Batch Time Analysis of Transactional Data

Description

Lenodo is a multinational e-commerce organization that sells products directly to consumers. The database administrator exports the data every night in a CSV file, but this export functionality is unused. Lenodo wants to use this data to uncover insights about the most-sold item and the countries where customers have bought this item.

You are a data analytics consultant, and you're asked to provide valuable insights and statistics across products, brands, categories, segments to the marketing, product, sales, and procurement teams and inform them about which product has the highest amount of sales and which product and its marketing needs the most improvement. These statistics will help to run effective digital marketing campaigns. The scope of this project is limited to data engineering and analysis.

Objective:

To use AWS Big Data stack for data engineering to analyze transactions, uncover patterns, and share actionable insights

Steps to perform

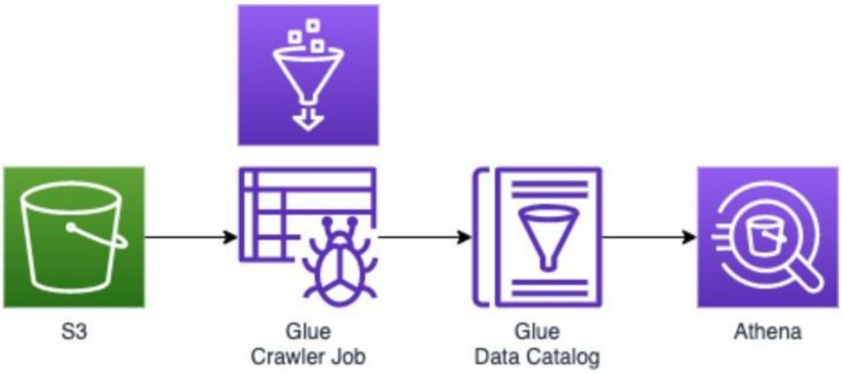
1.Create an S3 bucket with a unique name and upload the CSV file to the S3 bucket

2.Create a crawler to crawl the CSV data and generate a metadata catalog

3.Create a Glue job to transform the data into the Parquet format as CSV is not optimal for data warehouse queries

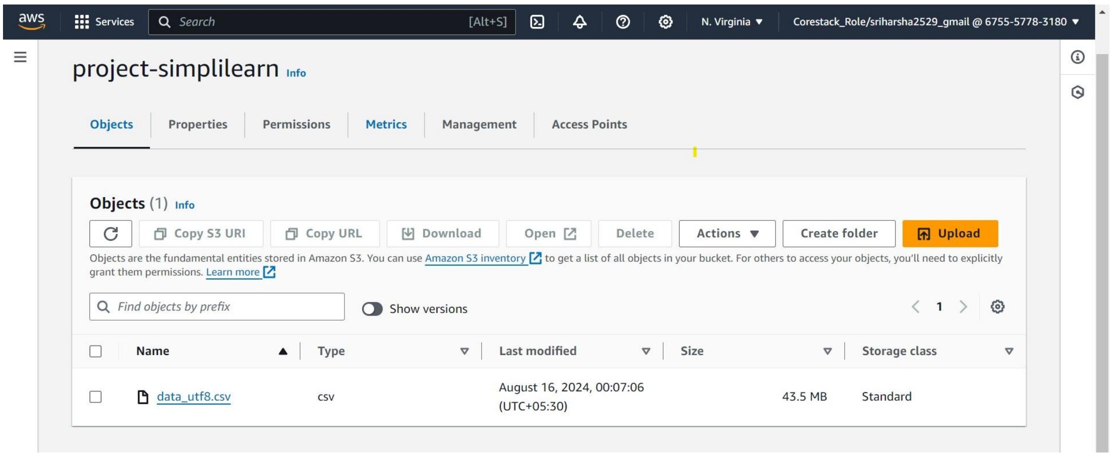
4.Add another crawler to crawl the Parquet data files to generate the metadata catalog of the Parquet file in order to query it with Athena

5.Query the data to identify the best-selling item and countries where customers have bought the most-sold item using Athena.



Batch Time Analysis of Transactional Data

Create a S3 Bucket and loaded data



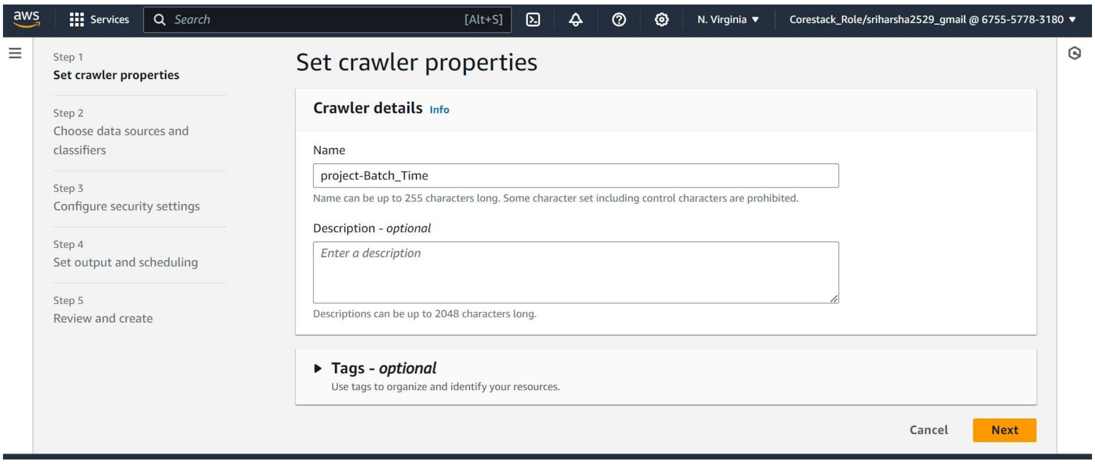
project-simplilearn Info

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
□	data_utf8.csv	csv	August 16, 2024, 00:07:06 (UTC+05:30)	43.5 MB	Standard

Step 2 : Created A crawler


aws
Services
Search [Alt+S]
N. Virginia ▾
Corestack_Role/sriharsha2529_gmail @ 6755-5778-3180 ▾
Step 1
Set crawler properties
Step 2
Choose data sources and classifiers
Step 3
Configure security settings
Step 4
Set output and scheduling
Step 5
Review and create
Cancel
Next

Set crawler properties

Crawler details Info

Name

project-Batch_Time

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

Description - *optional*

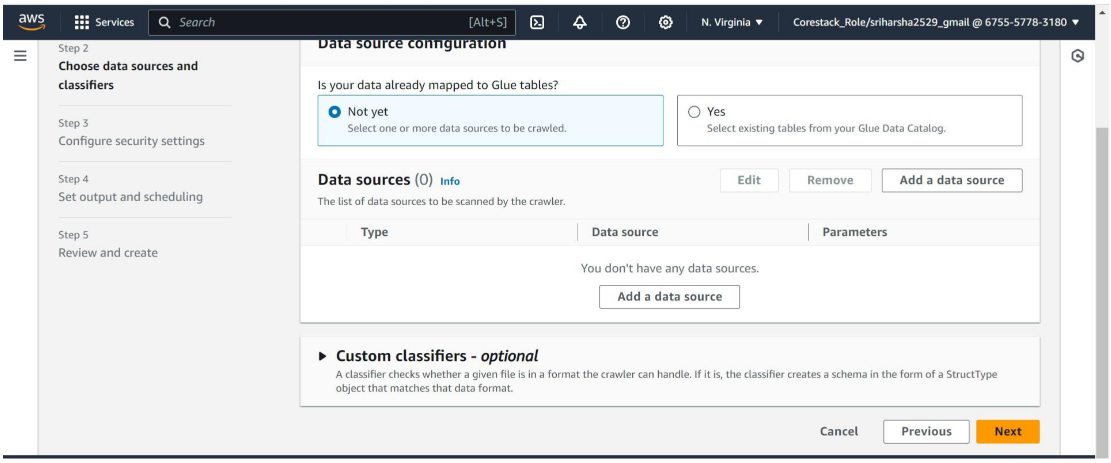
Enter a description

Descriptions can be up to 2048 characters long.

▸ Tags - *optional*

Use tags to organize and identify your resources.

Step 3: Add a Data source from S3


aws
Services
Search [Alt+S]
N. Virginia ▾
Corestack_Role/sriharsha2529_gmail @ 6755-5778-3180 ▾
Step 2
Choose data sources and classifiers
Step 3
Configure security settings
Step 4
Set output and scheduling
Step 5
Review and create
Edit
Remove
Add a data source
Cancel
Previous
Next

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 (0) Info

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

Type	Data source	Parameters

You don't have any data sources.

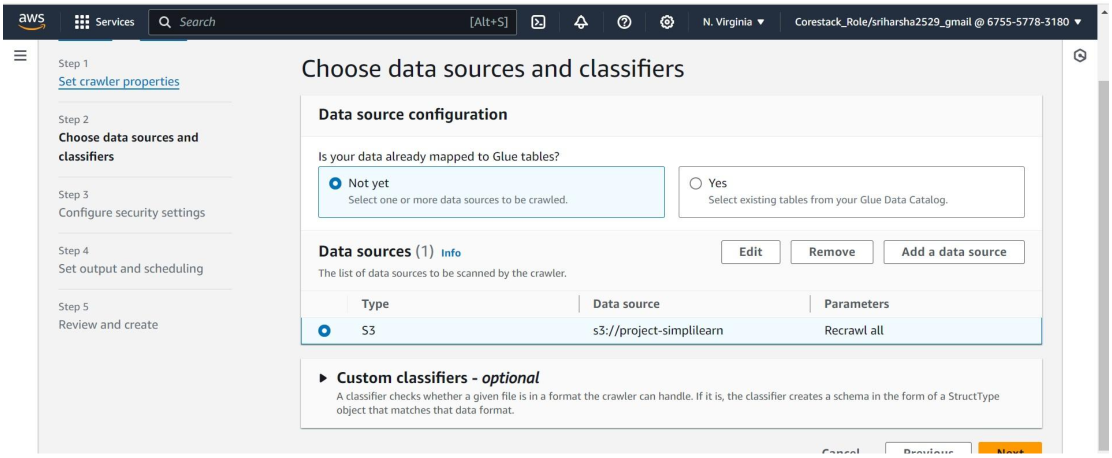
Add a data source

▸ 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.

Batch Time Analysis of Transactional Data

Step 4 : Data Source Added from S3



Step 1
Set crawler properties

Step 2
Choose data sources and classifiers

Step 3
Configure security settings

Step 4
Set output and scheduling

Step 5
Review and create

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.

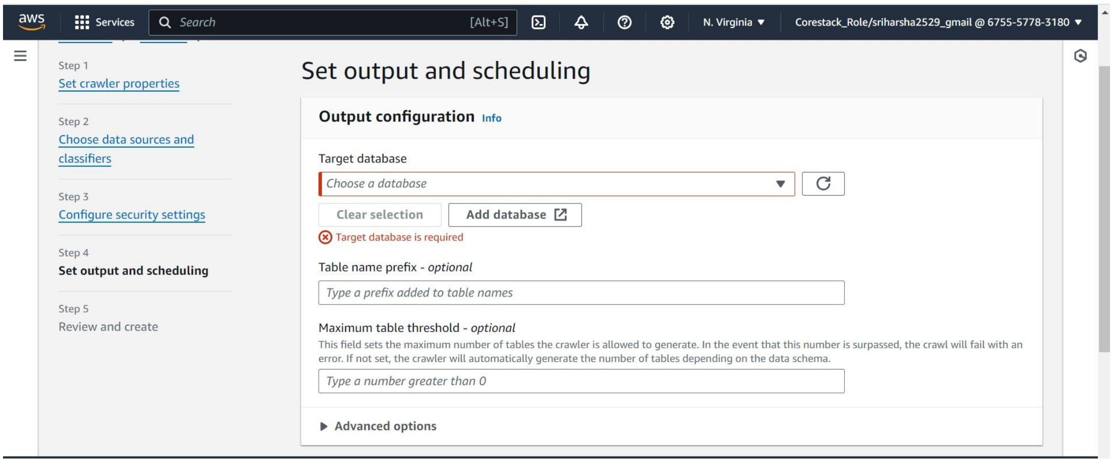
Data sources (1) Info

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

	Type	Data source	Parameters
●	S3	s3://project-simplilearn	Recrawl all

▶ 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.

Step 5 Data Base name



Step 1
Set crawler properties

Step 2
Choose data sources and classifiers

Step 3
Configure security settings

Step 4
Set output and scheduling

Step 5
Review and create

Set output and scheduling

Output configuration Info

Target database
Choose a database ▾

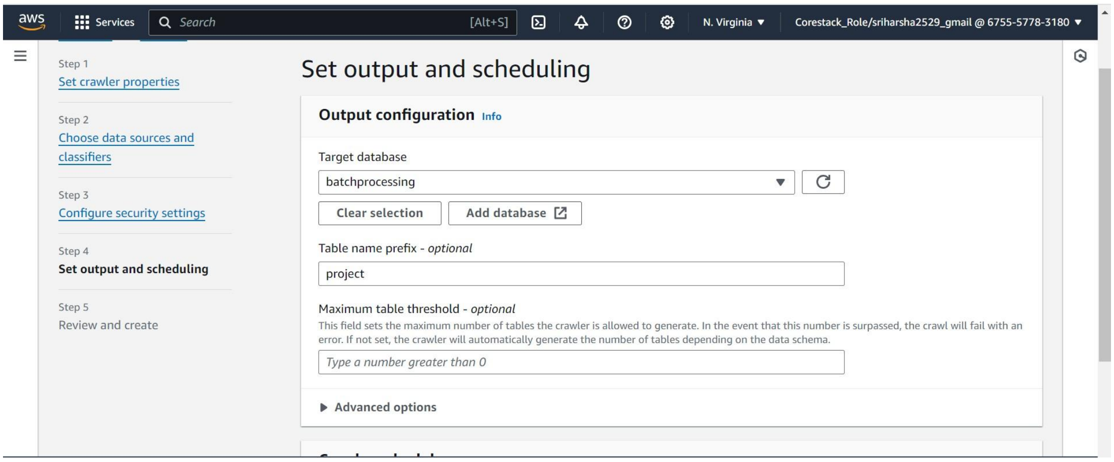
Target database is required

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

Step 6 database created



Step 1
Set crawler properties

Step 2
Choose data sources and classifiers

Step 3
Configure security settings

Step 4
Set output and scheduling

Step 5
Review and create

Set output and scheduling

Output configuration Info

Target database
batchprocessing ▾

Table name prefix - *optional*
project

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

Batch Time Analysis of Transactional Data

Step7schedule

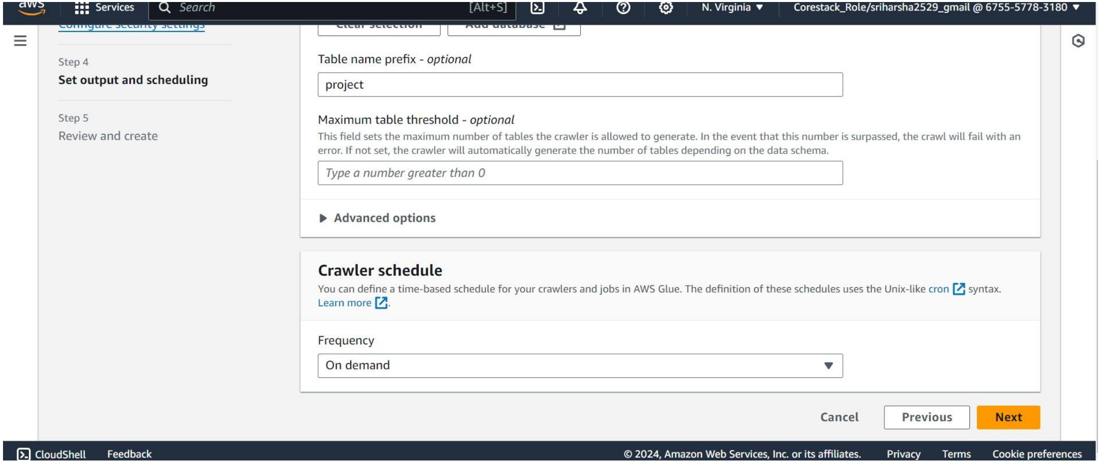


Table name prefix - *optional*

project

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

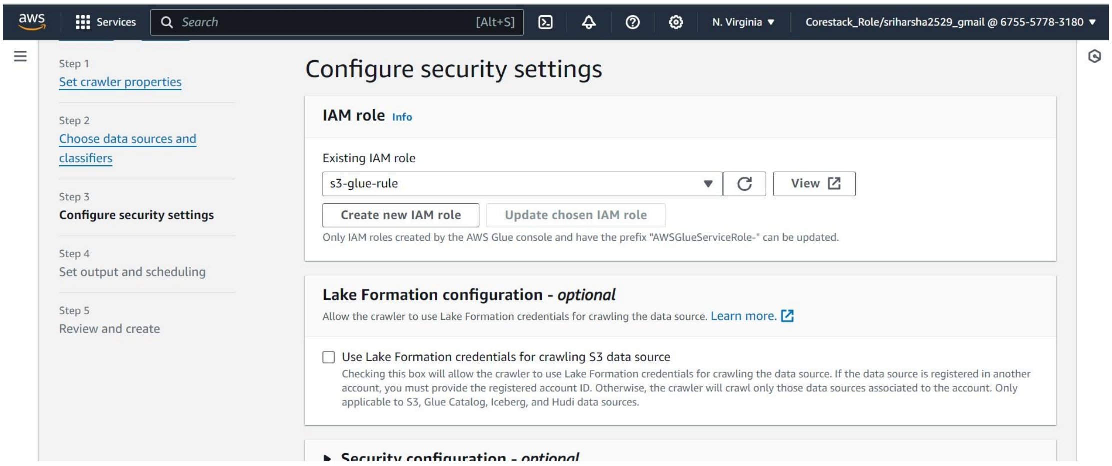
Crawler schedule

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

Step 8 Iam Rule creation



Configure security settings

IAM role

Existing IAM role

s3-glue-rule

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. Learn more.

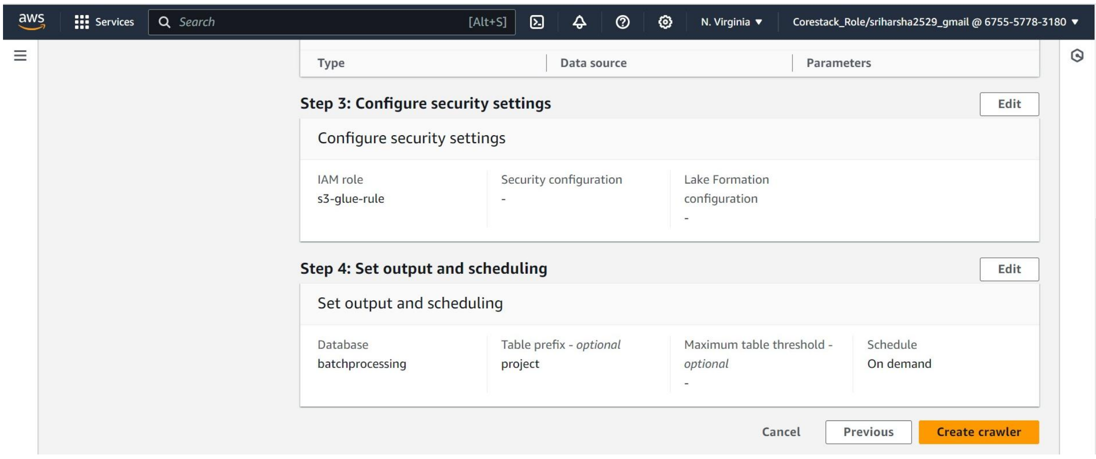
□ Use Lake Formation credentials for crawling S3 data source

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*

Batch Time Analysis of Transactional Data

Step 9 Preview and create



Type	Data source	Parameters

Step 3: Configure security settings

Configure security settings

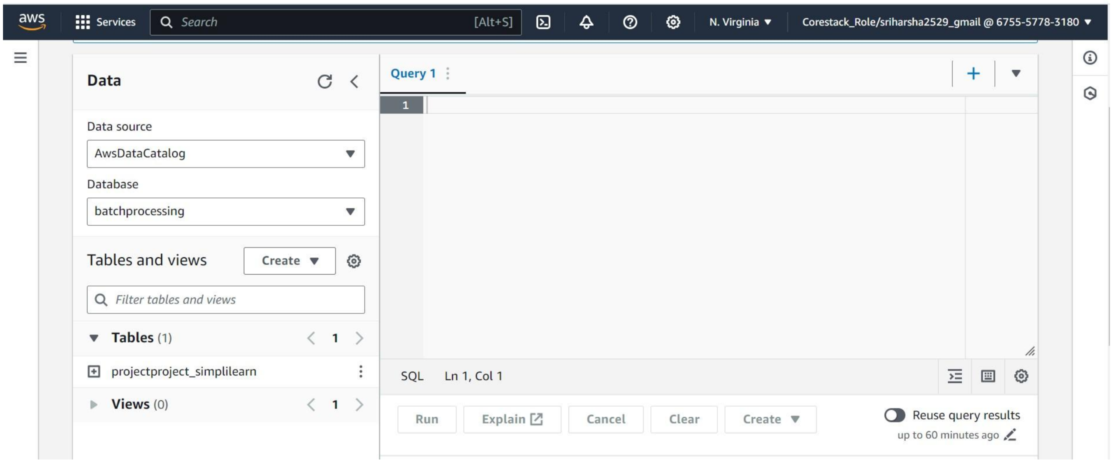
IAM role	Security configuration	Lake Formation configuration
s3-glue-rule	-	-

Step 4: Set output and scheduling

Set output and scheduling

Database	Table prefix - *optional*	Maximum table threshold - *optional*	Schedule
batchprocessing	project	-	On demand

Step 10 AThena



Data

Data source
AwsDataCatalog

Database
batchprocessing

Tables and views

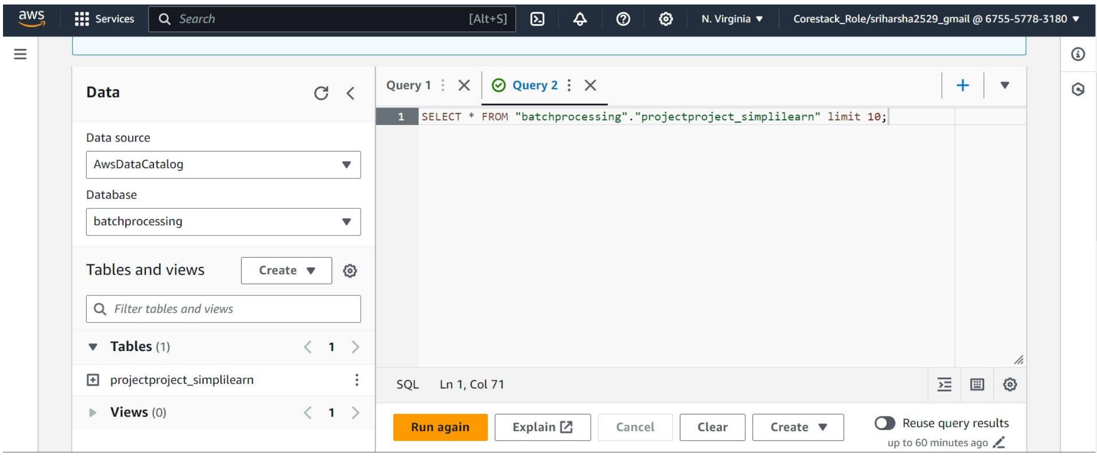
Tables (1)

projectproject_simplilearn

Views (0)

Batch Time Analysis of Transactional Data

Step 11 Quering

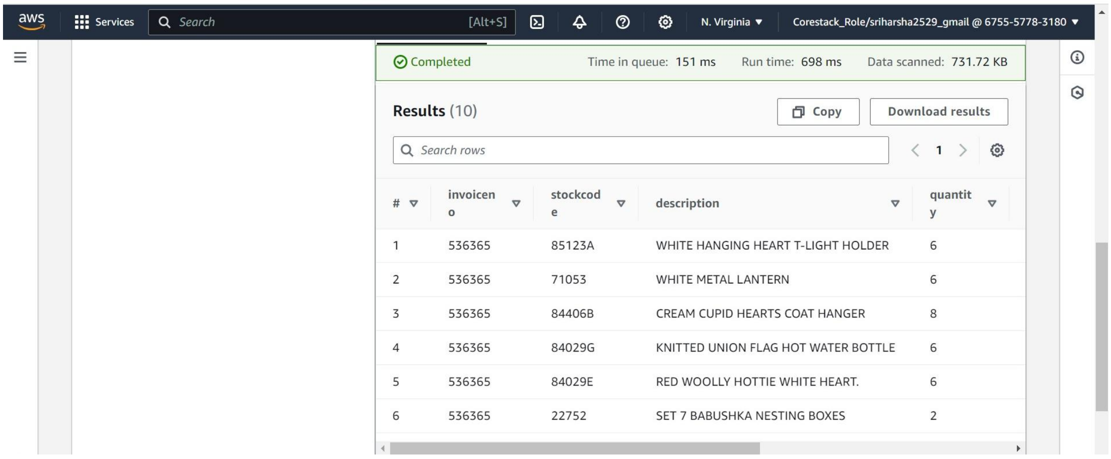


Data source: AwsDataCatalog

Database: batchprocessing

```
SELECT * FROM "batchprocessing"."projectproject_simplilearn" limit 10;
```

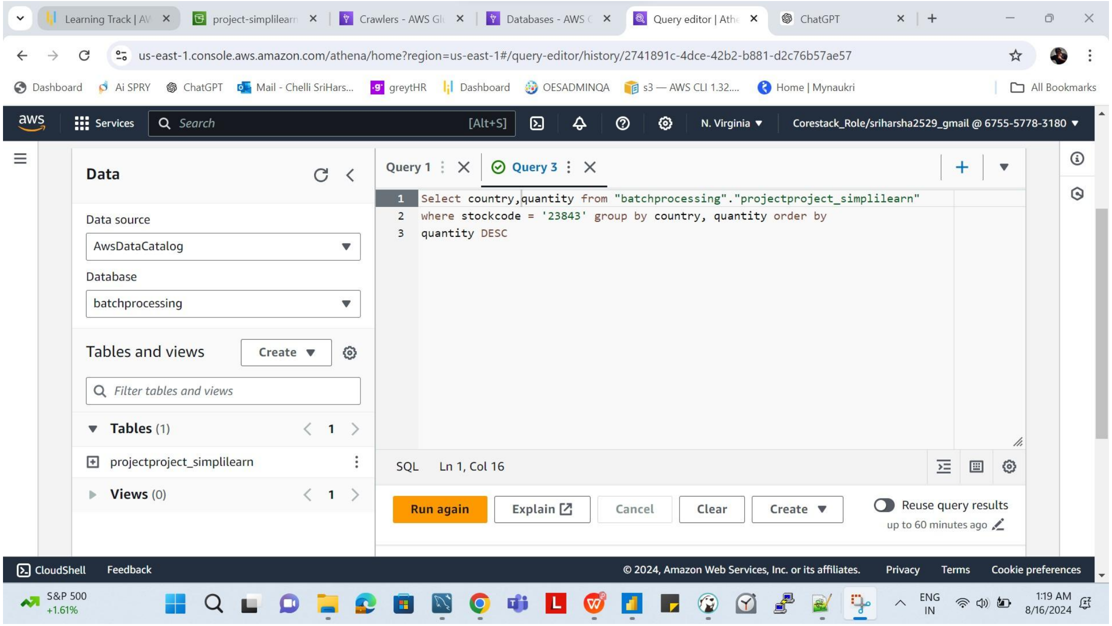
Step 12 result



Completed Time in queue: 151 ms Run time: 698 ms Data scanned: 731.72 KB

Results (10)

#	invoiceno	stockcode	description	quantity
1	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6
2	536365	71053	WHITE METAL LANTERN	6
3	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8
4	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6
5	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6
6	536365	22752	SET 7 BABUSHKA NESTING BOXES	2

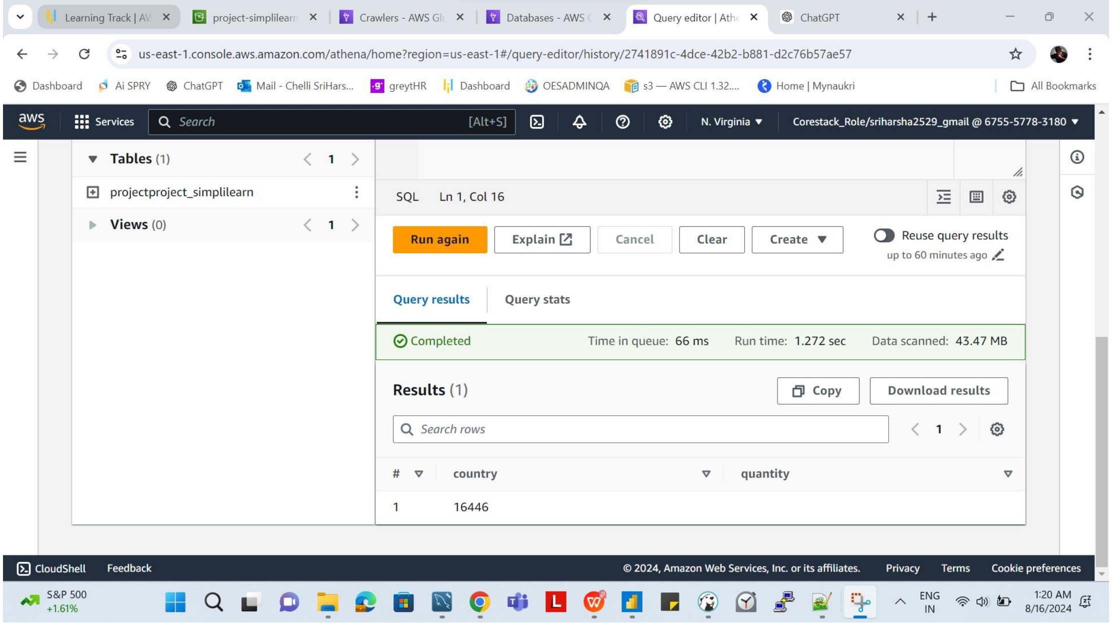


Data source: AwsDataCatalog

Database: batchprocessing

```
Select country,quantity from "batchprocessing"."projectproject_simplilearn"
where stockcode = '23843' group by country, quantity order by
quantity DESC
```

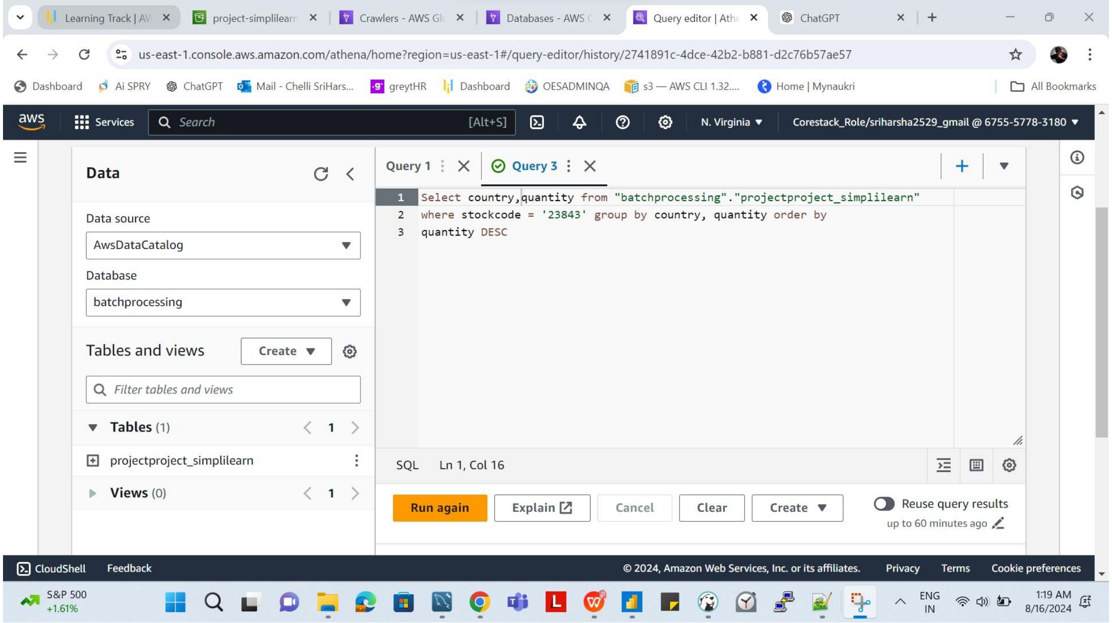
Batch Time Analysis of Transactional Data



Completed   Time in queue: 66 ms   Run time: 1.272 sec   Data scanned: 43.47 MB

Results (1)

#	country	quantity
1	16446	

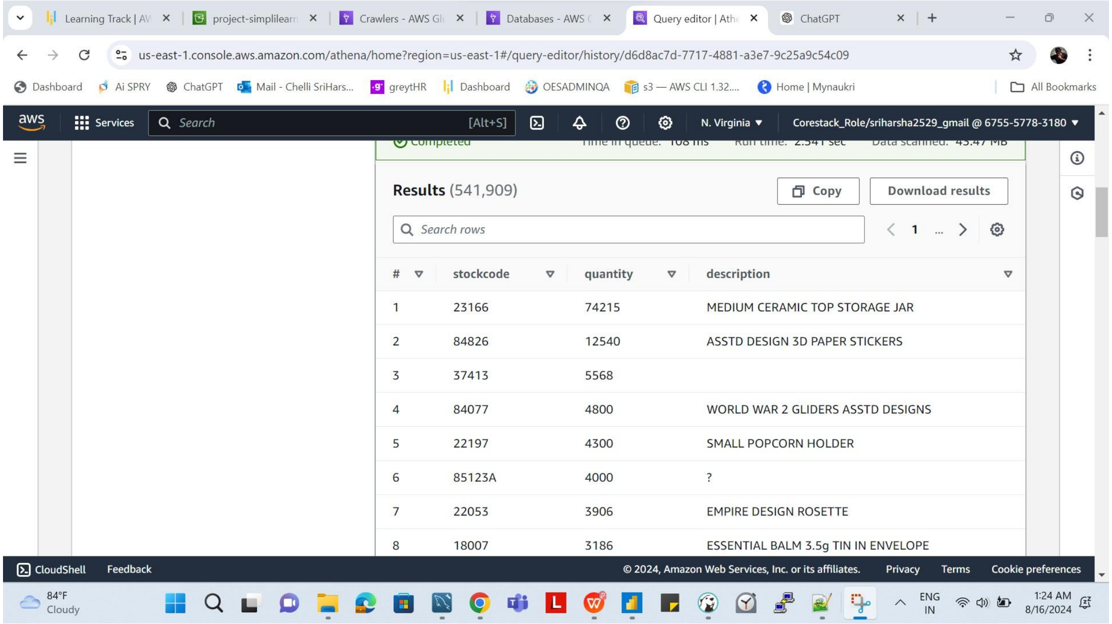


Data source: AwsDataCatalog

Database: batchprocessing

```
Select country,quantity from "batchprocessing"."projectproject_simplilearn"
where stockcode = '23843' group by country, quantity order by
quantity DESC
```

Batch Time Analysis of Transactional Data

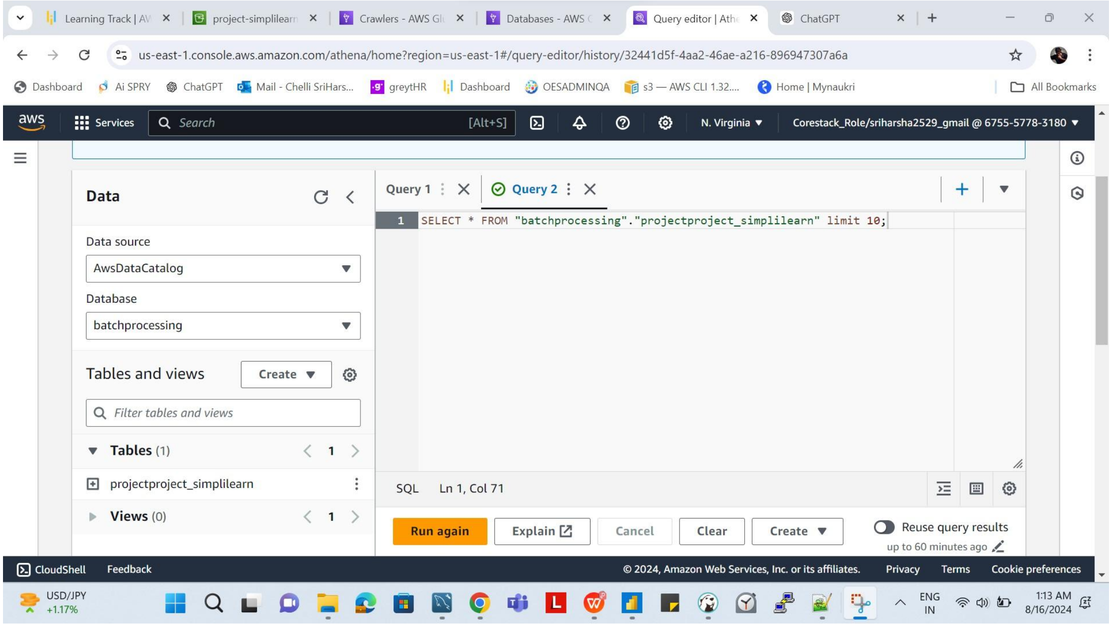


us-east-1.console.aws.amazon.com/athena/home?region=us-east-1#/query-editor/history/d6d8ac7d-7717-4881-a3e7-9c25a9c54c09

Completed Time in queue: 108 ms Run time: 2.341 sec Data scanned: 43.47 MB

Results (541,909)

#	stockcode	quantity	description
1	23166	74215	MEDIUM CERAMIC TOP STORAGE JAR
2	84826	12540	ASSTD DESIGN 3D PAPER STICKERS
3	37413	5568	
4	84077	4800	WORLD WAR 2 GLIDERS ASSTD DESIGNS
5	22197	4300	SMALL POPCORN HOLDER
6	85123A	4000	?
7	22053	3906	EMPIRE DESIGN ROSETTE
8	18007	3186	ESSENTIAL BALM 3.5g TIN IN ENVELOPE



us-east-1.console.aws.amazon.com/athena/home?region=us-east-1#/query-editor/history/32441d5f-4aa2-46ae-a216-896947307a6a

Data source: AwsDataCatalog

Database: batchprocessing

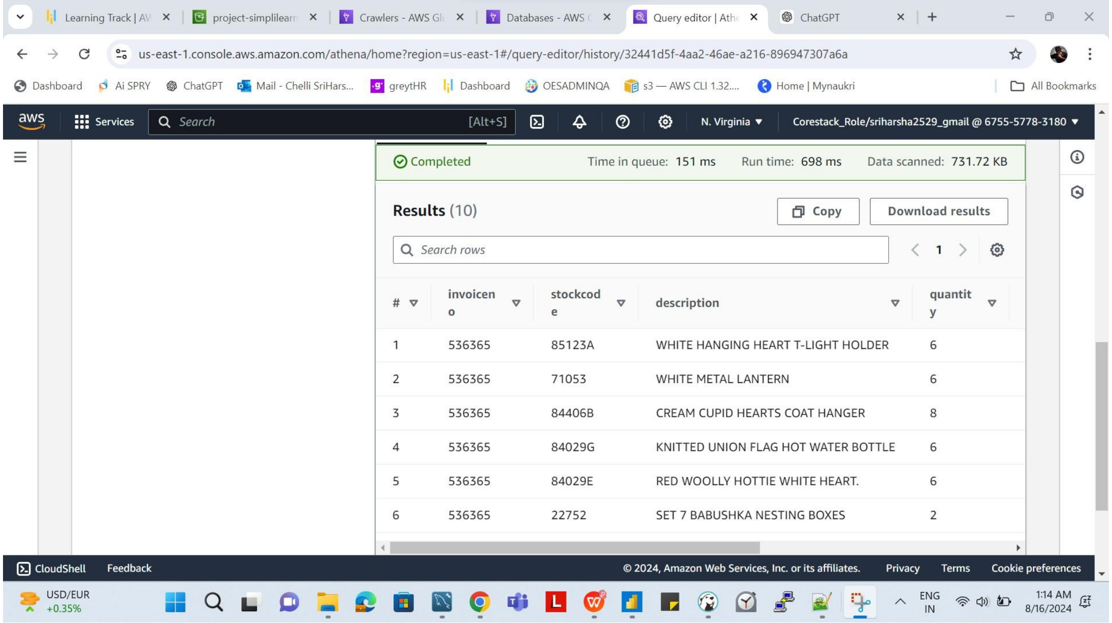
Tables (1): projectproject_simplilearn

Views (0)

```
SELECT * FROM "batchprocessing"."projectproject_simplilearn" limit 10;
```

Step 11 Quering

Batch Time Analysis of Transactional Data



Completed Time in queue: 151 ms Run time: 698 ms Data scanned: 731.72 KB

Results (10)

#	invoiceno	stockcode	description	quantity
1	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6
2	536365	71053	WHITE METAL LANTERN	6
3	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8
4	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6
5	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6
6	536365	22752	SET 7 BABUSHKA NESTING BOXES	2