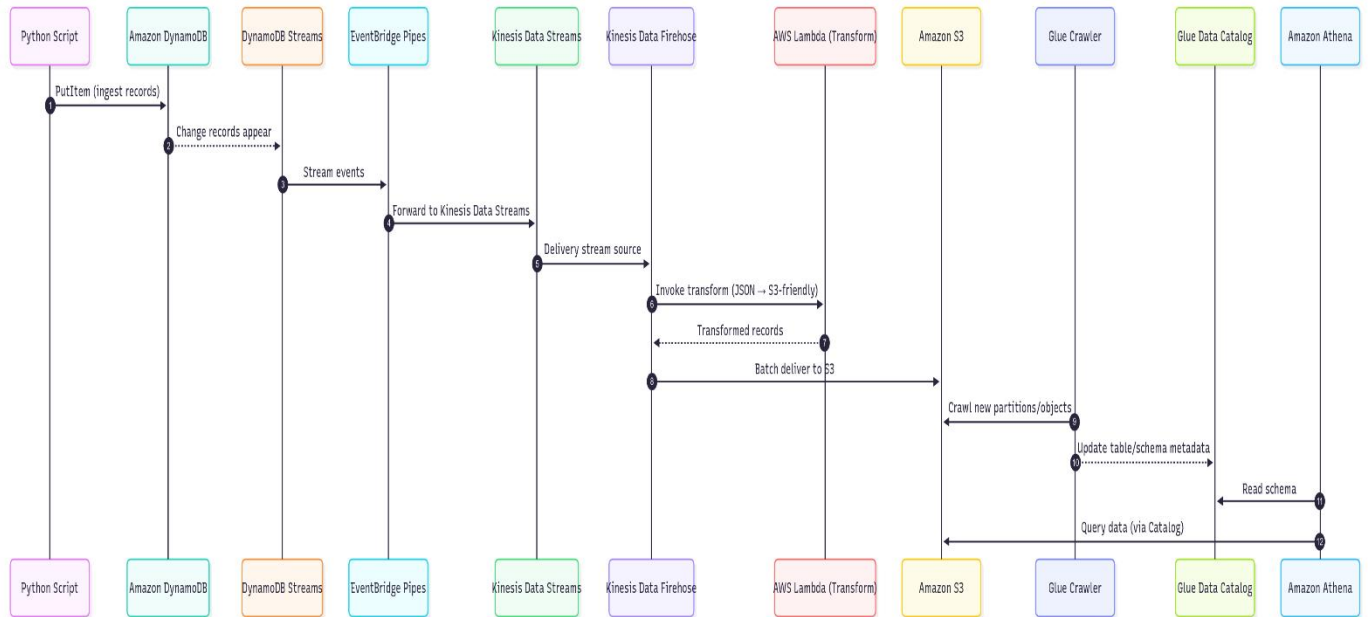


CDC Data Pipeline for Sales Data Analysis

This real time streaming pipeline aims to capture data changes in Dynamodb using Kinesis Data Streams, Kinesis Firehose, Glue, Lambda, S3 and Eventbridge. Finally, Athena is used to run queries for data analysis.

Architecture Diagram



Step 1: Create records in Dynamodb using Python script.

Python with boto3 pushes data to dynamodb.

Source

Target

Target [Info](#)

Send your event to an AWS service, an event bus, or an API destination.

Target

[kinesis-for-sales-data](#)

Amazon Kinesis stream

ARN

[arn:aws:kinesis:us-east-2:646047875925:stream/kinesis-for-sales-data](#)

Target Parameters

PartitionKey: \$.dynamodb.Keys.orderid.S

Settings

Encryption

Monitoring

Tags

Permissions

Execution Role

[service-role/Amazon_EventBridge_Pipe_sales-ingestion-dynamodb-to-kin_b477acd1](#)

IAM

Roles

Amazon_EventBridge_Pipe_sales-ingestion-dynamodb-to-kin_b477acd1

Identity and Access Management (IAM)

Search IAM

Dashboard

▼ Access management

User groups

Users

Roles

Policies

Identity providers

Account settings

Root access management [New](#)

▼ Access reports

Access Analyzer

Resource analysis [New](#)

Last activity

12 minutes ago

Maximum session duration

1 hour

Permissions

Trust relationships

Tags

Last Accessed

Revoke sessions

Permissions policies (4) [Info](#)

Search

Filter by Type

All types

☐

Policy name [?](#)

▲

☐

+

[AmazonDynamoDBFullAccess](#)

AWS managed

2

☐

+

[AmazonKinesisFullAccess](#)

AWS managed

4

☐

+

[DynamoDbPipeSourceTemplate-cc8c5949](#)

Customer managed

1

☐

+

[KinesisPipeTargetTemplate-4f26ee28](#)

Customer managed

1

Step 3: Create a Kinesis data stream to view the change data capture performed in DynamoDB

Choose the shard and accurate timestamp to view streamed data. Event name “Modify” implies the change made in data .

Amazon Kinesis

Dashboard

Data streams

Amazon Data Firehose [New](#)

Managed Apache Flink [New](#)

▼ Resources

CloudFormation templates

AWS Glue Schema Registry [New](#)

kinesis-for-sales-data

ACTIVE

On-demand

Data retention period
1 day

arn:aws:kinesis:us-east-2:646047875925:stream/kinesis-for-sales-data

September 15, 2025 at 01:19 EDT

Monitoring

Configuration

Enhanced fan-out (0)

Data viewer

Data analytics - new

Data stream sharing

EventBridge Pipelines

Shard

shardId-000000000001

Starting position

At timestamp

Start date

2025/09/15

Start time

22:00:00

Get records

Records (3)

Next records

Shard: shardId-000000000001 Starting position: At timestamp Timestamp: September 15, 2025 at 22:00:00 EDT

Find records

Partition key

Data

Approximate arrival timestamp

Sequence number

5097	("eventID":"7f335e5f273c9a90a3eb9731f...	September 15, 2025 at 22:01:52 EDT	4966704016339987045948449745346833...
8818	("eventID":"e20d59ff868832af3f04592f4...	September 15, 2025 at 22:01:53 EDT	4966704016339987045948449745401477...
3479	("eventID":"d92fda7d014fa9ab047cb5c2b...	September 15, 2025 at 22:01:54 EDT	4966704016339987045948449745471474...

Change data Capture

Record data

Sequence number

49667040163444471949881559382423358061850105594846904370

Shard ID

shardId-000000000003

Raw data

JSON

Copy

```
{
  "eventID": "1ed5db9dc70e69af4e0c00e0be4e1123",
  "eventName": "MODIFY",
  "eventVersion": "1.1",
  "eventSource": "aws:dynamodb",
  "awsRegion": "us-east-2",
  "dynamodb": {
    "ApproximateCreationDateTime": 1757989333,
    "Keys": {
      "orderid": {
```

Close

Step 4: Create Kinesis Data Firehose

The purpose of kinesis firehose is to Batch data coming from kinesis stream and dump into target. Here source is kinesis data stream and target is S3.

The screenshot shows the Amazon Data Firehose console for a stream named 'Kinesis-nrt-batch'. The left sidebar contains a 'Resources' section with links to 'What's new', 'Developer guide', and 'API reference'. The main content area is titled 'Kinesis-nrt-batch' and includes a 'Delete' button. Below the title is a 'Firehose stream details' section with the following information:

Status Active	Destination Amazon S3	Data transformation Enabled	Creation time September 15, 2025 at :
Source Amazon Kinesis Data Streams	ARN arn:aws:firehose:us-east-2:646047875925:deliverystream/Kinesis-nrt-batch	Dynamic partitioning Not enabled	Error logs status 0 Destination error logs

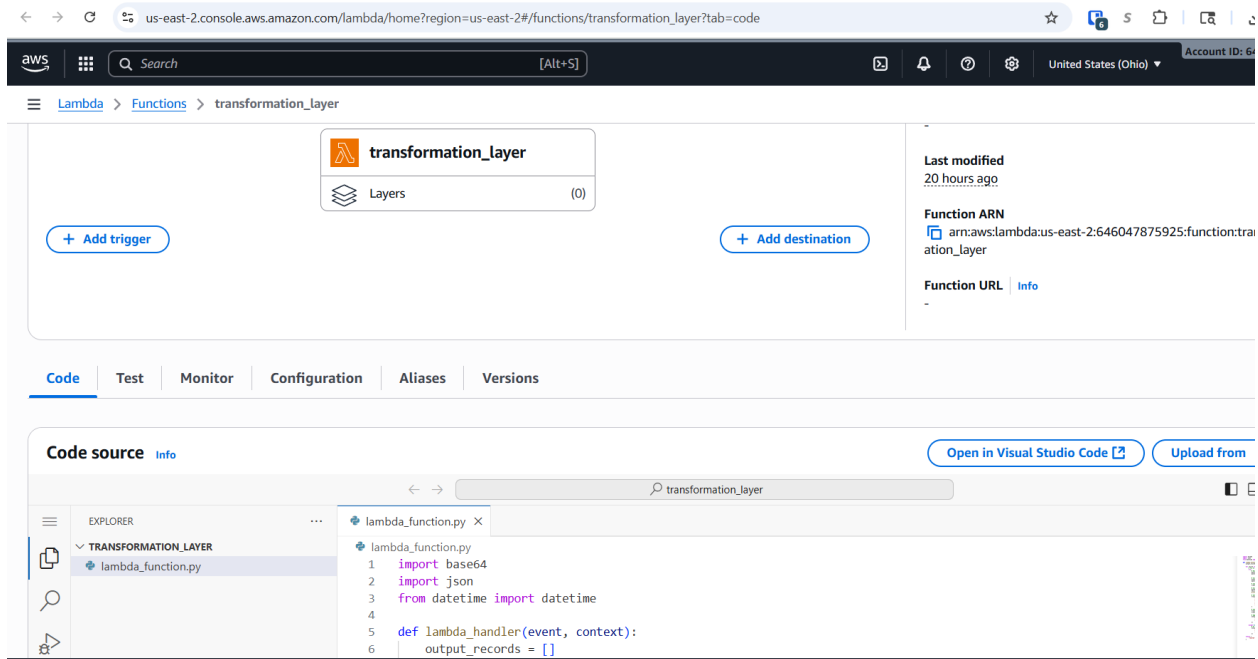
Step 5: Transformation using Lambda

Lambda function is used to transform the data from source Kinesis data stream and then push it back to Firehose. We use Lambda transformation since the data coming from Kinesis data stream is in JSON format and we want to push it to S3 in tabular format to write Athena queries. All kinds of transformations can be done using Lambda.

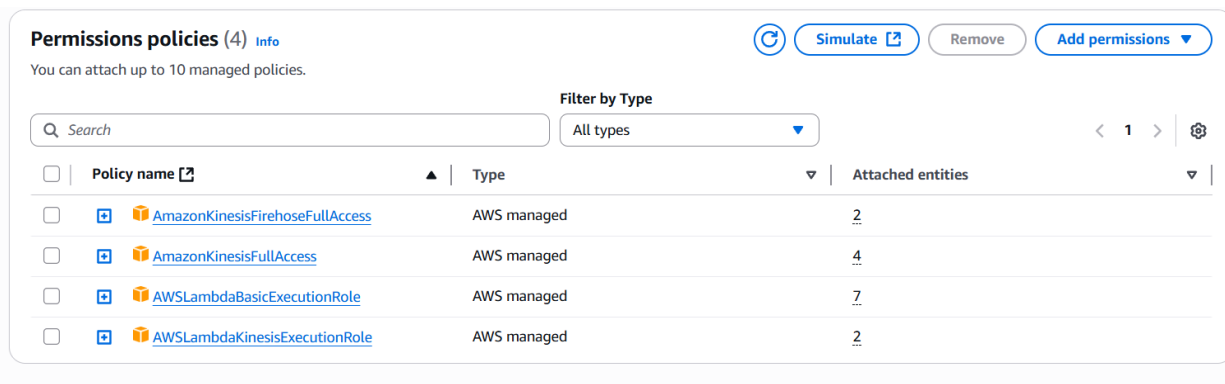
The screenshot shows the 'Configuration' page for the 'Kinesis-nrt-batch' stream. The left sidebar is the same as the previous screenshot. The main content area is titled 'Source settings' and includes a link to the 'Kinesis data stream' (kinesis-for-sales-data). Below this is a 'Transform and convert records' section with the following configuration:

Transform source records with AWS Lambda On	Lambda function transformation_layer	Runtime python3.13
Buffer size 1 MiB	Lambda function version \$LATEST	Timeout 15 minutes
Buffer interval 60 seconds	Description -	

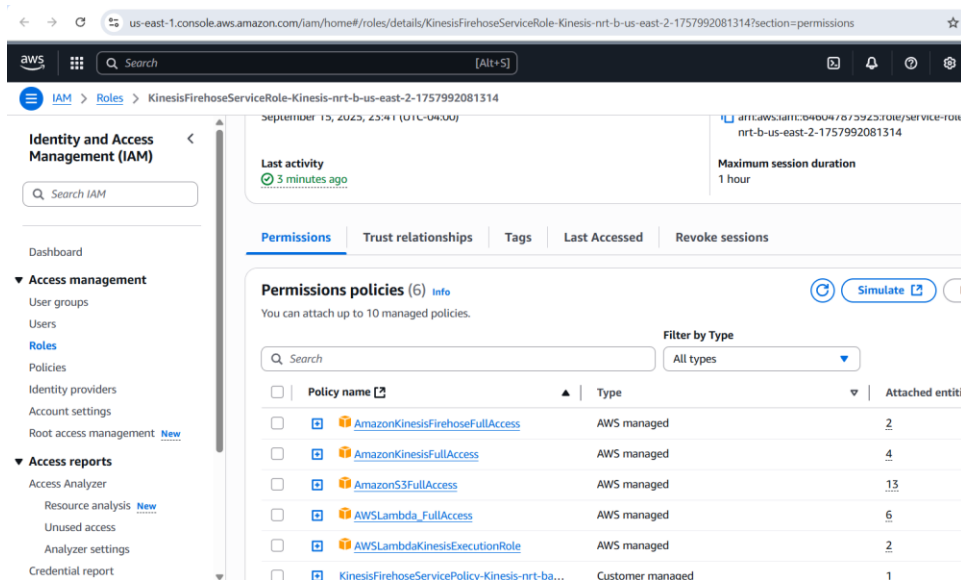
Below the table, there is a 'Convert record format' section which is currently 'Not enabled'.



Lambda code selects order id and other values from New image of kinesis data stream creates a new json and then pushes that back firehose which then dumps to S3. IAM roles for lambda function

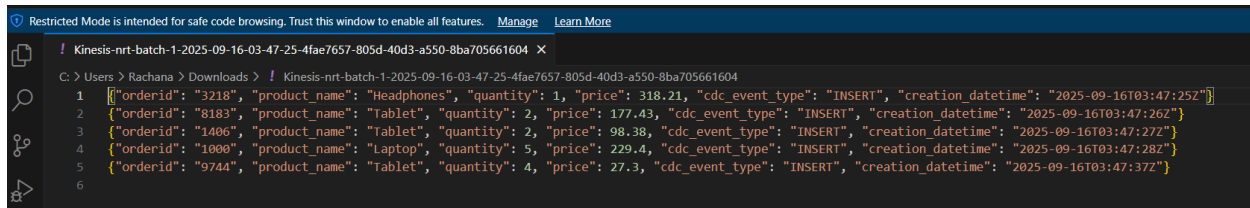
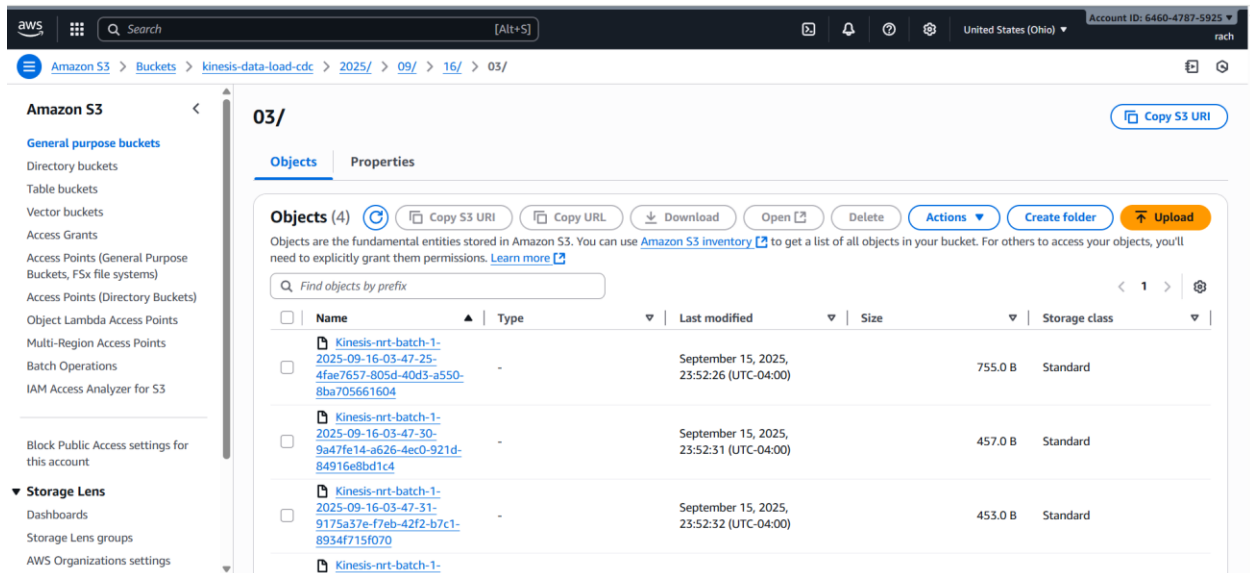


IAM role for kinesis firehose



Step 6: Target S3 bucket is created where data gets dumped in real time

S3 bucket contents in near real time



Step 7: Create a Glue crawler and catalog to crawl the contents in S3

Glue crawler to crawl data in S3 so that its ready for Athena querying

AWS Glue > Crawlers > kinesis-s3-crawler

kinesis-s3-crawler

Last updated (UTC)
September 16, 2025 at 23:46:15

Run crawler

Edit

Crawler properties

Name

kinesis-s3-crawler

IAM role

for_glue

Database

glue-sales-db

State

READY

Description

-

Security configuration

-

Lake Formation configuration

-

Table prefix

-

Maximum table threshold

-

Advanced settings

Crawler runs

Schedule

Data sources

Classifiers

Tags

Crawler runs (1)

Stop run

View CloudWatch logs

View run

The list of crawler runs for this crawler.

Filter data

Filter by a date and time range

< 1

	Start time (UTC)	End time (UTC)	Current/last duration	Status	DPU hours	Table changes
	September 16, 2025 at 22:17:36	September 16, 2025 at 22:18:23	46 s	Completed	0.141	1 table change, 1 partition change

Custom classifier for Glue crawler

AWS Glue > Classifiers > read-json

read-json

Last updated (UTC)
September 16, 2025 at 23:52:15

Edit

Delete

Classifier properties

Name

read-json

Json path

\$.orderid,\$.product_name,\$.quantity,\$.price

Last updated

September 16, 2025 at 22:14:22

Version

1

Glue table created

[AWS Glue](#) > [Tables](#) > 2025

2025

Database

[glue-sales-db](#)

Description

-

Last updated

September 16, 2025 at 22:18:22

JSON

Location

[s3://kinesis-data-load-cdc/2025/](#)

Connection

-

Column statistics

No statistics available

▶ Advanced properties

Schema

Partitions

Indexes

Column statistics - new

Schema (9)

View and manage the table schema.

Filter schemas

#	Column name	Data type	Partition key
1	orderid	string	-
2	product_name	string	-
3	quantity	int	-
4	price	double	-
5	cdc_event_type	string	-
6	creation_datetime	timestamp	-
7	partition_0	string	-
8	partition_1	string	-
9	partition_2	string	-

Step 8: Open Athena query editor to run queries against glue catalog table.

This data is used for analysis

Athena query

[Amazon Athena](#) > Query editor

Database

[glue-sales-db](#)

Tables and views

Create

Filter tables and views

Tables (1)

< 1 >

2025

Partitioned

Views (0)

< 1 >

SQL Ln 1, Col 22

Run again

Explain

Cancel

Clear

Create

Reuse query results

up to 60 minutes ago

Query results

Query stats

Completed

Time in queue: 95 ms

Run time: 786 ms

Data scanned: 2.07 K

Results (14)

Copy

Download results CSV

Search rows

#	orderid	product_name	quantity	price	cdc_event_type	creation_datetime	partition_0	partition_1
1	7090	Charger	4	272.98	INSERT	2025-09-16T03:47:30Z	09	16
2	3204	Tablet	4	333.23	INSERT	2025-09-16T03:47:36Z	09	16