

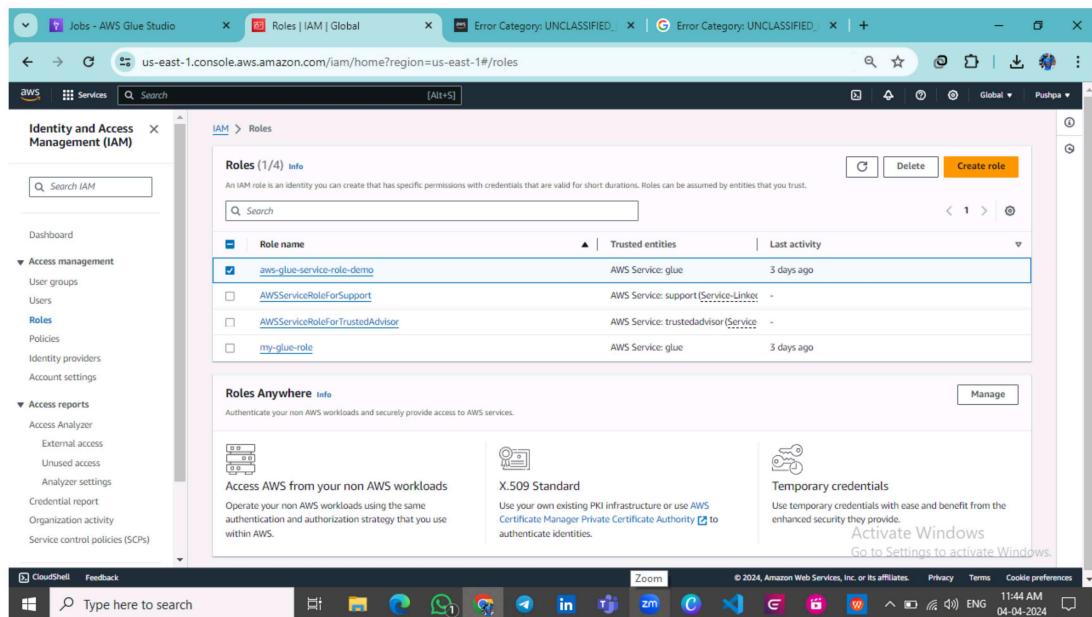
NAME : BELLAMKONDA PUSHPA SHREE

COURSE : AWS

PROJECT :AWS GLUE ETL

step-1

create an IAM role to access AWS Glue +EC2+CloudWatch+s3 here the role name is "aws-glue-service-role-demo"



step 1 upload source csv file to s3

2.1 create a bucket, here it is “aws-glue-demo-series”

The screenshot shows the AWS S3 console interface. On the left, there's a sidebar with options like Buckets, Access Grants, Access Points, Object Lambda Access Points, Multi-Region Access Points, Batch Operations, and IAM Access Analyzer for S3. Below that is a Storage Lens section. The main area is titled 'Amazon S3' and shows an 'Account snapshot' with a link to 'View Storage Lens dashboard'. Under 'General purpose buckets', there's a table with three entries:

Name	AWS Region	IAM Access Analyzer	Creation date
aws-glue-assets-102112187992-us-east-1	US East (N. Virginia) us-east-1	View analyzer for us-east-1	March 31, 2024, 11:33:03 (UTC+05:30)
aws-glue-demo-series	US East (N. Virginia) us-east-1	View analyzer for us-east-1	March 31, 2024, 14:58:26 (UTC+05:30)
pushpa-project-bucket	US East (N. Virginia) us-east-1	View analyzer for us-east-1	March 29, 2024, 14:22:04 (UTC+05:30)

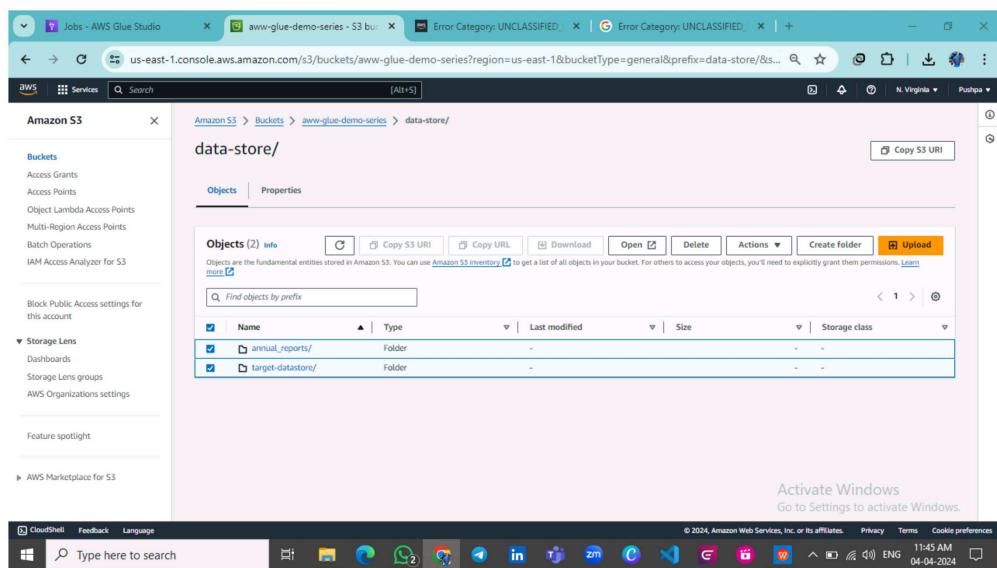
At the bottom of the table, there are buttons for 'Create bucket', 'Copy ARN', 'Empty', and 'Delete'. A search bar at the top of the main area says 'Find buckets by name'. The status bar at the bottom right shows 'Activate Windows Go to Settings to activate Windows.' and system information like 'CloudShell Feedback Language' and 'Type here to search'.

create one folder inside the bucket here it is “data-store”

The screenshot shows the AWS S3 console interface. On the left, the navigation pane is visible with sections like 'Buckets', 'Access Grants', 'Access Points', etc. The main content area shows the 'aww-glue-demo-series' bucket. At the top, there are tabs for 'Objects', 'Properties', 'Permissions', 'Metrics', 'Management', and 'Access Points'. Below these tabs, there is a toolbar with actions like 'Copy S3 URI', 'Download', 'Open', 'Delete', 'Actions', 'Create folder', and 'Upload'. A search bar is present above the object list. The object list table has columns for 'Name', 'Type', 'Last modified', 'Size', and 'Storage class'. It shows two entries: a folder named 'data-store/' and a file named 'run-1711889163953-part-+-00000'. The 'data-store/' folder is selected, indicated by a checked checkbox. At the bottom right of the page, there is a message: 'Activate Windows' and 'Go to Settings to activate Windows.'

Name	Type	Last modified	Size	Storage class
data-store/	Folder	-	-	-
run-1711889163953-part-+-00000	-	March 31, 2024, 18:16:23 (UTC+05:30)	994.5 KB	Standard

**creating sub folders inside it,they are
“annual_reports” and “target-
datastore”**



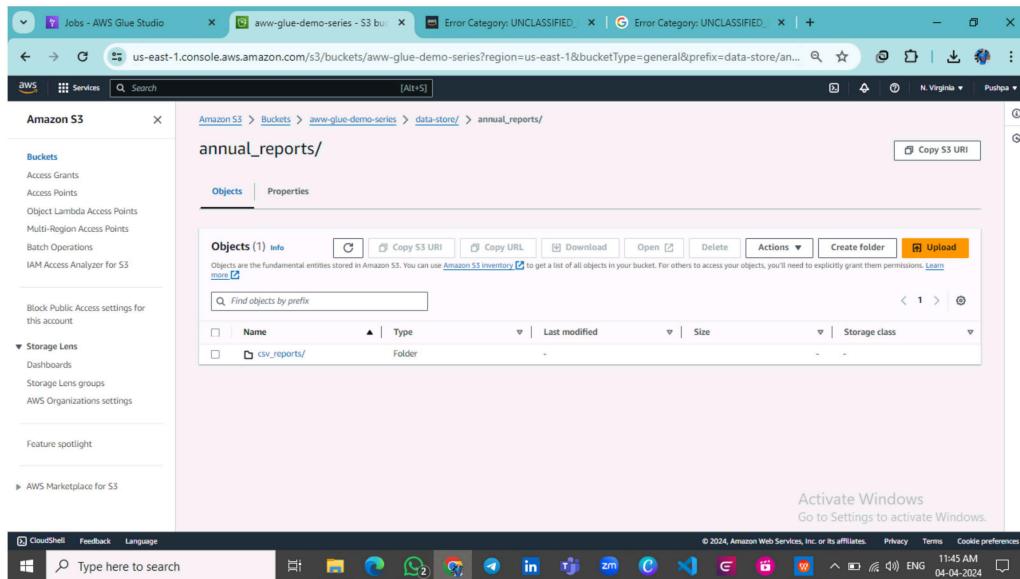
The screenshot shows the AWS S3 console interface. The left sidebar is titled "Amazon S3" and includes sections for Buckets, Access Grants, Access Points, Object Lambda Access Points, Multi-Region Access Points, Batch Operations, and IAM Access Analyzer for S3. It also has sections for Block Public Access settings for this account, Storage Lens, Dashboards, Storage Lens groups, AWS Organizations settings, Feature spotlight, and a link to the AWS Marketplace for S3.

The main content area shows a breadcrumb navigation path: Amazon S3 > Buckets > aww-glue-demo-series > data-store/. Below this, there is a table titled "Objects (2) info". The table has columns for Name, Type, Last modified, Size, and Storage class. Two entries are listed:

Name	Type	Last modified	Size	Storage class
annual_reports/	Folder	-	-	-
target-datastore/	Folder	-	-	-

At the bottom of the page, there is a message: "Activate Windows Go to Settings to activate Windows." The footer includes links for CloudShell, Feedback, Language, Privacy, Terms, and Cookie preferences, along with system status information like the date (04-04-2024) and time (11:45 AM).

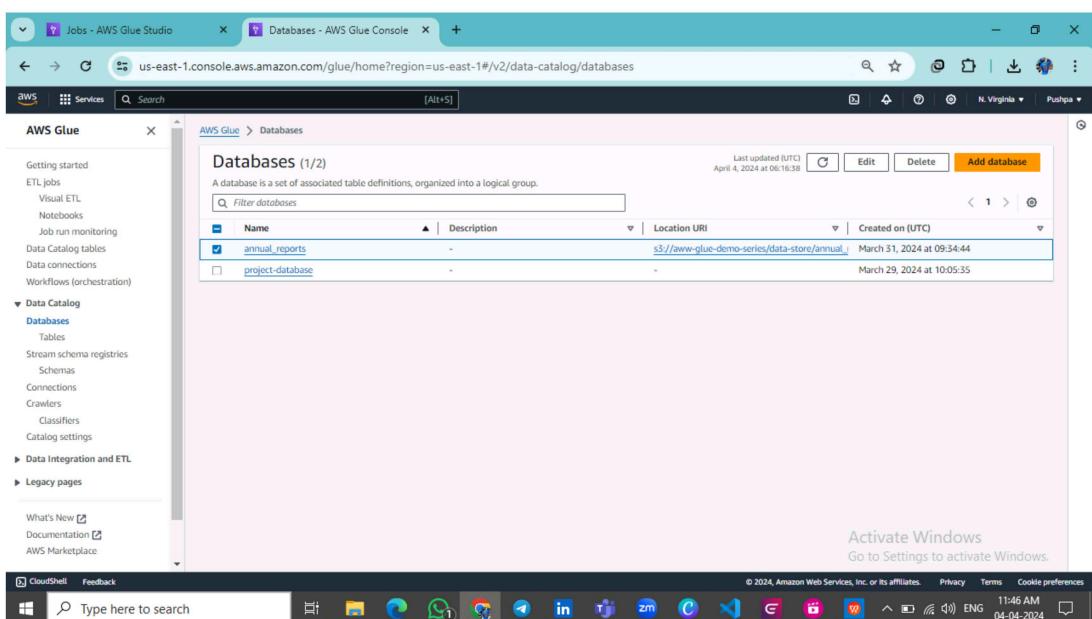
inside the “annual_reports” i upload the csv file



The screenshot shows the AWS S3 console interface. On the left, a sidebar menu includes 'Buckets', 'Storage Lens', and 'AWS Marketplace for S3'. The main area displays a list of objects in the 'annual_reports' folder of a specific bucket. The table has columns for 'Name', 'Type', 'Last modified', 'Size', and 'Storage class'. One object, 'csv_reports/', is listed as a folder. At the top of the main area, there are buttons for 'Actions', 'Create folder', and 'Upload'. A 'Copy S3 URI' button is also visible.

The “target-datastore” is to store the output csv file

step-3 Start the aws glue database,here the database name is “annual-reports”



The screenshot shows the AWS Glue Console interface. The left sidebar is titled "AWS Glue" and includes sections for "Getting started", "Data Catalog", "Data Integration and ETL", and "Legacy pages". Under "Data Catalog", there are links for "Tables", "Stream schema registries", "Schemas", "Connections", "Crawlers", "Classifiers", and "Catalog settings". The main content area is titled "Databases (1/2)" and contains a table with one row. The table has columns: Name, Description, Location URI, and Created on (UTC). The single row shows "annual_reports" with a Location URI of "s3://aws-glue-demo-series/data-store/annual/" and a Created on (UTC) of "March 31, 2024 at 09:34:44". The status bar at the bottom right indicates the date as "04-04-2024" and the time as "11:46 AM".

Name	Description	Location URI	Created on (UTC)
annual_reports	-	s3://aws-glue-demo-series/data-store/annual/	March 31, 2024 at 09:34:44

step-4 create and run glue crawl,here the crawler name is “annual-reports-crawl”

The screenshot shows the AWS Glue Console interface. The left sidebar is titled "AWS Glue" and includes sections for ETL jobs, Data Catalog tables, Data connections, Workflows (orchestration), Data Catalog, Databases, Tables, Stream schema registries, Schemas, Connections, Crawlers, Classifiers, Catalog settings, Data Integration and ETL, and Legacy pages. The "Crawlers" section is currently selected. The main content area is titled "Crawlers" and contains a table with the following data:

Name	State	Schedule	Last run	Last run timestamp	Log	Table changes from ...
annual-reports-crawl	Ready		Succeeded	March 31, 2024 at 09...	View log	1 created
last-crawler	Ready		Succeeded	March 29, 2024 at 11...	View log	1 created

At the top right of the table, there are buttons for "Action", "Run", and "Create crawler". The status bar at the bottom indicates "Activate Windows Go to Settings to activate Windows." and shows the date and time as "© 2024, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences 11:46 AM 04-04-2024".

Glue crawler is run successfully

The screenshot shows the AWS Glue Console interface. The left sidebar navigation includes 'AWS Glue' (selected), 'Jobs - AWS Glue Studio', 'Crawlers - AWS Glue Console', 'Data Catalog', 'Data Integration and ETL', and 'Legacy pages'. The main content area displays the 'annual-reports-crawl' crawler properties and its run history.

Crawler properties:

- Name: annual-reports-crawl
- IAM role: aws-glue-service-role-deme
- Database: annual_reports
- Description: -
- Security configuration: -
- Lake Formation configuration: -
- Table prefix: -
- State: READY
- Maximum table threshold: -

Crawler runs (1):

Start time (UTC)	End time (UTC)	Current/last duration	Status	DPU hours	Table changes
March 31, 2024 at 09:49:12	March 31, 2024 at 09:50:13	01 min	Completed	0.052	1 table change, 0 partition changes

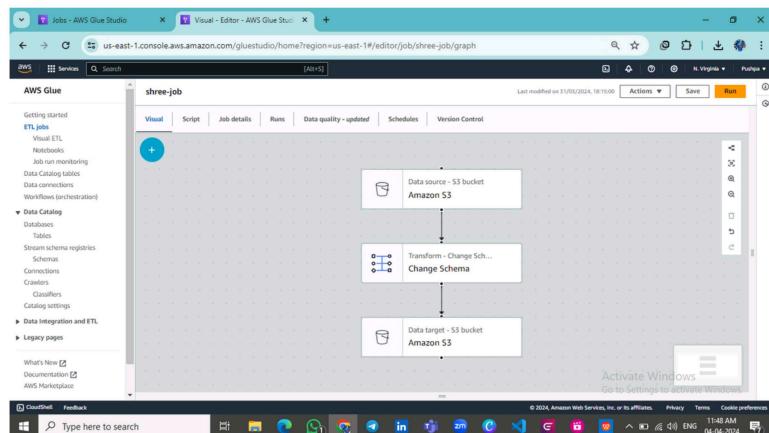
Below the table, there are buttons for 'Stop run', 'View CloudWatch logs', and 'View run details'. The status bar at the bottom indicates 'Activate Windows' and 'Go to Settings to activate Windows.'

step-5 creating glue jobs,here the job name is “ shree-job”

The screenshot shows the AWS Glue Studio interface for managing jobs. On the left, a sidebar lists various AWS Glue services: Getting started, ETL jobs (selected), Visual ETL, Notebooks, Job run monitoring, Data Catalog tables, Data connections, Workflows (orchestration), Data Catalog (selected), Databases, Tables, Stream schema registries, Schemas, Connections, Crawlers, Classifiers, Catalog settings, Data Integration and ETL, and Legacy pages. A 'What's New' section is also present. The main content area is titled 'AWS Glue Studio' and shows the 'Jobs' tab. It includes sections for 'Create job' (with options for Visual ETL, Notebook, and Script editor), 'Example jobs' (with a 'Create example job' button), and 'Your jobs (1)'. A table lists the single job: shree-job, which is a Glue ETL job last modified on 31/03/2024 at 18:15:00, using AWS Glue version 3.0. The bottom of the screen shows the Windows taskbar with various pinned icons.

Job name	Type	Last modified	AWS Glue version
shree-job	Glue ETL	31/03/2024, 18:15:00	3.0

visual of the job,"shree-job



-In this i selected s3 bucket as source in which the csv file is located

-And then i selected “change schema” to modify some values and column names

-And lastly i choose “s3 bucket ‘ as target to store the out put(i.e modified csv file)

The job i created “shree-job” has run successfully

Runs - Editor - AWS Glue Studio | us-east-1.console.aws.amazon.com/gluestudio/home?region=us-east-1#/editor/job/shree-job/runs

AWS Glue | shree-job | Last modified on 01/03/2024, 18:15:00 | Actions | Save | Run

Getting started | ETL jobs | Data Catalog | Data Integration and ETL | Legacy pages | What's New | Documentation | AWS Marketplace

Visual | Script | Job details | **Runs** | Data quality - updated | Schedules | Version Control

Job runs (1/2) | Info | Filter job runs by property

Run status	Retries	Start time (Local)	End time (Local)	Duration	Capacity (DPU)	Worker type	Glue version
Succeeded	0	04/04/2024 12:35:21	04/04/2024 12:36:38	1m 9s	10 DPU	G.1X	3.0
Succeeded	0	03/31/2024 18:15:06	03/31/2024 18:16:34	1m 20s	10 DPU	G.1X	3.0

Run details | Input arguments (10) | Continuous logs | Run Insights | Metrics | Spark UI

Job name: shree-job | Start time (Local): 04/04/2024 12:35:21 | Glue version: 3.0 | Last modified on (Local): 04/04/2024 12:36:38

Id: jr_81d864a4972784b3375a5b4d86bfd47f36ec231 | End time (Local): 04/04/2024 12:36:38 | Worker type: G.1X | Log group name: /aws-glue/jobs

Run status: Succeeded | Start-up time: 8 seconds | Max capacity: 10 DPU

CloudShell | Feedback | Type here to search | © 2024, Amazon Web Services, Inc. or its affiliates. | Privacy | Terms | Cookie preferences | 12:37 PM | ENG | 04-04-2024 | 10 Unsaved Job found. We found an unrun job. Do you wish to generate code? | Restore

The output file is successfully uploaded in the target s3 bucket folder

The screenshot shows the AWS S3 console interface. The left sidebar is titled 'Amazon S3' and includes sections for Buckets, Access Grants, Access Points, Object Lambda Access Points, Multi-Region Access Points, Batch Operations, IAM Access Analyzer for S3, Block Public Access settings for this account, Storage Lens (Dashboards, Storage Lens groups, AWS Organizations settings), Feature spotlight, and AWS Marketplace for S3.

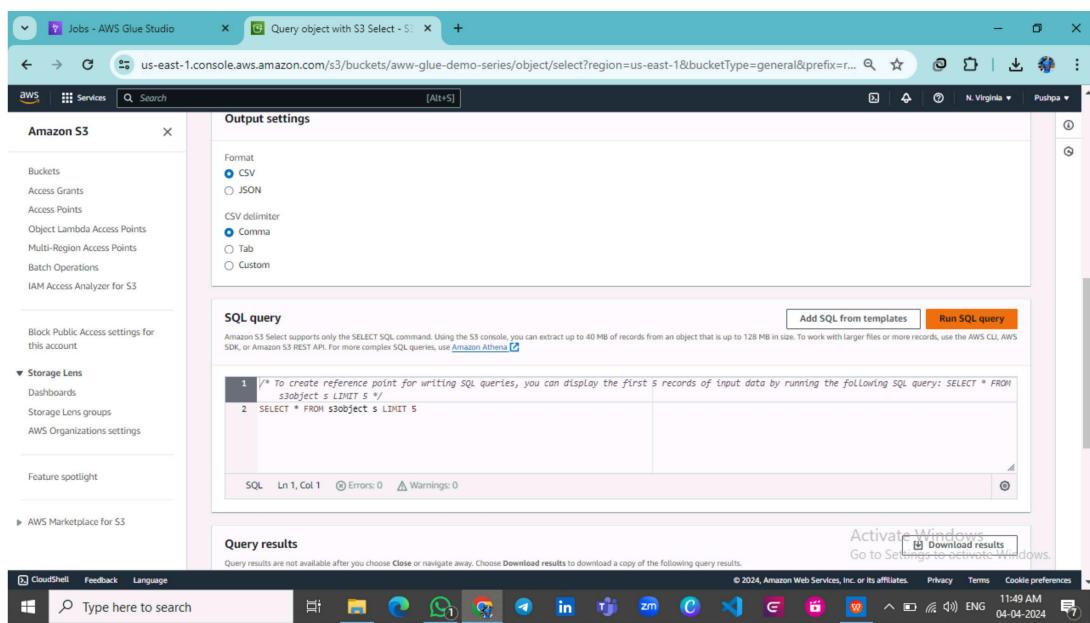
The main content area shows the 'Buckets' section for 'aww-glue-demo-series'. The 'Objects' tab is selected, displaying two items:

Name	Type	Last modified	Size
data-store/	Folder	-	-
run-1711889163953-part-r-00000	-	March 31, 2024, 18:16:23 (UTC+05:30)	-

Actions available for the selected object include: Copy S3 URI, Copy URL, Download, Open, Delete, Create folder, Upload, Share with a presigned URL, Calculate total size, Copy, Move, Initiate restore, Query with S3 Select, Edit actions, Rename object, Edit storage class, Edit server-side encryption, Edit metadata, Edit tags, and Activate Windows (with a note to go to Settings to activate Windows).

The status bar at the bottom indicates: © 2024, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences. 11:49 AM ENG 04-04-2024.

Running the sql query to get the output



Here the out put is produced with the modified values and column names

The screenshot shows the AWS Glue Studio interface with a query results page. The left sidebar is titled 'Amazon S3' and includes sections for Buckets, Access Grants, Access Points, Object Lambda Access Points, Multi-Region Access Points, Batch Operations, and IAM Access Analyzer for S3. Below these are sections for Storage Lens (Dashboards, Storage Lens groups, AWS Organizations settings) and Feature spotlight. At the bottom of the sidebar, there's a link to 'AWS Marketplace for S3'. The main content area is titled 'Query results' and displays a table of data. The table has the following columns: did, gender, scitizen, partner, dependents, tenure, phoneservice, multiplelines, internetservice, onlinesecurity, and onlinebackup. The data rows are:

did	gender	scitizen	partner	dependents	tenure	phoneservice	multiplelines	internetservice	onlinesecurity	onlinebackup
7590-VHVEG	Female	0	Yes	No	1	No	No phone service	DSL	No	Yes
5575-GNVEDE	Male	0	No	No	34	Yes	No	DSL	Yes	No
3668-QPYBK	Male	0	No	No	2	Yes	No	DSL	Yes	Yes
7795-CFOCW	Male	0	No	No	45	No	No phone service	DSL	Yes	No

Below the table, there are navigation arrows and a 'Download results' button. The status bar at the bottom right shows 'Activate Windows' and the date '04-04-2024'.