ETL Jobs

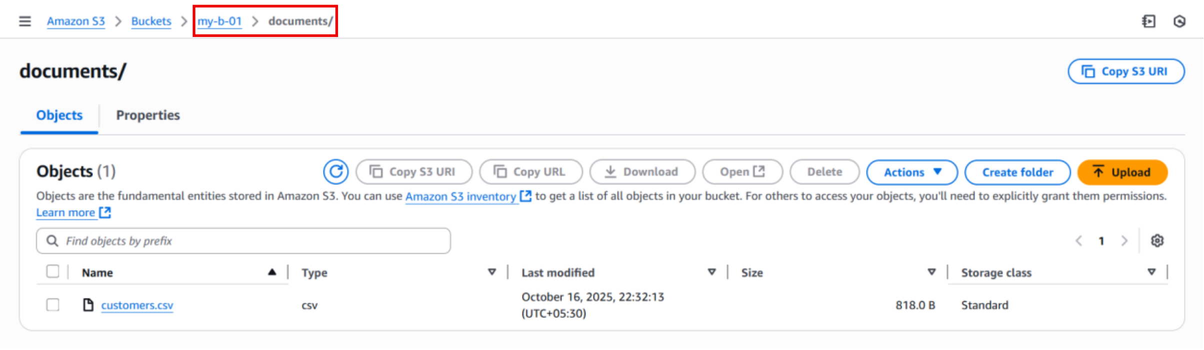
To Begin with the Lab

Summary of the Lab.

In this lab, an ETL job is created in AWS Glue to move data from one S3 folder to another. The source folder **"documents"** contains CSV files, and the target folder **"target-customer"** stores the transformed data. Using **Visual ETL**, an **S3 Source** and **S3 Target** are configured, with **Parquet** as the output format. The Data Catalog is updated to create and maintain table schemas. An **IAM role** with **S3FullAccess** is assigned to allow read/write access. After saving and running the job, permissions are verified if needed. Finally, the Parquet file appears in S3, and the table is added to the **Data Catalog**.

Prerequisite for the Lab:

- Folder created with the name "documents" inside the S3 Buckets with the files inside the folder.
- Create another folder in the same S3 bucket directory with the name target-customer.



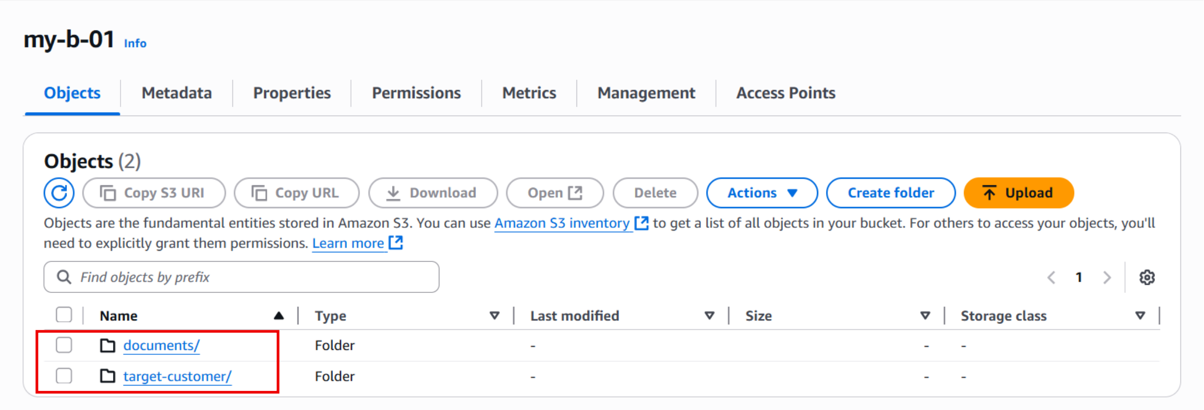
Amazon S3 > Buckets > my-b-01 > documents/

documents/

Objects (1)

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.csv	csv	October 16, 2025, 22:32:13 (UTC+05:30)	818.0 B	Standard



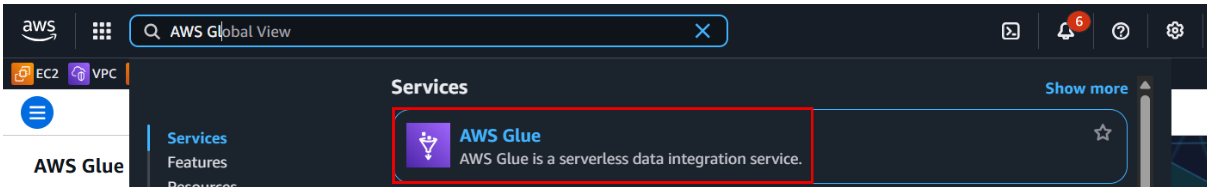
my-b-01 Info

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

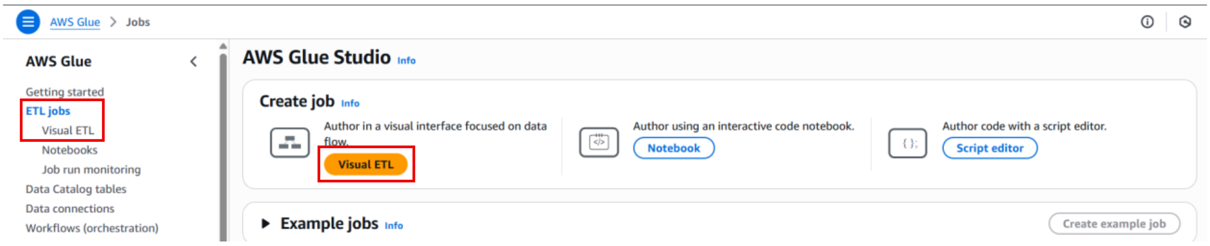
Name	Type	Last modified	Size	Storage class
documents/	Folder	-	-	-
target-customer/	Folder	-	-	-

- 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.

- In the **AWS Glue Console**, navigate to **ETL Jobs** and create a new job using **Visual ETL**.

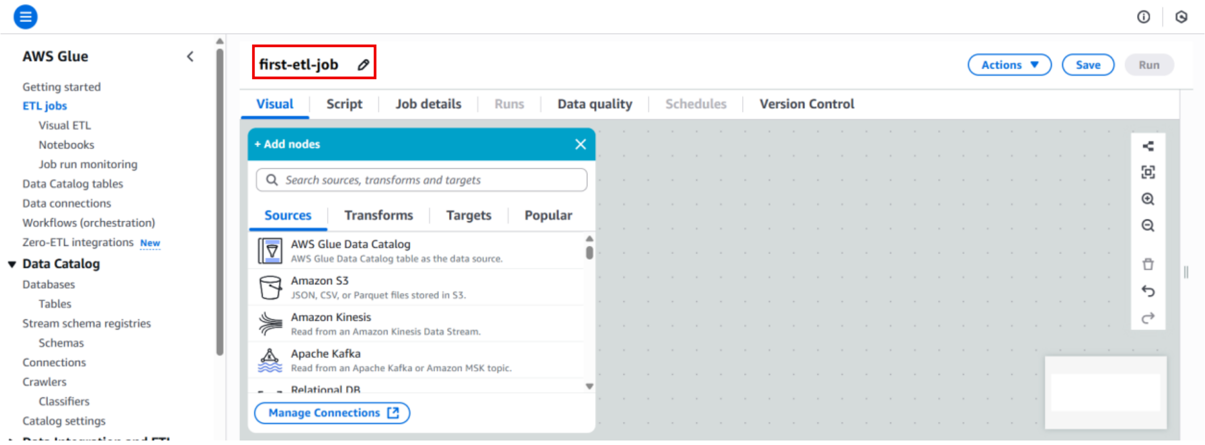


Author in a visual interface focused on data flow.

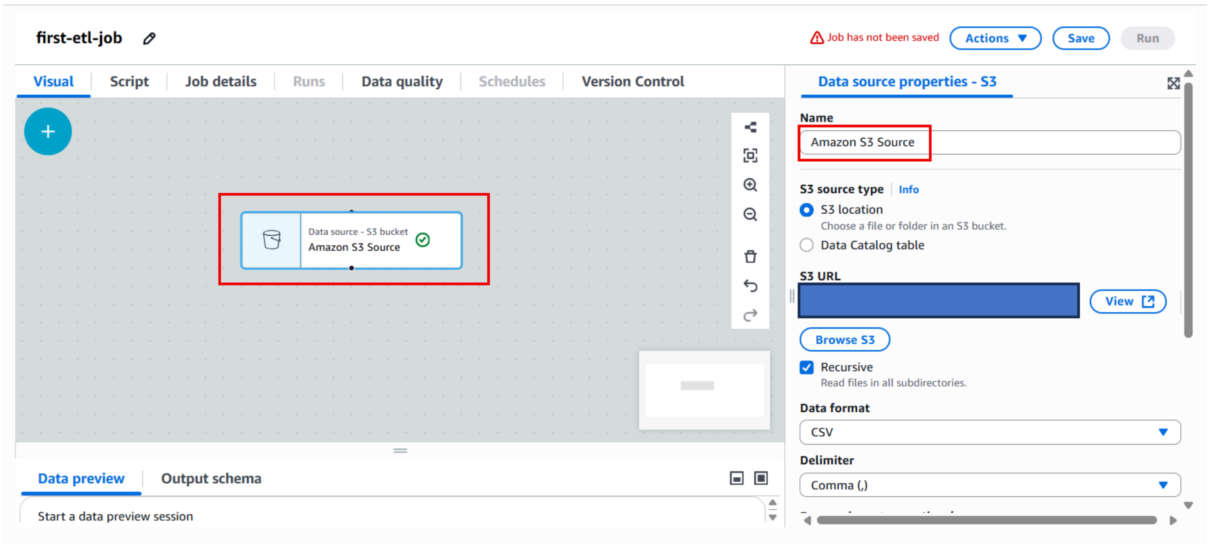
Author using an interactive code notebook.

Author code with a script editor.

- Name the job



- In the **Visual ETL editor**, add an **S3 Source** node and connect it to the folder containing the CSV file.
- Click on the Add node button and choose Amazon S3.
- We can rename the node and choose the path of the folder from the S3 bucket.
- Click Infer Schema



Name: Amazon S3 Source

S3 source type Info

● S3 location
Choose a file or folder in an S3 bucket.

○ Data Catalog table

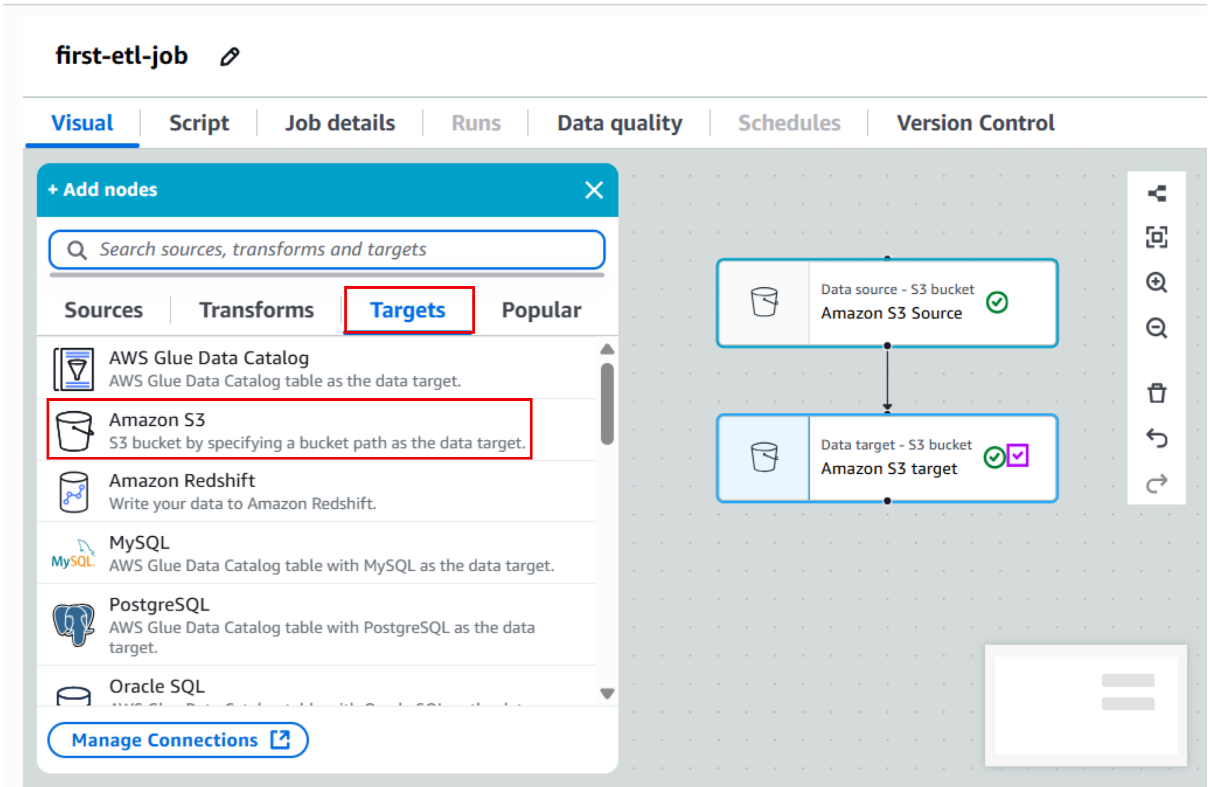
S3 URL: [redacted]

☑ Recursive
Read files in all subdirectories.

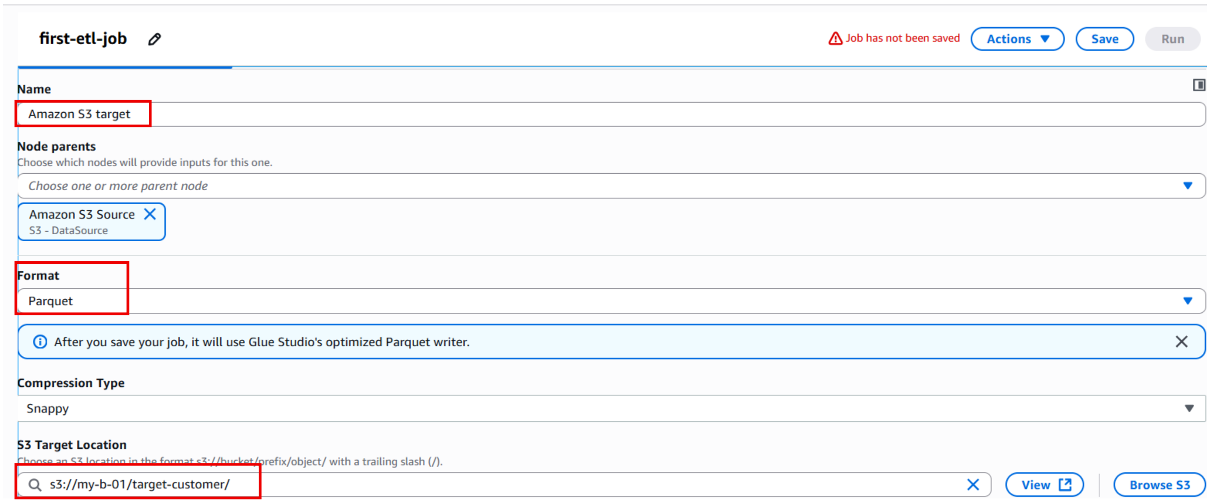
Data format: CSV

Delimiter: Comma (,)

- Click on the Targets and again select the Amazon S3 bucket.


first-etl-job
Visual
Script
Job details
Runs
Data quality
Schedules
Version Control
+ Add nodes
Search sources, transforms and targets
Sources
Transforms
Targets
Popular
AWS Glue Data Catalog
AWS Glue Data Catalog table as the data target.
Amazon S3
S3 bucket by specifying a bucket path as the data target.
Amazon Redshift
Write your data to Amazon Redshift.
MySQL
AWS Glue Data Catalog table with MySQL as the data target.
PostgreSQL
AWS Glue Data Catalog table with PostgreSQL as the data target.
Oracle SQL
Manage Connections
Data source - S3 bucket
Amazon S3 Source
Data target - S3 bucket
Amazon S3 target

- Rename it.
- Choose the format as Parquet.
- Choose the S3 Target Location
- In Data Catalog update options, choose the “Create a table in the Data Catalog and on subsequent runs, update the schema and add new partitions”
- Choose the Database and give the table name.



Name

Amazon S3 target

Node parents

Choose which nodes will provide inputs for this one.

Choose one or more parent node

Amazon S3 Source ✕
S3 - DataSource

Format

Parquet

ⓘ After you save your job, it will use Glue Studio's optimized Parquet writer.

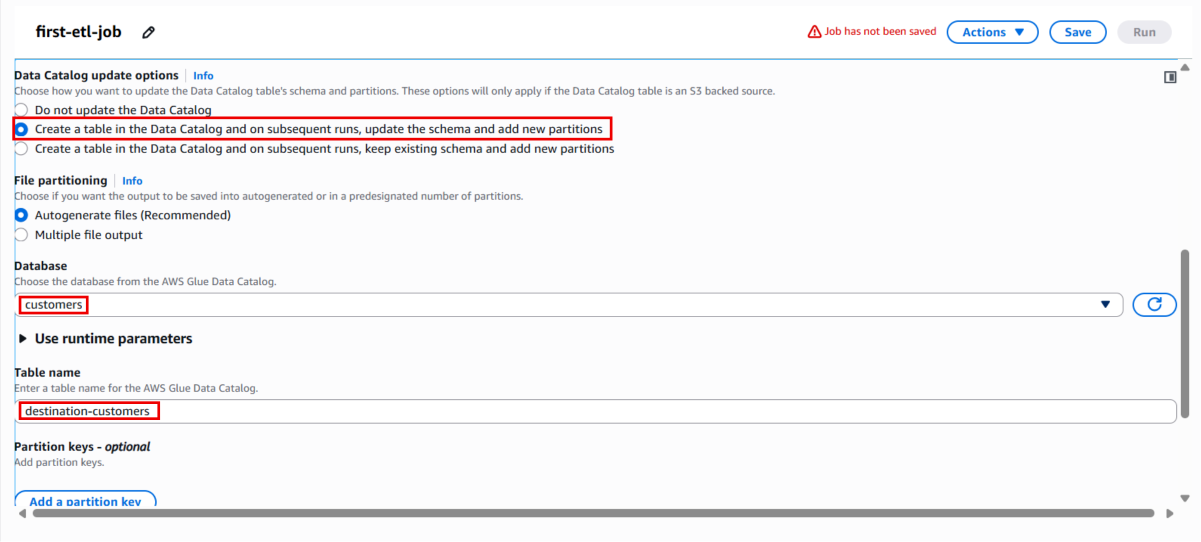
Compression Type

Snappy

S3 Target Location

Choose an S3 location in the format s3://bucket/prefix/object/ with a trailing slash (/).

s3://my-b-01/target-customer/



Data Catalog update options | Info

Choose how you want to update the Data Catalog table's schema and partitions. These options will only apply if the Data Catalog table is an S3 backed source.

○ Do not update the Data Catalog

● Create a table in the Data Catalog and on subsequent runs, update the schema and add new partitions

○ Create a table in the Data Catalog and on subsequent runs, keep existing schema and add new partitions

File partitioning | Info

Choose if you want the output to be saved into autogenerated or in a predesignated number of partitions.

● Autogenerate files (Recommended)

○ Multiple file output

Database

Choose the database from the AWS Glue Data Catalog.

customers

▶ Use runtime parameters

Table name

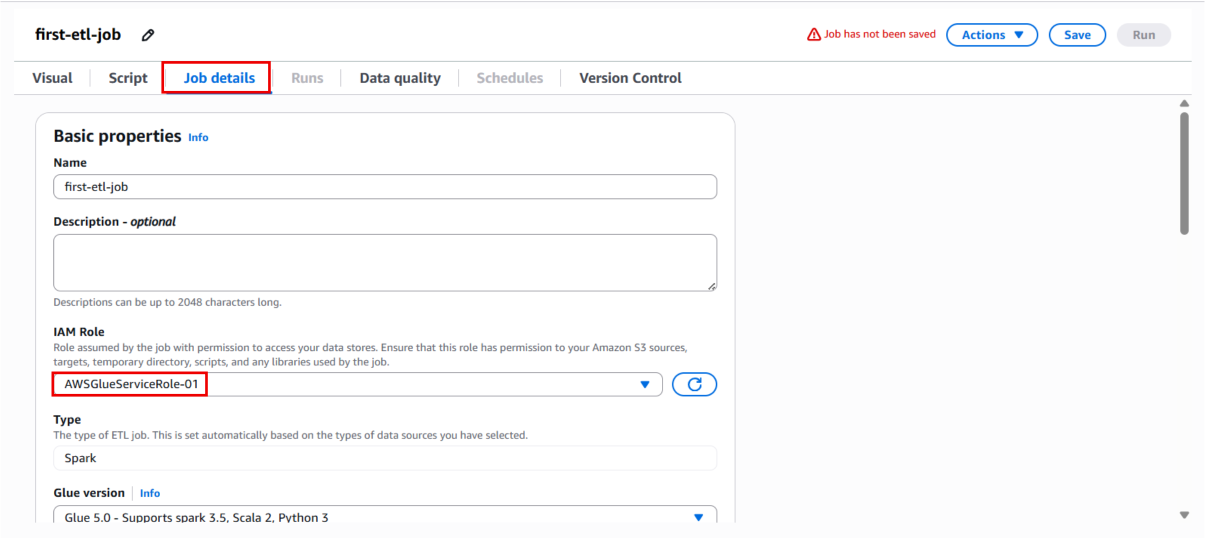
Enter a table name for the AWS Glue Data Catalog.

destination-customers

Partition keys - *optional*

Add partition keys.

- Now go to Job details, Assign an **IAM role** with permissions to read and write from S3 (e.g., attach AmazonS3FullAccess for this lab).
- Change the Requested number of workers.
- Then Click Save and Run.



Basic properties Info

Name

first-etl-job

Description - *optional*

Descriptions can be up to 2048 characters long.

IAM Role

Role assumed by the job with permission to access your data stores. Ensure that this role has permission to your Amazon S3 sources, targets, temporary directory, scripts, and any libraries used by the job.

AWSGlueServiceRole-01

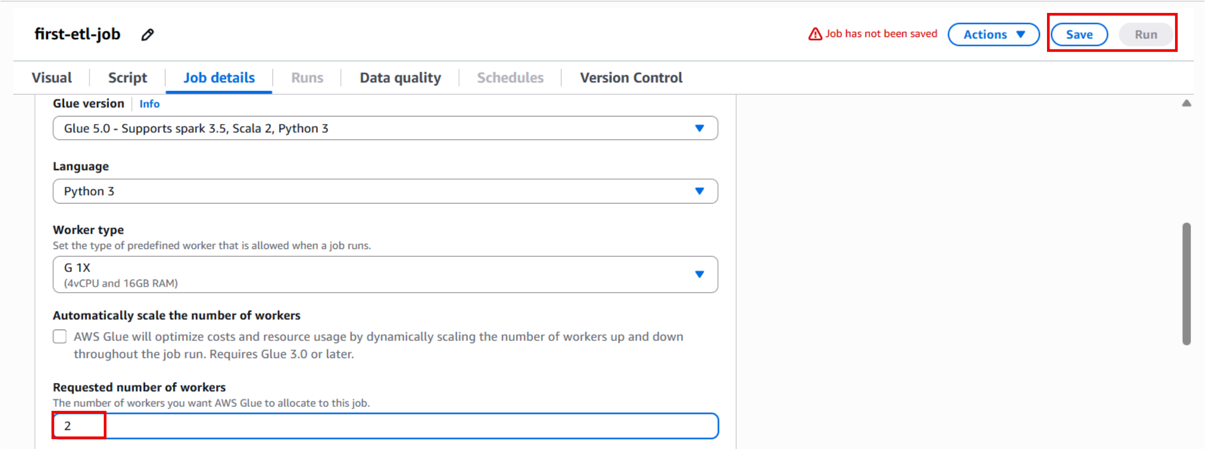
Type

The type of ETL job. This is set automatically based on the types of data sources you have selected.

Spark

Glue version | Info

Glue 5.0 - Supports spark 3.5, Scala 2, Python 3



Glue version Info
Glue 5.0 - Supports spark 3.5, Scala 2, Python 3

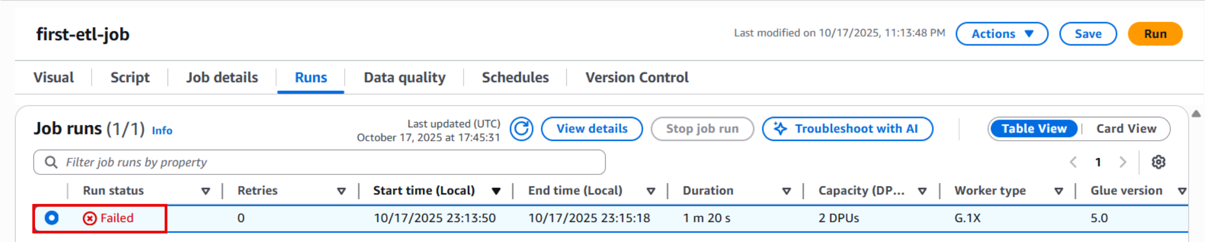
Language
Python 3

Worker type
Set the type of predefined worker that is allowed when a job runs.
G 1X
(4vCPU and 16GB RAM)

Automatically scale the number of workers
□ AWS Glue will optimize costs and resource usage by dynamically scaling the number of workers up and down throughout the job run. Requires Glue 3.0 or later.

Requested number of workers
The number of workers you want AWS Glue to allocate to this job.
2

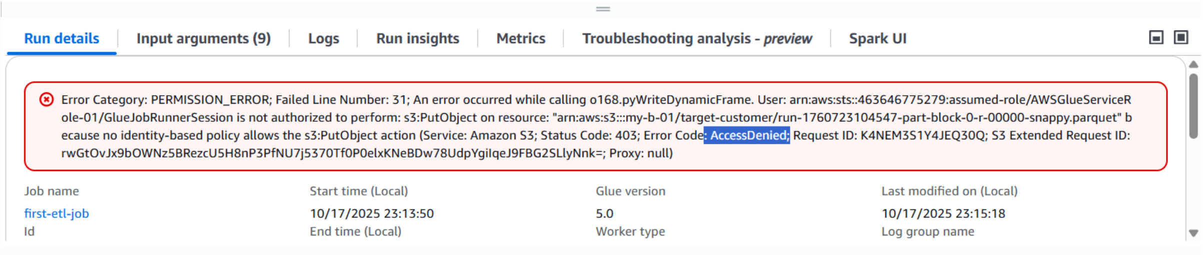
- Now we can go to Run details to check the status of the job.



Job runs (1/1) Info
Last updated (UTC)
October 17, 2025 at 17:45:31

	Run status	Retries	Start time (Local)	End time (Local)	Duration	Capacity (DP...	Worker type	Glue version
●	Failed	0	10/17/2025 23:13:50	10/17/2025 23:15:18	1 m 20 s	2 DPUs	G.1X	5.0

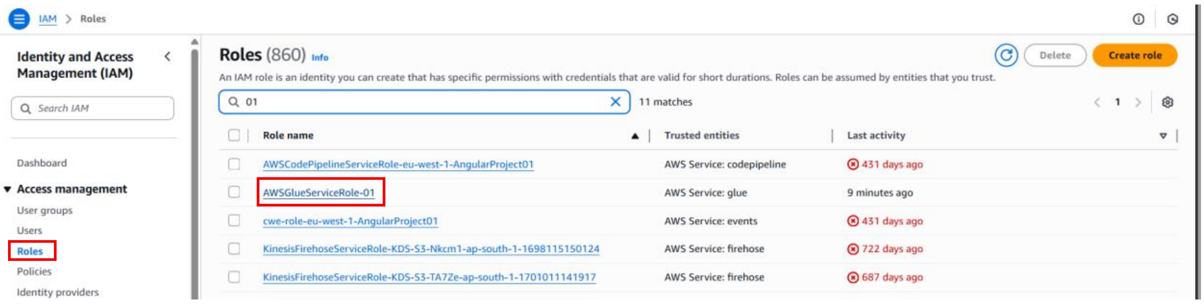
- If the job fails with "Access Denied," update the IAM role to include proper S3 permissions and rerun.



```
Error Category: PERMISSION_ERROR; Failed Line Number: 31; An error occurred while calling o168.pyWriteDynamicFrame. User: arn:aws:sts::463646775279:assumed-role/AWSGlueServiceRole-01/GlueJobRunnerSession is not authorized to perform: s3:PutObject on resource: "arn:aws:s3:::my-b-01/target-customer/run-1760723104547-part-block-0-r-00000-snappy.parquet" because no identity-based policy allows the s3:PutObject action (Service: Amazon S3; Status Code: 403; Error Code: AccessDenied; Request ID: K4NEM3S1Y4JEQ30Q; S3 Extended Request ID: rwGtOvJx9bOWNz5BRezcU5H8nP3PfNU7j5370Tf0P0elxKNeBDw78UdpYgilqeJ9FBG2SLlyNnk=; Proxy: null)
```

Job name	Start time (Local)	Glue version	Last modified on (Local)
first-etl-job	10/17/2025 23:13:50	5.0	10/17/2025 23:15:18
Id	End time (Local)	Worker type	Log group name

- Now Go to IAM role and give the permissions



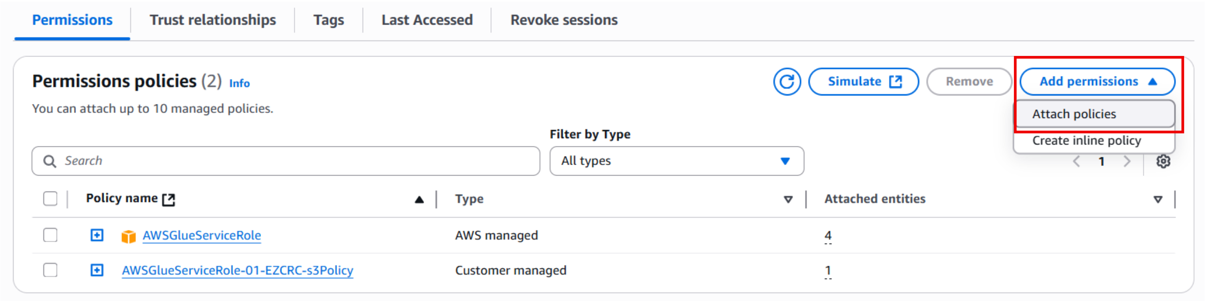
Roles (860) Info

An IAM role is an identity you can create that has specific permissions with credentials that are valid for short durations. Roles can be assumed by entities that you trust.

01 — 11 matches

	Role name	Trusted entities	Last activity
□	AWSCodePipelineServiceRole-eu-west-1-AngularProject01	AWS Service: codepipeline	431 days ago
□	AWSGlueServiceRole-01	AWS Service: glue	9 minutes ago
□	cwe-role-eu-west-1-AngularProject01	AWS Service: events	431 days ago
□	KinesisFirehoseServiceRole-KDS-S3-Nkcm1-ap-south-1-1698115150124	AWS Service: firehose	722 days ago
□	KinesisFirehoseServiceRole-KDS-S3-TA7Ze-ap-south-1-1701011141917	AWS Service: firehose	687 days ago

- Under Permissions, we have to attach policies for the S3 full access.

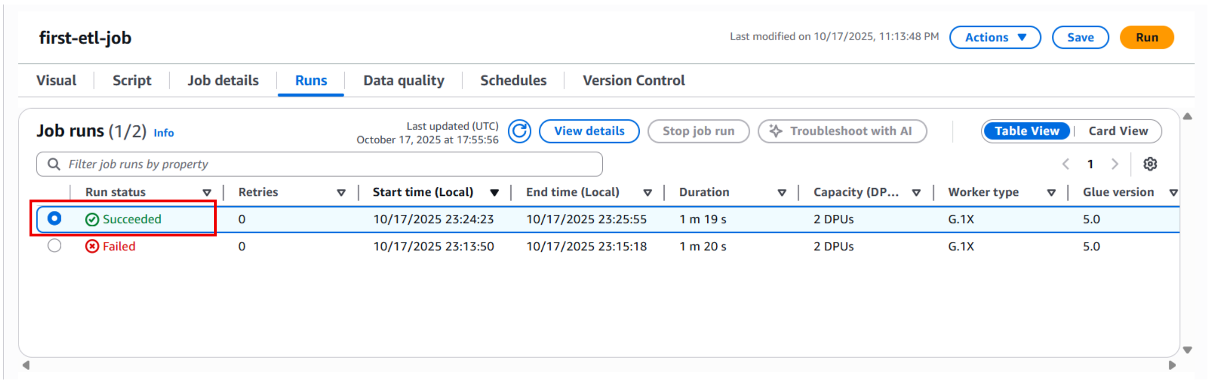


Permissions policies (2) Info

You can attach up to 10 managed policies.

Policy name	Type	Attached entities
AWSGlueServiceRole	AWS managed	4
AWSGlueServiceRole-01-EZCRC-s3Policy	Customer managed	1

- Go back to the Job and Rerun it.



first-etl-job

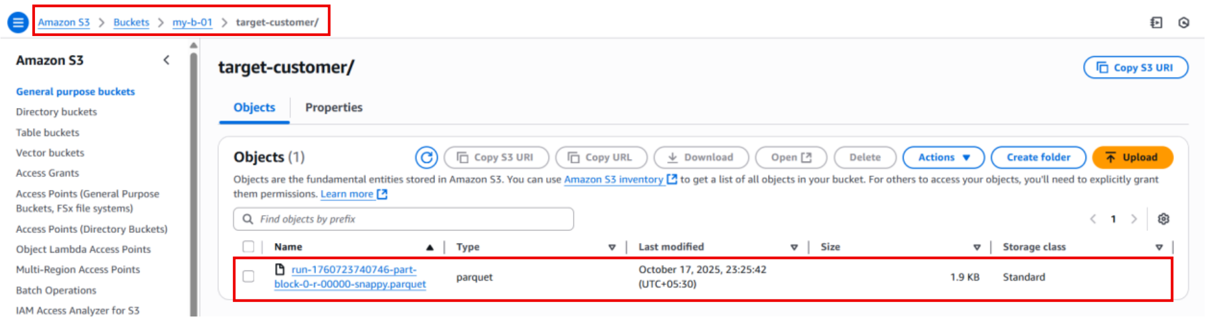
Last modified on 10/17/2025, 11:13:48 PM

Job runs (1/2) Info

Last updated (UTC)
October 17, 2025 at 17:55:56

	Run status	Retries	Start time (Local)	End time (Local)	Duration	Capacity (DP...	Worker type	Glue version
●	Succeeded	0	10/17/2025 23:24:23	10/17/2025 23:25:55	1 m 19 s	2 DPUs	G.1X	5.0
○	Failed	0	10/17/2025 23:13:50	10/17/2025 23:15:18	1 m 20 s	2 DPUs	G.1X	5.0

- Go back to your target-customer folder and check for the file created.



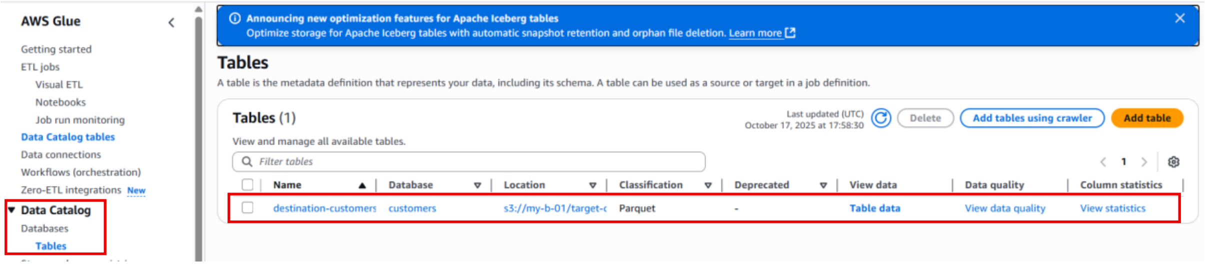
target-customer/

Objects (1)

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
run-1760723740746-part-block-0-r-00000-snappy.parquet	parquet	October 17, 2025, 23:25:42 (UTC+05:30)	1.9 KB	Standard

- Also, we can go the Data Catalog and in database, under tables we can see the table.



Announcing new optimization features for Apache Iceberg tables
Optimize storage for Apache Iceberg tables with automatic snapshot retention and orphan file deletion. Learn more

Tables

A table is the metadata definition that represents your data, including its schema. A table can be used as a source or target in a job definition.

Tables (1)

Last updated (UTC)
October 17, 2025 at 17:58:30

View and manage all available tables.

Name	Database	Location	Classification	Deprecated	View data	Data quality	Column statistics
destination-customers	customers	s3://my-b-01/target-c	Parquet	-	Table data	View data quality	View statistics