

Hands-on #12: AWS Services Lambda & Glue & Athena :

Serverless ETL Pipeline

Tarun Kumar Kanakala

801384590

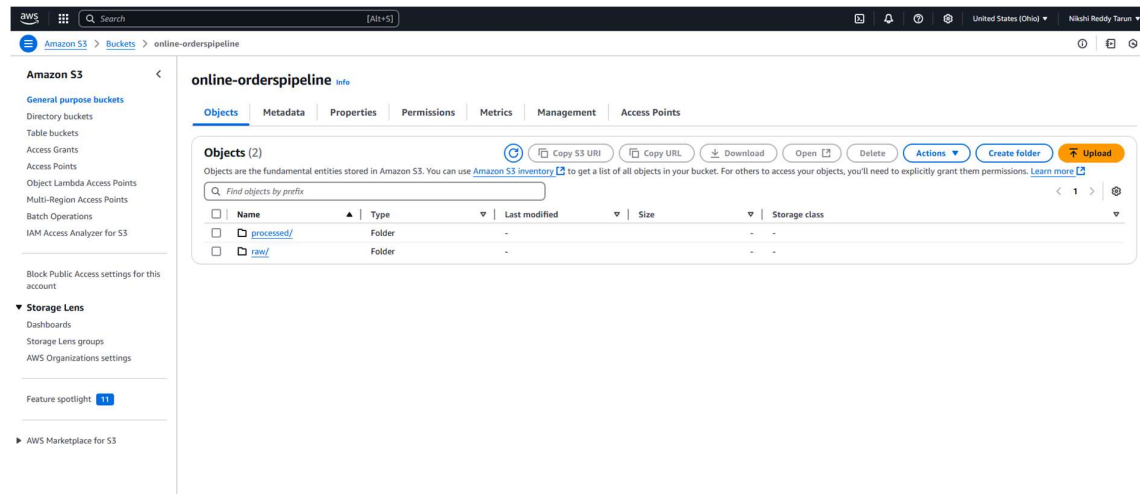
GITHUB LINK-

<https://github.com/K-Tarunkumar/ServerlessETLPipeline.git>

Step-by-step setup

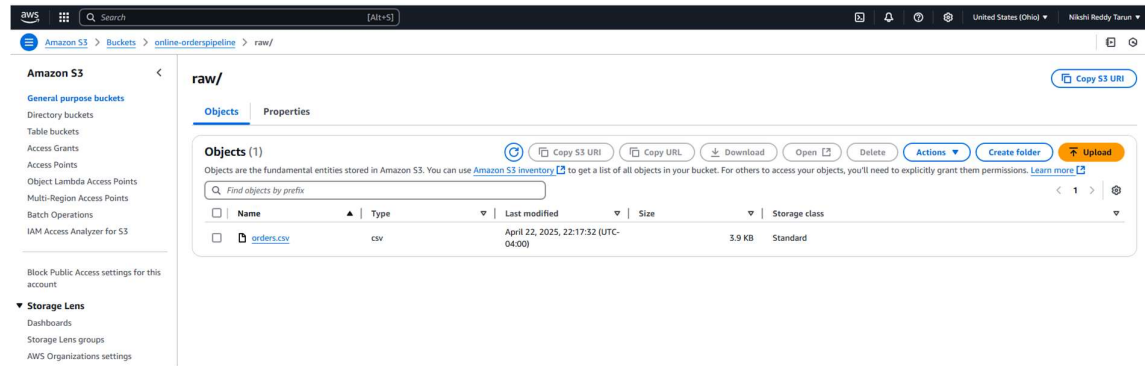
1. Create the bucket and folders

- Create an S3 bucket, for example online-orders-pipeline.
- Inside it, add two empty folders named raw/ and processed/.



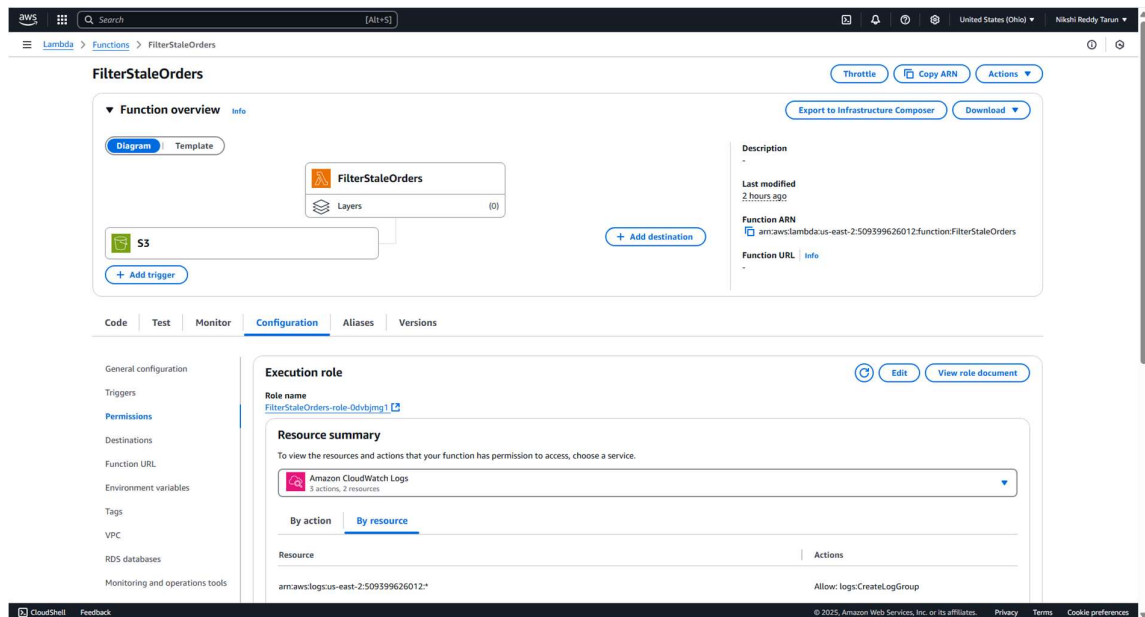
2. Generate sample data locally

Run the following Python script; it creates orders.csv with 100 random rows.



3. Create the Lambda function

Handler (Python 3.9 or newer)



4. Set up a Glue crawler

1. AWS Glue → Crawlers → Create.
2. Source: s3://<bucket>/processed/
3. Target database: create (or reuse) one, e.g. orders_db.
4. Table name: orders_processed.
5. Run the crawler after Lambda finishes.

crawler successfully starting
The following crawler is now starting: "Create crawler"

Create crawler Last updated (UTC) April 23, 2025 at 02:24:52 [Run crawler](#) [Edit](#) [Delete](#)

Crawler properties

Name Create crawler	IAM role AWSGlueServiceRole-order	Database orders_db	State READY
Description -	Security configuration -	Lake Formation configuration -	Table prefix orders_processed
Maximum table threshold -			

[Advanced settings](#)

Crawler runs | [Schedule](#) | [Data sources](#) | [Classifiers](#) | [Tags](#)

Crawler runs (2) Stop run View CloudWatch logs View run details

The list of crawler runs for this crawler.

Start time (UTC)	End time (UTC)	Current/last duration	Status	DPU hours	Table changes
April 23, 2025 at 02:24:54	April 23, 2025 at 02:25:39	44 s	Completed	0.126	-
April 23, 2025 at 02:17:56	April 23, 2025 at 02:18:41	45 s	Completed	0.132	1 table change, 2 partition changes

crawler successfully starting
The following crawler is now starting: "Create crawler"

Crawlers Last updated (UTC) April 23, 2025 at 02:59:15 [Action](#) [Run](#) [Create crawler](#)

A crawler connects to a data store, progresses through a prioritized list of classifiers to determine the schema for your data, and then creates metadata tables in your data catalog.

Crawlers (1) [info](#)

View and manage all available crawlers.

Name	State	Schedule	Last run	Last run timestamp	Log	Table changes from last run
Create crawler	Ready		Succeeded	April 23, 2025 at 02:24:54	View log	-

5. Query the data with Athena

Make sure Athena's query-result location is set, then run:

Amazon Athena Query editor

Data

Data source: AwsDataCatalog

Catalog: None

Database: orders_db

Tables and views: [Create](#)

Filter tables and views

Tables (1)

orders_processedonline_orderspipeline

Views (0)

SQL Ln 1, Col 76

```
SELECT * FROM "orders_db"."orders_processedonline_orderspipeline" limit 10;
```

[Run again](#) [Explain](#) [Cancel](#) [Clear](#) [Create](#)

Query results **Query stats**

Completed Time in queue: 58 ms Run time: 803 ms Data scanned: 3.88 KB

[Copy](#) [Download results CSV](#)

Results (10)

Search rows

#	orderid	customer	amount	status	orderdate	partition_0
1	O0001	Eve	380.12	cancelled	2025-03-27	raw
2	O0002	Bob	450.6	pending	2025-04-21	raw
3	O0003	Alice	319.59	pending	2025-02-21	raw
4	O0004	Bob	469.95	confirmed	2025-03-18	raw
5	O0005	Charlie	26.25	confirmed	2025-02-25	raw

aws [Search] [Alt+S] United States (Ohio) Nikshi Reddy Tarun

Amazon Athena Query editor

Data

Data source: AwsDataCatalog

Catalog: None

Database: orders_db

Tables and views: [Create](#)

Filter tables and views

Tables (1)

orders_processedonline_orderspipeline

Views (0)

SQL Ln 3, Col 46

```
1 SELECT SUM(amount) AS total_revenue
2 FROM "orders_db"."orders_processedonline_orderspipeline"
3 WHERE status IN ('confirmed', 'shipped');
```

[Run again](#) [Explain](#) [Cancel](#) [Clear](#) [Create](#)

Query results **Query stats**

Completed Time in queue: 61 ms Run time: 658 ms Data scanned: 3.88 KB

[Copy](#) [Download results CSV](#)

Results (1)

Search rows

#	total_revenue
1	13719.79

6. CloudWatch logs

CloudWatch Logs → Log groups → /aws/lambda/<function-name>

CloudWatch

Log groups

Log groups (2)

By default, we only load up to 10000 log groups.

Filter log groups or try prefix search

☐ Exact match

<input type="checkbox"/>	Log group	Log class	Anomaly d...	Data protection	Sensitive data co...	Retention	Metric filters	Contributor Insi...
<input type="checkbox"/>	/aws-glue/crawlers	Standard	Configure	-	-	Never expire	-	-
<input type="checkbox"/>	/aws/lambda/FilterStateOrders	Standard	Configure	-	-	Never expire	-	-

CloudWatch

Log groups

Log groups (2)

By default, we only load up to 10000 log groups.

Filter log groups or try prefix search

☐ Exact match

<input type="checkbox"/>	Log group	Log class	Anomaly d...	Data protection	Sensitive data co...	Retention	Metric filters	Contributor Insi...
<input type="checkbox"/>	/aws-glue/crawlers	Standard	Configure	-	-	Never expire	-	-
<input type="checkbox"/>	/aws/lambda/FilterStateOrders	Standard	Configure	-	-	Never expire	-	-

CloudWatch

Log groups

/aws-glue/crawlers

Actions

View in Logs Insights

Start tailing

Search log group

Log group details

Log class
Standard

ABN
[arn:aws:logsus-east-2:509399626012:log-group:/aws-glue/crawlers*](#)

Creation time
40 minutes ago

Retention
Never expire

Stored bytes
-

Metric filters
0

Subscription filters
0

Contributor Insights rules
-

KMS key ID
-

Anomaly detection
[Configure](#)

Data protection
-

Sensitive data count
-

Field indexes
[Configure](#)

Transformer
[Configure](#)

Log streams (1)

Filter log streams or try prefix search

☐ Exact match ☐ Show expired [Info](#)

Delete

Create log stream

Search all log streams

<input type="checkbox"/>	Log stream	Last event time
<input type="checkbox"/>	Create crawler	2025-04-23 02:25:59 (UTC)