

REDSHIFT DATA PIPELINE: FROM GENERATION TO DATABASE

INTRODUCTION

The "Redshift Data Pipeline: From Generation to Database" project showcases the end-to-end process of creating and populating a database on Amazon Redshift.

This documentation provides insights into the setup of a cloud environment, including the creation of a VPC, subnets, S3 bucket, IAM roles, security groups, and Redshift cluster. It covers the generation of sample data, formatting it, and loading it into the Redshift database. Additionally, for those interested in replicating or examining the infrastructure setup, I offer an Infrastructure as Code (IaC) option click [here](#). This project demonstrates the implementation of a robust data pipeline leveraging AWS services to efficiently manage and analyze datasets.

This demonstration walks through getting started with Amazon Redshift cluster and shows how to both load and query your data.

PREPARING THE ENVIRONMENT

A). Create a VPC.

The first step in configuring your environment is to create a virtual private cloud (VPC) to hold the resources for both your Amazon Elastic Compute Cloud (Amazon EC2) instance and Amazon Redshift database. (*The N.Virginia us-east-1 Region was used in this demonstration*).

In this project, the following settings were used:

- *Name*: Amazon Redshift Project
- *Location*: us-east-1
- *IPv4 CIDR block*: 10.60.0.0/16
- *Internet gateway*: Redshift IGW
- *Public route table*: Redshift Project Public
- *Private route table*: Redshift Project Private

aws

Services

Search

[Alt+S]

VPC

>

Your VPCs

>

Create VPC

Create VPC [Info](#)

A VPC is an isolated portion of the AWS Cloud populated by AWS objects, such as Amazon EC2 instances.

VPC settings

Resources to create [Info](#)

Create only the VPC resource or the VPC and other networking resources.

☒ VPC only

☐ VPC and more

Name tag - optional

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

Amazon Redshift Project

IPv4 CIDR block [Info](#)

☒ IPv4 CIDR manual input

☐ IPAM-allocated IPv4 CIDR block

IPv4 CIDR

10.60.0.0/16

CIDR block size must be between /16 and /28.

IPv6 CIDR block [Info](#)

☒ No IPv6 CIDR block

☐ IPAM-allocated IPv6 CIDR block

☐ Amazon-provided IPv6 CIDR block

☐ IPv6 CIDR owned by me

Tenancy [Info](#)

Default

Tags

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

Key

Value - optional

Q Name

X

Q Amazon Redshift Project

X

Remove tag

Add tag

You can add 49 more tags

Cancel

Create VPC

Figure 1. VPC Creation - After review click the 'create VPC' button.

aws Services Search [Alt+S]

VPC > Internet gateways > Create internet gateway

Create internet gateway [Info](#)

An internet gateway is a virtual router that connects a VPC to the internet. To create a new internet gateway specify the name for the gateway below.

Internet gateway settings

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

Tags - optional

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

Key	Value - optional	
<input type="text" value="Name"/>	<input type="text" value="Redshift IGW"/>	<input type="button" value="Remove"/>

You can add 49 more tags.

Figure 2. Internet Gateway creation.

aws Services Search [Alt+S]

VPC > Route tables > Create route table

Create route table [Info](#)

A route table specifies how packets are forwarded between the subnets within your VPC, the internet, and your VPN connection.

Route table settings

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

VPC
The VPC to use for this route table.

Tags

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

Key	Value - optional	
<input type="text" value="Name"/>	<input type="text" value="Redshift Project Public"/>	<input type="button" value="Remove"/>

You can add 49 more tags.

Figure 3. Public Route Table, click on the 'create route' table button.

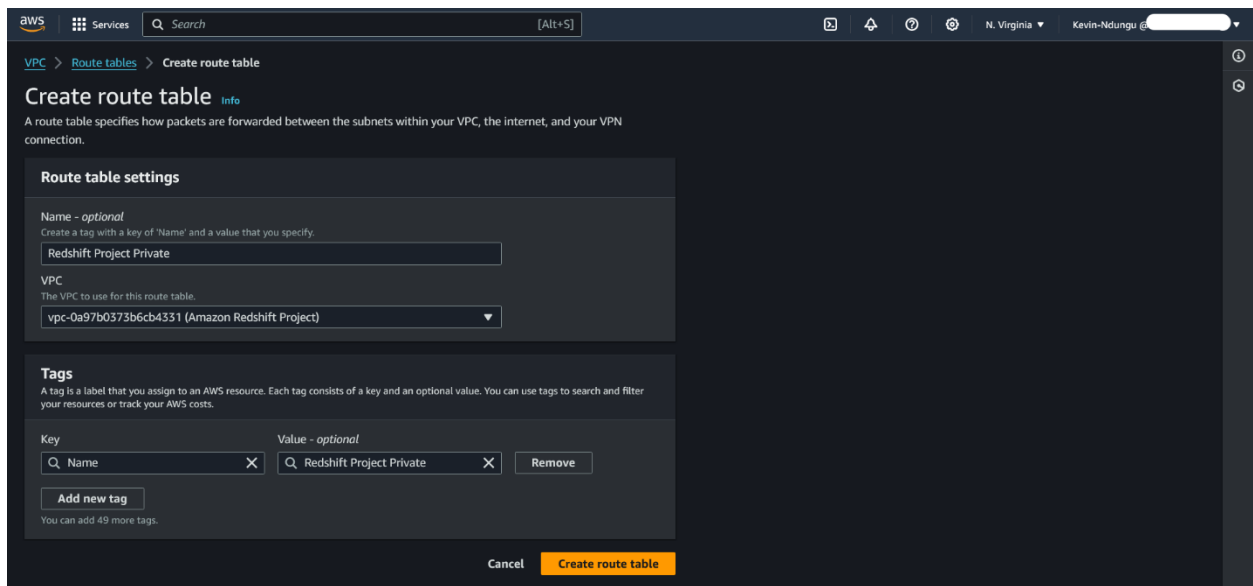


Figure 4. Private Route Table, click on the 'create route table' button.

B). Create the subnets.

Amazon Redshift is a managed service that connects to subnets in your VPC.

In this project, the following settings were used:

- **Private subnet 1**
 - *Name:* Redshift Private 01
 - *CIDR:* 10.60.101.0/24

aws Services Search [Alt+S]

VPC > Subnets > Create subnet

Create subnet [Info](#)

VPC

VPC ID
Create subnets in this VPC.

vpc-0a97b0373b6cb4331 (Amazon Redshift Project) ▼

Associated VPC CIDRs

IPv4 CIDRs
10.60.0.0/16

Subnet settings

Specify the CIDR blocks and Availability Zone for the subnet.

Subnet 1 of 1

Subnet name
Create a tag with a key of 'Name' and a value that you specify.

Redshift Private 01

The name can be up to 256 characters long.

Availability Zone [Info](#)
Choose the zone in which your subnet will reside, or let Amazon choose one for you.

US East (N. Virginia) / us-east-1a ▼

IPv4 VPC CIDR block [Info](#)
Choose the IPv4 VPC CIDR block to create a subnet in.

10.60.0.0/16 ▼

IPv4 subnet CIDR block

10.60.101.0/24 256 IPs

< > ^ v

▼ Tags - optional

Key	Value - optional	
Q Name X	Q Redshift Private 01 X	Remove

Add new tag

You can add 49 more tags.

Remove

Add new subnet

Cancel Create subnet

Figure 5. Click on the 'create subnet' button.

- **Private subnet 2**
 - *Name:* Redshift Private 02
 - *CIDR:* 10.60.102.0/24

aws Services Search [Alt+S]

VPC > Subnets > Create subnet

Create subnet [Info](#)

VPC

VPC ID
Create subnets in this VPC.

vpc-0a97b0373b6cb4331 (Amazon Redshift Project)

Associated VPC CIDRs

IPv4 CIDRs
10.60.0.0/16

Subnet settings

Specify the CIDR blocks and Availability Zone for the subnet.

Subnet 1 of 1

Subnet name
Create a tag with a key of 'Name' and a value that you specify.

Redshift Private 02
The name can be up to 256 characters long.

Availability Zone [Info](#)
Choose the zone in which your subnet will reside, or let Amazon choose one for you.

US East (N. Virginia) / us-east-1b

IPv4 VPC CIDR block [Info](#)
Choose the IPv4 VPC CIDR block to create a subnet in.

10.60.0.0/16

IPv4 subnet CIDR block

10.60.102.0/24 256 IPs

< > ^ v

▼ Tags - optional

Key	Value - optional	
Q Name	Q Redshift Private 02	Remove

Add new tag
You can add 49 more tags.

Remove

Add new subnet

Cancel Create subnet

Figure 6. Click on the 'create subnet' button.

C). Create an Amazon S3 bucket.

The data used for this demo is randomly generated and uploaded to an Amazon Simple Storage Service (Amazon S3) bucket.

In this project, the following settings were used:

- *Region:* us-east-1
- *Bucket:* redshift-database-demo (Make sure to use a unique bucket name)

The screenshot shows the AWS Management Console interface for creating a new S3 bucket. The breadcrumb navigation at the top indicates the path: Amazon S3 > Buckets > Create bucket. The main heading is 'Create bucket' with an 'Info' link. A sub-header states: 'Buckets are containers for data stored in S3. [Learn more](#)'. The 'General configuration' section includes a dropdown for 'AWS Region' set to 'US East (N. Virginia) us-east-1'. Under 'Bucket type', 'General purpose' is selected, with a description: 'Recommended for most use cases and access patterns. General purpose buckets are the original S3 bucket type. They allow a mix of storage classes that redundantly store objects across multiple Availability Zones.' The 'Directory - New' option is also visible. The 'Bucket name' field contains 'redshift-database-demo-254', with a note: 'Bucket name must be unique within the global namespace and follow the bucket naming rules. [See rules for bucket naming](#)'. Below this, a section titled 'Copy settings from existing bucket - optional' includes a 'Choose bucket' button and the format 's3://bucket/prefix'. The 'Object Ownership' section shows 'ACLs disabled (recommended)' selected, with a description: 'All objects in this bucket are owned by this account. Access to this bucket and its objects is specified using only policies.' The 'Block Public Access settings for this bucket' section has 'Block all public access' checked, with a note: 'Turning this setting on is the same as turning on all four settings below. Each of the following settings are independent of one another.'

Figure 7.

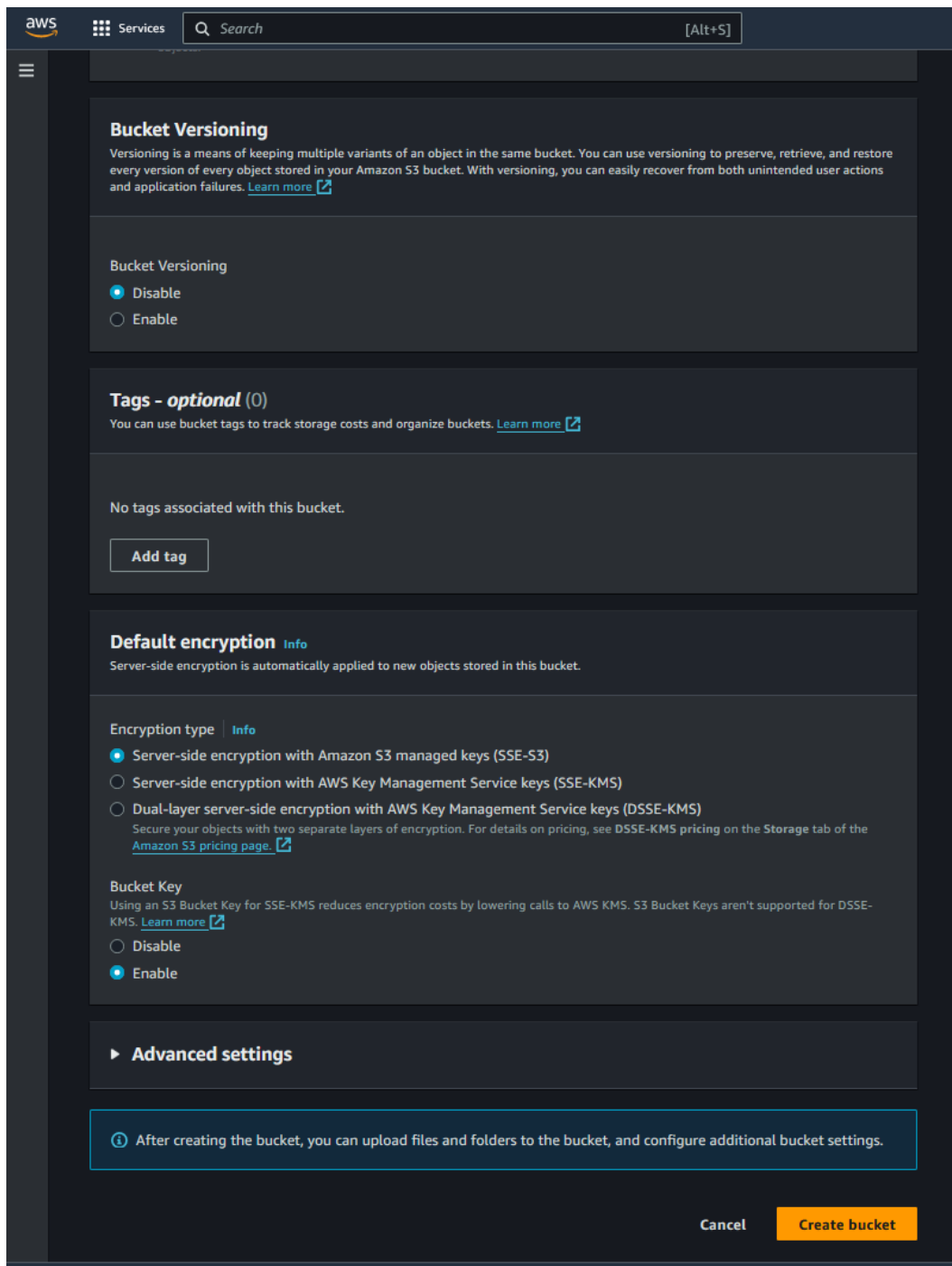


Figure 8.

D). Create an Amazon S3 endpoint.

By default, traffic between an Amazon Redshift cluster and Amazon S3 traverses the public AWS network. A VPC endpoint forces all COPY and UNLOAD traffic between your cluster and your data on Amazon S3 to stay in your VPC.

In this project, the following settings were used:

- *Service name:* com.amazonaws.us-east-1.s3
- *Endpoint type:* Gateway
- *Route table:* Redshift Project Private

Create endpoint

There are three types of VPC endpoints – Interface endpoints, Gateway Load Balancer endpoints, and Gateway endpoints. Interface endpoints and Gateway Load Balancer endpoints are powered by AWS PrivateLink, and use an Elastic Network Interface (ENI) as an entry point for traffic destined to the service. Interface endpoints are typically accessed using the public or private DNS name associated with the service, while Gateway endpoints and Gateway Load Balancer endpoints serve as a target for a route in your route table for traffic destined for the service.

Endpoint settings

Name tag: optional
Created a tag with a key of 'Name' and a value that you specify.
com.amazonaws.us-east-1.s3

Service category
Select the service category.

☒ AWS services
Services provided by Amazon

☐ PrivateLink Ready partner services
Services with an AWS Service Ready designation

☐ AWS Marketplace services
Services that you've purchased through AWS Marketplace

☐ EC2 Instance Connect Endpoint
An elastic network interface that allow you to connect to resources in a private subnet

☐ Other endpoint services
Find services shared with you by service name

Services (1/3)

Search

Type: Gateway X Clear filters

Service Name	Owner	Type
com.amazonaws.us-east-1.dynamodb	amazon	Gateway
com.amazonaws.us-east-1.s3	amazon	Gateway
com.amazonaws.us-east-1.s3express	amazon	Gateway

VPC
Select the VPC in which to create the endpoint.

VPC
The VPC in which to create your endpoint.
vpc-0a97b0373b6cb4331 (Amazon Redshift Project)

Figure 9. Part A.

Route tables (1/3)

Search

Name	Route Table ID	Main	Associated ID
Redshift Project Public	rtb-0b70d0c9472f18024	Yes	2 subnets
Redshift Project Private	rtb-0711614f40f5c33 Redshift Proj	No	-
Redshift Project Private	rtb-0711614f40f5c33 Redshift Proj	No	-

When you use an endpoint, the source IP addresses from your instances in your affected subnets for accessing the AWS service is 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.

Policy info
VPC endpoint policy controls access to the service.

☒ Full access
Provides access to any port or service within the VPC using IP addresses from any Amazon Web Services accounts for any resources in the Amazon Web Services service. All policies – AWS user policies, VPC endpoint policies, and Amazon Web Services service specific policies (e.g. Amazon S3 bucket policies, any IAM policy) – must grant the necessary permissions for access to succeed.

☐ Custom
Use the [policy generator](#) to generate a policy, then paste the generated policy below.

Tags

Key: Name Value: optional
com.amazonaws.us-east-1.s3 Remove

Add new tag
You can add 40 tags.

Cancel Create endpoint

Figure 10. Click the Create endpoint button.

NB: The Amazon Redshift cluster must be in the same Region as the S3 bucket.

E). Create an IAM role.

Amazon Redshift needs permission to copy data to and from S3 buckets. This access is granted using AWS Identity and Access Management (IAM) roles. A role defines a set of permissions for making AWS service requests.

In this project, the following settings were used:

- *Role name:* RedshiftDemoProject
- *Policy name:* RedshiftDemoProject

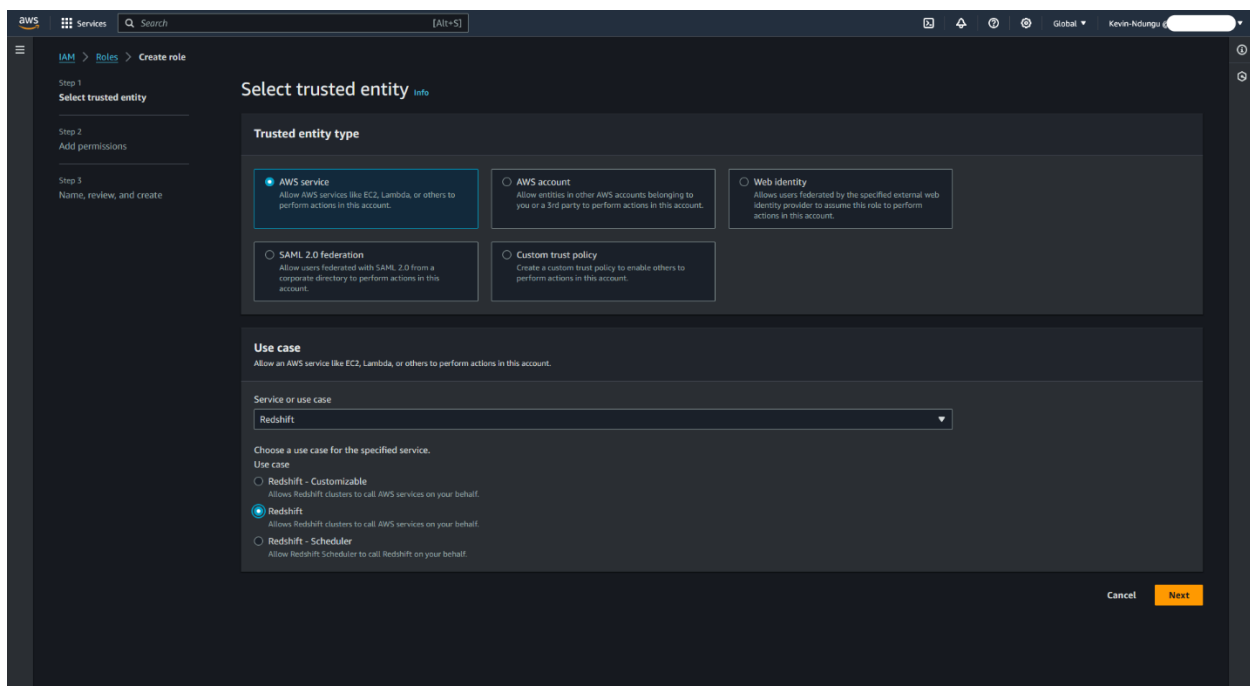


Figure 11.

In this example, the role and policy names are the same. This was for ease of use. They can be different.

The policy gives Amazon Redshift full access to a specific S3 bucket, redshift-database-demo, and the objects inside it.

It also has a trust relationship with `redshift.amazonaws.com`. The trust relationship defines who/what can assume a role. Because of this trust relationship, only Amazon Redshift can assume this role, and it has to be explicitly assigned to the cluster.

F). Create the security groups.

Security groups control access to EC2 instances and Amazon Redshift clusters. In this project, two security groups were created: one is for an EC2 instance and the other for the Amazon Redshift cluster.

In this project, the following settings were used:

- **Group 1**
 - *Name:* Redshift EC2
 - *Description:* EC2 access for Redshift Project
 - *Rule:* Allow SSH (port 22) from 0.0.0.0/0 (the internet)

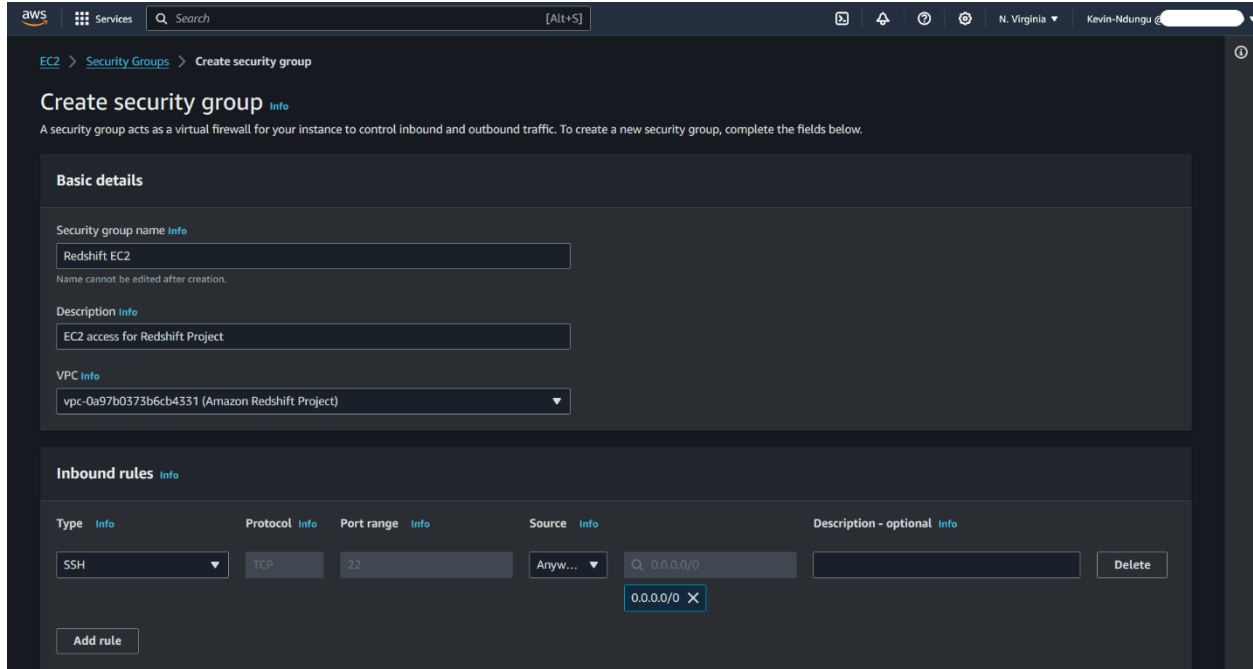


Figure 12.

Outbound rules [Info](#)

Type	Protocol	Port range	Destination	Description - optional
All traffic	All	All	Custom <input type="text" value="0.0.0.0/0"/>	<input type="text"/>

[Add rule](#) [Delete](#)

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

No tags associated with the resource.

[Add new tag](#)

You can add up to 50 more tags

[Cancel](#) [Create security group](#)

Figure 13. Click the Create security group button.

- **Group 2**
 - *Name:* Redshift Access
 - *Description:* Access to Redshift
 - *Rule:* Allow TCP port 5439 from the group "Redshift EC2"

Create security group [Info](#)

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](#)
Redshift Access
Name cannot be edited after creation.

Description [Info](#)
Access to Redshift

VPC [Info](#)
vpc-0a97b0373b6cb4331 (Amazon Redshift Project)

Inbound rules [Info](#)

Type	Protocol	Port range	Source	Description - optional
Redshift	TCP	5439	Anyw... <input type="text" value="0.0.0.0/0"/>	<input type="text"/>

[Add rule](#) [Delete](#)

Figure 14.

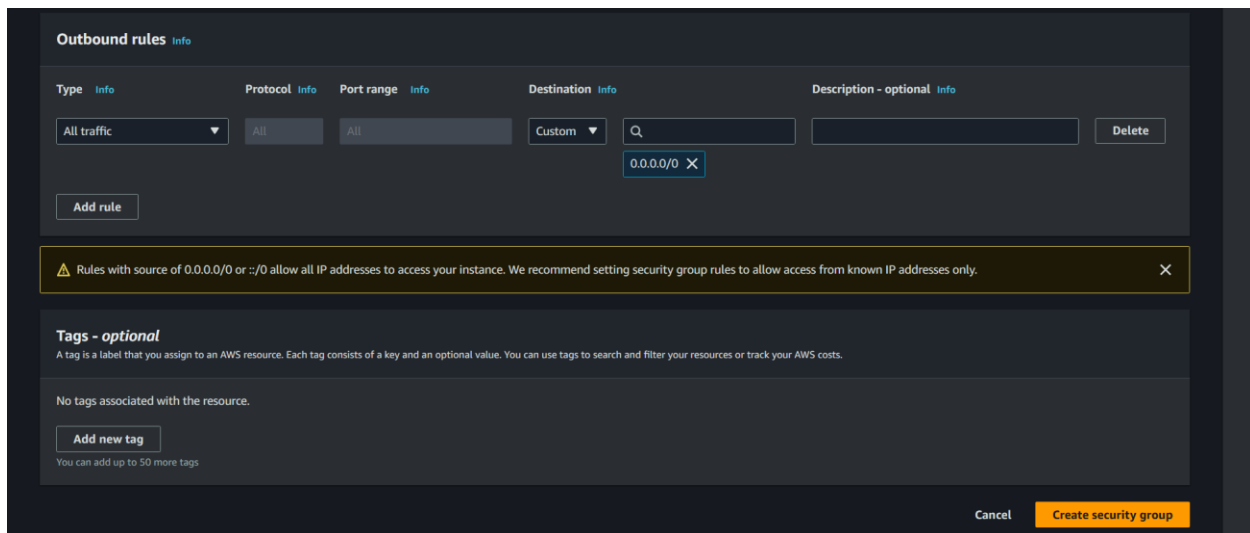


Figure 15. Click the Create security group button.

You could use a single group for both EC2 and Amazon Redshift access.

G). Create the subnet group.

It is recommended that you place your Amazon Redshift cluster in a private subnet.

In this project, the following settings were used:

- *Name:* Redshift-subnet-group
- *Description:* Private subnet access for Redshift
- *VPC:* Amazon Redshift Project
- Add all private subnets.

aws Services Search [Alt+S]

Amazon Redshift > Configurations > Subnet groups > Create subnet group

Create cluster subnet group Info

Cluster subnet group details

Name
You can't modify the name after your subnet group has been created.

Redshift-subnet-group

The name must be 1-255 characters. Valid characters are A-Z, a-z, 0-9, space, hyphen (-), underscore (_), and period (.).

Description

Private subnet access for Redshift.

Add subnets

VPC
Choose the VPC that contains the subnets that you want to include in your cluster subnet group.

vpc-0a97b0373b6cb4331

Add all the subnets for this VPC

Availability Zone Choose an Availability Zone **Subnet** Add subnet

Subnets in this cluster subnet group (2) Remove all

Availability Zone	Subnet ID	CIDR block	IPv6 CIDR block	Action
us-east-1b	subnet-0ab9d0...	10.60.102.0/24	-	Remove
us-east-1a	subnet-090950...	10.60.101.0/24	-	Remove

Cancel Create cluster subnet group

Figure 16. Click the Create cluster subnet group button.

Cluster subnet groups (1) Info Refresh Delete Actions Create cluster subnet group

Search Cluster subnet groups < 1 > ⌕

<input type="checkbox"/>	Name ▲	Status ▼	VPC ID ▼	Description
<input type="checkbox"/>	redshift-subnet-group 2 Subnets	Complete	vpc-0a97b0373b6cb4331	Private subnet access for Redshift.

Figure 17. Successful creation of the cluster subnet group.

Now at this stage we begin creation of the Amazon Redshift cluster.

[Alt+S]

Amazon Redshift > Clusters > Create cluster

Create cluster Info

Looking for free trial? Try Redshift Serverless. First-time Redshift Serverless customers receive a \$300 credit to use in their account.

Launch Redshift Serverless

Cluster configuration

Cluster identifier

This is the unique key that identifies a cluster.

myredshift-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 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.

2

Range (1-32)

Configuration summary Info

dc2.large | 2 nodes

\$365.00/month

Estimated on-demand compute price

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

[Learn more about pricing](#)

320 GB

Total compressed storage

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

Figure 18.

Choose the dc2.large node type and the number of nodes two.

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.

Database configurations

Admin user name

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

admin

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

Figure 19.

In the above diagram (Figure 19), you could opt to enable sample data generation. Configure the master's name and master password.

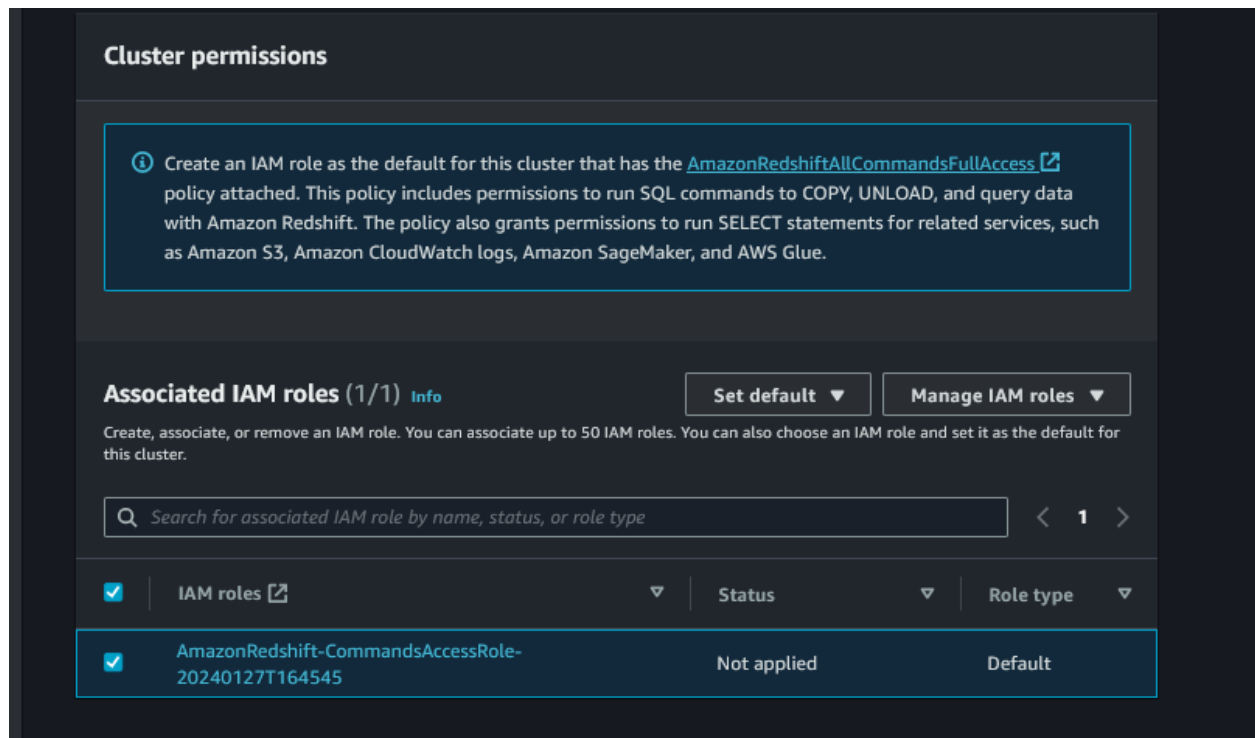


Figure 20.

Allocate the appropriate IAM role that was created earlier for Redshift's access to the S3 bucket.

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 and security** [Info](#)

Virtual private cloud (VPC)
This VPC defines the virtual networking environment for this cluster.

Amazon Redshift Project
vpc-0a97b0373b6cb4331

❗ You can't change the VPC associated with this cluster after the cluster has been created. [Learn more](#) about getting started cluster in vpc

VPC security groups
This VPC security group defines which subnets and IP ranges the cluster can use in the VPC. For more information, see [Learn more about Redshift clusters security groups](#)

Choose one or more security groups

Redshift Access
sg-035e34f420da64646

Cluster subnet group [Info](#)
Choose the Amazon Redshift subnet group to launch the cluster in.

redshift-subnet-group

Availability Zone
Specify the Availability Zone to create the cluster in. Otherwise, Amazon Redshift chooses an Availability Zone for you.

No preference

Enhanced VPC routing
Enabling this option routes network traffic between your cluster and data repositories through a VPC, instead of through the internet. [Learn more about getting started cluster in vpc](#)

☒ Turn off
☐ Turn on

Publicly accessible
For more information, see [Learn more about Redshift clusters security groups](#)

☐ Turn on Publicly accessible
Allow public connections to Amazon Redshift.

Figure 21.

Choose the VPC we created earlier (Amazon Redshift Project) and select Redshift Access as the VPC security group, deselect the default security group. Choose the redshift subnet group for the cluster and leave the rest of the option as is.

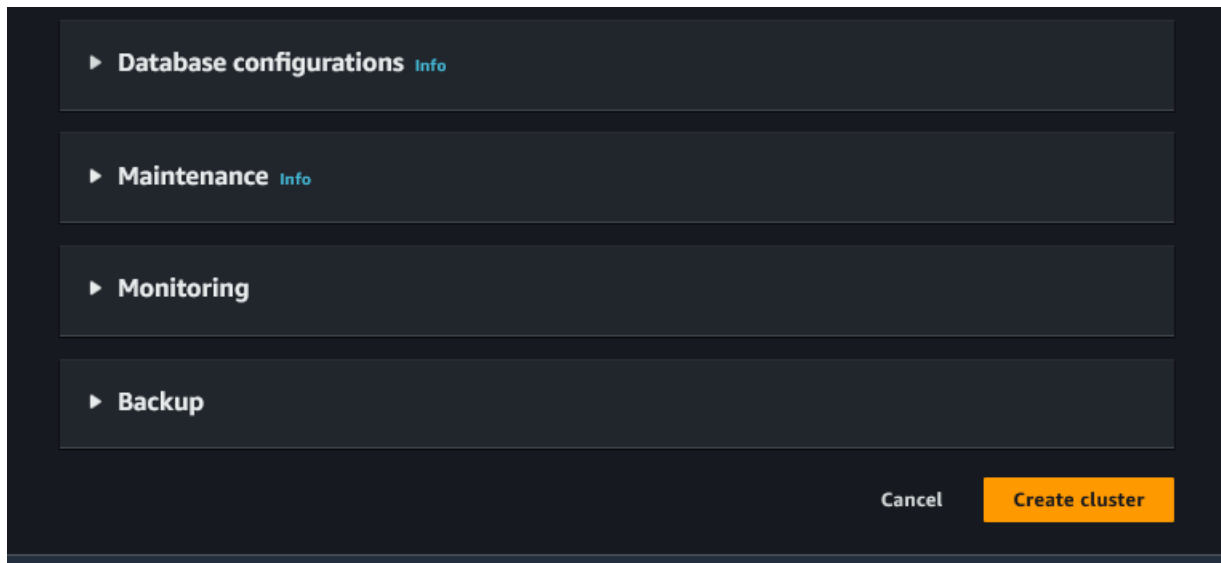


Figure 22. Leave this option in their default settings and click on the create cluster button.

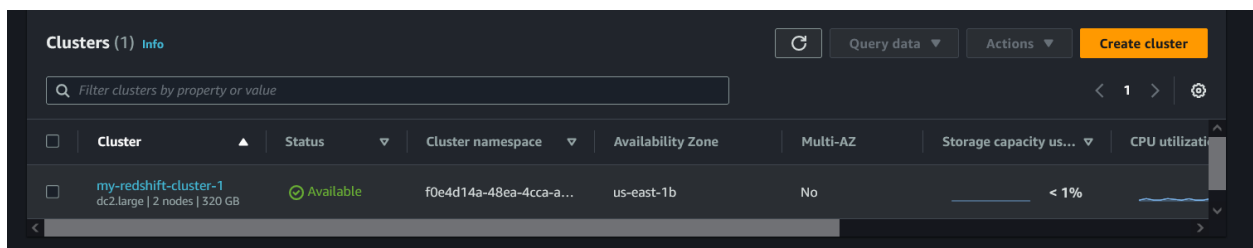


Figure 23. Wait until the cluster creation is complete and available.

INTERACTING WITH OUR NEWLY CREATED AMAZON REDSHIFT CLUSTER

On the left side of the console on the navigation panel, get to the query console by clicking on the query editor. (You can choose between the Query editor or Query editor v2).

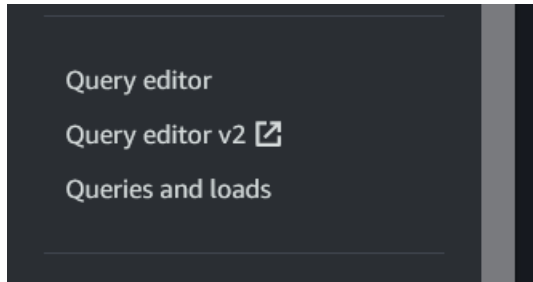


Figure 24.

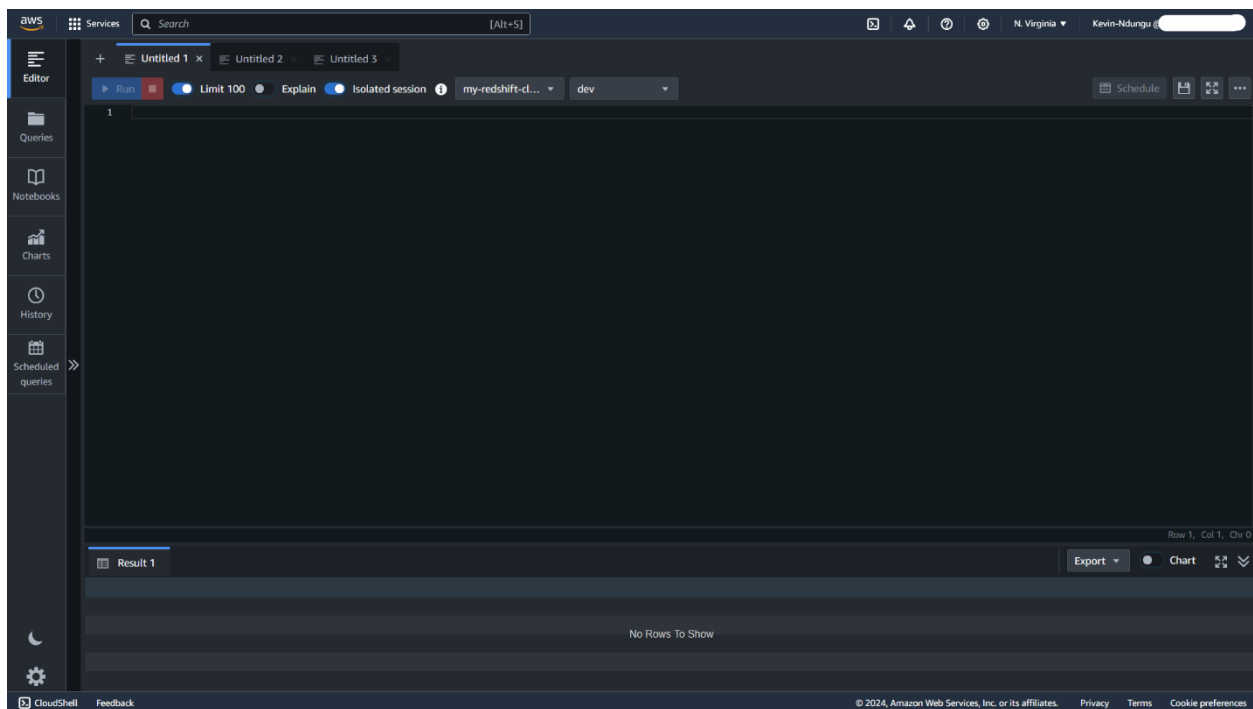


Figure 25.

Once the editor opens, the next step is to create a table (*named users*), in this case the below create table query was used:

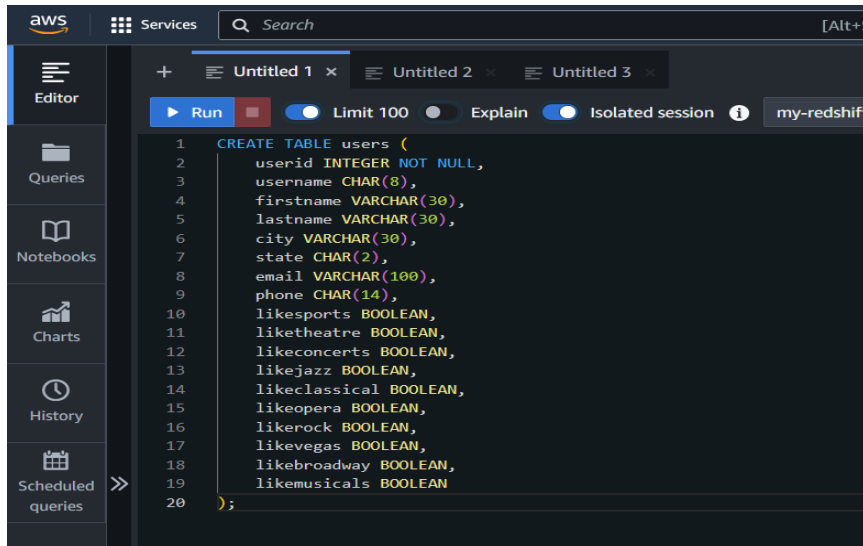


Figure 26. Click on the Run button at the top left to create table.

Once the table (*users*) is created it will be listed among other tables in the query's navigation panel as shown below:

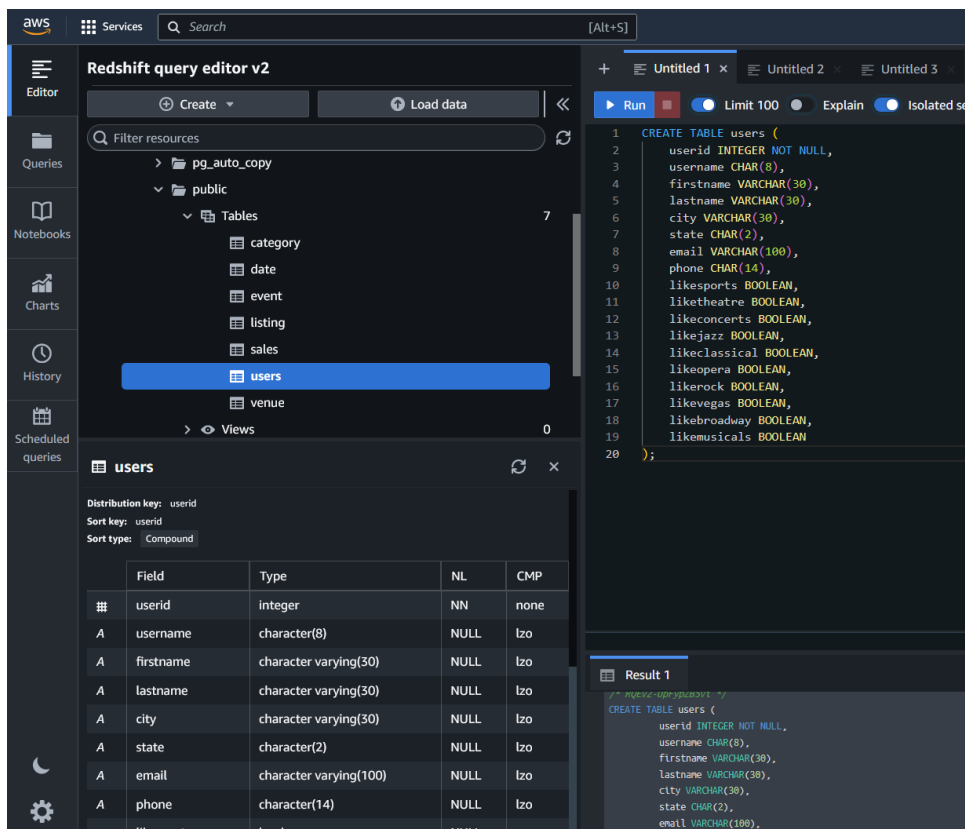


Figure 27.

Since the table is up, run the query in figure 28 was used to copy the data generated and stored in an S3 bucket to our newly created 'users' table.

So, I'm going to copy users' data from my S3 bucket. The credential that I passed is a role and it gives the ARN the Amazon resource name of the role. And in the file, itself is delimited by a comma and I ran the query.

(Note: Generate the data separately based on the information in the create table query and save it in a pipe delimited text or comma delimited format, store it in the S3 bucket created earlier).

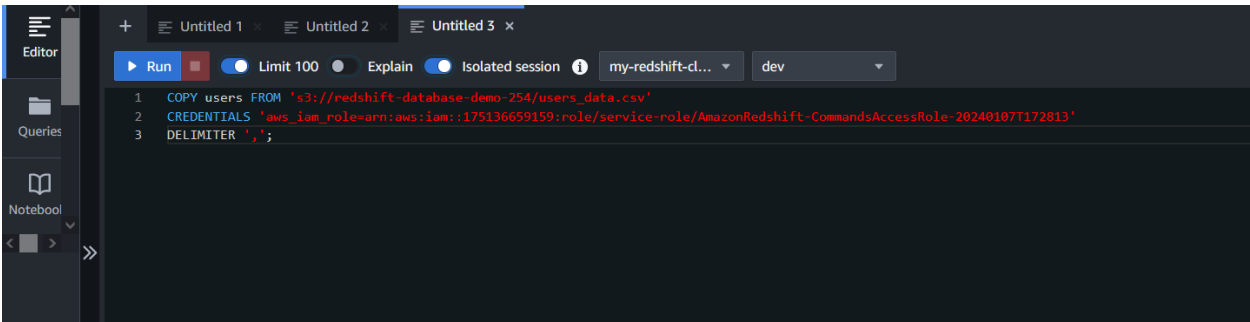


Figure 28.

Once the users table is populated, we can begin running some queries and get to view the results in the bottom display of the query editor.

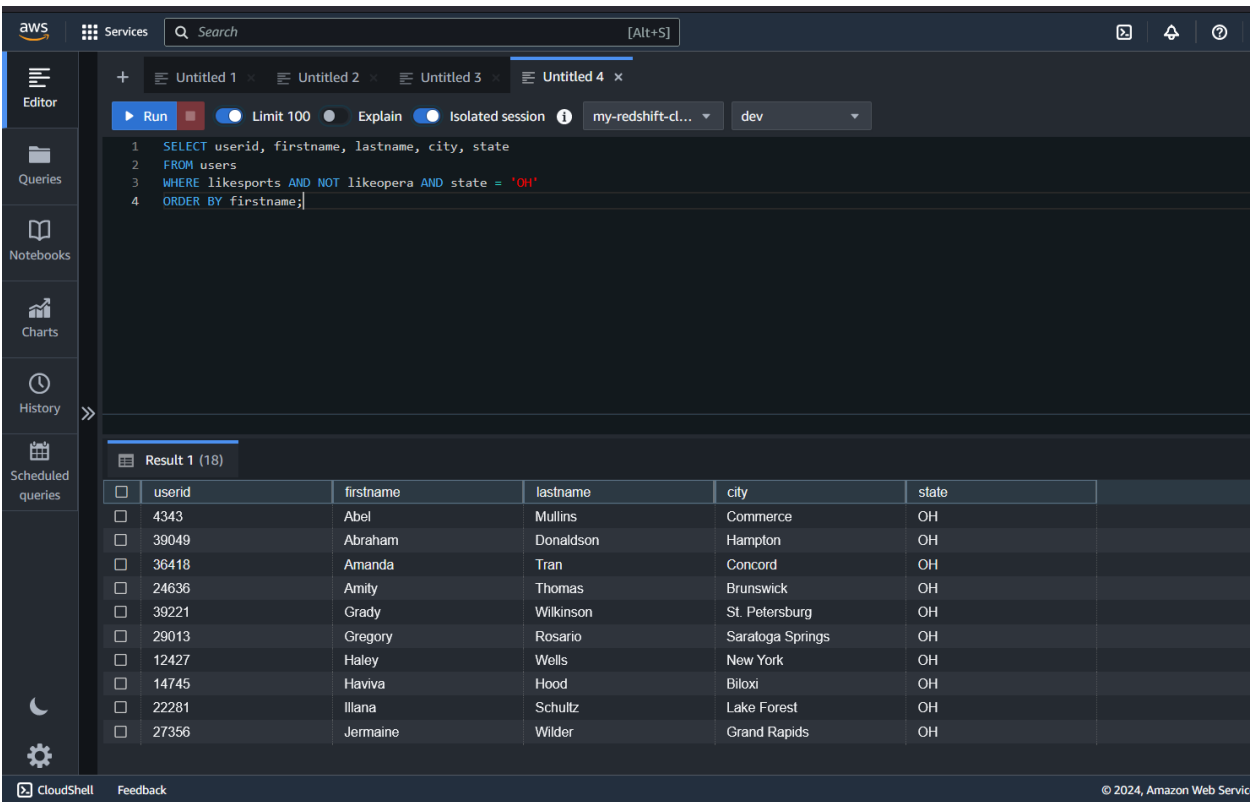


Figure 29.

In running the query code in the above diagram, we want to get information user ID, first name, last name, city state from users where people like sports and do not like opera and the state is Ohio and then giving an order by first name. The output is displayed in the Result tab below the editor area.

The screenshot shows the AWS Redshift console interface. On the left is a sidebar with navigation icons for Editor, Queries, Notebooks, Charts, History, and Scheduled queries. The main area is divided into two sections. The top section is the SQL editor, which contains a query with line numbers 1 through 8. The query is: `SELECT city, COUNT(*) AS count FROM users WHERE likejazz GROUP BY city ORDER BY count DESC LIMIT 10;`. Above the editor are tabs for 'Untitled 1' through 'Untitled 4', and a toolbar with buttons for 'Run', 'Limit 100', 'Explain', 'Isolated session', and a dropdown menu showing 'my-redshift-cl...' and 'dev'. The bottom section is the 'Result 1 (10)' tab, which displays a table with two columns: 'city' and 'count'. The table contains 10 rows of data, ordered by count in descending order.

city	count
Dover	33
Charleston	30
Hartford	28
Concord	27
Springfield	26
Richmond	24
Jackson	23
Orangeburg	23
Aliquippa	23
Columbus	23

Figure 30.

In the above diagram I wanted to see the top 10, I ran the query and I as shown the top 10 are 33 in Dover down to 23 in Columbus.

CONCLUSION

In conclusion, the Redshift Data Pipeline project has reached a significant milestone with the successful configuration and creation of the Amazon Redshift cluster. Throughout this documentation, we've meticulously detailed the setup and configuration of our cloud environment, laying the foundation for a robust data infrastructure. From the inception of our Virtual Private Cloud (VPC) to the establishment of connectivity between our Redshift cluster and Amazon S3, each step has been carefully orchestrated to ensure seamless data management and analysis.

We're now poised to leverage the power of Amazon Redshift for advanced analytics and insights generation. As we continue our journey, this project serves as a testament to the efficacy of cloud-based data pipelines in driving business intelligence and decision-making processes.