LFS303 - ENTERPRISE DATA LAKES IN LIFE SCIENCES

Overview

In this builder session, we will demonstrate how to leverage AWS Lake Formation and other AWS services to build an Enterprise Data Lake. We will understand how different user personas interact with the Data Lake through Lake Formation. We will see how to use Lake Formation to integrate data in S3 buckets. We will understand how to use Lake Formation Data Catalog to control data access. We will use AWS Athena and Redshift to perform ad-hoc analysis in the Data Lake under different personas to show how Lake Formation controls data access to the Data Lake.

Building Data Lake with AWS Lake Formation and connecting to it through data catalog, there are multiple advantages to this approach:

- Avoid building redundant copies of the same data and have a single control place for data and metadata access
- Reduce the amount of storage required and also lower security risk
- Keep the data secure with only authorized users having access to it
- Scale the storage and compute separately as needed, but still maintain a single control plane for data access

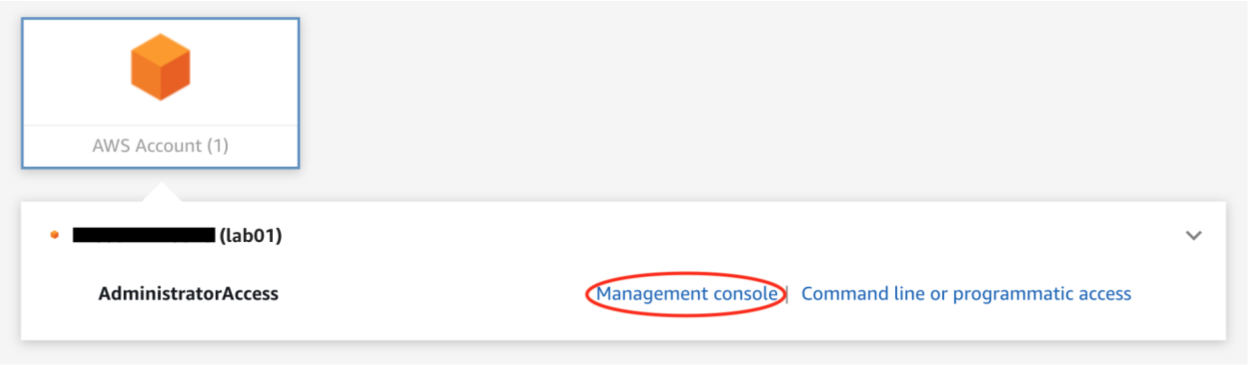
About the Personas in This Lab

Persona	Description
IAM Administrator (Superuser)	User who can create IAM users and roles and Amazon S3 buckets. Has the AdministratorAccess AWS managed policy. You need this permission to finish this lab.
Data Lake administrator	User who can access the data catalog, create databases, and grant Lake Formation permissions to other users. Has fewer IAM permissions than the IAM administrator, but enough to administer the data lake.
User with PHI access	Users who can access PHI data in the Data Lake.
User without PHI access	Users who can access non-PHI data in the Data Lake.
Workflow role	Service role with the required IAM policies to run a Glue workflow.

Log into provided Lab AWS account

Lab instructor will provide you login link and credentials for the lab.

Please use this link to access the lab login page and credentials provided by lab instructor to log into the lab account. Once logged in, click the Management console link to reach to the provide AWS lab account’s management console.



AWS Account (1)

[redacted] (lab01)

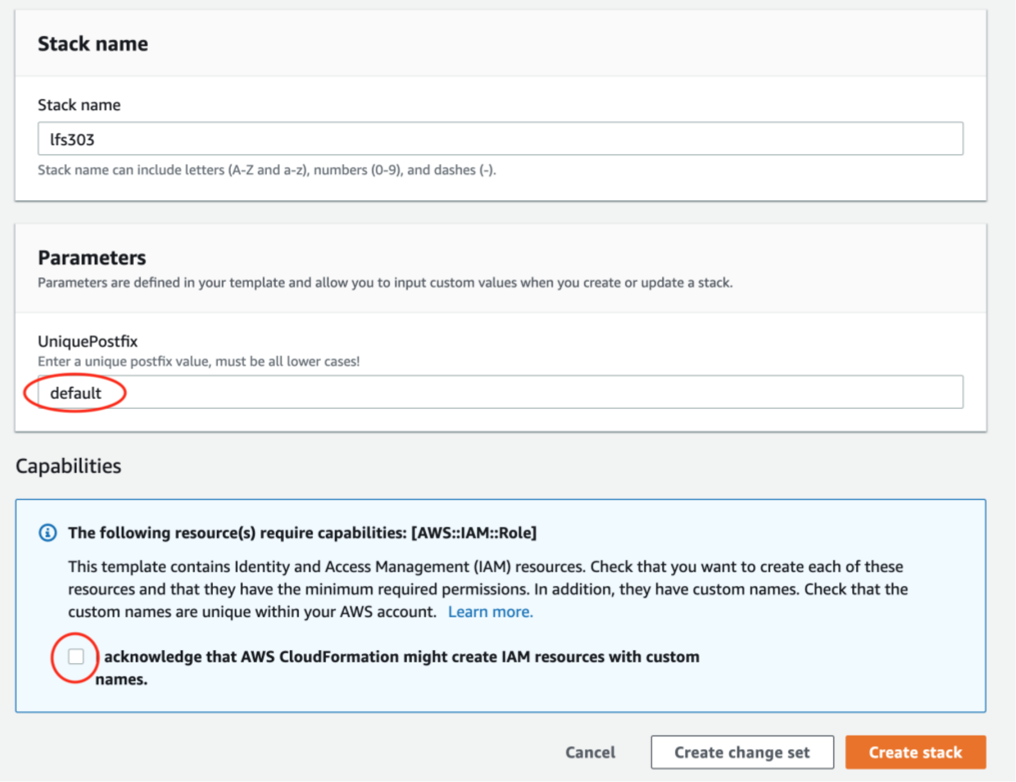
AdministratorAccess

Setup prerequisite with CloudFormation Template

Once you logged into the provided AWS account AWS console, you have administrator access in that account. Click the CloudFormation link below to create the resources needed for the lab including S3 Buckets, IAM roles and Cloud9 environment

[CloudFormation Template]

You could leave the stack name as the default or name it something you prefer. M**ake sure to change UniquePostfix value to something unique** and also check the checkbox below. Then click "Create Stack". It takes a couple minutes for the resources to be created.



Stack name

Stack name

lfs303

Stack name can include letters (A-Z and a-z), numbers (0-9), and dashes (-).

Parameters

Parameters are defined in your template and allow you to input custom values when you create or update a stack.

UniquePostfix

Enter a unique postfix value, must be all lower cases!

default

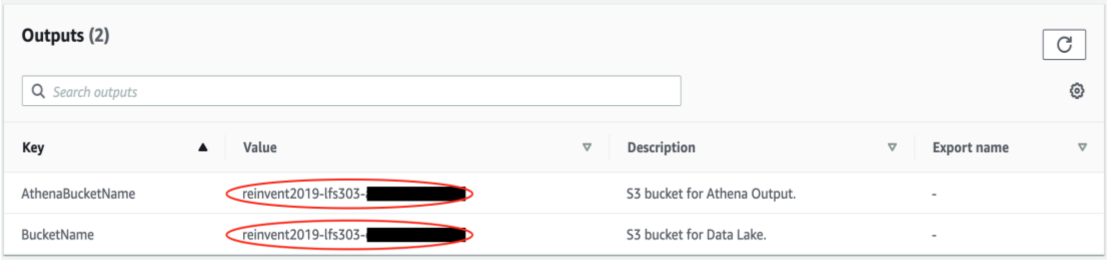
Capabilities

The following resource(s) require capabilities: [AWS::IAM::Role]

This template contains Identity and Access Management (IAM) resources. Check that you want to create each of these resources and that they have the minimum required permissions. In addition, they have custom names. Check that the custom names are unique within your AWS account. Learn more.

□ **acknowledge that AWS CloudFormation might create IAM resources with custom names.**

After the CloudFormation stack resources creation finished, go to the output tab and copy the value of <AthenaBucketName> and <BucketName> into a notepad. You will need the values for latter steps. If you go to S3 service now, you should see 2 buckets with those names created under your lab account.



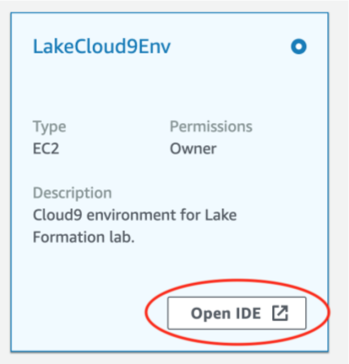
Outputs (2)

Search outputs

Key	Value	Description	Export name
AthenaBucketName	reinvent2019-lfs303-[redacted]	S3 bucket for Athena Output.	-
BucketName	reinvent2019-lfs303-[redacted]	S3 bucket for Data Lake.	-

Copy Lab data into S3 Bucket

Go to Cloud9 service and open the Cloud9 IDE.



LakeCloud9Env

Type: EC2

Permissions: Owner

Description: Cloud9 environment for Lake Formation lab.

Open a new terminal by clicking Windows -> New Terminal, and execute the commands below (replacing <BucketName> with your own S3 bucket name from previous step)

```
aws s3 cp s3://reinvent2019-lfs303/NONPHI.csv s3://<BucketName>/NONPHI/NONPHI.csv
aws s3 cp s3://reinvent2019-lfs303/PHI.csv s3://<BucketName>/PHI/PHI.csv
aws s3 ls s3://<BucketName> --recursive
```

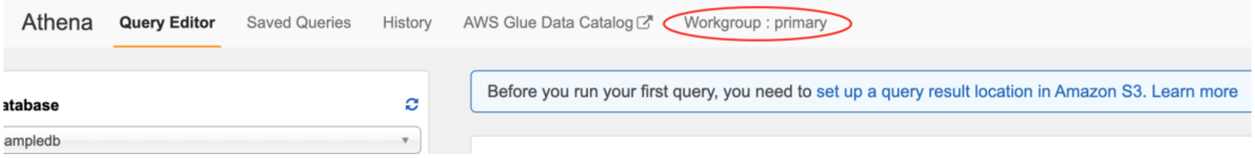
You should see the following if files copied successfully.

```
2019-11-23 01:41:53    9815312 NONPHI/NONPHI.csv
2019-11-23 01:41:54    4882781 PHI/PHI.csv
```

Closed the browser tab of Cloud9 IDE.

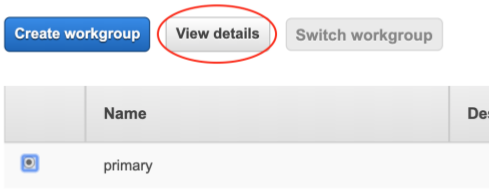
Config Athena Output location for primary workgroup

Go to Athena service in AWS Console, click “Get Started”. Before you could use Athena, you need to setup a query result location in S3 and we will use the S3 bucket we created earlier as the result bucket. Click “Workgroup : primary” tab



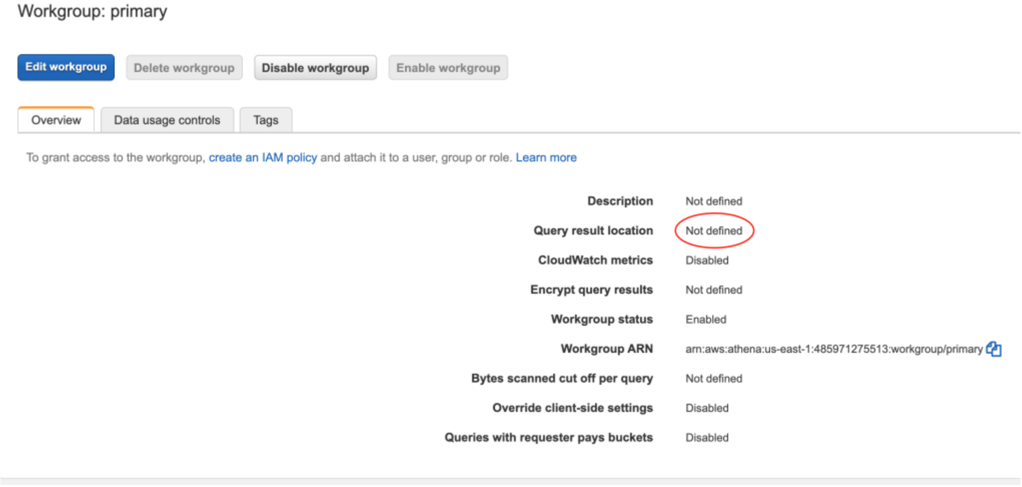
Before you run your first query, you need to set up a query result location in Amazon S3. Learn more

Choose primary group and click “View details”



	Name	De
●	primary	

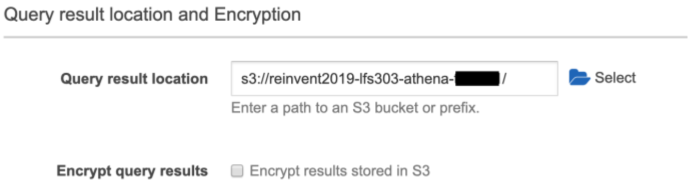
Notice the Query result location is not defined


Workgroup: primary

To grant access to the workgroup, create an IAM policy and attach it to a user, group or role. Learn more

Description	Not defined
Query result location	Not defined
CloudWatch metrics	Disabled
Encrypt query results	Not defined
Workgroup status	Enabled
Workgroup ARN	arn:aws:athena:us-east-1:485971275513:workgroup/primary
Bytes scanned cut off per query	Not defined
Override client-side settings	Disabled
Queries with requester pays buckets	Disabled

Click “Edit workgroup” and select the <AthenaBucketName> as the Query result location, then click “Save”


Query result location and Encryption

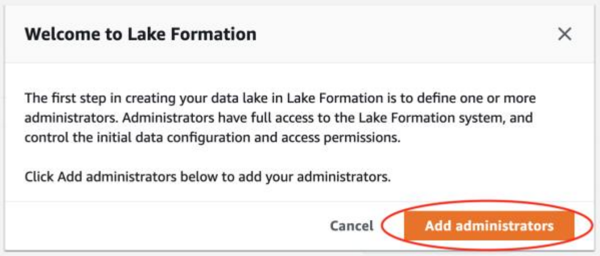
Query result location: s3://reinvent2019-lfs303-athena-[redacted]/

Enter a path to an S3 bucket or prefix.

Encrypt query results: □ Encrypt results stored in S3

Config Lake Formation

Go to Lake Formation service. You should see the following message asking you to add an administrator in Lake formation, click “Add Administrators”

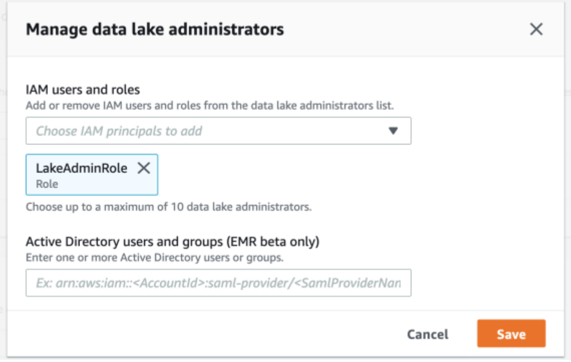


Welcome to Lake Formation

The first step in creating your data lake in Lake Formation is to define one or more administrators. Administrators have full access to the Lake Formation system, and control the initial data configuration and access permissions.

Click Add administrators below to add your administrators.

Choose “LakeAdminRole” in the dropdown and click save. You will now see the Lake Formation UI.



Manage data lake administrators

IAM users and roles
Add or remove IAM users and roles from the data lake administrators list.

Choose IAM principals to add

LakeAdminRole
Role

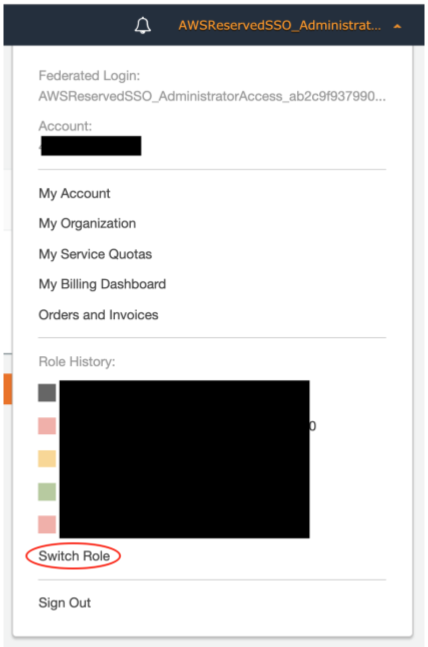
Choose up to a maximum of 10 data lake administrators.

Active Directory users and groups (EMR beta only)
Enter one or more Active Directory users or groups.

Ex: arn:aws:iam::<AccountId>:saml-provider/<SamlProviderNan

Note: switching roles in AWS Console in Lake Formation seems to be problematic. If you can’t switch role in Lake Formation AWS Console, change to another service and switch role there.

Copy the account number from the dropdown and click “Switch Role”



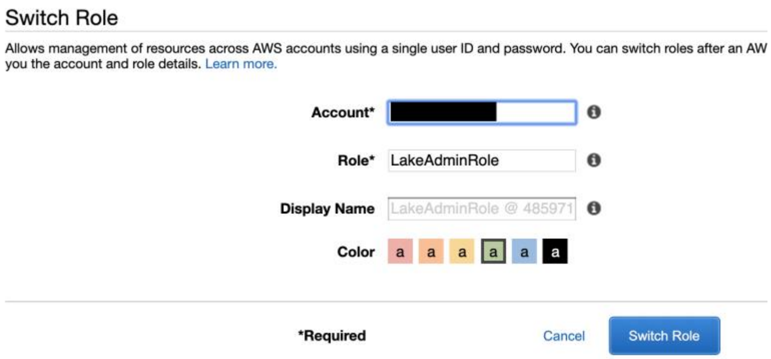
AWSReservedSSO_Administrat...

Federated Login:
AWSReservedSSO_AdministratorAccess_ab2c9f937990...

Account:
[redacted]

Role History:
[redacted]

Enter the account number after removing all dashes and set Role name to “LakeAdminRole”. Click “Switch Role”.



Switch Role

Allows management of resources across AWS accounts using a single user ID and password. You can switch roles after an AW you the account and role details. Learn more.

Account* [redacted]

Role* LakeAdminRole

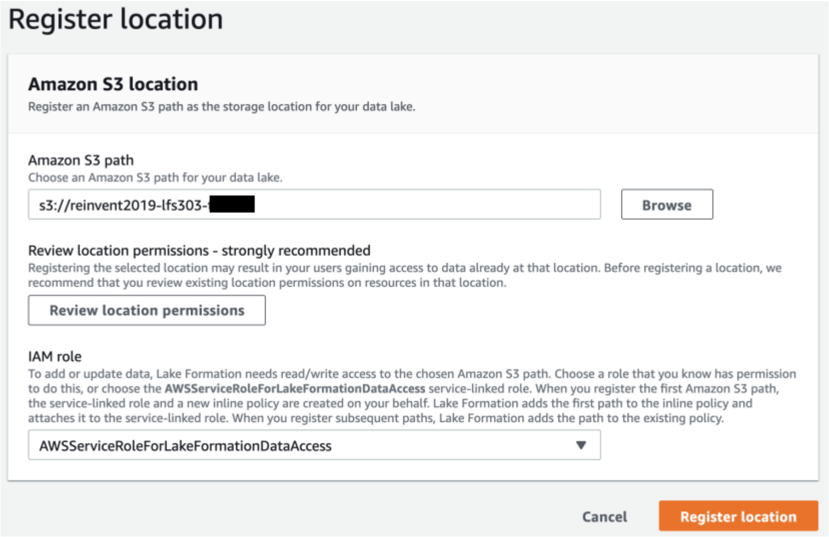
Display Name LakeAdminRole @ 485971

Color a a a a a a

*Required

Register your Amazon S3 storage

Lake Formation manages access to designated storage locations within Amazon S3. Register the storage locations that you want to be part of the data lake. Click “Data lake location” under Lake Formation and register the S3 bucket created by the CloudFormation template earlier.



Register location

Amazon S3 location

Register an Amazon S3 path as the storage location for your data lake.

Amazon S3 path

Choose an Amazon S3 path for your data lake.

s3://reinvent2019-lfs303-[redacted]

Review location permissions - strongly recommended

Registering the selected location may result in your users gaining access to data already at that location. Before registering a location, we recommend that you review existing location permissions on resources in that location.

IAM role

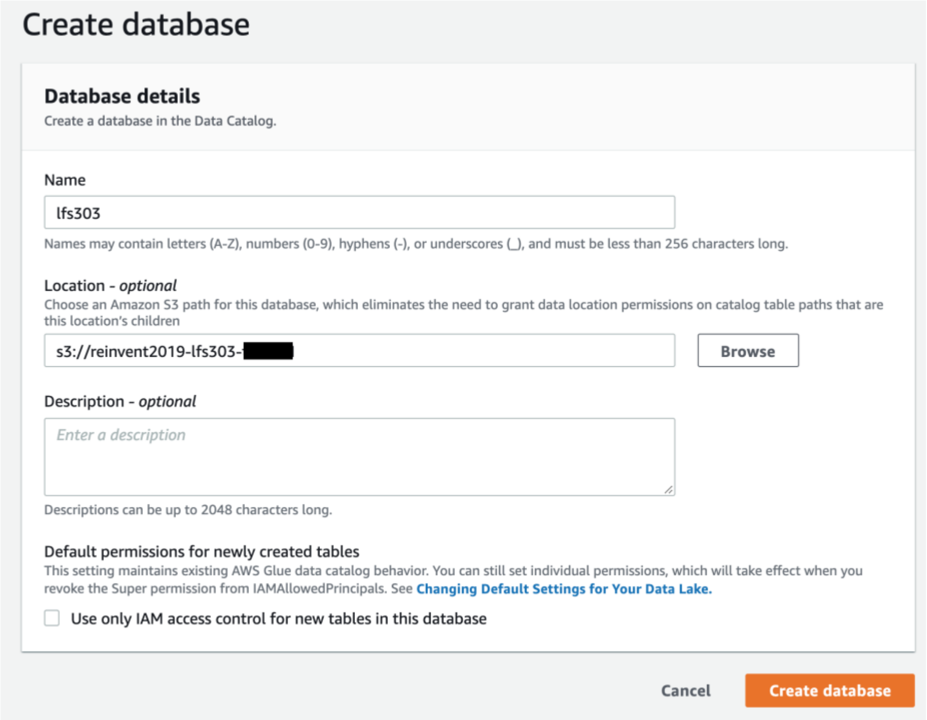
To add or update data, Lake Formation needs read/write access to the chosen Amazon S3 path. Choose a role that you know has permission to do this, or choose the **AWSServiceRoleForLakeFormationDataAccess** service-linked role. When you register the first Amazon S3 path, the service-linked role and a new inline policy are created on your behalf. Lake Formation adds the first path to the inline policy and attaches it to the service-linked role. When you register subsequent paths, Lake Formation adds the path to the existing policy.

AWSServiceRoleForLakeFormationDataAccess ▼

Create a database and Grant access to Crawler role

Lake Formation organizes data into a catalog of logical databases and tables. Create one or more databases and then automatically generate tables during data ingestion for common workflows.

Click “Database” under Lake Formation and create Database and choose the location as the S3 bucket created by CloudFormation template. Uncheck “Use only IAM access control for new tables in this database” if it is checked



Create database

Database details

Create a database in the Data Catalog.

Name

lfs303

Names may contain letters (A-Z), numbers (0-9), hyphens (-), or underscores (_), and must be less than 256 characters long.

Location - *optional*

Choose an Amazon S3 path for this database, which eliminates the need to grant data location permissions on catalog table paths that are this location's children

s3://reinvent2019-lfs303-[redacted]

Description - *optional*

Enter a description

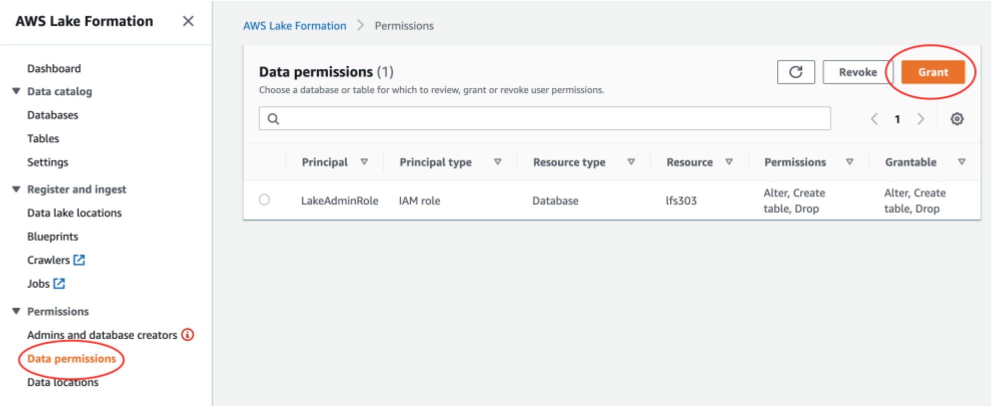
Descriptions can be up to 2048 characters long.

Default permissions for newly created tables

This setting maintains existing AWS Glue data catalog behavior. You can still set individual permissions, which will take effect when you revoke the Super permission from IAMAllowedPrincipals. See **Changing Default Settings for Your Data Lake.**

□ Use only IAM access control for new tables in this database

Click "Data permission" under Lake Formation and click "Grant" on the upper right corner.

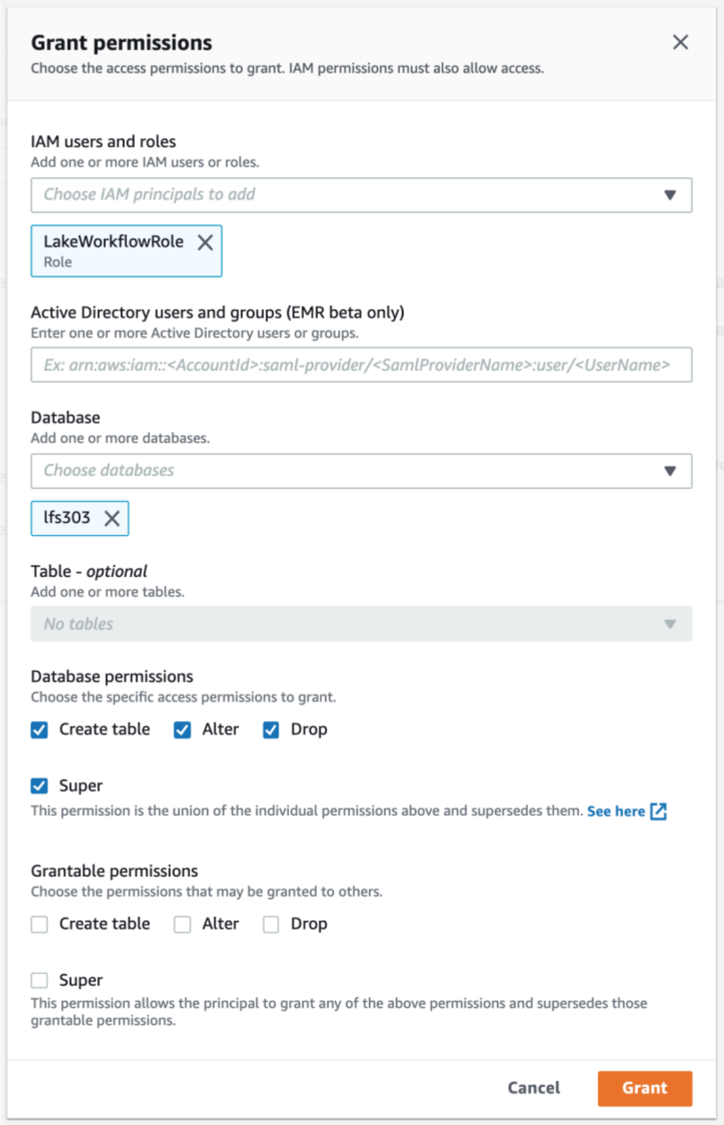


Data permissions (1)

Choose a database or table for which to review, grant or revoke user permissions.

	Principal	Principal type	Resource type	Resource	Permissions	Grantable
○	LakeAdminRole	IAM role	Database	lfs303	Alter, Create table, Drop	Alter, Create table, Drop

Grant "LakeWorkflowRole" permissions to create/alter/drop tables in the newly database.



Grant permissions

Choose the access permissions to grant. IAM permissions must also allow access.

IAM users and roles

Add one or more IAM users or roles.

Choose IAM principals to add

LakeWorkflowRole
Role

Active Directory users and groups (EMR beta only)

Enter one or more Active Directory users or groups.

Ex: arn:aws:iam::<AccountId>:saml-provider/<SamlProviderName>:user/<UserName>

Database

Add one or more databases.

Choose databases

lfs303

Table - *optional*

Add one or more tables.

No tables

Database permissions

Choose the specific access permissions to grant.

☑ Create table ☑ Alter ☑ Drop

☑ Super

This permission is the union of the individual permissions above and supersedes them. See here

Grantable permissions

Choose the permissions that may be granted to others.

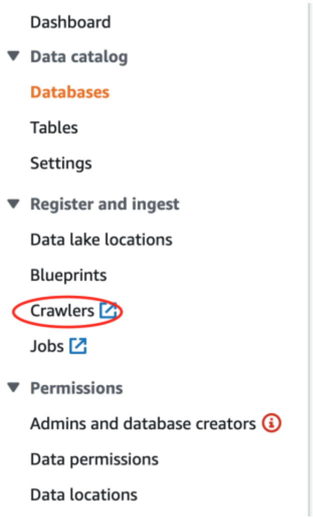
□ Create table □ Alter □ Drop

□ Super

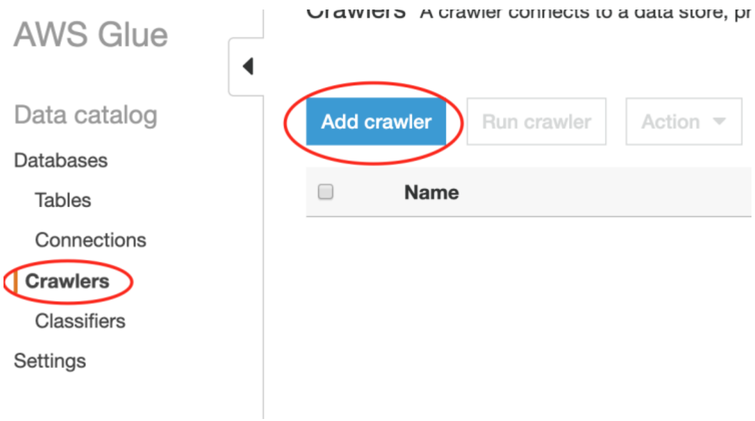
This permission allows the principal to grant any of the above permissions and supersedes those grantable permissions.

Use Glue Crawler to create tables automatically

Click “Crawler” under Lake Formation to open Glue Crawler UI in a new browser tab. Click “Get Started”



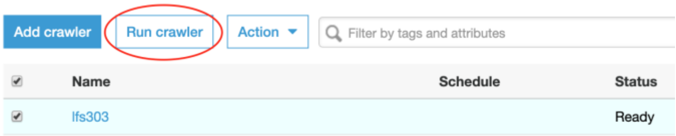
Click Crawler on the left and “Add crawler” to enter Crawler configuration screen



Put the following settings for the new crawler:

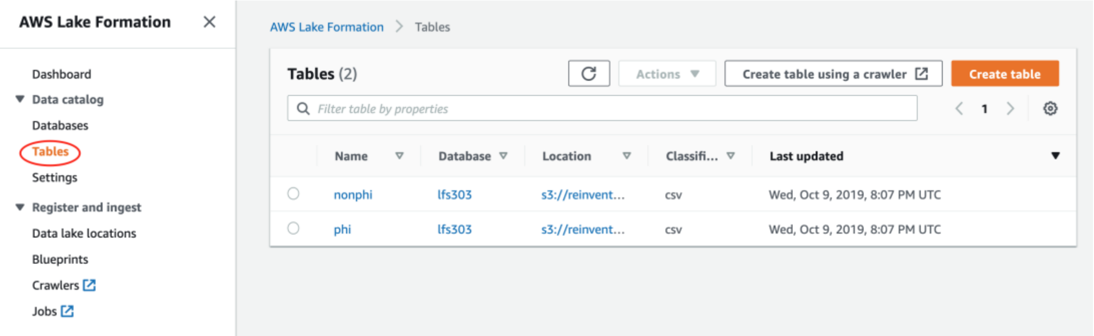
- Set Crawler name as “lfs303” or anything you like, then click “Next”.
- Click “Next”.
- Choose S3 as data store type and <BucketName> S3 bucket as the specified path
- Click “Next”.
- Choose “Choose an existing IAM role” and pick “LakeWorkflowRole” in the drop down, click “Next”.
- Click “Next”.
- Choose lfs303 (or the database name you named earlier) in Database name dropdown, click “Next”.
- Click “Finish”.

Choose the Crawler you just created and click “Run crawler”. The crawler should finish scanning the data and creating new tables based on the file’s schema in a couple minutes. You could examine the Crawler log to see what happened after it is done (need to switch back to the account admin user).



Name	Schedule	Status
lfs303		Ready

Close the Crawler browser tab and go back to Lake Formation tab. Click “Tables” under Lake Formation and you should see the 2 new tables created by Crawler.

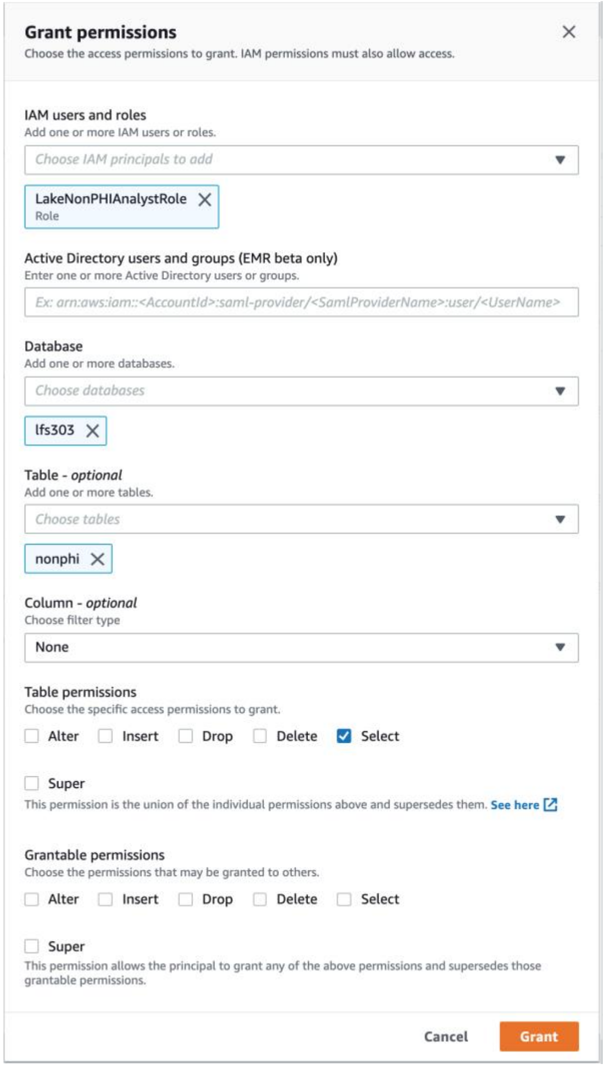


Tables (2)

Name	Database	Location	Classifi...	Last updated
nonphi	lfs303	s3://reinvent...	csv	Wed, Oct 9, 2019, 8:07 PM UTC
phi	lfs303	s3://reinvent...	csv	Wed, Oct 9, 2019, 8:07 PM UTC

Grant Data Permission to Analyst roles

Click “Data permission” under Lake formation and grant data access to “LakePHIAnalystRole” and “LakeNonPHIAnalystRole” with the following settings. (You can also experiment with the settings and observe the outcome)



Grant permissions ✕

Choose the access permissions to grant. IAM permissions must also allow access.

IAM users and roles
Add one or more IAM users or roles.
Choose IAM principals to add ▼
LakeNonPHIAnalystRole ✕
Role

Active Directory users and groups (EMR beta only)
Enter one or more Active Directory users or groups.
Ex: arn:aws:iam::<AccountId>:saml-provider/<SamlProviderName>:user/<UserName>

Database
Add one or more databases.
Choose databases ▼
lfs303 ✕

Table - *optional*
Add one or more tables.
Choose tables ▼
nonphi ✕

Column - *optional*
Choose filter type
None ▼

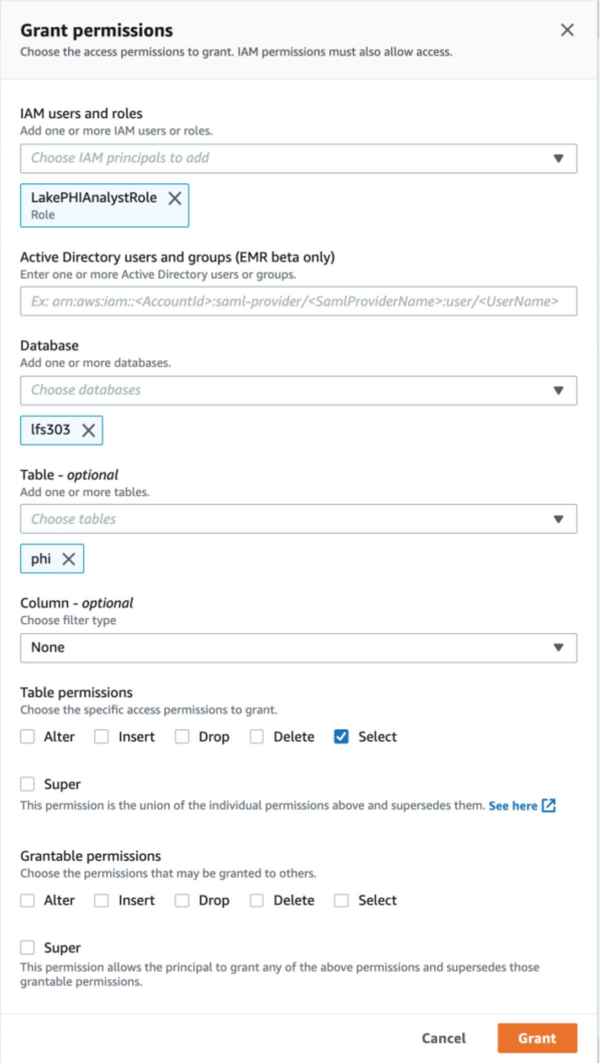
Table permissions
Choose the specific access permissions to grant.
□ Alter □ Insert □ Drop □ Delete ☑ Select

□ Super
This permission is the union of the individual permissions above and supersedes them. See here

Grantable permissions
Choose the permissions that may be granted to others.
□ Alter □ Insert □ Drop □ Delete □ Select

□ Super
This permission allows the principal to grant any of the above permissions and supersedes those grantable permissions.

Cancel Grant



Grant permissions ✕

Choose the access permissions to grant. IAM permissions must also allow access.

IAM users and roles
Add one or more IAM users or roles.
Choose IAM principals to add ▼
LakePHIAnalystRole ✕
Role

Active Directory users and groups (EMR beta only)
Enter one or more Active Directory users or groups.
Ex: arn:aws:iam::<AccountId>:saml-provider/<SamlProviderName>:user/<UserName>

Database
Add one or more databases.
Choose databases ▼
lfs303 ✕

Table - *optional*
Add one or more tables.
Choose tables ▼
phi ✕

Column - *optional*
Choose filter type
None ▼

Table permissions
Choose the specific access permissions to grant.
□ Alter □ Insert □ Drop □ Delete ☑ Select

□ Super
This permission is the union of the individual permissions above and supersedes them. See here

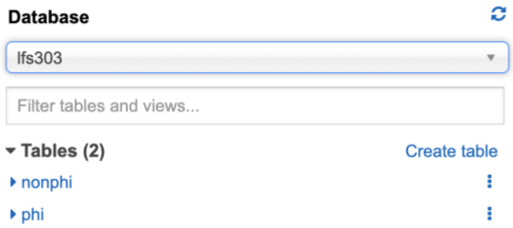
Grantable permissions
Choose the permissions that may be granted to others.
□ Alter □ Insert □ Drop □ Delete □ Select

□ Super
This permission allows the principal to grant any of the above permissions and supersedes those grantable permissions.

Cancel Grant

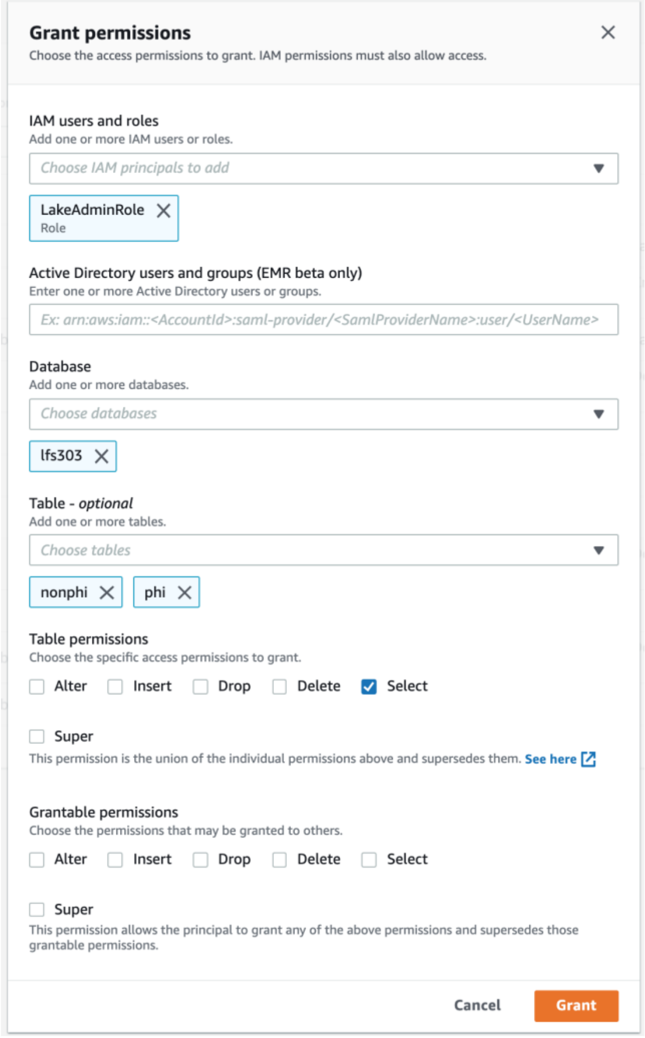
Query data in Athena

Switch to Athena service in AWS console under the current “LakeAdminRole” role. You should be able to see the 2 tables inside the database lfs303 (or the database you named earlier). But if you try to select the data in those tables, you will receive error messages because you don’t have the data permission under “LakeAdminRole” role. The error message is a bit misleading, but the cause is the lack of permission.



Database
lfs303 ▼
Filter tables and views...
▾ Tables (2) Create table
▸ nonphi
▸ phi

If you go back to Lake Formation and grant “LakeAdminRole” role the data permission to those tables, you will be able to view the data without errors.



Grant permissions

Choose the access permissions to grant. IAM permissions must also allow access.

IAM users and roles

Add one or more IAM users or roles.

Choose IAM principals to add

LakeAdminRole
Role

Active Directory users and groups (EMR beta only)

Enter one or more Active Directory users or groups.

Ex: arn:aws:iam::<AccountId>:saml-provider/<SamlProviderName>:user/<UserName>

Database

Add one or more databases.

Choose databases

lfs303

Table - *optional*

Add one or more tables.

Choose tables

nonphi

phi

Table permissions

Choose the specific access permissions to grant.

□ Alter □ Insert □ Drop □ Delete ☑ Select

□ Super

This permission is the union of the individual permissions above and supersedes them. See here

Grantable permissions

Choose the permissions that may be granted to others.

□ Alter □ Insert □ Drop □ Delete □ Select

□ Super

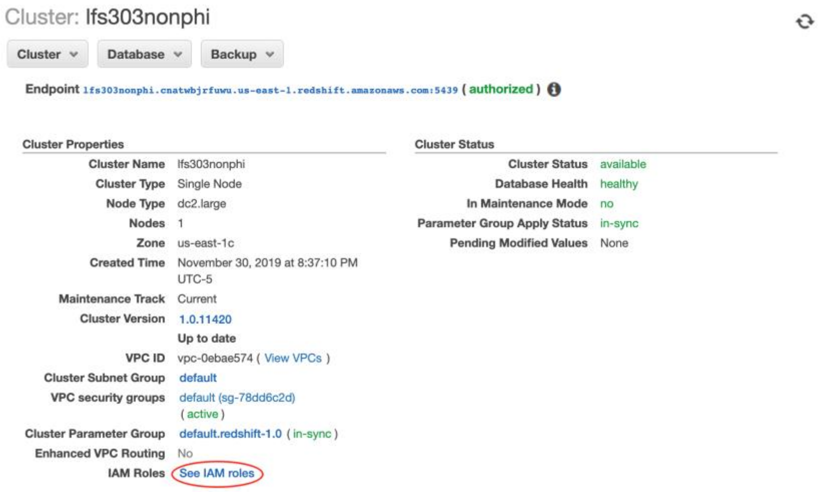
This permission allows the principal to grant any of the above permissions and supersedes those grantable permissions.

Then switch roles to “LakePHIAnalystRole” and “LakeNonPHIAnalystRole” and see what is the difference inside Athena and try to understand why.

Query Data in Redshift Spectrum (optional)

The data access control model is different in Redshift Spectrum than Athena. There are 2 Redshift clusters created by the CloudFormation template. One cluster assumes the “LakePHIAnalystRole” role for data access and the other assumes the “LakeNonPHIAnalystRole” role for data access. The clusters are named accordingly.

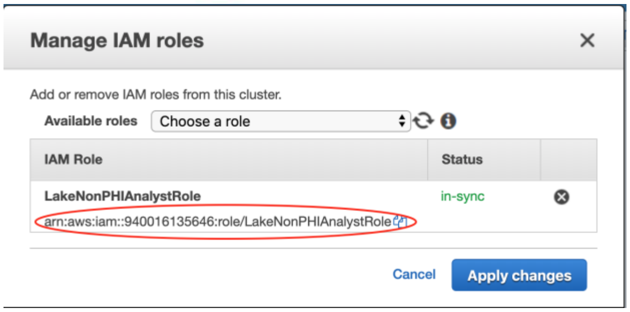
Switch back to the account admin user before move on. Switch to Redshift service in AWS Console and click "Clusters" on the left. Pick any of the 2 clusters and examine the IAM Roles property of the cluster. Notice it assume one of the data access roles.



Endpoint `lfs303nonphi.cnatwbjrfuwu.us-east-1.redshift.amazonaws.com:5439` (authorized)

Cluster Properties	
Cluster Name	lfs303nonphi
Cluster Type	Single Node
Node Type	dc2.large
Nodes	1
Zone	us-east-1c
Created Time	November 30, 2019 at 8:37:10 PM UTC-5
Maintenance Track	Current
Cluster Version	1.0.11420 Up to date
VPC ID	vpc-0ebae574 (View VPCs)
Cluster Subnet Group	default
VPC security groups	default (sg-78dd6c2d) (active)
Cluster Parameter Group	default.redshift-1.0 (in-sync)
Enhanced VPC Routing	No
IAM Roles	See IAM roles

Cluster Status	
Cluster Status	available
Database Health	healthy
In Maintenance Mode	no
Parameter Group Apply Status	in-sync
Pending Modified Values	None



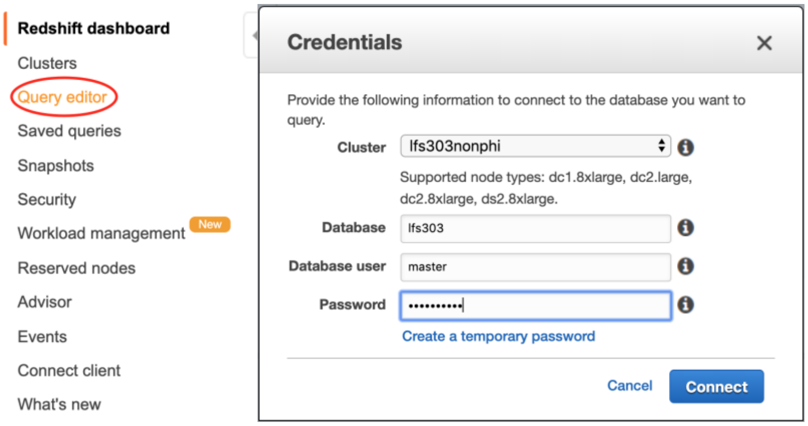
Add or remove IAM roles from this cluster.

Available roles: Choose a role

IAM Role	Status
LakeNonPHIAnalystRole arn:aws:iam::940016135646:role/LakeNonPHIAnalystRole	in-sync

Click "query editor" on the left. Pick any of the 2 clusters and put in the following login information

```
Database: lfs303
Database user: master
Password: Password#1
```



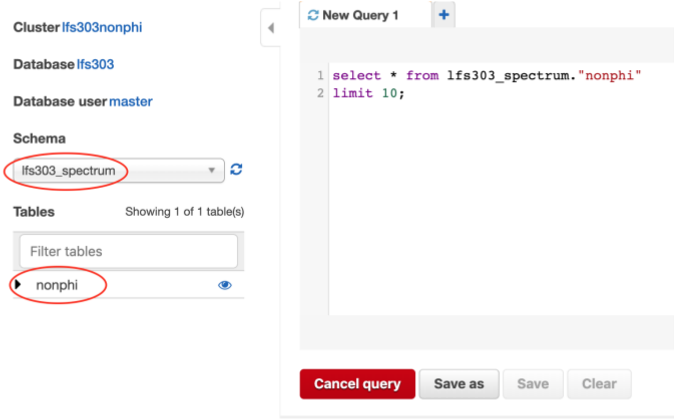
Provide the following information to connect to the database you want to query.

Cluster	lfs303nonphi Supported node types: dc1.8xlarge, dc2.large, dc2.8xlarge, ds2.8xlarge.
Database	lfs303
Database user	master
Password	••••••••••

Once in the query editor, create an external schema in Redshift from Lake Formation data catalog with following command. For “Role Name”, put in “LakePHIAnalystRole” or “LakeNonPHIAnalystRole” depends on which cluster it is.

```
create external schema lfs303_spectrum
from DATA CATALOG database 'lfs303'
iam_role 'arn:aws:iam::<account id>:role/<role name>';
```

Once the external schema is created successfully, you should see it inside the dropdown for schemas. Select it as the current schema, and you will see only the tables that is accessible to the role assumed by the Redshift cluster.



Cluster lfs303nonphi

Database lfs303

Database user master

Schema: lfs303_spectrum

Tables — Showing 1 of 1 table(s)

nonphi

```
select * from lfs303_spectrum."nonphi"
limit 10;
```

Cleanup (optional)

Remove Registered Data Location in Lake Formation.
Remove Database (lfs303) created in Lake Formation.
Remove crawler (lfs303) in Glue.
Empty both S3 buckets in S3.
Delete the CloudFormation stack (lfs303) created in Cloud Formation.