Design and implement a serverless data processing pipeline using cloud-native services (e.g., **AWS Glue**, Azure Data Factory, Google Dataflow).

Ebuka Obiakor – 10th March, 2024

What is AWS Glue?

AWS Glue is a serverless data integration service that makes it easy for analytics users to discover, prepare, move, and integrate data from multiple sources. You can use it for analytics, machine learning, and application development. It also includes additional productivity and data ops tooling for authoring, running jobs, and implementing business workflows.

***Key Features**:*

- ***Data Catalog**: The **AWS Glue Data Catalog** serves as a persistent metadata store. It helps organize and manage metadata related to your data sources, tables, and schemas.*
- ***Data Discovery**: Discover and explore data using crawlers that automatically infer schema and populate the catalog.*
- ***Data Preparation**: Use **AWS Glue jobs** to transform and clean data. You can create jobs visually with **AWS Glue Studio** or write custom ETL scripts.*
- ***Serverless Execution**: AWS Glue is serverless, meaning you don't need to manage infrastructure. It scales automatically based on your workload.*
- ***Integration with Other Services**: AWS Glue integrates seamlessly with other AWS services like Amazon S3, Amazon RDS, and Amazon Redshift.*

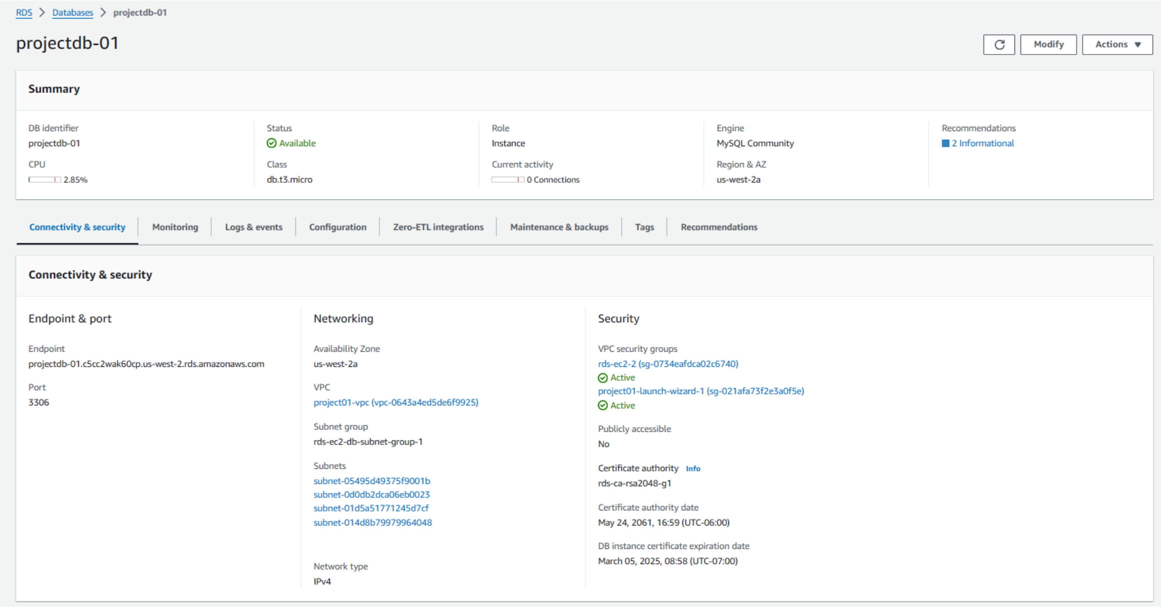
***Getting Started**:*

- ***AWS Glue Studio**: Use the visual job editor in **AWS Glue Studio** to build and monitor ETL jobs.* It simplifies the process of creating, running, and managing integration jobs
- ***ETL Scripts**: You can write custom ETL scripts in Python or Scala to perform data transformations.*
- ***Notebook-Based Jobs**: Create interactive jobs using Jupyter notebooks within **AWS Glue Studio**.*
- ***Local Development**: Develop and test AWS Glue jobs locally using interactive sessions.*

***Automation and Monitoring**:*

- ***Event-Based Triggers**: Automate jobs and crawlers based on events (e.g., file uploads).*
- ***Workflows**: Define workflows for ETL activities involving multiple crawlers, jobs, and triggers.*
- ***Monitoring Tools**: Monitor job runs using automated tools, the Apache Spark UI, and AWS CloudTrail.*

**The demo assumes existing Amazon RDS database exist. If not, you would need to create one. Both RDS and AWS Glue kept in the same region and VPC. **



projectdb-01

Summary

DB identifier: projectdb-01
CPU: 2.85%
Status: Available
Class: db.t3.micro
Role: Instance
Current activity: 0 Connections
Engine: MySQL Community
Region & AZ: us-west-2a
Recommendations: 2 Informational

Connectivity & security

Endpoint & port

Endpoint: projectdb-01.c5cc2wak60cp.us-west-2.rds.amazonaws.com
Port: 3306

Networking

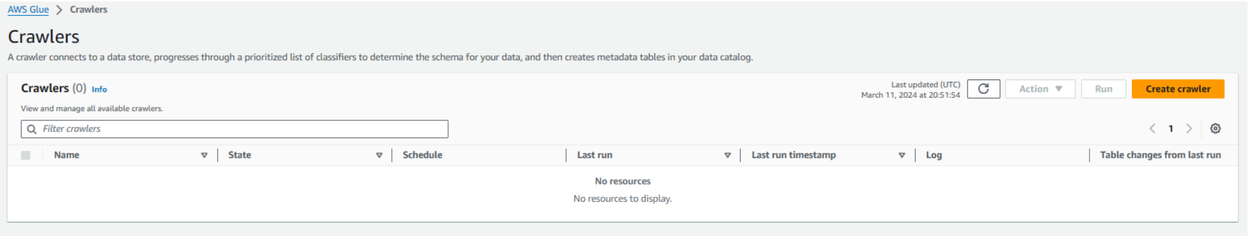
Availability Zone: us-west-2a
VPC: project01-vpc (vpc-0643a4ed5de6f9925)
Subnet group: rds-ec2-db-subnet-group-1
Subnets:
subnet-05495d49375f9001b
subnet-0d0db2dca06eb0023
subnet-01d5a51771245d7cf
subnet-014d8b79979964048
Network type: IPv4

Security

VPC security groups:
rds-ec2-2 (sg-0734eafdca02c6740)
Active
project01-launch-wizard-1 (sg-021afa73f2e3a0f5e)
Active
Publicly accessible: No
Certificate authority Info: rds-ca-rsa2048-g1
Certificate authority date: May 24, 2061, 16:59 (UTC-06:00)
DB instance certificate expiration date: March 05, 2025, 08:58 (UTC-07:00)

Steps to create an data catalog using crawlers.

1. Click on create crawler



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.

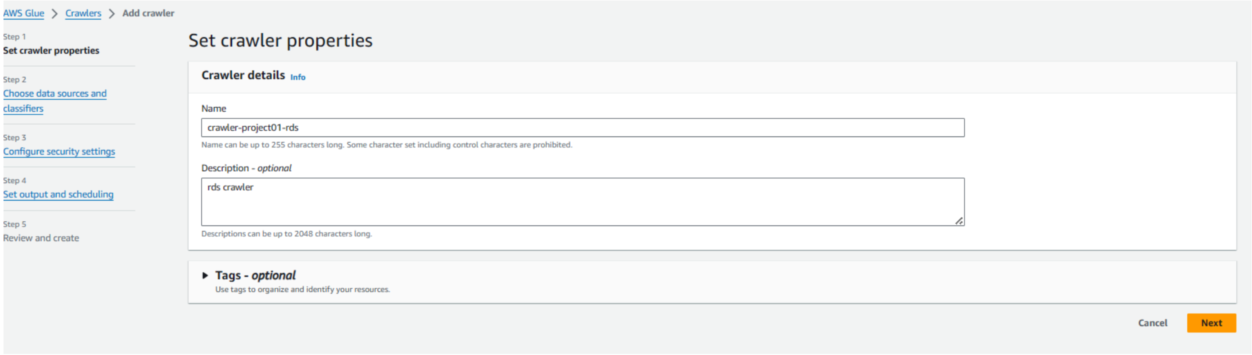
Crawlers (0)

View and manage all available crawlers.

Last updated (UTC)
March 11, 2024 at 20:51:54

Name	State	Schedule	Last run	Last run timestamp	Log	Table changes from last run

No resources
No resources to display.



Step 1
Set crawler properties

Step 2
Choose data sources and classifiers

Step 3
Configure security settings

Step 4
Set output and scheduling

Step 5
Review and create

Set crawler properties

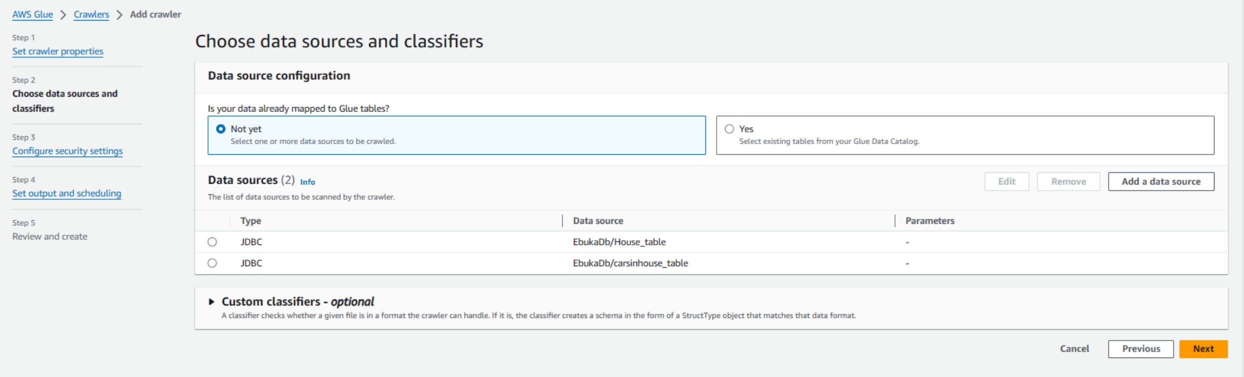
Crawler details

Name: crawler-project01-rds
Name can be up to 255 characters long. Some character set including control characters are prohibited.

Description - *optional*: rds crawler
Descriptions can be up to 2048 characters long.

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

2. Choose data sources and classifiers



AWS Glue > Crawlers > Add crawler

Step 1
Set crawler properties
Step 2
Choose data sources and classifiers
Step 3
Configure security settings
Step 4
Set output and scheduling
Step 5
Review and create

Choose data sources and classifiers

Data source configuration

Is your data already mapped to Glue tables?

● Not yet
Select one or more data sources to be crawled.

○ Yes
Select existing tables from your Glue Data Catalog.

Data sources (2) Info

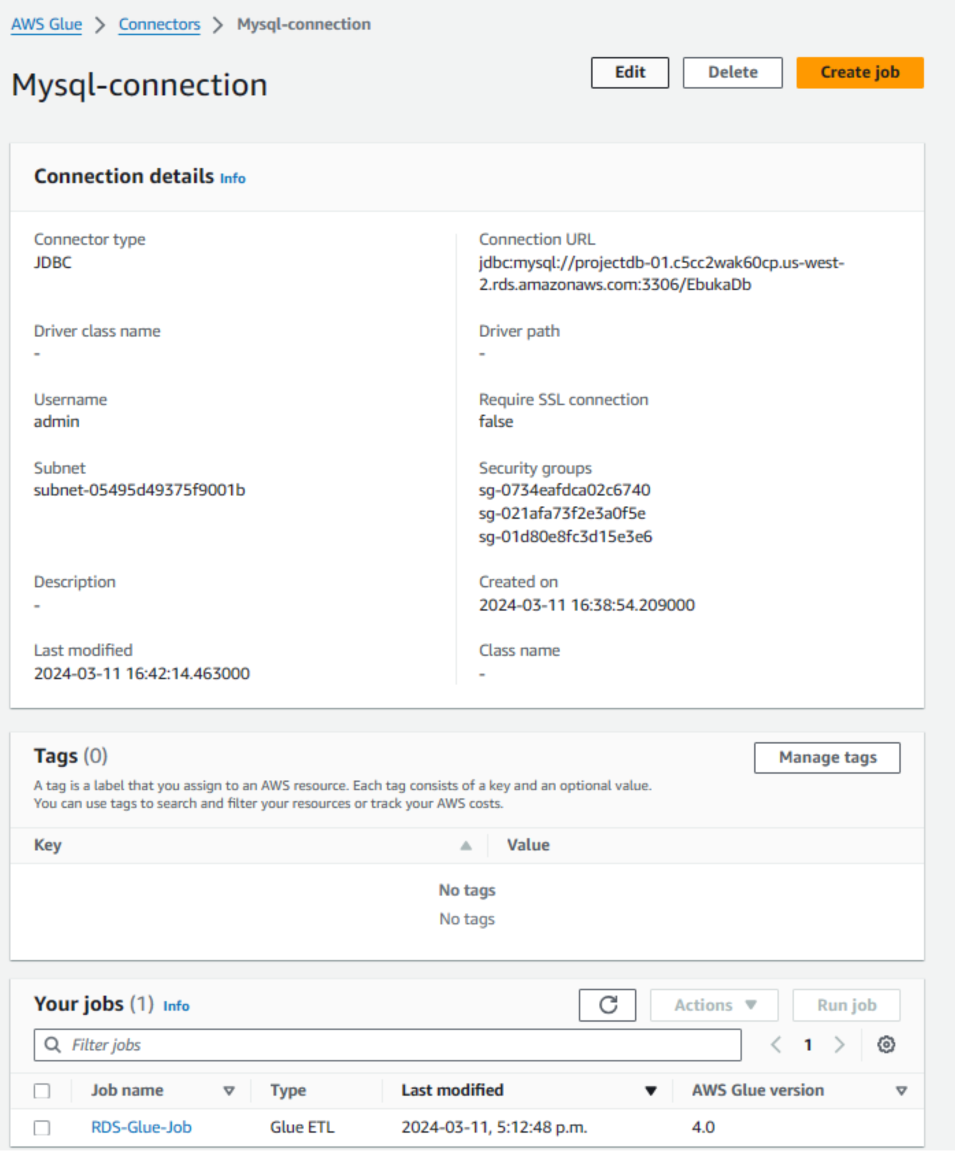
The list of data sources to be scanned by the crawler.

Edit
Remove
Add a data source

	Type	Data source	Parameters
○	JDBC	EbukaDb/House_table	-
○	JDBC	EbukaDb/carsinhouse_table	-

▸ **Custom classifiers - *optional***
A classifier checks whether a given file is in a format the crawler can handle. If it is, the classifier creates a schema in the form of a StructType object that matches that data format.

3. Create connection between AWS Glue and Amazon RDS using crawler



Mysql-connection

Connection details Info

Connector type JDBC	Connection URL jdbc:mysql://projectdb-01.c5cc2wak60cp.us-west-2.rds.amazonaws.com:3306/EbukaDb
Driver class name -	Driver path -
Username admin	Require SSL connection false
Subnet subnet-05495d49375f9001b	Security groups sg-0734eafdca02c6740 sg-021afa73f2e3a0f5e sg-01d80e8fc3d15e3e6
Description -	Created on 2024-03-11 16:38:54.209000
Last modified 2024-03-11 16:42:14.463000	Class name -

Tags (0)

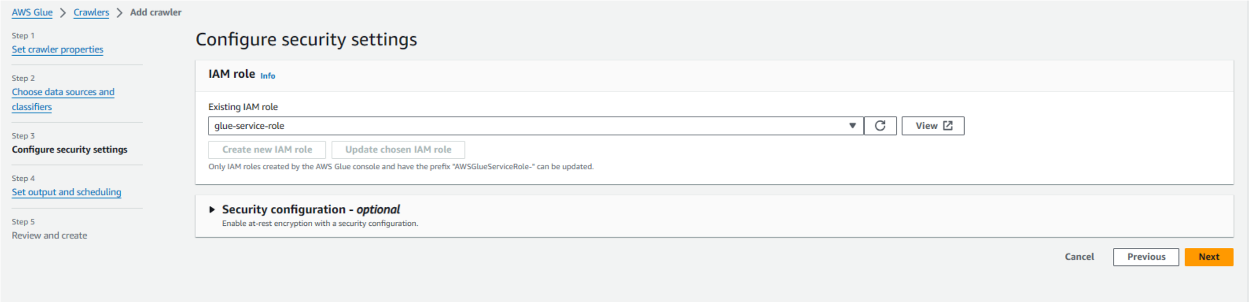
A tag is a label that you assign to an AWS resource. Each tag consists of a key and an optional value. You can use tags to search and filter your resources or track your AWS costs.

Key	Value
No tags No tags	

Your jobs (1) Info

□	Job name	Type	Last modified	AWS Glue version
□	RDS-Glue-Job	Glue ETL	2024-03-11, 5:12:48 p.m.	4.0

4. Configure security settings



Configure security settings

IAM role

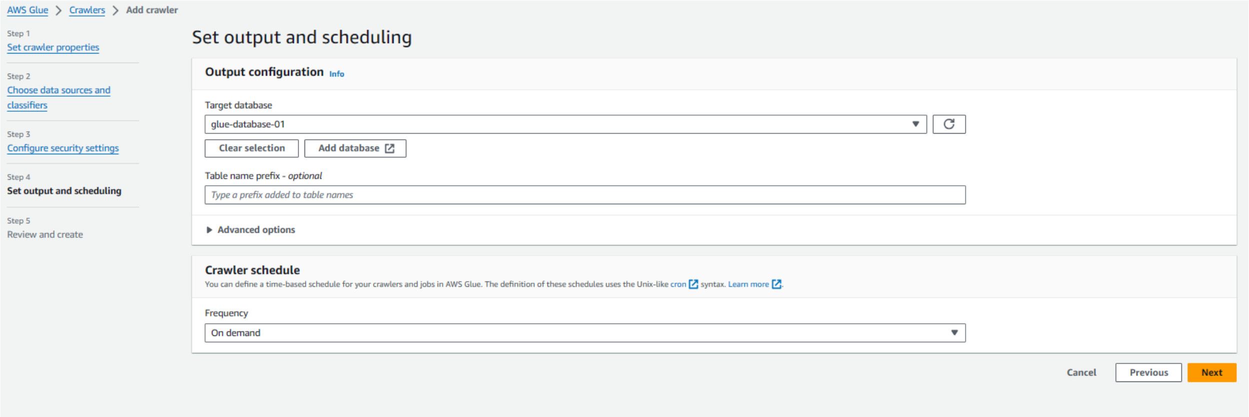
Existing IAM role: glue-service-role

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

▸ Security configuration - *optional*

Enable at-rest encryption with a security configuration.

5. Set output and scheduling


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
Info
Clear selection
Add database
Cancel
Previous
Next

Set output and scheduling

Output configuration

Target database: glue-database-01

Table name prefix - *optional*: *Type a prefix added to table names*

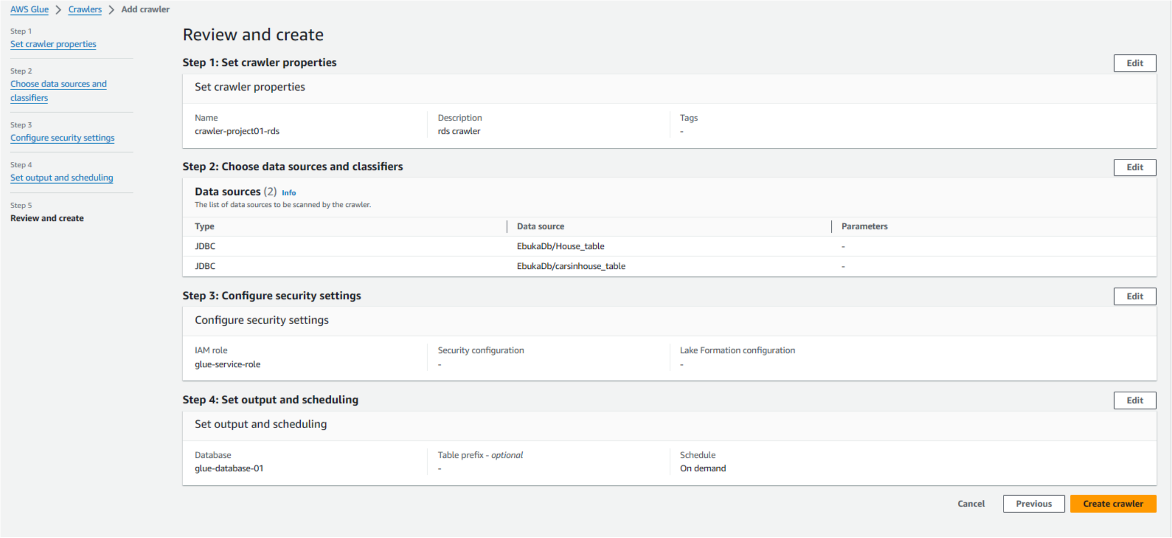
▸ Advanced options

Crawler schedule

You can define a time-based schedule for your crawlers and jobs in AWS Glue. The definition of these schedules uses the Unix-like cron syntax. Learn more.

Frequency: On demand

6. Review and create crawler


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
Edit
Edit
Edit
Edit
Info
Cancel
Previous
Create crawler

Review and create

Step 1: Set crawler properties

Set crawler properties

Name	Description	Tags
crawler-project01-rds	rds crawler	-

Step 2: Choose data sources and classifiers

Data sources (2)

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

Type	Data source	Parameters
JDBC	EbukaDb/House_table	-
JDBC	EbukaDb/carsinhouse_table	-

Step 3: Configure security settings

Configure security settings

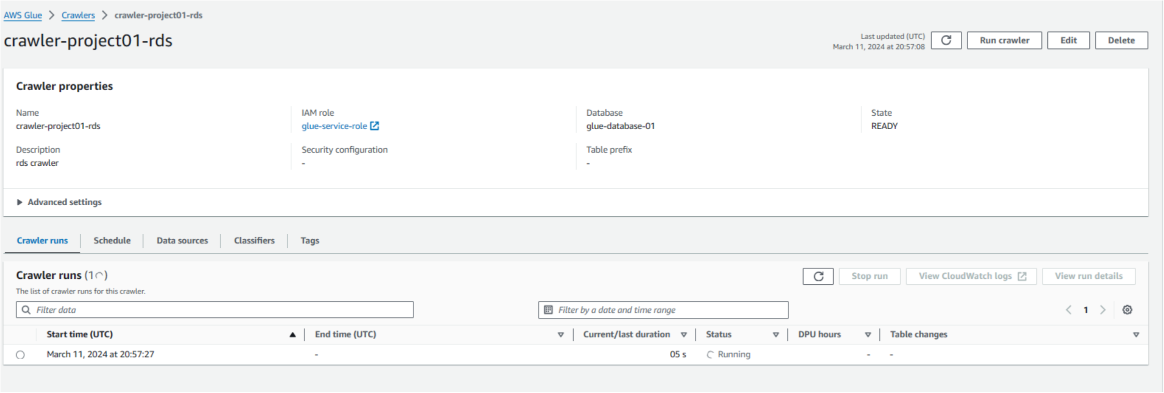
IAM role	Security configuration	Lake Formation configuration
glue-service-role	-	-

Step 4: Set output and scheduling

Set output and scheduling

Database	Table prefix - *optional*	Schedule
glue-database-01	-	On demand

7. Run crawler



AWS Glue > Crawlers > crawler-project01-rds

crawler-project01-rds

Last updated (UTC) March 11, 2024 at 20:57:08

Crawler properties

Name crawler-project01-rds	IAM role glue-service-role	Database glue-database-01	State READY
Description rds crawler	Security configuration -	Table prefix -	

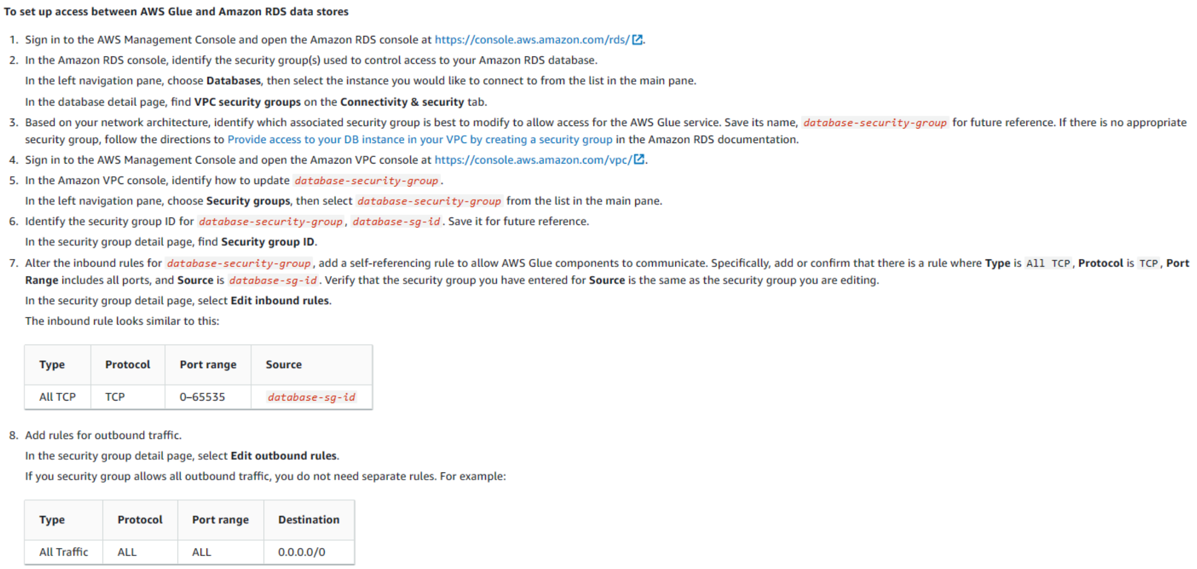
▶ Advanced settings

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
○	March 11, 2024 at 20:57:27	-	05 s	Running	-	-

8. Additional configuration needs to be set to allow RDS communicate with Glue over the network



To set up access between AWS Glue and Amazon RDS data stores

1. Sign in to the AWS Management Console and open the Amazon RDS console at https://console.aws.amazon.com/rds/.
2. In the Amazon RDS console, identify the security group(s) used to control access to your Amazon RDS database.

   In the left navigation pane, choose **Databases**, then select the instance you would like to connect to from the list in the main pane.

   In the database detail page, find **VPC security groups** on the **Connectivity & security** tab.
3. Based on your network architecture, identify which associated security group is best to modify to allow access for the AWS Glue service. Save its name, `database-security-group` for future reference. If there is no appropriate security group, follow the directions to Provide access to your DB instance in your VPC by creating a security group in the Amazon RDS documentation.
4. Sign in to the AWS Management Console and open the Amazon VPC console at https://console.aws.amazon.com/vpc/.
5. In the Amazon VPC console, identify how to update `database-security-group`.

   In the left navigation pane, choose **Security groups**, then select `database-security-group` from the list in the main pane.
6. Identify the security group ID for `database-security-group`, `database-sg-id`. Save it for future reference.

   In the security group detail page, find **Security group ID**.
7. Alter the inbound rules for `database-security-group`, add a self-referencing rule to allow AWS Glue components to communicate. Specifically, add or confirm that there is a rule where **Type** is `All TCP`, **Protocol** is `TCP`, **Port Range** includes all ports, and **Source** is `database-sg-id`. Verify that the security group you have entered for **Source** is the same as the security group you are editing.

   In the security group detail page, select **Edit inbound rules**.

   The inbound rule looks similar to this:

Type	Protocol	Port range	Source
All TCP	TCP	0–65535	`database-sg-id`

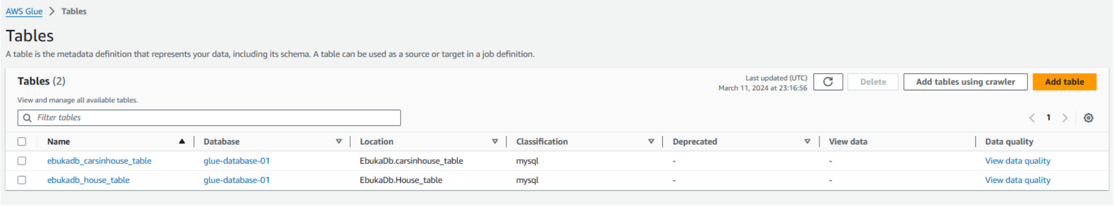
8. Add rules for outbound traffic.

   In the security group detail page, select **Edit outbound rules**.

   If you security group allows all outbound traffic, you do not need separate rules. For example:

Type	Protocol	Port range	Destination
All Traffic	ALL	ALL	0.0.0.0/0

9. Run crawler to create table with Data Catalog



AWS Glue > Tables

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.

Tables (2)

View and manage all available tables.

Last updated (UTC) March 11, 2024 at 23:16:56

□	Name	Database	Location	Classification	Deprecated	View data	Data quality
□	ebukadb_carsinhouse_table	glue-database-01	EbukaDb.carsinhouse_table	mysql	-	-	View data quality
□	ebukadb_house_table	glue-database-01	EbukaDb.House_table	mysql	-	-	View data quality

References

- https://docs.aws.amazon.com/glue/latest/dg/setup-vpc-for-glue-access.html
- https://docs.aws.amazon.com/glue/latest/dg/add-crawler.html
- https://docs.aws.amazon.com/glue/latest/dg/getting-started-iam-permissions.html

User Guides:

- ***AWS Glue Studio User Guide****: Learn how to use the visual interface for building ETL jobs.*
- ***AWS Glue Developer Guide****: Provides detailed instructions, features overview, and API references for developers.*
- ***AWS Glue CLI Reference****: Describes AWS CLI commands related to AWS Glue.*
- ***AWS Glue DataBrew Developer Guide****: Explore data preparation with ready-made transformations for analytics and ML*