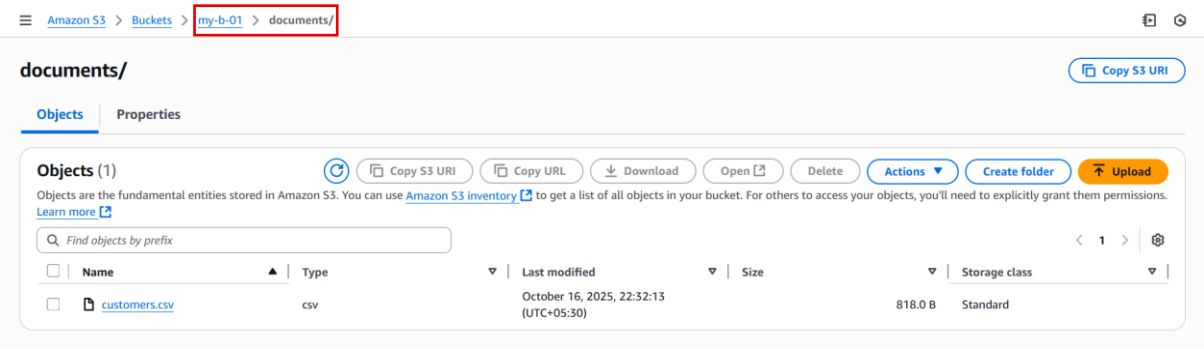
Setting Up Crawlers

To Begin with the Lab

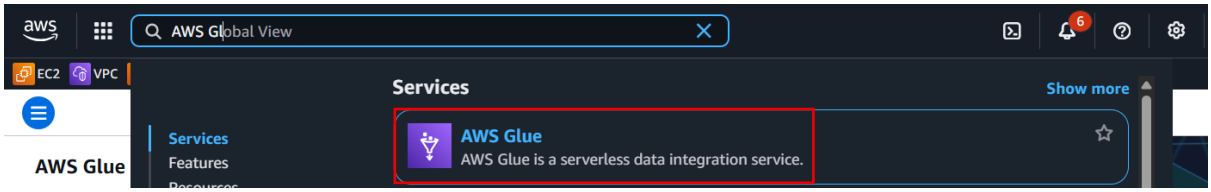
Summary of the Lab.

To set up an AWS Glue Crawler, first ensure you have an S3 bucket with a **"documents"** folder containing data files. In the AWS Console, go to **Glue Service** (same region as S3). Under **Data Catalog**, select **Crawlers → Create crawler**. Name it, add your **S3 data source**, and create or select an **IAM role** with S3 and Glue permissions. Then create a **database**, optionally add a table prefix, and complete the setup. Run the crawler to detect metadata. Once completed, view results under **Databases → your database → Tables** to see the created table and schema details.

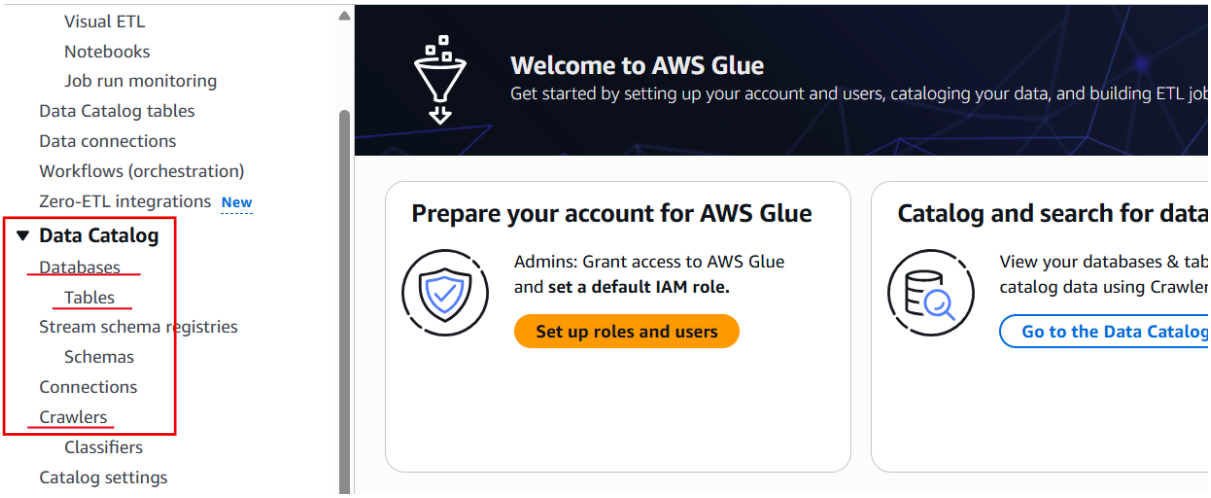
- Prerequisite for the Lab:
    - Folder created with the name "documents" inside the S3 Buckets with the files inside the folder.



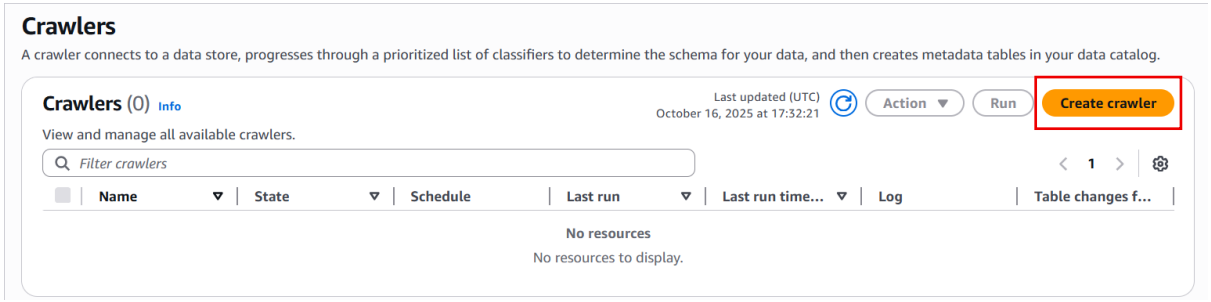
- Go to **AWS Console → Glue Service** (ensure you're in the same region as your S3 bucket).



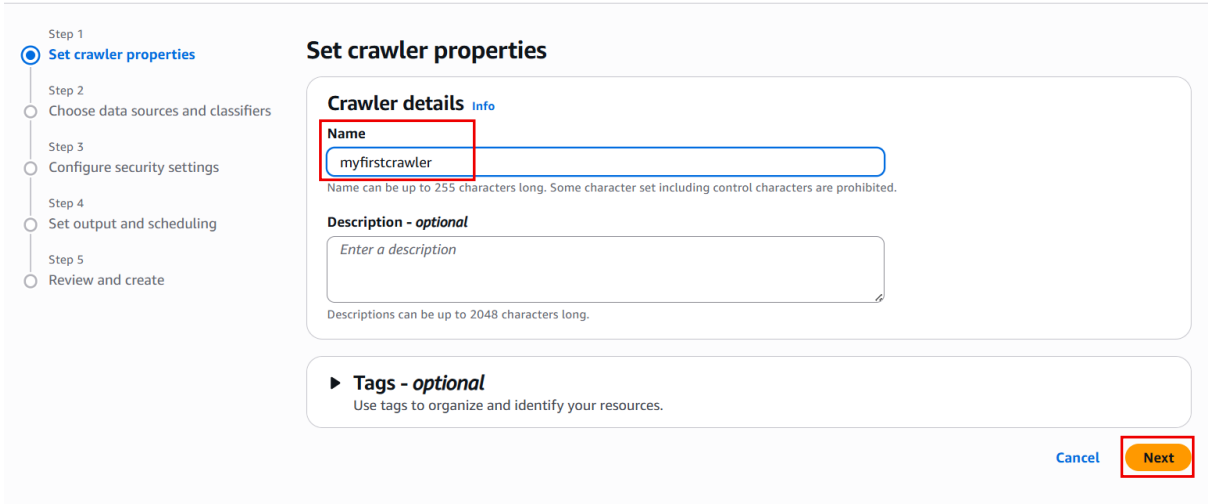
- In the left navigation panel, under **Data Catalog**, you'll find:
    - **Databases** → containers for tables
    - **Tables** → store metadata detected by crawlers
    - **Crawlers** → scan data sources and populate tables



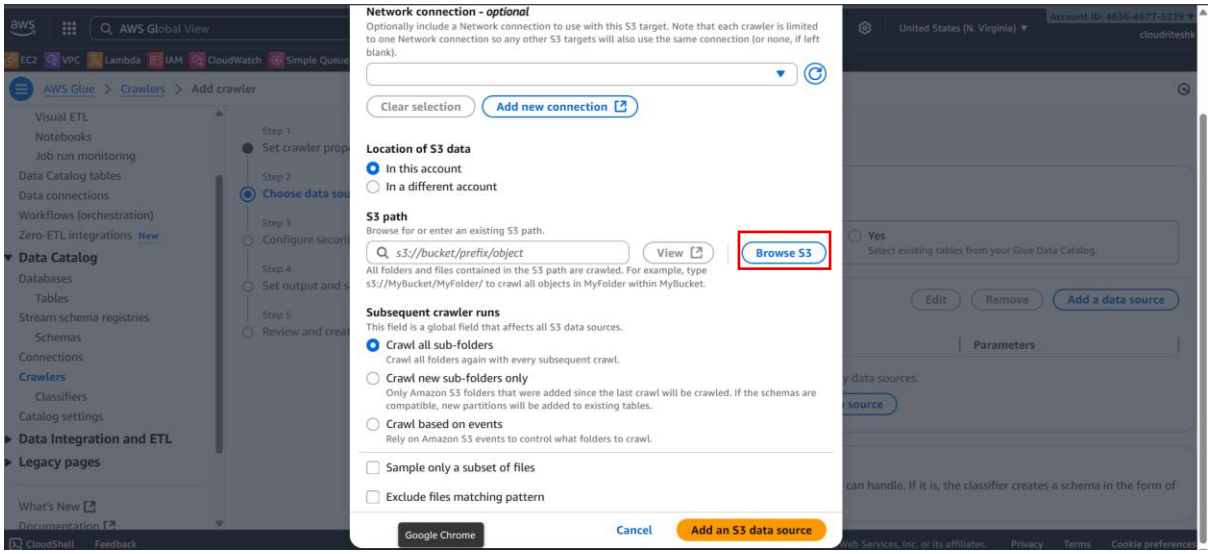
- Now Click on the Crawlers and then click on Create crawler.



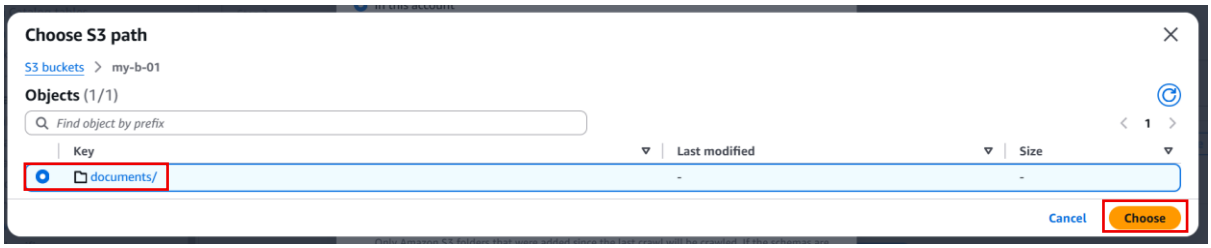
- Name the Crawler and click next.



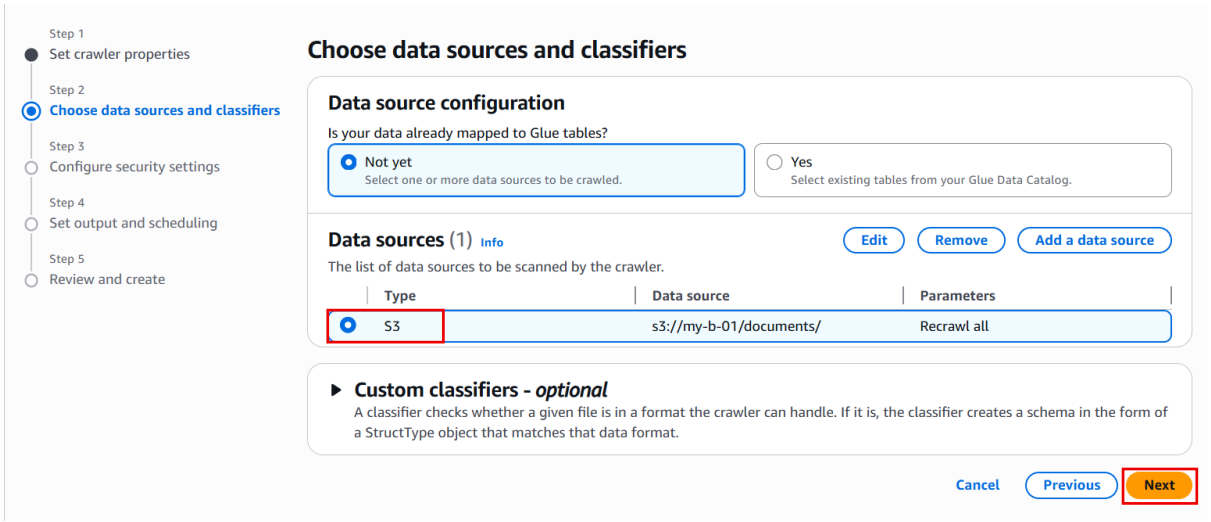
- Now click on the Add a data source.
- Click on the Browse S3



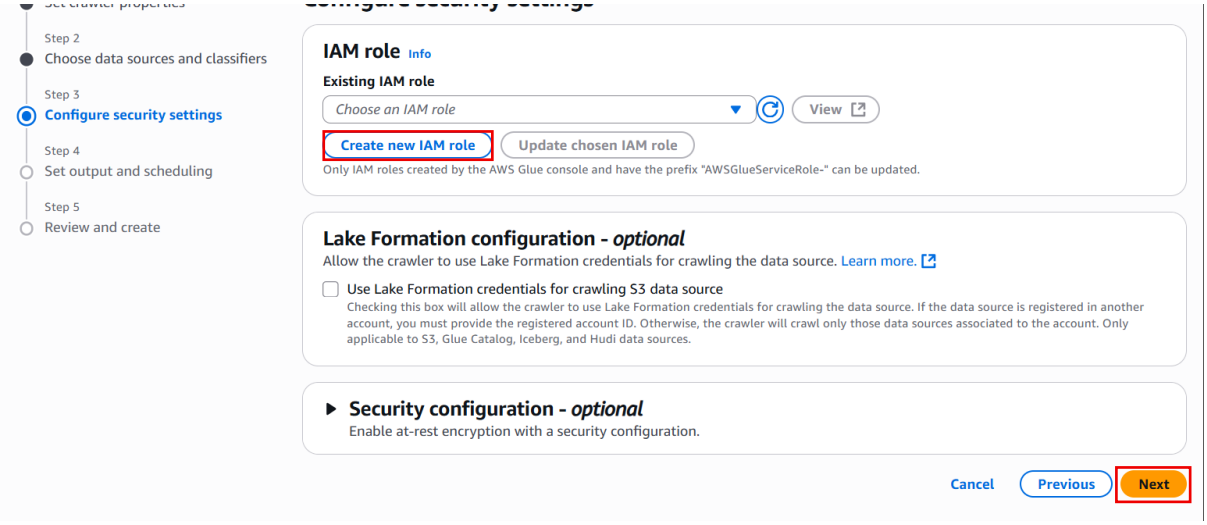
- Choose the folder and click choose.



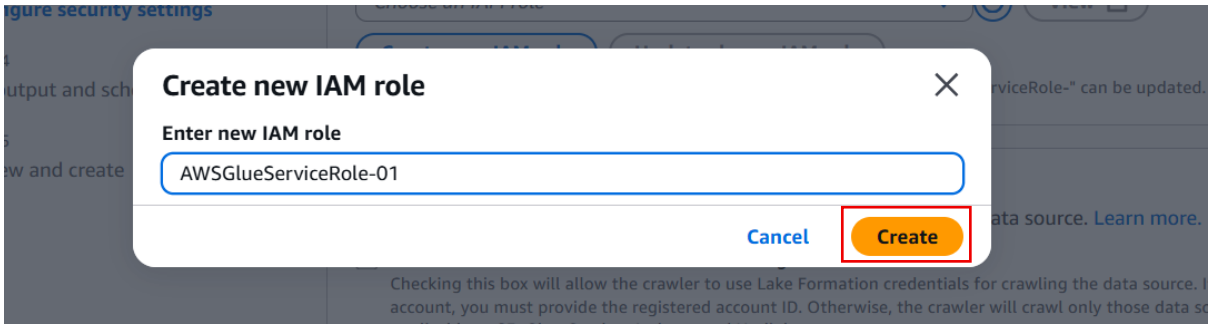
- Then, Click on Add an S3 data source.
- Check the Data Sources and Click Next.



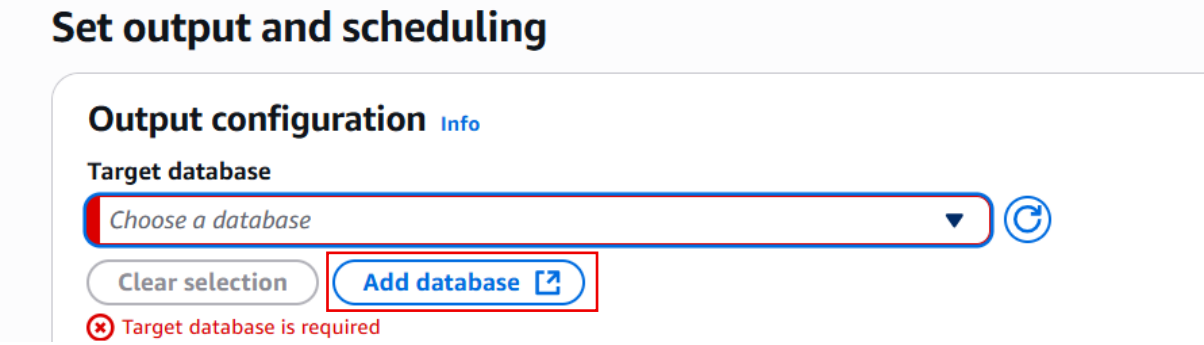
- Create an **IAM role** with S3 and Glue permissions
- Click Create new IAM role



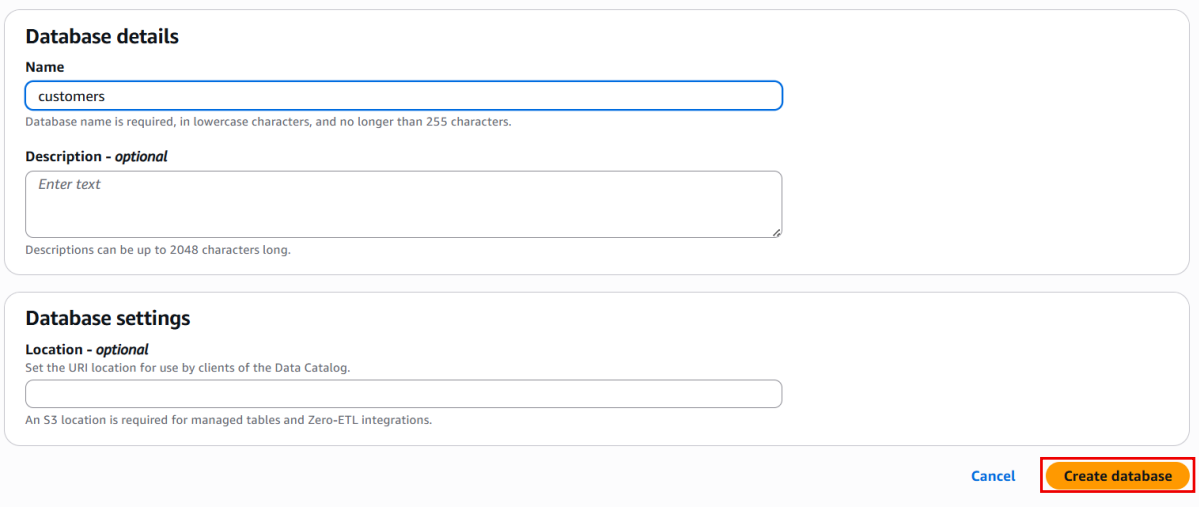
- If you want to rename, do it otherwise click on the Create.



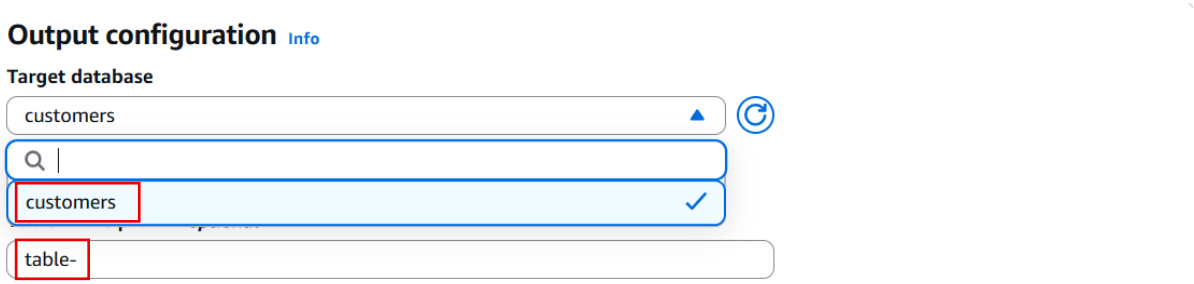
- Click Next.
- Click on the Add database



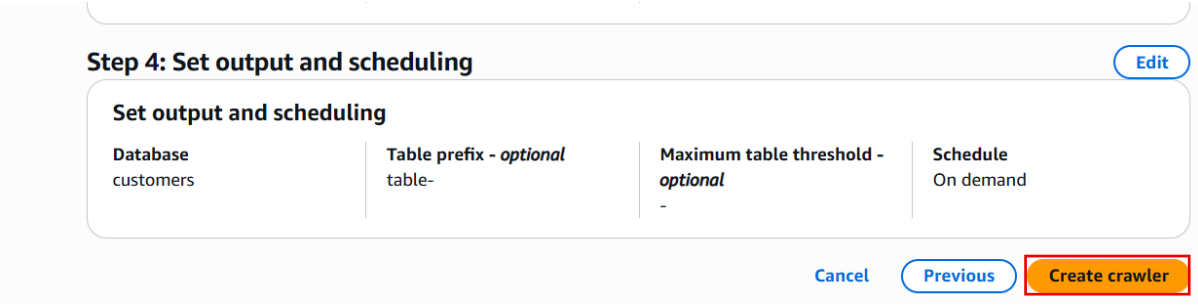
- Give the database name and click Create.



- Now go back to the clawer setup and refresh, you will see the database.
- Optional add a **table prefix** and click Next.

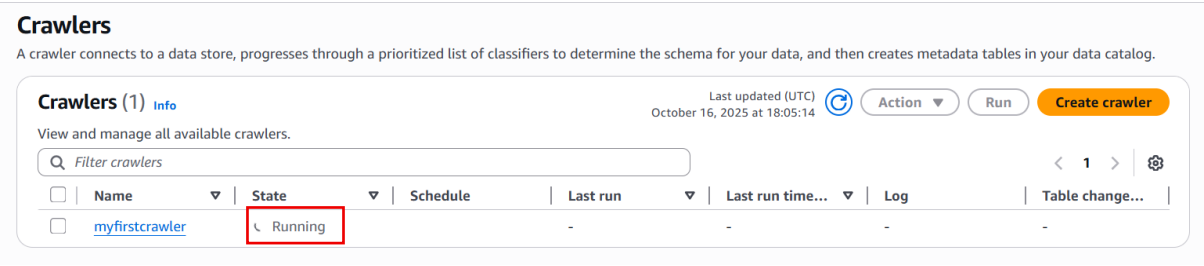


- Then Click on Create crawler.

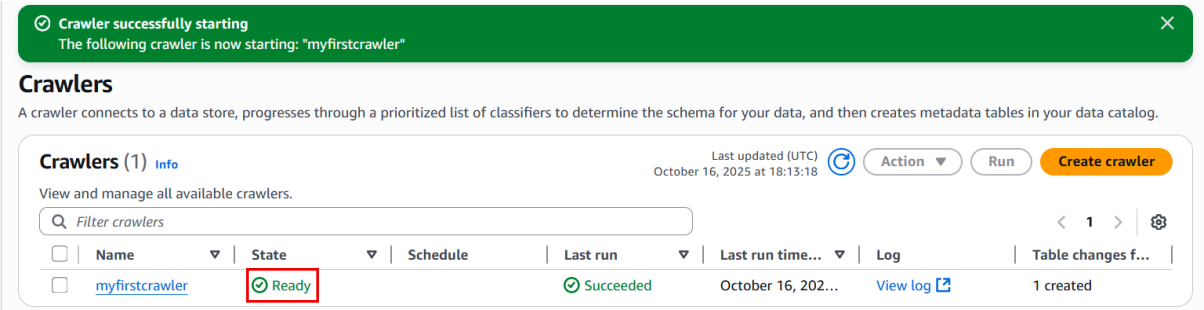


- After the crawler is create, Click on the Run Crawler.

- You can see crawler running.



- When it is completed, it will show ready.



- If you want to se the table detail, click on Go to **Databases → your database → Tables**.
- You'll see the new table created by the crawler.