Lesson 04 Demo 09

Building a Glue Data Catalog

Objective: To create a Glue Data Catalog using AWS Glue for seamless organization and efficient cataloging

Tools require: AWS Workspace

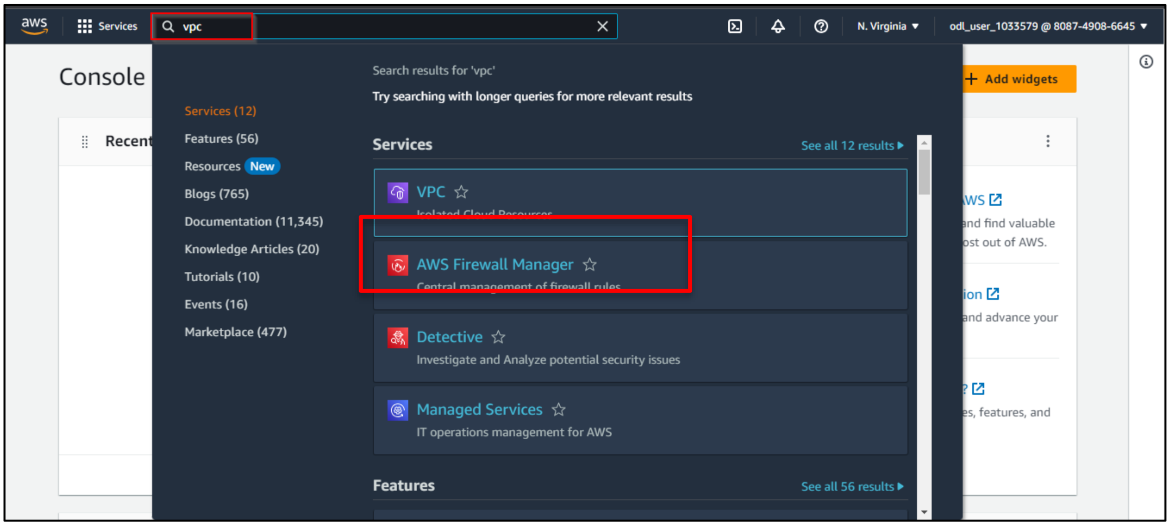
Prerequisites: AWS account

Steps to be followed:

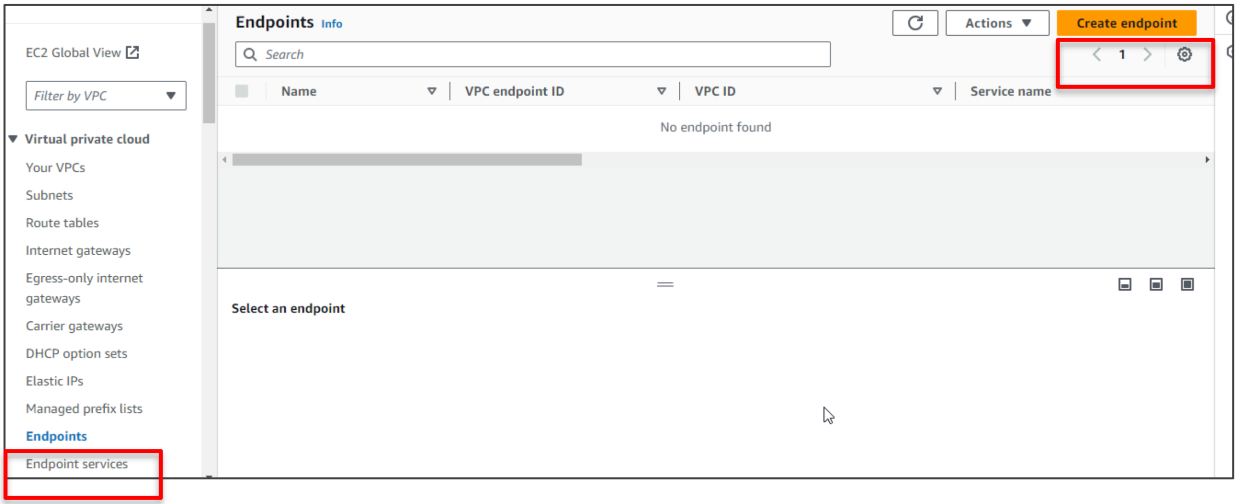
1. Create a VPC endpoint
2. Create a Glue Data Catalog
3. Add a data source to crawlers

Step 1: Create a VPC endpoint

1.1 Navigate to the AWS portal homepage, search for **VPC**, and click on it



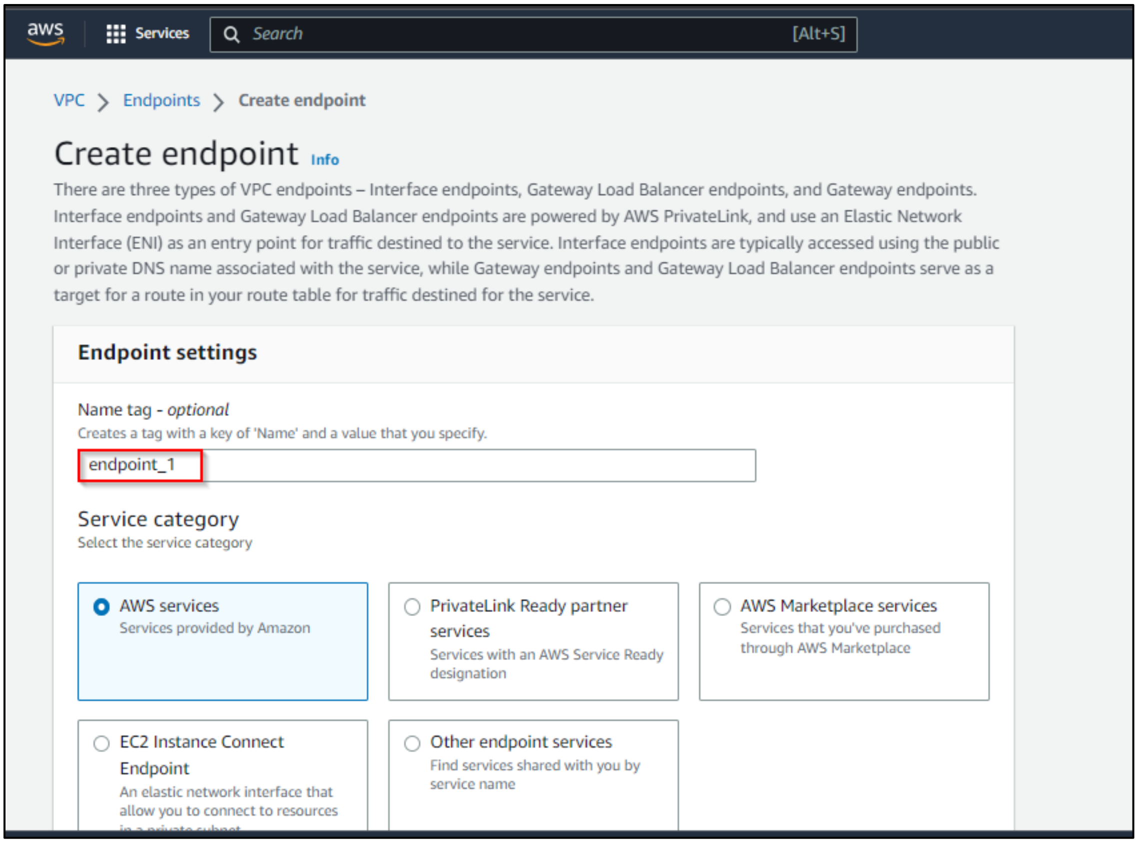
1.2 Click on **Endpoints** and then select **Create endpoint**



No endpoint found

Select an endpoint

1.3 Enter **endpoint_1** as the **Name tag**



Create endpoint Info

There are three types of VPC endpoints – Interface endpoints, Gateway Load Balancer endpoints, and Gateway endpoints. Interface endpoints and Gateway Load Balancer endpoints are powered by AWS PrivateLink, and use an Elastic Network Interface (ENI) as an entry point for traffic destined to the service. Interface endpoints are typically accessed using the public or private DNS name associated with the service, while Gateway endpoints and Gateway Load Balancer endpoints serve as a target for a route in your route table for traffic destined for the service.

Endpoint settings

Name tag - *optional*

Creates a tag with a key of 'Name' and a value that you specify.

endpoint_1

Service category

Select the service category

● AWS services
Services provided by Amazon

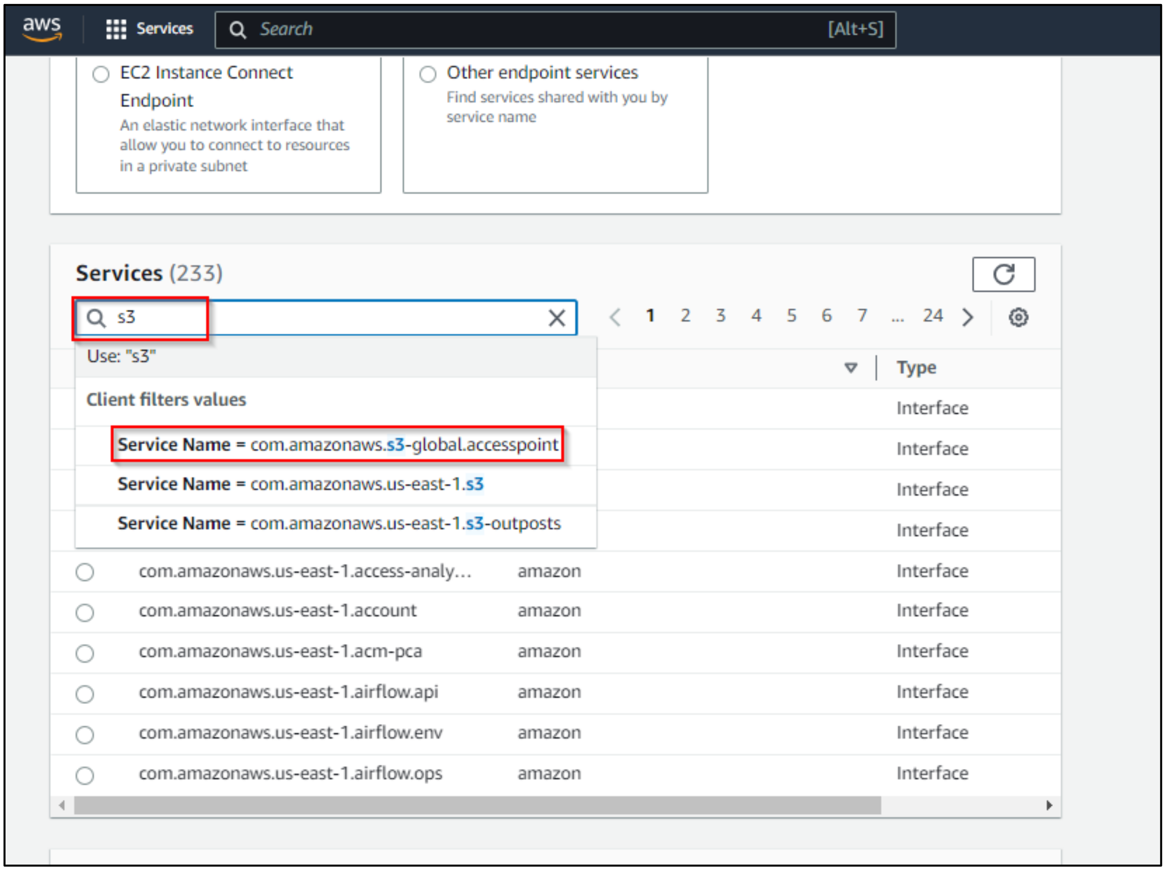
○ PrivateLink Ready partner services
Services with an AWS Service Ready designation

○ AWS Marketplace services
Services that you've purchased through AWS Marketplace

○ EC2 Instance Connect Endpoint
An elastic network interface that allow you to connect to resources in a private subnet

○ Other endpoint services
Find services shared with you by service name

1.4 Under **Services**, enter **s3** and choose **s3-global.accesspoint**



○ EC2 Instance Connect Endpoint
An elastic network interface that allow you to connect to resources in a private subnet

○ Other endpoint services
Find services shared with you by service name

Services (233)

s3

Use: "s3"

Client filters values

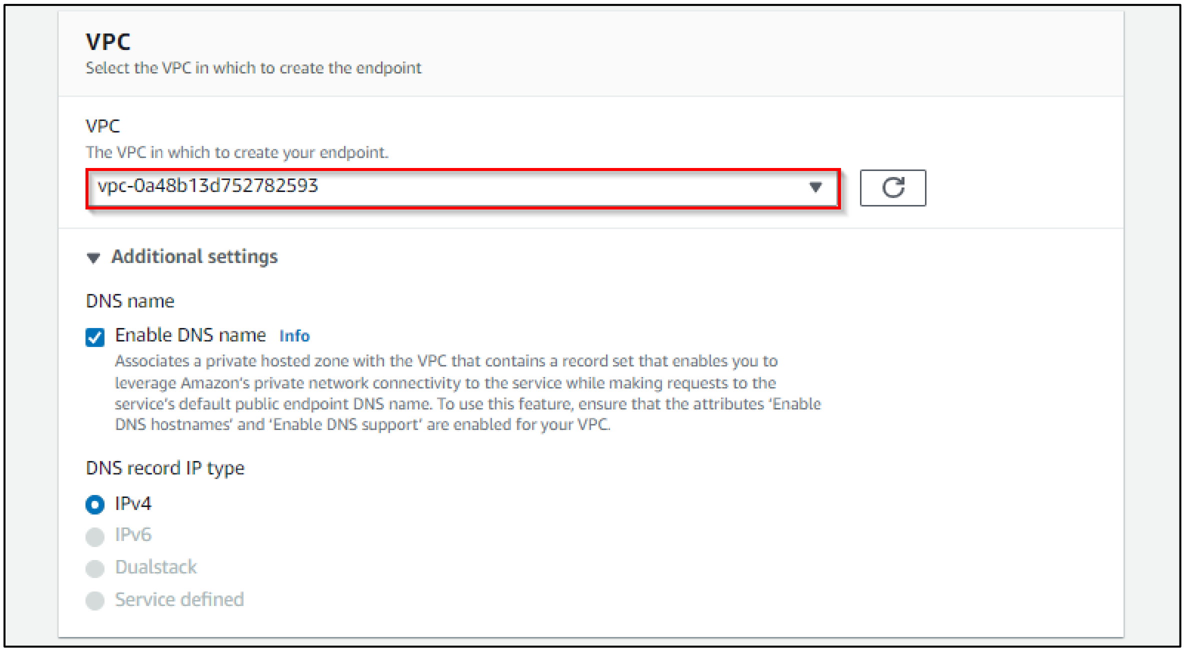
Service Name = com.amazonaws.s3-global.accesspoint

Service Name = com.amazonaws.us-east-1.s3

Service Name = com.amazonaws.us-east-1.s3-outposts

			Type
			Interface
			Interface
			Interface
			Interface
○	com.amazonaws.us-east-1.access-analy…	amazon	Interface
○	com.amazonaws.us-east-1.account	amazon	Interface
○	com.amazonaws.us-east-1.acm-pca	amazon	Interface
○	com.amazonaws.us-east-1.airflow.api	amazon	Interface
○	com.amazonaws.us-east-1.airflow.env	amazon	Interface
○	com.amazonaws.us-east-1.airflow.ops	amazon	Interface

1.5 Choose the default **VPC** option under **VPC**



VPC

Select the VPC in which to create the endpoint

VPC

The VPC in which to create your endpoint.

vpc-0a48b13d752782593

▼ Additional settings

DNS name

☑ Enable DNS name Info

Associates a private hosted zone with the VPC that contains a record set that enables you to leverage Amazon's private network connectivity to the service while making requests to the service's default public endpoint DNS name. To use this feature, ensure that the attributes 'Enable DNS hostnames' and 'Enable DNS support' are enabled for your VPC.

DNS record IP type

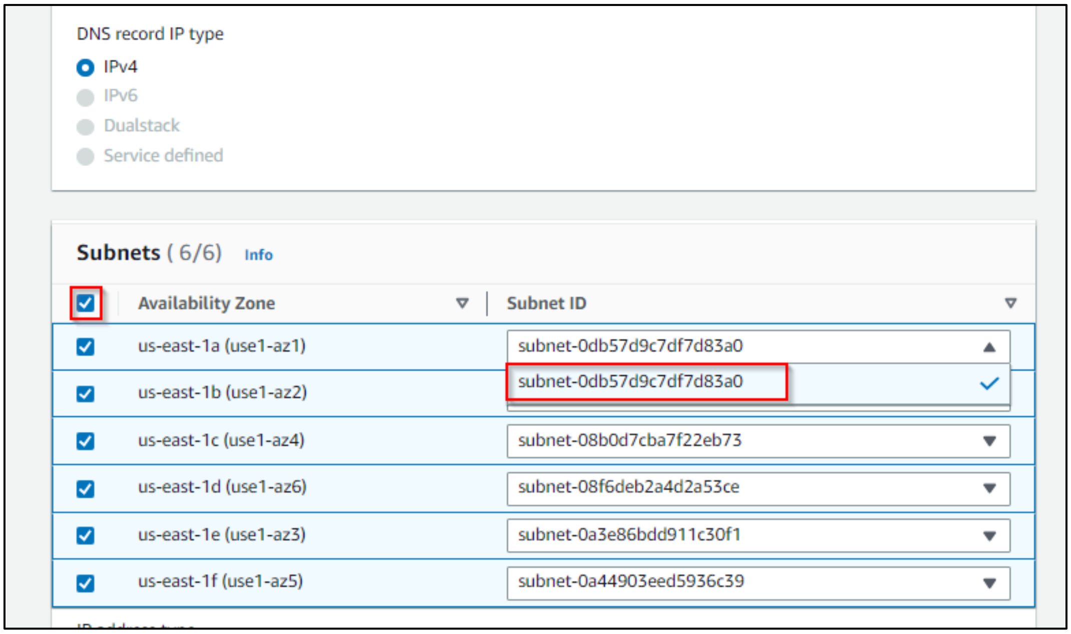
● IPv4

○ IPv6

○ Dualstack

○ Service defined

1.6 In **Subnets**, select the Availability Zone by clicking the checkbox next to it, and then click on the Subnet ID



DNS record IP type

● IPv4

○ IPv6

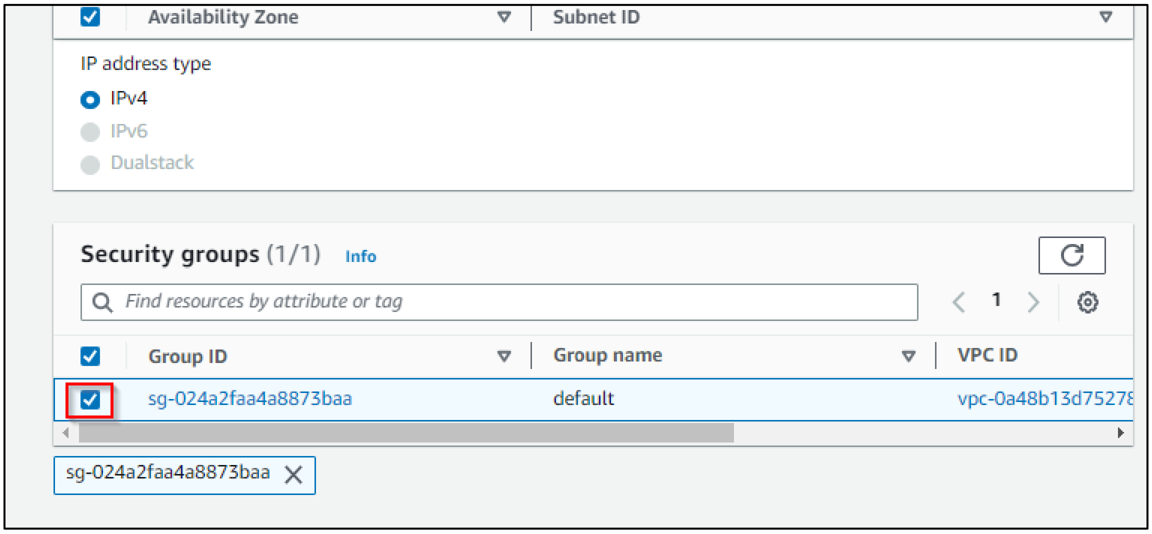
○ Dualstack

○ Service defined

Subnets (6/6) Info

☑	Availability Zone	Subnet ID
☑	us-east-1a (use1-az1)	subnet-0db57d9c7df7d83a0
☑	us-east-1b (use1-az2)	subnet-0db57d9c7df7d83a0 ✓
☑	us-east-1c (use1-az4)	subnet-08b0d7cba7f22eb73
☑	us-east-1d (use1-az6)	subnet-08f6deb2a4d2a53ce
☑	us-east-1e (use1-az3)	subnet-0a3e86bdd911c30f1
☑	us-east-1f (use1-az5)	subnet-0a44903eed5936c39

1.7 Select the appropriate **Group ID** under **Security groups**, and click on **Create endpoint**



☑	Availability Zone	Subnet ID

IP address type

● IPv4

○ IPv6

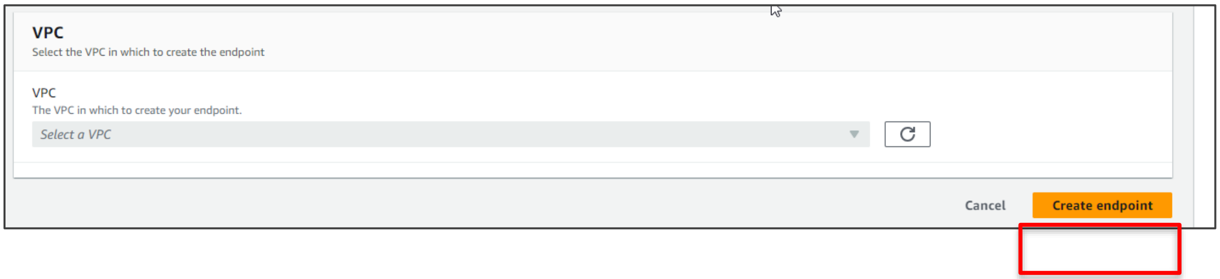
○ Dualstack

Security groups (1/1) Info

Find resources by attribute or tag

☑	Group ID	Group name	VPC ID
☑	sg-024a2faa4a8873baa	default	vpc-0a48b13d75278

sg-024a2faa4a8873baa ✕



VPC

Select the VPC in which to create the endpoint

VPC

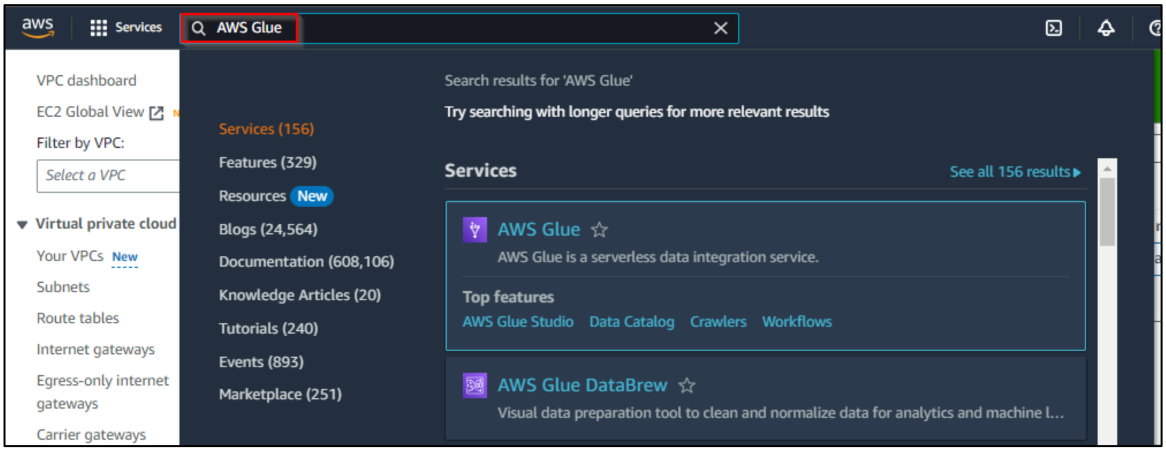
The VPC in which to create your endpoint.

Select a VPC

Cancel Create endpoint

Step 2: Create a Glue Data Catalog

2.1 Navigate to the AWS portal homepage and search for **AWS Glue**



AWS Glue

Search results for 'AWS Glue'

Try searching with longer queries for more relevant results

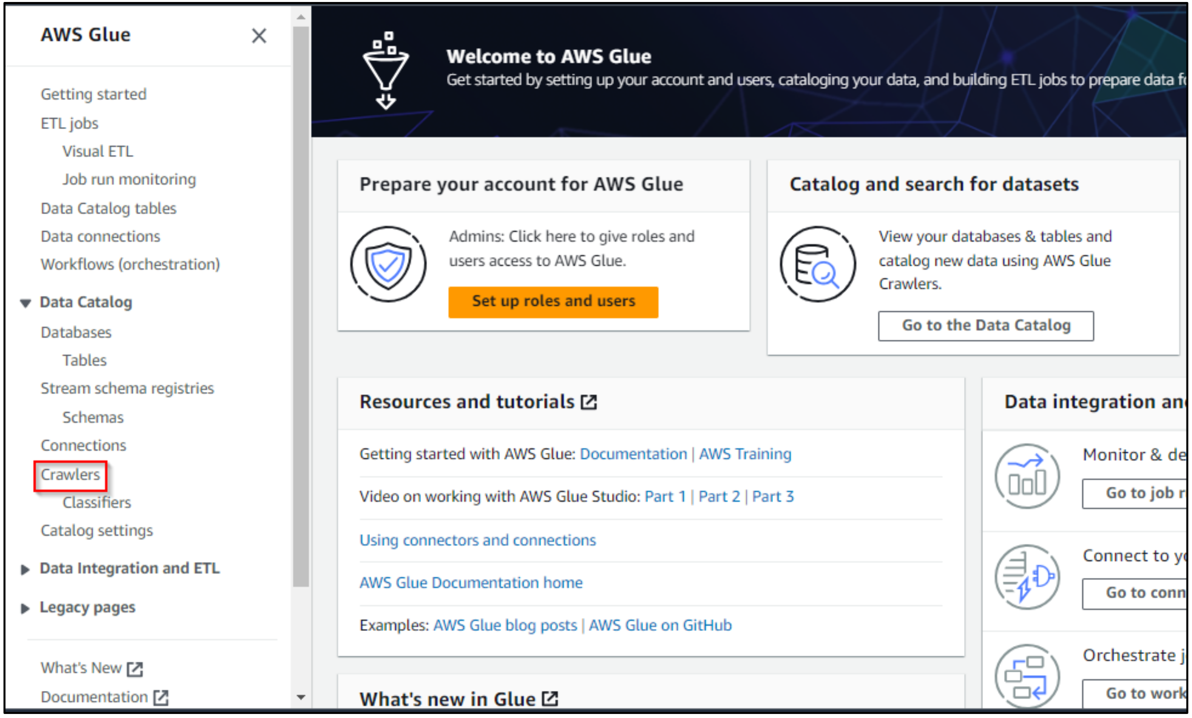
AWS Glue

AWS Glue is a serverless data integration service.

AWS Glue DataBrew

Visual data preparation tool to clean and normalize data for analytics and machine l...

2.2 Click on **Crawlers** under **Data Catalog**



AWS Glue ✕

Getting started
ETL jobs
Visual ETL
Job run monitoring
Data Catalog tables
Data connections
Workflows (orchestration)

▼ Data Catalog
Databases
Tables
Stream schema registries
Schemas
Connections
Crawlers
Classifiers
Catalog settings

▶ Data Integration and ETL

▶ Legacy pages

What's New
Documentation

Welcome to AWS Glue
Get started by setting up your account and users, cataloging your data, and building ETL jobs to prepare data f

Prepare your account for AWS Glue

Admins: Click here to give roles and users access to AWS Glue.

Set up roles and users

Catalog and search for datasets

View your databases & tables and catalog new data using AWS Glue Crawlers.

Go to the Data Catalog

Resources and tutorials

Getting started with AWS Glue: Documentation | AWS Training

Video on working with AWS Glue Studio: Part 1 | Part 2 | Part 3

Using connectors and connections

AWS Glue Documentation home

Examples: AWS Glue blog posts | AWS Glue on GitHub

Data integration an

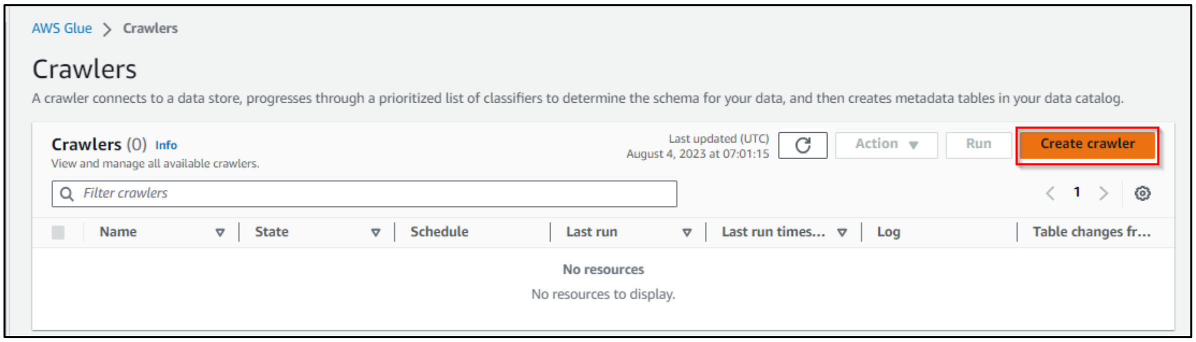
Monitor & de
Go to job r

Connect to y
Go to conn

Orchestrate j
Go to work

What's new in Glue

2.3 Select **Create crawler**



AWS Glue > Crawlers

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
View and manage all available crawlers.

Last updated (UTC)
August 4, 2023 at 07:01:15

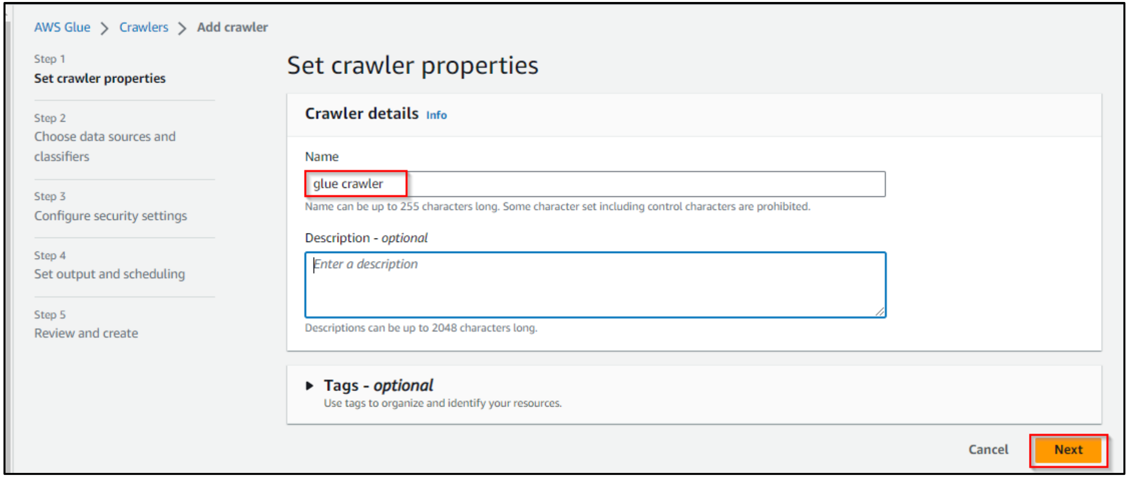
Action ▼ Run **Create crawler**

Filter crawlers

□	Name	State	Schedule	Last run	Last run times...	Log	Table changes fr...

No resources
No resources to display.

2.4 Enter the name as **glue crawler** and click **Next**



AWS Glue > Crawlers > Add crawler

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

Set crawler properties

Crawler details Info

Name

glue crawler

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

Description - *optional*

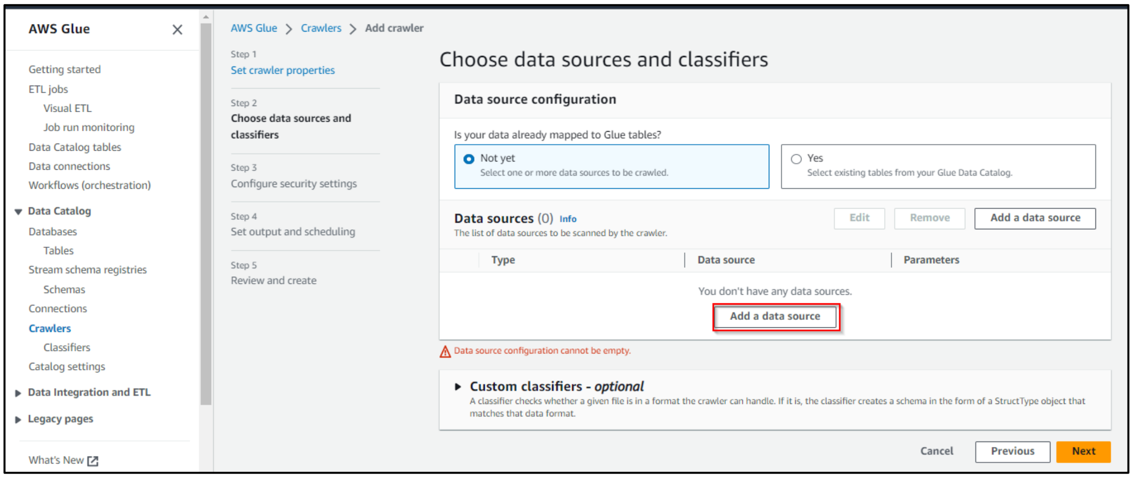
Enter a description

Descriptions can be up to 2048 characters long.

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

Cancel | Next

2.5 Click on **Add a data source**



AWS Glue ✕

Getting started
ETL jobs
Visual ETL
Job run monitoring
Data Catalog tables
Data connections
Workflows (orchestration)
▼ Data Catalog
Databases
Tables
Stream schema registries
Schemas
Connections
Crawlers
Classifiers
Catalog settings
▶ Data Integration and ETL
▶ Legacy pages
What's New

AWS Glue > Crawlers > Add crawler

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

Choose data sources and classifiers

Data source configuration

Is your data already mapped to Glue tables?

● Not yet
Select one or more data sources to be crawled.

○ Yes
Select existing tables from your Glue Data Catalog.

Data sources (0) Info
The list of data sources to be scanned by the crawler.

Edit | Remove | Add a data source

Type	Data source	Parameters

You don't have any data sources.

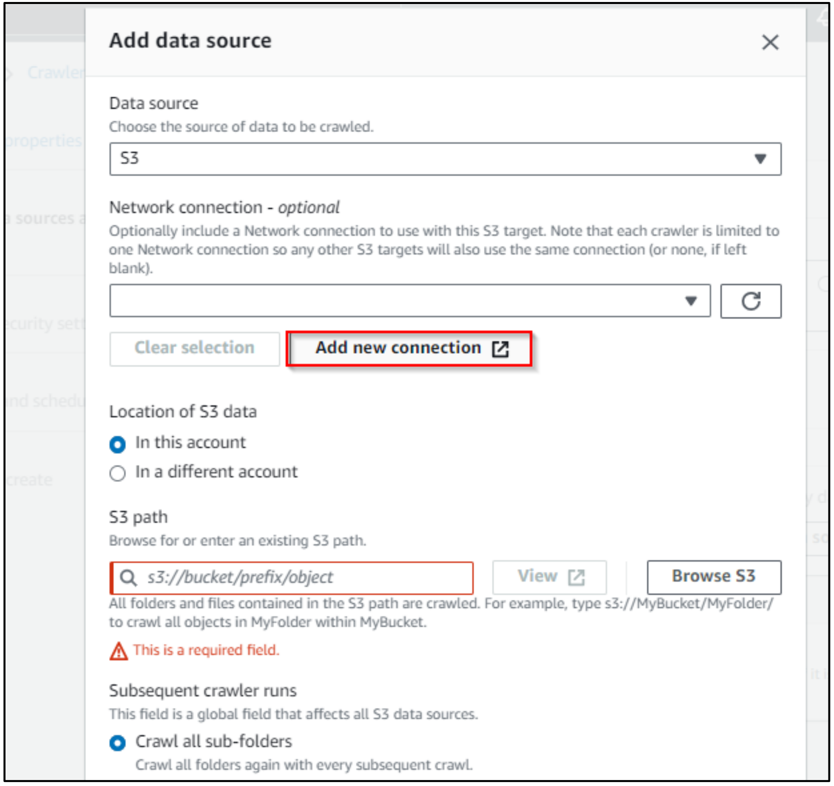
Add a data source

⚠ Data source configuration cannot be empty.

▶ **Custom classifiers - *optional***
A classifier checks whether a given file is in a format the crawler can handle. If it is, the classifier creates a schema in the form of a StructType object that matches that data format.

Cancel | Previous | Next

2.6 Click on **Add new connection**



Add data source

Data source

Choose the source of data to be crawled.

S3

Network connection - *optional*

Optionally include a Network connection to use with this S3 target. Note that each crawler is limited to one Network connection so any other S3 targets will also use the same connection (or none, if left blank).

Location of S3 data

● In this account

○ In a different account

S3 path

Browse for or enter an existing S3 path.

s3://bucket/prefix/object

All folders and files contained in the S3 path are crawled. For example, type s3://MyBucket/MyFolder/ to crawl all objects in MyFolder within MyBucket.

This is a required field.

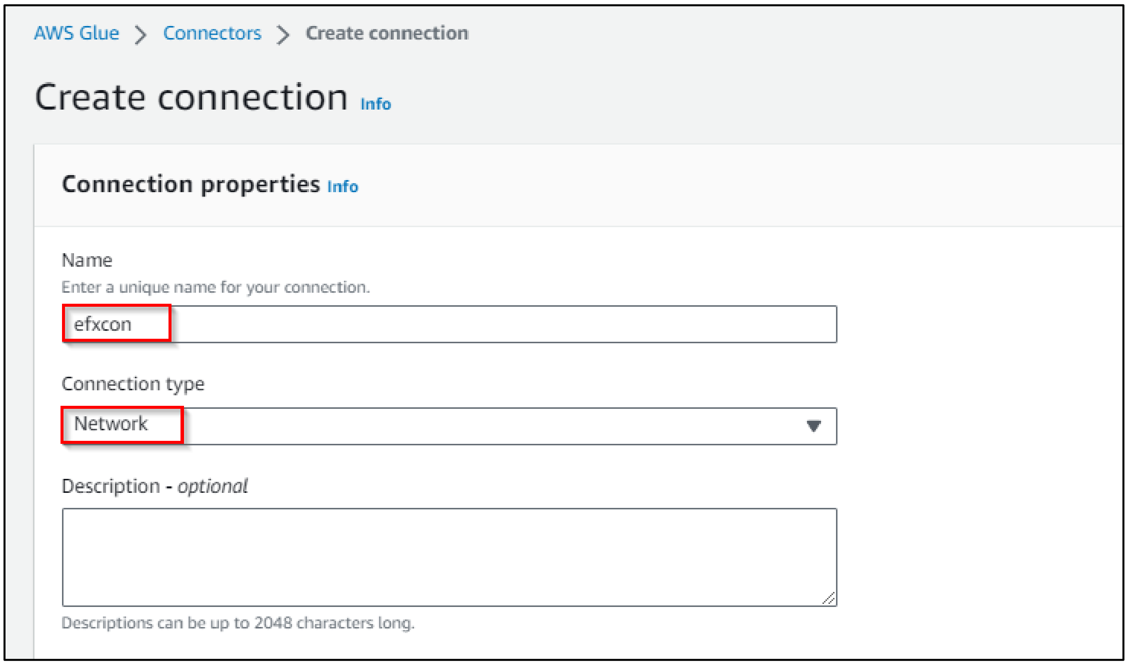
Subsequent crawler runs

This field is a global field that affects all S3 data sources.

● Crawl all sub-folders

Crawl all folders again with every subsequent crawl.

2.7 Enter the name as **efxcon** and select **Network** as the Connection type



Create connection Info

Connection properties Info

Name

Enter a unique name for your connection.

efxcon

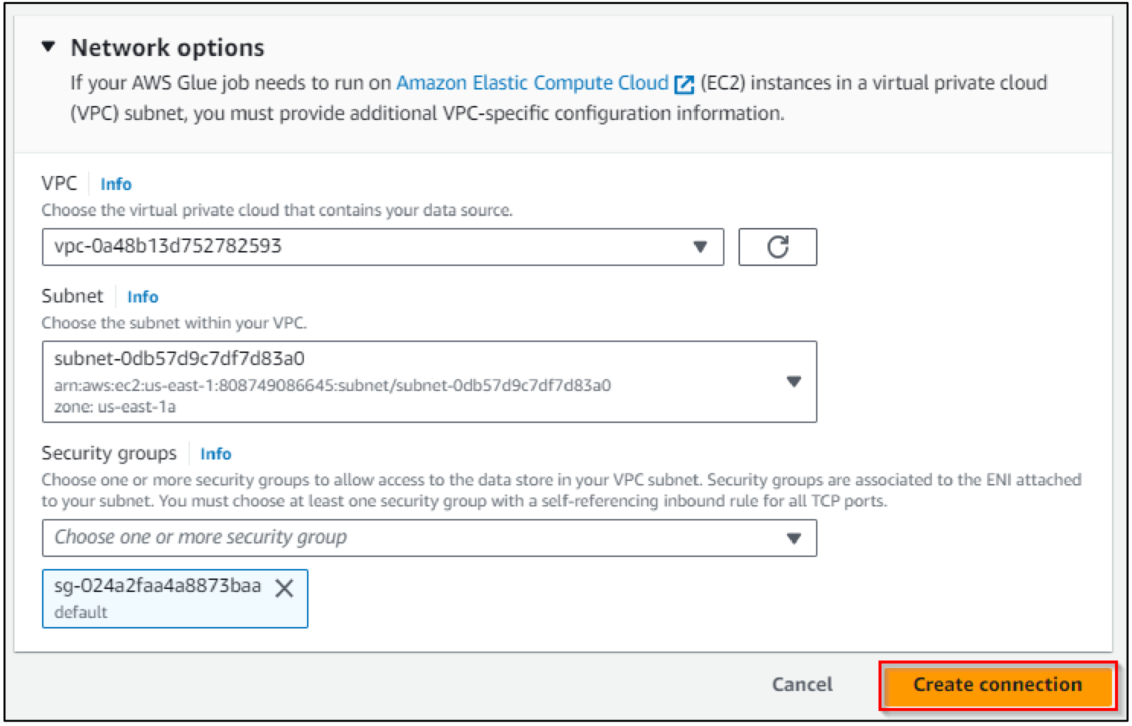
Connection type

Network

Description - *optional*

Descriptions can be up to 2048 characters long.

2.8 Configure the Network options as shown, and then click **Create connection**



Network options

If your AWS Glue job needs to run on Amazon Elastic Compute Cloud (EC2) instances in a virtual private cloud (VPC) subnet, you must provide additional VPC-specific configuration information.

VPC | Info

Choose the virtual private cloud that contains your data source.

vpc-0a48b13d752782593

Subnet | Info

Choose the subnet within your VPC.

subnet-0db57d9c7df7d83a0
arn:aws:ec2:us-east-1:808749086645:subnet/subnet-0db57d9c7df7d83a0
zone: us-east-1a

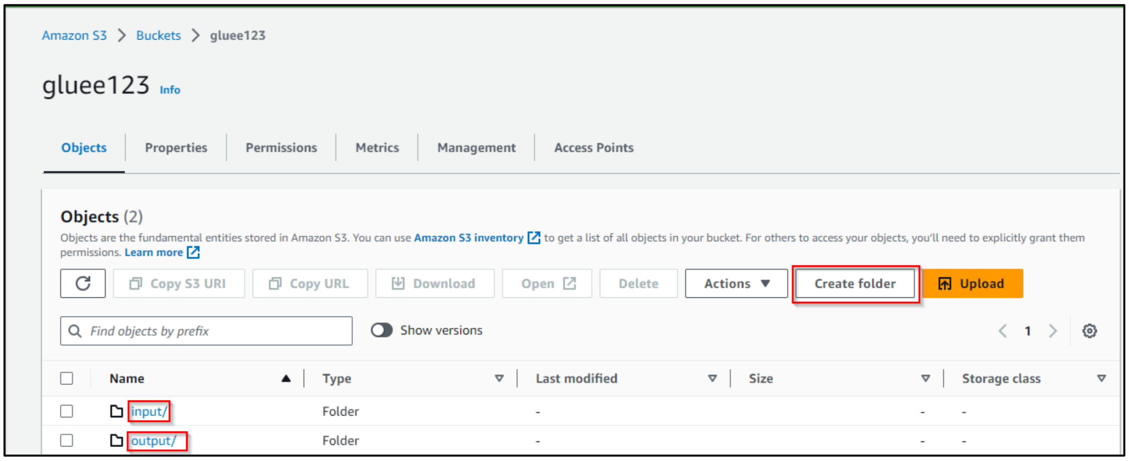
Security groups | Info

Choose one or more security groups to allow access to the data store in your VPC subnet. Security groups are associated to the ENI attached to your subnet. You must choose at least one security group with a self-referencing inbound rule for all TCP ports.

Choose one or more security group

sg-024a2faa4a8873baa
default

2.9 Create an S3 bucket named **gluee123** and **input** and **output** folders within it



Amazon S3 > Buckets > gluee123

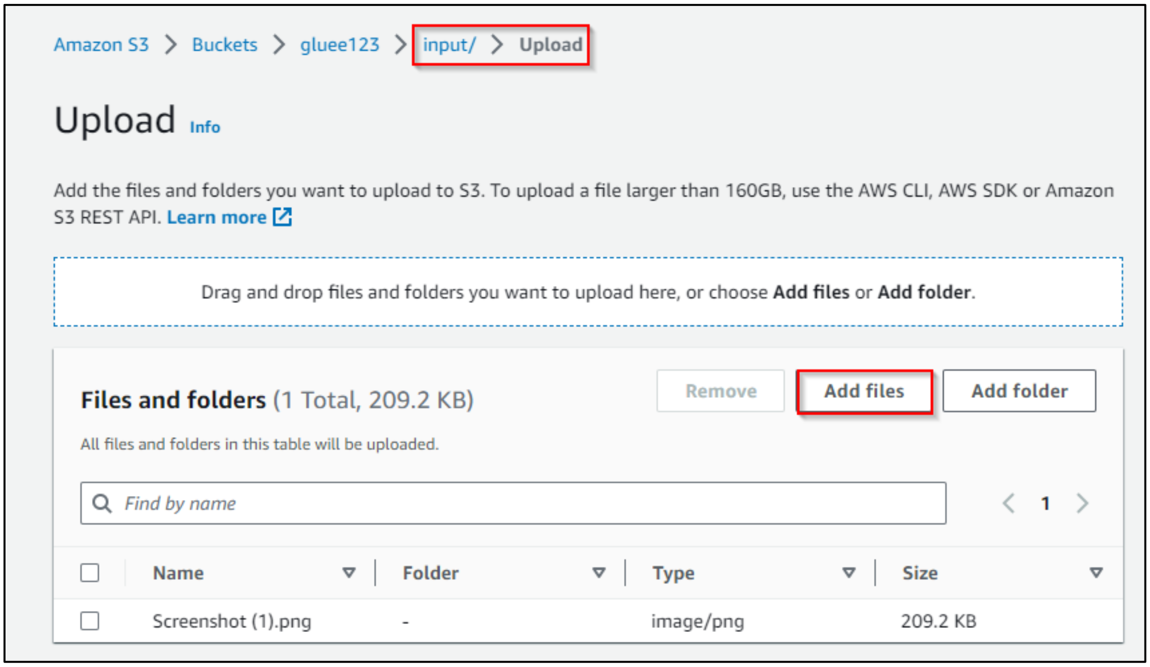
gluee123 Info

Objects (2)

Objects are the fundamental entities stored in Amazon S3. You can use **Amazon S3 inventory** to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. **Learn more**

□	Name	Type	Last modified	Size	Storage class
□	input/	Folder	-	-	-
□	output/	Folder	-	-	-

2.10 Navigate to the input folder, and either drag and drop or click **Add files** to upload an image



Amazon S3 > Buckets > gluee123 > input/ > Upload

Upload Info

Add the files and folders you want to upload to S3. To upload a file larger than 160GB, use the AWS CLI, AWS SDK or Amazon S3 REST API. Learn more

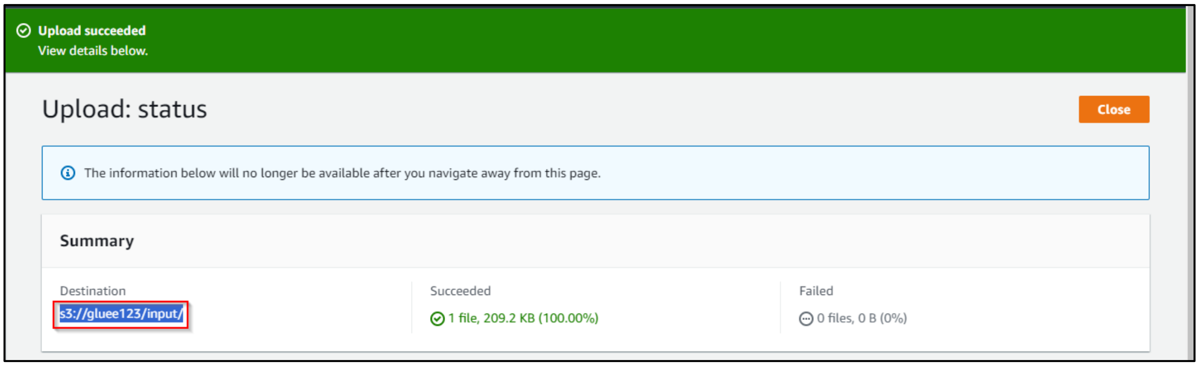
Drag and drop files and folders you want to upload here, or choose **Add files** or **Add folder**.

Files and folders (1 Total, 209.2 KB)

All files and folders in this table will be uploaded.

□	Name	Folder	Type	Size
□	Screenshot (1).png	-	image/png	209.2 KB

2.11 Copy the Destination path for use in the S3 path



Upload succeeded
View details below.

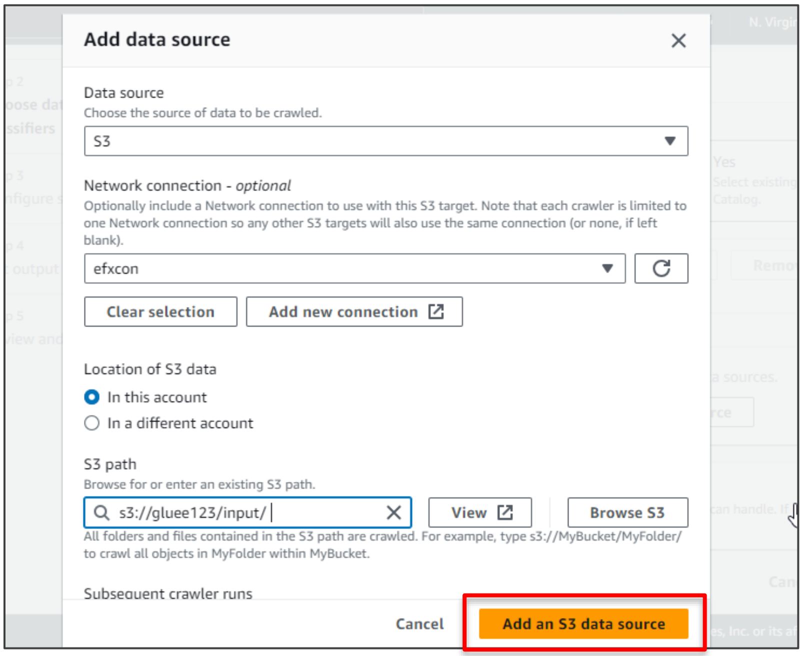
Upload: status

The information below will no longer be available after you navigate away from this page.

Summary

Destination	Succeeded	Failed
s3://gluee123/input/	1 file, 209.2 KB (100.00%)	0 files, 0 B (0%)

2.12 Paste the S3 path location as shown, and then click on **Add an S3 data source** in the **Add data source** tab



Add data source

Data source
Choose the source of data to be crawled.
S3

Network connection - *optional*
Optionally include a Network connection to use with this S3 target. Note that each crawler is limited to one Network connection so any other S3 targets will also use the same connection (or none, if left blank).
efxcon

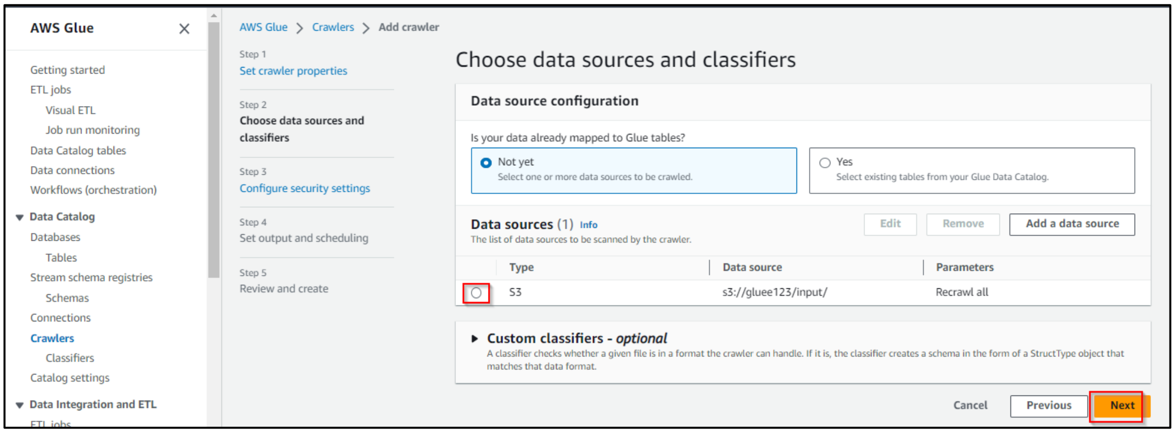
Location of S3 data
● In this account
○ In a different account

S3 path
Browse for or enter an existing S3 path.
s3://gluee123/input/
All folders and files contained in the S3 path are crawled. For example, type s3://MyBucket/MyFolder/ to crawl all objects in MyFolder within MyBucket.

Subsequent crawler runs

Step 3: Add a data source to Crawlers

3.1 Select **S3** type and click **Next**



Choose data sources and classifiers

Data source configuration

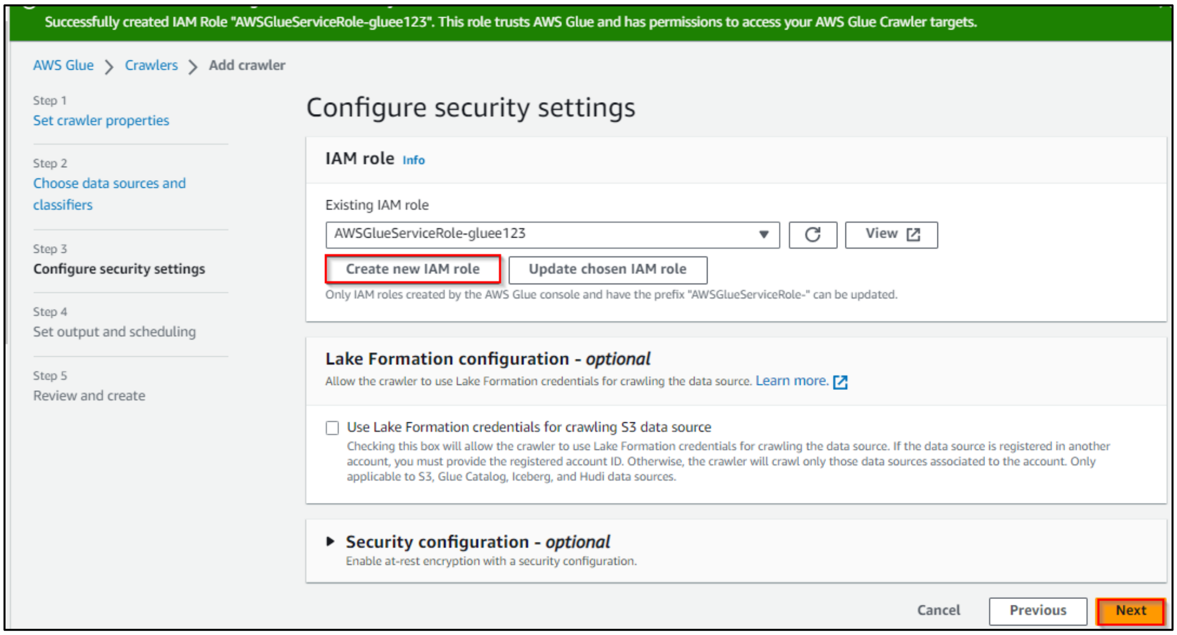
Is your data already mapped to Glue tables?
● Not yet
Select one or more data sources to be crawled.
○ Yes
Select existing tables from your Glue Data Catalog.

Data sources (1) Info
The list of data sources to be scanned by the crawler.

Type	Data source	Parameters
S3	s3://gluee123/input/	Recrawl all

Custom classifiers - *optional*
A classifier checks whether a given file is in a format the crawler can handle. If it is, the classifier creates a schema in the form of a StructType object that matches that data format.

3.2 Create a new IAM role named **gluee123**, and click **Next**



Successfully created IAM Role "AWSGlueServiceRole-gluee123". This role trusts AWS Glue and has permissions to access your AWS Glue Crawler targets.

AWS Glue > Crawlers > Add crawler

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

Configure security settings

IAM role Info

Existing IAM role

AWSGlueServiceRole-gluee123

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

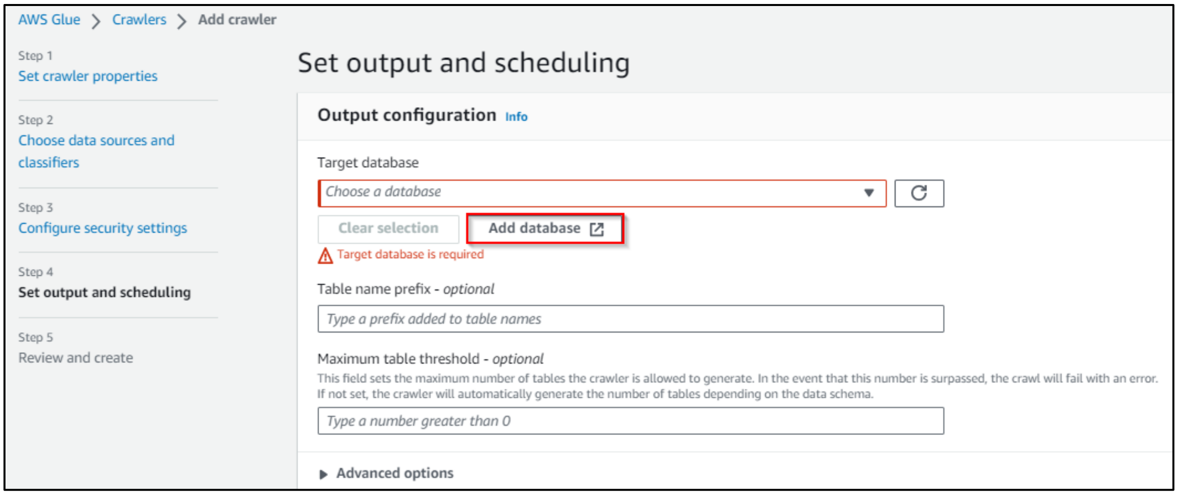
Checking this box will allow the crawler to use Lake Formation credentials for crawling the data source. If the data source is registered in another account, you must provide the registered account ID. Otherwise, the crawler will crawl only those data sources associated to the account. Only applicable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

▶ Security configuration - *optional*

Enable at-rest encryption with a security configuration.

Cancel | Previous | Next

3.3 Click on **Add database**



AWS Glue > Crawlers > Add crawler

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

Set output and scheduling

Output configuration Info

Target database

Choose a database

Clear selection | Add database

⚠ Target database is required

Table name prefix - *optional*

Type a prefix added to table names

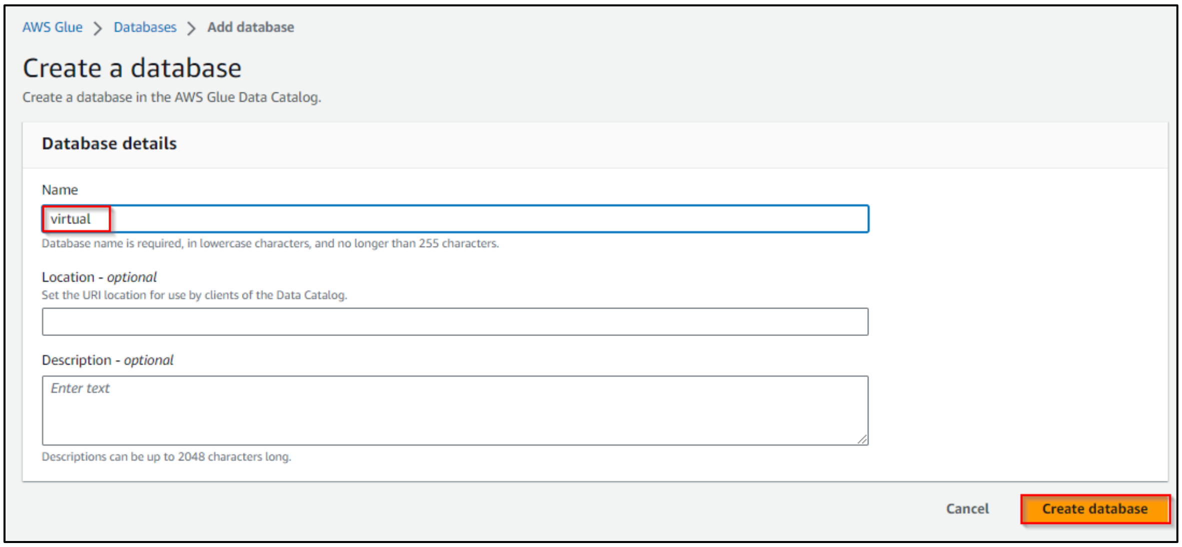
Maximum table threshold - *optional*

This field sets the maximum number of tables the crawler is allowed to generate. In the event that this number is surpassed, the crawl will fail with an error. If not set, the crawler will automatically generate the number of tables depending on the data schema.

Type a number greater than 0

▶ Advanced options

3.4 Name it **virtual**, and create the database



AWS Glue > Databases > Add database

Create a database

Create a database in the AWS Glue Data Catalog.

Database details

Name: virtual

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

Location - *optional*

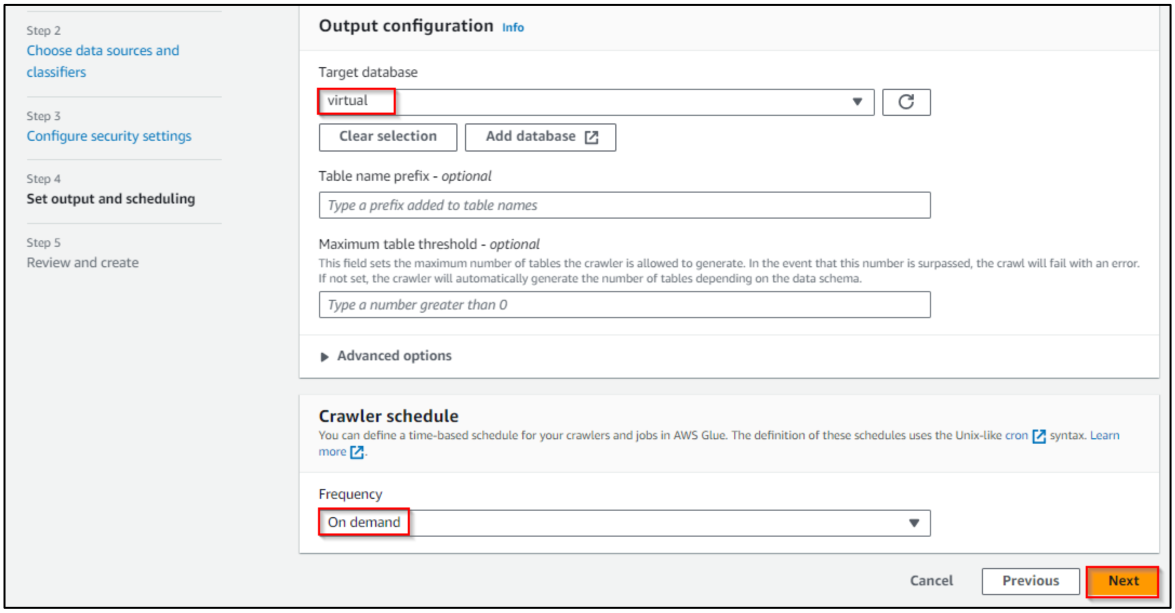
Set the URI location for use by clients of the Data Catalog.

Description - *optional*

Enter text

Descriptions can be up to 2048 characters long.

3.5 Select **virtual** from the database options, set the Frequency to **On demand**, and then click on **Next**



Step 2
Choose data sources and classifiers

Step 3
Configure security settings

Step 4
Set output and scheduling

Step 5
Review and create

Output configuration Info

Target database: virtual

Table name prefix - *optional*

Type a prefix added to table names

Maximum table threshold - *optional*

This field sets the maximum number of tables the crawler is allowed to generate. In the event that this number is surpassed, the crawl will fail with an error. If not set, the crawler will automatically generate the number of tables depending on the data schema.

Type a number greater than 0

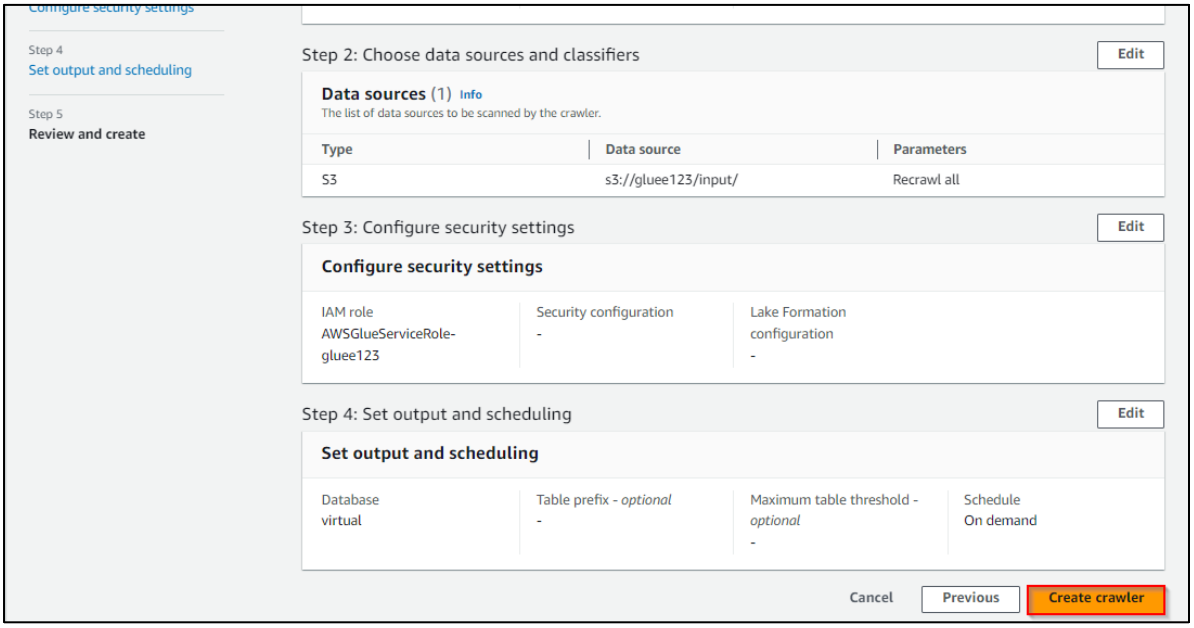
▶ Advanced options

Crawler schedule

You can define a time-based schedule for your crawlers and jobs in AWS Glue. The definition of these schedules uses the Unix-like cron syntax. Learn more.

Frequency: On demand

3.6 Review the settings and create the crawler by clicking **Create crawler**



Configure security settings

Step 4
Set output and scheduling

Step 5
Review and create

Step 2: Choose data sources and classifiers

Data sources (1) Info

The list of data sources to be scanned by the crawler.

Type	Data source	Parameters
S3	s3://gluee123/input/	Recrawl all

Step 3: Configure security settings

Configure security settings

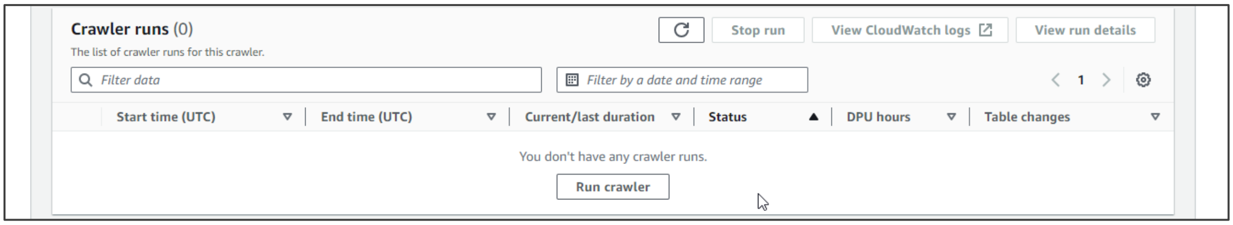
IAM role	Security configuration	Lake Formation configuration
AWSGlueServiceRole-gluee123	-	-

Step 4: Set output and scheduling

Set output and scheduling

Database	Table prefix - *optional*	Maximum table threshold - *optional*	Schedule
virtual	-	-	On demand

3.7 Click **Run crawler**

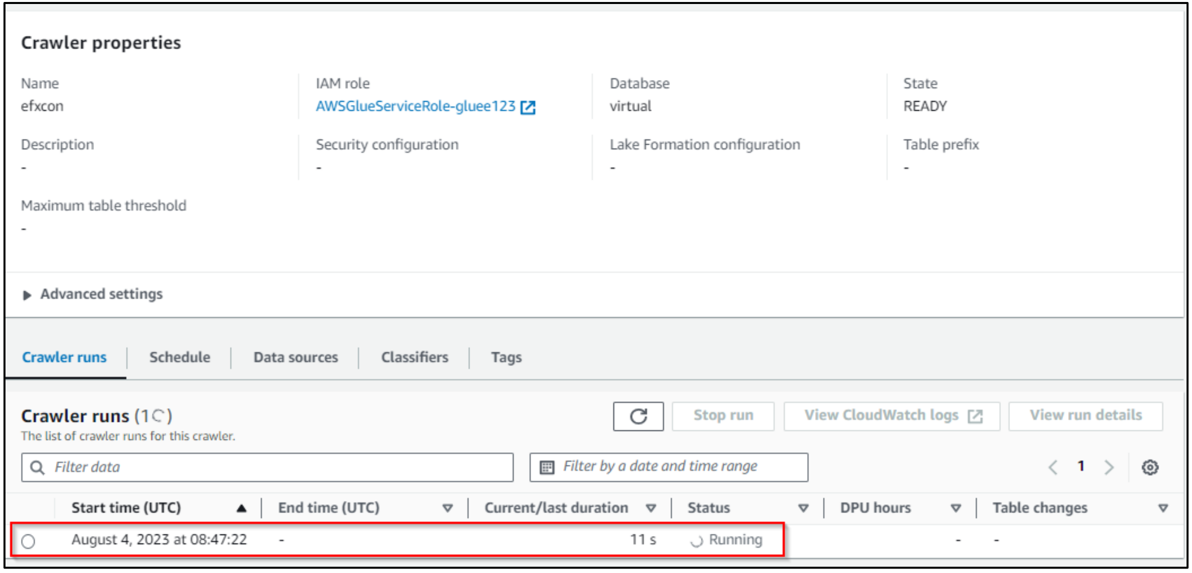


Crawler runs (0)

The list of crawler runs for this crawler.

Start time (UTC)	End time (UTC)	Current/last duration	Status	DPU hours	Table changes

You don't have any crawler runs.



Crawler properties

Name	IAM role	Database	State
efxcon	AWSGlueServiceRole-gluee123	virtual	READY

Description	Security configuration	Lake Formation configuration	Table prefix
-	-	-	-

Maximum table threshold
-

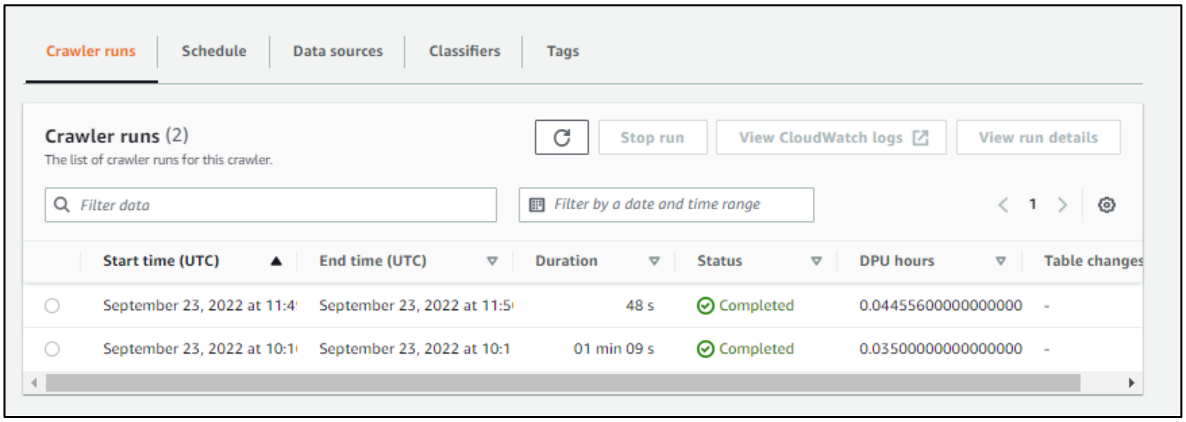
▶ Advanced settings

Crawler runs | Schedule | Data sources | Classifiers | Tags

Crawler runs (1)
The list of crawler runs for this crawler.

Stop run
View CloudWatch logs
View run details
Filter data
Filter by a date and time range
1

	Start time (UTC)	End time (UTC)	Current/last duration	Status	DPU hours	Table changes
○	August 4, 2023 at 08:47:22	-	11 s	Running	-	-



Crawler runs | Schedule | Data sources | Classifiers | Tags

Crawler runs (2)
The list of crawler runs for this crawler.

Stop run
View CloudWatch logs
View run details
Filter data
Filter by a date and time range
1

	Start time (UTC)	End time (UTC)	Duration	Status	DPU hours	Table changes
○	September 23, 2022 at 11:4	September 23, 2022 at 11:5	48 s	Completed	0.04455600000000000	-
○	September 23, 2022 at 10:1	September 23, 2022 at 10:1	01 min 09 s	Completed	0.03500000000000000	-

By following these steps, you have successfully set up a Glue Data Catalog, enhancing data management proficiency within your AWS environment.