**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.

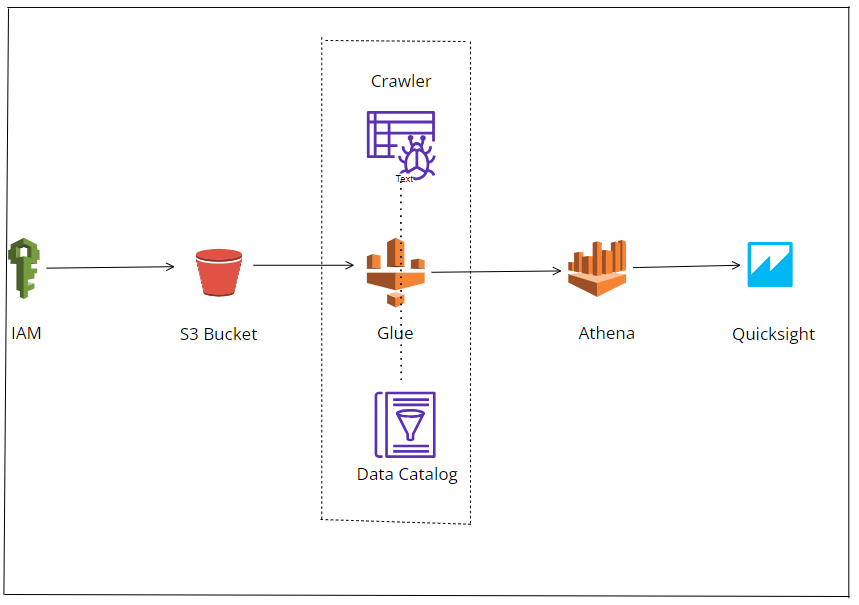
By the end of this article, data engineers, analysts, and anyone interested in harnessing the power of AWS for data processing and visualization will have a solid understanding of how to upload a CSV file into Amazon S3, utilize AWS Glue to crawl and create a database, leverage Amazon Athena for flexible 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:**

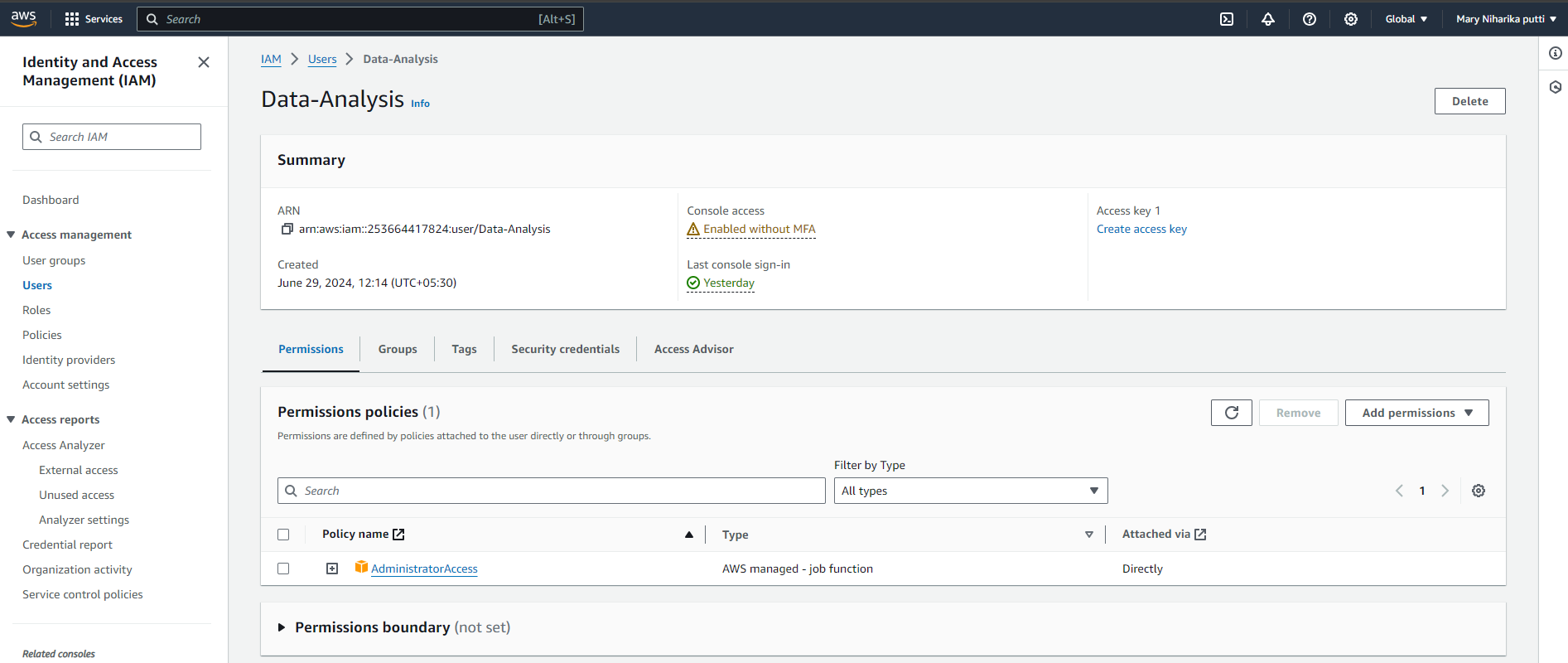
Before proceeding with the steps outlined in this article, ensure that you have the following prerequisites in place:

* 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.

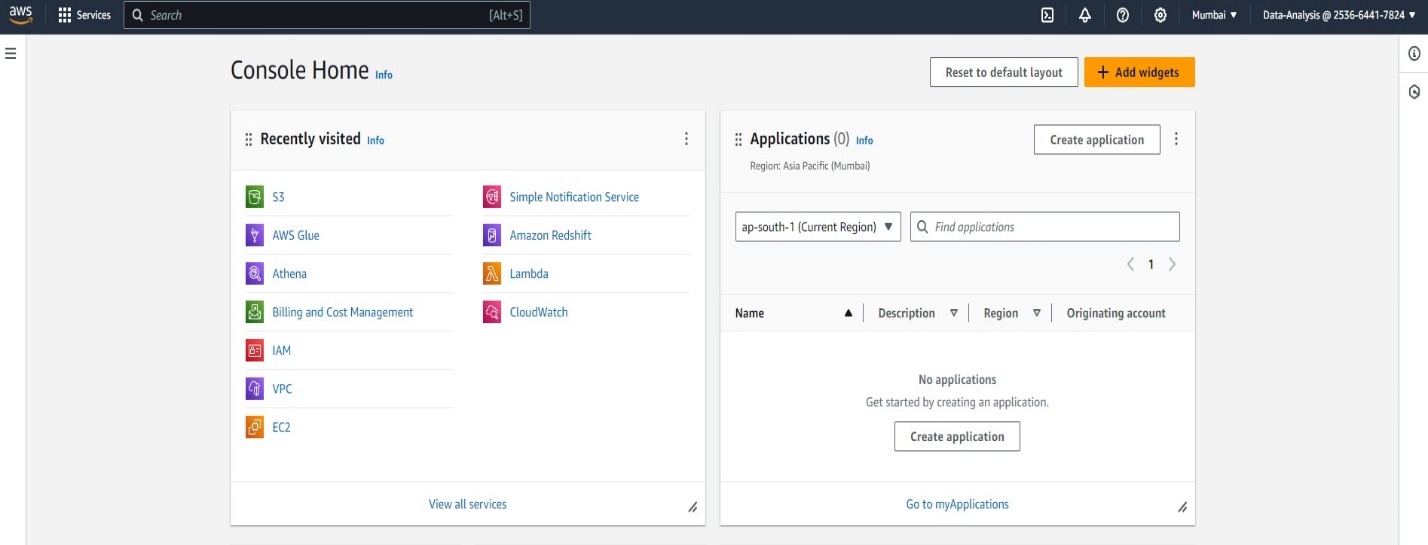
**CREATING IAM USER AND LOGIN:**

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:



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

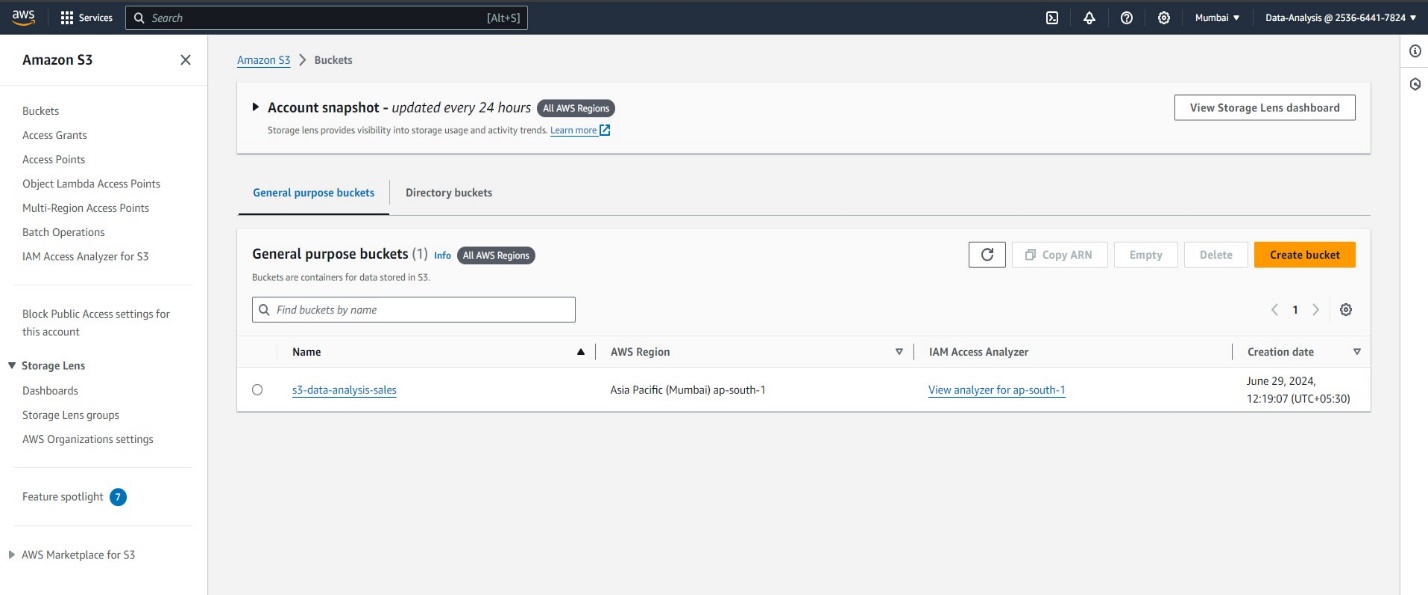


**UPLOADING DATA TO AMAZON S3:**

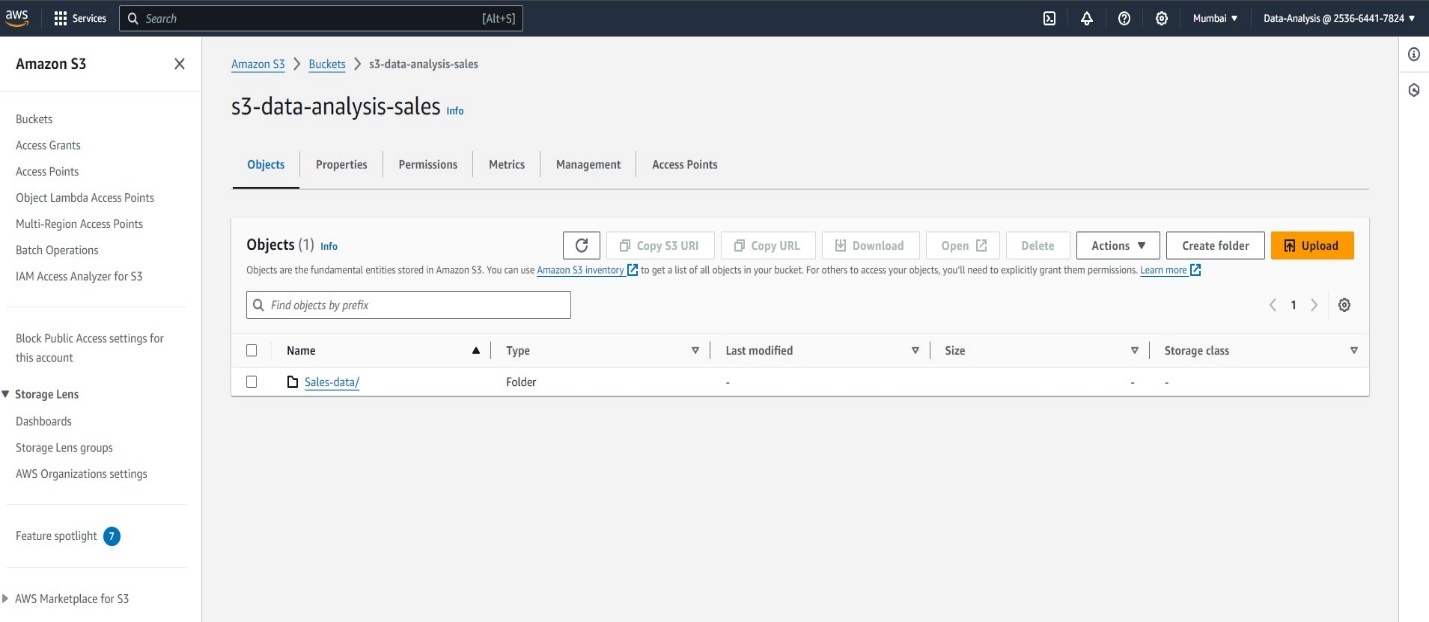
Amazon Simple Storage Service (S3) is a highly scalable and durable object storage service provided by AWS. It offers a secure and reliable solution for storing and retrieving data from anywhere on the web. Let’s explore how you can upload a CSV file to S3 and ensure efficient data organization.

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.

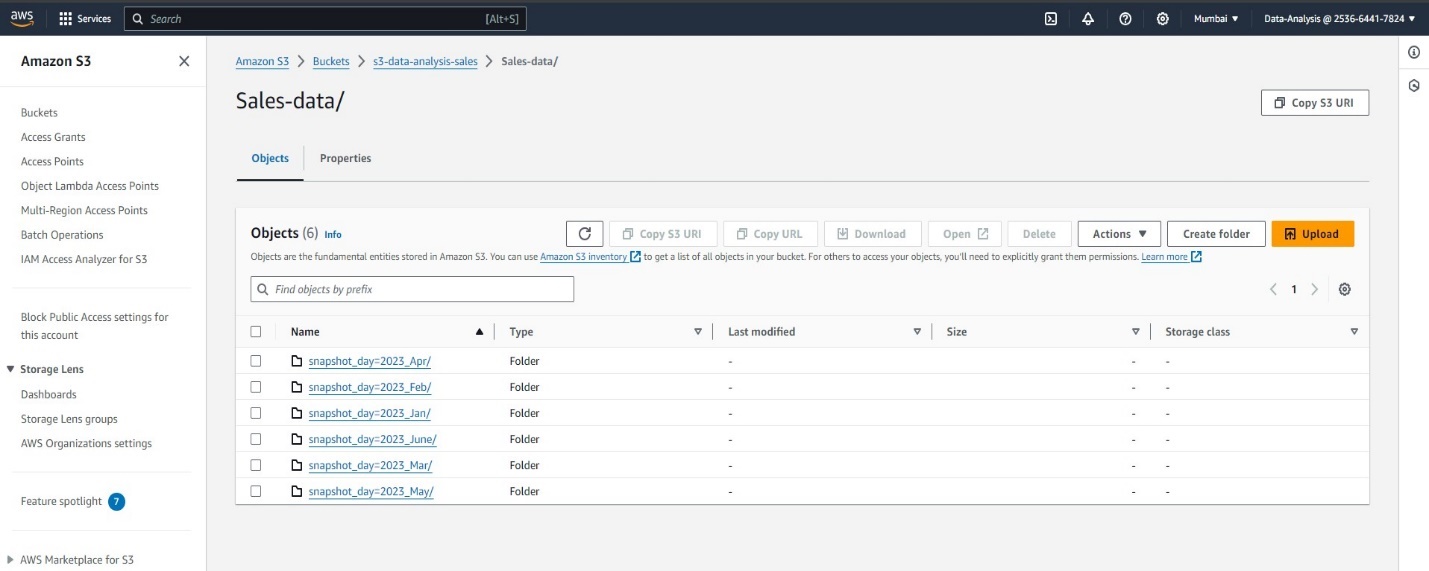
3. Created Bucket S3-data-analysis- sales:



4. Created folder as sales-data:



5. Uploaded Objects Sales.CSV:

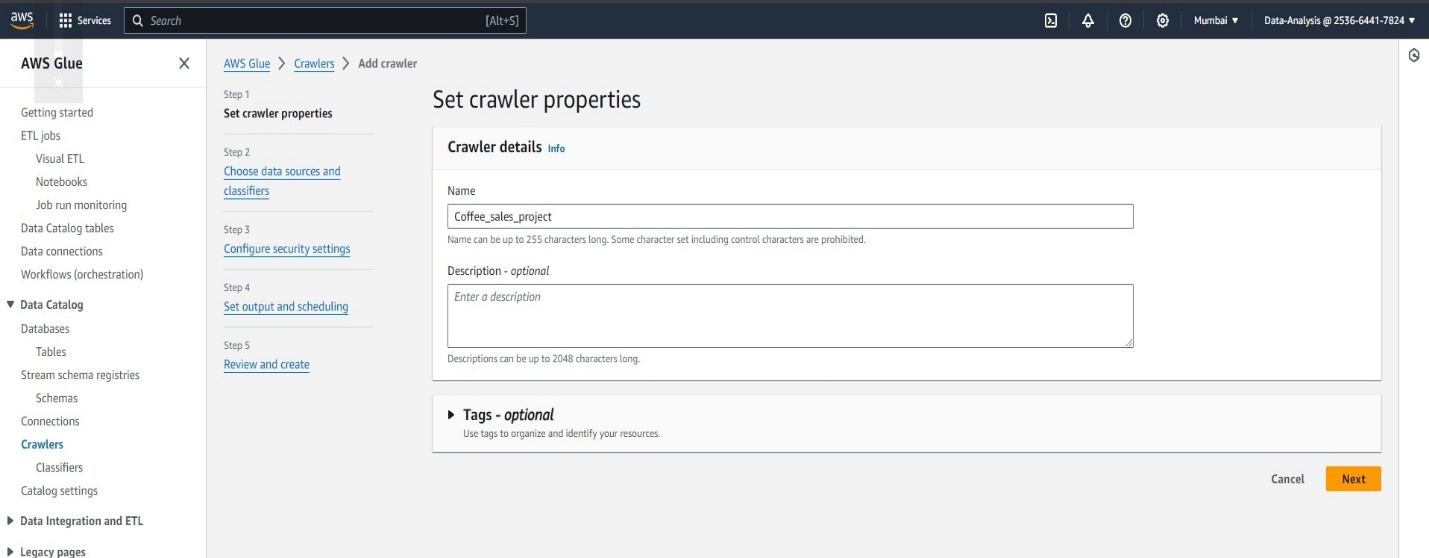
Here, I Have Created a snapshot\_day=2033\_jan to get partitions as per month. Then I have Uploaded my objects.csv file in to it.

**DATA CRAWLING WITH AWS GLUE:**

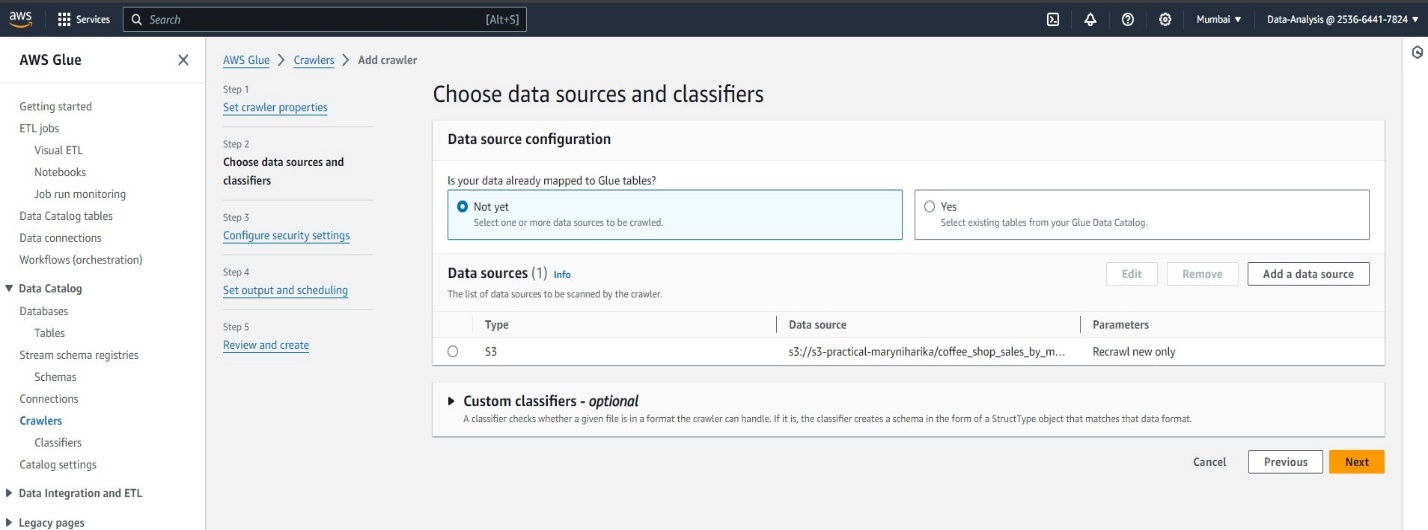
Data crawling is the process of automatically scanning and analyzing the data sources to infer the schema and structure of the data. It allows AWS Glue to understand the data format, columns, and data types without requiring manual intervention.

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:

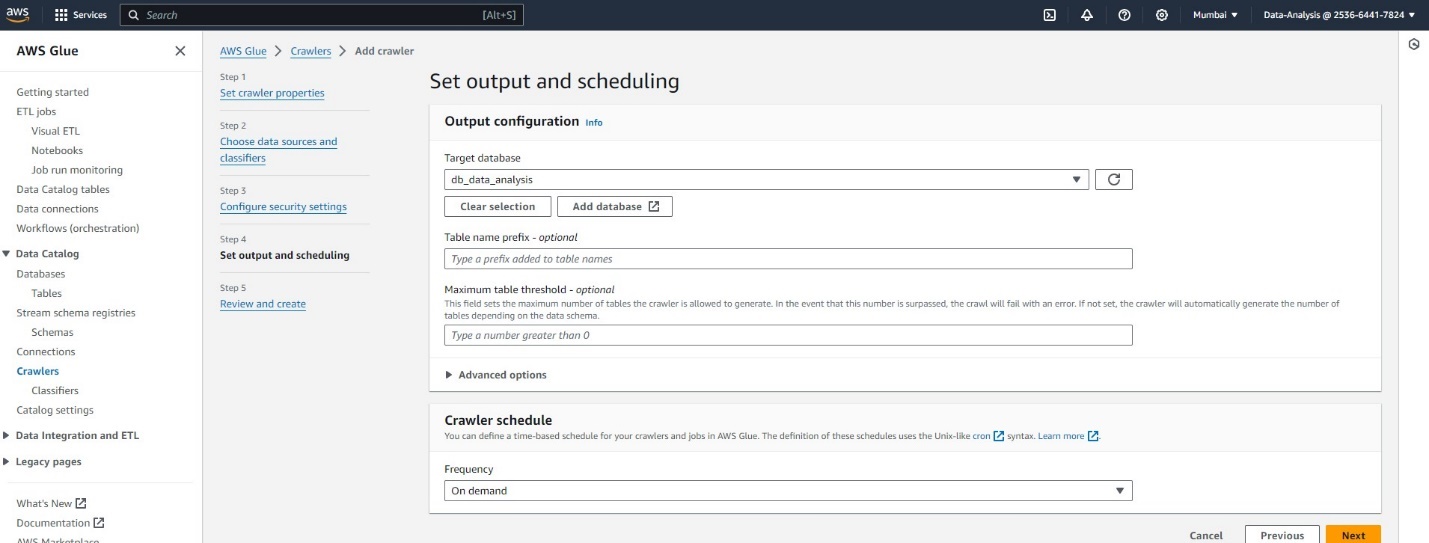


7. Adding source to the crawler:



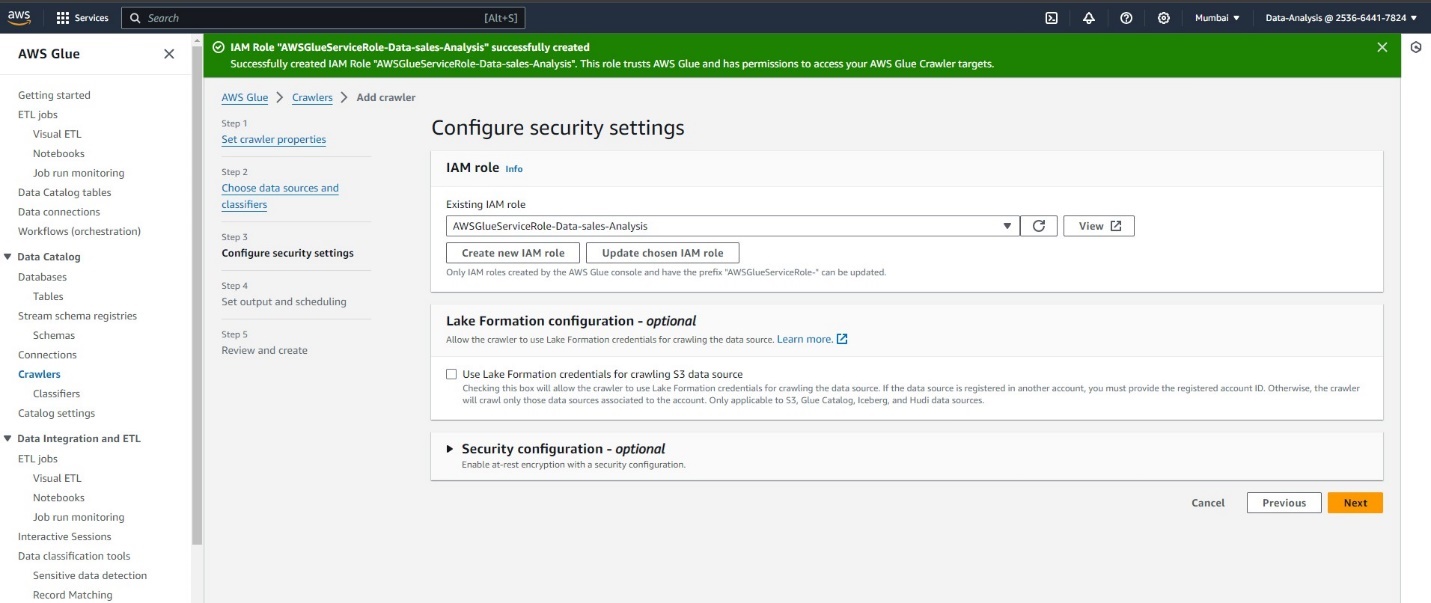
Next, we want to create a database, open the AWS Glue Console and navigate to the Data Catalog section, select Databases and proceed to creating a new database, click on add database on the top right-hand side of the console, key in details of the database such as name, location(optional) and description(optional),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

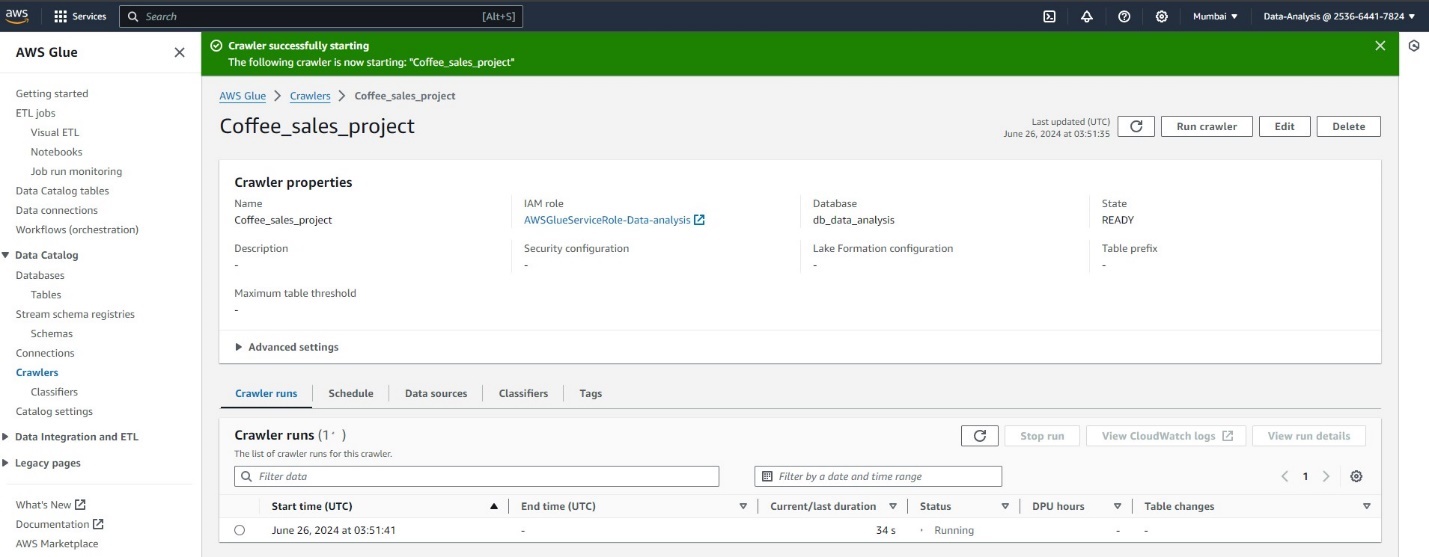


Configure the crawler’s settings, including IAM roles, database location, and frequency of running. e. Optionally, set up crawler-specific classifiers to handle special data formats or custom schema inference, then we Review the crawler configuration and click on “Finish” to create the crawler.

9. Creating I am Role:



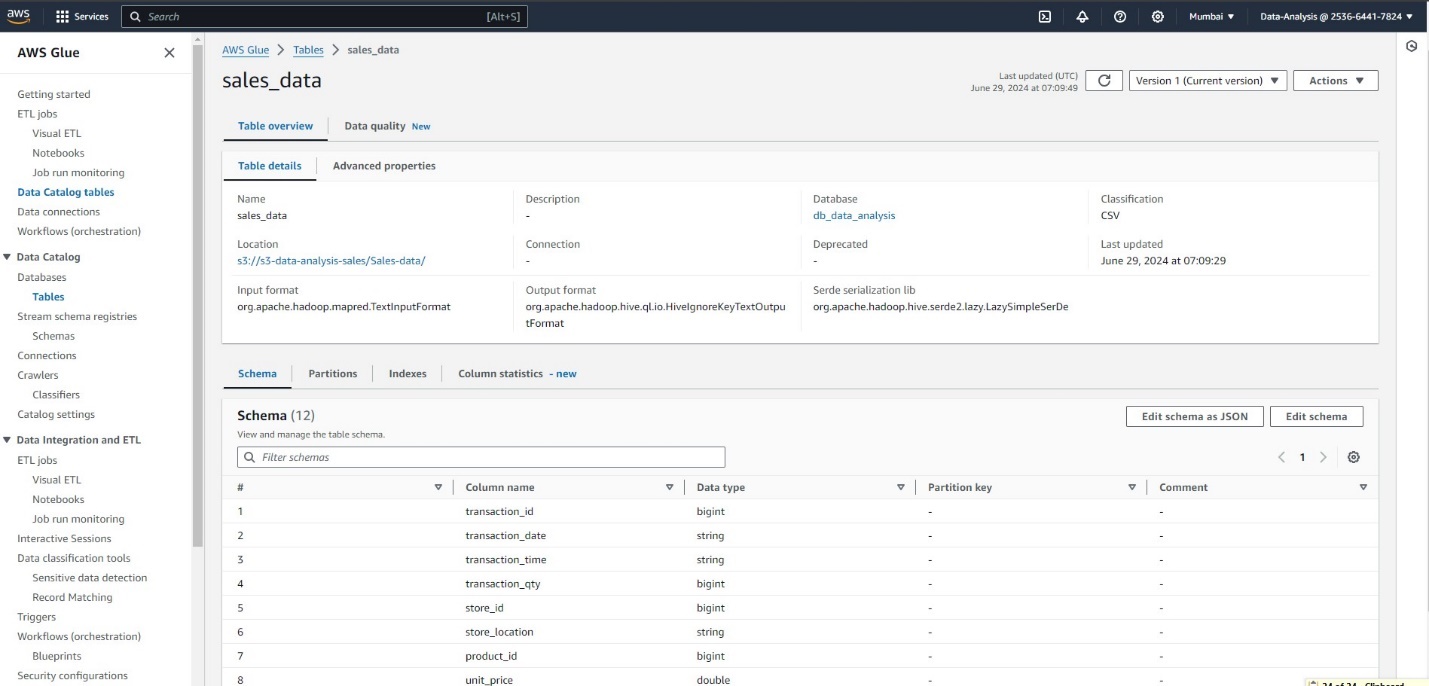
10. Crawler Running:



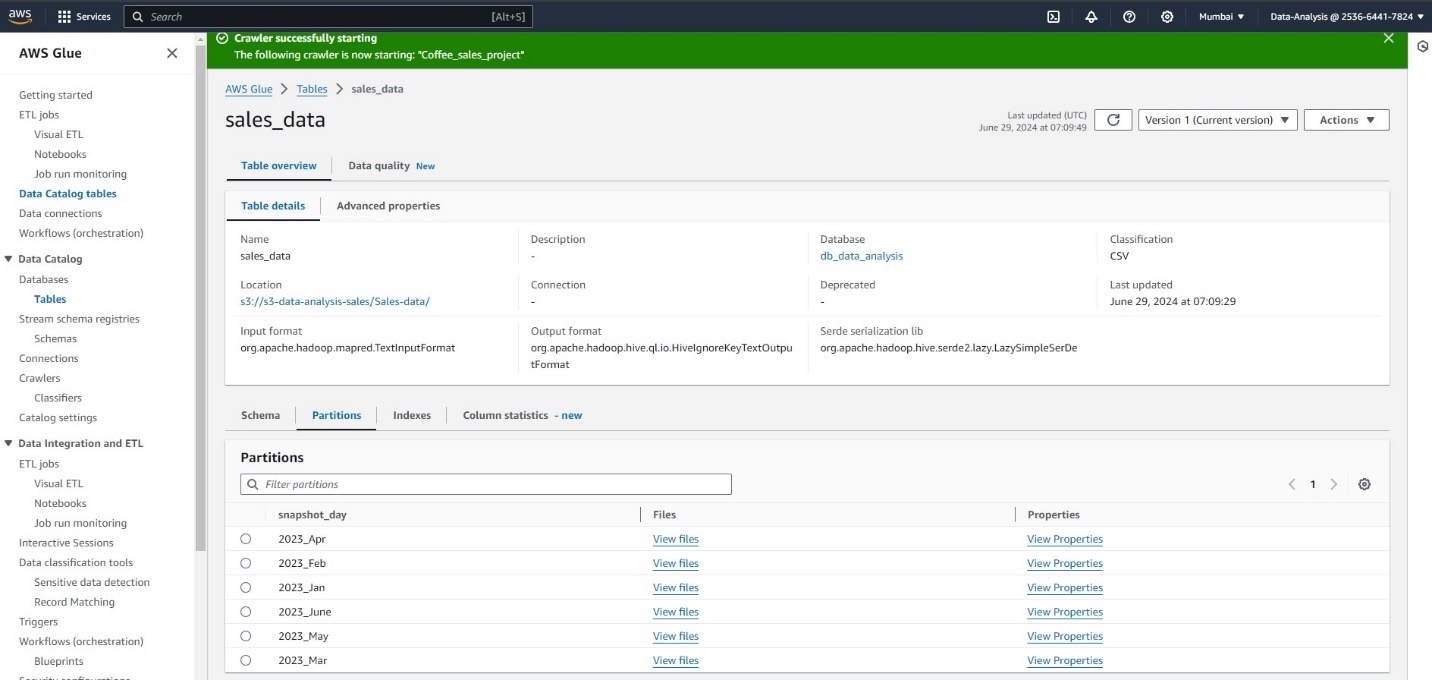
Automatic Schema and Table Definition Generation: Once the crawler is created and executed, AWS Glue automatically generates schema and table definitions based on the crawled data. It examines the structure of the CSV file, infers the data types, and creates metadata entries in the AWS Glue Data Catalog. 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:



12. Partitions on Aws Data Catalog:



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:**

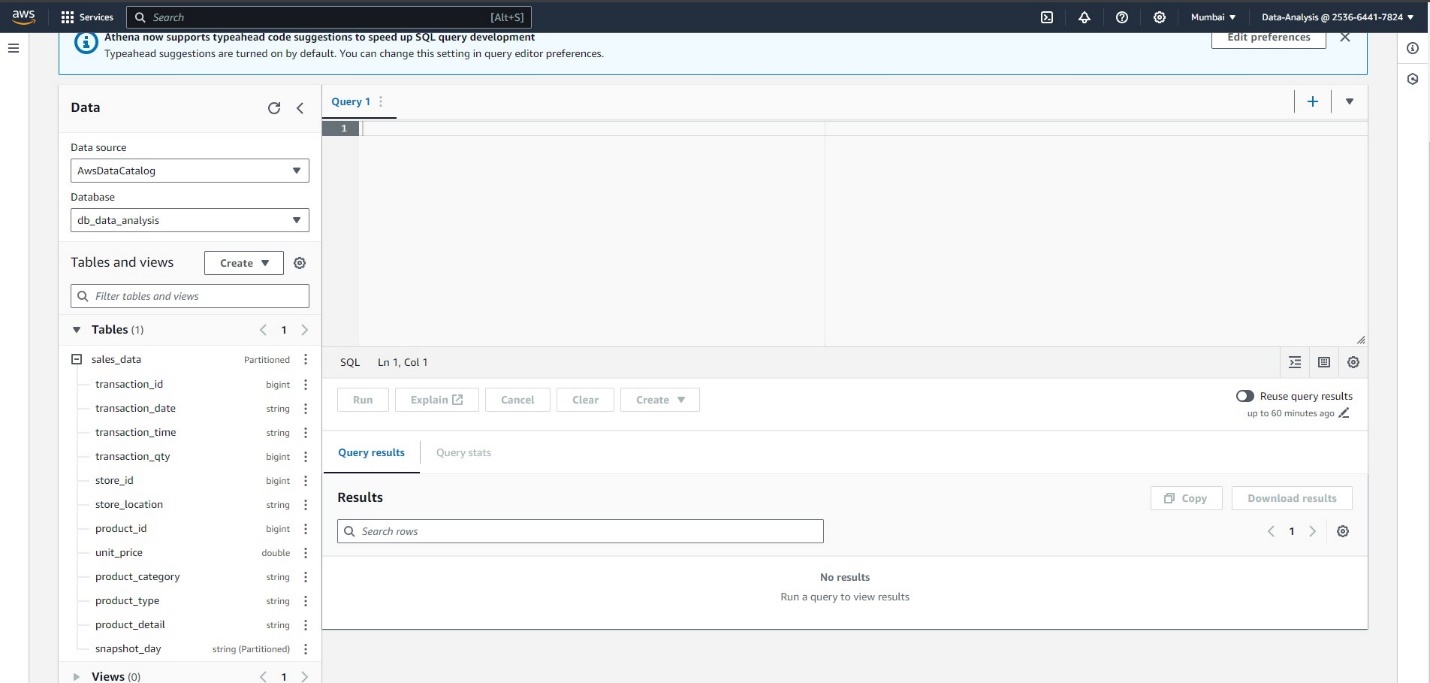
Once the data has been crawled and the schema is established in the AWS Glue Data Catalog, you can seamlessly use Athena to query the data stored in AWS Glue. Athena is a serverless query service that allows you to run standard SQL queries against various data sources.

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:

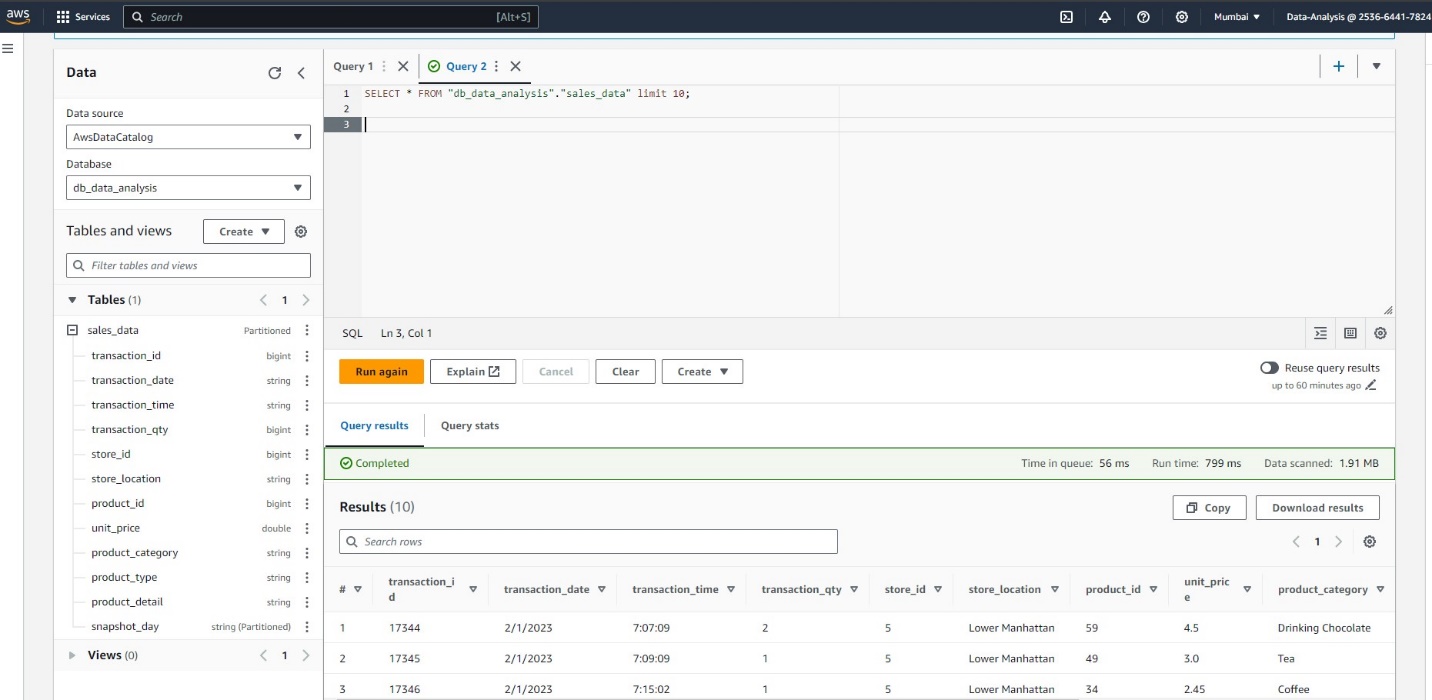
13. Querry Setup in Athena:



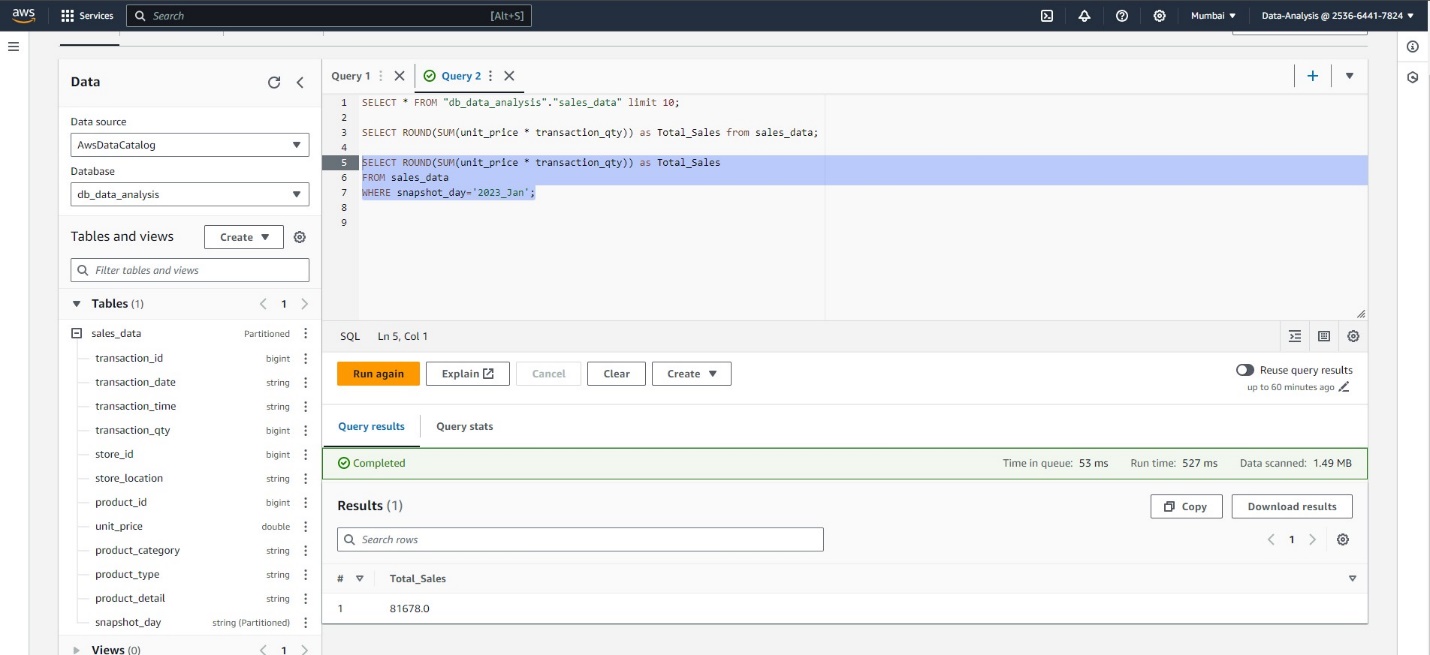
After running this query in the result section, we are able to see the output which is displayed on the picture below.

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:

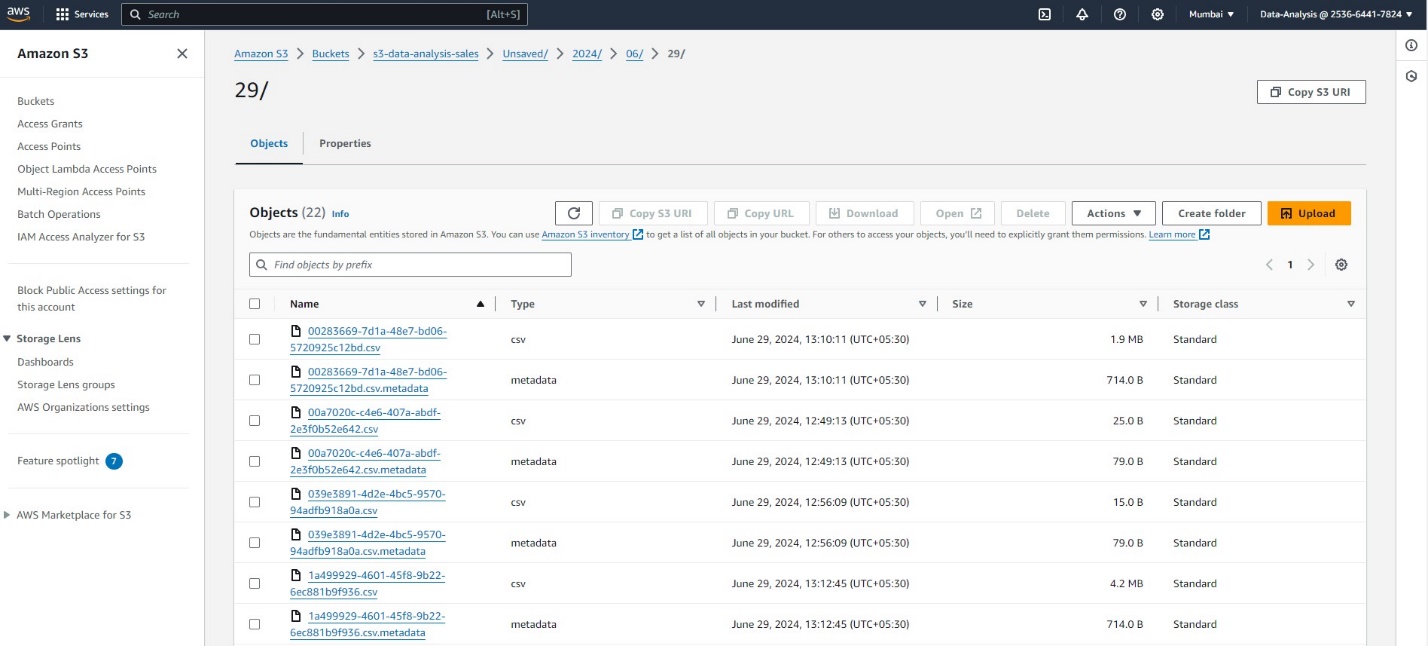


15.Filtering with partitions:



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.

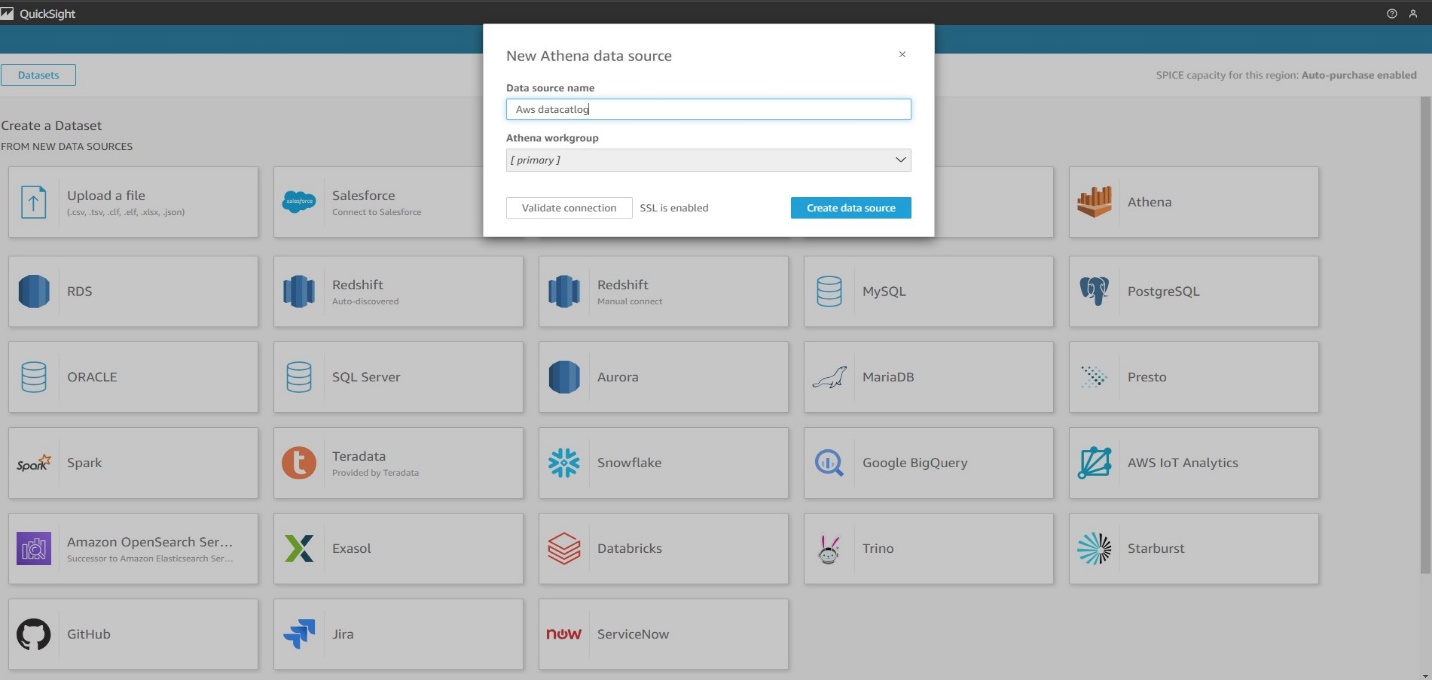
16. Querry Logs saved in S3:



**VISLAIZING DATA WITH AMAZON QUICKSIGHT:**

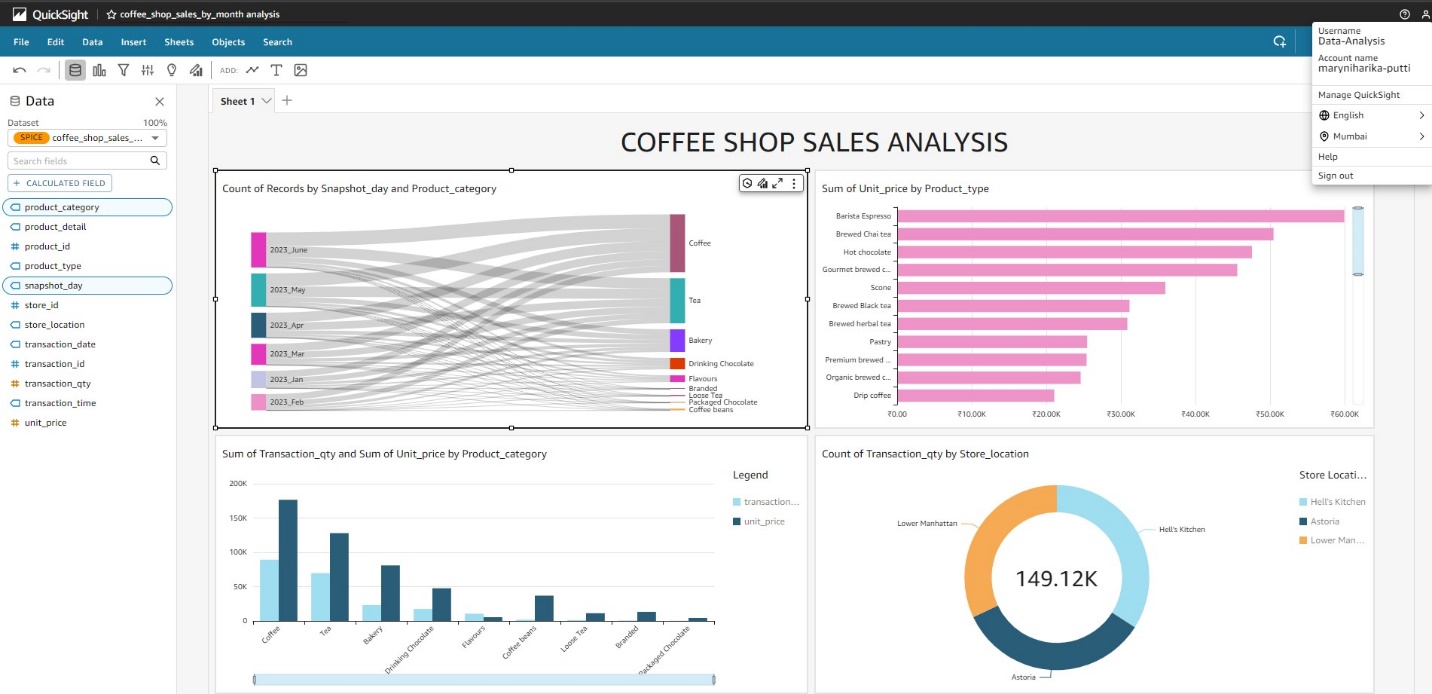
Amazon quick Sight provides a user-friendly interface for creating visualizations from various data sources, including CSV files. Follow the steps above to import your data, create a visualization, and customize it according to your requirements. Once you’re logged in to your quick Sight account, create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

17. Importing Data from Athena:



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