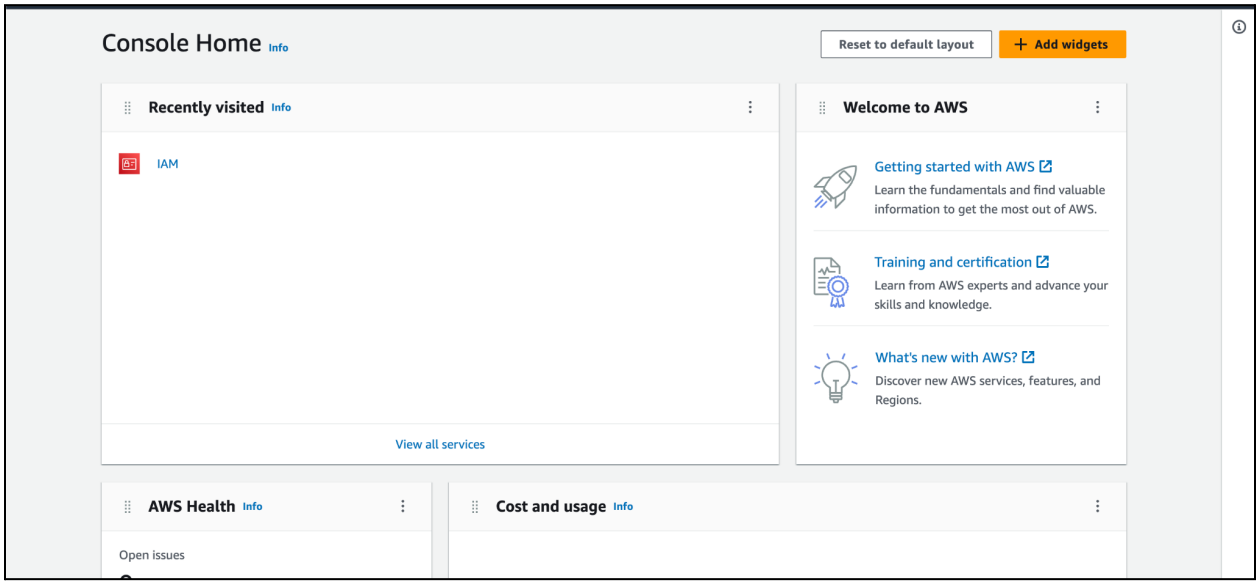
ETL AWS Project

Designed and deployed data pipelines to preprocess data, integrating data from various regions into data lakes, converting JSON and CSV data files into parquet formats and joining the data to make it available for the stakeholders (Data Analysts and Data Scientist) to be able to build reports on the cleaned data for business analysis.

Following are the steps:

1. Create IAM User

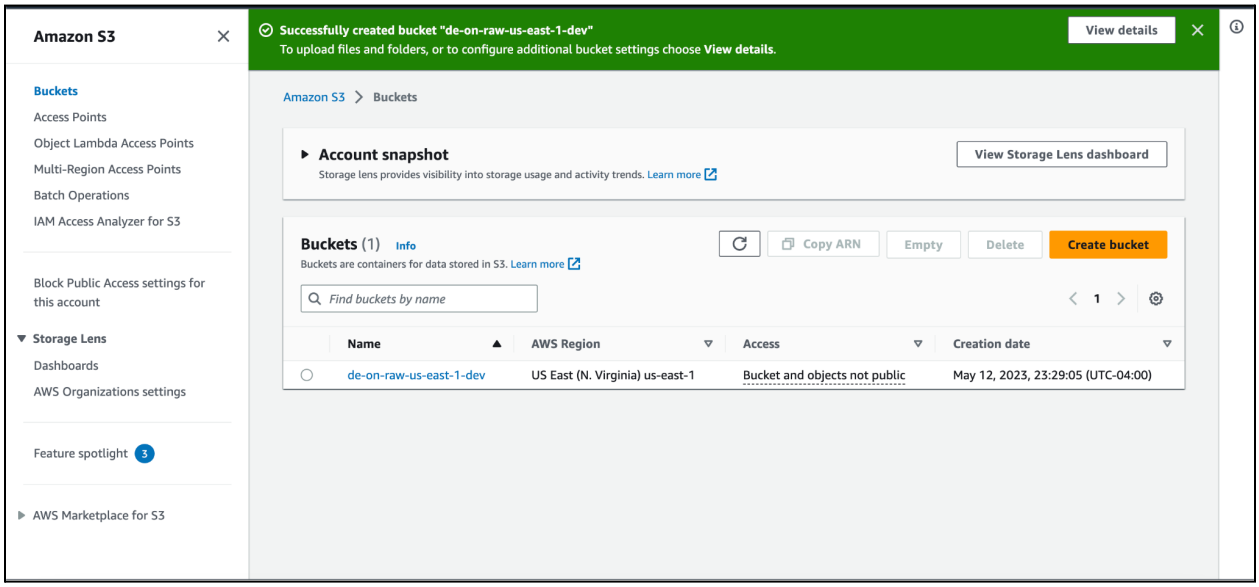


2. Configure CLI

- Downloaded AWS CLI tool.
- Open terminal check tool is installed by writing command:
  aws on terminal
- Write command
  aws configure
- Provide access key, secret access key, region and default output
- Check if it is configured or not by writing aws command. For eg:
  aws s3 ls

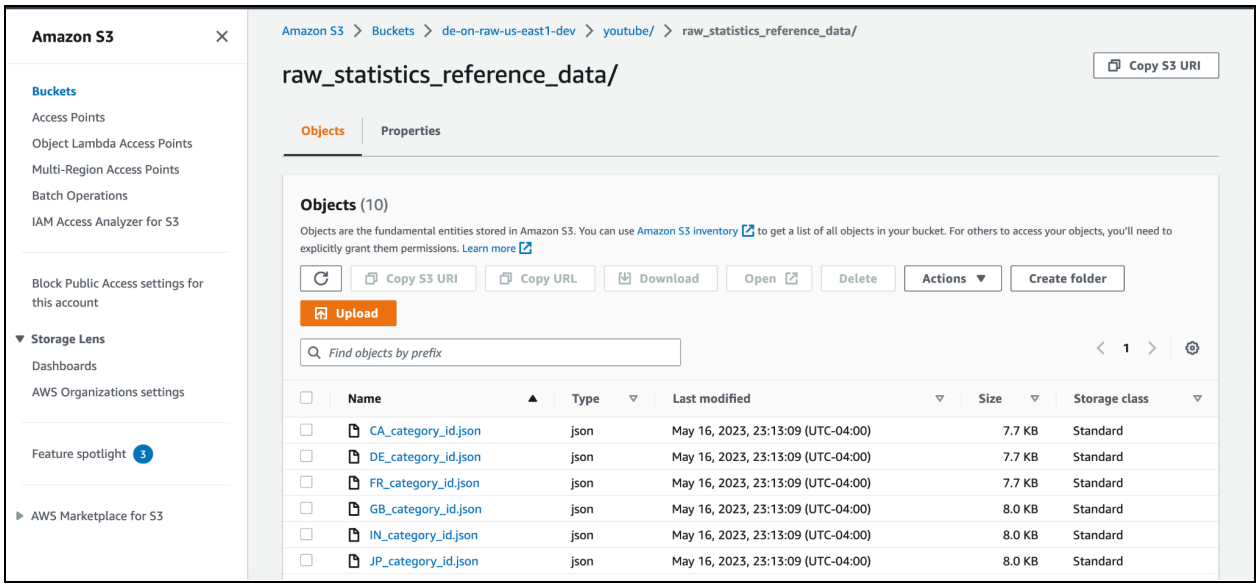
3. Upload Data files in S3 storage

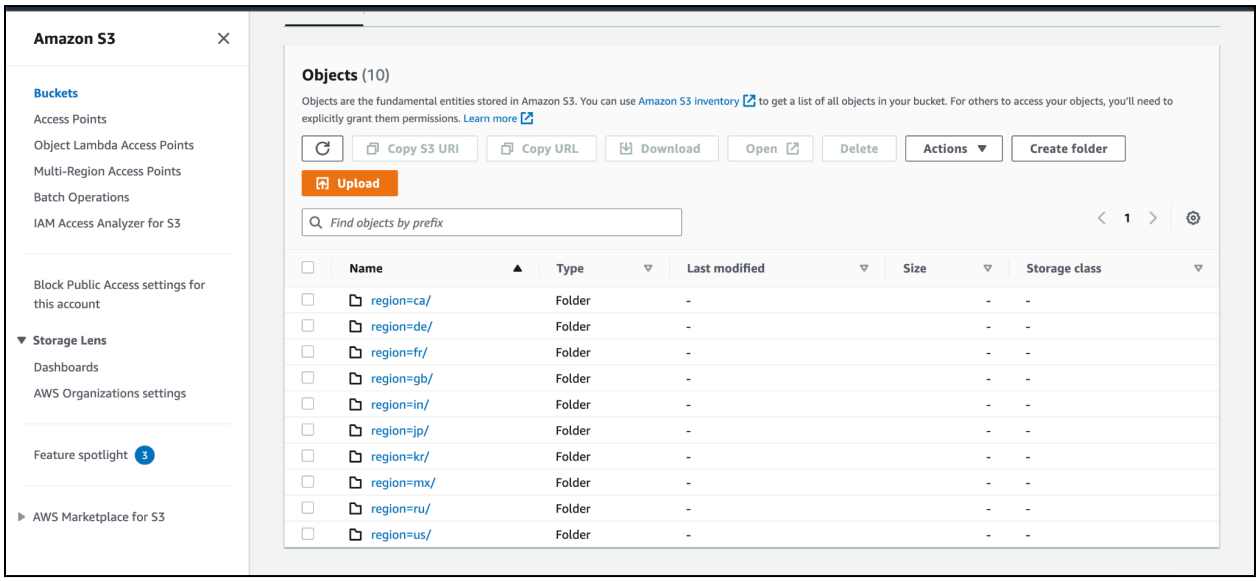
- Create Bucket



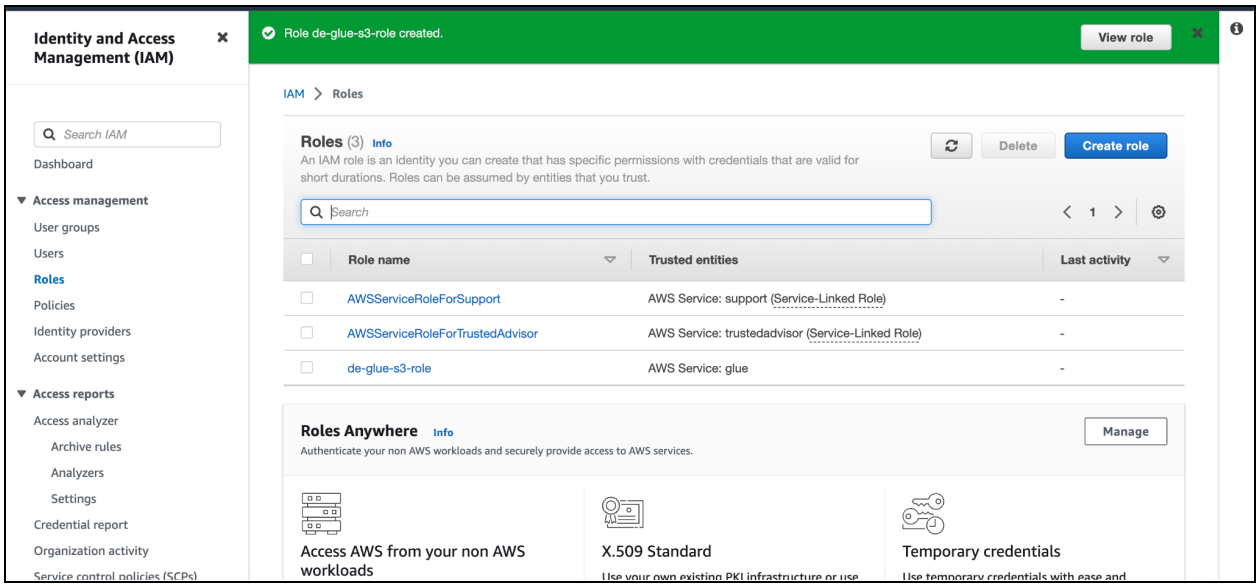
- Uploading data into bucket

```
 ETL_AWS % cd Data
 Data % cd archive
 archive % ls
Bvideos.csv             MX_category_id.json
N_category_id.json      MXvideos.csv
Nvideos.csv             RU_category_id.json
P_category_id.json      RUvideos.csv
Pvideos.csv             US_category_id.json
R_category_id.json      USvideos.csv
Rvideos.csv
 archive % aws s3 cp . s3://de-on-raw-us-east1-dev/youtube/raw_statistics_reference_data/ --recursive --exclude "*" --include "*.json"

o: 0

o: 0

####################################################

####################################################
arwin.so       0x000000010e067663 s_print_stack_trace + 19
lib            0x00007ff818d29c1d _sigtramp + 29
               0x00007ff818934000 0x0 + 140703540920320
               0x0000600001342ec8 0x0 + 105553136463560
 archive %
 archive % aws s3 cp . s3://de-on-raw-us-east1-dev/youtube/raw_statistics_reference_data/ --recursive --exclude "*" --include "*.json"

json to s3://de-on-raw-us-east1-dev/youtube/raw_statistics_reference_data/IN_category_id.json
json to s3://de-on-raw-us-east1-dev/youtube/raw_statistics_reference_data/CA_category_id.json
json to s3://de-on-raw-us-east1-dev/youtube/raw_statistics_reference_data/GB_category_id.json
json to s3://de-on-raw-us-east1-dev/youtube/raw_statistics_reference_data/DE_category_id.json
json to s3://de-on-raw-us-east1-dev/youtube/raw_statistics_reference_data/JP_category_id.json
json to s3://de-on-raw-us-east1-dev/youtube/raw_statistics_reference_data/US_category_id.json
json to s3://de-on-raw-us-east1-dev/youtube/raw_statistics_reference_data/FR_category_id.json
json to s3://de-on-raw-us-east1-dev/youtube/raw_statistics_reference_data/KR_category_id.json
json to s3://de-on-raw-us-east1-dev/youtube/raw_statistics_reference_data/MX_category_id.json
json to s3://de-on-raw-us-east1-dev/youtube/raw_statistics_reference_data/RU_category_id.json
 archive %
```

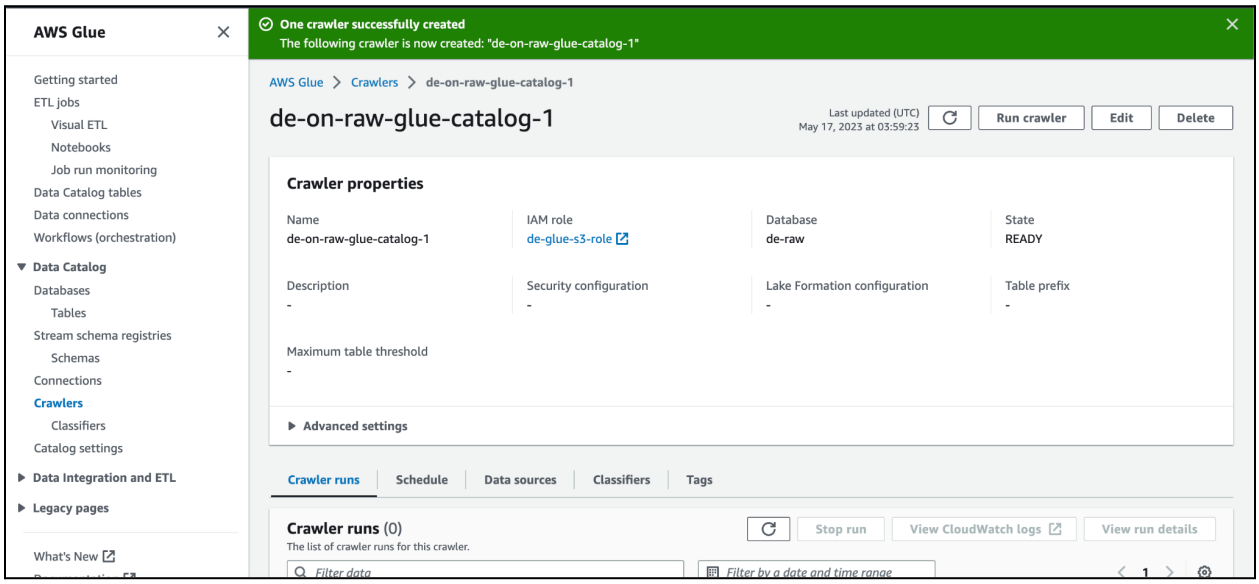


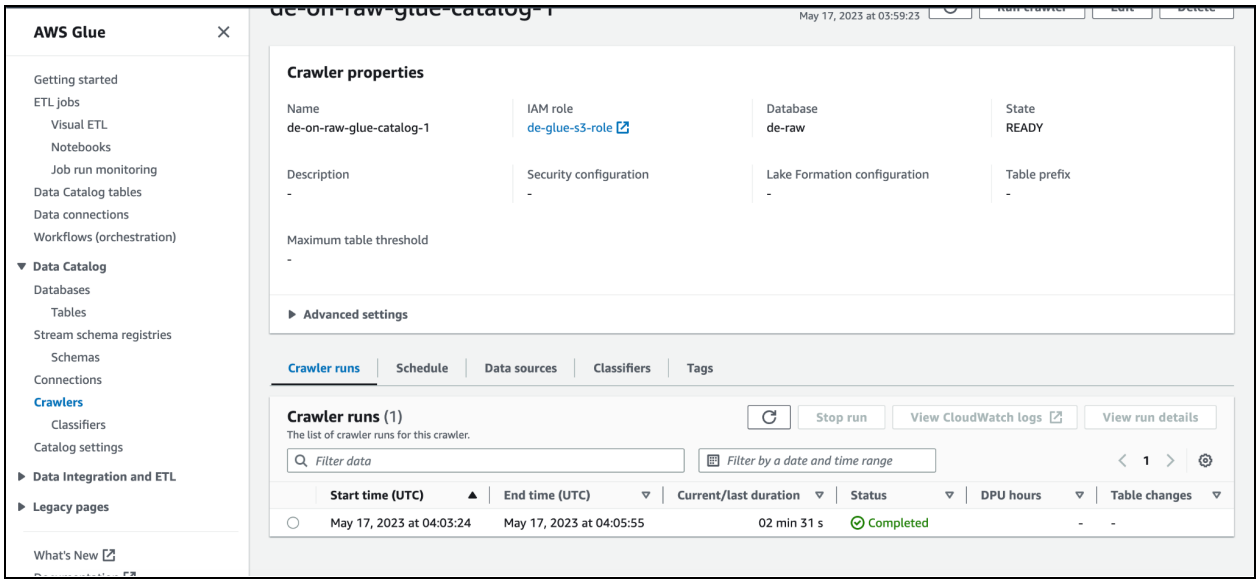


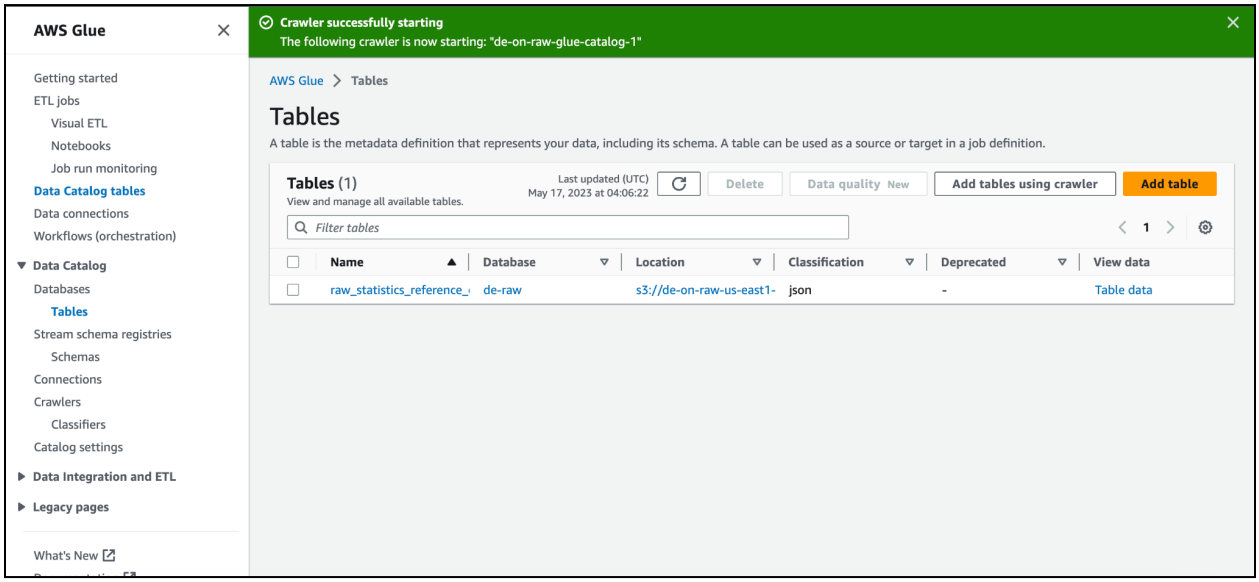
4. Create IAM roles for permissions

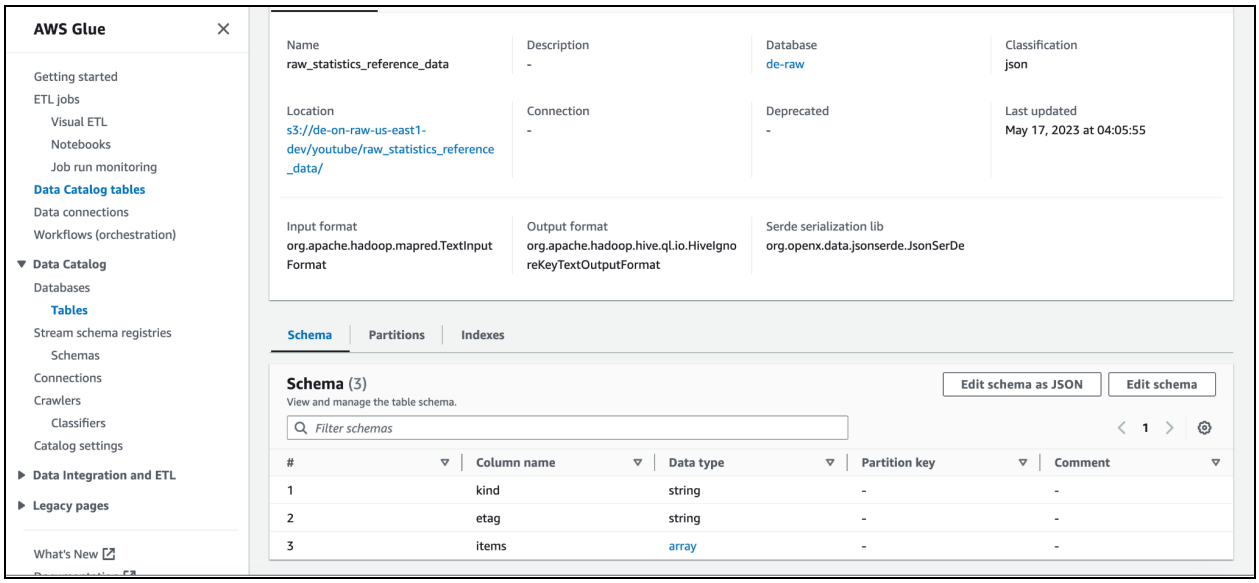


5. Creating AWS Glue Crawlers and Run Crawler Jobs



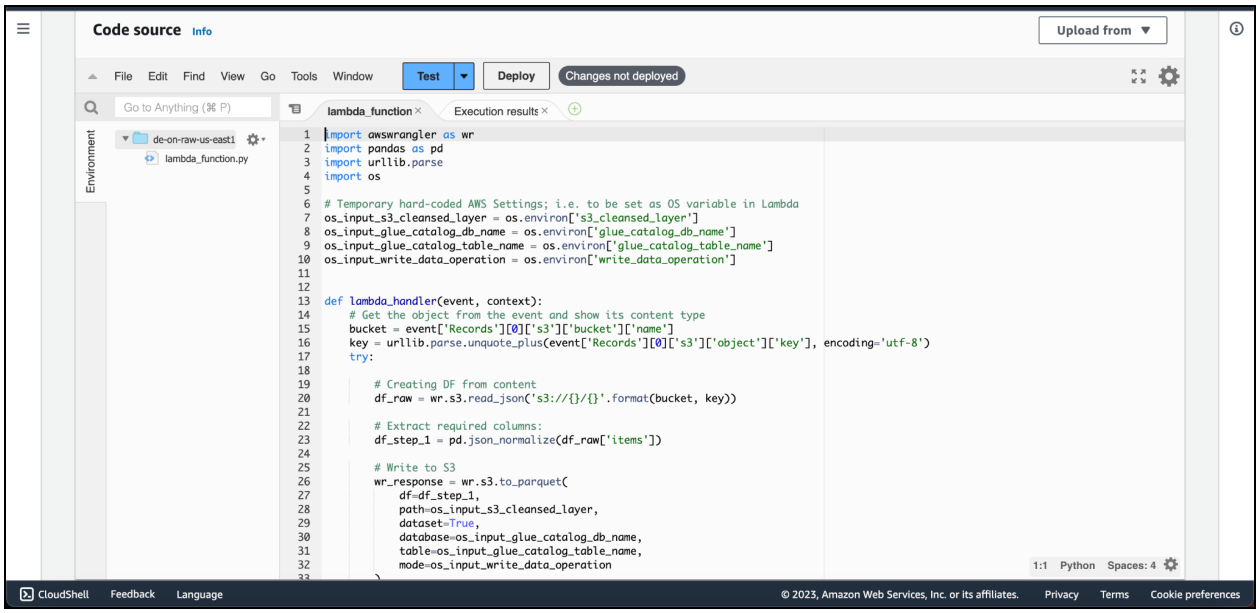

AWS Glue
Getting started
ETL jobs
Visual ETL
Notebooks
Job run monitoring
Data Catalog tables
Data connections
Workflows (orchestration)
Data Catalog
Databases
Tables
Stream schema registries
Schemas
Connections
Crawlers
Classifiers
Catalog settings
Data Integration and ETL
Legacy pages
What's New
de-on-raw-glue-catalog-1
May 17, 2023 at 03:59:23
Crawler properties
Name
de-on-raw-glue-catalog-1
IAM role
de-glue-s3-role
Database
de-raw
State
READY
Description
-
Security configuration
-
Lake Formation configuration
-
Table prefix
-
Maximum table threshold
-
Advanced settings
Crawler runs
Schedule
Data sources
Classifiers
Tags
Crawler runs (1)
The list of crawler runs for this crawler.
Stop run
View CloudWatch logs
View run details
Filter data
Filter by a date and time range
1
Start time (UTC)
End time (UTC)
Current/last duration
Status
DPU hours
Table changes
May 17, 2023 at 04:03:24
May 17, 2023 at 04:05:55
02 min 31 s
Completed
-
-

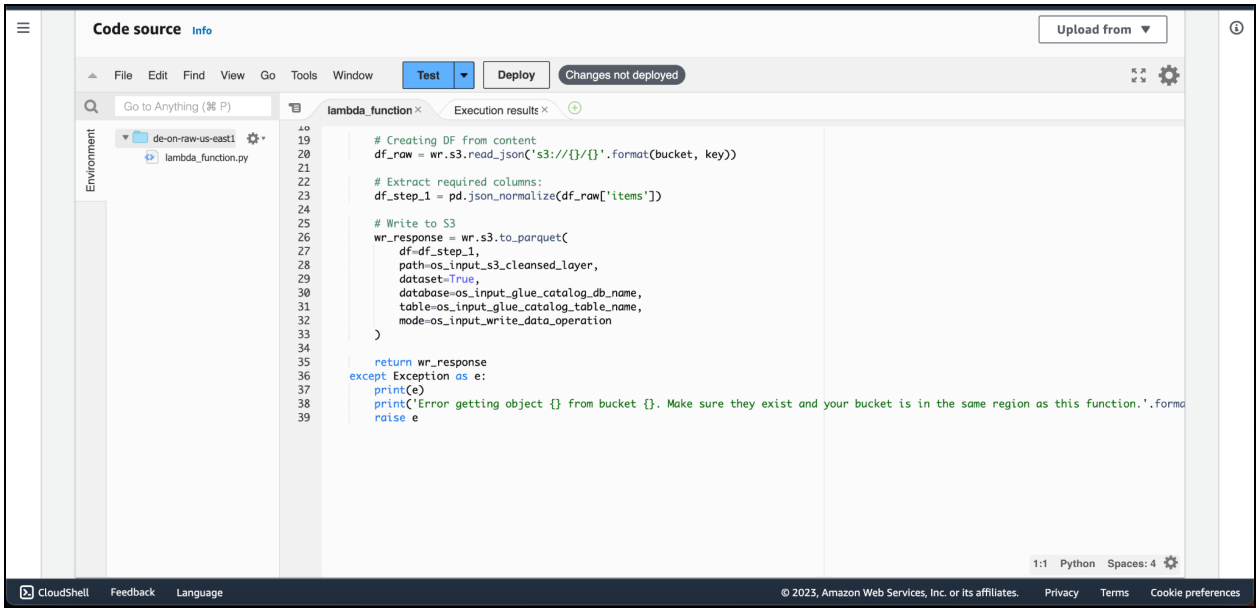

AWS Glue
Crawler successfully starting
The following crawler is now starting: "de-on-raw-glue-catalog-1"
Getting started
ETL jobs
Visual ETL
Notebooks
Job run monitoring
Data Catalog tables
Data connections
Workflows (orchestration)
Data Catalog
Databases
Tables
Stream schema registries
Schemas
Connections
Crawlers
Classifiers
Catalog settings
Data Integration and ETL
Legacy pages
What's New
AWS Glue
Tables
Tables
A table is the metadata definition that represents your data, including its schema. A table can be used as a source or target in a job definition.
Tables (1)
View and manage all available tables.
Last updated (UTC)
May 17, 2023 at 04:06:22
Delete
Data quality New
Add tables using crawler
Add table
Filter tables
1
Name
Database
Location
Classification
Deprecated
View data
raw_statistics_reference_
de-raw
s3://de-on-raw-us-east1-
json
-
Table data

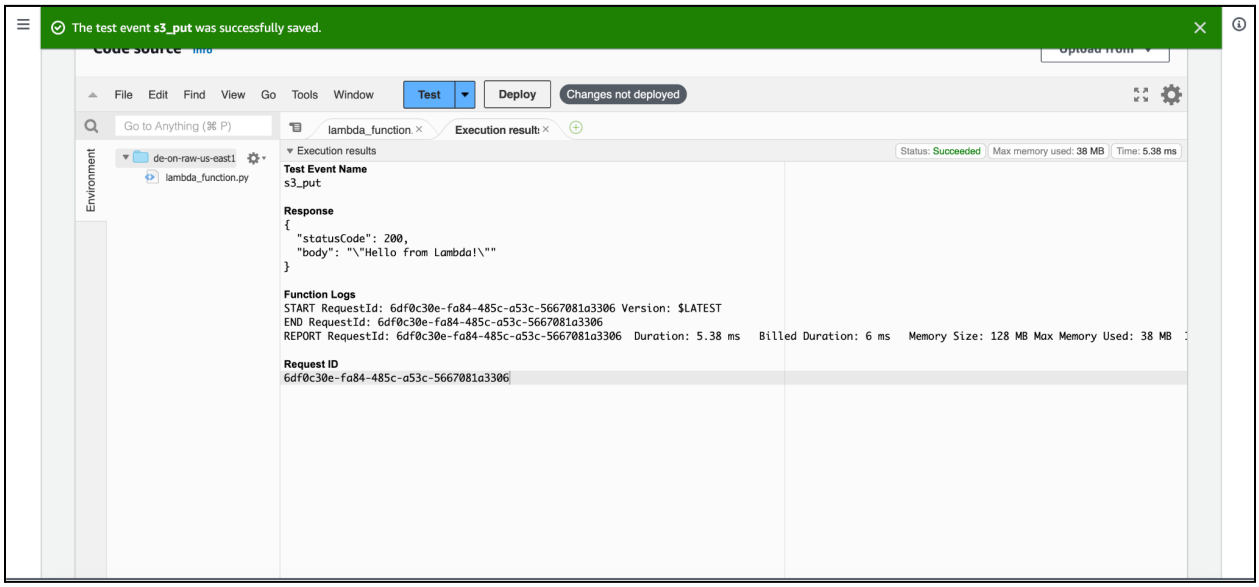


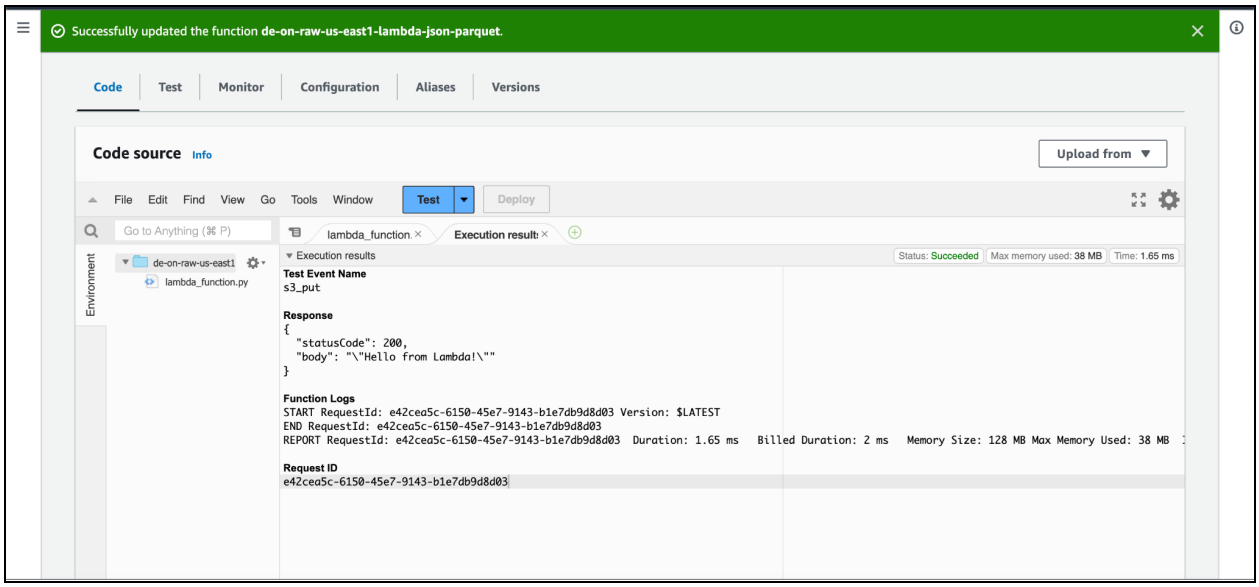
6. Create Lambda functions

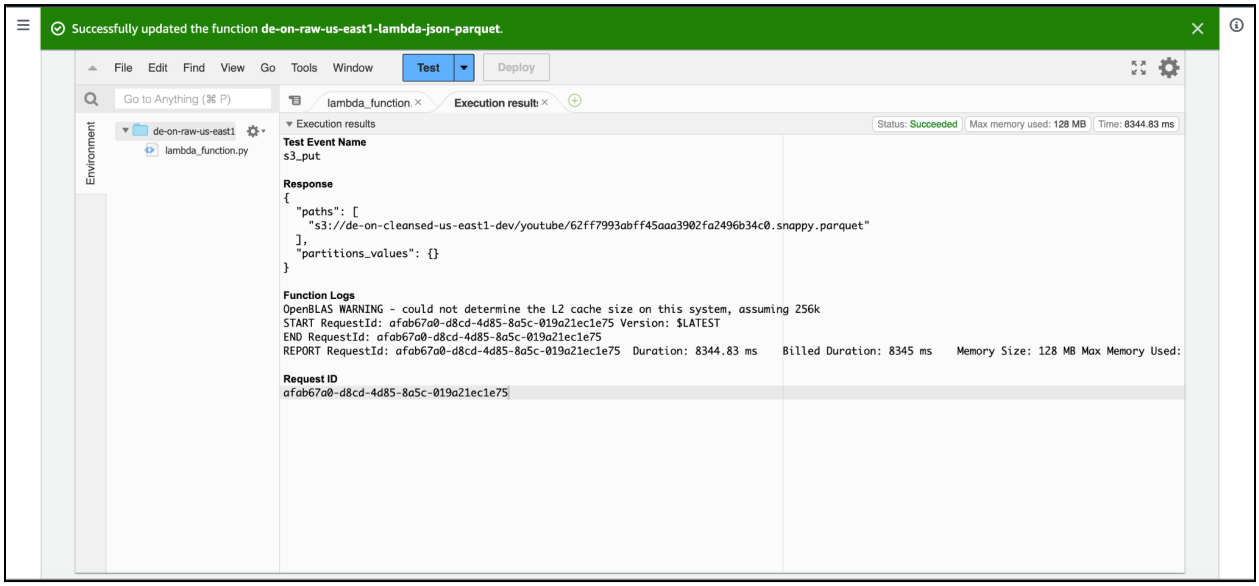
- To upload JSON data in tabular format



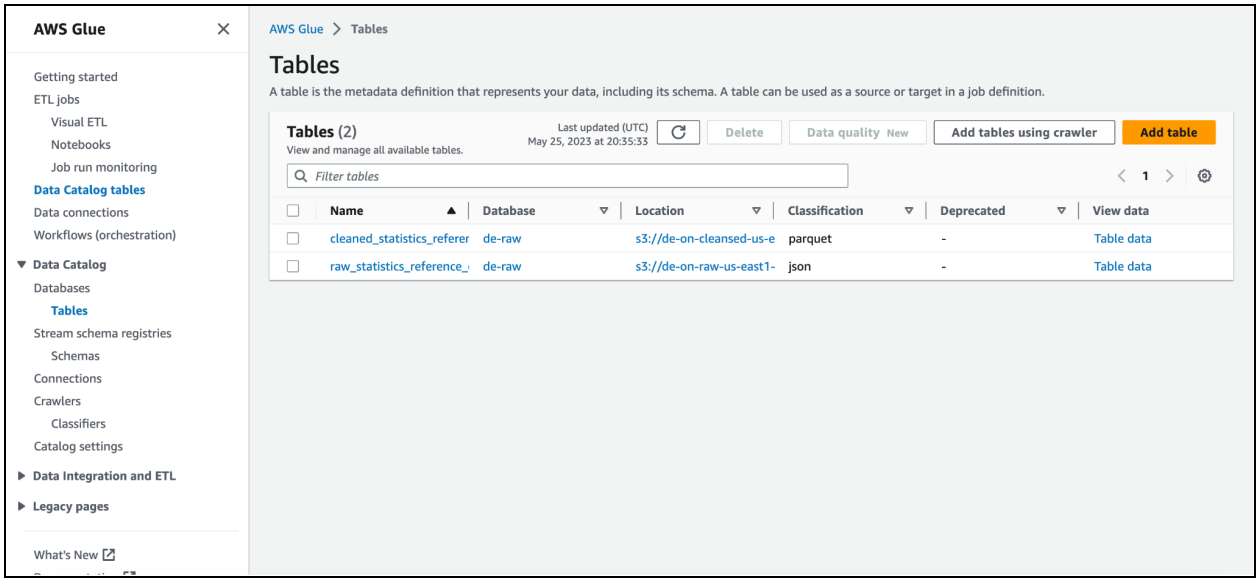

Code source Info
Upload from
File Edit Find View Go Tools Window
Test
Deploy
Changes not deployed
Go to Anything (⌘ P)
lambda_function
Execution result
Environment
de-on-raw-us-east1
lambda_function.py
# Creating DF from content
df_raw = wr.s3.read_json('s3://{}/{}'.format(bucket, key))
# Extract required columns:
df_step_1 = pd.json_normalize(df_raw['items'])
# Write to S3
wr_response = wr.s3.to_parquet(
df=df_step_1,
path=os_input_s3_cleansed_layer,
dataset=True,
database=os_input_glue_catalog_db_name,
table=os_input_glue_catalog_table_name,
mode=os_input_write_data_operation
)
return wr_response
except Exception as e:
print(e)
print('Error getting object {} from bucket {}. Make sure they exist and your bucket is in the same region as this function.'.forma
raise e
1:1 Python Spaces: 4
CloudShell
Feedback
Language
© 2023, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences

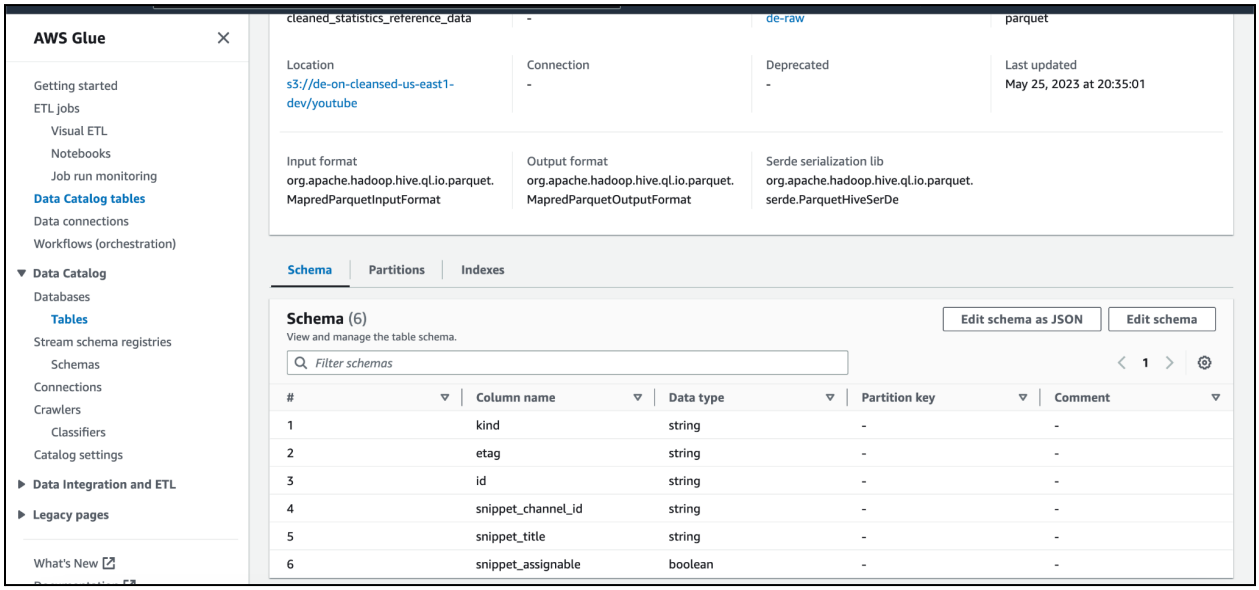

The test event s3_put was successfully saved.
File Edit Find View Go Tools Window
Test
Deploy
Changes not deployed
Go to Anything (⌘ P)
lambda_function
Execution result
Environment
de-on-raw-us-east1
lambda_function.py
Execution results
Status: Succeeded
Max memory used: 38 MB
Time: 5.38 ms
Test Event Name
s3_put
Response
{
"statusCode": 200,
"body": "\"Hello from Lambda!\""
}
Function Logs
START RequestId: 6df0c30e-fa84-485c-a53c-5667081a3306 Version: $LATEST
END RequestId: 6df0c30e-fa84-485c-a53c-5667081a3306
REPORT RequestId: 6df0c30e-fa84-485c-a53c-5667081a3306 Duration: 5.38 ms Billed Duration: 6 ms Memory Size: 128 MB Max Memory Used: 38 MB
Request ID
6df0c30e-fa84-485c-a53c-5667081a3306



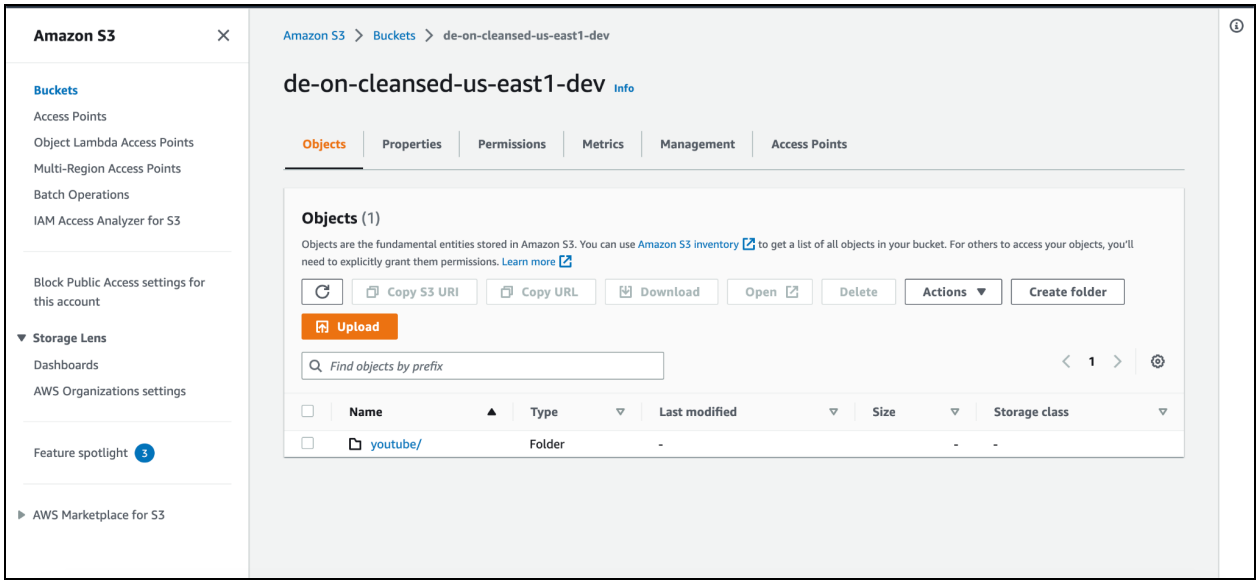


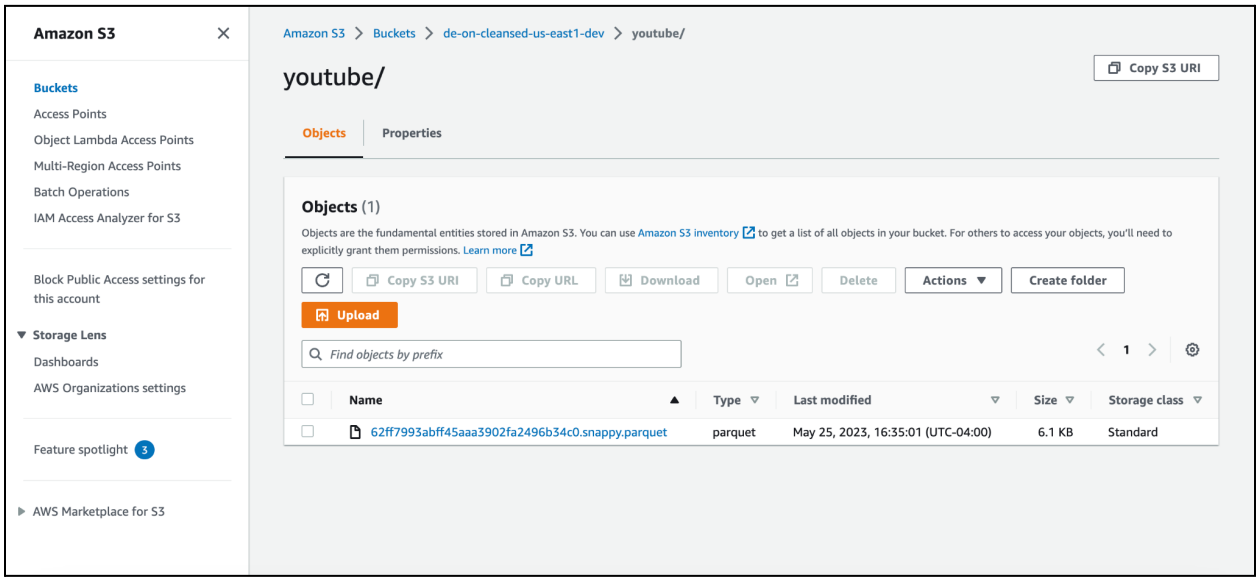
- Successfully created table in Glue and dataset in tabular format



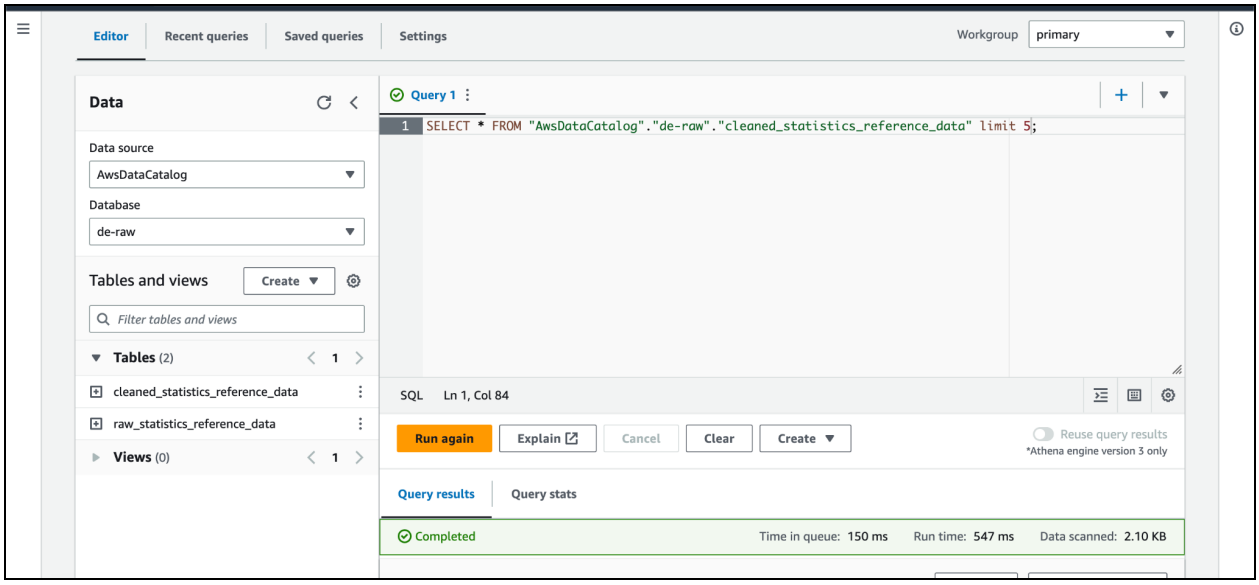


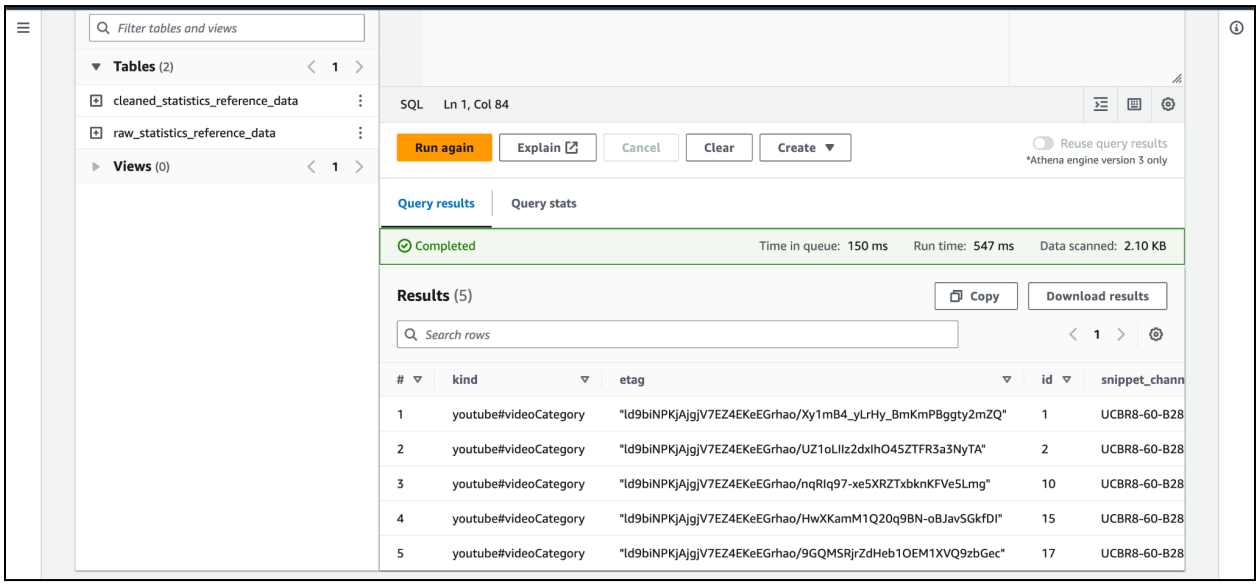
- Loaded Data in S3 in parquet format





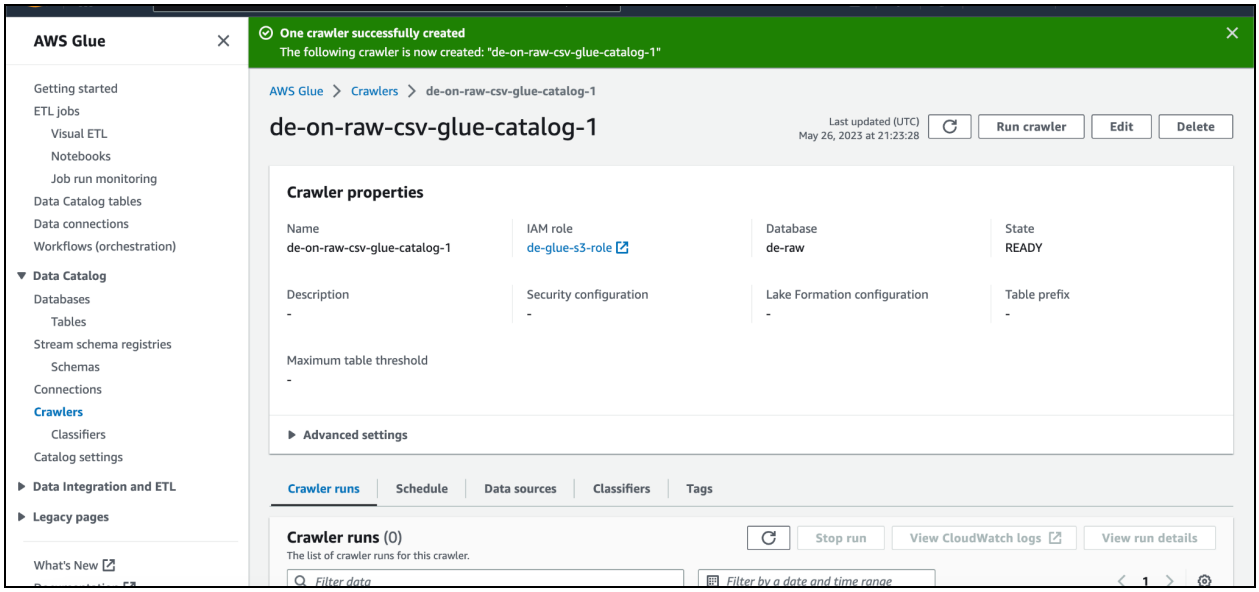
- Athena Query Output



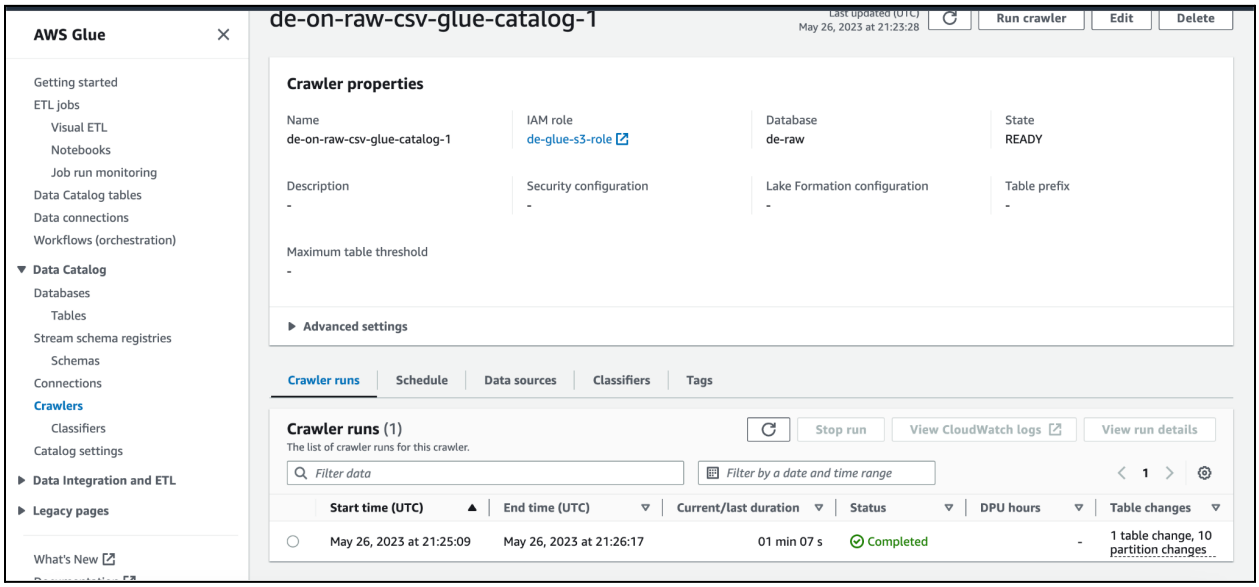


7. Load CSV datasets using Glue Job

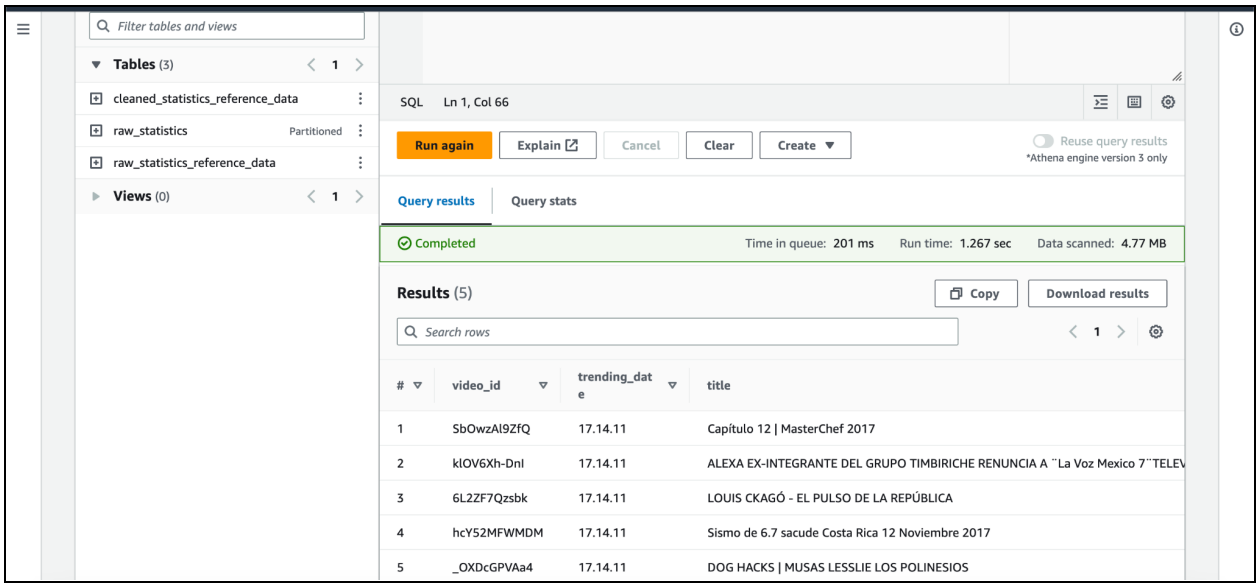
- Create Crawler



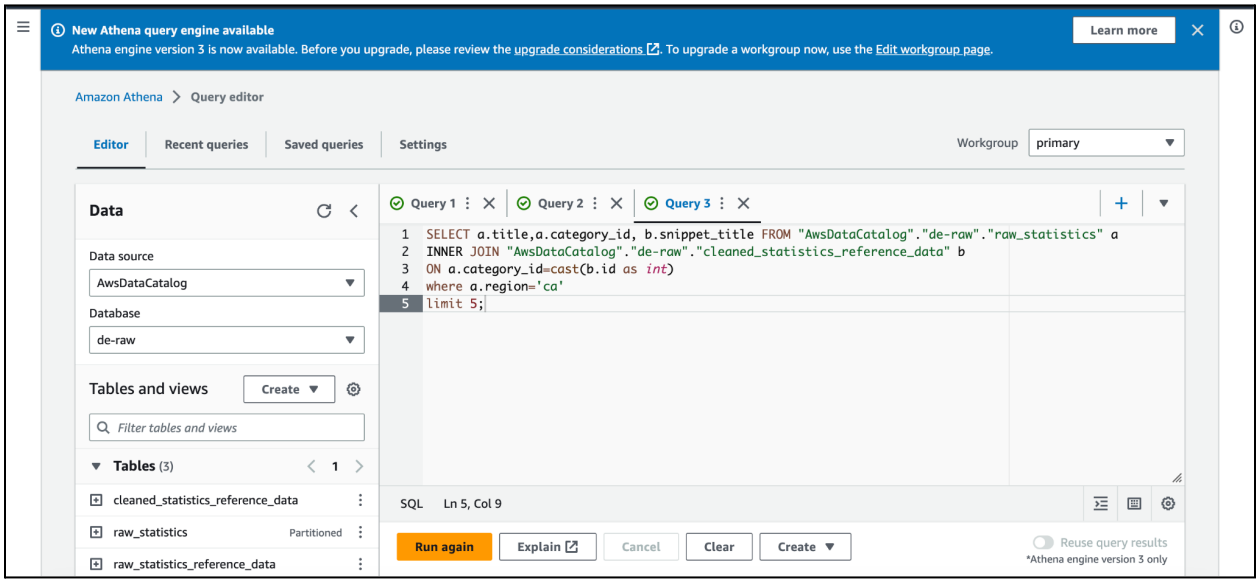
- Run Crawler

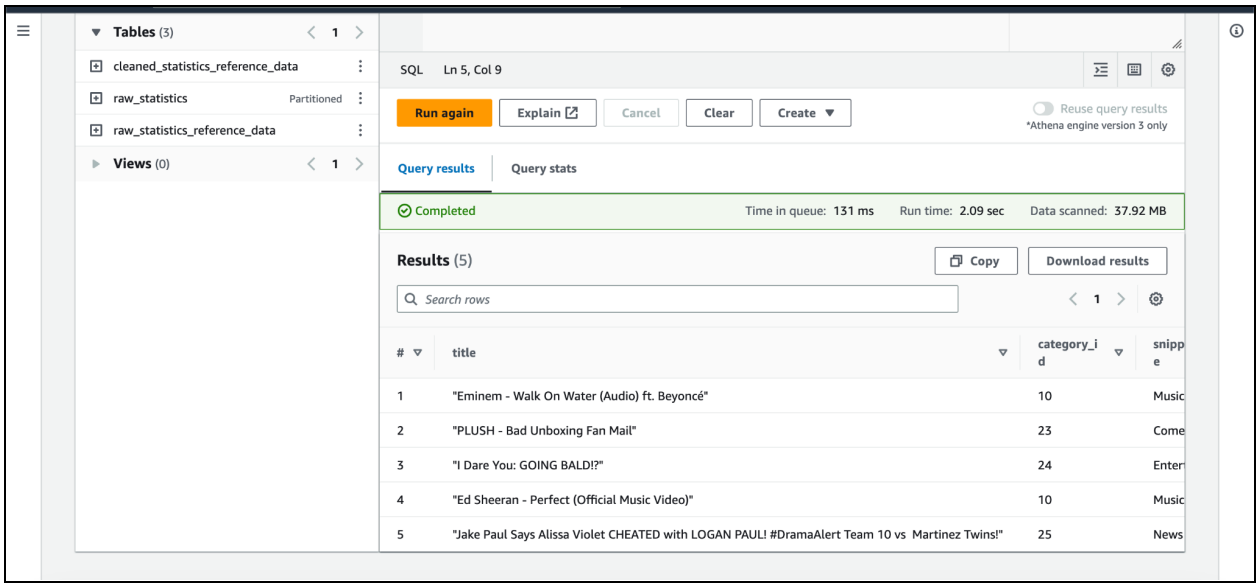


- Athena Output



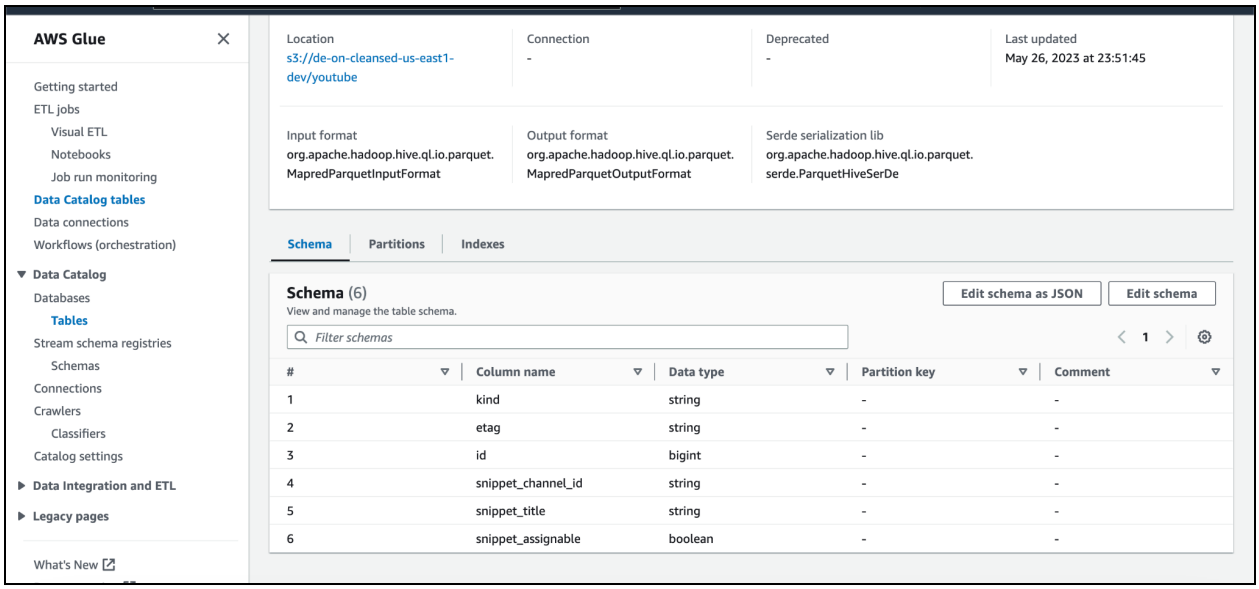
- Joining the tables and getting output



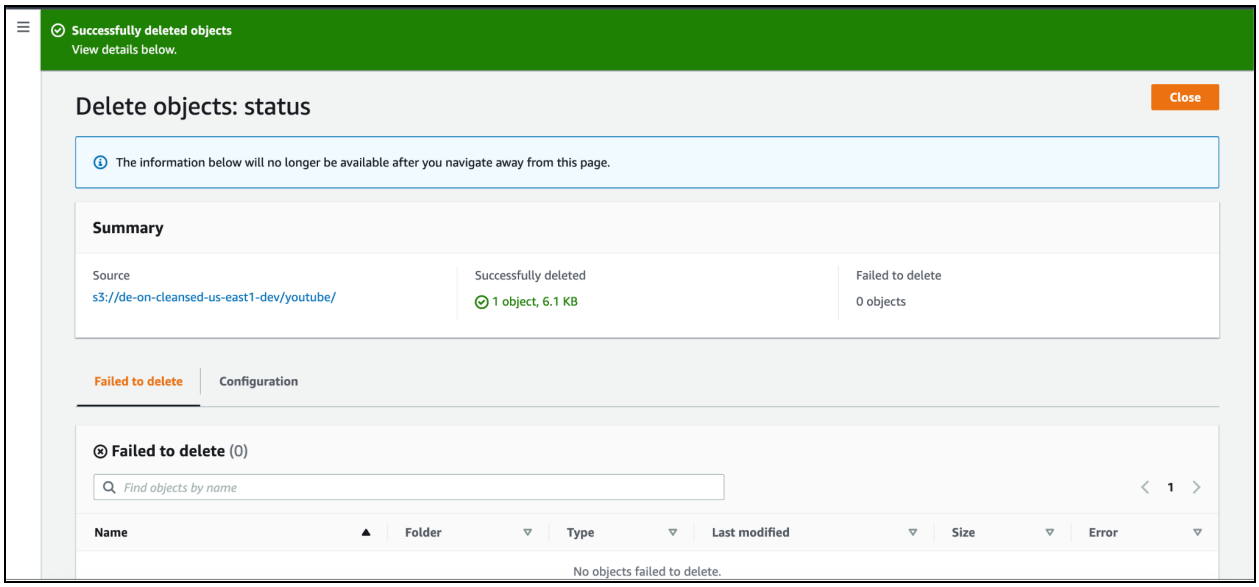


8. Changing the data type of id from String to BigInt

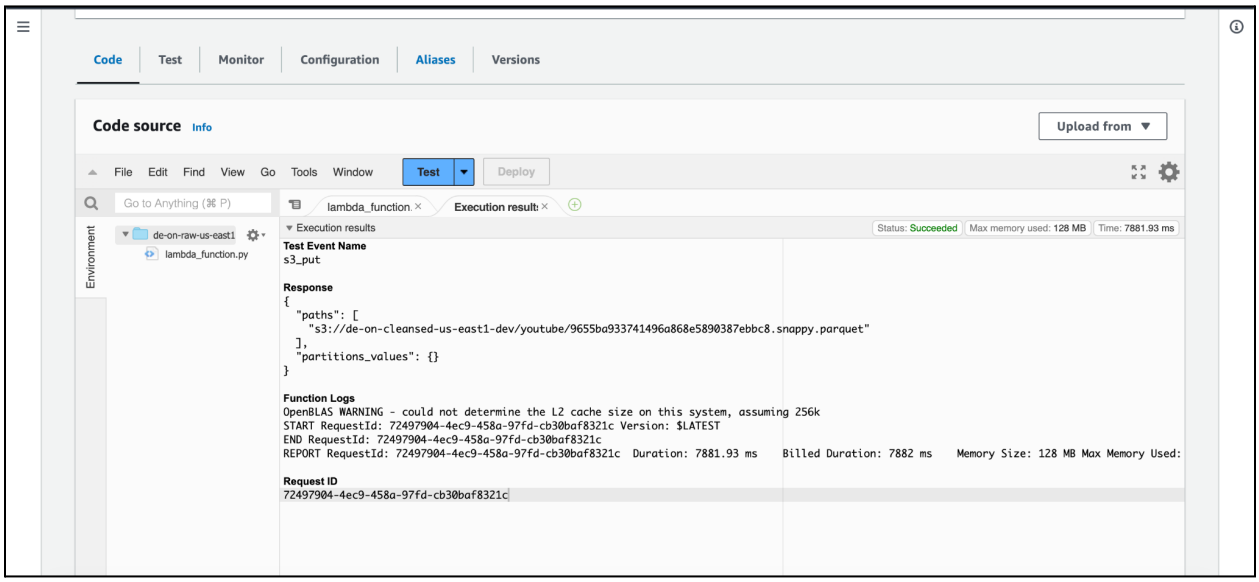
- Edit datatype in glue table



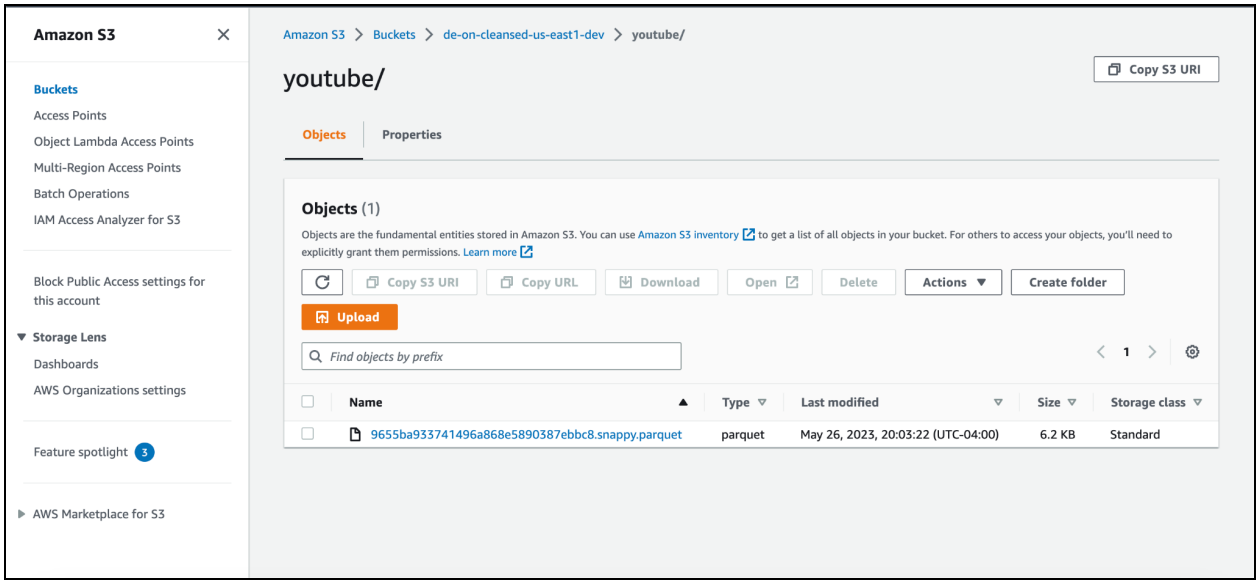
- Delete current parquet file in S3 Bucket



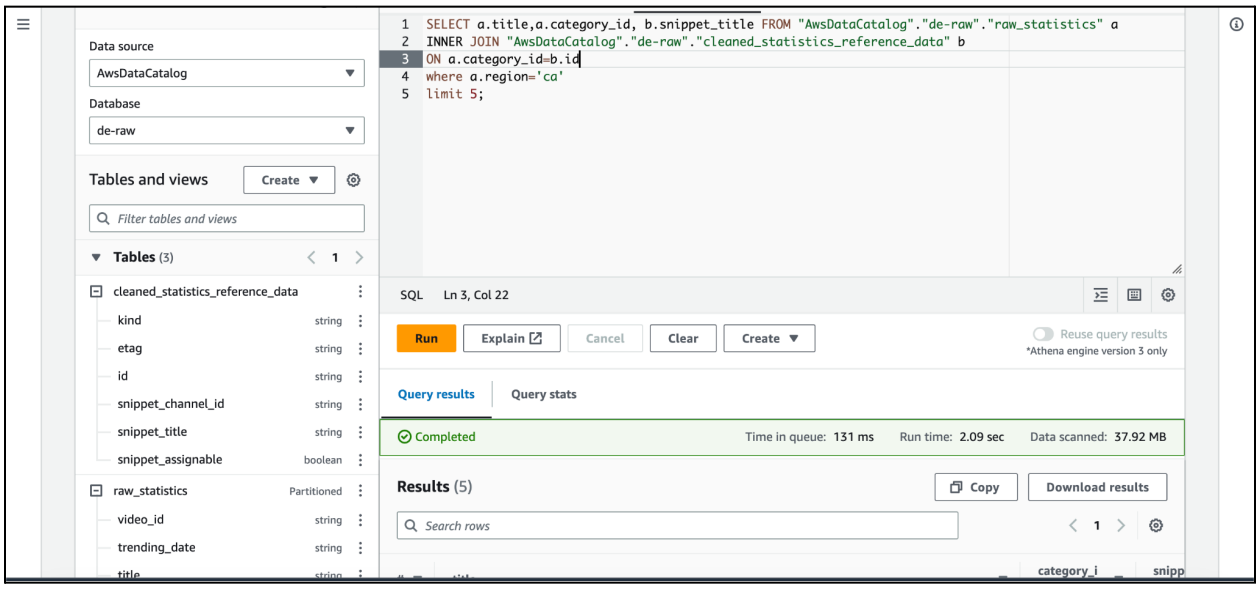
- Run Lambda function again

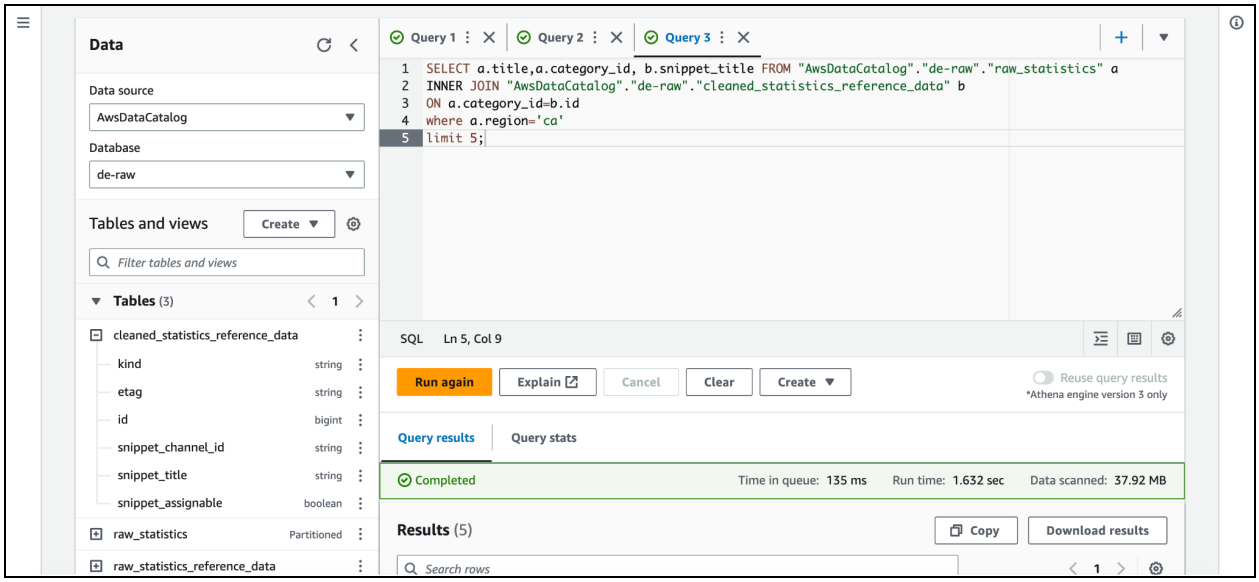


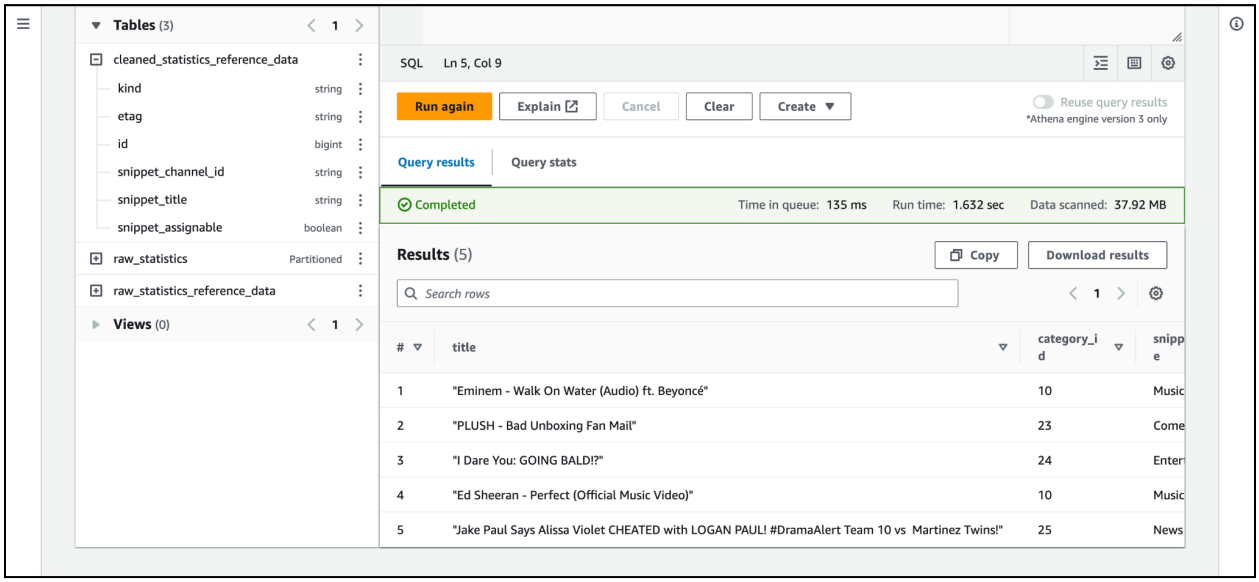
- New parquet file created in S3 bucket



- Successfully converted the data type

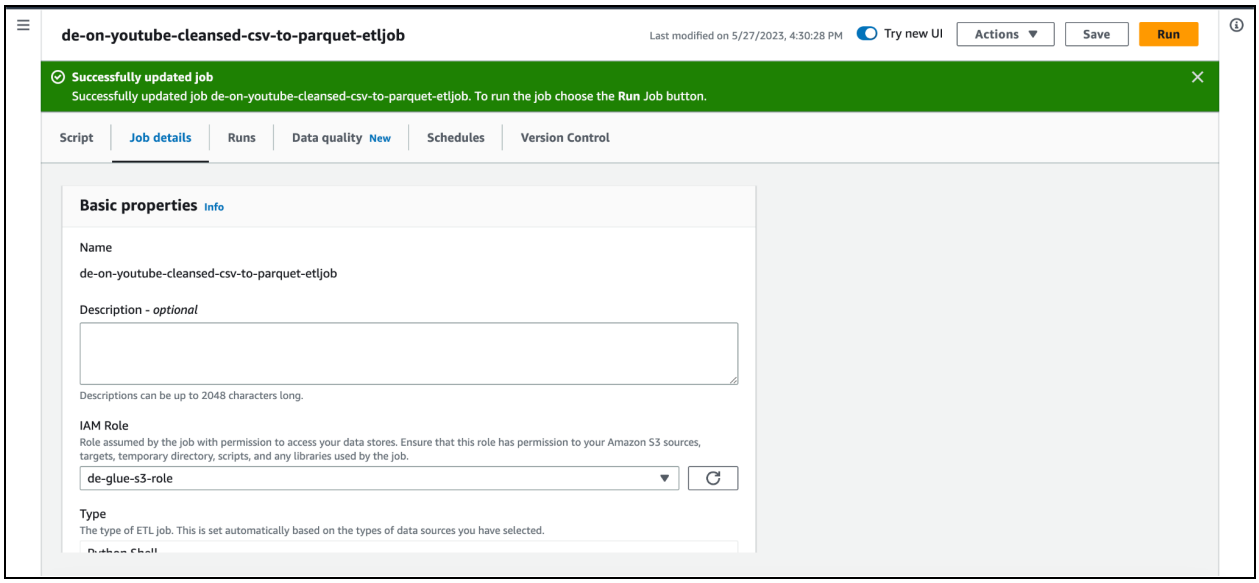


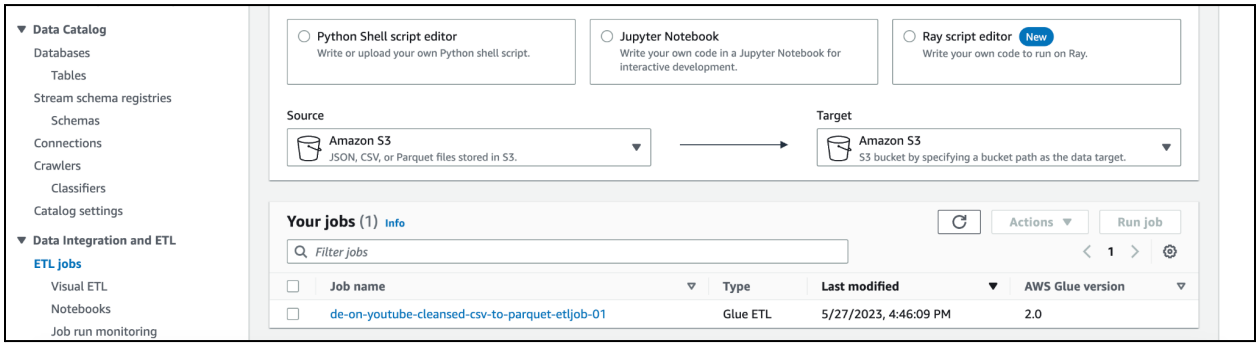




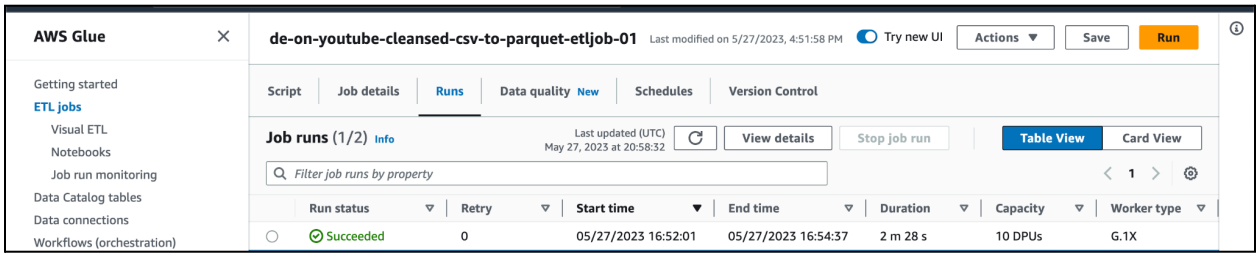
9. Convert all CSV data files into parquet format

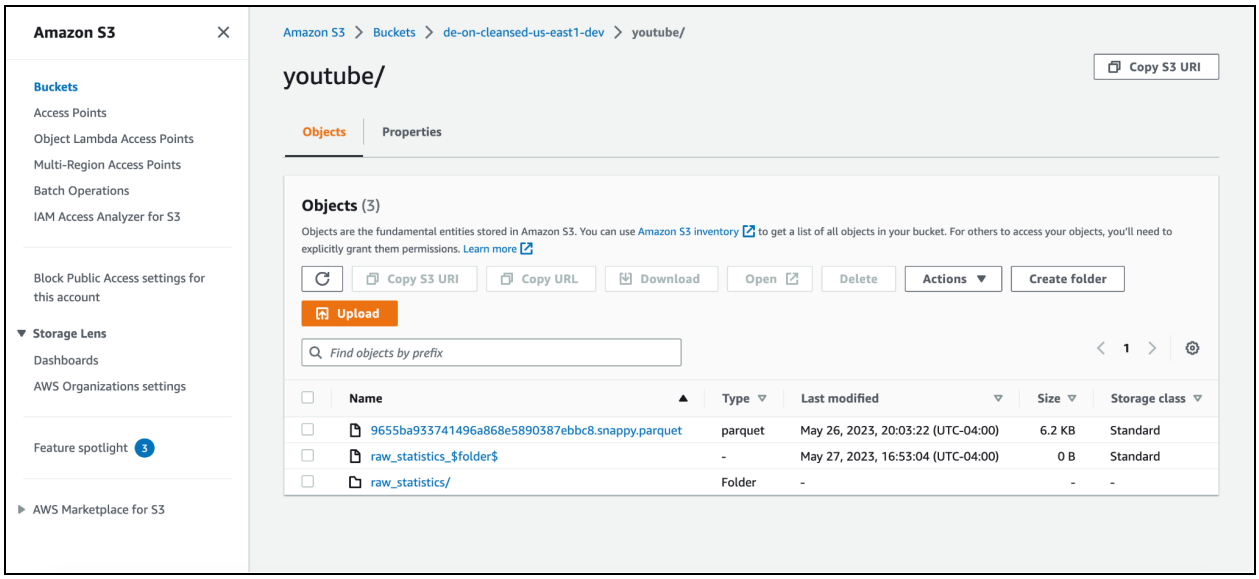
- Create ETL glue job

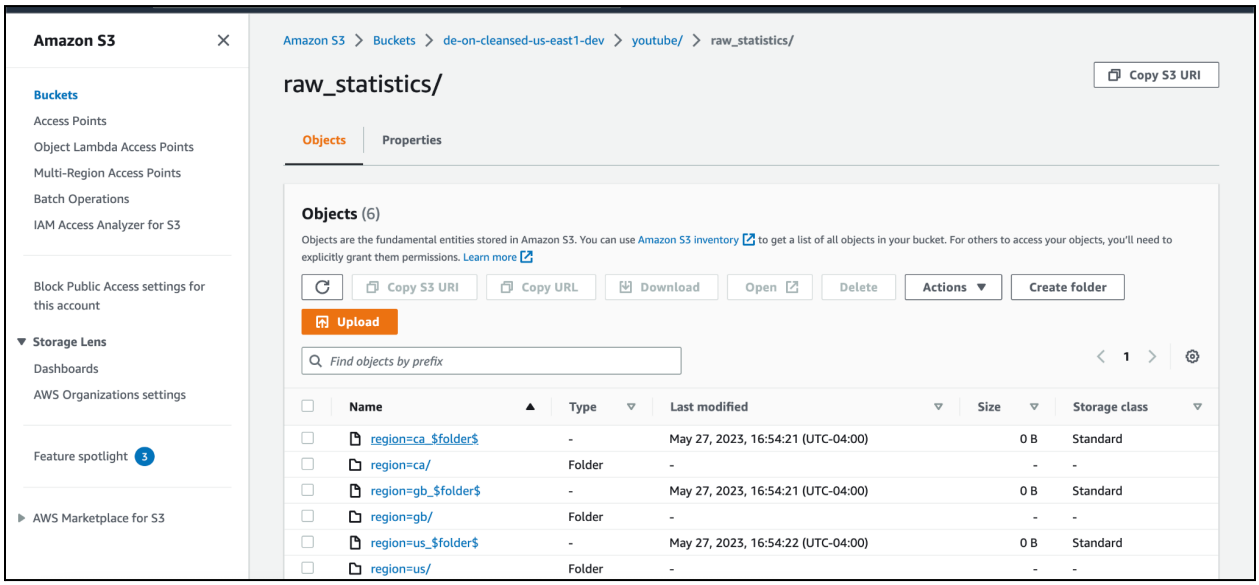

de-on-youtube-cleansed-csv-to-parquet-etljob
Last modified on 5/27/2023, 4:30:28 PM
Try new UI
Actions
Save
Run
Successfully updated job
Successfully updated job de-on-youtube-cleansed-csv-to-parquet-etljob. To run the job choose the Run Job button.
Script
Job details
Runs
Data quality New
Schedules
Version Control
Basic properties Info
Name
de-on-youtube-cleansed-csv-to-parquet-etljob
Description - optional
Descriptions can be up to 2048 characters long.
IAM Role
Role assumed by the job with permission to access your data stores. Ensure that this role has permission to your Amazon S3 sources, targets, temporary directory, scripts, and any libraries used by the job.
de-glue-s3-role
Type
The type of ETL job. This is set automatically based on the types of data sources you have selected.


Data Catalog
Databases
Tables
Stream schema registries
Schemas
Connections
Crawlers
Classifiers
Catalog settings
Data Integration and ETL
ETL jobs
Visual ETL
Notebooks
Job run monitoring
Python Shell script editor
Write or upload your own Python shell script.
Jupyter Notebook
Write your own code in a Jupyter Notebook for interactive development.
Ray script editor
New
Write your own code to run on Ray.
Source
Amazon S3
JSON, CSV, or Parquet files stored in S3.
Target
Amazon S3
S3 bucket by specifying a bucket path as the data target.
Your jobs (1) Info
Actions
Run job
Filter jobs
Job name
Type
Last modified
AWS Glue version
de-on-youtube-cleansed-csv-to-parquet-etljob-01
Glue ETL
5/27/2023, 4:46:09 PM
2.0

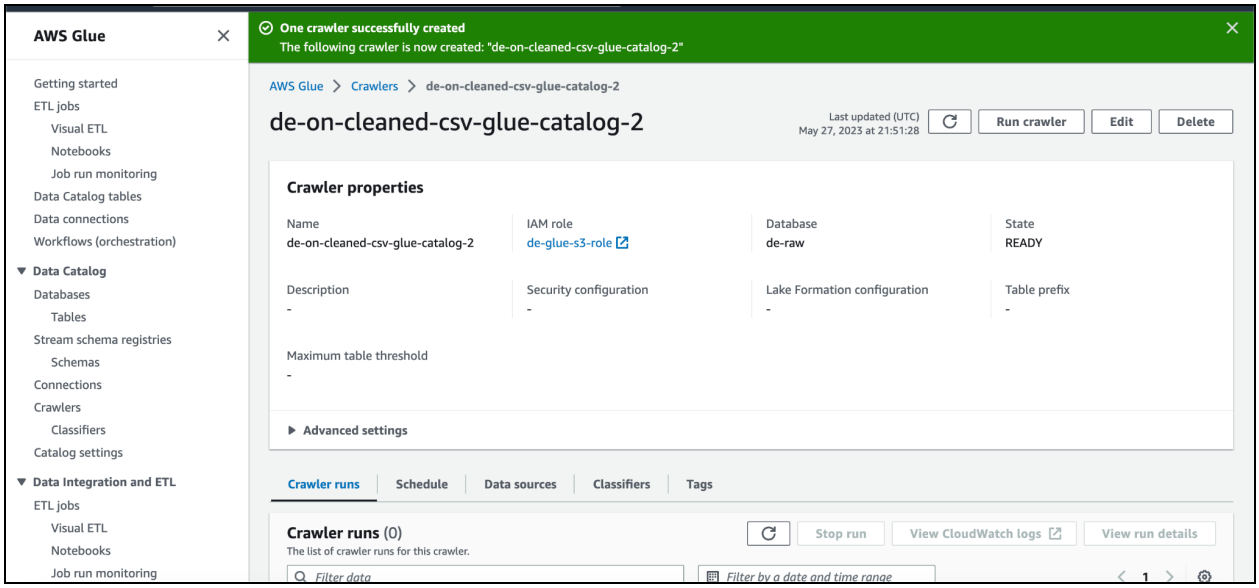
- Run Glue Job

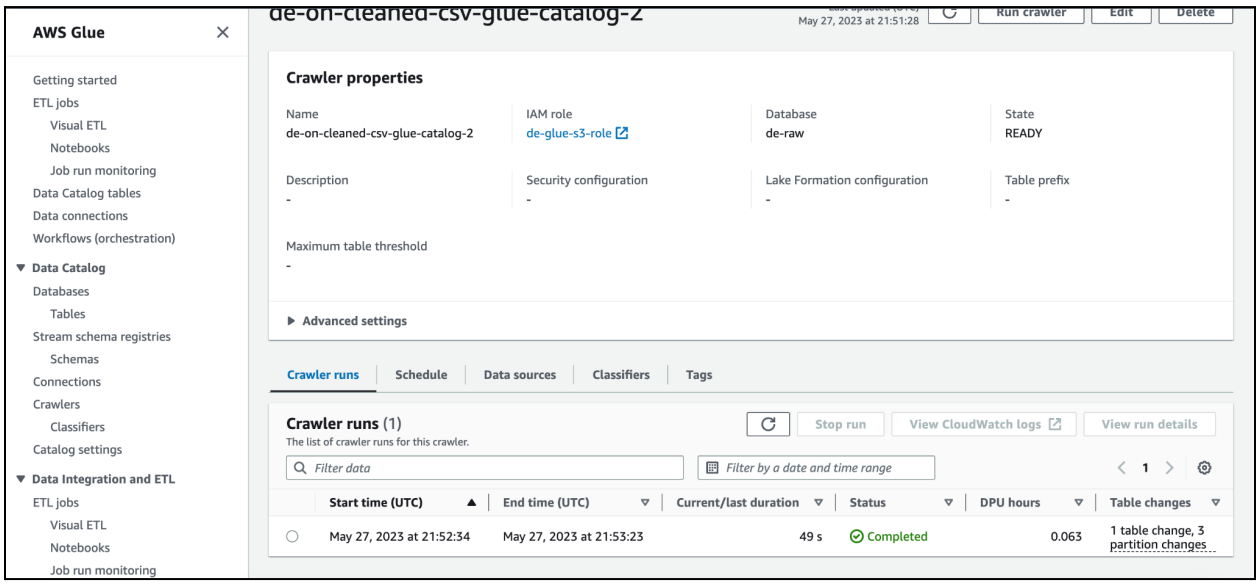

AWS Glue
Getting started
ETL jobs
Visual ETL
Notebooks
Job run monitoring
Data Catalog tables
Data connections
Workflows (orchestration)
de-on-youtube-cleansed-csv-to-parquet-etljob-01
Last modified on 5/27/2023, 4:51:58 PM
Try new UI
Actions
Save
Run
Script
Job details
Runs
Data quality New
Schedules
Version Control
Job runs (1/2) Info
Last updated (UTC)
May 27, 2023 at 20:58:32
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
Succeeded
0
05/27/2023 16:52:01
05/27/2023 16:54:37
2 m 28 s
10 DPUs
G.1X

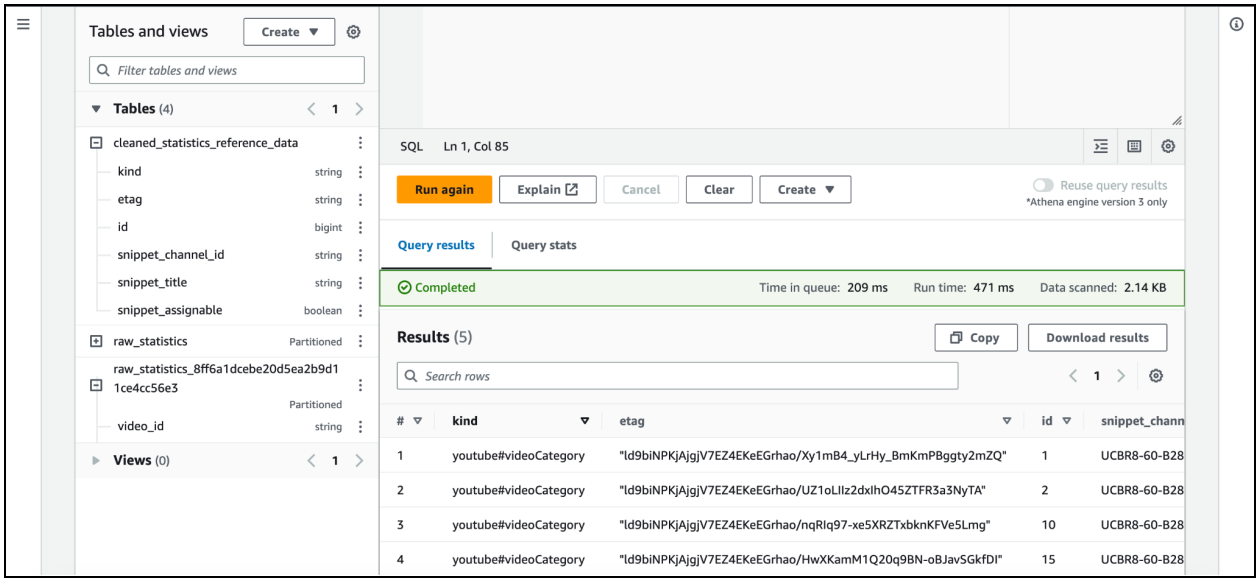




- Create crawlers for parquet files of regions data

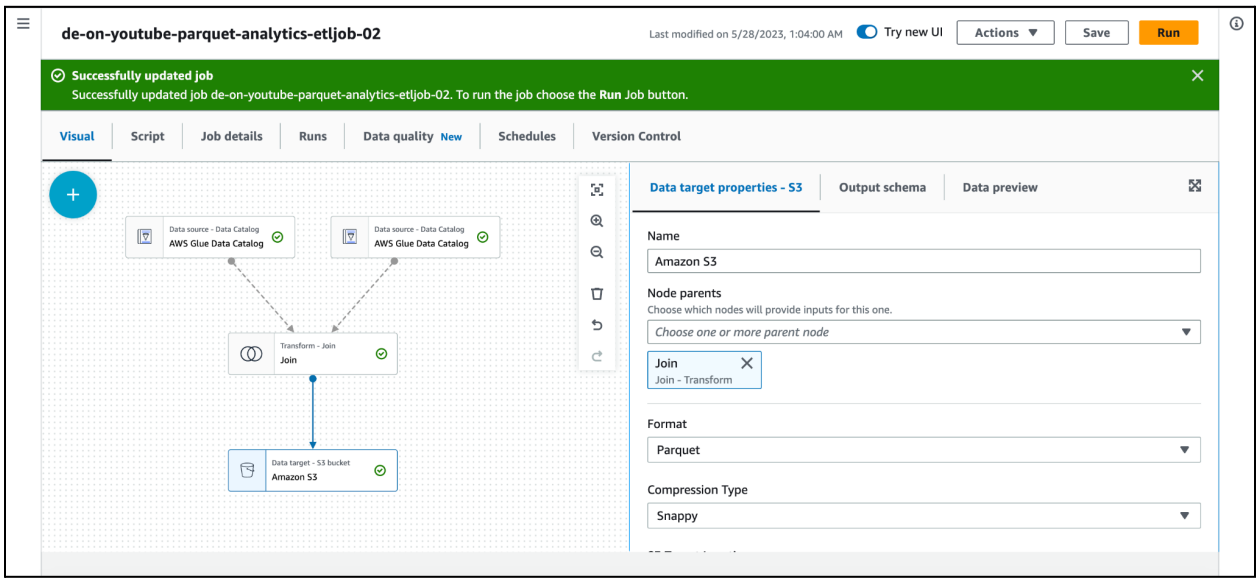

AWS Glue
Getting started
ETL jobs
Visual ETL
Notebooks
Job run monitoring
Data Catalog tables
Data connections
Workflows (orchestration)
Data Catalog
Databases
Tables
Stream schema registries
Schemas
Connections
Crawlers
Classifiers
Catalog settings
Data Integration and ETL
ETL jobs
Visual ETL
Notebooks
Job run monitoring
One crawler successfully created
The following crawler is now created: "de-on-cleaned-csv-glue-catalog-2"
AWS Glue > Crawlers > de-on-cleaned-csv-glue-catalog-2
de-on-cleaned-csv-glue-catalog-2
Last updated (UTC)
May 27, 2023 at 21:51:28
Run crawler
Edit
Delete
Crawler properties
Name
de-on-cleaned-csv-glue-catalog-2
IAM role
de-glue-s3-role
Database
de-raw
State
READY
Description
-
Security configuration
-
Lake Formation configuration
-
Table prefix
-
Maximum table threshold
-
Advanced settings
Crawler runs
Schedule
Data sources
Classifiers
Tags
Crawler runs (0)
The list of crawler runs for this crawler.
Stop run
View CloudWatch logs
View run details
Filter data
Filter by a date and time range
1


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
ETL jobs
Visual ETL
Notebooks
Job run monitoring
May 27, 2023 at 21:51:28
Run crawler
Edit
Delete
Crawler properties
Name
de-on-cleaned-csv-glue-catalog-2
IAM role
de-glue-s3-role
Database
de-raw
State
READY
Description
-
Security configuration
-
Lake Formation configuration
-
Table prefix
-
Maximum table threshold
-
Advanced settings
Crawler runs
Schedule
Data sources
Classifiers
Tags
Crawler runs (1)
The list of crawler runs for this crawler.
Stop run
View CloudWatch logs
View run details
Filter data
Filter by a date and time range
1
Start time (UTC)
End time (UTC)
Current/last duration
Status
DPU hours
Table changes
May 27, 2023 at 21:52:34
May 27, 2023 at 21:53:23
49 s
Completed
0.063
1 table change, 3 partition changes

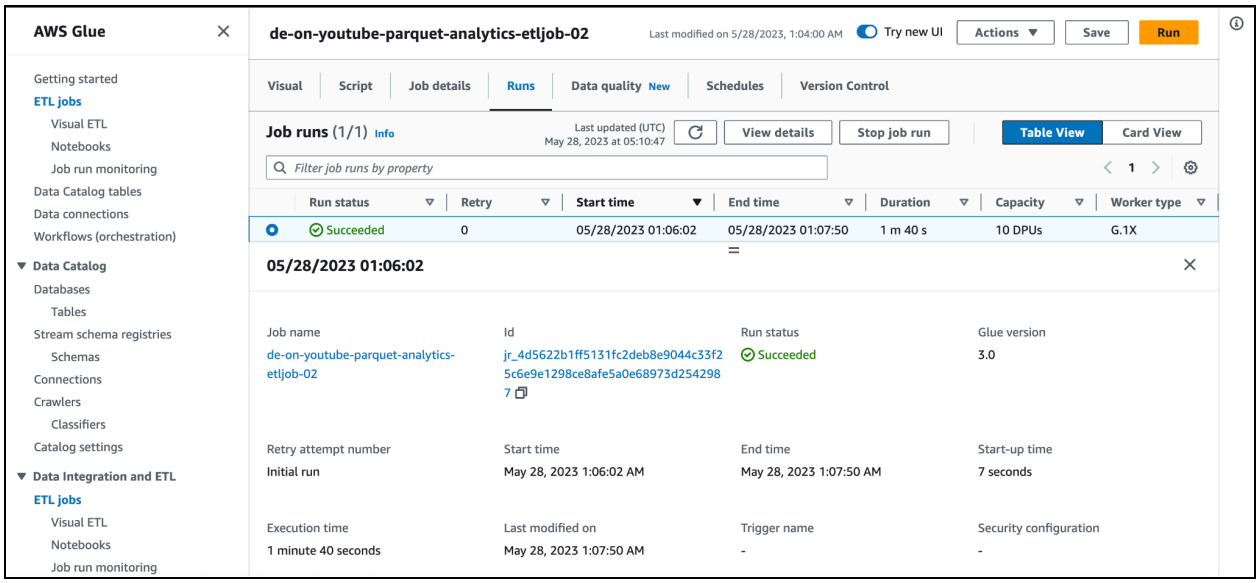


10. ETL pipeline for Reporting and Analytics

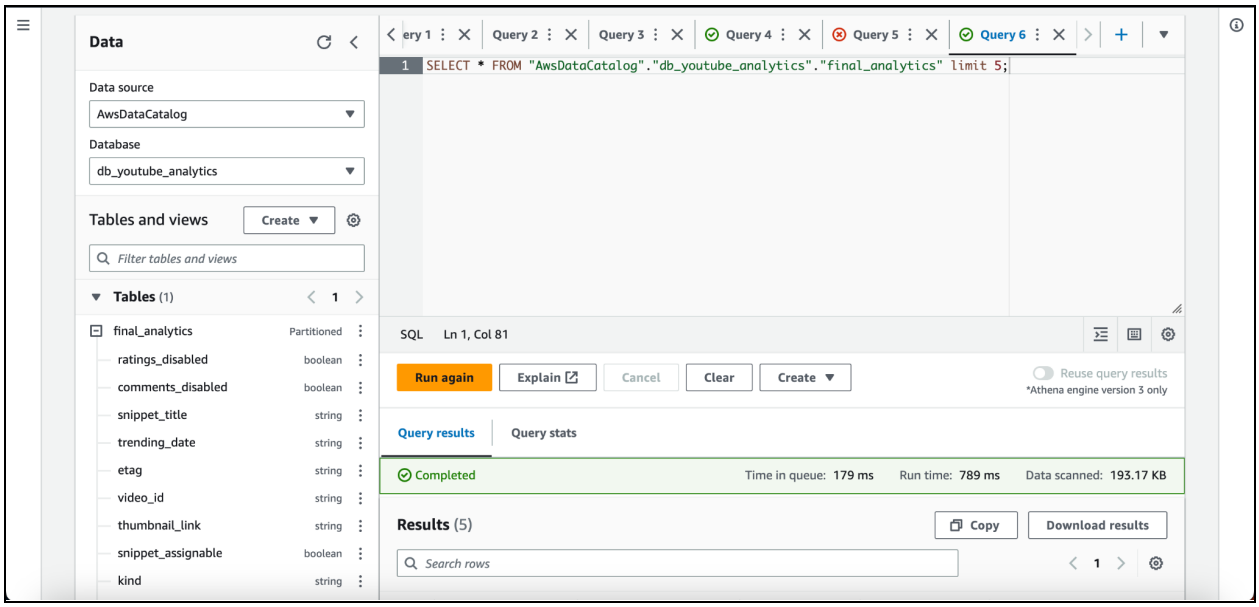
- Created ETL job for joining tables in AWS Glue

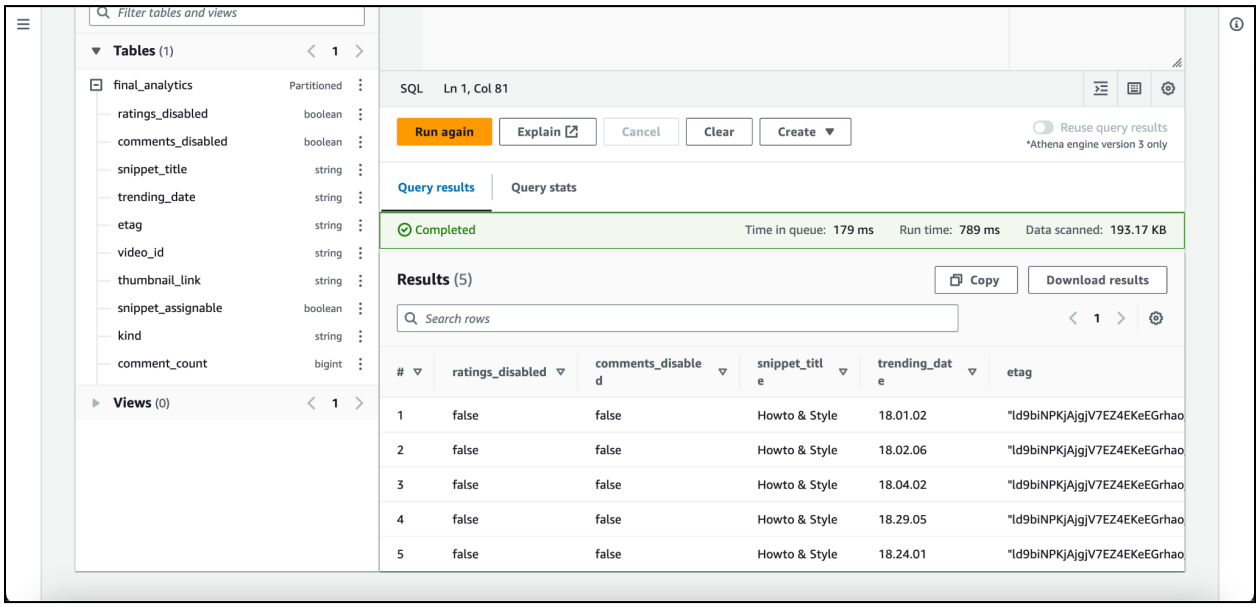


- Ran Job



- Athena Output of One Final table


Data
Data source
AwsDataCatalog
Database
db_youtube_analytics
Tables and views
Create
Filter tables and views
Tables (1)
final_analytics Partitioned
ratings_disabled boolean
comments_disabled boolean
snippet_title string
trending_date string
etag string
video_id string
thumbnail_link string
snippet_assignable boolean
kind string
Query 1
Query 2
Query 3
Query 4
Query 5
Query 6
1 SELECT * FROM "AwsDataCatalog"."db_youtube_analytics"."final_analytics" limit 5;
SQL Ln 1, Col 81
Run again
Explain
Cancel
Clear
Create
Reuse query results
*Athena engine version 3 only
Query results
Query stats
Completed
Time in queue: 179 ms
Run time: 789 ms
Data scanned: 193.17 KB
Results (5)
Copy
Download results
Search rows
1


Filter tables and views
Tables (1)
final_analytics Partitioned
ratings_disabled boolean
comments_disabled boolean
snippet_title string
trending_date string
etag string
video_id string
thumbnail_link string
snippet_assignable boolean
kind string
comment_count bigint
Views (0)
SQL Ln 1, Col 81
Run again
Explain
Cancel
Clear
Create
Reuse query results
*Athena engine version 3 only
Query results
Query stats
Completed
Time in queue: 179 ms
Run time: 789 ms
Data scanned: 193.17 KB
Results (5)
Copy
Download results
Search rows
1
ratings_disabled comments_disabled snippet_title trending_date etag
1 false false Howto & Style 18.01.02 "ld9biNPKjAjgjV7EZ4EKeEGrhao
2 false false Howto & Style 18.02.06 "ld9biNPKjAjgjV7EZ4EKeEGrhao
3 false false Howto & Style 18.04.02 "ld9biNPKjAjgjV7EZ4EKeEGrhao
4 false false Howto & Style 18.29.05 "ld9biNPKjAjgjV7EZ4EKeEGrhao
5 false false Howto & Style 18.24.01 "ld9biNPKjAjgjV7EZ4EKeEGrhao