Task Name: AWS Data Pipeline Setup for ETL Processing

Problem Statement

Organizations often struggle with managing and processing large volumes of data efficiently due to manual processes and a lack of automation. This leads to increased time for data handling and potential errors, hindering timely insights.

Objectives

1. **Automate Data Ingestion**: Use AWS Glue to automatically retrieve and catalog data from S3.
2. **Transform Data**: Create a Glue notebook to clean and restructure the data.
3. **Load Data**: Save the processed data back to a designated S3 bucket.
4. **Catalog Data**: Maintain an organized schema using AWS Glue for easy access.

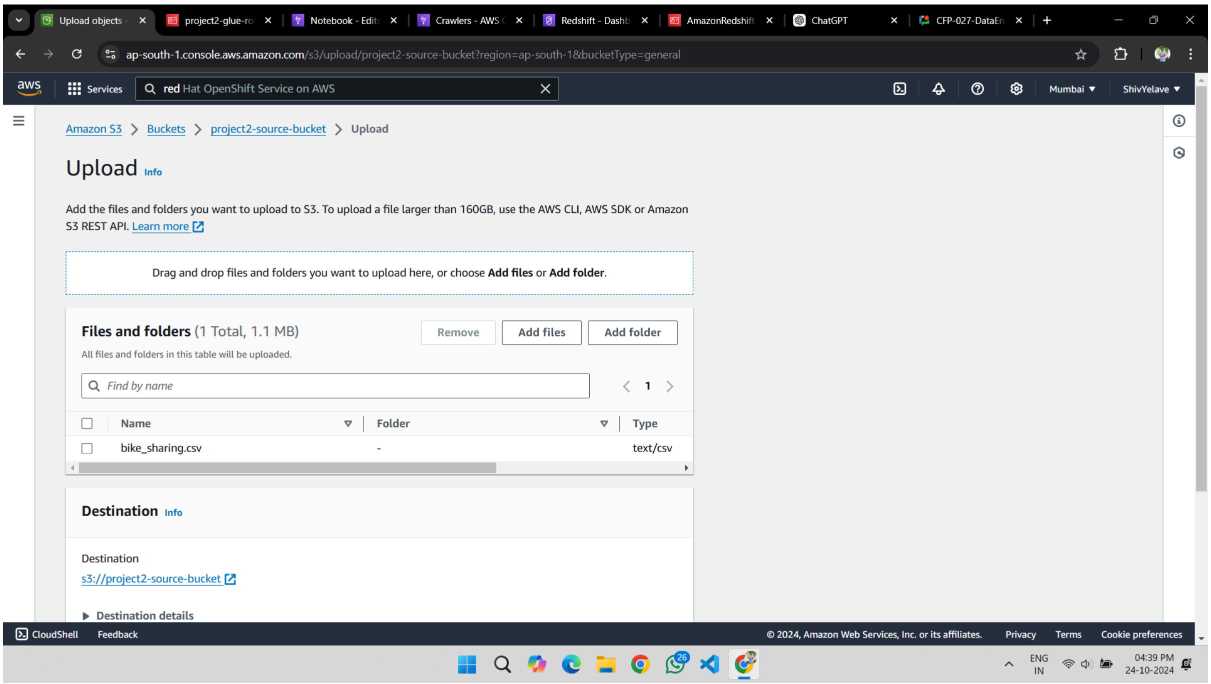
Overview:

This task involves creating an end-to-end data pipeline in AWS that retrieves data from an S3 bucket, processes it using AWS Glue, and stores the transformed data back in another S3 bucket. The pipeline includes the following components:

1. **Source S3 Bucket**: Stores the initial raw data that needs to be processed.
2. **AWS Glue Crawler**: Automatically discovers and catalogs the schema of the data in the source bucket, making it available for querying.
3. **AWS Glue Notebook Job**: Executes the ETL (Extract, Transform, Load) logic, fetching the data from the Glue catalog, performing necessary transformations, and writing the processed data back to the destination S3 bucket.
4. **Destination S3 Bucket**: Stores the final output of the ETL process, which can be used for further analysis or reporting.

Step 1: Set Up AWS S3 Buckets

1. **Create Source S3 Bucket**
   - Navigate to the S3 service in the AWS Management Console.
   - Click on "Create bucket."
   - Provide a unique bucket name (e.g., `source-bucket-name`) and configure settings as needed (e.g., versioning, encryption).



Amazon S3 > Buckets > project2-source-bucket > Upload

Upload Info

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

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

Files and folders (1 Total, 1.1 MB)

All files and folders in this table will be uploaded.

	Name	Folder	Type
□	bike_sharing.csv	-	text/csv

Destination Info

Destination
s3://project2-source-bucket

▶ Destination details

- ○ Click "Create bucket."

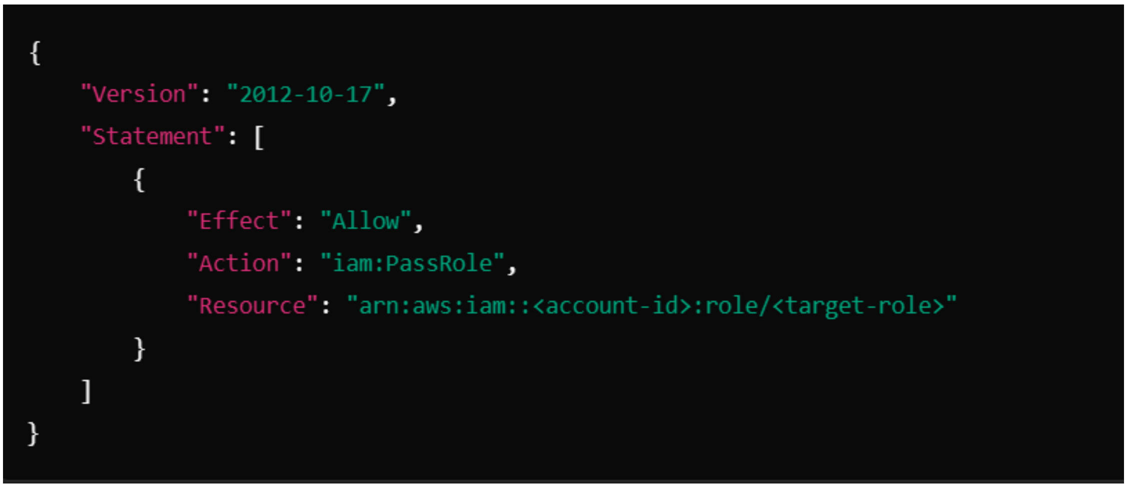
2. **Create Destination S3 Bucket**
   - ○ Repeat the steps above to create another bucket (e.g., `destination-bucket-name`).

Step 2: Configure IAM Roles and Permissions

1. **Create a New Policy**:
   - In the IAM console, click on **Policies** in the left sidebar.
   - Click on **Create policy**.

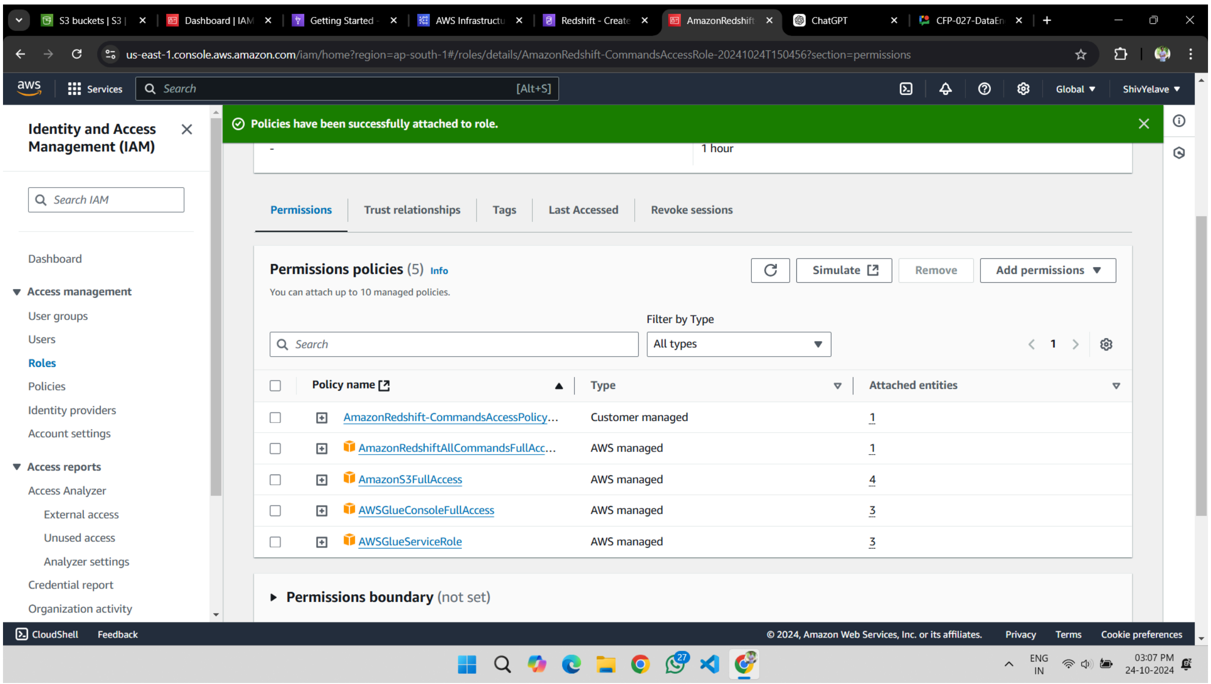
Define the Policy:

- Switch to the **JSON** tab and paste the following policy, which allows the Glue service to assume roles:



```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": "iam:PassRole",
            "Resource": "arn:aws:iam::<account-id>:role/<target-role>"
        }
    ]
}
```

2. **Create an IAM Role for AWS Glue**
   - Go to the IAM service in the AWS Management Console.
   - Click on "Roles" > "Create role."
   - Select "Glue" as the trusted entity.
   - Attach the following policies:
     - `AmazonS3FullAccess` (or a more restrictive policy specific to your buckets)
     - `AWSGlueServiceRole`
     - `Above Policy that we created.`
   - Click "Next: Tags," then "Next: Review."
   - Provide a name for the role (e.g., `GlueServiceRole`) and click "Create role."



Policies have been successfully attached to role.

Permissions policies (5) Info

You can attach up to 10 managed policies.

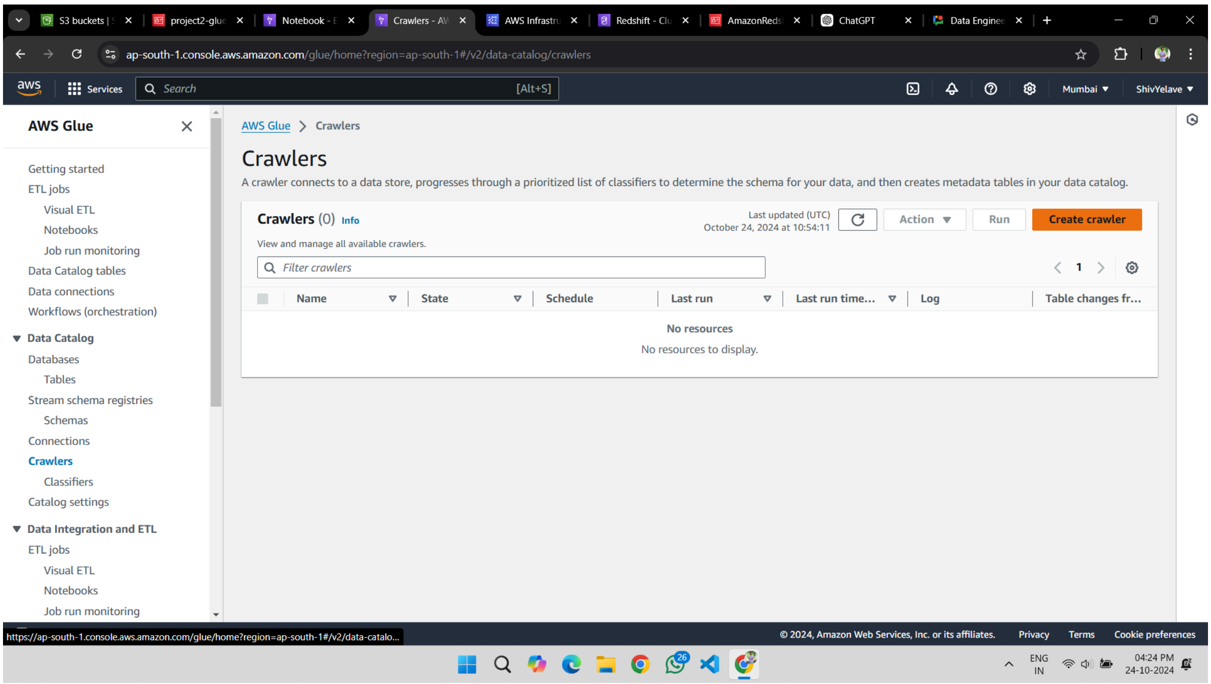
Filter by Type: All types

Policy name	Type	Attached entities
AmazonRedshift-CommandsAccessPolicy...	Customer managed	1
AmazonRedshiftAllCommandsFullAcc...	AWS managed	1
AmazonS3FullAccess	AWS managed	4
AWSGlueConsoleFullAccess	AWS managed	3
AWSGlueServiceRole	AWS managed	3

Permissions boundary (not set)

Step 3: Create an AWS Glue Crawler

1. **Navigate to AWS Glue Service**
   - In the AWS Management Console, go to the Glue service.
2. **Create a Crawler**
   - Click on "Crawlers" > "Add crawler."



ap-south-1.console.aws.amazon.com/glue/home?region=ap-south-1#/v2/data-catalog/crawlers

AWS Glue > Crawlers

Crawlers

A crawler connects to a data store, progresses through a prioritized list of classifiers to determine the schema for your data, and then creates metadata tables in your data catalog.

Crawlers (0) Info

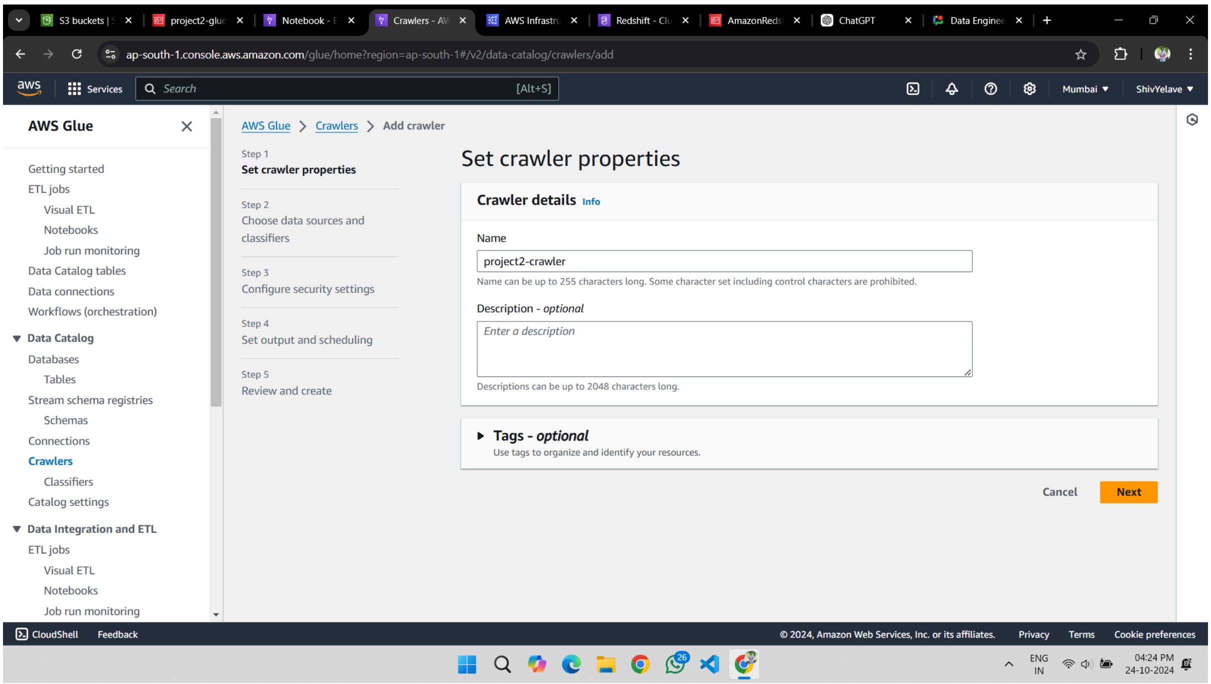
Last updated (UTC)
October 24, 2024 at 10:54:11

View and manage all available crawlers.

Name	State	Schedule	Last run	Last run time...	Log	Table changes fr...

No resources
No resources to display.

○ Name your crawler (e.g., `my-data-crawler`).


AWS Glue
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
Cancel
Next

ap-south-1.console.aws.amazon.com/glue/home?region=ap-south-1#/v2/data-catalog/crawlers/add

AWS Glue > Crawlers > Add crawler

Set crawler properties

Crawler details Info

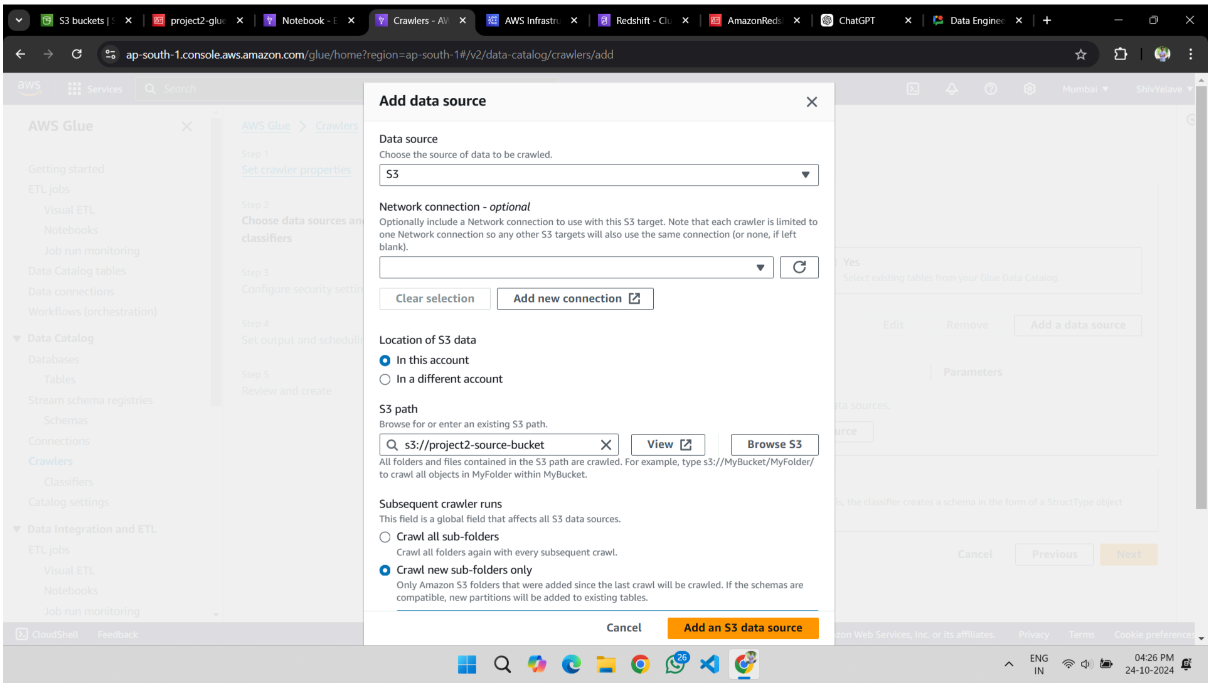
Name
project2-crawler
Name can be up to 255 characters long. Some character set including control characters are prohibited.

Description - *optional*
Enter a description
Descriptions can be up to 2048 characters long.

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

○ Choose "Data stores" as the crawler source type.

○ Select "S3" and specify your source bucket (e.g., `source-bucket-name`).



Add data source

Data source

Choose the source of data to be crawled.

S3

Network connection - *optional*

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

Location of S3 data

● In this account

○ In a different account

S3 path

Browse for or enter an existing S3 path.

s3://project2-source-bucket

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.

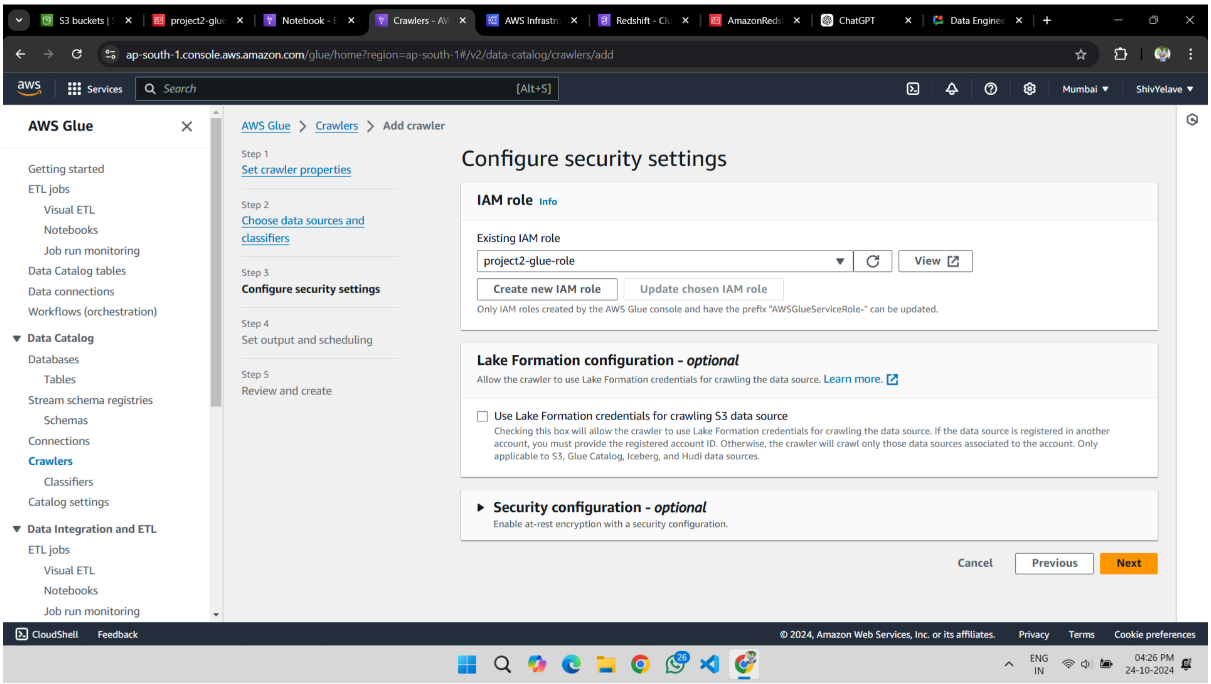
○ Crawl all sub-folders

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 compatible, new partitions will be added to existing tables.

- Specify the IAM role you created earlier (e.g., `GlueServiceRole`).



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

project2-glue-role

Only IAM roles created by the AWS Glue console and have the prefix "AWSGlueServiceRole-" can be updated.

Lake Formation configuration - *optional*

Allow the crawler to use Lake Formation credentials for crawling the data source. Learn more.

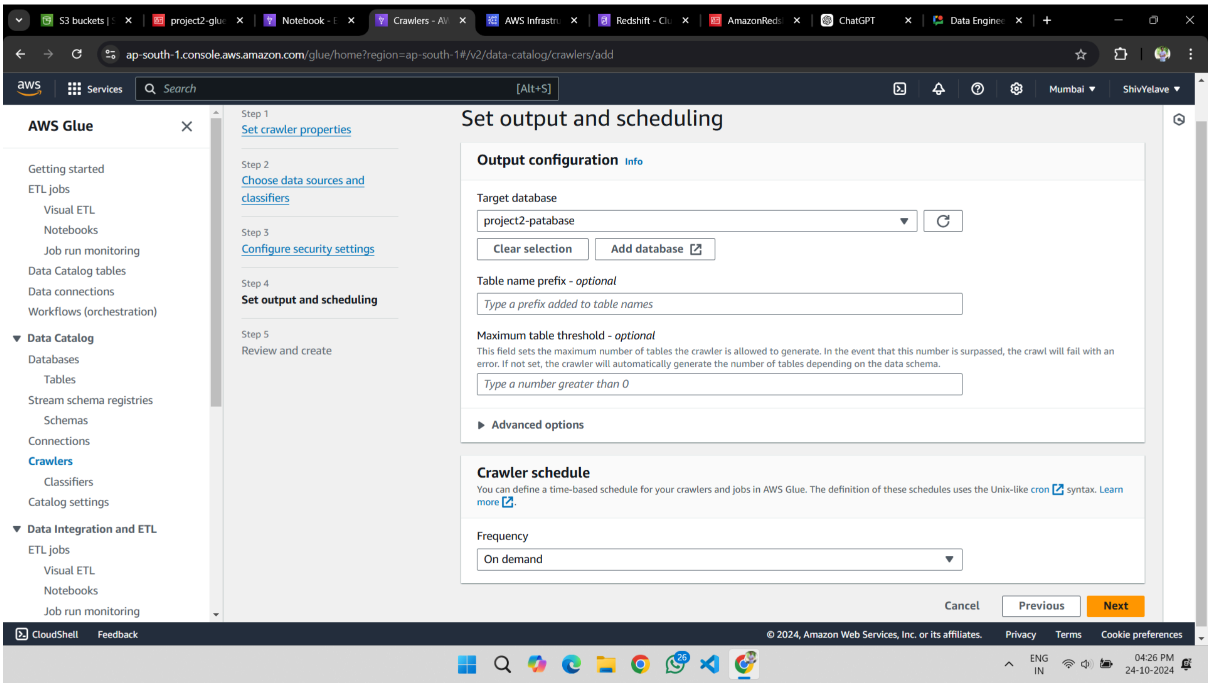
□ Use Lake Formation credentials for crawling S3 data source

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

▸ **Security configuration - *optional***

Enable at-rest encryption with a security configuration.

- Click "Next" and configure options for output (databases, tables).



Set output and scheduling

Output configuration Info

Target database: project2-patabase

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

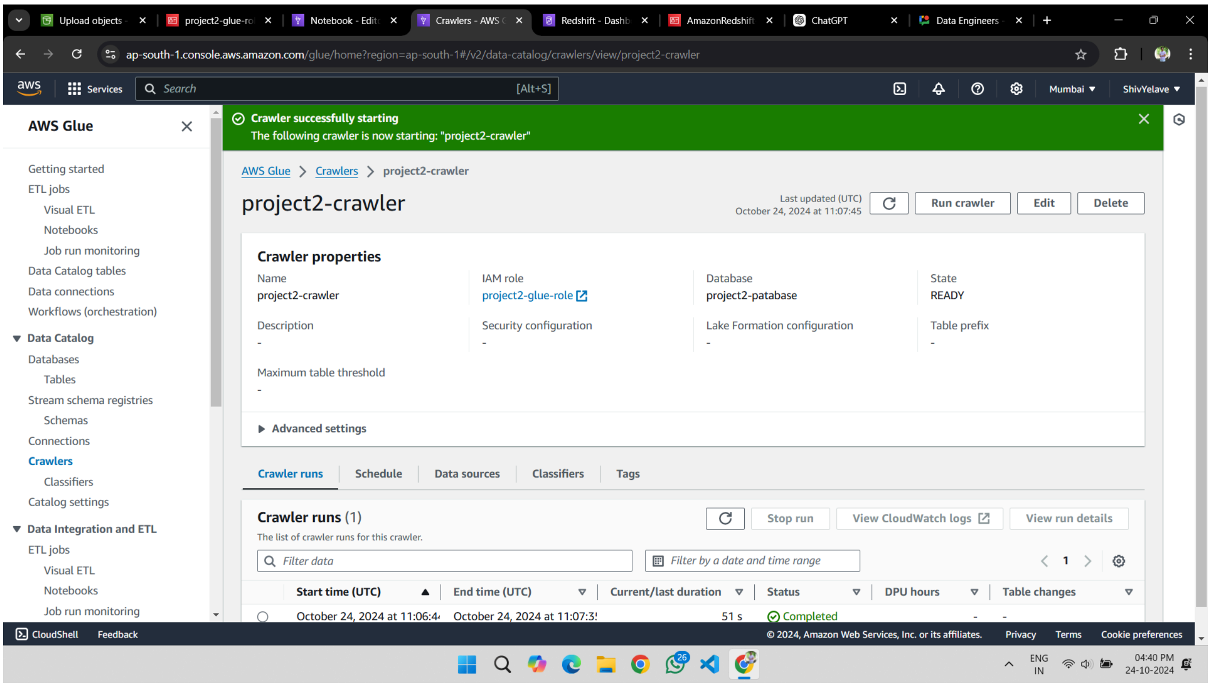
Crawler schedule

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

- ○ Review and finish creating the crawler.

3. **Run the Crawler**
    - ○ Go to the Crawlers page, select your crawler, and click "Run crawler."



Crawler successfully starting
The following crawler is now starting: "project2-crawler"

project2-crawler

Last updated (UTC)
October 24, 2024 at 11:07:45

Crawler properties

Name	IAM role	Database	State
project2-crawler	project2-glue-role	project2-patabase	READY
Description	**Security configuration**	**Lake Formation configuration**	**Table prefix**
-	-	-	-
Maximum table threshold			
-			

Crawler runs (1)

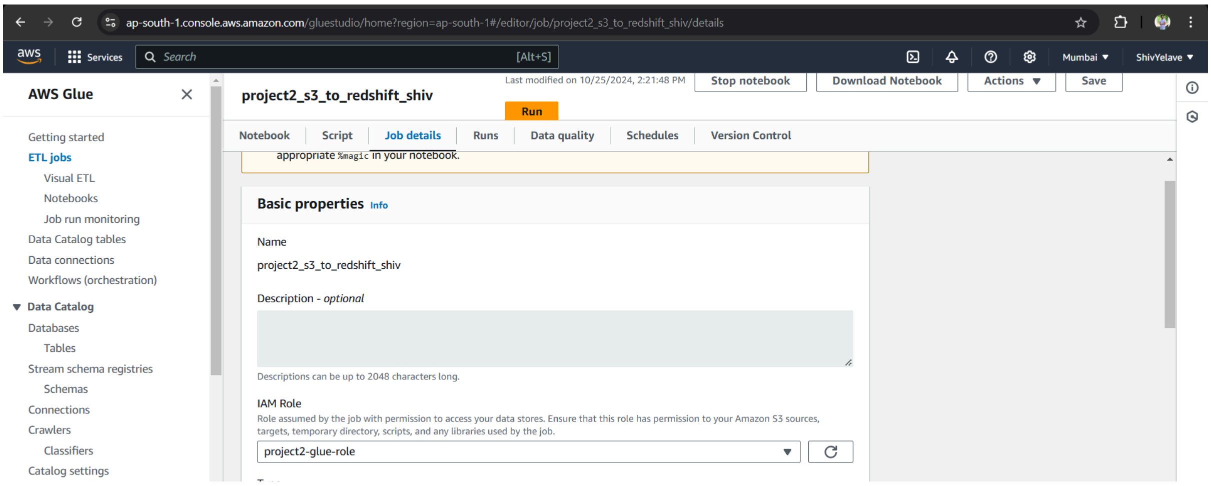
The list of crawler runs for this crawler.

Start time (UTC)	End time (UTC)	Current/last duration	Status	DPU hours	Table changes
October 24, 2024 at 11:06:4	October 24, 2024 at 11:07:3	51 s	Completed	-	-

Step 4: Create a Glue Notebook Job

1. **Navigate to Glue Jobs**
    - ○ In the Glue service, click on "Jobs" .
    - ○ Click on **Add notebook**.
    - ○ Name your job (e.g., `my-glue-notebook-job`).

- Choose the IAM role you created earlier (e.g., `GlueServiceRole`).



project2_s3_to_redshift_shiv

Last modified on 10/25/2024, 2:21:48 PM

appropriate %magic in your notebook.

Basic properties Info

Name
project2_s3_to_redshift_shiv

Description - *optional*

Descriptions can be up to 2048 characters long.

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.

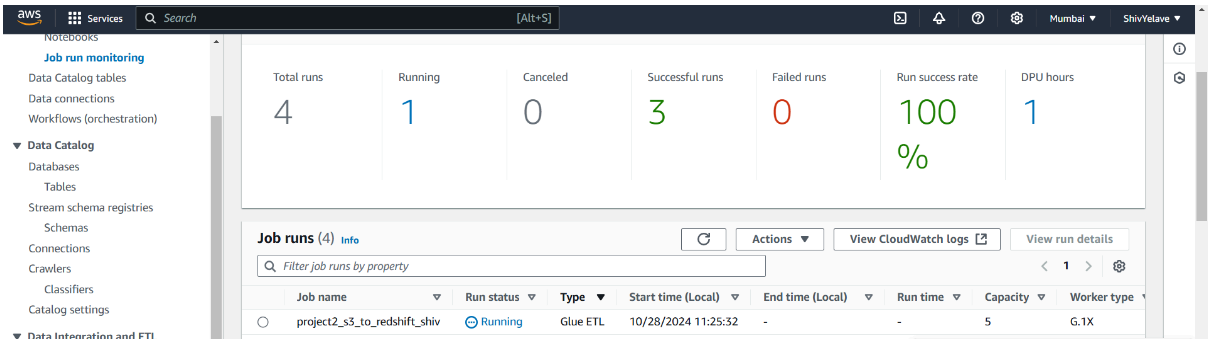
project2-glue-role

- Write and Save Glue Notebook Code.
- Click "Next" and configure any additional options (like job bookmarks).
- Review and create the job.

2. **Save the Notebook Job**
   - Save your notebook and ensure it's in the correct format.

Step 5: Monitor and Maintain the Pipeline

- Regularly check logs in CloudWatch for the Glue job and crawler.

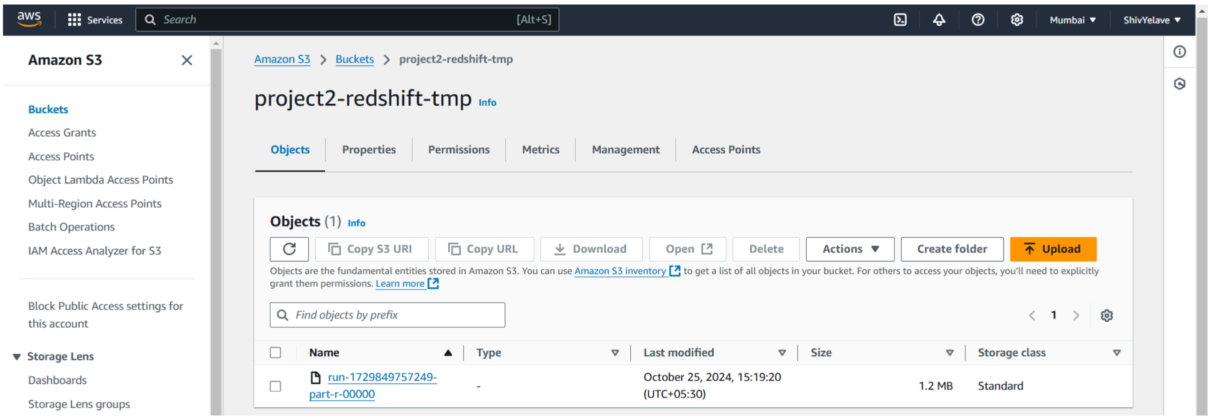


Total runs	Running	Canceled	Successful runs	Failed runs	Run success rate	DPU hours
4	1	0	3	0	100 %	1

Job runs (4) Info

	Job name	Run status	Type	Start time (Local)	End time (Local)	Run time	Capacity	Worker type
○	project2_s3_to_redshift_shiv	Running	Glue ETL	10/28/2024 11:25:32	-	-	5	G.1X

- Monitor data in the destination S3 bucket to ensure proper ETL execution.



project2-redshift-tmp Info

Objects (1) Info

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-1729849757249-part-r-00000	-	October 25, 2024, 15:19:20 (UTC+05:30)	1.2 MB	Standard