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

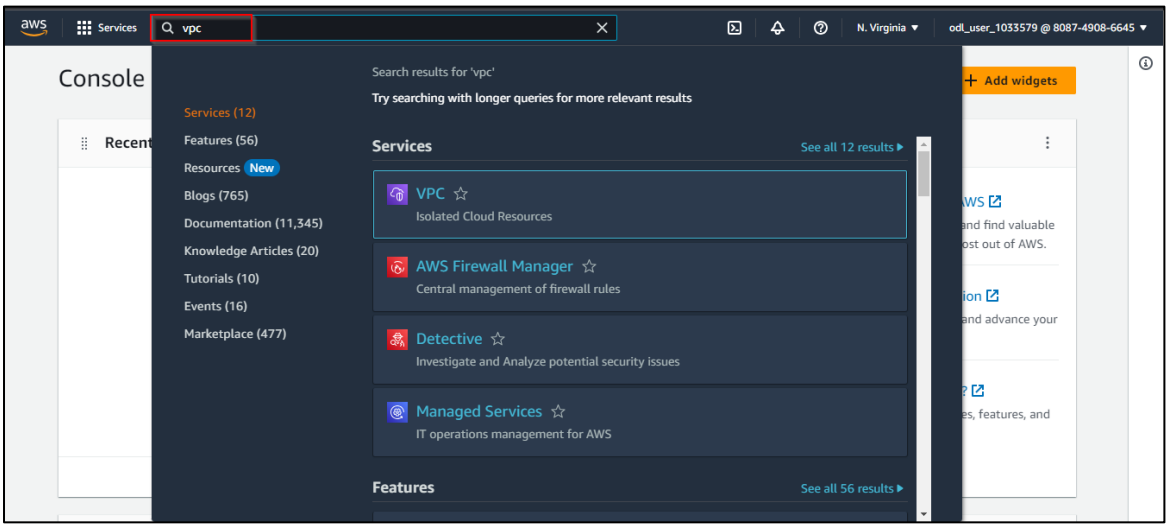
Prerequisites: AWS account

Steps to be followed:

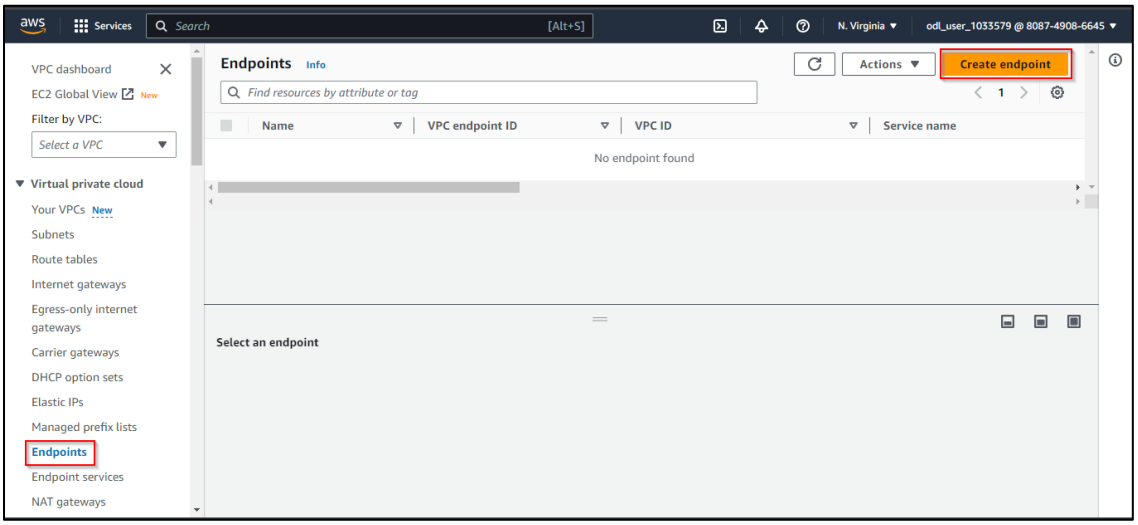
1. Create a VPC endpoint
2. Create a Glue Data Catalog
3. Add a data source to crawlers

Step 1: Create a VPC endpoint

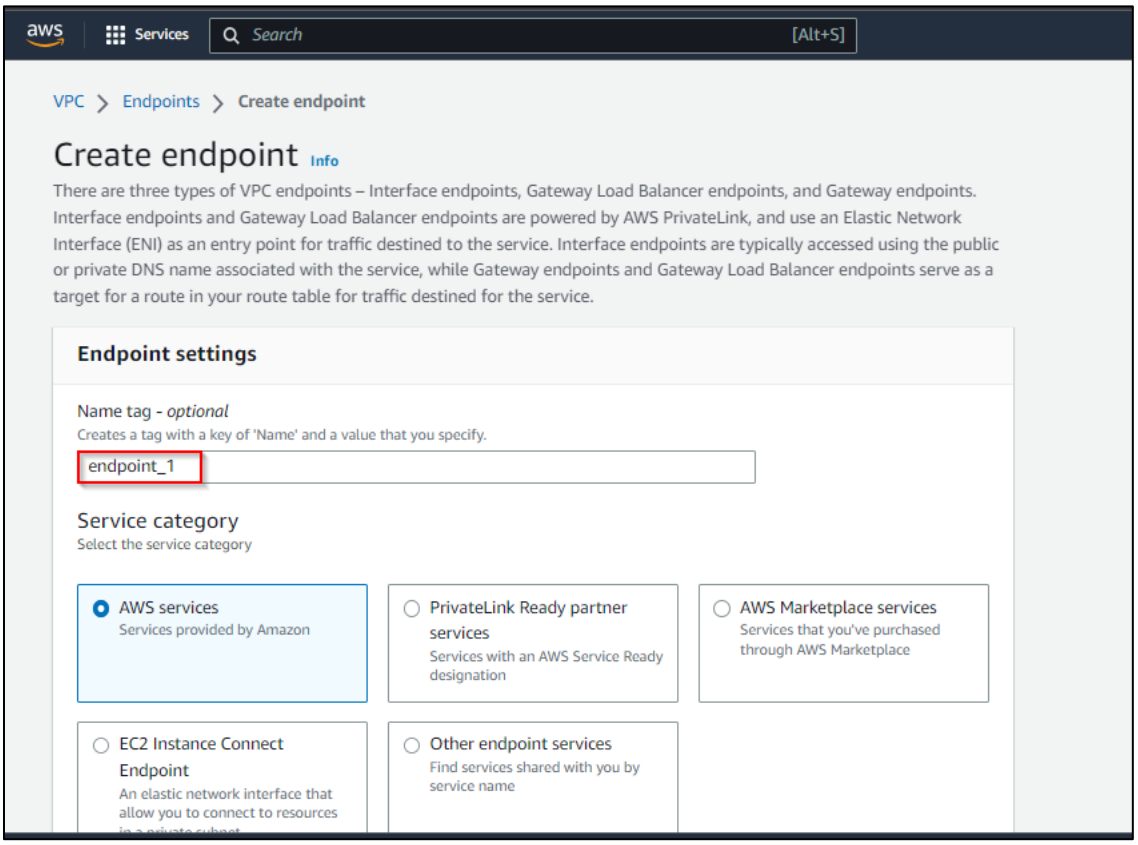
1.1 Navigate to the AWS portal homepage, search for **VPC**, and click on it



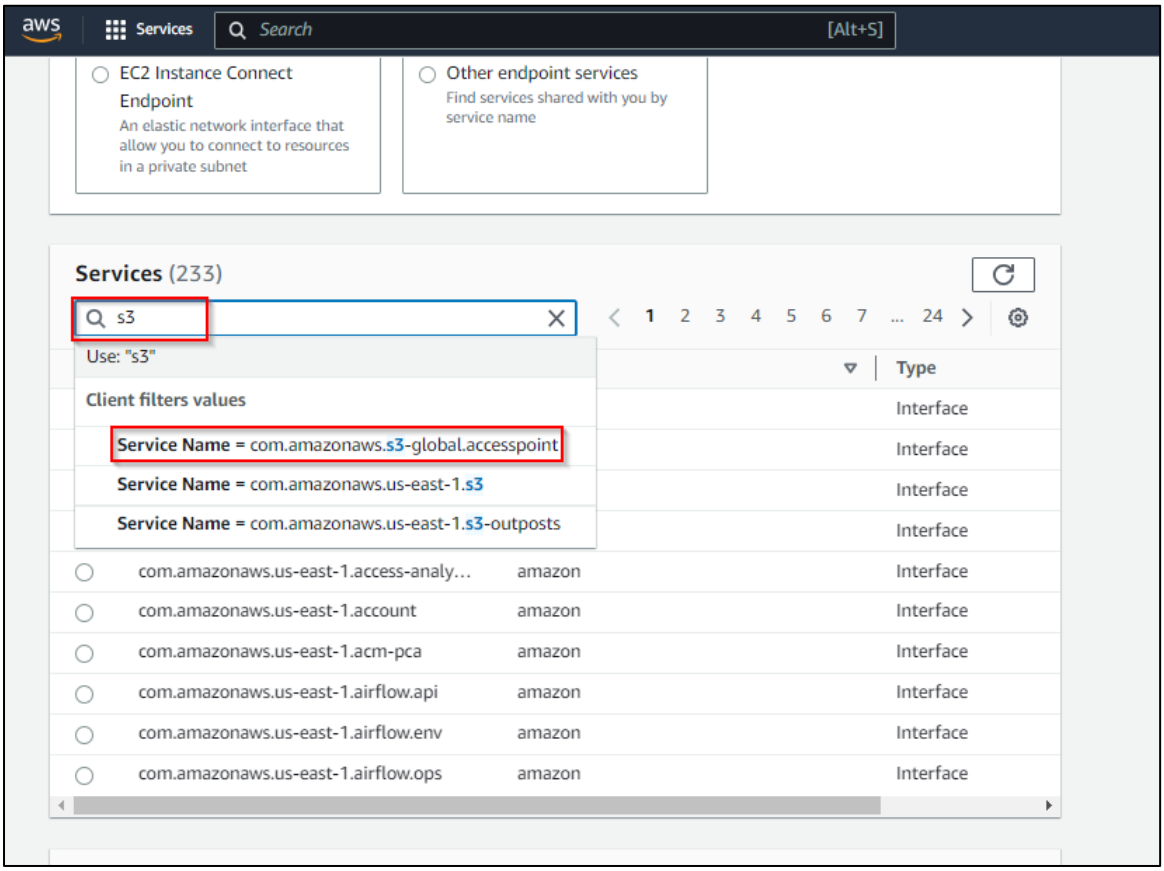
1.2 Click on **Endpoints** and then select **Create endpoint**



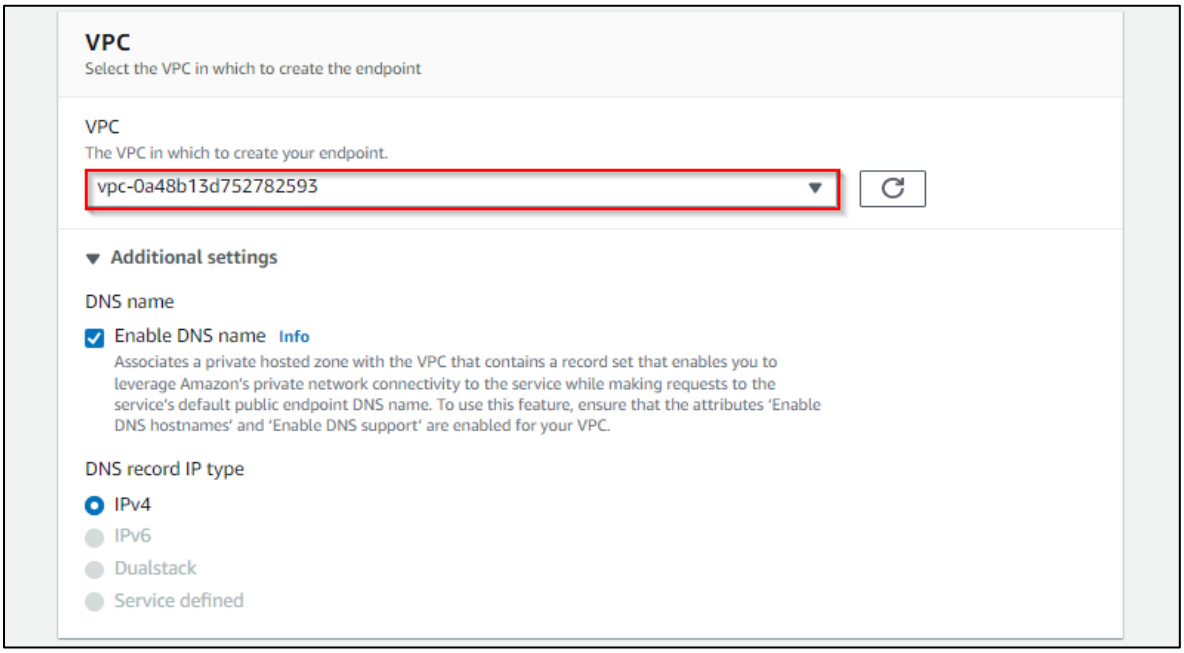
1.3 Enter **endpoint_1** as the **Name tag**



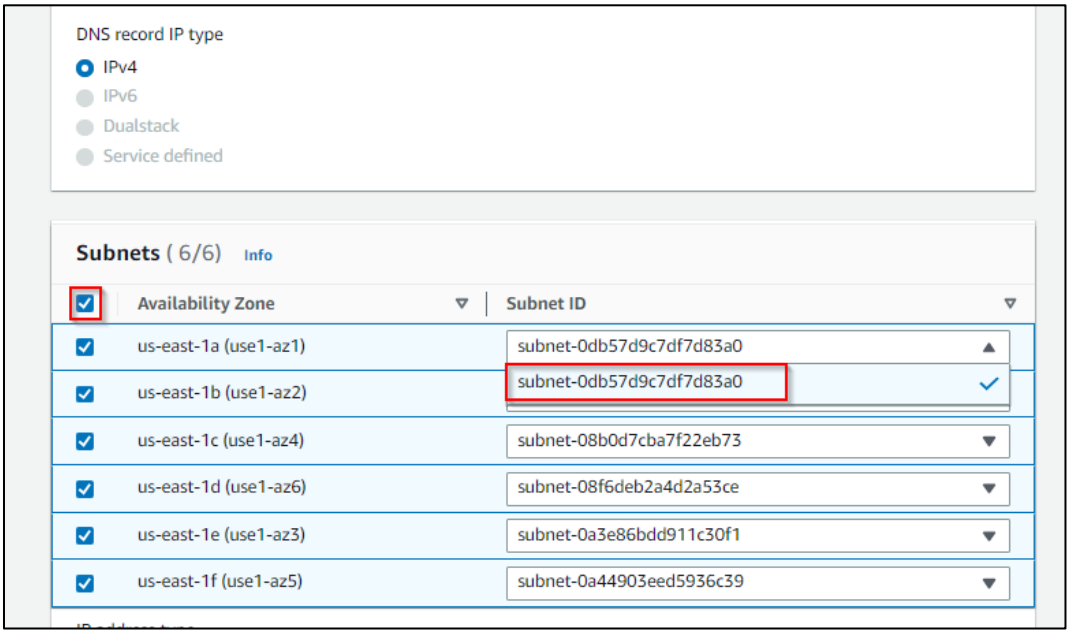
1.4 Under **Services**, enter **s3** and choose **s3-global.accesspoint**



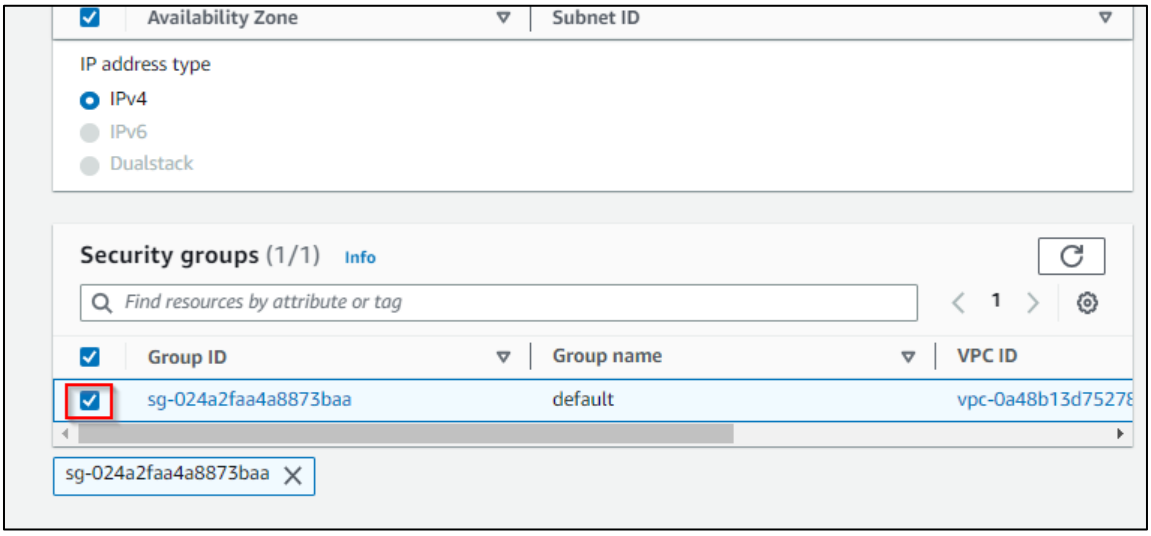
1.5 Choose the default **VPC** option under **VPC**



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

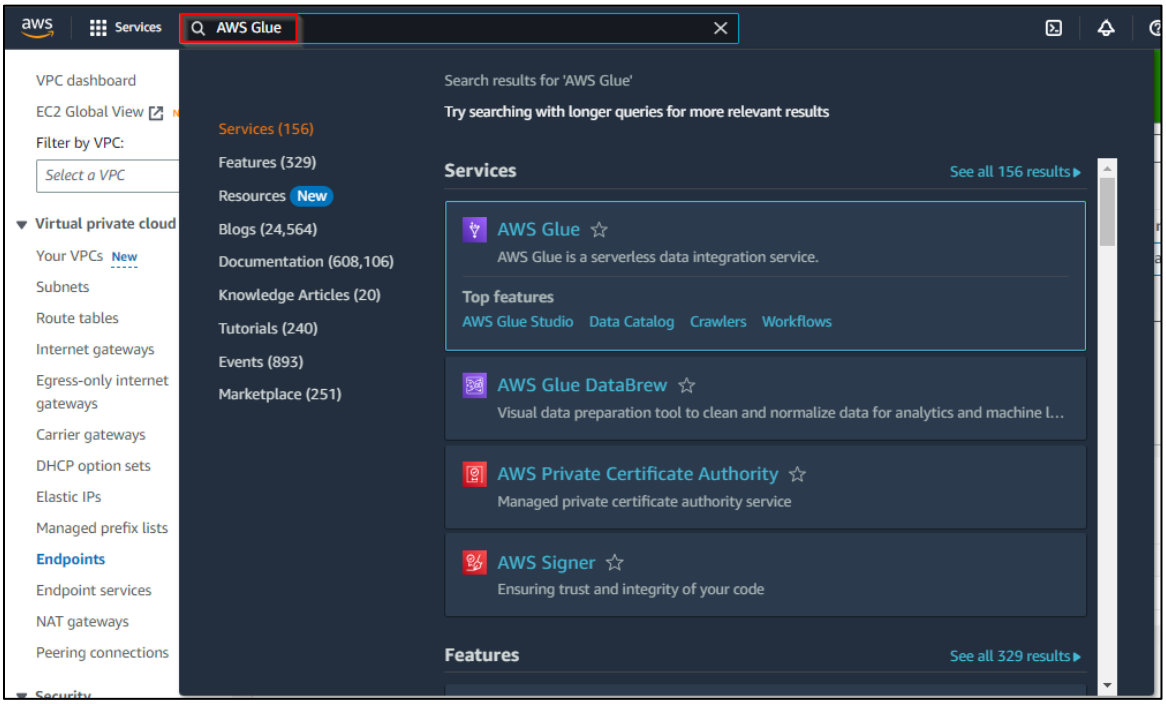


1.7 Select the appropriate **Group ID** under **Security groups**, and create the endpoint

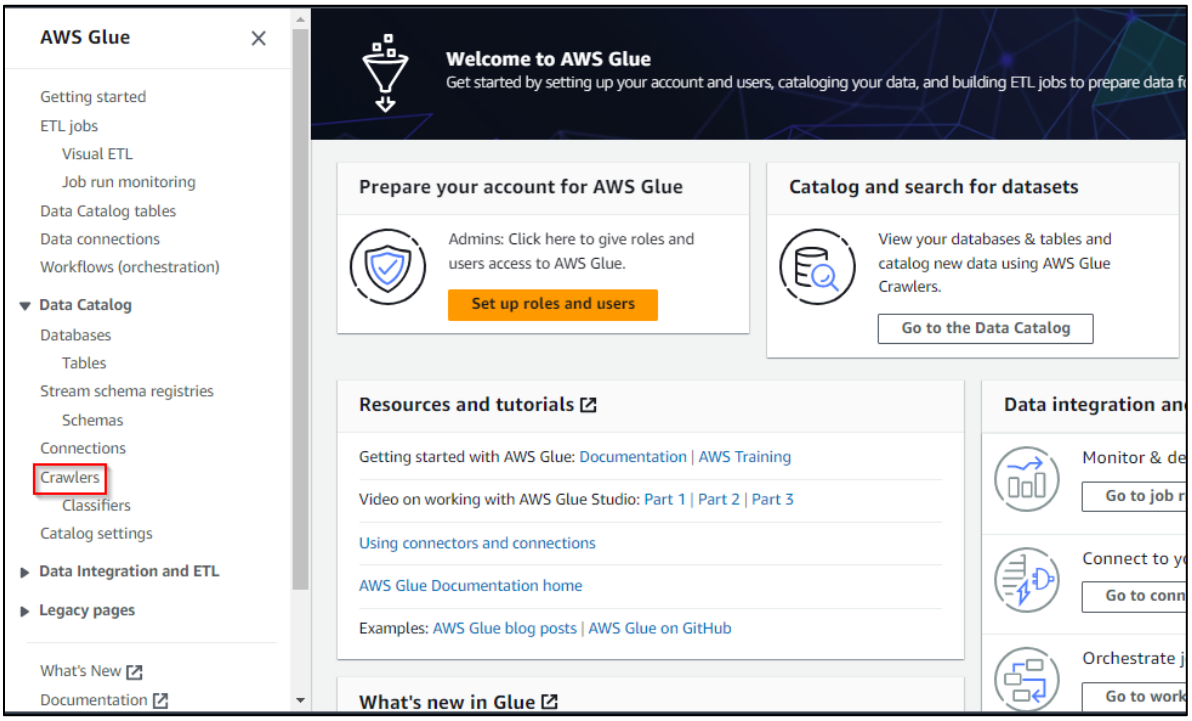


Step 2: Create a Glue Data Catalog

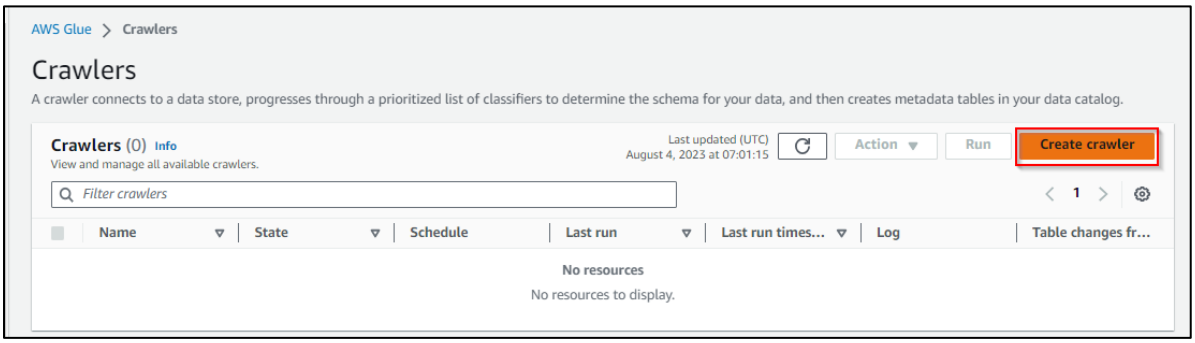
2.1 Navigate to the AWS portal homepage and search for **AWS Glue**



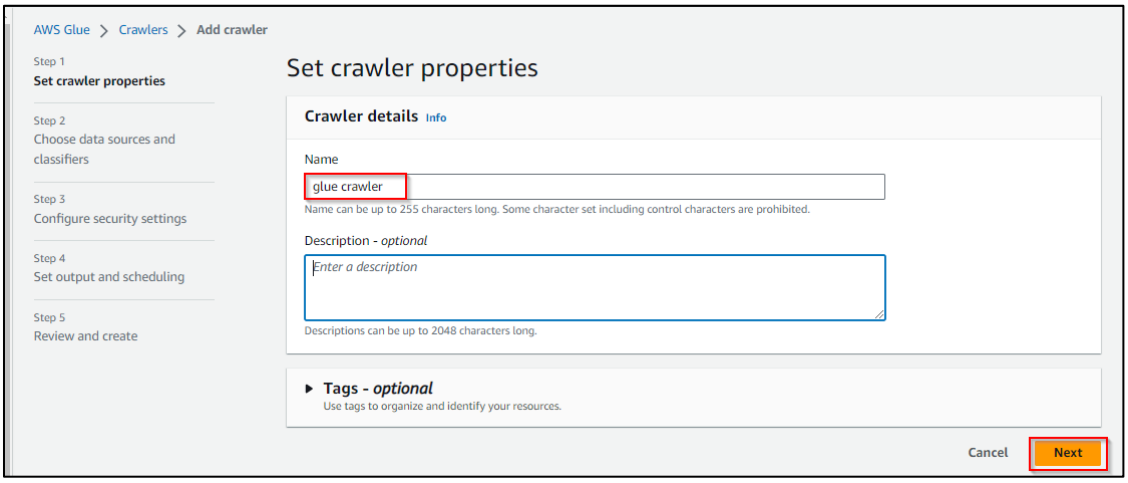
2.2 Click on **Crawlers** under **Data Catalog**



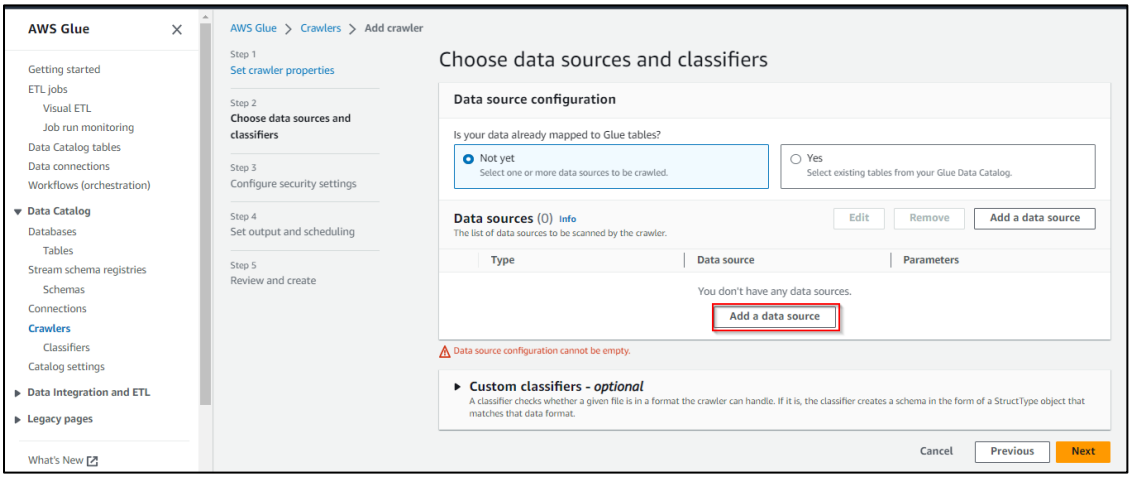
2.3 Select **Create crawler**



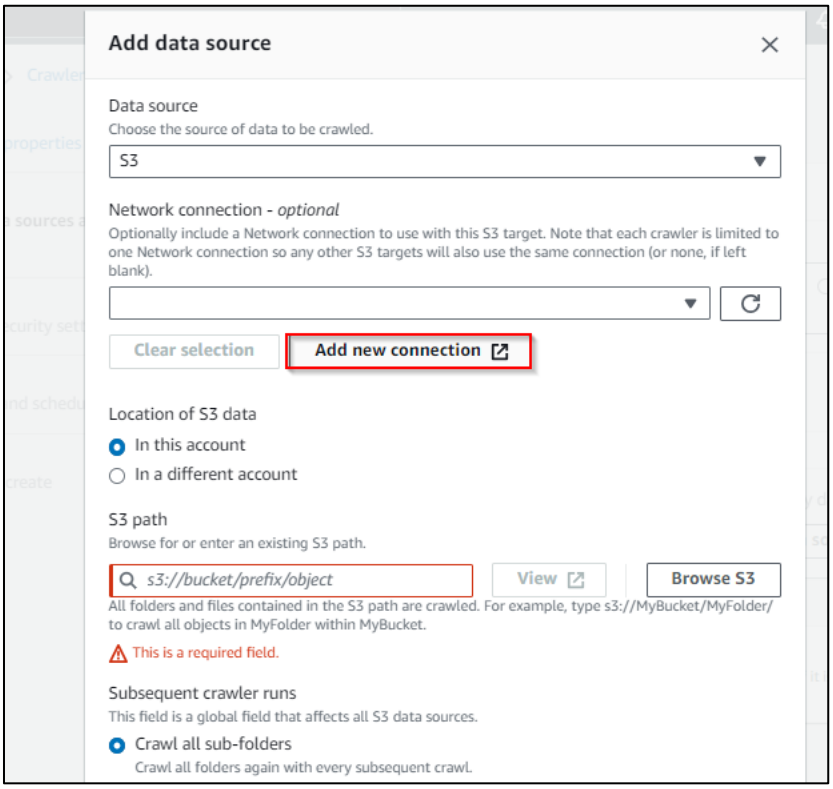
2.4 Enter the name as **glue crawler** and click **Next**



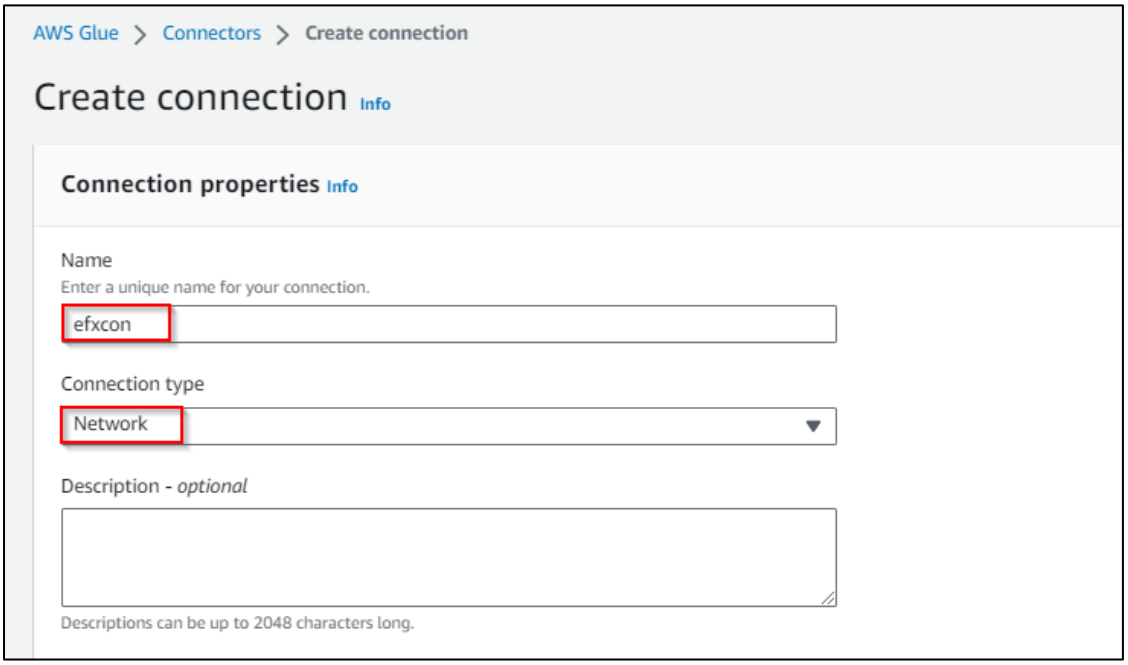
2.5 Click on **Add a data source**



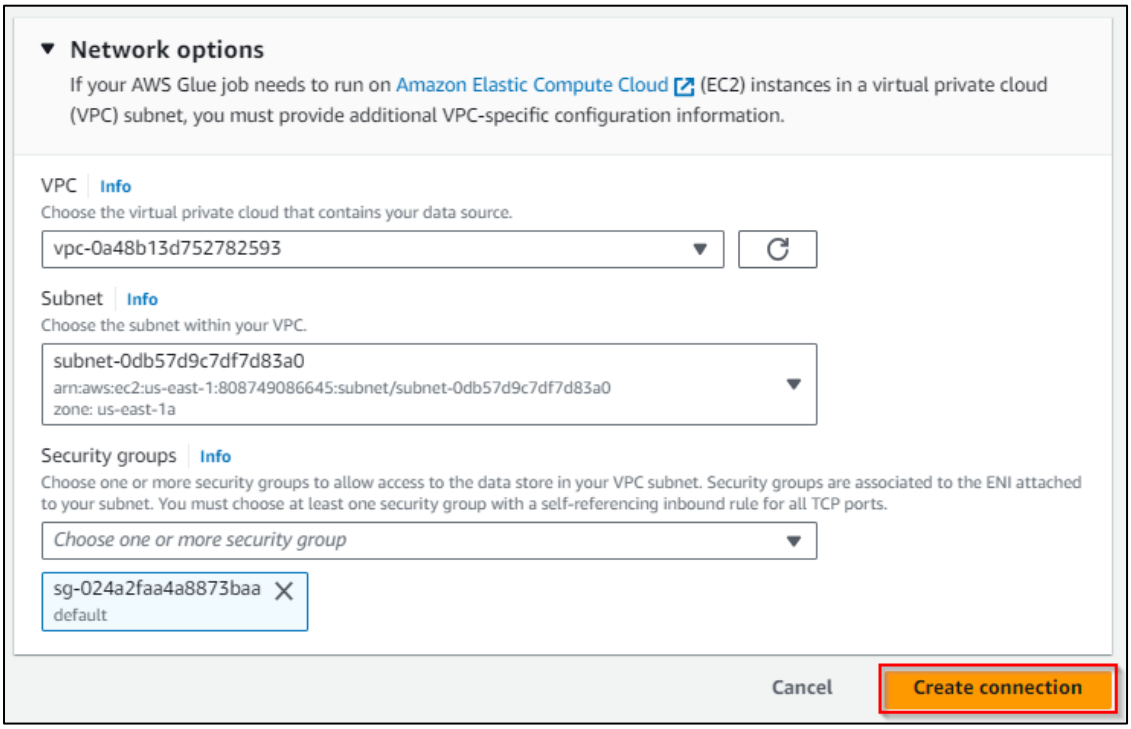
2.6 Click on **Add new connection**



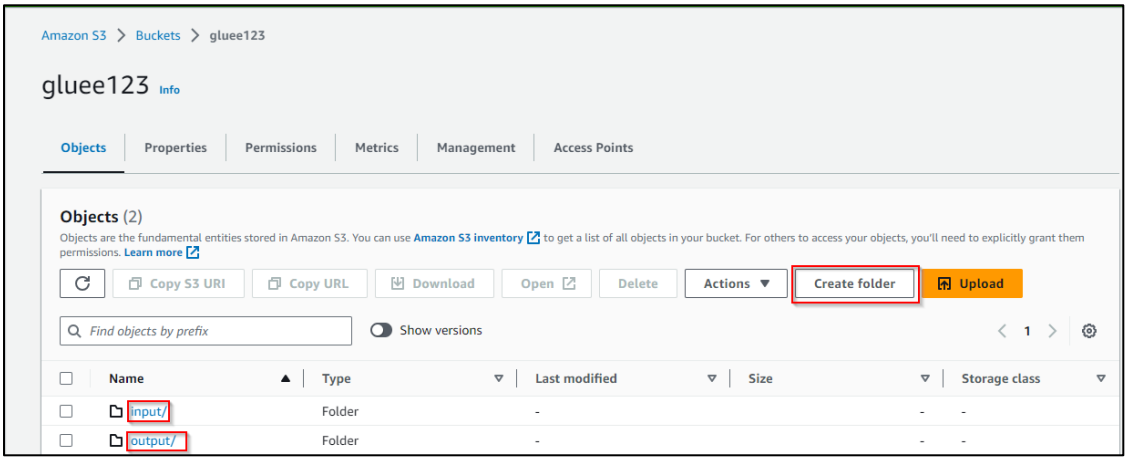
2.7 Enter the desired name and select the connection type as **Network**



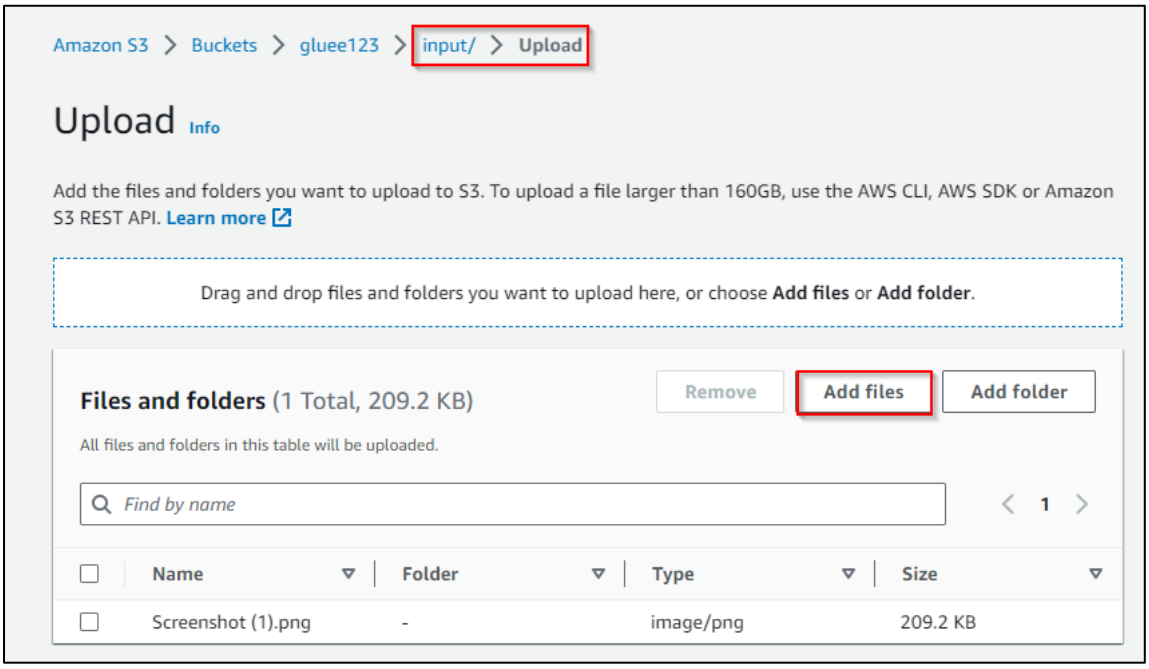
2.8 Configure the **Network options** as shown in the screenshot, and then click on and **Create connection**



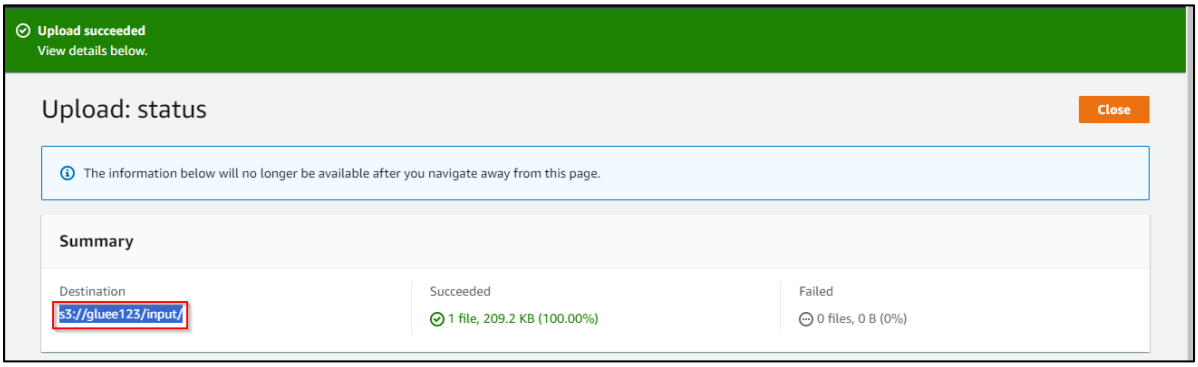
2.9 Create an S3 bucket named **gluee123** and **input** and **output** folders within it



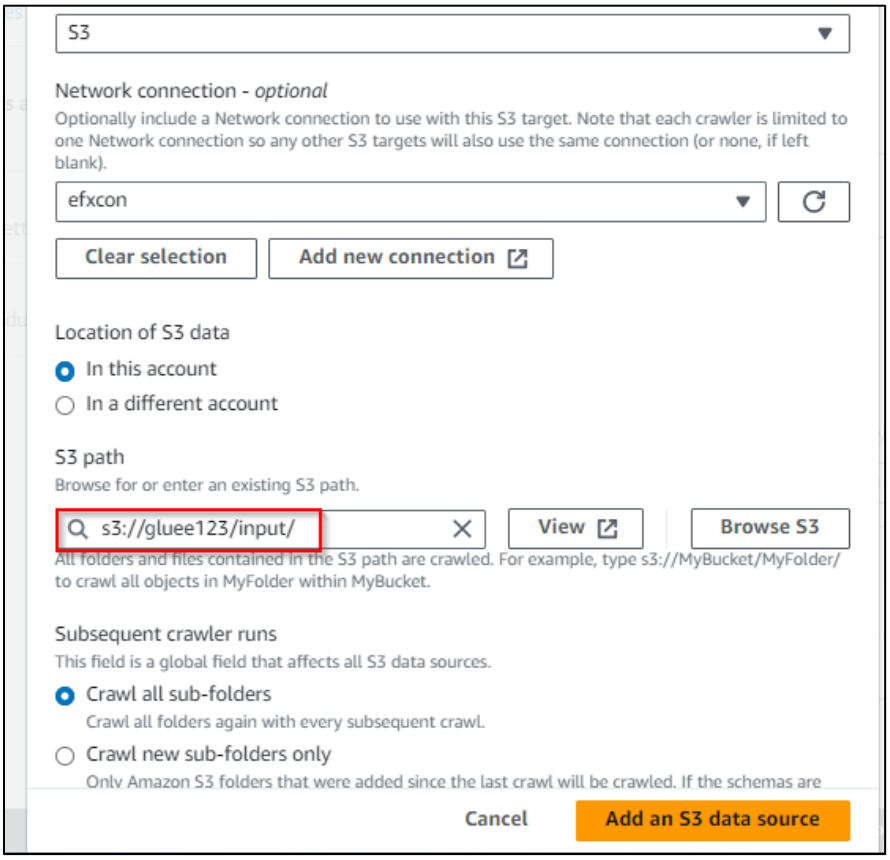
2.10 Navigate to the **input folder**, and drag and drop or click on **Add files** to upload an image



2.11 Copy the Destination path for use in the S3 path

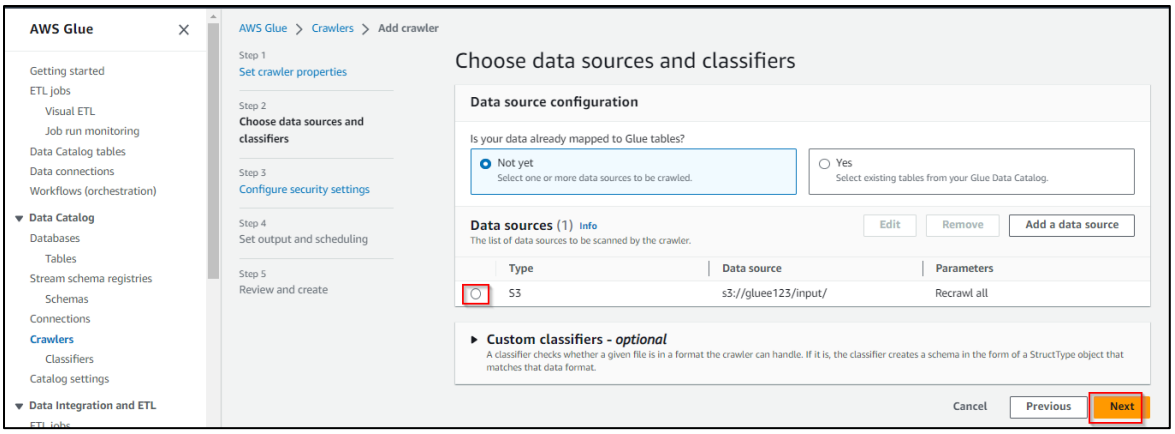


2.12 Paste the S3 path location as shown in the screenshot, and then click on **Add an S3 data source**

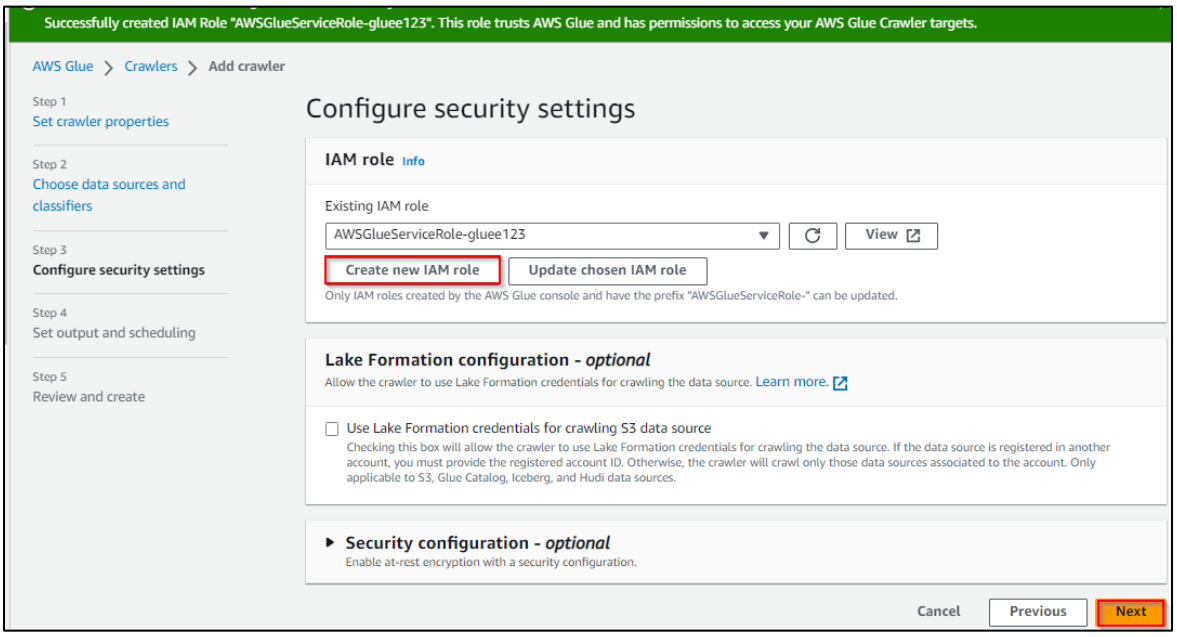


Step 3: Add data source to Crawlers

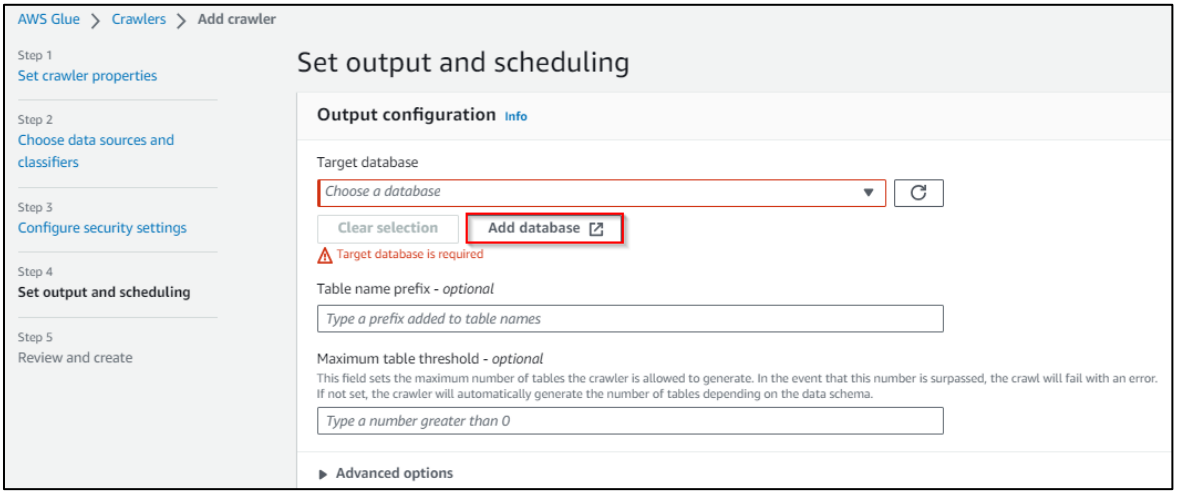
3.1 Select **S3** type and click **Next**



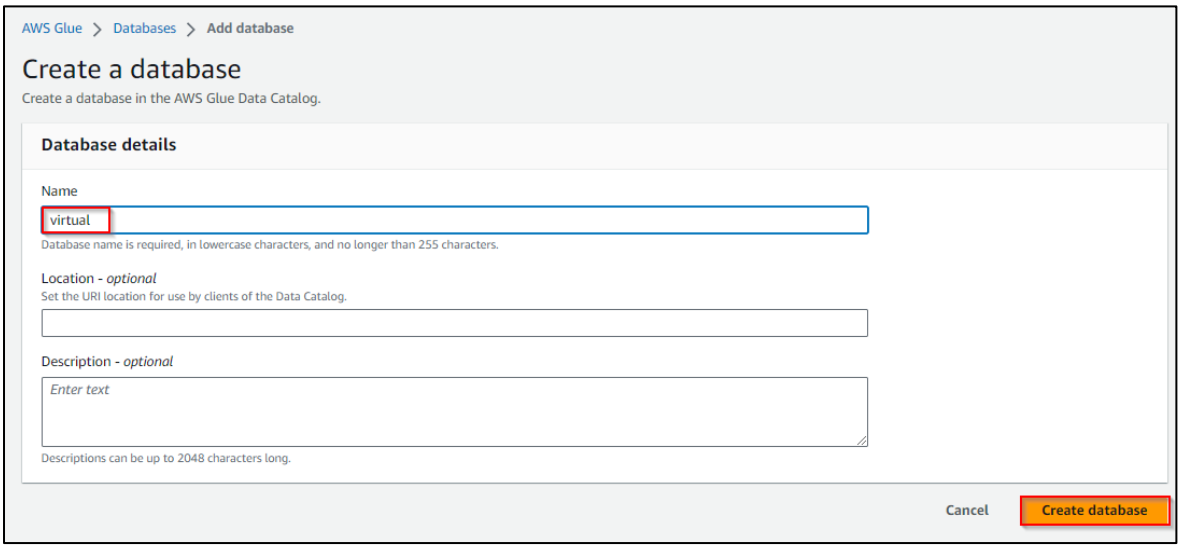
3.2 Create a new IAM role named **gluee123**, and click **Next**



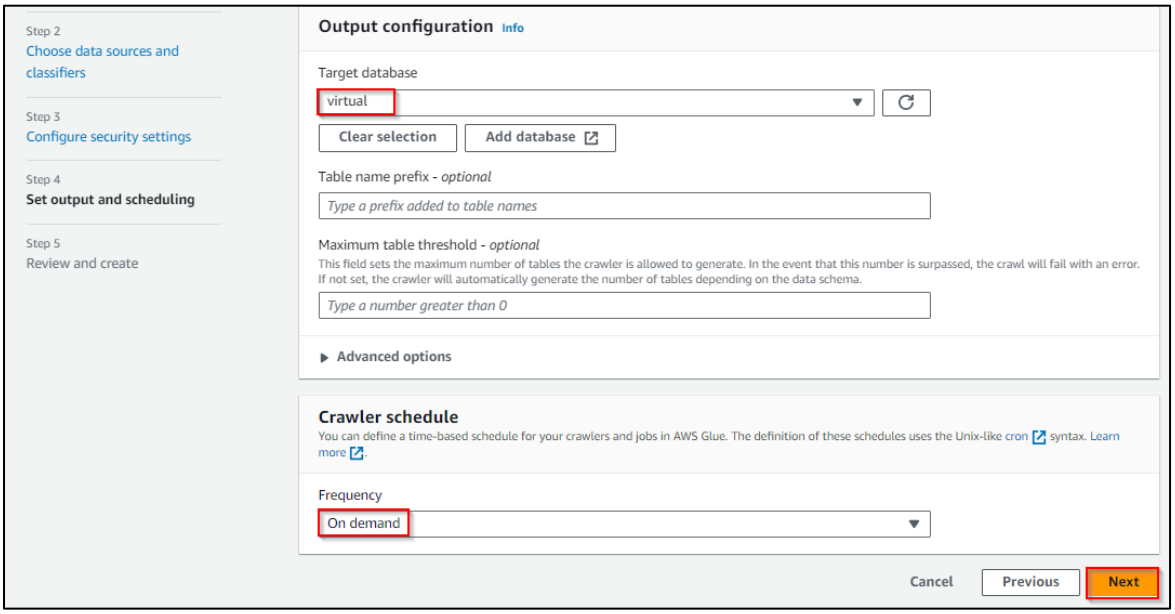
3.3 Click on **Add database**



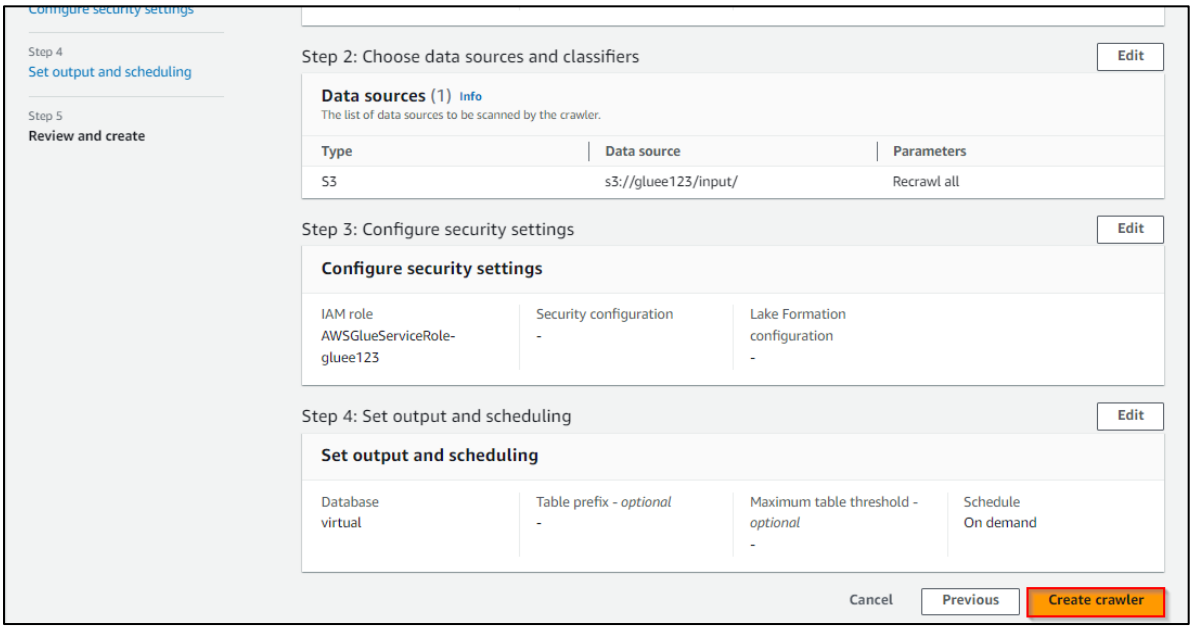
3.4 Name it **virtual**, and create the database



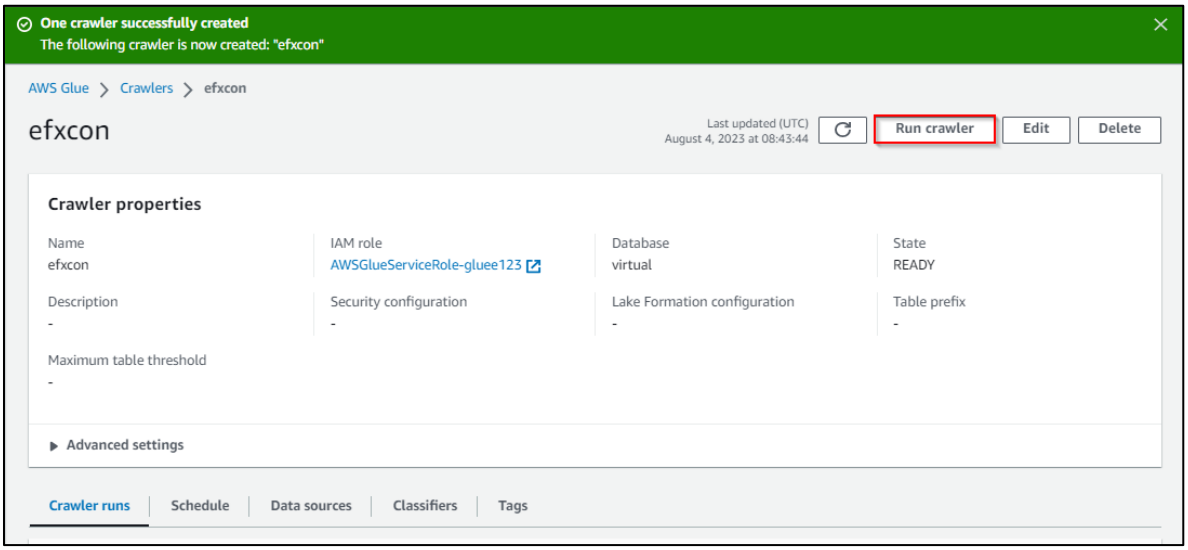
3.5 Select **virtual** from the database options, set the Frequency to **On demand**, and then click on **Next**



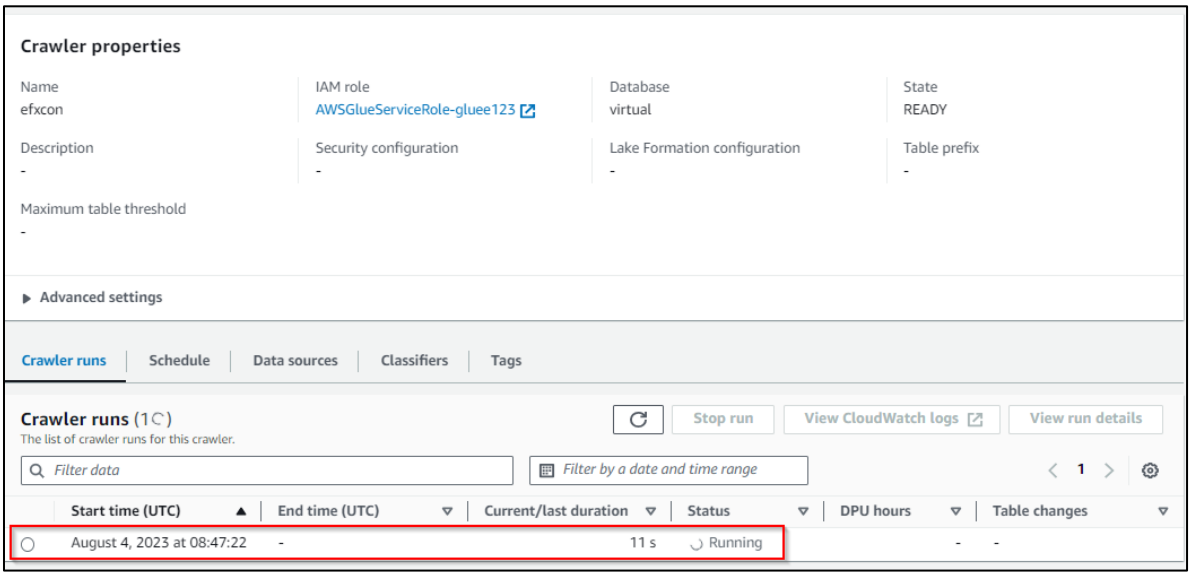
3.6 Review the settings, and create the crawler by clicking **Create crawler**

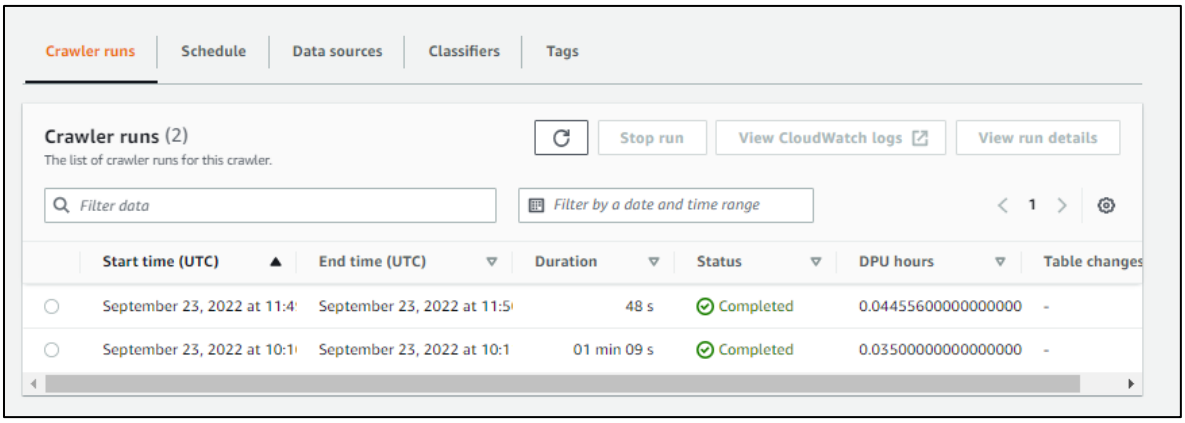


3.7 Click **Run crawler**



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.