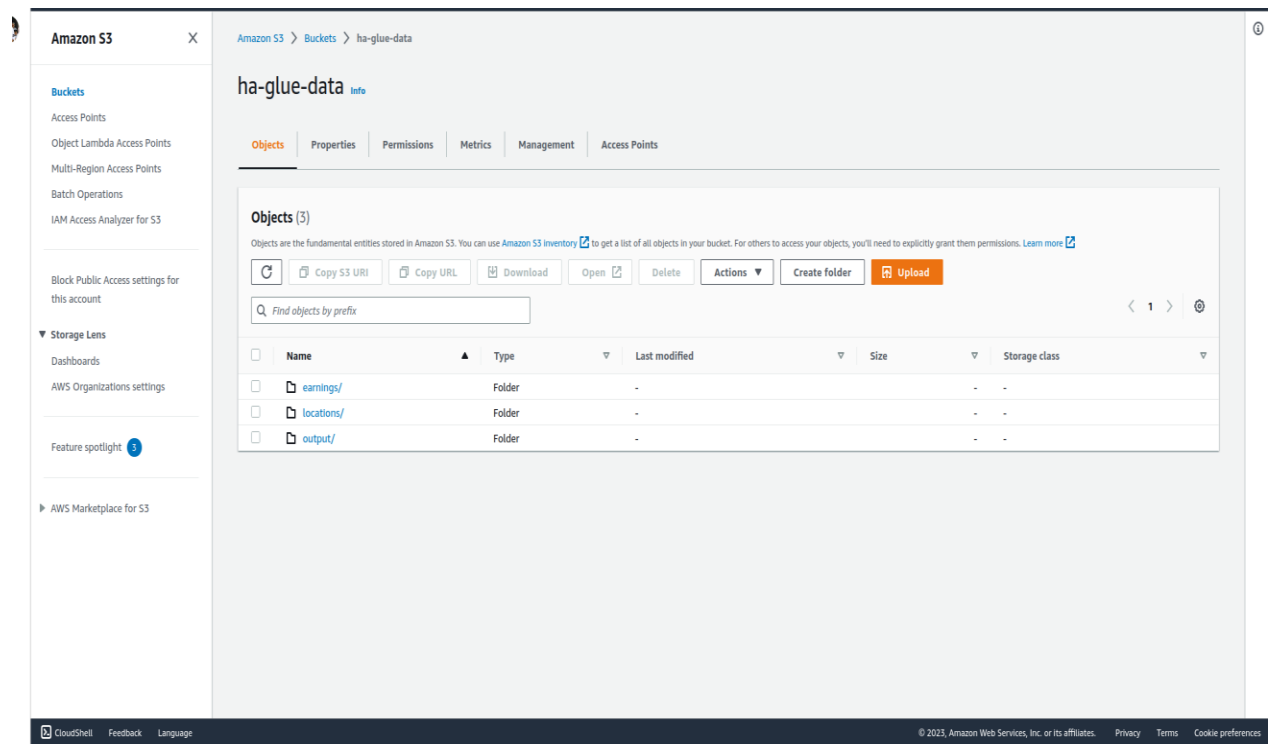


Members

Huzefa Anver (2303.KHI.DEG.002)

Syed Mohammad Anjil Hussain Rizvi (2303.KHI.DEG.031)

Ha-glue-data s3 bucket



This is our bucket containing 3 directories earnings and location containing our input data
And output containing our output data after processing it using glue job crawlers.

AWS Glue Crawler

The screenshot displays the AWS Glue console interface. On the left is a navigation sidebar with categories like 'Getting started', 'Data Catalog', and 'Data Integration and ETL'. The main panel is titled 'Crawlers' and includes a description: '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.' Below this is a table of crawlers. One crawler, 'ha_glue_crawler', is listed with a 'Ready' state and a 'Succeeded' last run. The table headers are: Name, State, Schedule, Last run, Last run timestamp, Log, and Table changes from la... The footer of the console shows '© 2023, Amazon Web Services, Inc. or its affiliates.' and links for 'Privacy', 'Terms', and 'Cookie preferences'.

Name	State	Schedule	Last run	Last run timestamp	Log	Table changes from la...
ha_glue_crawler	Ready		Succeeded	May 22, 2023 at 13:22:57	View log	2 created

As we can see our crawler is up and running, which we will use to extract the data from s3 bucket from multiple directories earnings and locations, process it and load it into our output folder

Aws glue job

The screenshot displays the AWS Glue console interface for a job named **ass-5.2_job**. The job is in a 'Successfully started' state. The 'Visual' tab is active, showing a workflow graph with the following components:

- Data sources:** Two 'Data source - S3 bucket Amazon S3' nodes at the top.
- Transform - Join:** A central 'Join' action receiving input from both data sources.
- Transform - SQL Query:** An action receiving input from the 'Join' transform, performing a 'SQL Query'.
- Data target:** A 'Data target - S3 bucket Amazon S3' node at the bottom receiving output from the 'SQL Query' transform.

On the right side, the 'Output schema' tab is selected, displaying the following schema:

Key	Data type	Partition
emp_id	long	-
location	string	-
.emp_id	long	-
earnings	long	-

This is the overall glue job we created. As we can see the graph, we have 2 sources of our data, both directing towards the action we created join which will combine the data and then we will perform some queries from combine data and then the queried data will be saved in our output folder.

Performing query

Transform

Output schema

Data preview

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.

```
1 SELECT
2   location,
3   AVG(earnings) AS average_earnings,
4   (AVG(earnings) - MIN(earnings)) / MIN(earnings) * 100 AS raise_percentage
5 FROM
6   myDataSource
7 GROUP BY
8   location:
```

This is our query we performed after combining the data

Query preview

The screenshot shows the AWS Glue console interface. On the left is a navigation menu with options like 'Getting started', 'ETL jobs', 'Visual ETL', 'Notebooks', 'Job run monitoring', 'Data Catalog tables', 'Data connections', 'Workflows (orchestration)', 'Data Catalog', 'Databases', 'Tables', 'Stream schema registries', 'Schemas', 'Connections', 'Crawlers', 'Classifiers', 'Catalog settings', 'Data Integration and ETL', 'AWS Glue Studio', 'Jobs', 'Interactive Sessions', 'Notebooks', 'Data classification tools', 'Sensitive data detection', and 'Record Matching'. The main panel displays a job named 'ass-5.2_job'. The 'Visual' tab is selected, showing a workflow diagram with three nodes: 'Data source - S3 bucket Amazon S3', 'Transform - Job Job', and 'Transform - SQL Query'. The 'Data preview' tab is active, showing a table with columns 'location', 'average_earnings', and 'raise_percentage'. The preview shows one row with values 'A', '5926.05', and '191.49286768322676'.

This is the preview we got after querying the data

Output directory

The screenshot shows the Amazon S3 console interface. On the left is a navigation menu with options like 'Buckets', 'Access Points', 'Object Lambda Access Points', 'Multi-Region Access Points', 'Batch Operations', 'IAM Access Analyzer for S3', 'Block Public Access settings for this account', 'Storage Lens', 'Dashboards', 'AWS Organizations settings', 'Feature spotlight', and 'AWS Marketplace for S3'. The main panel displays a bucket named 'run-1684765431236-part-00000'. The 'Objects' tab is active, showing a list of 36 objects. The objects are named 'run-1684765431236-part-00000' through 'run-1684765431236-part-00036'. The objects are all of type 'Standard' and have a size of 43.0 B. The last modified date is May 22, 2023, 18:50:39 (UTC+05:00).

Finally, as we can see our data is loaded in our output folder which we created in our s3 bucket