

**NAME: Mohammad Hussam (2303.KHI.DEG.020)**  
**PARING WITH : MAVIA ALAM KHAN(2302.KHI.DEG.017)**

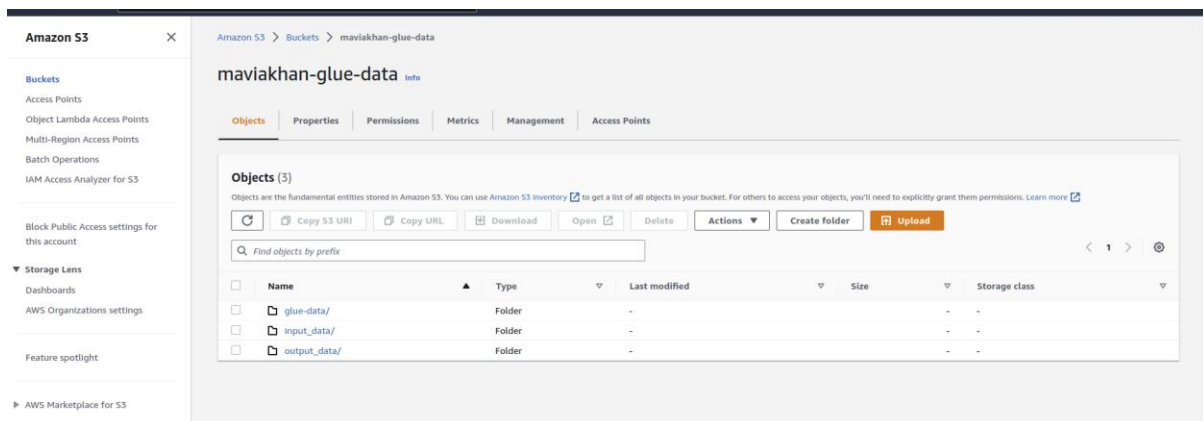
## **ASSIGNMENT NO :5.2**

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.

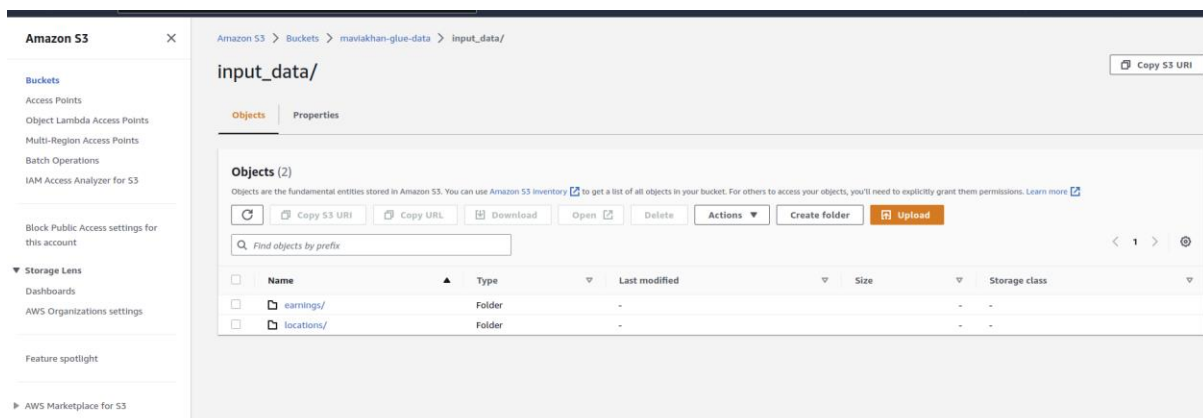
Create a Glue job that aggregates the data based on the office location to calculate average salaries and raise percentages for these locations.

## **SOLUTION**

First create a folder in s3 bucket name (input and output)



In input folder we store a two dataset earning.csv and location.csv



Now we create a crawler and extract the meta data

The screenshot shows the AWS Glue console interface. On the left is a navigation pane with categories like 'Getting started', 'Data Catalog', and 'Data Integration and ETL'. The main panel displays the configuration for the 'mavia\_combined\_employee\_earnings\_crawler'. A green banner at the top indicates 'Crawler successfully starting'. The crawler properties are as follows:

Property	Value
Name	mavia_combined_employee_earnings_crawler
IAM role	maviakhan-glue-role
Database	mavia-glue-database
State	READY
Description	-
Security configuration	-
Lake Formation configuration	-
Table prefix	mavia_
Maximum table threshold	-

Below the properties, the 'Crawler runs' tab shows a table of recent runs:

Start time (UTC)	End time (UTC)	Current/last duration	Status	DPU hours	Table changes
May 19, 2023 at 16:52:55	May 19, 2023 at 16:53:44	48 s	Completed	0.065	1 table change, 0 partition changes
May 19, 2023 at 11:55:06	May 19, 2023 at 11:55:56	49 s	Completed	0.067	3 table changes, 0 partition changes

The screenshot shows the AWS Glue console interface for the 'maviakhan\_s3\_earnings\_crawler'. A green banner at the top indicates 'Crawler successfully starting'. The crawler properties are as follows:

Property	Value
Name	maviakhan_s3_earnings_crawler
IAM role	maviakhan-glue-role
Database	mavia-glue-database
State	READY
Description	-
Security configuration	-
Lake Formation configuration	-
Table prefix	maviakhan
Maximum table threshold	-

Below the properties, the 'Crawler runs' tab shows a table of recent runs:

Start time (UTC)	End time (UTC)	Current/last duration	Status	DPU hours	Table changes
May 19, 2023 at 16:55:51	May 19, 2023 at 16:56:41	49 s	Completed	0.060	1 table change, 1 partition change
May 19, 2023 at 04:26:04	May 19, 2023 at 04:26:54	50 s	Completed	0.068	1 table change, 1 partition change

And after create the glue crawler we start creating the job

The screenshot displays the AWS Glue console interface. On the left, a navigation pane lists various ETL jobs and data catalog components. The main workspace shows a workflow diagram with the following components:

- Data source - S3 bucket AmazonLocation**: Connected to the **Transform - Join** node.
- Data source - S3 bucket AmazonEarning**: Connected to the **Transform - Join** node.
- Transform - Join**: A join node that receives input from both data sources.
- Transform - SQL Query**: A node that receives input from the join node and performs an SQL query.
- Data target - S3 bucket datatarget**: The final destination for the data, receiving input from the SQL query node.

On the right, the **Data source properties - S3** panel is open for the **AmazonLocation** source. The configuration includes:

- Name**: AmazonLocation
- S3 source type**: S3 location (selected)
- S3 location**: Choose a file or folder in an S3 bucket.
- S3 URL**: `s3://maviakhan-glue-data/input_data/locations/`
- Recursive**: ☒ (Read files in all subdirectories.)
- Data format**: CSV
- Delimiter**: Comma (,)
- Escape character - optional**: (Empty field)
- Quote character**: Double quote (")
- First line of source file contains column headers**: ☒

Now we create two s3 source one is employee earning data and location, we perform inner join on both data on emp\_id and after that prepare for querying

The screenshot displays the AWS Glue console interface for a workflow named **Assignmnt\_5.2**. The main workspace shows a workflow diagram with the following components:

- Data source - S3 bucket AmazonLocation**: Connected to the **Transform - Join** node.
- Data source - S3 bucket AmazonEarning**: Connected to the **Transform - Join** node.
- Transform - Join**: A join node that receives input from both data sources.
- Transform - SQL Query**: A node that receives input from the join node and performs an SQL query.
- Data target - S3 bucket datatarget**: The final destination for the data, receiving input from the SQL query node.

On the right, the **Transform** panel is open for the **Join** node. The configuration includes:

- Name**: Join
- Node parents**: Choose which nodes will provide inputs for this one. (AmazonLocation, AmazonEarning)
- Join type**: Inner join (selected)
- Join conditions**: Select a field from each parent node for the join condition. (AmazonEarning emp\_id = AmazonLocation emp\_id)

**AWS Glue** × **Assignmnt\_5.2** Last modified on 5/19/2023, 10:26:41 PM Try new UI End session Actions Save Run

Visual Script Job details Runs Schedules Version Control

Source Action Target Undo Redo Remove

Transform Output schema Data preview

```

graph TD
    A[Data source - S3 bucket AmazonLocation] --> C[Transform - Join Join]
    B[Data source - S3 bucket AmazonEarning] --> C
    C --> D[Transform - SQL Query SQL Query]
    D --> E[Data target - S3 bucket datatarget]
  
```

**Name**  
SQL Query

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

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

**Input sources**  
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;
9
  
```

**AWS Glue** × **Assignmnt\_5.2** Last modified on 5/19/2023, 10:26:41 PM Try new UI End session Actions Save Run

Visual Script Job details Runs Schedules Version Control

Source Action Target Undo Redo Remove

Data target properties - S3 Output schema Data preview

```

graph TD
    A[Data source - S3 bucket AmazonLocation] --> C[Transform - Join Join]
    B[Data source - S3 bucket AmazonEarning] --> C
    C --> D[Transform - SQL Query SQL Query]
    D --> E[Data target - S3 bucket datatarget]
  
```

**Name**  
datatarget

**Node parents**  
Choose which nodes will provide inputs for this one.  
Choose one or more parent node  
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://mavikhan-glue-data/output\_data/earningswithLocato 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**

Now here the queries based on the salaries and percentage of these locations.

