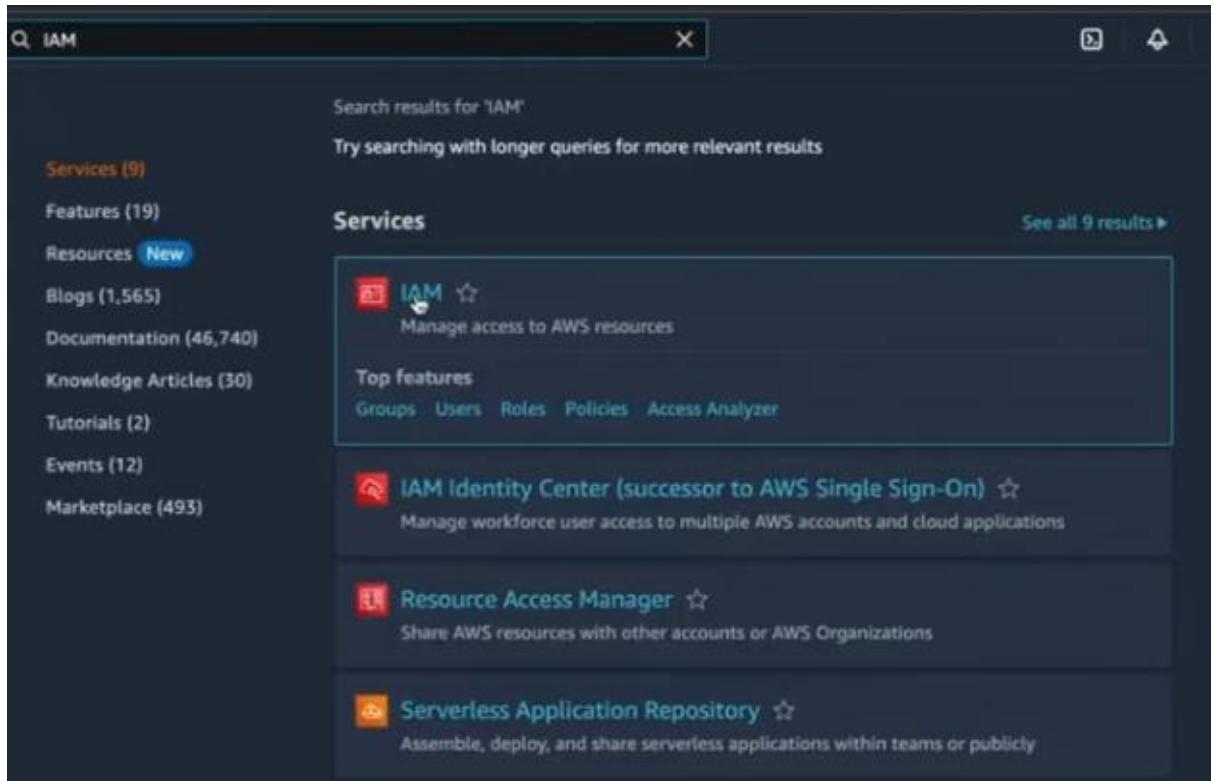


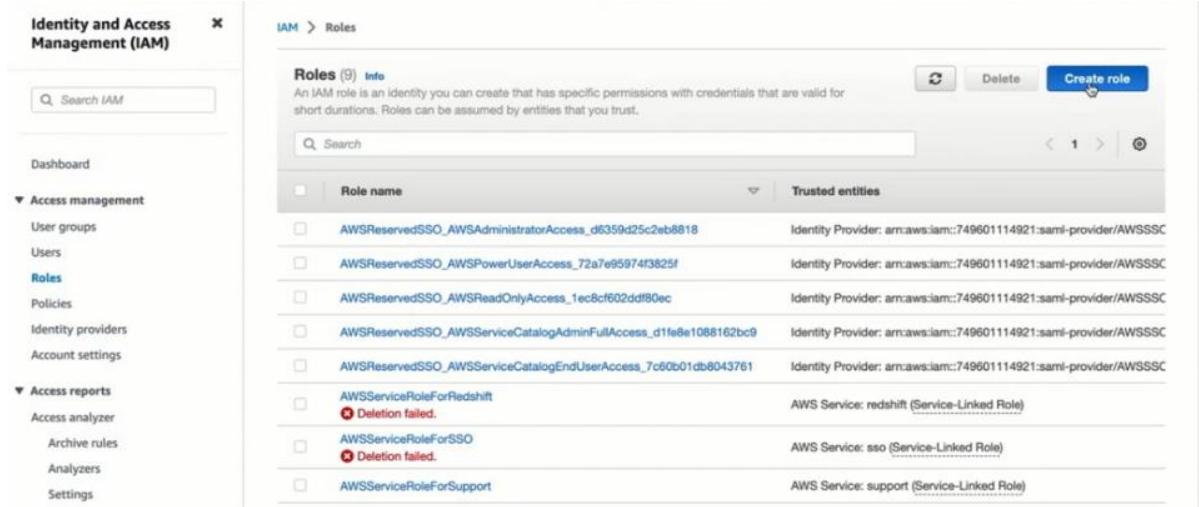
AWS GLUE FOR HARD ETL (EXTRACT, TRANSFORM, AND LOAD)

ACTIVITY

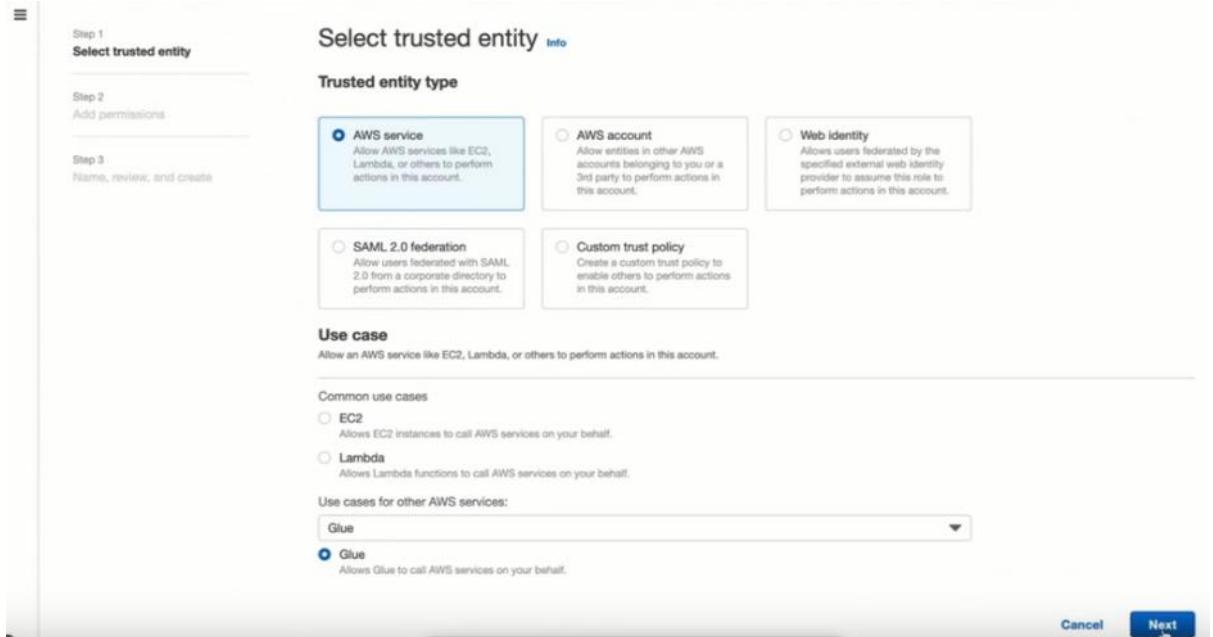
1. FIRST STEP IS TO CREATE THE ROLE FOR OUR AWS GLUE IN ORDER TO GET FULL PERMISSION. IN THE SERVICES, CHOOSE IAM.



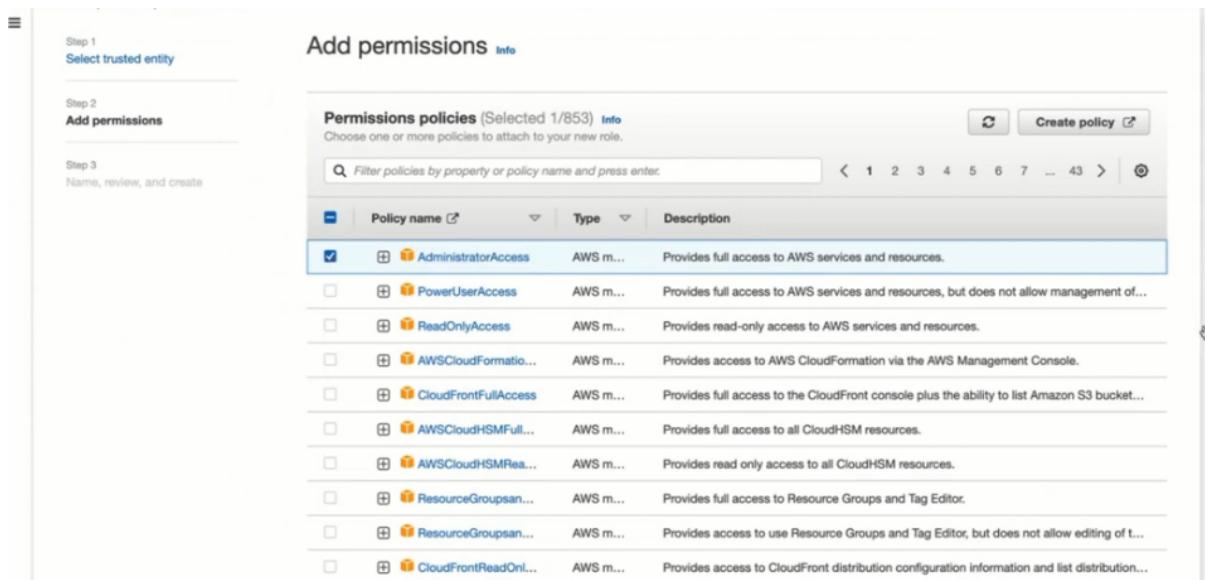
2. CLICK ROLES, THEN., CREATE ROLE.



3. CHOOSE AWS SERVICE IN THE TRUSTED ENTITY TYPE. AND IN THE USE CASES FOR OTHER AWS SERVICES, CHOOSE AND SELECT GLUE. CLICK, NEXT.



4. FOR TRAINING PURPOSES, WE WILL SELECT THE ADMINISTRATORACCESS POLICY WHICH WILL GIVE THE FULL ACCESS. SCROLL DOWN, THEN, CLICK, NEXT.



5. IN THE ROLE NAME, WE WILL CALL THIS, GLUEFULLACCESSROLE. ON THE DESCRIPTION, IT WILL BE, ALLOWS GLUE TO CALL AWS SERVICES ON YOUR BEHALF. SCROLL DOWN, THEN, CLICK CREATE ROLE.

Step 1
Select trusted entity

Step 2
Add permissions

Step 3
Name, review, and create

Name, review, and create

Role details

Role name
Enter a meaningful name to identify this role.
Maximum 64 characters. Use alphanumeric and '+-,@-_.' characters.

Description
Add a short explanation for this role.
Maximum 1000 characters. Use alphanumeric and '+-,@-_.' characters.

Step 1: Select trusted entities Edit

```
1- {
2-     "Version": "2012-10-17",
3-     "Statement": [
4-         {
5-             "Effect": "Allow",
6-             "Principal": {
7-                 "Service": "glue.amazonaws.com"
8-             },
9-             "Action": "sts:AssumeRole"
10-        }
11-    ]
12-}
```

AdministratorAccess AWS managed - job function Permissions policy Edit

Tags

Add tags - optional Info

Tags are key-value pairs that you can add to AWS resources to help identify, organize, or search for resources.

No tags associated with the resource.

Add tag You can add up to 50 more tags.

Cancel Previous Create role

6. THIS WILL NOW BE OUR GLUE ACCESS ROLE.

The screenshot shows the AWS Identity and Access Management (IAM) service. On the left, there's a navigation sidebar with options like Dashboard, User groups, Users, Roles (which is selected), Policies, Identity providers, and Account settings. Under 'Access management', there are sections for User groups, Users, Roles, Policies, Identity providers, and Account settings. Under 'Access reports', there are sections for Access analyzer, Archive rules, Analyzers, Settings, Credential report, Organization activity, and Service control policies (SCPs). The main content area is titled 'Roles (10)'. It contains a table with columns for 'Role name' and 'Trusted entities'. One row for 'AWSReservedSSO_AWSAdministratorAccess_d6359d25c2eb8818' is highlighted in blue. Other rows include 'AWSReservedSSO_AWSPowerUserAccess_72a7e95974f3825f', 'AWSReservedSSO_AWSReadOnlyAccess_1ec8cf602ddfb80ec', 'AWSReservedSSO_AWSServiceCatalogAdminFullAccess_d1fe8e1088162bc9', 'AWSReservedSSO_AWSServiceCatalogEndUserAccess_7c60b01db8043761', 'AWSServiceRoleForRedshift', 'AWSServiceRoleForSSO', 'AWSServiceRoleForSupport', 'AWSServiceRoleForTrustedAdvisor', and 'GlueFullAccessRole'. A success message at the top right says 'Role GlueFullAccessRole created.'

7. GO TO SERVICES, CLICK AMAZON S3 TO CRREATE OUR BUCKET.

The screenshot shows the AWS Services search interface. The search bar at the top has 'S3' typed into it. Below the search bar, there's a sidebar with sections for 'Hard ETL' (containing 'Unsafe jobs' and 'Visual'), 'Services (7)' (listing 'Features (19)', 'Resources (New)', 'Blogs (1,249)', 'Documentation (20,670)', 'Knowledge Articles (30)', 'Tutorials (12)', 'Events (26)', and 'Marketplace (1,165)'), and a large '+' button. The main content area shows search results for 'S3'. It includes a 'Services' section with a list of services and a 'Features' section with a list of features. A context menu is open over the 'Amazon S3' link in the 'Services' section, with options like 'Open Link in New Tab', 'Open Link in New Window', 'Open Link in Incognito Window', 'Save Link As...', 'Copy Link Address', 'Copy', 'Copy Link to Highlight', 'Search Google for "S3"', 'Print...', 'Translate Selection to English', 'Inspect', 'Speech', and 'Services'. A note at the bottom of the menu says 'Fully managed support for S3, FSx, FSx for Lustre, and FSx for Amazon OpenSearch Service'.

CLICK CREATE BUCKET.

The screenshot shows the Amazon S3 homepage under the 'Storage' category. The main heading is 'Amazon S3' with the subtext 'Store and retrieve any amount of data from anywhere'. Below this, a description states: 'Amazon S3 is an object storage service that offers industry-leading scalability, data availability, security, and performance.' To the right, there is a 'Create a bucket' button. A sidebar on the right is titled 'Pricing' and contains information about no minimum fees and a monthly bill calculator.

8. SINCE WE ARE BUILDING A RECOMMENDER SYSTEM, WE WILL GIVE A BUCKET NAME `recommender-system-8730872805`. THE NUMBERS GIVEN HERE ARE JUST RANDOM. THIS IS TO MAKE THAT THIS IS GLOBALLY UNIQUE NAME.

The screenshot shows the 'Create bucket' wizard in the AWS Management Console. The first step, 'General configuration', is shown. It includes fields for 'Bucket name' (containing 'recommender-system-8730872805'), 'AWS Region' (set to 'US East (N. Virginia) us-east-1'), and a 'Copy settings from existing bucket - optional' section with a 'Choose bucket' button. The second step, 'Object Ownership', is also visible, showing options for 'ACLs disabled (recommended)' or 'ACLs enabled'.

9. SCROLL DOWN, AND, CLICK CREATE BUCKET.

Tags (0) - optional
You can use bucket tags to track storage costs and organize buckets. [Learn more](#)

No tags associated with this bucket.

[Add tag](#)

Default encryption [Info](#)
Server-side encryption is automatically applied to new objects stored in this bucket.

Encryption key type [Info](#)

Amazon S3 managed keys (SSE-S3)
 AWS Key Management Service key (SSE-KMS)

Bucket Key
When KMS encryption is used to encrypt new objects in this bucket, the bucket key reduces encryption costs by lowering calls to AWS KMS.
[Learn more](#)

Disable
 Enable

Advanced settings

Info After creating the bucket, you can upload files and folders to the bucket, and configure additional bucket settings.

[Cancel](#) [Create bucket](#)

10. THERE YOU GO, WE HAVE THE BUCKET. INSIDE THE BUCKET, WE ARE GOING TO UPLOAD OUR CSV FILE. CLICK THE BUCKET, recommender-system-8730872805.

Amazon S3

Buckets

- Access Points
- Object Lambda Access Points
- Multi-Region Access Points
- Batch Operations
- IAM Access Analyzer for S3

Block Public Access settings for this account

Storage Lens

- Dashboards
- AWS Organizations settings

Feature spotlight [?](#)

Buckets (1) [Info](#)
Buckets are containers for data stored in S3. [Learn more](#)

Name	AWS Region	Access	Creation date
recommender-system-8730872805	US East (N. Virginia) us-east-1	Bucket and objects not public	June 14, 2023, 19:26:11 (UTC+02:00)

11. INSIDE THE BUCKET, WE ARE GOING TO UPLOAD OUR CSV FILE. CLICK UPLOAD.

The screenshot shows the Amazon S3 console interface. On the left, there's a sidebar with various links like 'Buckets', 'Access Points', and 'Storage Lens'. The main area shows the 'recommender-system-8730872805' bucket. At the top, there are tabs for 'Objects', 'Properties', 'Permissions', 'Metrics', 'Management', and 'Access Points'. Below the tabs, a section titled 'Objects (0)' explains what objects are and how to find them. It includes a search bar and a large button labeled 'Upload'. A message states 'No objects' and 'You don't have any objects in this bucket.' There's also a 'Create folder' button.

12. CLICK ADD FILES.

The screenshot shows the 'Upload' page within the 'recommender-system-8730872805' bucket. The top navigation bar includes 'Services' and a search bar. The main title is 'Upload' with a 'Info' link. A large dashed box is provided for dragging files or choosing them. Below it, a table titled 'Files and folders (0)' shows a single entry: 'All files and folders in this table will be uploaded.' with a 'Remove' button. A 'Find by name' search bar is present. The table headers are 'Name', 'Folder', 'Type', and 'Size'. A message at the bottom says 'No files or folders' and 'You have not chosen any files or folders to upload.' In the bottom right corner, there are 'Add files' and 'Add folder' buttons, with 'Add files' being highlighted. The 'Destination' section at the bottom shows the URL 's3://recommender-system-8730872805'.

13. CHOOSE THE SMALL DATA FILE. CLICK AND UPLOAD THE MOVIE_RATINGS.CSV FOR OUR S3 SOURCE.



14. YOU CAN SEE THAT THE MOVIE_RATINGS IS NOW UPLOADED.

The screenshot shows the AWS S3 console interface. At the top, a green header bar indicates "Upload succeeded". Below it, a summary table shows the upload status:

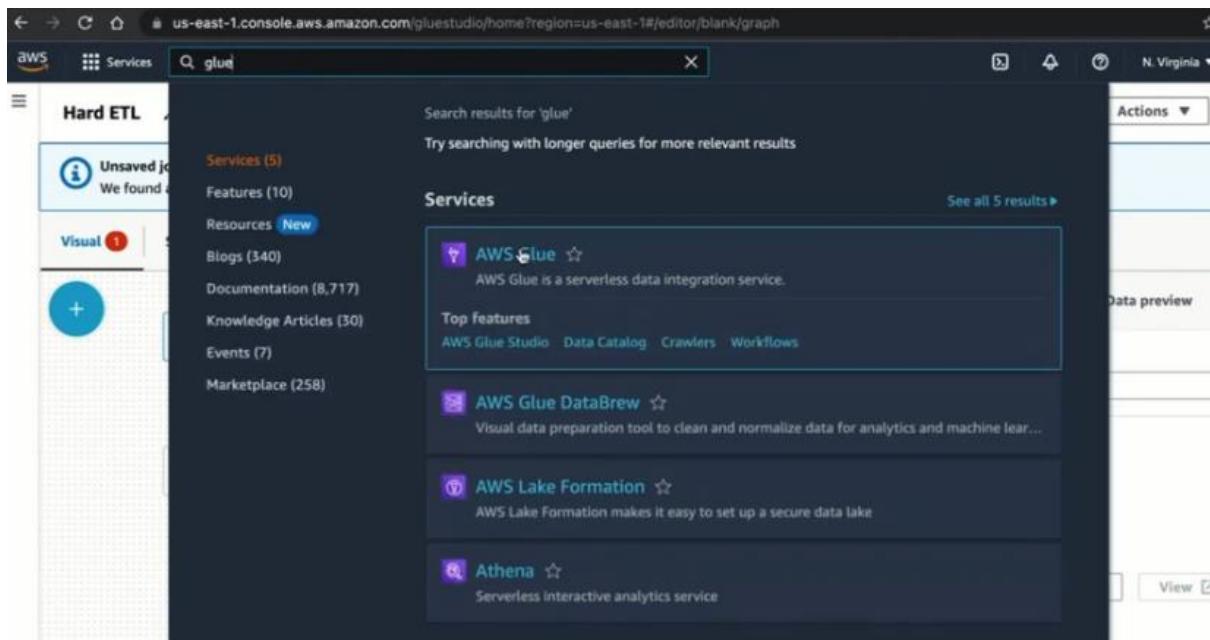
Destination	Succeeded	Failed
s3://recommender-system-8730872805	1 file, 2.3 MB (100.00%)	0 files, 0 B (0%)

Below the summary, there are tabs for "Files and folders" and "Configuration". The "Files and folders" tab is selected, showing a table of files:

Name	Type	Size	Status
movie_ratings.csv	text/csv	2.3 MB	Succeeded

At the bottom, the S3 bucket page shows the uploaded file "movie_ratings.csv" in the "Objects" section. The table includes columns: Name, Type, Last modified, Size, and Storage class.

15. WE NOW NEED TO CREATE THE DATA SOURCE IN THE DATA CATALOGUE OF GLUE. OPEN THE AWS GLUE.



16. GO TO THE DATA CATALOG, CLICK DATABASES. THEN, CLICK, ADD DATABASE. WE WILL BE CREATING ONE OF GLUE.

A screenshot of the AWS Glue Data Catalog Databases page. The left sidebar shows navigation options like 'Getting started', 'ETL jobs', 'Data Catalog tables', and 'Workflows (orchestration)'. Under 'Data Catalog', 'Database' is selected, showing 'Tables', 'Stream schema registries', 'Schemas', 'Connections', 'Crawlers', 'Classifiers', and 'Catalog settings'. The main content area shows a table for 'Databases (0)'. The table has columns for 'Name', 'Description', 'Location URI', and 'Created on (UTC)'. A note says 'No resources' and 'No resources to display.' At the top right, there are buttons for 'Edit', 'Delete', and 'Add database'.

17. SINCE WE WILL BE CREATING A DATABASE FOR THE MOVIE RATINGS, WE WILL NAME IT MOVIE-RATINGS-GLUE-DATABASE. THEN, CLICK CREATE DATABASE.

AWS Glue > Databases > Add database

Create a database

Create a database in the AWS Glue Data Catalog.

Database details

Name: movie-ratings-glue-database

Database name is required, in lowercase characters, and no longer than 255 characters.

Location - optional

Set the URI location for use by clients of the Data Catalog.

Description - optional

Enter text

Descriptions can be up to 2048 characters long.

Cancel Create database

18. THIS WILL NOW BE THE FIRST POINT IN THE DATA CATALOG. WE WILL THEN CREATE A TABLE WTIHIN THE MOVIE-RATINGS-GLUE-DATABASE DATABASE. THE TABLE IS GOING TO BE EXACTLY THE TABLE CONTAINING THE MOVIE RATINGS IN THE CSV FILE. TO DO SO, WE ARE GOING TO ENTER OUR DATABASE.

CLICK THE DATABASE movie-ratings-glue-database.

AWS Glue > Databases

Databases (1)

A database is a set of associated table definitions, organized into a logical group.

Name	Description	Location URI	Created on (UTC)
movie-ratings-glue-database	-	-	June 14, 2023 at 17:28:43

Last updated (UTC) June 14, 2023 at 17:28:45 Edit Delete Add database

19. CLICK ADD TABLES USING CRAWLER. BASICALLY, THE CRAWLER IS THE TOOL THAT CAN EXTRACT THE DATA IN A CSV FILE. IT EXTRACTS THE METADATA AND CAN ALSO POPULATE THE DATA IN THE CSV FILE.

AWS Glue > Databases > movie-ratings-glue-database

Database properties

Name	movie-ratings-glue-database	Description	-	Location	-	Created on (UTC)	June 14, 2023 at 17:28:45
------	-----------------------------	-------------	---	----------	---	------------------	---------------------------

Tables (0)

Last updated (UTC) June 14, 2023 at 17:29:03

Add tables using crawler Add table

No available tables

20. ENTER UNIQUE CRAWLER NAME. HERE, WE WILL CALL IT MOVE-RATINGS-CRAWLER. CLICK NEXT.

AWS Glue > Crawlers > Add crawler

Set crawler properties

Crawler details

Name: movie-ratings-crawler

Description - optional

Tags - optional

Cancel Next

21. CHOOSE NOT YET IN THE DATA SOURCE CONFIGURATION. THEN CLICK ADD A DATA SOURCE.

AWS Glue > Crawlers > Add crawler

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 (0)

Add a data source

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.

Cancel Previous Next

22. DATA SOURCE IS S3, LOCATION OF S3 DATA IS IN THIS ACCOUNT. THEN, BROWSE S3.

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).

Clear selection Add new connection

Location of S3 data

In this account

In a different account

S3 path

Browse for or enter an existing S3 path.

s3://bucket/prefix/object View Browse S3

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.

Crawl based on events

Rely on Amazon S3 events to control what folders to crawl.

Crawl only a subset of files

23. IN THE BROWSE S3, CHOOSE THE BUCKET LISTED BELOW, THE BUCKET WE CREATED AWHILE AGO.

Choose S3 path

S3 buckets

Buckets (1/1)

Find bucket

Name	Creation date
recommender-system-8730872805	2023-06-14T17:26:11.000Z

Cancel Choose

24. SCROLL DOWN, IN THE SUBSEQUENT CRAWLER RUNS, CHOOSE CRAWL ALL SUB-FOLDERS. THEN CLICK, 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.

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.
 Crawl based on events
Rely on Amazon S3 events to control what folders to crawl.

Sample only a subset of files
 Exclude files matching pattern

25. SELECT S3 IN THE DATA SOURCE. THEN, CLICK, NEXT.

AWS Glue X AWS Glue > Crawlers > Add crawler

Getting started
ETL jobs
Visual ETL
Notebooks
Job run monitoring
Data Catalog tables
Data connections
Workflows (orchestration)

▼ Data Catalog
Databases
Tables
Stream schema registries
Schemas
Connections
Crawlers
Classifiers
Catalog settings

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 (1) Info
The list of data sources to be scanned by the crawler.

Type	Data source	Parameters
S3	s3://recommender-system-8730...	Recrawl all

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.

26. ENTER THE IAM ROLE WE JUST CREATED. THIS IS TO GIVE US FULL PERMISSION FROM GLUE. SELECT THE ROLE THAT WE`VE CREATED, GLUEFULLACCESSROLE. THEN, CLICK NEXT.

AWS Glue

- Getting started
- ETL jobs
- Visual ETL
- Notebooks
- Job run monitoring
- Data Catalog tables
- Data connections
- Workflows (orchestration)

Data Catalog

- Databases
 - Tables
 - Stream schema registries
 - Schemas
 - Connections
 - Crawlers
 - Classifiers
 - Catalog settings

Data Integration and ETL

- ETL jobs
- Visual ETL
- Notebooks
- Job run monitoring

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

Choose an IAM role

[View](#)

AWSReservedSSO_AWSAdministratorAccess_d6359d25c2eb881

Provides full access to AWS services and resources

AWSReservedSSO_AWSPowerUserAccess_72a7e95974f3825f

Provides full access to AWS services and resources, but does not allow management of Users and groups

AWSReservedSSO_AWSReadOnlyAccess_1ecBcf602ddf80ec

This policy grants permission to view resources and basic metadata across all AWS services

AWSReservedSSO_AWSServiceCatalogAdminFullAccess_d1fe8e1088162bc9

Provides full access to AWS Service Catalog admin capabilities

AWSReservedSSO_AWSServiceCatalogEndUserAccess_7c60b01db8043761

Provides access to the AWS Service Catalog end user console

GlueFullAccessRole

Allows Glue to call AWS services on your behalf.

GlueFullAccessRole

Cancel Previous Next

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

GlueFullAccessRole [View](#)

Create new IAM 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 and Iceberg data sources.

► Security configuration - optional

Enable at-rest encryption with a security configuration.

Cancel Previous Next

27. IN THE TARGET DATABASE, SELECT THE ONE THAT WE CREATED.

AWS Glue

Getting started
ETL jobs
Visual ETL
Notebooks
Job run monitoring
Data Catalog tables
Data connections
Workflows (orchestration)

▼ Data Catalog
Databases
Tables
Stream schema registries
Schemas
Connections
Crawlers
Classifiers
Catalog settings

▼ Data Integration and ETL
ETL jobs
Visual ETL
Notebooks
Job run monitoring

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

Set output and scheduling

Output configuration Info

Target database
Choose a database

Table name prefix - optional

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.

► 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 THE FREQUENCY, WE KEEP IT ON DEMAND. THEN, CLICK NEXT.

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

Set output and scheduling

Output configuration Info

Target database
movie-ratings-glue-database

Table name prefix - optional

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.

► 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

28. REVIEW THE CONFIGURATION, THEN, CLICK CREATE CRAWLER.

AWS Glue

Getting started
ETL jobs
Visual ETL
Notebooks
Job run monitoring
Data Catalog tables
Data connections
Workflows (orchestration)

▼ Data Catalog
Databases
Tables
Stream schema registries
Schemas
Connections
Crawlers
Classifiers
Catalog settings

▼ Data Integration and ETL
ETL jobs
Visual ETL
Notebooks
Job run monitoring
Interactive Sessions
Data classification tools
Sensitive data detection

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

Name: movie-ratings-crawler
Description: -
Tags: -

Step 2: Choose data sources and classifiers

Data sources (1) Info
The list of data sources to be scanned by the crawler.

Type	Data source	Parameters
S3	s3://recommender-system-873087...	Recrawl all

Step 3: Configure security settings

Configure security settings

IAM role	Security configuration	Lake Formation configuration
GlueFullAccessRole	-	-

Step 4: Set output and scheduling

Set output and scheduling

Database	Table prefix - optional	Maximum table threshold - optional	Schedule
movie-ratings-glue-database	-	-	On demand

Cancel Previous Create crawler

29. WE NOW HAVE JUST CREATED OUR CRAWLER. WE WILL NOW RUN IT IN ORDER TO DO THIS EXTRACTION OF THE DATA AND PUT IT IN THE TABLE OF THIS DATABASE WE JUST CREATED WITHIN GLUE.

CLICK RUN CRAWLER.

AWS Glue

Getting started
ETL jobs
Visual ETL
Notebooks
Job run monitoring
Data Catalog tables
Data connections
Workflows (orchestration)

▼ Data Catalog
Databases
Tables
Stream schema registries
Schemas
Connections
Crawlers
Classifiers
Catalog settings

▼ Data Integration and ETL
ETL jobs
Visual ETL
Notebooks
Job run monitoring
Interactive Sessions
Data classification tools

One crawler successfully created
The following crawler is now created: "movie-ratings-crawler"

AWS Glue > Crawlers > movie-ratings-crawler

movie-ratings-crawler

Last updated (UTC)
June 14, 2023 at 17:31:22

Run crawler Edit Delete

Crawler properties

Name: movie-ratings-crawler	IAM role: GlueFullAccessRole	Database: movie-ratings-glue-database	State: READY
Description: -	Security configuration: -	Lake Formation configuration: -	Table prefix: -
Maximum table threshold: -			

Advanced settings

Crawler runs

(0) The list of crawler runs for this crawler.

Filter data

Start time (UTC) ▲ End time (UTC) ▼ Current/last duration ▼ Status ▼ DPU hours ▼ Table changes ▼

You don't have any crawler runs.

30. SELECT THE CRAWLER, THEN, CLICK, RUN.

The screenshot shows the AWS Glue interface with the 'Crawlers' section selected. A success message at the top states: 'One crawler successfully created' and 'The following crawler is now created: "movie-ratings-crawler"'. The main table lists one crawler named 'movie-ratings-cr...' with a status of 'Ready'. The table includes columns for Name, State, Schedule, Last run, Last run time, Log, and Table changes.

31. IT IS NOW SUCCESSFULLY RUNNING. IT WILL EXTRACT ALL THE DATA FROM THE CSV FILE TO PUT IT IN THE TABLE OF THE DATABASE WE CREATED. WE CREATED TWO THINGS SO FAR, THE DATABASE FIRST, THEN, WE ARE CREATING THE TABLE OF THIS DATABASE.

The screenshot shows the AWS Glue interface with the 'Crawlers' section selected. A progress message at the top says: 'Starting crawler' and 'Attempting to start run crawler "movie-ratings-crawler"'. The main table lists one crawler named 'movie-ratings-cr...' with a status of 'Ready'. The table includes columns for Name, State, Schedule, Last run, Last run time, Log, and Table changes.

IT IS NOW RUNNING AND IT WILL TAKE APPROX. 2 MINUTES. AT THE END, WE WILL SEE THAT WE HAVE THE MOVE RATINGS TABLE CONTAINING THE EXACT SAME DATA AS IN THE CSV FILE. AND THAT IS THE DATA CATALOG ELEMENT AND THE ELEMENT IS A TABLE WHICH WE WILL THEN CONNECT IN OUR ETL PROCESS.

The screenshot shows the AWS Glue interface with the 'Crawlers' section selected. A success message at the top says: 'Crawler successfully starting' and 'The following crawler is now starting: "movie-ratings-crawler"'. The main table lists one crawler named 'movie-ratings-cr...' with a status of 'Running'. The table includes columns for Name, State, Schedule, Last run, Last run time, Log, and Table changes.

ONCE IT IS STOPPING, IT MEANS THAT THE CRAWLING IS DONE AND SOON IT SHOULD SAY SUCCESSFUL.

Crawler successfully starting
The following crawler is now starting: "movie-ratings-crawler"

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 (1/1) Info		Last updated (UTC)	Action	Run	Create crawler		
View and manage all available crawlers.		June 14, 2023 at 17:53:43					
<input checked="" type="checkbox"/> Name		<input type="checkbox"/> State	<input type="checkbox"/> Schedule	<input type="checkbox"/> Last run	<input type="checkbox"/> Last run time...	<input type="checkbox"/> Log	<input type="checkbox"/> Table changes ...
<input checked="" type="checkbox"/> movie-ratings-cr...			-	-	-	-	1 created

Crawler successfully starting
The following crawler is now starting: "movie-ratings-crawler"

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 (1/1) Info		Last updated (UTC)	Action	Run	Create crawler		
View and manage all available crawlers.		June 14, 2023 at 17:34:50					
<input checked="" type="checkbox"/> Name		<input type="checkbox"/> State	<input type="checkbox"/> Schedule	<input type="checkbox"/> Last run	<input type="checkbox"/> Last run time...	<input type="checkbox"/> Log	<input type="checkbox"/> Tab
<input checked="" type="checkbox"/> movie-ratings-crawler				Succeeded	June 14, 2023 a...	View log	1 cr

32. ONCE READY OR SUCCESSFUL, GO TO THE DATA CATALOG, DATABASE, THEN, TABLES. REFRESH IT TO SEE THE TABLES. CLICK THE TABLE AVAILABLE IN THE LIST.

AWS Glue > **Crawler successfully starting**
The following crawler is now starting: "movie-ratings-crawler"

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 (1)		Last updated (UTC)		Delete	Data quality	New	Add tables using crawler	Action
View and manage all available tables.		June 14, 2023 at 17:55:14						
<input type="checkbox"/> Name		<input type="checkbox"/> Database	<input type="checkbox"/> Location	<input type="checkbox"/> Classification	<input type="checkbox"/> Deprecated	<input type="checkbox"/> View	<input type="checkbox"/> Table	
<input type="checkbox"/> recommender-system_8730872805		movie-ratings-glue-data	s3://recommender-system	csv	-	View	Table	

33. THIS IS NOW OUR TABLE OVERVIEW. THERE, WE CAN`T SEEM TO SEE THE VALUES WHICH GIVES OUR TABLE NOT VERY INTUITIVE. TO HAVE A GREAT LOOK OF OUR TABLE, WE WILL NEED TO USE ANOTHER SERVICE.

The screenshot shows the AWS Glue Table overview page. On the left, there's a sidebar with navigation links for AWS Glue, Data Catalog, and Data Integration. The main area displays a table with the following details:

Name	Description	Database	Classification
recommender_system_8730872805	-	movie-ratings-glue-database	csv

Below this, there's another table section with the following details:

Location	Connection	Deprecated	Last updated
s3://recommender-system-8730872805/	-	-	June 14, 2023 at 17:33:30

At the bottom, there's a schema table with the following columns:

#	Column name	Data type	Partition key	Comment
1	userid	bigint	-	-
2	movieid	bigint	-	-
3	rating	double	-	-
4	timestamp	bigint	-	-

34. IN THE SEARCH BAR, LOOK FOR ATHENA. THIS IS A QUERY SERVICE WHERE WE CAN MANIPULATE THE DATA SO WE CAN HAVE A GOOD LOOK OF OUR DATA.

The screenshot shows the AWS search interface with the query 'Athena' entered in the search bar. The results are categorized into Services and Features.

Services (1)

- Athena: Serverless interactive analytics service

Features (9)

- Data sources: Athena feature
- Workgroups: Athena feature
- Notebooks: Athena feature
- SQL queries: Athena Feature

HERE, YOU WILL SEE THE DATA SOURCE, DATABASE, AND THE TABLE THAT WE JUST CREATED.

The screenshot shows the Amazon Athena Query editor interface. In the top navigation bar, 'Amazon Athena' and 'Query editor' are selected. The 'Editor' tab is active. On the left, the 'Data' sidebar shows 'Data source' set to 'AwsDataCatalog' and 'Database' set to 'movie-ratings-glue-database'. Under 'Tables and views', there is a 'Create' button and a search bar. Below that, the 'Tables' section shows '(1)' and the table 'recommender_system_8730872805'. The main area is titled 'Query 1' with a line number '1'. It contains a SQL editor with the placeholder 'SQL Ln 1, Col 1' and a toolbar with 'Run', 'Explain', 'Cancel', 'Clear', and 'Create'. Below the SQL editor is the 'Query results' tab, which is currently selected, showing a 'Results' table with one row. There are buttons for 'Copy' and 'Download results'.

35. TO LOOK FOR THE TABLE, CLICK THE THREE DOTS ON THE RIGHT SIDE OF THE TABLE, THEN, CLICK PREVIEW TABLE.

This screenshot is similar to the previous one, showing the Amazon Athena Query editor. The 'Data' sidebar and 'Tables' section are identical. However, a context menu is open over the table 'recommender_system_8730872805'. The menu options include 'Run Query', 'Preview Table' (which is highlighted in blue), 'Generate table DDL', 'Insert', 'Insert into editor', 'Manage', 'Delete table', 'View properties', and 'View in Glue'. The main area below the menu is the same as the first screenshot, showing the 'Query results' tab with a single row of data.

36. THIS WILL BE THE RESULT OR THE PREVIEW OF THE TABLE. IN THE QUERY SECTION, WE LIMIT THE TABLE TO 10. YOU CAN SET THE LIMIT BASED ON YOUR PREFERENCE.

Amazon Athena > Query editor

Editor Recent queries Saved queries Settings Workgroup primary

Data

Data source: AwsDataCatalog
Database: movie-ratings-glue-database
Tables and views: Tables (1) Views (0)

Query 1 : X | **Query 2 : X**

```
1 SELECT * FROM "movie-ratings-glue-database"."recommender_system_8730872805" limit 10;
```

SQL Ln 1, Col 85

Run again Explain Cancel Clear Create Reuse query results *Athena engine version 3 only

Query results | Query stats

Completed Time in queue: 113 ms Run time: 623 ms Data scanned: 690.49 KB

Results (10)

Run again Explain Cancel Clear Create Reuse query results *Athena engine version 3 only

Query results | Query stats

Completed Time in queue: 113 ms Run time: 623 ms Data scanned: 690.49 KB

Results (10)

Search rows < 1 >

#	userid	movieid	rating	timestamp
1	1	110	1.0	1425941529
2	1	147	4.5	1425942435
3	1	858	5.0	1425941523
4	1	1221	5.0	1425941546
5	1	1246	5.0	1425941556
6	1	1968	4.0	1425942148
7	1	2762	4.5	1425941300
8	1	2918	5.0	1425941593
9	1	2959	4.0	1425941601
10	1	4226	4.0	1425942228

37. BACK INTO THE AWS GLUE, WE ARE DONE WITH THE FIRST DATA SOURCE, WE HAVE CREATED THE TABLE, AND THE TABLE IS POPULATED WITH THE DATA.

Crawler successfully starting
The following crawler is now starting: "movie-ratings-crawler"

AWS Glue > Tables > recommender_system_8730872805

recommender_system_8730872805

Last updated (UTC) June 14, 2023 at 17:33:30 Version 0 (Current version) Actions

Table overview | Data quality New

Table details | Advanced properties

Name recommender_system_8730872805	Description -	Database movie-ratings-glue-database	Classification CSV
Location s3://recommender-system-8730872805/	Connection -	Deprecated -	Last updated June 14, 2023 at 17:33:30
Input format org.apache.hadoop.mapred.TextInputFormat	Output format org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe	Serde serialization lib org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe	

Schema | Partitions | Indexes

Schema (4)
View and manage the table schema.

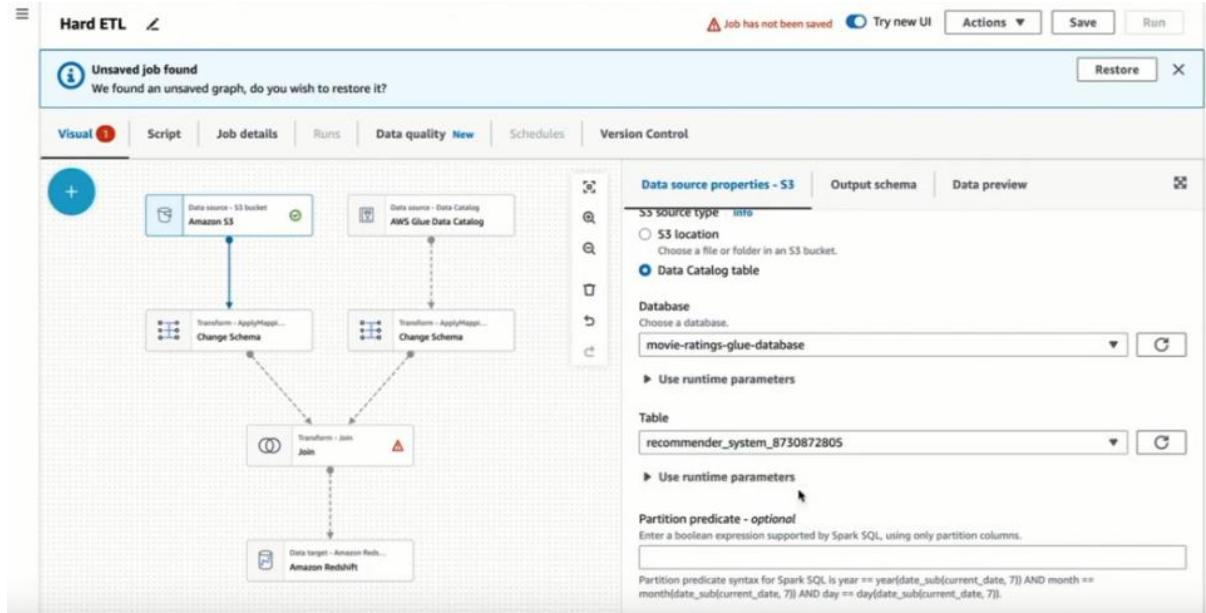
Edit schema as JSON | Edit schema | Filter schemas

38. WE NOW GO TO THE VISUAL EDITOR AND WE ARE GOING TO CONNECT THE DATA. WE JUST CREATED THE DATA SOURCE S3, BUT, NOW WE ARE GOING TO CONNECT IT.

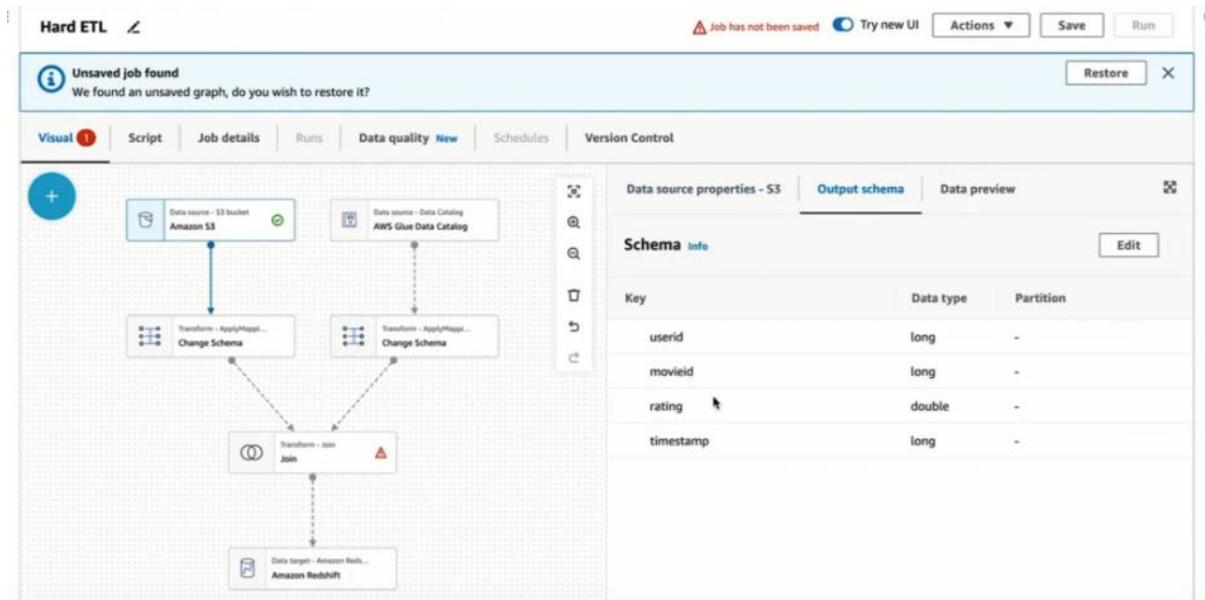
IN THE S3 SOURCE TYPE, CLICK DATA CATALOG TABLE.

IN THE DATABASE, CLICK THE DATABASE WE CREATED, MOVIE-RATINGS-GLUE-DATABASE.

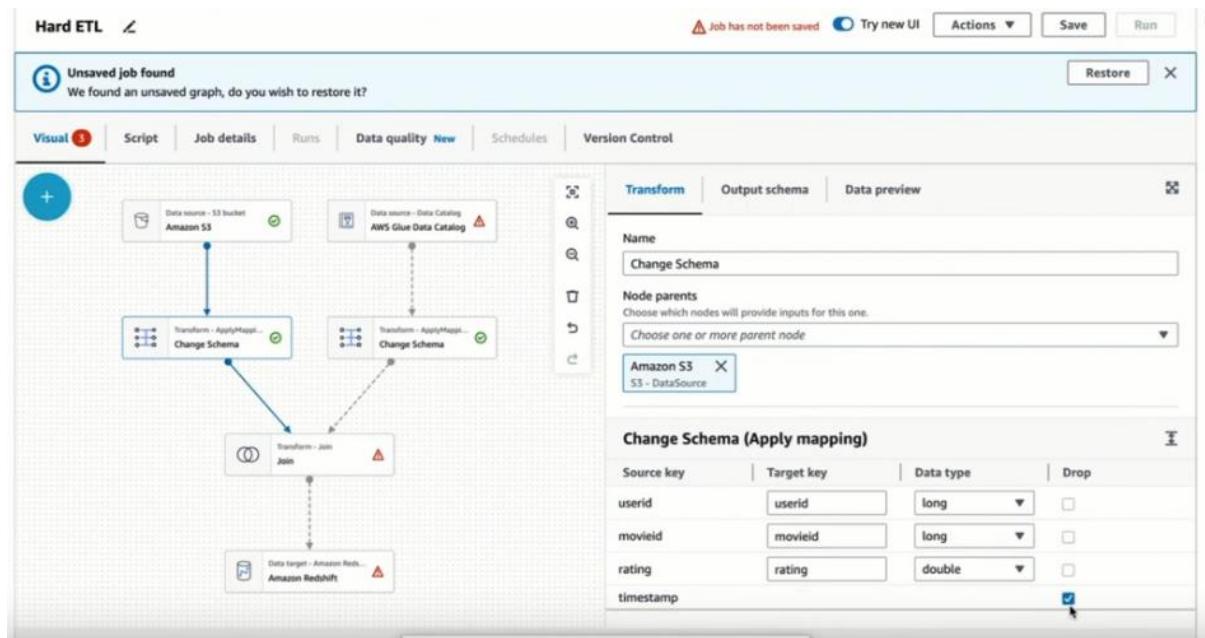
IN THE TABLE, CLICK THE TABLE THAT WAS CREATED.



39. IN THE OUTPUT SCHEMA, KEEP IT DEFAULT.

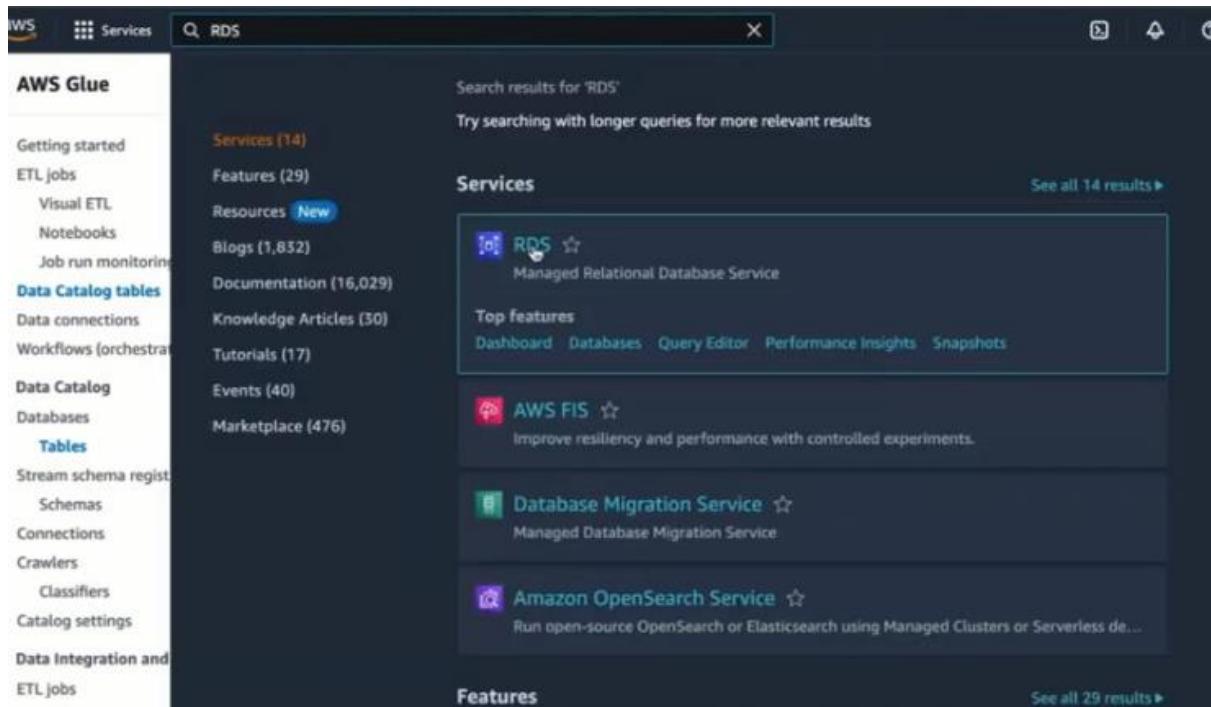


40. PROCEED TO THE TRANSFORM-APPLYMAPPING PROCEDURE. SINCE WE ARE BUILDING A RECOMMENDER SYSTEM, WE CAN CHECK OR CHANGE THE SCHEMA BASED ON THEIR RELEVANCE. HERE, WE CHOOSE TO DROP THE TIMESTAMP WHICH WE FOUND IRRELEVANT IN BUILDING THE RECOMMENDER SYSTEM.

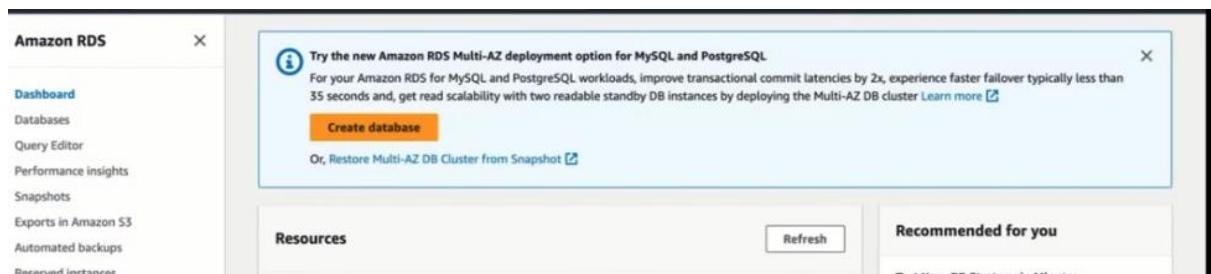


THIS IS THE KIND OF TRANSFORMATION WE CAN DO HERE WITH THIS APPLY MAPPING TRANSFORM. THE FIRST DATA SOURCE IS DONE. WE HAVE EXTRACTED IT PROPERLY WITH THE DATA CATALOG AND THEN WE HAVE TRANSFORMED IT PROPERLY THROUGH THE APPLY MAPPING TRANSFORM PROCEDURE.

41. PROCEED TO THE SECOND DATA SOURCE, THE RDS. IN THE SERVICES, LOOK FOR RDS.



42. CLICK CREATE DATABASE.



43. WE WILL CREATE A RDS DATABASE POWERED BY MYSQL. SO, CHOOSE EASY CREATE. THEN, CLICK MYSQL IN THE ENGINE TYPE OF CONFIGURATION.

Create database

Choose a database creation method Info

Standard create

You set all of the configuration options, including ones for availability, security, backups, and maintenance.

Easy create

Use recommended best-practice configurations. Some configuration options can be changed after the database is created.

Configuration

Engine type Info

Aurora (MySQL Compatible)



Aurora (PostgreSQL Compatible)



MySQL



MariaDB

PostgreSQL

Oracle

44. CHOOSE FREE TIER.

Edition

MySQL Community

DB instance size

Production

db.r6g.xlarge
4 vCPUs
32 GiB RAM
500 GiB
1.017 USD/hour

Dev/Test

db.r6g.large
2 vCPUs
16 GiB RAM
100 GiB
0.231 USD/hour

Free tier

db.t3.micro
2 vCPUs
1 GiB RAM
20 GiB
0.020 USD/hour

DB instance identifier

Type a name for your DB instance. The name must be unique across all DB instances owned by your AWS account in the current AWS Region.

45. GIVE DB INSTANCE IDENTIFIER A NAME. WE WILL CALL IT, customer-features-rds-db-instance. KEEP THE MASTER USERNAME AS ADMIN. CHOOSE THE PASSWORD, THEN, CONFIRM, IT. KEEP THE OTHERS AS DEFAULT. THEN CLICK CREATE.

DB instance identifier
Type a name for your DB instance. The name must be unique across all DB instances owned by your AWS account in the current AWS Region.

The DB instance identifier is case-insensitive, but is stored as all lowercase (as in "mydbinstance"). Constraints: 1 to 60 alphanumeric characters or hyphens. First character must be a letter. Can't contain two consecutive hyphens. Can't end with a hyphen.

Master username [Info](#)
Type a login ID for the master user of your DB instance.

1 to 16 alphanumeric characters. First character must be a letter.

Auto generate a password
Amazon RDS can generate a password for you, or you can specify your own password.

Master password [Info](#)

Constraints: At least 8 printable ASCII characters. Can't contain any of the following: / (slash), '(single quote)', "(double quote)" and @ (at sign).

Confirm master password [Info](#)

► Set up EC2 connection - optional
You can also set up a connection to an EC2 instance after creating the database. Go to the database list page or the database details page, choose **Actions**, and then choose **Set up to EC2 connection**.

► View default settings for Easy create
Easy create sets the following configurations to their default values, some of which can be changed later. If you want to change

46. THE CREATING STATUS WILL TAKE UP TO FEW MINUTES.

Amazon RDS

Databases

RDS > Databases

Databases (1)

DB identifier	Status	Role	Engine	Region & AZ	Size	Actions
customer-features-rds-db-instance	Creating	Instance	MySQL Community	-	db.t3.micro	-

Creating database customer-features-rds-db-instance

Your database might take a few minutes to launch.
You can use settings from customer-features-rds-db-instance to simplify configuration of suggested database add-ons while we finish creating your DB for you.

How was your experience creating an Amazon RDS database? [Provide feedback](#)

Consider creating a Blue/Green Deployment to minimize downtime during upgrades

You may want to consider using Amazon RDS Blue/Green Deployments and minimize your downtime during upgrades. A Blue/Green Deployment provides a staging environment for changes to production databases. [RDS User Guide](#) [Aurora User Guide](#)

The screenshot shows the Amazon RDS Databases page. At the top, a green banner indicates that the database 'customer-features-rds-db-instance' has been successfully created. It also includes a link to 'View connection details' and a feedback section. Below the banner, the main content area shows a table of databases. A tooltip suggests creating a Blue/Green deployment. The table lists one database entry:

DB identifier	Status	Role	Engine	Region & AZ	Size	Actions
customer-features-rds-db-instance	Backing-up	Instance	MySQL Community	us-east-1c	db.t3.micro	-

47. WE WILL NOW CREATE A TABLE CONTAINING THE CUSTOMER FEATURES DATA IN THE CSV FILE INTO THE DATABASE. WE CAN DO IT BY EITHER USING THE CLI OR AN EXTRA TOOL WHICH IS CALLED MYSQL WORKBENCH.

IF YOU DO NOT HAVE THE MYSQL WORKBENCH, YOU CAN GO TO GOOGLE AND SEARCH FOR MYSQL WORKBENCH, THEN, DOWNLOAD, THEN, INSTALL.

<https://dev.mysql.com/downloads/workbench/>

② MySQL Community Downloads

◀ MySQL Workbench

General Availability (GA) Releases Archives

MySQL Workbench 8.0.36

Select Operating System:

Microsoft Windows

Recommended Download:

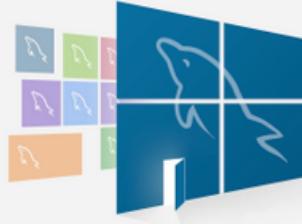
MySQL Installer
for Windows

All MySQL Products. For All Windows Platforms.
In One Package.

Starting with MySQL 5.6 the MySQL Installer package replaces the standalone MSI packages.

Windows (x86, 32 & 64-bit), MySQL Installer MSI

Go to Download Page >



Other Downloads:

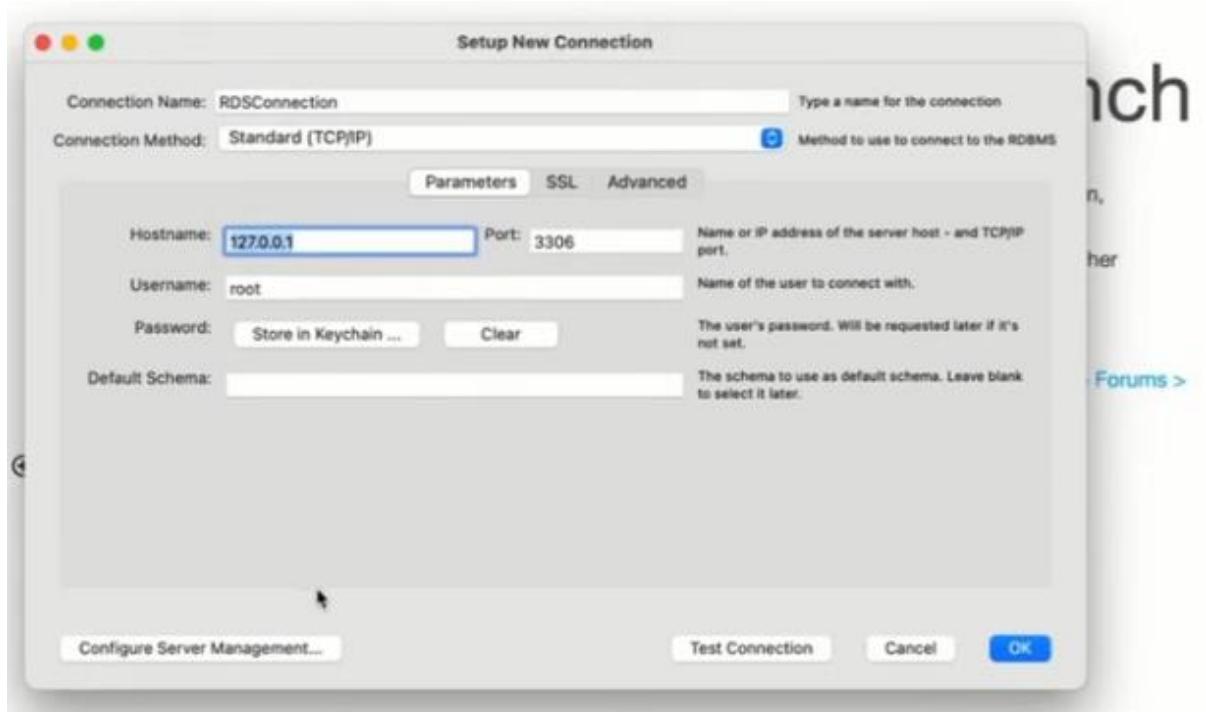
Windows (x86, 64-bit), MSI Installer (mysql-workbench-community-8.0.36-winx64.msi)	8.0.36	42.0M	Download
			MD5: 2156fe0cb6f5ed83908e4636ba86390a Signature

We suggest that you use the [MD5 checksums](#) and [GnuPG signatures](#) to verify the integrity of the packages you download.

- 48. ONCE INSTALLED, MAKE A CONNECTION BETWEEN MYSQL WORKBENCH AND RDS DATABASE. OPEN MYSQL WORKBENCH, CLICK THE PLUS (+) BUTTON TO CREATE THE CONNECTION.**



FOR THE CONNECTION NAME, MAKE IT RDSCONNECTION. HOST NAME IS ACCORDING TO YOU WHAT IS GOING TO BE. TO CHECK, GO TO YOUR AMAZON RDS, AND CHECK FOR THE ENDPOINT & PORT.



COPY THE ENDPOINT DETAILS. GO BACK TO THE MYSQL WORKBENCH, THEN, PASTE IT TO THE HOSTNAME. THE PORT 3306 IS THE PORT OF MY MYSQL, SO, IT WILL REMAIN THE SAME.

Amazon RDS

- Dashboard
- Databases**
- Query Editor
- Performance insights
- Snapshots
- Exports in Amazon S3
- Automated backups
- Reserved instances
- Proxies
- Subnet groups
- Parameter groups
- Option groups
- Custom engine versions
- Events
- Event subscriptions

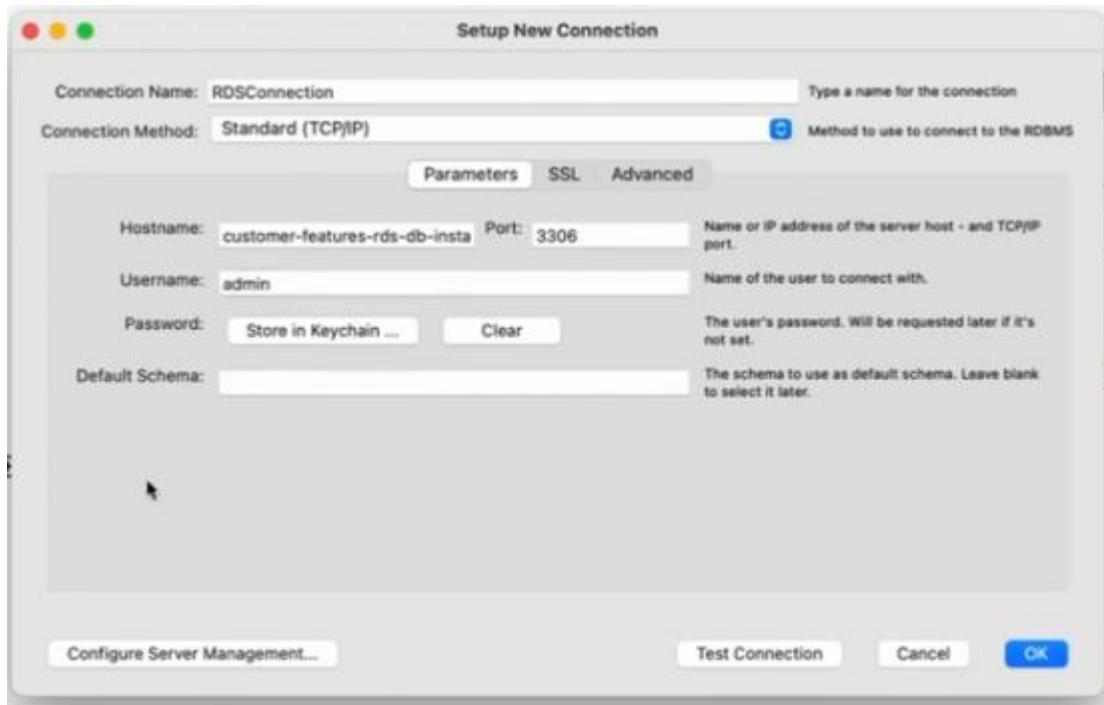
Connectivity & security | Monitoring | Logs & events | Configuration | Maintenance & backups | Tags

Connectivity & security

Endpoint & port	Networking	Security
Endpoint customer-features-rds-db-instance.crrwucvfcmbc.us-east-1.rds.amazonaws.com	Availability Zone us-east-1c	VPC security groups default (sg-00757adc39bf9625e) <input checked="" type="checkbox"/> Active
Port 3306	VPC vpc-0041e736a2d14cfa2	Publicly accessible No
	Subnet group default-vpc-0041e736a2d14cfa2	Certificate authority Info rds-ca-2019
	Subnets subnet-061069d72790ee8be subnet-0ab745a585106a535 subnet-0536055a78fadcc9a subnet-0c09e23eb5408adff0 subnet-066078a3fb43745dc subnet-0ff66ccebd5b8ad21	Certificate authority date August 22, 2024, 19:08 (UTC+02:00)
	Network type IPv4	DB instance certificate expiration date August 22, 2024, 19:08 (UTC+02:00)

FOR THE USERNAME, PUT admin, AS IT IS THE NAME WE PUT IN THE DB INSTANCE.

THEN, CICK, TEST CONNECTION TO TEST IF IT WORKS PROPERLY.



49. THE RESULT WILL GIVE US FAILED CONNECTION.



50. GO BACK TO AMAZON RDS, IF YOU NOTICE IN THE CONNECTIVITY & SECURITY, IN THE SECURITY FIELD - PUBLICLY ACCESSIBLE IS NO.

Endpoint & port	Networking	Security
Endpoint customer-features-rds-db-instance.crrwucvfcmbc.us-east-1.rds.amazonaws.com	Availability Zone us-east-1c	VPC security groups default (sg-00757adc39bf9625e) Active
Port 3306	VPC vpc-0041e736a2d14cfa2	Publicly accessible No
	Subnet group default-vpc-0041e736a2d14cfa2	Certificate authority rds-ca-2019
	Subnets subnet-061069d72790ee8be subnet-0ab745a585106a535 subnet-0536055a78fadcc9a subnet-0c09e23eb408adff0 subnet-066078a3fb45745dc subnet-0ff66ccebd5b8ad21	Certificate authority date August 22, 2024, 19:08 (UTC+02:00)
	Network type IPv4	DB instance certificate expiration date August 22, 2024, 19:08 (UTC+02:00)

51. WE HAVE TO CONFIGURE THE SECURITY – PUBLICLY AVAILABLE TO YES. TO DO THIS, CLICK MODIFY.

⌚ Successfully created database **customer-features-rds-db-instance**
You can use settings from customer-features-rds-db-instance to simplify configuration of [suggested database add-ons](#) while we finish creating your DB for you.
How was your experience creating an Amazon RDS database? [Provide feedback](#)

RDS > Databases > customer-features-rds-db-instance

customer-features-rds-db-instance

[Modify](#) [Actions ▾](#)

Summary			
DB identifier customer-features-rds-db-instance	CPU <div style="width: 5.54%;">5.54%</div>	Status Available	Class db.t3.micro
Role Instance	Current activity <div style="width: 0%;">0 Connections</div>	Engine MySQL Community	Region & AZ us-east-1c

[Connectivity & security](#) [Monitoring](#) [Logs & events](#) [Configuration](#) [Maintenance & backups](#) [Tags](#)

Connectivity & security

Endpoint & port Networking Security

SCROLL DOWN, AND FIND THE CONFIGURATION WHERE WE CAN SET THE PUBLIC ACCESS TO YES. IN THE CONNECTIVITY, CLICK ADDITIONAL CONFIGURATION. SELECT, PUBLICLY ACCESSIBLE.

Amazon RDS

Dashboard

Databases

- Query Editor
- Performance insights
- Snapshots
- Exports in Amazon S3
- Automated backups
- Reserved instances
- Proxies

Subnet groups

Parameter groups

Option groups

Custom engine versions

Events

Event subscriptions

Recommendations 1

Connectivity

Network type [Info](#)
To use dual-stack mode, make sure that you associate an IPv6 CIDR block with a subnet in the VPC you specify.

IPv4
Your resources can communicate only over the IPv4 addressing protocol.

Dual-stack mode
Your resources can communicate over IPv4, IPv6, or both.

DB subnet group
default-vpc-0041e736a2d14cfa2

Security group
List of DB security groups to associate with this DB instance.
[Choose security groups](#)

default X

Certificate authority [Info](#)
Using a server certificate provides an extra layer of security by validating that the connection is being made to an Amazon database. It does so by checking the server certificate that is automatically installed on all databases that you provision.

rds-ca-2019

Additional configuration

Public access

Choose security groups ▾

default X

Certificate authority [Info](#)

Using a server certificate provides an extra layer of security by validating that the connection is being made to an Amazon database. It does so by checking the server certificate that is automatically installed on all databases that you provision.

rds-ca-2019 ▾

▼ Additional configuration

Public access

Publicly accessible

RDS assigns a public IP address to the database. Amazon EC2 instances and other resources outside of the VPC can connect to your database. Resources inside the VPC can also connect to the database. Choose one or more VPC security groups that specify which resources can connect to the database.

Not publicly accessible

No IP address is assigned to the DB instance. EC2 instances and devices outside the VPC can't connect.

Database port

Specify the TCP/IP port that the DB instance will use for application connections. The application connection string must specify the port number. The DB security group and your firewall must allow connections to the port. [Learn more](#) ⓘ

3306

SCROLL DOWN, THEN SELECT, CONTINUE. YOU WILL SEE WHAT IS MODIFIED.

CLICK MODIFY DB INSTANCE.

Modify DB instance: customer-features-rds-db-instance

Summary of modifications

You are about to submit the following modifications. Only values that will change are displayed. Carefully verify your changes and click Modify DB Instance.

Attribute	Current value	New value
Public accessibility	No	Yes

Schedule modifications

When to apply modifications

- Apply during the next scheduled maintenance window

Current maintenance window: June 17, 2023 10:19 - 10:49 UTC+2

- Apply immediately

The modifications in this request and any pending modifications will be asynchronously applied as soon as possible, regardless of the maintenance window setting for this database instance.

[Cancel](#)
[Back](#)
[Modify DB instance](#)

Modify DB instance: customer-features-rds-db-instance

Summary of modifications

You are about to submit the following modifications. Only values that will change are displayed. Carefully verify your changes and click Modify DB Instance.

Attribute	Current value	New value
Public accessibility	No	Yes

Schedule modifications

When to apply modifications

- Apply during the next scheduled maintenance window

Current maintenance window: June 17, 2023 10:19 - 10:49 UTC+2

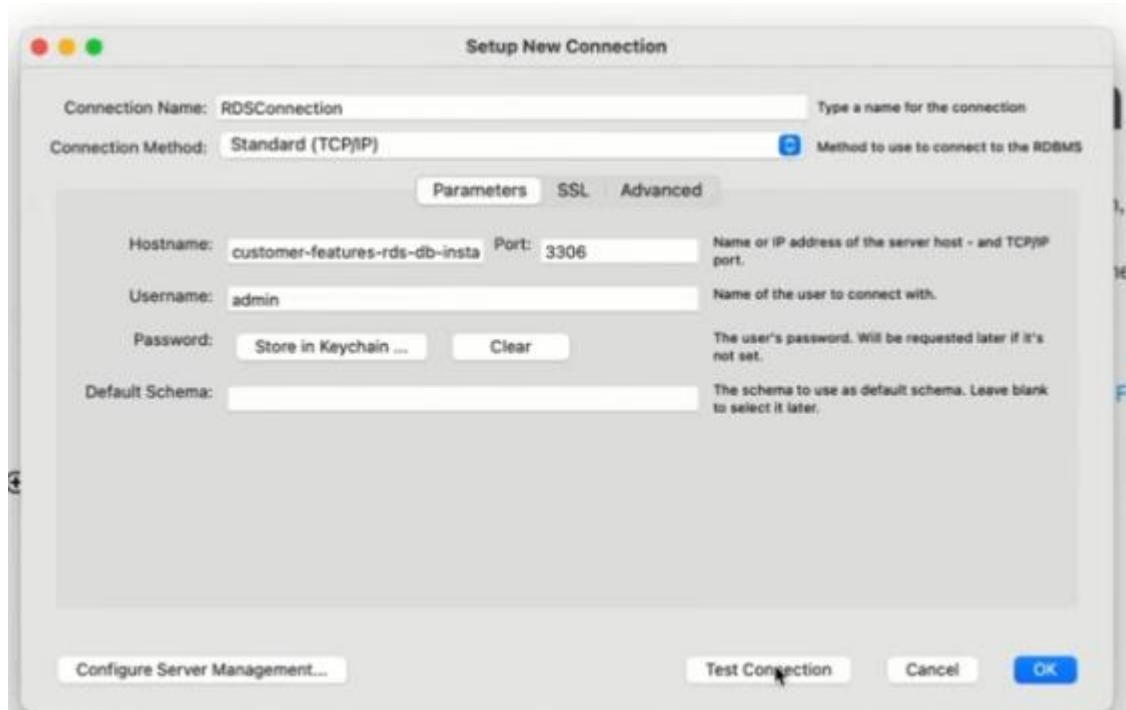
- Apply immediately

The modifications in this request and any pending modifications will be asynchronously applied as soon as possible, regardless of the maintenance window setting for this database instance.

SUCCESSFULLY MOODIED INSTANCE.

The screenshot shows the Amazon RDS console with a success message: "Successfully created database customer-features-rds-db-instance". It also displays a note about using a Blue/Green deployment for downtime minimization. Below this, the "Databases" section lists one instance: "customer-features-rds-db-instance" (Status: Available, Instance Type: MySQL Community, Region: us-east-1c, Size: db.t3.micro).

52. GO BACK TO MYSQL WORKBENCH, THEN, TEST THE CONNECTION.



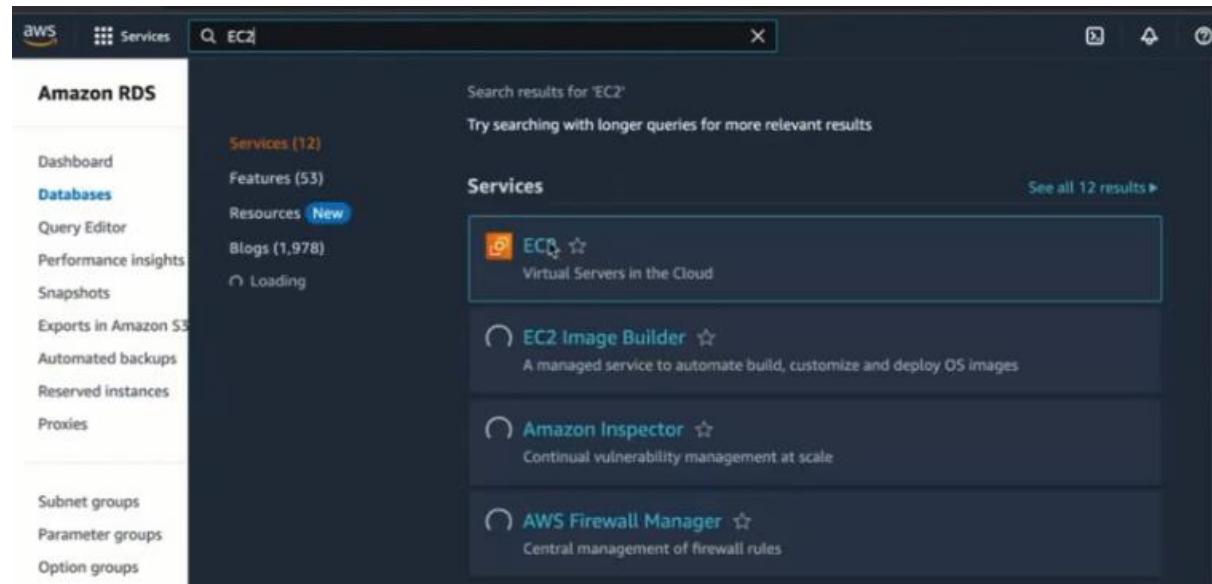
53. HERE, WE WILL GET THE FAILED CONNECTION AGAIN.



54. THIS IS BECAUSE WE FORGOT TO EDIT THE INBOUND RULE. SO, TO ACHIEVE THE CONNECTION, WE NEED TO FIX THE ACCESS ISSUE, SPECIFICALLY THE INBOUND RULES OF THE SECURITY GROUP ATTACHED TO THE RDS DATABASE INSTANCE.

SO NOW, WE ARE GOING TO EDIT THE INBOUND RULE IN ORDER TO OPEN THE PORT 3306, THE MYSQL PORT.

IN THE SEARCH BAR, GO TO EC2.



55. GO TO THE SECURITY GROUP AND CREATE SECURITY GROUP.

Name	Security group ID	Security group name	VPC ID	Description	Owner
-	sg-0ce2e3b4e3be04e52	SG-Open-HTTP	vpc-0041e736a2d14cfa2	Allows HTTP Traffic	7496011
-	sg-04c7ce4cb6a0e059f	launch-wizard-1	vpc-0041e736a2d14cfa2	launch-wizard-1 create...	7496011
-	sg-00757adc39bf9625e	default	vpc-0041e736a2d14cfa2	default VPC security gr...	7496011

56. WE WILL CALL IT, SG-OPEN-MYSQL. DESCRIPTION IS ALLOWS MYSQL ACCESS TO DEVELOPERS. AND MAKE THE VPC DEFAULT.

A security group acts as a virtual firewall for your instance to control inbound and outbound traffic. To create a new security group, complete the fields below.

Basic details

Security group name [Info](#)
SG-Open-MySQL
Name cannot be edited after creation.

Description [Info](#)
Allows MySQL Access to Developers

VPC [Info](#)
vpc-0041e736a2d14cfa2

Inbound rules [Info](#)

This security group has no inbound rules.

Add rule

57. CLICK ADD RULE IN THE INBOUND RULE. IN THE TYPE, SELECT CUSTOM TCP. PROTOCOL IS TCP. PORT IS 3306. DESTINATION IS ANYWHERE.

The screenshot shows the AWS VPC configuration interface. At the top, there's a search bar with the ID 'vpc-0041e736a2d14cfa2'. Below it, the 'Inbound rules' section is visible, showing a single rule: 'Custom TCP' on port 3306 from 'Anywhere' to '0.0.0.0/0'. An 'Add rule' button is present. The 'Outbound rules' section below it also shows a single rule: 'All traffic' on port All to 'Custom' destination '0.0.0.0/0'. An 'Add rule' button is also here.

58. MAKE THE OUTBOUND RULE DEFAULT. THEN, CLICK CREATE SECURITY GROUP.

This screenshot shows the 'Create security group' wizard. It's on the 'Configure security group' step. The 'Outbound rules' section is set up with 'All traffic' on port All to 'Custom' destination '0.0.0.0/0'. An 'Add rule' button is available. Below this, the 'Tags - optional' section indicates 'No tags associated with the resource.' and provides an 'Add new tag' button. At the bottom right, there are 'Cancel' and 'Create security group' buttons, with the latter being orange.

59. IT IS NOW IN THE SECURITY GROUP LIST. BUT, WE HAVE TO ATTACH IT TO THE RDS DATABASE INSTANCE.

New EC2 Experience Tell us what you think

EC2 Dashboard

EC2 Global View

Events

Limits

Instances

- Instances
- Instance Types
- Launch Templates
- Spot Requests
- Savings Plans
- Reserved Instances
- Dedicated Hosts

Security Groups (4) Info

Filter security groups

Name	Security group ID	Security group name	VPC ID	Description	Owner
-	sg-04c7ce4cb6a0e059f	launch-wizard-1	vpc-0041e736a2d14cfa2	launch-wizard-1 create...	7496011
-	sg-0ce2e3b4e3be04e92	SG-Open-HTTP	vpc-0041e736a2d14cfa2	Allows HTTP Traffic	7496011
-	sg-00757adc39bf9625e	default	vpc-0041e736a2d14cfa2	default VPC security gr...	7496011
-	sg-05eae22adcf0dc56	SG-Open-MySQL	vpc-0041e736a2d14cfa2	Allows MySQL Access t...	7496011

New EC2 Experience Tell us what you think

EC2 Dashboard

EC2 Global View

Events

Limits

Instances

- Instances
- Instance Types
- Launch Templates
- Spot Requests
- Savings Plans
- Reserved Instances
- Dedicated Hosts
- Scheduled Instances
- Capacity Reservations

Images

- AMIs
- AMI Catalog

Elastic Block Store

- Volumes

sg-05eae22adcf0dc56 - SG-Open-MySQL

Details

Security group name	sg-05eae22adcf0dc56	Description	vpc-0041e736a2d14cfa2
Owner	749601114921	Inbound rules count	1 Permission entry
		Outbound rules count	1 Permission entry

Inbound rules Outbound rules Tags

You can now check network connectivity with Reachability Analyzer

Inbound rules (1/1)

Name	Security group rule...	IP version	Type	Protocol	Port range
-	sgr-000a62b4b15cc1b...	IPv4	MySQL/Aurora	TCP	3306

60. GO BACK TO AMAZON RDS, THEN, CLICK MODIFY.

Amazon RDS

Dashboard

Databases

Query Editor

Performance insights

Snapshots

Exports in Amazon S3

Automated backups

Reserved instances

Proxies

Subnet groups

Parameter groups

Option groups

Custom engine versions

Events

Event subscriptions

Recommendations

RDS > Databases > customer-features-rds-db-instance

customer-features-rds-db-instance

Summary

DB identifier customer-features-rds-db-instance	CPU 3.38%	Status Available	Class db.t3.micro
Role Instance	Current activity 0 Connections	Engine MySQL Community	Region & AZ us-east-1c

Connectivity & security Monitoring Logs & events Configuration Maintenance & backups Tags

Connectivity & security

Endpoint & port	Networking	Security
Endpoint customer-features-rds-db-instance.crrwucvcmhc.us-east-1.rds.amazonaws.com	Availability Zone us-east-1c	VPC security groups default (sg-00757adc39bf9625e) Active
Port	VPC vpc-0041e736a2d14cfa2	Publicly accessible Yes

61. SCROLL DOWN AND GO TO CONNECTIVITY, THEN, SECURITY GROUP. THEN YOU WILL FIND THE ONE THAT WE JUST CREATED. CHOOSE DEAFULT AND SG-OPEN-MYSQL.

The screenshot shows the 'Connectivity' configuration page for an Amazon RDS instance. On the left, there's a sidebar with various navigation options like Dashboard, Databases, and Security groups. The main area is titled 'Connectivity' and contains settings for network type (IPv4 selected), DB subnet group (set to 'default-vpc-0041e736a2d14cfa2'), and security groups. Under 'Choose security groups', the 'default' group is checked, and 'SG-Open-MySQL' is also listed. A note at the bottom right says 'Changes made to an Amazon database are provisioned automatically.' At the bottom, there's a link for 'Additional configuration'.

62. SCROLL DOWN AND CLICK CONTINUE.

The screenshot shows the configuration page for a new Amazon RDS instance. The left sidebar has the same navigation as the previous screen. The main area includes sections for log types (Audit log, Error log, General log, Slow query log), IAM role (RDS service-linked role), maintenance (Enable auto minor version upgrade checked, weekly maintenance window from Saturday 08:00 UTC to Sunday 05:00 UTC), and deletion protection (Enable deletion protection unchecked). At the bottom right, there are 'Cancel' and 'Continue' buttons, with 'Continue' being highlighted.

63. HERE, YOU WILL SEE THE SUMMARY OF MODIFICATIONS. THEN, CLICK MODIFY DB INSTANCE.

The screenshot shows the 'Modify DB instance' page for the database 'customer-features-rds-db-instance'. On the left, there's a sidebar with various RDS management links like Dashboard, Databases, Query Editor, etc. The main area has a title 'Modify DB instance: customer-features-rds-db-instance'. Below it is a 'Summary of modifications' section with a note about upcoming changes. A table shows the modification details: changing the 'Security group' from 'default' to 'default, SG-Open-MySQL'. Under 'Schedule modifications', it shows two options: 'Apply during the next scheduled maintenance window' (selected) and 'Apply immediately'. The 'Apply during the next scheduled maintenance window' option includes a note about the current window being June 17, 2023, 10:19 - 10:49 UTC+2. The 'Modify DB instance' button at the bottom is highlighted in orange.

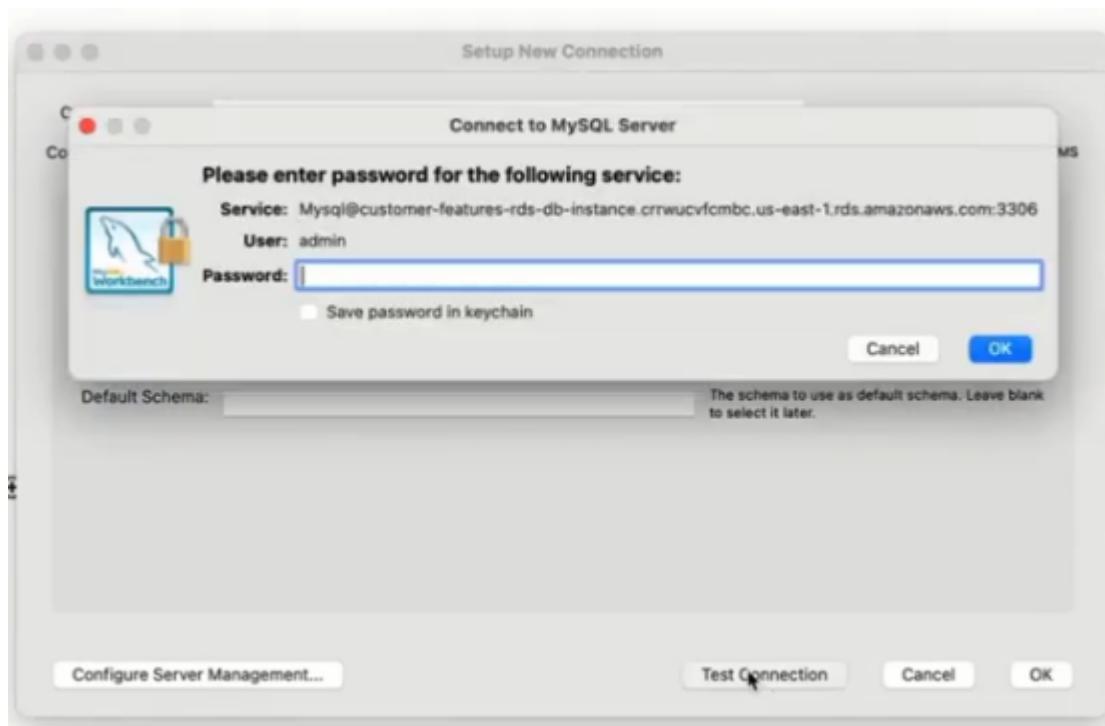
64. SUCCESSFULLY MODIFIED.

The screenshot shows the 'Databases' page with one entry: 'customer-features-rds-db-instance'. A green success message at the top says 'Successfully modified instance customer-features-rds-db-instance'. A tooltip suggests creating a Blue/Green deployment. The table lists the database details: Status (Available), Role (Instance), Engine (MySQL Community), Region & AZ (us-east-1c), Size (db.t3.micro), and Actions (1 Action). The 'Actions' column contains a dropdown menu with options like 'Edit', 'Delete', and 'Clone'.

65. GO BACK TO MYSQL WORKBENCH TO TEST THE CONNECTION.



CONGRATULATIONS! THE CONNECTION IS SUCCESSFUL! ENTER THE PASSWORD.

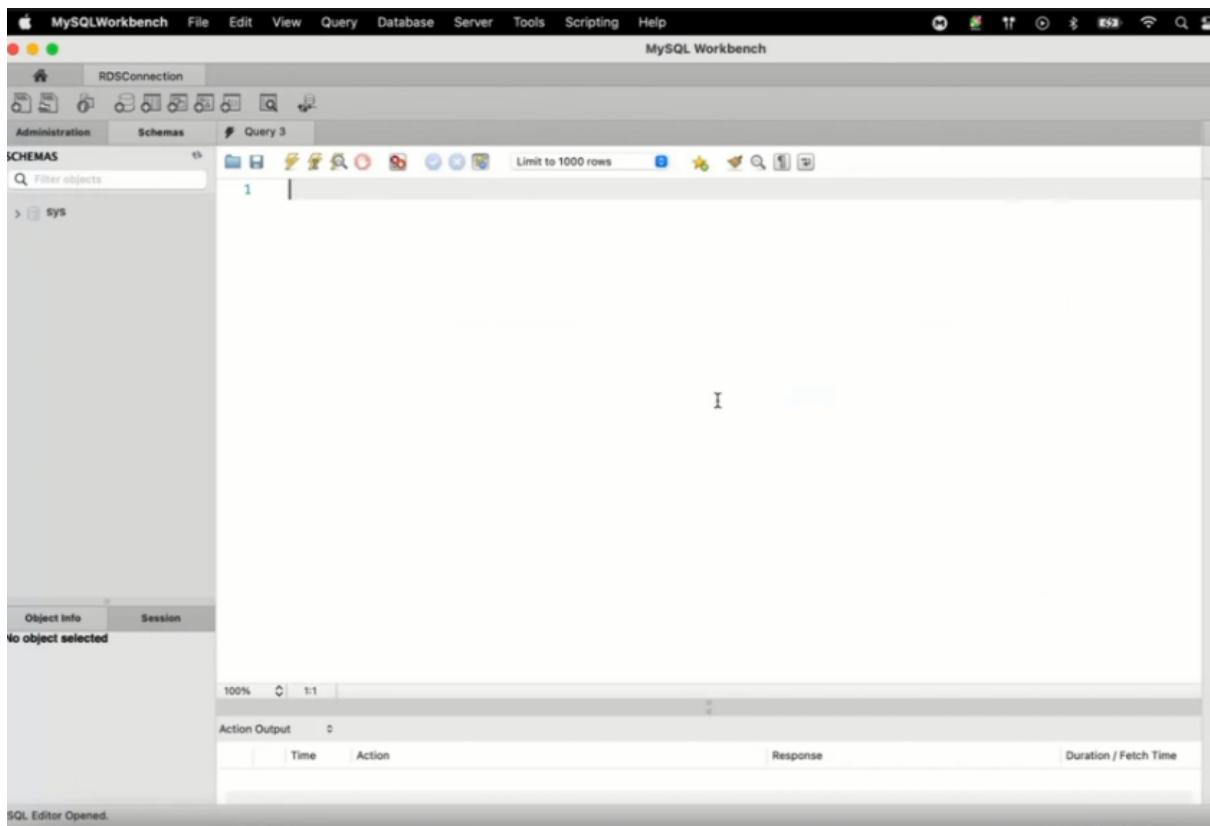




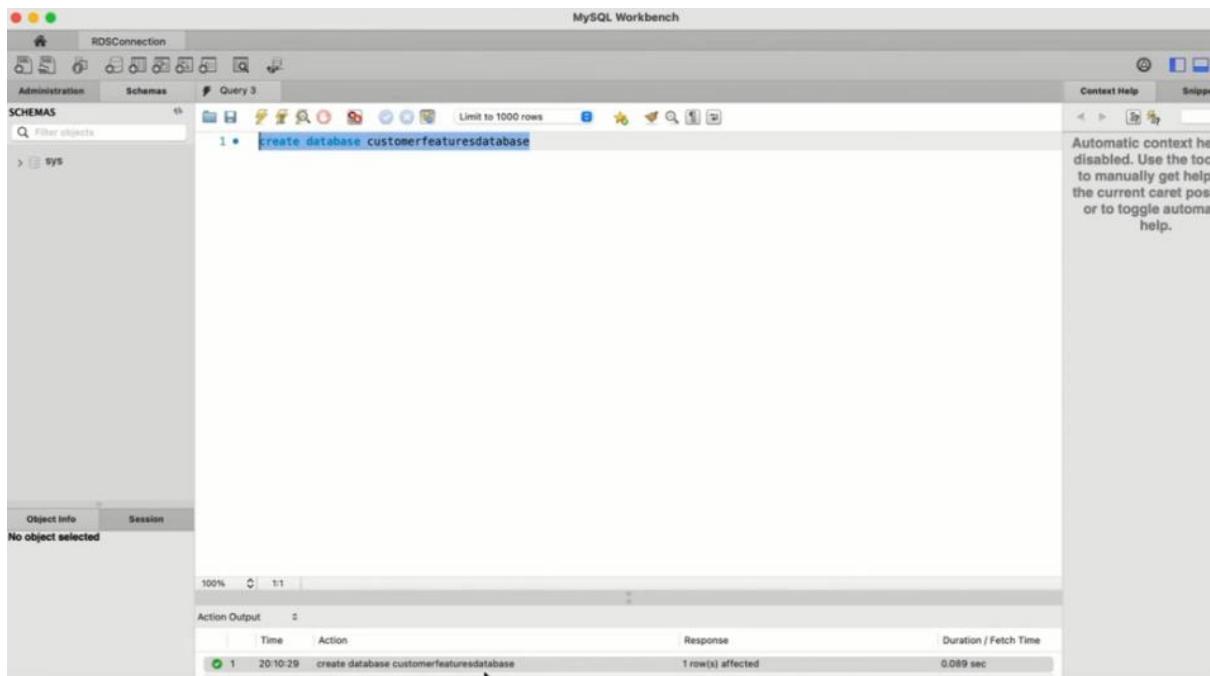
66. THE CONNECTION IS NOW ESTABLISHED, DOUBLE CLICK THE RDS CONNECTION. IT IS NOW OPENING THE SQL EDITOR.



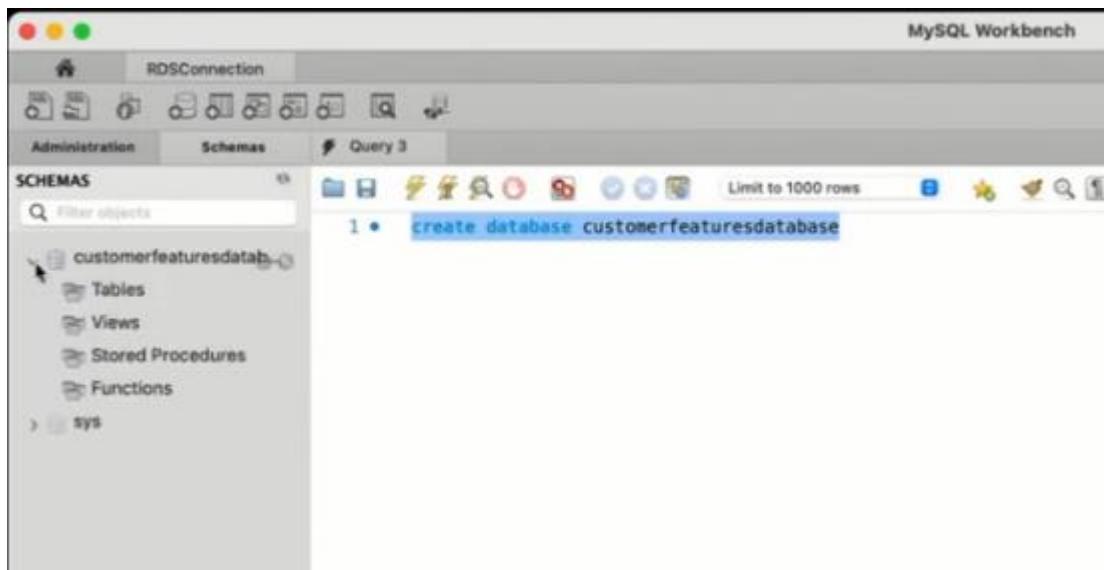
67. AND THIS IS WHERE WE ARE ABLE TO CREATE OUR TABLE WITHIN OUR RDS DATABASE INSTANCE.



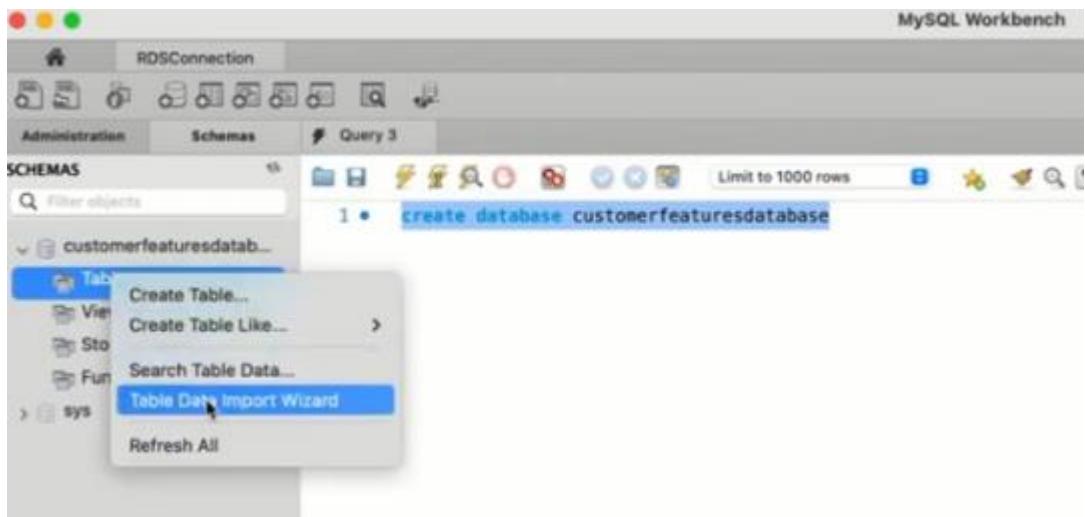
68. WE WILL NOW RUN A QUERY, CREATE DATABASE...THEN EXECUTE.



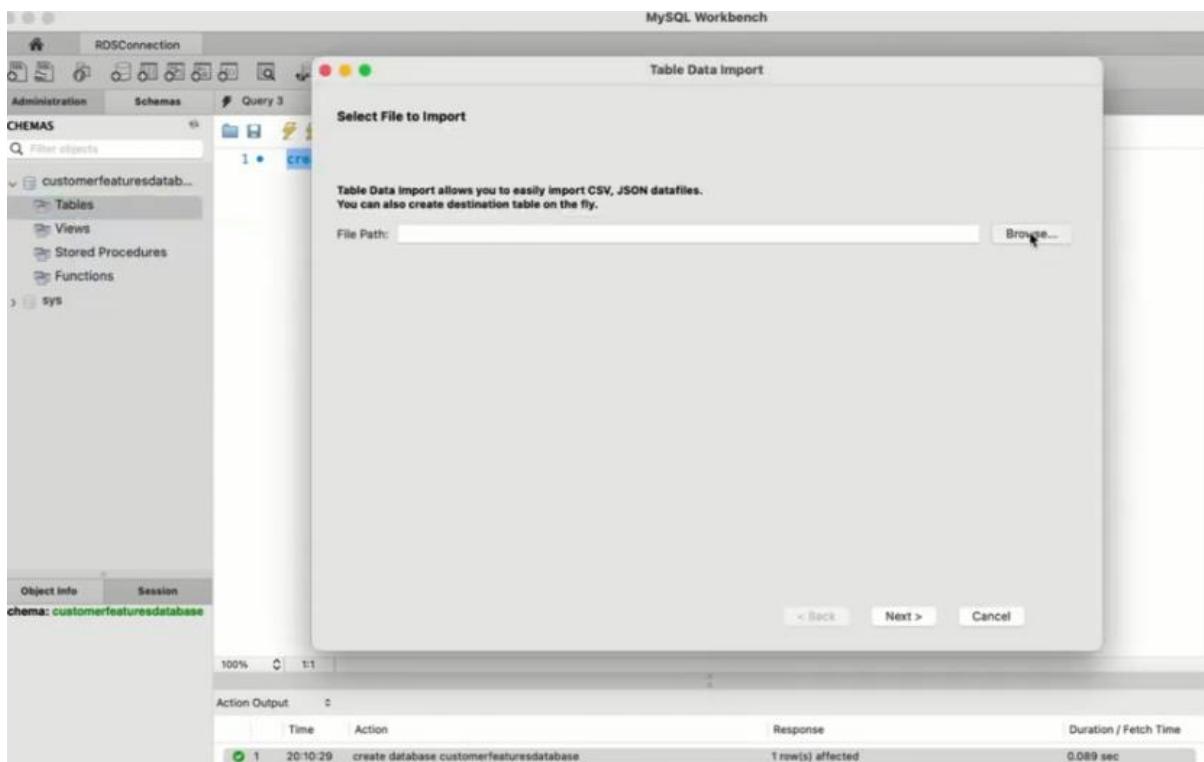
69. REFRESH THE SCHEMA TO SEE THE DATABASE.



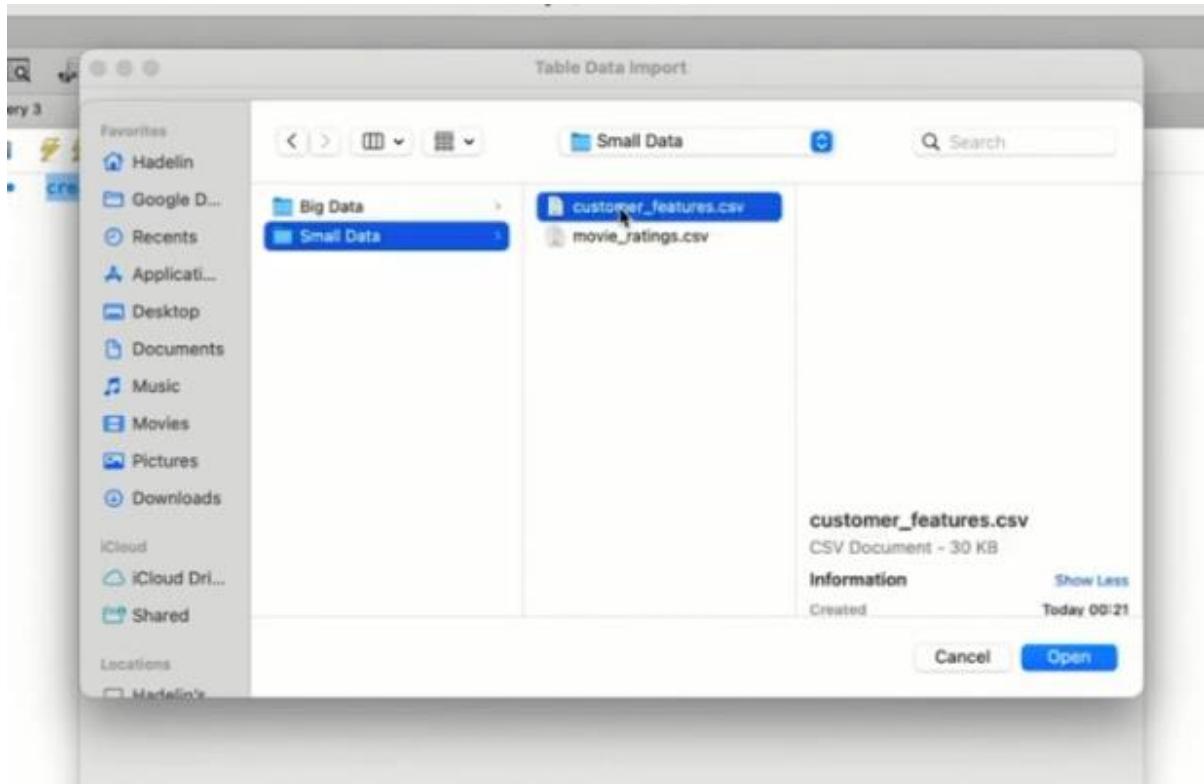
70. CLICK TABLE, CLICK RIGHT, SELECT TABLE DATA IMPORT WIZARD.



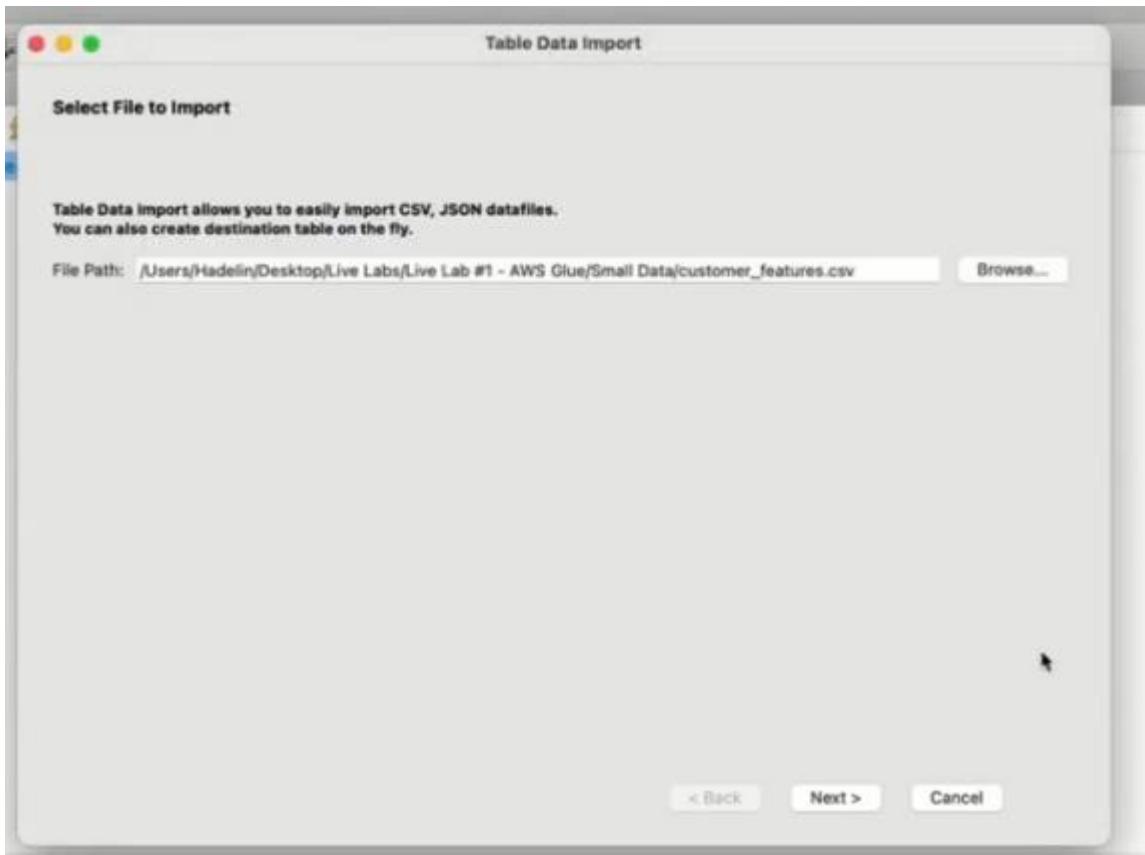
71. BROWSE THE MACHINE TO FIND THE CSV FILE.



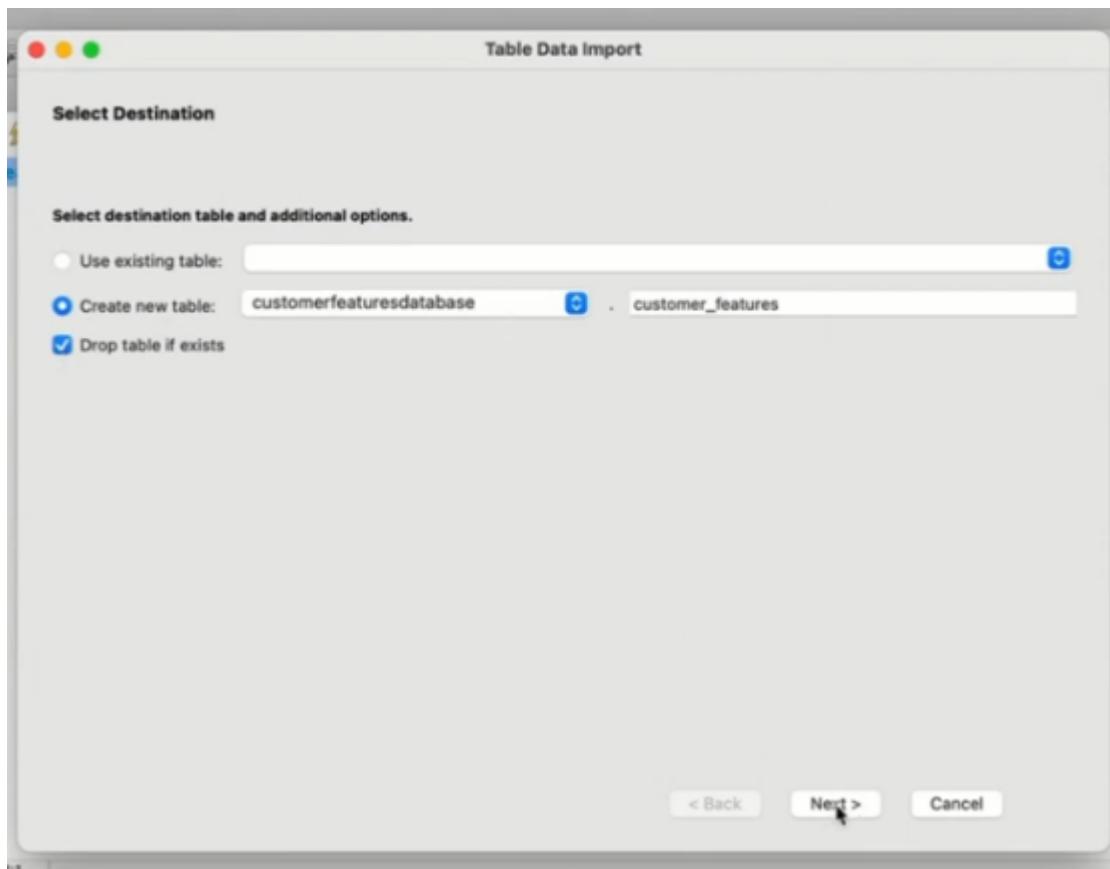
HERE, I CHOSE THE SMALL DATA, CUSTOMER FEATURES FILE.



72. CLICK NEXT.



73. YOU CAN CLICK THE DROP TABLE IF EXISTS IF YOU WANT TO. THEN, CLICK NEXT.



74. YOU CAN CHANGE THE TYPE IF YOU WANT. BUT, I JUST LEAVE IT AS IT IS. CLICK, NEXT.

The screenshot shows the 'Table Data Import' window with the title 'Configure Import Settings'. It displays detected file format as 'csv' and encoding as 'utf-8'. A table lists columns with their source column checked and field types: CustomerId (int), Surname (text), Gender (text), Age (int), and Geography (text). Below this, a preview table shows three rows of data:

CustomerId	Surname	Gender	Age	Geography	IsActiveMe...	HasPremiu...
1	Hargrave	Female	42	USA	1	1
2	Hill	Female	41	Canada	1	0
3	Onio	Female	42	USA	0	1

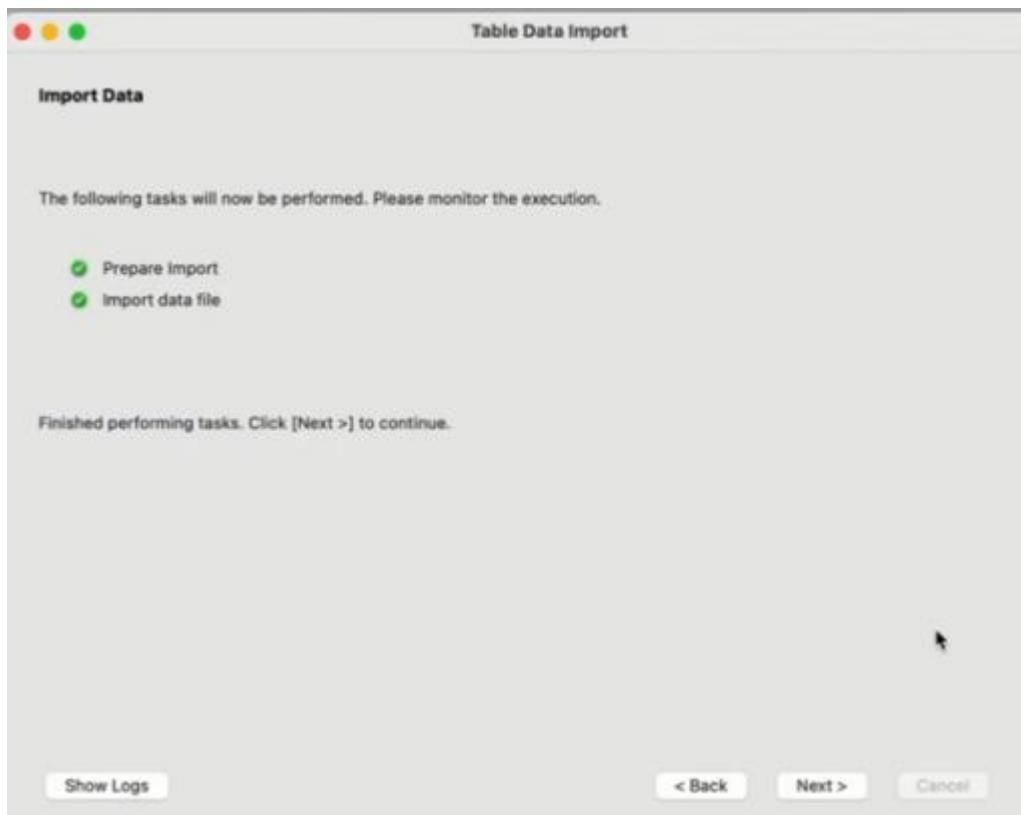
At the bottom, there are buttons for '< Back', 'Next >', and 'Cancel'.

75. IN THE IMPORT DATA, CLICK NEXT TO EXECUTE.

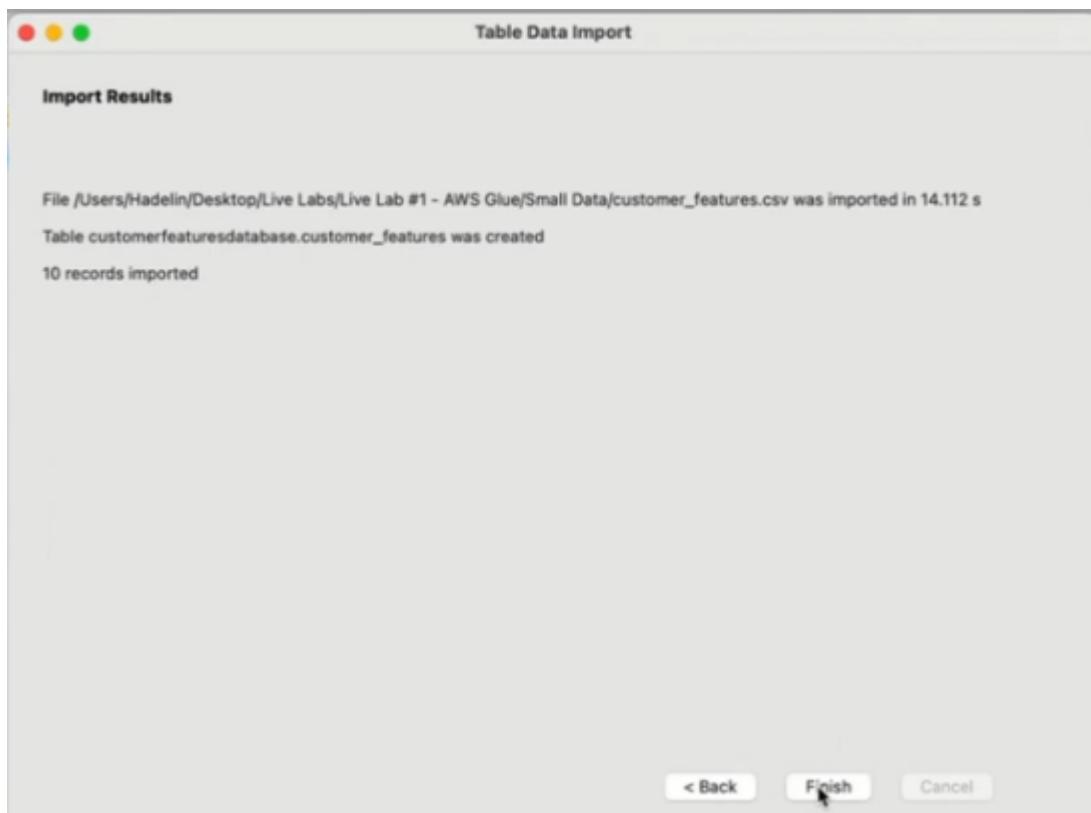


76. IT IS NOW BEGINNING THE IMPORT AND IT IS IMPORTING THE DATA TO POPULATE IT IN THE TABLE OF THE DATABASE IN THE RDS DATABASE INSTANCE THAT WE CREATED AND THAT WE CONNECTED SUCCESSFULLY TO MYSQL WORKBENCH.

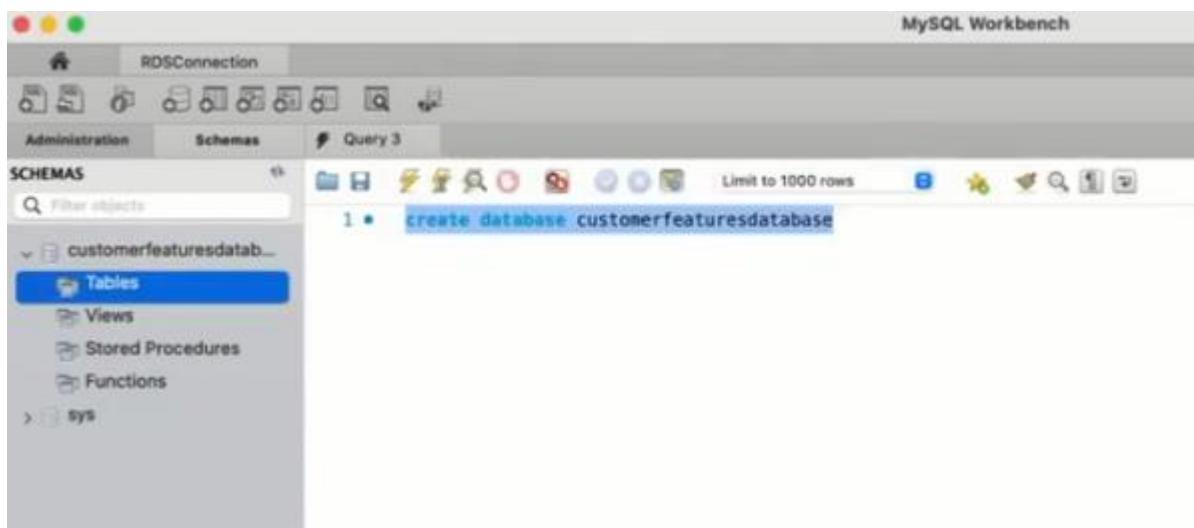
CLICK NEXT.



CLICK FINISH. THE RECORDED IS THEN FINISHED.



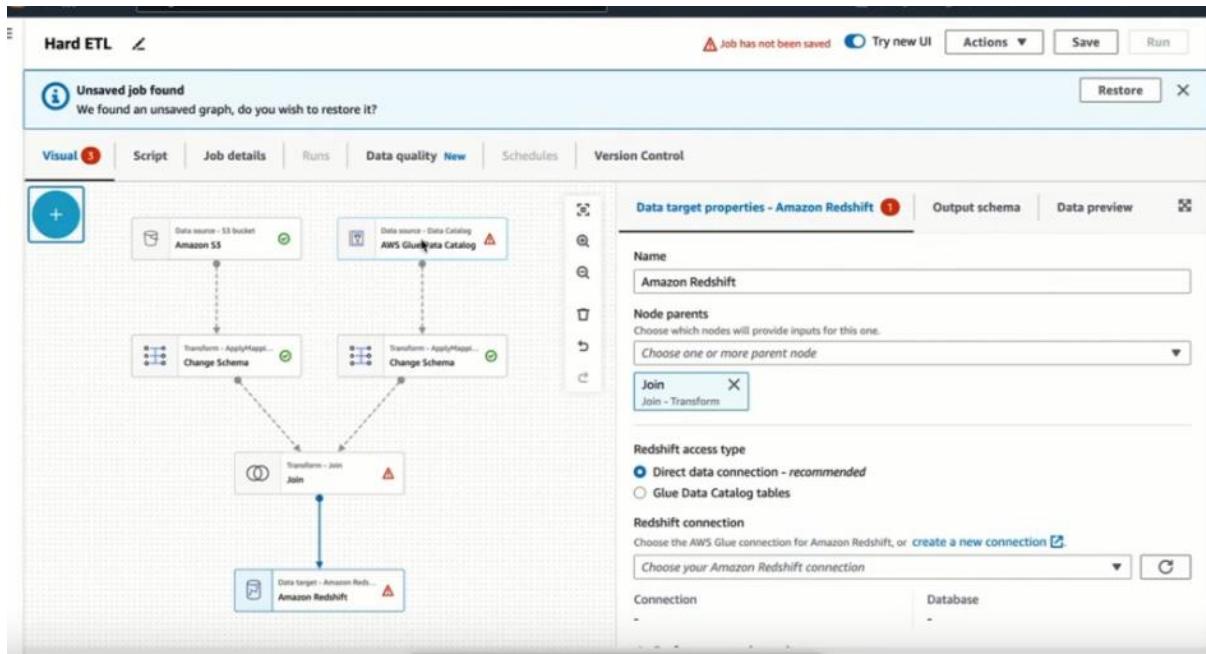
77. IN THE RDS DATABASE, WE HAVE THE DATA POPULATED.



78. GO BACK TO THE AMAZON RDS CONSOLE. OUR DATABASE INSTANCE HAS THE TABLE WELL POPULATED WITH THE DATA OF OUR CSV FILE CONTAINING THE CUSTOMER FEATURES.

A screenshot of the Amazon RDS console. The left sidebar shows "Databases" selected. The main area shows the "customer-features-rds-db-instance" database. The "Summary" tab is selected, displaying metrics like CPU usage (3.52%), Status (Available), and Engine (MySQL Community). The "Connectivity & security" tab is also visible, showing the endpoint URL and VPC security group information.

79. SO IN THE ETL PROCESS, WE HAVE DONE THE EXTRACT STEP. NOW, WE HAVE TO CONNECT OUR RDS TO THE ETL PROCESS. WE HAVE TO CONNECT THE RDS TO GLUE.



80. THE NEXT STEP THAT WE NEED TO DO IS TO CREATE AN ELEMENT FOR OUR RDS DATABASE INSTANCE, THEREFORE, AND NOW TABLE IN THE DATA CATAKOG. SO, WE ARE GOING BACK TO THE ETL, DATABASES, AND THIS TIME WE ARE GOING TO CREATE THE CUSTOMER FEATURES DATABASE.

CLICK ADD DATABASE.

Name	Description	Location URI	Created on (UTC)
movie-ratings-glue-database	-	-	June 14, 2023 at 17:28:43

81. NAME IT AS CUSTOMERR-FEATURES-GLUE-DATABASE. THEN CLICK CREATE DATABASE.

AWS Glue > Databases > Add database

Create a database

Create a database in the AWS Glue Data Catalog.

Database details

Name: customer-features-glue-database
Database name is required, in lowercase characters, and no longer than 255 characters.

Location - optional
Set the URI location for use by clients of the Data Catalog.

Description - optional
Enter text
Descriptions can be up to 2048 characters long.

Cancel **Create database**

82. THE SECOND DATABASE IS CREATED. BUT, WE NEED TO CONNECT THE RDS DATABASE THAT WE CREATED AND POPULATE IT TO MYSQL WORKBENCH TO THE GLUE DATABASE.

AWS Glue > Databases

Databases (2)

A database is a set of associated table definitions, organized into a logical group.

Name	Description	Location URI	Created on (UTC)
customer-features-glue-database	-	-	June 14, 2023 at 18:29:06
movie-ratings-glue-database	-	-	June 14, 2023 at 17:28:43

83. HERE, WE WILL USE THE CONNECTION INSTEAD OF THE CRAWLER. BECAUSE WE HAVE TO ESTABLISH THIS CONNECTION FROM RDS TO AWS GLUE.

GO TO AWS GLUE. DATA CATALOG, CONNECTIONS. CLICK CREAT CONNECTIONS.

The screenshot shows the AWS Glue Connectors page. On the left, there's a sidebar with navigation links for Getting started, ETL jobs, Notebooks, Job run monitoring, Data Catalog tables, Data connections, Workflows (orchestration), Data Catalog, Data connections, Connections, and Data Integration and ETL. The 'Data connections' link under 'Data Catalog' is highlighted. The main content area has two sections: 'Marketplace connectors' (with a 'Go to AWS Marketplace' button) and 'Custom connectors' (with a 'Create custom connector' button). Below these are two tables: one for 'Connectors (0)' and one for 'Connections (0)'. Both tables have columns for Name, Type, and Last modified, and include filters and actions buttons.

**84. ENTER A NAME FOR THE CONNECTION. WE CALL IT RDS CONNECTION.
THEN, CONNECTION TYPE IS AMAZON RDS.
IN DATABASE ENGINE, CHOOSE MYSQL.**

The screenshot shows the 'Create connection' page. The sidebar on the left is identical to the previous one. The main area is titled 'Create connection' and contains a 'Connection properties' section. It includes fields for 'Name' (set to 'RDSConnection'), 'Connection type' (set to 'Amazon RDS'), and 'Require SSL connection' (unchecked). Below this is a 'Database engine' dropdown set to 'MySQL'. There's also a 'Description - optional' field with a note about character limits. The bottom section is titled 'Connection access'.

85. IN CONNECTION ACCESS, DATABASE INSTANCES, REFRESH, THEN, SELECT YOUR RDS DATABASE INSTANCE.
IN THE DATABASE NAME, COPY THE DATABASE NAME IN THE MYSQL.
CONFIDENTIAL TYPE, USE THE USERNAME THAT WAS SET EARLIER WHICH IS THE ADMIN.
INPUT THE PASSWORD. THEN CLICK CREATE CONNECTION.

Connection access

Database instances
Provisioned Amazon Relational Database Service instances.
customer-features-rds-db-instance

Database name
customerfeaturesdatabase

Credential type
 Username and password
 Secret

Username
admin

Password

86. RDS CONNECTION IS NOW CREATED

AWS Glue

Getting started
ETL jobs
 Visual ETL
 Notebooks
 Job run monitoring
 Data Catalog tables
Data connections
 Workflows (orchestration)
Data Catalog
 Databases
 Tables
 Stream schema registries
 Schemas
Connections
 Crawlers
 Classifiers
 Catalog settings
Data Integration and ETL
 ETL jobs
 Visual ETL
 Notebooks
 Job run monitoring
 Interactive Sessions
 Data classification tools

Connectors info

Marketplace connectors
Subscribe to connectors from AWS partners to expand your data sources.

Custom connectors
Provide your own connector to expand your data sources. [Creating custom connectors](#)

Connectors (0) info
You can manage your connectors or use them to create connections.

 Filter connections by property < 1 >

Name	Type	Last modified
------	------	---------------

Connections (1) info
You can manage your connections or use a connection in a job.

 Filter connections by property < 1 >

Name	Type	Last modified
RDSConnection	JDBC	Jun 14, 2023

87. TEST THE CONNECTION. SELECT THE CONNECTION. CLICK ACTIONS. CLICK TEST CONNECTION.

The screenshot shows the 'Connections (1)' page in the AWS Glue console. There is one connection named 'by property'. The 'Type' is listed as 'JDBC' and the 'Last modified' date is 'Jun 14, 2023'. The 'Test connection' button is highlighted with a red box.

88. SELECT THE IAM ROLE – GLUEFULLACCESSROLE. SELECT CONFIRM.

The screenshot shows the 'Test Connection' dialog box. The 'IAM role' dropdown is set to 'GlueFullAccessRole'. The 'Confirm' button is highlighted with a red box.

89. HERE, WE WILL GET THE FAILED TEST CONNECTION. TO SOLVE THE ISSUE, WE HAVE TO FOLLOW WHAT IS SUGGESTED. BASED ON THE TROUBLESHOOT RECOMMENDATION, AT LEAST ONE SECURITY GROUP MUST OPEN ALL INGRESS PORTS AND TO LIMIT TRAFFIC, THE SOURCE SECURITY GROUP IN YOUR INBOUND RULE CAN BE RESTRICTED TO THE SAME SECURITY GROUP. SO, BASICALLY, IT IS REFERRING TO OUR RDS DATABASE INSTANCE,

The screenshot shows the 'Test Connection' dialog box with an error message: 'InvalidInputException: At least one security group must open all ingress ports. To limit traffic, the source security group in your inbound rule can be restricted to the same security group'. The 'Cancel' button is highlighted with a red box.

90. GO BACK TO THE RDS DATABASE INSTANCE. IN THE SECURITY FIELD, YOU WILL SEE THOSE TWO (2) VPC SECURITY GROUPS, THE ONE THAT WE`VE CREATED, SG-OPEN-MYSQL, AND THE DEFAULT ONE. BASED ON THE PREVIOUS TROUBLESHOOT SUGGESTIONS, WE NEED TO OPEN AT LEAST ONE SECURITY GROUP OF ALL INGRESS PORTS. SO, INSTEAD OF OPENING ONLY THE 3306 PORT, WE ARE GOING TO OPEN ALL PORTS IN ONE OF THE TWO SECURITY GROUPS.

91.

Endpoint & port	Networking	Security
Endpoint customer-features-rds-db-instance.crrwucvfcmbc.us-east-1.rds.amazonaws.com	Availability Zone us-east-1c	VPC security groups SG-Open-MySQL (sg-05eae22adcf0dc56) Active
Port 3306	VPC vpc-0041e736a2d14cfa2	default (sg-00757adc39bf9625e) Active
	Subnet group default-vpc-0041e736a2d14cfa2	Publicly accessible

92. GO BACK TO THE EC2 INSTANCE, SECURITY GROUPS. SELECT SG-OPEN-MYSQL. GO TO INBOUND RULES. THERE, WE ARE GOING TO ADD THE NEW RULE WHICH OPENS ALL PORTS.

Name	Security group ID	Security group name	VPC ID	Description	Owner
-	sg-04c7ce4cb6a0e059f	launch-wizard-1	vpc-0041e736a2d14cfa2	launch-wizard-1 create...	7496011
-	sg-0ce2e3b4e3be04e92	SG-Open-HTTP	vpc-0041e736a2d14cfa2	Allows HTTP Traffic	7496011
-	sg-00757adc39bf9625e	default	vpc-0041e736a2d14cfa2	default VPC security gr...	7496011
<input checked="" type="checkbox"/>	sg-05eae22adcf0dc56	SG-Open-MySQL	vpc-0041e736a2d14cfa2	Allows MySQL Access t...	7496011

93. IN THE INBOUND RULES, CLICK EDIT INBOUND RULES. THEN ADD RULE.

The screenshot shows the AWS EC2 Security Groups Inbound rules page. At the top, there is a message: "You can now check network connectivity with Reachability Analyzer" with a "Run Reachability Analyzer" button. Below this, the title "Inbound rules (1/1)" is displayed, along with a search bar and filter options. A table lists one rule: "sgr-000a62b4b15cc1b15" (Security group rule ID), "MySQL/Aurora" (Type), "TCP" (Protocol), "3306" (Port range), "Custom" (Source), and "Description - optional" (empty). There are "Edit inbound rules" and "Delete" buttons at the bottom of the table. The URL in the browser is "EC2 > Security Groups > sg-05eae22adcf0dcf56 - SG-Open-MySQL > Edit inbound rules".

94. IN THE ADD RULE, CLICK ALL TCP IN THE TYPE TO OPEN ALL PORTS FROM 0 – 65535. IN THE SOURCE, SELECT ANYWHERE-IPV4. CLICK SAVE RULES.

The screenshot shows the "Edit inbound rules" page with a new rule being added. The table has two rows: the first row is for the existing rule "sgr-000a62b4b15cc1b15" with source "Custom" and destination "0.0.0.0/0"; the second row is the new rule being added, with "All TCP" selected in the Type dropdown, "TCP" in the Protocol dropdown, "0 - 65535" in the Port range dropdown, and "Anywh..." in the Source dropdown. The "Description - optional" field is empty. The "Save rules" button is highlighted in orange at the bottom right. The URL in the browser is "EC2 > Security Groups > sg-05eae22adcf0dcf56 - SG-Open-MySQL > Edit inbound rules".

95. WE ARE NOW GOING TO SEE IF IT FIXES THE ISSUE. HERE, THE SECURITY GROUP IS MODIFIED.

The screenshot shows the AWS EC2 Security Groups interface. A success message at the top indicates that inbound security group rules were successfully modified on security group sg-05eae22adcf0dce56 (SG-Open-MySQL). The main table lists four security groups:

Name	Security group ID	Security group name	VPC ID	Description	Owner
sg-0ce2e3b4e3be04e92	SG-Open-HTTP	vpc-0041e736a2d14cfa2		Allows HTTP Traffic	7496011
sg-05eae22adcf0dce56	SG-Open-MySQL	vpc-0041e736a2d14cfa2		Allows MySQL Access t...	7496011
sg-04c7ce4cb6a0e059f	launch-wizard-1	vpc-0041e736a2d14cfa2		launch-wizard-1 create...	7496011
sg-00757adc39bf9625e	default	vpc-0041e736a2d14cfa2		default VPC security gr...	7496011

The selected security group is sg-05eae22adcf0dce56 - SG-Open-MySQL. The Inbound rules tab is active, showing two rules. A note says "You can now check network connectivity with Reachability Analyzer".

96. GO BACK TO CONNECTORS TO TEST THE CONNECTION AGAIN. IN THE CONNECTIONS, CLICK RDS CONNECTION, CLICK ACTIONS, CHOOSE TEST CONNECTION.

The screenshot shows the AWS Glue Connectors interface. The left sidebar includes sections for Getting started, ETL jobs, Data Catalog tables, Data connections, Workflows (orchestration), Data Catalog, Tables, Stream schema registries, Schemas, Connections, Crawlers, Classifiers, Catalog settings, Data Integration and ETL, ETL jobs, Visual ETL, Notebooks, Job run monitoring, Interactive Sessions, Data classification tools, and Sensitive data detection. The main area displays Marketplace connectors and Custom connectors. Under Connections, there is a table for managing connections:

Name	Type	Last modified
RDSConnection	JDBC	Jun 14, 2023

SELECT TEST CONNECTION.

The screenshot shows the 'Connections' list interface. At the top, there are buttons for 'Actions' (with a dropdown arrow), 'Create connection', and 'Create job'. Below this is a search bar with the placeholder 'View details by property' and a page navigation bar with a left arrow, a page number '1', and a right arrow. A table lists one connection:

	Type	Last modified
by property	JDBC	Jun 14, 2023

CHOOSE GLUEFULLACCESSROLE. THEN, CONFIRM.

The screenshot shows the 'Test Connection' dialog box. At the top, there is a 'Go to AWS Marketplace' link and a 'Create custom connector' button. Below this is a section titled 'IAM role' with the sub-instruction 'To test your connection, choose an IAM role'. A dropdown menu is open, showing 'Choose an option' at the top, followed by a search bar with a magnifying glass icon, and then 'GlueFullAccessRole' which is highlighted. A tooltip for 'GlueFullAccessRole' states: 'Allows Glue to call AWS services on your behalf.' At the bottom of the dialog are 'Cancel' and 'Confirm' buttons, with 'Confirm' being highlighted. The background shows the same 'Connections' list interface as the previous screenshot.

UNFORTUNATELY, WE HAVE ANOTHER MESSAGE ERROR. BASED ON THE TROUBLESHOOT MESSAGE, WE NEED TO CREATE A VPC S3 ENDPOINT.

The screenshot shows the 'Test Connection' dialog box again. This time, an error message is displayed in a red-bordered box: 'InvalidInputException: VPC S3 endpoint validation failed for SubnetId: subnet-0536055a78fadcc9a. VPC: vpc-0041e736a2d14cfa2. Reason: Could not find S3 endpoint or NAT gateway for subnetId: subnet-0536055a78fadcc9a in Vpc vpc-0041e736a2d14cfa2'. To the right of the message is a 'Troubleshoot' button with a magnifying glass icon. At the bottom of the dialog are 'Cancel' and 'Create job' buttons, with 'Create job' being highlighted. The background shows the same 'Connections' list interface as the previous screenshots.

97. GO BACK TO THE SEARCH BAR, TYPE VPC. THEN, RIGHT CLICK, OPEN LINK IN NEW TAB.

The screenshot shows the AWS Glue search interface with the query 'VPC' entered in the search bar. The results are categorized under 'Services'. A context menu is open over the 'Services' section, with the 'Open Link in New Tab' option highlighted. Other options in the menu include 'Open Link in New Window', 'Open Link in Incognito Window', 'Save Link As...', 'Copy Link Address', 'Copy', 'Copy Link to Highlight', 'Search Google for "VPC"', 'Print...', 'Translate Selection to English', 'Inspect', 'Speech', and 'Services'.

Search results for 'VPC'

Try searching with longer queries for more relevant results

See all 12 results ▶

Services (12)

Features (48)

Resources **New**

Blogs (740)

Documentation (11,232)

Knowledge Articles (30)

Tutorials (10)

Events (16)

Marketplace (452)

Top features

Your VPC

Isolate traffic between VPCs

Internet gateways

Egress-only internet gateways

IT operations management for AWS

See all 48 results ▶

Dashboard

VPC feature

VPC Reachability Analyzer

98. IN THE VCPC, GO TO THE ENDPOINTS. IT IS NORMAL NOT TO SEE ANY ENDPOINTS. BUT AS A SOLUTION TO THE CONNECTION, WE ARE GOING TO CREATE AN ENDPOINT.

The screenshot shows the AWS VPC Endpoints page. The left sidebar navigation includes 'VPC dashboard', 'EC2 Global View', 'Filter by VPC', 'Virtual private cloud' (with 'Your VPCs' selected), 'Endpoints' (selected), 'Endpoint services', 'NAT gateways', 'Peering connections', and 'Security'. The main content area is titled 'Endpoints' and shows a table with columns: Name, VPC endpoint ID, VPC ID, and Service name. A message at the top right says 'No endpoint found'. Below the table, a section titled 'Select an endpoint' is visible.

VPC dashboard

EC2 Global View **New**

Filter by VPC:

Select a VPC

Virtual private cloud

Your VPCs **New**

Subnets

Route tables

Internet gateways

Egress-only internet gateways

Carrier gateways

DHCP option sets

Elastic IPs

Managed prefix lists

Endpoints **New**

Endpoint services

NAT gateways

Peering connections

Security

Network ACLs

Security groups

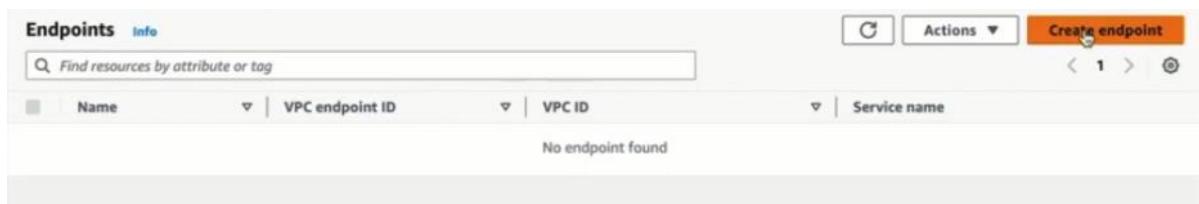
Endpoints **Info**

Find resources by attribute or tag

Name	VPC endpoint ID	VPC ID	Service name
No endpoint found			

Select an endpoint

99. CLICK CREATE ENDPOINTS



100. THE ENDPOINT SHOULD BE THE VPC ENDPOINT TO THE S3. SO, WE ARE GOING TO CALL IT VPC ENDPOINT TO S3.

A screenshot of the 'Create endpoint' wizard. The top navigation shows 'VPC > Endpoints > Create endpoint'. The main title is 'Create endpoint' with a blue 'Info' link. Below it is a descriptive text about VPC endpoint types. The first step, 'Endpoint settings', is active. It has a 'Name tag - optional' section with a text input containing 'VPC Endpoint to S3'. The 'Service category' section contains several options: 'AWS services' (selected, highlighted in blue), 'PrivateLink Ready partner services', 'AWS Marketplace services', 'EC2 Instance Connect Endpoint' (disabled, greyed out), and 'Other endpoint services'. Each option has a brief description below it.

101. IN THE SERVICE CATERGORY, SELECT AWS SERVICES.

Endpoint settings

Name tag - *optional*
Creates a tag with a key of 'Name' and a value that you specify.

VPC Endpoint to S3

Service category
Select the service category

<input checked="" type="radio"/> AWS services Services provided by Amazon	<input type="radio"/> PrivateLink Ready partner services Services with an AWS Service Ready designation	<input type="radio"/> AWS Marketplace services Services that you've purchased through AWS Marketplace
<input type="radio"/> EC2 Instance Connect Endpoint An elastic network interface that allow you to connect to resources in a private subnet	<input type="radio"/> Other endpoint services Find services shared with you by service name	

102. IN THE SERVICES, SELECT SERVICE NAME = COM.AMAZONAWS.US-EAST-1.S3

Services (225)		
<input type="text" value="S3"/> X C 1 2 3 4 5 6 7 ... 23 > ⑧		
Use: "S3"	Owner	Type
Client filters values		
Service Name = com.amazonaws.s3-global.accesspoint	amazon	Interface
Service Name = com.amazonaws.us-east-1.s3	amazon	Interface
Service Name = com.amazonaws.us-east-1.s3-outposts	amazon	Interface
com.amazonaws.us-east-1.acm-pca	amazon	Interface
com.amazonaws.us-east-1.airflow.api	amazon	Interface

103. SELECT GATEWAY

Services (1/2)		
<input type="text"/> Find resources by attribute or tag		
Service Name = com.amazonaws.us-east-1.s3 <input type="button" value="X"/>		<input type="button" value="Clear filters"/>
Service Name	Owner	Type
com.amazonaws.us-east-1.s3	amazon	Gateway
com.amazonaws.us-east-1.s3	amazon	Interface

104. SELECT THE VPC. REMEMBER TO CONNECT IT TO THE SUBNET.

VPC

vpc-0041e736a2d14cfa2 172.31.0.0/16	(default)
Select a VPC	<input type="button" value="▲"/>

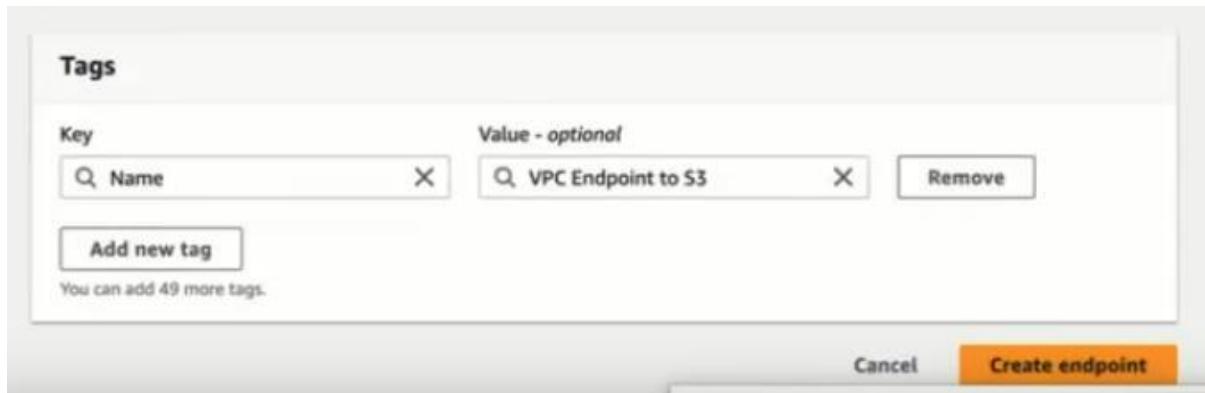
105. WE WILL CONNECT THE SUBNET THROUGH THE ROUTE TABLES BECAUSE THAT IS WHAT CONTAINS THE CONNECTION TO THE SUBNETS.

Route tables (1/1)

Name	Route Table ID	Main
-	rtb-05636a379b939ac57	Yes

Info: When you use an endpoint, the source IP addresses from your instances in your affected subnets for accessing the AWS service in the same region will be private IP addresses, not public IP addresses. Existing connections from your affected subnets to the AWS service that use public IP addresses may be dropped. Ensure that you don't have critical tasks running when you create or modify an endpoint.

106. CLICK CREATE ENDPOINT.



107. GO CHECK THE ENDPOINTS. HERE, WE SHOULD SEE THE ONE THAT WE HAVE CREATED.

The screenshot shows the AWS VPC Endpoints list. A success message at the top says 'Successfully created VPC endpoint vpce-0dfd95341ef6181af'. The main table lists one endpoint:

Name	VPC endpoint ID	VPC ID	Service name
VPC Endpoint to S3	vpce-0dfd95341ef6181af	vpc-0041e736a2d14cfa2	com.amazonaws.us-east-1.s3

Below the table is a detailed view for the endpoint 'vpce-0dfd95341ef6181af / VPC Endpoint to S3' with tabs for Details, Route tables, Policy, and Tags.

108. CHECK IF IT FIXES THE ISSUE. GO BACK TO THE CONNECTORS

AWS Glue

Getting started
ETL jobs
Visual ETL
Notebooks
Job run monitoring
Data Catalog tables
Data connections
Workflows (orchestration)

Data Catalog
Databases
Tables
Stream schema registries
Schemas
Connections
Crawlers
Classifiers
Catalog settings

Data Integration and ETL
ETL jobs
Visual ETL
Notebooks
Job run monitoring
Interactive Sessions

AWS Glue > Connectors

Connectors Info

Marketplace connectors
Subscribe to connectors from AWS partners to expand your data sources.
[Go to AWS Marketplace](#)

Custom connectors
Provide your own connector to expand your data sources. [Creating custom connectors](#)

[Create custom connector](#)

Connectors (0) Info
You can manage your connectors or use them to create connections.

[Actions](#)

Filter connections by property

Name	Type	Last modified

Connections (1) Info
You can manage your connections or use a connection in a job.

[Actions](#) [Create connection](#) [Create job](#)

Filter connections by property

Name	Type	Last modified
RDSConnection	JDBC	Jun 14, 2023

109. CLICK RDS CONNECTION, ACTION THE TEST CONNECTION.

Connections (1) Info
You can manage your connections or use a connection in a job.

[Actions](#) [Create connection](#) [Create job](#)

Filter connections by property

Name	Type	Last modified
RDSConnection	JDBC	Jun 14, 2023

Connections (1) Info
You can manage your connections or use a connection in a job.

[Actions](#) [Create connection](#) [Create job](#)

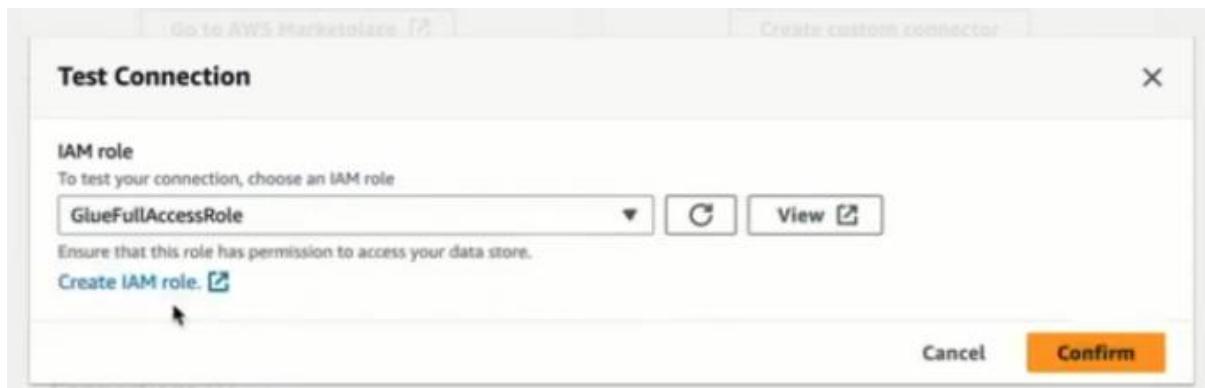
[View details](#) by property

[Delete](#) [Edit](#) [Test connection](#)

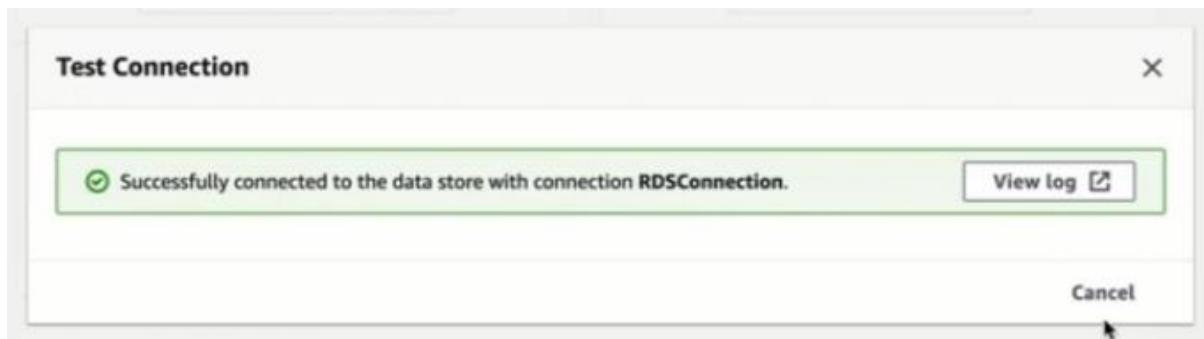
Filter connections by property

	Type	Last modified

110. SELECT GLUEFULLACCESS, THEN, CONFIRM.



111. IT SHOULD BE SUCCESSFULLY CONNECTED BY NOW. CONGRATULATIONS YOU DID THE HARD PART! WE CAN NOW PROCEED TO THE NEXT PROCEDURE WHICH IS TO HANDLE THE REDSHIFT.



112. IN THE AWS GLUE, SELECT DATA CATALOG, THEN, DATABASES. IN THE DATABASES, SELECT CUSTOMER-FEATURES-GLUE-DATABASE.

The screenshot shows the AWS Glue Data Catalog interface. On the left, there is a navigation sidebar with options like 'Getting started', 'ETL jobs', 'Visual ETL', 'Notebooks', 'Job run monitoring', 'Data Catalog tables', 'Data connections', 'Workflows (orchestration)', 'Data Catalog', 'Databases', 'Tables', 'Stream schema registries', 'Schemas', 'Connections', 'Crawlers', 'Classifiers', 'Catalog settings', and 'Data Integration and ETL'. The 'Databases' section is currently selected. The main area displays a table titled 'Databases (2)'. The table has columns: Name, Description, Location URI, and Created on (UTC). The first row, 'customer-features-glue-database', is highlighted in blue, indicating it is selected. The second row, 'movie-ratings-glue-database', is shown below it. The table includes a header bar with 'Last updated (UTC)' (June 14, 2023 at 18:51:27), 'Edit', 'Delete', and 'Add database' buttons, along with navigation arrows and a refresh icon.

113. IN THE CUSTOMER-FEATURES-GLUE-DATABASE, CLICK, ADD TABLES USING CRAWLER.

AWS Glue > Databases > customer-features-glue-database

customer-features-glue-database

Last updated (UTC) June 14, 2023 at 18:51:30 [Edit](#) [Delete](#)

Database properties

Name customer-features-glue-database	Description -	Location -	Created on (UTC) June 14, 2023 at 18:29:06
---	------------------	---------------	---

Tables (0) Last updated (UTC) June 14, 2023 at 18:51:31 [Edit](#) [Delete](#) [Data quality](#) [New](#) [Add tables using crawler](#) [Add table](#)

Filter tables

Name	Database	Location	Classification	Deprecated	View data
No available tables					

114. ENTER UNIQUE CRAWLER NAME. CLICK, NEXT.

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

Set crawler properties

Crawler details [Info](#)

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

Description - optional Descriptions can be up to 2048 characters long.

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

[Cancel](#) [Next](#)

115. CLICK NOT YET IN THE “IS YOUR DATA ALREADY MAPPED TO GLUE TABLES?”. THEN, ADD A DATA SOURCE.

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 (0) Info The list of data sources to be scanned by the crawler.

Type	Data source	Parameters
You don't have any data sources.		
Add a data source		

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.

[Cancel](#) [Previous](#) [Next](#)

116. IN THE DATA SOURCE, SELECT JDBC WHICH MEANS JAVA DATABASE CONNECTION.

Add data source

Data source Choose the source of data to be crawled.

- S3 JSON, CSV, or Parquet files stored in S3.
- JDBC JDBC stream as the data source
- DynamoDB Dynamo DB as the data source
- DocumentDB/MongoDB Mongo DB as the data source
- Delta Lake Delta Lake as the data source
- In a different account

S3 path Browse for or enter an existing S3 path.

s3://bucket/prefix/object [View](#) [Browse S3](#)

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.

117. IN THE CONNECTION, SELECT RDSCONNECTION THAT WE HAVE JUST CREATED AND TESTED SUCCESSFULLY.



118. IN THE INCLUDE PATH, WE HAVE TO MAKE SURE THAT WE INCLUDE THE RIGHT PATH. IT IS GOOD TO READ AND TAKE INTO CONSIDERATION WHAT IS WRITTEN BELOW.

Include path

Type a database/schema/table or database/table...

⚠ This is a required field.

You can substitute the percent (%) character for a schema or table. For databases that support schemas, enter MyDatabase/MySchema/% to match all tables in MySchema within MyDatabase. Oracle Database and MySQL don't support schema in the path; instead, enter MyDatabase/%. For Oracle database without SSL, MyDatabase can be either the system identifier (SID) or the service name (SERVICE_NAME). For Oracle database with SSL, MyDatabase must be the service name (SERVICE_NAME).

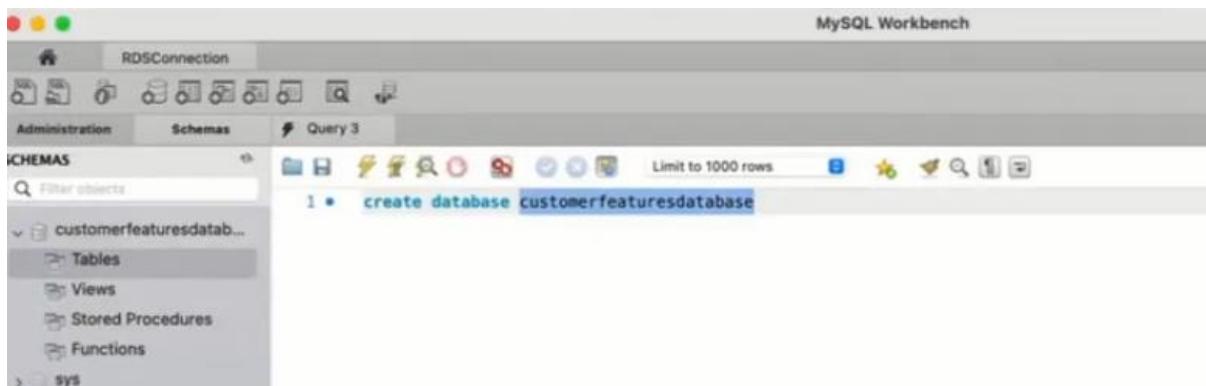
Additional metadata - optional

Select additional metadata properties for the crawler to crawl.

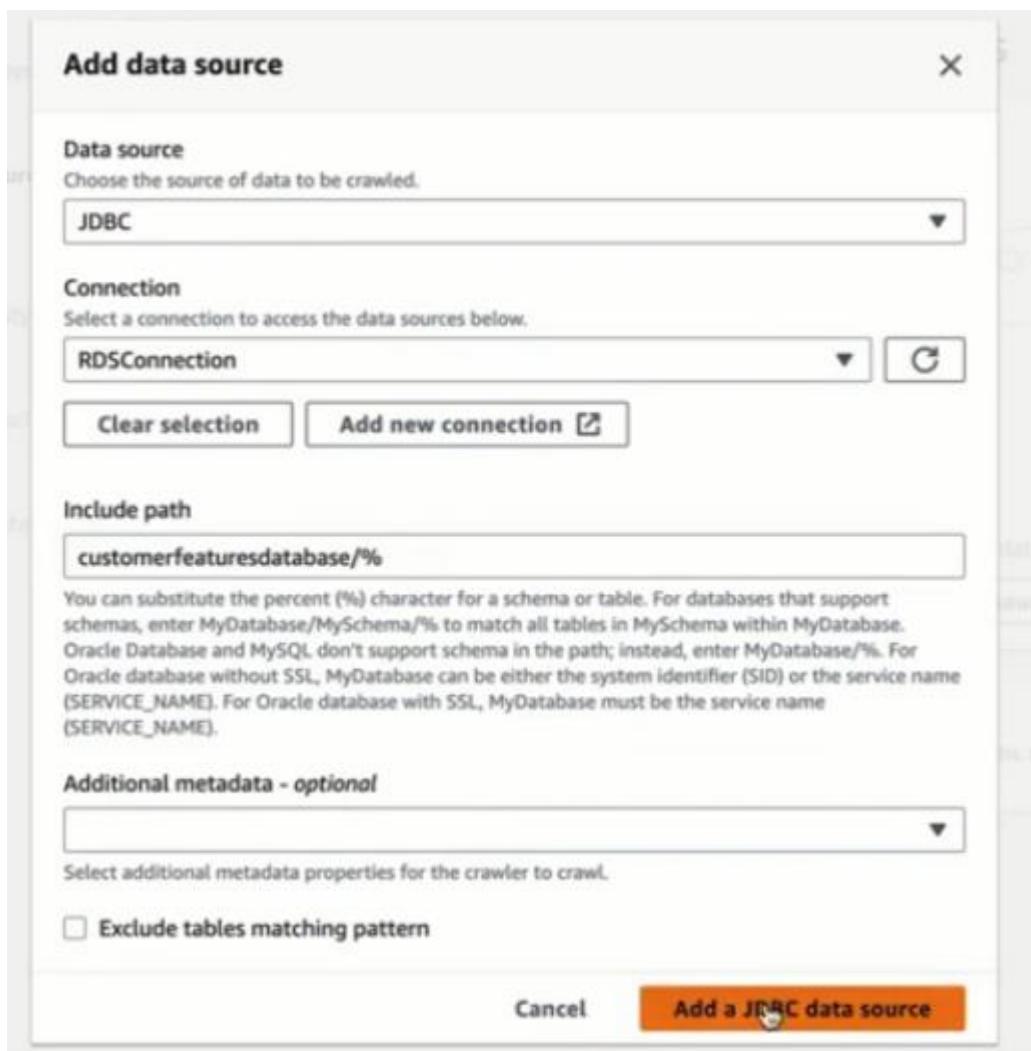
Exclude tables matching pattern

Cancel Add a JDBC data source

119. GO BACK TO THE MYSQL WORKBENCH TO DOUBLE CHECK THE DATABASE NAME THAT WE HAVE JUST CREATED PREVIOUSLY.



120. FOLLOW THE DATABASE NAME IN THE MYSQL AND THE INSTRUCTION GIVEN IN PUTTING THE NAME IN THE 'INCLUDE PATH'. FINALLY, CLICK ADD A JDBC DATA SOURCE.



121. SELECT JDBC. THEN, CLICK NEXT.

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 (1) Info

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

Type	Data source	Parameters
JDBC	customerfeaturesdatabase/%	-

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.

Cancel Previous Next

122. IN THE CONFIGURE SECURITY SETTINGS, EXISTING IAM ROLE, SELECT GLUEFULLACCESSROLE. THEN, CLICK NEXT.

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

Choose an IAM role	View
<input type="text"/> Q	<input type="button"/>
AWSReservedSSO_AWSAdministratorAccess_d6359d25c2eb881 8	serviceRole-* can be updated.
AWSReservedSSO_AWSPowerUserAccess_72a7e95974f3825f Provides full access to AWS services and resources, but does not allow management of Users and groups	
AWSReservedSSO_AWSReadOnlyAccess_1ec8cf602ddff80ec This policy grants permissions to view resources and basic metadata across all AWS services	
AWSReservedSSO_AWSServiceCatalogAdminFullAccess_d1fe8e1088162bc9 Provides full access to AWS Service Catalog admin capabilities	
AWSReservedSSO_AWSServiceCatalogEndUserAccess_7c60b01db8043761 Provides access to the AWS Service Catalog end user console	
GlueFullAccessRole Allows Glue to call AWS services on your behalf.	

Cancel Previous Next

Configure security settings

IAM role [Info](#)

Existing IAM role

GlueFullAccessRole



[View](#)

[Create new IAM role](#)

[Update chosen IAM 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.

[Cancel](#)

[Previous](#)

[Next](#)

123. IN THE TARGET DATABASE, SELECT CUSTOMER-FEATURES-GLEU-DATABASE. KEEP THE ON-DEMAND FREQUENCY OF THE CRAWLER SCHEDULE. IT MEANS THAT IT IS GOING TO CRAWL WHEN WE CLICK THE RUN BUTTON. THEN, CLICK NEXT.

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

Set output and scheduling

Output configuration [Info](#)

Target database

Choose a database

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

[Cancel](#) [Previous](#) [Next](#)

community.cloudwolf.com is sharing your screen. [Stop sharing](#) [Hide](#)

Target database

▼

Clear selection
Add database

Table name prefix - optional

Type a prefix added to table names

124. REVIEW, THEN, CLICK 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

Review and create

Step 1: Set crawler properties

Set crawler properties		
Name customer-features-crawler	Description	Tags

Step 2: Choose data sources and classifiers

Data sources (1) <small>Info</small>		
The list of data sources to be scanned by the crawler.		
Type	Data source	Parameters
JDBC	customerfeaturesdatabase/%	-

Step 3: Configure security settings

Configure security settings		
IAM role GlueFullAccessRole	Security configuration -	Lake Formation configuration -

Step 4: Set output and scheduling

Set output and scheduling		
Database customer-features-glue-database	Table prefix - optional -	Schedule On demand

Cancel
Previous
Create crawler

125. CRAWLER SUCCESSFULLY CREATED, BUT, WE NEED TO RUN IT. ON THE UPPER RIGHT SIDE, CLICK RUN CRAWLER.

Crawler properties

Name customer-features-crawler	IAM role GlueFullAccessRole	Database customer-features-glue-database	State READY
Description	Security configuration	Table prefix	

Crawler runs (0)
The list of crawler runs for this crawler.

Crawler runs (0)
Start time (UTC) ▲ End time (UTC) ▼ Current/last duration ▼ Status ▼ DPU hours ▼ Table changes ▼

You don't have any crawler runs.
Run crawler

126. OR YOU CAN CLICK THE CRAWLERS, REFRESH TO SEE THE 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 (1) Info
View and manage all available crawlers.

Name	State	Schedule	Last run	Last run time...	Log	Table changes ...
movie-ratings-cr...	Ready		Succeeded	June 14, 2023 a...	View log	1 created

127. SELECT THE NEW CRAWLER, CUSTOMER-FEATURES-CRAWLER. THEN,CLICK RUN.

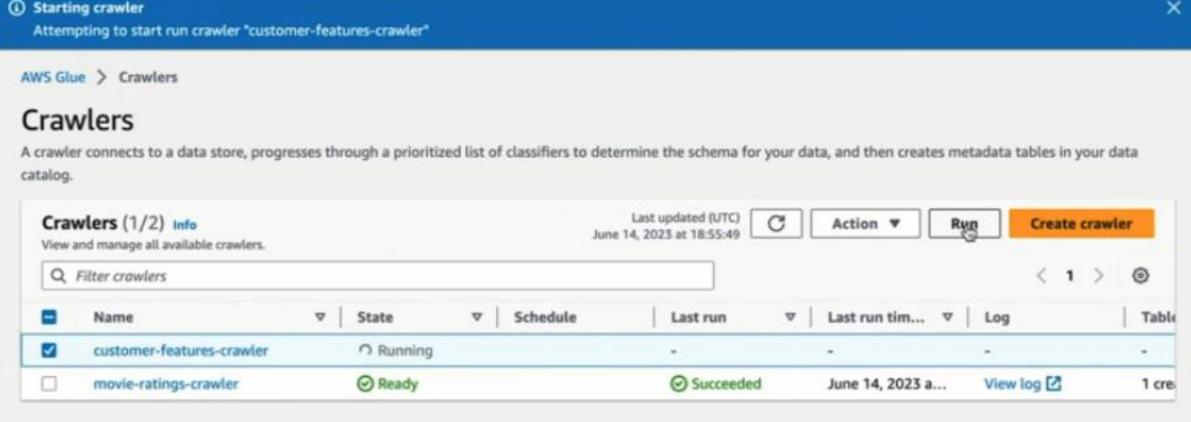
Crawlers
A crawler connects to a data store, progresses through a prioritized list of classifiers to determine the schema for your data, and then creates metadata tables in your data catalog.

Crawlers (1/2) Info
View and manage all available crawlers.

Name	State	Schedule	Last run	Last run time...	Log	Table changes ...
<input checked="" type="checkbox"/> customer-features-crawler	Ready		-	-	-	-
<input type="checkbox"/> movie-ratings-crawler	Ready		Succeeded	June 14, 2023 a...	View log	1 cre

128. AT THE TOP, YOU WILL SEE THE BLUE FLAG WHICH IS STATING THAT THE CRAWLER IS STARTING. IT IS GOING TO CRAWL THE DATA FROM THE RDS DATABASE TABLE THAT IS

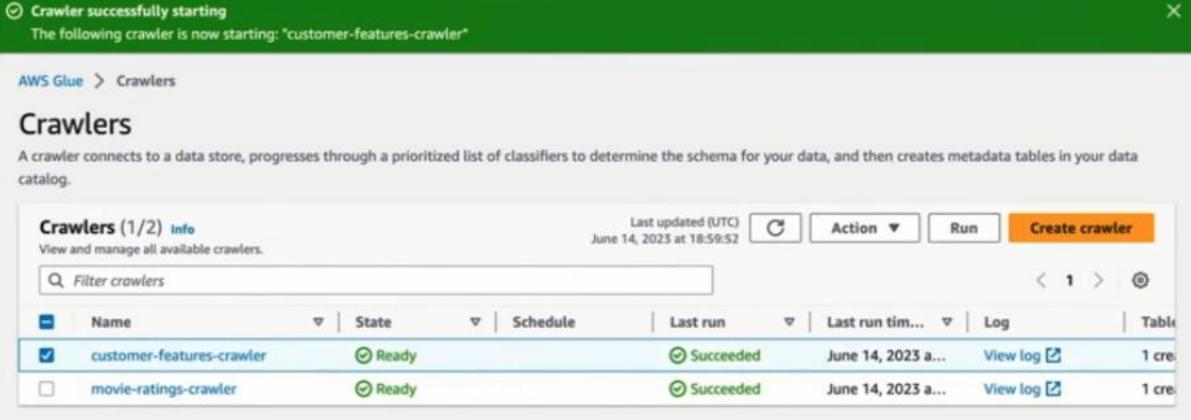
IN THE MYSQL WORKBENCH. IT IS NOW RUNNING AND EXTRACTING THE DATA AS IN THE FIRST STEP OF THE ETL PROCESS WHICH MEANS THAT IT IS IN THE EXTRACT STEP.



The screenshot shows the AWS Glue Crawlers interface. At the top, a blue header bar displays a progress message: "Starting crawler Attempting to start run crawler 'customer-features-crawler'". Below this, the main title "Crawlers" is shown, followed by a brief description: "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." A table titled "Crawlers (1/2) Info" lists two crawlers:

Name	State	Schedule	Last run	Last run time...	Log	Table
customer-features-crawler	Running	-	-	-	-	-
movie-ratings-crawler	Ready	-	Succeeded	June 14, 2023 a...	View log	1 cre

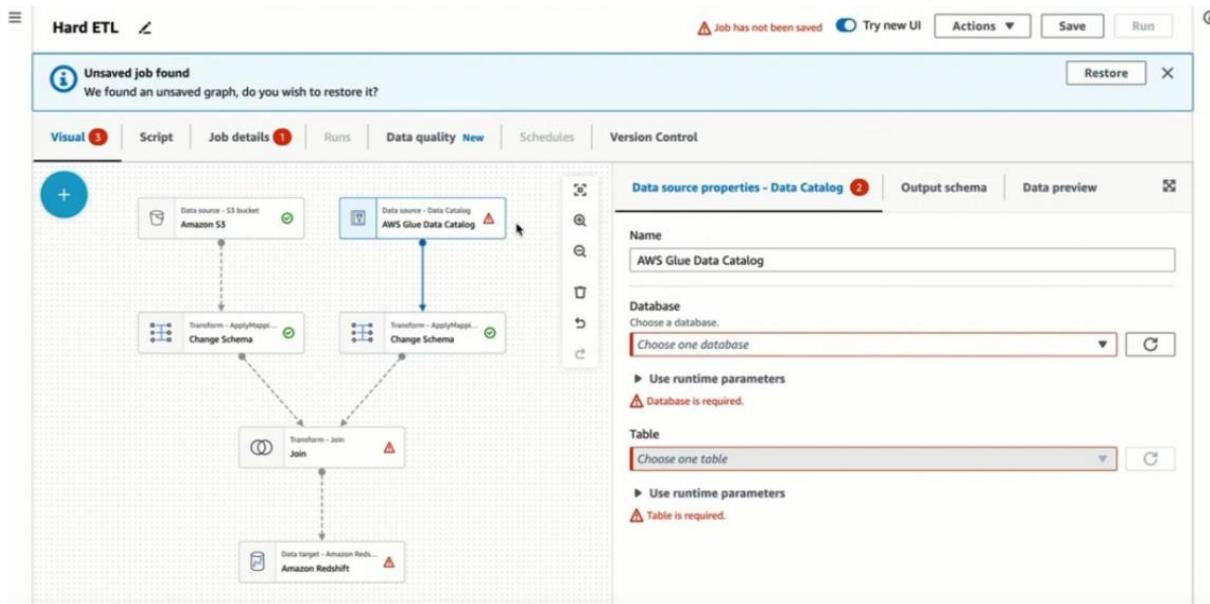
129. CRAWLER SUCCESSFULLY STARTING. THE NEXT STEP IS TO GO BACK TO THE VISUAL EDITOR.



The screenshot shows the AWS Glue Crawlers interface. A green success message at the top states: "Crawler successfully starting The following crawler is now starting: 'customer-features-crawler'". Below this, the main title "Crawlers" is shown, followed by a brief description: "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." A table titled "Crawlers (1/2) Info" lists two crawlers:

Name	State	Schedule	Last run	Last run time...	Log	Table
customer-features-crawler	Ready	-	Succeeded	June 14, 2023 a...	View log	1 cre
movie-ratings-crawler	Ready	-	Succeeded	June 14, 2023 a...	View log	1 cre

130. HERE, WE ARE GOING TO CONNECT THE DATA CATALOG ELEMENT RELATED TO OUR DS DATABASE IN THE ETL PROCESS.



131. IN THE NAME PORTION, MAKE SURE THAT THE AWS GLUE DATA CATALOG IS SELECTED. IN THE DATABASE, SELECT CUSTOMER-FEATURE-GLUE-DATABASE THAT WE CREATED.

Version Control

Data source properties - Data Catalog 2

Name
AWS Glue Data Catalog

Database
Choose a database.
Choose one database

Filter databases
customer-features-glue-database
movie-ratings-glue-database

► Use runtime parameters
⚠ Table is required.

132. IN THE TABLE, CHOOSE CUSTOMERFEATURESDATABASE_CUSTOMER_FEATURES.

Table

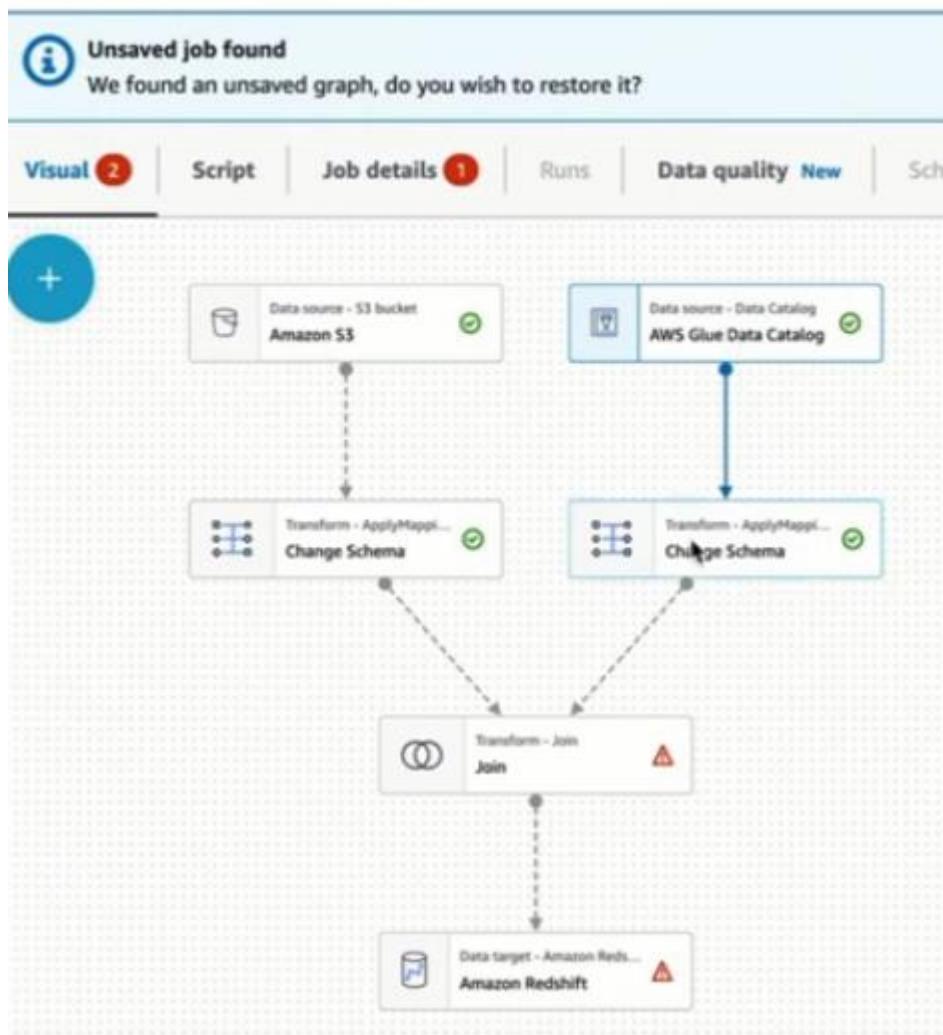
A screenshot of a user interface for selecting a table. At the top, there is a red-bordered input field labeled "Choose one table". Below it is a search bar with a magnifying glass icon and the placeholder text "Filter tables". A list of tables is displayed below the search bar, with the first item, "customerfeaturesdatabase_customer_features", highlighted in blue. To the right of the list is a small "C" icon inside a box.

133. IT SHOULD THEN LOOK LIKE THIS. THIS IS THE STEP OF THE EXTRACTION OF OUR SECOND DATA SOURCE RELATED TO RDS.

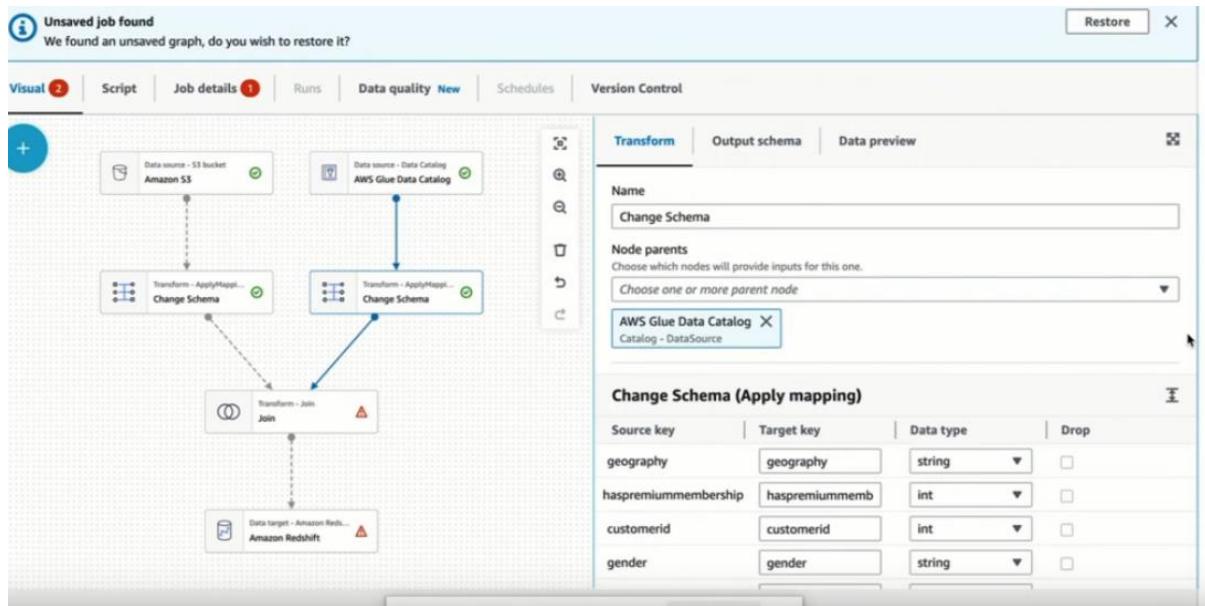
A screenshot of the "Data source properties - Data Catalog" tab in the AWS Glue Data Catalog configuration interface. The tab is selected, indicated by a blue border. There are three tabs at the top: "Data source properties - Data Catalog" (selected), "Output schema", and "Data preview". On the far right, there is a "Close" button. The interface is divided into sections: "Name" (containing "AWS Glue Data Catalog"), "Database" (containing "customer-features-glue-database" and a "Use runtime parameters" link), and "Table" (containing "customerfeaturesdatabase_customer_features" and a "Use runtime parameters" link). Each section has a dropdown arrow and a "Close" button to its right.

134. WE, THEN, MOVE ON TO THE TRANSFORM CHANGE SCHEMA PHASE.

Hard ETL ↴



LET'S CHECK IF WE ARE GOING TO TRANSFORM A BIT OUR DATA RELATED TO THE CUSTOMER FEATURES.



HERE, WE CAN CHANGE THE TARGET KEY AND DATA TYPES BASED ON THE PREFERENCE. BUT FOR THIS TRAINING PURPOSES, WE WILL PUT IT AS IT IS.

Version Control

Transform | Output schema | Data preview

Choose which nodes will provide inputs for this one. Choose one or more parent node

AWS Glue Data Catalog X Catalog - DataSource

Change Schema (Apply mapping)

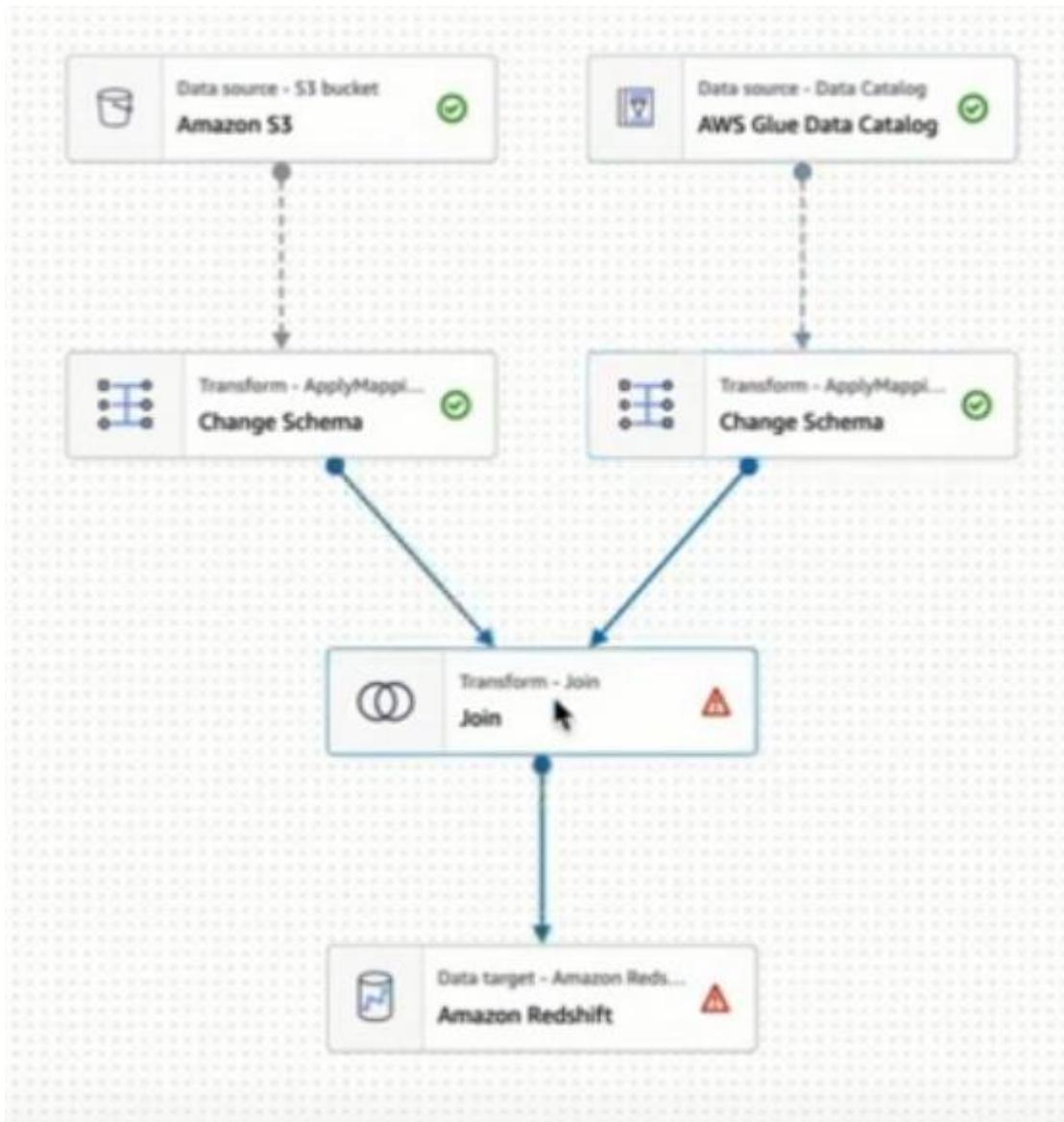
Source key	Target key	Data type	Drop
geography	geography	string	<input type="checkbox"/>
haspremiummembership	haspremiummemb	int	<input type="checkbox"/>
customerid	customerid	int	<input type="checkbox"/>
gender	gender	string	<input type="checkbox"/>
isactivemember	isactivemember	int	<input type="checkbox"/>
surname	surname	string	<input type="checkbox"/>
age	age	int	<input type="checkbox"/>

135. BUT TO TAKE INTO CONSIDERATION IN ORDER TO BUILD THE RECOMMENDER SYSTEM, WE CAN CHECK IF THERE IS ANY FEATURE OR VARIABLES THAT WE CAN DROP BECAUSE THEY ARE NOT RELEVANT TO MAKING A RECOMMENDER SYSTEM FOR MOVIES. FOR EXAMPLE, THE SURNAME MIGHT NOT BE RELEVANT, SO, WE ARE GOING TO DROP THIS VARIABLE.

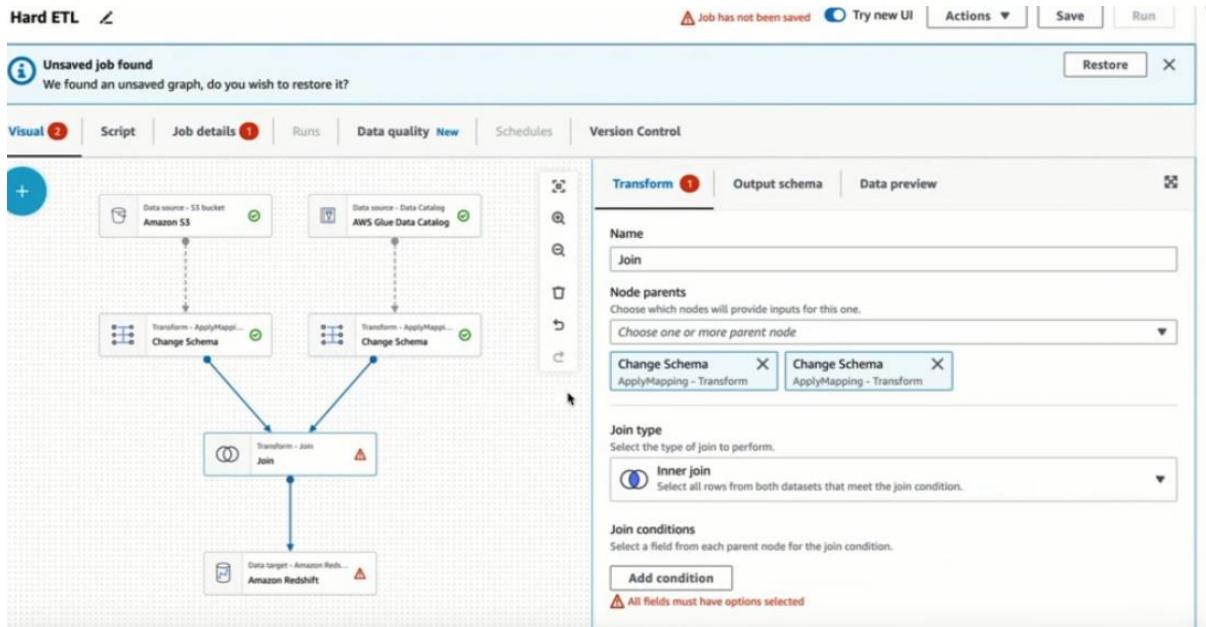
The screenshot shows the AWS Glue Data Catalog interface for transforming data. At the top, there are tabs for 'Transform', 'Output schema', and 'Data preview'. Below the tabs, under 'Node parents', it says 'Choose which nodes will provide inputs for this one.' and 'Choose one or more parent node'. A button labeled 'AWS Glue Data Catalog X Catalog - DataSource' is shown. The main area is titled 'Change Schema (Apply mapping)' and contains a table:

Source key	Target key	Data type	Drop
geography	geography	string	<input type="checkbox"/>
haspremiummembership	haspremiummemb	int	<input type="checkbox"/>
customerid	customerid	int	<input type="checkbox"/>
gender	gender	string	<input type="checkbox"/>
isactivemember	isactivemember	int	<input type="checkbox"/>
surname			<input checked="" type="checkbox"/>
age	age	int	<input type="checkbox"/>

136. WE WILL THEN PROCEED TO THE NEXT STEP, JOIN, WHICH IS STILL PART OF THE TRANSFORM STEP IN THE ETL PROCESS.



137. HERE, WE ARE GOING TO DO THE INNER JOIN TO JOIN THE RESULTED TABLES OF THE TRANSFORMATION COMING FROM THE S3, THE MOVIE RATINGS, AND THE ONE COMING FROM THE RDS DATABASE TABLE OF THE CUSTOMER FEATURE. WE ARE JOINING THESE TWO TABLE AND WE HAVE TO JOIN THEM THROUGH A COMMON DENOMINATOR OR A COMMON VARIABLE.



138. TO CHOOSE THE COMMON VARIABLE, WE NEED TO CLICK THE ADD CONDITION IN THE JOIN CONDITIONS FIELD.

Join type
Select the type of join to perform.

Inner join
Select all rows from both datasets that meet the join condition.

Join conditions
Select a field from each parent node for the join condition.

Add condition

⚠ All fields must have options selected

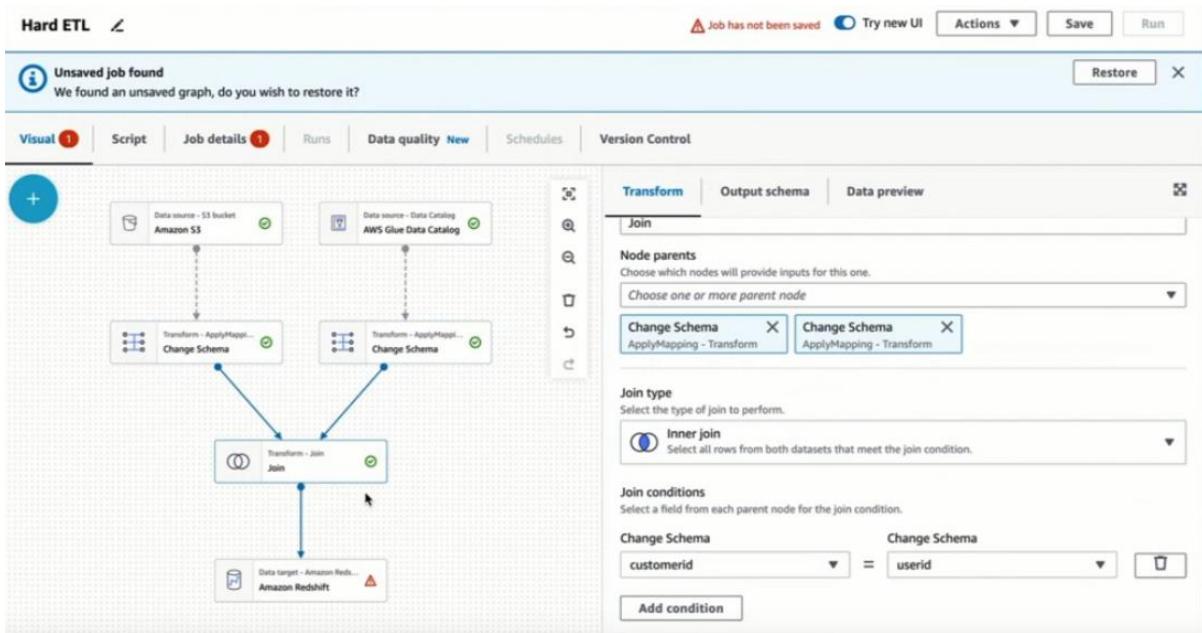
139. IN THE CHANGE SCHEMA, CHOOSE CUSTOMERID AND USERID. WE ARE GOING TO JOIN THEM THROUGH THIS COMMON DENOMINATOR OR VARIABLE WHICH ARE THE IDS OF THE CUSTOMERS OR USERS OF THIS MOVIE STREAMING PLATFORM.

Join conditions
Select a field from each parent node for the join condition.

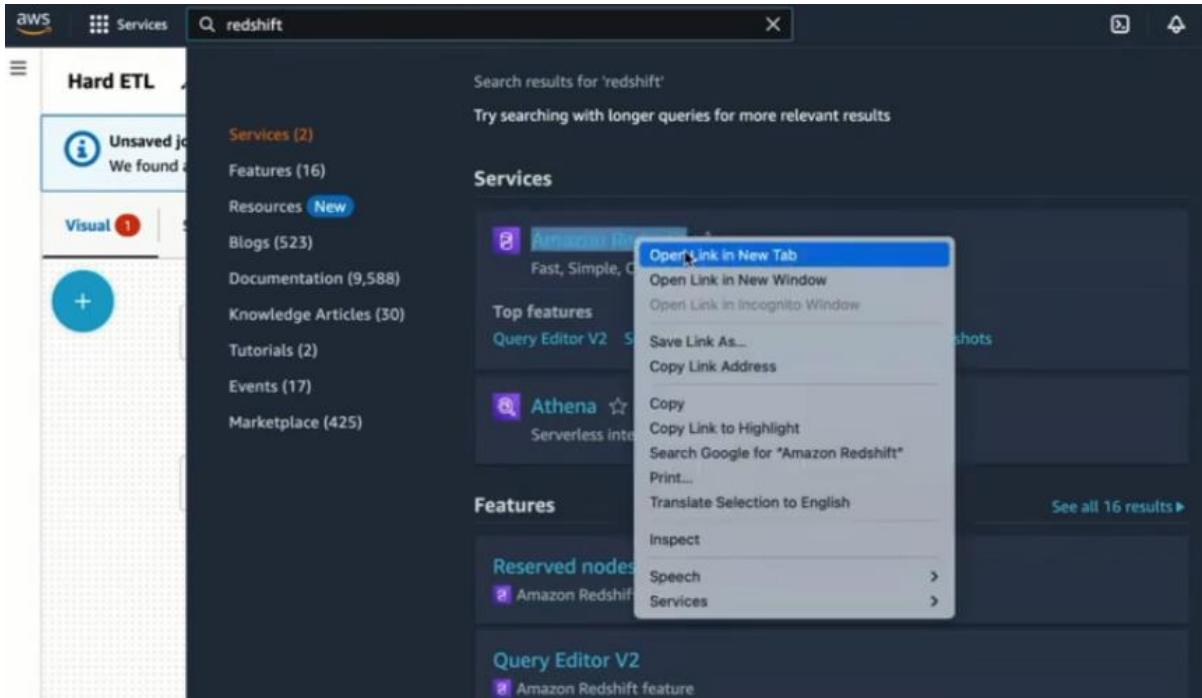
Change Schema customerid	=	Change Schema userid
------------------------------------	---	--------------------------------

Add condition

140. THIS WILL THEN VALIDATES THE TRANSFORM CELL. YOU ARE NOW DONE TO THE TRANSFORM PART OF THE ETL PROCESS. KEEP THE TAB OPEN.



141. MOVING ON TO THE FINAL STEP IS THE TARGET AMAZON REDSHIFT. IN THE SERVICES, TYPE REDSHIFT. OPEN IT IN THE NEW TAB.



142. CLICK THE TRY REDSHIFT SERVERLESS FREE TRIAL.



143. CLICK DEFAULT SETTINGS.

A screenshot of the 'Get started with Amazon Redshift Serverless' configuration page. At the top, it shows the URL 'Amazon Redshift Serverless > Get started with Amazon Redshift Serverless'. The main heading is 'Get started with Amazon Redshift Serverless' with an 'Info' link. Below it, a subtext explains that users will receive \$293.60 credit towards usage. Two options are presented: 'Use default settings' (selected) and 'Customize settings'. The 'Use default settings' option includes a note that default settings help get started and can be changed later. A 'How it works' section is partially visible at the bottom.

144. IN THE PERMISSIONS, WE HAVE TO CREATE ANOTHER ROLE.

Permissions

- ⓘ Associate an IAM role so that your serverless endpoint can LOAD and UNLOAD data. You can create an IAM role as the default for this configuration that has the [AmazonRedshiftAllCommandsFullAccess](#) policy attached. This policy includes permissions to run SQL commands to COPY, UNLOAD, and query data with Amazon Redshift Serverless. This policy also grants permissions to run SELECT statements for related services, such as Amazon S3, Amazon CloudWatch logs, Amazon SageMaker, and AWS Glue. You won't be able to run these SQL commands without an IAM role attached to your namespace.

Associated IAM roles (0)

Create, associate, or remove an IAM role. You can associate up to 50 IAM roles. You can also choose an IAM role and set it as the default.

[Set default](#)

[Manage IAM roles](#)

Search for associated IAM role by name, status, or role type

< 1 >

IAM roles	Status	Role type
No resources No associated IAM roles Associate IAM role		