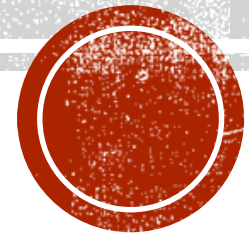
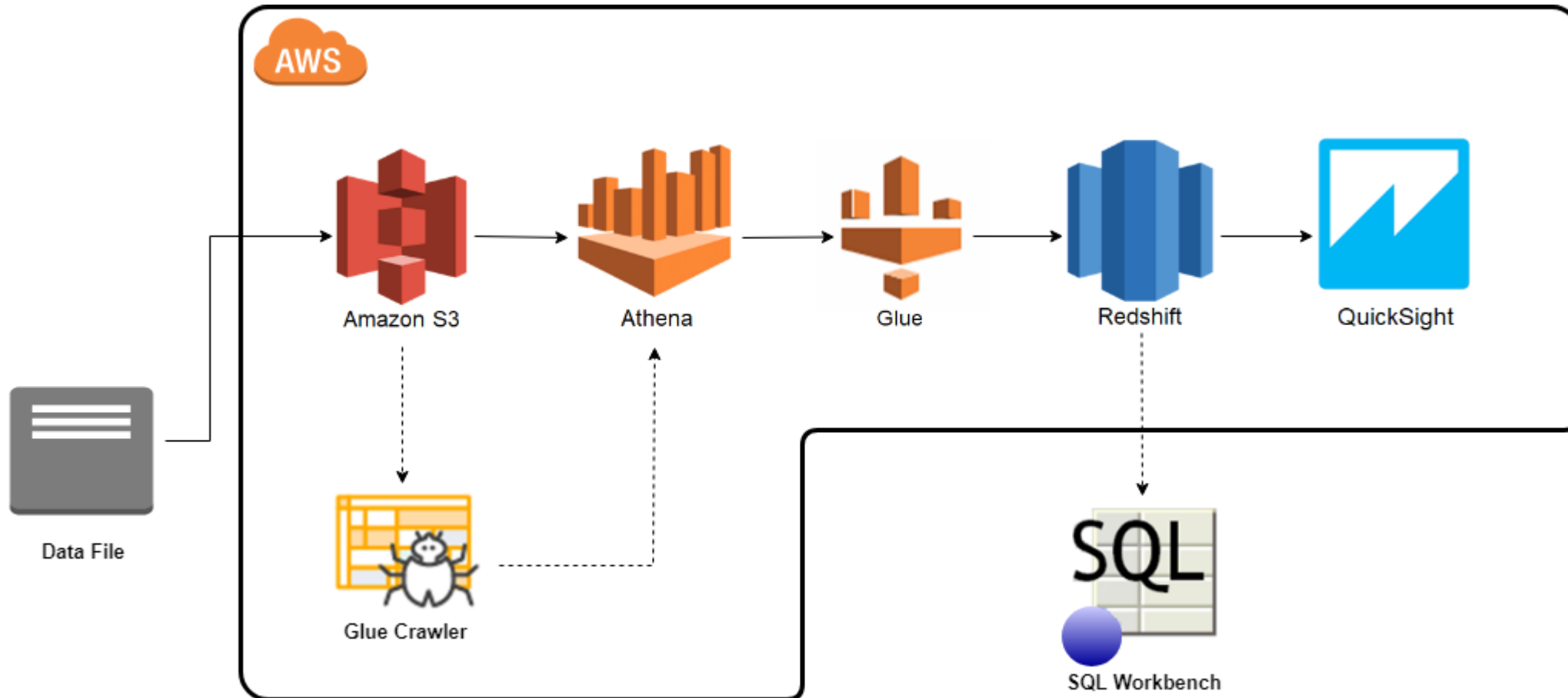


ARTS & CRAFTS WITH AWS GLUE

ETL Workshop



Amazon Web Services



AWS Glue

What is Glue?



AWS Glue

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



ETL

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



ETL

Original Data File

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

Example Business Requirements:

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

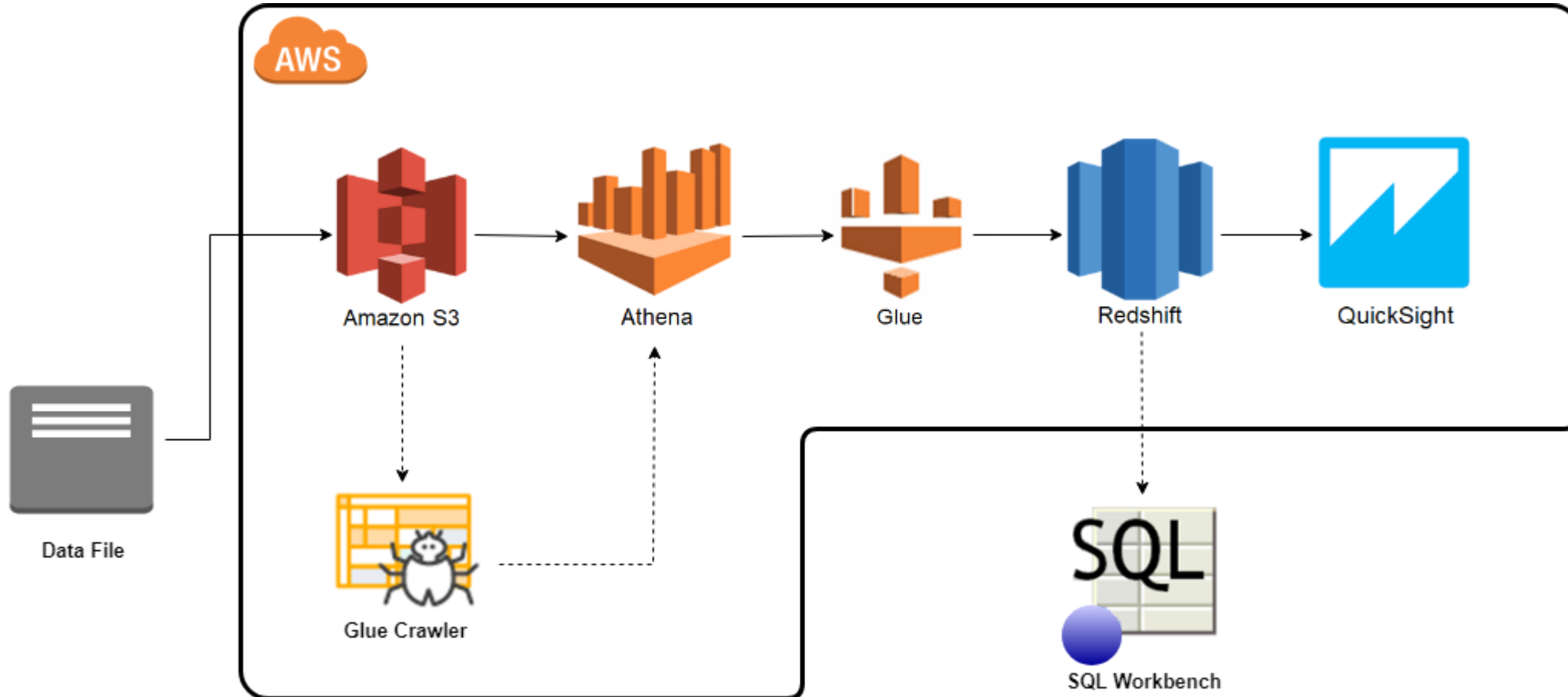


Glue Tutorial Prerequisites

- 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 dashboard

Clusters

Snapshots

Security

Parameter groups

Workload management

Reserved nodes

Advisor Beta

Events

Connect client

What's new

Launch cluster

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

Quick launch cluster

Launch cluster

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

Resources

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

Clusters (0)
Increase cluster limit


Security
Subnet groups (1)

Parameter groups (0)
Total Reservations (0)

Snapshots (0)
Manual (0)
Automated (0)

Events (0)
Event subscriptions (0)

Service health

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

› View complete service health details



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

Database name 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* Port number on which the database accepts connections.

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

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

Confirm password Confirm master user password



Redshift

— Create AWS Datawarehouse

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

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

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

Node type

dc2.large

Specifies the compute, memory, storage, and I/O capacity of the cluster's nodes.

CPU

7 EC2 Compute Units (2 virtual cores) per node

Memory

15.25 GiB per node

Storage

160GB SSD storage per node

I/O performance

Moderate

Cluster type

Single Node

Single Node clusters consist of a single node which performs both leader and compute functions.

Number of compute nodes*

1

Maximum

1

Minimum

1

Cancel

Previous

Continue

Redshift

└─ Create AWS Datawarehouse

Choose default VPC

Choose default
subnet group

Choose subnet
availability zone

Choose default
security group

Redshift dashboard

Launch your Amazon Redshift cluster - Advanced settings | [Switch to quick launch](#)

Clusters

Snapshot

CLUSTER DETAILS NODE CONFIGURATION ADDITIONAL CONFIGURATION REVIEW

Optionally, create a basic alarm for this cluster.



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

Optionally, select your maintenance track for this cluster.

Maintenance Track ☒ Current ☐ Trailing

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

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

Available IAM roles  



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

<input type="checkbox"/>	Cluster	Cluster Status	DB Health	Release Status	In Maintenance	Recent Events
<input checked="" type="checkbox"/>	glue-tutorial-jds	available	healthy	Up to date	no	3

Endpoint [glue-tutorial-jds.chtswcubv1n.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

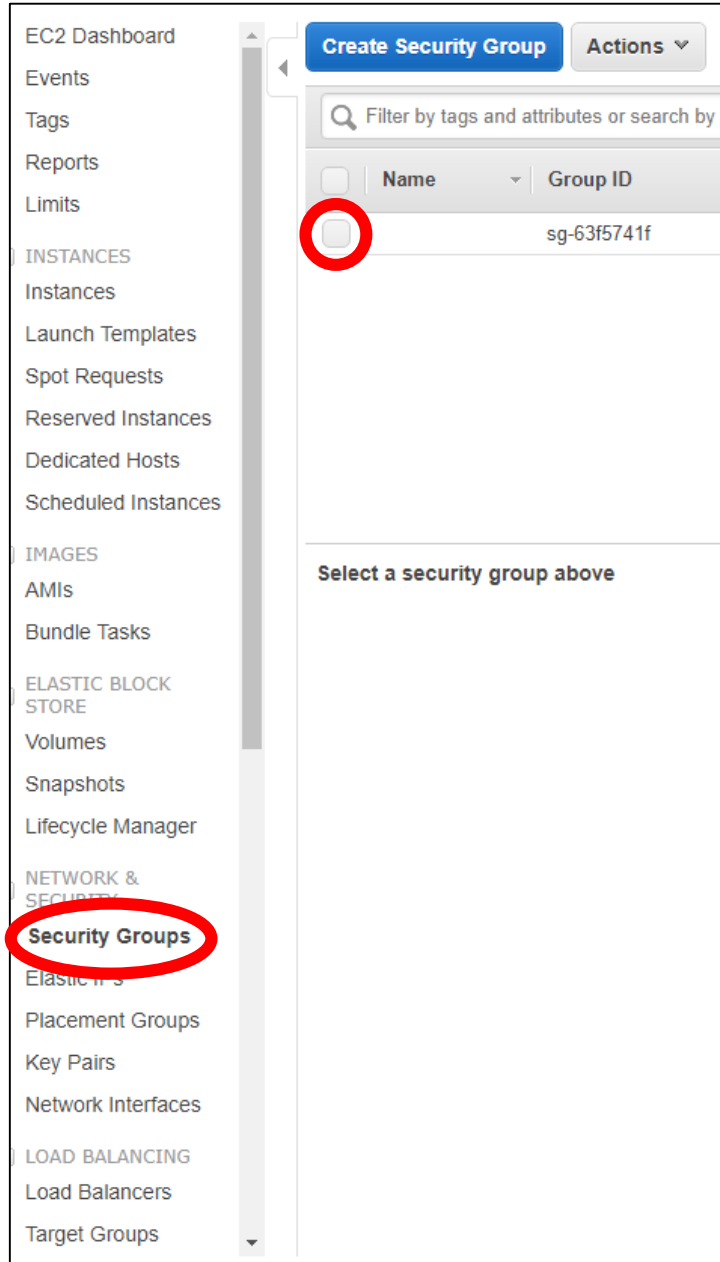
Tags

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 Dashboard

Create Security Group Actions ▾

Filter by tags and attributes or search by k

<input type="checkbox"/>	Name ▾	Group ID
<input type="checkbox"/>		sg-63f5741f

Select a security group above

EC2 Dashboard

Events

Tags

Reports

Limits

INSTANCES

Instances

Launch Templates

Spot Requests

Reserved Instances

Dedicated Hosts

Scheduled Instances

IMAGES

AMIs

Bundle Tasks

ELASTIC BLOCK STORE

Volumes

Snapshots

Lifecycle Manager

NETWORK & SECURITY

Security Groups

Elastic IP

Placement Groups

Key Pairs

Network Interfaces

LOAD BALANCING

Load Balancers

Target Groups



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

Edit inbound rules

Type ⓘ	Protocol ⓘ	Port Range ⓘ	Source ⓘ	Description ⓘ	
All traffic ▾	All	0 - 65535	Custom ▾ sg-63f5741f	e.g. SSH for Admin Desktop	✕
Redshift ▾	TCP	5439	Anywhere ▾ 0.0.0.0/0, ::/0	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.

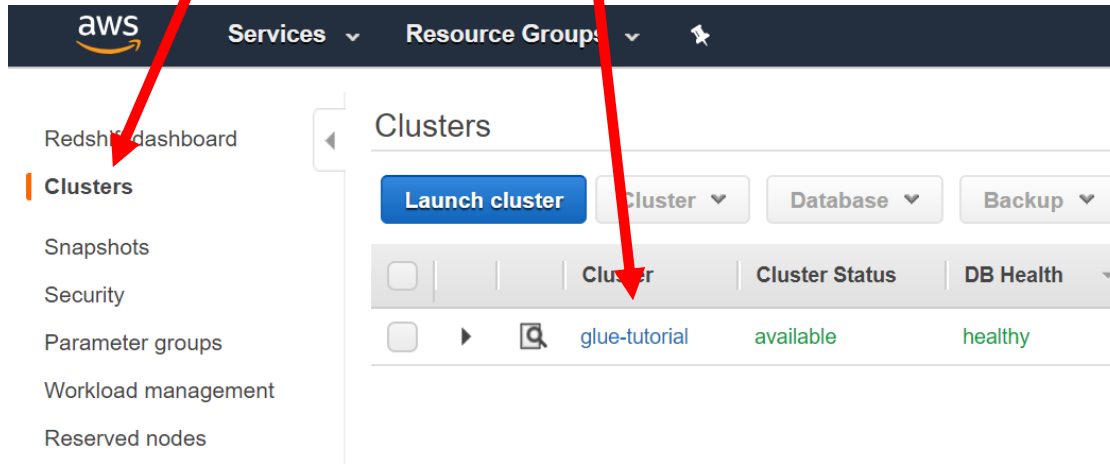
Cancel Save

Redshift

Connection

Go to Redshift and select 'Clusters'

Select glue-tutorial



Scroll down to Cluster Database Properties and copy the JDBC URL

Cluster: **glue-tutorial** Configuration Status Cluster Pe

Cluster Database Properties

Port	5439
Publicly Accessible	Yes
Database Name	sales
Master Username	master
Encrypted	No
JDBC URL	<code>jdbc:redshift://glue-tutorial.chafpggokoad.us-east-1.redshift.amazonaws.com:5439/sales</code>
ODBC URL	<code>Driver={Amazon Redshift (x64)}; Server=glue-tutorial.chafpggokoad.us-east-1.redshift.amazonaws.com; Database=sales; UID=master; PWD=insert_your_master_user_password_here; Port=5439</code>

Backup, Aud

Automated S
Cross-F

Capacity Details

Current Node Type	dc2.large
CPU	7 EC2 Compute Units (2 virtual cores)

SSH ingestio

Cluster public
ssh-rsa
AAAAA...

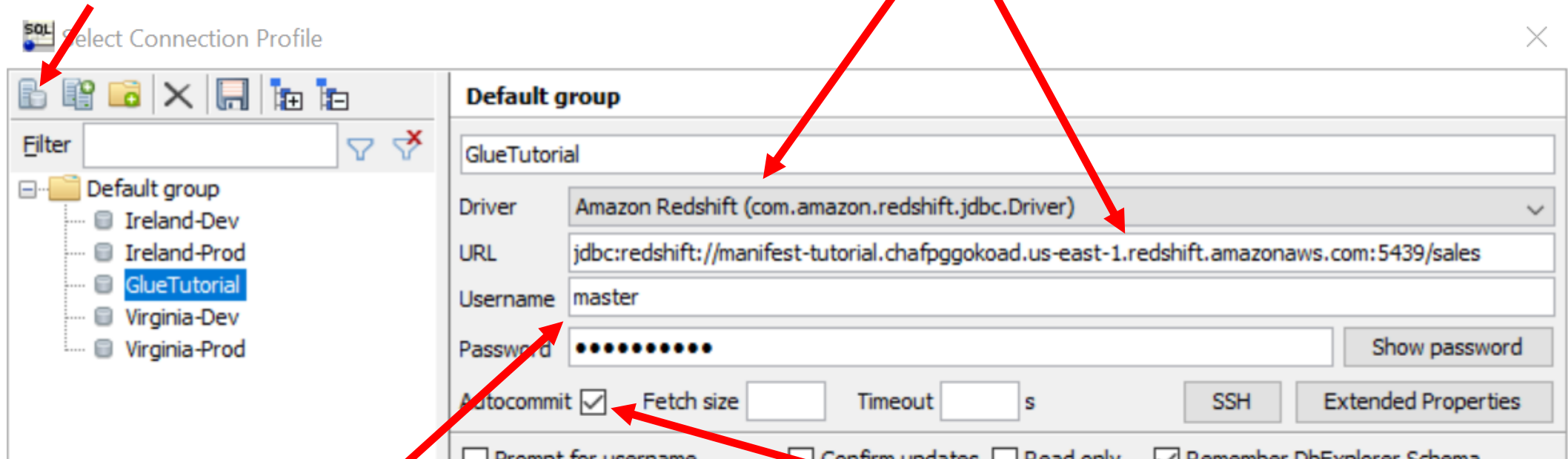




Connection

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



Select Connection Profile

Filter

- Default group
 - Ireland-Dev
 - Ireland-Prod
 - GlueTutorial
 - Virginia-Dev
 - Virginia-Prod

Default group

GlueTutorial

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

URL: jdbc:redshift://manifest-tutorial.chafpggokoad.us-east-1.redshift.amazonaws.com:5439/sales

Username: master

Password: [masked] Show password

Autocommit ☒ Fetch size [] Timeout [] s SSH Extended Properties

☐ Prompt for username ☐ Confirm updates ☐ Read only ☒ Remember DbExplorer Schema

☒ Save password ☐ Confirm DML without WHERE ☐ Store completion cache locally

☒ Separate connection per tab ☐ Rollback before disconnect ☐ Remove comments

☐ Ignore DROP errors ☐ Empty string is NULL ☐ Hide warnings

☐ Trim CHAR data ☒ Include NULL columns in INSERTs ☐ Check for uncommitted changes

Info Background [] [X] [] (None) Alternate Delimiter []

Workspace [] ...

Default directory [] ...

Main window icon [] ...

Macros [] ...

Tags []

Connect scripts Schema/Catalog Filter Variables Test

Manage Drivers Help OK Cancel

Test your connection



Lab 1

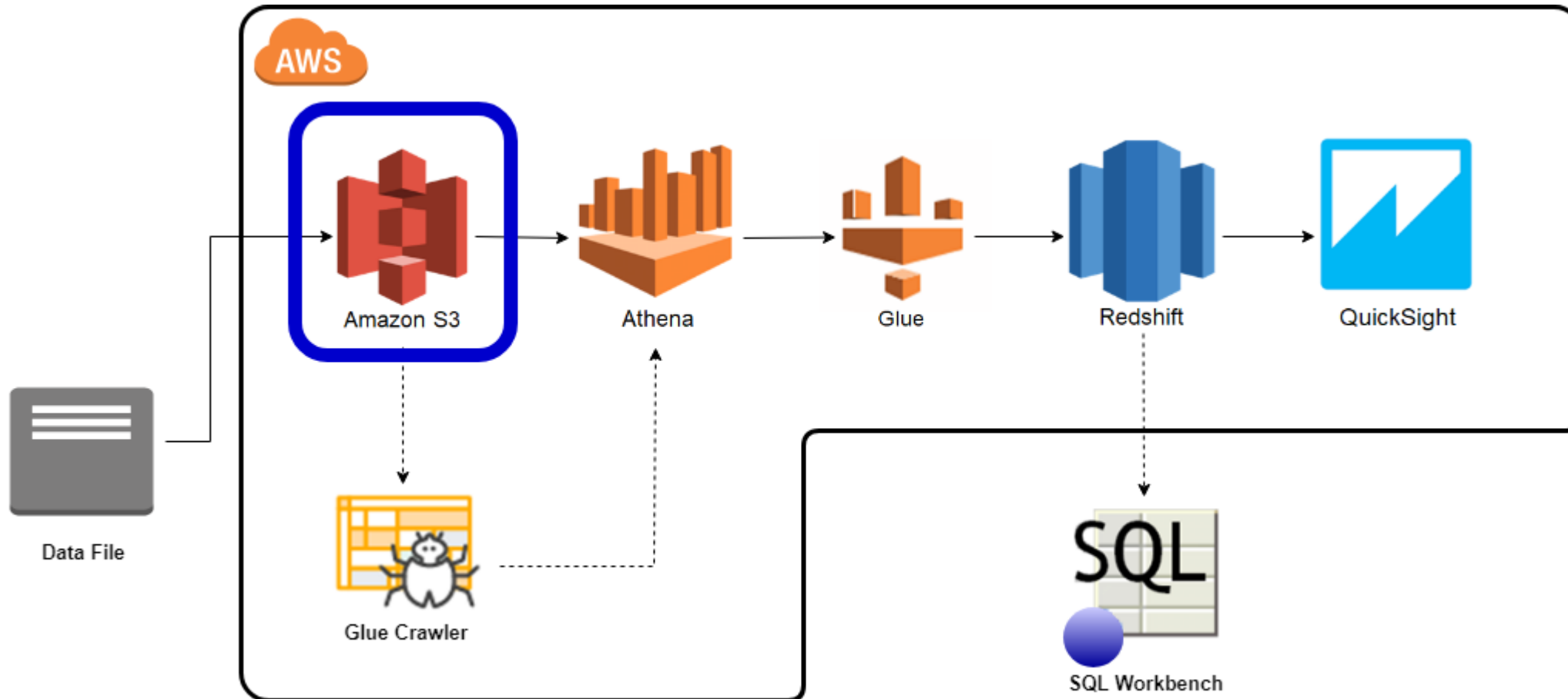
- Prerequisites
- Setup Redshift

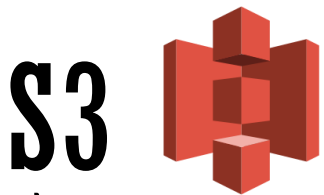
(Use US-EAST-1/N. Virginia Region)



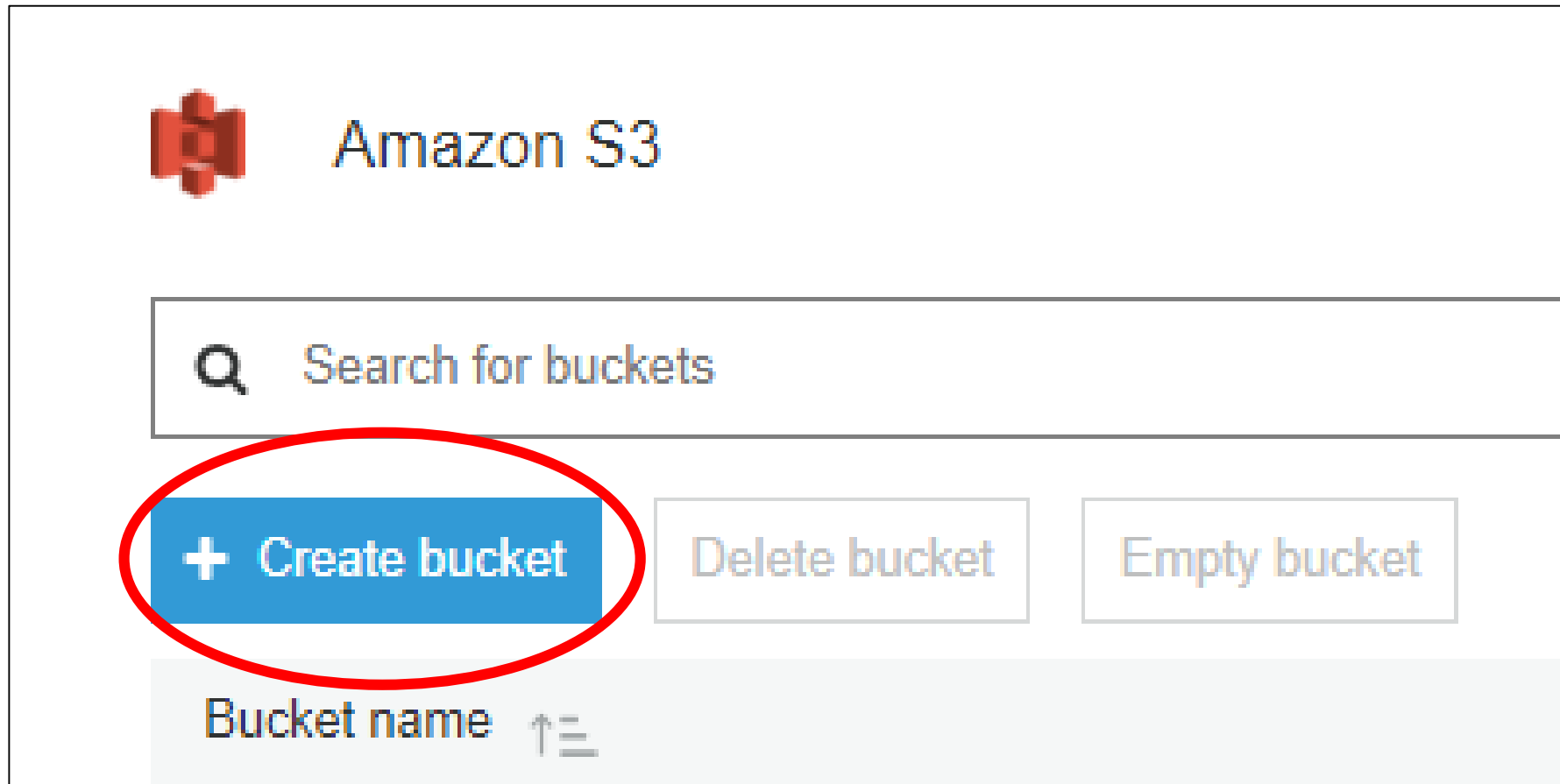
S3

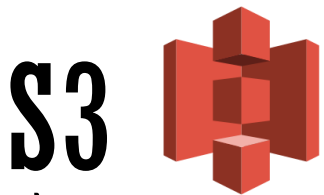
└ AWS Simple Storage Service





└ Create S3 bucket with AWS Console





Create S3 bucket with AWS Console

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

Create bucket

1 Name and region 2 Configure options 3 Set permissions 4 Review

Name and region

Bucket name ⓘ

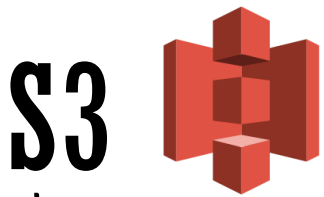
Enter DNS-compliant bucket name

Create Cancel Next

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

Specify the region





└─ **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





— Create S3 bucket folder

Create a folder
called
products_XXX

Upload + Create folder More ▾

☐ Name ↑ ▾

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:

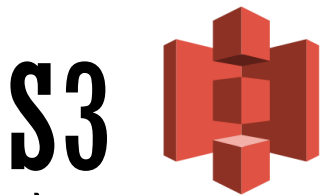
☒ None (Use bucket settings)

☐ 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)

☐ glue-scripts





Create S3 bucket folder

Create a folder
called glue-scripts

Upload + Create folder More ▾

☐ Name ↑

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)

☐ products.jar

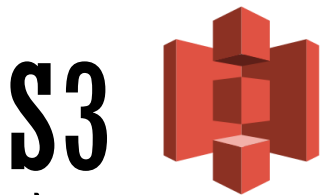




└ Add file to S3 bucket with AWS Console

Add file from repository called
“WA_Sales_Products_2012-14”

A screenshot of the AWS S3 'Upload' console. The interface has a blue header with the word 'Upload' and a close button. Below the header is a progress bar with four steps: '1 Select files' (highlighted), '2 Set permissions', '3 Set properties', and '4 Review'. The main area shows '1 Files' with a 'Size: 9.2 MB' and 'Target path: glue-tutorial-jds/products_jds/'. There is a '+ Add more files' link. Below that, a file named 'WA_Sales_Products_2012-14.csv' with a size of '- 9.2 MB' is listed, accompanied by a file icon and a close button. At the bottom, there are two buttons: 'Upload' (circled in red) and 'Next'.



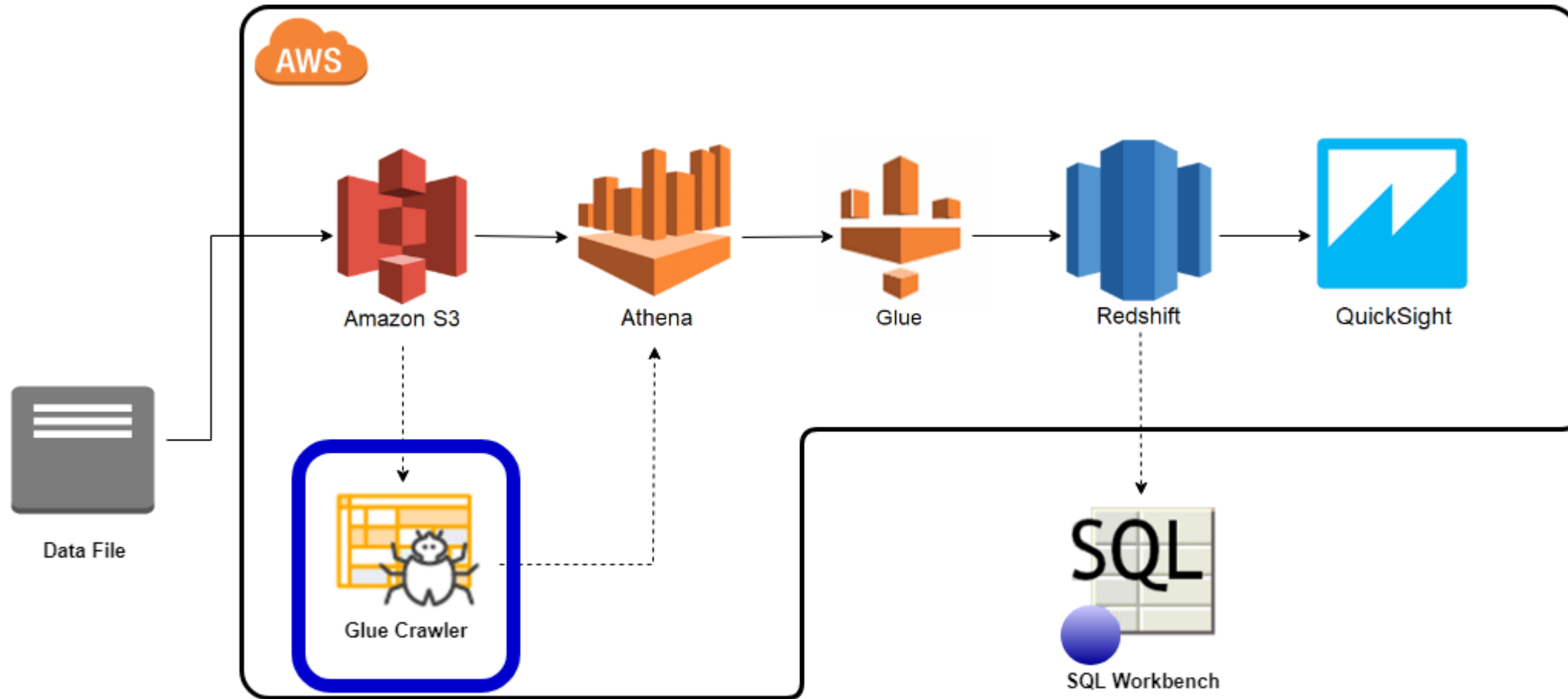
└─ **Add file to S3 bucket with AWS CLI***
(Alternative)

```
$ 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
```

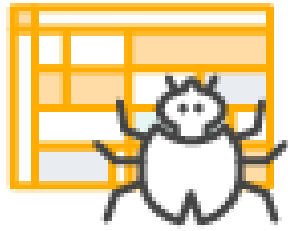
* Must install and set up AWS CLI in order to use this



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



Databases A database is a set of associated table definitions, organized into a logical group.

Add database

View tables

Action ▼

Create a new
Database

In the Glue Console
click on Databases



Glue



— Create Glue Database

Edit database

Database name

sales_jds

▼ Description and location (optional)

Location ⓘ

Enter location...

Description

Enter description...

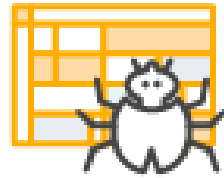
Apply

Give your database a name
“sales_XXX”



Glue Crawler

— Create Glue Crawler



Click on add tables to
create a table

Add tables

Action

Filter by attributes or search by keyword

Save view

Showing: 1 - 2 < >   

<input type="checkbox"/> Name	Database	Location	Classification	Last updated	Deprecated
<input type="checkbox"/> elb_logs	sampledb	s3://athena-examples-us-east-1/elb/plaintext	Unknown	26 July 2018 10:02 AM UTC-4	

Add tables ▾

Action ▾

Filter by attributes or search by keyword

Save view ▾

Showing: 1 - 2 < > ↺ ⚙ ?

Add tables using a crawler

Add table manually

elb_logs

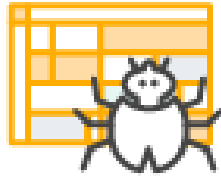
Database	Location	Classification	Last updated	Deprecated
sampledb	s3://athena-examples-us-east-1/elb/plaintext	Unknown	26 July 2018 10:02 AM UTC-4	

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	Classification
No items available	

Give your crawler a name,
glue-tutorial-XXX

▼ Grouping behavior for S3 data (optional)

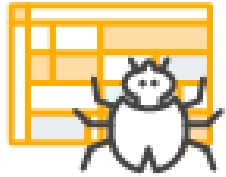
☐ Create a single schema for each S3 path

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



Glue Crawler



— Create Glue Crawler

Add a data store

Choose a data store

S3

Crawl data in

☒ Specified path

Include path

s3://glue-tutorial-jds/products_jds/

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

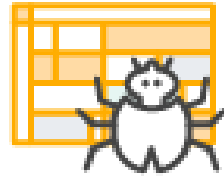
Choose where the table is going to look for data

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

Add another data store

☐ Yes

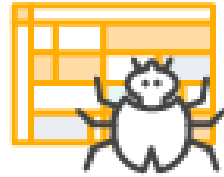
☒ No

Back Next



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-

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](#).

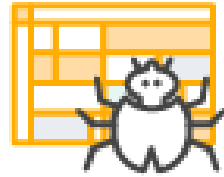
[Back](#)

[Next](#)



Glue Crawler

— Create Glue Crawler



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

Create a schedule for this crawler

Frequency

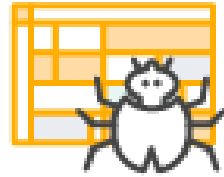
Run on demand

Back Next



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

Configure the crawler's output

Database ⓘ

glue-tutorial-jds

Add database

Prefix added to tables (optional) ⓘ

Type a prefix added to table names

▼ Configuration options (optional)

During the crawler run, all schema changes are logged.

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

☒ Update the table definition in the data catalog.

☐ Add new columns **only**.

☐ Ignore the change and don't update the table in the data catalog. ⓘ

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

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

☐ Delete tables and partitions from the data catalog.

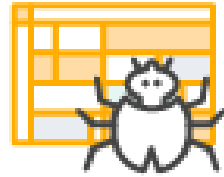
☐ Ignore the change and don't update the table in the data catalog.

☒ Mark the table as deprecated in the data catalog. ⓘ

Back Next



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

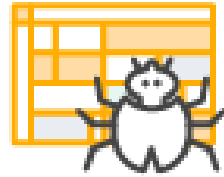
Mark the table as deprecated in the data catalog.

Back

Finish



Glue Crawler



— Run the Crawler to create Athena table

Run your crawler

Add crawler

Run crawler

Action

Filter by attributes

Showing: 1 - 1

<input checked="" type="checkbox"/>	Name	Schedule	Status	Logs	Last runtime	Median runtime	Tables updated	Tables added
<input checked="" type="checkbox"/>	glue-tutorial-jds		Ready		0 secs	0 secs	0	0

Select your crawler

Add tables

Action

Filter by attributes or search by keyword

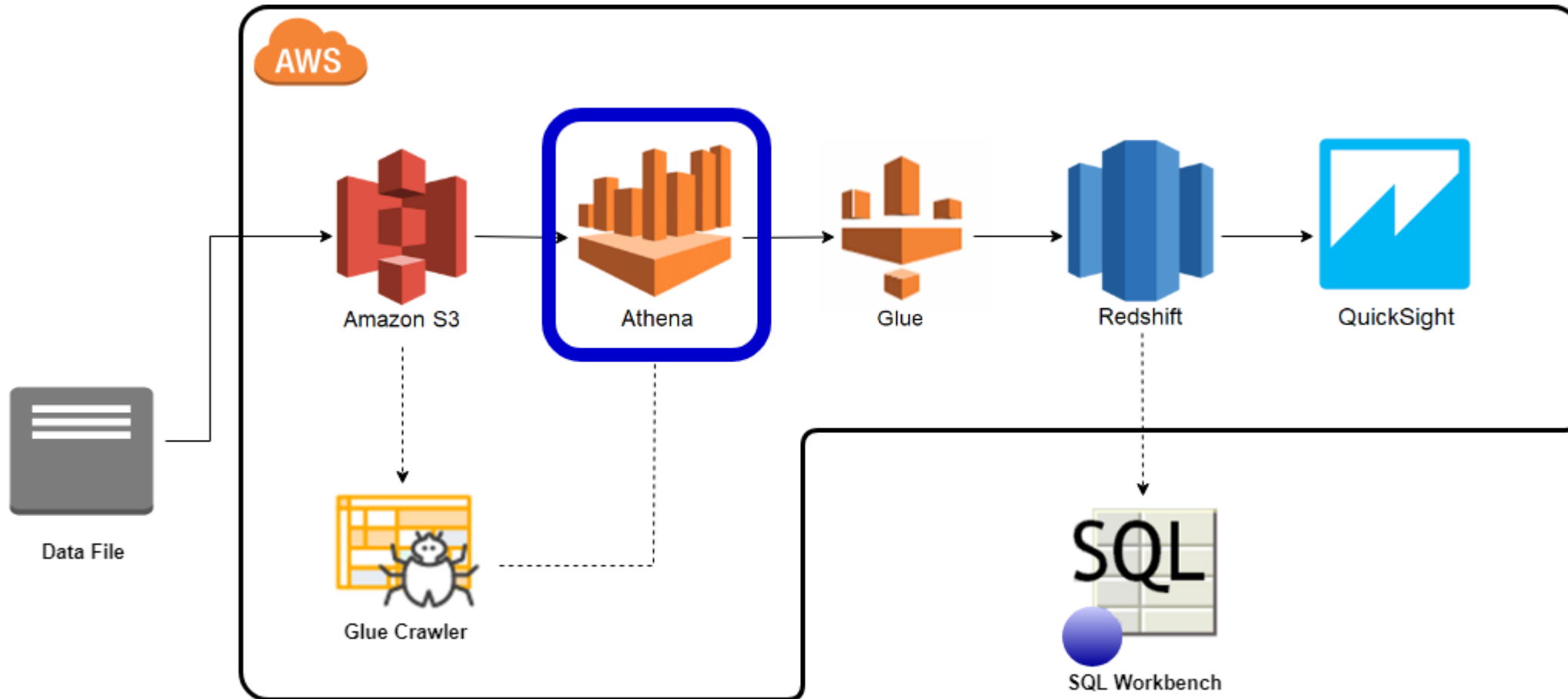
Save view

<div><div></div></div> Name	Database	Location	Classification	Last updated
<div><div></div></div> elb_logs	sampledb	s3://athena-examples-us-east-1/elb/plaintext	Unknown	26 July 2018 10:02 AM UTC-4
<div><div></div></div> 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





- 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 Query Editor interface showing a SQL query and its results.

Database: glue-tutorial-jds

Tables (2):

- file_ingestion
- glue_tutorial_jds
 - retailer country (string)
 - order method type (string)
 - retailer type (string)
 - product line (string)
 - product type (string)
 - product (string)
 - year (bigint)
 - quarter (string)
 - revenue (double)
 - quantity (bigint)
 - gross margin (double)

Views (0):

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

Query:

```
select * from glue_tutorial_jds limit 10
```

Run query **Save as** **Create view from query** (Run time: 1.58 seconds, Data scanned: 300.97KB) **Format query** **Clear**

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

Results:

	retailer country	order method type	retailer type	product line	product type	product	year	quarter	revenue	quantity	gross margin
1	United States	Fax	Outdoors Shop	Camping Equipment	Cooking Gear	TrailChef Deluxe Cook Set	2012	Q1 2012	59628.66	489	0.34754797
2	United States	Fax	Outdoors Shop	Camping Equipment	Cooking Gear	TrailChef Double Flame	2012	Q1 2012	35950.32	252	0.4742745
3	United States	Fax	Outdoors Shop	Camping Equipment	Tents	Star Dome	2012	Q1 2012	89940.48	147	0.35277197





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

Athena Query Editor interface showing a SQL query and its results.

Database: glue-tutorial-jds

Tables (1): glue_tutorial_sales

Views (0): You have not created any views. To create a view, run a query and click "Create view from query"

Query: 1 select count(*) from glue_tutorial_sales

Buttons: Run query, Save as, Create view from query

Metadata: (Run time: 1.56 seconds, Data scanned: 9.21MB)

Results:

	_col0
1	88475



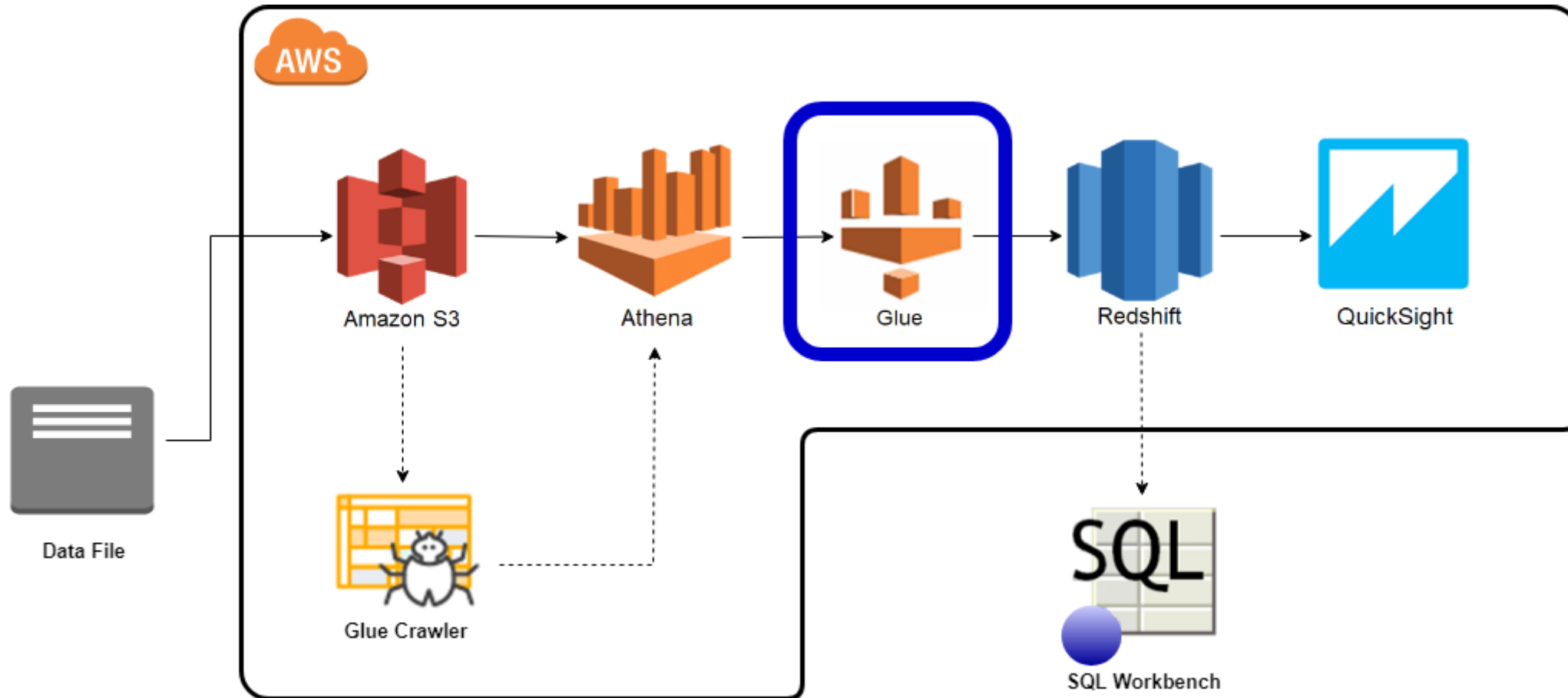
Lab 2

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

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



Glue

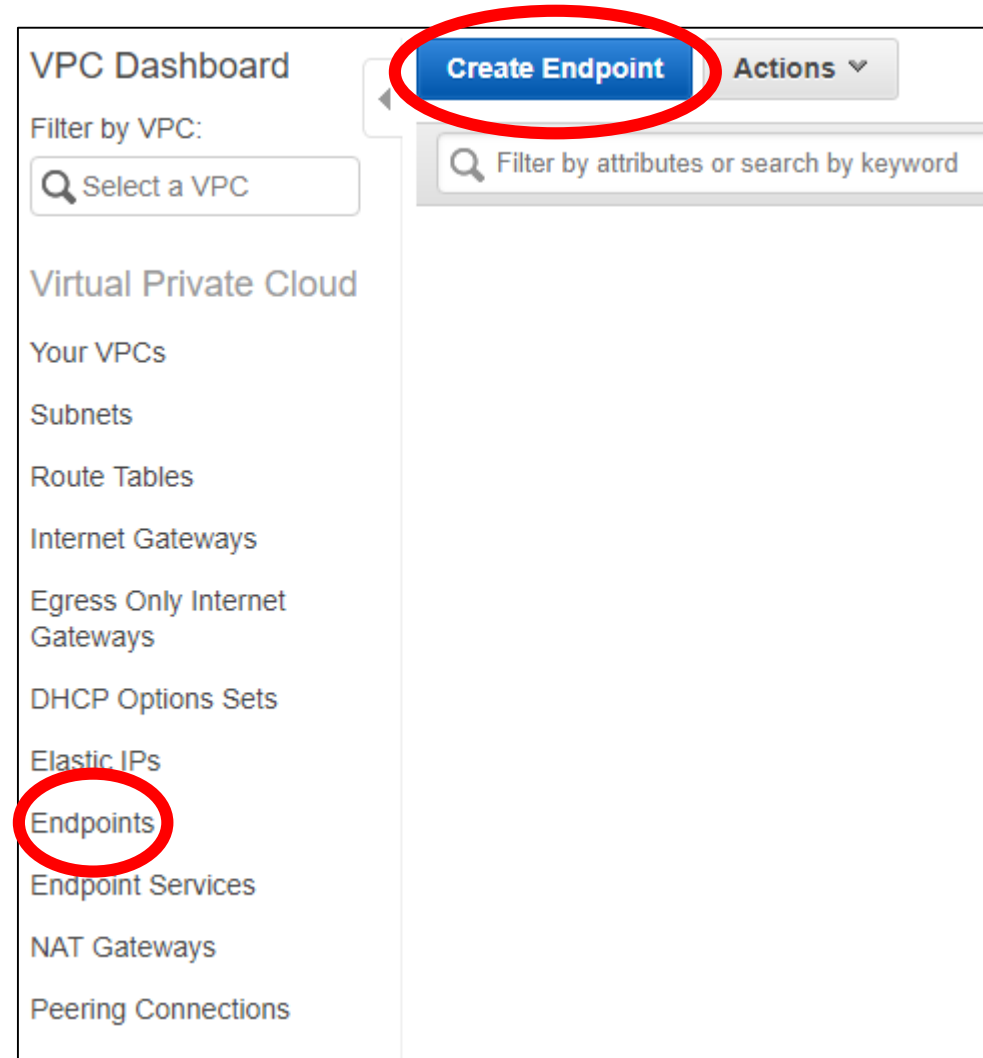


VPC



— Create a S3 endpoint

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



VPC



— Create a S3 endpoint

Select the S3
Service for Glue
to access S3

Service category ☒ AWS services ☐ Find service by name ☐ Your AWS Marketplace services

Service Name com.amazonaws.eu-west-1.s3 ⓘ

Filter by attributes | 1 to 21 of 21

	Service Name	Owner	Type
<input type="radio"/>	com.amazonaws.eu-west-1.ec2	amazon	Interface
<input type="radio"/>	com.amazonaws.eu-west-1.ec2messages	amazon	Interface
<input type="radio"/>	com.amazonaws.eu-west-1.elasticloadbal...	amazon	Interface
<input type="radio"/>	com.amazonaws.eu-west-1.events	amazon	Interface
<input type="radio"/>	com.amazonaws.eu-west-1.execute-api	amazon	Interface
<input type="radio"/>	com.amazonaws.eu-west-1.kinesis-streams	amazon	Interface
<input type="radio"/>	com.amazonaws.eu-west-1.kms	amazon	Interface
<input type="radio"/>	com.amazonaws.eu-west-1.logs	amazon	Interface
<input type="radio"/>	com.amazonaws.eu-west-1.monitoring	amazon	Interface
<input checked="" type="radio"/>	com.amazonaws.eu-west-1.s3	amazon	Gateway
<input type="radio"/>	com.amazonaws.eu-west-1.sagemaker.api	amazon	Interface
<input type="radio"/>	com.amazonaws.eu-west-1.sagemaker.ru...	amazon	Interface
<input type="radio"/>	com.amazonaws.eu-west-1.secretsmanager	amazon	Interface
<input type="radio"/>	com.amazonaws.eu-west-1.servicecatalog	amazon	Interface
<input type="radio"/>	com.amazonaws.eu-west-1.sns	amazon	Interface





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
<input checked="" type="checkbox"/> 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](#)






— Create a connection to Redshift

Go to Glue in services and
click on connections

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

[Add connection](#) [Test connection](#) [Action](#) Showing: 0 - 0 < > ↺ ⓘ

<input type="checkbox"/>	Name	Type	Date created	Last updated	Updated by
<div> You don't have any connections yet. Add connection</div>					

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



Glue



— Create a connection to Redshift

Set up your connection's properties.

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

Connection name

glue-tutorial-jds

Connection type

JDBC

Description (optional)

Enter description...

Next

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

Set up access to your data store.

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

JDBC URL ⓘ

JDBC syntax for most database engines is `jdbc:protocol://host:port/databasename`.

SQL Server syntax is `jdbc:sqlserver://host:port;databaseName=db_name`. Oracle syntax is `jdbc:oracle:thin://@host:port/service_name`. For more variations, see [Working with Connection](#).

Username

Password

VPC

Choose the VPC name that contains your data store.

Subnet

Choose the subnet within your VPC.

Security groups

Choose one or more security groups that allow access to the data store in your VPC. AWS Glue associates these security groups to the ENI attached to your subnet. To allow AWS Glue components to communicate and also prevent access from other networks, at least one chosen security group must specify a self-referencing inbound rule for all TCP ports.

<input checked="" type="checkbox"/> Group ID	Group name
<input checked="" type="checkbox"/> sg-045e2700f92cc3413	default



Glue



— **Test the connection to Redshift**

Add connection

Test connection

Action ▼

Showing: 1 - 1 < > ↺ ⓘ

<input checked="" type="checkbox"/>	Name	Type	Date created	Last updated	Updated by
<input checked="" type="checkbox"/>	glue-tutorial-jds	JDBC	27 July 2018 1:37 PM UTC-4	27 July 2018 1:37 PM UTC-4	





Test the connection to Redshift

Test connection

IAM role ⓘ

Choose an IAM role

- AWSGlueServiceRole-AWSGlueService-Glu... create
- AWSGlueServiceRole-DefaultRole
- AWSGlueServiceRole-glueServiceRole

Test connection

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 < > ↺ ⓘ

<input checked="" type="checkbox"/>	Name	Type	Date created	Last updated	Updated by
<input checked="" type="checkbox"/>	glue-tutorial-jds	JDBC	27 July 2018 1:37 PM UTC-4	27 July 2018 1:40 PM UTC-4	



<div><div>Add job</div><div>Action</div><div>Filter by attributes</div></div> <div>Showing: 1 - 1</div>				
<input type="checkbox"/> Name	ETL language	Script location	Last modified	Job bookmark
<input type="checkbox"/> sales_product_load	python	s3://manifest-glue-tutorial/glue_scripts/sales_pr...	26 July 2018 12:57 PM UTC-4	Disable

Click on jobs in the Glue Console





Give your job a name: glue-tutorial-XXX

The language used to write the script

Give your script a name glue-tutorial-XXX

The location where your script will be placed in S3

The image shows the "Job properties" form in the AWS Glue console. The form is titled "Job properties" and contains several sections. The "Name" field is set to "glue-tutorial-jds". The "IAM role" dropdown is set to "AWSGlueServiceRole-glueServiceRole". The "This job runs" section has three radio buttons: "A proposed script generated by AWS Glue", "An existing script that you provide", and "A new script to be authored by you", with the third option selected. The "ETL language" section has two radio buttons: "Python" (selected) and "Scala". The "Script file name" field is set to "glue-tutorial-jds". The "S3 path where the script is stored" field is set to "s3://glue-tutorial-jds/glue-scripts". The "Temporary directory" field is set to "s3://aws-glue-temporary-952552944372-us-east-1/jack.silverman". At the bottom, there is a section for "Advanced properties" and a section for "Script libraries and job parameters (optional)". Red arrows point from the surrounding text to specific fields in the form: from "Give your job a name: glue-tutorial-XXX" to the "Name" field; from "The language used to write the script" to the "ETL language" section; from "Give your script a name glue-tutorial-XXX" to the "Script file name" field; from "The location where your script will be placed in S3" to the "S3 path where the script is stored" field; from "Give your job a role to perform the actions necessary to run" to the "IAM role" dropdown; from "Create a new blank script" to the "A new script to be authored by you" radio button; and from "This is where a temporary script is generated when the script is being edited" to the "Temporary directory" field.

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 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

Dependent jars path

Referenced files path

Concurrent DPUs per job run ⓘ

Max concurrency ⓘ

Job timeout (minutes) ⓘ

Delay notification threshold (minutes) ⓘ

Number of retries

Job parameters

Key	Value
--REDSHIFT_DB_NAME	<input type="text" value="sales"/>
--SCHEMA_NAME	<input type="text" value="sales-jds"/>
--TABLE_NAME	<input type="text" value="products-jds"/>
--CATALOG_CONNECTION	<input type="text" value="glue-tutorial"/>
Type key...	<input type="text" value="Type value..."/>

Next

Parameters:

```
--REDSHIFT_DB_NAME  
    glue_tutorial_XXX  
--SCHEMA_NAME  
    sales-XXX  
--TABLE_NAME  
    products-XXX  
--CONNECTION_NAME  
    glue-tutorial-XXX
```



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

Connections

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

Showing: 1 - 1 < >

All connections

glue-tutorial-jds

Select

Showing: 0 - 0 < >

Required connections

No items selected

Add connection

Back

Next



Glue



— Create a Glue job

Job properties

Name	glue-tutorial-jds
IAM role	AWSGlueServiceRole-glueServiceRole
ETL language	python
Connections	glue-tutorial-jds
Path	s3://glue-tutorial-jds/glue-scripts/glue-tutorial-jds
Temporary directory	s3://aws-glue-temporary-952552944372-us-east-1/jack.silverman

▸ Advanced properties

▸ Script libraries and job parameters (optional)

Back

Save job and edit script



Glue



Writing the Script

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

```
1 import sys
2 from awsglue.transforms import *
3 from awsglue.utils import getResolvedOptions
4 from pyspark.context import SparkContext
5 from awsglue.context import GlueContext
6 from awsglue.dynamicframe import DynamicFrame
7 from awsglue.job import Job
8
9 args = getResolvedOptions(sys.argv, ['TempDir'])
10
11 sc = SparkContext()
12 glueContext = GlueContext(sc)
13 spark = glueContext.spark_session
14 job = Job(glueContext)
15 job.init(args['JOB_NAME'], args)
16
17
18
```

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
```

Include SQL
functions, types, and
datetime to use later

```
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)
```

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()
```

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("timestamp", 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





└─ 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





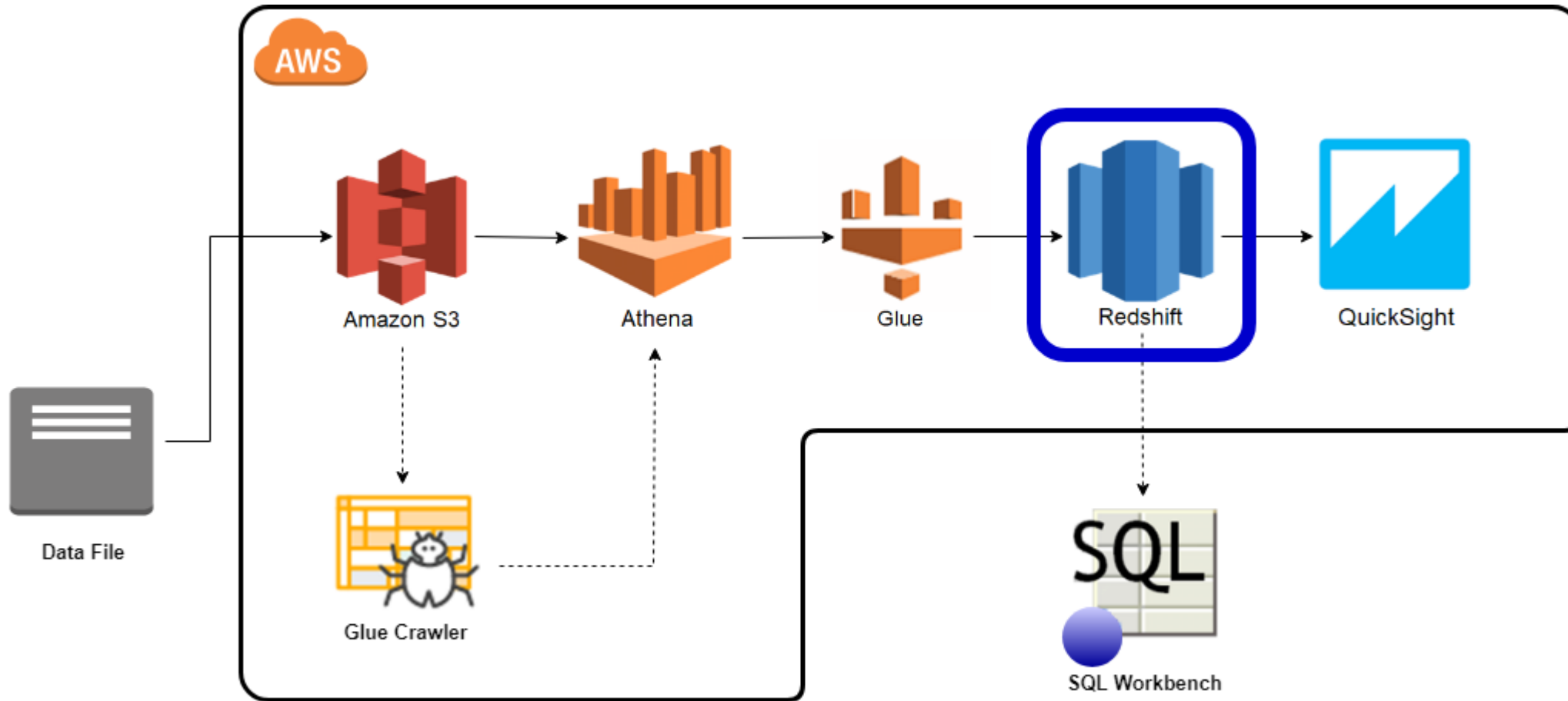
└ Writing the Script

```
...  
# 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



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
```





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. Job run events.

Add job

Action ▼

Filter by attributes

<input type="checkbox"/>	Name	Language	Script location
<input checked="" type="checkbox"/>	glue-tu		s3://glue-tutorial-jds/glue-
<input type="checkbox"/>	sales_		s3://manifest-glue-tutorial

Run job

Stop job run

Choose job triggers

Delete

Edit job

Edit script

Reset job bookmark

Create development endpoint

View run metrics

Script

Metrics

☒ glue-tutorial-jds
 python
 s3://glue-tutorial-jds

History

Details

Script

Metrics

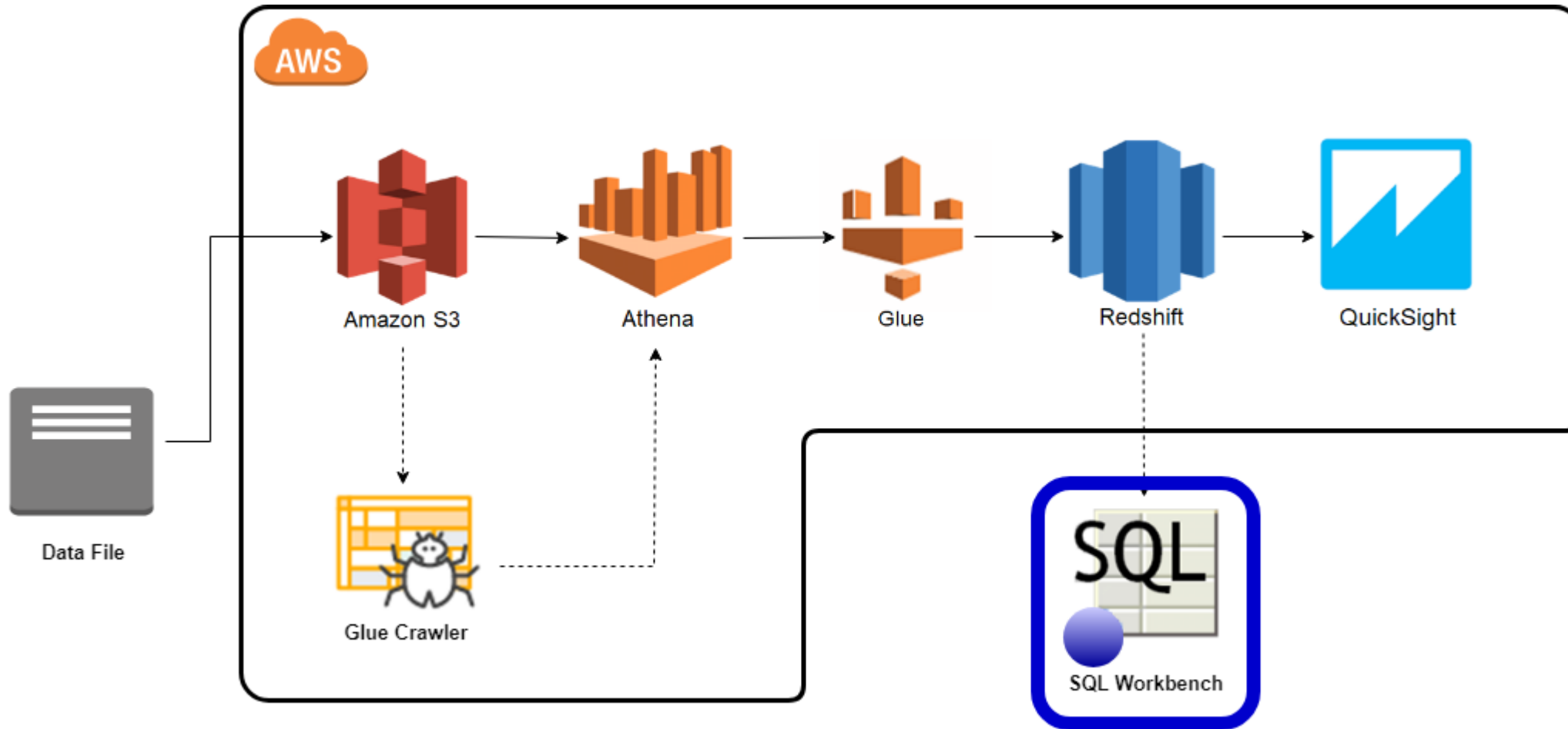
View run metrics

Run ID	Retry attempt	Run status	Error	Logs	Error log
<input type="radio"/> jr_1f154f2d70...	-	Succeeded		Logs	

When the job succeeds, check your Redshift table



SQL Workbench



Redshift



— Verify data in the table

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

2

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**

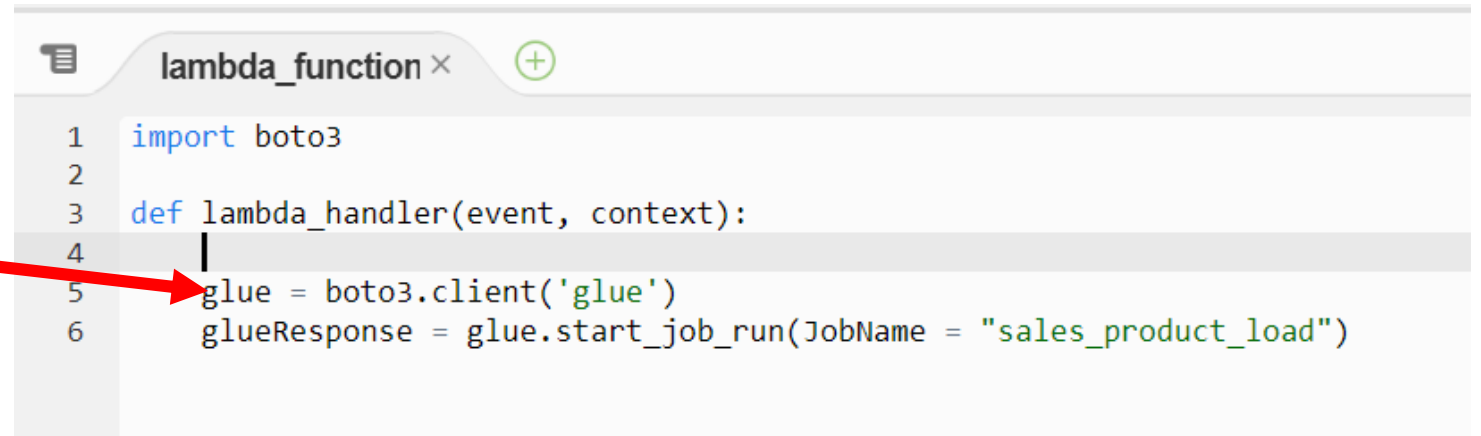
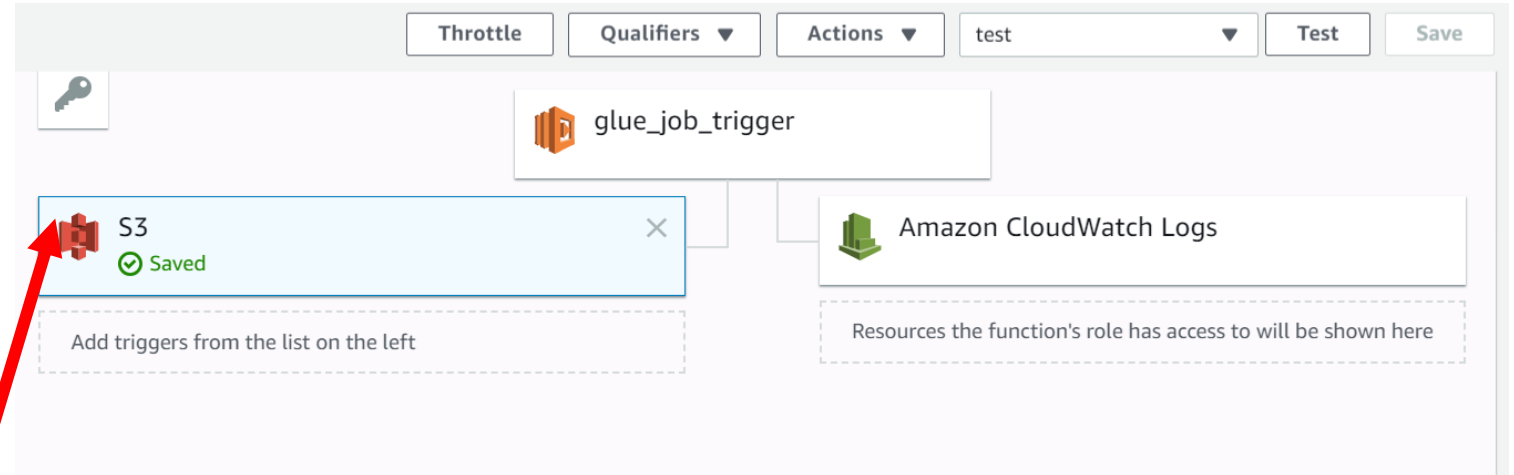
- **Create a Glue Trigger**
 - Automatically run the glue job
 - Run multiple different glue jobs
- **Control how resources can interact with other services**
- **Create reports for business analytics with the data that was loaded with the Glue job.**
- **Easily create, modify, and delete as well as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



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_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

<input type="checkbox"/>	filename ⓘ	glue_job ▾
<input type="checkbox"/>	WA_Sales_Products_2012-2014	sales_product_load



Glue Trigger

— Automatically run Glue job using Lambda

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

This returns the glue job associated with that file

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

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

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.

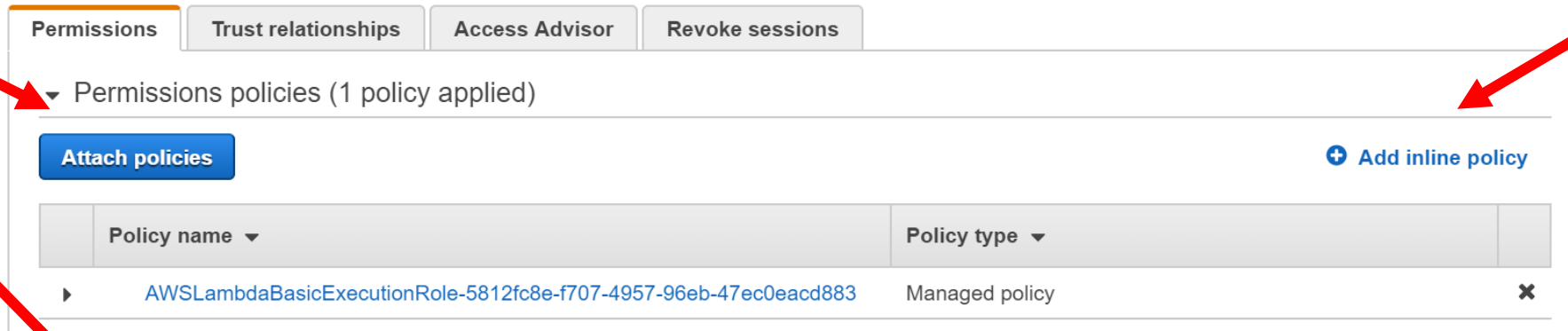
```
START RequestId: 2df6f8a8-95cb-11e8-aedb-510d0136df8b Version: $LATEST
An error occurred (AccessDeniedException) when calling the GetItem operation: User: arn:aws:sts::952552944372:assumed-role/lambda_basic_execution/glue_job_trigger is not authorized to perform: dynamodb:GetItem on resource: arn:aws:dynamodb:us-east-1:952552944372:table/glue_triggers: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

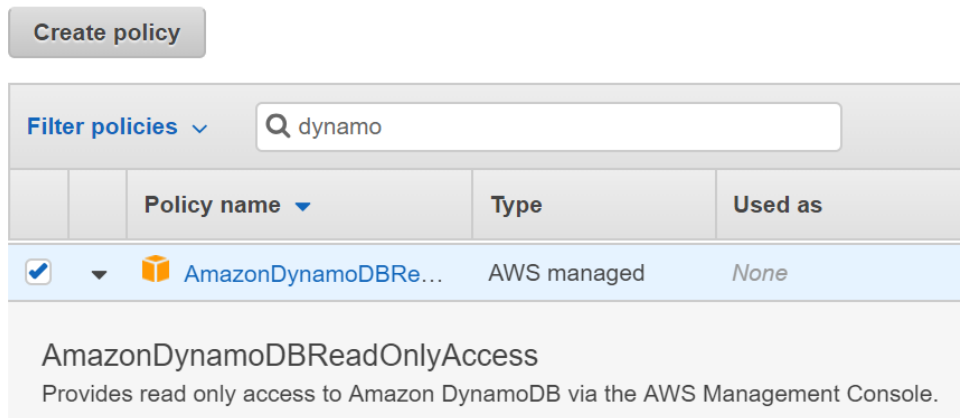
▼ Permissions policies (1 policy applied)

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

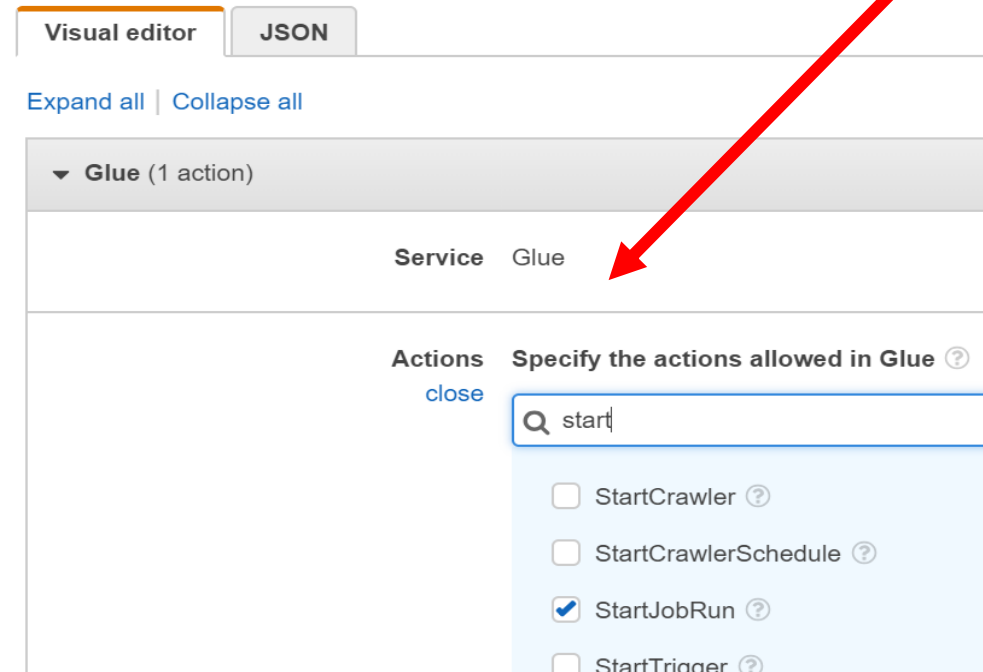


Create policy

Filter policies ▼ Q dynamo

	Policy name ▼	Type	Used as
<input checked="" type="checkbox"/>	 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

Q 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

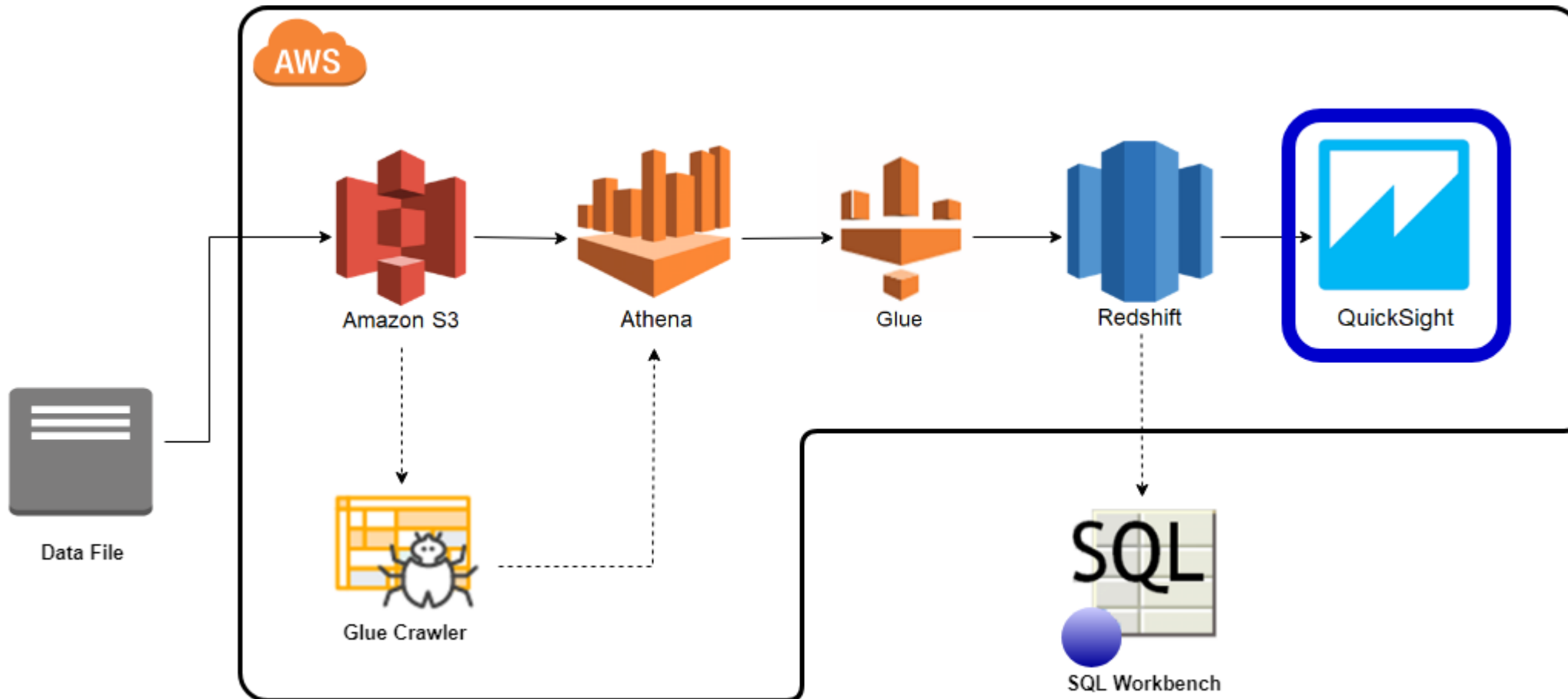
JavaScript

```
1  {
2
3    "Description" : "A text description for the template usage",
4
5    "Parameters": {
6
7      // A set of inputs used to customize the template per deployment
8
9    },
10
11    "Resources" : {
12
13      // The set of AWS resources and relationships between them
14
15    },
16
17    "Outputs" : {
18
19      // A set of values to be made visible to the stack creator
20
21    },
22
23    "AWSTemplateFormatVersion" : "2010-09-09"
24
25  }
```



QUICKSIGHT

— AWS Business Intelligence Tool

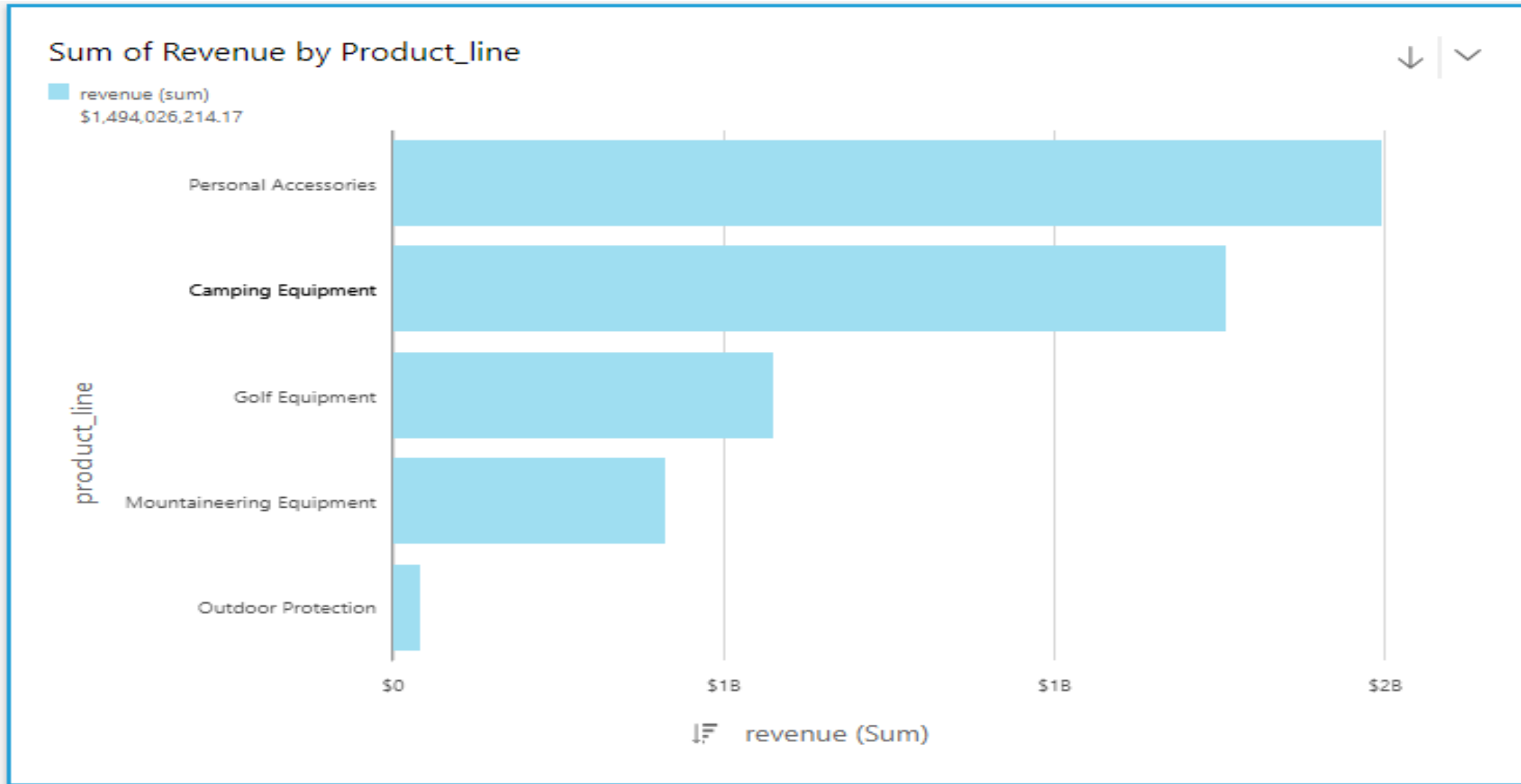


- Cloud based Business Intelligence reporting tool
- Build Reports from
 - Files in s3
 - Redshift
 - Athena



QUICKSIGHT

— AWS Business Intelligence Tool



Create Analysis

1. Create 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





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

AWS Account

952552944372

[Sign up for QuickSight](#)

To access QuickSight with a different account, [log in](#) again.



QUICKSIGHT

AWS Business Intelligence Tool

First author with 1GB SPICE	FREE	FREE
Team trial for 60 days (4 authors)*	FREE	FREE
Additional author per month (yearly)**	\$9	\$18
Additional author per month (monthly)**	\$12	\$24
Additional readers (Pay-per-Session)	N/A	\$0.30/session (max \$5/reader/month) ****
Additional SPICE per month	\$0.25 per GB	\$0.38 per GB
Single Sign On with SAML or OpenID Connect	✓	✓
Connect to spreadsheets, databases & business apps	✓	✓
Access data in Private VPCs		✓
Row-level security for dashboards		✓
Hourly refresh of SPICE data		✓
Secure data encryption at rest		✓
Connect to your Active Directory		✓
Use Active Directory Groups ***		✓
<p>* Trial authors are auto-converted to month-to-month subscription upon trial expiry</p> <p>** Each additional author includes 10GB of SPICE capacity</p> <p>*** Active Directory groups are available in accounts connected to Active Directory</p> <p>**** Sessions of 30-minute duration. Total charges for each reader are capped at \$5 per month. Conditions apply</p>		
<div>Continue</div>		



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

jsilverman@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
Enables QuickSight to auto-discover your Amazon S3 buckets

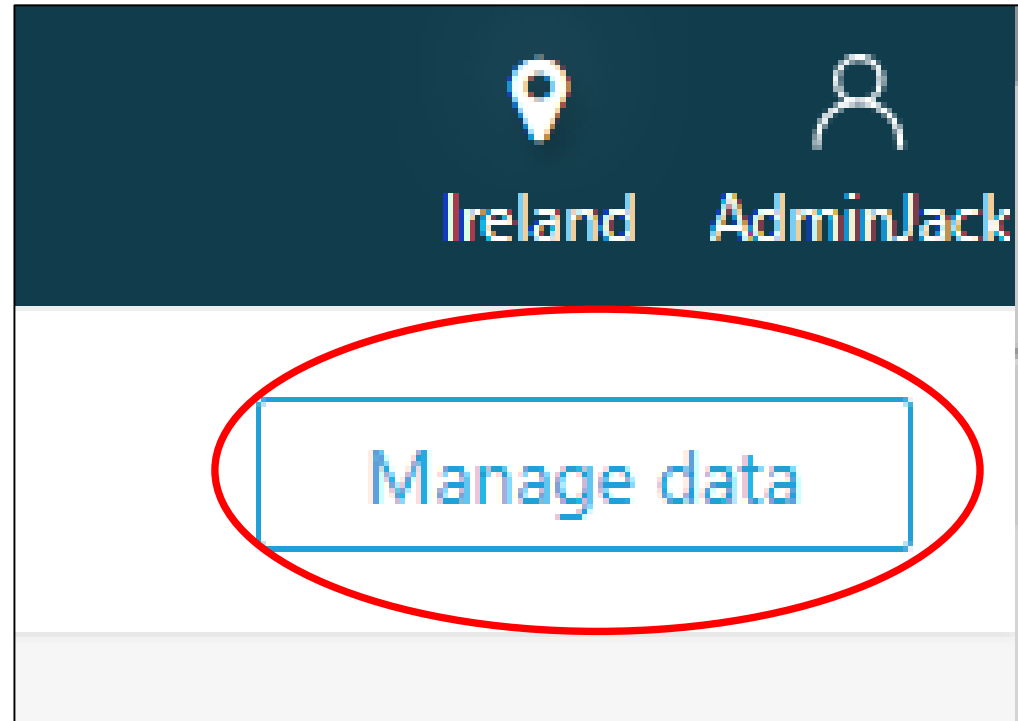
[Choose S3 buckets](#)

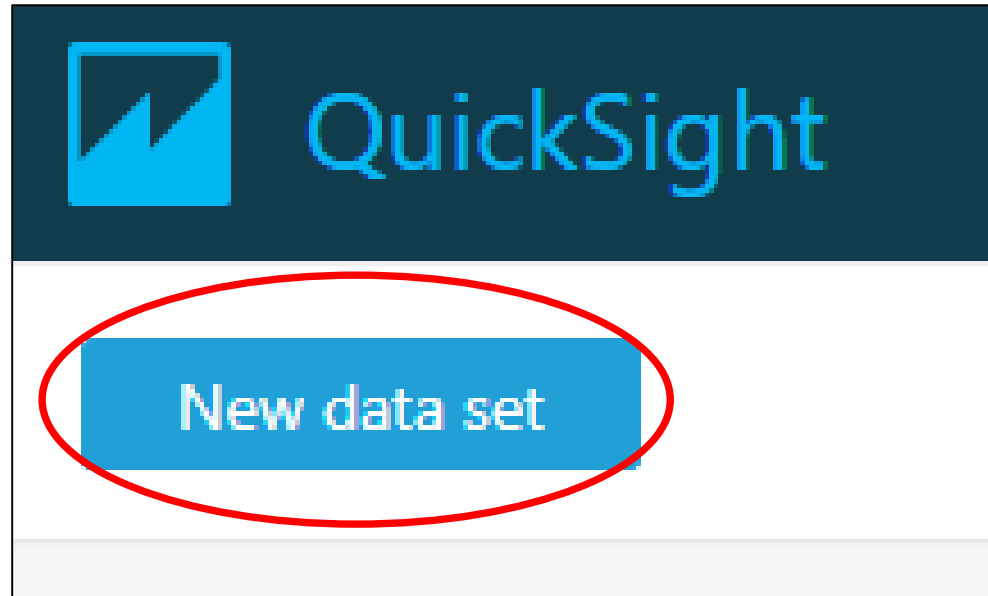
☐ Amazon S3 Storage Analytics
Enables QuickSight to visualize your S3 Storage Analytics data

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

Finish







Give you data set a name

This is the Redshift
endpoint without
port number

This information
comes from the
Redshift Cluster

New Redshift data source

Data source name

sales_jar

Connection type

Public network

Database server

glue-tutorial-jar.chtswcubv1n.eu-west-1.redshift.amazonaws.com

Port

5439

Database name

glue_tutorial

Username

master

Password

.....

Validated

SSL is enabled

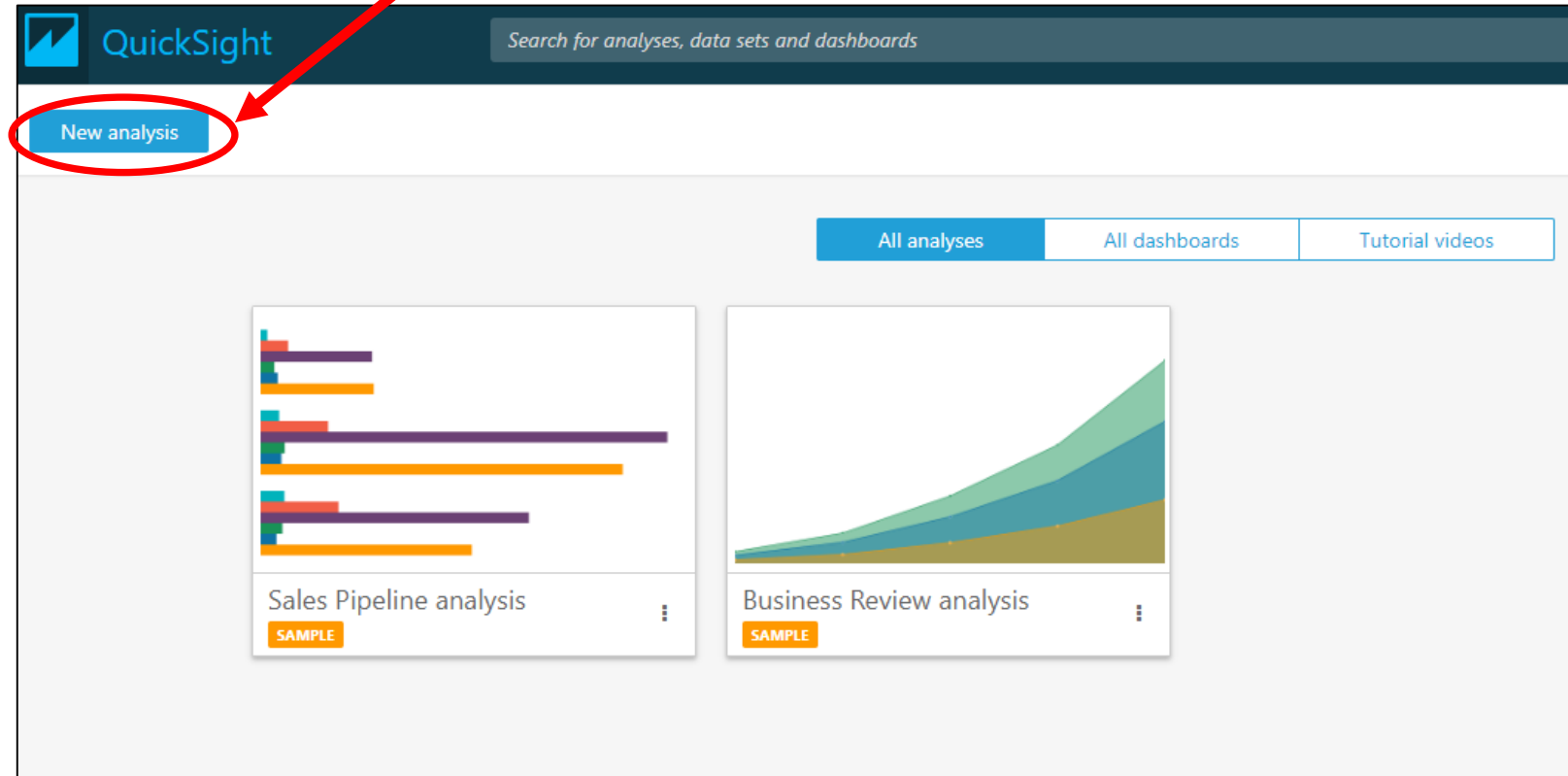
Create data source



QUICKSIGHT

— AWS Business Intelligence Tool

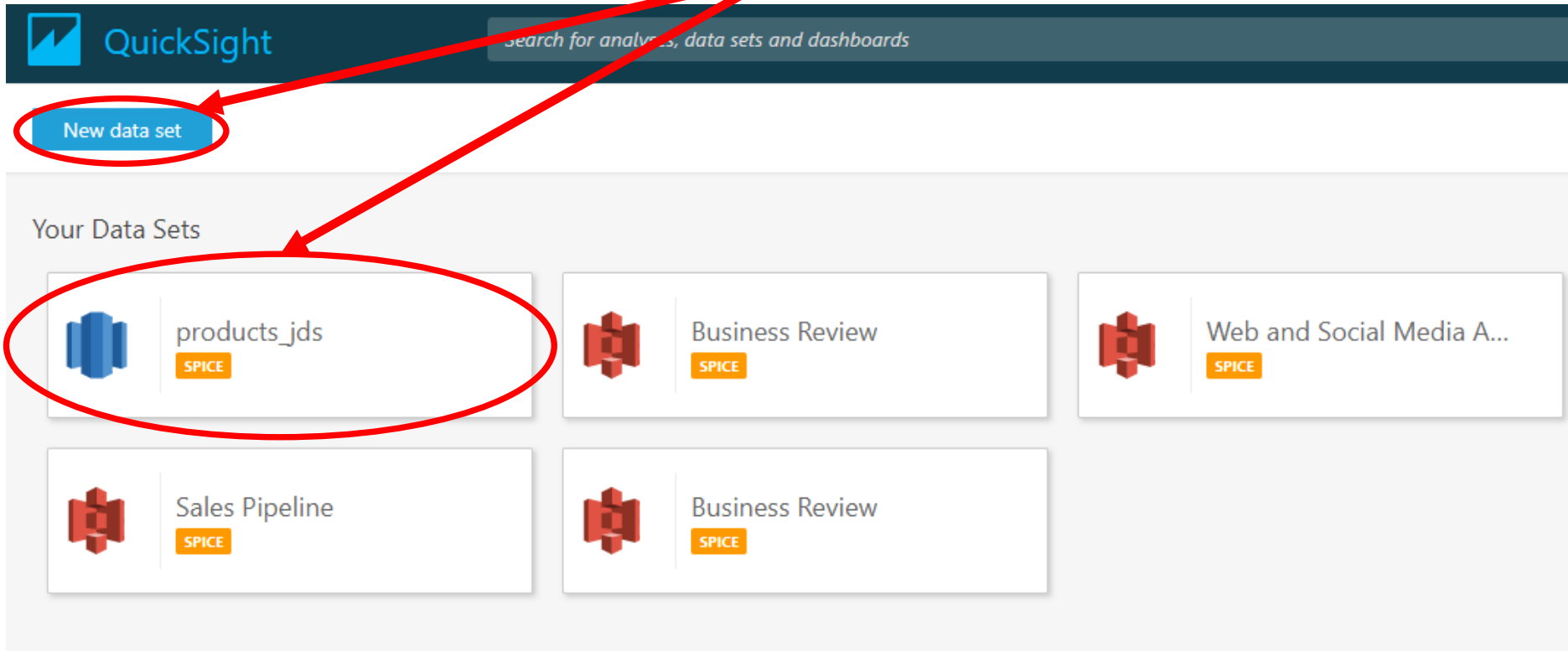
Select new Analysis



QUICKSIGHT

— AWS Business Intelligence Tool

Select product_jds (table in Redshift)



The screenshot shows the AWS QuickSight console interface. At the top, there is a dark blue header with the QuickSight logo and a search bar. Below the header, a red circle highlights the 'New data set' button. A red arrow points from the text 'Select product_jds (table in Redshift)' to this button. Another red arrow points from the same text to a card in the 'Your Data Sets' section. This card is also circled in red and contains the text 'products_jds' and a 'SPICE' label. Other cards in the section include 'Business Review', 'Web and Social Media A...', 'Sales Pipeline', and another 'Business Review'.

QuickSight Search for analytics, data sets and dashboards


New data set

Your Data Sets

- products_jds SPICE
- Business Review SPICE
- Web and Social Media A... SPICE
- Sales Pipeline SPICE
- Business Review SPICE



Select Create Analysis

 products_jds ×

SPICE Data Set 26MB

Import complete:
100% success
88475 rows were imported to SPICE
0 rows were skipped
[View summary](#)

Last refreshed: 20 hours ago

[Refresh Now](#) [Schedule refresh](#)

Data source name: Manifest Redshift
Database name: sales

[Delete data set](#) [Share](#)

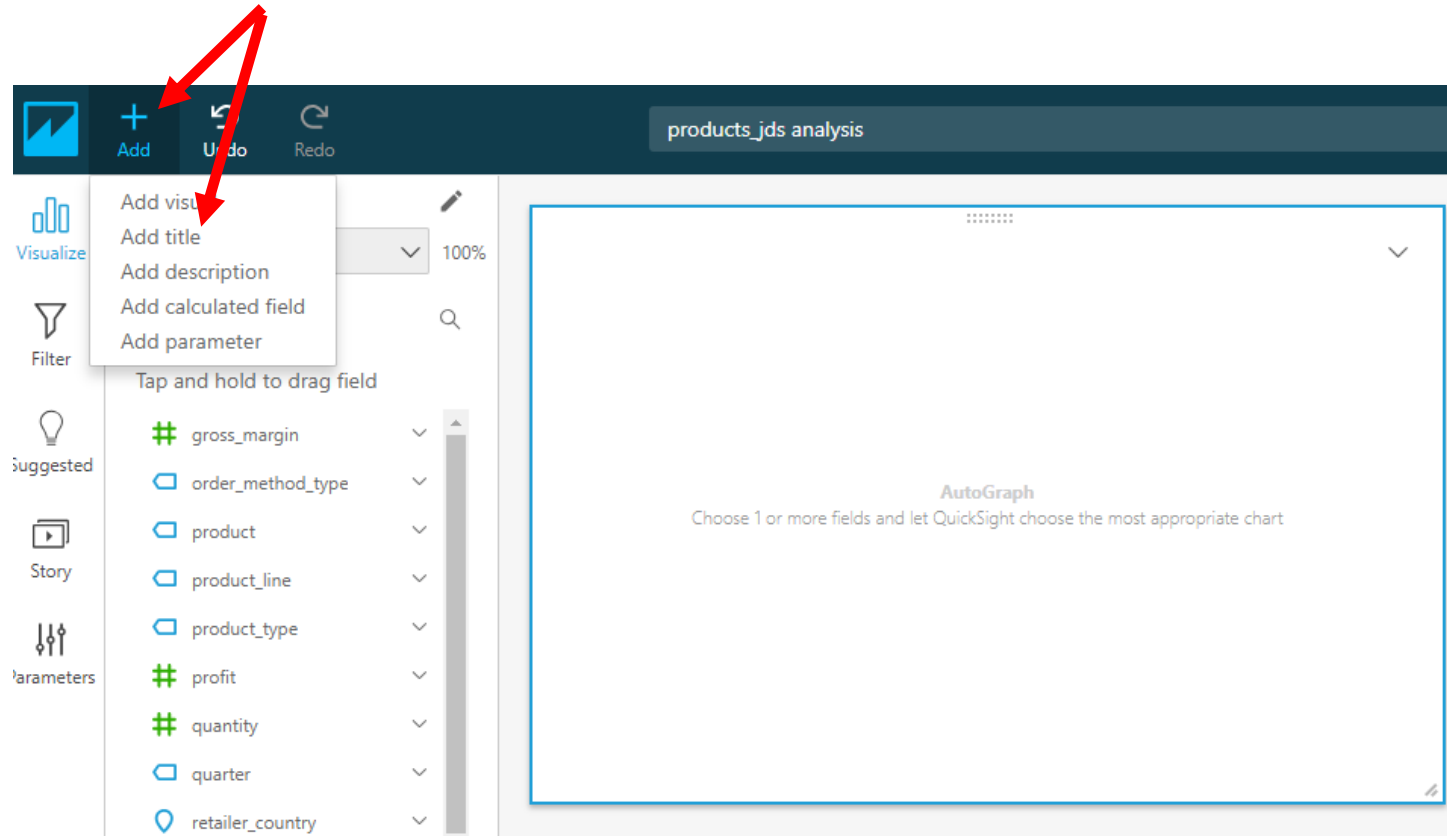
[Edit data set](#) [Duplicate data set](#) [Create analysis](#)



QUICKSIGHT

— AWS Business Intelligence Tool

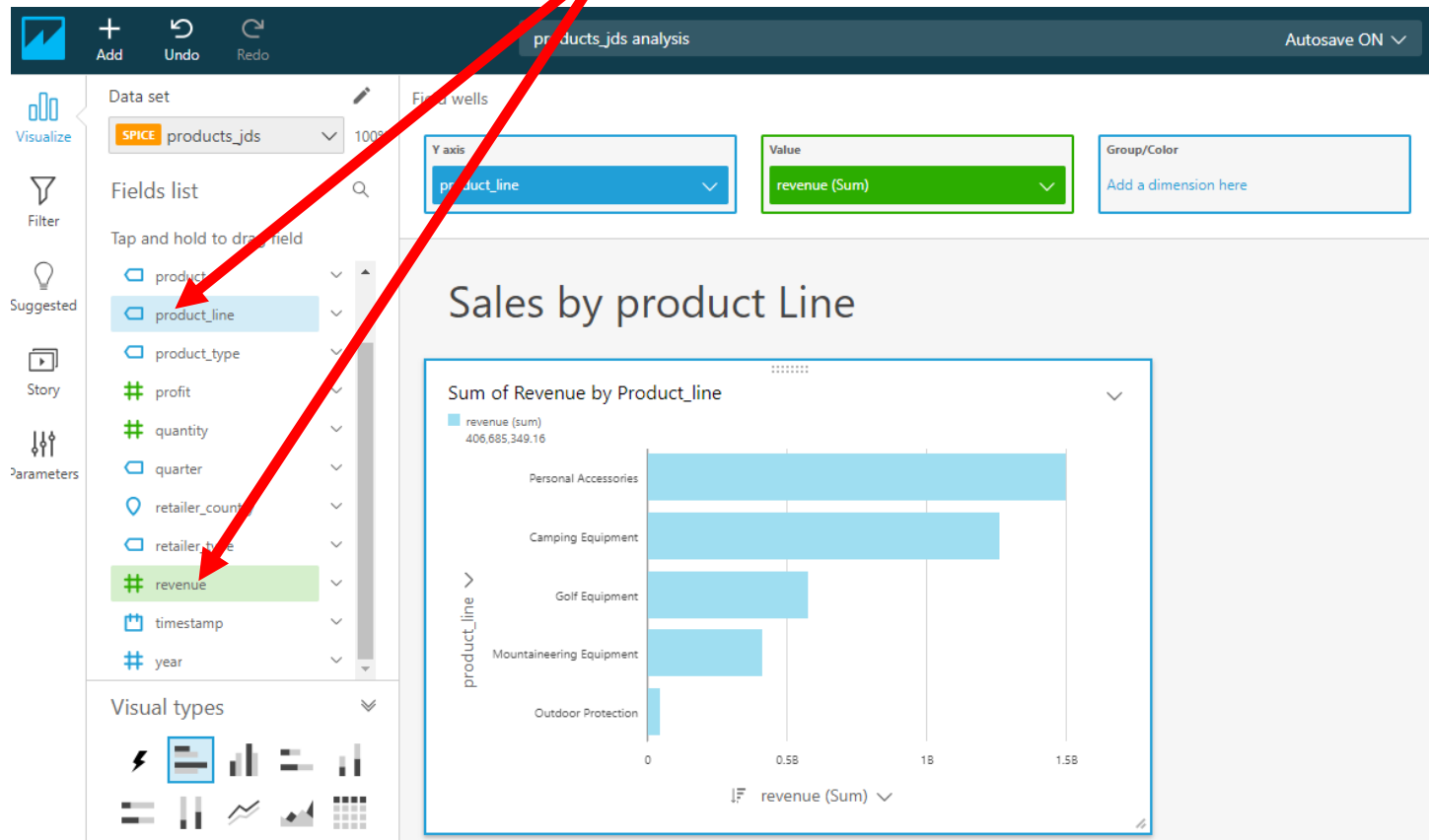
Select Add > Add Title



QUICKSIGHT

AWS Business Intelligence Tool

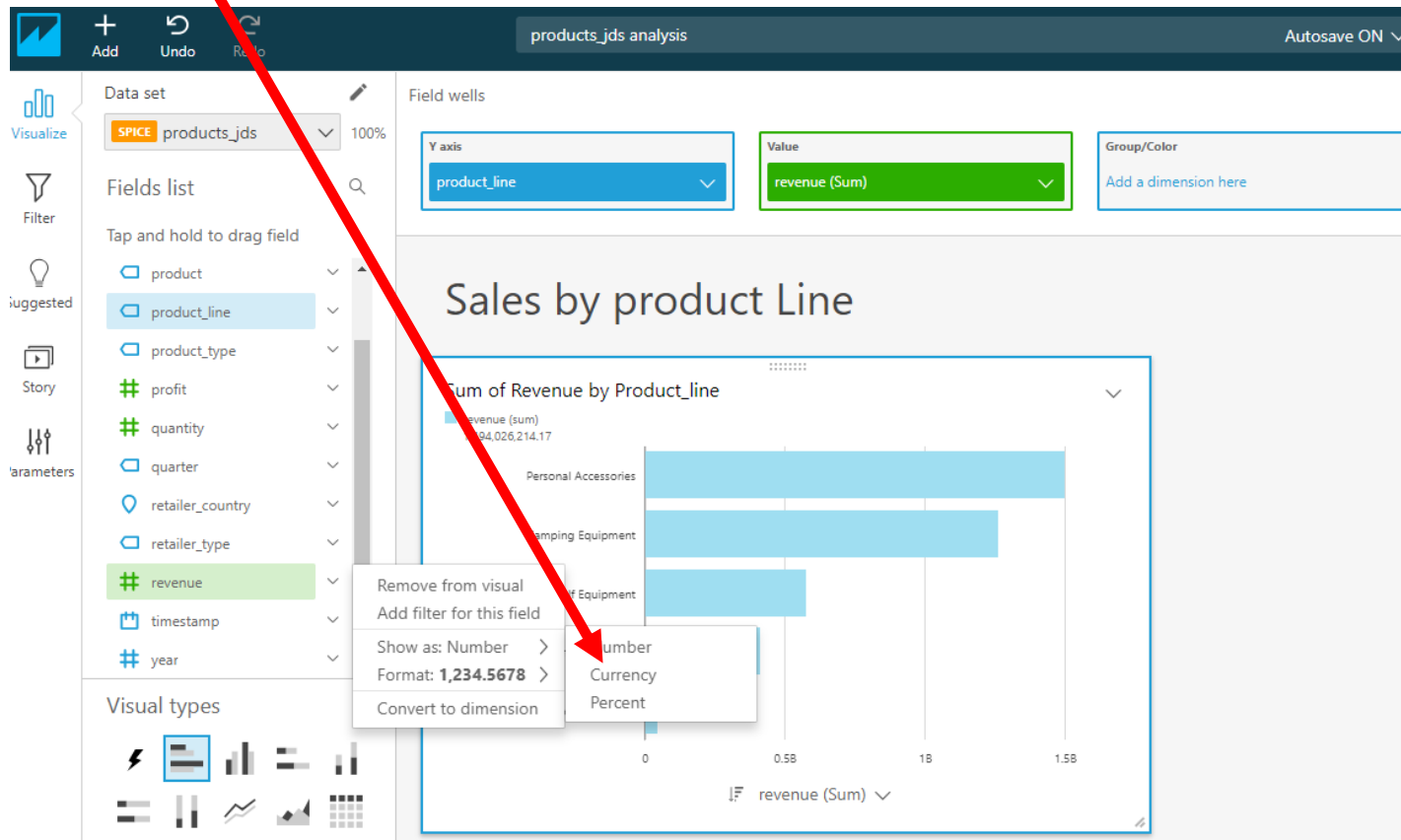
Choose product_line and revenue



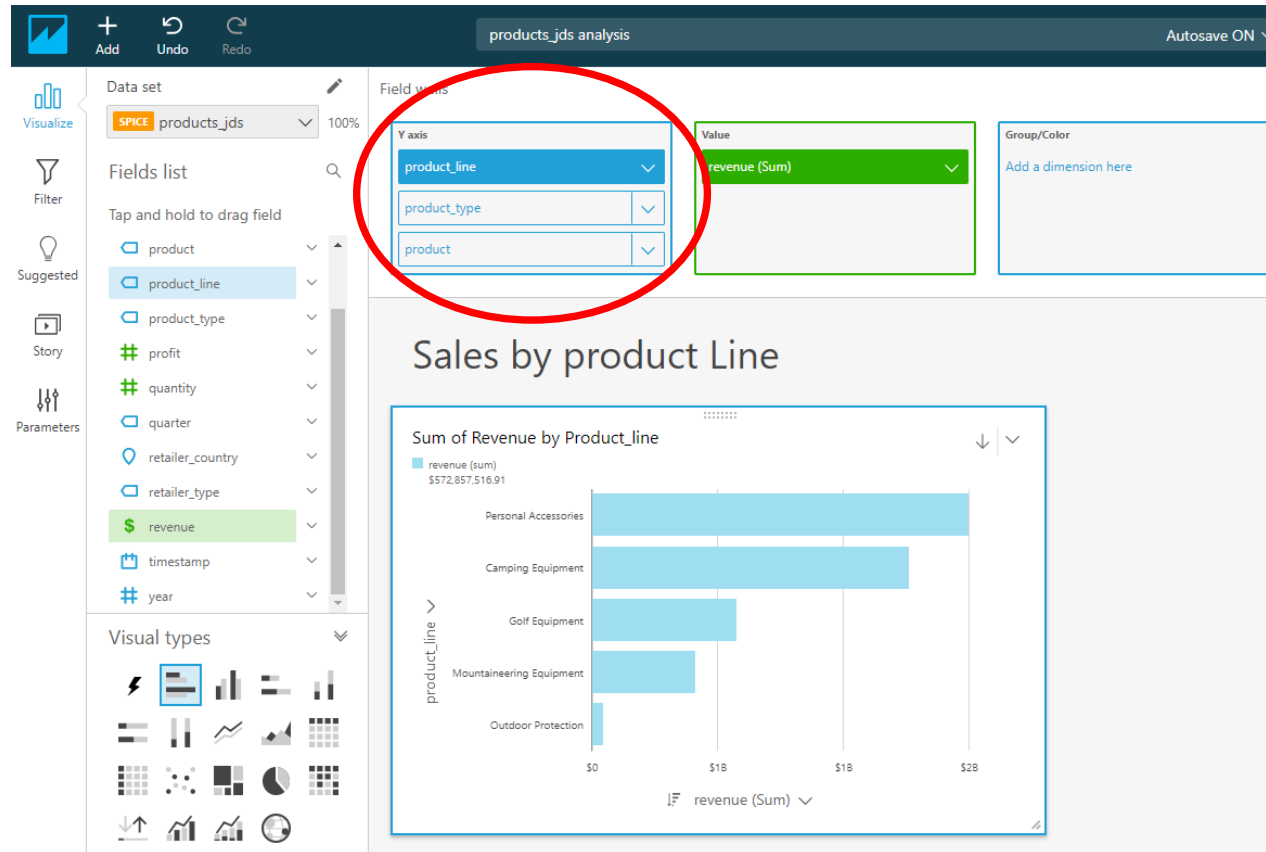
QUICKSIGHT

AWS Business Intelligence Tool

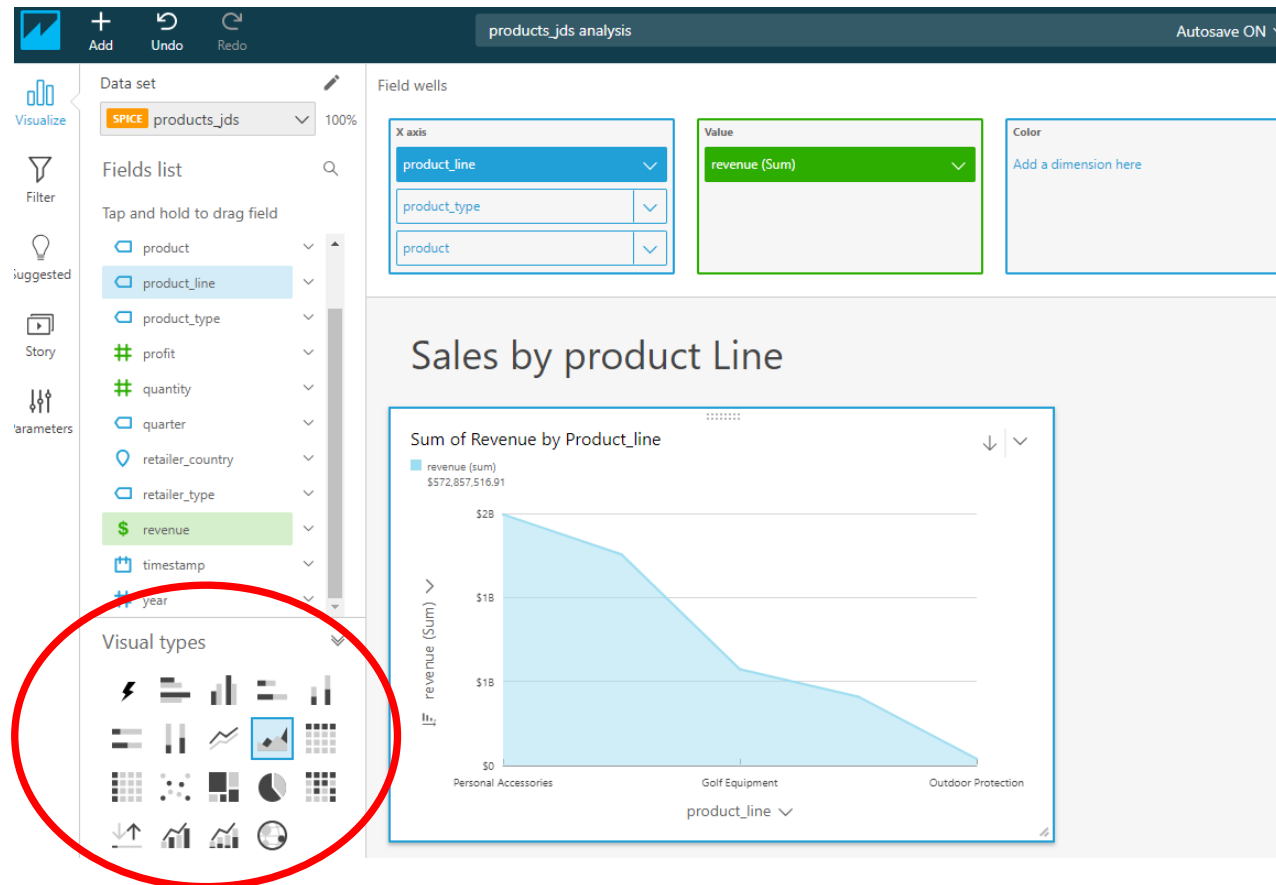
Change the format of Revenue to Currency



Add product_type and product as drill down layer



Change Visual Type

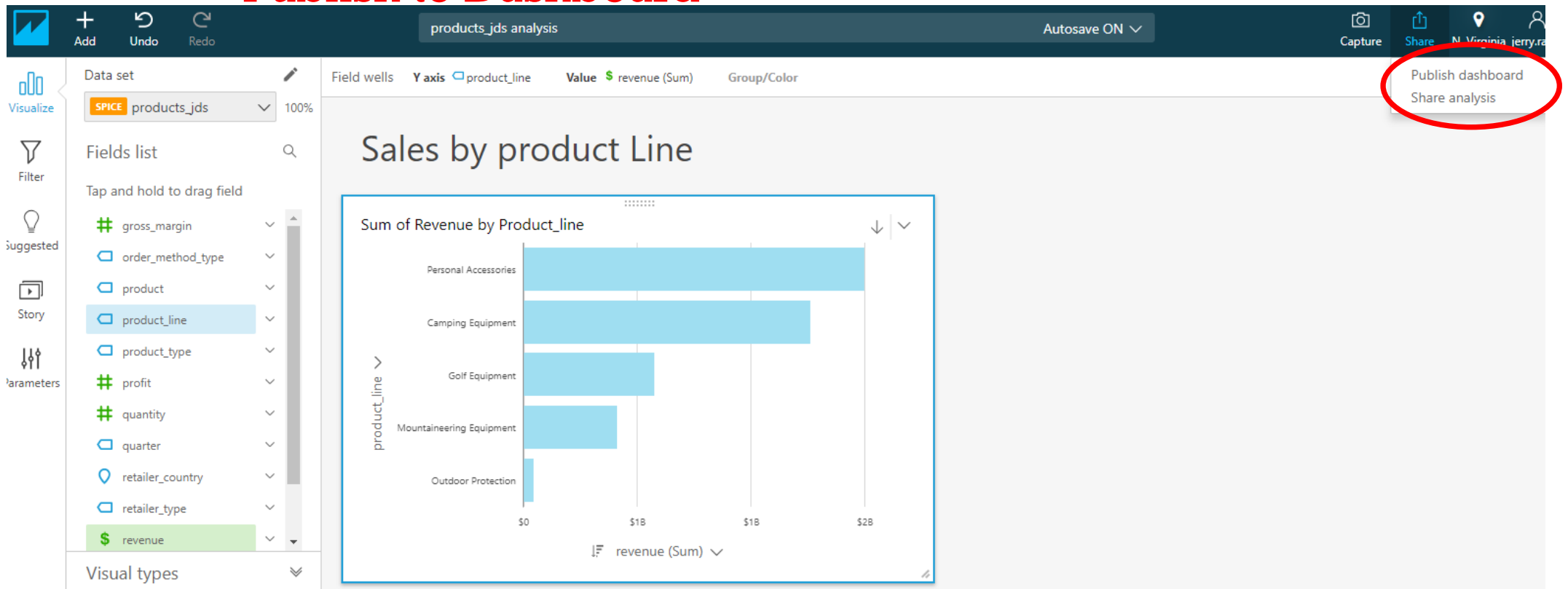


QUICKSIGHT

AWS Business Intelligence Tool



Publish to Dashboard



Name the Dashboard and select Publish dashboard

The screenshot shows the AWS QuickSight interface with a dashboard titled "Sales by Product Line". A modal dialog titled "Publish a dashboard" is open in the center. The dialog has two radio button options: "Publish new dashboard as" (which is selected) and "Replace an existing dashboard". Below the first option is a text input field containing "Sales by Product Line". Below the second option is a dropdown menu. At the bottom of the dialog are two buttons: "Cancel" and "Publish dashboard". The "Publish dashboard" button is circled in red. In the background, a bar chart is visible showing revenue by product line.

product_line	Sum of Revenue
Golf Equipment	\$18
Mountaineering Equipment	\$18
Outdoor Protection	\$28



Share the dashboard

Share dashboard with users

Select users in this account.

Search by email address

☐ Share with all users in this account

Name	Email	Permission	Role
------	-------	------------	------

Manage dashboard access

Share

Sum of Revenue by Product_line

product_line

Personal Accessories

Camping Equipment

Golf Equipment

Mountaineering Equipment

Outdoor Protection

revenue (Sum)



Lab 4

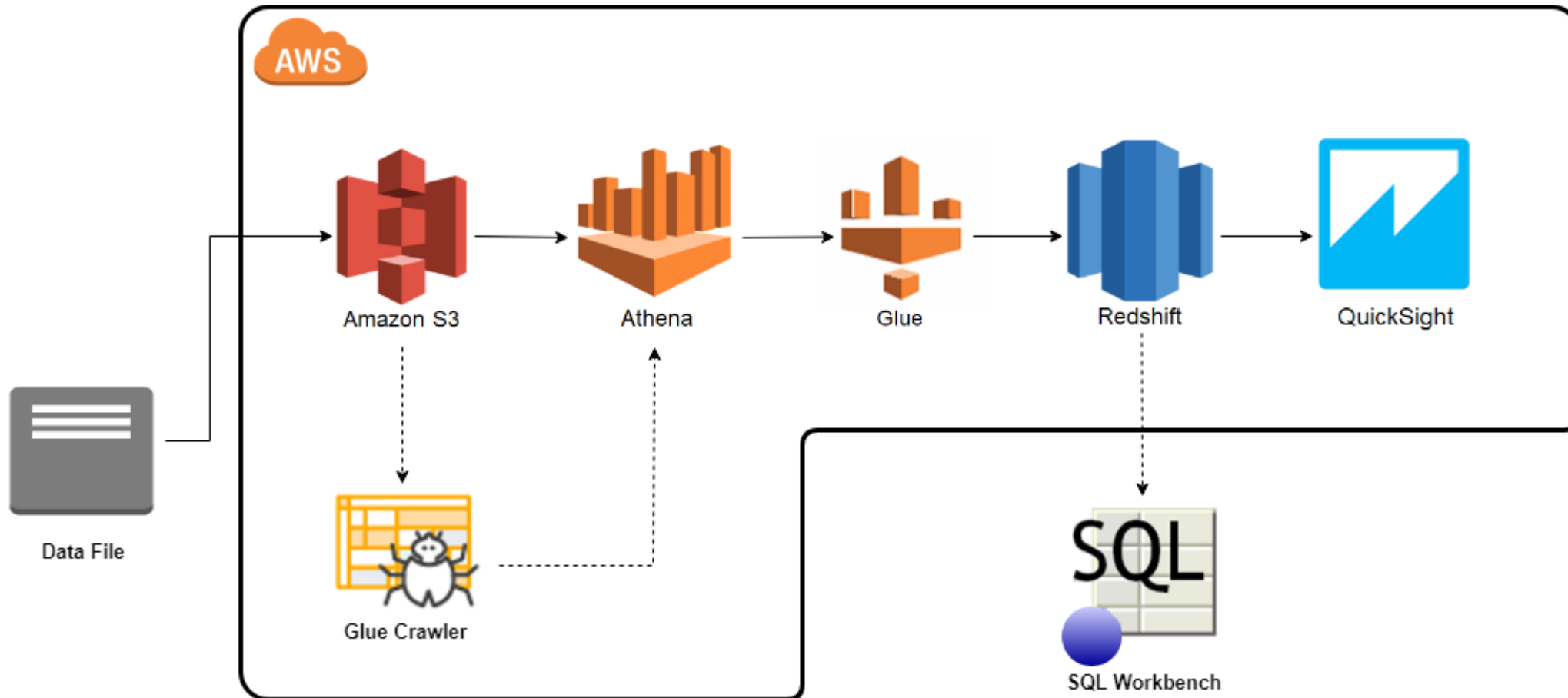
- Create Quicksight Account
- Create Dataset
- Create Analysis
- Publish to Dashboard

(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/>

