

Unit 5.2 Graded Assignment:

Syed Muhammad Raqim Ali Shah (2303.KHI.DEG.008)

Maaz Javaid Siddique (2303.KHI.DEG.004)

Qadeer Hussain (2303.KHI.DEG.006)

Daily Assignment :

Using the salary CSV as a base, prepare a new data file with employees' office locations. Make sure there are 5-6 distinct locations that are shared between employees.

Create a Glue job that aggregates the data based on the office location to calculate average salaries and raise percentages for these locations.

Answer:

1. Set up an AWS Glue job in the AWS Glue console

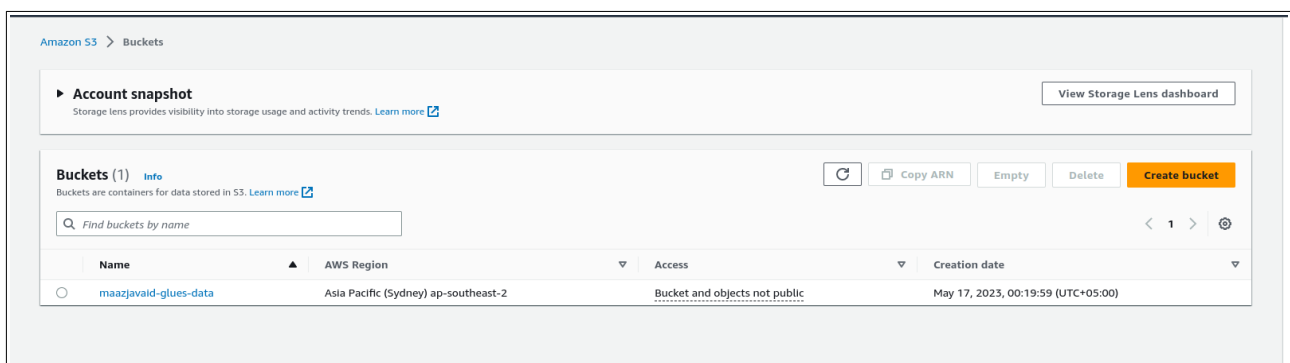
First of all we create the S3 bucket:

Q1. Why do we need to take this step?

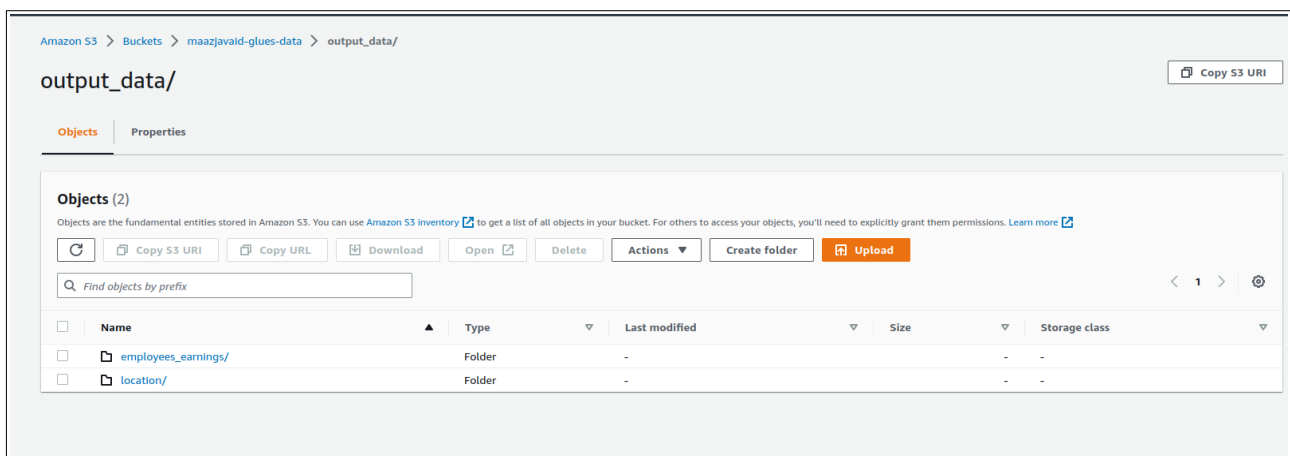
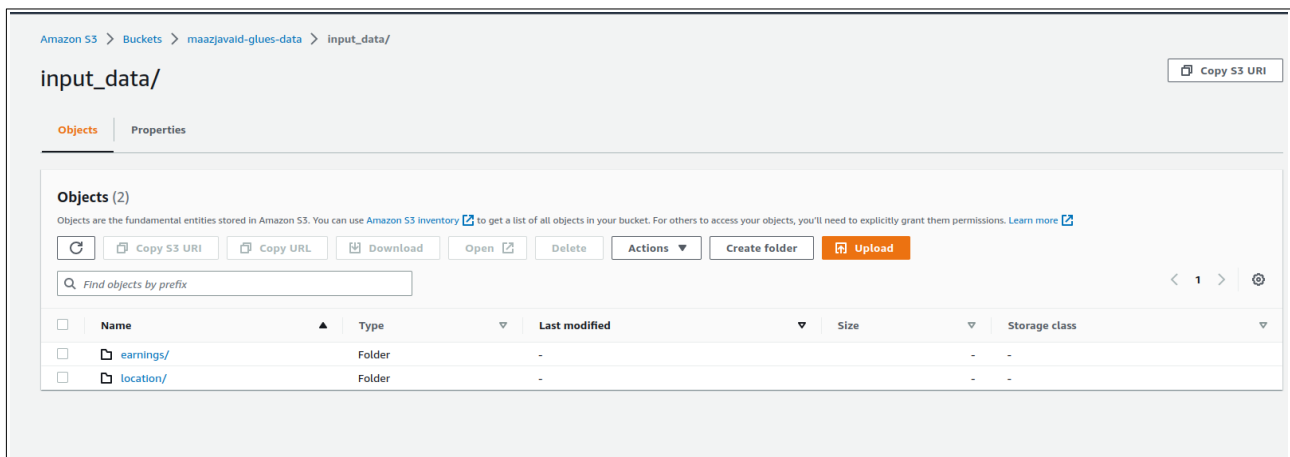
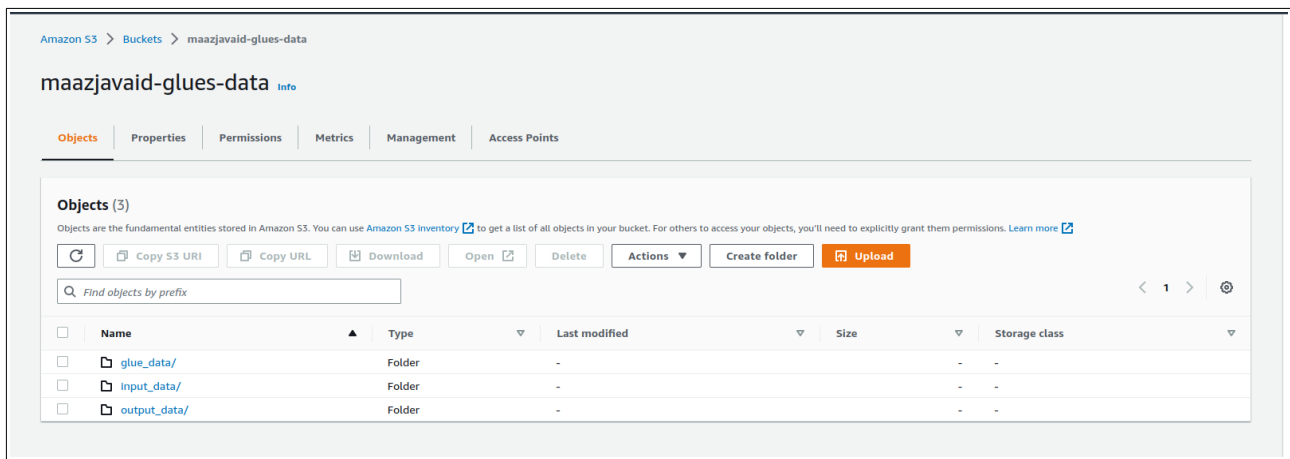
A. We need to specify the input and output data locations for the job. Typically, the data source (e.g. CSV file) and output data generated by the Glue job are stored in an S3 bucket.

Q2. What is this service's purpose?

A. S3 provides features such as durability, availability, security, and scalability. Durability ensures that data is stored redundantly across multiple devices in multiple facilities, making it highly resilient to data loss.



Then we create the directories on S3 bucket:



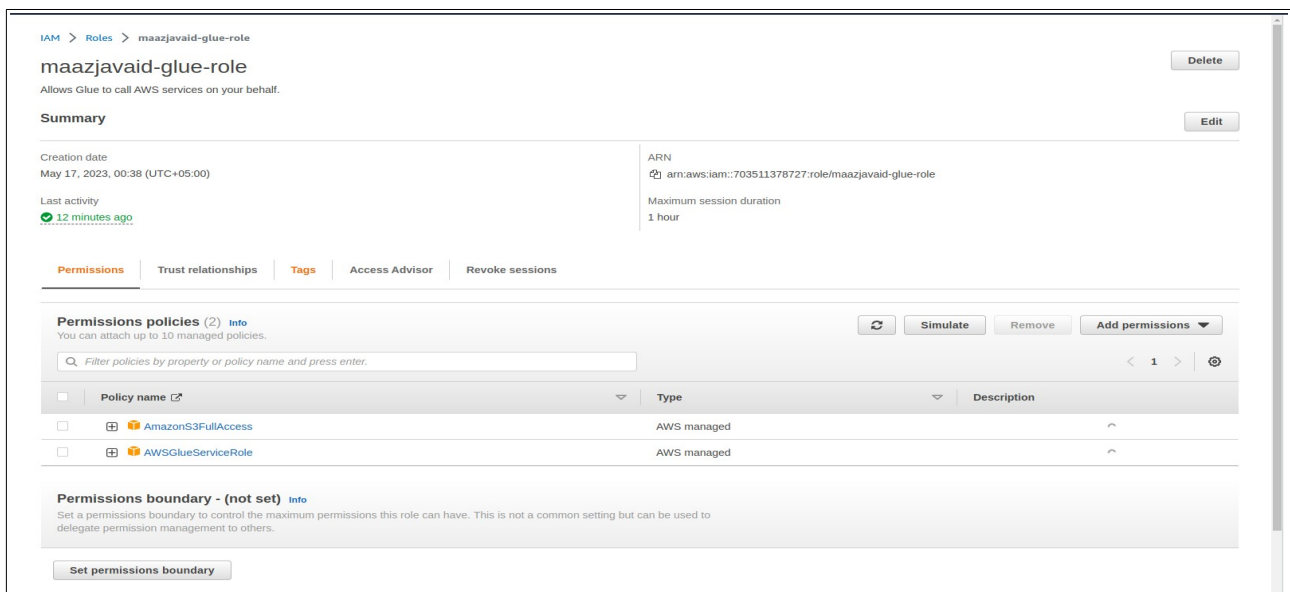
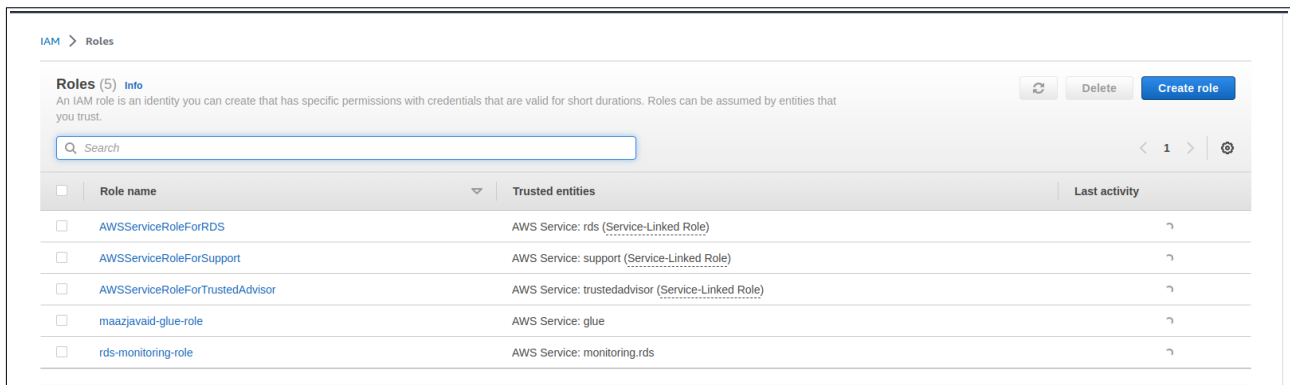
Then we go to IAM service and create the IAM Role:

Q1. Why do we need to take this step?

A. We need an IAM role to grant the Glue job permissions to access the input data stored in S3 and write the output data back to S3. The IAM role specifies the permissions and access controls that the Glue job can use when interacting with S3.

Q2. What is this service's purpose?

A. IAM provides several features such as access control, multi-factor authentication, password policies, and integration with other AWS services. With IAM, we can also create and manage IAM roles, which are a secure way to grant permissions to AWS services or resources without the need for access keys.



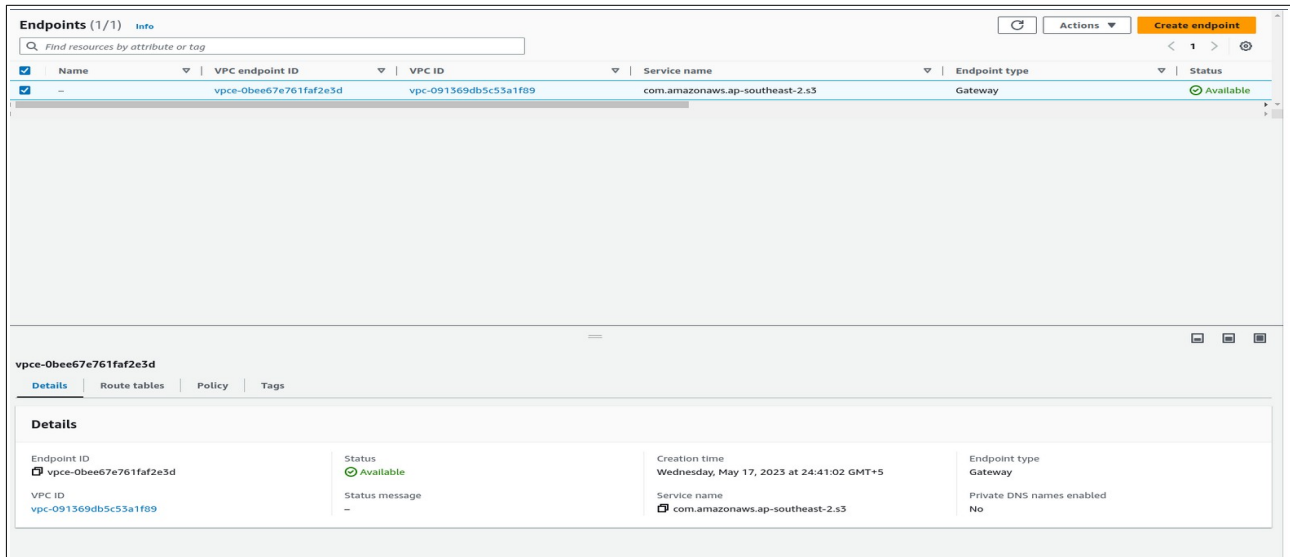
Create the VPC Endpoint:

Q1. Why do we need to take this step?

A. We need to create a VPC endpoint to allow the Glue job to access S3 securely. This is necessary because the input and output data for the Glue job are typically stored in an S3 bucket, which is an AWS service that resides in a VPC.

Q2. What is this service's purpose?

A. When we create a VPC endpoint, we can define which services or endpoint services are accessible from our VPC, as well as which VPCs and subnets are allowed to access the endpoint. VPC endpoints support traffic encryption using AWS PrivateLink and allow us to control access using VPC security groups and IAM policies.



Security group rules:

Q1. Why do we need to take this step?

A. By defining inbound and outbound traffic rules in security groups, we can control the network traffic to and from our AWS resources, such as EC2 instances, RDS databases, or load balancers. This helps to reduce the risk of unauthorized access, data breaches, and other security threats.

Q2. What is this service's purpose?

A. When we create security group rules, we define inbound and outbound traffic rules that control the network traffic to and from our AWS resources. This helps to protect our resources from unauthorized access, data breaches, and other security threats.

Security Groups (1/1) Info

Filter security groups

<input checked="" type="checkbox"/>	Name	Security group ID	Security group name	VPC ID	Description	Owner	Inbound rules count	Outbound rules co...
<input checked="" type="checkbox"/>	-	sg-0e167da7970bf2f2c	default	vpc-091369db5c53a1f89	default VPC security gr...	703511378727	2 Permission entries	1 Permission entry

sg-0e167da7970bf2f2c - default

Details | **Inbound rules** | Outbound rules | Tags

You can now check network connectivity with Reachability Analyzer [Run Reachability Analyzer](#)

Inbound rules (2)

Filter security group rules

<input type="checkbox"/>	Name	Security group rule...	IP version	Type	Protocol	Port range	Source	Description
<input type="checkbox"/>	-	sgr-0b6c7e430b90128...	IPv4	PostgreSQL	TCP	5432	0.0.0.0/0	Rule for external access
<input type="checkbox"/>	-	sgr-012be979989bb86ff	-	All TCP	TCP	0 - 65535	sg-0e167da7970bf2f2...	Rule for Glue Crawler

Security Groups (1/1) Info

Filter security groups

<input checked="" type="checkbox"/>	Name	Security group ID	Security group name	VPC ID	Description	Owner	Inbound rules count	Outbound rules co...
<input checked="" type="checkbox"/>	-	sg-0e167da7970bf2f2c	default	vpc-091369db5c53a1f89	default VPC security gr...	703511378727	2 Permission entries	1 Permission entry

sg-0e167da7970bf2f2c - default

Details | Inbound rules | **Outbound rules** | Tags

You can now check network connectivity with Reachability Analyzer [Run Reachability Analyzer](#)

Outbound rules (1/1)

Filter security group rules

<input checked="" type="checkbox"/>	Name	Security group rule...	IP version	Type	Protocol	Port range	Destination	Description
<input checked="" type="checkbox"/>	-	sgr-0114bab7e861b6c...	IPv4	All traffic	All	All	0.0.0.0/0	-

2.ETL in Glue

Create database:

Q1. Why do we need to take this step?

A. Creating a database in AWS Glue is an important step in preparing our data for ETL processing, as it can help with data organization, query performance, data discovery, and access control.

Q2. What is this service's purpose?

A. The purpose of creating a database in AWS Glue is to organize and manage our data sources and prepare them for ETL processing.

AWS Glue

Databases

Databases (1)

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

Filter databases

Name

▲

Description

▼

Location URI

▼

Created on (UTC)

▼

maazjavaid_glue_database

-

-

May 16, 2023 at 19:50:39

Last updated (UTC)
May 17, 2023 at 07:43:58

Edit

Delete

Add database

<

1

>

AWS Glue

>

Databases

>

maazjavaid_glue_database

maazjavaid_glue_database

Last updated (UTC)

May 17, 2023 at 07:44:07

Edit

Delete

Database properties

Name

maazjavaid_glue_database

Description

-

Location

-

Created on (UTC)

May 16, 2023 at 19:50:39

Tables (4)

View and manage all available tables.

Last updated (UTC)

May 17, 2023 at 07:44:10

Delete

Add tables using crawler

Add table

Q

Filter tables

<

1

>

<input type="checkbox"/>	Name	▲	Database	▼	Location	▼	Classification	▼	Deprecated	▼	View data
<input type="checkbox"/>	average_earnings		maazjavaid_glue_database		s3://maazjavaid-glues-data/output_d		parquet		-		Table data
<input type="checkbox"/>	maazjavaid_earnings		maazjavaid_glue_database		s3://maazjavaid-glues-data/input_dat		csv		-		Table data
<input type="checkbox"/>	maazjavaid_employees_earnings		maazjavaid_glue_database		s3://maazjavaid-glues-data/output_d		parquet		-		Table data
<input type="checkbox"/>	maazjavaid_location		maazjavaid_glue_database		s3://maazjavaid-glues-data/input_dat		csv		-		Table data

Creating Glue Crawlers:

Q1. Why do we need to take this step?

A. Creating Glue Crawlers in AWS Glue is an important step in preparing our data sources for ETL processing. It simplifies and automates the process of discovering and cataloging metadata about our data sources, which can save time, reduce errors, and improve data quality.

Q2. What is this service's purpose?

A. Glue Crawlers can also infer the schema of our data sources, which can simplify and automate the process of preparing data for ETL processing. By automatically discovering and cataloging metadata about our data sources, Glue Crawlers can help reduce errors and improve data quality.

AWS Glue

>

Crawlers

Crawlers

A crawler connects to a data store, progresses through a prioritized list of classifiers to determine the schema for your data, and then creates metadata tables in your data catalog.

Crawlers (3) info

View and manage all available crawlers.

Filter crawlers

Last updated (UTC)
May 17, 2023 at 07:44:43

Action

Run

Create crawler

<input type="checkbox"/>	Name		State		Schedule		Last run		Last run timestamp		Log	Table changes from last r...
<input type="checkbox"/>	maazjavaid_rds_employees...		Ready				Failed		May 17, 2023 at 05:11:51		View log	-
<input type="checkbox"/>	maazjavaid_s3_earnings_cr...		Ready				Succeeded		May 17, 2023 at 05:11:51		View log	1 updated
<input type="checkbox"/>	maazjavaid_s3_office_locat...		Ready				Succeeded		May 17, 2023 at 05:29:01		View log	1 created

Creating job:

Q1. Why do we need to take this step?

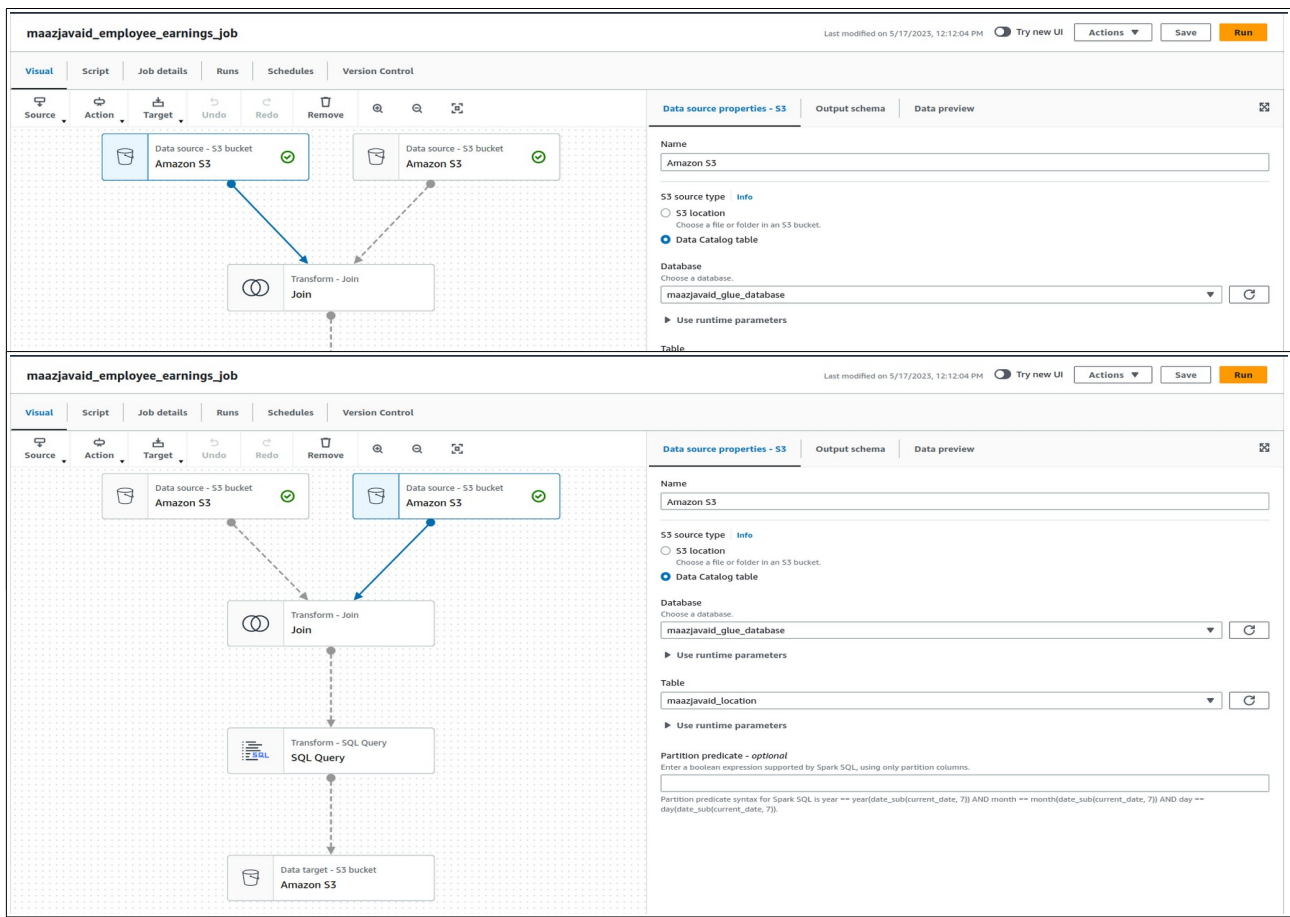
A. Creating a job in AWS Glue is necessary to define the ETL process that we want to run on our data. By creating a job, we can specify the data source that we want to use, the ETL script that we want to run, and the output data location where we want to write the transformed data.

Q2. What is this service's purpose?

A. AWS Glue's purpose is to simplify the ETL process and make it accessible to a wider range of users, regardless of their technical expertise. It provides a powerful and flexible platform for processing data at scale and preparing it for analysis, machine learning, and other downstream applications.

The screenshot shows the 'Create job' interface in AWS Glue Studio. At the top, there's a 'Create job' button and a 'Visual with a source and target' option selected. Below this, there are four tabs: 'Visual with a source and target', 'Visual with a blank canvas', 'Spark script editor', and 'Python Shell script editor'. The 'Visual with a source and target' tab is active, showing a 'Source' dropdown set to 'Amazon S3' and a 'Target' dropdown also set to 'Amazon S3'. Below the dropdowns, there's a 'Your jobs' section with a table of existing jobs.

Job name	Type	Last modified	AWS Glue version
maazjavaidd_employee_earnings_job	Glue ETL	5/17/2023, 12:12:04 PM	3.0



maazjavaid_employee_earnings_job

Last modified on 5/17/2023, 12:12:04 PMTry new UIActionsSaveRun

VisualScriptJob detailsRunsSchedulesVersion Control

SourceActionTargetUndoRedoRemove

Data source - S3 bucket
Amazon S3

Data source - S3 bucket
Amazon S3

Transform - Join
Join

Transform - SQL Query
SQL Query

Data target - S3 bucket
Amazon S3

Transform

Output schema

Data preview

Name

Join

Node parents

Choose which nodes will provide inputs for this one.

Choose one or more parent nodes

Amazon S3Amazon S3

S3 - DataSourceS3 - DataSource

The parents of this node have overlapping field names. AWS Glue Studio can add an Apply Mapping node to rename them and avoid downstream issues.

Custom prefix

Add a prefix to the field names of the parent node on the right

rightResolve it

Join type

Select the type of join to perform.

Inner join

Select all rows from both datasets that meet the join condition.

Join conditions

Select a field from each parent node for the join condition.

Amazon S3Amazon S3

emp_idemp_id

Add condition

maazjavaid_employee_earnings_job

Last modified on 5/17/2023, 12:12:04 PMTry new UIActionsSaveRun

VisualScriptJob detailsRunsSchedulesVersion Control

SourceActionTargetUndoRedoRemove

Data source - S3 bucket
Amazon S3

Data source - S3 bucket
Amazon S3

Transform - Join
Join

Transform - SQL Query
SQL Query

Data target - S3 bucket
Amazon S3

Transform

Output schema

Data preview

Name

SQL Query

Node parents

Choose which nodes will provide inputs for this one.

Choose one or more parent nodes

Join

Join - Transform

Associate an alias with each input source

Info

Edit the aliases used for the inputs to this node.

Input sources

SQL aliases

JoinmyDataSource

SQL query

Enter a SQL statement to add to your job.

1 SELECT
2 location,
3 AVG(earnings) AS average_earning,
4 (AVG(earnings)-MIN(earnings))/MAX(earnings)*100 AS raise_percentage
5 FROM
6 myDataSource
7 GROUP BY
8 location;
9

maazjavaidd_employee_earnings_job

Last modified on 5/17/2023, 12:12:04 PM

Try new UI

Actions

Save

Run

VisualScriptJob detailsRunsSchedulesVersion Control

SourceActionTargetUndoRedoRemove

Data source - S3 bucket
Amazon S3

Data source - S3 bucket
Amazon S3

Transform - Join
Join

Transform - SQL Query
SQL Query

Data target - S3 bucket
Amazon S3

Data target properties - S3

Output schema

Data preview

Name
Amazon S3

Node parents
Choose which nodes will provide inputs for this one.
Choose one or more parent nodes
SQL Query
SqlCode - Transform

Format
Parquet

Compression Type
Snappy

S3 Target Location
Choose an S3 location in the format s3://bucket/prefix/object/ with a trailing slash (/).
s3://maazjavaidd-glues-data/output_data/location/ViewBrowse S3

Data Catalog update options
Choose how you want to update the Data Catalog table's schema and partitions. These options will only apply if the Data Catalog table is an S3 backed source.
☐ Do not update the Data Catalog
☒ Create a table in the Data Catalog and on subsequent runs, update the schema and add new partitions
☐ Create a table in the Data Catalog and on subsequent runs, keep existing schema and add new partitions

Database
Choose the database from the AWS Glue Data Catalog.
maazjavaidd_glue_database

Use runtime parameters

maazjavaidd_employee_earnings_job

Last modified on 5/17/2023, 12:49:36 PM

Try new UI

End session

Actions

Save

Run

VisualScriptJob detailsRunsSchedulesVersion Control

SourceActionTargetUndoRedoRemove

Data source - S3 bucket
Amazon S3

Data source - S3 bucket
Amazon S3

Transform - Join
Join

Transform - SQL Query
SQL Query

Data target - S3 bucket
Amazon S3

Transform

Output schema

Data preview

Schema

Info

Use datapreview schema

Edit

Key	Data type
emp_id	long
earnings	long
date	string
emp_id	long
location	string

maazjavaidd_employee_earnings_job

Last modified on 5/17/2023, 12:49:36 PM

Try new UI

End session

Actions

Save

Run

VisualScriptJob detailsRunsSchedulesVersion Control

SourceActionTargetUndoRedoRemove

Data source - S3 bucket
Amazon S3

Data source - S3 bucket
Amazon S3

Transform - Join
Join

Transform - SQL Query
SQL Query

Data target - S3 bucket
Amazon S3

Transform

Output schema

Data preview

Data preview (5)

Info

Previewing 3 of 3 fields

Filter sample dataset

location	average_earning	raise_percentage
B	6086.875	184.30056048575452
C	5695.3	158.9949977262392
A	6217.975	205.85218888342354
D	5635.075	180.91101694915253
E	5503.4	154.31608133086874

