

## 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

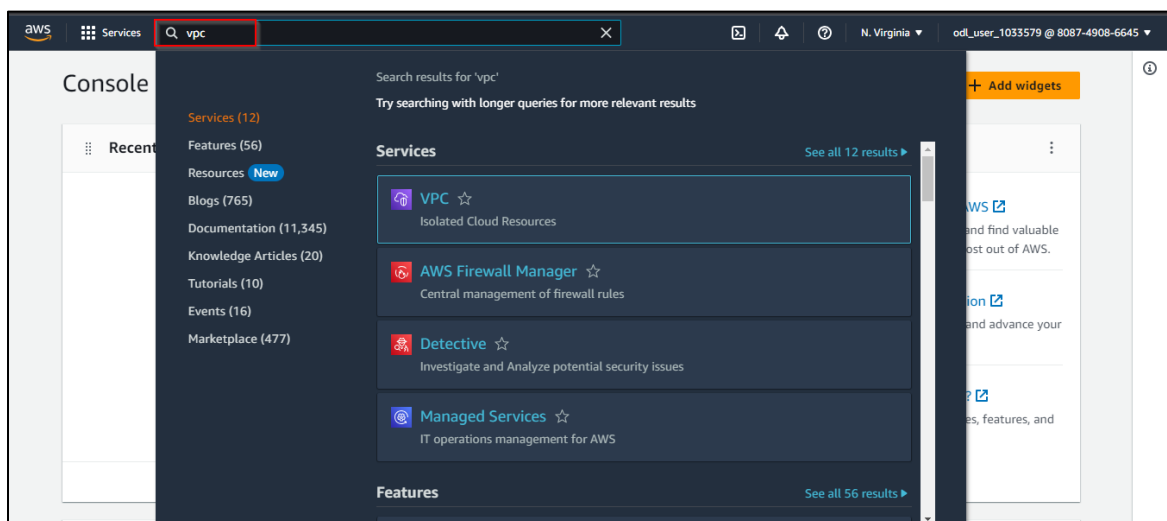
**Prerequisites:** AWS account

Steps to be followed:

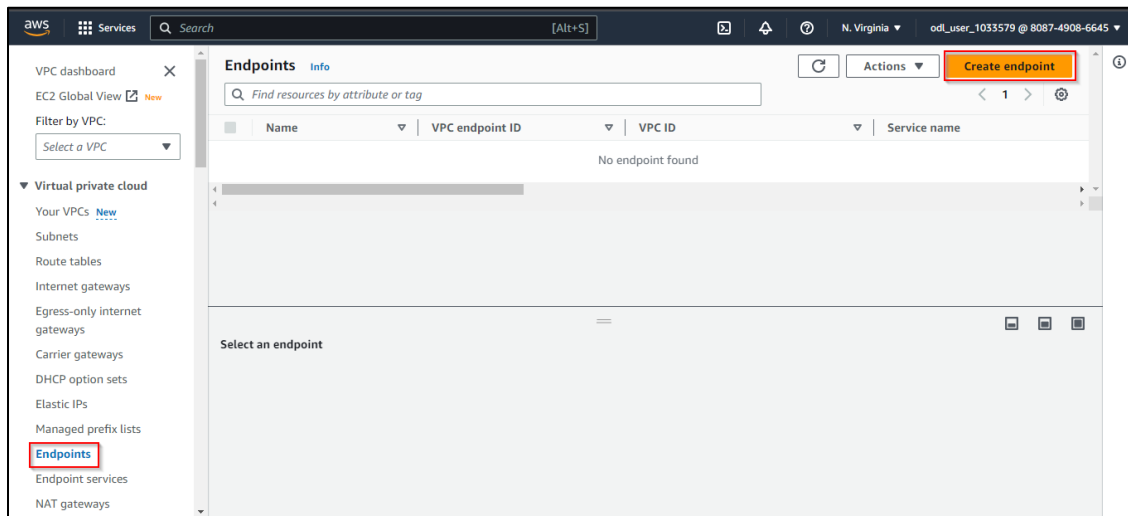
1. Create a VPC endpoint
2. Create a Glue Data Catalog
3. Add a data source to crawlers

#### Step 1: Create a VPC endpoint

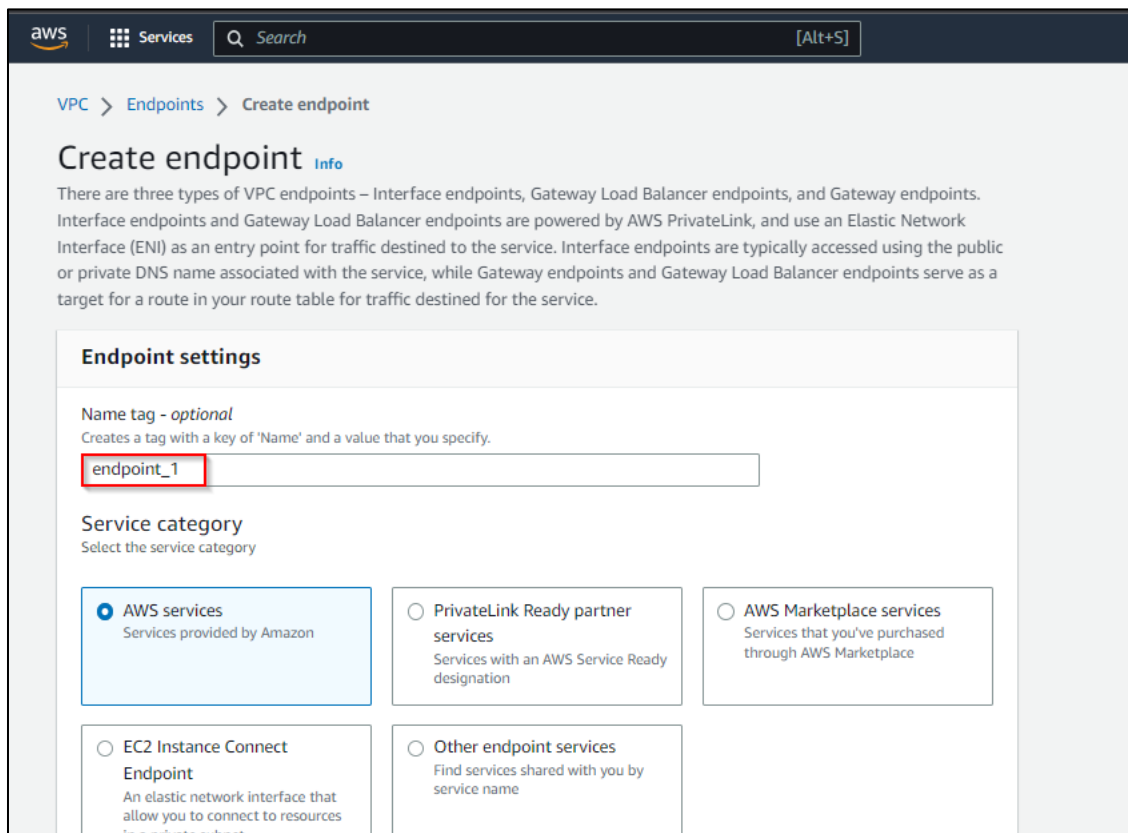
1.1 Navigate to the AWS portal homepage, search for **VPC**, and click on it



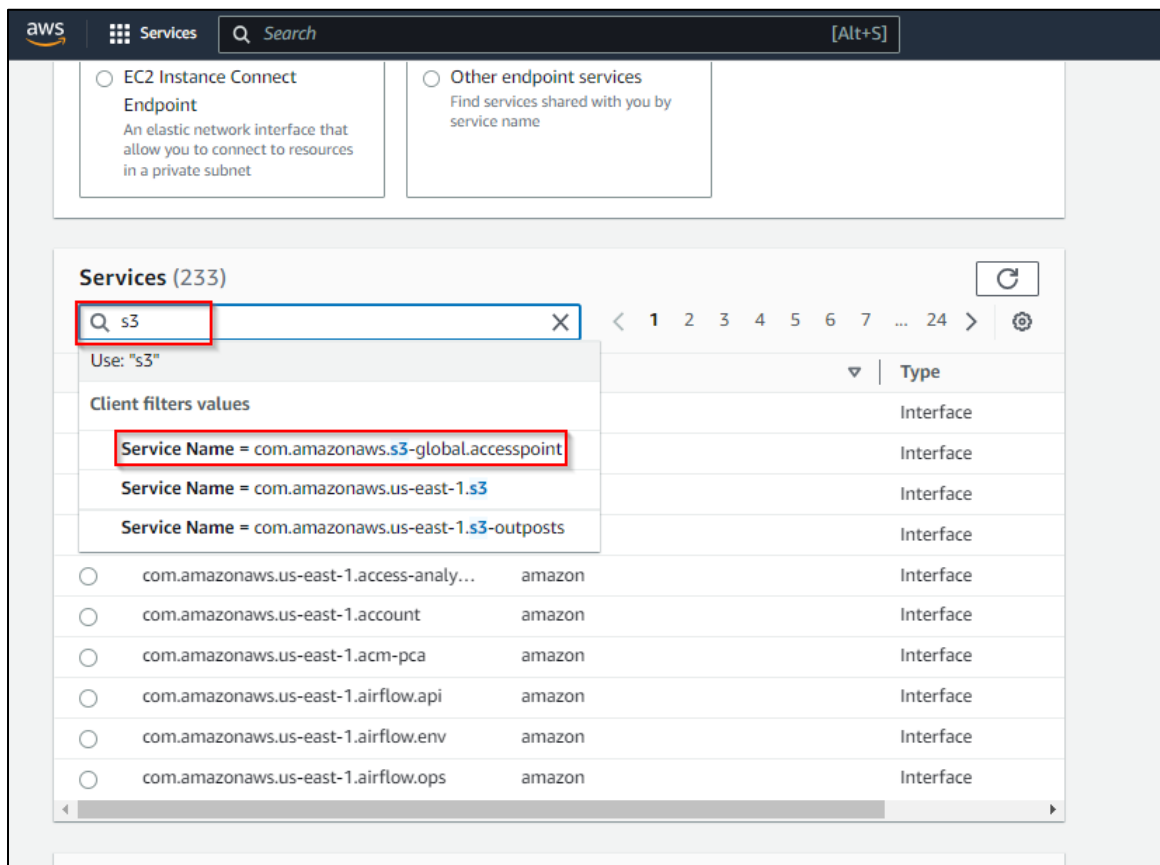
## 1.2 Click on **Endpoints** and then select **Create endpoint**



## 1.3 Enter **endpoint\_1** as the Name tag



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



## 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

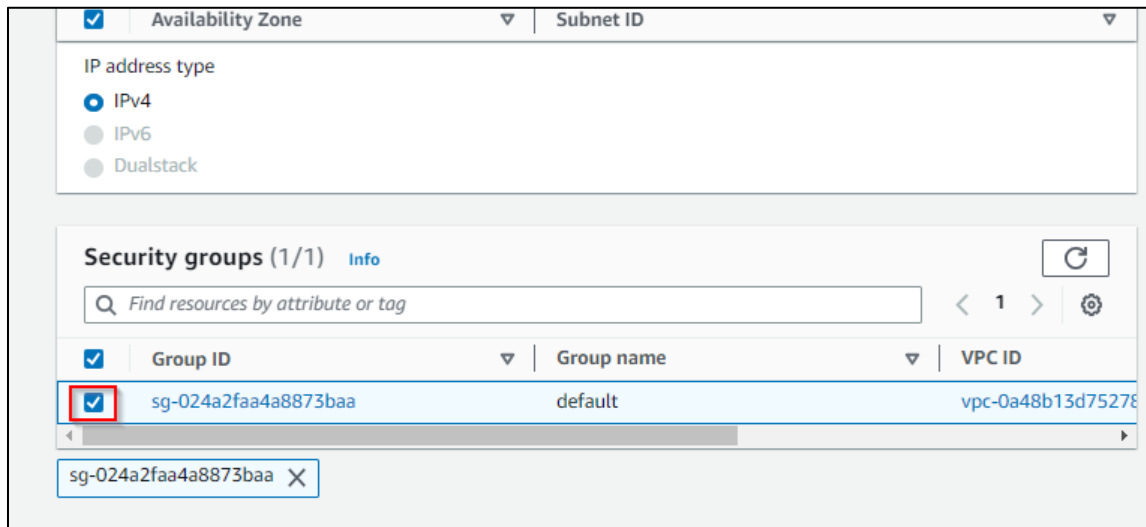
☐ Dualstack

☐ Service defined

### Subnets ( 6/6 ) [Info](#)

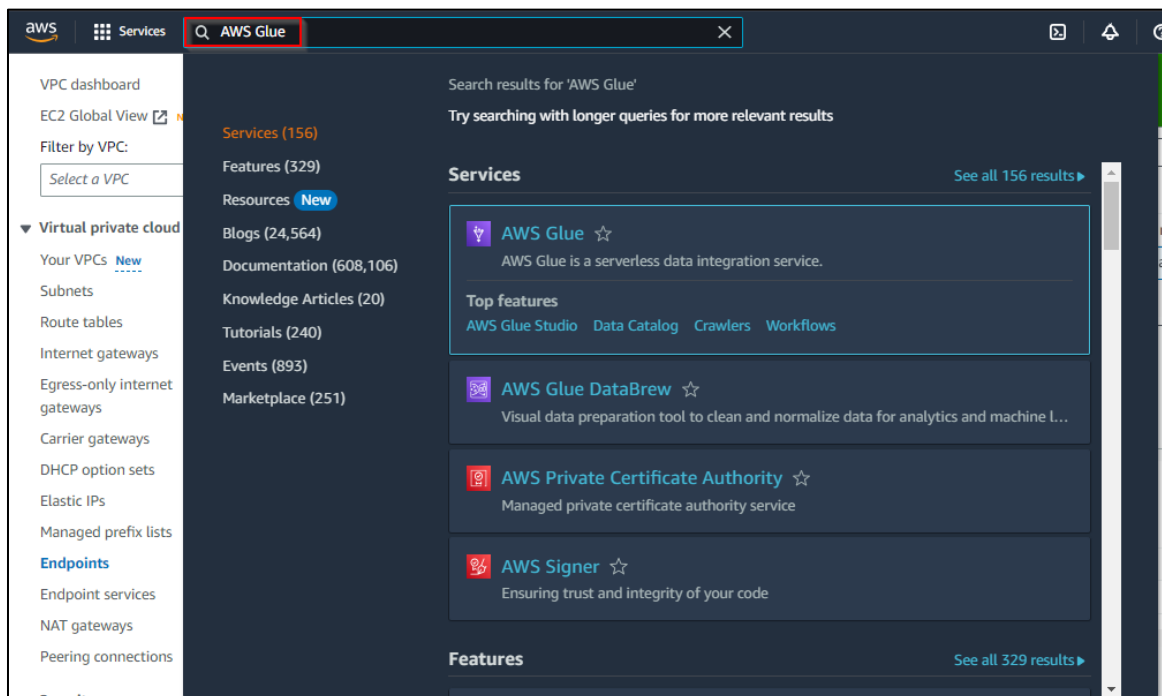
<input checked="" type="checkbox"/>	Availability Zone ▼	Subnet ID ▼
<input checked="" type="checkbox"/>	us-east-1a (use1-az1)	subnet-0db57d9c7df7d83a0 ▲
<input checked="" type="checkbox"/>	us-east-1b (use1-az2)	subnet-0db57d9c7df7d83a0 ✓
<input checked="" type="checkbox"/>	us-east-1c (use1-az4)	subnet-08b0d7cba7f22eb73 ▼
<input checked="" type="checkbox"/>	us-east-1d (use1-az6)	subnet-08f6deb2a4d2a53ce ▼
<input checked="" type="checkbox"/>	us-east-1e (use1-az3)	subnet-0a3e86bdd911c30f1 ▼
<input checked="" type="checkbox"/>	us-east-1f (use1-az5)	subnet-0a44903eed5936c39 ▼

## 1.7 Select the appropriate **Group ID** under **Security groups**, and create the endpoint

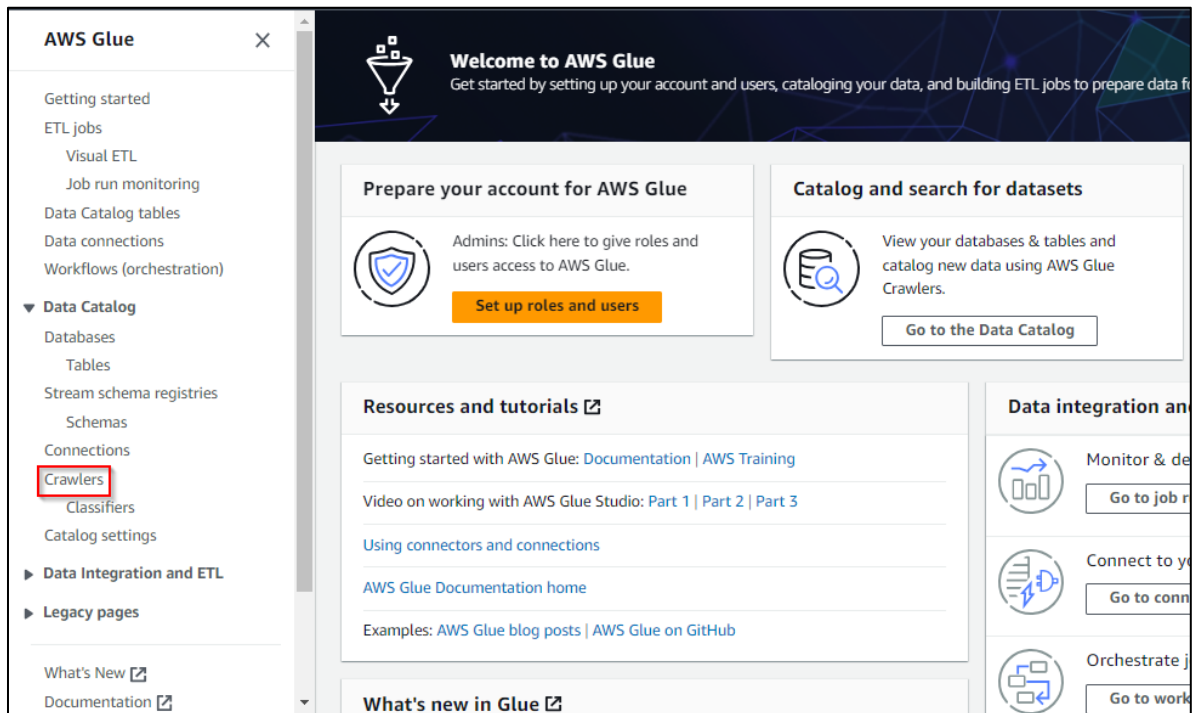


## Step 2: Create a Glue Data Catalog

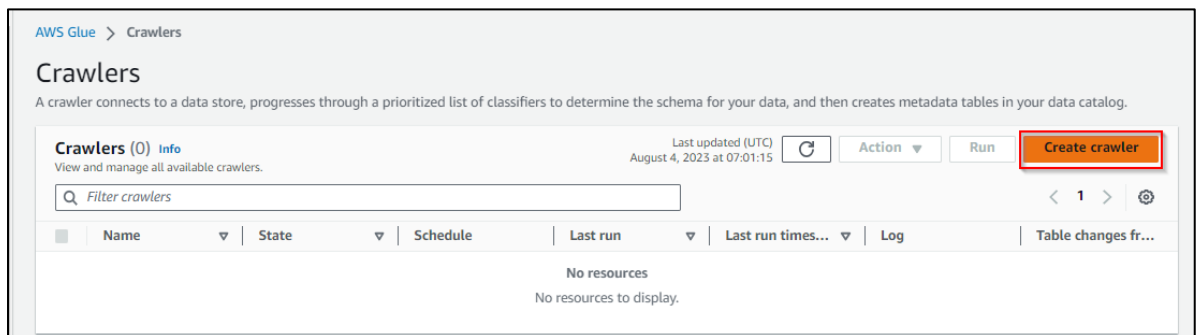
### 2.1 Navigate to the AWS portal homepage and search for **AWS Glue**



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



## 2.3 Select **Create crawler**



## 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 > 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](#) [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** [X]

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).  
[Dropdown] [Refresh]

[Clear selection] **Add new connection** [External Link Icon]

Location of S3 data  
☒ In this account  
☐ In a different account

S3 path  
Browse for or enter an existing S3 path.  
 [View [External Link Icon]] [Browse S3]  
 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 desired name and select the connection type as **Network**

AWS Glue > Connectors > Create connection

**Create connection** [Info]

**Connection properties** [Info]

Name  
Enter a unique name for your connection.

Connection type  
 ▼

Description - *optional*

Descriptions can be up to 2048 characters long.



## 2.8 Configure the **Network options** as shown in the screenshot, and then click on and **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

Cancel

Create connection

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

Amazon S3 > Buckets > gluee123

gluee123 [Info](#)

Objects

Properties

Permissions

Metrics

Management

Access Points

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](#)

↻

Copy S3 URI

Copy URL

Download

Open

Delete

Actions ▼

Create folder

Upload

Find objects by prefix

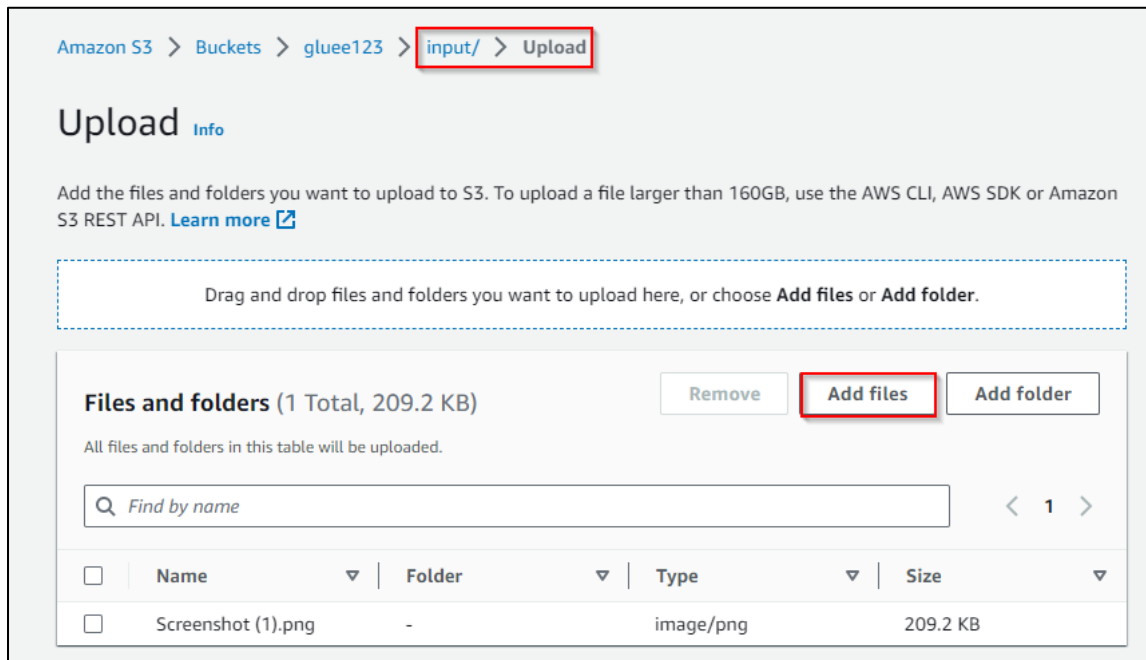
Show versions

< 1 >

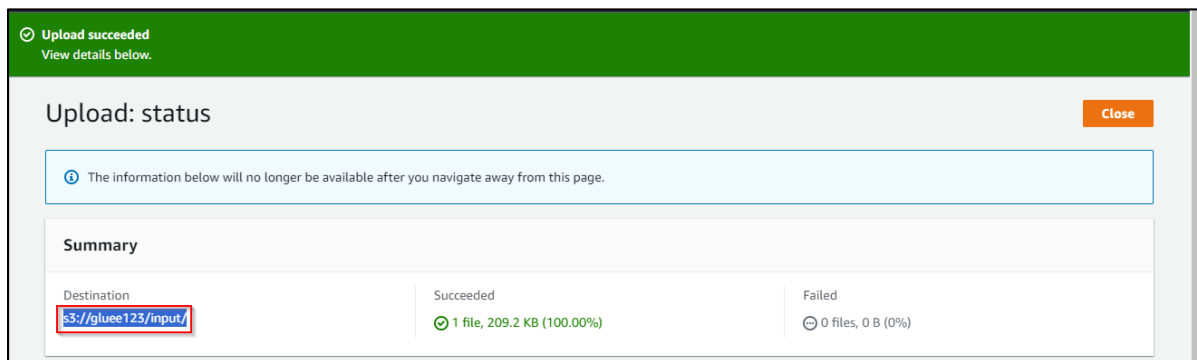
⚙

<input type="checkbox"/>	Name	Type	Last modified	Size	Storage class
<input type="checkbox"/>	input/	Folder	-	-	-
<input type="checkbox"/>	output/	Folder	-	-	-

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



2.11 Copy the Destination path for use in the S3 path



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

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

Clear selection

Add new connection

**Location of S3 data**

☒ In this account  
☐ In a different account

**S3 path**

Browse for or enter an existing S3 path.

View

Browse S3

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**

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

☒ Crawl all sub-folders  
 Crawl all folders again with every subsequent crawl.

☐ Crawl new sub-folders only  
 Only Amazon S3 folders that were added since the last crawl will be crawled. If the schemas are

Cancel

Add an S3 data source

## Step 3: Add data source to Crawlers

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

**AWS Glue** × **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 (1)** [Info](#) Edit Remove Add a data source

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

Type	Data source	Parameters
<input checked="" type="radio"/> <b>S3</b>	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.

Cancel Previous **Next**

### 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

↻ View

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  **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

Clear selection

Add database

Table name prefix - *optional*

Type a prefix added to table names

Maximum table threshold - *optional*

You can define 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

Cancel

Previous

Next

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 Edit

**Data sources (1) Info**

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

Type	Data source	Parameters
S3	s3://glue123/input/	Recrawl all

### Step 3: Configure security settings Edit

**Configure security settings**

IAM role AWSGlueServiceRole-glue123	Security configuration -	Lake Formation configuration -
--	-----------------------------	-----------------------------------

### Step 4: Set output and scheduling Edit

**Set output and scheduling**

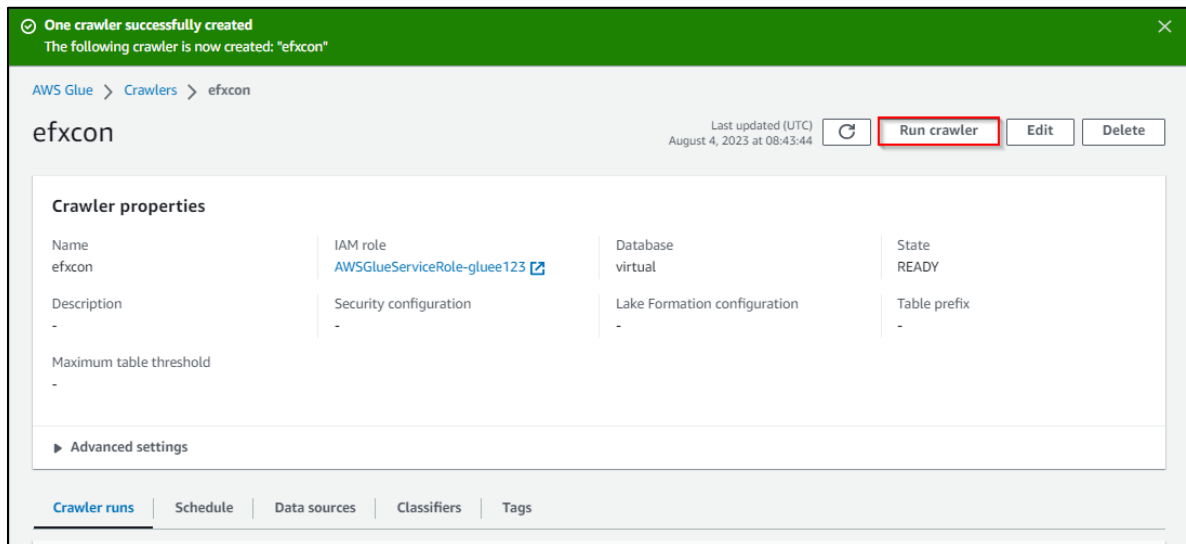
Database virtual	Table prefix - <i>optional</i> -	Maximum table threshold - <i>optional</i> -	Schedule On demand
---------------------	-------------------------------------	--	-----------------------

Cancel

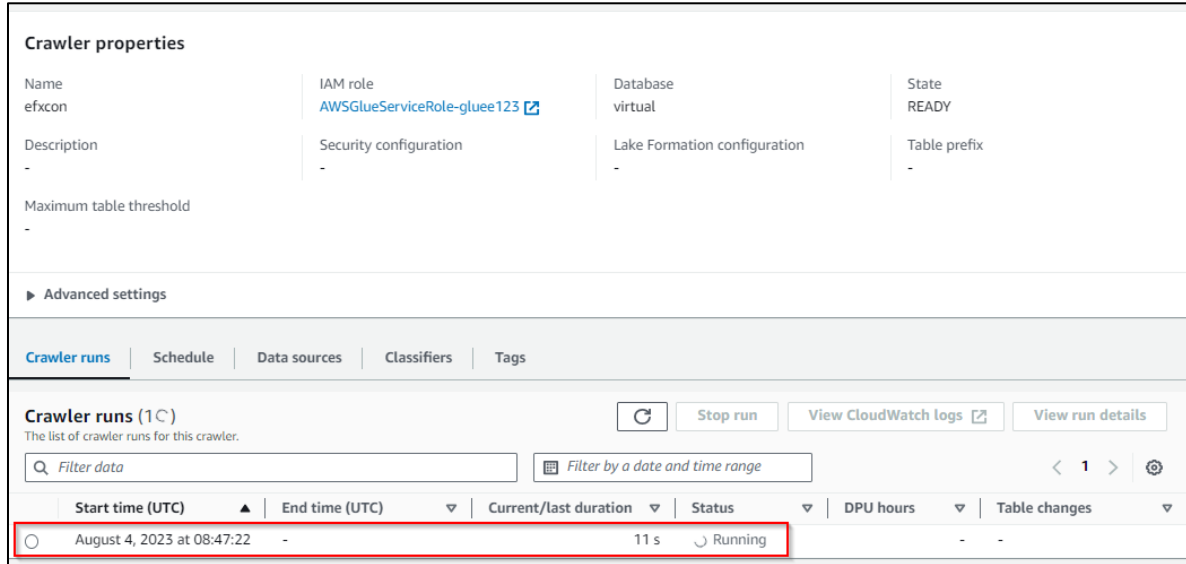
Previous

Create crawler

### 3.7 Click Run crawler



The **crawler** is running successfully, and the duration can be visible as shown in the screenshot below:



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