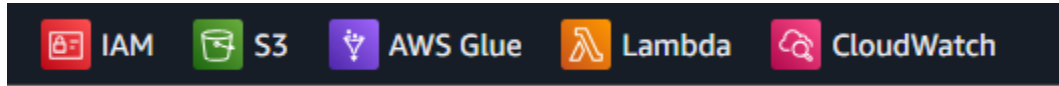


Data Pipeline on AWS

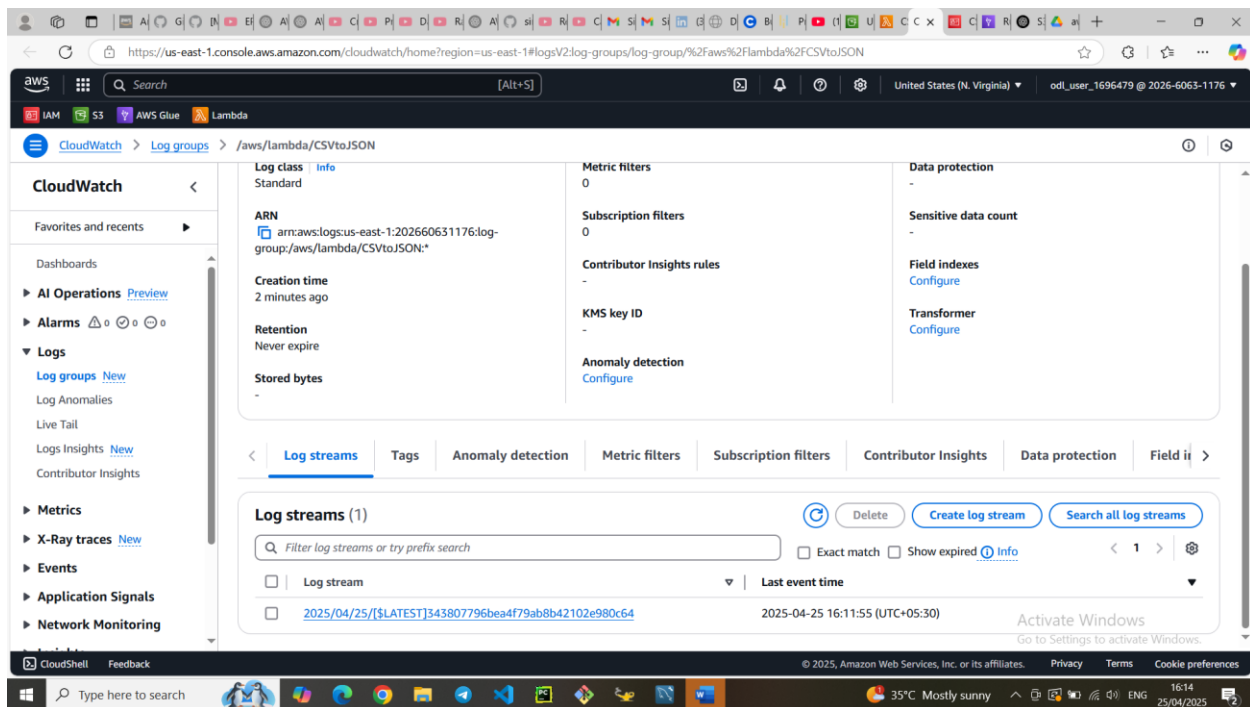
AWS services used:



- AWS Glue for creating pipeline
- Trigger AWS Glue using AWS Lambda Function
- Store data using amazon s3
- Use AWS IAM for policy
- Use AWS Cloudwatch for monitoring

Lambda Function: Python

```
import json
import boto3
glueClient = boto3.client('glue')
def lambda_handler(event, context):
    glueClient.start_job_run(JobName="csvjson")
    return "Job started"
```



CloudWatch Log events interface showing log data for the log group `/aws/lambda/CSVtoJSON`. The interface includes a search bar, filters, and a table of log events.

Log events

You can use the filter bar below to search for and match terms, phrases, or values in your log events. [Learn more about filter patterns](#)

Filter events - press enter to search

Clear 1m 30m 1h 12h Custom Local timezone

Display

Timestamp	Message
No older events at this moment. Retry	
2025-04-25T16:11:54.164+05:30	INIT_START Runtime Version: python:3.13.v38 Runtime Version ARN: arn:aws:lambda:us-east-1::runtime:9ed40a2fed561b26d389cb8d6a2a08d...
2025-04-25T16:11:54.728+05:30	START RequestId: a6670f08-7ecf-4065-8208-5f09959baa38 Version: \$LATEST
2025-04-25T16:11:55.197+05:30	END RequestId: a6670f08-7ecf-4065-8208-5f09959baa38
2025-04-25T16:11:55.197+05:30	REPORT RequestId: a6670f08-7ecf-4065-8208-5f09959baa38 Duration: 468.38 ms Billed Duration: 469 ms Memory Size: 128 MB Max Memory Us...
No newer events at this moment. Auto retry paused. Resume	

Glue Studio interface showing the `csvjson` job details. The interface includes tabs for Visual, Script, Job details, Runs, Data quality, Schedules, and Version Control.

csvjson Last modified on 25/04/2025, 15:56:43 [Actions](#) [Save](#) [Run](#)

Visual Script Job details **Runs** Data quality Schedules Version Control

Job runs (1/1) info Last updated (UTC) April 25, 2025 at 10:45:43 [View details](#) [Stop job run](#) [Troubleshoot with AI](#) [Table View](#) [Card View](#)

Filter job runs by property

Run status	Retries	Start time (Local)	End time (Local)	Duration	Capacity (DPUs)	Worker type	Glue version
Succeeded	0	04/25/2025 16:11:55	04/25/2025 16:13:10	1 m 7 s	2 DPUs	G.1X	5.0

Run details Input arguments (9) Logs Run insights Metrics Troubleshooting analysis - preview Spark UI [Rewind job bookmark](#)

Job name	Start time (Local)	Glue version	Last modified on (Local)
csvjson	04/25/2025 16:11:55	5.0	04/25/2025 16:13:10
ID	End time (Local)	Worker type	Log group name
j_r_159259bab9c1910d4527d27b270c9444e5d19725b19251e5a21c74be10323496	04/25/2025 16:13:10	G.1X	/aws-glue/jobs
Run status	Start-up time	Max capacity	Number of workers
Succeeded	8 seconds	2 DPUs	2
Retry attempt number	Execution time	Execution class	Timeout
Initial run	1 minute 7 seconds	Standard	480 minutes

https://us-east-1.console.aws.amazon.com/s3/buckets/json-bucket20?region=us-east-1&bucketType=general&tab=objects

aws lambda IAM S3 AWS Glue Lambda United States (N. Virginia) odl_user_1696479 @ 2026-6063-1176

Amazon S3 > Buckets > json-bucket20

json-bucket20 info

Objects Metadata Properties Permissions Metrics Management Access Points

Objects (1) [Copy S3 URI](#) [Copy URL](#) [Download](#) [Open](#) [Delete](#) [Actions](#) [Create folder](#) [Upload](#)

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

<input type="checkbox"/>	Name	Type	Last modified	Size	Storage class
<input type="checkbox"/>	run-AmazonS3_node1745576614474-1-part-r-00000	-	April 25, 2025, 16:12:59 (UTC+05:30)	985.0 B	Standard

Activate Windows
Go to Settings to activate Windows.

CloudShell Feedback © 2025, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences

https://us-east-1.console.aws.amazon.com/s3/buckets/pipelinedemo20script?region=us-east-1&bucketType=general&tab=objects

aws lambda IAM S3 AWS Glue Lambda United States (N. Virginia) odl_user_1696479 @ 2026-6063-1176

Amazon S3 > Buckets > pipelinedemo20script

pipelinedemo20script info

Objects Metadata Properties Permissions Metrics Management Access Points

Objects (1) [Copy S3 URI](#) [Copy URL](#) [Download](#) [Open](#) [Delete](#) [Actions](#) [Create folder](#) [Upload](#)

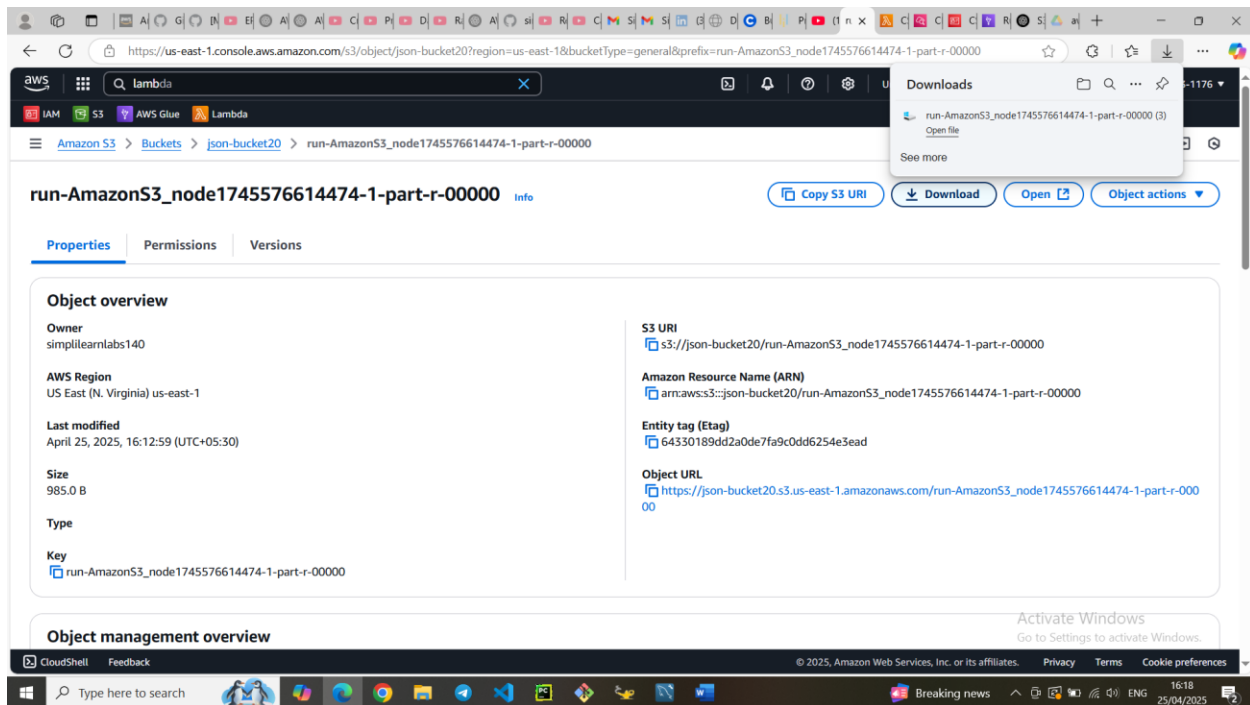
Objects are the fundamental entities stored in Amazon S3. You can use [Amazon S3 inventory](#) to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. [Learn more](#)

Find objects by prefix

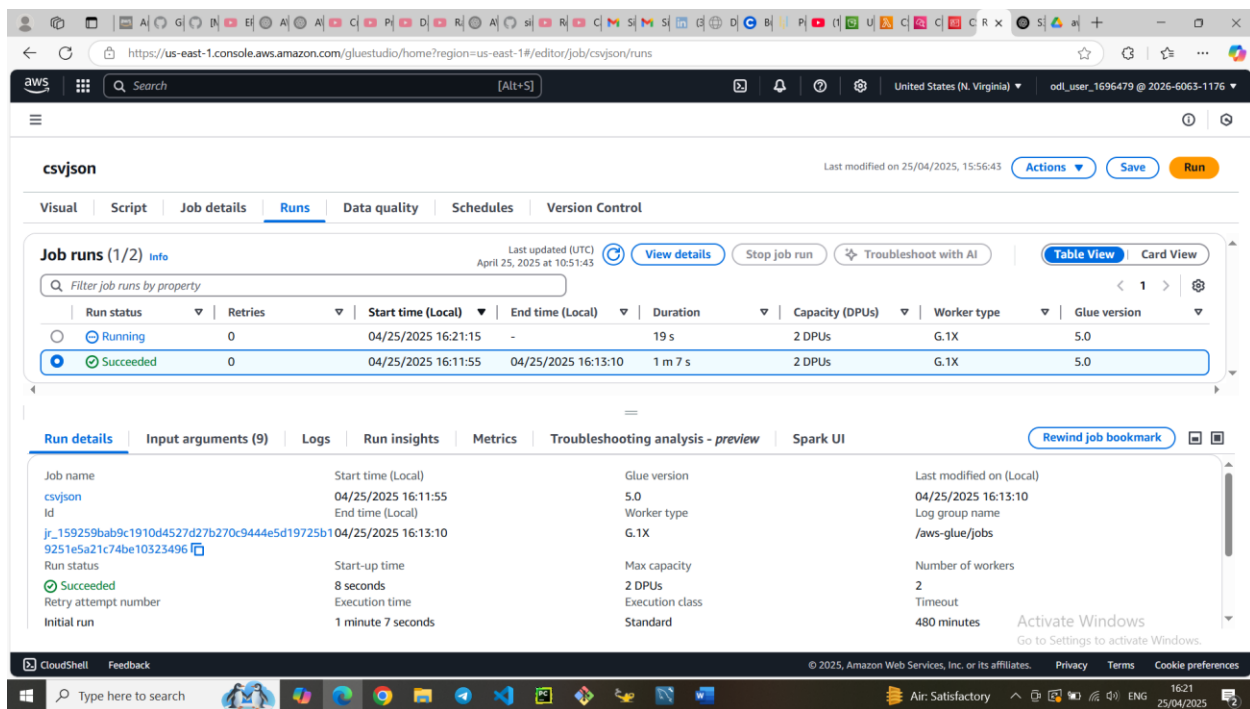
<input type="checkbox"/>	Name	Type	Last modified	Size	Storage class
<input type="checkbox"/>	csvjson.py	py	April 25, 2025, 15:56:44 (UTC+05:30)	1.6 KB	Standard

Activate Windows
Go to Settings to activate Windows.

CloudShell Feedback © 2025, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences



Adding 2nd CSV file in CSVinput S3 bucket



https://us-east-1.console.aws.amazon.com/gluestudio/home?region=us-east-1#/editor/job/csvjson/runs

aws [Search] [Alt+S] United States (N. Virginia) odt_user_1696479 @ 2026-6063-1176

csvjson

Last modified on 25/04/2025, 15:56:43 [Actions](#) [Save](#) [Run](#)

Visual | Script | Job details | **Runs** | Data quality | Schedules | Version Control

Job runs (1/2) info

Last updated (UTC) April 25, 2025 at 10:52:34 [View details](#) [Stop job run](#) [Troubleshoot with AI](#) [Table View](#) [Card View](#)

Filter job runs by property

Run status	Retries	Start time (Local)	End time (Local)	Duration	Capacity (DPUs)	Worker type	Glue version
<input checked="" type="radio"/> Succeeded	0	04/25/2025 16:21:15	04/25/2025 16:22:34	1 m 10 s	2 DPUs	G.1X	5.0
<input type="radio"/> Succeeded	0	04/25/2025 16:11:55	04/25/2025 16:13:10	1 m 7 s	2 DPUs	G.1X	5.0

Run details | Input arguments (9) | Logs | Run insights | Metrics | Troubleshooting analysis - preview | Spark UI [Rewind job bookmark](#)

Job name	Start time (Local)	Glue version	Last modified on (Local)
csvjson	04/25/2025 16:21:15	5.0	04/25/2025 16:22:34
Id	End time (Local)	Worker type	Log group name
jr_b154bfd50f6b2454345af9a885b4fdb308a251ad84c	04/25/2025 16:22:34	G.1X	/aws-glue/jobs
bd104730efdae5d99f1ed			
Run status	Start-up time	Max capacity	Number of workers
<input checked="" type="radio"/> Succeeded	8 seconds	2 DPUs	2
Retry attempt number	Execution time	Execution class	Timeout
Initial run	1 minute 10 seconds	Standard	480 minutes

Activate Windows
Go to Settings to activate Windows.

CloudShell Feedback © 2025, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences

https://us-east-1.console.aws.amazon.com/s3/buckets/json-bucket20?region=us-east-1&bucketType=general&tab=objects

aws [Search] lambda United States (N. Virginia) odt_user_1696479 @ 2026-6063-1176

Amazon S3 > Buckets > json-bucket20

json-bucket20 info

Objects | Metadata | Properties | Permissions | Metrics | Management | Access Points

Objects (1/2)

[Copy S3 URI](#) [Copy URL](#) [Download](#) [Open](#) [Delete](#) [Actions](#) [Create folder](#) [Upload](#)

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
<input type="checkbox"/> run-AmazonS3_node1745576614474-1-part-r-00000	-	April 25, 2025, 16:12:59 (UTC+05:30)	985.0 B	Standard
<input checked="" type="checkbox"/> run-AmazonS3_node1745576614474-2-part-r-00000	-	April 25, 2025, 16:22:23 (UTC+05:30)	357.0 B	Standard

Activate Windows
Go to Settings to activate Windows.

CloudShell Feedback © 2025, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences