**AMAZON REDSHIFT**

**What is Amazon redshift?**

Amazon redshift is an data warehouse in cloud, which process large amount of data in petabyte scale data warehouse service by Amazon Web Services (AWS). It is an efficient solution to collect and store all your data and enables you to analyze it using various business intelligence tools to acquire new insights for your business and customers.

**What is data warehouse?**

Data warehouse will have information that stores all the data from single or multiple resources, which has historical data for future analysis work to take information and decision for an organization.

**What is database?**

Database will have schema to create tables and column which has current data to run the business.

**Difference between database and data warehouse ?**

|  |  |
| --- | --- |
| **Database** | **Data warehouse** |
| **Day today activity** | **Historical process** |
| **will have current data** | **It will have historical data** |
| **size can be 100 mb** | **Size can be – 100 gb -100tb or petabyte** |
| **High performance** | **Highly flexible** |
| **structured and filtered in row and columns**  **database** | **Multi dimensional view** |
| **Used to run the business** | **Use to analyze the business** |
| **Used by DBA** | **Used by manager and analyst** |
| **Uses OLTP (online transcation processing)** | **Uses OLAP (online analytical processing)** |

**Benefits:**

1.faster performance

2.easy to setup, deploy and manage

3.cost-effective- no upfront cost

4.secure-encrypted by kms

**Use cases:**

1.Finance, healthcare, educational institute can use Redshift for analytics purposes. eg: to generate report, or create dashboard for last 10 to 15 years it will respond in sub second results for interactive analysis.

2. On premises data warehouse in order to make it fast, scalable, cost-effective can bring out by using redshift which automates the administrative task to scale, manage and maintain your data warehouse for time consuming labor intensive task.

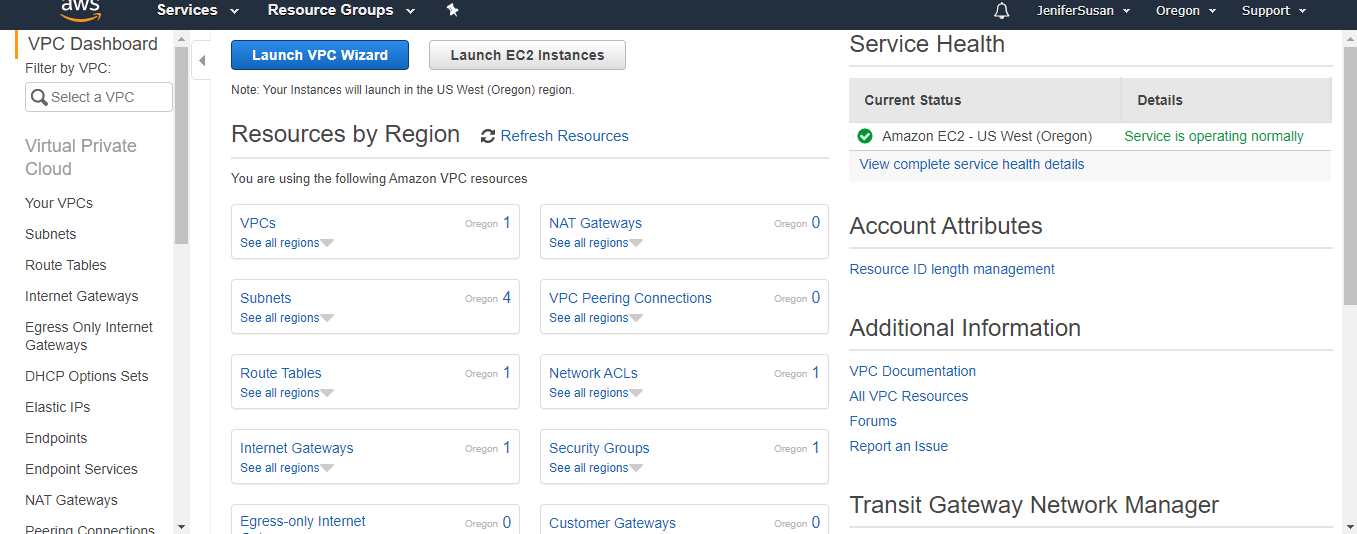
Steps to create:

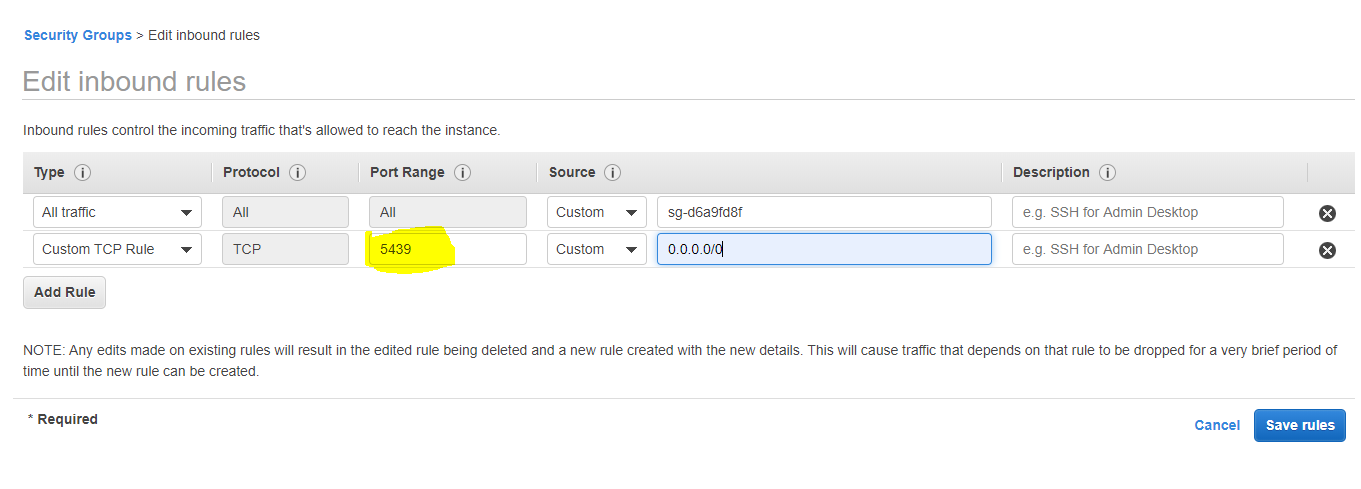
* **Step 1: Set Up Prerequisites**
* **Step 2: Create an IAM Role**
* **Step 3: Create a Sample Amazon Redshift Cluster**
* **Step 4: Authorize Access to the Cluster**
* **Step 5: Connect to the Sample Cluster and Run Queries**
* **Step 6: Load Sample Data from Amazon S3**
* **Step 7: Reset Your Environment**

**Step 1: Set Up Prerequisites**

Amazon Redshift uses port 5439 by default, the connection doesn't work if that port is not open in your firewall. You can't change the port number for your Amazon Redshift cluster after it is created. Thus, make sure that you specify an open port that works in your environment during the launch process.

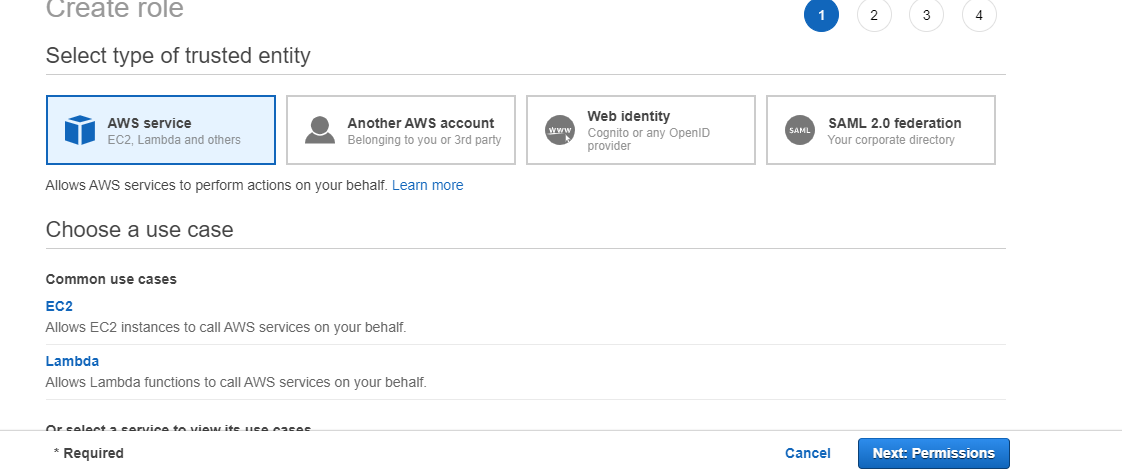
Open the security group – default vpc – edit inbound rules by adding port number 5439

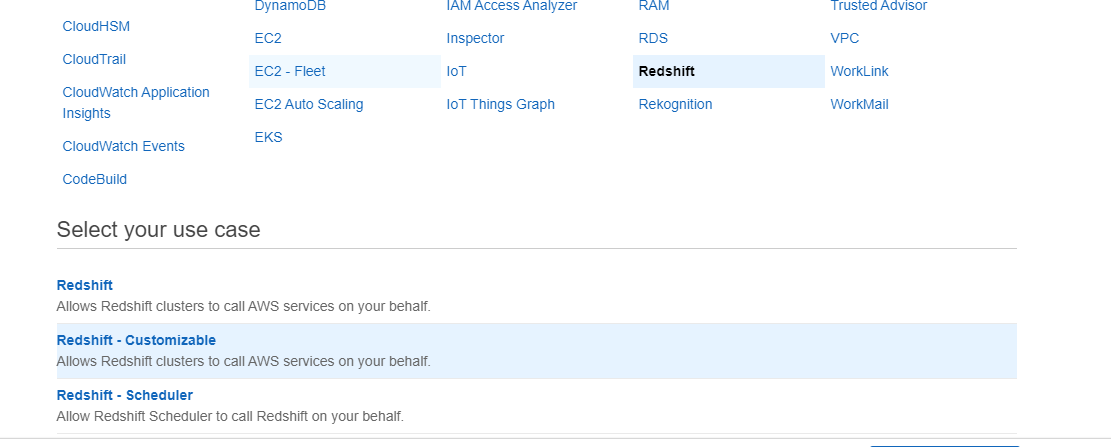


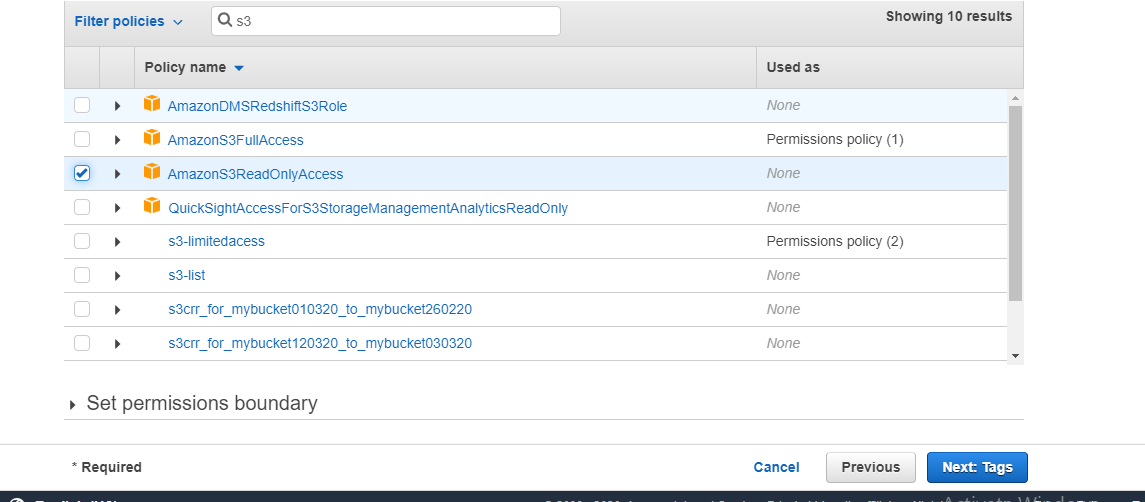


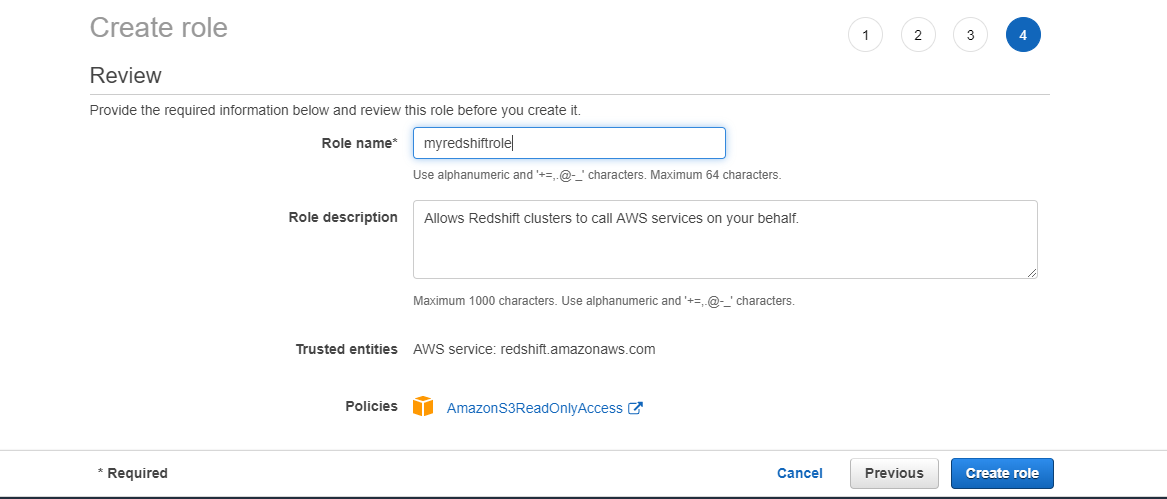
* **Step 2: Create an IAM Role**

1. Choose **Create role**.
2. In the **AWS Service** group, choose **Redshift.**
3. Under **Select your use case**, choose **Redshift - Customizable** then choose **Next: Permissions**.
4. On the **Attach permissions policies** page, choose **AmazonS3ReadOnlyAccess**. You can leave the default setting for **Set permissions boundary**. Then, choose **Next: Tags**.
5. The **Add tags** page appears. You can optionally add tags. Choose **Next: Review**.
6. For **Role name**, enter a name for your role. For this tutorial, enter **myRedshiftRole**.
7. Review the information, and then choose **Create Role**.
8. Choose the role name of the role you just created.
9. Copy the **Role ARN** to your clipboard—this value is the Amazon Resource Name (ARN) for the role that you just created. You use that value when you use the COPY command to load data in [Step 6: Load Sample Data from Amazon S3](https://docs.aws.amazon.com/redshift/latest/gsg/rs-gsg-create-sample-db.html).

****

****

****

****

* **Step 3: Create a Sample Amazon Redshift Cluster**

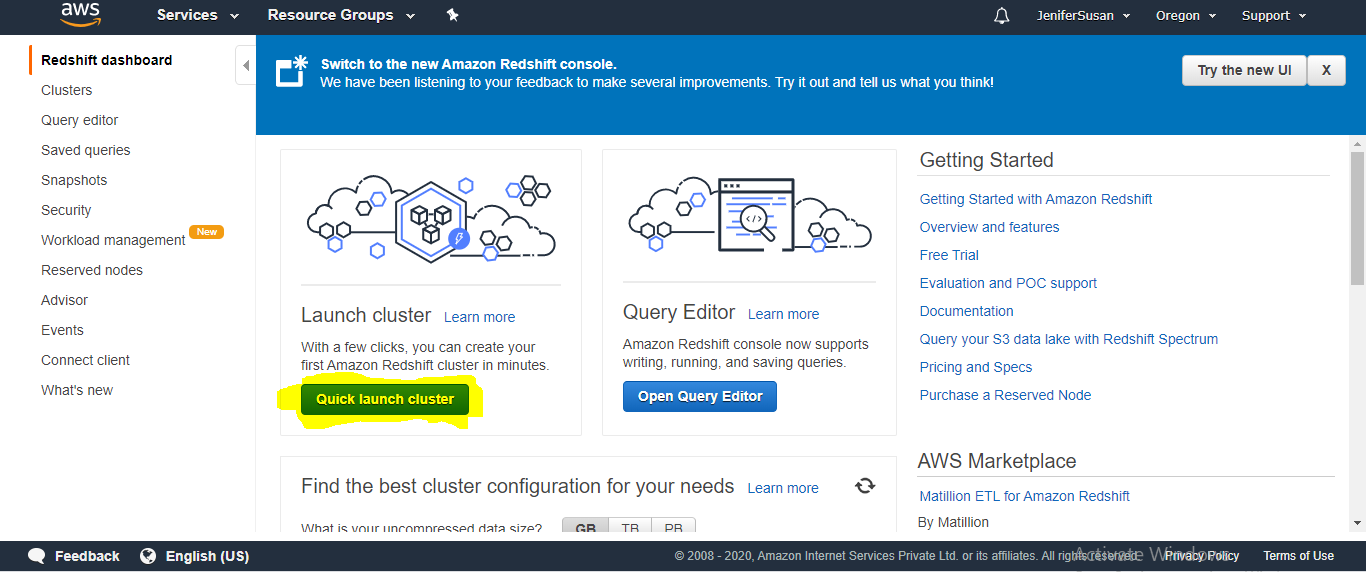
1. At top right, choose the AWS Region in which you want to create the cluster.
2. On the navigation menu, choose CLUSTERS, then choose Create cluster. The Create cluster page appears.
3. Choose dc2.large for the node type in the Compute optimized section. For Nodes, choose 2.
   * In the Cluster details section, specify values for Cluster identifier, Database port, Master user name, and Master user password.
   * Cluster identifier: redshift-cluster-1
   * Database port: Enter 5439.
   * Master user name: Enter awsuser.
   * Master user password**: Awsuser123**
4. In the Cluster permissions section, for Available IAM roles choose the IAM role that you previously created, myRedshiftRole. Then choose Add IAM role.
5. Choose Create cluster.

What is cluster ?

A cluster is the core unit of operations in the Amazon Redshift data warehouse. Each Redshift cluster is composed of two main components: Compute Node, which has its own dedicated CPU, memory, and disk storage. Compute nodes store data and execute queries and you can have many nodes in one cluster.

Redshift is meant to work in a Cluster formation. A typical Redshift Cluster has two or more Compute Nodes which are coordinated through a Leader Node. All client applications communicate with the cluster only with the Leader Node.

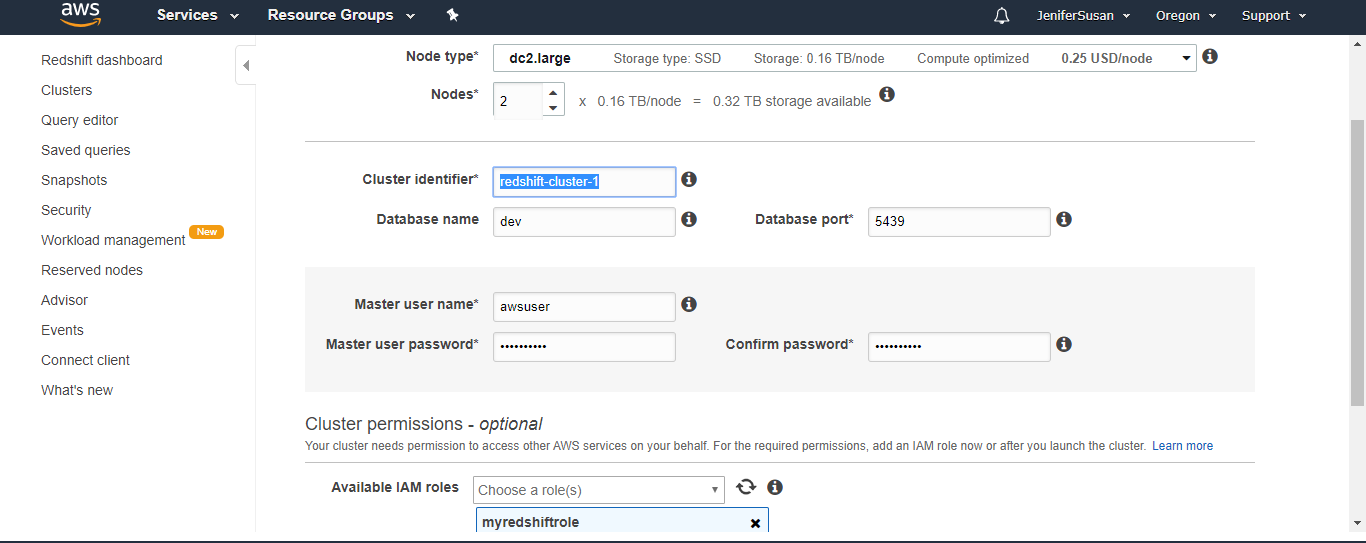
Choose Aws region as Oregon and click quick launch cluster

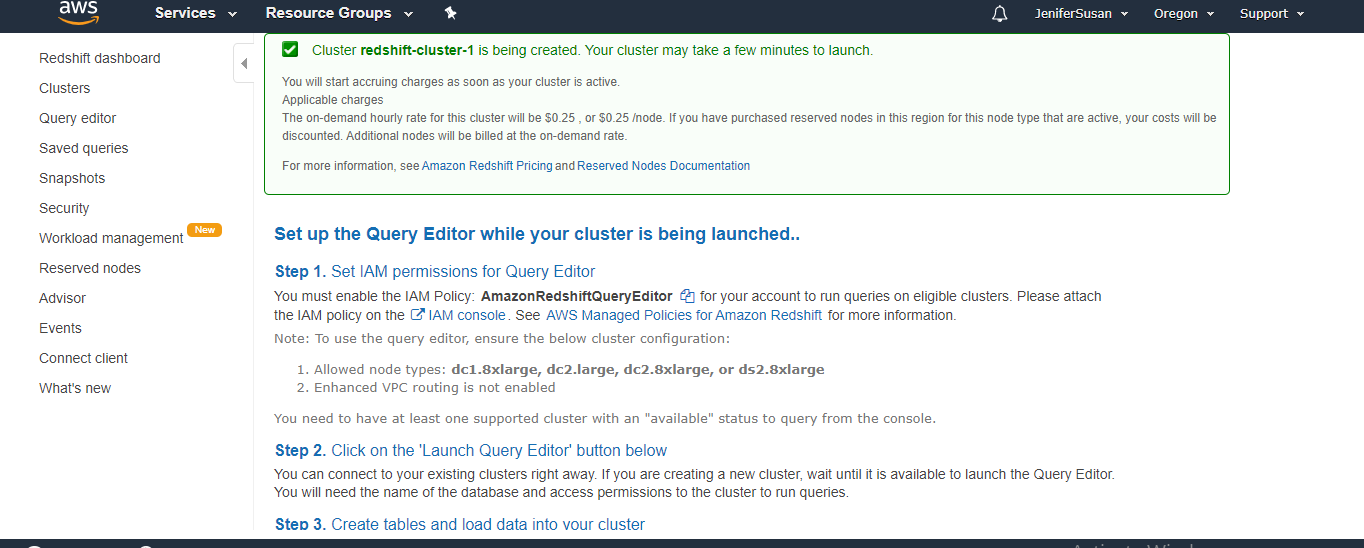


Enter the following details to create cluster :

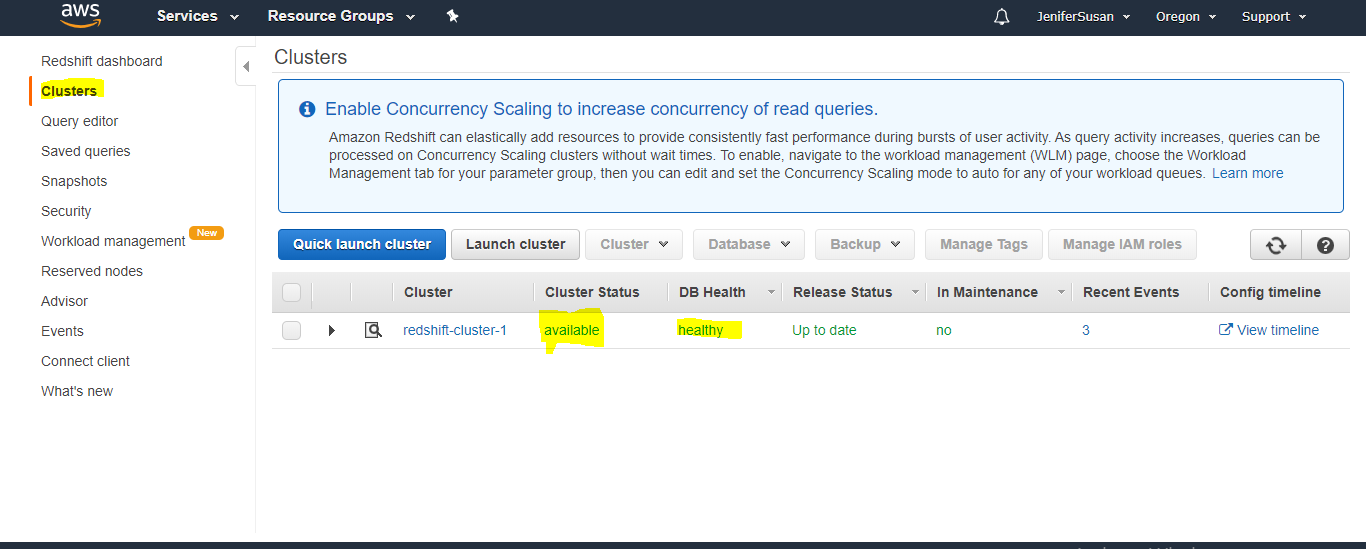
* + Cluster identifier: redshift-cluster-1
  + Database port: Enter 5439
  + Master user name: Enter awsuser
  + Master user password**: Awsuser123**

1. In the Cluster permissions section, for Available IAM roles choose the IAM role that you previously created, myRedshiftRole. Then choose Add IAM role.
2. Choose launch cluster.





# Step 4: Authorize Access to the Cluster



In the previous step, you launched your Amazon Redshift cluster. Before you can connect to the cluster, you need to configure a security group to authorize access.

To configure the VPC security group

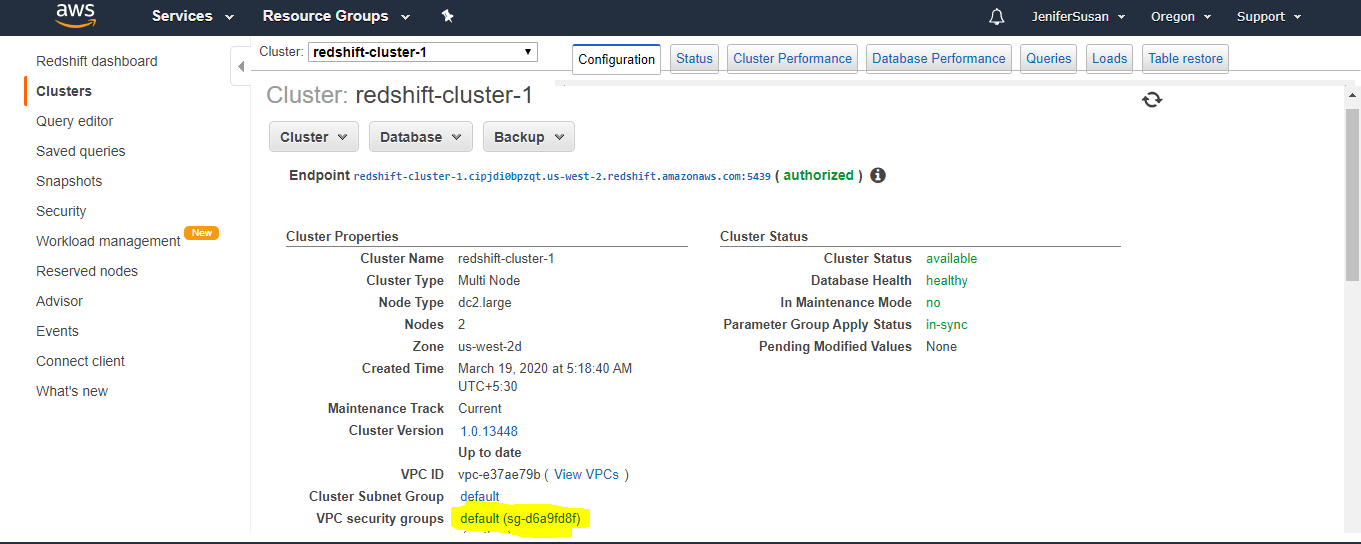
In the Amazon Redshift console, in the navigation pane, choose Clusters.

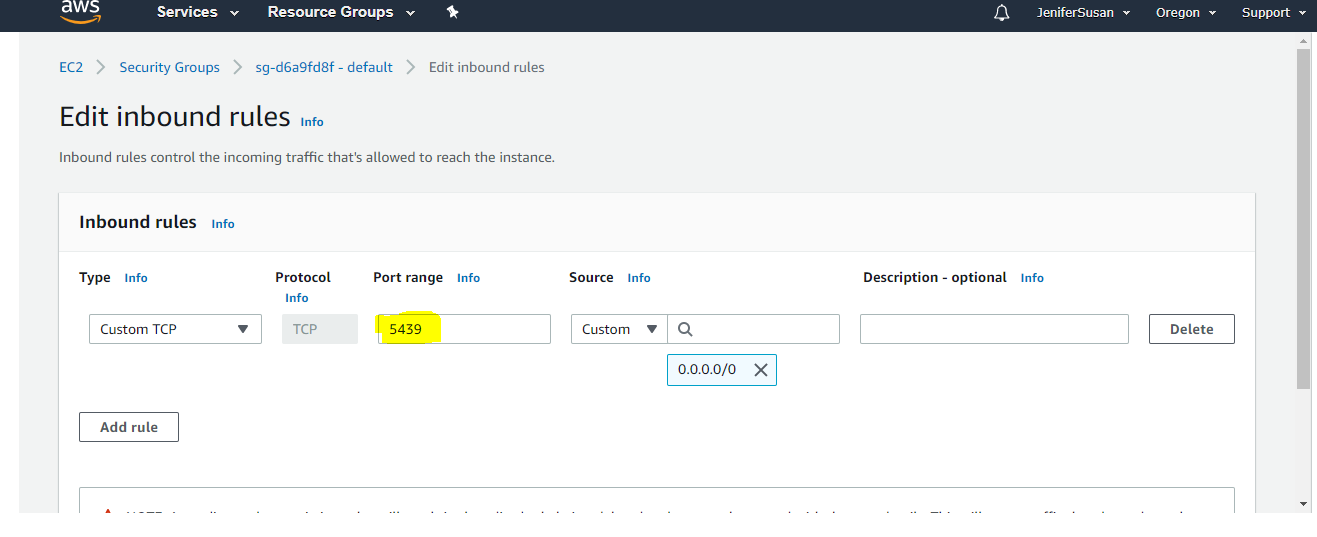
1. Choose redshift cluster1 to open it, and make sure that you are on the Configuration tab.
2. Under Cluster Properties, for VPC Security Groups, choose your security group.

After your security group opens in the Amazon EC2 console, choose the Inbound tab.

Choose Edit, Add Rule, and enter the following, then choose Save:

* Type: Custom TCP Rule.
* Protocol: TCP.
* Port Range: Enter the same port number that you used when you launched the cluster. The default port for Amazon Redshift is 5439, but your port might be different.
* Source: Select Custom, then enter 0.0.0.0/0.





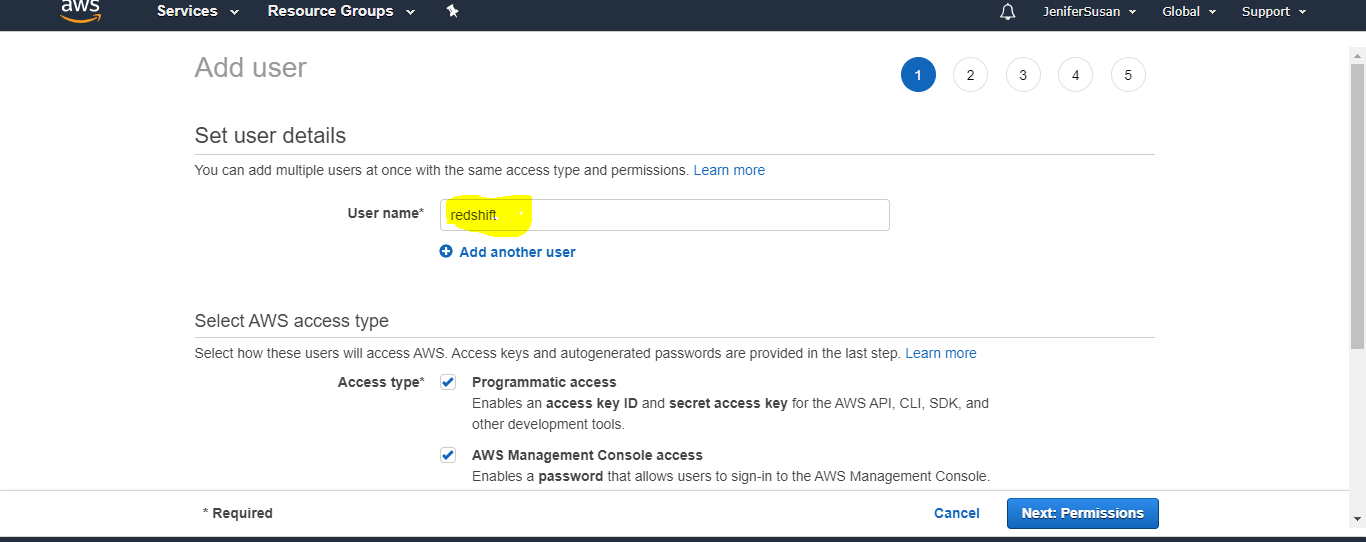
# Step 5: Connect to the Sample Cluster and Run Queries

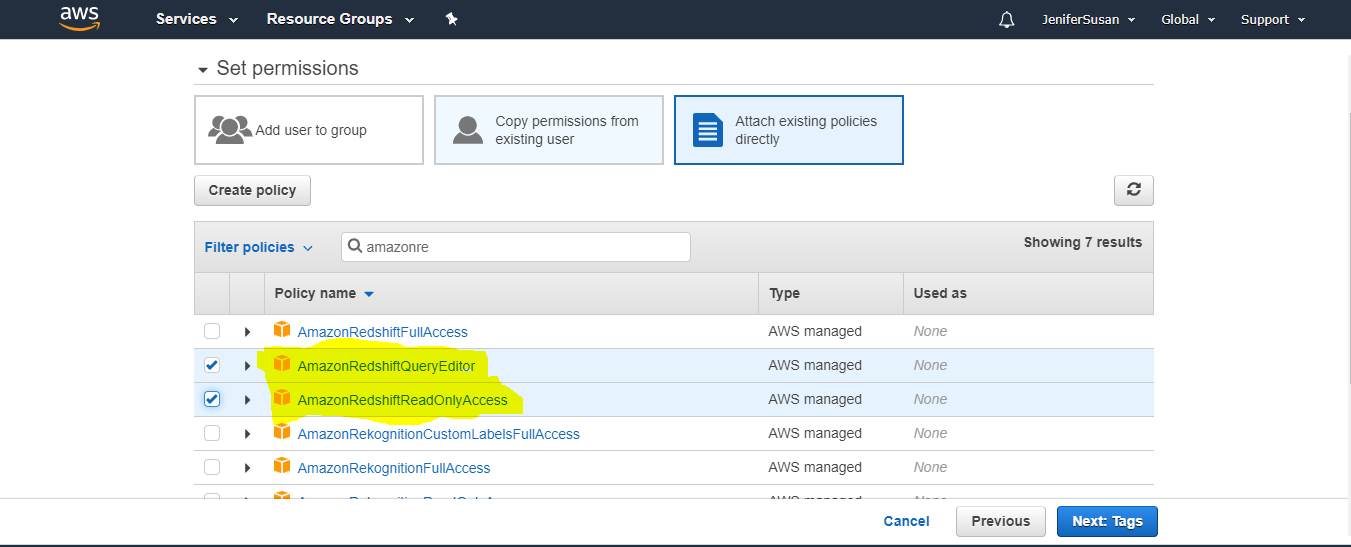
1.Querying database using query editor

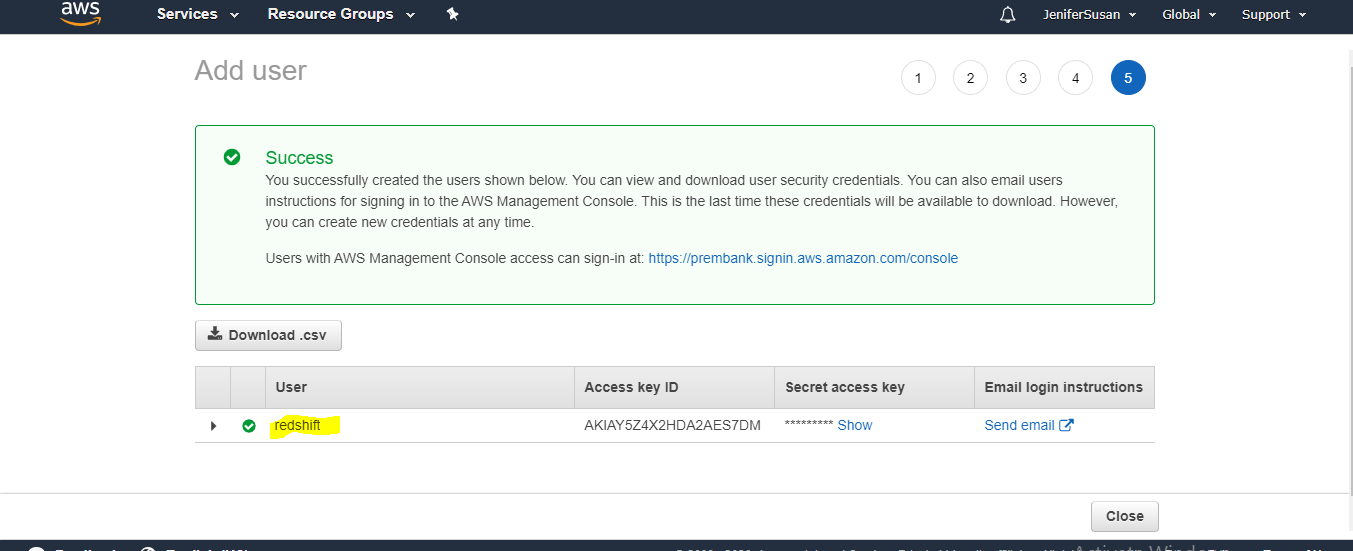
### 2.Enabling Access to the Query Editor

3.To attach the required IAM policies for the Query Editor

* Choose Users or create a new user in the name redshift
* Choose the user that needs access to the Query Editor.
* Choose Add permissions.
* Choose Attach existing policies directly.
* For Policy names, choose AmazonRedshiftQueryEditor and AmazonRedshiftReadOnlyAccess.
* Choose Next: Review.
* Choose Add permissions.







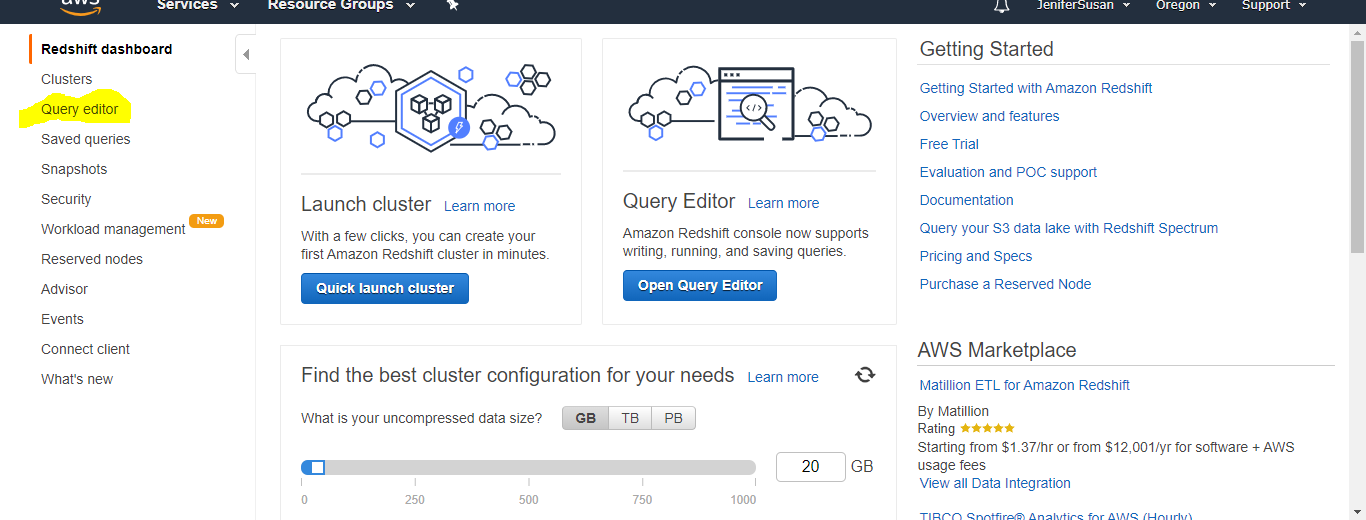
### Using the Query Editor

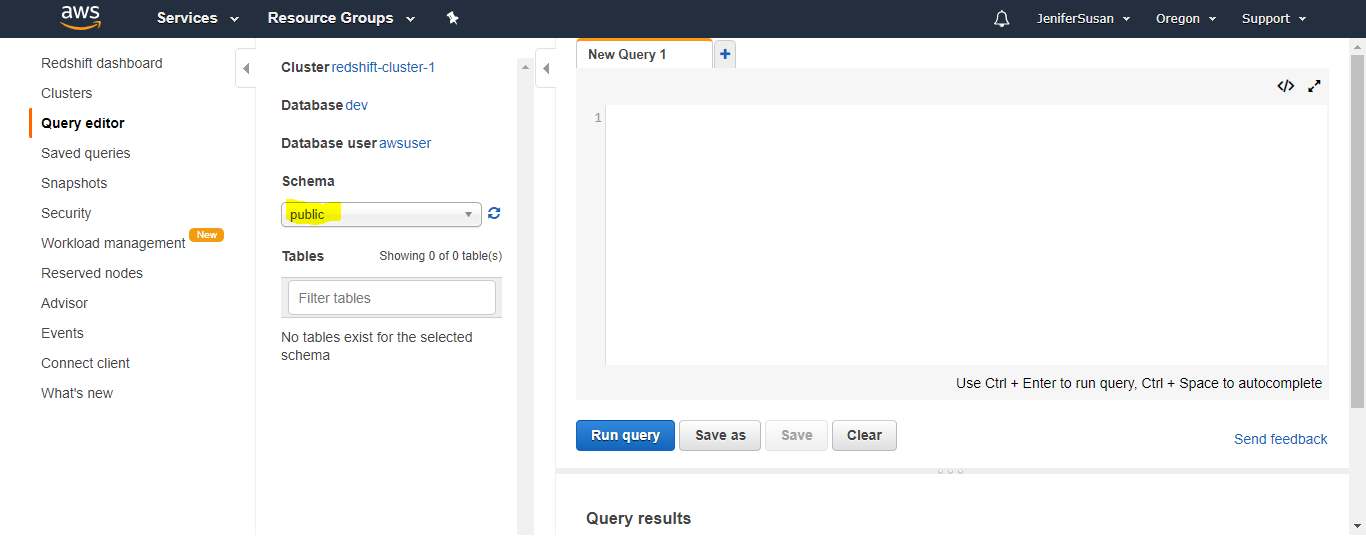
1. open the Amazon Redshift console
2. On the navigation menu, choose EDITOR, then connect to a database in your cluster.

On the Connect to database window, enter the values you used when you created the cluster as follows:

* + Cluster: Choose redshift cluster1
  + Database name: Enter dev
  + Database user: Enter awsuser
  + Database password: Awsuser123
  + Then choose Connect to database.

1. For Schema, choose public to create a new table based on that schema.
2. Enter the following in the query editor window, and choose Run query to create a new table.

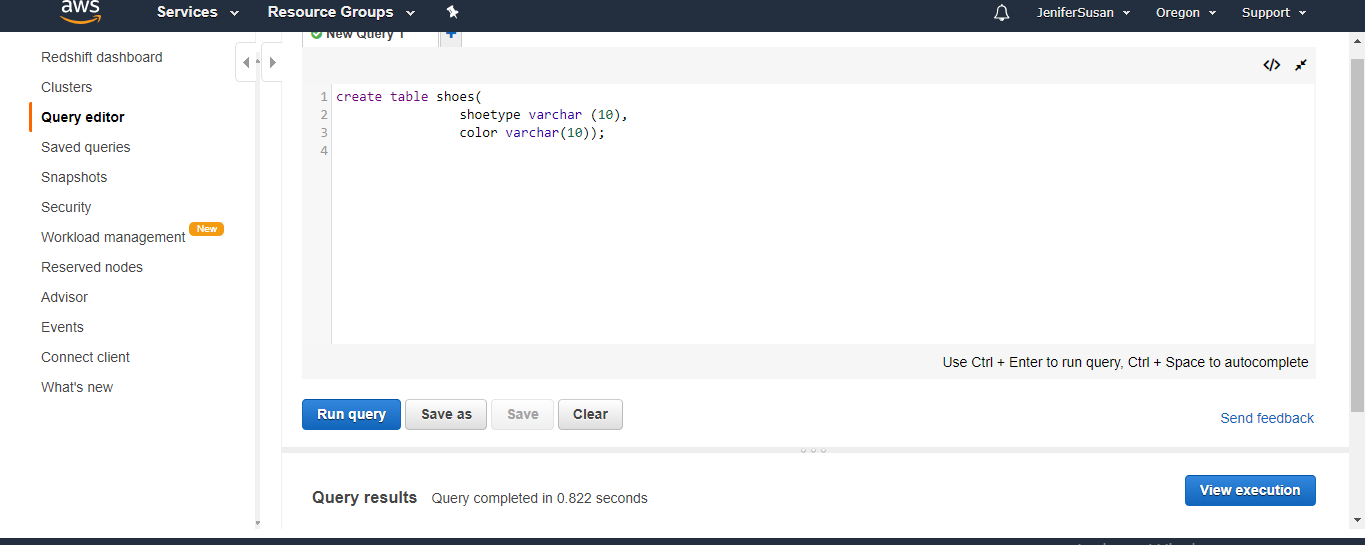




create table shoes(

shoetype varchar (10),

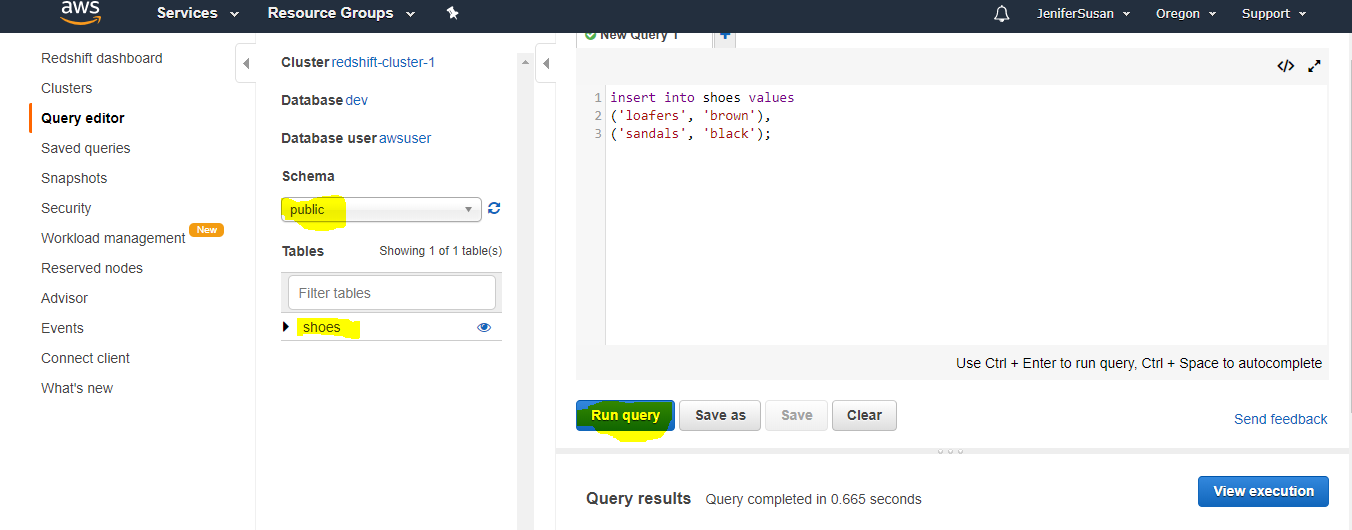
color varchar(10));



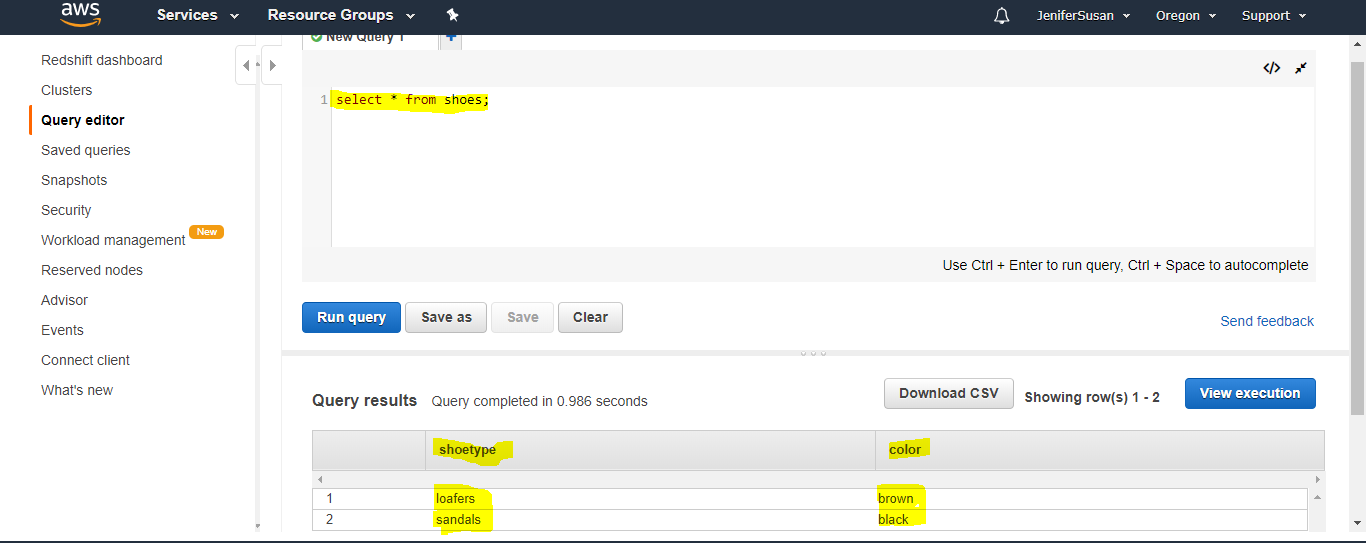
insert into shoes values

('loafers', 'brown'),

('sandals', 'black');



select \* from shoes;



# Step 6: Load Sample Data from Amazon S3

**To load sample data**

1. Create tables.

create tables in the dev database.

create table users(

userid integer not null distkey sortkey,

username char(8),

firstname varchar(30),

lastname varchar(30),

city varchar(30),

state char(2),

email varchar(100),

phone char(14),

likesports boolean,

liketheatre boolean,

likeconcerts boolean,

likejazz boolean,

likeclassical boolean,

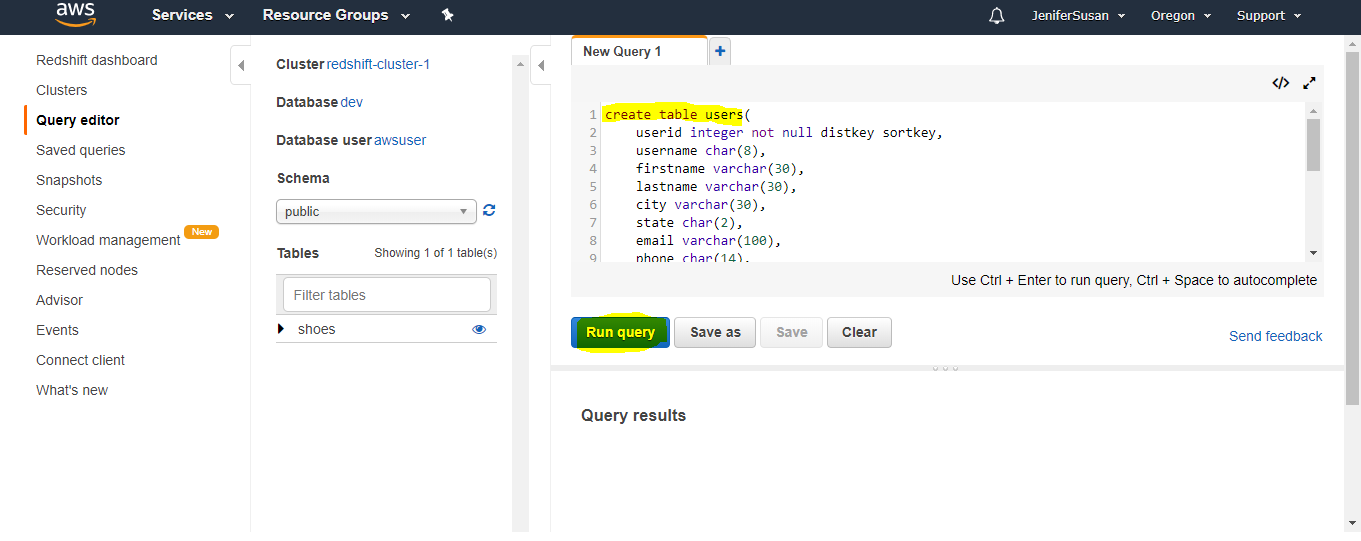
likeopera boolean,

likerock boolean,

likevegas boolean,

likebroadway boolean,

likemusicals boolean);



create table sales(

salesid integer not null,

listid integer not null distkey,

sellerid integer not null,

buyerid integer not null,

eventid integer not null,

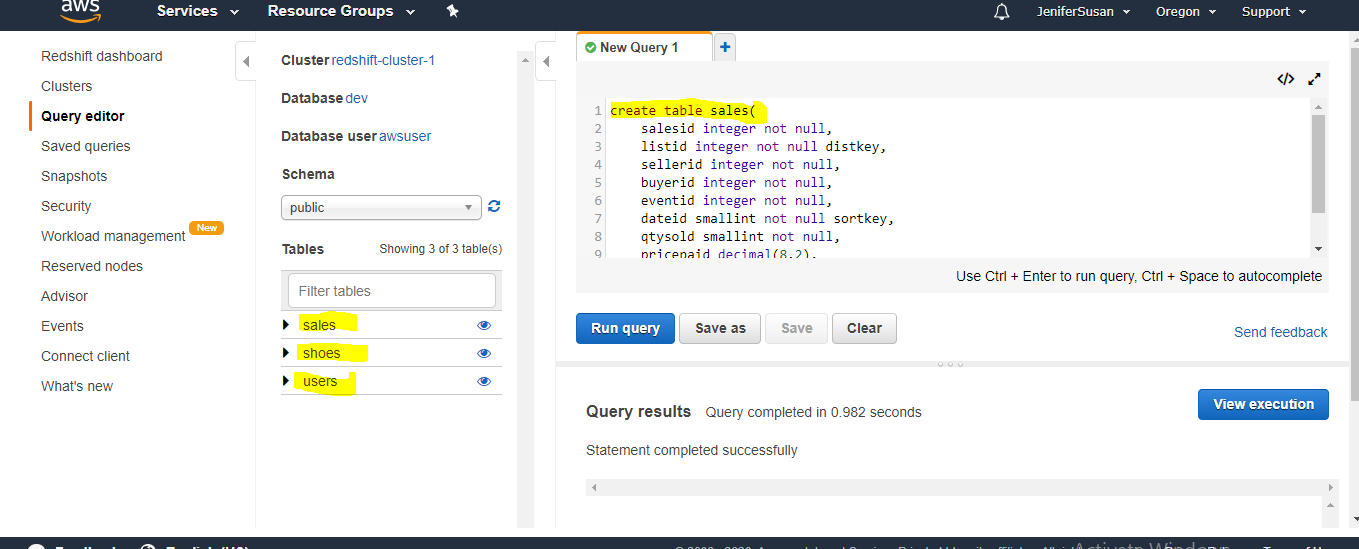
dateid smallint not null sortkey,

qtysold smallint not null,

pricepaid decimal(8,2),

commission decimal(8,2),

saletime timestamp);



create table venue(

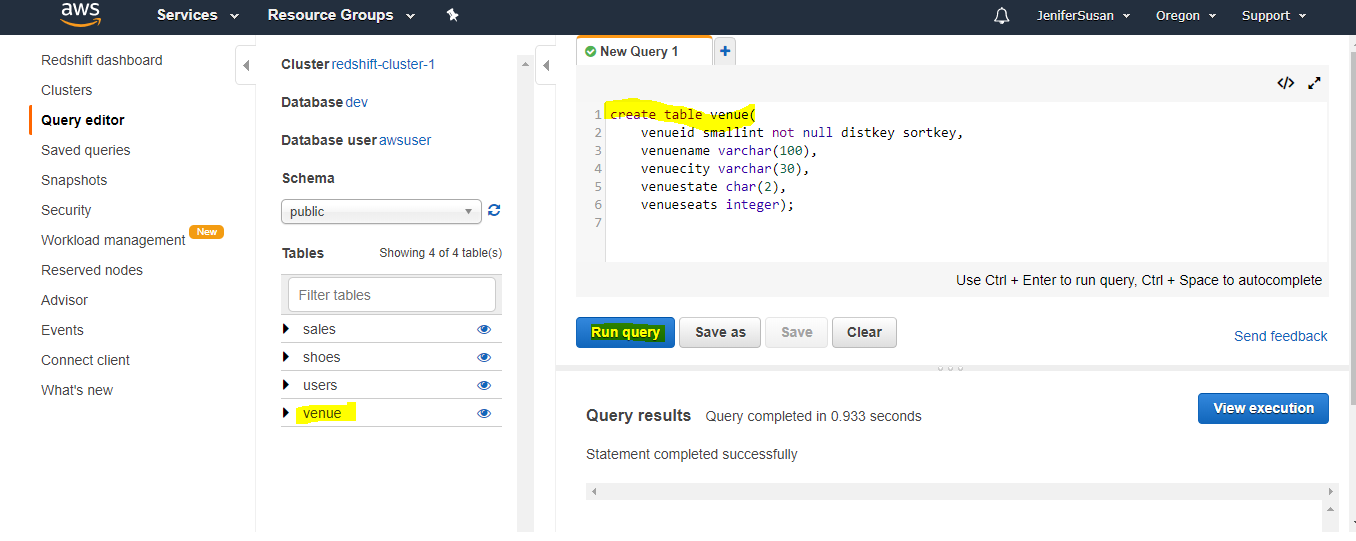
venueid smallint not null distkey sortkey,

venuename varchar(100),

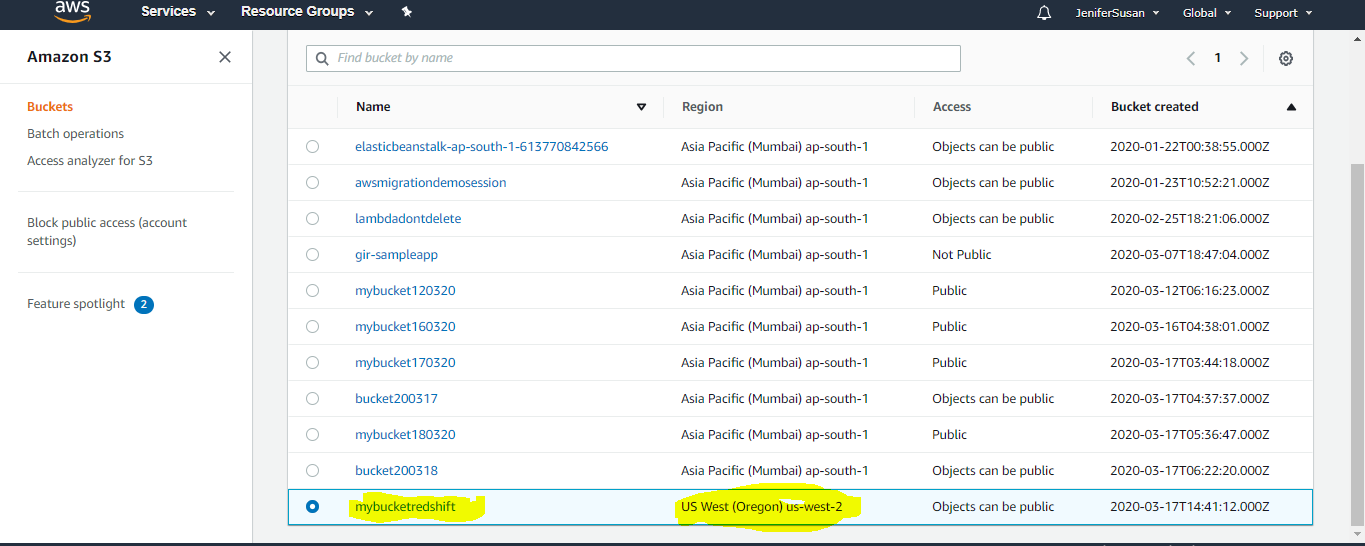
venuecity varchar(30),

venuestate char(2),

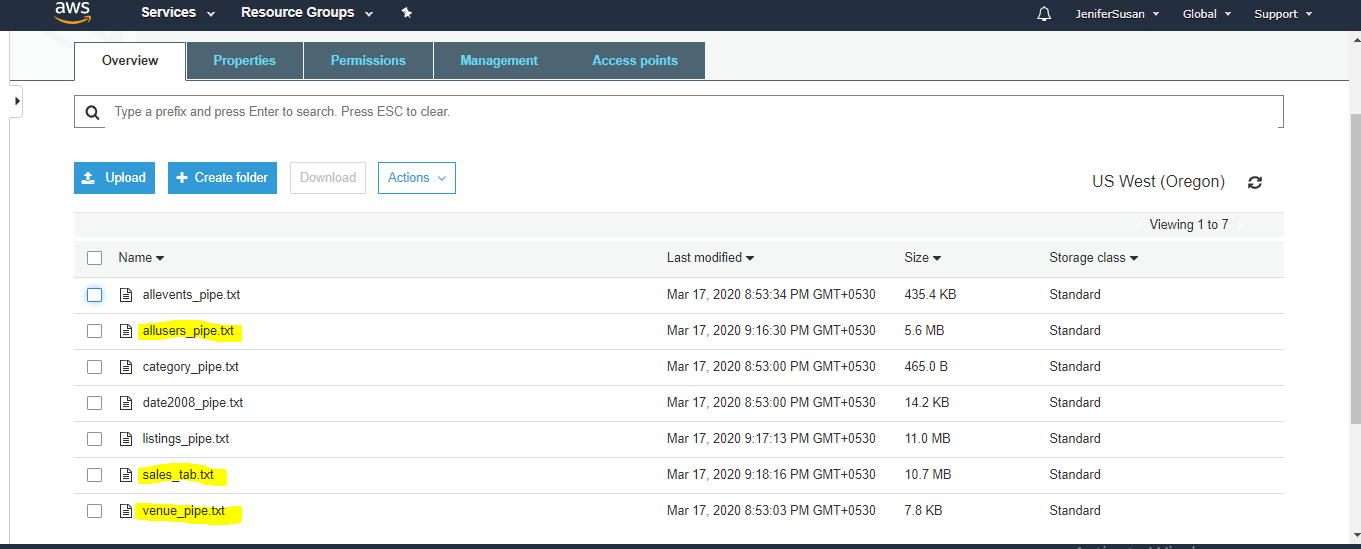
venueseats integer);



Create a newbucket name mybucketredshift in the same region where ur redshift cluster is created



Upload the sample data files in s3 individually.



Goto redshift console page in query editor

Add the following with copy commands

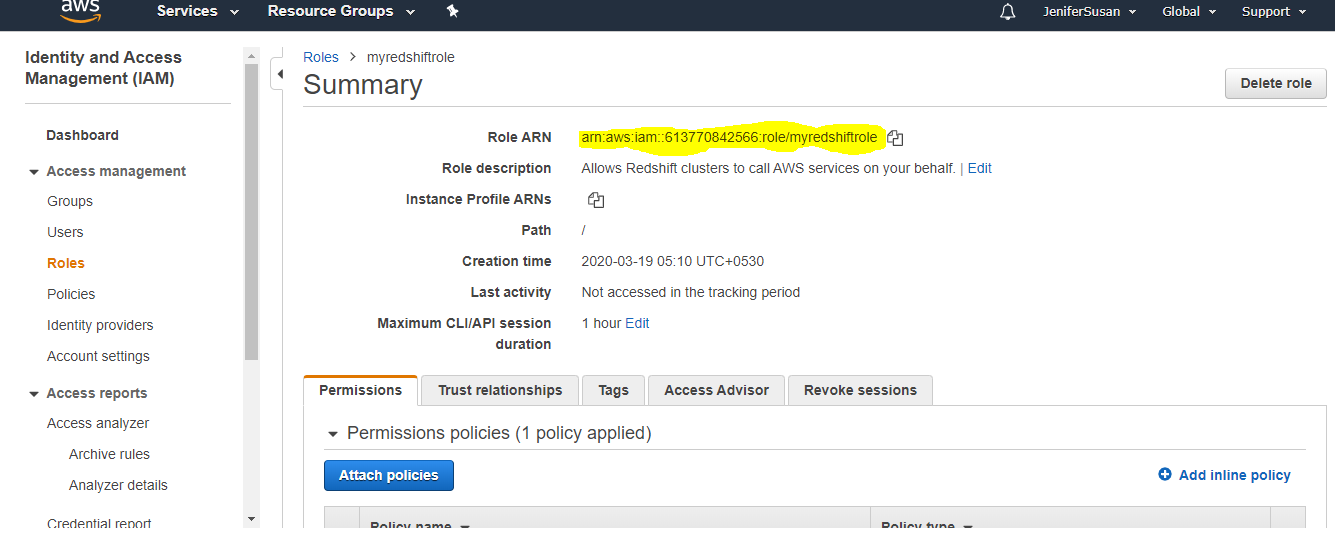
copy users from 's3://*<myBucket>*/allusers\_pipe.txt'

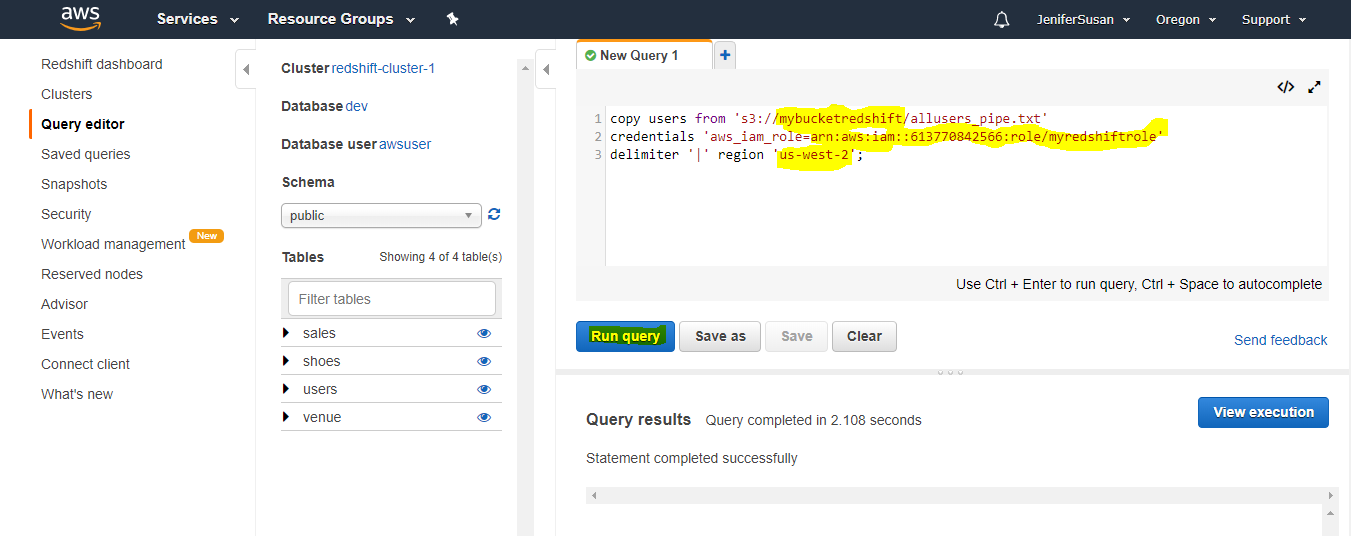
credentials 'aws\_iam\_role=*<iam-role-arn>*'

delimiter '|' region '*<aws-region>*';

change the bucket name and region name.

copy this ARN number of IAM role which you created earlier and paste it in query editor



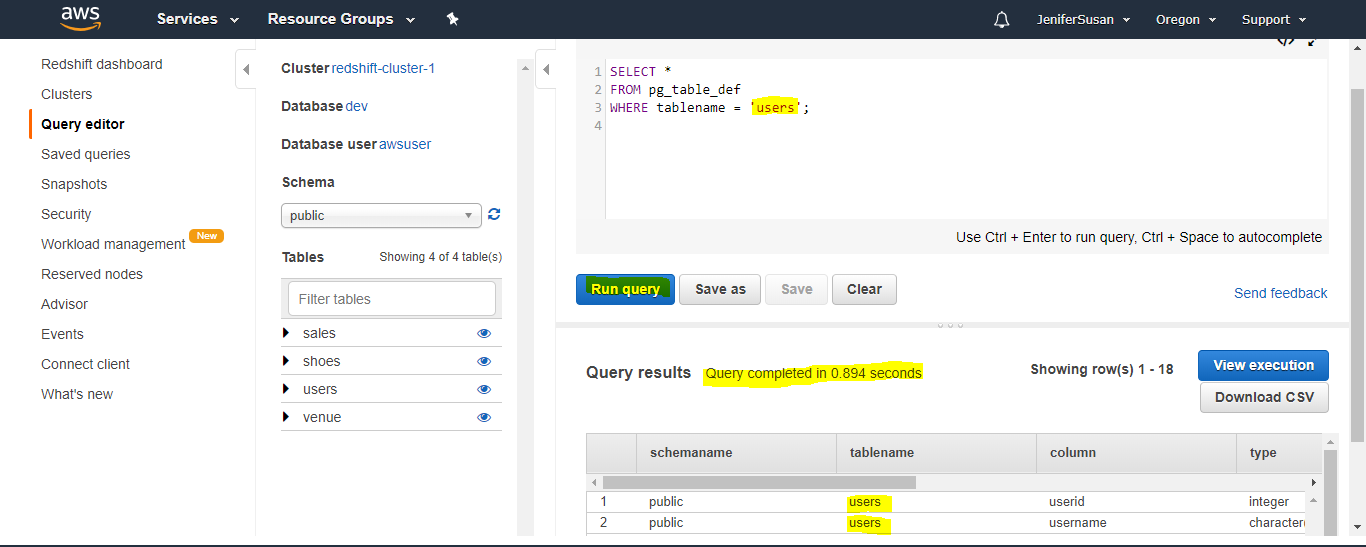


Now try to add some example queries.

SELECT \*

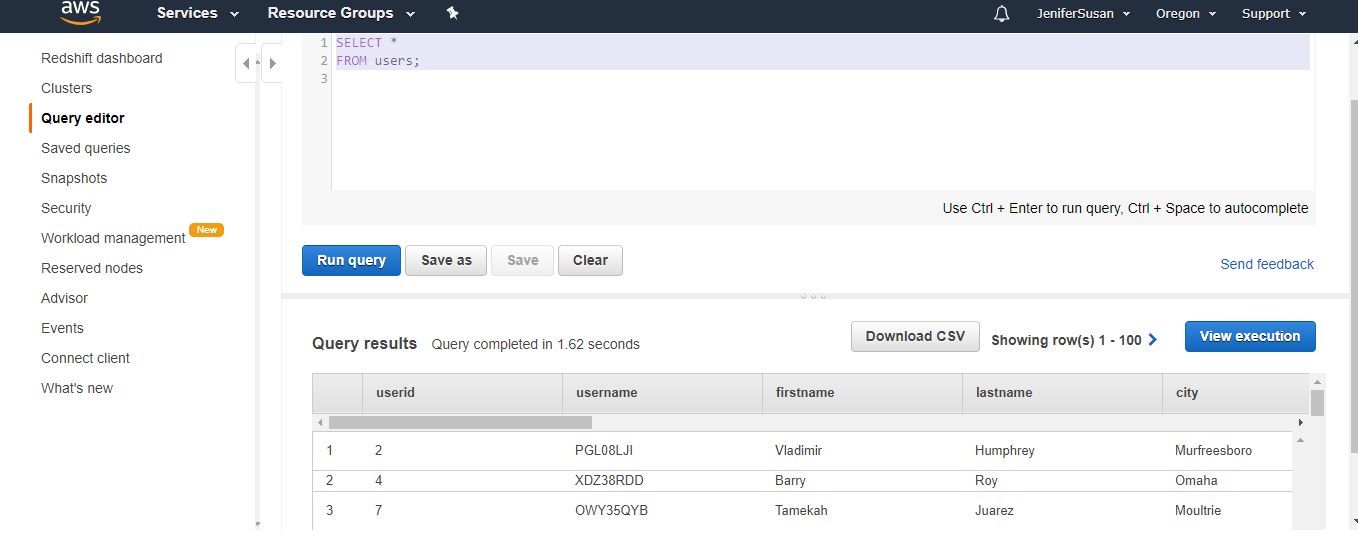
FROM pg\_table\_def

WHERE tablename = 'users';



SELECT \*

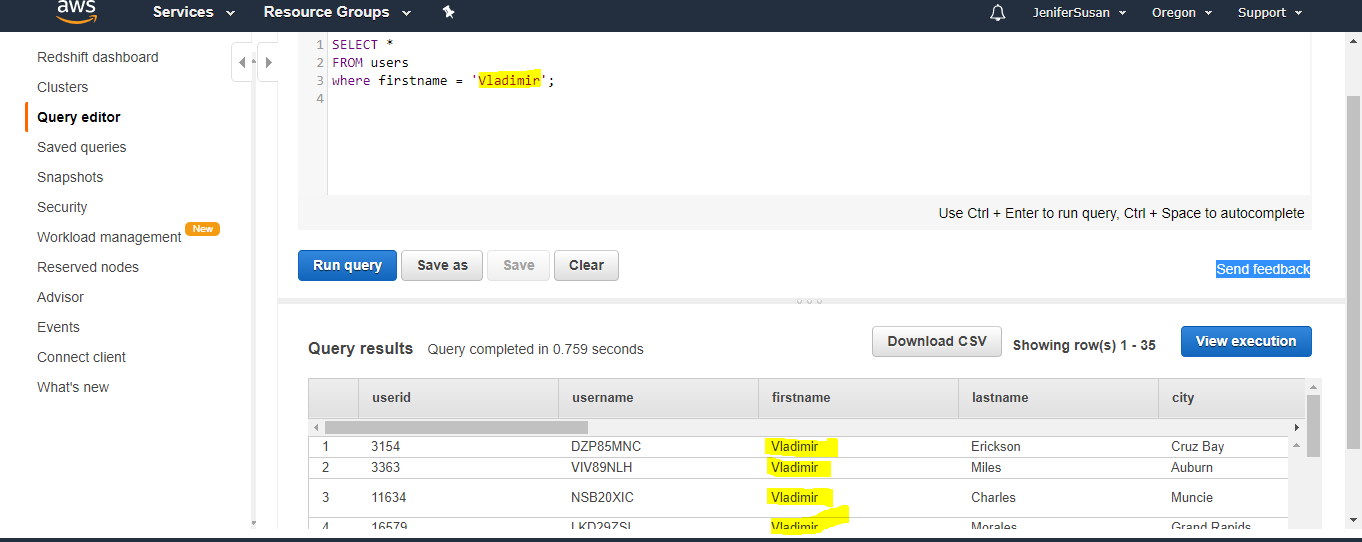
FROM users;



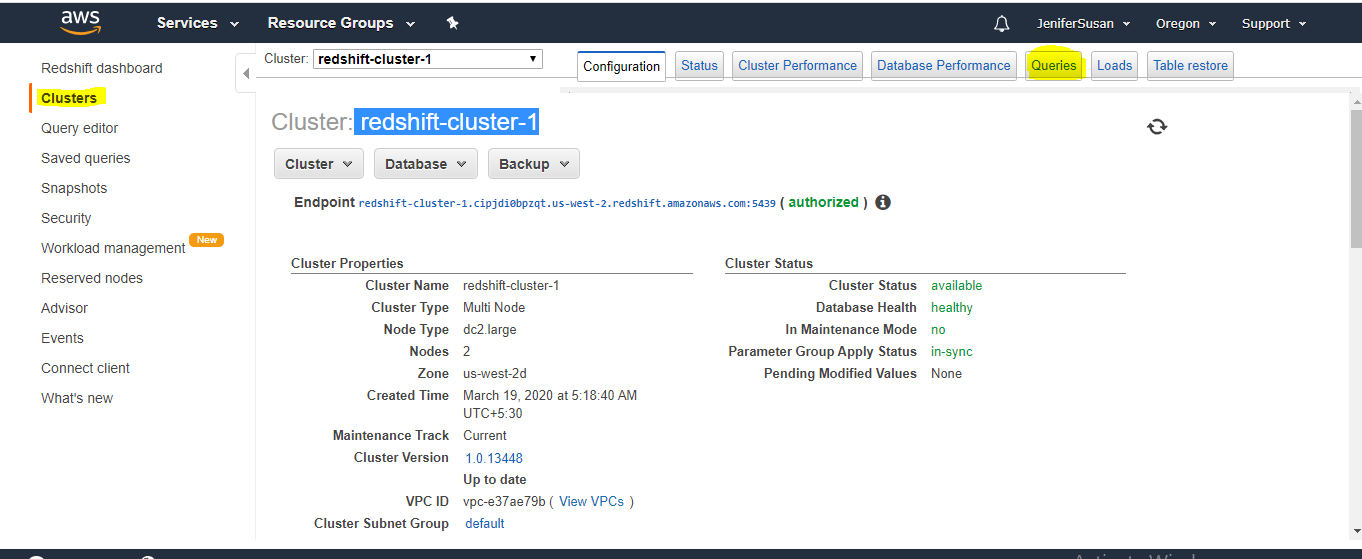
SELECT \*

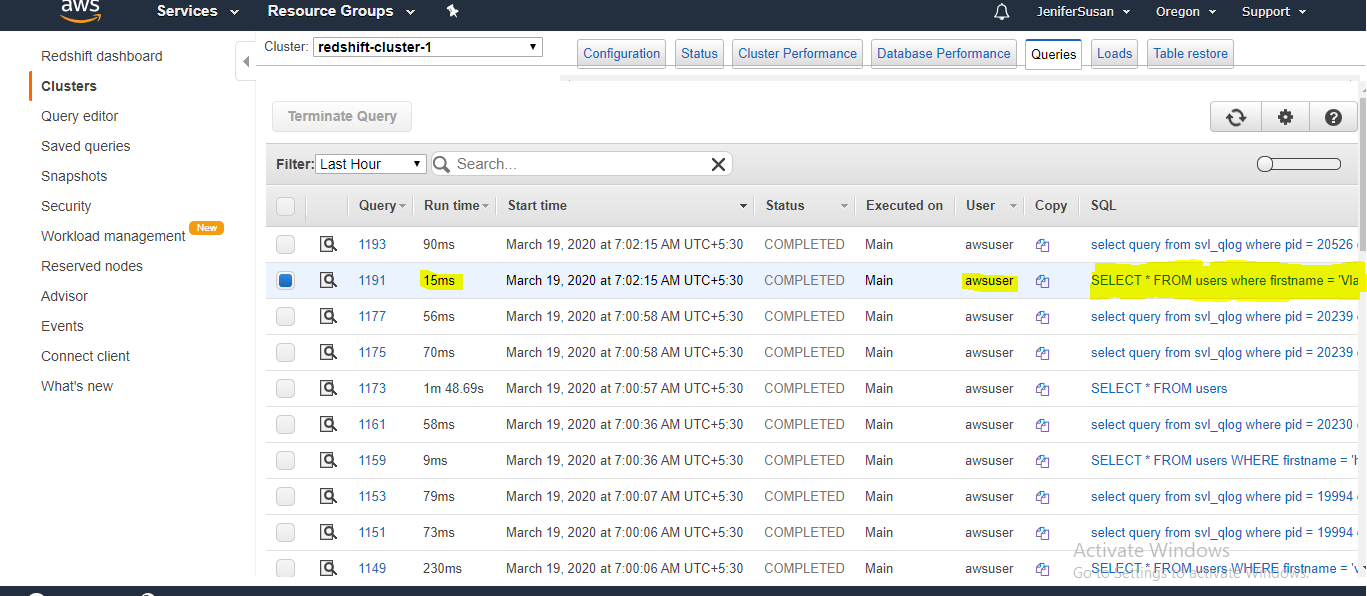
FROM users

where firstname = 'Vladimir';

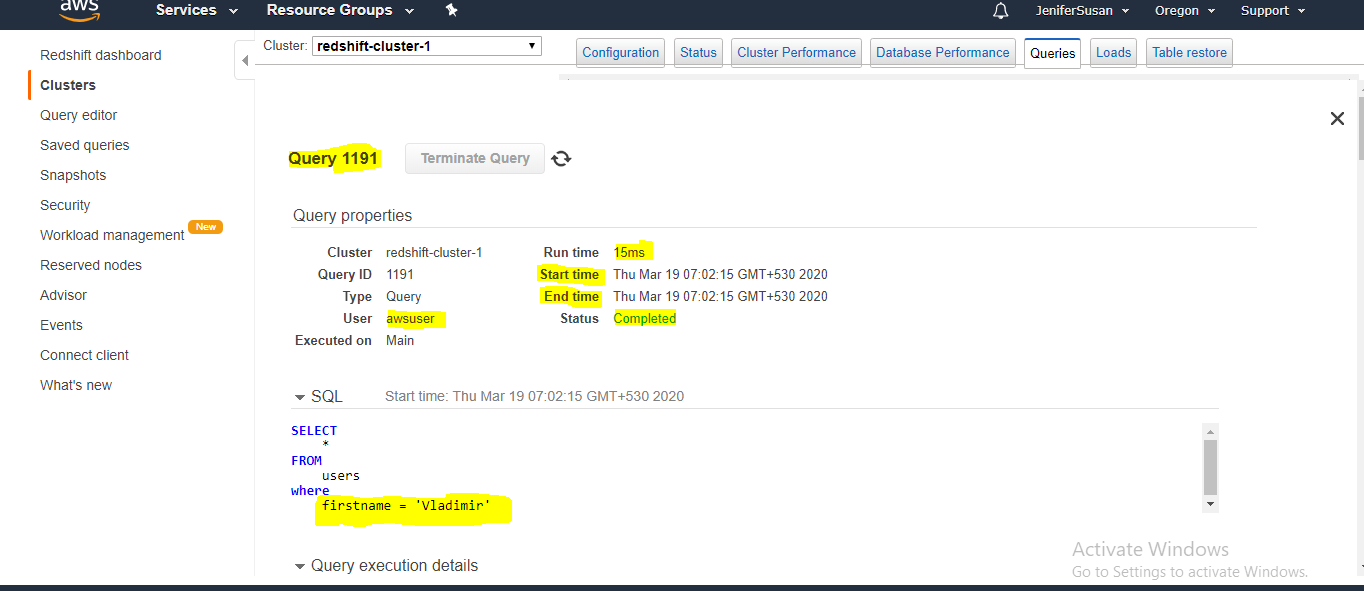


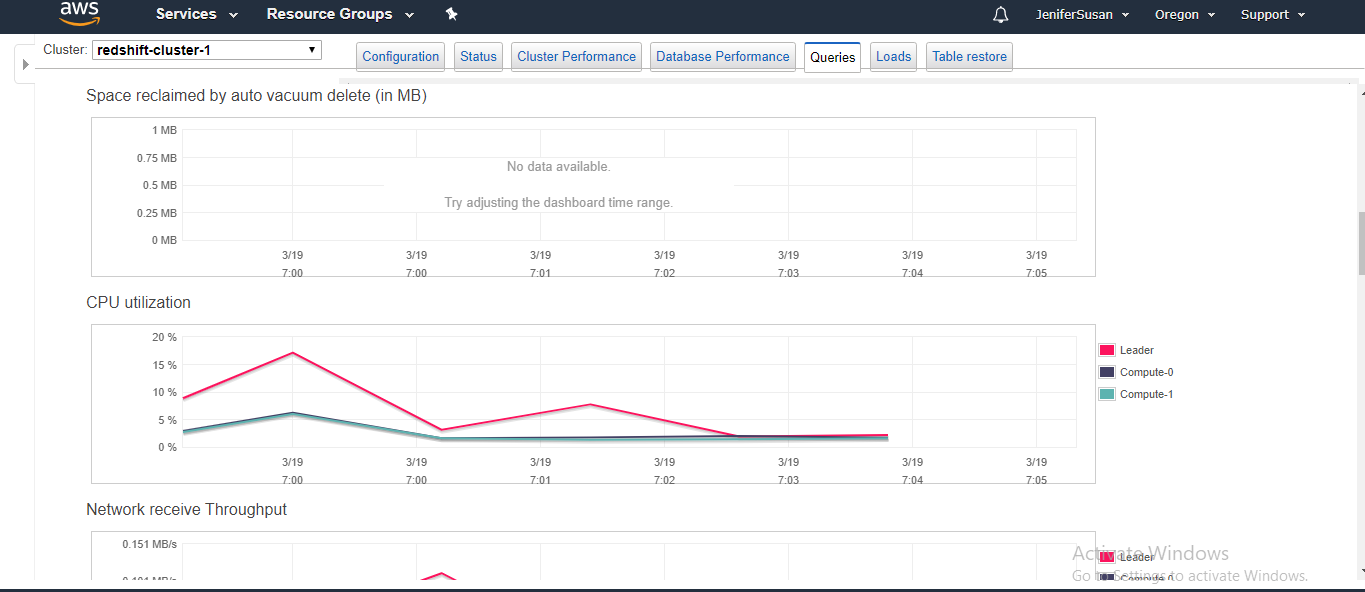
Goto back cluster  redshift-cluster-1 console page click queries





**To check the detail list of the particular query**





**Conculsion :**

1. you can analyses the particular query

2.who is the user who has long running query

3.who is the frequent user in the database

4.Historical data about the particular user.

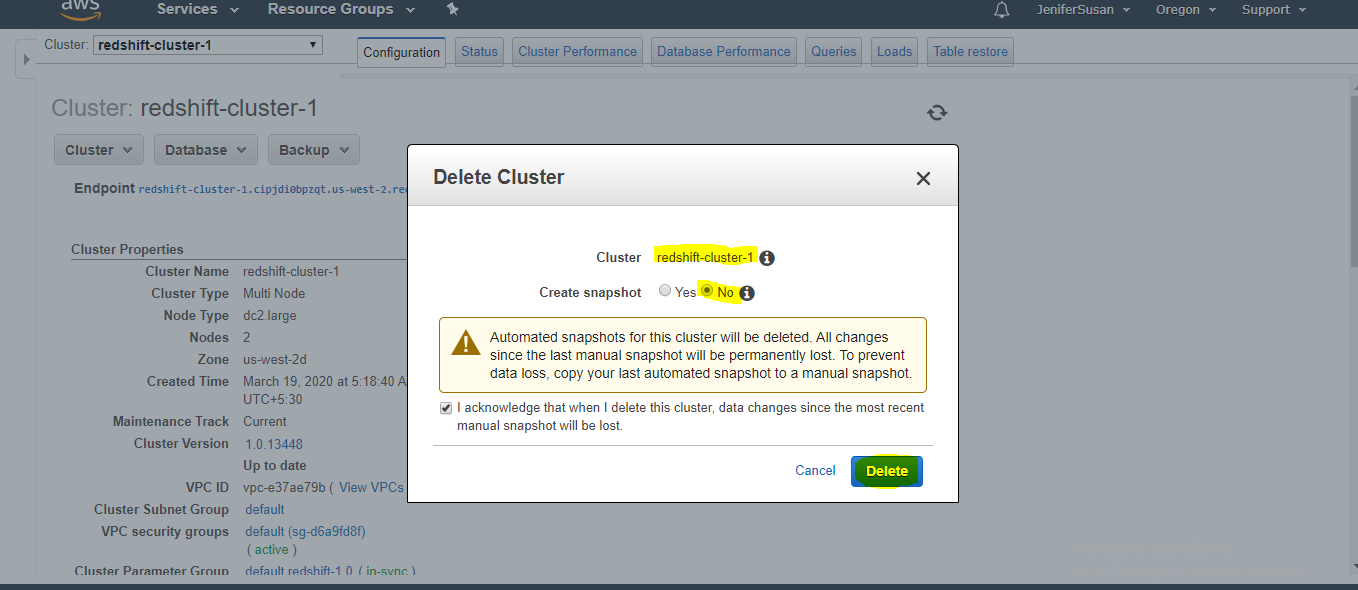
You can collect further information about the user for better analytical work to improve the business or decision in the organization.

# Step 7: Reset Your Environment

# Amazon Redshift costs around $0.25 per hour

To delete the cluster

1. On the navigation menu, choose CLUSTERS to display your list of clusters.
2. Choose the redshift-cluster-1 For Actions, choose Delete. The Delete cluster page appears.
3. Confirm the cluster to be deleted, then choose Delete cluster.



Interview questions

**1.What are the benefits of using AWS Redshift?**

Answer:

* We can run multiple queries on multiple nodes.
* We can use Postgresql, ODBC and JDBC.
* Automated backup
* Built-in security.
* When applications requires analytical function.
* Cost effective compared to traditional data warehousing technique.

#### What are the common features of Redshift?

AWS Redshift is a fully managed, petabyte-scale data warehouse service in the AWS, We can create a bunch of nodes once we have data warehouse created i.e AWS redshift Cluster. Now we can upload the set of data and perform the query for data analysis.

#### 3.What’s the use of redshift?

* Scalability up/down
* Pay for what ever we use

#### 4.Which is the database redshift is using?

Answer: Amazon Redshift is using PostgreSQL

#### 5.Why AWS redshift is named after Redshift?

Answer: Amazon Web Services given the name RedShift for Oracle trademark red.

#### 6.What is the difference between s3 and redshift?

Answer: AWS S3 is object based storage.  AWS Redshift is a fast, fully managed, petabyte-scale data warehouse.

#### 7.What is  cluster AWS cloud?

Answer: Cluster is grouping of similar services. You can create multiple clusters depends on the requirement on the services.

#### 8. Does redshift support unstructured data?

Answer: AWS Redshift is using  PostgreSQL supports only structured data.

#### 9.AWS Redshift is basically works on nodes and cluster and which of following option Redshift is based on?

Storage service  
database  
System Storage  
Answer: database

#### 10.What kind of application/services uses Redshift database?

Answer: Amazon Redshift is meant for services which are petabyte scale warehousing. Example: Big data analytics and OLAP. Redshift are fully managed and scalable in nature.

#### 11. What are the business intelligent tools to which Redshift can be integrated with?

Answer: Redshift can be integrated with Tableau, Jaspersoft, Microstrategy, Cognos Pentaho and Amazon QuickSight.

#### 12.How do we load data into Redshift?

Answer: Data is loaded from S3, DynamoDB,DMS and Read Replicas in RDS for example, when you have a RBS database but you want to do analytics on it to create a read replica,to pull that data from the read replica into Redshift and to do the analytics into Redshift.

#### 13.How many types of nodes supported by Redshift and what are the functions of nodes?

Answer: Redshift supports 2 nodes -leader node and compute node  
There is a leader node and the leader node is used to planning the queries and aggregate results across all compute nodes. So the compute nodes are going to actually be performing the queries and they will send the results back to the leader. If you have one node, then that nodes both a leader node and a compute node.