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 :

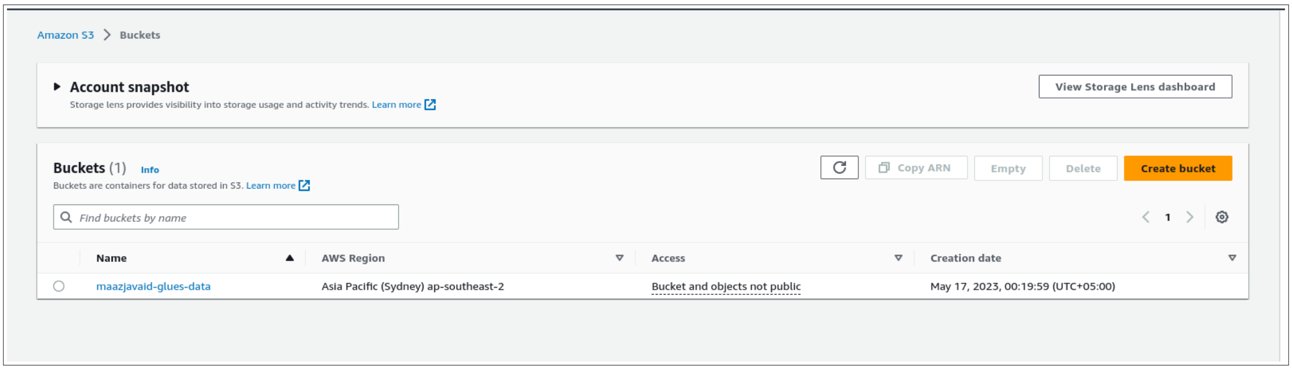
Using the salary CSV as a base, prepare a new data file with employees' office locations. Make sure there are 5-6 distinct locations that are shared between employees.

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:



Amazon S3 > Buckets

▸ Account snapshot

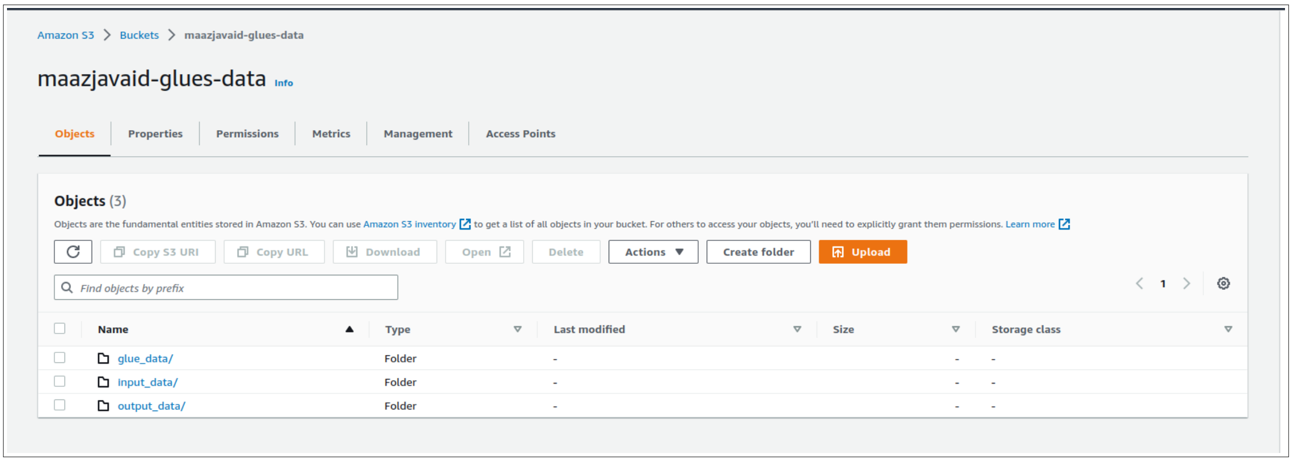
Storage lens provides visibility into storage usage and activity trends. Learn more

Buckets (1) Info

Buckets are containers for data stored in S3. Learn more

	Name	AWS Region	Access	Creation date
○	maazjavaid-glues-data	Asia Pacific (Sydney) ap-southeast-2	Bucket and objects not public	May 17, 2023, 00:19:59 (UTC+05:00)

Then we create the directories on S3 bucket:



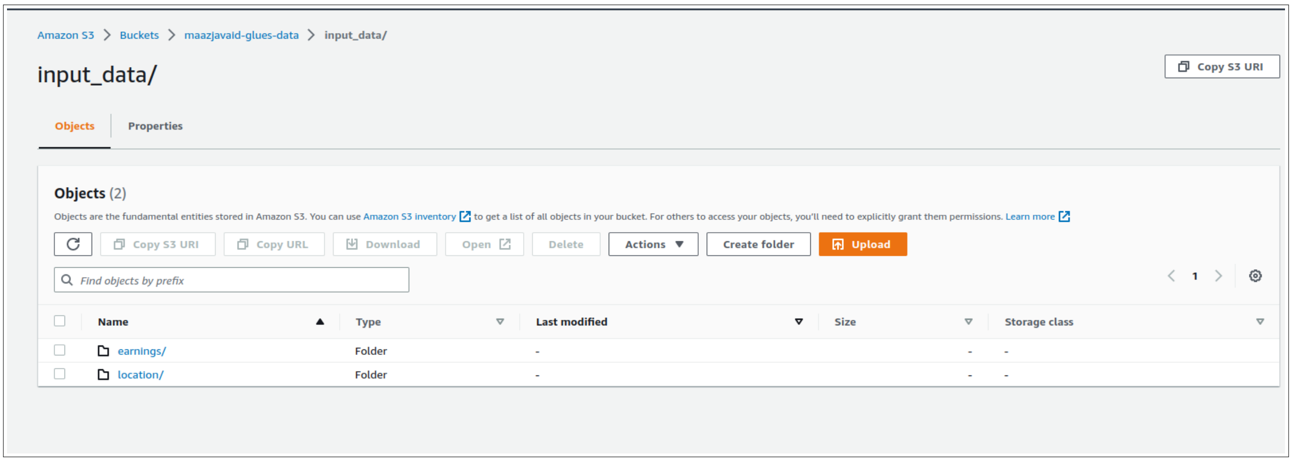
Amazon S3 > Buckets > maazjavaid-glues-data

maazjavaid-glues-data Info

Objects (3)

Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more

□	Name	Type	Last modified	Size	Storage class
□	glue_data/	Folder	-	-	-
□	input_data/	Folder	-	-	-
□	output_data/	Folder	-	-	-

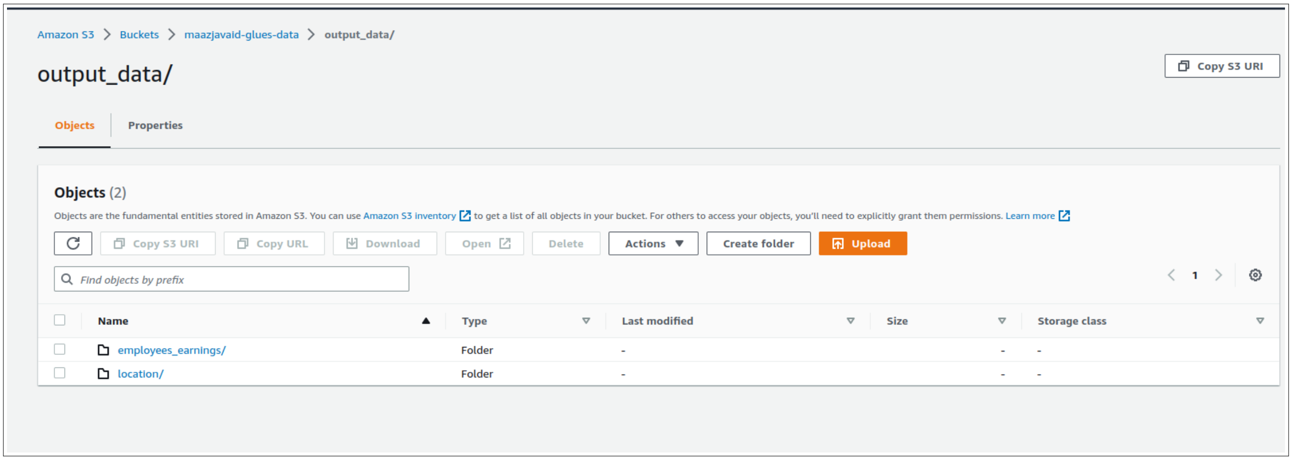


input_data/

Objects (2)

Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more

□	Name	Type	Last modified	Size	Storage class
□	earnings/	Folder	-	-	-
□	location/	Folder	-	-	-



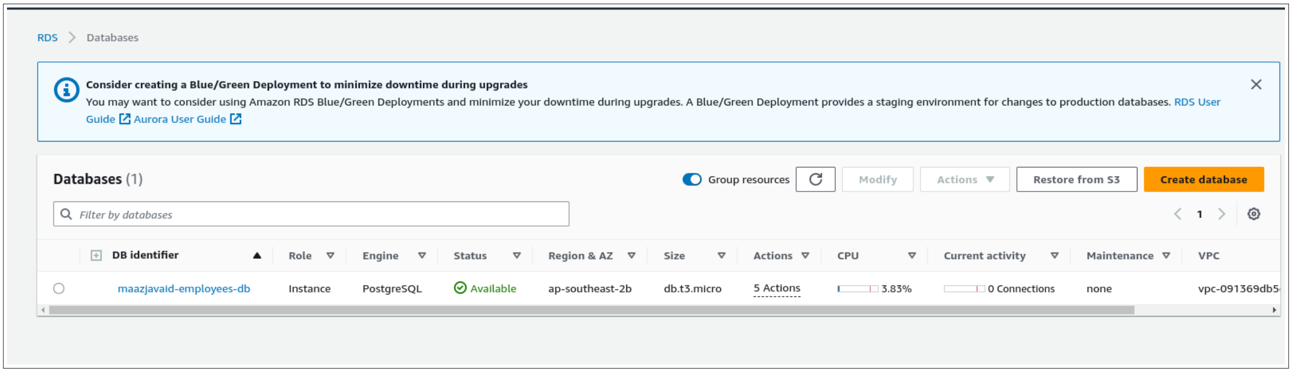
output_data/

Objects (2)

Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more

□	Name	Type	Last modified	Size	Storage class
□	employees_earnings/	Folder	-	-	-
□	location/	Folder	-	-	-

Create the RDS database:

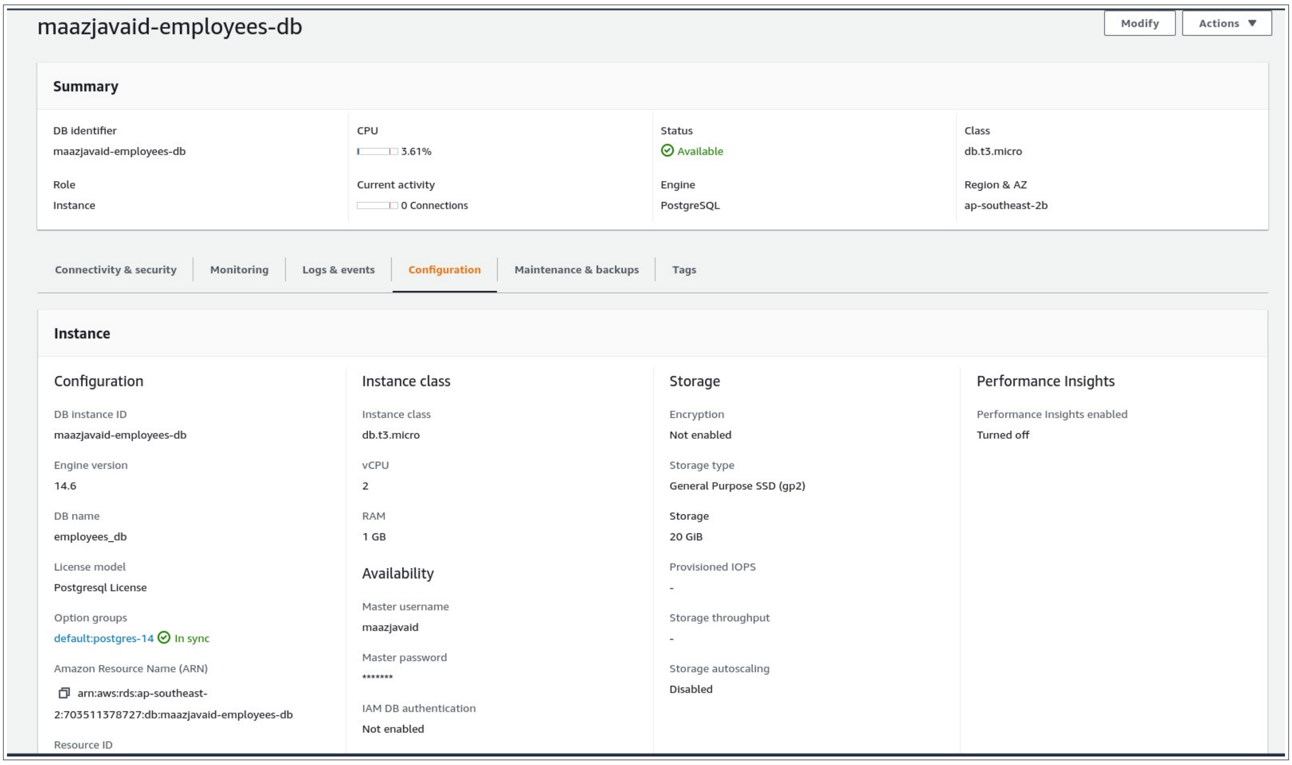


Consider creating a Blue/Green Deployment to minimize downtime during upgrades

You may want to consider using Amazon RDS Blue/Green Deployments and minimize your downtime during upgrades. A Blue/Green Deployment provides a staging environment for changes to production databases. RDS User Guide Aurora User Guide

Databases (1)

	DB identifier	Role	Engine	Status	Region & AZ	Size	Actions	CPU	Current activity	Maintenance	VPC
○	maazjavaid-employees-db	Instance	PostgreSQL	Available	ap-southeast-2b	db.t3.micro	5 Actions	3.83%	0 Connections	none	vpc-091369db5



maazjavaid-employees-db

Summary

DB identifier	CPU	Status	Class
maazjavaid-employees-db	3.61%	✓ Available	db.t3.micro
Role	Current activity	Engine	Region & AZ
Instance	0 Connections	PostgreSQL	ap-southeast-2b

Instance

Configuration

DB instance ID
maazjavaid-employees-db

Engine version
14.6

DB name
employees_db

License model
Postgresql License

Option groups
default:postgres-14 ✓ In sync

Amazon Resource Name (ARN)
arn:aws:rds:ap-southeast-2:703511378727:db:maazjavaid-employees-db

Resource ID

Instance class

Instance class
db.t3.micro

vCPU
2

RAM
1 GB

Availability

Master username
maazjavaid

Master password
****

IAM DB authentication
Not enabled

Storage

Encryption
Not enabled

Storage type
General Purpose SSD (gp2)

Storage
20 GiB

Provisioned IOPS
-

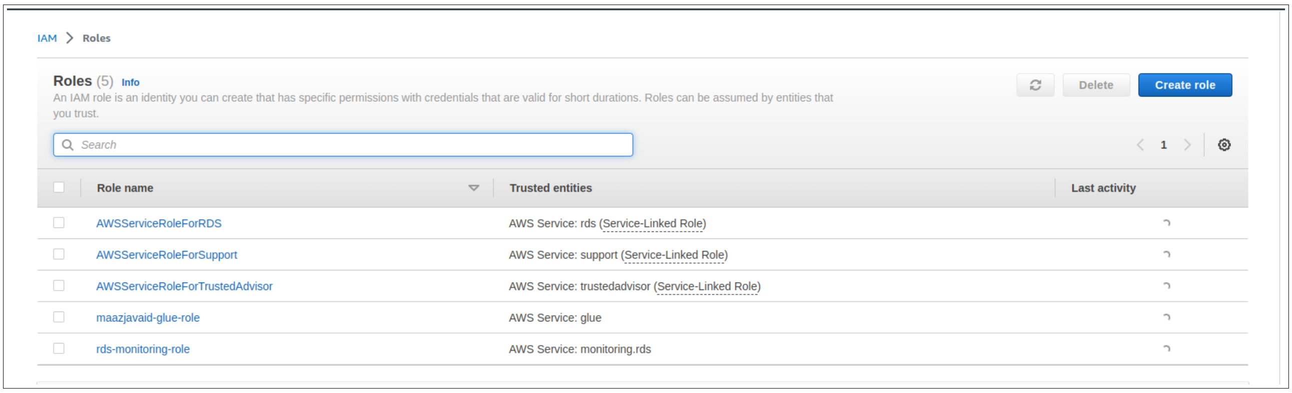
Storage throughput
-

Storage autoscaling
Disabled

Performance Insights

Performance Insights enabled
Turned off

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

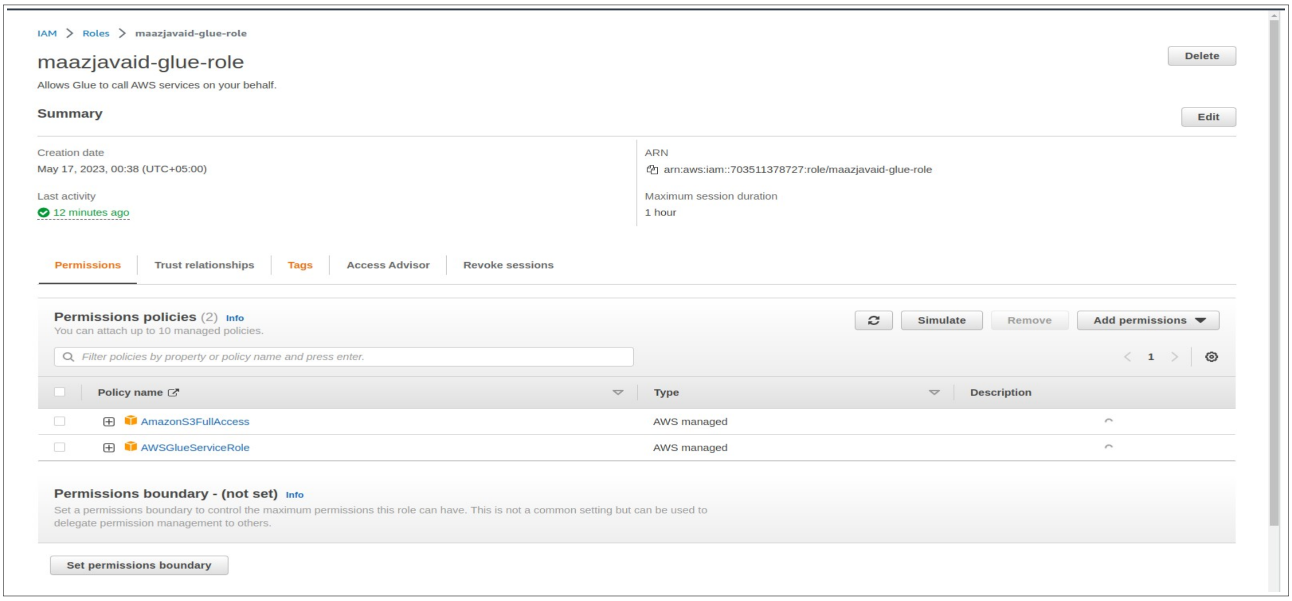


Roles (5) Info

An IAM role is an identity you can create that has specific permissions with credentials that are valid for short durations. Roles can be assumed by entities that you trust.

Search

□	Role name	Trusted entities	Last activity
□	AWSServiceRoleForRDS	AWS Service: rds (Service-Linked Role)	-
□	AWSServiceRoleForSupport	AWS Service: support (Service-Linked Role)	-
□	AWSServiceRoleForTrustedAdvisor	AWS Service: trustedadvisor (Service-Linked Role)	-
□	maazjavaid-glue-role	AWS Service: glue	-
□	rds-monitoring-role	AWS Service: monitoring.rds	-



maazjavaid-glue-role

Allows Glue to call AWS services on your behalf.

Summary

Creation date
May 17, 2023, 00:38 (UTC+05:00)

Last activity
✓ 12 minutes ago

ARN
arn:aws:iam::703511378727:role/maazjavaid-glue-role

Maximum session duration
1 hour

Permissions policies (2) Info
You can attach up to 10 managed policies.

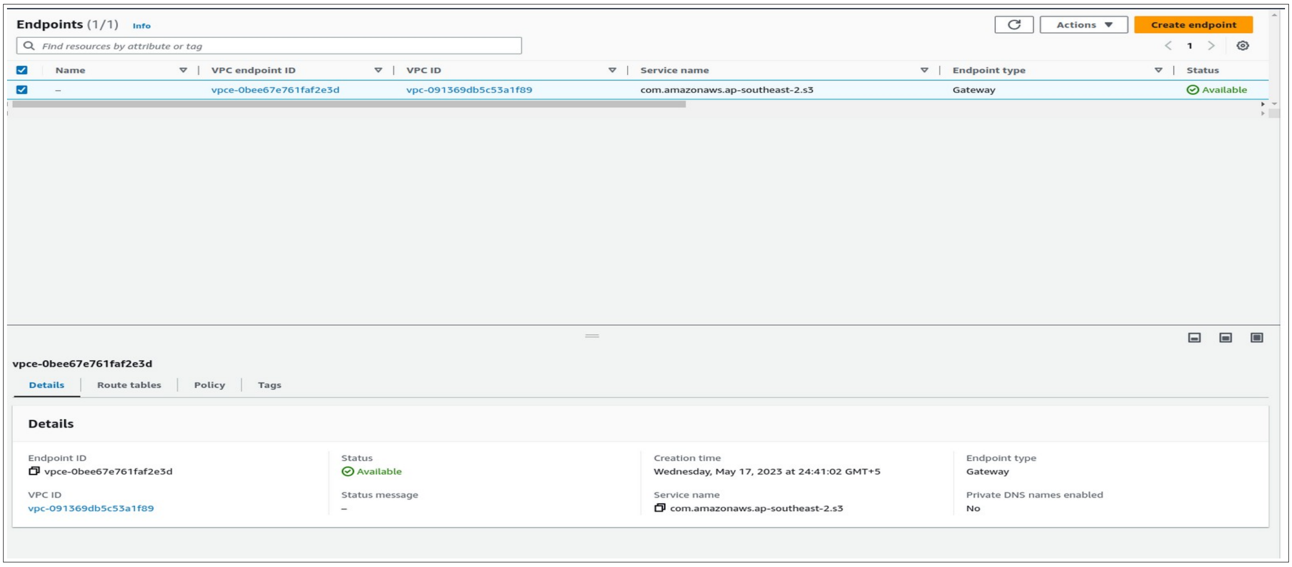
Filter policies by property or policy name and press enter.

□	Policy name	Type	Description
□	⊞ AmazonS3FullAccess	AWS managed	-
□	⊞ AWSGlueServiceRole	AWS managed	-

Permissions boundary - (not set) Info

Set a permissions boundary to control the maximum permissions this role can have. This is not a common setting but can be used to delegate permission management to others.

Create the VPC Endpoint:



Endpoints (1/1)

☑	Name	VPC endpoint ID	VPC ID	Service name	Endpoint type	Status
☑	–	vpce-0bee67e761faf2e3d	vpc-091369db5c53a1f89	com.amazonaws.ap-southeast-2.s3	Gateway	Available

vpce-0bee67e761faf2e3d

Details

Details

Endpoint ID
vpce-0bee67e761faf2e3d

VPC ID
vpc-091369db5c53a1f89

Status
Available

Status message
–

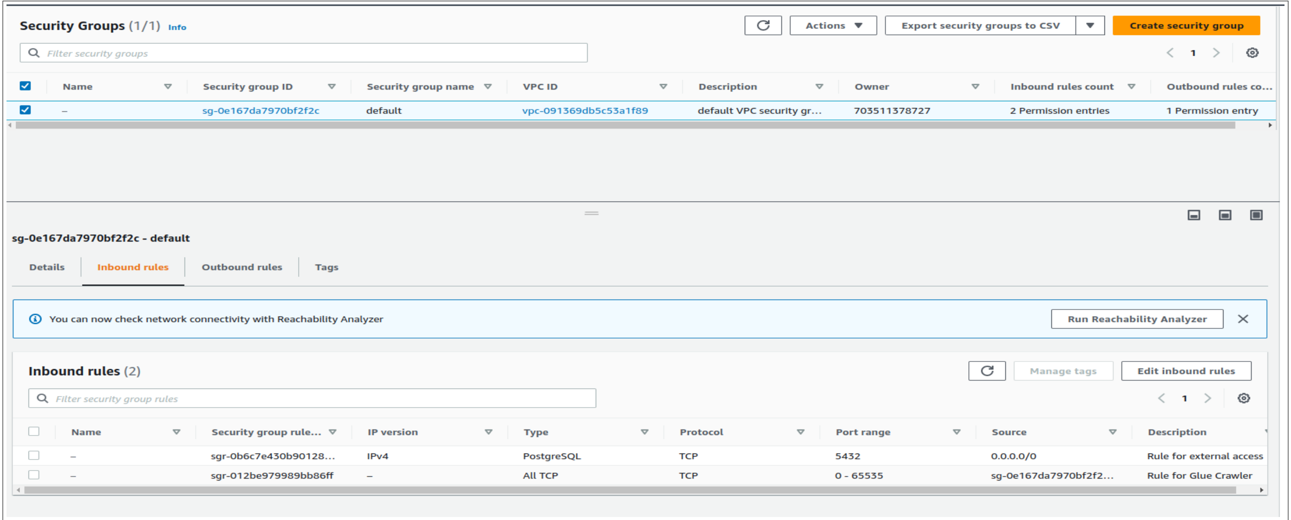
Creation time
Wednesday, May 17, 2023 at 24:41:02 GMT+5

Service name
com.amazonaws.ap-southeast-2.s3

Endpoint type
Gateway

Private DNS names enabled
No

Security group rules:



Security Groups (1/1)

☑	Name	Security group ID	Security group name	VPC ID	Description	Owner	Inbound rules count	Outbound rules co...
☑	–	sg-0e167da7970bf2f2c	default	vpc-091369db5c53a1f89	default VPC security gr...	703511378727	2 Permission entries	1 Permission entry

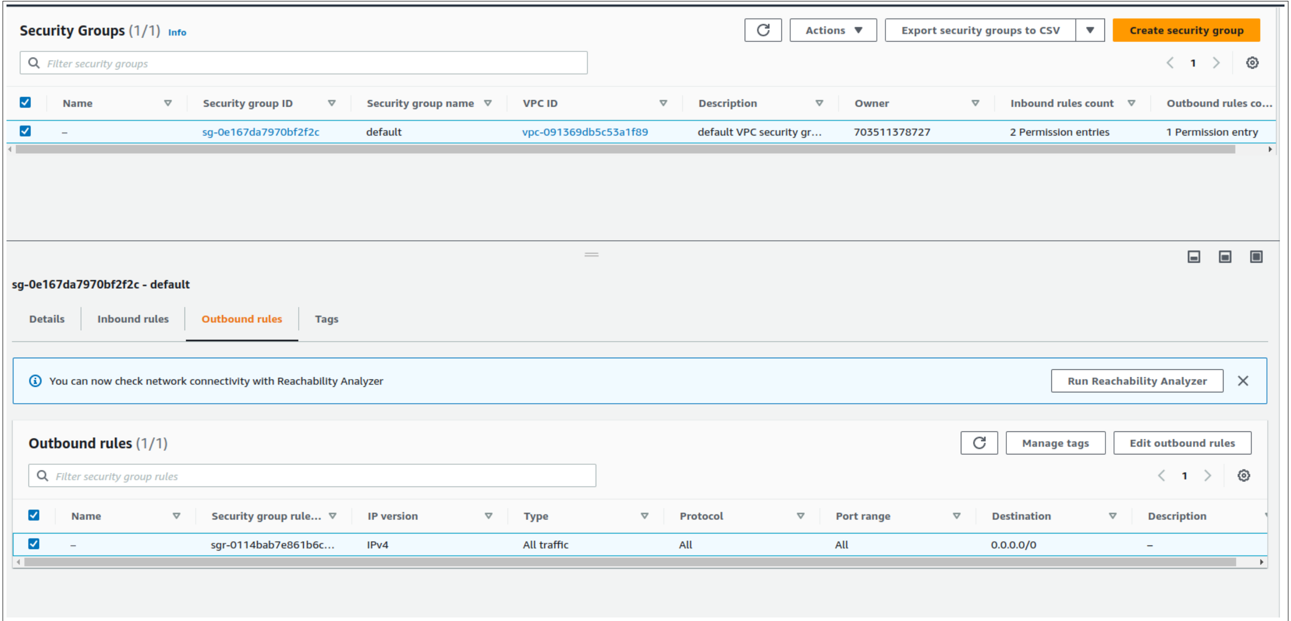
sg-0e167da7970bf2f2c - default

Inbound rules

You can now check network connectivity with Reachability Analyzer

Inbound rules (2)

□	Name	Security group rule...	IP version	Type	Protocol	Port range	Source	Description
□	–	sgr-0b6c7e430b90128...	IPv4	PostgreSQL	TCP	5432	0.0.0.0/0	Rule for external access
□	–	sgr-012be979989bb86ff	–	All TCP	TCP	0 - 65535	sg-0e167da7970bf2f2...	Rule for Glue Crawler



Security Groups (1/1)

☑	Name	Security group ID	Security group name	VPC ID	Description	Owner	Inbound rules count	Outbound rules co...
☑	–	sg-0e167da7970bf2f2c	default	vpc-091369db5c53a1f89	default VPC security gr...	703511378727	2 Permission entries	1 Permission entry

sg-0e167da7970bf2f2c - default

Outbound rules

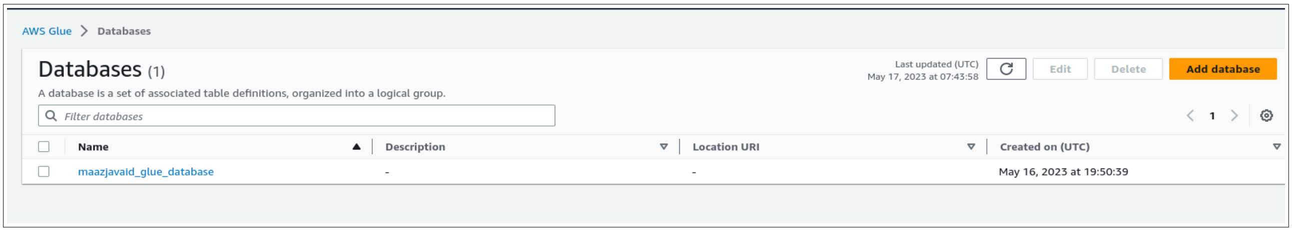
You can now check network connectivity with Reachability Analyzer

Outbound rules (1/1)

☑	Name	Security group rule...	IP version	Type	Protocol	Port range	Destination	Description
☑	–	sgr-0114bab7e861b6c...	IPv4	All traffic	All	All	0.0.0.0/0	–

2.ETL in Glue

Create database:

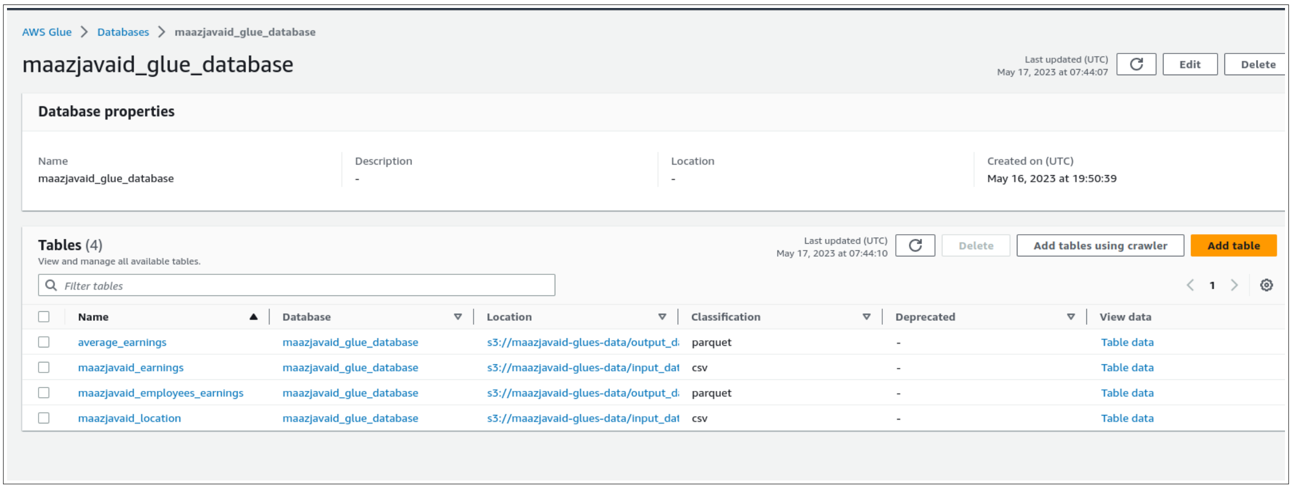


Databases (1)

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

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

□	Name	Description	Location URI	Created on (UTC)
□	maazjavaid_glue_database	-	-	May 16, 2023 at 19:50:39



maazjavaid_glue_database

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

Database properties

Name	Description	Location	Created on (UTC)
maazjavaid_glue_database	-	-	May 16, 2023 at 19:50:39

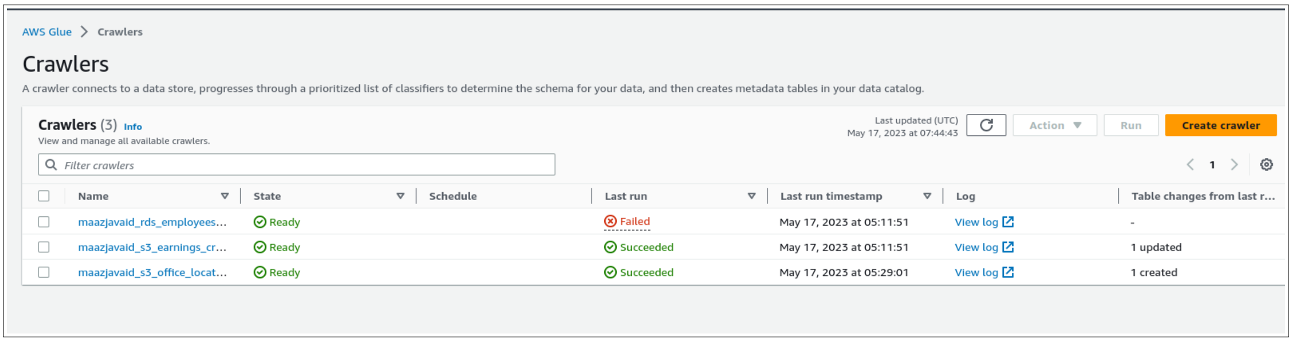
Tables (4)

View and manage all available tables.

Last updated (UTC) May 17, 2023 at 07:44:10

□	Name	Database	Location	Classification	Deprecated	View data
□	average_earnings	maazjavaid_glue_database	s3://maazjavaid-glues-data/output_d	parquet	-	Table data
□	maazjavaid_earnings	maazjavaid_glue_database	s3://maazjavaid-glues-data/input_dat	csv	-	Table data
□	maazjavaid_employees_earnings	maazjavaid_glue_database	s3://maazjavaid-glues-data/output_d	parquet	-	Table data
□	maazjavaid_location	maazjavaid_glue_database	s3://maazjavaid-glues-data/input_dat	csv	-	Table data

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.

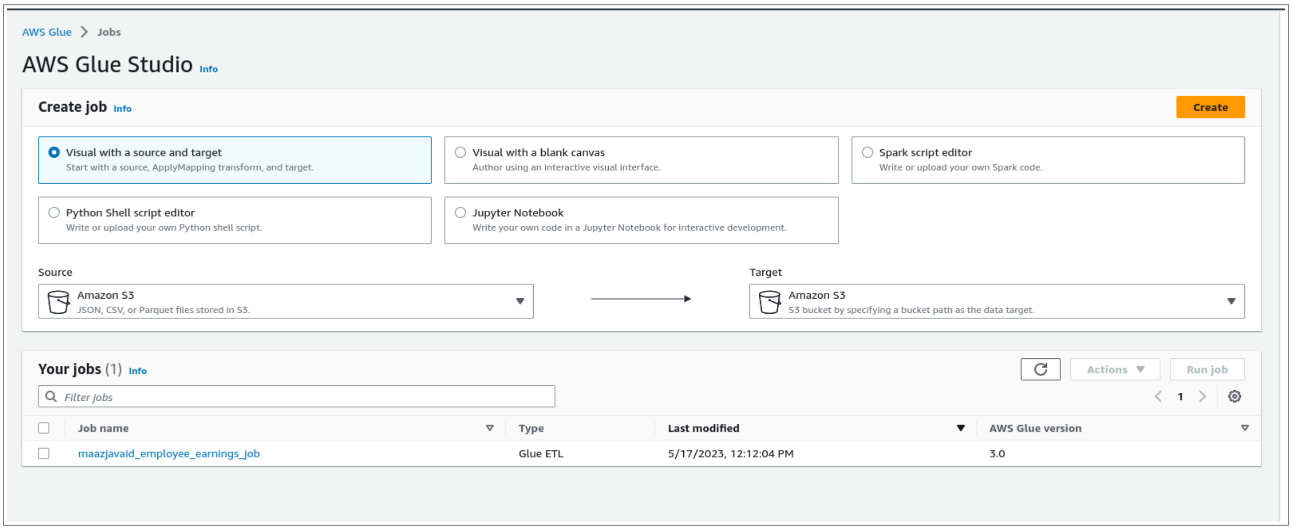
Crawlers (3)

View and manage all available crawlers.

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

□	Name	State	Schedule	Last run	Last run timestamp	Log	Table changes from last r...
□	maazjavaid_rds_employees...	✓ Ready		⊗ Failed	May 17, 2023 at 05:11:51	View log	-
□	maazjavaid_s3_earnings_cr...	✓ Ready		✓ Succeeded	May 17, 2023 at 05:11:51	View log	1 updated
□	maazjavaid_s3_office_locat...	✓ Ready		✓ Succeeded	May 17, 2023 at 05:29:01	View log	1 created

Create job:



AWS Glue Studio

Create job

● Visual with a source and target
Start with a source, ApplyMapping transform, and target.

○ Visual with a blank canvas
Author using an interactive visual interface.

○ Spark script editor
Write or upload your own Spark code.

○ Python Shell script editor
Write or upload your own Python shell script.

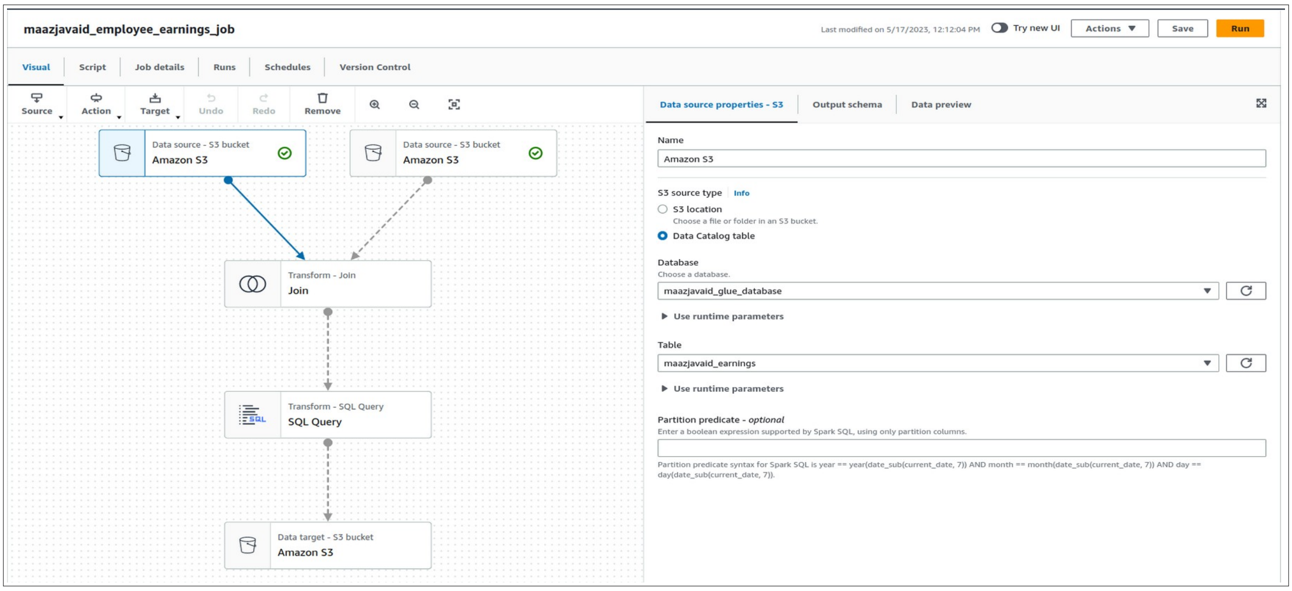
○ Jupyter Notebook
Write your own code in a Jupyter Notebook for interactive development.

Source: Amazon S3 — JSON, CSV, or Parquet files stored in S3.

Target: Amazon S3 — S3 bucket by specifying a bucket path as the data target.

Your jobs (1)

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



maazjavaid_employee_earnings_job

Name
Amazon S3

S3 source type Info
○ S3 location
Choose a file or folder in an S3 bucket.
● Data Catalog table

Database
Choose a database.
maazjavaid_glue_database

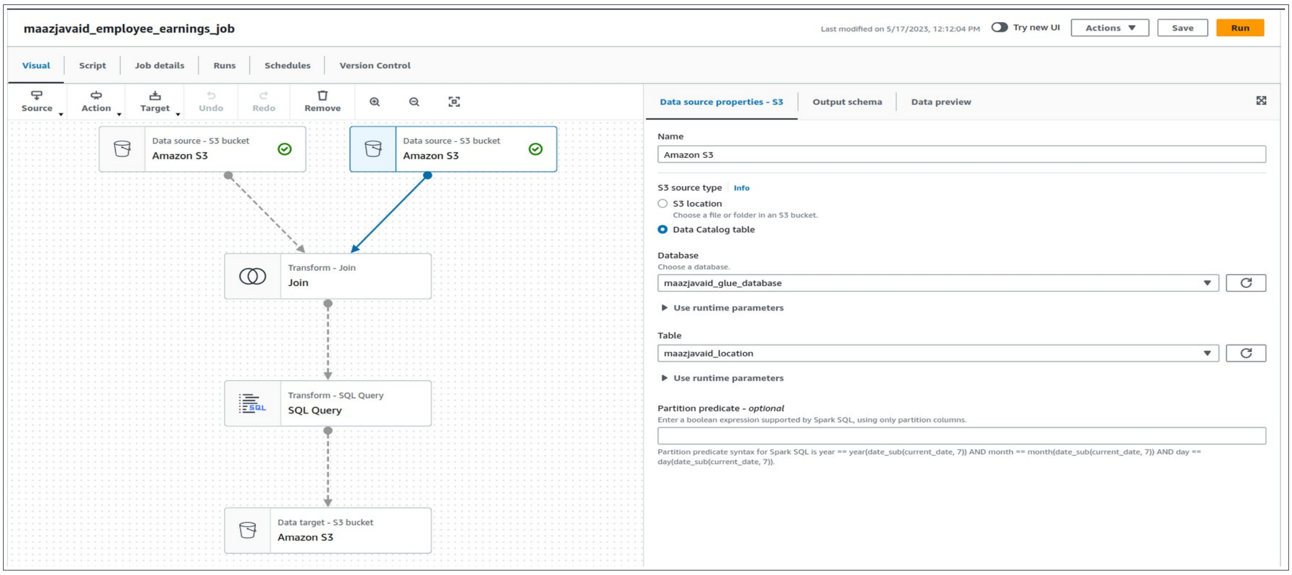
▶ Use runtime parameters

Table
maazjavaid_earnings

▶ Use runtime parameters

Partition predicate - *optional*
Enter a boolean expression supported by Spark SQL, using only partition columns.

Partition predicate syntax for Spark SQL is year == year(date_sub(current_date, 7)) AND month == month(date_sub(current_date, 7)) AND day == day(date_sub(current_date, 7)).



maazjavaid_employee_earnings_job

Name
Amazon S3

S3 source type Info
○ S3 location
Choose a file or folder in an S3 bucket.
● Data Catalog table

Database
Choose a database.
maazjavaid_glue_database

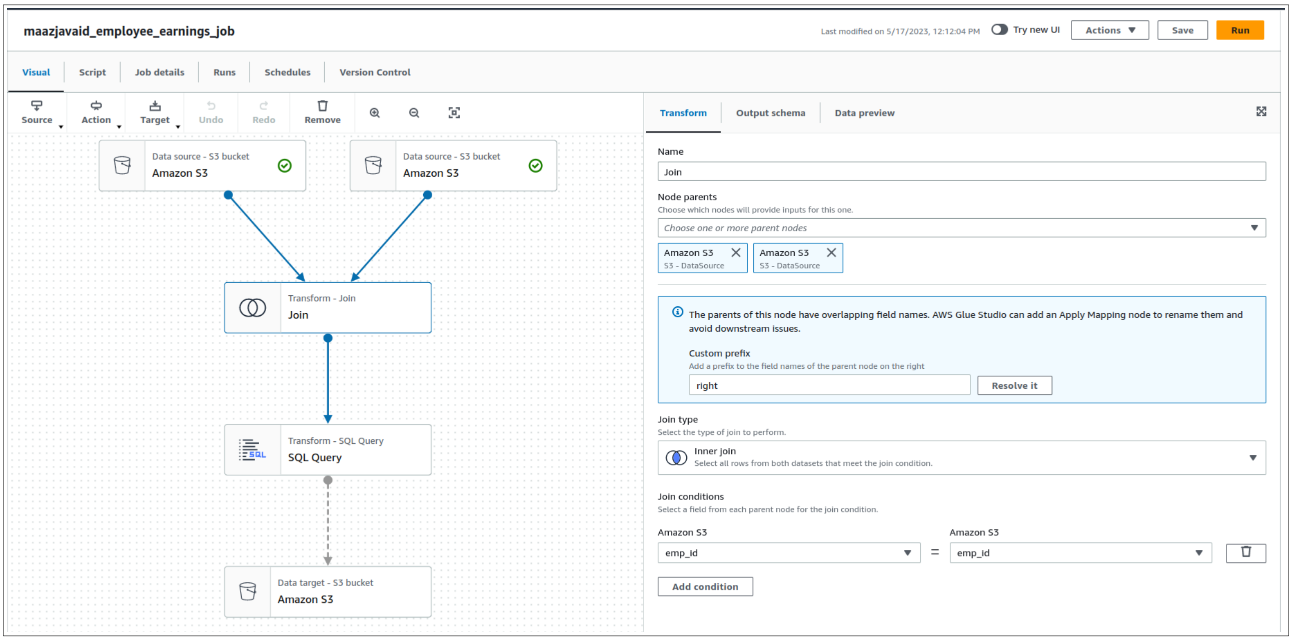
▶ Use runtime parameters

Table
maazjavaid_location

▶ Use runtime parameters

Partition predicate - *optional*
Enter a boolean expression supported by Spark SQL, using only partition columns.

Partition predicate syntax for Spark SQL is year == year(date_sub(current_date, 7)) AND month == month(date_sub(current_date, 7)) AND day == day(date_sub(current_date, 7)).



maazjavaid_employee_earnings_job

Name
Join

Node parents
Choose which nodes will provide inputs for this one.
Choose one or more parent nodes
Amazon S3 ✕ S3 - DataSource
Amazon S3 ✕ S3 - 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.

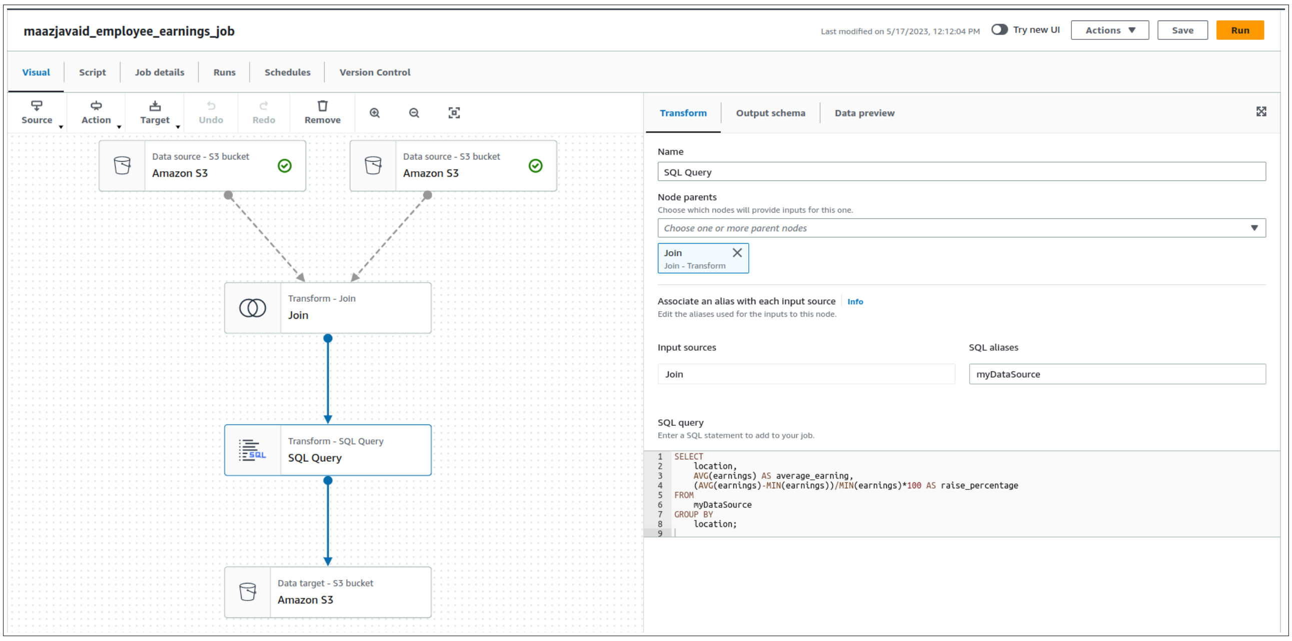
Custom prefix
Add a prefix to the field names of the parent node on the right
right
Resolve it

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 S3		Amazon S3
emp_id	=	emp_id

Add condition



Name: SQL Query

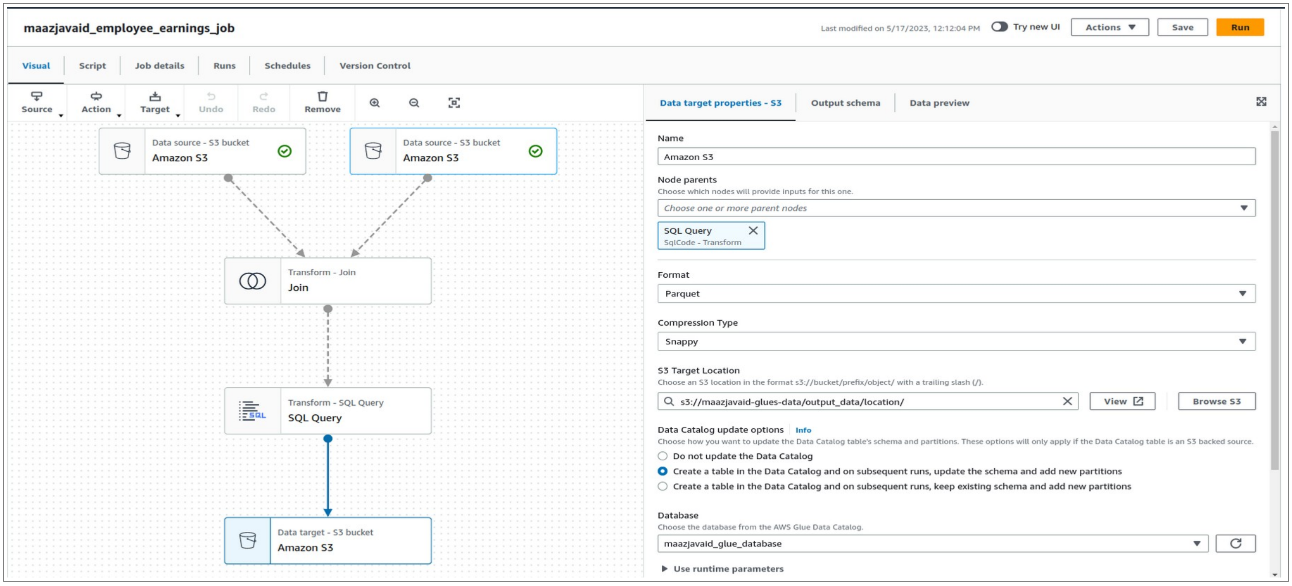
Node parents
Choose which nodes will provide inputs for this one.
Choose one or more parent nodes
Join
Join - Transform

Associate an alias with each input source Info
Edit the aliases used for the inputs to this node.

Input sources	SQL aliases
Join	myDataSource

SQL query
Enter a SQL statement to add to your job.

```sql
SELECT
    location,
    AVG(earnings) AS average_earning,
    (AVG(earnings)-MIN(earnings))/MIN(earnings)*100 AS raise_percentage
FROM
    myDataSource
GROUP BY
    location;
```



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

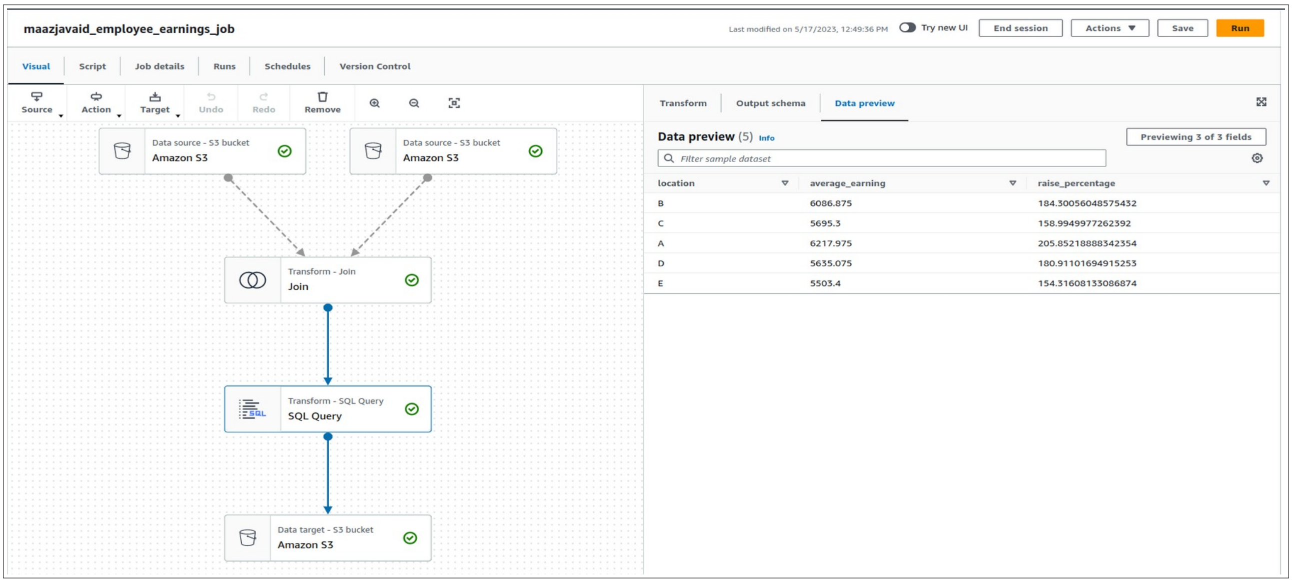
Compression Type: Snappy

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

Data Catalog update options Info
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

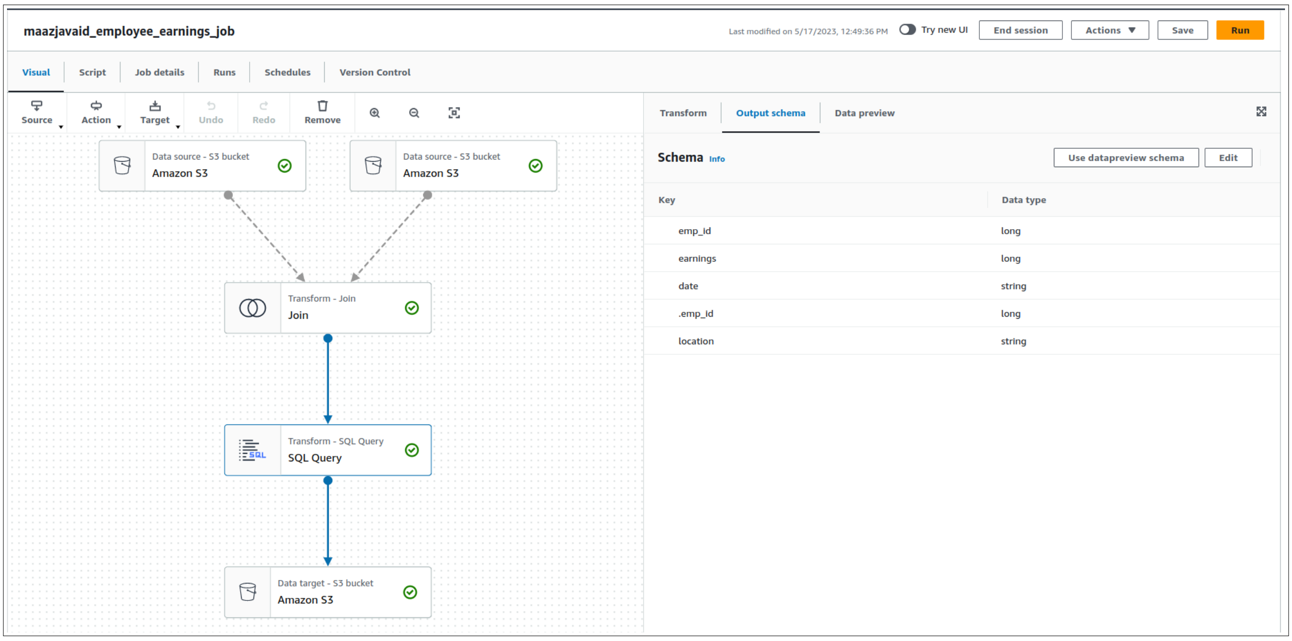
Database
Choose the database from the AWS Glue Data Catalog.
maazjavaid_glue_database

▶ Use runtime parameters



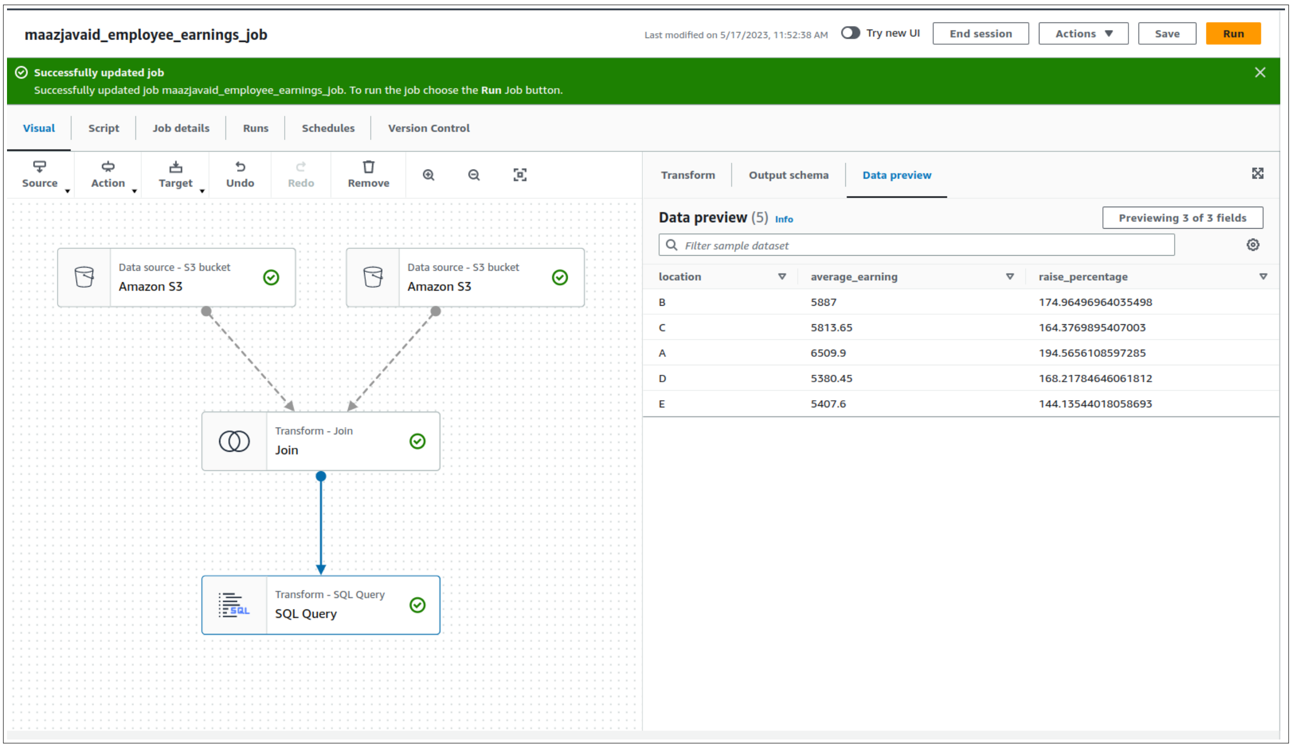
Data preview (5) Info
Previewing 3 of 3 fields
Filter sample dataset

location	average_earning	raise_percentage
B	6086.875	184.30056048575432
C	5695.3	158.9949977262392
A	6217.975	205.85218888342354
D	5635.075	180.91101694915253
E	5503.4	154.31608133086874



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

Successfully updated job:



Successfully updated job

Successfully updated job maazjavaid_employee_earnings_job. To run the job choose the **Run** Job button.

location	average_earning	raise_percentage
B	5887	174.96496964035498
C	5813.65	164.3769895407003
A	6509.9	194.5656108597285
D	5380.45	168.21784646061812
E	5407.6	144.13544018058693