Big Data: Amazon Athena, Glue, EMR

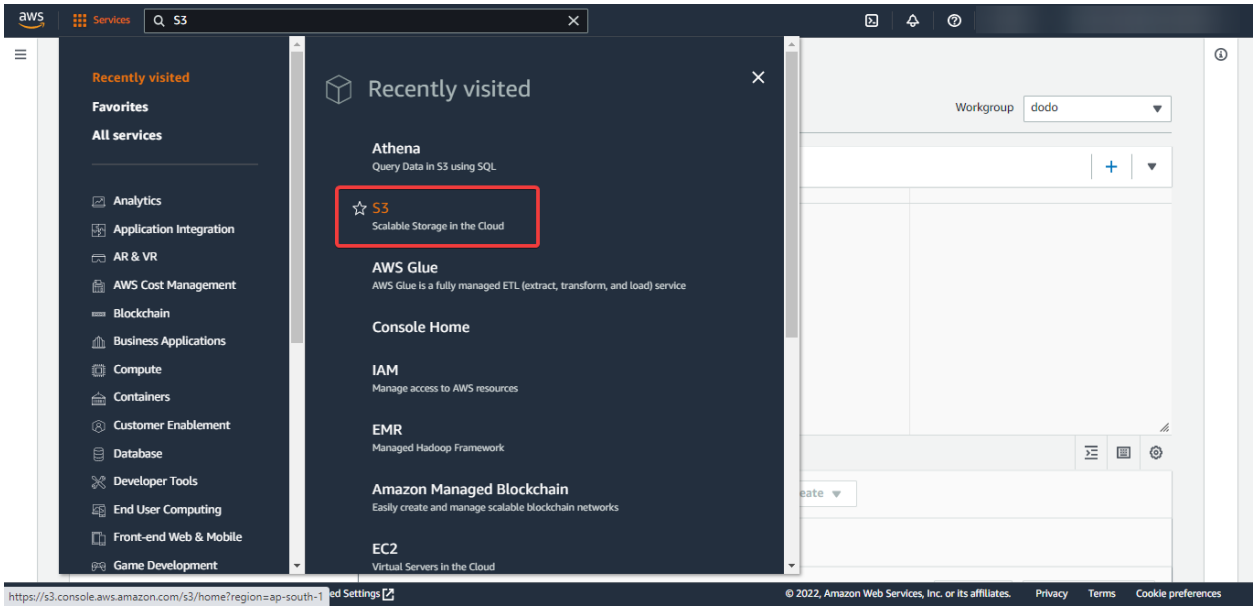
Table of Contents

Big Data: Amazon Athena, Glue, EMR

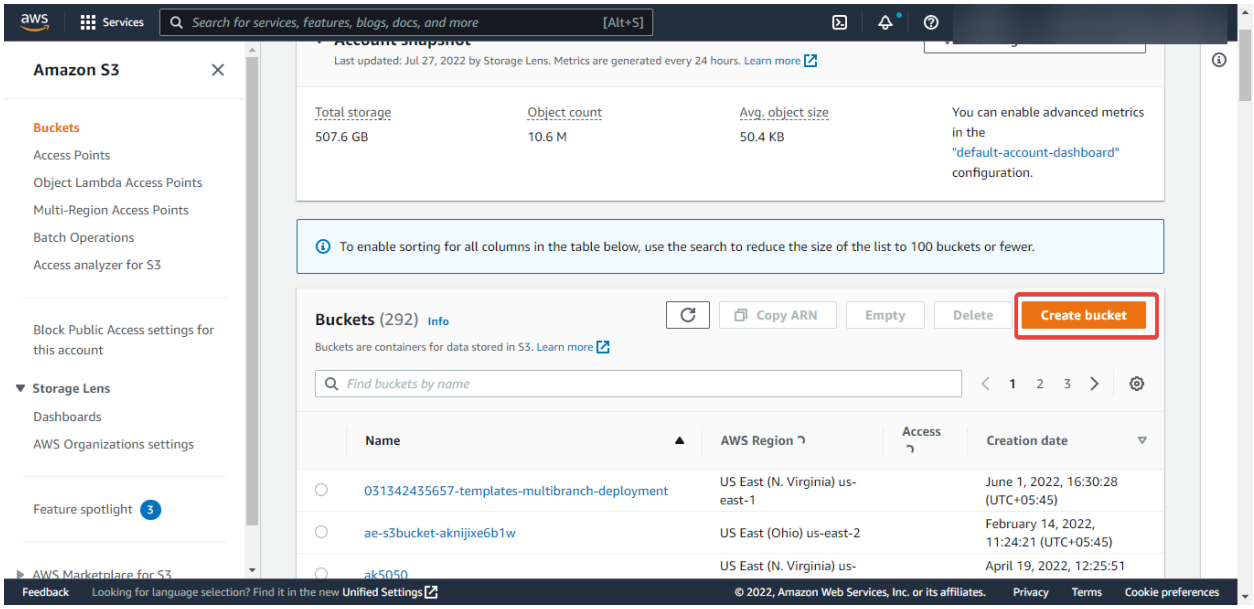
Query data from S3 files using Amazon Athena

Create a S3 bucket

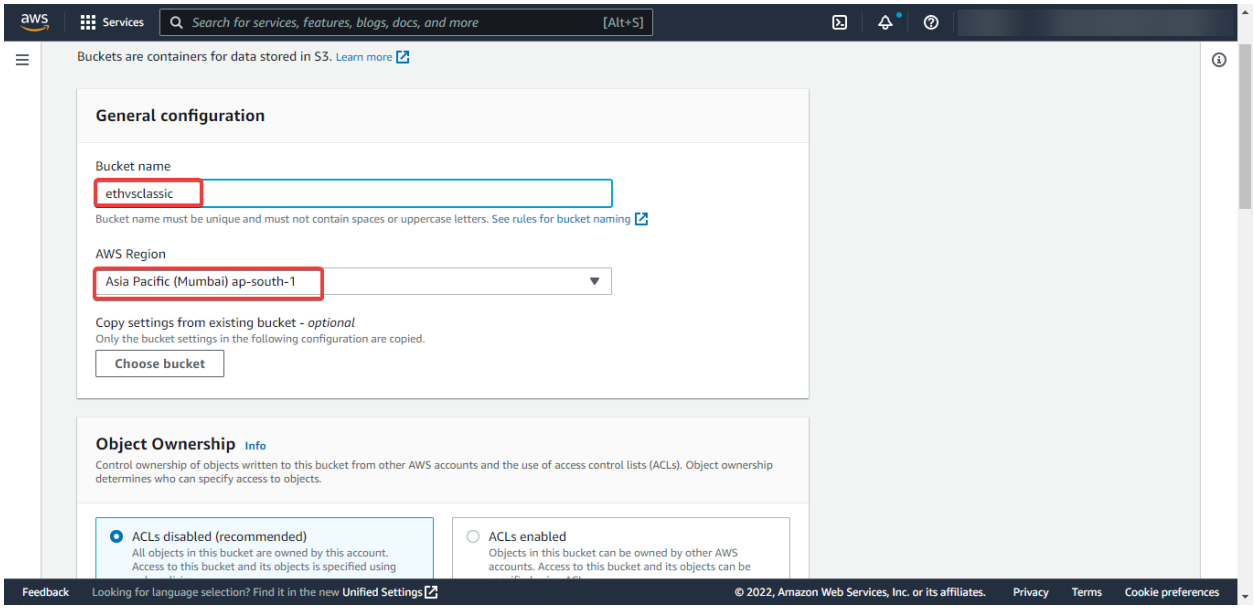
- Log in to **AWS Console**
- Select **S3.**



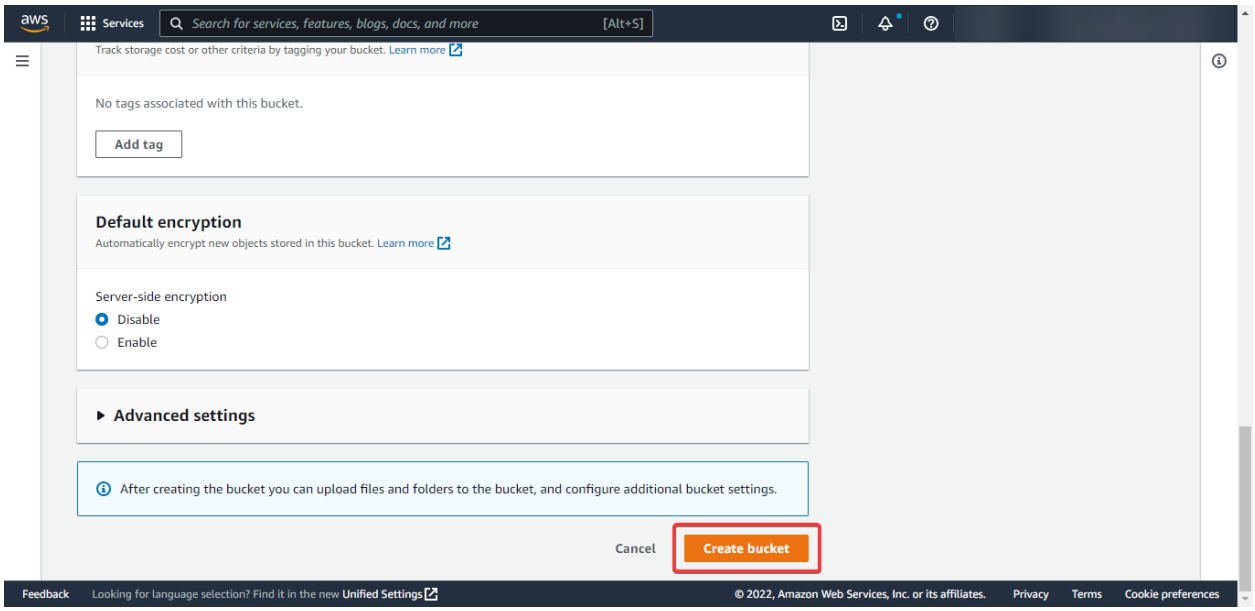
- Click **Create bucket**



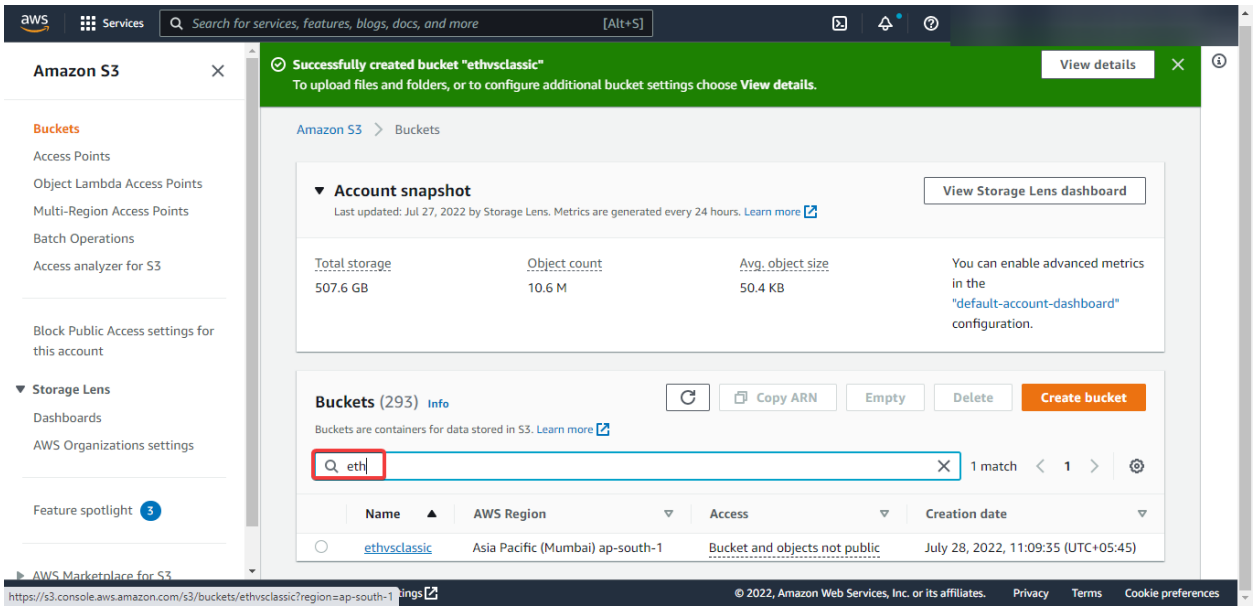
- Enter **Bucket name** ethvsclassic and select **AWS Region** Asia Pacific (Mumbai)ap-south-1



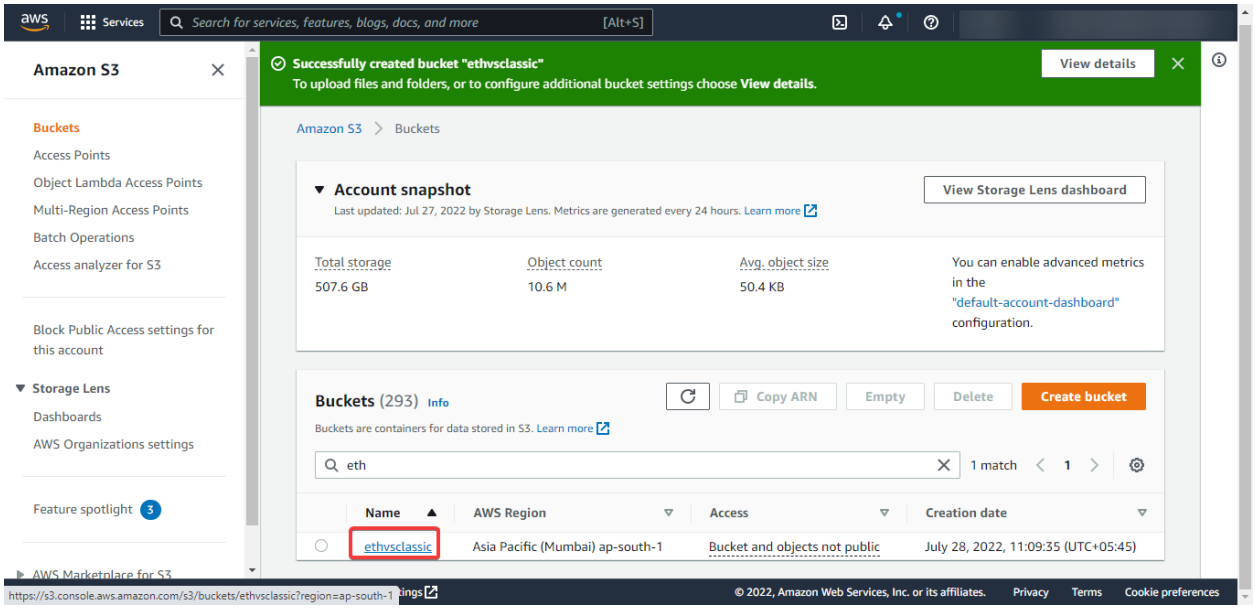
- Keep others default option.
- Once completed **Create bucket.**



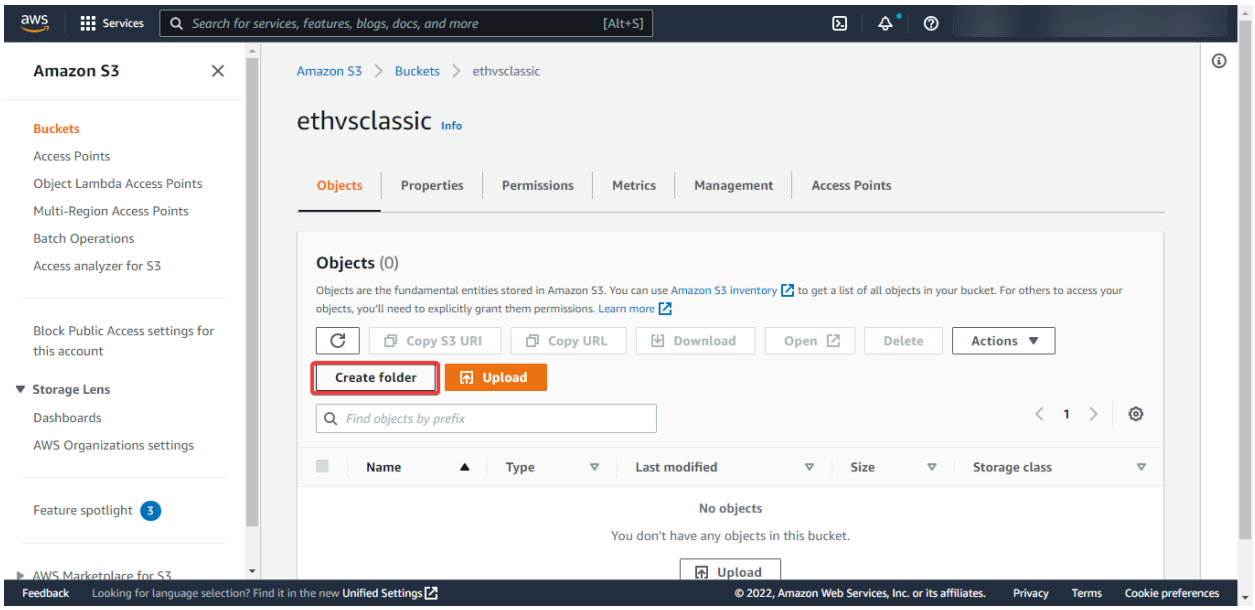
- Finally, you will see successfully bucket create at the top of the page.

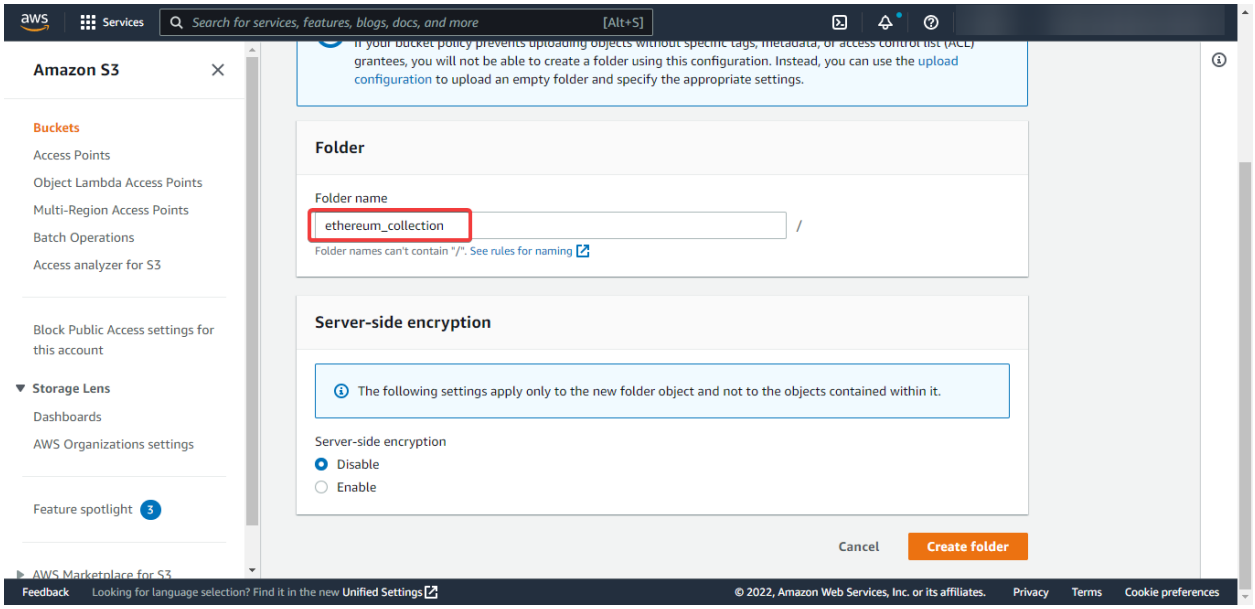


Click on your bucket **ethvsclassic.**

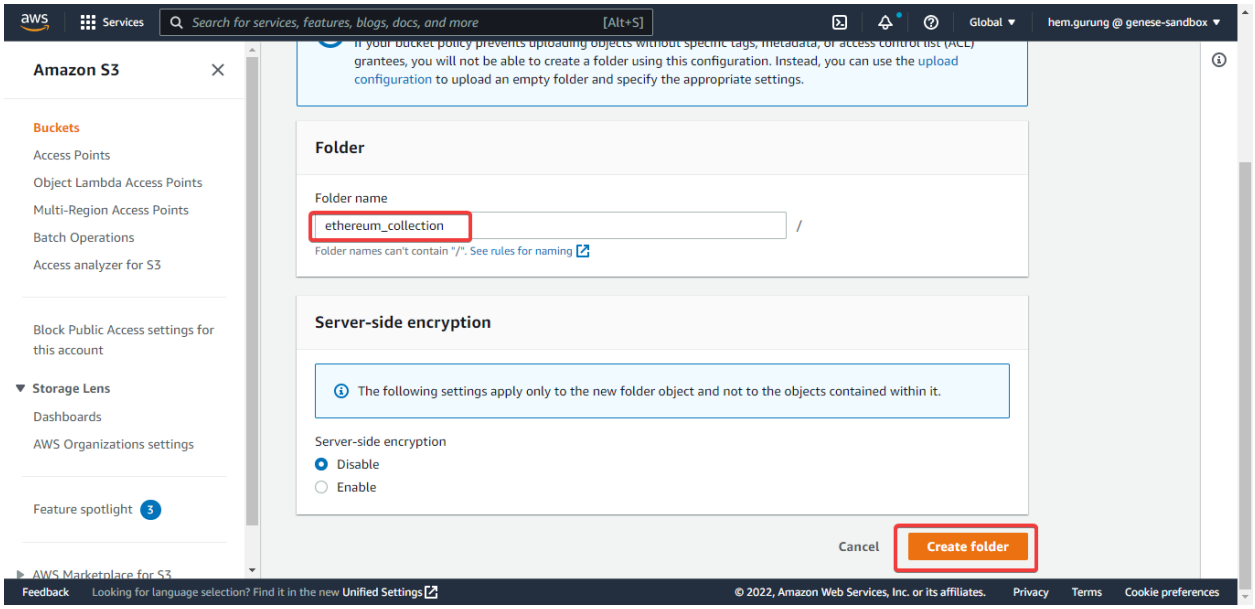


Click **create folder**.

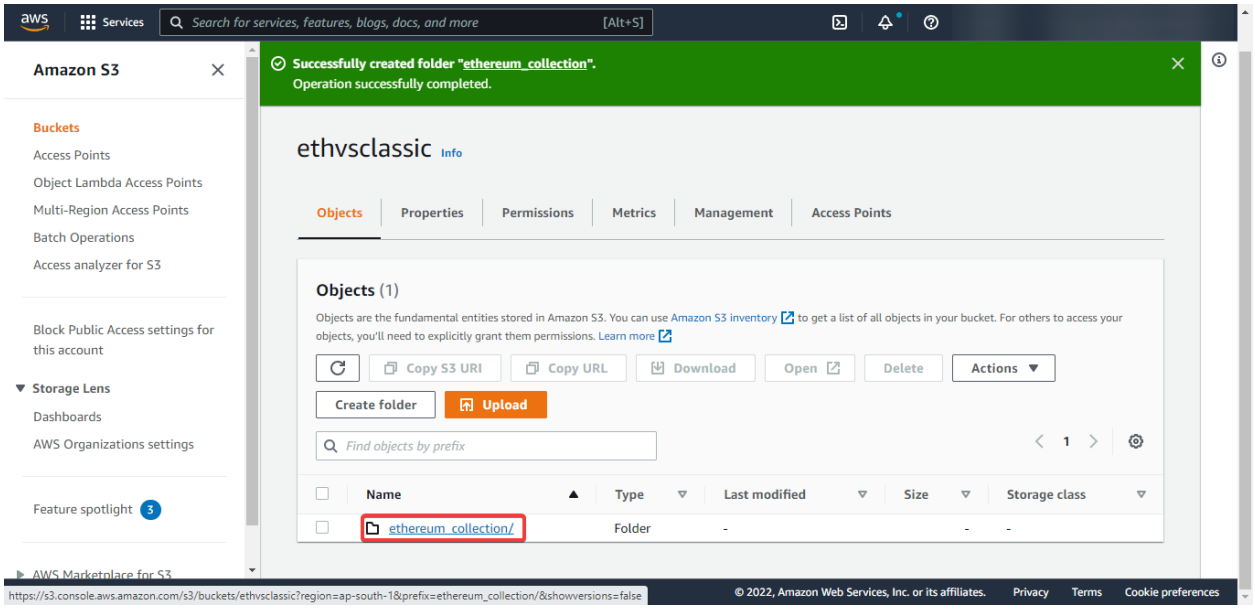




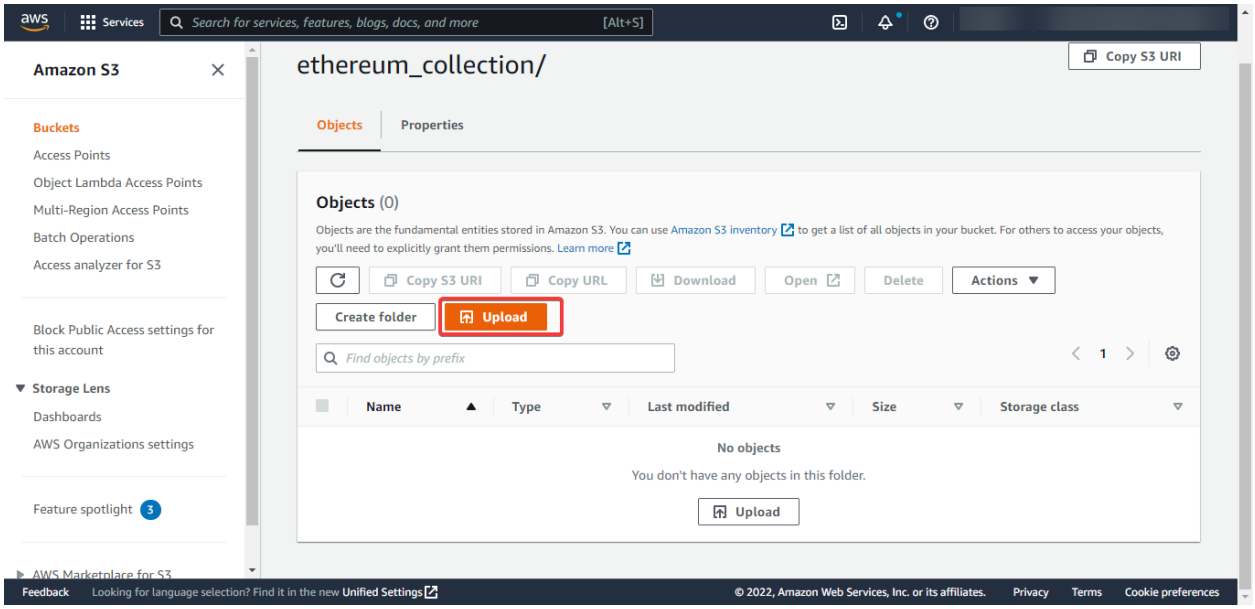
- Click **create folder**



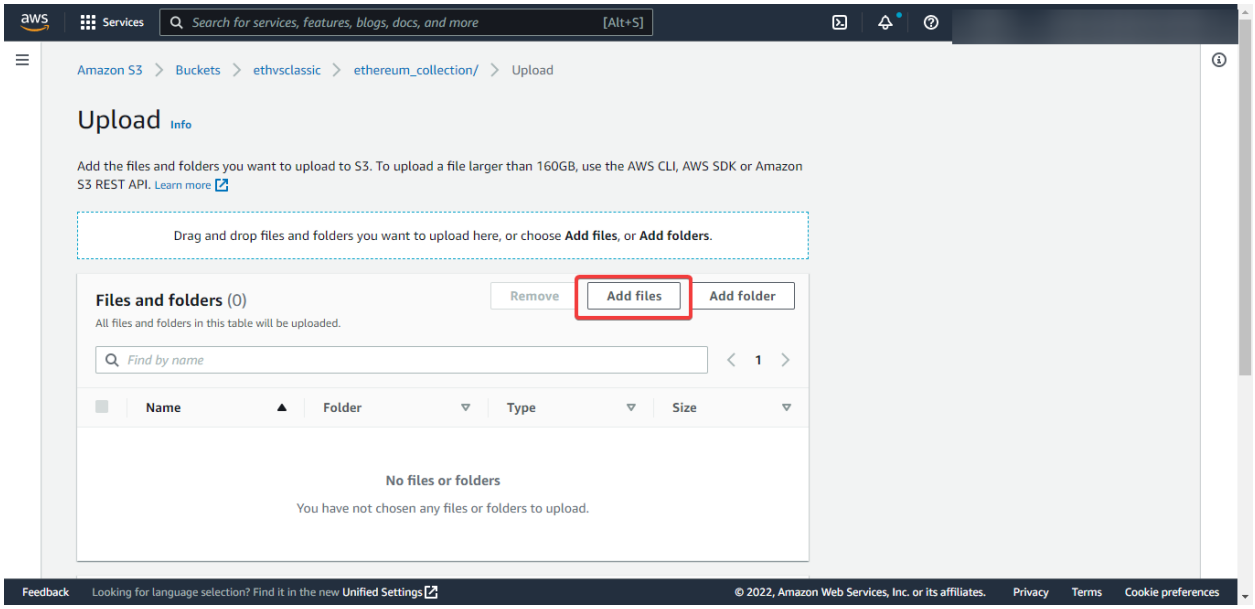
Open your folder **ethereum_collection**



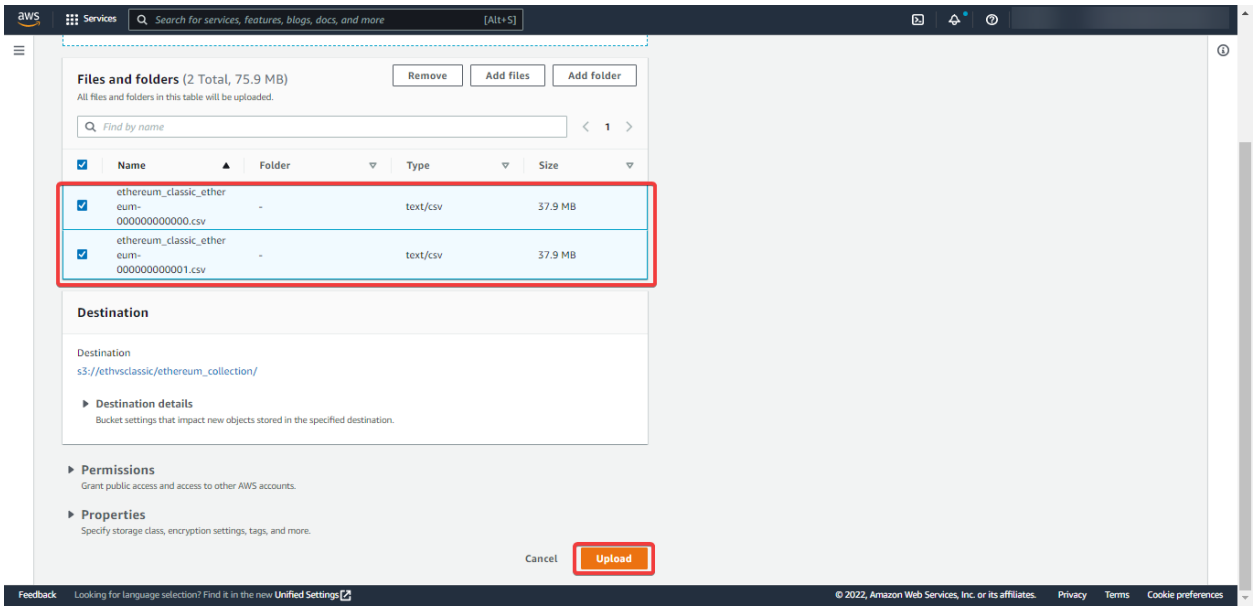
Upload files from your PC.



- **Add files**

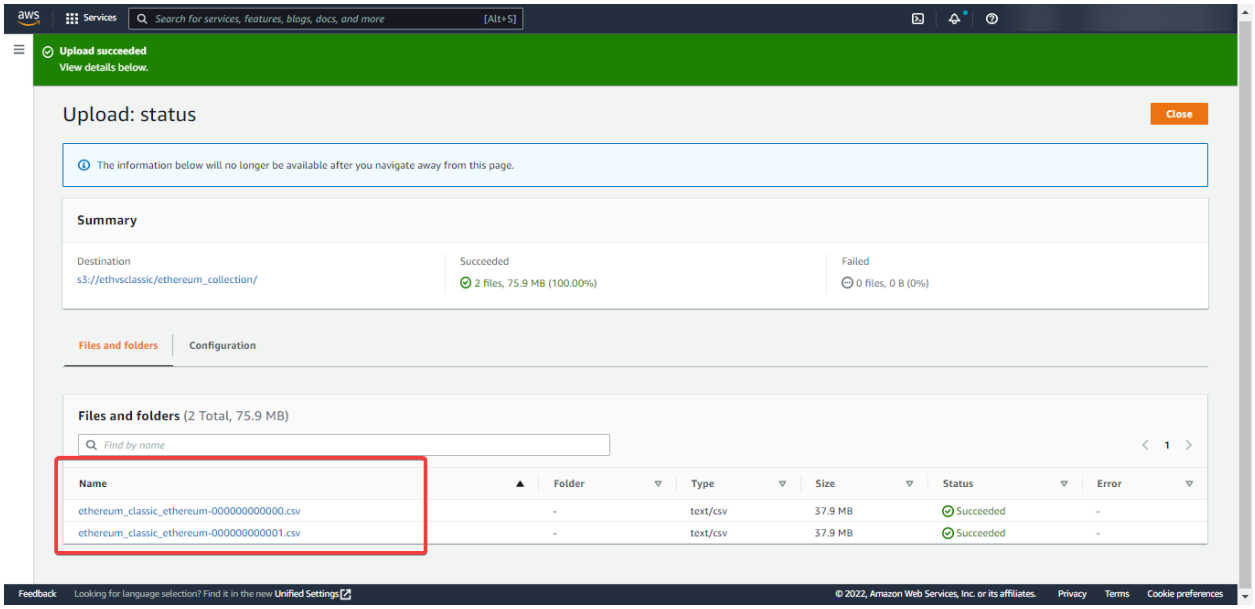


- Tick the uploaded files.
- Click **upload**.



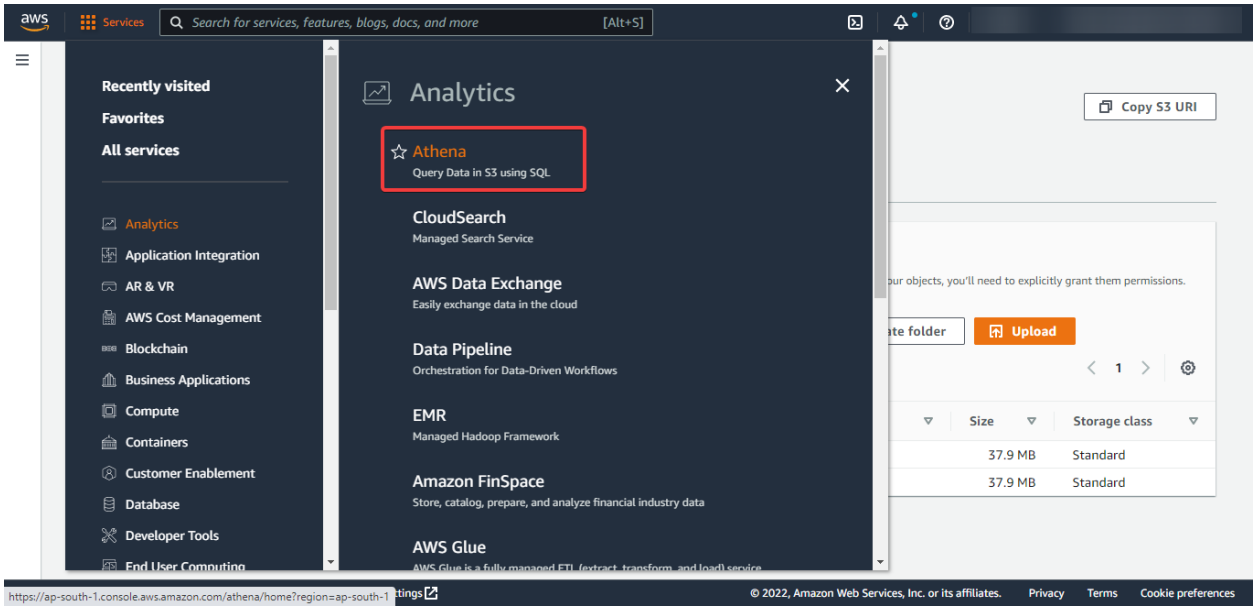
It makes take some time depend on file size.

- Finally, files have been uploaded in s3 bucket.


Services
Search for services, features, blogs, docs, and more
[Alt+S]
Upload succeeded
View details below.
Upload: status
Close
The information below will no longer be available after you navigate away from this page.
Summary
Destination
s3://ethvsclassic/ethereum_collection/
Succeeded
2 files, 75.9 MB (100.00%)
Failed
0 files, 0 B (0%)
Files and folders
Configuration
Files and folders (2 Total, 75.9 MB)
Find by name
1
Name
Folder
Type
Size
Status
Error
ethereum_classic_ethereum-000000000000.csv
-
text/csv
37.9 MB
Succeeded
-
ethereum_classic_ethereum-000000000001.csv
-
text/csv
37.9 MB
Succeeded
-
Feedback
Looking for language selection? Find it in the new Unified Settings
© 2022, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences

AWS Athena: Querying S3 data

- Go to **Amazon Athena** on AWS Console
- Select **Services** and click **Analytics**
- Select **Athena**



- Go to Query editor
- Click **Create** and select **S3 bucket data**

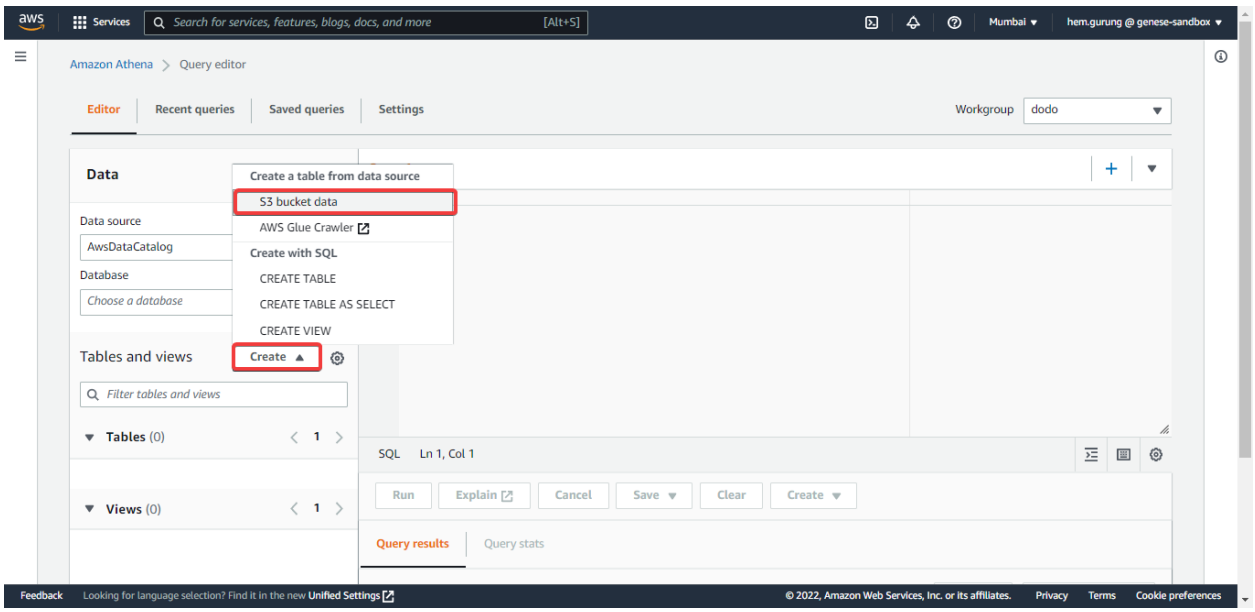
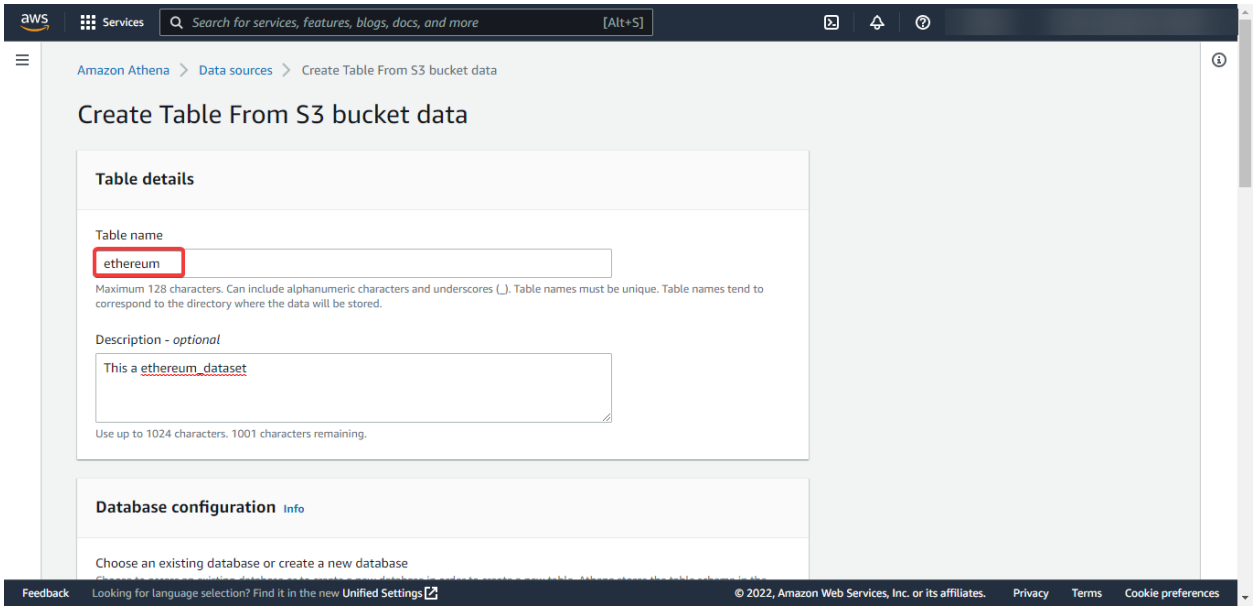
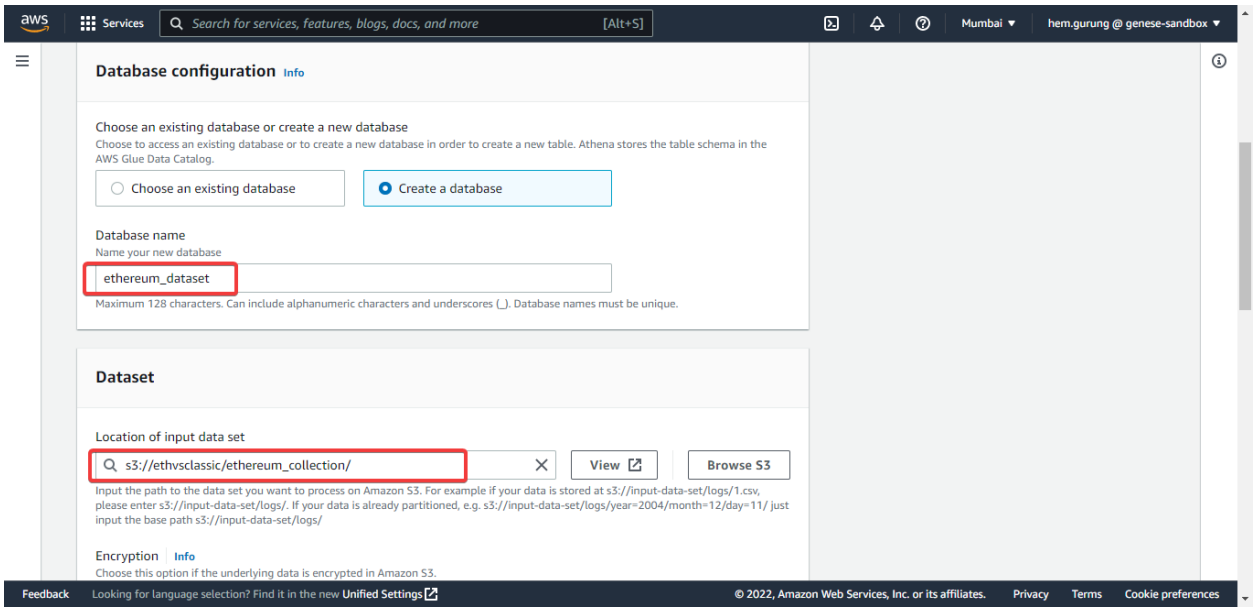


Table creation from S3 bucket

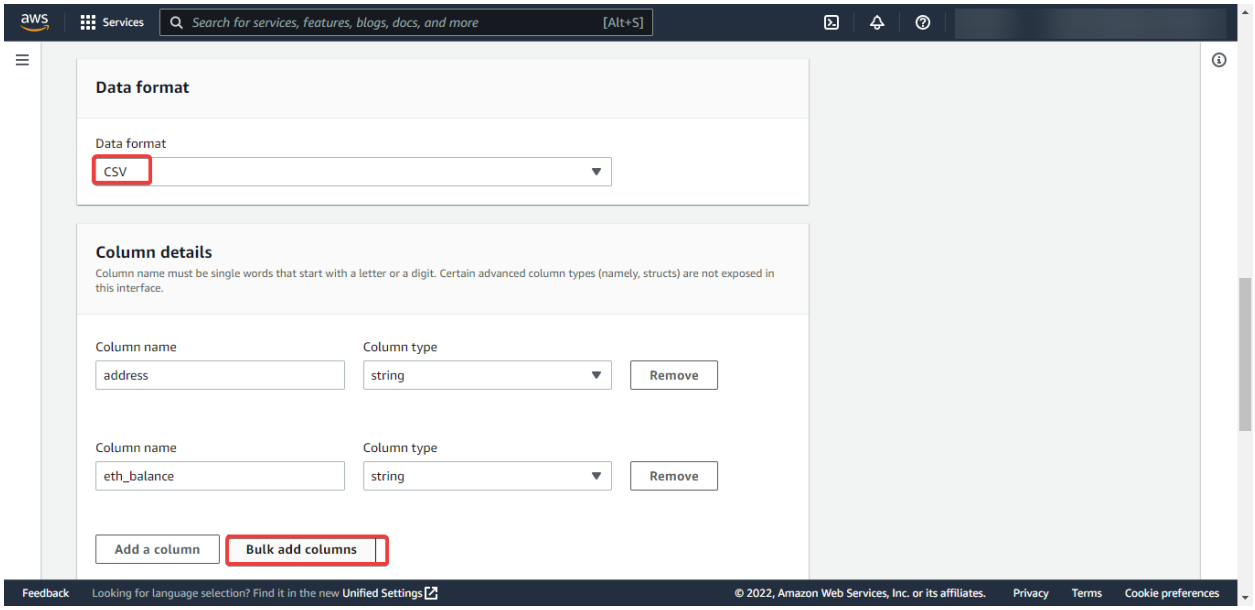
Create a **Table name : ethereum**



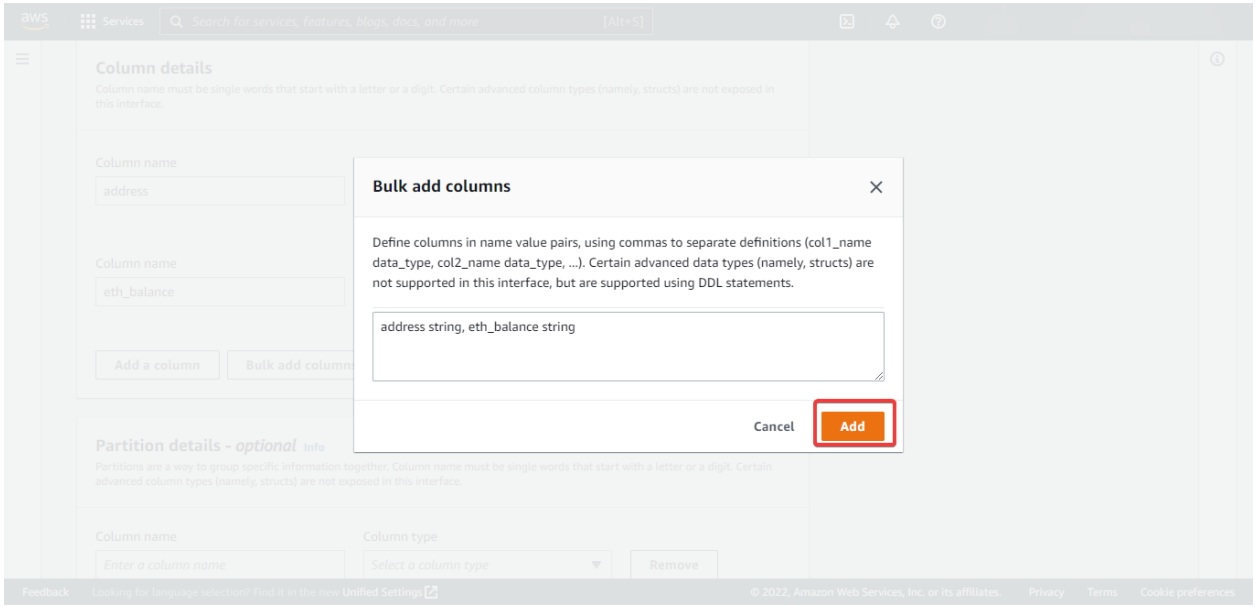
- Create **Database name**: ethereum_dataset and
- Input s3 dataset location s3://ethvclassic/ethereum_collection/



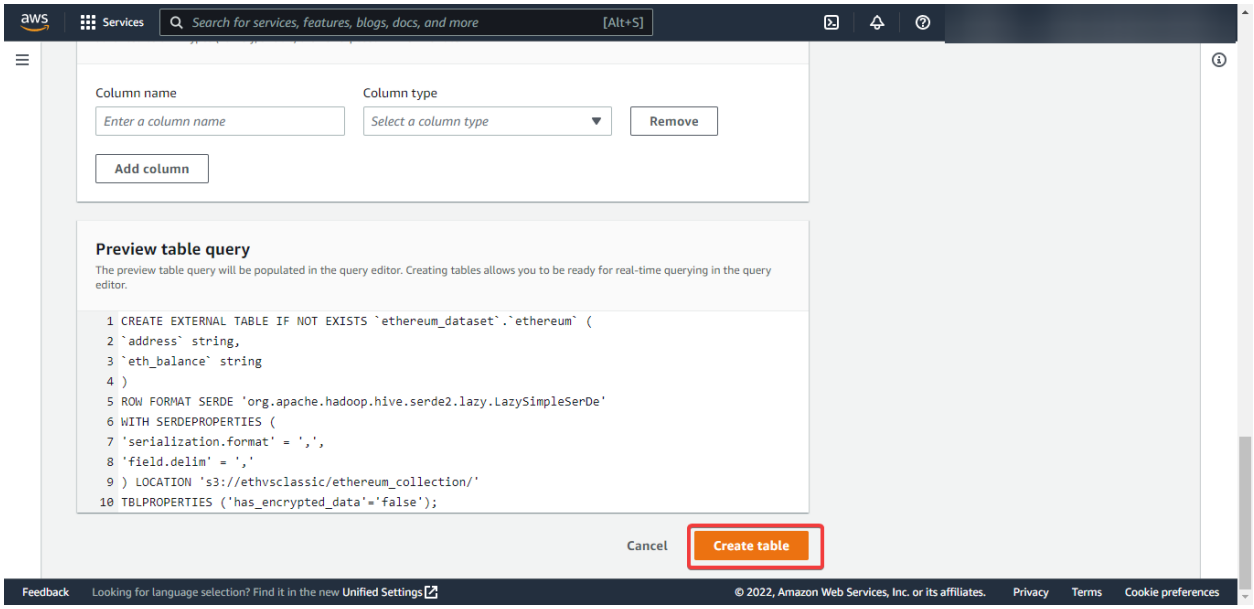
- Add **Data format** as CSV file
- Choose Bulk **add columns**



- **Add** Bulk add columns : address string, eth_balance string

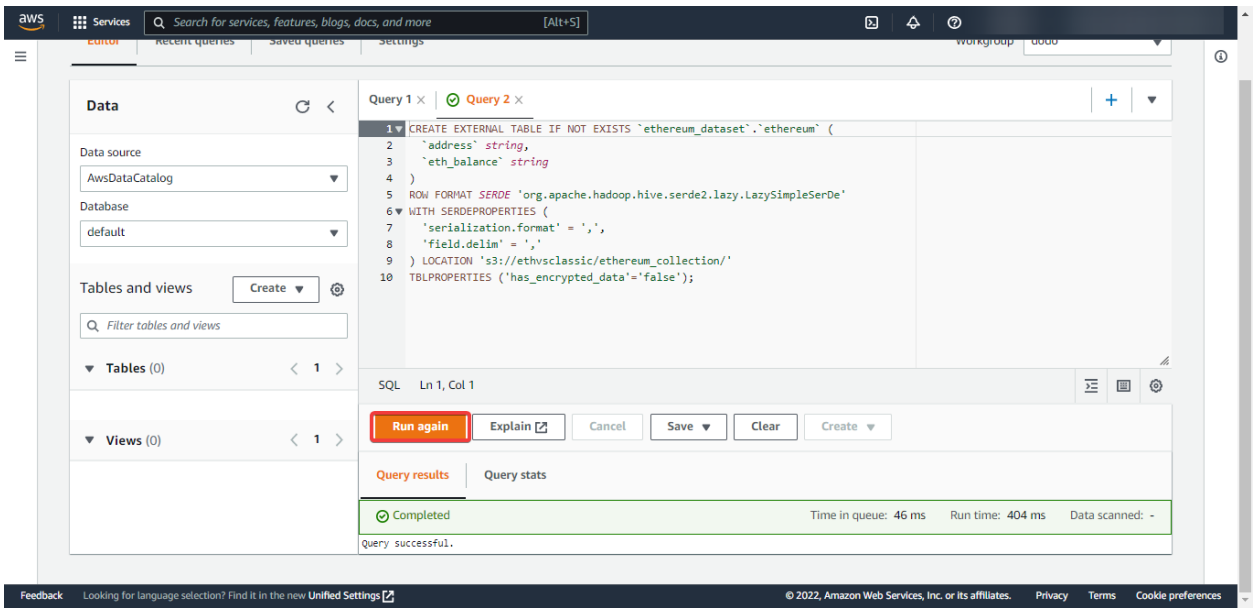


- Once completed **Create table**



Now Query will successfully run if there is already a workplace.

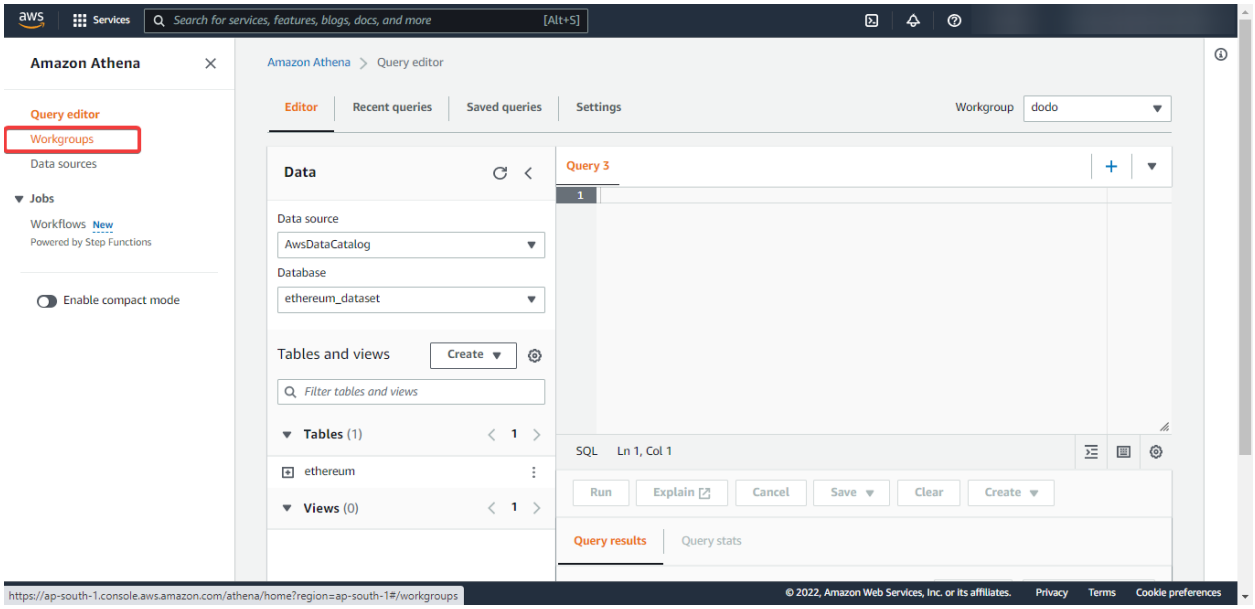
Note: If you have workplace issue setting follow below command.



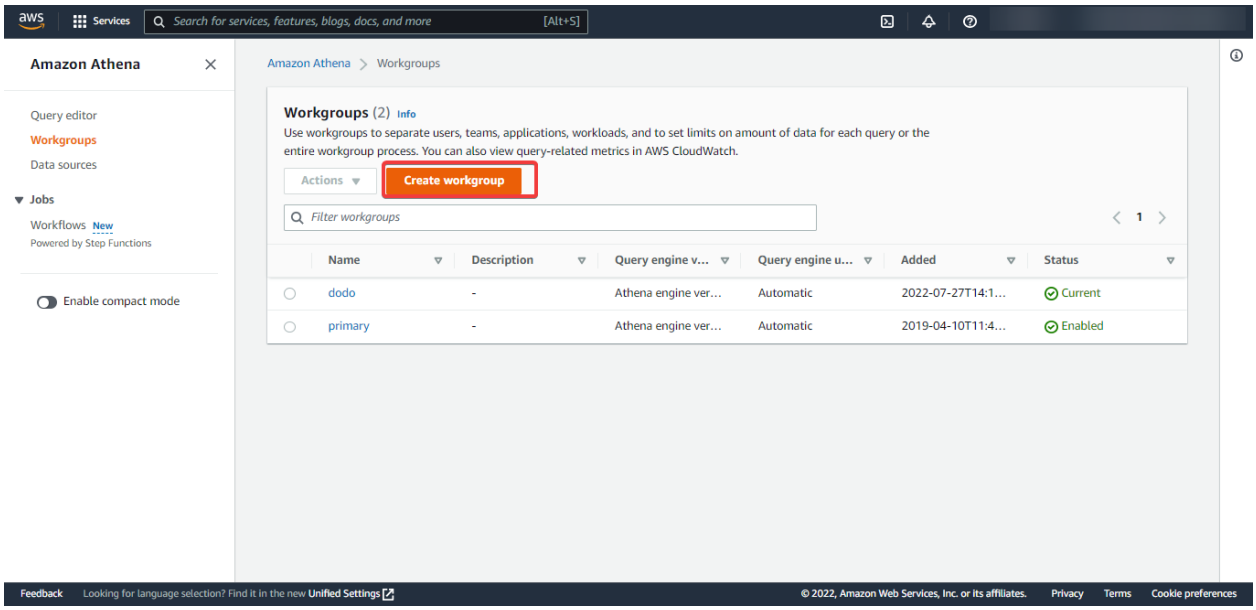
- Before running query, you have to create workgroup to save your output in separate bucket.

Workgroups Creation

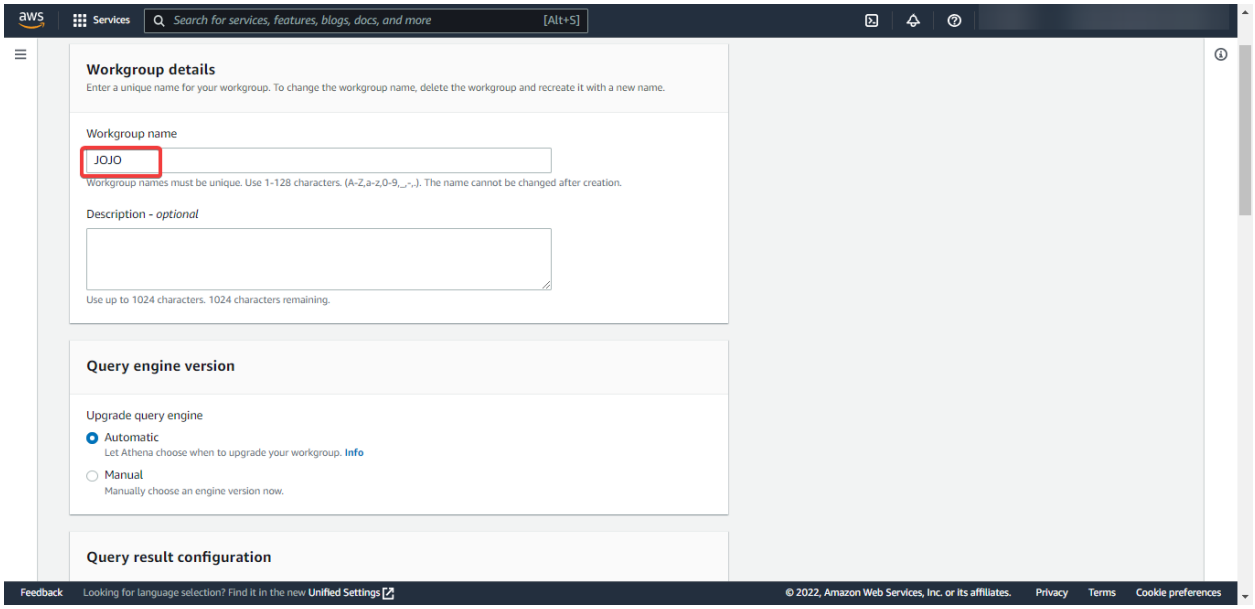
- Go to **workgroups** on left side of the **Amazon Athena** page.



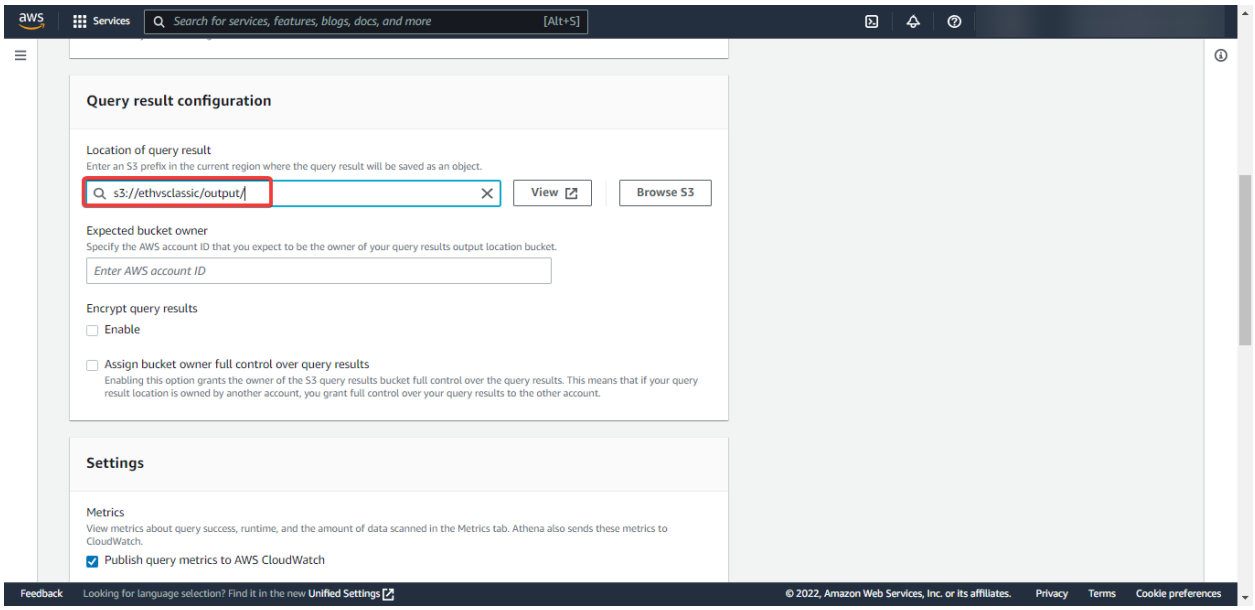
- Click **create workgroup**



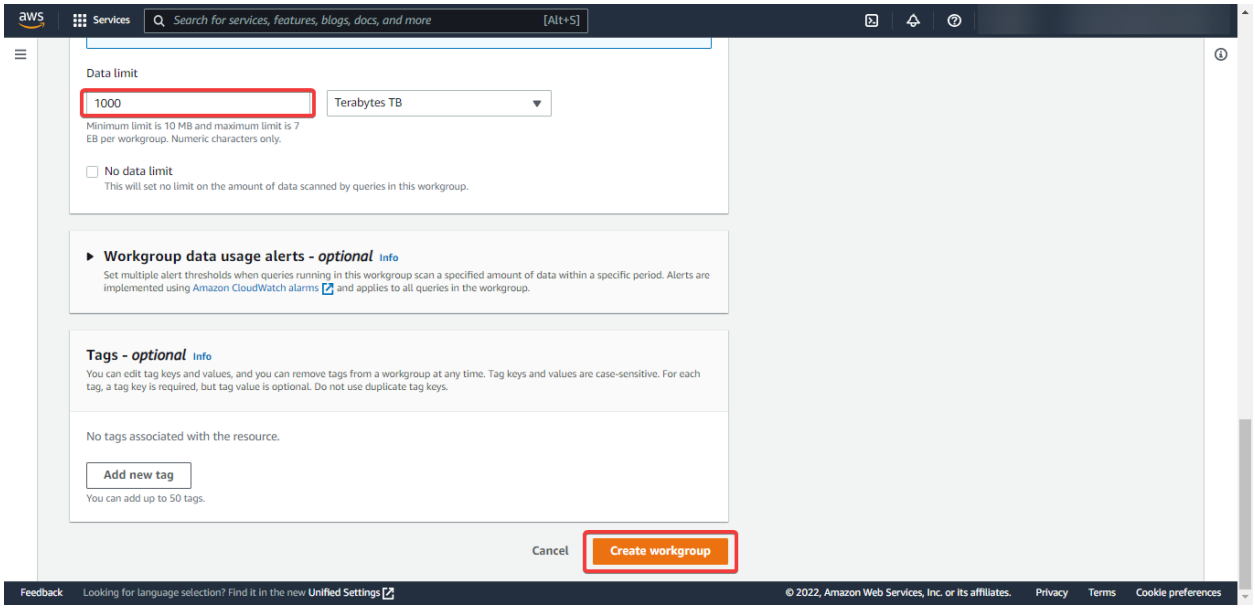
- Give a **Workgroup name** JOJO



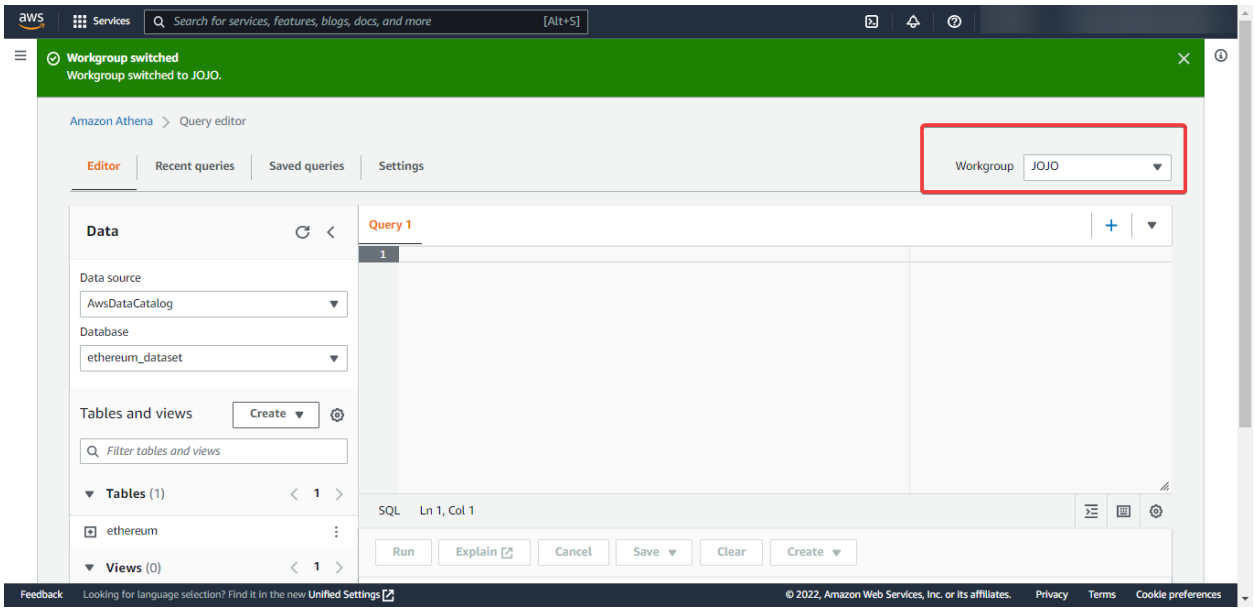
- Create a folder in S3 bucket where you want to save your output.
- Give s3 location to store query result **s3://ethvclassic/output/**



- Limit **Data limit** 1000 Terabytes TB
- Once completed click **Create workgroup.**

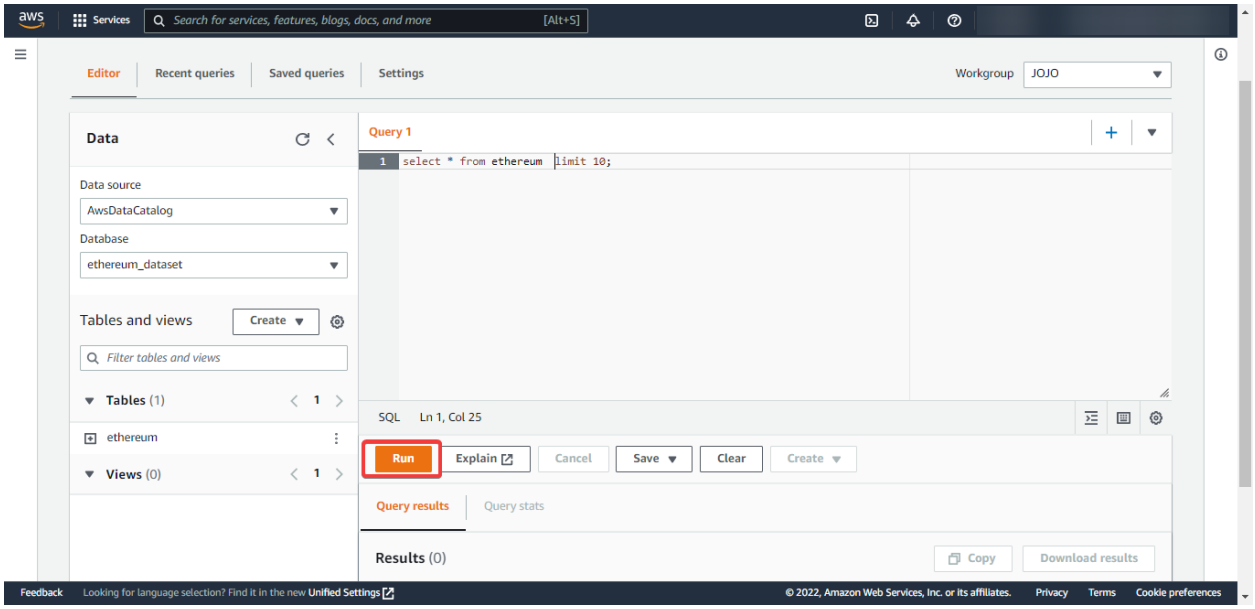


- Change your **workgroup** into JOJO on the right side of the Amazon Athena Query editor dashboard.

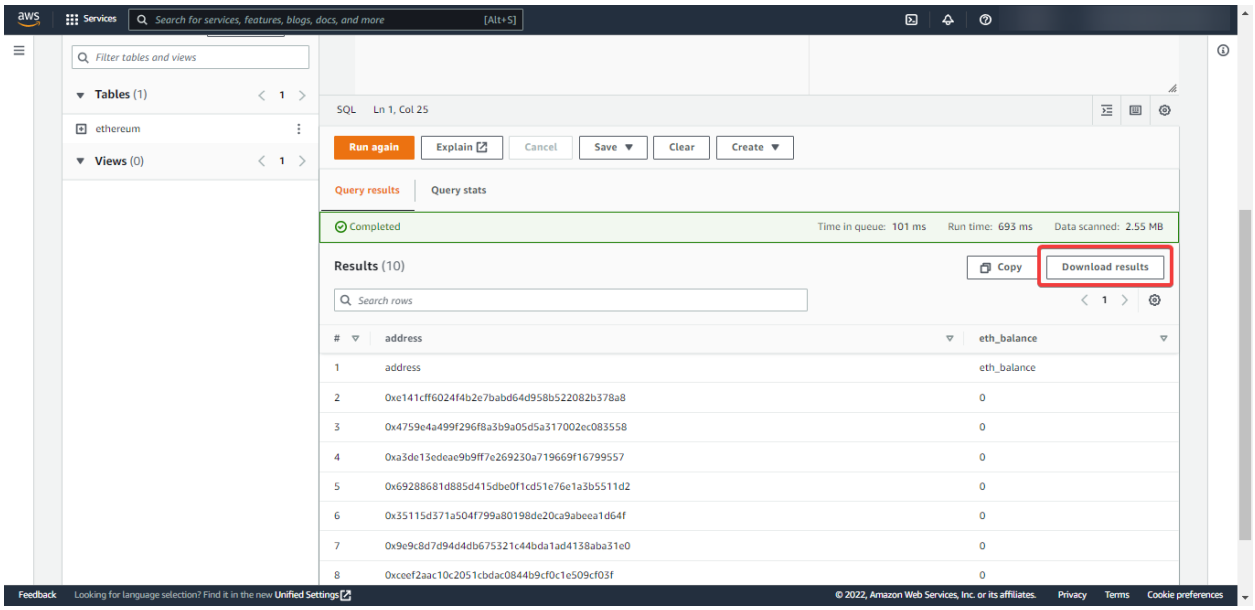


Command SQL query

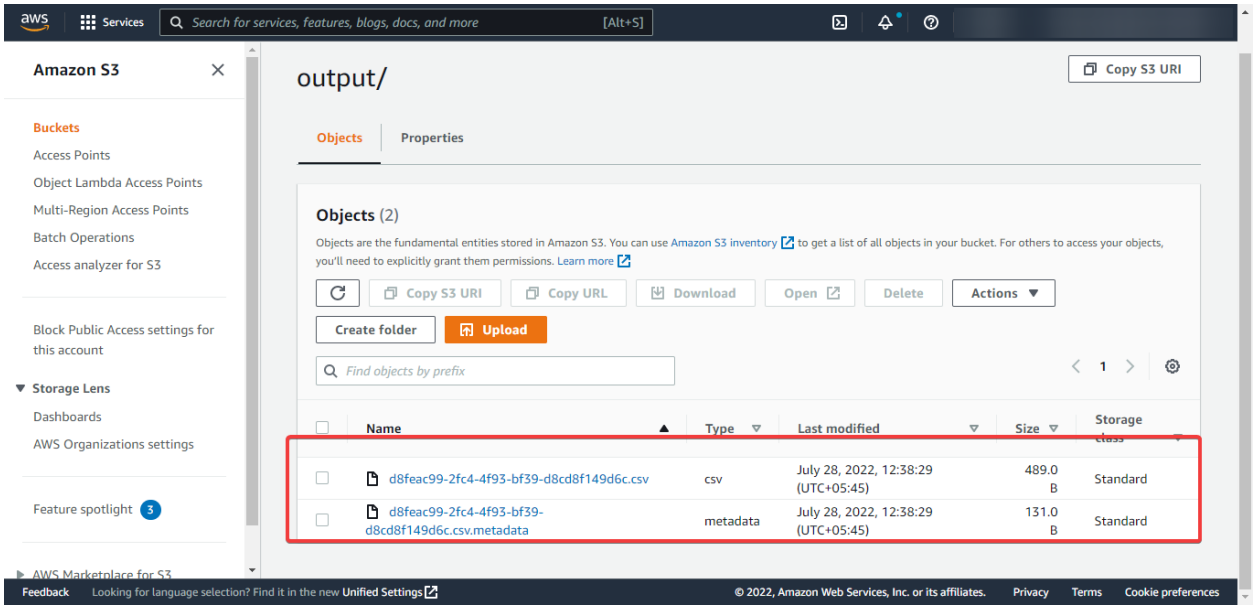
- Run the following command on query editor
  select * from ethereum limit 10;



- Query results are run successfully.
- Click **Download results**.



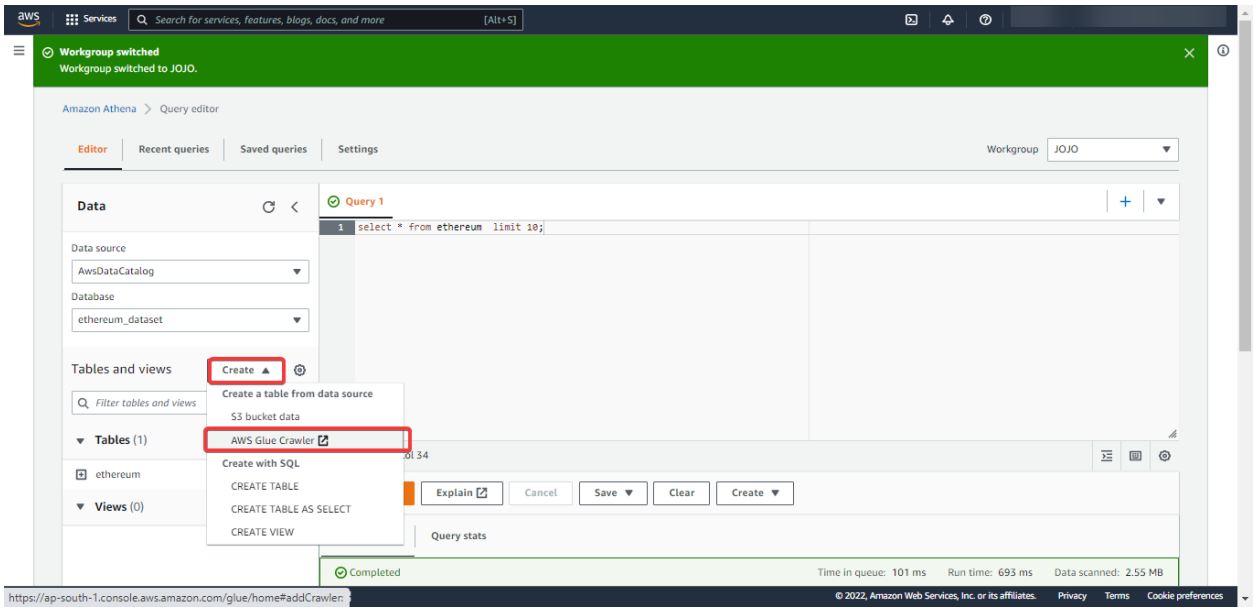
- Similarly, result is also stored automatically in S3 bucket folder.
  s3://ethvclassic/output/


aws
Services
Search for services, features, blogs, docs, and more
[Alt+S]
Amazon S3
Buckets
Access Points
Object Lambda Access Points
Multi-Region Access Points
Batch Operations
Access analyzer for S3
Block Public Access settings for this account
Storage Lens
Dashboards
AWS Organizations settings
Feature spotlight
3
AWS Marketplace for S3
output/
Copy S3 URI
Objects
Properties
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
1
Name
Type
Last modified
Size
Storage class
d8feac99-2fc4-4f93-bf39-d8cd8f149d6c.csv
csv
July 28, 2022, 12:38:29 (UTC+05:45)
489.0 B
Standard
d8feac99-2fc4-4f93-bf39-d8cd8f149d6c.csv.metadata
metadata
July 28, 2022, 12:38:29 (UTC+05:45)
131.0 B
Standard
Feedback
Looking for language selection? Find it in the new Unified Settings
© 2022, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences

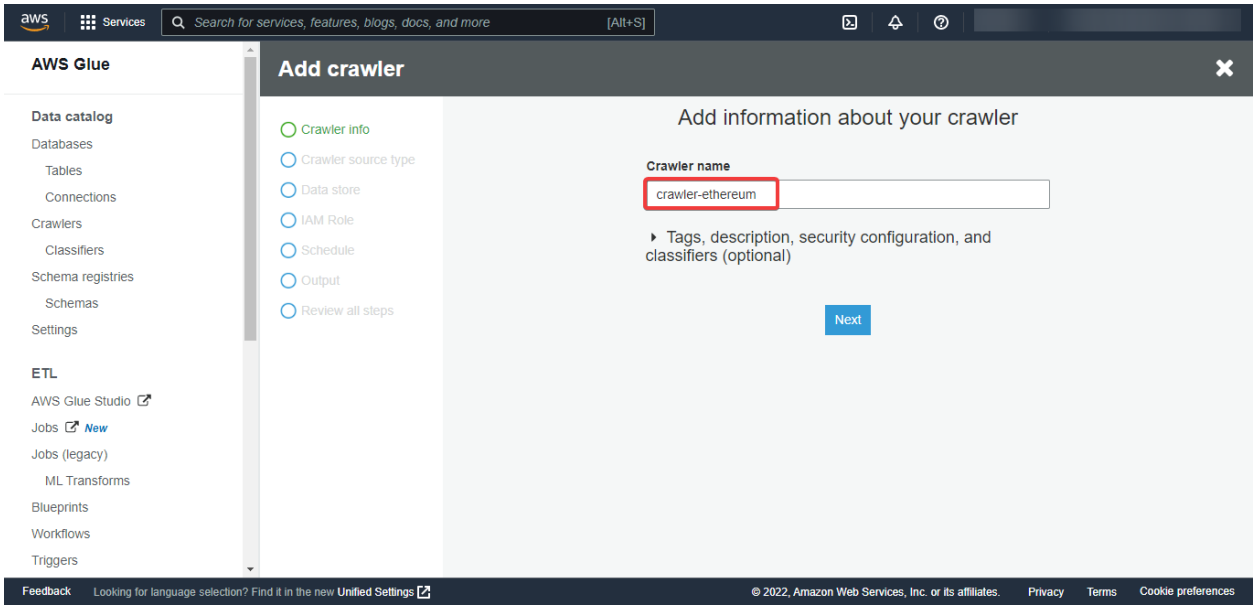
Query data on Amazon Athena from AWS Glue crawler

Create AWS Glue crawler

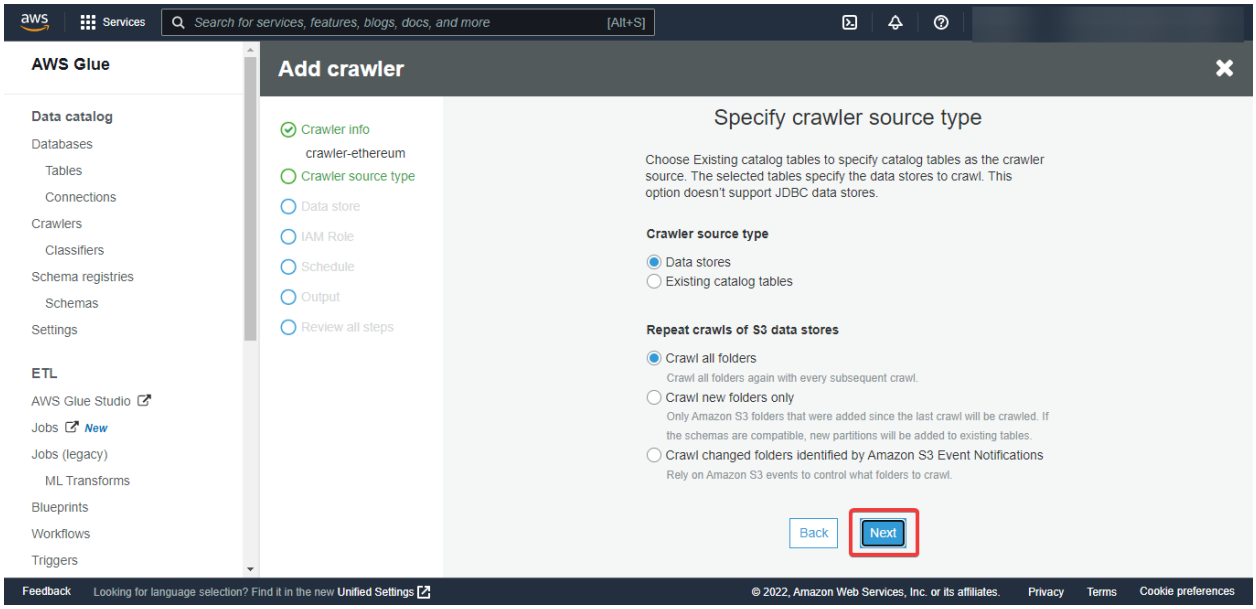
- Go to **Amazon Athena** query editor dashboard
- Click **Create** and select **AWS Glue Crawler**.



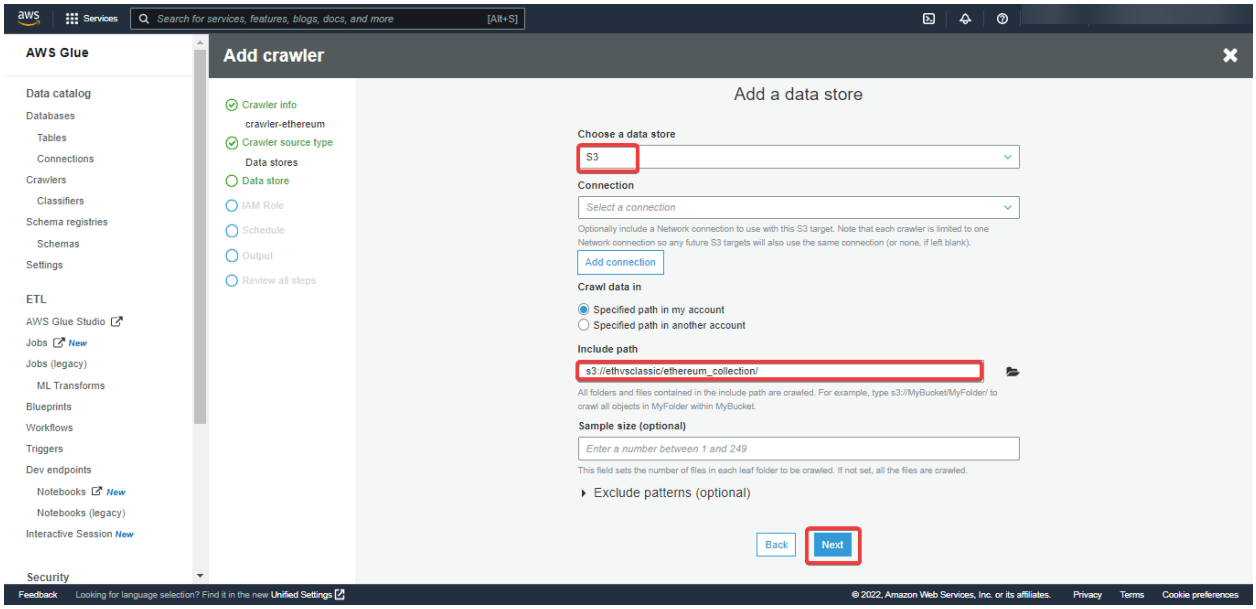
- **Add crawler**
- Create **crawler name** Crawler-ethereum



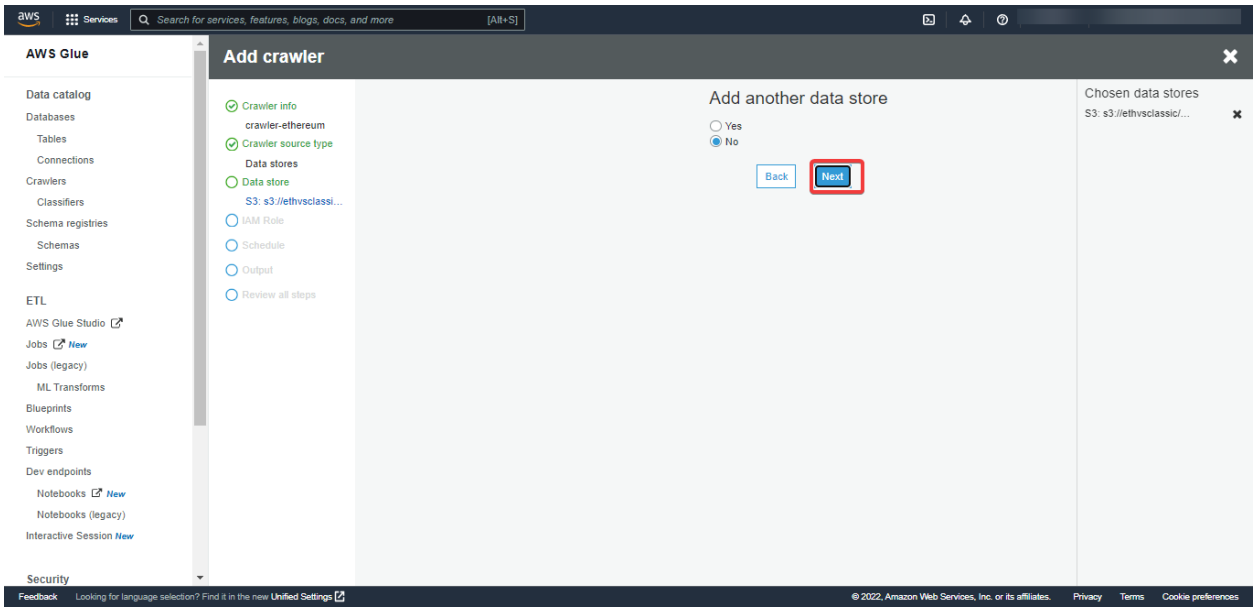
Click Next with default option.



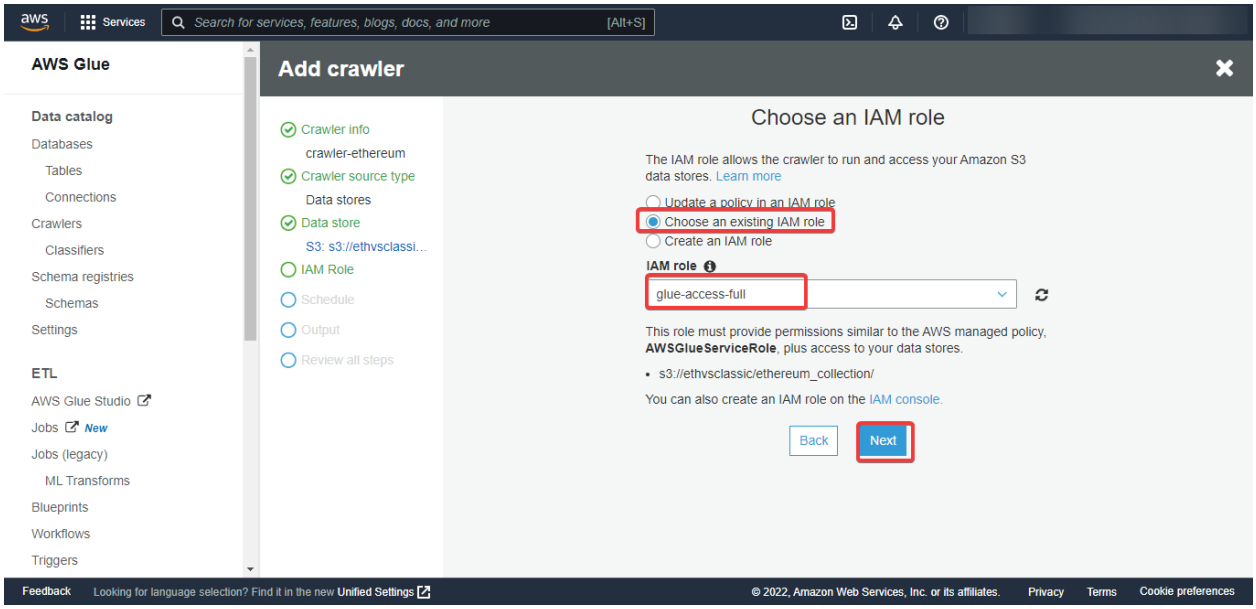
- **Choose a data store** as S3.
- **Include path** s3 dataset location s3://ethvsclassic/ethereum_collection/
- Click **Next**.



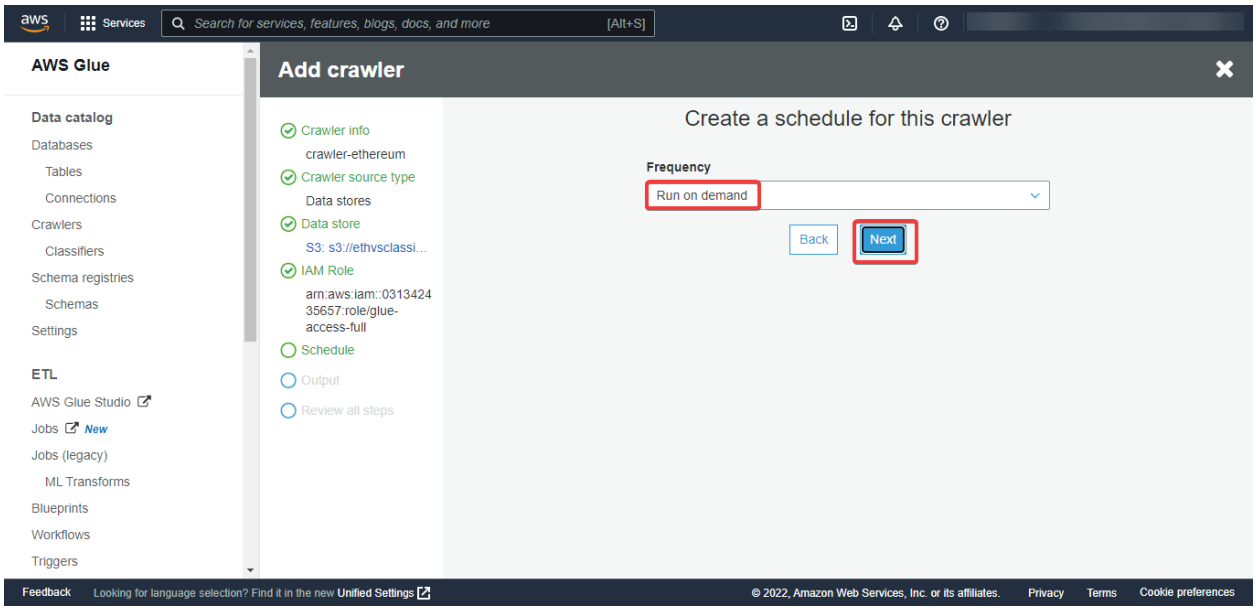
- **Add another data store** No
- **Click Next**



- **Choose an existing IAM role** that access glue. **See bottom page about creating IAM role.**
- Click **Next.**



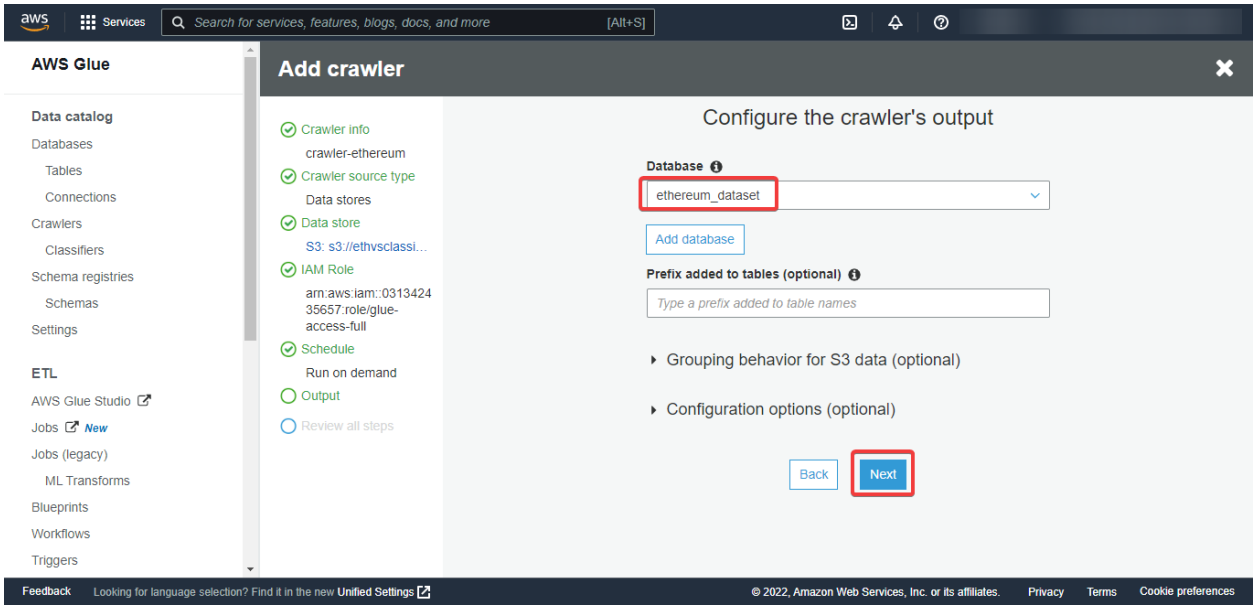
- Set Frequency as **Run-on demand.**
- Click **Next.**



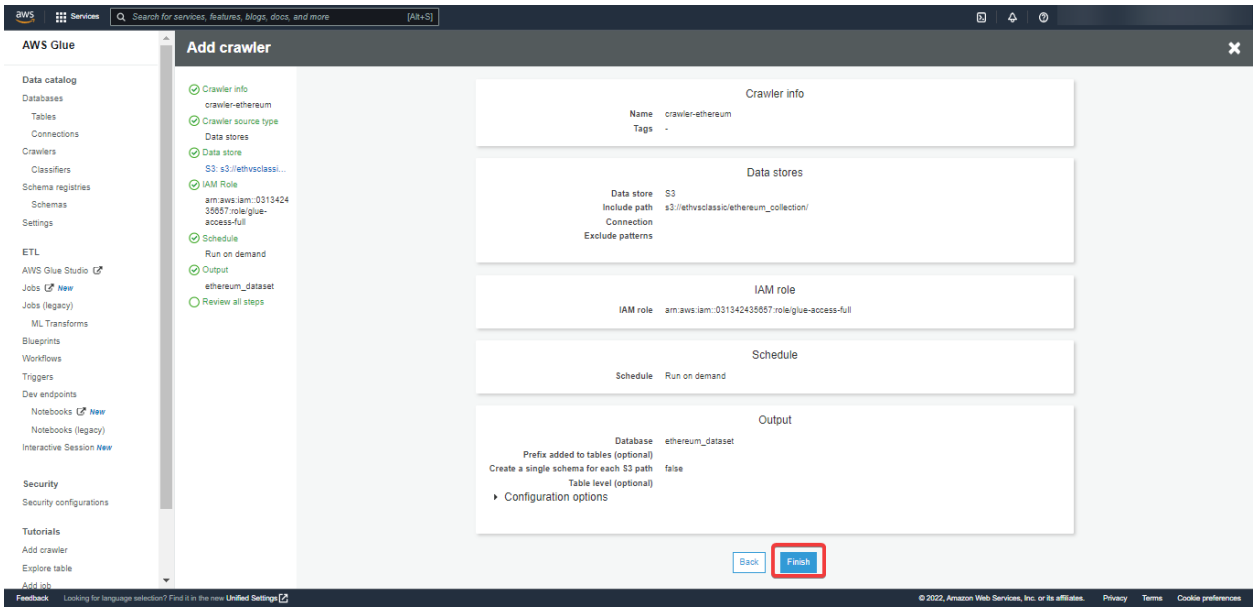
Choose exiting **Database** as ethereum_dataset(s3 bucket is already attached on aws Glue).

Or create a new Database on **Databases name** on AWS Glue **Data catalog.**

Click **Next.**

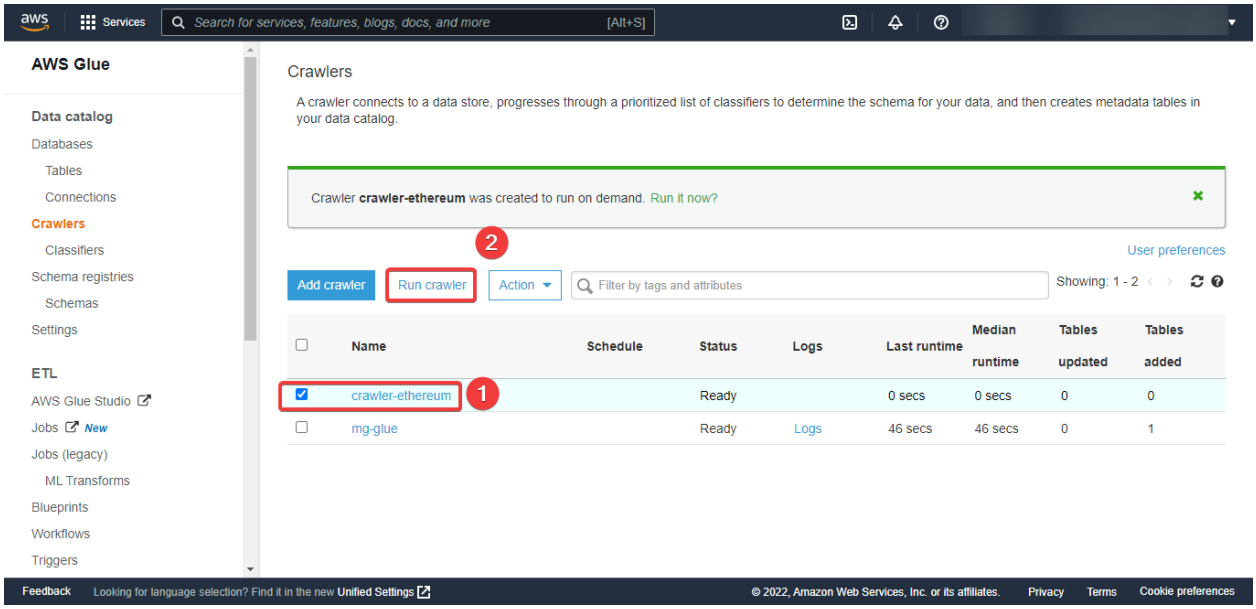


- Once you review all, click **Finish.**

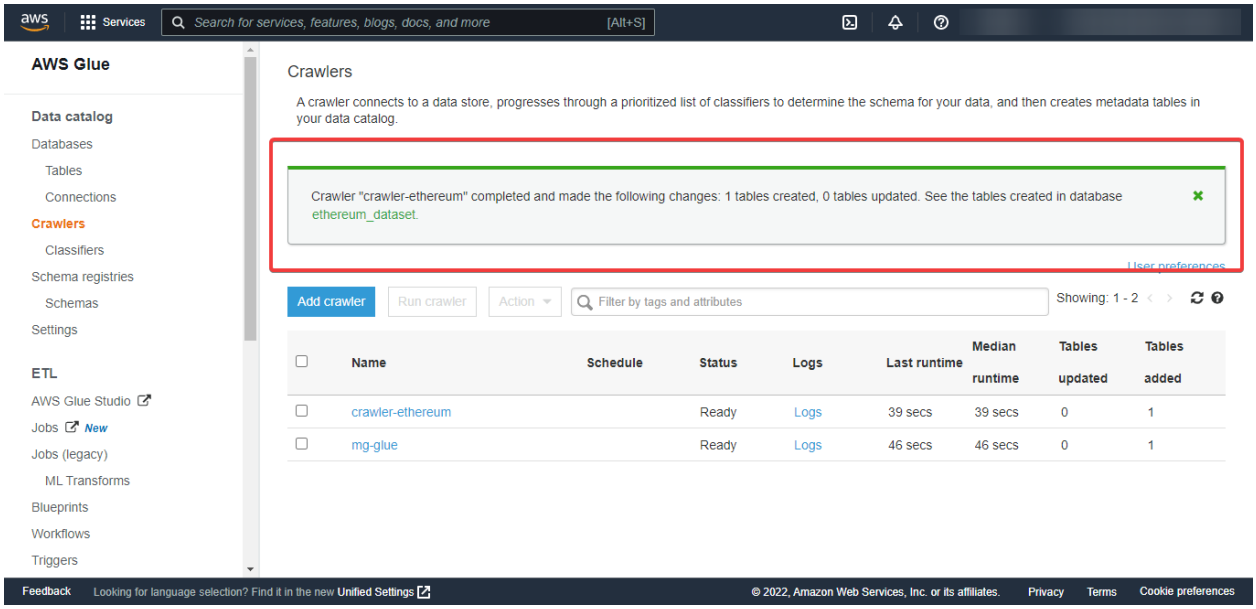


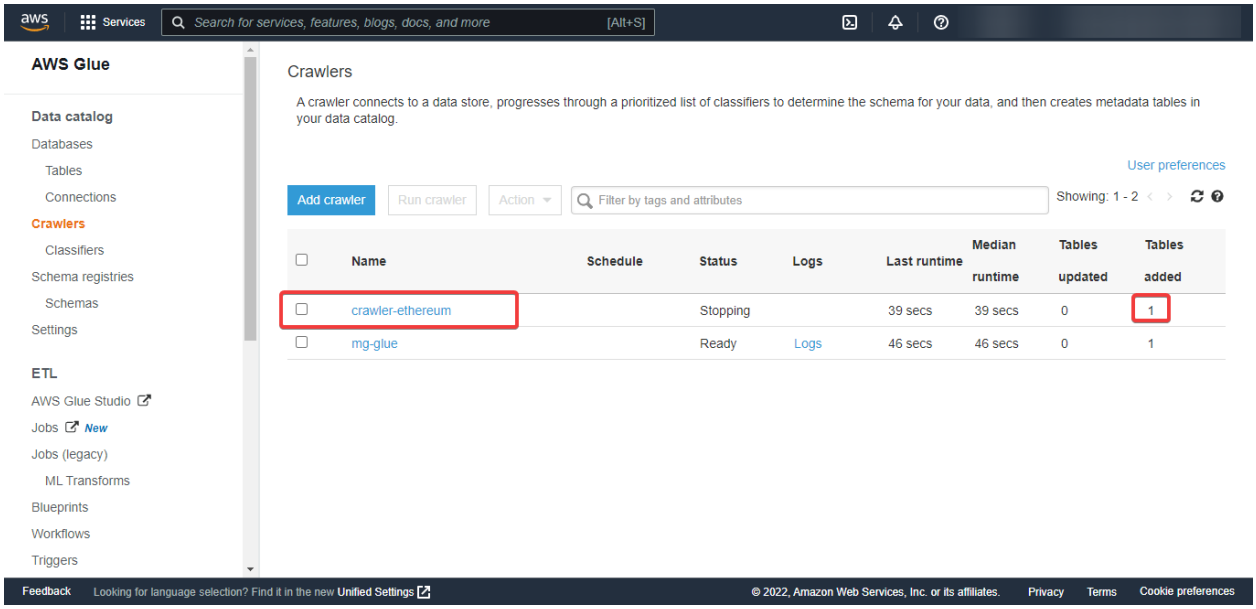
Run Crawler

- Select your **crawler** crawler-ethereum and click **Run crawler**.

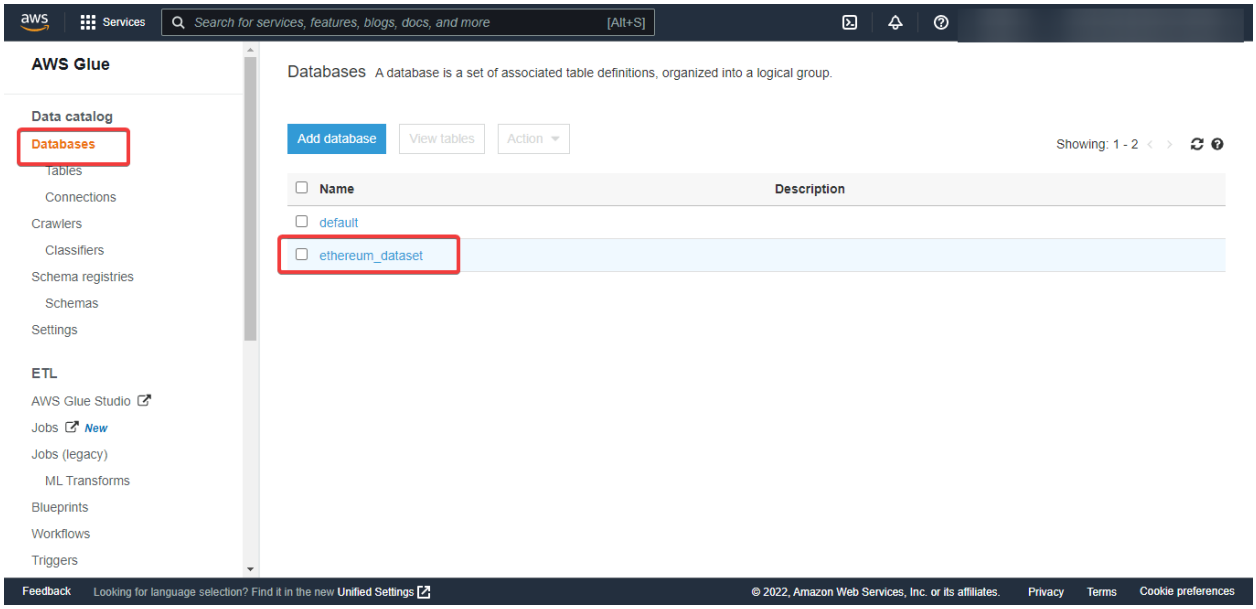


- Once a crawler is successful. A Table is added.

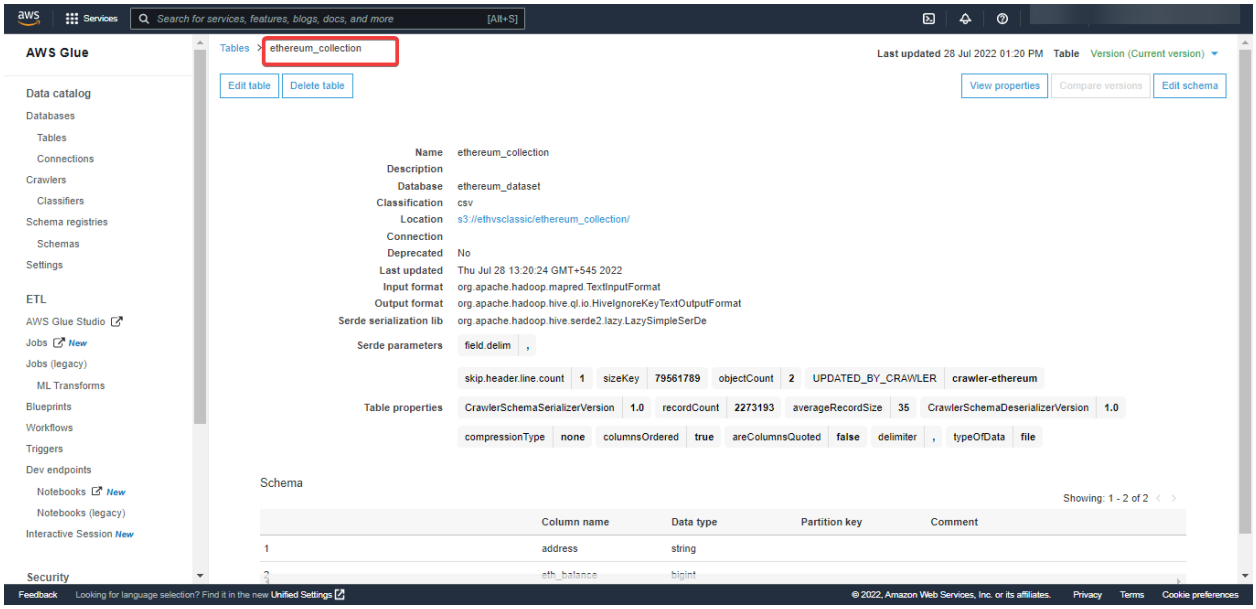




- Go to **databases** below AWS Glue Data catalog.
- Open your database folder ethereum_dataset

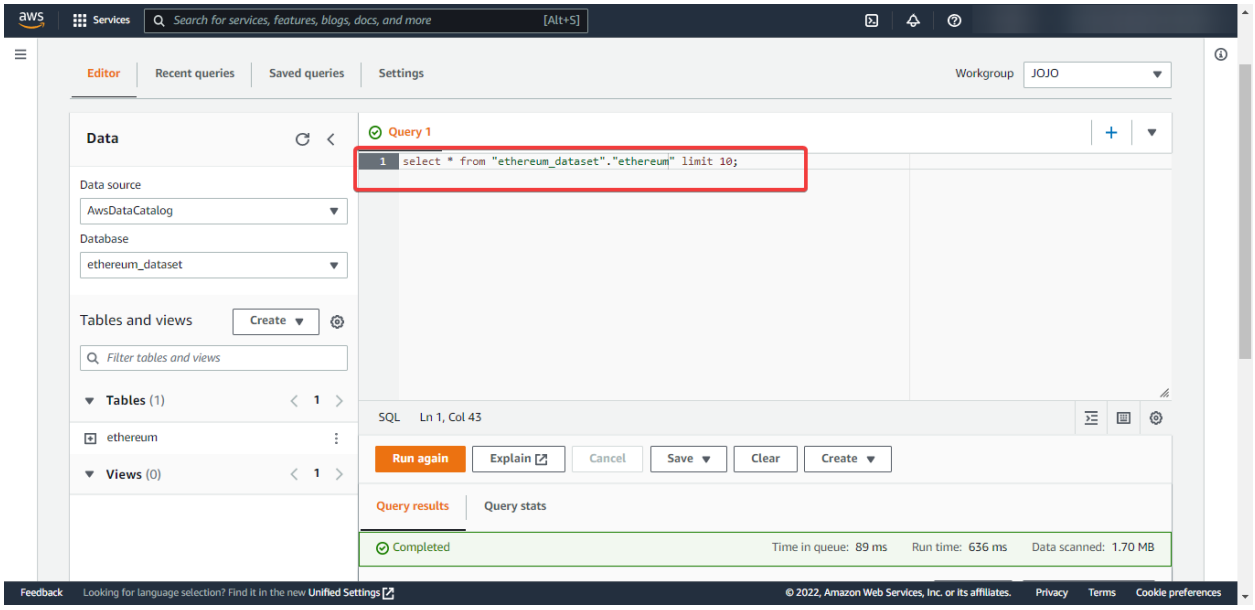


You will see your database collection.

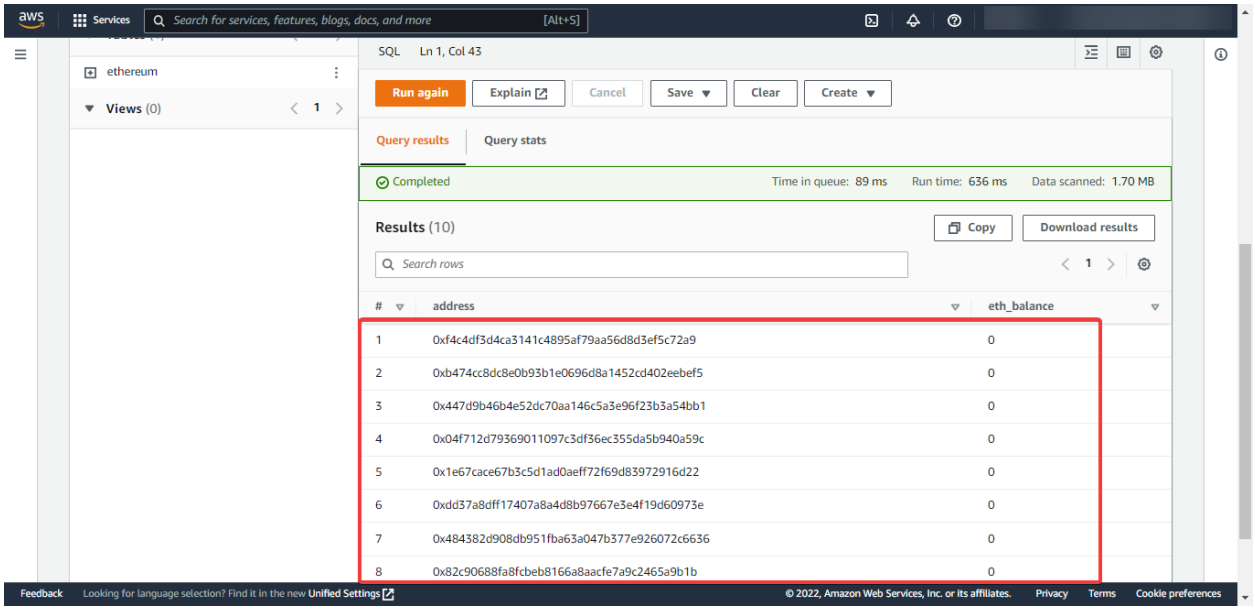


SQL query on Athena Editor

- Go to **AWS Athena query** dashboard and run given query



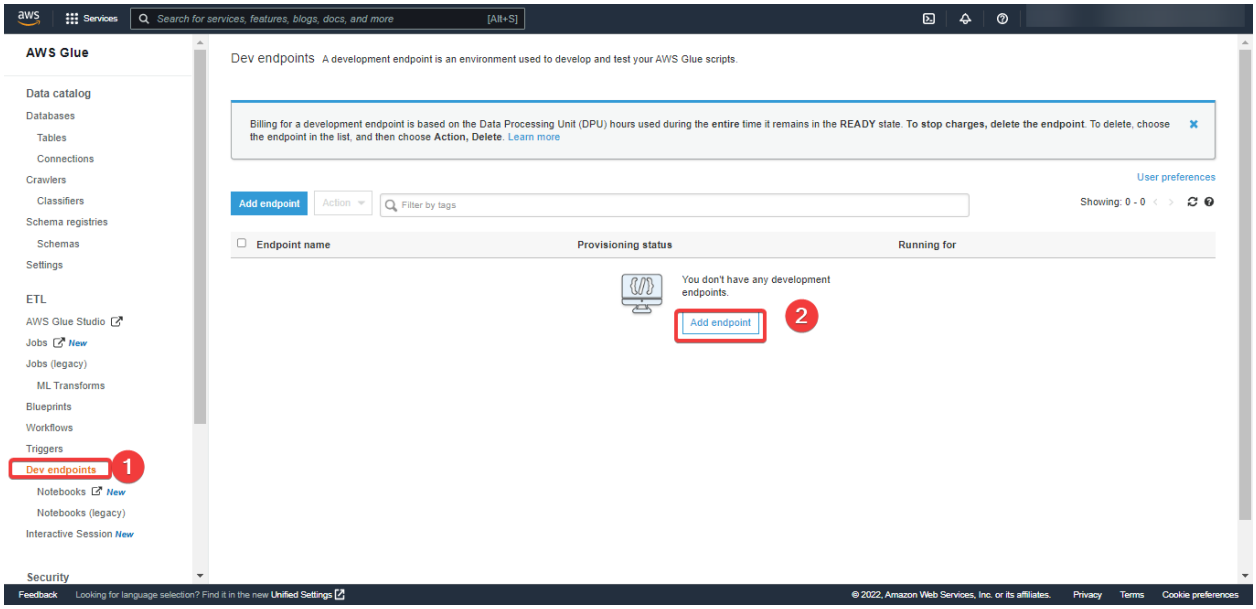
- The Result is shown below.



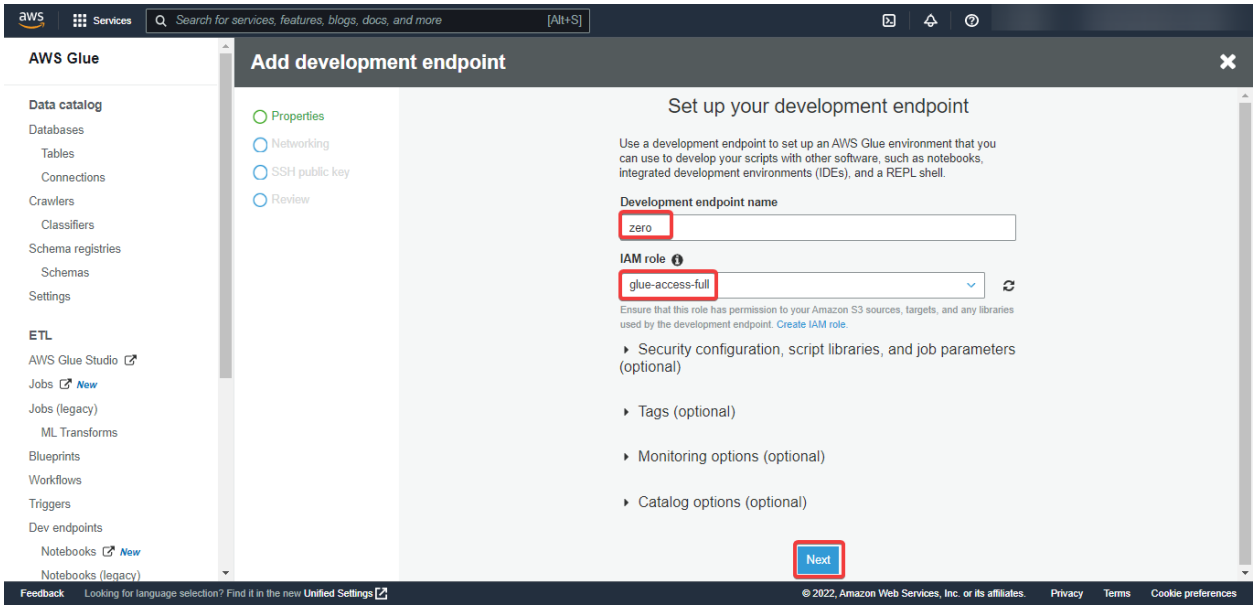
Run Notebook on AWS Glue

Add Dev Endpoints

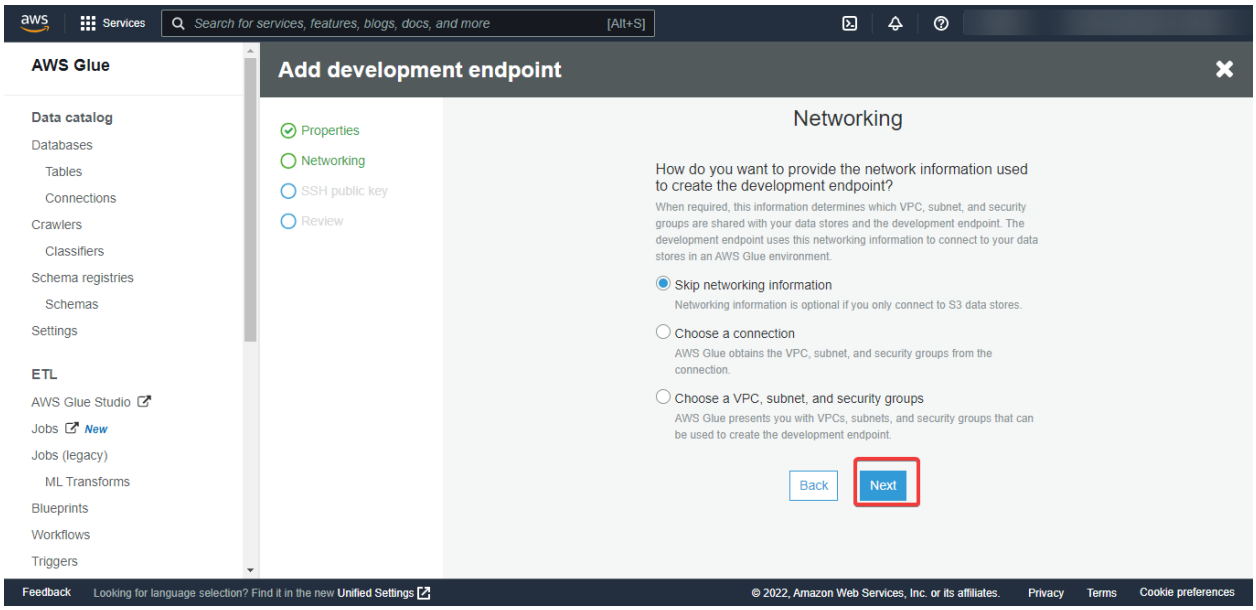
- Go to **AWS Glue**.
- Select **Dev endpoints** on left side of Data catalog
- **Add endpoint**



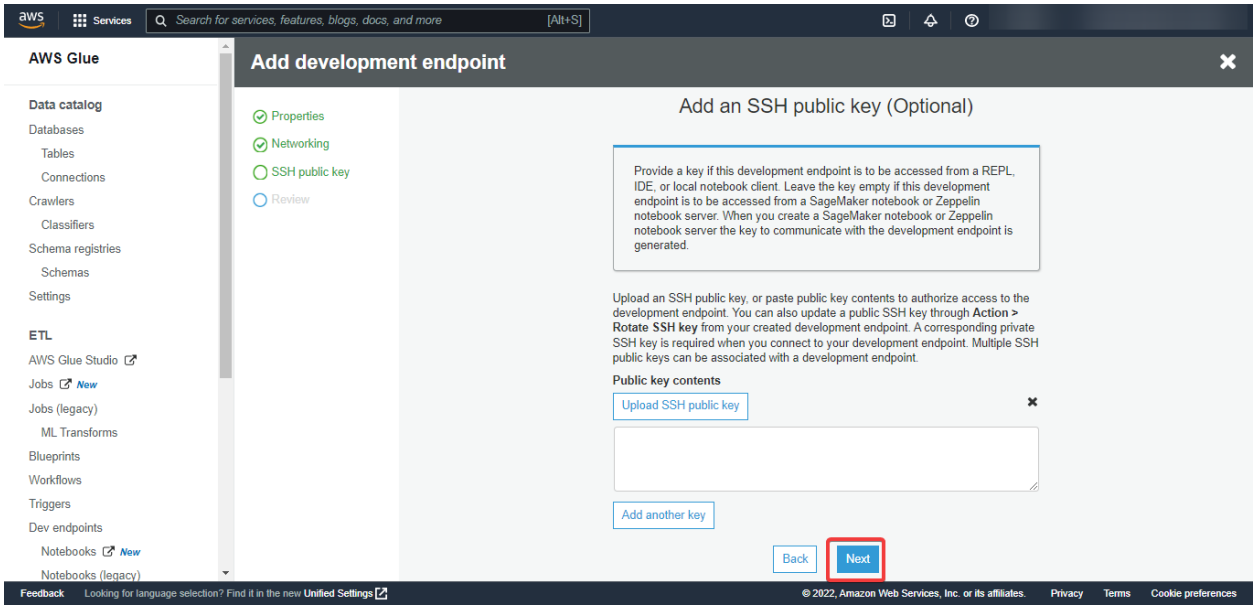
- Name **Development endpoint name** zero
- Create or choose **IAM role** that access Glue : glue-access-full
- Click **Next.**



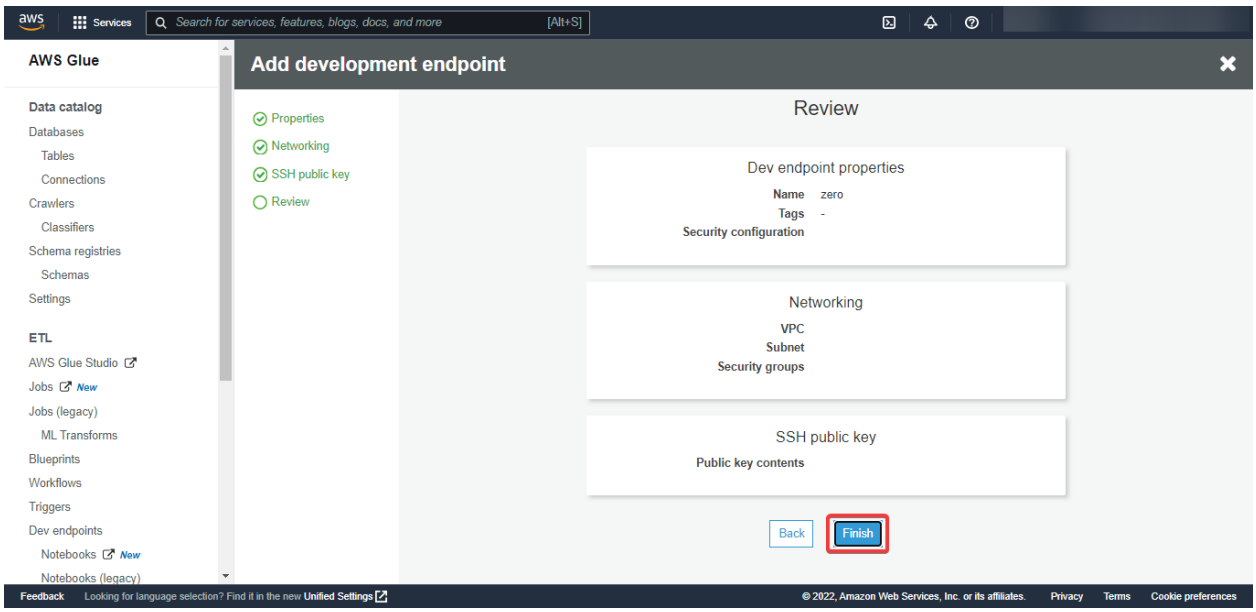
- Click **Next.**

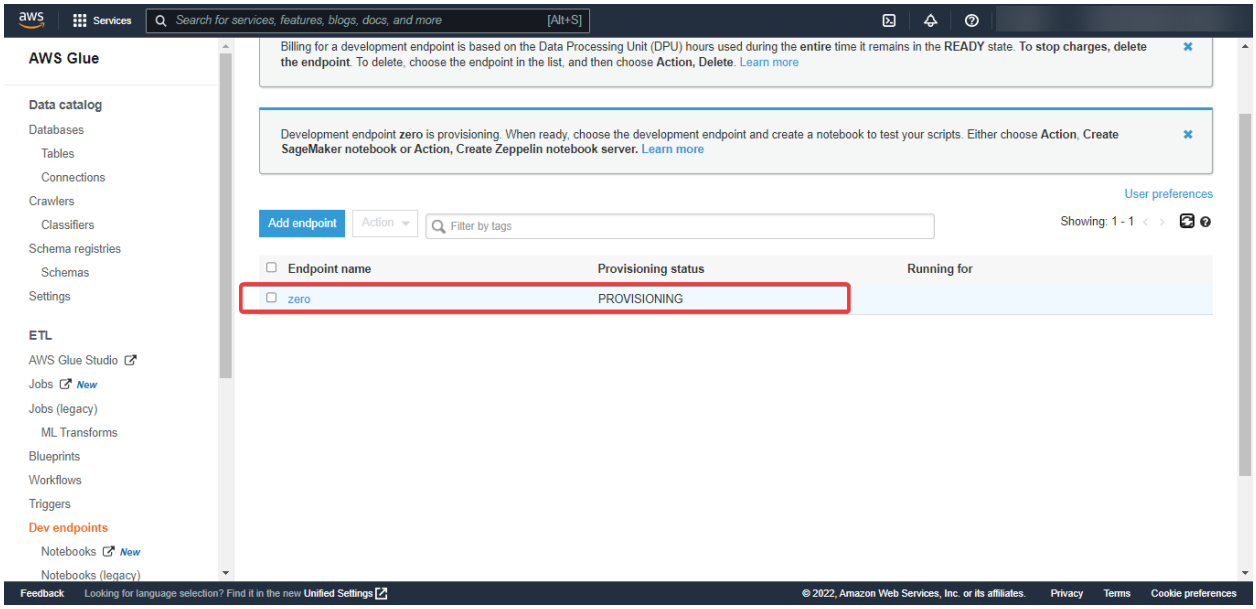


- Click **Next.**

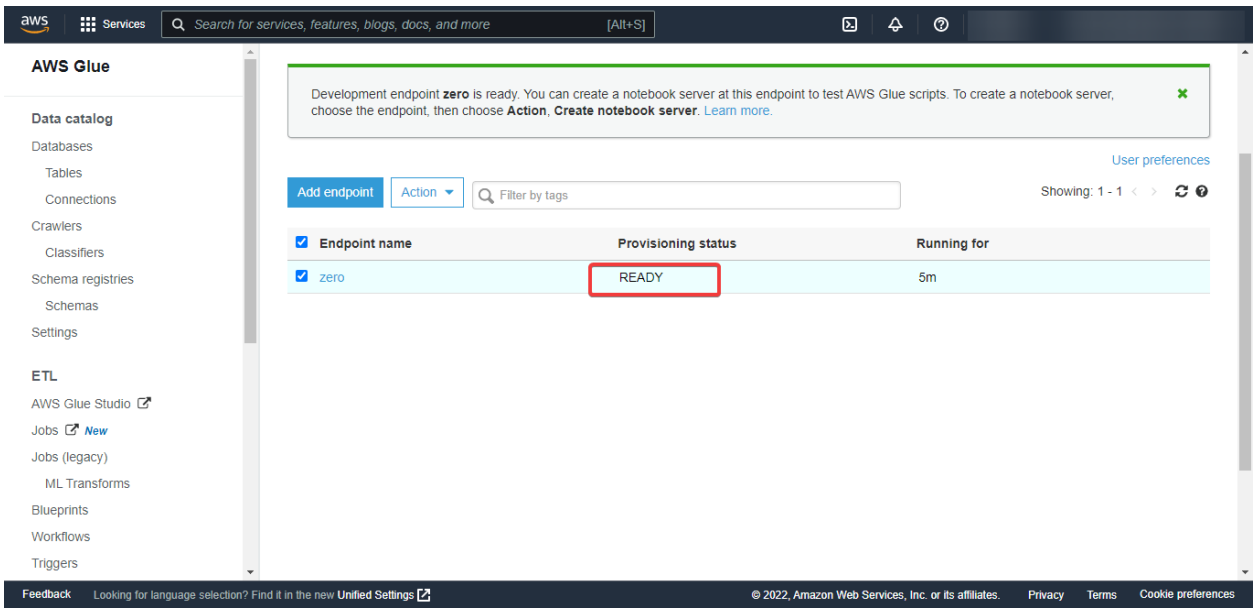


Once review, Click **Finish**.



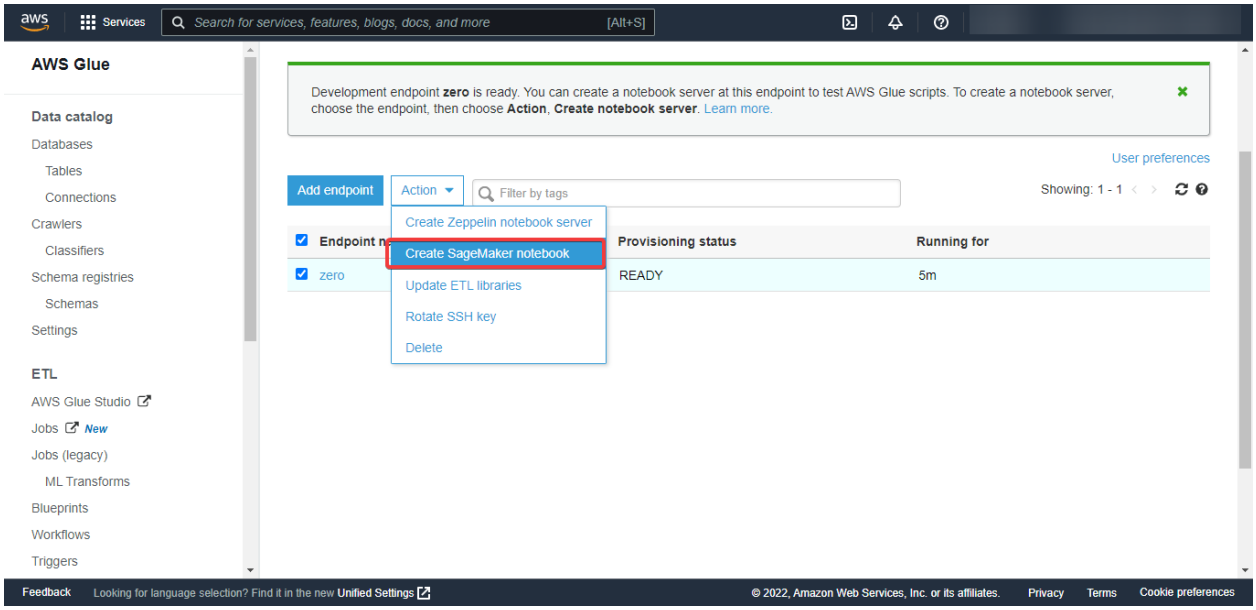


Now, your endpoint is ready.

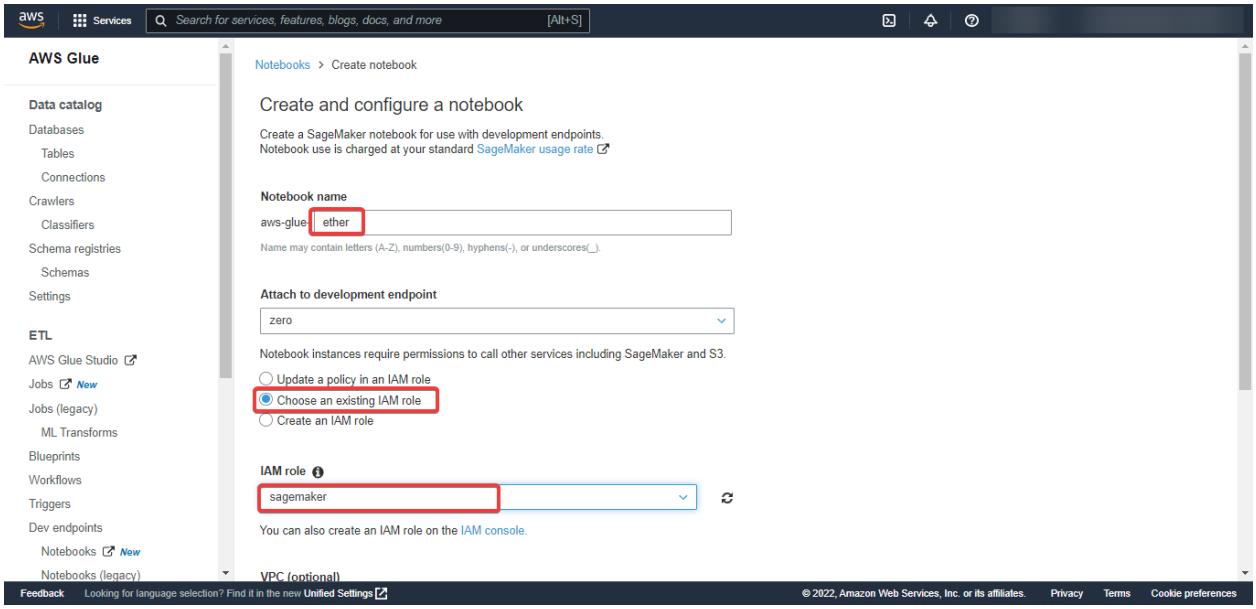


Create SageMaker notebook

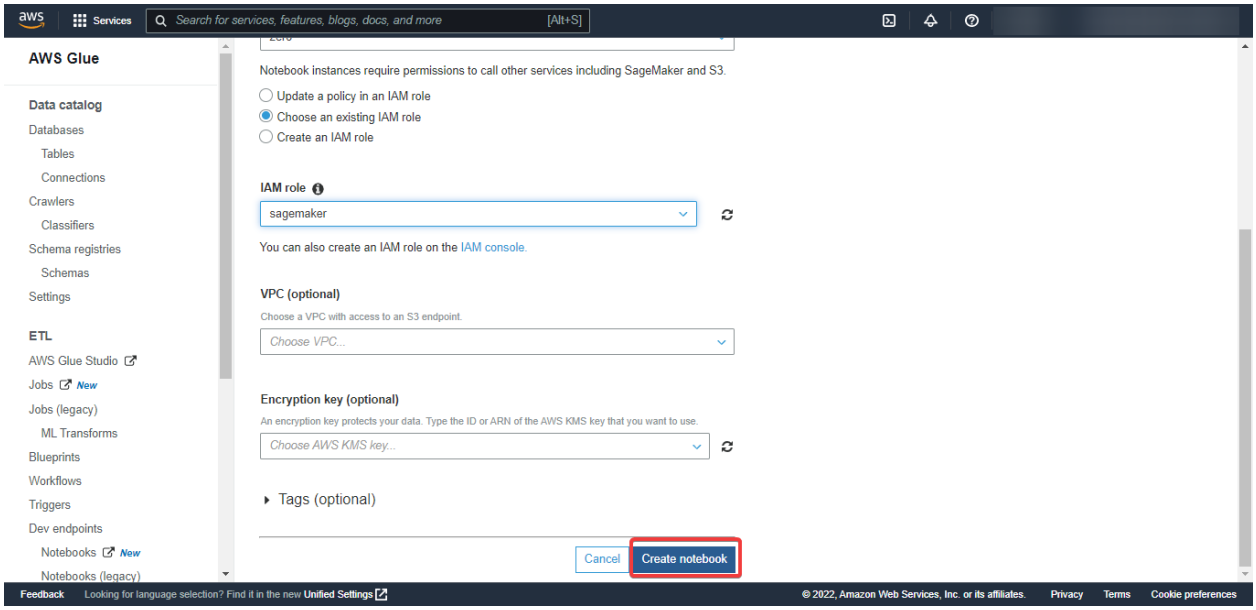
- Select your endpoint zero and click **create SageMaker notebook.**



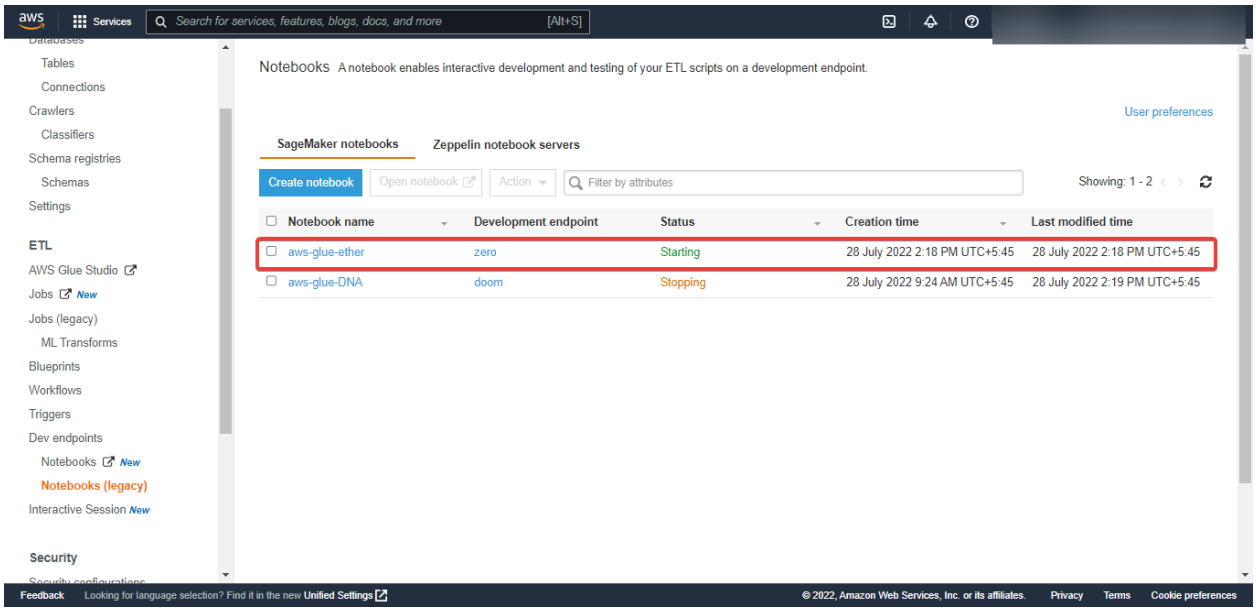
- Give Notebook name aws-glue-**ether**
- **Attach to development endpoint** zero as default we recently created.
- Choose or create IAM sagemaker role for glue. Choose an **existing IAM role.**



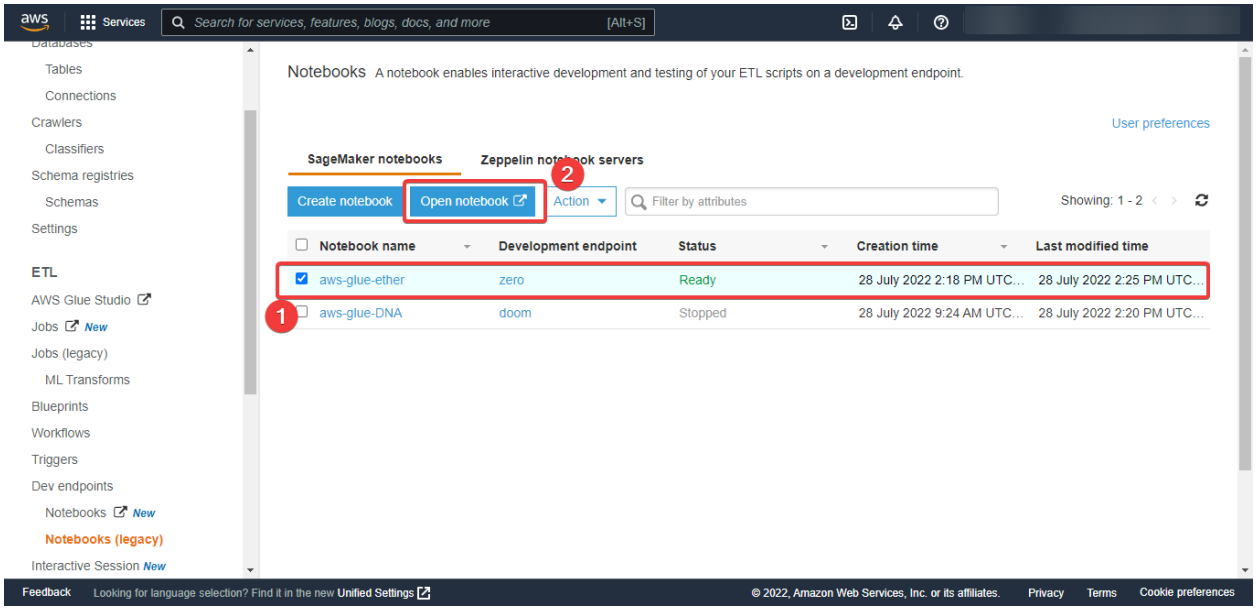
- Keep other options default
- Click **Create notebook**



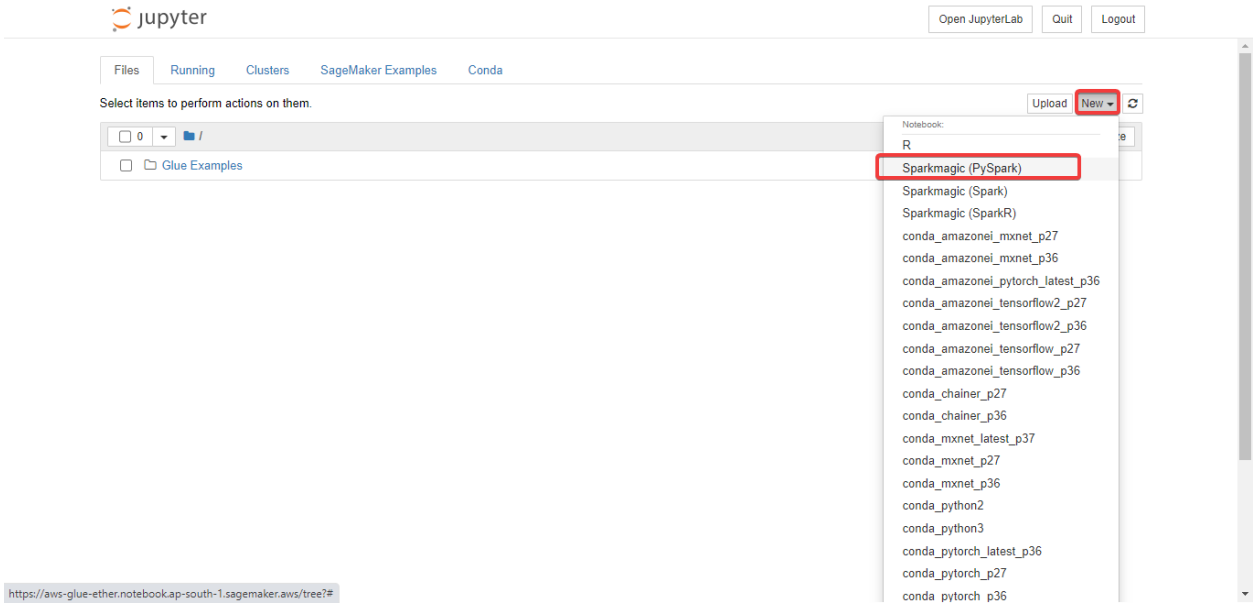
- Your aws-glue-ether notebook status is running.



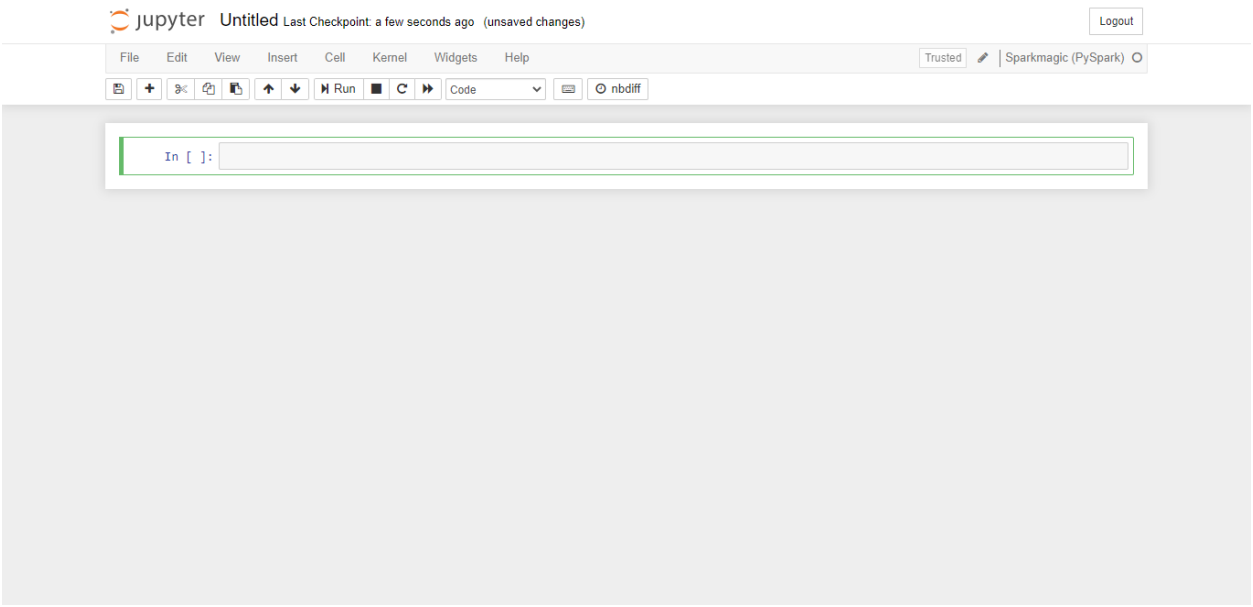
- Select your notebook aws-glue-ether and **open notebook**.



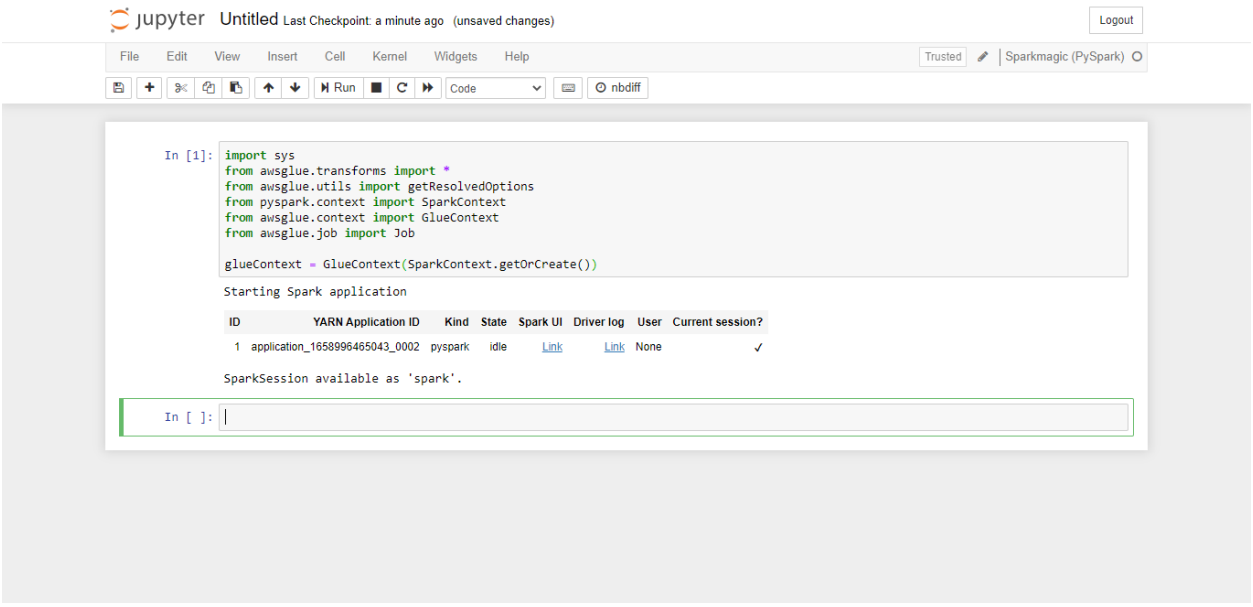
- Jupyter notebook is opened. Click **New** and Select **Pyspark**.



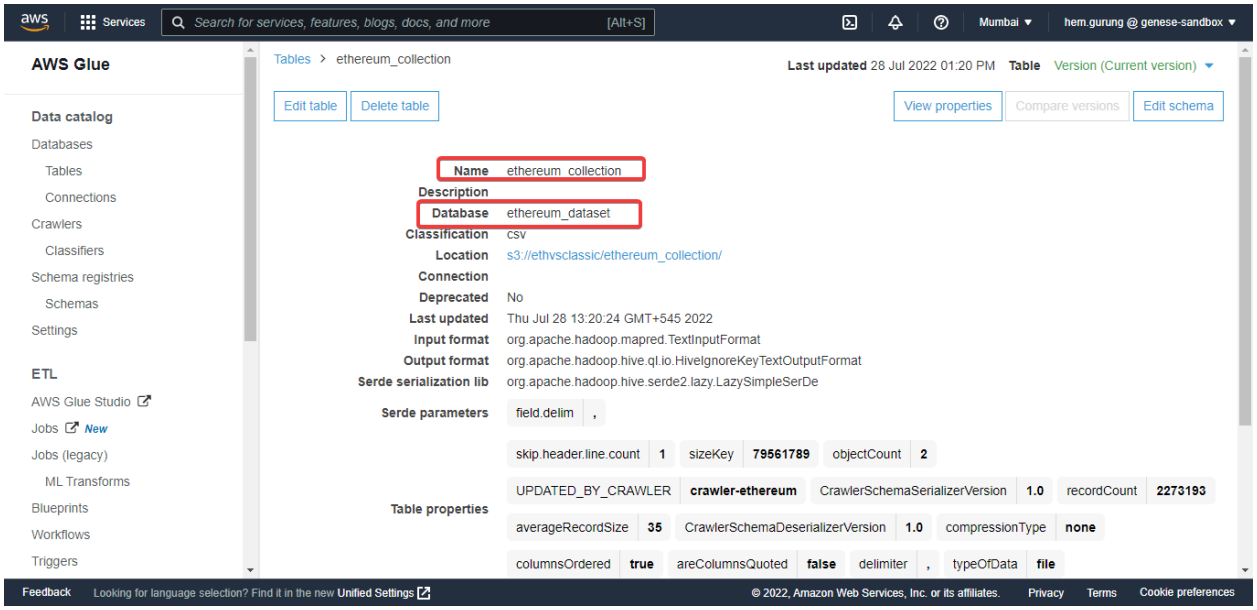
- Blank jupyter notebook will open with PySpark



Spark Session is successfully installed.



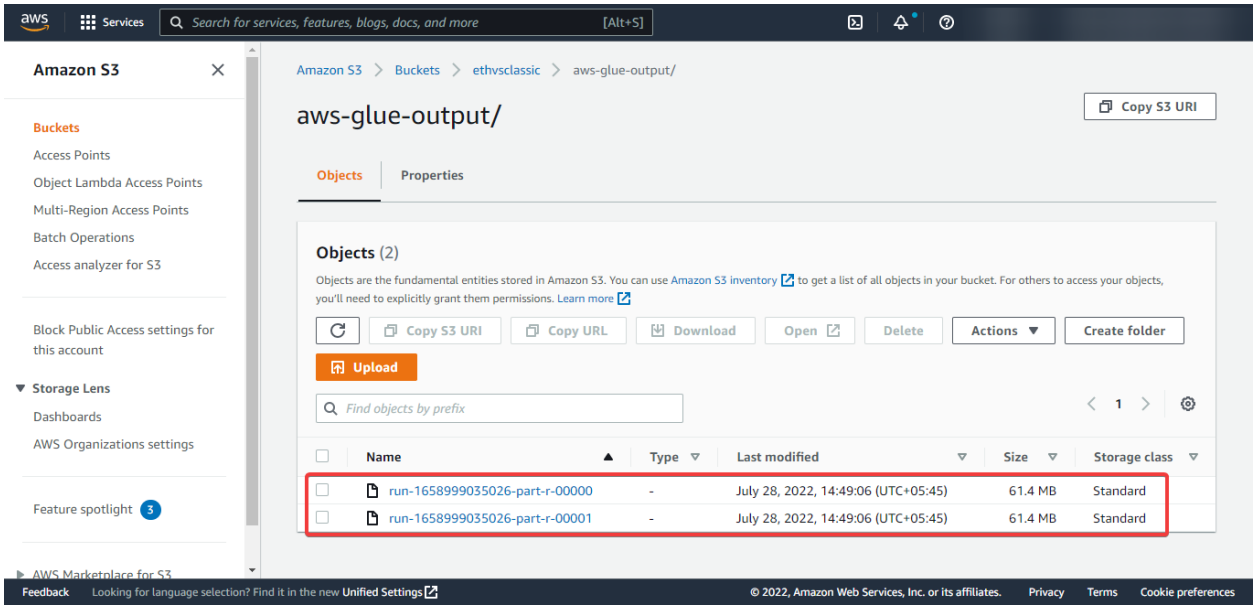
- Add Database **ethereum_dataset** and table name **ethereum_collection** from AWS Glue

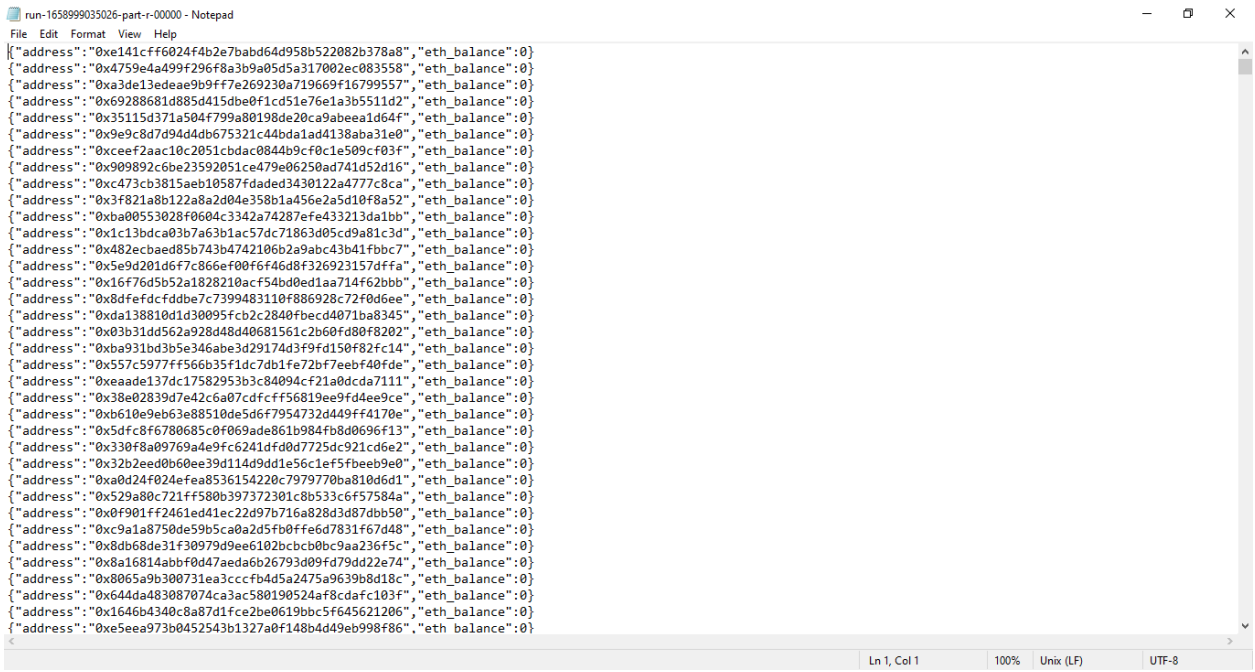


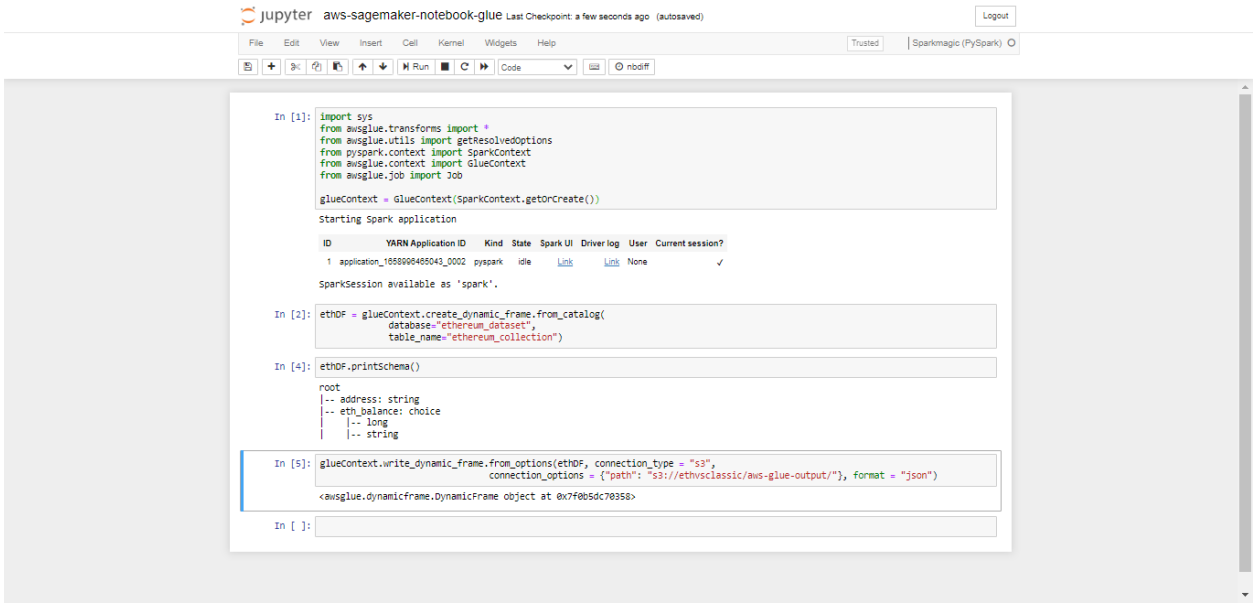

```
In [2]: ethDF = glueContext.create_dynamic_frame.from_catalog(
                    database="ethereum_dataset",
                    table_name="ethereum_collection")

In [4]: ethDF.printSchema()

root
|-- address: string
|-- eth_balance: choice
|    |-- long
|    |-- string

In [5]: glueContext.write_dynamic_frame.from_options(ethDF, connection_type = "s3",
                                    connection_options = {"path": "s3://ethvsclassic/aws-glue-output/"}, format = "json")

<awsglue.dynamicframe.DynamicFrame object at 0x7f0b5dc70358>

In [ ]:
```

Once completed **save** your notebook.

- Your **output** is also stored in **S3 bucket.**

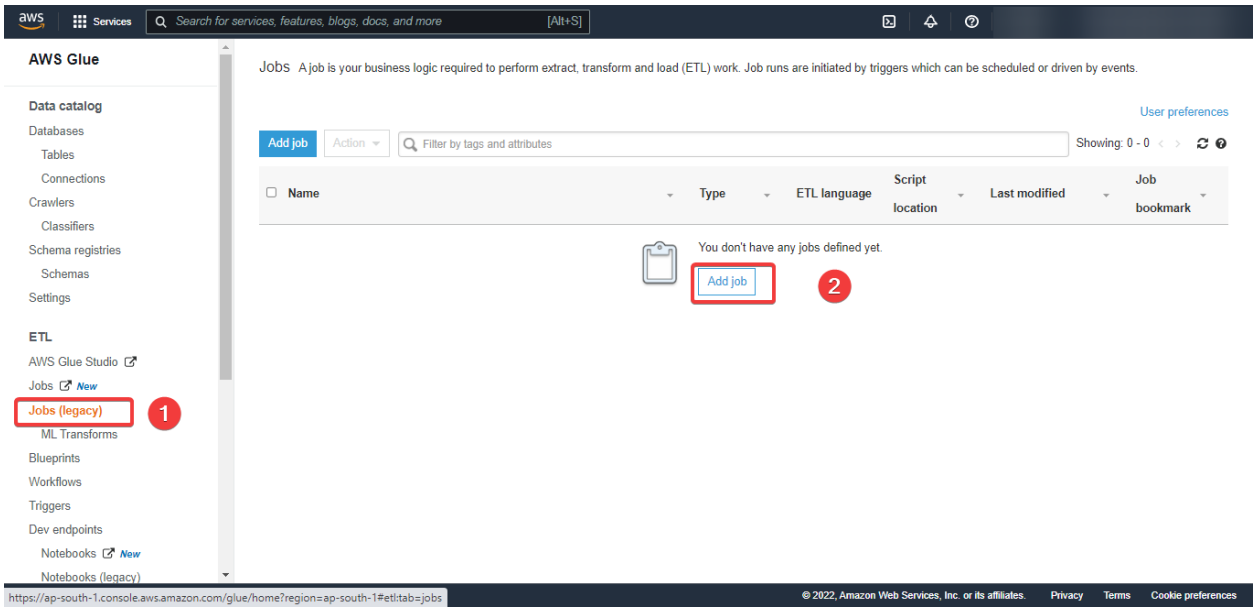

Services
Search for services, features, blogs, docs, and more
[Alt+S]
Amazon S3
Buckets
Access Points
Object Lambda Access Points
Multi-Region Access Points
Batch Operations
Access analyzer for S3
Block Public Access settings for this account
Storage Lens
Dashboards
AWS Organizations settings
Feature spotlight
Amazon S3 > Buckets > ethvsclassic > aws-glue-output/
aws-glue-output/
Copy S3 URI
Objects
Properties
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
Name
Type
Last modified
Size
Storage class
run-1658999035026-part-r-00000
July 28, 2022, 14:49:06 (UTC+05:45)
61.4 MB
Standard
run-1658999035026-part-r-00001
July 28, 2022, 14:49:06 (UTC+05:45)
61.4 MB
Standard
Feedback
Looking for language selection? Find it in the new Unified Settings
© 2022, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences


run-1658999035026-part-r-00000 - Notepad
File Edit Format View Help
{"address":"0xe141cff6024f4b2e7babd64d958b522082b378a8","eth_balance":0}
{"address":"0x4759e4a499f296f8a3b9a05d5a317002ec083558","eth_balance":0}
{"address":"0xa3de13edeae9b9ff7e269230a719669f16799557","eth_balance":0}
{"address":"0x69288681d885d415dbe0f1cd51e76e1a3b5511d2","eth_balance":0}
{"address":"0x35115d371a504f799a80198de20ca9abeea1d64f","eth_balance":0}
{"address":"0x9e9c8d7d94d4db675321c44bda1ad4138aba31e0","eth_balance":0}
{"address":"0xceef2aac10c2051cbdac0844b9cf0c1e509cf03f","eth_balance":0}
{"address":"0x909892c6be23592051ce479e06250ad741d52d16","eth_balance":0}
{"address":"0xc473cb3815aeb10587fdaded3430122a4777c8ca","eth_balance":0}
{"address":"0x3f821a8b122a8a2d04e358b1a456e2a5d10f8a52","eth_balance":0}
{"address":"0xba00553028f0604c3342a74287efe433213da1bb","eth_balance":0}
{"address":"0x1c13bdca03b7a63b1ac57dc71863d05cd9a81c3d","eth_balance":0}
{"address":"0x482ecbaed85b743b4742106b2a9abc43b41fbbc7","eth_balance":0}
{"address":"0x5e9d201d6f7c866ef00f6f46d8f326923157dffa","eth_balance":0}
{"address":"0x16f76d5b52a1828210acf54bd0ed1aa714f62bbb","eth_balance":0}
{"address":"0x8dfefdcfddbe7c7399483110f886928c72f0d6ee","eth_balance":0}
{"address":"0xda138810d1d30095fcb2c2840fbecd4071ba8345","eth_balance":0}
{"address":"0x03b31dd562a928d48d40681561c2b60fd80f8202","eth_balance":0}
{"address":"0xba931bd3b5e346abe3d29174d3f9fd150f82fc14","eth_balance":0}
{"address":"0x557c5977ff566b35f1dc7db1fe72bf7eebf40fde","eth_balance":0}
{"address":"0xeaade137dc17582953b3c84094cf21a0dcda7111","eth_balance":0}
{"address":"0x38e02839d7e42c6a07cdfcff56819ee9fd4ee9ce","eth_balance":0}
{"address":"0xb610e9eb63e88510de5d6f7954732d449ff4170e","eth_balance":0}
{"address":"0x5dfc8f6780685c0f069ade861b984fb8d0696f13","eth_balance":0}
{"address":"0x330f8a09769a4e9fc6241dfd0d7725dc921cd6e2","eth_balance":0}
{"address":"0x32b2eed0b60ee39d114d9dd1e56c1ef5fbeeb9e0","eth_balance":0}
{"address":"0xa0d24f024efea8536154220c7979770ba810d6d1","eth_balance":0}
{"address":"0x529a80c721ff580b397372301c8b533c6f57584a","eth_balance":0}
{"address":"0x0f901ff2461ed41ec22d97b716a828d3d87dbb50","eth_balance":0}
{"address":"0xc9a1a8750de59b5ca0a2d5fb0ffe6d7831f67d48","eth_balance":0}
{"address":"0x8db68de31f30979d9ee6102bcbcb0bc9aa236f5c","eth_balance":0}
{"address":"0x8a16814abbf0d47aeda6b26793d09fd79dd22e74","eth_balance":0}
{"address":"0x8065a9b300731ea3cccfb4d5a2475a9639b8d18c","eth_balance":0}
{"address":"0x644da483087074ca3ac580190524af8cdafc103f","eth_balance":0}
{"address":"0x1646b4340c8a87d1fce2be0619bbc5f645621206","eth_balance":0}
{"address":"0xe5eea973b0452543b1327a0f148b4d49eb998f86","eth_balance":0}
Ln 1, Col 1
100%
Unix (LF)
UTF-8


jupyter aws-sagemaker-notebook-glue Last Checkpoint: a few seconds ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Sparkmagic (PySpark)
Run
Code
nbdiff
In [1]:
import sys
from awsglue.transforms import *
from awsglue.utils import getResolvedOptions
from pyspark.context import SparkContext
from awsglue.context import GlueContext
from awsglue.job import Job
glueContext = GlueContext(SparkContext.getOrCreate())
Starting Spark application
ID YARN Application ID Kind State Spark UI Driver log User Current session?
1 application_1658906465043_0002 pyspark idle Link Link None ✓
SparkSession available as 'spark'.
In [2]:
ethDF = glueContext.create_dynamic_frame.from_catalog(
database="ethereum_dataset",
table_name="ethereum_collection")
In [4]:
ethDF.printSchema()
root
|-- address: string
|-- eth_balance: choice
| |-- long
| |-- string
In [5]:
glueContext.write_dynamic_frame.from_options(ethDF, connection_type = "s3",
connection_options = {"path": "s3://ethvsclassic/aws-glue-output/"}, format = "json")
<awsglue.dynamicframe.DynamicFrame object at 0x7f0b5dc70358>
In [ ]:

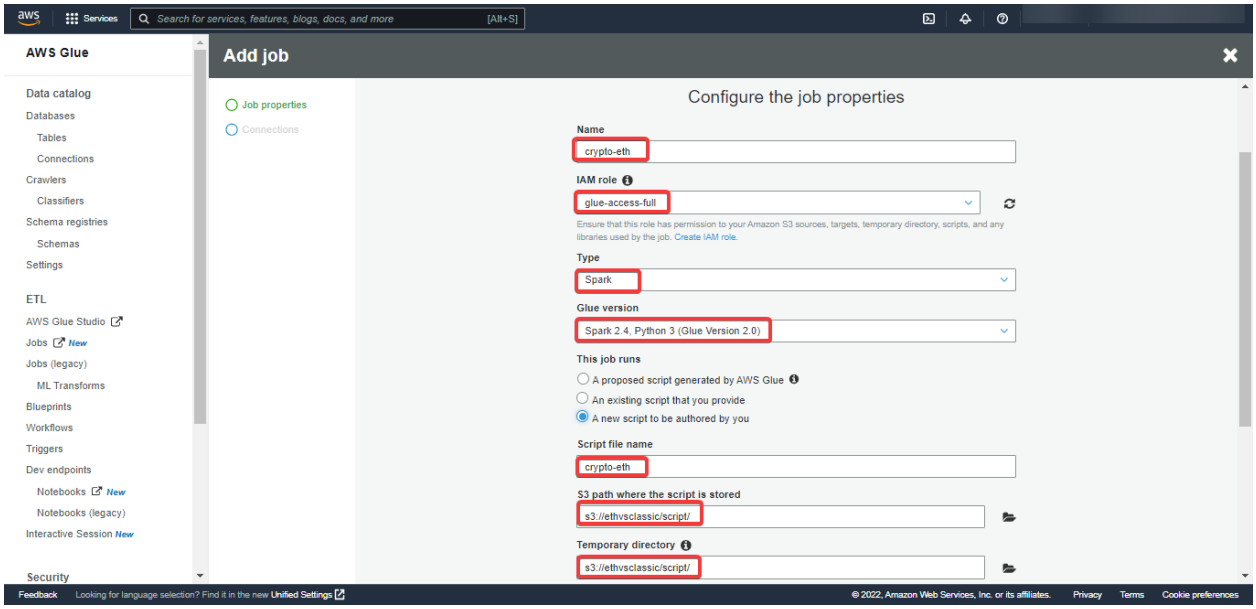
Run PySpark Job in AWS Glue

- Go to **AWS Glue** on AWS console.
- click **jobs(legacy)** on the ETL on the left side of **AWS Glue**
- Click Add job.

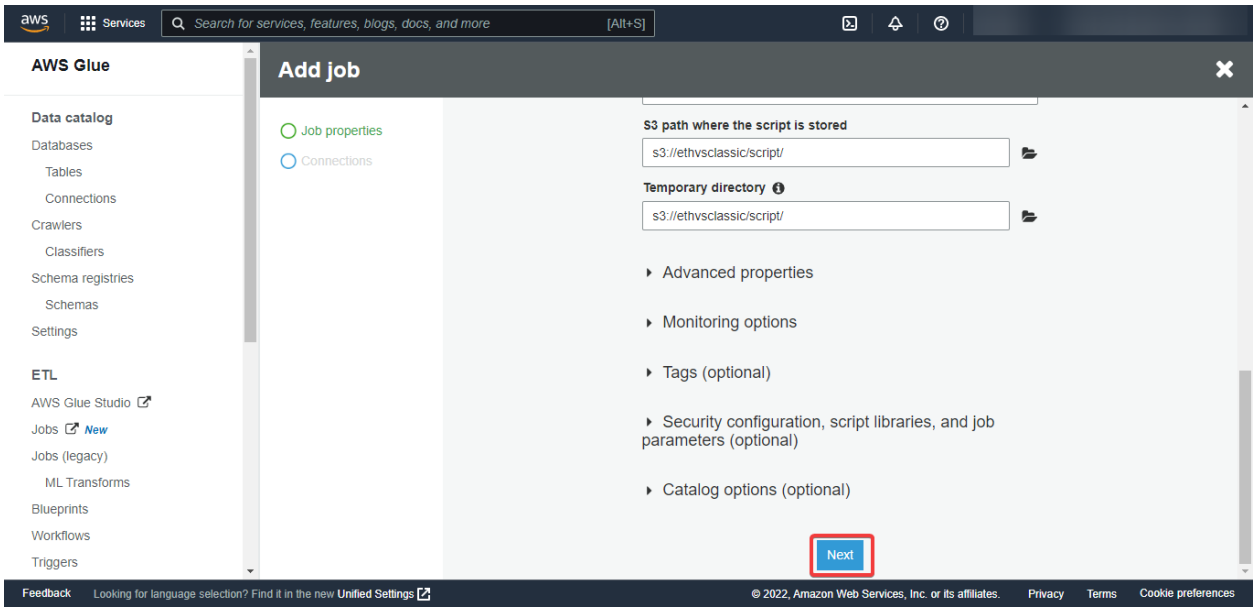


On the next screen,

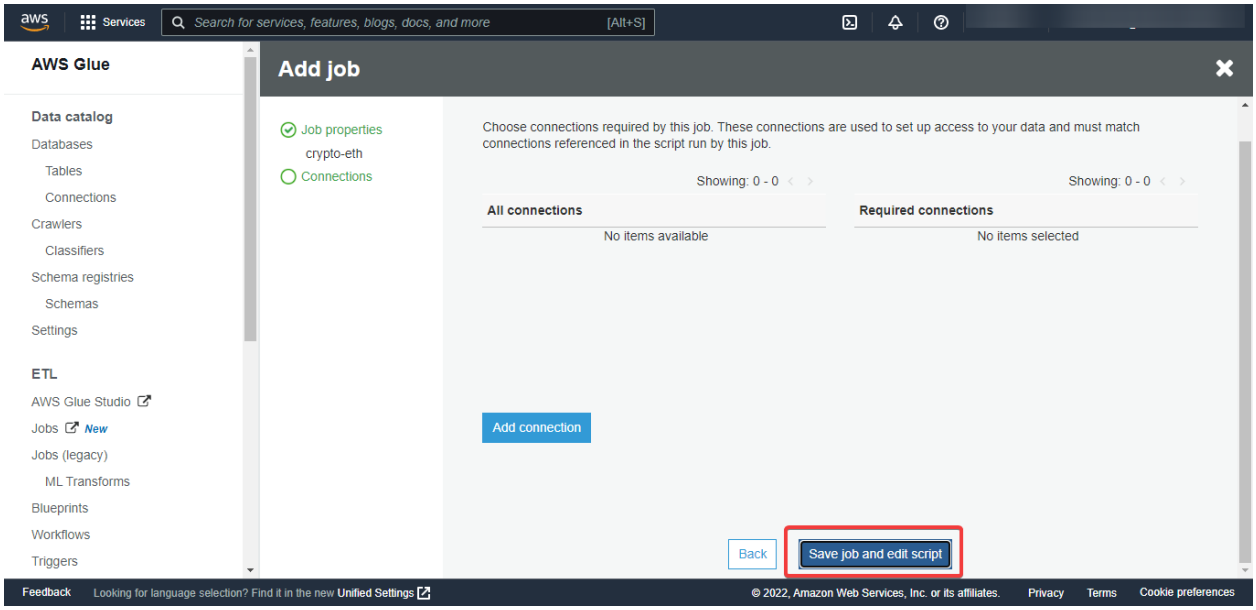
- Name Job crypto-eth
- Choose **IAM role** glue-access-full
- Type **Spark** with **Spark 2.Python 3 Version**
- Select a **new script to be authorized by you**.
- Input script file name **crypto-eth** as default job.
- Select S3 path where you want to store your script.
  S3://ethvsclassic/script/


aws
Services
Search for services, features, blogs, docs, and more
[Alt+S]
AWS Glue
Add job
Data catalog
Databases
Tables
Connections
Crawlers
Classifiers
Schema registries
Schemas
Settings
ETL
AWS Glue Studio
Jobs New
Jobs (legacy)
ML Transforms
Blueprints
Workflows
Triggers
Dev endpoints
Notebooks New
Notebooks (legacy)
Interactive Session New
Security
Job properties
Connections
Configure the job properties
Name
crypto-eth
IAM role
glue-access-full
Ensure that this role has permission to your Amazon S3 sources, targets, temporary directory, scripts, and any libraries used by the job. Create IAM role.
Type
Spark
Glue version
Spark 2.4, Python 3 (Glue Version 2.0)
This job runs
A proposed script generated by AWS Glue
An existing script that you provide
A new script to be authored by you
Script file name
crypto-eth
S3 path where the script is stored
s3://ethvsclassic/script/
Temporary directory
s3://ethvsclassic/script/
Feedback
Looking for language selection? Find it in the new Unified Settings
© 2022, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences

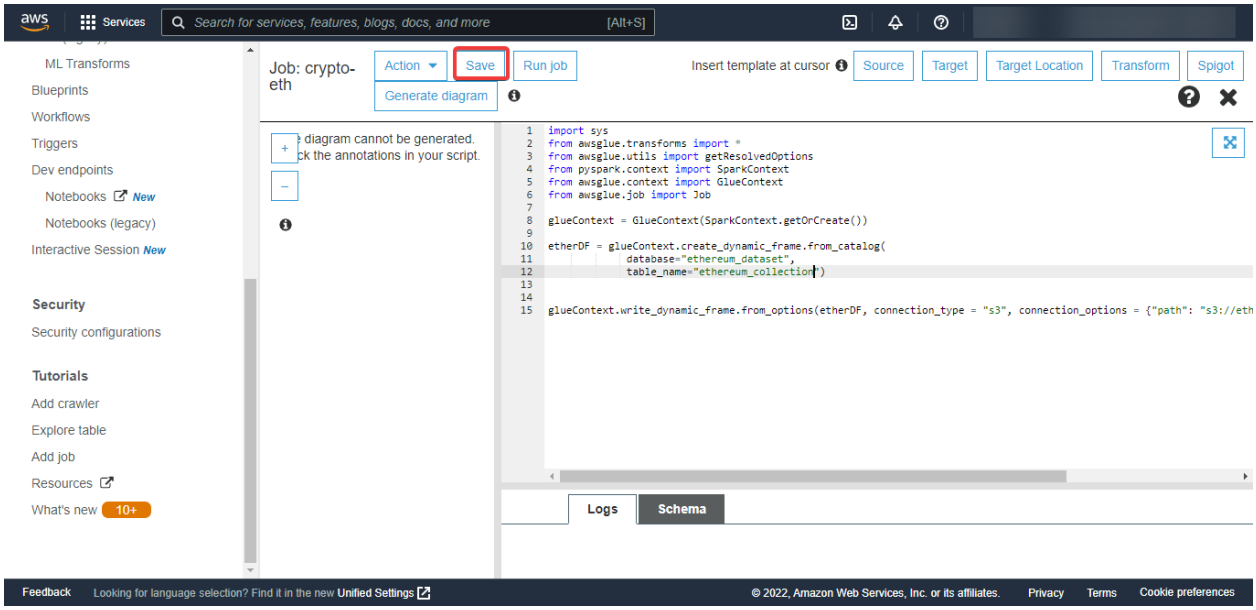
- Click **Next**


aws
Services
Search for services, features, blogs, docs, and more
[Alt+S]
AWS Glue
Add job
Data catalog
Databases
Tables
Connections
Crawlers
Classifiers
Schema registries
Schemas
Settings
ETL
AWS Glue Studio
Jobs New
Jobs (legacy)
ML Transforms
Blueprints
Workflows
Triggers
Job properties
Connections
S3 path where the script is stored
s3://ethvsclassic/script/
Temporary directory
s3://ethvsclassic/script/
Advanced properties
Monitoring options
Tags (optional)
Security configuration, script libraries, and job parameters (optional)
Catalog options (optional)
Next
Feedback
Looking for language selection? Find it in the new Unified Settings
© 2022, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences

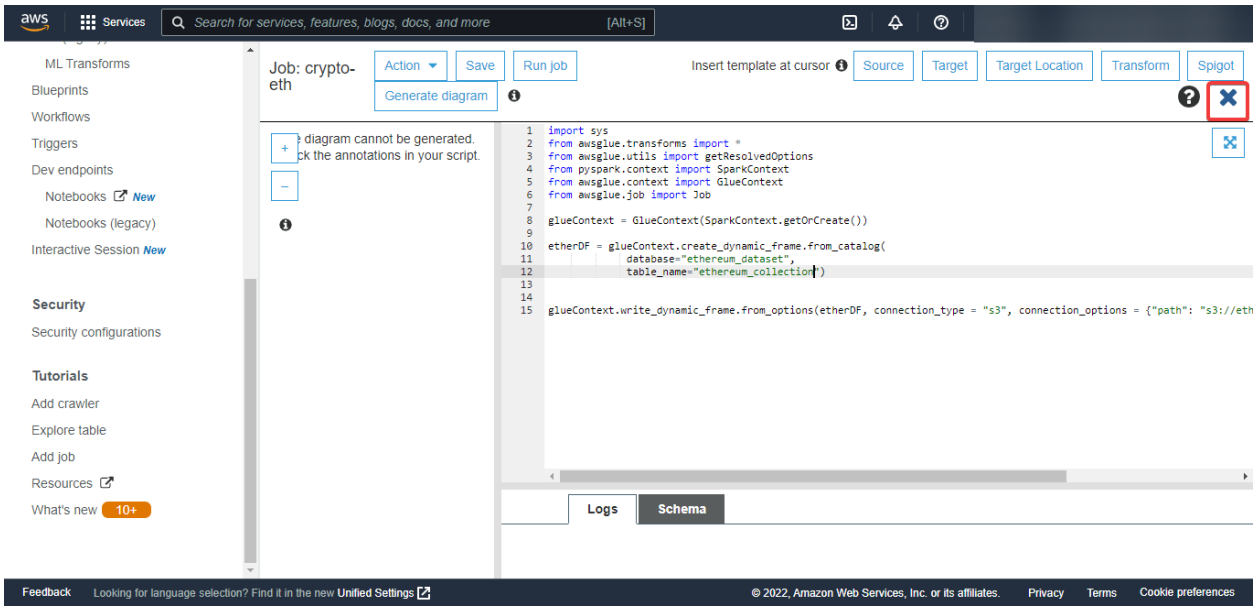
- Click **save job and edit script**



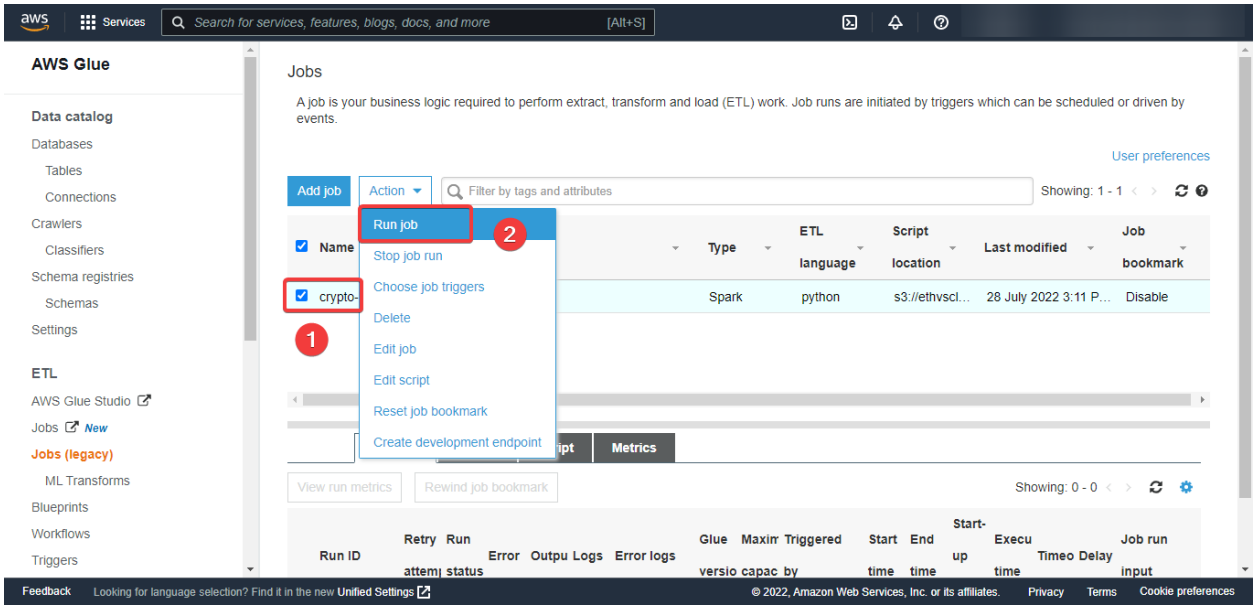
Copy paste the following code and click on the **Save** button. There is a bucket location **s3://ethvsclassic/glue-job/** mentioned in the script as the target location for the data. If you created bucket with a different name then use that bucket name.



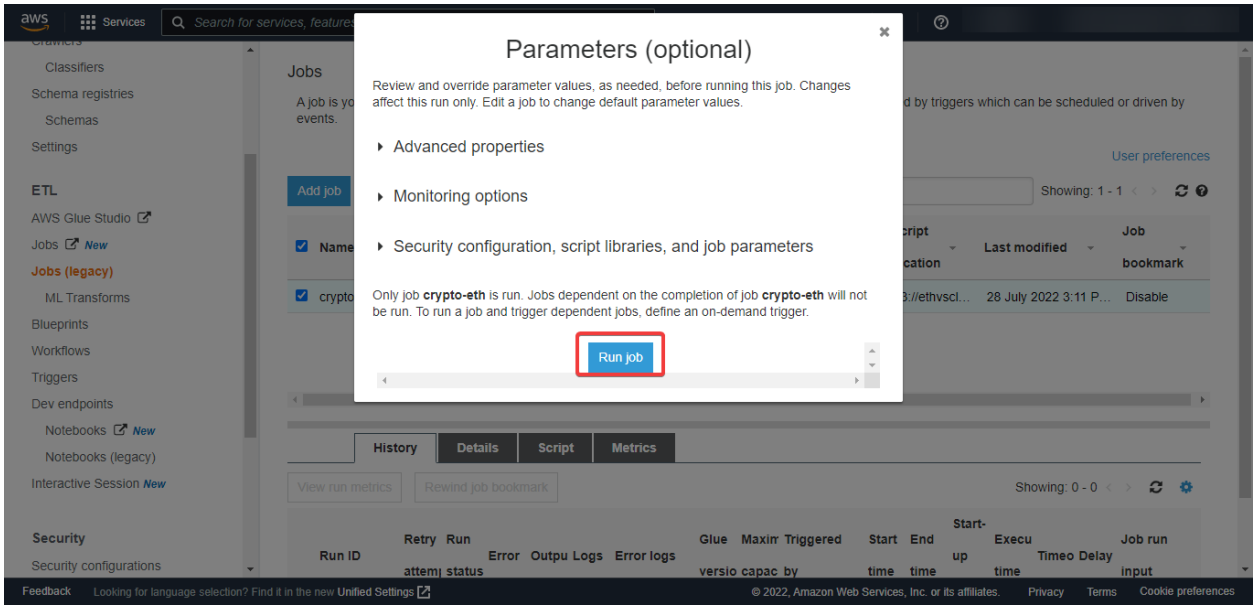
The job is saved. Click on the **X** button in the right top corner of the job definition to close the job definition.



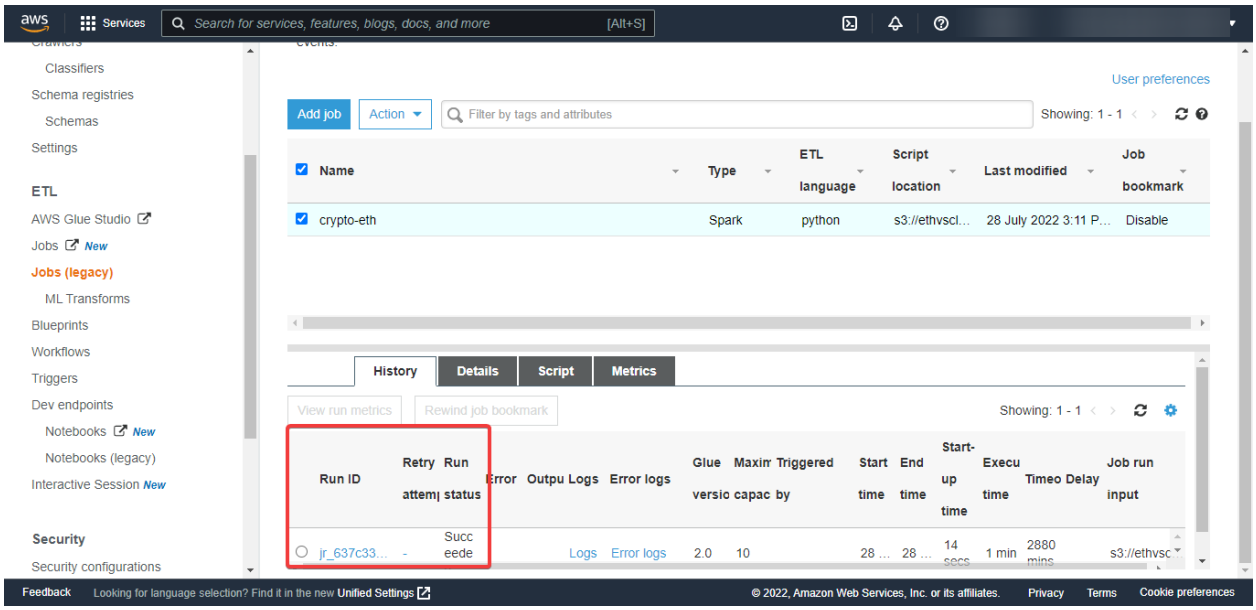
- Select your job and **Run job**



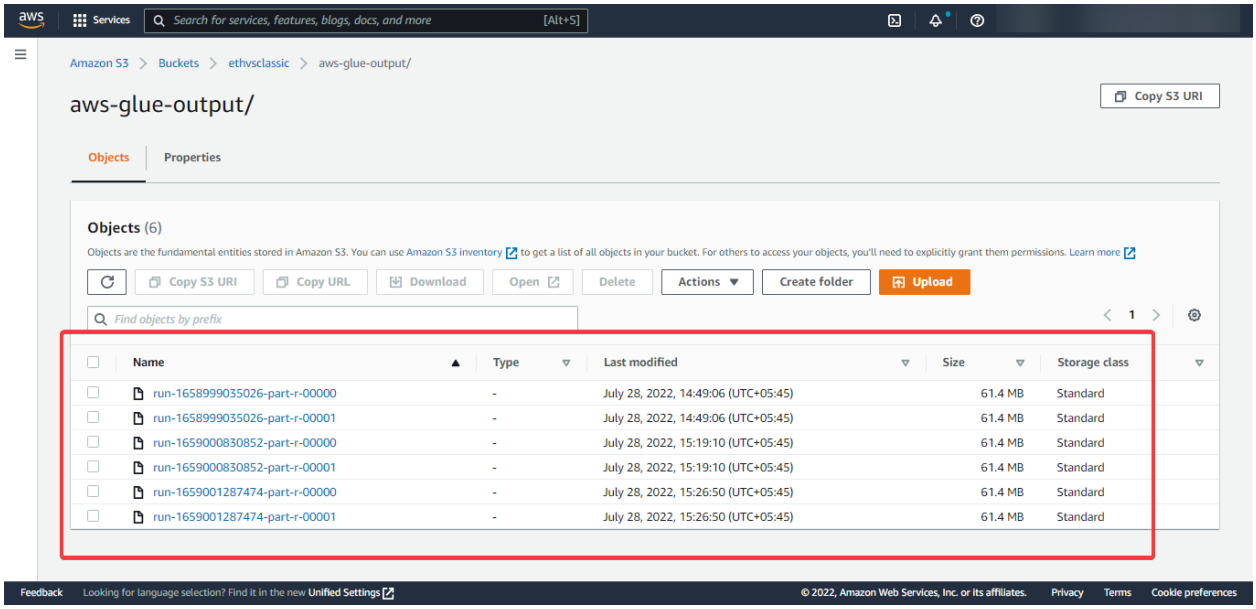
- Click **Run job** once a dialogue box is opened.



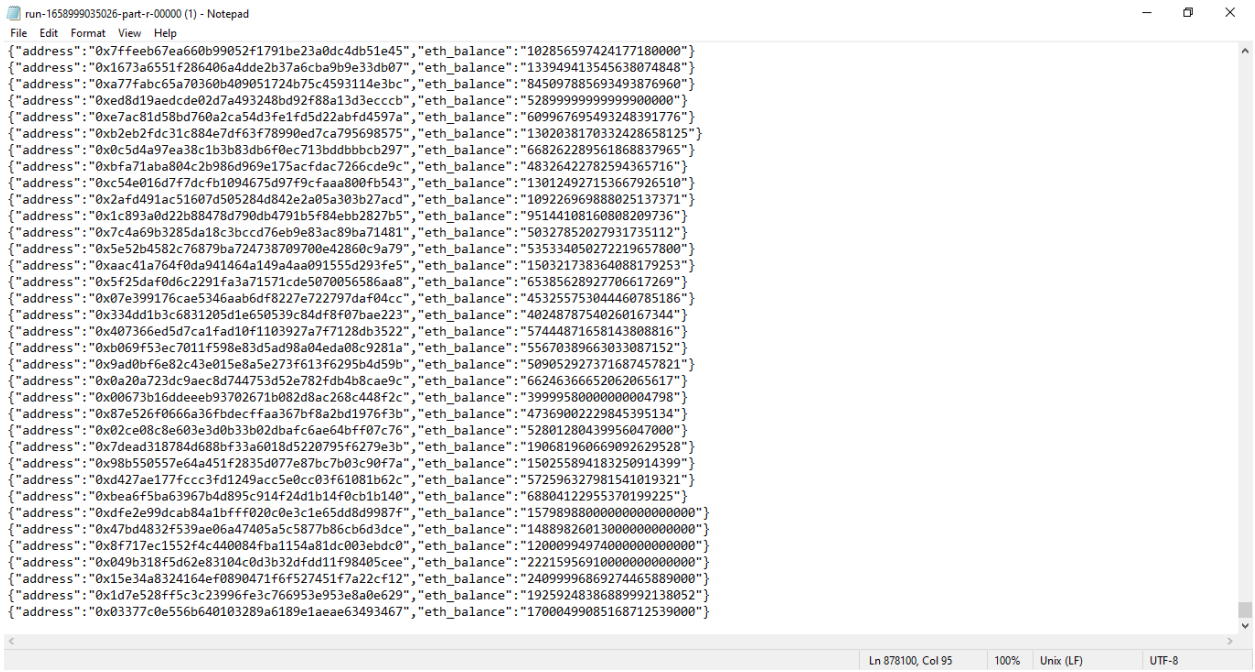
Wait till the job status changes to Succeeded. It might take some time in finishing the job.



Once the job completes, you can see the data created in the S3 location by the job.



- **Once you download, we can see our json file.**



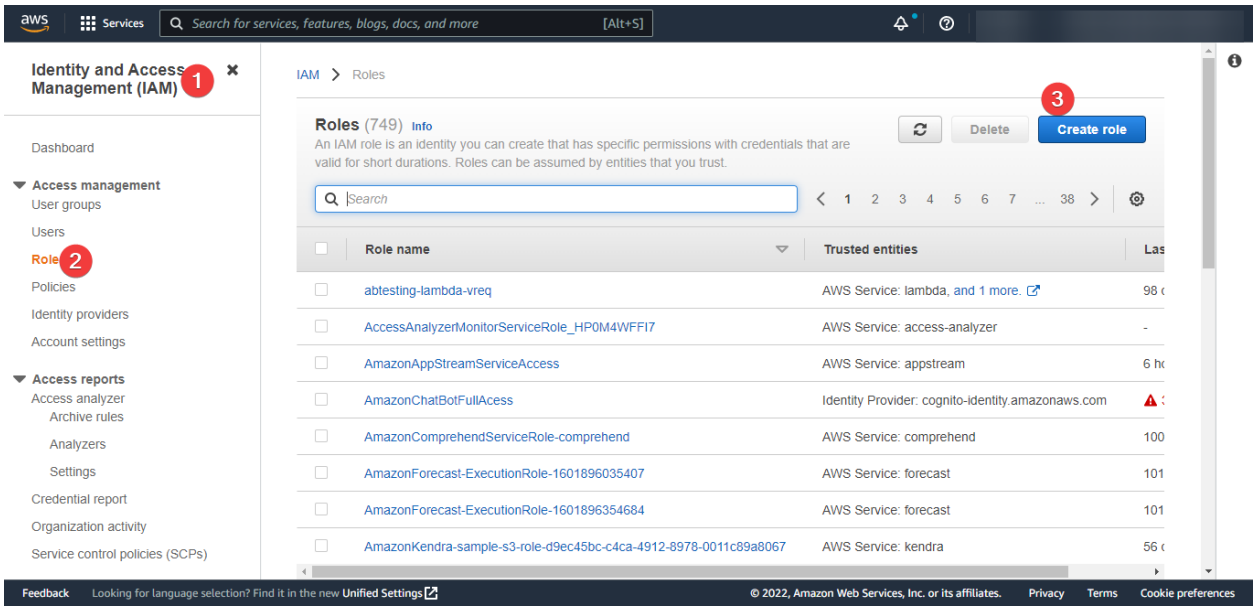
Using Amazon EMR with AWS Glue Catalog

- Make sure you have already created **IAM role, S3 bucket, glue database, glue cluster. Follow above procedure**

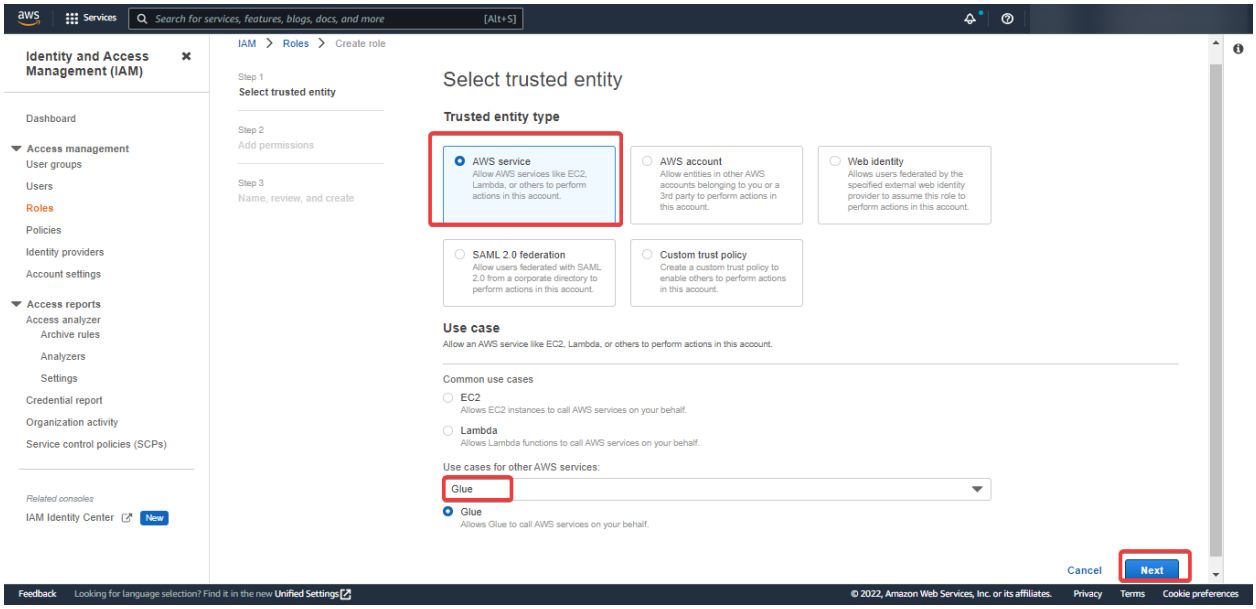
Create IAM Role for Glue

You start with creation of the IAM role which AWS Glue uses for the authorization to call other AWS Services.

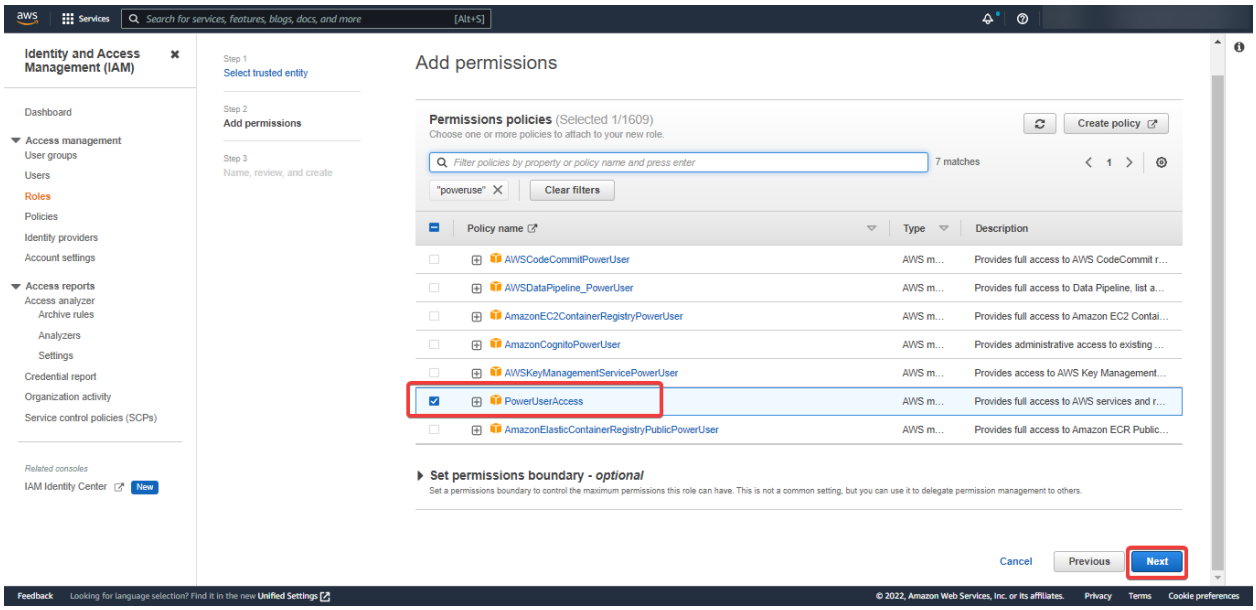
1. Login to the AWS Console and select **Mumbai** as the region.
2. Go to the IAM Management console and
3. click on the **Roles** menu in the left and then click on the **Create role** button.

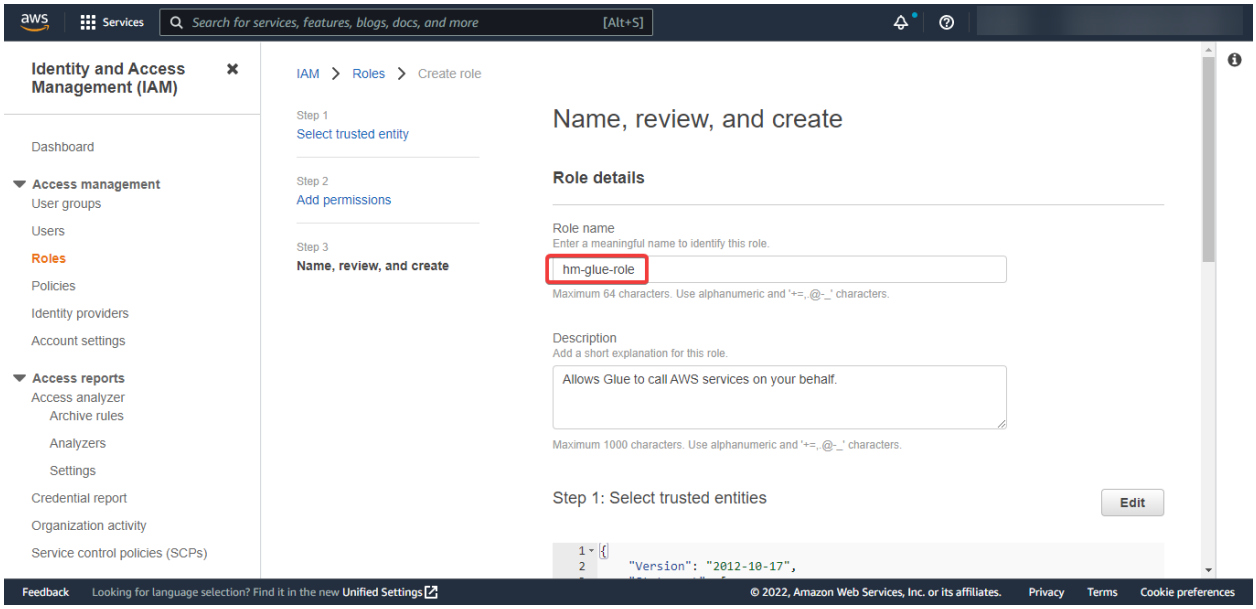


On the next screen, select **Glue** as the service and click on the **Next: Permissions** button.

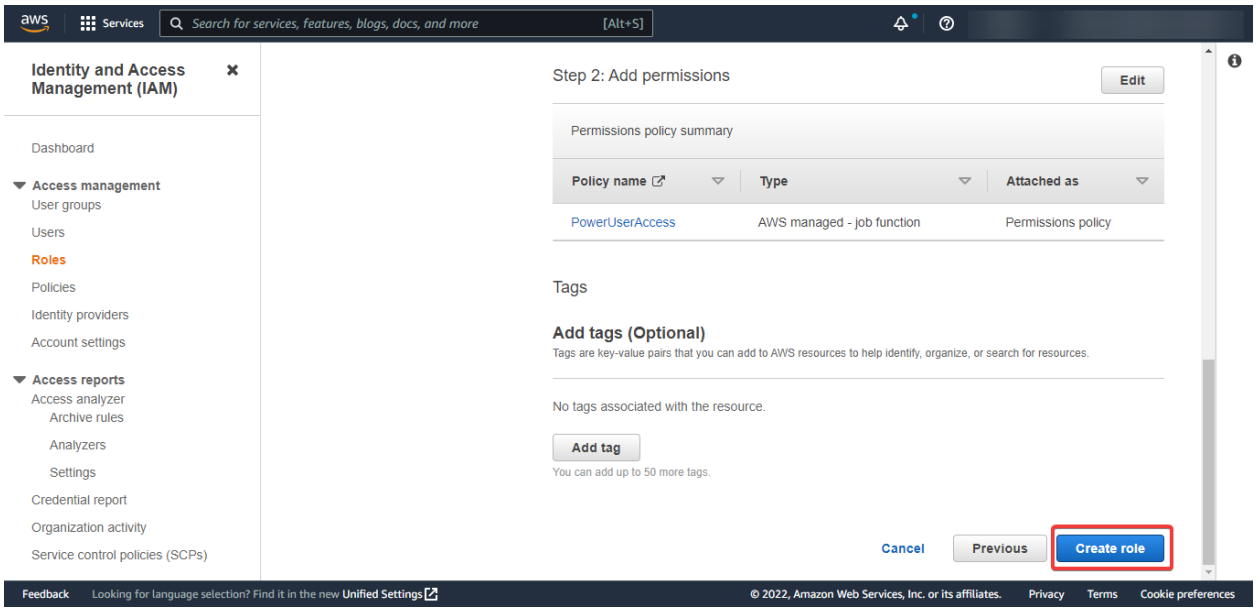


On the next screen, select **PowerUserAccess** as the policy and click on the **Next** button. The exercise is using power user permission but in actual production use it is recommended to use minimum required permission only.

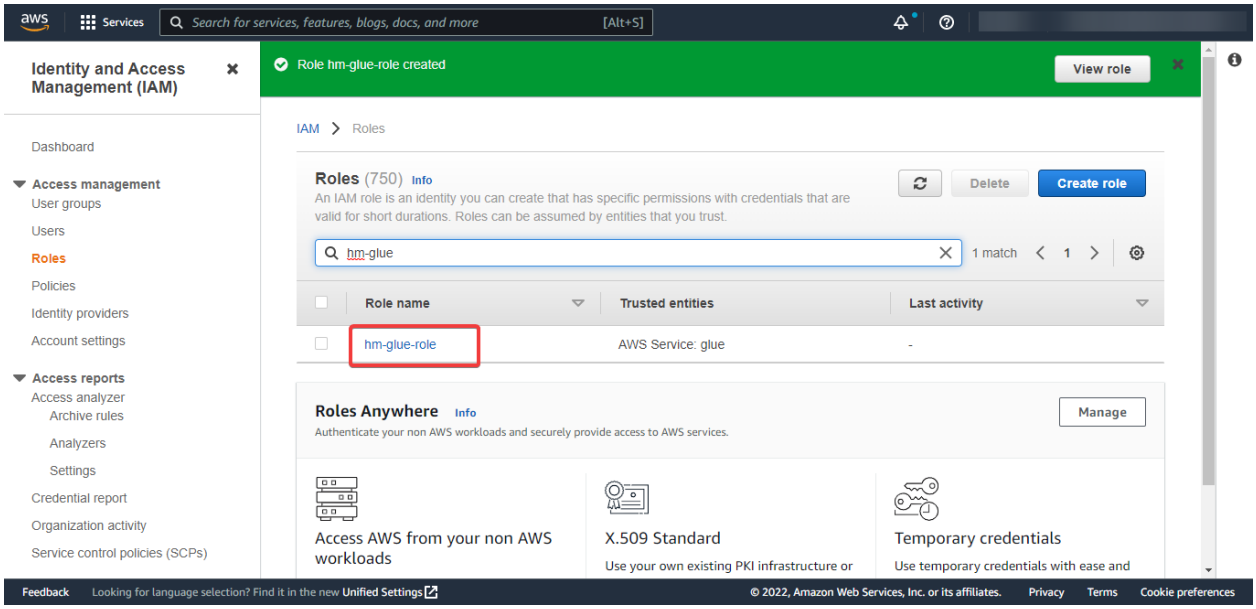




- On the next screen, click on the **Next: Review** button.
- On the next screen, type in **hm-glue-role** for the **Role name** and click on the **Create role** button.

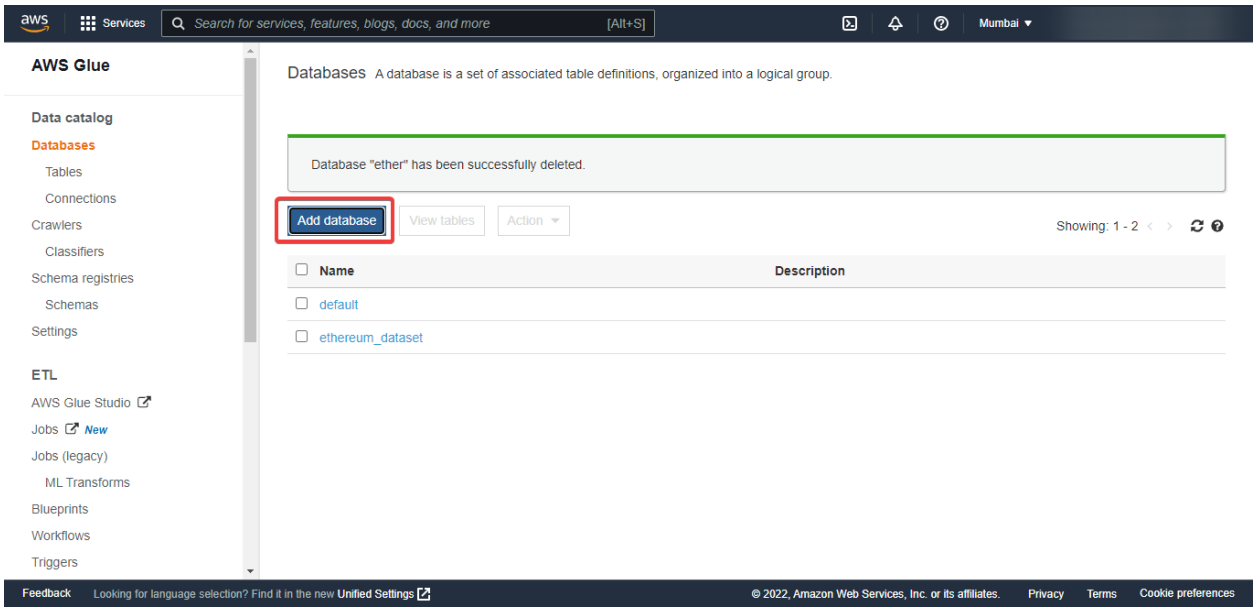


- IAM Glue Role is created now.

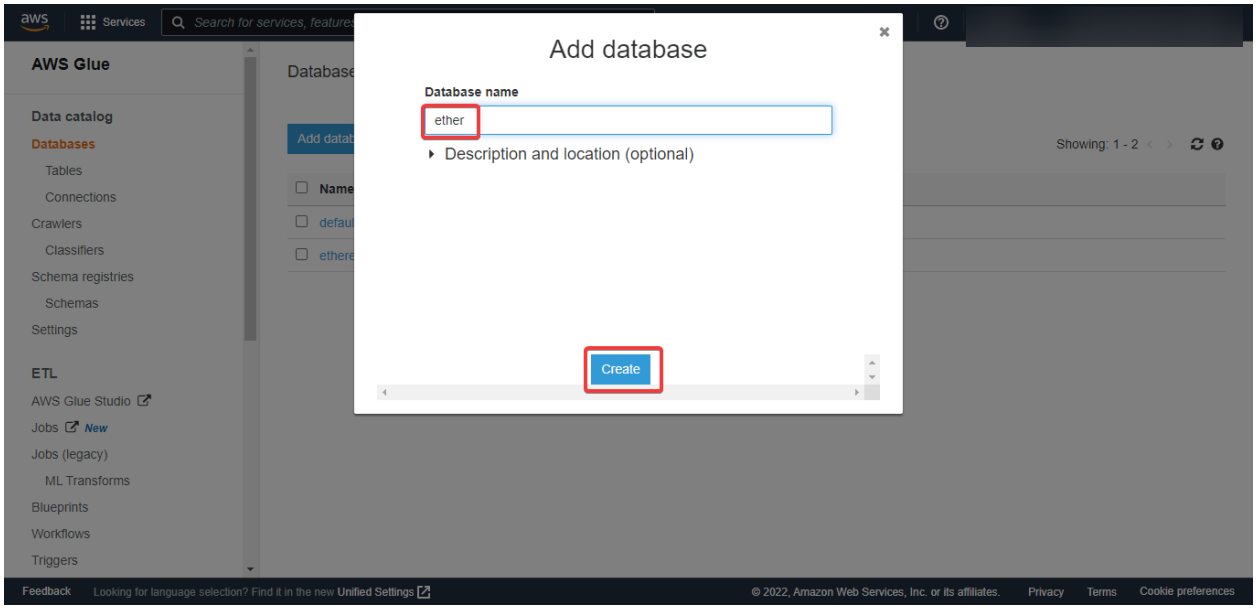


Create Glue database and glue cluster

- Click on Database on Data Catalog and Add database.



- Name a Database name **ether**
- Click **Create**

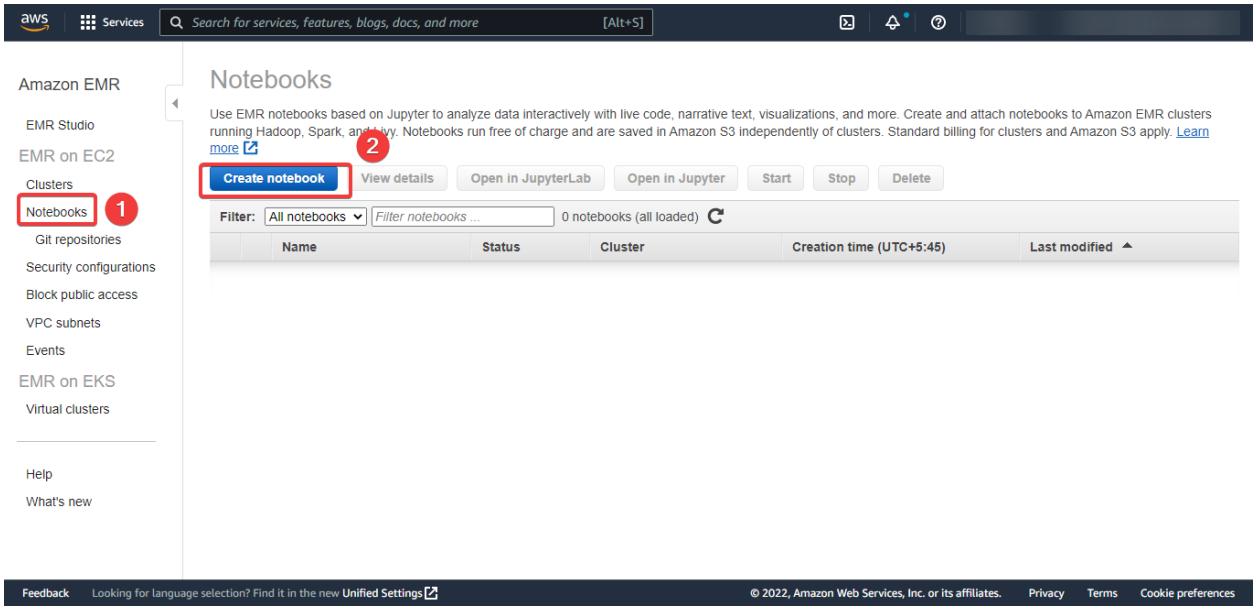


- Then add a crawler follow same process only change IAM role and name as hm-glue-role.
- Make sure you have S3 bucket, glue database, glue cluster as we already created above.

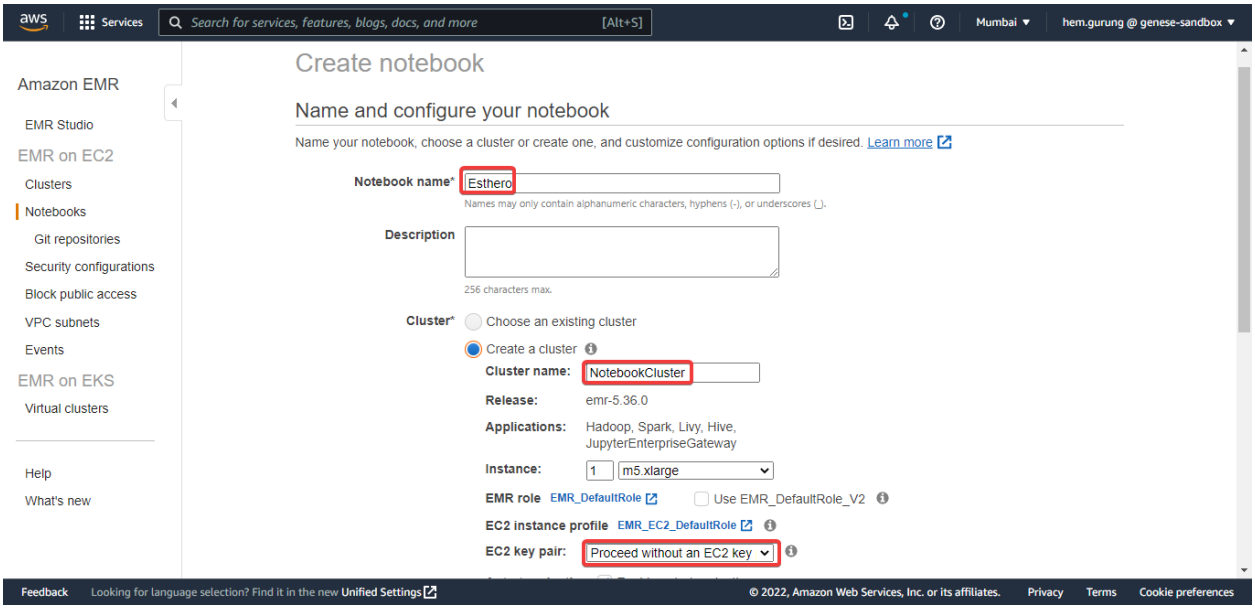
Launch EMR Cluster and Run Jupyter Notebook

You launch EMR cluster which is used to process data using Glue Data Catalog and PySpark code.

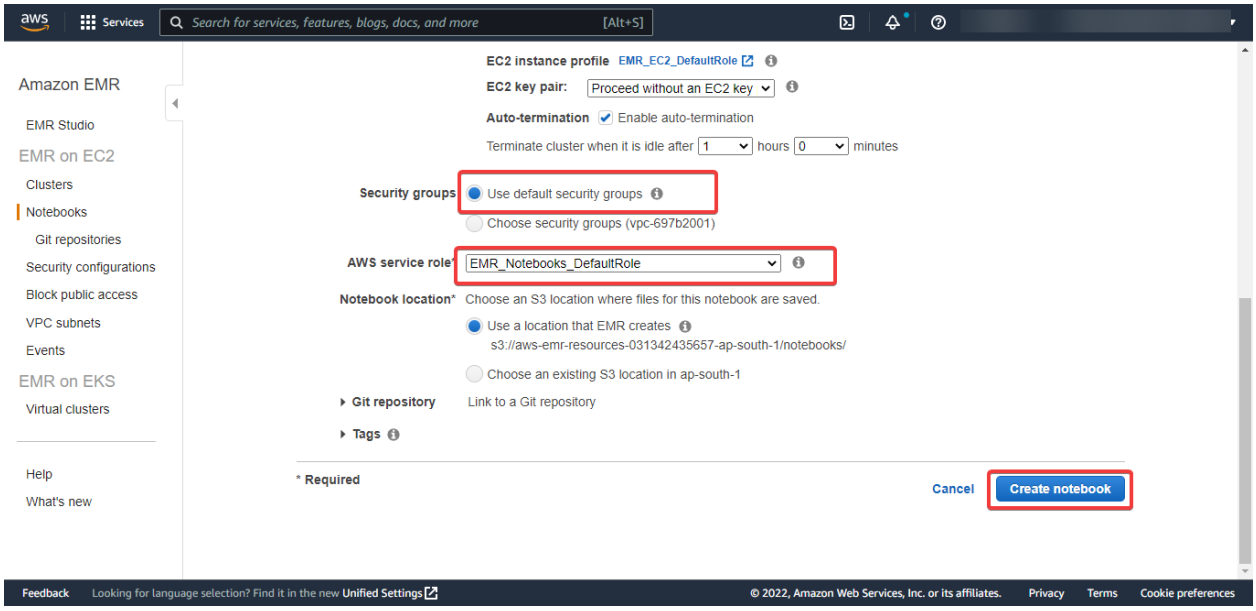
1. Go to the EMR Management console and click on the **Create notebook** button.



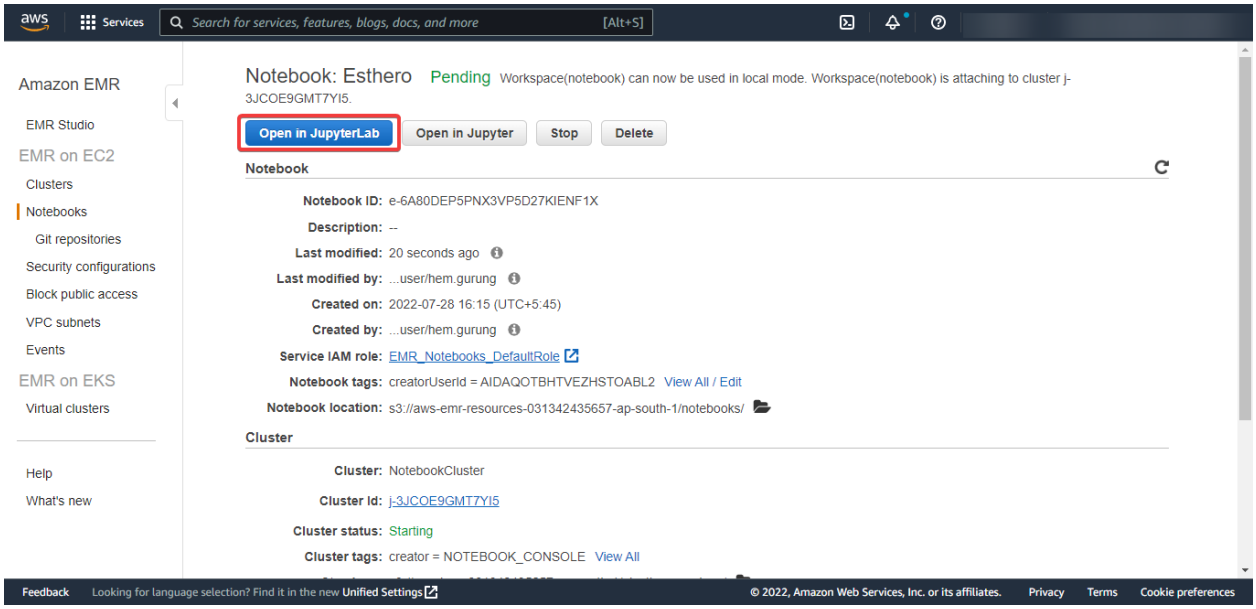
- Create a **Notebook name** **Esthero**
- Create a **cluster name** **NotebookCluster**
- Choose **instance** **1 mx5.xlarge size**
- Choose EC2 key pair **Proceed without EC2 key**



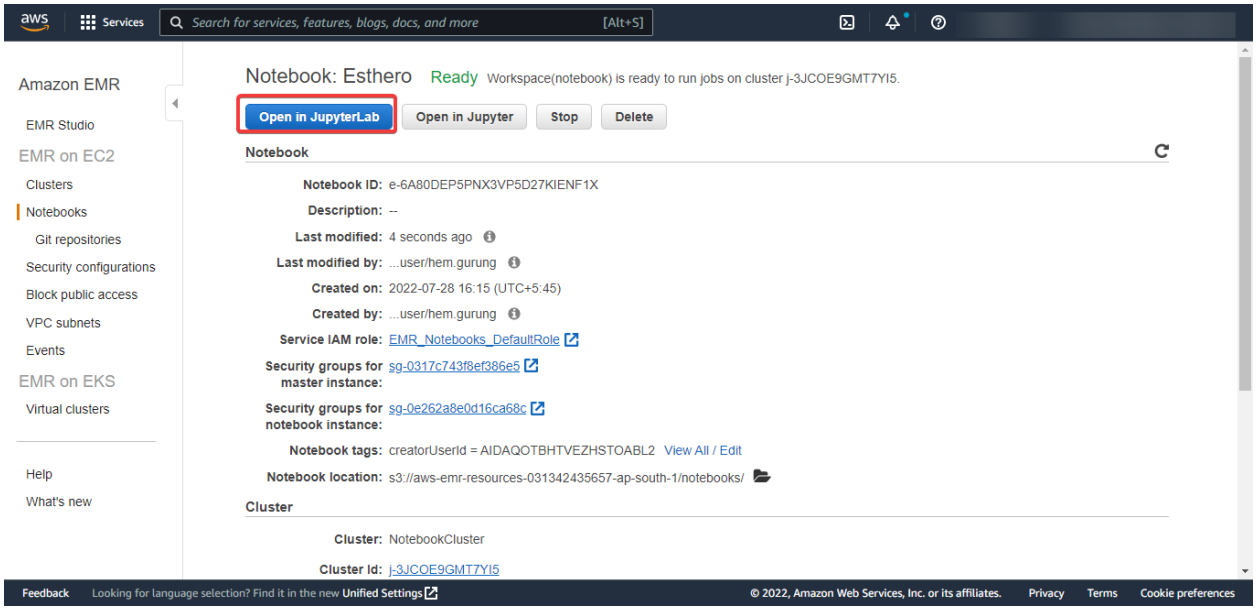
- Select **default security group** and **AWS service role as EMR_Notebook_DefaultRole**
- Click **Create notebook**



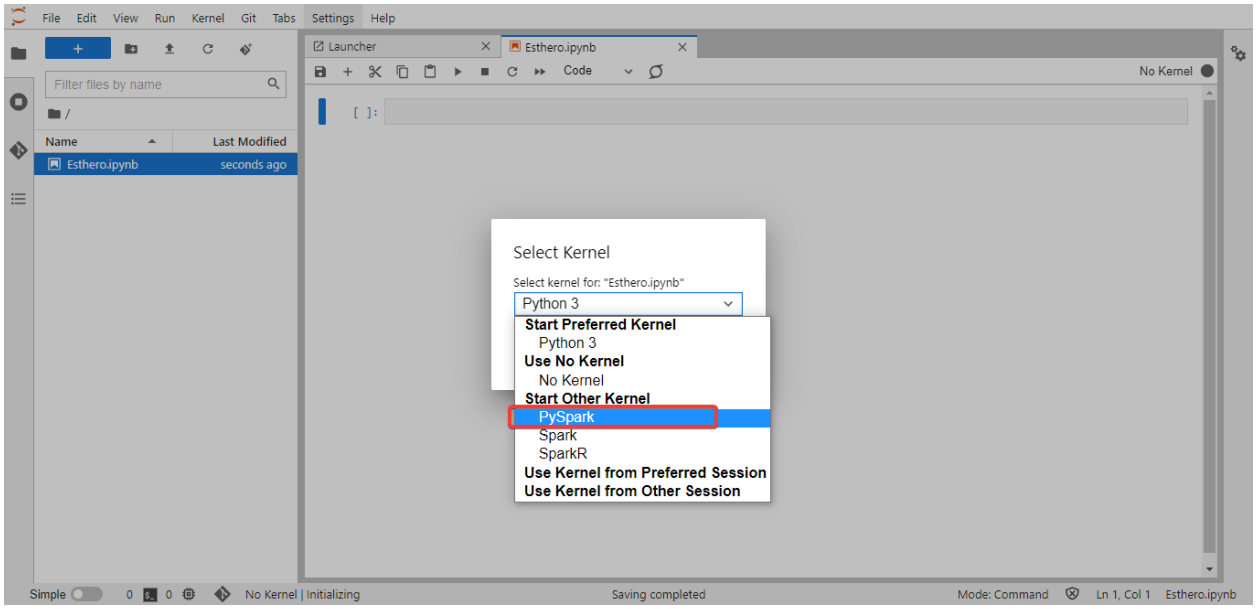
- **Notebook: Esthero** is still **pending.**
- Once successful click Open in **JupyterLab** or **Open in Jupyter**



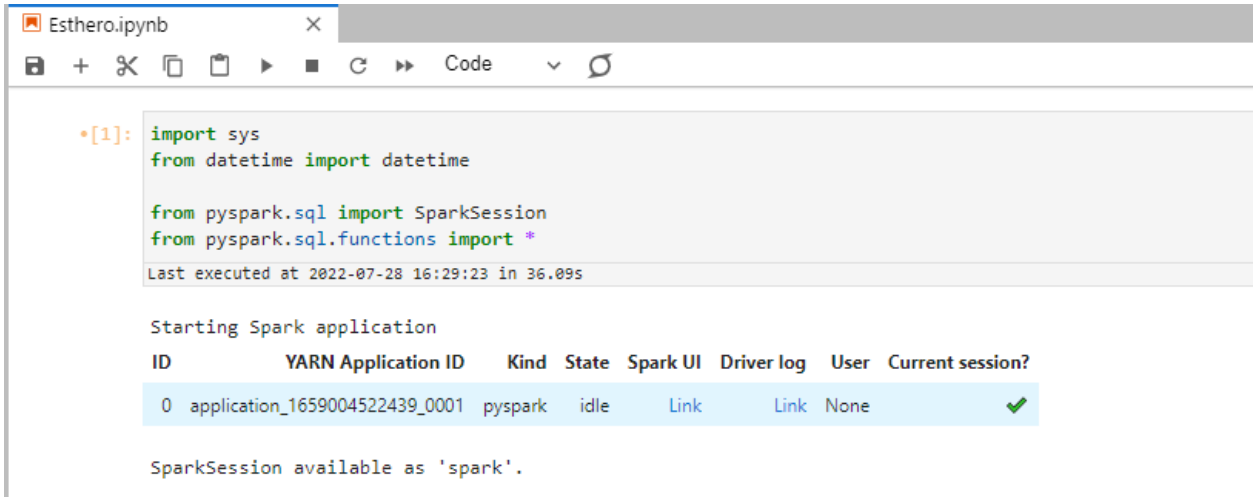
- You now run PySpark code to process S3 based data in the EMR Cluster using AWS Glue Catalog.
- In the EMR Management console, on the **Esthero** notebook page, click on the **Open in Jupyter** button.



It will open Notebook environment in a new browser tab or window. In Jupyter environment, click on **PySpark** option under the **Kernel** menu.



It will open Jupyter notebook IDE. It the cell, copy-paste the following code and run it. The code imports modules for the PySpark.



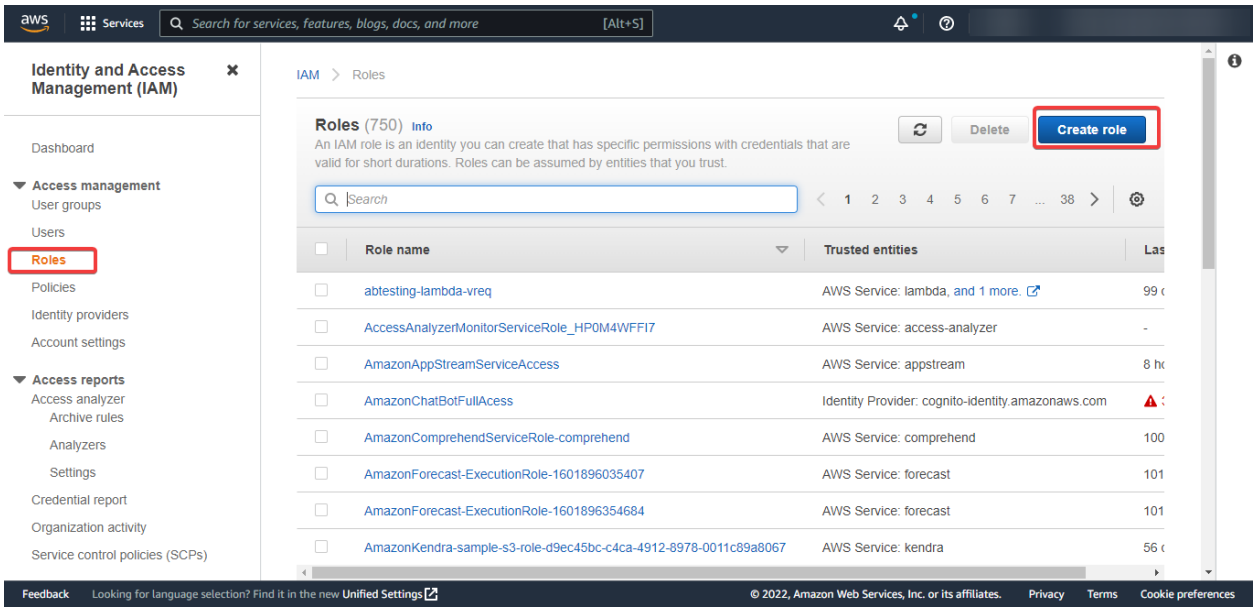
You copy-paste and run the following code to get the spark session.

```
[2]: spark = SparkSession\
         .builder\
         .appName("SparkETL")\
         .getOrCreate()
Last executed at 2022-07-28 16:30:25 in 1.38s
```

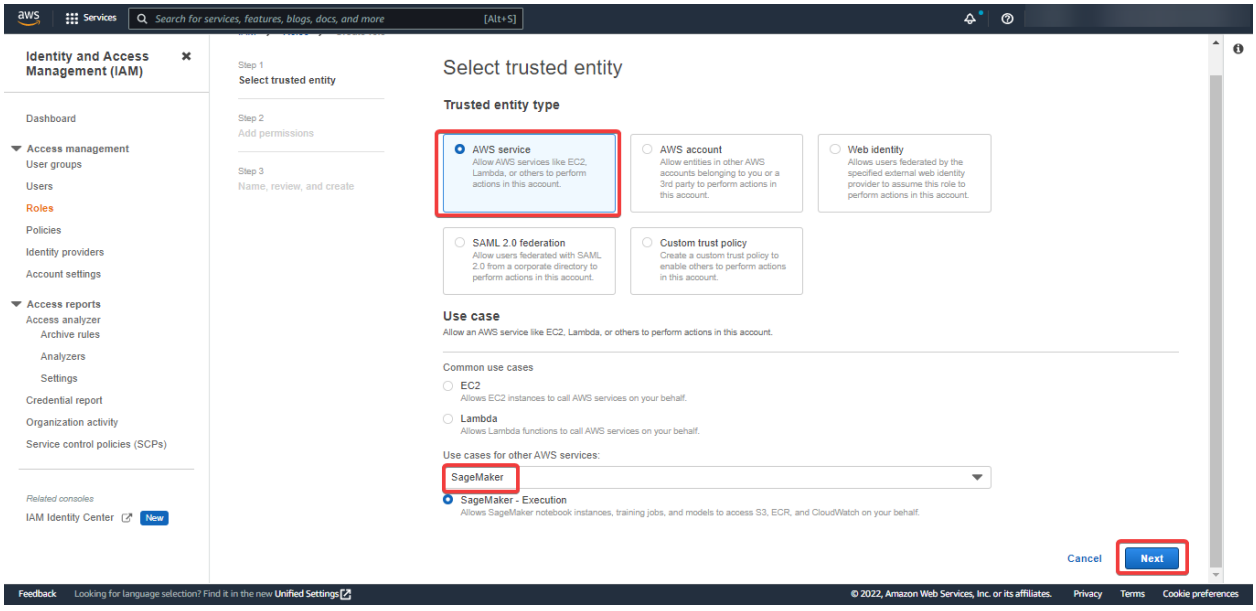
Now you can run your notebook.

IAM ROLE for Glue: Sagemaker

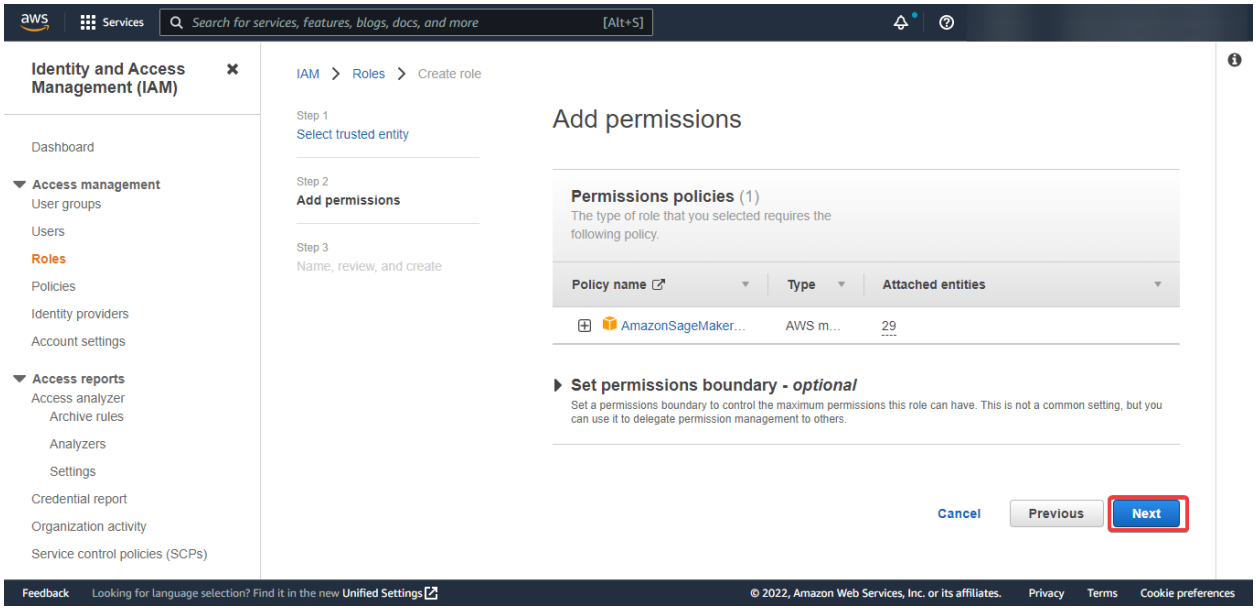
- Go **IAM Role** on aws console
- Select **Roles** on **IAM Dashboard**
- **Create role.**



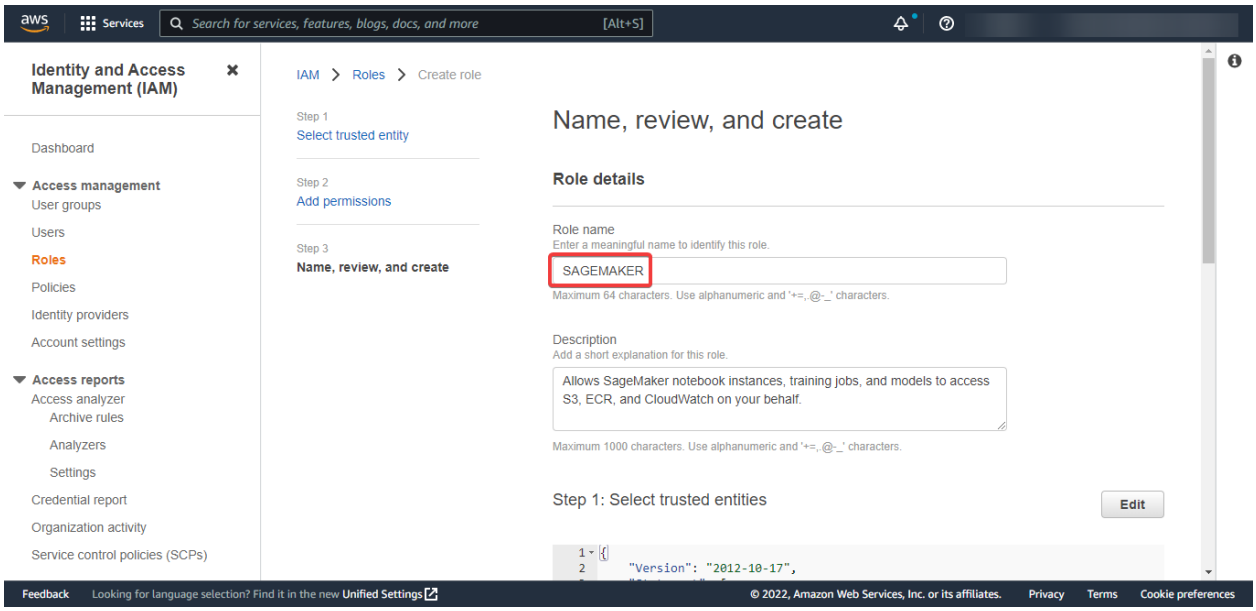
- Select **sagemaker** on **AWS service**.
- Click **Next**.



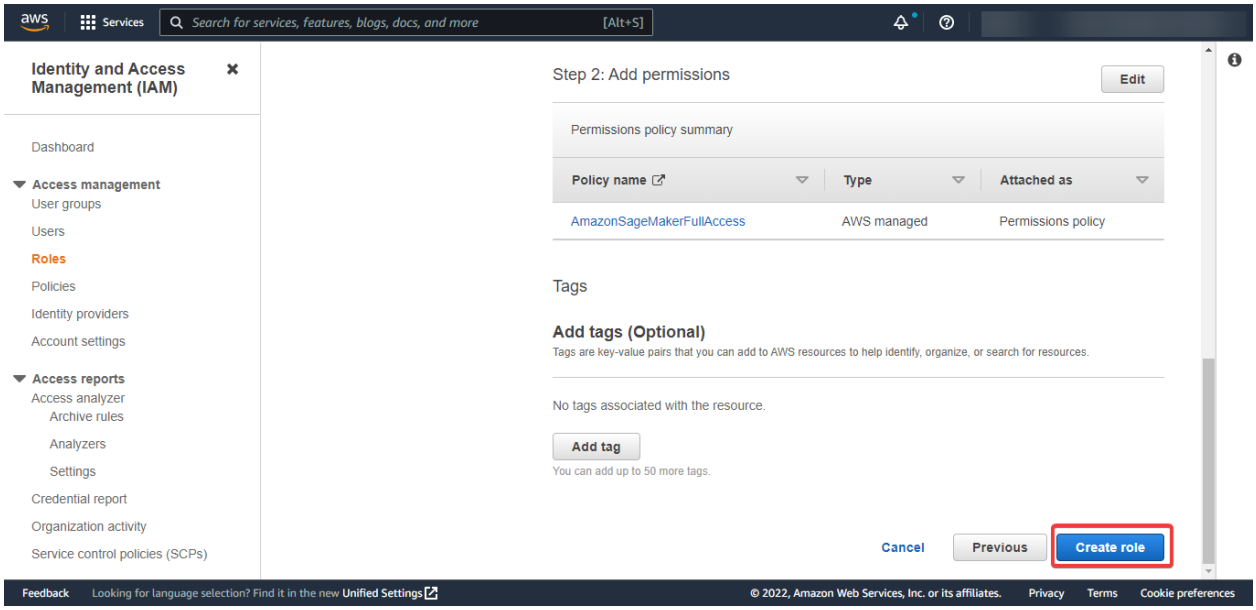
- Click **Next.**



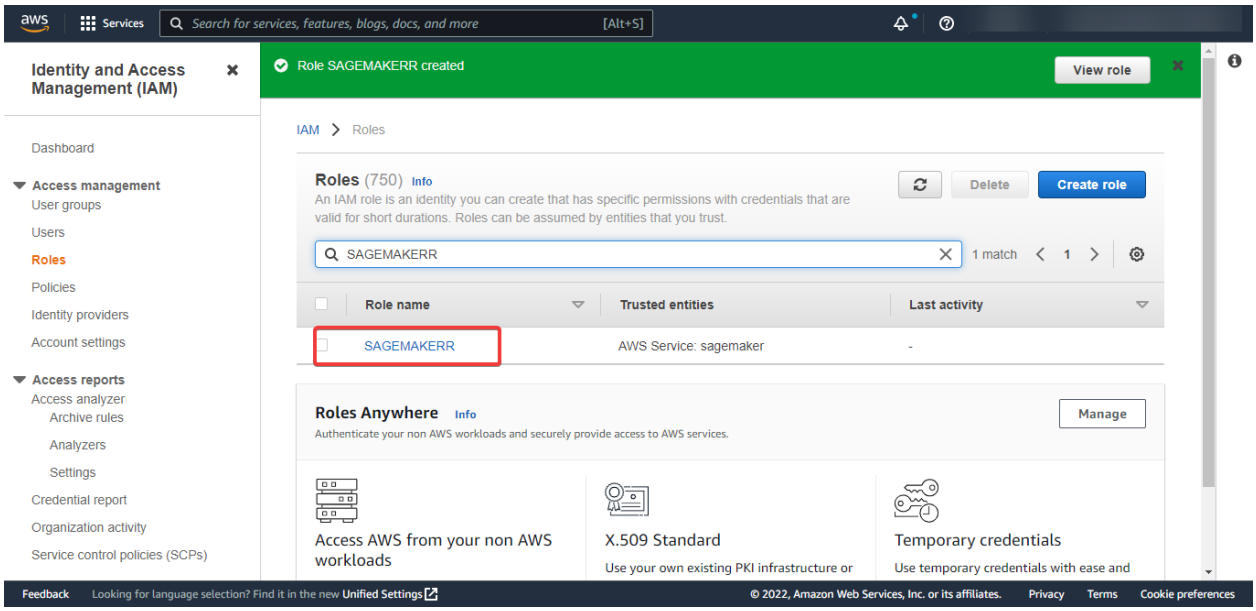
Create a **Role name** SAGEMAKER



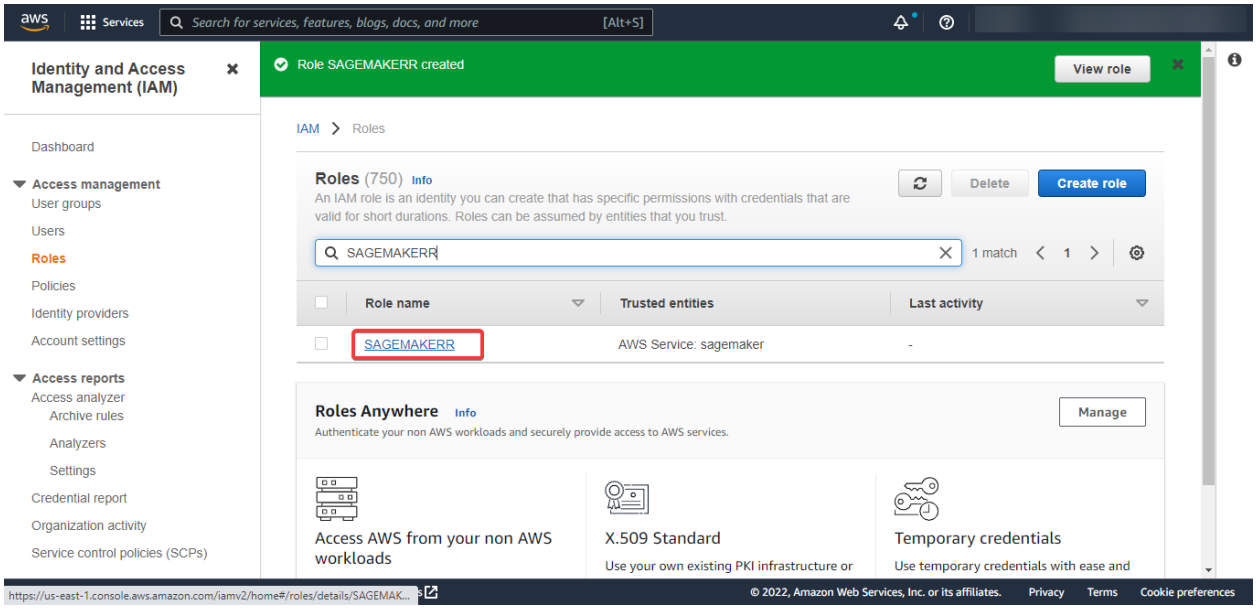
- Click **Create role**.



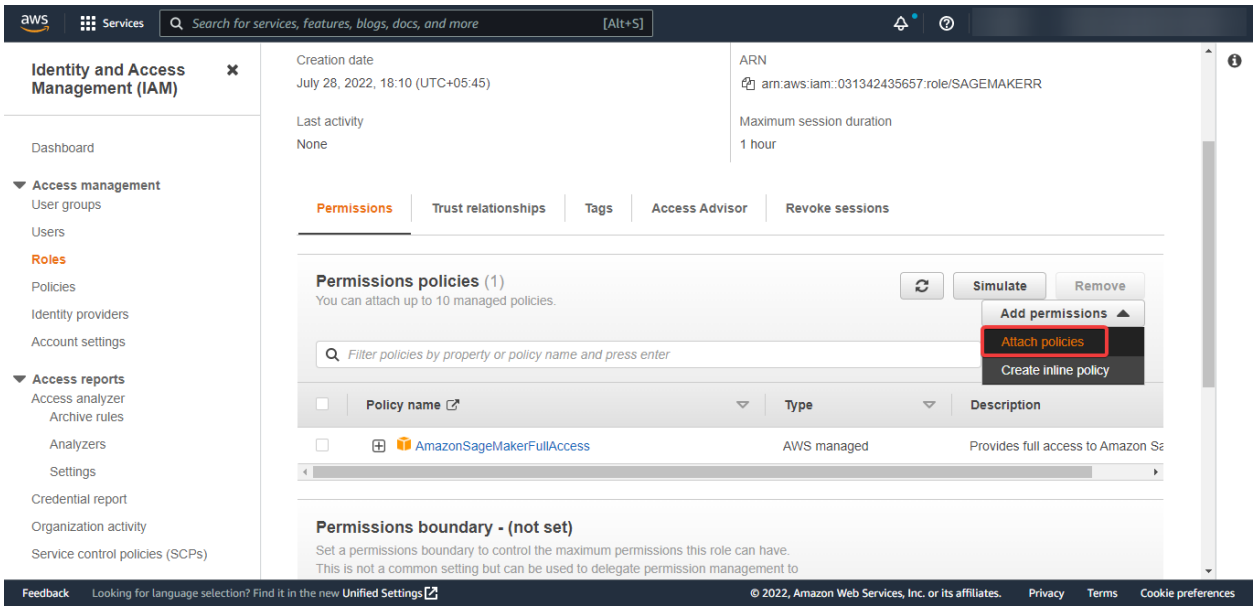
- Role SAGEMAKERR is created.



- Open your recently created sagemaker role **SAGEMAKERR** role.

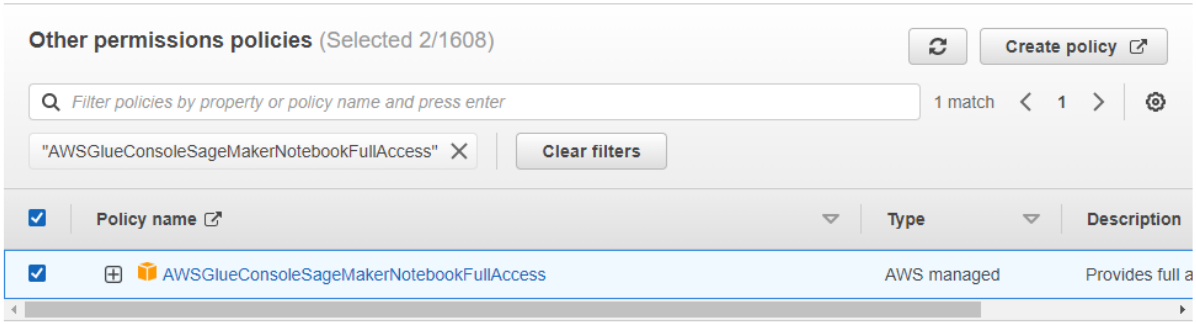


- Click **Add permission** and select **Attach policies**.

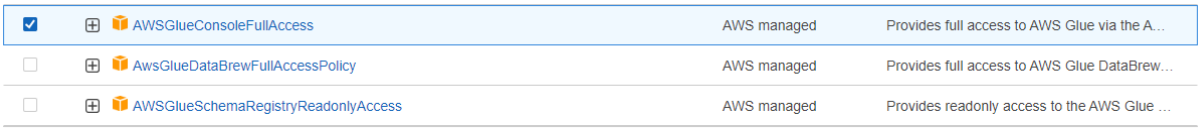


Following above procedure attach other policies such as

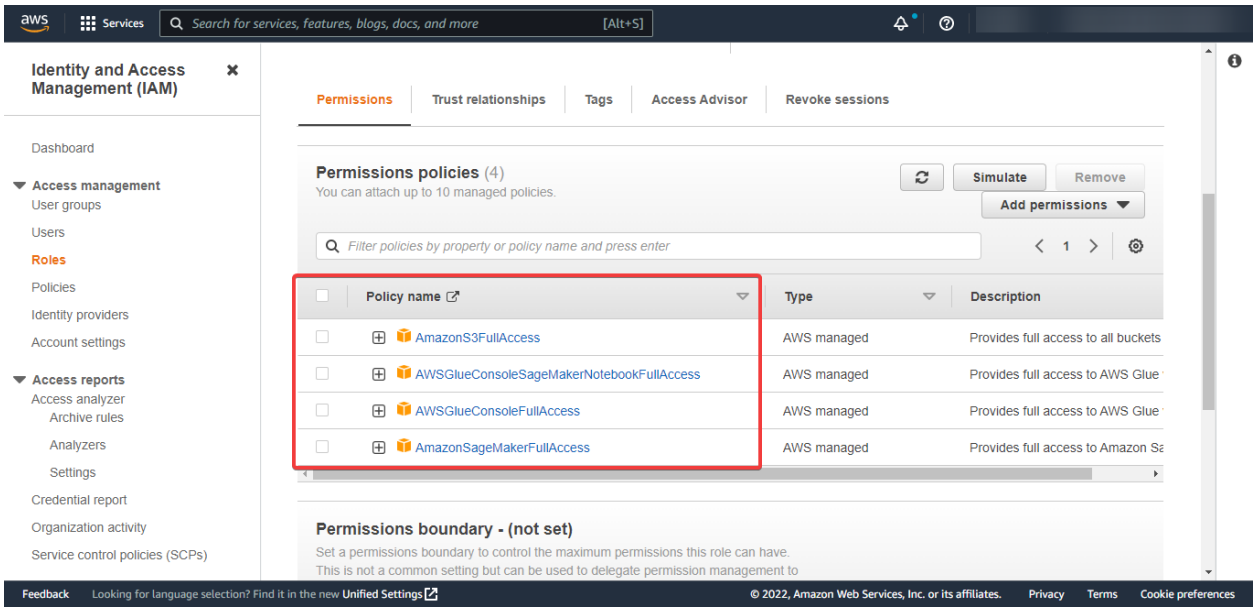
AWSGlueConsoleSagemakerNotebookFullAccess,
AmazonS3FullAccess,
AmazonSagemakerFullAccess
AWSGlueConsoleFullAccess


Other permissions policies (Selected 2/1608)
Create policy
Filter policies by property or policy name and press enter
1 match
1
"AWSGlueConsoleSageMakerNotebookFullAccess"
Clear filters
Policy name
Type
Description
AWSGlueConsoleSageMakerNotebookFullAccess
AWS managed
Provides full a


Cancel
Attach policies

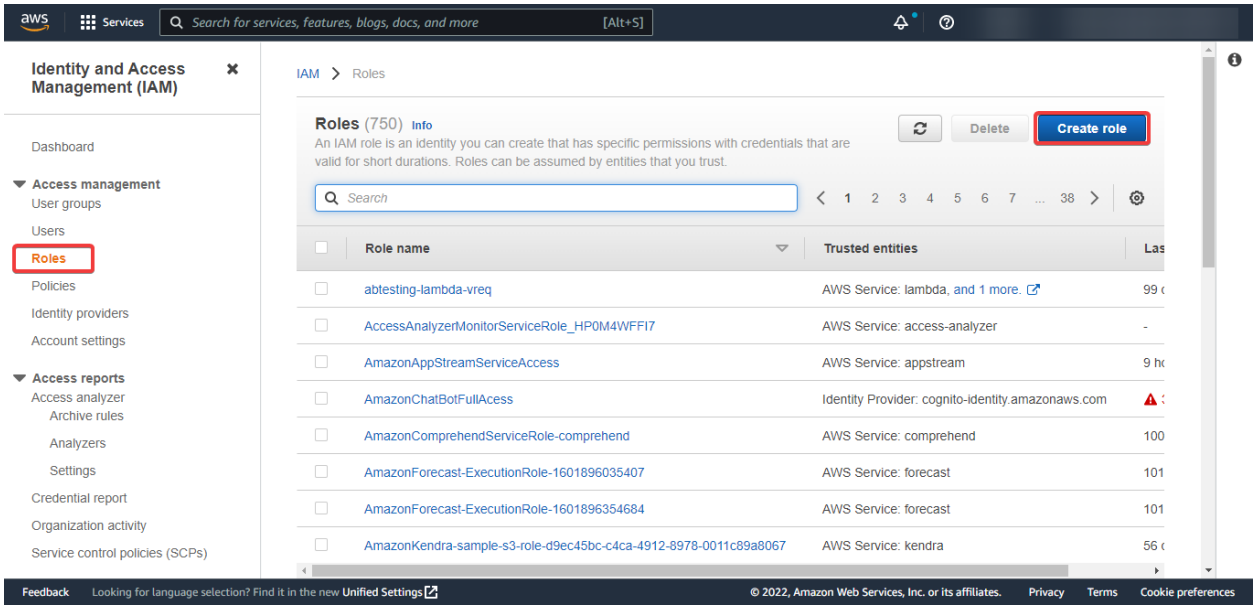

AWSGlueConsoleFullAccess
AWS managed
Provides full access to AWS Glue via the A...
AwsGlueDataBrewFullAccessPolicy
AWS managed
Provides full access to AWS Glue DataBrew...
AWSGlueSchemaRegistryReadonlyAccess
AWS managed
Provides readonly access to the AWS Glue ...


Cancel
Attach policies

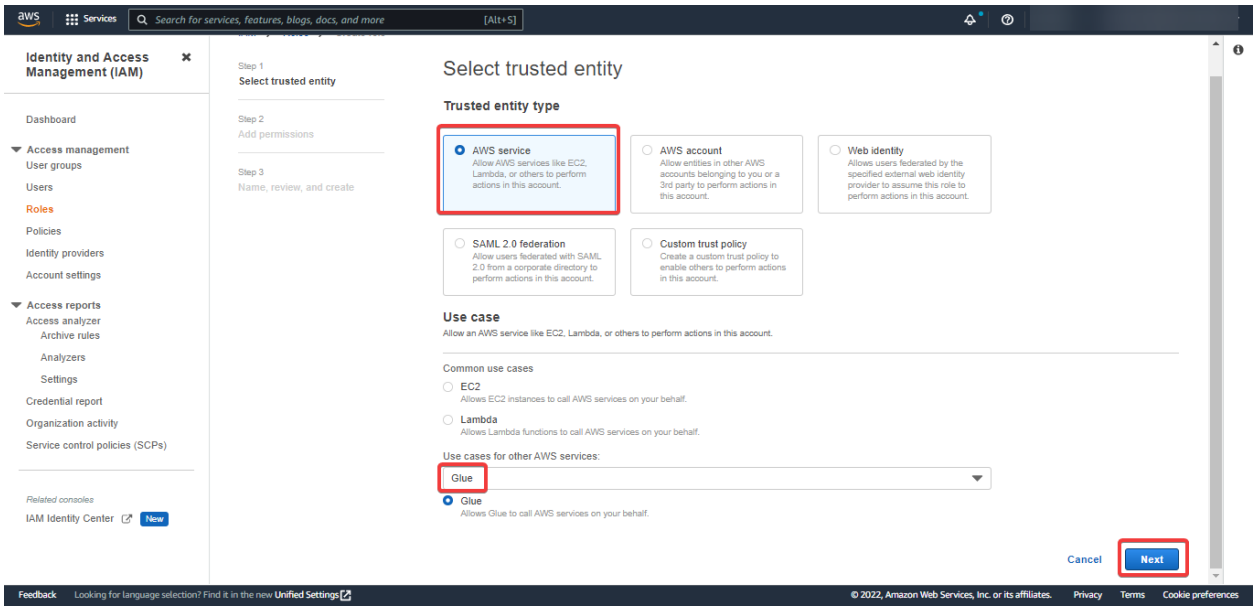

Services
Search for services, features, blogs, docs, and more
[Alt+S]
Identity and Access Management (IAM)
Dashboard
Access management
User groups
Users
Roles
Policies
Identity providers
Account settings
Access reports
Access analyzer
Archive rules
Analyzers
Settings
Credential report
Organization activity
Service control policies (SCPs)
Permissions
Trust relationships
Tags
Access Advisor
Revoke sessions
Permissions policies (4)
You can attach up to 10 managed policies.
Simulate
Remove
Add permissions
Filter policies by property or policy name and press enter
Policy name
Type
Description
AmazonS3FullAccess
AWS managed
Provides full access to all buckets
AWSGlueConsoleSageMakerNotebookFullAccess
AWS managed
Provides full access to AWS Glue
AWSGlueConsoleFullAccess
AWS managed
Provides full access to AWS Glue
AmazonSageMakerFullAccess
AWS managed
Provides full access to Amazon Sa
Permissions boundary - (not set)
Set a permissions boundary to control the maximum permissions this role can have.
This is not a common setting but can be used to delegate permission management to
Feedback
Looking for language selection? Find it in the new Unified Settings
© 2022, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences

Create IAM role for Glue: Crawler

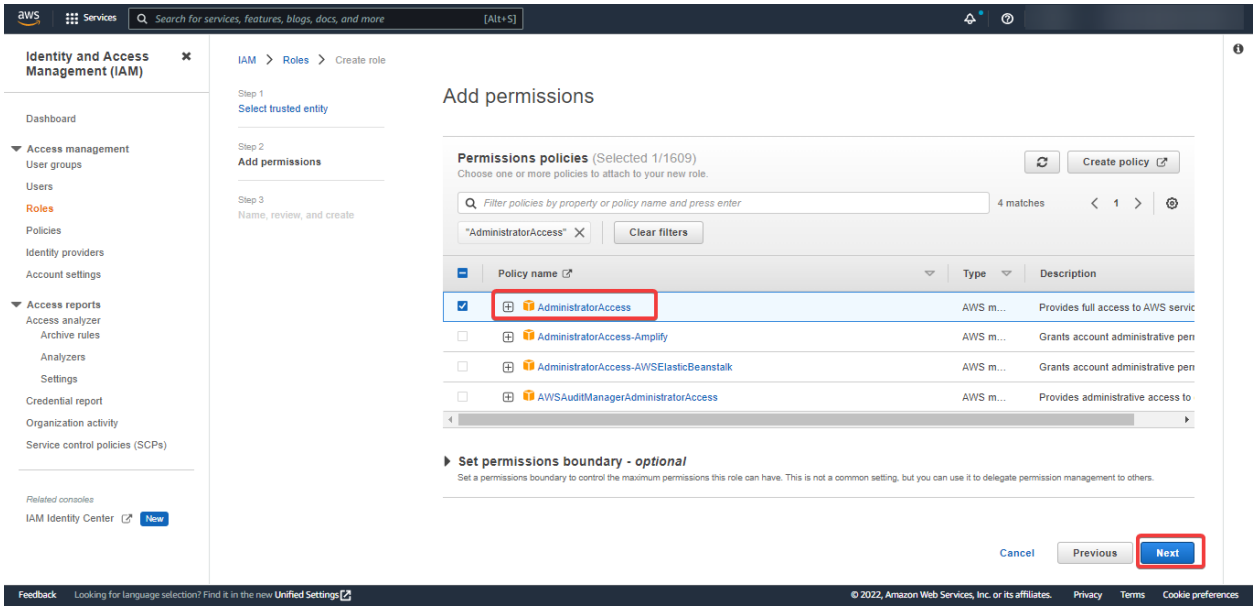
- Go to **IAM** dashboard
- Go to **Roles** and **create role**.



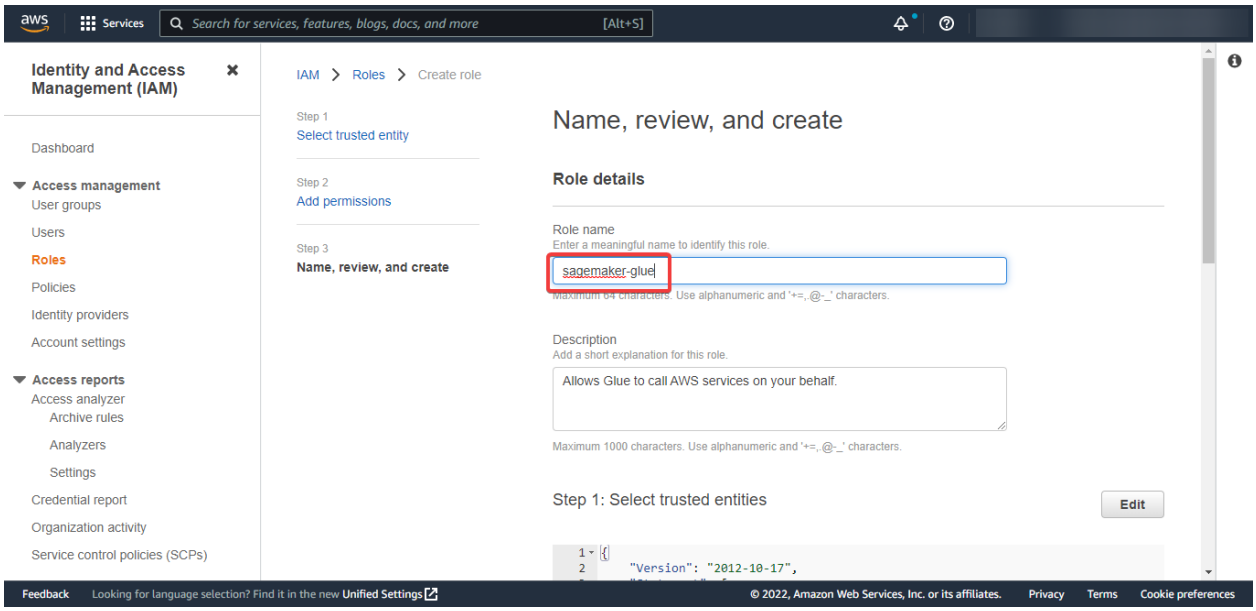
- Select **Glue** on **AWS service**
- Click **Next**



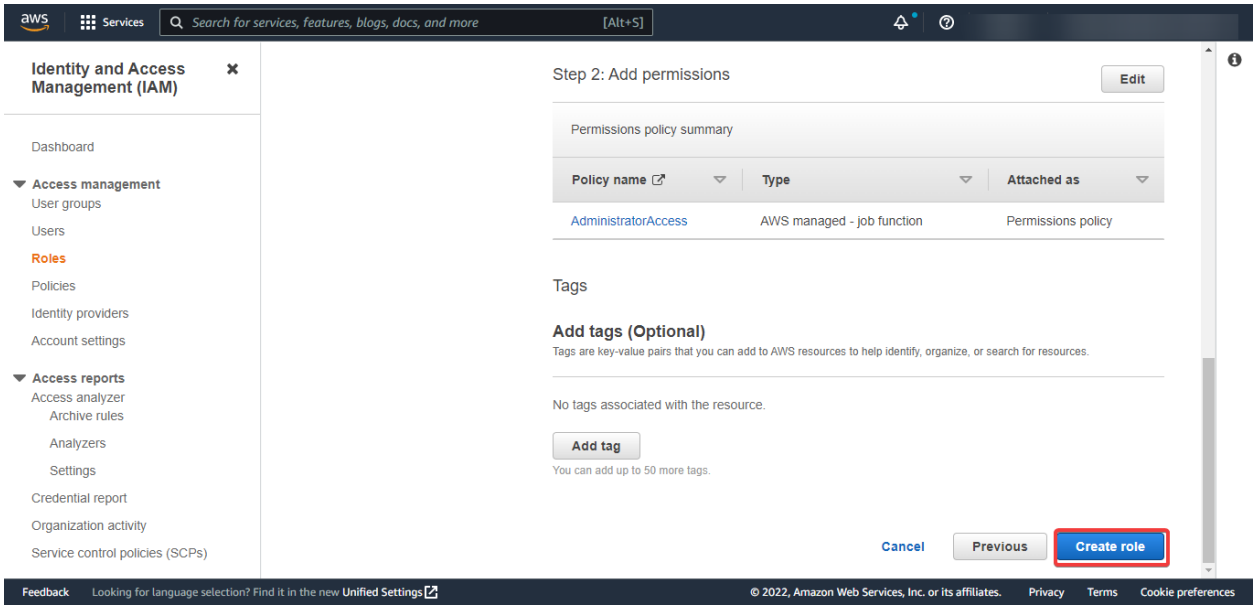
- Add **AdministerAccess** policy.
- Click **Next**



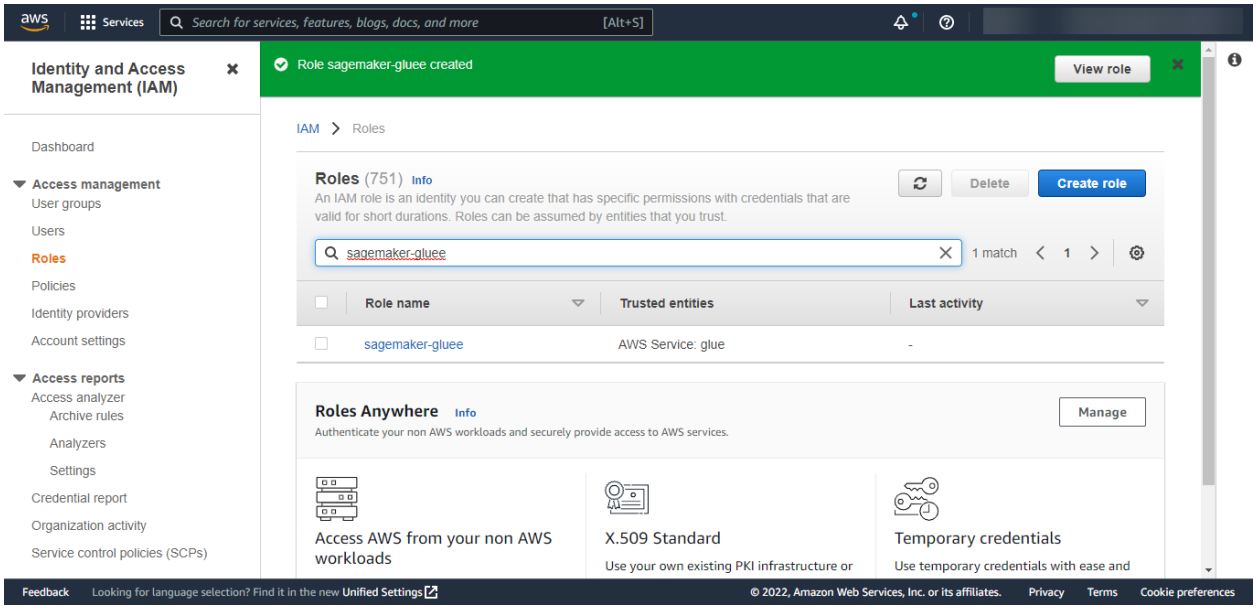
- Create **role name** sagemaker-glue



- Click **create role.**



- Sagemaker-IAM-role is created



CleanUp

Once you completed your task donot forget to follow this step in order to save from extra fee on using aws services.

- Delete your Glue Job in AWS Glue Console.
- Stop and delete notebook instance AWS Glue Console.
- Delete developer endpoint in AWS Glue Console.
- Delete database in AWS Glue and athena Console.

- Delete s3 bucket in S3 Management Console. If you created bucket with a different name then delete that one.
- Delete IAM Roles your IAM Management Console.