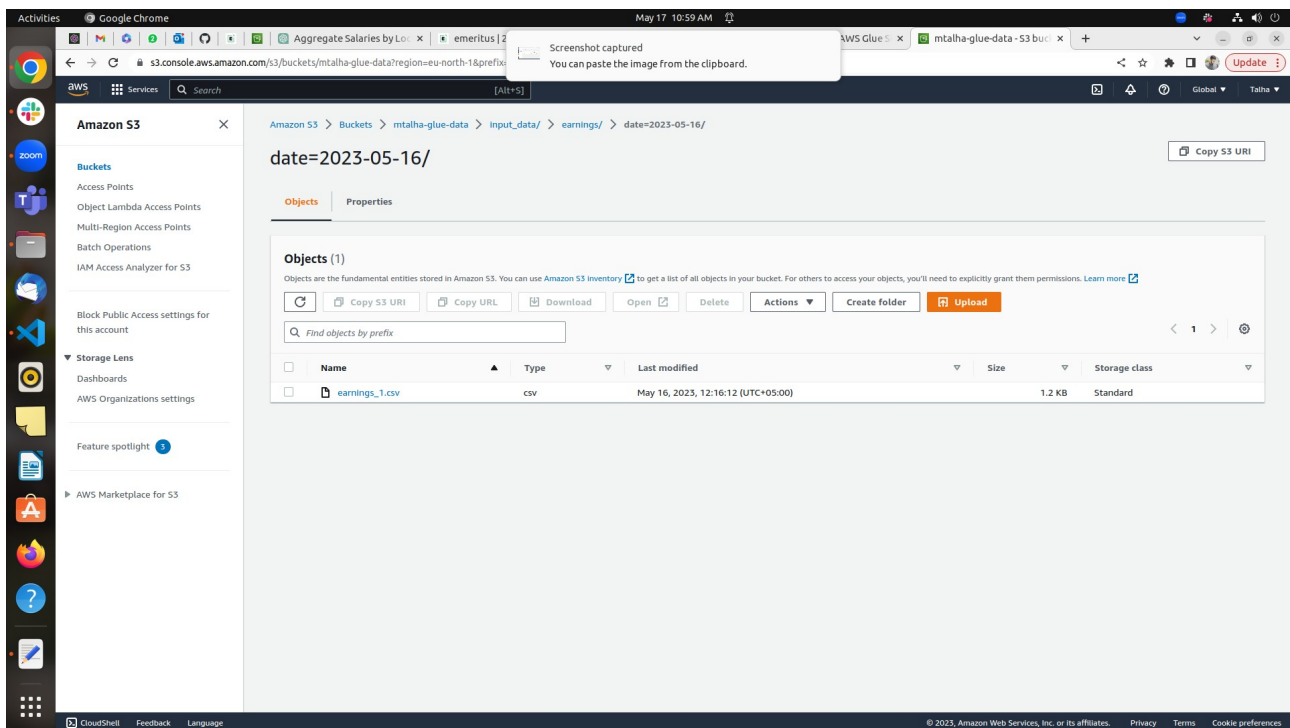
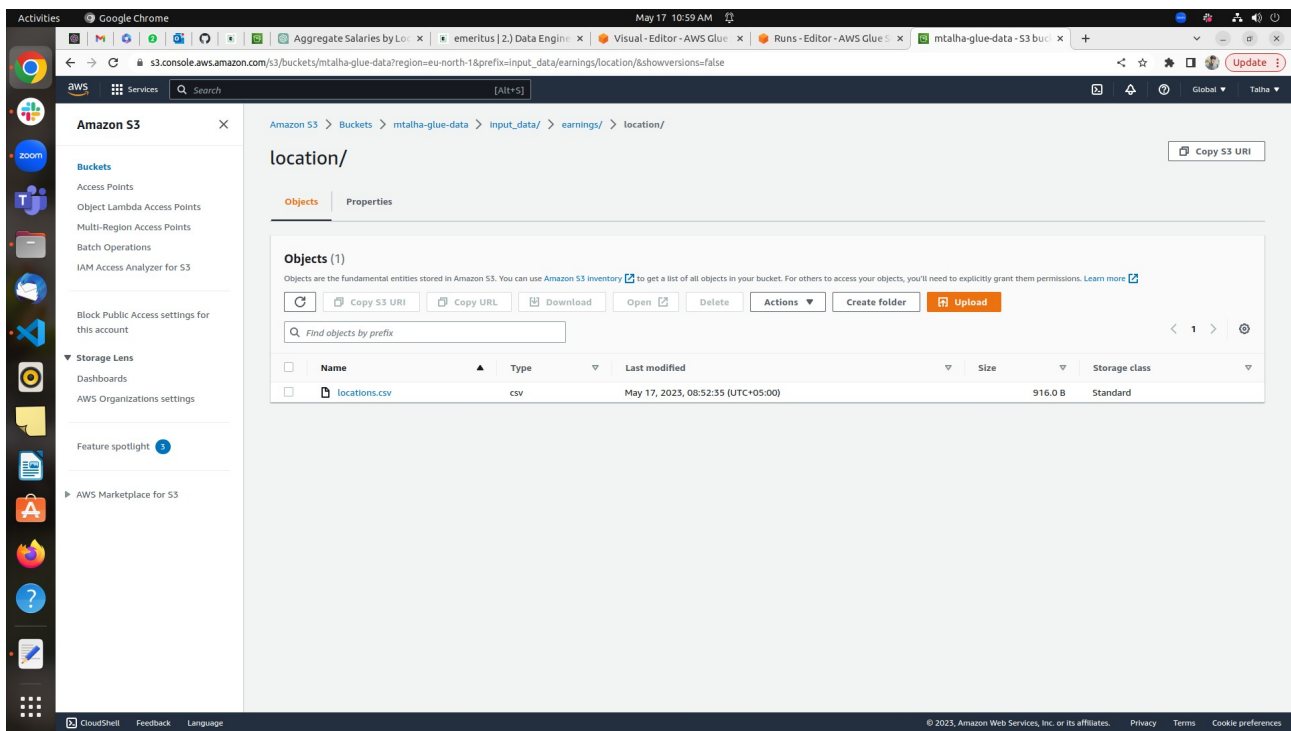
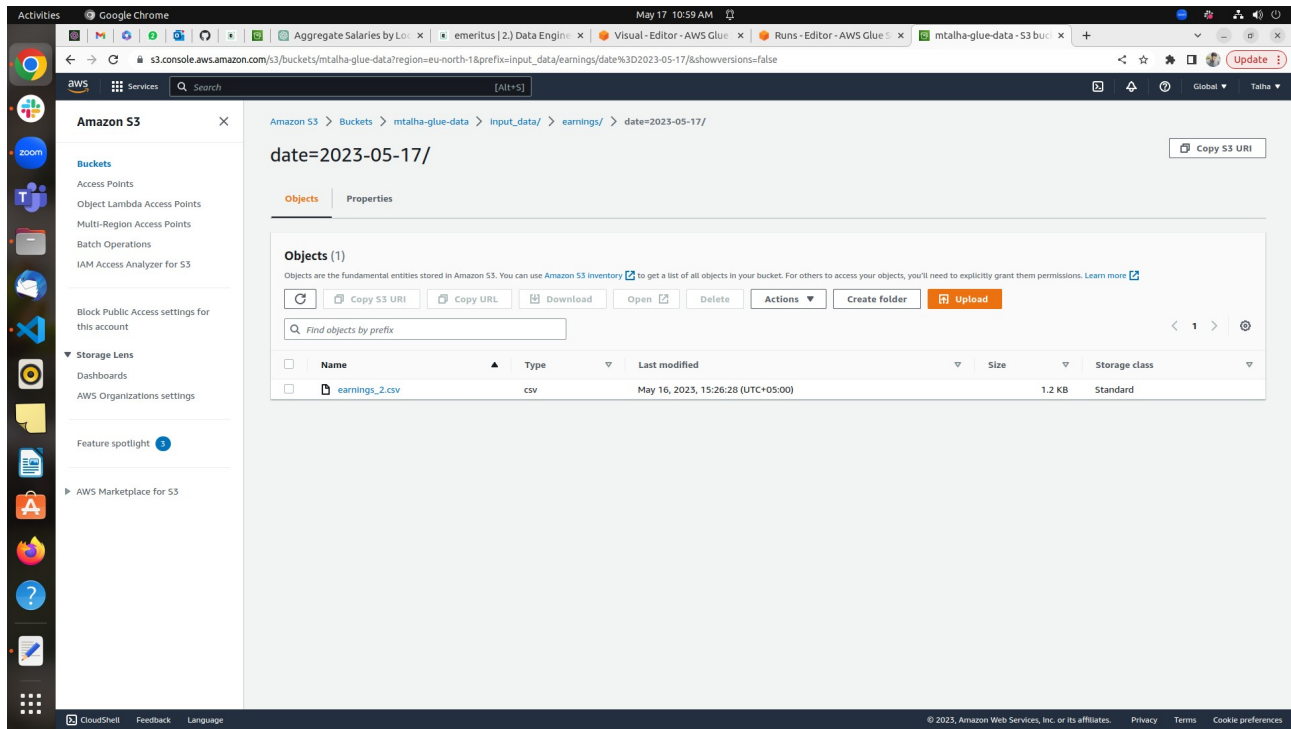


**Using the earnings CSV as a base, prepare a new data file with employees' office locations. Make sure there are 5-6 distinct locations that are shared between employees.**

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

## Uploading data into S3 bucket





Created Crawler

Activities Google Chrome May 17 10:59 AM

eu-north-1.console.aws.amazon.com/glue/home?region=eu-north-1#/v2/data-catalog/crawlers

**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

CloudShell Feedback Language

© 2023, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences

One crawler successfully deleted  
The following crawler is now deleted: "mtalha\_rds\_crawler"

AWS Glue > Crawlers

**Crawlers** info  
View and manage all available crawlers.

Filter crawlers

Table with 8 columns: Name, State, Schedule, Last run, Last run timestamp, Log, Table changes...  
Rows: mतालha\_s3\_earnings\_crawler (Ready, Succeeded, May 16, 2023 at 10:33:09, 1 updated), mतालha\_s3\_location (Ready, Succeeded, May 17, 2023 at 04:05:31, -)

## Created Glue Job

Activities Slack May 17 11:03 AM

eu-north-1.console.aws.amazon.com/gluestudio/home?region=eu-north-1#/editor/job/mtalha\_employee\_earnings\_job/graph

**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

CloudShell Feedback Language

© 2023, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences

mtalha\_employee\_earnings\_job  
Last modified on 5/17/2023, 11:01:27 AM Try new UI End session Actions

Successfully updated job  
Successfully updated job mतालha\_employee\_earnings\_job. To run the job choose the Run Job button.

Visual Script Job details Runs Schedules Version Control

Source Action Target Undo Redo Remove

Diagram showing data flow: Data source - S3 bucket Amazon S3 → Transform - Join → Transform - SQL Query → Data target - S3 bucket Amazon S3

Transform - SQL Query  
SQL Query

Data preview (5) info  
Previewing 2 of 2 fields  
Filter sample dataset

location	avg_earning
B	6286.75
C	5576.95
A	5926.05
D	5889.7
E	5599.2

# Running the job

The screenshot shows the AWS Glue console interface for the job `mtalha_employee_earnings_job`. The interface includes a left-hand navigation menu with options like 'Getting started', 'Data Catalog', and 'Legacy pages'. The main content area displays the 'Runs' tab for the job, showing a table of job runs. The most recent run is 'Failed'.

**Job runs (1/26)**

Run status	Retry	Start time	End time	Duration	Capacity	Worker type	Glue version
Failed	0	05/17/2023 11:01:28	05/17/2023 11:02:40	1 m 5 s	3 DPU's	G.1X	3.0
Failed	0	05/17/2023 10:56:39	05/17/2023 10:57:40	54 s	3 DPU's	G.1X	3.0
Succeeded	0	05/17/2023 10:49:16	05/17/2023 10:50:17	54 s	3 DPU's	G.1X	3.0
Succeeded	0	05/17/2023 10:45:02	05/17/2023 10:46:06	57 s	3 DPU's	G.1X	3.0
Failed	0	05/17/2023 10:43:23	05/17/2023 10:44:18	47 s	3 DPU's	G.1X	3.0

**05/17/2023 10:58:36**

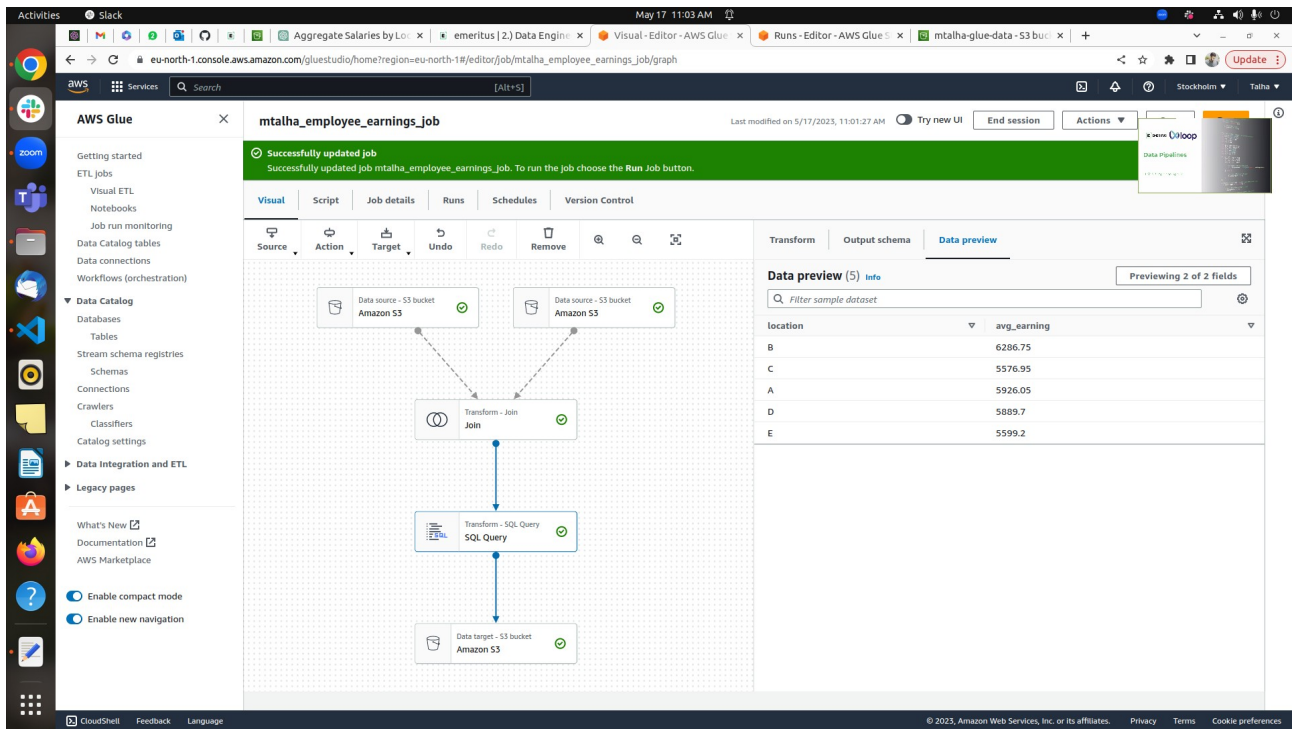
**AnalysisException: Expressions referencing the outer query are not supported outside of WHERE/HAVING clauses:**

Job name	Id	Run status	Glue version
mtalha_employee_earnings_job	j_4ecc3ab6bda4cf06513a0a271e08e04951ffc61f89737e20bd16bd595acca	Failed	3.0

Retry attempt number	Start time	End time	Start-up time
Initial run	May 17, 2023 10:58:36 AM	May 17, 2023 10:59:28 AM	7 seconds

Execution time	Last modified on	Trigger name	Security configuration
45 seconds	May 17, 2023 10:59:28 AM	-	-

# Query output showing the result Avg salary based on the location



The screenshot displays the AWS Glue console interface for the job `mtalha_employee_earnings_job`. The workflow is visualized in the 'Visual' tab, showing two data sources (Amazon S3 buckets) feeding into a 'Transform - Join' node, which then feeds into a 'Transform - SQL Query' node, and finally into a 'Data target - S3 bucket'.

The 'Data preview' tab is active, showing a table with the following data:

location	avg_earning
B	6286.75
C	5576.95
A	5926.05
D	5889.7
E	5599.2

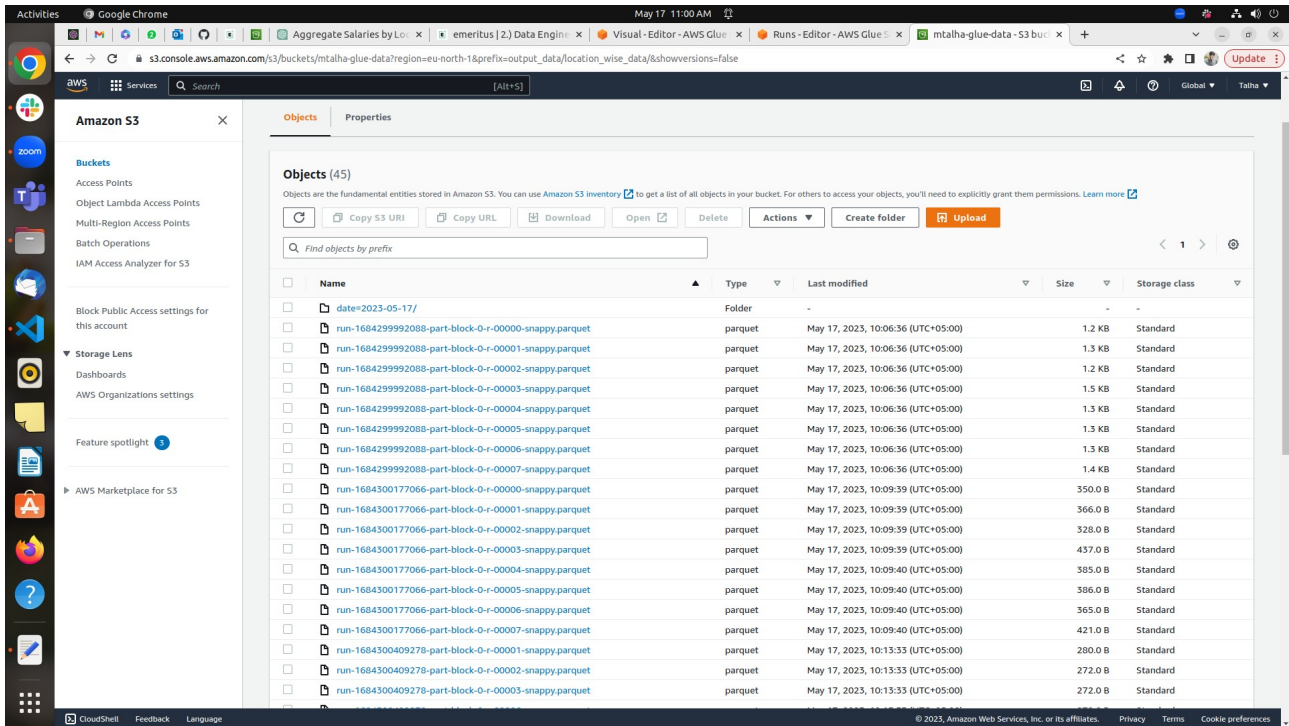
# Query output showing the result raise percentage based on the location

The screenshot displays the AWS Glue console interface. At the top, a green banner indicates the job 'mtalha\_employee\_earnings\_job' was successfully started. The main workspace shows a visual workflow with three data sources (Amazon S3 buckets) feeding into a 'Transform - Join' node, which then feeds into a 'Transform - SQL Query' node, finally outputting to a 'Data target - S3 bucket'. The 'Data preview' tab on the right shows a table with columns 'location', 'earnings\_avg', and 'percentage'.

location	earnings_avg	percentage
B	6086.875	184.30056048575432
C	5695.3	158.9949977262392
A	6217.975	205.85218888342354
D	5635.075	180.91101694915253
E	5503.4	154.31608133086874



# Saving the data



The screenshot displays the Amazon S3 console interface in a web browser. The left sidebar shows the navigation menu with options like Buckets, Access Points, and Storage Lens. The main content area is titled 'Objects (45)' and shows a list of objects in a bucket. The objects are listed in a table with columns for Name, Type, Last modified, Size, and Storage class. The objects are named with a date and a unique identifier, and they are all of type 'parquet'.

Name	Type	Last modified	Size	Storage class
date=2023-05-17/	Folder	-	-	-
run-168429992088-part-block-0-r-00000-snappy.parquet	parquet	May 17, 2023, 10:06:36 (UTC+05:00)	1.2 KB	Standard
run-168429992088-part-block-0-r-00001-snappy.parquet	parquet	May 17, 2023, 10:06:36 (UTC+05:00)	1.3 KB	Standard
run-168429992088-part-block-0-r-00002-snappy.parquet	parquet	May 17, 2023, 10:06:36 (UTC+05:00)	1.2 KB	Standard
run-168429992088-part-block-0-r-00003-snappy.parquet	parquet	May 17, 2023, 10:06:36 (UTC+05:00)	1.5 KB	Standard
run-168429992088-part-block-0-r-00004-snappy.parquet	parquet	May 17, 2023, 10:06:36 (UTC+05:00)	1.3 KB	Standard
run-168429992088-part-block-0-r-00005-snappy.parquet	parquet	May 17, 2023, 10:06:36 (UTC+05:00)	1.3 KB	Standard
run-168429992088-part-block-0-r-00006-snappy.parquet	parquet	May 17, 2023, 10:06:36 (UTC+05:00)	1.3 KB	Standard
run-168429992088-part-block-0-r-00007-snappy.parquet	parquet	May 17, 2023, 10:06:36 (UTC+05:00)	1.4 KB	Standard
run-1684300177066-part-block-0-r-00000-snappy.parquet	parquet	May 17, 2023, 10:09:39 (UTC+05:00)	350.0 B	Standard
run-1684300177066-part-block-0-r-00001-snappy.parquet	parquet	May 17, 2023, 10:09:39 (UTC+05:00)	366.0 B	Standard
run-1684300177066-part-block-0-r-00002-snappy.parquet	parquet	May 17, 2023, 10:09:39 (UTC+05:00)	328.0 B	Standard
run-1684300177066-part-block-0-r-00003-snappy.parquet	parquet	May 17, 2023, 10:09:39 (UTC+05:00)	437.0 B	Standard
run-1684300177066-part-block-0-r-00004-snappy.parquet	parquet	May 17, 2023, 10:09:40 (UTC+05:00)	385.0 B	Standard
run-1684300177066-part-block-0-r-00005-snappy.parquet	parquet	May 17, 2023, 10:09:40 (UTC+05:00)	386.0 B	Standard
run-1684300177066-part-block-0-r-00006-snappy.parquet	parquet	May 17, 2023, 10:09:40 (UTC+05:00)	365.0 B	Standard
run-1684300177066-part-block-0-r-00007-snappy.parquet	parquet	May 17, 2023, 10:09:40 (UTC+05:00)	421.0 B	Standard
run-1684300409278-part-block-0-r-00001-snappy.parquet	parquet	May 17, 2023, 10:13:33 (UTC+05:00)	280.0 B	Standard
run-1684300409278-part-block-0-r-00002-snappy.parquet	parquet	May 17, 2023, 10:13:33 (UTC+05:00)	272.0 B	Standard
run-1684300409278-part-block-0-r-00003-snappy.parquet	parquet	May 17, 2023, 10:13:33 (UTC+05:00)	272.0 B	Standard