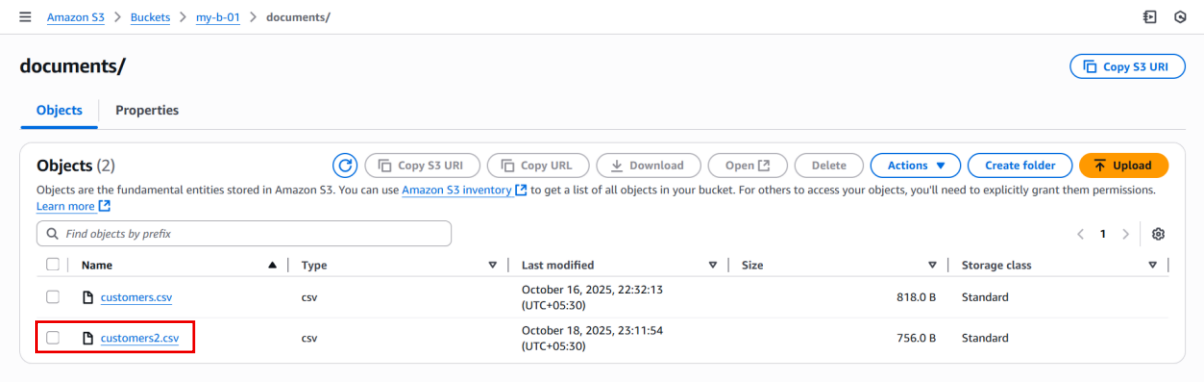
Stateless Data Ingestion in Glue

To Begin with the Lab

Summary of the Lab

In this lab, we demonstrated stateless data ingestion in AWS Glue. After uploading a new CSV file to the S3 source folder, the existing ETL job was rerun. The job reprocessed all source files instead of only the new one, causing duplicate records in the target data. This behaviour illustrates that without state tracking; Glue reloads all data every time the job runs. Using Athena, the duplication was confirmed by querying and ordering records by ID. Stateless ingestion processes the entire dataset on each run, making it inefficient for incremental or large-scale data updates.

- Prerequisites
    - You must already have:
    - An S3 bucket with a documents folder containing a CSV file.
    - A target folder in the same bucket (e.g., target-customer).
    - An existing AWS Glue ETL job configured to move data from documents to target-customer.
    - A table in the Glue Data Catalog created by your ETL job.
- Go to the S3 console.
- Navigate to your documents folder (source location).
- Click Upload → Add files → select an additional CSV file (e.g., customers2.csv).
- Click Upload to complete.



- Under **Tables**, select the table created by your ETL job
- Note the **number of records** currently in the table

Results (11)

Copy | Download results CSV

Search rows

1

#	id	name	age	address	city	state	email
1	1	John Smith	25	123 Main St.	New York	NY	johnsmith@example.com
2	2	Jane Doe	30	456 Oak Street	Los Angeles	CA	janedoe@example.com
3	3	Mark Johnson	40	789 Pine Court	San Francisco	CA	markjohnson@example.com
4	4	Alice Ali	30	123 Elm Street	Chicago	IL	aliceali@example.com
5	5	Robert Yang	22	789 Cedar Ln	Miami	FL	robertyang@example.com
6	6	Sarah Smith	28	123 Aspen Street	New York	NY	sarahsmith@example.com
7	7	David Ramirez	35	456 Walnut Lane	Houston	TX	davidramirez@example.com
8	8	Anna Alanson	27	777 Chestnut Avenue	Los Angeles	CA	annaalanson@example.com
9	9	Michael Jones	31	456 Willow Street	San Francisco	CA	michaeljones@example.com
10	10	Linda Lovegood	29	789 Sycamore Ave	Chicago	IL	lindalovegood@example.com
11	11	Sarah Smith	28	123 Aspen Street	New York	NY	sarahsmith@example.com

- Select your existing ETL job.
- Click Run to execute it again.
- Wait for the job to complete successfully.
- Go back to Athena and re-run your previous query.
- Observe that records are duplicated (e.g., now 32 rows instead of 21) as first file had 11 record and second file has 10 records but as we rerun the job again so it took the previous 10 records again.
- This shows that the ETL job reloaded all data (stateless ingestion).
- Use an ORDER BY clause to confirm duplicate IDs

Completed | **Time in queue:** 78 ms | **Run time:** 475 ms | **Data scanned:** 3.42 KB

Results (32)

Copy | Download results CSV

Search rows

1

#	id	name	age	address	city	state	email
1	1	John Smith	25	123 Main St.	New York	NY	johnsmith@example.com
2	2	Jane Doe	30	456 Oak Street	Los Angeles	CA	janedoe@example.com
3	3	Mark Johnson	40	789 Pine Court	San Francisco	CA	markjohnson@example.com
4	4	Alice Ali	30	123 Elm Street	Chicago	IL	aliceali@example.com
5	5	Robert Yang	22	789 Cedar Ln	Miami	FL	robertyang@example.com