AWS Data Pipeline Setup for ETL Processing

Problem Statement

Organizations often struggle with managing and processing large volumes of data efficiently due to manual processes and a lack of automation. This leads to increased time for data handling and potential errors, hindering timely insights.

Objectives

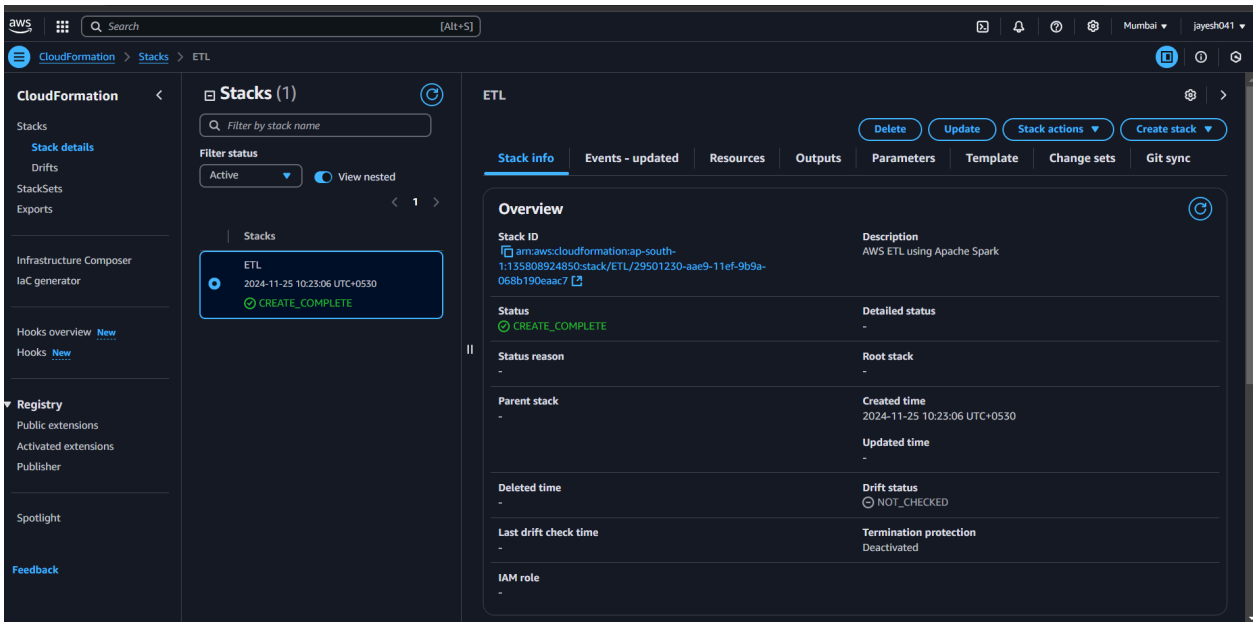
1. Set up the necessary AWS infrastructure to process and analyze data.
2. Automate data cataloguing using AWS Glue.
3. Perform ETL (Extract, Transform, Load) operations to prepare the data for analysis.
4. Validate and derive insights from the processed data using Redshift Query Editor v2.

AWS Services Used

- **Amazon S3**: For data storage.
- **Amazon Redshift**: For data warehousing and analysis.
- **AWS Glue**: For data cataloguing and ETL operations.
- **Amazon Redshift Query Editor v2**: For querying and validating data.
- **IAM (Identity and Access Management)**: For role and policy management.
- **CloudFormation**: Infrastructure as code (IaC) to automate resource creation.

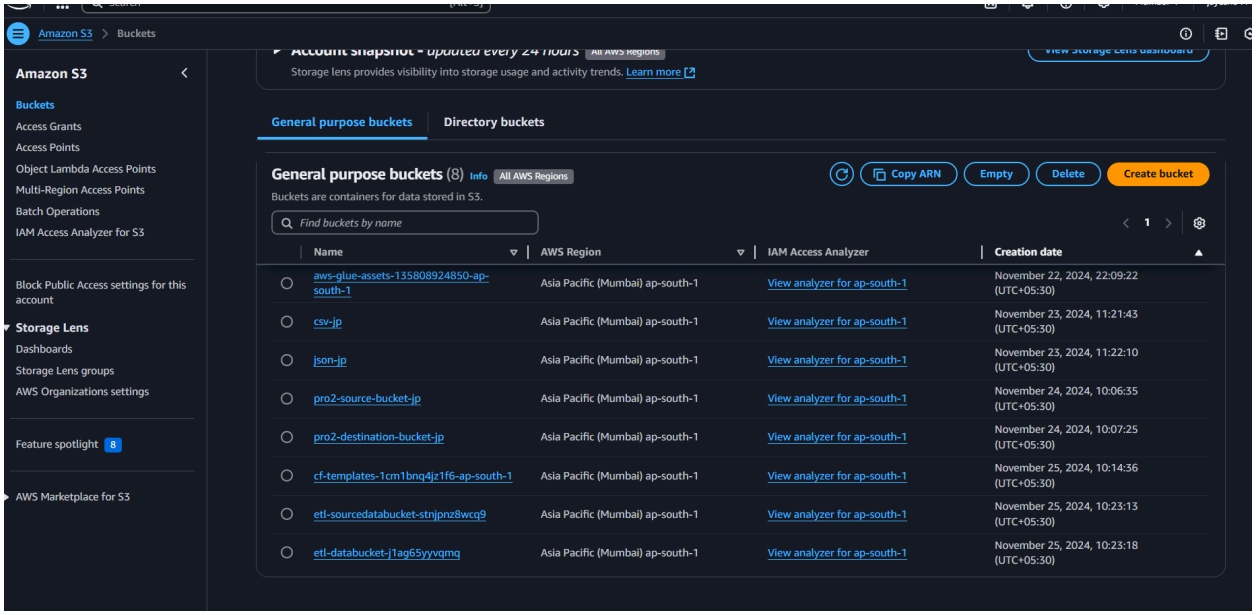
Step 1 : - Infrastructure Setup with CloudFormation

AWS CloudFormation is an Infrastructure-as-Code (IaC) service that simplifies the management of AWS resources by automating their creation and configuration.



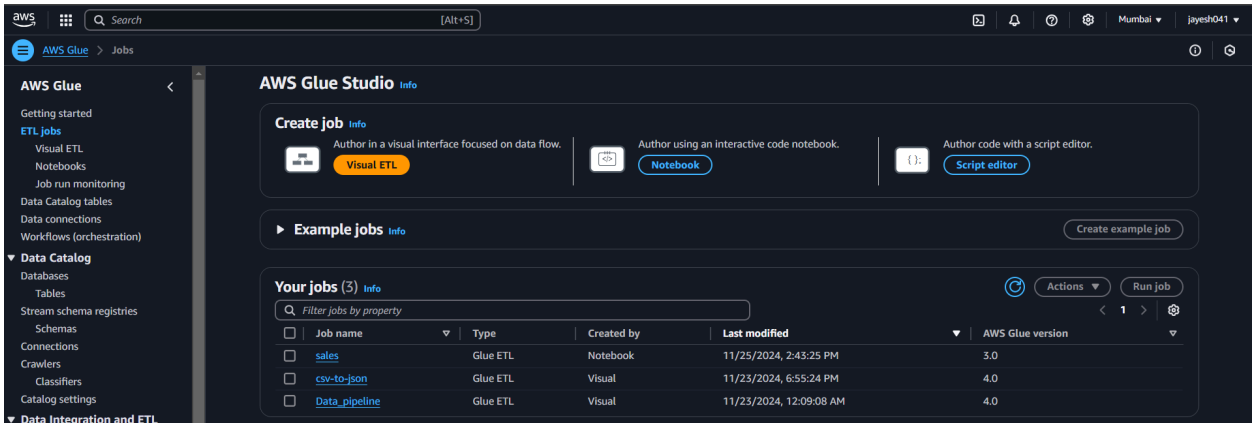
Step 2 :- Now we need to check for the S3 bucket, Redshift cluster ,Glue and its connection .

1 :- Check S3 Buckets .

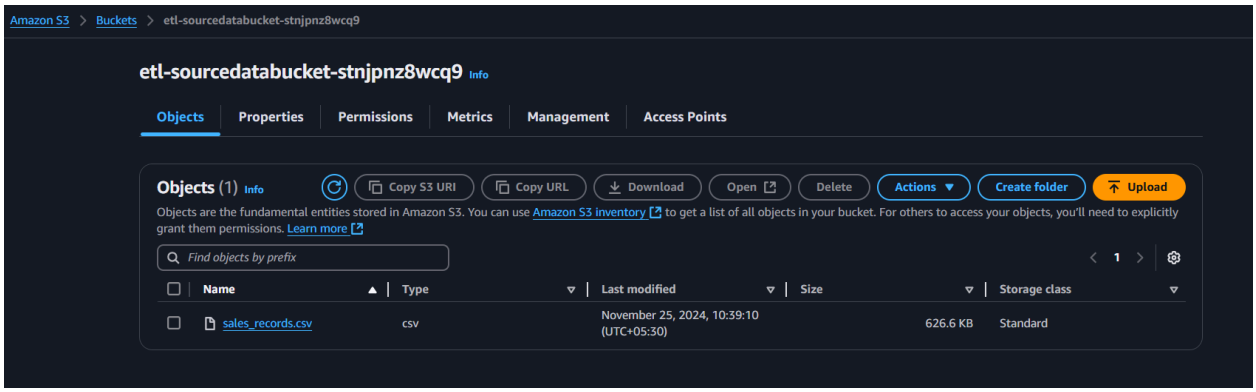


2 :- Check Redshift Cluster

3 : - Check AWS Glue

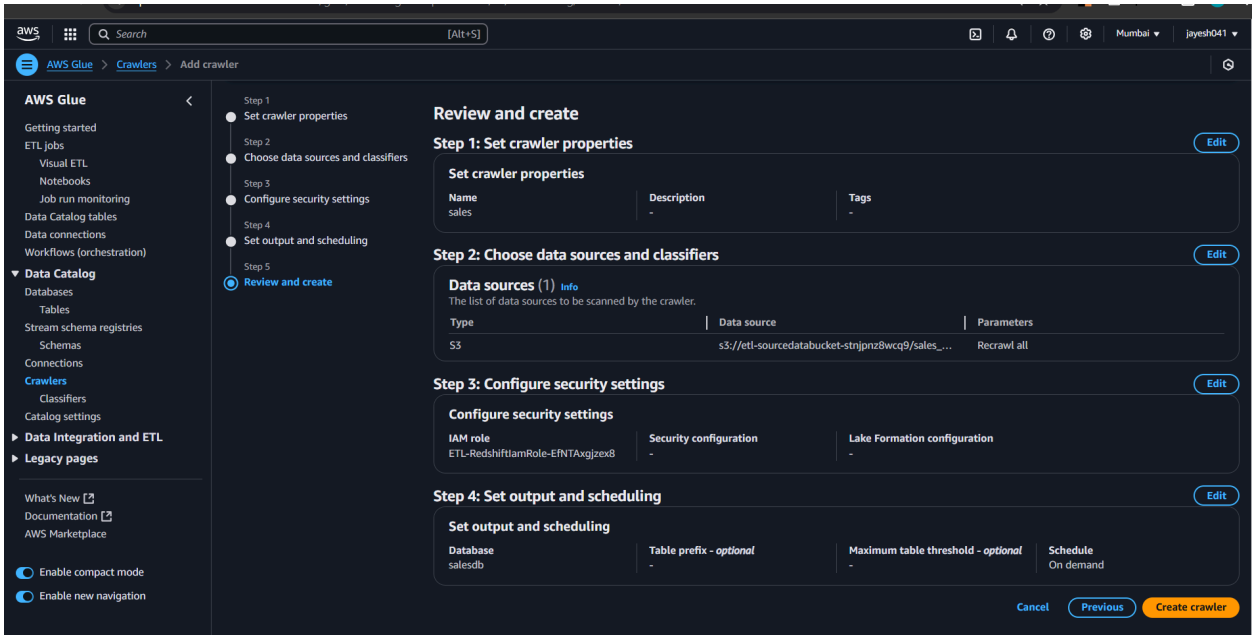


Step 3 :- Upload Dataset in S3



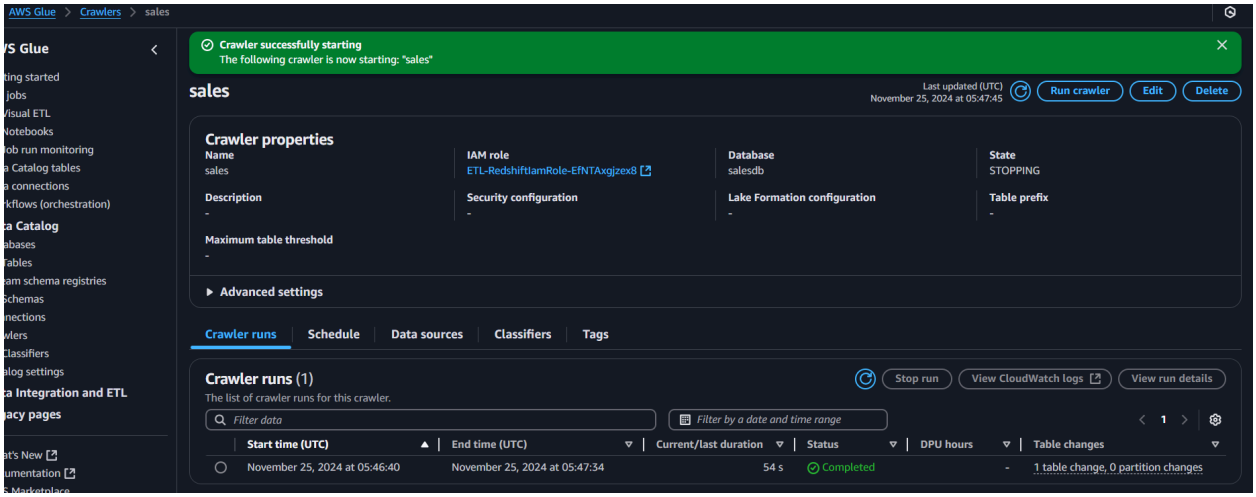
Step 4 :- Create an AWS Glue Crawler

- **Navigate to AWS Glue Service**
  In the AWS Management Console, go to the Glue service.
- **Create a Crawler**
  - Click on "Crawlers" > "Add crawler.
  - Name your crawler (e.g., my-data-crawler).
  - Choose "Data stores" as the crawler source type.
  - Select "S3" and specify your source bucket (e.g., source-bucket-name).
  - Specify the IAM role you created earlier (e.g., GlueServiceRole).
  - Click "Next" and configure options for output (databases, tables).
  - Review and finish creating the crawler.



- **Run the Crawler**

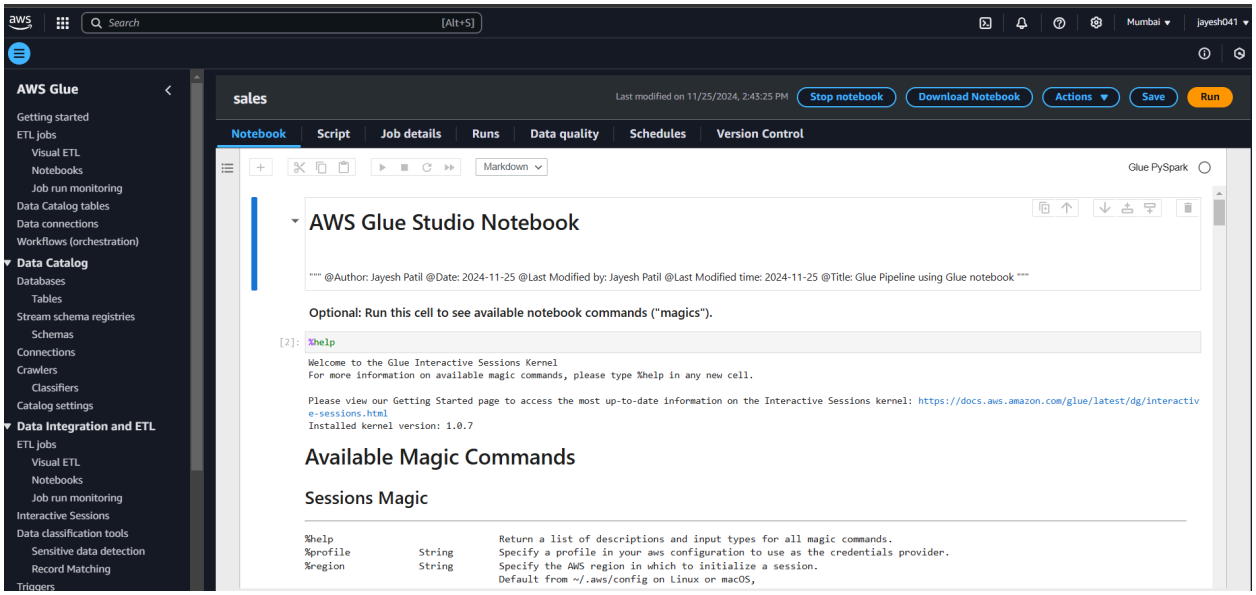
- Go to the Crawlers page, select your crawler, and click "Run crawler."



Step 5 :- Create a Glue Notebook Job

1. Navigate to Glue Jobs

- In the Glue service, click on "Jobs" .
- Click on **Add notebook**.
- Name your job (e.g., my-glue-notebook-job).○ Choose the IAM role you created earlier (e.g., GlueServiceRole)
- Write and Save Glue Notebook Code.
- Click "Next" and configure any additional options (like job bookmarks).
- Review and create the job.



2. **Save the Notebook Job**

○ Save your notebook and ensure it’s in the correct format.

Step 6 :- check the data in redshift Cluster

- Open the **Redshift Cluster** in the AWS Management Console.
- Navigate to the **Query Editor v2** section.
- If not already connected to the database, click on **Edit Connection** to establish a connection.
- To connect, locate your cluster and go to the **Properties** section.
- In the **Properties** section, find the details for **User** and **Database Name**.
- Use these details to configure the connection and connect to your database.
- In Query Editor v2 add the **database name** and **the username**. Now try to execute some query and check the output.