AWS DATA ANALYSIS USING SALES DATA

INTRODUCTION:

In today’s digital landscape, data has become the lifeblood of organizations, driving critical decisions and providing valuable insights. However, the sheer volume and complexity of data can pose challenges when it comes to processing and deriving meaningful information from it. That’s where effective data processing and visualization techniques come into play.

Enter Amazon Web Services (AWS), a leading cloud computing platform that offers a comprehensive suite of services for data management, processing, and visualization. With AWS, organizations can harness the power of scalable and cost-effective solutions to unlock the true potential of their data.

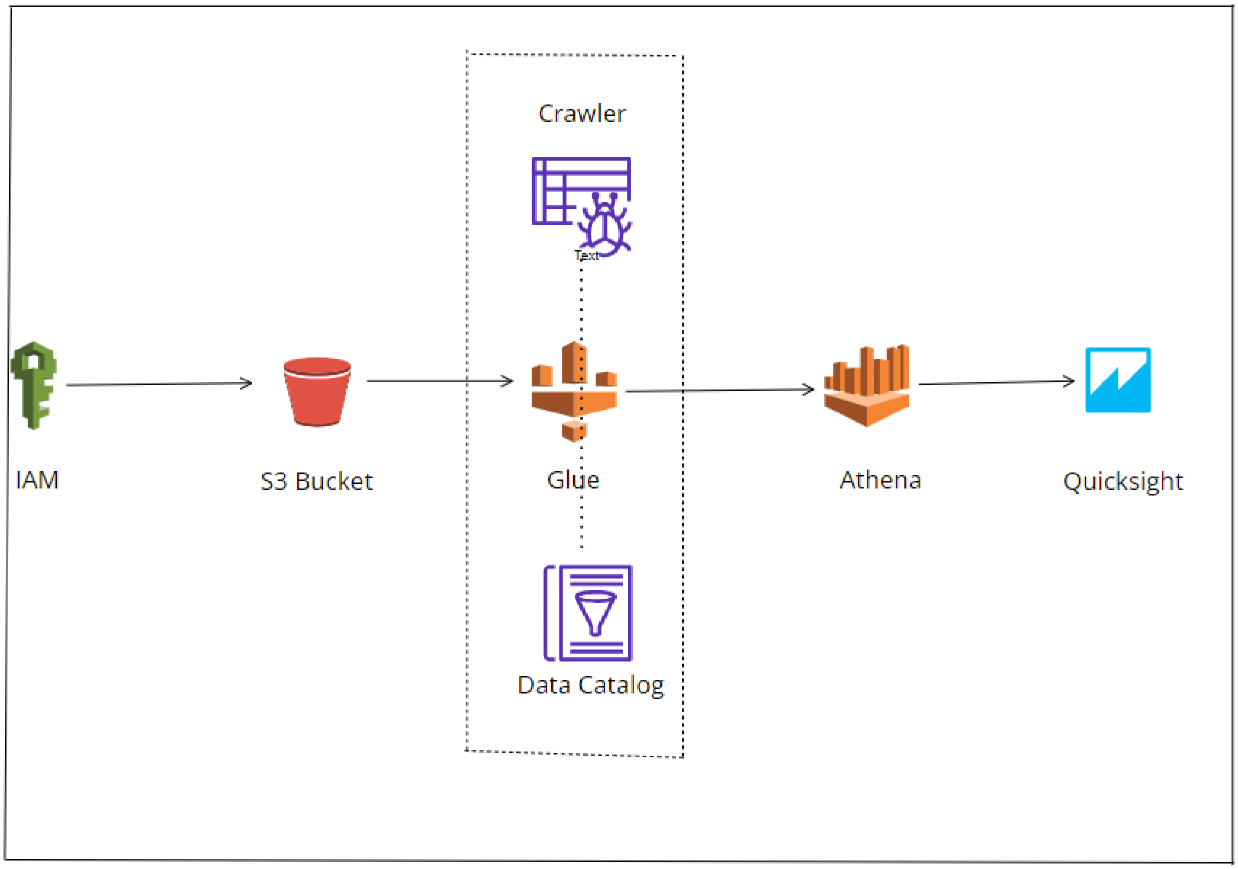
In this article, we will take you on a journey through the process of leveraging AWS services to transform raw data into actionable insights. Specifically, we will focus on the seamless integration of services like Amazon S3, AWS Glue, Amazon Athena, and Amazon Quick Sight to facilitate data processing and visualization.

By the end of this article, data engineers, analysts, and anyone interested in harnessing the power of AWS for data processing and visualization will have a solid understanding of how to upload a CSV file into Amazon S3, utilize AWS Glue to crawl and create a database, leverage Amazon Athena for flexible querying, and visualize the data as interactive graphs using Amazon Quick Sight.

The benefits of using AWS services for this process are manifold. Firstly, AWS provides scalable infrastructure, ensuring that your data processing and visualization needs can easily grow with your business. Secondly, it offers a cost-effective solution, as you only pay for the resources you use. Additionally, AWS services seamlessly integrate with each other, allowing for a streamlined and efficient workflow.

So whether you are a data engineer seeking to optimize data processing pipelines, an analyst looking for better ways to derive insights, or simply someone interested in exploring the capabilities of AWS for data management, this article is for you.
Now, let’s dive into the step-by-step process of analyzing airport data using various services on Aws such as s3, glue, Athena and quick sight.

ARCHITETURE OF AWS DATA ANALYSIS:



PREREQUISITES:

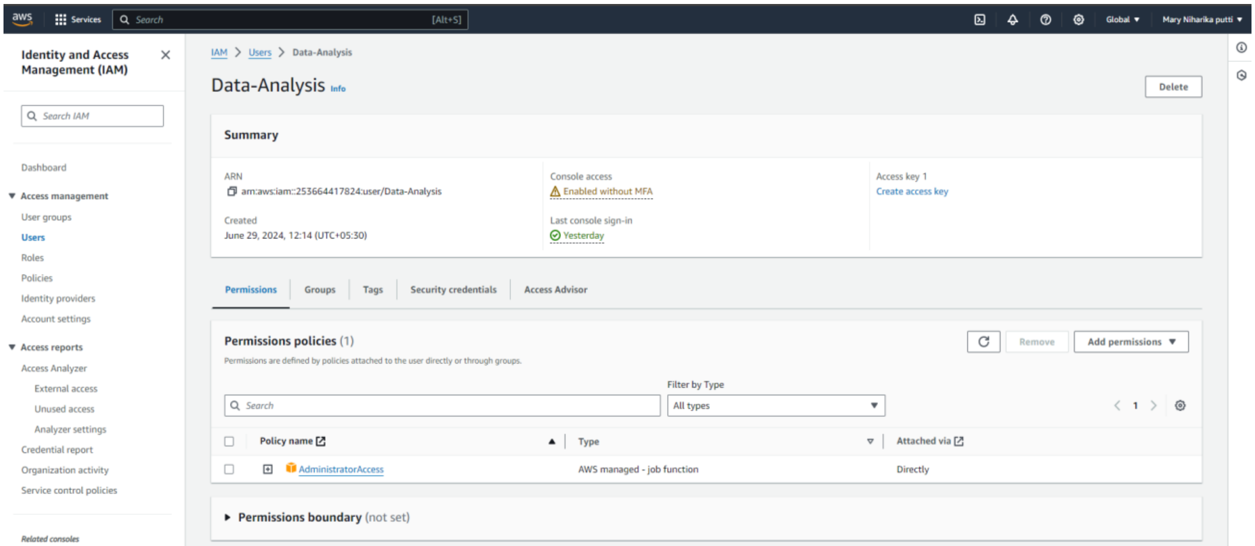
Before proceeding with the steps outlined in this article, ensure that you have the following prerequisites in place:

- Identity access management (IAM) is user is an entity created in aws that provides a way to interact with aws resources.
- Simple Storage Service (S3) is an online storage service where you can store and retrieve any type of data on the web regardless of the time and place.
- AWS Glue is a powerful service that allows you to discover, catalog, and transform data from various sources, making it easier to prepare data for analytics.
- Amazon Athena is an interactive serverless service that can be used to query and analyze raw data using standard SQL.
- Quick Sight is an AWS based Business Intelligence and visualization tool that is used to visualize data, perform ad hoc analysis and get business insights from our data.

CREATING IAM USER AND LOGIN:

As part of our AWS setup for data analysis, I created an IAM user named Data-analysis and granted it Administrator Access. This user has full administrative permissions, enabling it to manage and access all AWS resources. With the Data-analysis IAM user, I performed various data analysis tasks, ensuring seamless integration and control over AWS services. This setup allows for efficient and secure handling of data analysis operations, leveraging the comprehensive capabilities provided by AWS.

1.Created I AM User Data-Analysis:



Data-Analysis Info

Summary

ARN
arn:aws:iam::253664417824:user/Data-Analysis

Created
June 29, 2024, 12:14 (UTC+05:30)

Console access
Enabled without MFA

Last console sign-in
Yesterday

Access key 1
Create access key

Permissions policies (1)

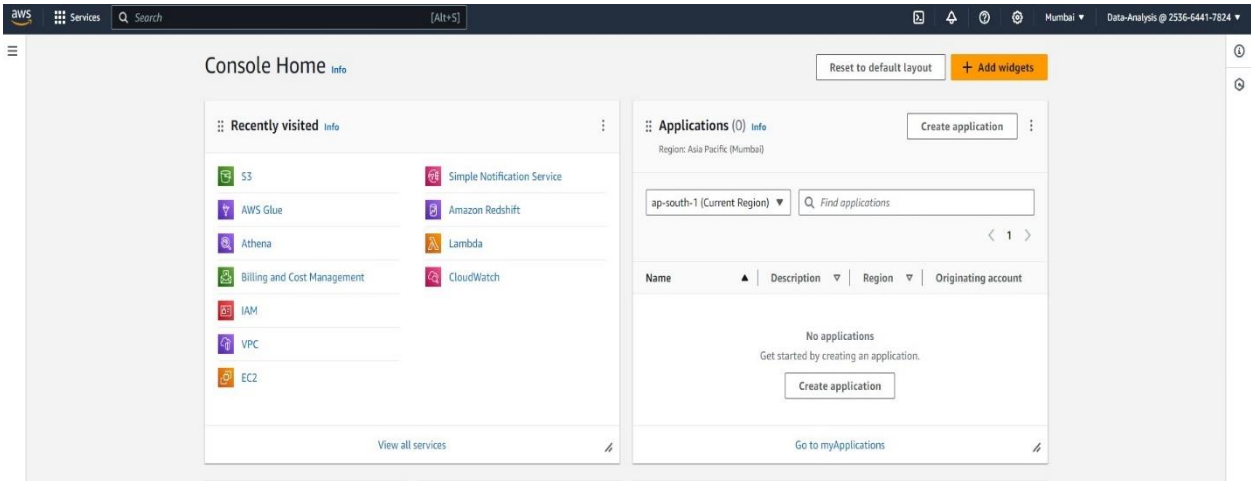
Permissions are defined by policies attached to the user directly or through groups.

Filter by Type
All types

	Policy name	Type	Attached via
□	AdministratorAccess	AWS managed - job function	Directly

Permissions boundary (not set)

2. Login as I AM User to perform Data Analysis:



Console Home Info

Recently visited Info

- S3
- AWS Glue
- Athena
- Billing and Cost Management
- IAM
- VPC
- EC2
- Simple Notification Service
- Amazon Redshift
- Lambda
- CloudWatch

Applications (0) Info

Region: Asia Pacific (Mumbai)

ap-south-1 (Current Region)

Find applications

Name	Description	Region	Originating account

No applications

Get started by creating an application.

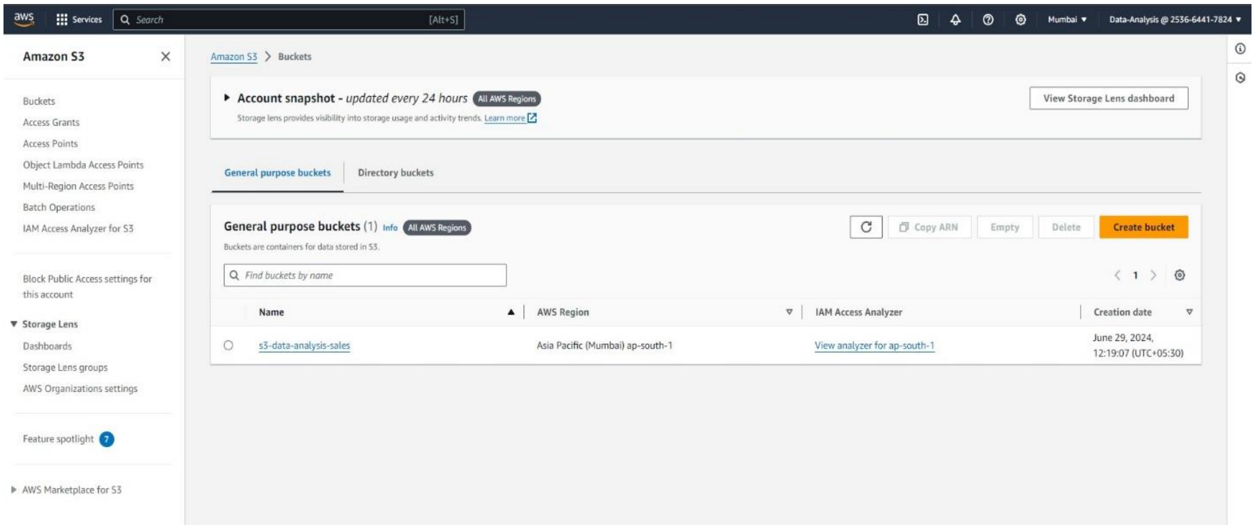
Create application

UPLOADING DATA TO AMAZON S3:

Amazon Simple Storage Service (S3) is a highly scalable and durable object storage service provided by AWS. It offers a secure and reliable solution for storing and retrieving data from anywhere on the web. Let's explore how you can upload a CSV file to S3 and ensure efficient data organization.

Uploading a CSV file to S3: You can upload a CSV file to S3 using various methods but for this project we are going to be working with AWS Management Console. We begin with Logging in to the AWS Management Console and navigate to the buckets section under s3. let's proceed to uploading our CSV file, but before that lets create a new bucket or we can use an existing bucket.

3. Created Bucket S3-data-analysis- sales:



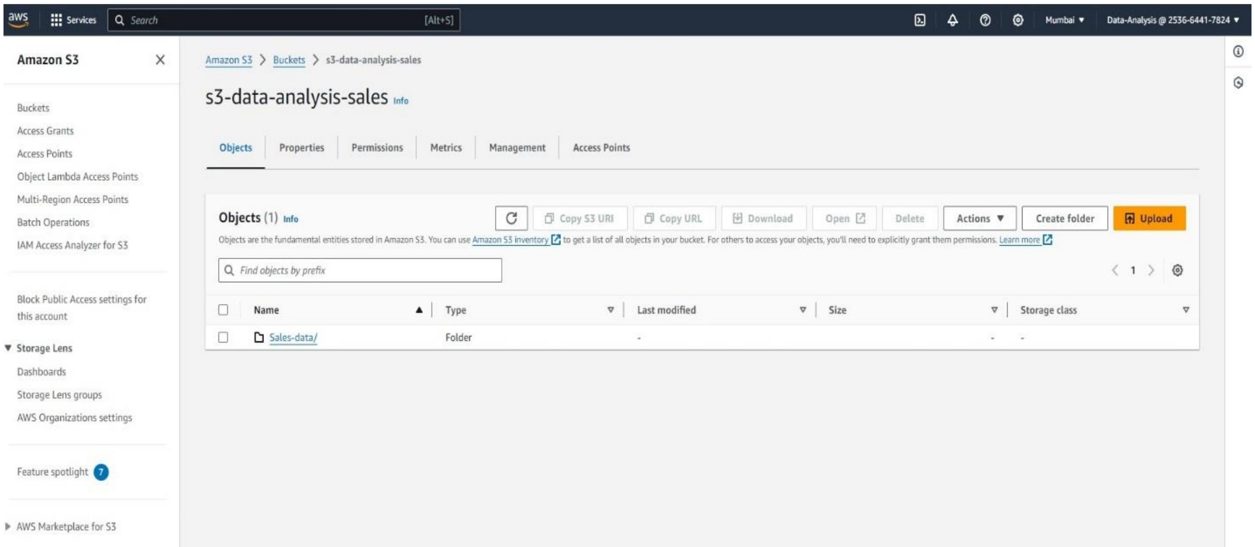
Storage lens provides visibility into storage usage and activity trends. Learn more

General purpose buckets (1) Info All AWS Regions

Buckets are containers for data stored in S3.

	Name	AWS Region	IAM Access Analyzer	Creation date
○	s3-data-analysis-sales	Asia Pacific (Mumbai) ap-south-1	View analyzer for ap-south-1	June 29, 2024, 12:19:07 (UTC+05:30)

4. Created folder as sales-data:

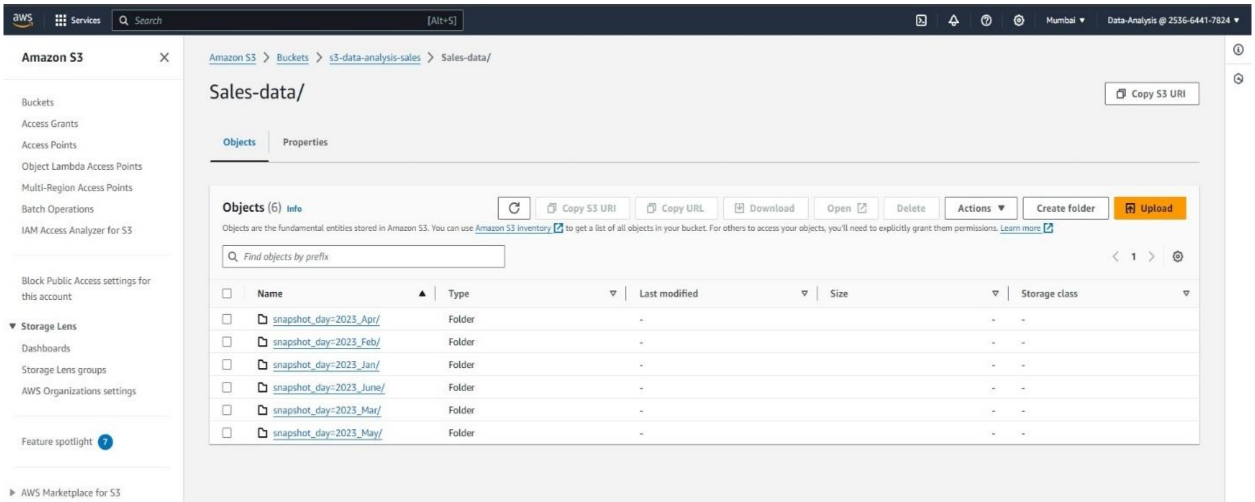


Objects (1) Info

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

□	Name	Type	Last modified	Size	Storage class
□	Sales-data/	Folder	-	-	-

5. Uploaded Objects Sales.CSV:



Sales-data/

Objects (6) Info

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

□	Name ▲	Type	Last modified	Size	Storage class
□	snapshot_day=2023_Apr/	Folder	-	-	-
□	snapshot_day=2023_Feb/	Folder	-	-	-
□	snapshot_day=2023_Jan/	Folder	-	-	-
□	snapshot_day=2023_June/	Folder	-	-	-
□	snapshot_day=2023_Mar/	Folder	-	-	-
□	snapshot_day=2023_May/	Folder	-	-	-

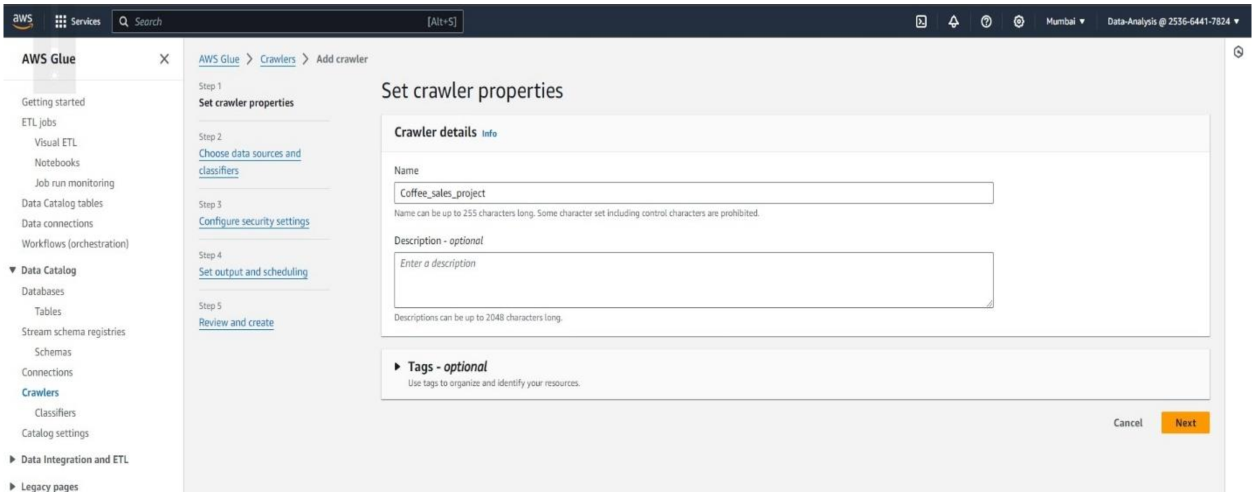
Here, I Have Created a snapshot_day=2033_jan to get partitions as per month. Then I have Uploaded my objects.csv file in to it.

DATA CRAWLING WITH AWS GLUE:

Data crawling is the process of automatically scanning and analyzing the data sources to infer the schema and structure of the data. It allows AWS Glue to understand the data format, columns, and data types without requiring manual intervention.

To create a crawler in AWS Glue and discover the uploaded CSV file, follow these steps: on the AWS Glue Console and navigate to the Crawlers section under databases, next we are going to Click on “Add crawler” to create a new crawler. We will be prompted on steps to create the crawler which involves Provide a name for the crawler, we using “Coffe_sales_project”, the next step involves choosing the data source and locations to crawl, Select the S3 bucket containing the uploaded CSV file. and proceed to the next step.

6. Creating Crawler:



Set crawler properties

Crawler details Info

Name

Coffee_sales_project

Name can be up to 255 characters long. Some character set including control characters are prohibited.

Description - *optional*

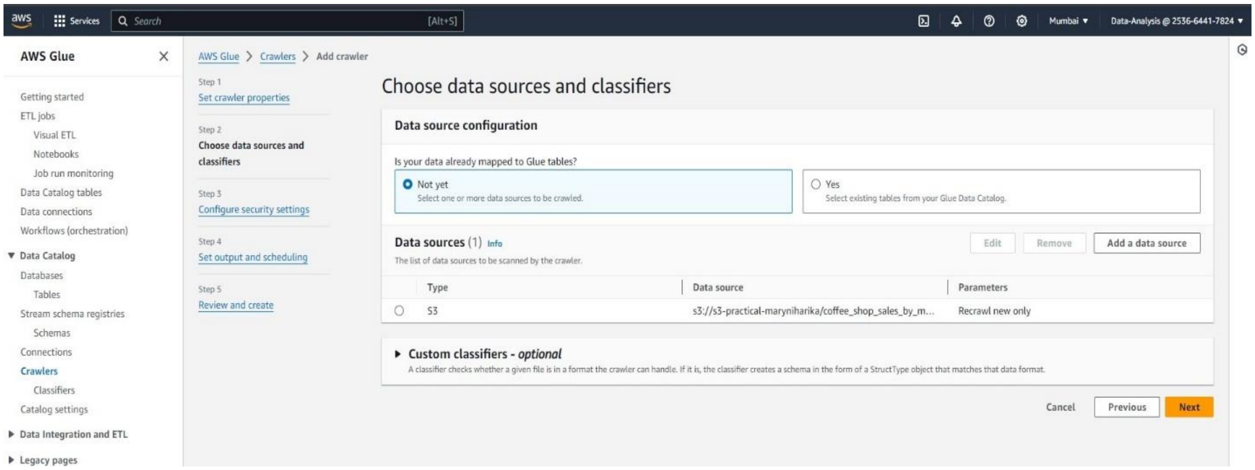
Enter a description

Descriptions can be up to 2048 characters long.

▸ Tags - *optional*

Use tags to organize and identify your resources.

7. Adding source to the crawler:



Choose data sources and classifiers

Data source configuration

Is your data already mapped to Glue tables?

● Not yet
Select one or more data sources to be crawled.

○ Yes
Select existing tables from your Glue Data Catalog.

Data sources (1) Info

The list of data sources to be scanned by the crawler.

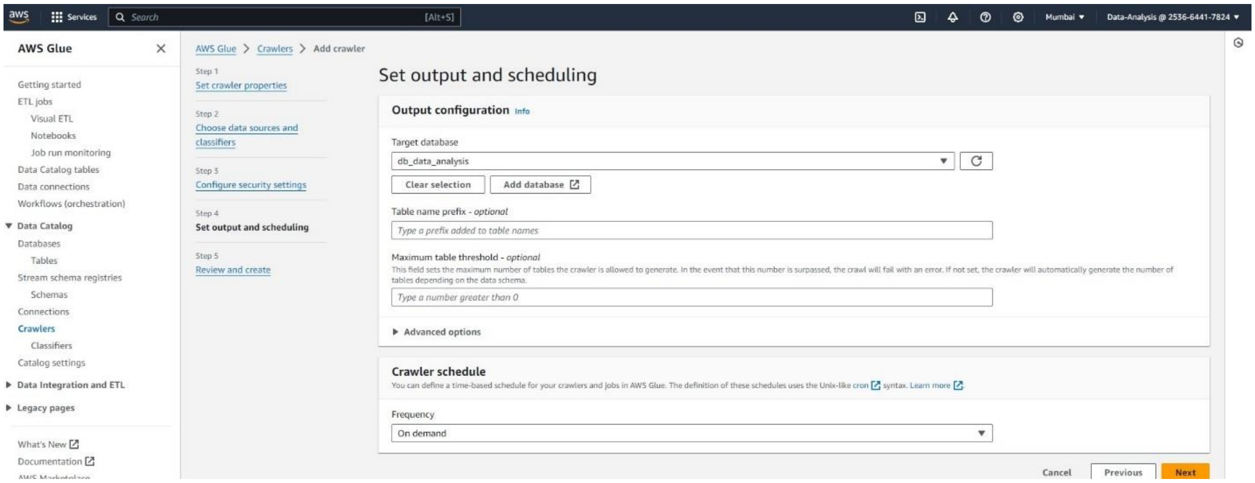
	Type	Data source	Parameters
○	S3	s3://s3-practical-maryniharika/coffee_shop_sales_by_m...	Recrawl new only

▸ Custom classifiers - *optional*

A classifier checks whether a given file is in a format the crawler can handle. If it is, the classifier creates a schema in the form of a StructType object that matches that data format.

Next, we want to create a database, open the AWS Glue Console and navigate to the Data Catalog section, select Databases and proceed to creating a new database, click on add database on the top right-hand side of the console, key in details of the database such as name, location(optional) and description(optional), then create database. After creating the database head back to Databases to check for the newly created database.

8. Creating Database add to the crawler:



Set output and scheduling

Output configuration Info

Target database: db_data_analysis

Table name prefix - *optional*

Type a prefix added to table names

Maximum table threshold - *optional*

This field sets the maximum number of tables the crawler is allowed to generate. In the event that this number is surpassed, the crawl will fail with an error. If not set, the crawler will automatically generate the number of tables depending on the data schema.

Type a number greater than 0

▸ Advanced options

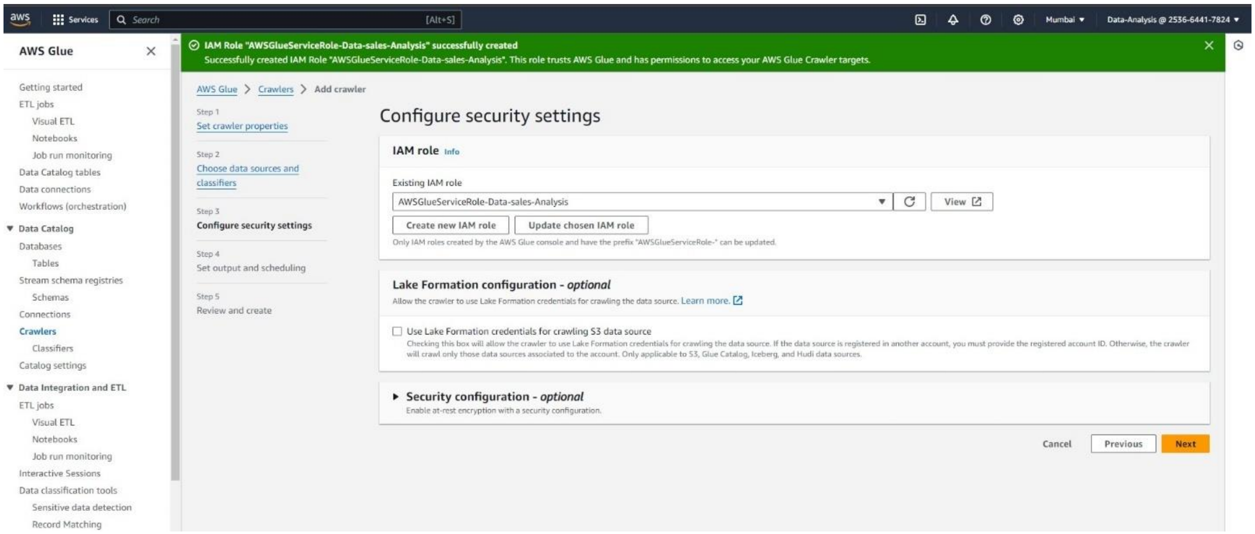
Crawler schedule

You can define a time-based schedule for your crawlers and jobs in AWS Glue. The definition of these schedules uses the Unix-like cron syntax. Learn more

Frequency: On demand

Configure the crawler’s settings, including IAM roles, database location, and frequency of running. e. Optionally, set up crawler-specific classifiers to handle special data formats or custom schema inference, then we Review the crawler configuration and click on “Finish” to create the crawler.

9. Creating I am Role:



IAM Role "AWSGlueServiceRole-Data-sales-Analysis" successfully created
Successfully created IAM Role "AWSGlueServiceRole-Data-sales-Analysis". This role trusts AWS Glue and has permissions to access your AWS Glue Crawler targets.

Configure security settings

IAM role

Existing IAM role: AWSGlueServiceRole-Data-sales-Analysis

Only IAM roles created by the AWS Glue console and have the prefix "AWSGlueServiceRole-" can be updated.

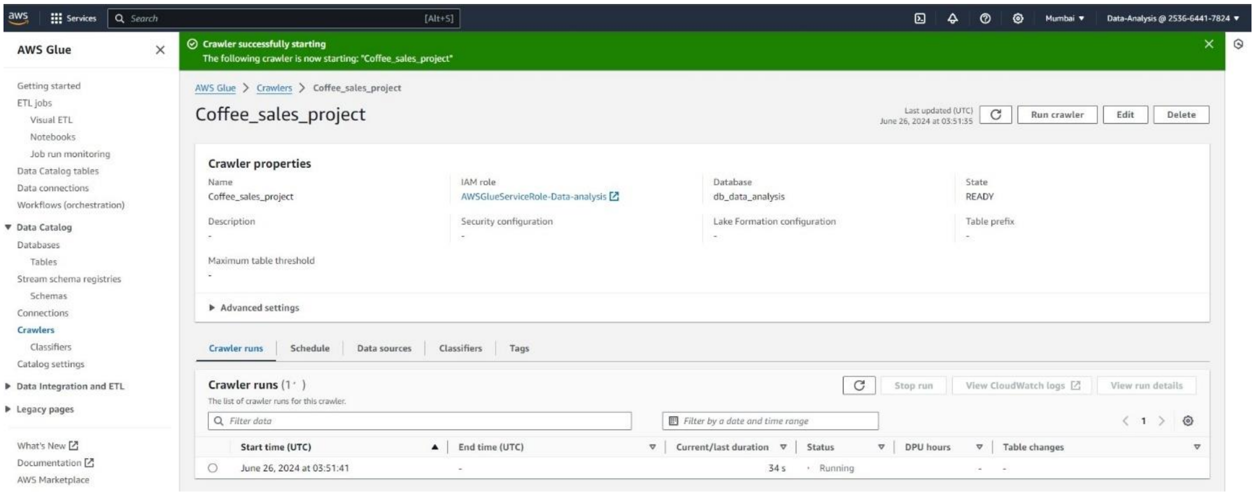
Lake Formation configuration - *optional*

Allow the crawler to use Lake Formation credentials for crawling the data source.

☐ Use Lake Formation credentials for crawling S3 data source
Checking this box will allow the crawler to use Lake Formation credentials for crawling the data source. If the data source is registered in another account, you must provide the registered account ID. Otherwise, the crawler will crawl only those data sources associated to the account. Only applicable to S3, Glue Catalog, Iceberg, and Hudi data sources.

▸ Security configuration - *optional*
Enable at-rest encryption with a security configuration.

10. Crawler Running:



Crawler successfully starting
The following crawler is now starting: "Coffee_sales_project"

Coffee_sales_project

Last updated (UTC) June 26, 2024 at 03:51:35

Crawler properties

Name: Coffee_sales_project
IAM role: AWSGlueServiceRole-Data-analysis
Database: db_data_analysis
State: READY
Description: -
Security configuration: -
Lake Formation configuration: -
Table prefix: -
Maximum table threshold: -

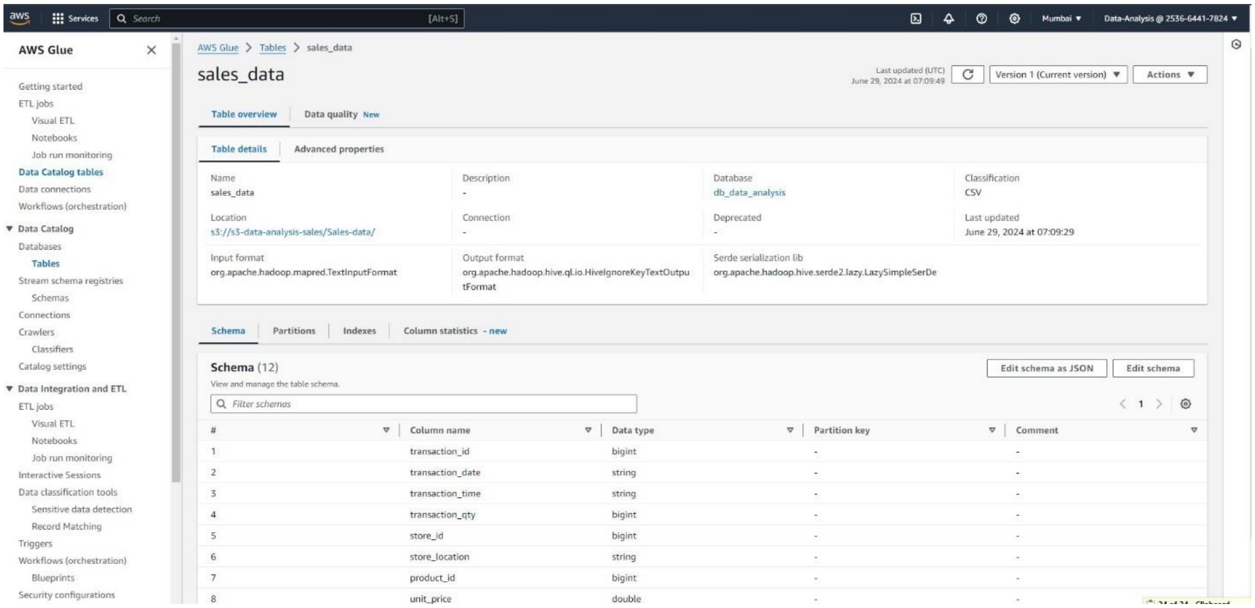
Crawler runs (1ˑ)
The list of crawler runs for this crawler.

	Start time (UTC)	End time (UTC)	Current/last duration	Status	DPU hours	Table changes
○	June 26, 2024 at 03:51:41	-	34 s	Running	-	-

Automatic Schema and Table Definition Generation: Once the crawler is created and executed, AWS Glue automatically generates schema and table definitions based on the crawled data. It examines the structure of the CSV file, infers the data types, and creates metadata entries in the AWS Glue Data Catalog. This process eliminates the need for manually defining the schema and table structures, saving time and ensuring accuracy.

The generated schema information, including column names, data types, and partition keys if applicable, can be further customized or enhanced in the AWS Glue Data Catalog as needed.

11. Schema Information in Aws Data Catalog:



sales_data

Last updated (UTC)
June 29, 2024 at 07:09:49

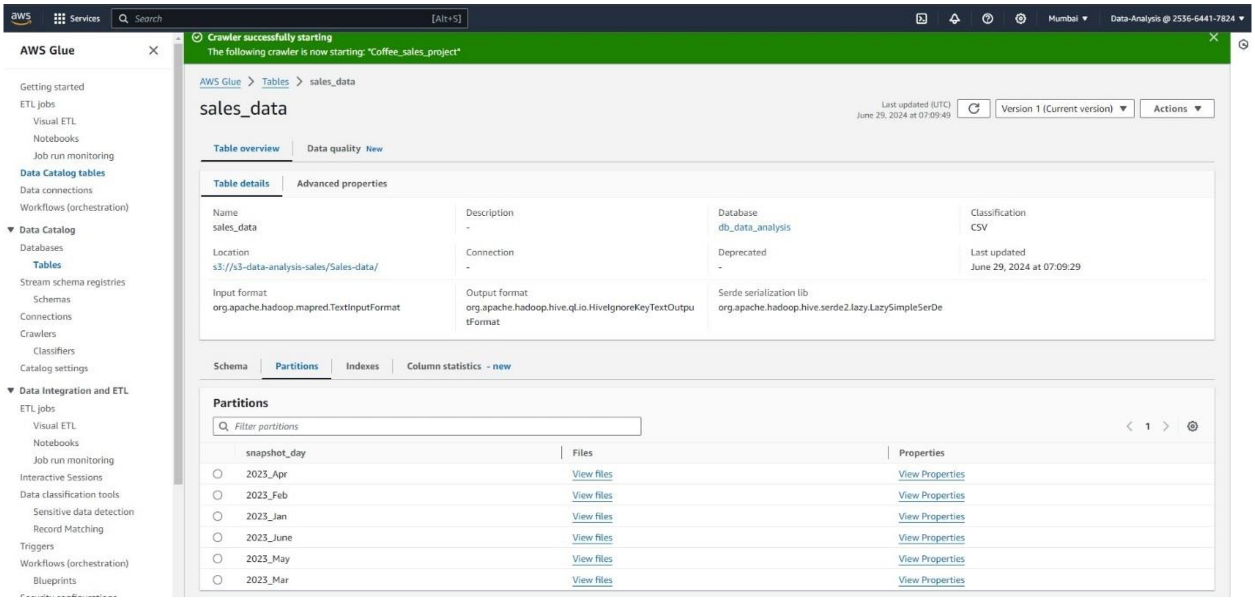
Name sales_data	Description -	Database db_data_analysis	Classification CSV
Location s3://s3-data-analysis-sales/Sales-data/	Connection -	Deprecated -	Last updated June 29, 2024 at 07:09:29
Input format org.apache.hadoop.mapred.TextInputFormat	Output format org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat	Serde serialization lib org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe	

Schema (12)

View and manage the table schema.

#	Column name	Data type	Partition key	Comment
1	transaction_id	bigint	-	-
2	transaction_date	string	-	-
3	transaction_time	string	-	-
4	transaction_qty	bigint	-	-
5	store_id	bigint	-	-
6	store_location	string	-	-
7	product_id	bigint	-	-
8	unit_price	double	-	-

12. Partitions on Aws Data Catalog:



Crawler successfully starting
The following crawler is now starting: "Coffee_sales_project"

sales_data

Last updated (UTC)
June 29, 2024 at 07:09:49

Name sales_data	Description -	Database db_data_analysis	Classification CSV
Location s3://s3-data-analysis-sales/Sales-data/	Connection -	Deprecated -	Last updated June 29, 2024 at 07:09:29
Input format org.apache.hadoop.mapred.TextInputFormat	Output format org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat	Serde serialization lib org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe	

Partitions

	snapshot_day	Files	Properties
○	2023_Apr	View files	View Properties
○	2023_Feb	View files	View Properties
○	2023_Jan	View files	View Properties
○	2023_June	View files	View Properties
○	2023_May	View files	View Properties
○	2023_Mar	View files	View Properties

As per the Objects we gave on S3 the Partitions name Created as Snapshot_day in the aws glue catalog as per the on demand it will crawl the data uploaded new in the folder.

QUERYING DATA WITH AMAZON ATHENA:

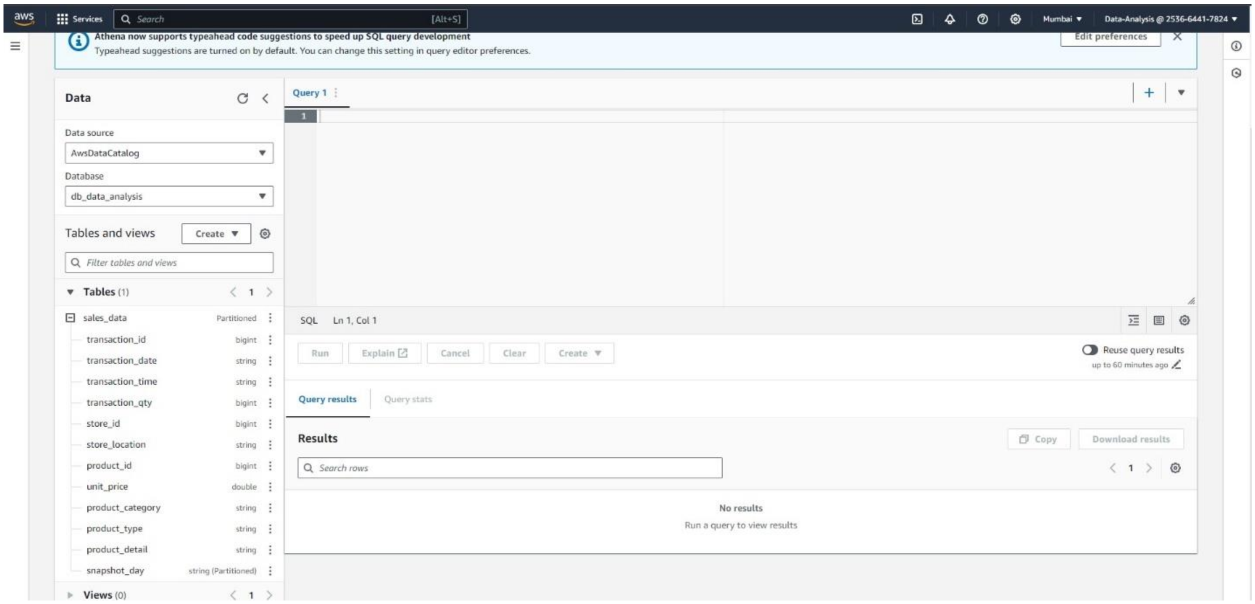
Once the data has been crawled and the schema is established in the AWS Glue Data Catalog, you can seamlessly use Athena to query the data stored in AWS Glue. Athena is a serverless query service that allows you to run standard SQL queries against various data sources.

We start by Opening the Athena Console and navigate to the Query Editor. The data source and a database is required which in our case AWS glue data-catalog (data source) along with the database created in the above step, will be displayed In the Query Editor, the TABLE for this project will be our S3 bucket where the data is stored.

Execute the query on the table in Athena: When executing a query in Athena, it utilizes the metadata stored in the Glue Data Catalog to understand the structure of the data and optimize the query plan. Athena leverages the schema information to perform predicate push-down, column pruning, and other optimizations, resulting in faster query execution.

What's left now is to query the table and see if the configuration is proper. To test this out, we'll run this SQL query on the table:

13. Querry Setup in Athena:



Athena now supports typeahead code suggestions to speed up SQL query development
Typeahead suggestions are turned on by default. You can change this setting in query editor preferences.

Data source: AwsDataCatalog

Database: db_data_analysis

Tables (1)

sales_data	Partitioned
transaction_id	bigint
transaction_date	string
transaction_time	string
transaction_qty	bigint
store_id	bigint
store_location	string
product_id	bigint
unit_price	double
product_category	string
product_type	string
product_detail	string
snapshot_day	string (Partitioned)

Views (0)

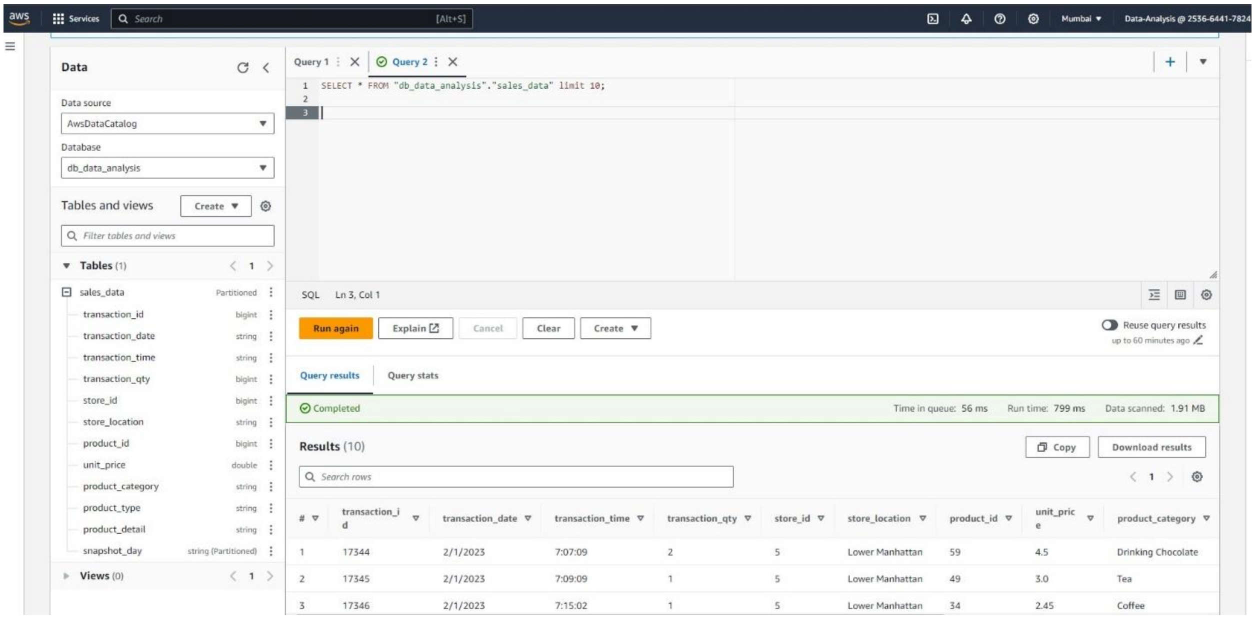
Reuse query results up to 60 minutes ago

No results
Run a query to view results

After running this query in the result section, we are able to see the output which is displayed on the picture below.

Athena returns the query results, which can be downloaded or further analyzed using various visualization or reporting tools, although query results are saved automatically for every query that runs regardless of whether the query itself was downloaded or not.

14. Querry Result in Athena:



Data

Data source: AwsDataCatalog

Database: db_data_analysis

Tables and views

Tables (1)

sales_data	Partitioned
transaction_id	bigint
transaction_date	string
transaction_time	string
transaction_qty	bigint
store_id	bigint
store_location	string
product_id	bigint
unit_price	double
product_category	string
product_type	string
product_detail	string
snapshot_day	string (Partitioned)

Views (0)

```
SELECT * FROM "db_data_analysis"."sales_data" limit 10;
```

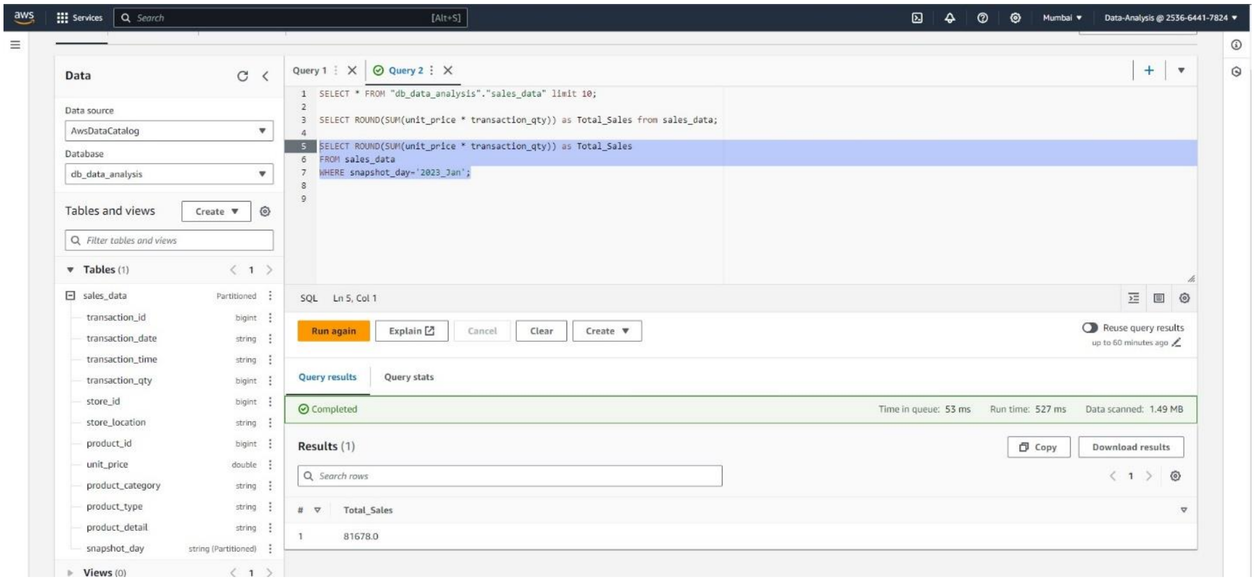
SQL Ln 3, Col 1

Completed Time in queue: 56 ms Run time: 799 ms Data scanned: 1.91 MB

Results (10)

#	transaction_id	transaction_date	transaction_time	transaction_qty	store_id	store_location	product_id	unit_price	product_category
1	17344	2/1/2023	7:07:09	2	5	Lower Manhattan	59	4.5	Drinking Chocolate
2	17345	2/1/2023	7:09:09	1	5	Lower Manhattan	49	3.0	Tea
3	17346	2/1/2023	7:15:02	1	5	Lower Manhattan	34	2.45	Coffee

15.Filtering with partitions:



```
SELECT * FROM "db_data_analysis"."sales_data" limit 10;

SELECT ROUND(SUM(unit_price * transaction_qty)) as Total_Sales from sales_data;

SELECT ROUND(SUM(unit_price * transaction_qty)) as Total_Sales
FROM sales_data
WHERE snapshot_day='2023_Jan';
```

SQL Ln 5, Col 1

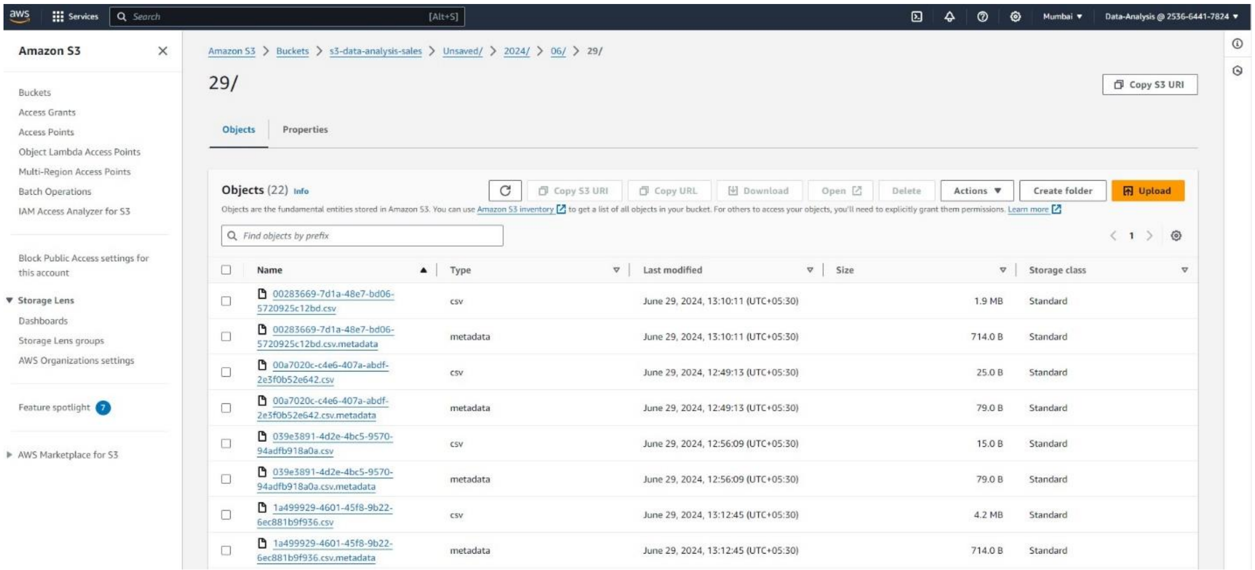
Completed Time in queue: 53 ms Run time: 527 ms Data scanned: 1.49 MB

Results (1)

#	Total_Sales
1	81678.0

As per the Partitions we got in Aws Glue data catalog we can filter the queries, then it scans the particular folder. This optimizes the query and executes.

16. Querry Logs saved in S3:



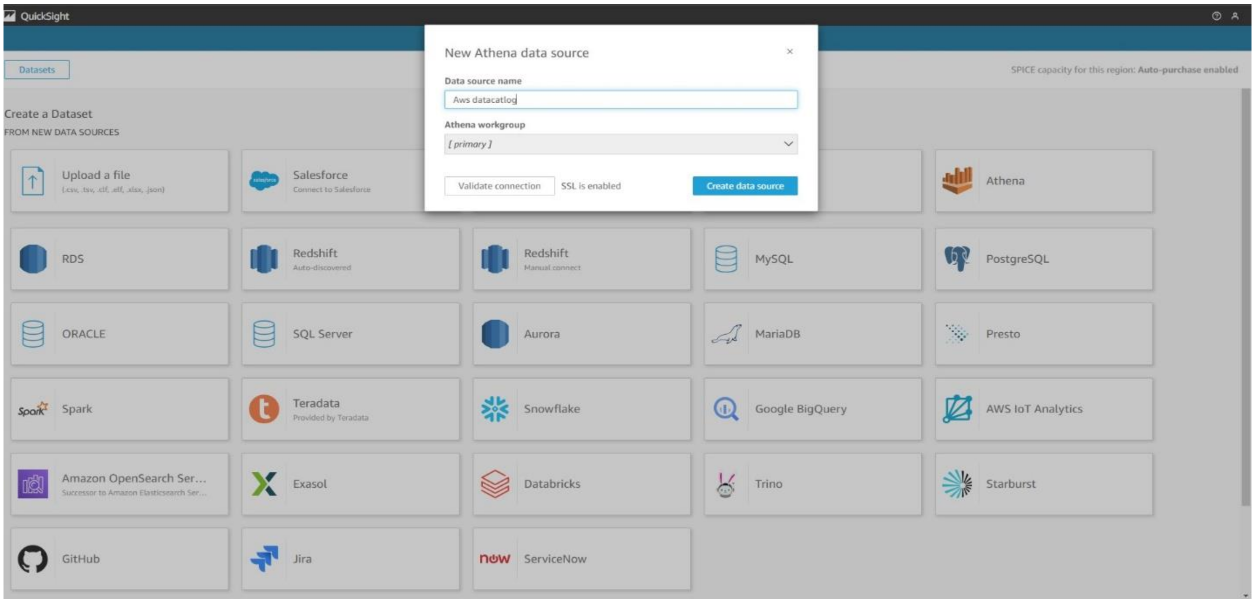
Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more

☐	Name	Type	Last modified	Size	Storage class
☐	00283669-7d1a-48e7-bd06-5720925c12bd.csv	csv	June 29, 2024, 13:10:11 (UTC+05:30)	1.9 MB	Standard
☐	00283669-7d1a-48e7-bd06-5720925c12bd.csv.metadata	metadata	June 29, 2024, 13:10:11 (UTC+05:30)	714.0 B	Standard
☐	00a7020c-c4e6-407a-abdf-2e3f0b52e642.csv	csv	June 29, 2024, 12:49:13 (UTC+05:30)	25.0 B	Standard
☐	00a7020c-c4e6-407a-abdf-2e3f0b52e642.csv.metadata	metadata	June 29, 2024, 12:49:13 (UTC+05:30)	79.0 B	Standard
☐	039e3891-4d2e-4bc5-9570-94adfb918a0a.csv	csv	June 29, 2024, 12:56:09 (UTC+05:30)	15.0 B	Standard
☐	039e3891-4d2e-4bc5-9570-94adfb918a0a.csv.metadata	metadata	June 29, 2024, 12:56:09 (UTC+05:30)	79.0 B	Standard
☐	1a499929-4601-45f8-9b22-6ec881b9f936.csv	csv	June 29, 2024, 13:12:45 (UTC+05:30)	4.2 MB	Standard
☐	1a499929-4601-45f8-9b22-6ec881b9f936.csv.metadata	metadata	June 29, 2024, 13:12:45 (UTC+05:30)	714.0 B	Standard

VISLAIZING DATA WITH AMAZON QUICKSIGHT:

Amazon quick Sight provides a user-friendly interface for creating visualizations from various data sources, including CSV files. Follow the steps above to import your data, create a visualization, and customize it according to your requirements. Once you're logged in to your quick Sight account, create a new analysis by clicking on "New analysis" on the quick Sight homepage.

17. Importing Data from Athena:



New Athena data source

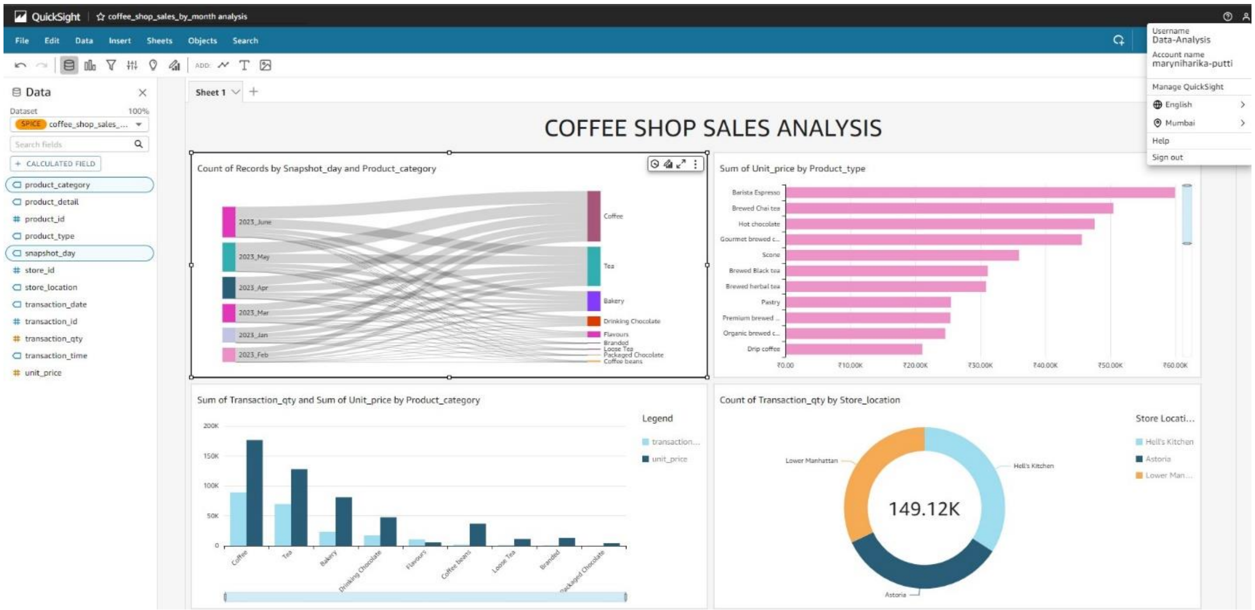
Data source name: Aws datacatlog

Athena workgroup: [primary]

SSL is enabled

Configure data fields: quick Sight will automatically detect the fields in your CSV file. Verify that the data types and field settings are correct. Make any necessary adjustments by clicking on the field names. Create a visualization: After your data is configured, click on the “Visualize” button to start creating your graph. Choose the appropriate visualization type for your data, such as a bar chart, Sankey chart, column chart, or scatter plot. Select data fields: In the visualization editor, drag and drop the relevant data fields from the field list onto the appropriate areas of the chart. For example, you might place a numeric field on the y-axis and a time-based field on the x-axis.

18. Reporting the sales visualization:



CONCLUSION:

The seamless integration of AWS Athena and AWS Glue enables organizations to leverage schema and crawled data for efficient and optimized database queries. By combining the power of metadata-driven analysis and query execution, businesses can uncover valuable insights, make informed decisions, and drive success in the data-driven era.

PREPARED BY

MARYNIHARIKA PUTTI