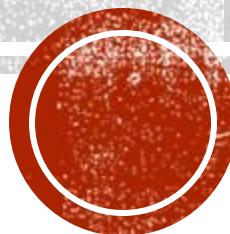
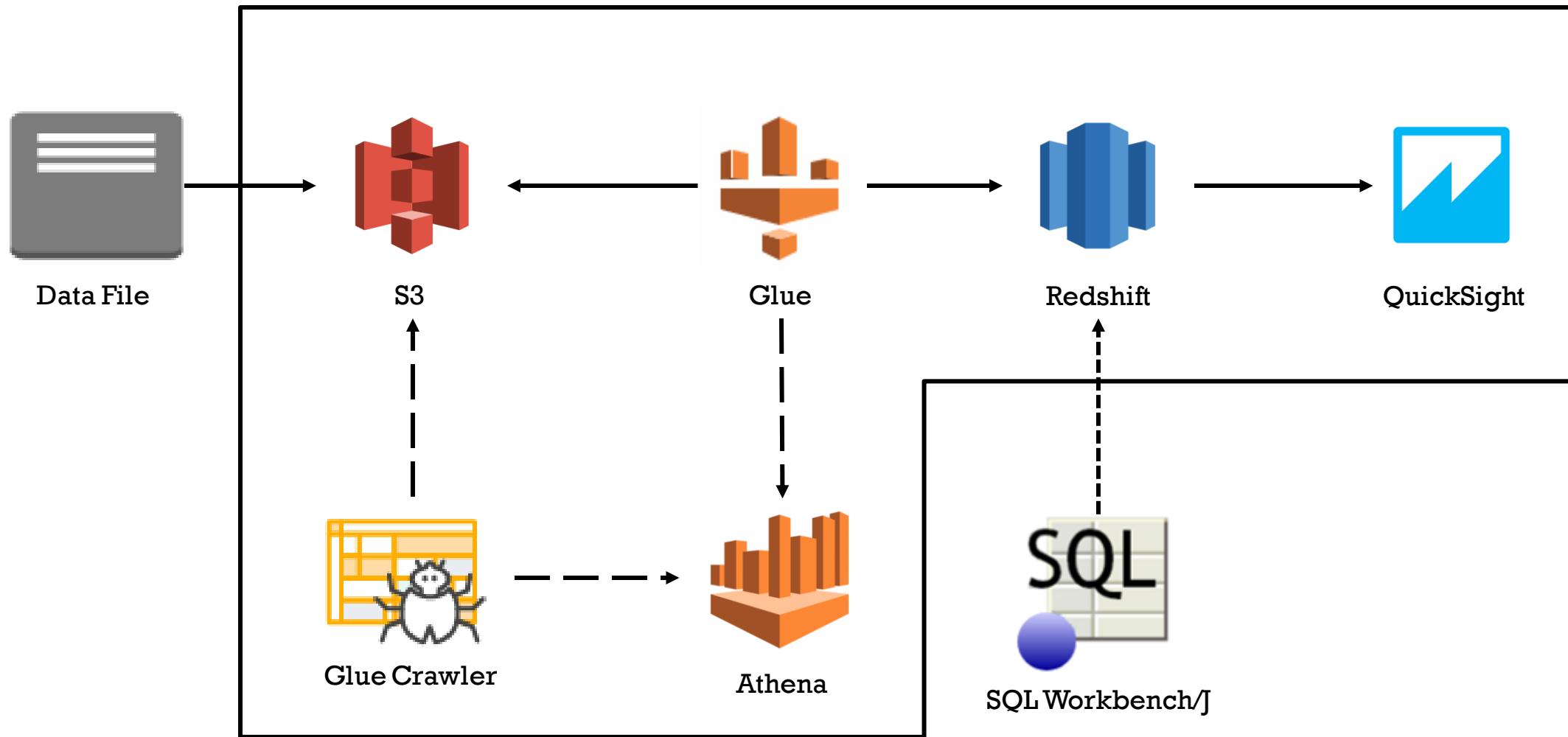


ARTS & CRAFTS WITH AWS GLUE

ETL Workshop

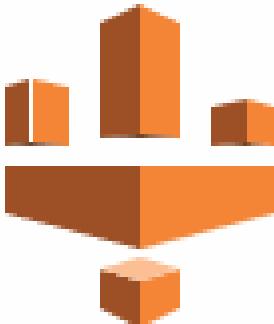


Amazon Web Services



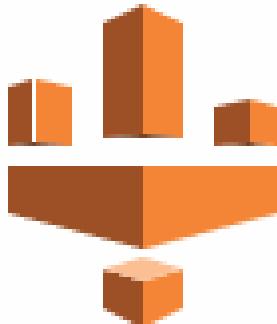
AWS Glue

What is Glue?



AWS Glue

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



ETL

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



ETL

Original Data File(WA_Sales_Products.csv)

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

Example Business Requirements:

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



Why use Glue?

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



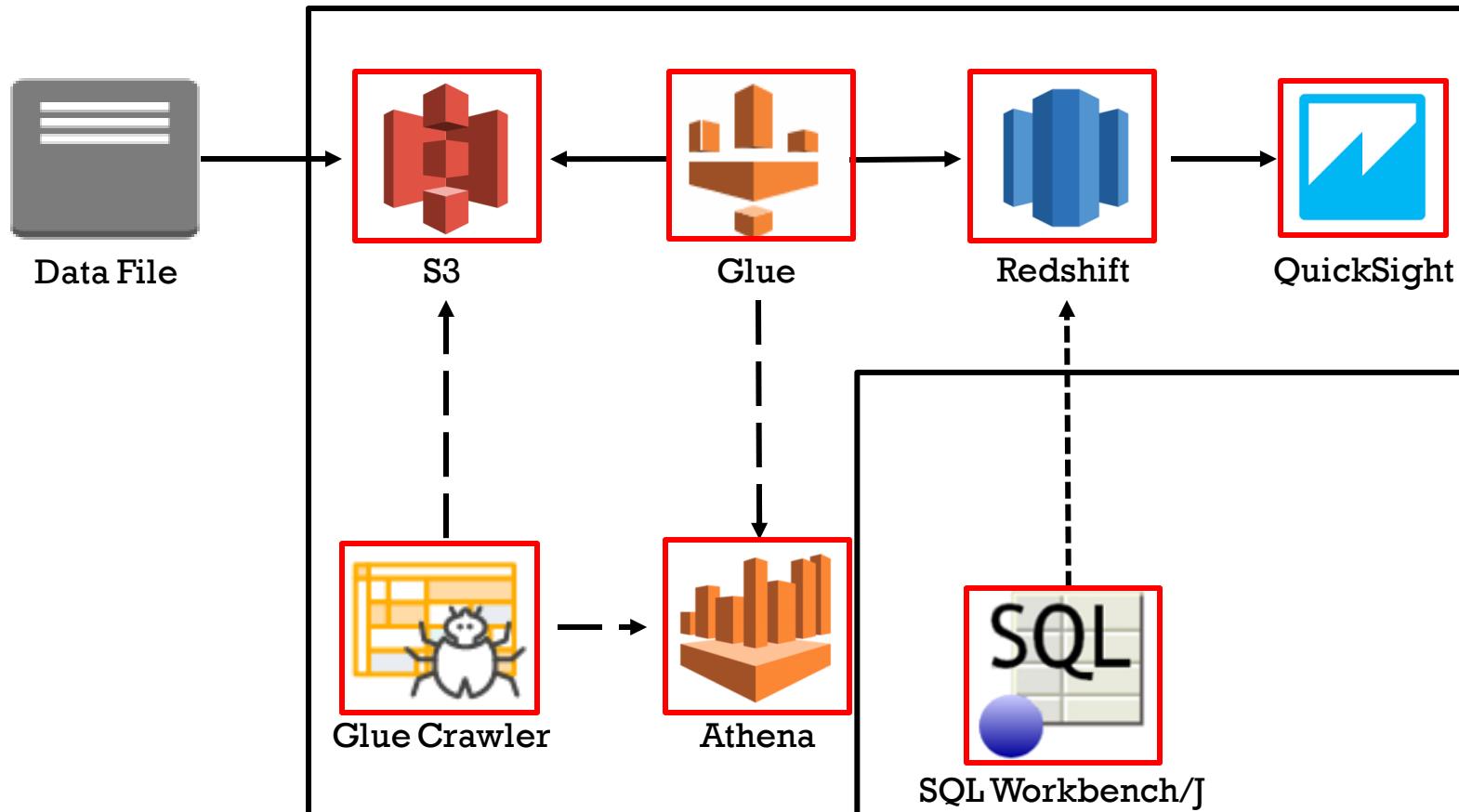
AWS vs. Hadoop

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

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



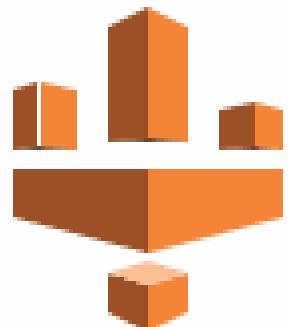
GLUE TUTORIAL OVERVIEW



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

Glue Tutorial Prerequisites

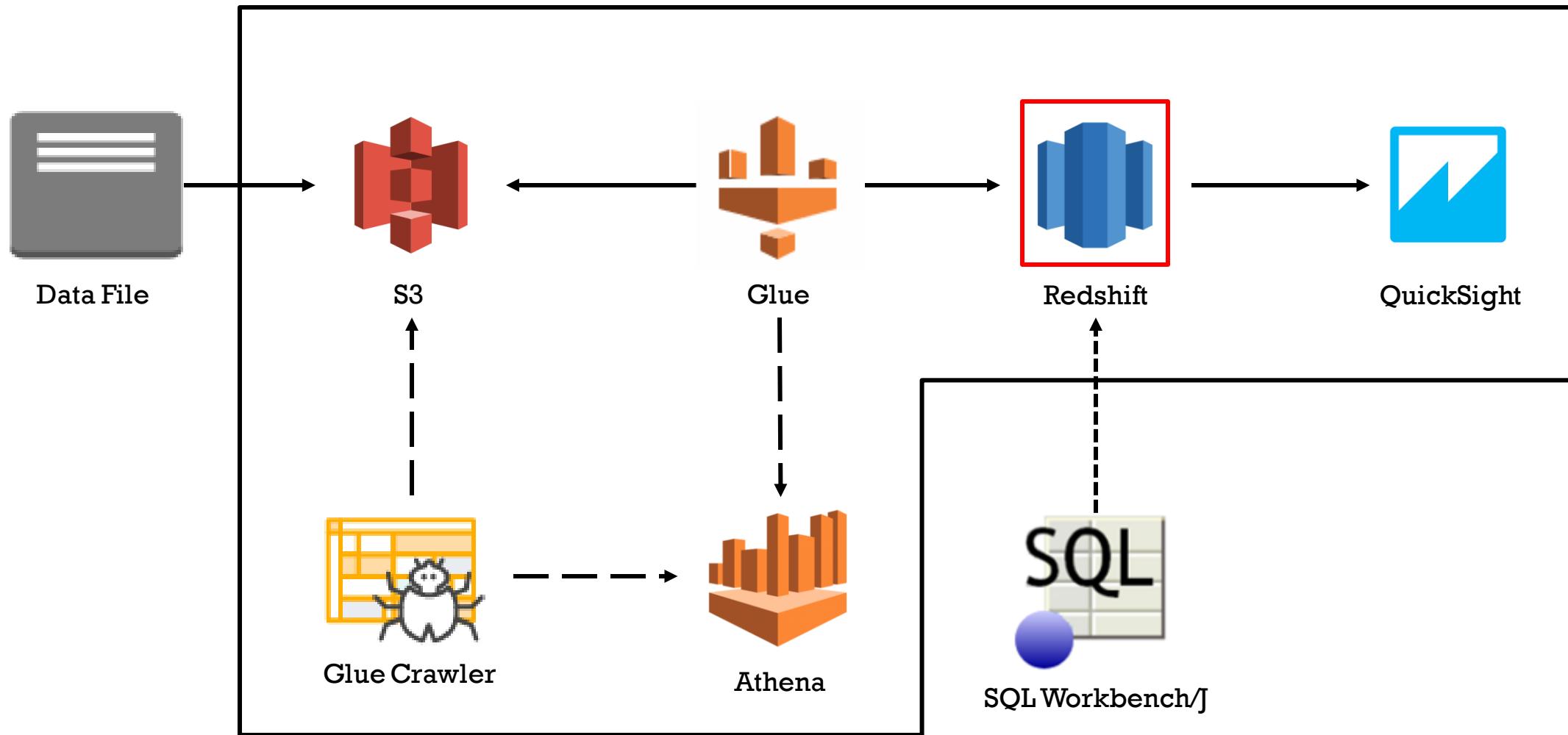
- Prerequisites :
 - Setup AWS Account (Personal or Temporary)
 - <https://portal.aws.amazon.com/billing/signup>
 - <http://escab.juddsolutions.com/registration/create>
 - Clone or save git repository <https://github.com/jayzandme/aws-glue-tutorial.git>
 - download SQL Workbench/j <https://www.sql-workbench.eu/>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



Redshift



└ Create AWS Data Warehouse





Create AWS Data Warehouse

The screenshot shows the Amazon Redshift console interface. On the left, there's a vertical navigation bar with icons for DASHBOARD, CLUSTERS (which is selected and highlighted in orange), QUERIES, EDITOR, CONFIG, MARKETPLACE, and ADVISOR. The main content area features a large heading "Amazon Redshift" and a sub-headline: "Fast, simple, cost-effective data warehouse that can extend queries to your data lake". Below this, a paragraph explains: "Amazon Redshift is a fast, scalable data warehouse that makes it simple and cost-effective to analyze all of your data across your data warehouse and data lake." To the right, there's a callout box titled "Create cluster" with the subtext: "With a few clicks, you can create your first Amazon Redshift cluster in minutes." A prominent orange button labeled "Create cluster" is at the bottom of this box, and it is circled in red. At the very bottom of the page, there's a section titled "Pricing and cost". The top navigation bar includes the AWS logo, "Services", "Resource Groups", a search bar, and user account information for "James-AWS", "Ohio", and "Support".

Redshift



Create AWS Data Warehouse

DASHBOARD

CLUSTERS Create cluster

QUERIES

EDITOR

CONFIG

MARKETPLACE

ADVISOR

ALARMS

GRID

Create cluster

Cluster configuration

Calculate the best configuration for your needs

Node type
Choose a node type that meets your CPU, RAM, storage capacity, and drive type requirements.

Recommended

RA3		
High performance with scalable managed storage		
<input type="radio"/>	ra3.16xlarge	\$13.04/node/hour Managed storage: up to 64 TB /node
\$0.024/GB/month		

DC2

DC2		
High performance with fixed local SSD storage		
<input checked="" type="radio"/>	dc2.large	\$0.25/node/hour Storage: 160 GB/node
<input type="radio"/>	dc2.8xlarge	\$4.80/node/hour Storage: 2.6 TB/node

dc2.large Selected

dc2.large
Storage: 160 GB/node

dc2.8xlarge
Storage: 2.6 TB/node

dc2.large
2 vCPU (gen 2)

Redshift



Create AWS Data Warehouse

EVENTS

WHAT'S NEW

▶ Show legacy dense storage node types

Nodes
Enter the number of nodes that you need.

1

Range (1-32)

Configuration summary
dc2.large | 1 node

\$180.00/month
Estimated compute price
Save more than 60% of your costs by purchasing reserved nodes.
[Learn more](#)

160 GB
Total compressed storage
The total storage capacity for the cluster if you deploy the number of nodes that you chose.





Create AWS Data Warehouse

Specify Cluster Name



Cluster details

Cluster identifier

This is the unique key that identifies a cluster.

The identifier must be from 1 to 63 characters. Valid characters are a-z (lowercase only) and - (hyphen).

Database port (optional)

Port number of the port where the database accepts inbound connections.



The port must be numeric (1150-65535).

Create a user



Master user name

Enter a login ID for the master user of your DB instance.

The name must be 1-128 alphanumeric characters, and it can't be a **reserved word**.

Create a password
for the user



Master user password



Show password

- The value must be 8-64 characters.
- The value must contain at least one uppercase letter.
- The value must contain at least one lowercase letter.
- The value must contain at least one number.
- The master password can only contain ASCII characters (ASCII codes 33-126), except ' (single quotation mark), " (double quotation mark), /, \, or @.



Redshift



Create AWS Data Warehouse

Unclick 'Use defaults'

Take note of the VPC id for later

Change 'Publicly Accessible' to Yes

Additional configurations Use defaults

These configurations are optional, and default settings have been defined to help you get started with your cluster. Turn off "Use defaults" to modify these settings now.

▼ Network and security

Virtual private cloud (VPC)
This VPC defines the virtual networking environment for this cluster.

Default VPC
vpc-b2fb56da

(i) You cannot edit the VPC once the cluster has been created. [Learn more](#) X

Cluster subnet group
Choose the Amazon Redshift subnet group to launch the cluster in.

default

Availability Zone
Specify the Availability Zone that you want the cluster to be created in. Otherwise, Amazon Redshift chooses an Availability Zone for you.

No preference

VPC security group
This VPC security group defines which subnets and IP ranges the cluster can use in the VPC.

Choose one or more security groups

default
sg-797ba212

Publicly accessible
Allow instances and devices outside the VPC connect to your database

No Yes



Create AWS Data Warehouse

Give your cluster a
Database to start with

▼ Database configurations

Database name (optional)
Specify a database name to create an additional database.

The name must be 1-64 alphanumeric characters (lowercase only), and it can't be a [reserved word](#).

Parameter groups
Defines database parameter and query queues for all the databases.

Default parameter group for redshift-1.0

Encryption
Encrypt all data on your cluster.
 Disabled
 Use AWS Key Management Service (AWS KMS)
 Use a hardware security module (HSM)



Redshift



Create AWS Data Warehouse

Amazon Redshift > Clusters

Clusters (1)

All status	Search			
Cluster	Status	Storage capacity us...	CPU utilization	Snapshots
glue-tutorial-xxx dc2.large 1 node 160 GB	Available	< 1%	33%	4 snapshots

DASHBOARD

CLUSTERS

QUERIES

EDITOR



Lab 1

- Launch Redshift cluster
(Use US-EAST-2/Ohio Region)



EC2

Edit Security Groups

In a new tab go to
the EC2 service

The screenshot shows the AWS EC2 console interface. On the left, there's a navigation sidebar with various services like Dedicated Hosts, Capacity Reservations, IMAGES (AMIs), ELASTIC BLOCK STORE (Volumes, Snapshots), NETWORK & SECURITY (Security Groups, Elastic IPs, Placement Groups, Key Pairs). The 'Security Groups' link is circled in orange. The main content area shows a table of security groups with columns: Name, Group ID, Group Name, VPC ID, and Owner. One row is selected, and its Group ID ('sg-797ba212') is highlighted with a red circle. Below the table, a modal window is open for the selected security group ('sg-797ba212'). The modal has tabs for Description, Inbound (which is active and highlighted with an orange border), Outbound, and Tags. The 'Edit' button within the Inbound tab is also circled in orange. At the bottom of the modal, there are fields for Type (All traffic), Protocol (All), Port Range (All), Source (sg-797ba212 (default)), and Description.

Name	Group ID	Group Name	VPC ID	Owner
<input checked="" type="checkbox"/>	sg-797ba212	default	vpc-b2fb56da	681132037743
<input type="checkbox"/>	sg-7bde0610	jlz-db-sg	vpc-f2fa579a	681132037743
<input type="checkbox"/>	sg-9970a9f2	jlz-webapp-sg	vpc-f2fa579a	681132037743
<input type="checkbox"/>	sg-a078a1cb	default	vpc-f2fa579a	681132037743

Security Group: sg-797ba212

Description **Inbound** Outbound Tags

Edit

Type <small>i</small>	Protocol <small>i</small>	Port Range <small>i</small>	Source <small>i</small>	Description <small>i</small>
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.

Cancel **Save**



Redshift



Connection

Go to Redshift and select 'Clusters'

Amazon Redshift > Clusters

Clusters (1)

All status Search

Cluster	Status	Storage capacity us...	CPU utilization	Snapshots
glue-tutorial-xxx dc2.large 1 node 160 GB	Available	< 1%	33%	4 snapshots

● Available dc2.large 1 total nodes

Query monitoring Cluster performance Maintenance and monitoring Backup Properties Schedule

Cluster permissions

Select Properties

Select glue-tutorial-xxx

Endpoint Copy

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

▼ View all connection details.

JDBC URL Copy

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

Scroll down to Cluster Database Properties and copy the JDBC URL



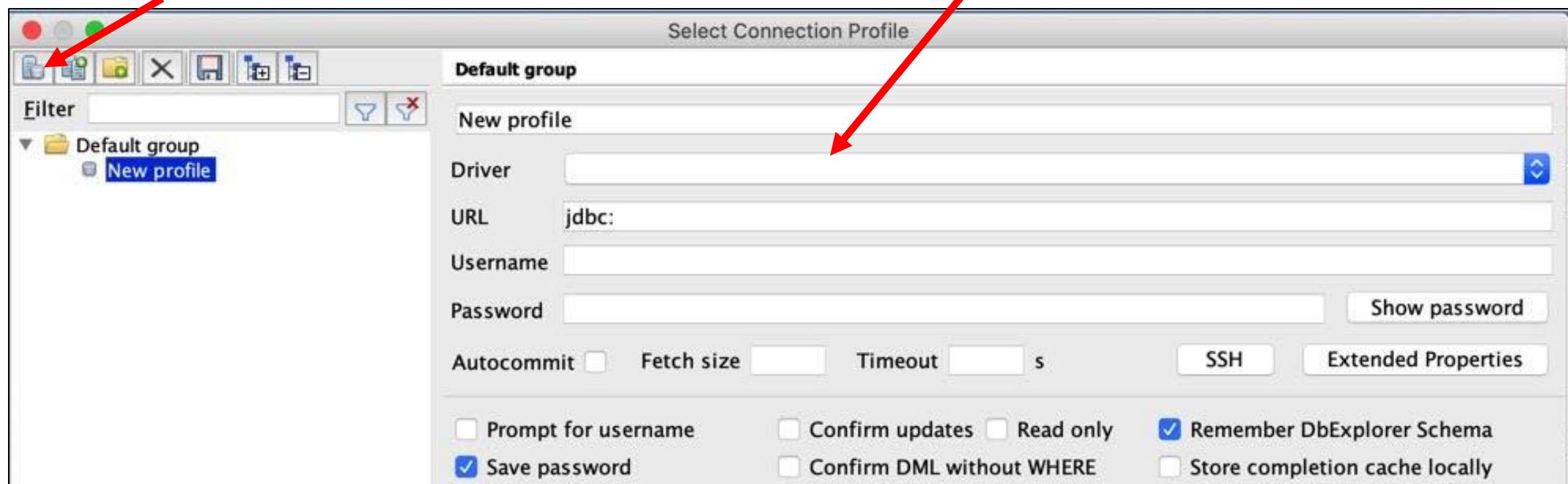
Redshift



└ Connection

Open SQL Workbench and select
Create a new connection

Set the Driver to Amazon Redshift

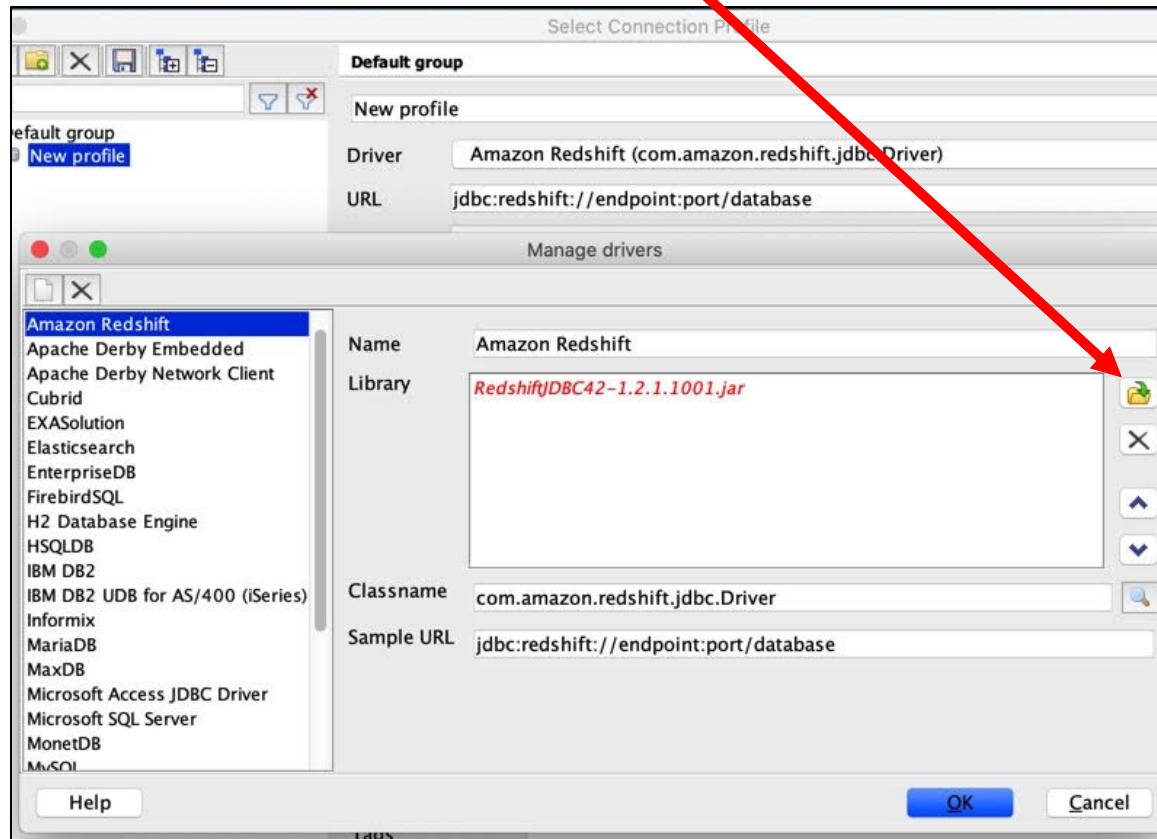


Redshift

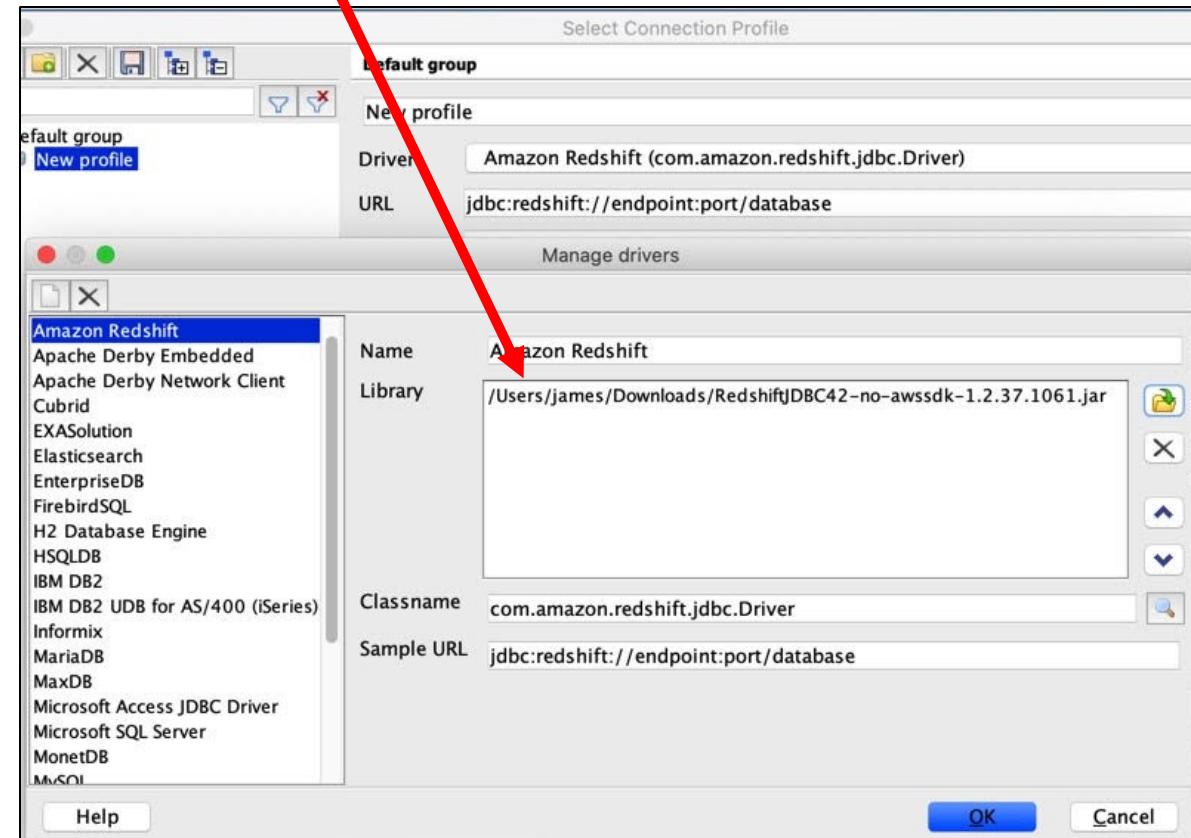
Connection



Choose the file



Select the Redshift JDBC Driver you downloaded earlier

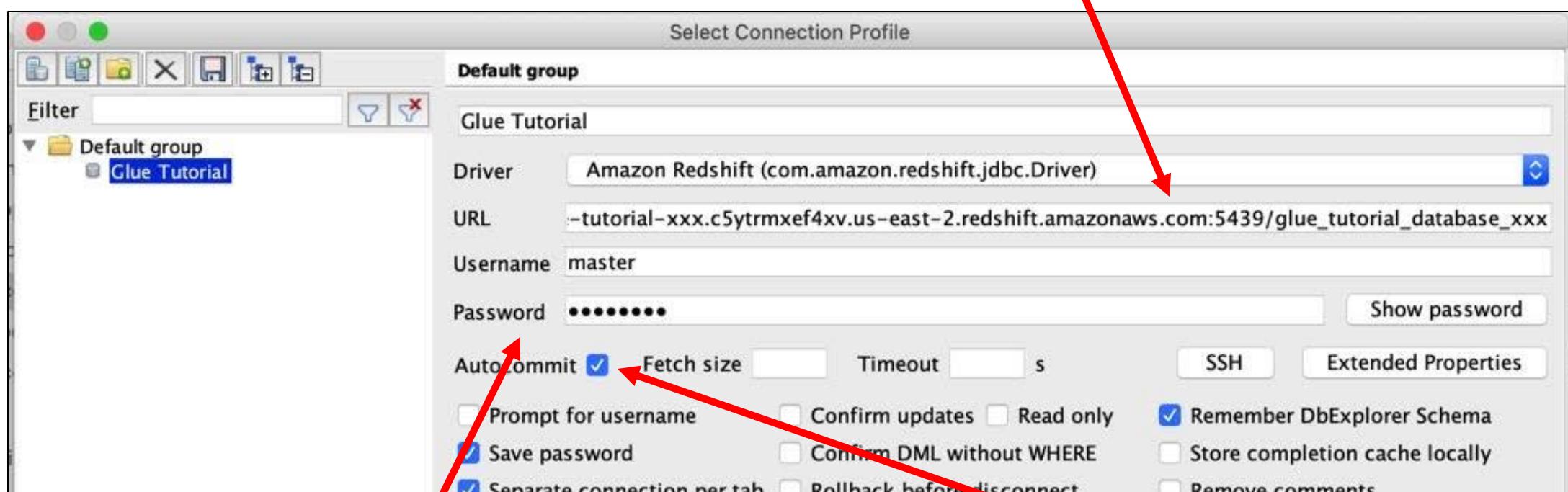


Redshift



Connection

Paste the JDBC URL

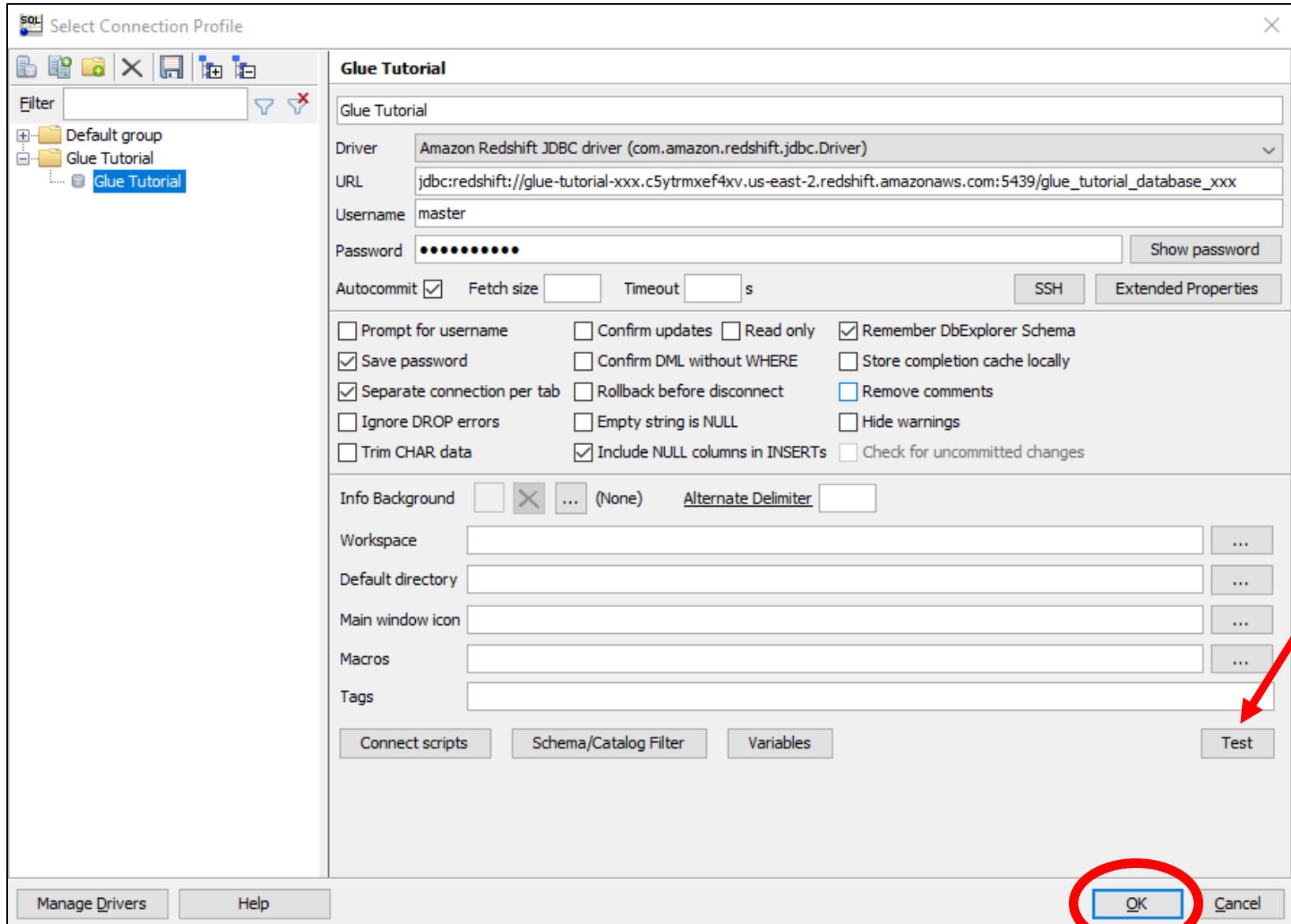


Enter the username and password that you created

Select Autocommit

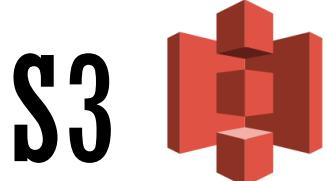
Redshift

Connection

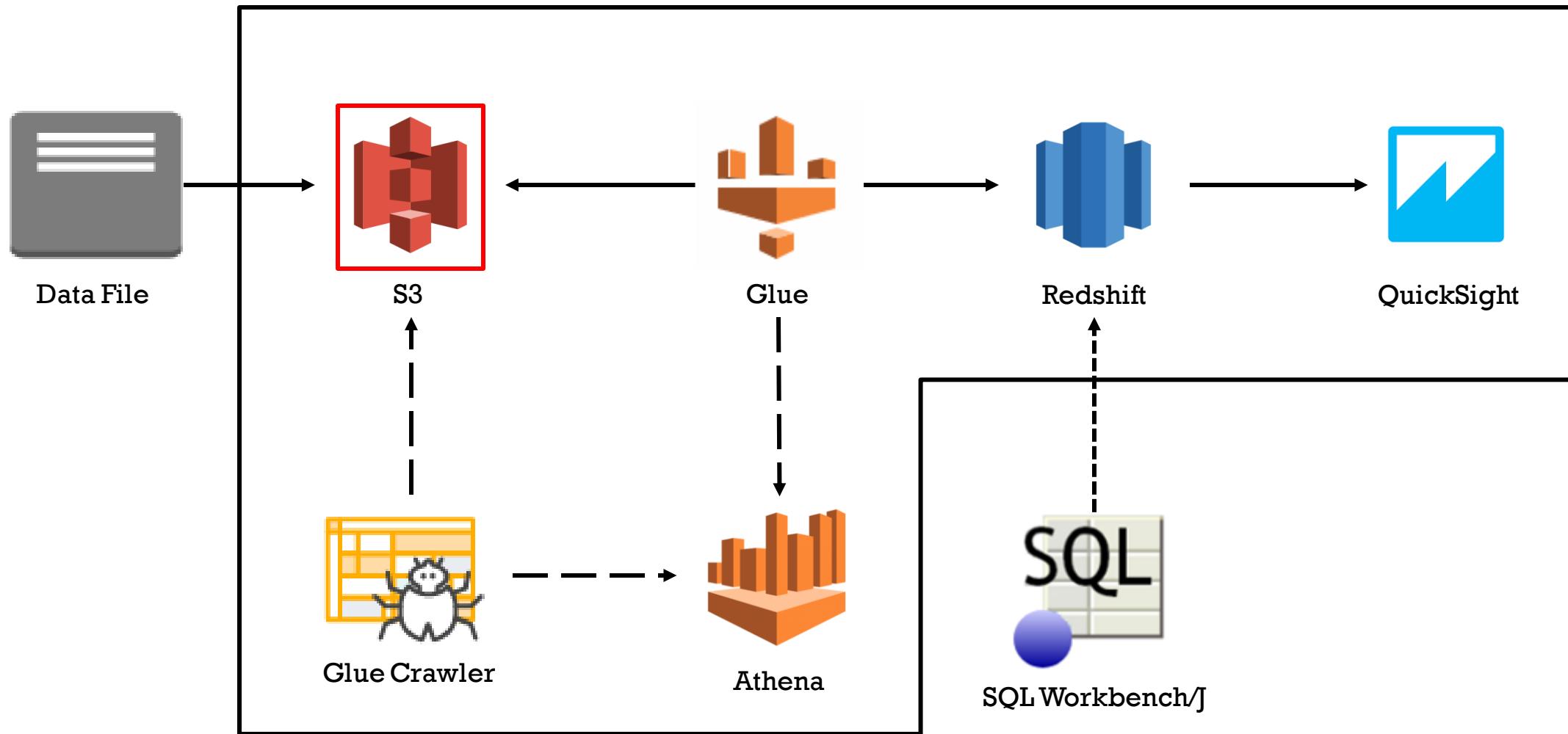


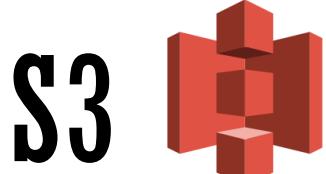
Test your
connection





└ Create S3 bucket with AWS Console





Create S3 bucket with AWS Console

Amazon S3 Documentation

Buckets Batch operations Access analyzer for S3 Block public access (account settings) Feature spotlight 2

S3 buckets

Search for buckets All access types

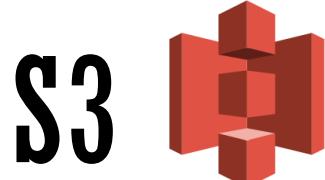
+ Create bucket Edit public access settings Empty Delete 0 Buckets 0 Regions

You do not have any buckets. Here is how to get started with Amazon S3.

Create a new bucket Upload your data Set up your permissions

A screenshot of the AWS S3 console. On the left, there's a sidebar with links for Buckets, Batch operations, Access analyzer for S3, Block public access (account settings), and Feature spotlight (with 2 notifications). The main area is titled 'S3 buckets'. It features a search bar, a dropdown for 'All access types', and a prominent blue button labeled '+ Create bucket' which is circled in red. Below the button, it says 'You do not have any buckets. Here is how to get started with Amazon S3.' There are three icons at the bottom: 'Create a new bucket' (a bucket with a cloud), 'Upload your data' (a bucket with an upward arrow), and 'Set up your permissions' (two user silhouettes with a plus sign).

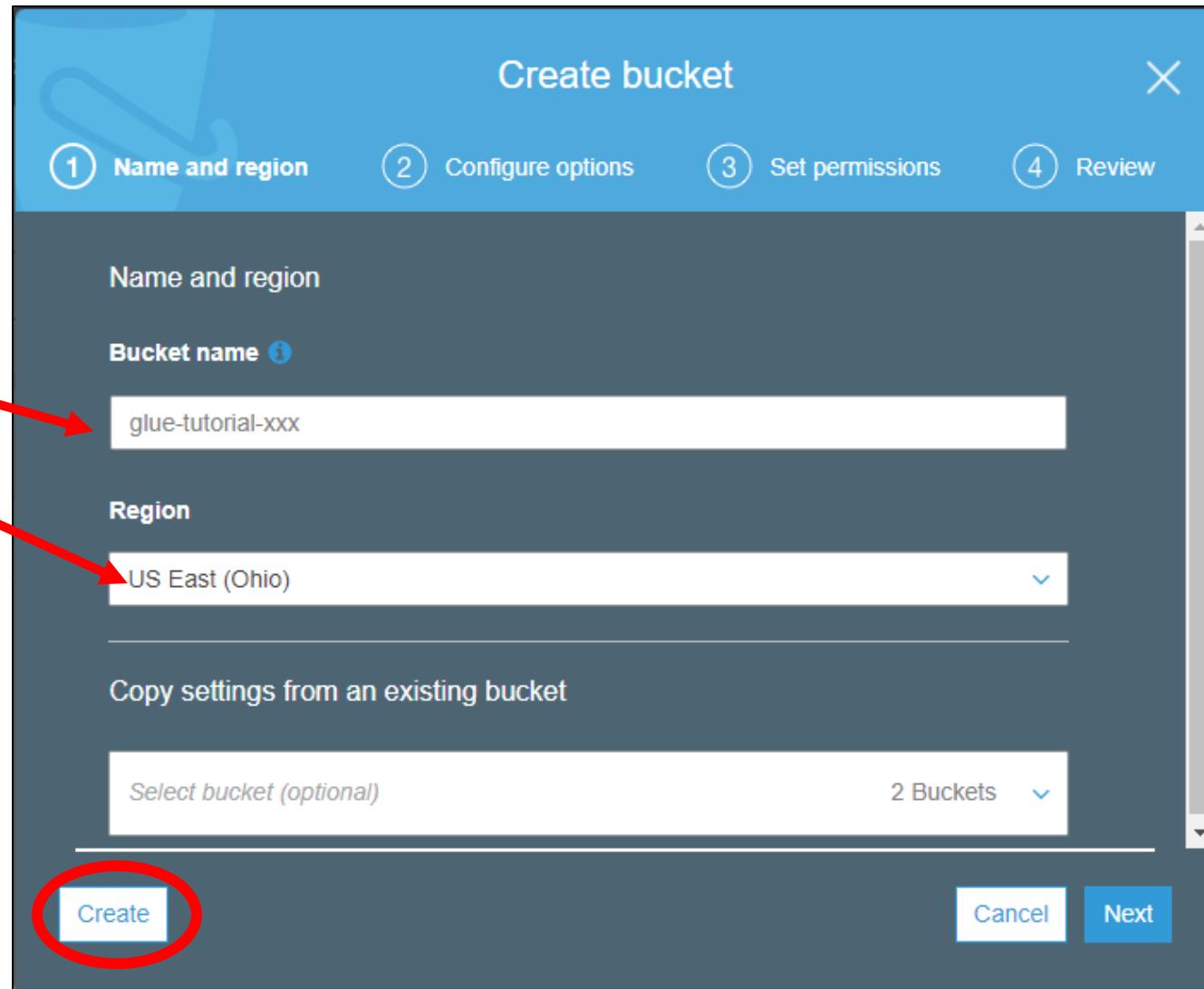




Create S3 bucket with AWS Console

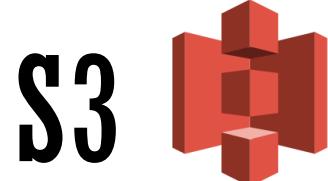
Give your S3
bucket a name
Use glue-tutorial-
XXX

Specify the region



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





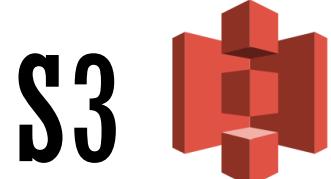
└ Create S3 bucket with AWS CLI*

(Alternative)

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

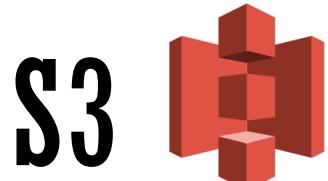
* Must install and set up AWS CLI in order to use this





Create S3 bucket with AWS Console

A screenshot of the AWS S3 console. At the top left is the 'Amazon S3' logo. To its right is a blue 'Discover' button. Below the logo is a search bar with the placeholder 'Search for buckets'. Underneath the search bar are three buttons: a blue '+ Create bucket' button, and two grey 'Delete bucket' and 'Empty bucket' buttons. To the right of these buttons, the text '1 Buckets 0 Public' is displayed. The main content area shows a table with one row. The first column contains a small blue bucket icon followed by the bucket name 'glue-tutorial-xxx'. The second column, 'Access', shows 'Not public *'. The third column, 'Region', shows 'US East (Ohio)'. A red oval highlights the 'glue-tutorial-xxx' bucket name.



Create S3 bucket folder

Amazon S3 > glue-tutorial-xxx

Overview Properties Permissions

Upload **+ Create folder** More ▾

Name ↑ Last modified

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:

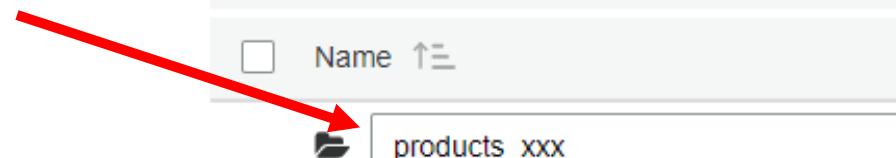
None (Use bucket settings)
 AES-256

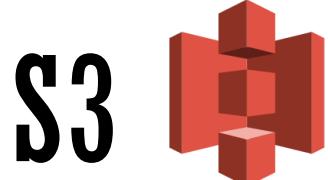
Use Server-Side Encryption with Amazon S3-Managed Keys (SSE-S3)

AWS-KMS

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

Save Cancel





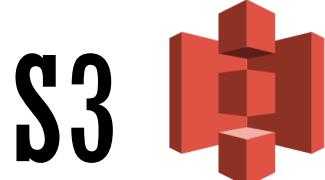
Create S3 bucket with AWS Console

The screenshot shows the AWS S3 console interface. At the top, there are three buttons: 'Upload', 'Create folder', and 'More'. Below these buttons is a search bar. The main area displays a list of objects. The first object is a folder named 'products_XXX', which is circled in red. The columns in the list are 'Name', 'Last modified', and 'Size'.

Name	Last modified	Size
products_XXX	--	--

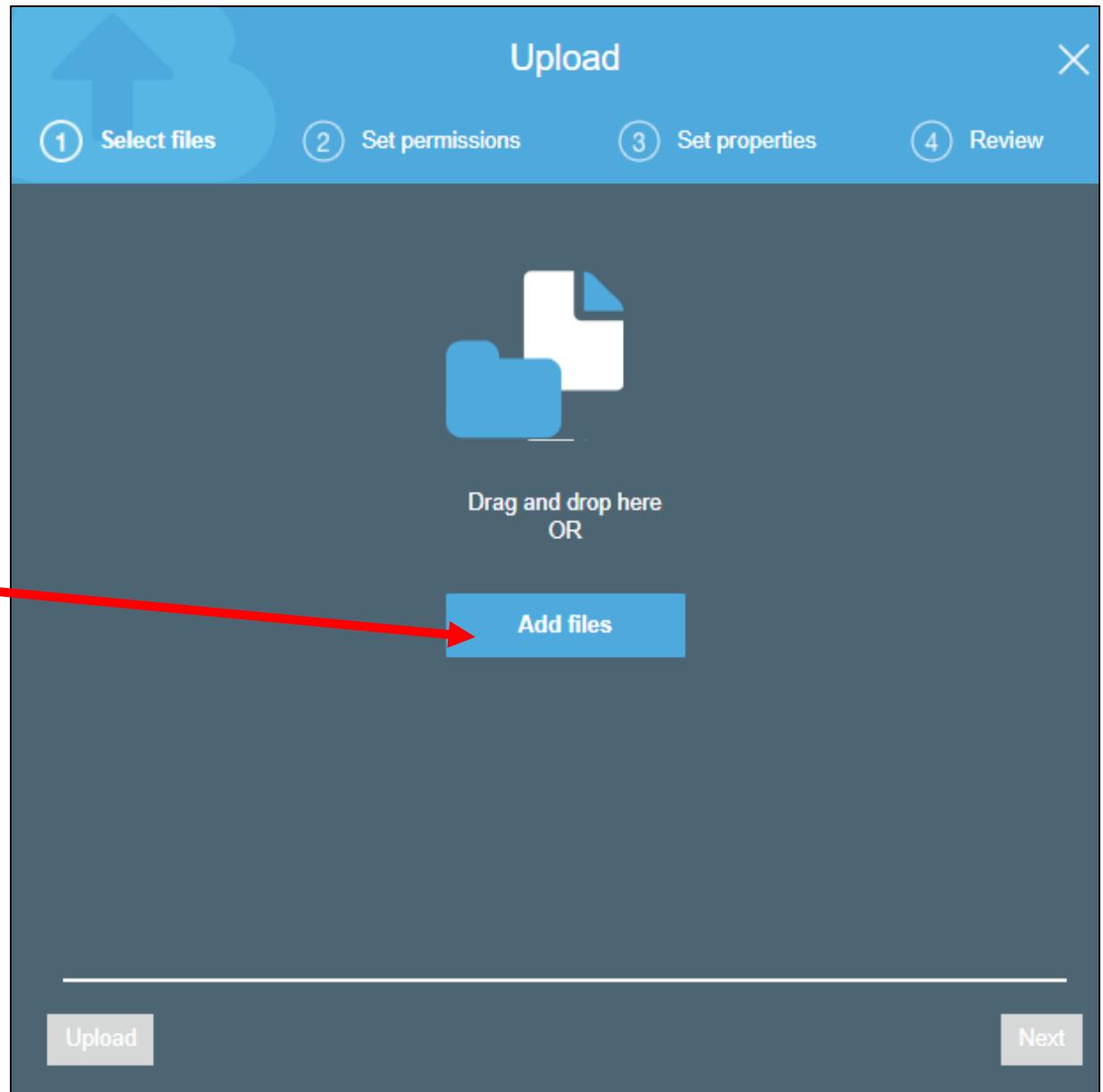
The screenshot shows the AWS S3 console interface for a specific bucket named 'glue-tutorial-XXX'. It includes a breadcrumb navigation path: 'Amazon S3 > glue-tutorial-XXX / products_XXX'. Below the path is an 'Overview' section. A search bar is present at the bottom of the page. At the very bottom, there are three buttons: 'Upload' (circled in red), 'Create folder', and 'More'.

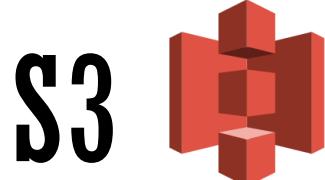




Add file to S3 bucket with AWS Console

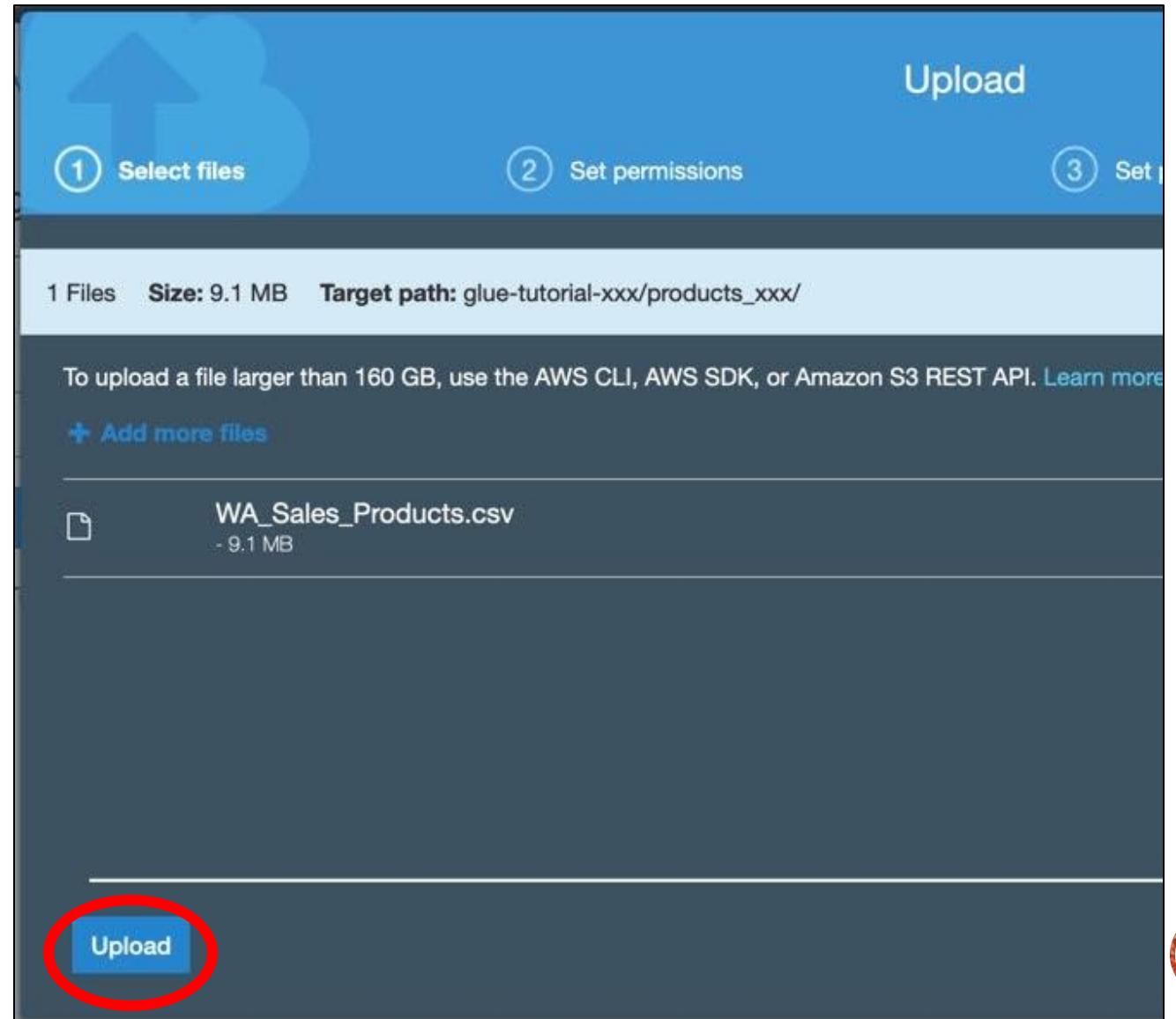
Add file from repository
called
“WA_Sales_Products.csv”

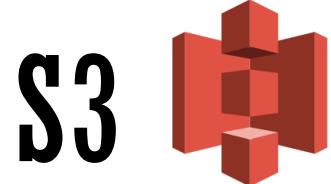




Add file to S3 bucket with AWS Console

Add file from repository
called
“WA_Sales_Products.csv”





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

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

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



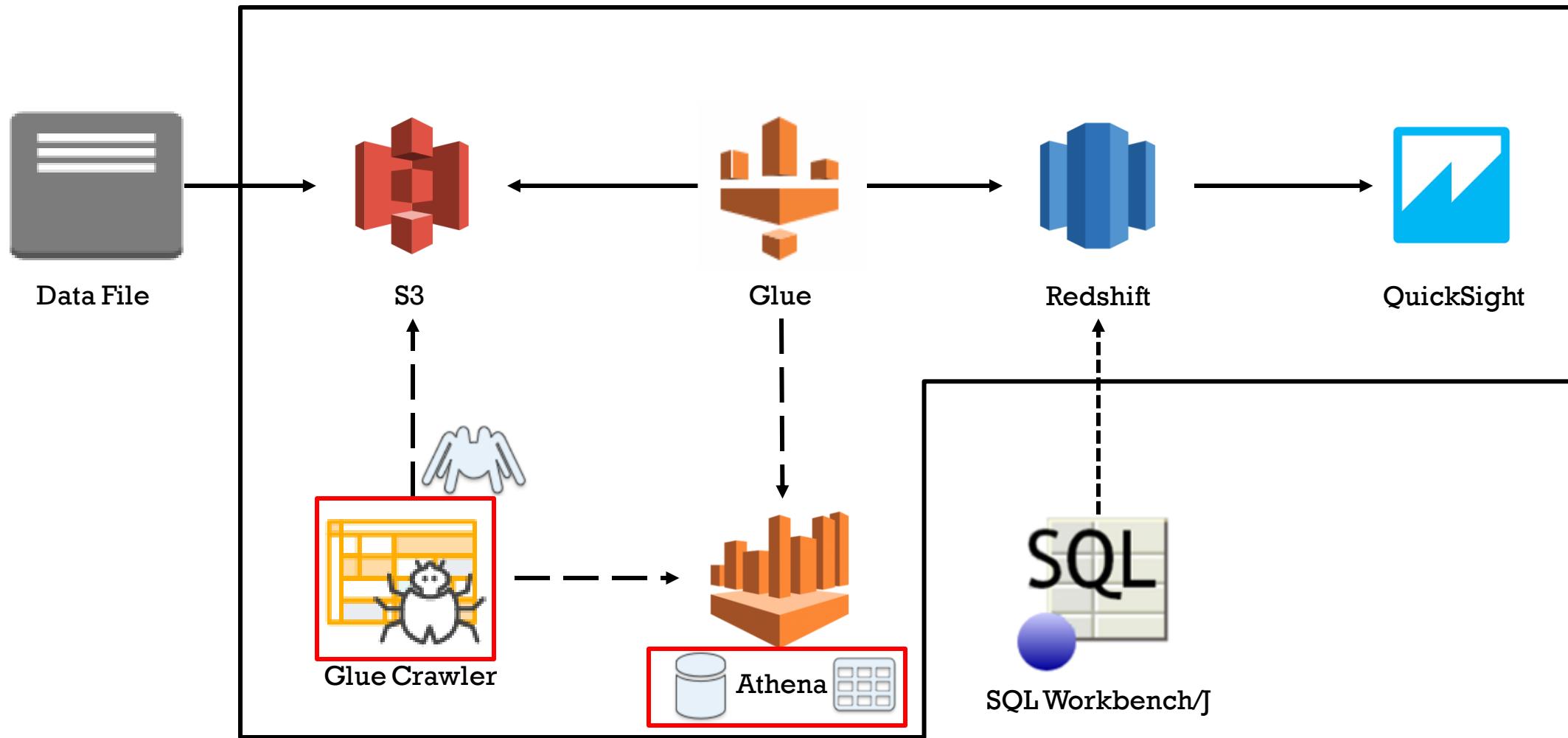
Lab 2

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

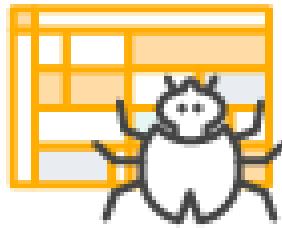
(Use US-EAST-2/Ohio Region)



Glue Crawler



Glue Crawler



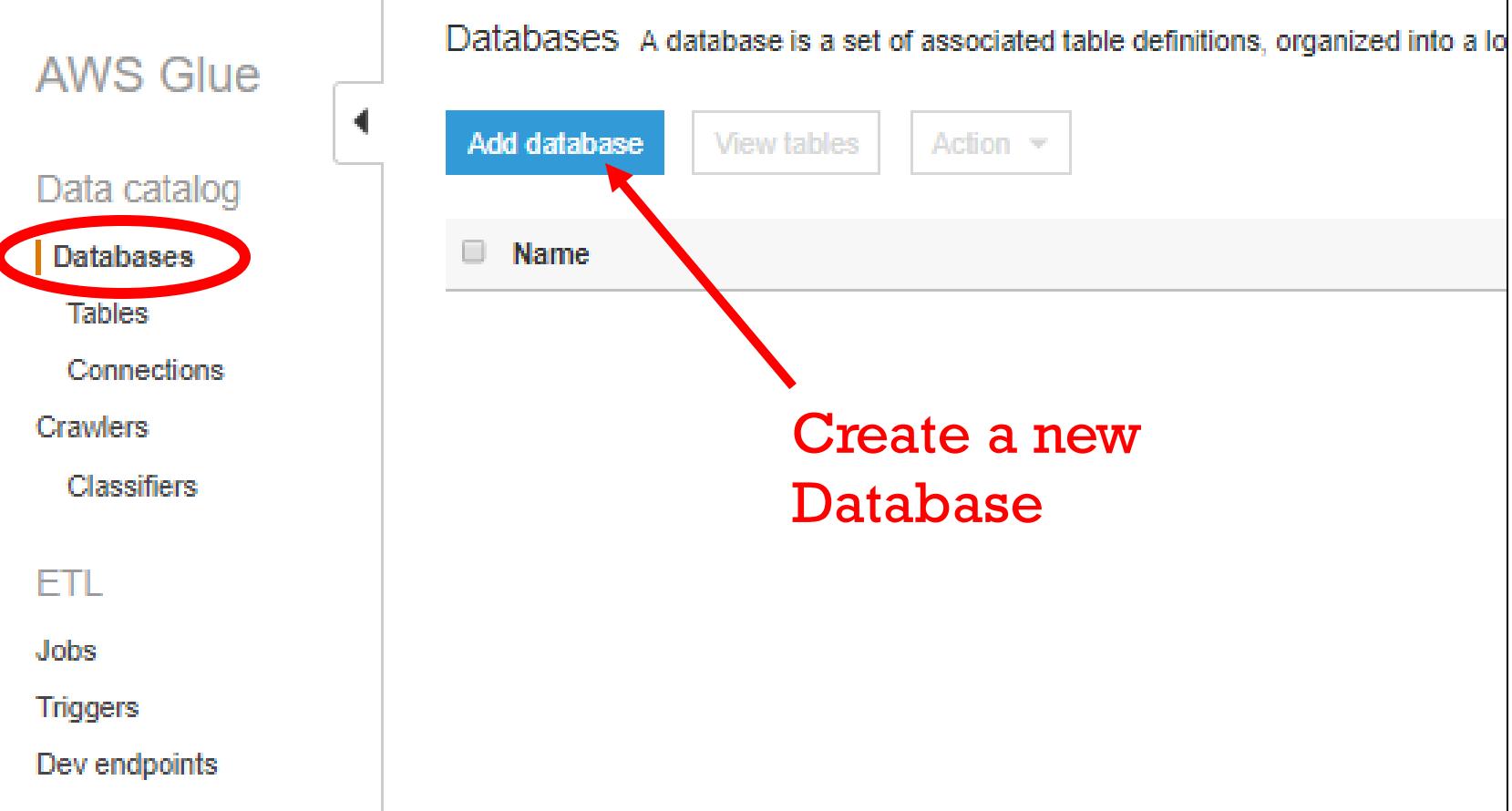
- Scans file data to create metadata
- Determines column names and data types
- Creates a Glue Table



Glue



Create Glue Database



The screenshot shows the AWS Glue Data Catalog interface. On the left, there is a sidebar with the following navigation options:

- AWS Glue
- Data catalog
- Databases (this option is highlighted with a red oval)
- Tables
- Connections
- Crawlers
- Classifiers
- ETL
- Jobs
- Triggers
- Dev endpoints

The main content area is titled "Databases" with a sub-instruction: "A database is a set of associated table definitions, organized into a logical schema." Below this, there are three buttons: "Add database" (highlighted with a blue box), "View tables", and "Action". A red arrow points from the "Add database" button to the text "Create a new Database".





Give your database a name
“glue_database_XXX”

Add database

Database name

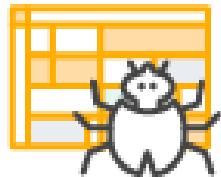
glue_database_xxx

Description and location (optional)

Create

A screenshot of a "Add database" form. The title "Add database" is at the top center. Below it is a "Database name" label with a text input field containing "glue_database_xxx". Underneath the input field is a link labeled "Description and location (optional)". At the bottom right is a blue "Create" button, which is circled in red. A red arrow points from the text "glue_database_XXX" in the first slide to the "Database name" input field in this screenshot.

Glue Crawler



Create Table with Glue Crawler

AWS Glue

- Data catalog
- Databases
- Tables** (highlighted with a red oval)
- Connections
- Crawlers
- Classifiers
- ETL
- Jobs
- Triggers
- Dev endpoints

Add tables ▾ Action ▾ Filter by attributes or search by keyword

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

Add tables ▾ Action ▾ Filter by attributes or search by keyword

Add tables using a crawler

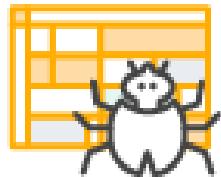
Add table manually

Database	Location	Classification
sampledb	s3://athena-examples-us-east-1/elb/plain/text	Unknown

Click on add tables to create a table

Create a table using a crawler

Glue Crawler



└ Create Table with Glue Crawler

Give your crawler a name,
glue-tutorial-XXX

Add information about your crawler

Crawler name

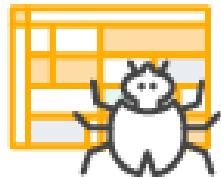
A red arrow points from the text "glue-tutorial-XXX" in the previous slide to this input field.

► Tags, description, security configuration, and classifiers
(optional)

Next

A red oval highlights the "Next" button at the bottom of the form.

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-xoo/products_xoo

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.

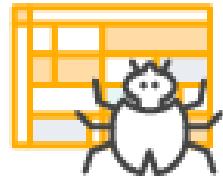
Exclude patterns (optional)

Back Next

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

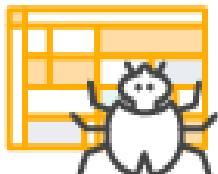
Add another data store

Yes
 No

[Back](#) [Next](#)



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-

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`

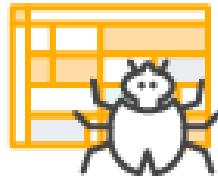
You can also create an IAM role on the [IAM console](#).

[Back](#) [Next](#)

Give your
role a name



Glue Crawler



Create Table with Glue Crawler

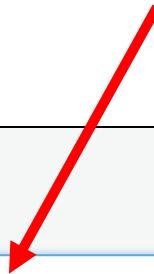
Your crawler can run on
either a timed schedule
or on demand

Create a schedule for this crawler

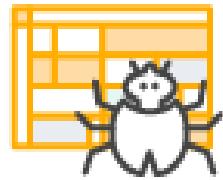
Frequency

Run on demand

Back **Next**



Glue Crawler



Create Table with Glue Crawler

Choose the database you created for your table to be in

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

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

Configure the crawler's output

Database i
glue_database_xxx

Add database

Prefix added to tables (optional) i
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. i

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

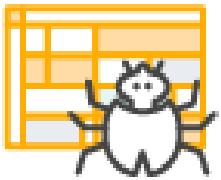
How should AWS Glue handle deleted objects in the data store?

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. i

Back **Next**



Glue Crawler



Create Table with Glue Crawler

Crawler info

Name: glue_tutorial_xxx
Tags: -

IAM role

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

Schedule

Schedule: Run on demand

Output

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

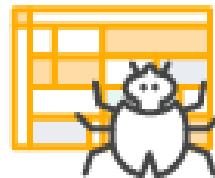
▼ Configuration options

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

[Back](#) [Finish](#)



Glue Crawler



Create Table with Glue Crawler

Run your crawler

Crawler glueTutorialxxx was created to run on demand. [Run it now?](#)

Add crawler Run crawler Action ▾ Filter by tags and attributes User preferences Showing: 1 - 1

Name	Schedule	Status	Logs	Last runtime	Median runtime	Tables updated	Tables added
glueTutorialxxx		Ready		0 secs	0 secs	0	0

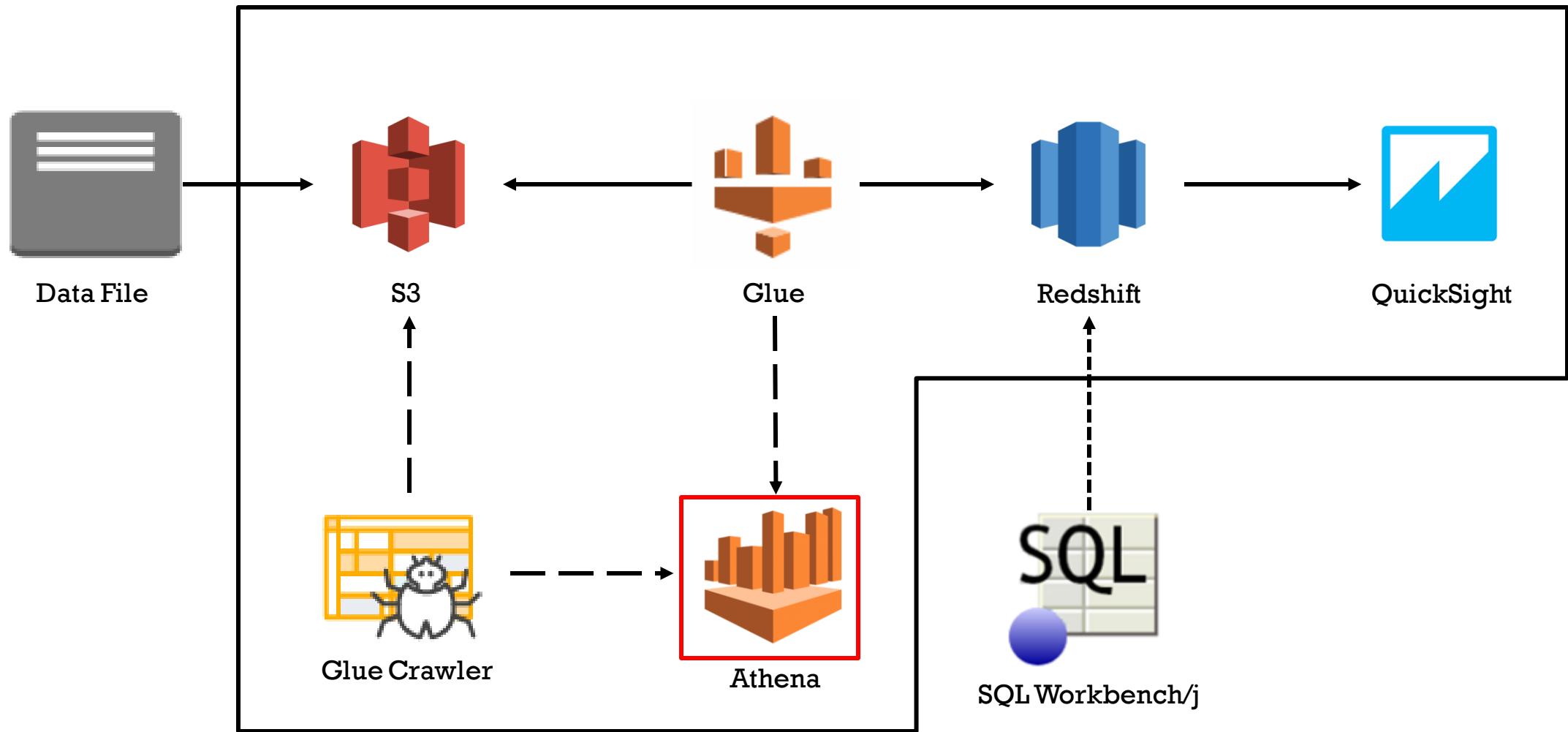
Tables A table is the metadata definition that represents your data, including its schema. A table can be used as a source or target in a job definition.

Add tables Action Filter by attributes or search by keyword

Name	Database	Location	Classification
products_xxx	glueDatabasexxx	s3://glue-tutorial-xxx/products_xxx/	csv

Your table should be in the Tables tab

Athena



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;**

Database

glue_database_xxx

Filter tables and views...

Tables (1)

products_xxx

Views (0)

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

New query 1

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

Run query Save as Create view from query (Run time: 1.44 seconds, Data scanned: 298.47KB) Format query

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	Tent	One Person Tent

Athena



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

The screenshot shows the Amazon Athena console interface. On the left, the 'Database' sidebar is set to 'glue_database_xxx'. Below it, under 'Tables (1)', 'products_xxx' is listed. Under 'Views (0)', there is a note: 'You have not created any views. To create a view, run a query and click "Create view from query"'. In the main area, a 'New query 1' tab is active, containing the SQL query: `1 SELECT COUNT(*)
2 FROM products_xxx;`. Below the query are buttons for 'Run query', 'Save as', and 'Create view from query'. A status message indicates '(Run time: 1.71 seconds, Data scanned: 9.21MB)'. At the bottom, the 'Results' section displays the output: `_col0
1 88475`.



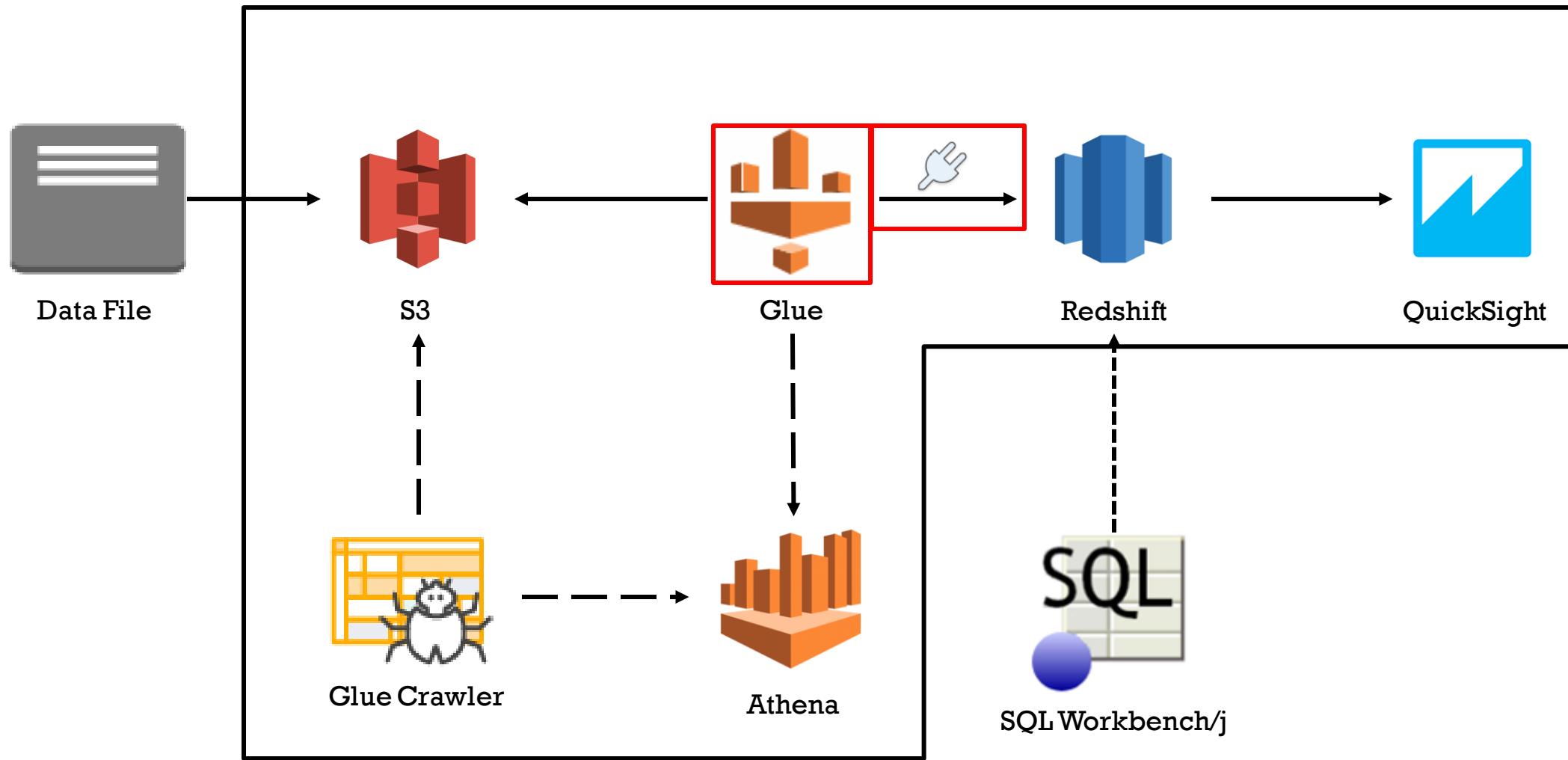
Lab 3

- Create/Run Glue Crawler
- Query using Athena

(Use US-EAST-2/Ohio Region)



Glue



VPC



Create a S3 endpoint

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

The screenshot shows the AWS VPC Dashboard. At the top right, there is a blue button labeled "Create Endpoint" which is circled in red. Below it is a search bar with the placeholder "Filter by attributes or search by keyword". On the left, there is a sidebar with several options: "Virtual Private Cloud", "Your VPCs", "Subnets", "Route Tables", "Internet Gateways", "Egress Only Internet Gateways", "DHCP Options Sets", "Elastic IPs", "Endpoints", "Endpoint Services", "NAT Gateways", and "Peering Connections". The "Endpoints" link is also circled in red.



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 [i](#)

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.elasticloadbal... a	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... a	amazon	Interface



VPC



Create a S3 endpoint

Choose VPC

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.

Route Table ID	Main	Associated With
rtb-35cf515d	<input checked="" type="checkbox"/>	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": "*"  
    }  
  ]  
}
```

[Cancel](#)[Create endpoint](#)



Create a connection to Redshift

AWS Glue

- Data catalog
- Databases
- Tables
- Connections** (This item is circled in red)
- Crawlers
- Classifiers
- Settings

ETL

- Jobs
- Triggers
- Dev endpoints

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

Add connection	Test connection	Action ▾
<input type="checkbox"/> Name	Type	Date created
		Last updated

You don't have any connections yet.

Add connection

Click on “Add Connection” to
create a connection to the
Redshift cluster





Create a connection to Redshift

Name of the connection:
glue-tutorial-XXX

Set up your connection's properties.

For more information, see [Working with Connections](#).

Connection name

Connection type

Require SSL connection
Fail if unable to connect over SSL

Description (optional)

Next

The connection type
should be Redshift





Create a connection to Redshift

Set up access to your data store.

For more information, see [Working with Connections](#).

Cluster

Database name

Username

Password

Back **Next**

The screenshot shows the first step of the 'Create a connection to Redshift' wizard. It asks for access to a data store. The 'Cluster' dropdown is set to 'glue-tutorial-xxx'. The 'Database name' field contains 'glueTutorialDatabase_XXX'. The 'Username' field contains 'master'. The 'Password' field contains '.....'. At the bottom, there are 'Back' and 'Next' buttons, with 'Next' being highlighted by a red circle.

Name of the cluster:
glue-tutorial-XXX

Name of the database:
glueTutorialDatabase_XXX

Username and
password created
for Redshift



Glue



Create a connection to Redshift

Connection properties

Name	glueTutorial_xxx
Type	Amazon Redshift
Require SSL connection	false
Description (optional)	-

Connection access

Cluster	glue-tutorial-xxx
Username	master
VPC Id	vpc-b2fb56da
Subnet	subnet-963d18db
Security groups	sg-797ba212

[Back](#) [Finish](#)





Test the connection to Redshift

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

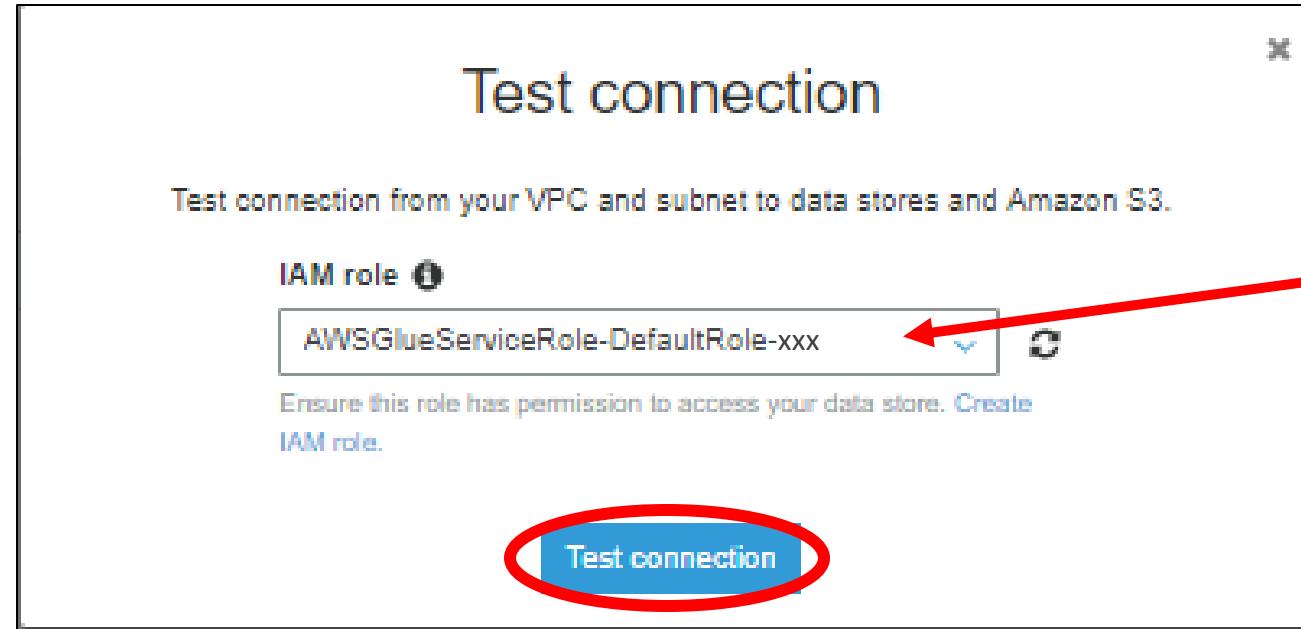
Add connection	Test connection	Action ▾	
Name	Type	Date created	Last updated
<input checked="" type="checkbox"/> glueTutorial_xxx	JDBC	22 August 2018 1:44 PM UTC-4	22 August 2018 1:44 PM UTC-4

Select newly created
connection





Test the connection to Redshift



Select your recently created IAM role





Test the connection to Redshift

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

glueTutorial_xxx connected successfully to your instance.

Add connection

Test connection

Action ▾

<input checked="" type="checkbox"/> Name	Type	Date created
glueTutorial_xxx	JDBC	22 August 2018 1:44 PM UTC-4



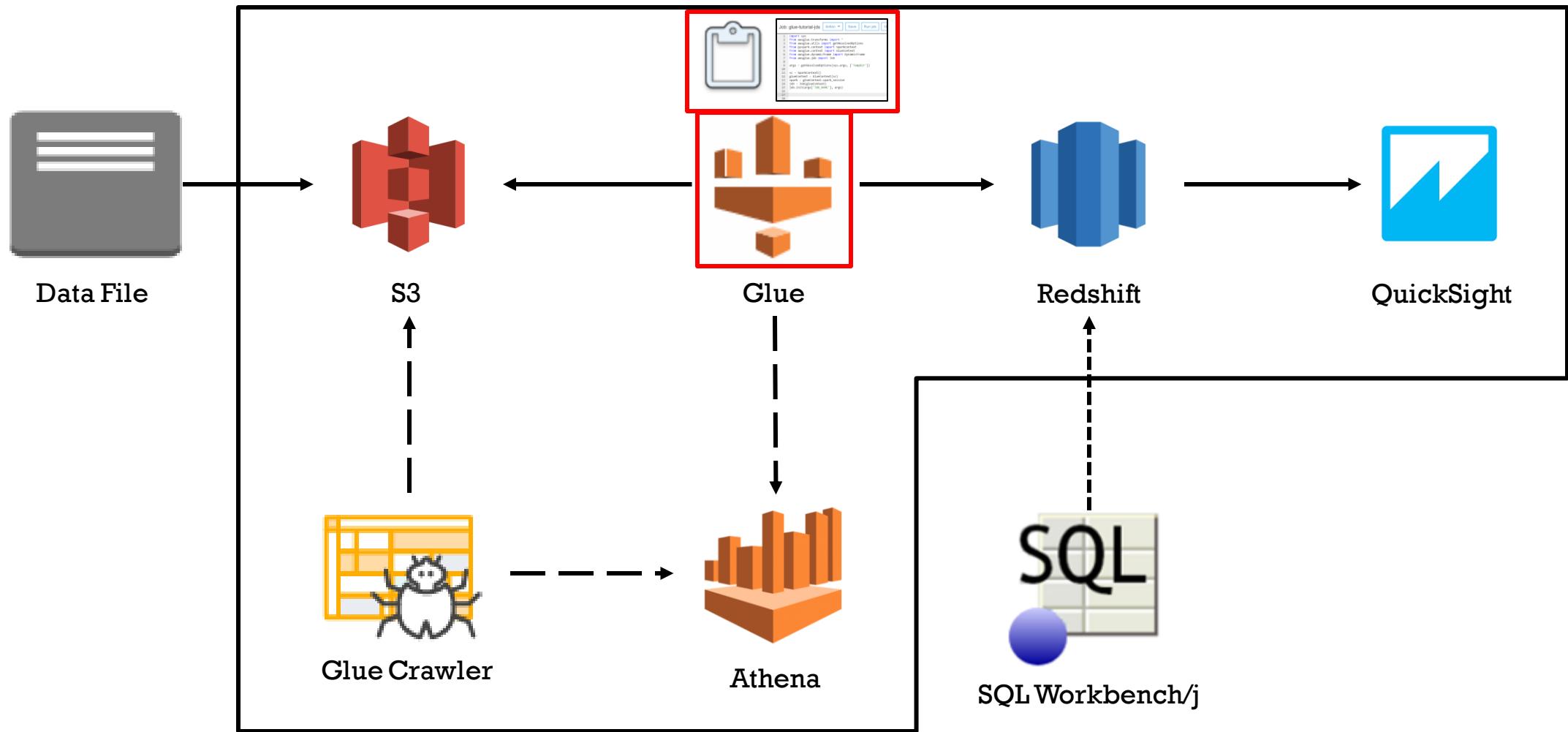
Lab 4

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

(Use US-EAST-2/ Ohio Region)



Glue



Glue



Create a Glue job

AWS Glue

Data catalog

Databases

- Tables
- Connections
- Crawlers
- Classifiers

ETL

- Jobs
- Triggers
- Dev endpoints

Add job Action Filter by attributes

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. Add job





The language used to write the script

Give your script a name
glueTutorial_XXX

The location where your script will be placed in S3

Give your job a name: glueTutorial_XXX

Configure the job properties

Name: glueTutorial_XXX

IAM role: AWSGlueServiceRole-DefaultRole-XXX

Type: Spark

Glue version: Spark 2.4, Python 3 (Glue version 1.0)

This job runs:

- A proposed script generated by AWS Glue
- An existing script that you provide
- A new script to be authored by you

Script file name: glueTutorial_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

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

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

▼ Security configuration, script libraries, and job parameters (optional)

Security configuration ⓘ

None

The security configuration specifies how the script is encrypted using server-side encryption with AWS KMS-managed keys (SSE-KMS) or Amazon S3-managed encryption keys (SSE-S3).

Server-side encryption

Python library path

s3://bucket/prefix/object

Dependent jars path

s3://bucket/prefix/object

Referenced files path

s3://bucket/prefix/object

Worker type ⓘ

Standard

Maximum capacity ⓘ

2

Max concurrency ⓘ

1

Job timeout (minutes) ⓘ

15

The default is 2,880 minutes (48 hours).

Delay notification threshold (minutes) ⓘ



Glue



Create a Glue job

Job parameters

Key	Value
--REDSHIFT_DB_NAME	glueTutorial_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	glueTutorial_xxx
Type key...	Type value...

Next

Parameterize values to
be used in the script

Parameters:

```
--REDSHIFT_DB_NAME
glueTutorial_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
glueTutorial_xxx
```



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 < >

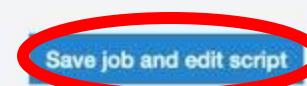
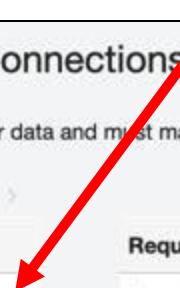
All connections	Required connections
glueTutorial_xxx	glueTutorial_xxx

Select

Add connection

Back

Save job and edit script



Glue



Writing the Script

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

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

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

Include SQL functions, types, and datetime to use later

Add the parameters that were passed into the Glue job

Glue



Writing the Script

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

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

The data will be written to the datasource as a DynamicFrame

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



Glue



Writing the Script

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

sourcedata needs to be set to a Data Frame

```
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")
```

Convert back to a Dynamic Frame

This is where the transformations happen



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



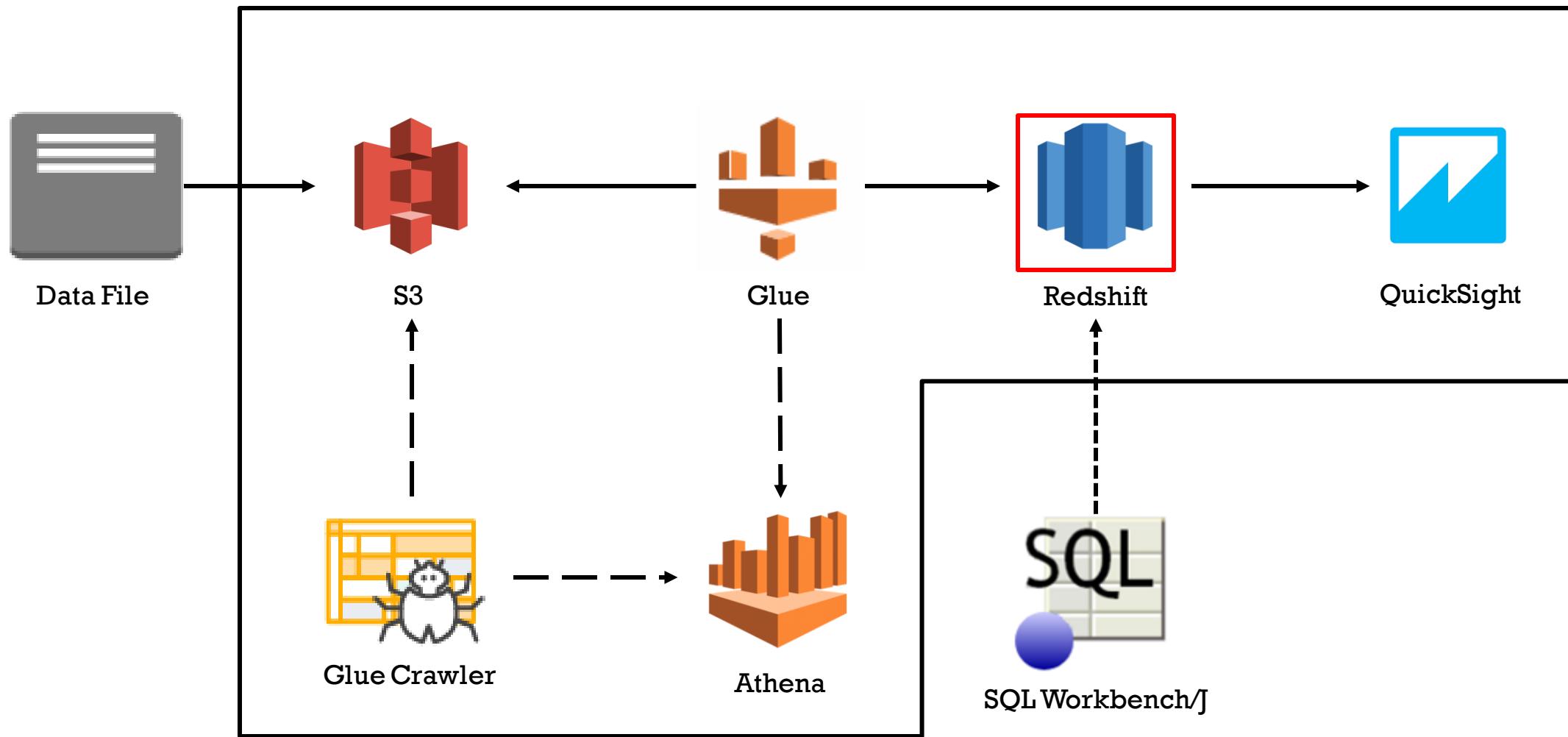
Writing the Script

```
...  
# 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



Redshift



Create table

Copy the SQL script from the repository into SQL Workbench

SQL Workbench/J GlueTutorial - Default.wksp

File Edit View Data SQL Macros Workspace Tool

Statement 1 Database Explorer 2

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

Add your own initials to the schema and table names

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

SQL Workbench/J GlueTutorial - Default.wksp

File Edit View Data SQL Macros Workspace Tools Help

Statement 1 Database Explorer 2

```
1 SELECT * FROM sales_XXX.products_XXX LIMIT 50;
2 |
```

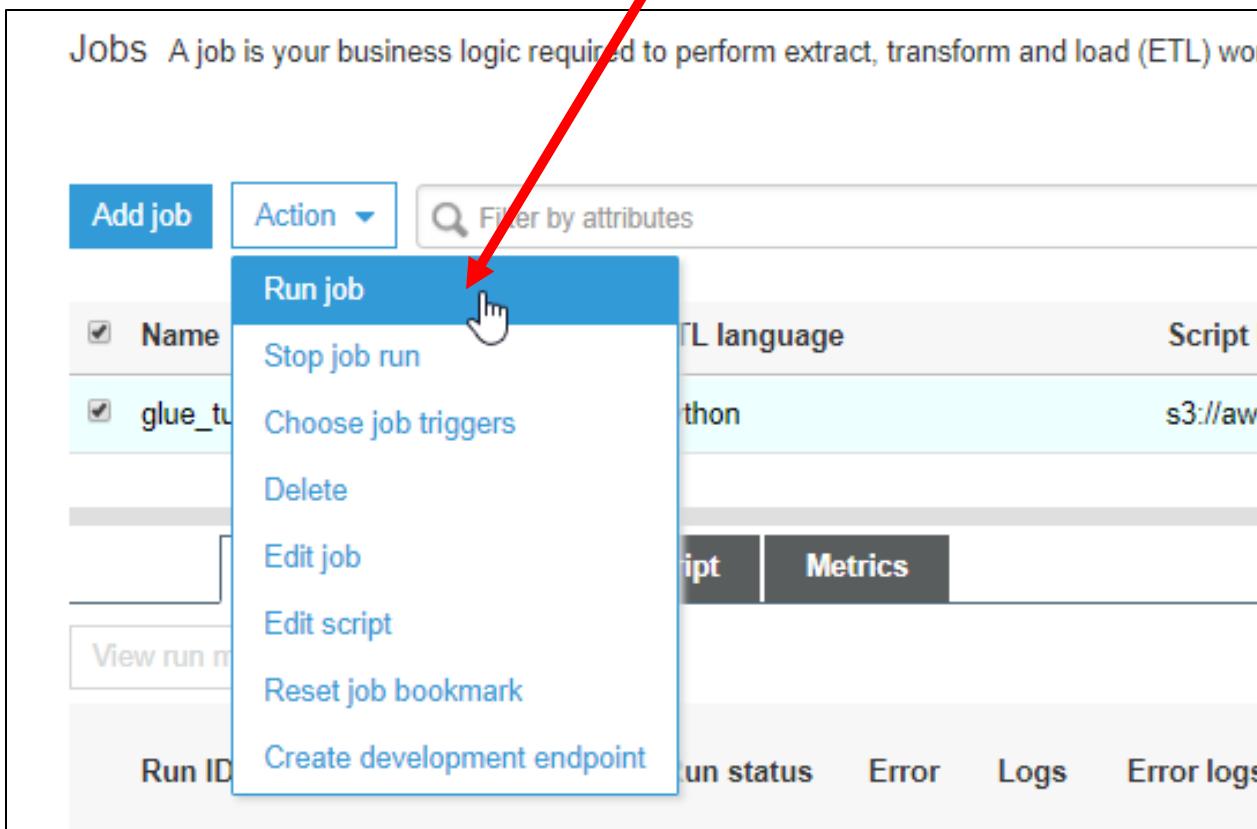


Glue



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) work.

Add job Action ▾ Filter by attributes

Name glue_tutorial_XXX

Run job 

Stop job run

Choose job triggers

Delete

Edit job

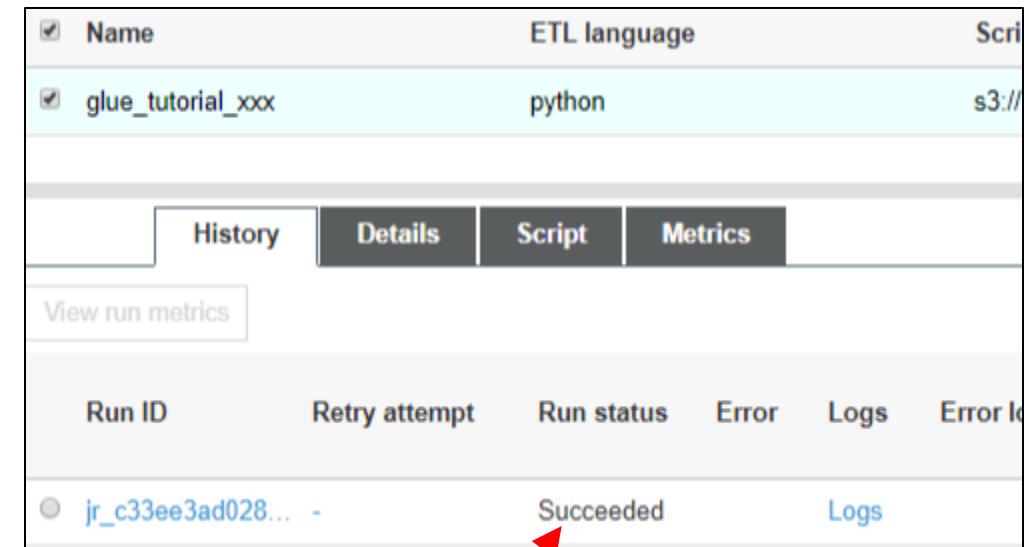
Edit script

Reset job bookmark

Create development endpoint

Run ID

Run status Error Logs Error logs



Name	ETL language	Script location			
glue_tutorial_XXX	python	s3://awslabs/amazon-glue-tutorials/glue-tutorial-XXX.py			
History					
Run ID	Retry attempt	Run status	Error	Logs	Error logs
jr_c33ee3ad028...	-	Succeeded		Logs	

When the job succeeds,
check your Redshift table



Redshift



Verify data in the table

```
1 SELECT *
2 FROM sales_redshift_schema_XXX.products_redshift_table_XXX LIMIT 100;
3
4
```

Result 1 Messages

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



Lab 5

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

(Use US-EAST-2/Ohio Region)



Enhancements

└ Improve the versatility of your Glue job

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



Glue Trigger



Automatically run Glue job using Lambda – a serverless function

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

Do this using a Lambda.

Name the lambda and choose Python as the language

Lambda > Functions

Functions (5)

Filter by tags and attributes or search by keyword

Function name	Description	Runtime	Code size	Last modified
---------------	-------------	---------	-----------	---------------

Create function

Basic information

Function name
Enter a name that describes the purpose of your function.
glue_job_trigger_xxx

Use only letters, numbers, hyphens, or underscores with no spaces.

Runtime [Info](#)
Choose the language to use to write your function.
Python 3.8

Permissions [Info](#)
Lambda will create an execution role with permission to upload logs to Amazon CloudWatch Logs. You can configure and modify permissions further when you add triggers.

▶ Choose or create an execution role

Glue Trigger



Automatically run Glue job using Lambda – a serverless function

▼ Choose or create an execution role

Execution role
Choose a role that defines the permissions of your function. To create a custom role, go to the [IAM console](#).

Create a new role with basic Lambda permissions
 Use an existing role
 Create a new role from AWS policy templates

i Role creation might take a few minutes. Please do not delete the role or edit the trust or permissions policies in this role.

Name the role and choose permissions for S3

Role name
Enter a name for your new role.

Use only letters, numbers, hyphens, or underscores with no spaces.

Policy templates - optional [Info](#)
Choose one or more policy templates.

Amazon S3 object read-only permissions X
S3

Glue Trigger



Automatically run Glue job
using Lambda – a serverless function

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

glue_job_trigger_xxx

Throttle Qualifiers Actions

Configuration Permissions Monitoring

Designer

+ Add trigger

glue_job_trigger_xxx

Layers (0)

Add trigger

Trigger configuration

S3 aws storage

Bucket
Please select the S3 bucket that serves as the event source. The bucket must be in the same region as the function.
glue-tutorial-xxx

Event type
Select the events that you want to have trigger the Lambda function. You can optionally set up a prefix or suffix for an event. However, for each bucket, individual events cannot have multiple configurations with overlapping prefixes or suffixes that could match the same object key.
All object create events

Prefix
Enter a single optional prefix to limit the notifications to objects with keys that start with matching characters.
products_xxx/

Suffix
Enter a single optional suffix to limit the notifications to objects with keys that end with matching characters.
e.g. .jpg

Lambda will add the necessary permissions for Amazon S3 to invoke your Lambda function from this trigger. [Learn more](#) about the Lambda permissions model.

Enable trigger
Enable the trigger now, or create it in a disabled state for testing (recommended).

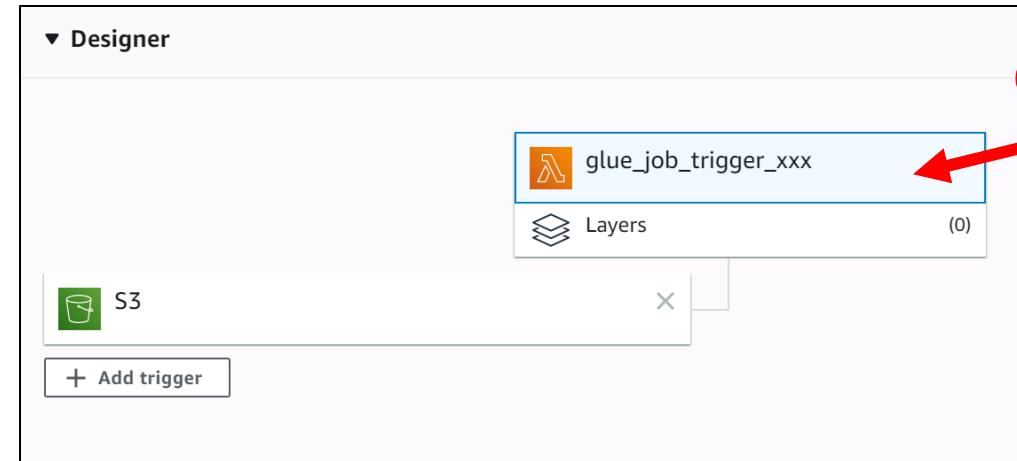
Cancel Add

Glue Trigger



Automatically run Glue job
using Lambda – a serverless function

Then make the Lambda run
the Glue job.



Add code
under
'lambda_
function'

The screenshot shows the configuration page for the Lambda function 'glue_job_trigger_xxx'. At the top, there are tabs for 'Throttle', 'Qualifiers', 'Actions', 'Select a test event', 'Test', and a prominent orange 'Save' button which is circled in red. Below these tabs, there are sections for 'Code entry type' (set to 'Edit code inline'), 'Runtime' (set to 'Python 3.8'), and 'Handler' (set to 'lambda_function.lambda_handler'). The main area is a code editor titled 'lambda_function'. It contains Python code that imports boto3 and defines a lambda_handler function. An arrow points from the red 'Add code under "lambda_function"' annotation to the code editor. The code is as follows:

```
import boto3

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

```
import boto3

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



Glue Trigger



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

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

The screenshot shows the 'Create table' wizard in the Amazon DynamoDB console. On the left, a sidebar lists options like 'Dashboard', 'Tables', 'Backups', 'Reserved capacity', 'Preferences', 'DAX', 'Clusters', 'Subnet groups', 'Parameter groups', and 'Events'. The main area is titled 'Create table' and contains a brief introduction to DynamoDB. A large red circle highlights the 'Create table' button. Below it, the 'Create DynamoDB table' step is shown. It asks for a 'Table name*' (set to 'glue_triggers') and a 'Primary key*' (set to 'filename'). A red double-headed arrow points between the table name and the primary key fields. Under 'Table settings', there's a note about default settings and a checked checkbox for 'Use default settings'.



Glue Trigger



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

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

Create item

Tree ▾ ▲ ▼

▼ Item {2}

- + filename String : WA_Sales_Products
- + glue_job String : glueTutorial_xxx

glue_triggers Close

Overview Items Metrics Alarms Capacity Indexes Global Tab

Create item Actions ▾

Scan: [Table] glue_triggers: filename ^

Scan [Table] glue_triggers: filename

+ Add filter

Start search

	filename ⓘ	glue_job
<input type="checkbox"/>	WA_Sales_Products	glueTutorial_xxx



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

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

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

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

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



Glue Trigger

IAM Roles determine how a resource can interact with other services

Log output

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

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

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



Glue Trigger

IAM Roles determine how a resource can interact with other services

Identity and Access Management (IAM)

Create role Delete role

Role name Trusted entities

Role name	Trusted entities
AWSRXWLambdaFunction-role-5ni...	AWS service: lambda
lambda_jlz	AWS service: lambda
lambda_xxx	AWS service: lambda

Dashboard

Access management

Groups

Users

Roles (circled in red)

Policies

Permissions Trust relationships Access Advisor Revoke sessions

Permissions policies (1 policy applied)

Attach policies

Policy name Policy type

AWSLambdaBasicExecutionRole-5812fc8e-f707-4957-96eb-47ec0eacd883 Managed policy

Attach Permissions

Create policy

Filter policies

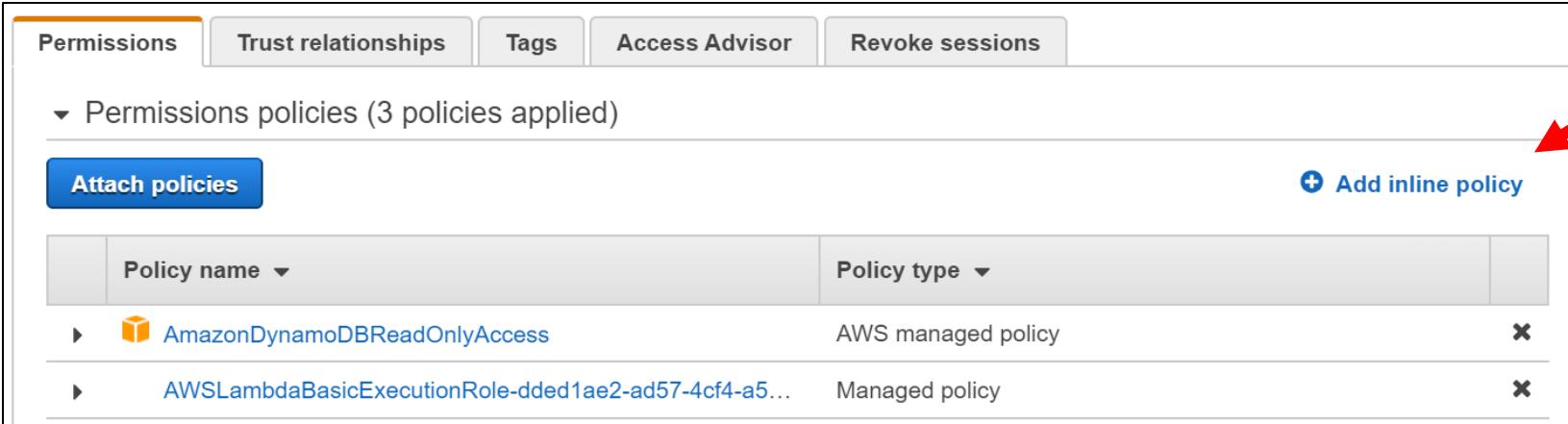
Policy name	Type	Used as
AmazonDynamoDBRe...	AWS managed	None

AmazonDynamoDBReadOnlyAccess

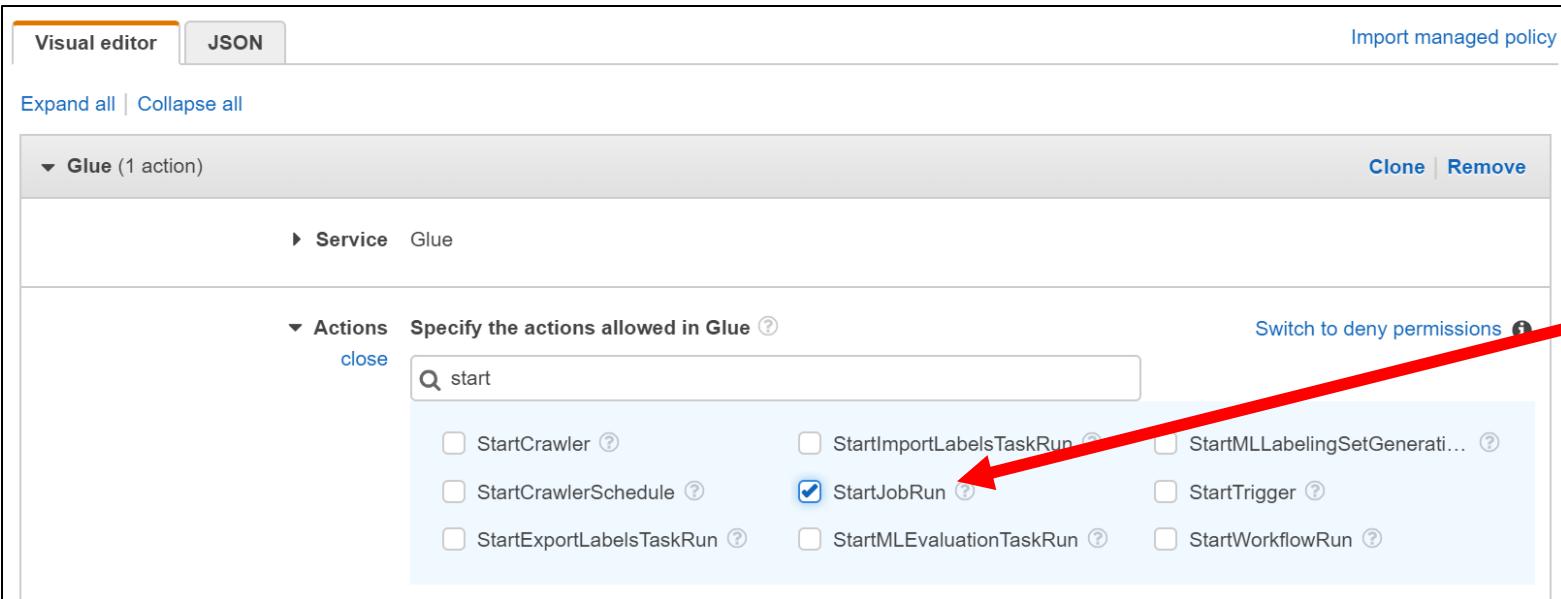
Provides read only access to Amazon DynamoDB via the AWS Management Console.

Glue Trigger

IAM Roles determine how a resource can interact with other services



The screenshot shows the 'Permissions' tab of an IAM role configuration. It displays three applied policies: 'AmazonDynamoDBReadOnlyAccess' (AWS managed policy) and 'AWSLambdaBasicExecutionRole-dded1ae2-ad57-4cf4-a5...' (Managed policy). A red arrow points from the top right towards the 'Add inline policy' button.



The screenshot shows the 'Visual editor' tab of a policy's actions section. Under the 'Glue' service, the 'Actions' section lists several options. A red arrow points from the bottom right towards the 'StartJobRun' checkbox, which is checked. Other visible actions include StartCrawler, StartImportLabelsTaskRun, StartMLLabelingSetGenerati..., StartTrigger, StartWorkflowRun, StartExportLabelsTaskRun, StartMLEvaluationTaskRun, and StartCrawlerSchedule.



Glue Trigger



Test the Lambda

Type a prefix and press Enter to search. Press ESC to clear.

Upload Create folder Download Actions ▾

Name WA_Sales_Products.csv

Add tags Make public Rename Delete Undo delete Cut Copy

A screenshot of the AWS Glue Data Catalog interface. On the left, there's a search bar and three buttons: 'Upload', 'Create folder', and 'Download'. Below these are two filter dropdowns: 'Name' and 'ETL language'. A blue box highlights the 'Name' dropdown, which contains the text 'WA_Sales_Products.csv'. To the right of the filters is a 'Actions' dropdown menu with several options: 'Add tags', 'Make public', 'Rename' (which is highlighted with a red arrow), 'Delete', 'Undo delete', 'Cut', and 'Copy'. A red diagonal line connects the 'Rename' option in the catalog to the 'Metrics' tab in the adjacent screenshot.

Name	ETL language	Script			
glueTutorialxxx	python	s3://			
History	Details	Script	Metrics		
View run metrics					
Run ID	Retry attempt	Run status	Error	Logs	Error logs
jr_c33ee3ad028...	-	Running	x	Logs	

A screenshot of the AWS Glue job details page. At the top, there are tabs for 'History', 'Details', 'Script', and 'Metrics', with 'Metrics' being the active tab (indicated by a red arrow). Below the tabs is a section titled 'View run metrics'. Further down, there's a table with columns: 'Run ID', 'Retry attempt', 'Run status', 'Error', 'Logs', and 'Error logs'. A single row is shown with the Run ID 'jr_c33ee3ad028...', a Retry attempt of '-', a Run status of 'Running', an Error of 'x', and Logs and Error logs links.



Lab 6

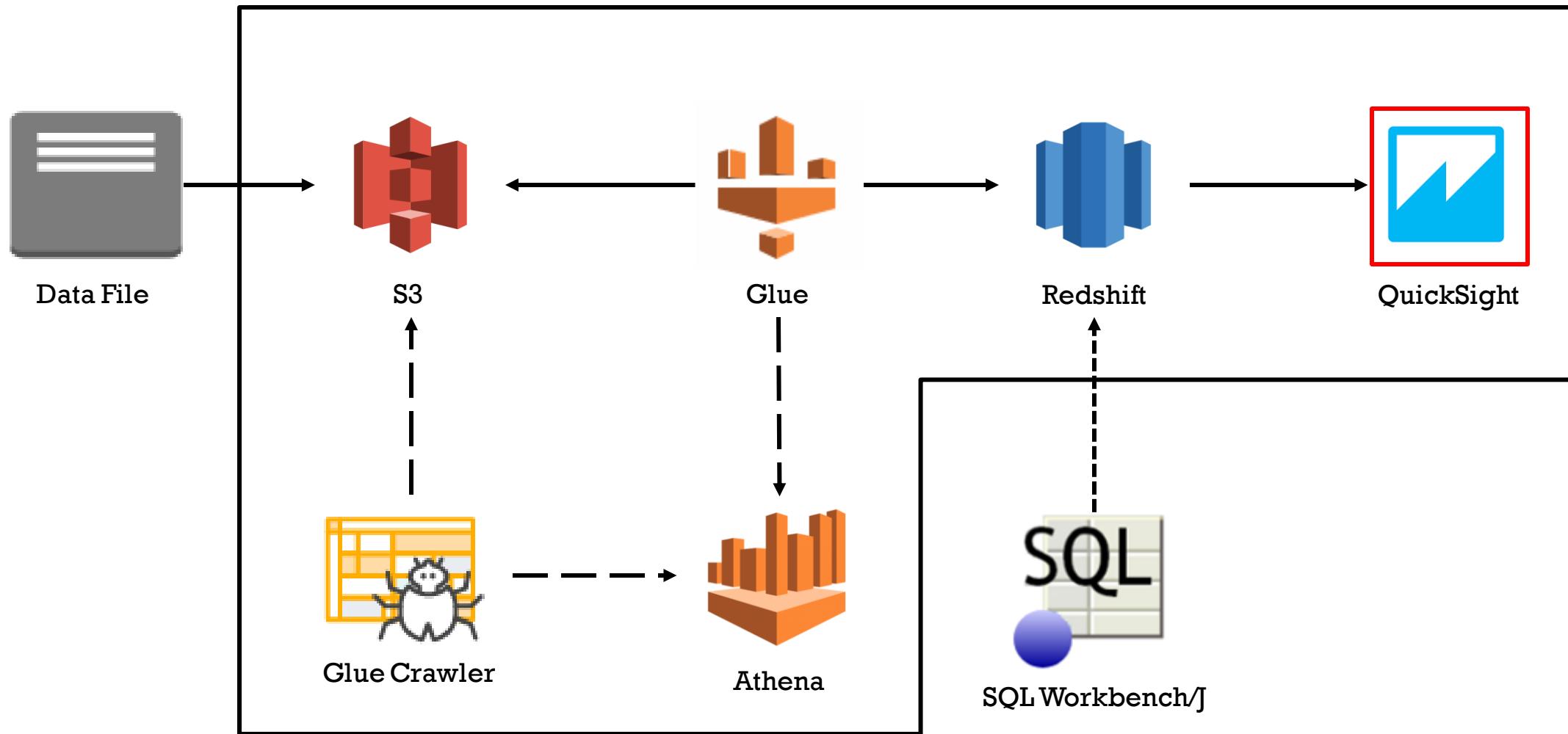
- Create Lambda
- Create DynamoDB
- Edit IAM Role
- Test Lambda

(Use US-EAST-2/Ohio Region)



QUICKSIGHT

AWS Business Intelligence Tool





AWS Business Intelligence Tool

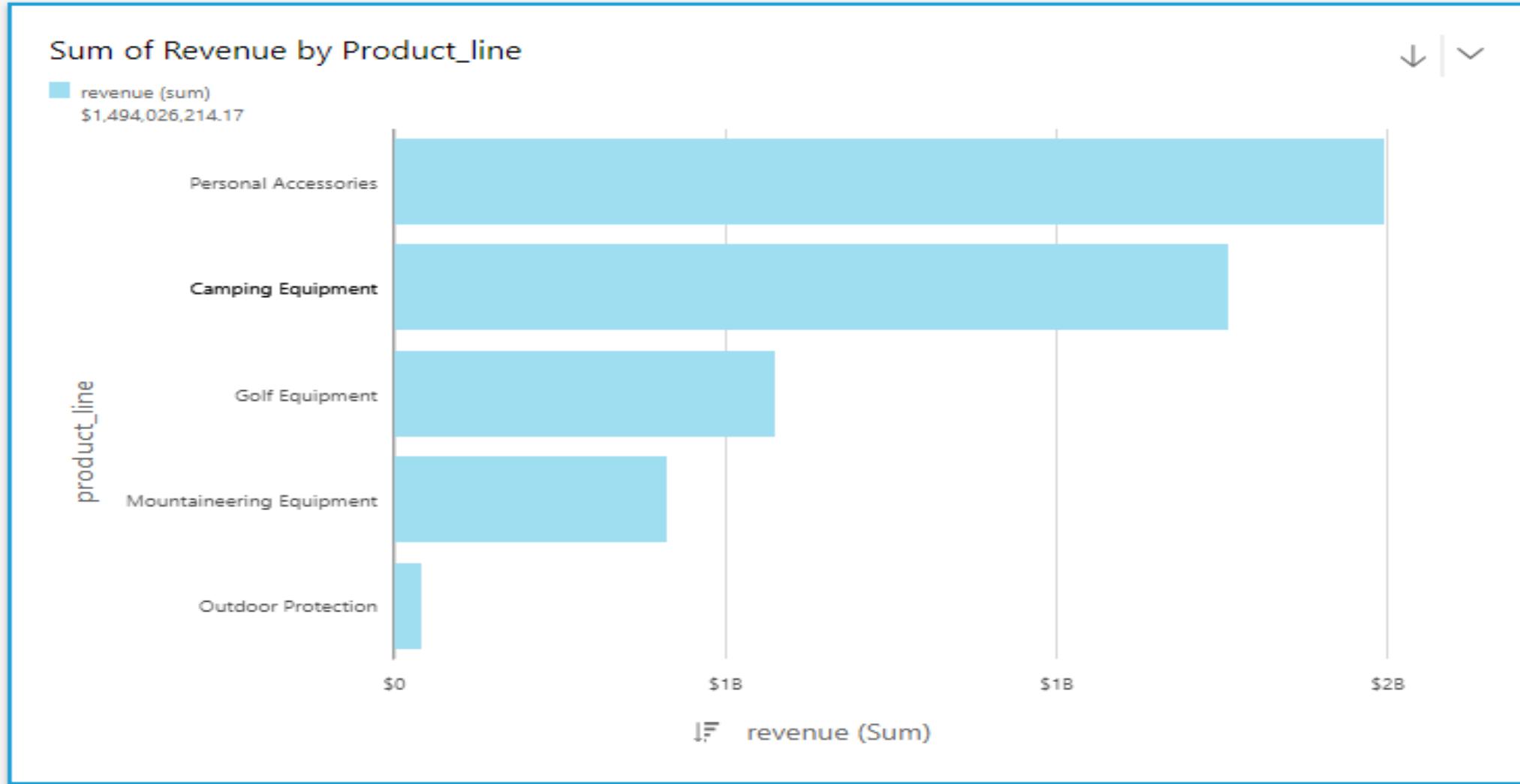
- Cloud based Business Intelligence reporting tool
- Build Reports from
 - Files in S3
 - Redshift
 - Athena



QUICKSIGHT



AWS Business Intelligence Tool





Create Analysis

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



QUICKSIGHT



AWS Business Intelligence Tool

Edit inbound rules

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

Add Rule

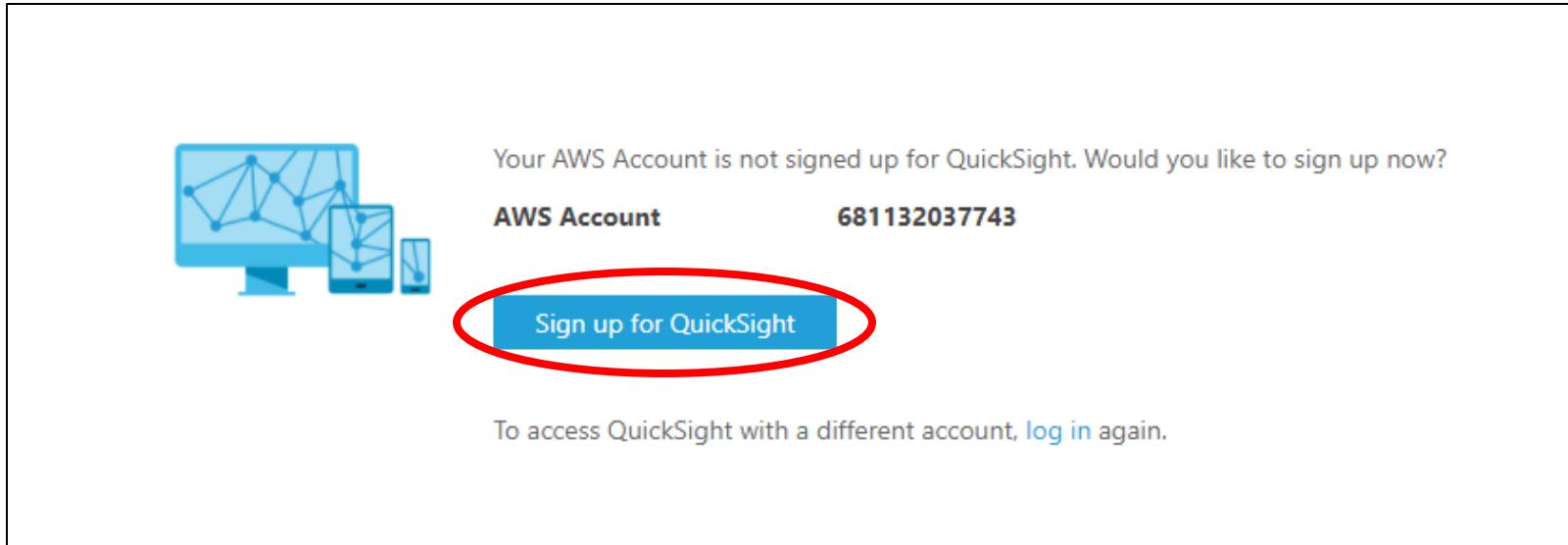
NOTE: Any edits made on existing rules will result in the edited rule being deleted and a new rule created with the new details. This will cause traffic that depends on that rule to be dropped for a very brief period of time until the new rule can be created.

Cancel **Save**

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
james-zhang i
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) ▼ i

> 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 Choose 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

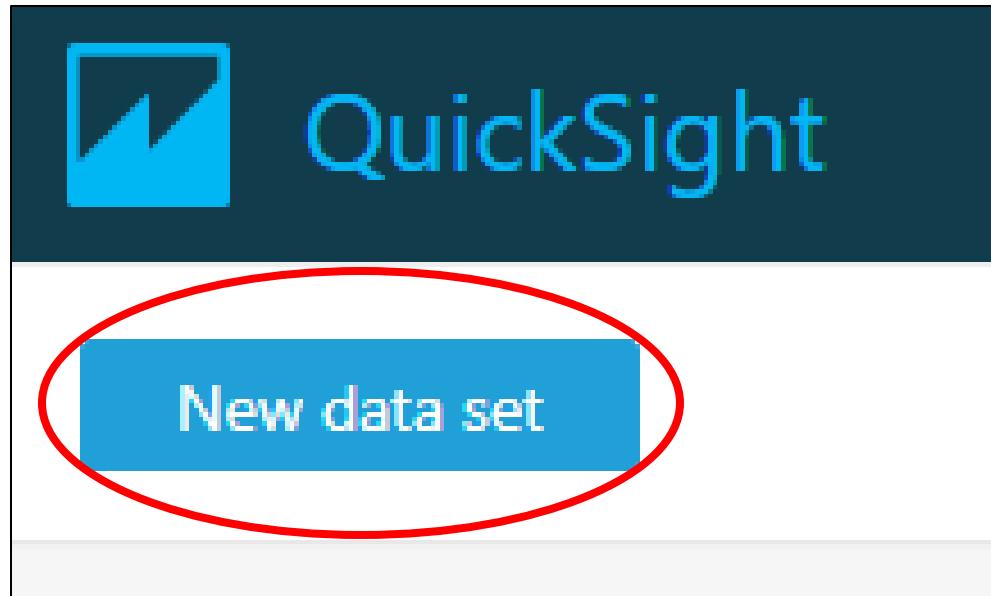


AWS Business Intelligence Tool



Manage data





QUICKSIGHT



AWS Business Intelligence Tool

Create a Data Set

FROM NEW DATA SOURCES

 Upload a file .csv, .tsv, .clf, .elf, .xlsx, .json	 Salesforce Connect to Salesforce	 S3 Analytics	 S3
 Athena	 RDS	 Redshift Auto-discovered	 Redshift Manual connect
 MySQL	 PostgreSQL	 SQL Server	 Aurora



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 X

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: glueTutorial_database_xxx

Username: master

Password: [REDACTED]

Validate connection SSL is enabled Create data source

QUICKSIGHT



AWS Business Intelligence Tool

Choose your table X

[sales_XXX](#)

Schema: contain sets of tables.

[sales_redshift_schema_XXX](#) ▼

Tables: contain the data you can visualize.

[products_redshift_table_XXX](#)

Select

Edit/Preview data

Use custom SQL



QUICKSIGHT



AWS Business Intelligence Tool

Finish data set creation

X

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

Edit/Preview data

Visualize



QUICKSIGHT



AWS Business Intelligence Tool

Select Add > Add title

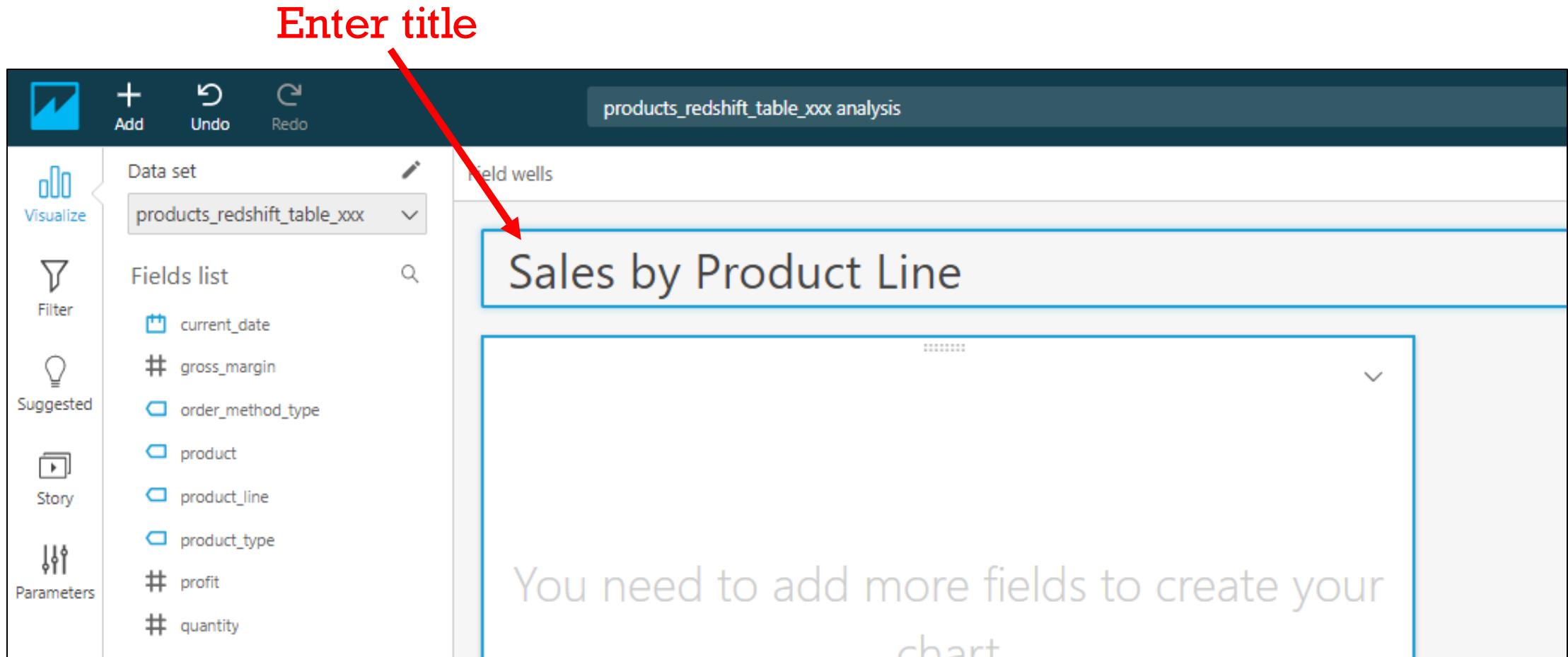
The screenshot shows the QuickSight interface. The top navigation bar includes a logo, a search bar, and a user profile icon. Below the search bar, there's a toolbar with icons for 'Add', 'Undo', and 'Redo'. A red arrow points to the 'Add' icon. A dropdown menu is open from this icon, with 'Add title' highlighted by another red arrow. The menu also contains 'Add visual', 'Add description', 'Add calculated field', and 'Add parameter'. To the left, a sidebar lists 'Visualize', 'Filter', 'Suggested', and 'Story' options. The main workspace is titled 'products_redshift_table_xxx analysis'. In the bottom right corner of the workspace, there's a section labeled 'AutoGraph' with the instruction 'Choose 1 or more fields and let QuickSight choose the most app'. At the bottom right of the slide, there's a small circular red icon.

QUICKSIGHT



AWS Business Intelligence Tool

Enter title



products_redshift_table_xxx analysis

Data set: products_redshift_table_xxx

Field wells

Sales by Product Line

You need to add more fields to create your chart

Visualize

Add Undo Redo

Fields list

- current_date
- # gross_margin
- order_method_type
- product
- product_line
- product_type
- # profit
- # quantity

Filter

Suggested

Story

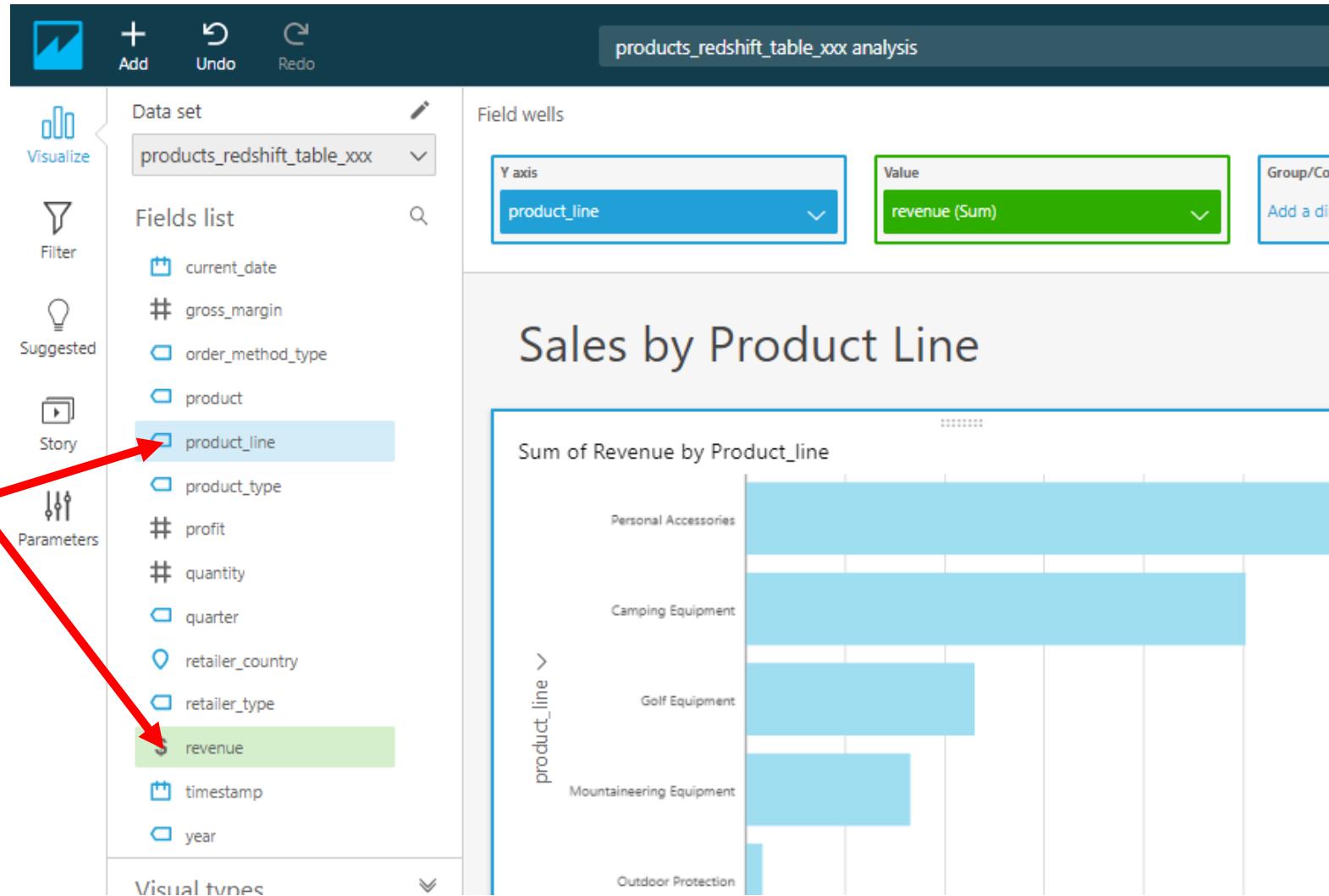
Parameters



QUICKSIGHT



AWS Business Intelligence Tool



The screenshot shows the AWS Quicksight interface. On the left, there's a sidebar with various icons and a list of fields. A red arrow points from the text "Choose product_line and revenue" to the "product_line" and "revenue" fields in the sidebar. The main area displays a bar chart titled "Sales by Product Line" with the subtitle "Sum of Revenue by Product_line". The chart shows revenue for different product lines: Personal Accessories, Camping Equipment, Golf Equipment, Mountaineering Equipment, and Outdoor Protection.

Choose
product_line
and revenue

product_line

revenue

Y axis: product_line

Value: revenue (Sum)

Sales by Product Line

Sum of Revenue by Product_line

product_line	Revenue (Sum)
Personal Accessories	High
Camping Equipment	Medium-High
Golf Equipment	Medium-Low
Mountaineering Equipment	Low-Medium
Outdoor Protection	Very Low

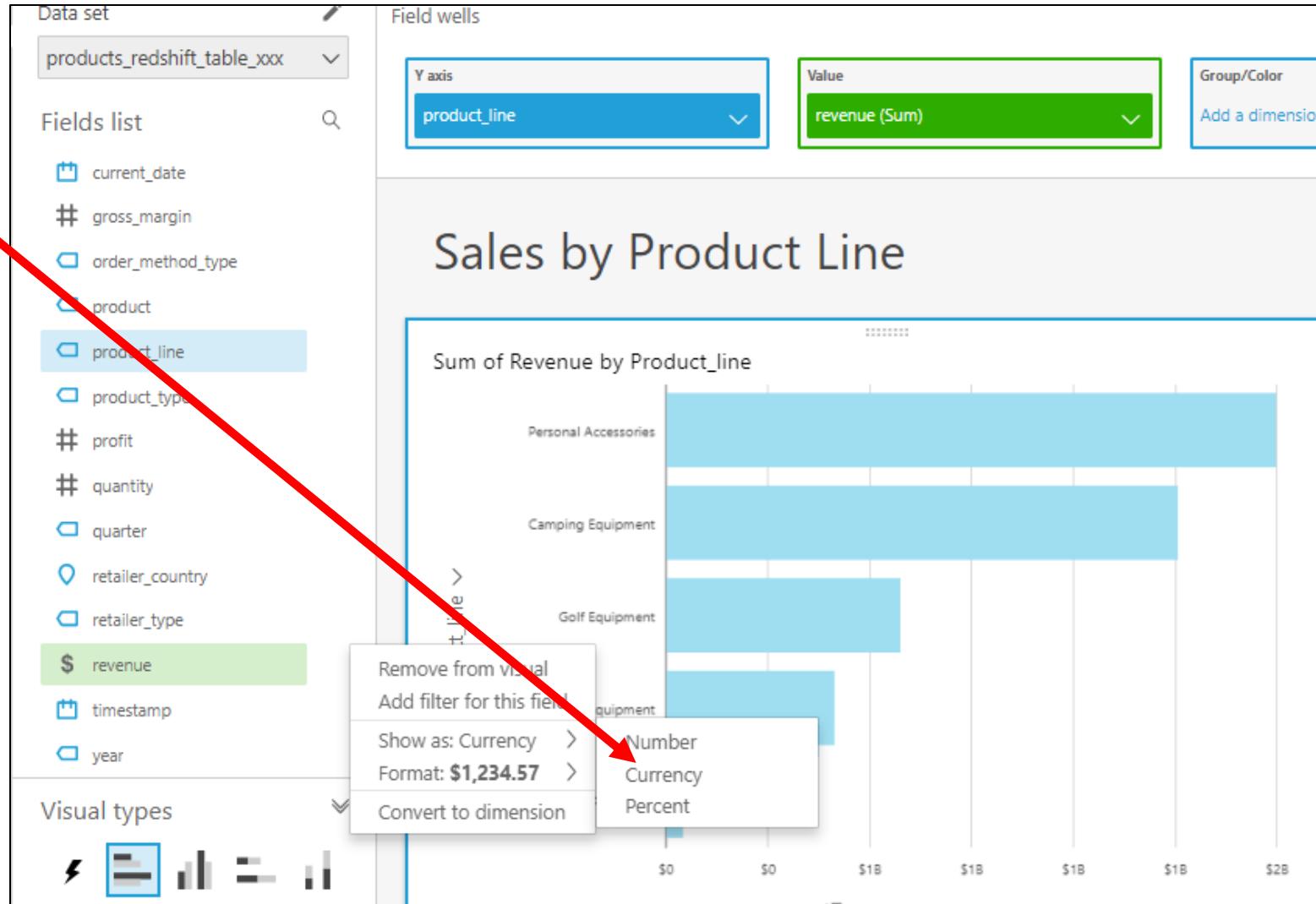


QUICKSIGHT



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

Sales by Product Line

Sum of Revenue by Product_line

product_line	Sum of Revenue
Personal Accessories	\$3.5B
Camping Equipment	\$2.8B
Golf Equipment	\$1.8B
Mountaineering Equipment	\$0.8B
Outdoor Protection	\$0.2B

product_line >

product_line

Visual types

Fields list

Data set

product_line

product_type

product

revenue

current_date

gross_margin

order_method_type

product

product_line

product_type

profit

quantity

quarter

retailer_country

retailer_type

timestamp

year

Y axis

Value

Group/C

Add a d

Visualize

Add

Undo

Redo

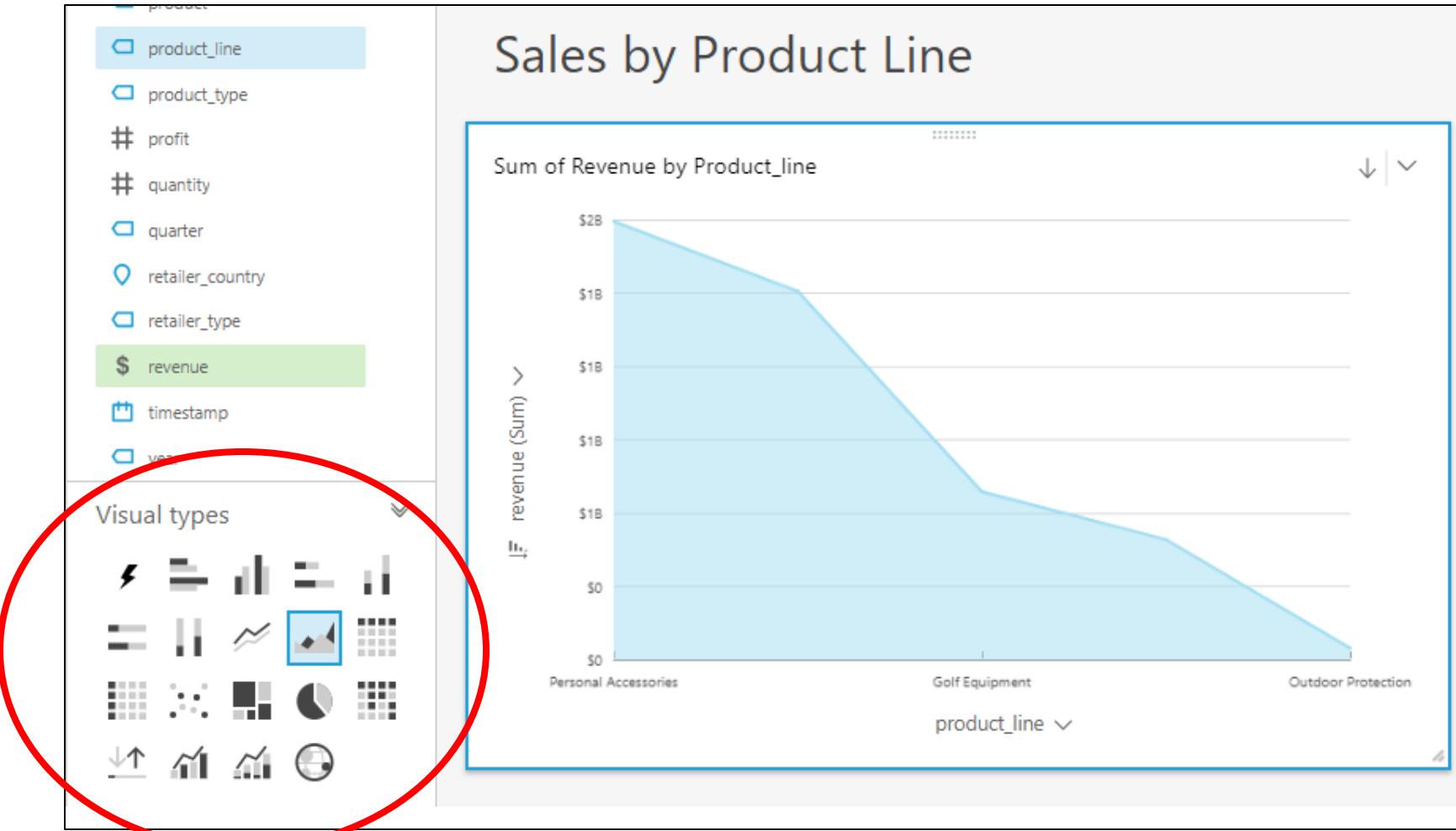
products_redshift_table_xxx analysis

QUICKSIGHT



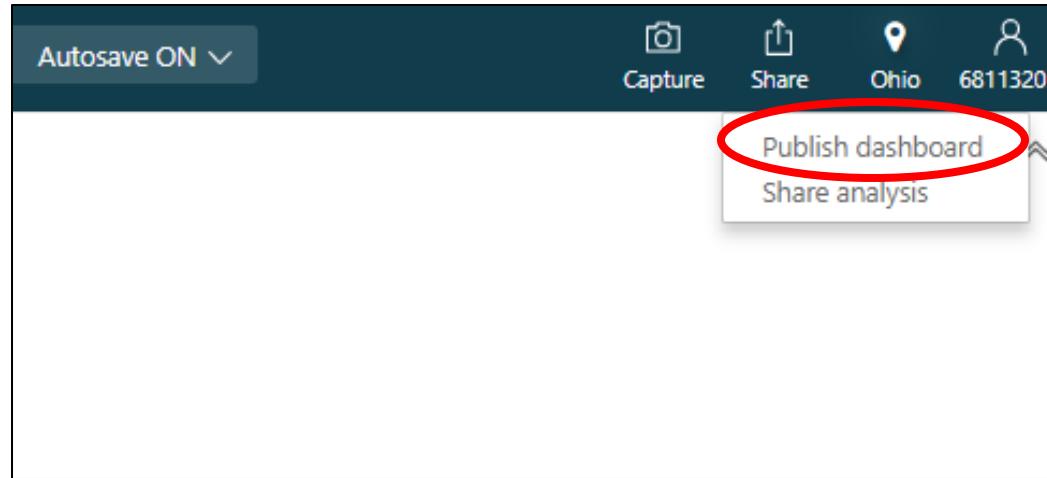
AWS Business Intelligence Tool

Change Visual Type



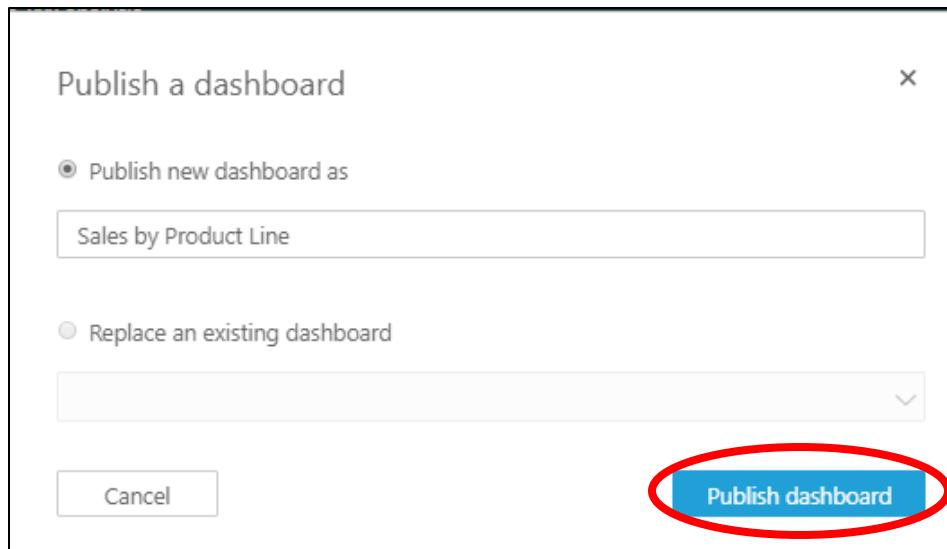


Publish to Dashboard





Name the Dashboard and select Publish dashboard



Lab 7

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

(Use US-EAST-2/Ohio Region)



CLOUDFORMATION

└ Templates

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



CLOUDFORMATION



└ Templates

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

Parameters:

GlueDatabaseName:

Type: String

Default: glue_database_XXX

GlueConnectionName:

Type: String

Default: glueTutorial_XXX

RedshiftDBName:

Type: String

Default: glueTutorial_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: glueTutorial

ScriptLocation:

Type: String

Default: "s3://glueTutorial-XXX/products_XXX"

Resources:

MyJob:

Type: AWS::Glue::Job

Properties:

Command:

Name: glueetl

ScriptLocation: !Ref ScriptLocation

AllocatedCapacity: 2

DefaultArguments:

--REDSHIFT_DB_NAME": !Ref RedshiftDBName

--SCHEMA_NAME": !Ref SchemaName

--REDSHIFT_TABLE_NAME": !Ref RedshiftTableName

--GLUE_TABLE_NAME": !Ref GlueTableName

--CONNECTION_NAME": !Ref GlueConnectionName

--GLUE_DB_NAME": !Ref GlueDatabaseName

ExecutionProperty:

MaxConcurrentRuns: 2

Connections: !Ref GlueConnectionName

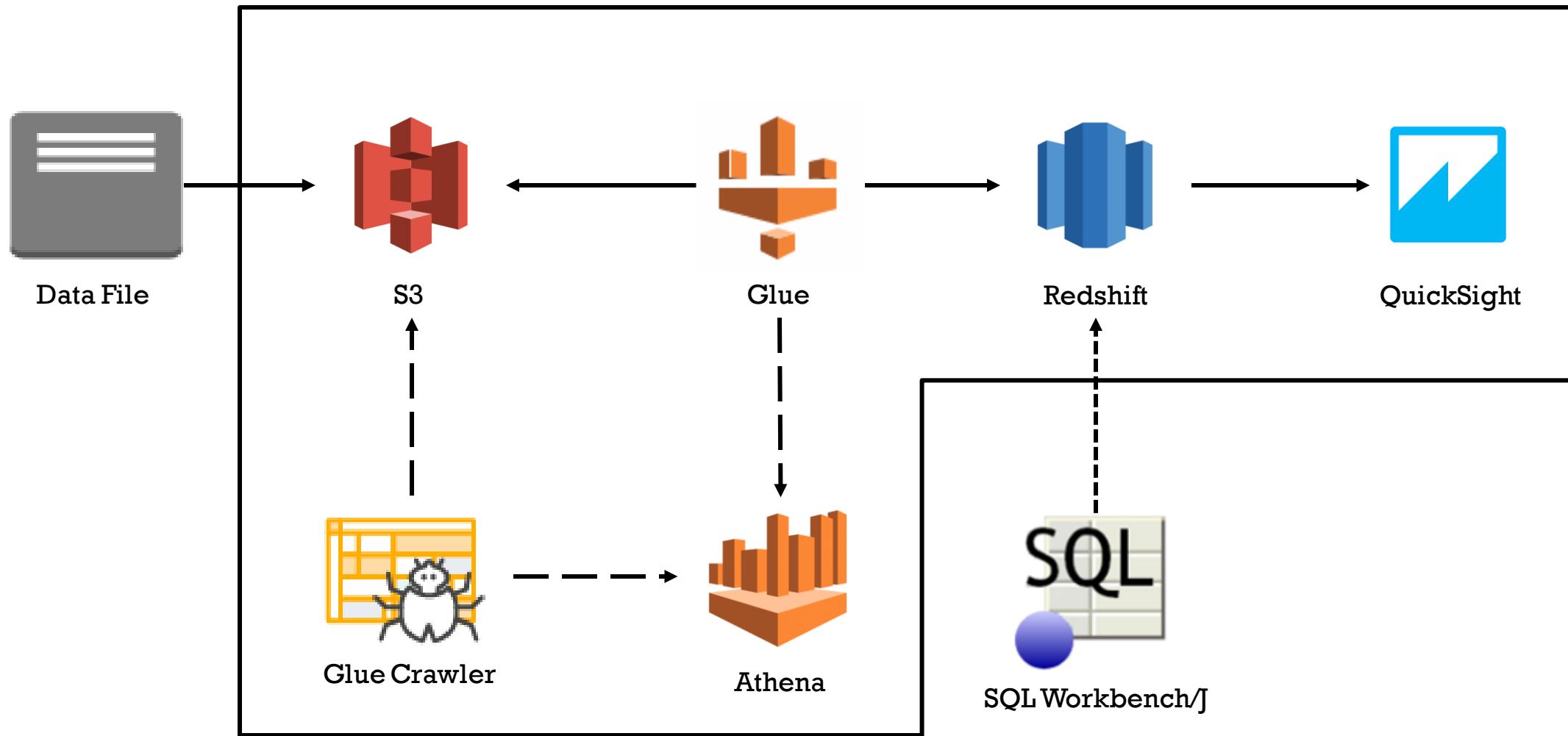
MaxRetries: 0

Name: !Ref GlueJobName



SUMMARY

AWS Data Workflow



Conclusion

Glue - AWS ETL Tool

Simple –

Use AWS for your ETL job

Less Setup

Flexible –

Good for developers as well as non-developers

Customizable

Cost Effective –

Cheaper than other ETL tools

Pay only when you use Glue



CLEAN UP

└ AWS

Delete the following resources:

Redshift Cluster *

S3 Bucket *

QuickSight Account *

DynamoDB Table*

Glue Job

Glue Database

Glue Table

Glue Connection

Lambda

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



RESOURCES

└ AWS Business Intelligence Tool

AWS Glue Documentation

<https://aws.amazon.com/glue/>

Pricing

Informatica

https://aws.amazon.com/marketplace/pp/B0752DY9DV?qid=1534179668153&sr=0-1&ref=srh_res_product_title

Glue

<https://aws.amazon.com/glue/pricing/>

Matillion

<https://aws.amazon.com/marketplace/pp/B010ED5YF8>

AWS Services Documentation

<https://aws.amazon.com/documentation/>

Hadoop vs AWS

<https://www.trustradius.com/compare-products/amazon-web-services-vs-hadoop>

<https://databricks.com/blog/2017/05/31/top-5-reasons-for-choosing-s3-over-hdfs.html>

<https://data-flair.training/blogs/13-limitations-of-hadoop/>



THANK YOU

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

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

