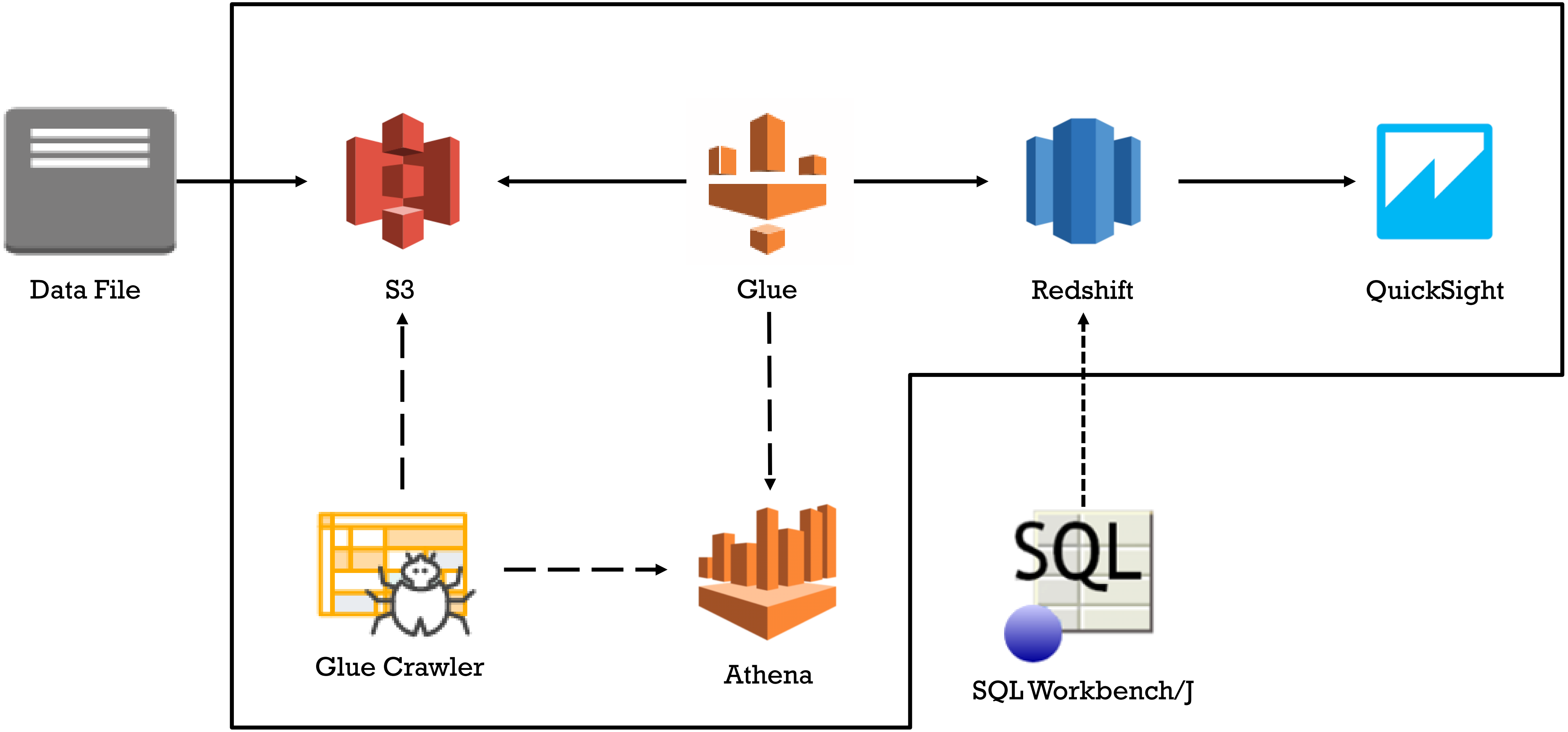
ARTS & CRAFTS WITH AWS GLUE

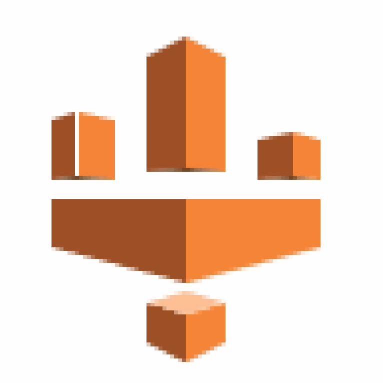
ETL Workshop

Amazon Web Services



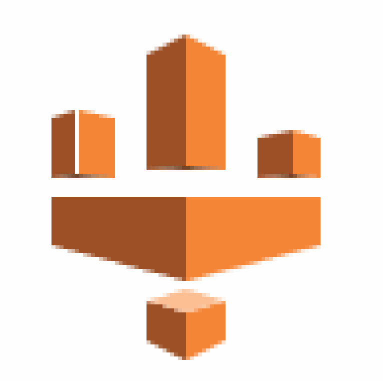
AWS Glue

What is Glue?



AWS Glue

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

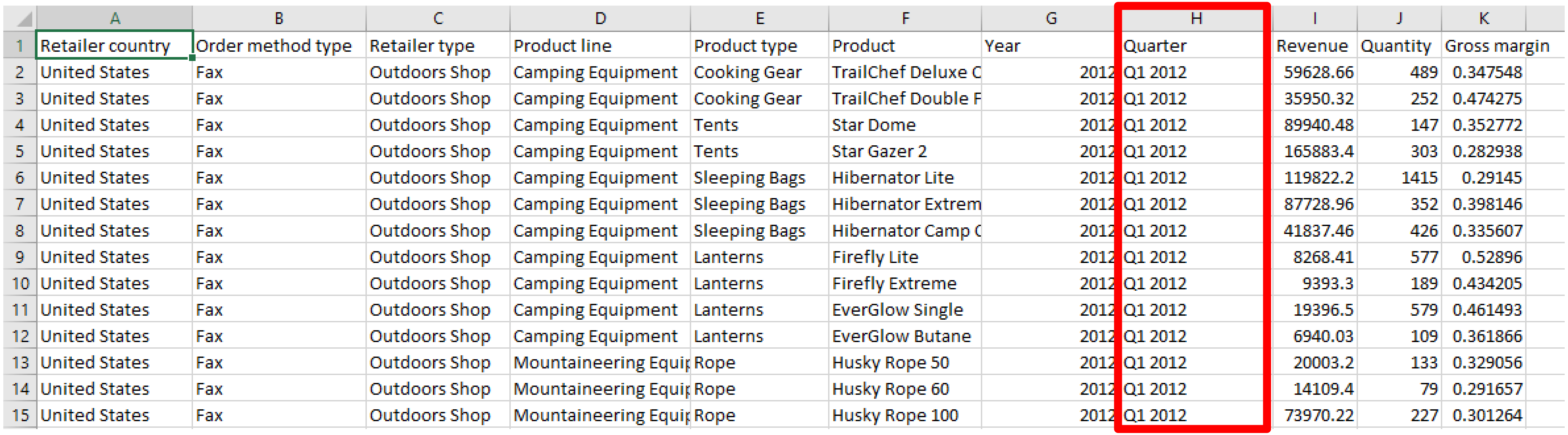


ETL

- 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



	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

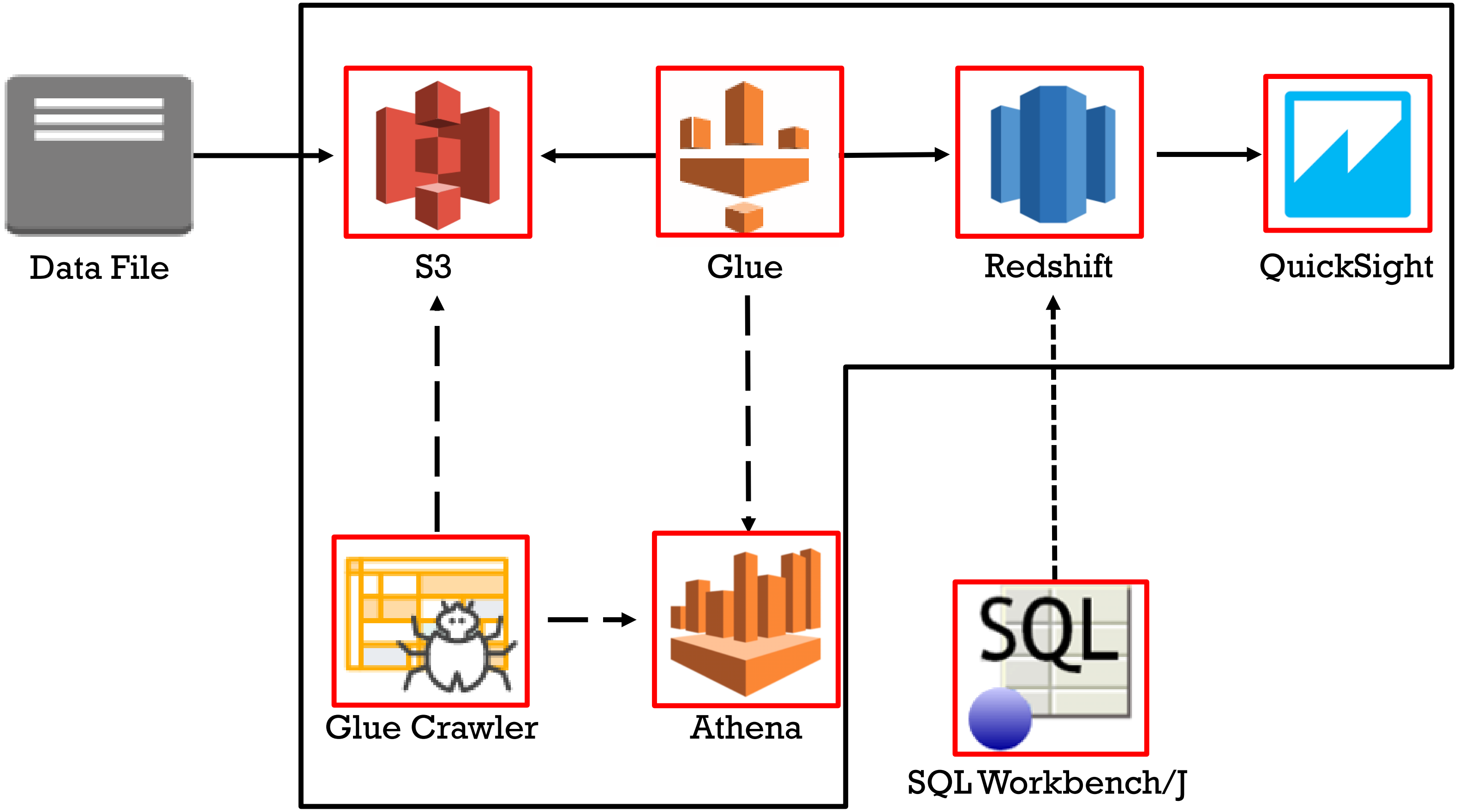
Why use Glue?

- 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

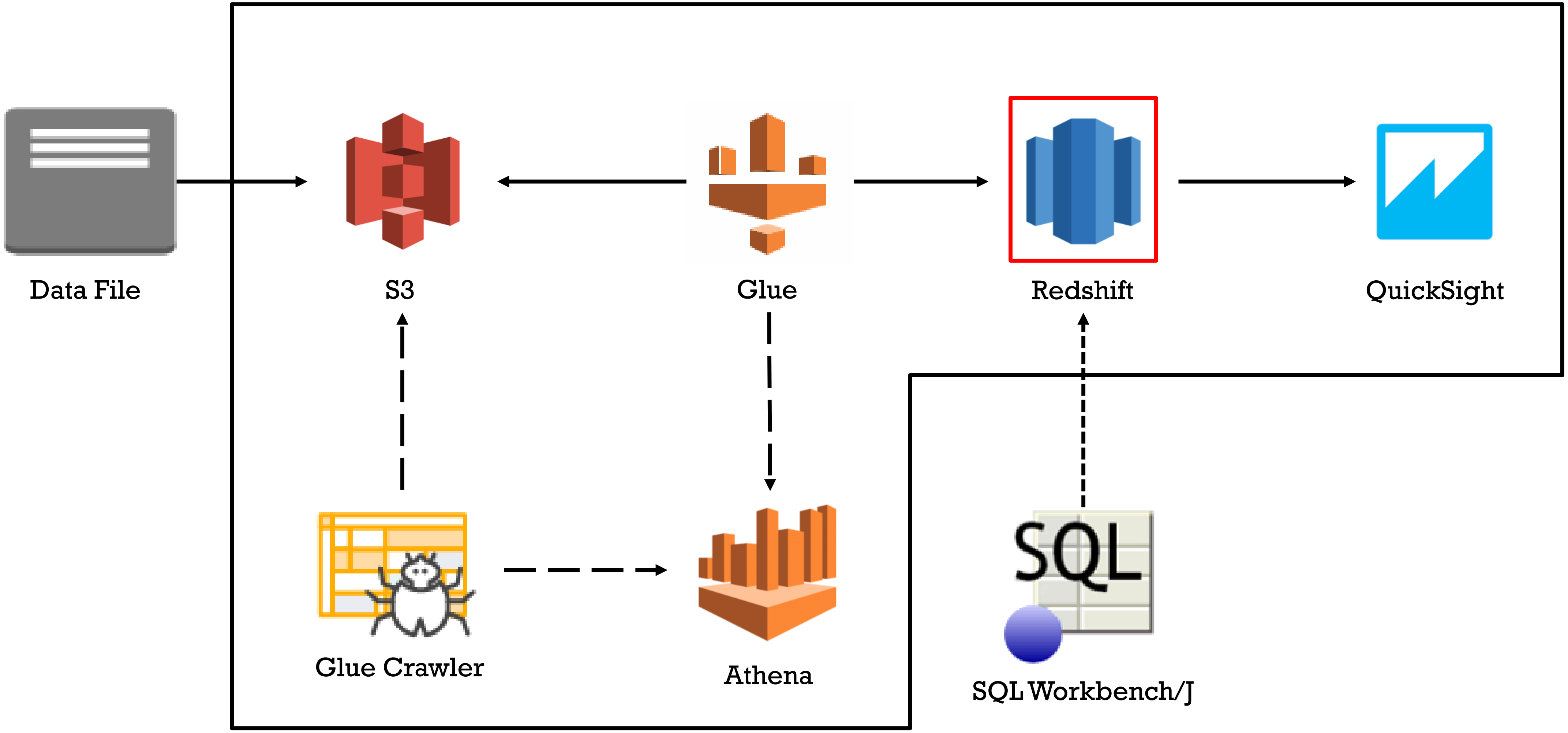
- Setup Redshift Cluster
- S3 bucket for storing the file
- Athena table to access data in file
- Glue connection
- Glue job
- Connect To Redshift in SQL Workbench
- Create Redshift table
- Run Glue job
- QuickSight

Glue Tutorial Prerequisites

- Prerequisites :
  - Setup AWS Account
  - Clone or save git repository https://github.com/jackdsilverman/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

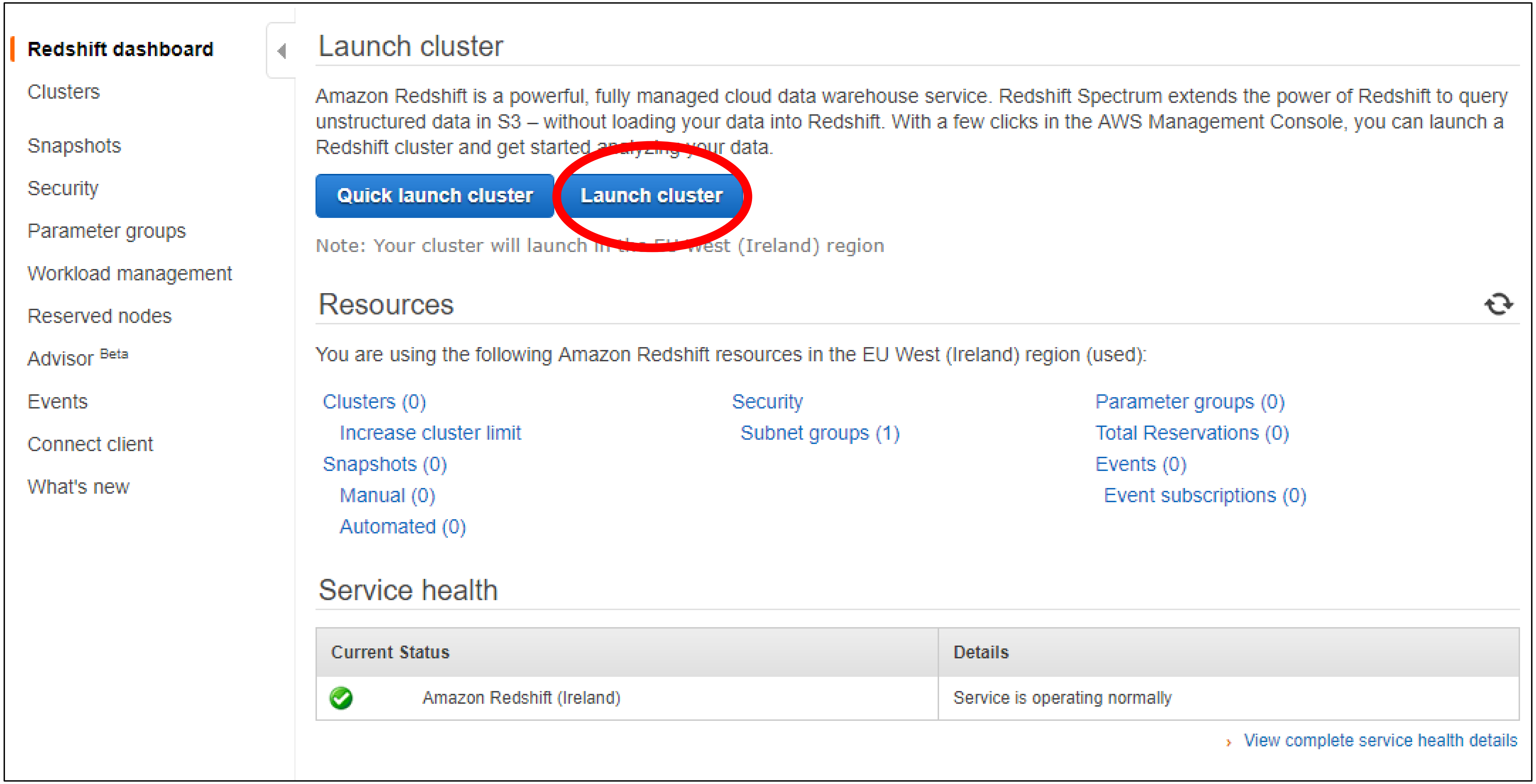
Redshift

└── **Create AWS Data Warehouse**



Redshift

└─ **Create AWS Data Warehouse**



Launch cluster

Amazon Redshift is a powerful, fully managed cloud data warehouse service. Redshift Spectrum extends the power of Redshift to query unstructured data in S3 – without loading your data into Redshift. With a few clicks in the AWS Management Console, you can launch a Redshift cluster and get started analyzing your data.

Note: Your cluster will launch in the EU West (Ireland) region

Resources

You are using the following Amazon Redshift resources in the EU West (Ireland) region (used):

Service health

Current Status	Details
✓ Amazon Redshift (Ireland)	Service is operating normally

Redshift

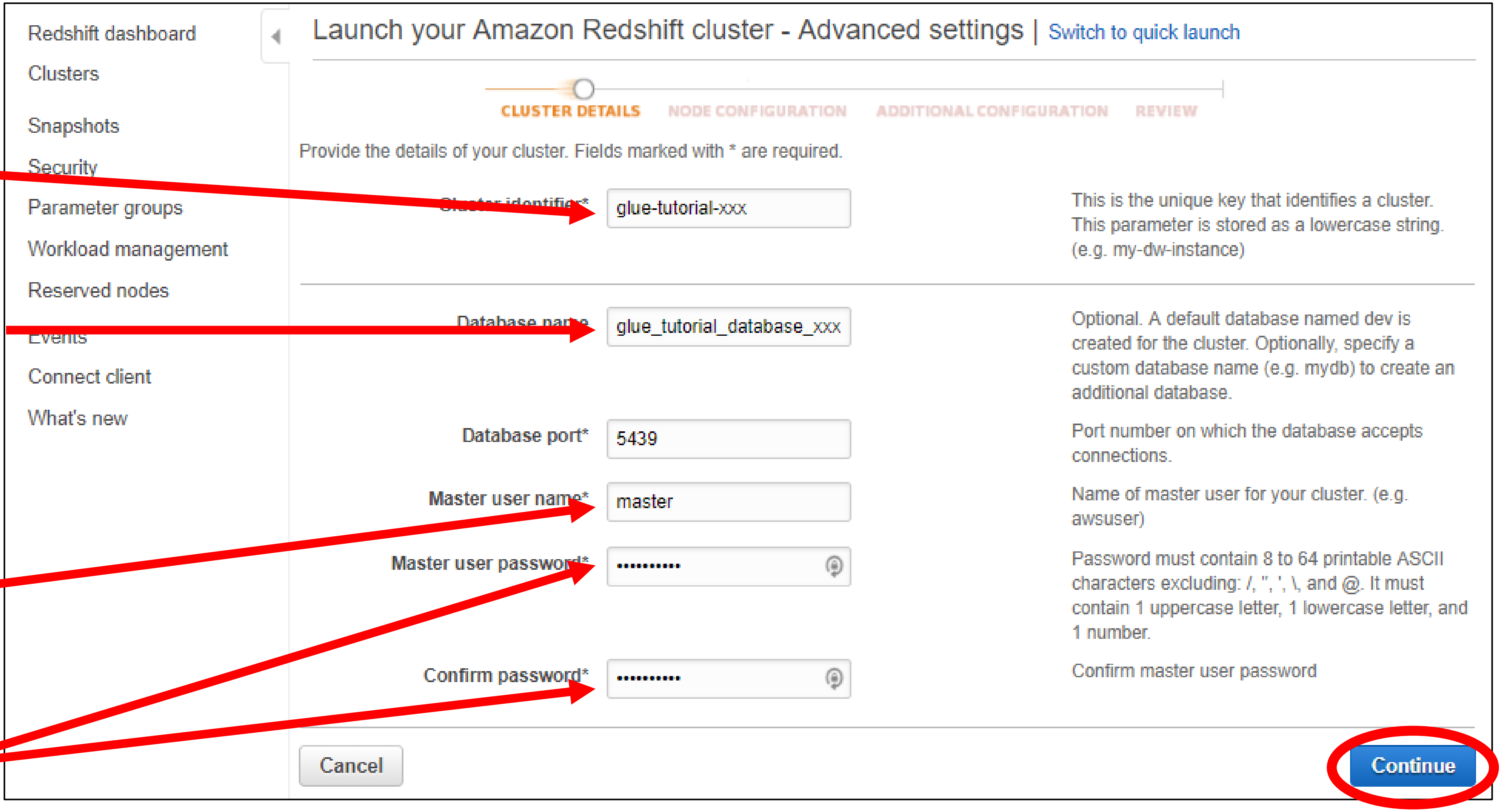
└── **Create AWS Data Warehouse**

Specify Cluster Name

Give your cluster a Database to start with

Create a user

Create a password for the user



Launch your Amazon Redshift cluster - Advanced settings | Switch to quick launch

Provide the details of your cluster. Fields marked with * are required.

Cluster identifier* glue-tutorial-xxx

This is the unique key that identifies a cluster. This parameter is stored as a lowercase string. (e.g. my-dw-instance)

Database name glue_tutorial_database_xxx

Optional. A default database named dev is created for the cluster. Optionally, specify a custom database name (e.g. mydb) to create an additional database.

Database port* 5439

Port number on which the database accepts connections.

Master user name* master

Name of master user for your cluster. (e.g. awsuser)

Master user password* ••••••••••

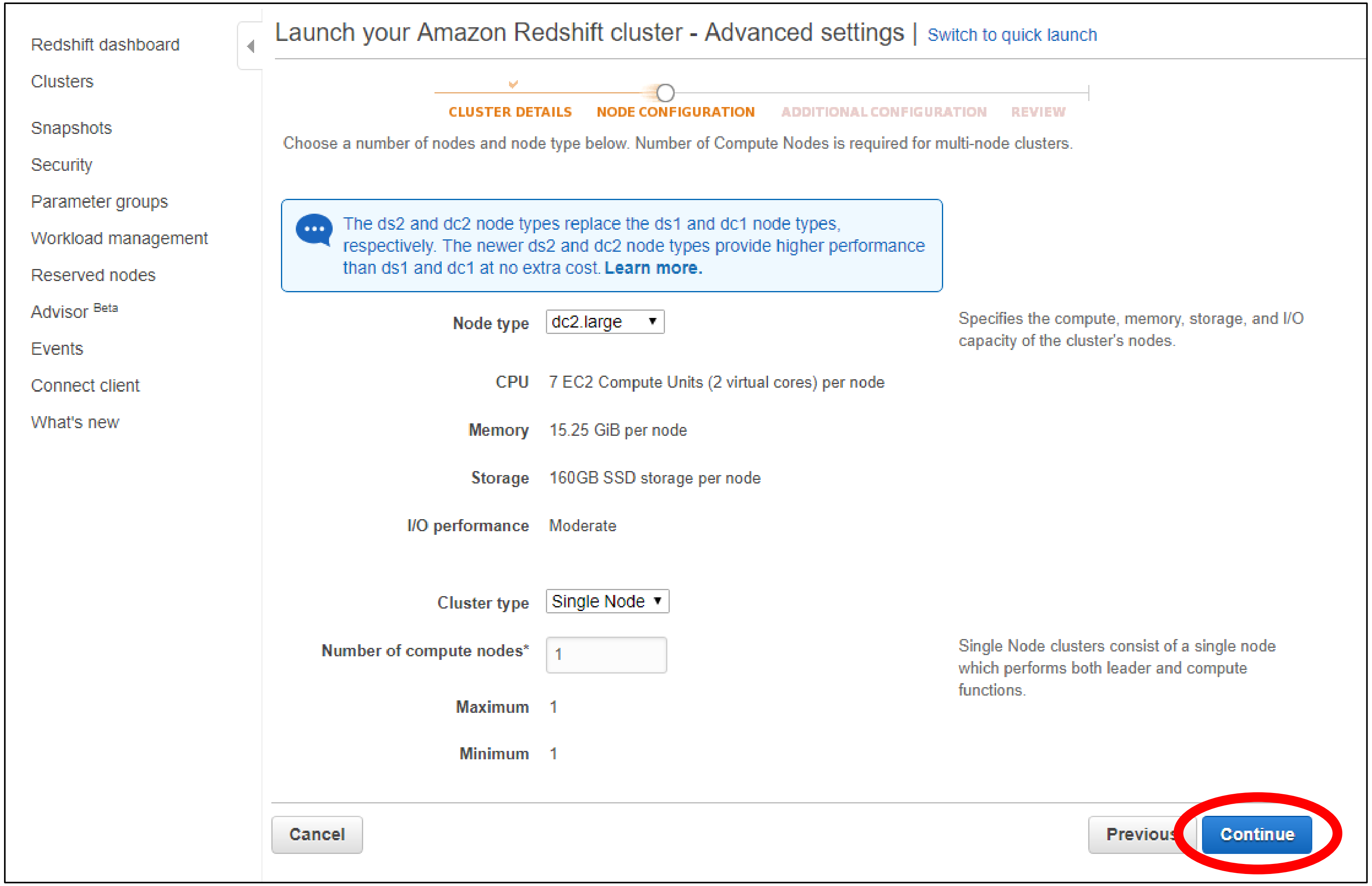
Password must contain 8 to 64 printable ASCII characters excluding: /, ", ', \, and @. It must contain 1 uppercase letter, 1 lowercase letter, and 1 number.

Confirm password* ••••••••••

Confirm master user password

Redshift

└── Create AWS Data Warehouse



Launch your Amazon Redshift cluster - Advanced settings | Switch to quick launch

Choose a number of nodes and node type below. Number of Compute Nodes is required for multi-node clusters.

The ds2 and dc2 node types replace the ds1 and dc1 node types, respectively. The newer ds2 and dc2 node types provide higher performance than ds1 and dc1 at no extra cost. **Learn more.**

Node type	dc2.large ▾	Specifies the compute, memory, storage, and I/O capacity of the cluster's nodes.
CPU	7 EC2 Compute Units (2 virtual cores) per node	
Memory	15.25 GiB per node	
Storage	160GB SSD storage per node	
I/O performance	Moderate	
Cluster type	Single Node ▾	
Number of compute nodes*	1	Single Node clusters consist of a single node which performs both leader and compute functions.
Maximum	1	
Minimum	1	

Redshift

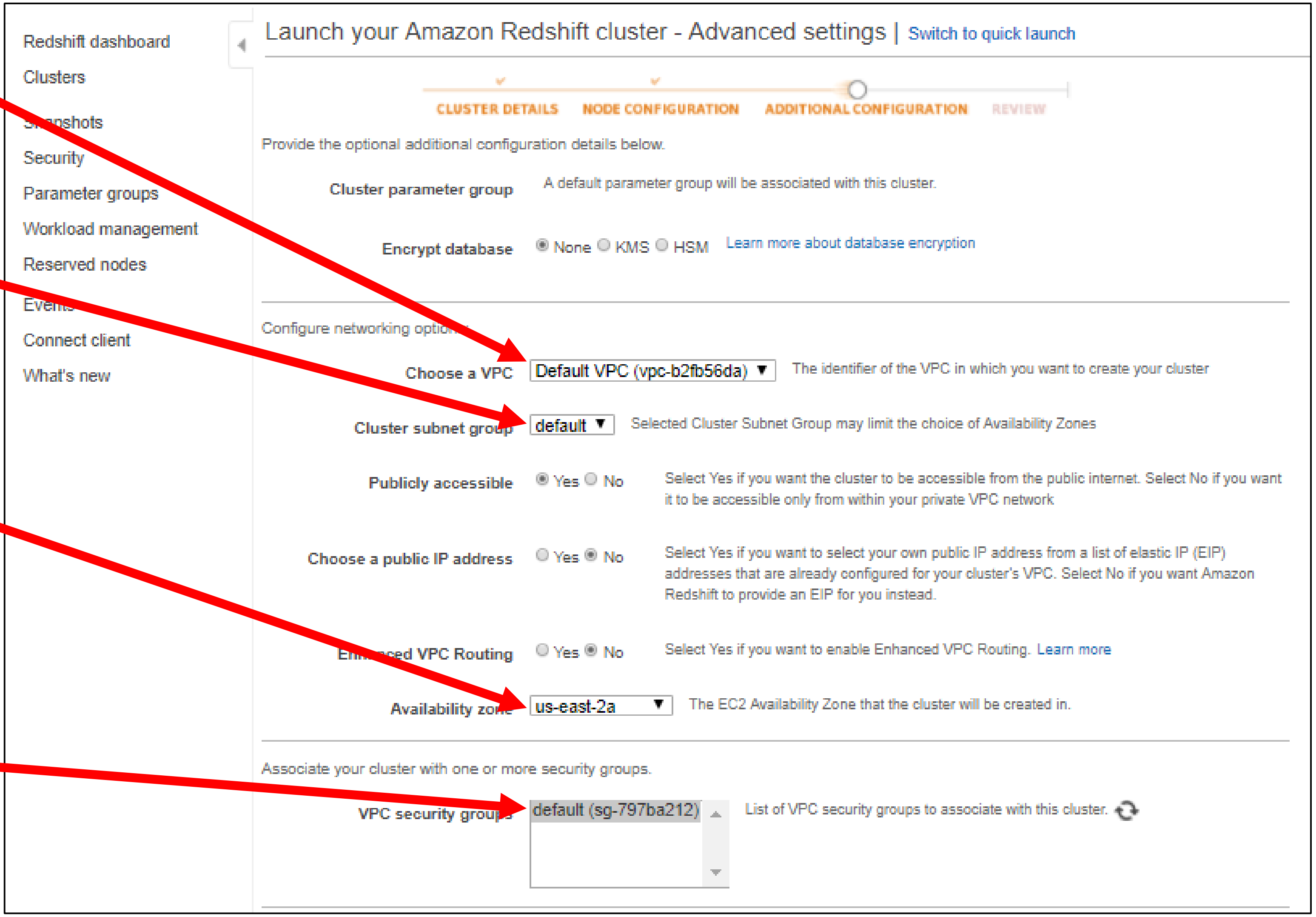
└── **Create AWS Data Warehouse**

Choose default VPC

Choose default subnet group

Choose subnet availability zone

Choose default security group



Redshift dashboard
Clusters
Snapshots
Security
Parameter groups
Workload management
Reserved nodes
Events
Connect client
What's new

Launch your Amazon Redshift cluster - Advanced settings | Switch to quick launch

CLUSTER DETAILS — NODE CONFIGURATION — ADDITIONAL CONFIGURATION — REVIEW

Provide the optional additional configuration details below.

Cluster parameter group A default parameter group will be associated with this cluster.

Encrypt database ● None ○ KMS ○ HSM Learn more about database encryption

Configure networking options:

Choose a VPC Default VPC (vpc-b2fb56da) ▼ The identifier of the VPC in which you want to create your cluster

Cluster subnet group default ▼ Selected Cluster Subnet Group may limit the choice of Availability Zones

Publicly accessible ● Yes ○ No Select Yes if you want the cluster to be accessible from the public internet. Select No if you want it to be accessible only from within your private VPC network

Choose a public IP address ○ Yes ● No Select Yes if you want to select your own public IP address from a list of elastic IP (EIP) addresses that are already configured for your cluster's VPC. Select No if you want Amazon Redshift to provide an EIP for you instead.

Enhanced VPC Routing ○ Yes ● No Select Yes if you want to enable Enhanced VPC Routing. Learn more

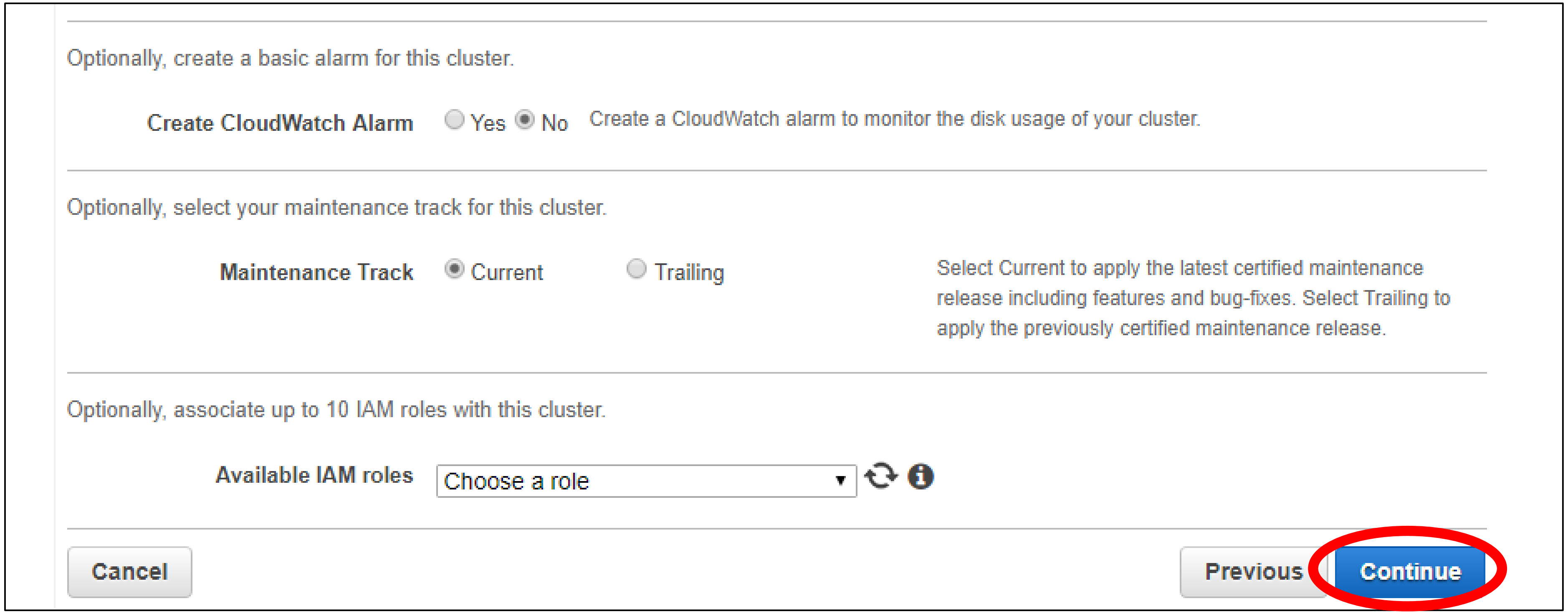
Availability zone us-east-2a ▼ The EC2 Availability Zone that the cluster will be created in.

Associate your cluster with one or more security groups.

VPC security groups default (sg-797ba212) List of VPC security groups to associate with this cluster.

Redshift

└── Create AWS Data Warehouse



Optionally, create a basic alarm for this cluster.

Create CloudWatch Alarm ○ Yes ● No Create a CloudWatch alarm to monitor the disk usage of your cluster.

Optionally, select your maintenance track for this cluster.

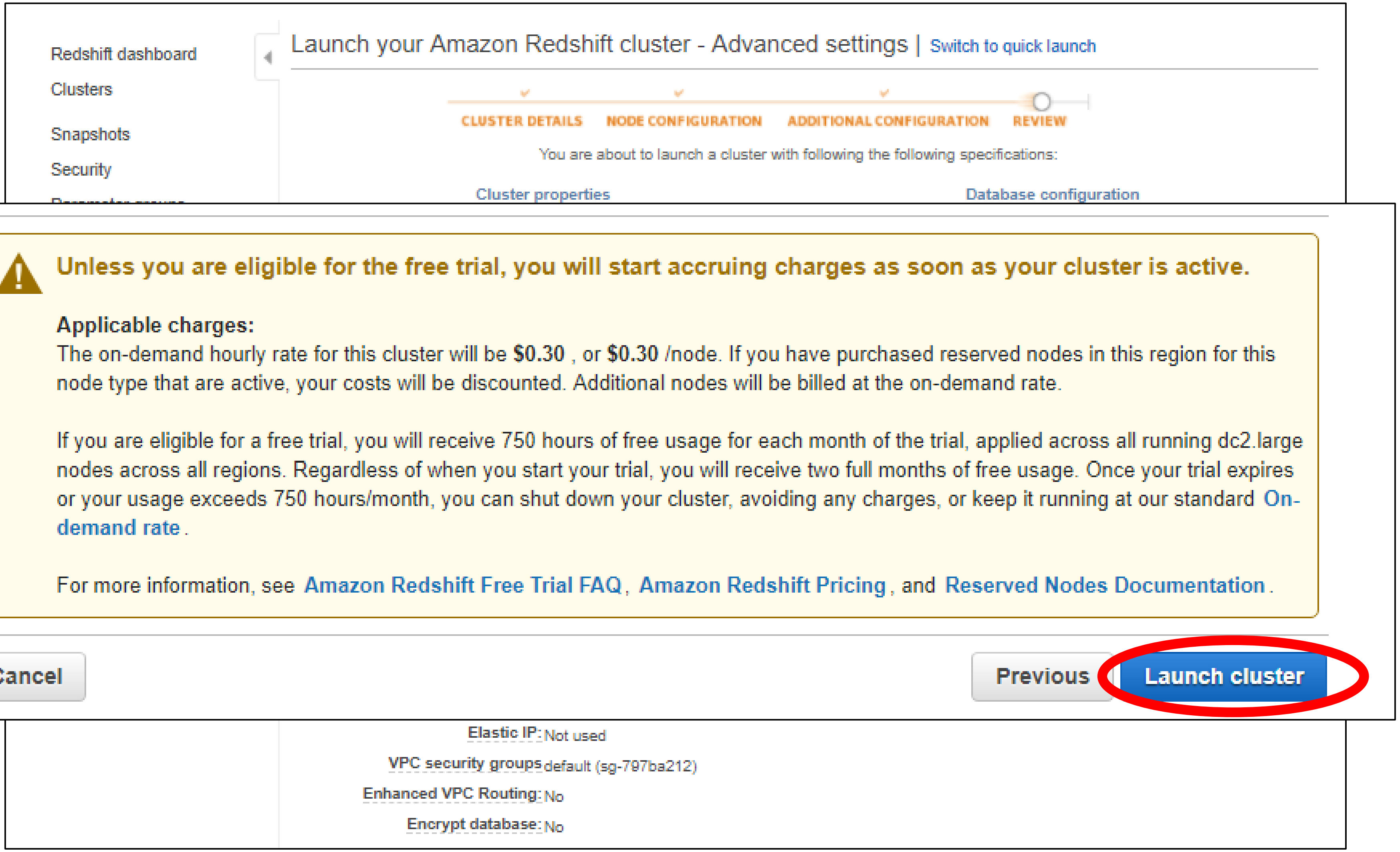
Maintenance Track ● Current ○ Trailing Select Current to apply the latest certified maintenance release including features and bug-fixes. Select Trailing to apply the previously certified maintenance release.

Optionally, associate up to 10 IAM roles with this cluster.

Available IAM roles: Choose a role ▾

Redshift

└── Create AWS Data Warehouse



Launch your Amazon Redshift cluster - Advanced settings | Switch to quick launch

You are about to launch a cluster with following the following specifications:

Cluster properties

Database configuration

Unless you are eligible for the free trial, you will start accruing charges as soon as your cluster is active.

Applicable charges:
The on-demand hourly rate for this cluster will be **$0.30** , or **$0.30** /node. If you have purchased reserved nodes in this region for this node type that are active, your costs will be discounted. Additional nodes will be billed at the on-demand rate.

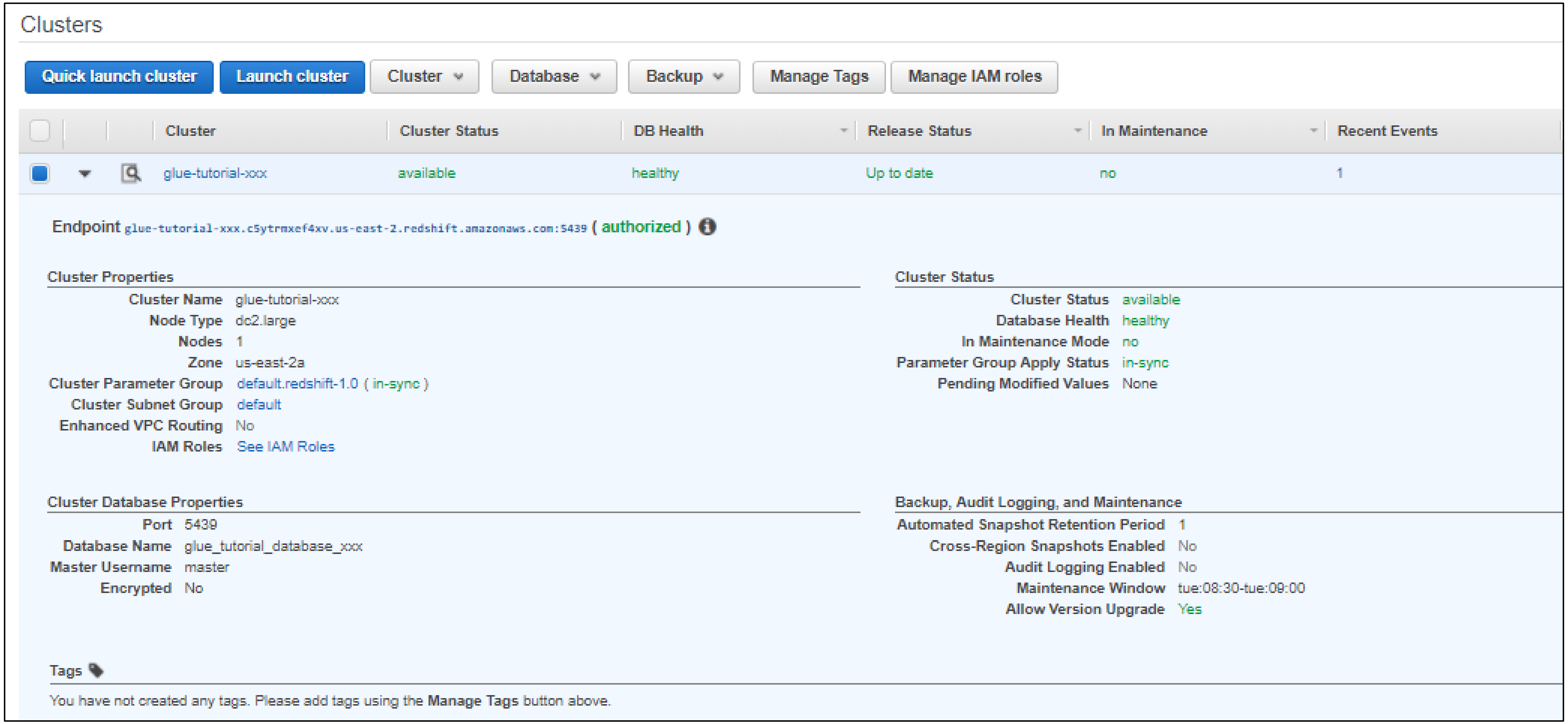
If you are eligible for a free trial, you will receive 750 hours of free usage for each month of the trial, applied across all running dc2.large nodes across all regions. Regardless of when you start your trial, you will receive two full months of free usage. Once your trial expires or your usage exceeds 750 hours/month, you can shut down your cluster, avoiding any charges, or keep it running at our standard On-demand rate .

For more information, see Amazon Redshift Free Trial FAQ , Amazon Redshift Pricing , and Reserved Nodes Documentation .

Elastic IP: Not used
VPC security groups default (sg-797ba212)
Enhanced VPC Routing: No
Encrypt database: No

Redshift

└── Create AWS Data Warehouse



Clusters

Cluster	Cluster Status	DB Health	Release Status	In Maintenance	Recent Events
glue-tutorial-xxx	available	healthy	Up to date	no	1

Endpoint `glue-tutorial-xxx.c5ytrmxef4xv.us-east-2.redshift.amazonaws.com:5439` (authorized)

Cluster Properties

Cluster Name	glue-tutorial-xxx
Node Type	dc2.large
Nodes	1
Zone	us-east-2a
Cluster Parameter Group	default.redshift-1.0 (in-sync)
Cluster Subnet Group	default
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

Cluster Database Properties

Port	5439
Database Name	glue_tutorial_database_xxx
Master Username	master
Encrypted	No

Backup, Audit Logging, and Maintenance

Automated Snapshot Retention Period	1
Cross-Region Snapshots Enabled	No
Audit Logging Enabled	No
Maintenance Window	tue:08:30-tue:09:00
Allow Version Upgrade	Yes

Tags

You have not created any tags. Please add tags using the **Manage Tags** button above.

Lab 1

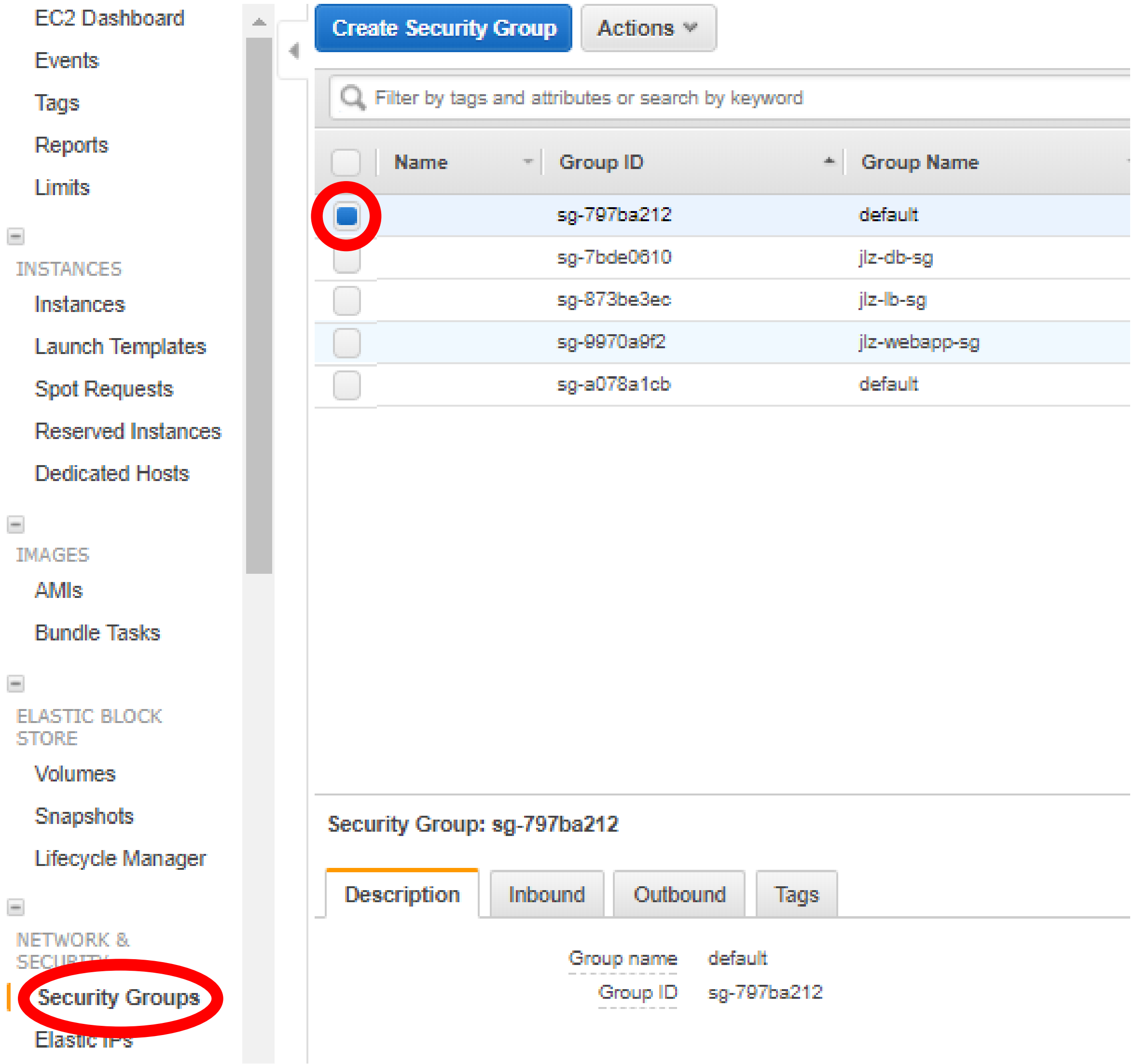
- Launch Redshift cluster

(Use US-EAST-2/Ohio Region)

EC2

└─Edit Security Groups

In a new tab go to the EC2 service



	Name	Group ID	Group Name
☑		sg-797ba212	default
□		sg-7bde0610	jlz-db-sg
□		sg-873be3ec	jlz-lb-sg
□		sg-9970a9f2	jlz-webapp-sg
□		sg-a078a1cb	default

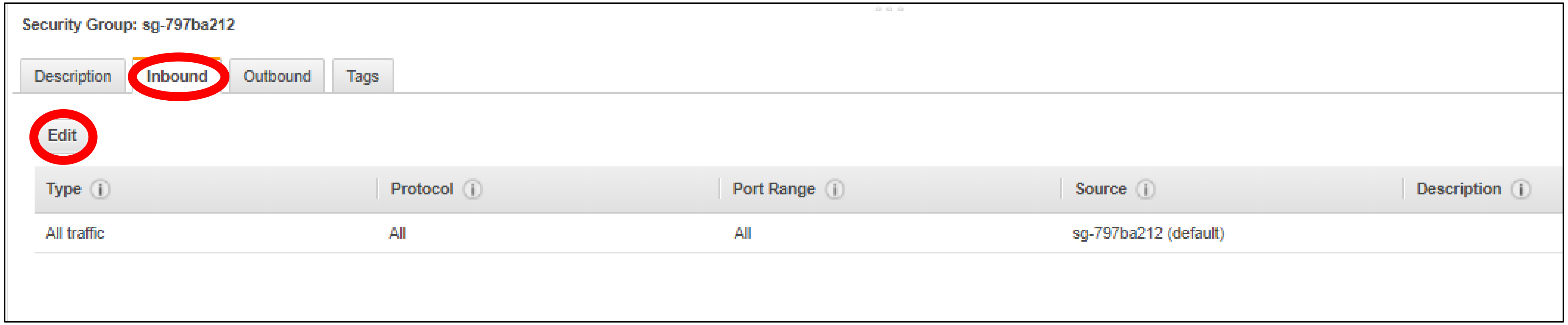
Security Group: sg-797ba212

Group name default

Group ID sg-797ba212

EC2

└─Edit Security Groups



Security Group: sg-797ba212

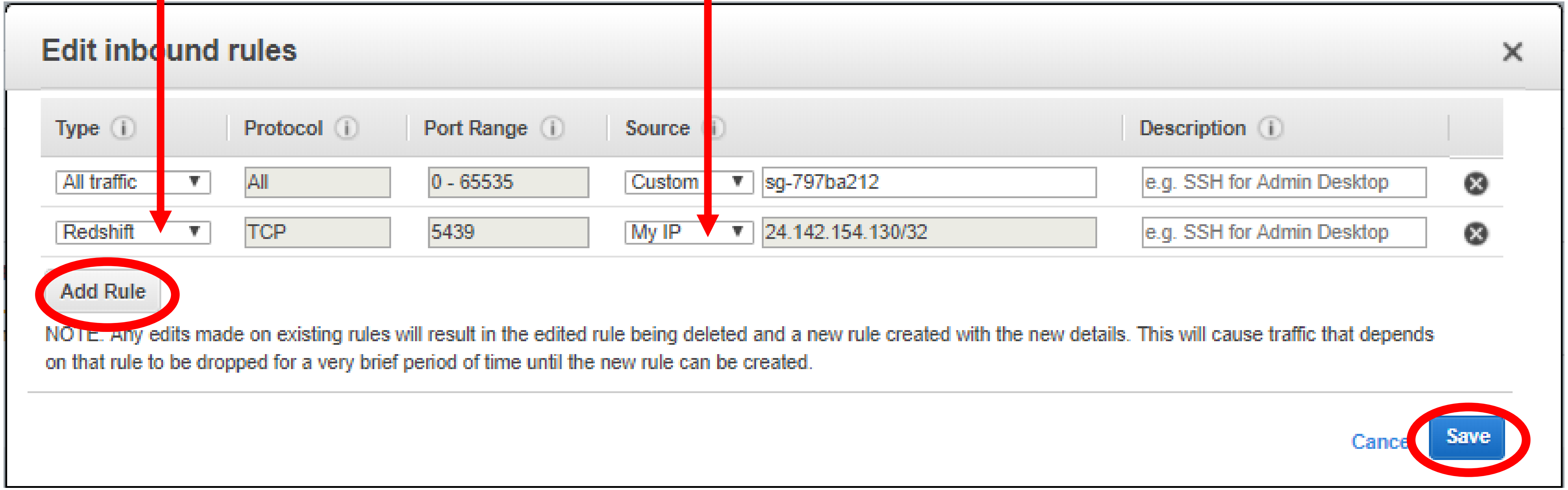
Type	Protocol	Port Range	Source	Description
All traffic	All	All	sg-797ba212 (default)	

EC2

└─Edit Security Groups

Choose Redshift Type

Specifies who has access to the Redshift cluster



Edit inbound rules

Type	Protocol	Port Range	Source		Description
All traffic	All	0 - 65535	Custom	sg-797ba212	e.g. SSH for Admin Desktop
Redshift	TCP	5439	My IP	24.142.154.130/32	e.g. SSH for Admin Desktop

Add Rule

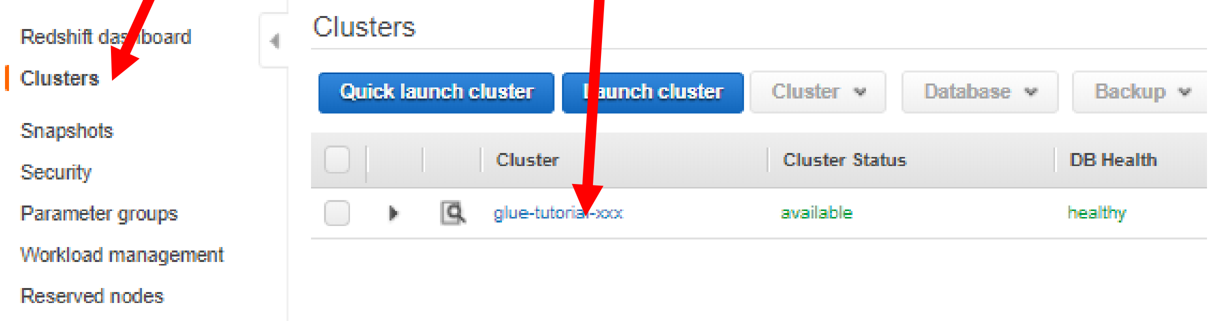
NOTE: Any edits made on existing rules will result in the edited rule being deleted and a new rule created with the new details. This will cause traffic that depends on that rule to be dropped for a very brief period of time until the new rule can be created.

Redshift

└─**Connection**

Go to Redshift and select ‘Clusters’

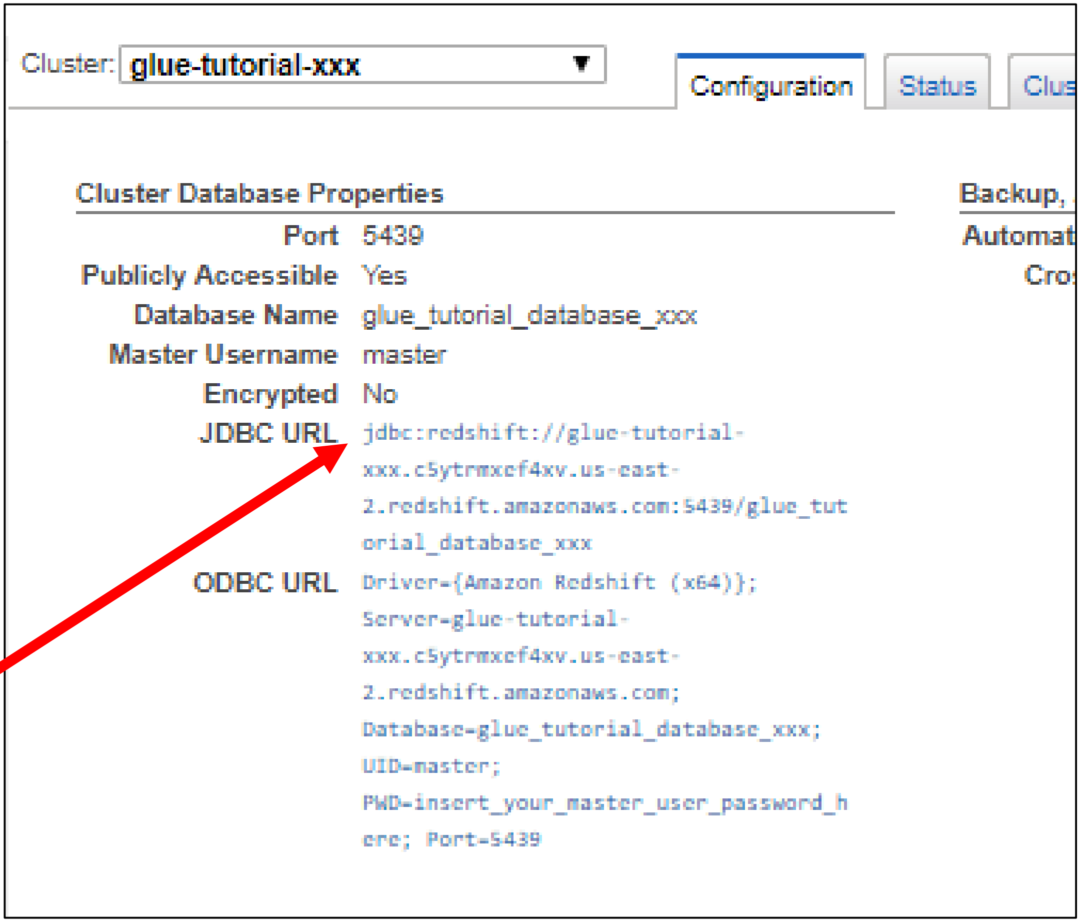
Select glue-tutorial



Clusters

Cluster	Cluster Status	DB Health
glue-tutorial-xxx	available	healthy

Scroll down to Cluster Database Properties and copy the JDBC URL



Cluster: glue-tutorial-xxx

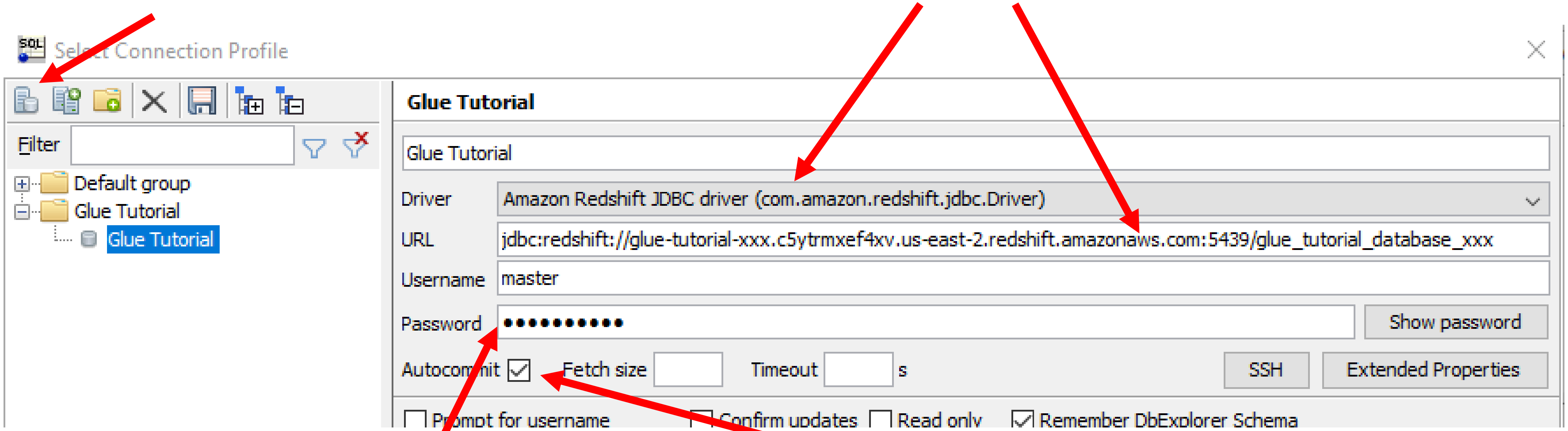
Cluster Database Properties

Port	5439
Publicly Accessible	Yes
Database Name	glue_tutorial_database_xxx
Master Username	master
Encrypted	No
JDBC URL	jdbc:redshift://glue-tutorial-xxx.c5ytrmxef4xv.us-east-2.redshift.amazonaws.com:5439/glue_tutorial_database_xxx
ODBC URL	Driver={Amazon Redshift (x64)}; Server=glue-tutorial-xxx.c5ytrmxef4xv.us-east-2.redshift.amazonaws.com; Database=glue_tutorial_database_xxx; UID=master; PWD=insert_your_master_user_password_here; Port=5439

Redshift
└─Connection

Open SQL Workbench and select Create a new connection

Set the Driver to Amazon Redshift and paste the JDBC URL



Glue Tutorial

Driver: Amazon Redshift JDBC driver (com.amazon.redshift.jdbc.Driver)

URL: jdbc:redshift://glue-tutorial-xxx.c5ytrmxef4xv.us-east-2.redshift.amazonaws.com:5439/glue_tutorial_database_xxx

Username: master

Password: •••••••••••

Autocommit ☑ Fetch size Timeout s

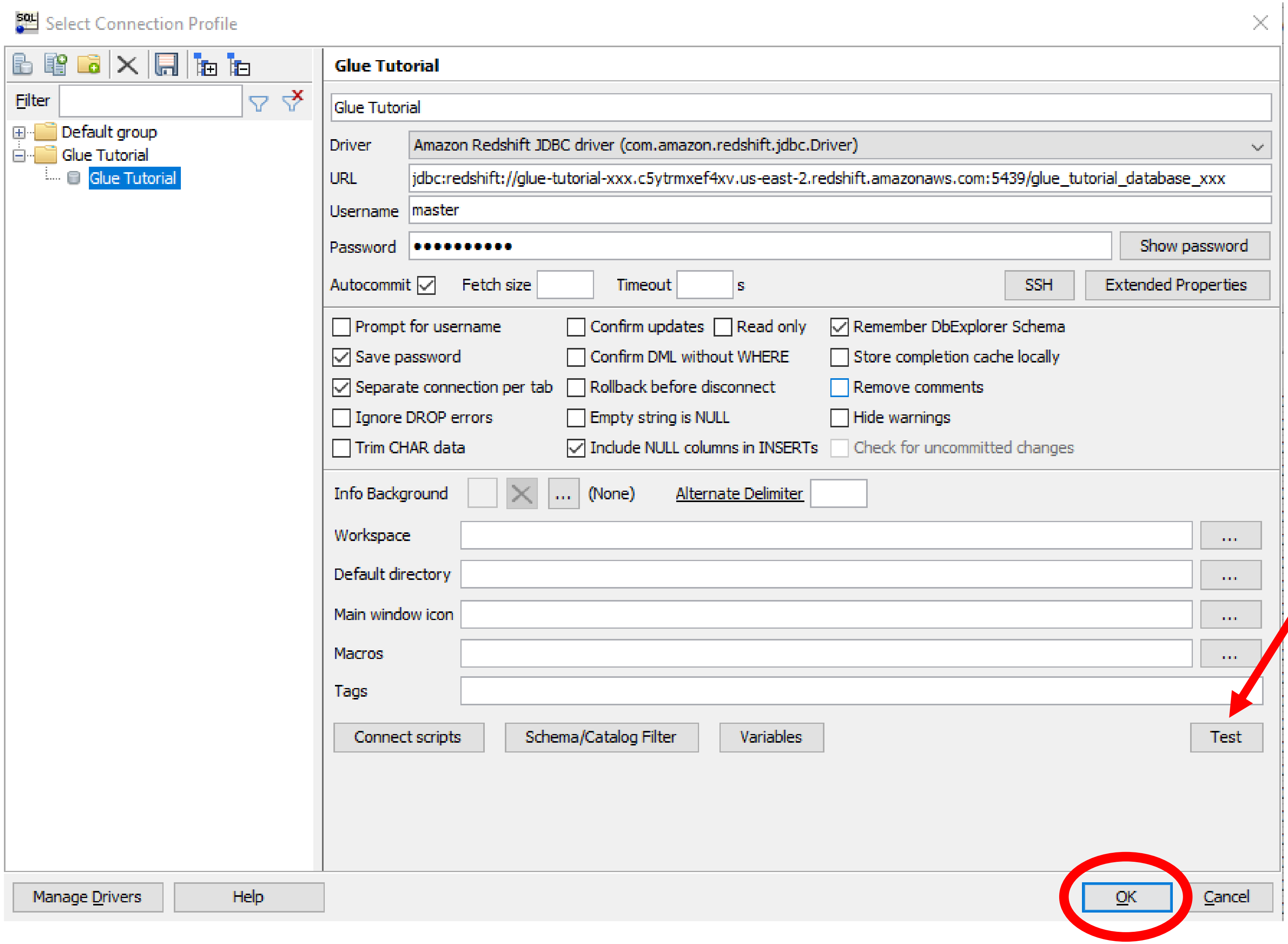
□ Prompt for username □ Confirm updates □ Read only ☑ Remember DbExplorer Schema

The username and password that was created

Select Autocommit

Redshift

└─Connection



Select Connection Profile

Glue Tutorial

Driver: Amazon Redshift JDBC driver (com.amazon.redshift.jdbc.Driver)

URL: jdbc:redshift://glue-tutorial-xxx.c5ytrmxef4xv.us-east-2.redshift.amazonaws.com:5439/glue_tutorial_database_xxx

Username: master

Password: ••••••••••

Autocommit ☑ Fetch size Timeout s

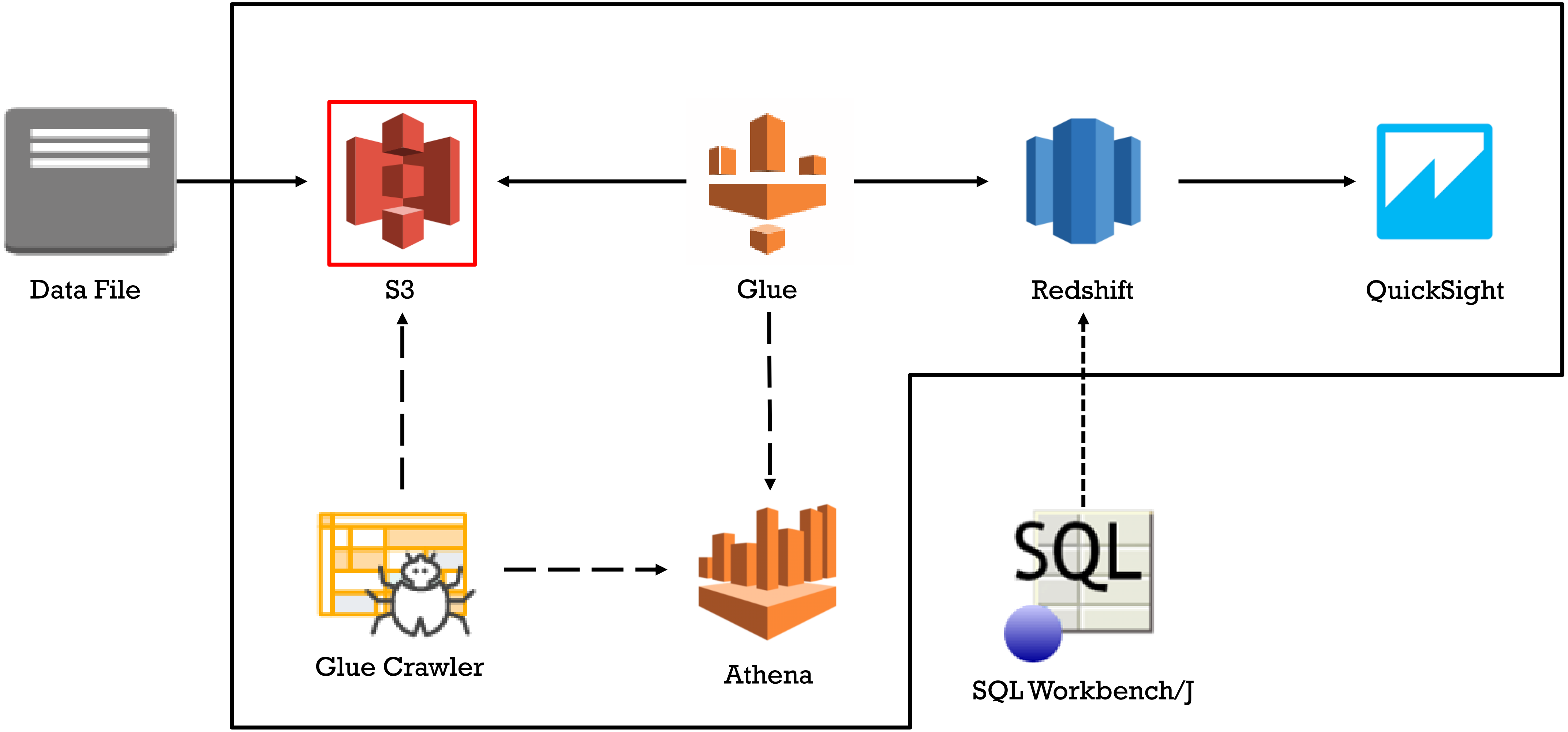
□ Prompt for username	□ Confirm updates □ Read only	☑ Remember DbExplorer Schema
☑ Save password	□ Confirm DML without WHERE	□ Store completion cache locally
☑ Separate connection per tab	□ Rollback before disconnect	□ Remove comments
□ Ignore DROP errors	□ Empty string is NULL	□ Hide warnings
□ Trim CHAR data	☑ Include NULL columns in INSERTs	□ Check for uncommitted changes

Info Background (None) Alternate Delimiter

Test your connection

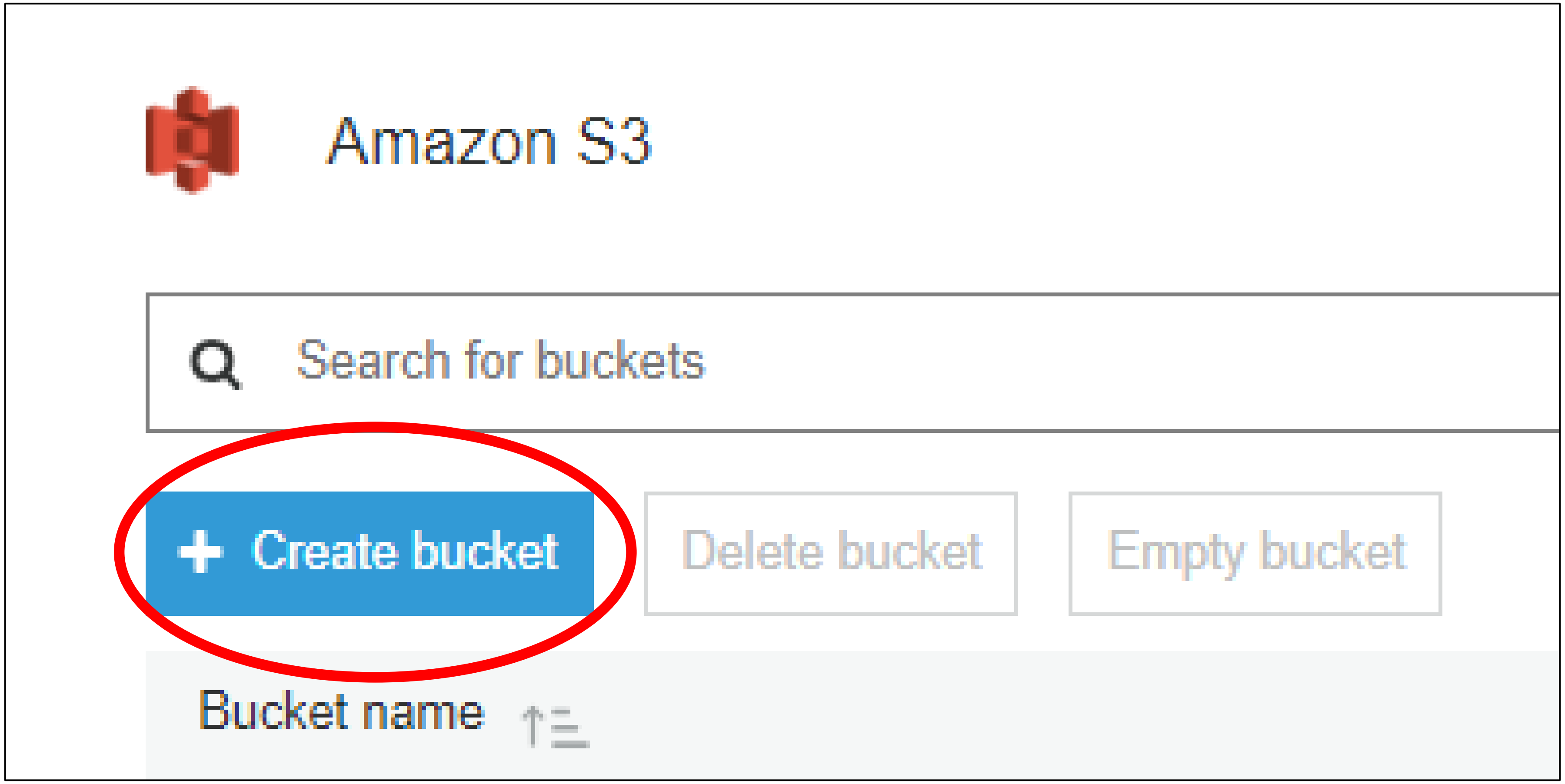
S3

└─**Create S3 bucket with AWS Console**



S3

└─**Create S3 bucket with AWS Console**

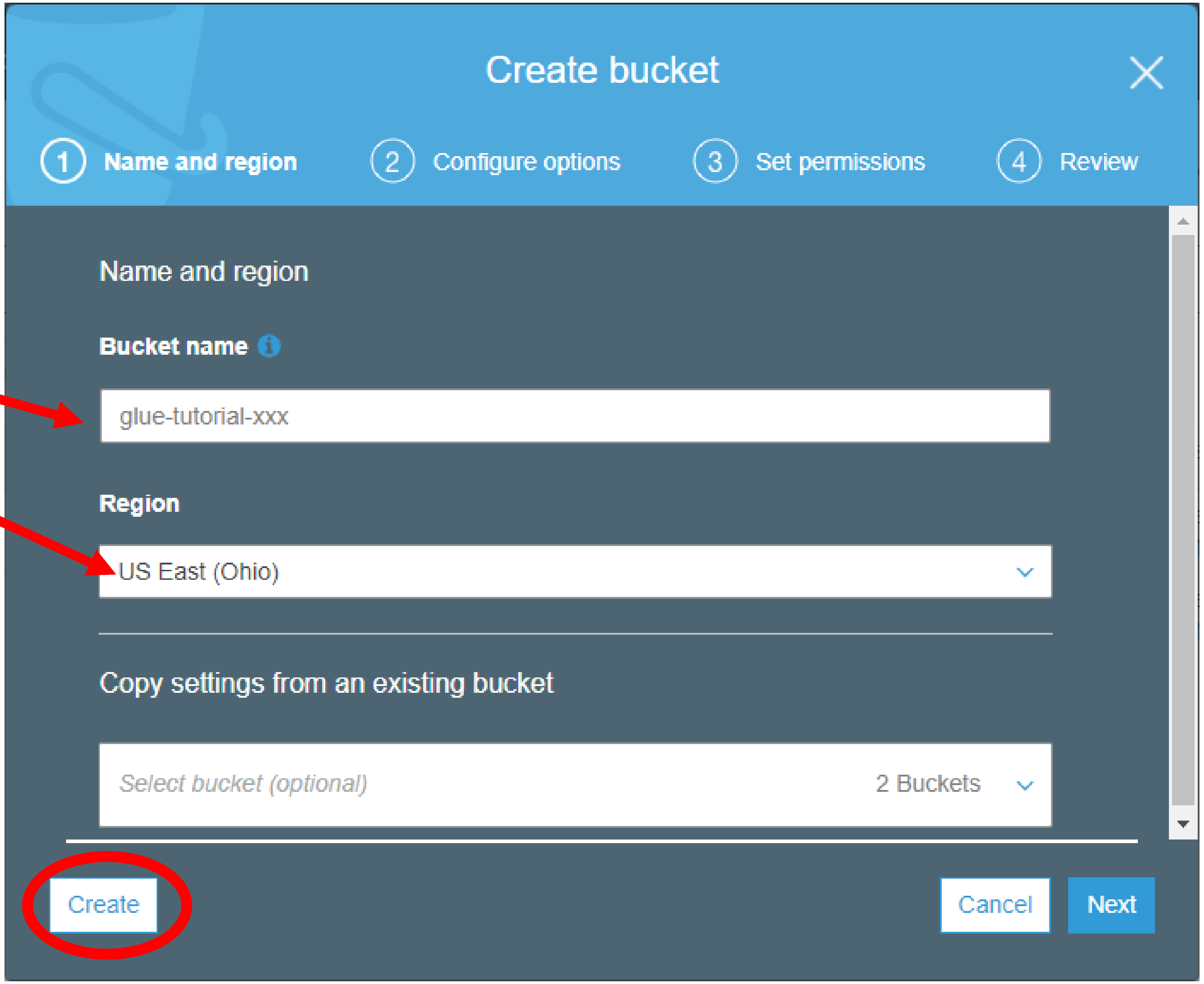


S3

└─Create S3 bucket with AWS Console

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

Specify the region



Name and region

Bucket name: glue-tutorial-xxx

Region: US East (Ohio)

Copy settings from an existing bucket

Select bucket (optional) — 2 Buckets

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

S3

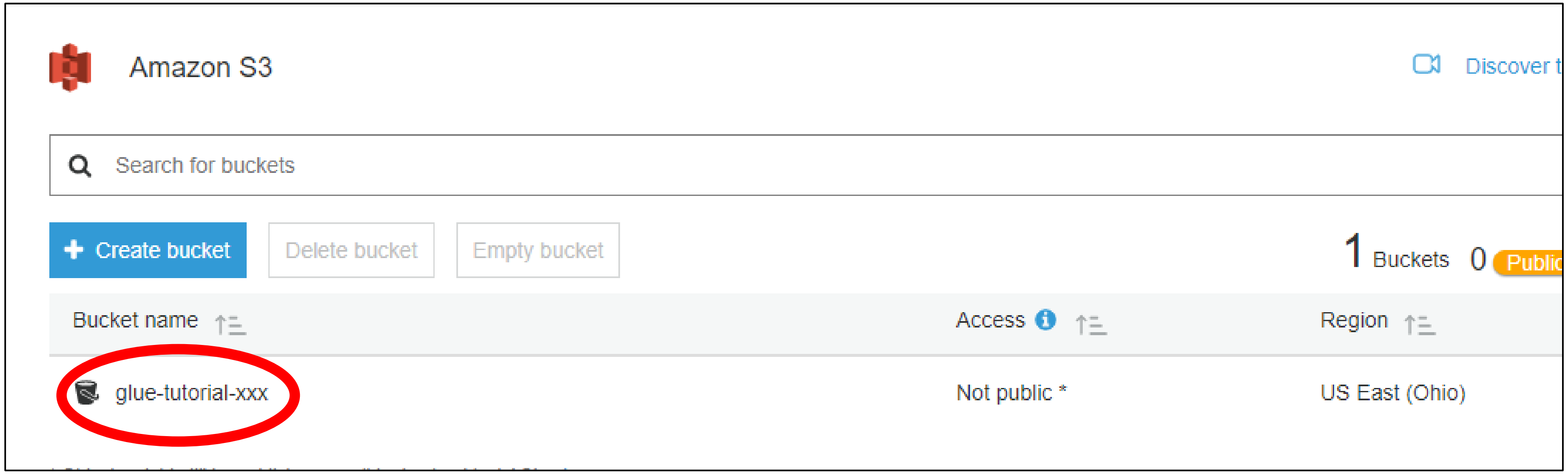
Create S3 bucket with AWS CLI* (Alternative)

```
$ aws s3api create-bucket --bucket  glue-tutorial-XXX --region
us-east-1
```

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

S3

└─Create S3 bucket with AWS Console

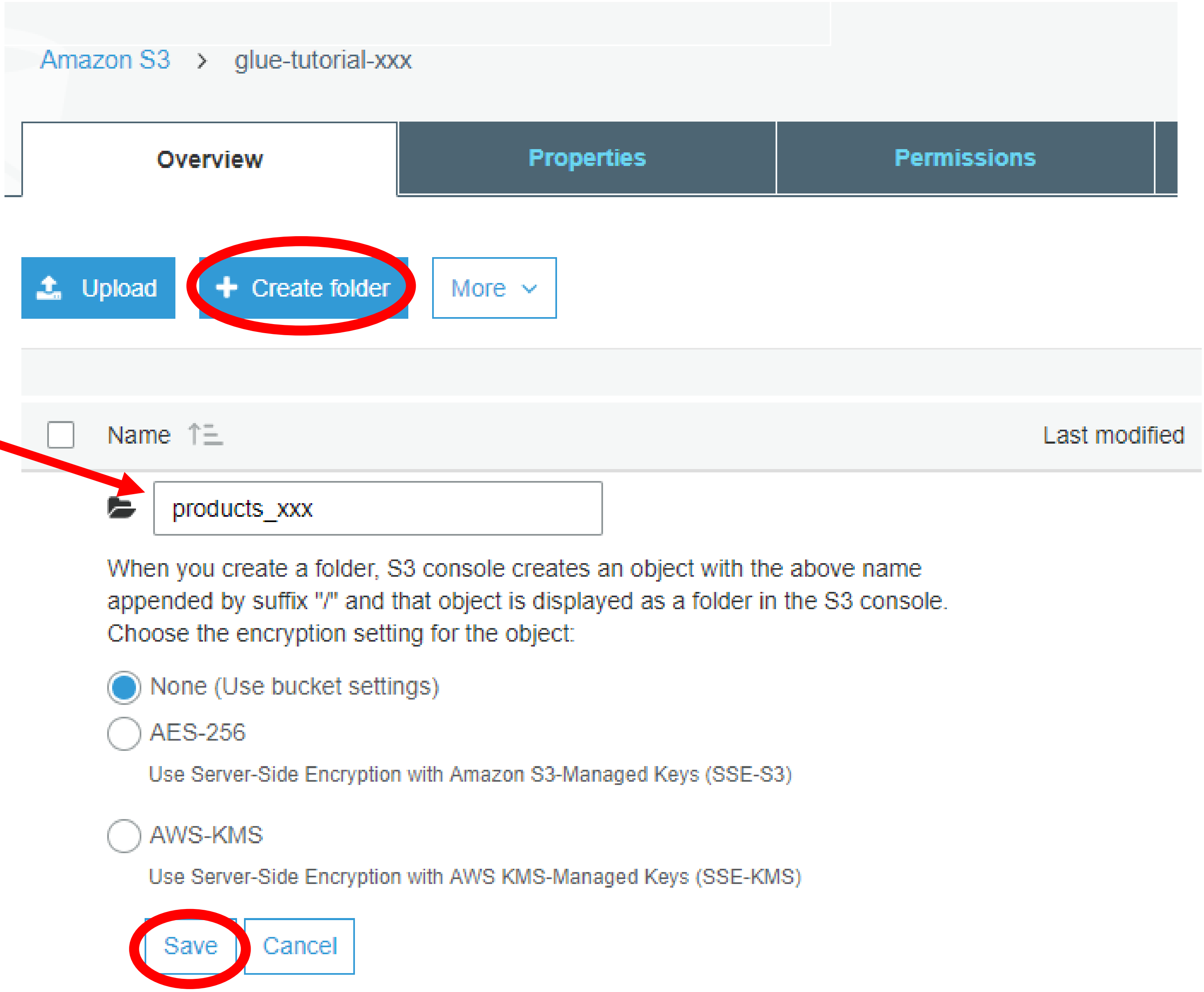


Bucket name	Access	Region
glue-tutorial-xxx	Not public *	US East (Ohio)

S3

└─**Create S3 bucket folder**

Create a folder called products_XXX



products_xxx

When you create a folder, S3 console creates an object with the above name appended by suffix "/" and that object is displayed as a folder in the S3 console. Choose the encryption setting for the object:

● None (Use bucket settings)

○ AES-256

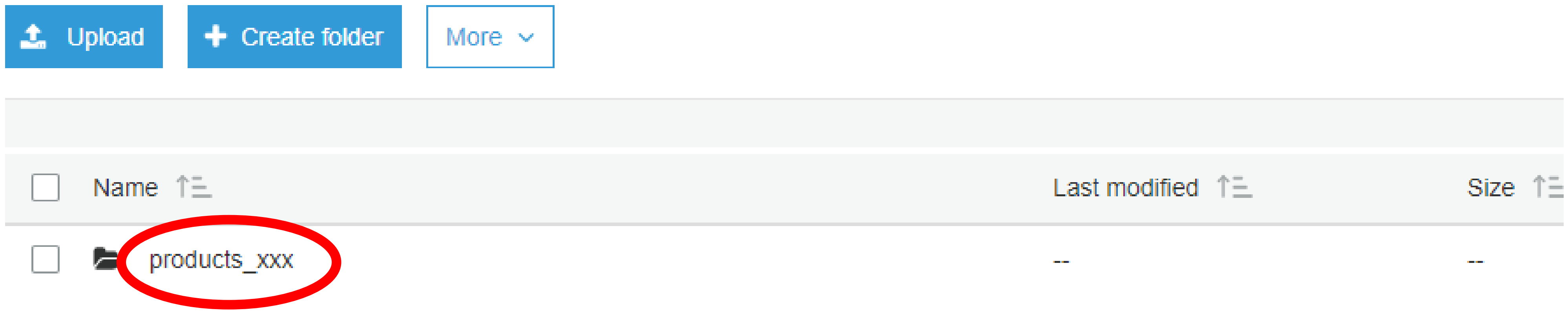
Use Server-Side Encryption with Amazon S3-Managed Keys (SSE-S3)

○ AWS-KMS

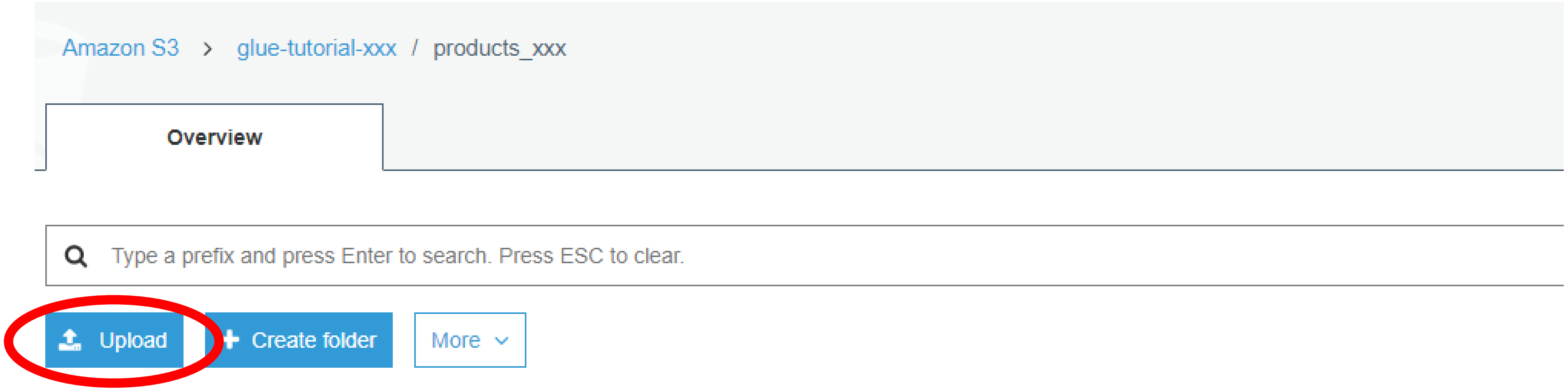
Use Server-Side Encryption with AWS KMS-Managed Keys (SSE-KMS)

S3

Create S3 bucket with AWS Console

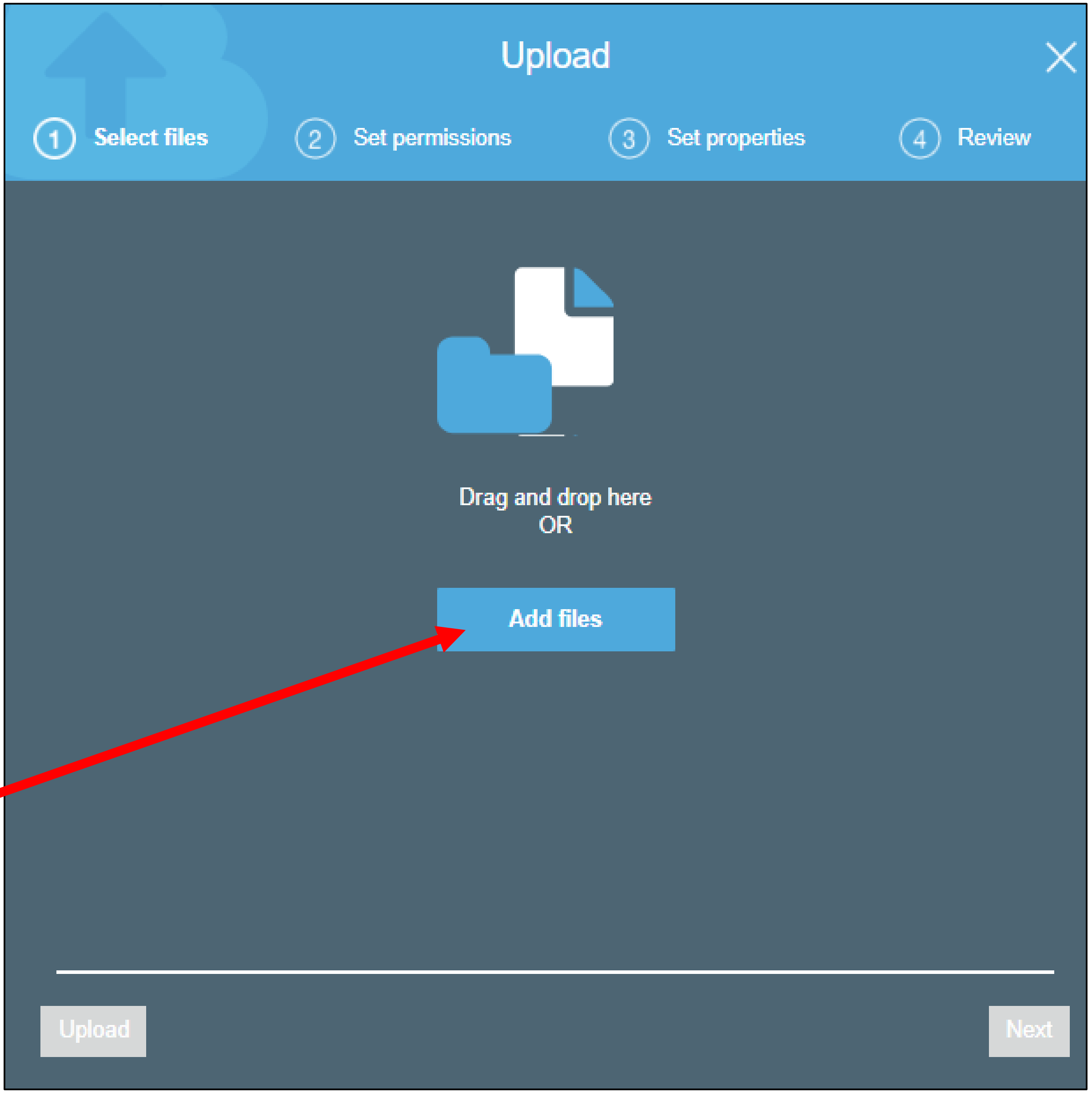


□	Name	Last modified	Size
□	products_xxx	--	--


Amazon S3 › glue-tutorial-xxx / products_xxx
Overview
Type a prefix and press Enter to search. Press ESC to clear.
Upload
Create folder
More

S3

└─Add file to S3 bucket with AWS Console

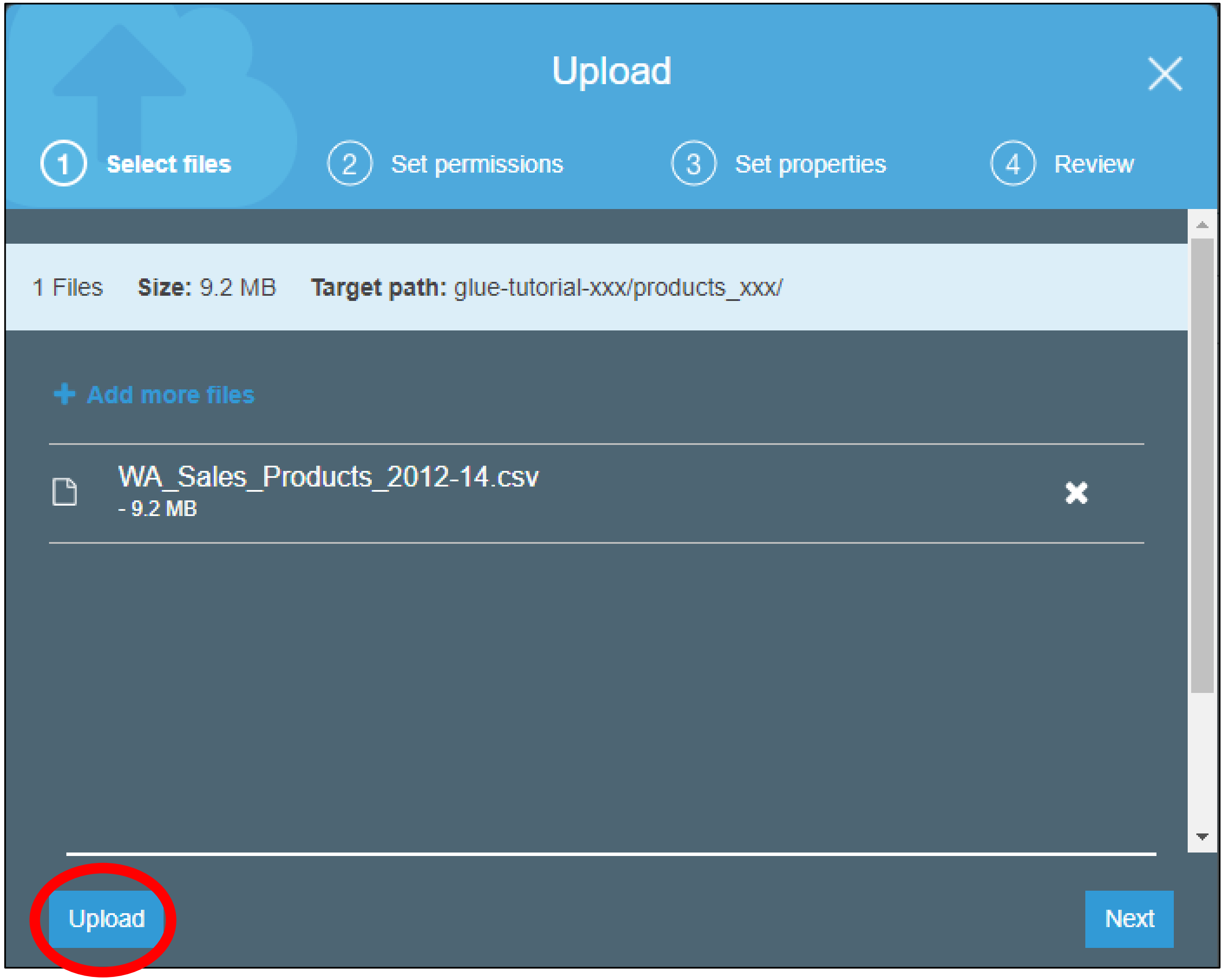


Add file from repository called "WA_Sales_Products_2012-14"

S3

└─**Add file to S3 bucket with AWS Console**

Add file from repository called "WA_Sales_Products_2012-14"



Upload

1 Files **Size:** 9.2 MB **Target path:** glue-tutorial-xxx/products_xxx/

WA_Sales_Products_2012-14.csv
- 9.2 MB

S3

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

```
$ aws s3 cp <your-file-path>/aws-glue-tutorial/WA_Sales_Products_2012-14.csv s3://glue-tutorial-XXX/products_XXX/
```

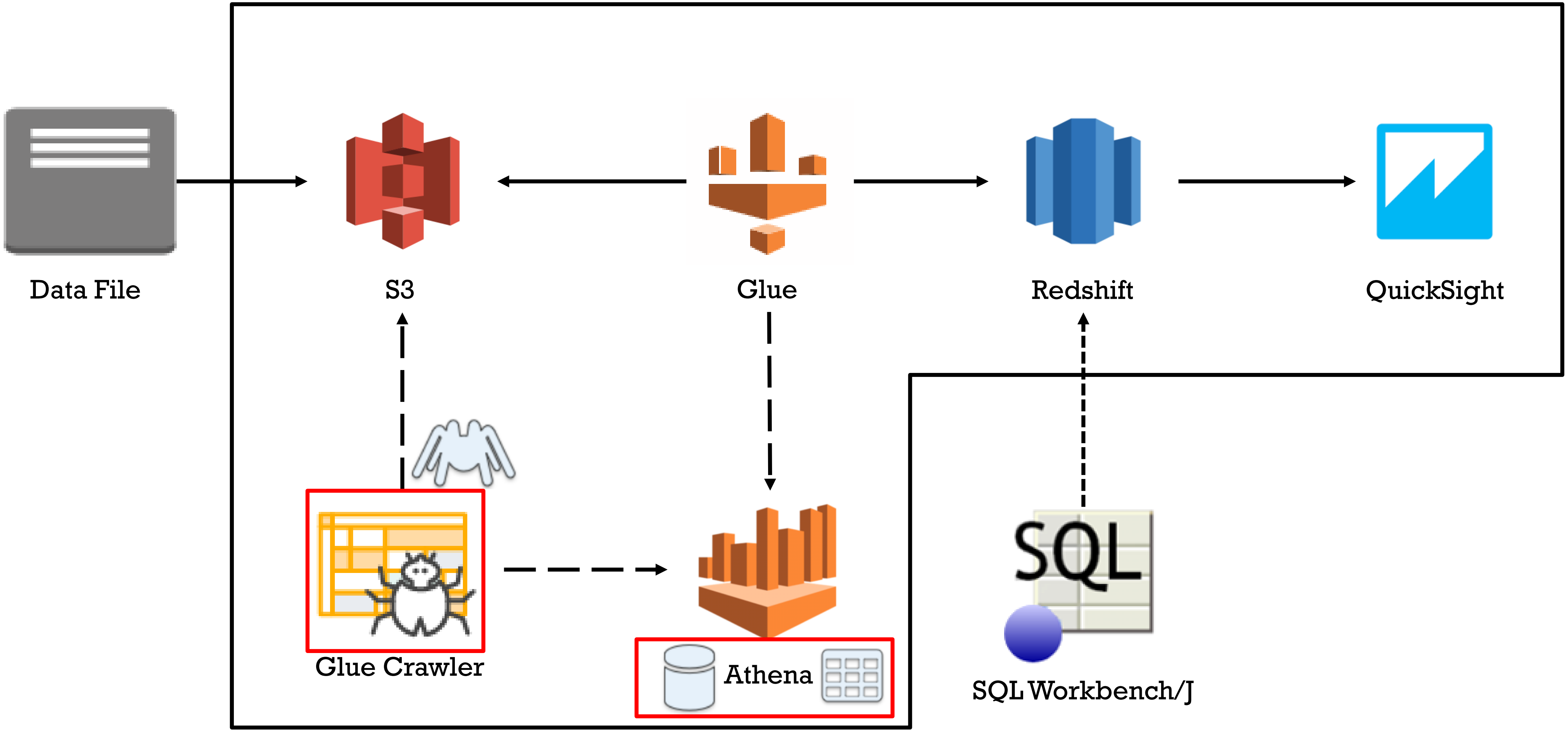
* Must install and set up AWS CLI in order to use this

Lab 2

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

(Use US-EAST-2/Ohio Region)

Glue Crawler

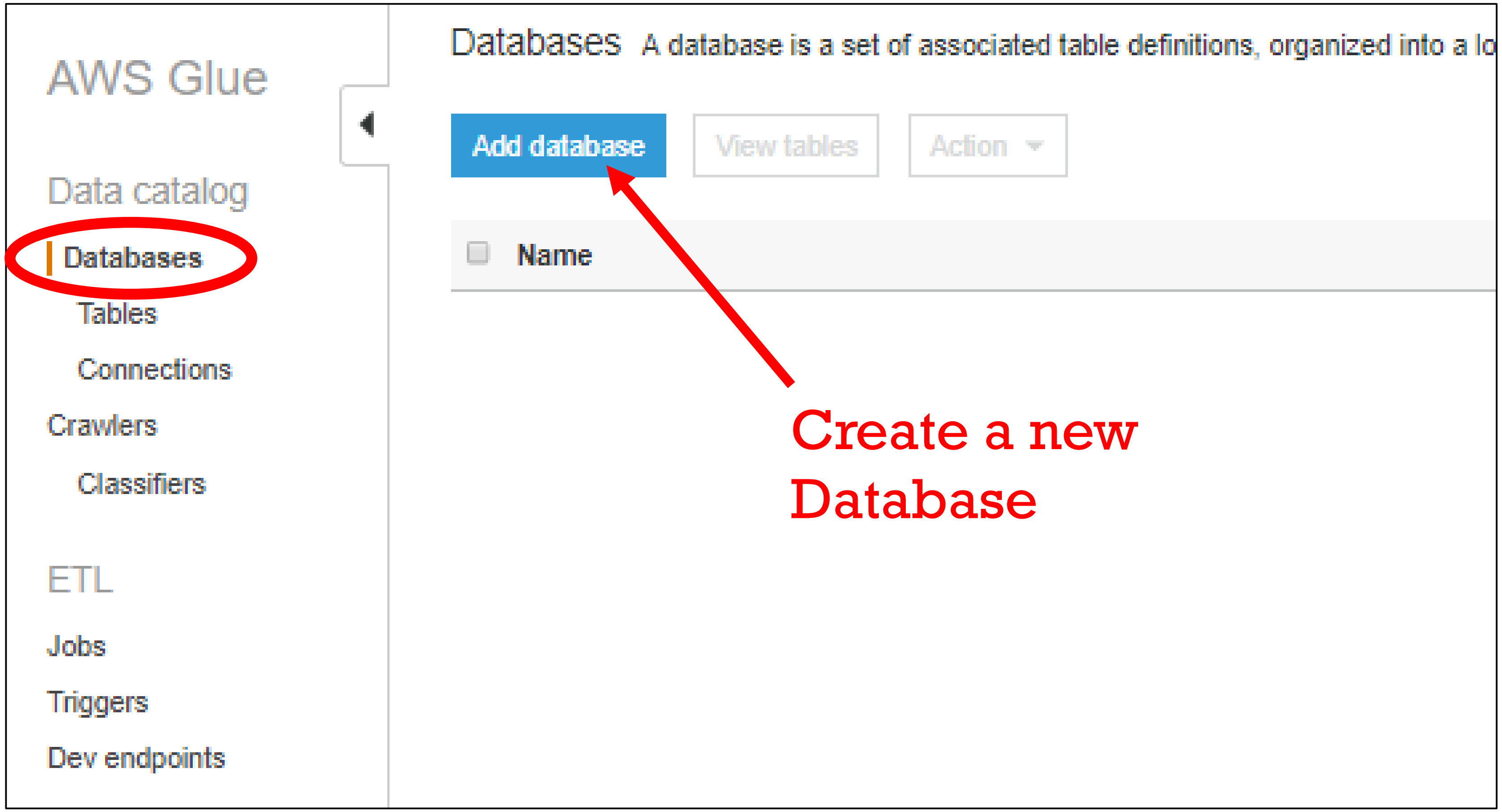


Glue Crawler

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

Glue

Create Glue Database



AWS Glue

Data catalog

Databases

Tables

Connections

Crawlers

Classifiers

ETL

Jobs

Triggers

Dev endpoints

Databases A database is a set of associated table definitions, organized into a lo

Add database View tables Action ▾

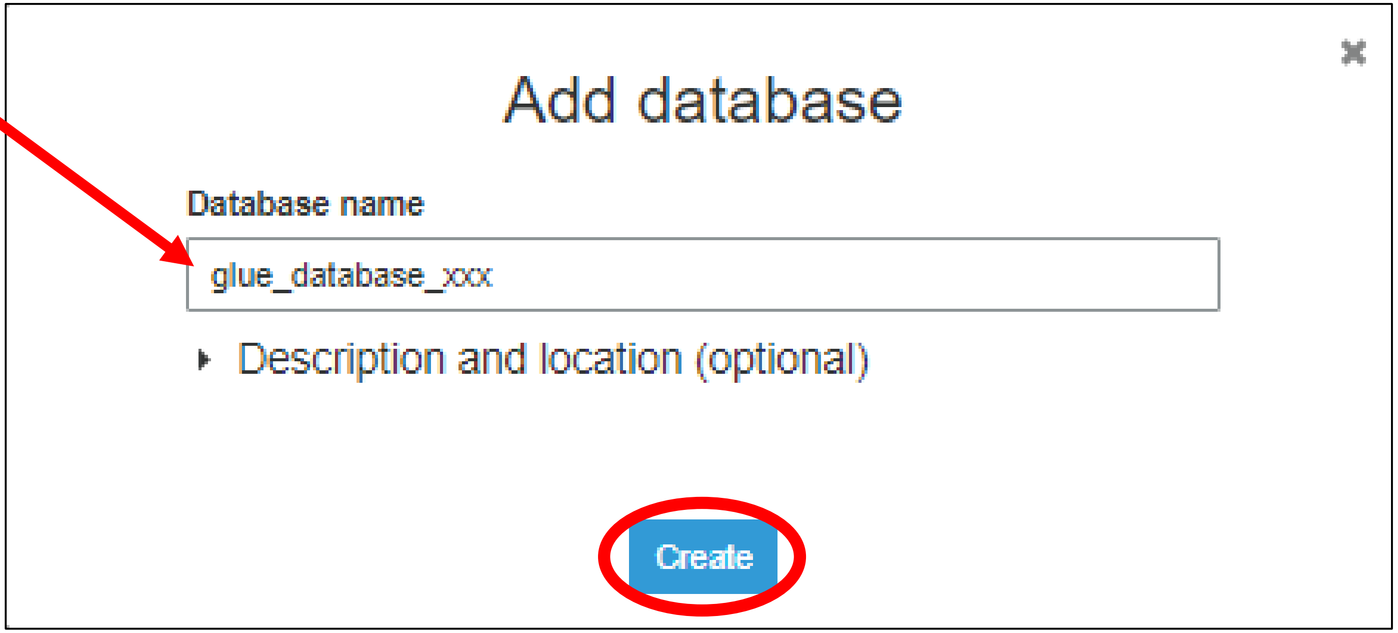
□ Name

Create a new Database

Glue

└─Create Glue Database

Give your database a name “glue_database_XXX”

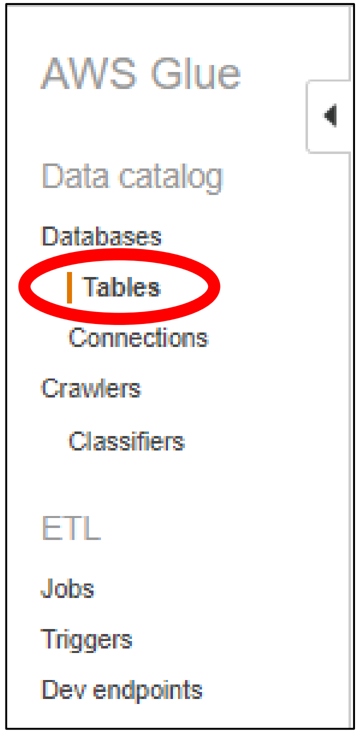


Add database

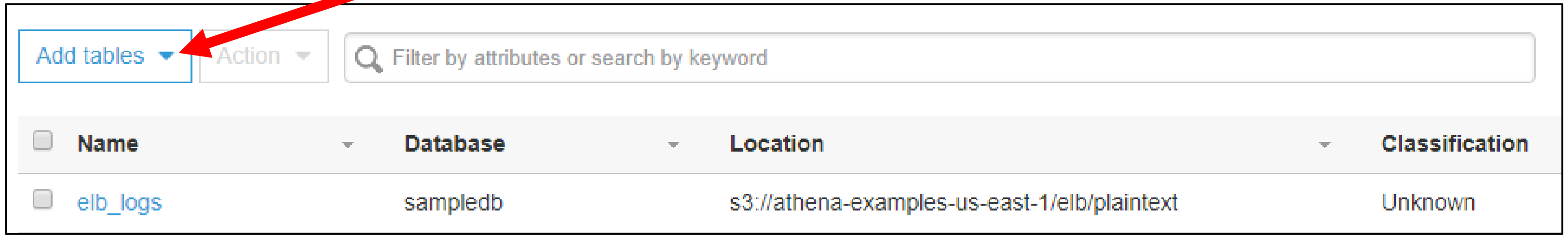
Database name: glue_database_xxx

Glue Crawler

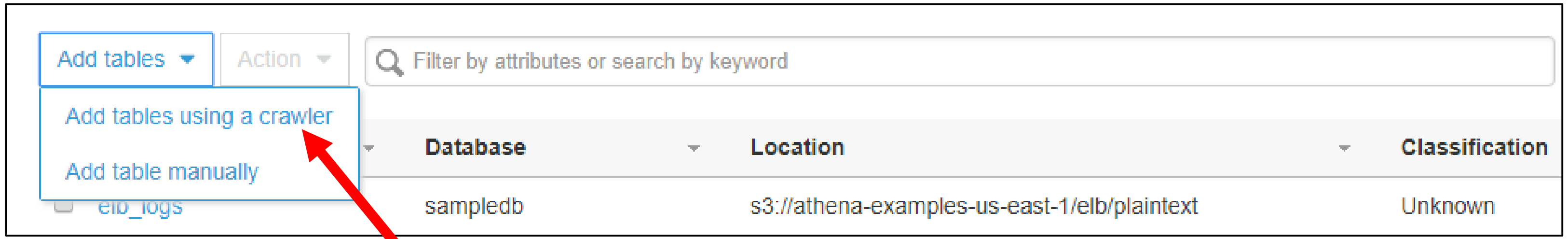
Create Table with Glue Crawler



Click on add tables to create a table



Name	Database	Location	Classification
elb_logs	sampledb	s3://athena-examples-us-east-1/elb/plaintext	Unknown


Add tables
Action
Filter by attributes or search by keyword
Add tables using a crawler
Add table manually

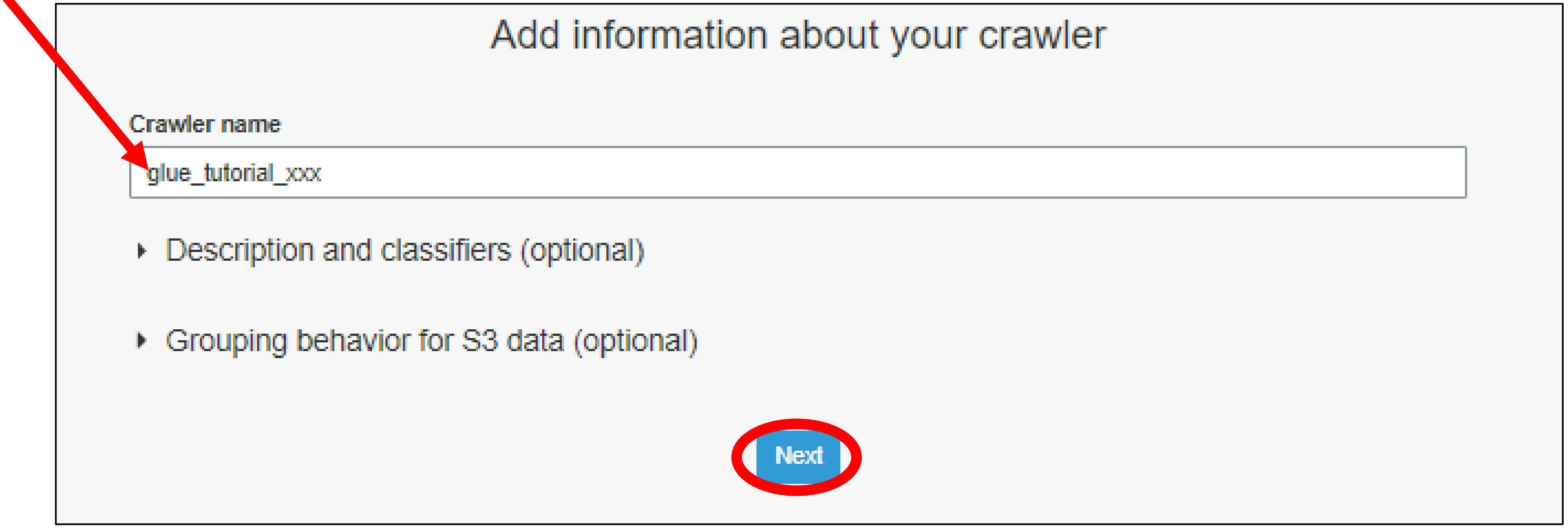
Name	Database	Location	Classification
elb_logs	sampledb	s3://athena-examples-us-east-1/elb/plaintext	Unknown

Create a table using a crawler

Glue Crawler

└─Create Table with Glue Crawler

Give your crawler a name, glue-tutorial-XXX



Add information about your crawler

Crawler name

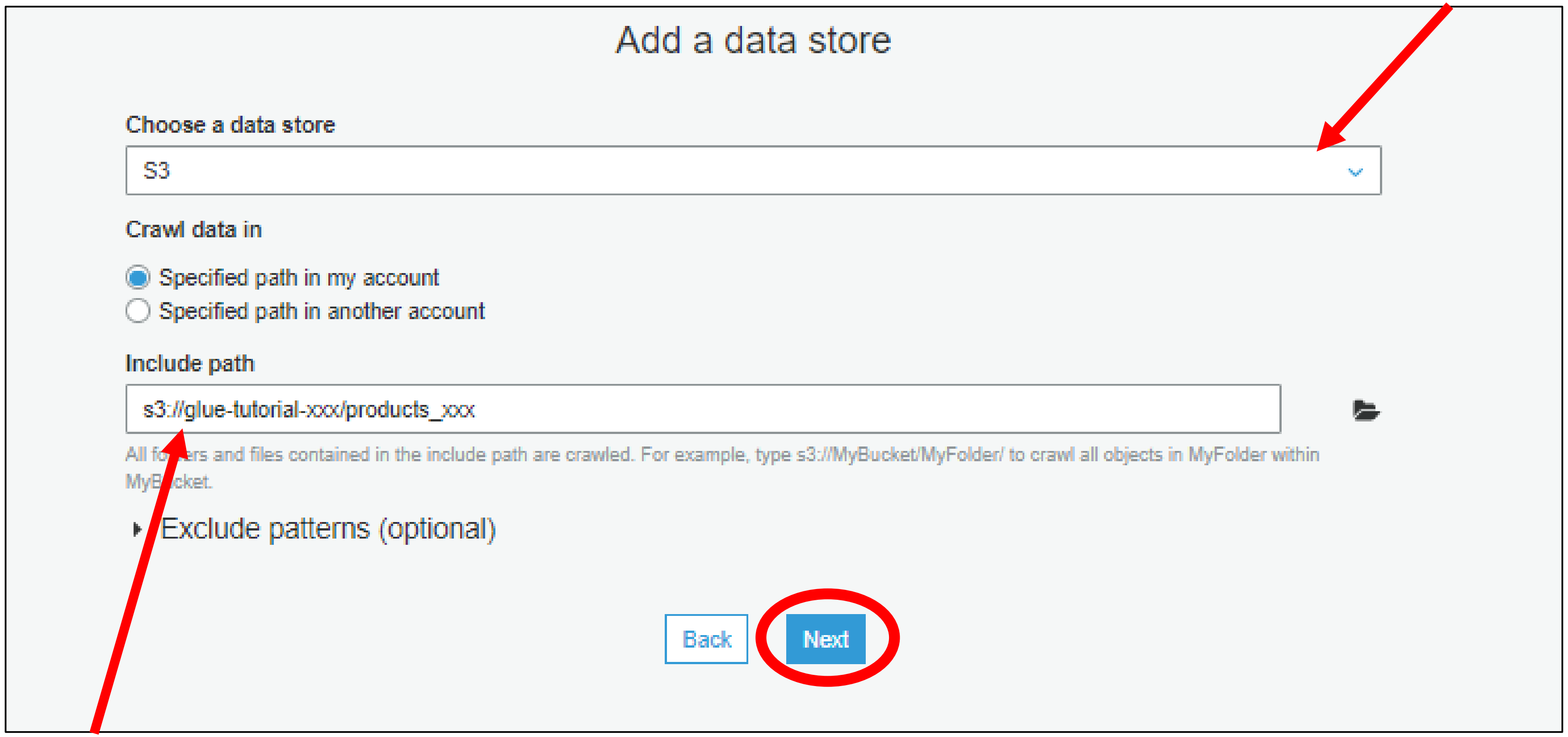
glue_tutorial_xxx

▸ Description and classifiers (optional)
▸ Grouping behavior for S3 data (optional)
Next

Glue Crawler

└─Create Table with Glue Crawler

Choose where the table is going to look for data



Add a data store

Choose a data store

S3

Crawl data in

● Specified path in my account
○ Specified path in another account

Include path

s3://glue-tutorial-xxx/products_xxx

All folders and files contained in the include path are crawled. For example, type s3://MyBucket/MyFolder/ to crawl all objects in MyFolder within MyBucket.

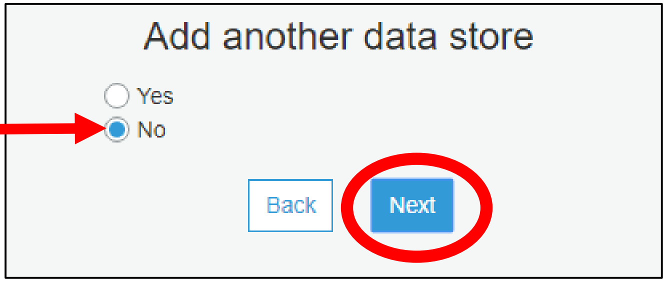
▸ Exclude patterns (optional)

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



Add another data store

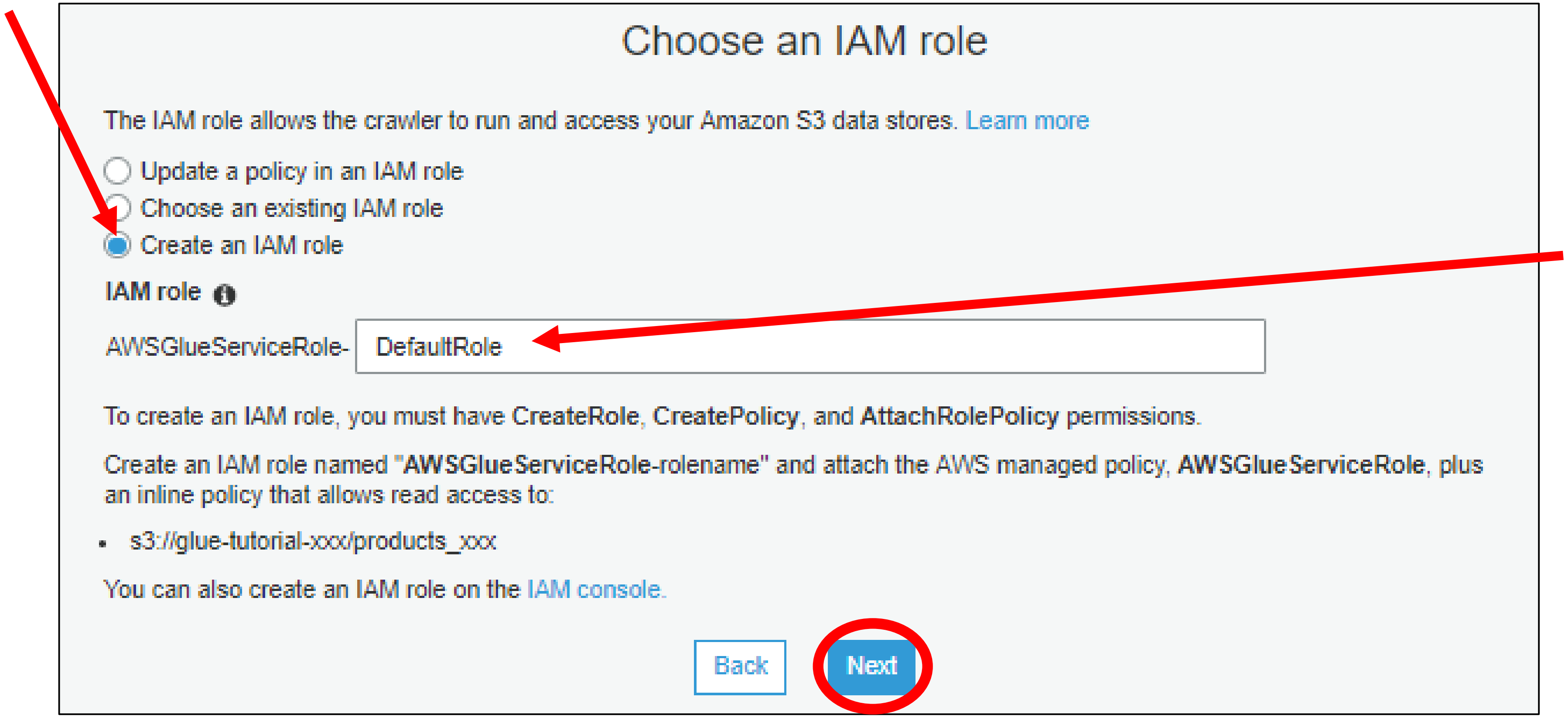
○ Yes

● No

Glue Crawler

└─Create Table with Glue Crawler

Need to create role to access S3 bucket



Choose an IAM role

The IAM role allows the crawler to run and access your Amazon S3 data stores. Learn more

○ Update a policy in an IAM role
○ Choose an existing IAM role
● Create an IAM role

IAM role

AWSGlueServiceRole- DefaultRole

To create an IAM role, you must have **CreateRole**, **CreatePolicy**, and **AttachRolePolicy** permissions.

Create an IAM role named "**AWSGlueServiceRole**-rolename" and attach the AWS managed policy, **AWSGlueServiceRole**, plus an inline policy that allows read access to:

- s3://glue-tutorial-xxx/products_xxx

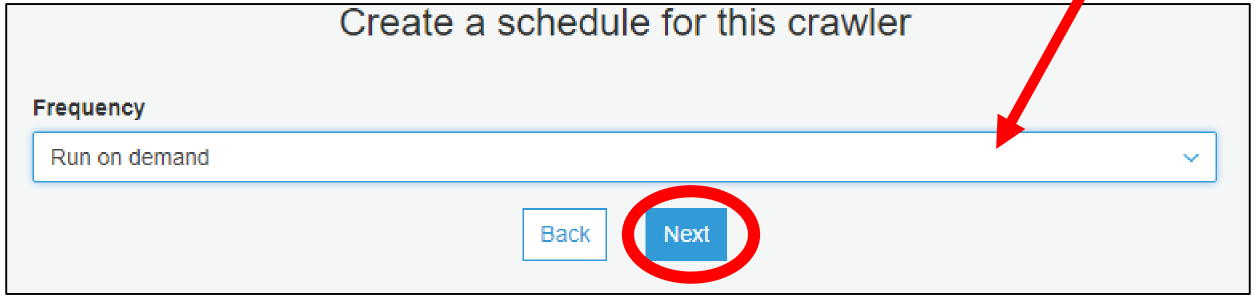
You can also create an IAM role on the IAM console.

Give your role a name

Glue Crawler

└─Create Table with Glue Crawler

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



Create a schedule for this crawler

Frequency

Run on demand

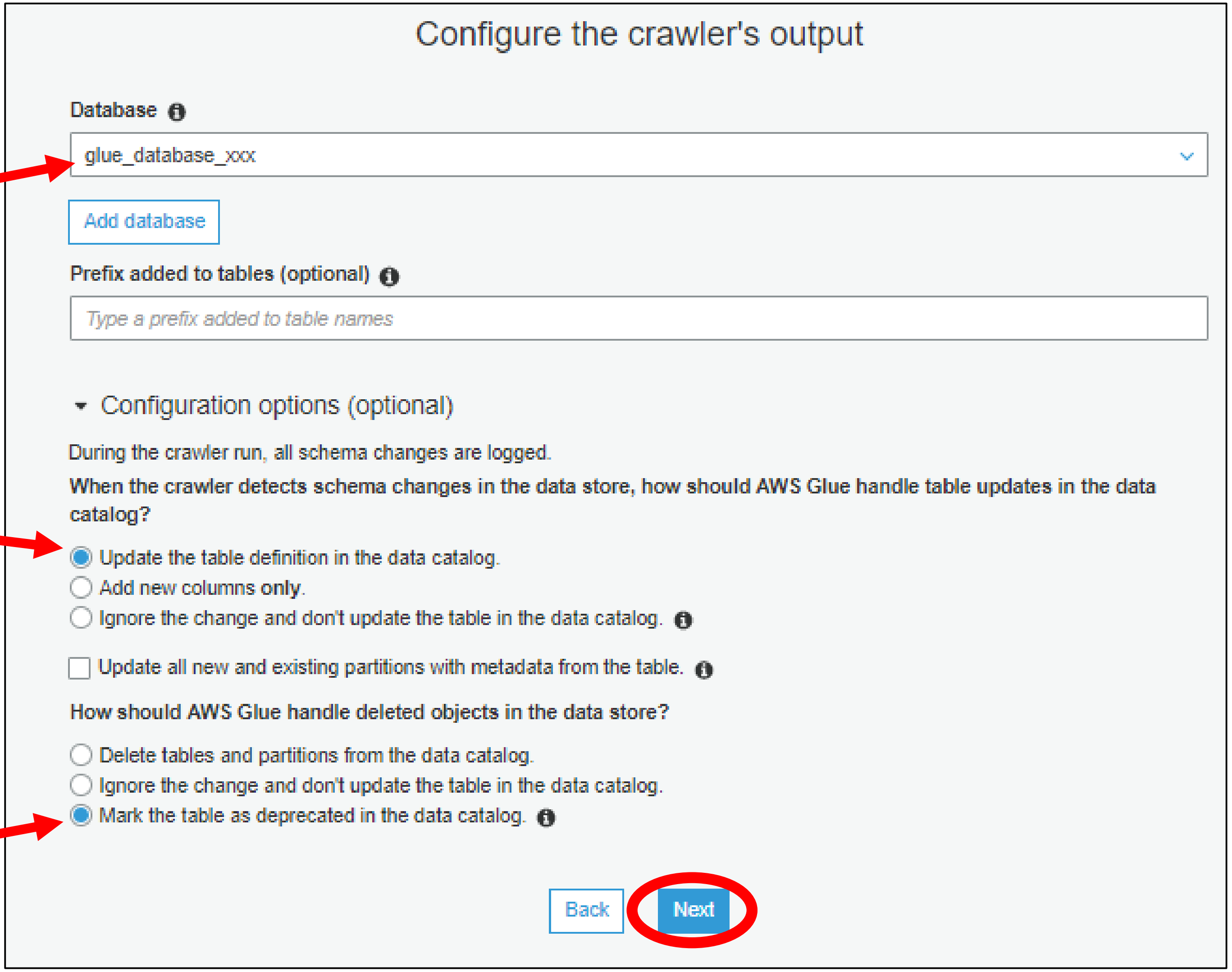
Glue Crawler

└─Create Table with Glue Crawler

Choose the database you created for the database your table will live in

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

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



Configure the crawler's output

Database

glue_database_xxx

Prefix added to tables (optional)

Type a prefix added to table names

Configuration options (optional)

During the crawler run, all schema changes are logged.

When the crawler detects schema changes in the data store, how should AWS Glue handle table updates in the data catalog?

● Update the table definition in the data catalog.
○ Add new columns only.
○ Ignore the change and don't update the table in the data catalog.

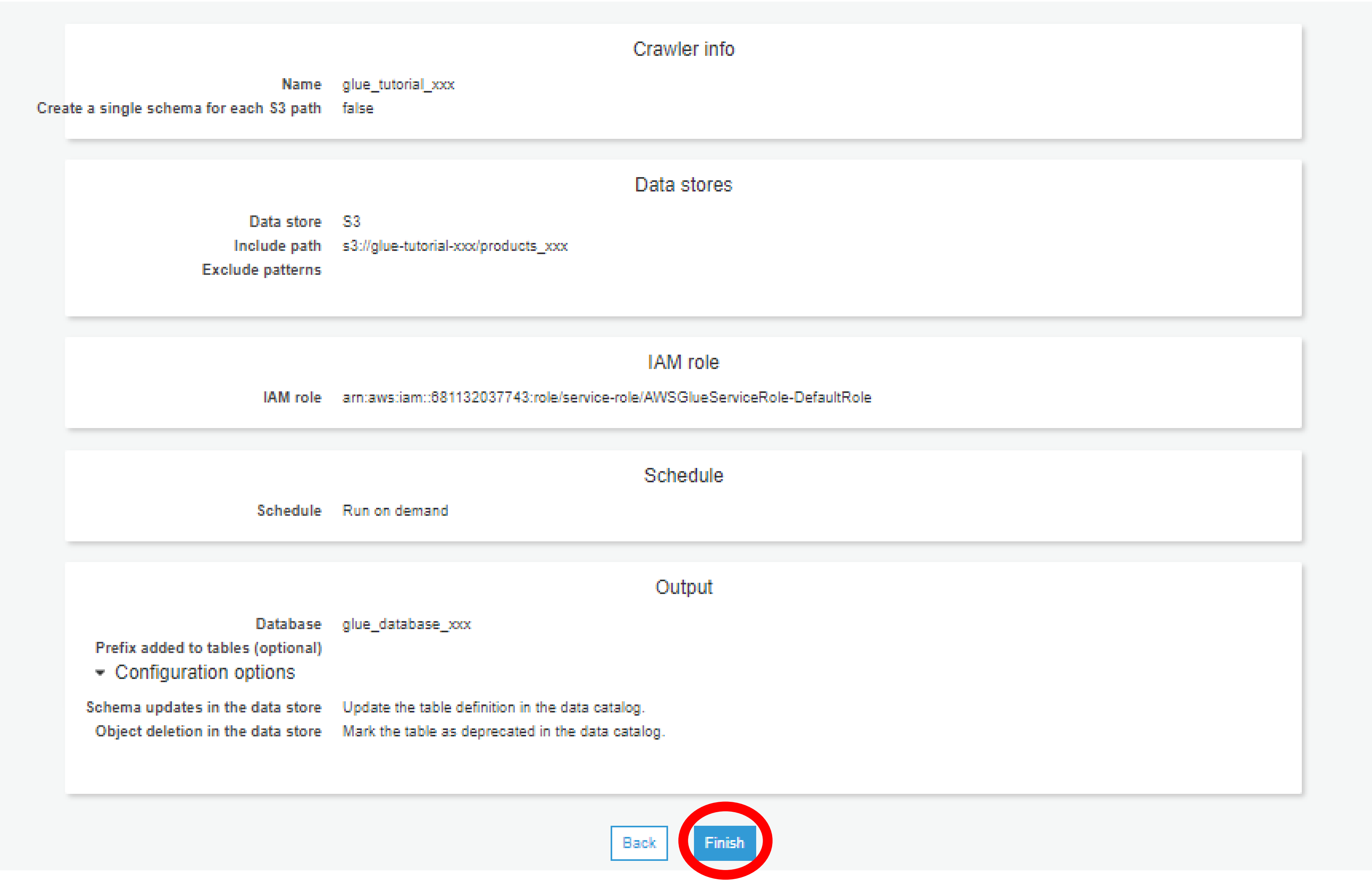
□ Update all new and existing partitions with metadata from the table.

How should AWS Glue handle deleted objects in the data store?

○ Delete tables and partitions from the data catalog.
○ Ignore the change and don't update the table in the data catalog.
● Mark the table as deprecated in the data catalog.

Glue Crawler

└─Create Table with Glue Crawler



Crawler info

Name glue_tutorial_xxx
Create a single schema for each S3 path false

Data stores

Data store S3
Include path s3://glue-tutorial-xxx/products_xxx
Exclude patterns

IAM role

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

Schedule

Schedule Run on demand

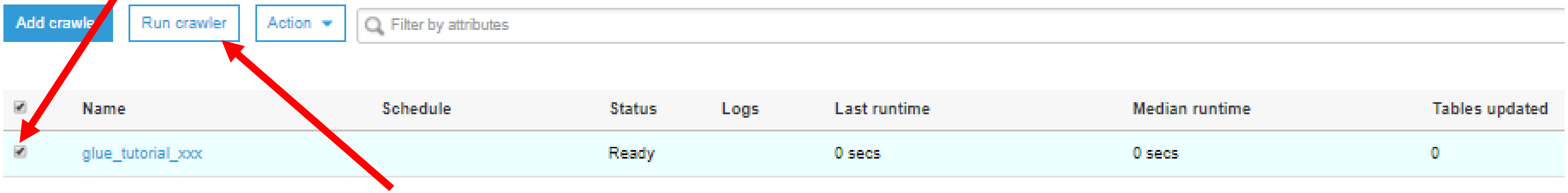
Output

Database glue_database_xxx
Prefix added to tables (optional)
▾ 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.

Glue Crawler

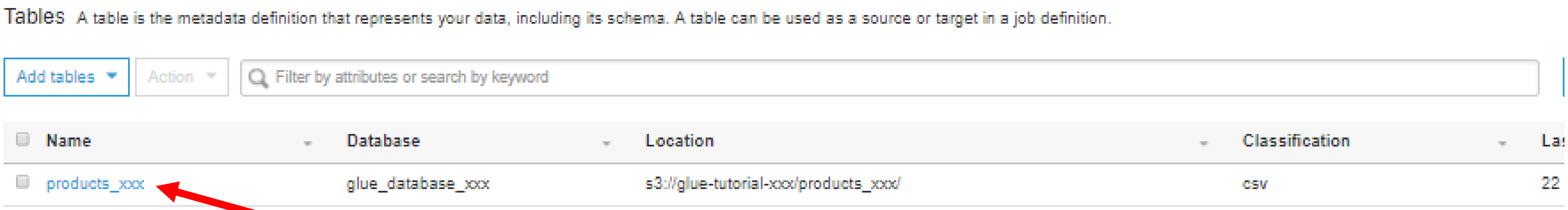
Create Table with Glue Crawler

Select your crawler



☑	Name	Schedule	Status	Logs	Last runtime	Median runtime	Tables updated
☑	glue_tutorial_xxx		Ready		0 secs	0 secs	0

Run your crawler

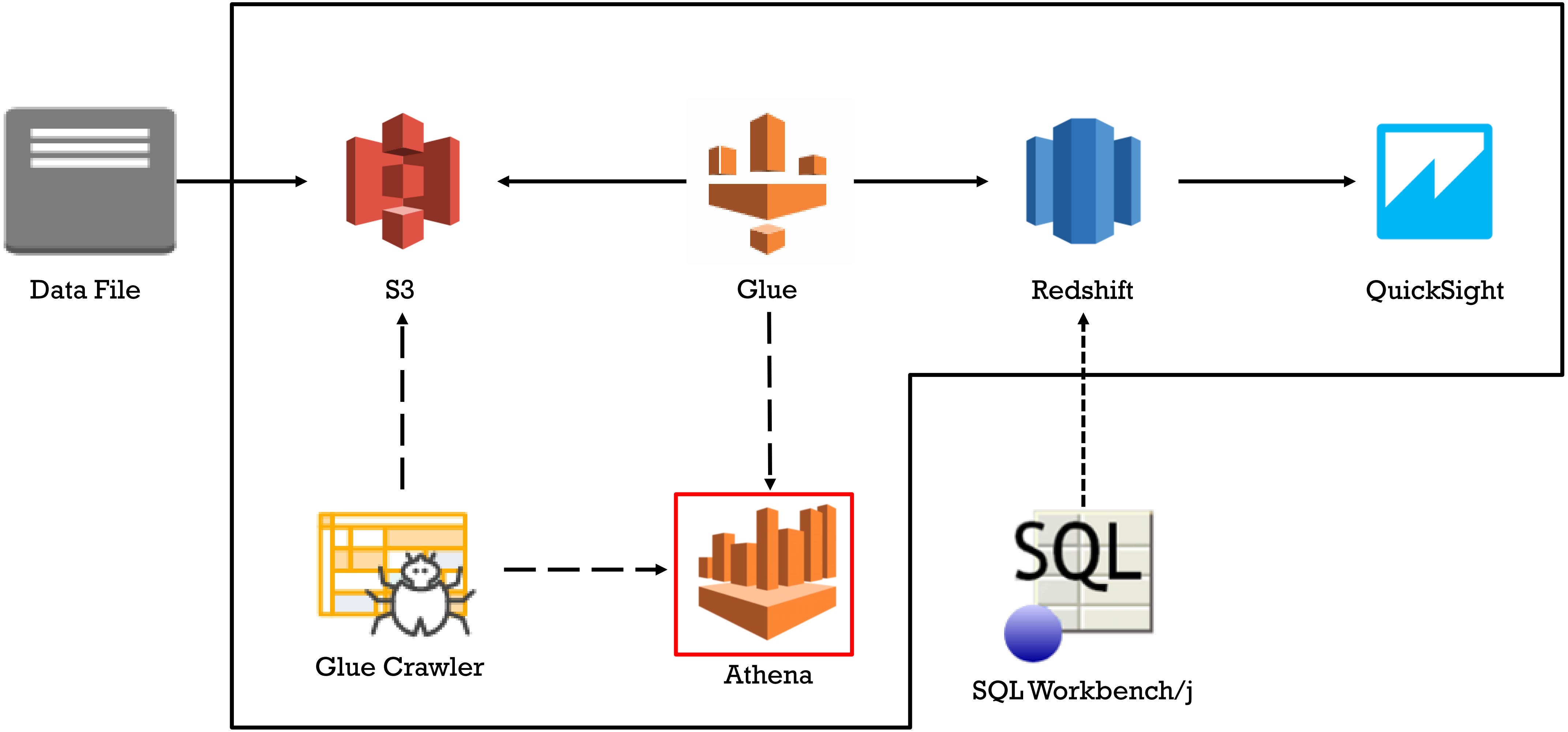


Tables A table is the metadata definition that represents your data, including its schema. A table can be used as a source or target in a job definition.

□	Name	Database	Location	Classification	La
□	products_xxx	glue_database_xxx	s3://glue-tutorial-xxx/products_xxx/	csv	22

Your table should be in the Tables tab

Athena

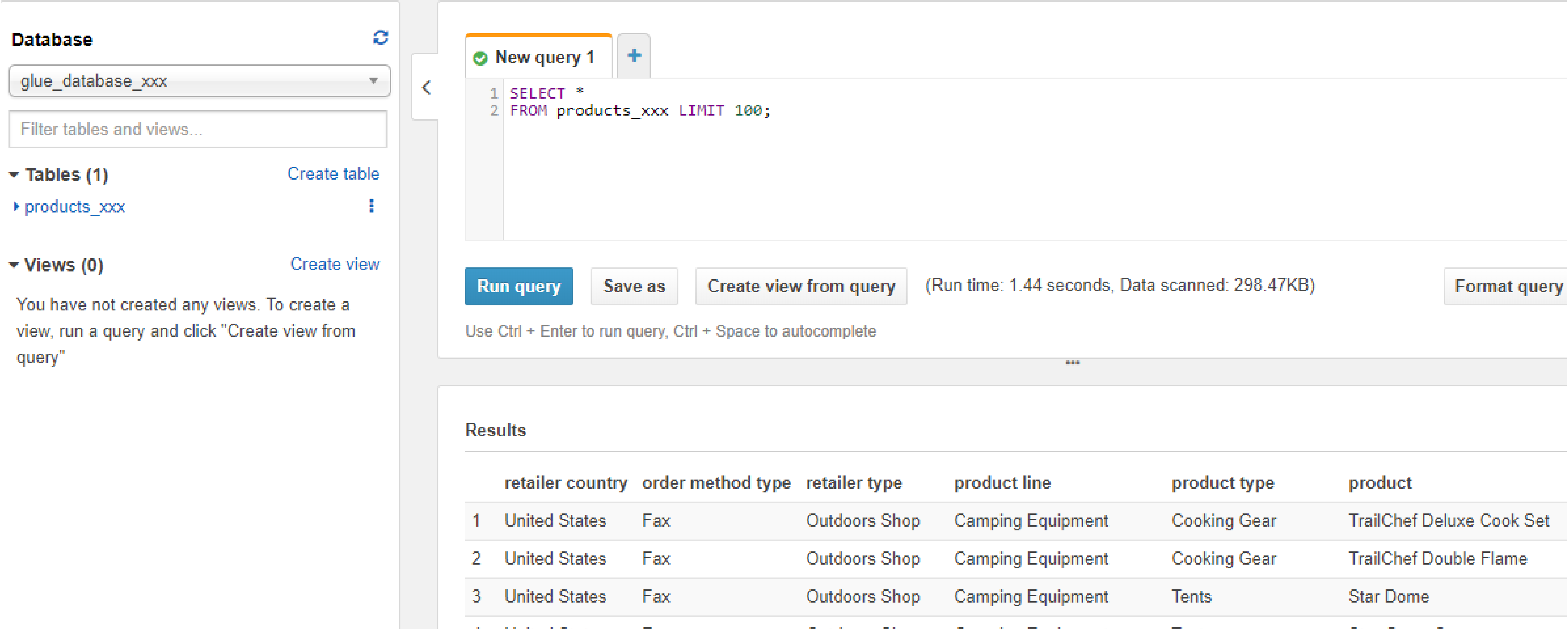

Data File
S3
Glue
Redshift
QuickSight
Glue Crawler
Athena
SQL
SQL Workbench/j

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;



You have not created any views. To create a view, run a query and click "Create view from query"

```
SELECT *
FROM products_xxx LIMIT 100;
```

(Run time: 1.44 seconds, Data scanned: 298.47KB)

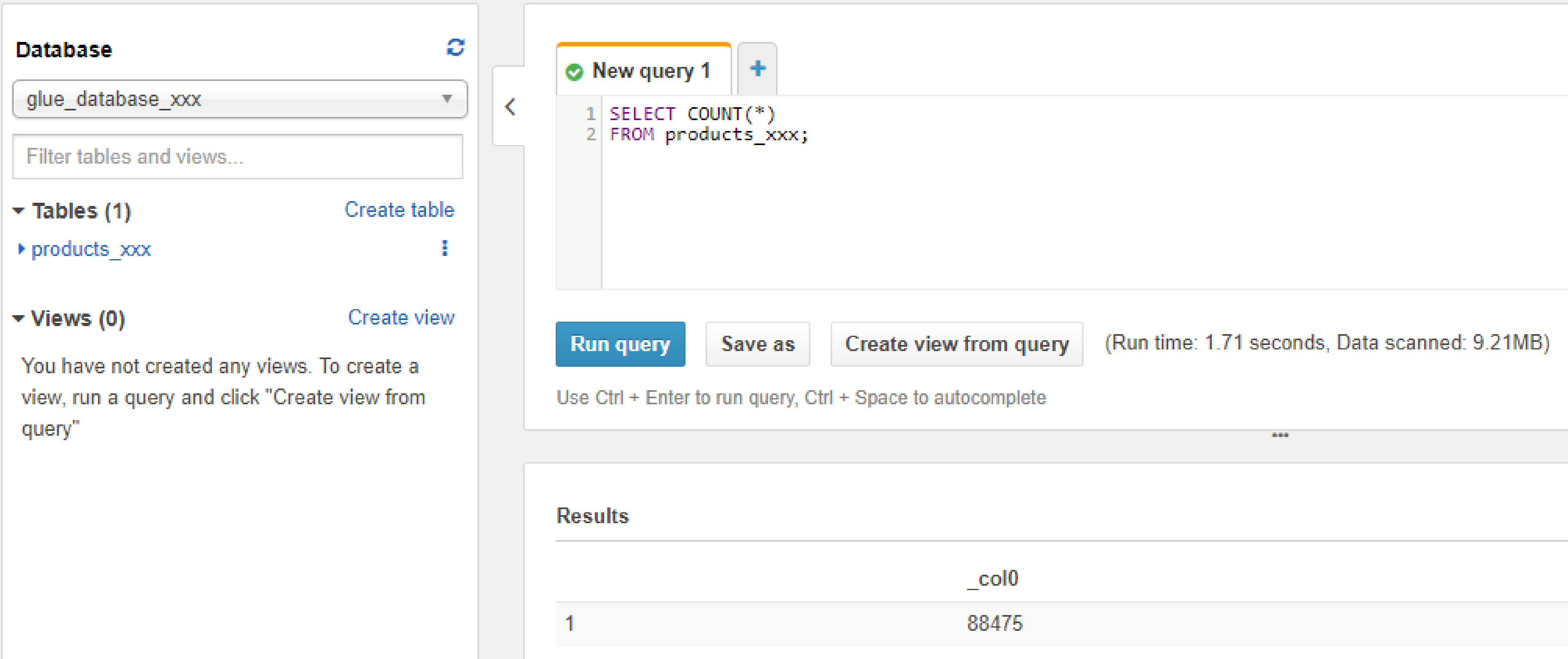
Use Ctrl + Enter to run query, Ctrl + Space to autocomplete

Results

	retailer country	order method type	retailer type	product line	product type	product
1	United States	Fax	Outdoors Shop	Camping Equipment	Cooking Gear	TrailChef Deluxe Cook Set
2	United States	Fax	Outdoors Shop	Camping Equipment	Cooking Gear	TrailChef Double Flame
3	United States	Fax	Outdoors Shop	Camping Equipment	Tents	Star Dome

Athena

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



Database: glue_database_xxx

You have not created any views. To create a view, run a query and click "Create view from query"

```
SELECT COUNT(*)
FROM products_xxx;
```

(Run time: 1.71 seconds, Data scanned: 9.21MB)

Use Ctrl + Enter to run query, Ctrl + Space to autocomplete

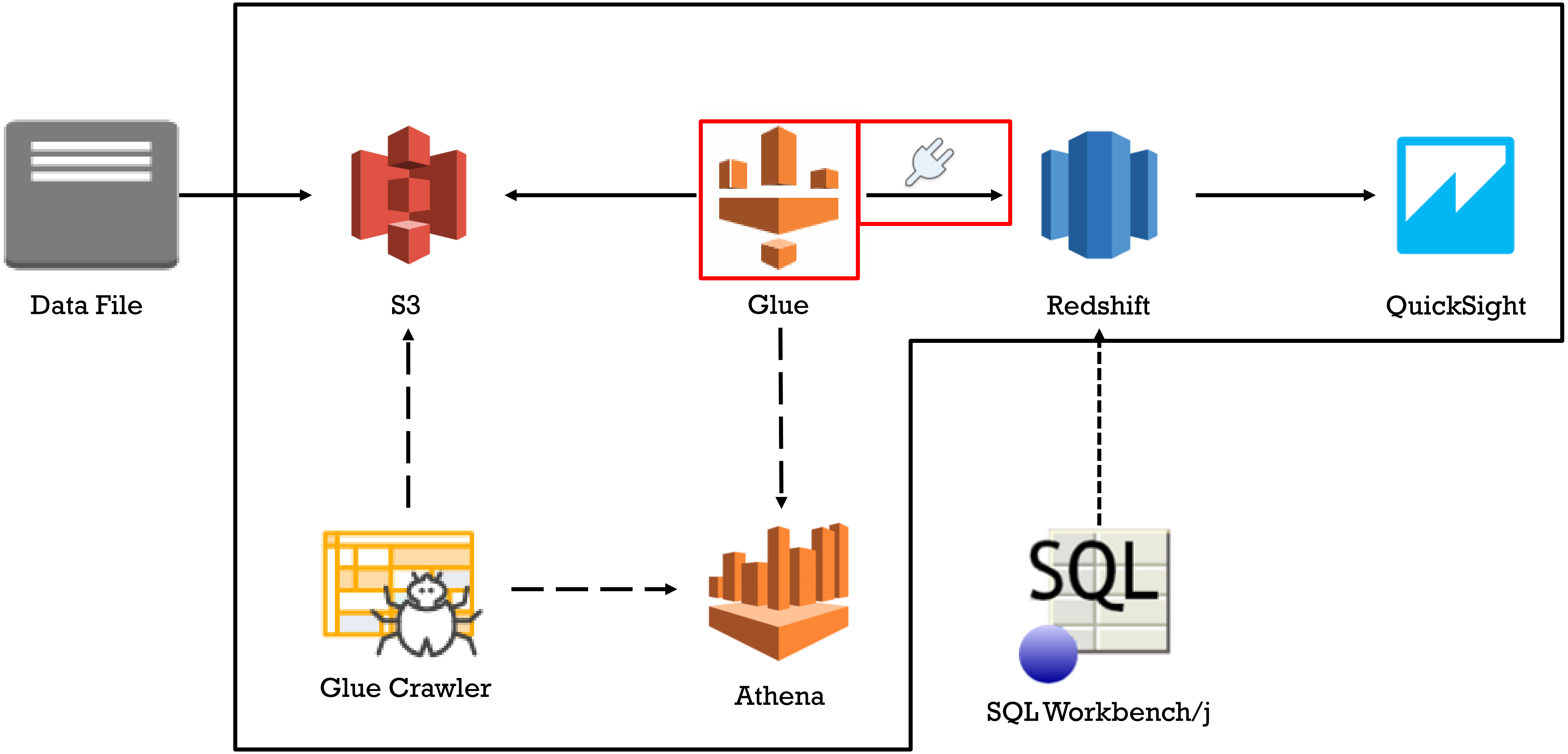
	_col0
1	88475

Lab 3

- Create/Run Glue Crawler
- Query Athena

(Use US-EAST-2/Ohio Region)

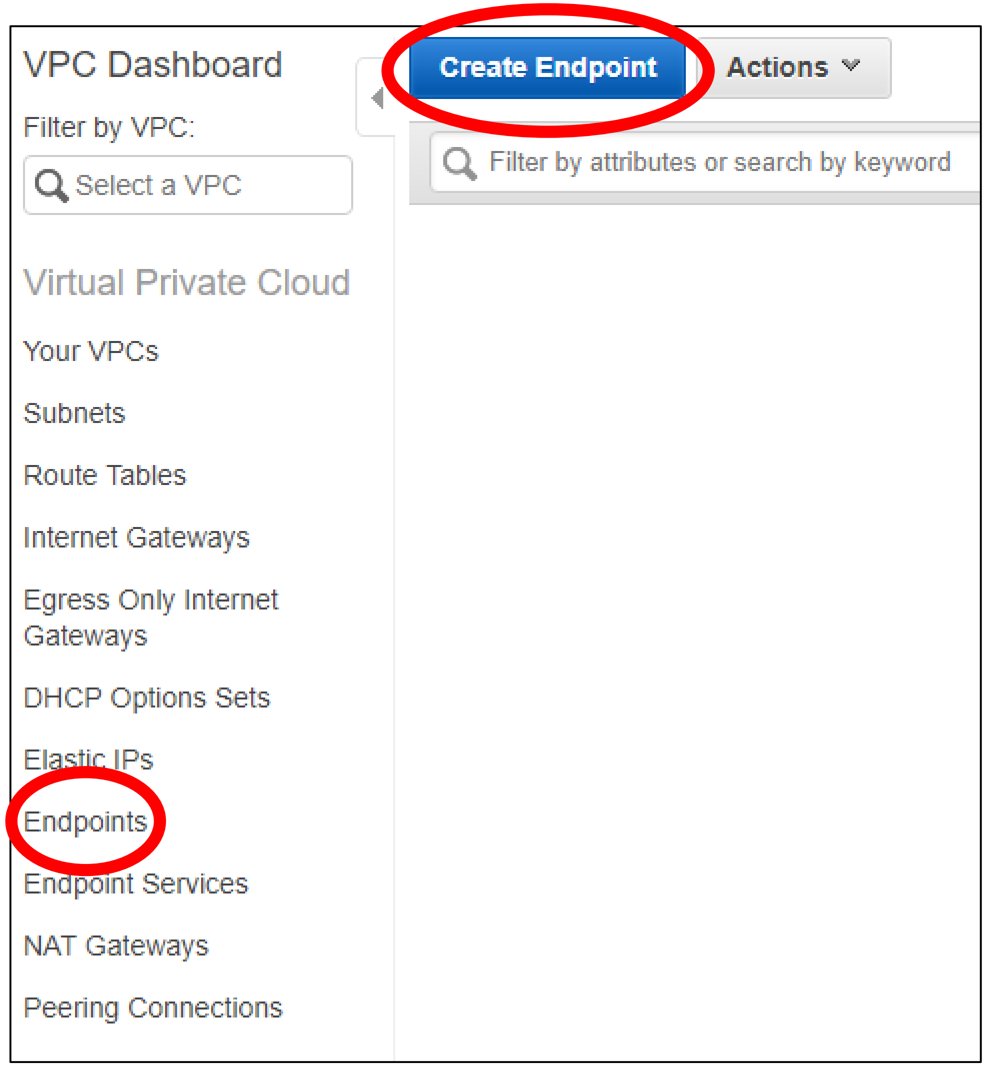
Glue


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

VPC

└─Create a S3 endpoint

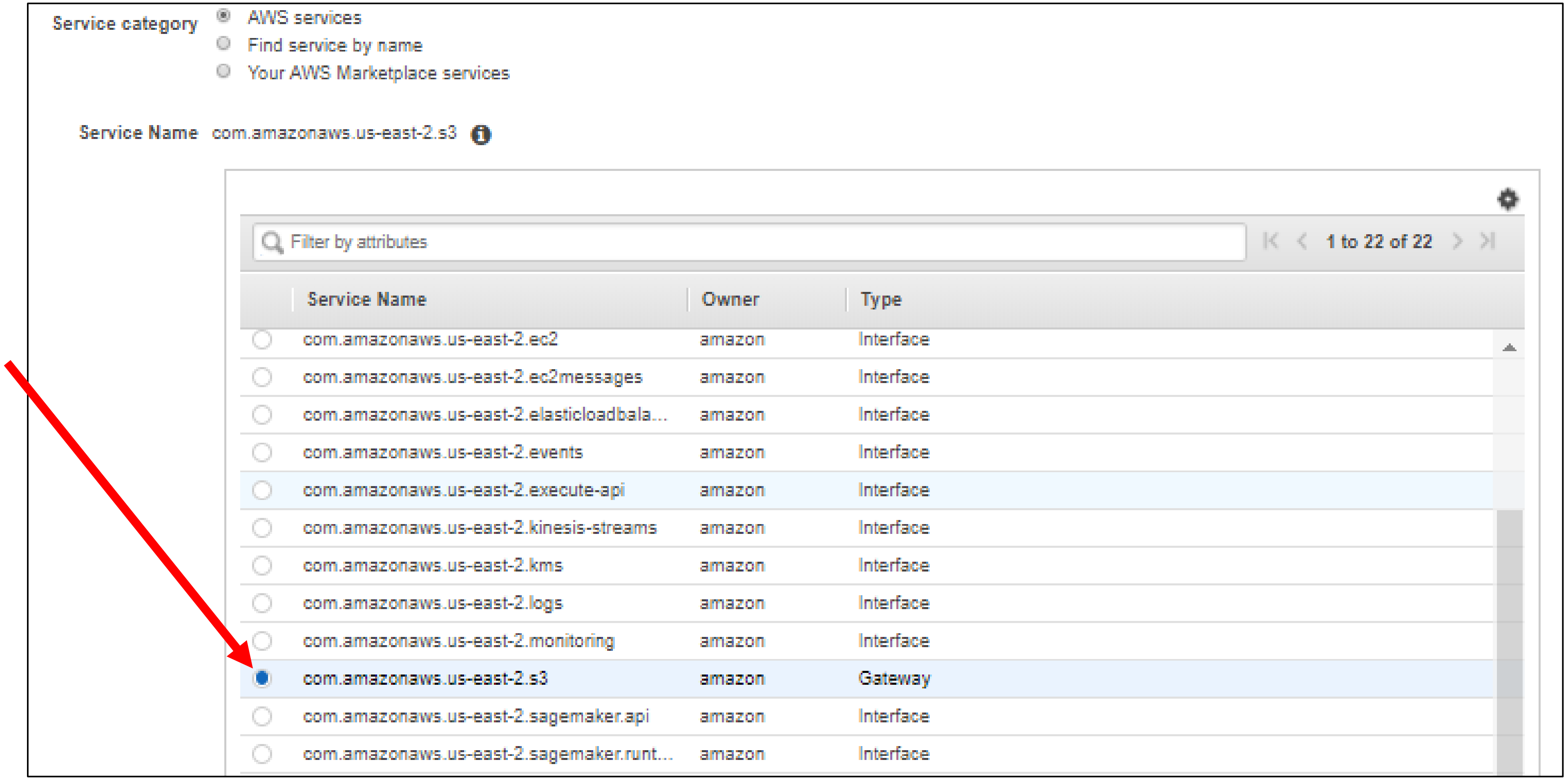
We need to create a S3 endpoint for Glue to access S3



VPC

└─Create a S3 endpoint

Select the S3 Service for Glue to access S3



Service category ● AWS services
○ Find service by name
○ Your AWS Marketplace services

Service Name com.amazonaws.us-east-2.s3

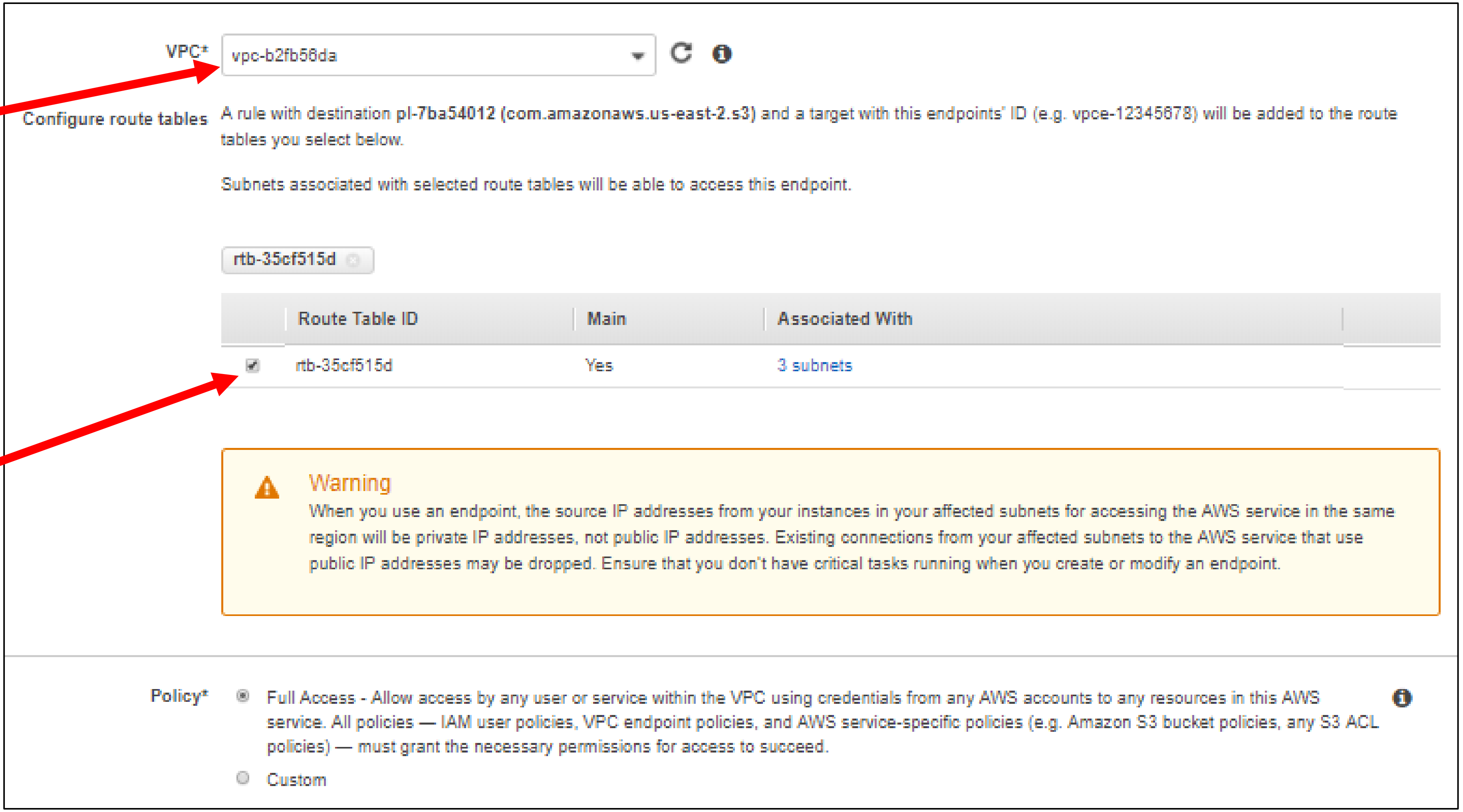
	Service Name	Owner	Type
○	com.amazonaws.us-east-2.ec2	amazon	Interface
○	com.amazonaws.us-east-2.ec2messages	amazon	Interface
○	com.amazonaws.us-east-2.elasticloadbala...	amazon	Interface
○	com.amazonaws.us-east-2.events	amazon	Interface
○	com.amazonaws.us-east-2.execute-api	amazon	Interface
○	com.amazonaws.us-east-2.kinesis-streams	amazon	Interface
○	com.amazonaws.us-east-2.kms	amazon	Interface
○	com.amazonaws.us-east-2.logs	amazon	Interface
○	com.amazonaws.us-east-2.monitoring	amazon	Interface
●	com.amazonaws.us-east-2.s3	amazon	Gateway
○	com.amazonaws.us-east-2.sagemaker.api	amazon	Interface
○	com.amazonaws.us-east-2.sagemaker.runt...	amazon	Interface

VPC

└─Create a S3 endpoint

Choose VPC

Choose to add to the Route Table



VPC* vpc-b2fb56da

Configure route tables: A rule with destination pl-7ba54012 (com.amazonaws.us-east-2.s3) and a target with this endpoints' ID (e.g. vpce-12345678) will be added to the route tables you select below.

Subnets associated with selected route tables will be able to access this endpoint.

rtb-35cf515d

	Route Table ID	Main	Associated With
☑	rtb-35cf515d	Yes	3 subnets

Warning
When you use an endpoint, the source IP addresses from your instances in your affected subnets for accessing the AWS service in the same region will be private IP addresses, not public IP addresses. Existing connections from your affected subnets to the AWS service that use public IP addresses may be dropped. Ensure that you don't have critical tasks running when you create or modify an endpoint.

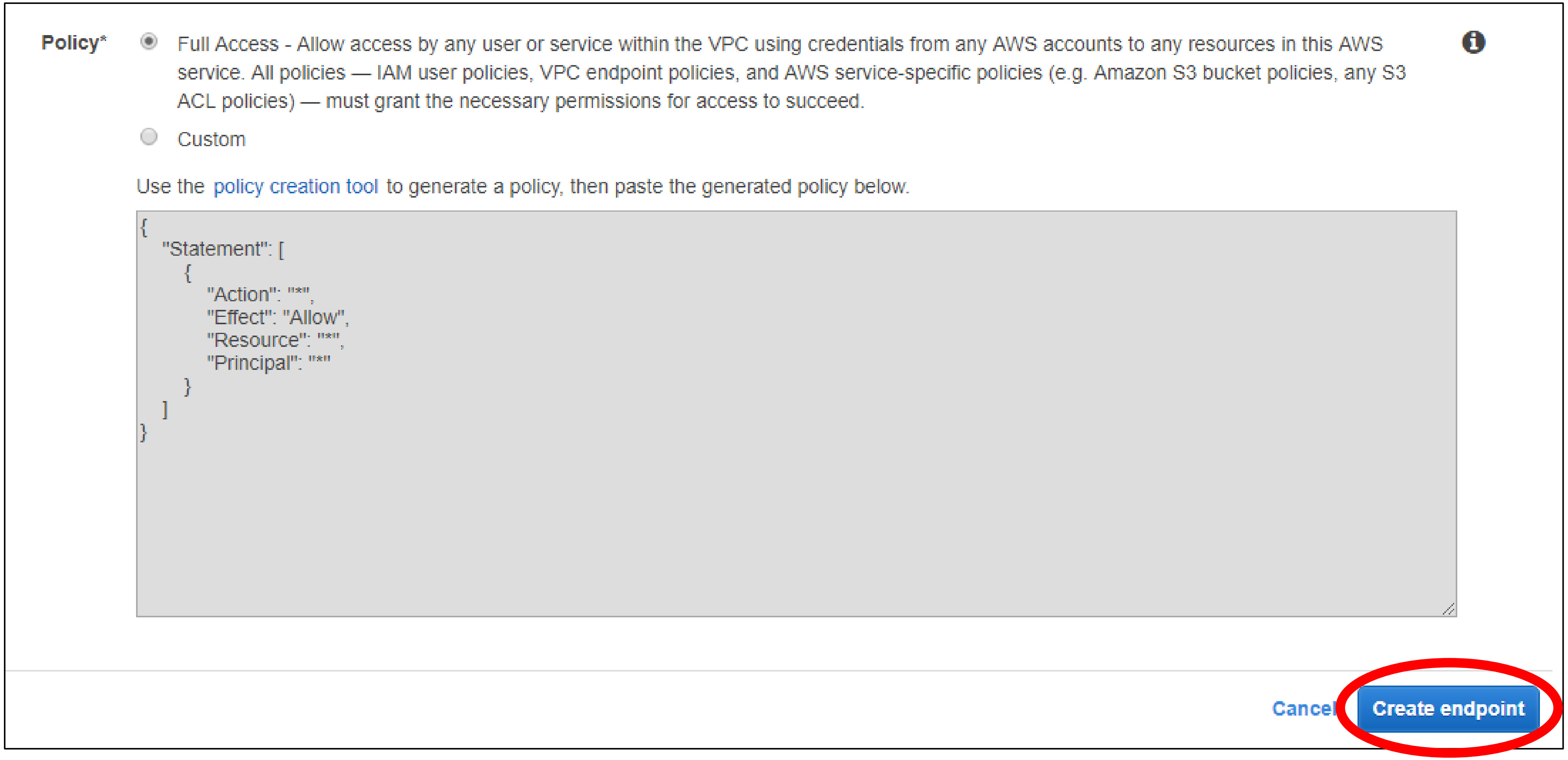
Policy*

● Full Access - Allow access by any user or service within the VPC using credentials from any AWS accounts to any resources in this AWS service. All policies — IAM user policies, VPC endpoint policies, and AWS service-specific policies (e.g. Amazon S3 bucket policies, any S3 ACL policies) — must grant the necessary permissions for access to succeed.

○ Custom

VPC

└─Create a S3 endpoint



Policy* ● Full Access - Allow access by any user or service within the VPC using credentials from any AWS accounts to any resources in this AWS service. All policies — IAM user policies, VPC endpoint policies, and AWS service-specific policies (e.g. Amazon S3 bucket policies, any S3 ACL policies) — must grant the necessary permissions for access to succeed.

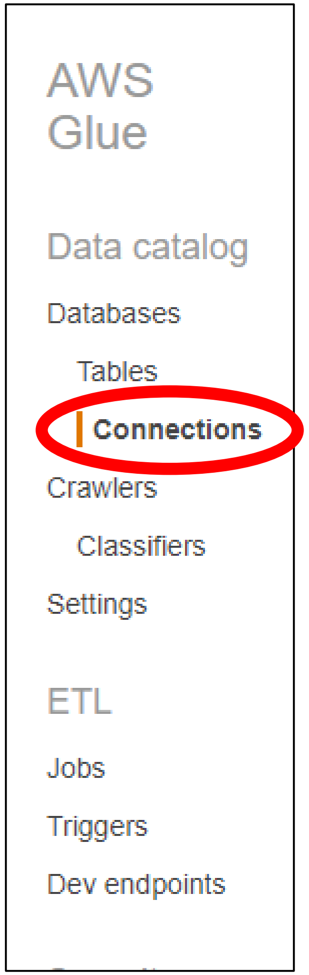
○ Custom

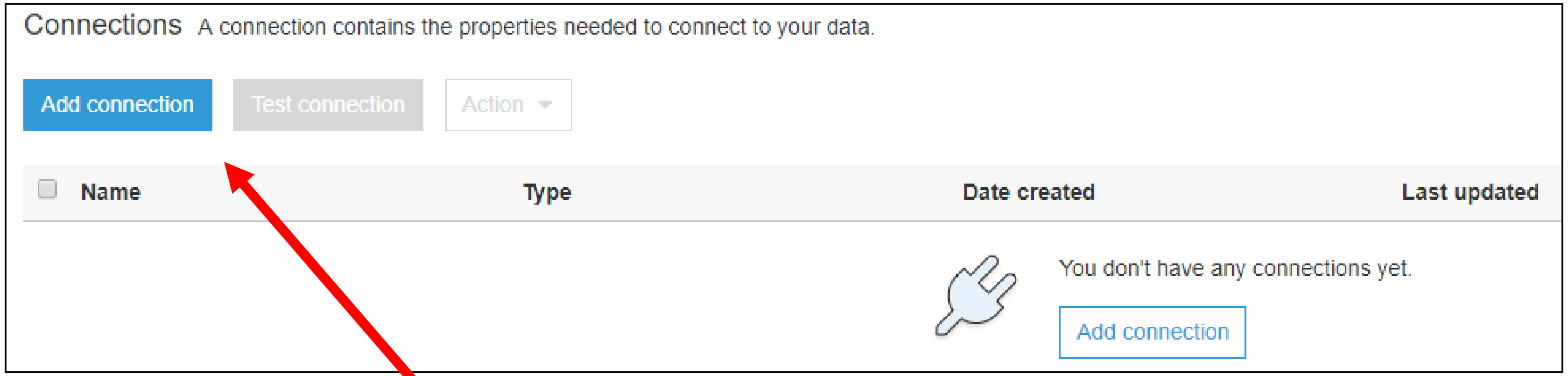
Use the policy creation tool to generate a policy, then paste the generated policy below.

```
{
   "Statement": [
      {
         "Action": "*",
         "Effect": "Allow",
         "Resource": "*",
         "Principal": "*"
      }
   ]
}
```

Glue

Create a connection to Redshift





Connections A connection contains the properties needed to connect to your data.

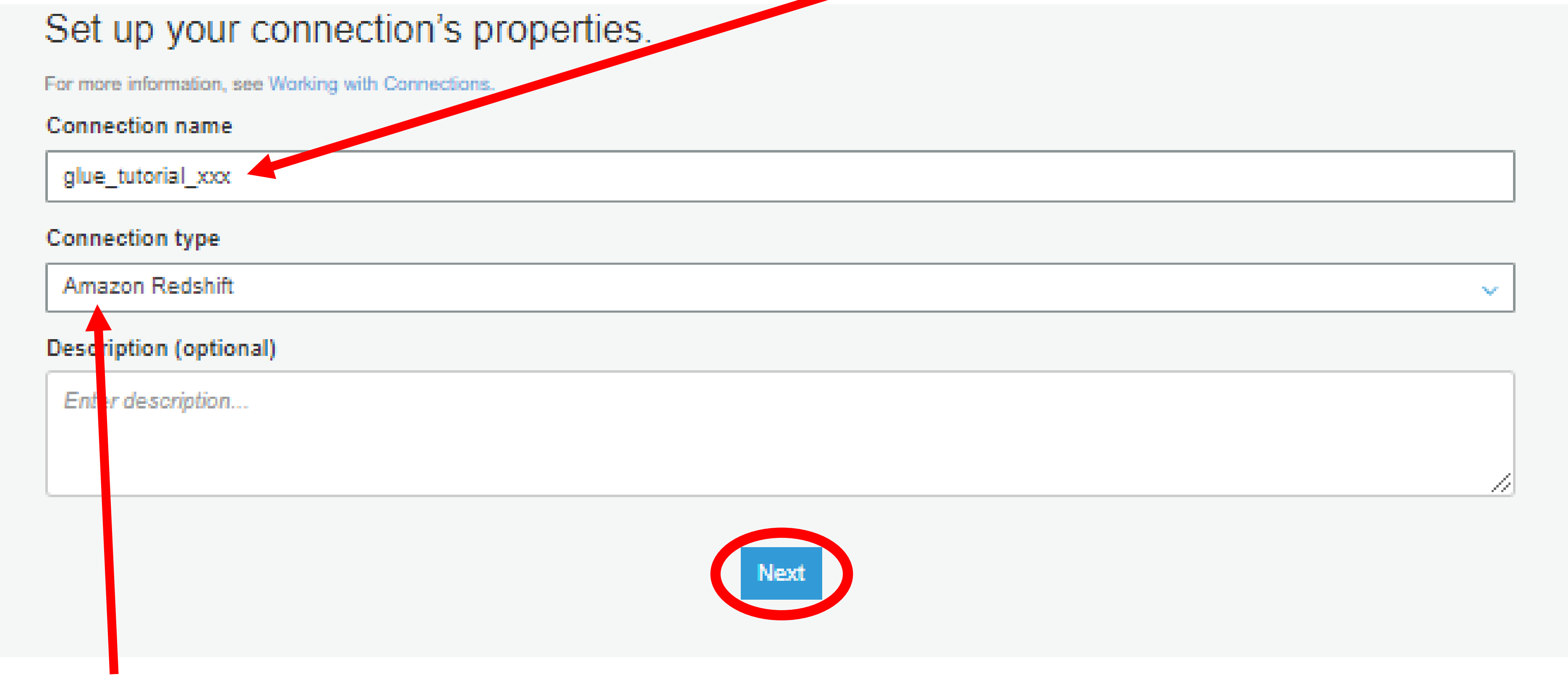
Name	Type	Date created	Last updated

You don't have any connections yet.

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

Glue

Create a connection to Redshift



Set up your connection's properties.

For more information, see Working with Connections.

Connection name

glue_tutorial_xxx

Connection type

Amazon Redshift

Description (optional)

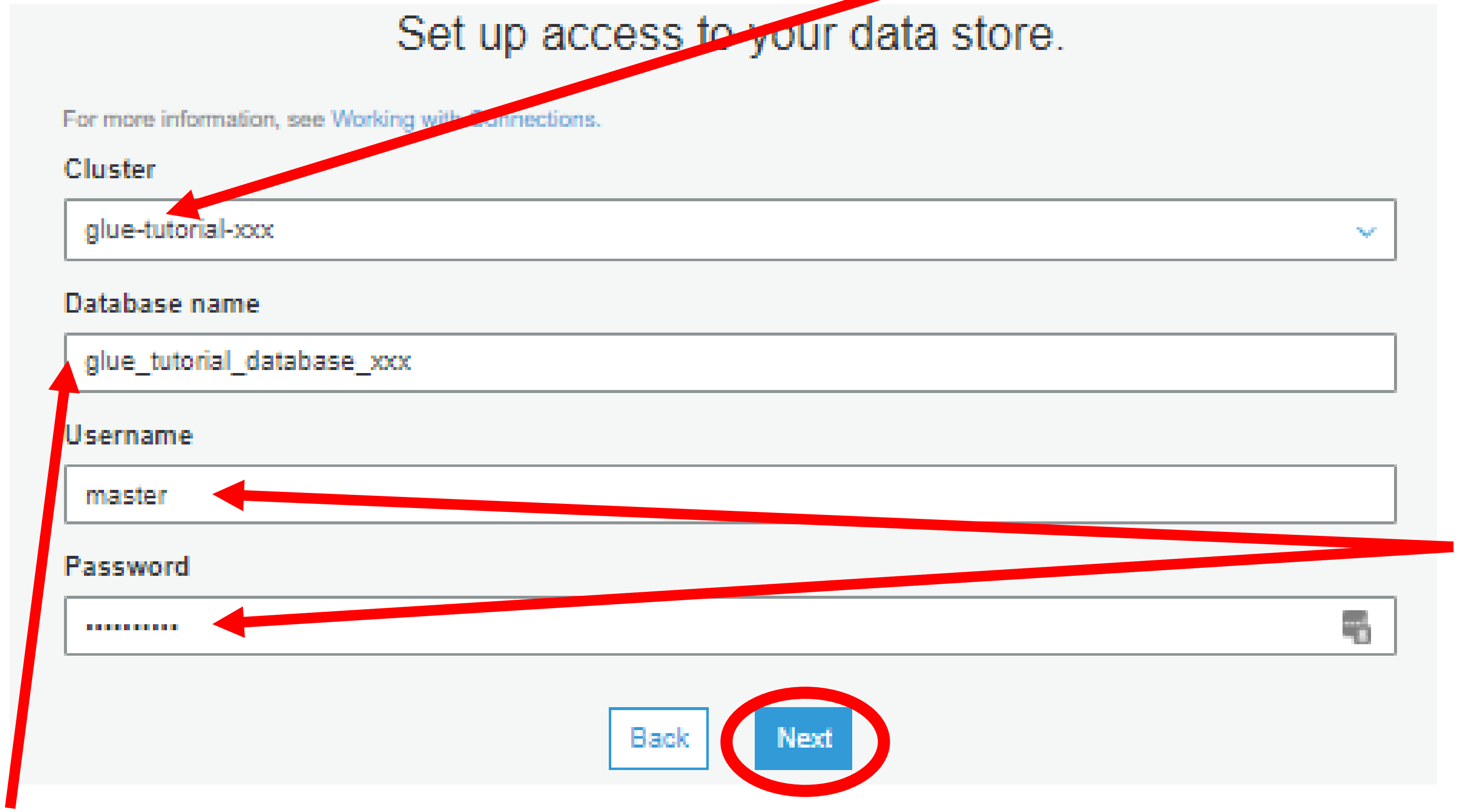
Enter description...

Name of the connection: glue-tutorial-XXX

The connection type should be Redshift

Glue

└─Create a connection to Redshift



Set up access to your data store.

For more information, see Working with Connections.

Cluster

glue-tutorial-xxx

Database name

glue_tutorial_database_xxx

Username

master

Password

••••••••••

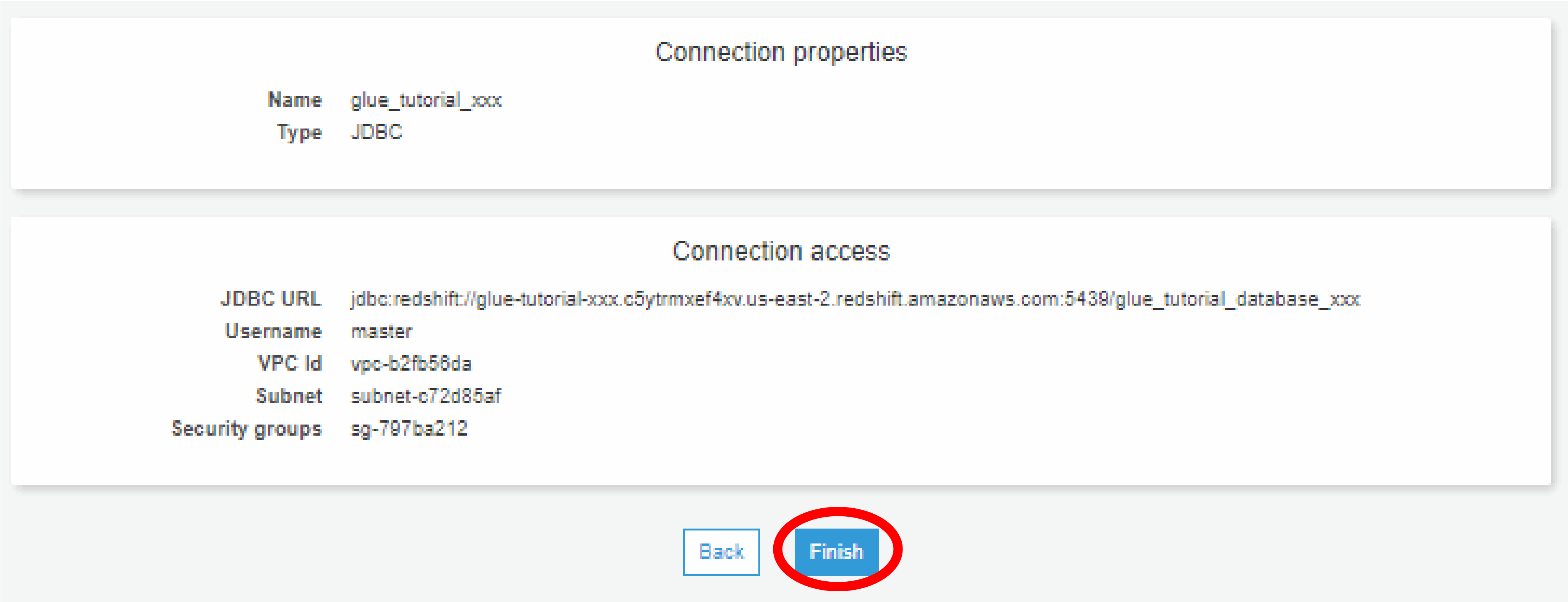
Name of the cluster: glue-tutorial-XXX

Username and password created for Redshift

Name of the database: glue_tutorial_database_xxx

Glue

└─Create a connection to Redshift



Connection properties

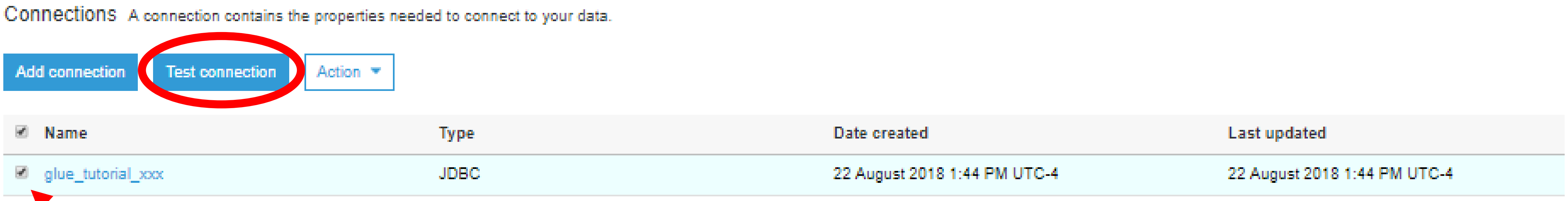
Name	glue_tutorial_xxx
Type	JDBC

Connection access

JDBC URL	jdbc:redshift://glue-tutorial-xxx.c5ytrmxef4xv.us-east-2.redshift.amazonaws.com:5439/glue_tutorial_database_xxx
Username	master
VPC Id	vpc-b2fb56da
Subnet	subnet-c72d85af
Security groups	sg-797ba212

Glue

Test the connection to Redshift



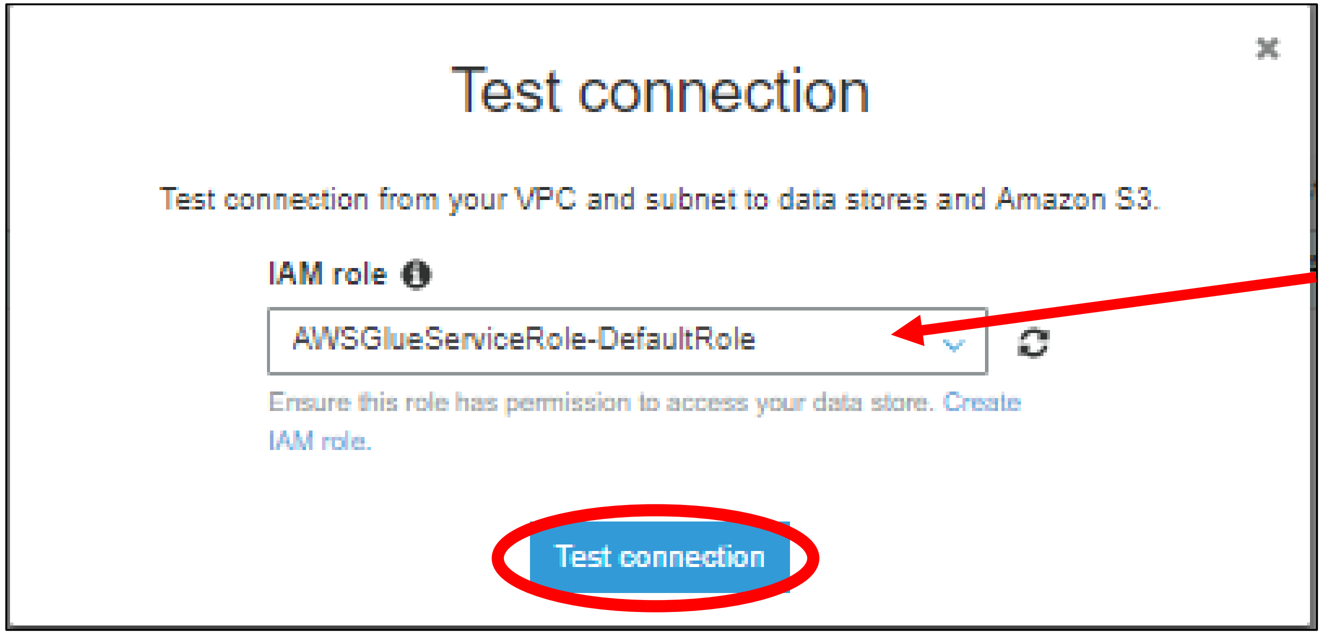
Connections A connection contains the properties needed to connect to your data.

☑	Name	Type	Date created	Last updated
☑	glue_tutorial_xxx	JDBC	22 August 2018 1:44 PM UTC-4	22 August 2018 1:44 PM UTC-4

Select newly created connection

Glue

Test the connection to Redshift



Test connection

Test connection from your VPC and subnet to data stores and Amazon S3.

IAM role

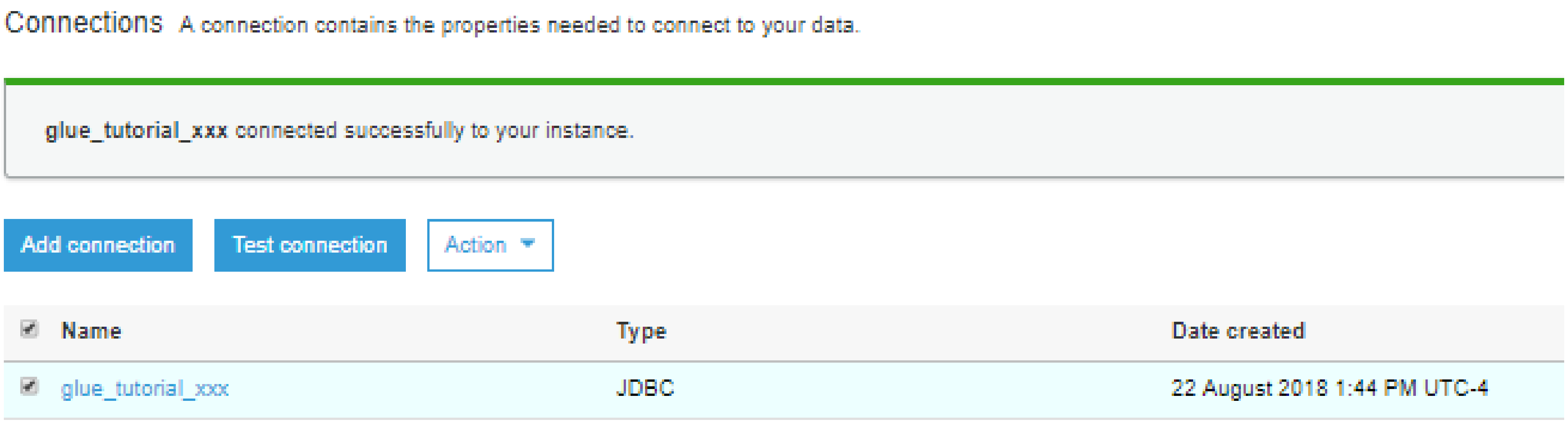
AWSGlueServiceRole-DefaultRole

Ensure this role has permission to access your data store. Create IAM role.

Select your recently created IAM role

Glue

Test the connection to Redshift



Connections A connection contains the properties needed to connect to your data.

glue_tutorial_xxx connected successfully to your instance.

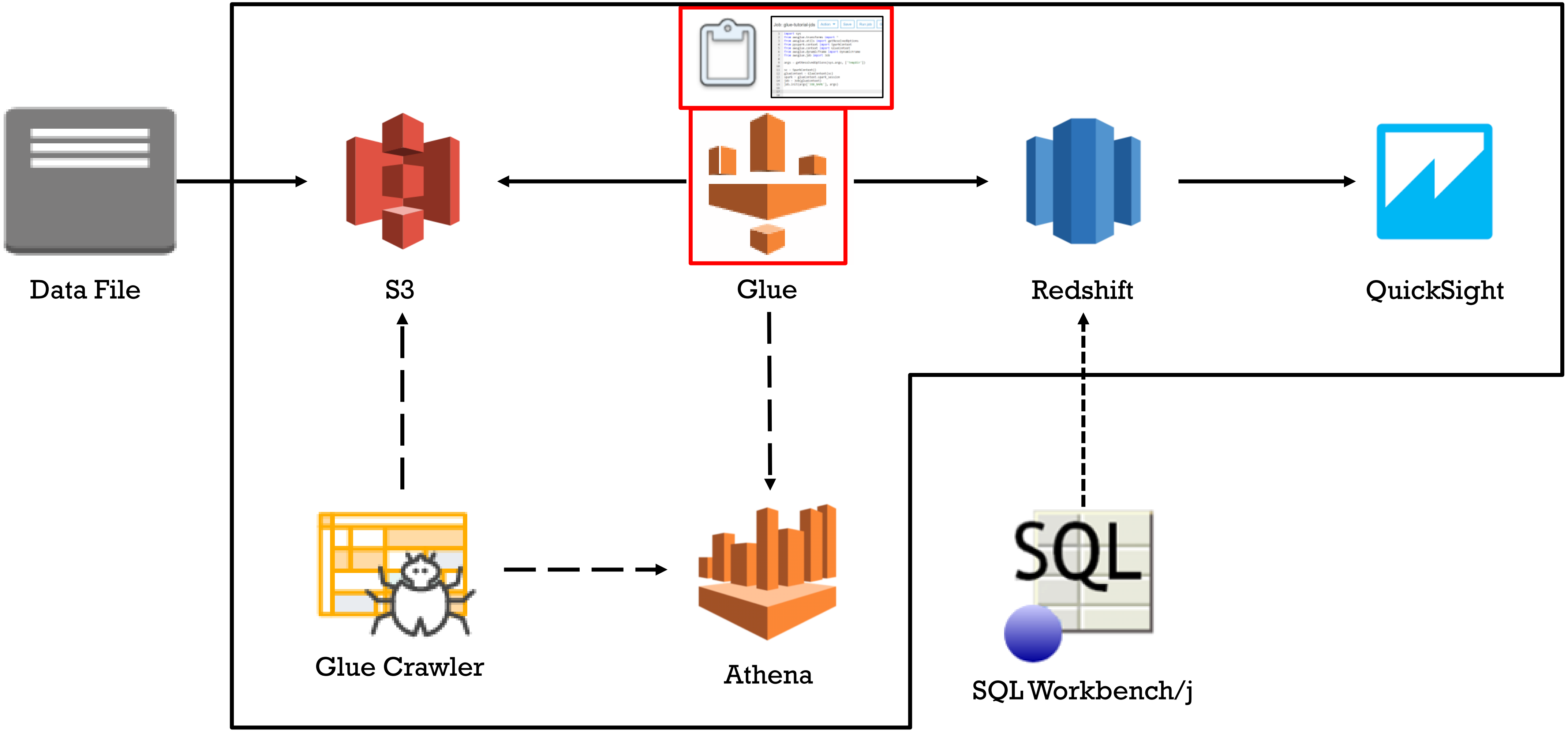
☑	Name	Type	Date created
☑	glue_tutorial_xxx	JDBC	22 August 2018 1:44 PM UTC-4

Lab 4

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

(Use US-EAST-1/N. Virginia Region)

Glue



Data File

S3

Glue

Redshift

QuickSight

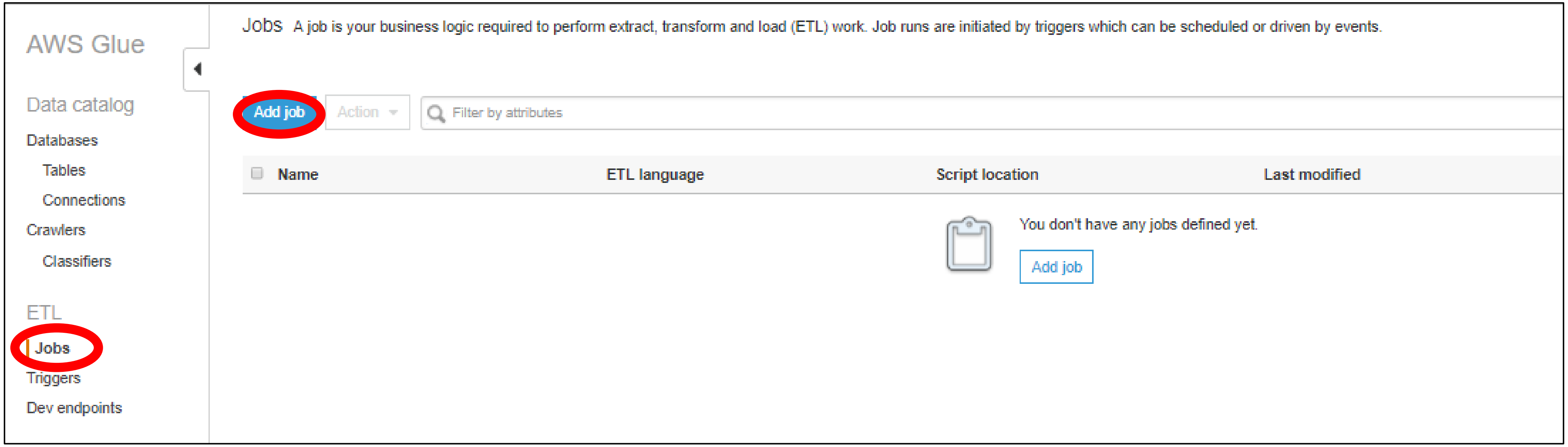
Glue Crawler

Athena

SQL Workbench/j

Glue

Create a Glue job



Jobs A job is your business logic required to perform extract, transform and load (ETL) work. Job runs are initiated by triggers which can be scheduled or driven by events.

□ Name	ETL language	Script location	Last modified

You don't have any jobs defined yet.

Glue

└─Create a Glue job

Give your job a name: glue-tutorial-XXX

The language used to write the script

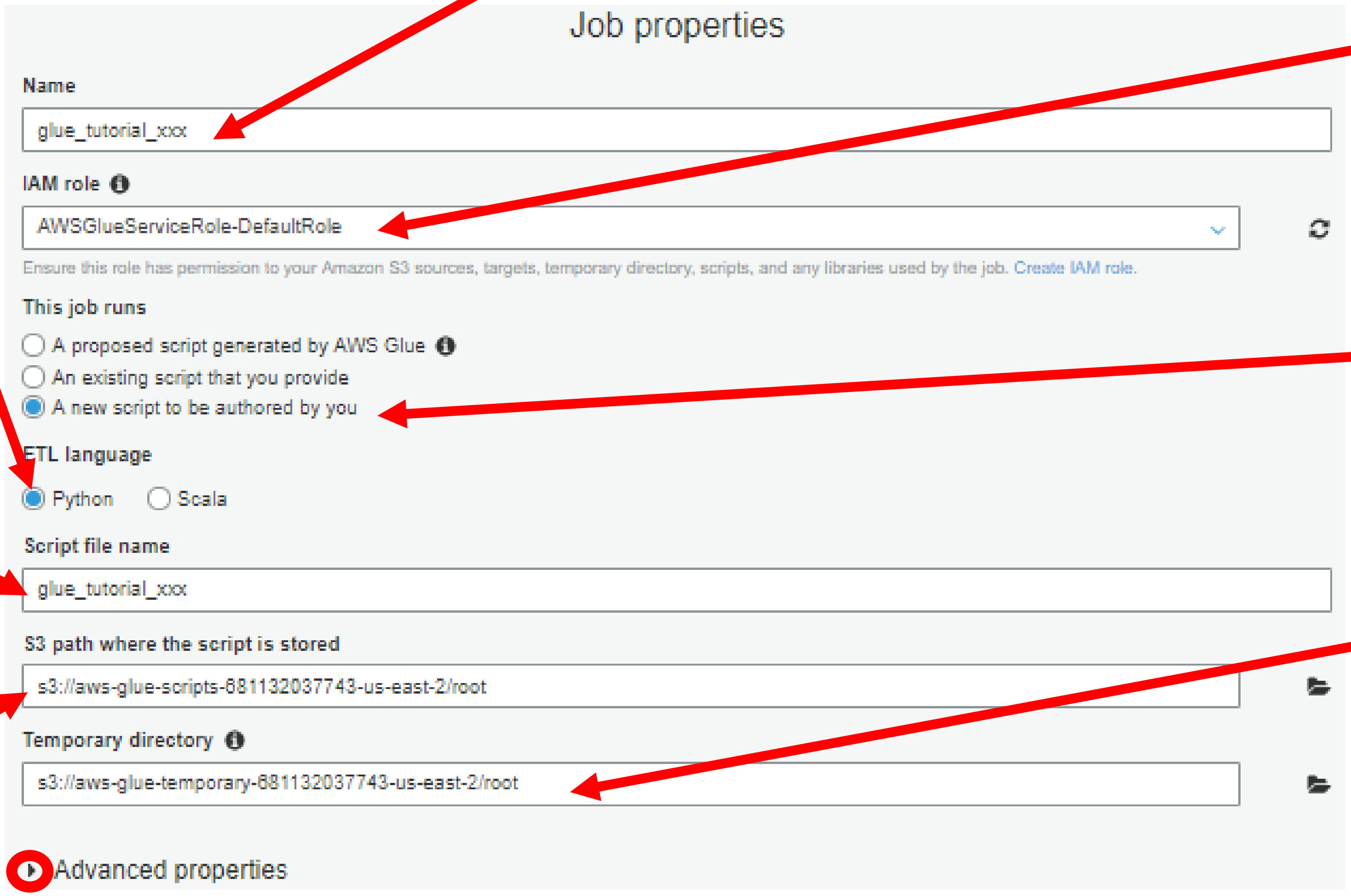
Give your script a name glue-tutorial-XXX

The location where your script will be placed in S3

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

Create a new blank script

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



Job properties

Name

glue_tutorial_xxx

IAM role

AWSGlueServiceRole-DefaultRole

Ensure this role has permission to your Amazon S3 sources, targets, temporary directory, scripts, and any libraries used by the job. Create IAM role.

This job runs

○ A proposed script generated by AWS Glue

○ An existing script that you provide

● A new script to be authored by you

ETL language

● Python ○ Scala

Script file name

glue_tutorial_xxx

S3 path where the script is stored

s3://aws-glue-scripts-681132037743-us-east-2/root

Temporary directory

s3://aws-glue-temporary-681132037743-us-east-2/root

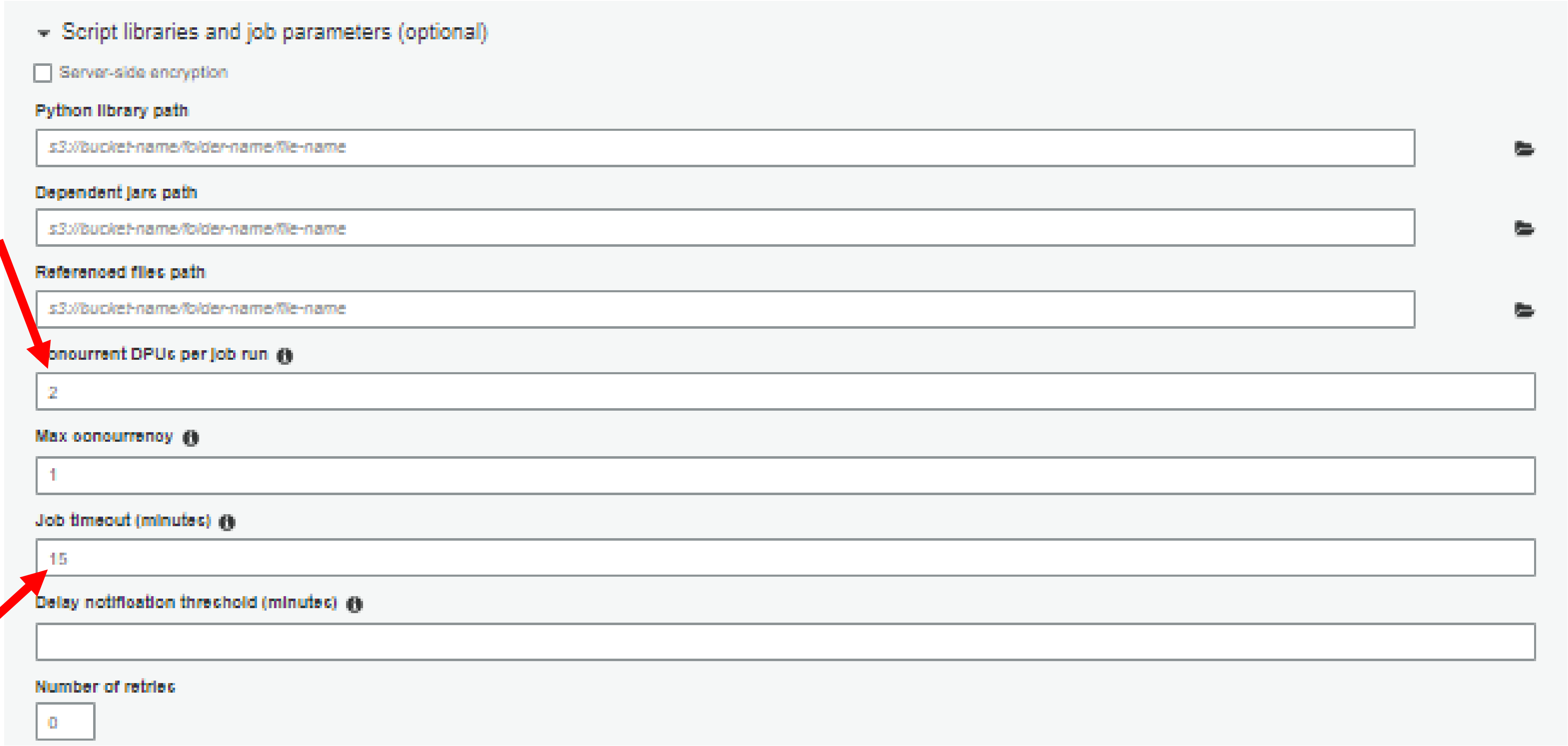
Advanced properties

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



▾ Script libraries and job parameters (optional)

☐ Server-side encryption

Python library path: s3://bucket-name/folder-name/file-name

Dependent jars path: s3://bucket-name/folder-name/file-name

Referenced files path: s3://bucket-name/folder-name/file-name

Concurrent DPUs per job run: 2

Max concurrency: 1

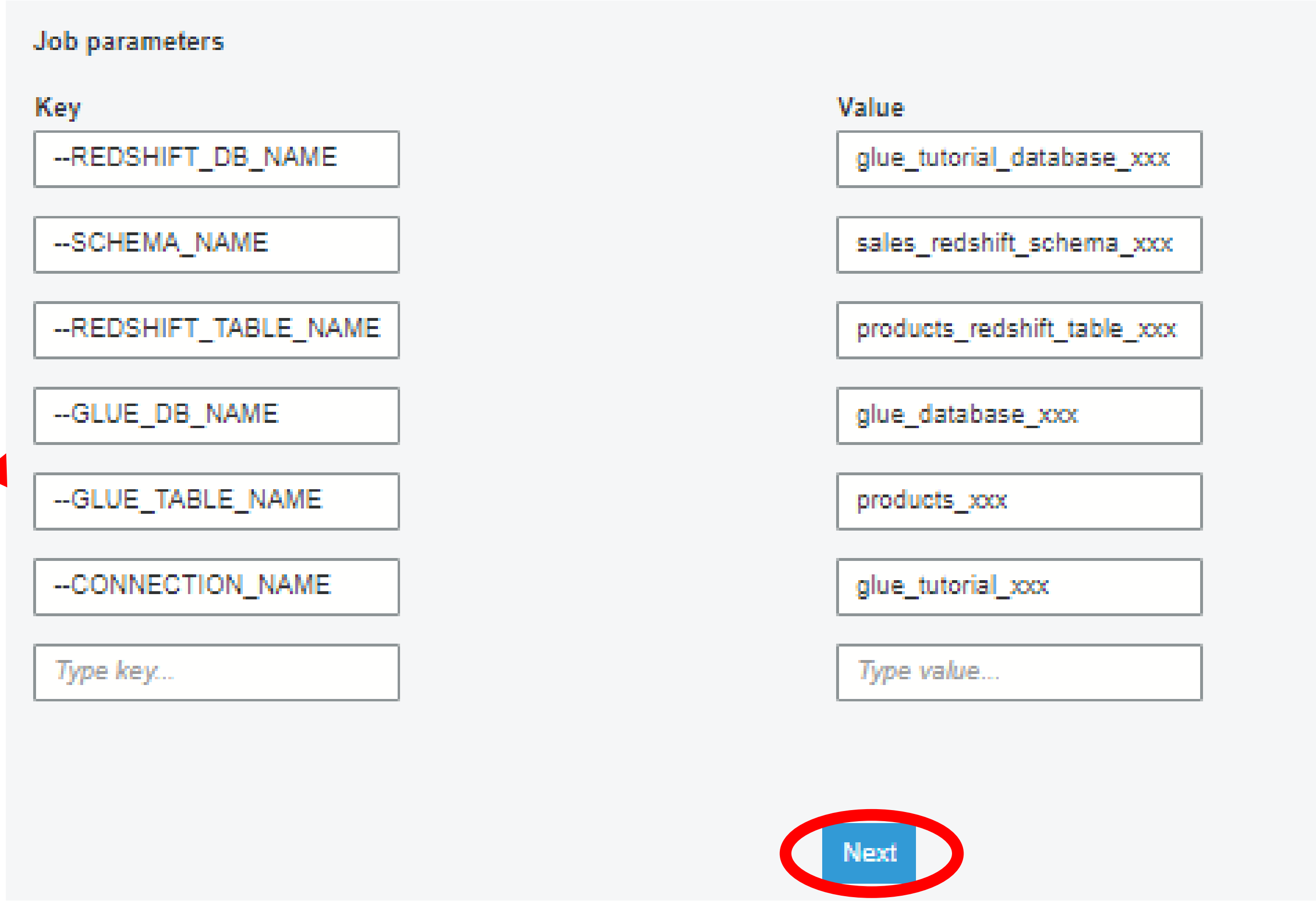
Job timeout (minutes): 15

Delay notification threshold (minutes):

Number of retries: 0

Glue

└─Create a Glue job



Job parameters

Key	Value
--REDSHIFT_DB_NAME	glue_tutorial_database_xxx
--SCHEMA_NAME	sales_redshift_schema_xxx
--REDSHIFT_TABLE_NAME	products_redshift_table_xxx
--GLUE_DB_NAME	glue_database_xxx
--GLUE_TABLE_NAME	products_xxx
--CONNECTION_NAME	glue_tutorial_xxx
Type key...	*Type value...*

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

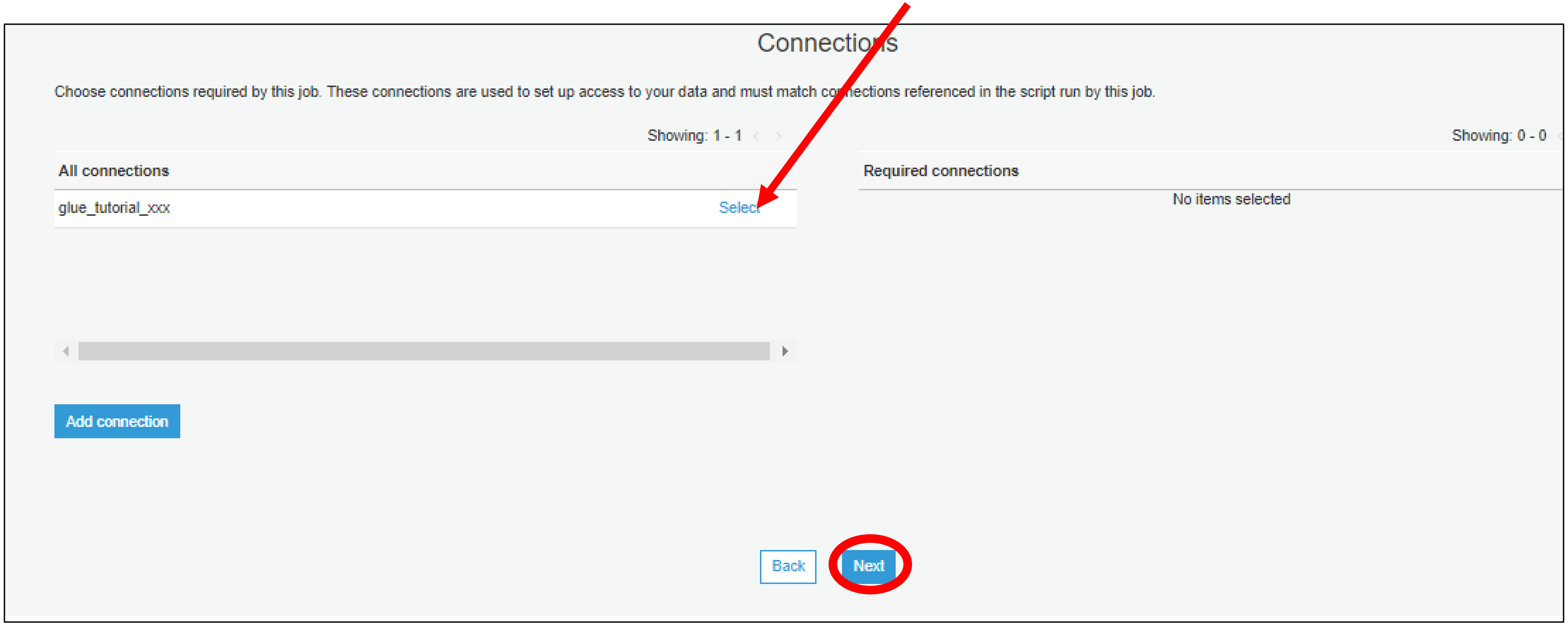
--CONNECTION_NAME
glue_tutorial_xxx

Parameterize values to be used in the script

Glue

└─Create a Glue job

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



Connections

Choose connections required by this job. These connections are used to set up access to your data and must match connections referenced in the script run by this job.

Showing: 1 - 1

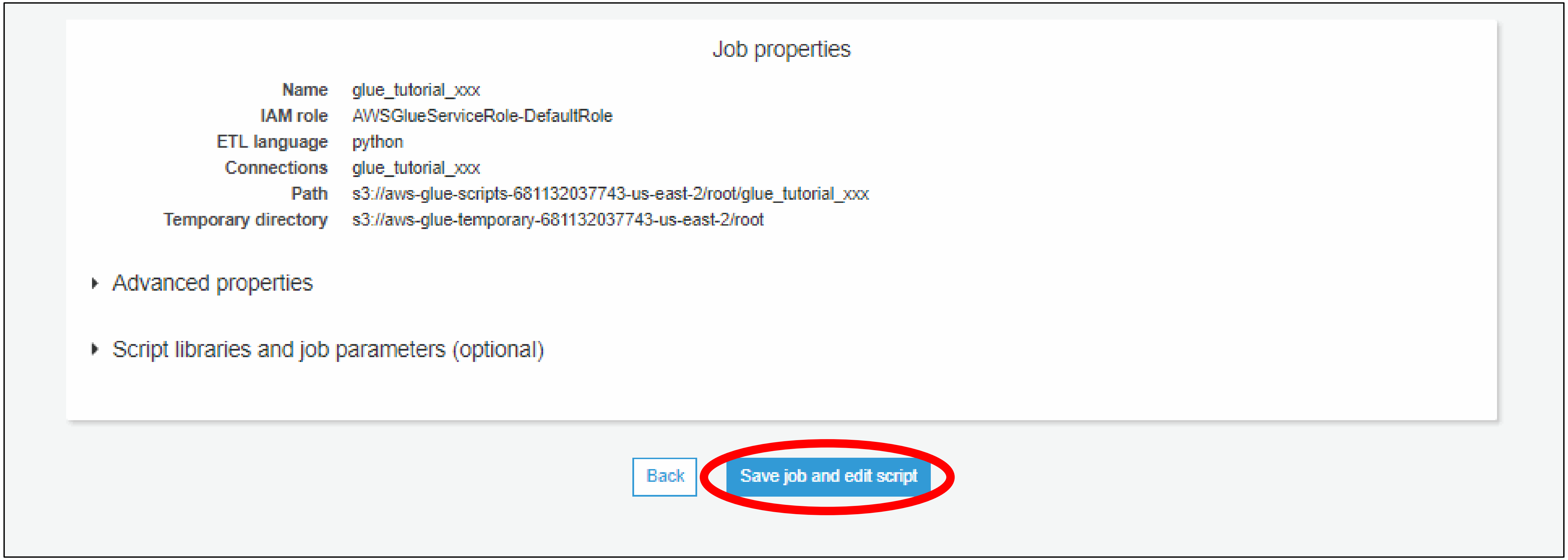
All connections	
glue_tutorial_xxx	Select

Showing: 0 - 0

Required connections
No items selected

Glue

Create a Glue job

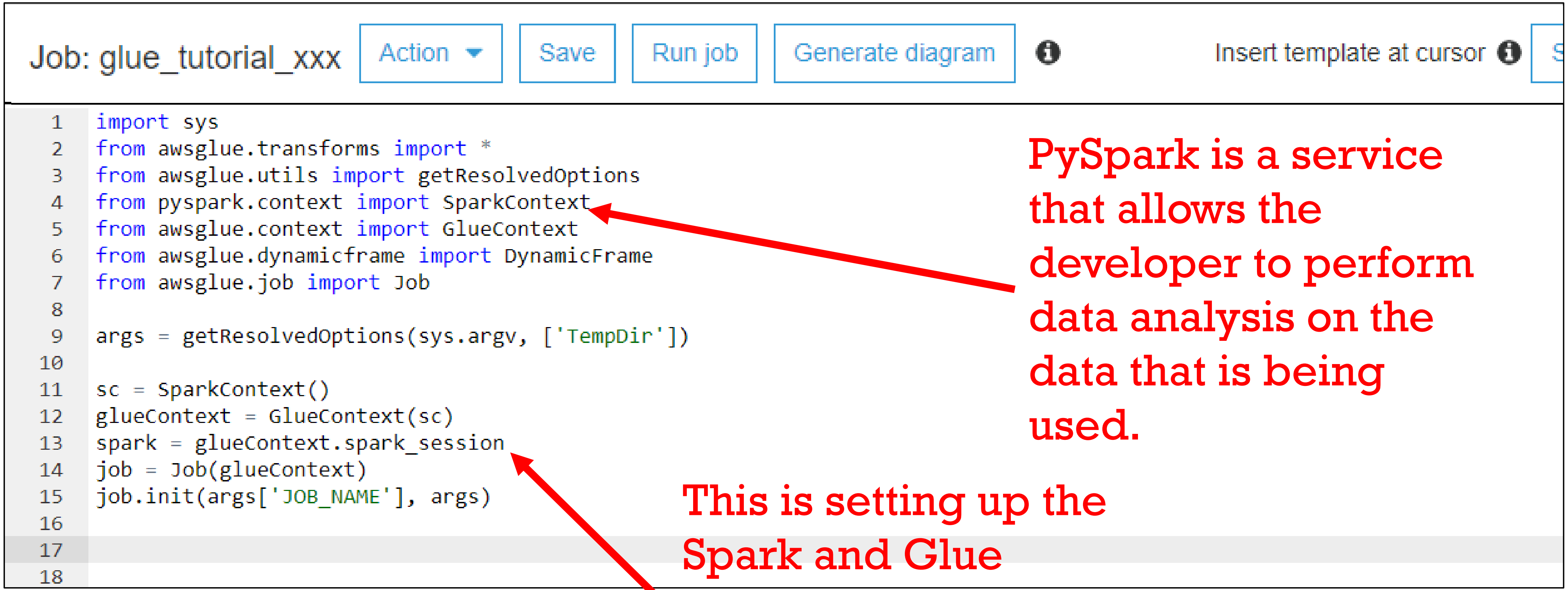


Job properties

Name	glue_tutorial_xxx
IAM role	AWSGlueServiceRole-DefaultRole
ETL language	python
Connections	glue_tutorial_xxx
Path	s3://aws-glue-scripts-681132037743-us-east-2/root/glue_tutorial_xxx
Temporary directory	s3://aws-glue-temporary-681132037743-us-east-2/root

Glue

Writing the Script



```python
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

args = getResolvedOptions(sys.argv, ['TempDir'])

sc = SparkContext()
glueContext = GlueContext(sc)
spark = glueContext.spark_session
job = Job(glueContext)
job.init(args['JOB_NAME'], args)


```

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

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

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’])
sc = SparkContext()
glueContext = GlueContext(sc)
spark = glueContext.spark_session
job = Job(glueContext)
job.init(args['JOB_NAME'], args)
```

Include SQL functions, types, and datetime to use later

Add the parameters that were passed into the Glue job

Glue

└─Writing the Script

```
...
job.init(args['JOB_NAME'], args)

datasource =
glueContext.create_dynamic_frame.from_catalog(
    database = args[‘GLUE_DB_NAME'],
    table_name = args[‘GLUBE_TABLE_NAME']
)
```

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

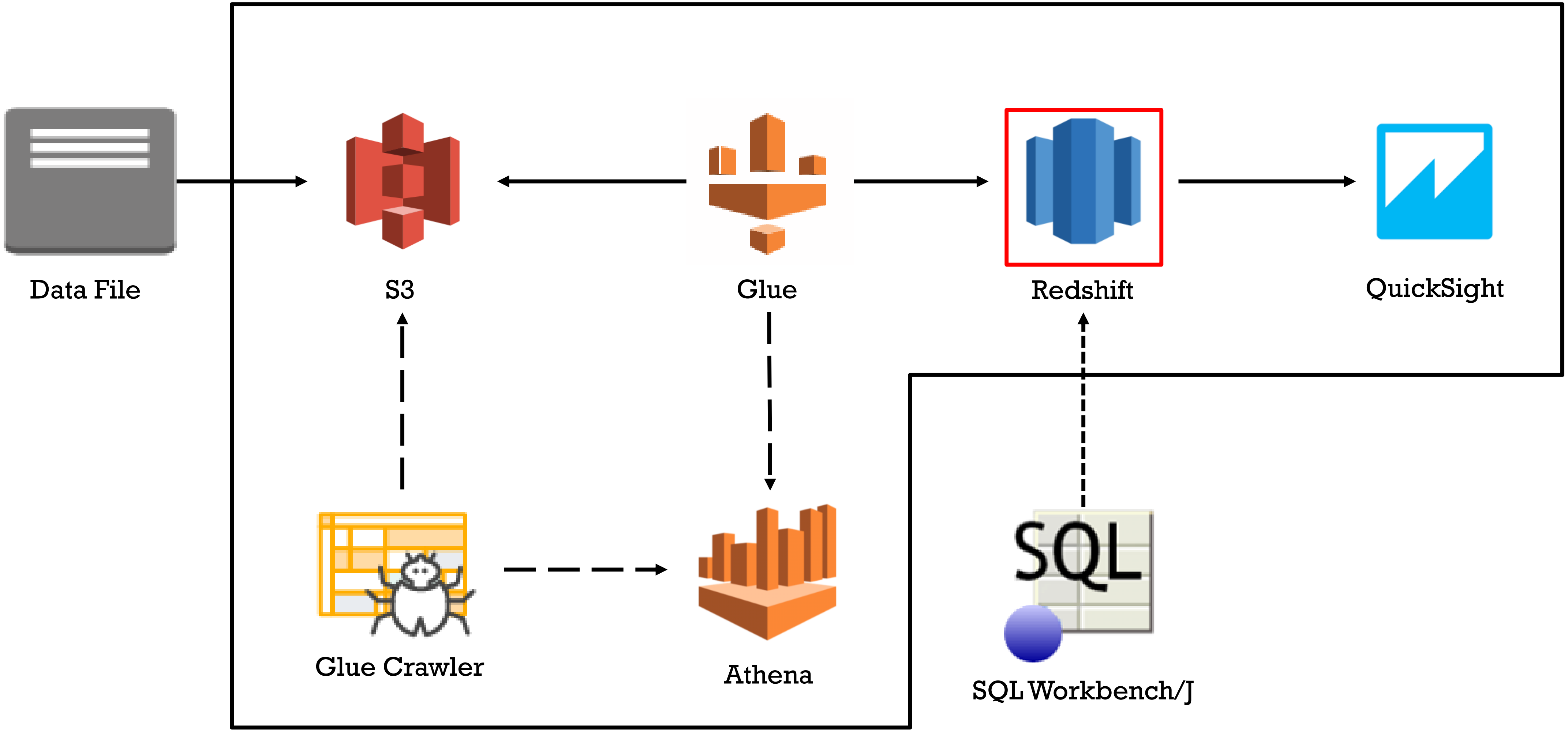
Glue

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

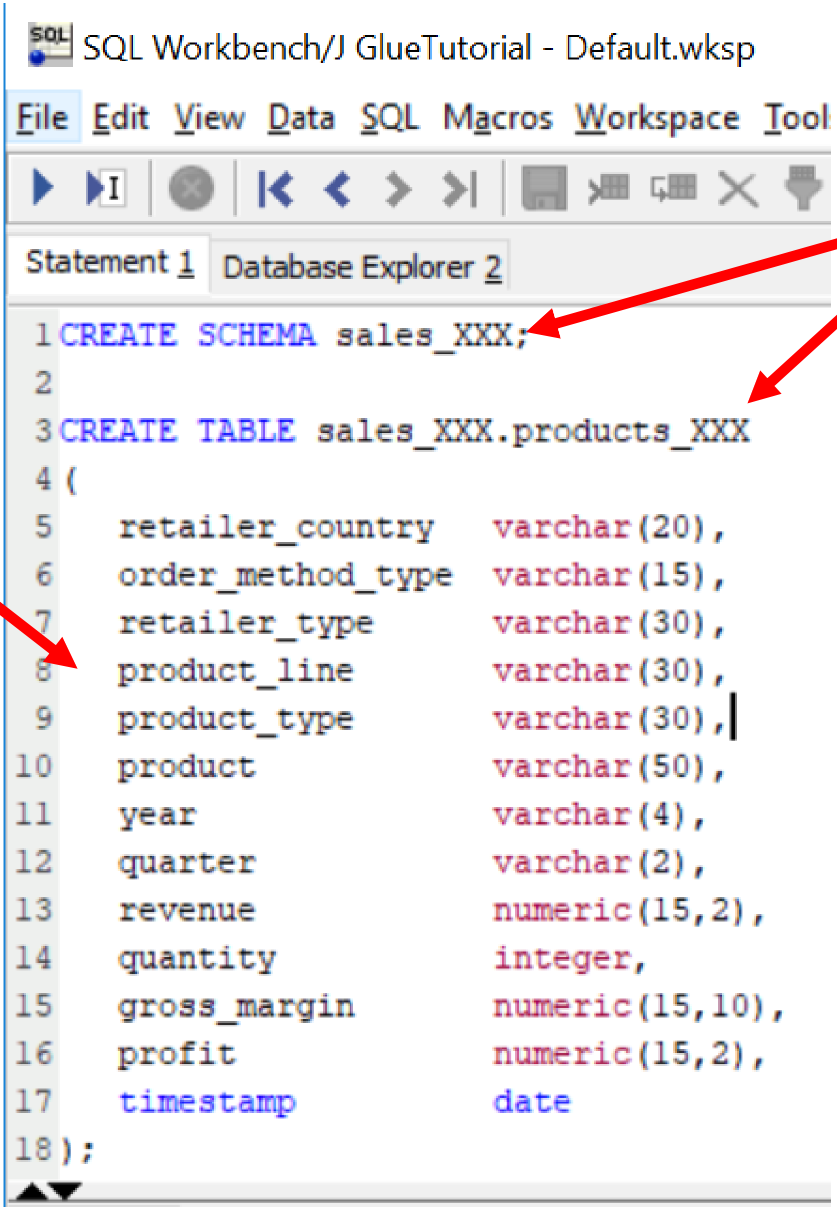
Redshift


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

Redshift

└─**Create table**

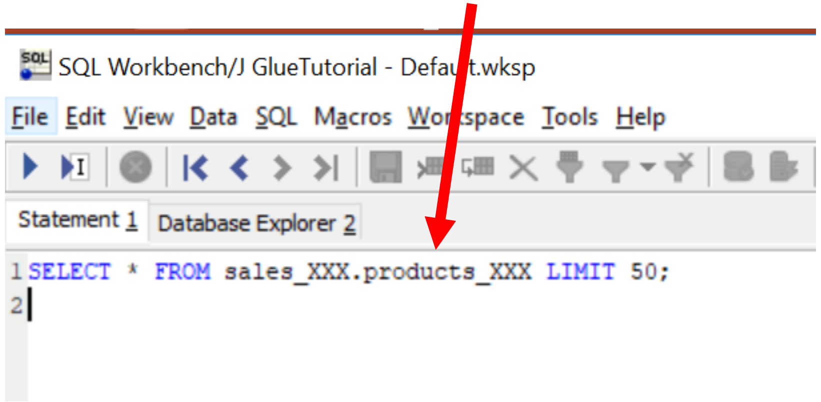
Copy the SQL script from the repository into SQL Workbench


SQL Workbench/J GlueTutorial - Default.wksp
File Edit View Data SQL Macros Workspace Tool
Statement 1
Database Explorer 2

```
CREATE SCHEMA sales_XXX;

CREATE TABLE sales_XXX.products_XXX
(
   retailer_country    varchar(20),
   order_method_type   varchar(15),
   retailer_type       varchar(30),
   product_line        varchar(30),
   product_type        varchar(30),
   product             varchar(50),
   year                varchar(4),
   quarter             varchar(2),
   revenue             numeric(15,2),
   quantity            integer,
   gross_margin        numeric(15,10),
   profit              numeric(15,2),
   timestamp           date
);
```

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

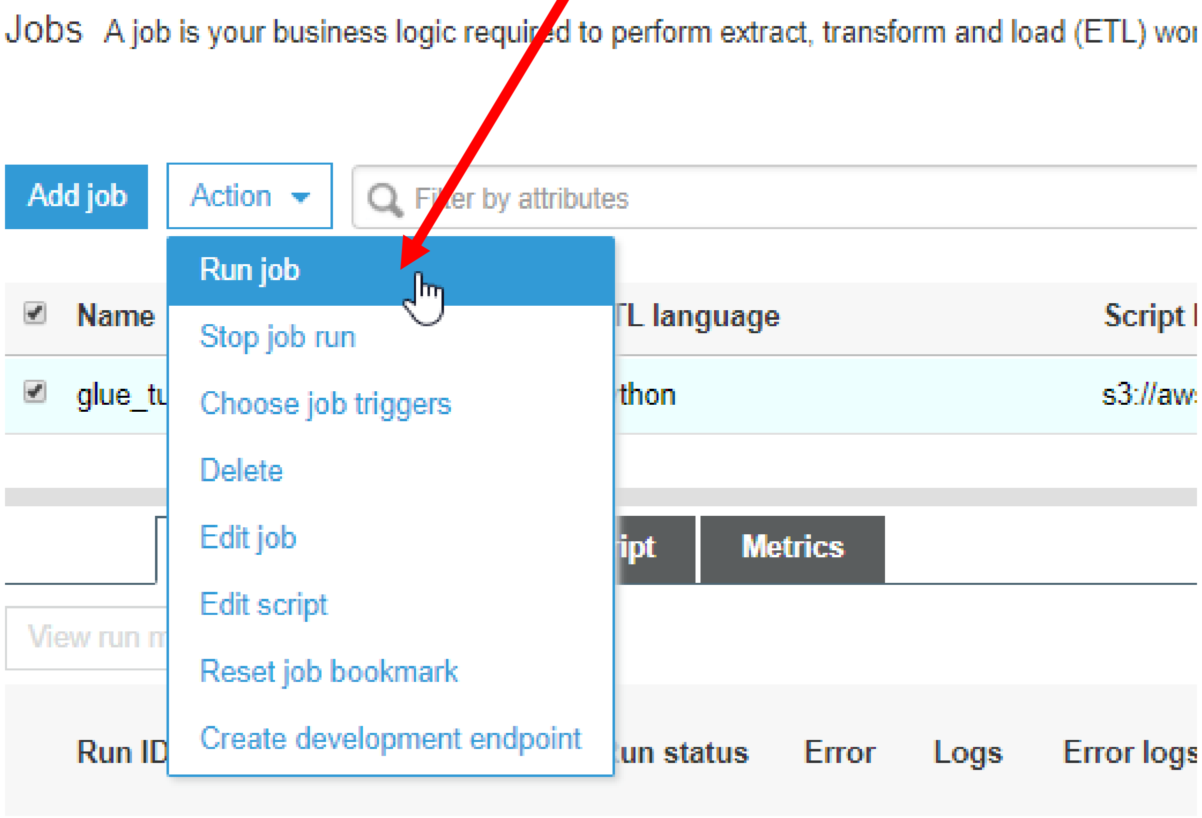

SQL Workbench/J GlueTutorial - Default.wksp
File Edit View Data SQL Macros Workspace Tools Help
Statement 1
Database Explorer 2

```
SELECT * FROM sales_XXX.products_XXX LIMIT 50;
```

Glue

Run the Glue job

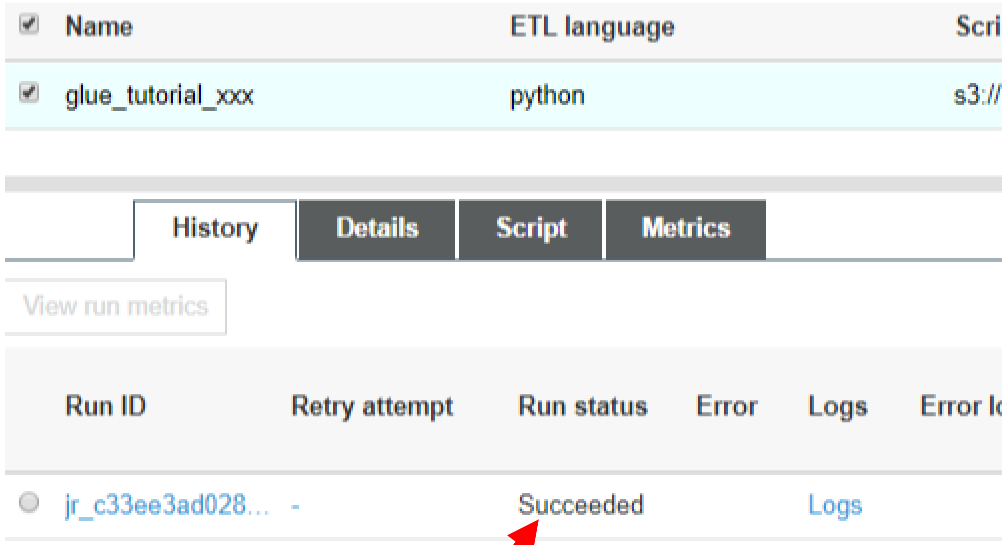
Go back to Glue and run your Glue job



Jobs A job is your business logic required to perform extract, transform and load (ETL) wor

☑	Name	TL language	Script l
☑	glue_tu	thon	s3://aw:

Run ID	un status	Error	Logs	Error logs

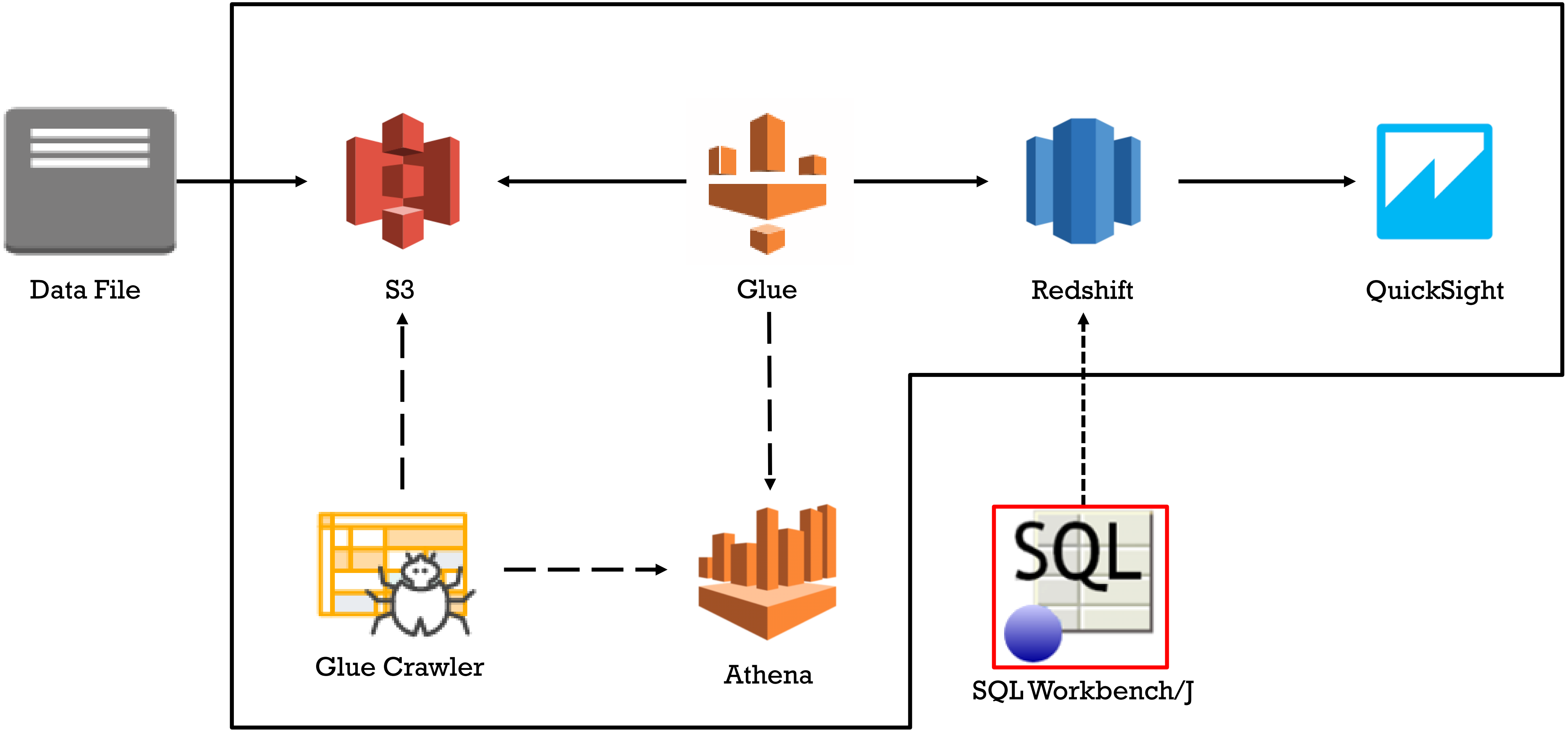


☑	Name	ETL language	Scri
☑	glue_tutorial_xxx	python	s3://

	Run ID	Retry attempt	Run status	Error	Logs	Error lo
○	jr_c33ee3ad028...	-	Succeeded		Logs	

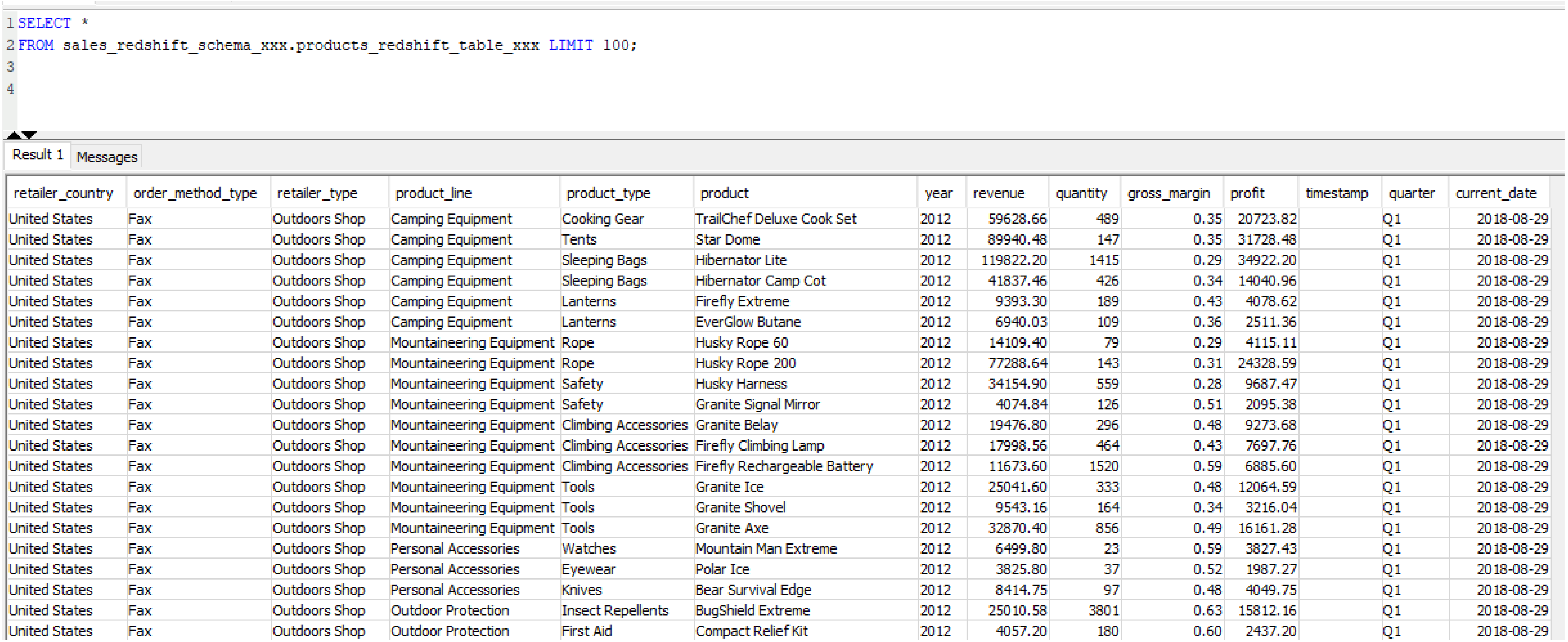
When the job succeeds, check your Redshift table

SQL Workbench



Redshift

└─Verify data in the table



```
SELECT *
FROM sales_redshift_schema_xxx.products_redshift_table_xxx LIMIT 100;
```

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

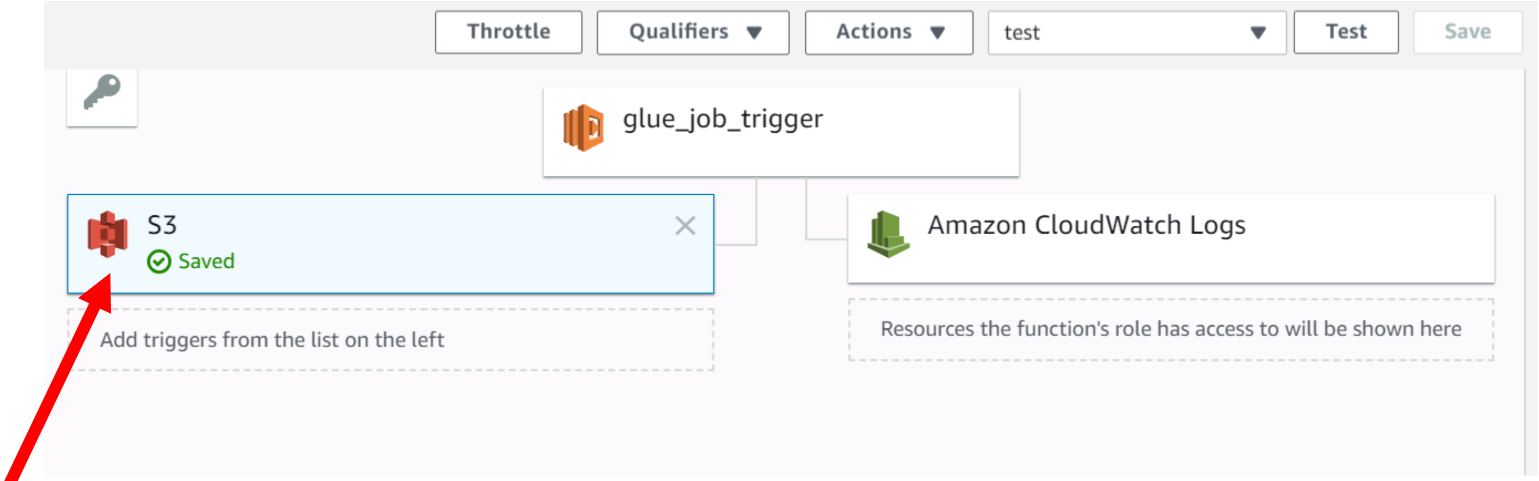
└─**Improve the versatility of your Glue job**

- 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

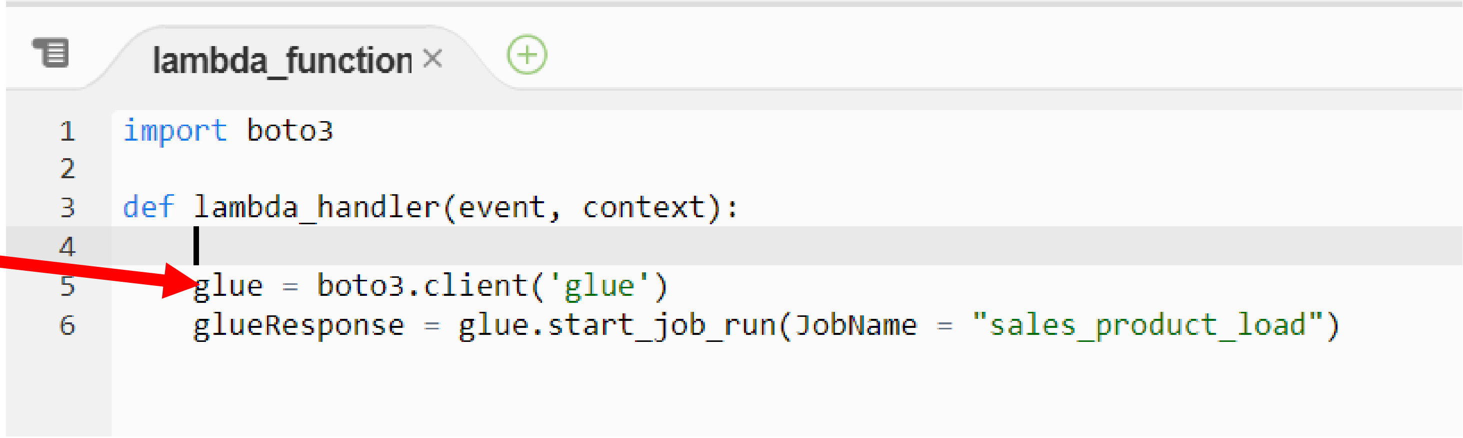
└─**Automatically run Glue job using Lambda – a serverless function**

- Instead of running the Glue job manually, have it run automatically when a file is added to S3
- Use a Lambda
- You can set a Lambda to run when a file lands in an S3 bucket
- Then make the Lambda run the Glue job



Add triggers from the list on the left

Resources the function's role has access to will be shown here

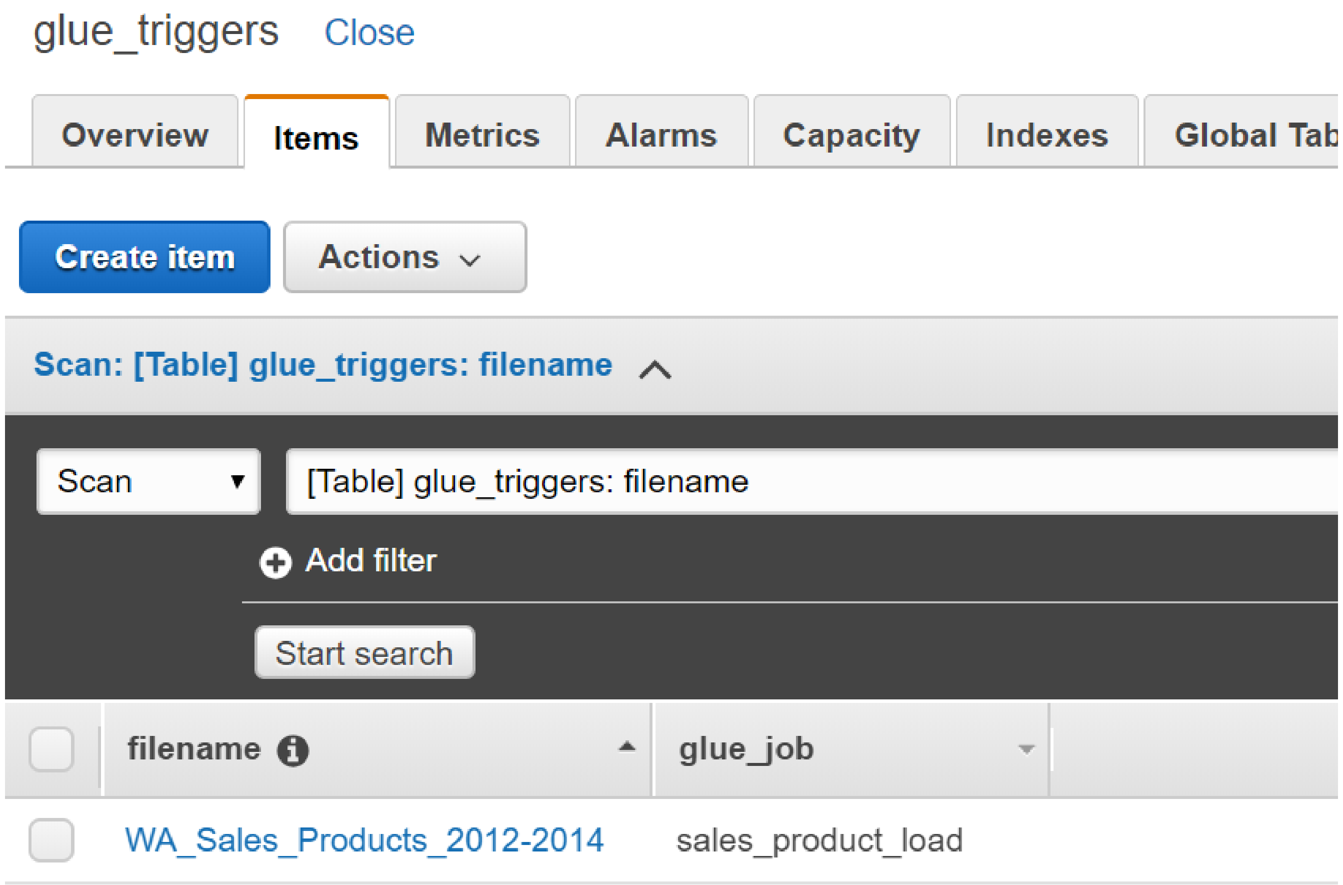


```python
import boto3

def lambda_handler(event, context):

    glue = boto3.client('glue')
    glueResponse = glue.start_job_run(JobName = "sales_product_load")
```

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

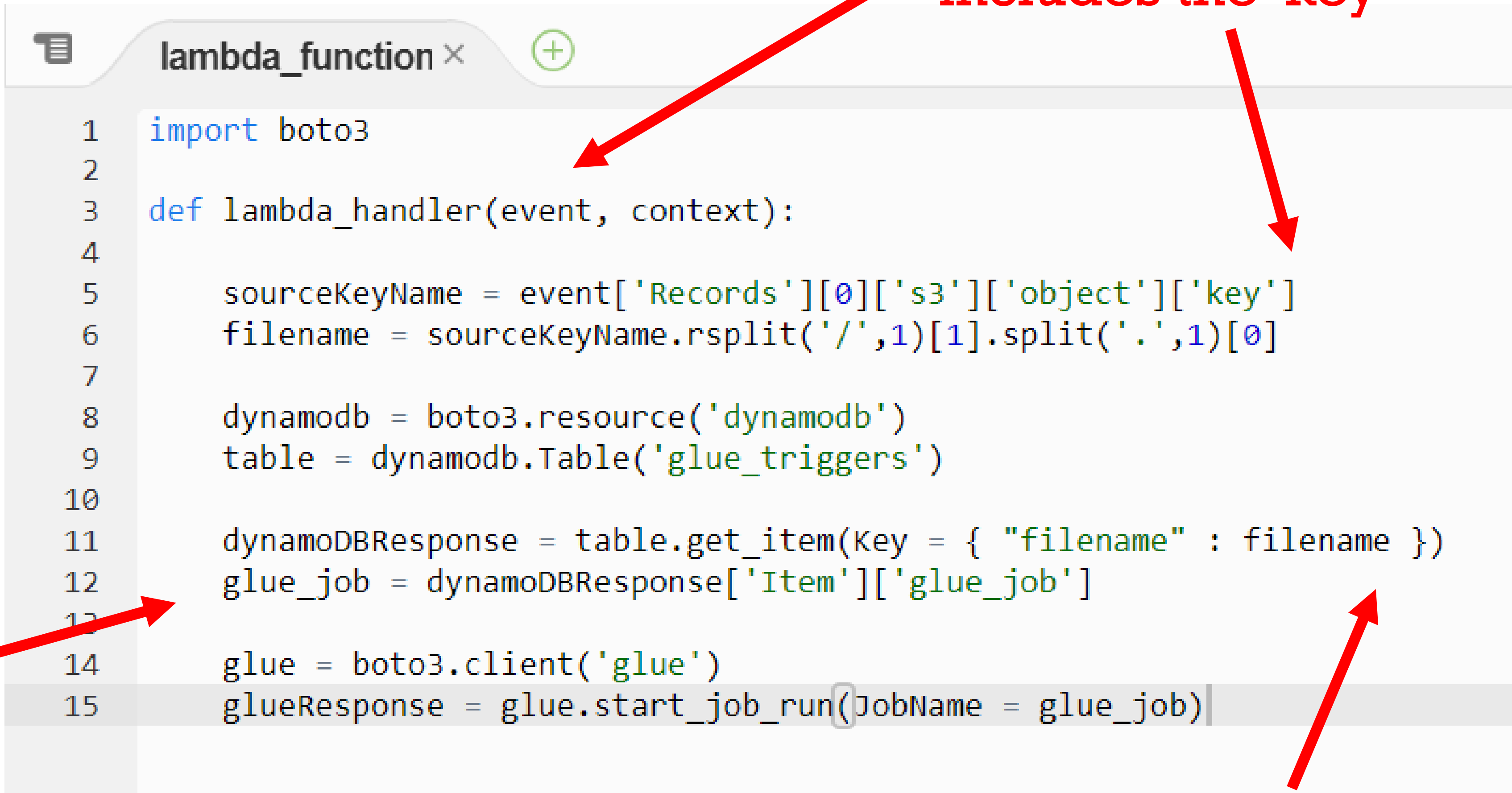


	filename	glue_job
□	WA_Sales_Products_2012-2014	sales_product_load

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



```
import boto3

def lambda_handler(event, context):

    sourceKeyName = event['Records'][0]['s3']['object']['key']
    filename = sourceKeyName.rsplit('/',1)[1].split('.',1)[0]

    dynamodb = boto3.resource('dynamodb')
    table = dynamodb.Table('glue_triggers')

    dynamoDBResponse = table.get_item(Key = { "filename" : filename })
    glue_job = dynamoDBResponse['Item']['glue_job']

    glue = boto3.client('glue')
    glueResponse = glue.start_job_run(JobName = glue_job)
```

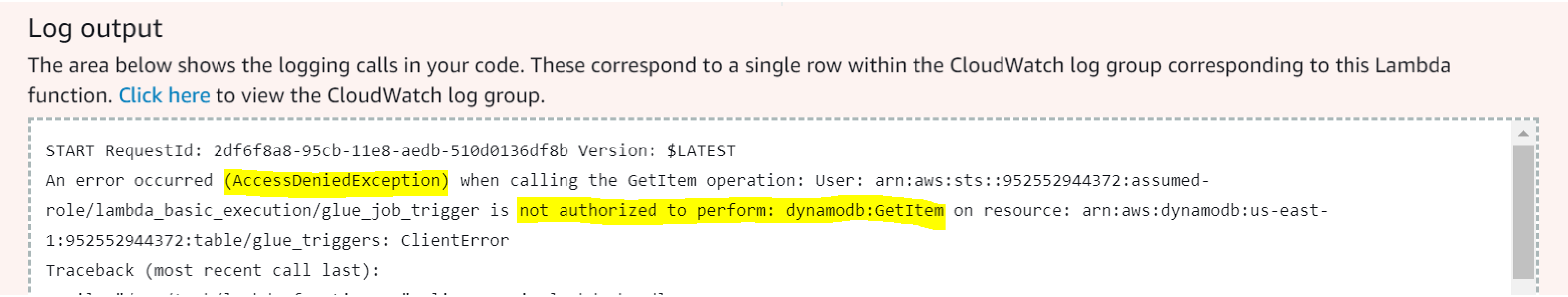
Lambda receives an event from S3, which includes the 'key'

This returns the Glue job associated with that file

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

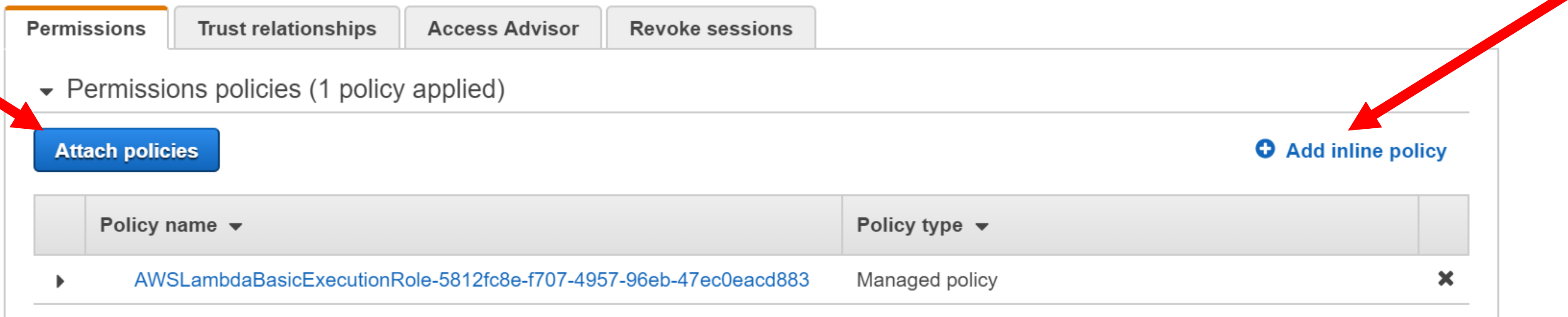
The area below shows the logging calls in your code. These correspond to a single row within the CloudWatch log group corresponding to this Lambda function. Click here to view the CloudWatch log group.

```
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):
```

- If you made the lambda from the previous slides, you would 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**



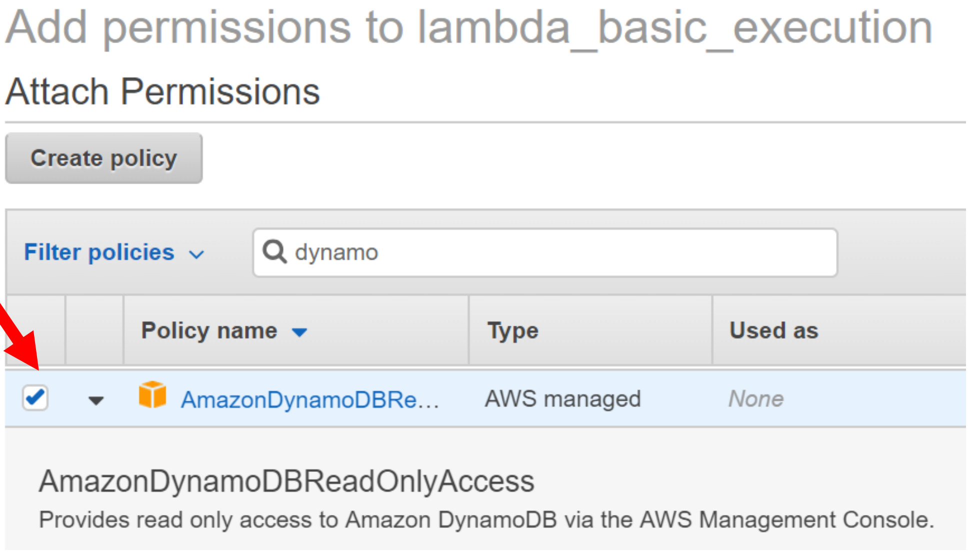
Permissions | Trust relationships | Access Advisor | Revoke sessions

▾ Permissions policies (1 policy applied)

Attach policies

⊕ Add inline policy

	Policy name ▾	Policy type ▾	
▸	AWSLambdaBasicExecutionRole-5812fc8e-f707-4957-96eb-47ec0eacd883	Managed policy	✖



Add permissions to lambda_basic_execution

Attach Permissions

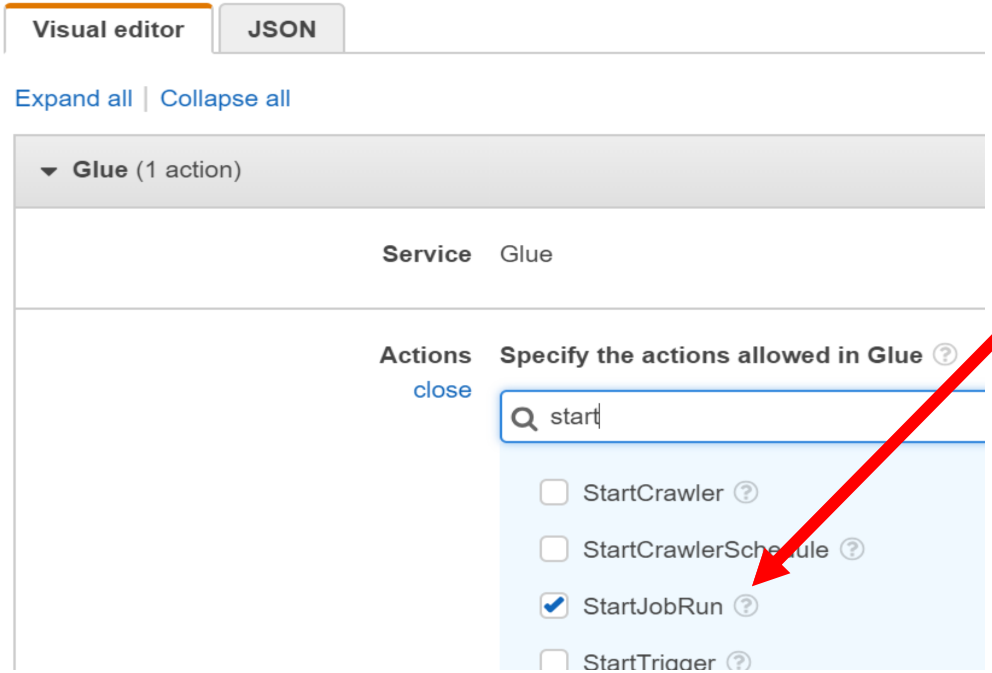
Create policy

Filter policies ⌄ dynamo

		Policy name ▾	Type	Used as
☑	▾	AmazonDynamoDBRe…	AWS managed	None

AmazonDynamoDBReadOnlyAccess

Provides read only access to Amazon DynamoDB via the AWS Management Console.



Visual editor | JSON

Expand all | Collapse all

▾ **Glue** (1 action)

Service Glue

Actions **Specify the actions allowed in Glue** ?

close

start

□ StartCrawler ?

□ StartCrawlerSchedule ?

☑ StartJobRun ?

□ StartTrigger ?

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"
```

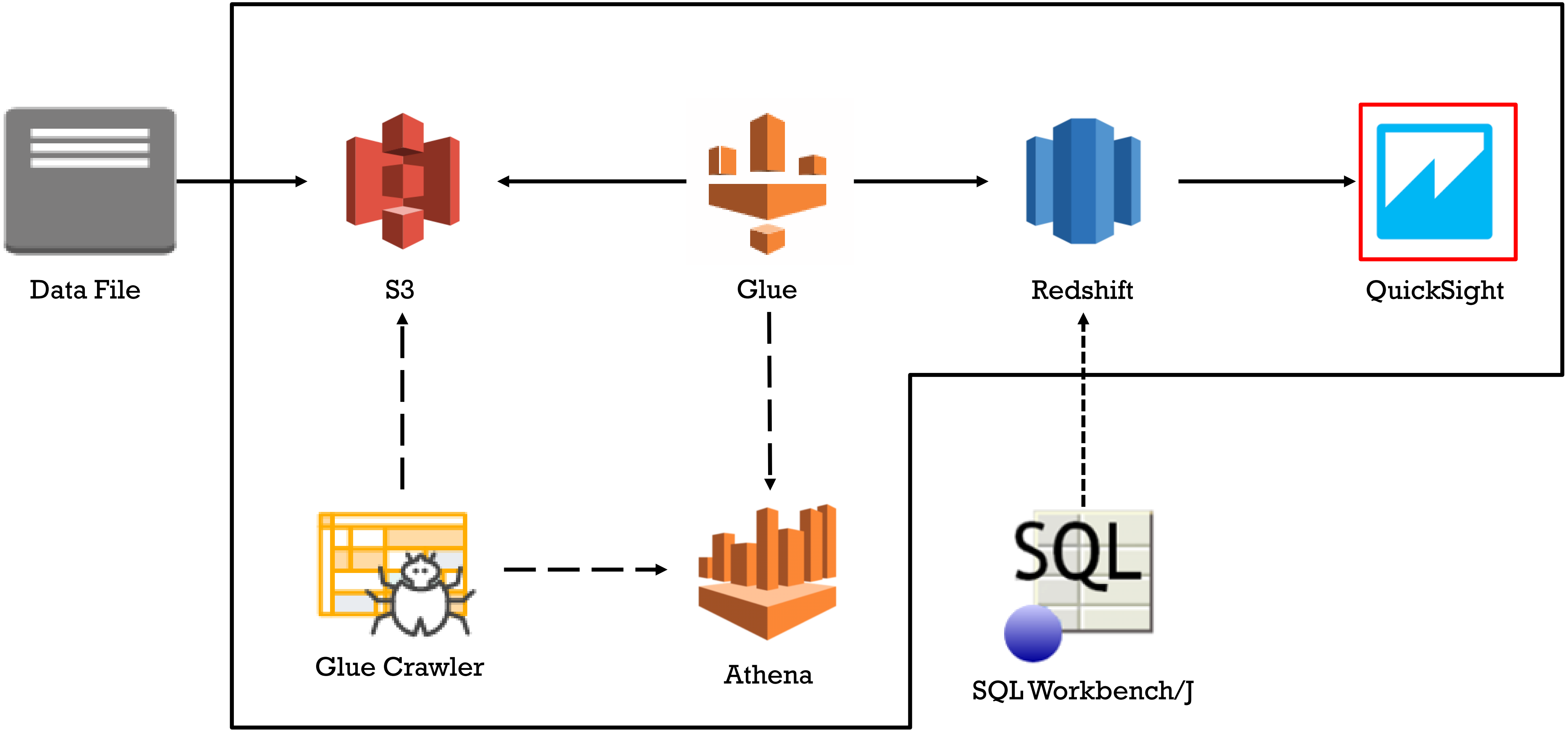
CLOUDFORMATION

Templates

```
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
```

QUICKSIGHT

AWS Business Intelligence Tool



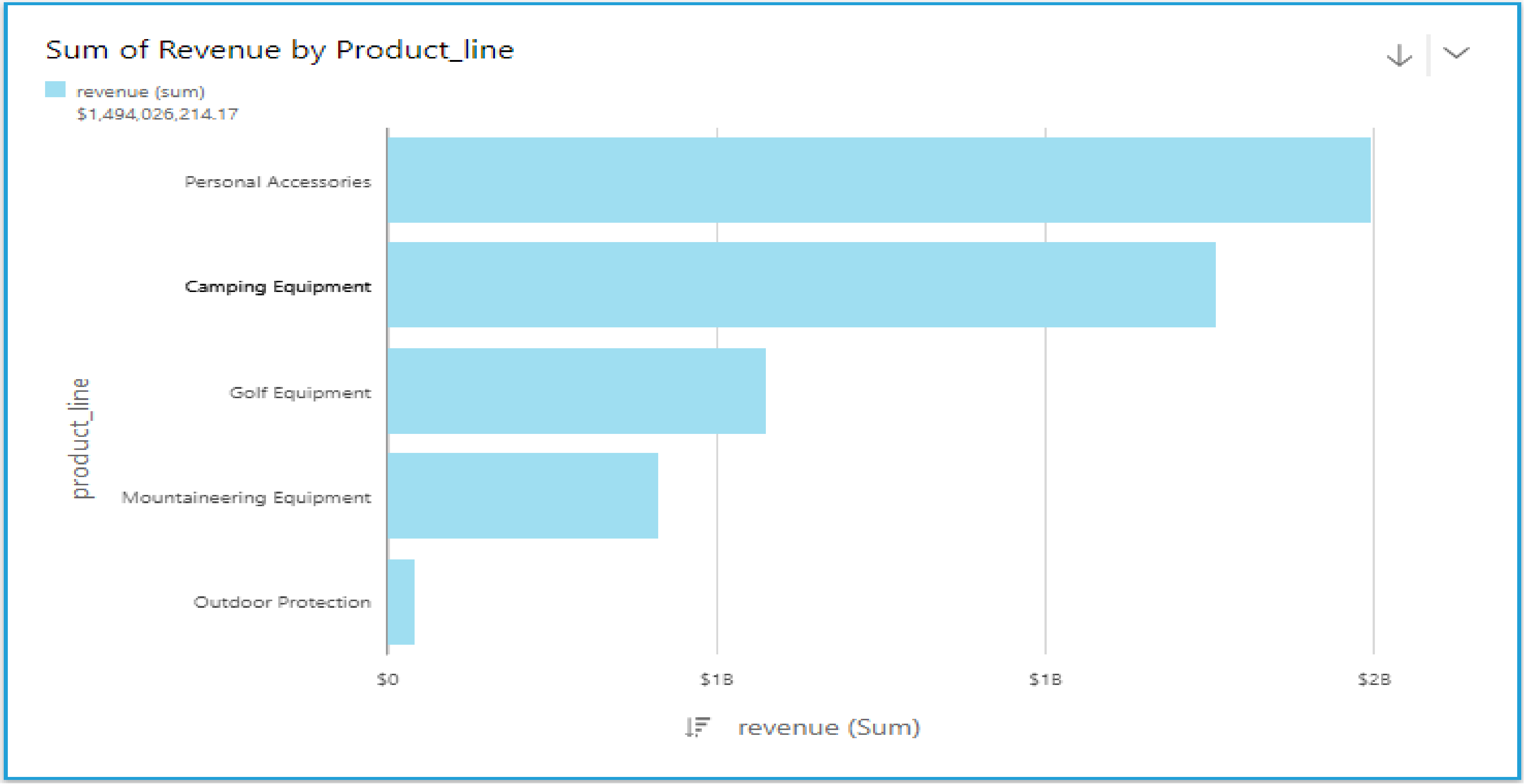
QUICKSIGHT

AWS Business Intelligence Tool

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

QUICKSIGHT

AWS Business Intelligence Tool



QUICKSIGHT

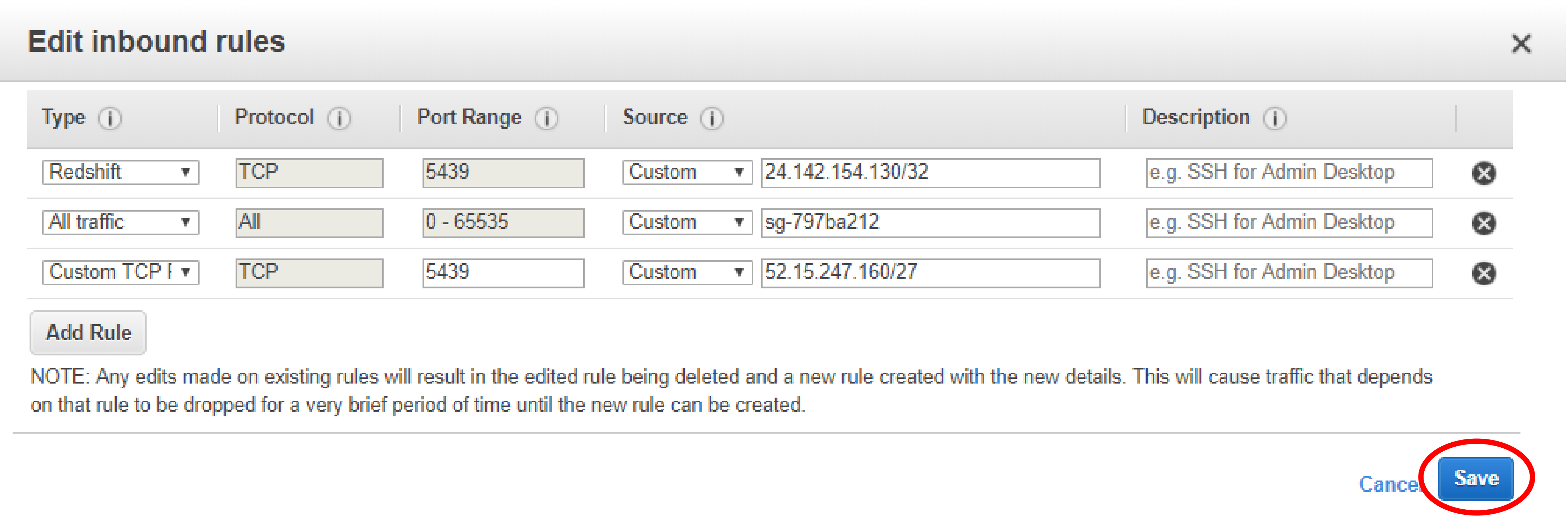
└AWS Business Intelligence Tool

Create Analysis

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

QUICKSIGHT

AWS Business Intelligence Tool



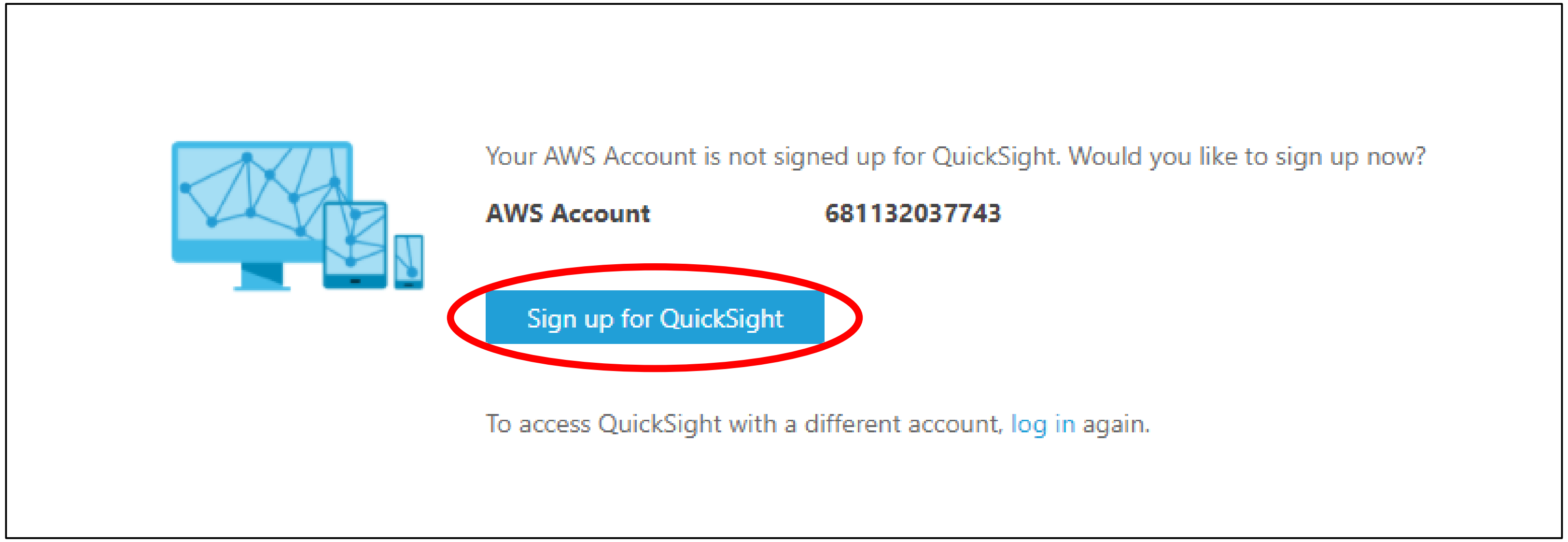
Edit inbound rules

Type	Protocol	Port Range	Source		Description	
Redshift ▾	TCP	5439	Custom ▾	24.142.154.130/32	e.g. SSH for Admin Desktop	⊗
All traffic ▾	All	0 - 65535	Custom ▾	sg-797ba212	e.g. SSH for Admin Desktop	⊗
Custom TCP F ▾	TCP	5439	Custom ▾	52.15.247.160/27	e.g. SSH for Admin Desktop	⊗

NOTE: Any edits made on existing rules will result in the edited rule being deleted and a new rule created with the new details. This will cause traffic that depends on that rule to be dropped for a very brief period of time until the new rule can be created.

QUICKSIGHT

AWS Business Intelligence Tool



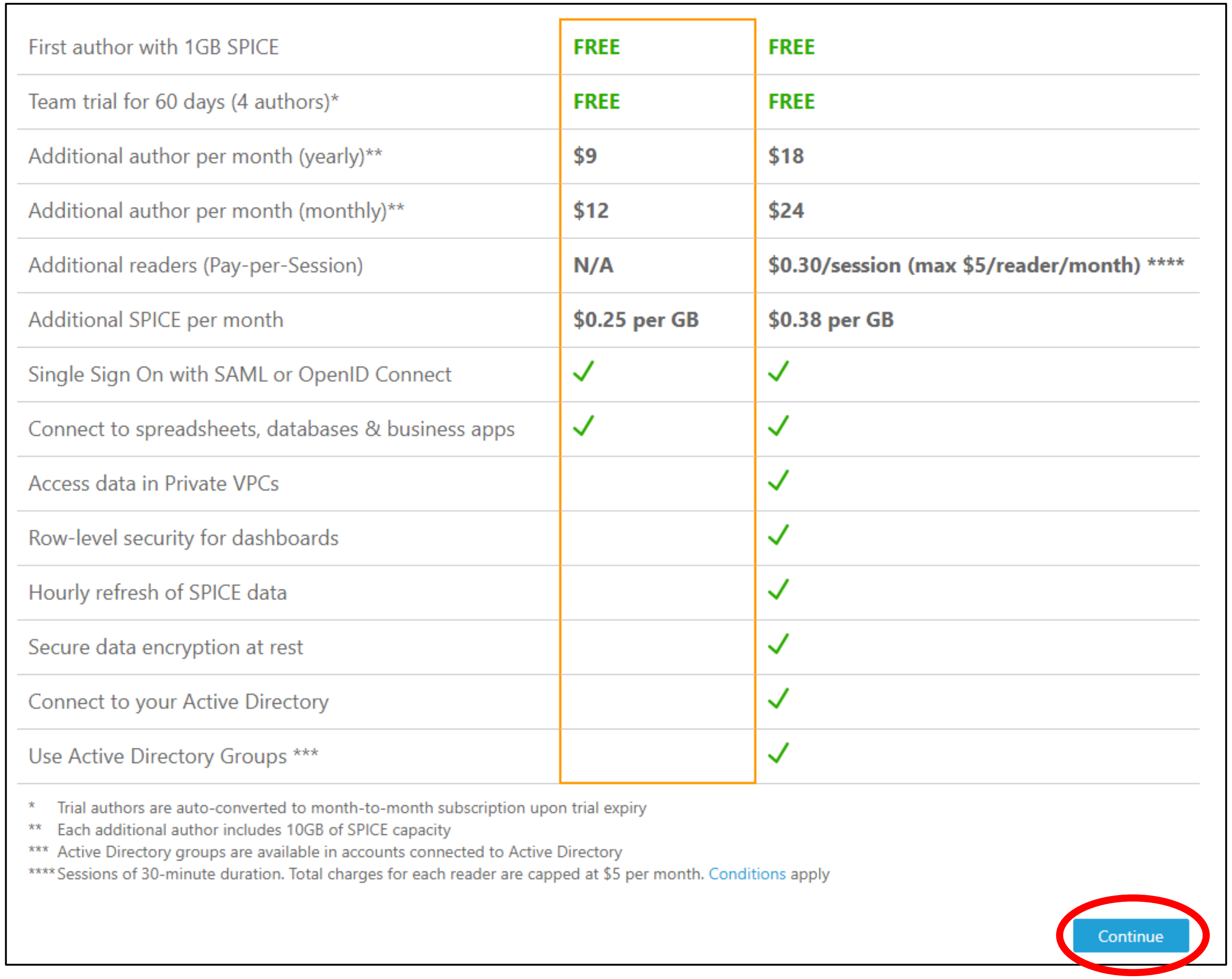
Your AWS Account is not signed up for QuickSight. Would you like to sign up now?

AWS Account **681132037743**

To access QuickSight with a different account, log in again.

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

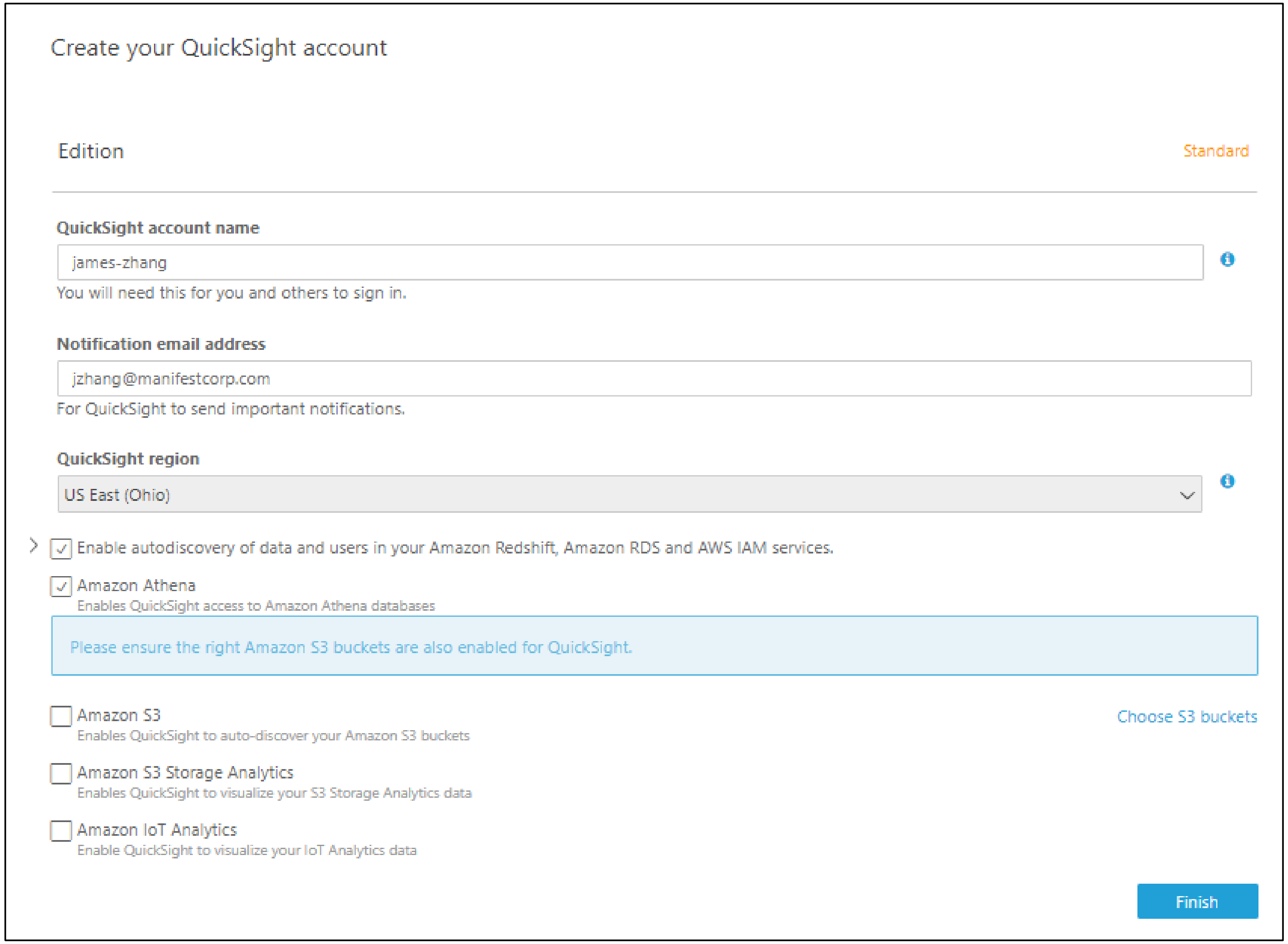
** Each additional author includes 10GB of SPICE capacity

*** 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

QUICKSIGHT

AWS Business Intelligence Tool



Create your QuickSight account

Edition

QuickSight account name

james-zhang

You will need this for you and others to sign in.

Notification email address

jzhang@manifestcorp.com

For QuickSight to send important notifications.

QuickSight region

US East (Ohio)

☑ Enable autodiscovery of data and users in your Amazon Redshift, Amazon RDS and AWS IAM services.

☑ Amazon Athena
Enables QuickSight access to Amazon Athena databases

Please ensure the right Amazon S3 buckets are also enabled for QuickSight.

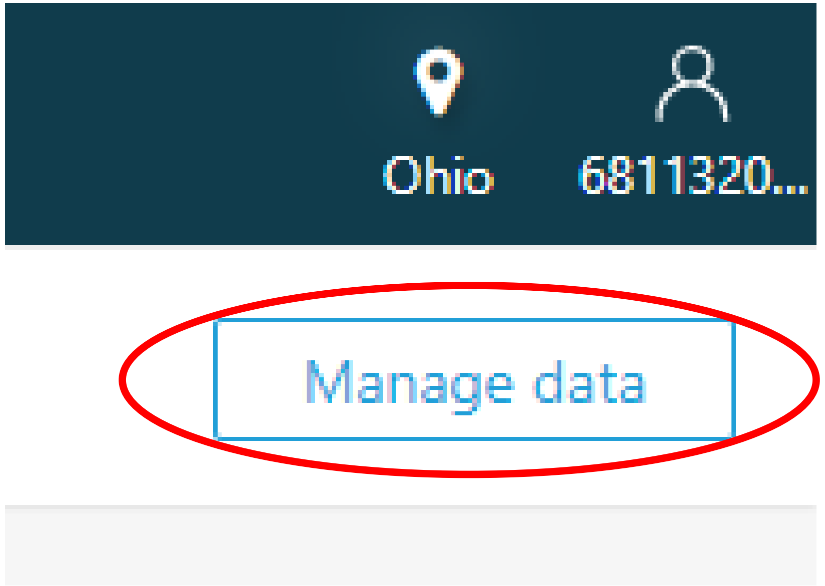
□ Amazon S3
Enables QuickSight to auto-discover your Amazon S3 buckets

□ Amazon S3 Storage Analytics
Enables QuickSight to visualize your S3 Storage Analytics data

□ Amazon IoT Analytics
Enable QuickSight to visualize your IoT Analytics data

QUICKSIGHT

└─AWS Business Intelligence Tool

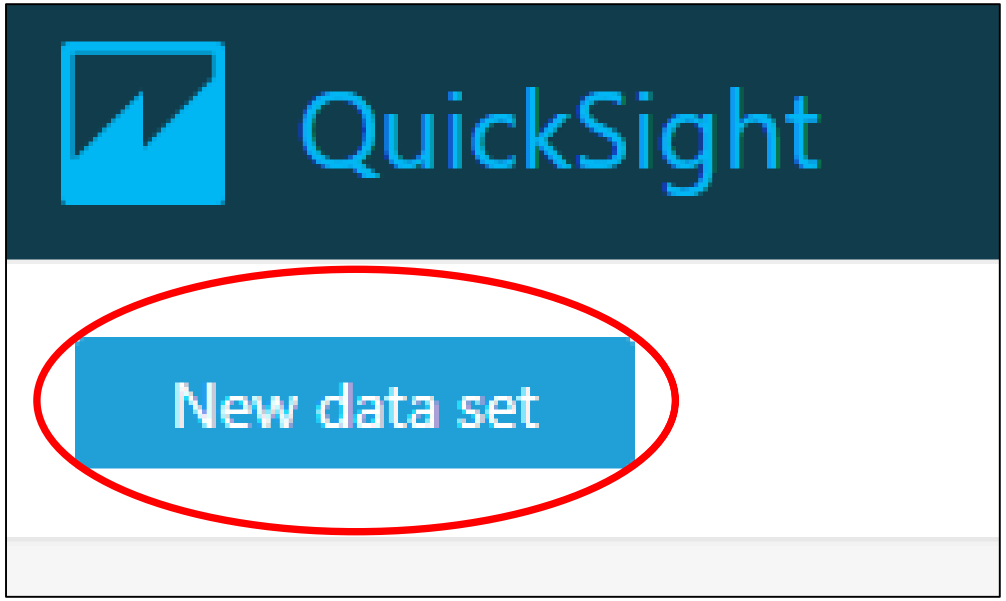


Ohio

6811320...

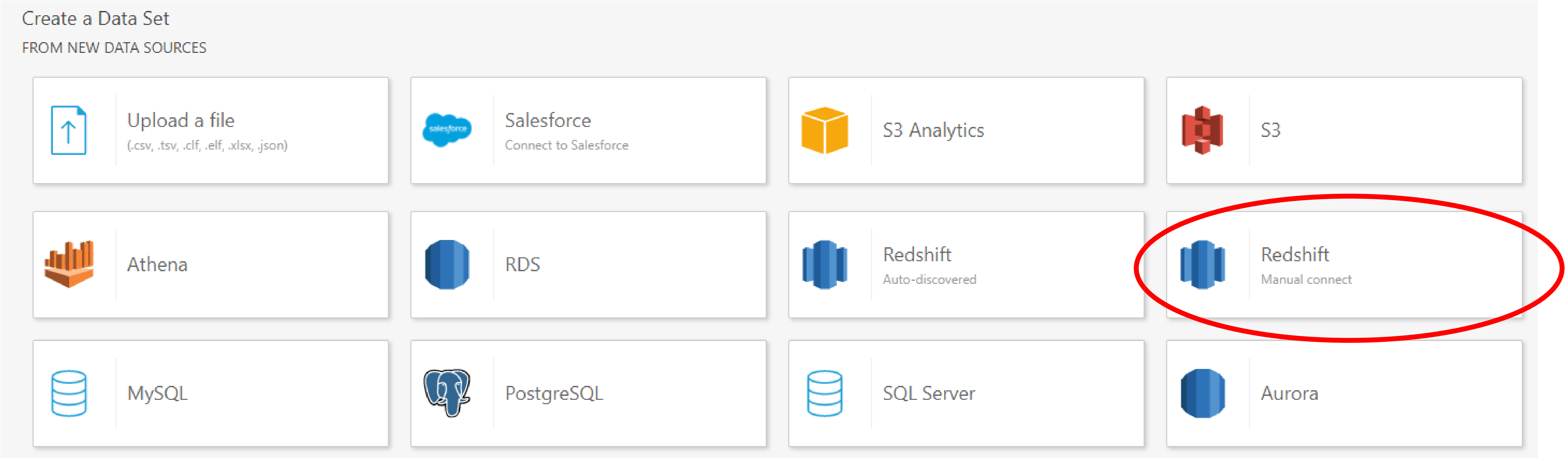
QUICKSIGHT

AWS Business Intelligence Tool



QUICKSIGHT

AWS Business Intelligence Tool

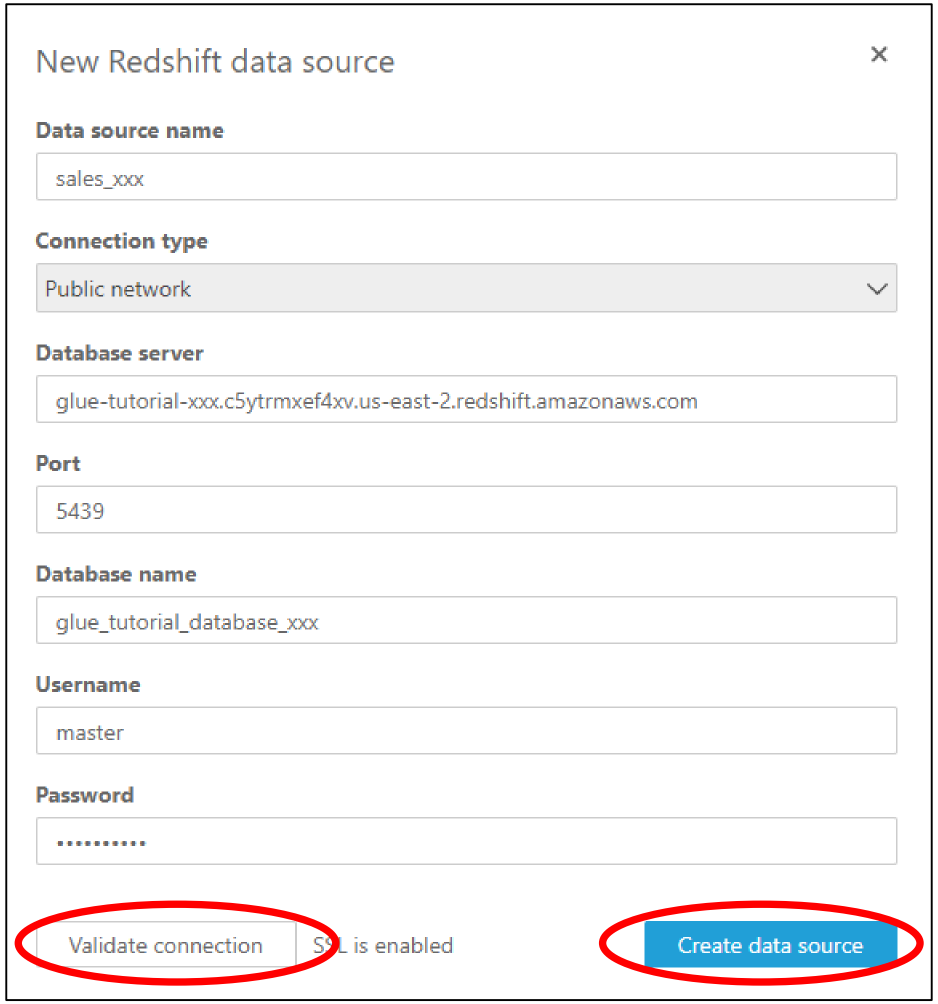


Create a Data Set

FROM NEW DATA SOURCES

QUICKSIGHT

AWS Business Intelligence Tool



New Redshift data source

Data source name: sales_xxx

Connection type: Public network

Database server: glue-tutorial-xxx.c5ytrmxef4xv.us-east-2.redshift.amazonaws.com

Port: 5439

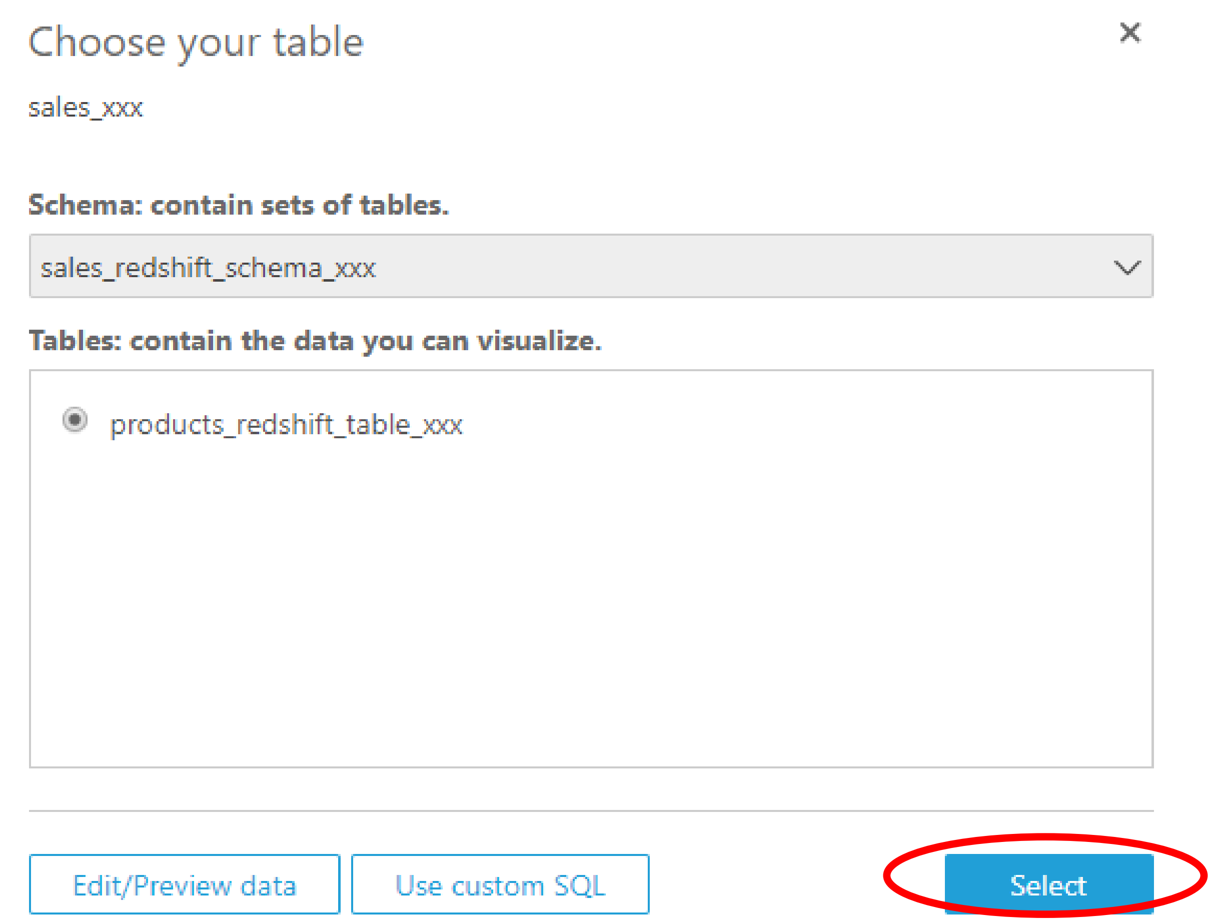
Database name: glue_tutorial_database_xxx

Username: master

Password: ••••••••••

QUICKSIGHT

└─AWS Business Intelligence Tool



Choose your table

sales_xxx

Schema: contain sets of tables.

sales_redshift_schema_xxx

Tables: contain the data you can visualize.

● products_redshift_table_xxx

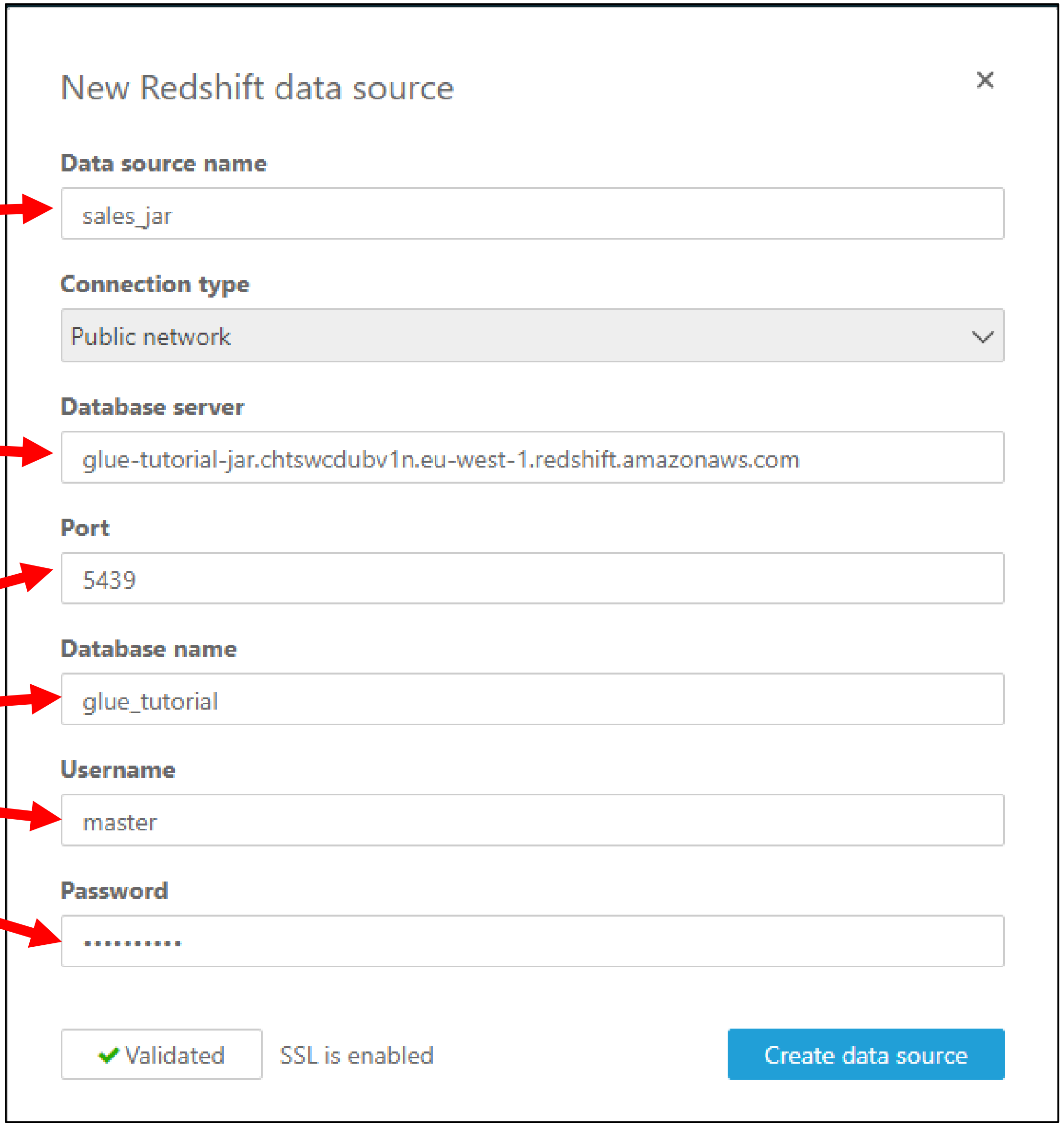
QUICKSIGHT

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



New Redshift data source

Data source name: sales_jar

Connection type: Public network

Database server: glue-tutorial-jar.chtswcdubv1n.eu-west-1.redshift.amazonaws.com

Port: 5439

Database name: glue_tutorial

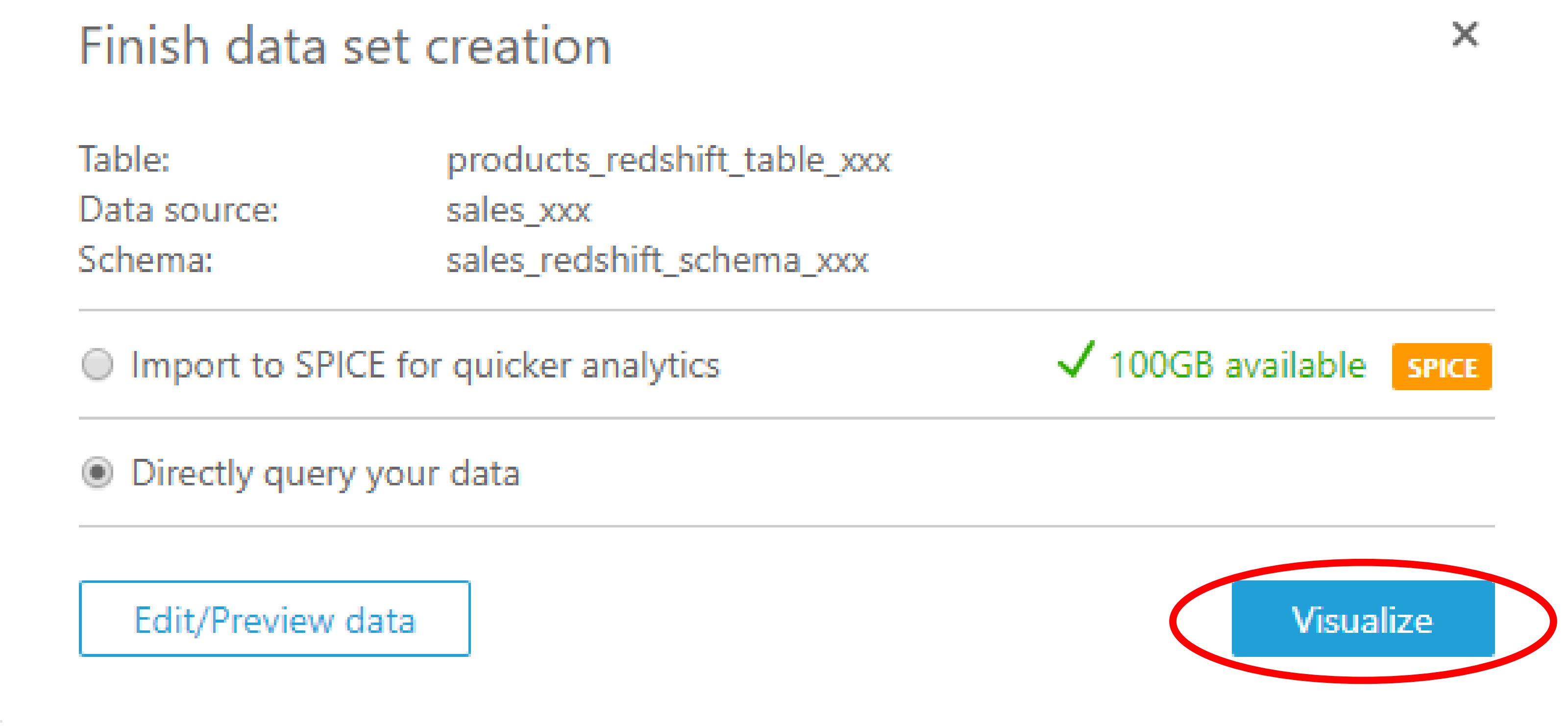
Username: master

Password: ••••••••••

✓ Validated SSL is enabled

QUICKSIGHT

AWS Business Intelligence Tool



Finish data set creation

Table:	products_redshift_table_xxx
Data source:	sales_xxx
Schema:	sales_redshift_schema_xxx

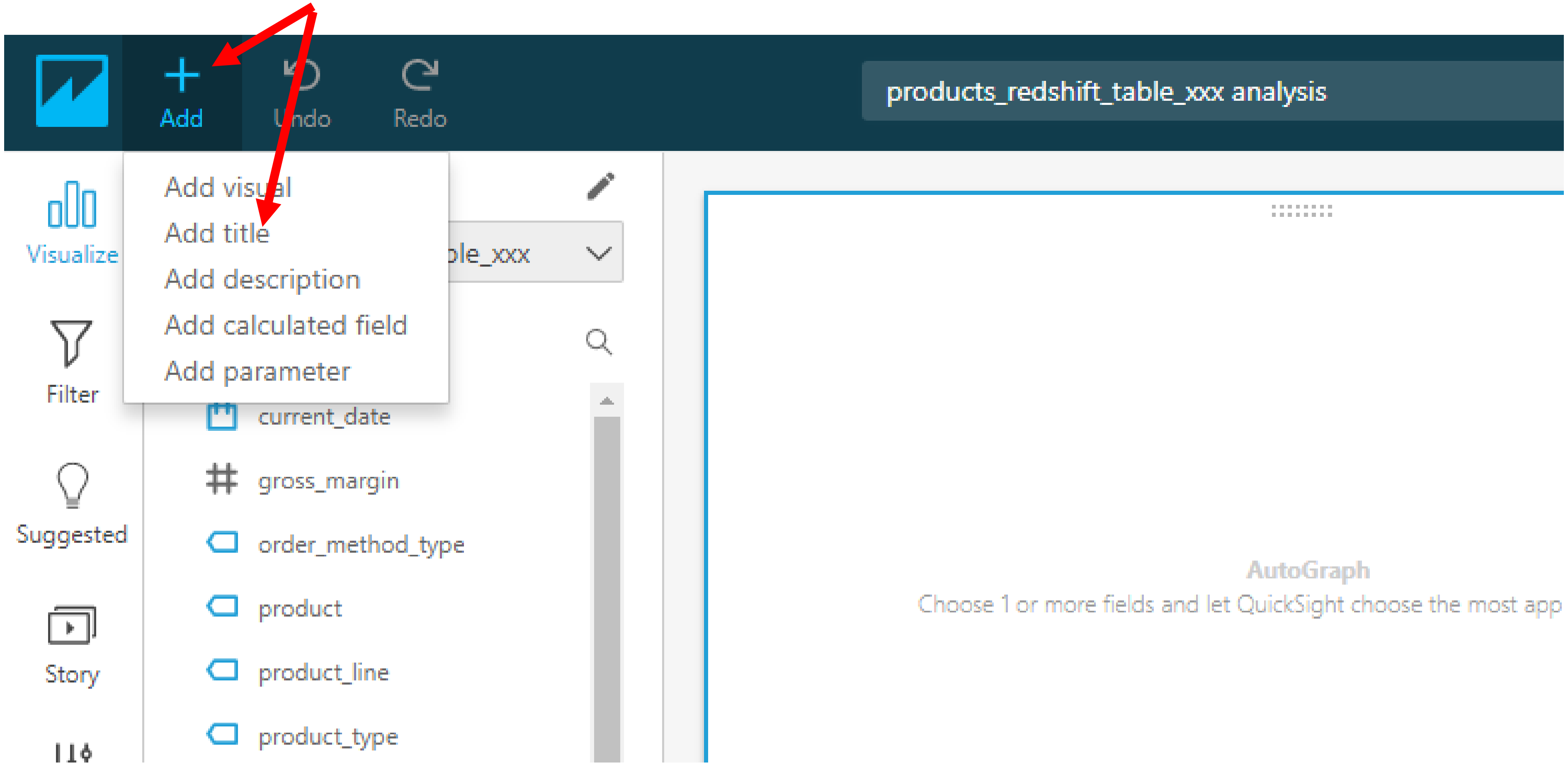
○ Import to SPICE for quicker analytics ✓ 100GB available SPICE

● Directly query your data

QUICKSIGHT

AWS Business Intelligence Tool

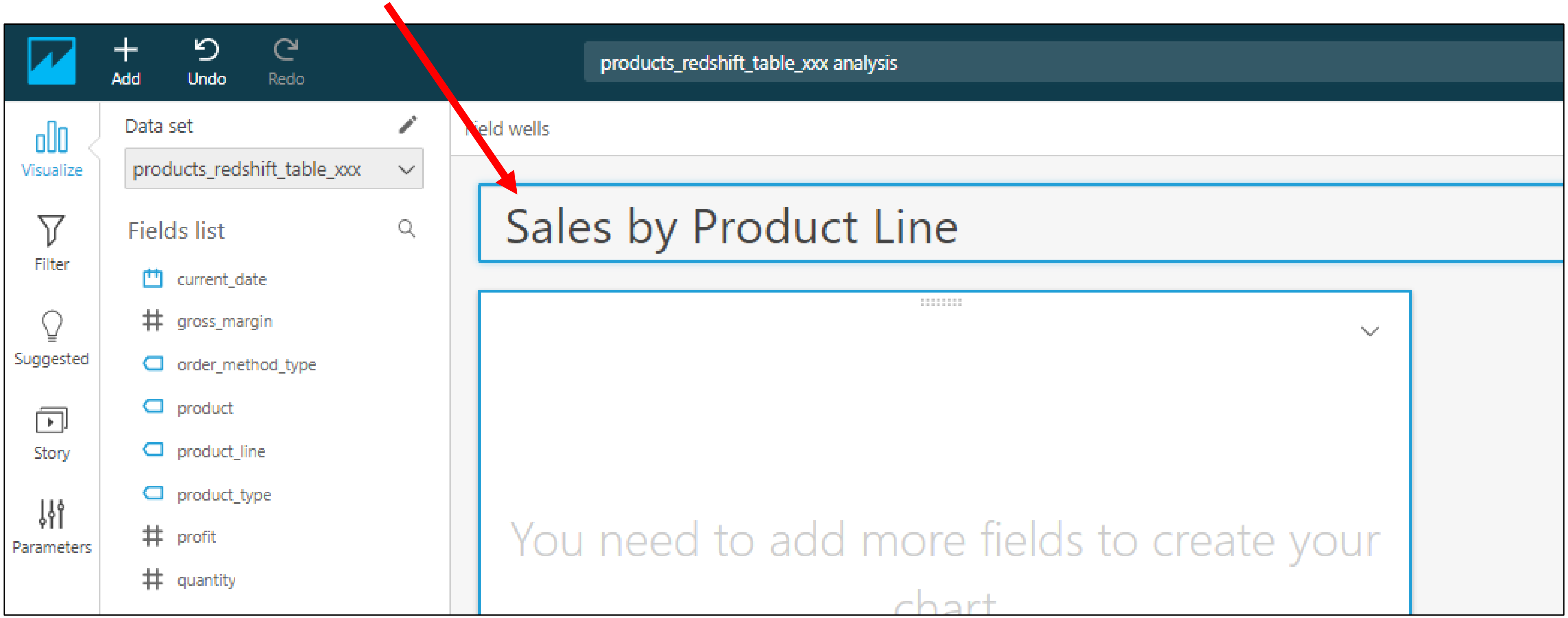
Select Add > Add title



QUICKSIGHT

AWS Business Intelligence Tool

Enter title

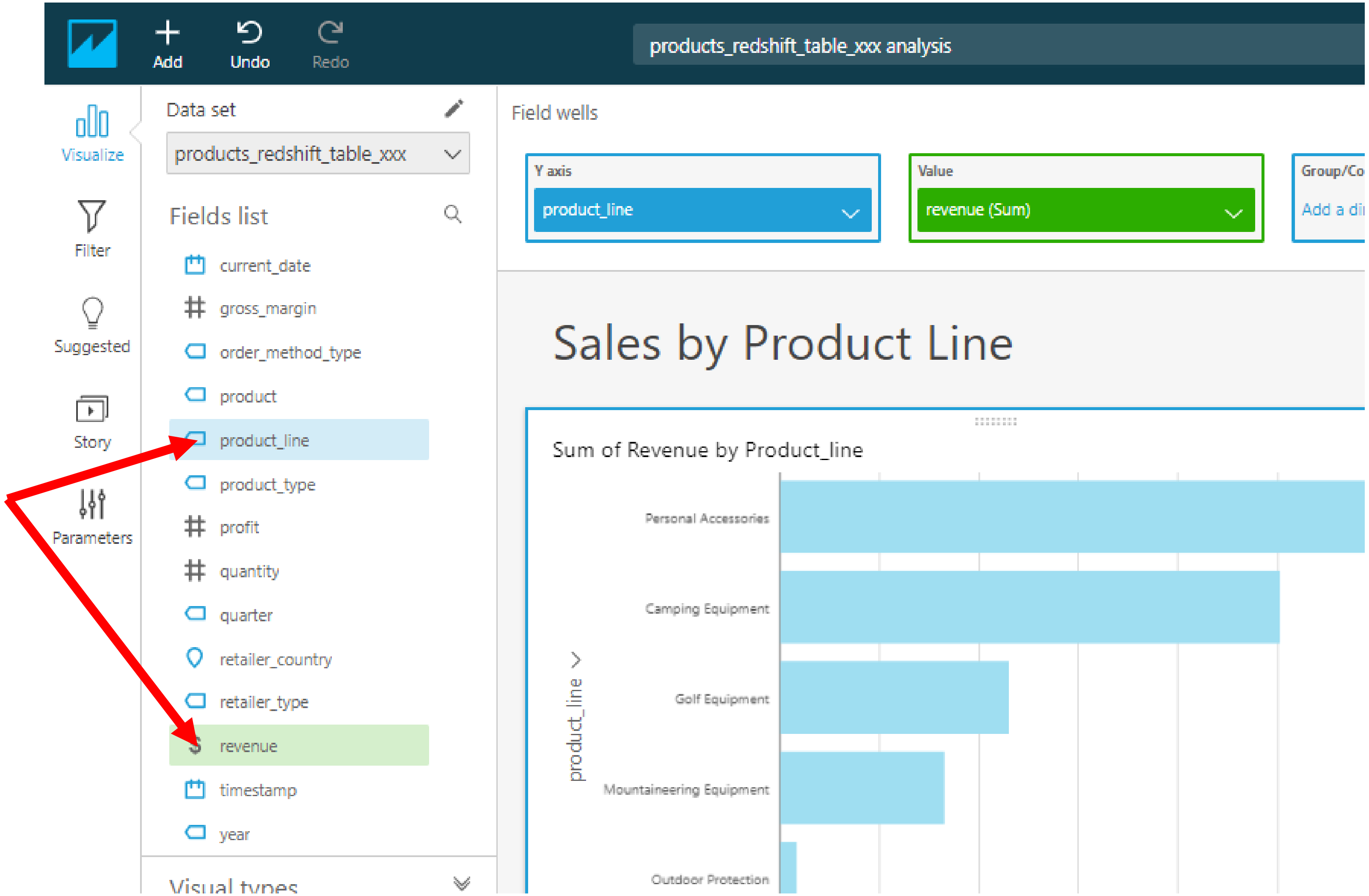


Sales by Product Line

You need to add more fields to create your chart

QUICKSIGHT

AWS Business Intelligence Tool



products_redshift_table_xxx analysis

Data set: products_redshift_table_xxx

Field wells

Y axis: product_line

Value: revenue (Sum)

Sales by Product Line

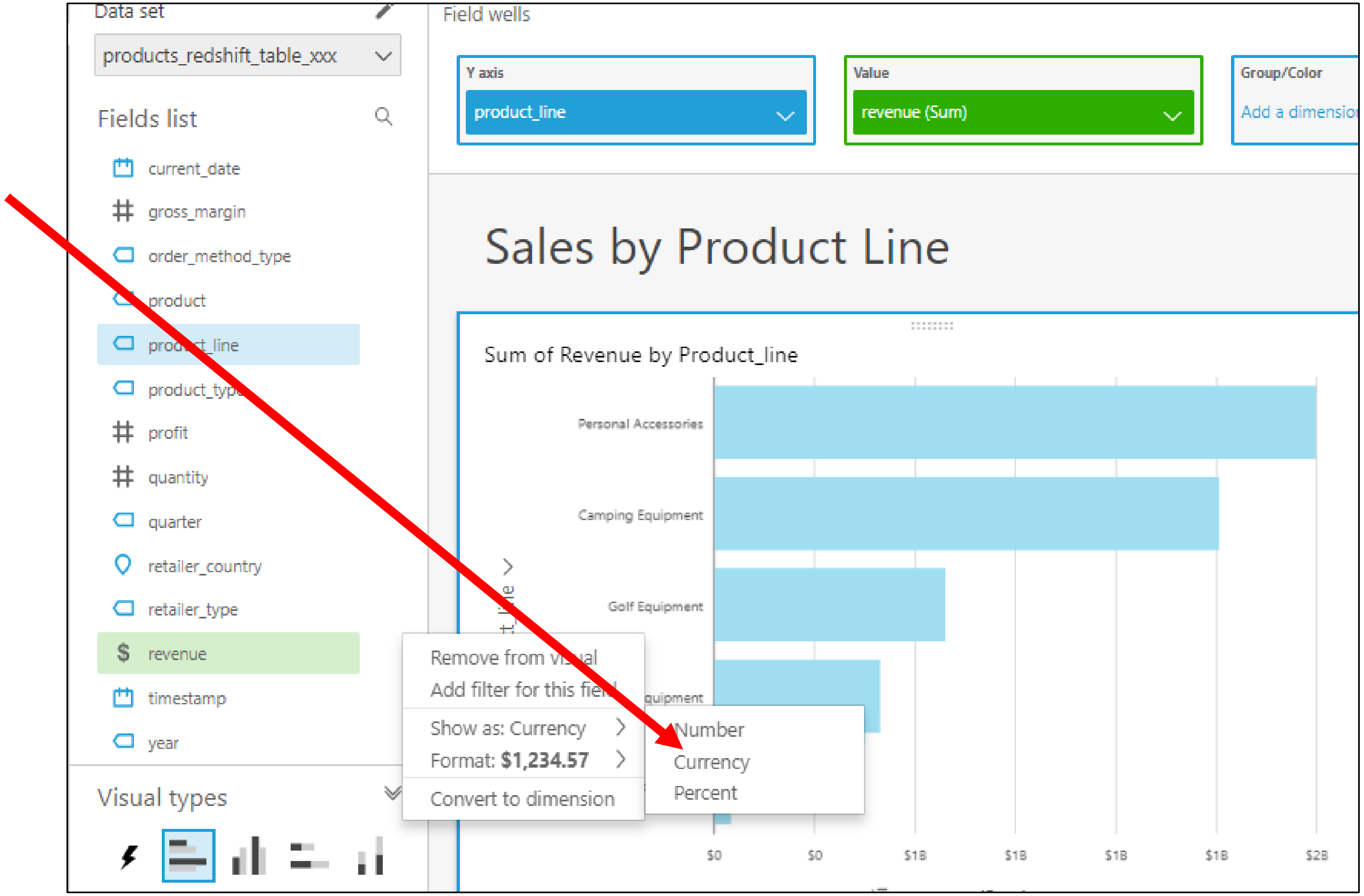
Sum of Revenue by Product_line

Choose product_line and revenue

QUICKSIGHT

AWS Business Intelligence Tool

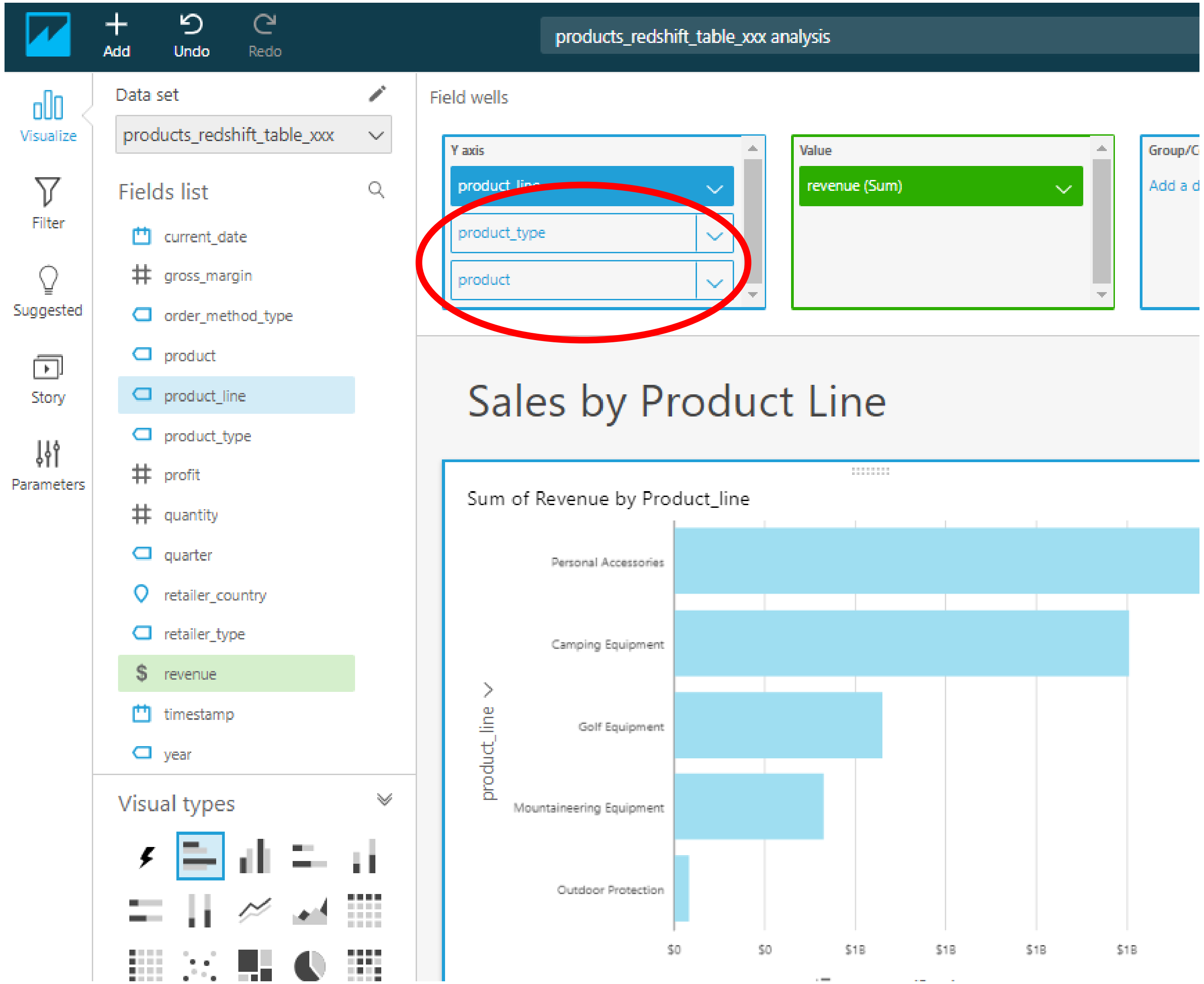
Change the format of Revenue to Currency



QUICKSIGHT

AWS Business Intelligence Tool

Add product_type and product as drill down layer



products_redshift_table_xxx analysis

Data set: products_redshift_table_xxx

Fields list

- current_date
- gross_margin
- order_method_type
- product
- product_line
- product_type
- profit
- quantity
- quarter
- retailer_country
- retailer_type
- revenue
- timestamp
- year

Field wells

Y axis: product_line, product_type, product

Value: revenue (Sum)

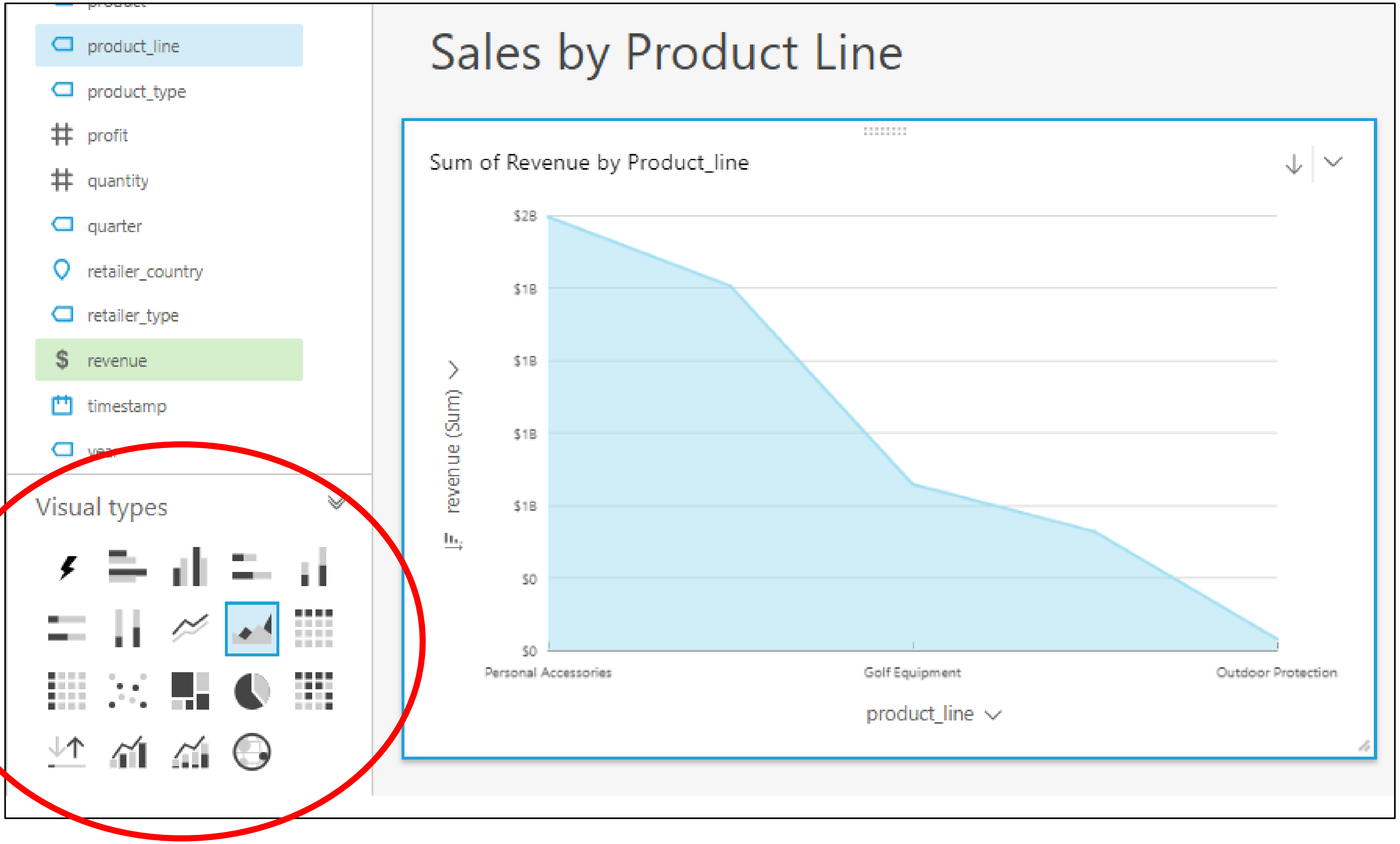
Group/C… Add a d…

Sales by Product Line

QUICKSIGHT

AWS Business Intelligence Tool

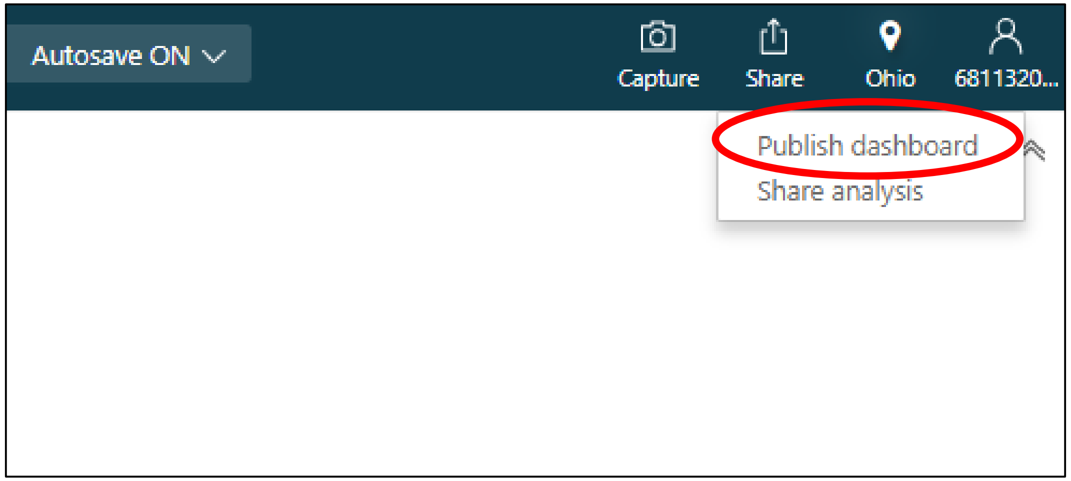
Change Visual Type



QUICKSIGHT

AWS Business Intelligence Tool

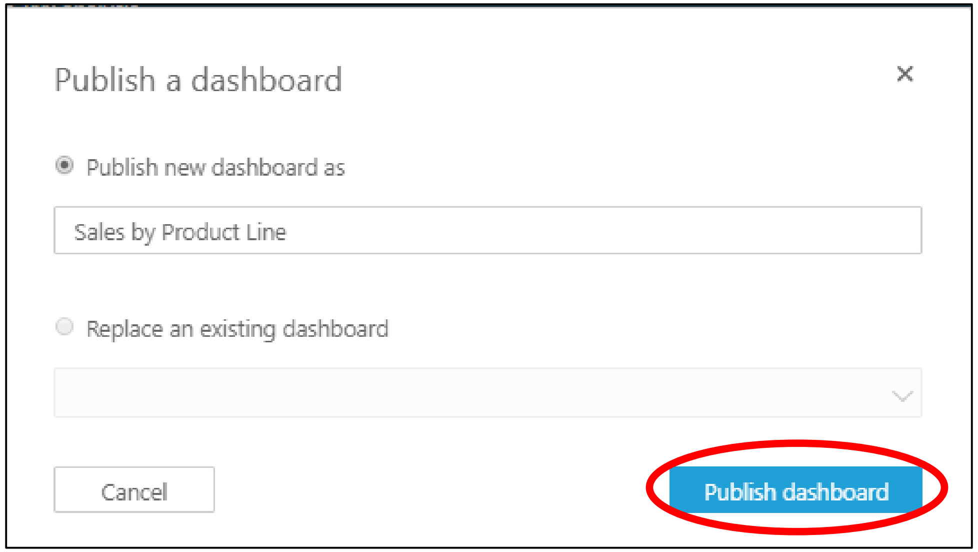
Publish to Dashboard



QUICKSIGHT

AWS Business Intelligence Tool

Name the Dashboard and select Publish dashboard



Publish a dashboard

● Publish new dashboard as

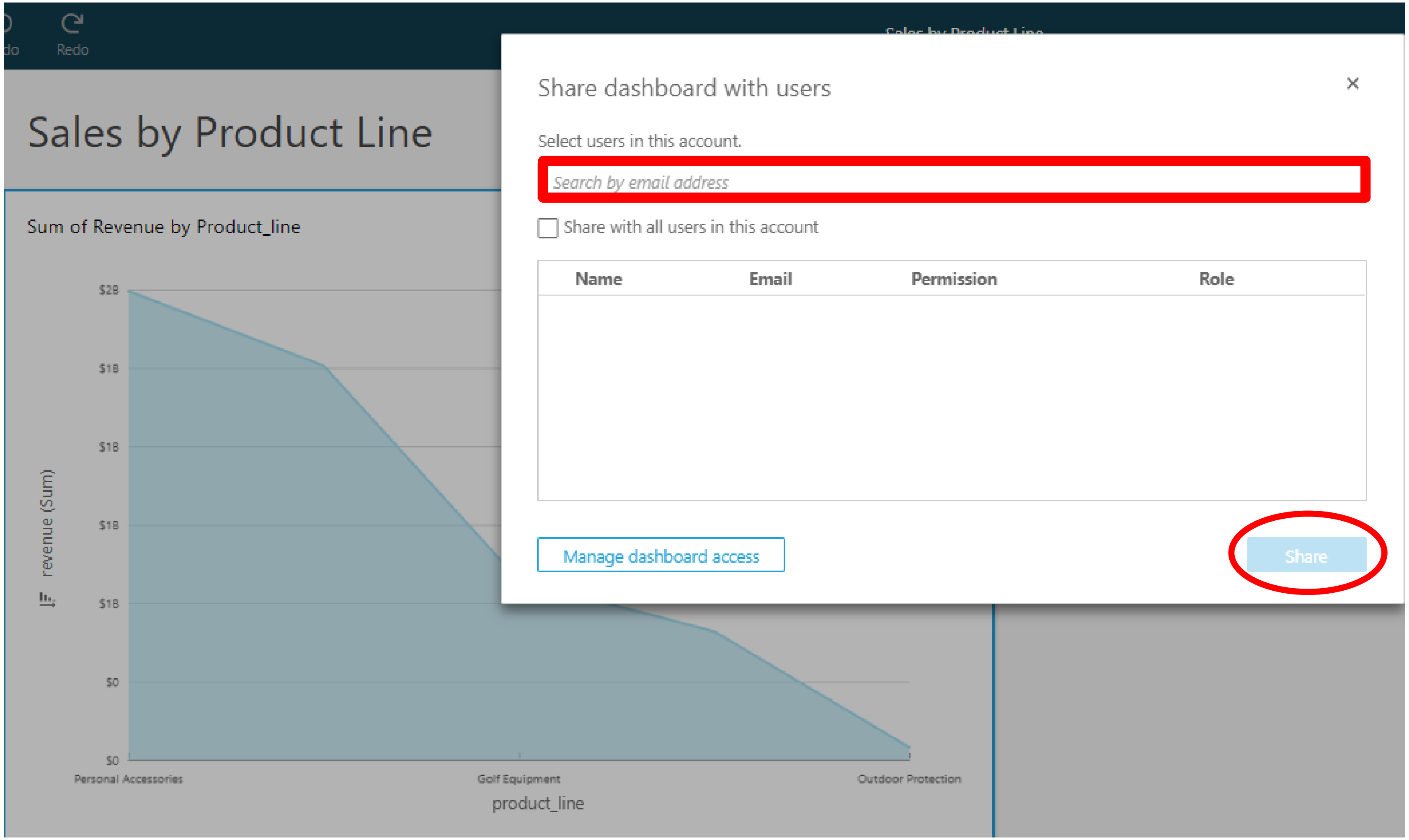
Sales by Product Line

○ Replace an existing dashboard

QUICKSIGHT

AWS Business Intelligence Tool

Share the dashboard



Share dashboard with users ×

Select users in this account.

Search by email address

☐ Share with all users in this account

Name	Email	Permission	Role

Manage dashboard access

Share

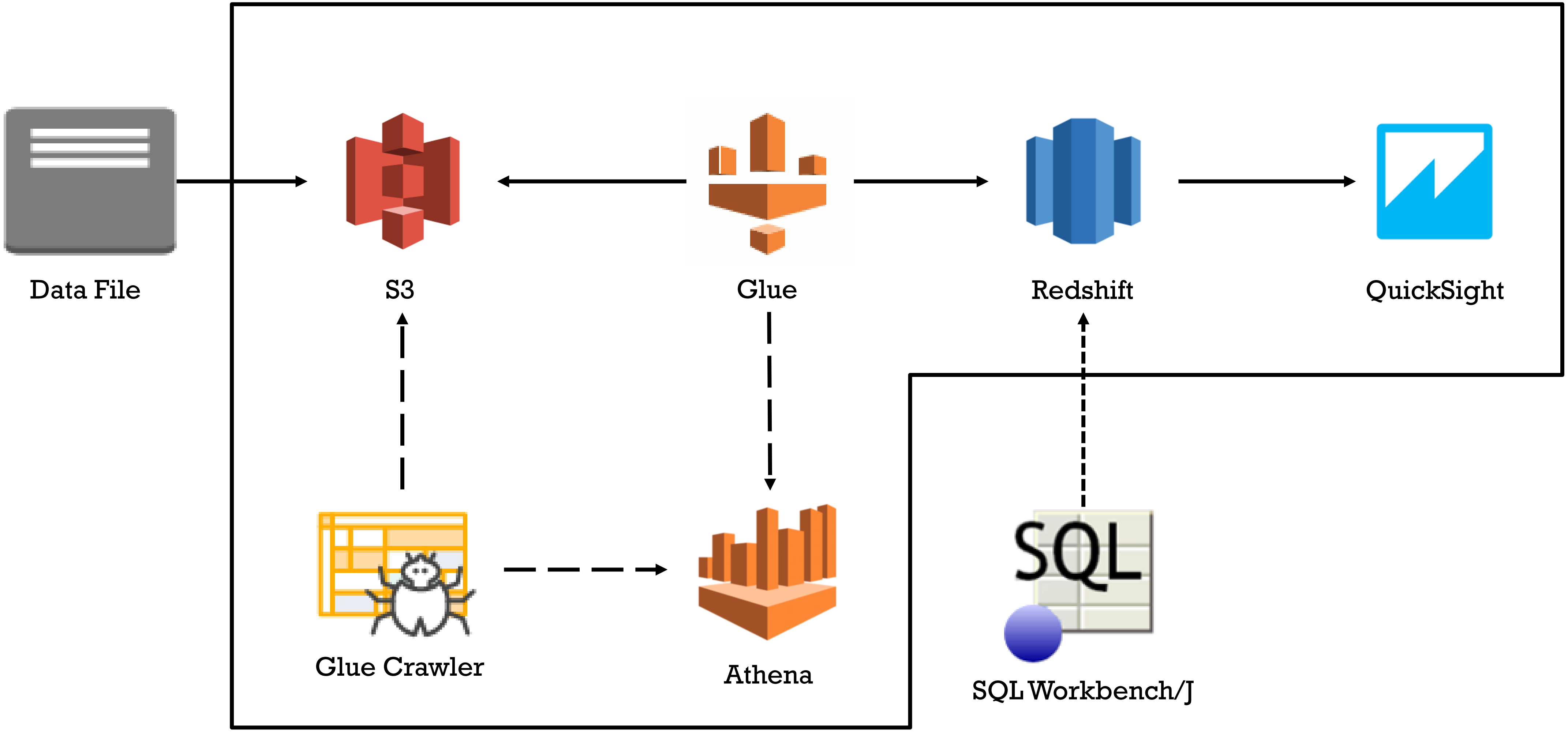
Lab 6

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

(Use US-EAST-2/Ohio Region)

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 *
- Glue Job
- Glue Database
- Glue Table
- Glue Connection

* 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/