

Project Report: AWS Serverless Data Pipeline – Online Orders Demo

By- Vivek Sai Chinna Burada

Project Overview

This project demonstrates a cost-effective serverless data pipeline using AWS services to process online order data. The pipeline ingests a CSV file of orders, filters outdated pending/cancelled entries, catalogs the cleaned data, and allows ad-hoc querying.

AWS Services Used

- **Amazon S3** – Stores raw and processed order files.
 - **AWS Lambda** – Processes data to filter unwanted records.
 - **AWS Glue** – Crawls the processed data and creates a catalog.
 - **Amazon Athena** – Executes SQL queries on cleaned data.
 - **Amazon CloudWatch** – Logs Lambda execution details and metrics.
-

Goal & Data Flow

1. Upload `orders.csv` to S3 → raw/
2. S3 trigger invokes Lambda
3. Lambda filters out pending/cancelled orders older than 30 days
4. Lambda saves filtered data to S3 → processed/
5. Glue crawler catalogs the cleaned data
6. Athena queries cleaned data

A screenshot of a web browser displaying the AWS S3 console. The URL in the address bar is https://us-east-1.console.aws.amazon.com/s3/buckets/handson14?region=us-east-1&bucketType=general&tab=objects. The page title is 'handson14 - AWS S3'. The navigation bar shows 'Amazon S3 > Buckets > handson14'. Below the navigation, there are tabs for Objects, Metadata, Properties, Permissions, Metrics, Management, and Access Points. The 'Objects' tab is selected. A sub-header 'Objects (3)' is shown. Below it, a note says '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.' There is a link 'Learn more'. A search bar 'Find objects by prefix' is present. A table lists the objects: 'processed/' (Folder), 'queries_output/' (Folder), and 'raw/' (Folder). Each row has a checkbox, a name, type, last modified, size, and storage class columns. Action buttons at the top include Copy S3 URI, Copy URL, Download, Open, Delete, Actions (with a dropdown arrow), Create folder, and Upload. At the bottom right of the table area, there are navigation arrows and a settings gear icon. The footer includes links for CloudShell, Feedback, Privacy, Terms, and Cookie preferences, along with a copyright notice: © 2025, Amazon Web Services, Inc. or its affiliates.

S3 Bucket Setup

- Bucket name: handson14
- Folders:
 - raw/: contains uploaded orders.csv
 - processed/: receives Lambda output filtered_orders.csv

The screenshot shows the AWS S3 console interface. The URL in the browser is <https://us-east-1.console.aws.amazon.com/s3/buckets/handson14?region=us-east-1&bucketType=general&prefix=processed/&showversions=false>. The page displays a single object named "filtered_orders.csv" in the "processed" folder of the "handson14" bucket. The object is a CSV file, last modified on April 22, 2025, at 13:19:24 (UTC-04:00), and has a size of 5.4 KB. The storage class is Standard. There are buttons for Actions (Copy S3 URI, Copy URL, Download, Open, Delete, Create folder, Upload), a search bar, and navigation controls.

Name	Type	Last modified	Size	Storage class
filtered_orders.csv	csv	April 22, 2025, 13:19:24 (UTC-04:00)	5.4 KB	Standard

✍ Data Generation (Local)

Python script generated `orders.csv` with 100 random orders and statuses (`confirmed`, `shipped`, `pending`, `cancelled`).

The screenshot shows a spreadsheet interface with a toolbar at the top. The toolbar includes icons for View, Zoom, Add Category, Pivot Table, Insert, Table, Chart, Text, Shape, Media, Comment, Share, Format, and Organize. A message bubble says "Table data was imported and can be adjusted." The sheet name is "Sheet 1".

filtered_orders (1)

OrderID	Customer	Amount	Status	OrderDate
00001	Bob	478.65	shipped	2025-02-10
00002	Alice	481.53	confirmed	2025-03-14
00003	Diana	230.68	confirmed	2025-04-21
00004	Diana	278.81	confirmed	2025-03-26
00006	Bob	308.72	confirmed	2025-03-10
00007	Eve	221.16	confirmed	2025-03-22
00009	Charlie	269.78	shipped	2025-02-21
00010	Eve	416.44	shipped	2025-02-14
00011	Alice	438.24	shipped	2025-02-22
00012	Charlie	361.27	shipped	2025-02-13
00014	Charlie	383.52	confirmed	2025-04-21
00015	Charlie	119.23	shipped	2025-01-29
00017	Bob	491.37	shipped	2025-04-08
00018	Charlie	97.72	shipped	2025-04-01
00019	Charlie	464.88	shipped	2025-02-15
00021	Alice	147.31	confirmed	2025-01-30
00022	Charlie	187.01	shipped	2025-02-24
00024	Eve	484.45	confirmed	2025-03-08
00025	Bob	307.55	confirmed	2025-04-04
00029	Bob	25.12	pending	2025-04-04
00030	Alice	157.44	shipped	2025-04-19
00031	Diana	219.11	cancelled	2025-03-27
00032	Bob	274.62	shipped	2025-02-23
00034	Bob	338.59	shipped	2025-01-27
00035	Bob	147.83	pending	2025-03-29
00037	Eve	318.62	shipped	2025-02-16
00038	Charlie	446.5	confirmed	2025-02-14
00042	Diana	267.57	confirmed	2025-02-19
00043	Diana	259.2	confirmed	2025-02-25

🧠 Lambda Function – Core Logic

- Triggered by S3 event (raw/ prefix)
- Filters out:
 - Orders with **status = pending/cancelled**
 - AND **older than 30 days**
- Writes filtered CSV to processed/ folder

Screenshot of the AWS CloudWatch Log Events page for the log group /aws/lambda/ecommerce_analyst. The page shows log events from April 22, 2025, with a timestamp range from 2025-04-22T17:19:22.825Z to 2025-04-22T17:19:23.246Z. The log entries include various Lambda function logs, such as access denied errors, file processing logs, and report generation logs.

Timestamp	Message
2025-04-22T17:19:22.825Z	[ERROR] ClientError: An error occurred (AccessDenied) when calling the GetObject operation: User: arn:aws:sts::...
2025-04-22T17:19:23.116Z	END RequestId: a1bc1609-0e18-40d2-9ba8-4b1c0d5c1178 Duration: 298.20 ms Billed Duration: 299 ms Memory Size:...
2025-04-22T17:19:23.116Z	REPORT RequestId: a1bc1609-0e18-40d2-9ba8-4b1c0d5c1178 Duration: 298.20 ms Billed Duration: 299 ms Memory Size:...
2025-04-22T17:19:23.127Z	START RequestId: 63f29701-ef32-4855-b1ba-87864614b007 Version: \$LATEST
2025-04-22T17:19:23.127Z	Lambda triggered by S3 event.
2025-04-22T17:19:23.127Z	Incoming file: raw/orders.csv
2025-04-22T17:19:23.127Z	Successfully read file from S3: orders.csv
2025-04-22T17:19:23.127Z	Processing records...
2025-04-22T17:19:23.127Z	Total records processed: 200
2025-04-22T17:19:23.127Z	Records filtered out: 61
2025-04-22T17:19:23.127Z	Records kept: 139
2025-04-22T17:19:23.127Z	Filtered file successfully written to S3: processed/filtered_orders.csv
2025-04-22T17:19:23.246Z	END RequestId: 63f29701-ef32-4855-b1ba-87864614b007 Duration: 419.69 ms Billed Duration: 420 ms Memory Size:...
2025-04-22T17:19:23.246Z	REPORT RequestId: 63f29701-ef32-4855-b1ba-87864614b007 Duration: 419.69 ms Billed Duration: 420 ms Memory Size:...

Screenshot of the AWS CloudWatch Log Events page for the log group /aws-glue/crawlers/orders_crawler. The page shows log events from April 22, 2025, with a timestamp range from 2025-04-22T17:35:09.372Z to 2025-04-22T17:36:01.140Z. The log entries detail the execution of a crawler named "orders_crawler", including benchmarking, classification results, and table creation logs.

Timestamp	Message
2025-04-22T17:35:09.372Z	[84cf904b-7a06-461d-a31a-a0b604732b38] BENCHMARK : Running Start Crawl for Crawler orders_crawler
2025-04-22T17:35:48.439Z	[84cf904b-7a06-461d-a31a-a0b604732b38] BENCHMARK : Classification complete, writing results to database ecomm
2025-04-22T17:35:48.456Z	[84cf904b-7a06-461d-a31a-a0b604732b38] INFO : Crawler configured with Configuration {"Version":1.0,"CreatePartiti...
2025-04-22T17:35:59.992Z	[84cf904b-7a06-461d-a31a-a0b604732b38] INFO : Created table processed in database ecomm
2025-04-22T17:36:01.117Z	[84cf904b-7a06-461d-a31a-a0b604732b38] BENCHMARK : Finished writing to Catalog
2025-04-22T17:36:01.139Z	[84cf904b-7a06-461d-a31a-a0b604732b38] INFO : Run Summary For TABLE:
2025-04-22T17:36:01.140Z	[84cf904b-7a06-461d-a31a-a0b604732b38] INFO : ADD: 1

AWS Glue – Data Catalog

- Glue crawler setup:
 - **Source:** s3://handson14/processed/
 - **Target DB:** orders_crawler
 - **Table name:** processed
- Run once after Lambda execution

The screenshot shows the AWS Glue Data Catalog interface. On the left, there's a sidebar with navigation links for AWS Glue, Data Catalog, Data Integration and ETL, and Legacy pages. The main area displays the 'Table details' for a table named 'processed'. It shows the table is associated with a database 'ecommerce' and a location 's3://handson14/processed/'. The schema section shows five columns: 'orderid' (string), 'customer' (string), 'amount' (double), 'status' (string), and 'orderdate' (string). The schema is defined as follows:

#	Column name	Data type	Partition key	Comment
1	orderid	string	-	-
2	customer	string	-	-
3	amount	double	-	-
4	status	string	-	-
5	orderdate	string	-	-

Athena Queries & Output

1. Recent 10 Orders

```
sql
CopyEdit
SELECT *
FROM orders_processed
ORDER BY orderdate DESC
LIMIT 10;
```

2. Total Revenue of Fulfilled Orders

```
sql
CopyEdit
```

```
SELECT SUM(amount) AS total_revenue
FROM orders_processed
WHERE status IN ('confirmed', 'shipped');
```

The screenshot shows the Amazon Athena Query Editor interface. On the left, there's a sidebar for 'Data' configuration, including 'Data source' set to 'AwsDataCatalog', 'Catalog' set to 'None', and 'Database' set to 'ecomm'. Below that is a 'Tables and views' section with a 'Create' button and a search bar. Under 'Tables (1)', there's a single entry for 'processed'. The main area contains two queries:

- Query 1:** SELECT round(SUM(amount),2) AS total_revenue FROM processed WHERE status IN ('confirmed', 'shipped');
- Query 2:** (This part is currently empty)

Below the queries is a 'SQL' panel showing 'Ln 3, Col 46'. At the bottom of the editor are buttons for 'Run again', 'Explain', 'Cancel', 'Clear', and 'Create'. A note indicates that typeahead suggestions are turned on by default. The status bar at the bottom right shows 'Completed', 'Time in queue: 131 ms', 'Run time: 550 ms', and 'Data scanned: 5.38 KB'.

The screenshot shows the Amazon S3 console. On the left, there's a navigation sidebar with sections for 'Amazon S3', 'General purpose buckets', 'Storage Lens', 'Feature spotlight', and 'AWS Marketplace for S3'. The main area displays a list of objects in the 'queries_output' folder of the 'handson14' bucket. The objects are:

Name	Type	Last modified	Size	Storage class
0493724a-695f-4751-a5ae-cdbe7be62cda.csv	csv	April 22, 2025, 13:50:01 (UTC-04:00)	26.0 B	Standard
0493724a-695f-4751-a5ae-cdbe7be62cda.csv.metadata	metadata	April 22, 2025, 13:50:01 (UTC-04:00)	83.0 B	Standard
1ccfac00-0cf1-44fb-b253-8d4159e2c26e.csv	csv	April 22, 2025, 13:49:39 (UTC-04:00)	37.0 B	Standard
1ccfac00-0cf1-44fb-b253-8d4159e2c26e.csv.metadata	metadata	April 22, 2025, 13:49:39 (UTC-04:00)	83.0 B	Standard
76b27b9f-54fd-4815-b08c-7340224e9f5e.csv	csv	April 22, 2025, 13:48:30 (UTC-04:00)	536.0 B	Standard
76b27b9f-54fd-4815-b08c-7340224e9f5e.csv.metadata	metadata	April 22, 2025, 13:48:30 (UTC-04:00)	261.0 B	Standard
c19626b3-60ff-4e86-bc82-d03d8f46edeb.csv	csv	April 22, 2025, 13:46:10 (UTC-04:00)	6.6 KB	Standard
c19626b3-60ff-4e86-bc82-d03d8f46edeb.csv.metadata	metadata	April 22, 2025, 13:46:11 (UTC-04:00)	261.0 B	Standard

The status bar at the bottom right shows 'Completed', 'Time in queue: 131 ms', 'Run time: 550 ms', and 'Data scanned: 5.38 KB'.

Mac OS X desktop showing two windows:

Numbers window (Sheet 1):

orderid	customer	amount	status	orderdate
00003	Diana	230.68	confirmed	2025-04-21
00063	Diana	305.64	cancelled	2025-04-21
00129	Alice	206.43	shipped	2025-04-21
00014	Charlie	383.52	confirmed	2025-04-21
00030	Alice	157.44	shipped	2025-04-19
00103	Alice	226.29	cancelled	2025-04-19
00197	Diana	186.6	pending	2025-04-19
00168	Eve	312.9	confirmed	2025-04-18
00134	Eve	65.21	pending	2025-04-16
00051	Diana	255.48	pending	2025-04-16

CloudWatch window:

- Log groups: New
- Log Anomalies
- Live Tail
- Logs Insights: New
- Contributor Insights
- Metrics
- X-Ray traces: New
- Events
- Application Signals
- Network Monitoring: ?
- Insights
- Settings
- Telemetry config: New
- Getting Started
- cloudShell
- Feedback

Mac OS X desktop showing two windows:

Numbers window (Sheet 1):

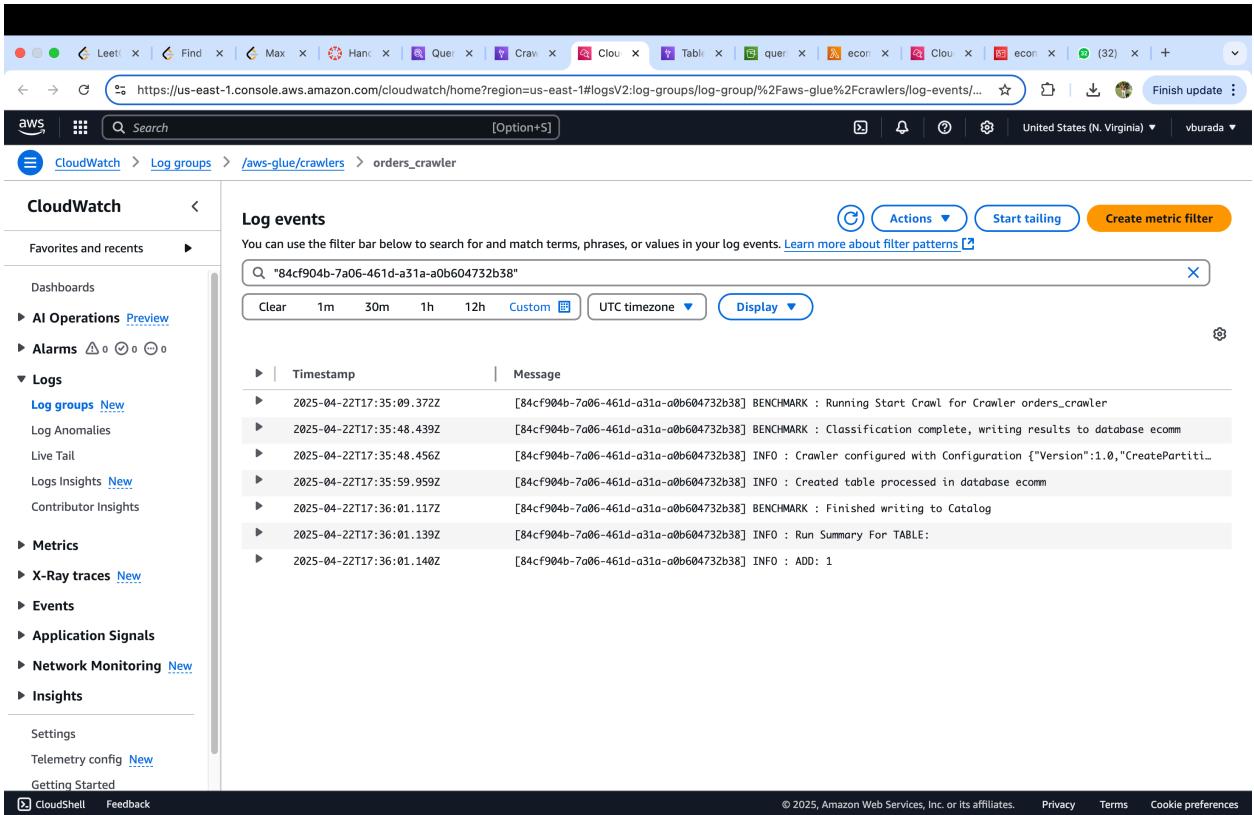
0493724a-69
total_revenue
30205.1

CloudWatch window (partially visible):

CloudWatch Logs

Each Lambda run logs:

- File name and path
- Number of records kept vs. filtered
- Success or errors



The screenshot shows the AWS CloudWatch Logs console. The left sidebar navigation includes CloudWatch, Favorites and recents, Dashboards, AI Operations (Preview), Alarms, Logs (Log groups New, Log Anomalies, Live Tail, Logs Insights New, Contributor Insights), Metrics, X-Ray traces New, Events, Application Signals, Network Monitoring New, and Insights. Under Logs, the 'Log groups' section is selected. The main area displays 'Log events' with a search bar containing the query "84cf904b-7a06-461d-a31a-a0b604732b38". Below the search bar are buttons for Clear, 1m, 30m, 1h, 12h, Custom, UTC timezone, and Display. A 'Actions' button, 'Start tailing' button, and 'Create metric filter' button are also present. The log events table has columns for Timestamp and Message. The table contains the following data:

Timestamp	Message
2025-04-22T17:35:09.372Z	[84cf904b-7a06-461d-a31a-a0b604732b38] BENCHMARK : Running Start Crawl for Crawler orders_crawler
2025-04-22T17:35:48.439Z	[84cf904b-7a06-461d-a31a-a0b604732b38] BENCHMARK : Classification complete, writing results to database ecomm
2025-04-22T17:35:48.456Z	[84cf904b-7a06-461d-a31a-a0b604732b38] INFO : Crawler configured with Configuration {"Version":1.0,"CreatePartiti...
2025-04-22T17:35:59.959Z	[84cf904b-7a06-461d-a31a-a0b604732b38] INFO : Created table processed in database ecomm
2025-04-22T17:36:01.117Z	[84cf904b-7a06-461d-a31a-a0b604732b38] BENCHMARK : Finished writing to Catalog
2025-04-22T17:36:01.139Z	[84cf904b-7a06-461d-a31a-a0b604732b38] INFO : Run Summary For TABLE:
2025-04-22T17:36:01.140Z	[84cf904b-7a06-461d-a31a-a0b604732b38] INFO : ADD: 1

Summary

Stage	Outcome
Data Ingestion	orders.csv uploaded to S3
Data Processing	Lambda filters unwanted rows
Cataloging	Glue crawler creates Athena table
Querying	Athena SQL queries executed
Monitoring	CloudWatch logs Lambda execution