

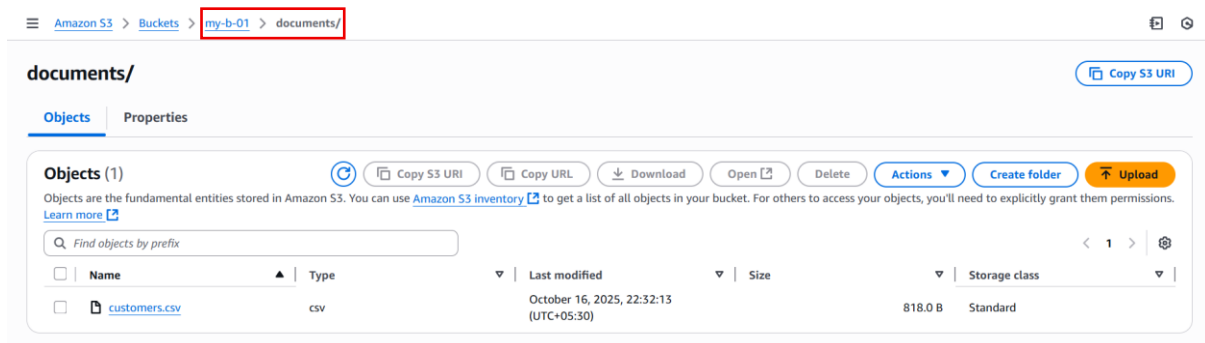
Setting Up Crawlers

To Begin with the Lab

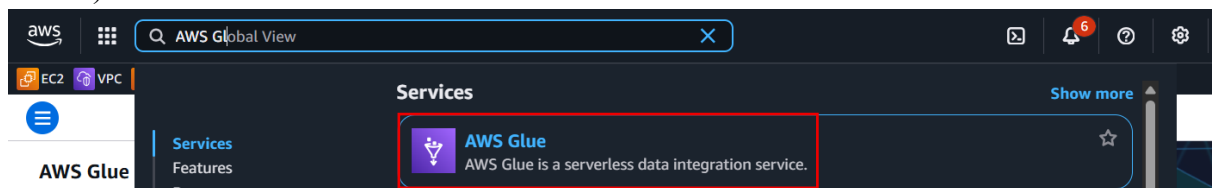
Summary of the Lab.

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

Visual ETL
Notebooks
Job run monitoring
Data Catalog tables
Data connections
Workflows (orchestration)
Zero-ETL integrations [New](#)

▼ **Data Catalog**
[Databases](#)
[Tables](#)
 Stream schema registries
 Schemas
 Connections
[Crawlers](#)
 Classifiers
 Catalog settings

Welcome to AWS Glue

Get started by setting up your account and users, cataloging your data, and building ETL jobs.

Prepare your account for AWS Glue

Admins: Grant access to AWS Glue and **set a default IAM role.**

[Set up roles and users](#)

Catalog and search for data

View your databases & tables, catalog data using Crawler

[Go to the Data Catalog](#)

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

Crawlers

A crawler connects to a data store, progresses through a prioritized list of classifiers to determine the schema for your data, and then creates metadata tables in your data catalog.

Crawlers (0) [Info](#) Last updated (UTC) October 16, 2025 at 17:32:21 [Action](#) [Run](#) [Create crawler](#)

View and manage all available crawlers.

<input type="checkbox"/>	Name	State	Schedule	Last run	Last run time...	Log	Table changes f...
No resources							
No resources to display.							

- Name the Crawler and click next.

Set crawler properties

Step 1: **Set crawler properties**
 Step 2: Choose data sources and classifiers
 Step 3: Configure security settings
 Step 4: Set output and scheduling
 Step 5: Review and create

Crawler details [Info](#)

Name

Name can be up to 255 characters long. Some character set including control characters are prohibited.

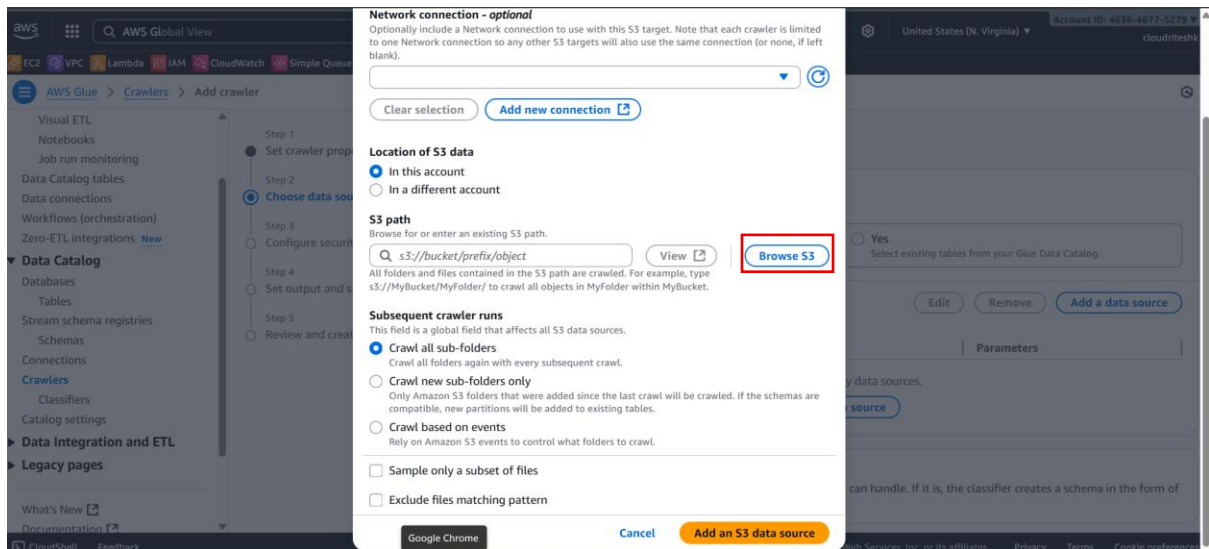
Description - optional

Descriptions can be up to 2048 characters long.

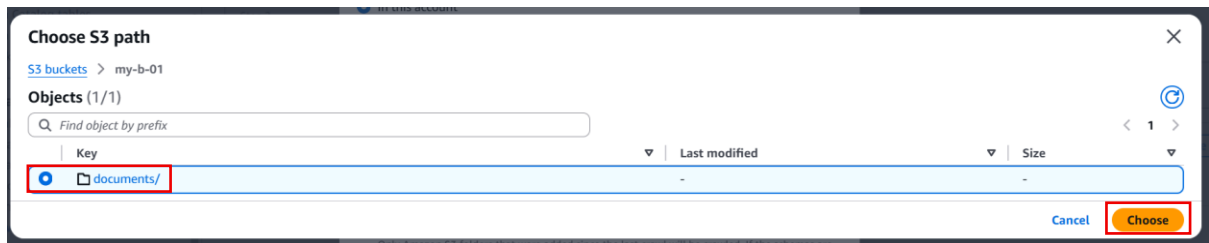
► **Tags - optional**
Use tags to organize and identify your resources.

[Cancel](#) [Next](#)

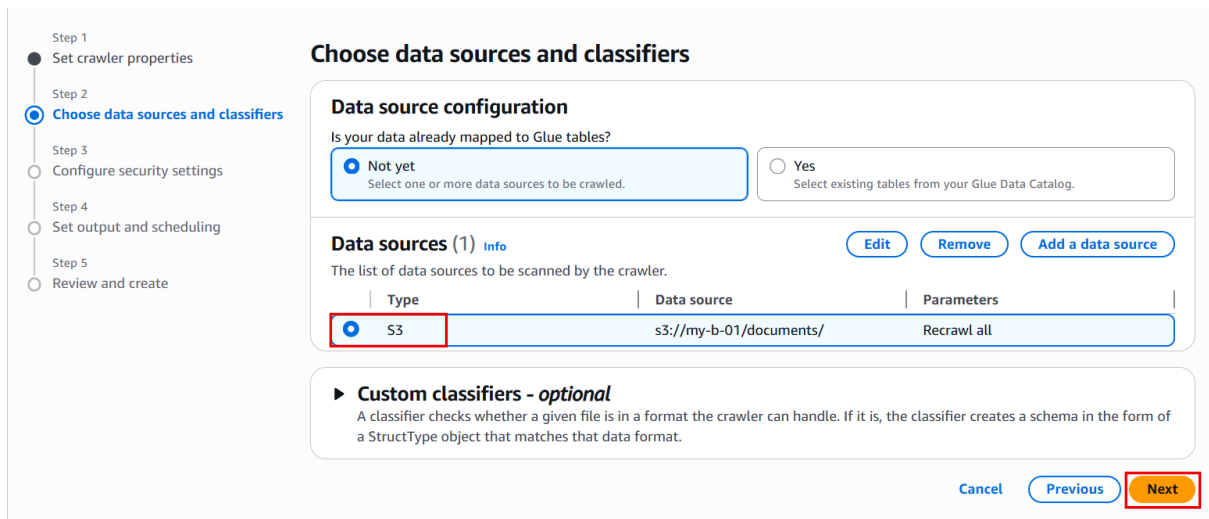
- Now click on the Add a data source.
- Click on the Browse S3



- 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

Step 2
● Choose data sources and classifiers

Step 3
● **Configure security settings**

Step 4
○ Set output and scheduling

Step 5
○ Review and create

IAM role [Info](#)

Existing IAM role

Choose an IAM role ▼ 🔄 View [🔗](#)

Create new IAM role Update chosen IAM role

Only IAM roles created by the AWS Glue console and have the prefix "AWSGlueServiceRole-" can be updated.

Lake Formation configuration - optional

Allow the crawler to use Lake Formation credentials for crawling the data source. [Learn more.](#) [🔗](#)

☐ **Use Lake Formation credentials for crawling S3 data source**

Checking this box will allow the crawler to use Lake Formation credentials for crawling the data source. If the data source is registered in another account, you must provide the registered account ID. Otherwise, the crawler will crawl only those data sources associated to the account. Only applicable to S3, Glue Catalog, Iceberg, and Hudi data sources.

▶ Security configuration - optional

Enable at-rest encryption with a security configuration.

[Cancel](#) [Previous](#) [Next](#)

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

Create new IAM role ✕

Enter new IAM role

[Cancel](#) [Create](#)

Checking this box will allow the crawler to use Lake Formation credentials for crawling the data source. If the data source is registered in another account, you must provide the registered account ID. Otherwise, the crawler will crawl only those data sources associated to the account. Only applicable to S3, Glue Catalog, Iceberg, and Hudi data sources.

- Click Next.
- Click on the Add database

Set output and scheduling

Output configuration [Info](#)

Target database

🔄

[Clear selection](#) [Add database \[🔗\]\(#\)](#)

❌ **Target database is required**

- Give the database name and click Create.

Database details

Name

Database name is required, in lowercase characters, and no longer than 255 characters.

Description - optional

Enter text

Descriptions can be up to 2048 characters long.

Database settings

Location - optional

Set the URI location for use by clients of the Data Catalog.

An S3 location is required for managed tables and Zero-ETL integrations.

Cancel
Create database

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

Output configuration [Info](#)

Target database

- Then Click on Create crawler.

Step 4: Set output and scheduling

[Edit](#)

Set output and scheduling

Database customers	Table prefix - optional table-	Maximum table threshold - optional -	Schedule On demand
-----------------------	-----------------------------------	---	-----------------------

Cancel
Previous
Create crawler

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

One crawler successfully created
The following crawler is now created: "myfirstcrawler"

myfirstcrawler

Last updated (UTC)
October 16, 2025 at 18:03:25

[Run crawler](#)
[Edit](#)
[Delete](#)

- You can see crawler running.

Crawlers

A crawler connects to a data store, progresses through a prioritized list of classifiers to determine the schema for your data, and then creates metadata tables in your data catalog.

Crawlers (1) [Info](#) Last updated (UTC) October 16, 2025 at 18:05:14 [Action](#) [Run](#) [Create crawler](#)

View and manage all available crawlers.

<input type="checkbox"/>	Name	State	Schedule	Last run	Last run time...	Log	Table change...
<input type="checkbox"/>	myfirstcrawler	Running		-	-	-	-

- When it is completed, it will show ready.

Crawler successfully starting
The following crawler is now starting: "myfirstcrawler"

Crawlers

A crawler connects to a data store, progresses through a prioritized list of classifiers to determine the schema for your data, and then creates metadata tables in your data catalog.

Crawlers (1) [Info](#) Last updated (UTC) October 16, 2025 at 18:13:18 [Action](#) [Run](#) [Create crawler](#)

View and manage all available crawlers.

<input type="checkbox"/>	Name	State	Schedule	Last run	Last run time...	Log	Table changes f...
<input type="checkbox"/>	myfirstcrawler	Ready		Succeeded	October 16, 202...	View log	1 created

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

Schema (7) [Edit schema as JSON](#) [Edit schema](#)

View and manage the table schema.

#	Column name	Data type	Partition key	Comment
1	id	bigint	-	-
2	name	string	-	-
3	age	bigint	-	-
4	address	string	-	-
5	city	string	-	-
6	state	string	-	-
7	email	string	-	-