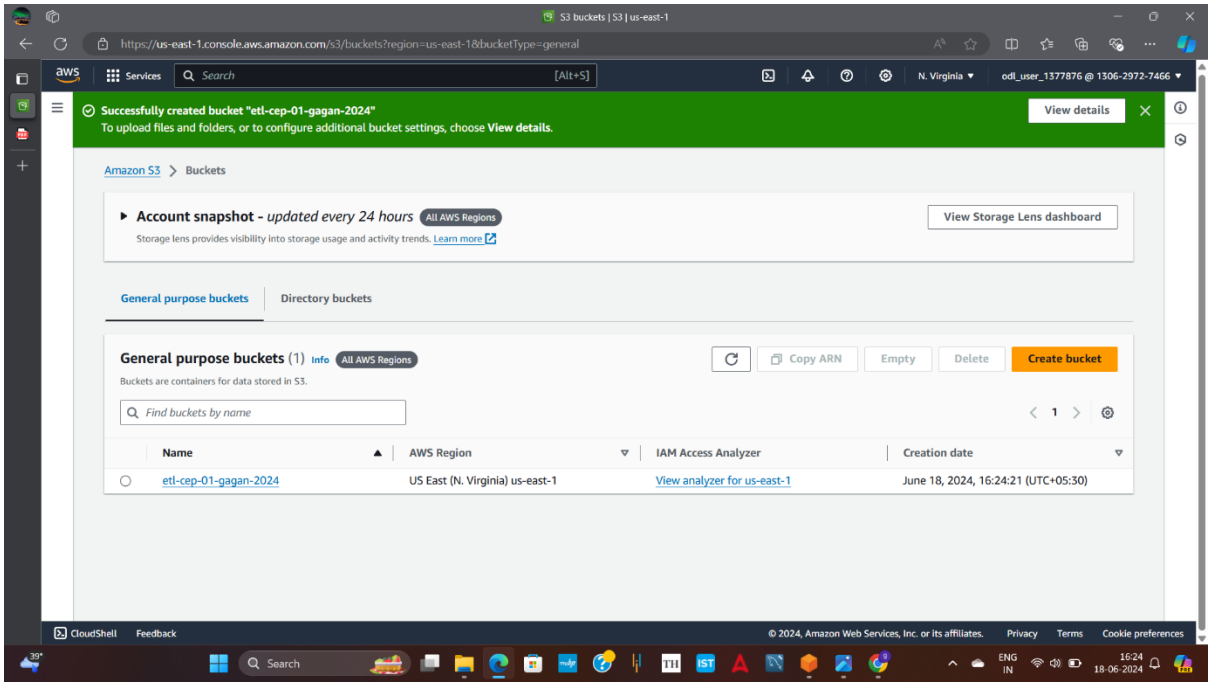
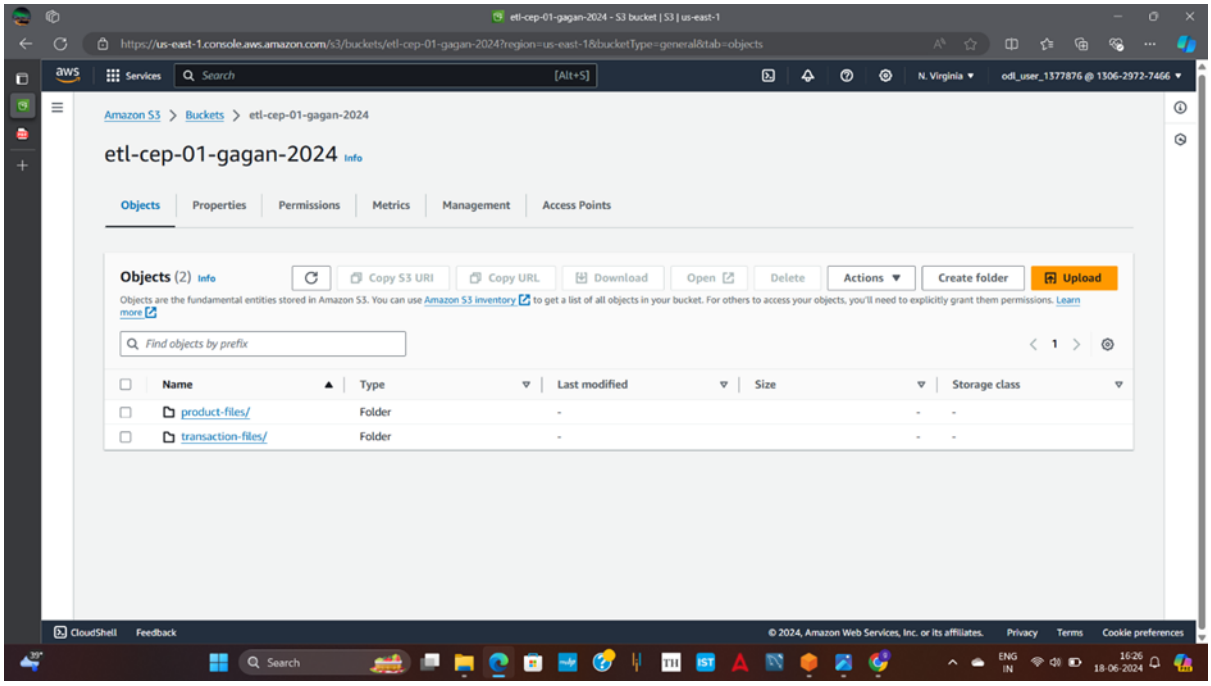
COURSE END PROJECT(ETL)

GAGAN SHRIVASTAVA

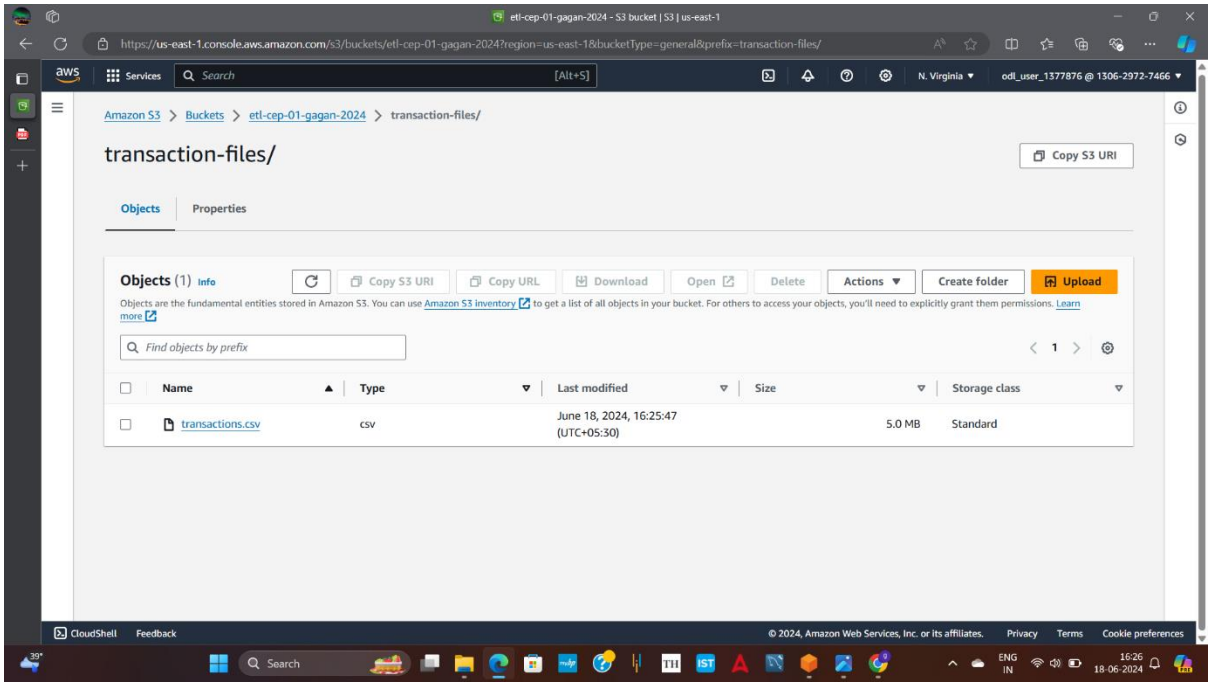
>bucket created as -etl-cep-gagan-2024 in S3 bucket.



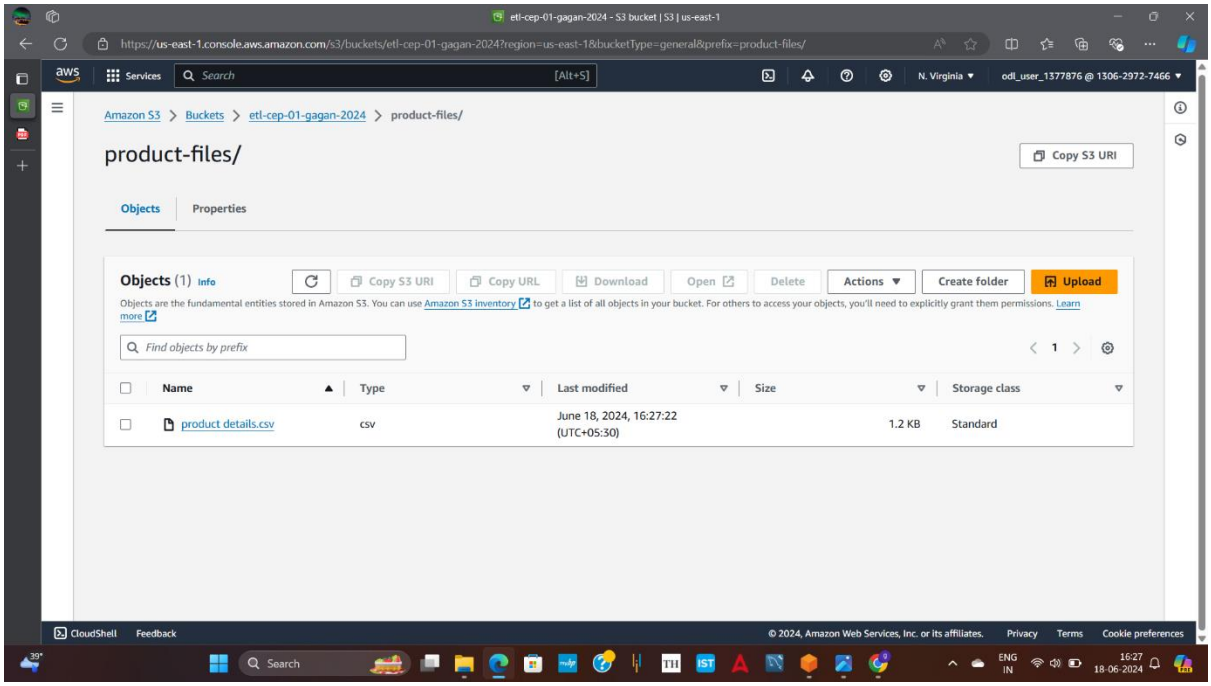
>created transaction files and product files inside bucket.



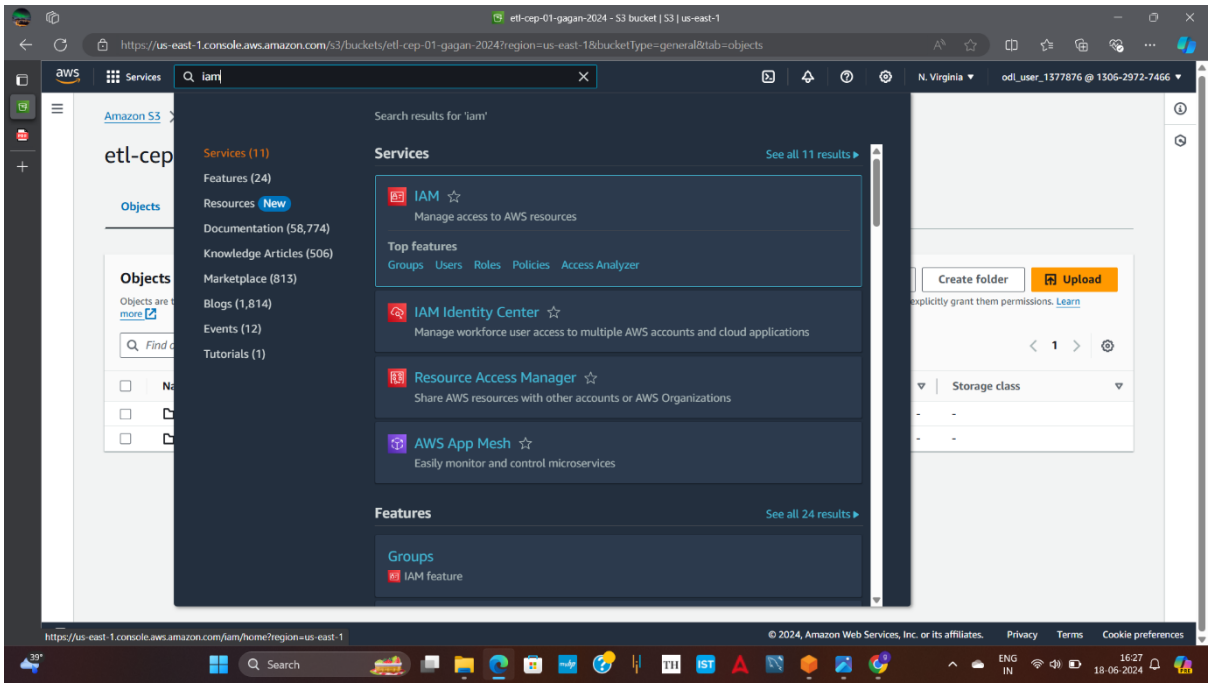
>upload transaction csv data file inside transaction file.



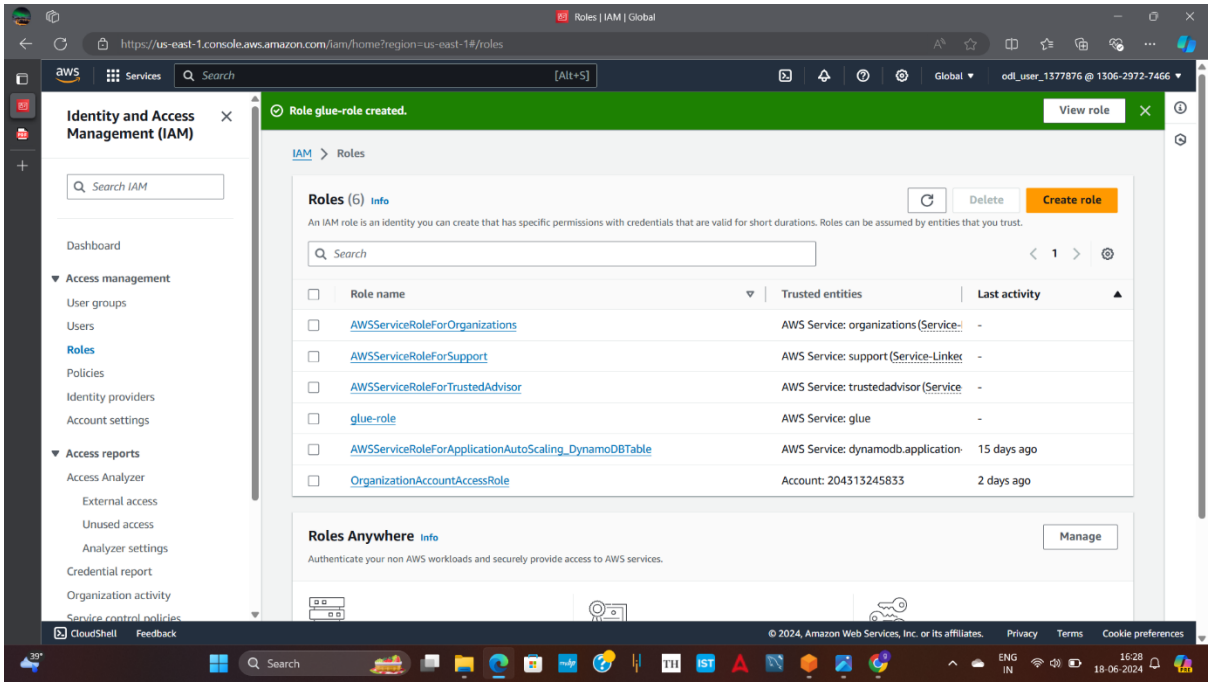
>Upload product csv data file inside product file.



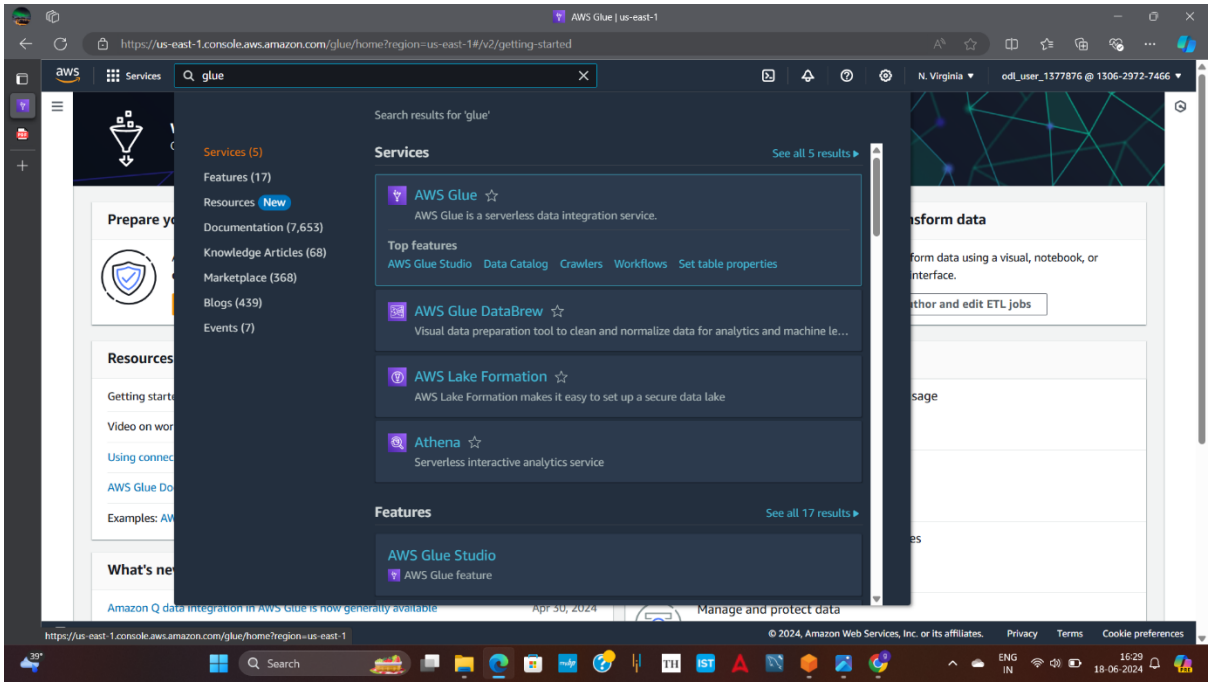
>Search iam in search bar and then click it for creating a role.



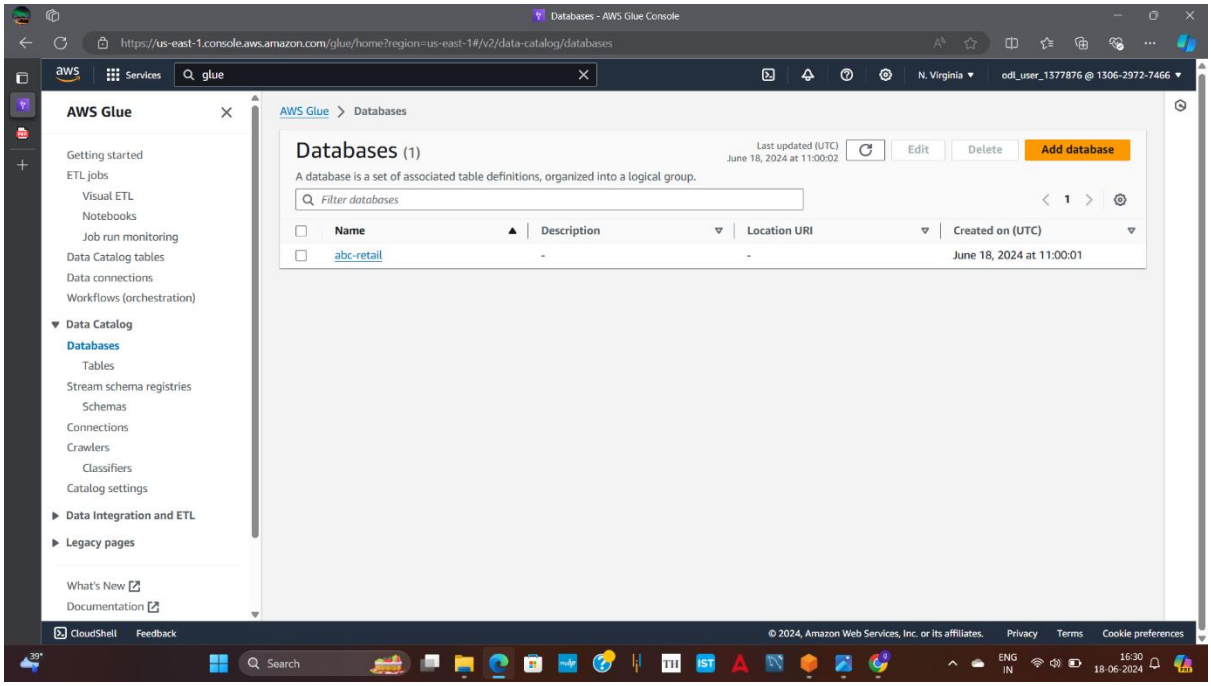
>create a role as- glue-role to get permission and access to data for glue.



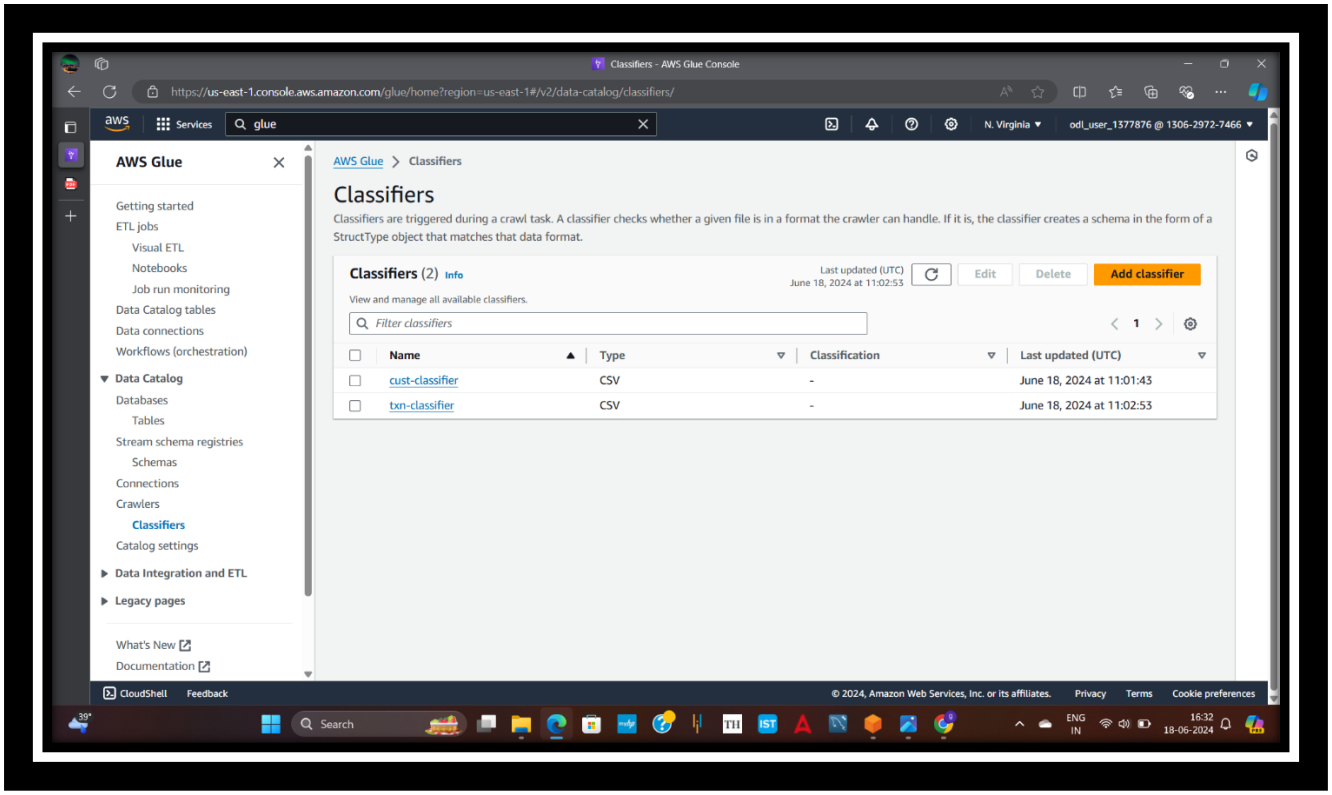
>search aws glue in search bar and then click it.



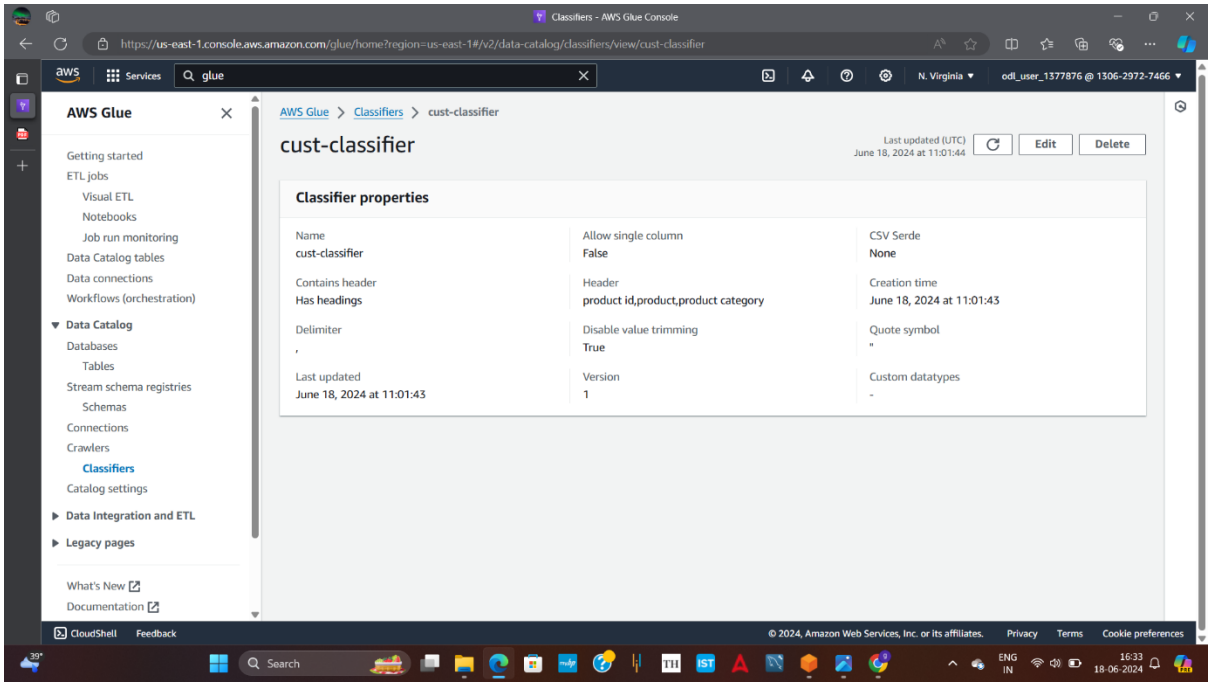
>create database as- abc-retail in databases under data catalog.



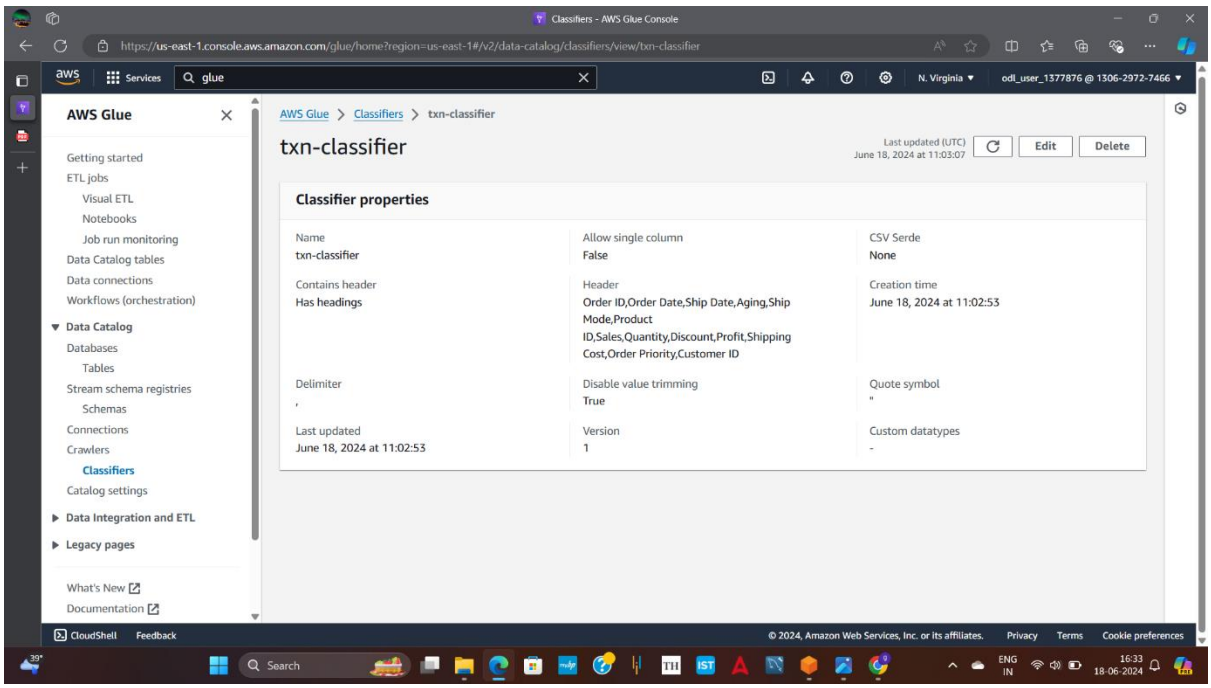
>Create two separate classifiers as cust-classifiers for product files and txn-classifier for transaction files.



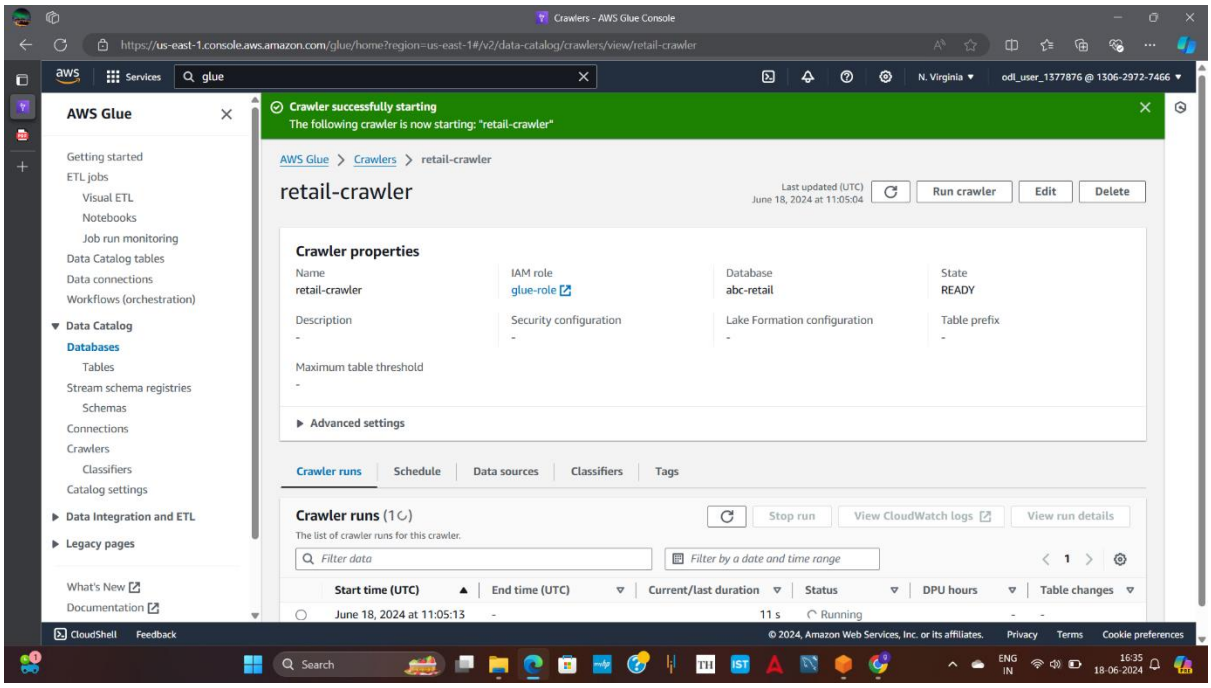
>cust-classifier properties.



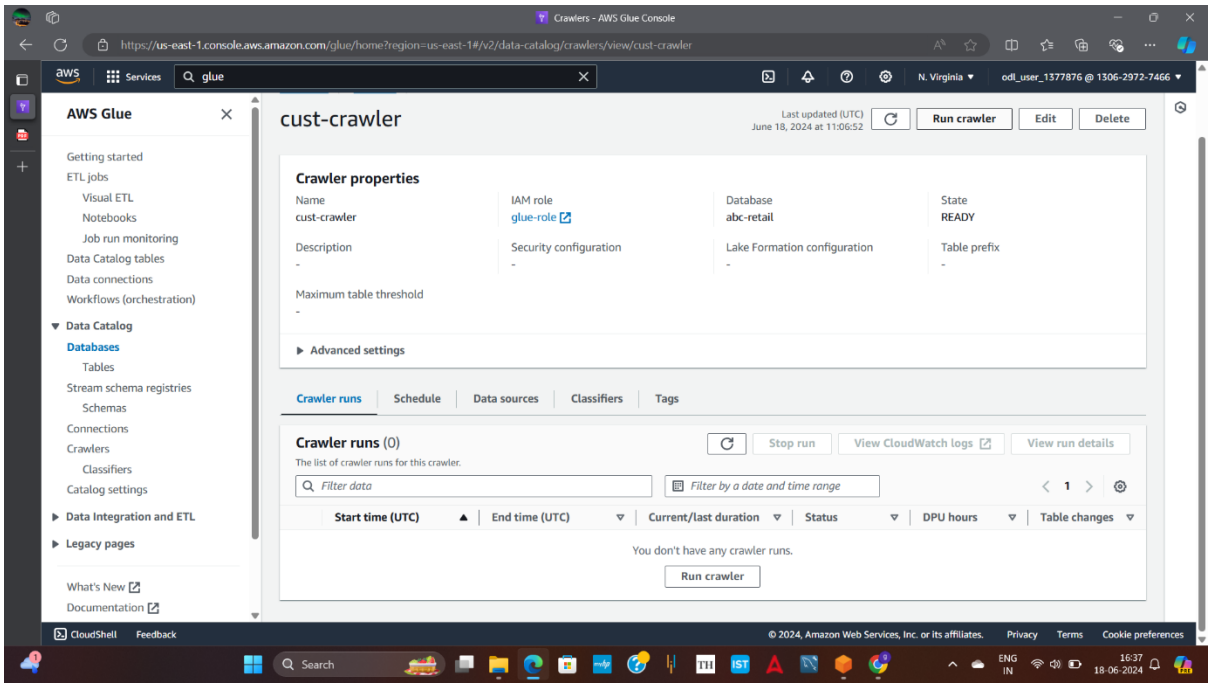
>txn-classifier properties.



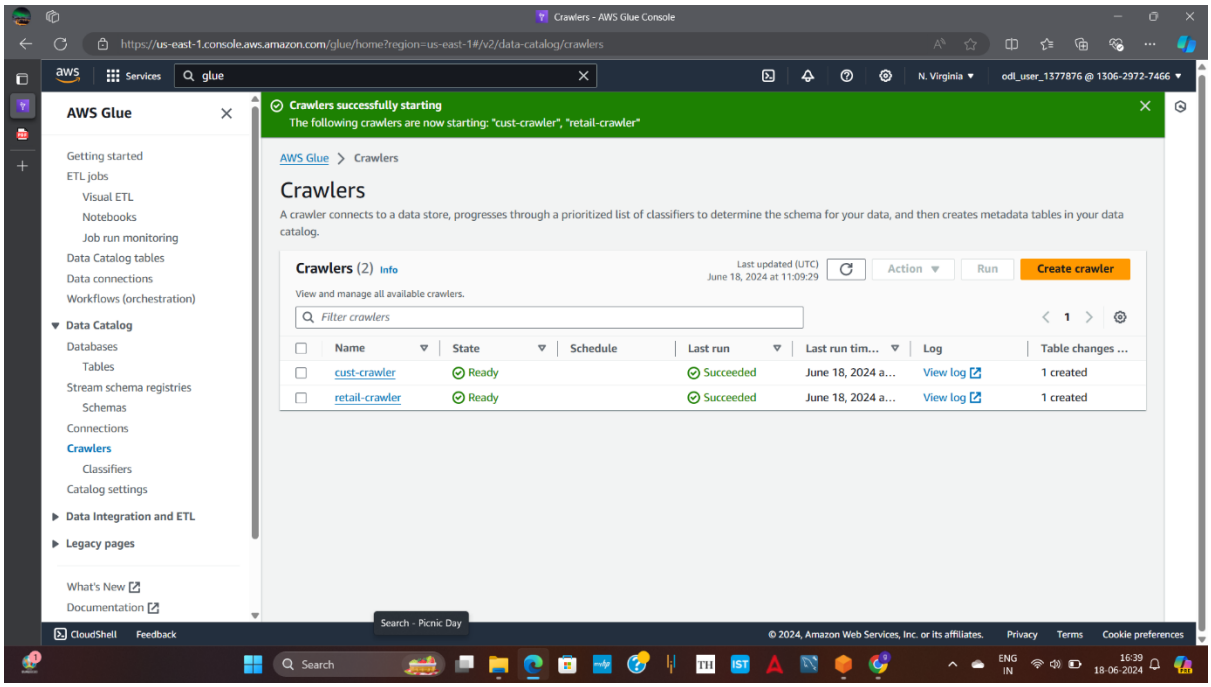
>Inside abc-retail database create crawler as retail-crawler for transaction files then run crawler.



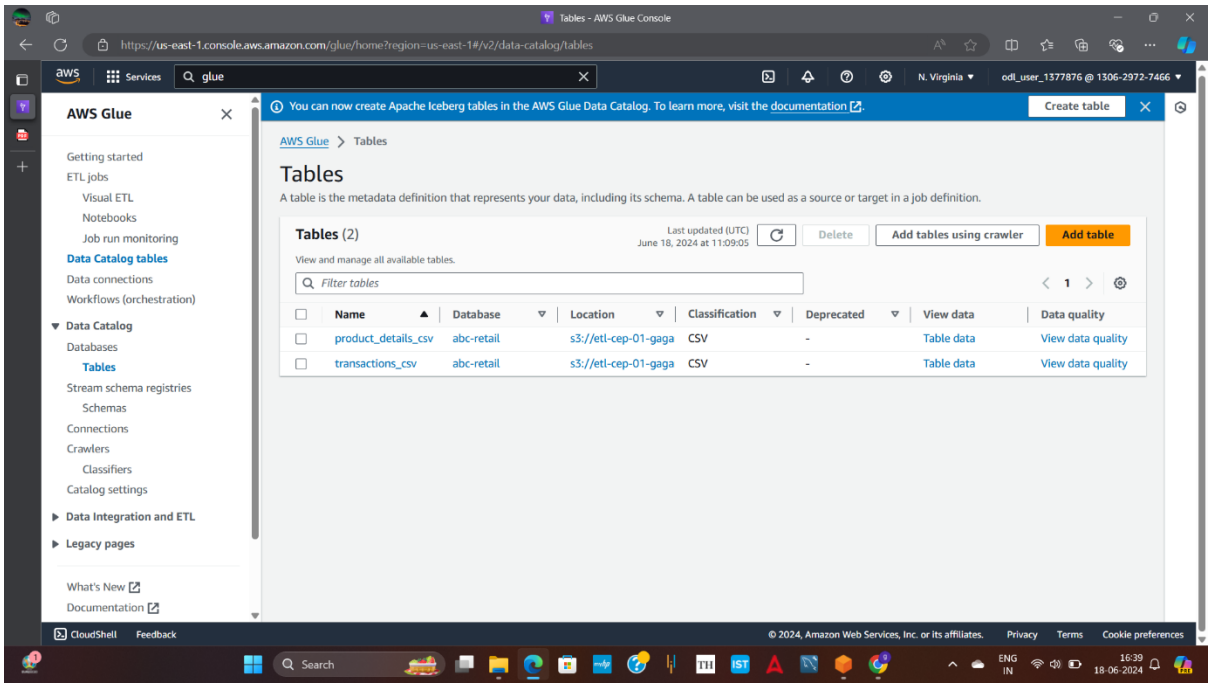
>create another crawler inside abc-retail database as cust-crawler for product files then run it as well.



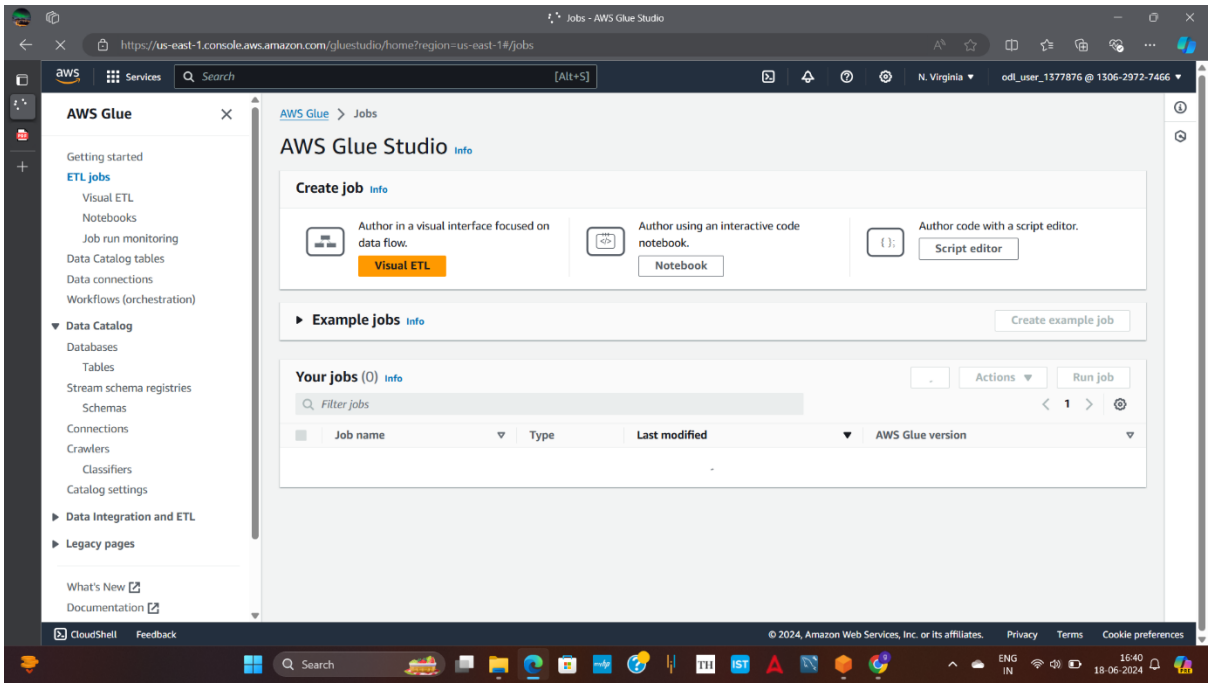
>After run wait for some time then go to the crawler under data catalog to see both crawler status as ready.



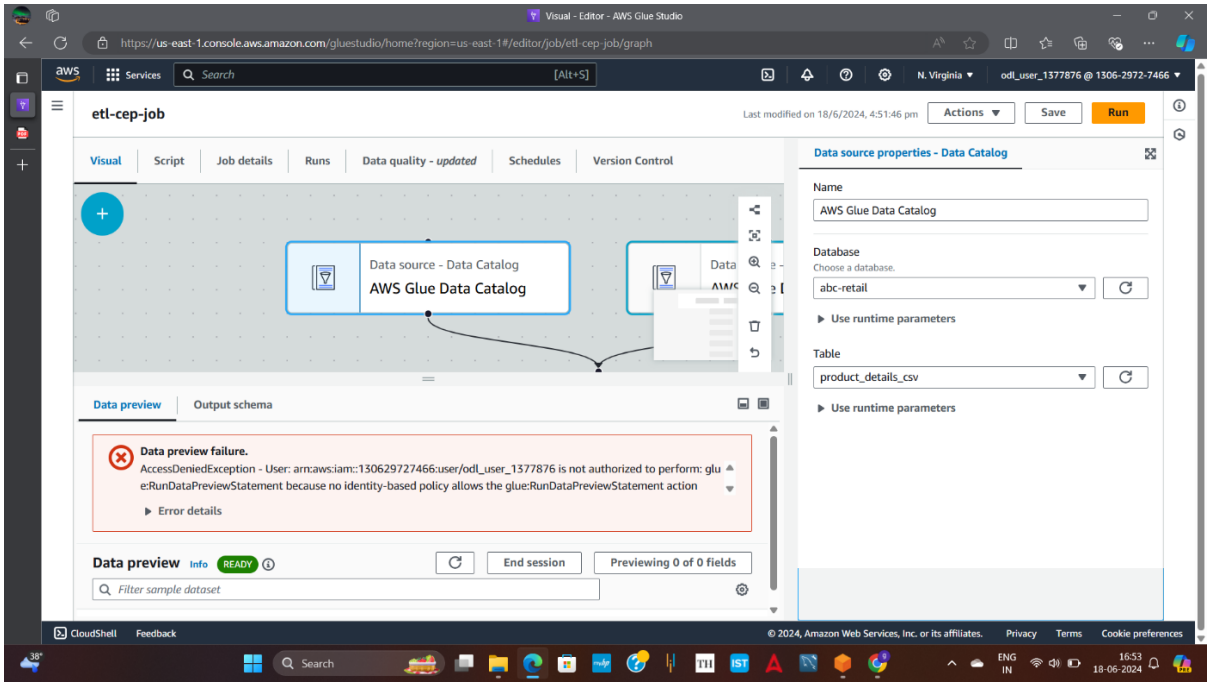
>Go to the table under data catalog to see both table got created.



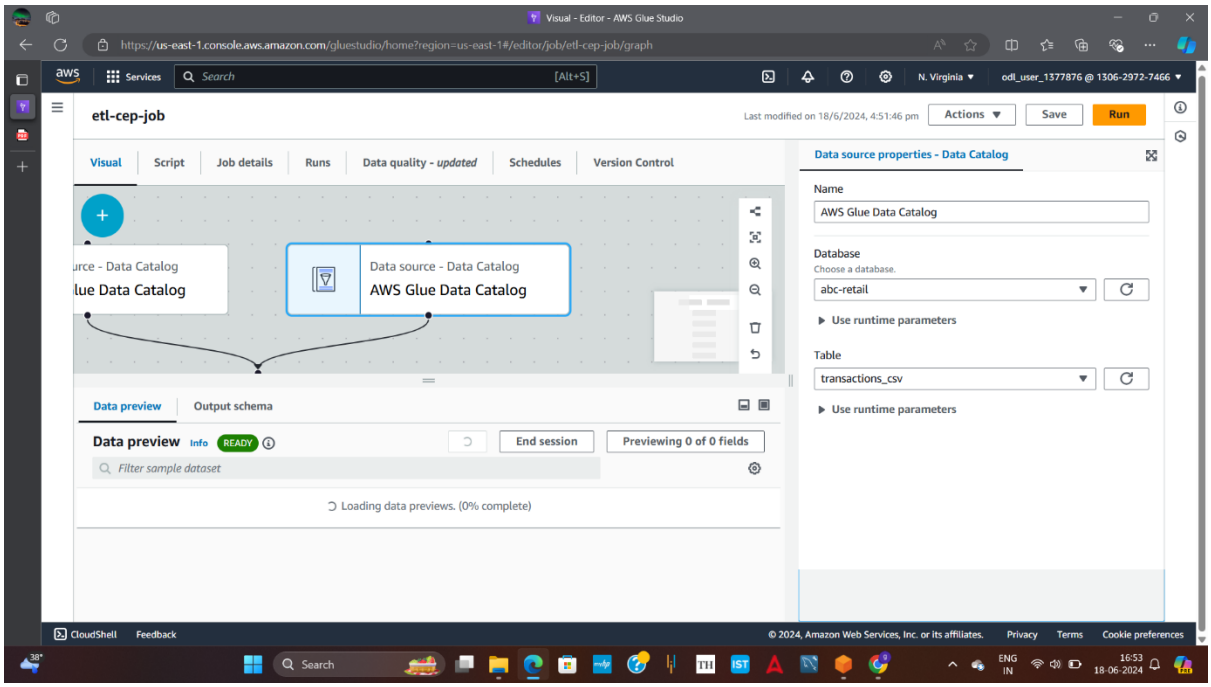
>Go to the etl jobs under which click on visual etl.



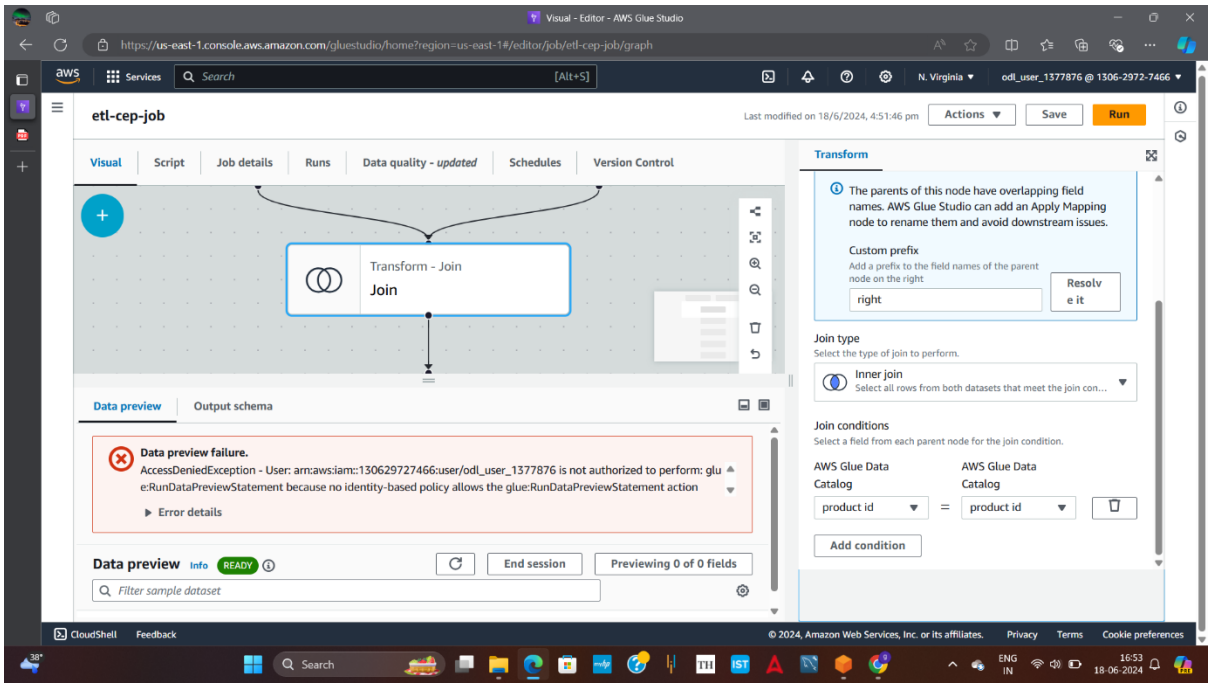
>create block of aws glue data catalog(extract) for product csv data file.

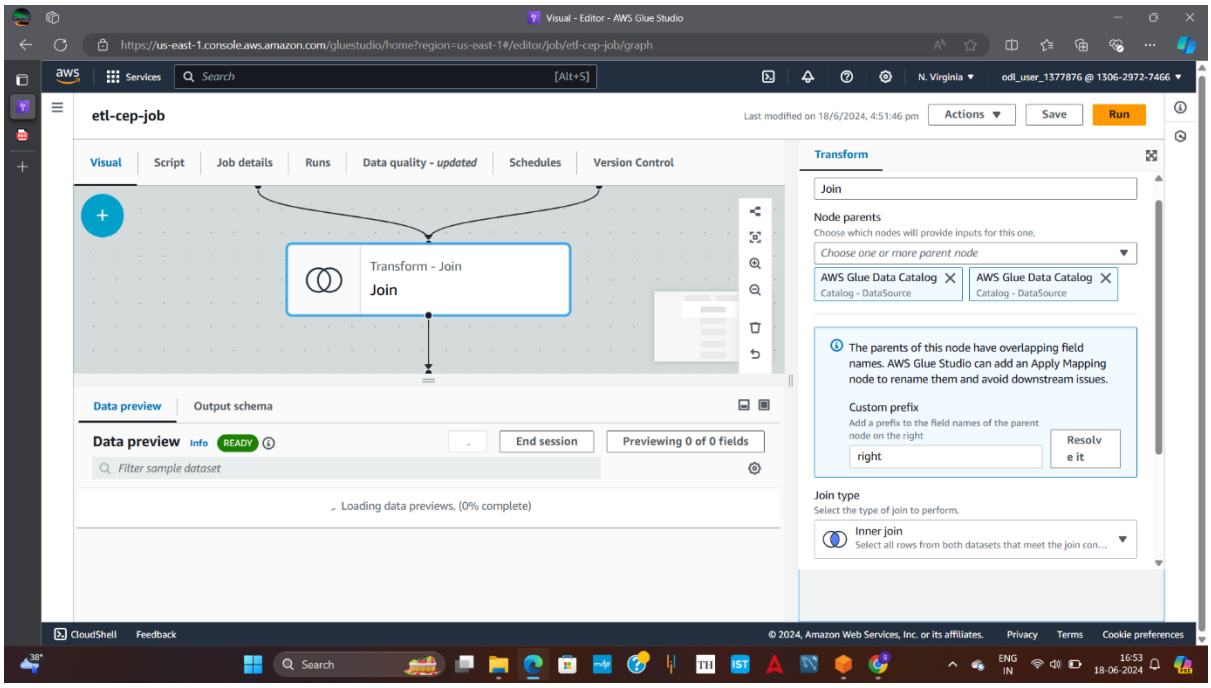


>Create another block of aws glue data catalog(extract) for transaction csv data file.

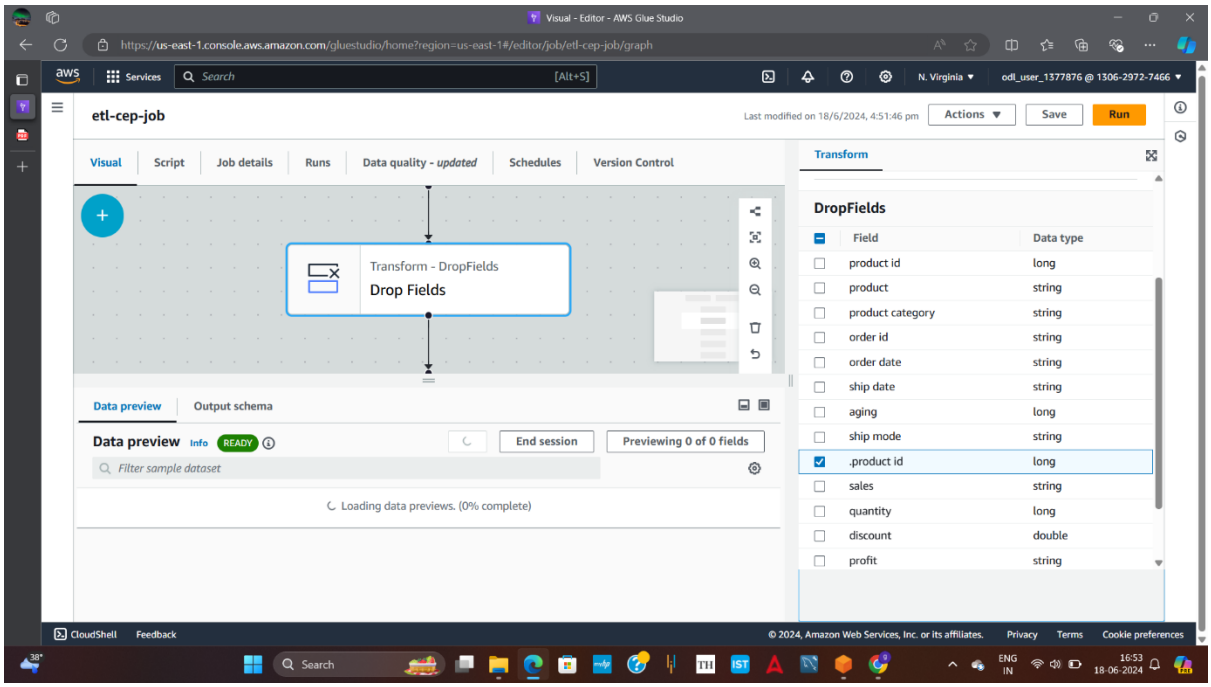


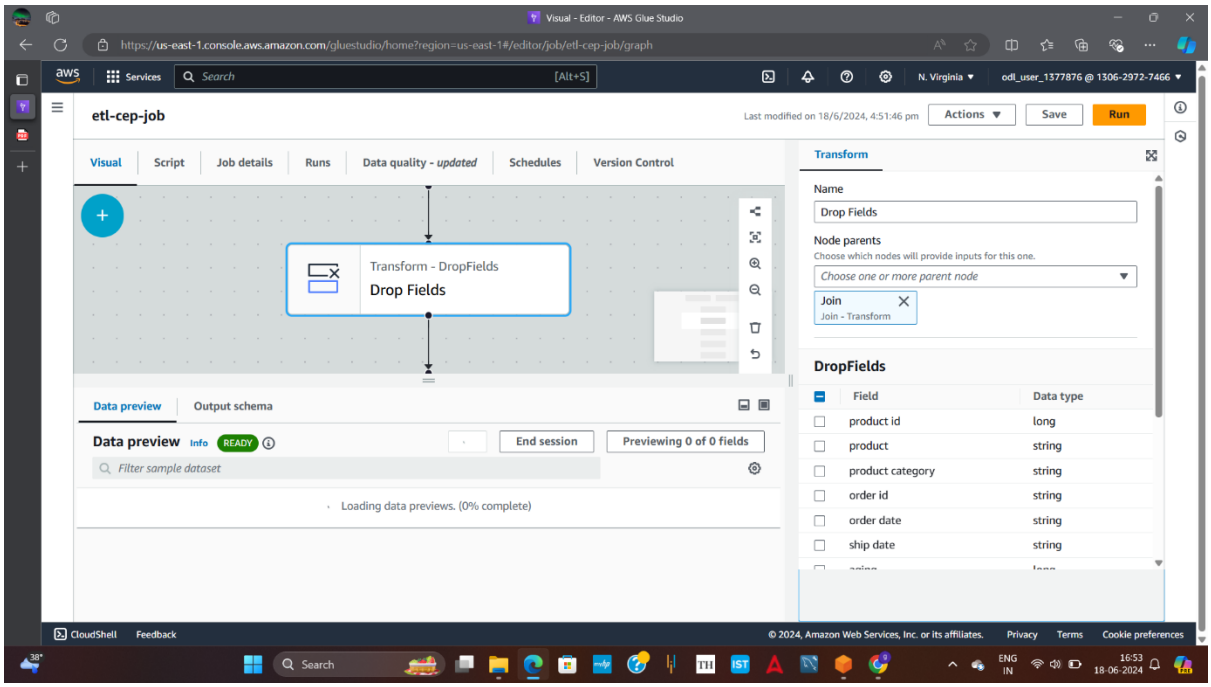
>join(transform) both the data catalog(inner join) on the basis of product id.



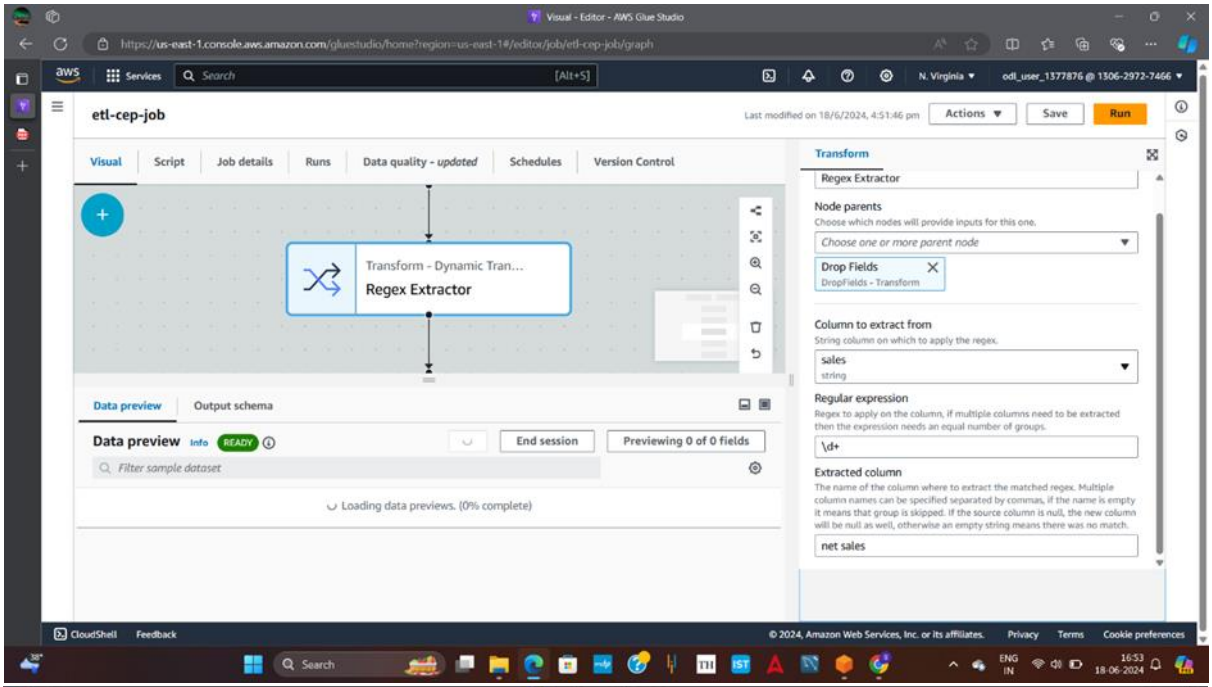


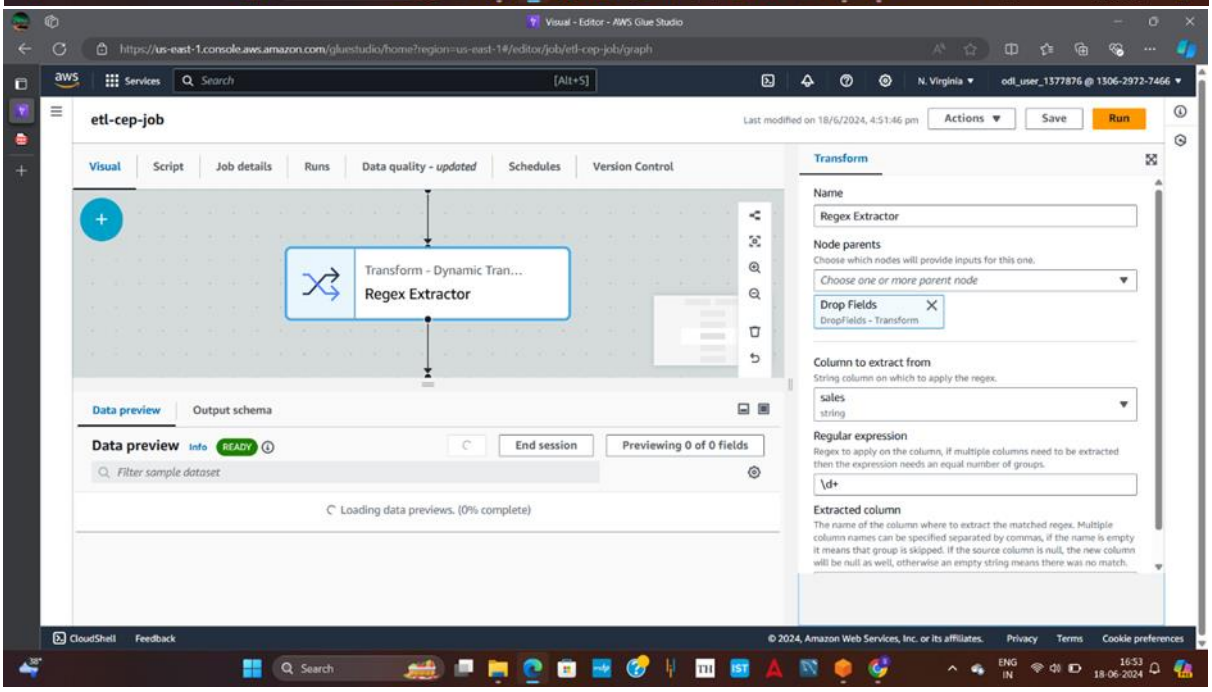
>To drop(transform) the extra field – product id because of join.



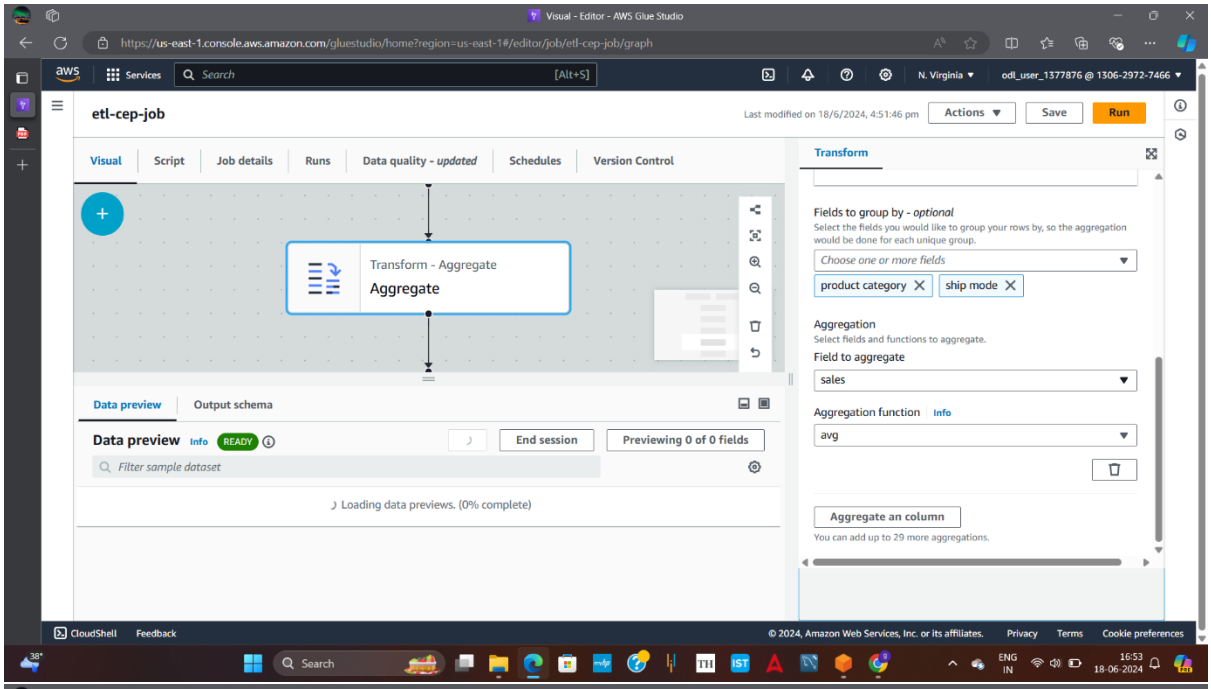


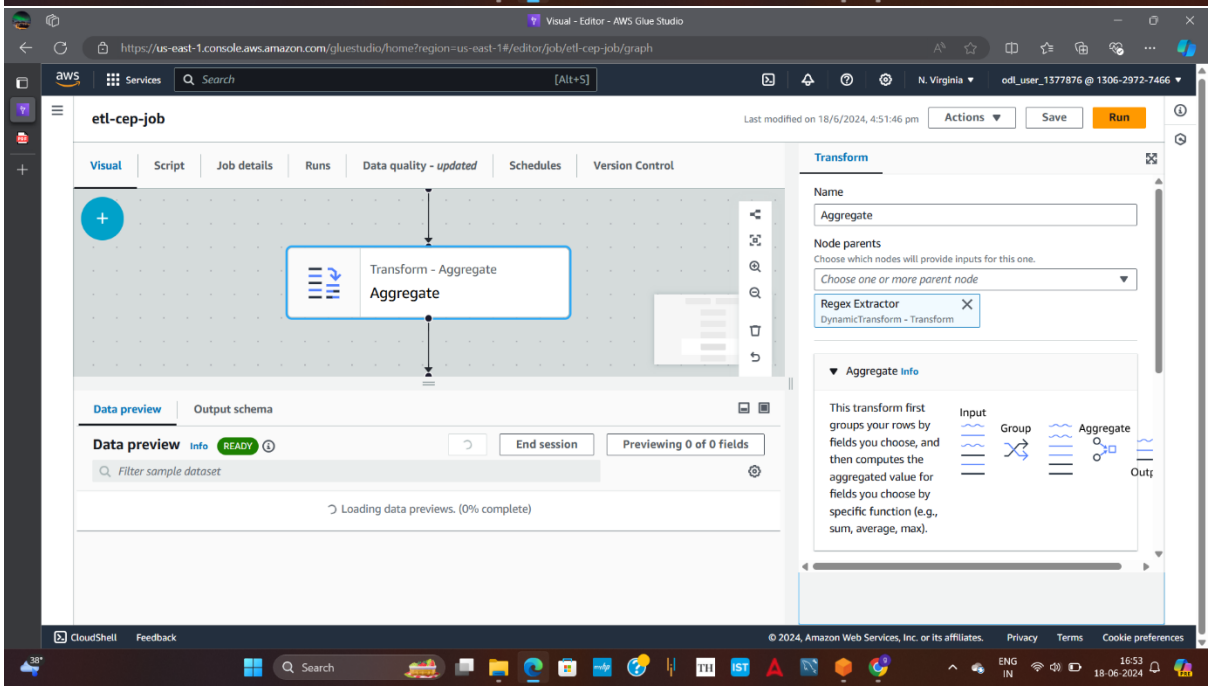
>Add regex extractor(transform) to cleanse the data.



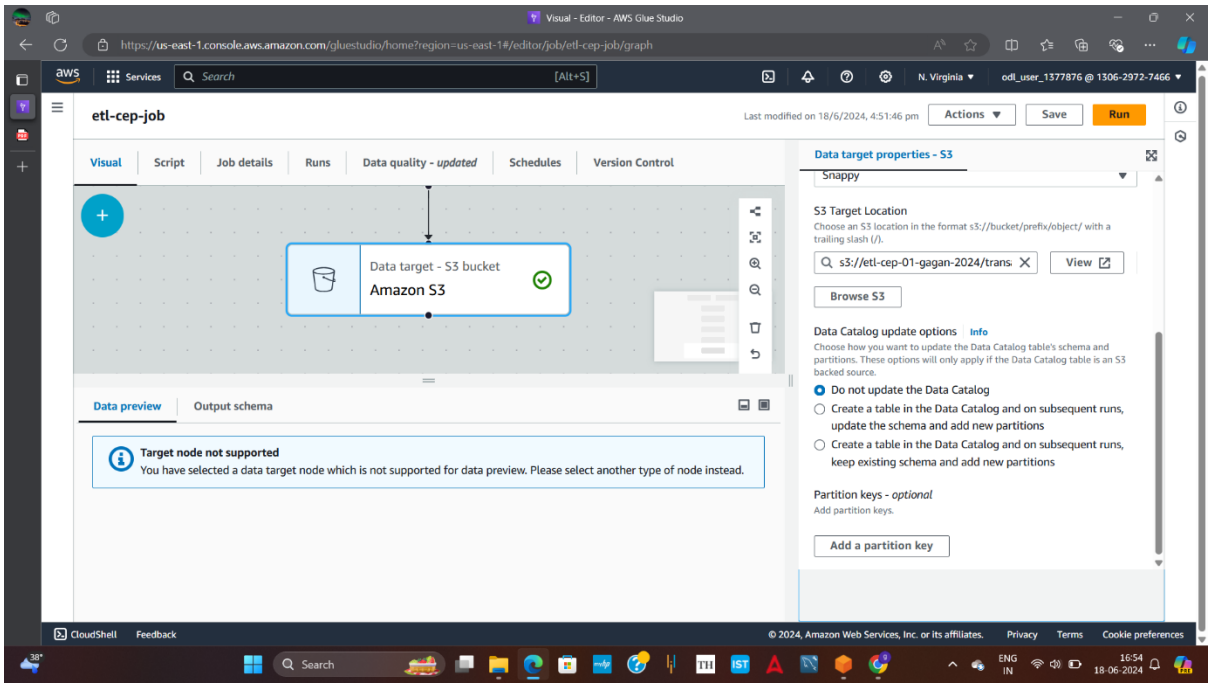


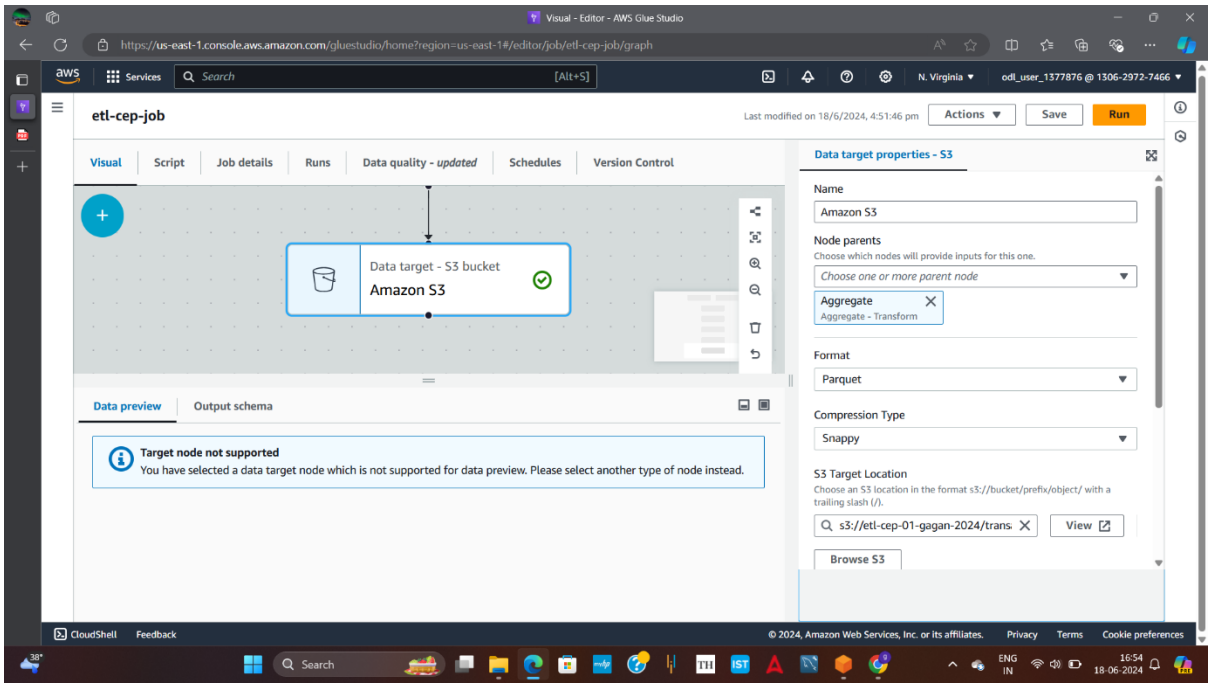
>Add aggregator(transform) to apply group by and aggregate fuctions on data as follows.





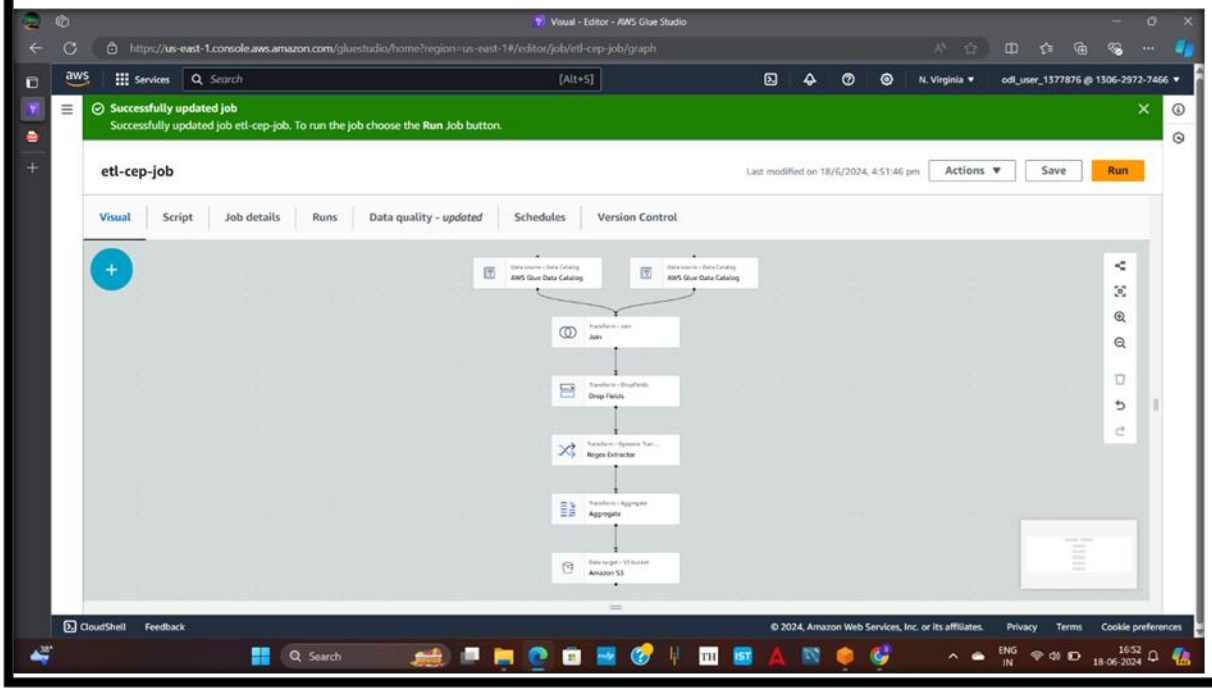
>Target location amazon s3(load) the data we load in the block and choose target location to see the output.



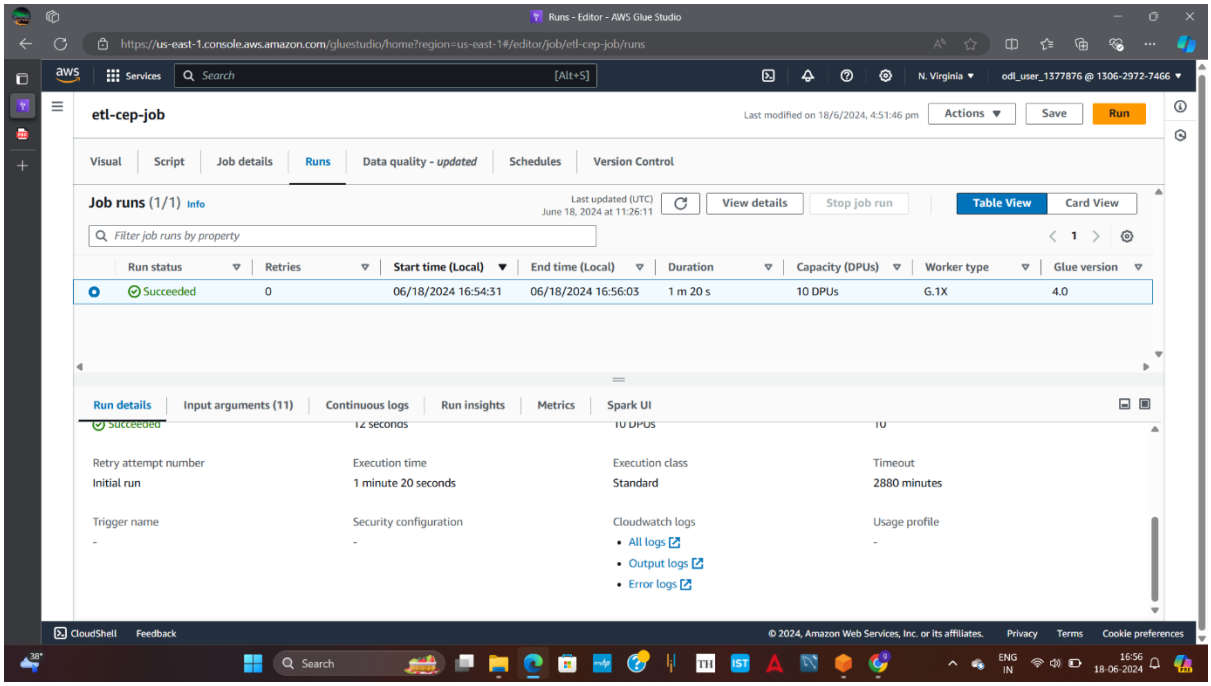


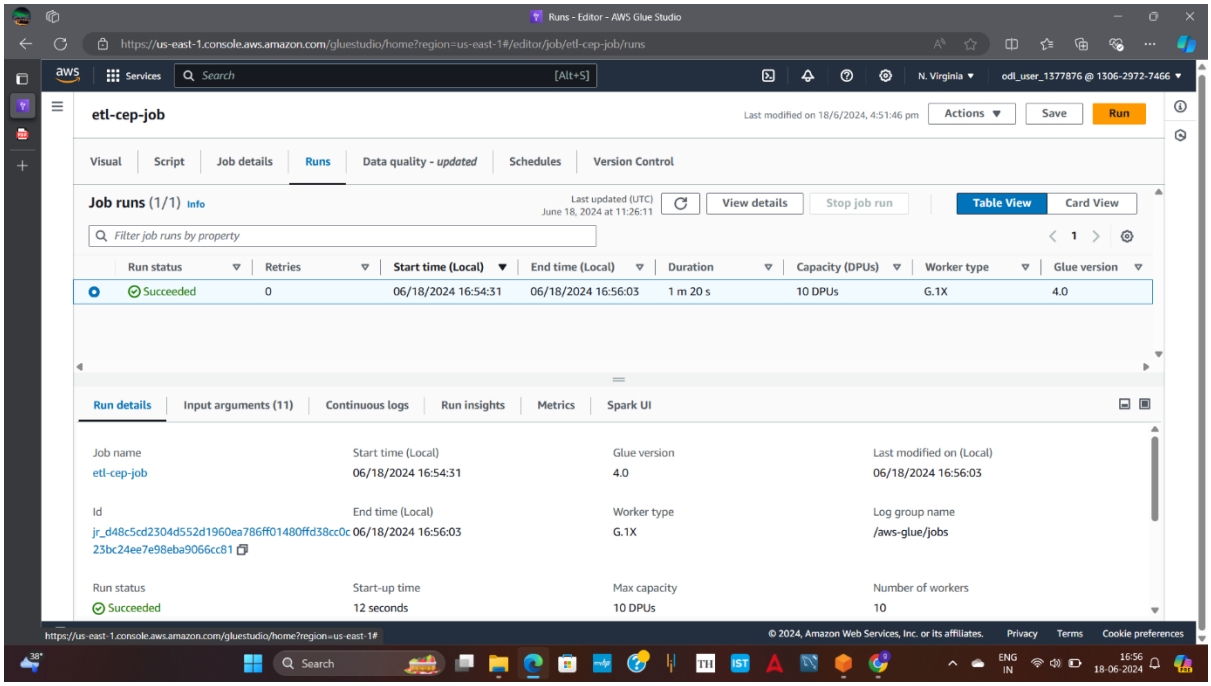
the output location is not as per the cep as create diff bucket for output location because it is not transferred there so I transferred the output in location of transaction file under etl-cep-01-gagan-2024.

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

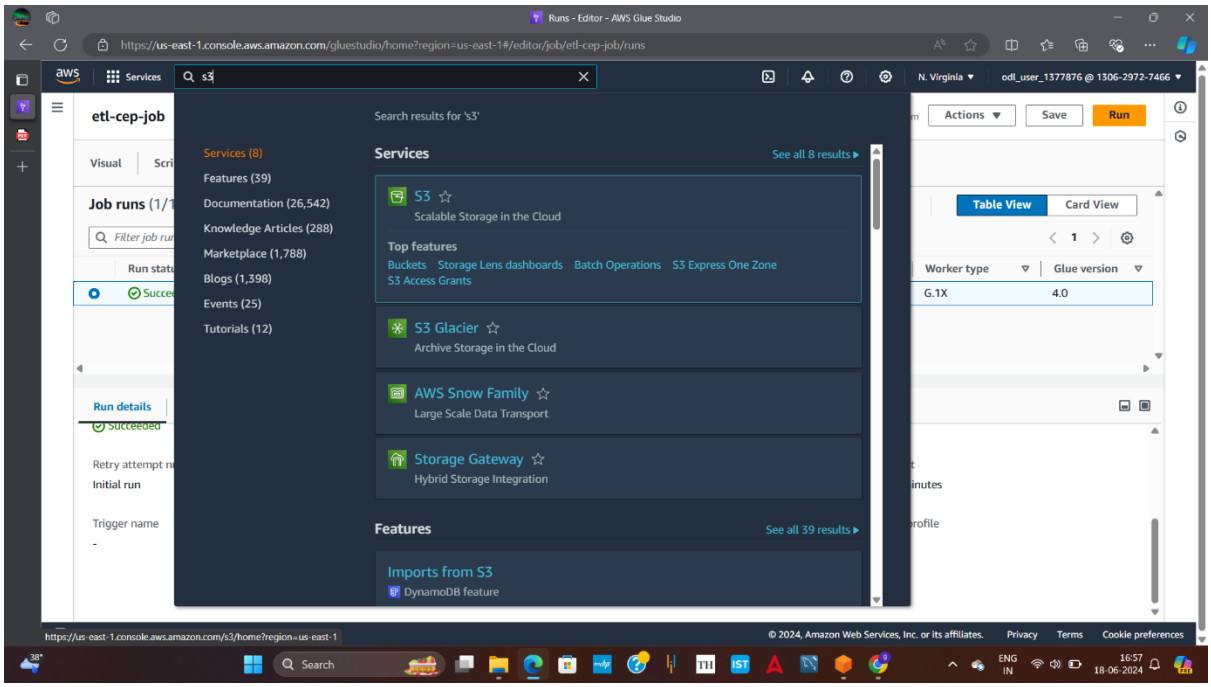


>See the etl-cep-job details as succeeded.

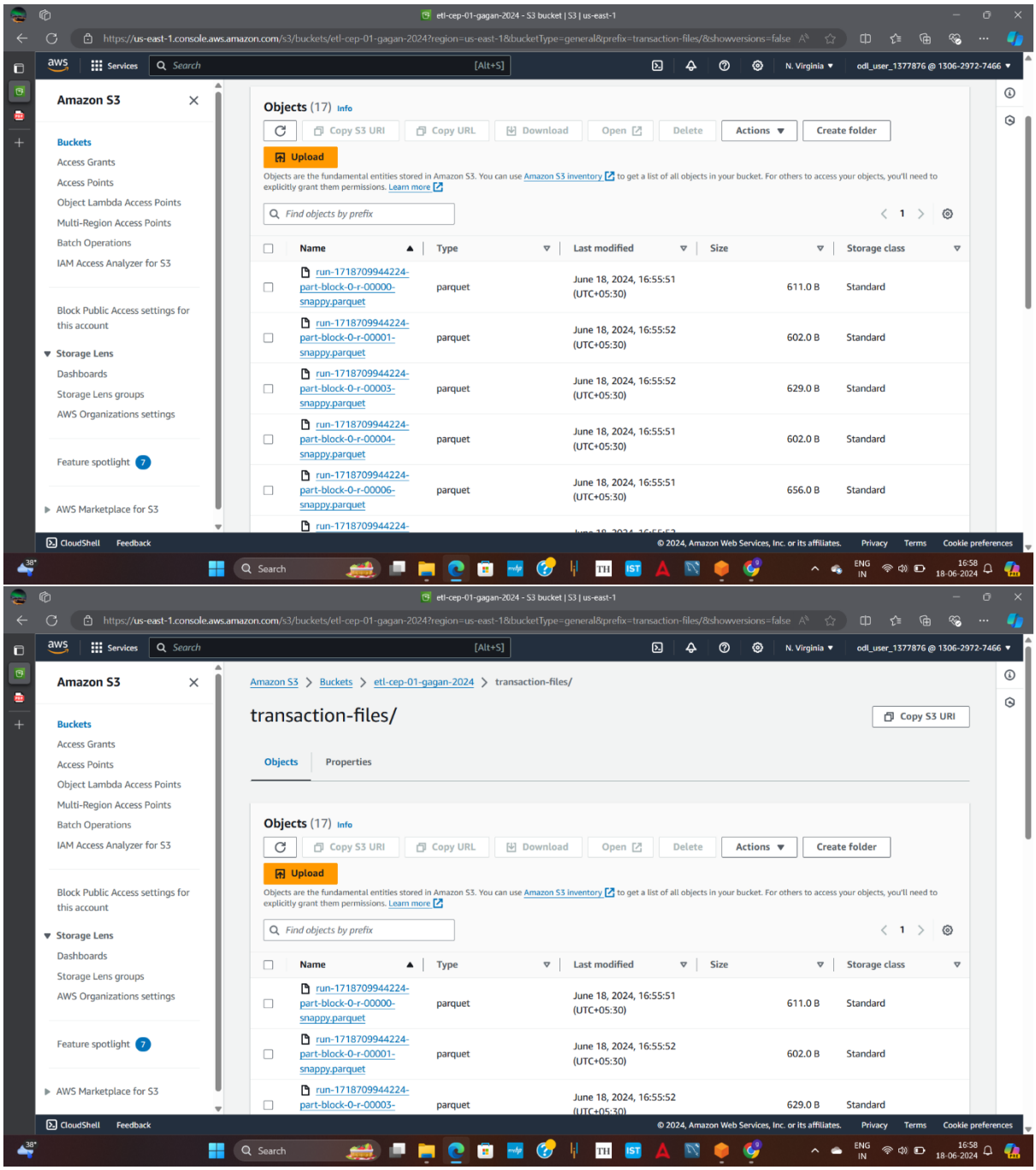




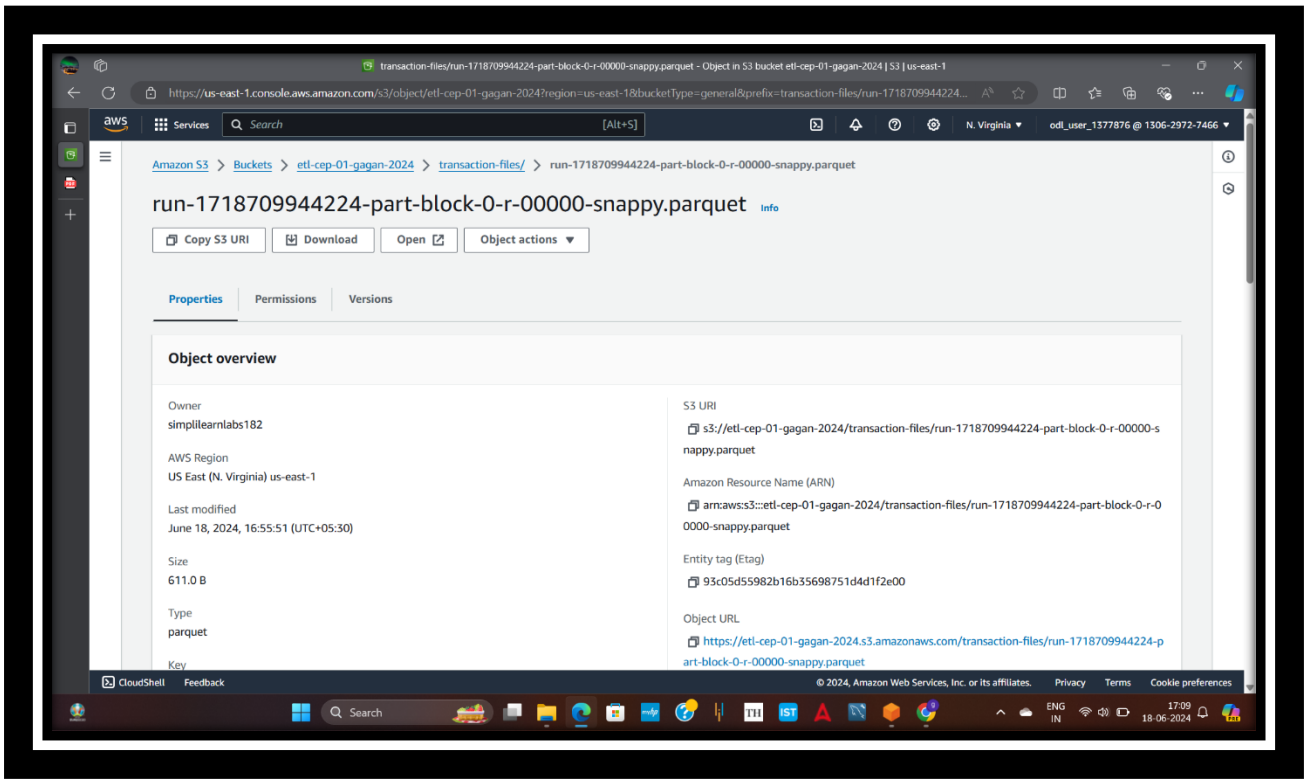
>Search s3 on search bar then click it to go to the buckets.



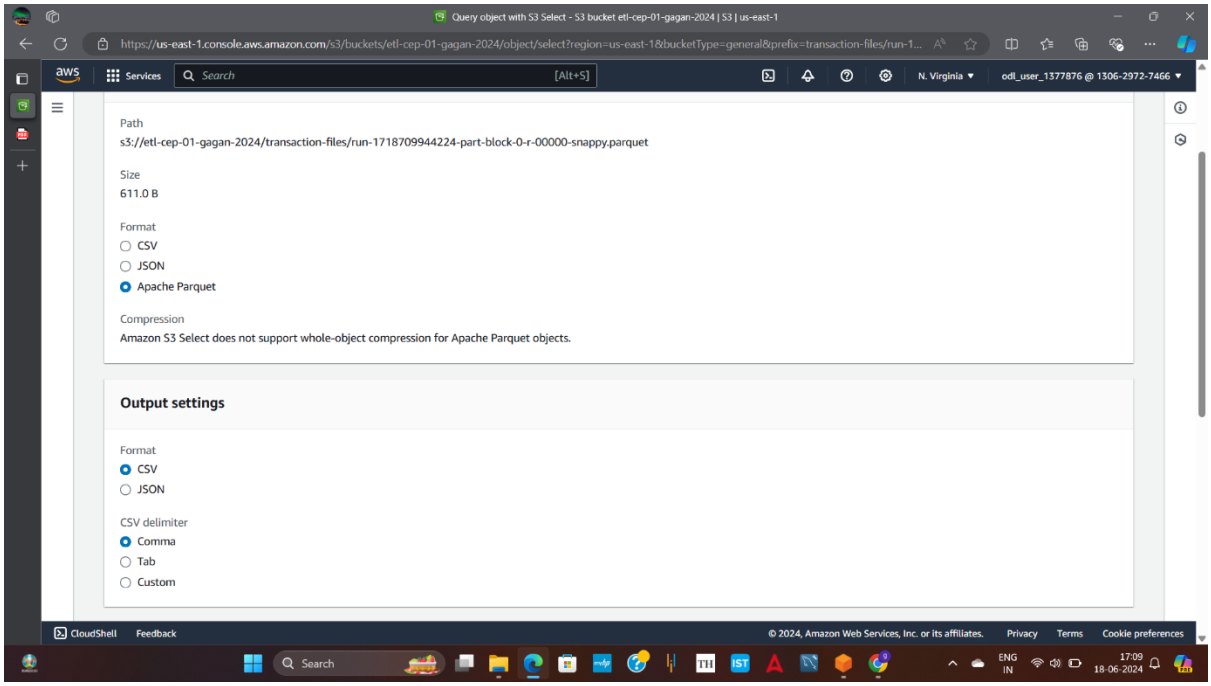
>Click on transaction file under etl-cep-01-gagan-2024 bucket to see the output of etl-cep-job.



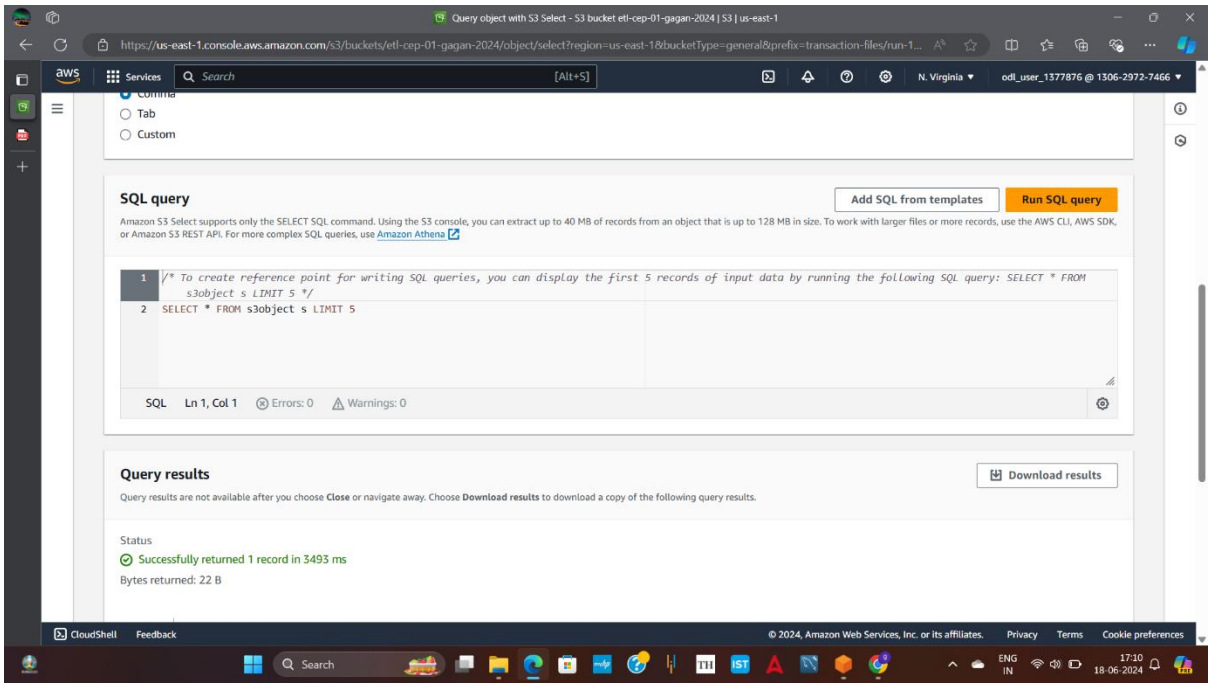
>Click on -part-block-0-r-00000-snappy.parquet file then click on object actions to run sql query.



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



>Click run on sql query.



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