## 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

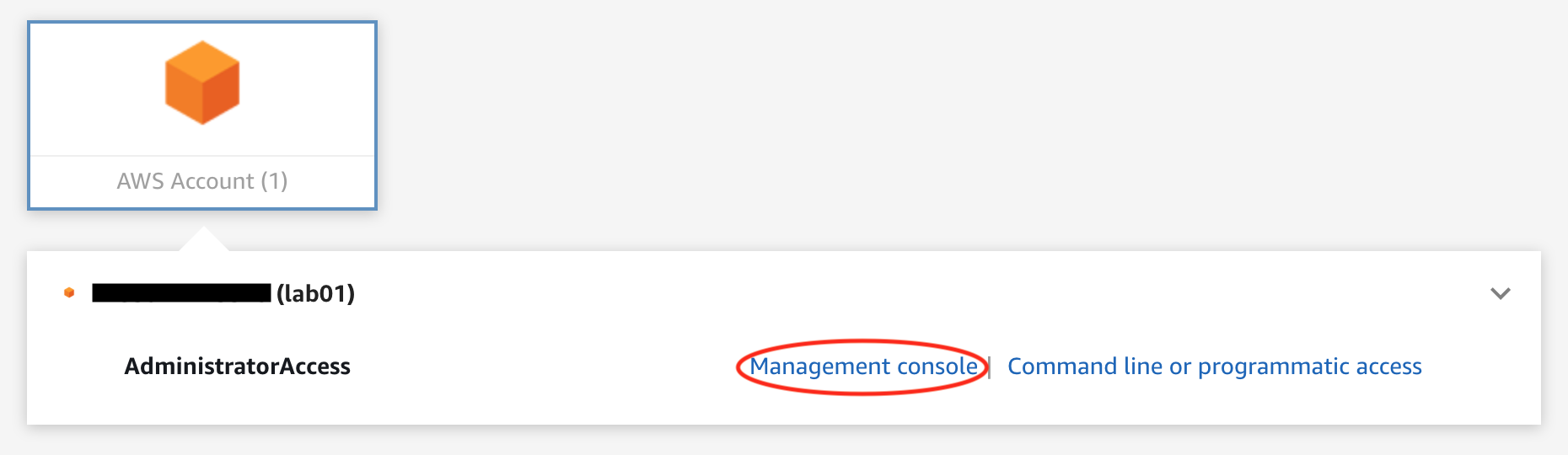
### About the Personas in This Lab

|  |  |
| --- | --- |
| **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](https://qiwanglab.awsapps.com/start) 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.

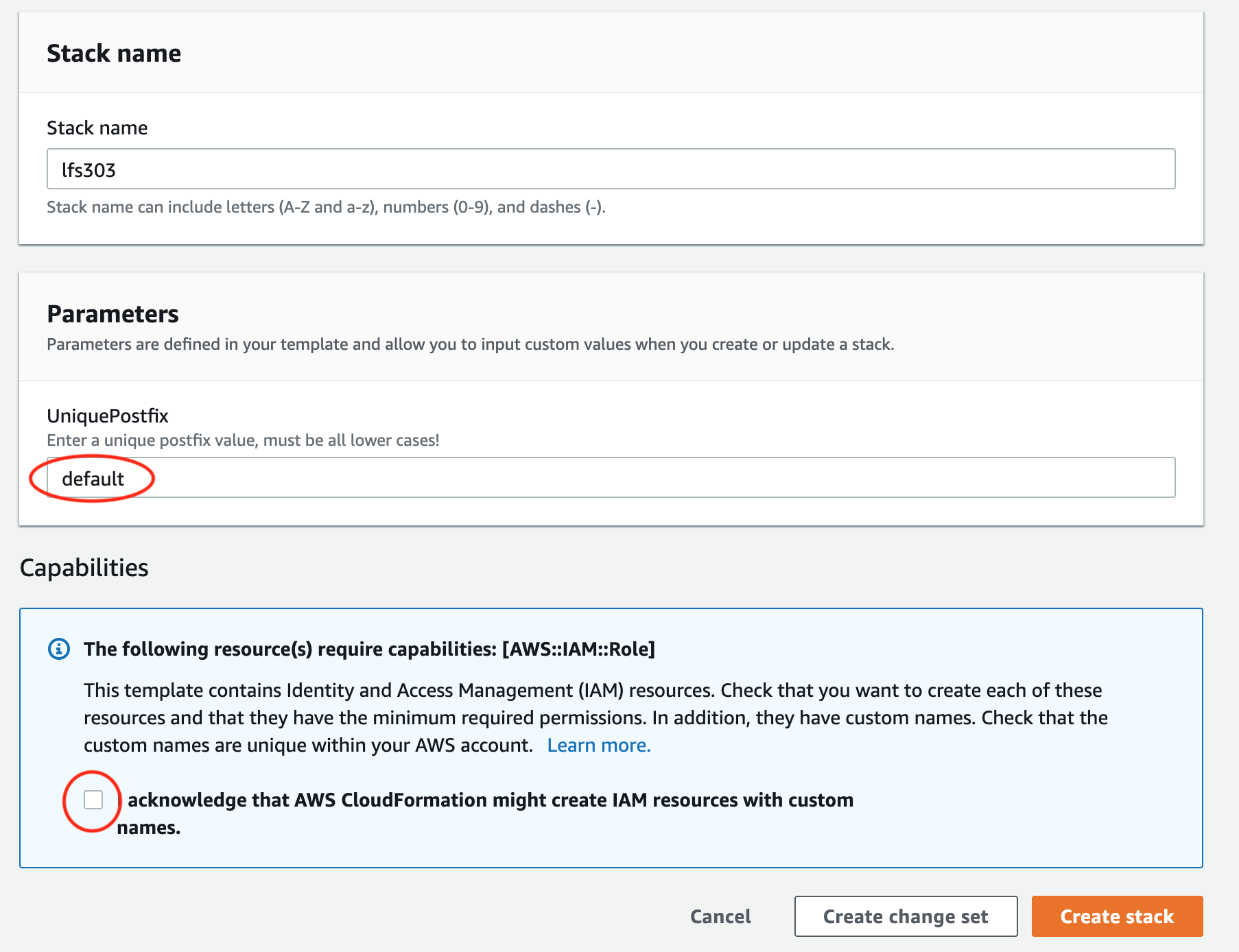


### Setup prerequisite with CloudFormation Template

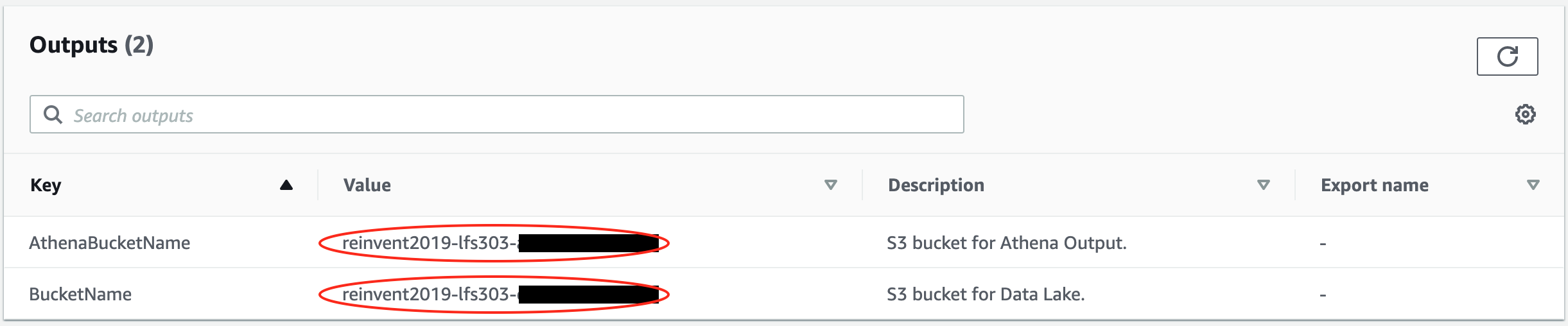
Once you logged into the provided AWS account AWS console, you have administrator access in that account. Click the CloudFormation link below to create the resources needed for the lab including S3 Buckets, IAM roles and Cloud9 environment

[CloudFormation Template](https://us-east-1.console.aws.amazon.com/cloudformation/home?region=us-east-1#/stacks/create/review?templateURL=https://reinvent2019-lfs303.s3.amazonaws.com/cloudformation_lfs303.json&stackName=lfs303)

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.

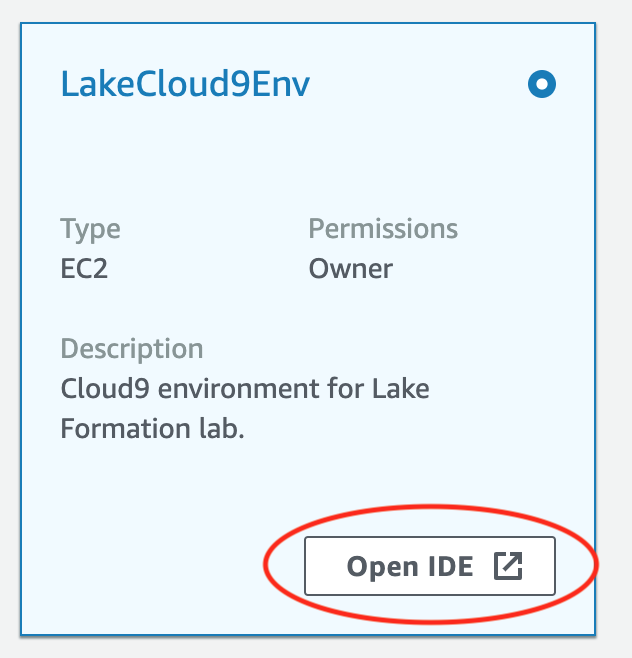


After the CloudFormation stack resources creation finished, go to the output tab and copy the value of <AthenaBucketName> and <BucketName> into a notepad. You will need the values for latter steps. If you go to S3 service now, you should see 2 buckets with those names created under your lab account.



### Copy Lab data into S3 Bucket

Go to Cloud9 service and open the Cloud9 IDE.



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.

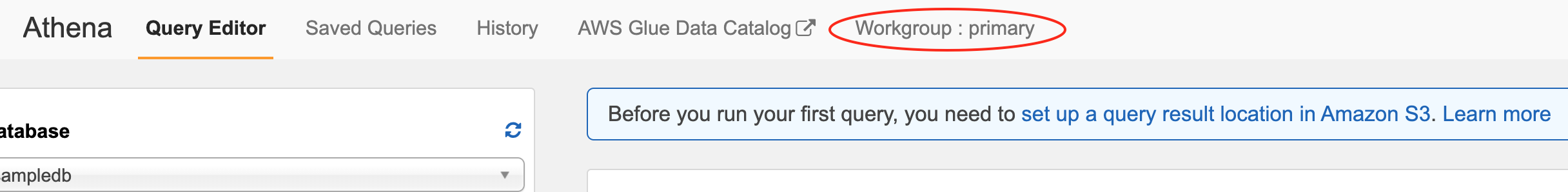
2019-11-23 01:41:53 9815312 NONPHI/NONPHI.csv

2019-11-23 01:41:54 4882781 PHI/PHI.csv

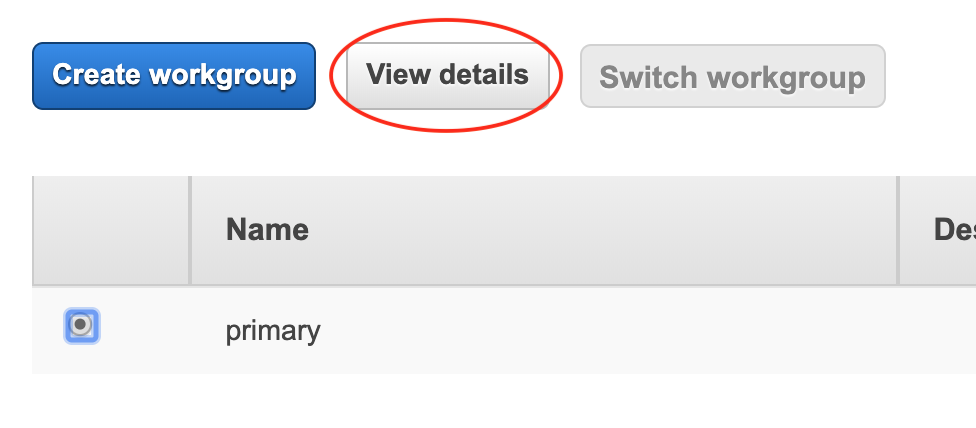
Closed the browser tab of Cloud9 IDE.

### Config Athena Output location for primary workgroup

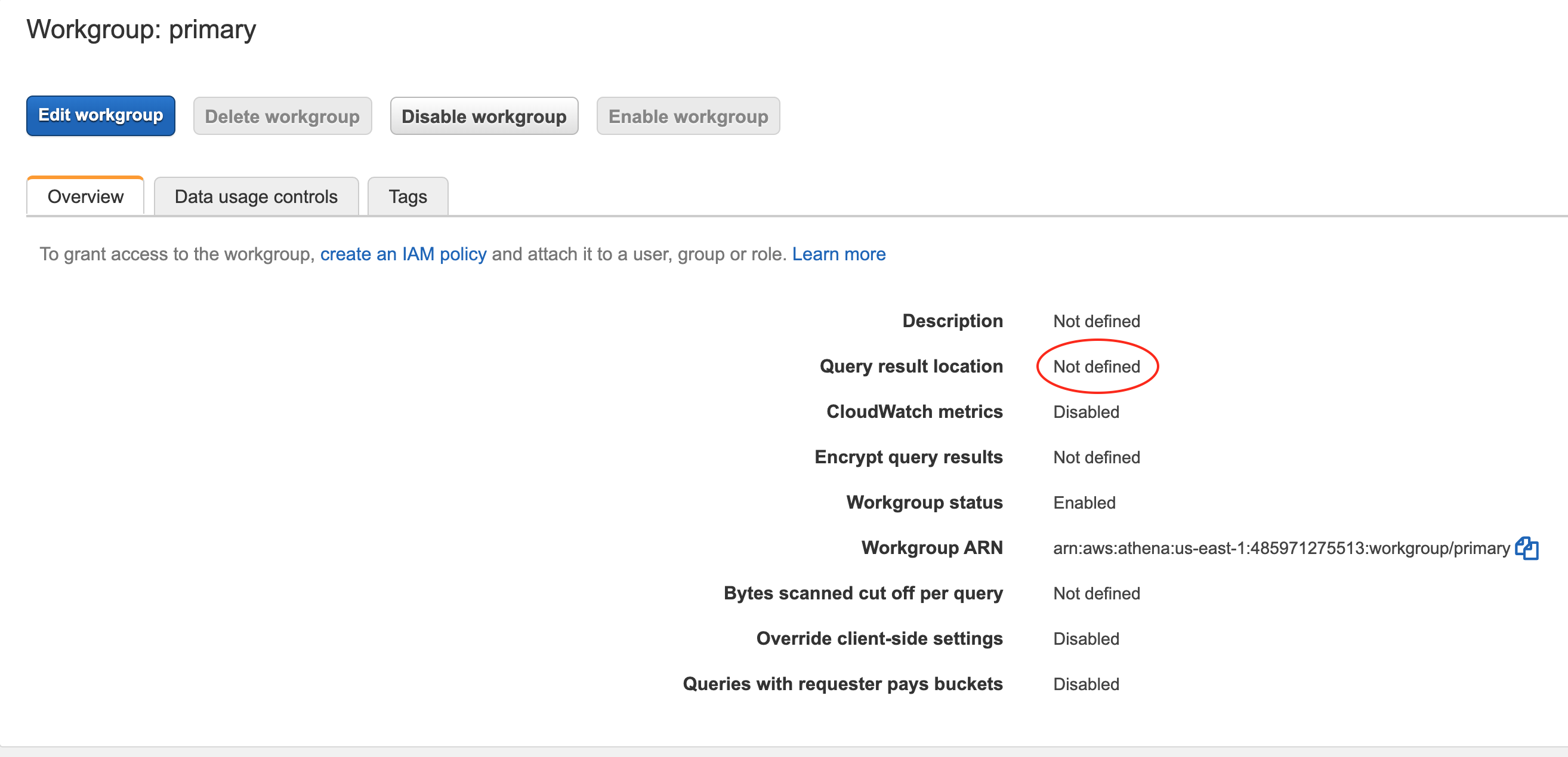
Go to Athena service in AWS Console, click “Get Started”. Before you could use Athena, you need to setup a query result location in S3 and we will use the S3 bucket we created earlier as the result bucket. Click “Workgroup : primary” tab



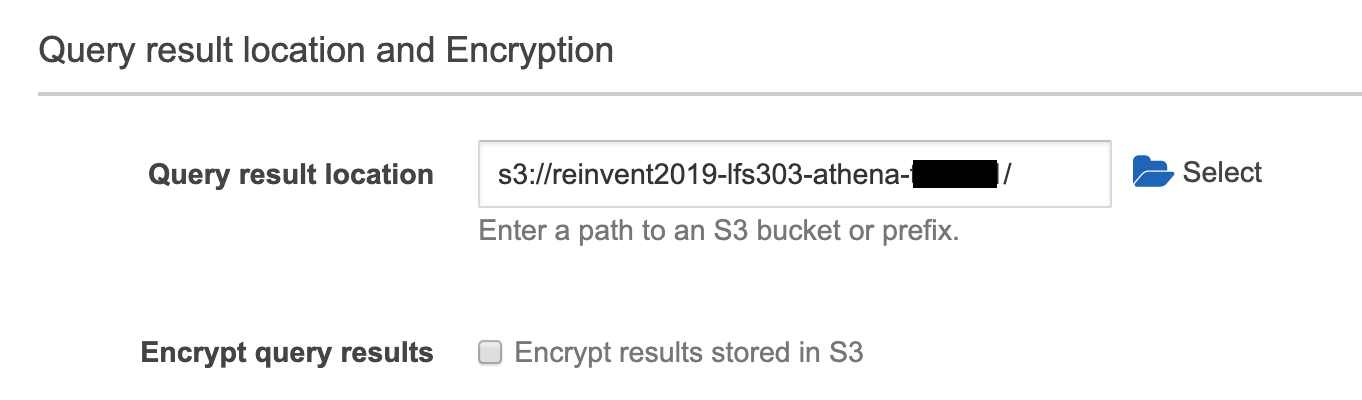
Choose primary group and click “View details”



Notice the Query result location is not defined

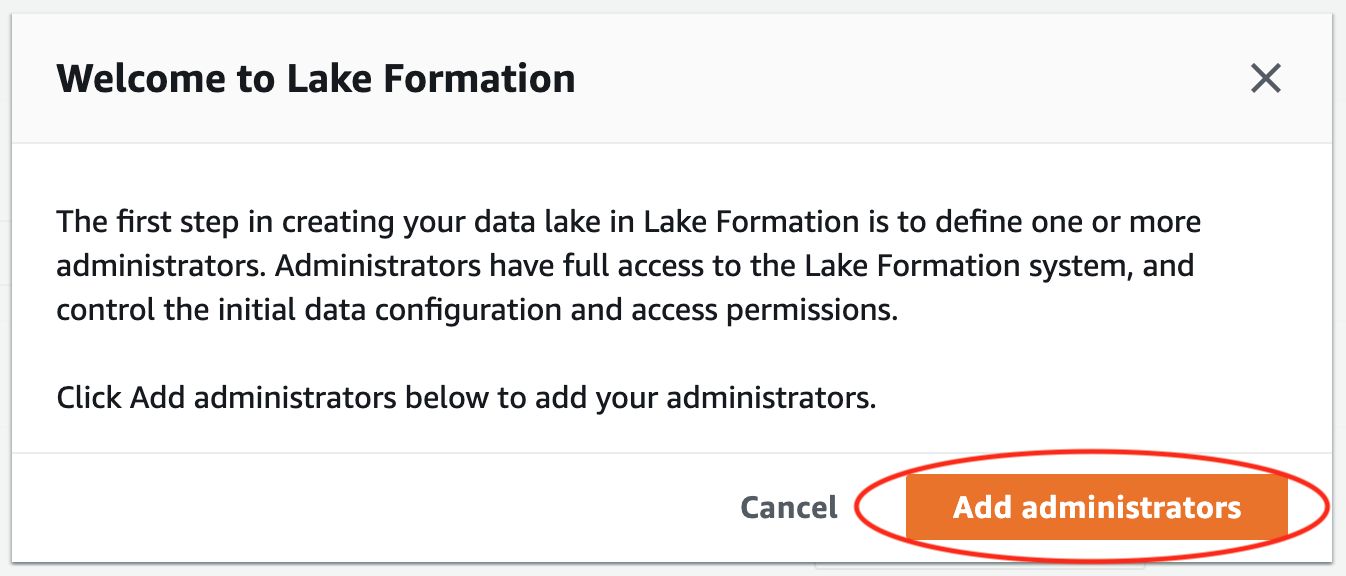


Click “Edit workgroup” and select the <AthenaBucketName> as the Query result location, then click “Save”

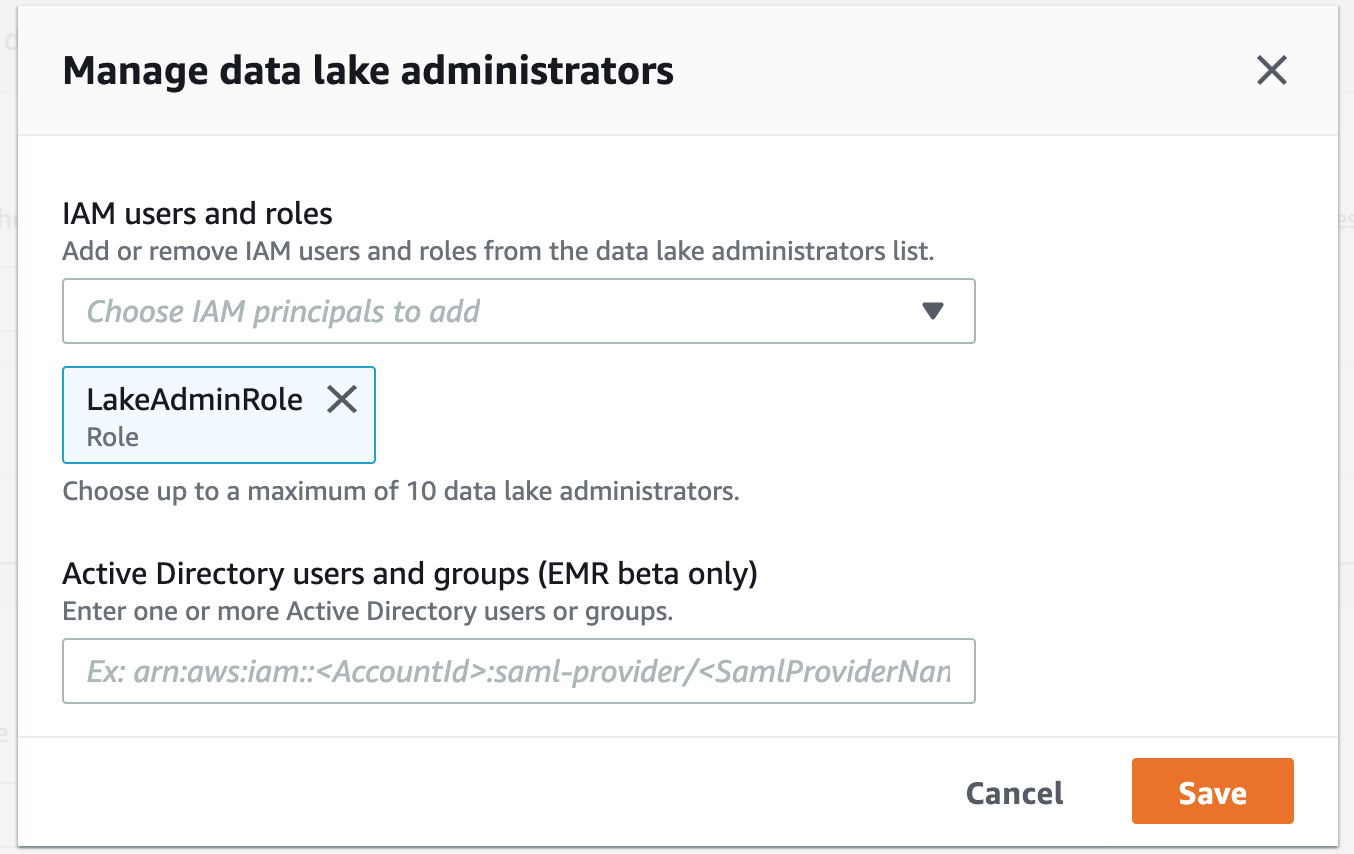


### Config Lake Formation

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

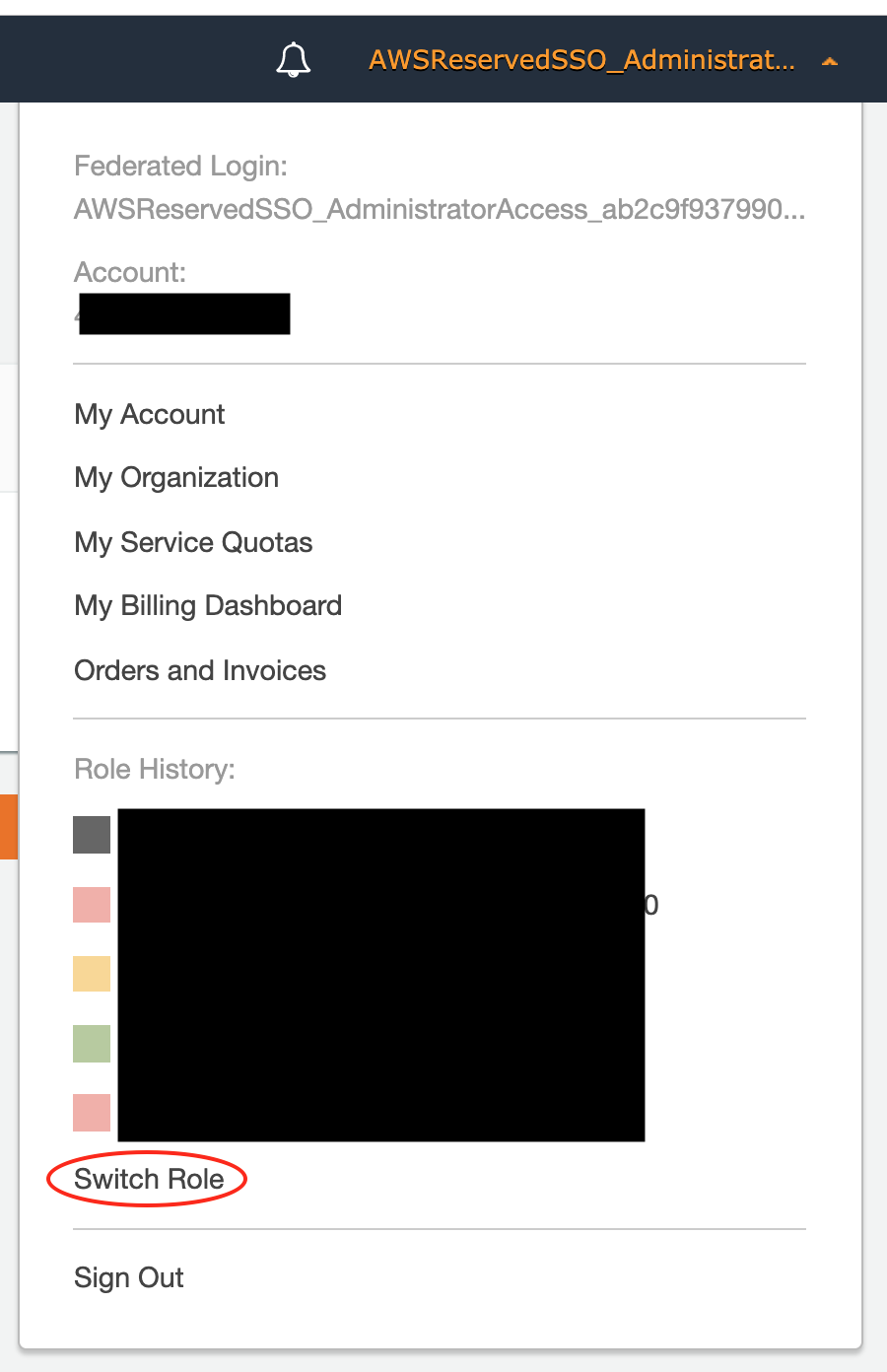


Choose “LakeAdminRole” in the dropdown and click save. You will now see the Lake Formation UI.

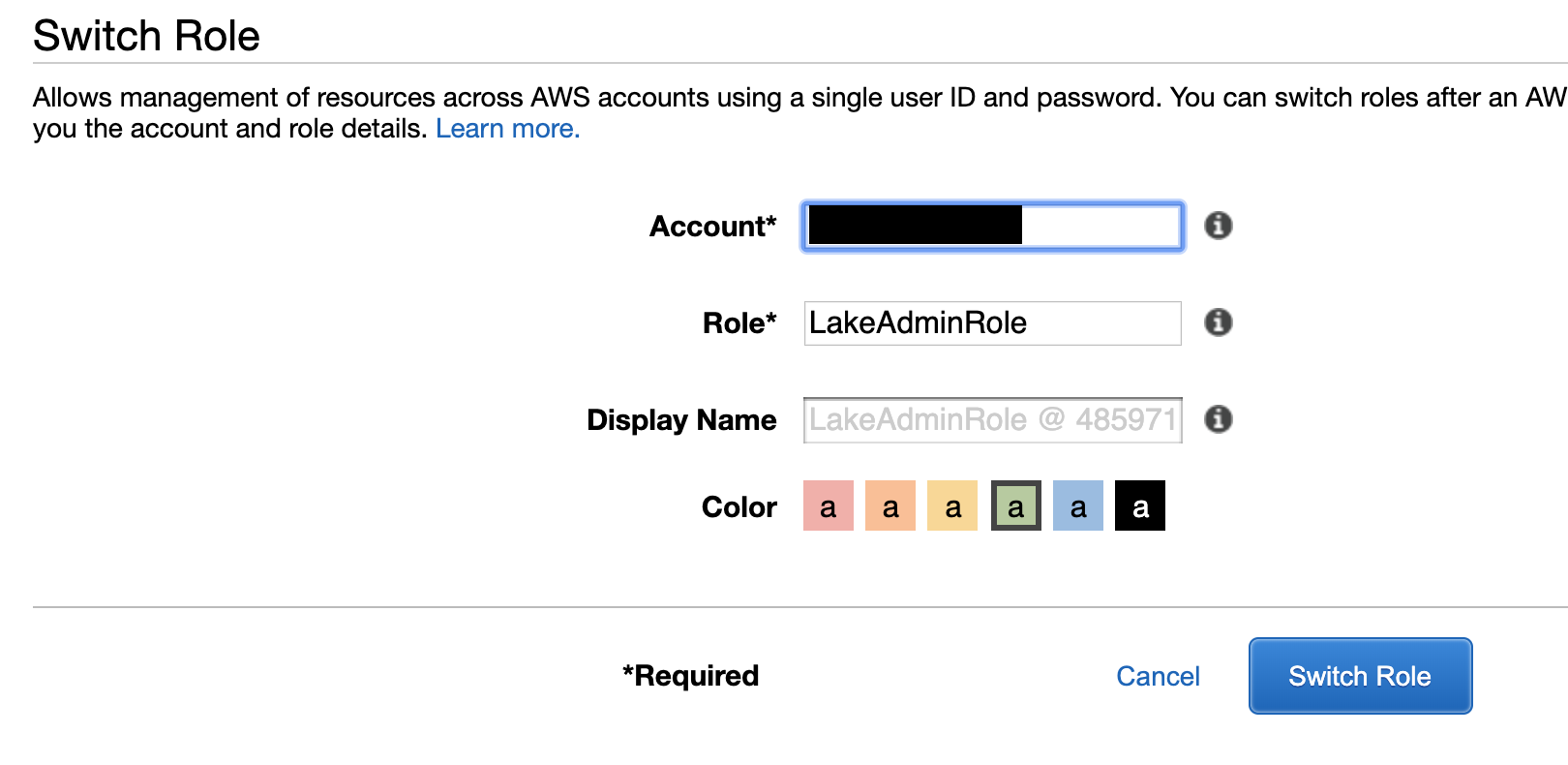


Note: switching roles in AWS Console in Lake Formation seems to be problematic. If you can’t switch role in Lake Formation AWS Console, change to another service and switch role there.

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

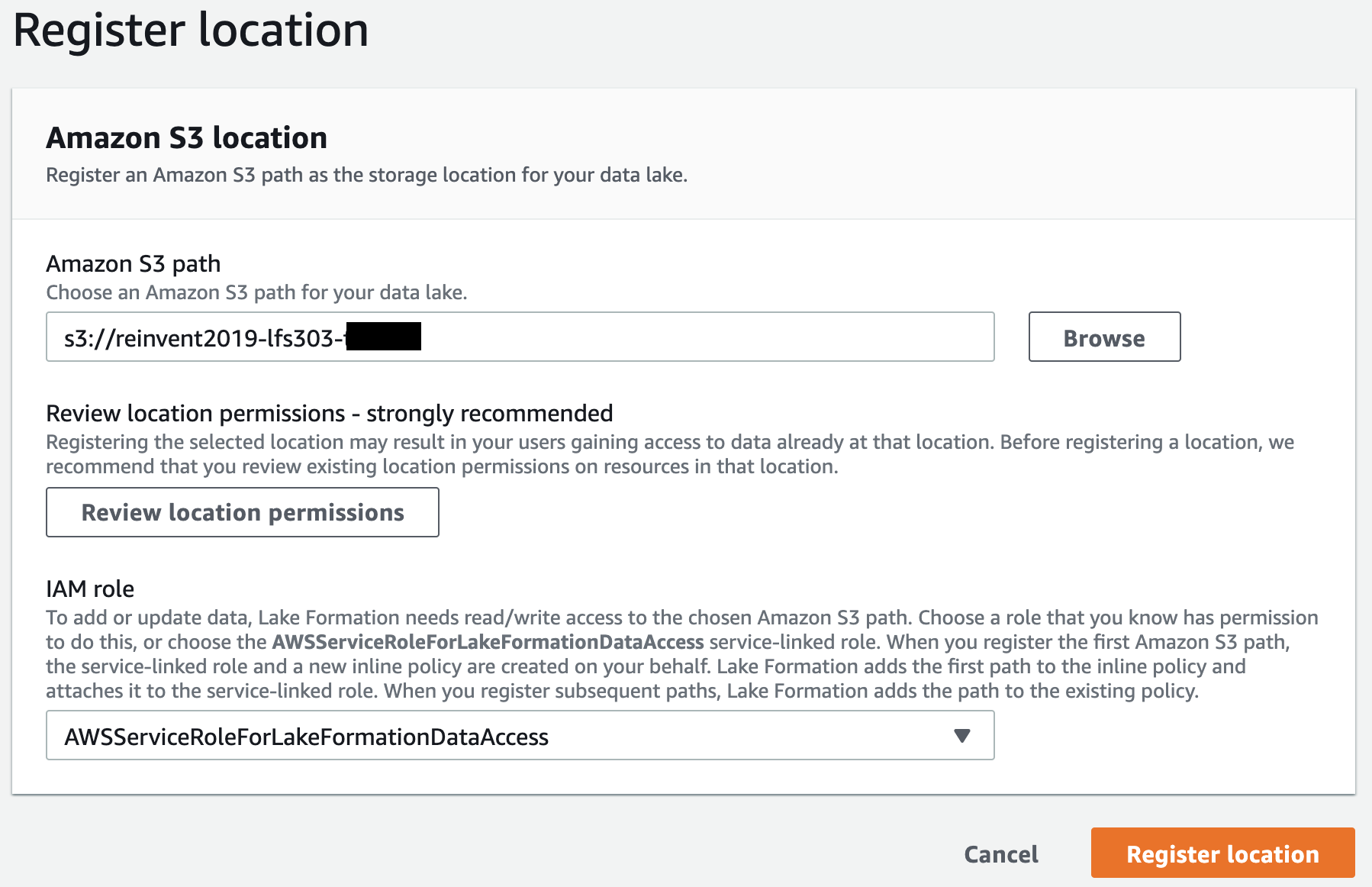


Enter the account number after removing all dashes and set Role name to “LakeAdminRole”. Click “Switch Role”.



### Register your Amazon S3 storage

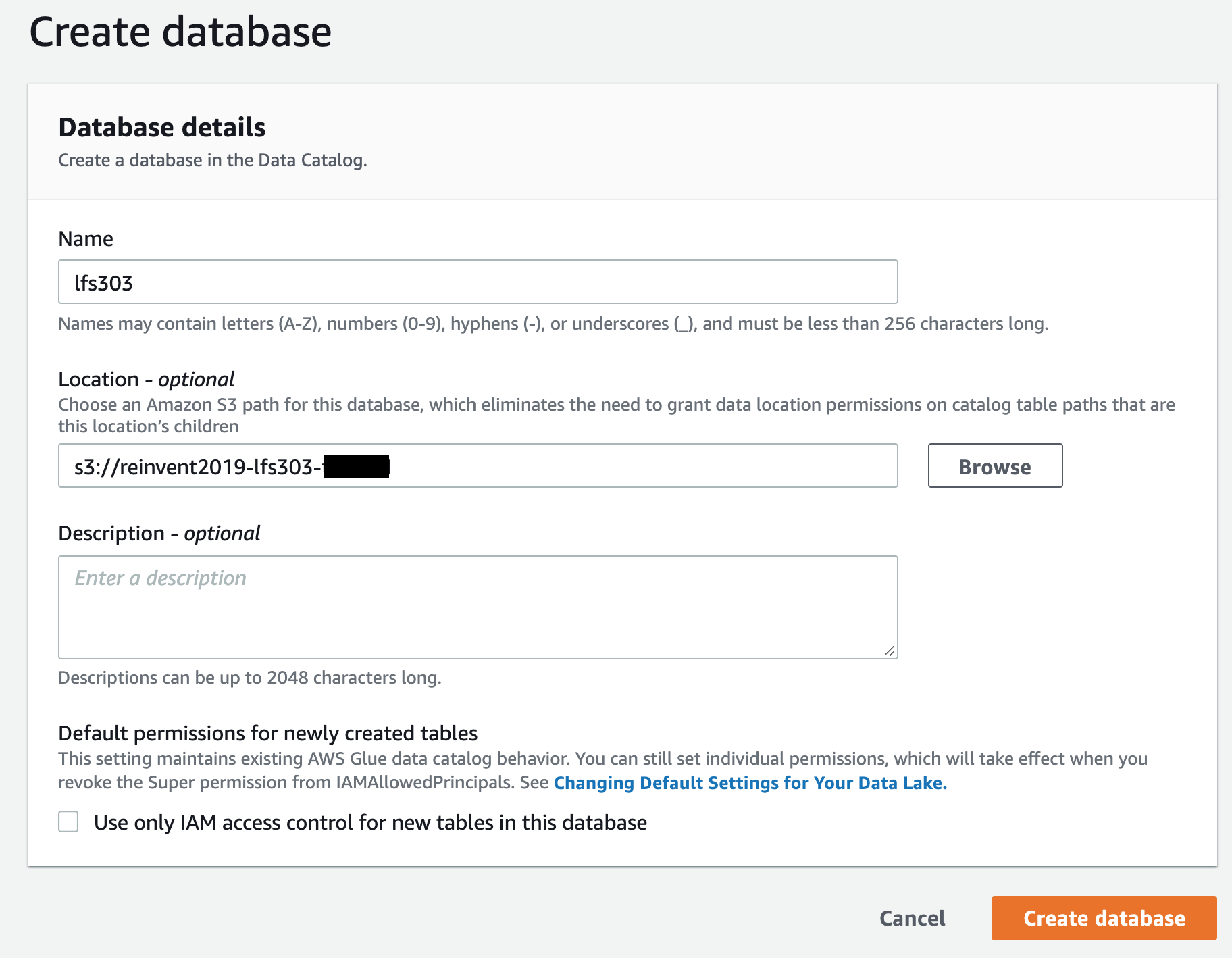
Lake Formation manages access to designated storage locations within Amazon S3. Register the storage locations that you want to be part of the data lake. Click “Data lake location” under Lake Formation and register the S3 bucket created by the CloudFormation template earlier.



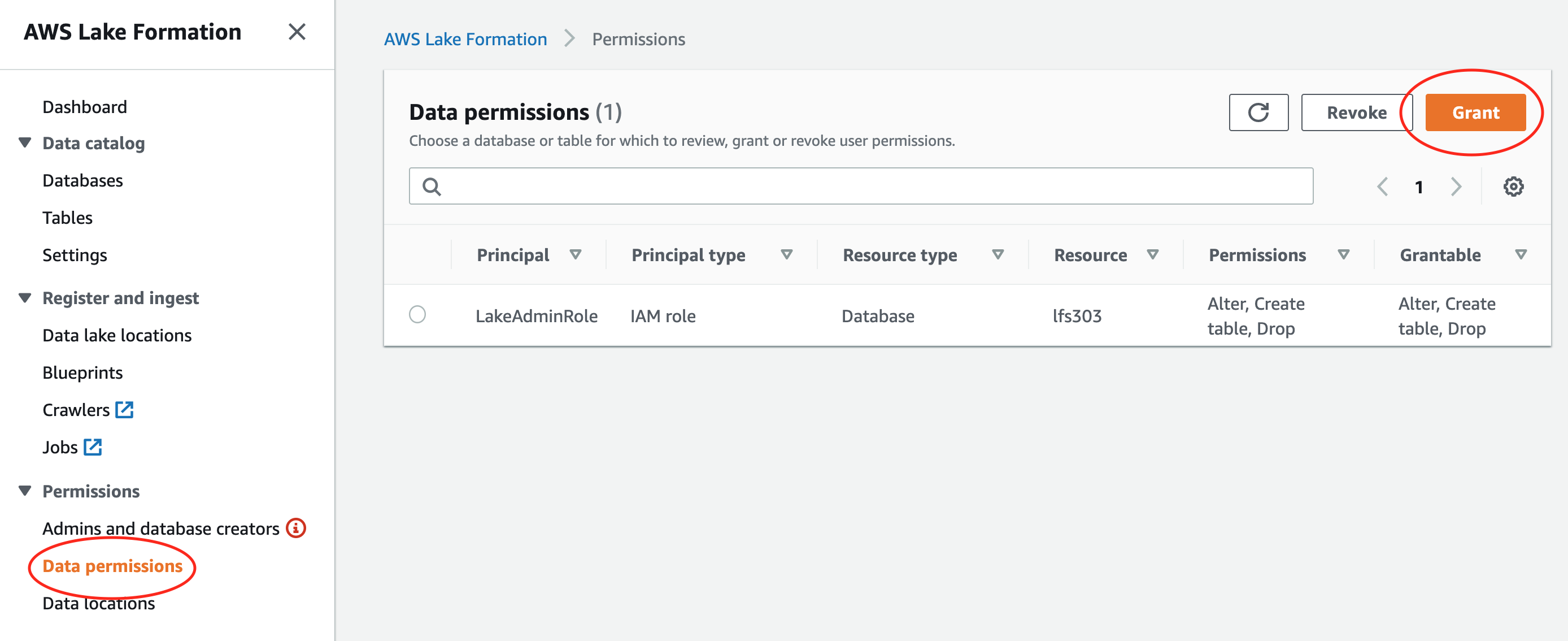
### Create a database and Grant access to Crawler role

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.

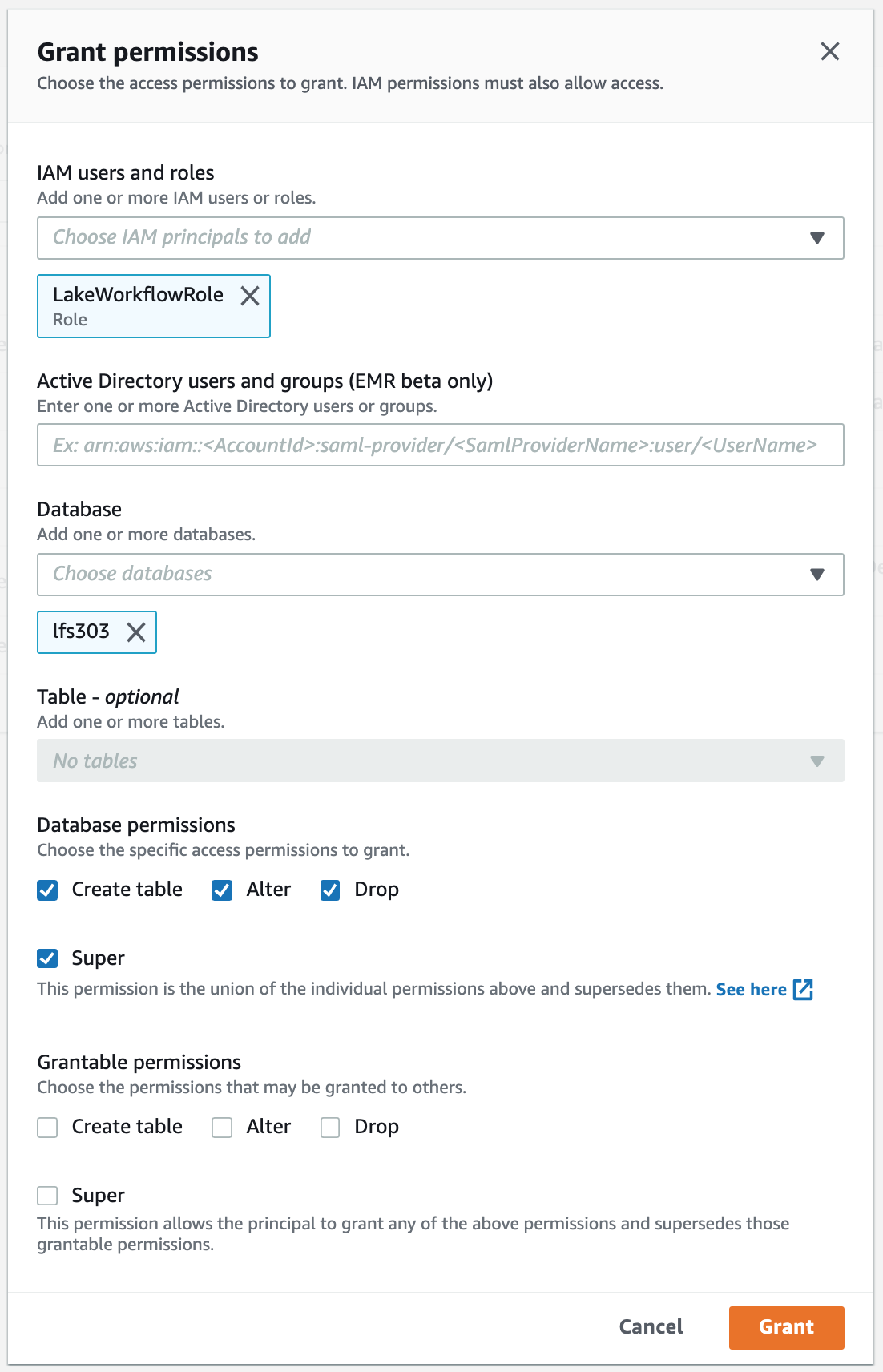
Click “Database” under Lake Formation and create Database and choose the location as the S3 bucket created by CloudFormation template. Uncheck “Use only IAM access control for new tables in this database” if it is checked



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

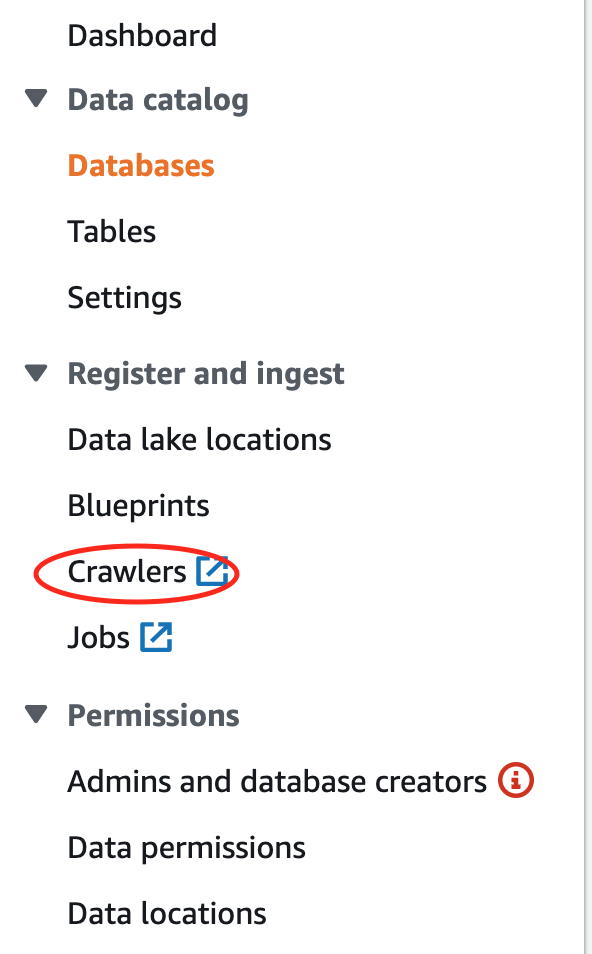


Grant “LakeWorkflowRole” permissions to create/alter/drop tables in the newly database.

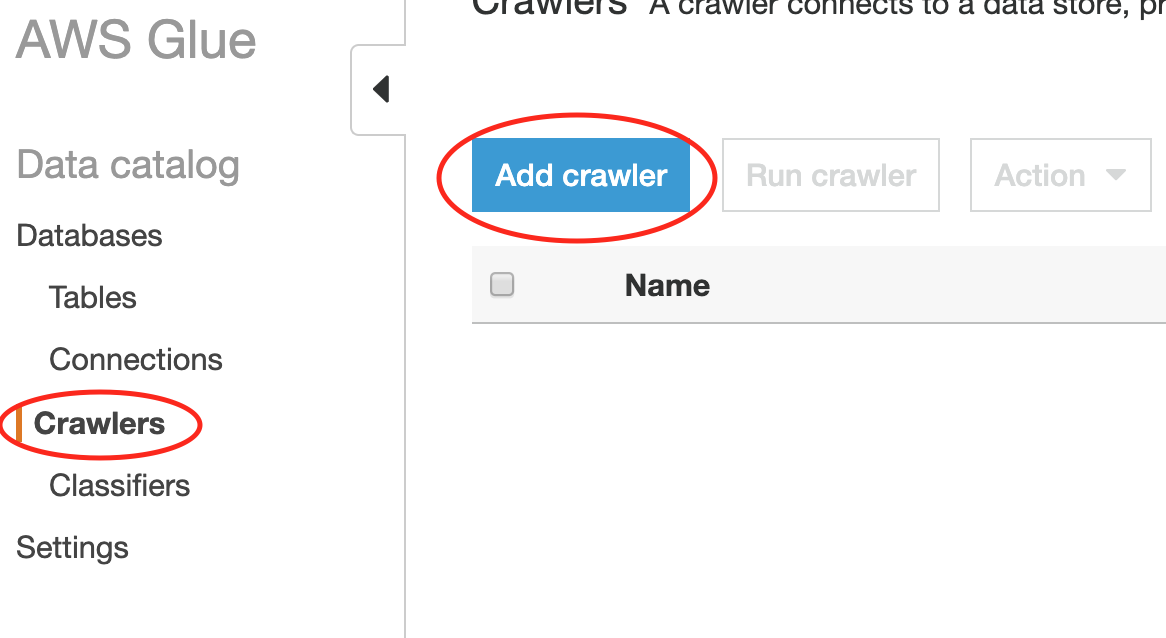


### Use Glue Crawler to create tables automatically

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



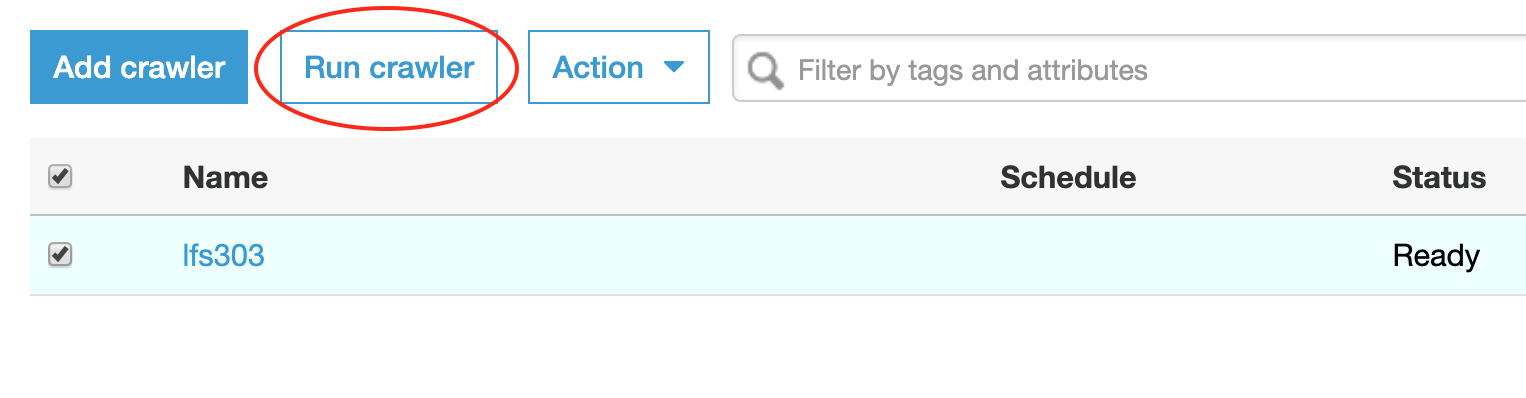
Click Crawler on the left and “Add crawler” to enter Crawler configuration screen



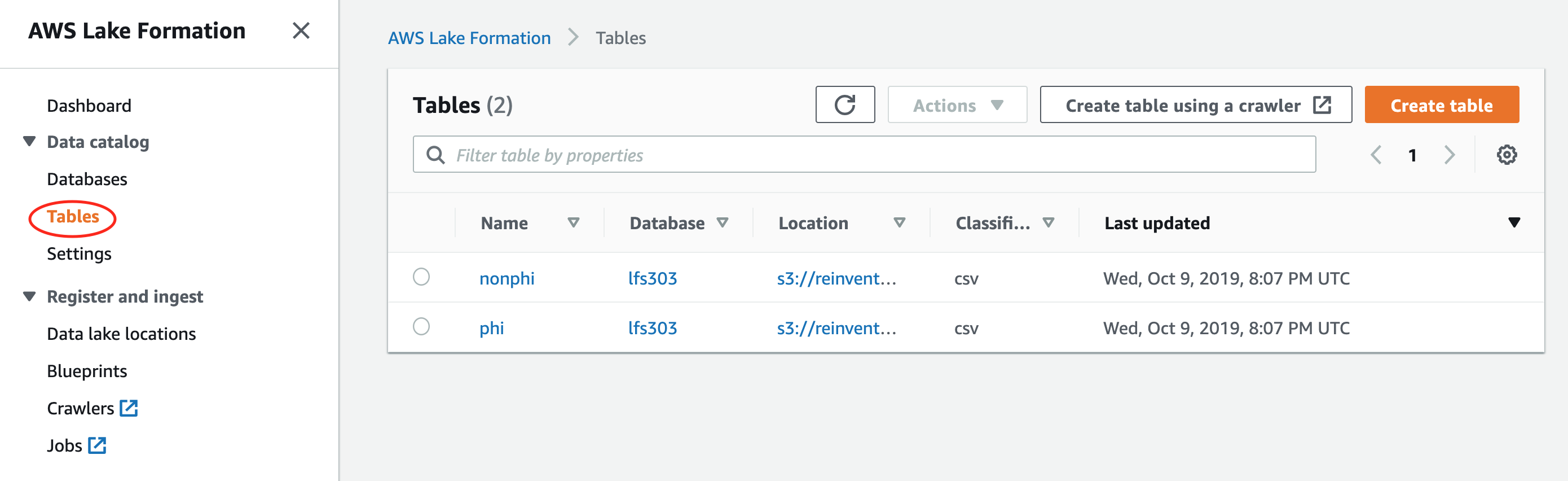
Put the following settings for the new crawler:

* 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).

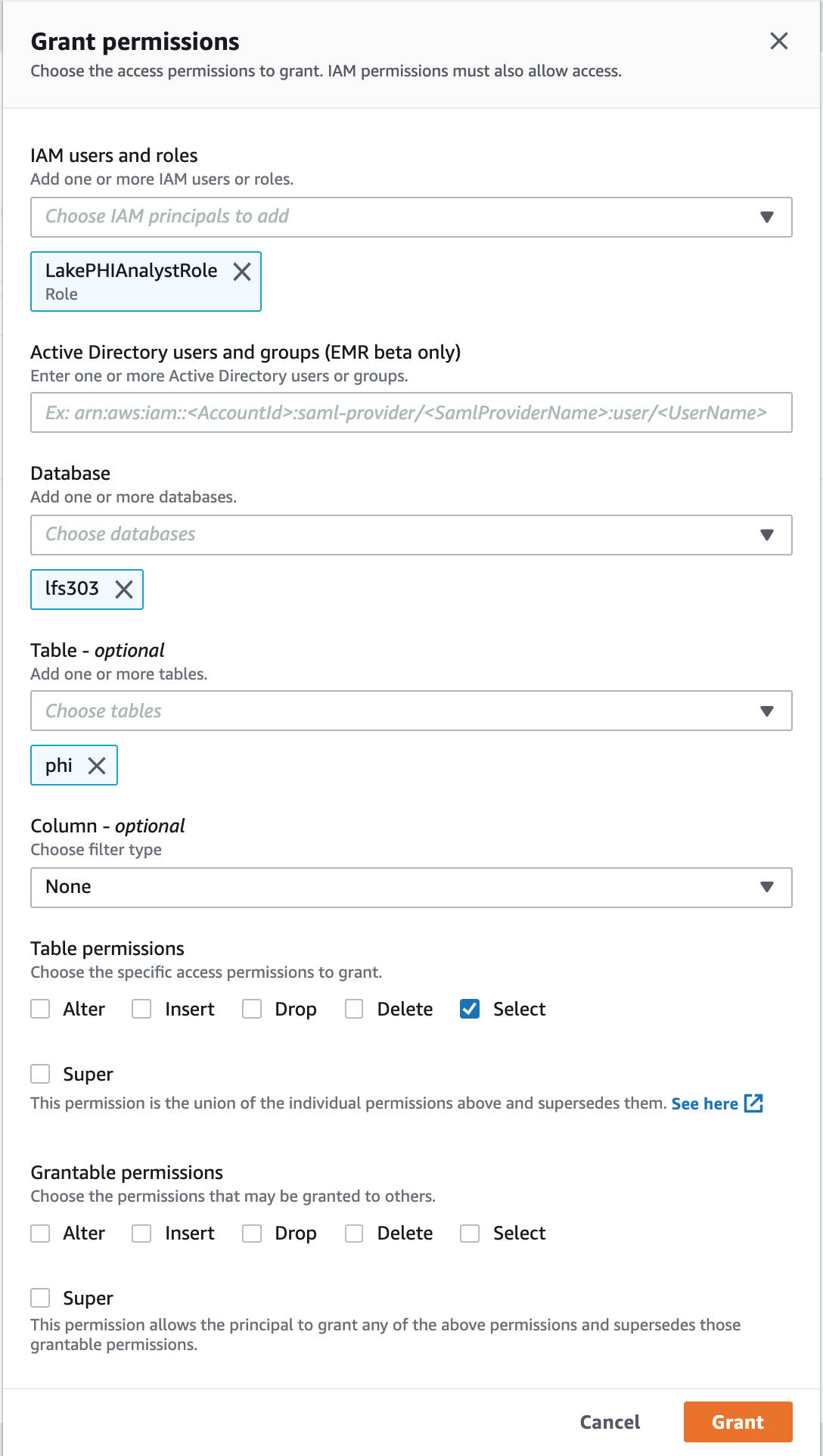
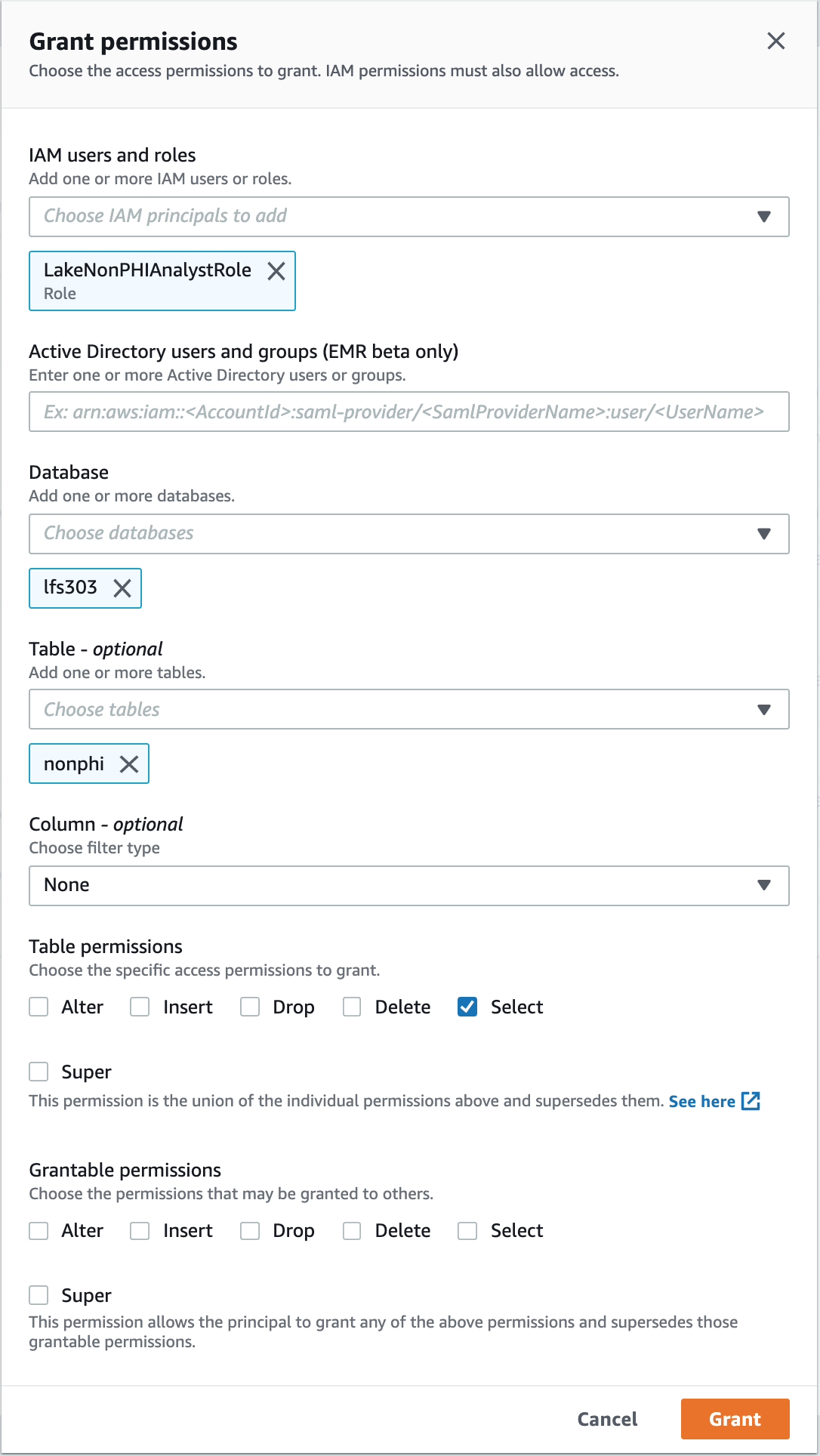


Close the Crawler browser tab and go back to Lake Formation tab. Click “Tables” under Lake Formation and you should see the 2 new tables created by Crawler.



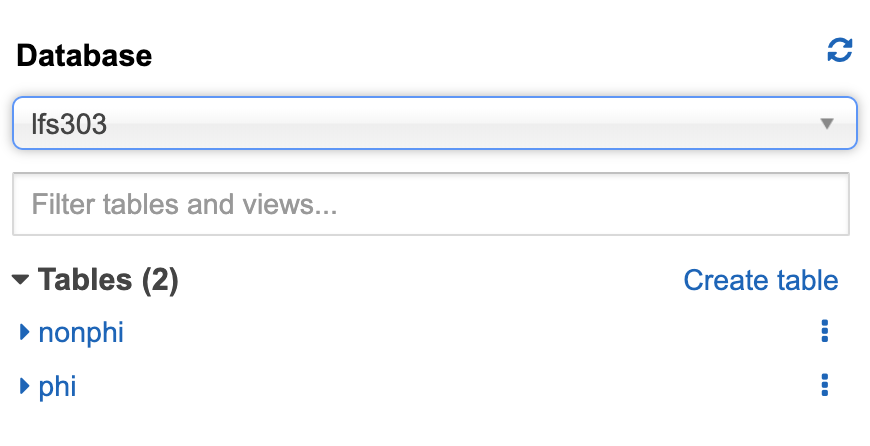
### Grant Data Permission to Analyst roles

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)

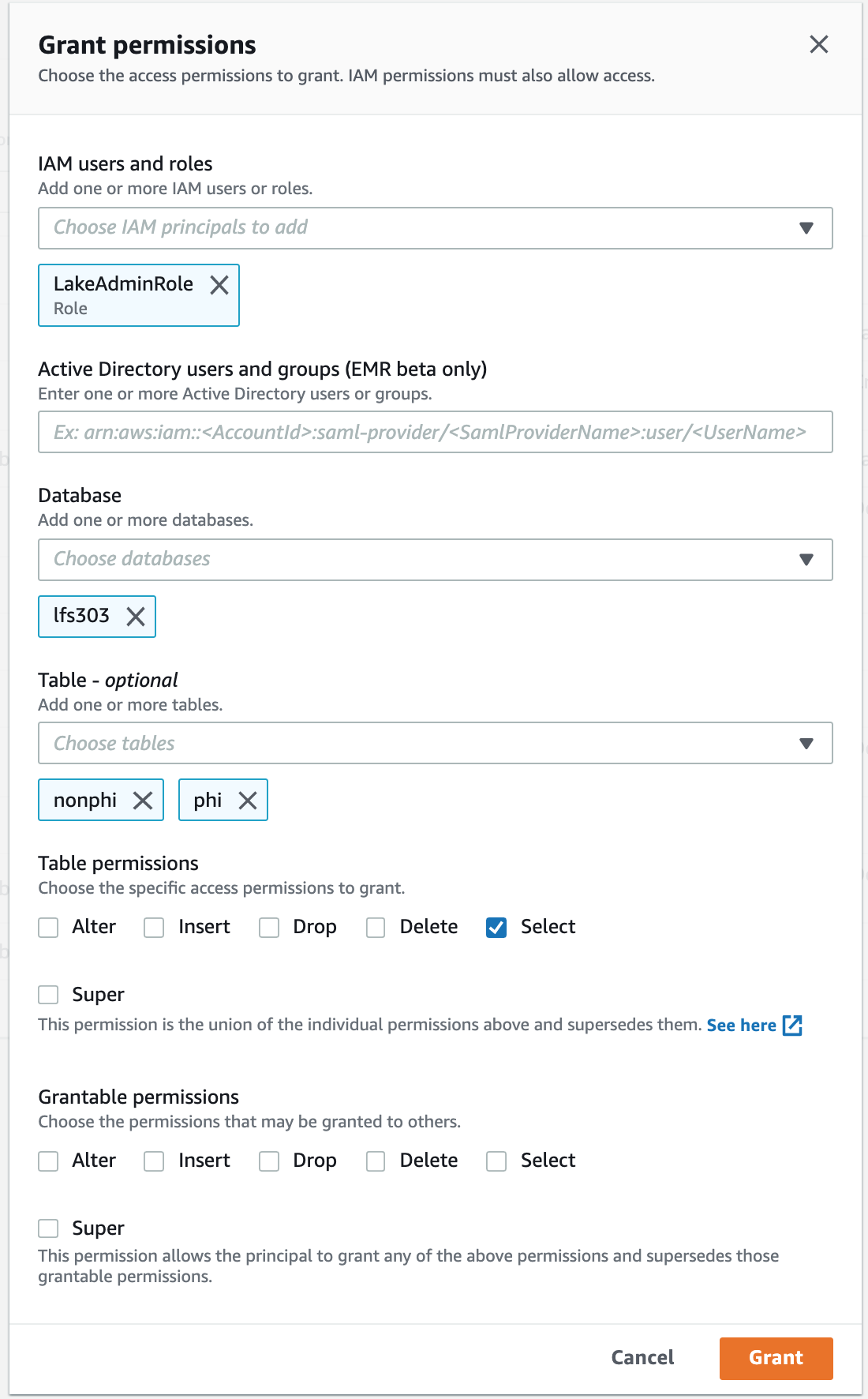


### Query data in Athena

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.



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.

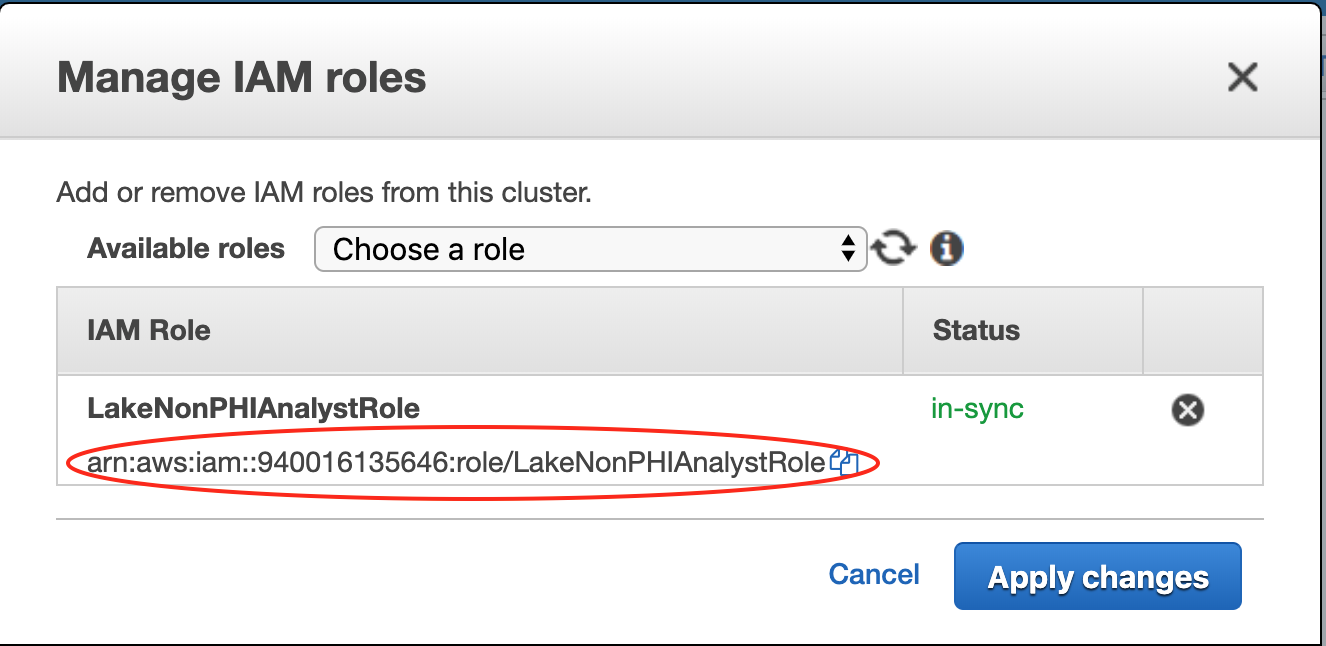
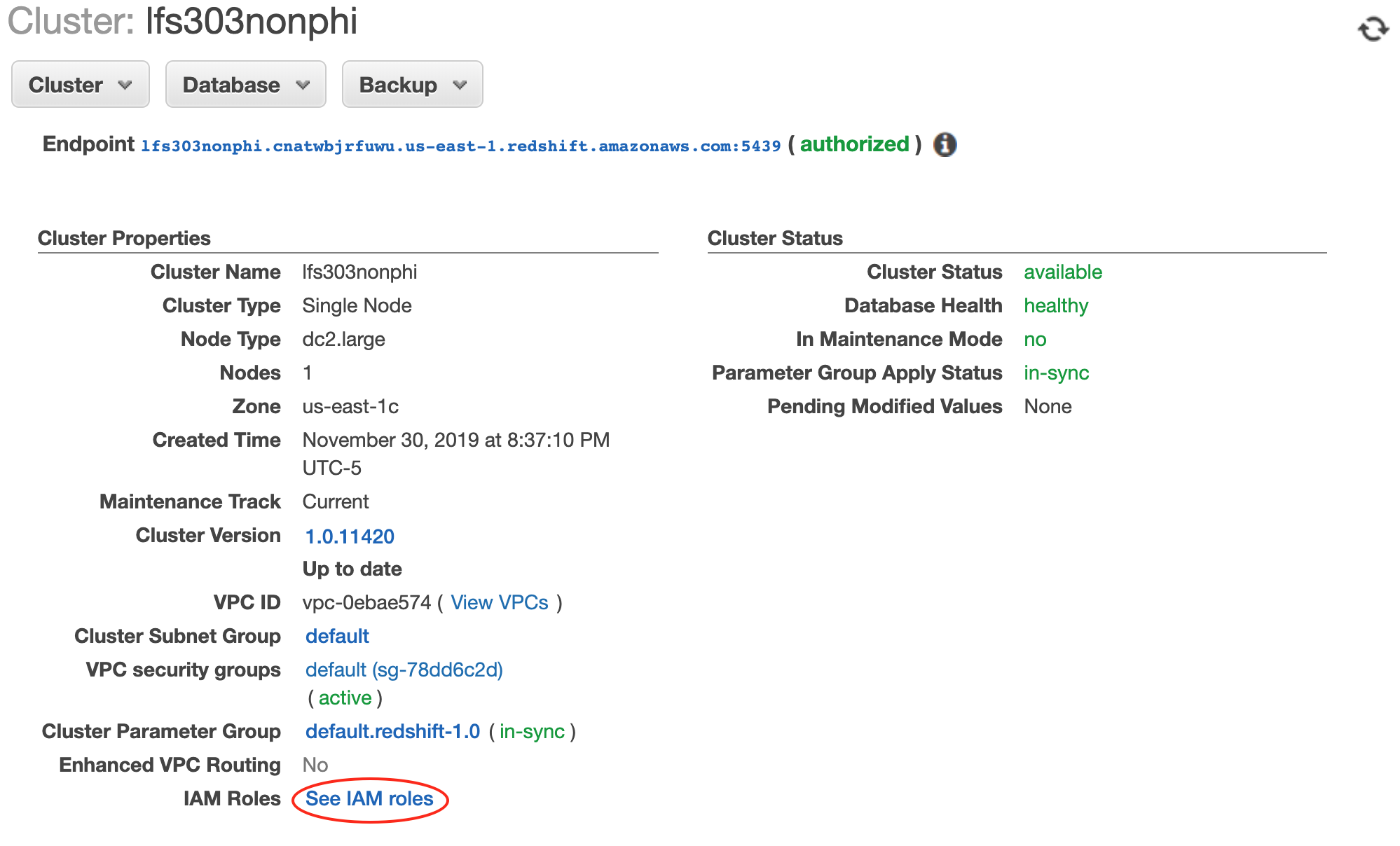


Then switch roles to “LakePHIAnalystRole” and “LakeNonPHIAnalystRole” and see what is the difference inside Athena and try to understand why.

### 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.

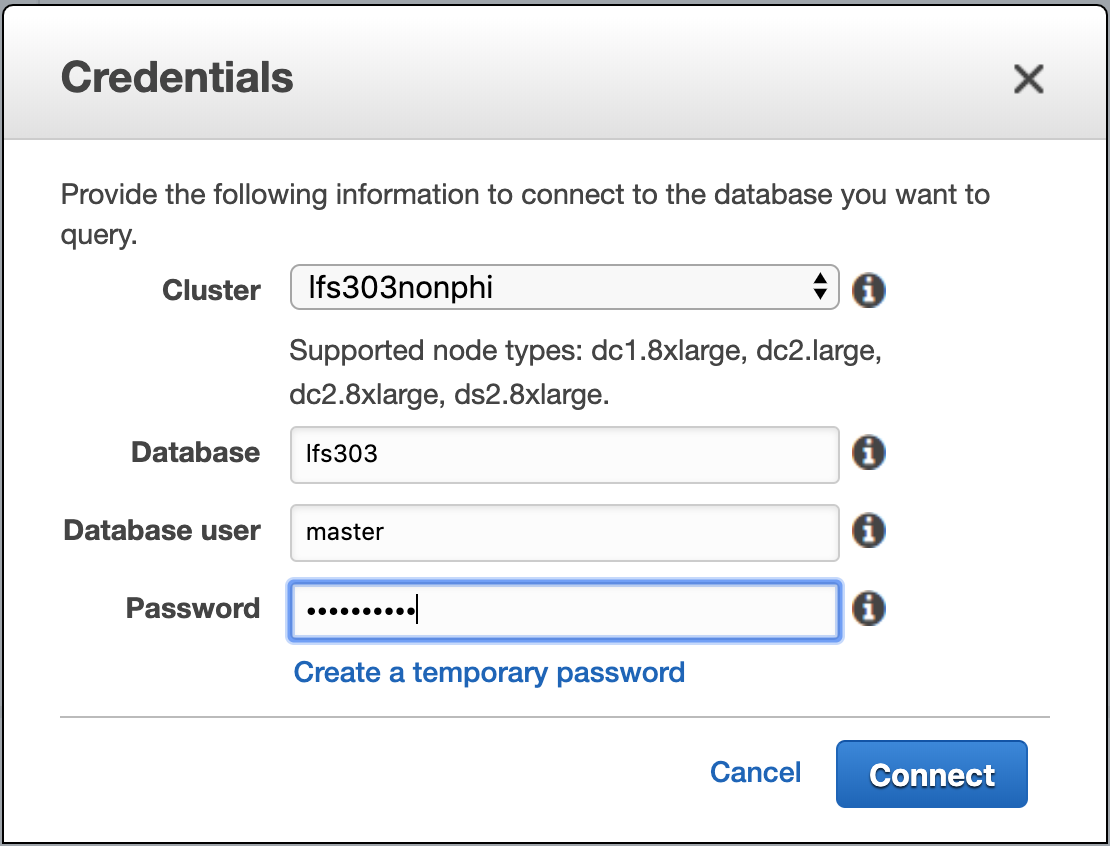
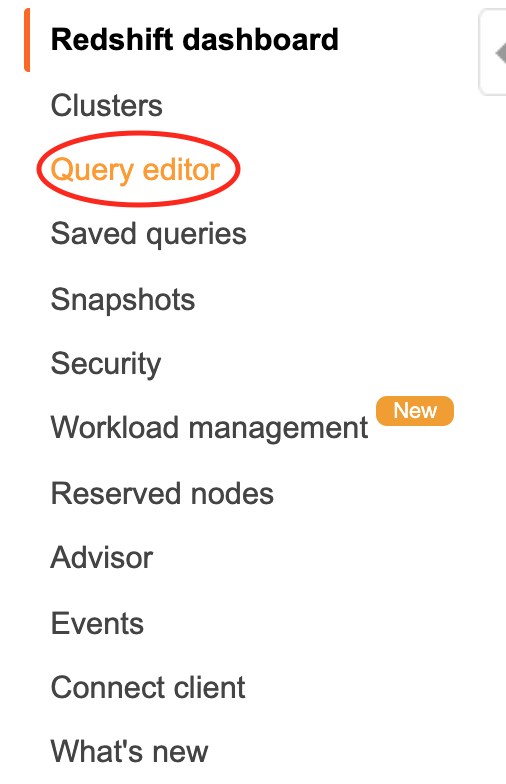


Click “query editor” on the left. Pick any of the 2 clusters and put in the following login information

Database: lfs303

Database user: master

Password: Password#1



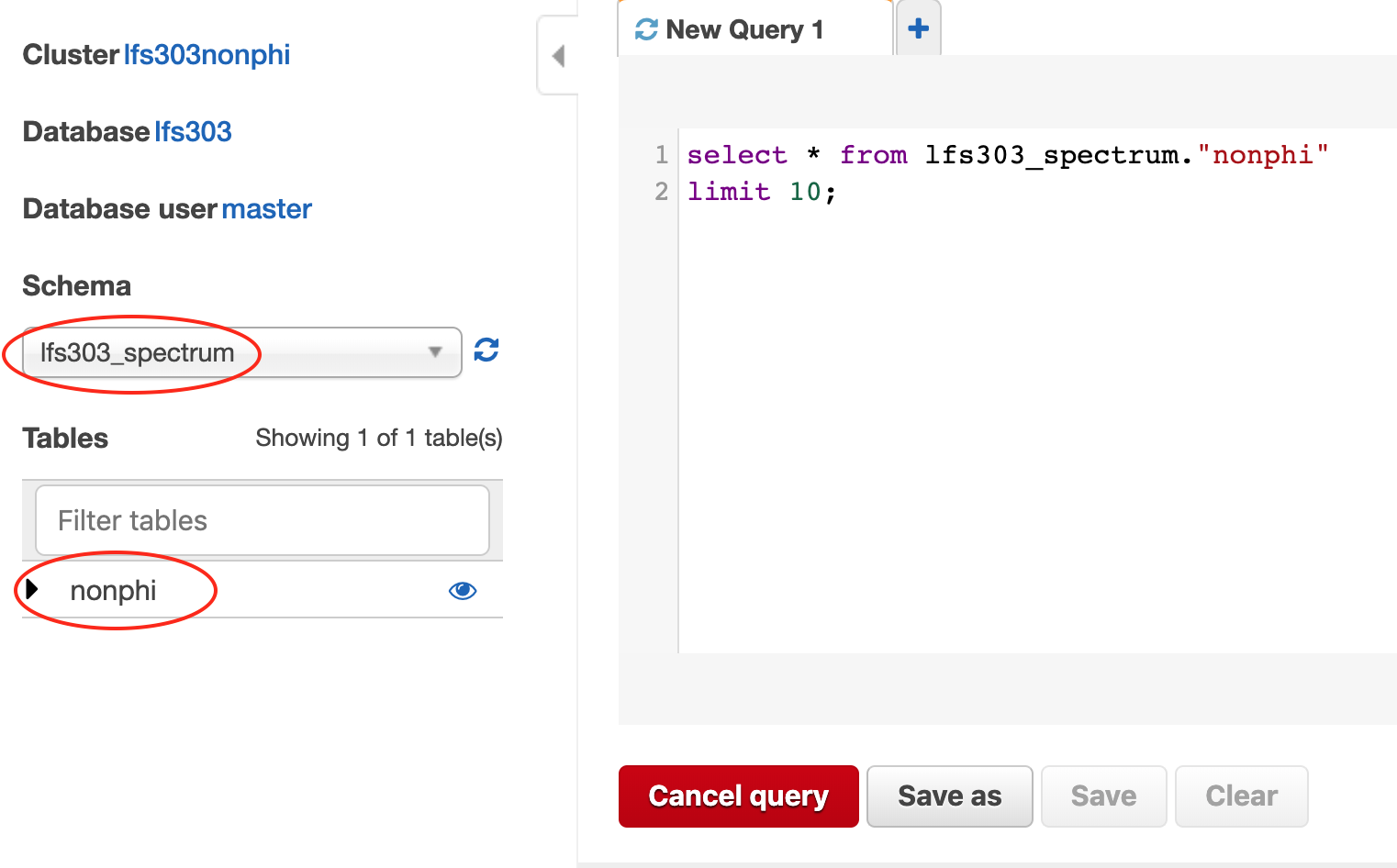
Once in the query editor, create an external schema in Redshift from Lake Formation data catalog with following command. For “Role Name”, put in “LakePHIAnalystRole” or “LakeNonPHIAnalystRole” depends on which cluster it is.

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.



### 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.