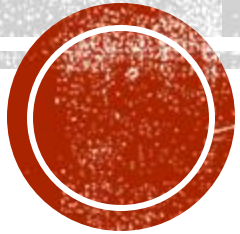
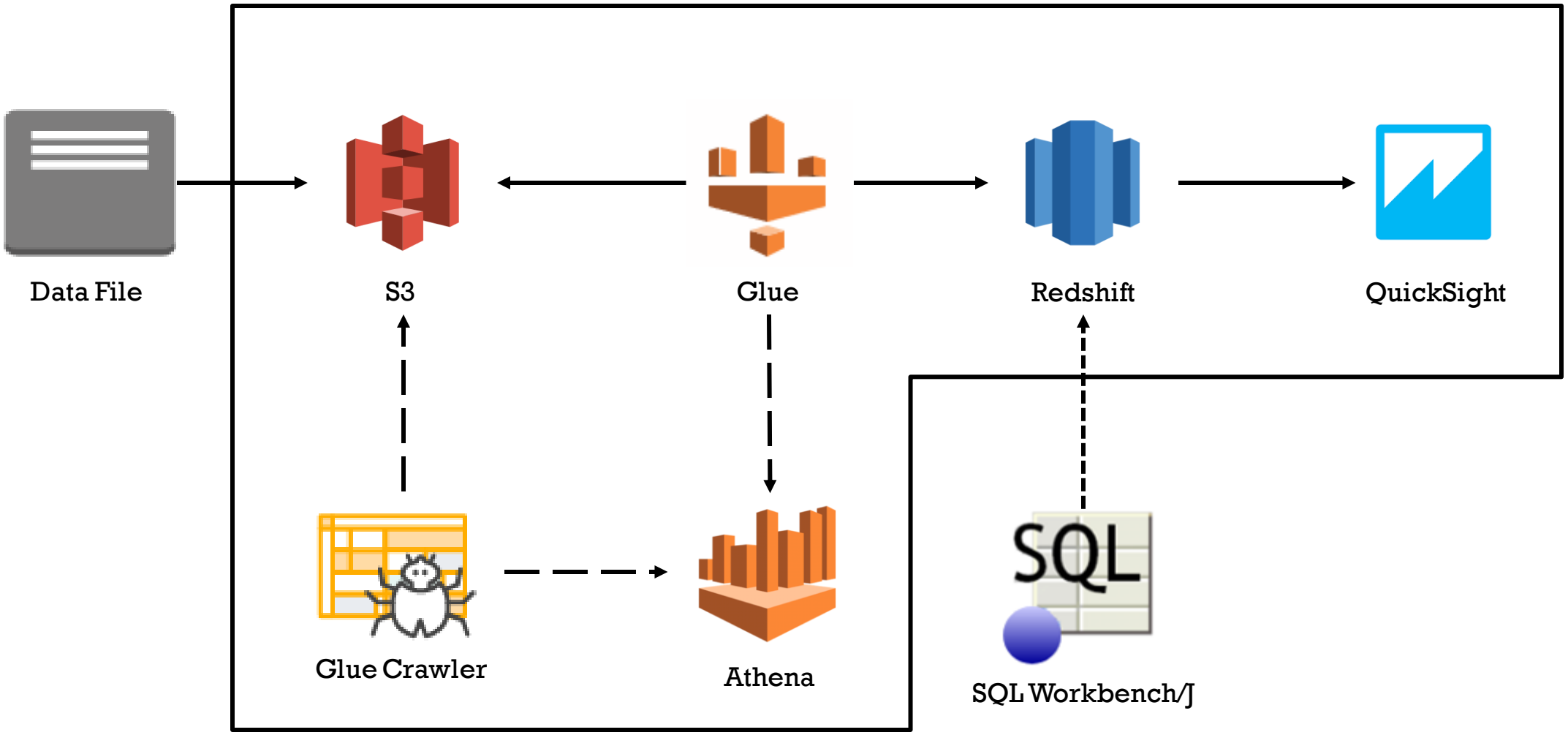
ARTS & CRAFTS WITH AWS GLUE



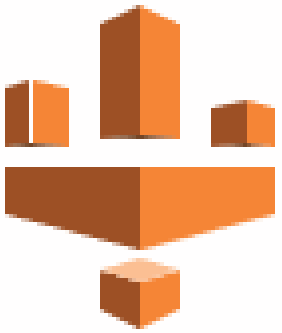
ETL Workshop

Amazon Web Services



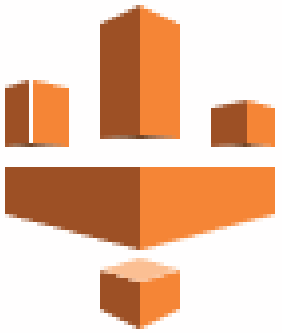
AWS Glue

What is Glue?



AWS Glue

- Amazon Web Services tool to Extract, Transform, and Load(ETL)
- Used to prepare data for business analytics



ETL

- Extract: Pull data from a source
  - Files
  - Database
  - Reporting Tool
- Transform: Modify the data to fit your needs
  - Add new columns like data source or timestamp
  - Remove unwanted data
  - Alter data with calculations
- Load: Store in your database

ETL

Original Data File(WA_Sales_Products.csv)

	A	B	C	D	E	F	G	H	I	J	K
1	Retailer country	Order method type	Retailer type	Product line	Product type	Product	Year	Quarter	Revenue	Quantity	Gross margin
2	United States	Fax	Outdoors Shop	Camping Equipment	Cooking Gear	TrailChef Deluxe C	2012	Q1 2012	59628.66	489	0.347548
3	United States	Fax	Outdoors Shop	Camping Equipment	Cooking Gear	TrailChef Double F	2012	Q1 2012	35950.32	252	0.474275
4	United States	Fax	Outdoors Shop	Camping Equipment	Tents	Star Dome	2012	Q1 2012	89940.48	147	0.352772
5	United States	Fax	Outdoors Shop	Camping Equipment	Tents	Star Gazer 2	2012	Q1 2012	165883.4	303	0.282938
6	United States	Fax	Outdoors Shop	Camping Equipment	Sleeping Bags	Hibernator Lite	2012	Q1 2012	119822.2	1415	0.29145
7	United States	Fax	Outdoors Shop	Camping Equipment	Sleeping Bags	Hibernator Extrem	2012	Q1 2012	87728.96	352	0.398146
8	United States	Fax	Outdoors Shop	Camping Equipment	Sleeping Bags	Hibernator Camp C	2012	Q1 2012	41837.46	426	0.335607
9	United States	Fax	Outdoors Shop	Camping Equipment	Lanterns	Firefly Lite	2012	Q1 2012	8268.41	577	0.52896
10	United States	Fax	Outdoors Shop	Camping Equipment	Lanterns	Firefly Extreme	2012	Q1 2012	9393.3	189	0.434205
11	United States	Fax	Outdoors Shop	Camping Equipment	Lanterns	EverGlow Single	2012	Q1 2012	19396.5	579	0.461493
12	United States	Fax	Outdoors Shop	Camping Equipment	Lanterns	EverGlow Butane	2012	Q1 2012	6940.03	109	0.361866
13	United States	Fax	Outdoors Shop	Mountaineering Equip	Rope	Husky Rope 50	2012	Q1 2012	20003.2	133	0.329056
14	United States	Fax	Outdoors Shop	Mountaineering Equip	Rope	Husky Rope 60	2012	Q1 2012	14109.4	79	0.291657
15	United States	Fax	Outdoors Shop	Mountaineering Equip	Rope	Husky Rope 100	2012	Q1 2012	73970.22	227	0.301264

Example Business Requirements:

- Remove the Year from Quarter
- Add a profit column from revenue * gross margin columns
- Add a current date column

Why use Glue?

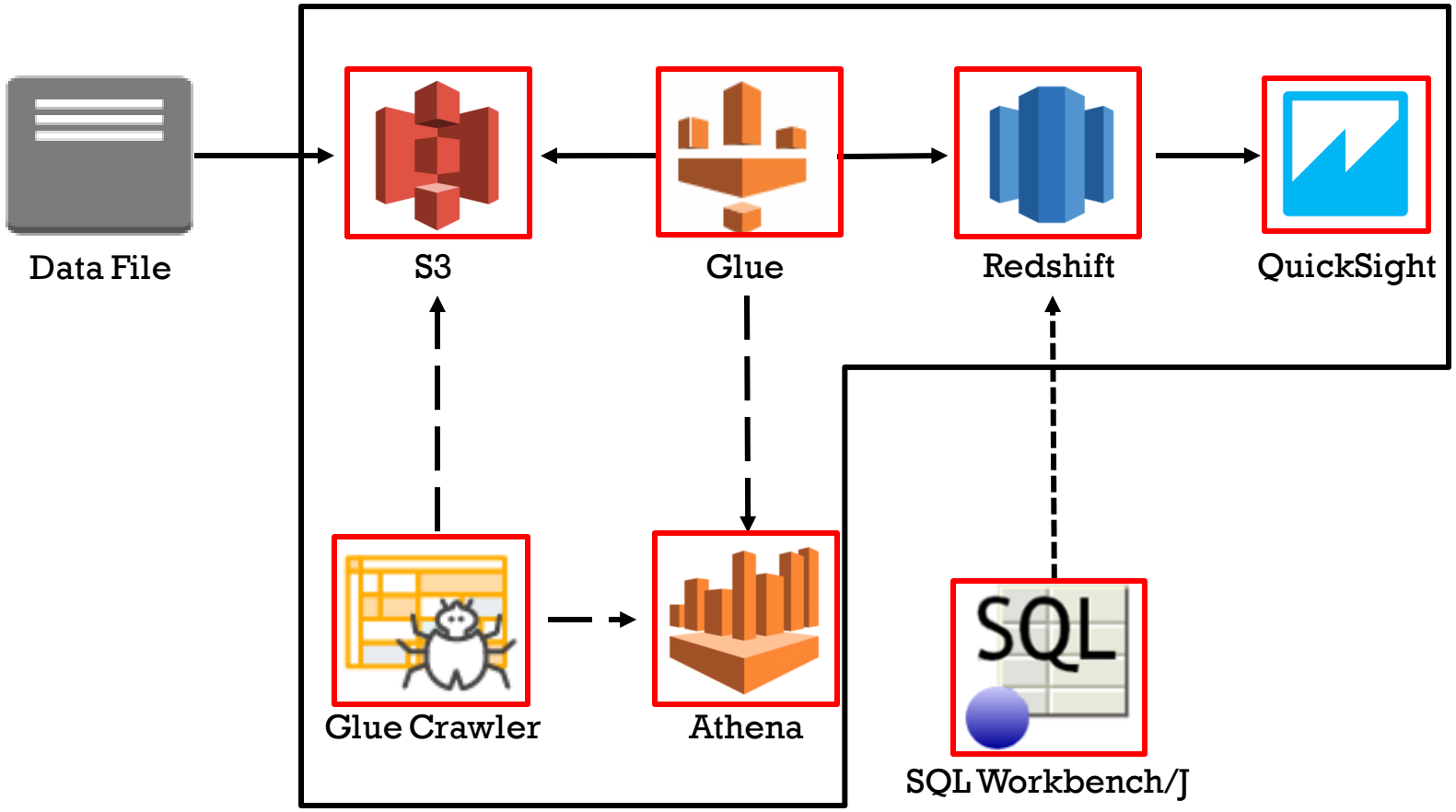
- Serverless
  - companies do not have to invest and maintain on premise servers
- Easily scalable
  - adjust storage needs up and down based on need
- Cost Effective – Glue is cheaper than other ETL Services
  - Only pay when being used, where Matillion and Informatica charge hourly or yearly
  - Matillion: $2.74 per hour (m4.large EC2), Informatica $3.66 per hour (m4.large EC2), Glue $0.44 per DPU-Hour
- Code based (Python or Scala) so you can do anything you can program
- Easy integration with other AWS tools
- Automatic error handling and logging

AWS vs. Hadoop

Hadoop – A popular platform used to store and transform big data

- AWS is more flexible – scale up or down storage based on need
- AWS is less complex – no need to set up and maintain servers
- AWS cheaper
  - Start up cost
  - Maintenance cost
  - Pay as you go
- Hadoop has challenges handling a lot of small files
- AWS – End to End solution for data needs
  - Storage
  - Transform
  - Business Intelligence
- ETL & ELT(AWS) vs. ELT(Hadoop)
- Durability
  - Data stored in multiple locations within region
  - If a location fails data is still available

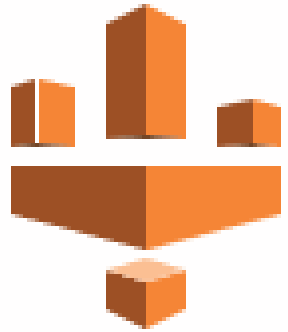
GLUE TUTORIAL OVERVIEW



- Setup Redshift Cluster
- S3 bucket for storing the file
- Athena table to access data in file
- Glue connection
- Glue job
- Connect To Redshift in SQL Workbench
- Create Redshift table
- Run Glue job
- QuickSight

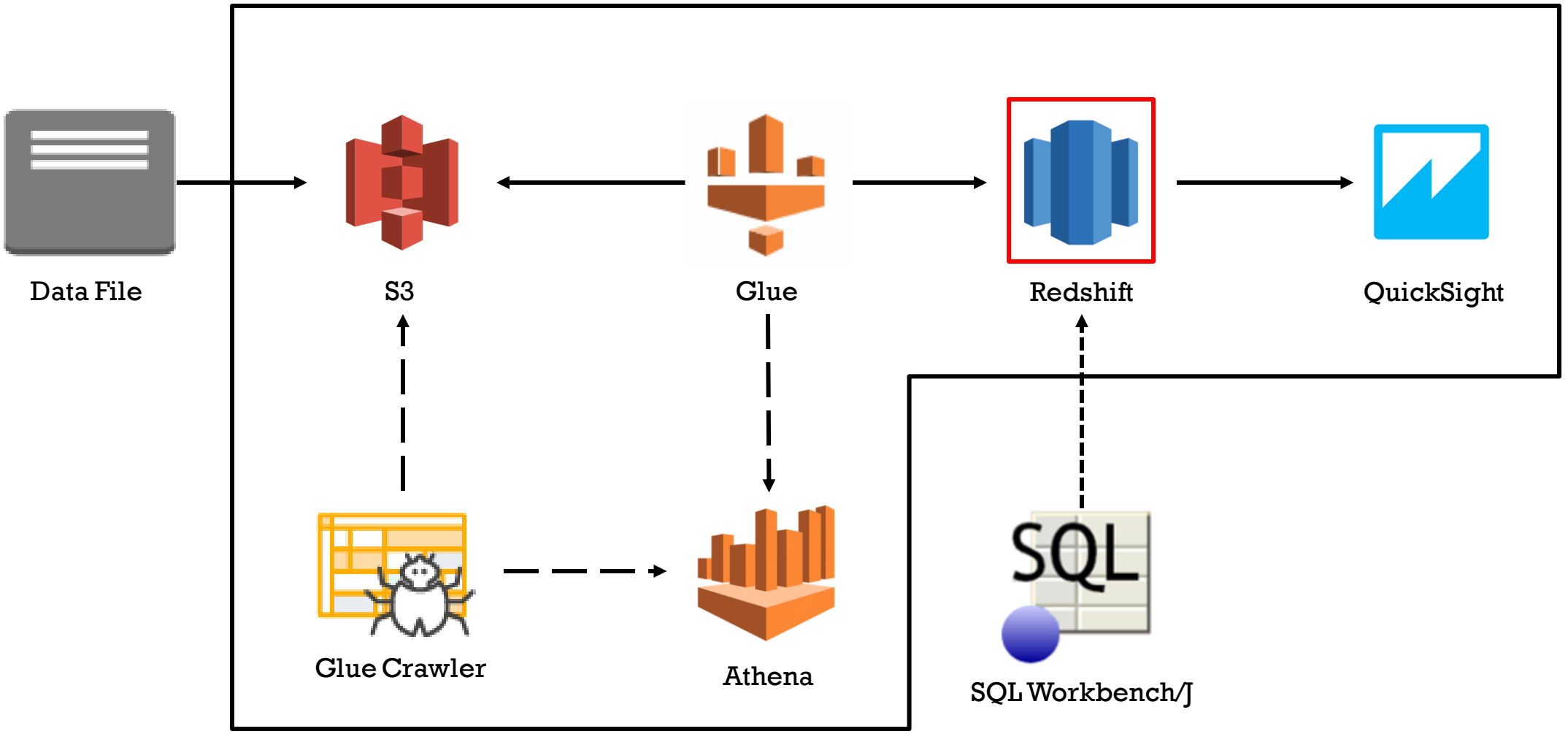
Glue Tutorial Prerequisites

- Prerequisites :
  - Setup AWS Account (Personal or Temporary)
    - https://portal.aws.amazon.com/billing/signup
    - http://escab.juddsolutions.com/registration/create
  - Clone or save git repository https://github.com/jayzandme/aws-glue-tutorial.git
  - download SQL Workbench/j https://www.sql-workbench.eu/
  - download Redshift JDBC driver https://docs.aws.amazon.com/redshift/latest/mgmt/configure-jdbc-connection.html#download-jdbc-driver



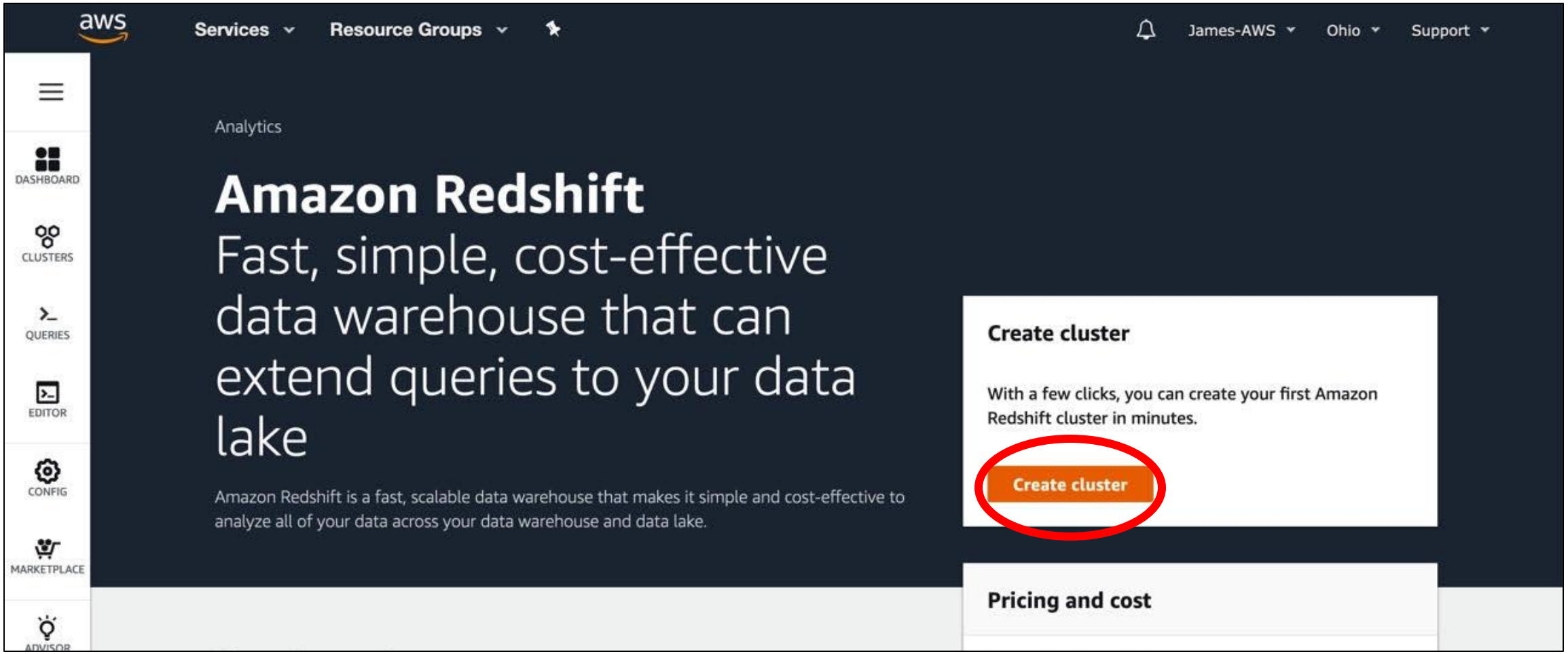
Redshift

Create AWS Data Warehouse



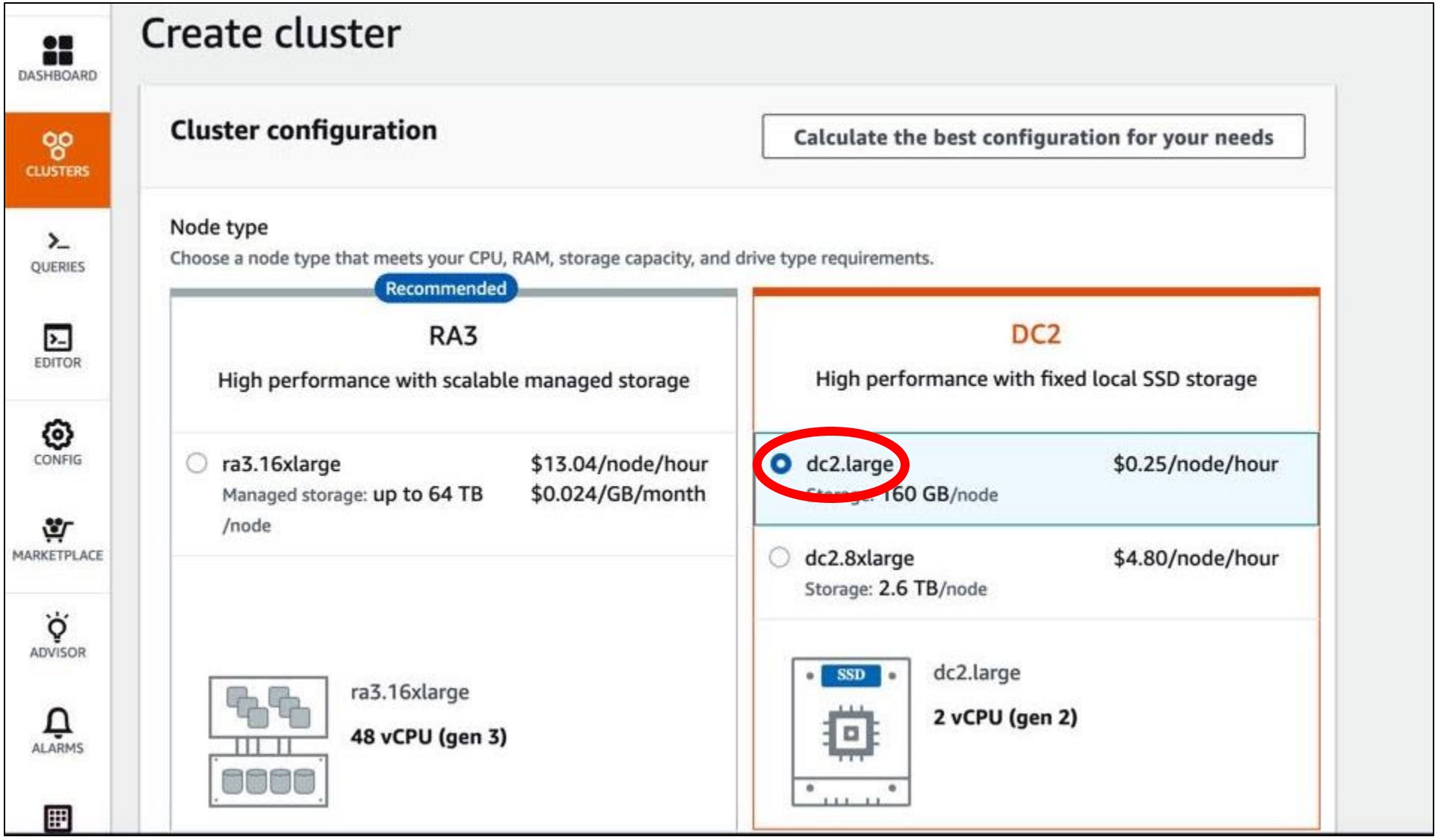
Redshift

Create AWS Data Warehouse



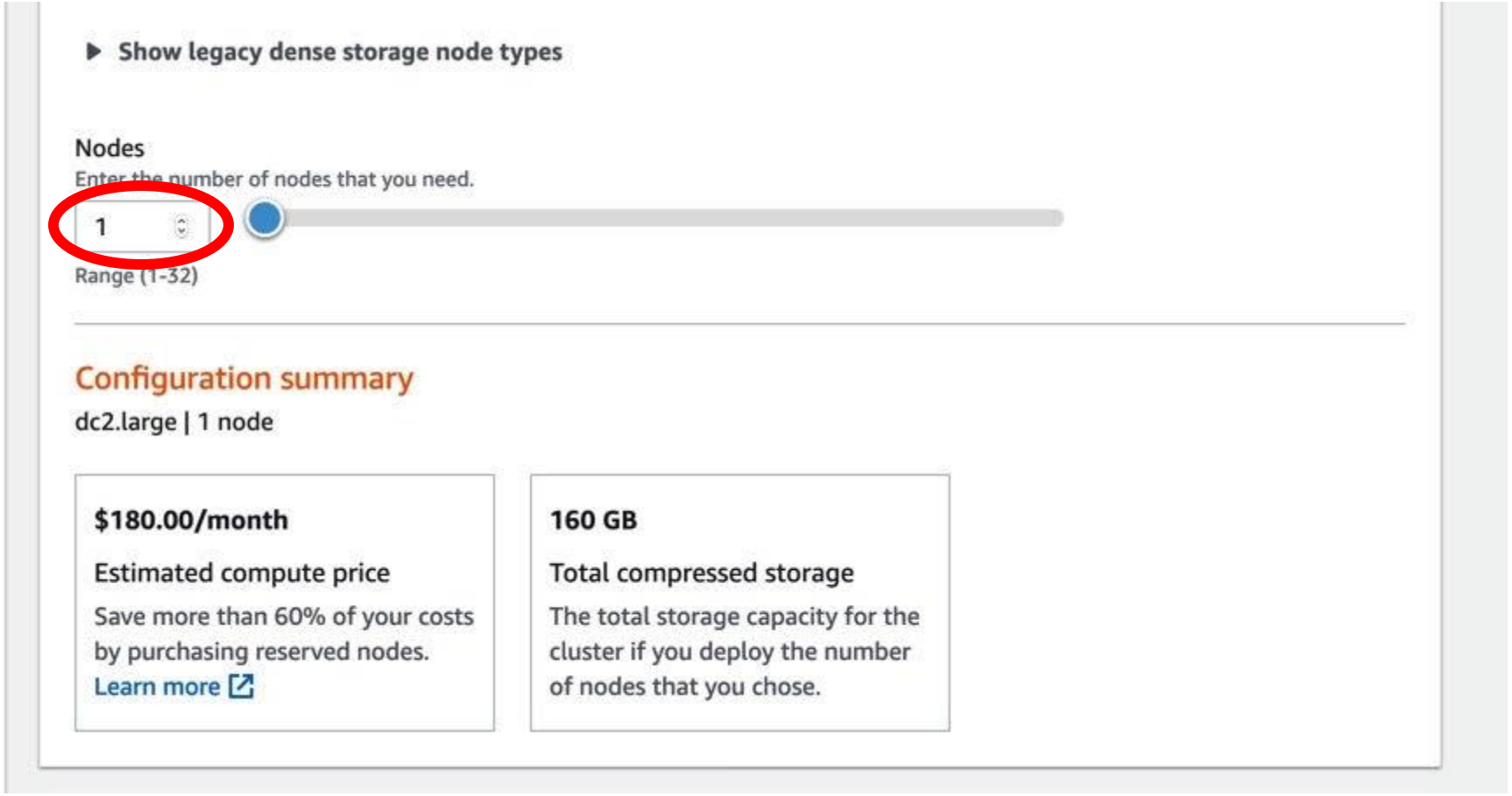
Redshift

Create AWS Data Warehouse



Redshift

Create AWS Data Warehouse



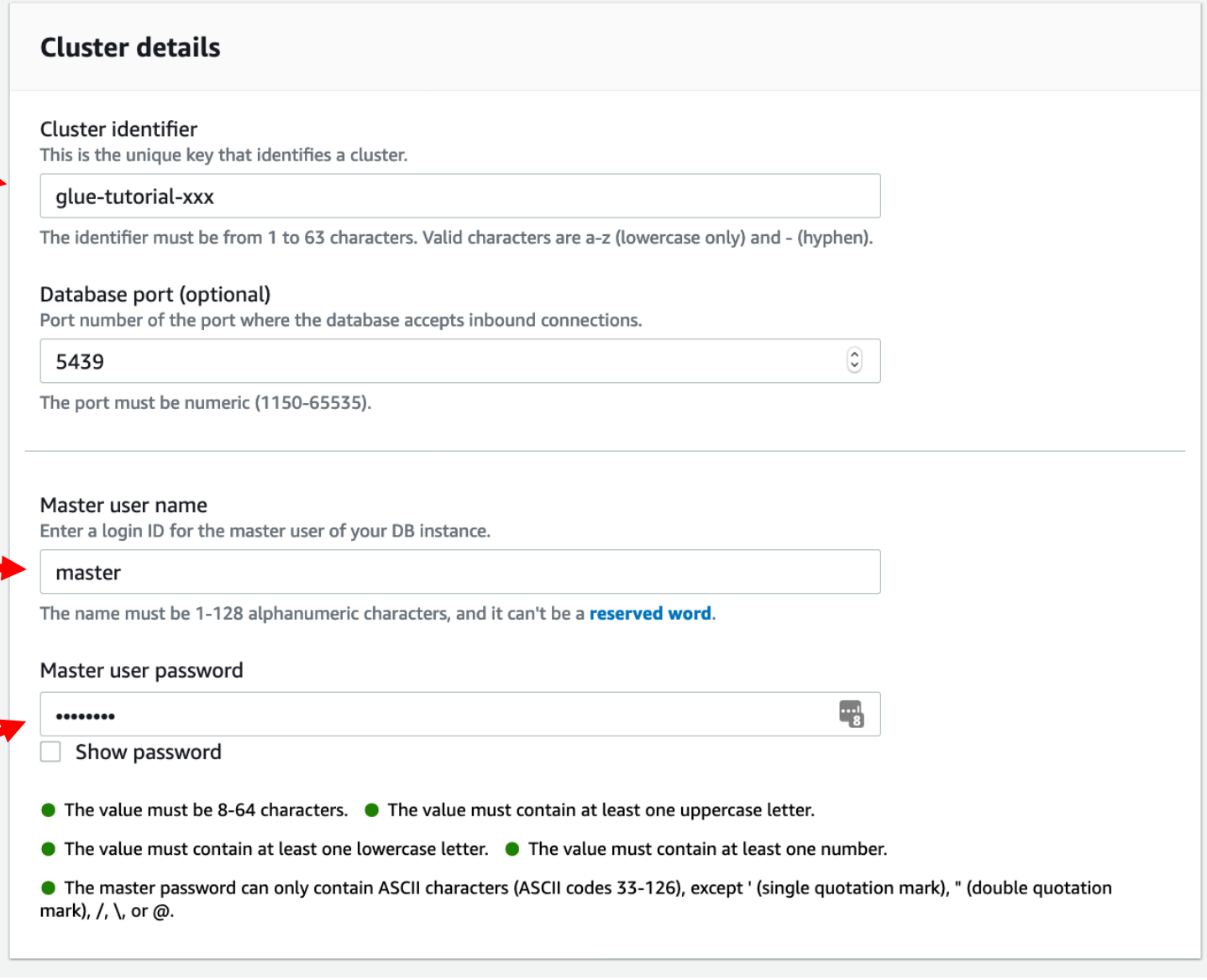
Redshift

Create AWS Data Warehouse

Specify Cluster Name

Create a user

Create a password for the user



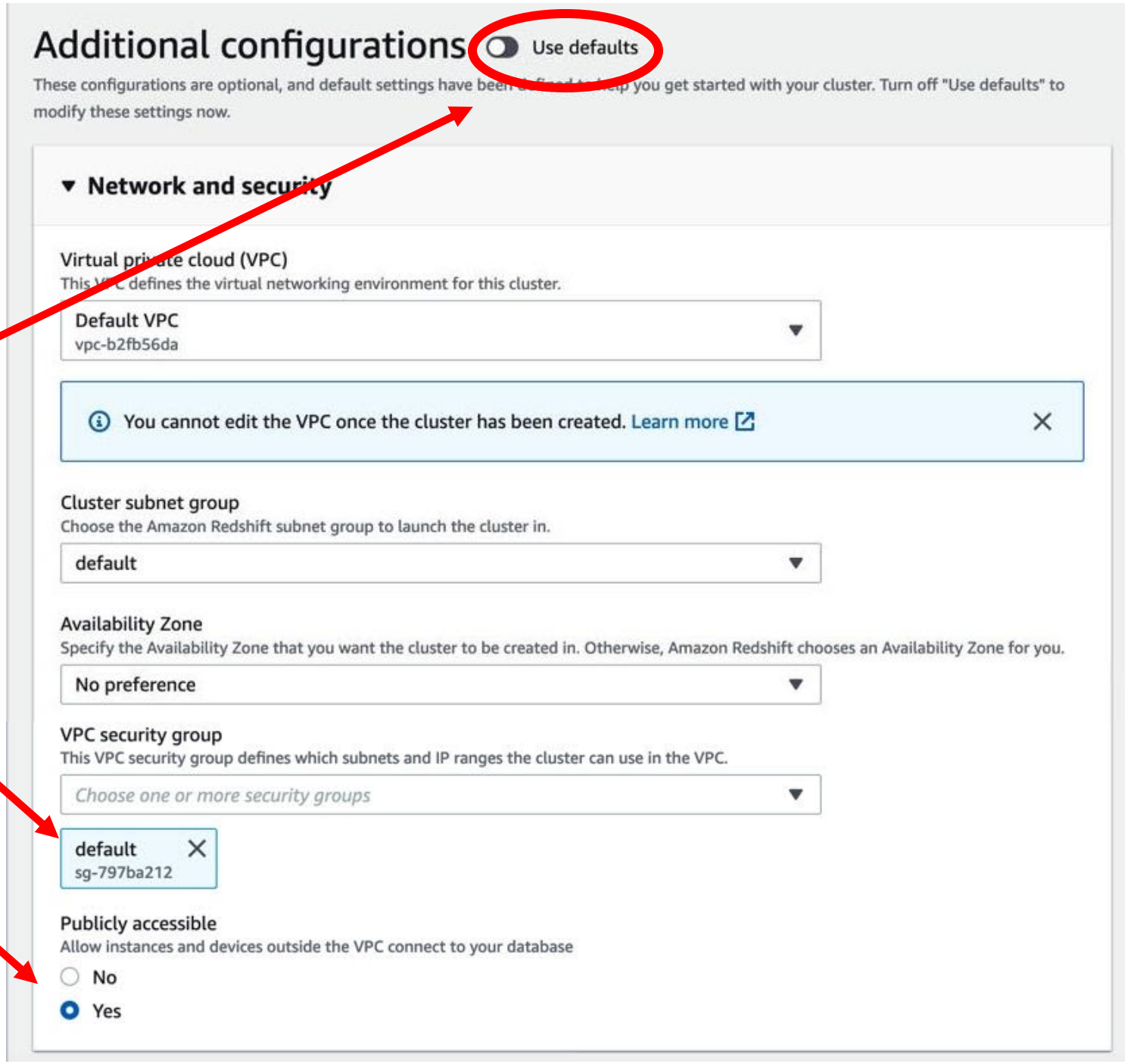
Redshift

Create AWS Data Warehouse

Unclick 'Use defaults'

Take note of the VPC id for later

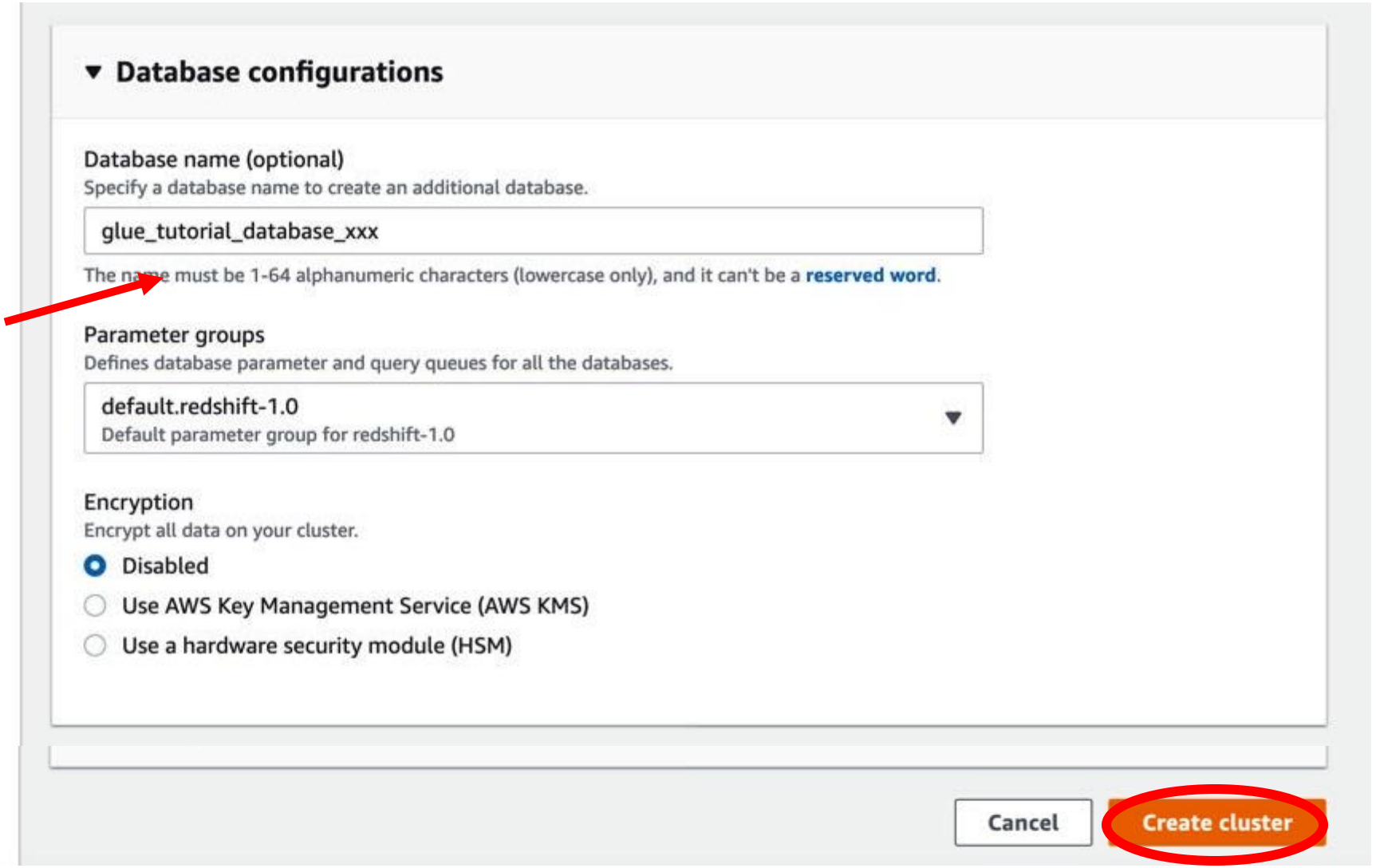
Change 'Publicly Accessible' to Yes



Redshift

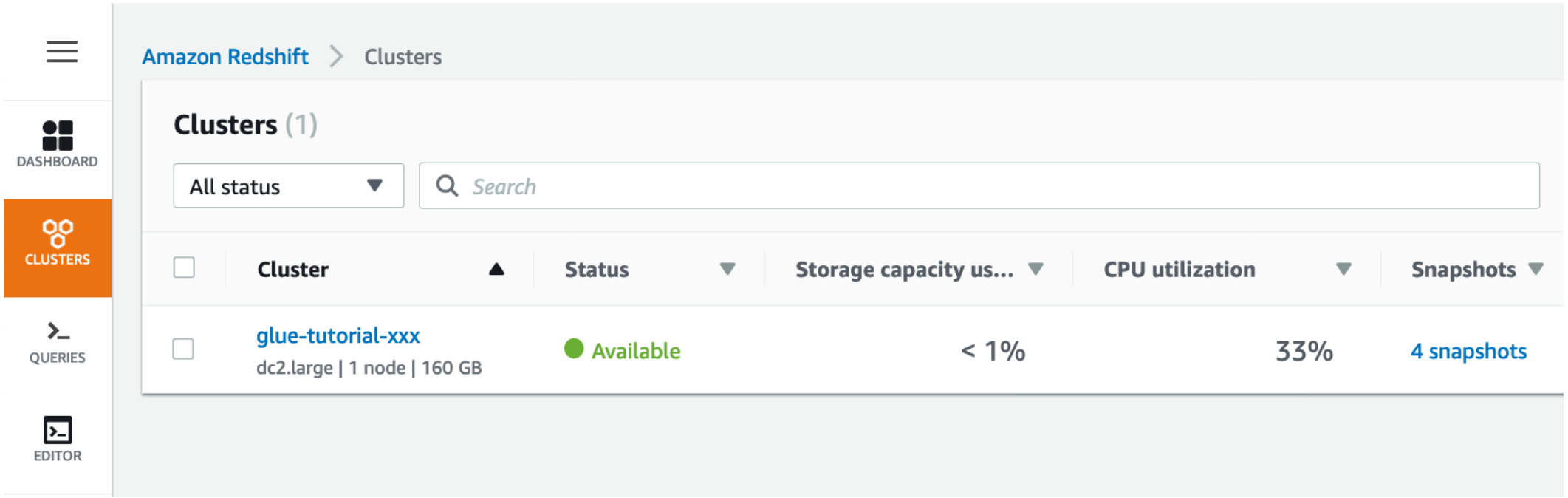
Create AWS Data Warehouse

Give your cluster a Database to start with



Redshift

Create AWS Data Warehouse



Lab 1

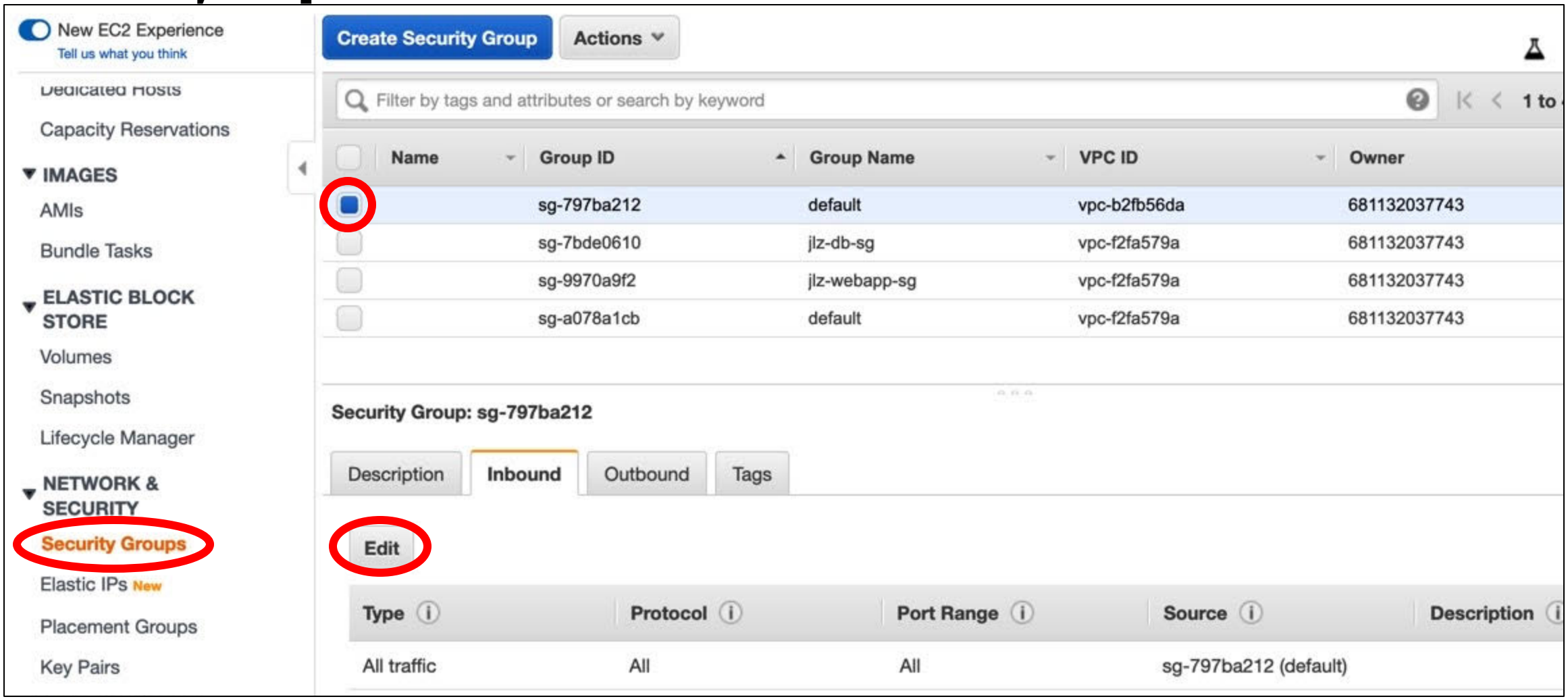
- Launch Redshift cluster

(Use US-EAST-2/Ohio Region)

EC2

Edit Security Groups

In a new tab go to the EC2 service



EC2

Edit Security Groups

Choose Redshift Type

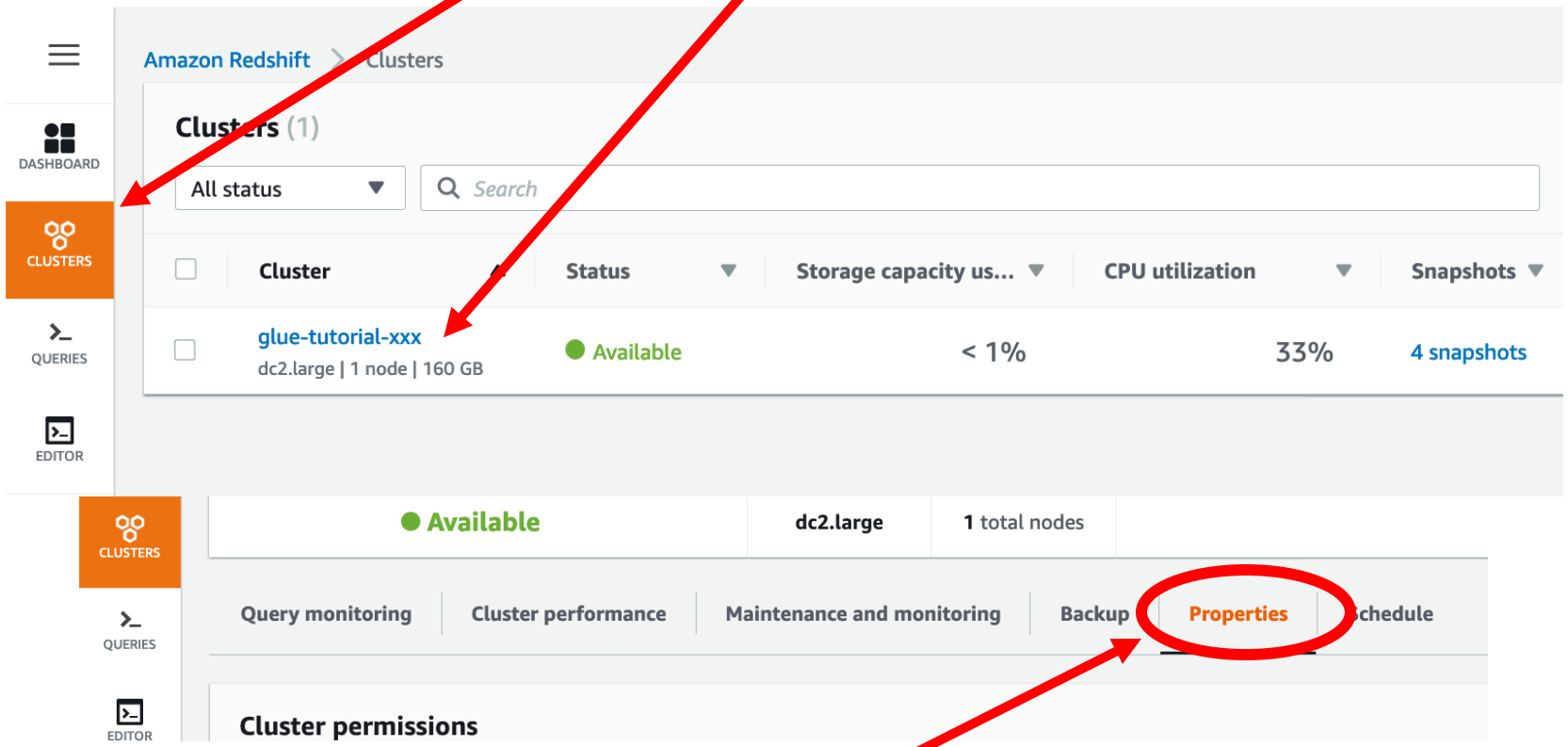
Specifies who has access to the Redshift cluster

Redshift

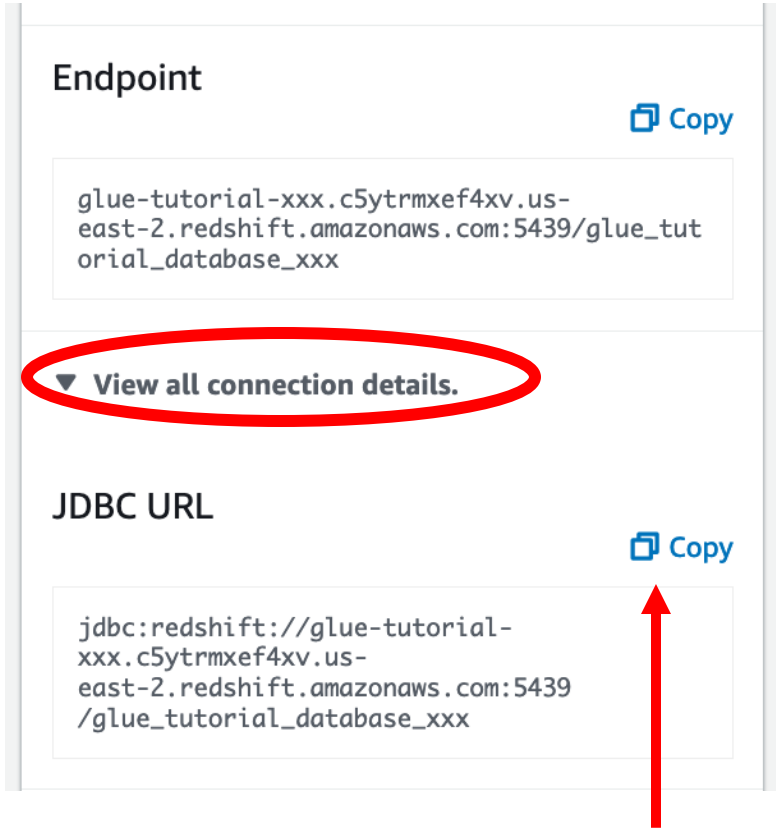
Connection

Go to Redshift and select ‘Clusters’

Select glue-tutorial-xxx



Select Properties



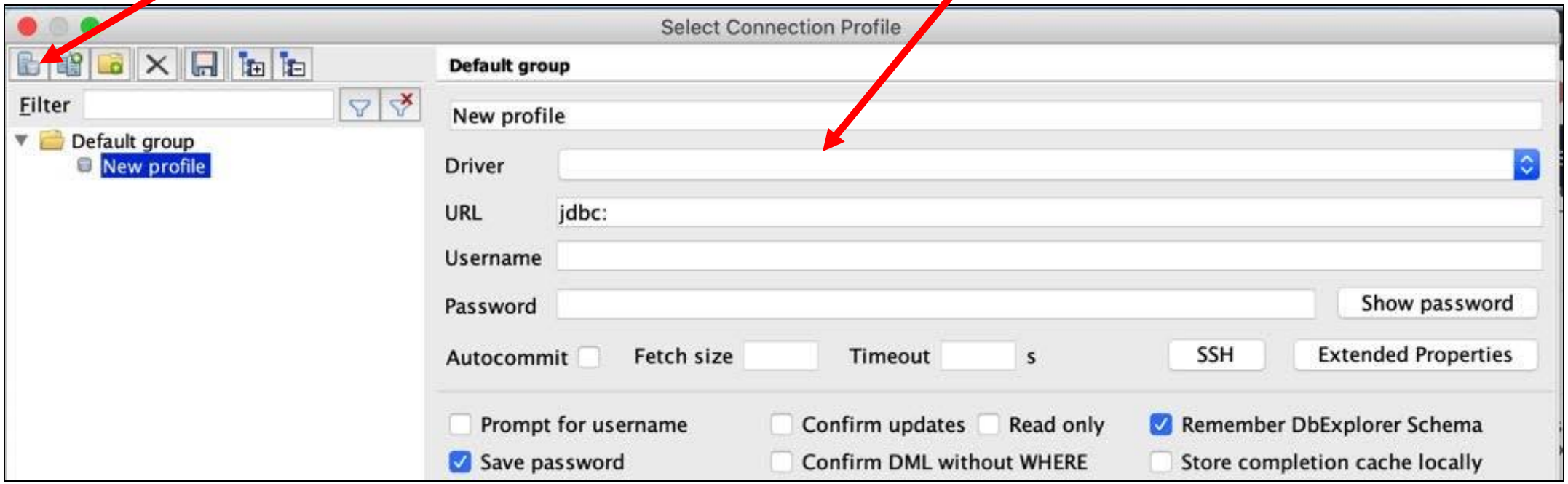
Scroll down to Cluster Database Properties and copy the JDBC URL

Redshift

Connection

Open SQL Workbench and select Create a new connection

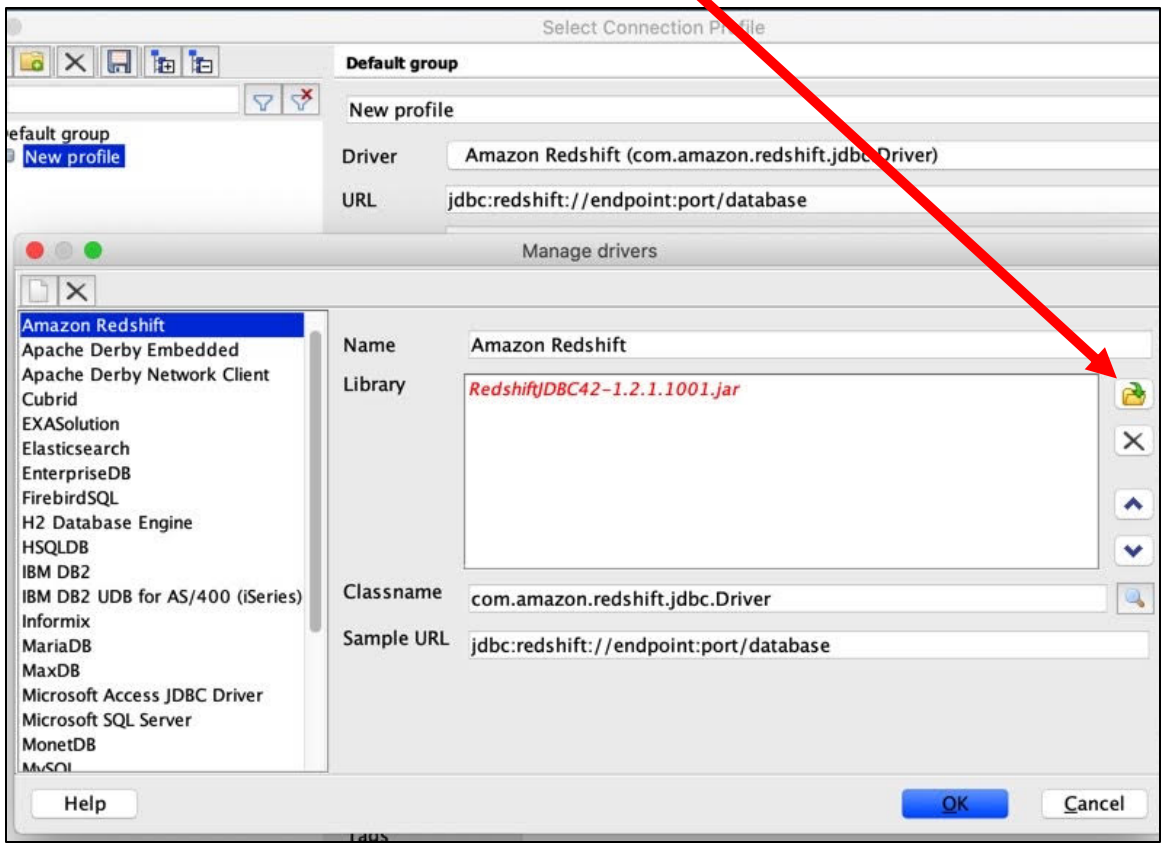
Set the Driver to Amazon Redshift



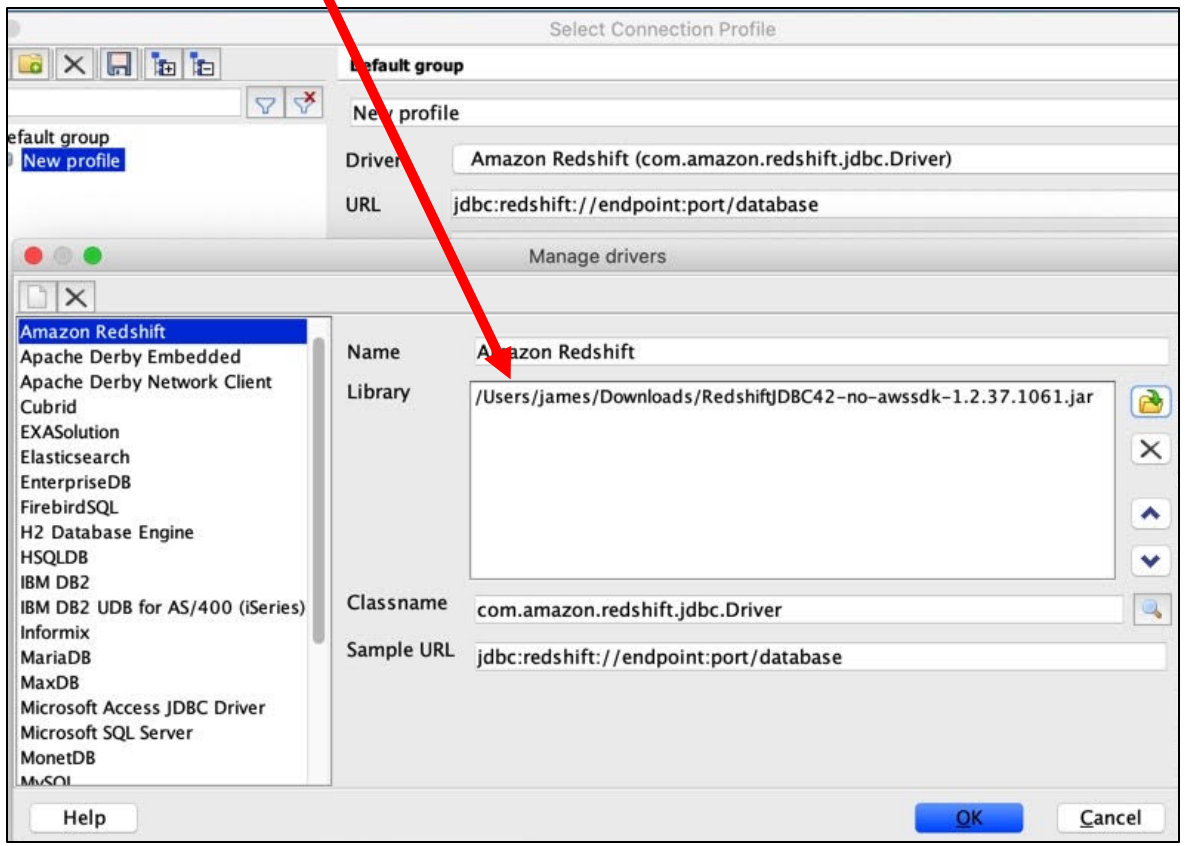
Redshift

Connection

Choose the file



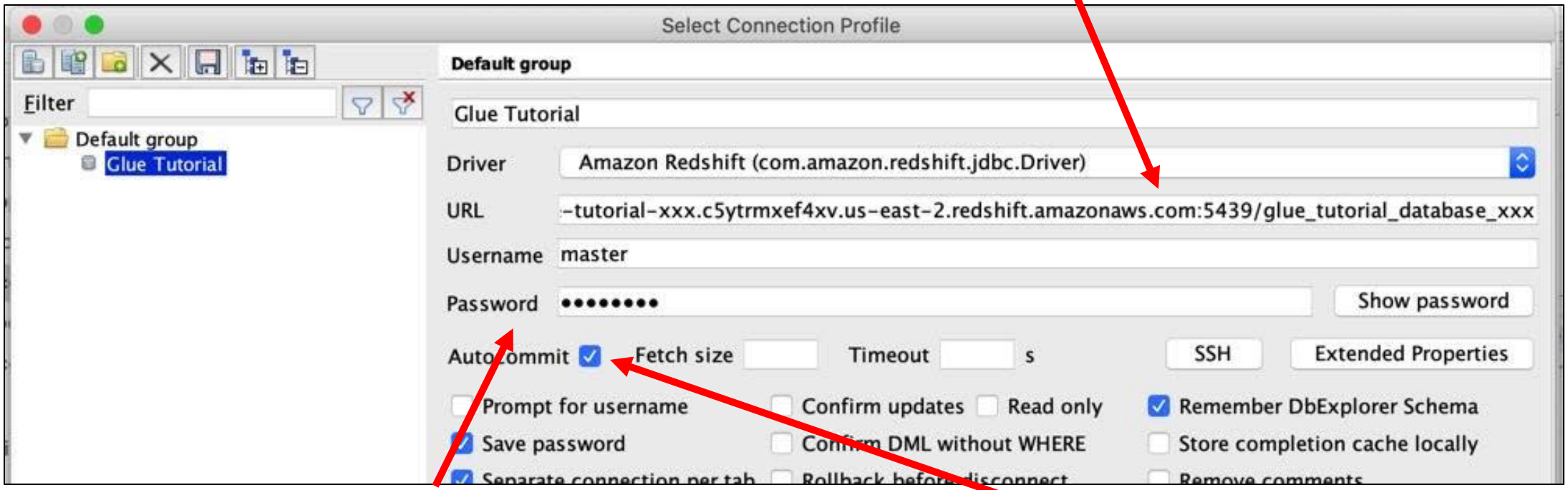
Select the Redshift JDBC Driver you downloaded earlier



Redshift

Connection

Paste the JDBC URL

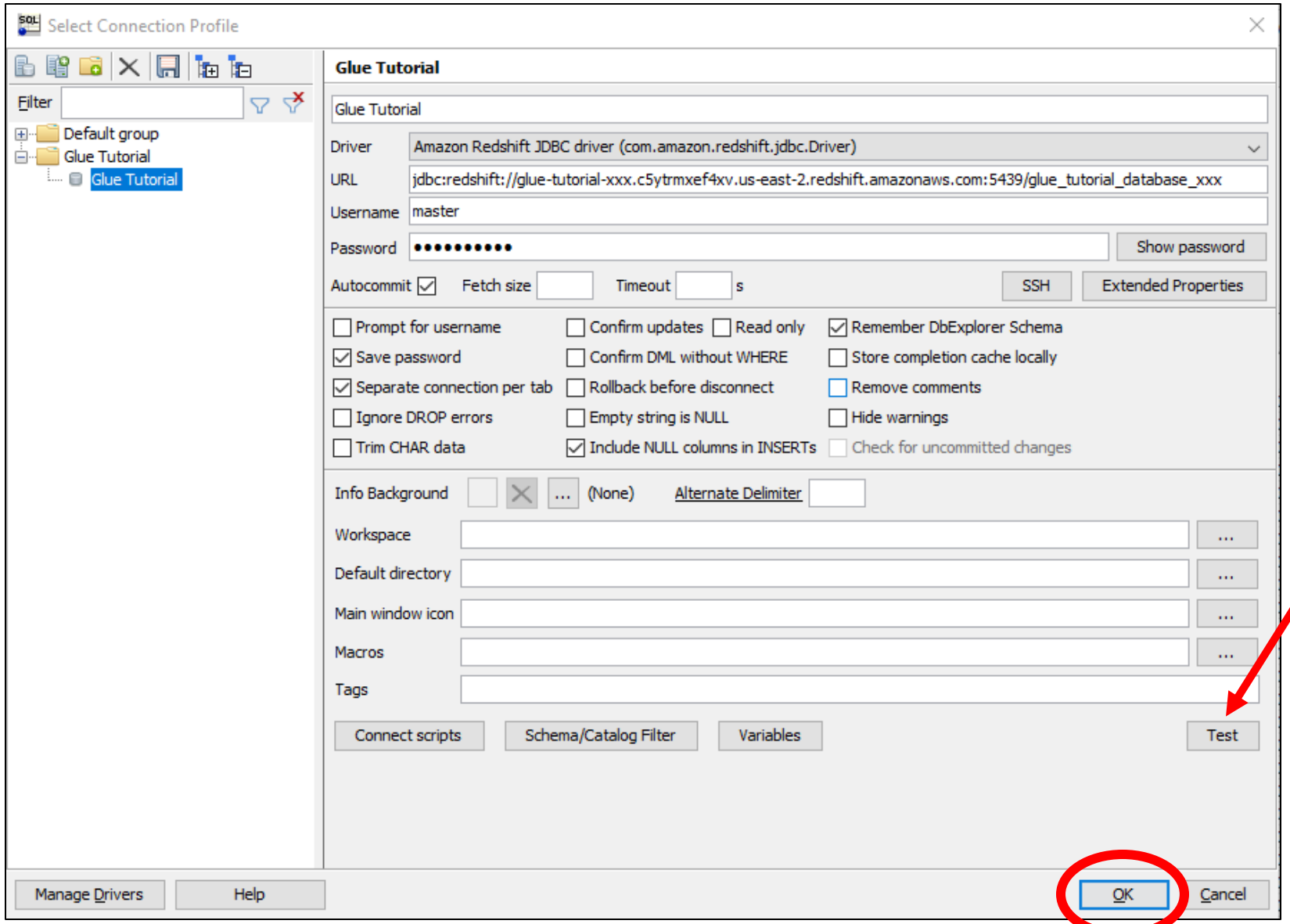


Enter the username and password that you created

Select Autocommit

Redshift

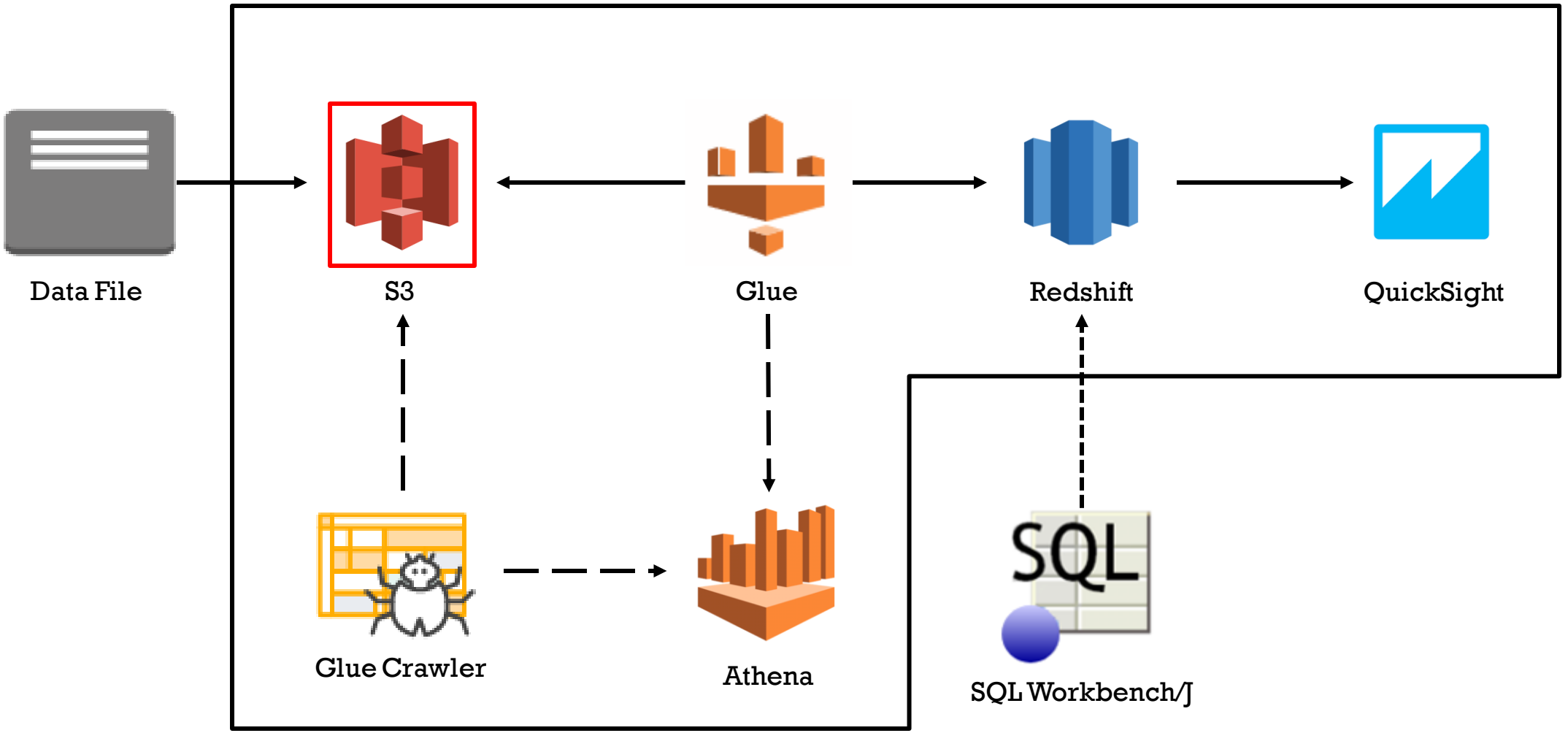
Connection



Test your connection

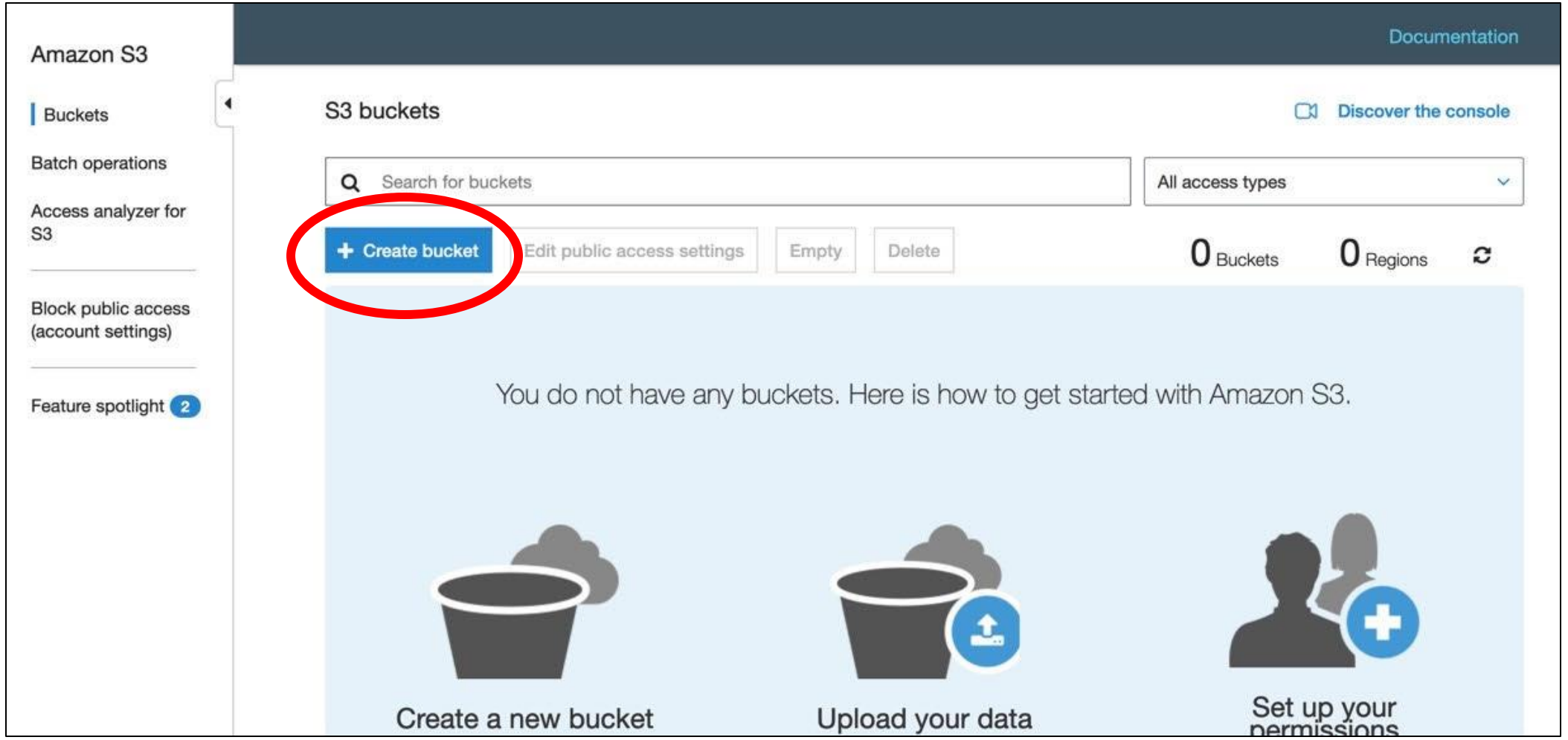
S3

Create S3 bucket with AWS Console



S3

Create S3 bucket with AWS Console

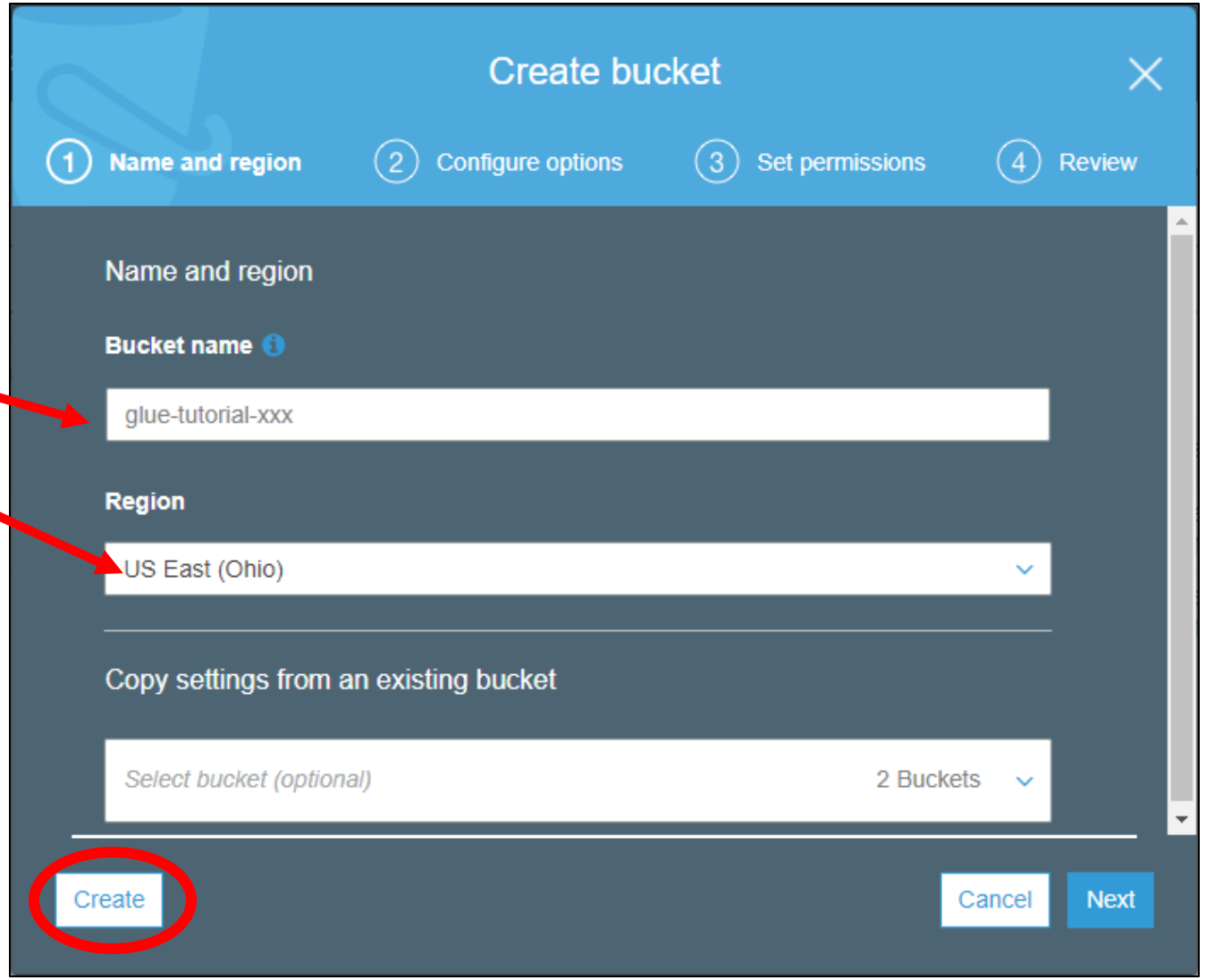


S3

Create S3 bucket with AWS Console

Give your S3 bucket a name
Use glue-tutorial-XXX

Specify the region



Your bucket name needs to be unique because these are accessible across all regions and by potentially everyone

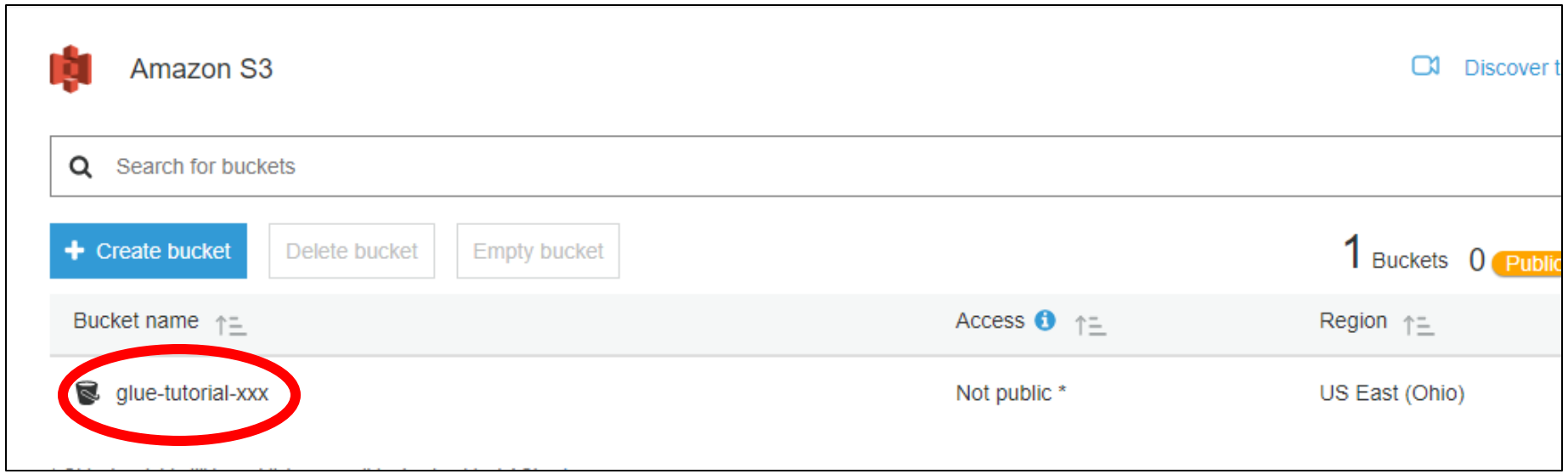
Create S3 bucket with AWS CLI* (Alternative)

```
$ aws s3api create-bucket --bucket  glue-tutorial-xxx --region
us-east-2
```

* Must install and set up AWS CLI in order to use this

S3

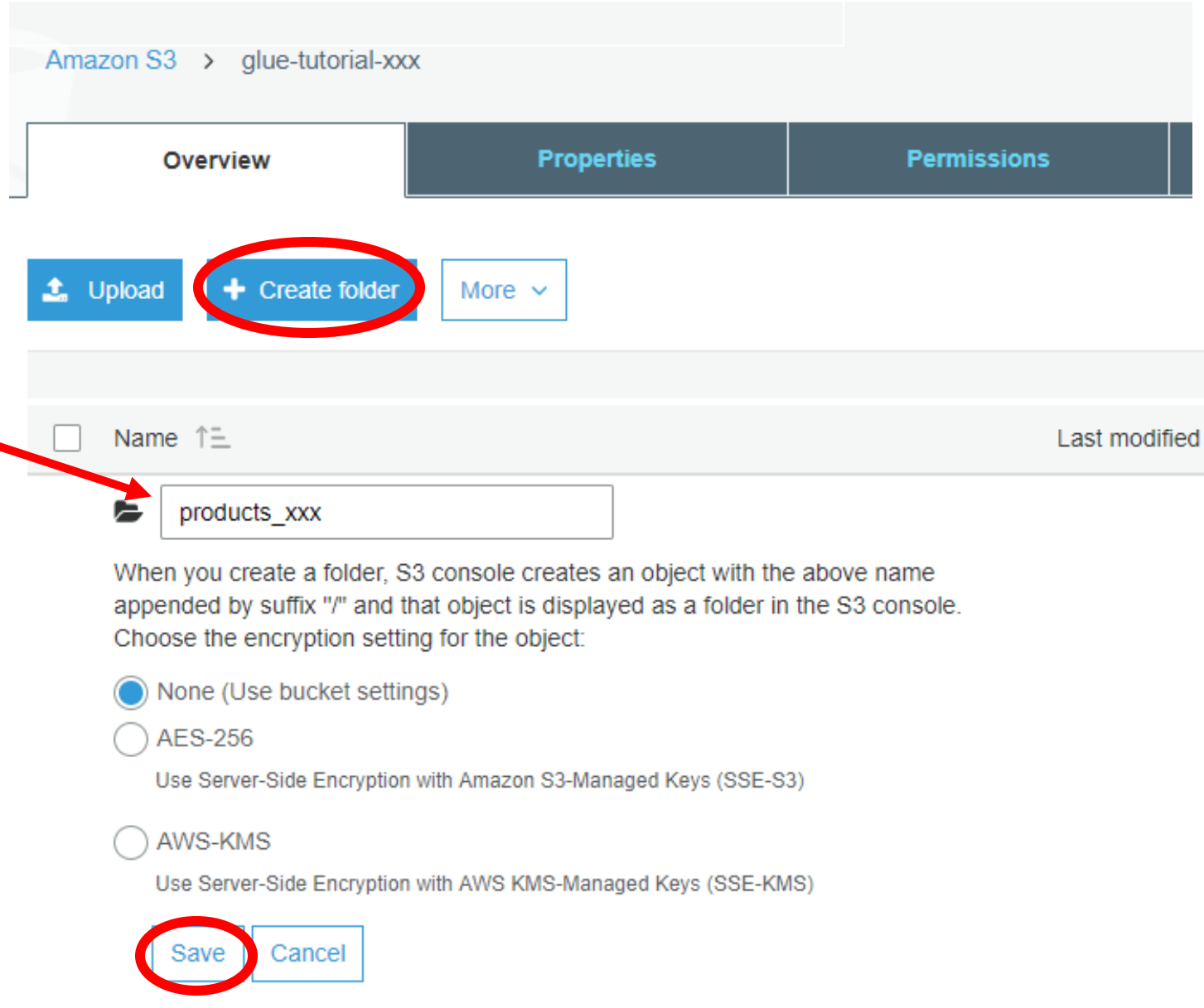
Create S3 bucket with AWS Console



S3

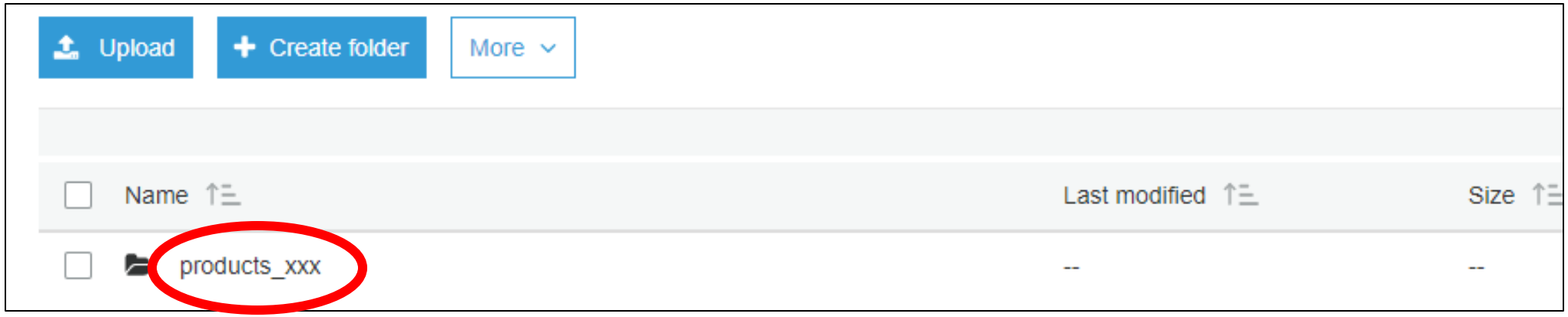
Create S3 bucket folder

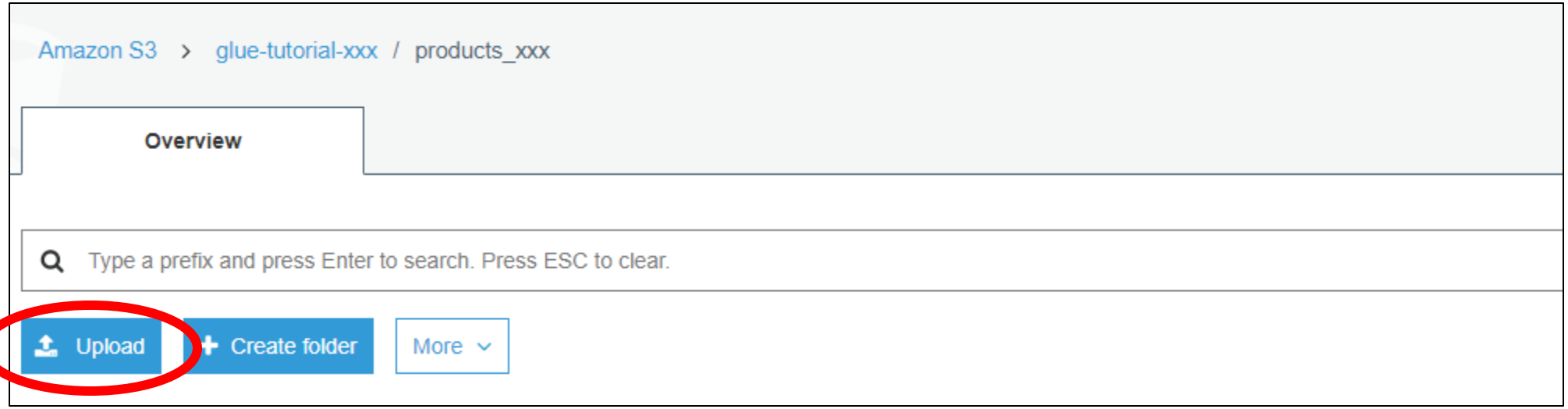
Create a folder called products_XXX



S3

Create S3 bucket with AWS Console

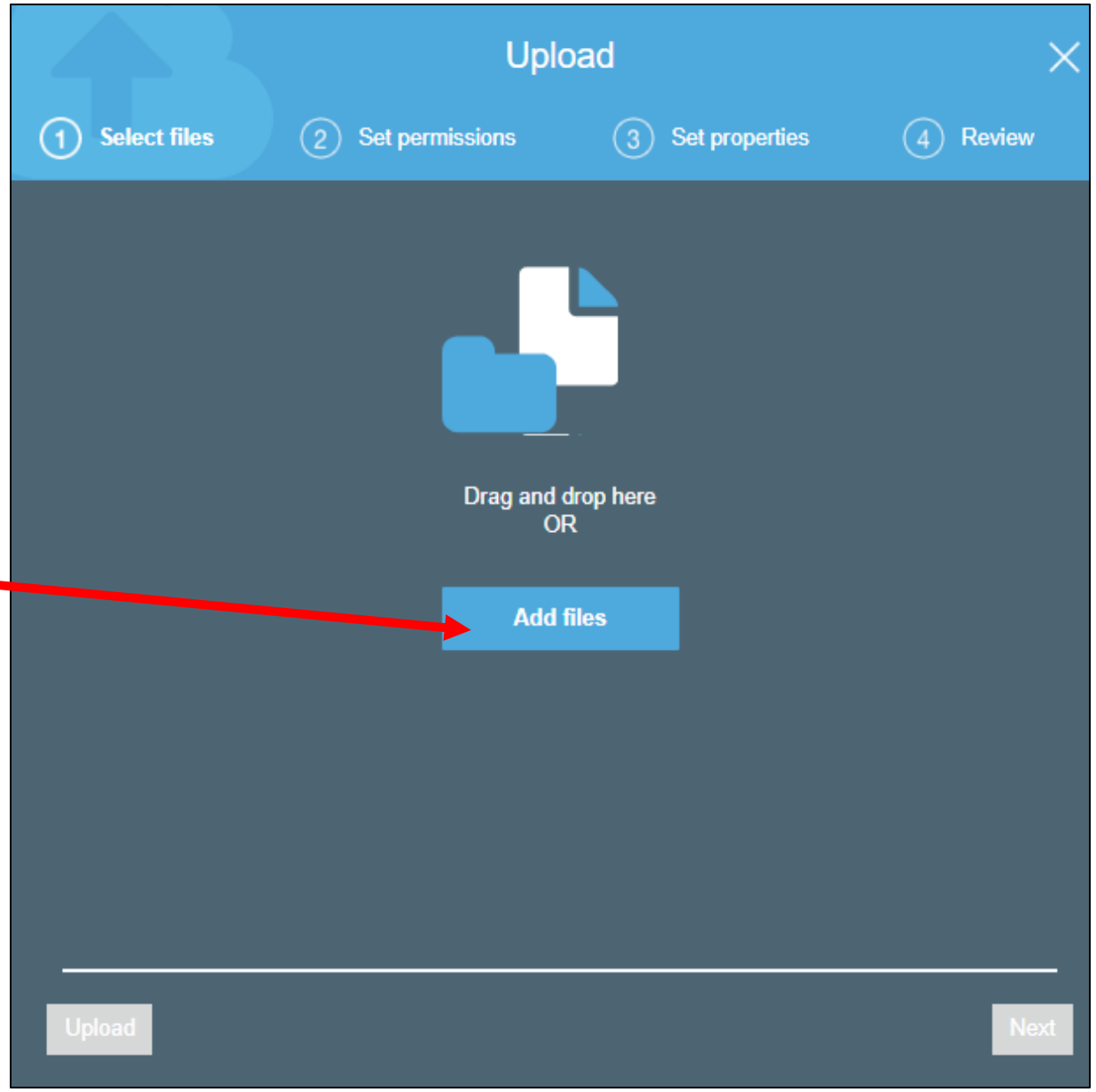




S3

Add file to S3 bucket with AWS Console

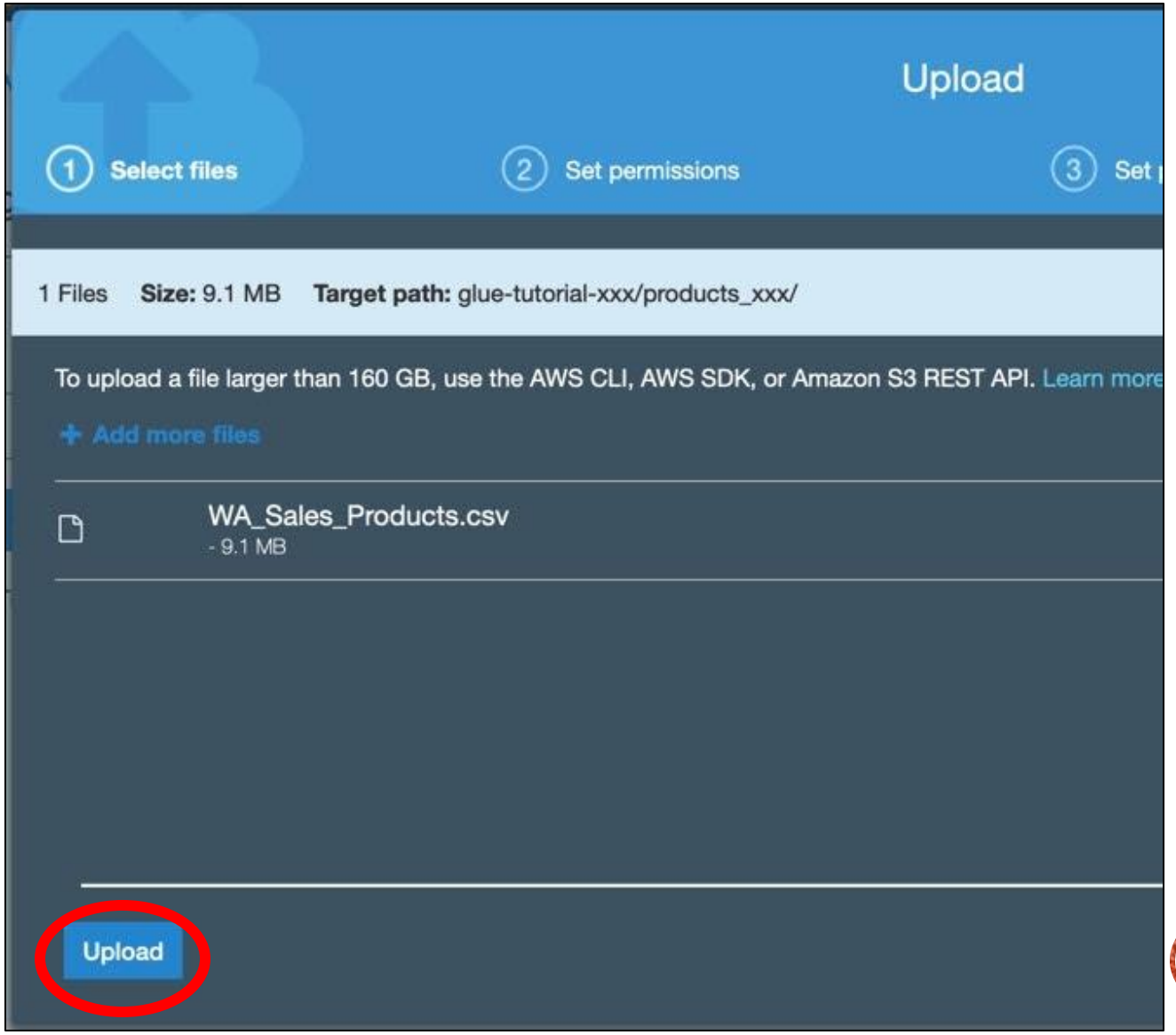
Add file from repository called "WA_Sales_Products.csv"



S3

Add file to S3 bucket with AWS Console

Add file from repository called "WA_Sales_Products.csv"



S3

Add file to S3 bucket with AWS CLI* (Alternative)

```
$ aws s3 cp <your-file-path>/aws-glue-
tutorial/WA_Sales_Products.csv s3://glue-tutorial-
XXX/products_XXX/
```

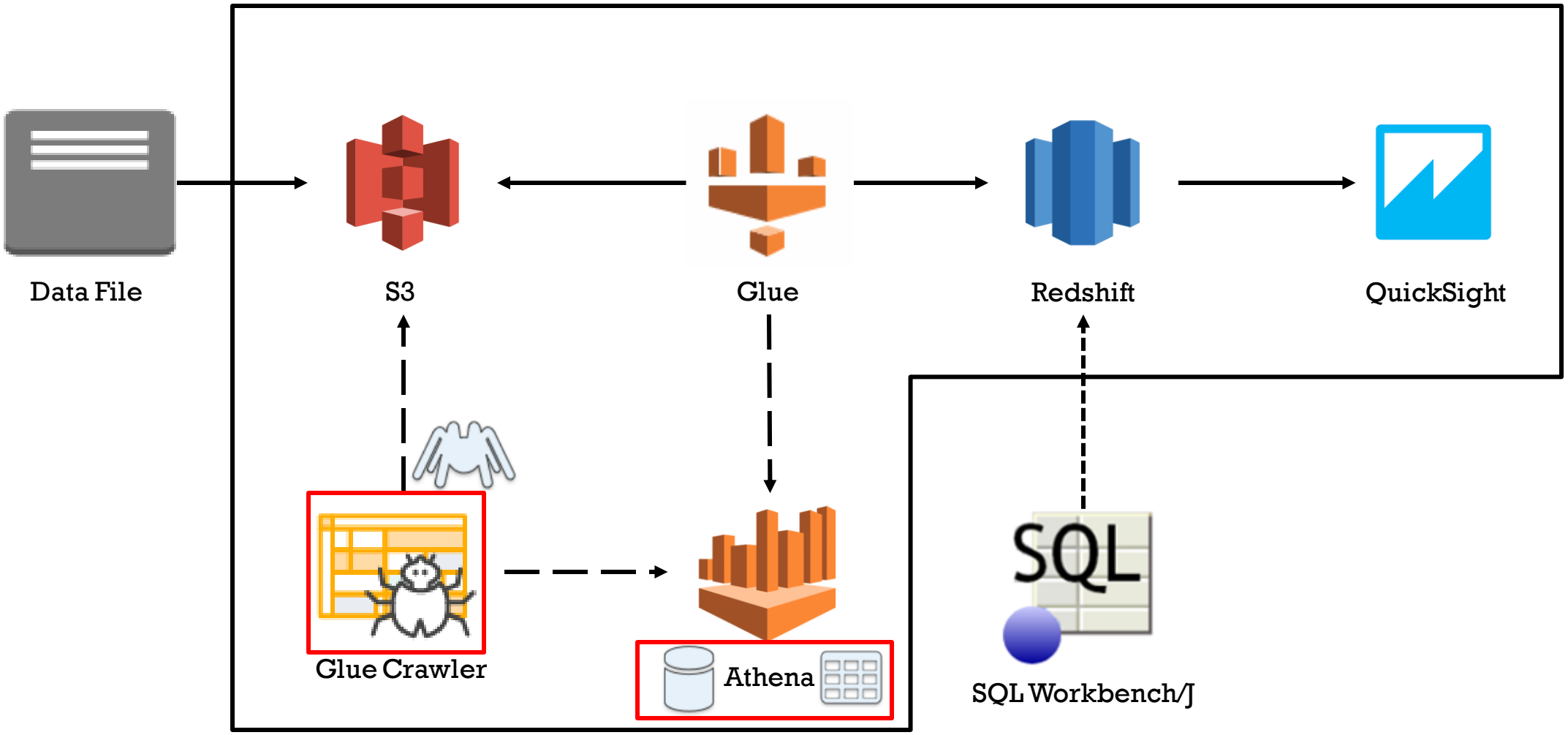
* Must install and set up AWS CLI in order to use this

Lab 2

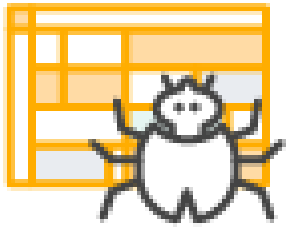
- Test Redshift Connection
- Create S3 bucket
- Add file to S3 bucket

(Use US-EAST-2/Ohio Region)

Glue Crawler



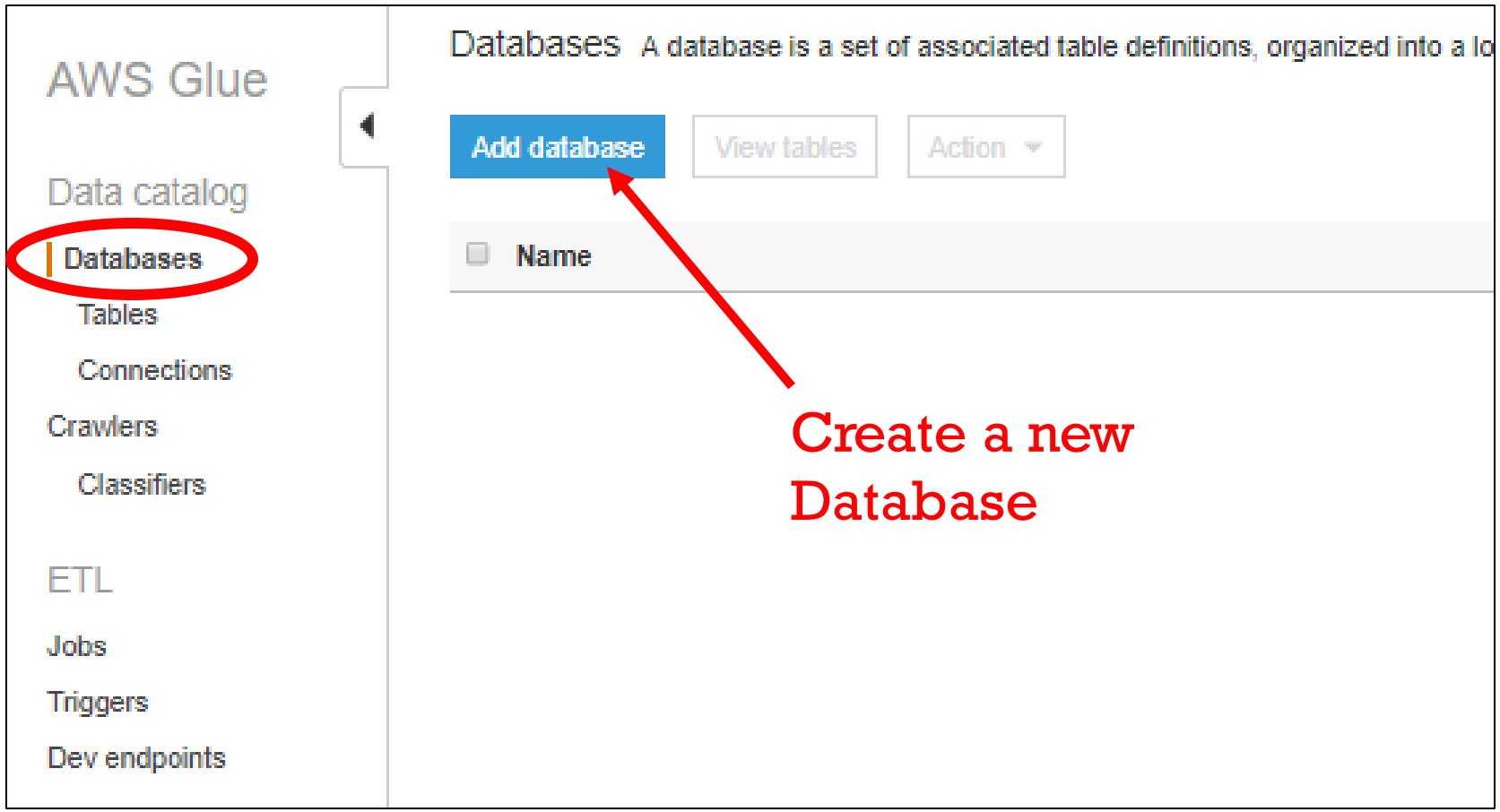
Glue Crawler



- Scans file data to create metadata
- Determines column names and data types
- Creates a Glue Table

Glue

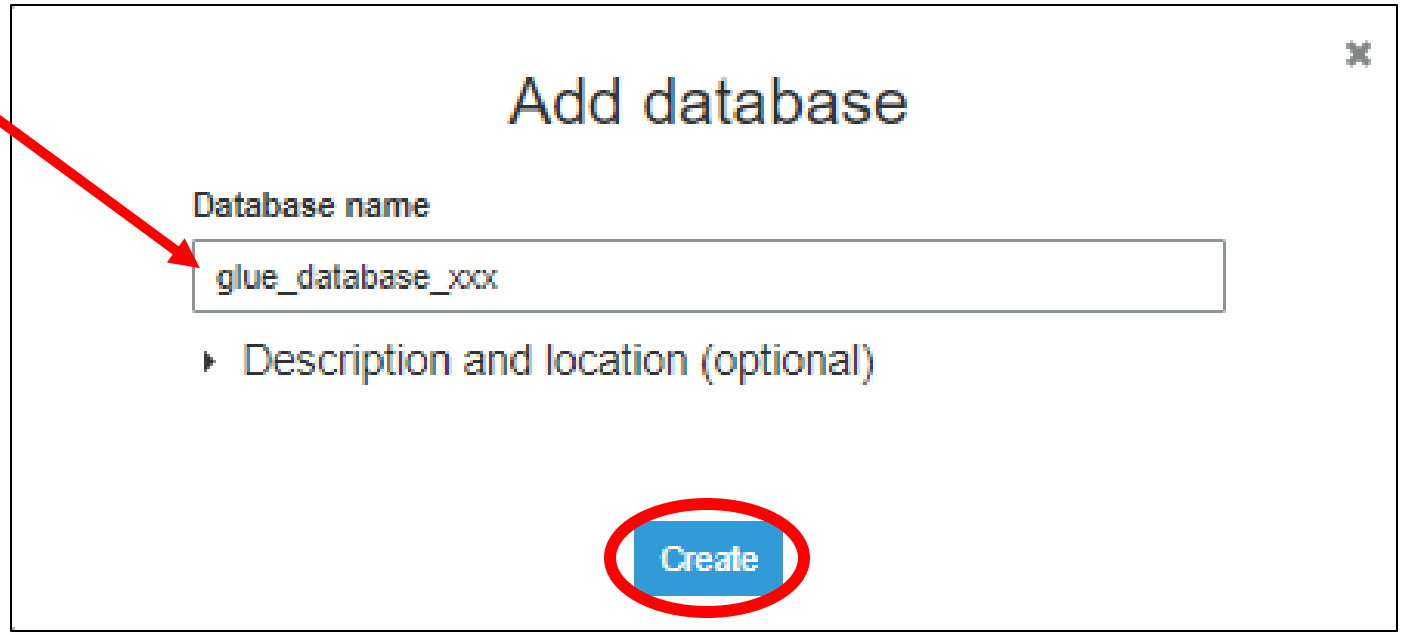
Create Glue Database



Glue

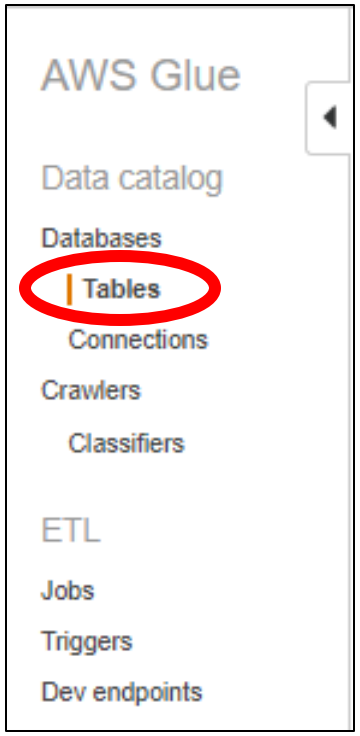
Create Glue Database

Give your database a name
"glue_database_XXX"

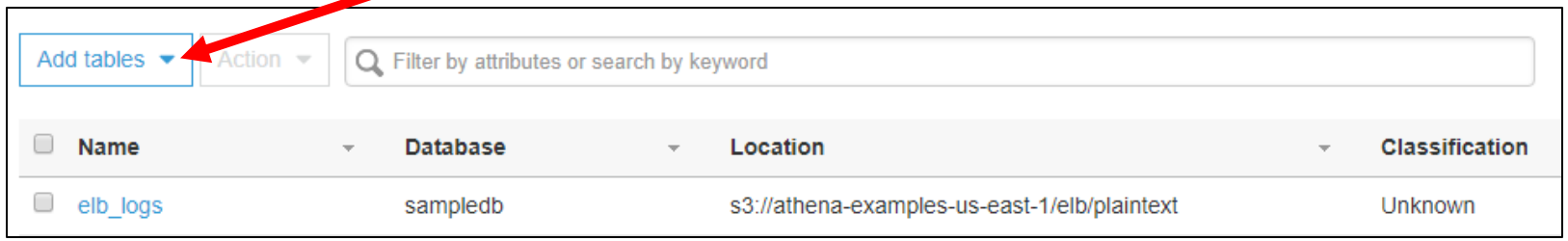


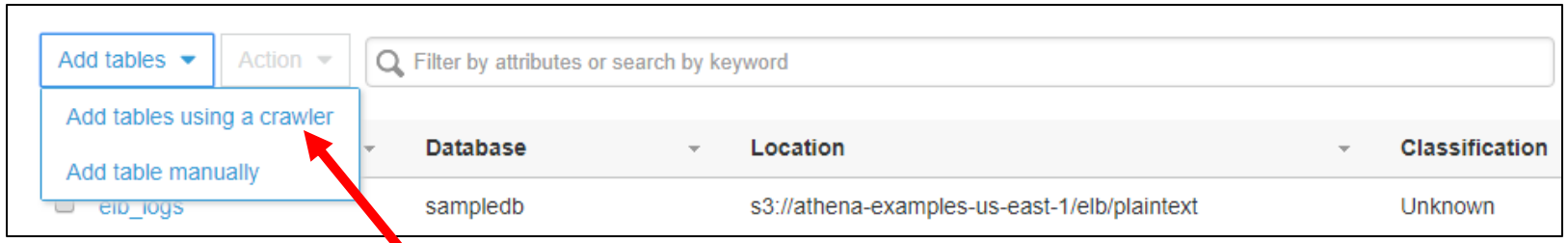
Glue Crawler

Create Table with Glue Crawler



Click on add tables to create a table



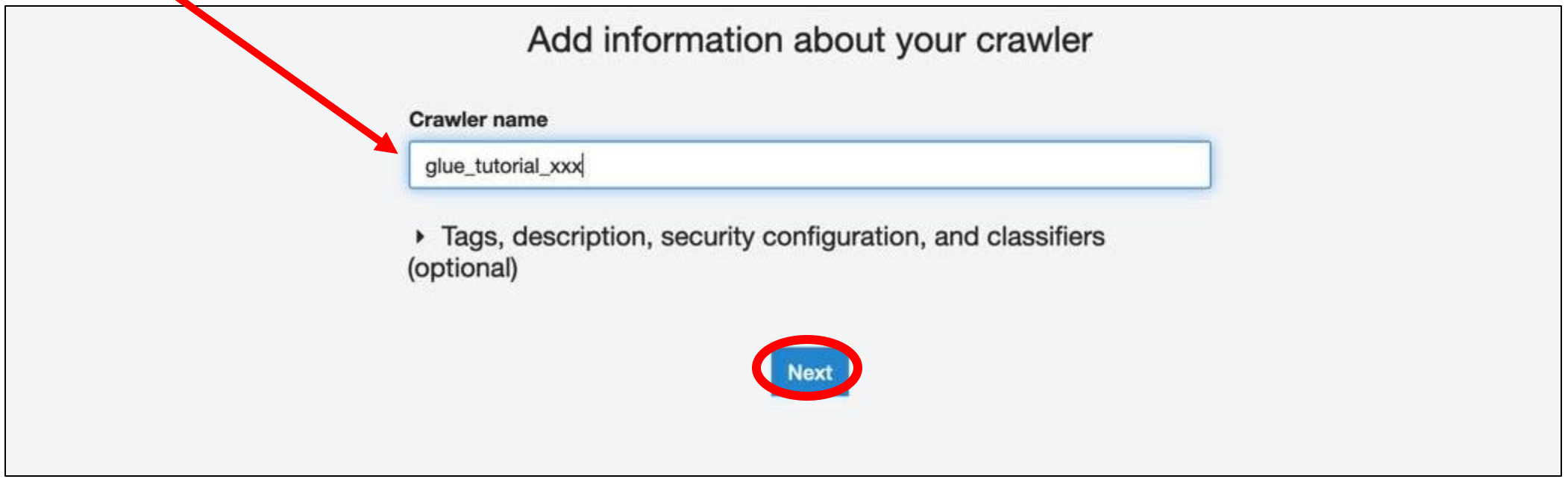


Create a table using a crawler

Glue Crawler

Create Table with Glue Crawler

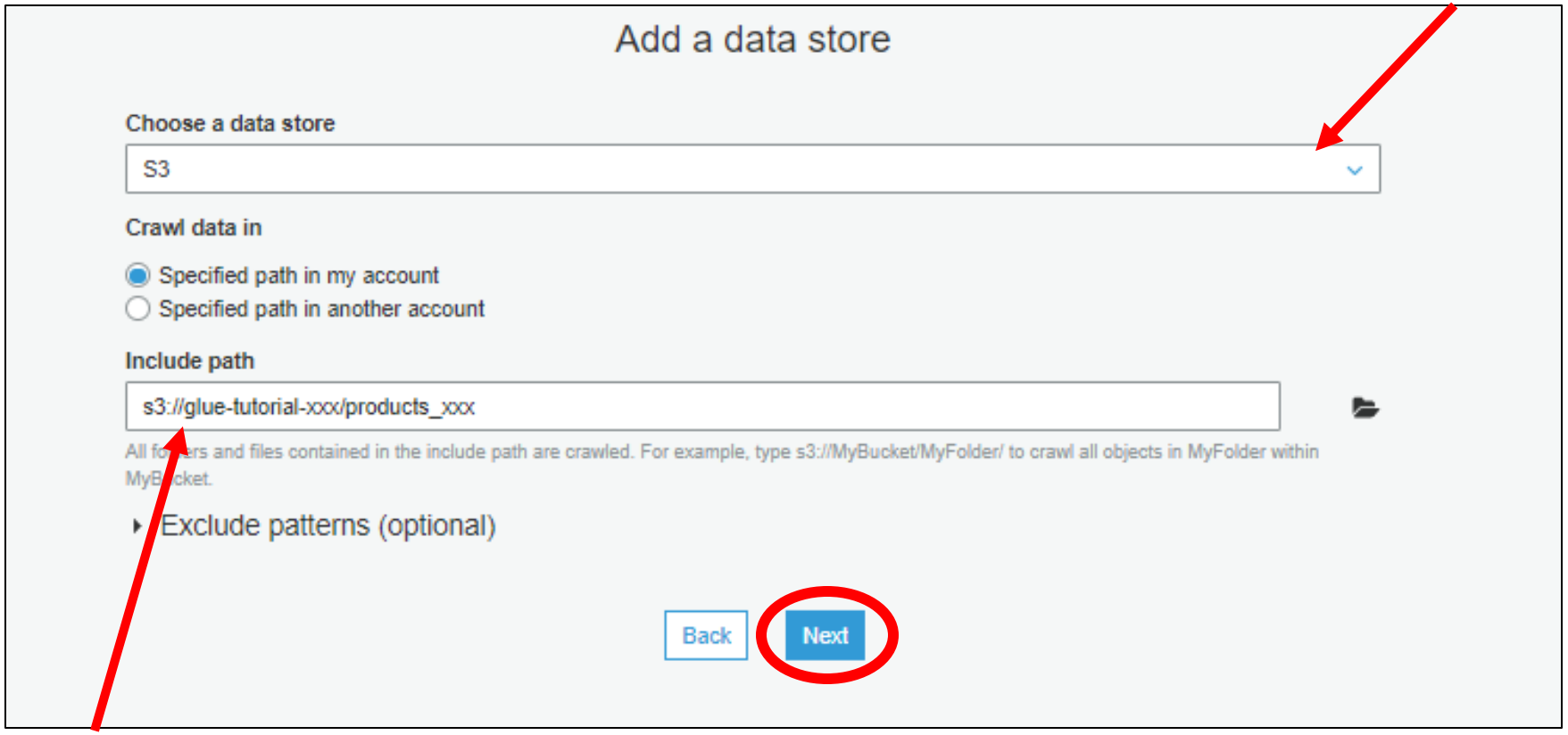
Give your crawler a name,
glue-tutorial-XXX



Glue Crawler

Create Table with Glue Crawler

Choose where the table is going to look for data

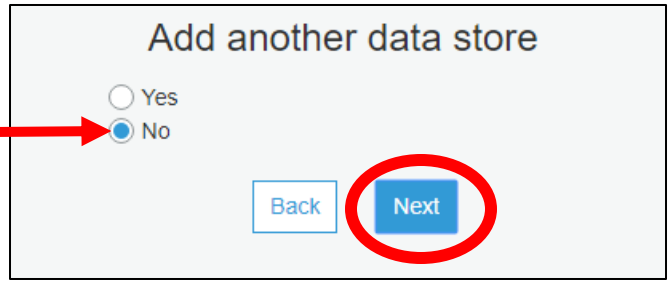


Specify the path for the table to search for in S3

Glue Crawler

Create Table with Glue Crawler

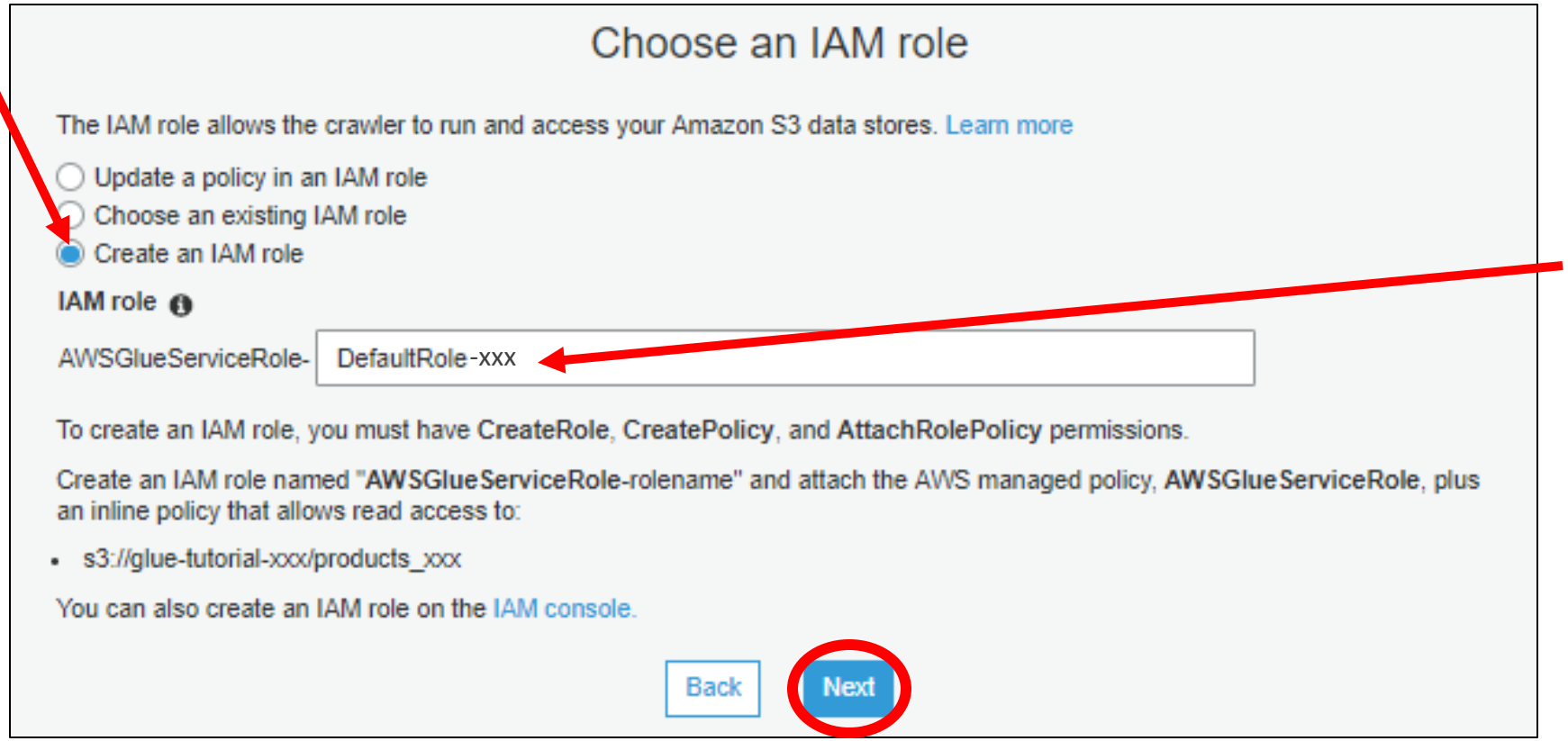
We do not want to add another source of data



Glue Crawler

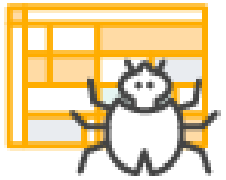
Create Table with Glue Crawler

Need to create role
to access S3 bucket



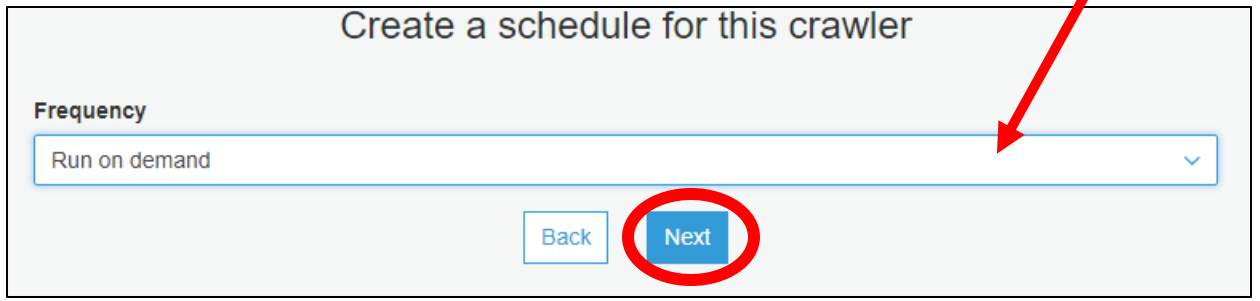
Give your
role a name

Glue Crawler



Create Table with Glue Crawler

Your crawler can run on either a timed schedule or on demand



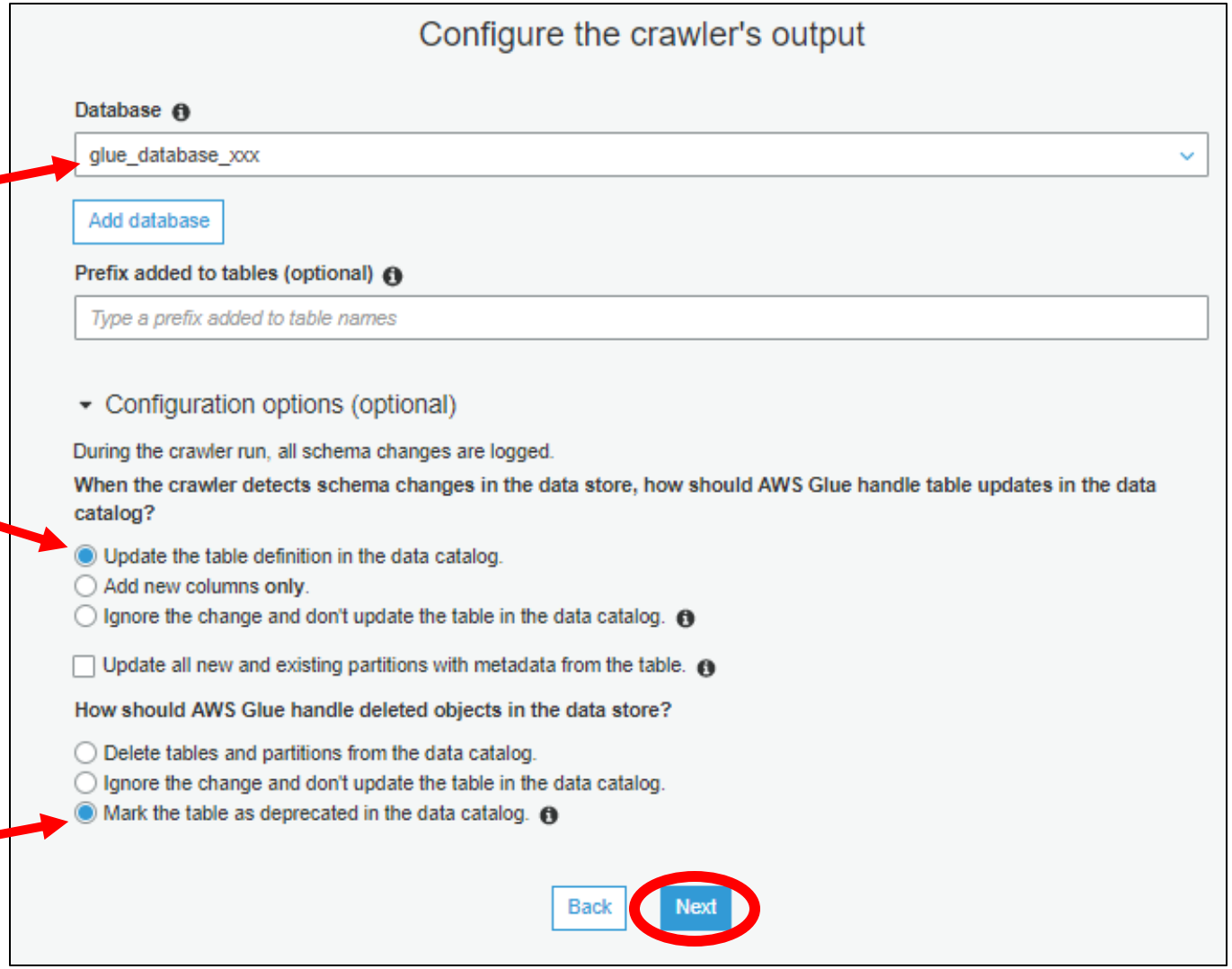
Glue Crawler

Create Table with Glue Crawler

Choose the database you created for your table to be in

The crawler will update the table if there is a change in the data

This will leave the table where it is but mark it as deprecated



Glue Crawler

Create Table with Glue Crawler

Crawler info

Name	glue_tutorial_xxx
Tags	-

IAM role

IAM role	arn:aws:iam::681132037743:role/service-role/AWSGlueServiceRole-DefaultRole

Schedule

Schedule	Run on demand

Output

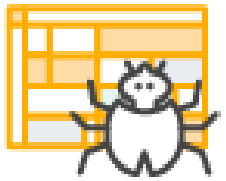
Database	glue_database_xxx
Prefix added to tables (optional)	
Create a single schema for each S3 path	false

▾ Configuration options

Schema updates in the data store	Update the table definition in the data catalog.
Object deletion in the data store	Mark the table as deprecated in the data catalog.

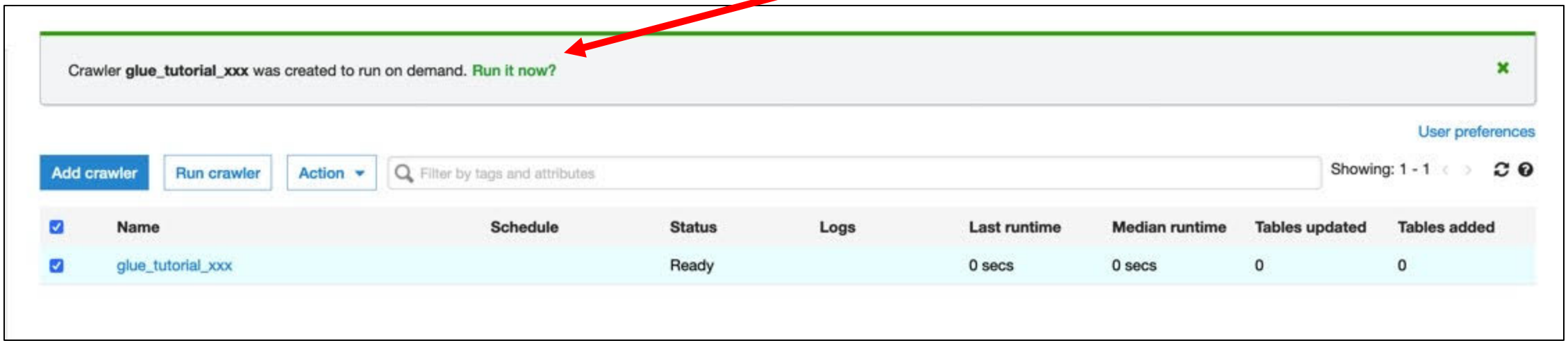
Back Finish

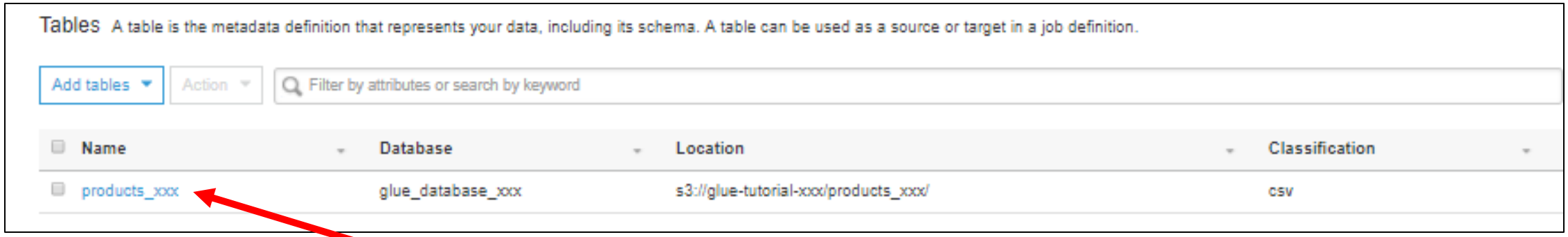
Glue Crawler



Create Table with Glue Crawler

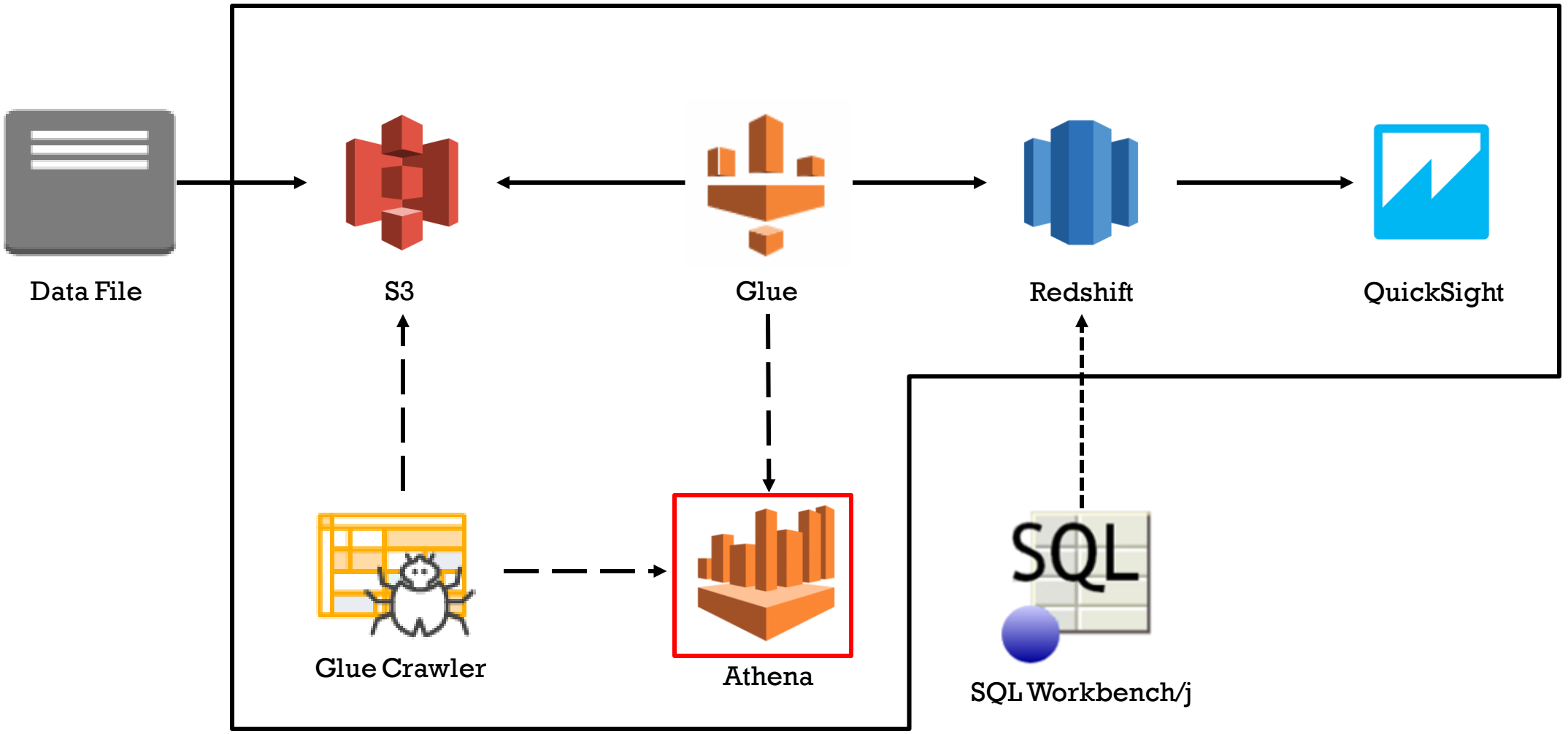
Run your crawler





Your table should be in the Tables tab

Athena

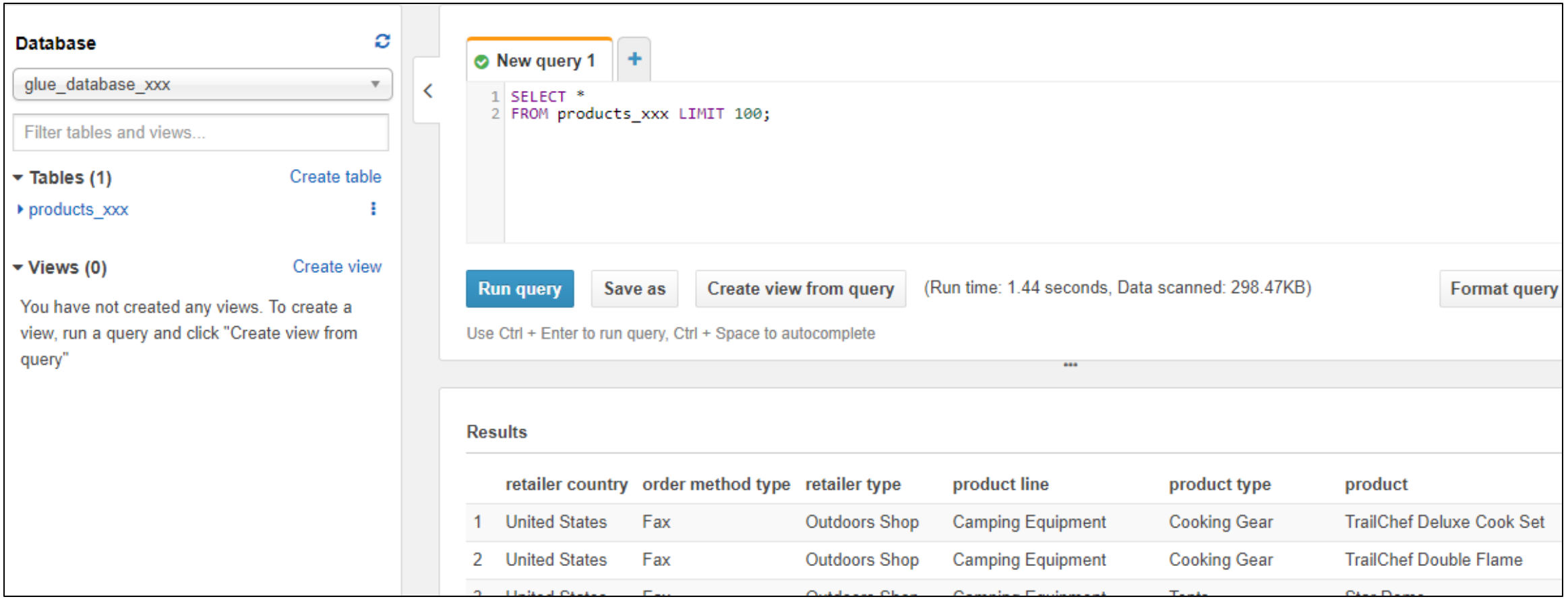

Data File
S3
Glue
Redshift
QuickSight
Glue Crawler
Athena
SQL
SQL Workbench/j

Athena

- Interactive query service used to analyze data
  - Data stored in S3
  - Run queries to verify your data is stored correctly

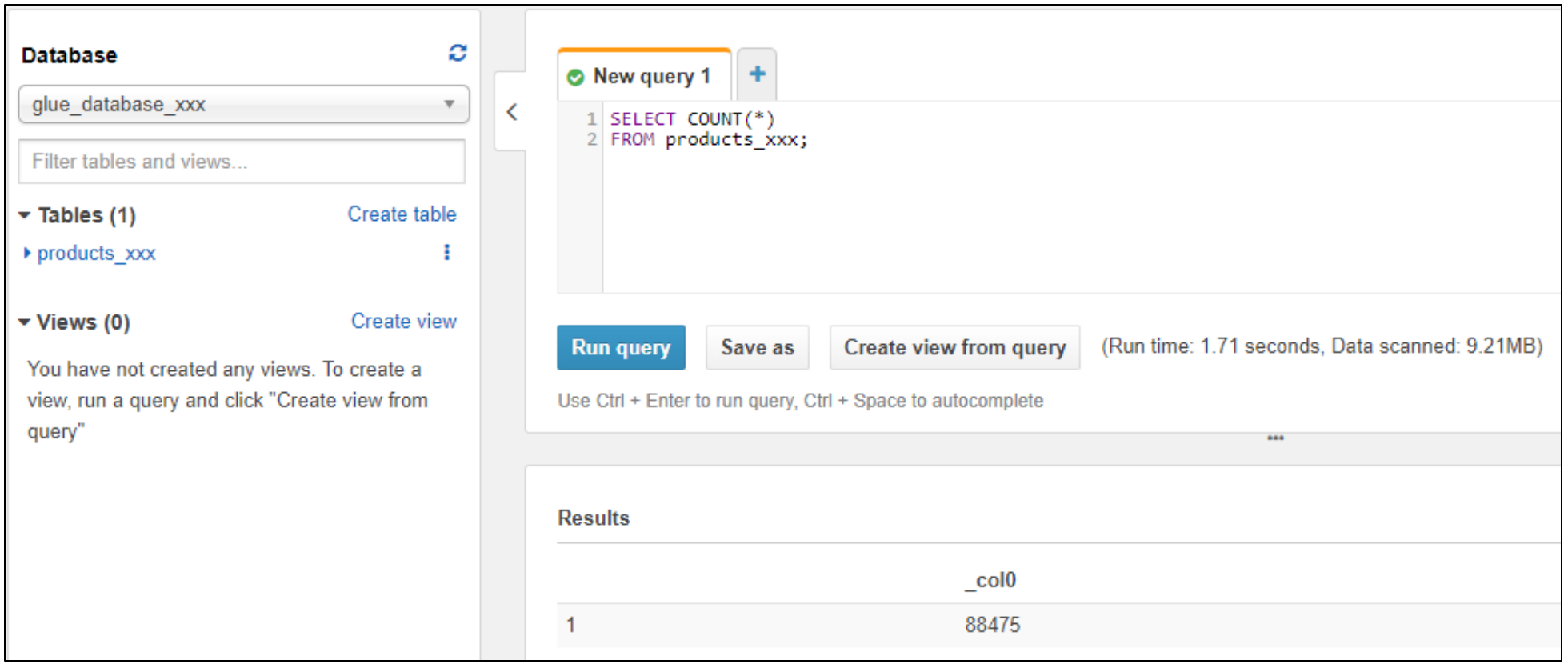
Athena

- Run an SQL select query to verify data populating correctly
- SELECT * FROM products_xxx LIMIT 100;



Athena

- Run an SQL count query to verify all data is there
- SELECT COUNT(*) FROM products_xxx;

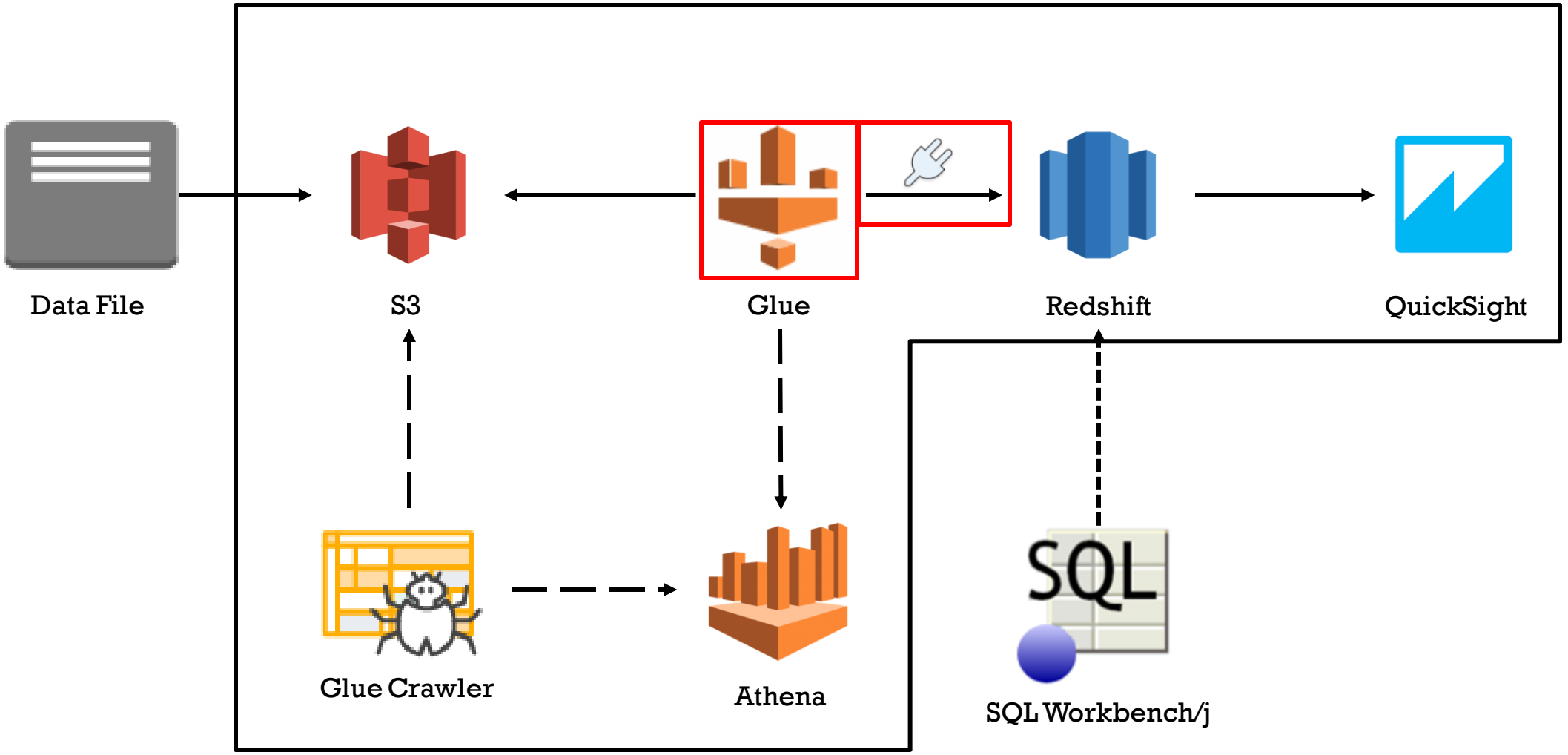


Lab 3

- Create/Run Glue Crawler
- Query using Athena

(Use US-EAST-2/Ohio Region)

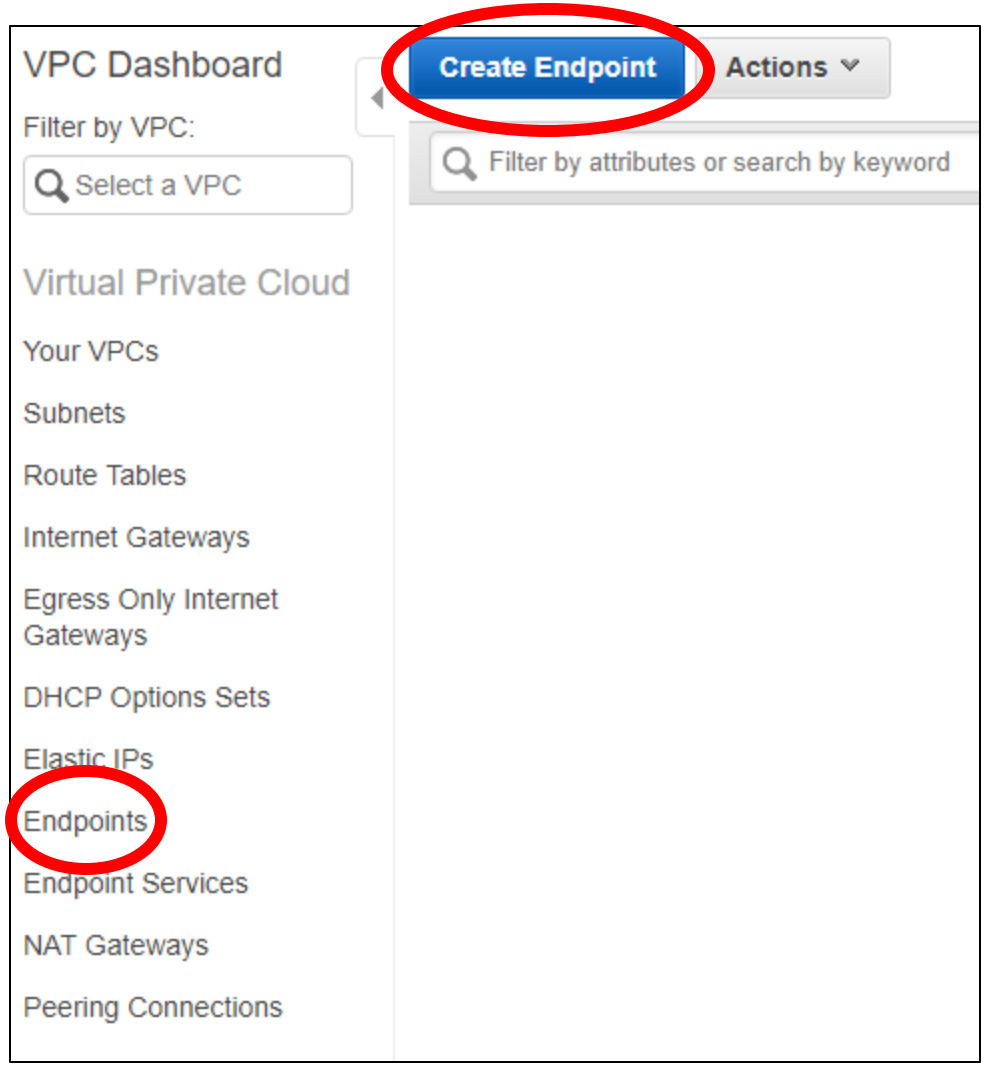
Glue


Data File
S3
Glue
Redshift
QuickSight
Glue Crawler
Athena
SQL
SQL Workbench/j

VPC

Create a S3 endpoint

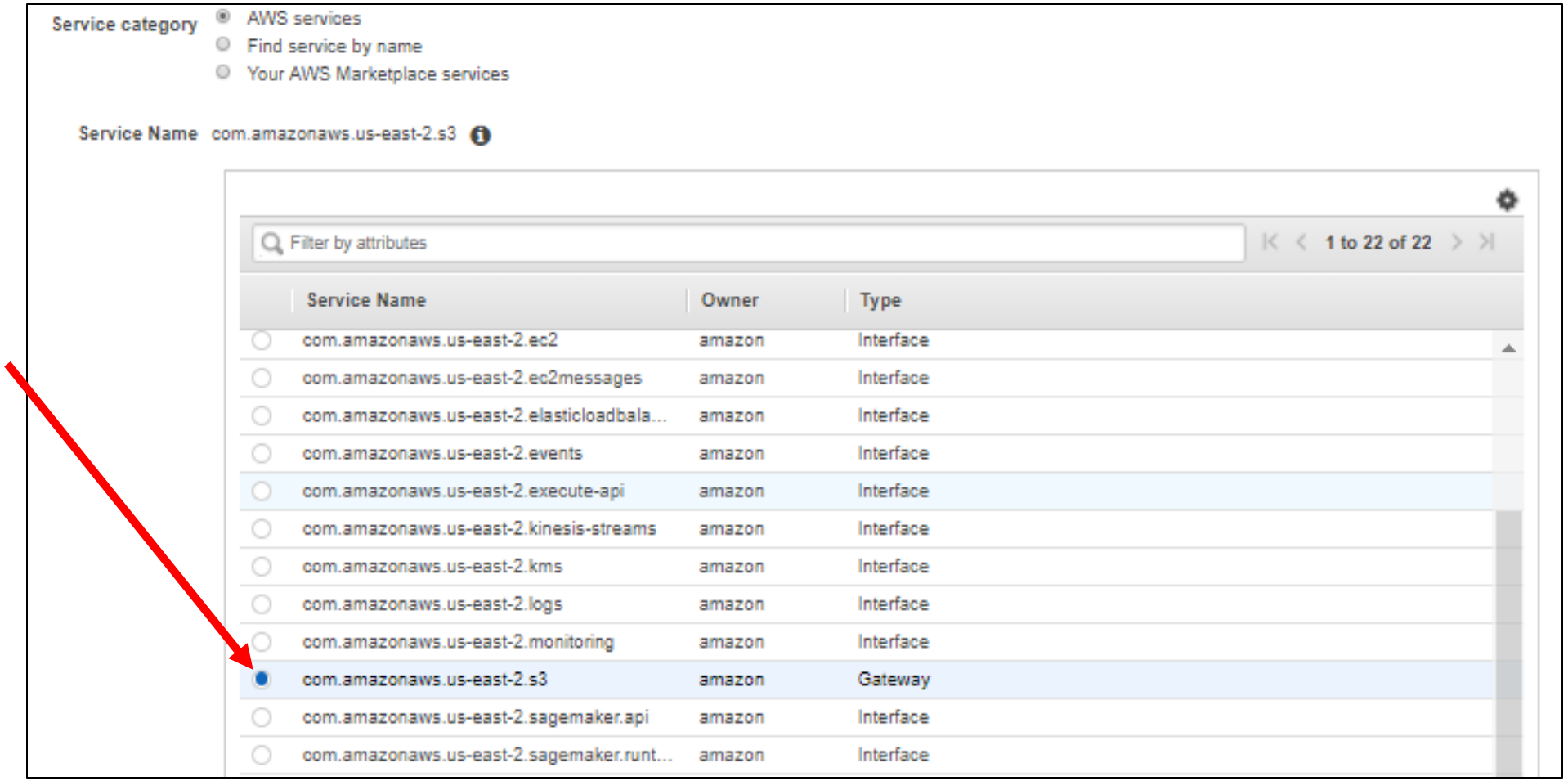
We need to create a S3 endpoint for Glue to access S3



VPC

Create a S3 endpoint

Select the S3 Service for Glue to access S3

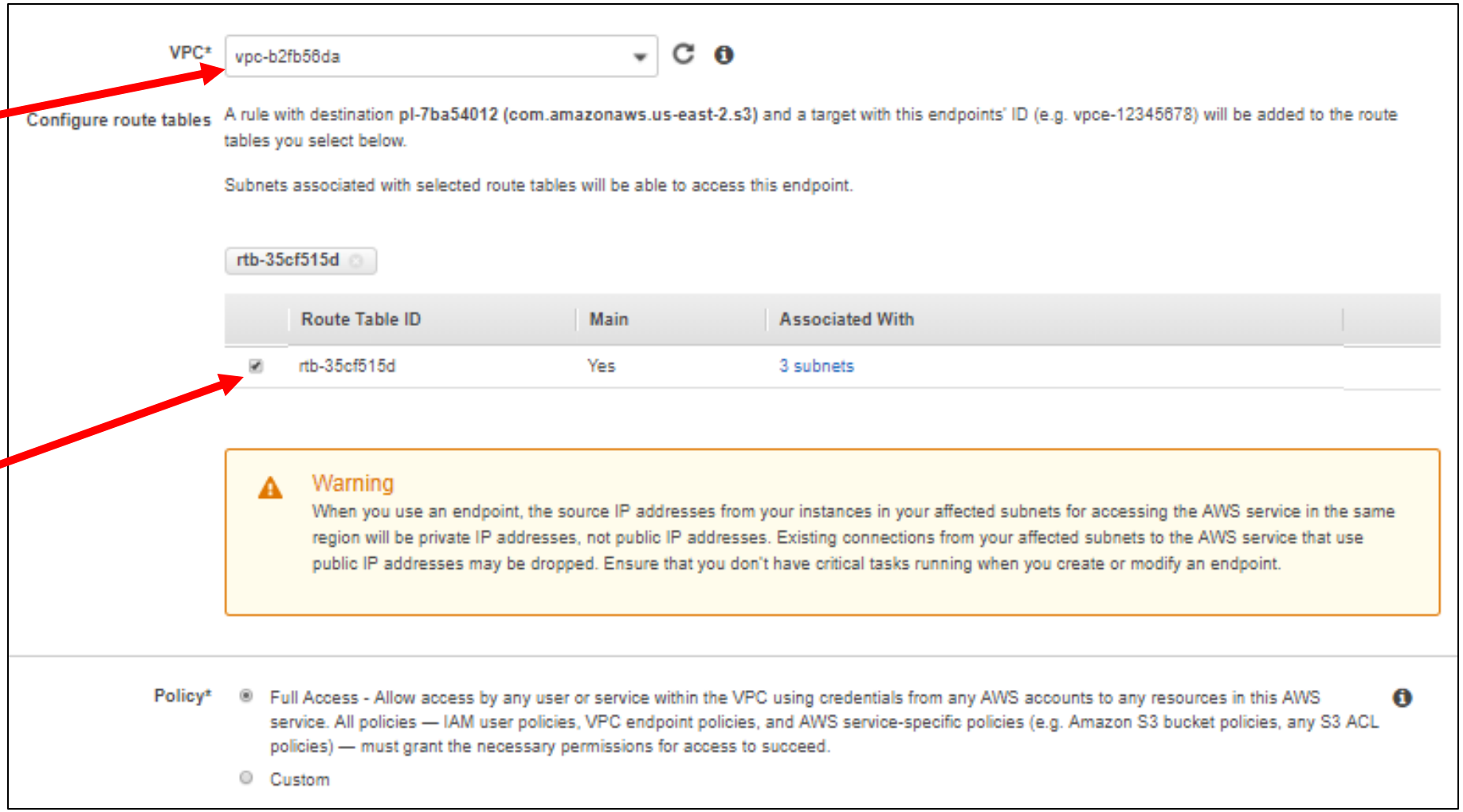


VPC

Create a S3 endpoint

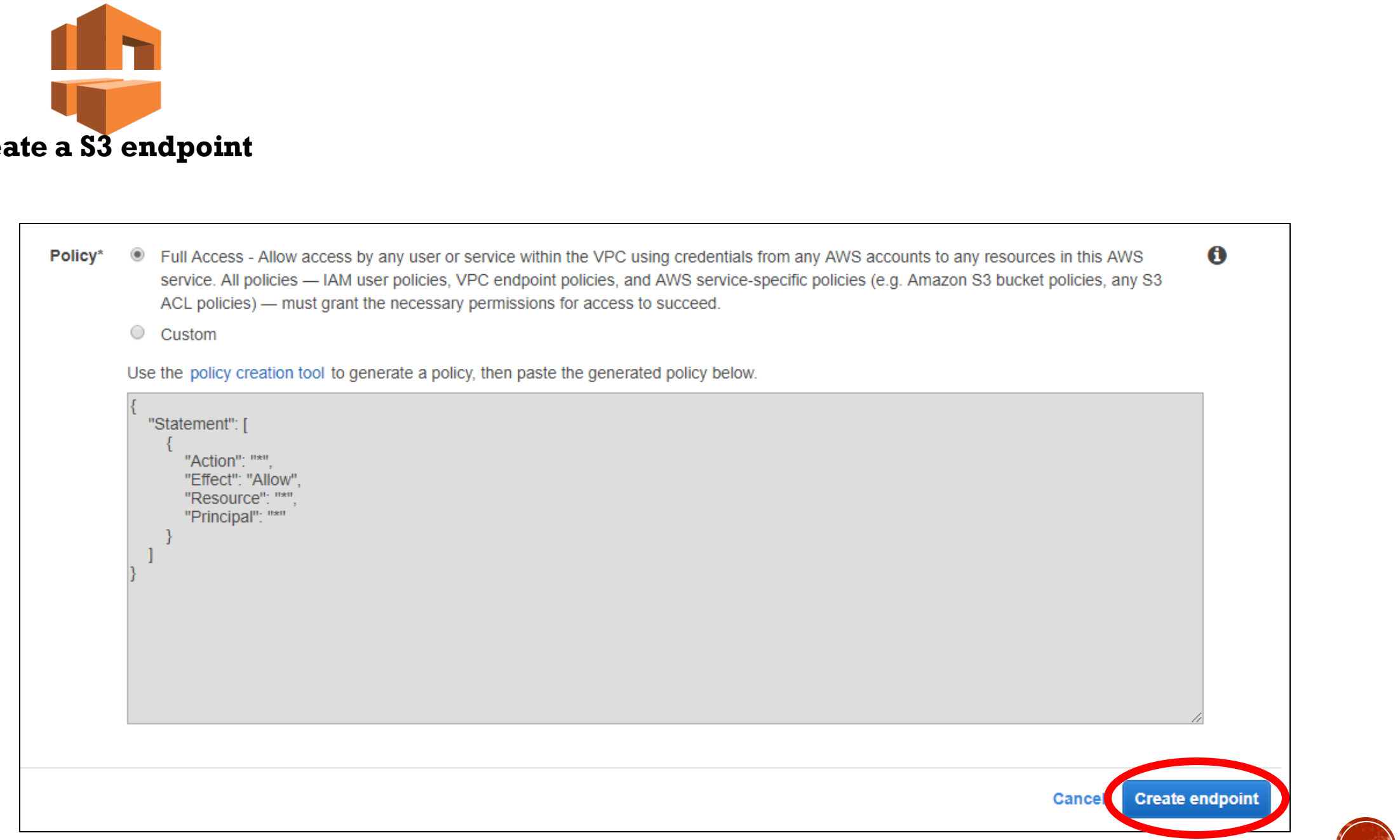
Choose VPC

Choose to add to the Route Table



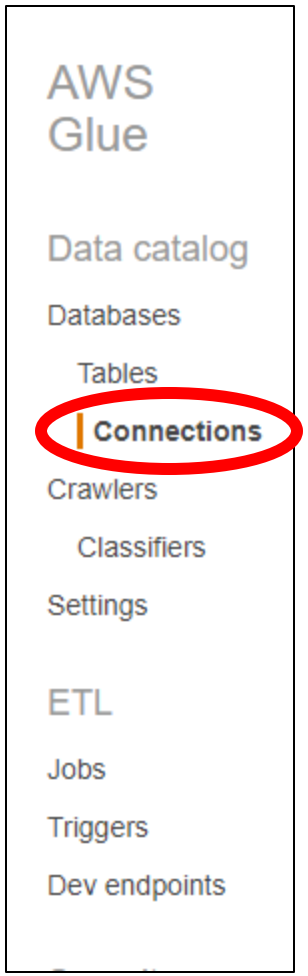
VPC

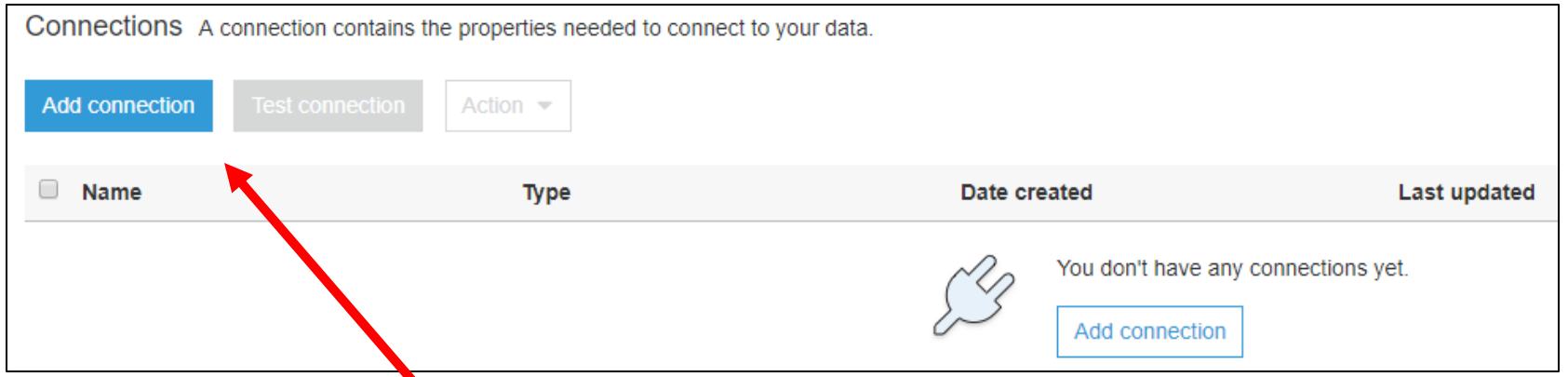
Create a S3 endpoint



Glue

Create a connection to Redshift



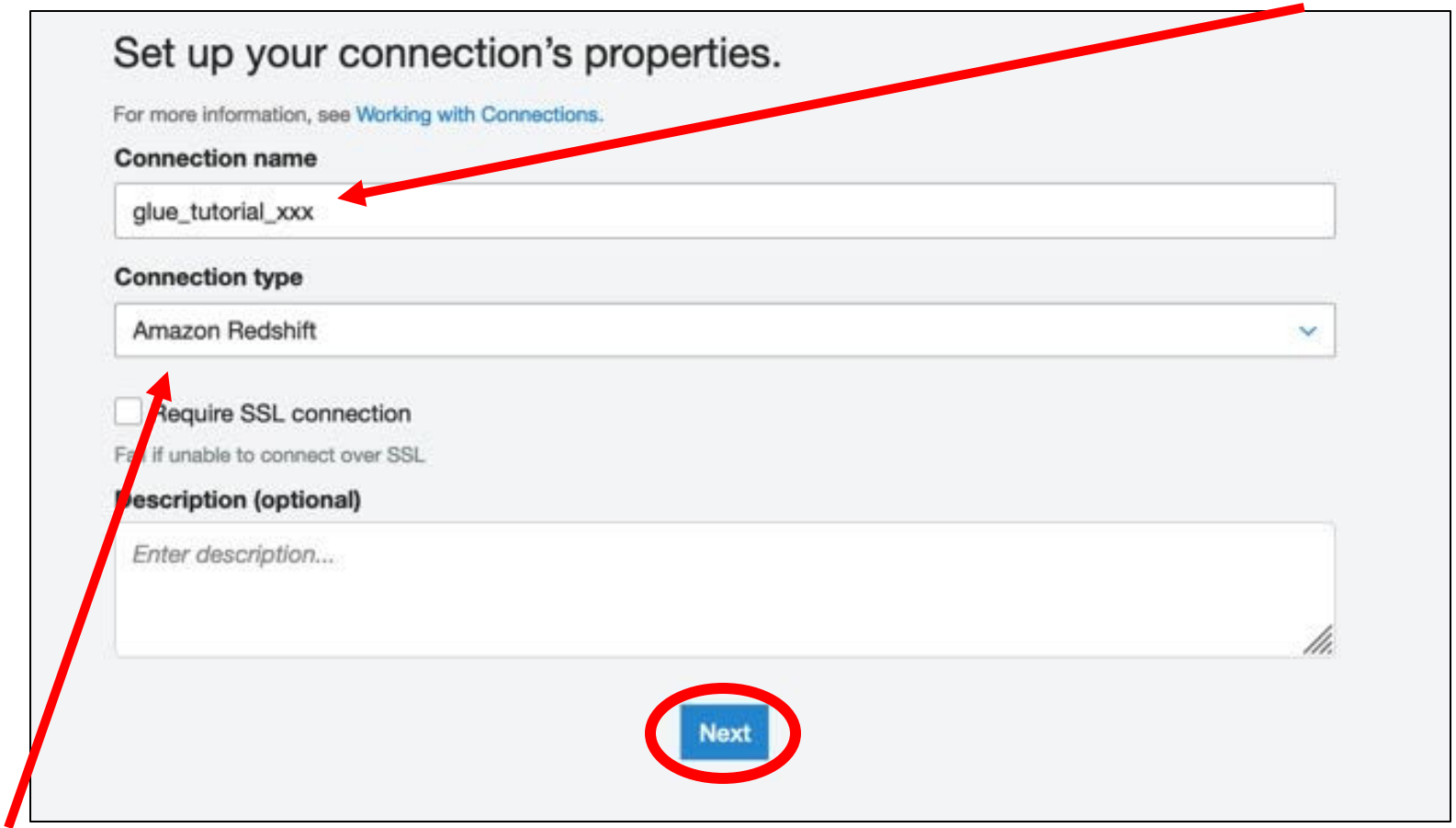


Click on "Add Connection" to create a connection to the Redshift cluster

Glue

Create a connection to Redshift

Name of the connection:
glue-tutorial-XXX

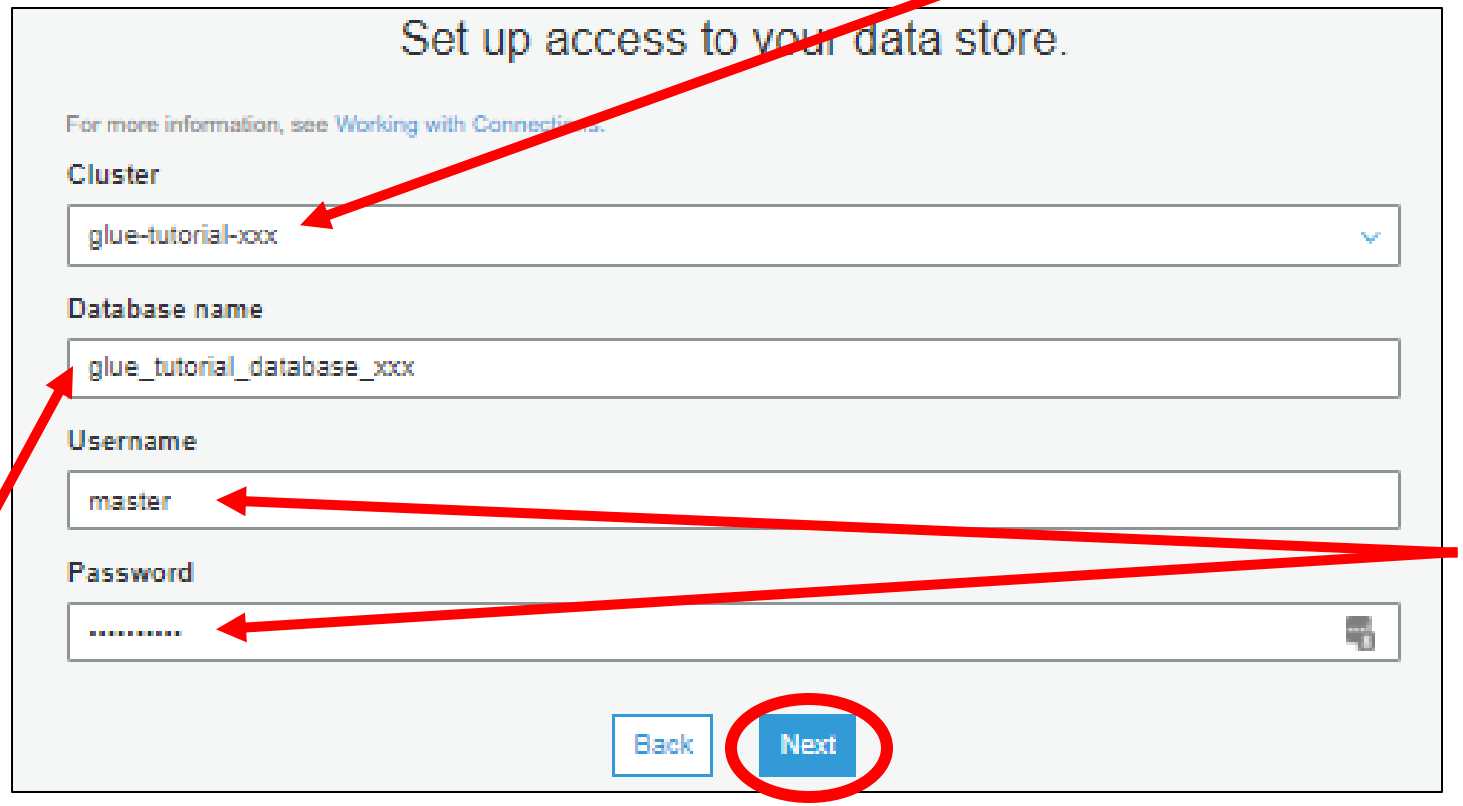


The connection type
should be Redshift

Glue

Create a connection to Redshift

Name of the cluster:
glue-tutorial-XXX

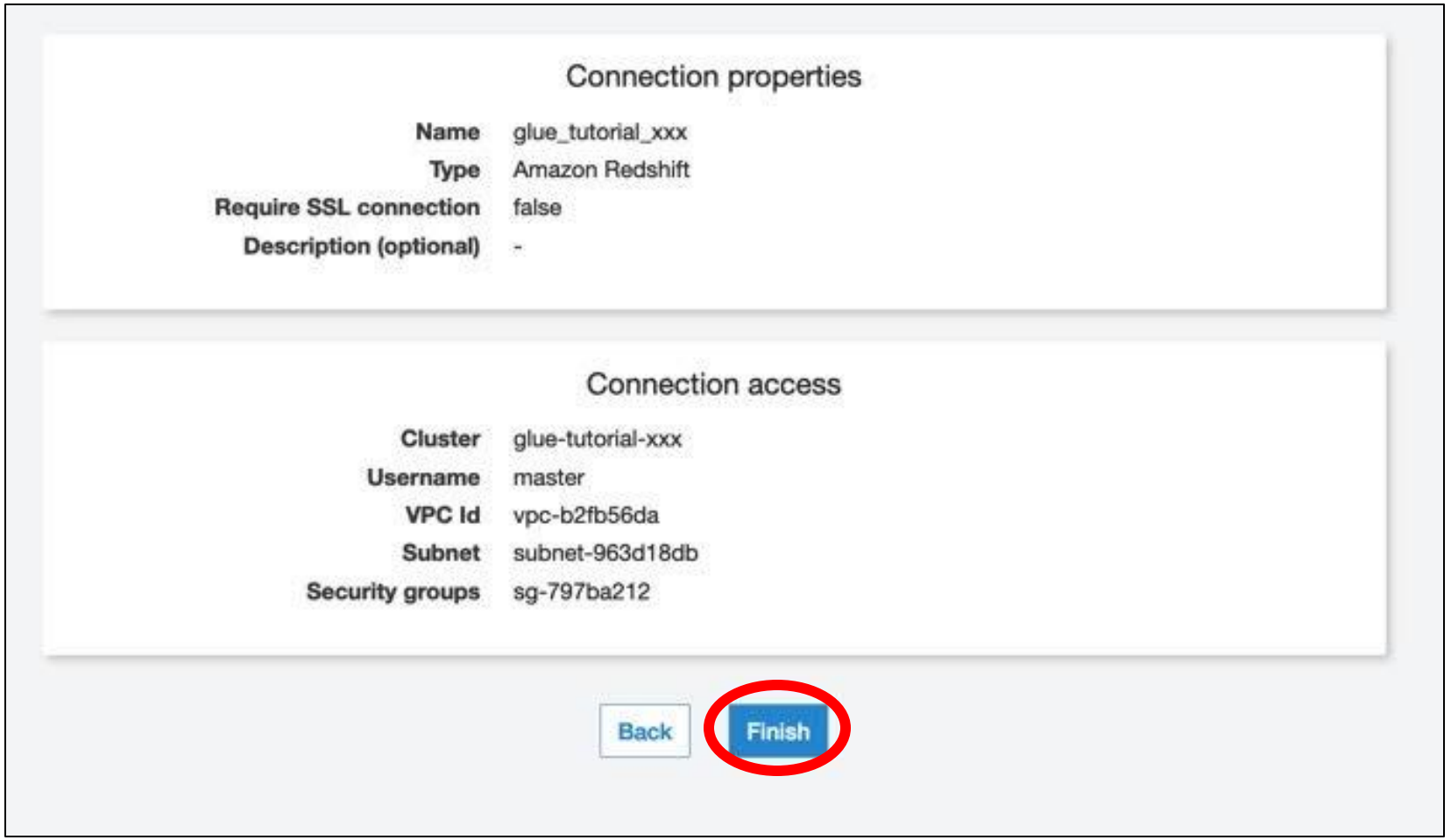


Username and password created for Redshift

Name of the database:
glue_tutorial_database_xxx

Glue

Create a connection to Redshift



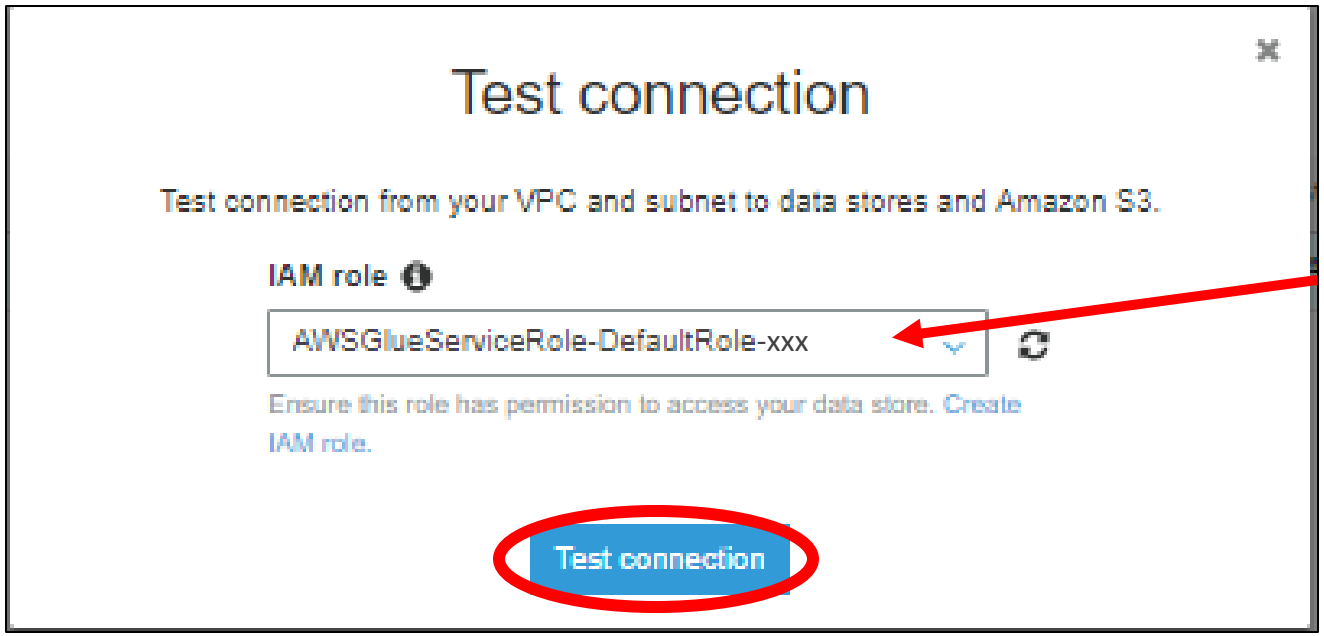
Glue

Test the connection to Redshift

Select newly created connection

Glue

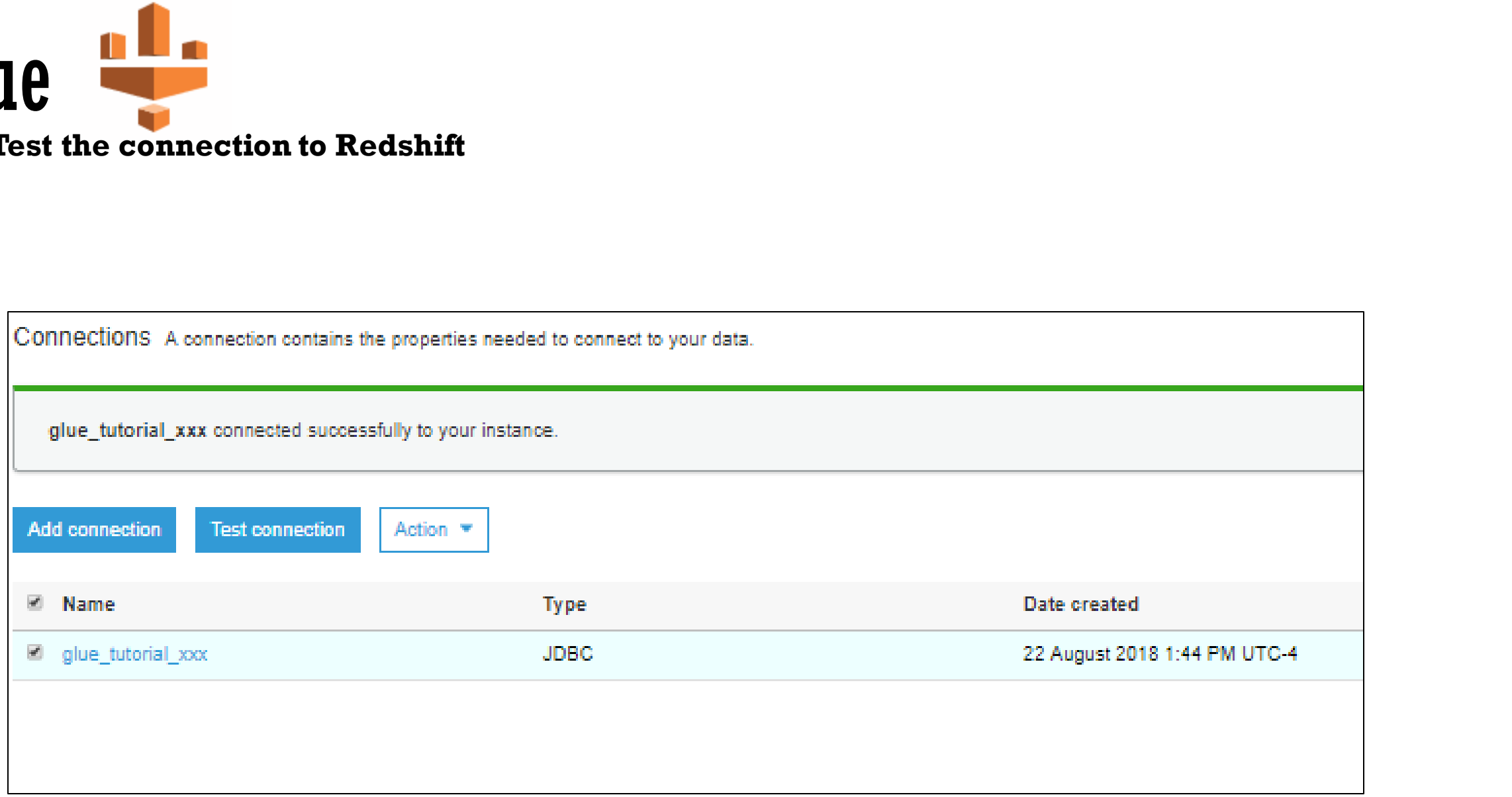
Test the connection to Redshift



Select your recently created IAM role

Glue

Test the connection to Redshift

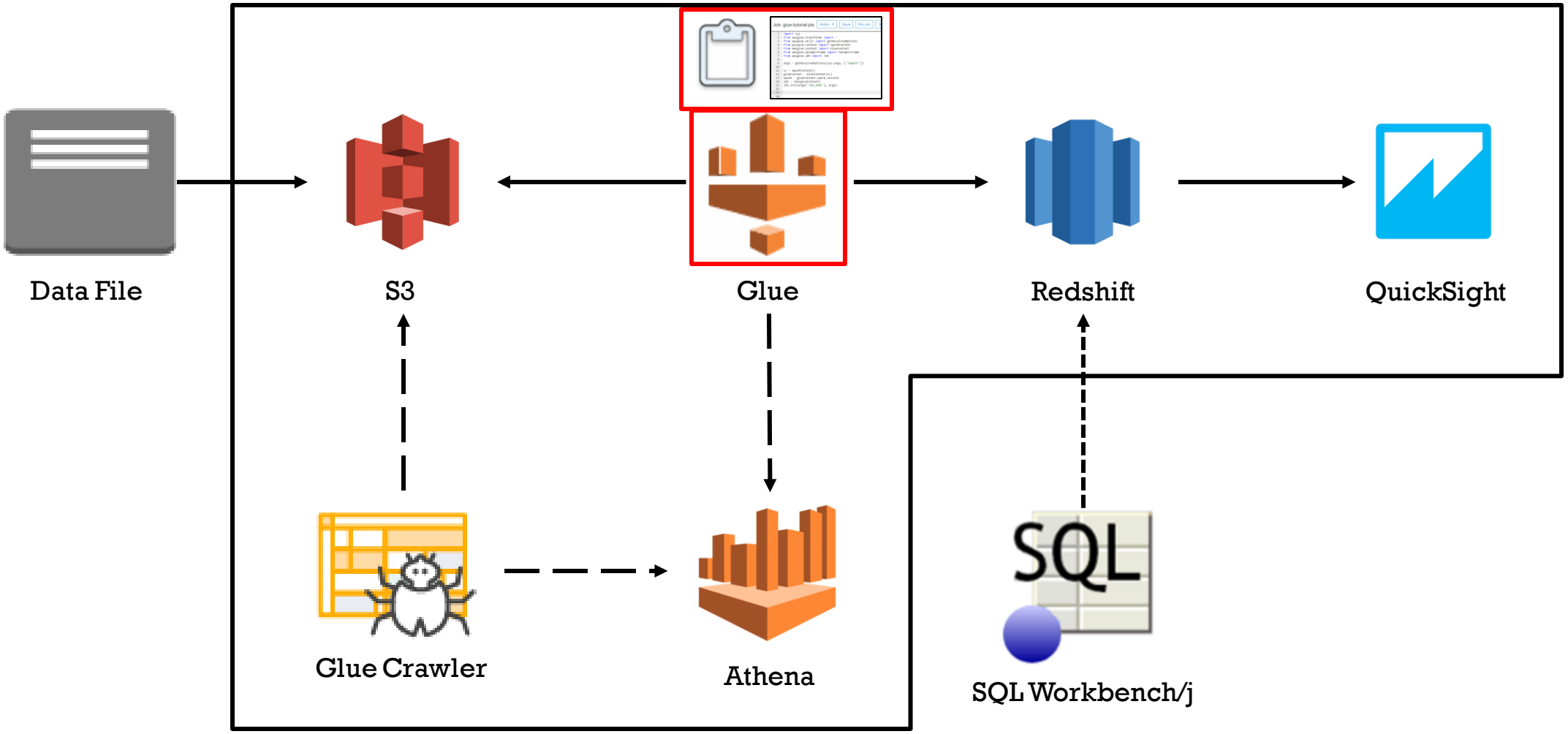


Lab 4

- Create S3 Endpoint
- Add Redshift Connection
- Test Redshift Connection

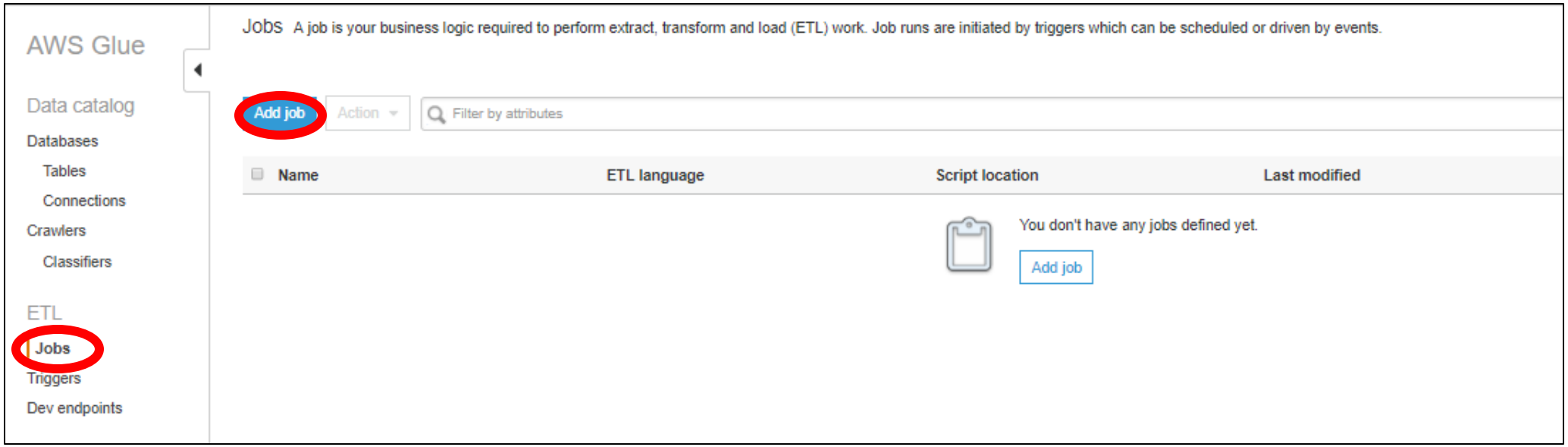
(Use US-EAST-2/ Ohio Region)

Glue


Data File
S3
Glue
Redshift
QuickSight
Glue Crawler
Athena
SQL
SQL Workbench/j

Glue

Create a Glue job



Glue

Create a Glue job

Give your job a name: glue_tutorial_XXX

The language used to write the script

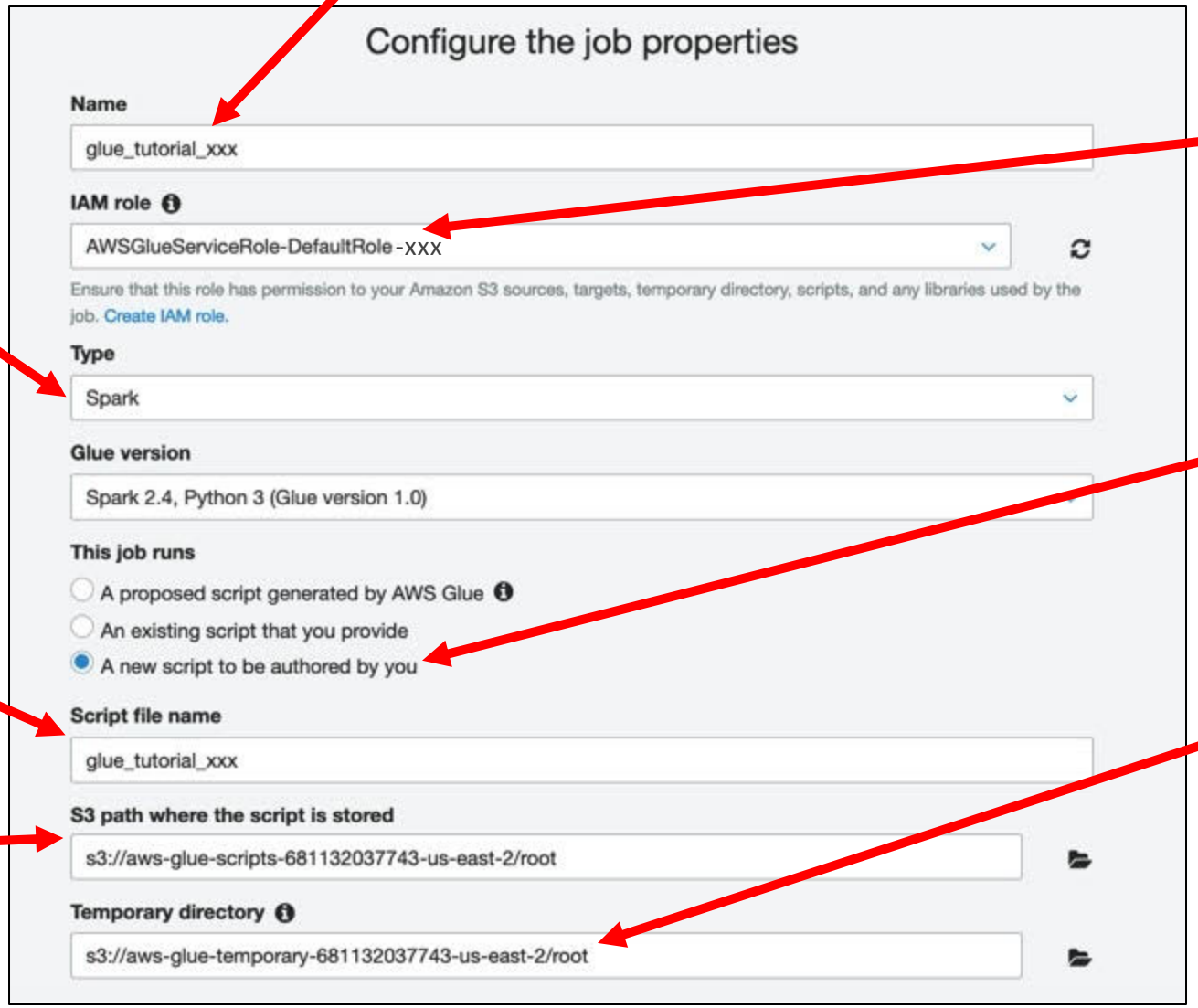
Give your job a role to perform the actions necessary to run

Give your script a name glue_tutorial_XXX

Create a new blank script

The location where your script will be placed in S3

This is where a temporary script is generated when the script is being edited

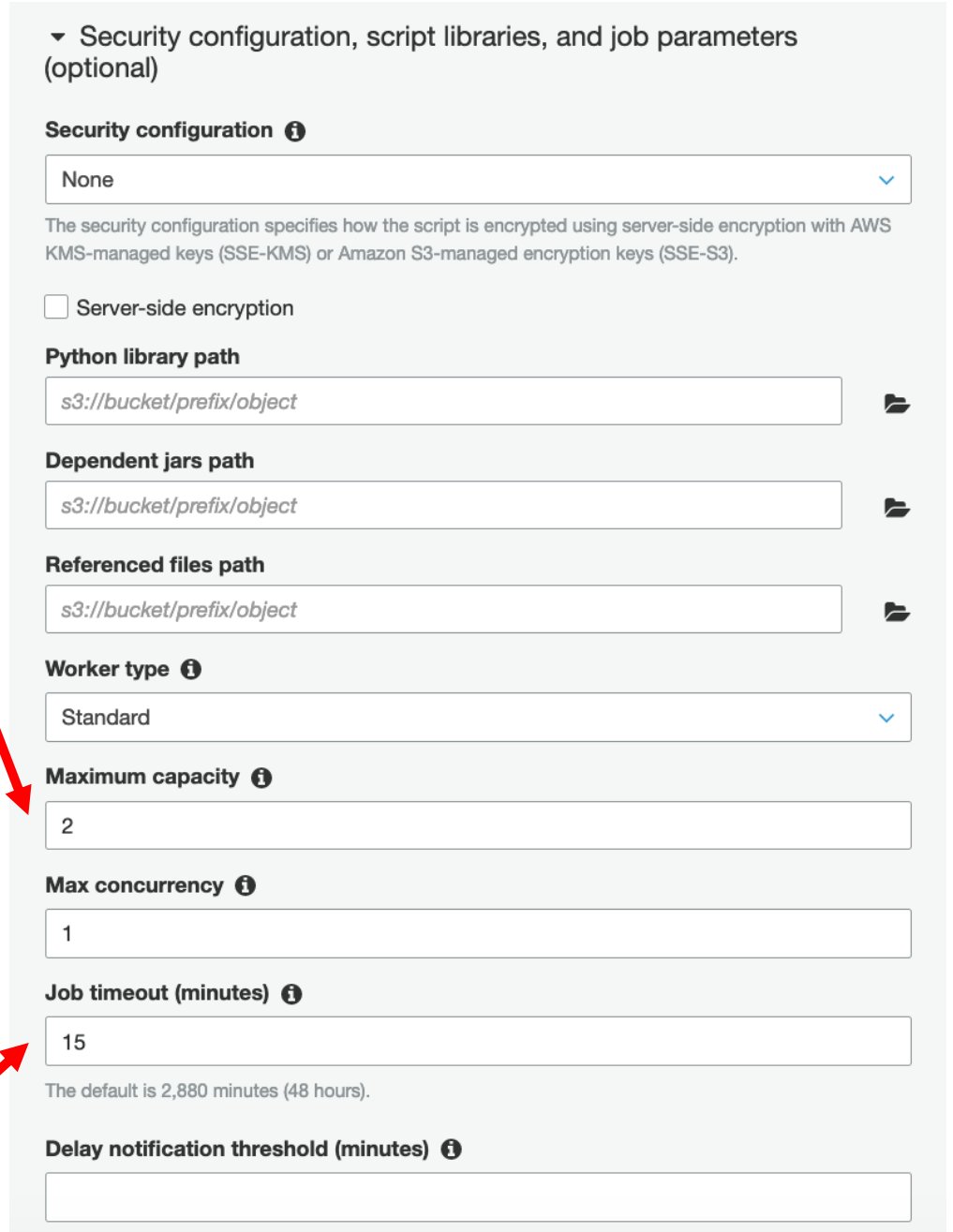


Glue

Create a Glue job

DPU = Data Processing Unit. Glue jobs are charged per DPU hour. Change to 2

Job automatically stops after set time



Glue

Create a Glue job

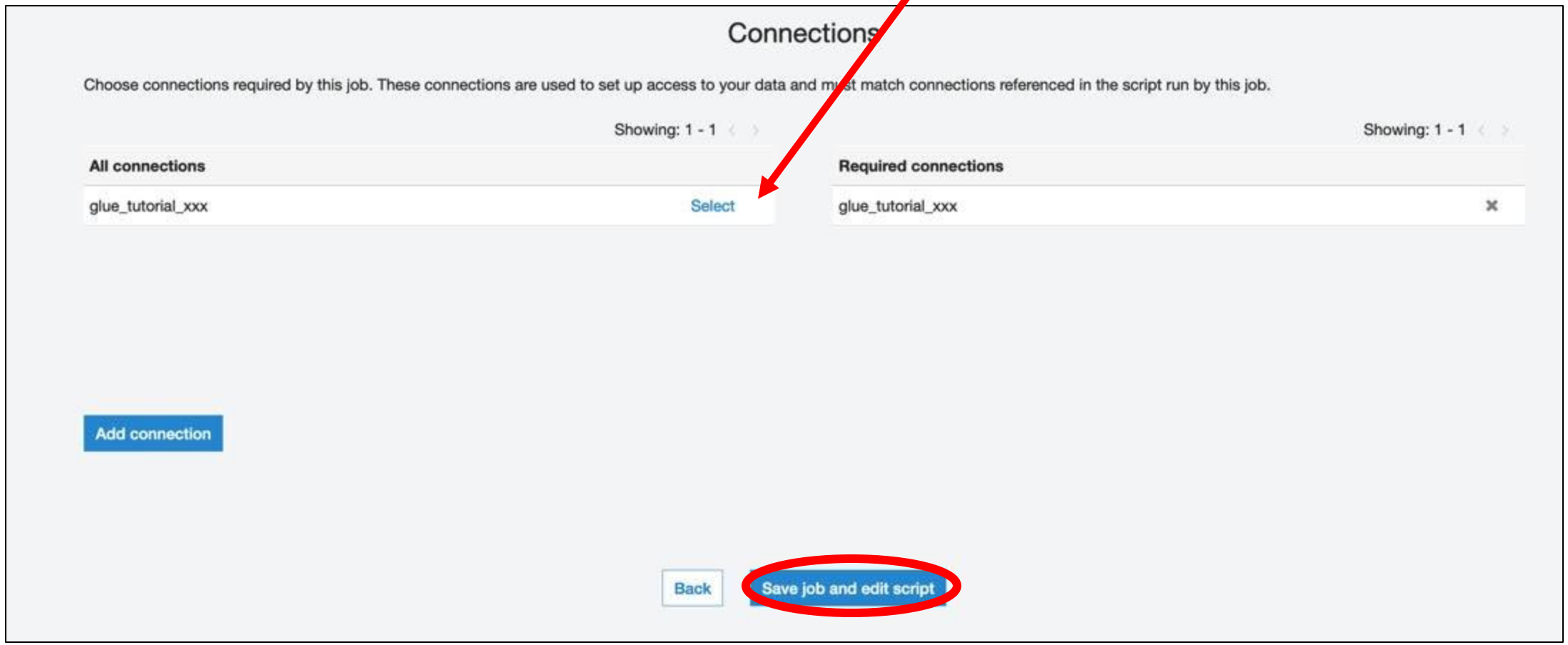
Parameterize values to be used in the script

Parameters:
--REDSHIFT_DB_NAME
glue_tutorial_database_xxx
--REDSHIFT_TABLE_NAME
products_redshift_table_xxx
--SCHEMA_NAME
sales_redshift_schema_xxx
--GLUE_DB_NAME
glue_database_xxx
--GLUE_TABLE_NAME
products_xxx
--CONNECTION_NAME
glue_tutorial_xxx

Glue

Create a Glue job

Select the Redshift connection that you want to use: glue-tutorial-XXX



Glue

Writing the Script

```
import sys
from awsglue.transforms import *
from awsglue.utils import getResolvedOptions
from pyspark.context import SparkContext
from awsglue.context import GlueContext
from awsglue.dynamicframe import DynamicFrame
from awsglue.job import Job
from pyspark.sql.functions import *
from pyspark.sql.types import *
from datetime import datetime

args = getResolvedOptions(sys.argv, ['TempDir', 'JOB_NAME', 'REDSHIFT_DB_NAME',
'REDSHIFT_TABLE_NAME', 'GLUE_DB_NAME', 'GLUE_TABLE_NAME', 'SCHEMA_NAME',
'CONNECTION_NAME'])
```

PySpark is a service that allows the developer to perform data analysis on the data that is being used.

Include SQL functions, types, and datetime to use later

Add the parameters that were passed into the Glue job

Glue

Writing the Script

```
…
sc = SparkContext()
glueContext = GlueContext(sc)
spark = glueContext.spark_session
job = Job(glueContext)
job.init(args['JOB_NAME'], args)
datasource =
glueContext.create_dynamic_frame.from_catalog(
    database = args[‘GLUE_DB_NAME'],
    table_name = args[‘GLUE_TABLE_NAME']
)
```

This is setting up the Spark and Glue environment to be able to interact with the data

The data will be written to the datasource as a DynamicFrame

These are the database and the table that we created in Glue

Glue

Writing the Script

```
...
# Convert to PySpark Data Frame
sourcedata = datasource.toDF()

split_col = split(sourcedata["quarter"], " ")
sourcedata = sourcedata.withColumn("quarter new", split_col.getItem(0))
sourcedata = sourcedata.withColumn("profit", col("revenue")*col("gross margin"))
sourcedata = sourcedata.withColumn("current date", current_date())

# Convert back to Glue Dynamic Frame
datasource = DynamicFrame.fromDF(sourcedata, glueContext, "datasource")
```

sourcedata needs to be set to a Data Frame

This is where the transformations happen

Convert back to a Dynamic Frame

Glue

Writing the Script

```
...
applymapping = ApplyMapping.apply(
    frame = datasource,
    mappings = [
        ("retailer country", "string", "retailer_country", "varchar(20)"),
        ("order method type", "string", "order_method_type", "varchar(15)"),
        ("retailer type", "string", "retailer_type", "varchar(30)"),
        ("product line", "string", "product_line", "varchar(30)"),
        ("product type", "string", "product_type", "varchar(30)"),
        ("product", "string", "product", "varchar(50)"),
        ("year", "bigint", "year", "varchar(4)"),
        ("quarter new", "string", "quarter", "varchar(2)"),
        ("revenue", "double", "revenue", "numeric"),
        ("quantity", "bigint", "quantity", "integer"),
        ("gross margin", "double", "gross_margin", "decimal(15,10)"),
        ("profit", "double", "profit", "numeric"),
        (“current date", "date", “current_date", "date")
    ]
```

This is how the data in the DynamicFrame will be mapped to the columns in Redshift

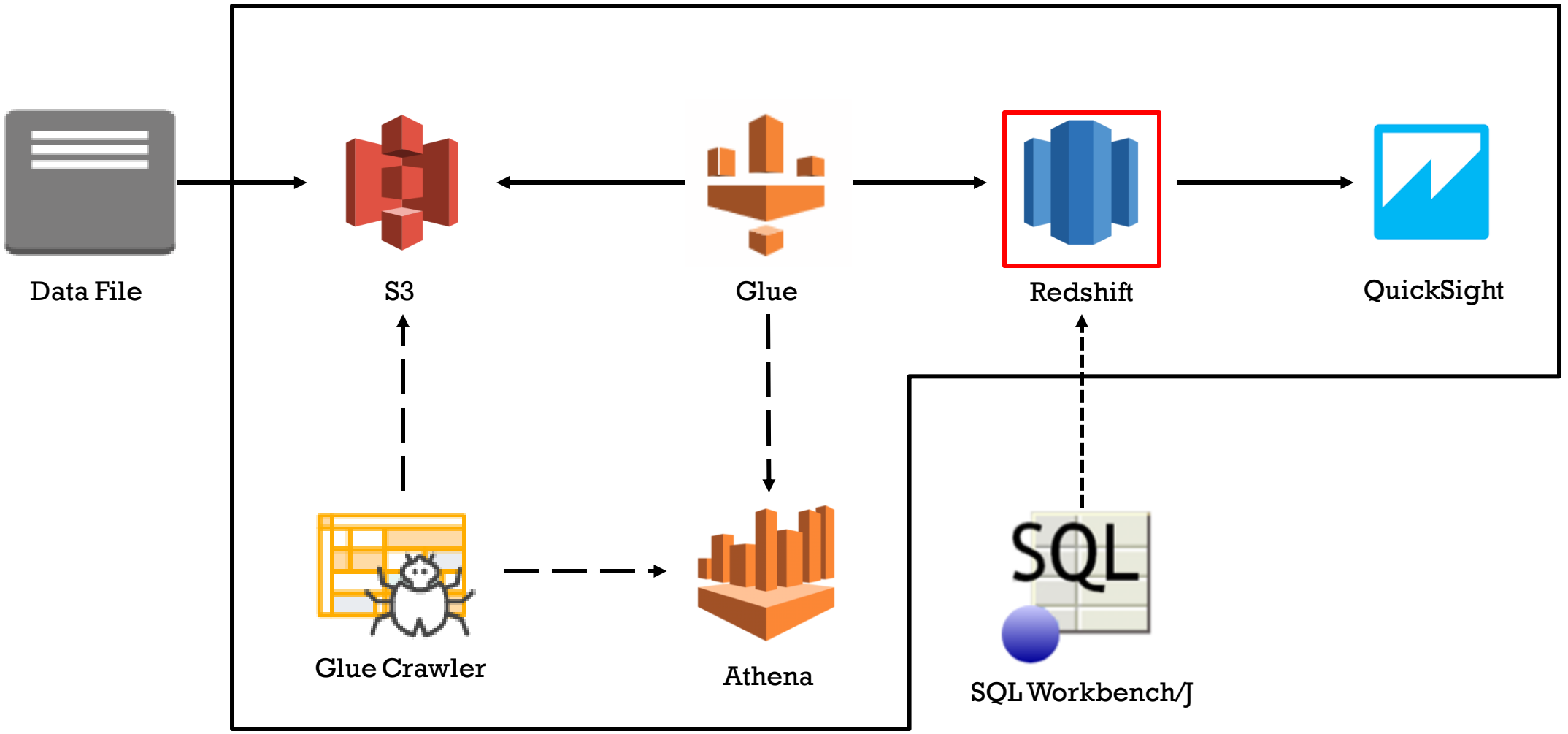
Glue

Writing the Script

```
...
# datasink (loading) using spark
datasink = glueContext.write_dynamic_frame.from_jdbc_conf(
    frame = applymapping,
    catalog_connection = args['CONNECTION_NAME'],
    connection_options = {
        "dbtable": "{}.{}".format(args['SCHEMA_NAME'], args['‘REDSHIFT_TABLE_NAME']),
        "database": args['REDSHIFT_DB_NAME']
    },
    redshift_tmp_dir = args["TempDir"]
)
```

The datasink will connect to Redshift using the parameters given and load the data to Redshift

Redshift


Data File
S3
Glue
Redshift
QuickSight
Glue Crawler
Athena
SQL
SQL Workbench/J

Redshift

Create table

Copy the SQL script from the repository into SQL Workbench

SQL Workbench/J GlueTutorial - Default.wksp

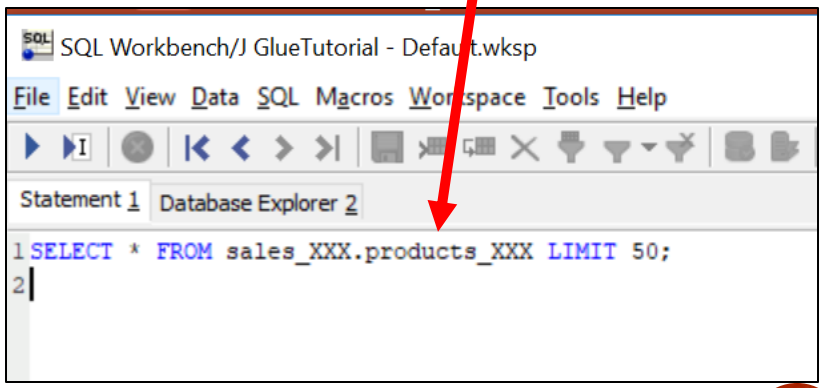
File Edit View Data SQL Macros Workspace Tool

Statement 1 | Database Explorer 2

```
1 CREATE SCHEMA sales_XXX;
2
3 CREATE TABLE sales_XXX.products_XXX
4 (
5    retailer_country    varchar(20),
6    order_method_type   varchar(15),
7    retailer_type       varchar(30),
8    product_line        varchar(30),
9    product_type        varchar(30),
10   product             varchar(50),
11   year                varchar(4),
12   quarter             varchar(2),
13   revenue             numeric(15,2),
14   quantity            integer,
15   gross_margin        numeric(15,10),
16   profit              numeric(15,2),
17   timestamp           date
18 );
```

Add your own initials to the schema and table names

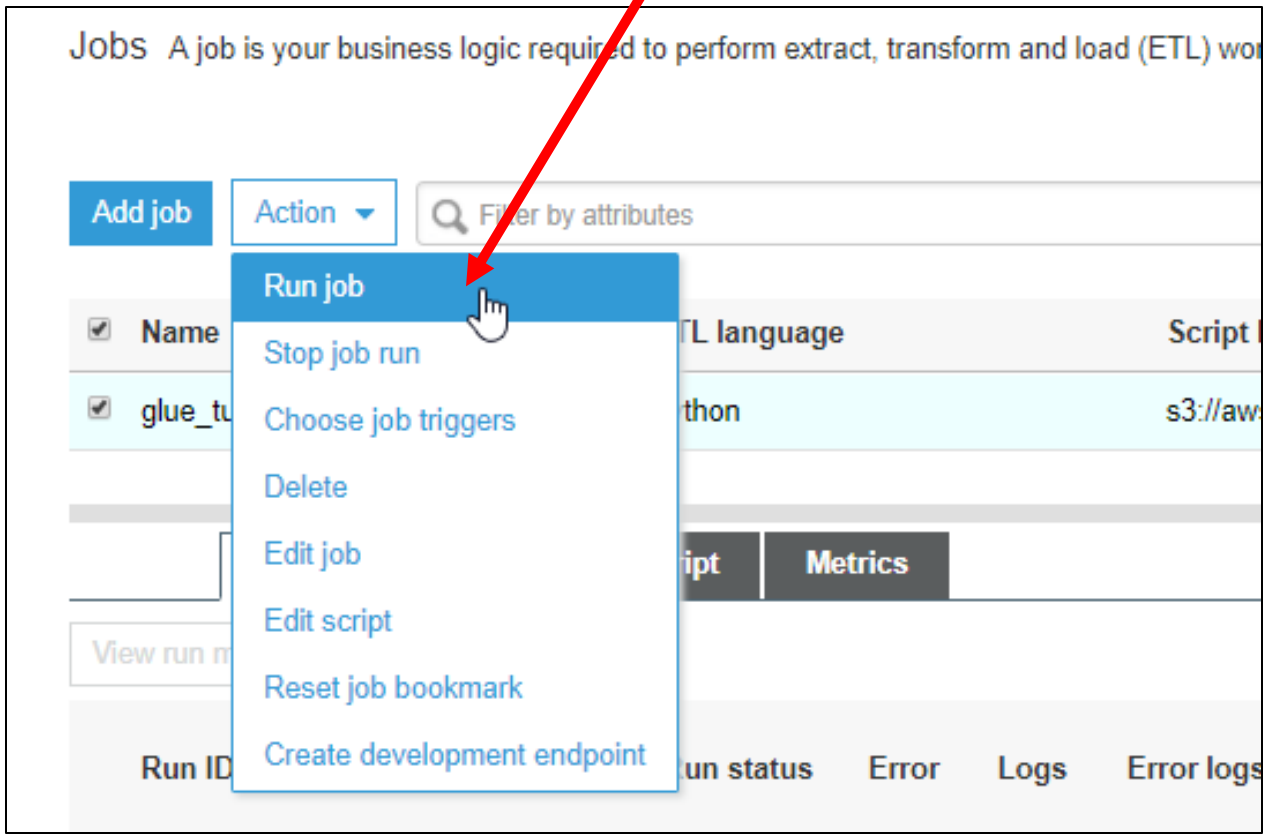
Run a SELECT to make sure your table was made and nothing is in it

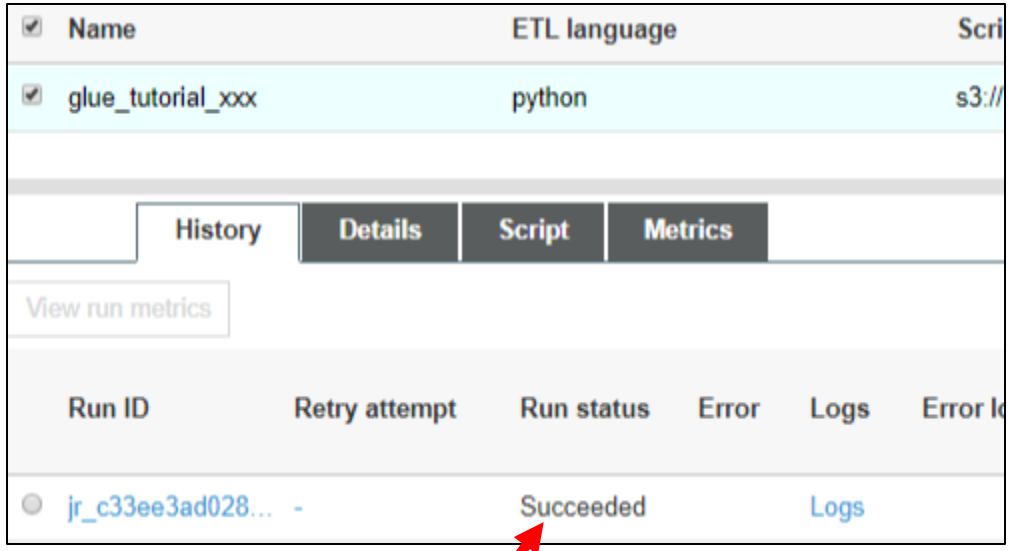


Glue

Run the Glue job

Go back to Glue and run your Glue job





When the job succeeds, check your Redshift table

Redshift

Verify data in the table

```
1 SELECT *
2 FROM sales_redshift_schema_xxx.products_redshift_table_xxx LIMIT 100;
3
4
```

Result 1 | Messages

retailer_country	order_method_type	retailer_type	product_line	product_type	product	year	revenue	quantity	gross_margin	profit	timestamp	quarter	current_date
United States	Fax	Outdoors Shop	Camping Equipment	Cooking Gear	TrailChef Deluxe Cook Set	2012	59628.66	489	0.35	20723.82		Q1	2018-08-29
United States	Fax	Outdoors Shop	Camping Equipment	Tents	Star Dome	2012	89940.48	147	0.35	31728.48		Q1	2018-08-29
United States	Fax	Outdoors Shop	Camping Equipment	Sleeping Bags	Hibernator Lite	2012	119822.20	1415	0.29	34922.20		Q1	2018-08-29
United States	Fax	Outdoors Shop	Camping Equipment	Sleeping Bags	Hibernator Camp Cot	2012	41837.46	426	0.34	14040.96		Q1	2018-08-29
United States	Fax	Outdoors Shop	Camping Equipment	Lanterns	Firefly Extreme	2012	9393.30	189	0.43	4078.62		Q1	2018-08-29
United States	Fax	Outdoors Shop	Camping Equipment	Lanterns	EverGlow Butane	2012	6940.03	109	0.36	2511.36		Q1	2018-08-29
United States	Fax	Outdoors Shop	Mountaineering Equipment	Rope	Husky Rope 60	2012	14109.40	79	0.29	4115.11		Q1	2018-08-29
United States	Fax	Outdoors Shop	Mountaineering Equipment	Rope	Husky Rope 200	2012	77288.64	143	0.31	24328.59		Q1	2018-08-29
United States	Fax	Outdoors Shop	Mountaineering Equipment	Safety	Husky Harness	2012	34154.90	559	0.28	9687.47		Q1	2018-08-29
United States	Fax	Outdoors Shop	Mountaineering Equipment	Safety	Granite Signal Mirror	2012	4074.84	126	0.51	2095.38		Q1	2018-08-29
United States	Fax	Outdoors Shop	Mountaineering Equipment	Climbing Accessories	Granite Belay	2012	19476.80	296	0.48	9273.68		Q1	2018-08-29
United States	Fax	Outdoors Shop	Mountaineering Equipment	Climbing Accessories	Firefly Climbing Lamp	2012	17998.56	464	0.43	7697.76		Q1	2018-08-29
United States	Fax	Outdoors Shop	Mountaineering Equipment	Climbing Accessories	Firefly Rechargeable Battery	2012	11673.60	1520	0.59	6885.60		Q1	2018-08-29
United States	Fax	Outdoors Shop	Mountaineering Equipment	Tools	Granite Ice	2012	25041.60	333	0.48	12064.59		Q1	2018-08-29
United States	Fax	Outdoors Shop	Mountaineering Equipment	Tools	Granite Shovel	2012	9543.16	164	0.34	3216.04		Q1	2018-08-29
United States	Fax	Outdoors Shop	Mountaineering Equipment	Tools	Granite Axe	2012	32870.40	856	0.49	16161.28		Q1	2018-08-29
United States	Fax	Outdoors Shop	Personal Accessories	Watches	Mountain Man Extreme	2012	6499.80	23	0.59	3827.43		Q1	2018-08-29
United States	Fax	Outdoors Shop	Personal Accessories	Eyewear	Polar Ice	2012	3825.80	37	0.52	1987.27		Q1	2018-08-29
United States	Fax	Outdoors Shop	Personal Accessories	Knives	Bear Survival Edge	2012	8414.75	97	0.48	4049.75		Q1	2018-08-29
United States	Fax	Outdoors Shop	Outdoor Protection	Insect Repellents	BugShield Extreme	2012	25010.58	3801	0.63	15812.16		Q1	2018-08-29
United States	Fax	Outdoors Shop	Outdoor Protection	First Aid	Compact Relief Kit	2012	4057.20	180	0.60	2437.20		Q1	2018-08-29

Lab 5

- Create Glue Job
- Redshift Schema and Table
- Run Glue Job
- Query Redshift

(Use US-EAST-2/Ohio Region)

Enhancements

Improve the versatility of your Glue job

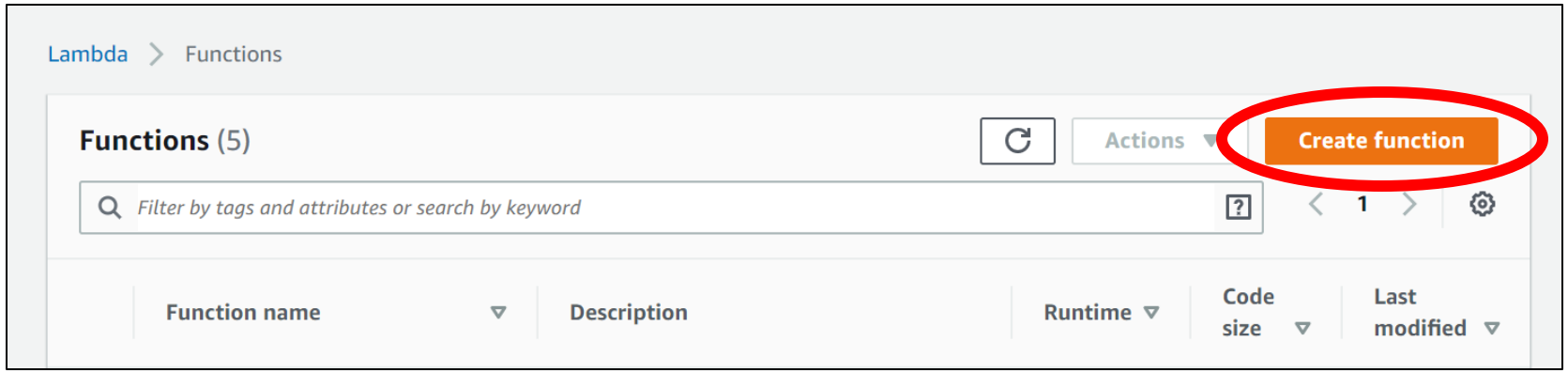
- Create a Glue Trigger
  - Automatically run the Glue job
  - Run multiple different Glue jobs
- Control how resources can interact with other services
- Create reports for business analytics with the data that was loaded with the Glue job.
- Easily create, modify, and delete as well as move Glue jobs with a template

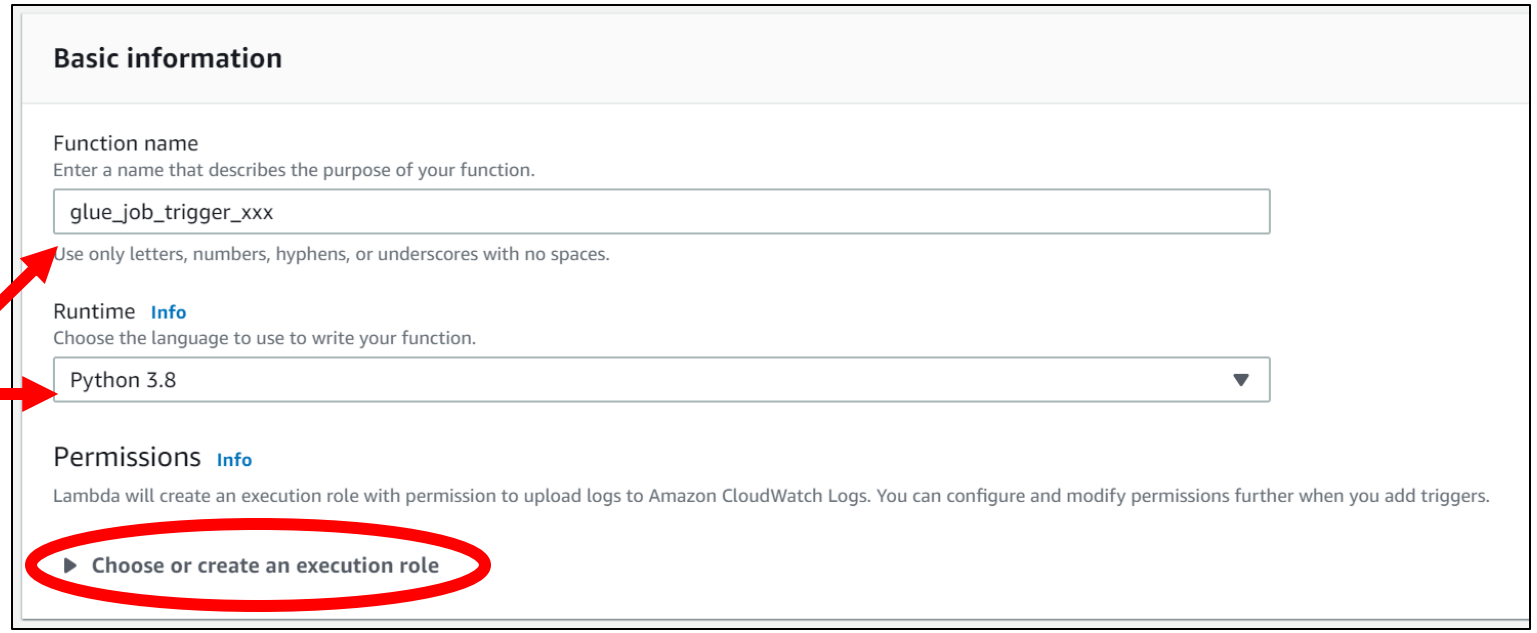
Glue Trigger

Automatically run Glue job using Lambda – a serverless function

Instead of running the Glue job manually, have it run automatically when a file is added to S3.

Do this using a Lambda.



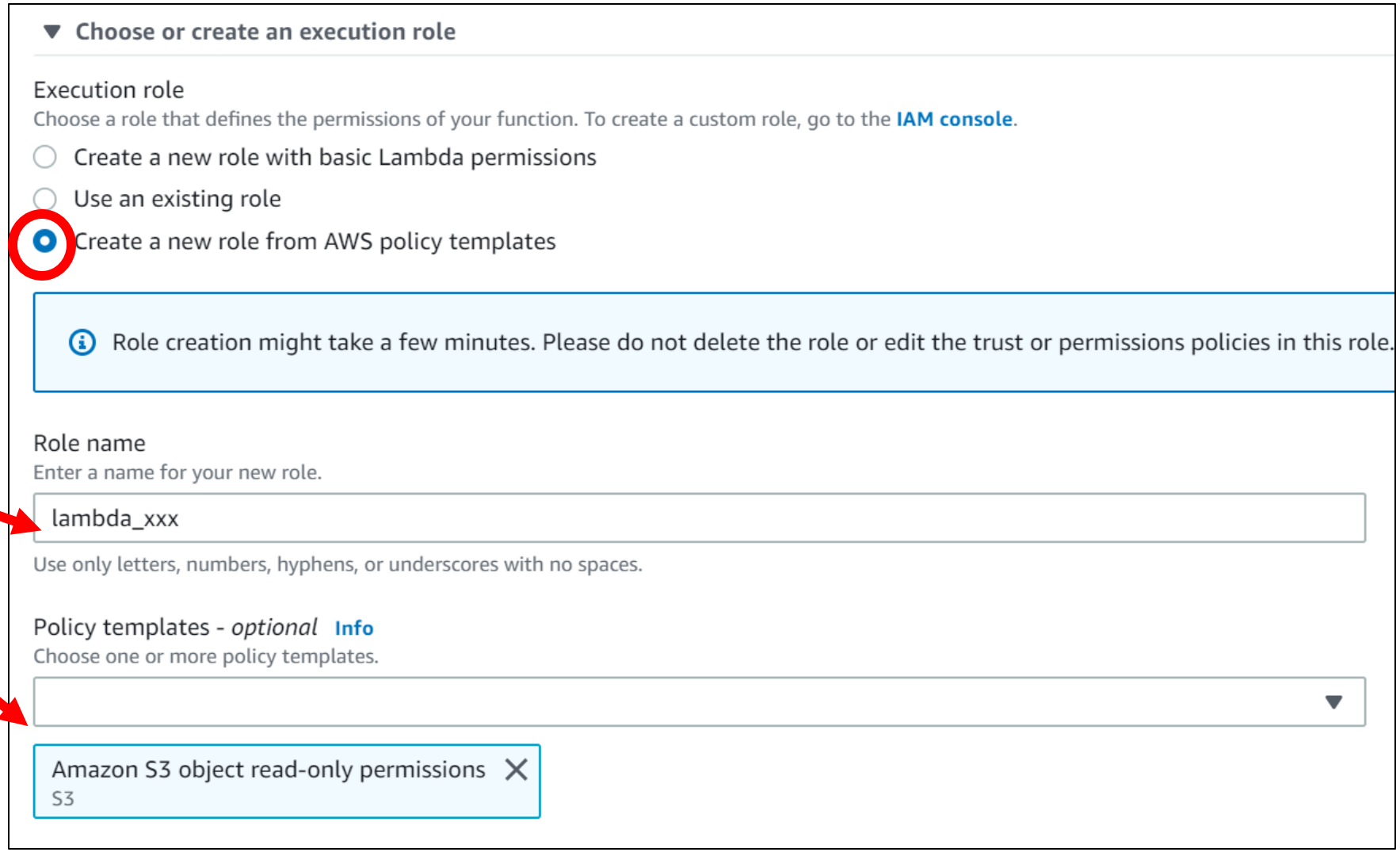


Name the lambda and choose Python as the language

Glue Trigger

Automatically run Glue job using Lambda – a serverless function

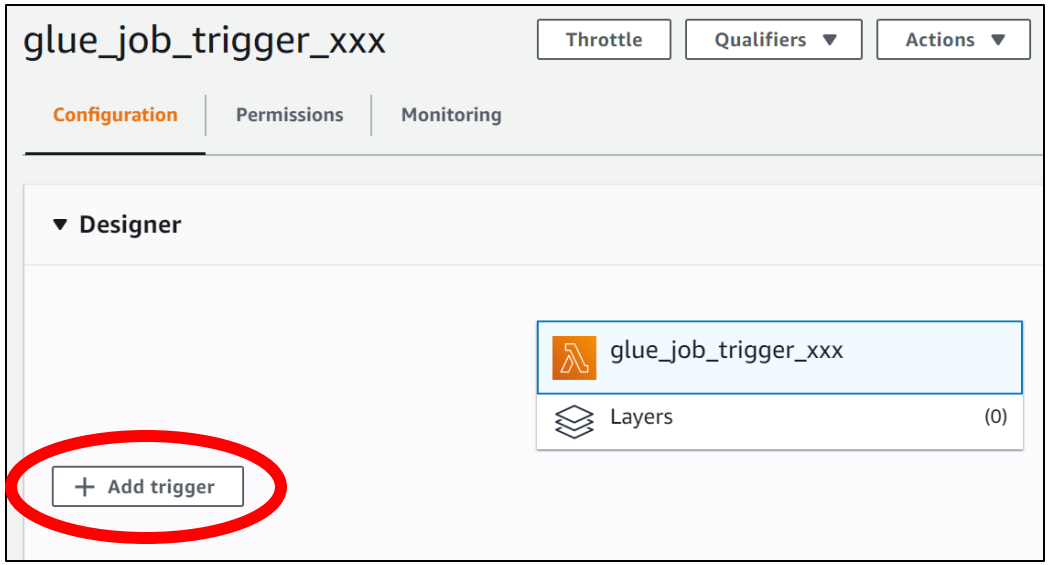
Name the role and choose permissions for S3

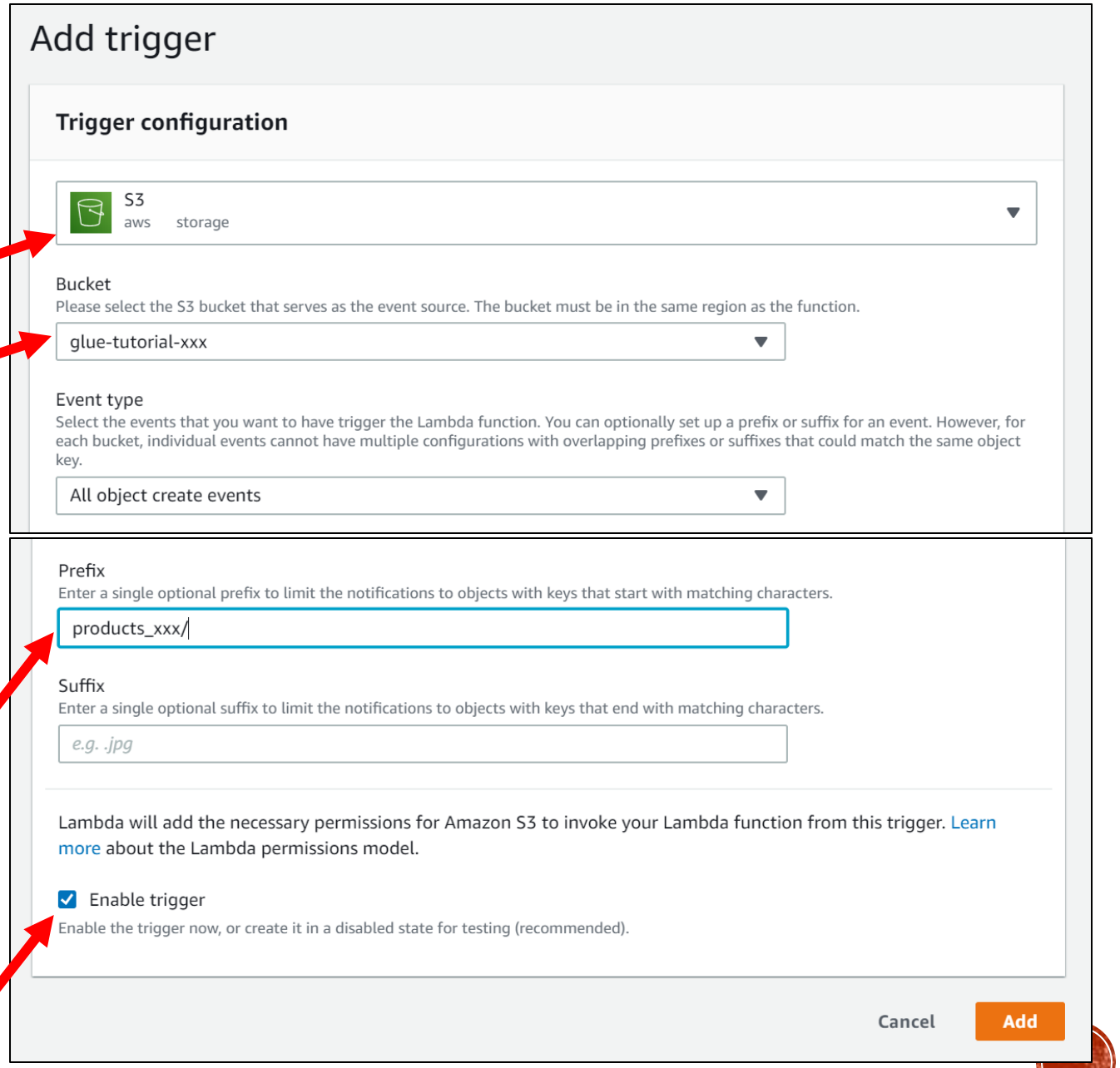


Glue Trigger

└─Automatically run Glue job using Lambda – a serverless function

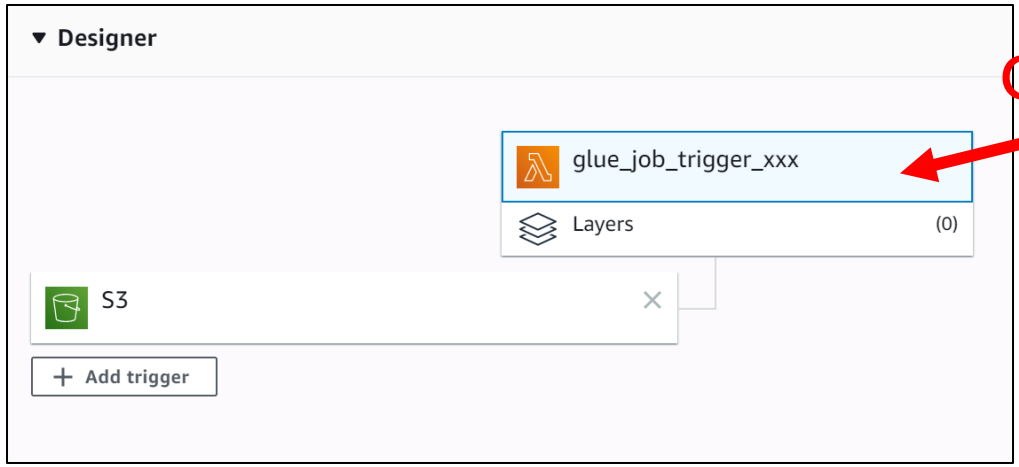
You can set a Lambda to run when a file lands in an S3 bucket





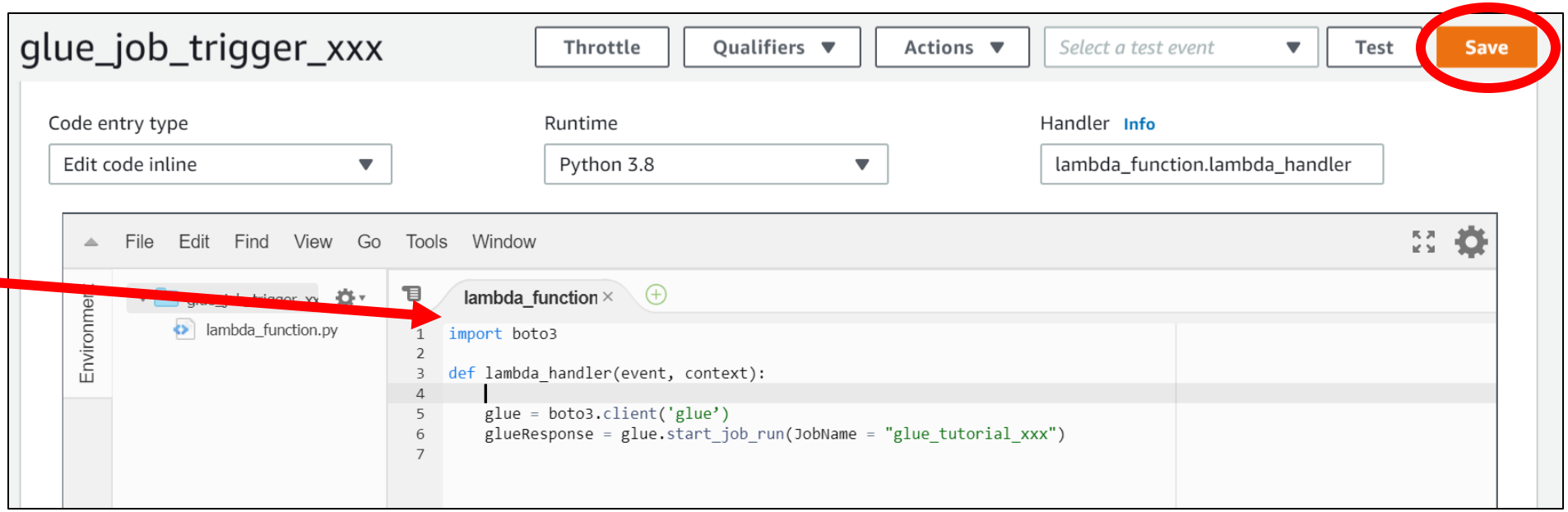
Glue Trigger

└Automatically run Glue job using Lambda – a serverless function



Then make the Lambda run the Glue job.

Add code under 'lambda_function'

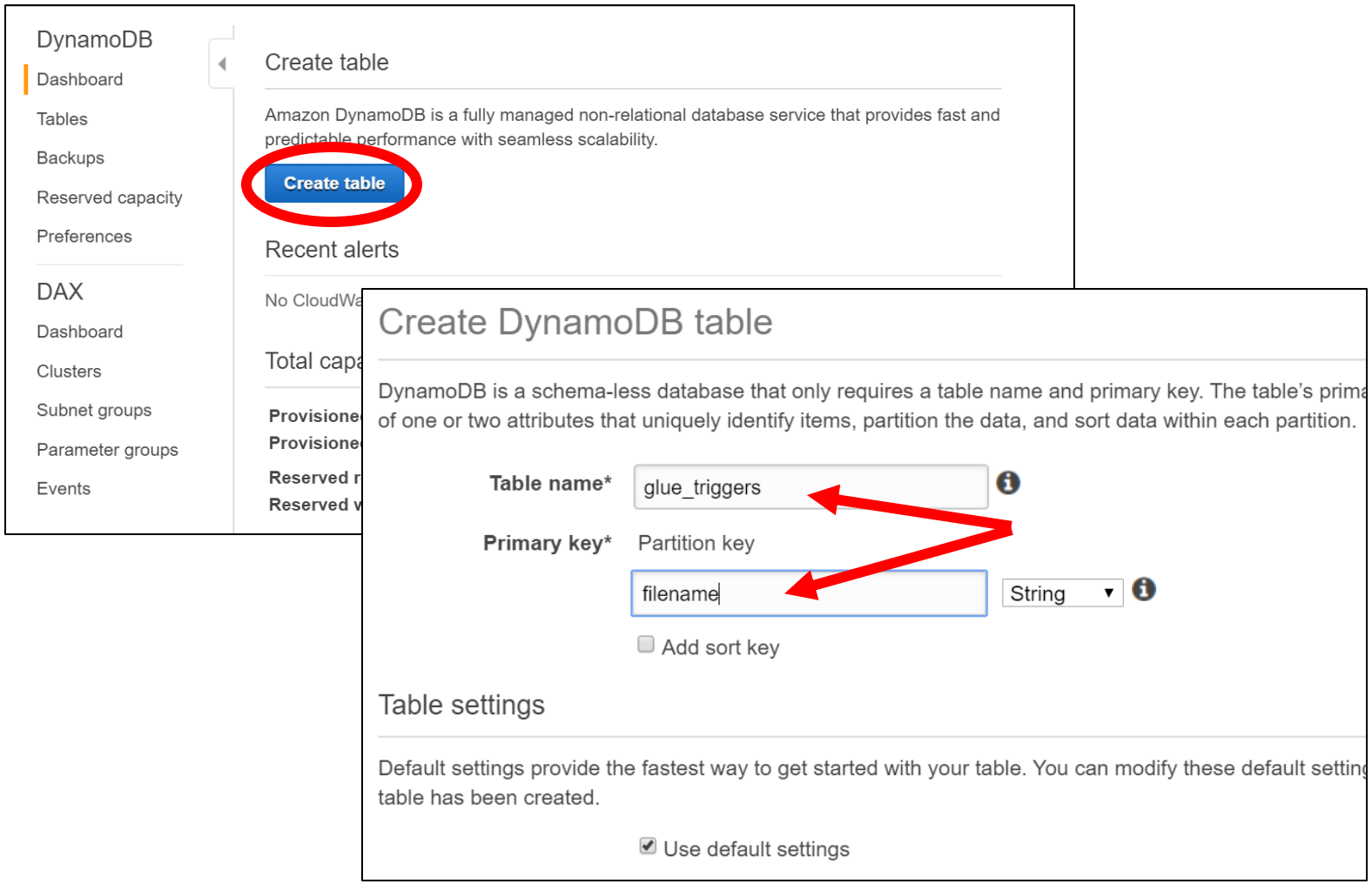



```
import boto3
def lambda_handler(event, context):
    glue = boto3.client('glue')
    glueResponse = glue.start_job_run(JobName = "glue_tutorial_xxx")
```

Glue Trigger

Run multiple different Glue jobs with DynamoDB – a non-relational database

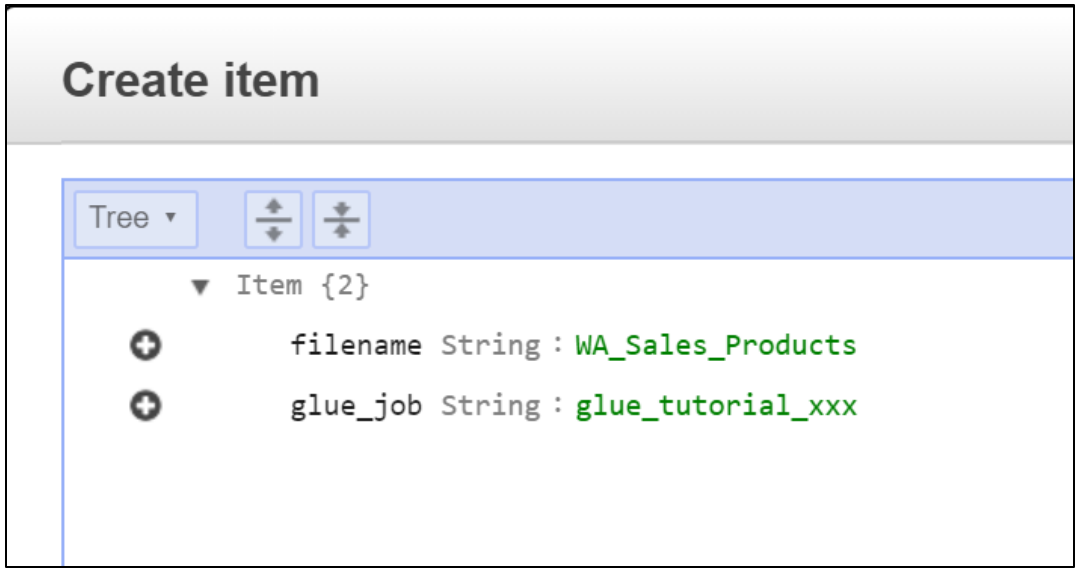
- The Lambda currently can only run one Glue job
- It would be better if it could run different Glue jobs based on the file.
- We could store that information in a DynamoDB table

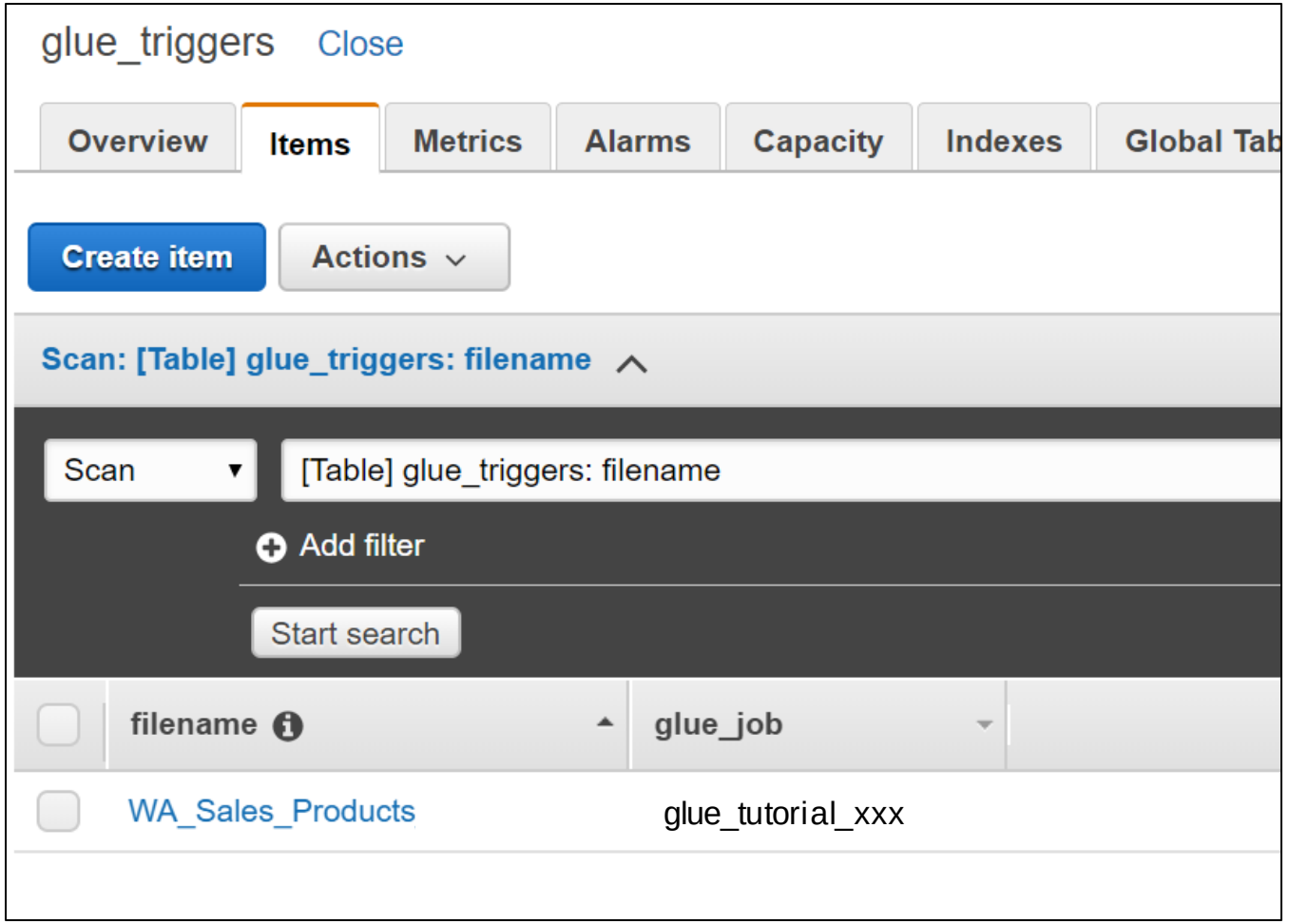


Glue Trigger

Run multiple different Glue jobs with DynamoDB – a non-relational database

- Enter the filename and the glue job that loads that file





Glue Trigger

Automatically run Glue job using Lambda

- The Lambda can look up the filename in the DynamoDB table to find which Glue job to run

Lambda receives an event from S3, which includes the 'key'

lambda_function

```
1  import boto3
2
3  def lambda_handler(event, context):
4
5      source_key_name = event['Records'][0]['s3']['object']['key']
6      filename = source_key_name.rsplit('/',1)[1].split('. ',1)[0]
7
8      dynamodb = boto3.resource('dynamodb')
9      table = dynamodb.Table('glue_triggers')
10
11     dynamodb_response = table.get_item(Key = {'filename': filename})
12     glue_job = dynamodb_response['Item']['glue_job']
13
14     glue = boto3.client('glue')
15     glue_response = glue.start_job_run(JobName = glue_job)
```

This returns the Glue job associated with that file, then we trigger that job

We get the filename from the key, then search the DynamoDB table with it

Glue Trigger

└─IAM Roles determine how a resource can interact with other services

Log output

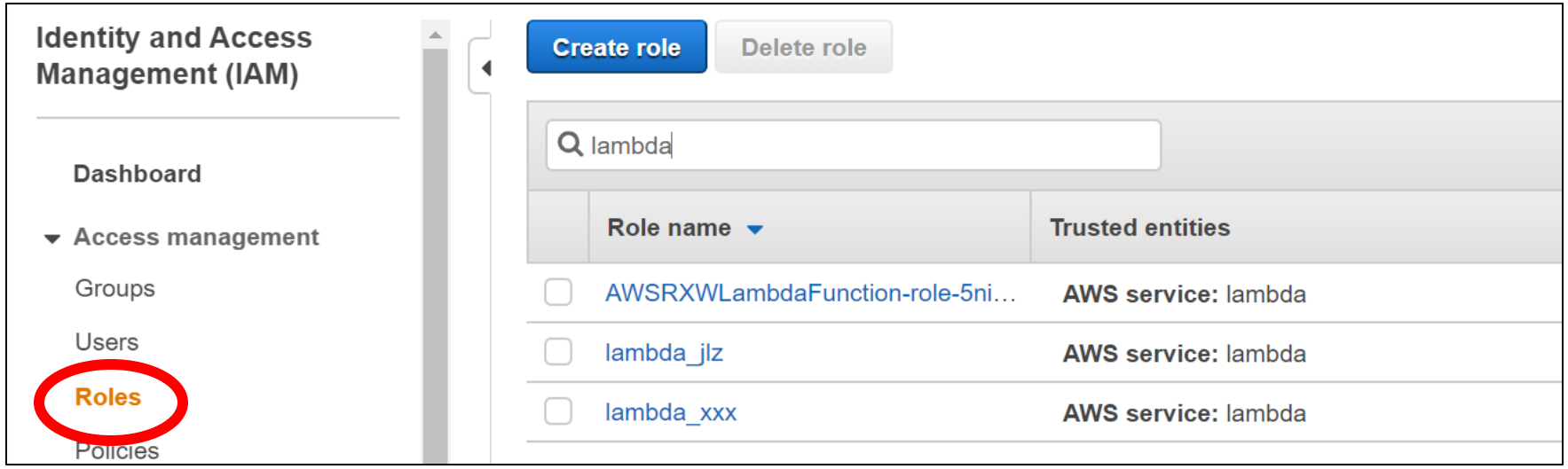
The area below shows the logging calls in your code. These correspond to a single row within the CloudWatch log group corresponding to this Lambda function. Click here to view the CloudWatch log group.

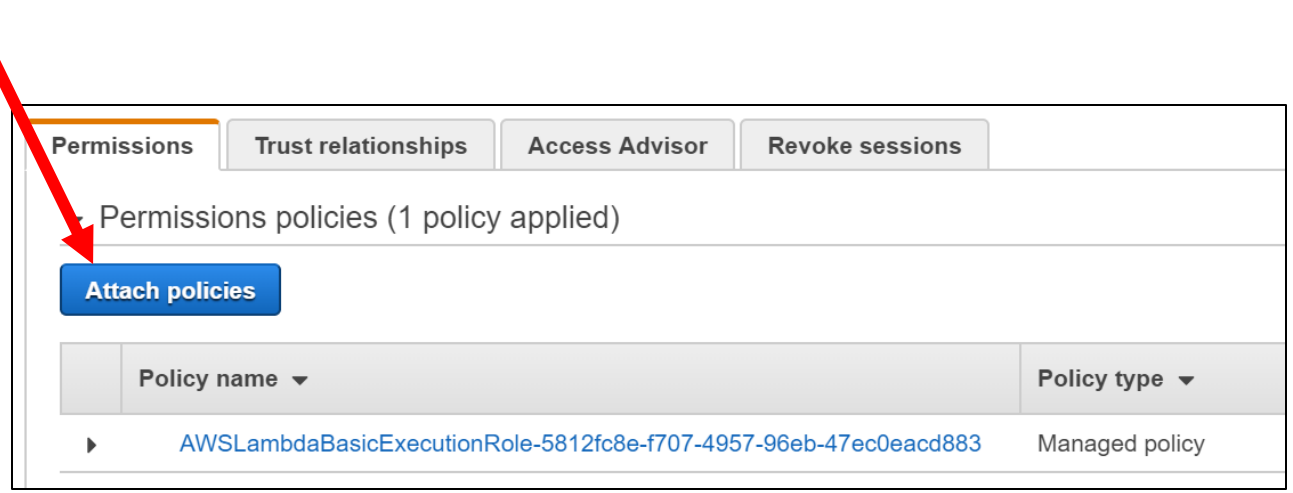
```
START RequestId: 2df6f8a8-95cb-11e8-aedb-510d0136df8b Version: $LATEST
An error occurred (AccessDeniedException) when calling the GetItem operation: User: arn:aws:sts::952552944372:assumed-
role/lambda_basic_execution/glue_job_trigger is not authorized to perform: dynamodb:GetItem on resource: arn:aws:dynamodb:us-east-
1:952552944372:table/glue_triggers: ClientError
Traceback (most recent call last):
```

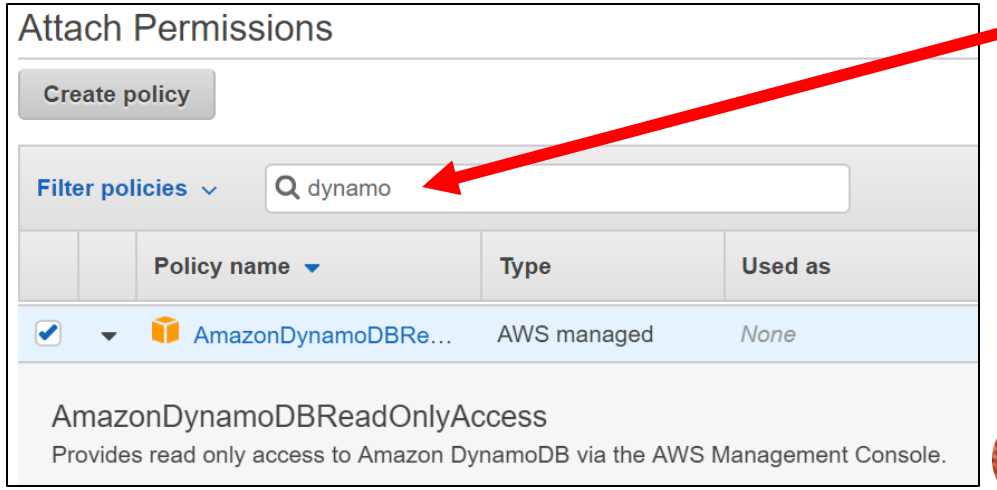
- However when we try to run this lambda, we get an AccessDeniedException
- We need to add permission to the Lambda's IAM Role to access DynamoDB and Glue

Glue Trigger

└─IAM Roles determine how a resource can interact with other services

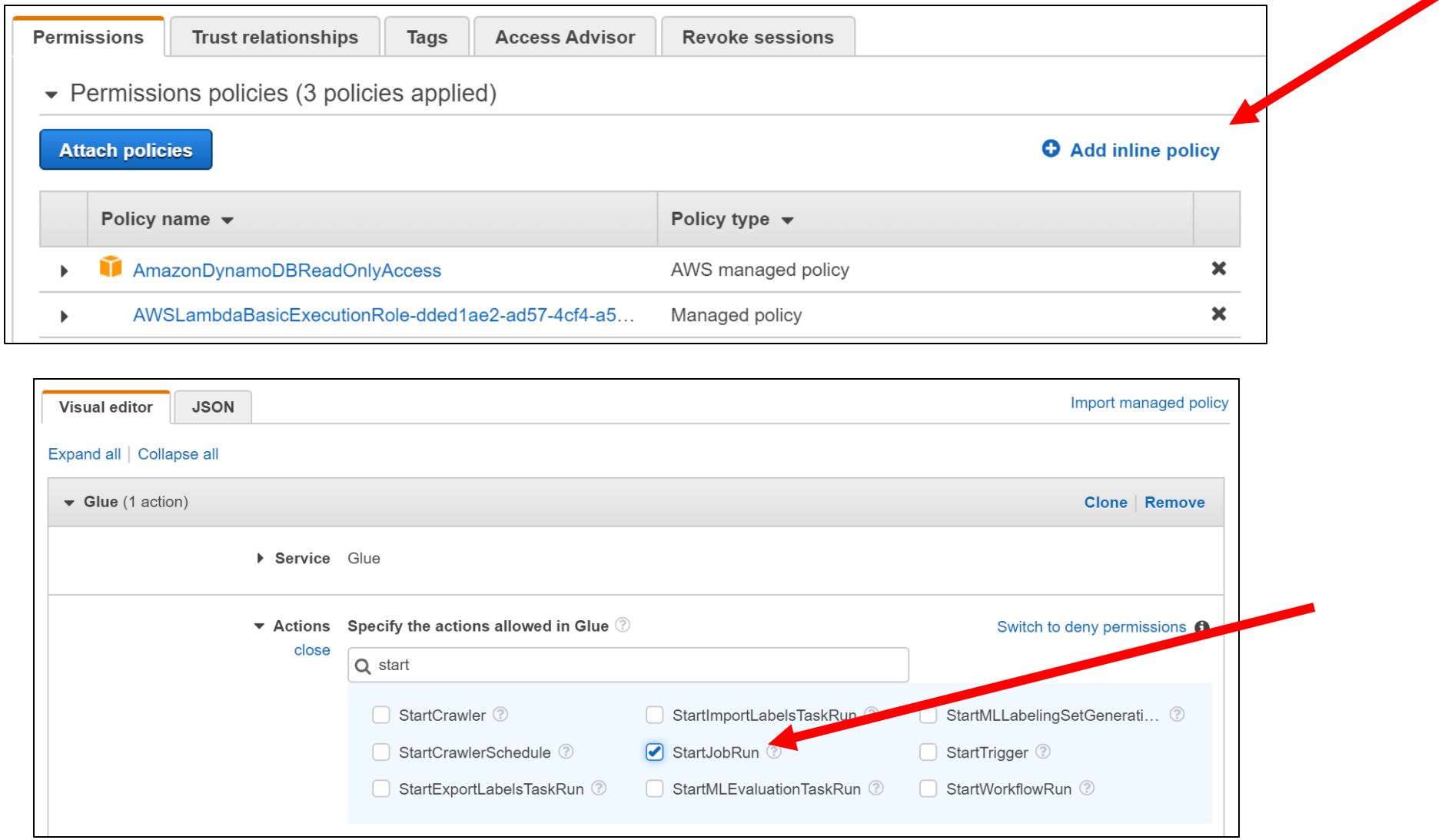






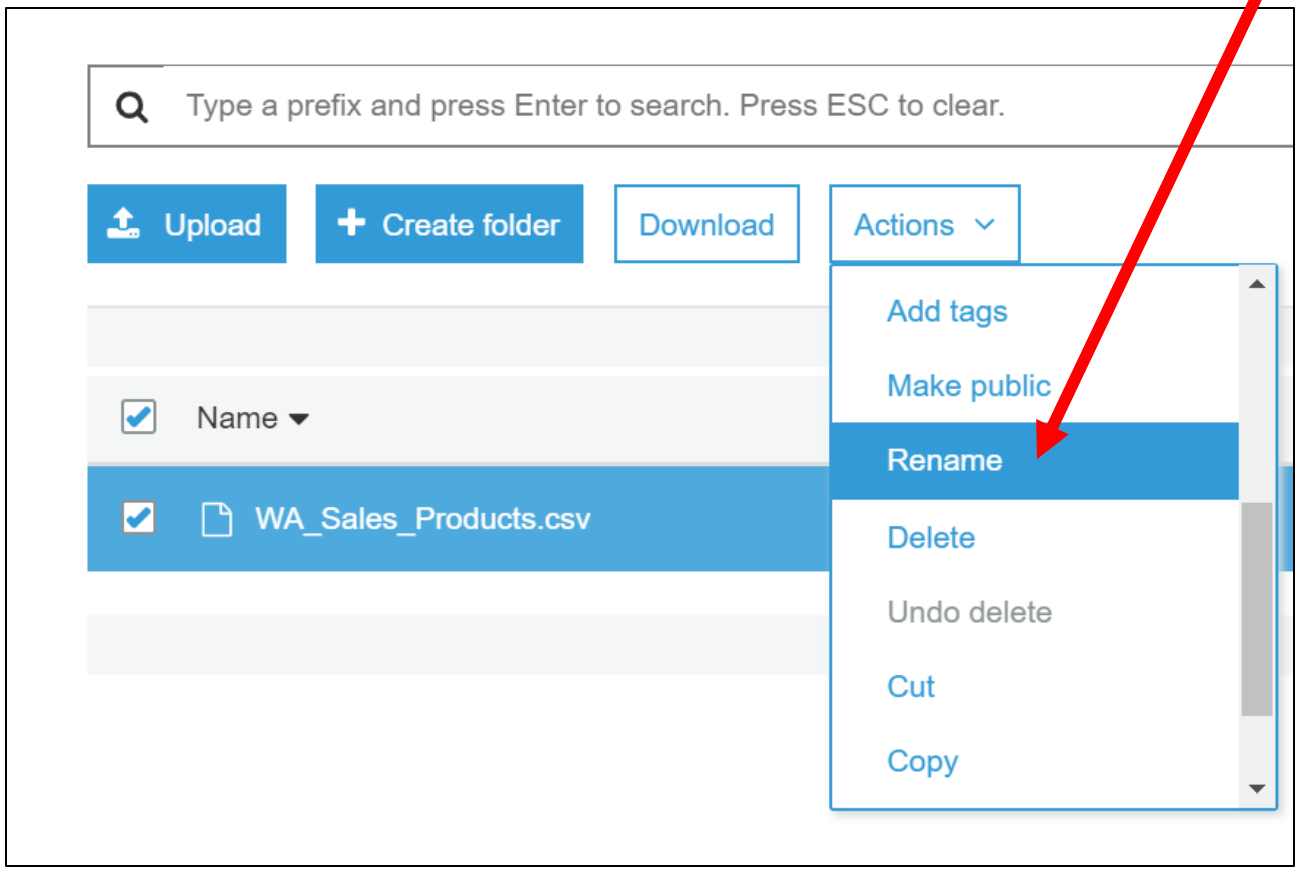
Glue Trigger

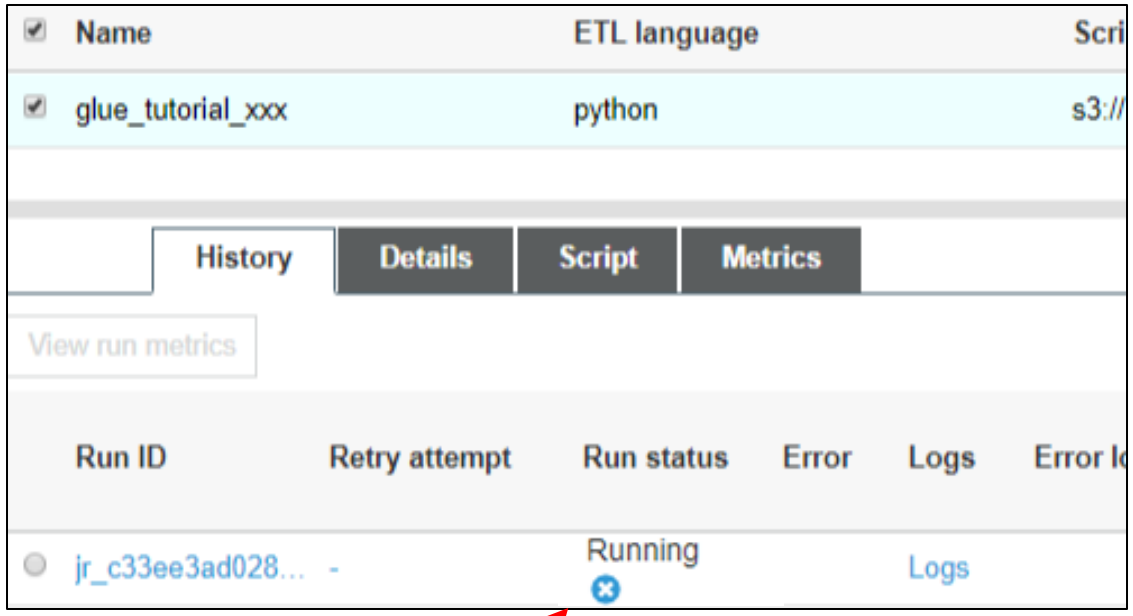
IAM Roles determine how a resource can interact with other services



Glue Trigger

Test the Lambda





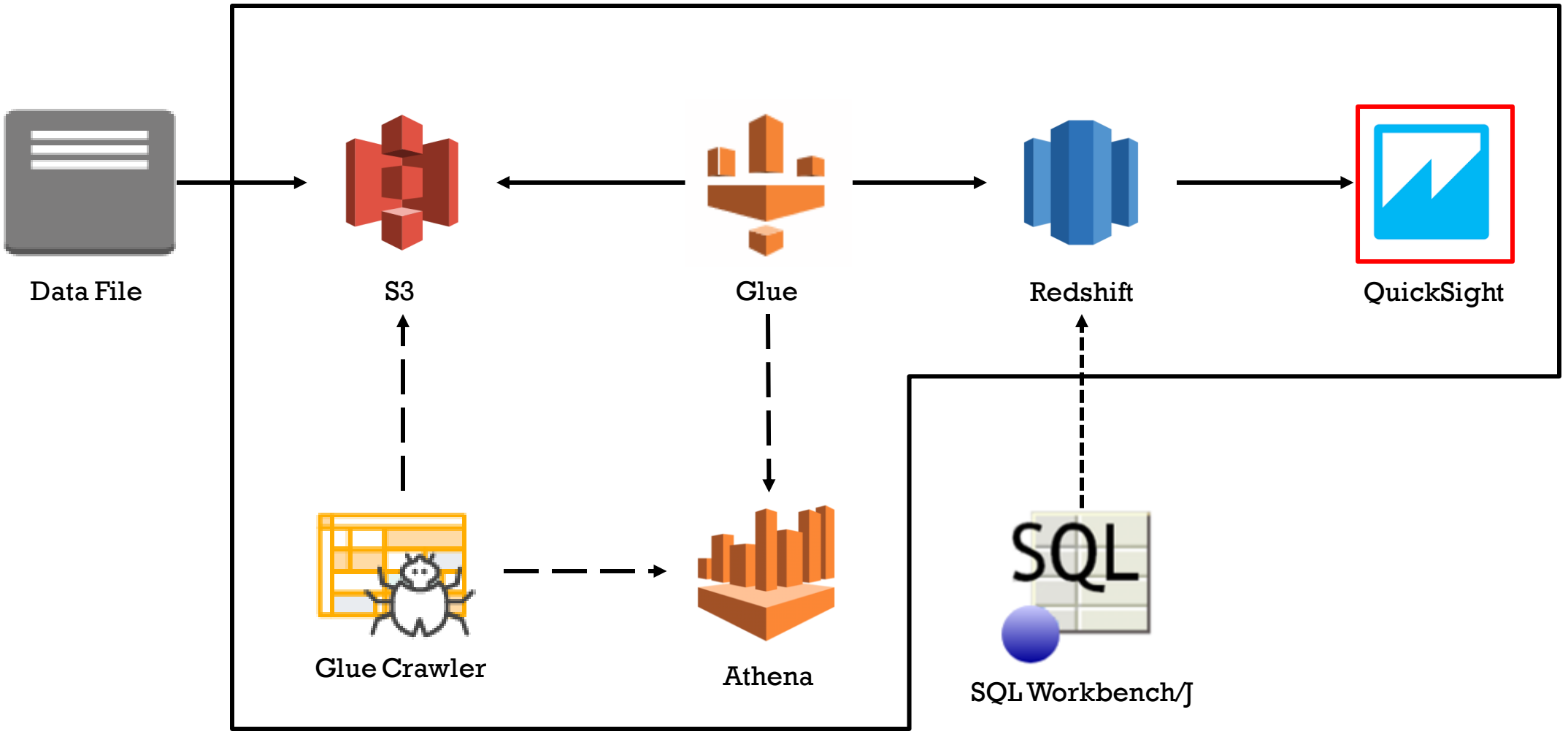
Lab 6

- Create Lambda
- Create DynamoDB
- Edit IAM Role
- Test Lambda

(Use US-EAST-2/Ohio Region)

QUICKSIGHT

AWS Business Intelligence Tool



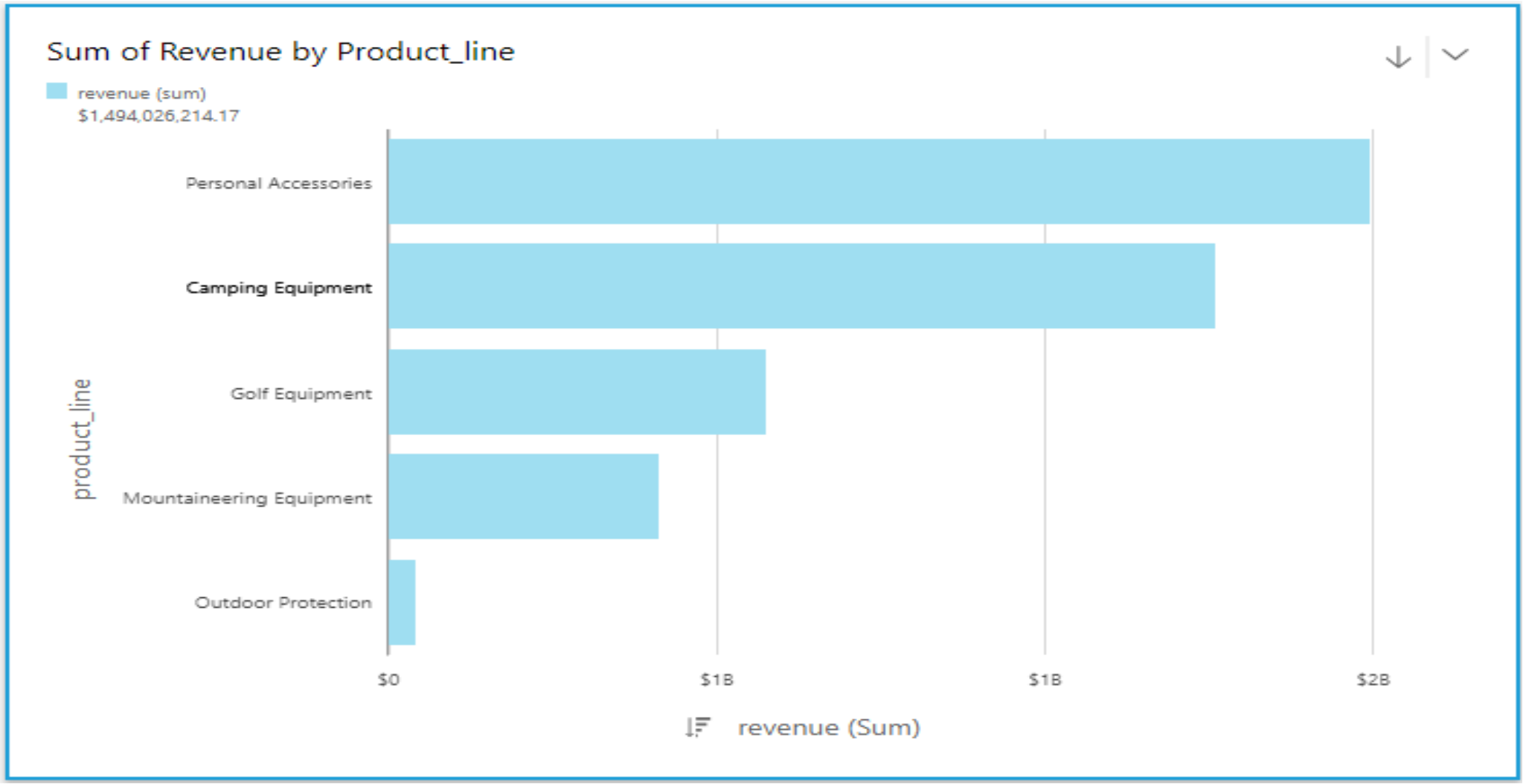
QUICKSIGHT

AWS Business Intelligence Tool

- Cloud based Business Intelligence reporting tool
- Build Reports from
  - Files in S3
  - Redshift
  - Athena

QUICKSIGHT

AWS Business Intelligence Tool



QUICKSIGHT

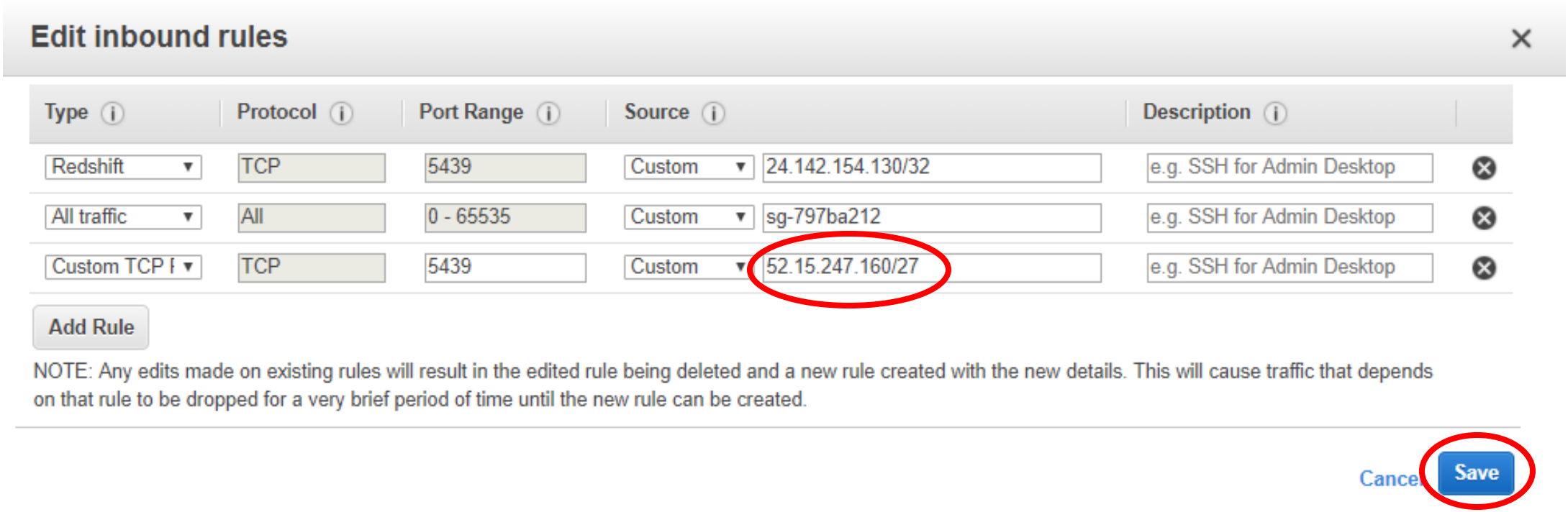
AWS Business Intelligence Tool

Create Analysis

1. Create/Select data set
2. Select fields
3. Set field format
4. Add drill down layer
5. Select/change visual type
6. Publish to the dashboard

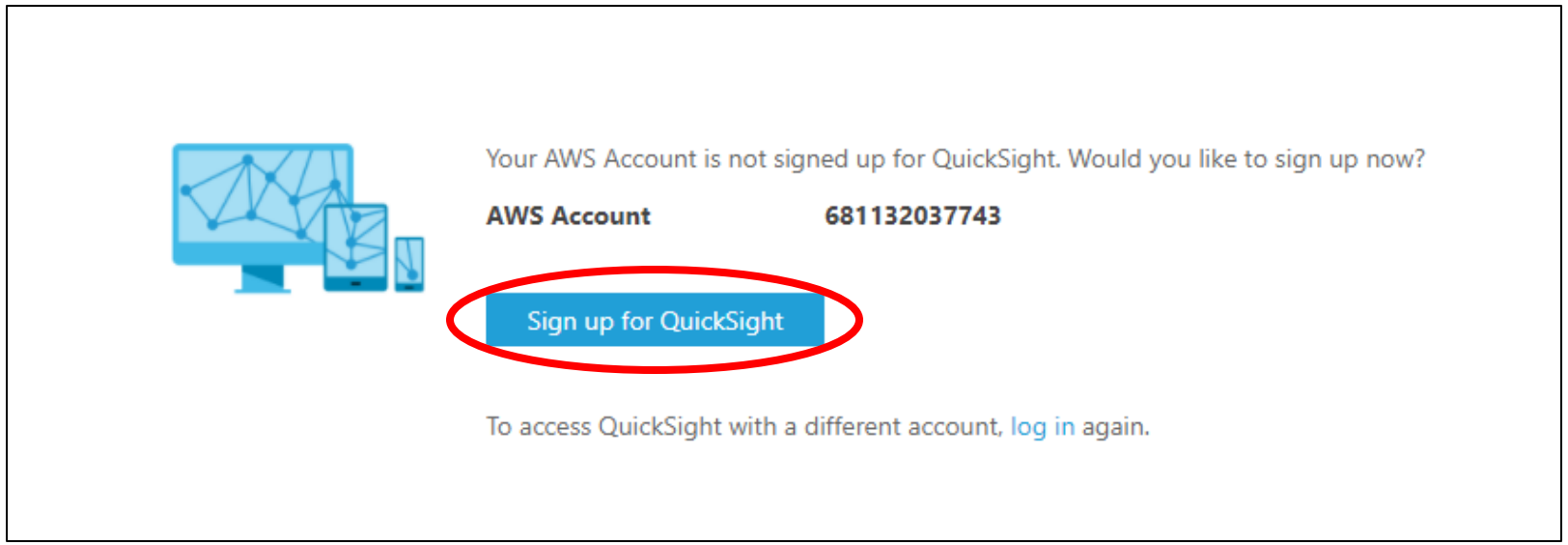
QUICKSIGHT

AWS Business Intelligence Tool



QUICKSIGHT

AWS Business Intelligence Tool



QUICKSIGHT

AWS Business Intelligence Tool

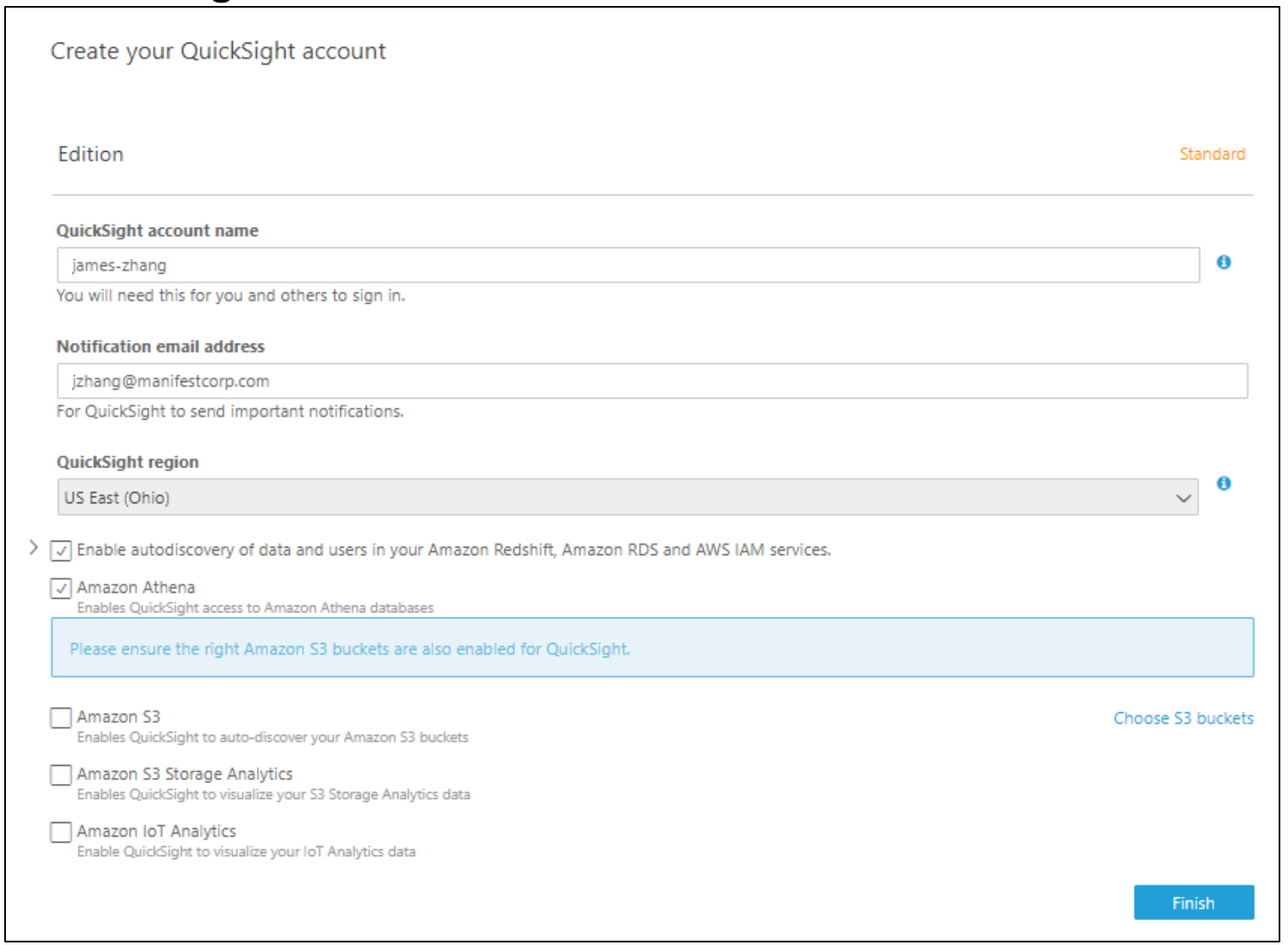
First author with 1GB SPICE	**FREE**	**FREE**
Team trial for 60 days (4 authors)*	**FREE**	**FREE**
Additional author per month (yearly)**	**$9**	**$18**
Additional author per month (monthly)**	**$12**	**$24**
Additional readers (Pay-per-Session)	**N/A**	**$0.30/session (max $5/reader/month) ******
Additional SPICE per month	**$0.25 per GB**	**$0.38 per GB**
Single Sign On with SAML or OpenID Connect	✓	✓
Connect to spreadsheets, databases & business apps	✓	✓
Access data in Private VPCs		✓
Row-level security for dashboards		✓
Hourly refresh of SPICE data		✓
Secure data encryption at rest		✓
Connect to your Active Directory		✓
Use Active Directory Groups ***		✓

* Trial authors are auto-converted to month-to-month subscription upon trial expiry
** Each additional author includes 10GB of SPICE capacity
*** Active Directory groups are available in accounts connected to Active Directory
**** Sessions of 30-minute duration. Total charges for each reader are capped at $5 per month. Conditions apply

Continue

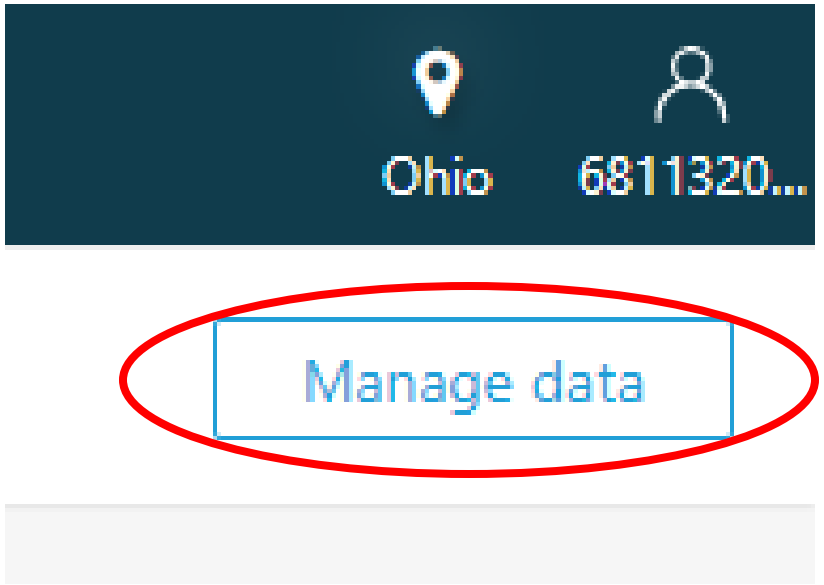
QUICKSIGHT

AWS Business Intelligence Tool



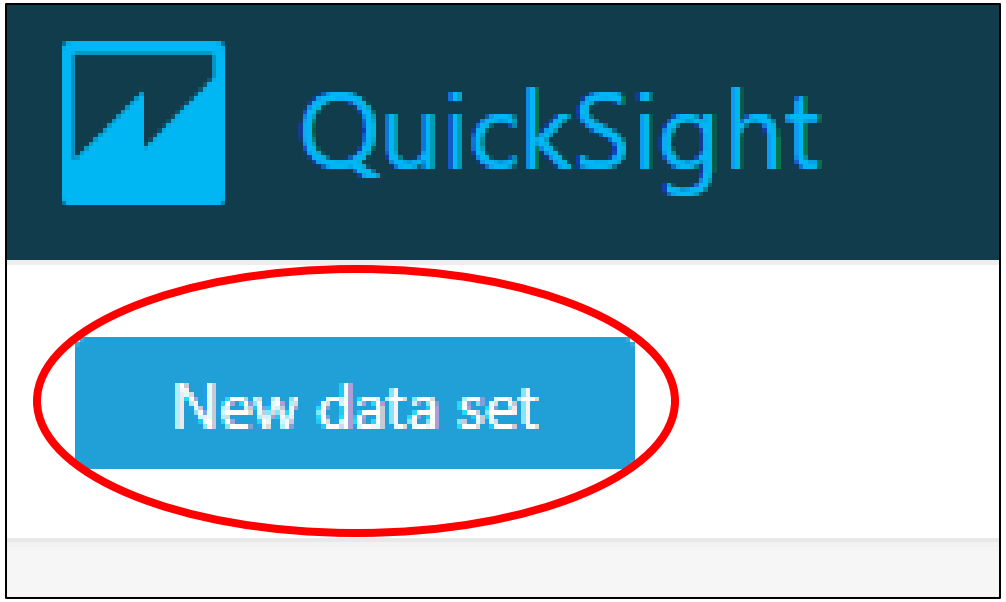
QUICKSIGHT

AWS Business Intelligence Tool



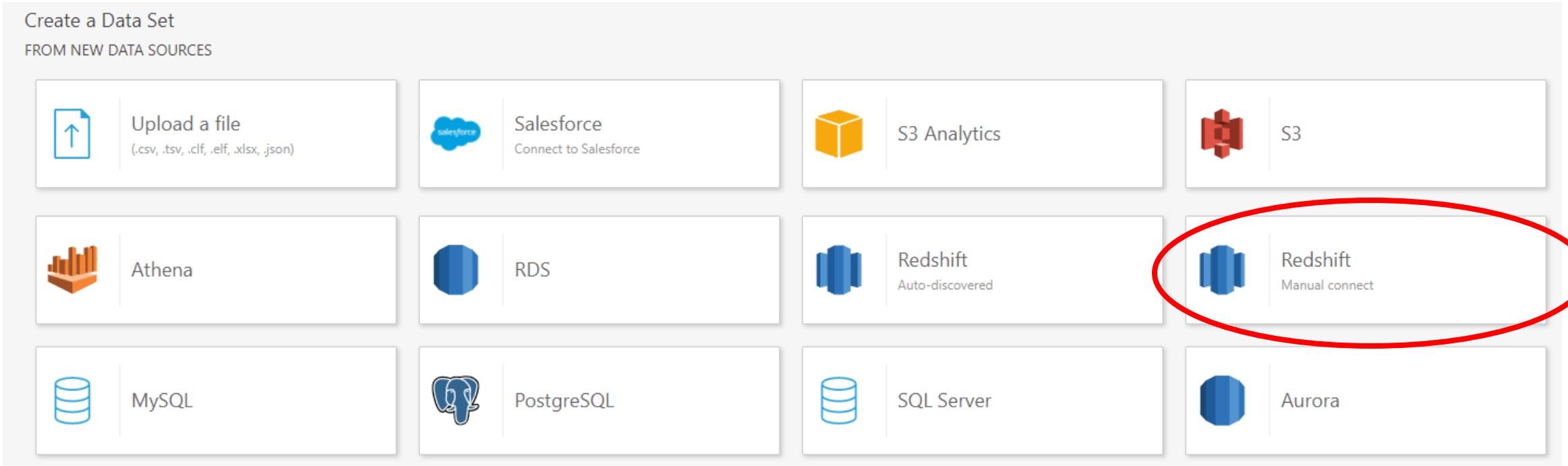
QUICKSIGHT

AWS Business Intelligence Tool



QUICKSIGHT

AWS Business Intelligence Tool



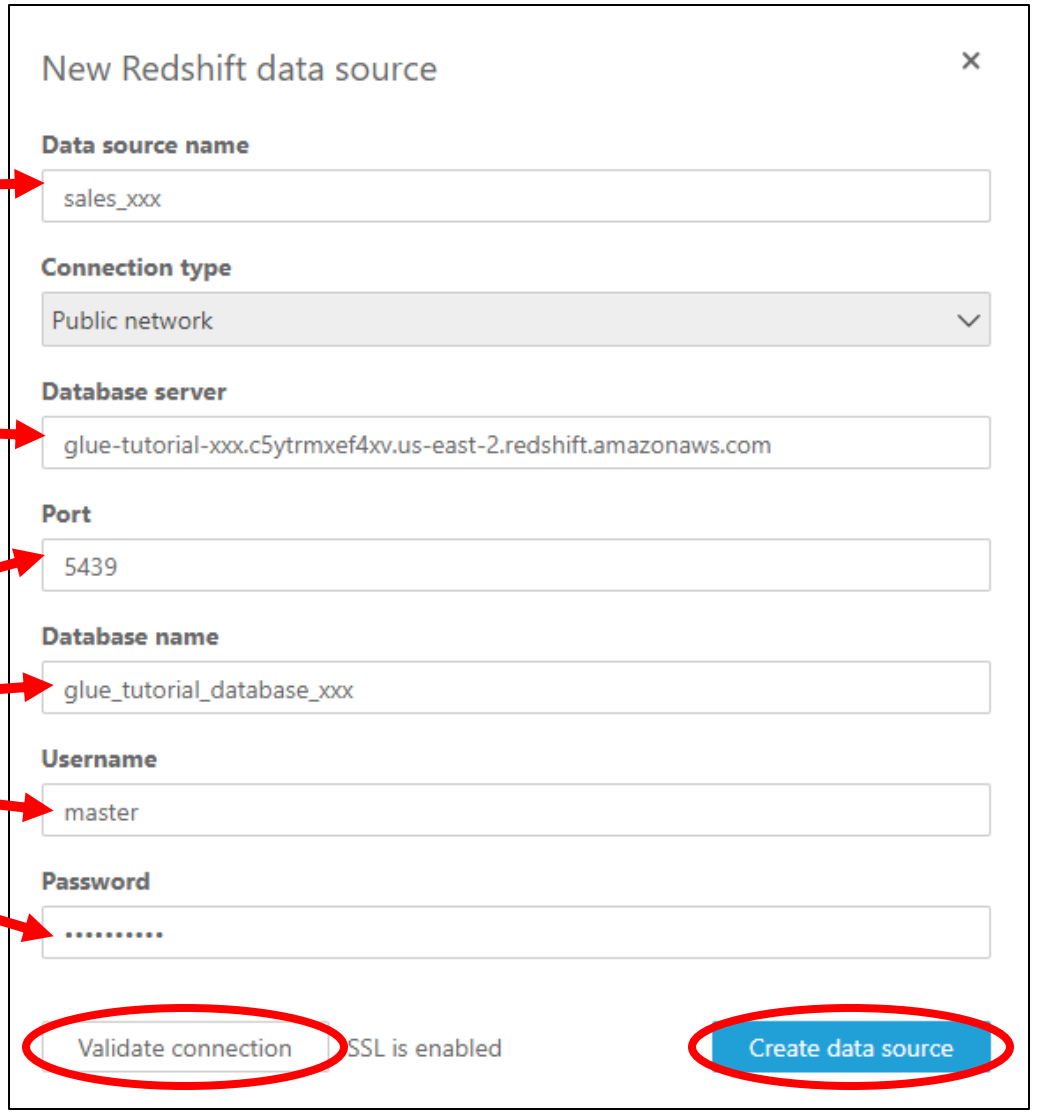
QUICKSIGHT

AWS Business Intelligence Tool

Give your data source a name

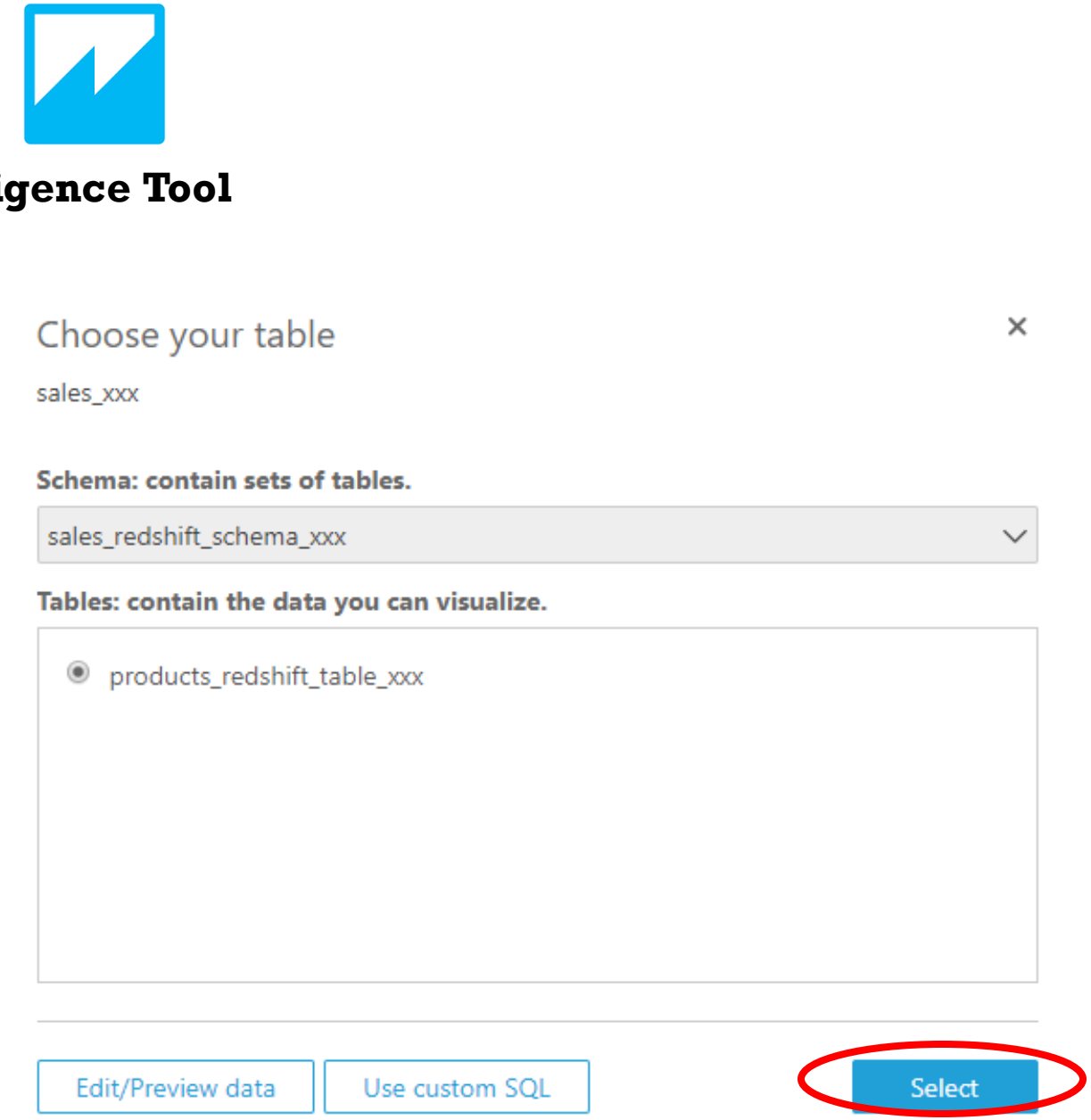
This is the Redshift endpoint without port number

This information comes from the Redshift Cluster



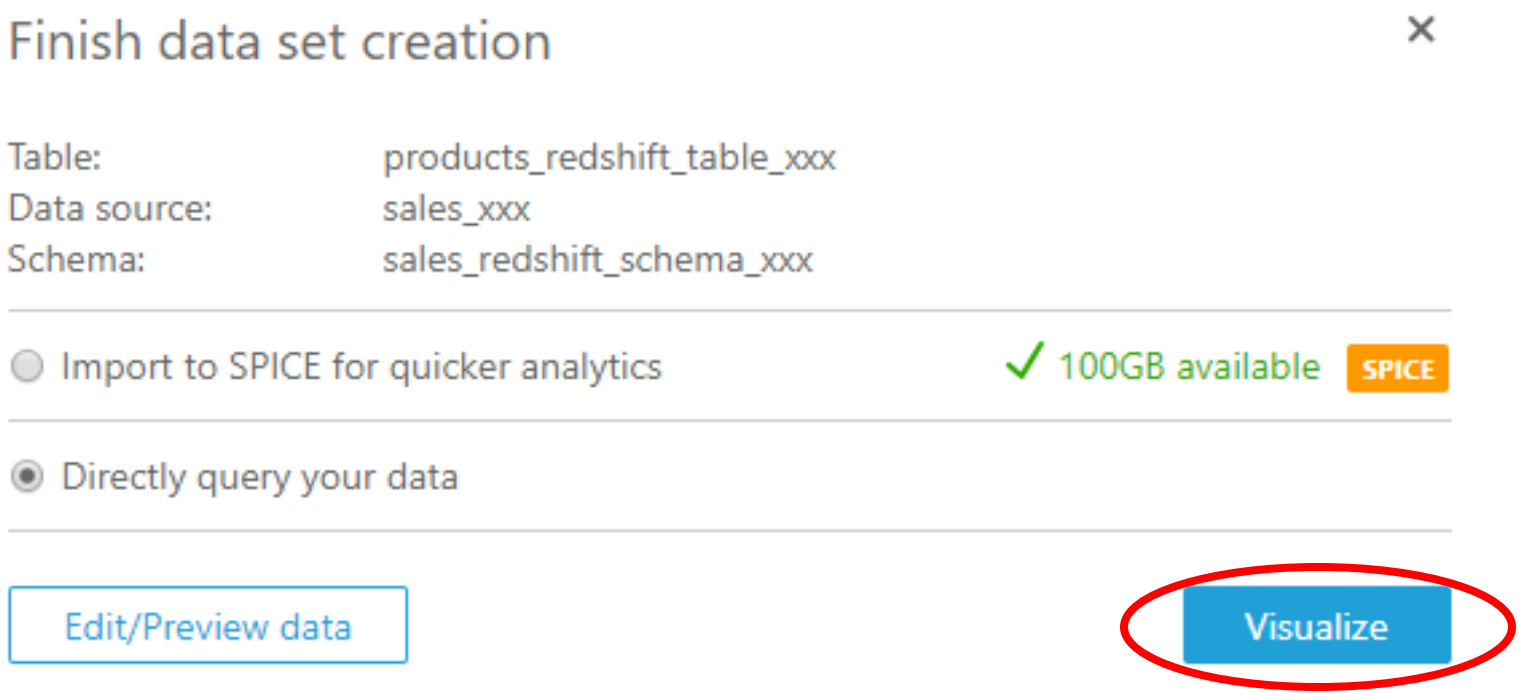
QUICKSIGHT

AWS Business Intelligence Tool



QUICKSIGHT

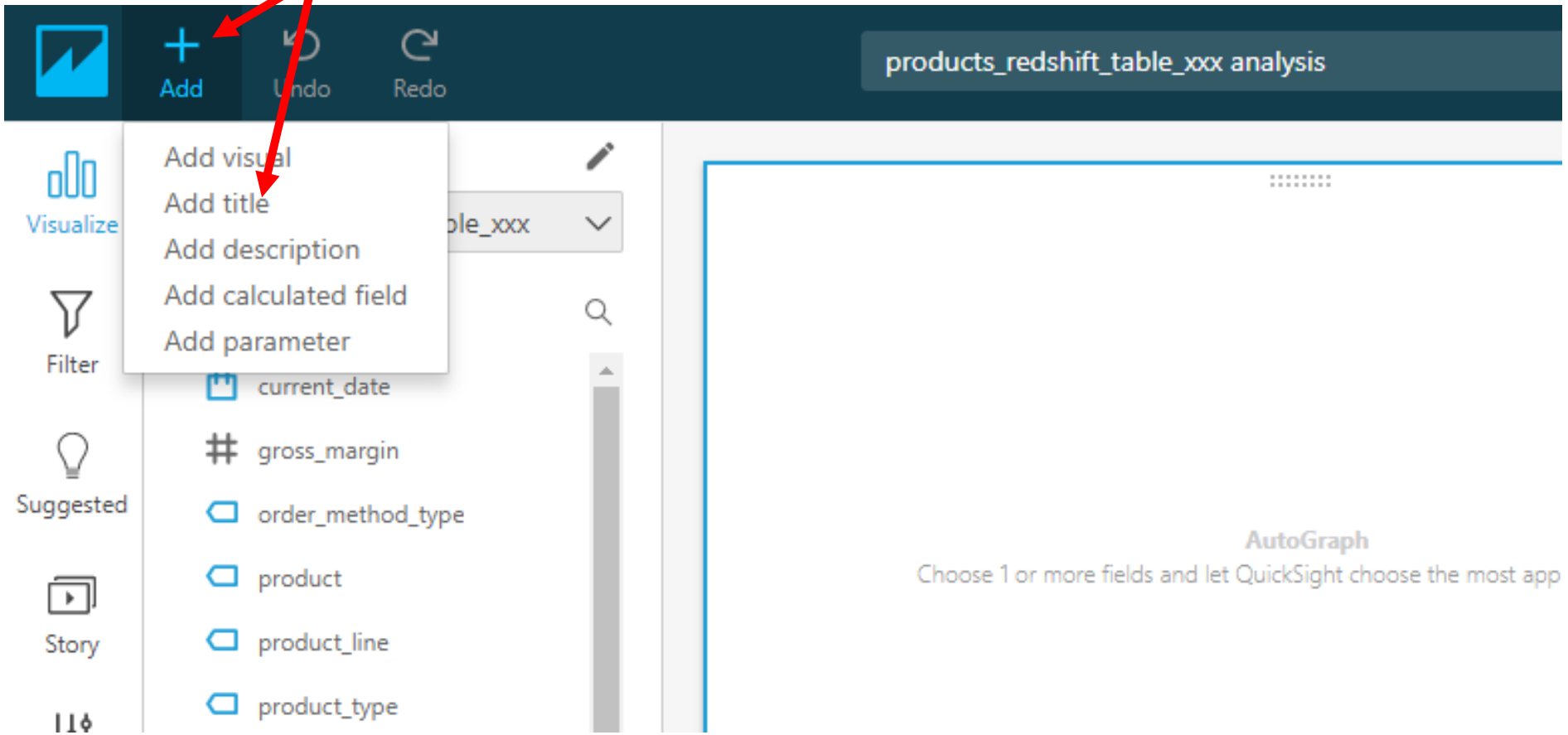
AWS Business Intelligence Tool



QUICKSIGHT

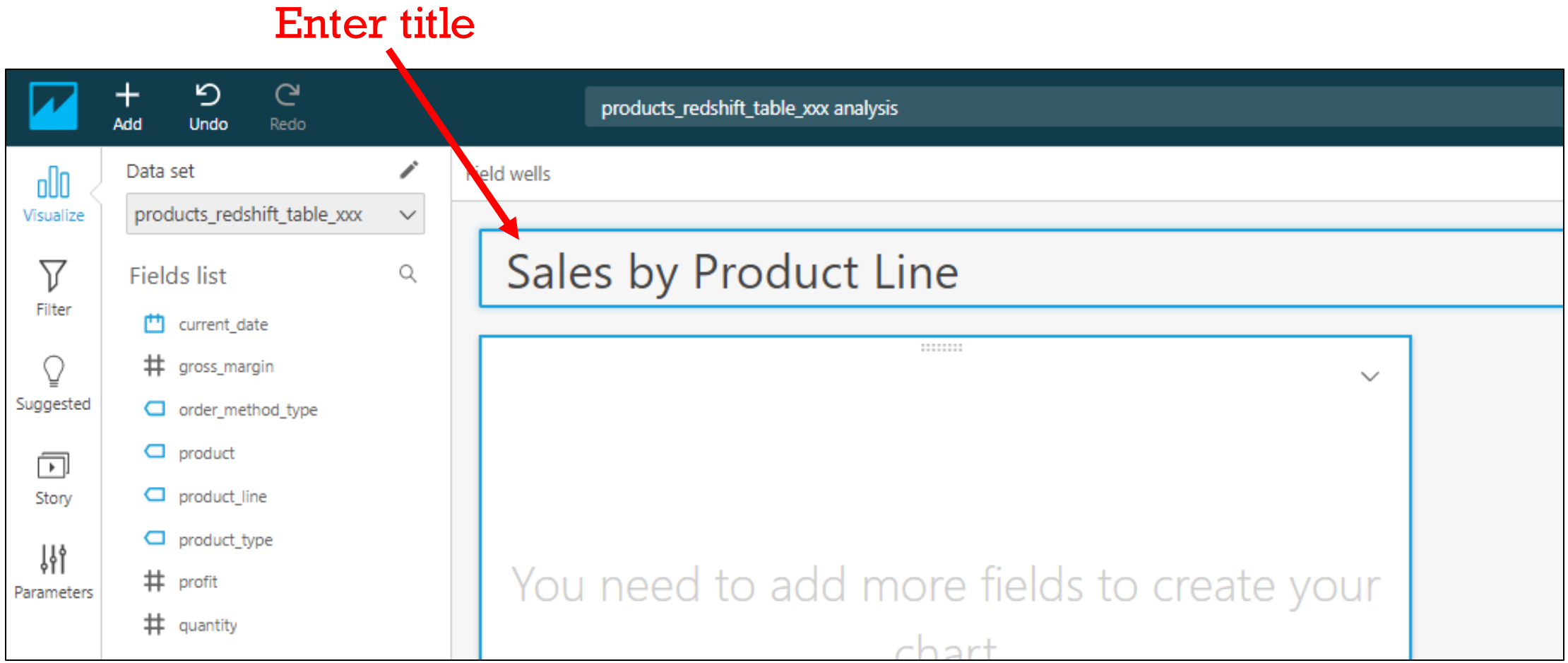
AWS Business Intelligence Tool

Select Add > Add title



QUICKSIGHT

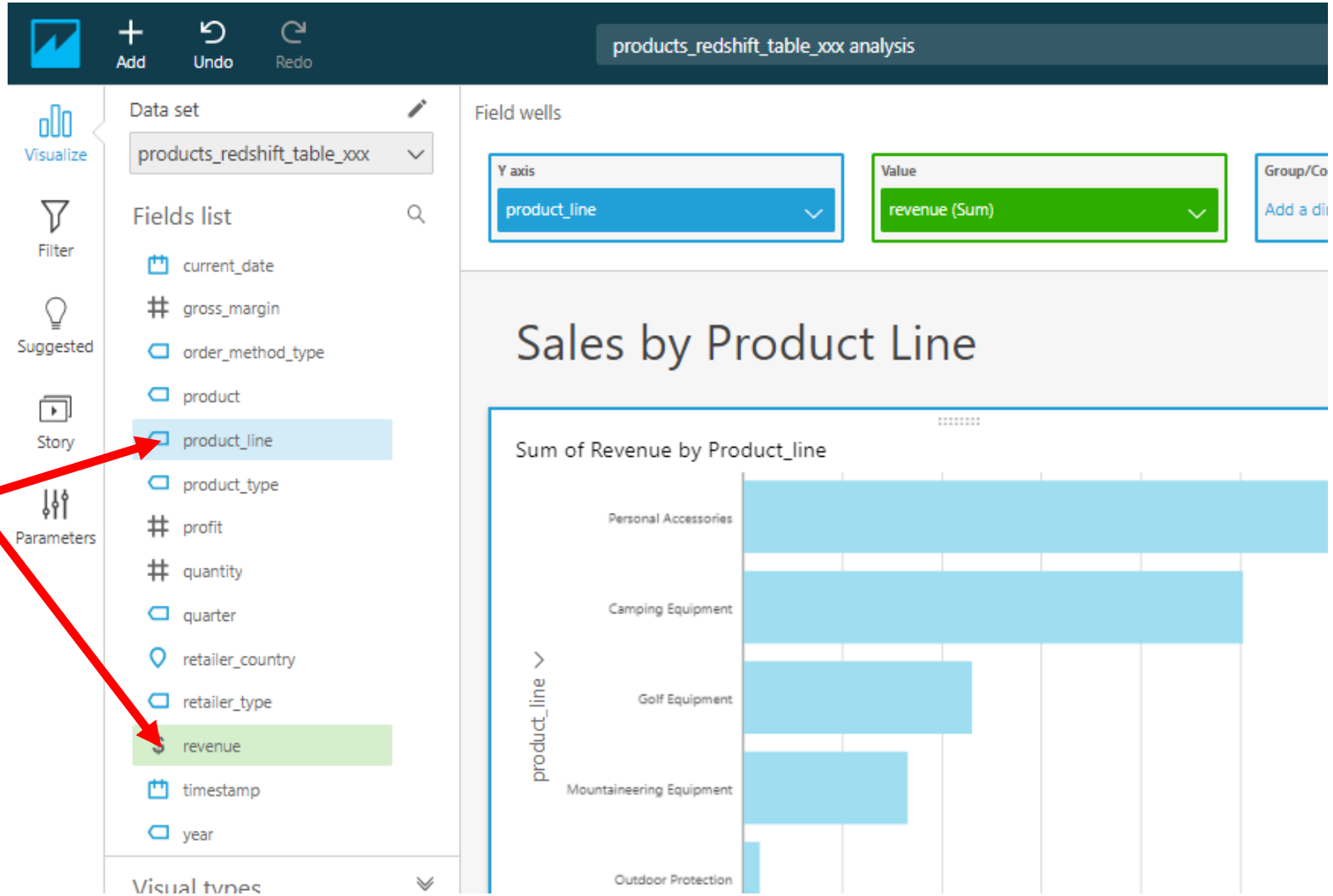
AWS Business Intelligence Tool



QUICKSIGHT

AWS Business Intelligence Tool

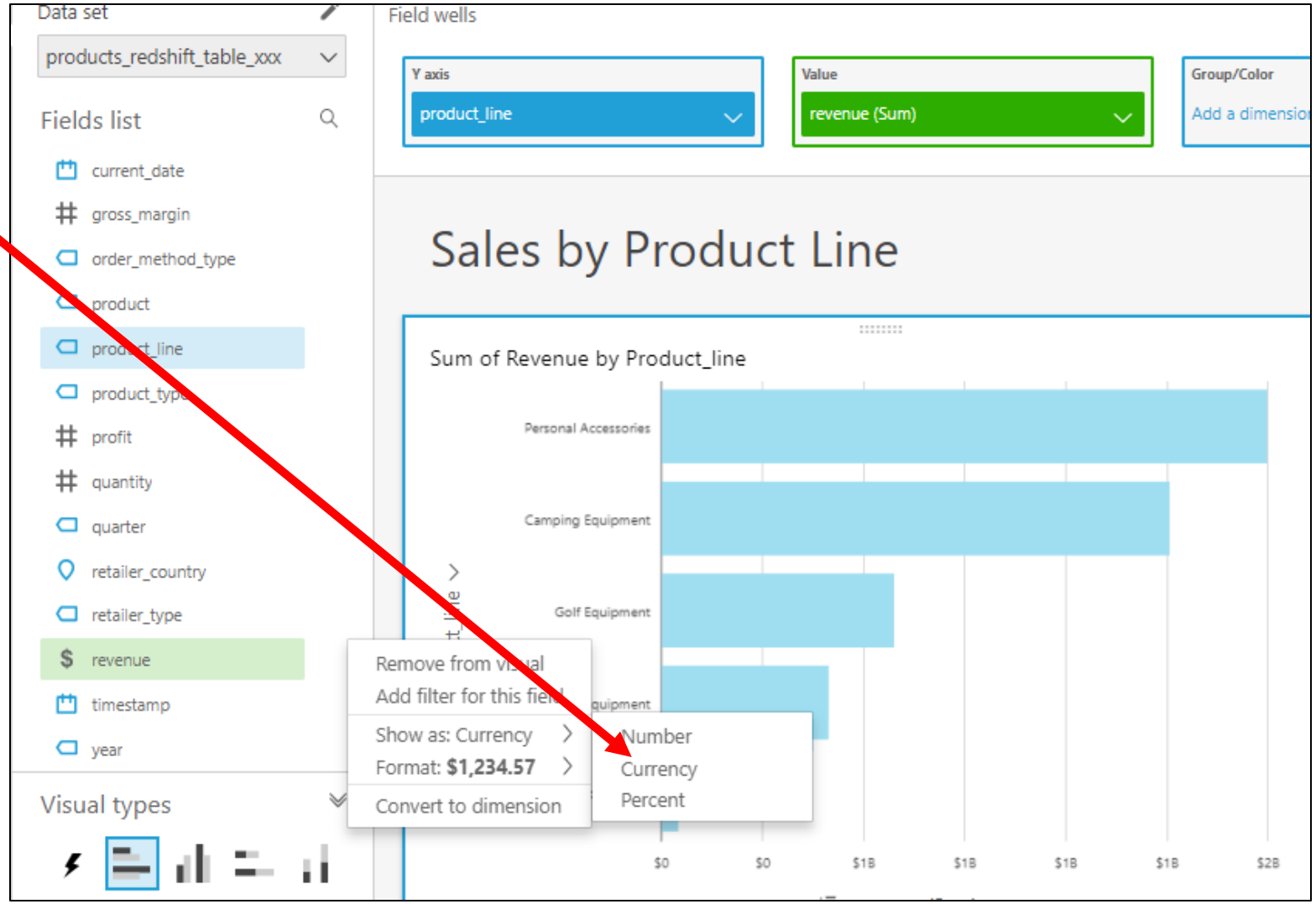
Choose product_line and revenue



QUICKSIGHT

AWS Business Intelligence Tool

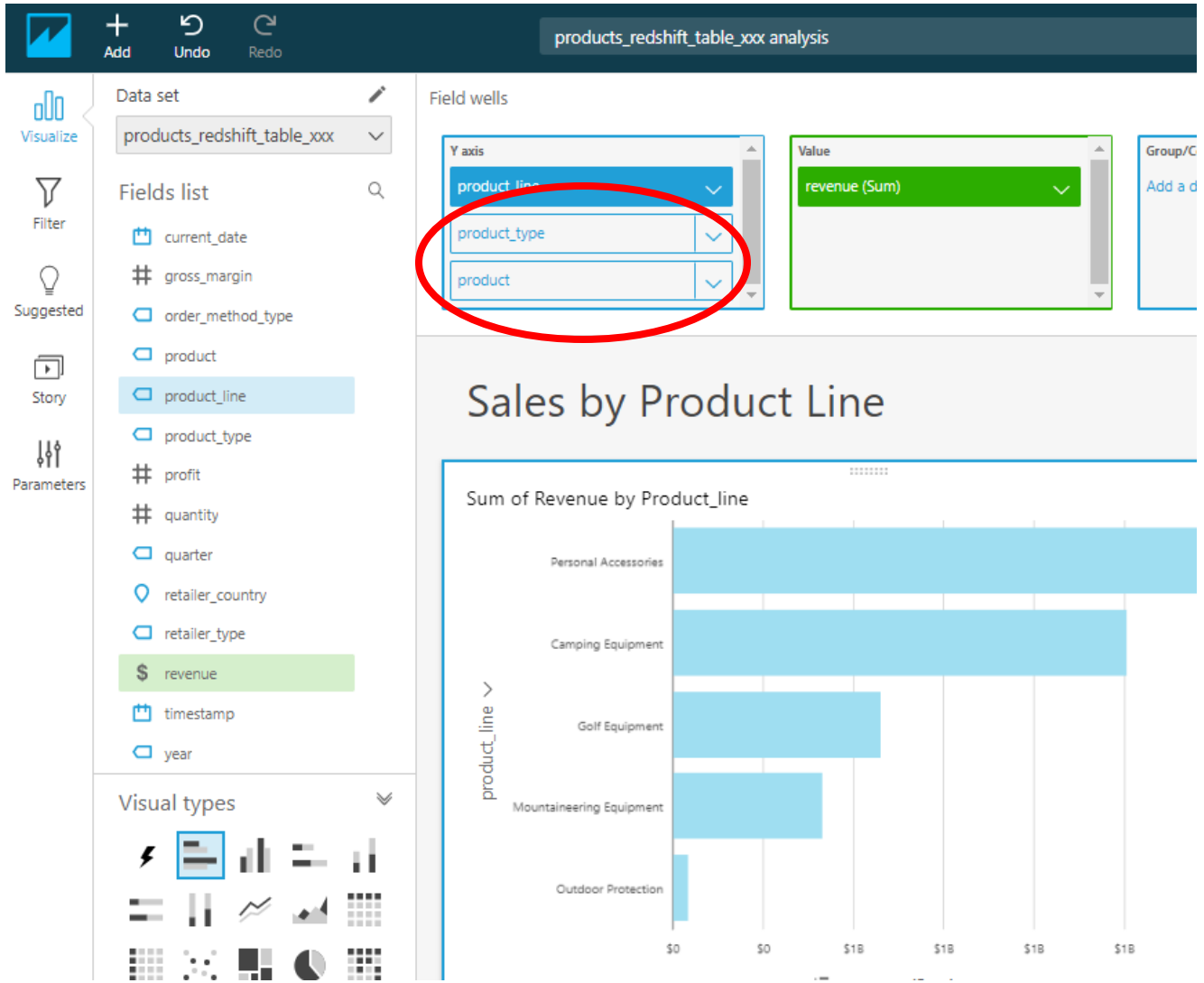
Change the format of Revenue to Currency



QUICKSIGHT

AWS Business Intelligence Tool

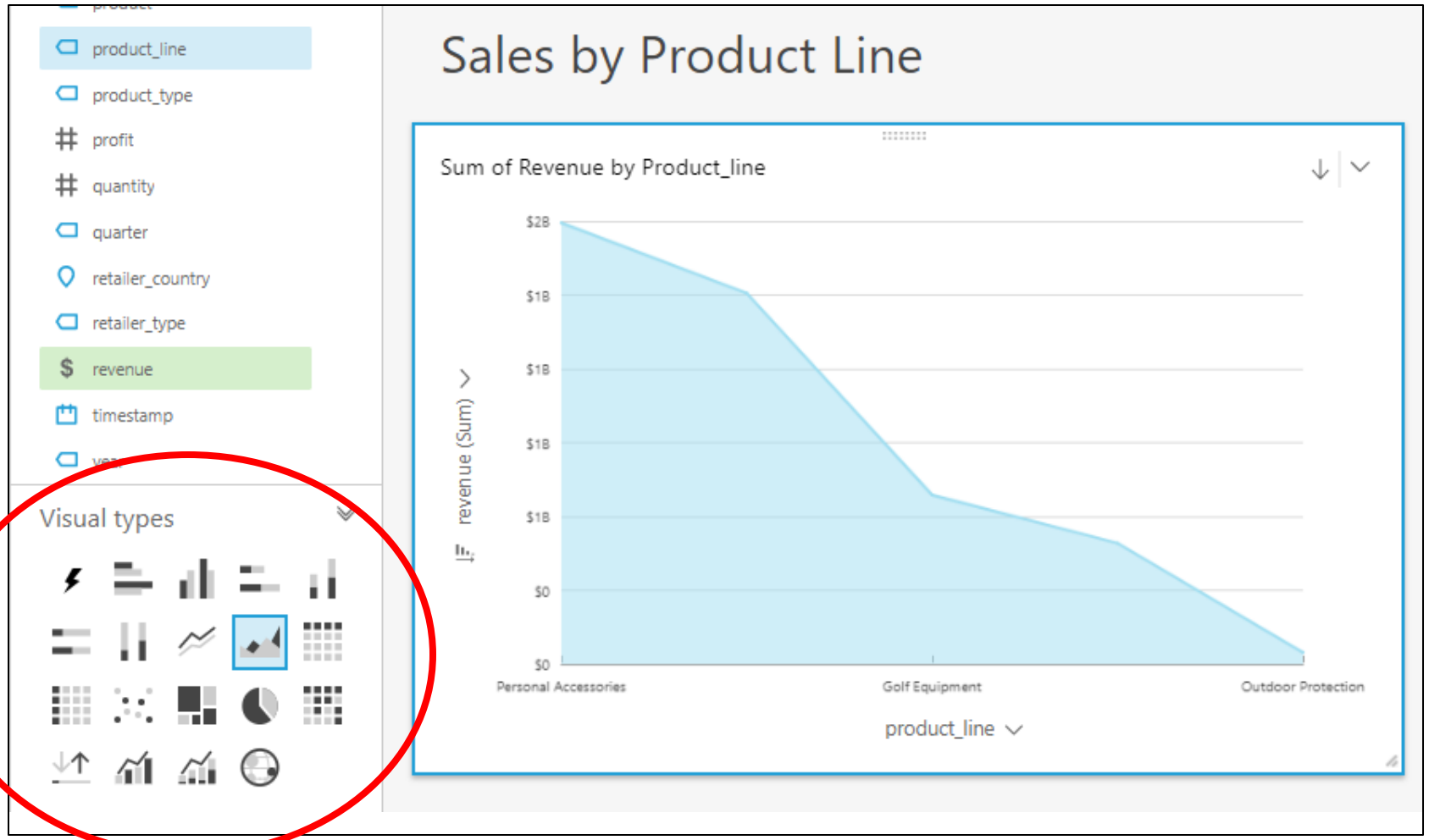
Add product_type and product as drill down layer



QUICKSIGHT

AWS Business Intelligence Tool

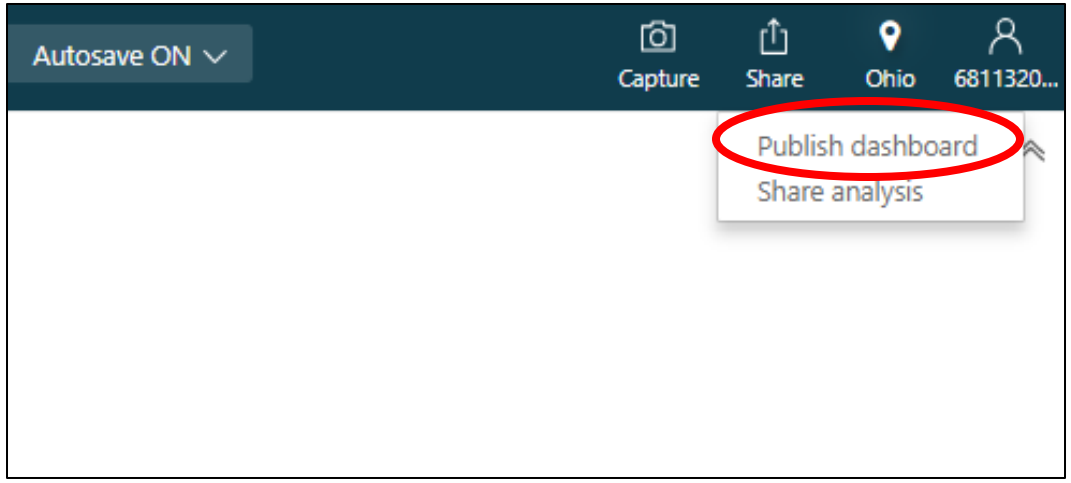
Change Visual Type



QUICKSIGHT

AWS Business Intelligence Tool

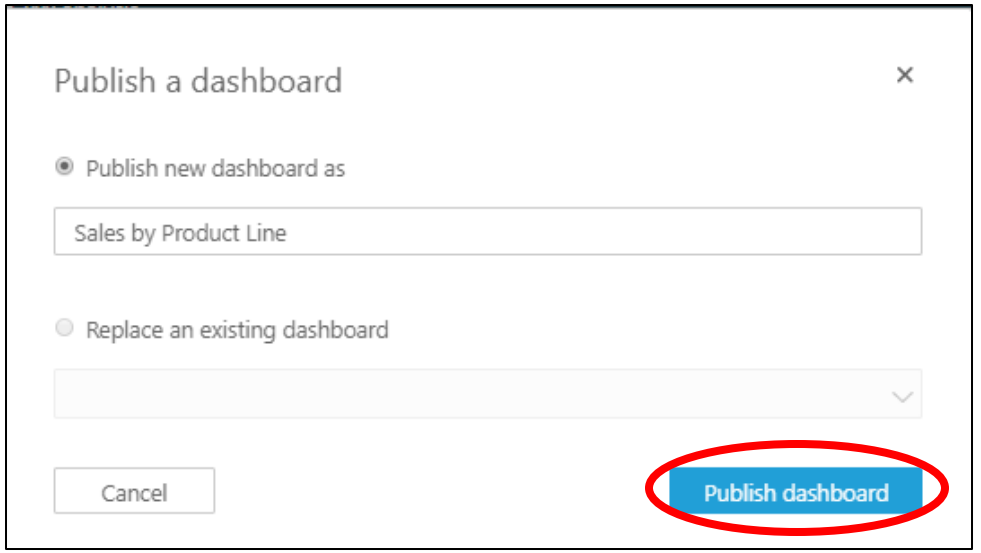
Publish to Dashboard



QUICKSIGHT

AWS Business Intelligence Tool

Name the Dashboard and select Publish dashboard



Lab 7

- Create QuickSight Account
- Create Dataset
- Create Analysis
- Publish to Dashboard
- Create More Analyses

(Use US-EAST-2/Ohio Region)

CLOUDFORMATION 

Templates

- Template used build the infrastructure for AWS resources
- Use Case:
  - Build Glue job through Cloud Formation vs Glue console
  - Advantages
    - Easy to modify
    - Easy to create multiple Glue jobs with similar patterns
    - Easy to delete multiple related resources at once
    - Easy to deploy to a different account

CLOUDFORMATION

└─Templates

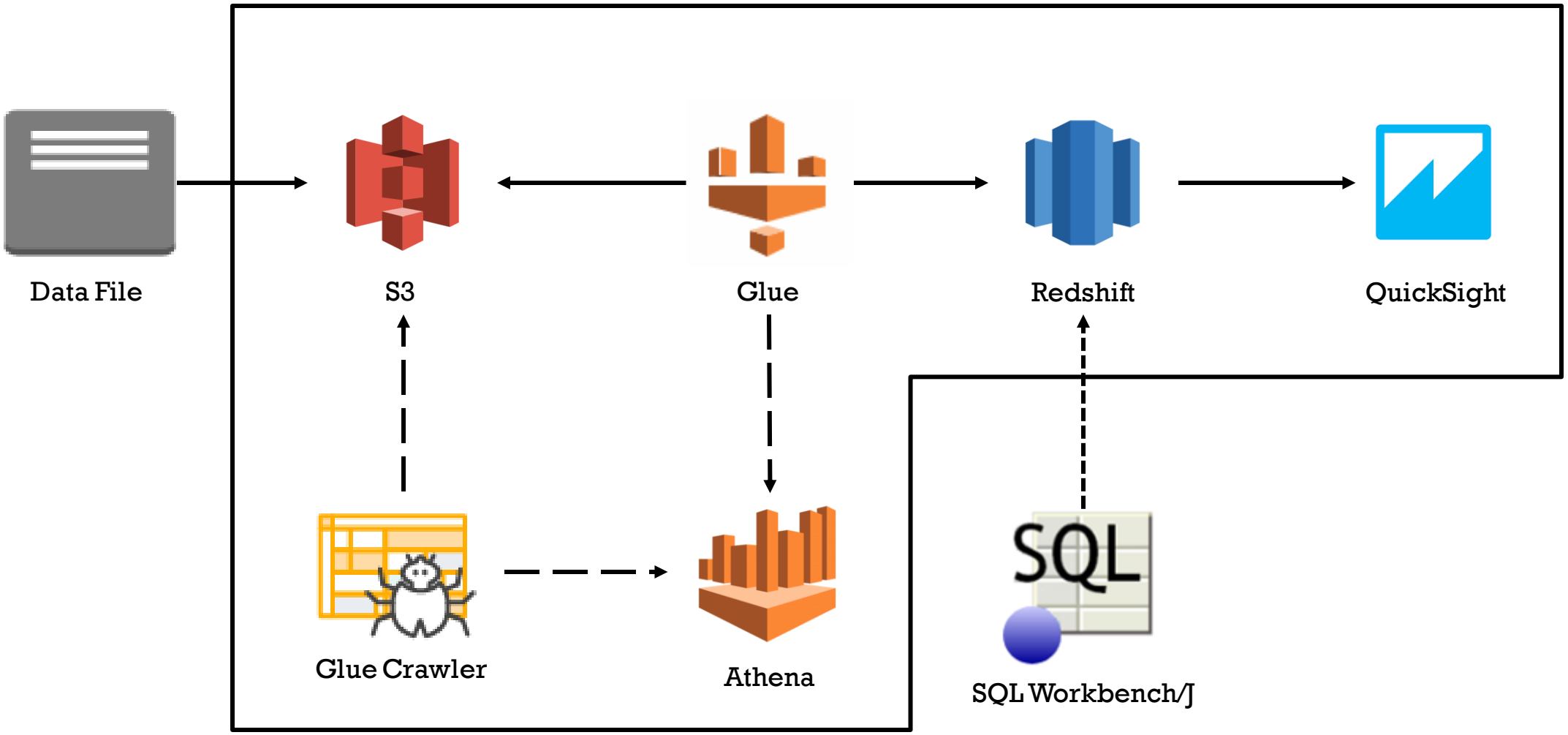
```
AWSTemplateFormatVersion:"2010-09-09“

Parameters:
  GlueDatabaseName:
    Type:String
    Default:glue_database_XXX
  GlueConnectionName:
    Type:String
    Default:glue_tutorial_XXX
  RedshiftDBName:
    Type:String
    Default:glue_tutorial_database_XXX
  SchemaName:
    Type:String
    Default:sales_redshift_schema_XXX
  RedshiftTableName:
    Type:String
    Default:products_redshift_table_XXX
  GlueTableName:
    Type:String
    Default:products_glue_table_XXX
  GlueJobName:
    Type:String
    Default:glue_tutorial
  ScriptLocation:
    Type:String
    Default:"s3://glue-tutorial-XXX/products_XXX"

Resources:
  MyJob:
    Type:AWS::Glue::Job
    Properties:
      Command:
        Name:glueetl
      ScriptLocation:!Ref ScriptLocation
      AllocatedCapacity:2
      DefaultArguments:
        "--REDSHIFT_DB_NAME":!Ref RedshiftDBName
        "--SCHEMA_NAME":!Ref SchemaName
        "--REDSHIFT_TABLE_NAME":!Ref RedshiftTableName
        "--GLUE_TABLE_NAME":!Ref GlueTableName
        "--CONNECTION_NAME":!Ref GlueConnectionName
        "--GLUE_DB_NAME":!Ref GlueDatabaseName
      ExecutionProperty:
        MaxConcurrentRuns:2
      Connections:!Ref GlueConnectionName
      MaxRetries:0
      Name:!Ref GlueJobName
```

SUMMARY

AWS Data Workflow



Conclusion

Glue - AWS ETL Tool

Simple –

Use AWS for your ETL job
Less Setup

Flexible –

Good for developers as well as non-developers
Customizable

Cost Effective –

Cheaper than other ETL tools
Pay only when you use Glue

CLEAN UP

AWS

Delete the following resources:

- Redshift Cluster *
- S3 Bucket *
- QuickSight Account *
- DynamoDB Table*
- Glue Job
- Glue Database
- Glue Table
- Glue Connection
- Lambda

* These services will accrue charges to your AWS account if not removed

RESOURCES

AWS Business Intelligence Tool

AWS Glue Documentation
https://aws.amazon.com/glue/

Pricing
Informatica
https://aws.amazon.com/marketplace/pp/B0752DY9DV?qid=1534179668153&sr=0-1&ref_=srh_res_product_title
Glue
https://aws.amazon.com/glue/pricing/
Matillion
https://aws.amazon.com/marketplace/pp/B010ED5YF8

AWS Services Documentation
https://aws.amazon.com/documentation/

Hadoop vs AWS
https://www.trustradius.com/compare-products/amazon-web-services-vs-hadoop
https://databricks.com/blog/2017/05/31/top-5-reasons-for-choosing-s3-over-hdfs.html
https://data-flair.training/blogs/13-limitations-of-hadoop/

THANK YOU

- Lydia White
  - Email: lwhite@manifestcorp.com
  - LinkedIn: https://www.linkedin.com/in/lydia-white-a30044138/

- James Zhang
  - Email: jzhang@manifestcorp.com
  - LinkedIn: https://www.linkedin.com/in/jameslzhang/