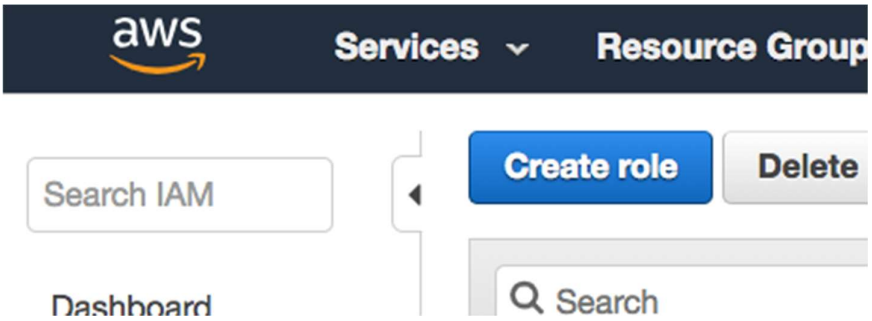
CLOUD Data Warehouse Setup

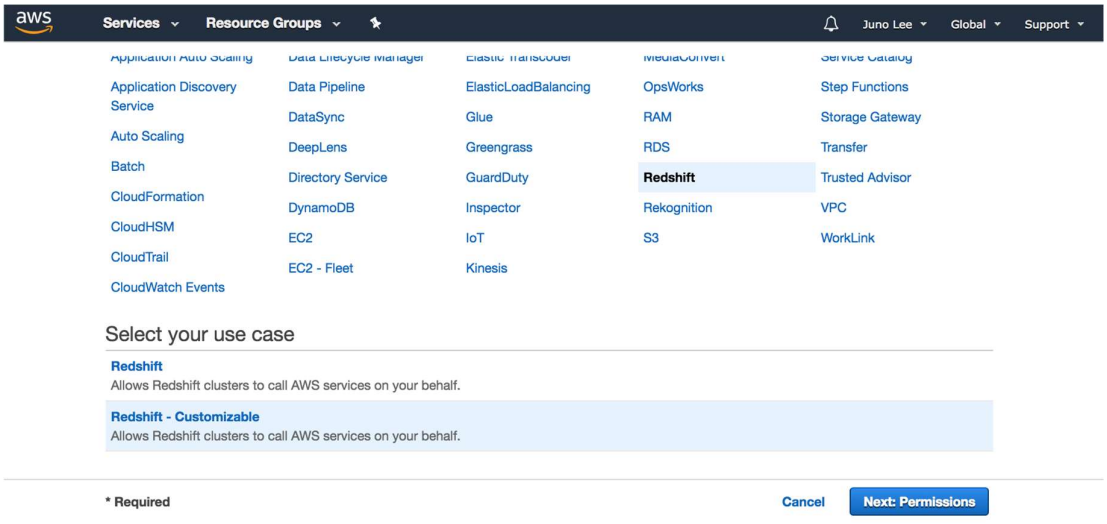
Create an IAM Role

Here, you'll create an IAM role that you will later attach to your Redshift cluster to enable your cluster to load data from Amazon S3 buckets. Read more about IAM roles and Redshift **here**.

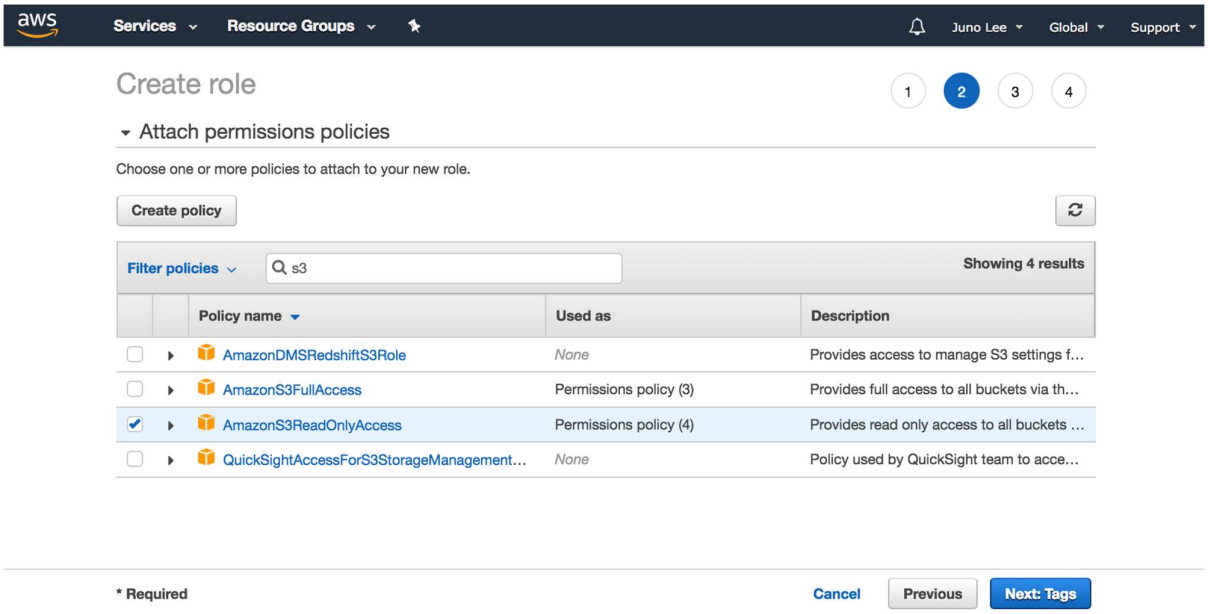
1. Sign in to the AWS Management Console and open the IAM console at **https://console.aws.amazon.com/iam/**.
2. In the left navigation pane, choose **Roles**.
3. Choose **Create role**.



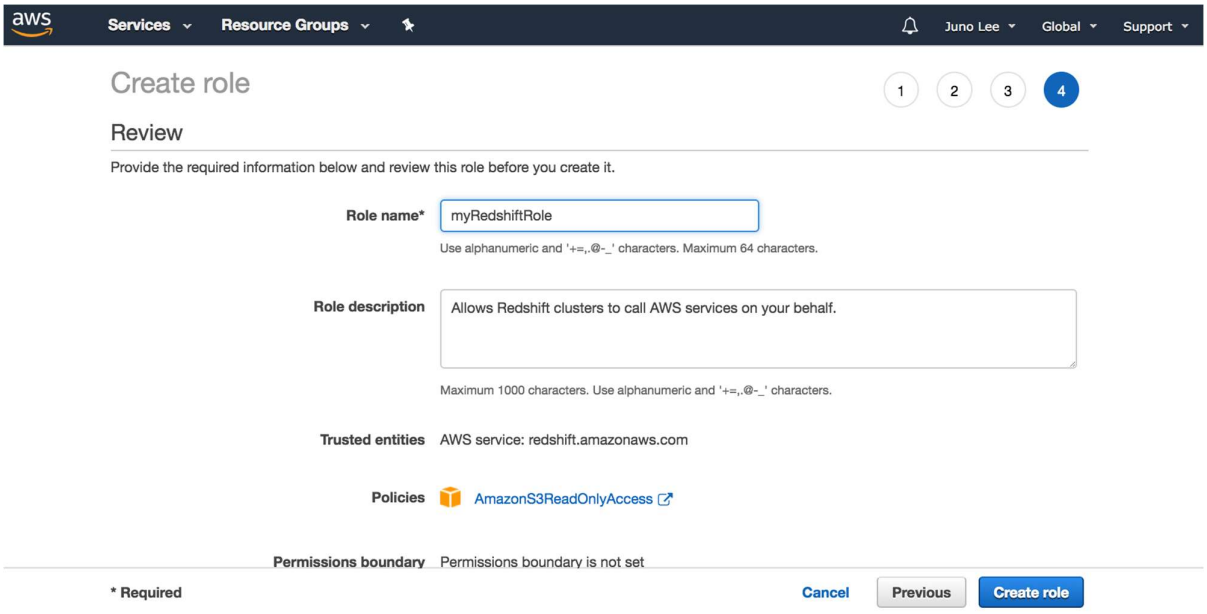
4. In the **AWS Service** group, choose **Redshift**.
5. Under **Select your use case**, choose **Redshift - Customizable**, and then **Next: Permissions**.



6. On the **Attach permissions policies** page, choose **AmazonS3ReadOnlyAccess**, and then choose **Next: Tags**.
7. Skip this page and choose **Next: Review**.



8. For **Role name**, enter `myRedshiftRole`, and then choose **Create Role**.

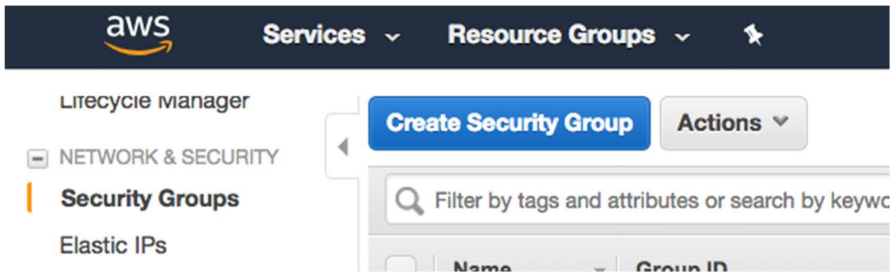


You can now attach this role when you launch a new cluster or attach it to an existing cluster. In the next page, you'll attach the role to a new cluster.

Create Security Group

Here, you'll create a security group you will later use to authorize access to your Redshift cluster.

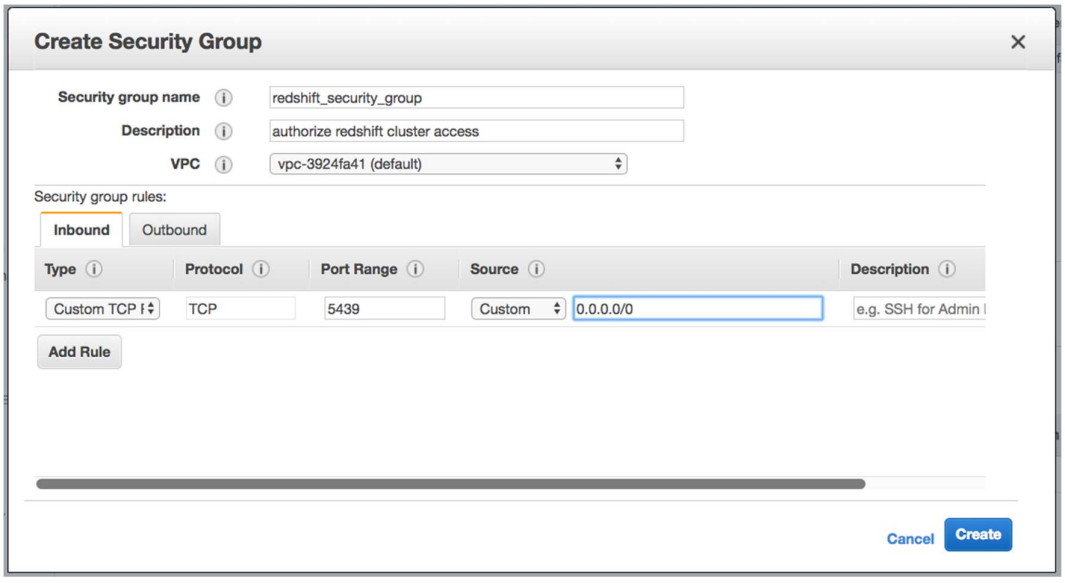
1. Go to your [Amazon EC2 console] and under **Network and Security** in the left navigation pane, select **Security Groups**.
2. Choose the **Create Security Group** button.



3. Enter `redshift_security_group` for **Security group name**.
4. Enter "authorize redshift cluster access" for **Description**.
5. Select the **Inbound** tab under **Security group rules**.
6. Click on **Add Rule** and enter the following values:

- **Type**: Custom TCP Rule.
- **Protocol**: TCP.
- **Port Range**: `5439`. The default port for Amazon Redshift is 5439, but your port might be different. See note on determining your firewall rules on the earlier "AWS Setup Instructions" page in this lesson.
- **Source**: select Custom IP, then type 0.0.0.0/0.
  Important: Using 0.0.0.0/0 is not recommended for anything other than demonstration purposes because it allows access from any computer on the internet. In a real environment, you would create inbound rules based on your own network settings.

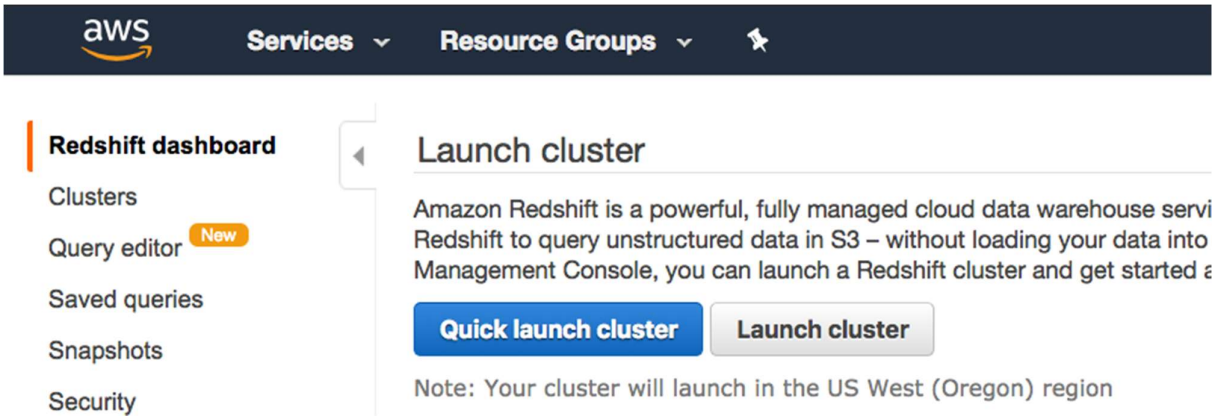
7. Choose **Create**.



Launch a Redshift Cluster

WARNING: The cluster that you are about to launch will be live, and you will be charged the standard Amazon Redshift usage fees for the cluster until you delete it. **Make sure to delete your cluster each time you're finished working to avoid large, unexpected costs for yourself.** Instructions on deleting your cluster are included on the last page in this lesson. You can always launch a new cluster, so don't leave your Redshift cluster running overnight or throughout the week if you don't need to.

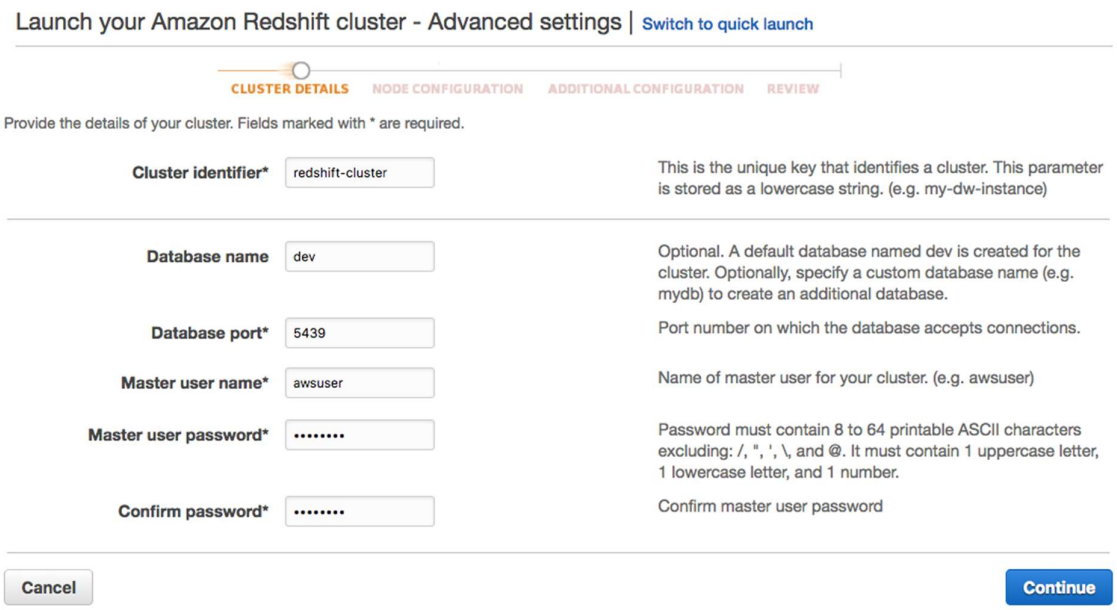
1. Sign in to the AWS Management Console and open the Amazon Redshift console at **https://console.aws.amazon.com/redshift/**.
2. On the Amazon Redshift Dashboard, choose **Launch cluster**.



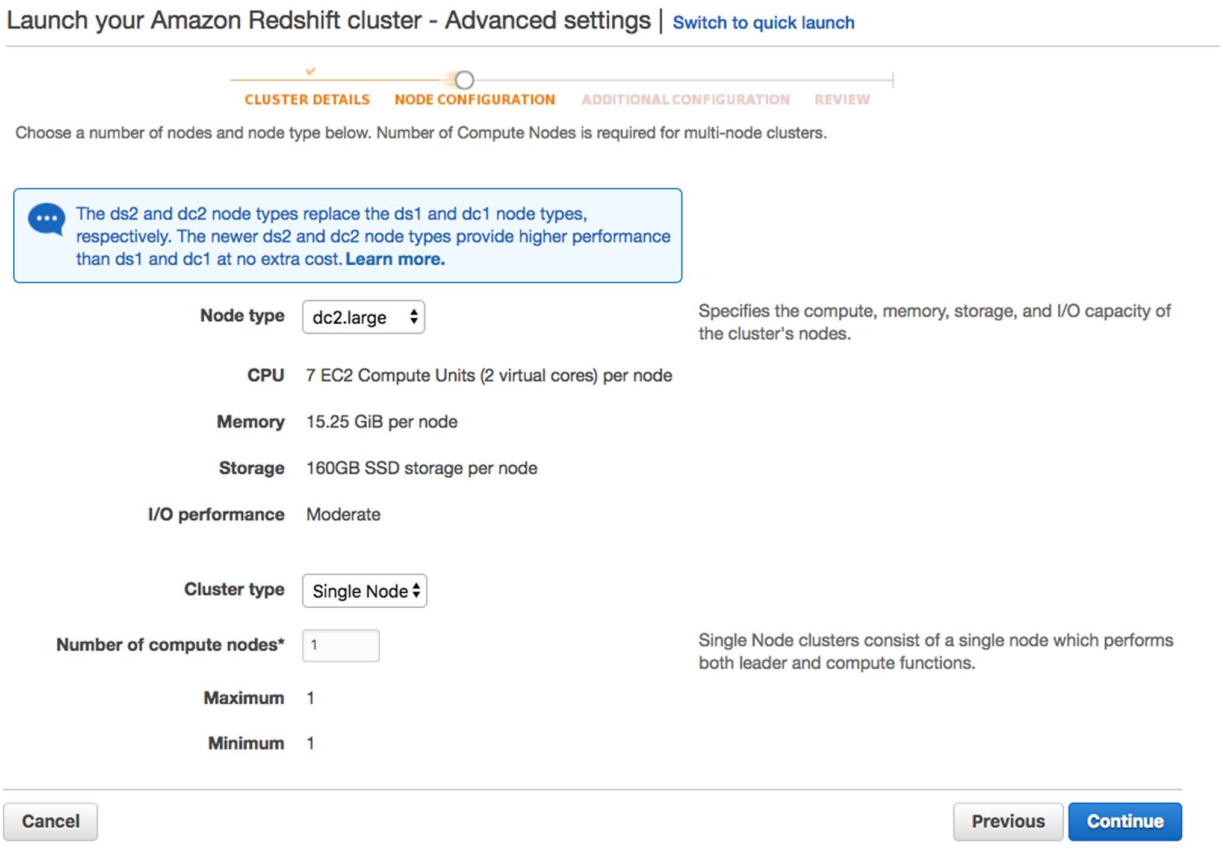
3. On the Cluster details page, enter the following values and then choose Continue:

- **Cluster identifier**: Enter `redshift-cluster`.
- **Database name**: Enter `dev`.
- **Database port**: Enter `5439`.
- **Master user name**: Enter `awsuser`.
- **Master user password** and **Confirm password**: Enter a password for the master user account.

Please note: We **strongly advise** you to keep these passwords closely guarded, including not putting them in your GitHub public repo, etc.

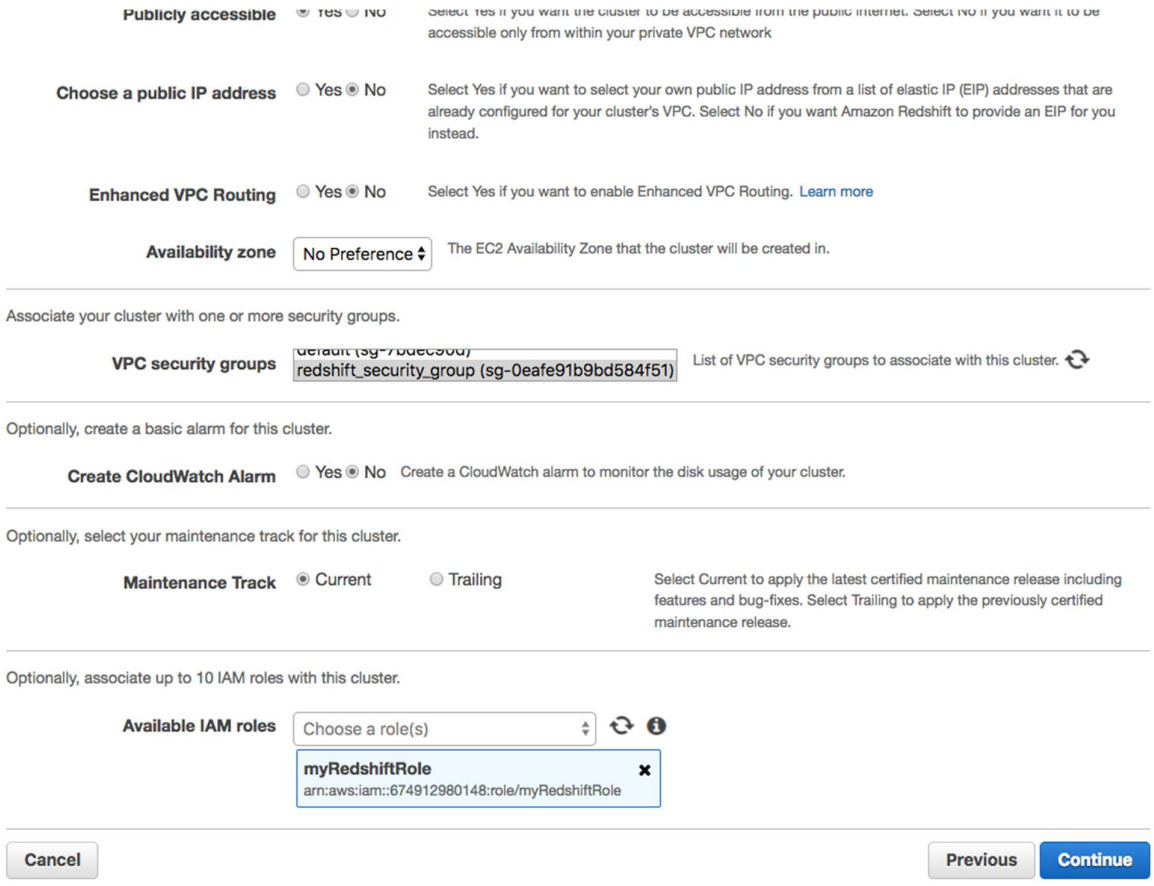


4. On the Node Configuration page, accept the default values and choose **Continue**.

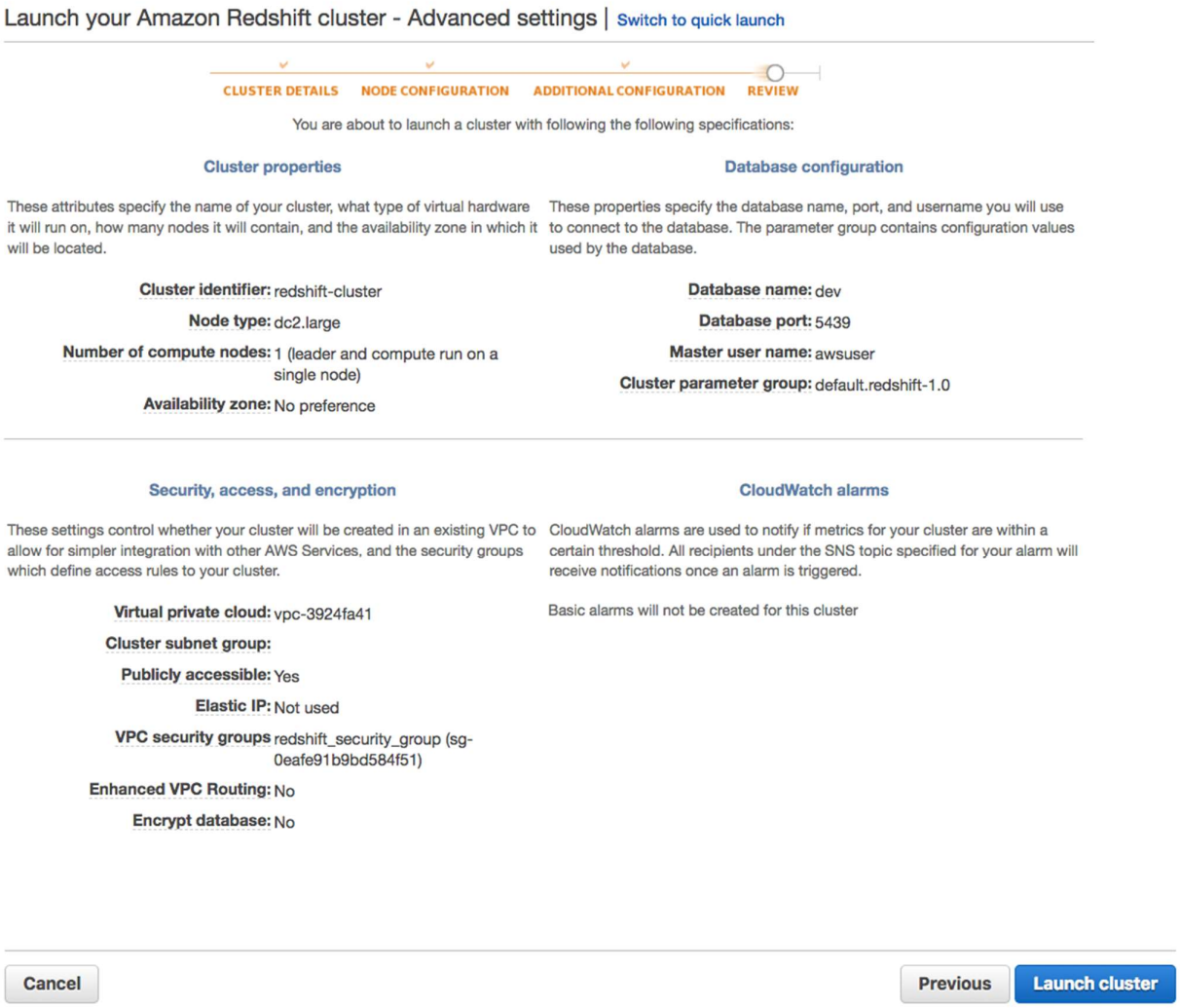


5. On the Additional Configuration page, enter the following values:

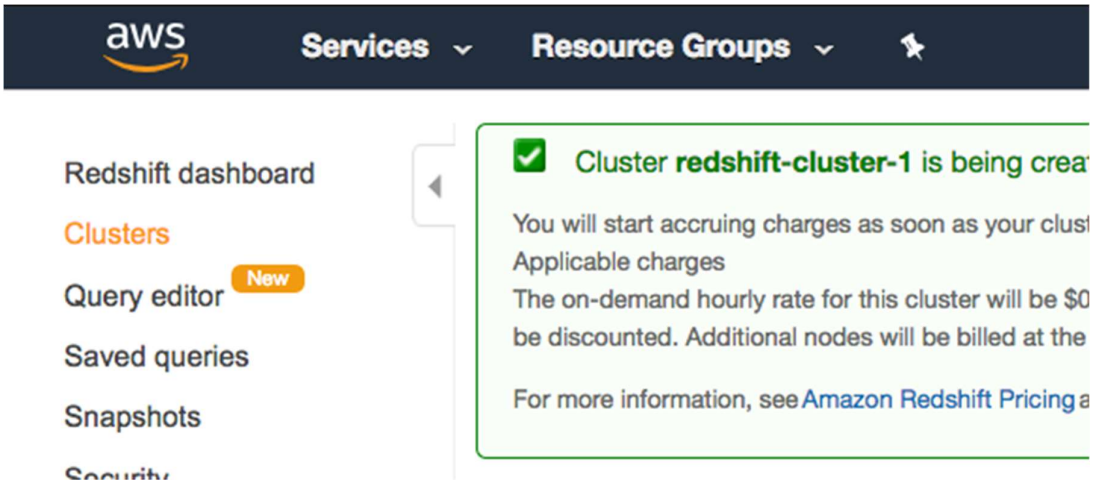
- **VPC security groups**: redshift_security_group
- **Available IAM roles**: myRedshiftRole
  Choose **Continue**.



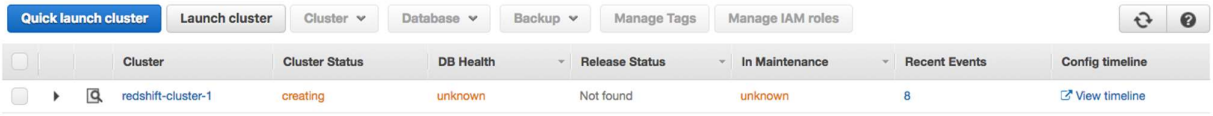
6. Review your Cluster configuration and choose **Launch cluster**.

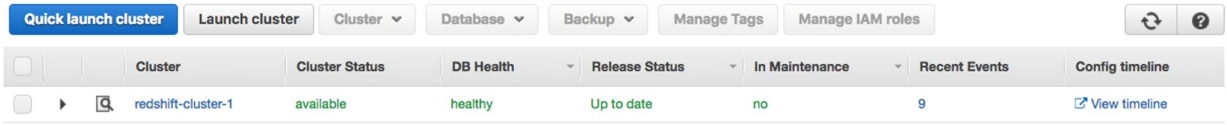


7. A confirmation page will appear and the cluster will take a few minutes to finish. Choose **Clusters** in the left navigation pane to return to the list of clusters.



8. On the Clusters page, look at the cluster that you just launched and review the **Cluster Status** information. Make sure that the **Cluster Status** is **available** and the **Database Health** is **healthy** before you try to connect to the database later. You can expect this to take 5-10 minutes.

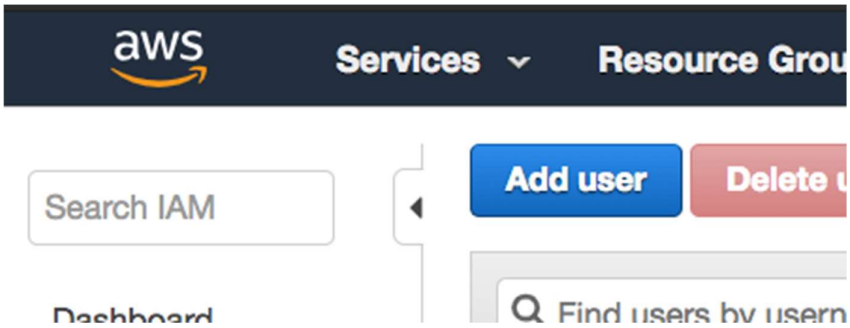




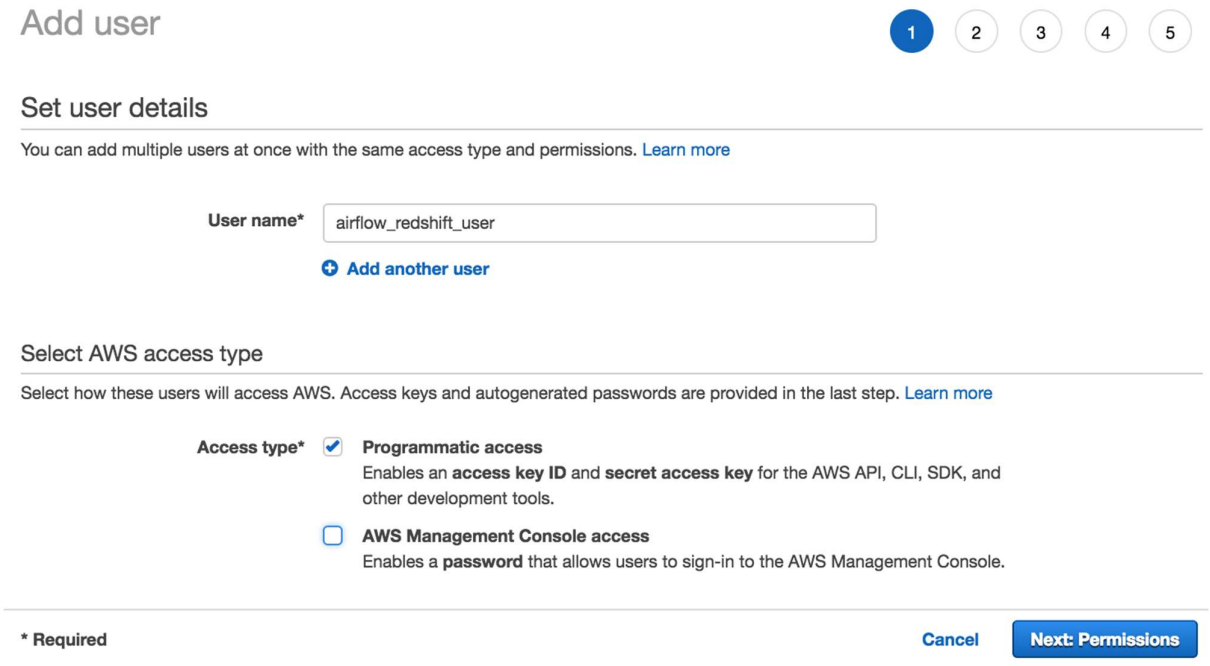
Create an IAM User

Here, you'll create an IAM user that you will use to access your Redshift cluster.

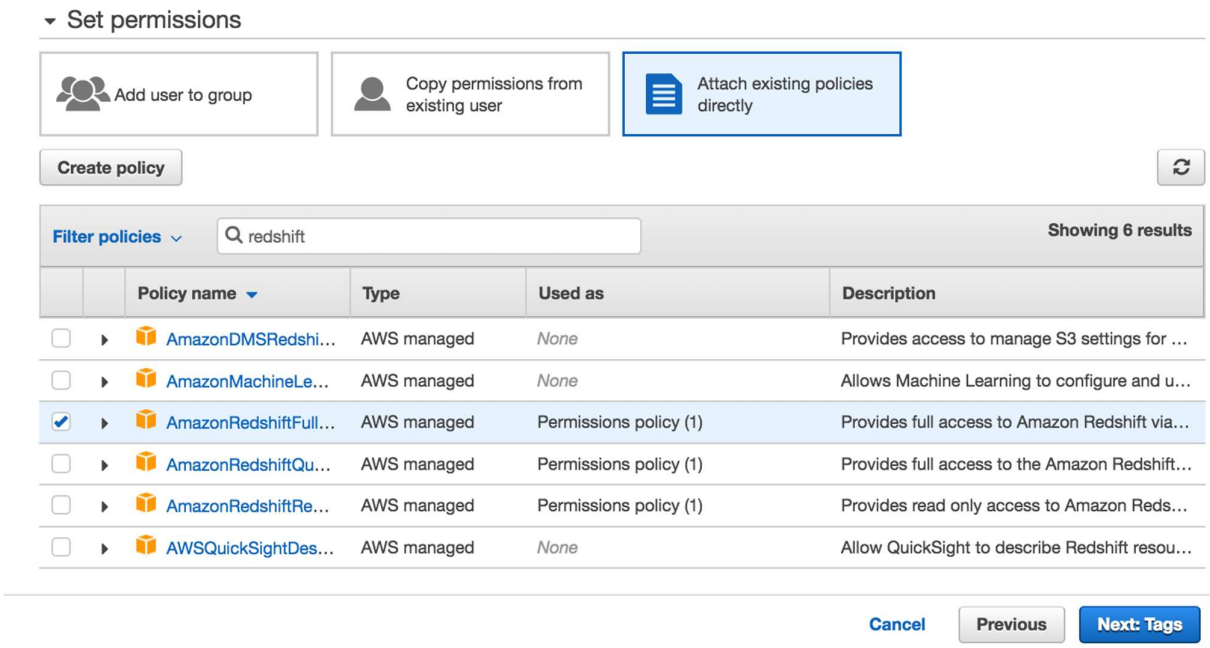
1. Sign in to the AWS Management Console and open the IAM console at **https://console.aws.amazon.com/iam/**.
2. In the left navigation pane, choose **Users**.
3. Choose **Add User**.

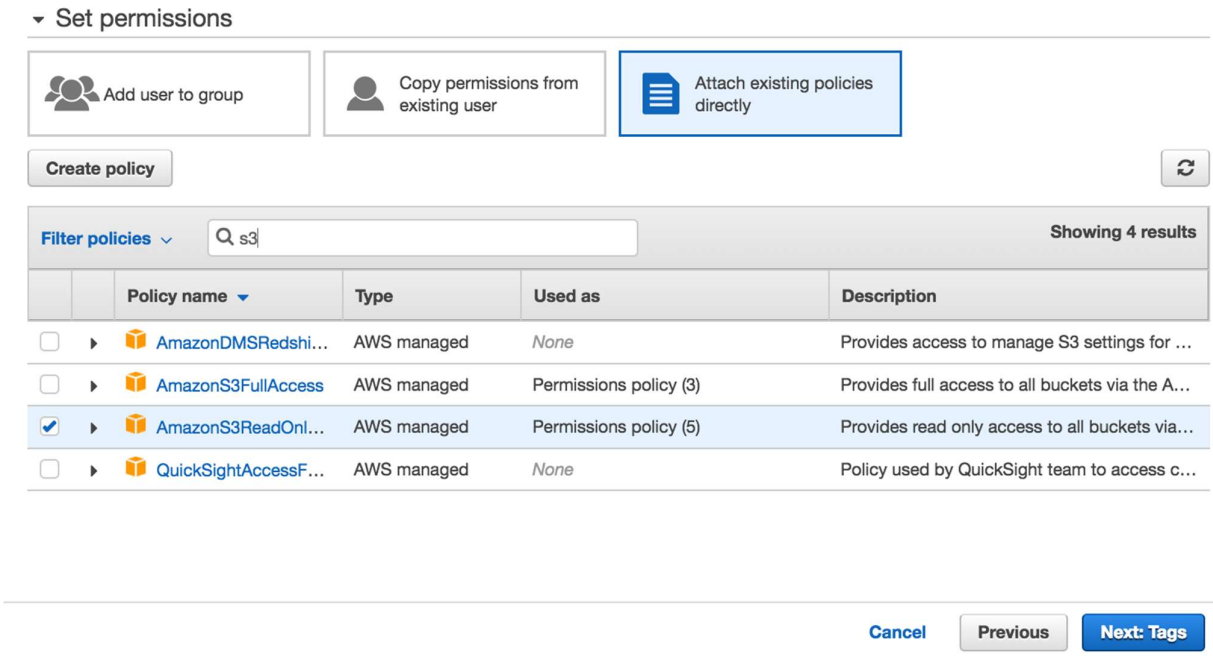


4. Enter a name for your user (e.g. `airflow_redshift_user`)
5. Choose **Programmatic access**, then choose **Next: Permissions**.

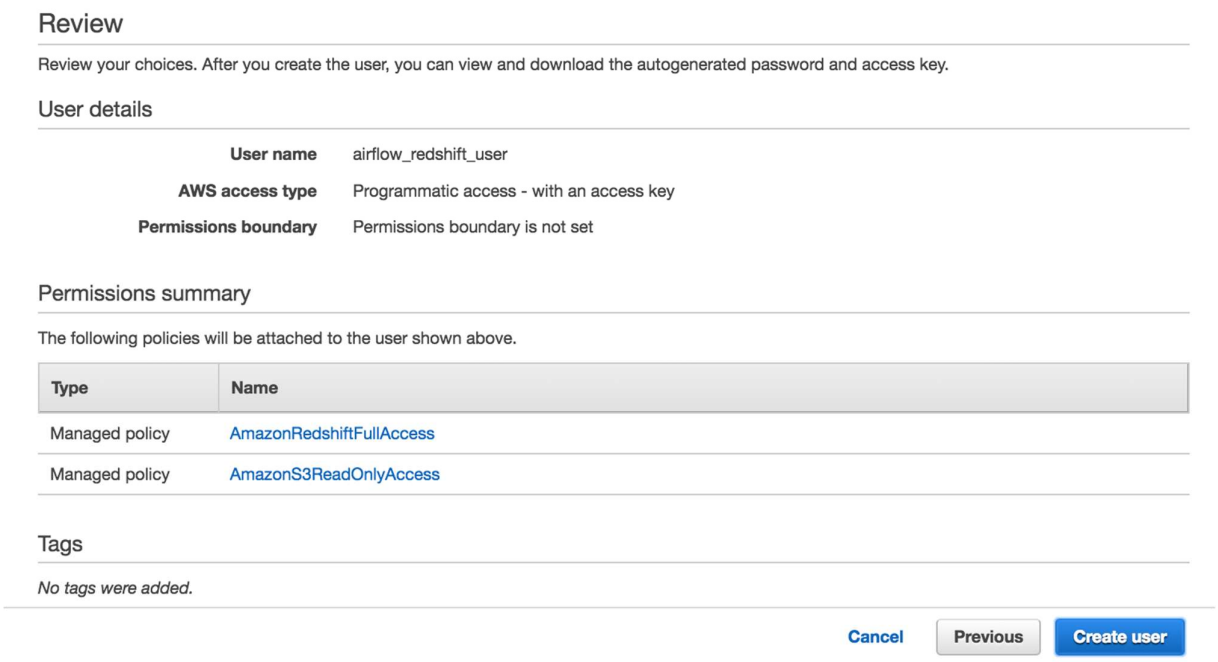


6. Choose **Attach existing policies directly**.
7. Search for redshift and select **AmazonRedshiftFullAccess**. Then, search for S3 and select **AmazonS3ReadOnlyAccess**. After selecting both policies, choose **Next: Tags**.
8. Skip this page and choose **Next: Review**.

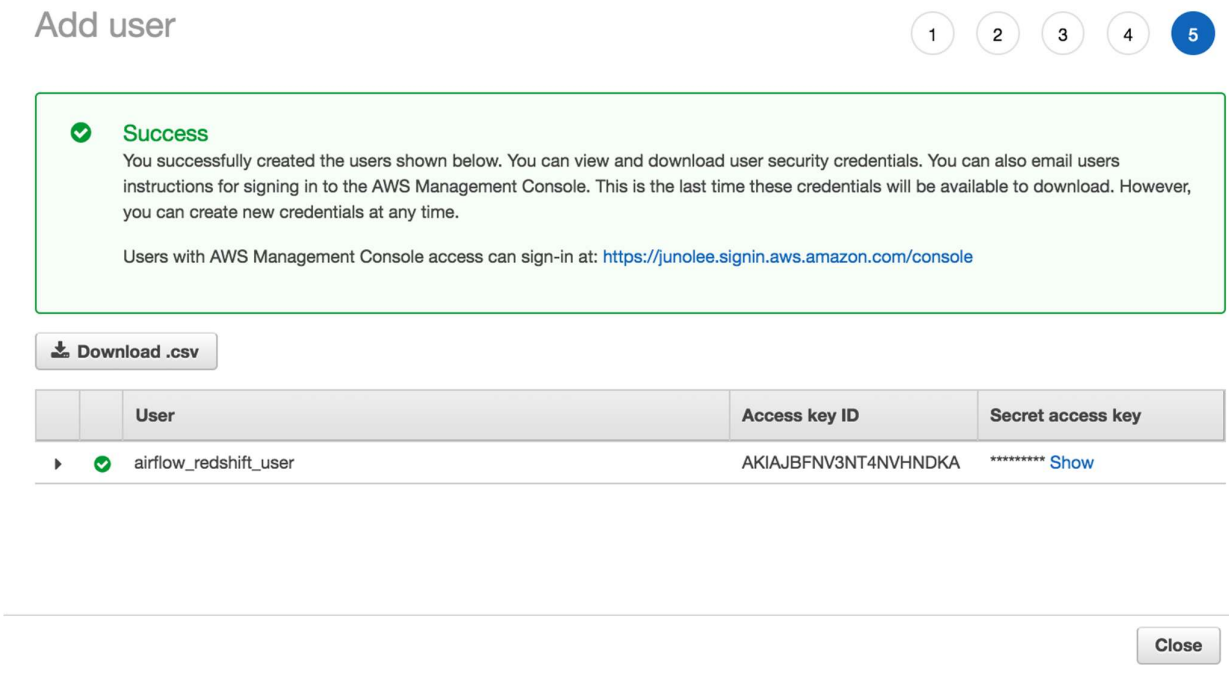




9. Review your choices and choose **Create user**.



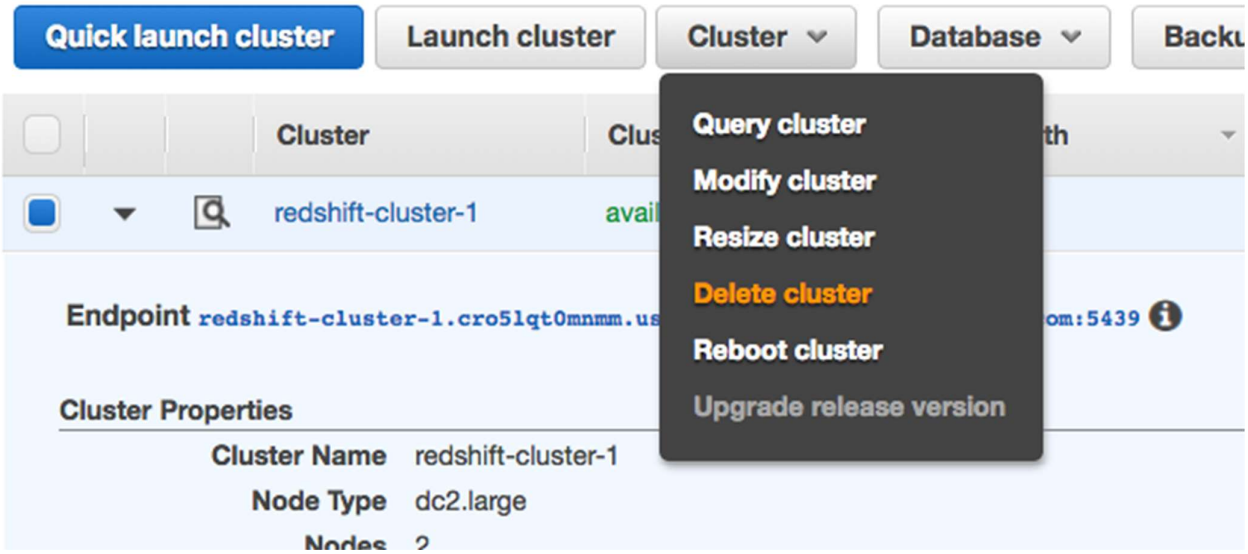
10. **Save your credentials!** This is the only time you can view or download these credentials on AWS. Choose **Download .csv** to download these credentials and then save this file to a safe location. You'll need to copy and paste this **Access key ID** and **Secret access key** in the next step. We **strongly advise** you to keep this **Access key ID** and **Secret access key** closely guarded, including not putting them in a GitHub public repo, etc.



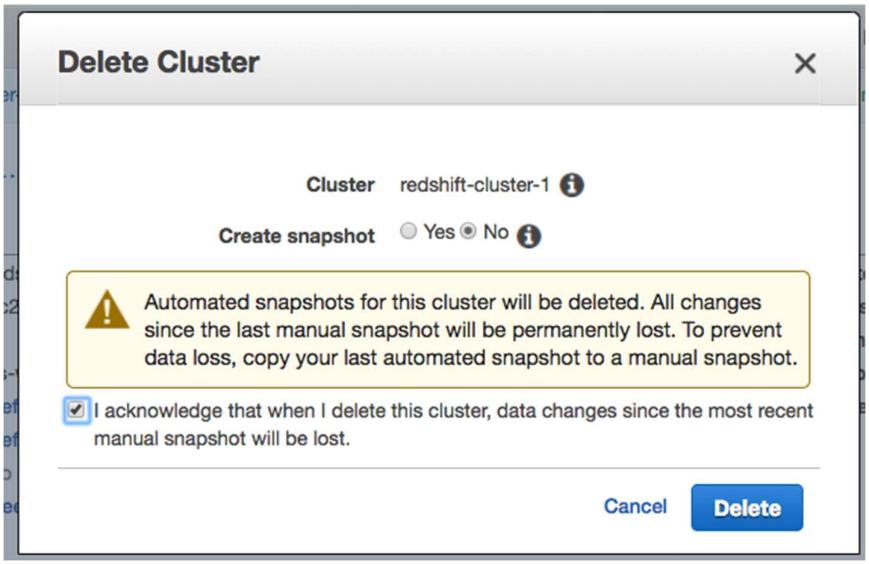
Delete a Redshift Cluster

Make sure to delete your cluster each time you're finished working to avoid large, unexpected costs. You can always launch a new cluster, so don't leave it running overnight or throughout the week if you don't need to.

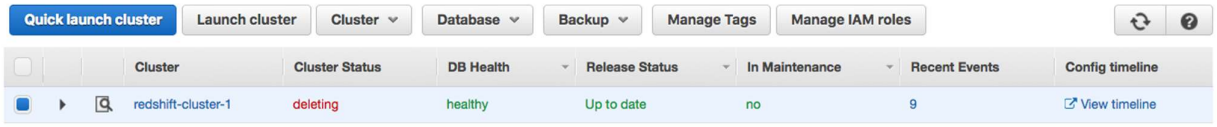
1. On the **Clusters** page of your Amazon Redshift console, click on the box next to your cluster to select it, and then click on **Cluster** > **Delete cluster**.



2. You can choose **No** for **Create snapshot**, check the box that you acknowledge this, and then choose **Delete**.

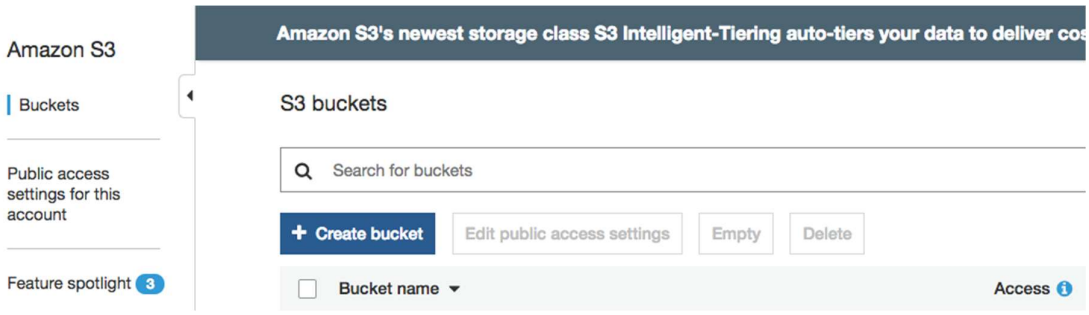


3. Your cluster will change it's status to **deleting**, and then disappear from your Cluster list once it's finished deleting. You'll no longer be charged for this cluster.



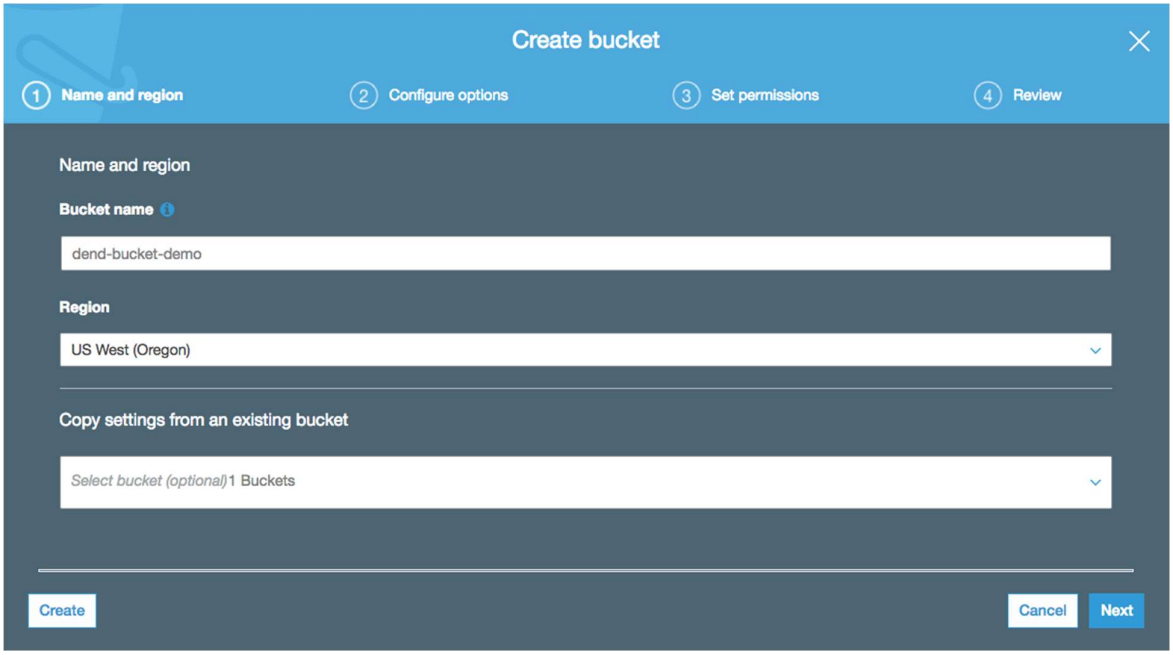
Create an S3 Bucket

1. Go to the **Amazon S3 Console** and select **Create bucket**.

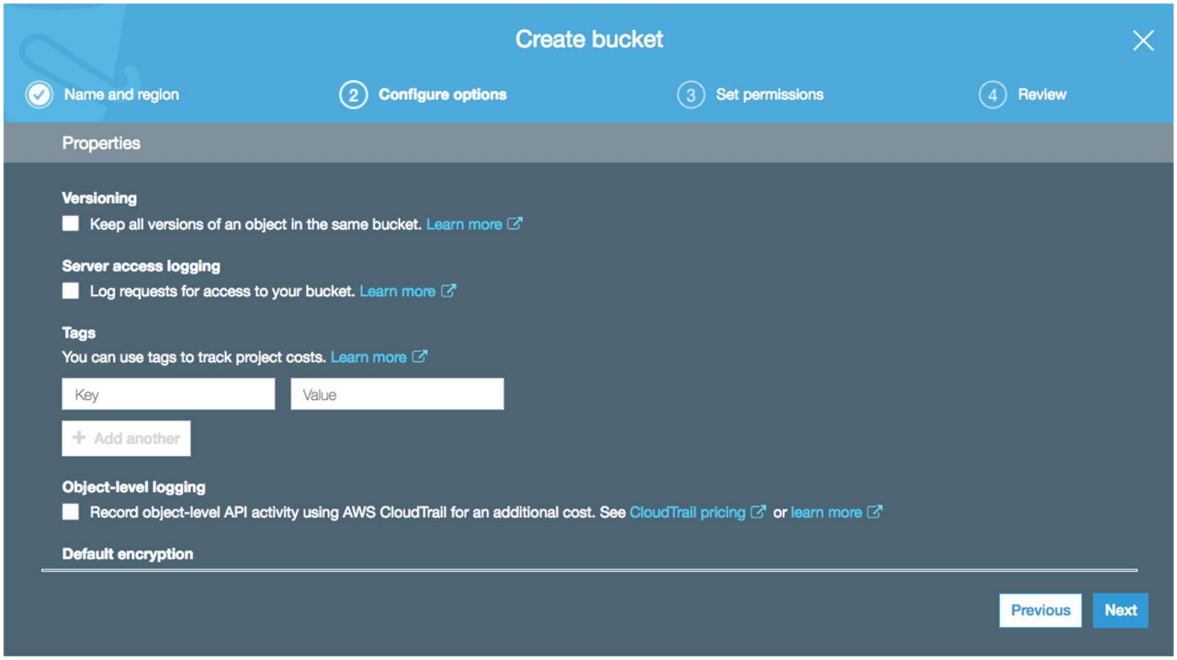


2. Enter a name for your bucket and select the region you'd like to create it in.

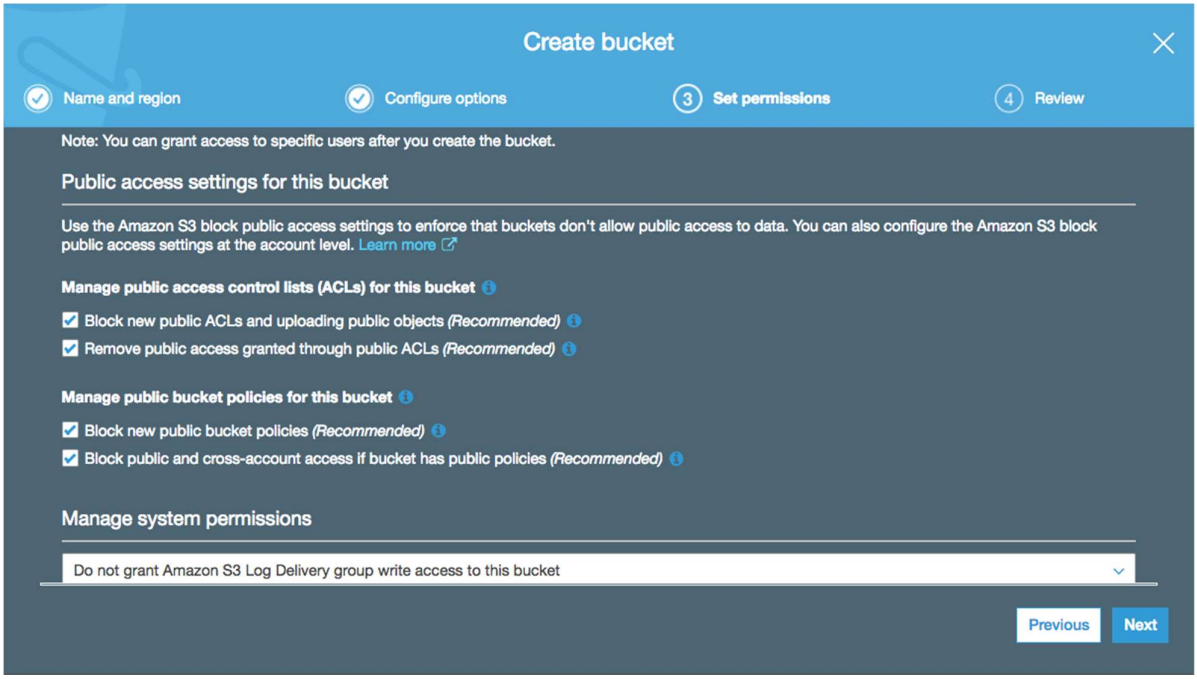
You won't be able to use the same name entered in the screenshot below - the names of all existing buckets on Amazon S3 are unique. You won't be able to change this name later, so choose one that makes sense for the content you'll have in it. This bucket name will be included in any URLs pointing to objects you add in your bucket.



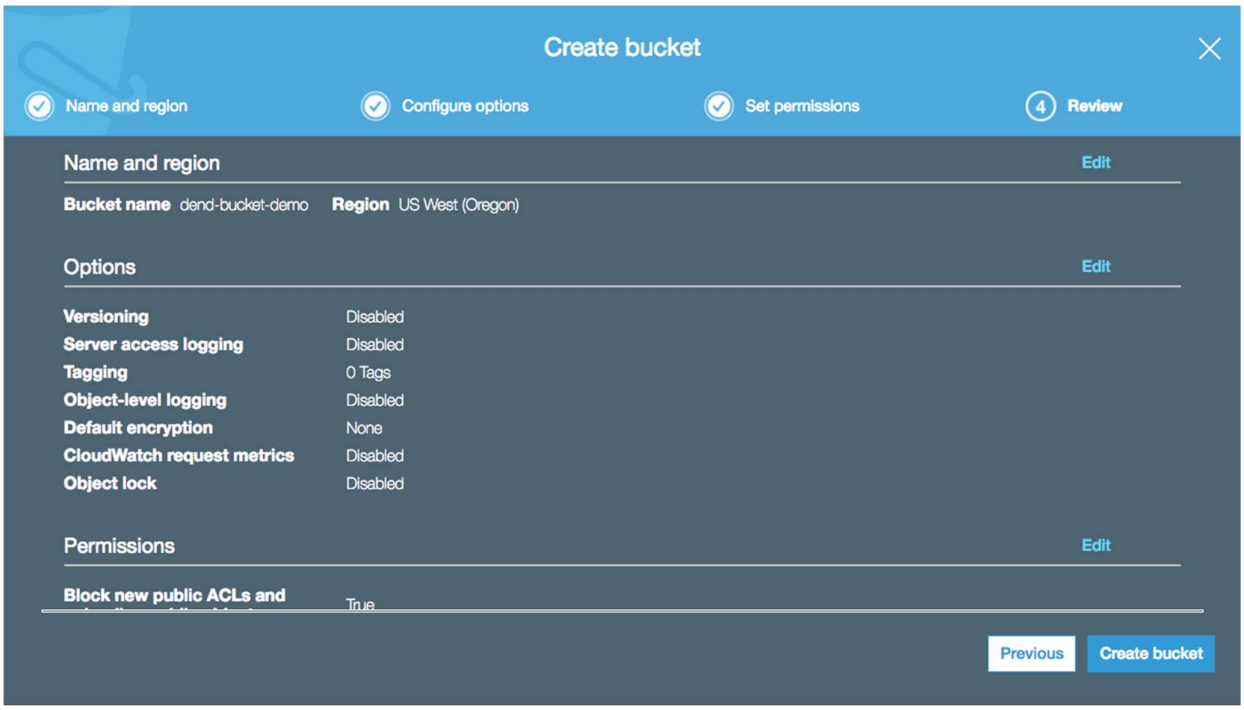
3. Keep the default settings and select **Next**.



4. Specify **public access settings** for this bucket. For example, unchecking all of these boxes would allow anyone to be able to access this bucket. Be careful with this - you may end up having to pay lots of fees in data transfers from your bucket if you share this link and many people access large amounts of data with it. For this demo, we will leave all of these boxes checked.



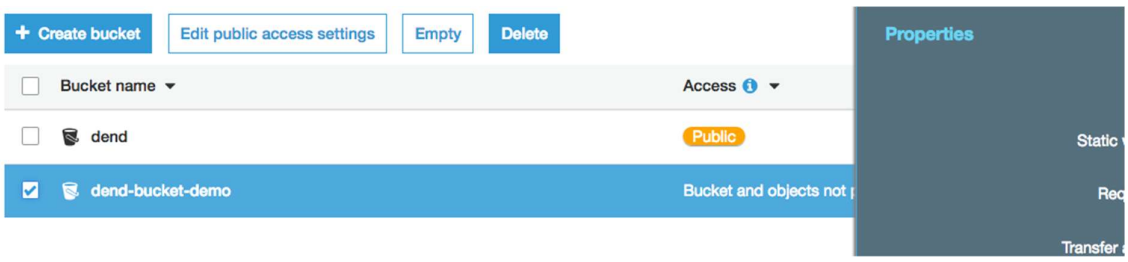
5. Review your settings and select **Create bucket**.



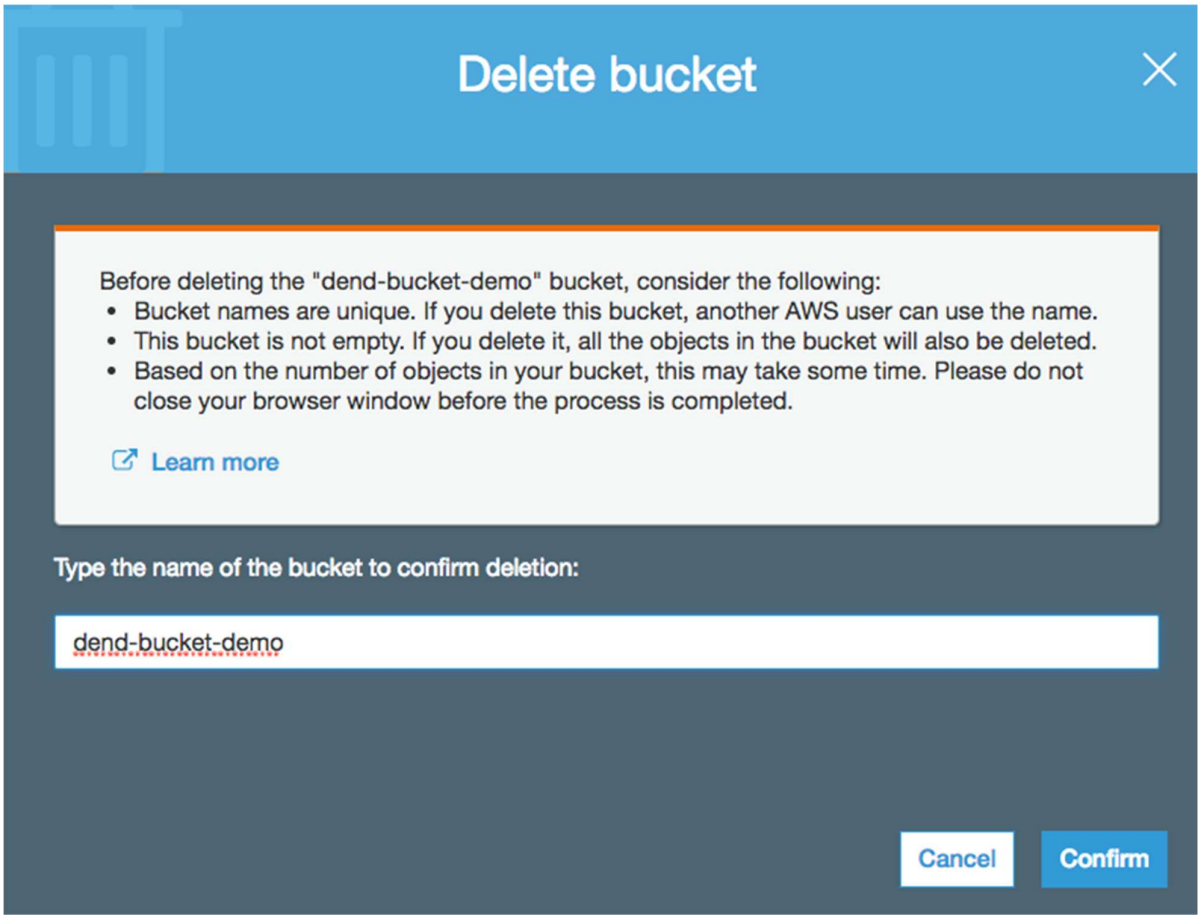
6. You should now see your bucket in your list. Any buckets you've made public will be labeled so here.

Delete S3 Bucket

To delete the bucket, select the bucket and click **Delete**.



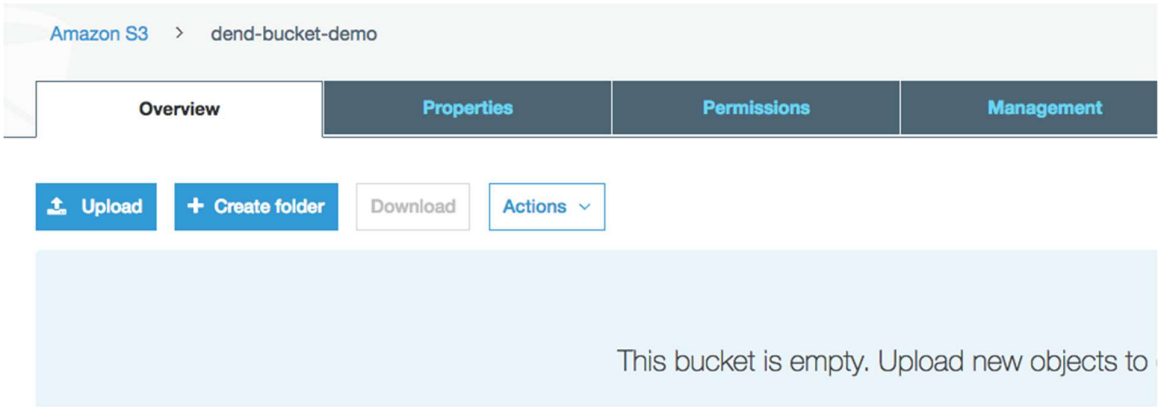
Then, enter the name of the bucket and select **Confirm**.



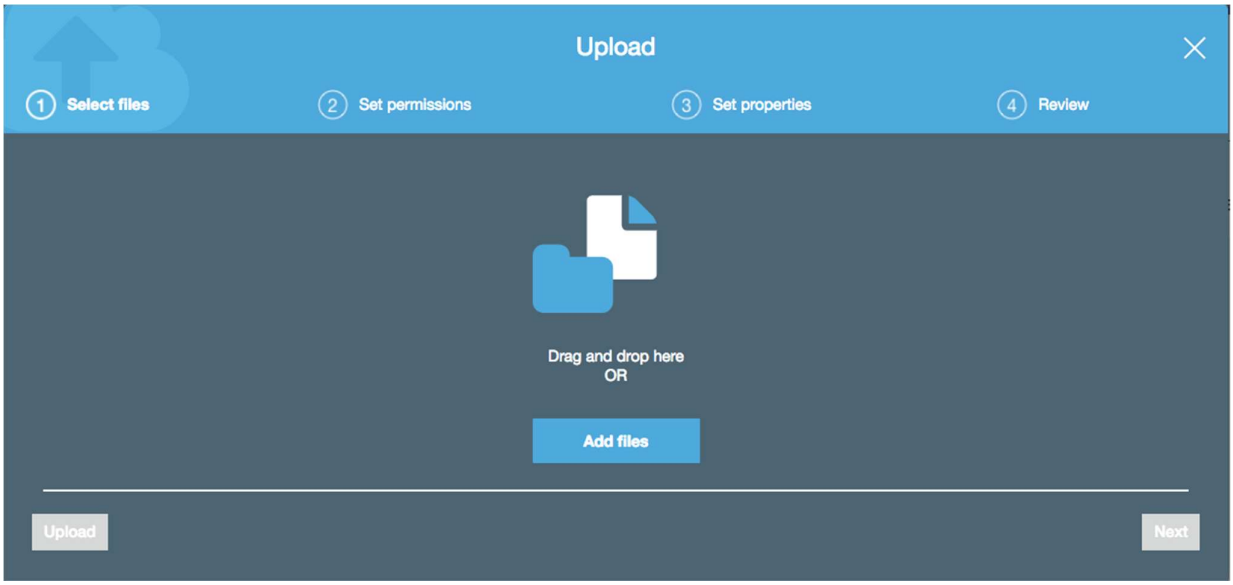
Upload to S3 Bucket

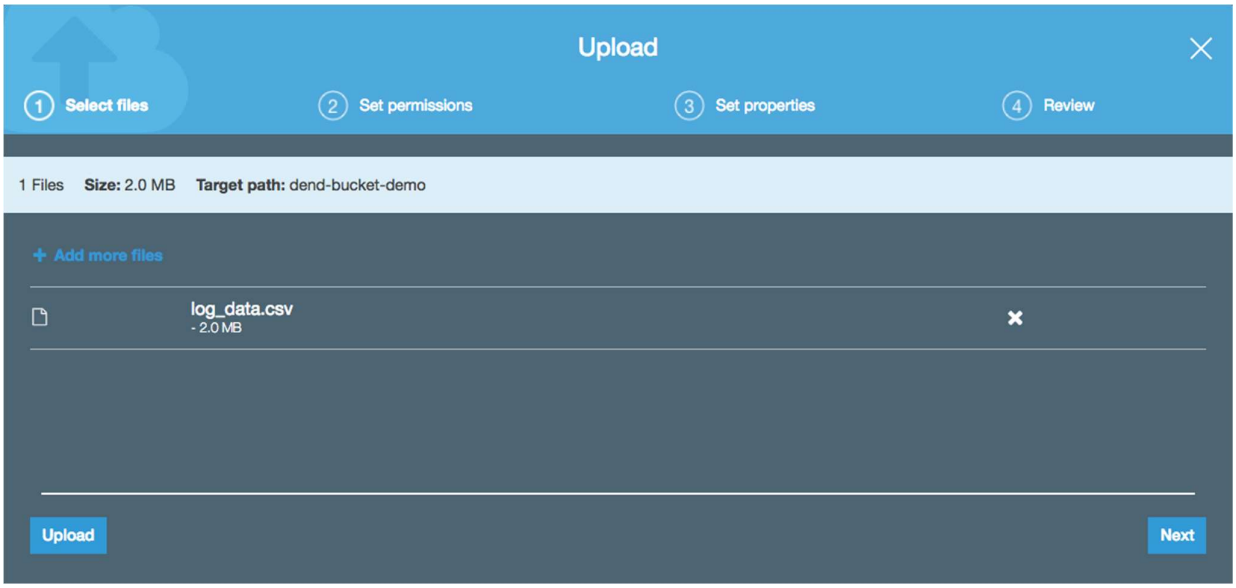
For this exercise, download and upzip `log_data` from the Resources tab in the classroom.

1. Click on the bucket you just created and select **Upload** on the top left.

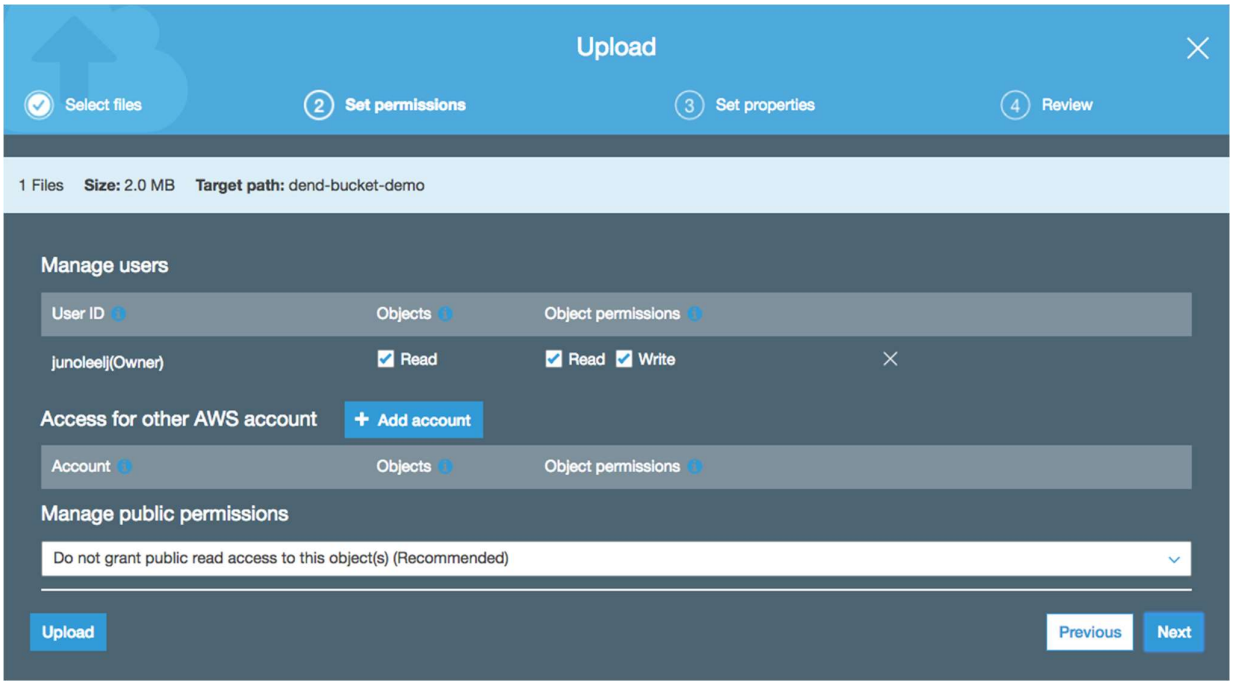


2. Select **Add files** and choose the `log_data.csv` file you downloaded from this classroom.

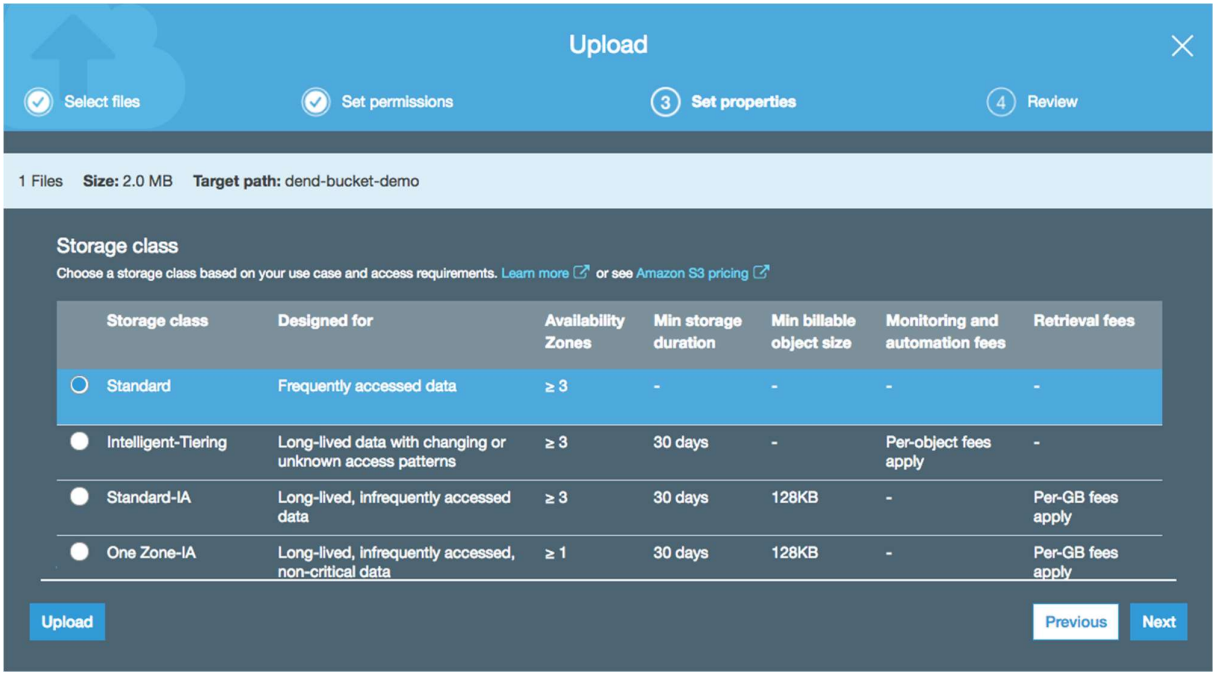




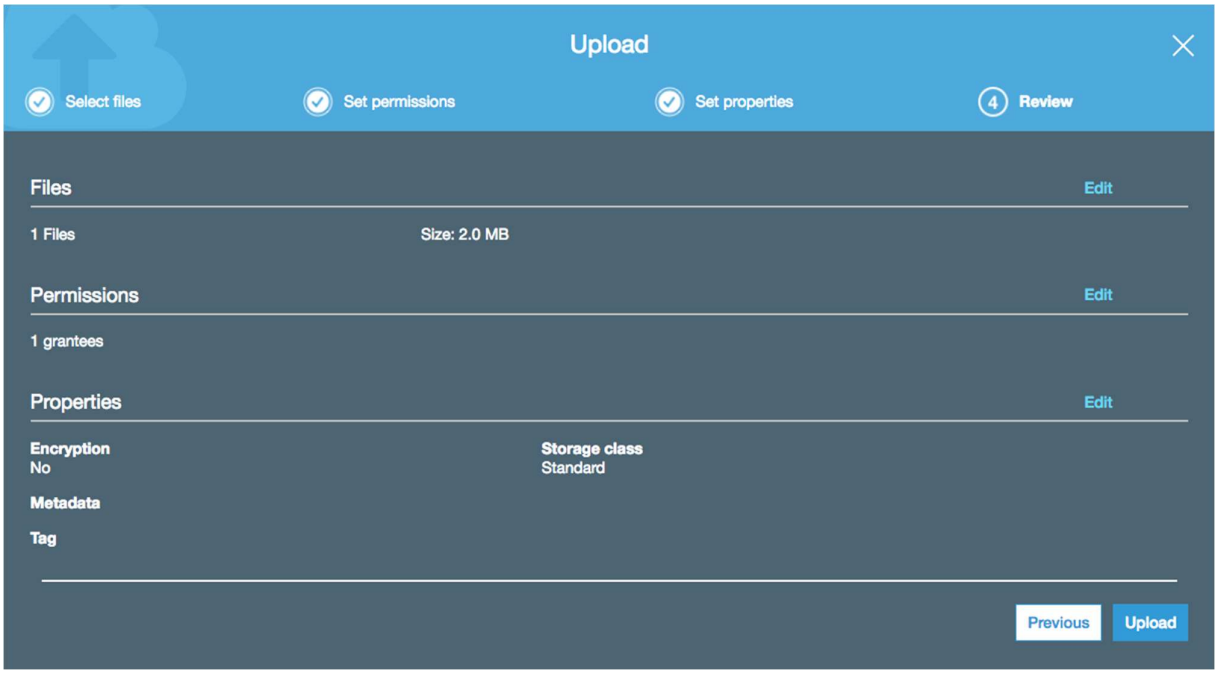
3. Leave the default settings and select **Next**.



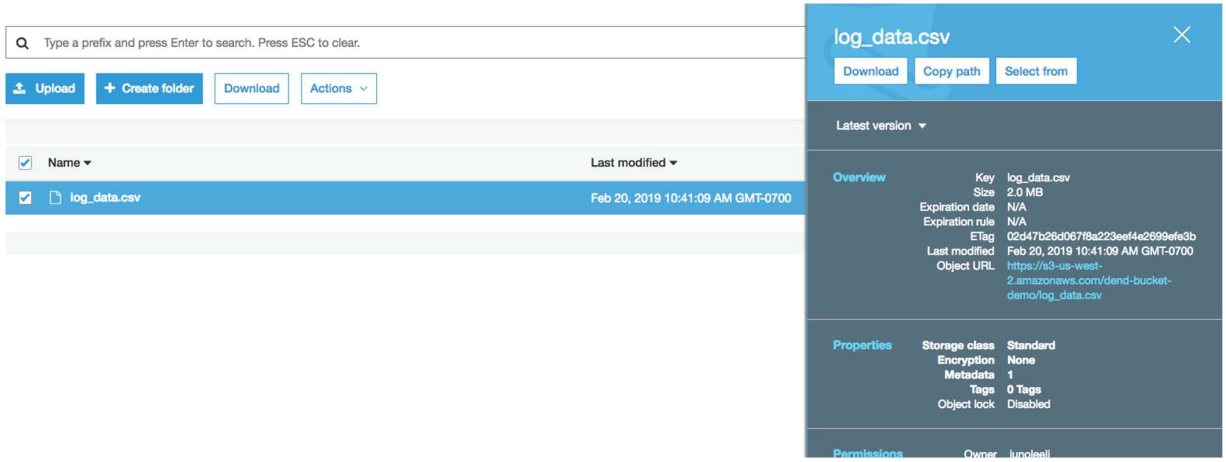
4. Choose the **Standard** storage class and select **Next**.



5. Review your settings and select **Upload**.

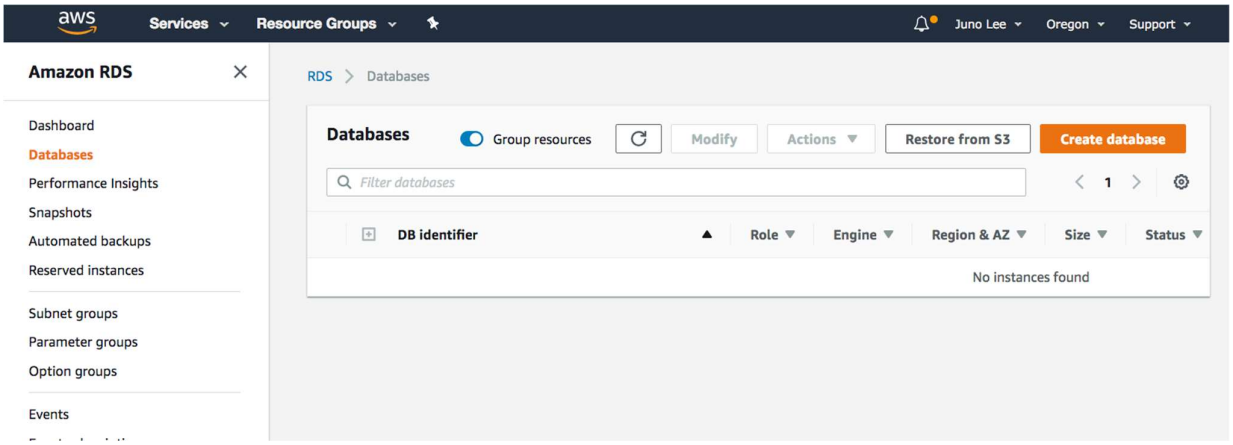


6. You should be able to see details on the file you just uploaded by selecting it in your bucket.

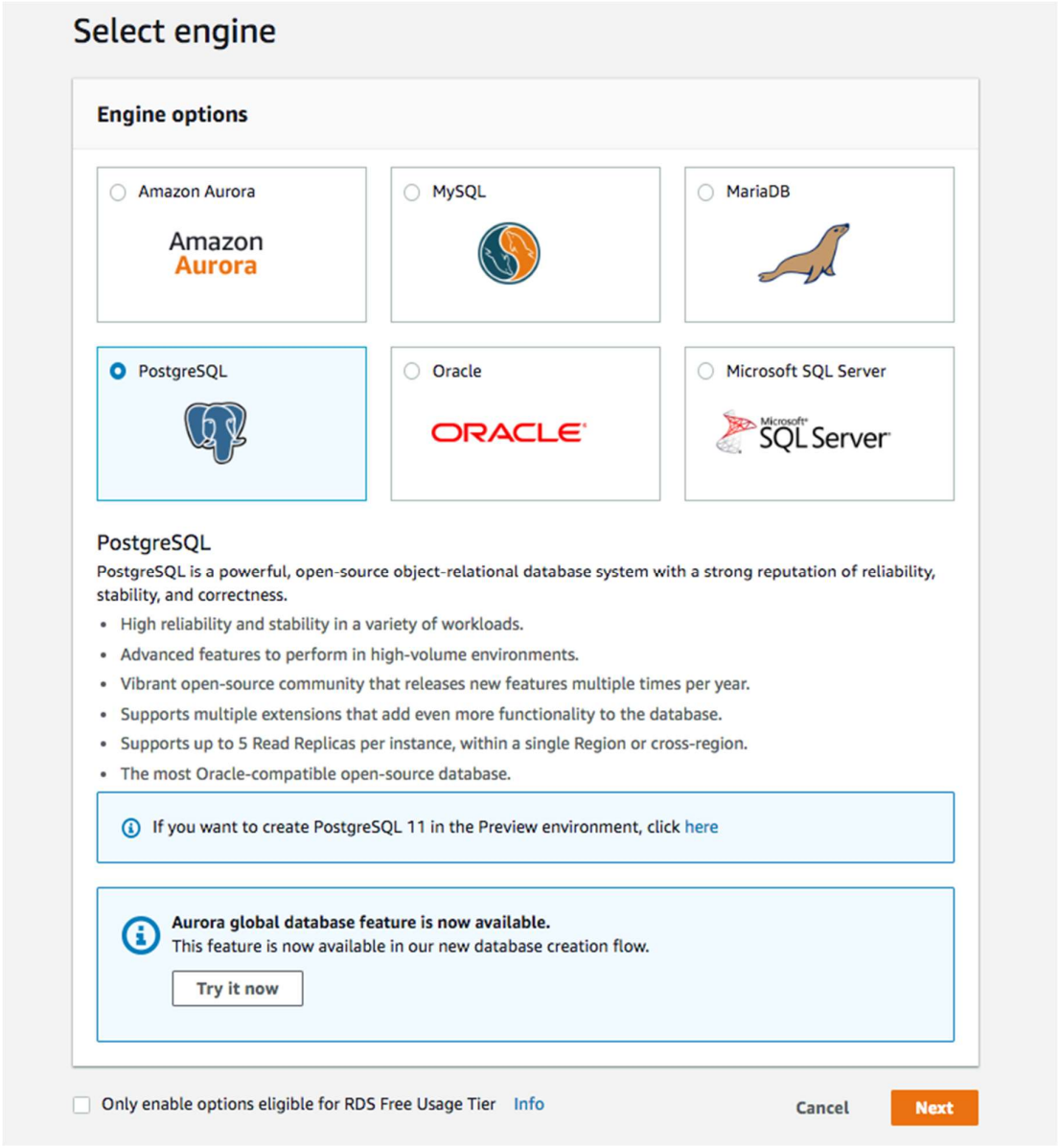

Type a prefix and press Enter to search. Press ESC to clear.
Upload
Create folder
Download
Actions
Name
Last modified
log_data.csv
Feb 20, 2019 10:41:09 AM GMT-0700
log_data.csv
Download
Copy path
Select from
Latest version
Overview
Key log_data.csv
Size 2.0 MB
Expiration date N/A
Expiration rule N/A
ETag 02d47b26d067f8a223eef4e2699efe3b
Last modified Feb 20, 2019 10:41:09 AM GMT-0700
Object URL https://s3-us-west-2.amazonaws.com/dend-bucket-demo/log_data.csv
Properties
Storage class Standard
Encryption None
Metadata 1
Tags 0 Tags
Object lock Disabled
Permissions
Owner

Create a PostgreSQL DB Instance using RDS

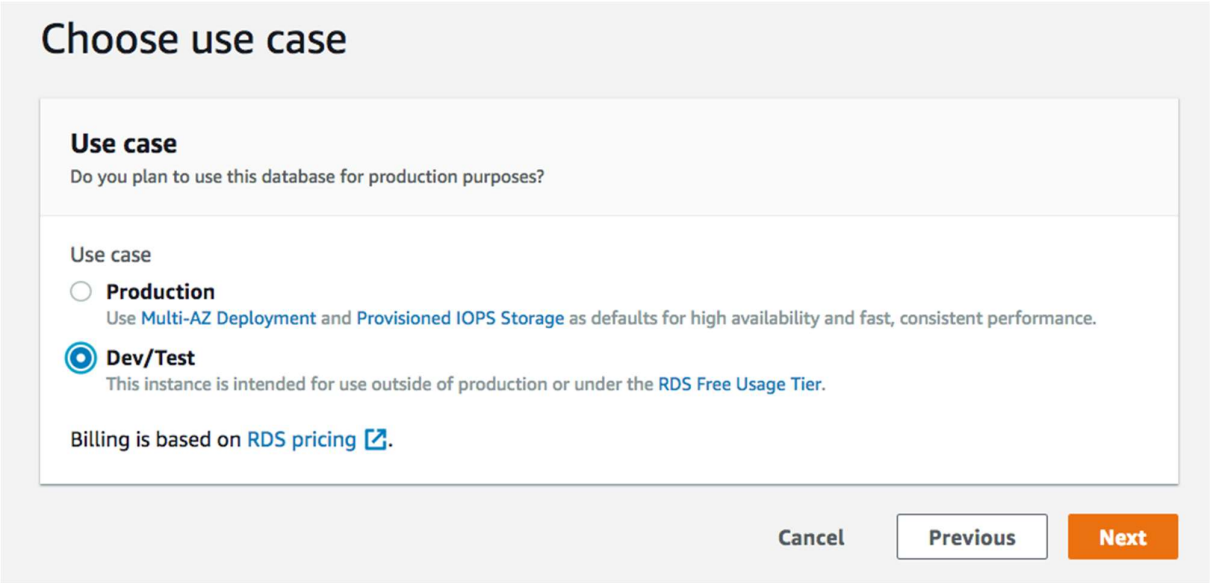
1. Go to the **Amazon RDS console** and click on **Databases** on the left navigation pane. Choose what region you'd like to create this database in on the right of the top menu bar.
2. Click on the **Create Database** button.



3. Select **PostgreSQL** on the Select Engine page.

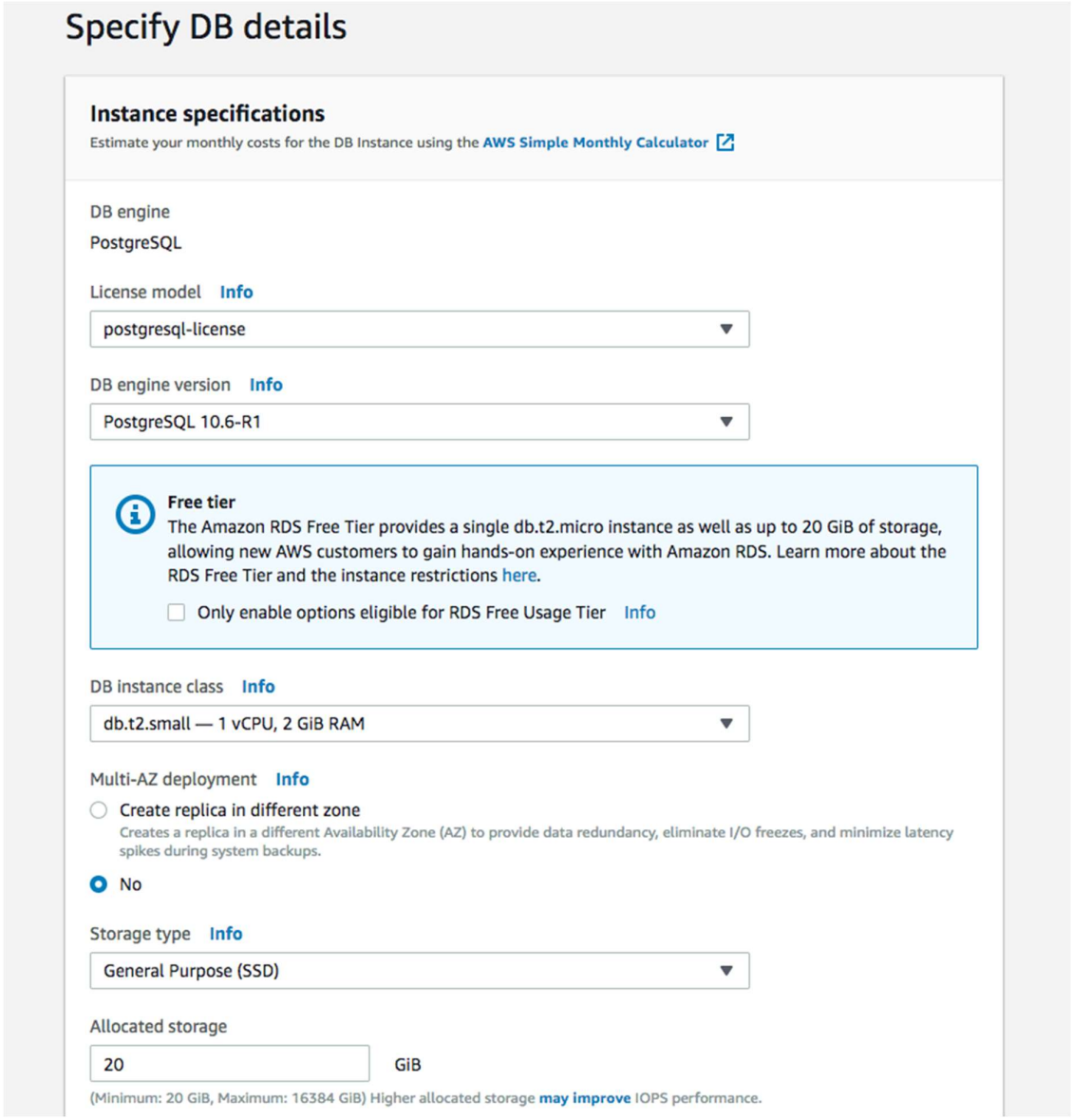


4. Since this is for demonstration purposes, select **Dev/Test** under Use case.

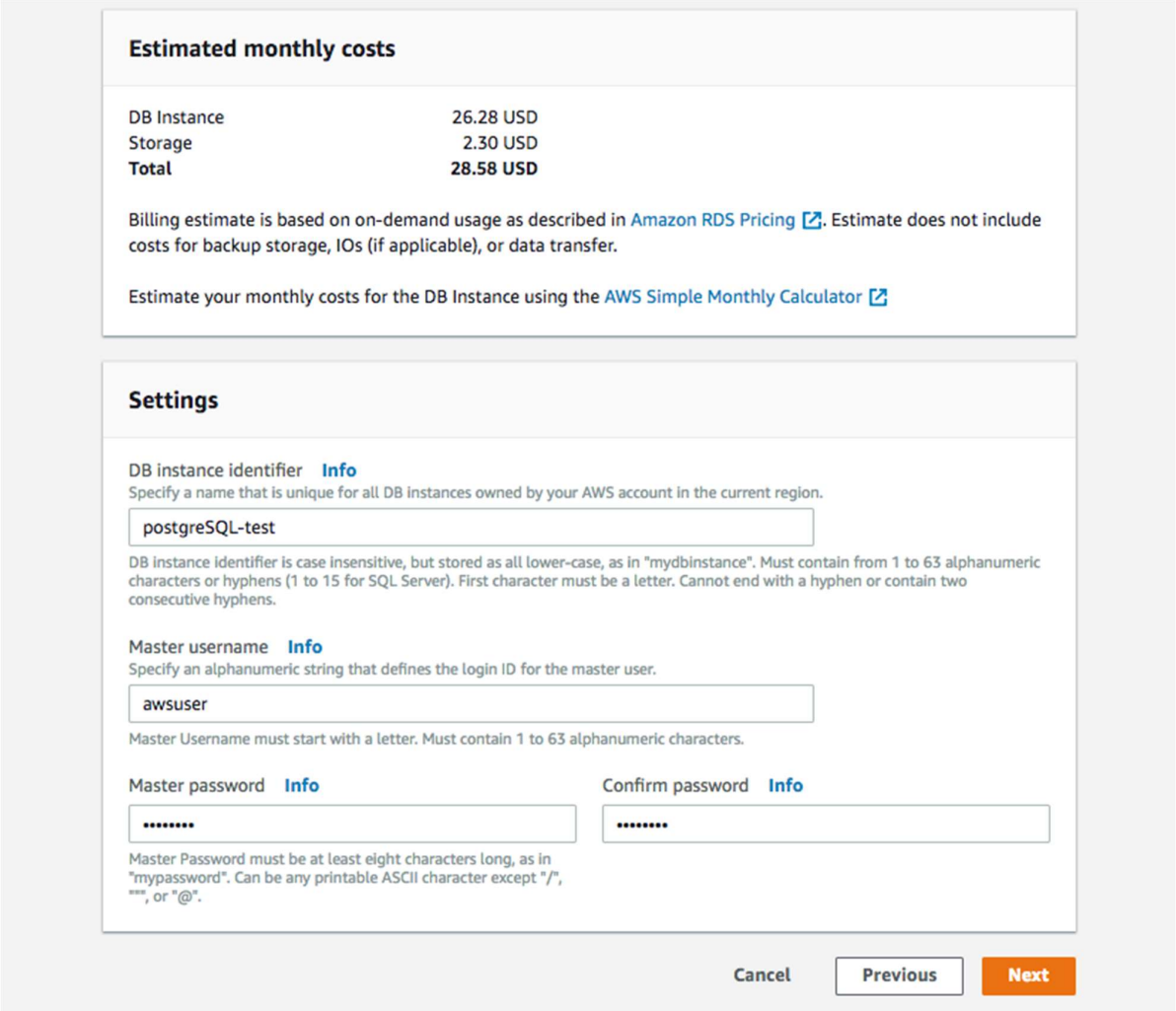


5. Next, is a long **Specify DB details** page. You can leave the default values (shown below) for most of these settings. Just make the following choices:

- For **DB instance class**, select `db.t2.small`



- For DB instance identifier, enter `postgreSQL-test` or another name of your choice
- Enter a master username and password



- Leave the default values for the next few sections.

Configure advanced settings

Network & Security

Virtual Private Cloud (VPC) **Info**

VPC defines the virtual networking environment for this DB instance.

Default VPC (vpc-2cc4274b)

Only VPCs with a corresponding DB subnet group are listed.

Subnet group **Info**

DB subnet group that defines which subnets and IP ranges the DB instance can use in the VPC you selected.

default

Public accessibility **Info**

- (•) Yes
  EC2 instances and devices outside of the VPC hosting the DB instance will connect to the DB instances. You must also select one or more VPC security groups that specify which EC2 instances and devices can connect to the DB instance.
- ( ) No
  DB instance will not have a public IP address assigned. No EC2 instance or devices outside of the VPC will be able to connect.

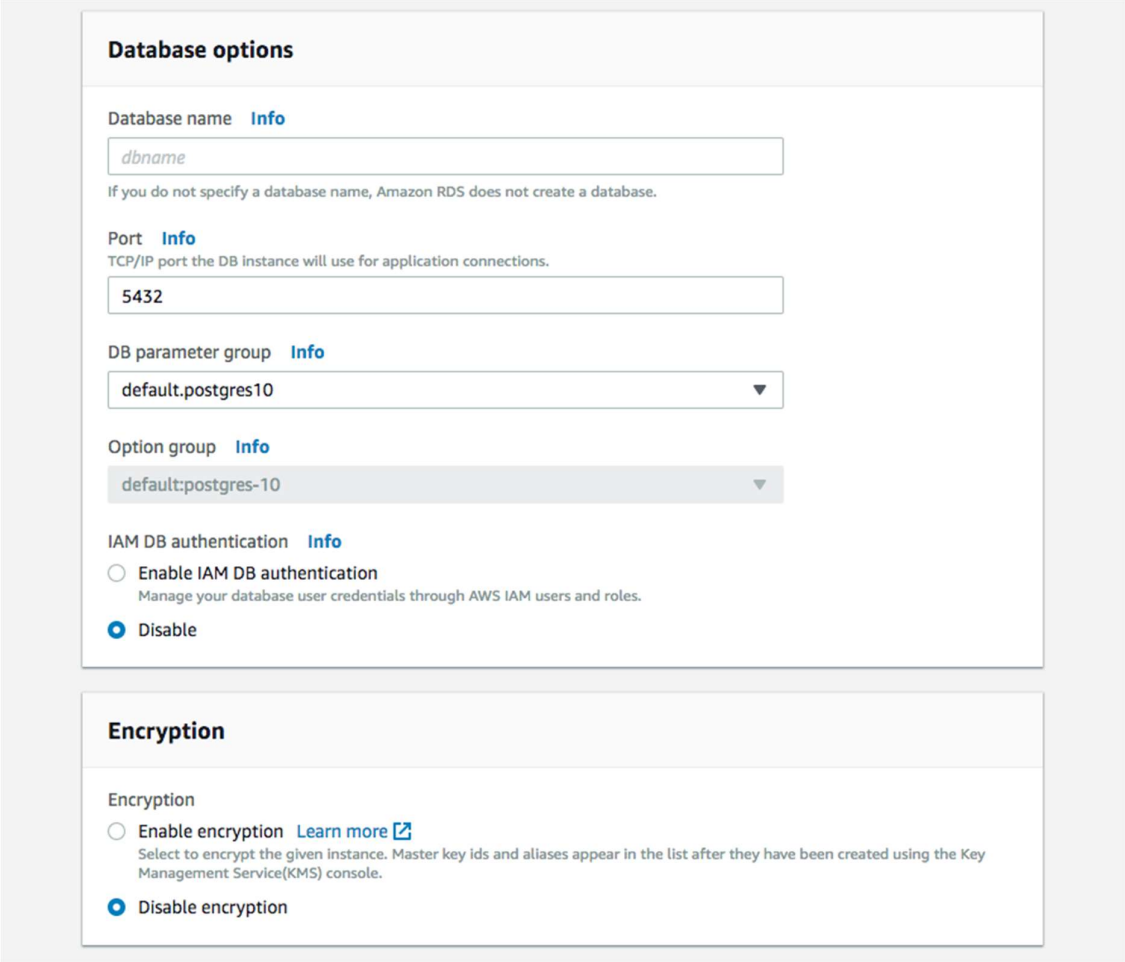
Availability zone **Info**

No preference

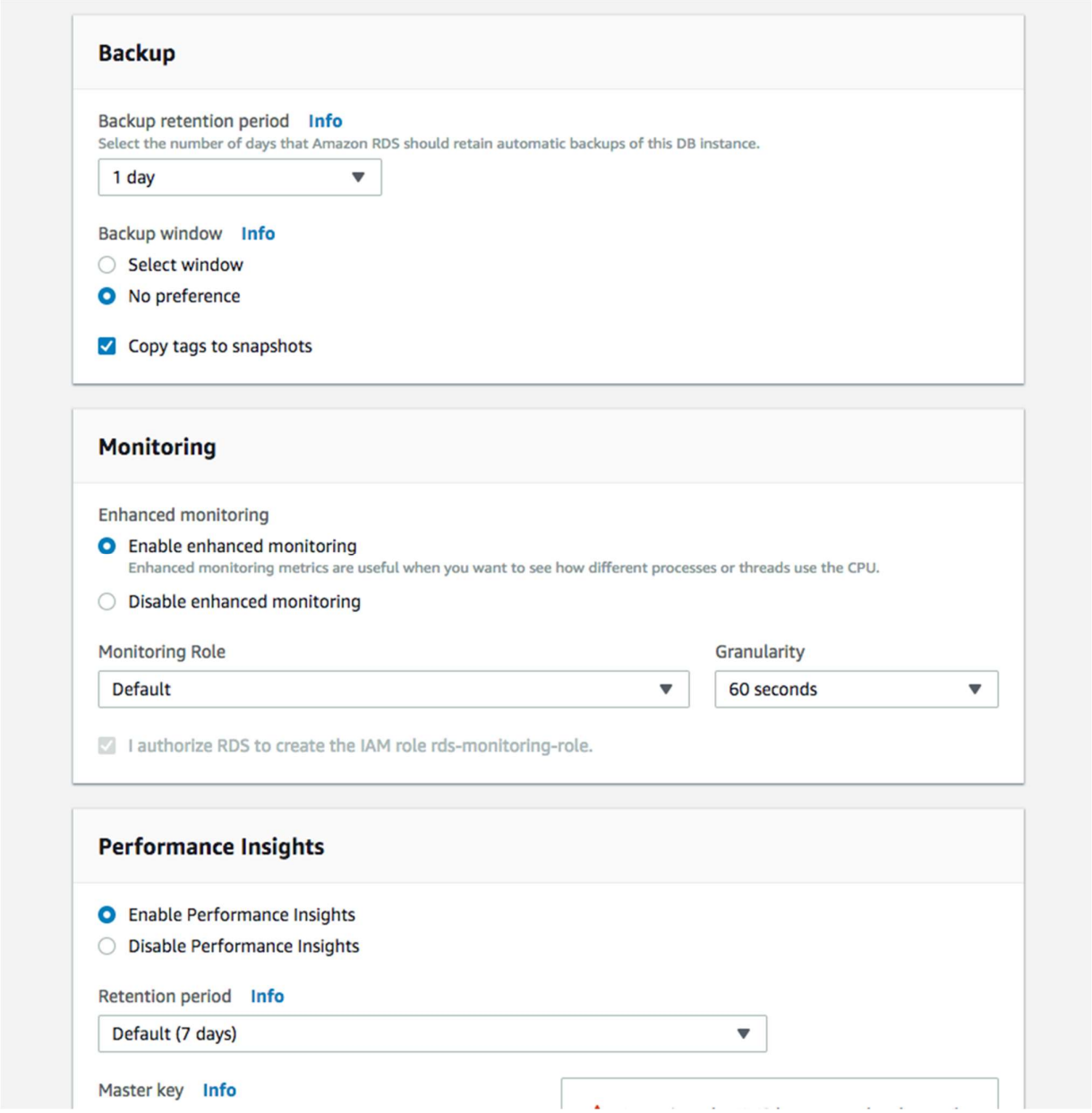
VPC security groups

Security groups have rules authorizing connections from all the EC2 instances and devices that need to access the DB instance.

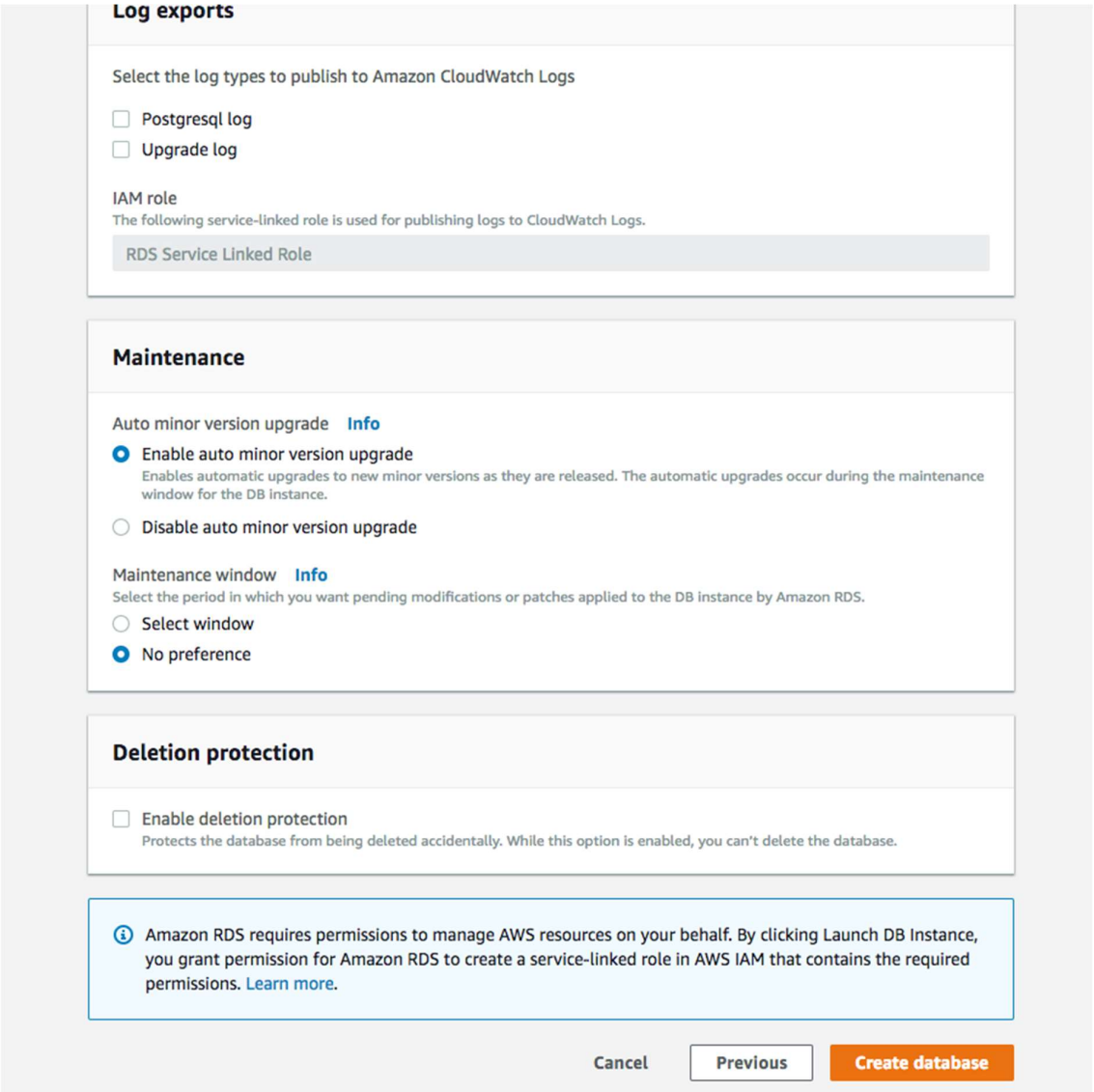
- (•) Create new VPC security group
- ( ) Choose existing VPC security groups



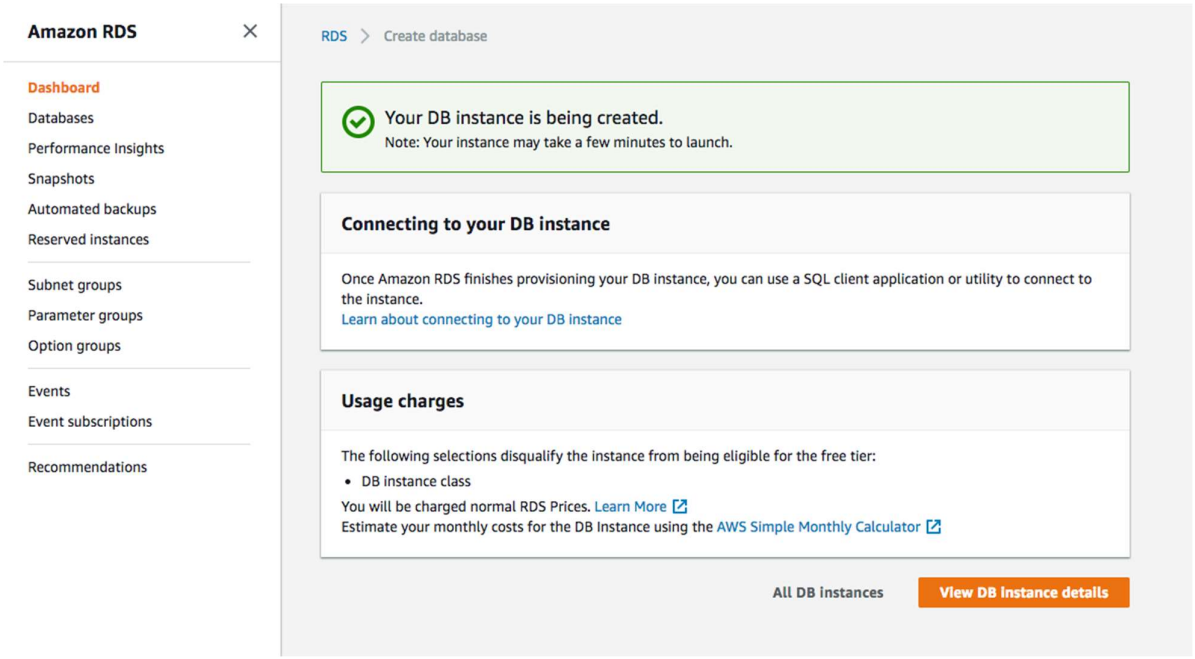
- In the **Backup** section and select `1 day` since this is for demonstration purposes.



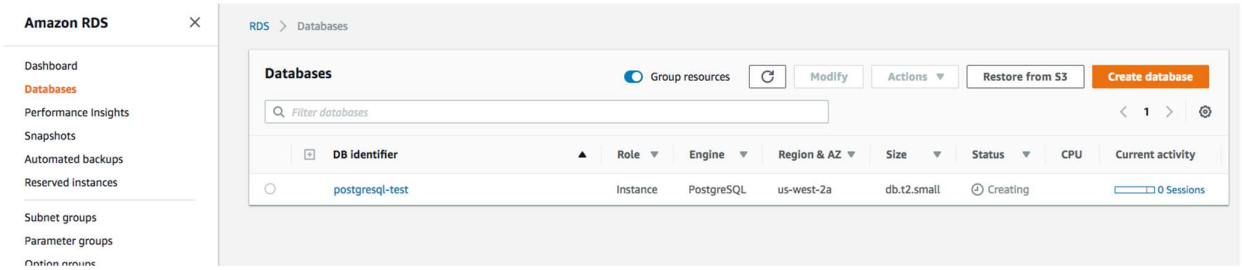
- Leave the default values for the rest and click on **Create database** on the bottom right.



You should land on a confirmation page.



6. Click **Databases** on the left navigation pane to return to your list of databases. You should see your newly created database with the status **Creating**.



7. Wait a few minutes for this to change to the status **Available**.