

Batch Time Analysis of Transactional Data

Christina Gryś

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 (ensure that the file is in UTF-8 format only)
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

Create a Bucket: **aws-grysc**

Amazon S3

Successfully created bucket "aws-grysc"
To upload files and folders, or to configure additional bucket settings choose [View details](#).

Replicate your data to any storage class to save on storage costs. [Get started](#)

Amazon S3 > Buckets

Account snapshot
Last updated: May 10, 2022 by Storage Lens. Metrics are generated every 24 hours. [Learn more](#)

[View Storage Lens dashboard](#)

Total storage	Object count	Avg. object size	You can enable advanced metrics in the "default-account-dashboard" configuration.
46.9 MB	3	15.6 MB	

Buckets (1) [Info](#)

Buckets are containers for data stored in S3. [Learn more](#)

	Name	AWS Region	Access	Creation date
<input type="radio"/>	aws-grysc	US East (N. Virginia) us-east-1	Bucket and objects not public	May 17, 2022, 19:47:45 (UTC-04:00)

Create A Folder: dataset.csv

The screenshot shows the AWS Management Console interface. At the top, a green notification banner states: "Successfully created folder 'dataset.csv'. Operation successfully completed." Below this, a blue banner prompts the user to "Replicate your data to any storage class to save on storage costs." The main content area displays the "aws-grysc" bucket details. The "Objects" tab is selected, showing a list of objects. A table lists the objects, with one entry: "dataset.csv/" of type "Folder". The table has columns for Name, Type, Last modified, Size, and Storage class. The left sidebar shows the "Amazon S3" service selected, with options for Buckets, Access Points, and Storage Lens.

Upload the provided file: data_utf8.csv

The screenshot shows the "Upload: status" page in the AWS Management Console. A green notification banner at the top says "Upload succeeded". The page displays a summary of the upload: "Destination: s3://aws-grysc/dataset.csv/", "Succeeded: 1 file, 43.5 MB (100.00%)", and "Failed: 0 files, 0 B (0%)". Below the summary, the "Files and folders" tab is selected, showing a table of the uploaded file. The table has columns for Name, Folder, Type, Size, Status, and Error. The entry "data_utf8.csv" is listed with a status of "Succeeded".

Add Database: lenodo

The screenshot shows the AWS Glue console. The "Databases" section is active, displaying a list of databases. A table lists the databases, with one entry: "lenodo". The table has columns for Name and Description. The left sidebar shows the "AWS Glue" service selected, with options for Data catalog, Databases, Tables, Connections, Crawlers, and Classifiers.

Add Crawler: lenodo-crawler

aws

Services

Search for services, features, blogs, docs, and more

[Option+S]

📄

🔔

🔍

N. Virginia

Corestack_Role/grysc@gmail 1868-3307-7210

AWS Glue

Data catalog

Databases

Tables

Connections

Crawlers

Classifiers

Schema registries

Schemas

Settings

Crawlers

A crawler connects to a data store, progresses through a prioritized list of classifiers to determine the schema for your data, and then creates metadata tables in your data catalog.

Crawler lenodo-crawler was created to run on demand. [Run it now?](#)

[User preferences](#)

Add crawler

Run crawler

Action

Filter by tags and attributes

Showing: 1 - 1

🔄

🔍

<input checked="" type="checkbox"/>	Name	Schedule	Status	Logs	Last runtime	Median runtime	Tables updated	Tables added
<input checked="" type="checkbox"/>	lenodo-crawler		Ready		0 secs	0 secs	0	0

Services, features, blogs, docs, and more

[Option+S]

📄

🔔

🔍

N

Crawlers > lenodo-crawler

Run crawler

Edit

Name	lenodo-crawler
Description	
Create a single schema for each S3 path	false
Security configuration	
Tags	-
State	Ready
Schedule	
Last updated	Tue May 17 19:55:52 GMT-400 2022
Date created	Tue May 17 19:55:52 GMT-400 2022
Database	lenodo
Table level	
Service role	service-role/AWSGlueServiceRole-lenodo-project
Selected classifiers	
Data store	S3
Include path	s3://aws-grysc/dataset.csv
Connection	
Exclude patterns	
Configuration options	
Schema updates in the data store	Update the table definition in the data catalog.
Object deletion in the data store	Mark the table as deprecated in the data catalog.

Crawlers A crawler connects to a data store, progresses through a prioritized list of classifiers to determine the schema for your data, and then creates metadata tables in your data catalog.

[Add crawler](#)
[Run crawler](#)
Action ▾

Showing: 1 - 1 < > ↺ ⓘ

<input type="checkbox"/>	Name	Schedule	Status	Logs	Last runtime	Median runtime	Tables updated	Tables added
<input type="checkbox"/>	lenodo-crawler		Ready	Logs	52 secs	52 secs	0	1

Tables A table is the metadata definition that represents your data, including its schema. A table can be used as a source or target in a job definition.

Add tables ▾
Action ▾

Save view ▾
Showing: 1 - 1 < > ↺ ⚙ ⓘ

<input type="checkbox"/>	Name ▾	Database ▾	Location ▾	Classification	Last updated ▾	Deprecated ▾
<input type="checkbox"/>	dataset_csv	lenodo	s3://aws-grsync/dataset.csv/	csv	17 May 2022 7:57 AM UTC-4	

AWS
Search for services, features, blogs, docs, and more
[Option+S]

AWS Glue

- Data catalog
- Databases
- Tables
- Connections
- Crawlers
- Classifiers
- Schema registries
- Schemas
- Settings

- ETL
- AWS Glue Studio [↗](#)
- Jobs [↗](#) - New
- Jobs (legacy)
- ML Transforms
- Blueprints
- Workflows
- Triggers
- Dev endpoints
- Notebooks
- Interactive Session - New

- Security
- Security configurations

- Tutorials
- Add crawler
- Explore table
- Add job
- Resources [↗](#)
- What's new 10+

Last updated 17 May 2022 07:57 PM **Table** Version (Current version) ▼

Edit table Delete table
View properties
Compare versions
Edit schema

Name dataset_csv

Description

Database lenodo

Classification csv

Location s3://aws-grysc/dataset.csv

Connection

Deprecated No

Last updated Tue May 17 19:57:06 GMT-400 2022

Input format org.apache.hadoop.mapred.TextInputFormat

Output format org.apache.hadoop.hive.q.l.o.HiveIgnoreKeyTextOutputFormat

Serde serialization lib org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe

Serde parameters field.delim ,

Table properties	
CrawlerSchemaSerializerVersion	1.0 recordCount 555861 averageRecordSize 82 CrawlerSchemaDeserializerVersion 1.0
compressionType	none columnsOrdered true areColumnsQuoted false delimiter , typeOfData file

Schema

Showing: 1 - 8 of 8 < >

	Column name	Data type	Partition key	Comment
1	invoiceno	string		
2	stockcode	string		
3	description	string		
4	quantity	bigint		
5	invokedate	string		
6	unitprice	double		
7	customerid	bigint		
8	country	string		

Feedback
Looking for language selection? Find it in the new Unified Settings [↗](#)

© 2022 Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences

Configurations of the data source, transformation layer, and data target all added.

aws Services Search for services, features, blogs, docs, and more [Option+S] N. Virginia Corestack_Role/grys1cm_gmail @ 1868-3307-7210

CSV-to-Parquet Last modified on 5/17/2022, 8:06:17 PM Save Delete Run

Successfully updated Job
Successfully updated job dataset_csv. To run the job choose the Run Job button.

Visual Script Job details Runs Schedules

Source Transform Target Undo Redo Remove

Node properties Data target properties - S3 Output schema Data preview

Format
Parquet

Compression Type
Snappy

S3 Target Location
Choose an S3 location in the format s3://bucket/prefix/object/ with a trailing slash (/).
s3://aws-grysc/dataset.csv/ View Browse S3

Data Catalog update options Info
Choose how you want to update the Data Catalog table's schema and partitions. These options will only apply if the Data Catalog table is an S3 backed source.
☒ Do not update the Data Catalog
☐ Create a table in the Data Catalog and on subsequent runs, update the schema and add new partitions
☐ Create a table in the Data Catalog and on subsequent runs, keep existing schema and add new partitions

Partition keys - optional
Add partition keys.
Add a partition key

Visual workflow diagram showing the data flow: Data source - S3 bucket Amazon S3 → Transform - ApplyMapping Apply Mapping → Data target - S3 bucket Amazon S3.

Run the job

aws Services Search for services, features, blogs, docs, and more [Option+S] N. Virginia Corestack_Role/grys1cm_gmail @ 1868-3307-7210

CSV-to-Parquet Last modified on 5/17/2022, 8:06:17 PM Save Delete Run

Successfully started Job
Successfully started job CSV-to-Parquet. Navigate to [Run Details](#) for more details.

Visual Script Job details Runs Schedules

Recent job runs (5) Info

Find runs

May 17, 2022 8:42:28 PM Stop Job run

Job name	Id	Run status	Glue version
CSV-to-Parquet	jr_aca14b002c5dedb994e8e167be0a1779f9ff59372947e2be6780553420eac8f4	Running	3.0
Retry attempt number	Start time	End time	Start-up time
Initial run	May 17, 2022 8:42:28 PM		0
Execution time	Last modified on	Trigger name	Security configuration
0 seconds	May 17, 2022 8:42:28 PM	-	-
Timeout	Max capacity	Number of workers	Worker type
2880 minutes	10 DPUs	10	G.1X
Execution class	Log group name	Cloudwatch logs	Performance and debugging recommendations
-	/aws-glue/jobs	<ul style="list-style-type: none">All logsOutput logsError logs	<ul style="list-style-type: none">View in CloudWatch

Athena Query: Output Final Results

The screenshot shows the Amazon Athena console interface. On the left, the 'Query editor' sidebar is visible with options for 'Workgroups', 'Data sources', and 'Jobs'. The main area is titled 'Editor' and shows a query execution summary for 'Query 1'. The query is a simple SELECT statement: `select * from dataset_csv limit 10;`. The execution status is 'Completed' with a run time of 0.997 seconds and 153.85 KB of data scanned. Below the summary, the 'Results (10)' are displayed in a table format. The table has columns: #, Invoiceno, stockcode, description, quantity, and Invoicedate. The first row of results is visible, showing data for 'from auctonius transforme immet'.

#	Invoiceno	stockcode	description	quantity	Invoicedate
1			from auctonius transforme immet		

Results: Best Selling Item = World War 2 Gliders Asstd Designs and sold: 53847 copies

The screenshot shows the Amazon Athena console interface. On the left, the 'Query editor' sidebar is visible. The main area is titled 'Editor' and shows a query execution summary for 'Query 1'. The query is a more complex SELECT statement: `-- BEST-SELLING ITEM
SELECT
stockcode,
Description,
SUM(Quantity) AS total_sold_items
FROM dataset_csv
GROUP BY 1, 2
ORDER BY 3 DESC
LIMIT 1`. The execution status is 'Completed' with a run time of 1.204 seconds and 44.03 MB of data scanned. Below the summary, the 'Results (1)' are displayed in a table format. The table has columns: #, stockcode, Description, and total_sold_items. The first row of results is visible, showing data for '84077 WORLD WAR 2 GLIDERS ASSTD DESIGNS' with a total of 53847 copies sold.

#	stockcode	Description	total_sold_items
1	84077	WORLD WAR 2 GLIDERS ASSTD DESIGNS	53847

Results for Countries where customers have bought the most-sold item

Amazon Athena

Query editor

Workgroups

Data sources

Jobs

Workflows [New](#)

Powered by Step Functions

Enable compact mode

Editor

Recent queries

Saved queries

Settings

Workgroup: primary

Data

Data source: AwsDataCatalog

Database: lenodo

Tables and views

Create

Filter tables and views

Tables (1)

dataset_csv

invoiceno: string

stockcode: string

description: string

quantity: bigint

invoicedate: string

unitprice: double

customerid: bigint

country: string

Views (0)

Query 1

```
1 --COUNTRIES WHERE CUSTOMERS HAVE BOUGHT THE MOST-SOLD ITEM
2 SELECT
3   country,
4   SUM(quantity) AS total_sold_items
5 FROM dataset_csv
6 WHERE stockcode = '84077'
7 GROUP BY 1
8 ORDER BY 2 DESC
9 ;
```

SQL Ln 3, Col 13

Run again Cancel Save Clear Create

Completed Time in queue: 0.137 sec Run time: 1.05 sec Data scanned: 44.03 MB

Results (14)

Copy Download results

Search rows

#	country	total_sold_items
1	United Kingdom	48326

SQL Ln 3, Col 13

Run again Cancel Save Clear Create

Completed Time in queue: 0.137 sec Run time: 1.05 sec Data scanned: 44.03 MB

Results (14)

Copy Download results

Search rows

#	country	total_sold_items
1	United Kingdom	48326
2	Sweden	2304
3	EIRE	816
4	Japan	577
5	France	528
6	Spain	288
7	Canada	288
8	Denmark	144
9	Switzerland	144
10	Hong Kong	144
11	Germany	96
12	Unspecified	96
13	Portugal	48
14	Norway	48