

COURSE END PROJECT(ETL)

GAGAN SHRIVASTAVA

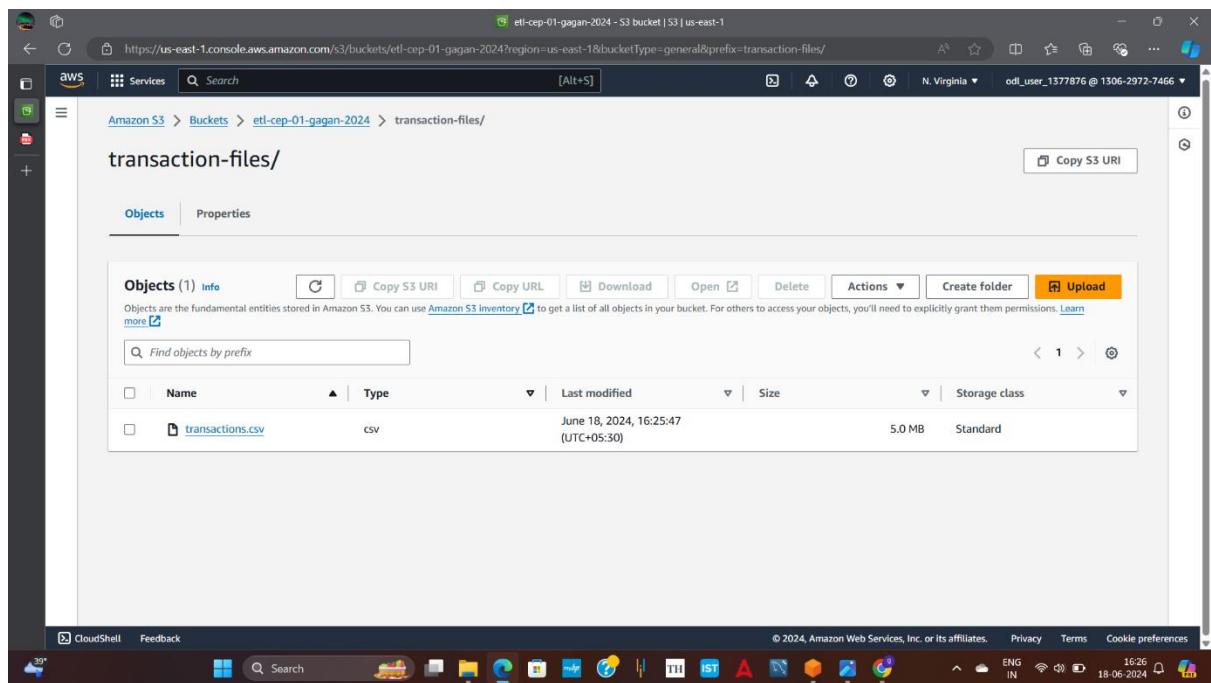
>bucket created as -etl-cep-gagan-2024 in S3 bucket.

The screenshot shows the AWS S3 console with a green success message at the top: "Successfully created bucket 'etl-cep-01-gagan-2024'". Below it, an "Account snapshot" section displays storage usage and activity trends. The main area lists "General purpose buckets" with one entry: "etl-cep-01-gagan-2024" (US East (N. Virginia) us-east-1). The browser status bar indicates the URL is https://us-east-1.console.aws.amazon.com/s3/buckets?region=us-east-1&bucketType=general.

>created transaction files and product files inside bucket.

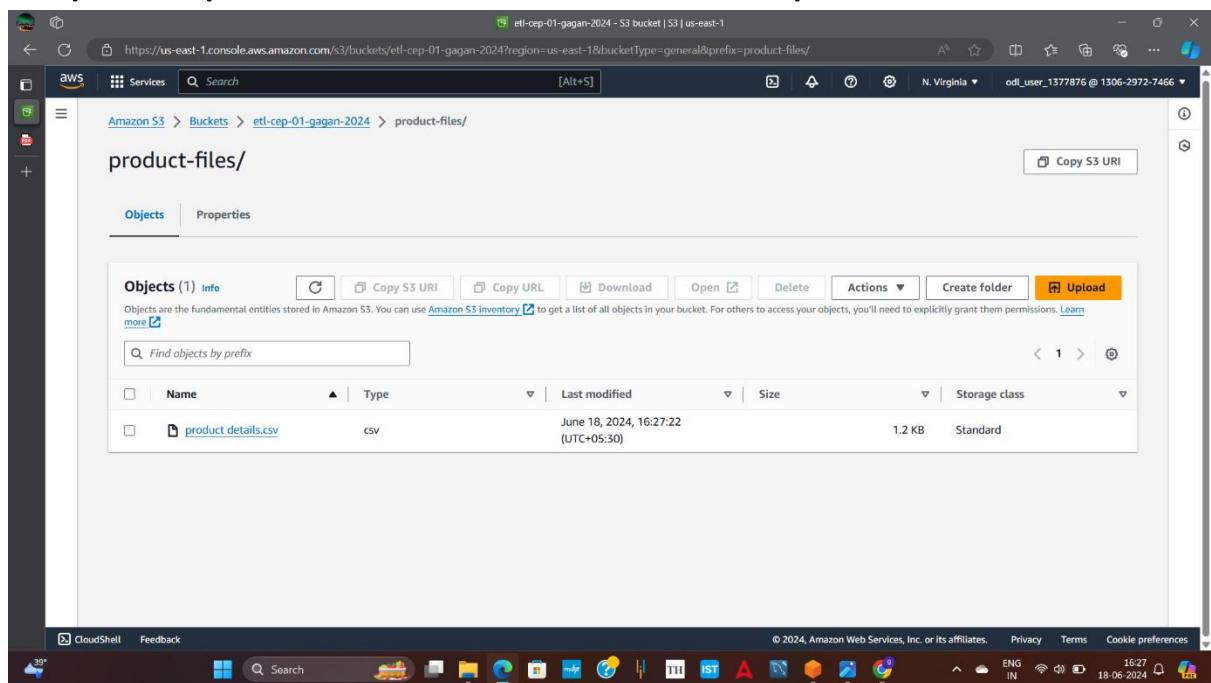
The screenshot shows the AWS S3 console for the bucket "etl-cep-01-gagan-2024". The "Objects" tab is selected, showing two items: "product-files/" and "transaction-files/". Both are listed as Folders. The browser status bar indicates the URL is https://us-east-1.console.aws.amazon.com/s3/buckets/etl-cep-01-gagan-2024?region=us-east-1&bucketType=general&tab=objects.

>upload transaction csv data file inside transaction file.



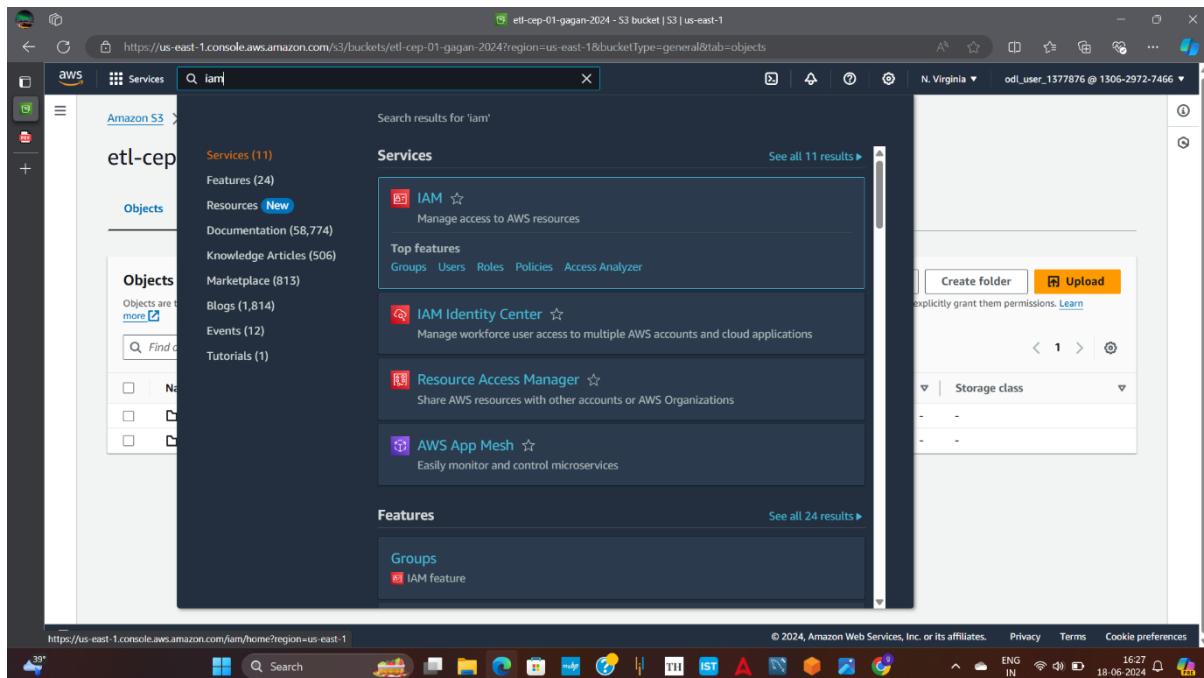
The screenshot shows the AWS S3 console interface. The URL in the address bar is <https://us-east-1.console.aws.amazon.com/s3/buckets/etl-cep-01-gagan-2024?region=us-east-1&bucketType=general&prefix=transaction-files/>. The page displays the contents of the 'transaction-files/' folder. There is one object listed: 'transactions.csv' (Type: csv, Last modified: June 18, 2024, 16:25:47 (UTC+05:30), Size: 5.0 MB, Storage class: Standard). The 'Actions' menu is visible above the table, and the 'Upload' button is highlighted.

>Upload product csv data file inside product file.



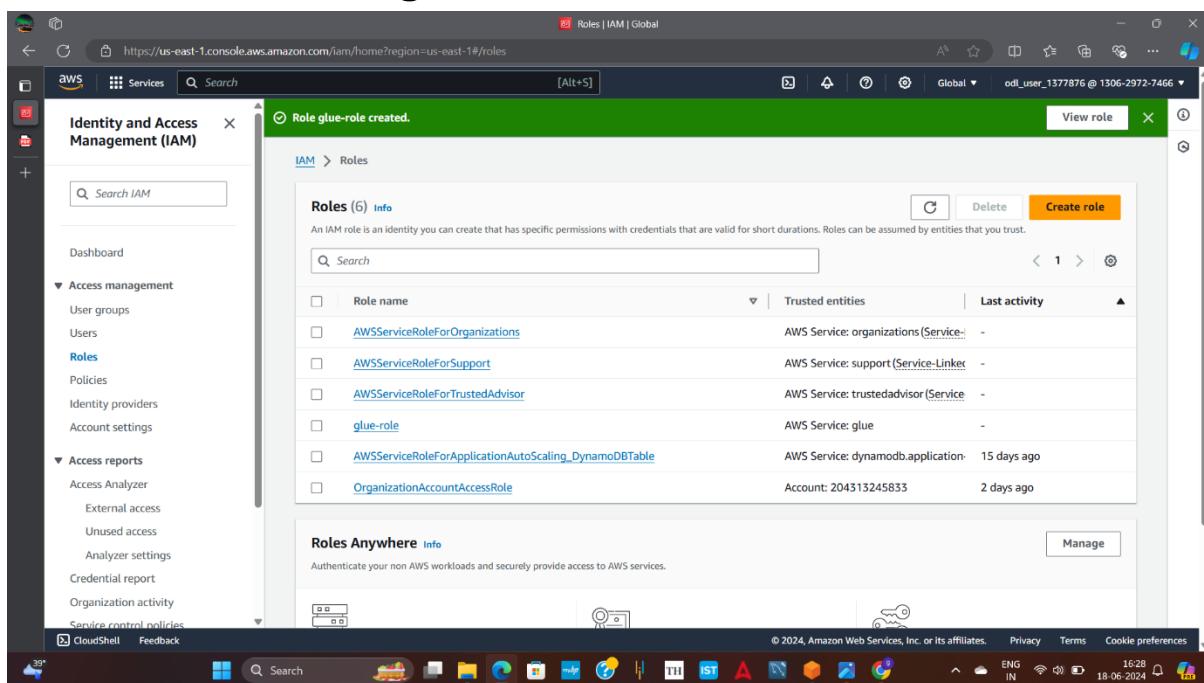
The screenshot shows the AWS S3 console interface. The URL in the address bar is <https://us-east-1.console.aws.amazon.com/s3/buckets/etl-cep-01-gagan-2024?region=us-east-1&bucketType=general&prefix=product-files/>. The page displays the contents of the 'product-files/' folder. There is one object listed: 'product details.csv' (Type: csv, Last modified: June 18, 2024, 16:27:22 (UTC+05:30), Size: 1.2 KB, Storage class: Standard). The 'Actions' menu is visible above the table, and the 'Upload' button is highlighted.

>Search iam in search bar and then click it for creating a role.



The screenshot shows the AWS Lambda console with a search bar at the top containing 'iam'. Below the search bar, there are two main sections: 'Services' and 'Features'. The 'Services' section is expanded, showing the 'IAM' service listed first. The 'IAM' service card includes the description 'Manage access to AWS resources' and 'Top features: Groups, Users, Roles, Policies, Access Analyzer'. Other services listed include 'IAM Identity Center', 'Resource Access Manager', and 'AWS App Mesh'. The 'Features' section below shows 'Groups' and 'IAM feature'. The left sidebar shows the Lambda function 'etl-cep' and its objects. The bottom of the screen shows the AWS navigation bar with links for 'Roles | IAM | Global', 'View role', and other account information.

>create a role as- glue-role to get permission and access to data for glue.



The screenshot shows the AWS IAM console with a success message 'Role glue-role created.' at the top. The left sidebar shows the 'Identity and Access Management (IAM)' section with options like 'Dashboard', 'Access management', 'Access reports', and 'CloudShell'. The 'Roles' section is selected. The main area shows a table titled 'Roles (6) Info' with columns for 'Role name', 'Trusted entities', and 'Last activity'. The table lists six roles: 'AWSServiceRoleForOrganizations', 'AWSServiceRoleForSupport', 'AWSServiceRoleForTrustedAdvisor', 'glue-role', 'AWSServiceRoleForApplicationAutoScaling_DynamoDBTable', and 'OrganizationAccountAccessRole'. The 'glue-role' row shows 'AWS Service: glue' under 'Trusted entities' and '15 days ago' under 'Last activity'. Below the table, there is a section titled 'Roles Anywhere' with a 'Manage' button.

>search aws glue in search bar and then click it.

The screenshot shows the AWS Glue search results page. The search term 'glue' is entered in the search bar at the top. The results are categorized into 'Services' and 'Features'. Under 'Services', there are five items: AWS Glue, AWS Glue DataBrew, AWS Lake Formation, and Athena. Under 'Features', there is one item: AWS Glue Studio. To the right of the search results, there is a sidebar titled 'Transform data' which includes sections for 'Author and edit ETL jobs' and 'Usage'.

>create database as- abc-retail in databases under data catalog.

The screenshot shows the AWS Glue Databases page. In the top right corner, there is a large orange button labeled 'Add database'. Below this button, a table lists existing databases. There is one database entry: 'abc-retail' with a 'Name' column value of 'abc-retail', a 'Created on (UTC)' column value of 'June 18, 2024 at 11:00:01', and a 'Last updated (UTC)' column value of 'June 18, 2024 at 11:50:02'. The left sidebar contains navigation links for AWS Glue, Data Catalog (with sub-links for Databases, Tables, Stream schema registries, Schemas, Connections, Crawlers, Classifiers, Catalog settings), Data Integration and ETL, Legacy pages, and various documentation links.

>Create two separate classifiers as cust-classifiers for product files and txn-classifier for transaction files.

The screenshot shows the AWS Glue Console interface. The left sidebar is titled "AWS Glue" and includes sections for Getting started, ETL jobs, Data Catalog tables, Data connections, Workflows (orchestration), Data Catalog (with sub-options like Databases, Tables, Stream schema registries, Schemas, Connections, Crawlers, and Classifiers), Data Integration and ETL, Legacy pages, What's New, Documentation, CloudShell, and Feedback. The "Classifiers" section under Data Catalog is currently selected. The main content area is titled "Classifiers" and contains a brief description: "Classifiers are triggered during a crawl task. A classifier checks whether a given file is in a format the crawler can handle. If it is, the classifier creates a schema in the form of a StructType object that matches that data format." Below this is a table titled "Classifiers (2) Info". The table has columns for Name, Type, Classification, and Last updated (UTC). It lists two entries: "cust-classifier" (CSV, last updated June 18, 2024 at 11:01:43) and "txn-classifier" (CSV, last updated June 18, 2024 at 11:02:53). The table includes a search bar labeled "Filter classifiers" and navigation buttons for "1" and "2". The top of the browser window shows the URL "https://us-east-1.console.aws.amazon.com/glue/home?region=us-east-1#/v2/data-catalog/classifiers/" and the AWS logo. The top right corner shows the region "N. Virginia", the user "odl_user_1377876 @ 1506-2972-7466", and the date "June 18, 2024". The bottom of the screen shows the Windows taskbar with various pinned icons and the system tray indicating the date and time as "18-06-2024 16:32".

Name	Type	Classification	Last updated (UTC)
cust-classifier	CSV	-	June 18, 2024 at 11:01:43
txn-classifier	CSV	-	June 18, 2024 at 11:02:53

>cust-classifier properties.

The screenshot shows the AWS Glue console interface. On the left, there's a navigation sidebar with options like 'Getting started', 'ETL jobs', 'Visual ETL', 'Notebooks', 'Job run monitoring', 'Data Catalog tables', 'Data connections', 'Workflows (orchestration)', 'Data Catalog' (expanded), 'Databases', 'Tables', 'Stream schema registries', 'Schemas', 'Connections', 'Crawlers', 'Classifiers' (selected), 'Catalog settings', 'Data Integration and ETL', and 'Legacy pages'. The main content area is titled 'AWS Glue > Classifiers > cust-classifier' and shows the 'cust-classifier' properties in a table:

Classifier properties		
Name	Allow single column	CSV Serde
cust-classifier	False	None
Contains header	Header	Creation time
Has headings	product id,product.product category	June 18, 2024 at 11:01:43
Delimiter	Disable value trimming	Quote symbol
,	True	"
Last updated	Version	Custom datatypes
June 18, 2024 at 11:01:43	1	-

At the top right, there are 'Edit' and 'Delete' buttons. Above the table, it says 'Last updated (UTC) June 18, 2024 at 11:01:44'. The bottom of the screen shows a Windows taskbar with various icons and system status.

>txn-classifier properties.

This screenshot is similar to the previous one but for the 'txn-classifier'. The navigation sidebar is identical. The main content area is titled 'AWS Glue > Classifiers > txn-classifier' and shows the 'txn-classifier' properties in a table:

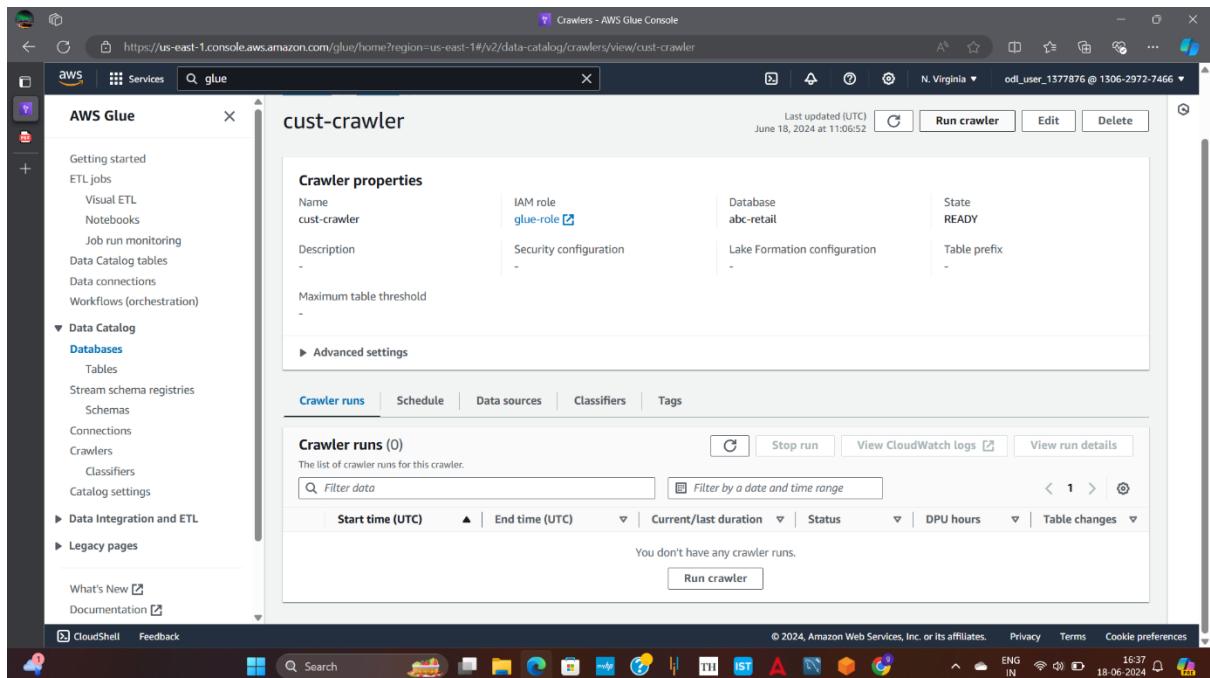
Classifier properties		
Name	Allow single column	CSV Serde
txn-classifier	False	None
Contains header	Header	Creation time
Has headings	Order ID,Order Date,Ship Date,Aging,Ship Mode,Product ID,Sales,Quantity,Discount,Profit,Shipping Cost,Order Priority,Customer ID	June 18, 2024 at 11:02:53
Delimiter	Disable value trimming	Quote symbol
,	True	"
Last updated	Version	Custom datatypes
June 18, 2024 at 11:02:53	1	-

At the top right, there are 'Edit' and 'Delete' buttons. Above the table, it says 'Last updated (UTC) June 18, 2024 at 11:03:07'. The bottom of the screen shows a Windows taskbar with various icons and system status.

>Inside abc-retail database create crawler as retail-crawler for transaction files then run crawler.

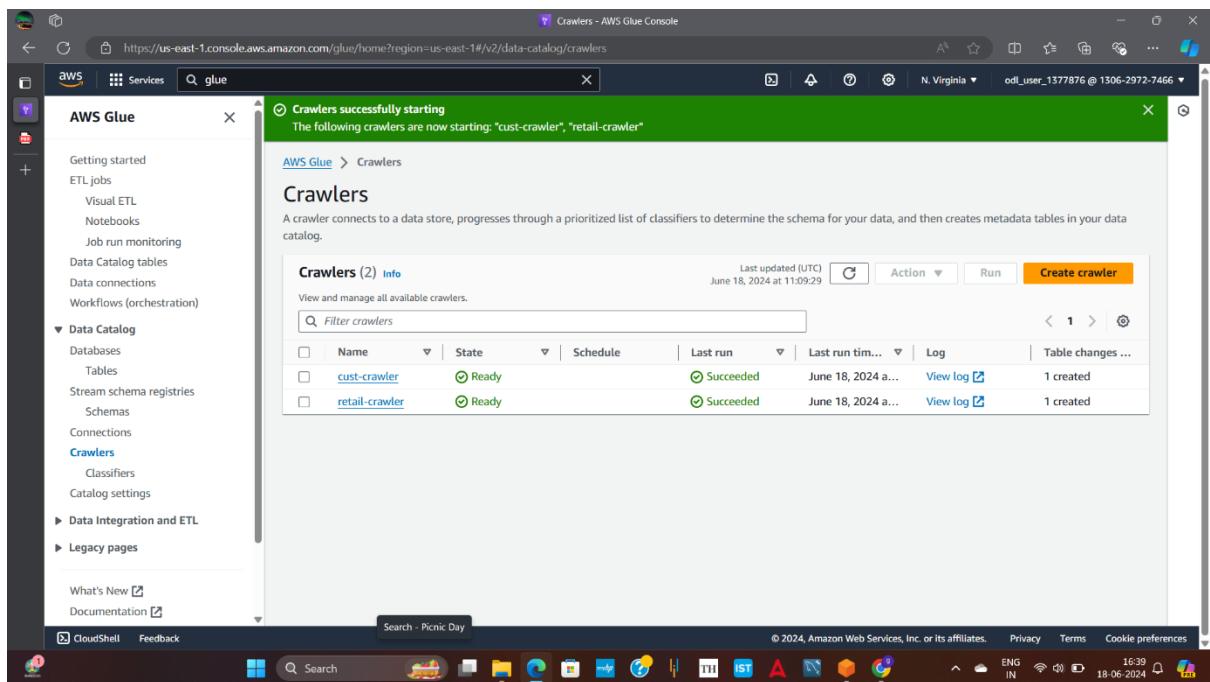
The screenshot shows the AWS Glue console with the URL <https://us-east-1.console.aws.amazon.com/glue/home?region=us-east-1#/v2/data-catalog/crawlers/view/retail-crawler>. A green banner at the top indicates "Crawler successfully starting" and "The following crawler is now starting: 'retail-crawler'". The main area displays the "retail-crawler" properties, including its name, IAM role (glue-role), database (abc-retail), and state (READY). The "Crawler runs" section shows one run that has just started, with a duration of 11s and a status of "Running". The browser's taskbar at the bottom shows various open tabs and icons.

>create another crawler inside abc-retail database as cust-crawler for product files then run it as well.



The screenshot shows the AWS Glue console interface. On the left, there's a navigation sidebar with options like 'Getting started', 'ETL jobs', 'Data Catalog', 'Data Integration and ETL', and 'Legacy pages'. Under 'Data Catalog', 'Databases' is selected. In the main area, a crawler named 'cust-crawler' is being configured. The 'Crawler properties' section includes fields for Name ('cust-crawler'), IAM role ('glue-role'), Database ('abc-retail'), and State ('READY'). Below this, there are sections for Description, Security configuration, Lake Formation configuration, and Table prefix. A 'Run crawler' button is visible at the top right. The 'Crawler runs' section shows 0 runs, with a note: 'You don't have any crawler runs.' A 'Run crawler' button is also present here. The bottom of the screen shows the standard Windows taskbar with various icons.

>After run wait for some time then go to the crawler under data catalog to see both crawler status as ready.

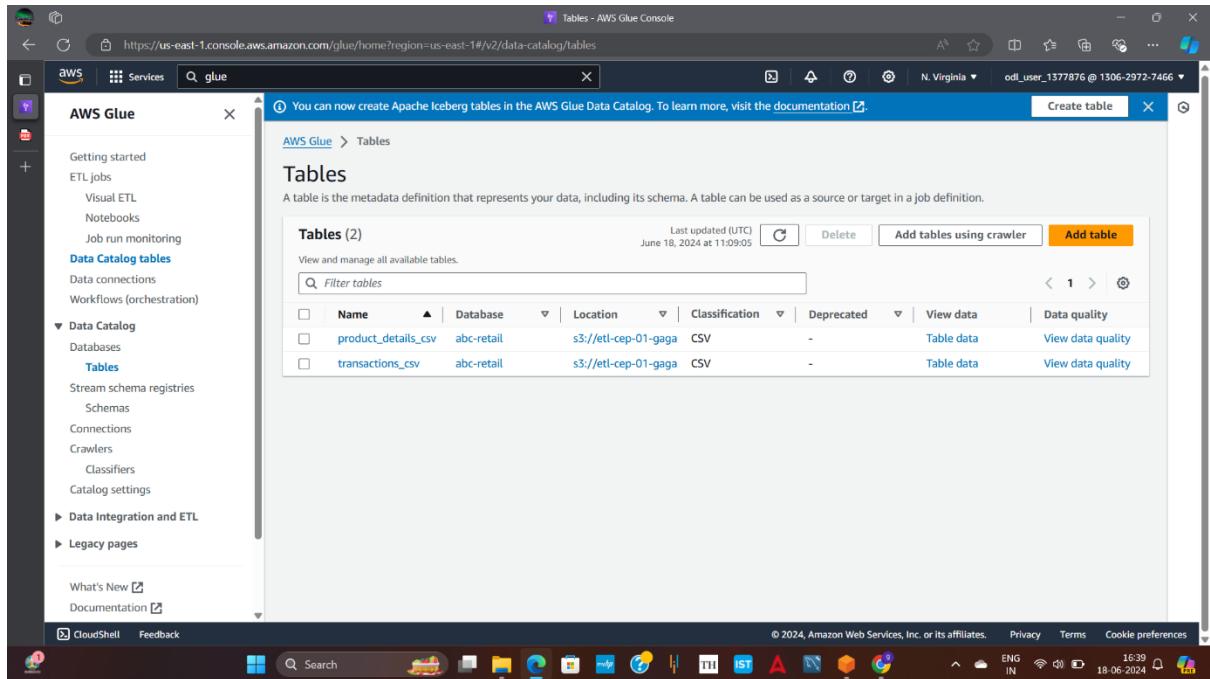


The screenshot shows the 'Crawlers' page in the AWS Glue console. At the top, a green banner says 'Crawlers successfully starting' followed by 'The following crawlers are now starting: "cust-crawler", "retail-crawler"'. Below this, the 'Crawlers' section has a brief description: 'A crawler connects to 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.' There are two crawlers listed in the 'Crawlers (2) info' table:

Name	State	Last run	Last run time...	Log	Table changes ...
cust-crawler	Ready	Succeeded	June 18, 2024 a...	View log	1 created
retail-crawler	Ready	Succeeded	June 18, 2024 a...	View log	1 created

The table includes columns for Name, State, Last run, Last run time..., Log, and Table changes The 'Action' and 'Run' buttons are also visible at the top of the table. The bottom of the screen shows the standard Windows taskbar.

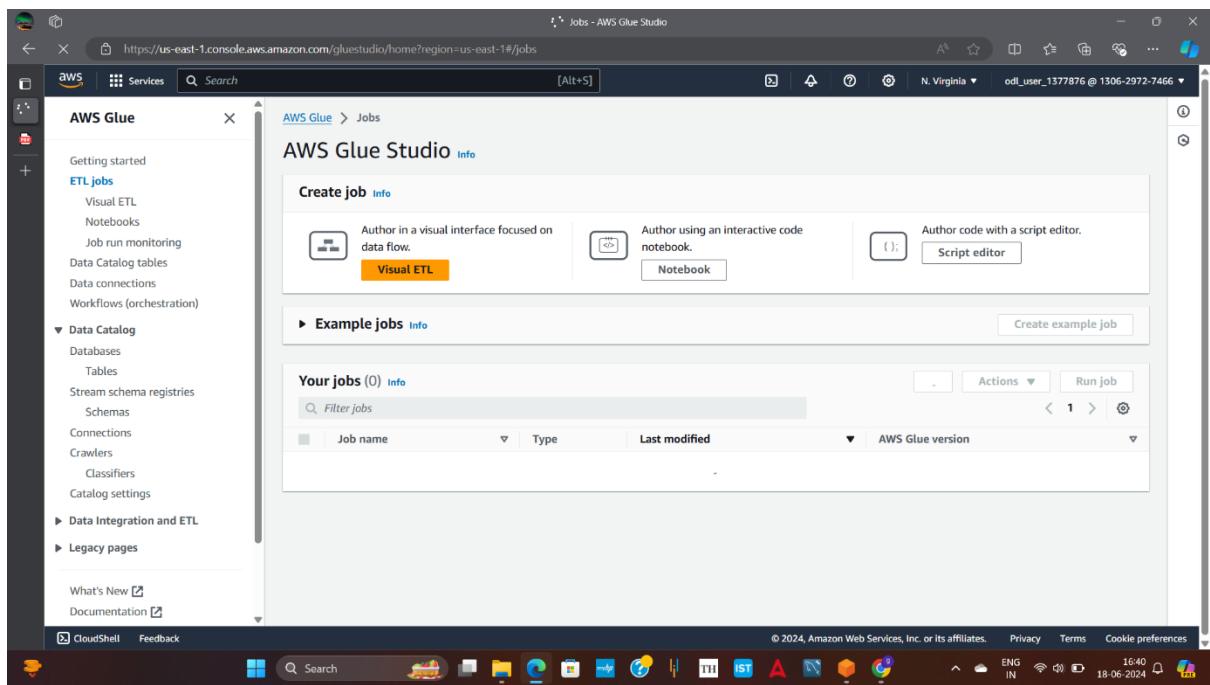
>Go to the table under data catalog to see both table got created.



The screenshot shows the AWS Glue Data Catalog Tables page. The left sidebar has sections for Getting started, ETL jobs, Data Catalog tables, Data Catalog, Databases, and Tables. Under Tables, it lists Stream schema registries, Schemas, Connections, Crawlers, Classifiers, and Catalog settings. The main content area is titled 'Tables' and shows a table with two rows:

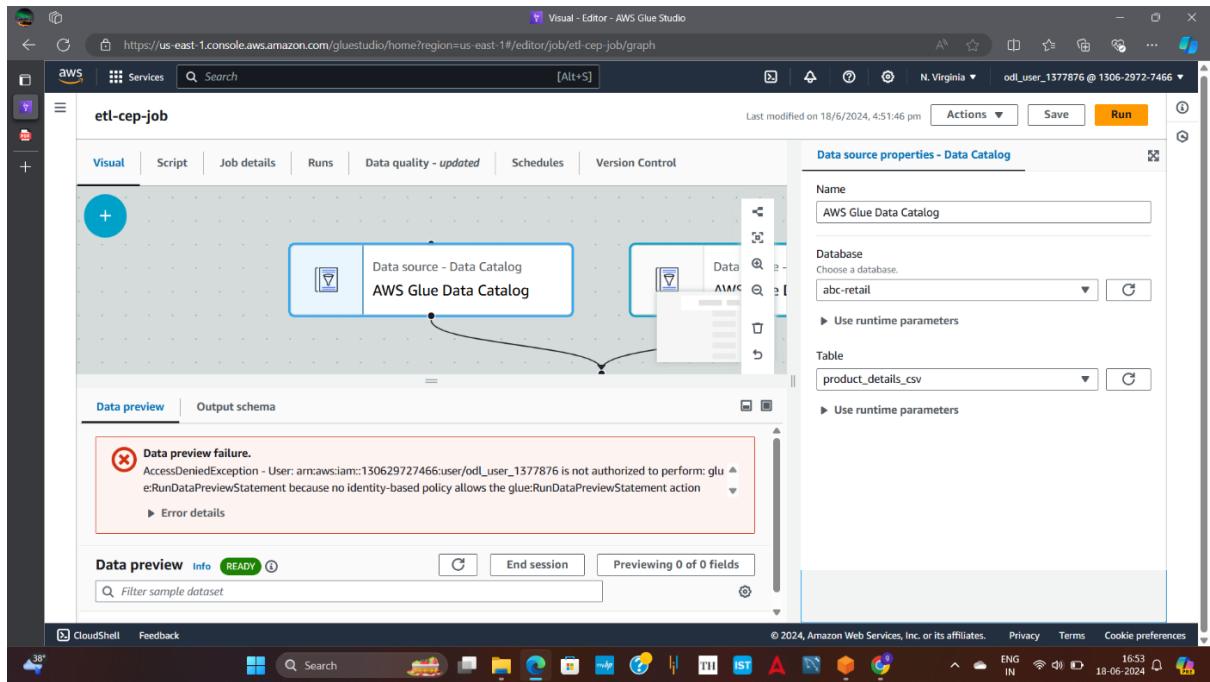
Name	Database	Location	Classification	Deprecated	View data	Data quality
product_details_csv	abc-retail	s3://etl-cep-01-gaga	CSV	-	Table data	View data quality
transactions_csv	abc-retail	s3://etl-cep-01-gaga	CSV	-	Table data	View data quality

>Go to the etl jobs under which click on visual etl.

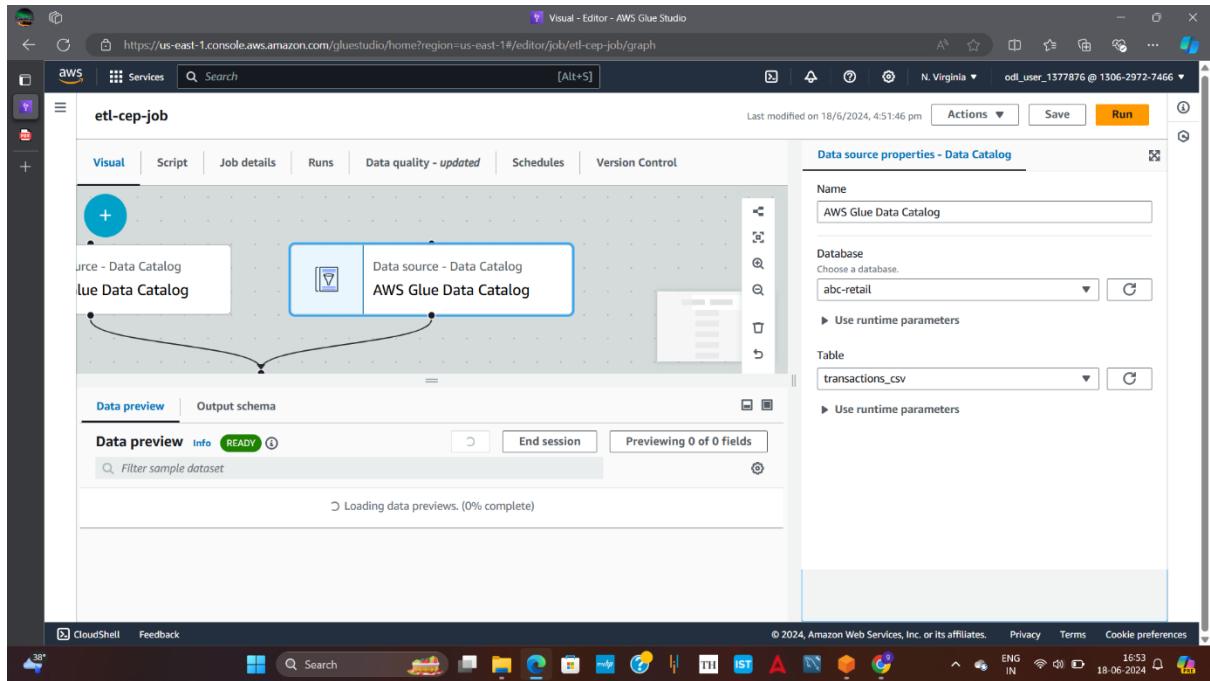


The screenshot shows the AWS Glue Studio Jobs page. The left sidebar has sections for Getting started, ETL jobs, Data Catalog tables, Data Catalog, Databases, Tables, Stream schema registries, Schemas, Connections, Crawlers, Classifiers, Catalog settings, Data Integration and ETL, and Legacy pages. The main content area is titled 'Create job' and shows three options: 'Visual ETL' (selected), 'Notebook', and 'Script editor'. Below this is a section for 'Example jobs' and 'Your jobs (0)'. The 'Your jobs' table has columns for Job name, Type, Last modified, and AWS Glue version.

>create block of aws glue data catalog(extract) for product csv data file.



>Create another block of aws glue data catalog(extract) for transaction csv data file.



>join(transform) both the data catalog(inner join) on the basis of product id.

The screenshot displays two separate sessions of the AWS Glue Studio visual editor. Both sessions show a single 'Transform - Join' node in the center of a workflow graph. The left side of each screen has tabs for Visual, Script, Job details, Runs, Data quality - updated, Schedules, and Version Control. The right side contains a 'Transform' configuration panel.

Top Session (Left):

- Data preview:** Shows a red error box indicating a "Data preview failure" due to an AccessDeniedException.
- Join conditions:** Shows a comparison between "product id" from "AWS Glue Data Catalog" and "product id" from "AWS Glue Data Catalog".
- Join type:** Set to "Inner join".

Bottom Session (Right):

- Data preview:** Shows a green "READY" status.
- Join conditions:** Shows a comparison between "product id" from "AWS Glue Data Catalog" and "product id" from "AWS Glue Data Catalog - DataSource".
- Join type:** Set to "Inner join".

>To drop(transform) the extra field – product id because of join.

The screenshot shows two side-by-side configurations of an AWS Glue Studio job named "etl-cep-job".

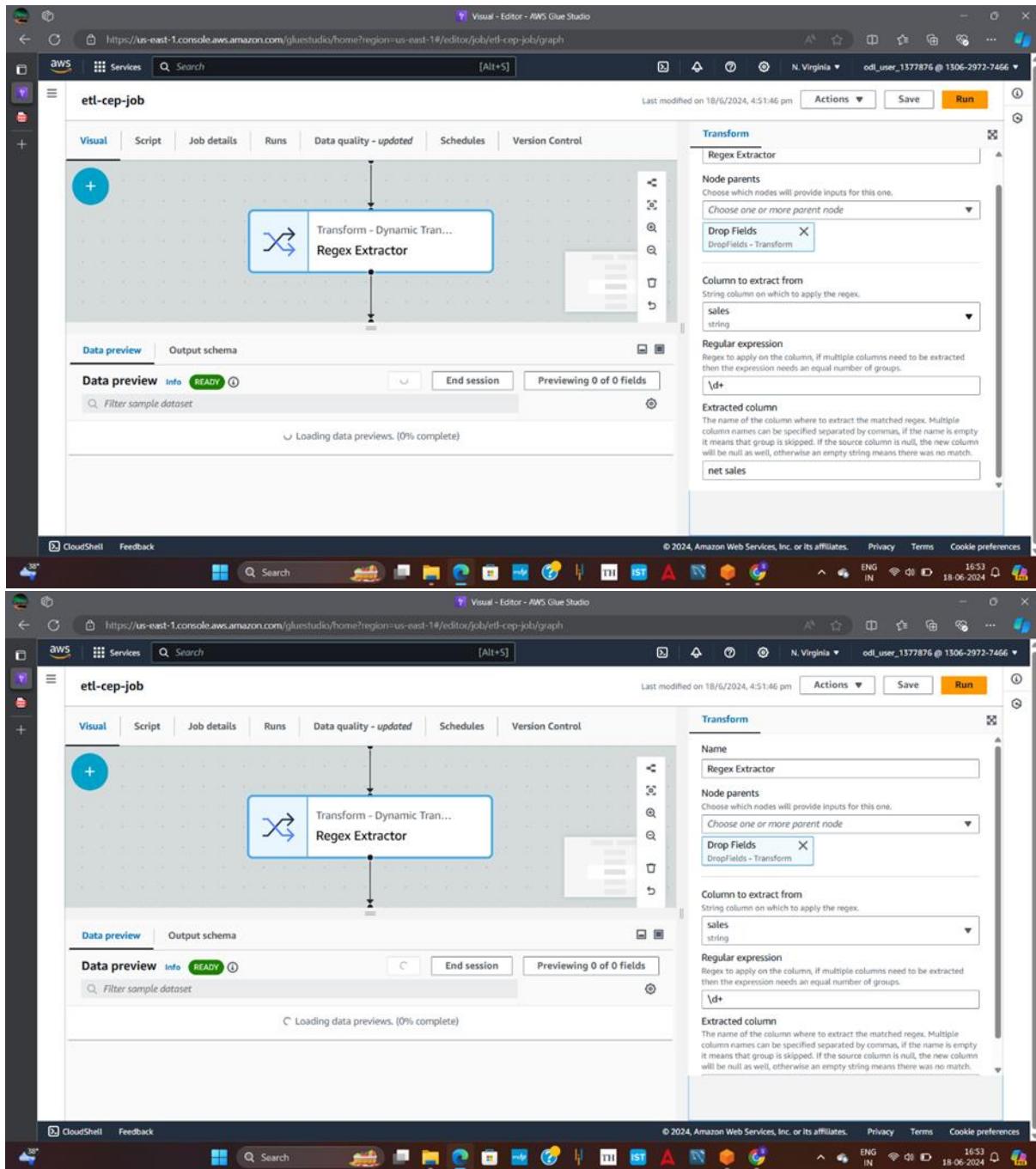
Left Configuration: This configuration includes a "Drop Fields" transform node. In the "Transform" panel, under "DropFields", the "product id" field is selected and checked. The output schema table shows the following fields:

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

Right Configuration: This configuration also includes a "Drop Fields" transform node. In the "Transform" panel, under "DropFields", the "product id" field is deselected (unchecked). The output schema table shows the following fields:

Field	Data type
product id	long
product	string
product category	string
order id	string
order date	string
ship date	string
aging	long
ship mode	string
sales	string
quantity	long
discount	double
profit	string

>Add regex extractor(transform) to cleanse the data.



>Add aggregator(transform) to apply group by and aggregate fuctions on data as follows.

The screenshot displays two side-by-side configurations of an AWS Glue Studio job named "etl-cep-job".

Top Configuration:

- Fields to group by - optional:** Choose one or more fields: product category, ship mode.
- Field to aggregate:** sales
- Aggregation function:** avg
- Aggregate an column:** You can add up to 29 more aggregations.

Bottom Configuration:

- Name:** Aggregate
- Node parents:** Choose one or more parent node: Regex Extractor, DynamicTransform - Transform.
- Aggregate Info:** This transform first groups your rows by fields you choose, and then computes the aggregated value for fields you choose by specific function (e.g., sum, average, max).
- Input:** Input flow (represented by blue arrows).
- Group:** Grouping node (represented by a blue square).
- Aggregate:** Aggregation node (represented by a blue circle).
- Output:** Output flow (represented by blue arrows).

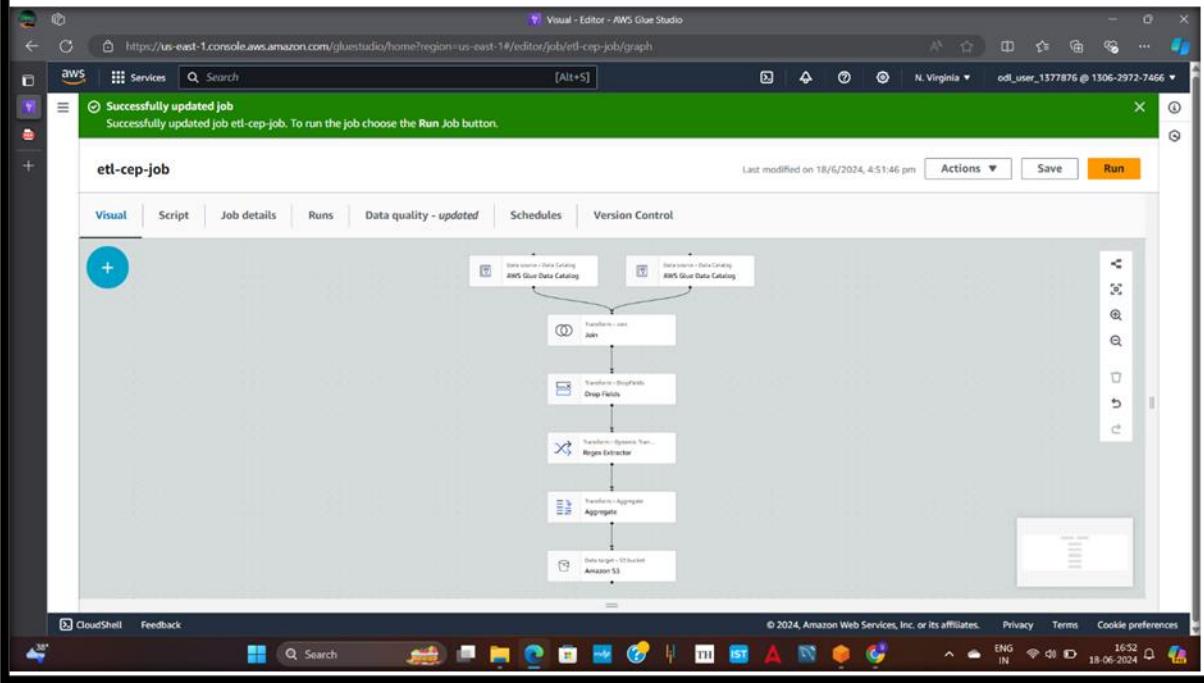
>Target location amazon s3(load) the data we load in the block and choose target location to see the output.

The screenshot displays two separate AWS Glue Studio job configurations, both titled "etl-cep-job".

- Top Configuration:**
 - Data target properties - S3:** Compression type is set to "Snappy".
 - S3 Target Location:** The target location is specified as "s3://etl-cep-01-gagan-2024/trans".
 - Note:** A message box states: "Target node not supported. You have selected a data target node which is not supported for data preview. Please select another type of node instead."
- Bottom Configuration:**
 - Data target properties - S3:** Format is set to "Parquet" and compression type is "Snappy".
 - S3 Target Location:** The target location is specified as "s3://etl-cep-01-gagan-2024/trans".
 - Note:** A message box states: "Target node not supported. You have selected a data target node which is not supported for data preview. Please select another type of node instead."

the output location is not as per the cep as create diff bucket for output location because it is not transferred there so I transferred the output in location of transaction file under etl-cep-01-gagan-2024.

>name etl-job as -etl-cep-job then save and run.



>See the etl-cep-job details as succeeded.

The screenshot shows two identical views of the AWS Glue Studio 'Runs' editor for the 'etl-cep-job'. Both views display a table of job runs and detailed run information for the most recent successful run.

Job runs (1/1) Info

Run status	Retries	Start time (Local)	End time (Local)	Duration	Capacity (DPUs)	Worker type	Glue version
Succeeded	0	06/18/2024 16:54:31	06/18/2024 16:56:03	1 m 20 s	10 DPUs	G.1X	4.0

Run details

Retry attempt number	Execution time	Execution class	Timeout
Initial run	1 minute 20 seconds	Standard	2880 minutes
Trigger name	Security configuration	Cloudwatch logs	Usage profile
-	-	<ul style="list-style-type: none">All logsOutput logsError logs	-

Run details

Job name	Start time (Local)	Glue version	Last modified on (Local)
etl-cep-job	06/18/2024 16:54:31	4.0	06/18/2024 16:56:03
Id	End time (Local)	Worker type	Log group name
jr_d48c5cd2304d552d1960ea786ff01480ffd38cc0c	06/18/2024 16:56:03	G.1X	/aws-glue/jobs
Run status	Start-up time	Max capacity	Number of workers
Succeeded	12 seconds	10 DPUs	10

>Search s3 on search bar then click it to go to the buckets.

The screenshot shows the AWS Glue Studio interface with a search bar at the top containing the query 's3'. Below the search bar, there are two main sections: 'Services' and 'Features'.

Services (8 results):

- S3: Scalable Storage in the Cloud (Top feature)
- S3 Glacier: Archive Storage in the Cloud
- AWS Snow Family: Large Scale Data Transport
- Storage Gateway: Hybrid Storage Integration

Features (39 results):

- Imports from S3: DynamoDB feature

On the right side of the interface, there is a panel titled 'Actions' with a 'Run' button, and another panel showing 'Worker type: G.1X' and 'Glue version: 4.0'.

>Click on transaction file under etl-cep-01-gagan-2024 bucket to see the output of etl-cep-job.

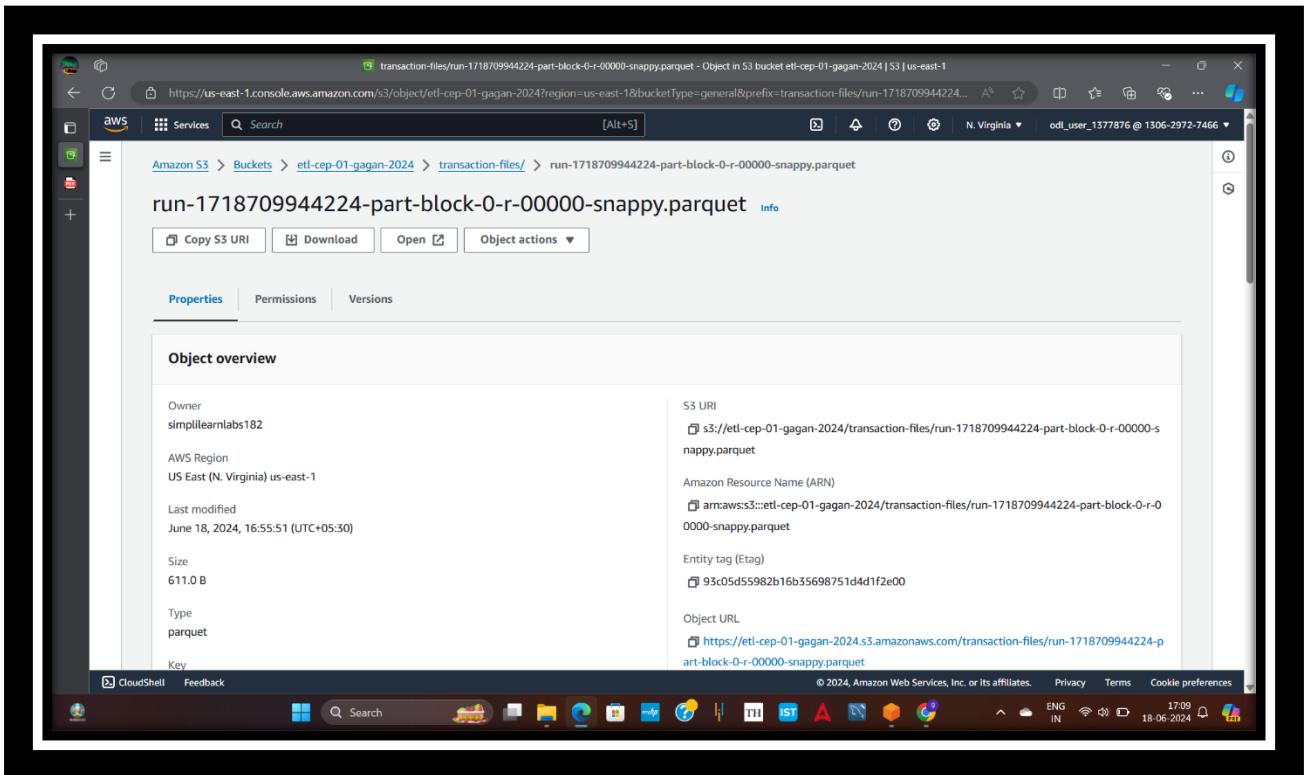
Objects (17) [Info](#)

Name	Type	Last modified	Size	Storage class
run-1718709944224-part-block-0-r-00000-snappy.parquet	parquet	June 18, 2024, 16:55:51 (UTC+0:30)	611.0 B	Standard
run-1718709944224-part-block-0-r-00001-snappy.parquet	parquet	June 18, 2024, 16:55:52 (UTC+0:30)	602.0 B	Standard
run-1718709944224-part-block-0-r-00003-snappy.parquet	parquet	June 18, 2024, 16:55:52 (UTC+0:30)	629.0 B	Standard
run-1718709944224-part-block-0-r-00004-snappy.parquet	parquet	June 18, 2024, 16:55:51 (UTC+0:30)	602.0 B	Standard
run-1718709944224-part-block-0-r-00006-snappy.parquet	parquet	June 18, 2024, 16:55:51 (UTC+0:30)	656.0 B	Standard
run-1718709944224-		June 18, 2024, 16:55:52		

CloudShell Feedback

CloudShell Feedback

>Click on -part-block-0-r-00000-snappy.parquet file
then click on object actions to run sql query.



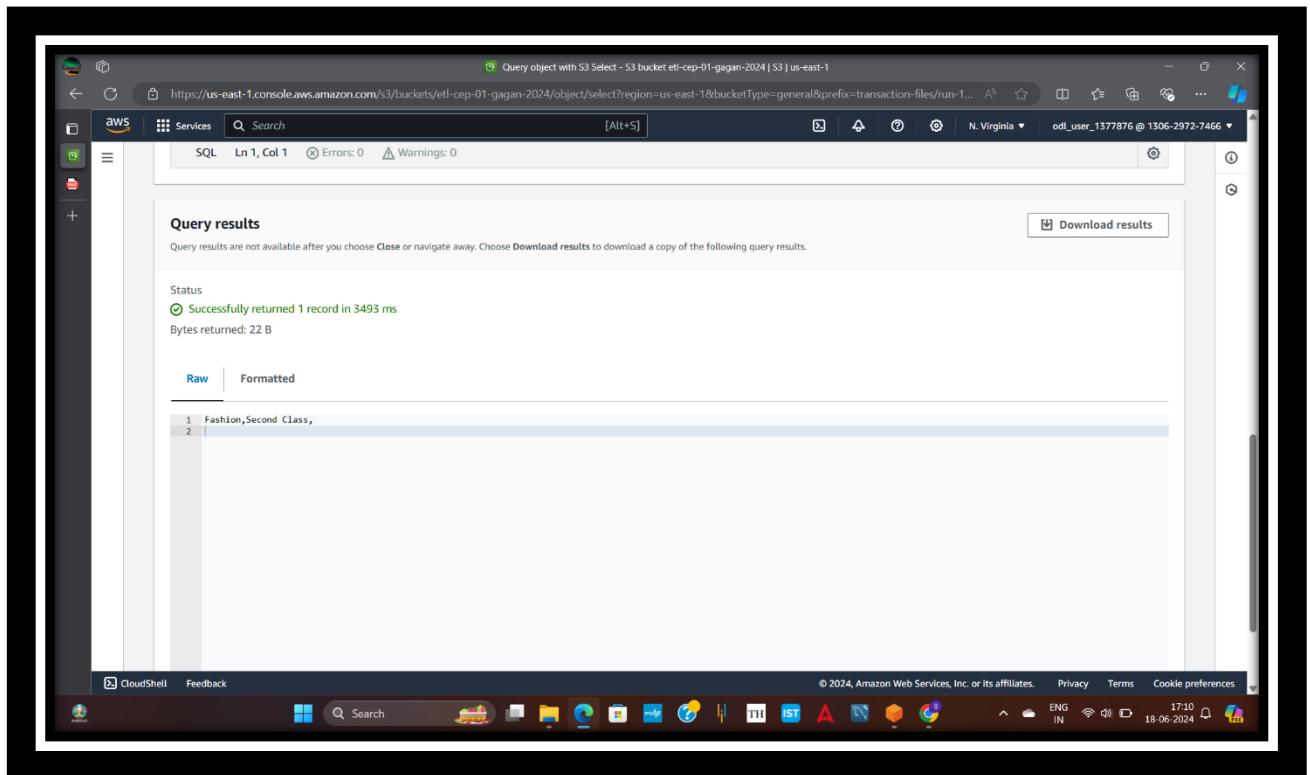
>Set the input -output setting as given in image.

The screenshot shows the AWS S3 Select configuration page for a specific object. The top section displays the object path: `s3://etl-cep-01-gagan-2024/transaction-files/run-1718709944224-part-block-0-r-00000-snappy.parquet`, size: 611.0 B, and format: Apache Parquet. A note states: "Amazon S3 Select does not support whole-object compression for Apache Parquet objects." Below this is the "Output settings" section, which is currently set to CSV format with a comma delimiter. The bottom of the screen shows a standard Windows taskbar with various icons and system status.

>Click run on sql query.

The screenshot shows the AWS S3 Select SQL query execution interface. In the "SQL query" section, the user has entered the following SQL command: `SELECT * FROM s3object s LIMIT 5`. The "Run SQL query" button is highlighted in orange. Below the query, the status message indicates: "Successfully returned 1 record in 3493 ms". The bottom section, "Query results", shows the status "Status" and the message "Successfully returned 1 record in 3493 ms". The Windows taskbar at the bottom is visible.

>See the query result as follows as status-successfully can see the result in raw as well as formatted form.



The screenshot shows the AWS Lambda SQL interface. The URL in the browser is <https://us-east-1.console.aws.amazon.com/s3/buckets/etl-cep-01-gagan-2024/object/select?region=us-east-1&bucketType=general&prefix=transaction-files/run-1...>. The page displays a "Query results" section with the following content:

Status
Successfully returned 1 record in 3493 ms
Bytes returned: 22 B

Raw | Formatted

1	Fashion,Second Class,
2	

Below the table, there are "CloudShell" and "Feedback" buttons. The bottom of the screen shows a Windows taskbar with various icons.