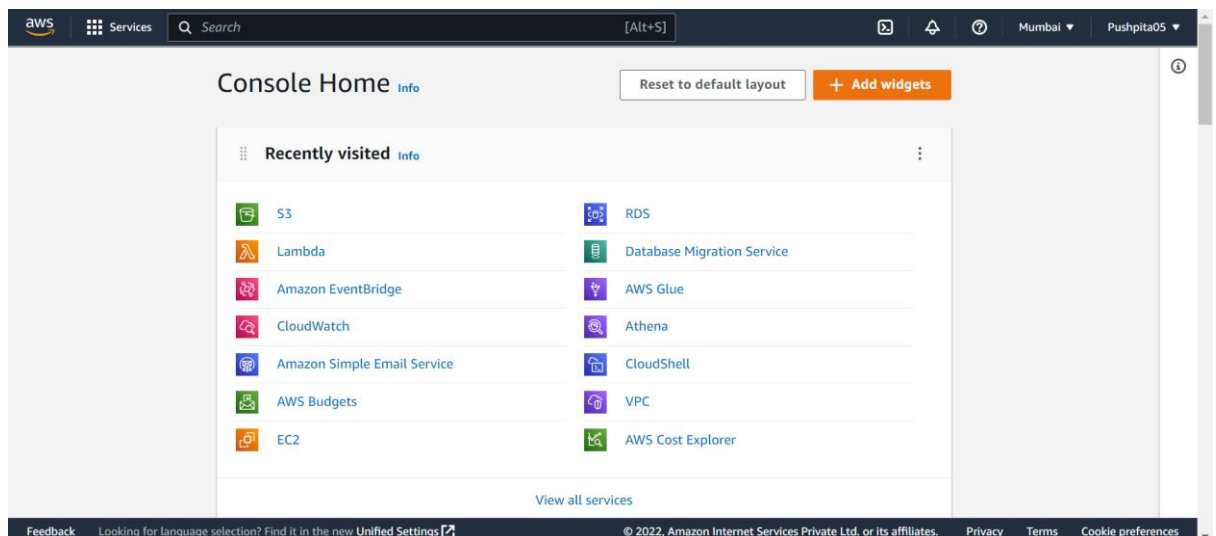


Experiment 6 - Querying Data in S3 with Amazon Athena

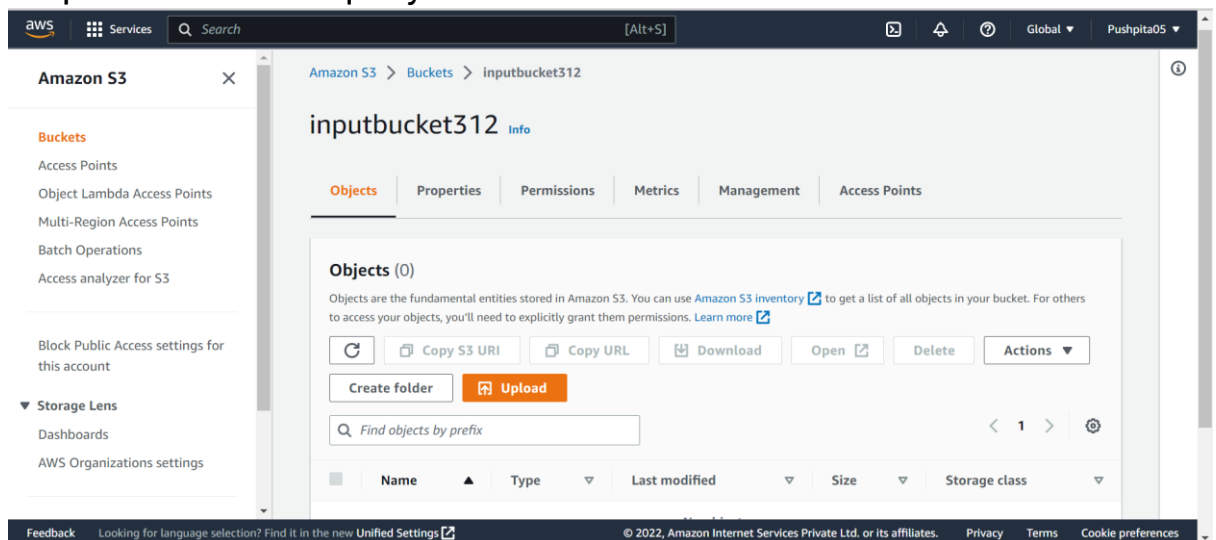
Aim: AWS Athena to query JSON/CSV files located in an s3 bucket

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.



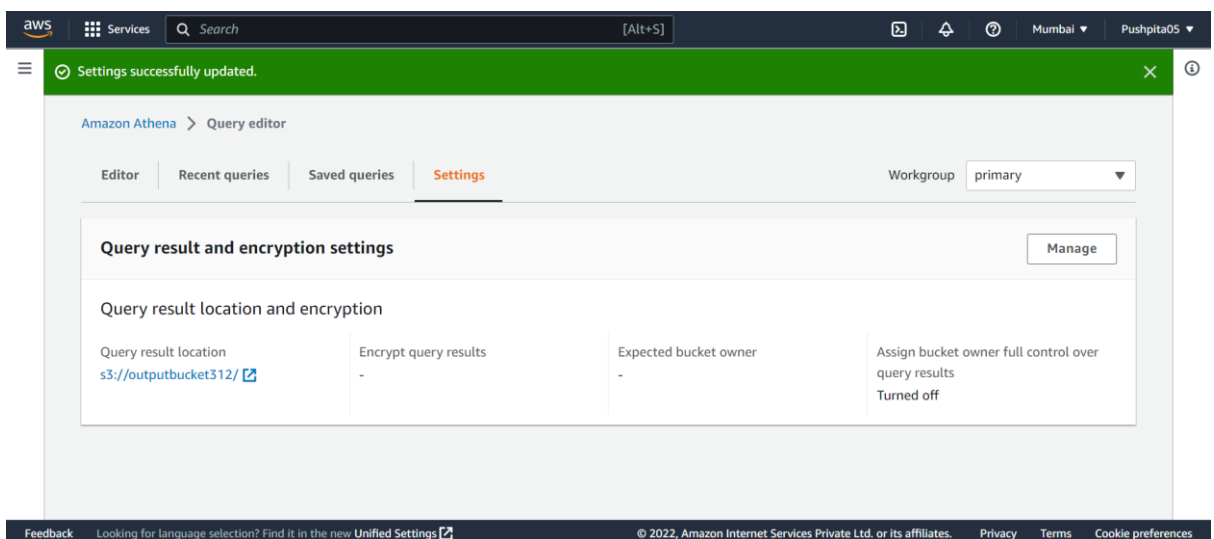
3. Upload a sample dataset (json, csv, tsv, etc) file in your AWS S3 bucket.

This document is read-only. Sign in to create, edit, and share documents. [Sign in](#)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	user_id	cellphone_rating															
2	0	30	1														
3	0	5	3														
4	0	10	9														
5	0	9	3														
6	0	23	2														
7	0	8	2														
8	0	22	1														
9	0	16	2														
10	0	19	1														
11	0	3	10														
12	1	7	8														
13	1	31	7														
14	1	18	5														
15	1	3	10														
16	1	32	6														
17	1	28	8														
18	1	16	7														
19	1	15	8														
20	1	4	7														
21	1	8	8														
22	6	13	5														
23	6	31	6														
24	6	23	8														

cellphones ratings

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.

AWS Glue

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.

[Add crawler](#) [Run crawler](#) [Action](#) [User preferences](#)

	Name	Schedule	Status	Logs	Last runtime	Median runtime	Tables updated	Tables added
You don't have any crawlers yet. Add crawler								

[Feedback](#) [Looking for language selection? Find it in the new Unified Settings](#) © 2022, Amazon Internet Services Private Ltd. or its affiliates. [Privacy](#) [Terms](#) [Cookie preferences](#)

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

Add crawler

☒ Crawler info
mycrawler

☒ Crawler source type
Data stores

☐ Data store

☐ IAM Role

☐ Schedule

☐ Output

☐ Review all steps

Crawl data in

☒ Specified path in my account

☐ Specified path in another account

Include path

s3://inputbucket312

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

Sample size (optional)

Enter a number between 1 and 249

This field sets the number of files in each leaf folder to be crawled. If not set, all the files are crawled.

▸ Exclude patterns (optional)

[Back](#) [Next](#)

[Feedback](#) [Looking for language selection? Find it in the new Unified Settings](#) © 2022, Amazon Internet Services Private Ltd. or its affiliates. [Privacy](#) [Terms](#) [Cookie preferences](#)

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 new sub-folders only
Crawl based on events

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)

AWS
Services
Search
[Alt+S]

Mumbai
Pushpita05

AWS Glue

Data Catalog
Databases
Tables
Stream schema registries
Schemas
Connections
Crawlers
Classifiers
Catalog settings
Data Integration and ETL
AWS Glue Studio
Jobs
Interactive Sessions
Notebooks

Add crawler

Crawler info
mycrawler

Crawler source type
Data stores
S3: s3://inputbucket...

IAM Role
arn:aws:iam::436748659018:role/service-role/AWSGlueServiceRole-iamcrawler1

Schedule
Run on demand

Output
mydatabase

Review all steps

Schedule
Schedule
Run on demand

Output

Database
mydatabase

Prefix added to tables (optional)

Table threshold (optional)

Create a single schema for each S3 path
false

Table level (optional)

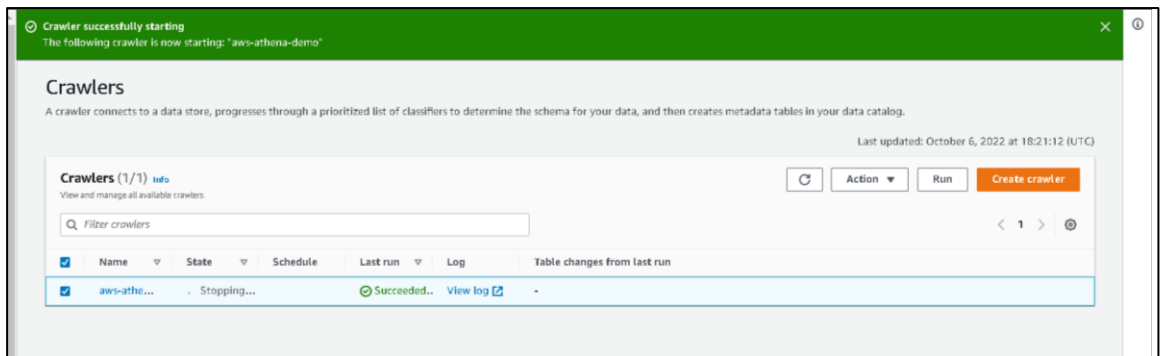
Configuration options

Back
Finish

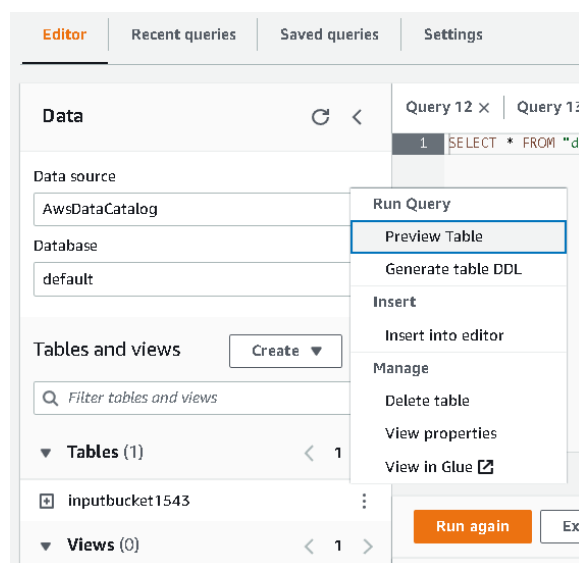
Feedback
Looking for language selection? Find it in the new Unified Settings

© 2022, Amazon Internet Services Private Ltd. or its affiliates.
Privacy
Terms
Cookie preferences

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.

Services

Search for services, features, blogs, docs, and more

[Alt+S]

Amazon S3

Buckets

Access Points

Object Lambda Access Points

Multi-Region Access Points

Batch Operations

Access analyzer for S3

Block Public Access settings for this account

Storage Lens

Cashboards

AWS Organizations settings

Feature spotlight

AWS Marketplace for S3

Amazon S3 > Buckets > awss3output1543

awss3output1543

Info

Objects

Properties

Permissions

Metrics

Management

Access Points

Objects (40)

Objects are the fundamental entities stored in Amazon S3. You can use [Amazon S3 inventory](#) to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. [Learn more](#)

Copy S3 URI

Copy URL

Download

Open

Delete

Actions

Create folder

Upload

Find objects by prefix

	Name	Type	Last modified	Size	Storage class
	03c21af6-13a2-4e19-b0d4-63b08c11cf50.csv	csv	October 6, 2022, 22:30:24 (UTC+05:30)	20.0 B	Standard
	03c21af6-13a2-4e19-b0d4-63b08c11cf50.csv.metadata	metadata	October 6, 2022, 22:30:24 (UTC+05:30)	196.0 B	Standard
	07aee09e-3b7d-4e4a-b701-d6306cd682f9.csv	csv	October 6, 2022, 22:30:27 (UTC+05:30)	20.0 B	Standard
	07aee09e-3b7d-4e4a-b701-d6306cd682f9.csv.metadata	metadata	October 6, 2022, 22:30:27 (UTC+05:30)	196.0 B	Standard
	0abee03c-616b-495e-a0db-3509243f6218.csv	csv	October 6, 2022, 23:25:55 (UTC+05:30)	70.0 B	Standard
	0abee03c-616b-495e-a0db-3509243f6218.csv.metadata	metadata	October 6, 2022, 23:25:55 (UTC+05:30)	196.0 B	Standard
	0afb301c-c388-4ffc-84e1-77c1ab61c4c3.csv	csv	October 6, 2022, 23:23:50 (UTC+05:30)	70.0 B	Standard
	0afb301c-c388-4ffc-84e1-77c1ab61c4c3.csv.metadata	metadata	October 6, 2022, 23:23:51 (UTC+05:30)	196.0 B	Standard
	1b125e40-7004-4d5c-93d2-cb70e6708175.csv	csv	October 6, 2022, 22:19:35 (UTC+05:30)	87.0 B	Standard

Amazon Athena > Query editor

Editor

Recent queries

Saved queries

Settings

Workgroup: test

Data

Data source: AwsDataCatalog

Database: default

Tables and views

Filter tables and views

Tables (1)

inputbucket1543

Views (0)

Query 12 x

Query 13 x

Query 14 x

1 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)

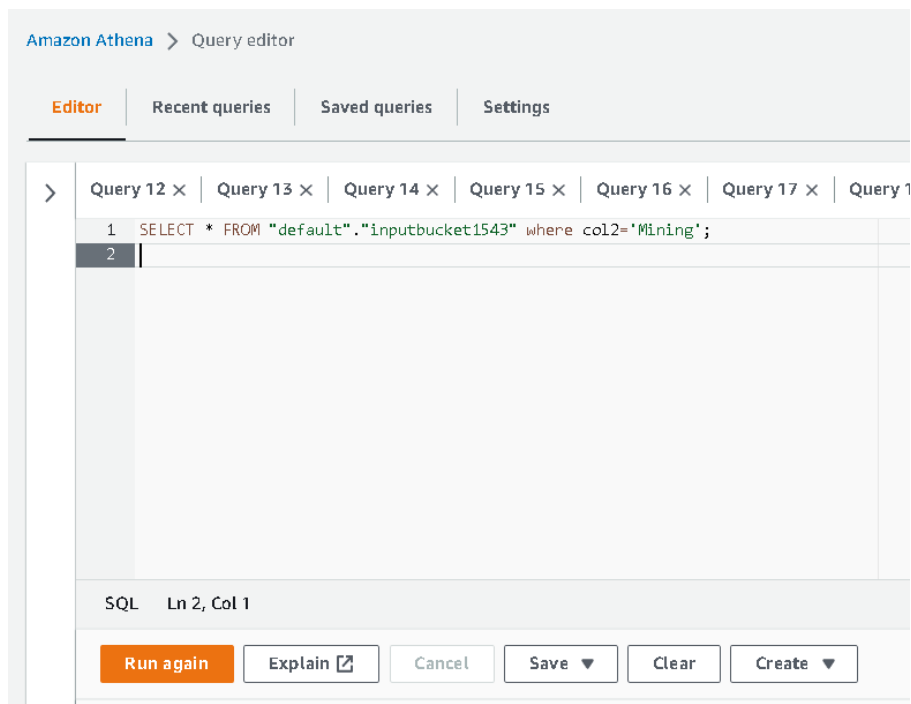
Copy

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.



The screenshot shows the Amazon Athena Query Results interface. At the top, there's a status bar with buttons for 'Run again', 'Explain', 'Cancel', 'Save', 'Clear', and 'Create'. Below it, there's a 'Query results' tab and a 'Query stats' tab. The 'Query results' tab is active, showing a green status bar with 'Completed' and performance metrics: 'Time in queue: 175 ms', 'Run time: 743 ms', and 'Data scanned: 1.04 MB'. Below the status bar, there's a 'Results (100+)' section with a search bar and a table of results. The table has columns for row number, col0, col1, col2, col3, and col4. The results are as follows:

#	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.

Run againExplainCancelSaveClearCreate

Query resultsQuery stats

CompletedTime in queue: 175 msRun time: 743 msData scanned: 1.04 MB

Results (100+)

Search rows

CopyDownload results

< 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	