

# 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.

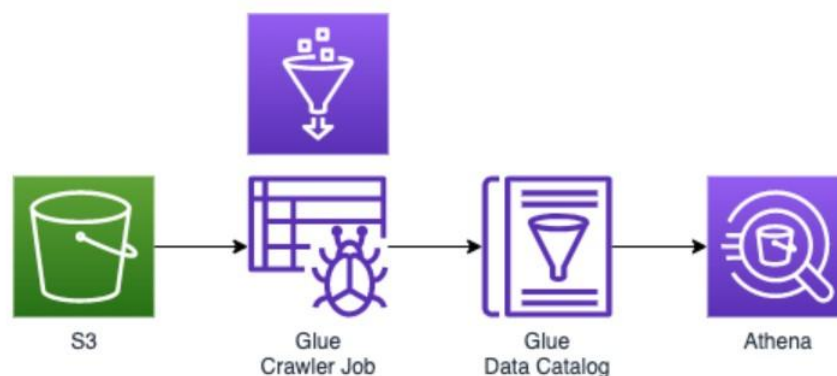
You are a data analytics consultant, and you're asked to provide valuable insights and statistics across products, brands, categories, segments to the marketing, product, sales, and procurement teams and inform them about which product has the highest amount of sales and which product and its marketing 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

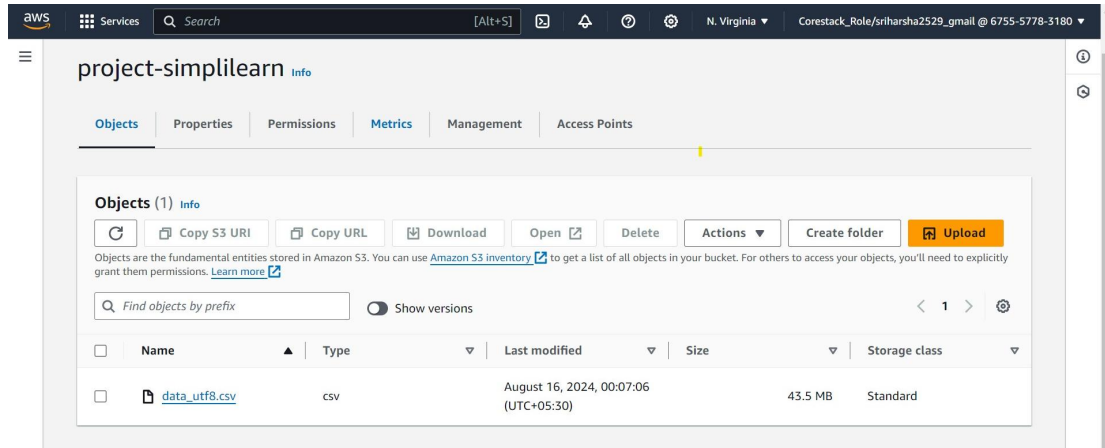
## Steps to perform

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.

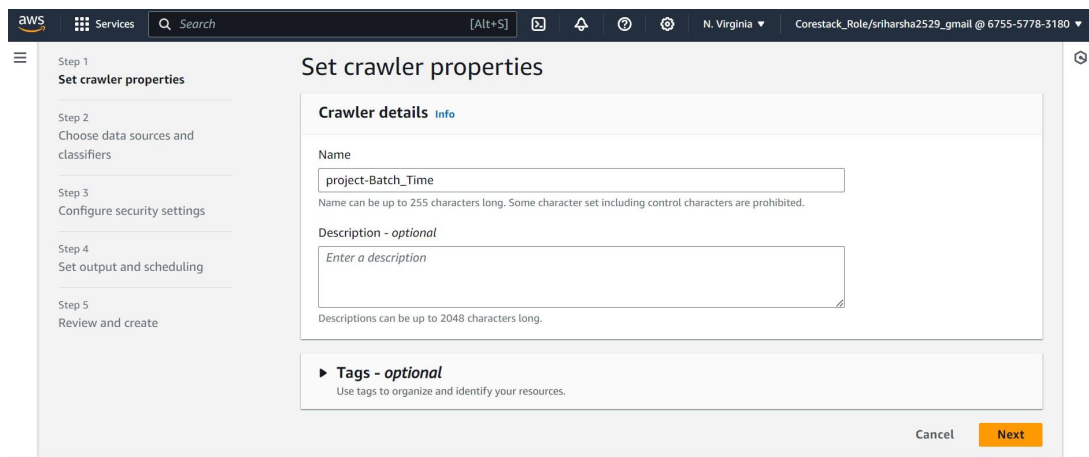


# Batch Time Analysis of Transactional Data

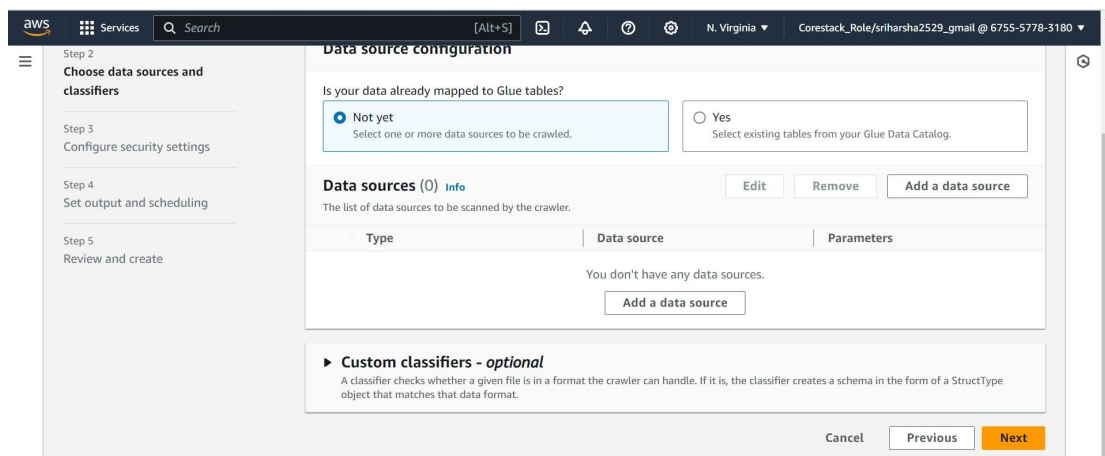
Create a S3 Bucket and loaded data



Step 2 : Created A crawler



Step 3: Add a Data source from S3



# Batch Time Analysis of Transactional Data

## Step 4 : Data Source Added from S3

The screenshot shows the AWS Glue console interface for Step 4, "Choose data sources and classifiers". The left sidebar lists the steps: Step 1 (Set crawler properties), Step 2 (Choose data sources and classifiers), Step 3 (Configure security settings), Step 4 (Set output and scheduling), and Step 5 (Review and create). The main content area is titled "Choose data sources and classifiers" and contains a "Data source configuration" section. This section asks "Is your data already mapped to Glue tables?" with two radio buttons: "Not yet" (selected) and "Yes". Below this, there is a table titled "Data sources (1)" with columns "Type", "Data source", and "Parameters". The table contains one entry: "S3" with data source "s3://project-simplilearn" and parameters "Recrawl all". There are buttons for "Edit", "Remove", and "Add a data source". At the bottom, there is a section for "Custom classifiers - optional" with a description of what a classifier does.

**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
<input checked="" type="radio"/> 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

The screenshot shows the AWS Glue console interface for Step 5, "Set output and scheduling". The left sidebar lists the steps: Step 1 (Set crawler properties), Step 2 (Choose data sources and classifiers), Step 3 (Configure security settings), Step 4 (Set output and scheduling), and Step 5 (Review and create). The main content area is titled "Set output and scheduling" and contains an "Output configuration" section. This section has a "Target database" dropdown menu with the text "Choose a database" and a refresh button. Below this are buttons for "Clear selection" and "Add database". A red error message "Target database is required" is displayed. There is also a "Table name prefix - optional" text input field with the placeholder "Type a prefix added to table names". Below that is a "Maximum table threshold - optional" text input field with the placeholder "Type a number greater than 0". At the bottom, there is a section for "Advanced options".

**Set output and scheduling**

**Output configuration** [Info](#)

Target database

Choose a database

Clear selection Add 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

This screenshot is identical to the previous one, showing the "Set output and scheduling" step. However, the "Target database" dropdown menu now shows "batchprocessing" as the selected option. The error message "Target database is required" is no longer present. The "Table name prefix" is set to "project" and the "Maximum table threshold" is still empty.

**Set output and scheduling**

**Output configuration** [Info](#)

Target database

batchprocessing

Clear selection Add database

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

The screenshot shows the 'Set output and scheduling' step in the AWS Glue console. The left sidebar contains a navigation menu with 'Step 4: Set output and scheduling' selected, and 'Step 5: Review and create' below it. The main content area has a 'Table name prefix - optional' field with the value 'project'. Below it is a 'Maximum table threshold - optional' field with the placeholder text 'Type a number greater than 0'. A section titled 'Advanced options' is collapsed. The 'Crawler schedule' section is expanded, showing a 'Frequency' dropdown menu set to 'On demand'. At the bottom right are 'Cancel', 'Previous', and 'Next' buttons. The footer includes 'CloudShell', 'Feedback', and copyright information for Amazon Web Services.

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

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

Cancel Previous Next

CloudShell Feedback © 2024, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences

## Step 8 iam Rule creation

The screenshot shows the 'Configure security settings' step in the AWS Glue console. The left sidebar contains a navigation menu with 'Step 3: Configure security settings' selected, and 'Step 4: Set output and scheduling' below it. The main content area has a section titled 'IAM role' with an 'Info' link. It shows an 'Existing IAM role' dropdown menu with the value 's3-glue-rule' and a 'View' link. Below it are 'Create new IAM role' and 'Update chosen IAM role' buttons. A note states: 'Only IAM roles created by the AWS Glue console and have the prefix "AWSGlueServiceRole-" can be updated.' The 'Lake Formation configuration - optional' section is expanded, showing a checkbox for 'Use Lake Formation credentials for crawling S3 data source'. Below it is a note: '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.' At the bottom is a section titled 'Security configuration - optional'.

Configure security settings

IAM role Info

Existing IAM role  
s3-glue-rule View

Create new IAM role Update chosen IAM role

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

The screenshot shows the AWS Glue console interface. At the top, there's a navigation bar with the AWS logo, 'Services', a search bar, and user information. Below this, a sidebar on the left contains a hamburger menu. The main content area is divided into two steps:

**Step 3: Configure security settings** (with an 'Edit' button)

Configure security settings

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

**Step 4: Set output and scheduling** (with an 'Edit' button)

Set output and scheduling

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

At the bottom right, there are three buttons: 'Cancel', 'Previous', and 'Create crawler' (highlighted in orange).

## Step 10 Athena

The screenshot shows the AWS Athena console interface. At the top, there's a navigation bar with the AWS logo, 'Services', a search bar, and user information. Below this, a sidebar on the left contains a hamburger menu. The main content area is divided into two sections:

**Data** (with a refresh icon and a back arrow)

Data source: AwsDataCatalog

Database: batchprocessing

Tables and views: Create (dropdown), Settings (gear icon)

Filter tables and views: (search bar)

**Tables (1)** (with a back arrow and a forward arrow)

projectproject\_simplilearn (with a dropdown menu)

**Views (0)** (with a back arrow and a forward arrow)

Below the table list, there's a query editor area with a tab labeled 'Query 1'. The editor shows a single line of SQL: 'Ln 1, Col 1'. At the bottom of the editor, there are buttons: 'Run', 'Explain', 'Cancel', 'Clear', and 'Create' (dropdown). To the right of the buttons, there's a toggle switch for 'Reuse query results up to 60 minutes ago'.

# Batch Time Analysis of Transactional Data

## Step 11 Querying

The screenshot shows the AWS Athena Query Editor interface. On the left, the 'Data' pane shows the 'Data source' as 'AwsDataCatalog' and the 'Database' as 'batchprocessing'. The 'Tables and views' section shows a table named 'projectproject\_simplilearn'. The main query editor displays the following SQL query:

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

At the bottom, there are buttons for 'Run again', 'Explain', 'Cancel', 'Clear', and 'Create'. A status bar indicates 'Reuse query results up to 60 minutes ago'.

## Step 12 result

The screenshot shows the AWS Athena Query Editor interface displaying the results of Query 2. The status bar indicates 'Completed' with a 'Time in queue: 151 ms', 'Run time: 698 ms', and 'Data scanned: 731.72 KB'. The 'Results (10)' section shows a table with 6 rows and 5 columns: '#', 'invoice number', 'stockcode', 'description', and 'quantity'.

#	invoice number	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

The screenshot shows the AWS Athena Query Editor interface displaying Query 3. The 'Data' pane shows the 'Data source' as 'AwsDataCatalog' and the 'Database' as 'batchprocessing'. The main query editor displays the following SQL query:

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

At the bottom, there are buttons for 'Run again', 'Explain', 'Cancel', 'Clear', and 'Create'. A status bar indicates 'Reuse query results up to 60 minutes ago'.

# Batch Time Analysis of Transactional Data

The screenshot shows the AWS Athena Query Editor interface. The left sidebar displays the 'Tables (1)' section with 'projectproject\_simplilearn' and 'Views (0)'. The main area shows the SQL query editor with the text 'SQL Ln 1, Col 16'. Below the editor, the 'Query results' tab is active, showing a 'Completed' status with 'Time in queue: 66 ms', 'Run time: 1.272 sec', and 'Data scanned: 43.47 MB'. The 'Results (1)' section shows a table with one row and two columns: 'country' and 'quantity'.

#	country	quantity
1	16446	

The screenshot shows the AWS Athena Query Editor interface. The left sidebar displays the 'Data' section with 'Data source' set to 'AwsDataCatalog', 'Database' set to 'batchprocessing', and 'Tables and views' section. The main area shows the SQL query editor with the text 'SQL Ln 1, Col 16'. Below the editor, the 'Query results' tab is active, showing a 'Query 3' status with 'Time in queue: 66 ms', 'Run time: 1.272 sec', and 'Data scanned: 43.47 MB'. The 'Results (1)' section shows a table with one row and two columns: 'country' and 'quantity'.

#	country	quantity
1	16446	

# Batch Time Analysis of Transactional Data

The screenshot shows the AWS Athena console displaying the results of a query. The query has completed, with a time in queue of 108 ms and a run time of 2.341 sec. The results are shown in a table with 541,909 rows. The table has four columns: #, stockcode, quantity, and description. The first 8 rows of data are as follows:

#	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

The screenshot shows the AWS Athena console with the query editor open. The query is as follows:

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

The console also shows the data source as AwsDataCatalog, the database as batchprocessing, and the table as projectproject\_simplilearn. The query is ready to be executed.

## Step 11 Quering



# Batch Time Analysis of Transactional Data

Learning Track | AV

project-simplilearn

Crawlers - AWS Glue

Databases - AWS Glue

Query editor | Athena

ChatGPT

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

DashboardAi SPRYChatGPTMail - Chelli SriHars...greythRDashboarDOESADMINQA s3 — AWS CLI 1.32...Home | MynaukriAll Bookmarks

awsServicesSearch[Alt+S]

N. VirginiaCorestack\_Role/sriharsha2529\_gmail @ 6755-5778-3180

CompletedTime in queue: 151 msRun time: 698 msData scanned: 731.72 KB

Results (10)

CopyDownload results

Search rows

#	invoice no	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

CloudShellFeedback

© 2024, Amazon Web Services, Inc. or its affiliates. PrivacyTermsCookie preferences

USD/EUR+0.35%WindowsTaskbarIconsSystem Tray1:14 AM 8/16/2024