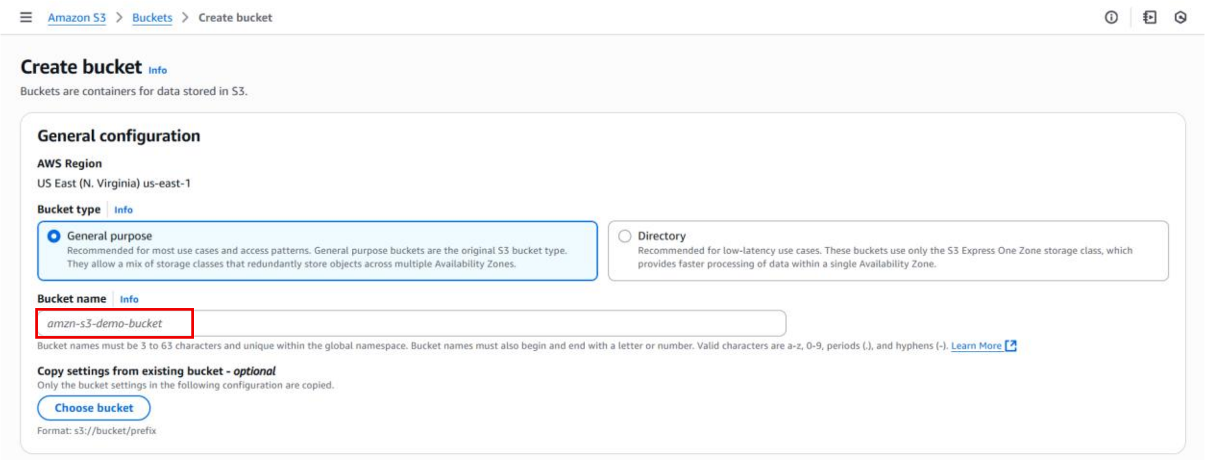
Partitioning with Glue

To Begin with the Lab

Summary of the Lab

In this lab, you learn how to implement **data partitioning using AWS Glue and Athena**. You create an S3 bucket with subfolders (London, New York, Tokyo) representing partitions and upload corresponding sales data. Using AWS Glue, a crawler is configured to detect data and automatically create partitions in the Glue Data Catalog. Then, in Athena, you query the data and update metadata using the **ALTER TABLE ADD PARTITION** command to include new folders like Tokyo.

- **Create an S3 Bucket**
- Go to **AWS Management Console → S3 → Create bucket**.
- Enter a unique name.
- Keep other settings default and **create the bucket**.



Create bucket

Buckets are containers for data stored in S3.

General configuration

AWS Region

US East (N. Virginia) us-east-1

Bucket type

● General purpose
Recommended for most use cases and access patterns. General purpose buckets are the original S3 bucket type. They allow a mix of storage classes that redundantly store objects across multiple Availability Zones.

○ Directory
Recommended for low-latency use cases. These buckets use only the S3 Express One Zone storage class, which provides faster processing of data within a single Availability Zone.

Bucket name

amzn-s3-demo-bucket

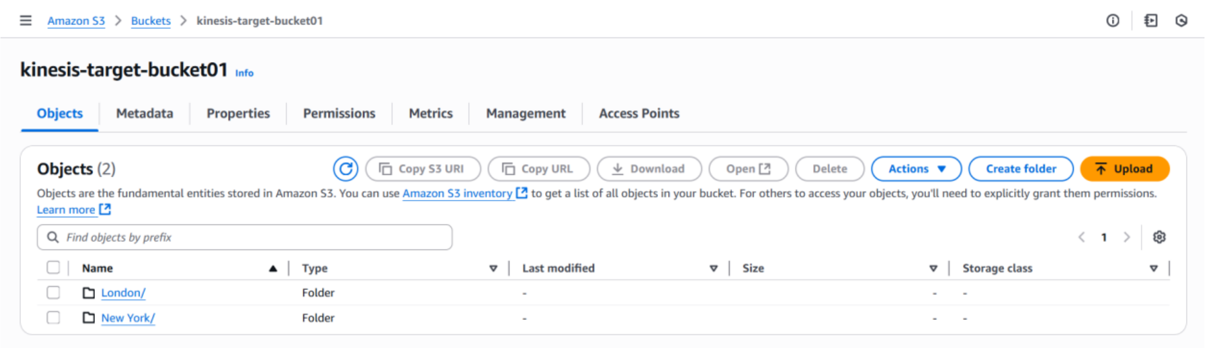
Bucket names must be 3 to 63 characters and unique within the global namespace. Bucket names must also begin and end with a letter or number. Valid characters are a-z, 0-9, periods (.), and hyphens (-). Learn More

Copy settings from existing bucket - *optional*

Only the bucket settings in the following configuration are copied.

Format: s3://bucket/prefix

- Inside the bucket, create folders for each location:
  - London/
  - NewYork/
- These folders will act as **partition keys**.



kinesis-target-bucket01

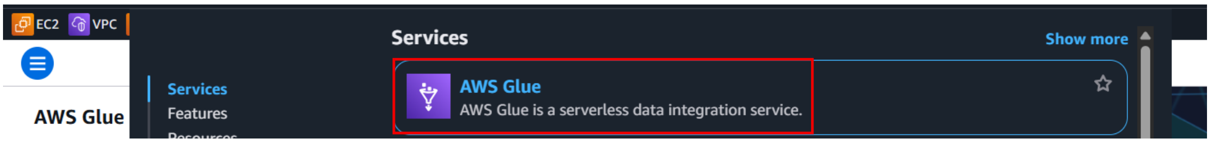
Objects (2)

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
London/	Folder	-	-	-
New York/	Folder	-	-	-

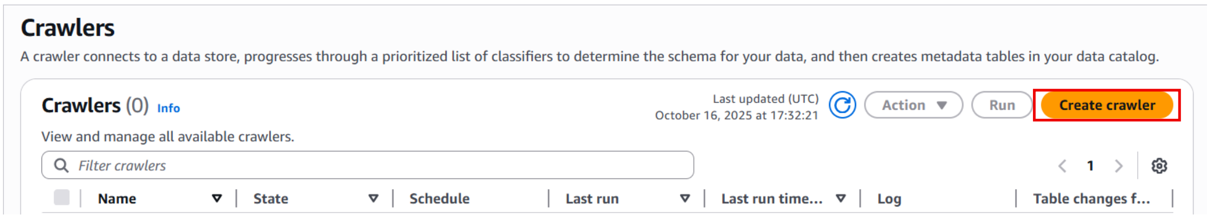
- Upload your sales data files into each folder:
- sales_london.csv → into the **London** folder.

- sales_newyork.csv → into the **New York** folder.
- Go to **AWS Glue → Crawlers → Create crawler**.
- Go to **AWS Console → Glue Service** (ensure you're in the same region as your S3 bucket).



AWS Glue is a serverless data integration service.

- Now Click on the Crawlers and then click on Create crawler.


Crawlers

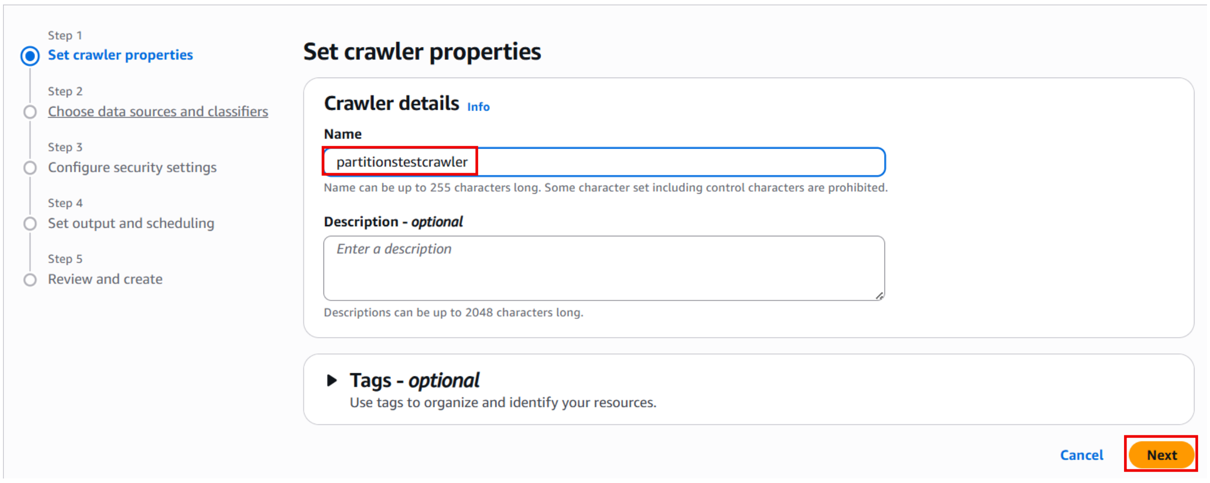
A crawler connects to a data store, progresses through a prioritized list of classifiers to determine the schema for your data, and then creates metadata tables in your data catalog.

Crawlers (0) Info

Last updated (UTC)
October 16, 2025 at 17:32:21

View and manage all available crawlers.

- Name the Crawler and click next.


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 crawler properties

Crawler details Info

Name: partitionstesttcrawler

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

Description - *optional*

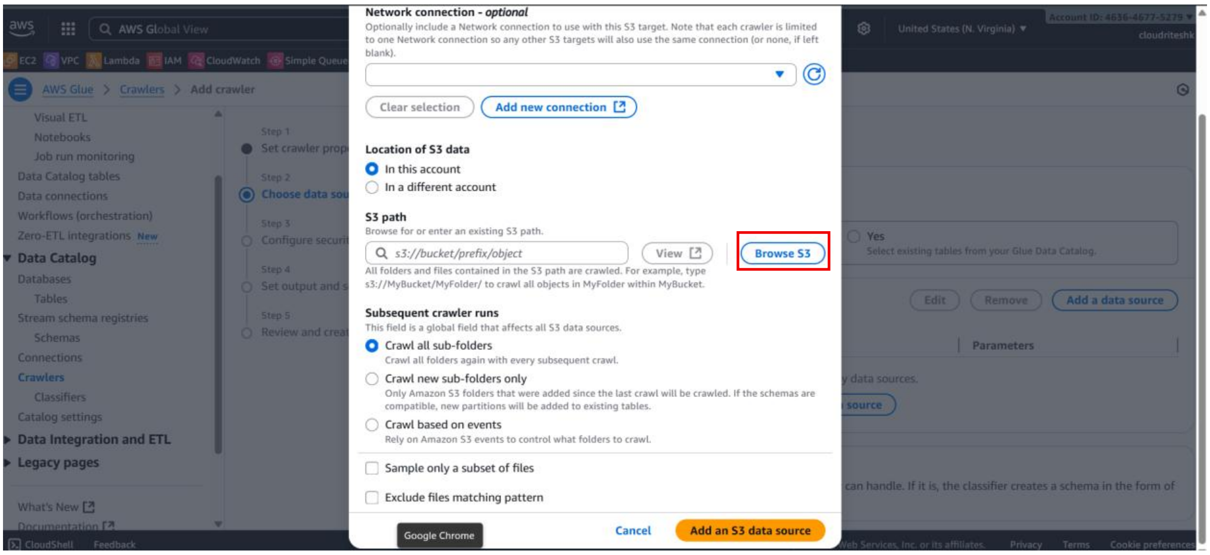
Enter a description

Descriptions can be up to 2048 characters long.

Tags - *optional*

Use tags to organize and identify your resources.

- Now click on the Add a data source.
- Click on the Browse S3



Network connection - *optional*

Optionally include a Network connection to use with this S3 target. Note that each crawler is limited to one Network connection so any other S3 targets will also use the same connection (or none, if left blank).

Clear selection | Add new connection

Location of S3 data

● In this account

○ In a different account

S3 path

Browse for or enter an existing S3 path.

s3://bucket/prefix/object | View | Browse S3

All folders and files contained in the S3 path are crawled. For example, type s3://MyBucket/MyFolder/ to crawl all objects in MyFolder within MyBucket.

Subsequent crawler runs

This field is a global field that affects all S3 data sources.

● Crawl all sub-folders
Crawl all folders again with every subsequent crawl.

○ Crawl new sub-folders only
Only Amazon S3 folders that were added since the last crawl will be crawled. If the schemas are compatible, new partitions will be added to existing tables.

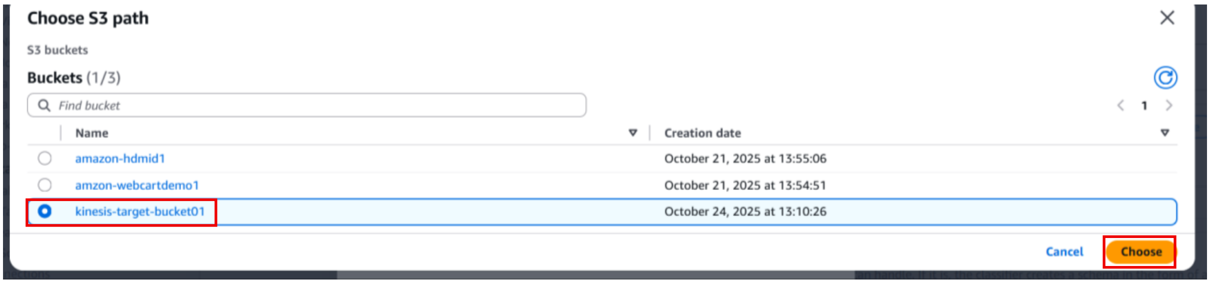
○ Crawl based on events
Rely on Amazon S3 events to control what folders to crawl.

□ Sample only a subset of files

□ Exclude files matching pattern

Cancel | Add an S3 data source

- Choose the folder and click choose.



Choose S3 path

S3 buckets

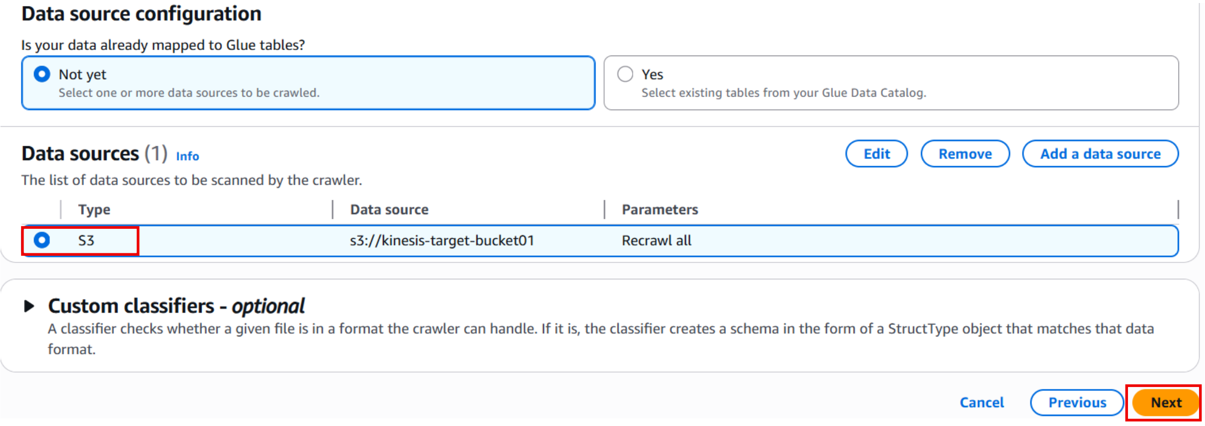
Buckets (1/3)

Find bucket

	Name	Creation date
○	amazon-hdmid1	October 21, 2025 at 13:55:06
○	amzon-webcartdemo1	October 21, 2025 at 13:54:51
●	kinesis-target-bucket01	October 24, 2025 at 13:10:26

Cancel | Choose

- Then, Click on Add an S3 data source.
- Check the Data Sources and Click 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 (1) Info

Edit | Remove | Add a data source

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

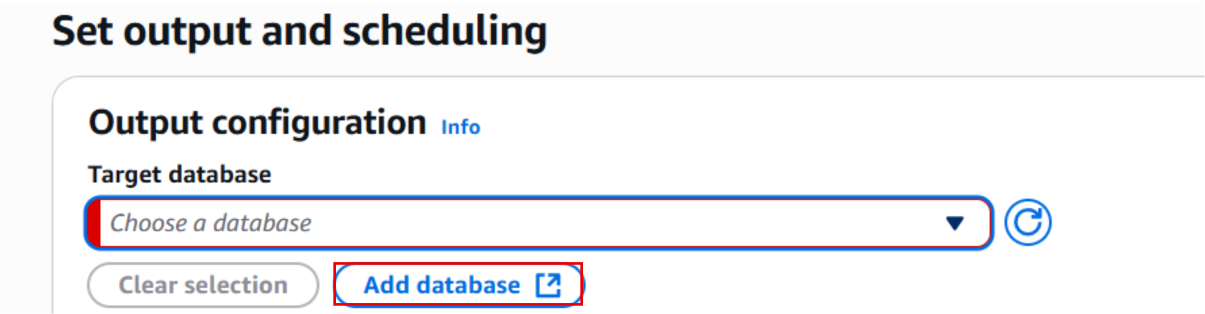
	Type	Data source	Parameters
●	S3	s3://kinesis-target-bucket01	Recrawl all

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.

Cancel | Previous | Next

- Select an existing role with S3 and Glue permissions
- Click on the Add database



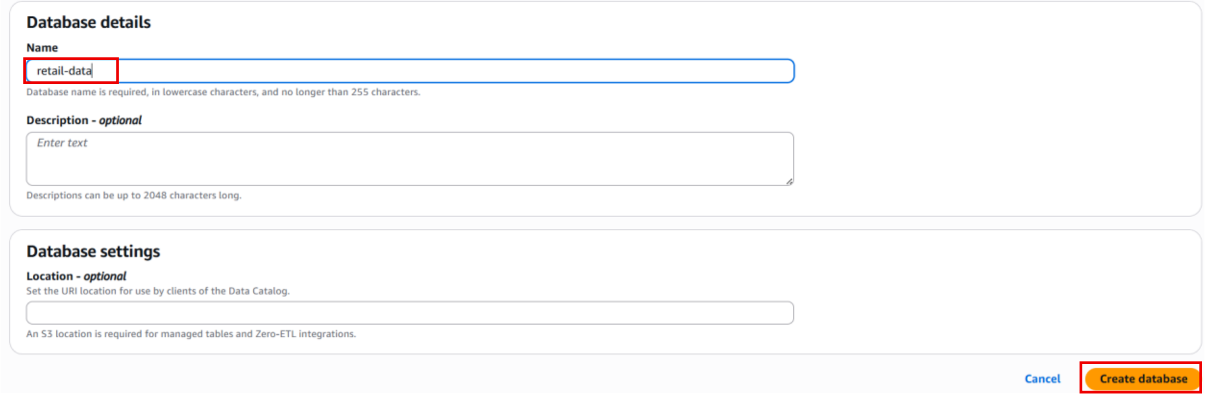
Set output and scheduling

Output configuration Info

Target database

Choose a database

- Give the database name and click Create.



Database details

Name

retail-data

Database name is required, in lowercase characters, and no longer than 255 characters.

Description - *optional*

Enter text

Descriptions can be up to 2048 characters long.

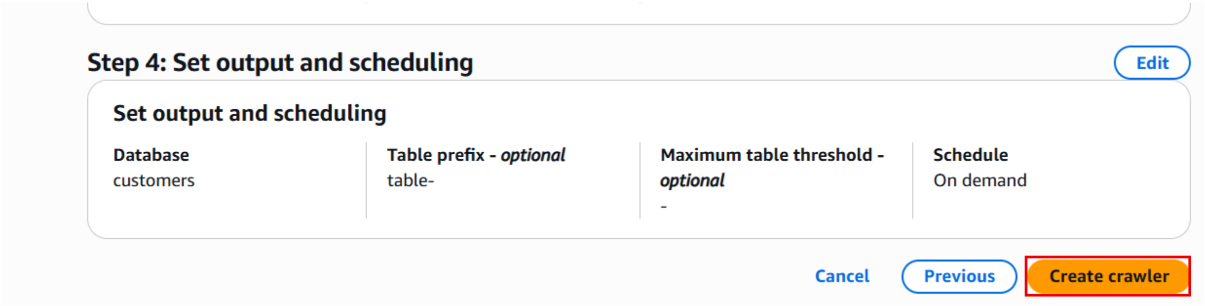
Database settings

Location - *optional*

Set the URI location for use by clients of the Data Catalog.

An S3 location is required for managed tables and Zero-ETL integrations.

- Now go back to the clawer setup and refresh, you will see the database.
- Optional add a **table prefix** and click Next.
- Then Click on Create crawler.

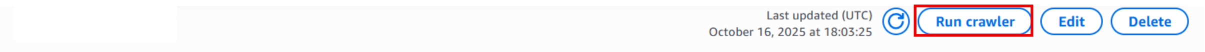


Step 4: Set output and scheduling

Set output and scheduling

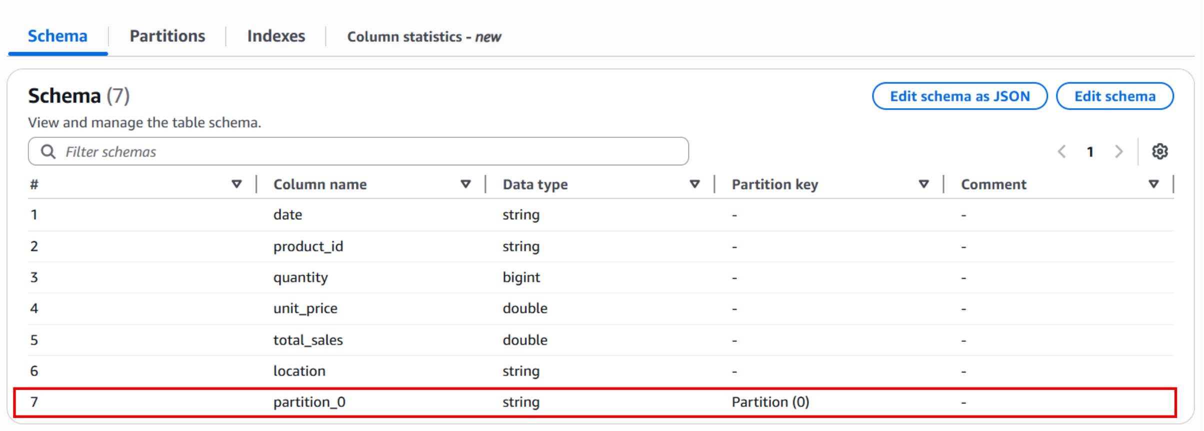
Database	Table prefix - *optional*	Maximum table threshold - *optional*	Schedule
customers	table-	-	On demand

- After the crawler is create, Click on the Run Crawler.



Last updated (UTC)
October 16, 2025 at 18:03:25

- Go to **AWS Glue → Tables**.
- Open your new table and verify:
  - Columns detected from CSV files.
  - **Partition key** (e.g., partition_0) added automatically.

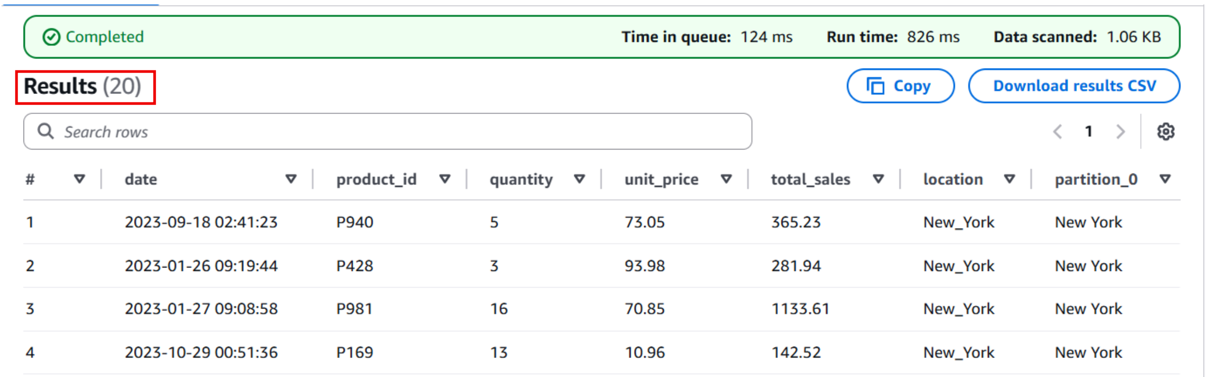


Schema (7)

View and manage the table schema.

#	Column name	Data type	Partition key	Comment
1	date	string	-	-
2	product_id	string	-	-
3	quantity	bigint	-	-
4	unit_price	double	-	-
5	total_sales	double	-	-
6	location	string	-	-
7	partition_0	string	Partition (0)	-

- Open **Athena** and select your Glue database.



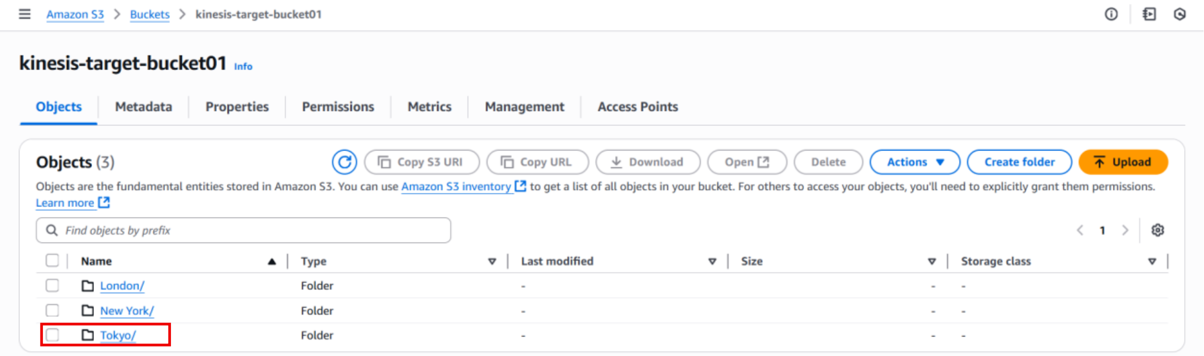
Completed　Time in queue: 124 ms　Run time: 826 ms　Data scanned: 1.06 KB

Results (20)

#	date	product_id	quantity	unit_price	total_sales	location	partition_0
1	2023-09-18 02:41:23	P940	5	73.05	365.23	New_York	New York
2	2023-01-26 09:19:44	P428	3	93.98	281.94	New_York	New York
3	2023-01-27 09:08:58	P981	16	70.85	1133.61	New_York	New York
4	2023-10-29 00:51:36	P169	13	10.96	142.52	New_York	New York

- Add a New Partition Folder

In S3, create a new folder (e.g., Tokyo/) and upload sales_tokyo.csv.



kinesis-target-bucket01 Info

Objects (3)

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
□ London/	Folder	-	-	-
□ New York/	Folder	-	-	-
□ Tokyo/	Folder	-	-	-

- Run the same Athena query — **Tokyo data won't appear yet**.
- Re-run the crawler and data will appear.
- **Update Metadata for New Partition**
- Go to **Athena** and run:
  - `ALTER TABLE your_table_name`
  - `ADD PARTITION (partition_0='Tokyo')`
  - `LOCATION 's3://your-bucket-name/Tokyo/';`
- Re-run your query → **Tokyo data now appears**.