Experiment : 6

Title : Querying Data in S3 with Amazon Athena

Date: 06/10/2022

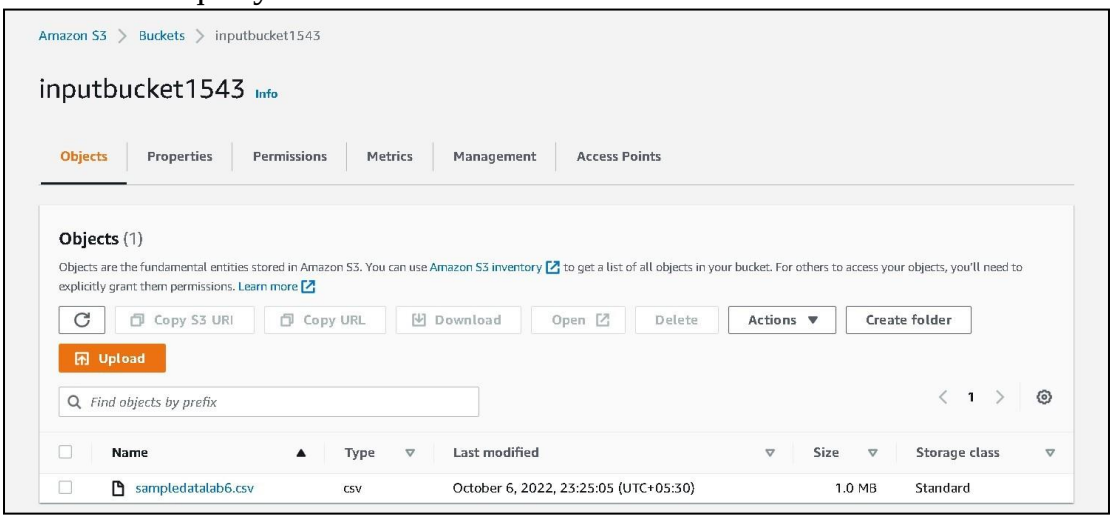
Aim : AWS Athena to query JSON/CSV files located in an s3 bucket

Pre-requisites : AWS Console, Amazon S3, Amazon Athena, Amazon Crawler, sampledata.csv file

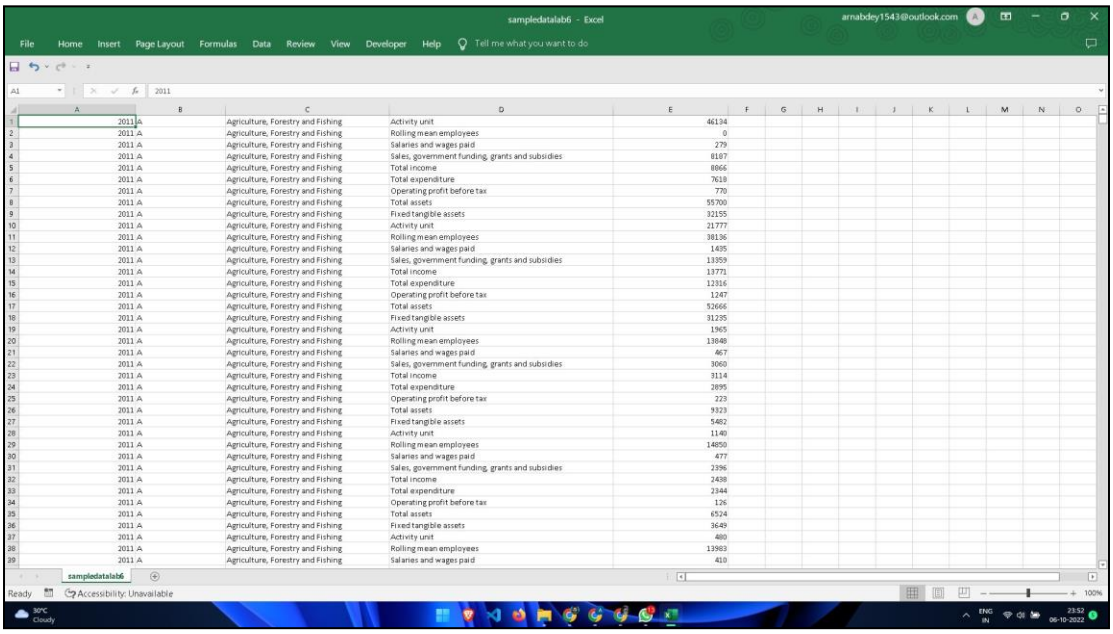
Procedure :

Steps:

1. Login to your AWS IAM account.
2. Create two buckets, one bucket for input data file and another for output result of the query.

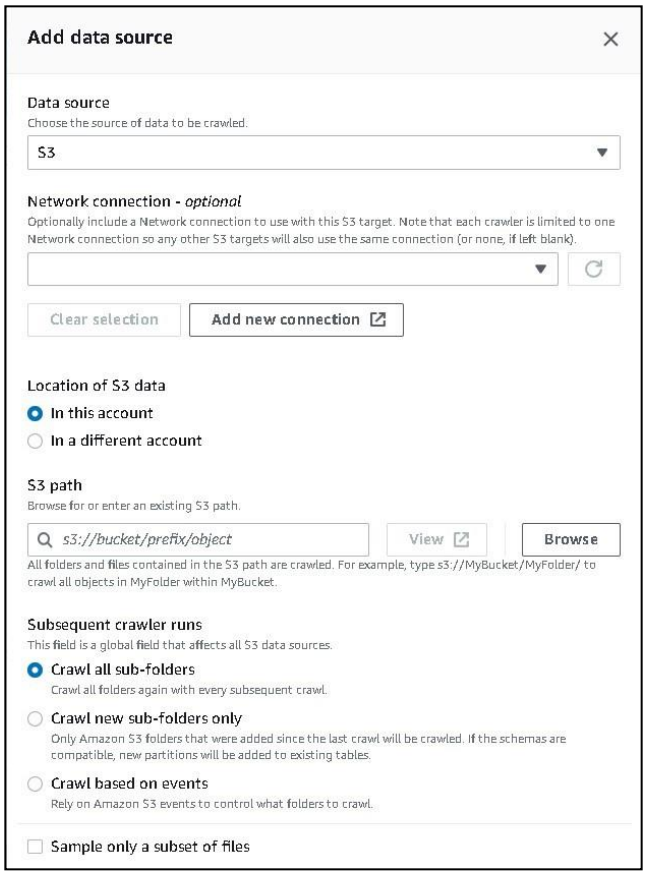


3. Upload a sample dataset (json, csv, tsv, etc) file in your AWS S3 bucket.



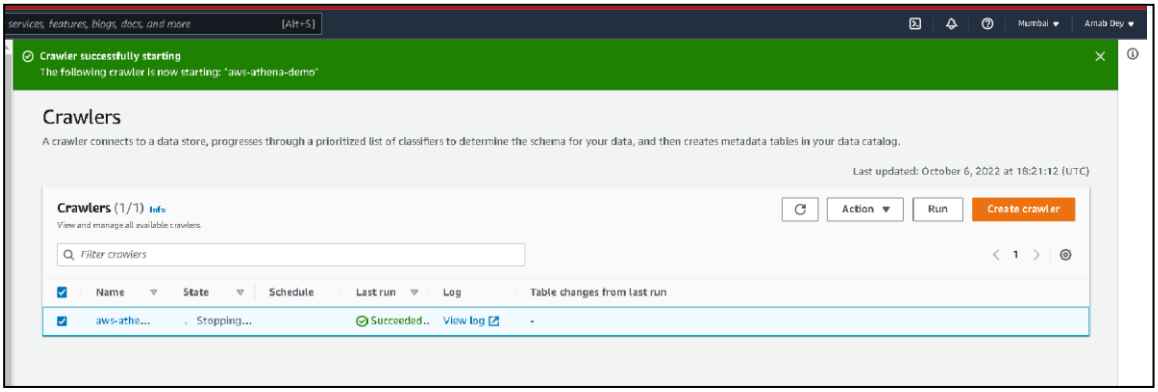
4. Go to AWS Athena.
5. Firstly, create a workgroup (workgroup is nothing but a kind of a container where our athena service stores the temporary data).
6. Give workgroup name, description, query result location, data usage limit.

7. Go to query editor panel, then go the settings, switch to your custom made workgroup from the primary. There are two ways to query s3 dataset –
   - Using aws glue crawler which inspect the json object within the source data bucket and then connect that to a pseudo table in athena.
   - The other alternative is to use a manual process where you specify the names and the types of each column.
8. Click on create drop down button, go for Aws Glue Crawler.
9. Create aws Crawler, enter crawler details – name, description, tags.
10. Add data source of the S3 bucket input file.

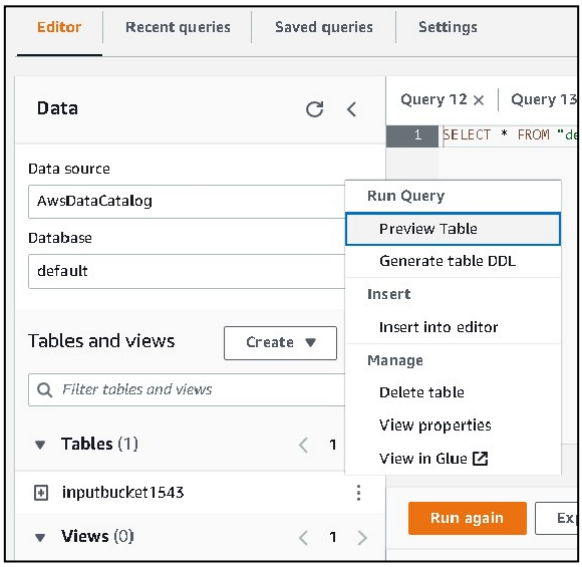


11. Add Database (default)
12. Create a IAM role to read the S3 contents.

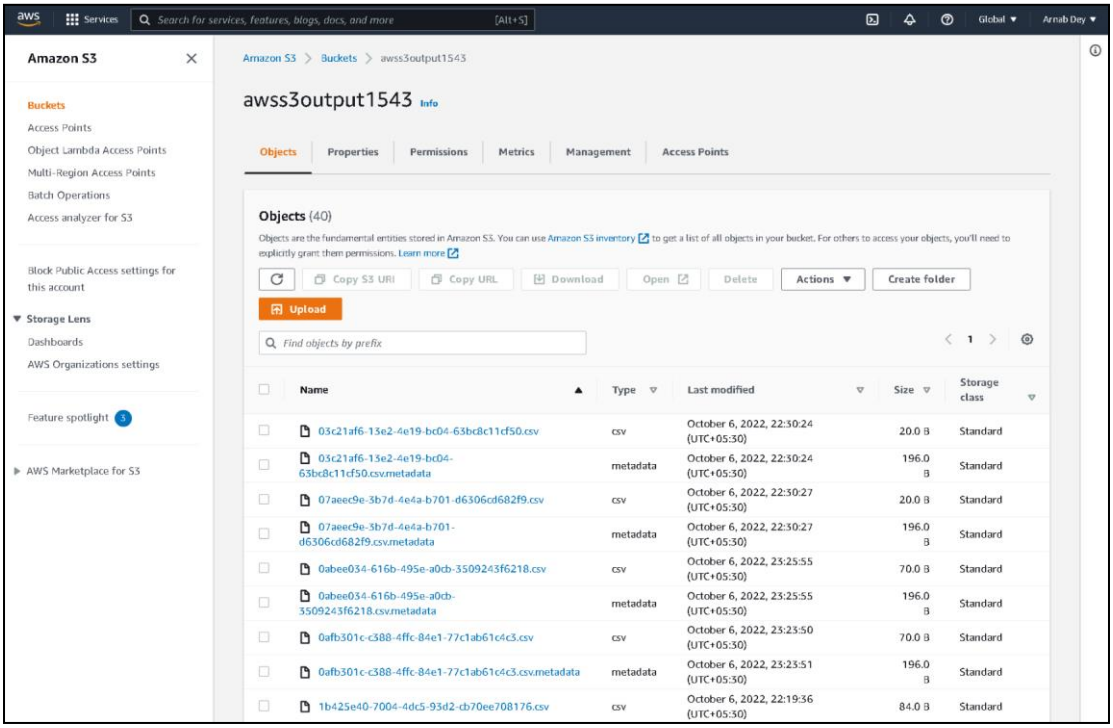
    (for other details, you can go with default configuration)
13. Run your crawler.

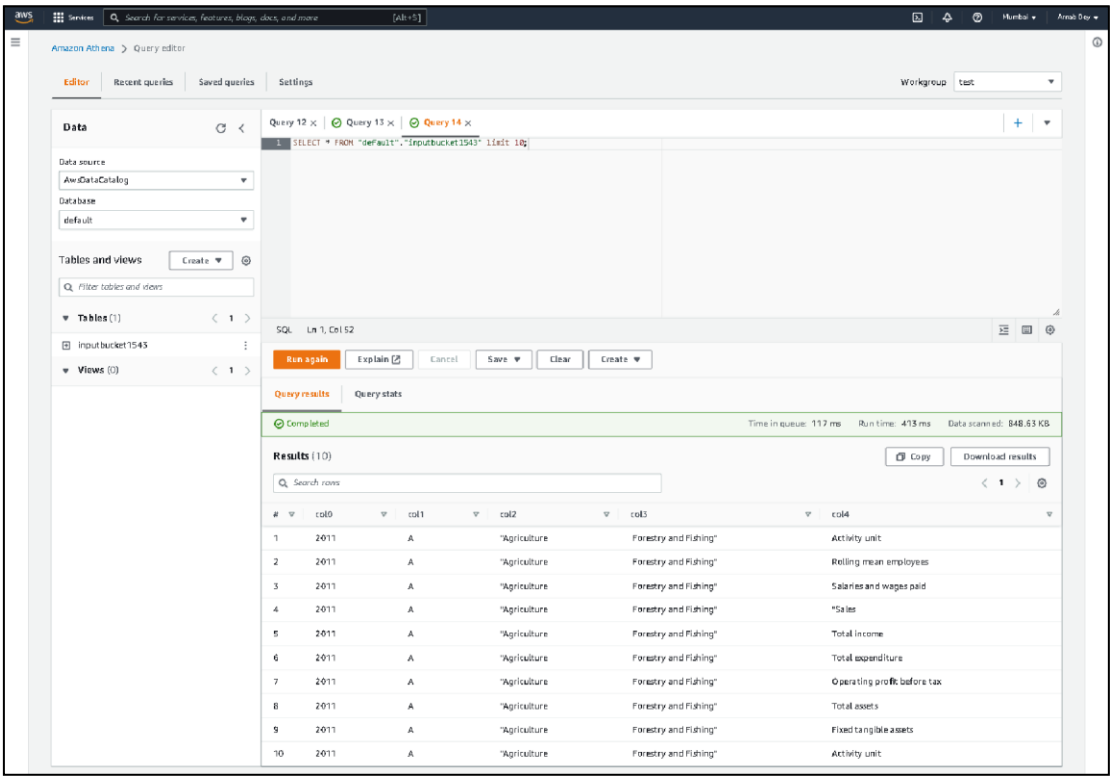


14. We see the table created in aws Athena, click on option button and preview table.

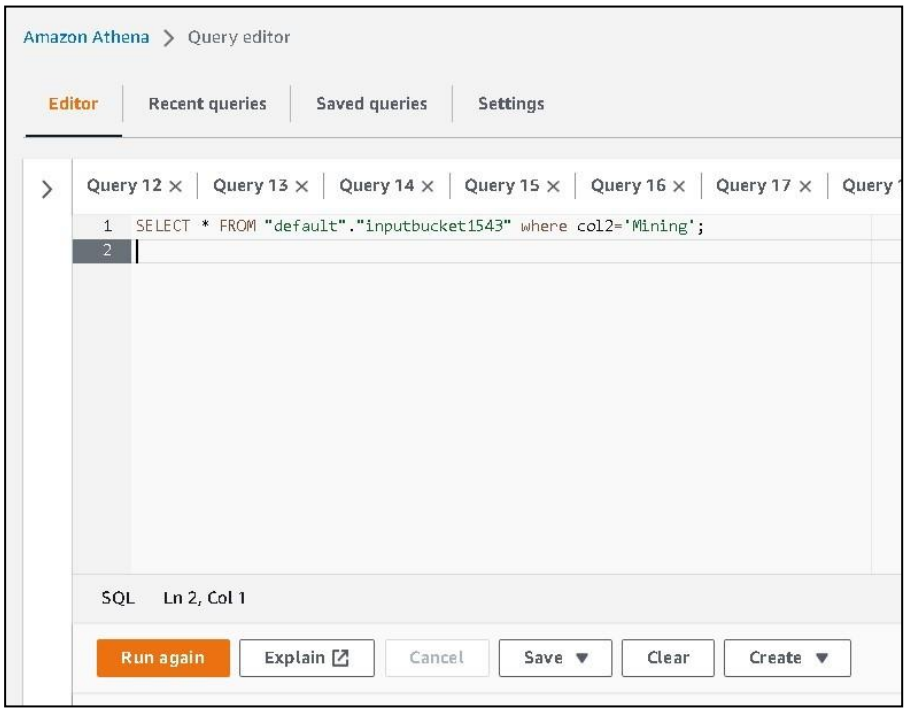


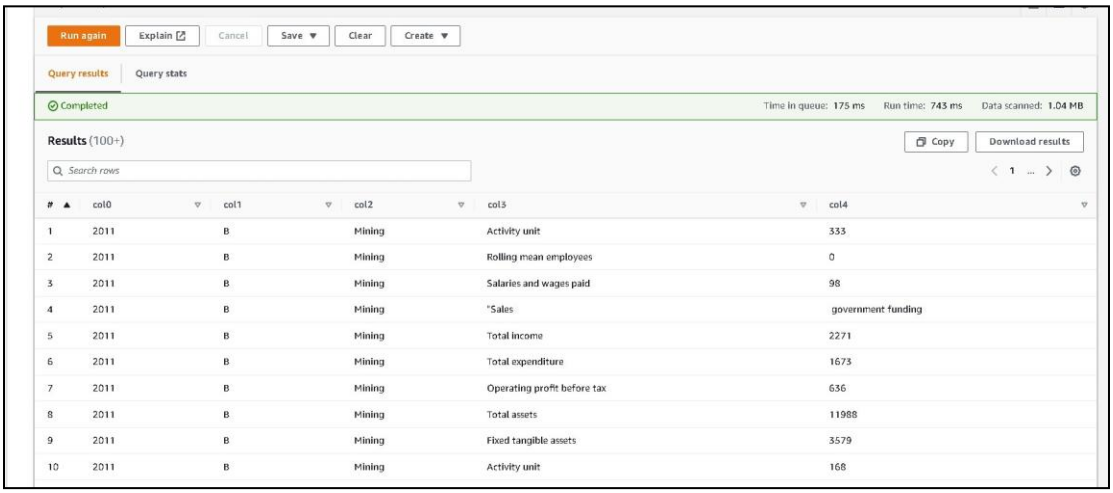
15. We get the data with the columns that were specified in the CSV file in S3.




Amazon Athena
Query editor
Editor
Recent queries
Saved queries
Settings
Workgroup
test
Data
Data source
AwsDataCatalog
Database
default
Tables and views
Create
Tables (1)
inputbucket1543
Views (0)
Query 12
Query 13
Query 14
SELECT * FROM "default"."inputbucket1543" limit 10;
SQL Ln 1, Col 52
Run again
Explain
Cancel
Save
Clear
Create
Query results
Query stats
Completed
Time in queue: 117 ms
Run time: 413 ms
Data scanned: 848.63 KB
Results (10)
Copy
Download results
Search rows
col0 col1 col2 col3 col4
1 2011 A "Agriculture Forestry and Fishing" Activity unit
2 2011 A "Agriculture Forestry and Fishing" Rolling mean employees
3 2011 A "Agriculture Forestry and Fishing" Salaries and wages paid
4 2011 A "Agriculture Forestry and Fishing" "Sales
5 2011 A "Agriculture Forestry and Fishing" Total income
6 2011 A "Agriculture Forestry and Fishing" Total expenditure
7 2011 A "Agriculture Forestry and Fishing" Operating profit before tax
8 2011 A "Agriculture Forestry and Fishing" Total assets
9 2011 A "Agriculture Forestry and Fishing" Fixed tangible assets
10 2011 A "Agriculture Forestry and Fishing" Activity unit

16. Run Query.


Amazon Athena
Query editor
Editor
Recent queries
Saved queries
Settings
Query 12
Query 13
Query 14
Query 15
Query 16
Query 17
1 SELECT * FROM "default"."inputbucket1543" where col2='Mining';
2
SQL Ln 2, Col 1
Run again
Explain
Cancel
Save
Clear
Create



Result:

We have successfully used AWS Athena to query JSON/CSV files located in an s3 bucket by setting up an Athena Database and Table using AWS Glue's Crawler.