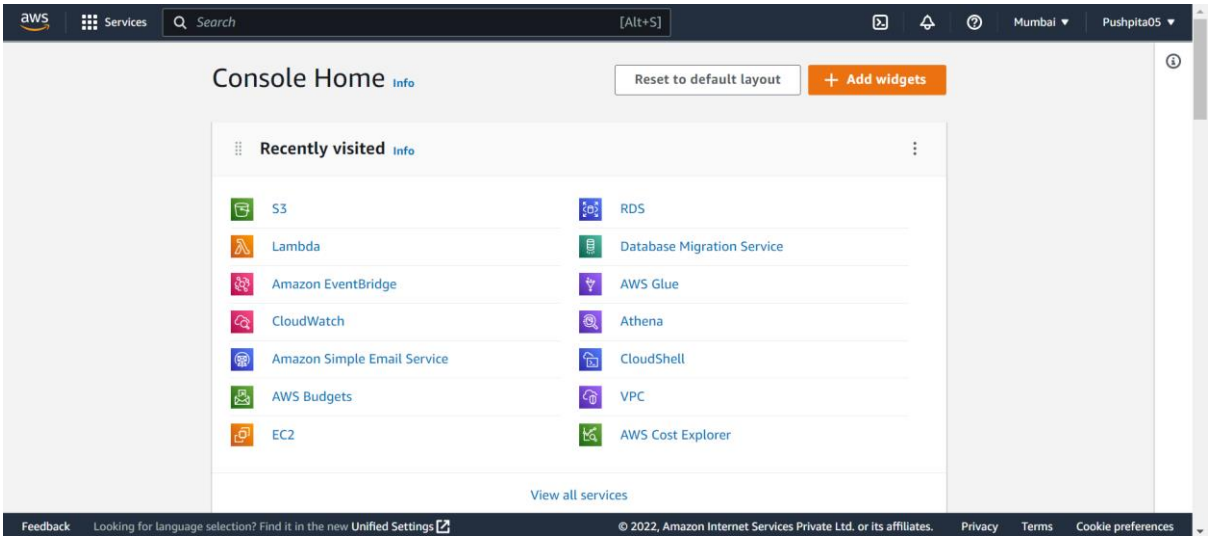
Experiment 6 - Querying Data in S3 with Amazon Athena

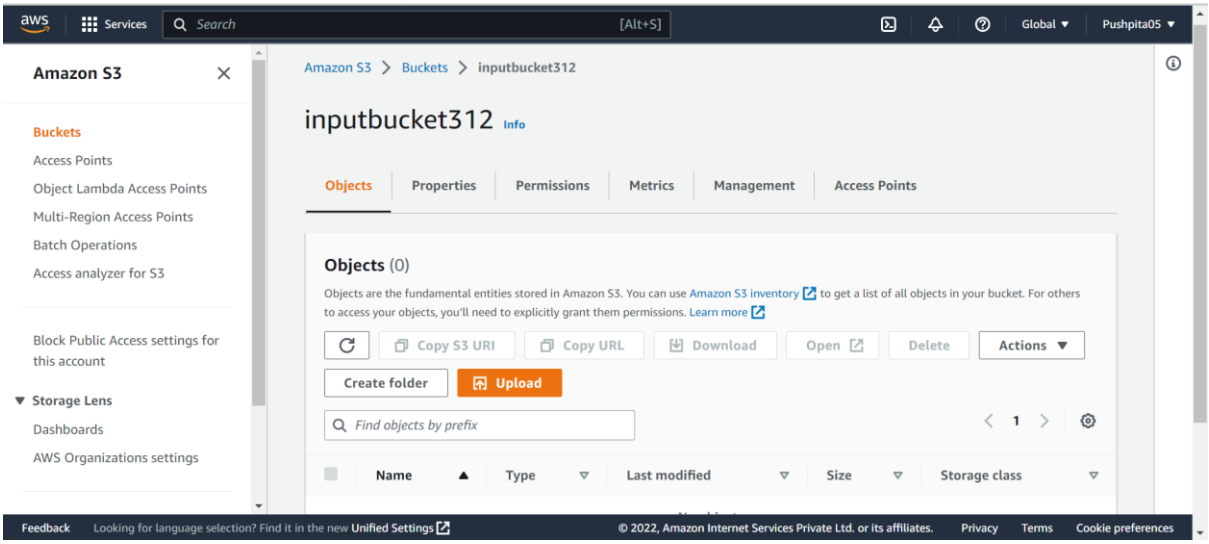
Aim: AWS Athena to query JSON/CSV files located in an s3 bucket

Procedure:

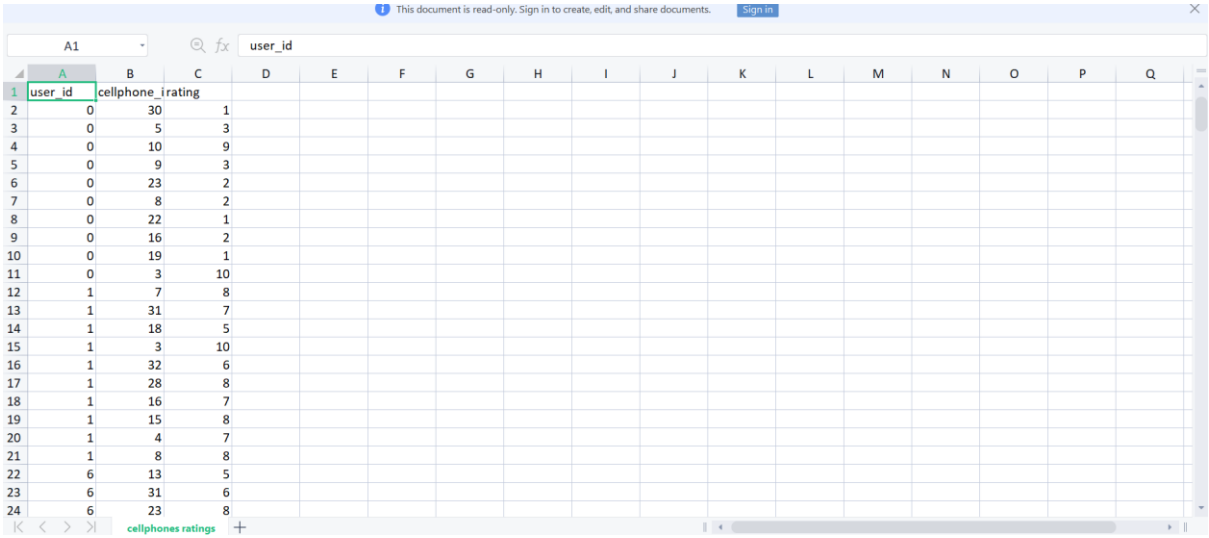
1. Firstly, open the AWS console homepage on browser (https://aws.amazon.com/console/).



2. Create two buckets, one bucket for input data file and another for output result of the query.

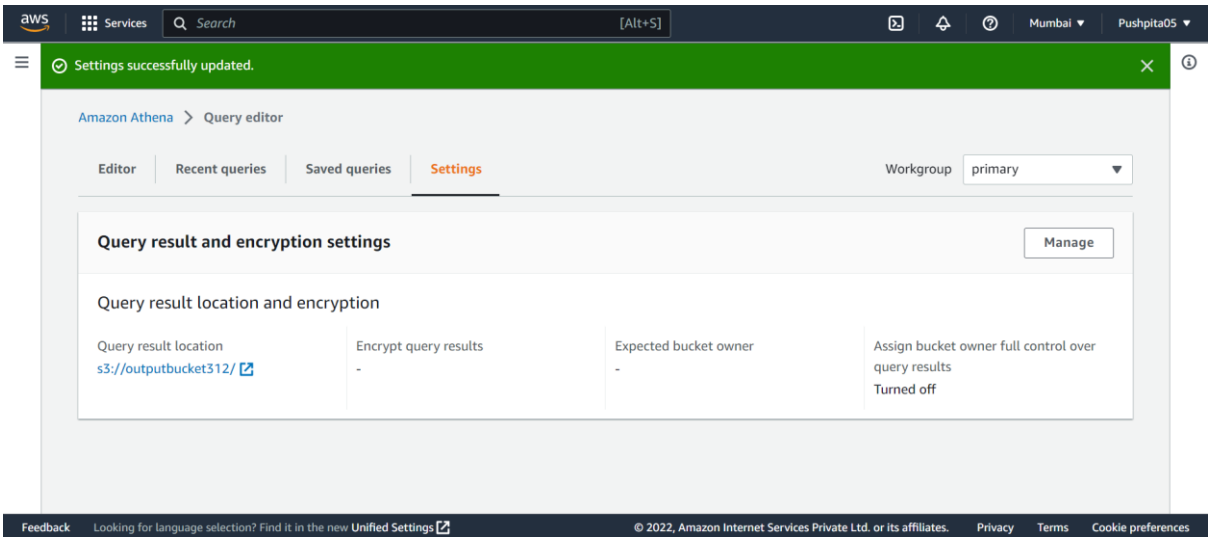


3. Upload a sample dataset (json, csv, tsv, etc) file in your AWS S3 bucket.



user_id	cellphone_i	rating
0	30	1
0	5	3
0	10	9
0	9	3
0	23	2
0	8	2
0	22	1
0	16	2
0	19	1
0	3	10
1	7	8
1	31	7
1	18	5
1	3	10
1	32	6
1	28	8
1	16	7
1	15	8
1	4	7
1	8	8
6	13	5
6	31	6
6	23	8

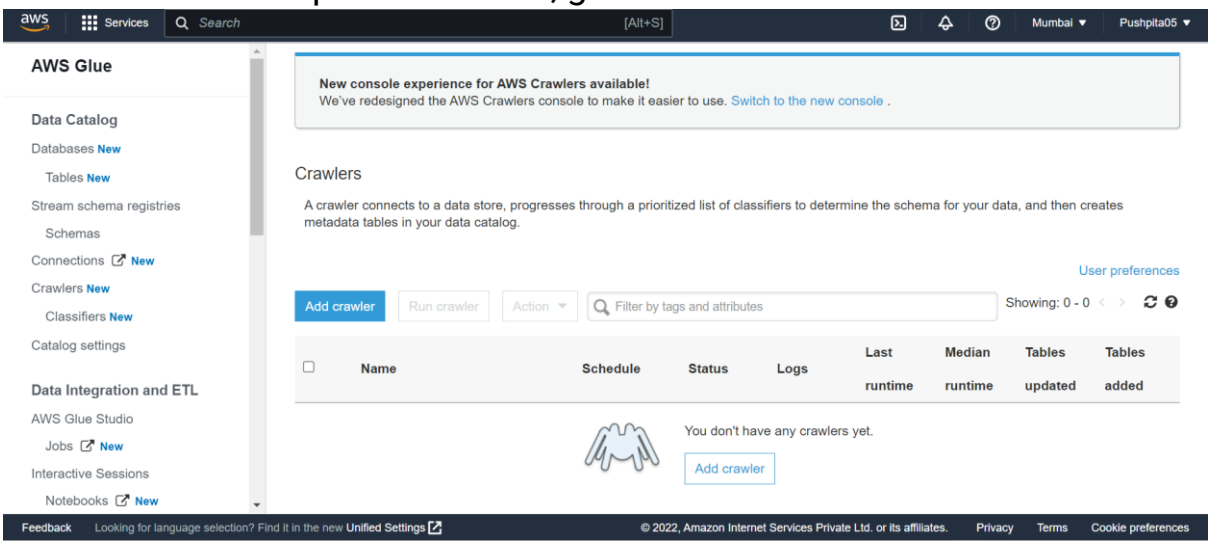
4. Go to AWS Athena.
5. Firstly, create a workgroup (workgroup is nothing but a kind of a container where our athena service stores the temporary data).
6. Give workgroup name, description, query result location, data usage limit.



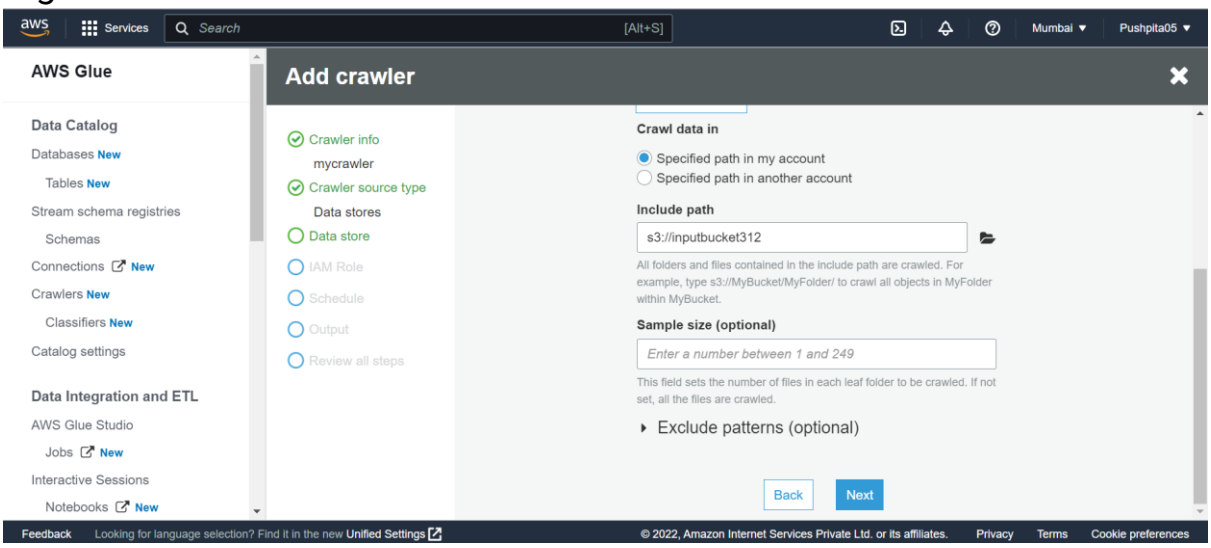
7. Go to query editor panel, then go the settings, switch to your custom-made workgroup from the primary.
   There are two ways to query s3 dataset –
   - Using aws glue crawler which inspect the json object within the source data bucket and then connect that to a pseudo table in athena.

- The other alternative is to use a manual process where you specify the names and the types of each column.

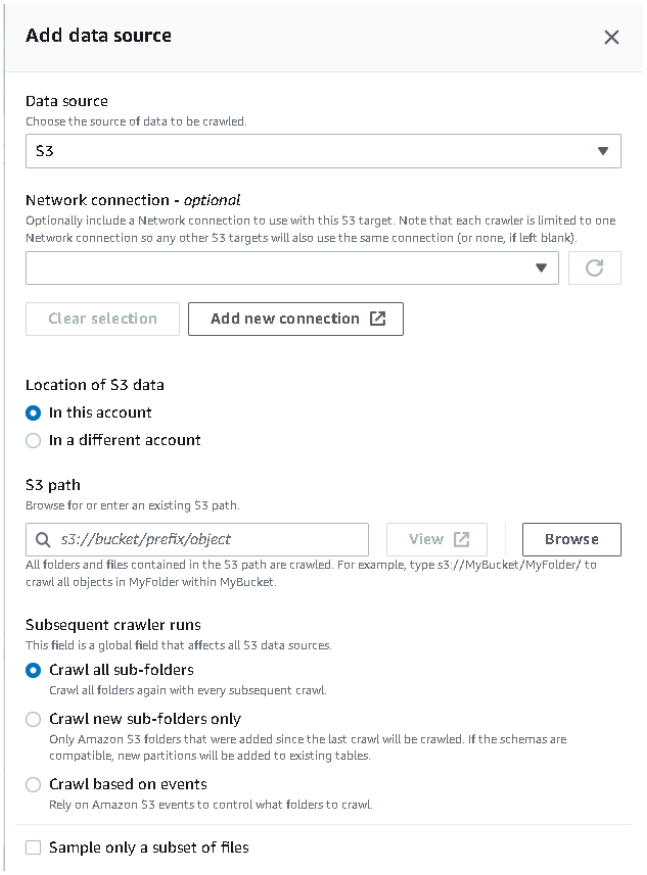
8. Click on create drop down button, go for Aws Glue Crawler.



9. Create aws Crawler, enter crawler details – name, description, tags.

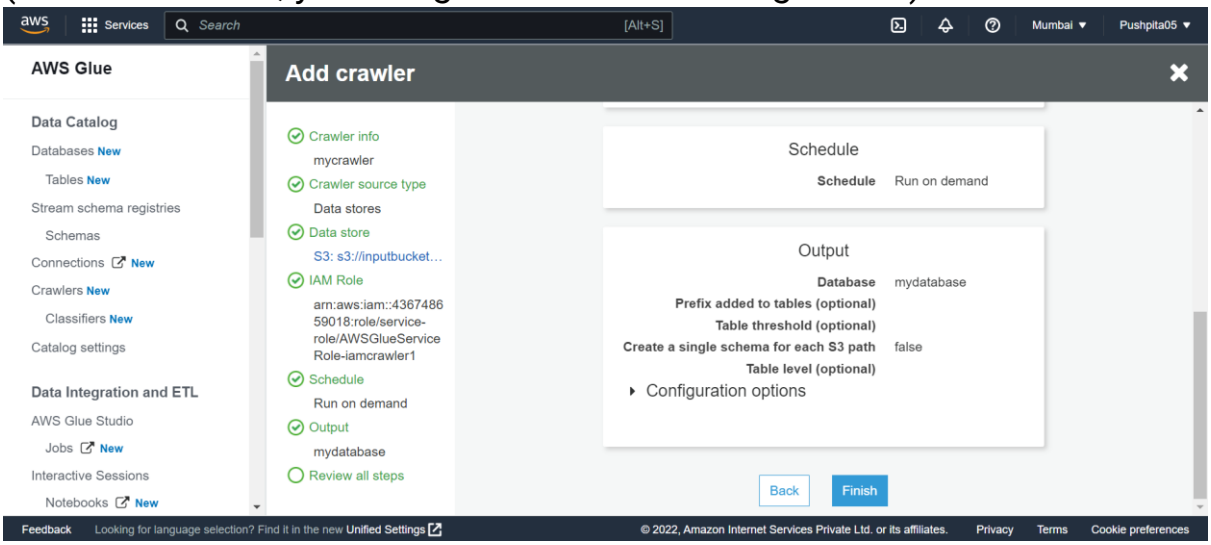


10. Add data source of the S3 bucket input file.

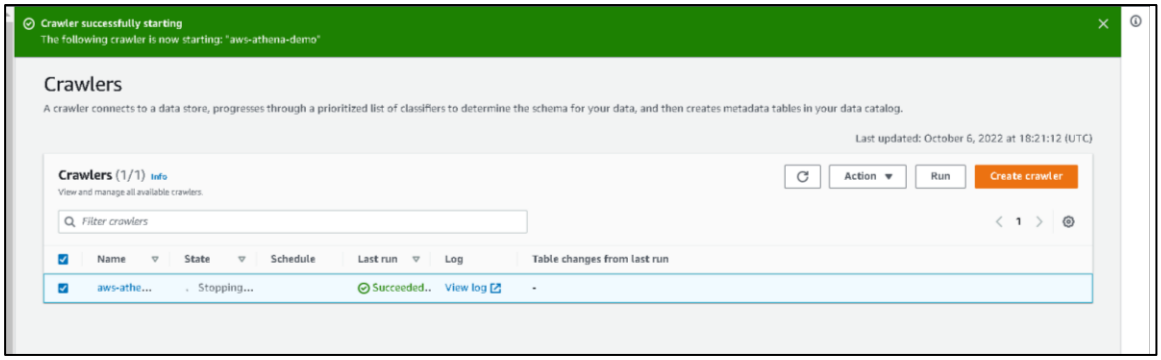


11. Add Database (default)
12. Create a IAM role to read the S3 contents.

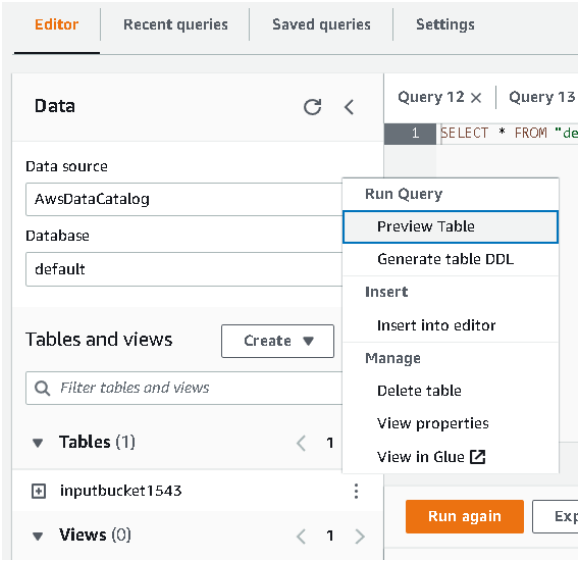
(for other details, you can go with default configuration)



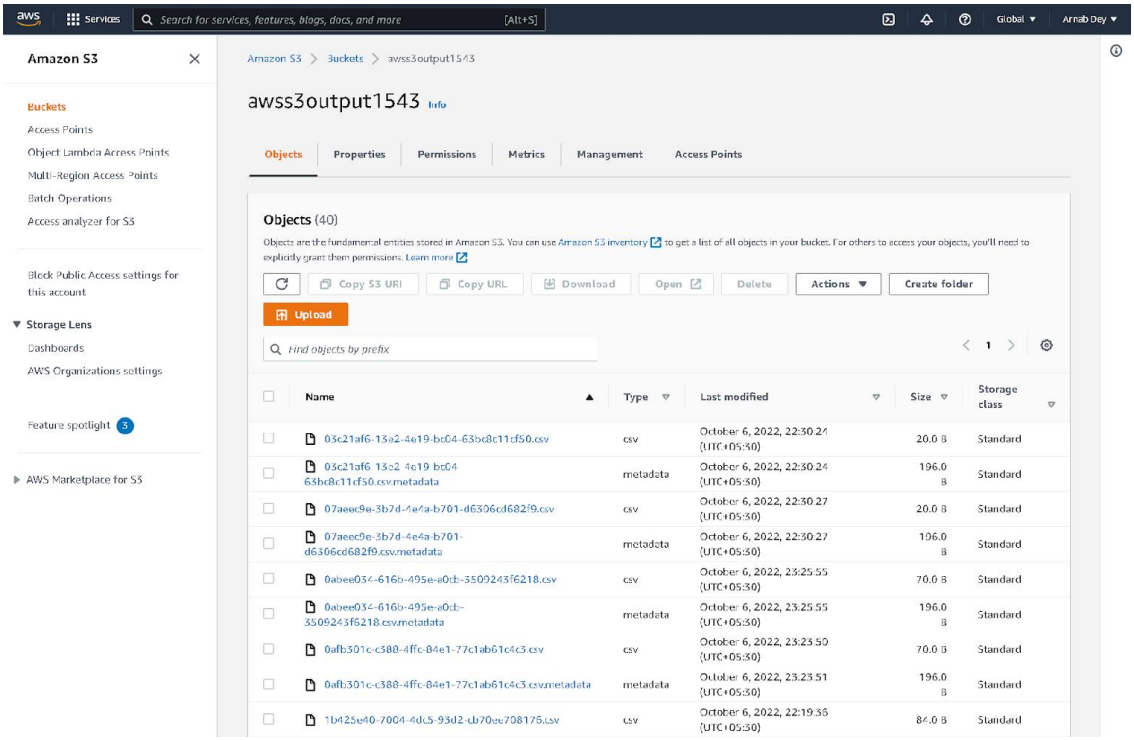
13. Run your crawler.

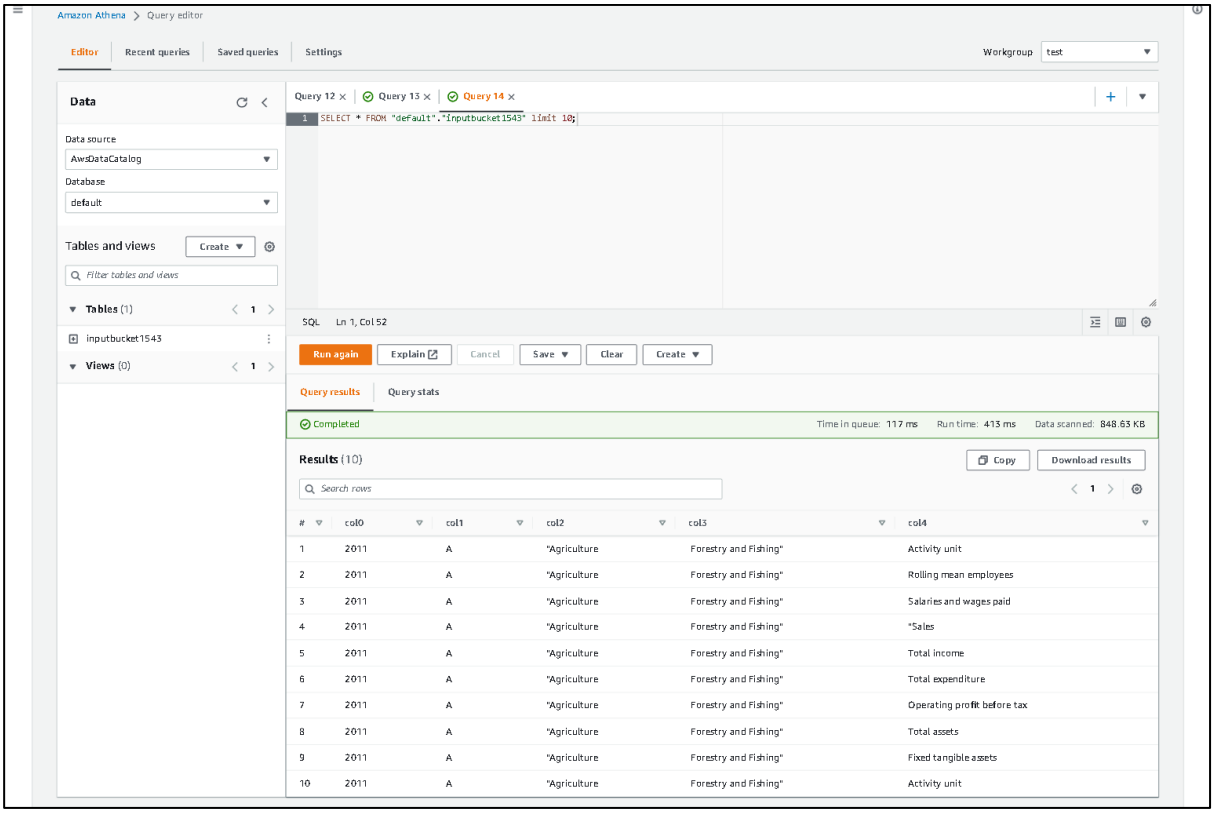


14. We see the table created in aws Athena, click on option button and preview table.



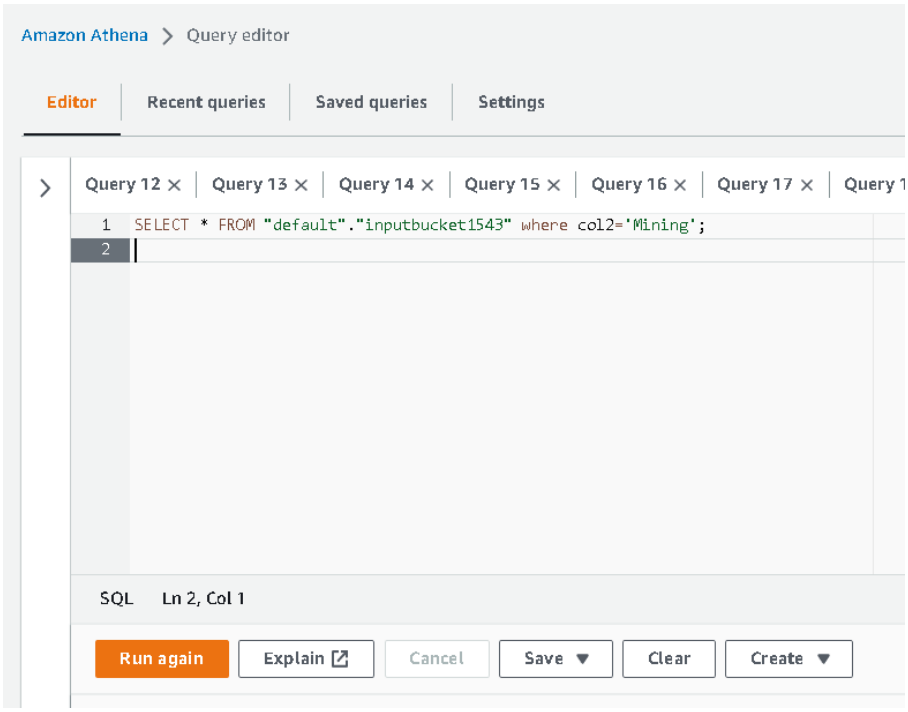
15. We get the data with the columns that were specified in the CSV file in S3.

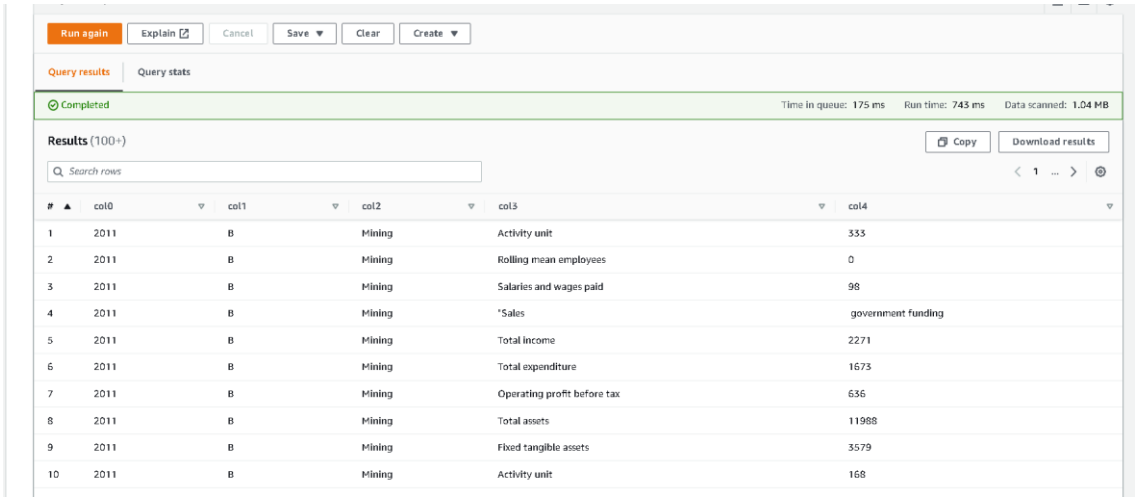




#	col0	col1	col2	col3	col4
1	2011	A	"Agriculture	Forestry and Fishing"	Activity unit
2	2011	A	"Agriculture	Forestry and Fishing"	Rolling mean employees
3	2011	A	"Agriculture	Forestry and Fishing"	Salaries and wages paid
4	2011	A	"Agriculture	Forestry and Fishing"	"Sales
5	2011	A	"Agriculture	Forestry 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.

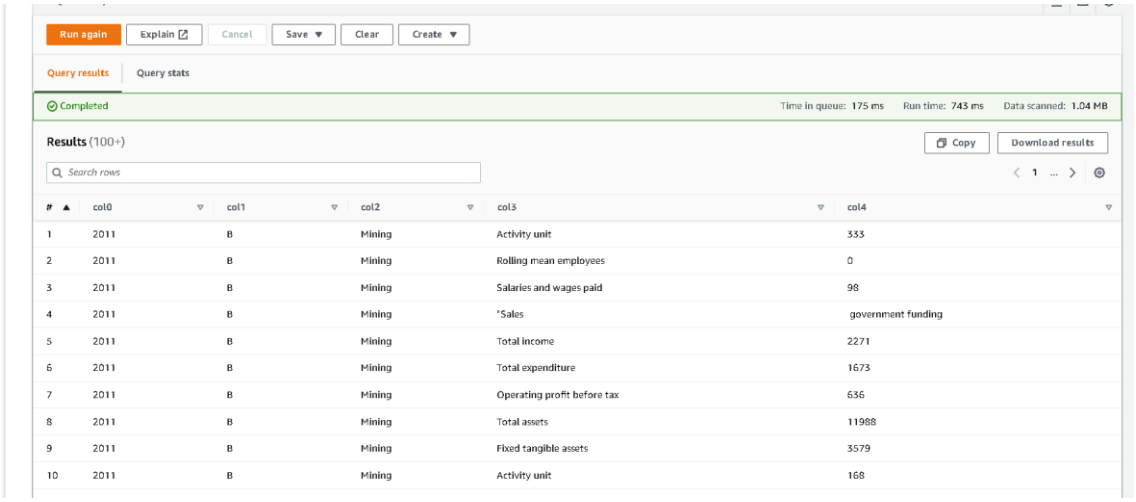




#	col0	col1	col2	col3	col4
1	2011	B	Mining	Activity unit	333
2	2011	B	Mining	Rolling mean employees	0
3	2011	B	Mining	Salaries and wages paid	98
4	2011	B	Mining	'Sales	government funding
5	2011	B	Mining	Total income	2271
6	2011	B	Mining	Total expenditure	1673
7	2011	B	Mining	Operating profit before tax	636
8	2011	B	Mining	Total assets	11988
9	2011	B	Mining	Fixed tangible assets	3579
10	2011	B	Mining	Activity unit	168

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.



#	col0	col1	col2	col3	col4
1	2011	B	Mining	Activity unit	333
2	2011	B	Mining	Rolling mean employees	0
3	2011	B	Mining	Salaries and wages paid	98
4	2011	B	Mining	*Sales	government funding
5	2011	B	Mining	Total income	2271
6	2011	B	Mining	Total expenditure	1673
7	2011	B	Mining	Operating profit before tax	636
8	2011	B	Mining	Total assets	11988
9	2011	B	Mining	Fixed tangible assets	3579
10	2011	B	Mining	Activity unit	168