1. Login to the AWS Console:

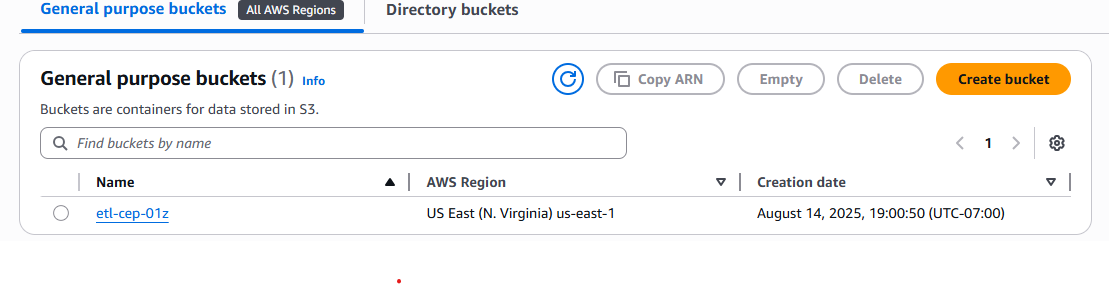
• Open your web browser and navigate to the AWS Management Console

• Log in with your AWS account credentials

• Navigate to S3

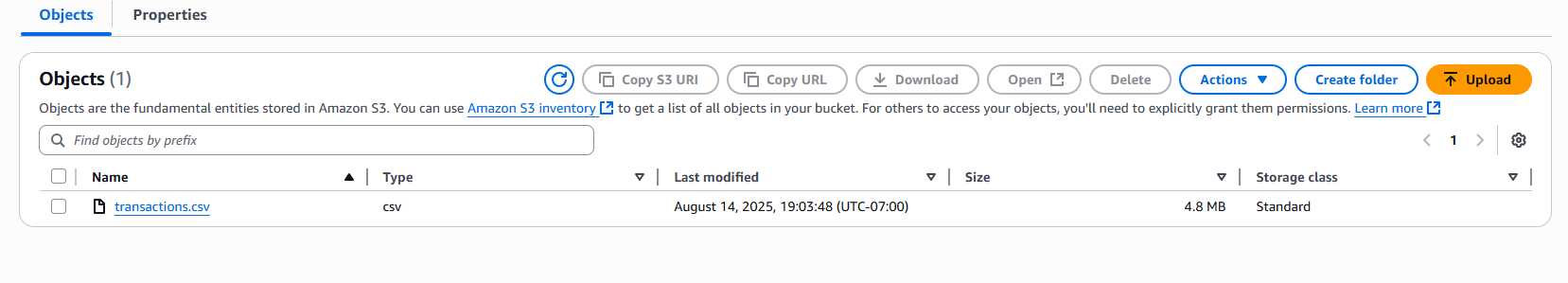
• Click on Create bucket and add the bucket name as etl-cep-01.

Scroll down the screen and click on the Create bucket button.

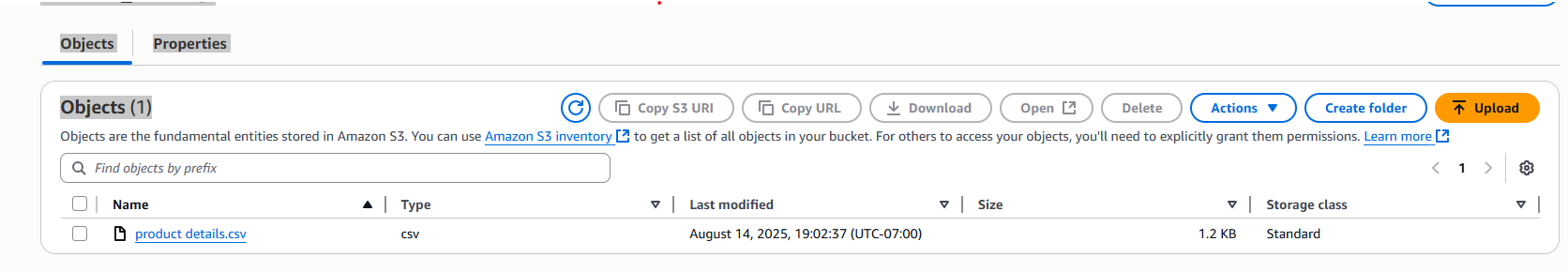


1. Create two subfolders inside the etl-cep-01 bucket

• Click on Create folder, add the folder name as transaction-files for the first folder and product-files for the second folder. Scroll down the screen and click on the Create folder button.



• Inside product-files folder, click on Upload then click on Add files. Select the product details file from your local system and click on Open. Scroll down and click on Upload. Repeat the same step for transaction files and upload the transactions dataset.



1. Create a new bucket

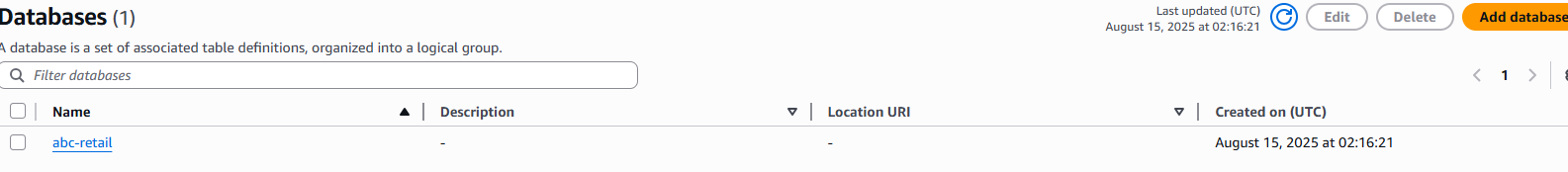
• Create a new bucket named etl-cep-output-01 just as you did above

1. Navigate to AWS Glue and create a new database

• In the AWS management console, search for AWS Glue and select it

• In the AWS Glue console, navigate to the Databases section

• Click on Add database and provide the name as abc-retail. Scroll down and click on Create database



5. Set up two classifiers to read transaction data and product data

• Navigate to the Data Catalog, click on Classifiers then click on Add classifier

• Fill the first classifier details as given below, then click on Create

o Classifier name as cust classifier

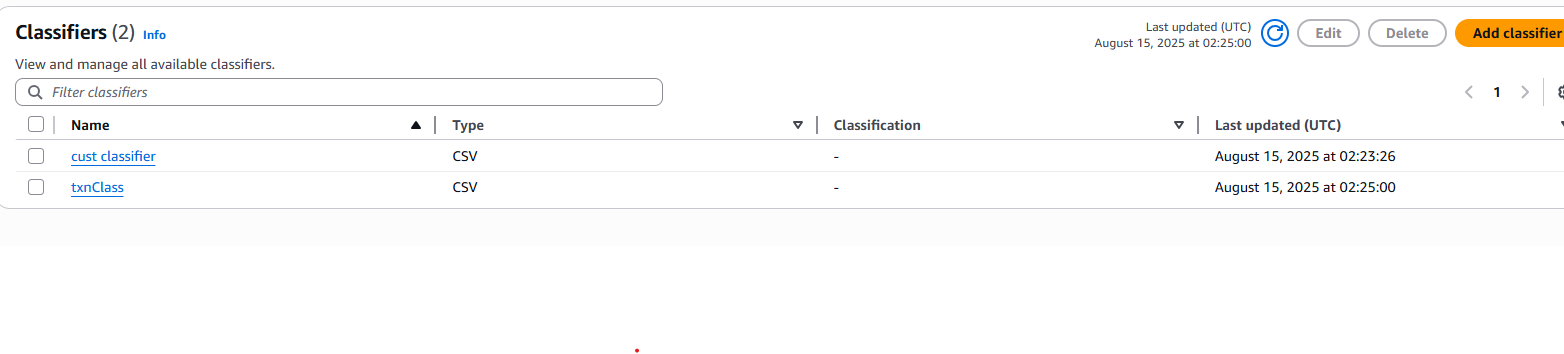
o Classifier type and properties as CSV

o CSV Serde – optional as None

o Column delimiter as comma(,)

o Quote symbol as Double-quote(“)

o Column headings as Has headings



• Fill in the second classifier details as given below, then click on Create

o Classifier name as txnClass

o Classifier type and properties as CSV

o CSV Serde – optional as None o Column delimiter as comma(,)

o Quote symbol as Double-quote(“)

o Column headings as Has headings and fill in the details as given below:

o Order ID, Order Date, Ship Date, Aging, Ship Mode, Product ID, Sales, Quantity, Discount, Profit, Shipping Cost, Order Priority, Customer ID

1. Create IAM role

• In the AWS management console, search for IAM service and select it

• Navigate to Roles and click on Create role

• Select AWS service as the Trusted entity type and Glue as the Use case then click on Next

• Select the Administrator Access policy. Scroll down and click on Next.

• Enter the name as glue-role. Scroll down and click on Create role.

7. Set up a Crawler

• Navigate to AWS Glue and click on Databases from the Data Catalog and select abc-retail database

• Click on Add tables using a crawler

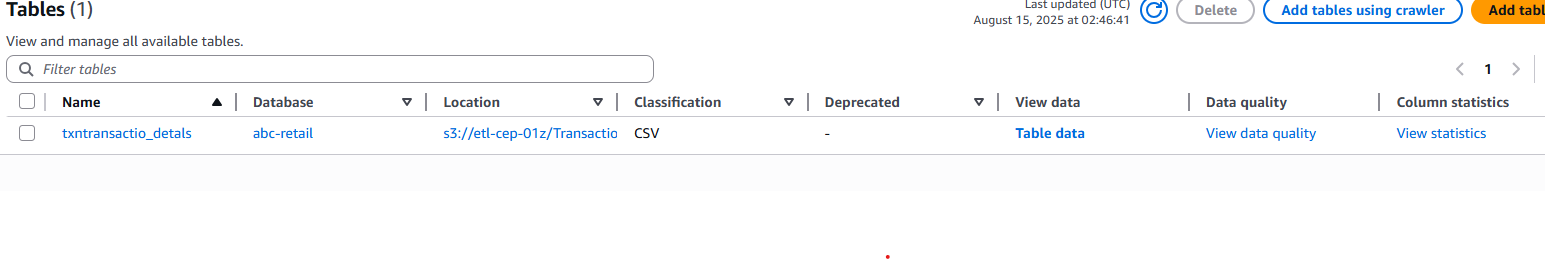
• Enter the name as retail-crawl and click on Next

• Click on Add a data source

• Click on Browse S3 and click on etl-cep-01 then select transaction files/ and click on Choose

• Click on Add an S3 data source

• Choose classifier as txnClass from the drop down of custom classifiers – optional and click on Next

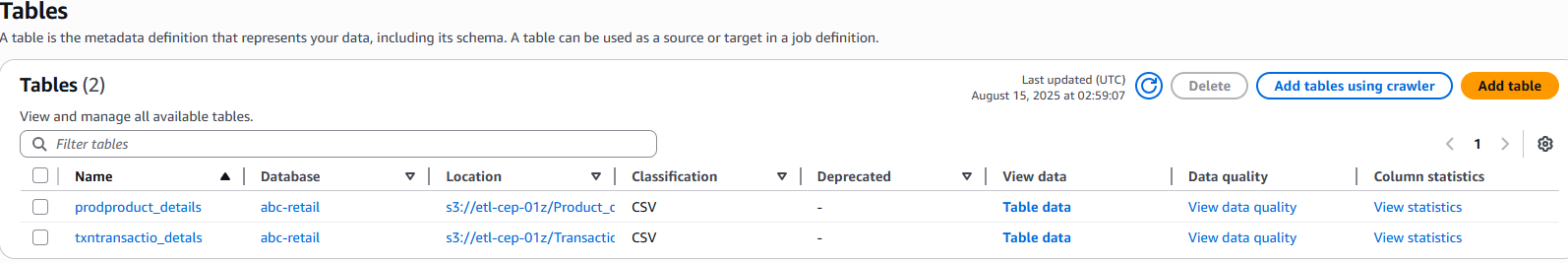


• Choose glue-role in the IAM role section and click on Next

• Choose Target database as abc-retail and enter the table name prefix as txn and click on Next

• Click on Create crawler

• Click on Run crawler Note: Repeat above steps for other Product dataset as well. While choosing classifier choose cust classifier.



8. Create ETL job

• Navigate to AWS Glue, click on ETL jobs, and click on Visual ETL

• In the Add nodes, double-click on AWS Glue Data Catalog • Select Join from the add nodes and link Join to both the AWS Glue Data Catalog

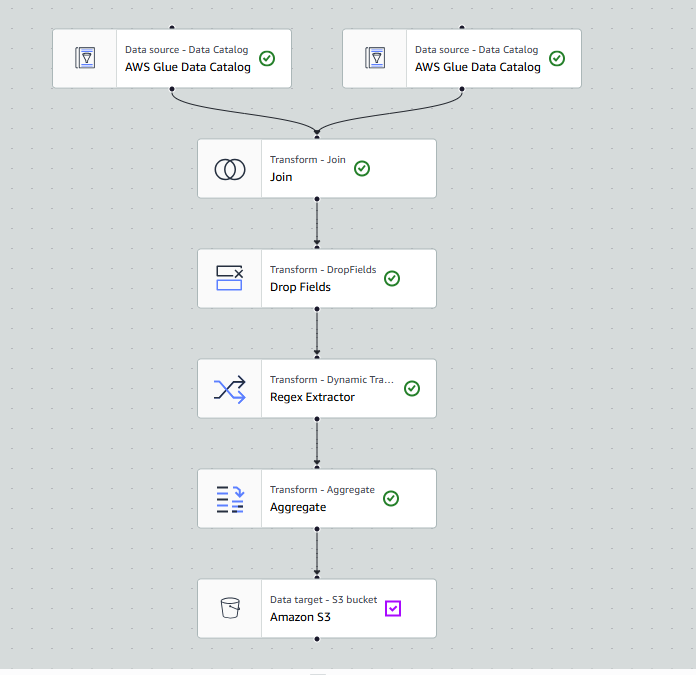
• Click on the Join box, and then select Drop Fields from the Add nodes

• Click on the Drop Fields box, and then select Regex Extractor from the Add nodes • Click on the Regex Extractor box, and then select Aggregate from the Add nodes

• Click on the Aggregate box, and then select Amazon S3 from the Targets in Add nodes • Click on the first AWS Glue Data Catalog box, and select abc-retail in the Database dropdown and select txntransaction files under the Table dropdown

• Click on the second AWS Glue Data Catalog box and select abc-retail in the Database dropdown and select product files under Table dropdown

• Select the Join box and add both the AWS Glue Data Catalog in the node parents. Select Inner join in the Join type and in the Join conditions box select product id in both the AWS Glue Data Catalog boxes



• Click on the Drop Fields box and in the Drop Fields section select product id as it appears twice

• We need to extract sales values as it has $ symbol in it. Click on Regex Extractor and fill the following fields:

o Column to extract from as sales

o Regular expression as \d+

o Extracted column as Net Sales

Note: Before selecting the newly generated column in the aggregation section, you must wait for the Regex Extractor data preview to load fully.

• Now, let’s create our summary report using Aggregate block. Fill the following fields:

o Fields to group by as product category and ship mode

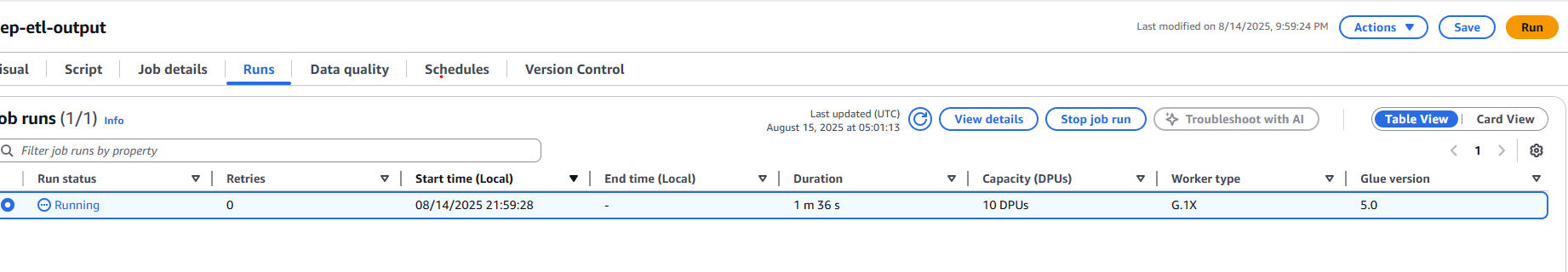
o Field to aggregate as sales

o Aggregation function as avg

• Click on Amazon S3 block then click on Browse S3 and select etl-cep output-01

• On the top left corner click on Untitled job and give the etl-cep-job, click on Save and the click on Run

• Check the progress in Runs



9. Check the output and run the sql query

• Navigate to S3 bucket to check output in cep-etl-output-01 bucket

• Click on run-1715077987267-part-block-0-r-00000-snappy. Parquet

• Navigate to Object actions and select Query with S3 Select

• In the Output settings select CSV format and click on Run SQL query

• The output appears as shown below

