

AWS Redshift

Introduction

Amazon Redshift is a fully managed, petabyte-scale data warehouse service in the cloud. It enables users to analyze large datasets using standard SQL and existing Business Intelligence (BI) tools.

Key Features

1. Scalability

- **Massively Parallel Processing (MPP):** Distributes data and query load across multiple nodes.
- **Elastic Resize:** Adjusts the size of your data warehouse cluster with minimal downtime.

2. Performance

- **Columnar Storage:** Stores data in a columnar format, optimizing I/O and reducing storage costs.
- **Data Compression:** Automatically compresses data to save storage and improve query performance.
- **Query Optimization:** Advanced query optimization techniques enhance performance.

3. Security

- **Data Encryption:** Supports encryption at rest and in transit using AWS Key Management Service (KMS).
- **Network Isolation:** Integration with Amazon VPC provides enhanced security through network isolation.
- **Access Control:** Role-based access control allows fine-grained permissions for users and groups.

4. Cost-Effective

- **Pay-as-you-go Pricing:** Only pay for what you use with no upfront costs.
- **Reserved Instances:** Option to reserve instances for a lower hourly rate.

5. Integration with AWS Ecosystem

- **Seamless Integration:** Works with AWS services like S3, DynamoDB, and AWS Glue for data ingestion and ETL processes.
- **Data Lake Integration:** Easily integrates with data lakes for advanced analytics.

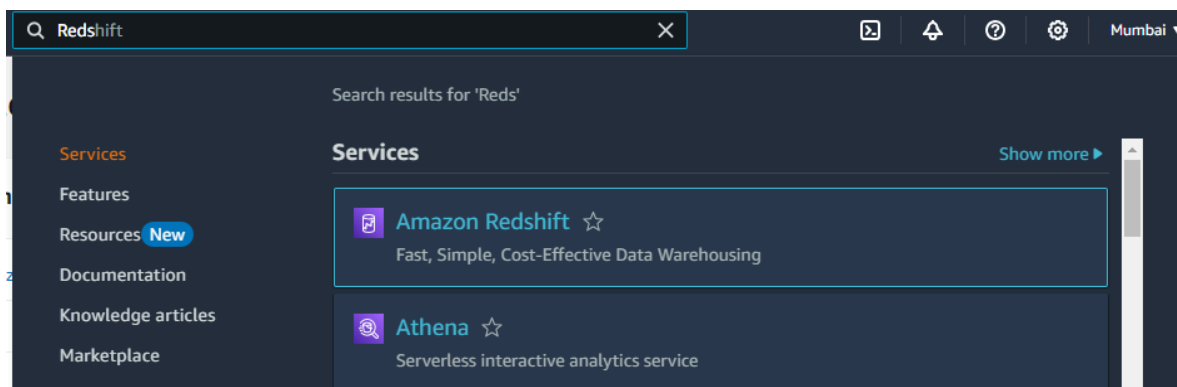
Use Cases

- **Business Intelligence:** Run complex queries and analytics to drive business decisions.
- **Data Warehousing:** Consolidate data from multiple sources for reporting and analysis.
- **ETL Processes:** Efficiently extract, transform, and load large datasets.

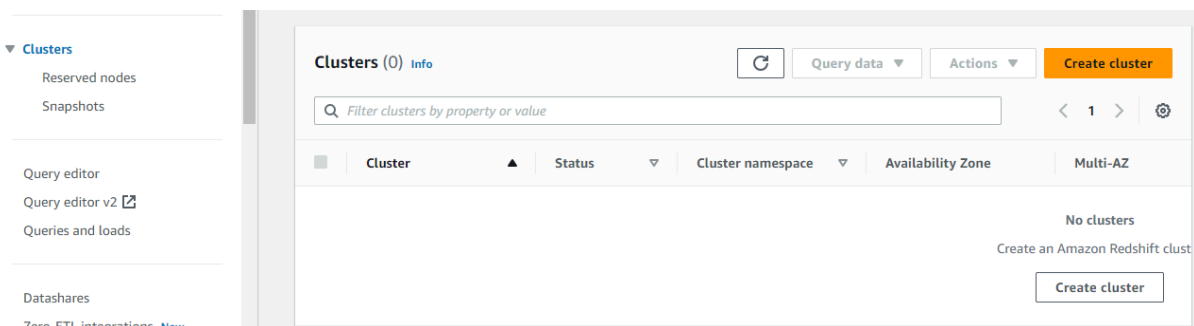
AWS Redshift Cluster

1. Create an Amazon Redshift Cluster

1. **Log in to AWS Management Console:**
 - Go to the [AWS Management Console](#).
2. **Navigate to Amazon Redshift:**
 - Search for and select **Amazon Redshift** from the services menu.



3. **Create a Cluster:**
 - Click **Create cluster**.



- **Cluster Identifier:** Choose a unique name (e.g., `deven-redshift-cluster-1`).

Amazon Redshift > Clusters > Create cluster

Create cluster [Info](#)

Cluster configuration

Cluster identifier
This is the unique key that identifies a cluster.

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:** Select the appropriate node type for your workload (e.g., `dc2.large`), and Start with one node for practice.

Node type [Info](#)
Choose a node type that meets your CPU, RAM, storage capacity, and drive type requirements.

Number of nodes
Enter the number of nodes that you need.

Range (1-32)

Configuration summary [Info](#)
dc2.large | 1 node

<p>\$229.95/month</p> <p>Estimated on-demand compute price</p> <p>Save more than 60% of your costs by purchasing reserved nodes.</p> <p>Learn more about pricing ↗</p>	<p>160 GB</p> <p>Total compressed storage</p> <p>The total storage capacity for the cluster if you deploy the number of nodes that you chose.</p>
---	--

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.

- **Master Username:** Choose an admin username and **Master Password:** Set a password and confirm it.

Database configurations

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

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

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

4. Configure Security Settings:

- Choose or create a new VPC and configure security groups to allow access to the cluster (e.g., open port 5439 for PostgreSQL).

Associated IAM roles (0) [Info](#) Set default Manage IAM roles

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 for this cluster.

☐ IAM roles No resources No associated IAM roles Associate IAM role

Associate IAM roles ×

IAM roles
Choose from existing IAM roles. You can associate up to 50 IAM roles with this cluster.

 1 match

☒ IAM roles deven_redshift_s3_access

Cancel Associate IAM roles

5. **Launch the Cluster:**

- Review your settings and click **Create cluster**. Wait for the cluster status to change to **Available**.

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-07cf6b1448b22f0c5)** and **default** subnet.

Security

Using **default (sg-07427a59ceee2774b)** 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.

Cancel

Create cluster

Clusters (1) Info

Query data

Actions

Create cluster

Filter clusters by property or value

< 1 >

<input type="checkbox"/>	Cluster	Status	Cluster namespace	Availability Zone	Multi-AZ
<input type="checkbox"/>	deven-redshift-cluster-1 dc2.large 1 node 160 GB	Available	17f29f3a-b541-40a3-...	ap-south-1b	No

- Click on **Query data**

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

deven-redshift-cluster-1

Actions

Edit

Add partner integration

Query data

General information Info

Cluster identifier

deven-redshift-cluster-1

Status

Available

Node type

dc2.large

Endpoint

deven-redshift-cluster-1.cp6cnywaoe5m.ap-south-1.redshift.amazonaws.com:5439/

dev

Custom domain name

-

Date created

September 16, 2024, 17:19 (UTC+05:30)

Number of nodes

1

Cluster namespace ARN

arn:aws:redshift:ap-south-1:980921736064:namespace:17f29f3a-b541-40a3-9ca2-b1fdc0d8454

Storage used

-

Patch version

Patch 184

JDBC URL

jdbc:redshift://deven-redshift-cluster-1.cp6cnywaoe5m.ap-south-1.redshift.amazonaws.com:5439/

dev

Cluster configuration

Production

Multi-AZ

No

ODBC URL

Driver={Amazon Redshift (x64)}; Server=deven-redshift-

- In Editor click connect to database and make connection with through your created master username and password

[Amazon Redshift](#) > Query editor

Editor

Query history

Saved queries

Scheduled queries

Connect to a database to run queries and view results.

Resources Info

⌂

⌂

⌂

Select schema

Select a schema to view data tables.

Status - database - user -

Connect to database

Query 1 +

1

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

deven-redshift-cluster-1 (Available)

Database name

dev

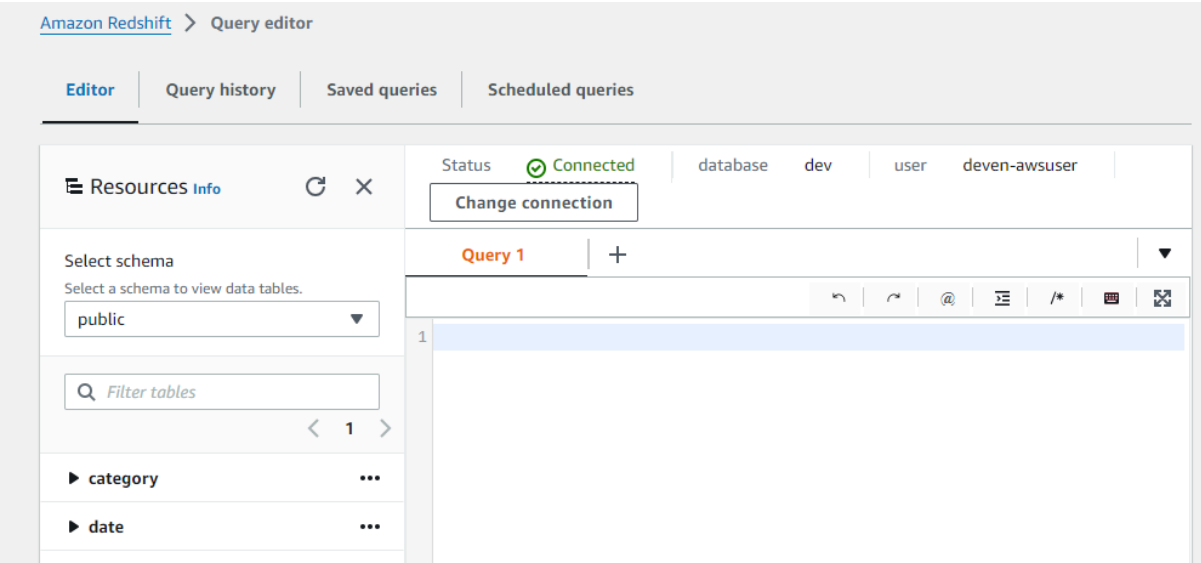
Database user

User name authorized to access your database.

deven-awsuser

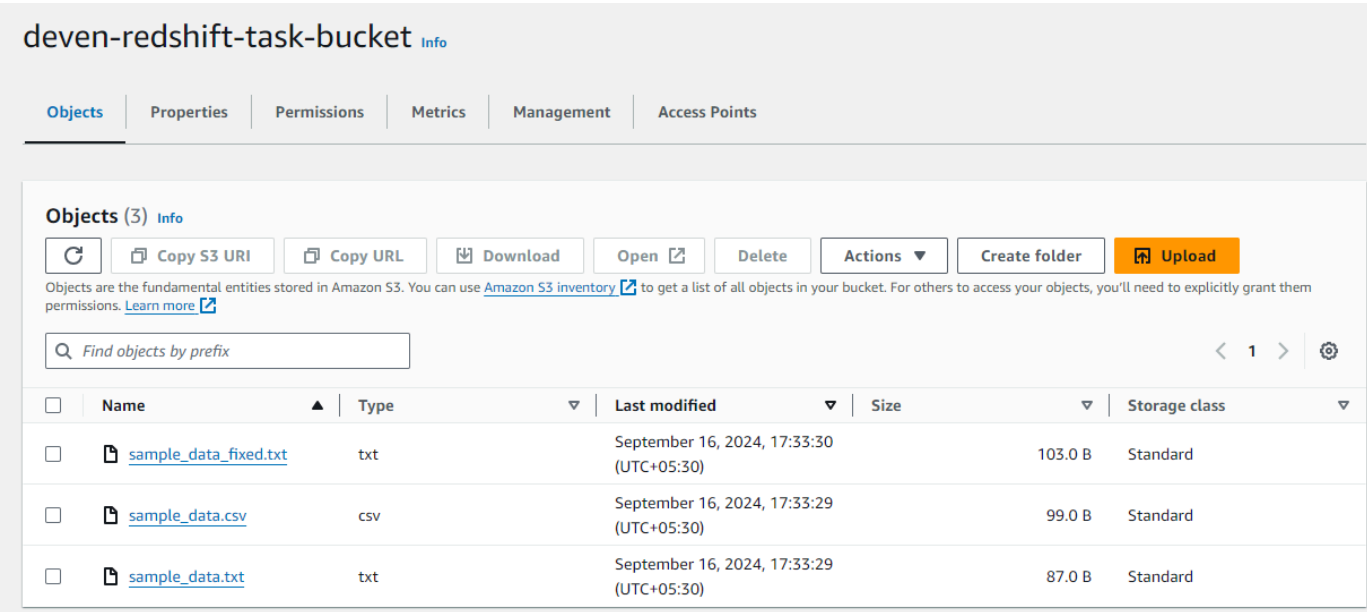
Cancel

Connect



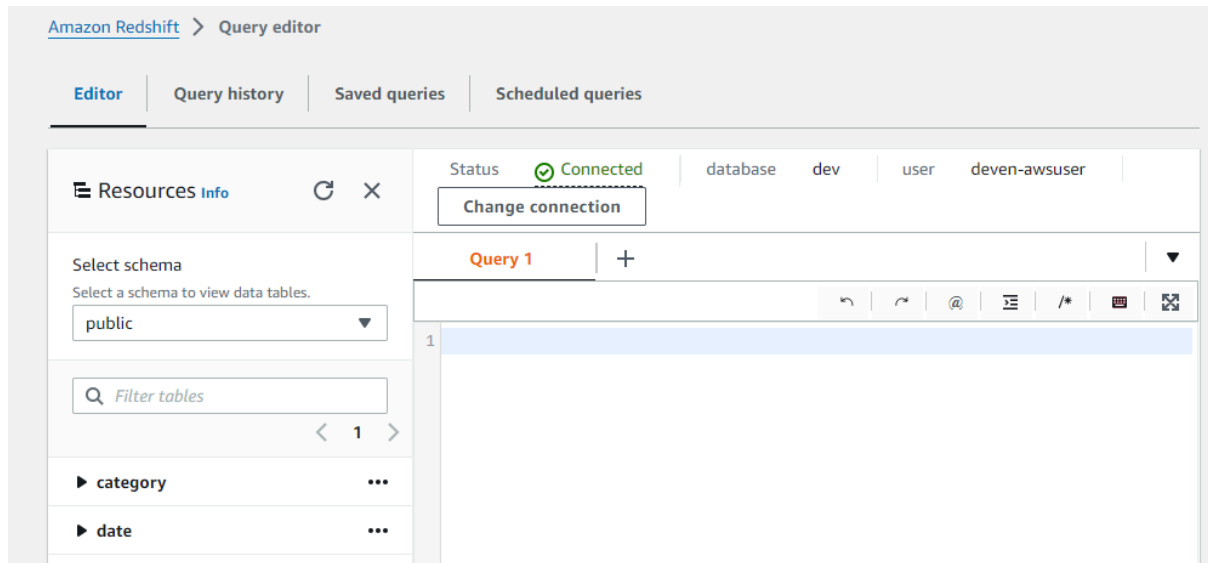
2. Upload Data Files to S3

- **CSV File:** `sample_data.csv`
- **Character-Delimited File:** `sample_data.txt` (using a character like `|` for delimiting)
- **Fixed Width File:** `sample_data_fixed.txt` (columns have fixed widths)



4. Create Tables in the Redshift Database

1. Run Query in Redshift Editor:



Create Tables: Write SQL commands to create tables in your Redshift database. For example:

code

-- For CSV file

```
CREATE TABLE csv_table (  
    id INT,  
    name VARCHAR(100),  
    age INT  
);
```

-- For character-delimited file (using | as delimiter)

```
CREATE TABLE char_delimited_table (  
    id INT,  
    name VARCHAR(100),  
    age INT  
);
```

-- For fixed width file (assuming fixed width for each column)

```
CREATE TABLE fixed_width_table (  
    id INT,  
    name VARCHAR(100),  
    age INT  
);
```



```

1 CREATE TABLE csv_table (
2     id INT,
3     name VARCHAR(100),
4     age INT
5 );
6 CREATE TABLE char_delimited_table (
7     id INT,
8     name VARCHAR(100),
9     age INT
10 );
11 CREATE TABLE fixed_width_table (
12     id INT,
13     name VARCHAR(100),
14     age INT
15 );
16

```

Result:

<input type="text" value="table"/>		2	id I
		3	name
		4	age
3 tables		5);
		6	CREATE T
▶ char_delimited_table	...	7	id I
▼ csv_table	...	8	name
id		9	age
name		10);
age		11	CREATE T
▶ fixed_width_table	...	12	id I
		13	name
		14	age
		15);
		16	

5. Transfer Data with COPY Commands

Write COPY Commands:

For CSV:

code

```
COPY csv_table
```

```
FROM 's3://deven-redshift-task-bucket/sample_data.csv'
```

```
IAM_ROLE
```

```
'arn:aws:iam::your-account-id:role/deven_redshift_s3_access'
```

```
CSV;
```

For **Character-Delimited** (e.g., using |):

code

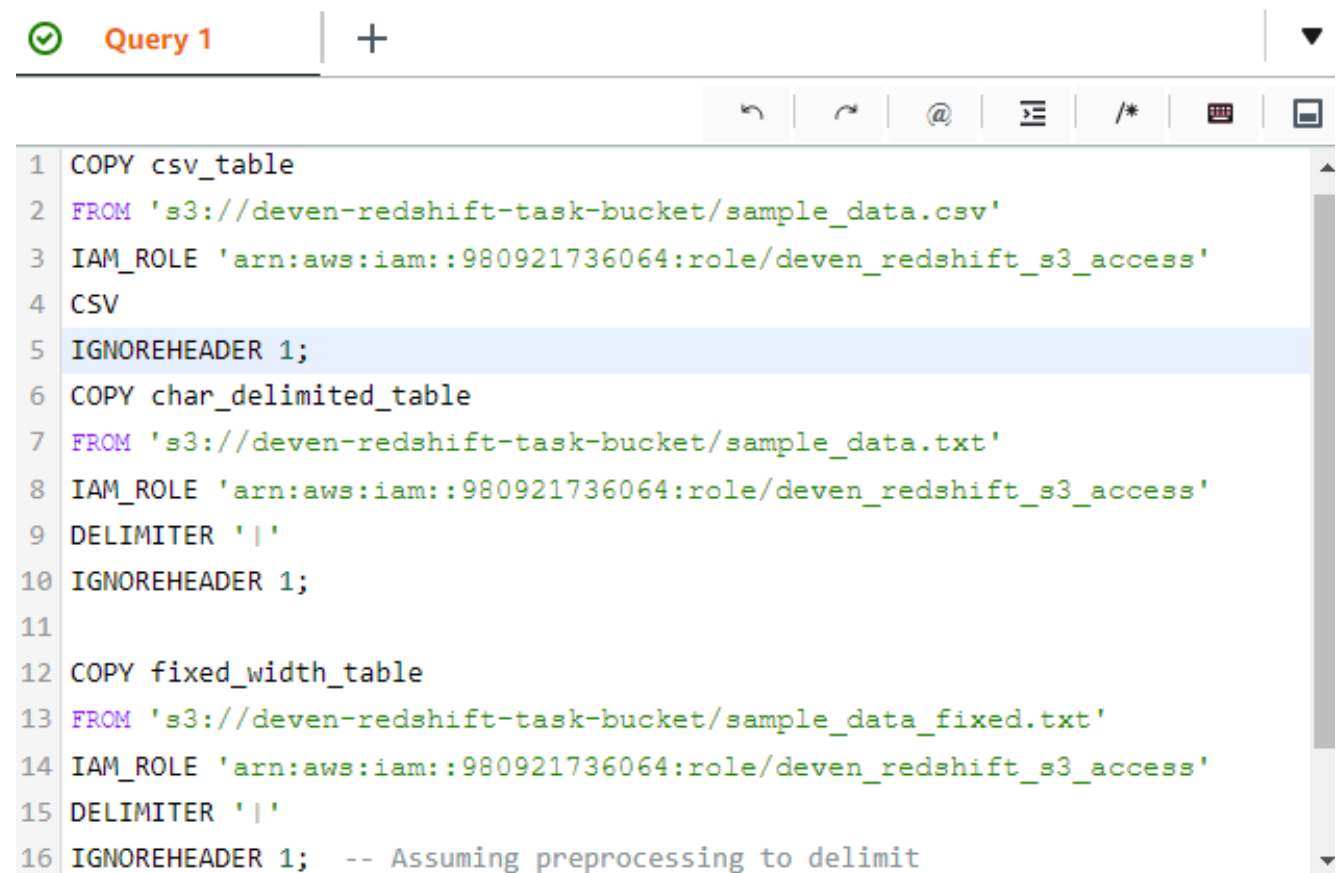
```
COPY char_delimited_table
FROM 's3://deven-redshift-task-bucket/sample_data.txt'
IAM_ROLE
'arn:aws:iam::your-account-id:role/deven_redshift_s3_access'
DELIMITER '|';
```

For **Fixed Width** (Redshift doesn't support fixed-width directly; you might need to preprocess the data into a delimited format before loading):

code

```
COPY fixed_width_table
FROM 's3://deven-redshift-task-bucket/sample_data_fixed.txt'
IAM_ROLE
'arn:aws:iam::your-account-id:role/deven_redshift_s3_access'
FIXEDWIDTH '1,16,2';
```

Note- Above is old method for fixed-width but still it is running.....



The screenshot shows the AWS Redshift Query Editor interface. At the top, there is a tab labeled 'Query 1' with a green checkmark icon. Below the tab is a toolbar with icons for undo, redo, search, and other editor functions. The main area displays a SQL query with line numbers 1 through 16. The query is as follows:

```
1 COPY csv_table
2 FROM 's3://deven-redshift-task-bucket/sample_data.csv'
3 IAM_ROLE 'arn:aws:iam::980921736064:role/deven_redshift_s3_access'
4 CSV
5 IGNOREHEADER 1;
6 COPY char_delimited_table
7 FROM 's3://deven-redshift-task-bucket/sample_data.txt'
8 IAM_ROLE 'arn:aws:iam::980921736064:role/deven_redshift_s3_access'
9 DELIMITER '|'
10 IGNOREHEADER 1;
11
12 COPY fixed_width_table
13 FROM 's3://deven-redshift-task-bucket/sample_data_fixed.txt'
14 IAM_ROLE 'arn:aws:iam::980921736064:role/deven_redshift_s3_access'
15 DELIMITER '|'
16 IGNOREHEADER 1; -- Assuming preprocessing to delimit
```

6. Analyze the Database


- 1. **Run Queries:**
 - Use SQL queries to analyze the data. For example:

```
code
-- Sample query to check data
SELECT * FROM csv_table;
```

<

1

>



id	name	age
1	John Doe	28
2	Jane Smith	34
3	Bob Johnson	45
4	Emily Davis	29
5	Michael Brown	38

```
-- Aggregate data
SELECT AVG(age) AS average_age FROM csv_table;
```

Rows returned (1)

Export ▼

Q

Search rows

< 1 >

⚙

average_age

▼

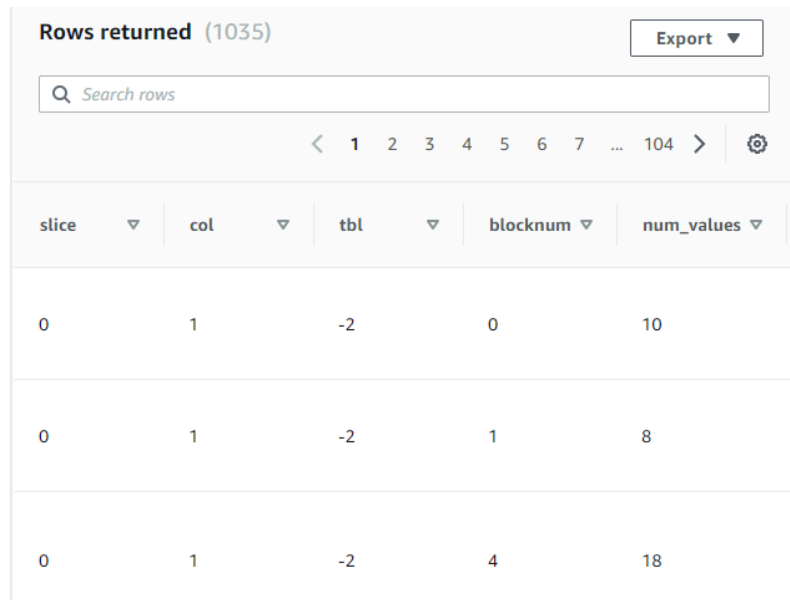
34

2. Use Redshift Performance Tools:

Review the performance using system tables and views:

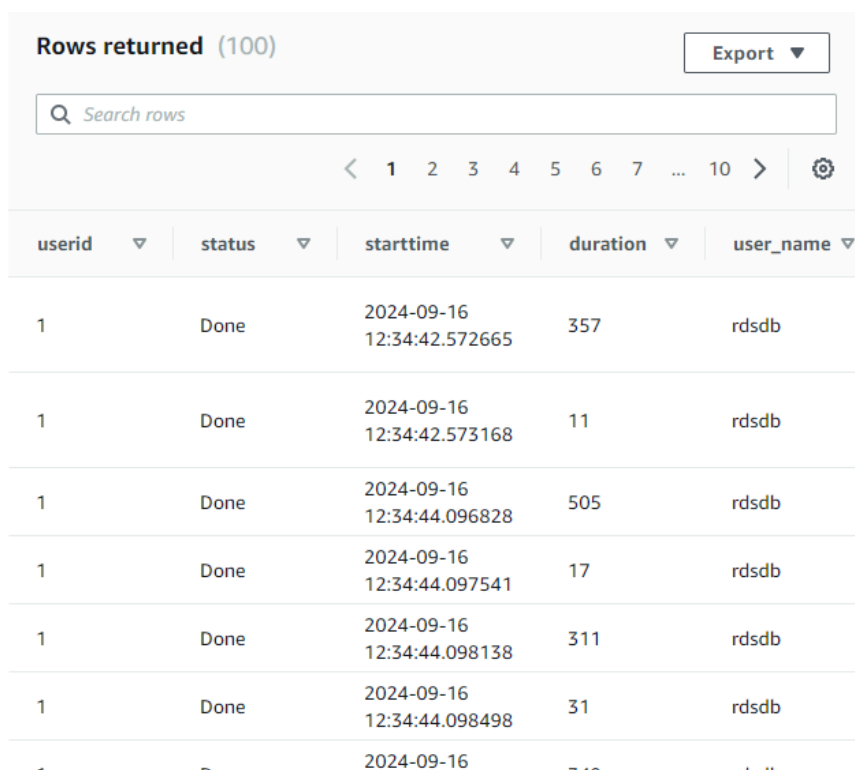
code

```
SELECT * FROM stv_blocklist; -- Check block-level data distribution
```



slice	col	tbl	blocknum	num_values
0	1	-2	0	10
0	1	-2	1	8
0	1	-2	4	18

```
SELECT * FROM stv_recents; -- Recent queries and their performance
```



userid	status	starttime	duration	user_name
1	Done	2024-09-16 12:34:42.572665	357	rdsdb
1	Done	2024-09-16 12:34:42.573168	11	rdsdb
1	Done	2024-09-16 12:34:44.096828	505	rdsdb
1	Done	2024-09-16 12:34:44.097541	17	rdsdb
1	Done	2024-09-16 12:34:44.098138	311	rdsdb
1	Done	2024-09-16 12:34:44.098498	31	rdsdb

3. Visualize Data:

- Optionally, you can use Amazon QuickSight or other BI tools to create dashboards and visualizations from your Redshift data.