AWS FINAL PROJECT (Batch Time Analysis of Transactional Data)

[Author: Sai Kumar Reddy]

DESCRIPTION

Lenodo is a multinational e-commerce organization that sells products directly to consumers. The database administrator exports the data every night in a CSV file, but this export functionality is unused. Lenodo wants to use this data to uncover insights about the most-sold item and the countries where customers have bought this item.

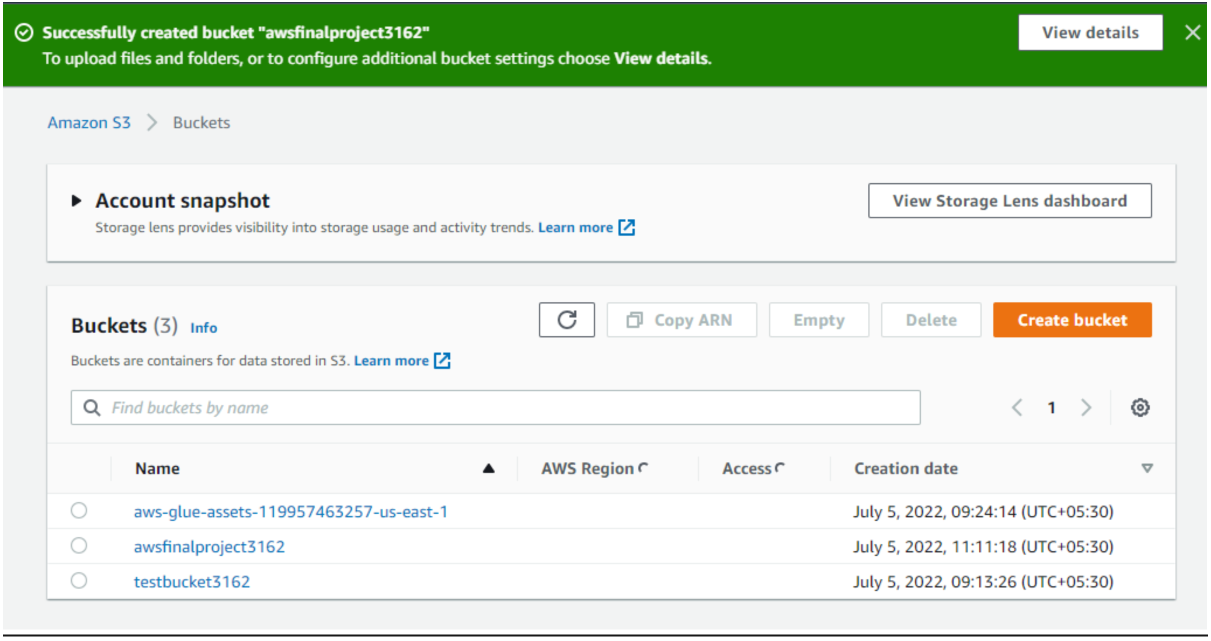
You are a data analytics consultant, and you're asked to provide valuable insights and statistics across products, brands, categories, segments to the marketing, product, sales, and procurement teams and inform them about which product has the highest amount of sales and which product and its marketing needs the most improvement. These statistics will help to run effective digital marketing campaigns. The scope of this project is limited to data engineering and analysis.

Objective:

To use AWS Big Data stack for data engineering to analyze transactions, uncover patterns, and share actionable insights.

Steps to perform:

1. **Create an S3 bucket with a unique name and upload the CSV file to the S3 bucket (ensure that the file is in UTF-8 format only)**



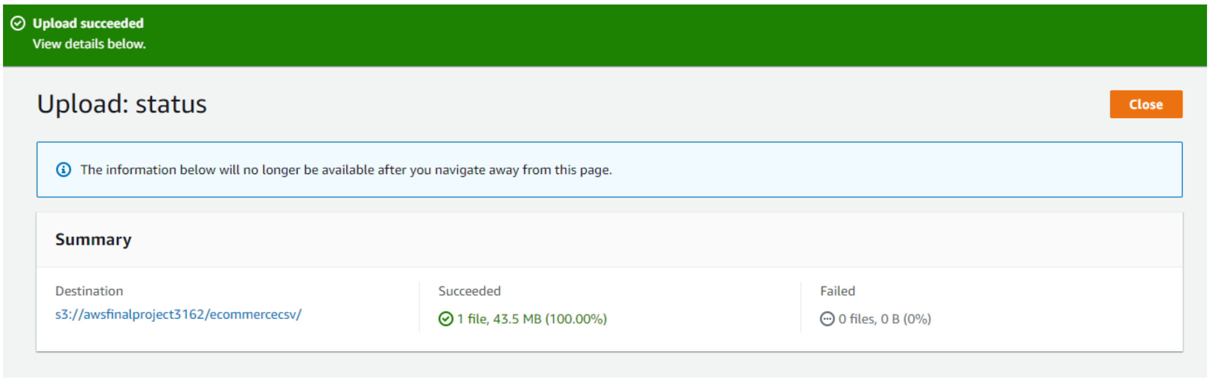
Successfully created bucket "awsfinalproject3162"
To upload files and folders, or to configure additional bucket settings choose **View details**.

Amazon S3 > Buckets

Account snapshot
Storage lens provides visibility into storage usage and activity trends. Learn more

Buckets (3) Info
Buckets are containers for data stored in S3. Learn more

Name	AWS Region	Access	Creation date
aws-glue-assets-119957463257-us-east-1			July 5, 2022, 09:24:14 (UTC+05:30)
awsfinalproject3162			July 5, 2022, 11:11:18 (UTC+05:30)
testbucket3162			July 5, 2022, 09:13:26 (UTC+05:30)



Upload succeeded
View details below.

Upload: status

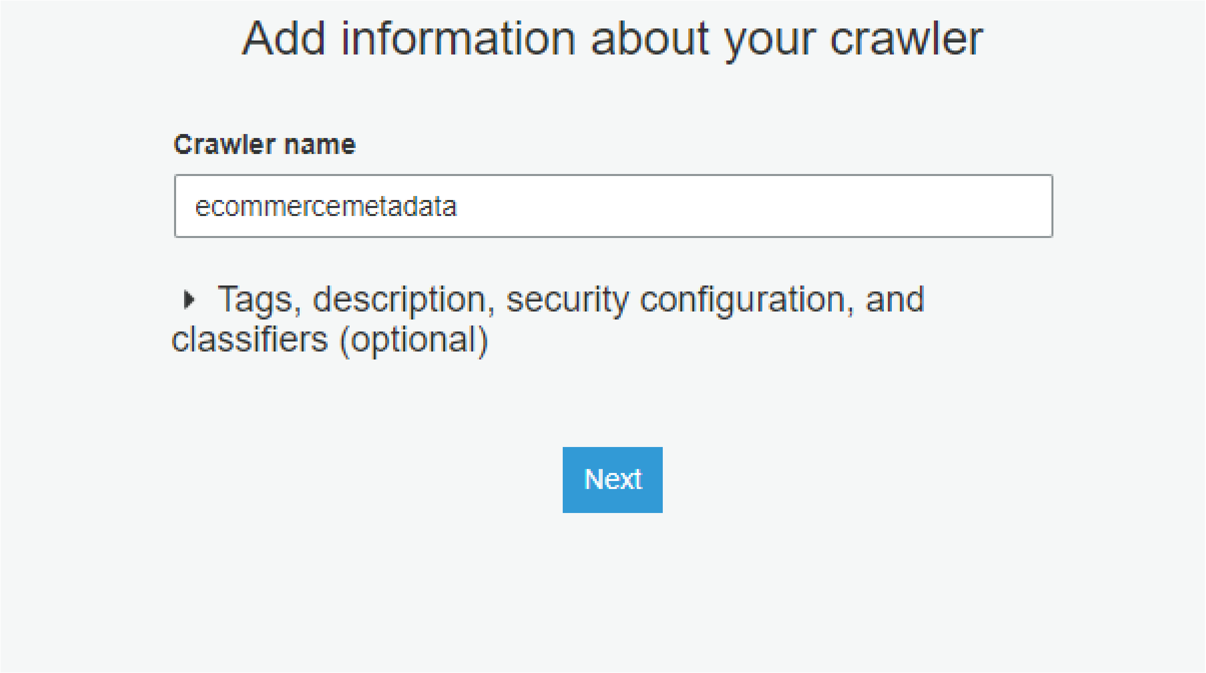
The information below will no longer be available after you navigate away from this page.

Summary

Destination	Succeeded	Failed
s3://awsfinalproject3162/ecommercecsv/	1 file, 43.5 MB (100.00%)	0 files, 0 B (0%)

2.Create a crawler to crawl the CSV data and generate a metadata catalog

- **STEP 1: Add Crawler**

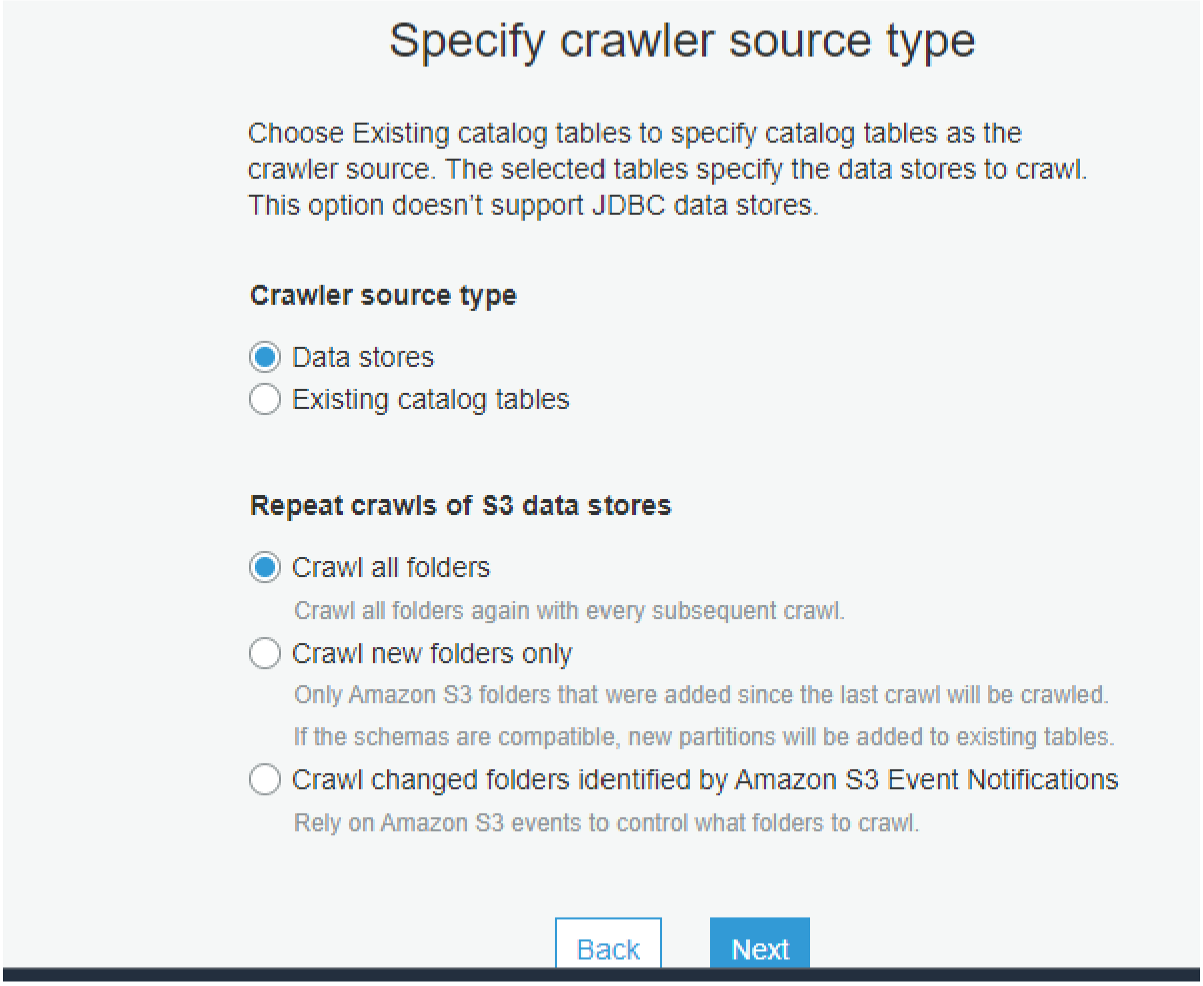


Add information about your crawler

Crawler name: ecommercemetadata

Tags, description, security configuration, and classifiers (optional)
Next

- **STEP 2: Specify Crawler Source Type.**



Specify crawler source type

Choose Existing catalog tables to specify catalog tables as the crawler source. The selected tables specify the data stores to crawl. This option doesn't support JDBC data stores.

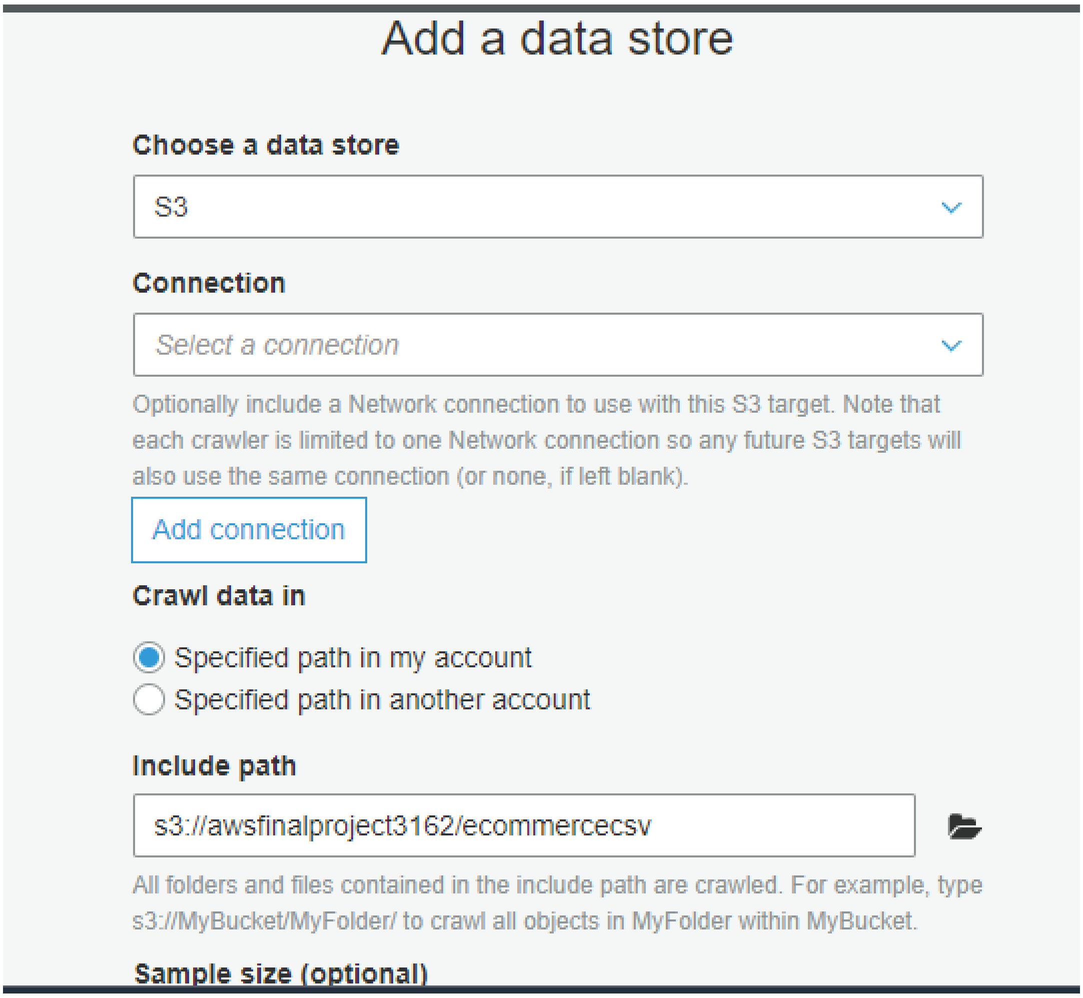
Crawler source type

● Data stores
○ Existing catalog tables

Repeat crawls of S3 data stores

● Crawl all folders
Crawl all folders again with every subsequent crawl.
○ Crawl new folders only
Only Amazon S3 folders that were added since the last crawl will be crawled. If the schemas are compatible, new partitions will be added to existing tables.
○ Crawl changed folders identified by Amazon S3 Event Notifications
Rely on Amazon S3 events to control what folders to crawl.

- **STEP 3: Add a data store**



Add a data store

Choose a data store

S3

Connection

Select a connection

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

Crawl data in

● Specified path in my account
○ Specified path in another account

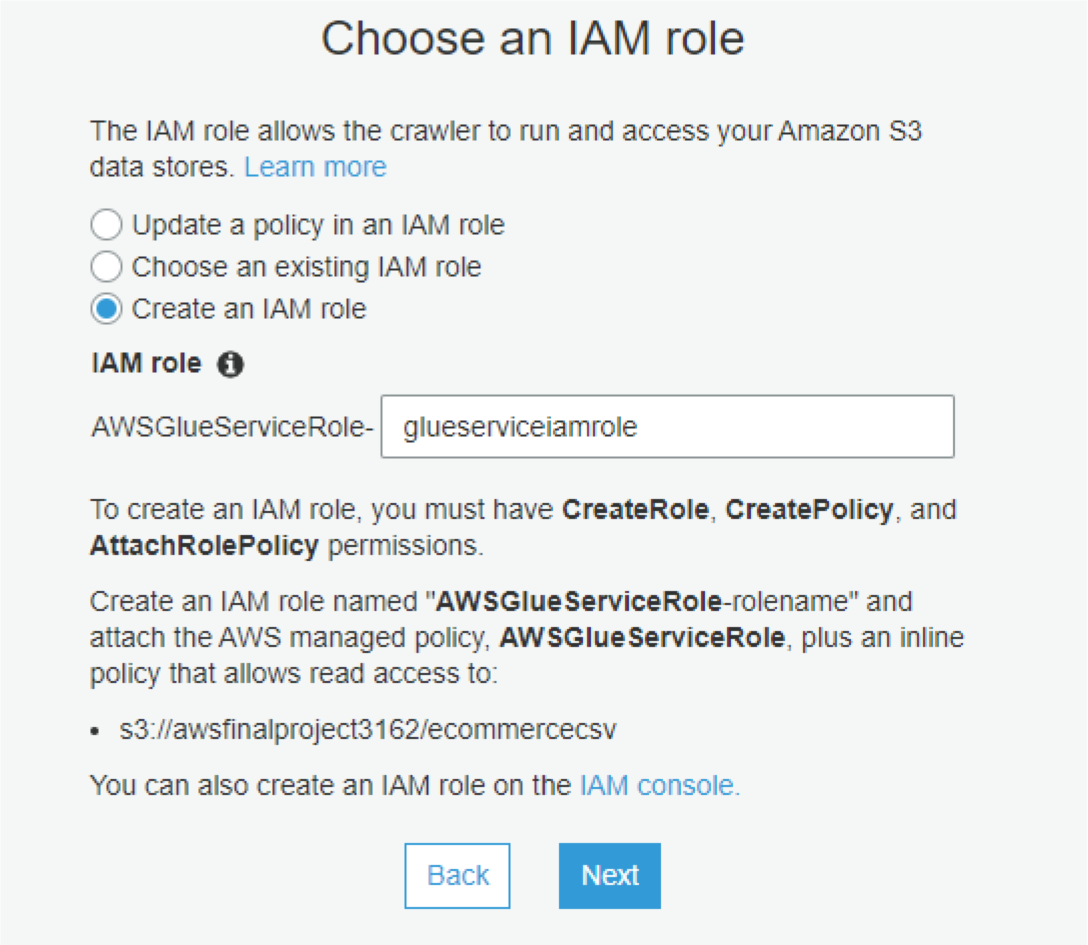
Include path

s3://awsfinalproject3162/ecommercecsv

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

Sample size (optional)

Step 4: choose I AM Role.



Choose an IAM role

The IAM role allows the crawler to run and access your Amazon S3 data stores. Learn more

○ Update a policy in an IAM role
○ Choose an existing IAM role
● Create an IAM role

IAM role

AWSGlueServiceRole- glueserviceiamrole

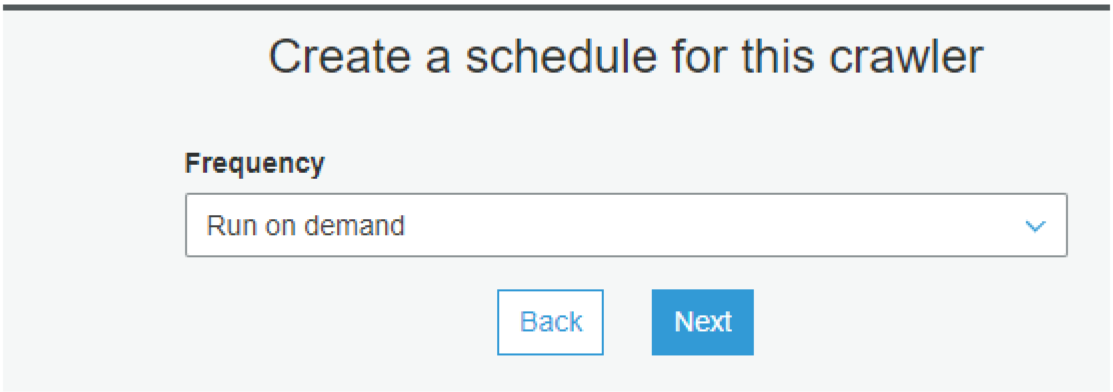
To create an IAM role, you must have **CreateRole**, **CreatePolicy**, and **AttachRolePolicy** permissions.

Create an IAM role named "**AWSGlueServiceRole**-rolename" and attach the AWS managed policy, **AWSGlueServiceRole**, plus an inline policy that allows read access to:

- s3://awsfinalproject3162/ecommercecsv

You can also create an IAM role on the IAM console.

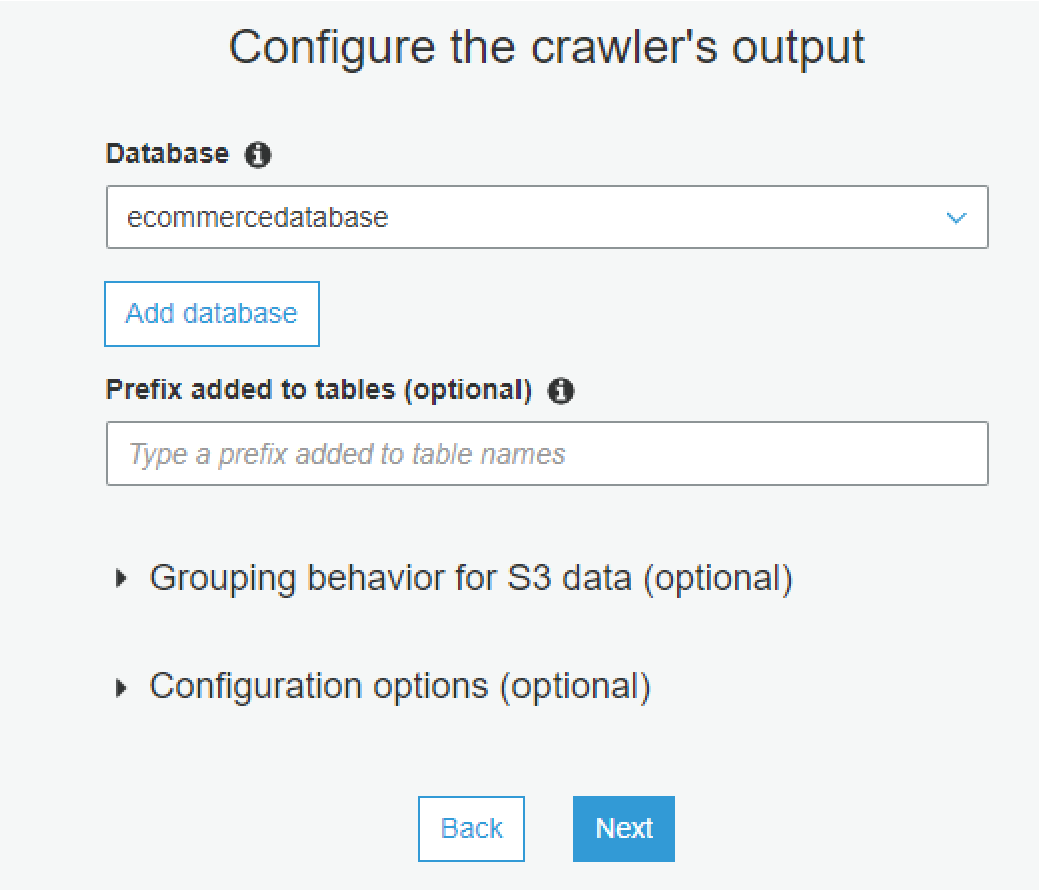
- **Step 5 : Create a schedule for this crawler.**



Create a schedule for this crawler

Frequency

Run on demand

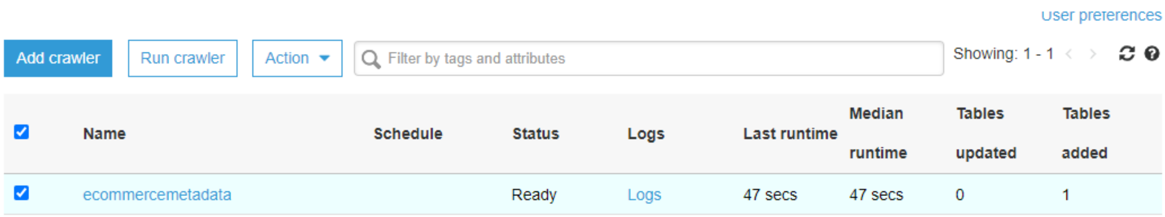


Configure the crawler's output

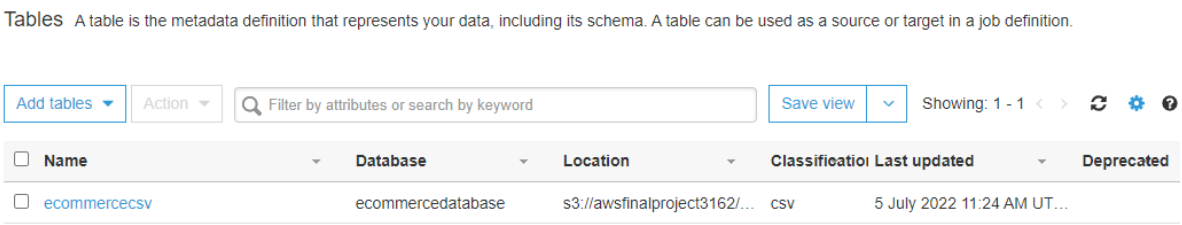
Database: ecommercedatabase

Prefix added to tables (optional): Type a prefix added to table names

▸ Grouping behavior for S3 data (optional)
▸ Configuration options (optional)
Back
Next



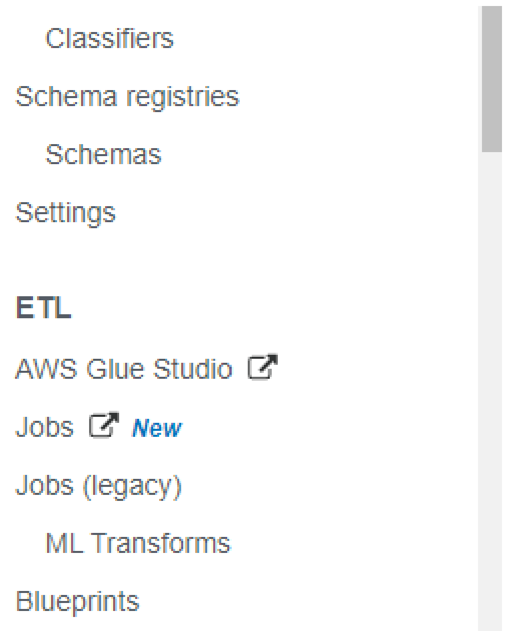
☑	Name	Schedule	Status	Logs	Last runtime	Median runtime	Tables updated	Tables added
☑	ecommercemetadata		Ready	Logs	47 secs	47 secs	0	1



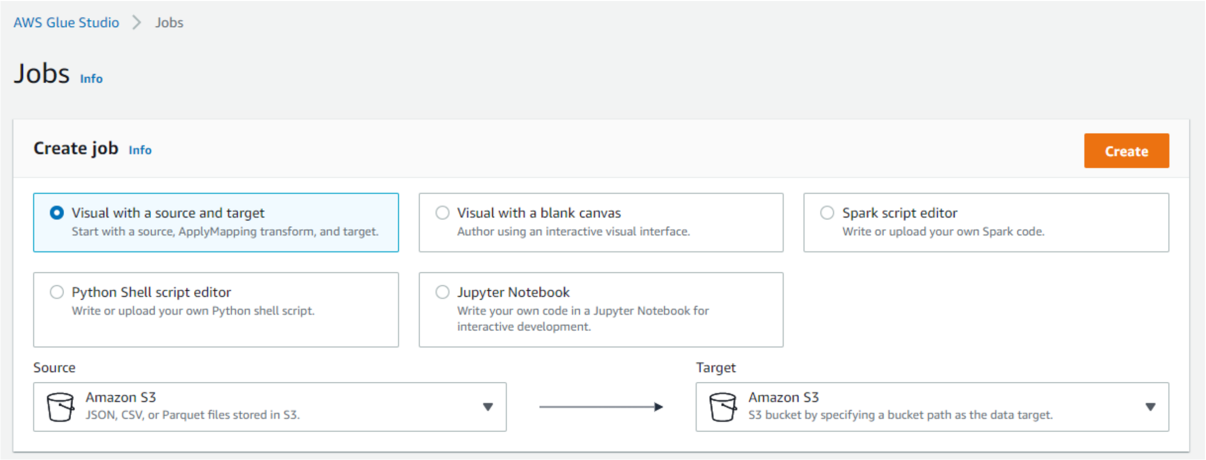
Tables A table is the metadata definition that represents your data, including its schema. A table can be used as a source or target in a job definition.

☐	Name	Database	Location	Classificatioı	Last updated	Deprecated
☐	ecommercecsv	ecommercedatabase	s3://awsfinalproject3162/...	csv	5 July 2022 11:24 AM UT...	

3.Create a Glue job to transform the data into the Parquet format as CSV is not optimal for data warehouse queries



Create a Glue ETL JOB.



● Visual with a source and target
Start with a source, ApplyMapping transform, and target.

○ Visual with a blank canvas
Author using an interactive visual interface.

○ Spark script editor
Write or upload your own Spark code.

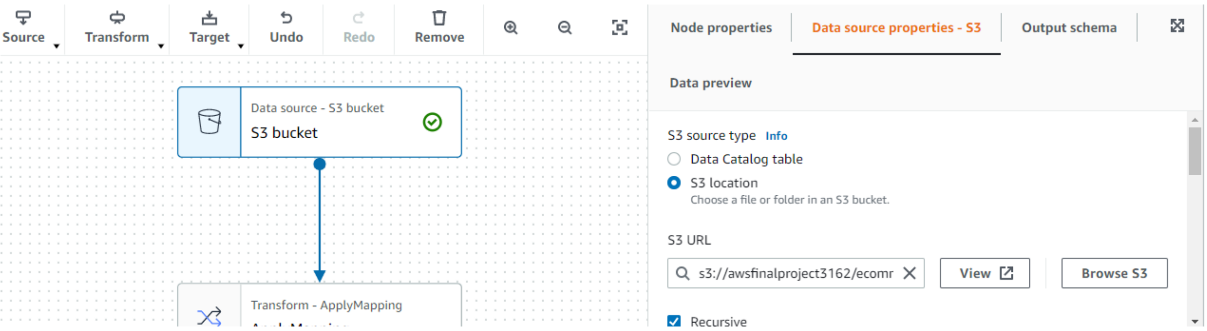
○ Python Shell script editor
Write or upload your own Python shell script.

○ Jupyter Notebook
Write your own code in a Jupyter Notebook for interactive development.

Source: Amazon S3 — JSON, CSV, or Parquet files stored in S3.

Target: Amazon S3 — S3 bucket by specifying a bucket path as the data target.

Source is S3 csvfile.

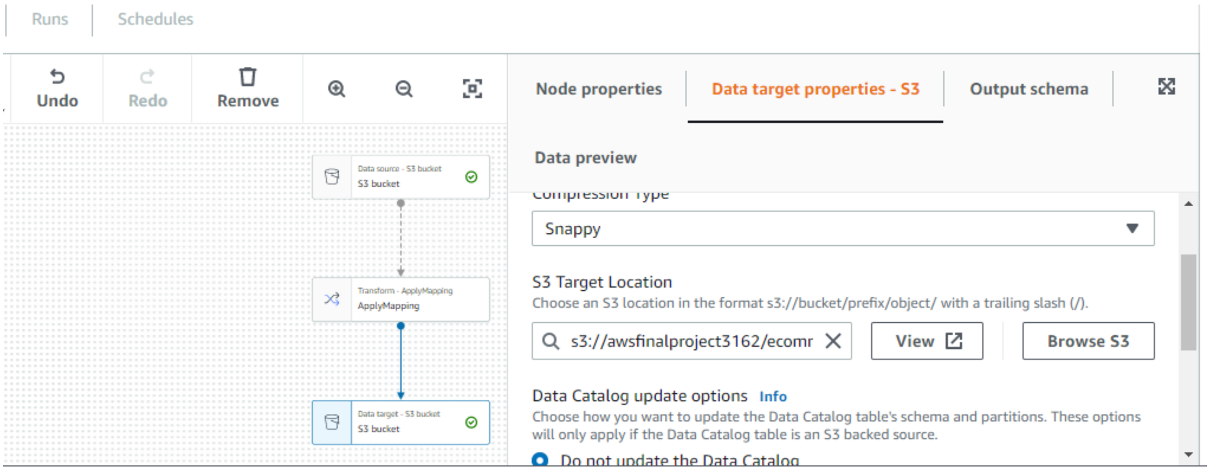


S3 source type Info
○ Data Catalog table
● S3 location
Choose a file or folder in an S3 bucket.

S3 URL: s3://awsfinalproject3162/ecomr

☑ Recursive

Destination is S3 and Parquet File with snappy compression Type.



Compression Type

Snappy

S3 Target Location

Choose an S3 location in the format s3://bucket/prefix/object/ with a trailing slash (/).

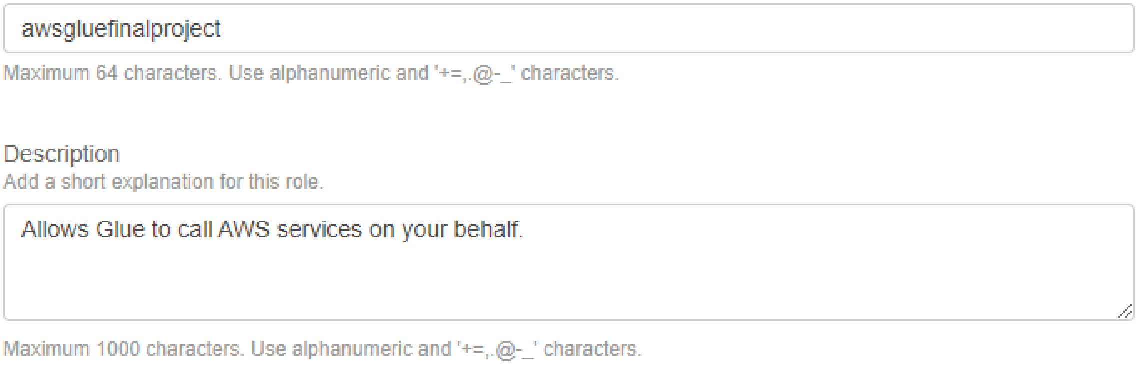
s3://awsfinalproject3162/ecomr

Data Catalog update options

Choose how you want to update the Data Catalog table's schema and partitions. These options will only apply if the Data Catalog table is an S3 backed source.

● Do not update the Data Catalog

Provide IAM ROLE like (AWS GLUE FULL ACCESS, S3)



awsgluefinalproject

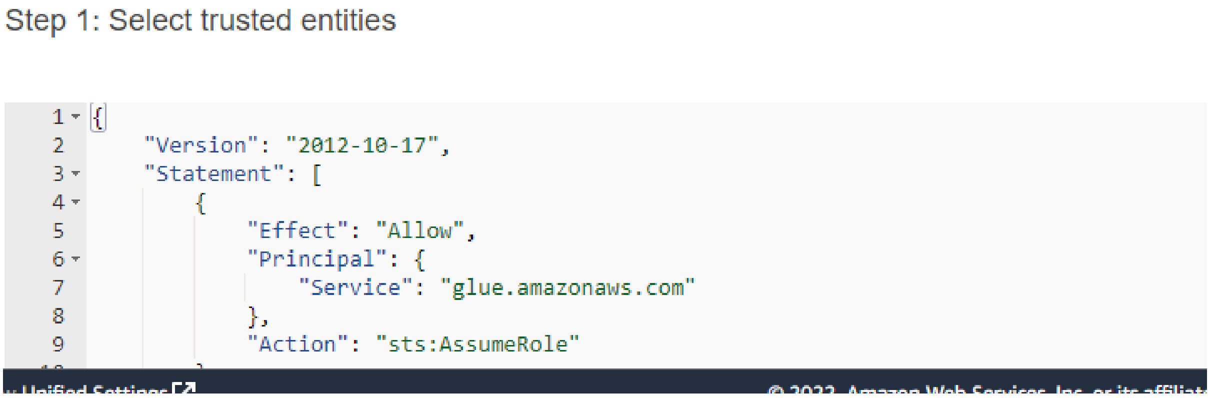
Maximum 64 characters. Use alphanumeric and '+=,.@-_' characters.

Description

Add a short explanation for this role.

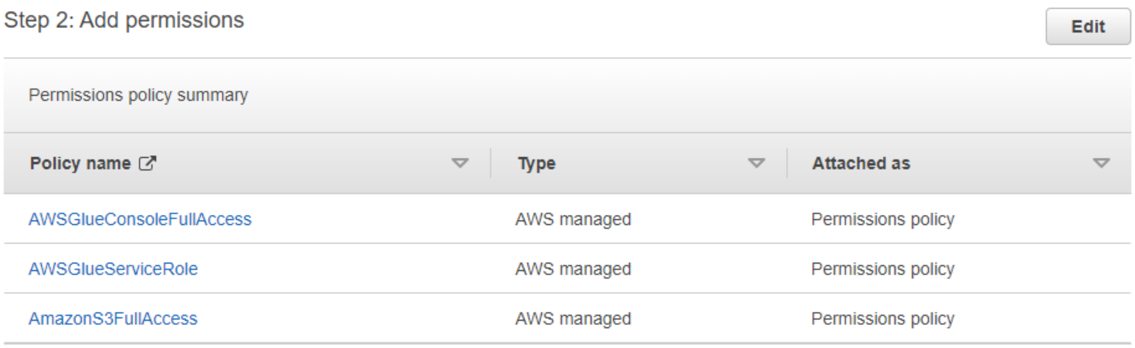
Allows Glue to call AWS services on your behalf.

Maximum 1000 characters. Use alphanumeric and '+=,.@-_' characters.



Step 1: Select trusted entities

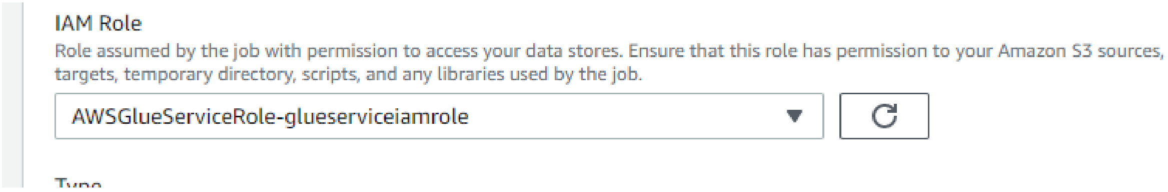
```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Principal": {
                "Service": "glue.amazonaws.com"
            },
            "Action": "sts:AssumeRole"
```



Step 2: Add permissions

Permissions policy summary

Policy name	Type	Attached as
AWSGlueConsoleFullAccess	AWS managed	Permissions policy
AWSGlueServiceRole	AWS managed	Permissions policy
AmazonS3FullAccess	AWS managed	Permissions policy

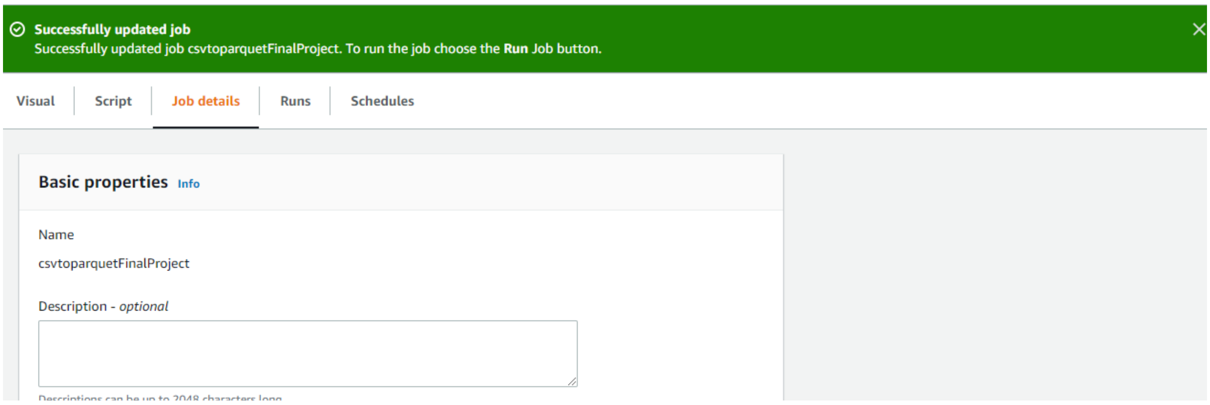


IAM Role

Role assumed by the job with permission to access your data stores. Ensure that this role has permission to your Amazon S3 sources, targets, temporary directory, scripts, and any libraries used by the job.

AWSGlueServiceRole-glueserviceiamrole

Type



Successfully updated job

Successfully updated job csvtoparquetFinalProject. To run the job choose the **Run** Job button.

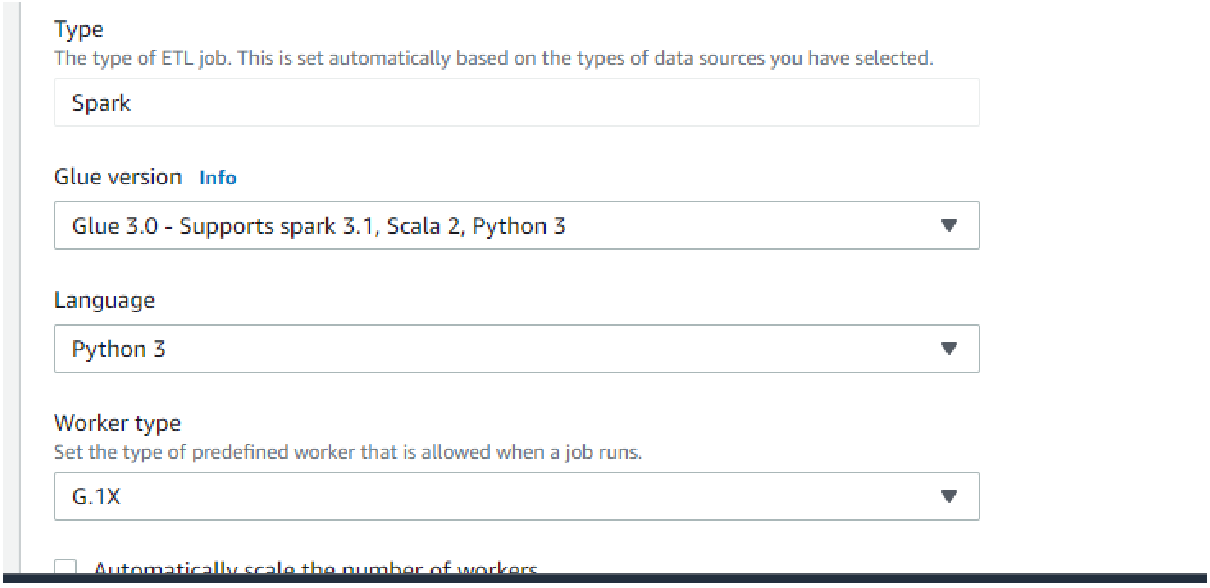
Basic properties

Name

csvtoparquetFinalProject

Description - *optional*

Descriptions can be up to 2048 characters long.



Type

The type of ETL job. This is set automatically based on the types of data sources you have selected.

Spark

Glue version

Glue 3.0 - Supports spark 3.1, Scala 2, Python 3

Language

Python 3

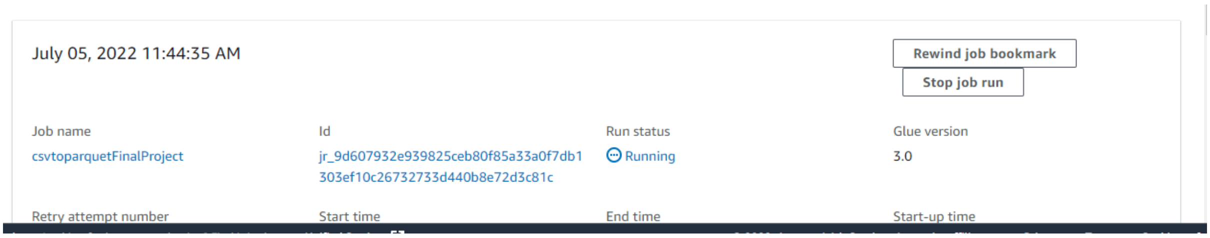
Worker type

Set the type of predefined worker that is allowed when a job runs.

G.1X

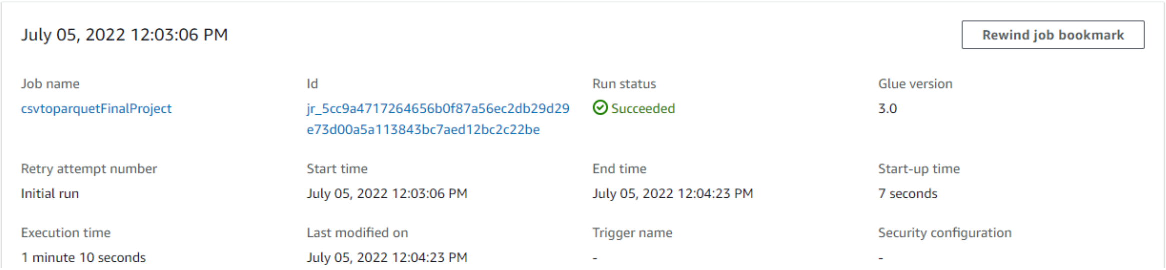
□ Automatically scale the number of workers

We need to run the Glue ETL JOB.



July 05, 2022 11:44:35 AM

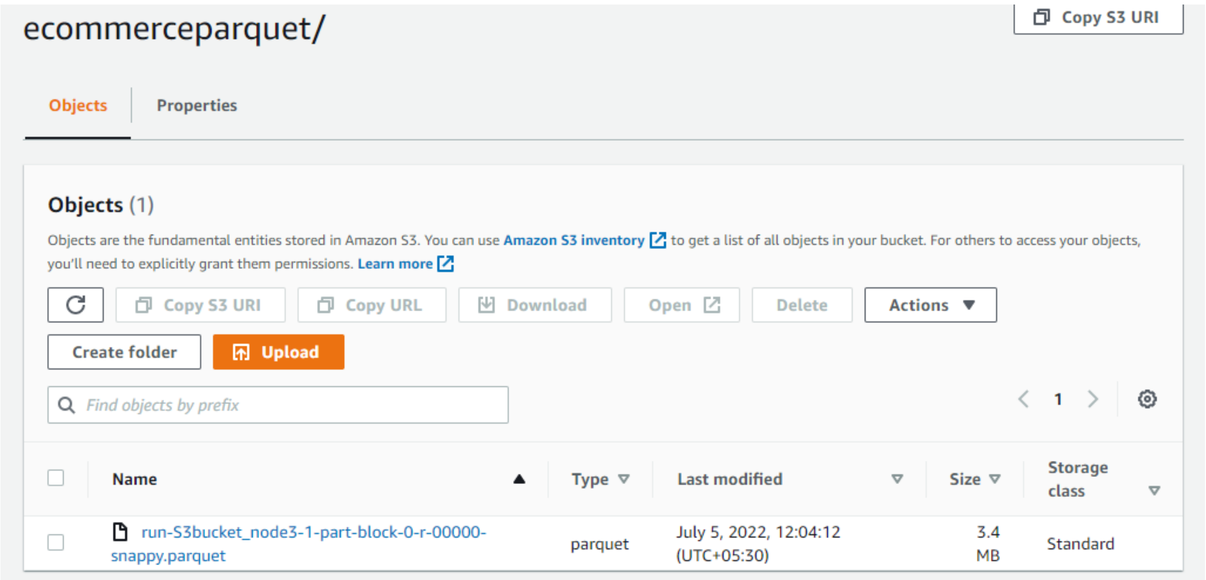
Job name	Id	Run status	Glue version
csvtoparquetFinalProject	jr_9d607932e939825ceb80f85a33a0f7db1303ef10c26732733d440b8e72d3c81c	Running	3.0
Retry attempt number	Start time	End time	Start-up time



July 05, 2022 12:03:06 PM

Job name	Id	Run status	Glue version
csvtoparquetFinalProject	jr_5cc9a4717264656b0f87a56ec2db29d29e73d00a5a113843bc7aed12bc2c22be	✓ Succeeded	3.0
Retry attempt number	**Start time**	**End time**	**Start-up time**
Initial run	July 05, 2022 12:03:06 PM	July 05, 2022 12:04:23 PM	7 seconds
Execution time	**Last modified on**	**Trigger name**	**Security configuration**
1 minute 10 seconds	July 05, 2022 12:04:23 PM	-	-

Output File in the form of



ecommerceparquet/

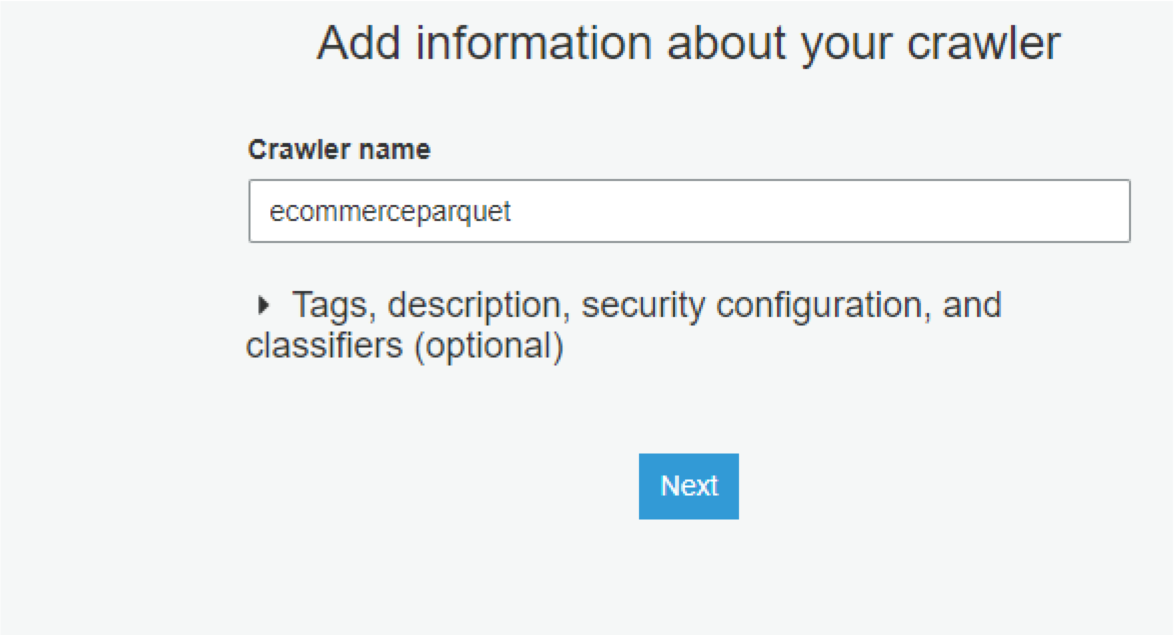
Objects (1)

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 ▽
□	run-S3bucket_node3-1-part-block-0-r-00000-snappy.parquet	parquet	July 5, 2022, 12:04:12 (UTC+05:30)	3.4 MB	Standard

4.Add another crawler to crawl the Parquet data files to generate the metadata catalog of the Parquet file in order to query it with Athena

- **Add another Crawler Name (ecommerceparquet)**



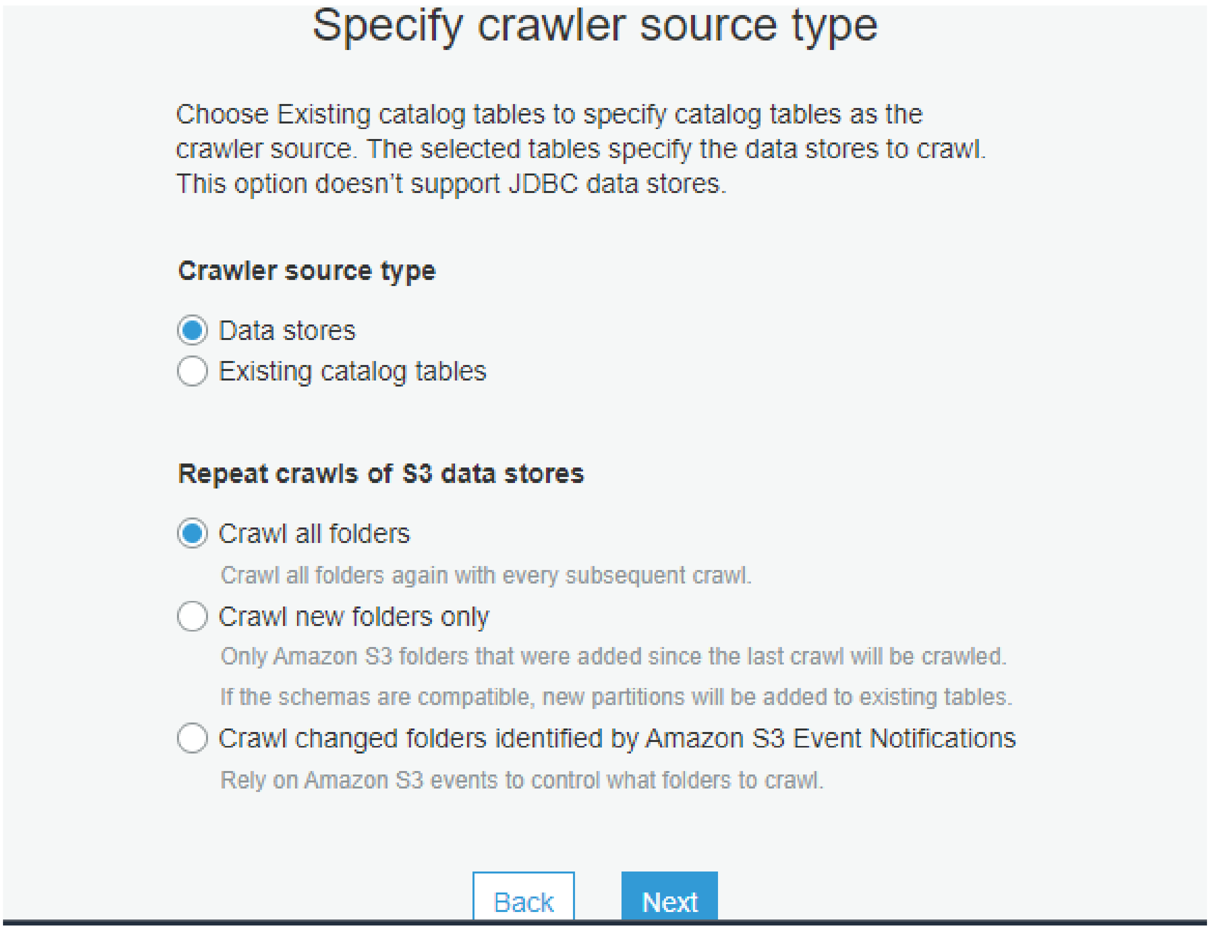
Add information about your crawler

Crawler name

ecommerceparquet

▸ Tags, description, security configuration, and classifiers (optional)
Next

- **Specify Crawler Source Type**



Specify crawler source type

Choose Existing catalog tables to specify catalog tables as the crawler source. The selected tables specify the data stores to crawl. This option doesn't support JDBC data stores.

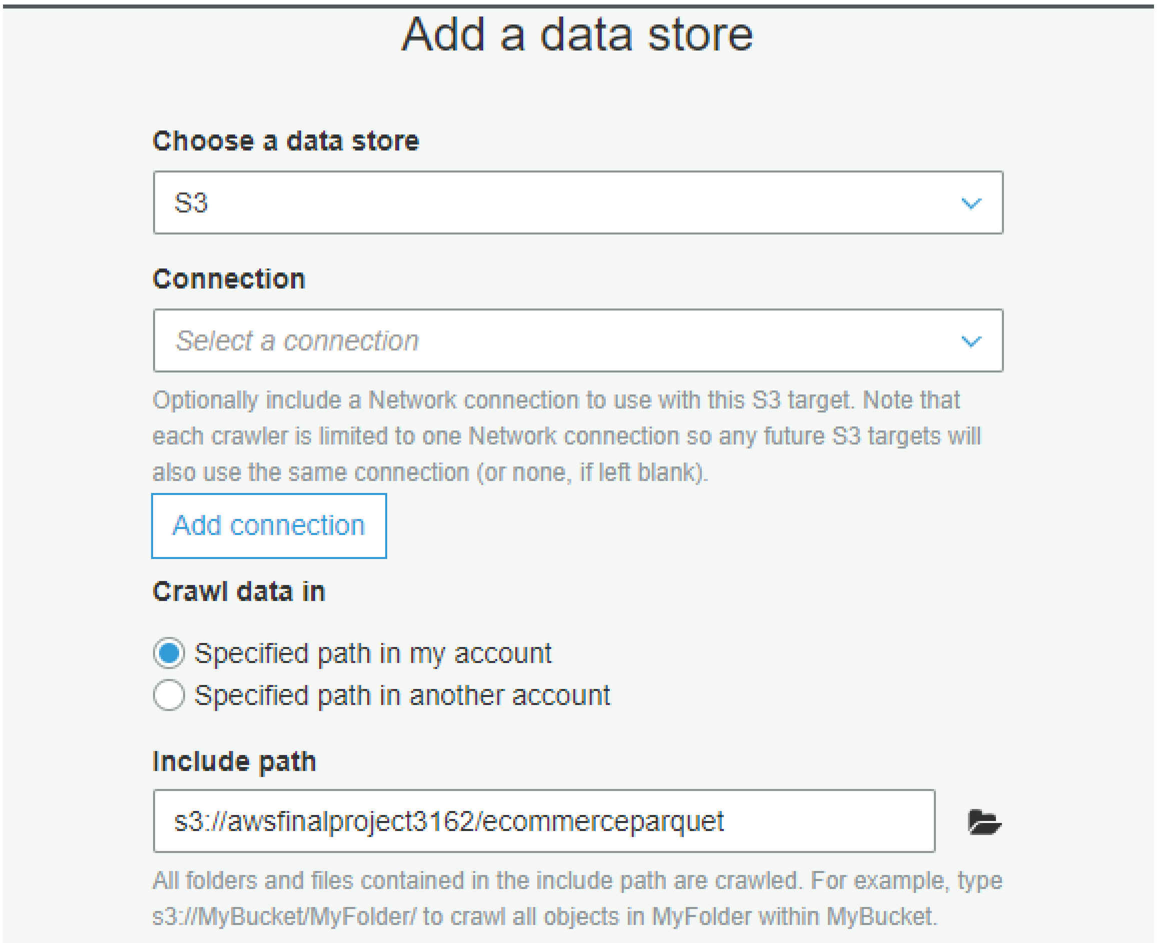
Crawler source type

● Data stores
○ Existing catalog tables

Repeat crawls of S3 data stores

● Crawl all folders
Crawl all folders again with every subsequent crawl.
○ Crawl new folders only
Only Amazon S3 folders that were added since the last crawl will be crawled.
If the schemas are compatible, new partitions will be added to existing tables.
○ Crawl changed folders identified by Amazon S3 Event Notifications
Rely on Amazon S3 events to control what folders to crawl.

- **Add a data store**



Add a data store

Choose a data store

S3

Connection

Select a connection

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

Crawl data in

● Specified path in my account

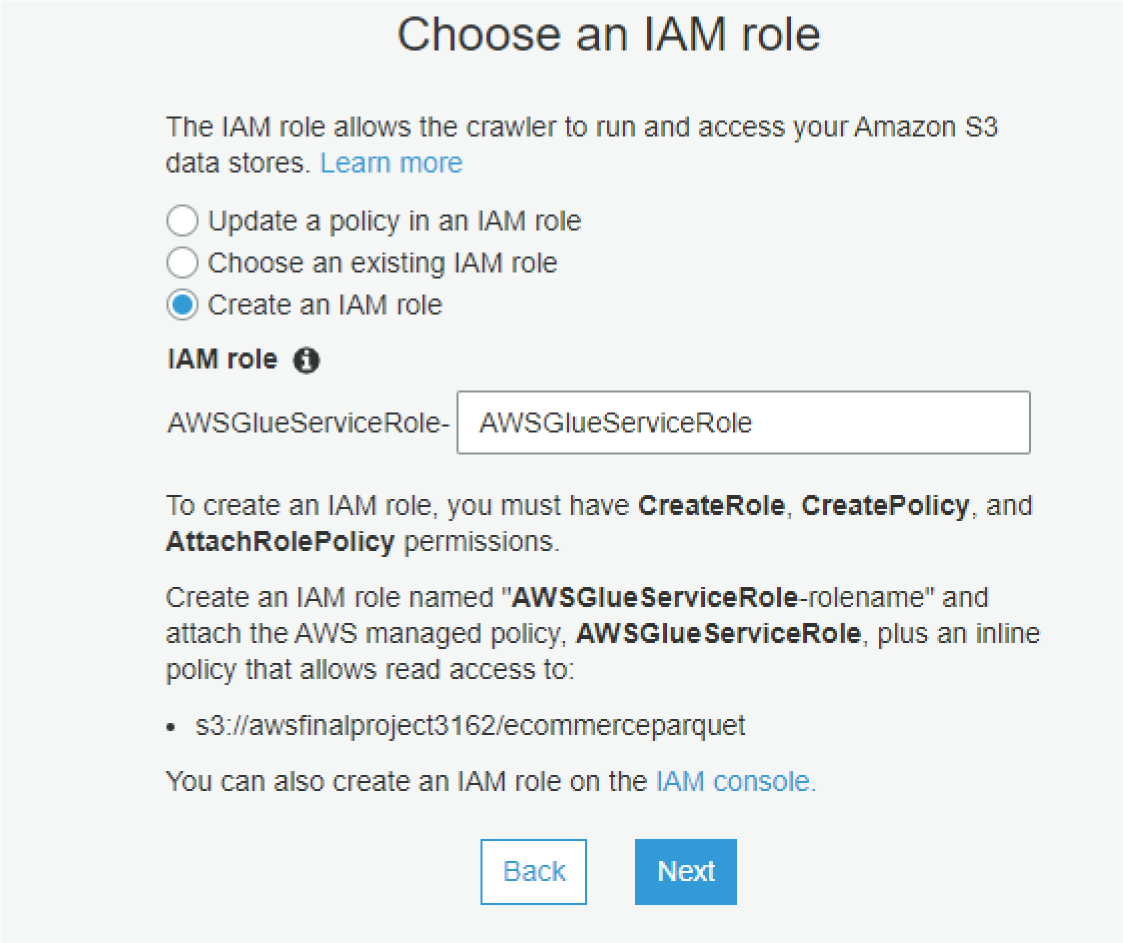
○ Specified path in another account

Include path

s3://awsfinalproject3162/ecommerceparquet

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

- **Choose an IAM role**



Choose an IAM role

The IAM role allows the crawler to run and access your Amazon S3 data stores. Learn more

○ Update a policy in an IAM role
○ Choose an existing IAM role
● Create an IAM role

IAM role

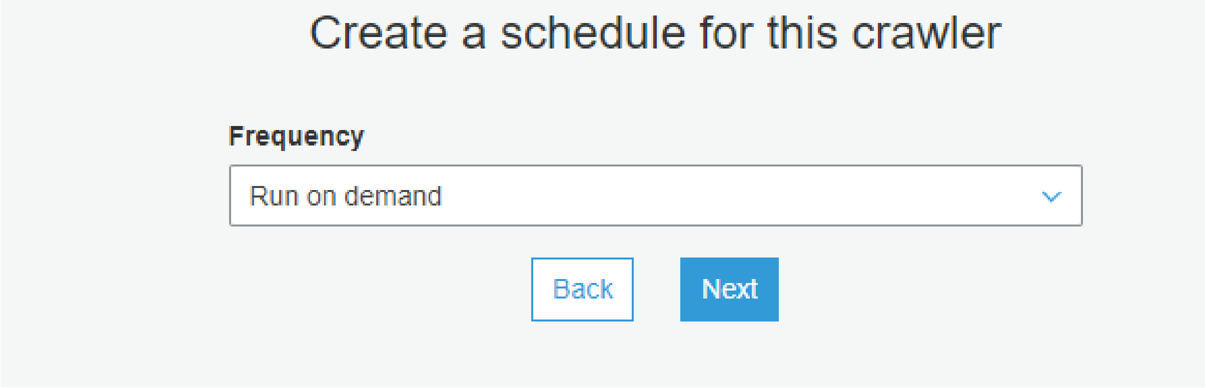
AWSGlueServiceRole- AWSGlueServiceRole

To create an IAM role, you must have **CreateRole**, **CreatePolicy**, and **AttachRolePolicy** permissions.

Create an IAM role named "**AWSGlueServiceRole**-rolename" and attach the AWS managed policy, **AWSGlueServiceRole**, plus an inline policy that allows read access to:

- s3://awsfinalproject3162/ecommerceparquet

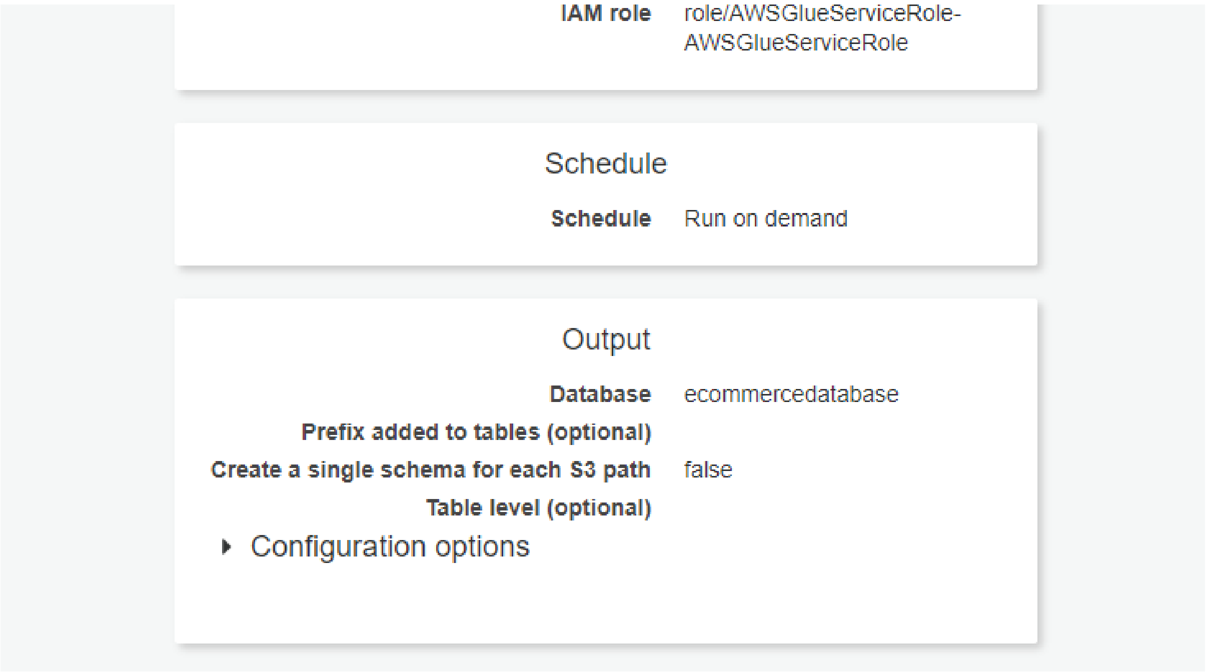
You can also create an IAM role on the IAM console.



Create a schedule for this crawler

Frequency

Run on demand



IAM role	role/AWSGlueServiceRole-AWSGlueServiceRole

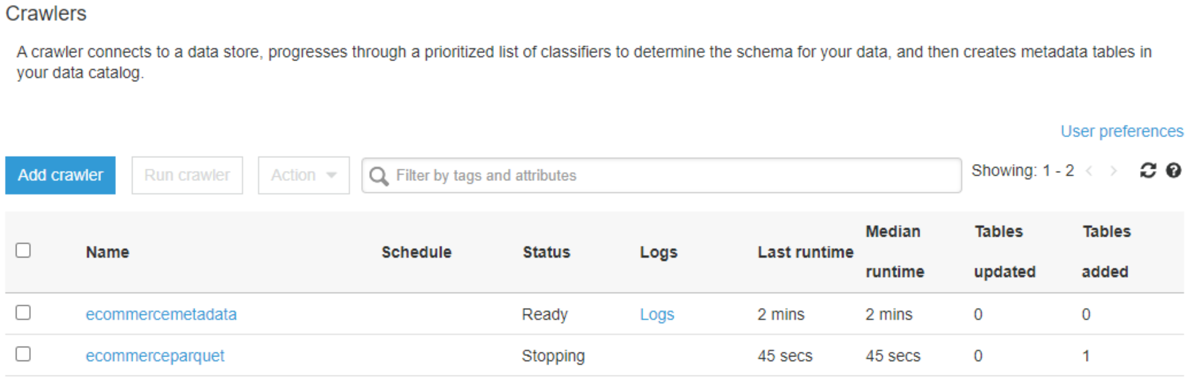
Schedule

Schedule	Run on demand

Output

Database	ecommercedatabase
Prefix added to tables (optional)	
Create a single schema for each S3 path	false
Table level (optional)	

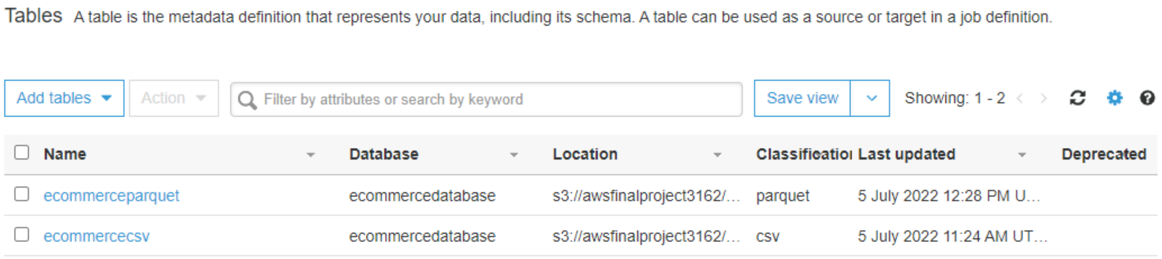
➢ Run the GLUE ETL JOB.



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.

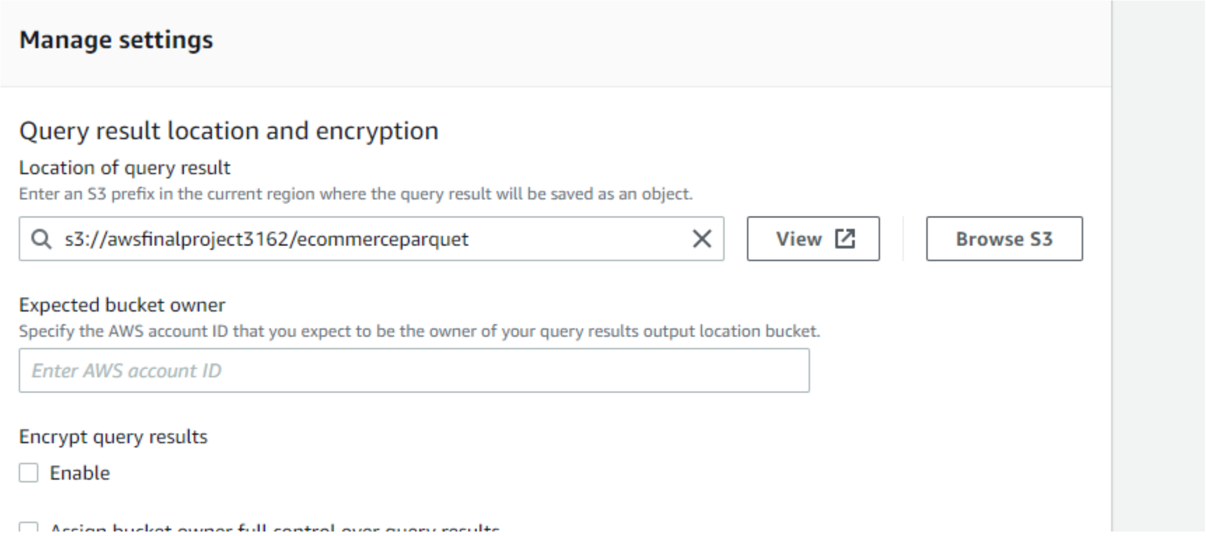
☐	Name	Schedule	Status	Logs	Last runtime	Median runtime	Tables updated	Tables added
☐	ecommercemetadata		Ready	Logs	2 mins	2 mins	0	0
☐	ecommerceparquet		Stopping		45 secs	45 secs	0	1



Tables A table is the metadata definition that represents your data, including its schema. A table can be used as a source or target in a job definition.

☐	Name	Database	Location	Classificatioı	Last updated	Deprecated
☐	ecommerceparquet	ecommercedatabase	s3://awsfinalproject3162/...	parquet	5 July 2022 12:28 PM U...	
☐	ecommercecsv	ecommercedatabase	s3://awsfinalproject3162/...	csv	5 July 2022 11:24 AM UT...	

5.Query the data to identify the best-selling item and countries where customers have bought the most-sold item using Athena



Manage settings

Query result location and encryption

Location of query result

Enter an S3 prefix in the current region where the query result will be saved as an object.

s3://awsfinalproject3162/ecommerceparquet

Expected bucket owner

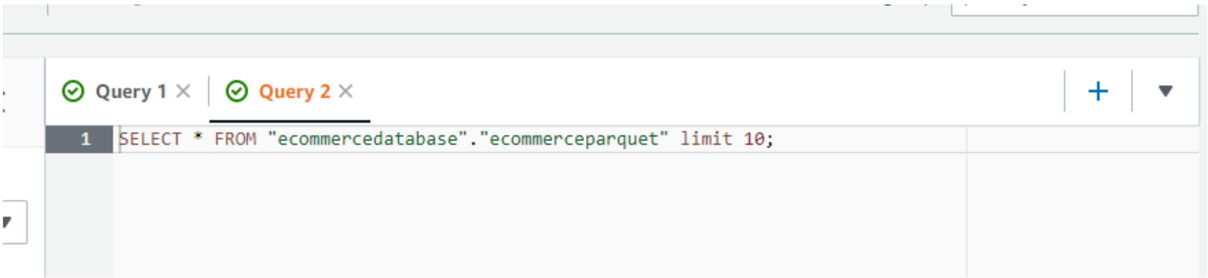
Specify the AWS account ID that you expect to be the owner of your query results output location bucket.

Enter AWS account ID

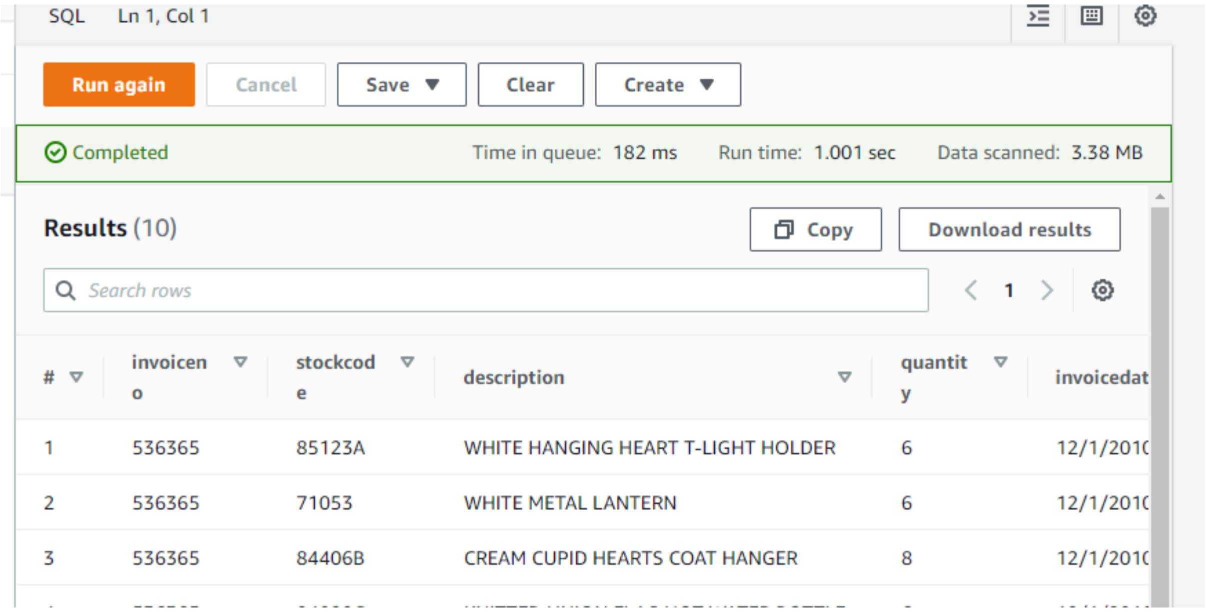
Encrypt query results

□ Enable

➢ **Check the table ecommerceparquet in athena.**



```
SELECT * FROM "ecommercedatabase"."ecommerceparquet" limit 10;
```

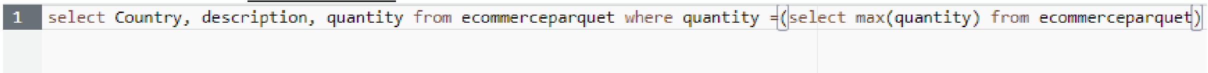


Completed   Time in queue: 182 ms   Run time: 1.001 sec   Data scanned: 3.38 MB

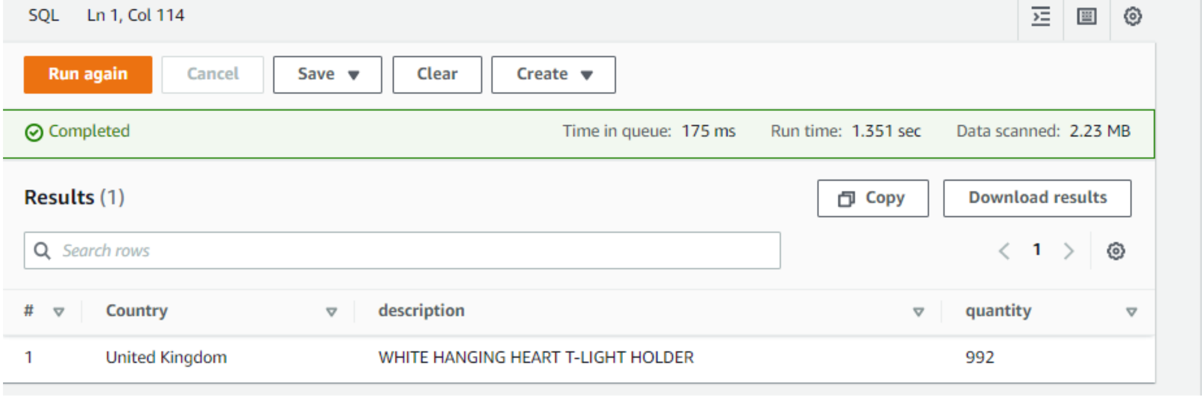
Results (10)

#	invoiceno	stockcode	description	quantity	invoicedat
1	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	12/1/2010
2	536365	71053	WHITE METAL LANTERN	6	12/1/2010
3	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	12/1/2010

➢ Final Query:



```
select Country, description, quantity from ecommerceparquet where quantity =(select max(quantity) from ecommerceparquet)
```



Completed   Time in queue: 175 ms   Run time: 1.351 sec   Data scanned: 2.23 MB

Results (1)

#	Country	description	quantity
1	United Kingdom	WHITE HANGING HEART T-LIGHT HOLDER	992