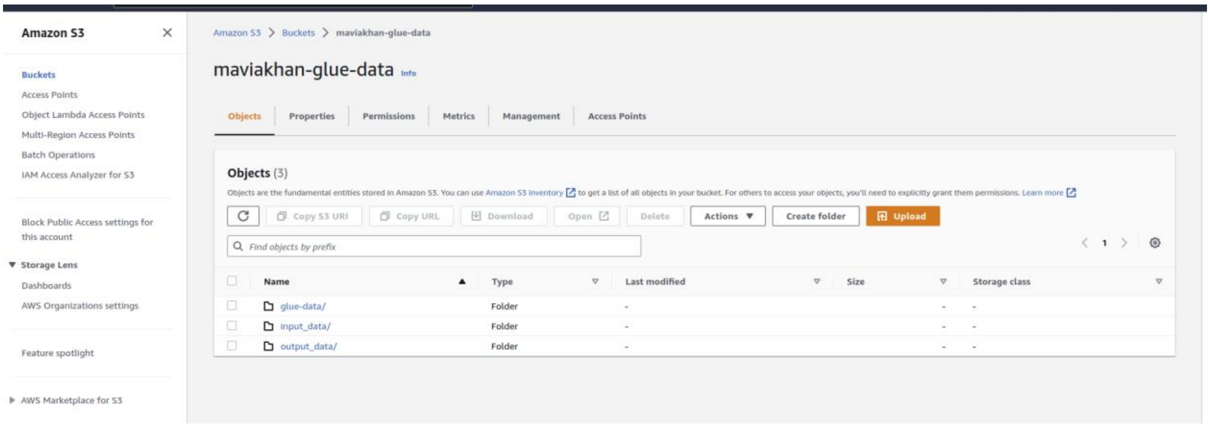
NAME: Mohammad Hussam (2303.KHI.DEG.020)
PARING WITH : MAVIA ALAM KHAN(2302.KHI.DEG.017)
&
AQSA TAUHEED(2303.KHI.DEG.011)

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)



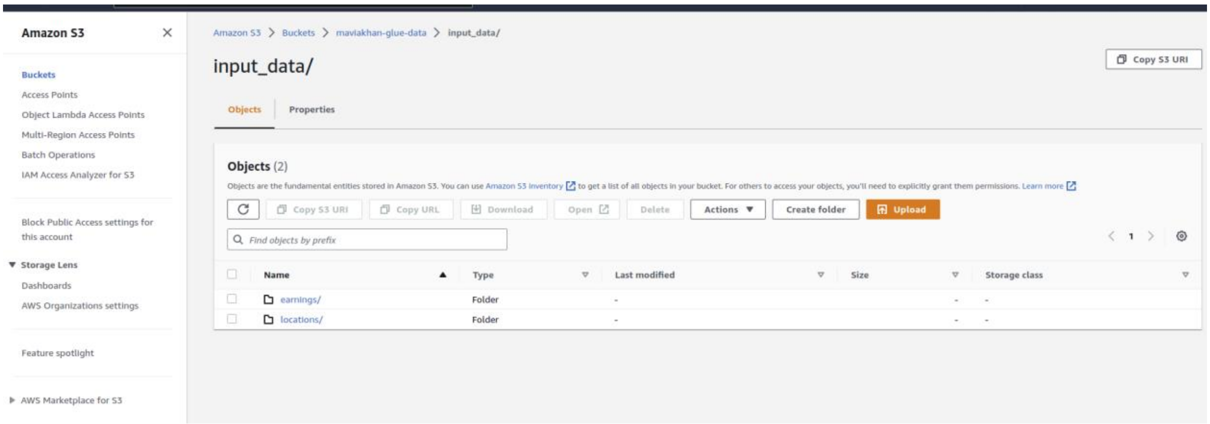
maviakhan-glue-data

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	-	-	-

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



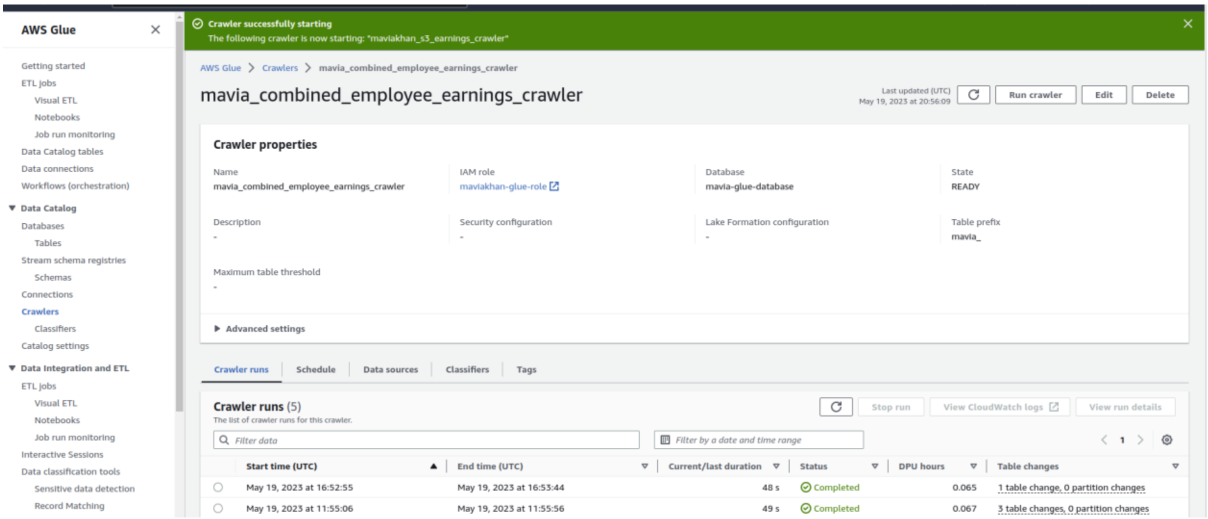
input_data/

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	-	-	-
□	locations/	Folder	-	-	-

Now we create a crawler and extract the meta data



Crawler successfully starting
The following crawler is now starting: "maviakhan_s3_earnings_crawler"

AWS Glue > Crawlers > mavia_combined_employee_earnings_crawler

mavia_combined_employee_earnings_crawler

Last updated (UTC)
May 19, 2023 at 20:56:09

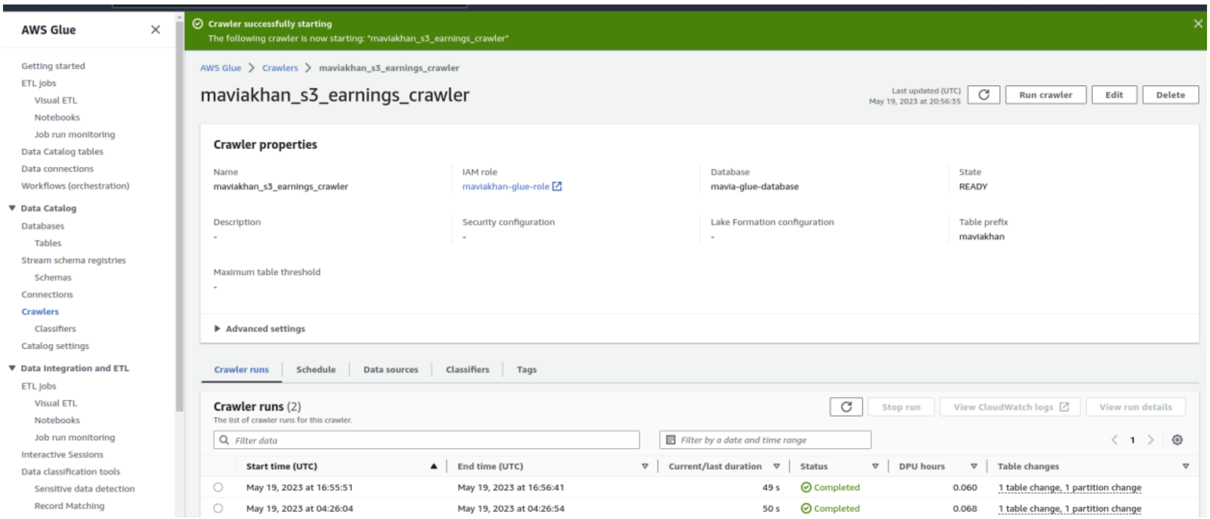
Crawler properties

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

▶ Advanced settings

Crawler runs (5)
The list of crawler runs for this crawler.

	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



Crawler successfully starting
The following crawler is now starting: "maviakhan_s3_earnings_crawler"

AWS Glue > Crawlers > maviakhan_s3_earnings_crawler

maviakhan_s3_earnings_crawler

Last updated (UTC)
May 19, 2023 at 20:56:35

Crawler properties

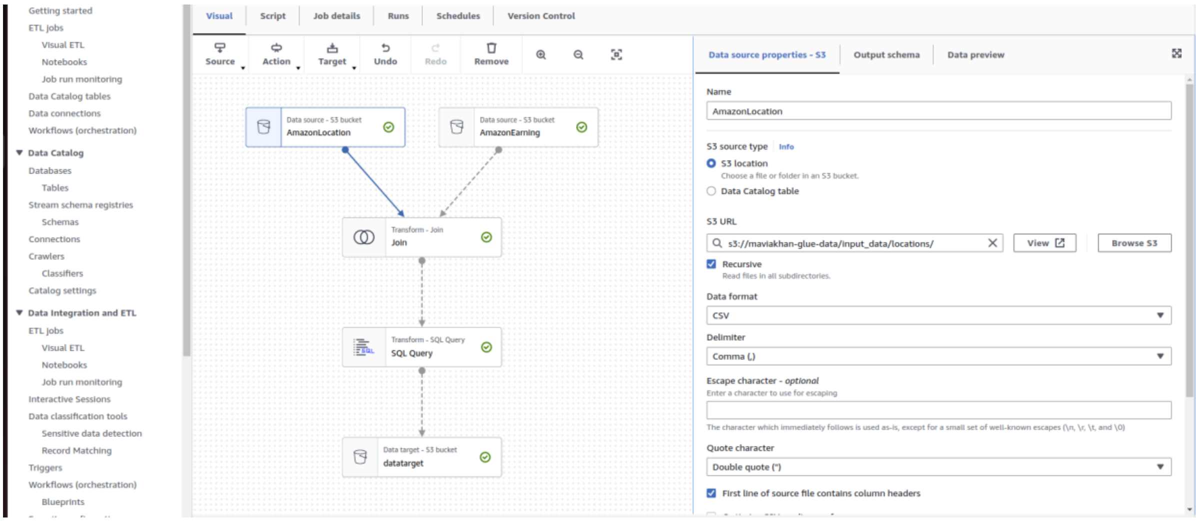
Name	IAM role	Database	State
maviakhan_s3_earnings_crawler	maviakhan-glue-role	mavia-glue-database	READY
Description	**Security configuration**	**Lake Formation configuration**	**Table prefix**
-	-	-	maviakhan
Maximum table threshold			
-			

▶ Advanced settings

Crawler runs (2)
The list of crawler runs for this crawler.

	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



Name: AmazonLocation

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

S3 URL: s3://maviakhan-glue-data/input_data/locations/
View
Browse S3

☑ Recursive
Read files in all subdirectories.

Data format: CSV

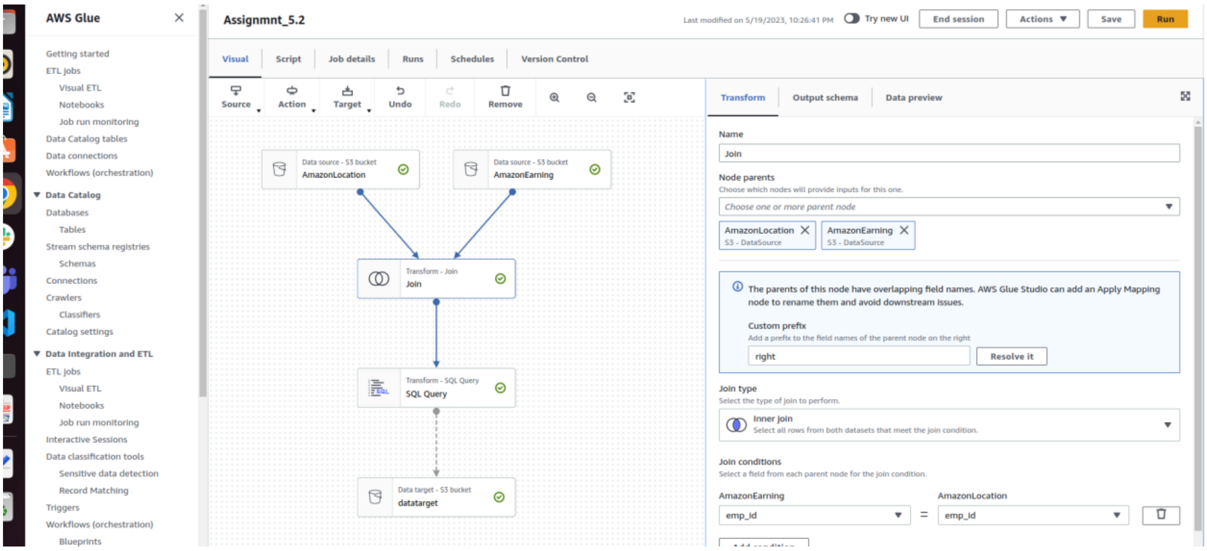
Delimiter: Comma (,)

Escape character - *optional*
Enter a character to use for escaping
The character which immediately follows is used as-is, except for a small set of well-known escapes (\n, \r, \t, and \0)

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



Assignmnt_5.2

Last modified on 5/19/2023, 10:26:41 PM

Name: Join

Node parents
Choose which nodes will provide inputs for this one.
Choose one or more parent node
AmazonLocation (S3 - DataSource)
AmazonEarning (S3 - DataSource)

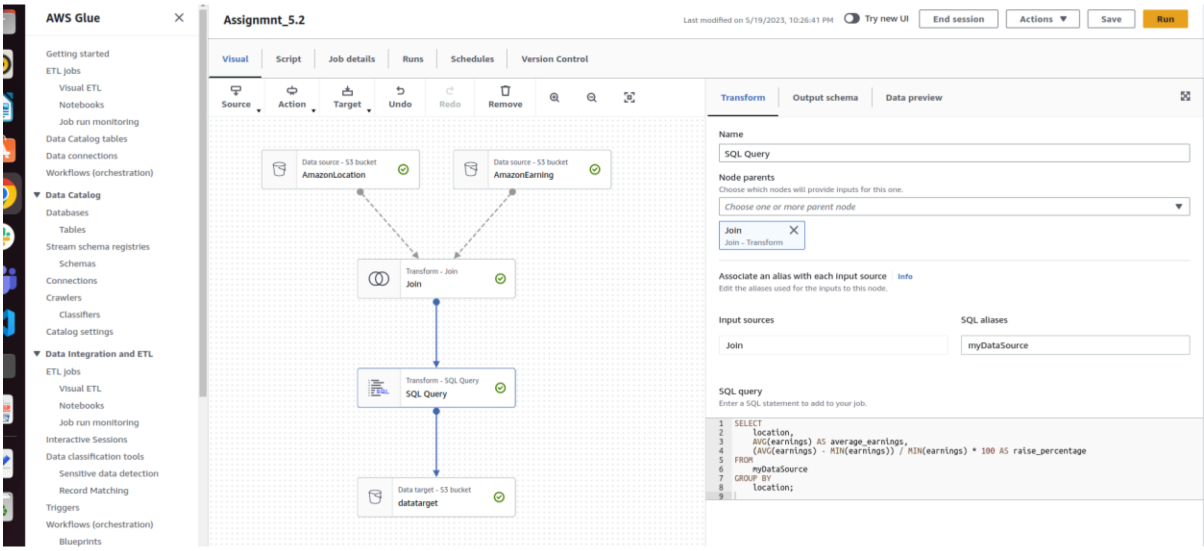
The parents of this node have overlapping field names. AWS Glue Studio can add an Apply Mapping node to rename them and avoid downstream issues.

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.

AmazonEarning		AmazonLocation
emp_id	=	emp_id



Assignmnt_5.2

Last modified on 5/19/2023, 10:26:41 PM

Name: SQL Query

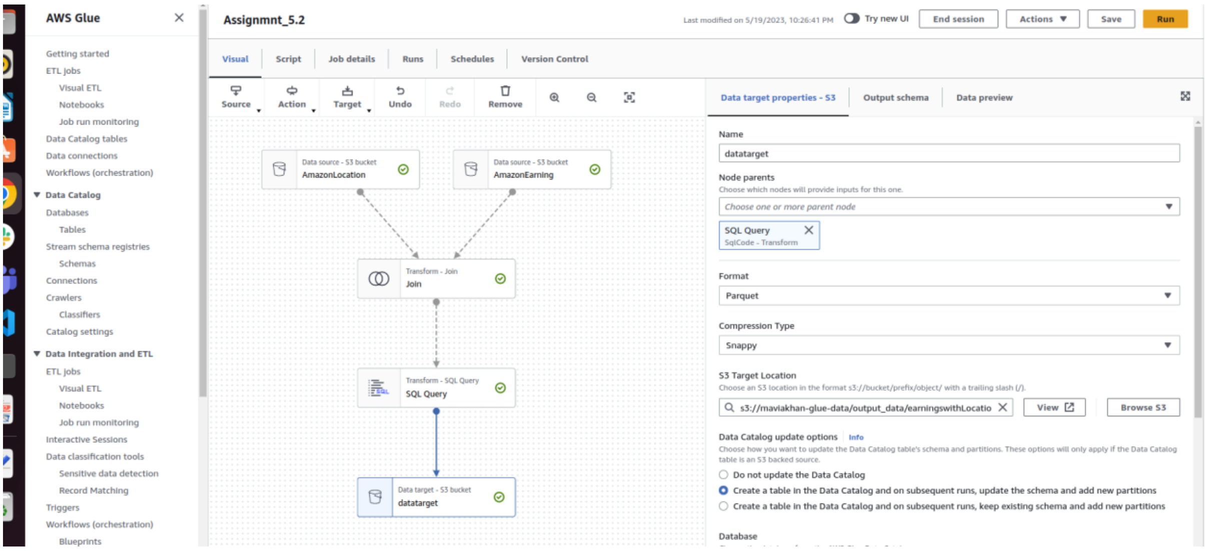
Node parents
Choose which nodes will provide inputs for this one.
Choose one or more parent node
Join (Join - Transform)

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

Input sources	SQL aliases
Join	myDataSource

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

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



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://maviakhan-glue-data/output_data/earningswithLocatio

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.

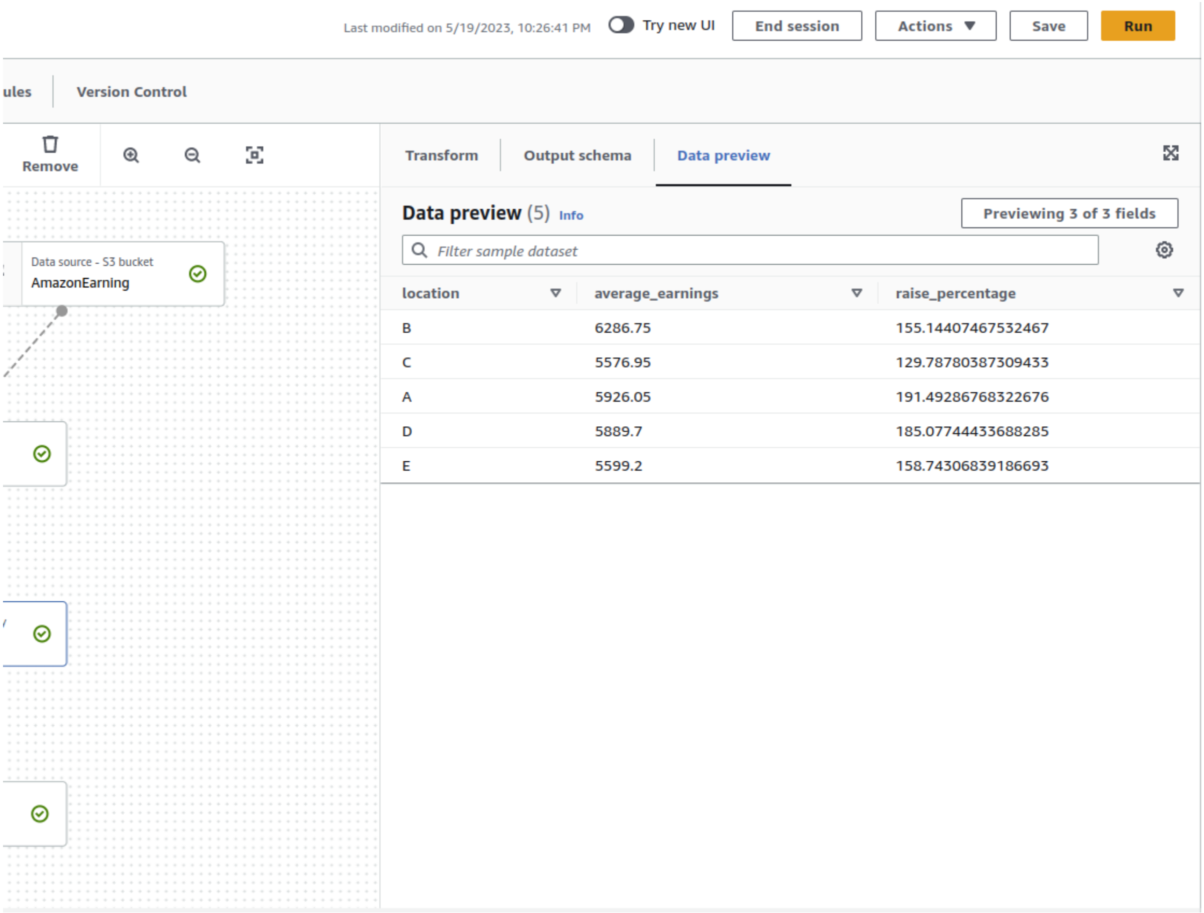
○ Do not update the Data Catalog

● 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

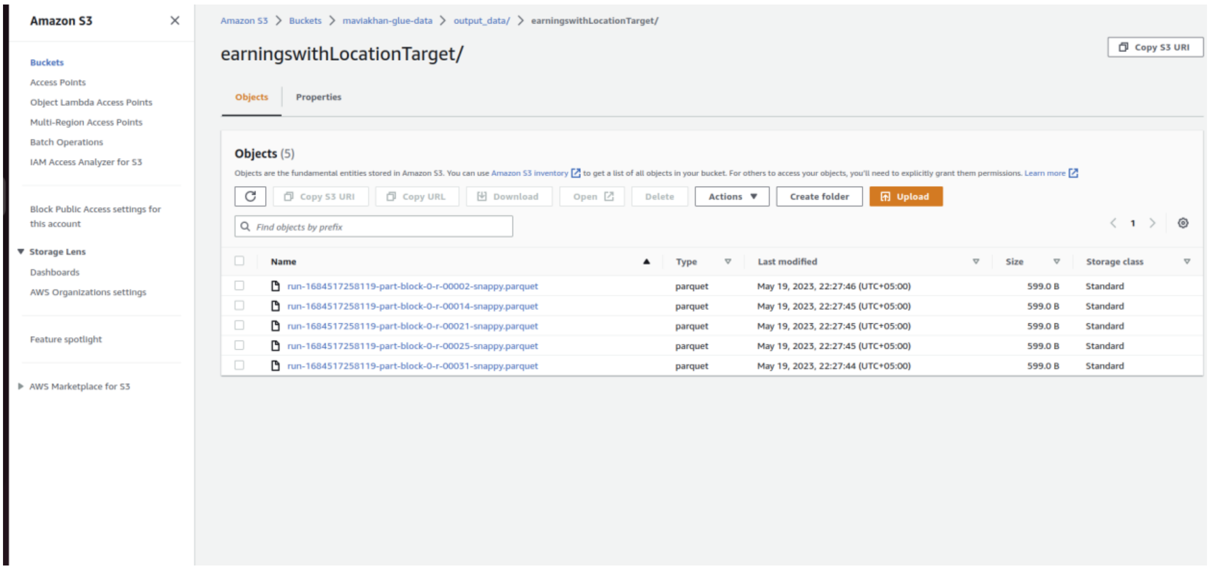
Now here the queries based on the salaries and percentage of these locations.



Data preview (5) Info

location	average_earnings	raise_percentage
B	6286.75	155.14407467532467
C	5576.95	129.78780387309433
A	5926.05	191.49286768322676
D	5889.7	185.07744433688285
E	5599.2	158.74306839186693

Now finally we load the data and show in the output folder


Amazon S3
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
AWS Marketplace for S3
Amazon S3 > Buckets > maviakhan-glue-data > output_data/ > earningswithLocationTarget/
Copy S3 URI
Objects
Properties
Copy S3 URI
Copy URL
Download
Open
Delete
Actions
Create folder
Upload
Find objects by prefix
1

earningswithLocationTarget/

Objects (5)

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
□	run-1684517258119-part-block-0-r-00002-snappy.parquet	parquet	May 19, 2023, 22:27:46 (UTC+05:00)	599.0 B	Standard
□	run-1684517258119-part-block-0-r-00014-snappy.parquet	parquet	May 19, 2023, 22:27:45 (UTC+05:00)	599.0 B	Standard
□	run-1684517258119-part-block-0-r-00021-snappy.parquet	parquet	May 19, 2023, 22:27:45 (UTC+05:00)	599.0 B	Standard
□	run-1684517258119-part-block-0-r-00025-snappy.parquet	parquet	May 19, 2023, 22:27:45 (UTC+05:00)	599.0 B	Standard
□	run-1684517258119-part-block-0-r-00031-snappy.parquet	parquet	May 19, 2023, 22:27:44 (UTC+05:00)	599.0 B	Standard