

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/>) and select S3.

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

Name	AWS Region	Access	Creation date
inputbucketakil	Asia Pacific (Mumbai) ap-south-1	Bucket and objects not public	November 24, 2022, 18:24:17 (UTC+05:30)
outputbucketakil	Asia Pacific (Mumbai) ap-south-1	Bucket and objects not public	November 24, 2022, 18:27:57 (UTC+05:30)

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

The screenshot shows a spreadsheet application window titled 'sample'. The table has a header row with columns 'name', 'marks', and 'Marks2'. The data rows are labeled 'a' through 'e' and contain the following values:

name	marks	Marks2
a	1	3
b	2	3
c	3	3
d	4	3
e	5	3

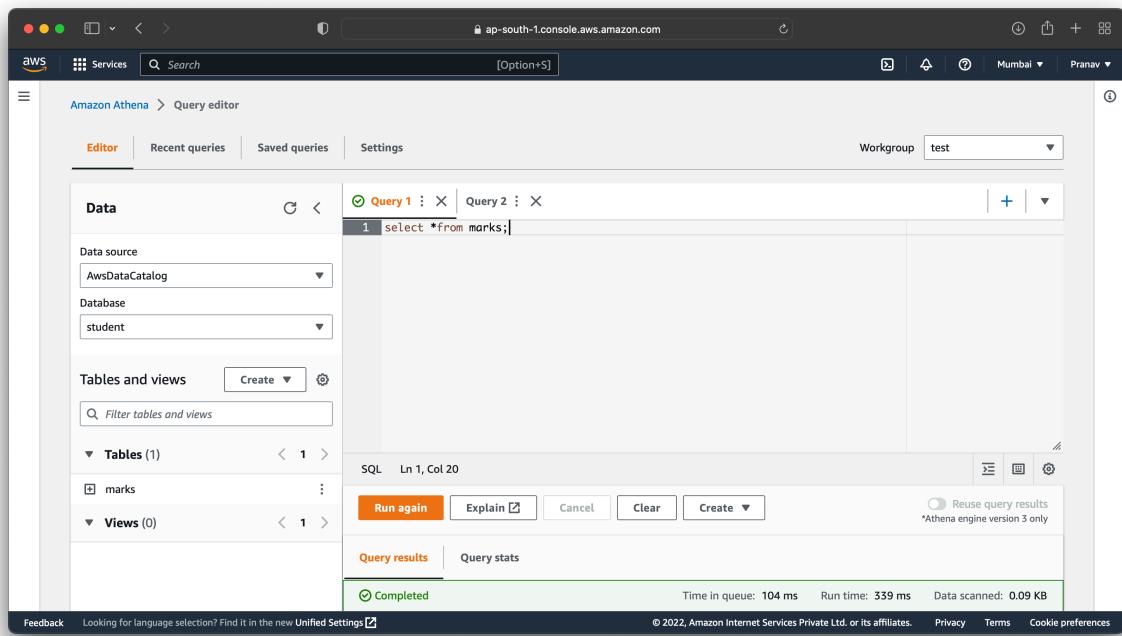
4. Go to AWS Athena.
5. Firstly, create a workgroup (workgroup is nothing but a kind of a container where our athena service stores the temporary data).

The screenshot shows the AWS Athena Workgroups page. The left sidebar includes 'Query editor', 'Workgroups' (which is selected), and 'Data sources'. The main area displays 'Workgroups (2) Info' with a note about using workgroups for separating users, teams, applications, workloads, and setting data limits. A 'Create workgroup' button is present. The table lists two workgroups:

Name	Description	Query engine ...	Query engine ...	Created on	Status
primary	-	Athena engine v...	Automatic	2022-11-14T23:...	Turned On
test	-	Athena engine v...	Automatic	2022-11-24T18:...	Current

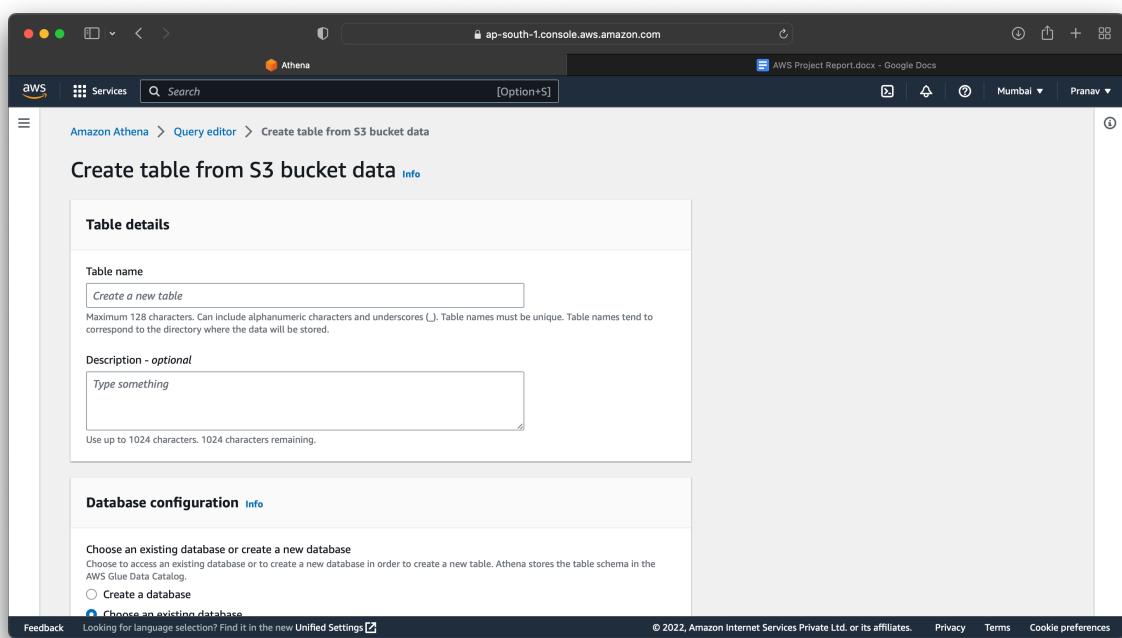
6. Give workgroup name, description, query result location, data usage limit.

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
 - Select S3 Bucket



The screenshot shows the AWS Athena Query Editor interface. At the top, there are tabs for 'Editor', 'Recent queries', 'Saved queries', and 'Settings'. A dropdown menu for 'Workgroup' is set to 'test'. The main area has two tabs: 'Query 1' (highlighted) and 'Query 2'. The SQL query in 'Query 1' is: 'select * from marks;'. On the left, the 'Data' pane shows the 'Data source' as 'AwsDataCatalog', 'Database' as 'student', and 'Tables and views' section with 'marks' selected. Below the tables is a 'Views (0)' section. The bottom of the editor shows the status bar with 'Completed' status, 'Time in queue: 104 ms', 'Run time: 339 ms', and 'Data scanned: 0.09 KB'.

8. Create a table name and the database.



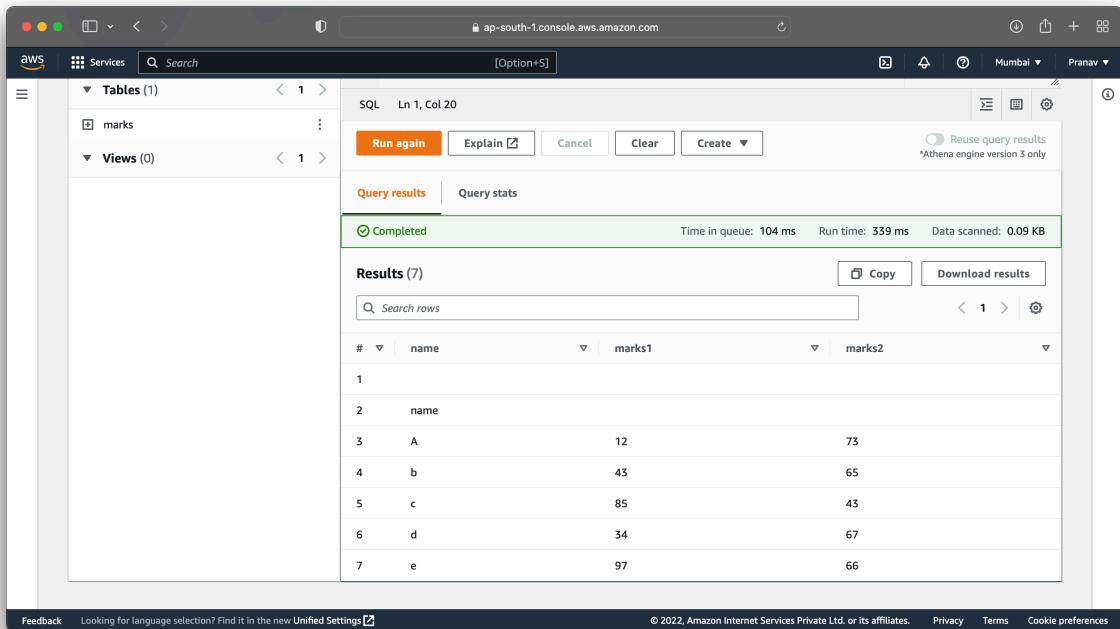
The screenshot shows the 'Create table from S3 bucket data' configuration page. It has two main sections: 'Table details' and 'Database configuration'. In the 'Table details' section, there is a 'Table name' field with the placeholder 'Create a new table'. Below it, a note says: 'Maximum 128 characters. Can include alphanumeric characters and underscores (_). Table names must be unique. Table names tend to correspond to the directory where the data will be stored.' In the 'Description - optional' field, there is a placeholder 'Type something'. A note below it says: 'Use up to 1024 characters. 1024 characters remaining.' In the 'Database configuration' section, there is a note: 'Choose an existing database or create a new database. Choose to access an existing database or to create a new database in order to create a new table. Athena stores the table schema in the AWS Glue Data Catalog.' There are two radio button options: 'Create a database' (selected) and 'Choose an existing database'.

9. Select the input and output source with respect to the respective buckets created earlier.

10. Now go to the Query Editor and select the created database.

The screenshot shows the AWS Lambda console interface. At the top, there's a navigation bar with tabs for 'Lambda' (selected), 'Functions', 'Logs', 'Metrics', and 'Actions'. Below the navigation bar, there's a search bar and a 'Create function' button. The main content area displays a table with columns: 'Name', 'Runtime', 'Last modified', 'Status', and 'Actions'. One row in the table is highlighted in blue, showing details for a function named 'HelloWorldFunction'. On the right side of the screen, there's a sidebar with options like 'AWS Lambda Metrics' and 'AWS Lambda Metrics Insights'.

11. Run the desired queries and verify the output.



The screenshot shows the AWS Athena Query Editor interface. On the left, there's a sidebar with 'Tables (1)' containing a single entry 'marks'. Below it is 'Views (0)'. The main area has a toolbar with 'Run again' (orange), 'Explain', 'Cancel', 'Clear', and 'Create'. To the right of the toolbar are buttons for 'Reuse query results' and a note 'Athena engine version 3 only'. Below the toolbar, tabs for 'Query results' and 'Query stats' are visible, with 'Query results' being active. A green bar indicates the query is 'Completed'. Below this, the 'Results (7)' section displays a table with 7 rows. The columns are '#', 'name', 'marks1', and 'marks2'. The data is as follows:

#	name	marks1	marks2
1			
2	name		
3	A	12	73
4	b	43	65
5	c	85	43
6	d	34	67
7	e	97	66

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.