AWS RedShift Tasks

Introduction

Amazon Redshift is a fully managed, petabyte-scale data warehouse service designed for analytical workloads. Here are the key features and concepts to know:

1. Architecture

- **Cluster-Based**: Redshift uses a cluster architecture composed of one leader node and multiple compute nodes.
- **Leader Node**: Manages query coordination, parsing, and optimization.
- **Compute Nodes**: Store data and perform query execution. They are divided into slices for parallel processing.

2. Data Storage

- **Columnar Storage**: Redshift stores data in a columnar format, which optimizes I/O and improves performance for analytical queries.
- **Compression**: Supports various compression algorithms to reduce storage costs and improve query performance.

3. Performance Optimization

- **Distribution Styles**: Data can be distributed among nodes using different styles (KEY, EVEN, ALL) to optimise query performance.
- **Sort Keys**: Define how data is sorted, improving performance for certain query types.
- **Concurrency Scaling**: Automatically adds transient capacity to handle bursts of queries without impacting performance.

4. SQL Interface

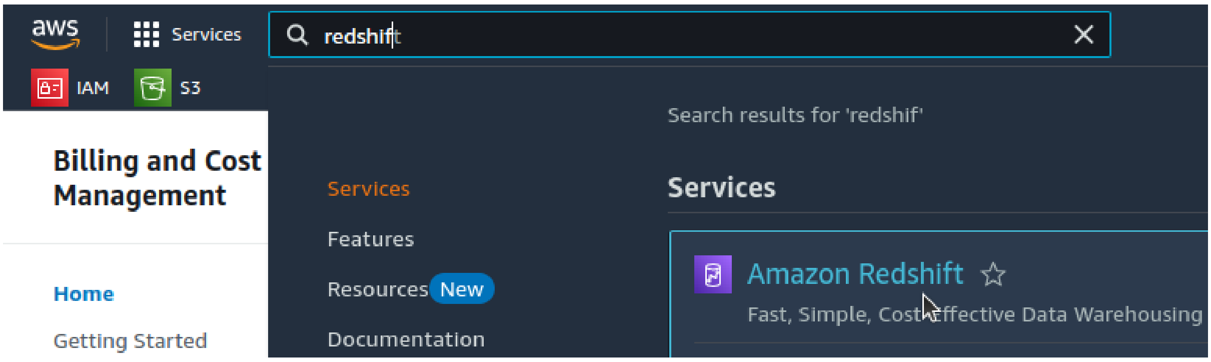
- Supports standard SQL and integrates with various business intelligence tools, enabling users to run complex queries and generate reports.

5. Data Loading

- **COPY Command**: Efficiently loads large amounts of data from Amazon S3, Amazon DynamoDB, or other data sources.
- **Data Formats**: Supports various formats, including CSV, JSON, Parquet, and Avro.

Create RDS Cluster

Navigate to Amazon Redshift



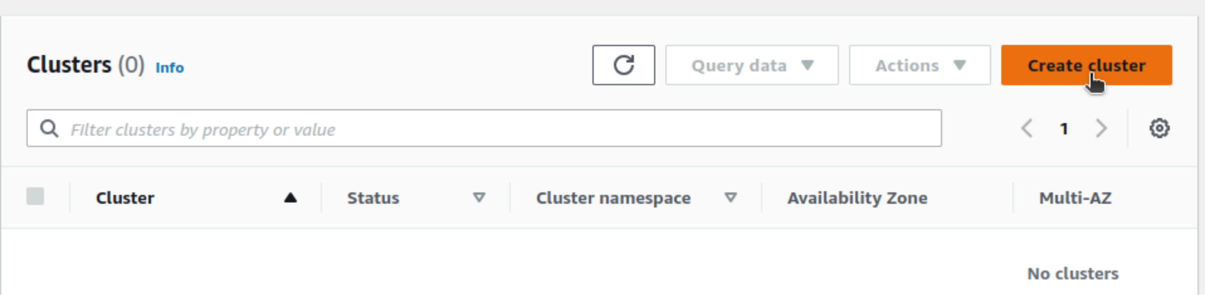
Search results for 'redshif'

Services

Amazon Redshift ☆

Fast, Simple, Cost-effective Data Warehousing

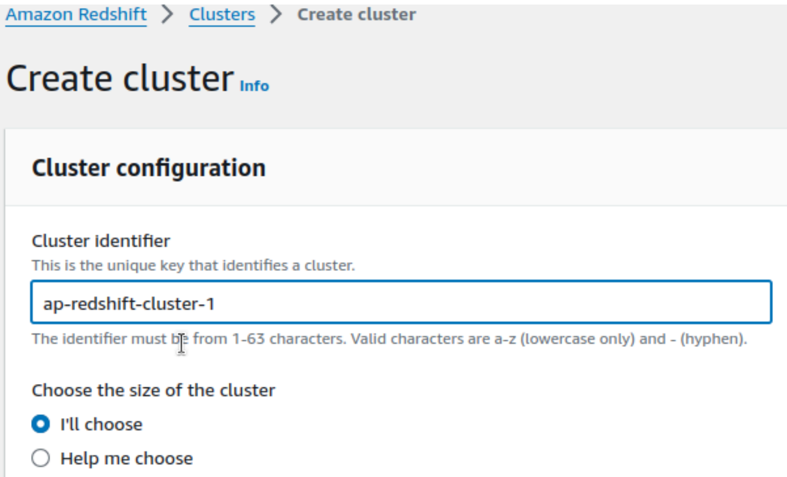
Click on Create Cluster.



□	Cluster	Status	Cluster namespace	Availability Zone	Multi-AZ
					No clusters

Configure your cluster:

Cluster Identifier: Give your cluster a name



Cluster configuration

Cluster identifier

This is the unique key that identifies a cluster.

ap-redshift-cluster-1

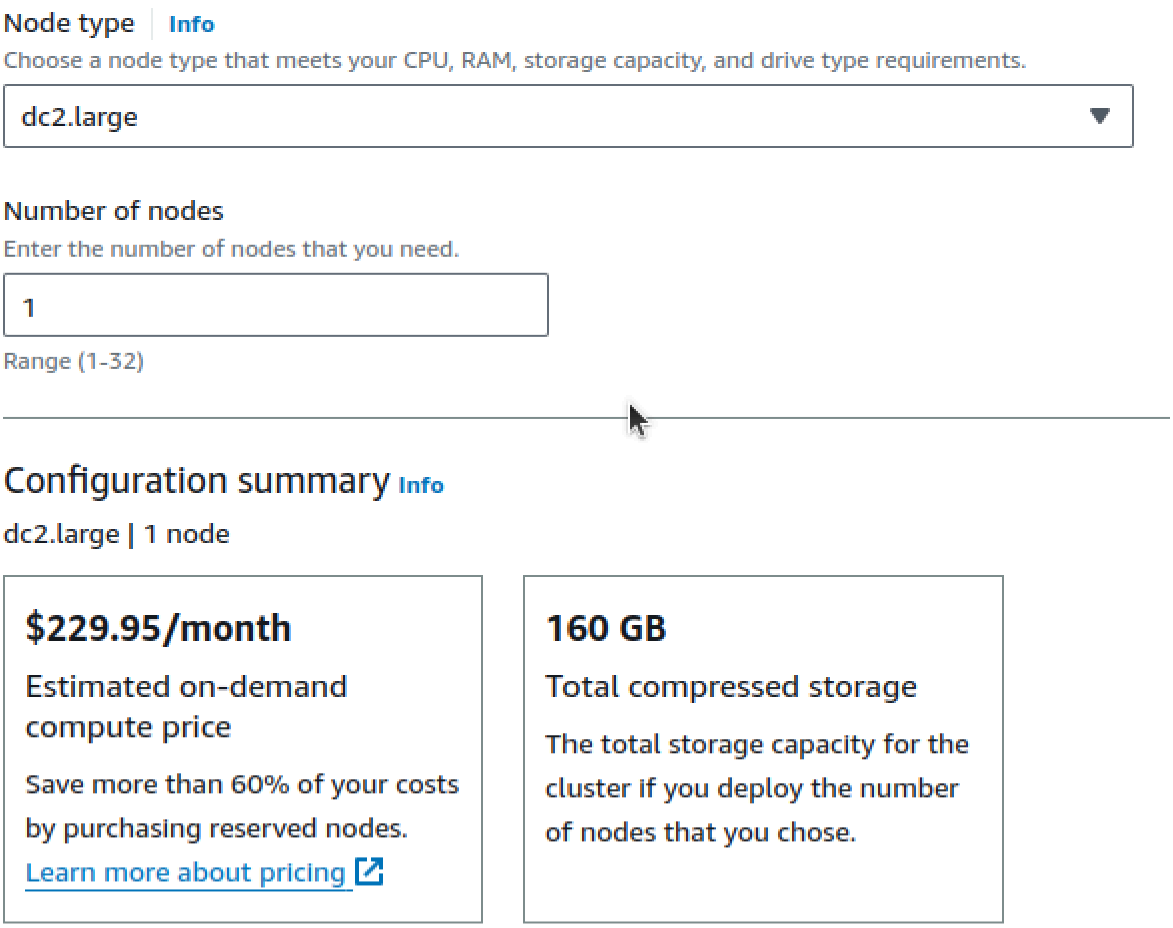
The identifier must be from 1-63 characters. Valid characters are a-z (lowercase only) and - (hyphen).

Choose the size of the cluster

● I'll choose

○ Help me choose

Node Type: Choose the instance type based on your performance needs.



Node type Info

Choose a node type that meets your CPU, RAM, storage capacity, and drive type requirements.

dc2.large ▼

Number of nodes

Enter the number of nodes that you need.

1

Range (1-32)

Configuration summary Info

dc2.large | 1 node

$229.95/month

Estimated on-demand compute price

Save more than 60% of your costs by purchasing reserved nodes.

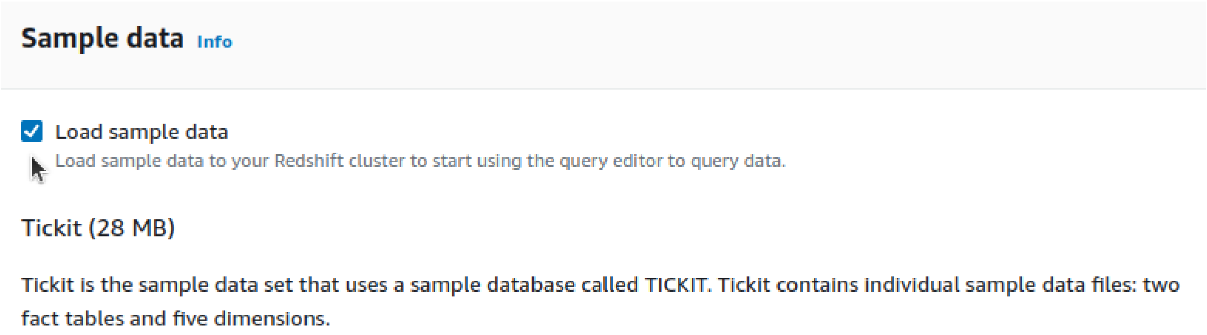
Learn more about pricing

160 GB

Total compressed storage

The total storage capacity for the cluster if you deploy the number of nodes that you chose.

Load Sample Data



Sample data Info

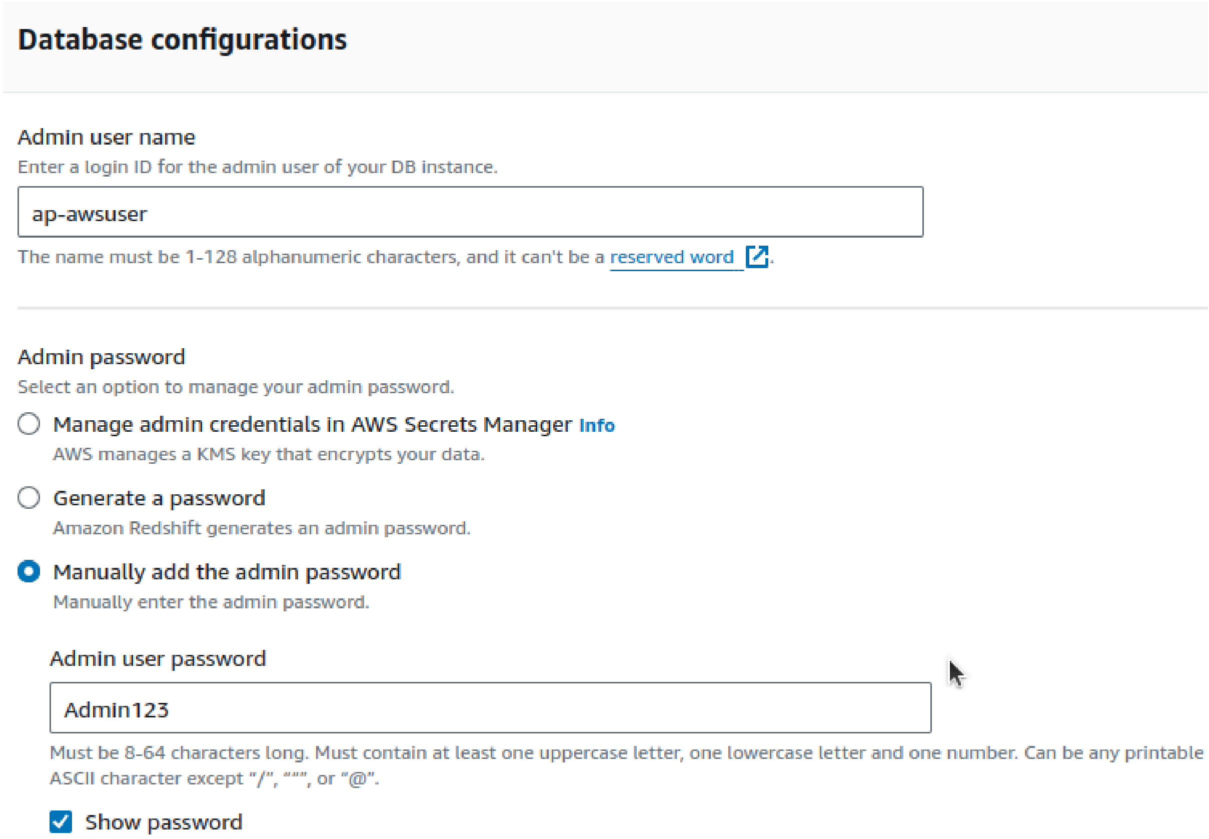
☑ Load sample data

Load sample data to your Redshift cluster to start using the query editor to query data.

Tickit (28 MB)

Tickit is the sample data set that uses a sample database called TICKIT. Tickit contains individual sample data files: two fact tables and five dimensions.

Configure database



Database configurations

Admin user name

Enter a login ID for the admin user of your DB instance.

ap-awsuser

The name must be 1-128 alphanumeric characters, and it can't be a reserved word.

Admin password

Select an option to manage your admin password.

○ Manage admin credentials in AWS Secrets Manager Info

AWS manages a KMS key that encrypts your data.

○ Generate a password

Amazon Redshift generates an admin password.

● Manually add the admin password

Manually enter the admin password.

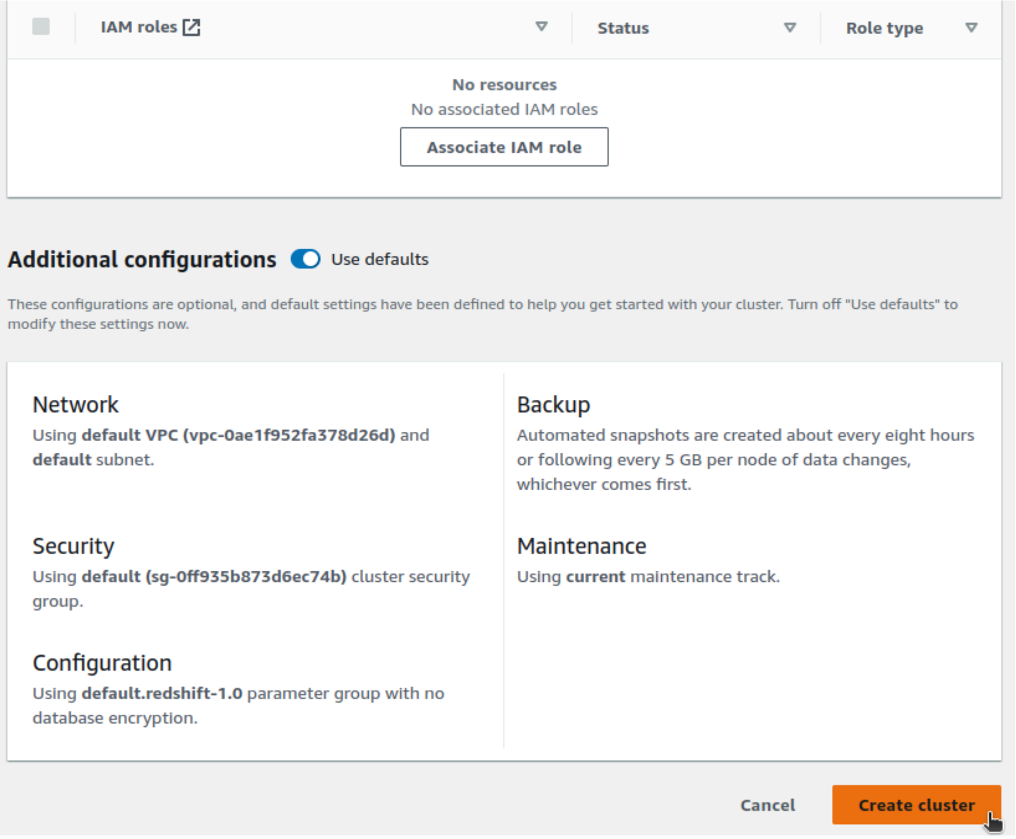
Admin user password

Admin123

Must be 8-64 characters long. Must contain at least one uppercase letter, one lowercase letter and one number. Can be any printable ASCII character except "/", """, or "@".

☑ Show password

Configure additional settings (like VPC, IAM roles) if necessary.



□	IAM roles	Status	Role type

No resources

No associated IAM roles

Associate IAM role

Additional configurations

● Use defaults

These configurations are optional, and default settings have been defined to help you get started with your cluster. Turn off "Use defaults" to modify these settings now.

Network

Using **default VPC (vpc-0ae1f952fa378d26d)** and **default** subnet.

Security

Using **default (sg-0ff935b873d6ec74b)** cluster security group.

Configuration

Using **default.redshift-1.0** parameter group with no database encryption.

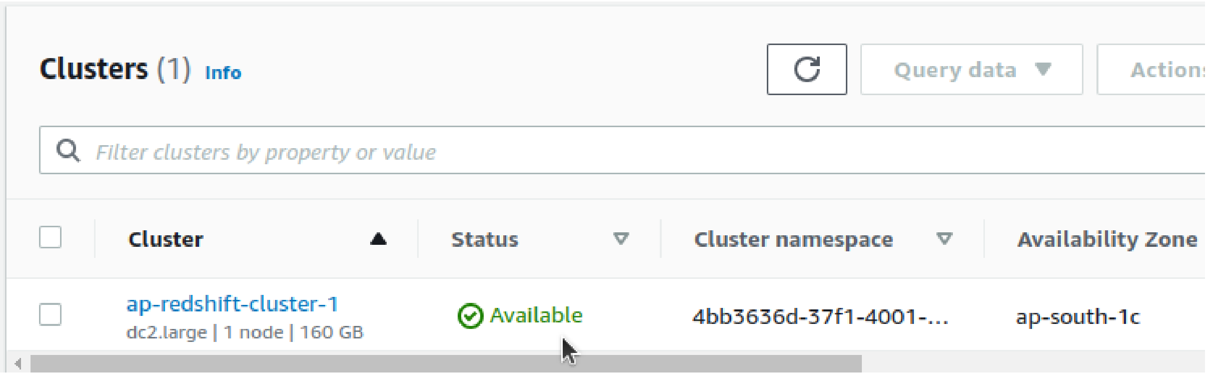
Backup

Automated snapshots are created about every eight hours or following every 5 GB per node of data changes, whichever comes first.

Maintenance

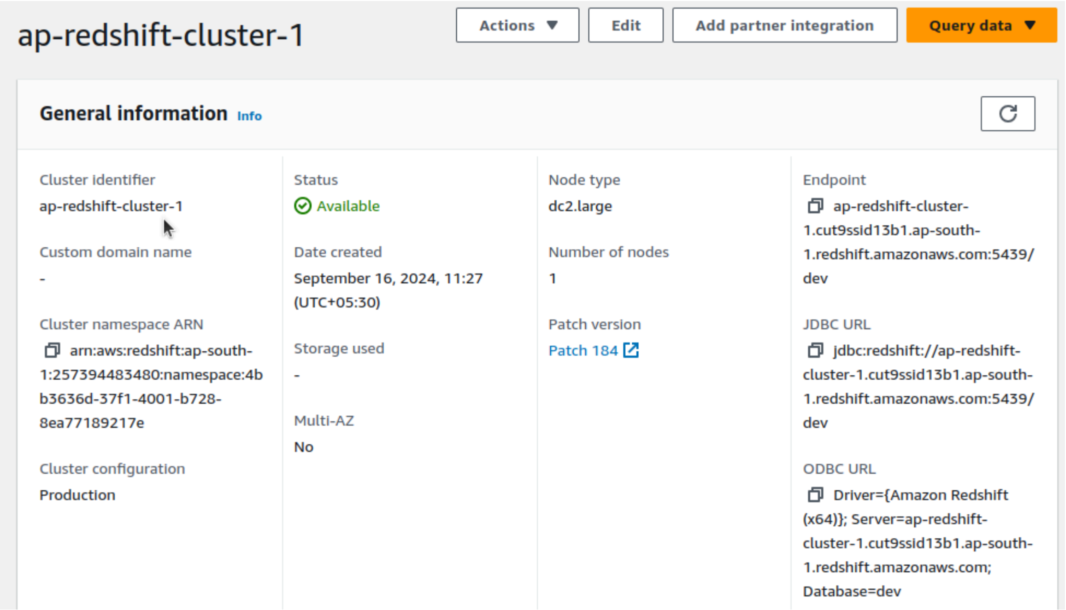
Using **current** maintenance track.

Cluster Created



□	Cluster ▲	Status ▽	Cluster namespace ▽	Availability Zone
□	ap-redshift-cluster-1 dc2.large | 1 node | 160 GB	✓ Available	4bb3636d-37f1-4001-...	ap-south-1c

View cluster information



Cluster identifier
ap-redshift-cluster-1

Custom domain name
-

Cluster namespace ARN
arn:aws:redshift:ap-south-1:257394483480:namespace:4bb3636d-37f1-4001-b728-8ea77189217e

Cluster configuration
Production

Status
✓ Available

Date created
September 16, 2024, 11:27 (UTC+05:30)

Storage used
-

Multi-AZ
No

Node type
dc2.large

Number of nodes
1

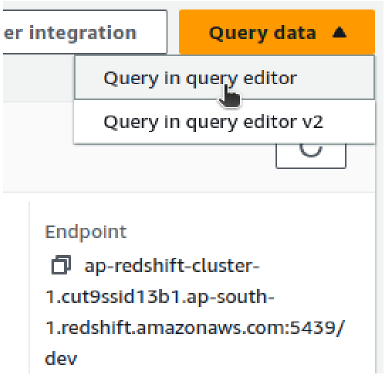
Patch version
Patch 184

Endpoint
ap-redshift-cluster-1.cut9ssid13b1.ap-south-1.redshift.amazonaws.com:5439/dev

JDBC URL
jdbc:redshift://ap-redshift-cluster-1.cut9ssid13b1.ap-south-1.redshift.amazonaws.com:5439/dev

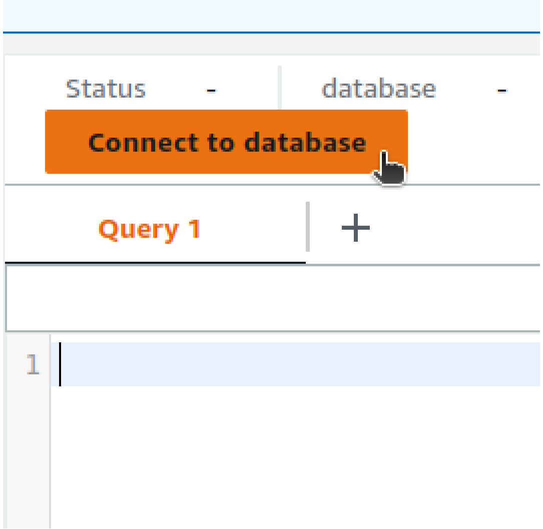
ODBC URL
Driver={Amazon Redshift (x64)}; Server=ap-redshift-cluster-1.cut9ssid13b1.ap-south-1.redshift.amazonaws.com; Database=dev

Open Query Editor

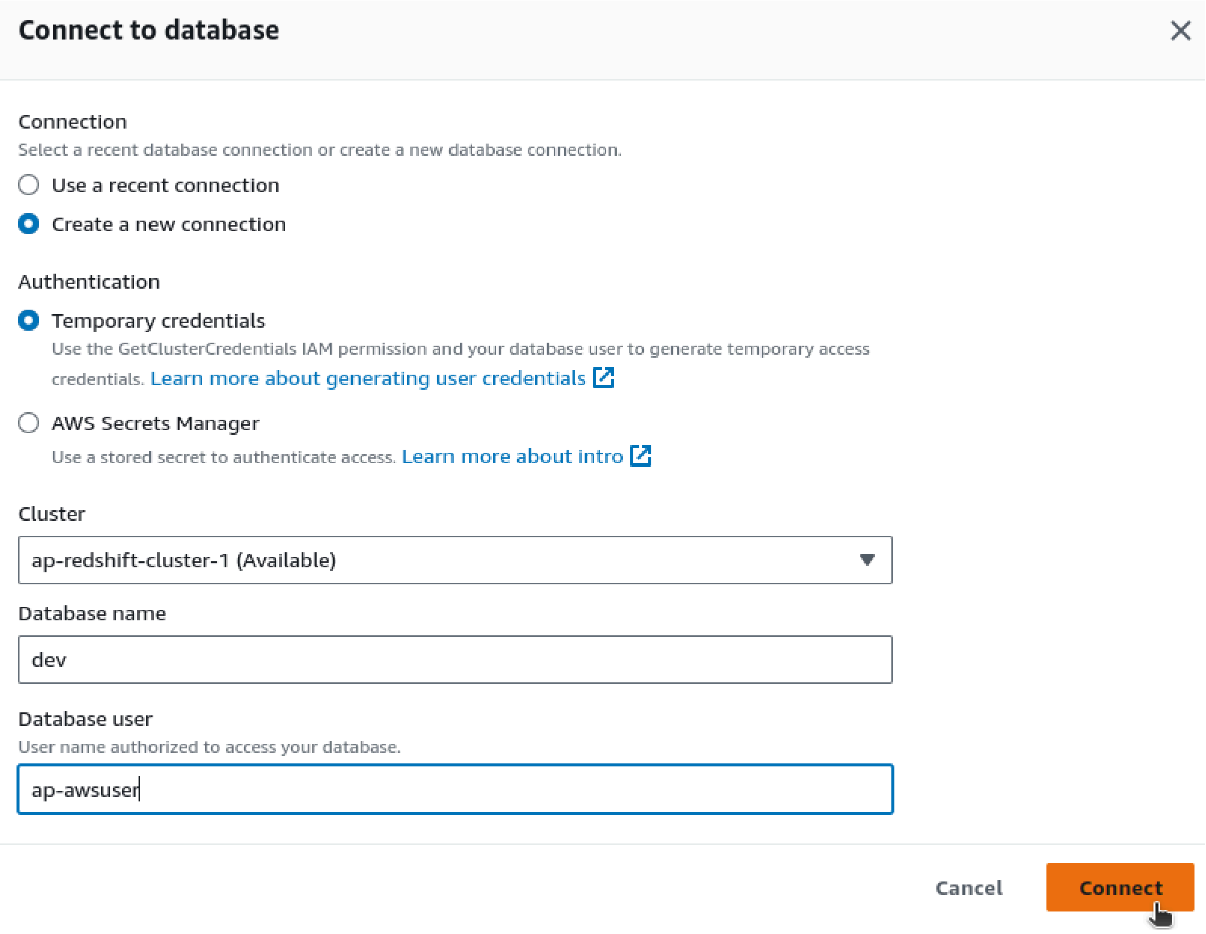


Endpoint
ap-redshift-cluster-1.cut9ssid13b1.ap-south-1.redshift.amazonaws.com:5439/dev

Connect to Your Redshift Cluster



Connect to Database



Connect to database

Connection

Select a recent database connection or create a new database connection.

○ Use a recent connection

● Create a new connection

Authentication

● Temporary credentials

Use the GetClusterCredentials IAM permission and your database user to generate temporary access credentials. Learn more about generating user credentials

○ AWS Secrets Manager

Use a stored secret to authenticate access. Learn more about intro

Cluster

ap-redshift-cluster-1 (Available)

Database name

dev

Database user

User name authorized to access your database.

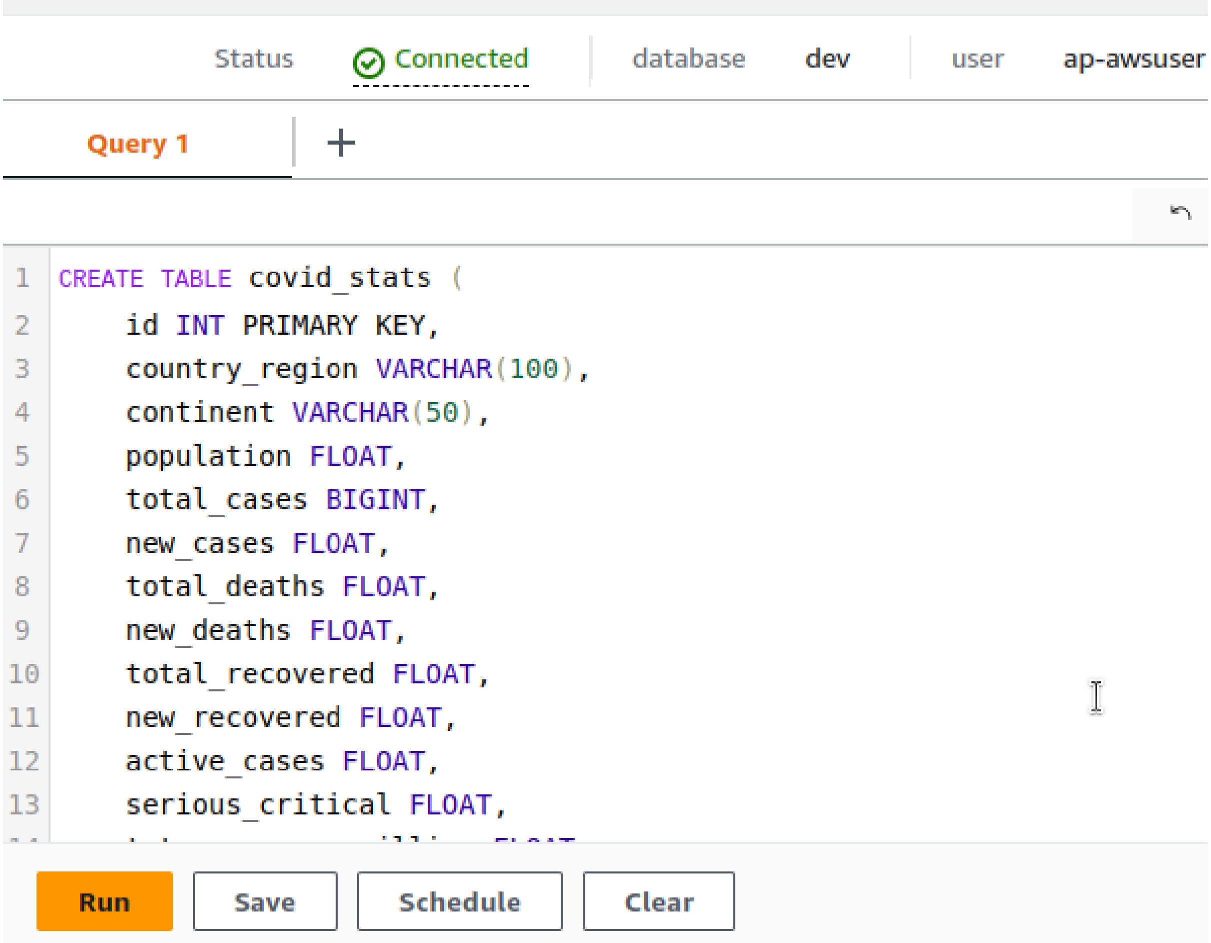
ap-awsuser

Writing query

Define Table Schema in Redshift

Based on the dataset structure, you can create a Redshift table with an `id` column as the primary key.

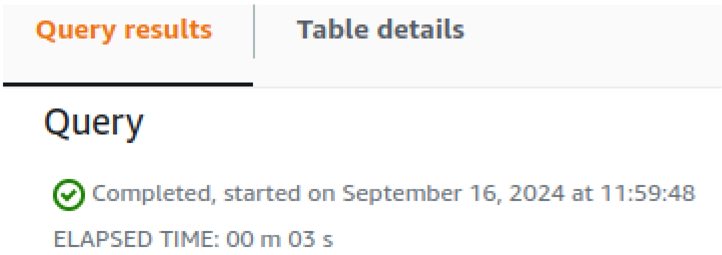
```
CREATE TABLE covid_stats (
    id INT PRIMARY KEY,
    country_region VARCHAR(100),
    continent VARCHAR(50),
    population FLOAT,
    total_cases BIGINT,
    new_cases FLOAT,
    total_deaths FLOAT,
    new_deaths FLOAT,
    total_recovered FLOAT,
    new_recovered FLOAT,
    active_cases FLOAT,
    serious_critical FLOAT,
    tot_cases_per_million FLOAT,
    deaths_per_million FLOAT,
    total_tests FLOAT,
    tests_per_million FLOAT,
    who_region VARCHAR(50)
);
```



Status ✓ Connected | database dev | user ap-awsuser

```
CREATE TABLE covid_stats (
    id INT PRIMARY KEY,
    country_region VARCHAR(100),
    continent VARCHAR(50),
    population FLOAT,
    total_cases BIGINT,
    new_cases FLOAT,
    total_deaths FLOAT,
    new_deaths FLOAT,
    total_recovered FLOAT,
    new_recovered FLOAT,
    active_cases FLOAT,
    serious_critical FLOAT,
```

Query Output



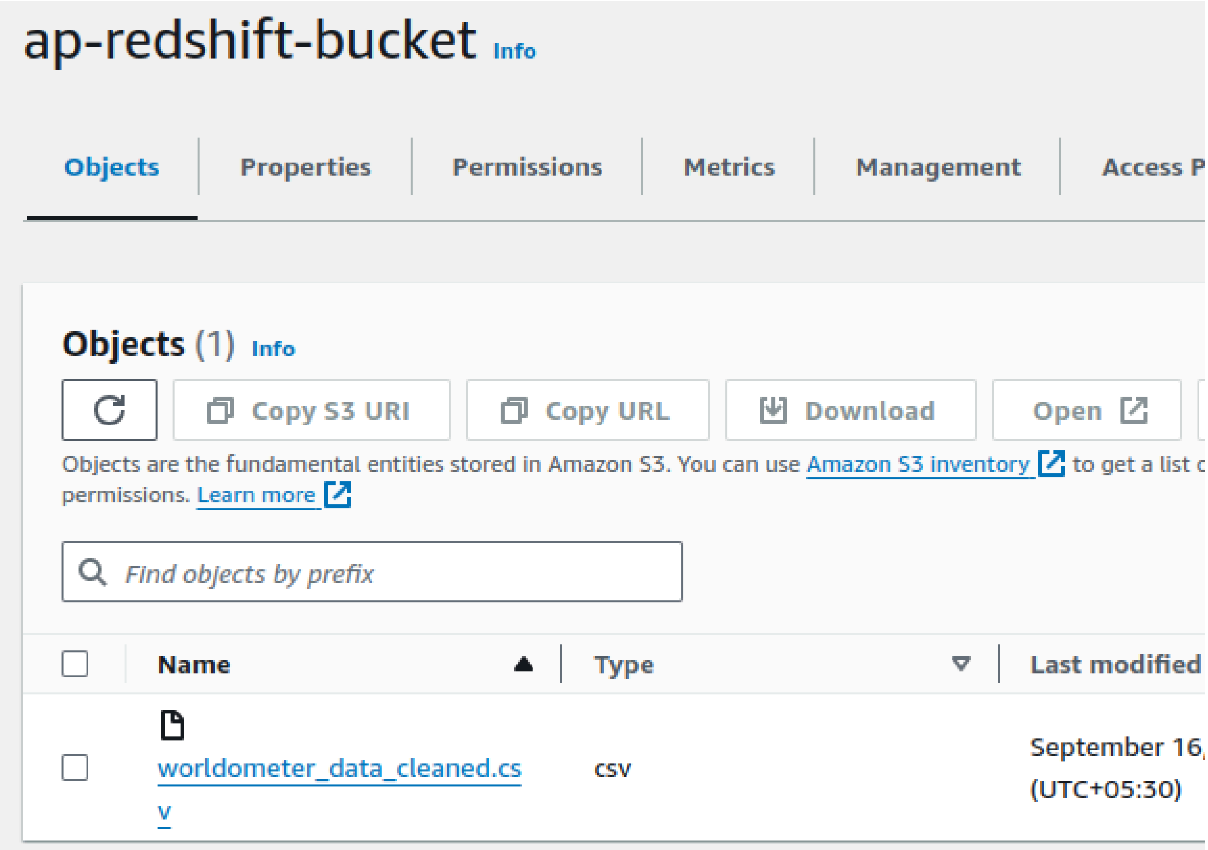
Query

✓ Completed, started on September 16, 2024 at 11:59:48

ELAPSED TIME: 00 m 03 s

Create a s3 bucket

Creating a s3 bucket with a name ‘ap-redshift-bucket’ and adding a csv file

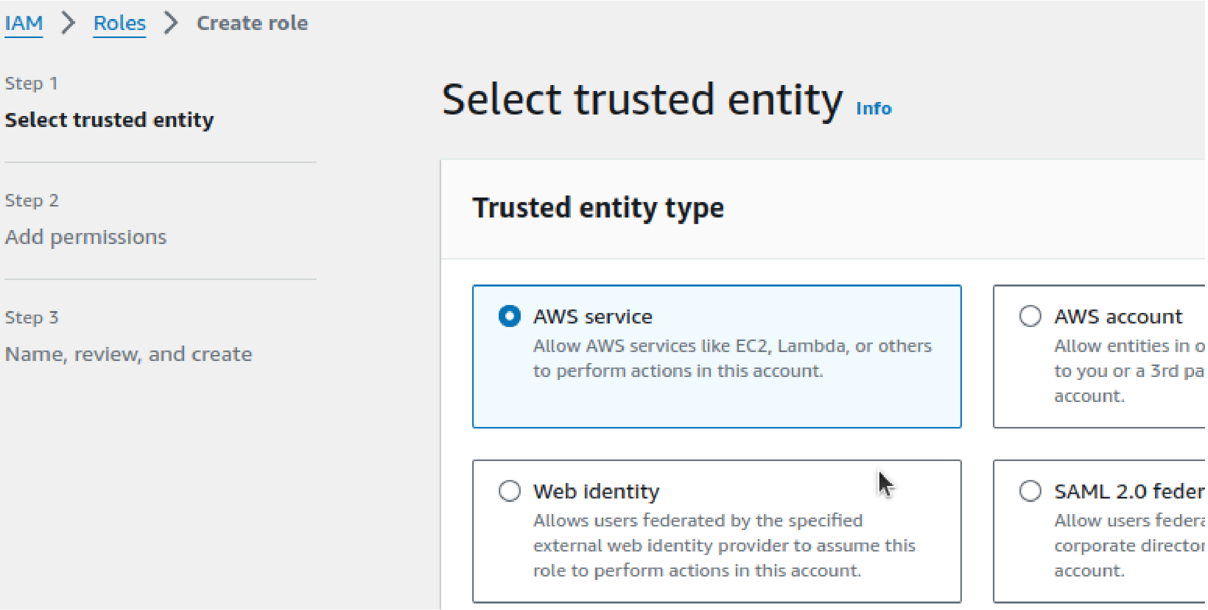


Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list c permissions. Learn more

□	Name ▲	Type ▽	Last modified
□	worldometer_data_cleaned.csv	csv	September 16, (UTC+05:30)

Creating a Role

Go to the IAM Console in AWS.



Step 1
Select trusted entity

Step 2
Add permissions

Step 3
Name, review, and create

Select trusted entity

Trusted entity type

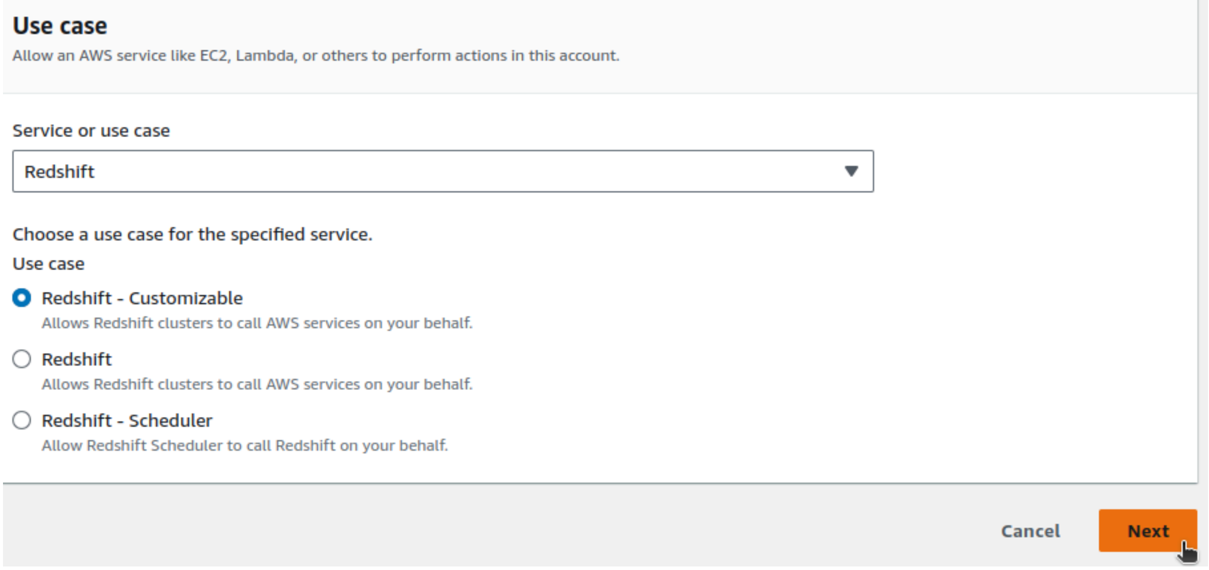
● AWS service
Allow AWS services like EC2, Lambda, or others to perform actions in this account.

○ AWS account
Allow entities in o
to you or a 3rd pa
account.

○ Web identity
Allows users federated by the specified external web identity provider to assume this role to perform actions in this account.

○ SAML 2.0 feder
Allow users federa
corporate director
account.

Specify Service name



Use case
Allow an AWS service like EC2, Lambda, or others to perform actions in this account.

Service or use case
Redshift ▼

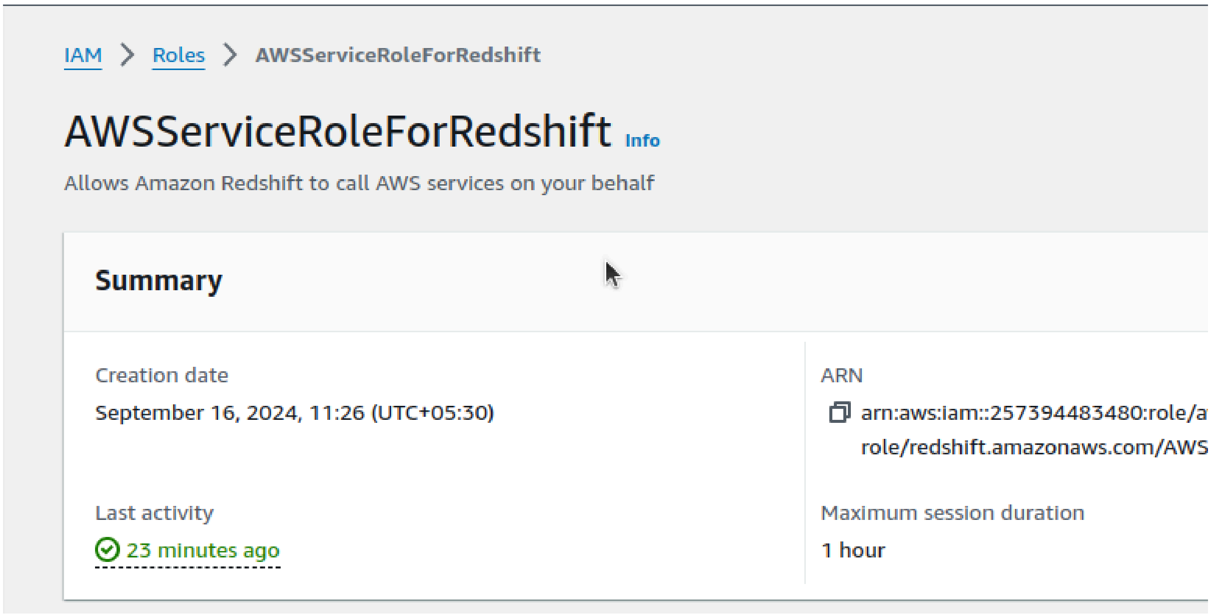
Choose a use case for the specified service.
Use case

● Redshift - Customizable
Allows Redshift clusters to call AWS services on your behalf.

○ Redshift
Allows Redshift clusters to call AWS services on your behalf.

○ Redshift - Scheduler
Allow Redshift Scheduler to call Redshift on your behalf.

Role Created



AWSServiceRoleForRedshift Info

Allows Amazon Redshift to call AWS services on your behalf

Summary

Creation date
September 16, 2024, 11:26 (UTC+05:30)

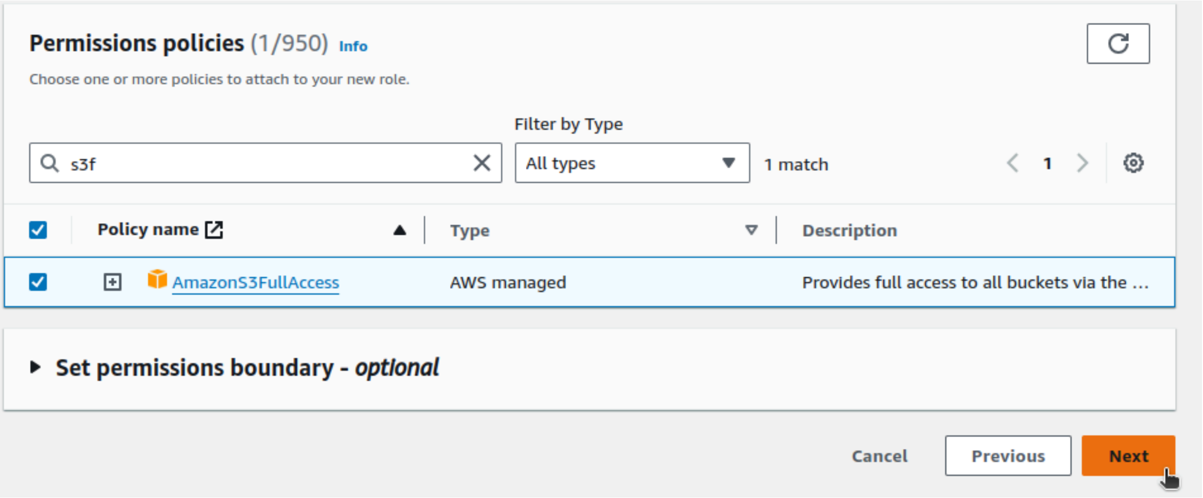
ARN
arn:aws:iam::257394483480:role/a
role/redshift.amazonaws.com/AWS

Last activity
✓ 23 minutes ago

Maximum session duration
1 hour

Add permission

Attach the **AmazonS3ReadOnlyAccess** policy to allow Redshift to read from S3.



Permissions policies (1/950) Info

Choose one or more policies to attach to your new role.

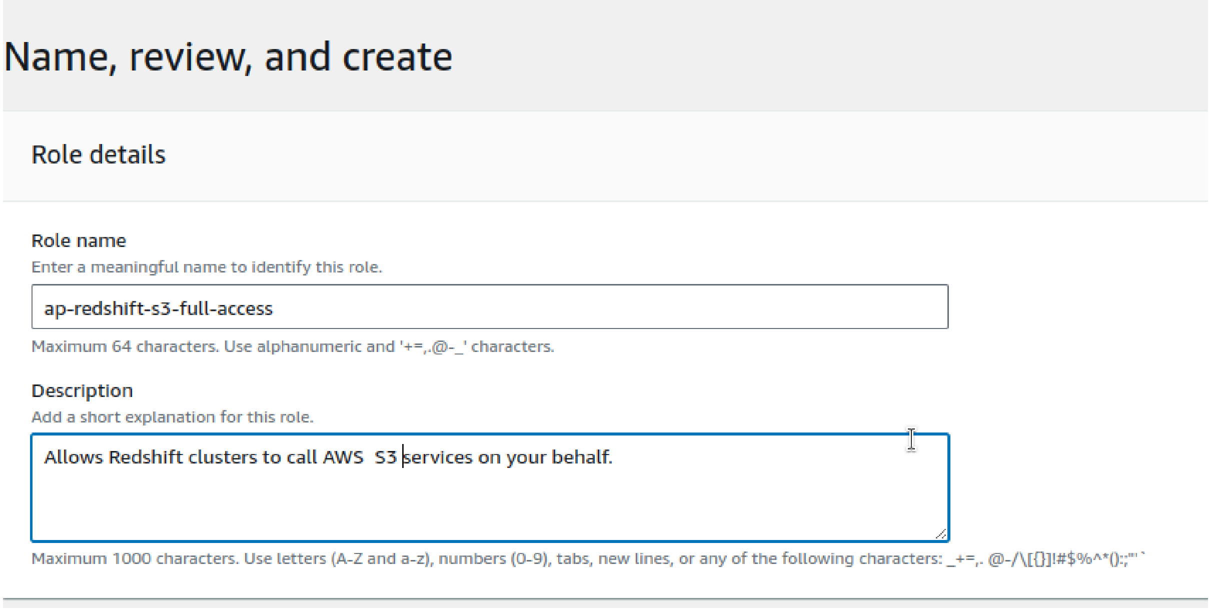
Filter by Type

s3f | All types | 1 match

☑	Policy name	Type	Description
☑	AmazonS3FullAccess	AWS managed	Provides full access to all buckets via the ...

▸ Set permissions boundary - *optional*

Give name to the role



Name, review, and create

Role details

Role name

Enter a meaningful name to identify this role.

ap-redshift-s3-full-access

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

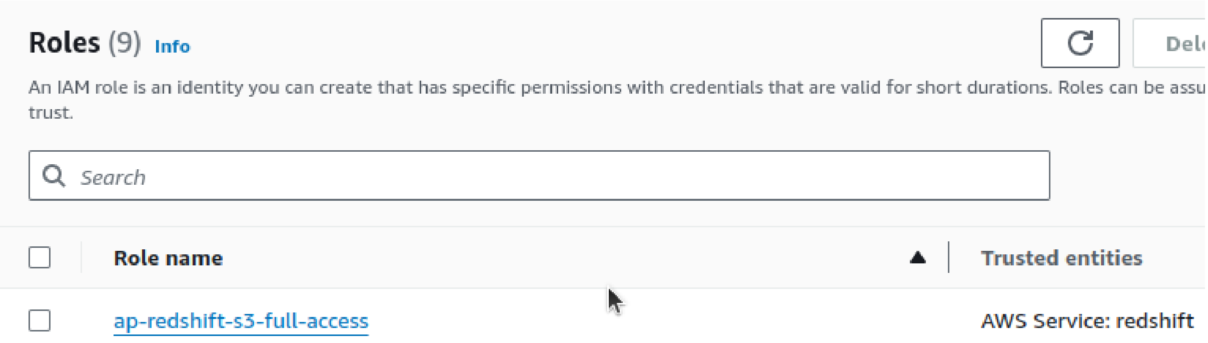
Description

Add a short explanation for this role.

Allows Redshift clusters to call AWS S3 services on your behalf.

Maximum 1000 characters. Use letters (A-Z and a-z), numbers (0-9), tabs, new lines, or any of the following characters: _+=,. @-/\[{}]!#$%^*();"' `

Role Added Successfully



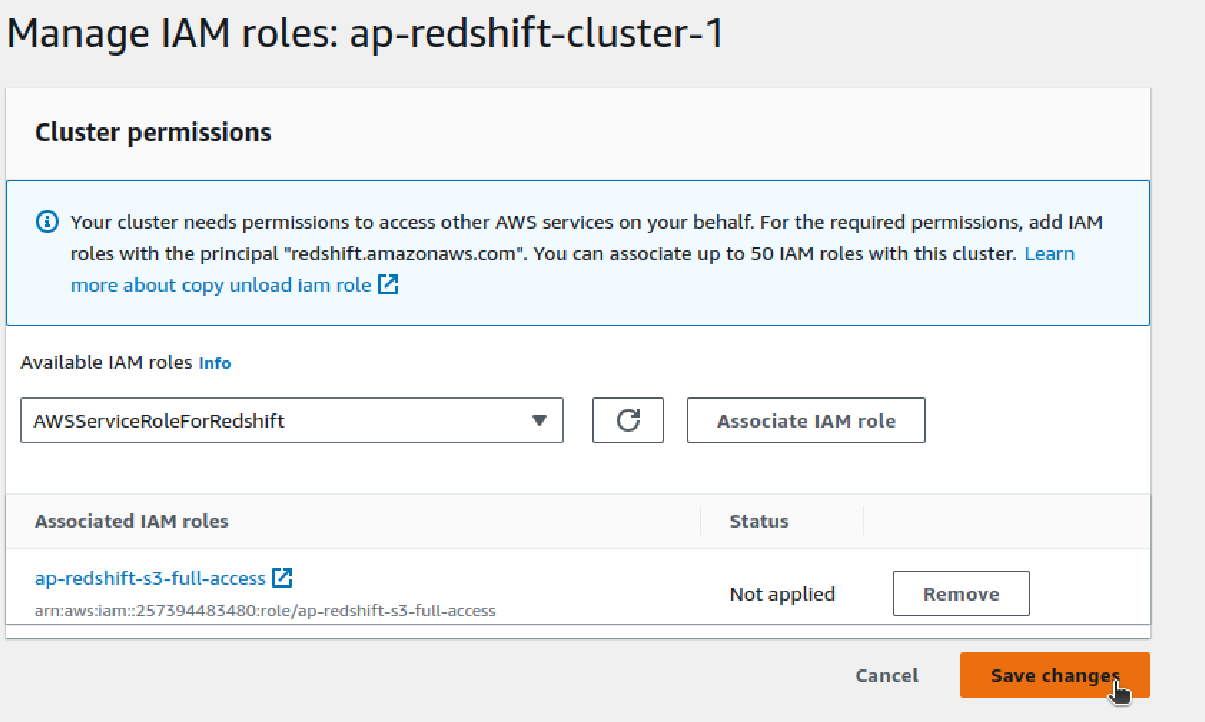
Roles (9) Info

An IAM role is an identity you can create that has specific permissions with credentials that are valid for short durations. Roles can be assu trust.

□	Role name ▲	Trusted entities
□	ap-redshift-s3-full-access	AWS Service: redshift

Adding Role to the Redshift Cluster

- Go to the **Redshift Console**.
- Select your cluster and choose **Manage IAM Roles**.
- Attach the IAM role created in step 1 to your Redshift cluster.



Manage IAM roles: ap-redshift-cluster-1

Cluster permissions

ⓘ Your cluster needs permissions to access other AWS services on your behalf. For the required permissions, add IAM roles with the principal "redshift.amazonaws.com". You can associate up to 50 IAM roles with this cluster. Learn more about copy unload iam role

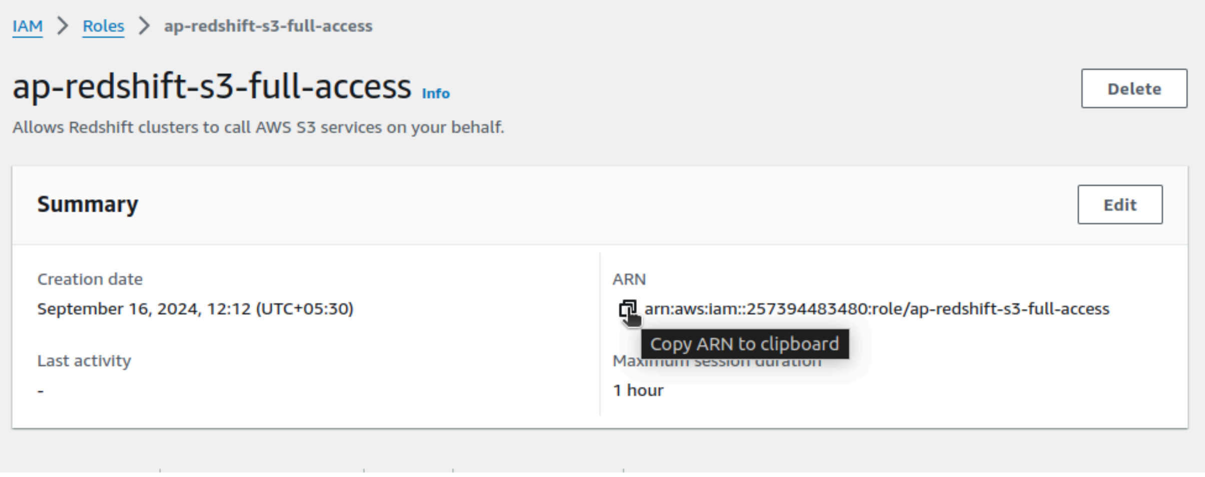
Available IAM roles Info

AWSServiceRoleForRedshift ▼

Associated IAM roles	Status	
ap-redshift-s3-full-access arn:aws:iam::257394483480:role/ap-redshift-s3-full-access	Not applied	Remove

Copy Role ARN

From IAM console copy role arn



IAM > Roles > ap-redshift-s3-full-access

ap-redshift-s3-full-access Info

Allows Redshift clusters to call AWS S3 services on your behalf.

Summary

Creation date
September 16, 2024, 12:12 (UTC+05:30)

Last activity
-

ARN
arn:aws:iam::257394483480:role/ap-redshift-s3-full-access

Copy ARN to clipboard

Maximum session duration
1 hour

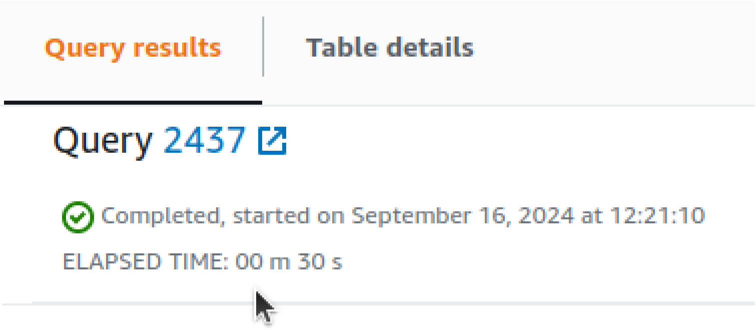
Copy Commands

In the COPY command, replace `'your-iam-role'` with the actual ARN of the IAM role you created:

Code

```
COPY covid_stats
FROM 's3://ap-redshift-bucket/worldometer_data_cleaned.csv'
IAM_ROLE 'arn:aws:iam::257394483480:role/ap-redshift-s3-full-access'
CSV
IGNOREHEADER 1;
```

Query Output

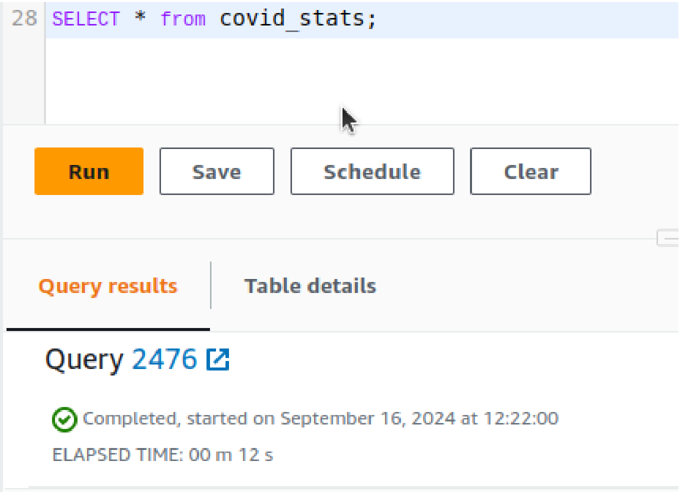


Query 2437

Completed, started on September 16, 2024 at 12:21:10

ELAPSED TIME: 00 m 30 s

Displaying Table Data

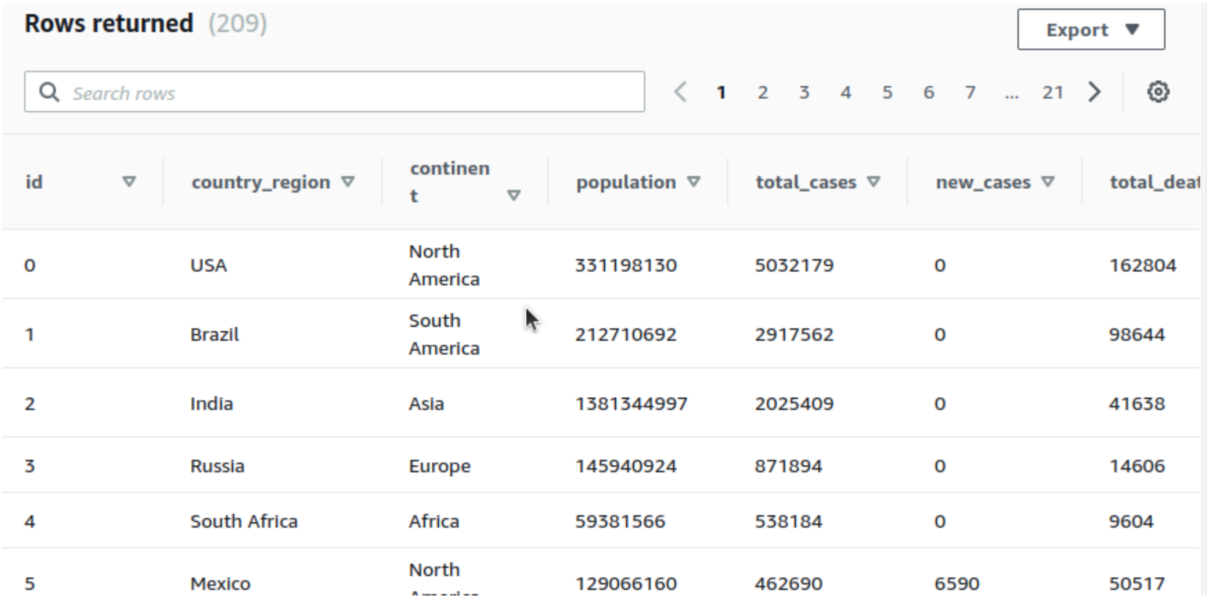


```
SELECT * from covid_stats;
```

Query 2476

Completed, started on September 16, 2024 at 12:22:00

ELAPSED TIME: 00 m 12 s



id	country_region	continent	population	total_cases	new_cases	total_deat
0	USA	North America	331198130	5032179	0	162804
1	Brazil	South America	212710692	2917562	0	98644
2	India	Asia	1381344997	2025409	0	41638
3	Russia	Europe	145940924	871894	0	14606
4	South Africa	Africa	59381566	538184	0	9604
5	Mexico	North	129066160	462690	6590	50517

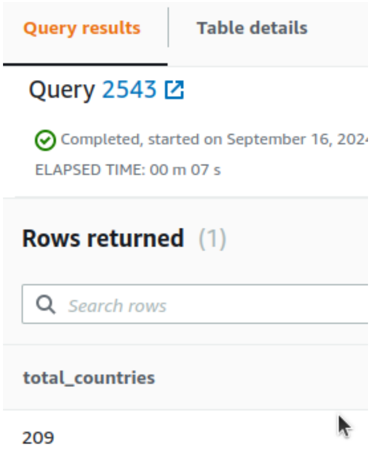
Analysis Tasks

Performing Analysis on the table

Query 1

```
--Get total number of countries and regions
SELECT COUNT(DISTINCT country_region) AS total_countries
FROM covid_stats;
```

Query Output



Completed, started on September 16, 202

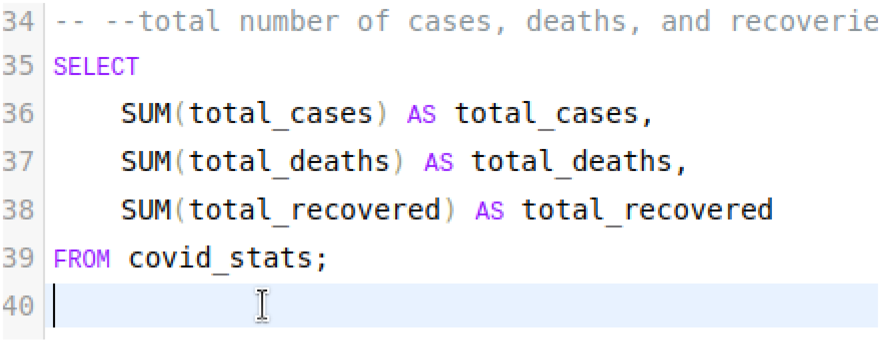
ELAPSED TIME: 00 m 07 s

Rows returned (1)

total_countries
209

Query 2

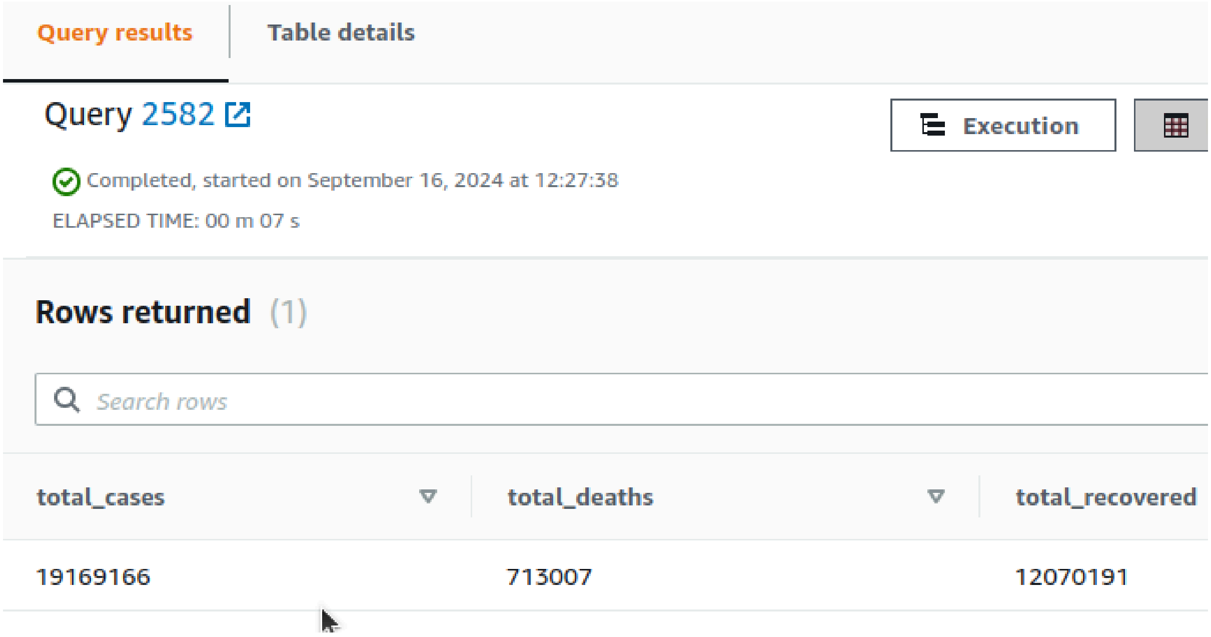
```
-- --total number of cases, deaths, and recoveries
--SELECT
--    SUM(total_cases) AS total_cases,
--    SUM(total_deaths) AS total_deaths,
--    SUM(total_recovered) AS total_recovered
--FROM covid_stats;
```



```
-- --total number of cases, deaths, and recoverie
SELECT
    SUM(total_cases) AS total_cases,
    SUM(total_deaths) AS total_deaths,
    SUM(total_recovered) AS total_recovered
FROM covid_stats;

```

Query Output



Query 2582

Completed, started on September 16, 2024 at 12:27:38

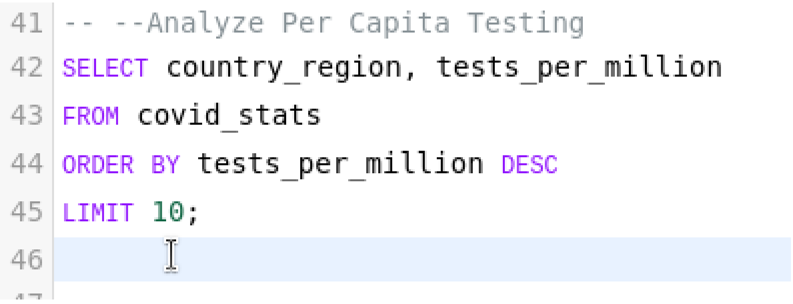
ELAPSED TIME: 00 m 07 s

Rows returned (1)

total_cases	total_deaths	total_recovered
19169166	713007	12070191

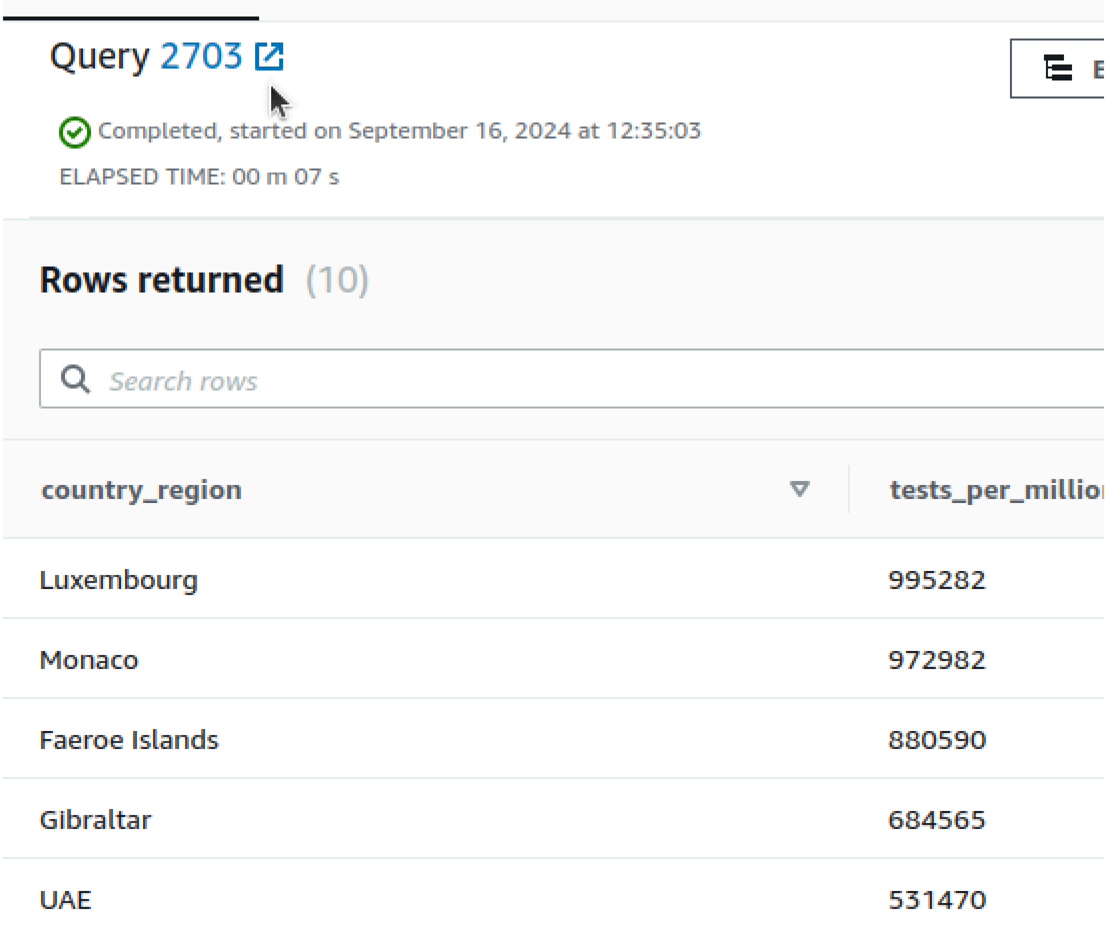
Query 3

```
-- --Analyze Per Capita Testing
SELECT country_region, tests_per_million
FROM covid_stats
ORDER BY tests_per_million DESC
LIMIT 10;
```



```
-- --Analyze Per Capita Testing
SELECT country_region, tests_per_million
FROM covid_stats
ORDER BY tests_per_million DESC
LIMIT 10;
```

Query Output



Query 2703

Completed, started on September 16, 2024 at 12:35:03

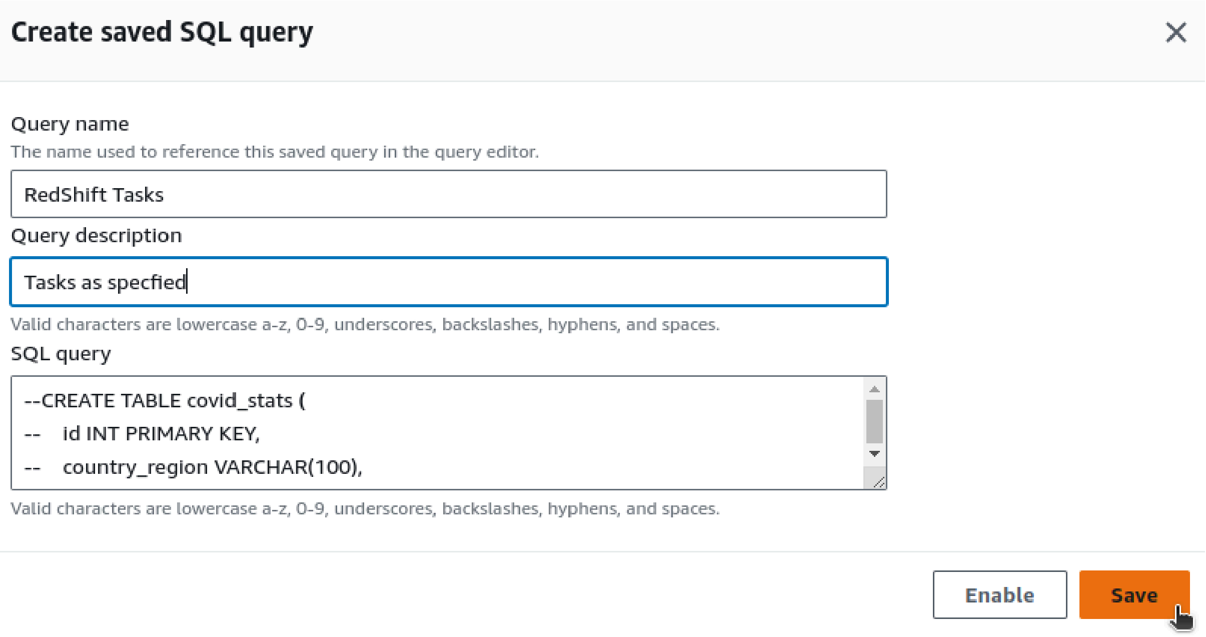
ELAPSED TIME: 00 m 07 s

Rows returned (10)

country_region	tests_per_millio
Luxembourg	995282
Monaco	972982
Faeroe Islands	880590
Gibraltar	684565
UAE	531470

Save Query

You can save query by clicking on 'Save'



Create saved SQL query

Query name

The name used to reference this saved query in the query editor.

RedShift Tasks

Query description

Tasks as specfied

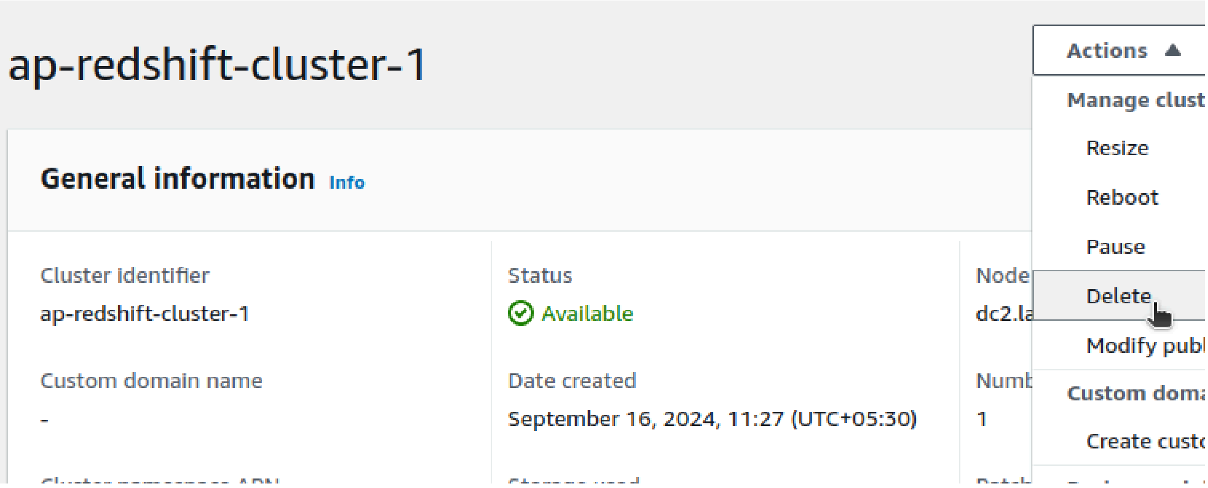
Valid characters are lowercase a-z, 0-9, underscores, backslashes, hyphens, and spaces.

SQL query

```
--CREATE TABLE covid_stats (
--    id INT PRIMARY KEY,
--    country_region VARCHAR(100),
```

Valid characters are lowercase a-z, 0-9, underscores, backslashes, hyphens, and spaces.

Delete Cluster



ap-redshift-cluster-1

General information Info

Cluster identifier: ap-redshift-cluster-1

Status: Available

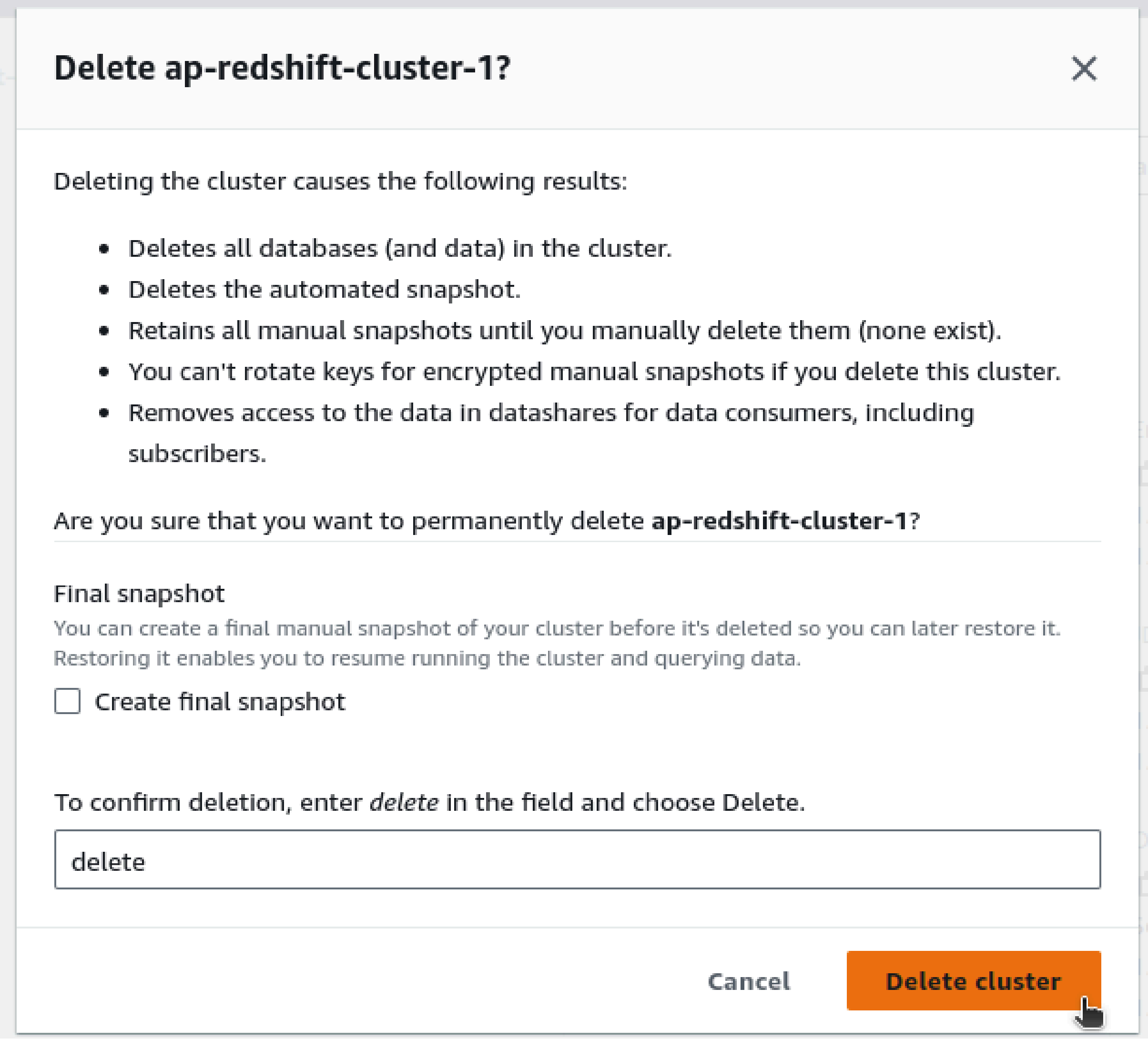
Node: dc2.la

Custom domain name: -

Date created: September 16, 2024, 11:27 (UTC+05:30)

Numb: 1

Confirm Delete



Delete ap-redshift-cluster-1?

Deleting the cluster causes the following results:

- Deletes all databases (and data) in the cluster.
- Deletes the automated snapshot.
- Retains all manual snapshots until you manually delete them (none exist).
- You can't rotate keys for encrypted manual snapshots if you delete this cluster.
- Removes access to the data in datashares for data consumers, including subscribers.

Are you sure that you want to permanently delete **ap-redshift-cluster-1**?

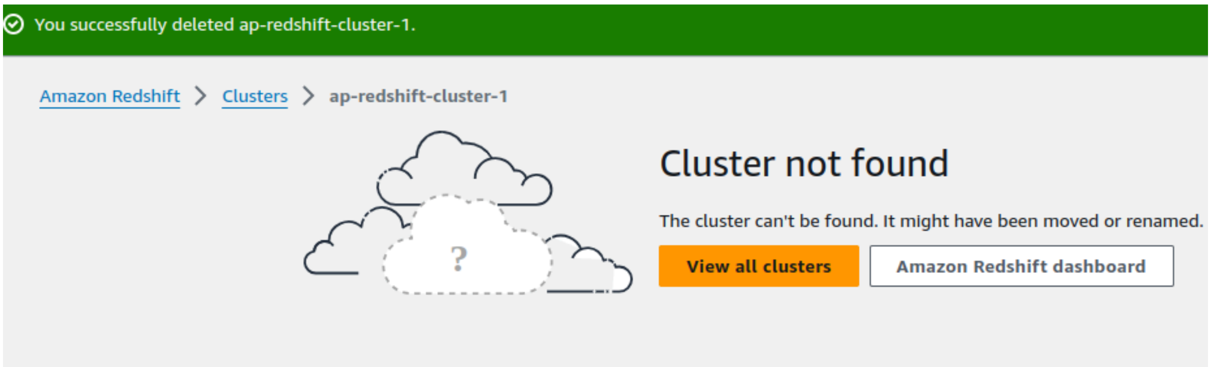
Final snapshot

You can create a final manual snapshot of your cluster before it's deleted so you can later restore it. Restoring it enables you to resume running the cluster and querying data.

□ Create final snapshot

To confirm deletion, enter *delete* in the field and choose Delete.

delete



You successfully deleted ap-redshift-cluster-1.

Amazon Redshift > Clusters > ap-redshift-cluster-1

Cluster not found

The cluster can't be found. It might have been moved or renamed.