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.

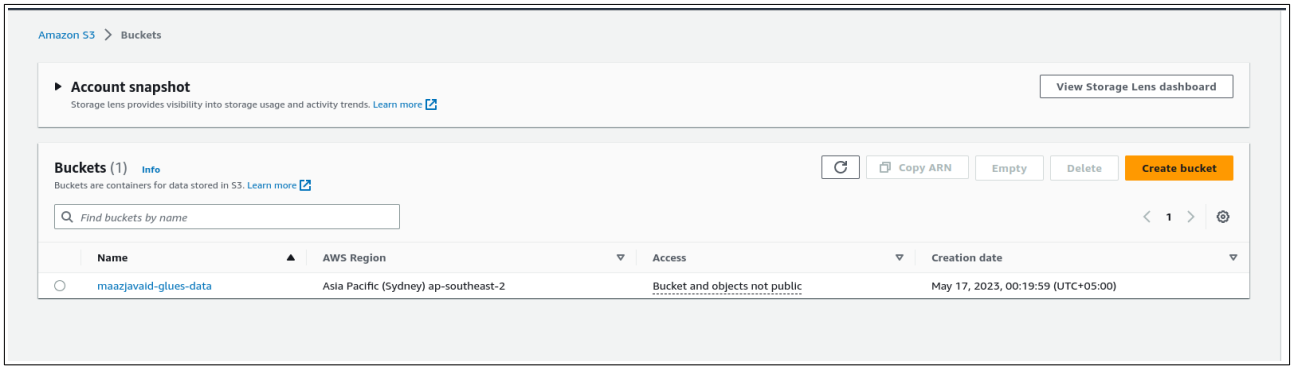
Answer:

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

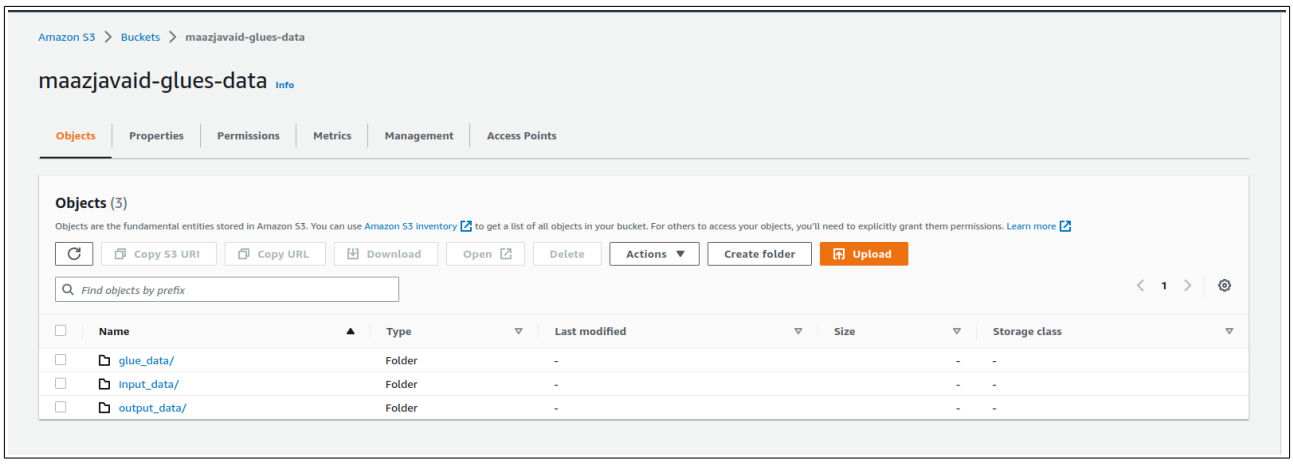
First of all we create the S3 bucket:

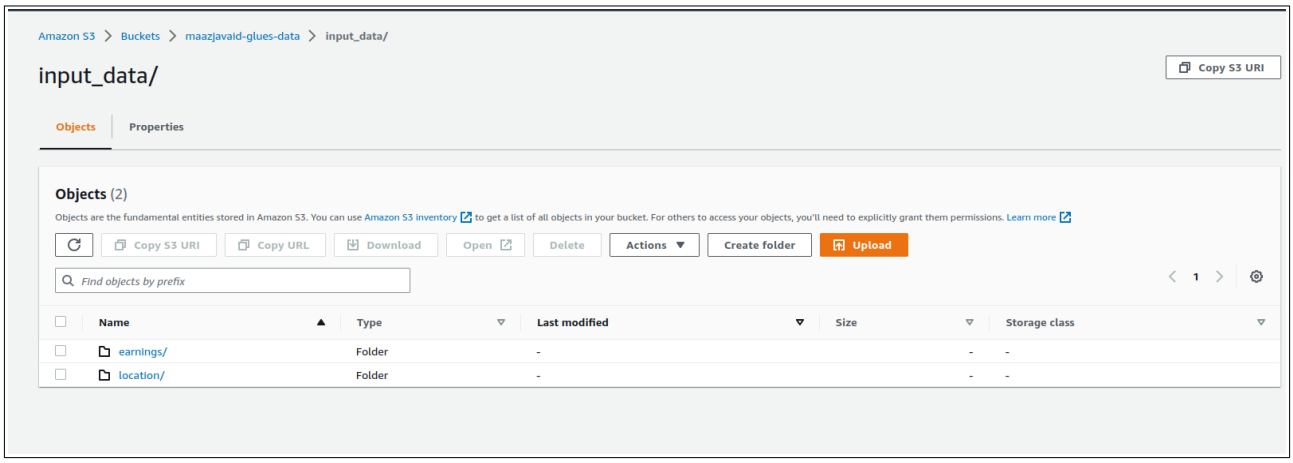
Q1. Why do we need to take this step?
A. We need to specify the input and output data locations for the job. Typically, the data source (e.g. CSV file) and output data generated by the Glue job are stored in an S3 bucket.

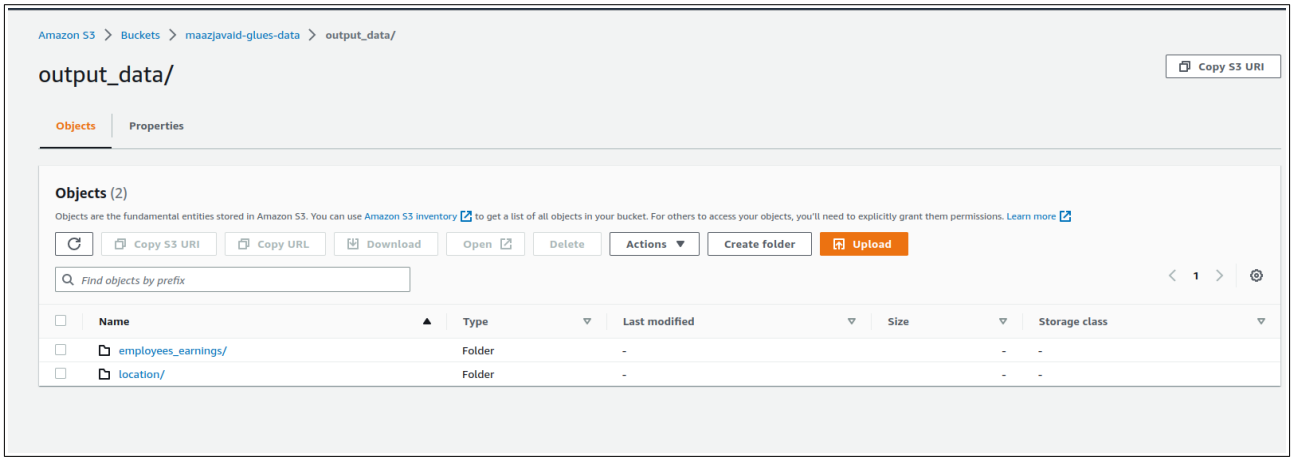
Q2. What is this service's purpose?
A. S3 provides features such as durability, availability, security, and scalability. Durability ensures that data is stored redundantly across multiple devices in multiple facilities, making it highly resilient to data loss.



Then we create the directories on S3 bucket:







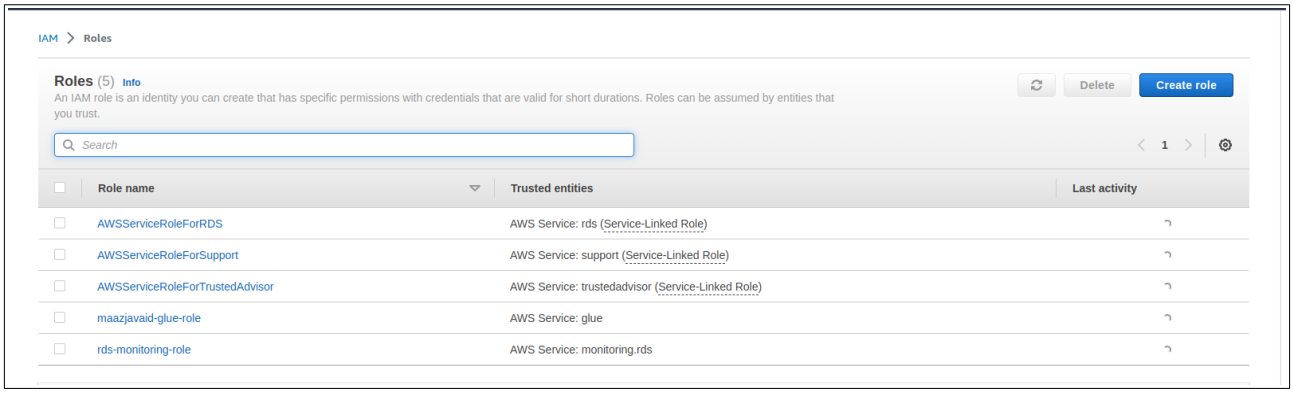
Then we go to IAM service and create the IAM Role:

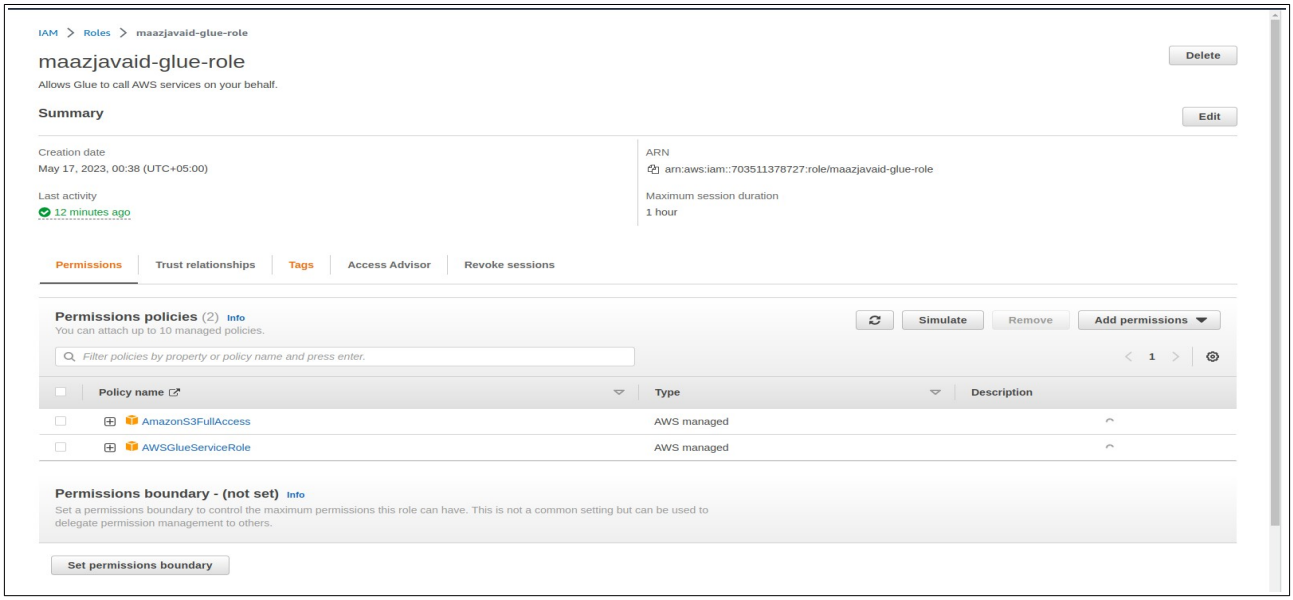
Q1. Why do we need to take this step?

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.





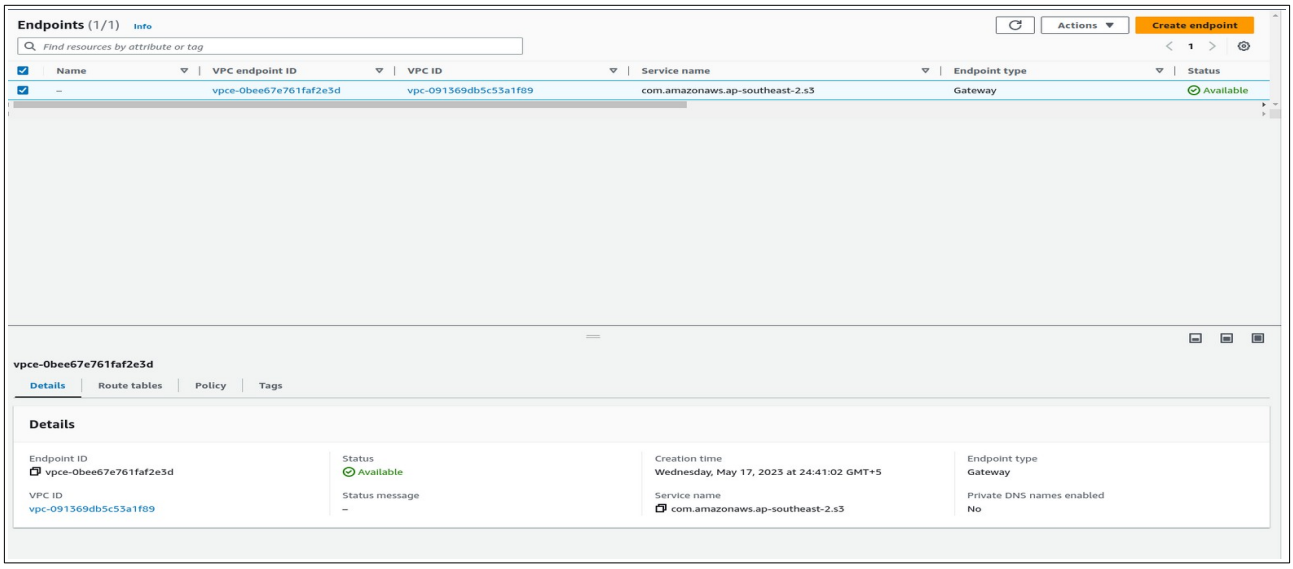
Create the VPC Endpoint:

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.



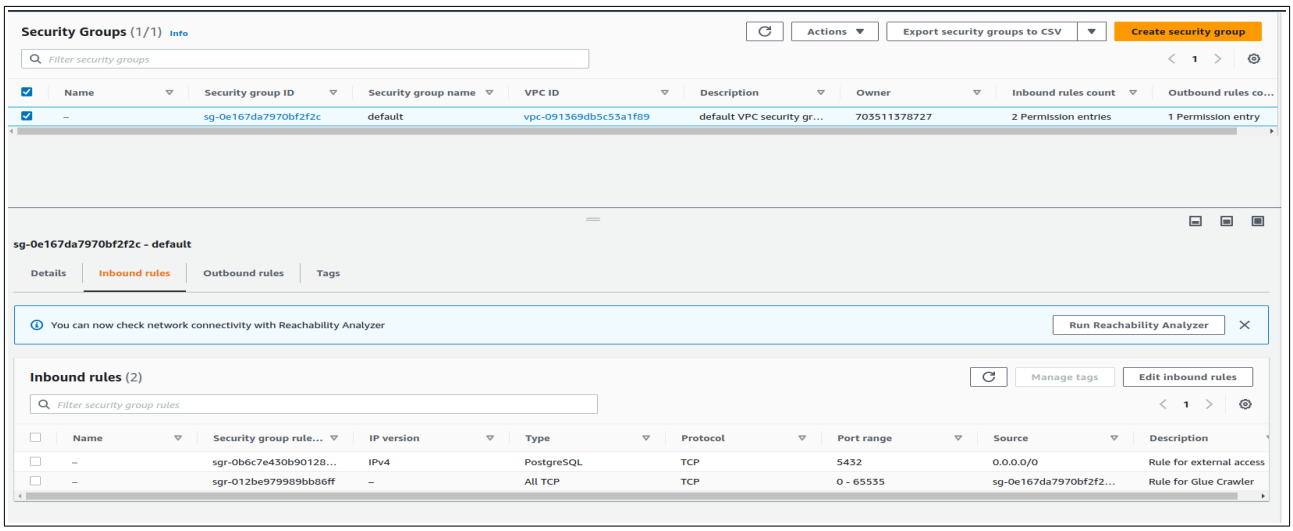
Security group rules:

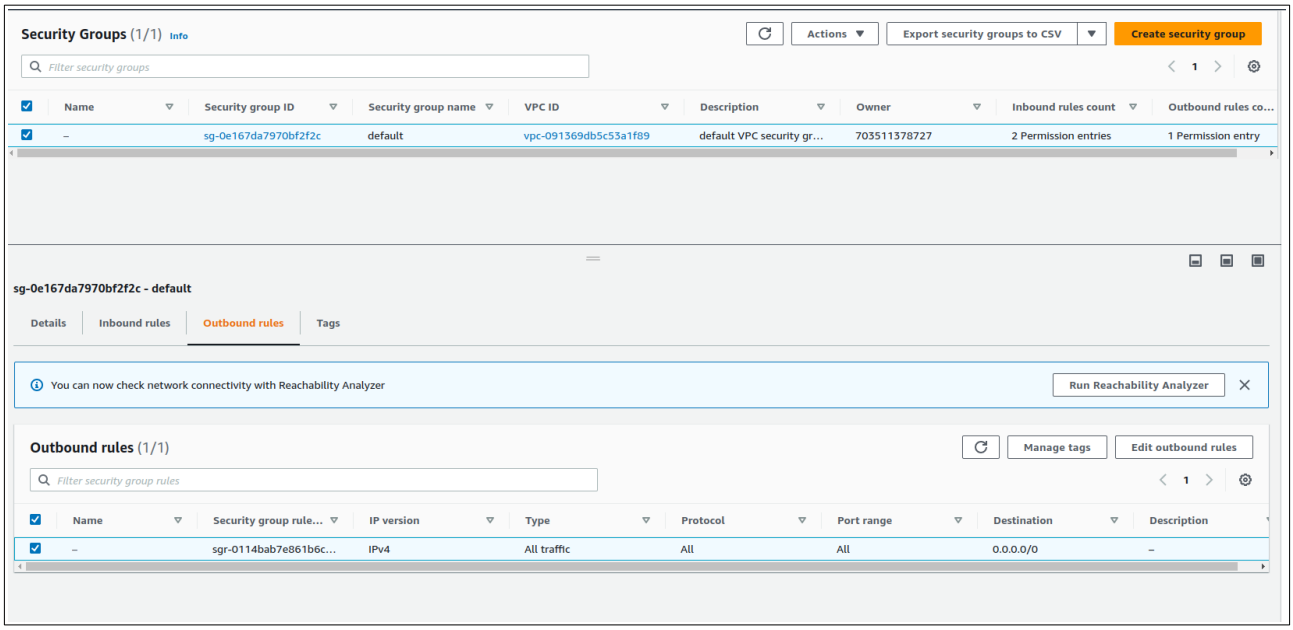
Q1. Why do we need to take this step?

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.





2.ETL in Glue

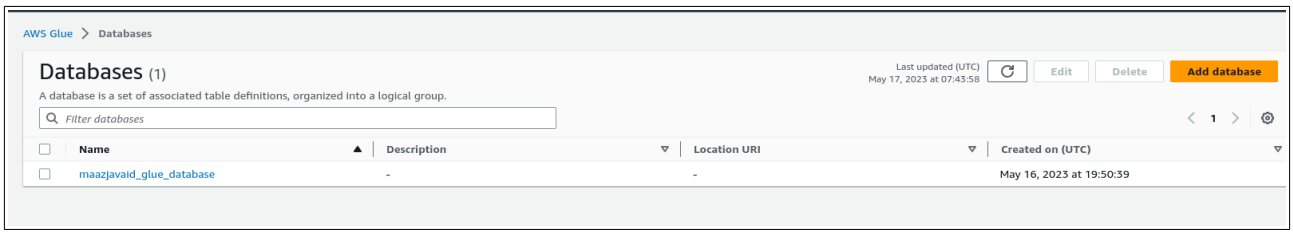
Create database:

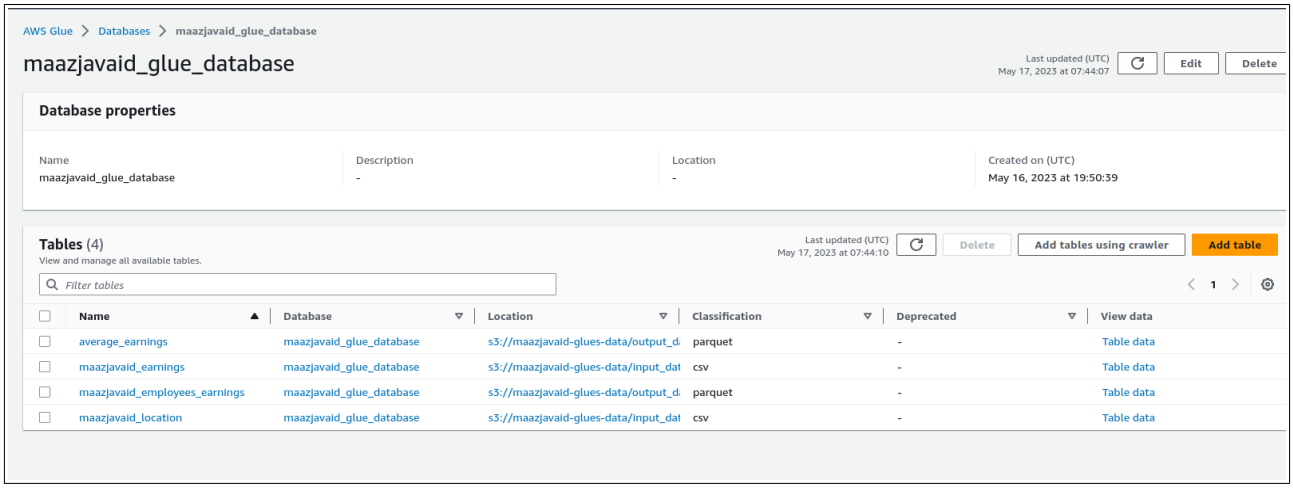
Q1. Why do we need to take this step?

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.





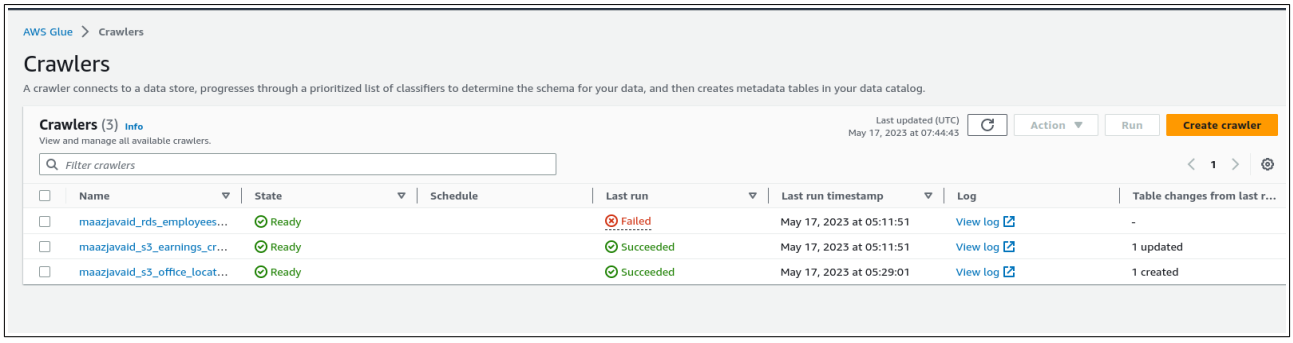
Creating Glue Crawlers:

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.



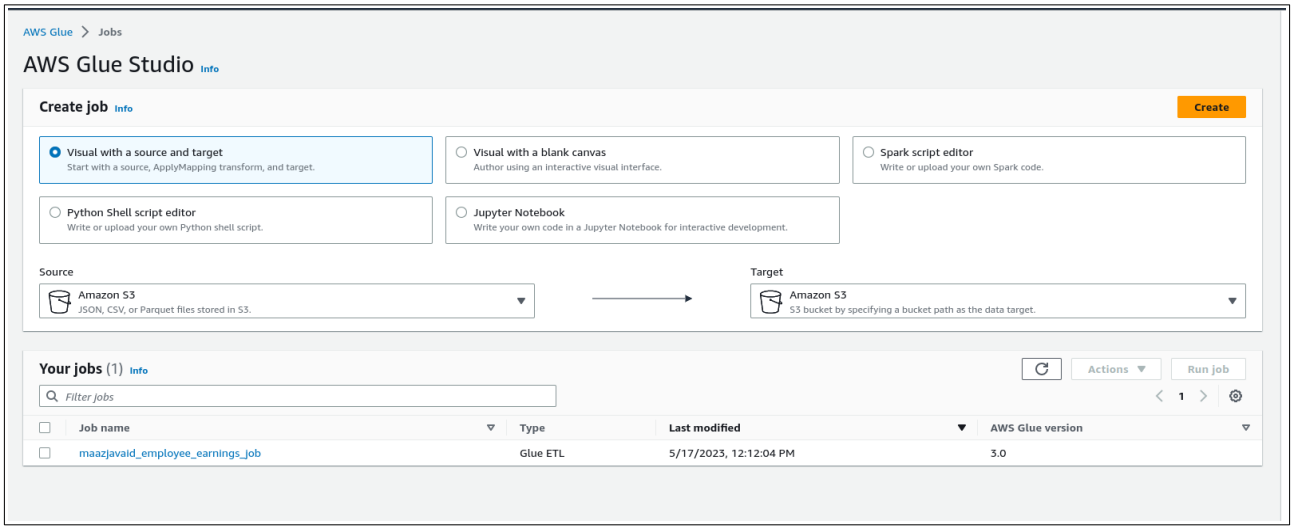
Creating job:

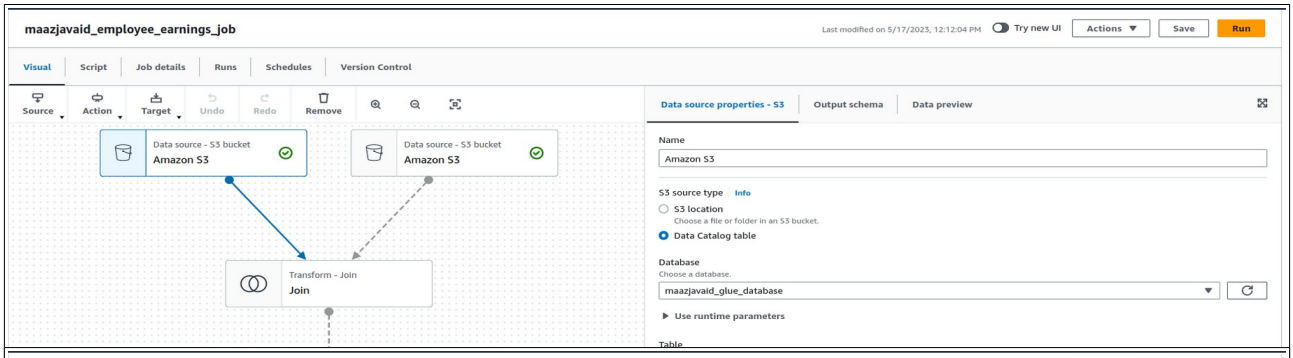
Q1. Why do we need to take this step?

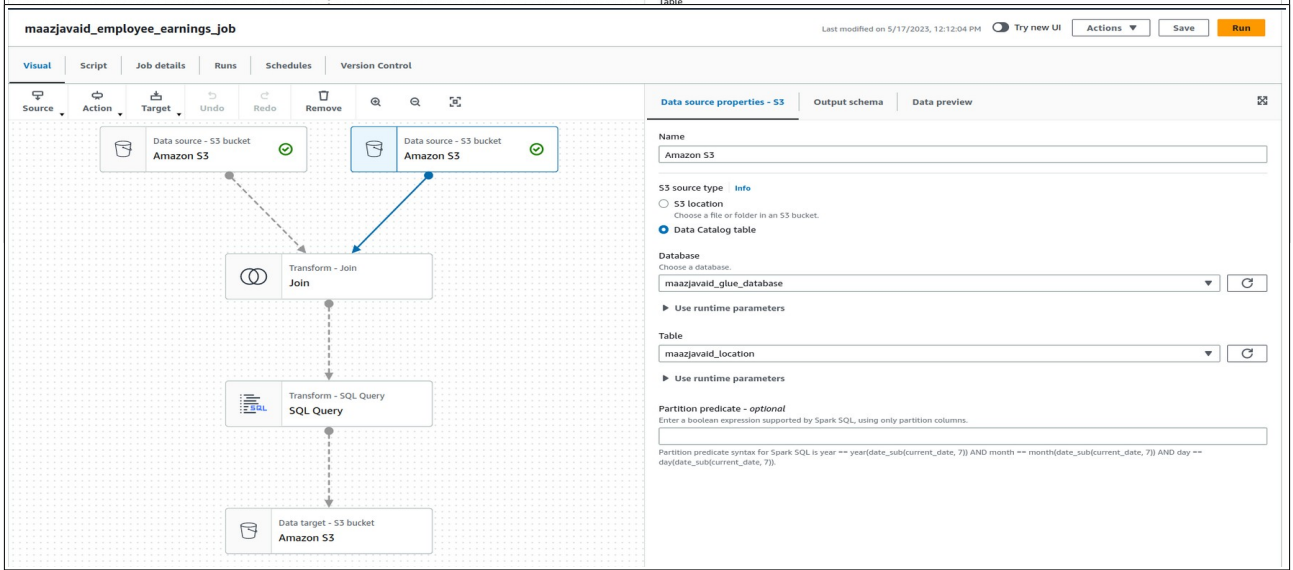
A. Creating a job in AWS Glue is necessary to define the ETL process that we want to run on our data. By creating a job, we can specify the data source that we want to use, the ETL script that we want to run, and the output data location where we want to write the transformed data.

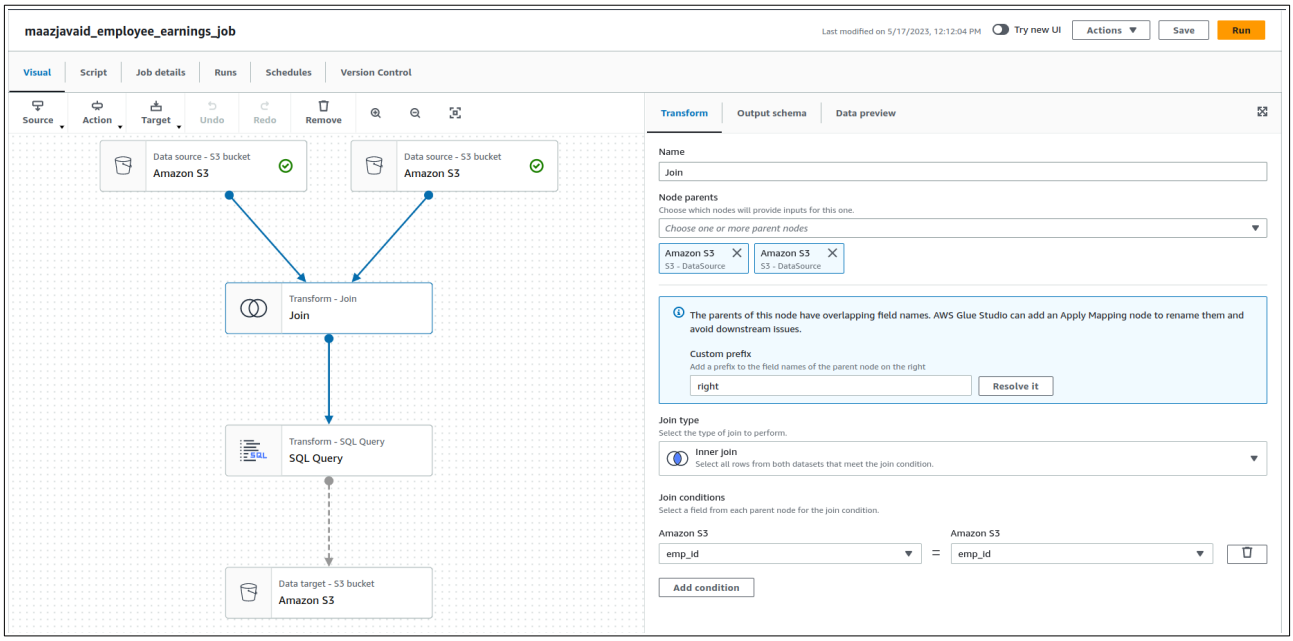
Q2. What is this service's purpose?

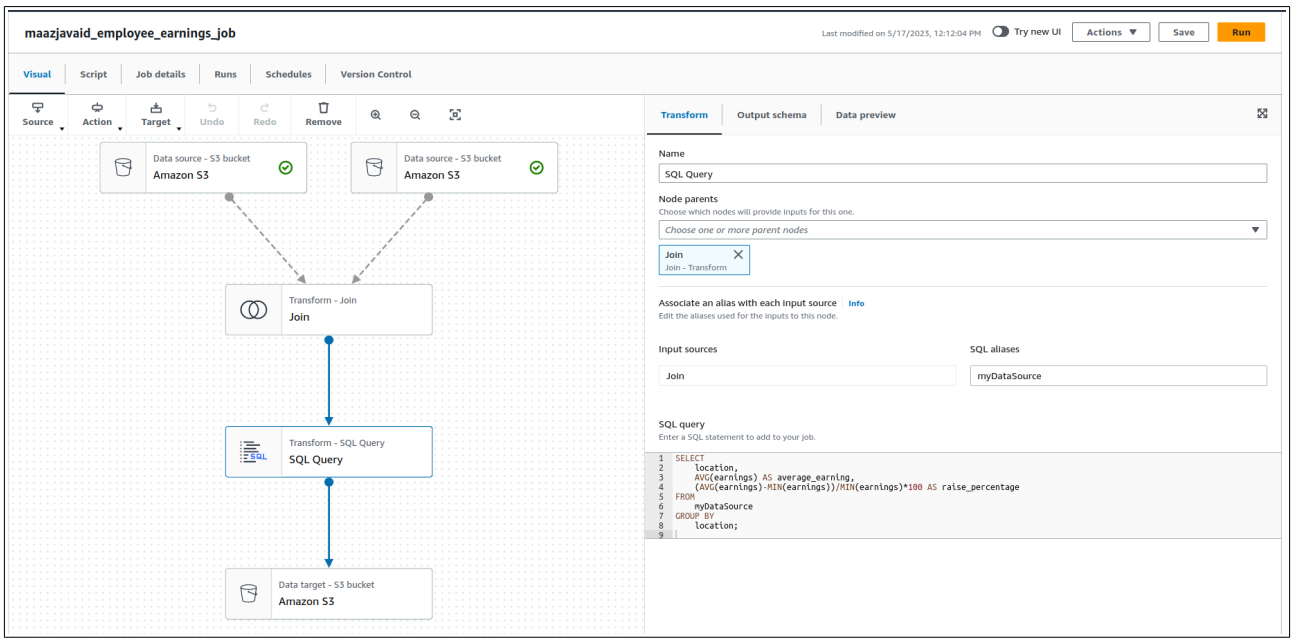
A. AWS Glue's purpose is to simplify the ETL process and make it accessible to a wider range of users, regardless of their technical expertise. It provides a powerful and flexible platform for processing data at scale and preparing it for analysis, machine learning, and other downstream applications.

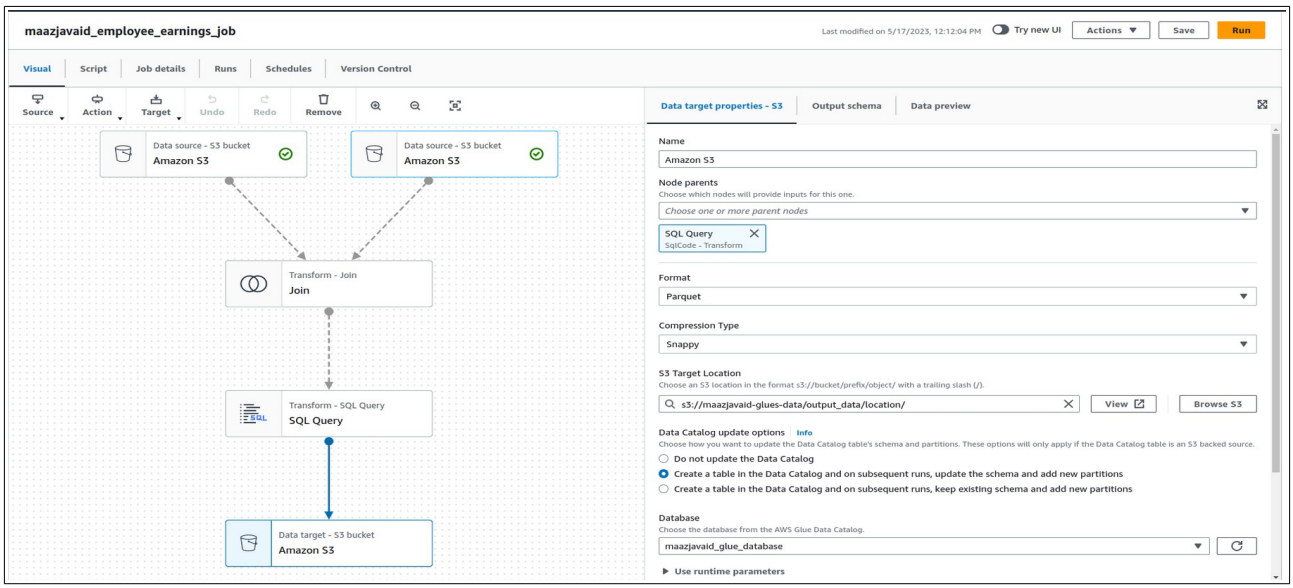


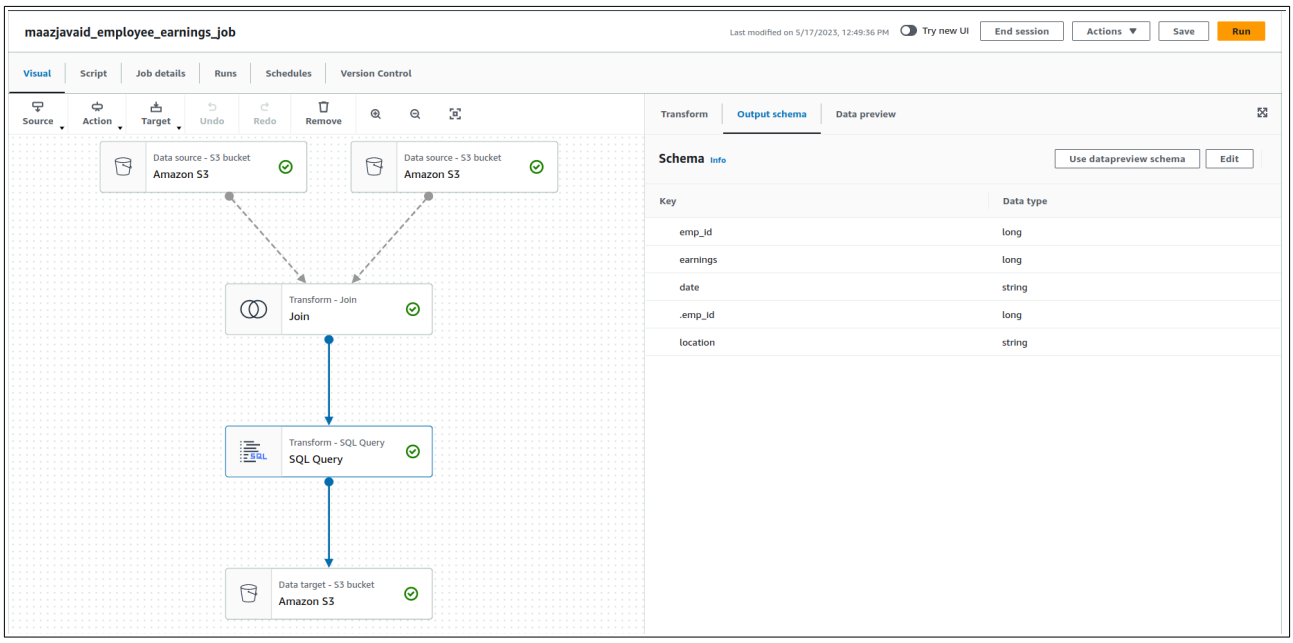


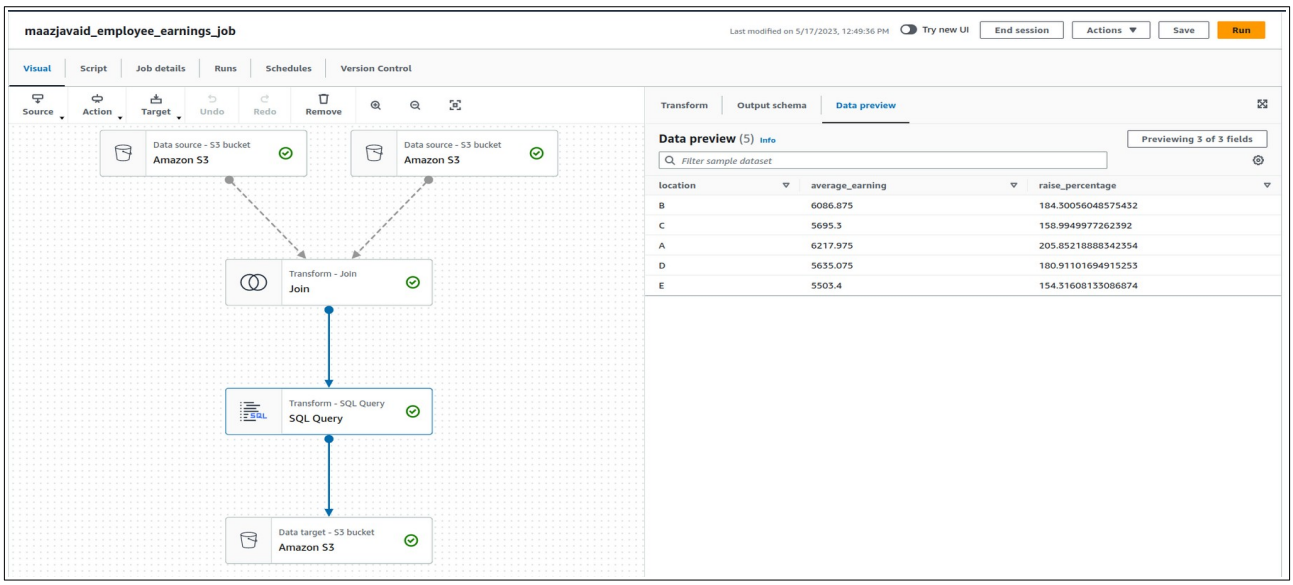



maazjavaid_employee_earnings_job
Last modified on 5/17/2023, 12:12:04 PM
Try new UI
Actions
Save
Run
Visual
Script
Job details
Runs
Schedules
Version Control
Source
Action
Target
Undo
Redo
Remove
Data source - S3 bucket
Amazon S3
Data source - S3 bucket
Amazon S3
Transform - Join
Join
Transform - SQL 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 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


maazjavaid_employee_earnings_job
Last modified on 5/17/2023, 12:12:04 PM
Try new UI
Actions
Save
Run
Visual
Script
Job details
Runs
Schedules
Version Control
Source
Action
Target
Undo
Redo
Remove
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
Join
myDataSource
SQL query
Enter a SQL statement to add to your job.
SELECT
location,
AVG(earnings) AS average_earning,
(AVG(earnings)-MIN(earnings))/MIN(earnings)*100 AS raise_percentage
FROM
myDataSource
GROUP BY
location;


maazjavaid_employee_earnings_job
Last modified on 5/17/2023, 12:12:04 PM
Try new UI
Actions
Save
Run
Visual
Script
Job details
Runs
Schedules
Version Control
Source
Action
Target
Undo
Redo
Remove
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://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


maazjavaid_employee_earnings_job
Last modified on 5/17/2023, 12:49:36 PM
Try new UI
End session
Actions
Save
Run
Visual
Script
Job details
Runs
Schedules
Version Control
Source
Action
Target
Undo
Redo
Remove
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


maazjavaid_employee_earnings_job
Last modified on 5/17/2023, 12:49:36 PM
Try new UI
End session
Actions
Save
Run
Visual
Script
Job details
Runs
Schedules
Version Control
Source
Action
Target
Undo
Redo
Remove
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
184.30056048575432
C
5695.3
158.9949977262592
A
6217.975
205.85218888342354
D
5635.075
180.91101694915253
E
5503.4
154.31608133086874