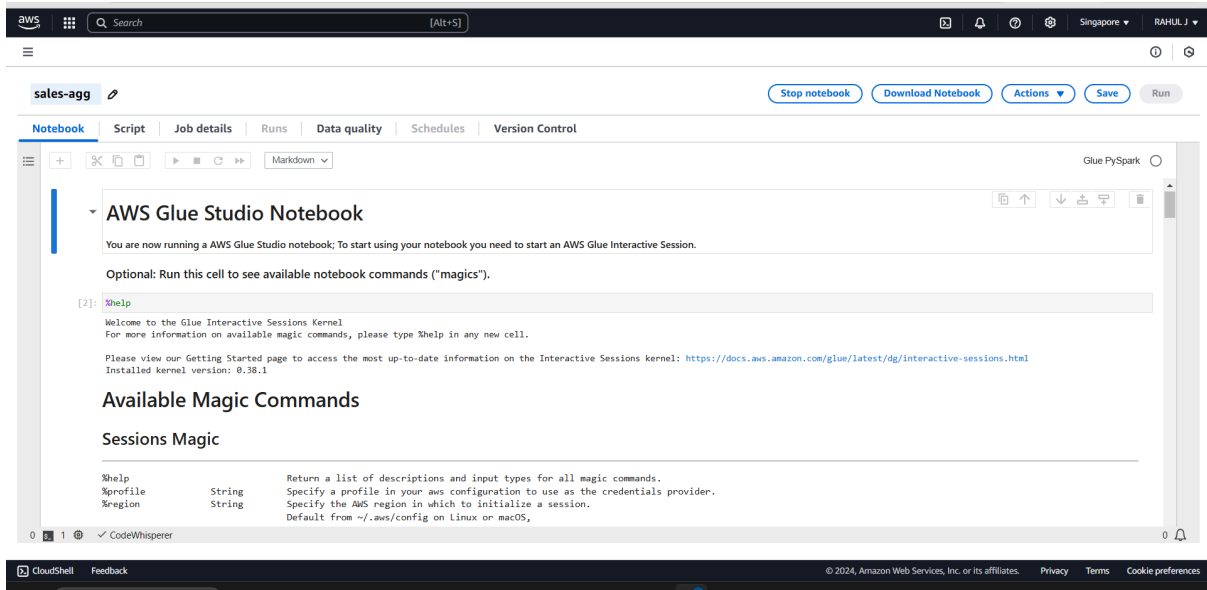


## Creating an ETL Jobs

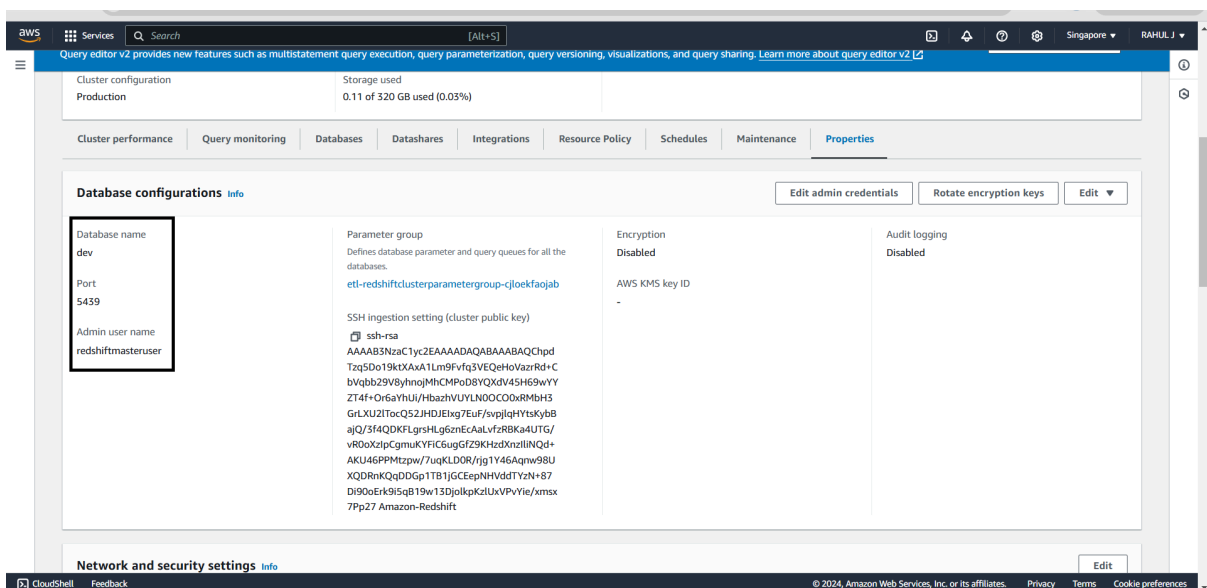
Now go to the **ETL Jobs** in choose **notebook** if we have an existing **IPYNB** file then we can upload it from our local machine and choose the **IAM Role** for the notebook



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

CreateLoad data

Filter resources

etl-redshift-cluster

Untitled 1

RunLimit 100ExplainIsolated session

eti-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