

# Retail Data Management

MADE BY- Ashish Chamel

COURSE- Extract, Transform, and Load (ETL)

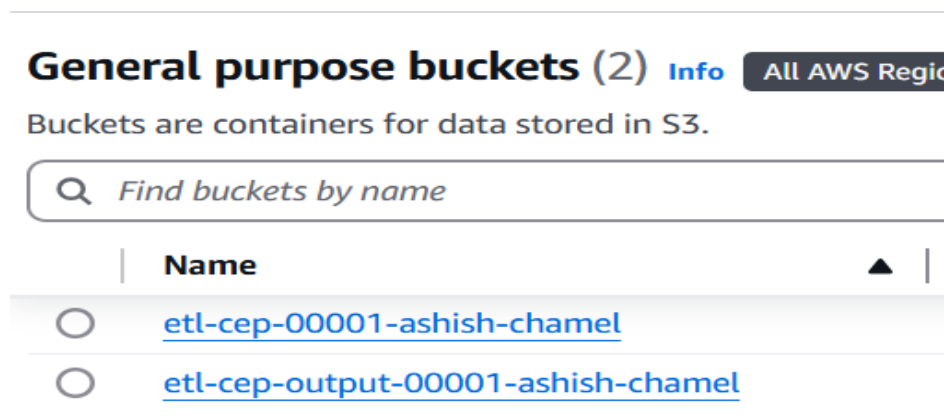
DATE OF SUBMISSION-05/04/2025

## Step-1

1. Created buckets in s3-

input bucket- etl-cep-00001-ashish-chamel (input files)

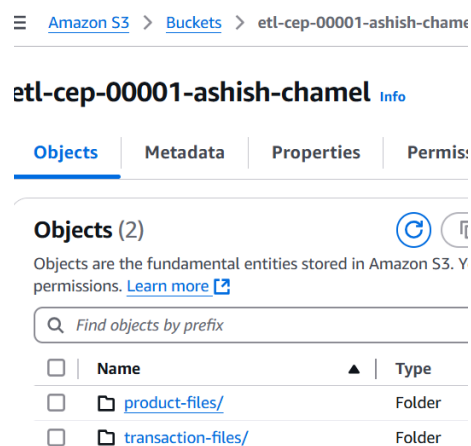
output bucket- etl-cep-output-00001-ashish-chamel(output from a etl job)



## Step-2

1. Created subfolders inside etl-cep-00001-ashish-chamel ie-

- Product-files
- transaction-files



2. Uploaded product details.csv , transaction.csv to corresponding folders ie product files and transaction files

Amazon S3 > Buckets > etl-cep-00001-ashish-chamel > product-files/

## product-files/

Objects

Properties

### Objects (1)



Copy S3 URI



Objects are the fundamental entities stored in Amazon S3. You can use [Amazon S3](#) permissions. [Learn more](#)

Find objects by prefix



Name



Type



[product details.csv](#)

csv



Upload succeeded

For more information, see the **Files and folders** table.



After you navigate away from this page, the following information is preserved:

### Summary

Destination

[s3://etl-cep-00001-ashish-chamel/transaction-files/](#)

Files and folders

Configuration

### Files and folders (1 total, 4.8 MB)

Find by name

Name

Folder

[transactions.csv](#)

-

Step-3

- 1. Created a new database ie. abc-retail

Databases (1)

A database is a set of associated table definitions, organized into a logical group.

Filter databases

<input type="checkbox"/>	Name		Description
<input type="checkbox"/>	<a href="#">abc-retail</a>		-

Step-4

- 1. Here I setup 2 classifiers ie cust\_classifier & txn class,to read transaction data and product data.

Classifiers

Classifiers are triggered during a crawl task. A classifier checks \ StructType object that matches that data format.

Classifiers (2) Info

View and manage all available classifiers.

Filter classifiers

<input type="checkbox"/>	Name		Type
<input type="checkbox"/>	<a href="#">cust_classifier</a>		CSV
<input type="checkbox"/>	<a href="#">txnClass</a>		CSV

- 2. Created the 1<sup>st</sup> classifier ie cust\_classifier with following settings
  - Classifier type and properties as CSV
  - CSV Serde – optional as None
  - Column delimiter as comma(,)
  - Quote symbol as Double-quote(“)
  - Column headings as Has headings

cust\_classifier

Last updated (UTC)  
April 2, 2025 at 12:42:51

Classifier properties

<b>Name</b> cust_classifier	<b>Allow single column</b> False	<b>CSV Serde</b> None
<b>Contains header</b> Has headings	<b>Header</b> -	<b>Creation time</b> April 2, 2025 at 12:42:48
<b>Delimiter</b> ,	<b>Disable value trimming</b> True	<b>Quote symbol</b> "
<b>Last updated</b> April 2, 2025 at 12:42:48	<b>Version</b> 1	<b>Custom datatypes</b> -

3. Created the 2<sup>nd</sup> classifier ie txnClass with following settings
- Classifier type and properties as CSV o CSV Serde – optional as None
  - Column delimiter as comma(,)
  - Quote symbol as Double-quote(“)
  - Column headings as Has headings ie  
Order ID, Order Date, Ship Date, Aging, Ship Mode, Product ID, Sales, Quantity, Discount, Profit, Shipping Cost, Order Priority, Customer ID

txnClass

Last updated (UTC)  
April 2, 2025 at 12:47:07

Classifier properties

<b>Name</b> txnClass	<b>Allow single column</b> False	<b>CSV Serde</b> None
<b>Contains header</b> Has headings	<b>Header</b> Order ID,Order Date,Ship Date,Aging,Ship Mode,Product ID,Sales,Quantity,Discount,Profit,Shipping Cost,Order Priority,Customer ID	<b>Creation time</b> April 2, 2025 at 12:46:58
<b>Delimiter</b> ,	<b>Disable value trimming</b> True	<b>Quote symbol</b> "
<b>Last updated</b> April 2, 2025 at 12:46:58	<b>Version</b> 1	<b>Custom datatypes</b> -

Step-5

1. Created a IAM role with administrator access as TrustedEntityType

TrustedEntityType

Info

Allows Glue to call AWS services on your behalf.

Summary

Creation date

April 02, 2025, 18:22 (UTC+05:30)

Last activity

-

Permissions

Trust relationships

Tags

Last Access

Permissions policies (1)

Info

You can attach up to 10 managed policies.

Search

Policy name

AdministratorAccess

AWS manag

## Step-6

1. Setting up two crawlers ie. retail-crawl and product-crawl for extracting metadata
2. Setup of a retail-crawl

### Choose data sources and classifiers

#### Data source configuration

Is your data already mapped to Glue tables?

☒ Not yet

Select one or more data sources to be crawled.

☐ Yes

Select existing tables from you

#### Data sources (1) [Info](#)

[Edit](#)

The list of data sources to be scanned by the crawler.

Type	Data source	Parameters
<input type="radio"/> S3	s3://etl-cep-00001-ashish-cha...	Recrawl all

#### ▼ Custom classifiers - *optional*

A classifier checks whether a given file is in a format the crawler can handle. If it is, the classifier creates a object that matches that data format.

##### Custom classifiers [Info](#)

Select one or more classifiers to use with this crawler.

Choose one or more classifiers



txnClass [×](#)

[Clear selection](#)

[Add new classifier](#) [↗](#)

tables with automatic snapshot retention and orphan file deletion. [Learn more](#) [↗](#)

### Configure security settings

#### IAM role [Info](#)

Existing IAM role

TrustedEntityType



[View](#) [↗](#)

[Create new IAM role](#)

[Update chosen IAM role](#)

Only IAM roles created by the AWS Glue console and have the prefix "AWSGlueServiceRole-" can be updated.

#### Lake Formation configuration - *optional*

Allow the crawler to use Lake Formation credentials for crawling the data source. [Learn more.](#) [↗](#)

☐ Use Lake Formation credentials for crawling S3 data source

Checking this box will allow the crawler to use Lake Formation credentials for crawling the data source. If the data source is the registered account ID. Otherwise, the crawler will crawl only those data sources associated to the account. Only applicable for S3 data sources.

#### ► Security configuration - *optional*

Enable at-rest encryption with a security configuration.

### Output configuration [Info](#)

Target database

abc-retail



[Clear selection](#)

[Add database](#) [↗](#)

Table name prefix - *optional*

txn|

3. Retail-crawl created and ran to extract metadata into database ie. abc-retail

One crawler successfully created  
The following crawler is now created: "retail-crawl"

retail-crawl

Last updated (UTC)  
April 2, 2025 at 13:01:33

Run

Crawler properties

Name retail-crawl	IAM role <a href="#">TrustedEntityType</a>	Database abc-retail	State READY
Description -	Security configuration -	Lake Formation configuration -	Table prefix txn
Maximum table threshold -			

retail-crawl

Crawler properties

Name retail-crawl	IAM role <a href="#">TrustedEntityType</a>	Database abc-retail
Description -	Security configuration -	Lake Formation configuration -
Maximum table threshold -		

► Advanced settings

Crawler runs

Schedule

Data sources

Classifiers

Tags

Crawler runs (1)

The list of crawler runs for this crawler.

Filter data

Filter by a date and time range

Start time (UTC)	End time (UTC)	Current/last duration	Status
April 2, 2025 at 13:02:22	April 2, 2025 at 13:04:24	02 min 02 s	Completed

4. Creating a 2<sup>nd</sup> crawler ie product-crawl to extract the metadata.

## Set crawler properties

### Crawler details [Info](#)

Name

product-crawl

Name can be up to 255 characters long. Some character set i

Description - optional

Enter a description

Descriptions can be up to 2048 characters long.

## Choose data sources and classifiers

### Data source configuration

Is your data already mapped to Glue tables?

☒ Not yet  
Select one or more data sources to be crawled.

☐ Yes  
Select existing tables

### Data sources (1) [Info](#)

The list of data sources to be scanned by the crawler.

	Type	Data source
<input type="radio"/>	S3	s3://etl-cep-00001-ashish-chamel/prod...

### ▼ Custom classifiers - *optional*

A classifier checks whether a given file is in a format the crawler can handle. If it is, the classifier object that matches that data format.

#### Custom classifiers [Info](#)

Select one or more classifiers to use with this crawler.

Choose one or more classifiers

cust\_classifier

Clear selection

Add new classifier

## Configure security settings

### IAM role [Info](#)

Existing IAM role

TrustedEntityType

Create new IAM role

Update chosen IAM role

Only IAM roles created by the AWS Glue console and have the prefix "AWSGlueServiceRole-" can be updated.

One crawler successfully created  
The following crawler is now created: "product-crawl"

## product-crawl

### Crawler properties

Name  
product-crawl

IAM role  
[TrustedEntityType](#)

Description  
-

Security configuration  
-

Maximum table threshold  
-

- Product-crawl successfully created and ran to extract metadata in the database abc-retail

🟢 **Crawler successfully starting**  
The following crawler is now starting: "product-crawl"

### product-crawl

April

**Crawler properties**

**Name**  
product-crawl

**Description**  
-

**Maximum table threshold**  
-

► **Advanced settings**

**IAM role**  
[TrustedEntityType](#)

**Security configuration**  
-

**Database**  
abc-retail

**Lake Formation configuration**  
-

**Crawler runs** | Schedule | Data sources | Classifiers | Tags

**Crawler runs (1)** 🔄 Stop

The list of crawler runs for this crawler.

🔍  📅

	Start time (UTC) ▲	End time (UTC) ▼	Current/last duration ▼	Status ▼
<input type="radio"/>	April 2, 2025 at 13:11:37	April 2, 2025 at 13:12:54	01 min 17 s	🟢 Completed

## Step-7

- Creation of a job in AWS GLUE .
- Importing AWS Glue Data Catalog from data sources and setting up data catalog 1 where
  - table I choose custproduct\_files
  - database as abc-retail
  -

The screenshot shows the AWS Glue console interface. On the left, a job configuration canvas displays a data source icon labeled "Data source - Data Catalog" with a green checkmark. A line connects this source to a job node. On the right, the "Data source properties - Data Catalog" panel is open, showing the following configuration:

- Name:** AWS Glue Data Catalog
- Database:** abc-retail (selected from a dropdown menu)
- Table:** custproduct\_files (selected from a dropdown menu)

- Importing AWS Glue Data Catalog from data sources and setting up data catalog 2 where in
  - table I choose txnttransaction\_files
  - and database as abc-retail

The screenshot shows the AWS Glue console interface, similar to the previous one, but with a different table selected. The "Data source properties - Data Catalog" panel on the right shows the following configuration:

- Name:** AWS Glue Data Catalog
- Database:** abc-retail (selected from a dropdown menu)
- Table:** txnttransaction\_files (selected from a dropdown menu)



#### 4. Importing & setting up Join from Transforms linking node parents as AWS Glue Data Catalog-1 and AWS Glue Data Catalog-1

- Join type as inner join
- join conditions as productid = product id

The screenshot shows the AWS Glue console interface. On the left, a workflow diagram displays a 'Transform - Join' node with a green checkmark. The right pane is titled 'Transform' and contains the following configuration:

- Name:** Join
- Node parents:** Choose one or more parent node. Two nodes are selected: 'AWS Glue Data Catalog - Catalog - DataSource' and 'AWS Glue Data Catalog - Catalog - DataSource'.
- Join type:** Inner join. Select all rows from both datasets that meet the join condition.
- Join conditions:** Select a field from each parent node for the join condition. The condition is set to 'productid = product id'.

At the bottom of the right pane, there is an 'Add condition' button.

#### 5. Importing & setting up drop fields from Transform Node parent as Join

- selecting productid.

The screenshot shows the AWS Glue console interface. On the left, a workflow diagram displays a 'Transform - DropFields' node with a green checkmark. The right pane is titled 'Field' and contains a table of fields to be dropped.

Field	Data type
<input checked="" type="checkbox"/> productid	long
<input type="checkbox"/> product	string
<input type="checkbox"/> product category	string
<input type="checkbox"/> order id	string
<input type="checkbox"/> order date	string
<input type="checkbox"/> ship date	string
<input type="checkbox"/> aging	long
<input type="checkbox"/> ship mode	string
<input type="checkbox"/> product id	long
<input type="checkbox"/> sales	long
<input type="checkbox"/> quantity	long
<input type="checkbox"/> discount	double
<input type="checkbox"/> profit	string
<input type="checkbox"/> shipping cost	string
<input type="checkbox"/> order priority	string
<input type="checkbox"/> customer id	string

At the top of the right pane, there is a warning message: 'Job has not been saved'. Below the table, there are buttons for 'Actions', 'Save', and 'Run'. At the bottom of the right pane, there is a 'Start session' button.

6. Importing & setting up Regex Extractor from transform- choosing node parents as Drop Fields.

- And Column to extract from as sales
- Regular expression as \d+
- Extracted column as NetSales

The screenshot displays the Alteryx interface with a workflow canvas on the left and a configuration panel on the right. The workflow canvas shows a 'Transform - Dynamic Transform' node with a 'Regex Extractor' sub-node. The configuration panel on the right is titled 'Transform' and contains the following settings:

- Name:** Regex Extractor
- Node parents:** Choose which nodes will provide inputs for this one.   
Choose one or more parent node   
Drop Fields   
DropFields - Transform
- Column to extract from:** String column on which to apply the regex.   
sales   
long
- Regular expression:** Regex to apply on the column, if multiple columns need to be extracted then the expression needs an equal number of groups.   
\d+
- Extracted column:** The name of the column where to extract the matched regex. Multiple column names can be specified separated by commas, if the name is empty it means that group is skipped. If the source column is null, the new column will be null as well, otherwise an empty string means there was no match.   
NetSales

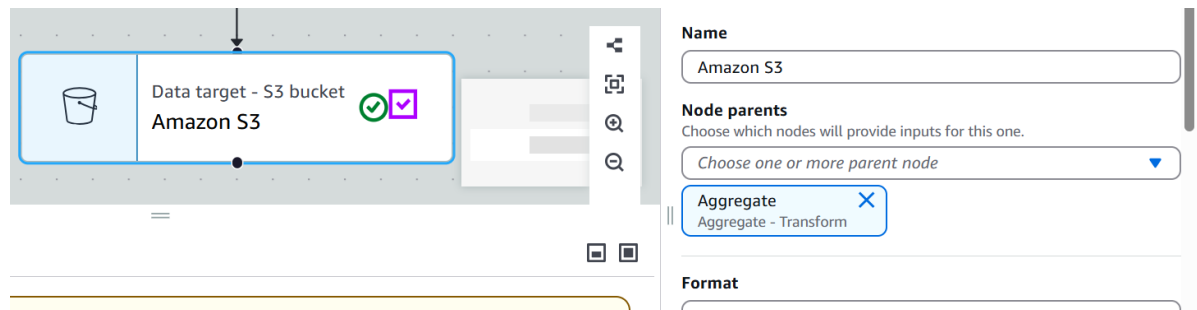
7. Importing & setting up aggregate from transform choosing node as Regex extractor

- Fields to group by as product category and ship mode
- In aggregation
- Field to aggregate as sales and aggregation function as avg

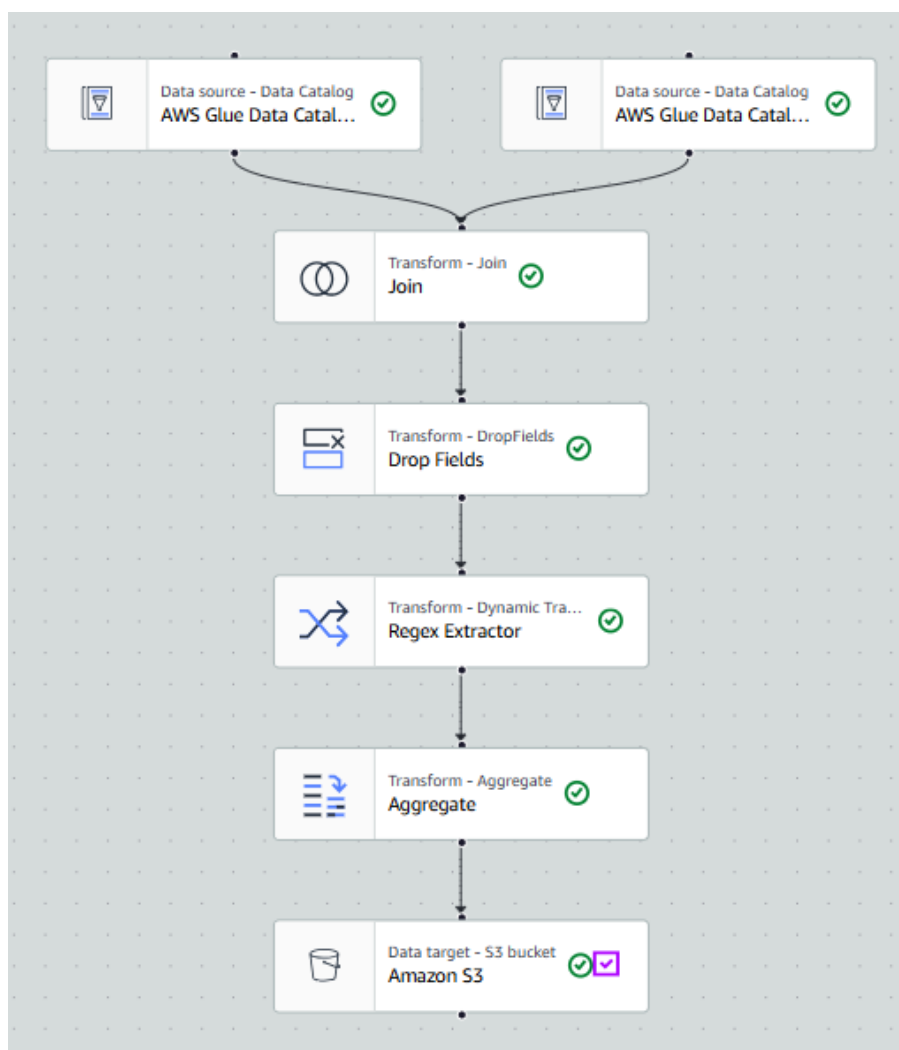
The screenshot displays the Alteryx interface with a workflow canvas on the left and a configuration panel on the right. The workflow canvas shows a 'Transform - Aggregate' node. The configuration panel on the right is titled 'Transform' and contains the following settings:

- Name:** Aggregate
- Node parents:** Choose which nodes will provide inputs for this one.   
Choose one or more parent node   
Regex Extractor   
DynamicTransform - Transform
- Aggregate Info:** (Expandable section)
- Fields to group by - optional:** Select the fields you would like to group your rows by, so the aggregation would be done for each unique group.   
Choose one or more fields   
product category   
ship mode
- Aggregation:** Select fields and functions to aggregate.   
**Field to aggregate:** sales   
**Aggregation function:** avg

8. Importing & setting up amazon s3 node parents as aggregate  
Location as s3:// etl-cep-output-00001-ashish-chamel



9. Final outlook of model of a etl job



10. Etl-cep-job-ashish (ETL) job ran successfully

etl-cep-job-ashish

Visual

Script

Job details

Runs

Data quality

Schedules

Version Control

Job runs (1/1) info

Last updated (UTC)  
April 2, 2025 at 13:43:27

View details

Stop job run

Filter job runs by property

Run status	Retries	Start time (Local)	End time (Local)	Duration	Capacity (DPUs)	Worker typ
Succeeded	0	04/02/2025 19:10:56	04/02/2025 19:12:28	1 m 25 s	10 DPUs	G.1X

Step-8

1. This is the output after successful completion of ETL job in our output bucket ie etl-cep-output-00001-ashish-chamel

etl-cep-output-00001-ashish-chamel info

Objects

Metadata

Properties

Permissions

Metrics

Management

Access Points

Objects (16)

Copy S3 URI

Copy URL

Download

Open

Delete

Actions

Create folder

Find objects by prefix

Name	Type	Last modified	Size	Storage class
run-1743601328427-part-block-0-r-00000-snappy.parquet	parquet	April 2, 2025, 19:12:17 (UTC+05:30)	701.0 B	Standard
run-1743601328427-part-block-0-r-00001-snappy.parquet	parquet	April 2, 2025, 19:12:18 (UTC+05:30)	692.0 B	Standard
run-1743601328427-part-block-0-r-00003-snappy.parquet	parquet	April 2, 2025, 19:12:17 (UTC+05:30)	719.0 B	Standard
run-1743601328427-part-block-0-r-00004-snappy.parquet	parquet	April 2, 2025, 19:12:16 (UTC+05:30)	692.0 B	Standard
run-1743601328427-part-block-0-r-00006-snappy.parquet	parquet	April 2, 2025, 19:12:18 (UTC+05:30)	746.0 B	Standard
run-1743601328427-part-block-0-r-00008-snappy.parquet	parquet	April 2, 2025, 19:12:18 (UTC+05:30)	818.0 B	Standard
run-1743601328427-part-block-0-r-00009-snappy.parquet	parquet	April 2, 2025, 19:12:17 (UTC+05:30)	791.0 B	Standard
run-1743601328427-part-block-0-r-00011-snappy.parquet	parquet	April 2, 2025, 19:12:17 (UTC+05:30)	773.0 B	Standard
run-1743601328427-part-block-0-r-00012-snappy.parquet	parquet	April 2, 2025, 19:12:18 (UTC+05:30)	800.0 B	Standard

2. Running the sql query to see the results

SQL query

Add SQL from templates

Run SQL query

Amazon S3 Select supports only the SELECT SQL command. Using the S3 console, you can extract up to 40 MB of records from an object that is up to 128 MB in size. To work with larger files or more records, use the AWS CLI, AWS SDK, or Amazon S3 REST API. For more complex SQL queries, use Amazon Athena

1 /\* To create reference point for writing SQL queries, you can display the first 5 records of input data by running the following SQL query: SELECT \* FROM s3object s LIMIT 5 \*/

2 SELECT \* FROM s3object s LIMIT 5

SQL

Ln 2, Col 1

Errors: 0

Warnings: 0

Query results

Download results

Query results are not available after you choose Close or navigate away. Choose Download results to download a copy of the following query results.

Status

Successfully returned 1 record in 736 ms

Bytes returned: 39 B

Raw

Formatted

Fashion

Second Class

1274.702380952381

Product Category	Ship Mode	Average Sales
Fashion	Second Class	1274.702380952381