Customer Sales Data Transformation with AWS Glue

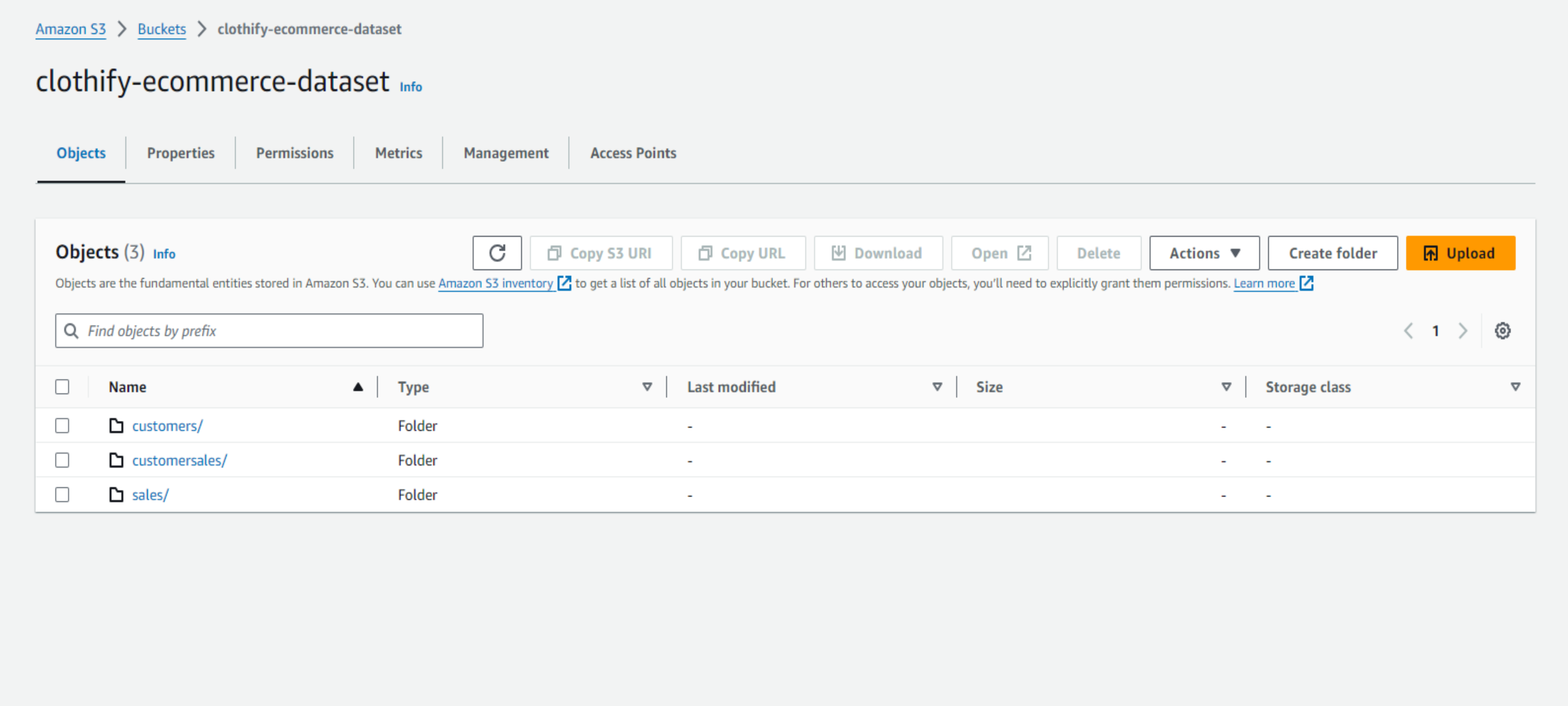
Sameer Singh
https://www.linkedin.com/in/sameer-singh-data/

Overview

Built an ETL pipeline using AWS Glue, seamlessly handling data from ingestion through transformation and storage in Amazon S3.

- ✓ Leveraged AWS services for integrated customer and sales data analysis.
- ✓ Utilized Amazon S3 for efficient data storage.
- ✓ Employed AWS Glue for metadata management, ETL processes, and data transformation.
- ✓ Combined datasets, apply filtering, and enhance data clarity with AWS Glue.
- ✓ Stored transformed data in S3, organized using Parquet files.
- ✓ Utilized Amazon Athena for efficient querying, facilitating comprehensive analysis.

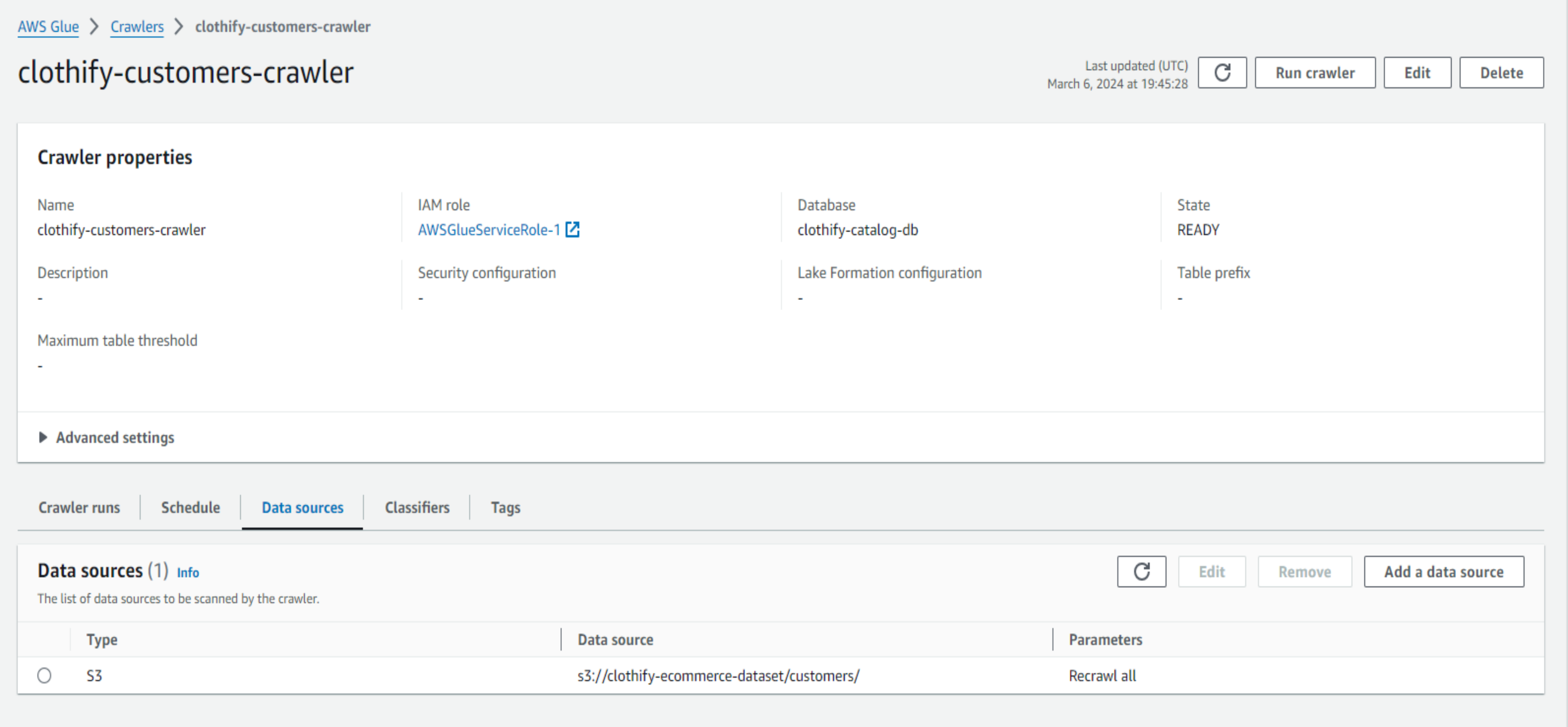
Created three S3 folders in buckets: two dedicated to customers and sales data, and one for storing Glue job results.



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
□	customers/	Folder	-	-	-
□	customersales/	Folder	-	-	-
□	sales/	Folder	-	-	-

Utilized Glue Crawlers to catalog metadata for Customer Dataset and creating Data Catalog tables.



clothify-customers-crawler

Crawler properties

Name clothify-customers-crawler	IAM role AWSGlueServiceRole-1	Database clothify-catalog-db	State READY
Description -	Security configuration -	Lake Formation configuration -	Table prefix -
Maximum table threshold -			

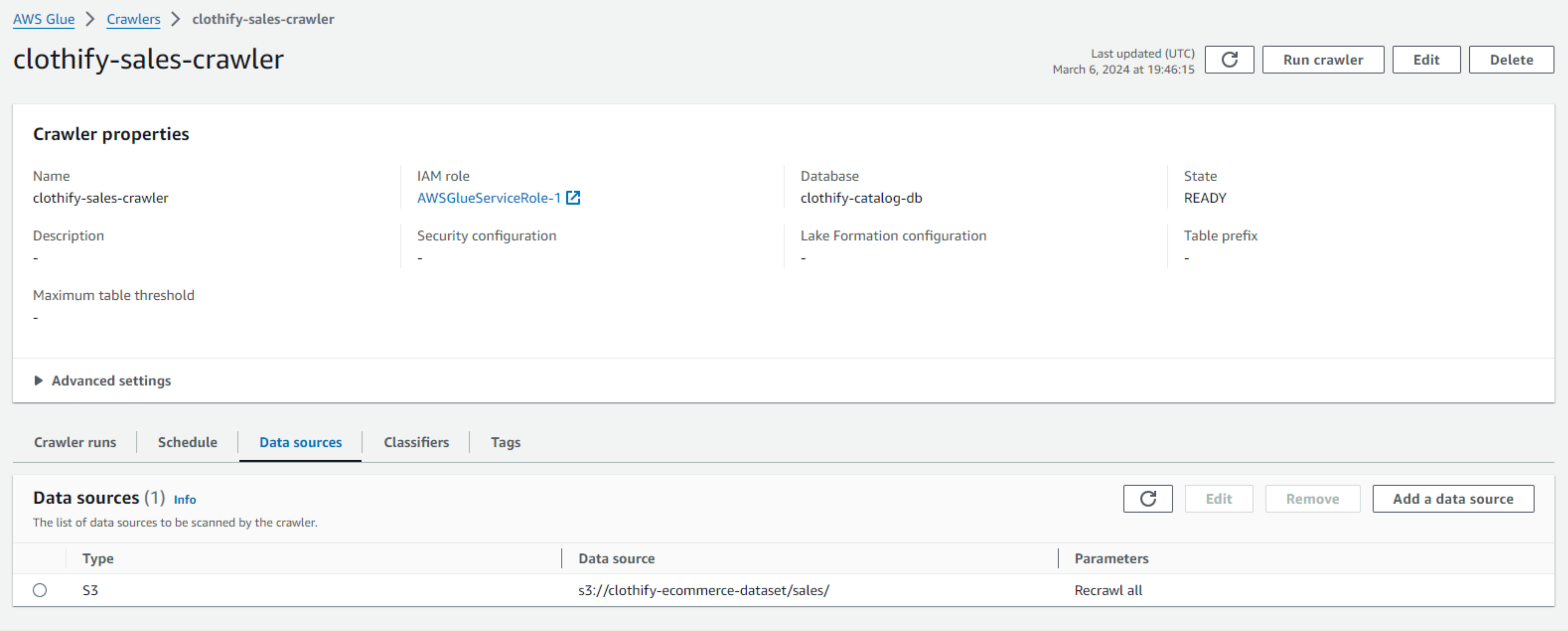
▶ Advanced settings

Data sources (1) Info

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

	Type	Data source	Parameters
○	S3	s3://clothify-ecommerce-dataset/customers/	Recrawl all

Same Process for Sales Dataset and creating Data Catalog tables.



clothify-sales-crawler

Last updated (UTC)
March 6, 2024 at 19:46:15

Crawler properties

Name: clothify-sales-crawler

IAM role: AWSGlueServiceRole-1

Database: clothify-catalog-db

State: READY

Description: -

Security configuration: -

Lake Formation configuration: -

Table prefix: -

Maximum table threshold: -

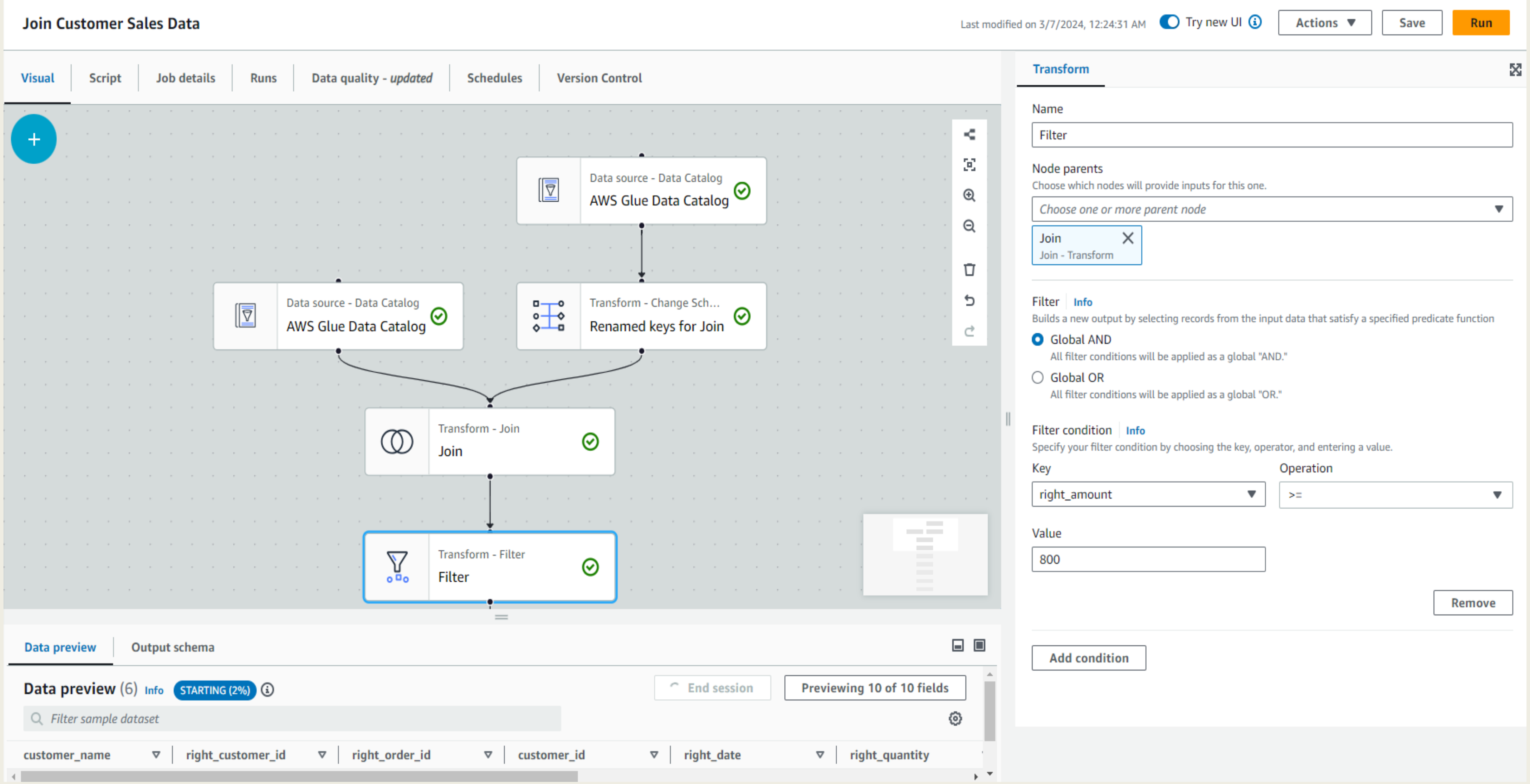
▶ Advanced settings

Data sources (1)

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

	Type	Data source	Parameters
○	S3	s3://clothify-ecommerce-dataset/sales/	Recrawl all

Developed a Glue ETL job using two Data Catalog tables for customers and sales datasets.



Last modified on 3/7/2024, 12:24:31 AM

Transform

Name: Filter

Node parents
Choose which nodes will provide inputs for this one.
Choose one or more parent node ▼
Join ✕
Join - Transform

Filter Info
Builds a new output by selecting records from the input data that satisfy a specified predicate function

● Global AND
All filter conditions will be applied as a global "AND."

○ Global OR
All filter conditions will be applied as a global "OR."

Filter condition Info
Specify your filter condition by choosing the key, operator, and entering a value.

Key: right_amount ▼
Operation: >= ▼
Value: 800

Data preview (6) Info

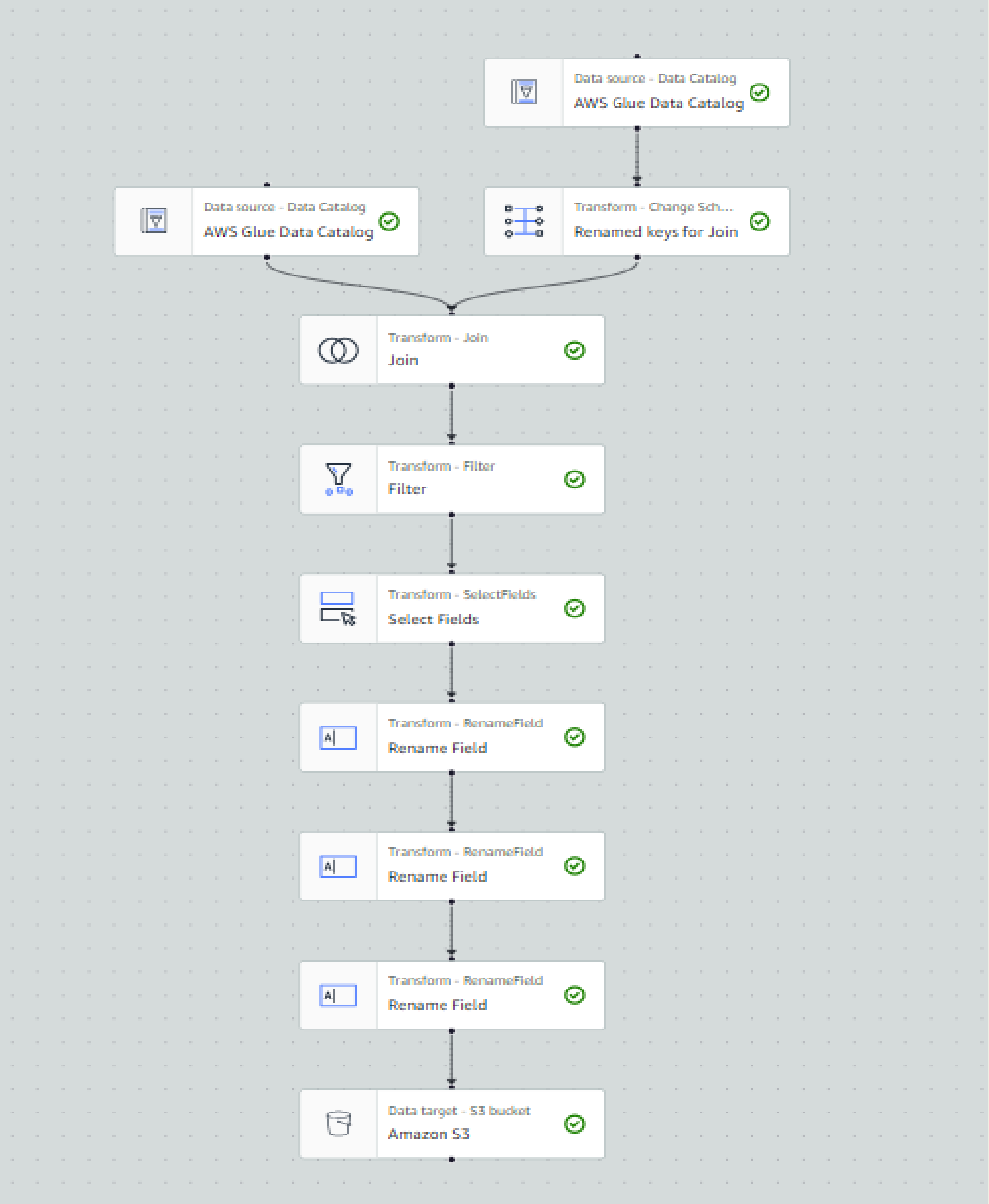
customer_name	right_customer_id	right_order_id	customer_id	right_date	right_quantity

Utilized AWS Glue ETL Job with metadata from Data Catalog tables.

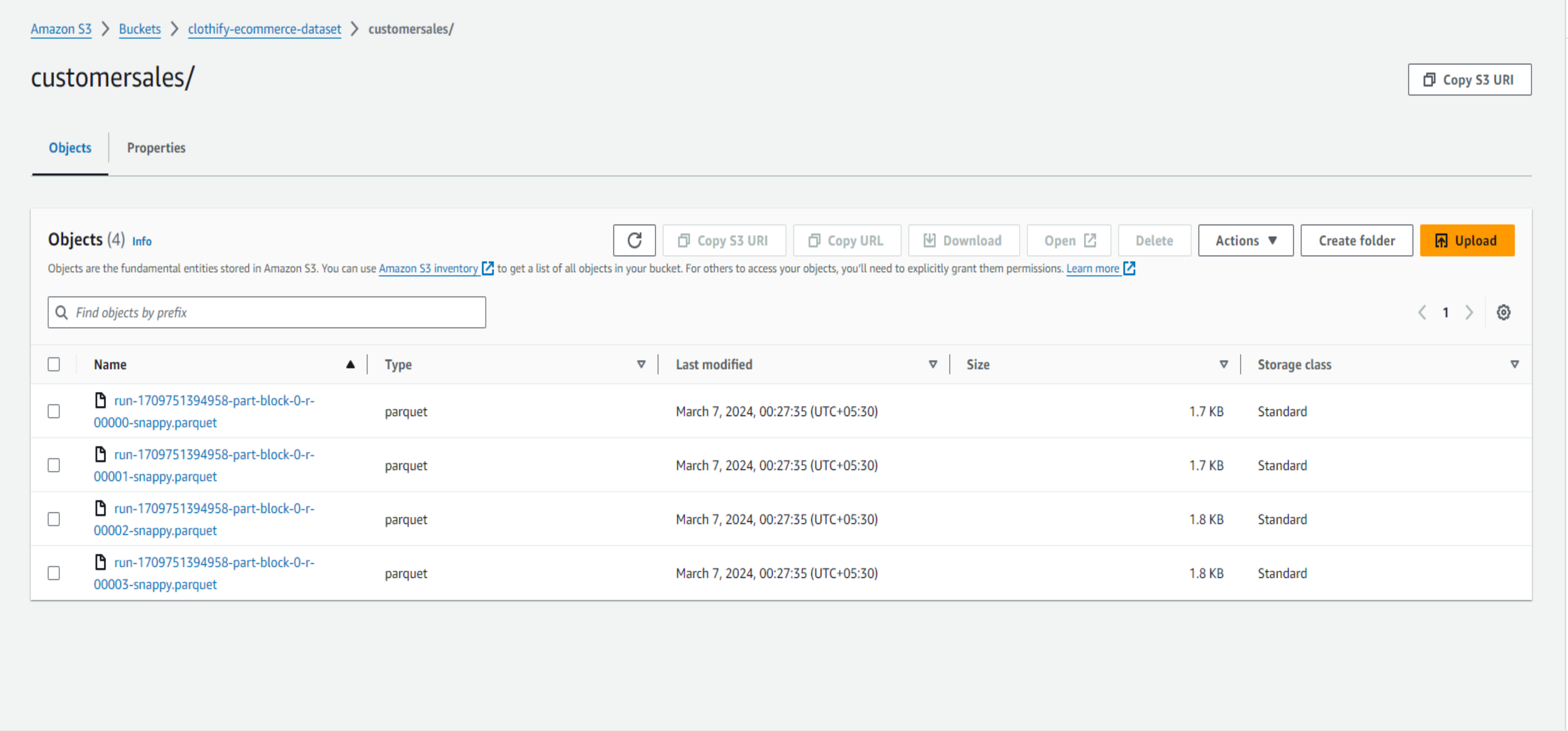
Executed a join operation on 'customer id' and applied a filter for orders exceeding $800.

Implemented data transformation by renaming key sales fields to 'order date', 'order ID', and 'order amount.'

Organized and stored the transformed dataset in the 'customer-sales' folder within the specified S3 bucket.



Result from Glue Job Stored in customersales folder in s3 bucket



Amazon S3 > Buckets > clothify-ecommerce-dataset > customersales/

customersales/

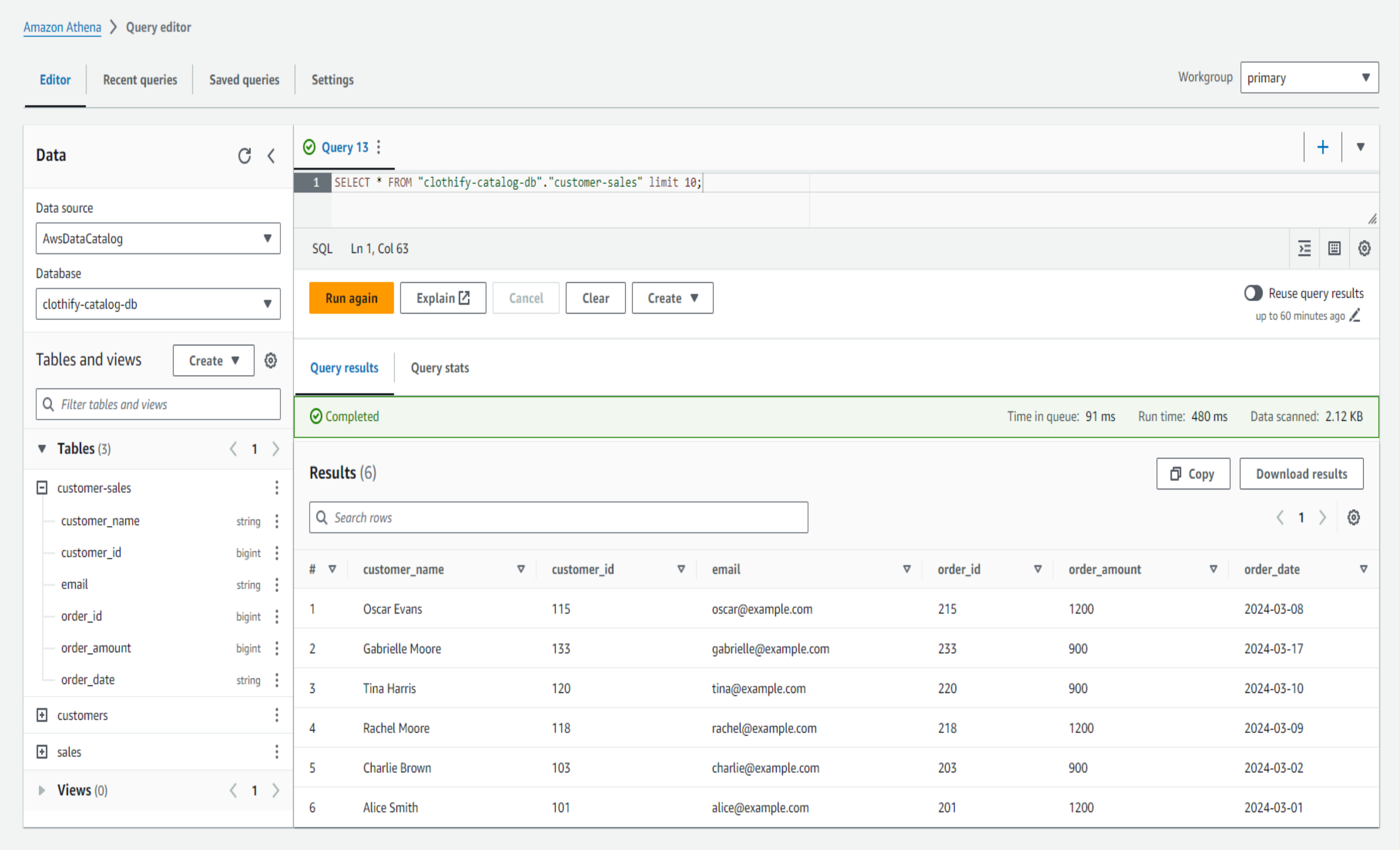
Objects (4) Info

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

Find objects by prefix

□	Name ▲	Type	Last modified	Size	Storage class
□	run-1709751394958-part-block-0-r-00000-snappy.parquet	parquet	March 7, 2024, 00:27:35 (UTC+05:30)	1.7 KB	Standard
□	run-1709751394958-part-block-0-r-00001-snappy.parquet	parquet	March 7, 2024, 00:27:35 (UTC+05:30)	1.7 KB	Standard
□	run-1709751394958-part-block-0-r-00002-snappy.parquet	parquet	March 7, 2024, 00:27:35 (UTC+05:30)	1.8 KB	Standard
□	run-1709751394958-part-block-0-r-00003-snappy.parquet	parquet	March 7, 2024, 00:27:35 (UTC+05:30)	1.8 KB	Standard

Leveraged Athena to query the combined dataset stored in Parquet format.



Data source: AwsDataCatalog

Database: clothify-catalog-db

Tables (3)

- customer-sales

customer_name	string
customer_id	bigint
email	string
order_id	bigint
order_amount	bigint
order_date	string

- customers
- sales

Views (0)

```
SELECT * FROM "clothify-catalog-db"."customer-sales" limit 10;
```

SQL Ln 1, Col 63

Completed   Time in queue: 91 ms   Run time: 480 ms   Data scanned: 2.12 KB

Results (6)

#	customer_name	customer_id	email	order_id	order_amount	order_date
1	Oscar Evans	115	oscar@example.com	215	1200	2024-03-08
2	Gabrielle Moore	133	gabrielle@example.com	233	900	2024-03-17
3	Tina Harris	120	tina@example.com	220	900	2024-03-10
4	Rachel Moore	118	rachel@example.com	218	1200	2024-03-09
5	Charlie Brown	103	charlie@example.com	203	900	2024-03-02
6	Alice Smith	101	alice@example.com	201	1200	2024-03-01