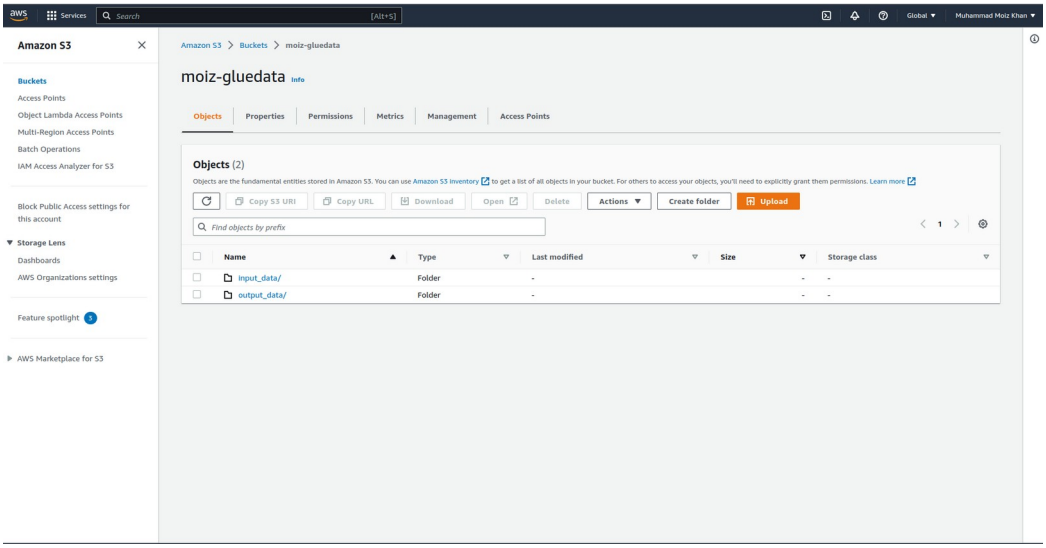
Talha Khan (2303.009.KHI.DEG)
Muhammad Moiz Khan (2303.022.KHI.DEG)

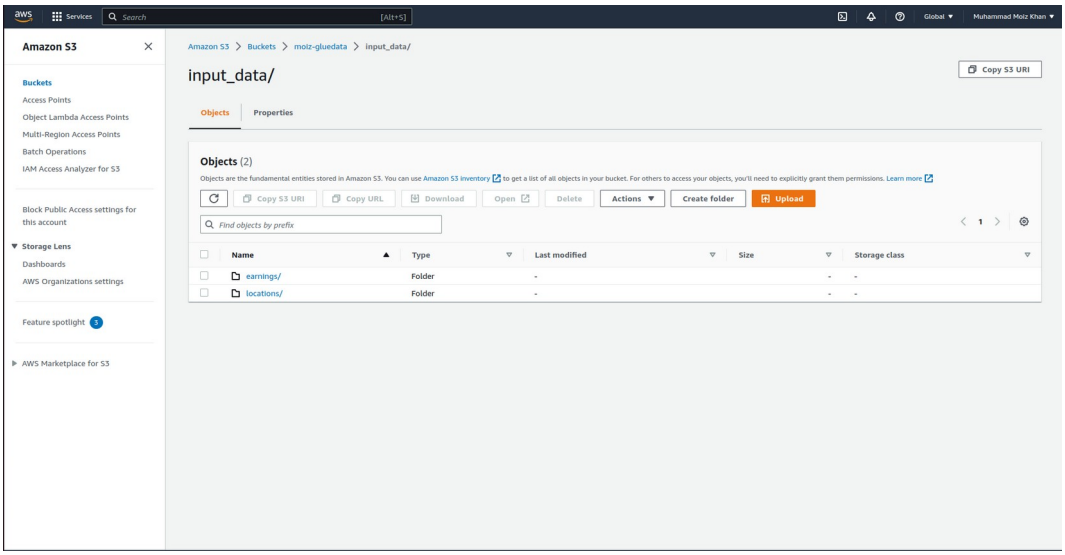
ASSIGNMNET 5.2

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.

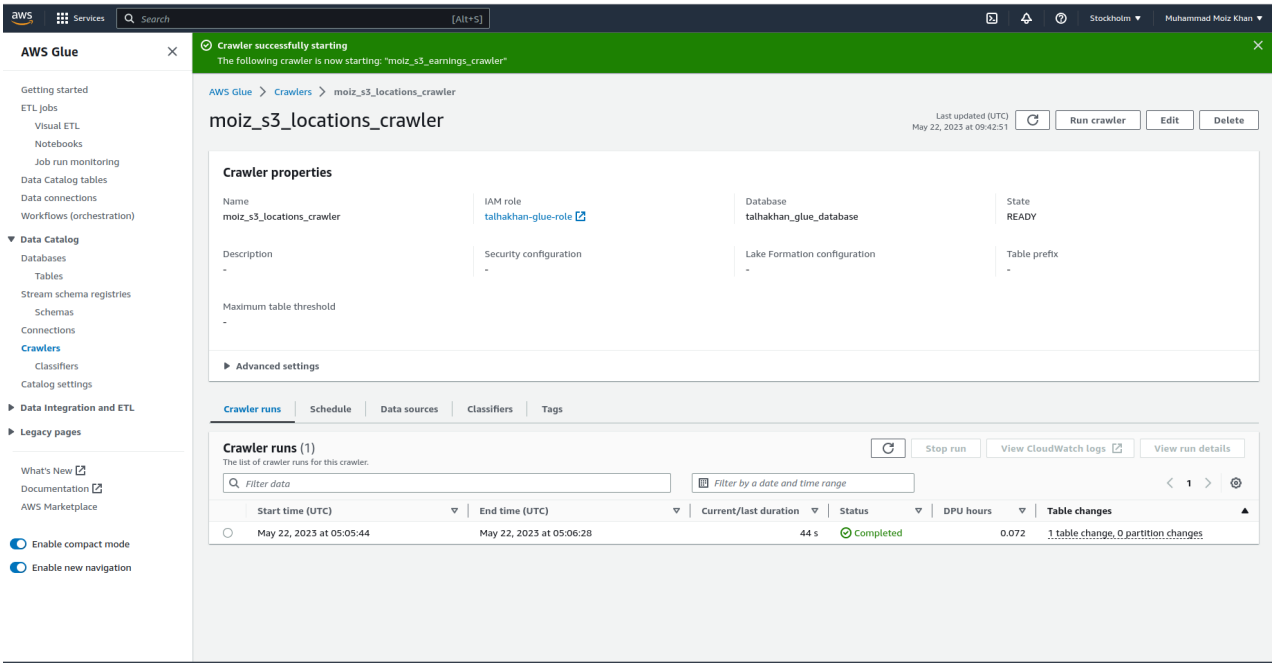
1) First create a folder in s3 bucket name (input and output data):

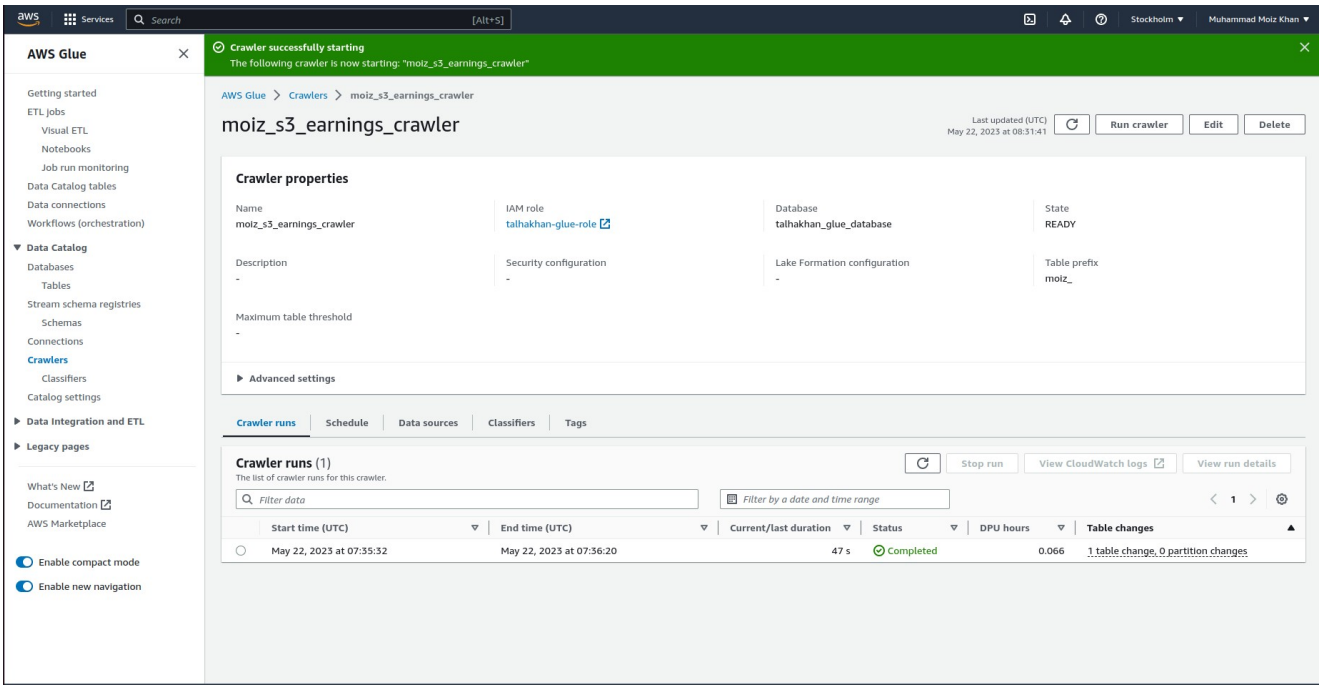


2) Next, we store a two dataset earning.csv and location.csv in input folder:

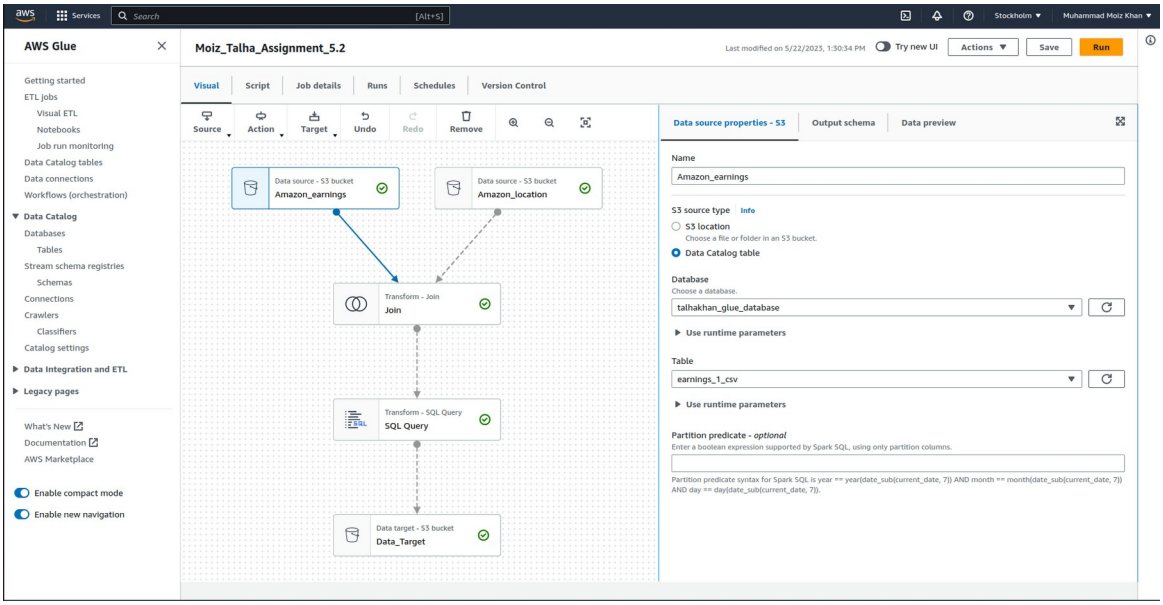


3) Now we create two crawlers named earnings and locations respectively, and extract the meta data :

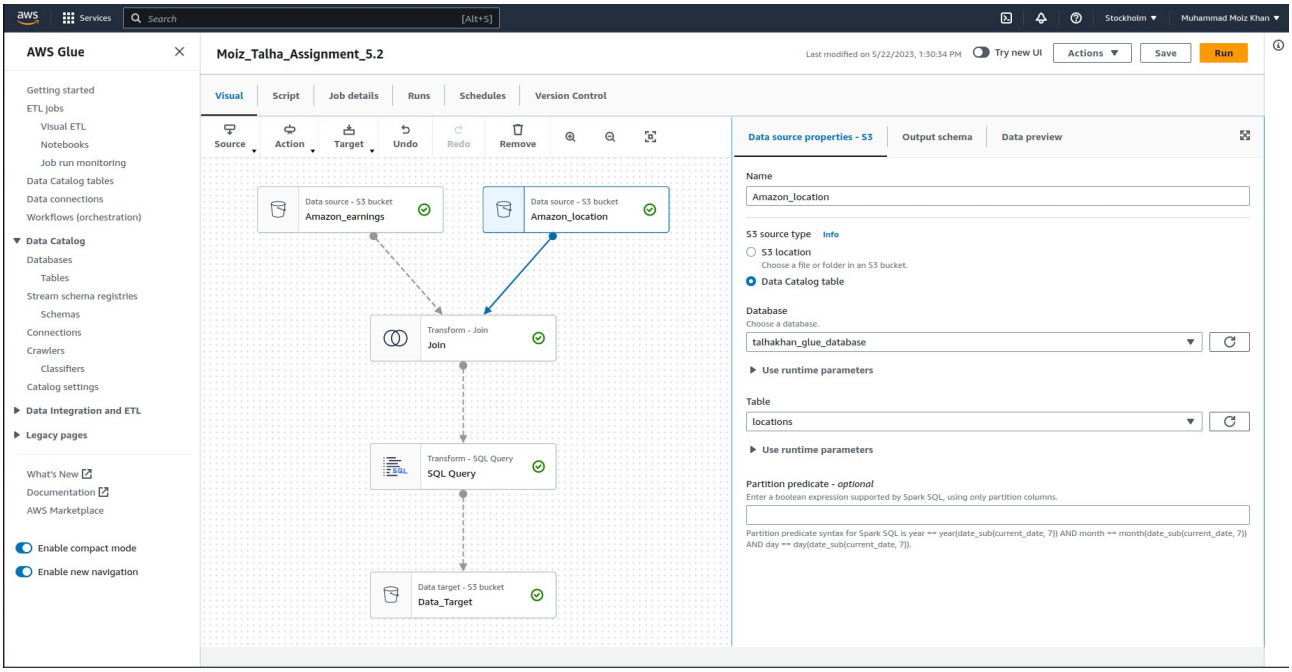


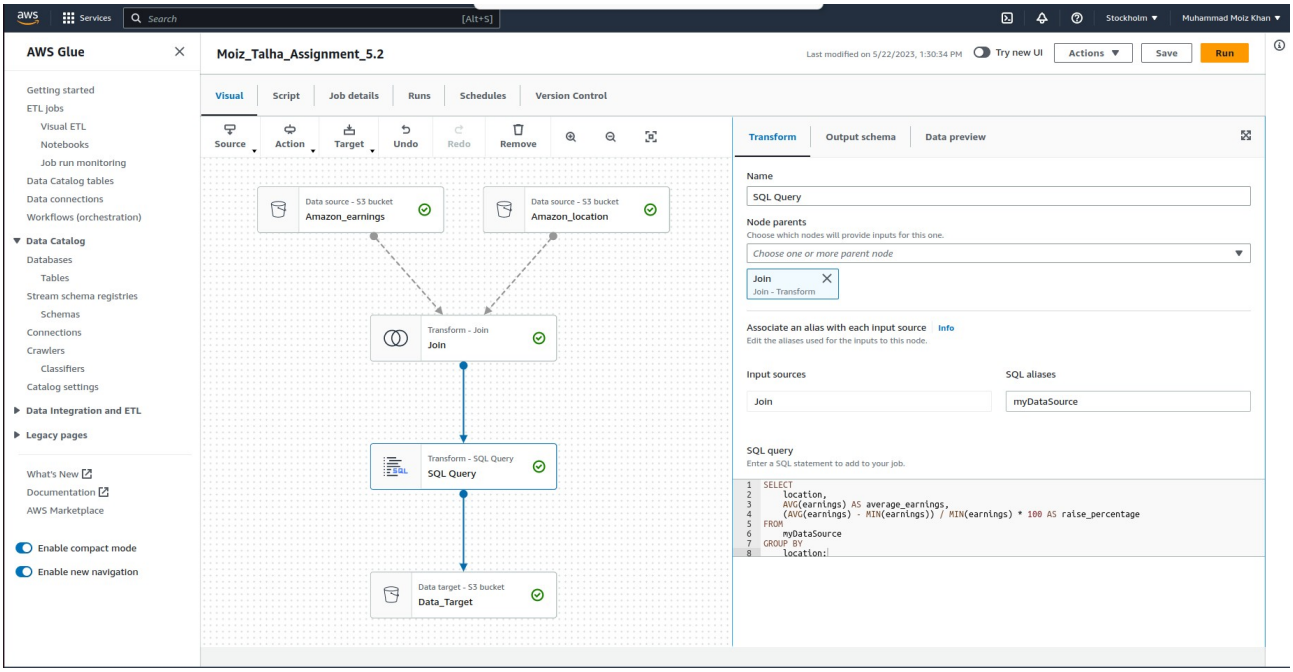


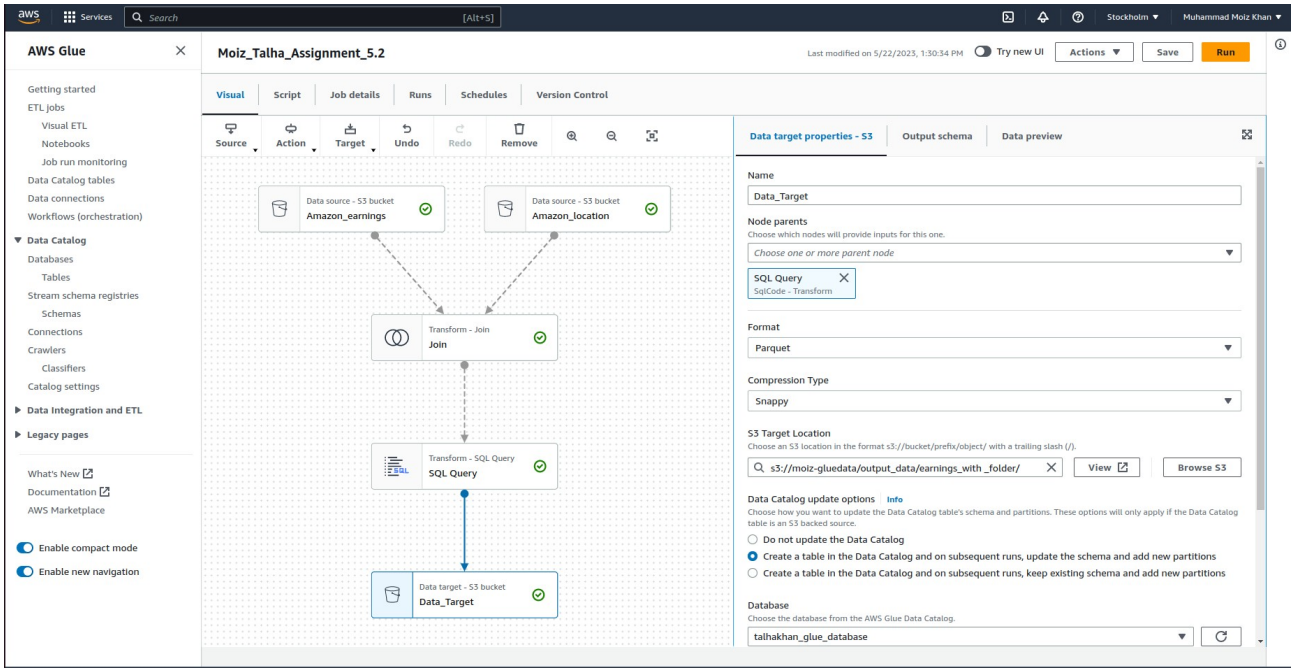
4) After creating the glue crawler we start creating the job:



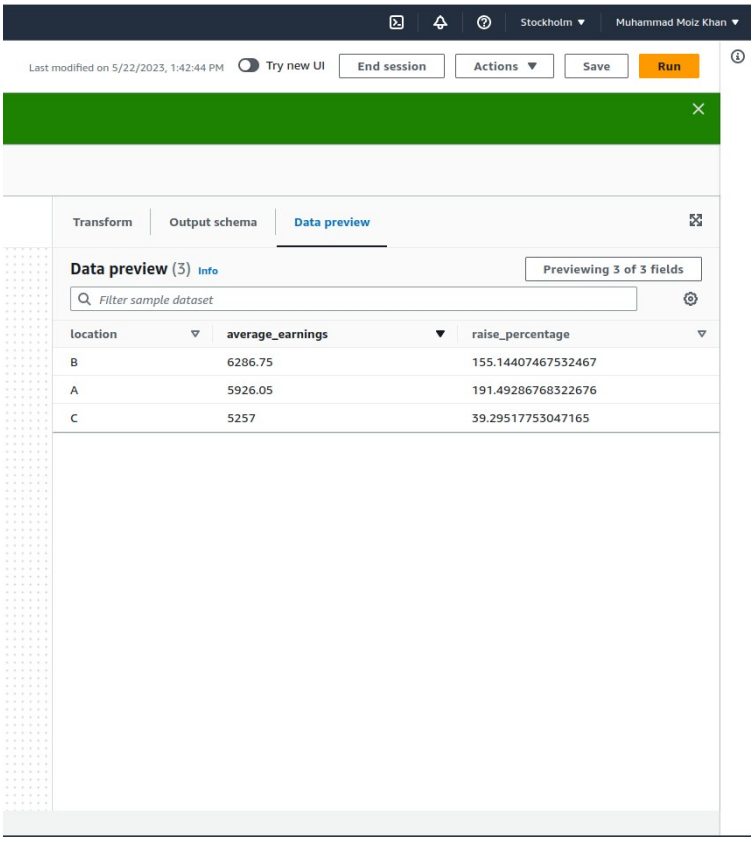
5) Now we create two s3 source one is employee earning data and location,we perform inner join on both data on emp_id and after that prepare for querying :

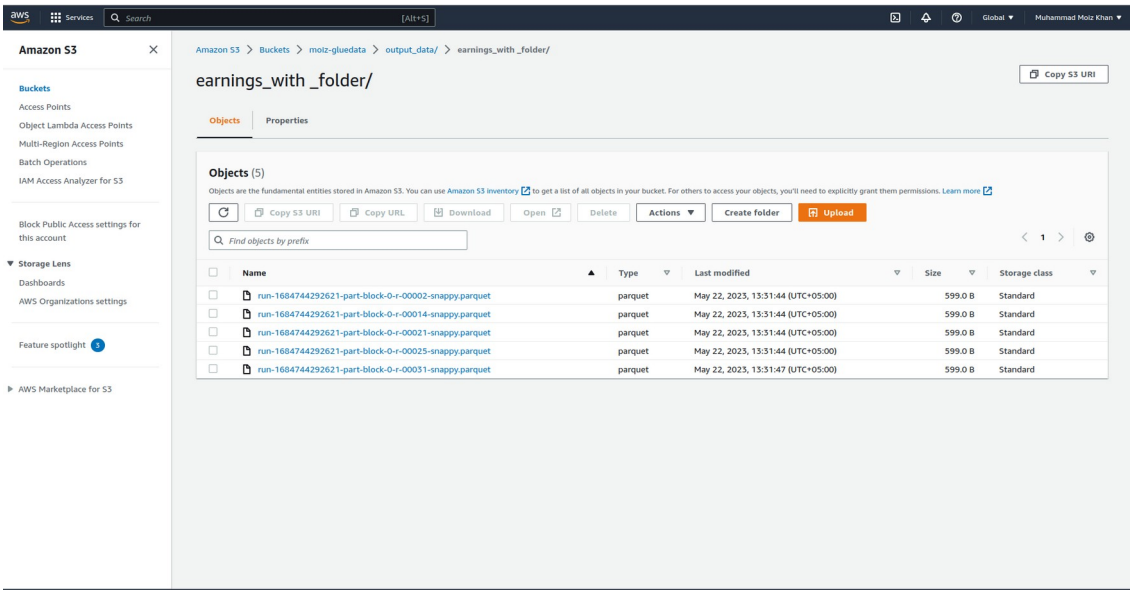


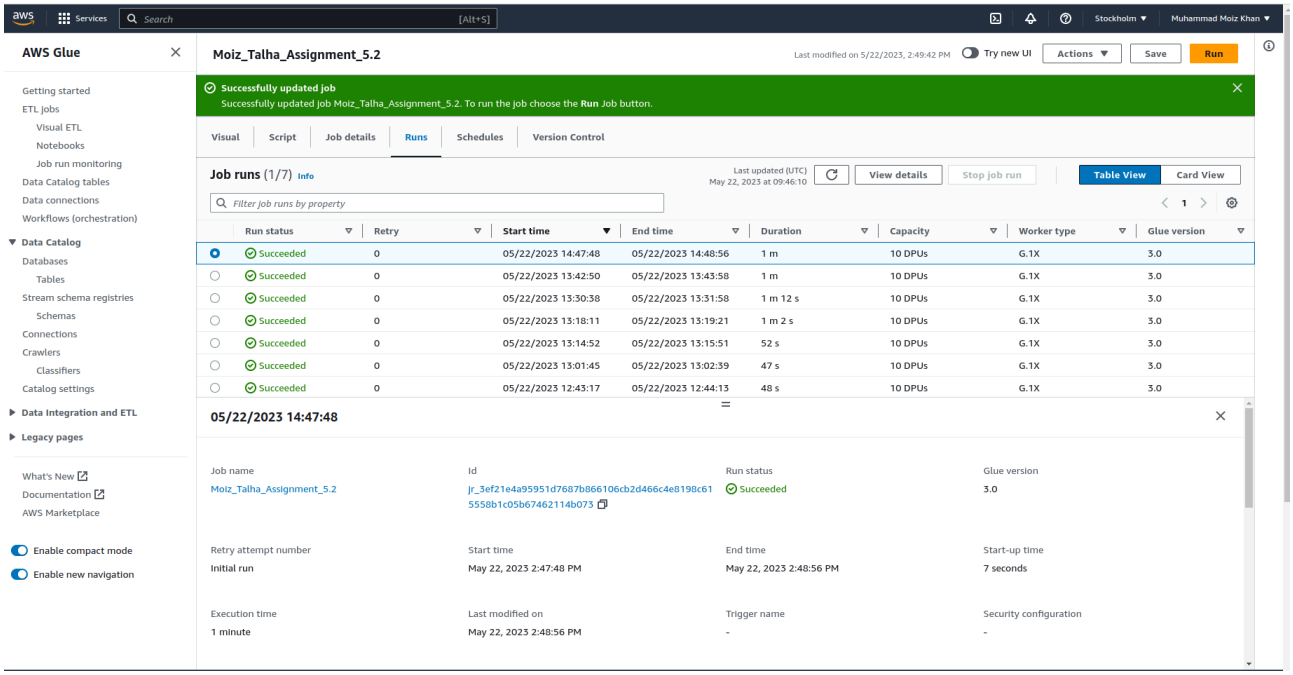

aws
Services
Search
[Alt+S]
Stockholm
Muhammad Moiz Khan
AWS Glue
Moiz_Talha_Assignment_5.2
Last modified on 5/22/2023, 1:30:34 PM
Try new UI
Actions
Save
Run
Getting started
ETL jobs
Visual ETL
Notebooks
Job run monitoring
Data Catalog tables
Data connections
Workflows (orchestration)
Data Catalog
Databases
Tables
Stream schema registries
Schemas
Connections
Crawlers
Classifiers
Catalog settings
Data Integration and ETL
Legacy pages
What's New
Documentation
AWS Marketplace
Enable compact mode
Enable new navigation
Visual
Script
Job details
Runs
Schedules
Version Control
Source
Action
Target
Undo
Redo
Remove
Data source - S3 bucket
Amazon_earnings
Data source - S3 bucket
Amazon_location
Transform - Join
Join
Transform - SQL Query
SQL Query
Data target - S3 bucket
Data_Target
Transform
Output schema
Data preview
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;


aws
Services
Search
[Alt+S]
Stockholm
Muhammad Moiz Khan
AWS Glue
Moiz_Talha_Assignment_5.2
Last modified on 5/22/2023, 1:30:34 PM
Try new UI
Actions
Save
Run
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
Legacy pages
What's New
Documentation
AWS Marketplace
Enable compact mode
Enable new navigation
Visual
Script
Job details
Runs
Schedules
Version Control
Source
Action
Target
Undo
Redo
Remove
Data source - S3 bucket
Amazon_earnings
Data source - S3 bucket
Amazon_location
Transform - Join
Join
Transform - SQL Query
SQL Query
Data target - S3 bucket
Data_Target
Data target properties - S3
Output schema
Data preview
Name
Data_Target
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://moiz-gluedata/output_data/earnings_with _folder/
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.
talhakhan_glue_database

6) Now finally we load the data and show in the output folder :






aws
Services
Search
[Alt+S]
Stockholm
Muhammad Moiz Khan
AWS Glue
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
Legacy pages
What's New
Documentation
AWS Marketplace
Enable compact mode
Enable new navigation
Moiz_Talha_Assignment_5.2
Last modified on 5/22/2023, 2:49:42 PM
Try new UI
Actions
Save
Run
Successfully updated job
Successfully updated job Moiz_Talha_Assignment_5.2. To run the job choose the Run Job button.
Visual
Script
Job details
Runs
Schedules
Version Control
Job runs (1/7) Info
Last updated (UTC)
May 22, 2023 at 09:46:10
View details
Stop job run
Table View
Card View
Filter job runs by property
Run status
Retry
Start time
End time
Duration
Capacity
Worker type
Glue version
Succeeded 0 05/22/2023 14:47:48 05/22/2023 14:48:56 1 m 10 DPUs G.1X 3.0
Succeeded 0 05/22/2023 13:42:50 05/22/2023 13:43:58 1 m 10 DPUs G.1X 3.0
Succeeded 0 05/22/2023 13:30:38 05/22/2023 13:31:58 1 m 12 s 10 DPUs G.1X 3.0
Succeeded 0 05/22/2023 13:18:11 05/22/2023 13:19:21 1 m 2 s 10 DPUs G.1X 3.0
Succeeded 0 05/22/2023 13:14:52 05/22/2023 13:15:51 52 s 10 DPUs G.1X 3.0
Succeeded 0 05/22/2023 13:01:45 05/22/2023 13:02:39 47 s 10 DPUs G.1X 3.0
Succeeded 0 05/22/2023 12:43:17 05/22/2023 12:44:13 48 s 10 DPUs G.1X 3.0
05/22/2023 14:47:48
Job name
Moiz_Talha_Assignment_5.2
Id
jr_3ef21e4a95951d7687b866106cb2d466c4e8198c615558b1c05b67462114b073
Run status
Succeeded
Glue version
3.0
Retry attempt number
Initial run
Start time
May 22, 2023 2:47:48 PM
End time
May 22, 2023 2:48:56 PM
Start-up time
7 seconds
Execution time
1 minute
Last modified on
May 22, 2023 2:48:56 PM
Trigger name
-
Security configuration
-