

## Experiment : 6

Title : Querying Data in S3 with Amazon Athena

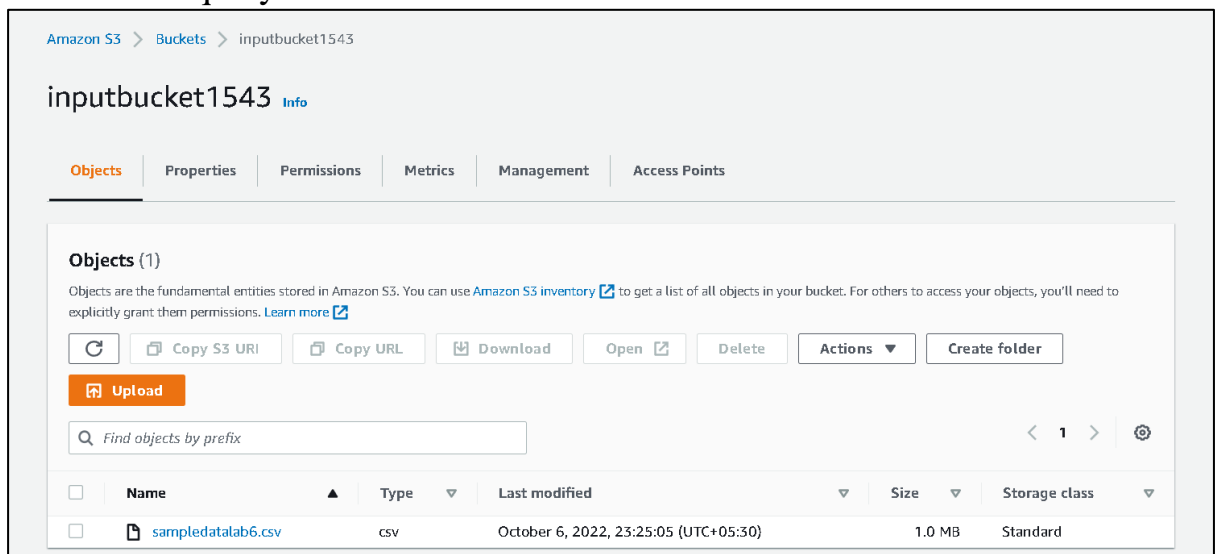
**Aim** : AWS Athena to query JSON/CSV files located in an s3 bucket

**Pre-requisites** : AWS Console, Amazon S3, Amazon Athena, Amazon Crawler, sampledatab6.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.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	2011.A	Agriculture, Forestry and Fishing	Activity unit		46134										
2	2011.A	Agriculture, Forestry and Fishing	Rolling mean employees		0										
3	2011.A	Agriculture, Forestry and Fishing	Salaries and wages paid		279										
4	2011.A	Agriculture, Forestry and Fishing	Sales, government funding, grants and subsidies		61877										
5	2011.A	Agriculture, Forestry and Fishing	Total income		8866										
6	2011.A	Agriculture, Forestry and Fishing	Total expenditure		7618										
7	2011.A	Agriculture, Forestry and Fishing	Operating profit before tax		770										
8	2011.A	Agriculture, Forestry and Fishing	Total assets		59700										
9	2011.A	Agriculture, Forestry and Fishing	Fixed tangible assets		32155										
10	2011.A	Agriculture, Forestry and Fishing	Activity unit		21777										
11	2011.A	Agriculture, Forestry and Fishing	Rolling mean employees		88136										
12	2011.A	Agriculture, Forestry and Fishing	Salaries and wages paid		1495										
13	2011.A	Agriculture, Forestry and Fishing	Sales, government funding, grants and subsidies		13959										
14	2011.A	Agriculture, Forestry and Fishing	Total income		13771										
15	2011.A	Agriculture, Forestry and Fishing	Total expenditure		12314										
16	2011.A	Agriculture, Forestry and Fishing	Operating profit before tax		1247										
17	2011.A	Agriculture, Forestry and Fishing	Total assets		52646										
18	2011.A	Agriculture, Forestry and Fishing	Fixed tangible assets		31235										
19	2011.A	Agriculture, Forestry and Fishing	Activity unit		1965										
20	2011.A	Agriculture, Forestry and Fishing	Rolling mean employees		13840										
21	2011.A	Agriculture, Forestry and Fishing	Salaries and wages paid		467										
22	2011.A	Agriculture, Forestry and Fishing	Sales, government funding, grants and subsidies		3060										
23	2011.A	Agriculture, Forestry and Fishing	Total income		3114										
24	2011.A	Agriculture, Forestry and Fishing	Total expenditure		2995										
25	2011.A	Agriculture, Forestry and Fishing	Operating profit before tax		223										
26	2011.A	Agriculture, Forestry and Fishing	Total assets		9923										
27	2011.A	Agriculture, Forestry and Fishing	Fixed tangible assets		5482										
28	2011.A	Agriculture, Forestry and Fishing	Activity unit		1140										
29	2011.A	Agriculture, Forestry and Fishing	Rolling mean employees		14850										
30	2011.A	Agriculture, Forestry and Fishing	Salaries and wages paid		477										
31	2011.A	Agriculture, Forestry and Fishing	Sales, government funding, grants and subsidies		2396										
32	2011.A	Agriculture, Forestry and Fishing	Total income		2438										
33	2011.A	Agriculture, Forestry and Fishing	Total expenditure		2344										
34	2011.A	Agriculture, Forestry and Fishing	Operating profit before tax		126										
35	2011.A	Agriculture, Forestry and Fishing	Total assets		6524										
36	2011.A	Agriculture, Forestry and Fishing	Fixed tangible assets		3649										
37	2011.A	Agriculture, Forestry and Fishing	Activity unit		480										
38	2011.A	Agriculture, Forestry and Fishing	Rolling mean employees		13983										
39	2011.A	Agriculture, Forestry and Fishing	Salaries and wages paid		410										

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.

Amazon Athena > Query editor

Editor | Recent queries | Saved queries | **Settings**

Workgroup: test

**Query result and encryption settings** [Manage]

**Query result location and encryption**

Query result location	Encrypt query results	Expected bucket owner	Assign bucket owner full control over query results
s3://aws3output1543/	Turned off		Turned off

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.

Add data source

Data source

Choose the source of data to be crawled.

S3

Network connection - optional

Optionally include a Network connection to use with this S3 target. Note that each crawler is limited to one Network connection so any other S3 targets will also use the same connection (or none, if left blank).

Clear selection
Add new connection

Location of S3 data

In this account
In a different account

S3 path

Browse for or enter an existing S3 path.

s3://bucket/prefix/object
View
Browse

All folders and files contained in the S3 path are crawled. For example, type s3://MyBucket/MyFolder/ to crawl all objects in MyFolder within MyBucket.

Subsequent crawler runs

This field is a global field that affects all S3 data sources.

Crawl all sub-folders
Crawl all folders again with every subsequent crawl.
Crawl new sub-folders only
Only Amazon S3 folders that were added since the last crawl will be crawled. If the schemas are compatible, new partitions will be added to existing tables.
Crawl based on events
Rely on Amazon S3 events to control what folders to crawl.

Sample only a subset of files

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.

Crawler successfully starting
The following crawler is now starting: "aws-athena-demo"

Crawlers

A crawler connects to a data store, progresses through a prioritized list of classifiers to determine the schema for your data, and then creates metadata tables in your data catalog.

Last updated: October 6, 2022 at 18:21:12 (UTC)

Crawlers (1/1)

View and manage all available crawlers.

Filter crawlers

☒

aws-athena...

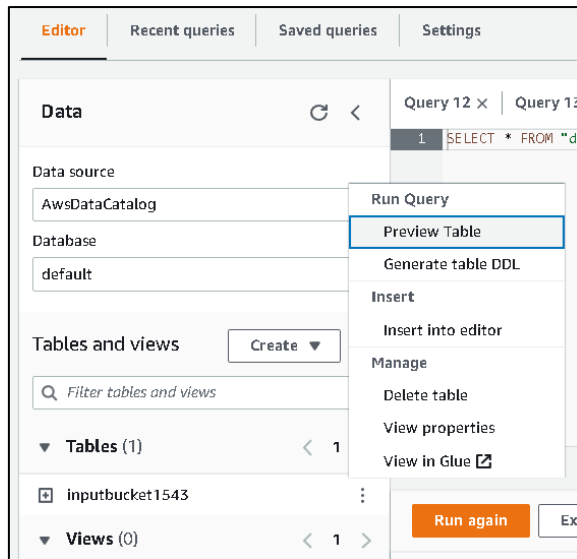
Stopping...

Succeeded...

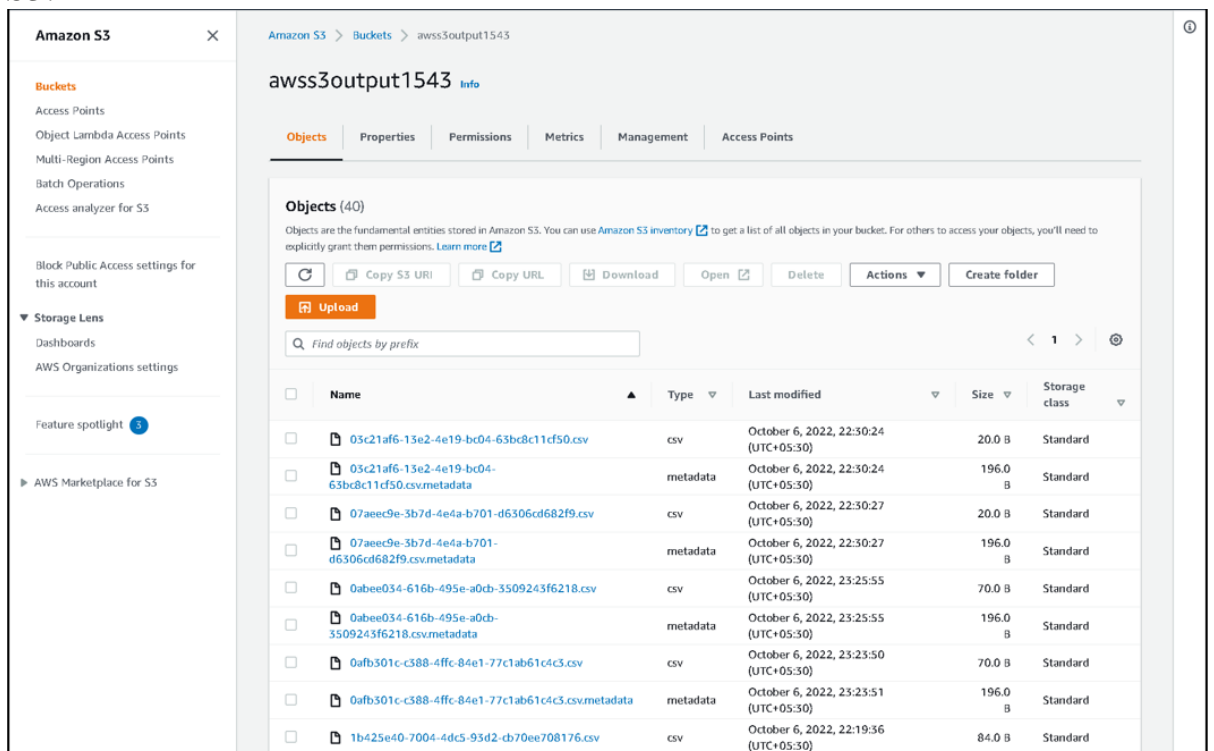
View log

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

Jayanth Sai Yarlagadda  
RA2011028010054



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 Filter tables and views

Tables (1): Inputbucket1543 Views (0)

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)

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 x Query 13 x Query 14 x Query 15 x Query 16 x Query 17 x Query

1 SELECT \* FROM "default"."inputbucket1543" where col2='Mining';

2

SQL Ln 2, Col 1

Run again Explain Cancel Save Clear Create

Run again

Explain

Cancel

Save

Clear

Create

Query results

Query stats

Completed

Time in queue: 175 ms

Run time: 743 ms

Data scanned: 1.04 MB

Results (100+)

Copy

Download results

Search rows

< 1 ... > @

#	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.