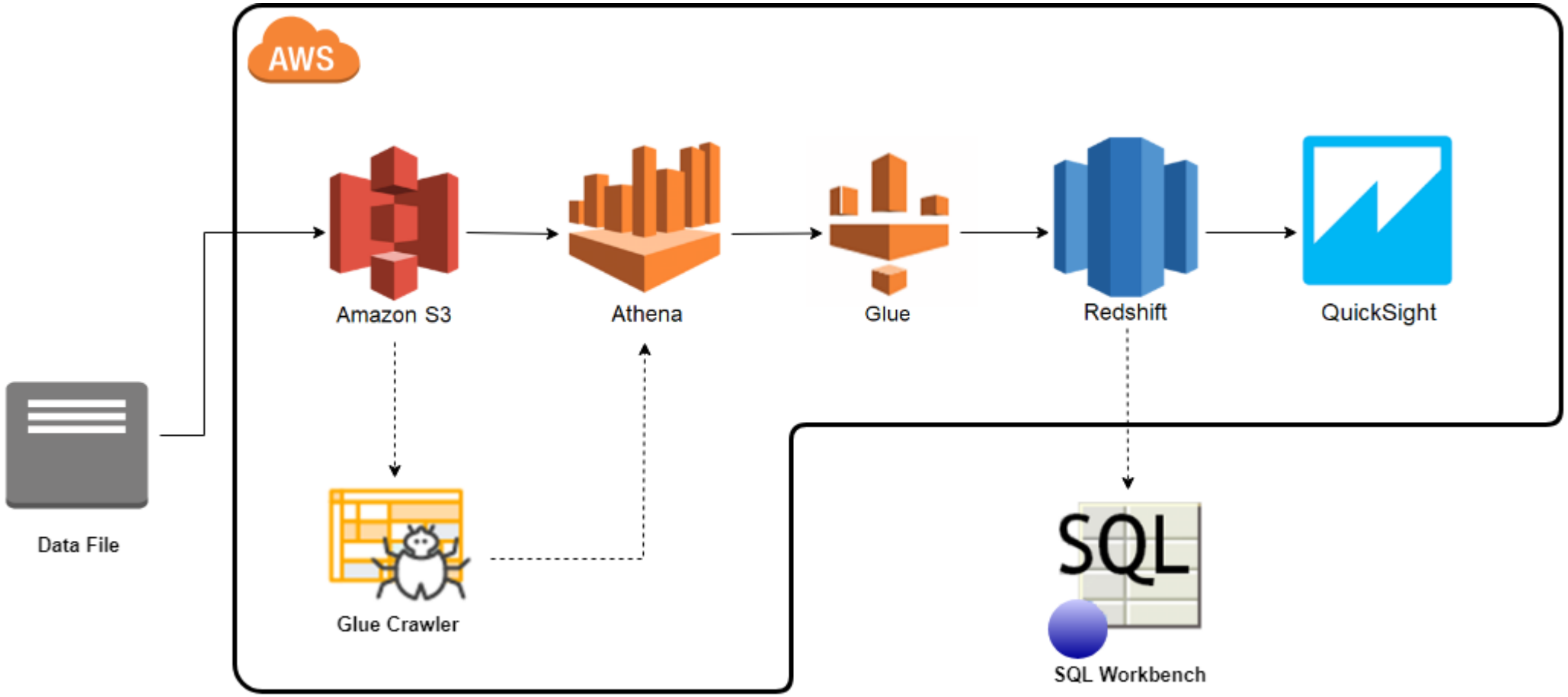
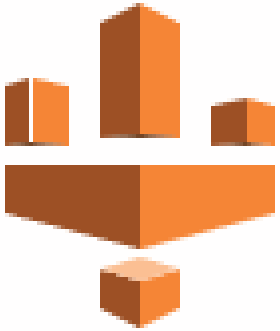
ARTS & CRAFTS WITH AWS GLUE

ETL Workshop

Amazon Web Services



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
- Adding a timestamp.

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 Software library used to store and transform large amounts of 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

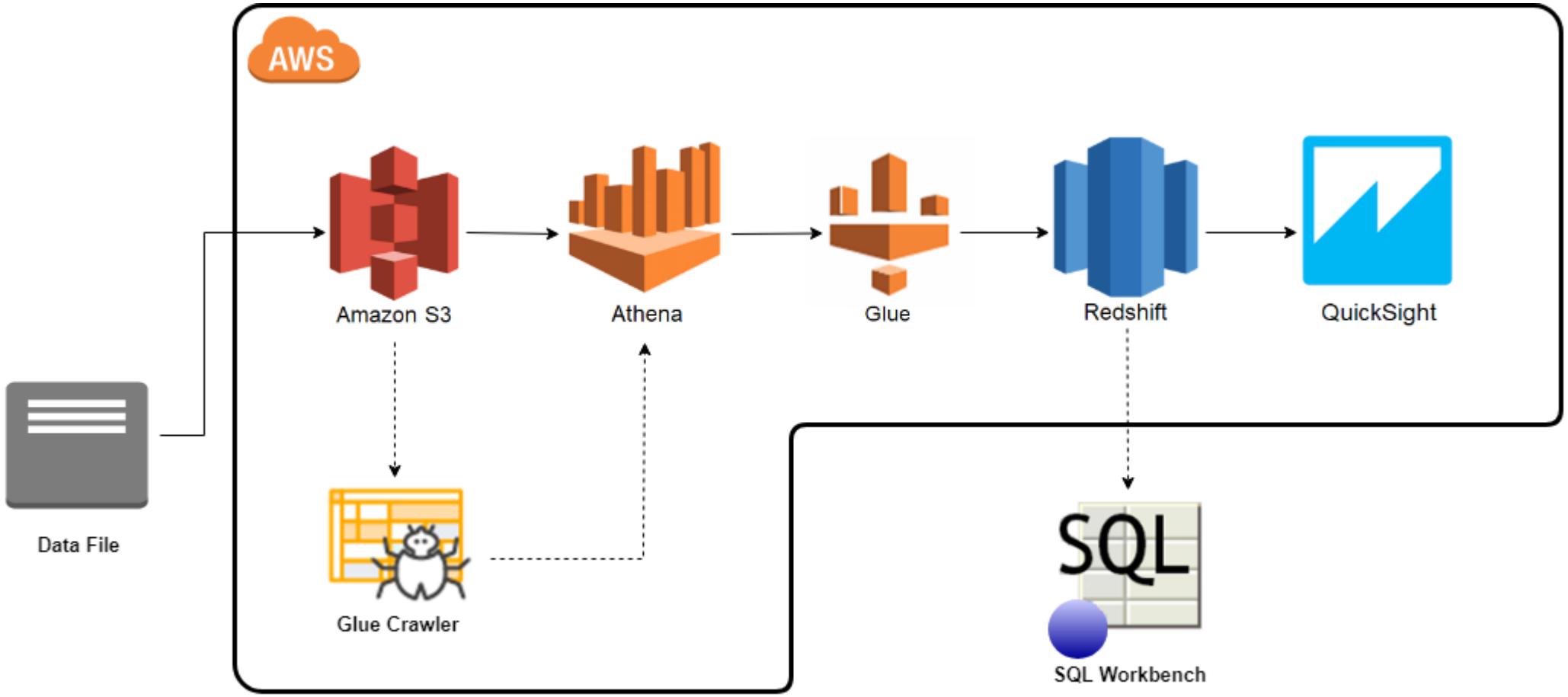
Glue Tutorial Overview

- Setup Redshift Cluster
- S3 bucket for storing the file
- Athena table to access data in file
- Glue connection
- Glue job
- Redshift connection
- Redshift tables
- Run glue job
- QuickSight

Glue Tutorial Prerequisites

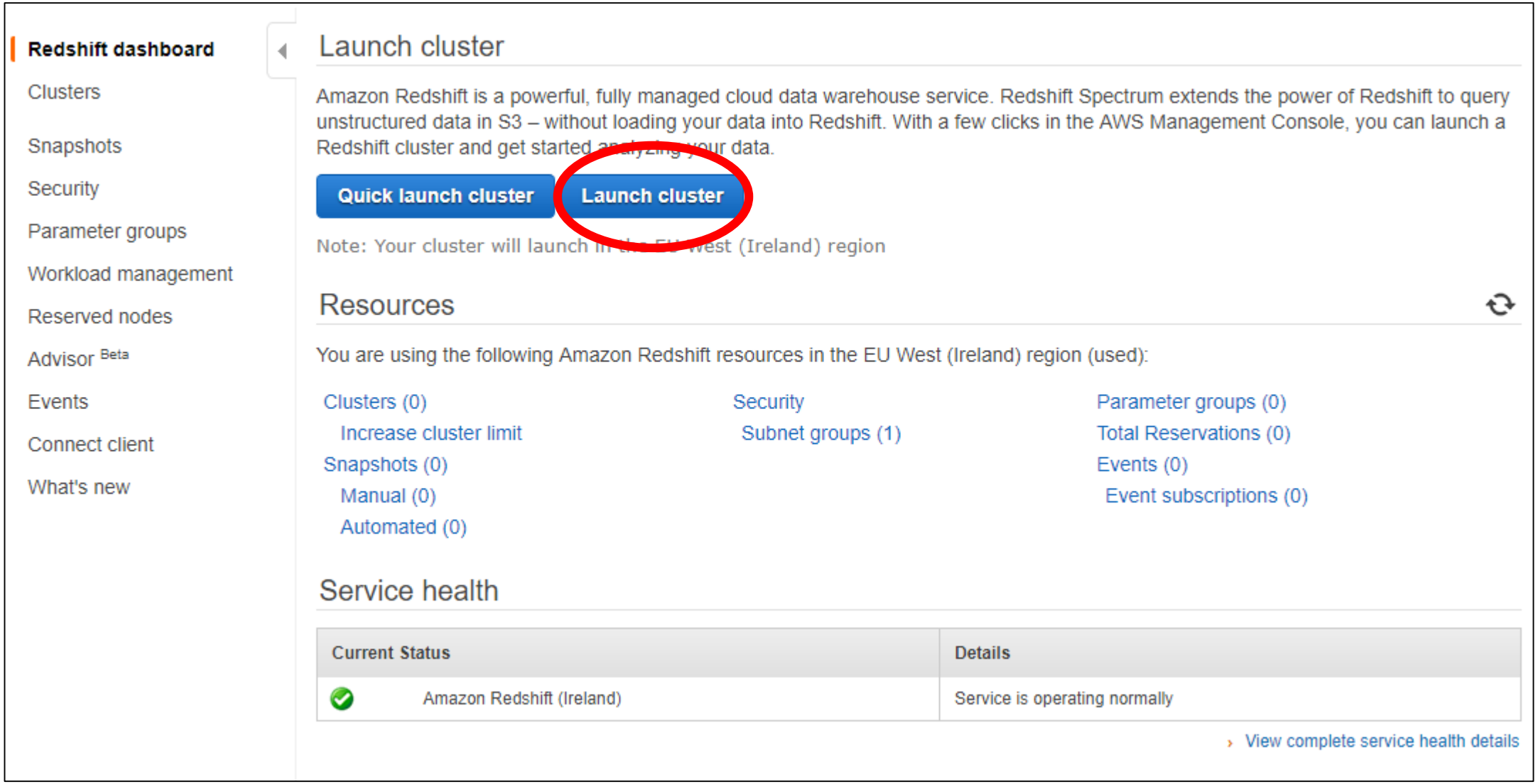
- Prerequisites :
  - Setup AWS Account
  - Clone or save git repository link https://github.com/jackdsilverman/aws-glue-tutorial.git
  - download SQL Workbench/j https://www.sql-workbench.eu/
  - download Redshift jdbc driver https://docs.aws.amazon.com/redshift/latest/mgmt/configure-jdbc-connection.html#download-jdbc-driver

Amazon Web Services



Redshift

Create AWS Data Warehouse



Redshift

Create AWS Datawarehouse

Specify Cluster Name

Give your cluster a Database to start with

Create a user

Create a password for the user

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

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

CLUSTER DETAILS NODE CONFIGURATION ADDITIONAL CONFIGURATION REVIEW

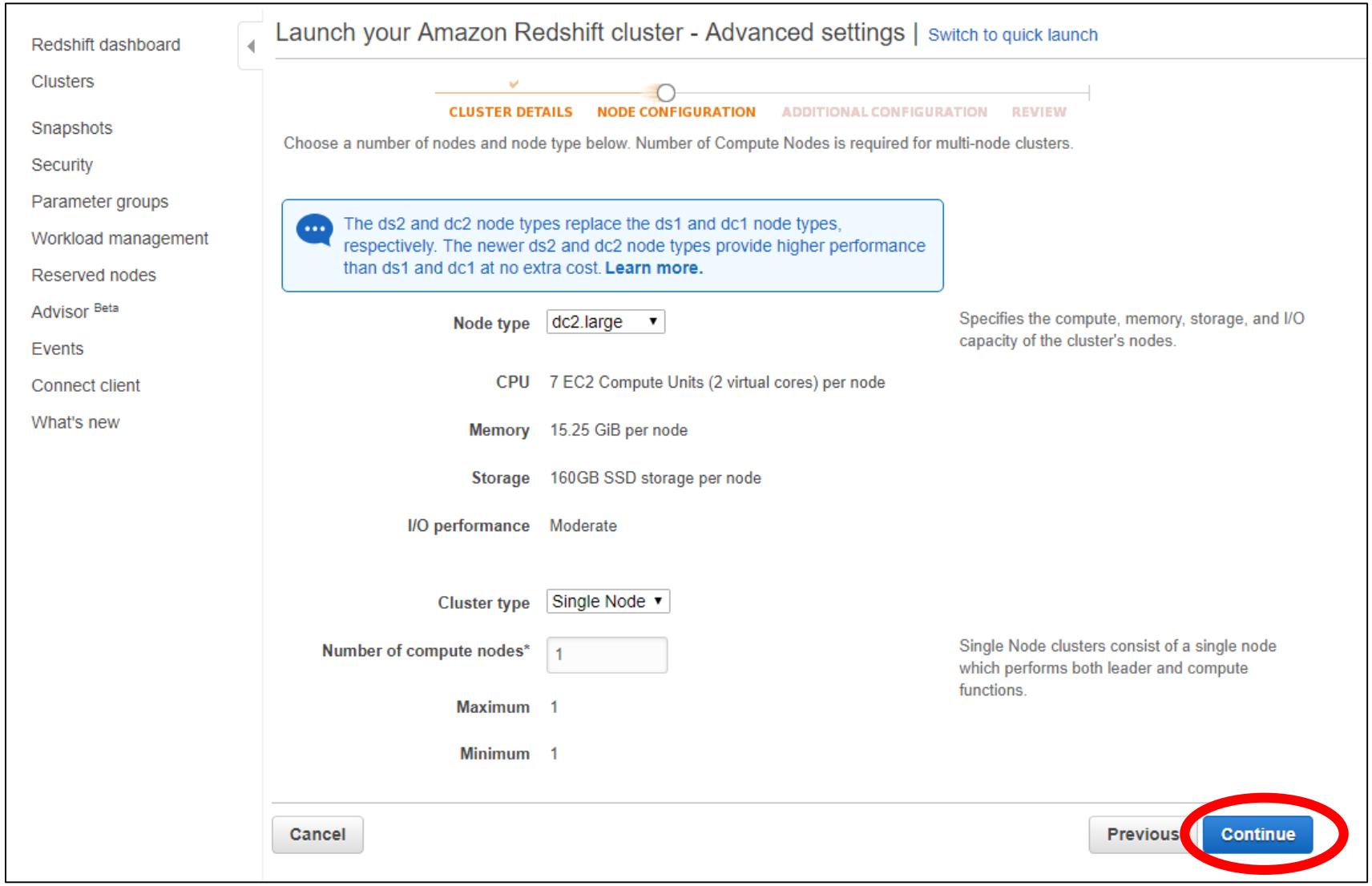
Provide the details of your cluster. Fields marked with * are required.

Cluster identifier*	glue-tutorial-jds	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	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

Cancel Continue

Redshift

Create AWS Datawarehouse



Redshift

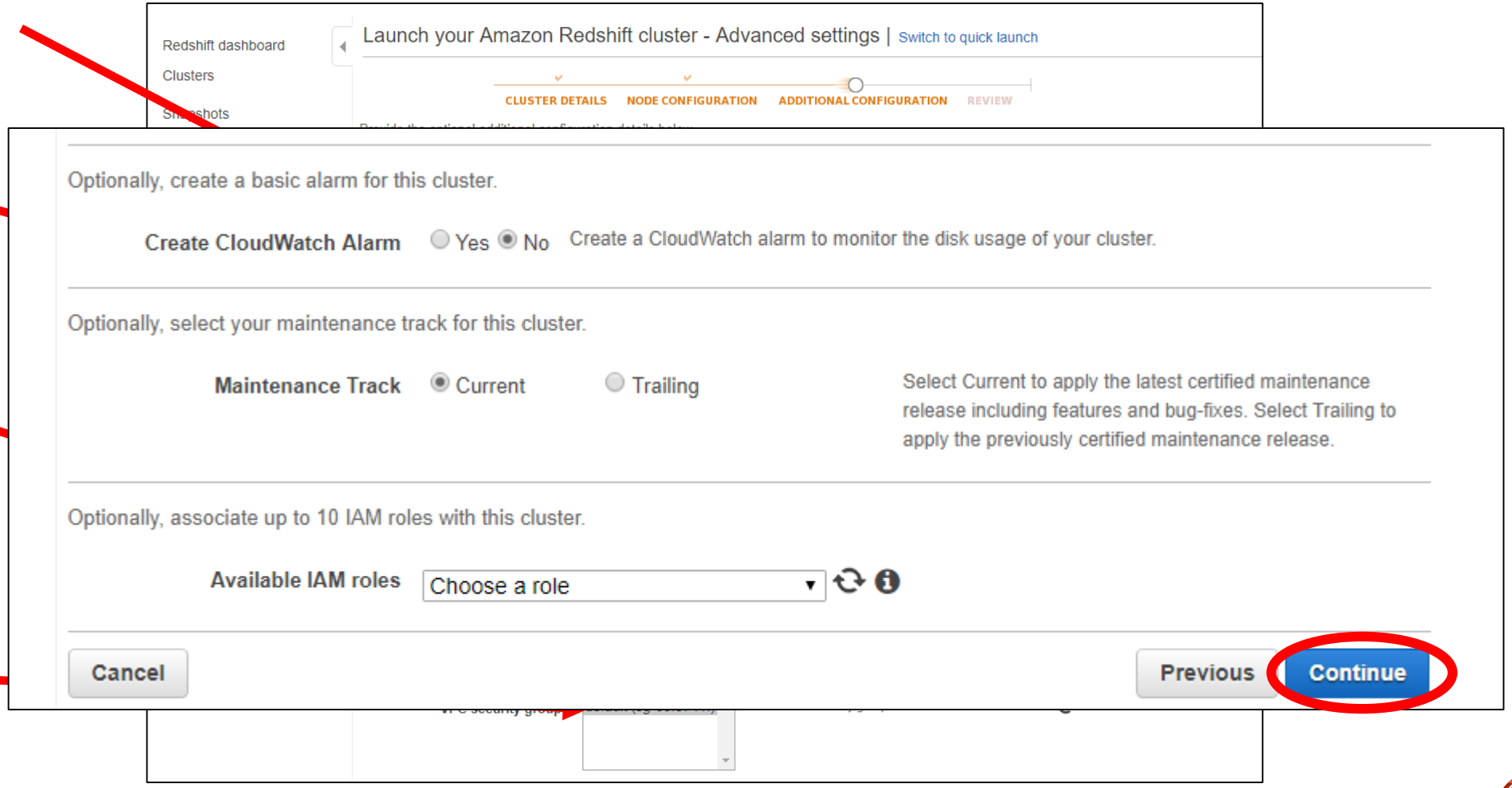
Create AWS Datawarehouse

Choose default VPC

Choose default subnet group

Choose subnet availability zone

Choose default security group



Redshift

Create AWS Datawarehouse

Redshift dashboard
Clusters
Snapshots
Security
Parameter groups

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

CLUSTER DETAILS NODE CONFIGURATION ADDITIONAL CONFIGURATION REVIEW

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 .

Cancel

Previous

Launch cluster

Elastic IP: Not used

VPC security groups default (sg-63f5741f)

Enhanced VPC Routing: No

Encrypt database: No

Redshift

Create AWS Data warehouse

Clusters

Quick launch cluster | Launch cluster | Cluster | Database | Backup | Manage Tags | Manage IAM roles

Cluster	Cluster Status	DB Health	Release Status	In Maintenance	Recent Events
glue-tutorial-jds	available	healthy	Up to date	no	3

Endpoint glue-tutorial-jds.chtswcdubv1n.eu-west-1.redshift.amazonaws.com:5439 (authorized)

Cluster Properties

Cluster Name	glue-tutorial-jds
Node Type	dc2.large
Nodes	1
Zone	eu-west-1a
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
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	sat:22:30-sat:23:00
Allow Version Upgrade	Yes

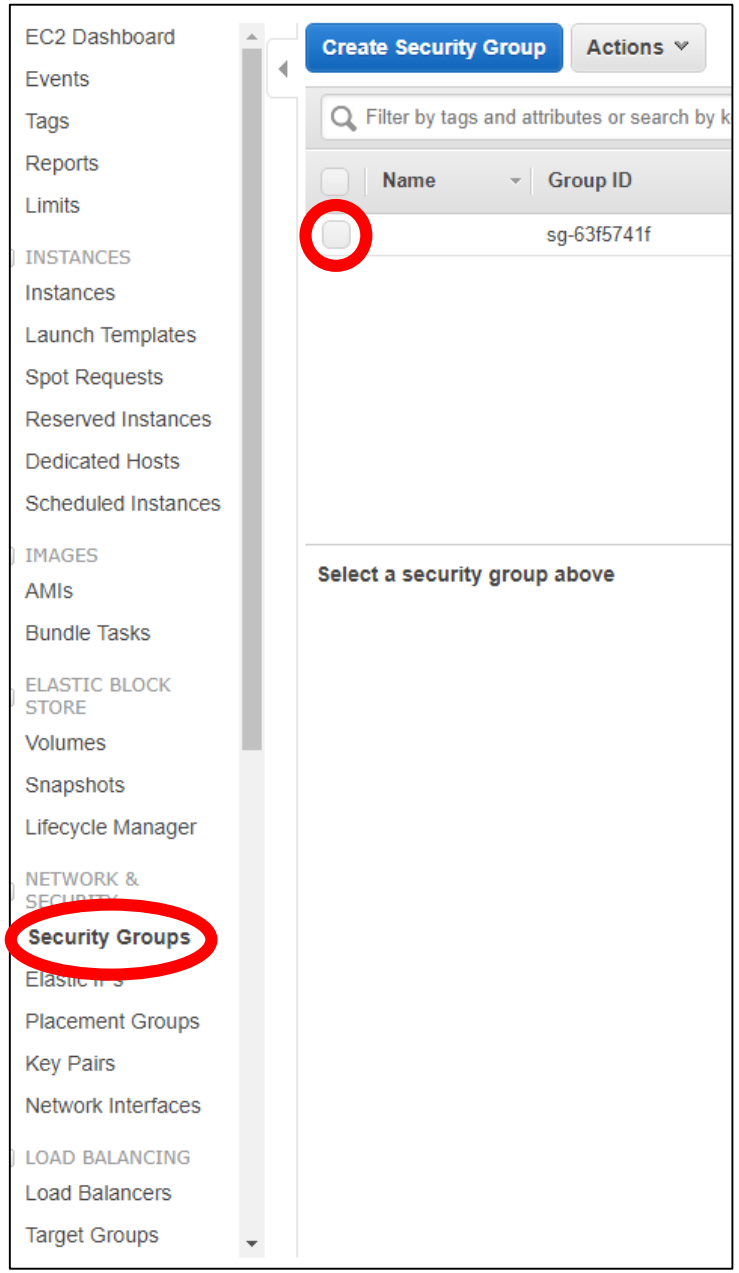
Tags

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

EC2

Edit Security Groups

In a new tab go to the EC2 service



EC2

Edit Security Groups

Security Group: sg-63f5741f

Description | **Inbound** | Outbound | Tags

Edit

Type	Protocol	Port Range	Source	Description
Redshift	TCP	5439	0.0.0.0/0	
Redshift	TCP	5439	::/0	
All traffic	All	All	sg-63f5741f (default)	

EC2

└─**Edit Security Groups**

Choose Redshift Type

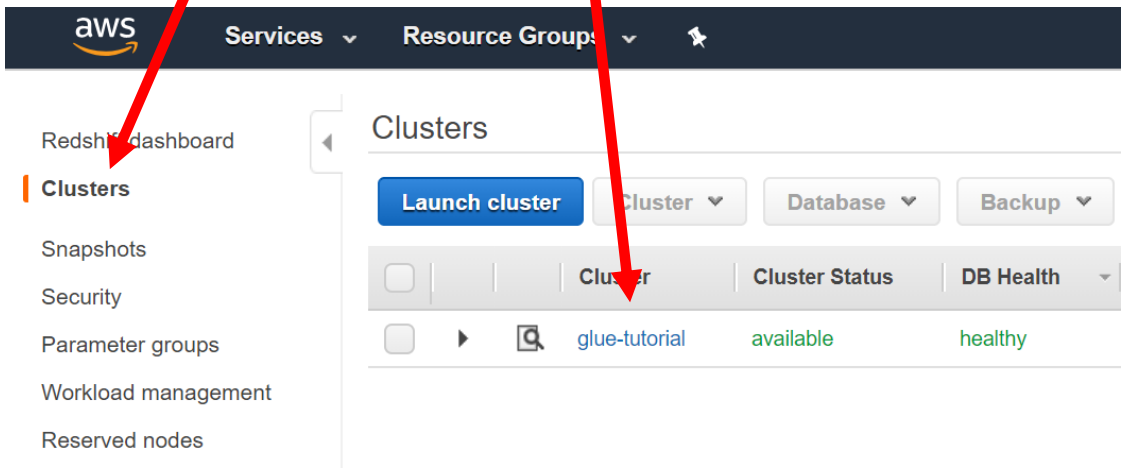
This gives everyone access to Redshift cluster

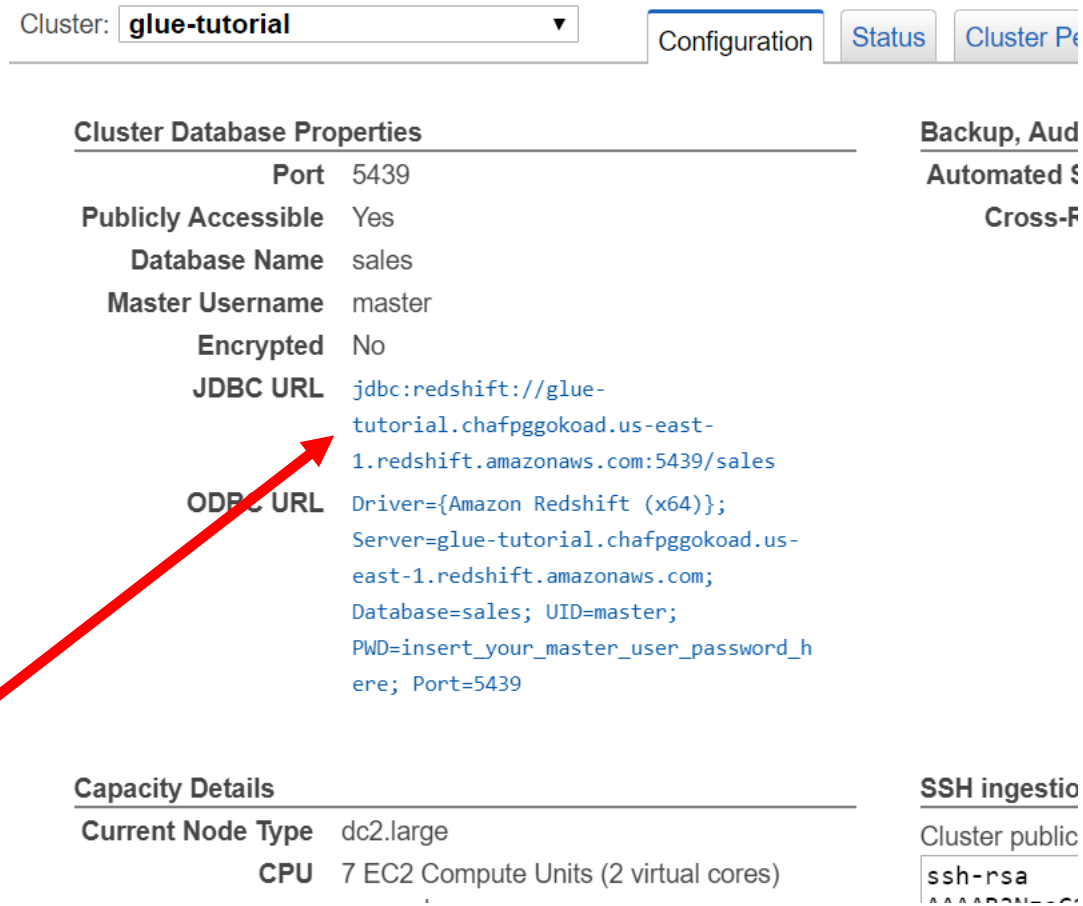
Redshift

Connection

Go to Redshift and select 'Clusters'

Select glue-tutorial





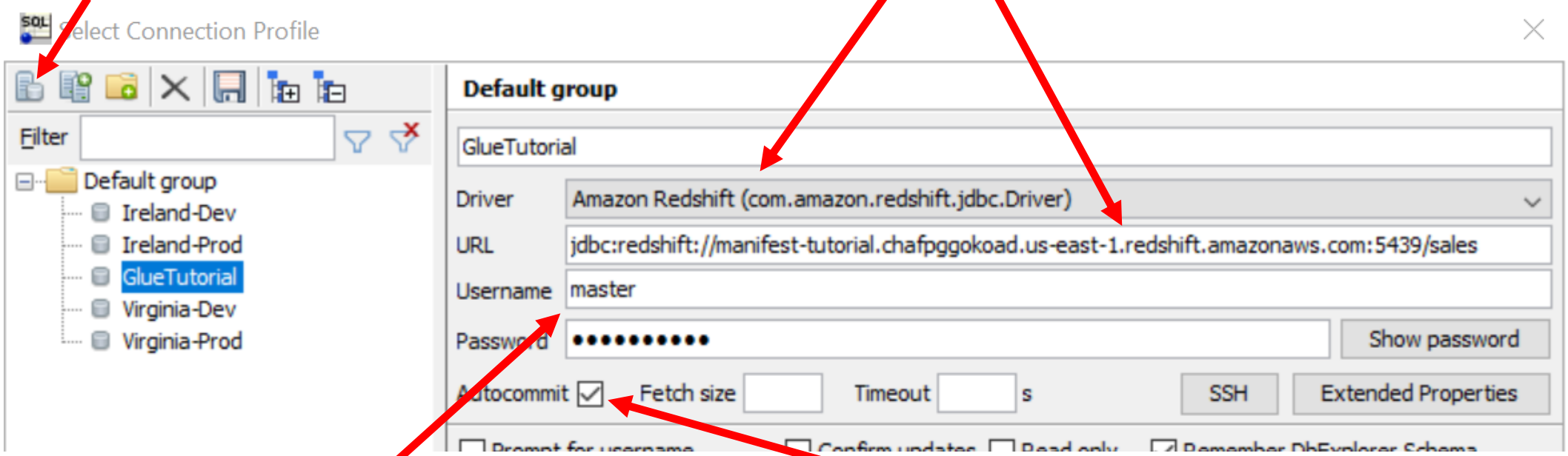
Scroll down to Cluster Database Properties and copy the JDBC URL

Redshift

Connection

Open SQL Workbench and select Create a new connection

Set the Driver to Amazon Redshift and paste the JDBC URL

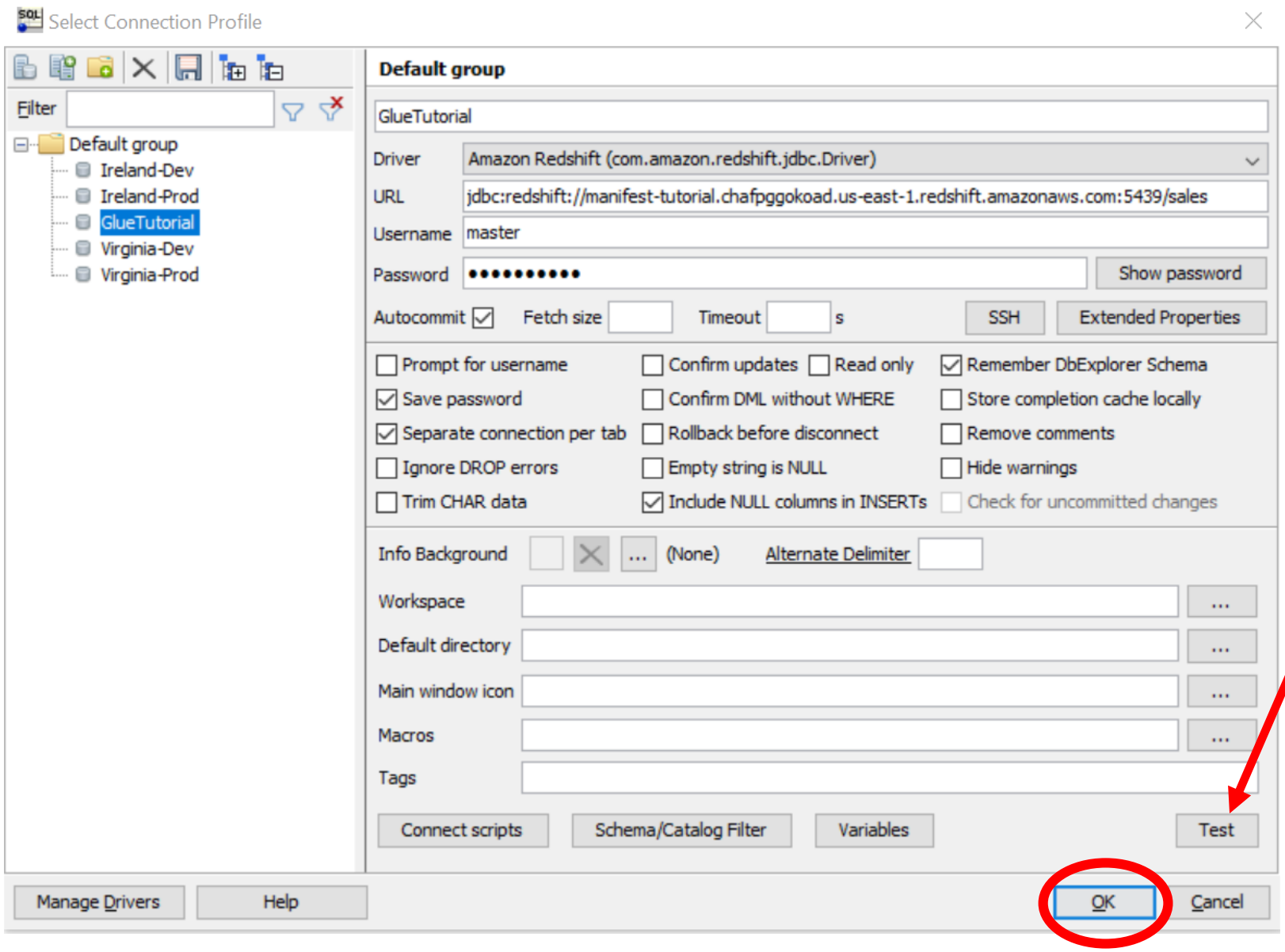


The username and password that was created

Select Autocommit

Redshift

└─Connection



Test your connection

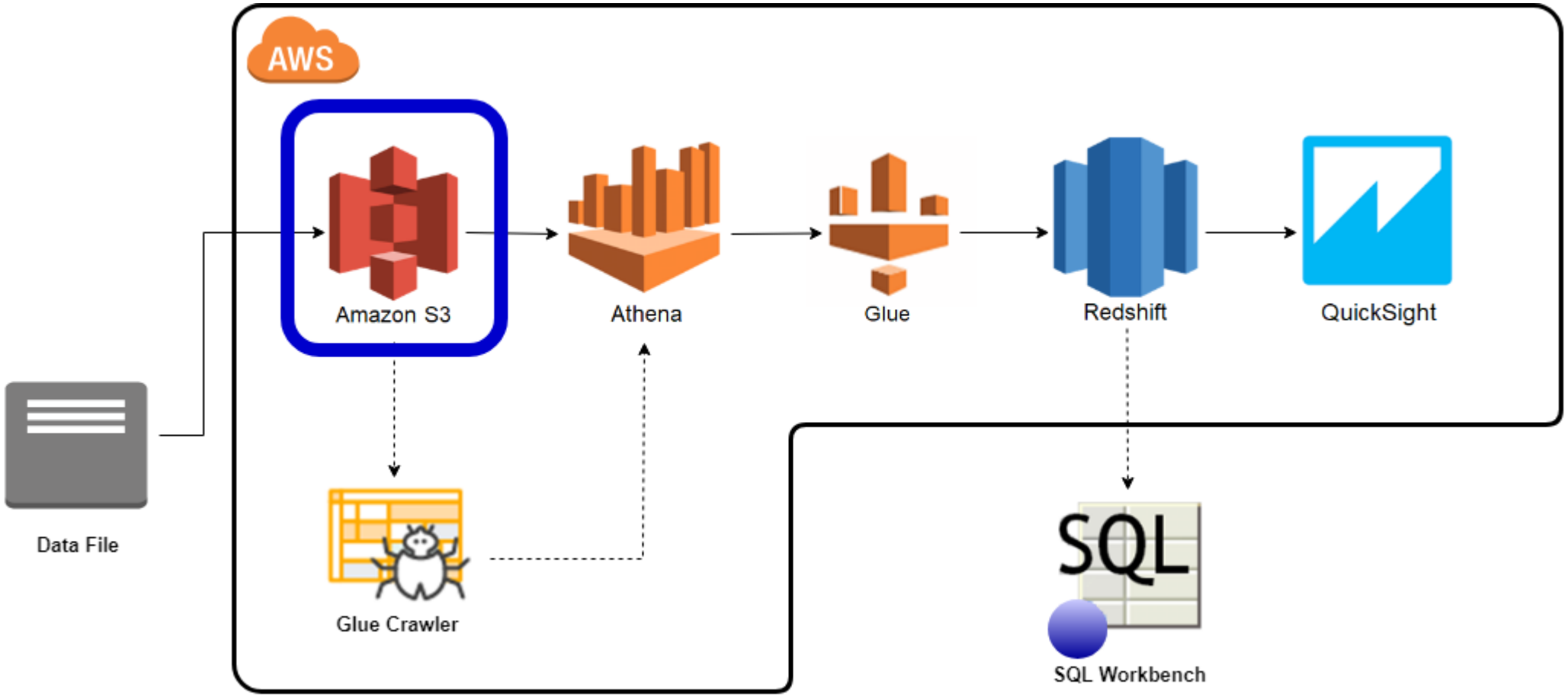
Lab 1

- Prerequisites
- Setup Redshift

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

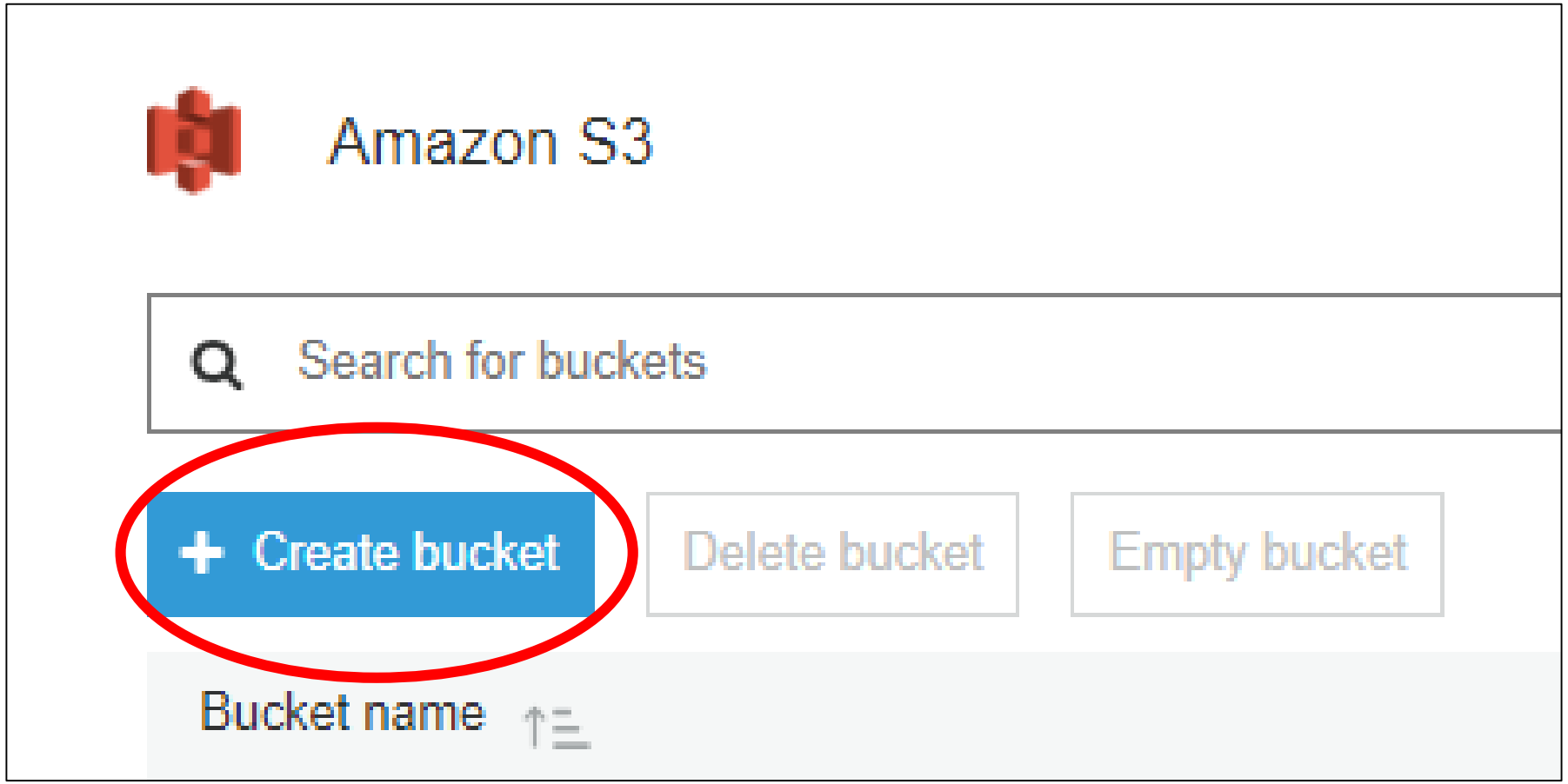
S3

AWS Simple Storage Service



S3

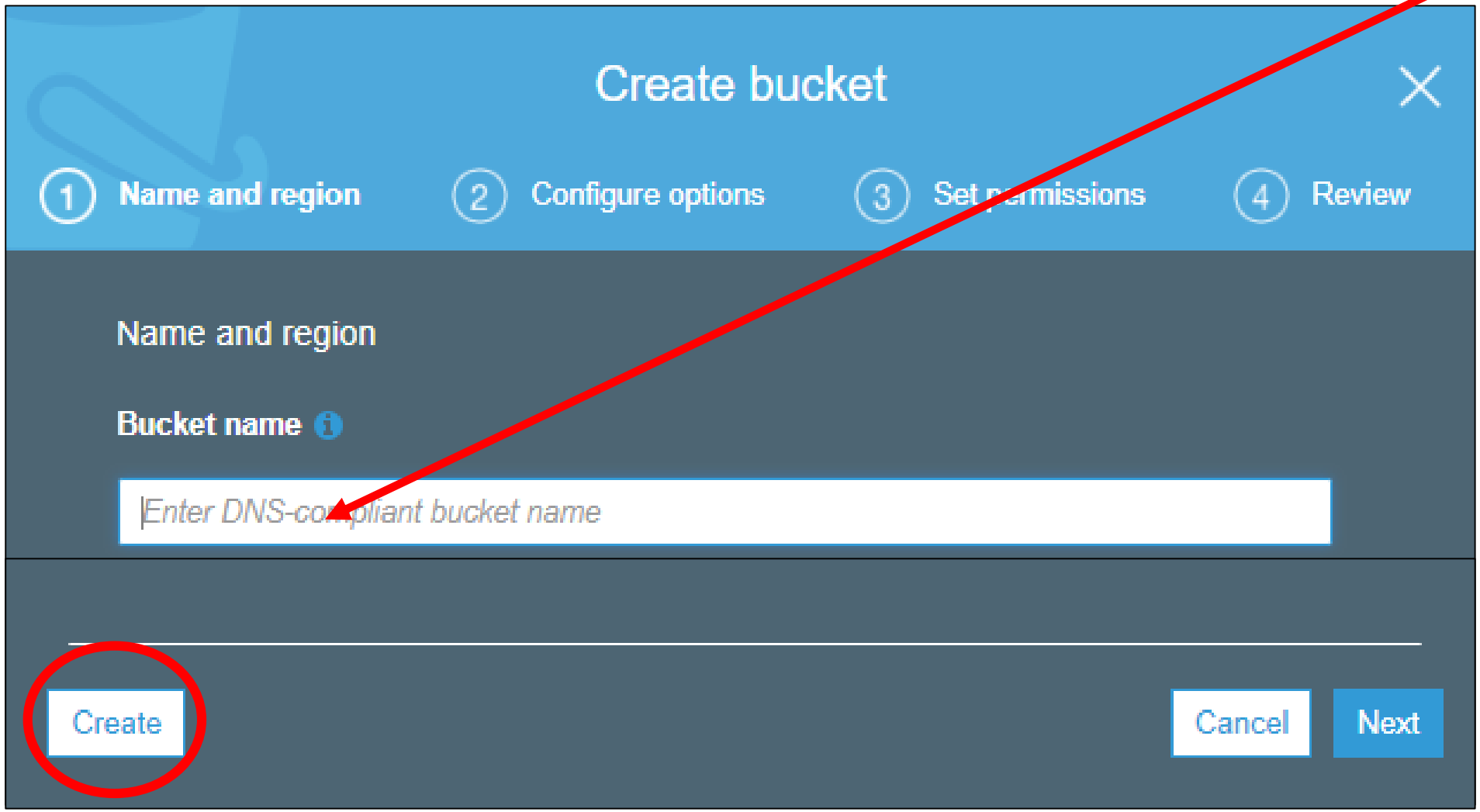
Create S3 bucket with AWS Console



S3

Create S3 bucket with AWS Console

Give your s3 bucket a name
Use glue-tutorial-XXX



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

Specify the region

S3

Create S3 bucket with AWS CLI* (Alternative)

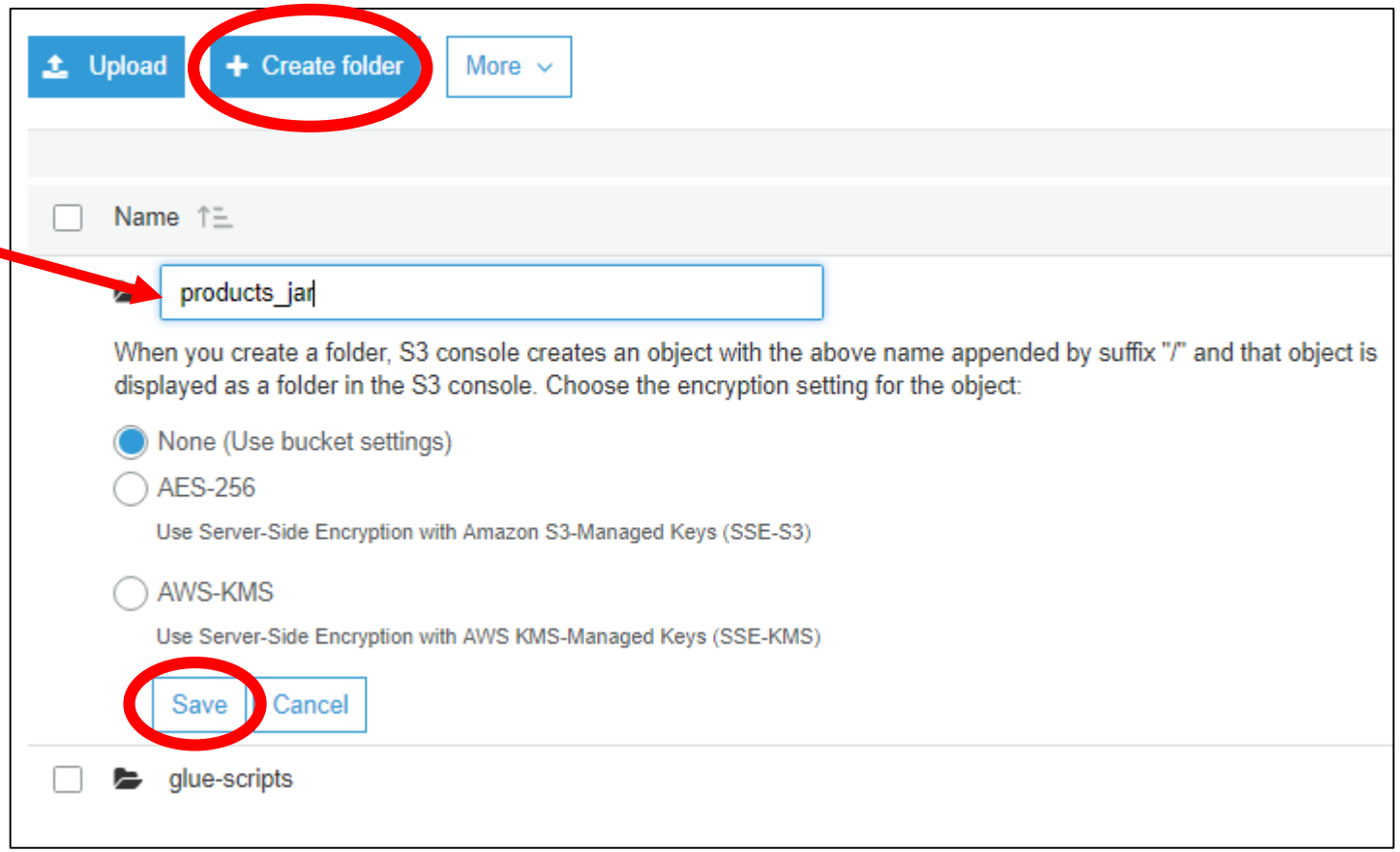
```
$ aws s3api create-bucket --bucket  glue-tutorial-xxx --region
us-east-1
```

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

S3

Create S3 bucket folder

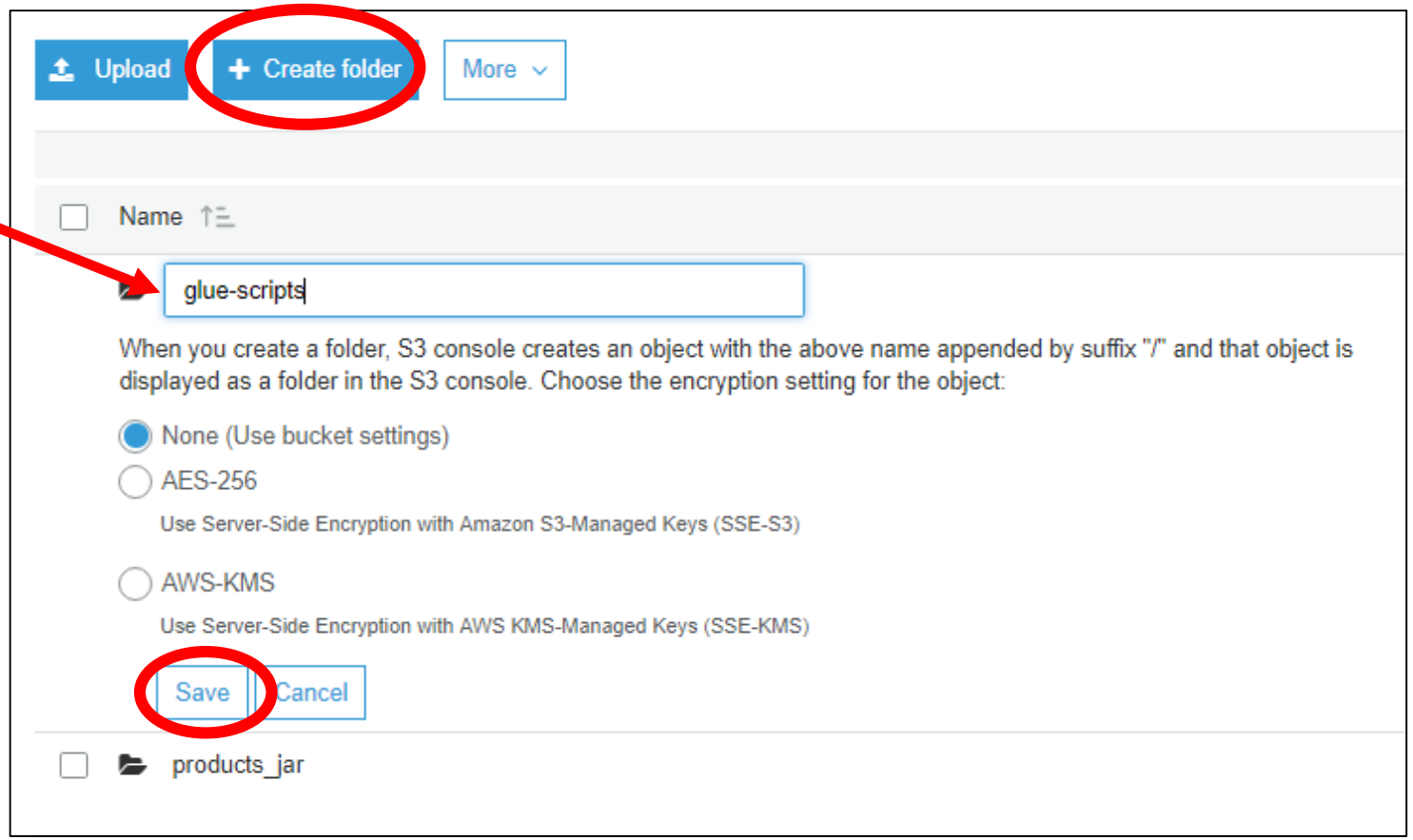
Create a folder called products_XXX



S3

Create S3 bucket folder

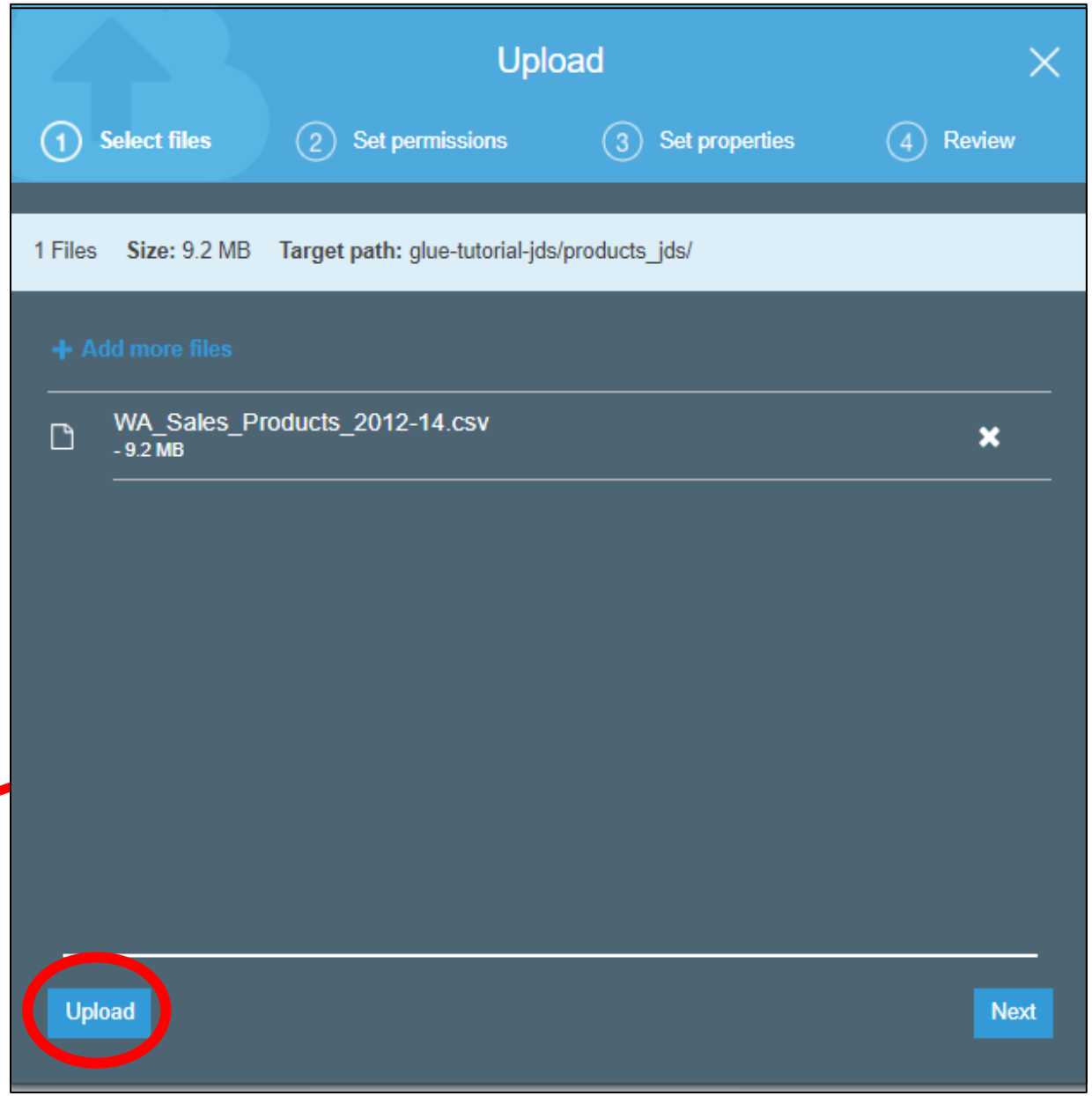
Create a folder called glue-scripts



S3

Add file to S3 bucket with AWS Console

Add file from repository called "WA_Sales_Products_2012-14"



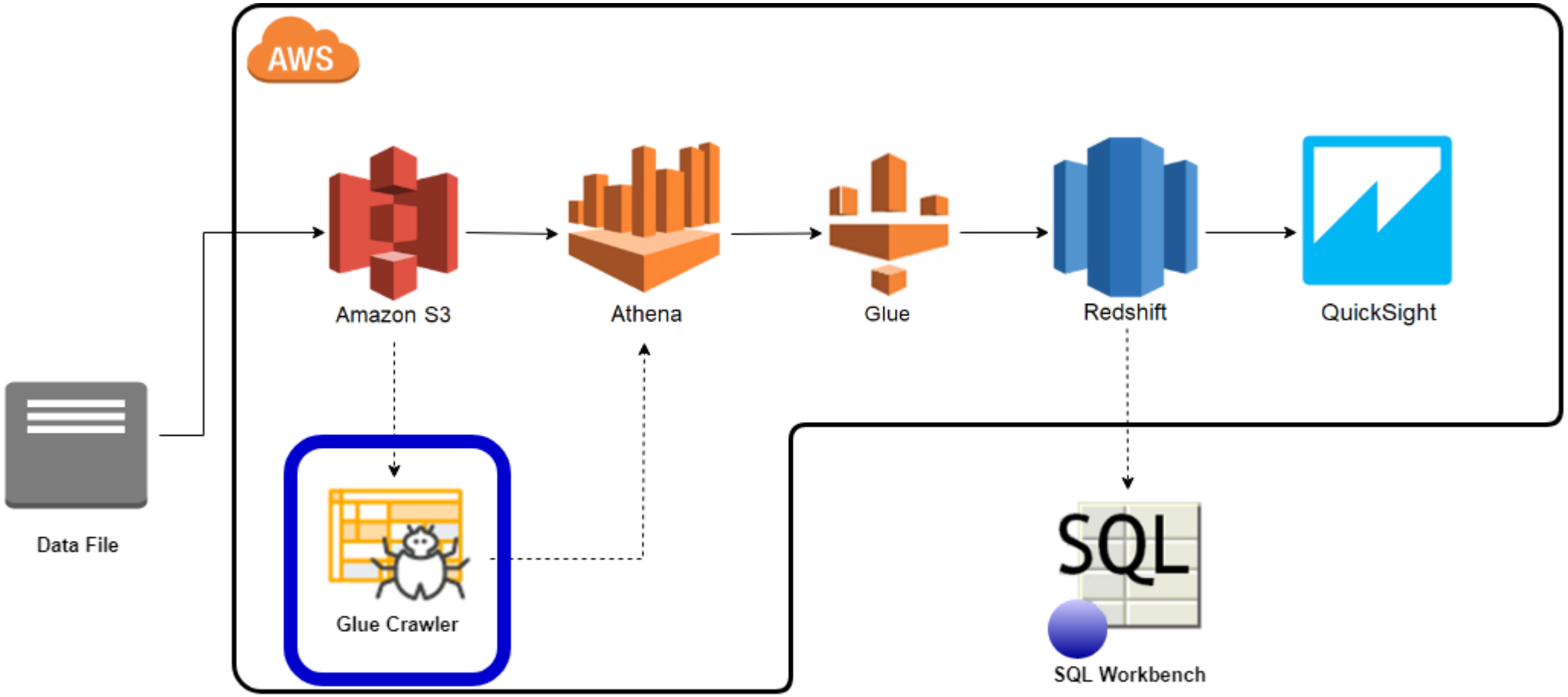
S3

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

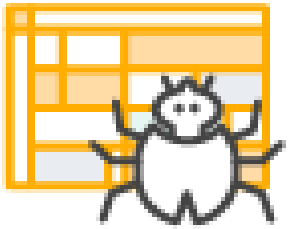
```
$ aws s3 cp <your-file-path>/aws-glue-tutorial/WA_Sales_Products_2012-14.csv s3://glue-tutorial-xxx/products_xxx/WA_Sales_Products_2012-14.csv
```

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

Glue Crawler



Glue Crawler



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

Glue

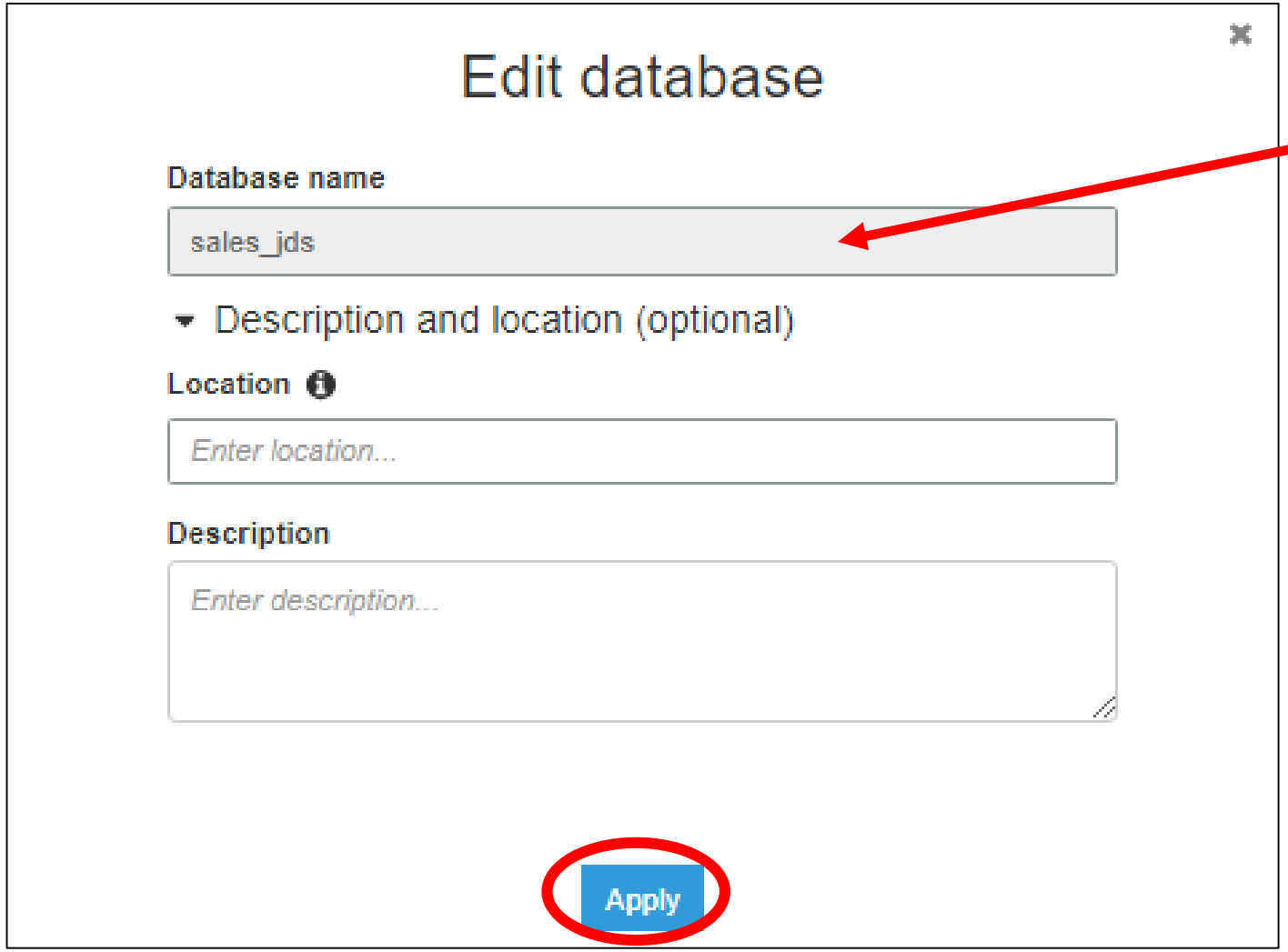
Create Glue Database

Create a new Database

In the Glue Console click on Databases

Glue

Create Glue Database



Give your database a name "sales_XXX"

Glue Crawler

Create Glue Crawler

Click on add tables to create a table

Create a table using a crawler

Glue Crawler

Create Glue Crawler

Add information about your crawler

Crawler name

glue-tutorial-jds

Description and classifiers (optional)

Description

Enter description...

Classifiers infer the schema of your data. AWS Glue tries to match your data with custom classifiers in the order listed. The first classifier to recognize your data is used. Built-in classifiers are used if you do not supply a classifier that matches.

Custom classifiers Showing: 0 - 0

Classifier	Classification
No items available	

Selected classifiers

Classifier	Clas
No clas	

Grouping behavior for S3 data (optional)

☐ Create a single schema for each S3 path

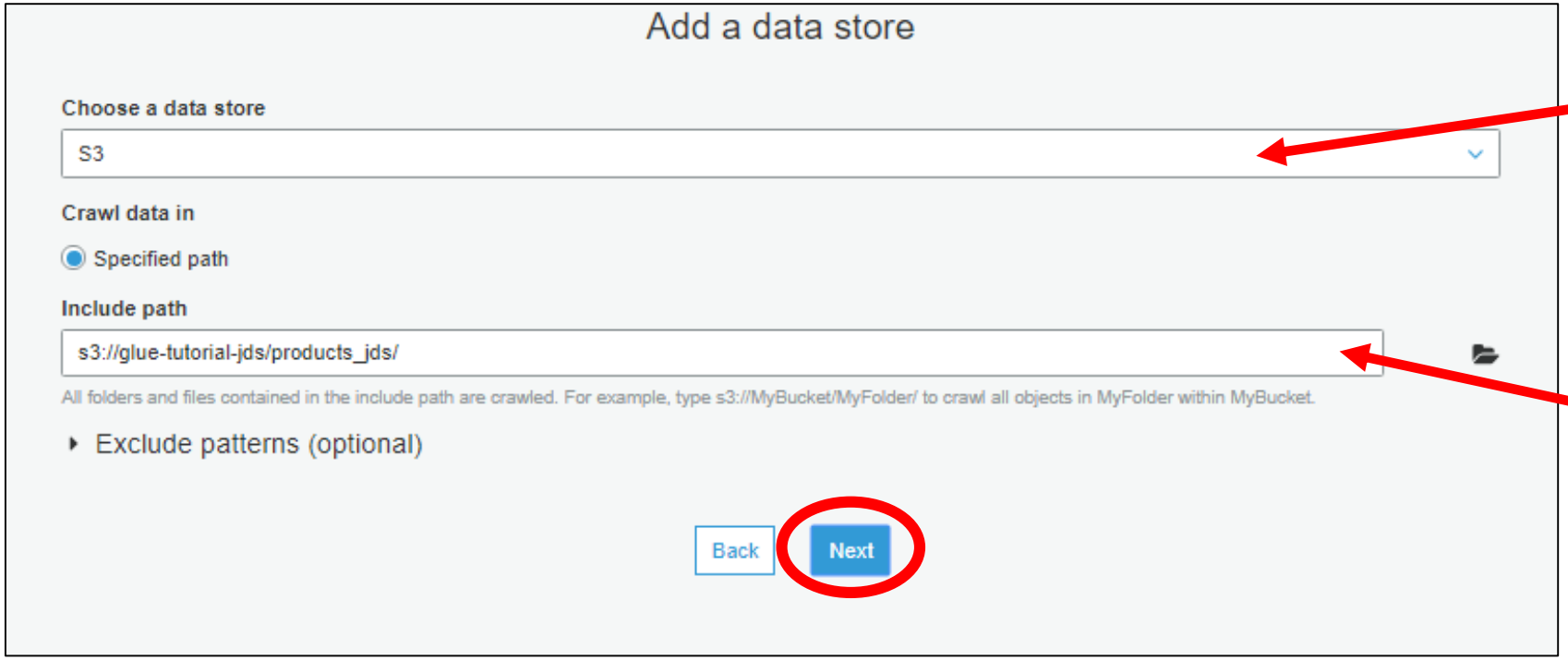
By default, when a crawler defines tables for data stored in S3, it considers both data compatibility and schema similarity. Select this check box to group compatible schemas into a single table definition across all S3 objects under the provided include path. Other criteria will still be considered to determine proper grouping. Learn more

Next

Give your crawler a name, glue-tutorial-XXX

Glue Crawler

Create Glue Crawler



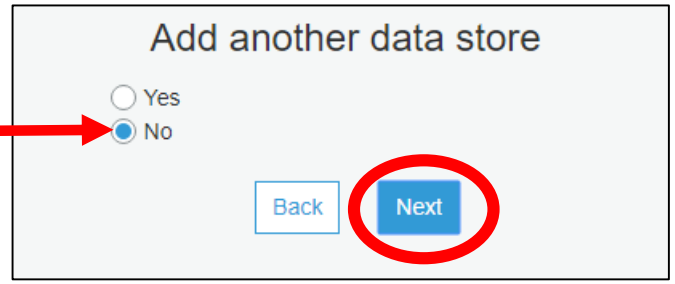
Choose where the table is going to look for data

Specify the path for the table to search for in s3

Glue Crawler

Create Glue Crawler

We do not want to add another source of data



Glue Crawler

└─**Create Glue Crawler**

Need to create role to access S3 bucket

Give your role a name

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-jar/products_jar

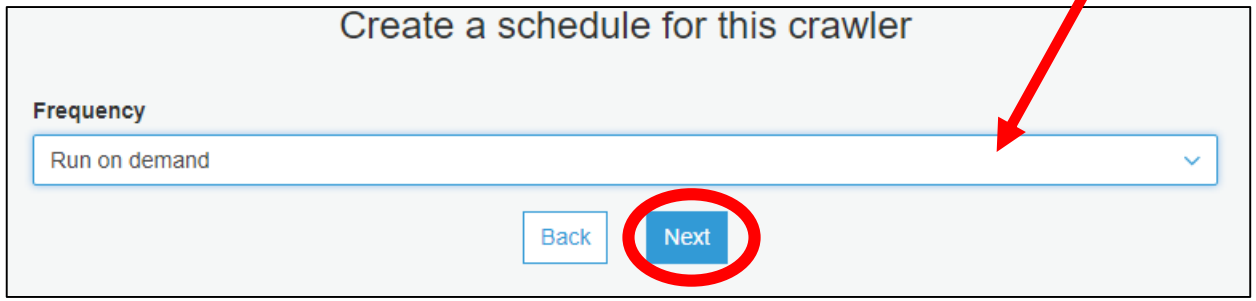
You can also create an IAM role on the IAM console.

Back Next

Glue Crawler

Create Glue Crawler

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



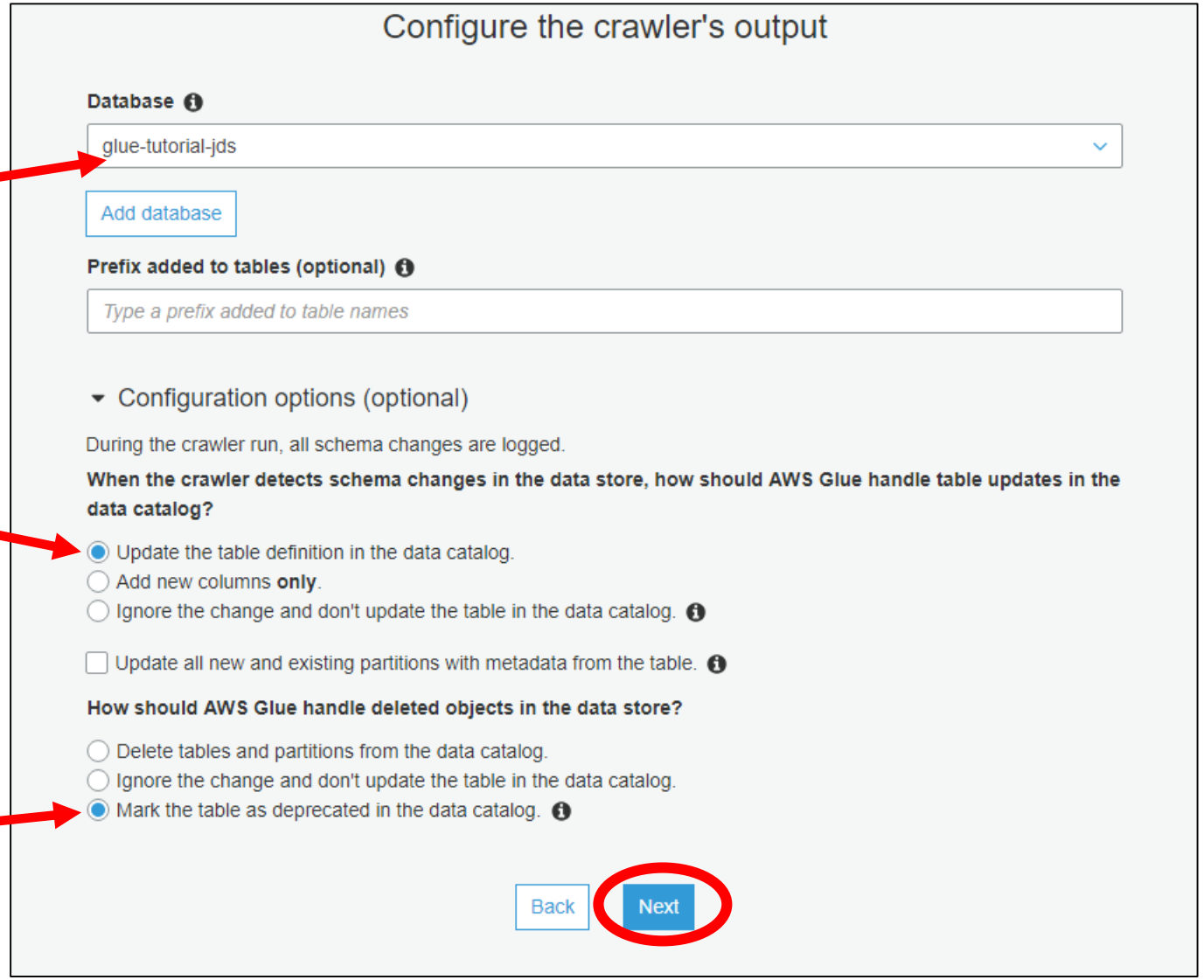
Glue Crawler

└─Create 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



Glue Crawler

Create Glue Crawler

Crawler info

Name	glue-tutorial-jds
Create a single schema for each S3 path	false

Data stores

Data store	S3
Include path	s3://glue-tutorial-jds/products_jds/
Exclude patterns	

IAM role

IAM role	arn:aws:iam::952552944372:role/AWSGlueServiceRole-glueServiceRole

Schedule

Schedule	Run on demand

Output

Database	sales_jds
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.

Back Finish

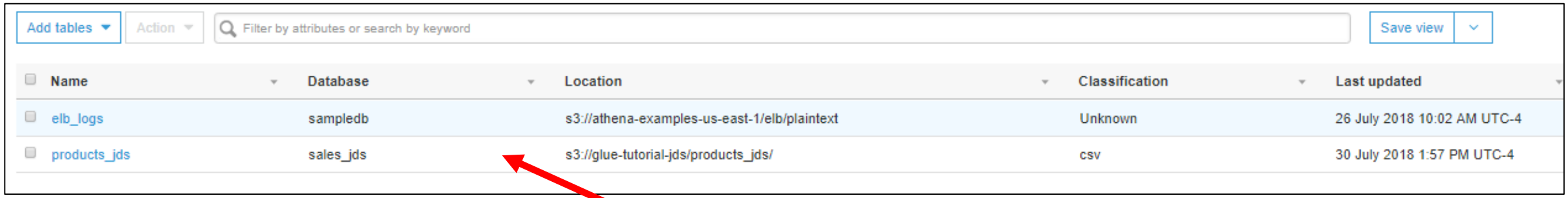
Glue Crawler

└─**Run the Crawler to create Athena table**

Run your crawler

Name	Schedule	Status	Logs	Last runtime	Median runtime	Tables updated	Tables added
glue-tutorial-jds		Ready		0 secs	0 secs	0	0

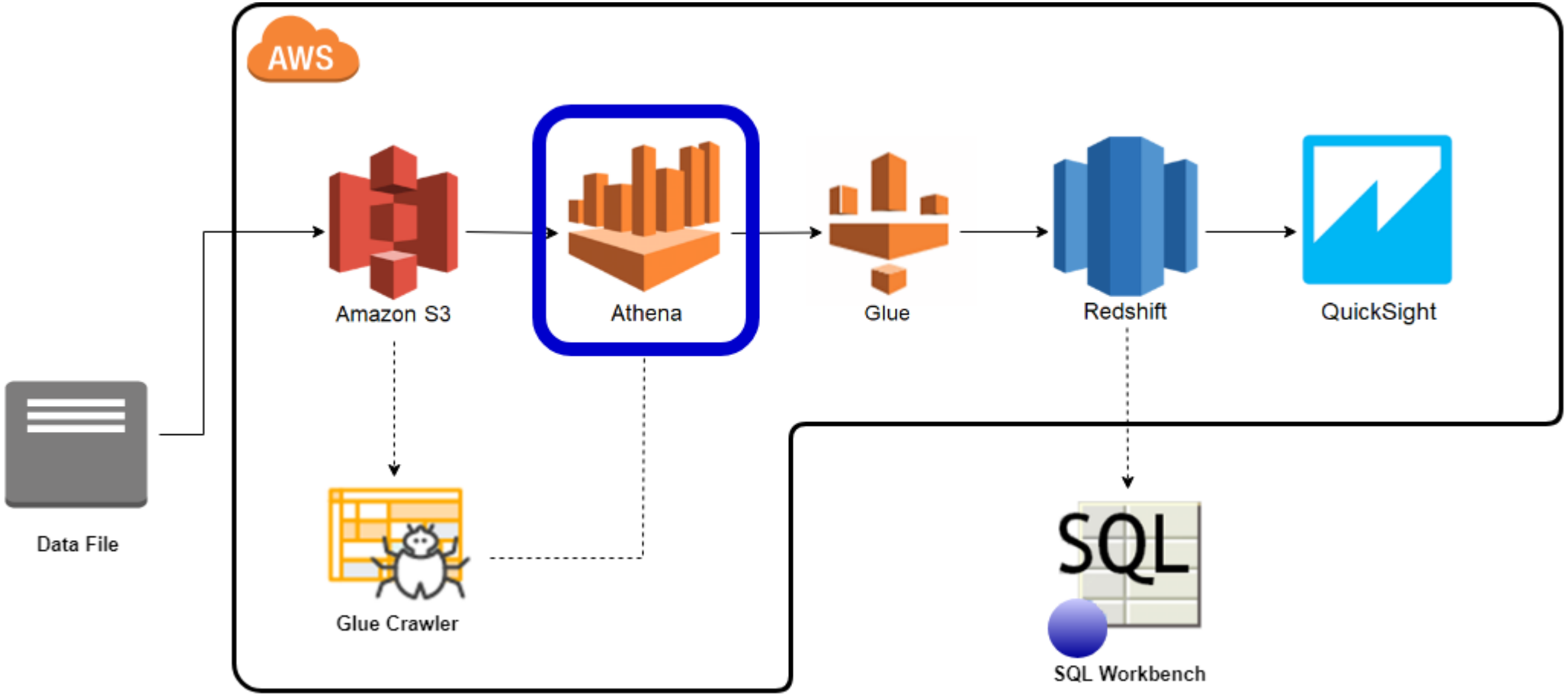
Select your crawler



Name	Database	Location	Classification	Last updated
elb_logs	sampledb	s3://athena-examples-us-east-1/elb/plaintext	Unknown	26 July 2018 10:02 AM UTC-4
products_jds	sales_jds	s3://glue-tutorial-jds/products_jds/	csv	30 July 2018 1:57 PM UTC-4

Your table should be in the table tab

Athena

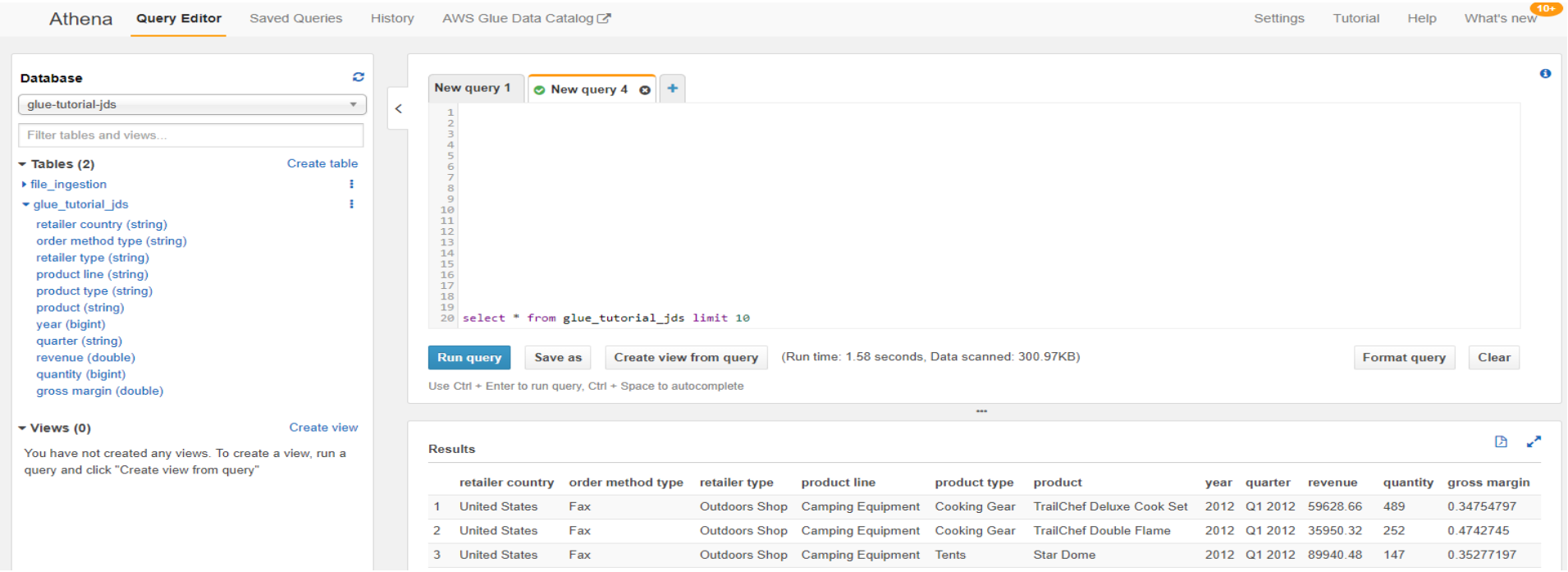

AWS
Amazon S3
Athena
Glue
Redshift
QuickSight
Data File
Glue Crawler
SQL
SQL Workbench

Athena

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

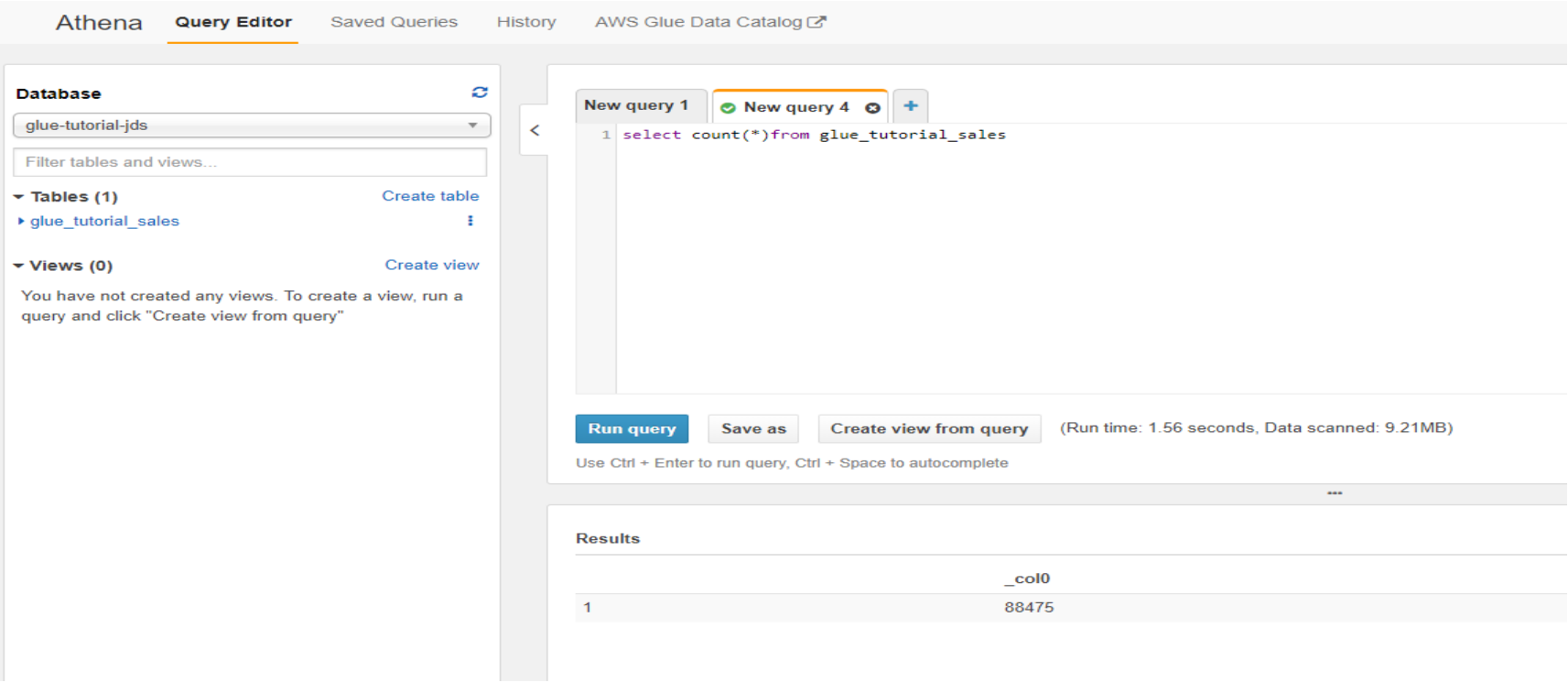
Athena

- 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;

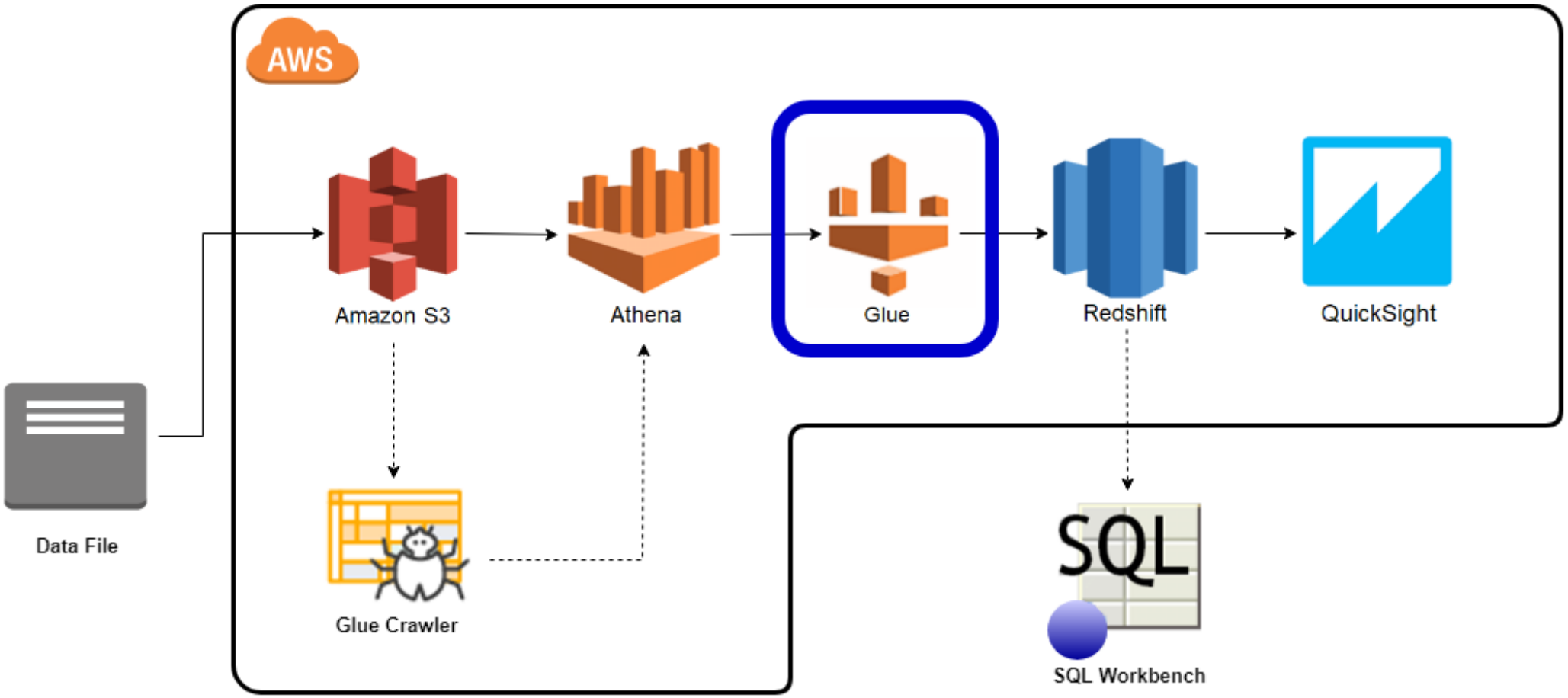


Lab 2

- Create S3 bucket
- Put file in S3
- Create Glue Crawler
- Query Athena

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

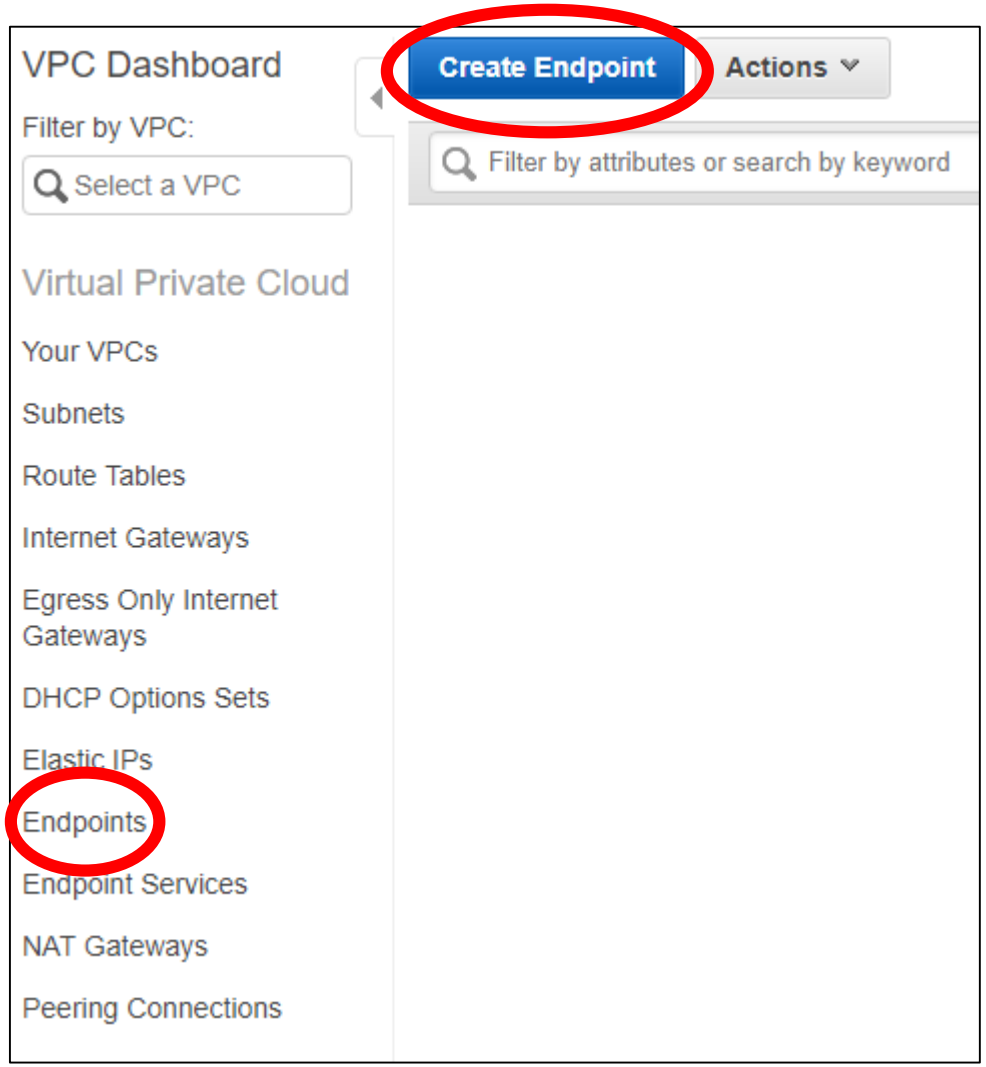
Glue


AWS
Amazon S3
Athena
Glue
Redshift
QuickSight
Data File
Glue Crawler
SQL
SQL Workbench

VPC

Create a S3 endpoint

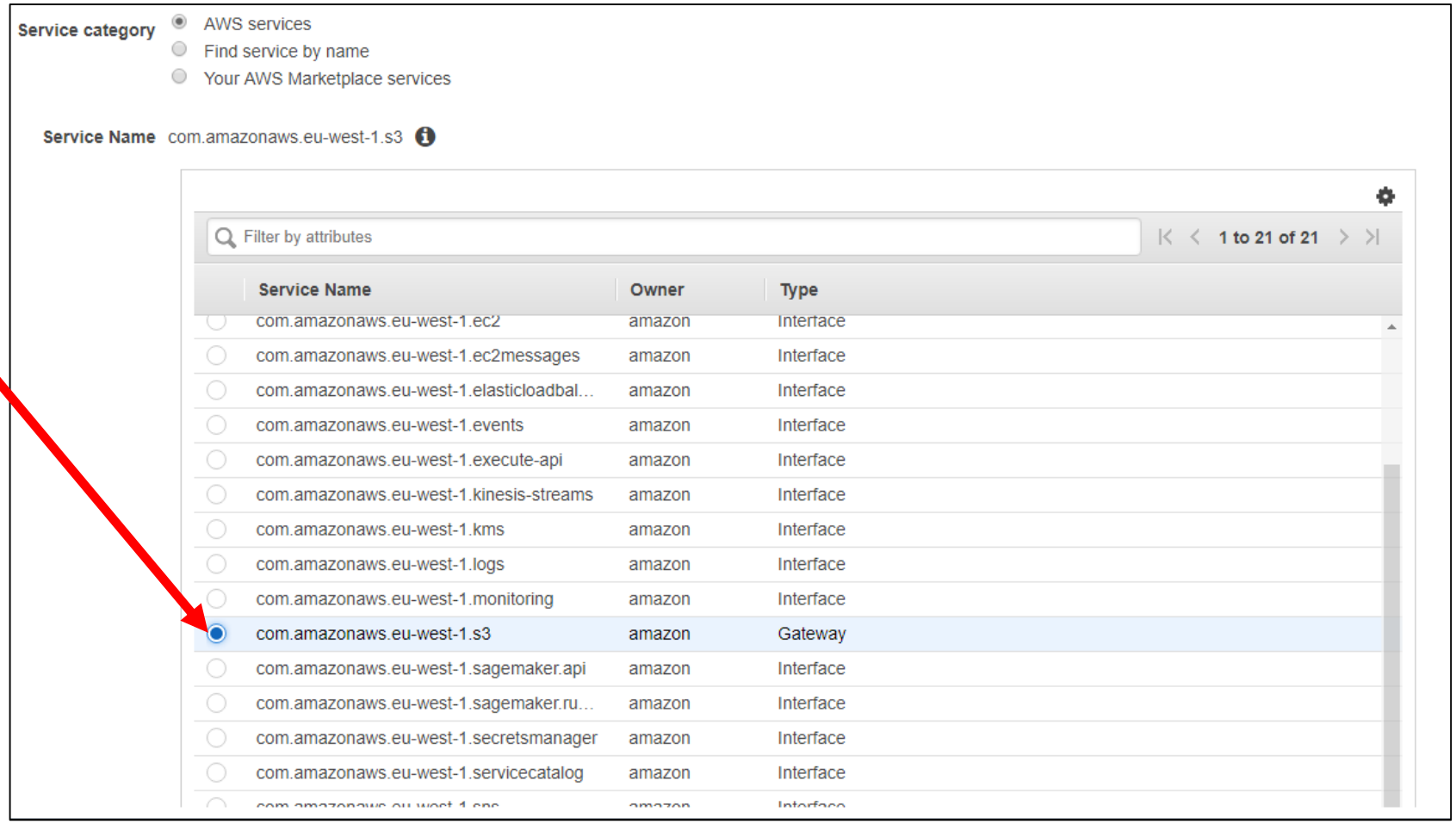
We need to create a S3 endpoint for Glue to access S3



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* vpc-c5500ea3

Configure route tables A rule with destination **pl-6da54004 (com.amazonaws.eu-west-1.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-e5ffb99c

	Route Table ID	Main	Associated With
☑	rtb-e5ffb99c	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

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

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

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.

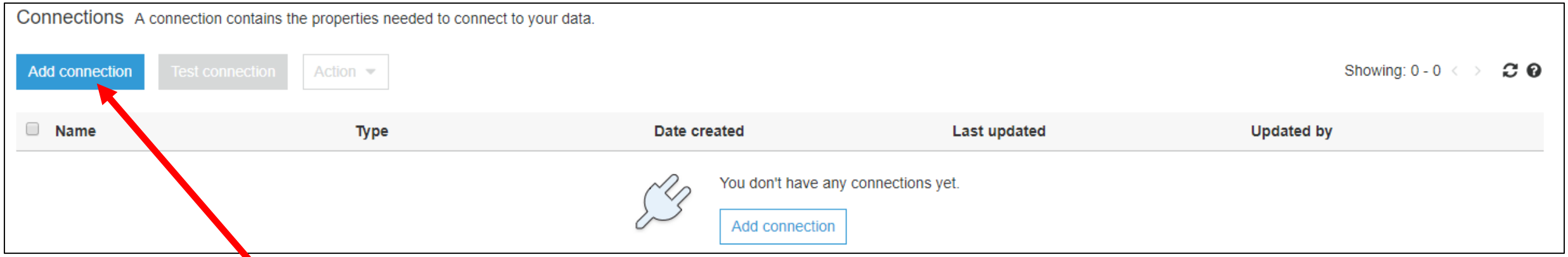
```
{
   "Statement": [
      {
         "Action": "*",
         "Effect": "Allow",
         "Resource": "*",
         "Principal": "*"
      }
   ]
}
```

Cancel **Create endpoint**

Glue

Create a connection to Redshift

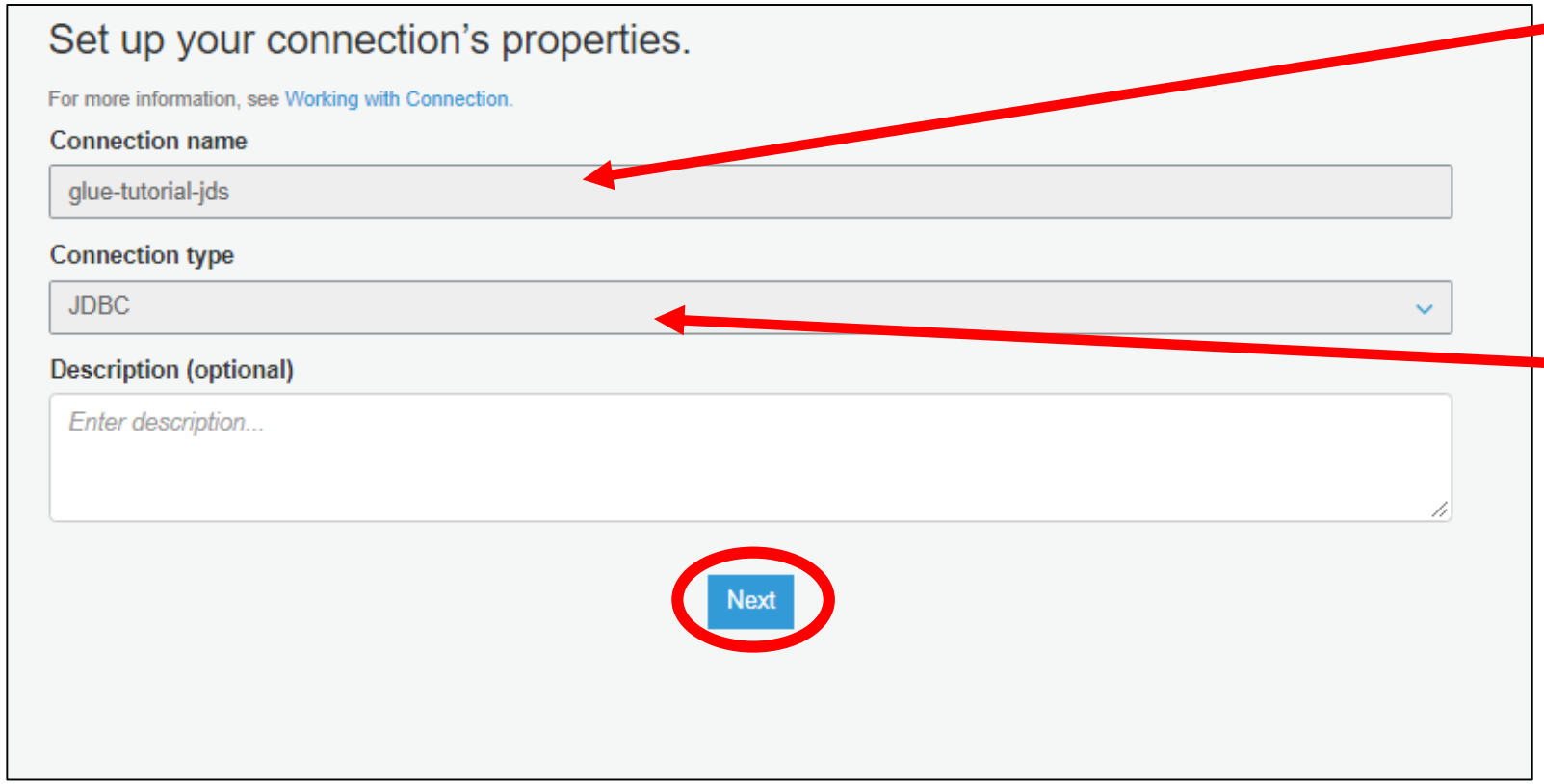
Go to Glue in services and click on connections



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 JDBC

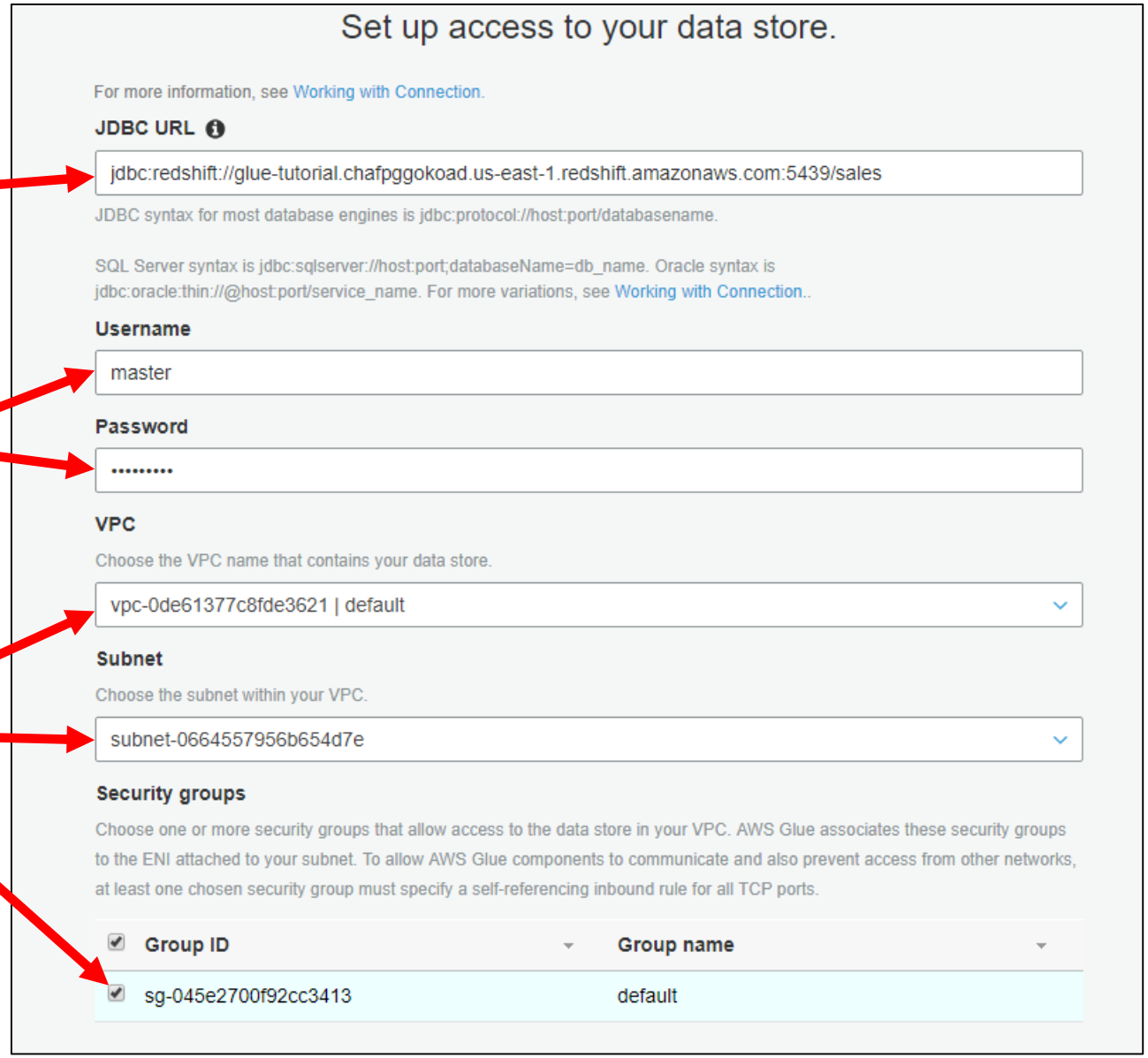
Glue

Create a connection to Redshift

This is the Redshift cluster url

Username and Password for Redshift

This is the VPC/Subnet/Security Group used in your Redshift cluster



Glue

Test the connection to Redshift

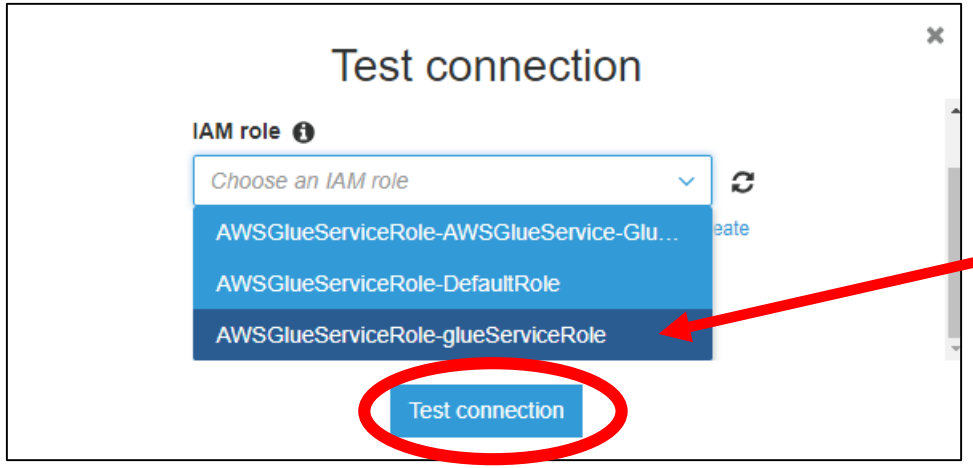
Add connection | Test connection | Action

Showing: 1 - 1

Name	Type	Date created	Last updated	Updated by
glue-tutorial-jds	JDBC	27 July 2018 1:37 PM UTC-4	27 July 2018 1:37 PM UTC-4	

Glue

Test the connection to Redshift



Select this IAM role

Glue

Test the connection to Redshift

glue-tutorial-jds connected successfully to your instance.

Add connection | Test connection | Action

Showing: 1 - 1

Name	Type	Date created	Last updated	Updated by
glue-tutorial-jds	JDBC	27 July 2018 1:37 PM UTC-4	27 July 2018 1:40 PM UTC-4	

Glue

Create a Glue job

Click on jobs in the Glue Console

Glue

Create a Glue job

Give your job a name: glue-tutorial-XXX

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

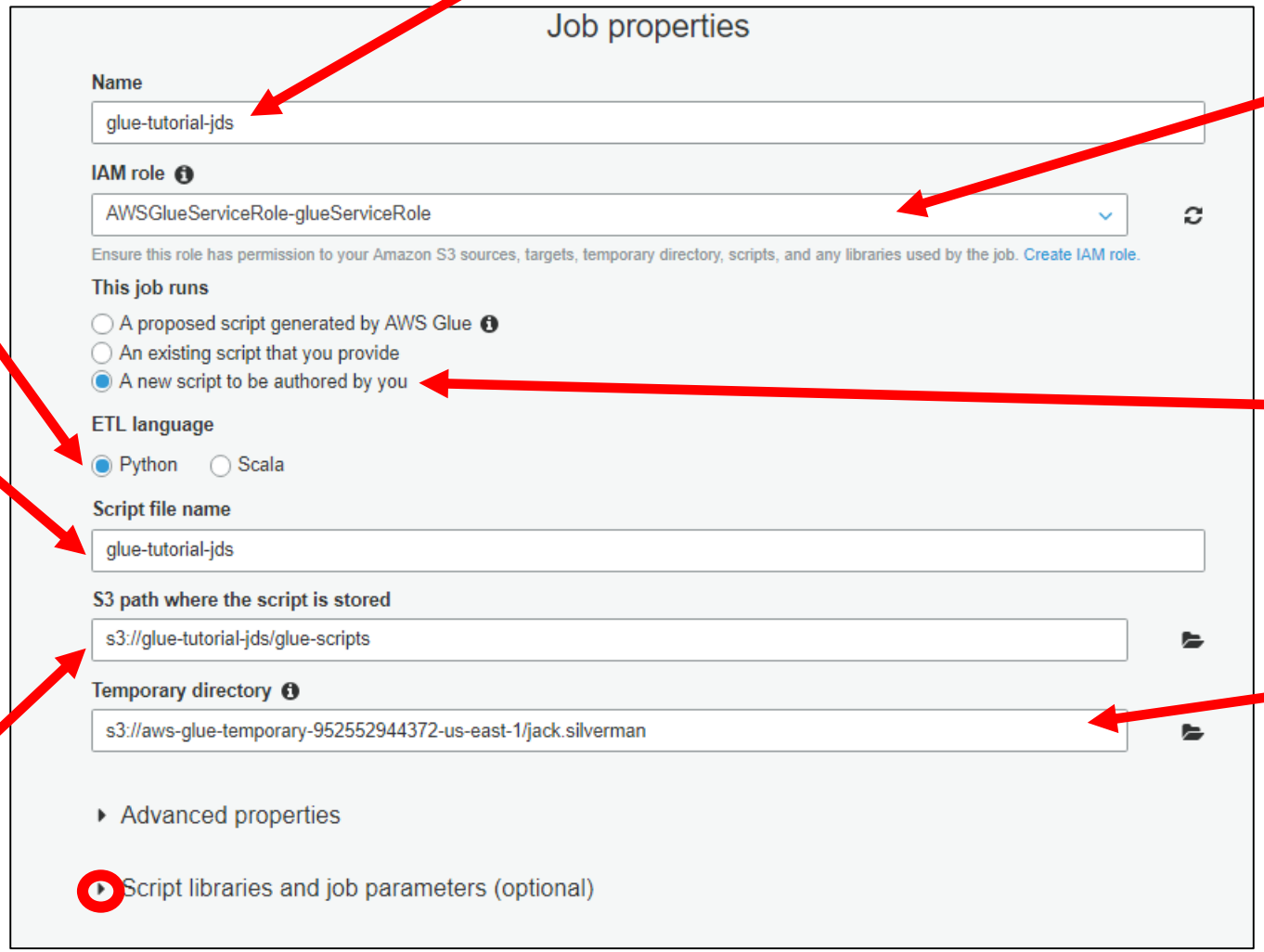
The language used to write the script

Give your script a name glue-tutorial-XXX

Create a new blank script

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

The location where your script will be placed in S3



Glue

Create a connection to the database

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

Job automatically stops after set time

Parameterize values to be used in the script

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

Job parameters

Key	Value
--REDSHIFT_DB_NAME	sales
--SCHEMA_NAME	sales-jds
--TABLE_NAME	products-jds
--CATALOG_CONNECTION_	glue-tutorial
Type key...	Type value...

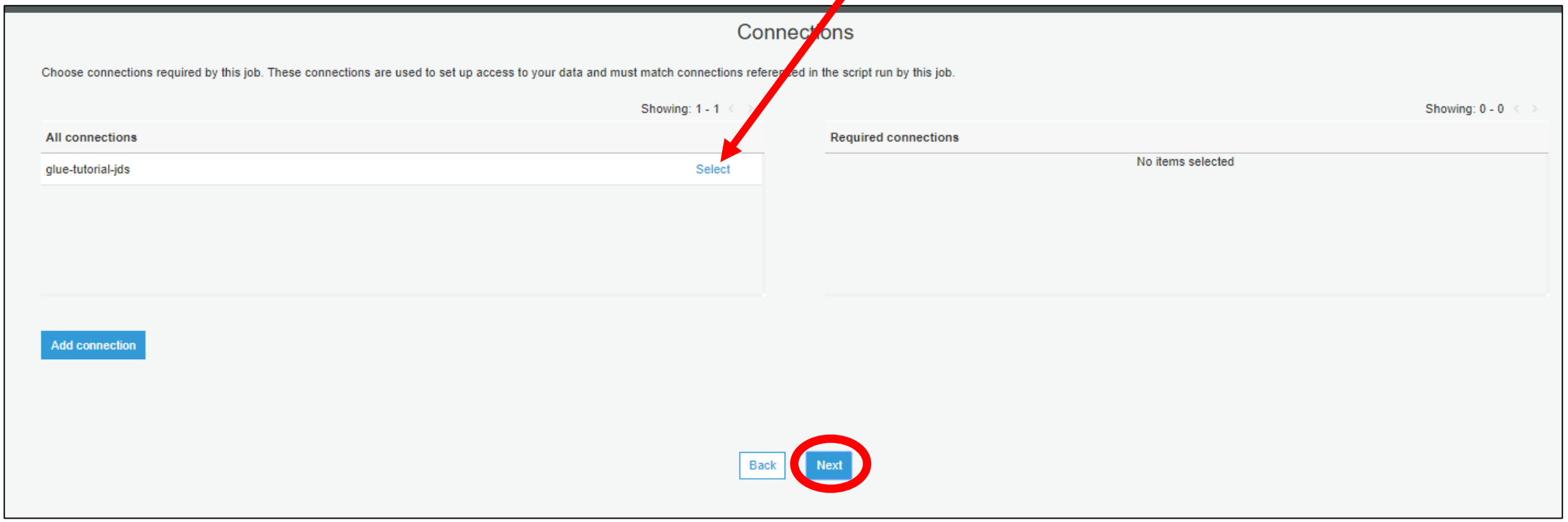
Next

Parameters:
--REDSHIFT_DB_NAME
glue_tutorial_XXX
--SCHEMA_NAME
sales-XXX
--TABLE_NAME
products-XXX
--CONNECTION_NAME
glue-tutorial-XXX

Glue

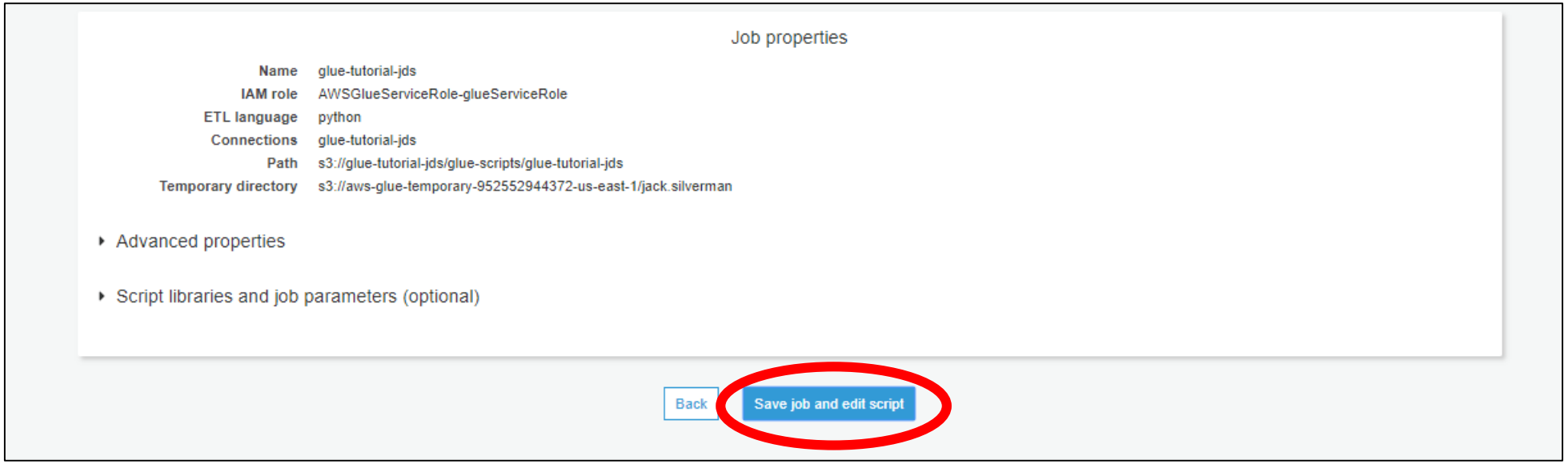
Create a Glue job

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



Glue

Create a Glue job



Glue

Writing the Script

Job: glue-tutorial-jds | Action | Save | Run job | Generate diagram | Insert template at cursor | Source

```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', 'TABLE_NAME', 'SCHEMA_NAME',
'REDSHIFT_DB_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['SCHEMA_NAME'],
    table_name = args['TABLE_NAME'],
    transformation_ctx = ‘datasource’
)
```

The data will be written to the datasource as a DynamicFrame

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

Glue uses frames and will know in what order to do things according to the transformation_ctx

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("timestamp", 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"),
        ("timestamp", "date", "timestamp", "date")
    ],
    transformation_ctx = "applymapping")
```

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

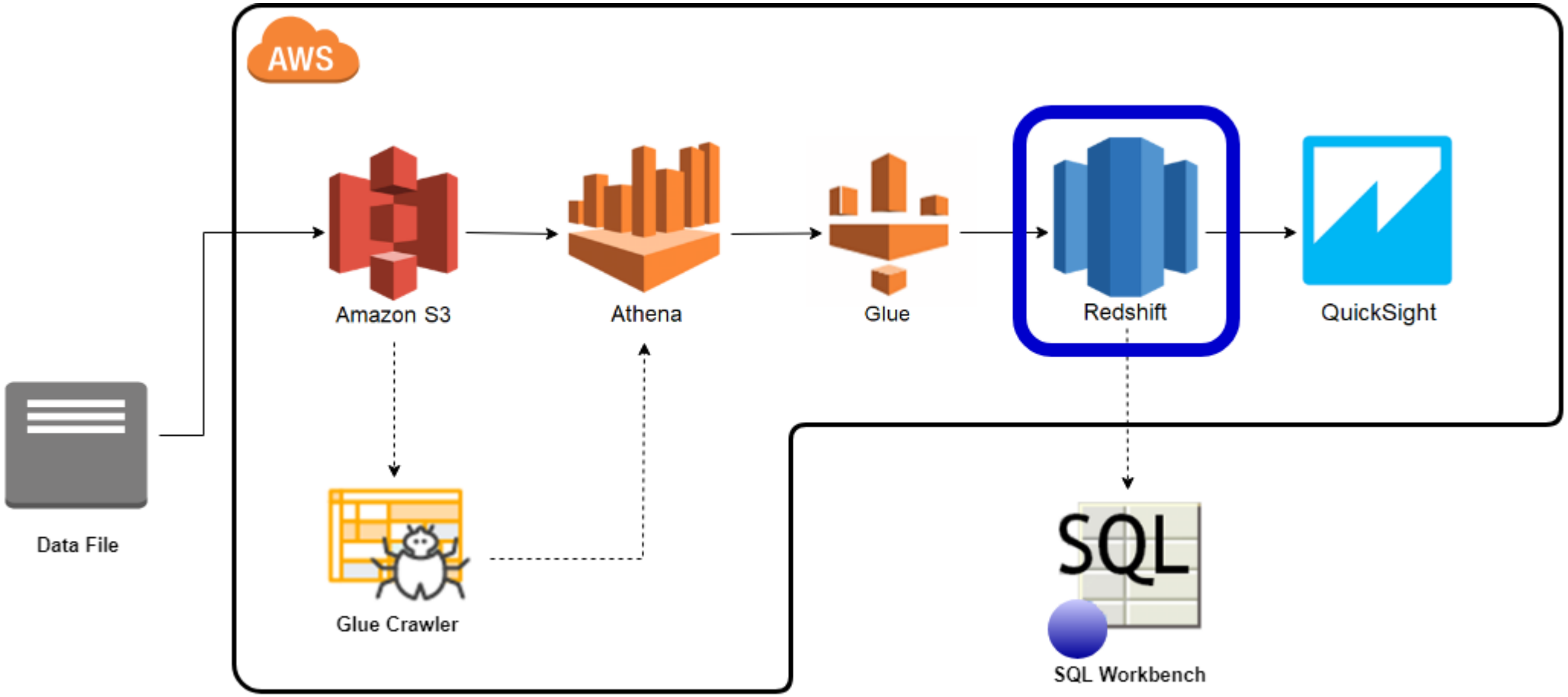
Glue

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['TABLE_NAME']),
        "database": args['REDSHIFT_DB_NAME']
    },
    redshift_tmp_dir = args["TempDir"],
    transformation_ctx = "datasink")
```

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

Redshift


AWS
Amazon S3
Athena
Glue
Redshift
QuickSight
Data File
Glue Crawler
SQL
SQL Workbench

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

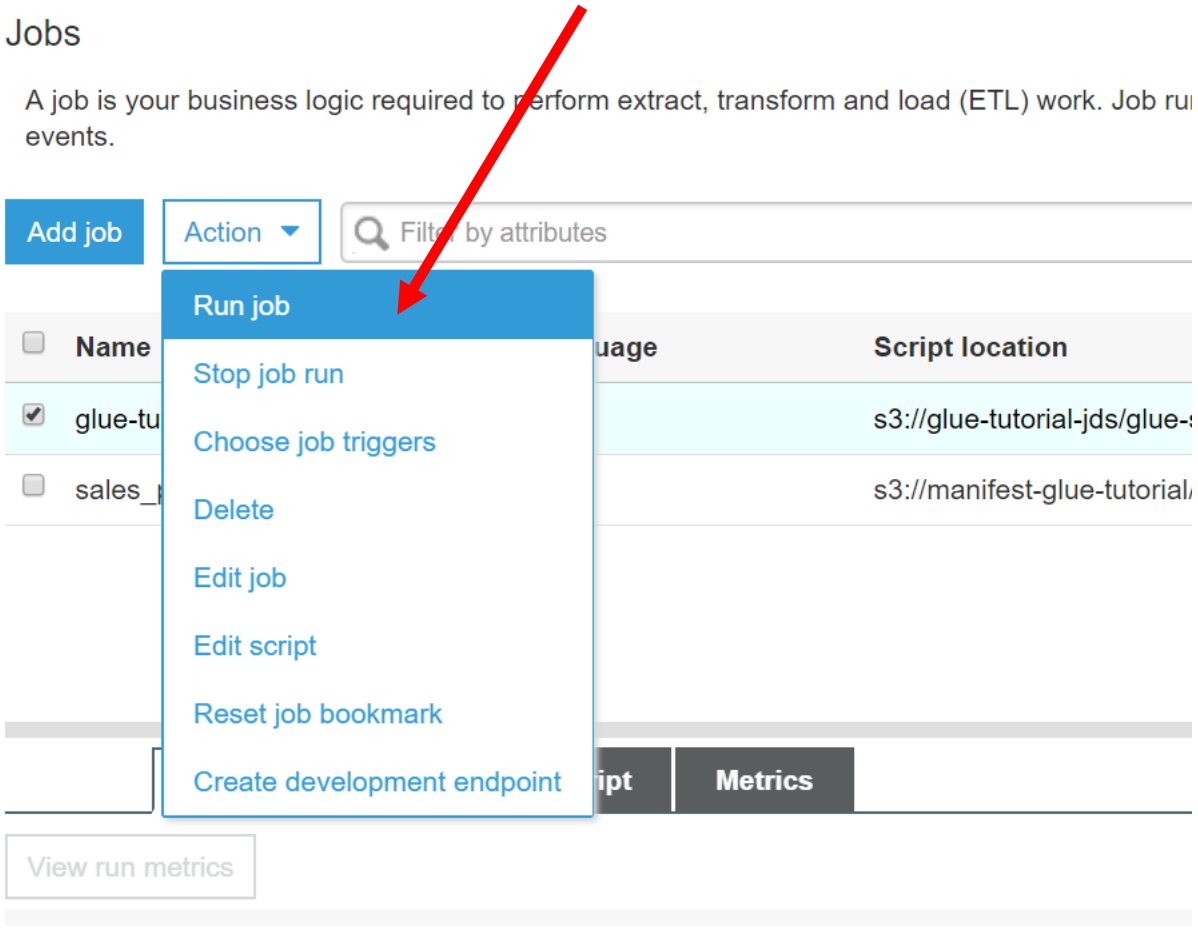
Statement 1 | Database Explorer 2

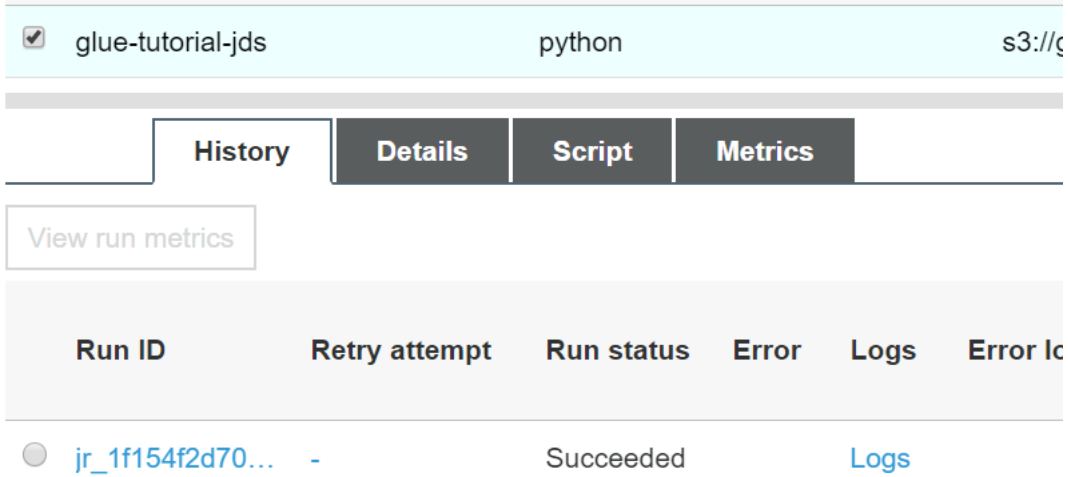
```
SELECT * FROM sales_XXX.products_XXX LIMIT 50;
```

Glue

Run the Glue job

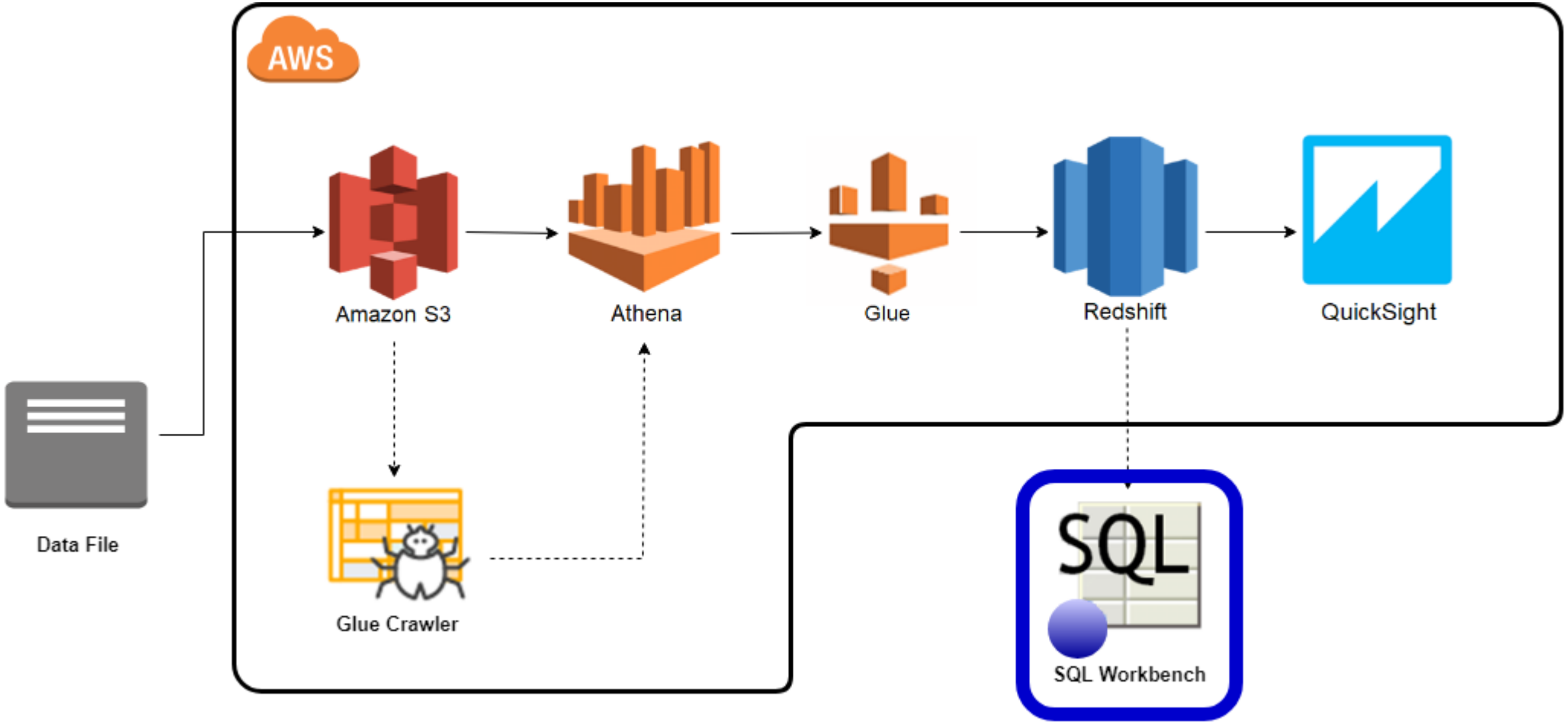
Go back to Glue and run your Glue job





When the job succeeds, check your Redshift table

SQL Workbench



Redshift

Verify data in the table

```
SELECT * FROM sales_jds.products_jds LIMIT 50;
```

Result 1 | Messages

retailer_country	order_method_type	retailer_type	product_line	product_type	product	year	quarter	revenue	quantity
United States	Fax	Outdoors Shop	Camping Equipment	Cooking Gear	TrailChef Deluxe Cook Set	2012	Q1	59628.66	489
United States	Fax	Outdoors Shop	Camping Equipment	Tents	Star Dome	2012	Q1	89940.48	147
United States	Fax	Outdoors Shop	Camping Equipment	Sleeping Bags	Hibernator Lite	2012	Q1	119822.20	1415
United States	Fax	Outdoors Shop	Camping Equipment	Sleeping Bags	Hibernator Camp Cot	2012	Q1	41837.46	426
United States	Fax	Outdoors Shop	Camping Equipment	Lanterns	Firefly Extreme	2012	Q1	9393.30	189
United States	Fax	Outdoors Shop	Camping Equipment	Lanterns	EverGlow Butane	2012	Q1	6940.03	109
United States	Fax	Outdoors Shop	Mountaineering Equipment	Rope	Husky Rope 60	2012	Q1	14109.40	79
United States	Fax	Outdoors Shop	Mountaineering Equipment	Rope	Husky Rope 200	2012	Q1	77288.64	143
United States	Fax	Outdoors Shop	Mountaineering Equipment	Safety	Husky Harness	2012	Q1	34154.90	559
United States	Fax	Outdoors Shop	Mountaineering Equipment	Safety	Granite Signal Mirror	2012	Q1	4074.84	126
United States	Fax	Outdoors Shop	Mountaineering Equipment	Climbing Accessories	Granite Belay	2012	Q1	19476.80	296
United States	Fax	Outdoors Shop	Mountaineering Equipment	Climbing Accessories	Firefly Climbing Lamp	2012	Q1	17998.56	464
United States	Fax	Outdoors Shop	Mountaineering Equipment	Climbing Accessories	Firefly Rechargeable Battery	2012	Q1	11673.60	1520
United States	Fax	Outdoors Shop	Mountaineering Equipment	Tools	Granite Ice	2012	Q1	25041.60	333
United States	Fax	Outdoors Shop	Mountaineering Equipment	Tools	Granite Shovel	2012	Q1	9543.16	164
United States	Fax	Outdoors Shop	Mountaineering Equipment	Tools	Granite Axe	2012	Q1	32870.40	856
United States	Fax	Outdoors Shop	Personal Accessories	Watches	Mountain Man Extreme	2012	Q1	6499.80	23
United States	Fax	Outdoors Shop	Personal Accessories	Eyewear	Polar Ice	2012	Q1	3825.80	37
United States	Fax	Outdoors Shop	Personal Accessories	Knives	Bear Survival Edge	2012	Q1	8414.75	97
United States	Fax	Outdoors Shop	Outdoor Protection	Insect Repellents	BugShield Extreme	2012	Q1	25010.58	3801
United States	Fax	Outdoors Shop	Outdoor Protection	First Aid	Compact Relief Kit	2012	Q1	4057.20	180
United States	Telephone	Golf Shop	Personal Accessories	Watches	Infinity	2012	Q1	11000.00	50

Lab 3

- Glue Connection
- Glue Database
- Glue Job
- Redshift Schema and Table
- Run Glue Job
- Query Redshift

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

Enhancements

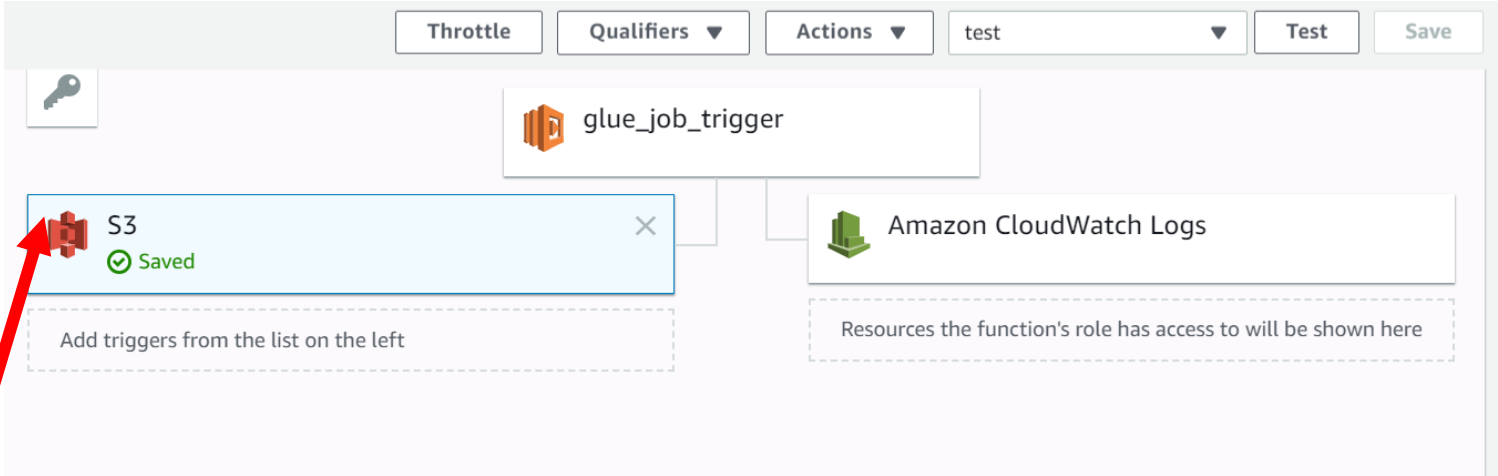
└─**Improve the versatility of your glue job**

- 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



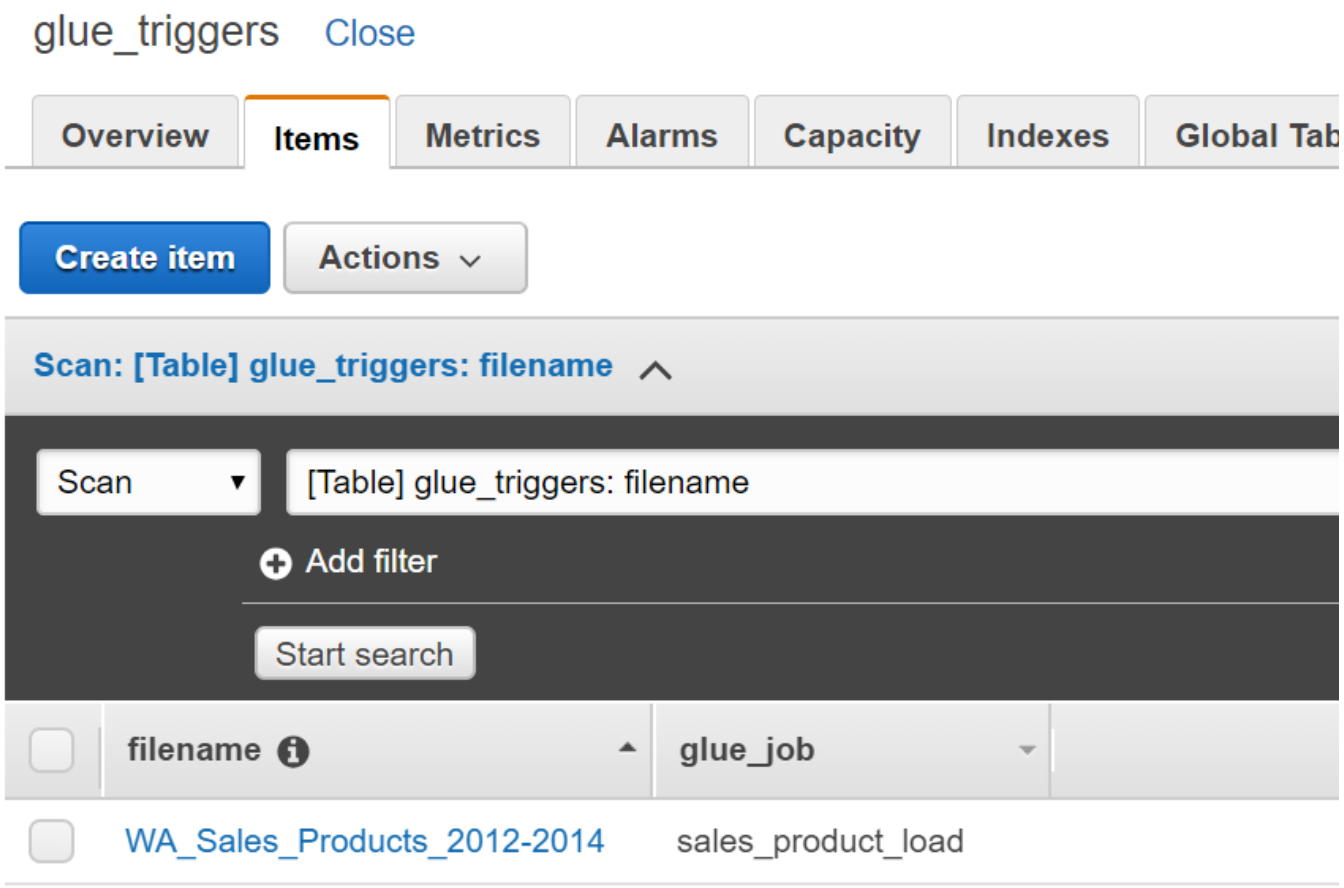
lambda_function

```
1  import boto3
2
3  def lambda_handler(event, context):
4
5      glue = boto3.client('glue')
6      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



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

```python
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)
```

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

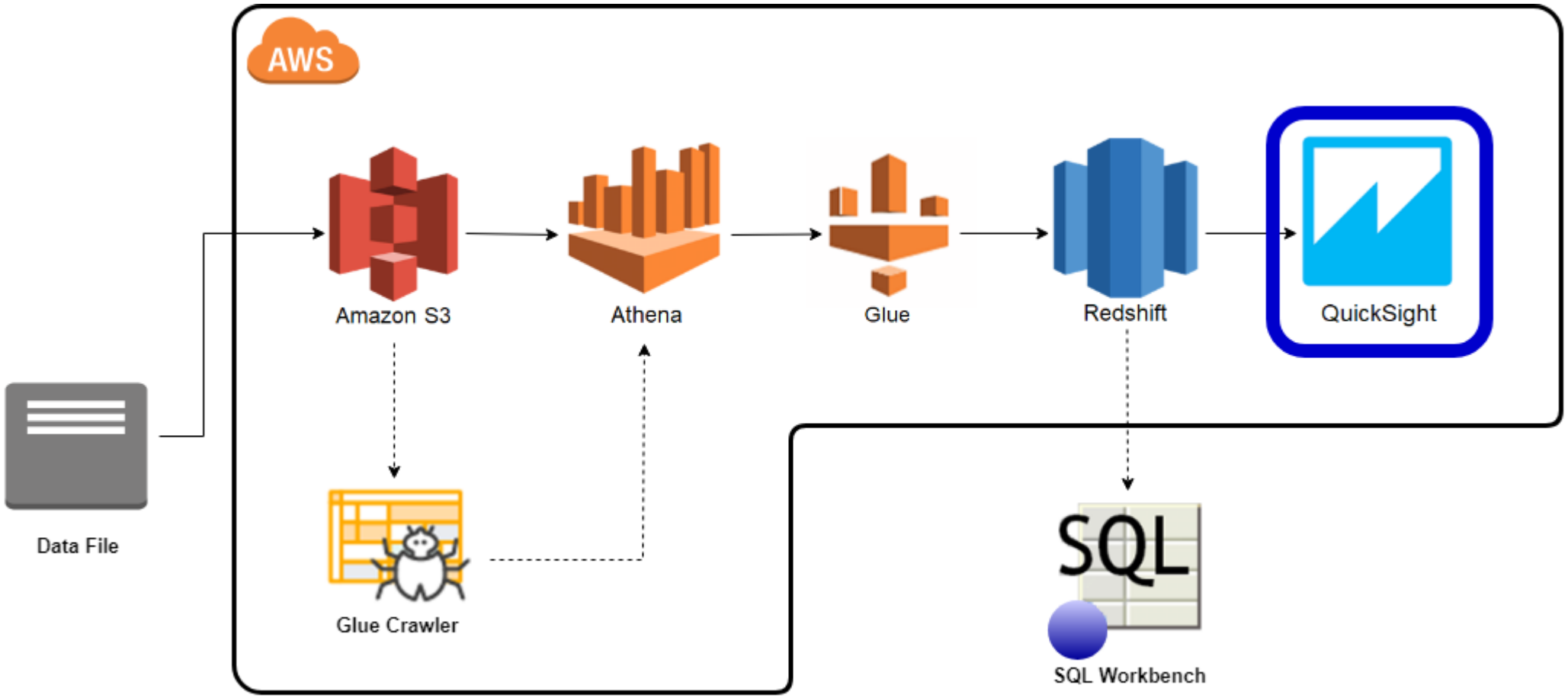
Templates

JavaScript

```javascript
{

    "Description" : "A text description for the template usage",

    "Parameters": {

        // A set of inputs used to customize the template per deployment

    },

    "Resources" : {

        // The set of AWS resources and relationships between them

    },

    "Outputs" : {

        // A set of values to be made visible to the stack creator

    },

    "AWSTemplateFormatVersion" : "2010-09-09"

}
```

QUICKSIGHT

AWS Business Intelligence Tool



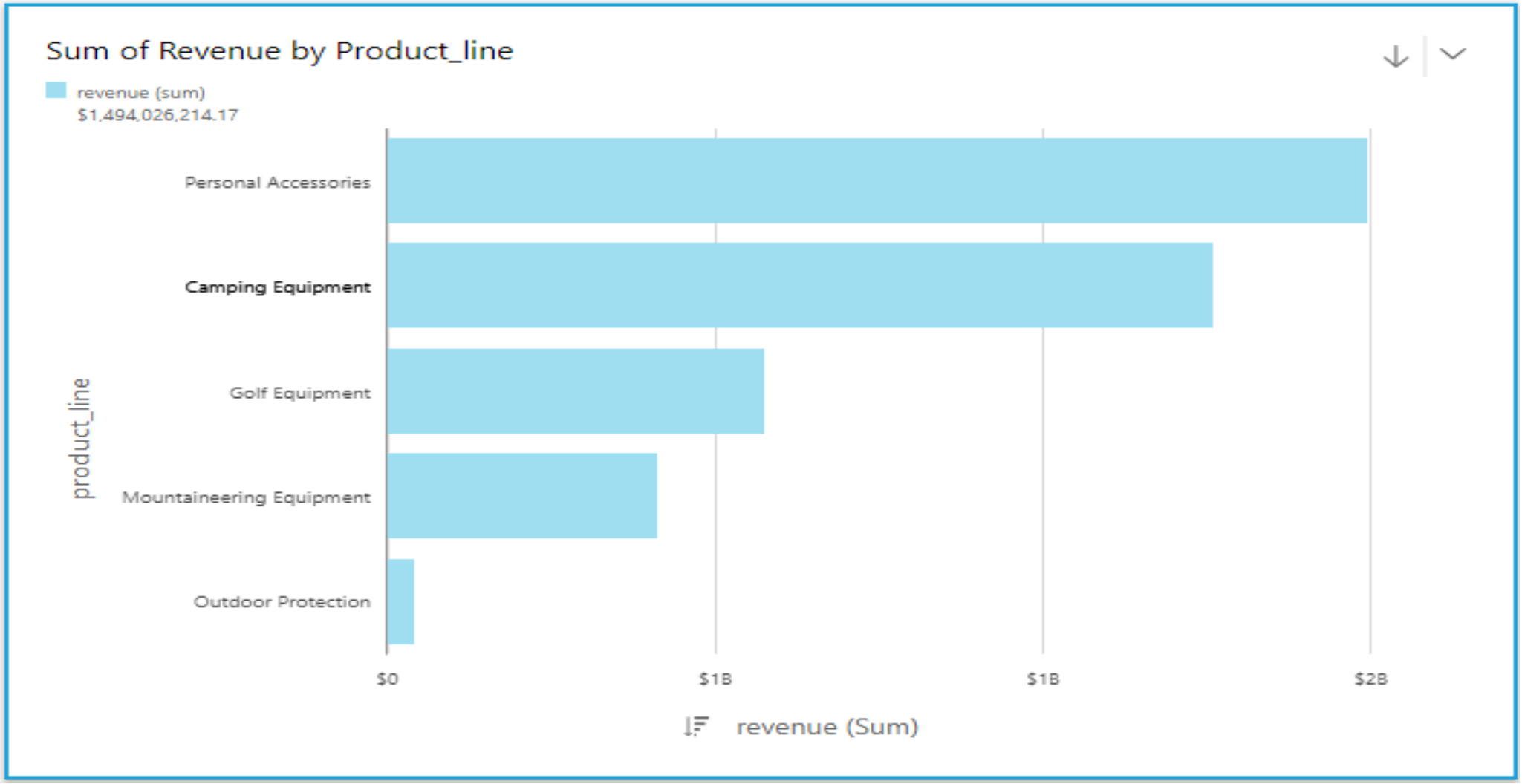
QUICKSIGHT

AWS Business Intelligence Tool

- Cloud based Business Intelligence reporting tool
- Build Reports from
  - Files in s3
  - Redshift
  - Athena

QUICKSIGHT

AWS Business Intelligence Tool



QUICKSIGHT

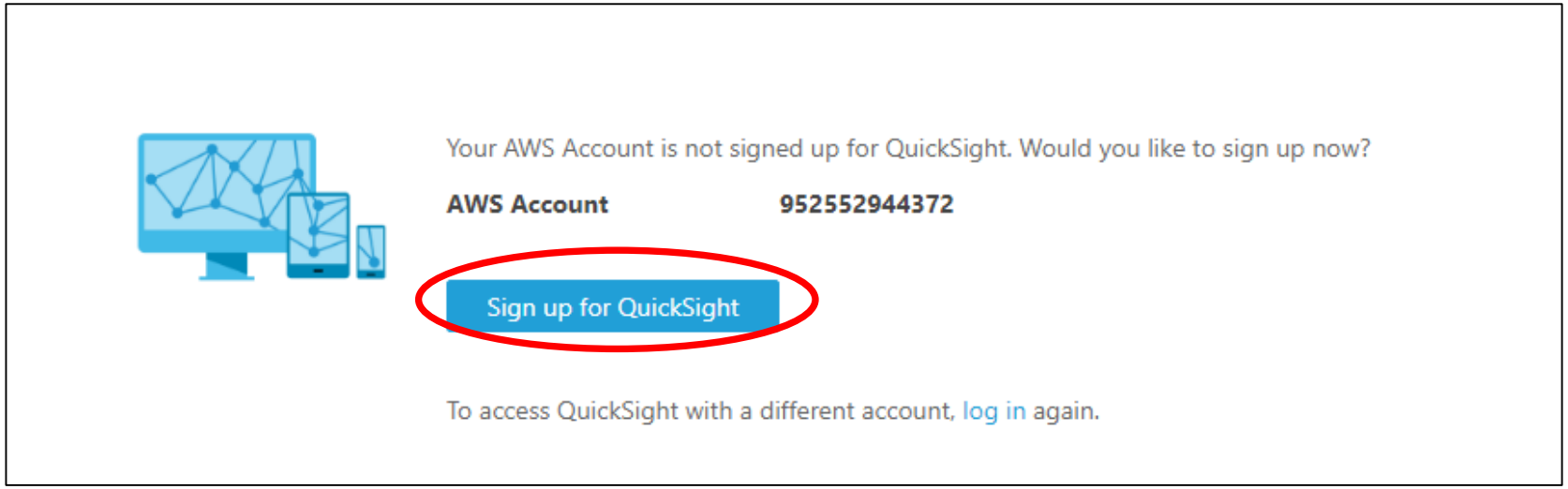
AWS Business Intelligence Tool

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



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

Create your QuickSight account

Edition — Standard

QuickSight account name

jack_silverman

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

Notification email address

jsilveman@manifestcorp.com

For QuickSight to send important notifications.

QuickSight capacity region

EU (Ireland)

☑ 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 — Choose S3 buckets
Enables QuickSight to auto-discover your Amazon S3 buckets

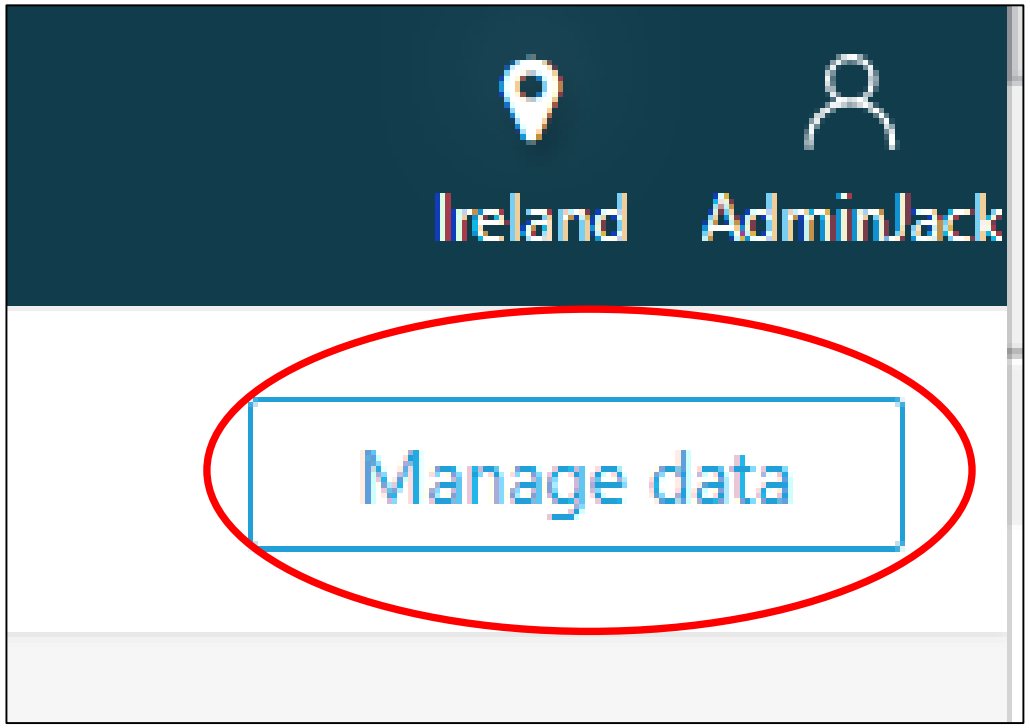
☐ Amazon S3 Storage Analytics
Enables QuickSight to visualize your S3 Storage Analytics data

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

Finish

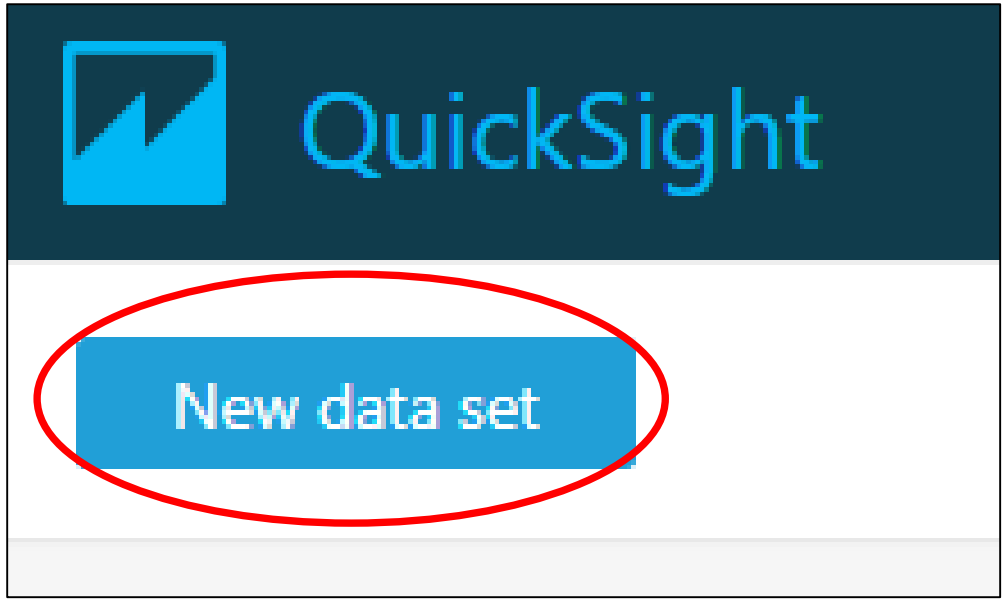
QUICKSIGHT

AWS Business Intelligence Tool



QUICKSIGHT

AWS Business Intelligence Tool



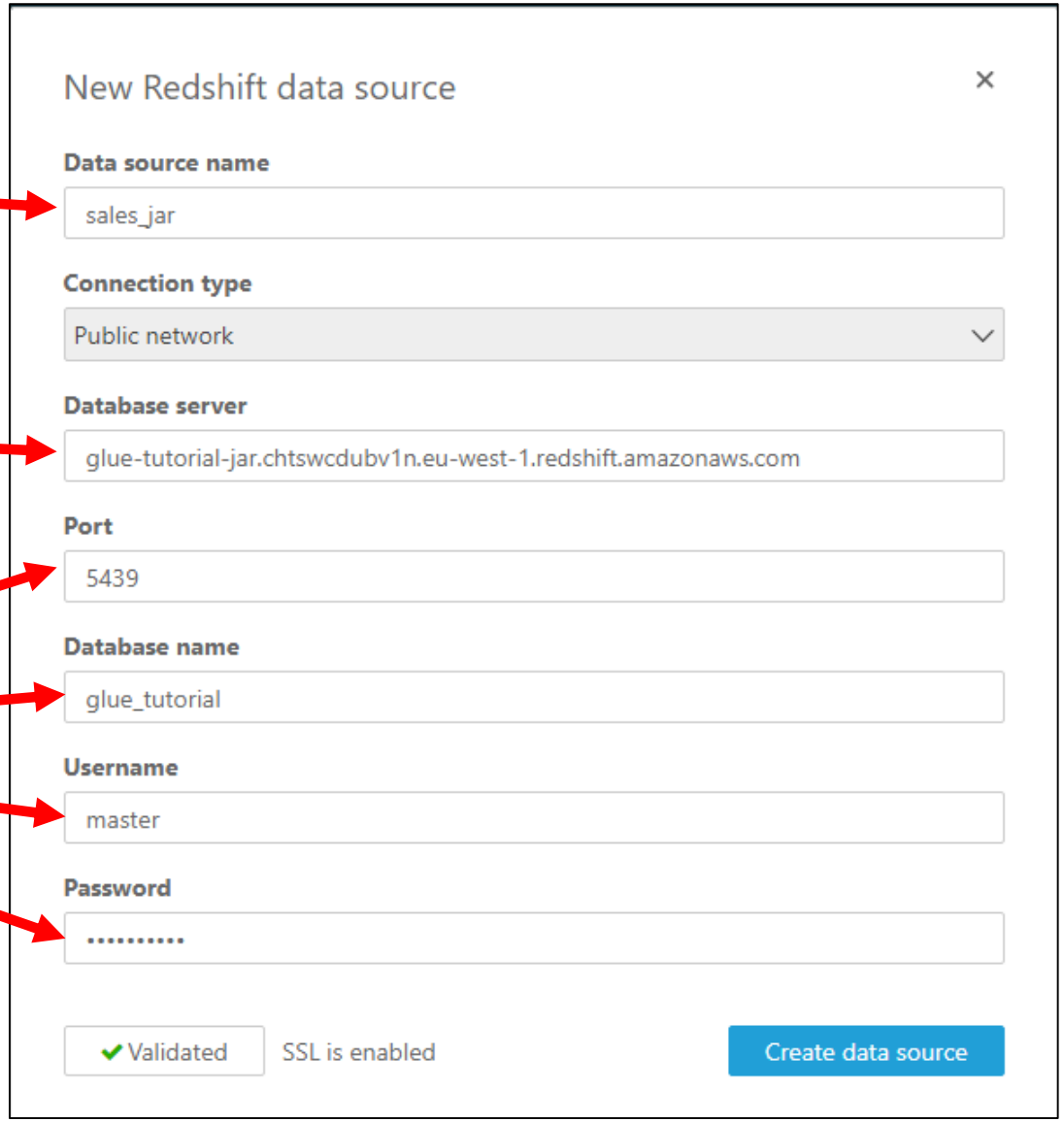
QUICKSIGHT

AWS Business Intelligence Tool

Give you data set a name

This is the Redshift endpoint without port number

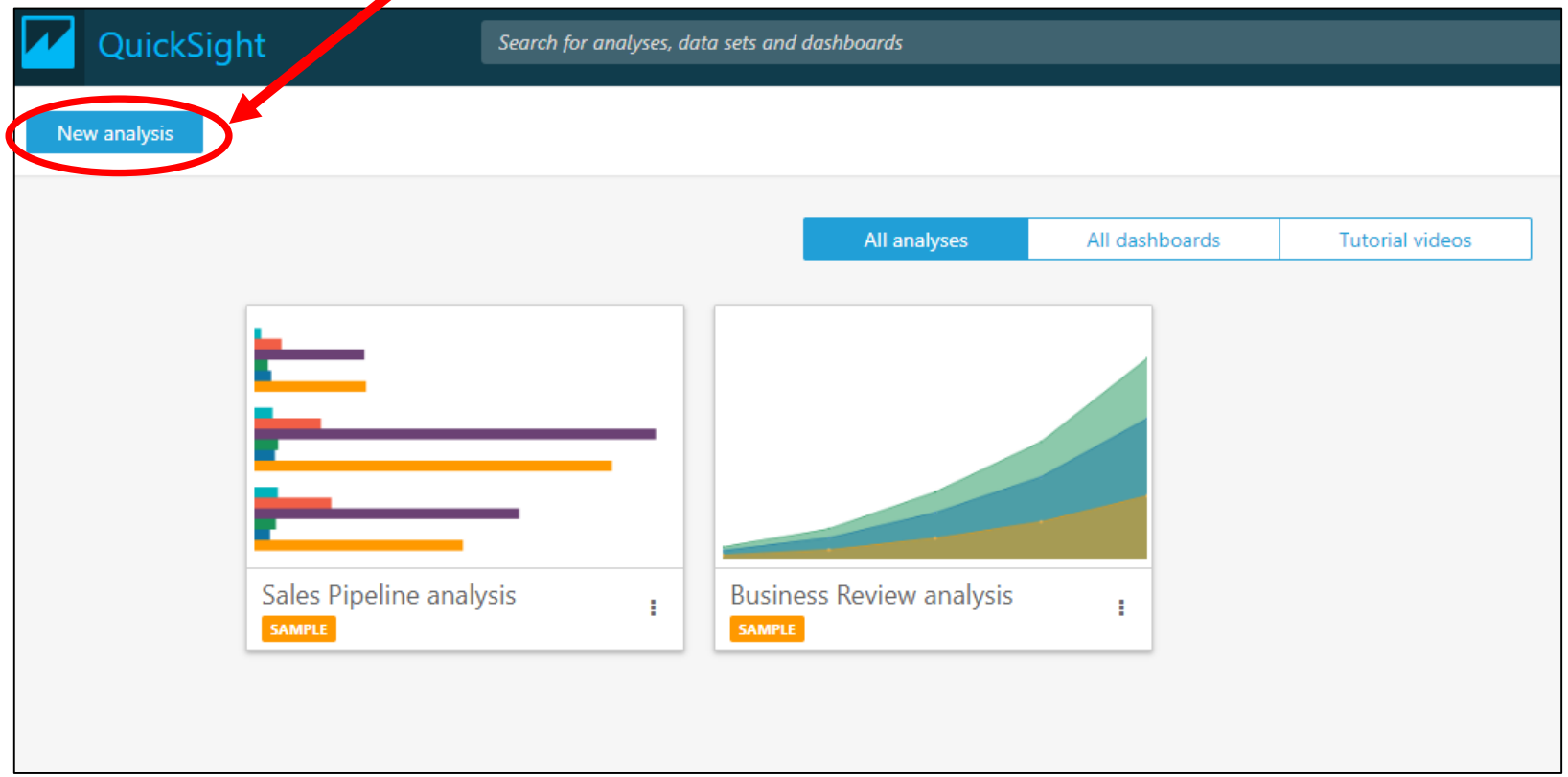
This information comes from the Redshift Cluster



QUICKSIGHT

AWS Business Intelligence Tool

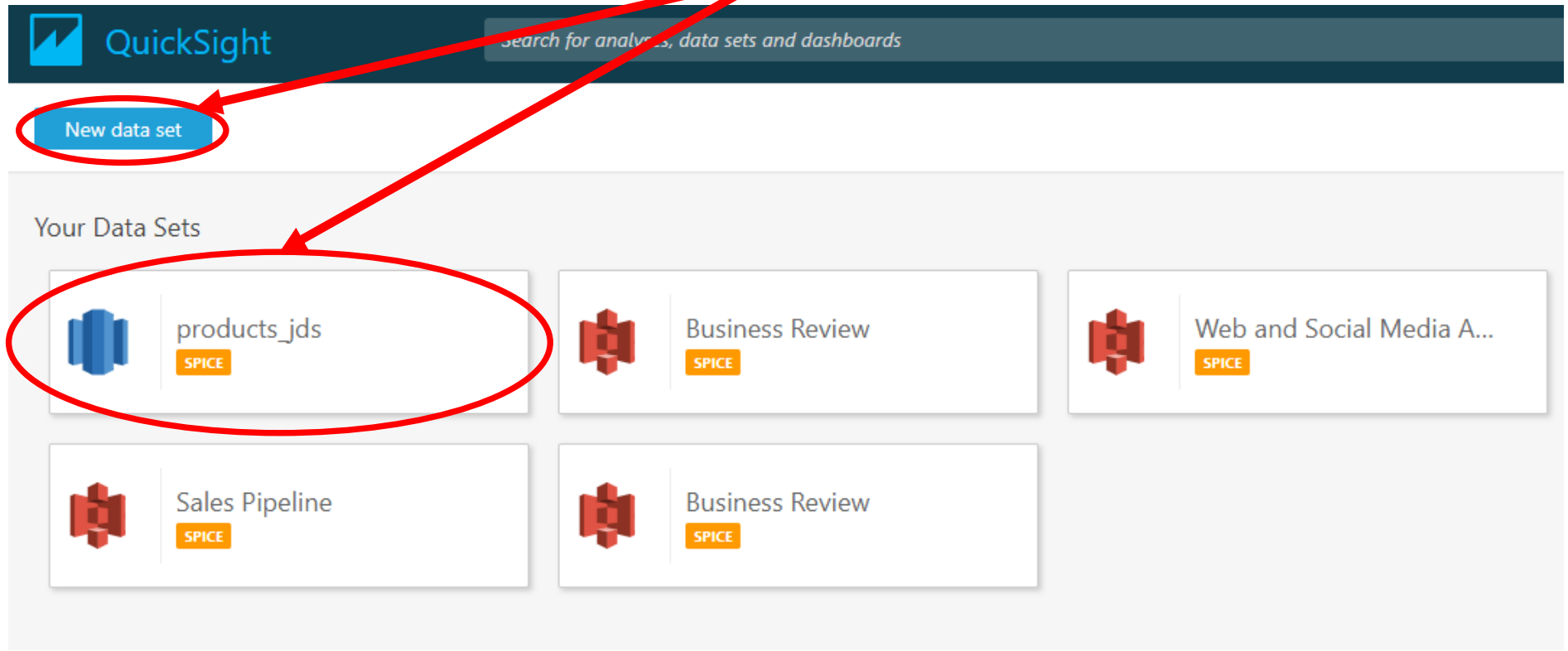
Select new Analysis



QUICKSIGHT

AWS Business Intelligence Tool

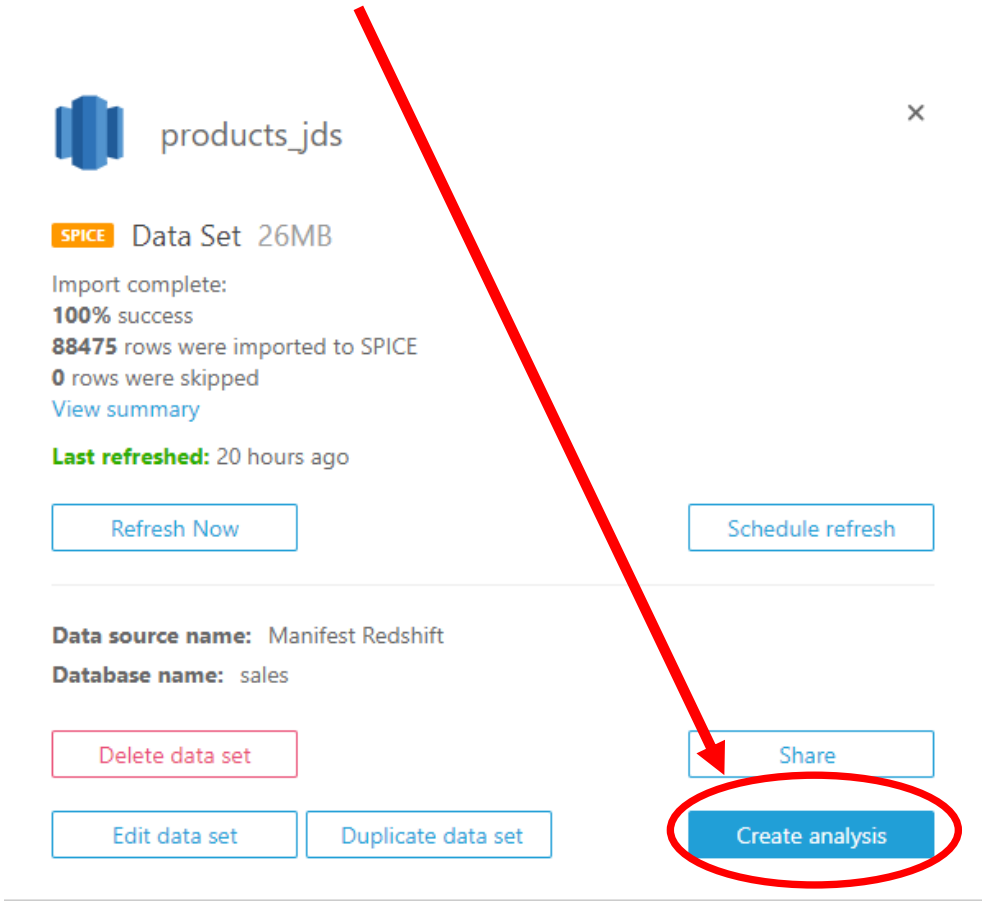
Select product_jds (table in Redshift)



QUICKSIGHT

AWS Business Intelligence Tool

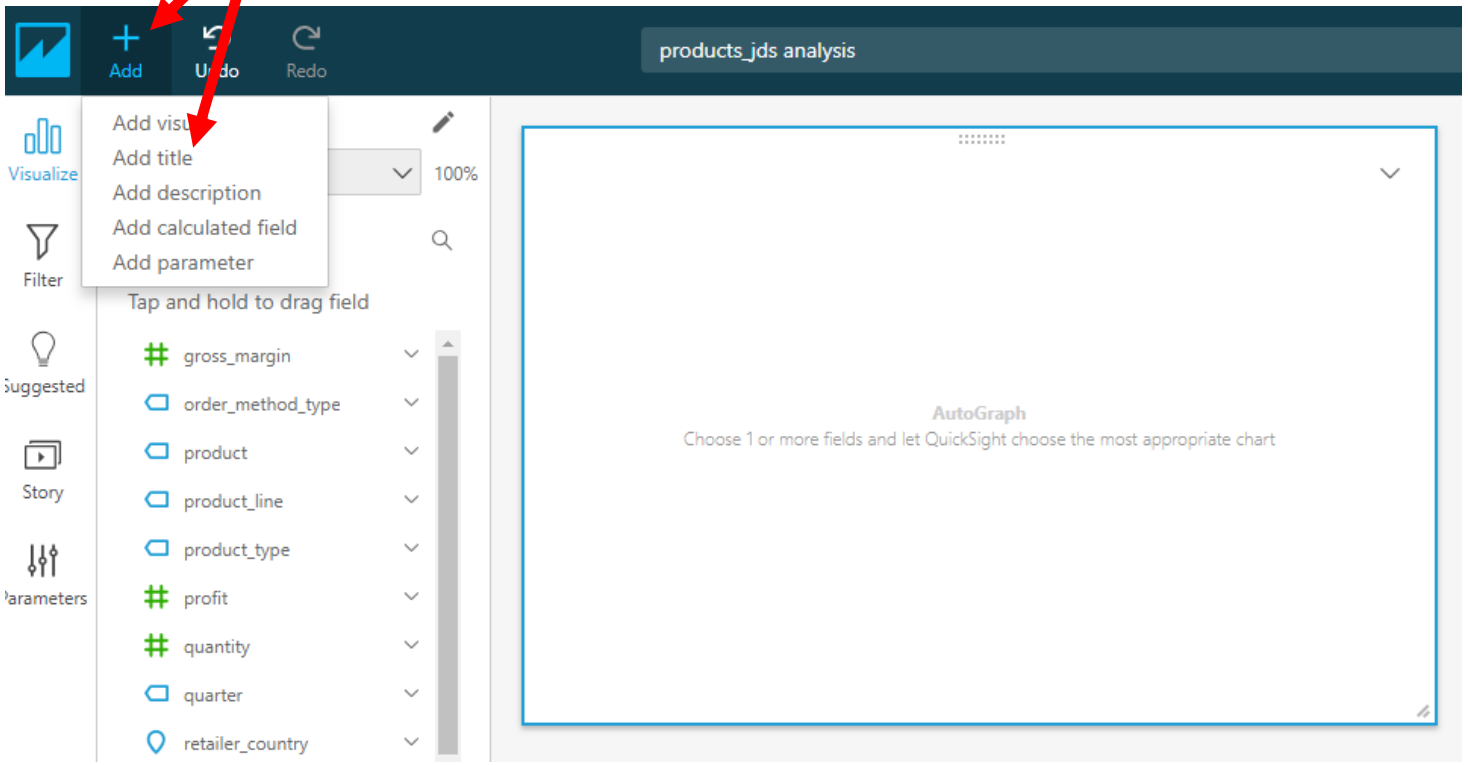
Select Create Analysis



QUICKSIGHT

AWS Business Intelligence Tool

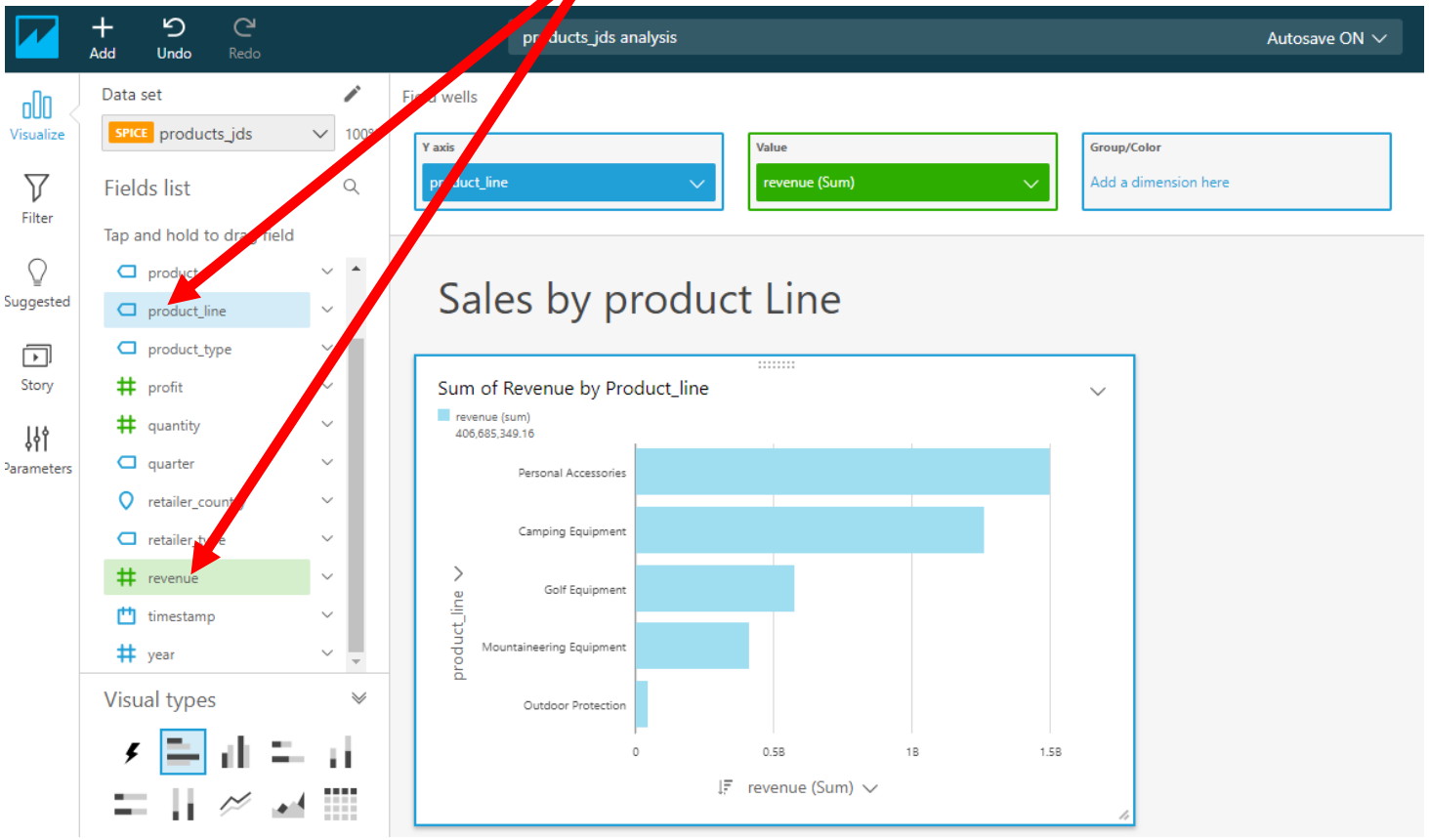
Select Add > Add Title



QUICKSIGHT

AWS Business Intelligence Tool

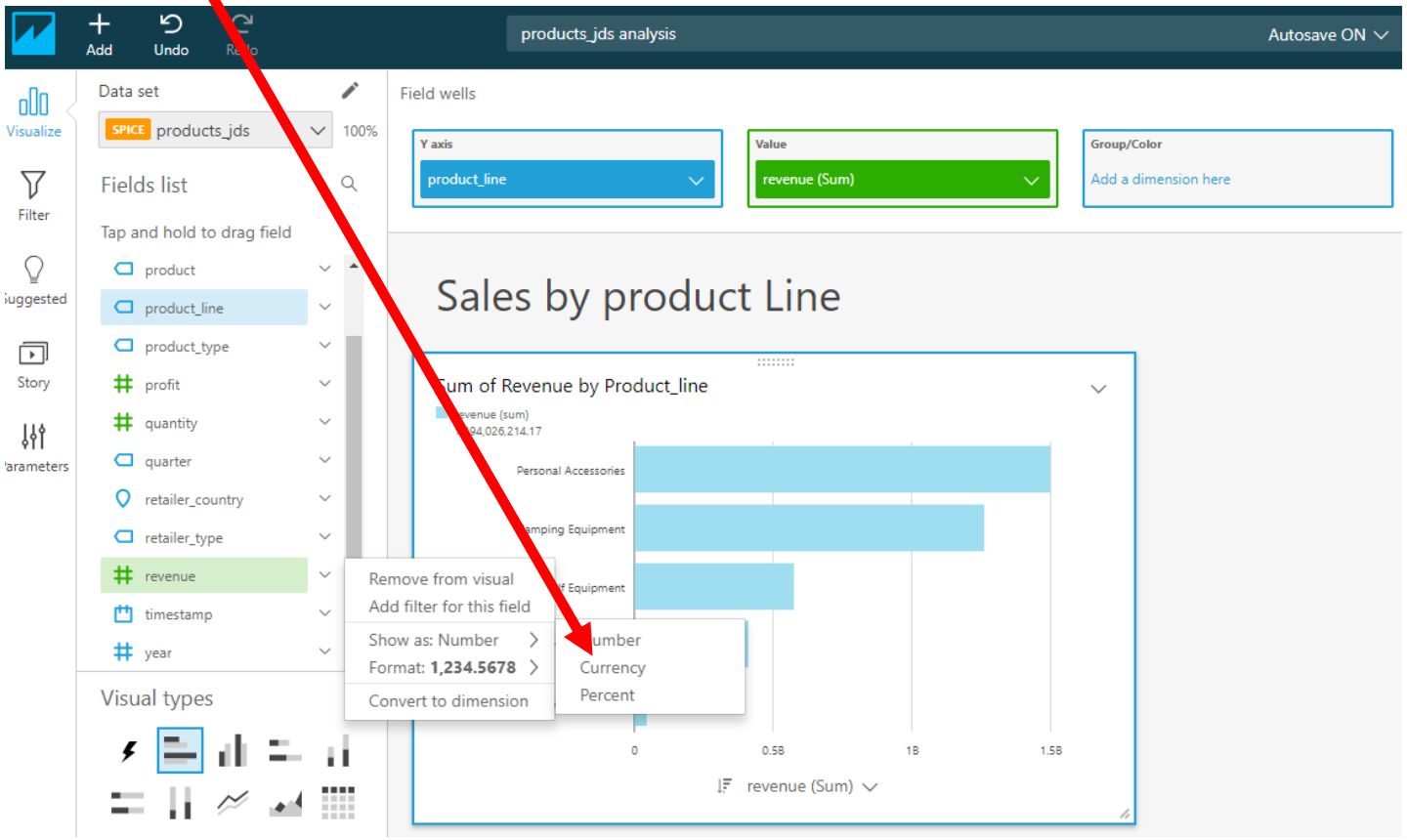
Choose product_line and revenue



QUICKSIGHT

AWS Business Intelligence Tool

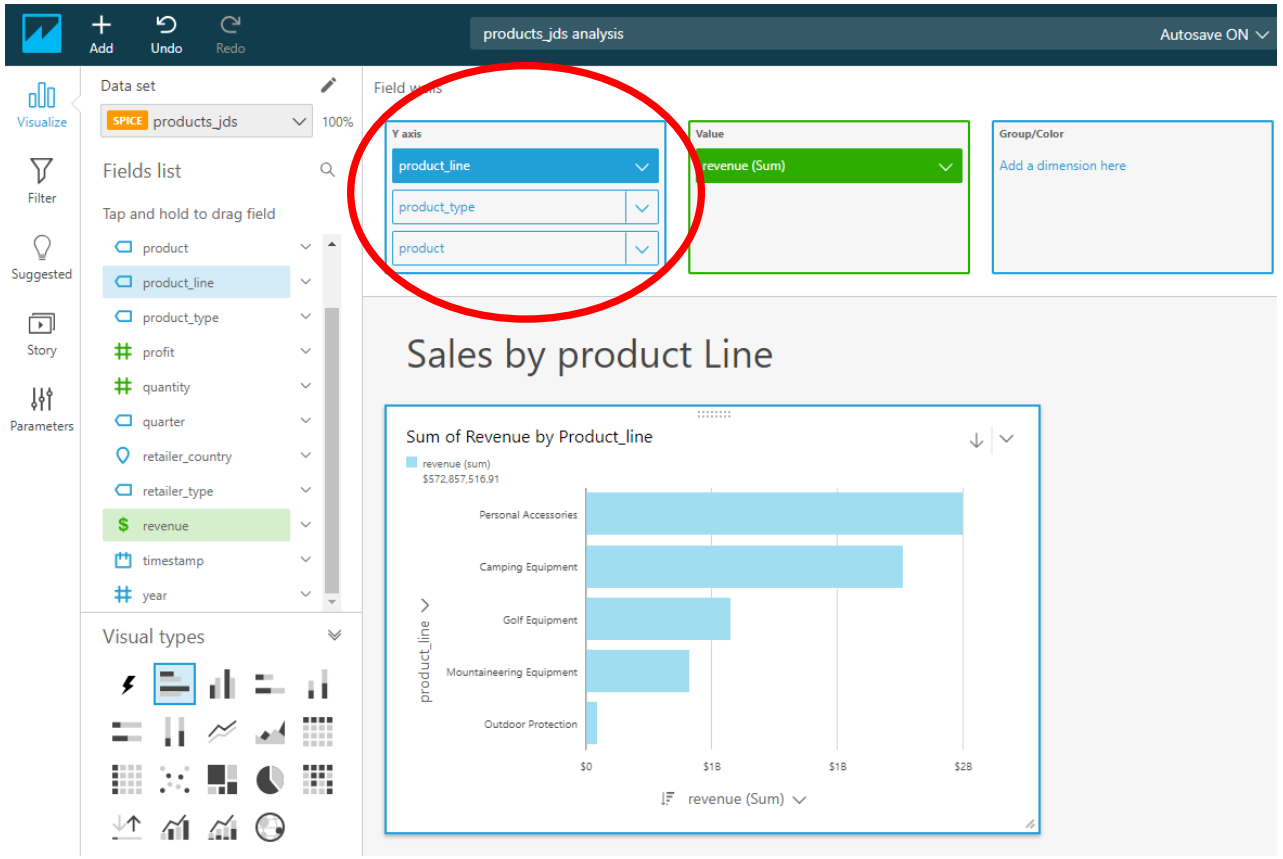
Change the format of Revenue to Currency



QUICKSIGHT

AWS Business Intelligence Tool

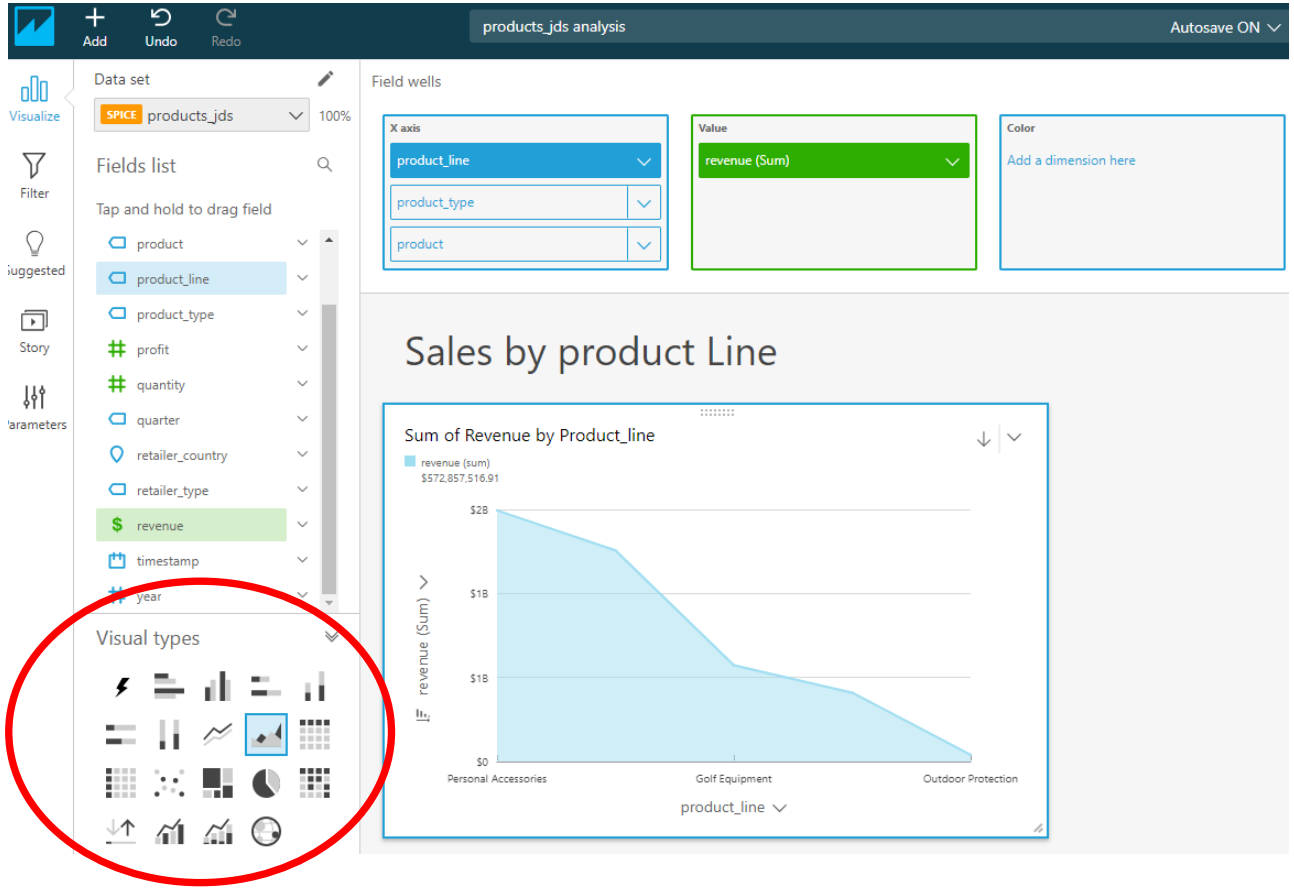
Add product_type and product as drill down layer



QUICKSIGHT

AWS Business Intelligence Tool

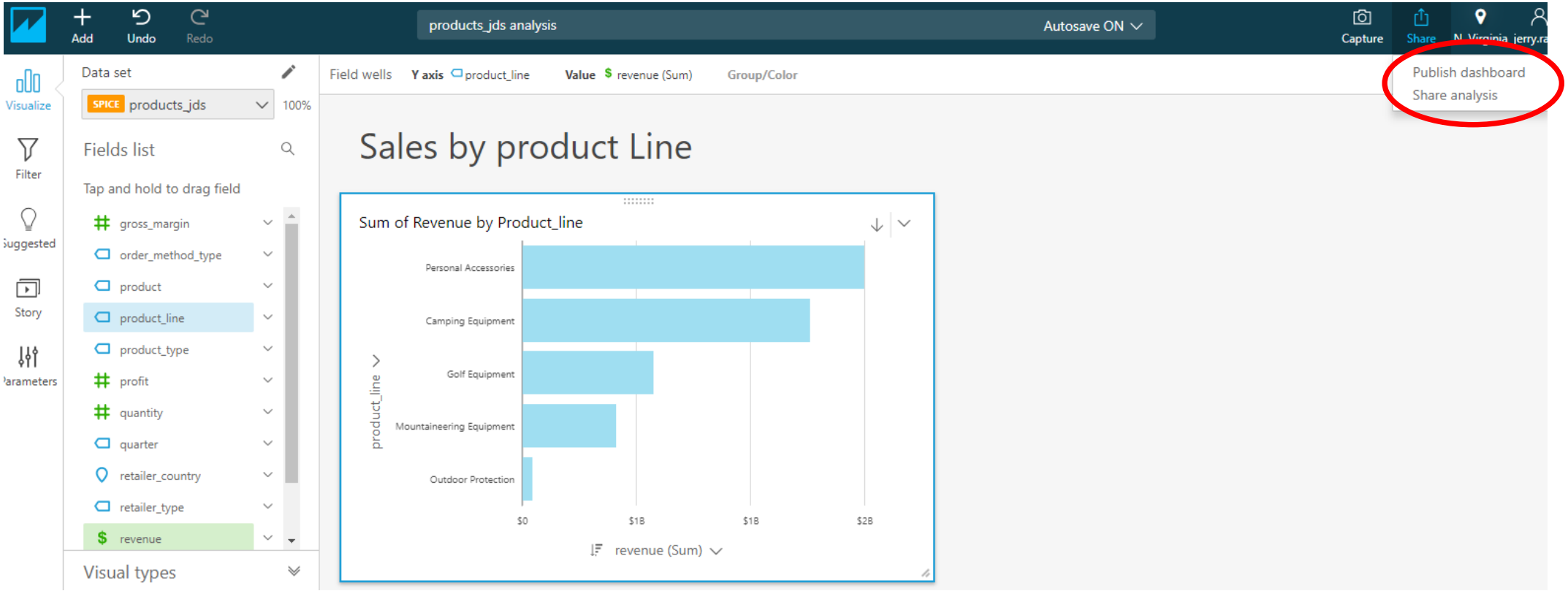
Change Visual Type



QUICKSIGHT

AWS Business Intelligence Tool

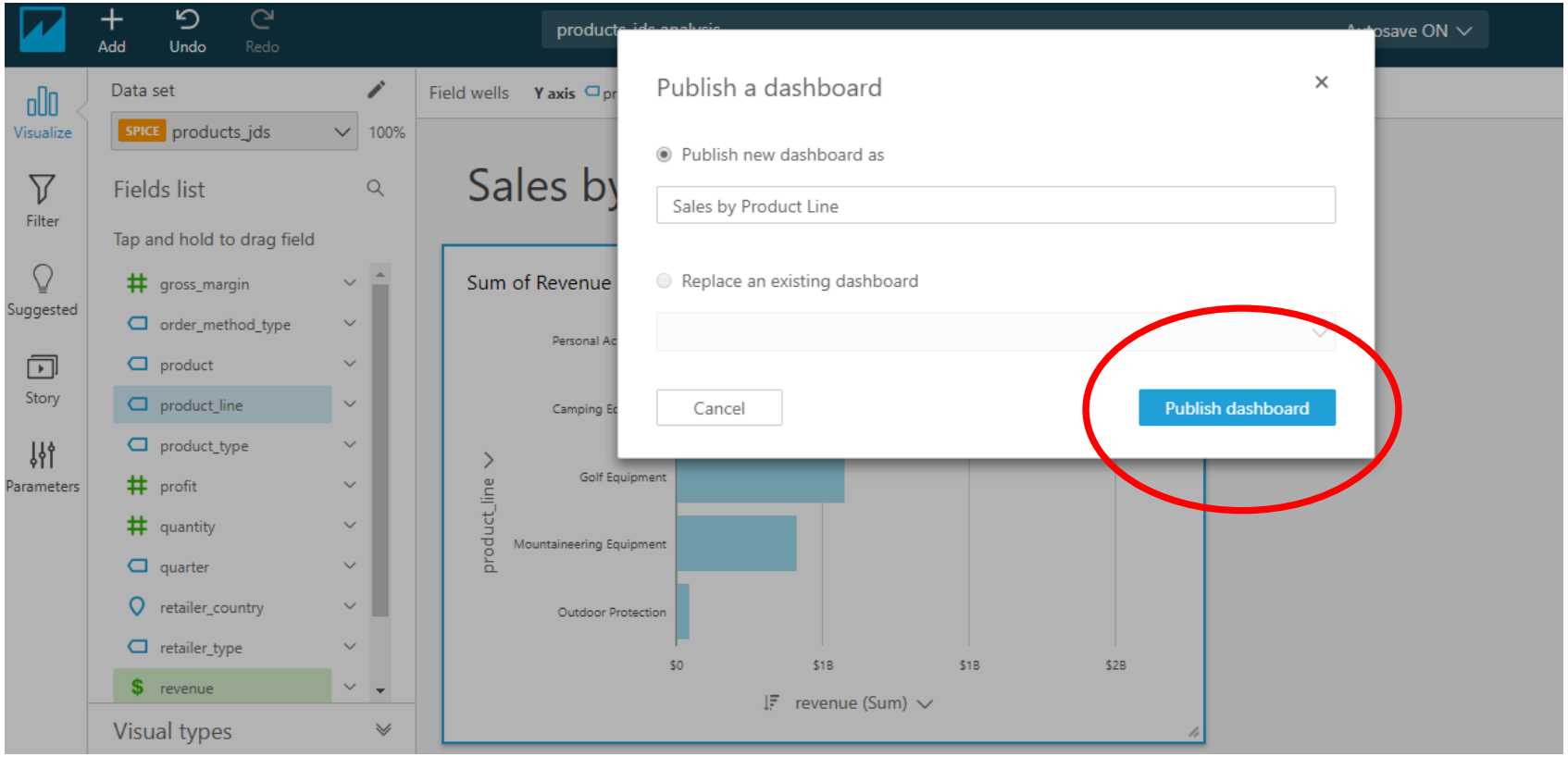
Publish to Dashboard



QUICKSIGHT

AWS Business Intelligence Tool

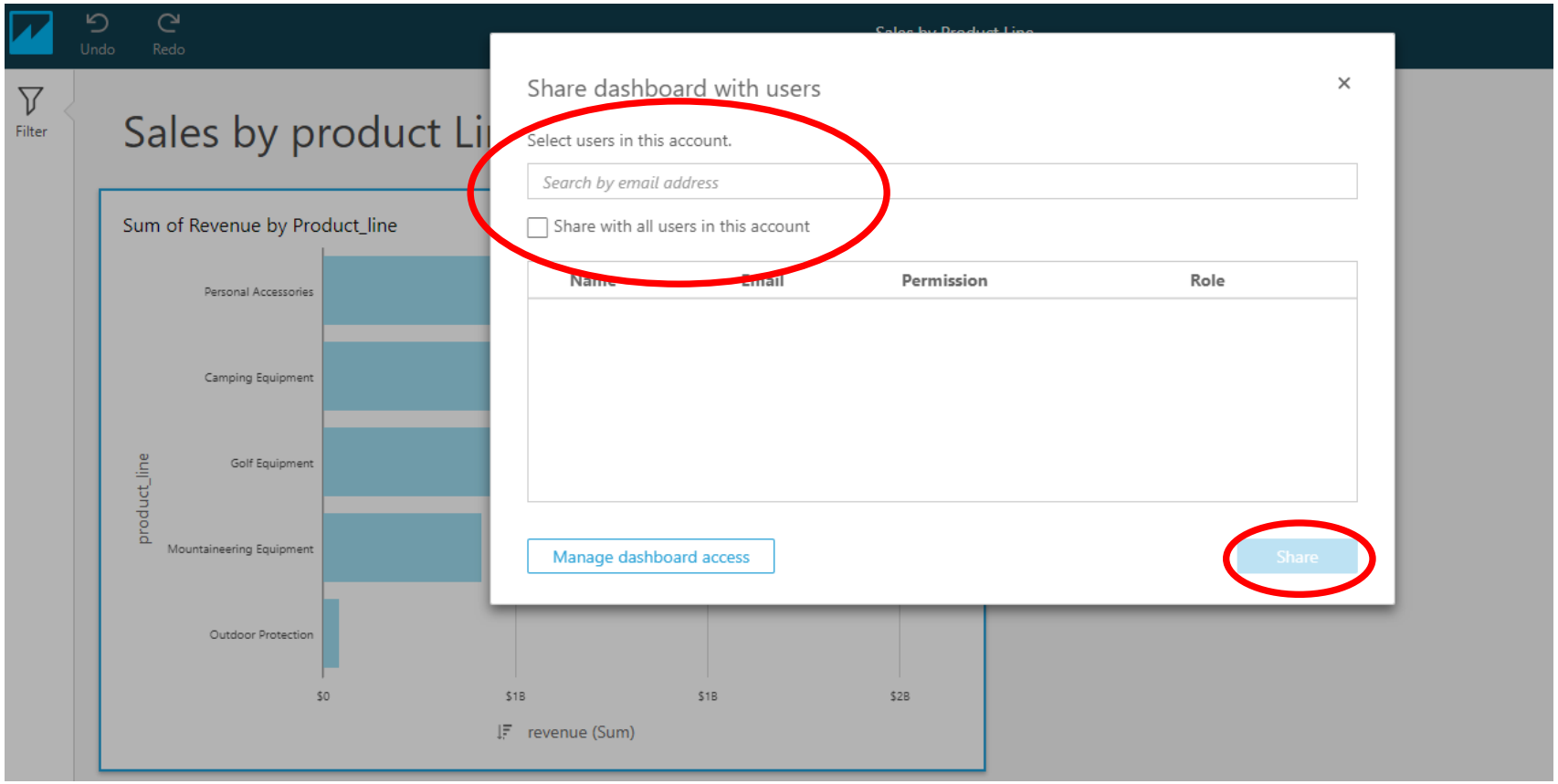
Name the Dashboard and select Publish dashboard



QUICKSIGHT

AWS Business Intelligence Tool

Share the dashboard



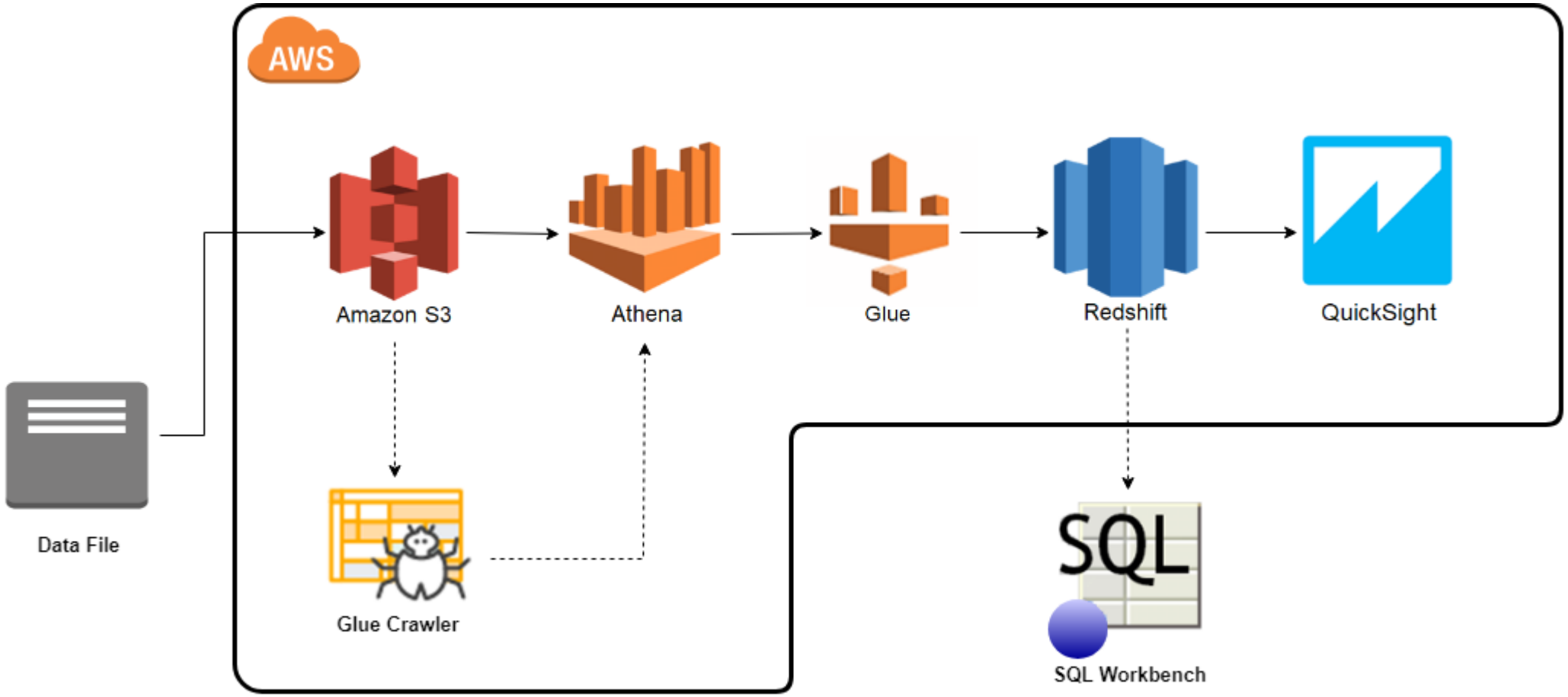
Lab 4

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

(Use US-EAST-1/N. Virginia 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 *
- Glue Job
- Glue Database
- Glue Table
- Glue Connection

* Redshift and S3 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/