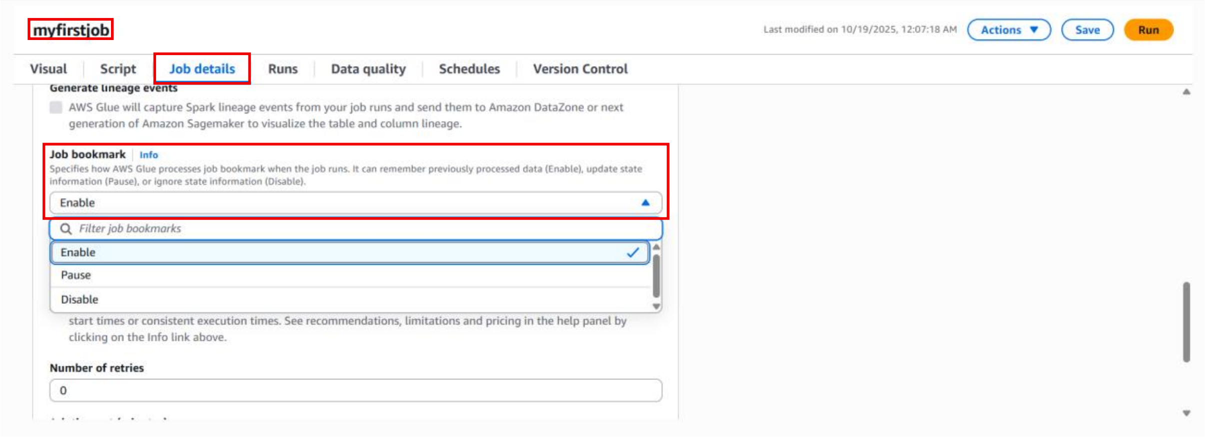
Stateful Ingestion with Bookmarks

To Begin with the Lab

Summary of the lab

In this lab, we implemented **stateful data ingestion** in AWS Glue using **job bookmarks**. After enabling bookmarks in the ETL job settings, Glue began tracking previously processed data. The job was first run with one CSV file, creating a Parquet file in the target folder. After uploading a new CSV file, rerunning the job processed only the new data. Verifying in Athena showed 21 records (11 original + 10 new) without duplicates. This demonstrated how **Glue bookmarks** prevent reprocessing old data, ensuring efficient **incremental data loading** and maintaining data consistency between source and target folders.

- Prerequisites
    - Have an S3 bucket with two folders:
    - documents → contains CSV files (source data).
    - target → will store transformed data (output).
    - An ETL job in AWS Glue must already exist to move data from documents → target.
    - The AWS Glue IAM role should have AmazonS3FullAccess permissions.
- Go to AWS Console → AWS Glue → ETL Jobs.
- Select your existing ETL job.
- Click Job Details.
- Scroll down to the Job bookmarks section.
- Choose Enable under Specifies how AWS Glue processes job bookmarks.
- Click Save to update the job configuration.



Generate lineage events

□ AWS Glue will capture Spark lineage events from your job runs and send them to Amazon DataZone or next generation of Amazon Sagemaker to visualize the table and column lineage.

Job bookmark Info

Specifies how AWS Glue processes job bookmark when the job runs. It can remember previously processed data (Enable), update state information (Pause), or ignore state information (Disable).

Enable ▲

Filter job bookmarks

Enable ✓

Pause

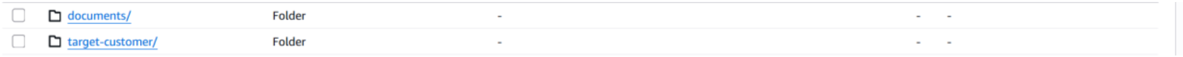
Disable

start times or consistent execution times. See recommendations, limitations and pricing in the help panel by clicking on the Info link above.

Number of retries

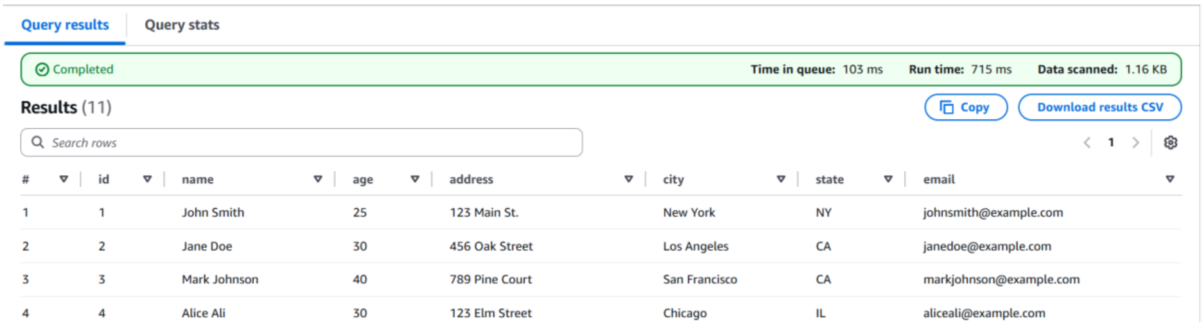
0

- Open the S3 Console.
- Go to your target-customer folder and delete all existing files (to start fresh).
- In the documents folder, keep only the first CSV file(customers.csv) and delete any others.



□	documents/	Folder	-	-	-
□	target-customer/	Folder	-	-	-

- Go back to AWS Glue → Jobs.
- Select your ETL job and click Run.
- Wait for the job status to show Succeeded.
- Check your target folder — it should now have a Parquet file containing data from the first CSV.
- Verify in Athena or the Glue Data Catalog that records have been loaded (e.g., 11 records).

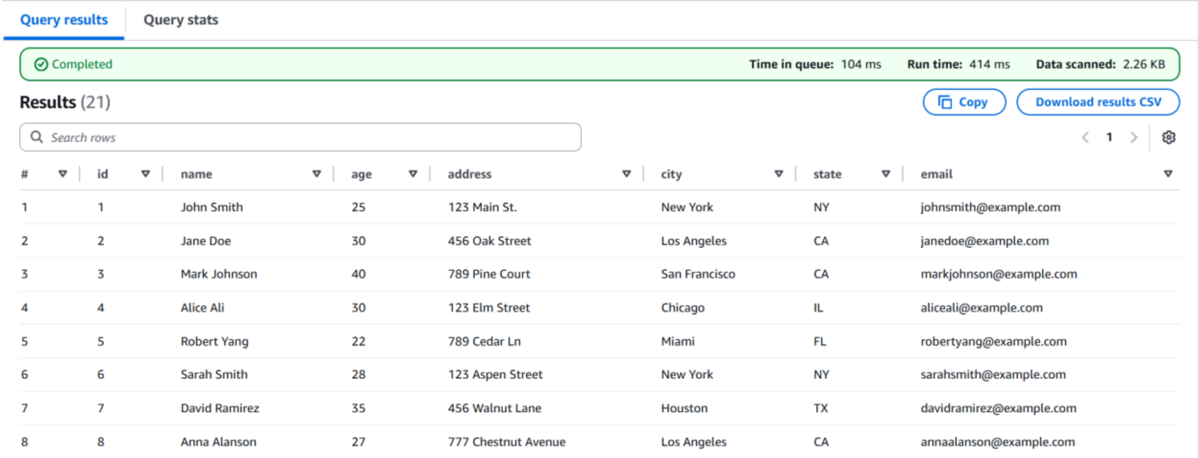


Completed   Time in queue: 103 ms   Run time: 715 ms   Data scanned: 1.16 KB

Results (11)

#	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

- Go to the documents folder in S3.
- Upload a new CSV file (e.g., customers2.csv).
- This represents newly arriving data.
- Go back to AWS Glue → Jobs and rerun the same job.
- Wait for it to finish successfully.
- Check your target folder — it should now contain the updated Parquet file.
- Open Athena and query the table linked to your target folder.
- You should now see 21 total records (11 old + 10 new), without duplicates.
- This confirms that bookmarks worked correctly — only new files were ingested.



Completed   Time in queue: 104 ms   Run time: 414 ms   Data scanned: 2.26 KB

Results (21)

#	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