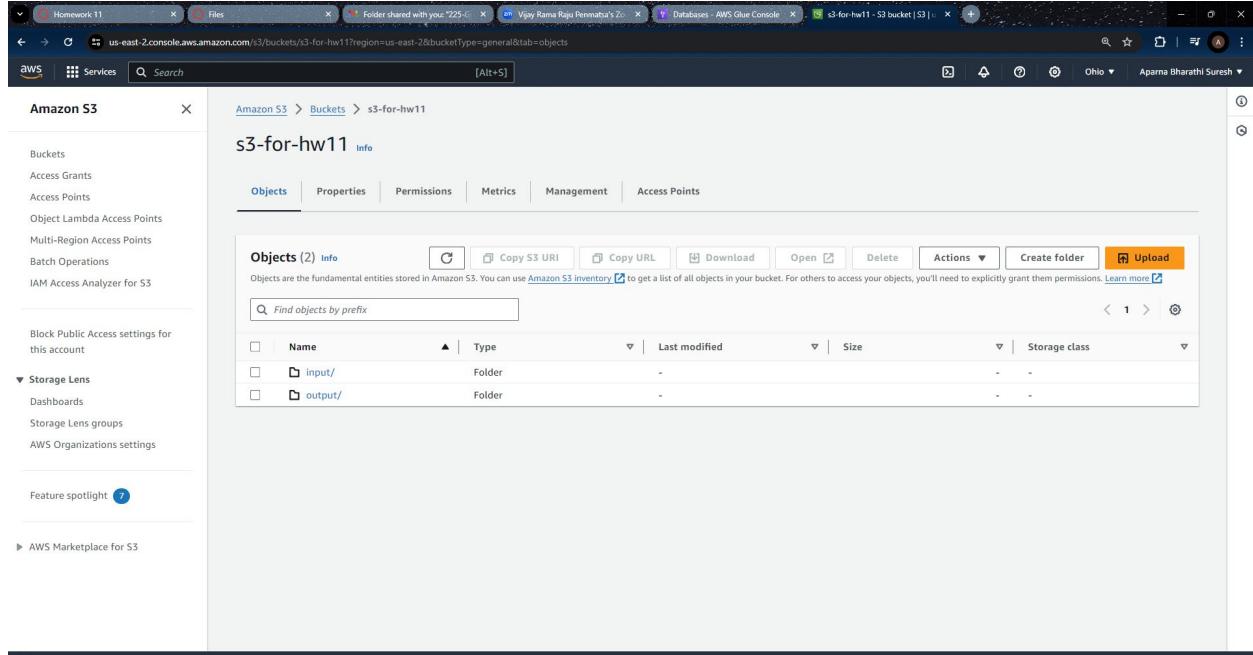


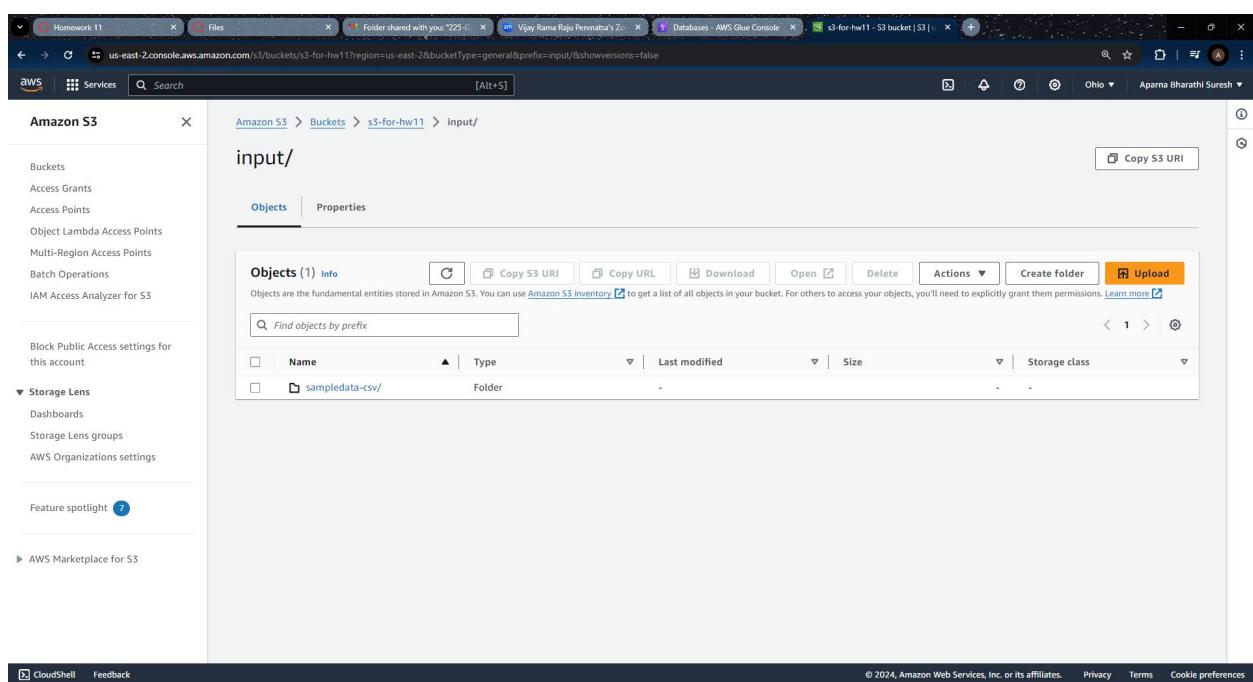
# Homework11\_AparnaBharathiSuresh

1. Load sampledata.csv Download sampledata.csv in an S3 bucket.



The screenshot shows the AWS S3 console interface. The left sidebar is collapsed, and the main area displays the 's3-for-hw11' bucket. The 'Objects' tab is selected, showing two folders: 'input/' and 'output/'. The 'input/' folder contains one object named 'sampledata-csv/'. A 'Copy S3 URI' button is visible above the object list.

Name	Type	Last modified	Size	Storage class
sampledata-csv/	Folder	-	-	-

This screenshot shows the contents of the 'input/' folder within the 's3-for-hw11' bucket. It contains a single object named 'sampledata-csv/'. The 'Properties' tab is selected, and a 'Copy S3 URI' button is visible above the object list.

Name	Type	Last modified	Size	Storage class
sampledata-csv/	Folder	-	-	-

Files Homework 11 Vijay Rama Raju Penmata's Zone Upload objects - S3 bucket s3-for-hw11

aws Services Search [Alt+S]

Upload succeeded View details below.

### Upload: status

The information below will no longer be available after you navigate away from this page.

#### Summary

Destination	Succeeded	Failed
s3://s3-for-hw11/input/sampleddata-csv/	1 file, 973.0 B (100.00%)	0 files, 0 B (0%)

Files and folders Configuration

#### Files and folders (1 Total, 973.0 B)

Name	Folder	Type	Size	Status	Error
sampleddata ...	-	text/csv	973.0 B	Succeeded	-

CloudShell Feedback © 2024, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences

This screenshot shows the AWS S3 upload confirmation page. It displays a green success message: "Upload succeeded" with a link to "View details below.". Below this, there's a summary table showing the destination bucket and the status of the upload. A main table lists the uploaded file, detailing its name, type (text/csv), size (973.0 B), status (Succeeded), and error (none). The interface includes tabs for "Files and folders" and "Configuration". At the bottom, there are links for CloudShell, Feedback, and copyright information.

Homework 11 Files Folder shared with you \*225 Vijay Rama Raju Penmata's Zone Databases - AWS Glue Console s3-for-hw11 - S3 bucket [S3]

aws Services Search [Alt+S]

Amazon S3 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 7

AWS Marketplace for S3

Amazon S3 > Buckets > s3-for-hw11 > input/ > sampleddata-csv/

sampleddata-csv/

Objects (1) Info Copy S3 URI Copy URL Download Open Delete Actions Create folder Upload

Objects are the fundamental entities stored in Amazon S3. You can use [Amazon S3 inventory](#) to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. [Learn more](#)

Name	Type	Last modified	Size	Storage class
sampleddata.csv	CSV	May 1, 2024, 10:30:53 (UTC-07:00)	973.0 B	Standard

CloudShell Feedback © 2024, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences

This screenshot shows the AWS S3 bucket contents page for the "sampleddata-csv/" folder. It displays a single object named "sampleddata.csv" with the following details: Type: CSV, Last modified: May 1, 2024, 10:30:53 (UTC-07:00), Size: 973.0 B, Storage class: Standard. The page includes a sidebar with various AWS services like Buckets, Access Grants, and Storage Lens. At the bottom, there are links for CloudShell, Feedback, and copyright information.

## 2. Using AWS Glue, create a crawler to import the data from S3 to your database tables.

The screenshot shows the AWS Glue Data Catalog Databases page. The left sidebar includes sections for Getting started, ETL jobs, Data Catalog (Tables, Stream schema registries, Schemas, Connections, Crawlers, Classifiers, Catalog settings), Data Integration and ETL, Legacy pages, What's New, Documentation, CloudShell, and Feedback. The main content area displays a table titled "Databases (2)". The table has columns for Name, Description, Location URI, and Created on (UTC). The entries are:

Name	Description	Location URI	Created on (UTC)
input	-	-	May 1, 2024 at 00:27:16
output	-	-	May 1, 2024 at 00:27:25

The screenshot shows the AWS IAM Roles page. The left sidebar includes sections for Identity and Access Management (IAM) (Dashboard, Access management, User groups, Users, Roles, Policies, Identity providers, Account settings), Access reports (Access Analyzer, External access, Unused access, Analyzer settings), Credential report, and Organization activity. The main content area shows a success message: "Role role\_hw\_11 created." Below it is a table titled "Roles (6) Info". The table has columns for Role name, Trusted entities, and Last activity. The entries are:

Role name	Trusted entities	Last activity
AWSServiceRoleForRDS	AWS Service: rds (Service-Linked Role)	1 hour ago
AWSServiceRoleForRedshift	AWS Service: redshift (Service-Linked Role)	7 days ago
AWSServiceRoleForSupport	AWS Service: support (Service-Linked Role)	-
AWSServiceRoleForTrustedAdvisor	AWS Service: trustedadvisor (Service-Linked Role)	-
rds-monitoring-role	AWS Service: monitoring.rds	-
role_hw_11	AWS Service: glue	-

**AWS Glue**

**Review and create**

**Step 1: Set crawler properties**

Name: glue-for-hw11

**Step 2: Choose data sources and classifiers**

Type: S3  
Data source: s3://s3-for-hw11/input/sampledata-csv/sampledata.csv  
Parameters: Recrawl all

**Step 3: Configure security settings**

IAM role: role-for-hw11

**Step 4: Set output and scheduling**

Database input: Table prefix - optional: -  
Maximum table threshold - optional: -  
Schedule: On demand

**Create crawler**

**Crawler successfully starting**  
The following crawler is now starting: "glue-for-hw11"

**glue-for-hw11**

**Crawler properties**

Name: glue-for-hw11	IAM role: role-for-hw11	Database input: -	State: READY
Description: -	Security configuration: -	Lake Formation configuration: -	Table prefix: -
Maximum table threshold: -			

**Crawler runs (1)**

Start time (UTC)	End time (UTC)	Current/last duration	Status	DPU hours	Table changes
May 1, 2024 at 17:45:09	May 1, 2024 at 17:46:08	58 s	Completed	-	-

Screenshot of the AWS Glue Console showing the 'input' database properties. The database has a single table named 'sampledata\_csv'.

**Database properties**

Name	Description	Location	Created on (UTC)
input	-	-	May 1, 2024 at 17:34:57

**Tables (1)**

Name	Database	Location	Classification	Deprecated	View data	Data quality
sampledata_csv	input	s3://s3-for-hw11/input/sample	CSV	-	Table data	View data quality

Screenshot of the AWS Glue Table Detail view for 'sampledata\_csv'.

**Crawler successfully starting**  
The following crawler is now starting: "glue-for-hw11"

**sampledata\_csv**

**Table overview**

Name	Description	Database	Classification
sampledata_csv	-	input	CSV

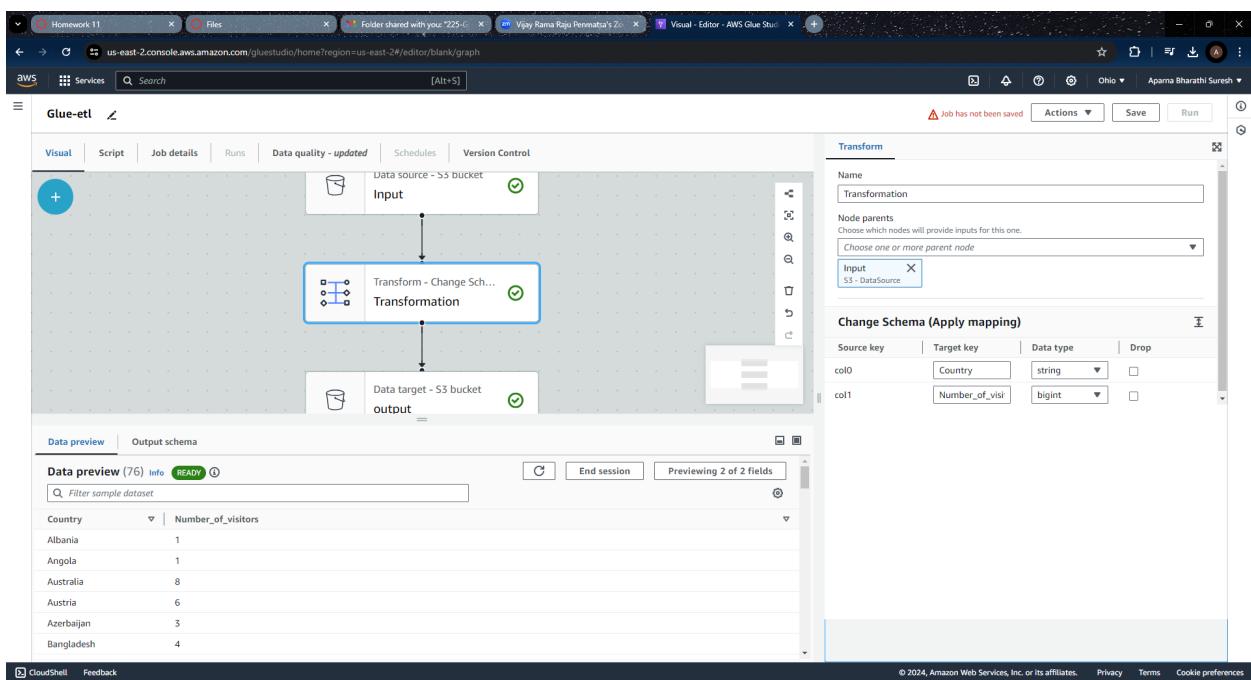
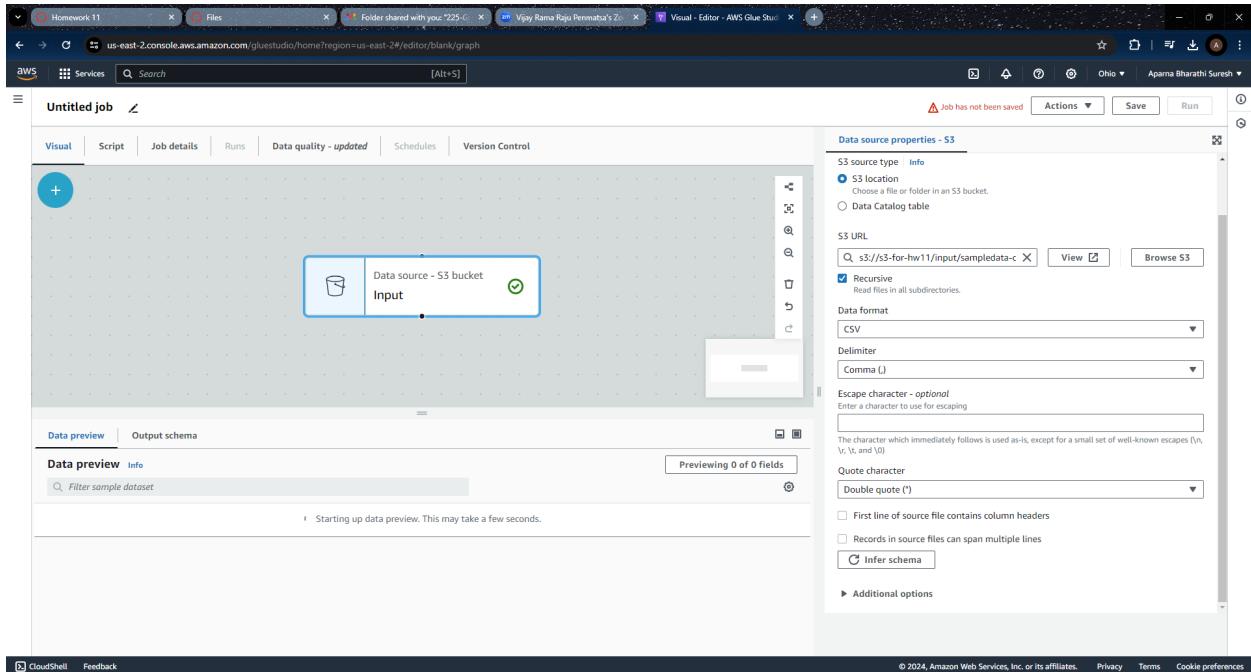
**Table details**

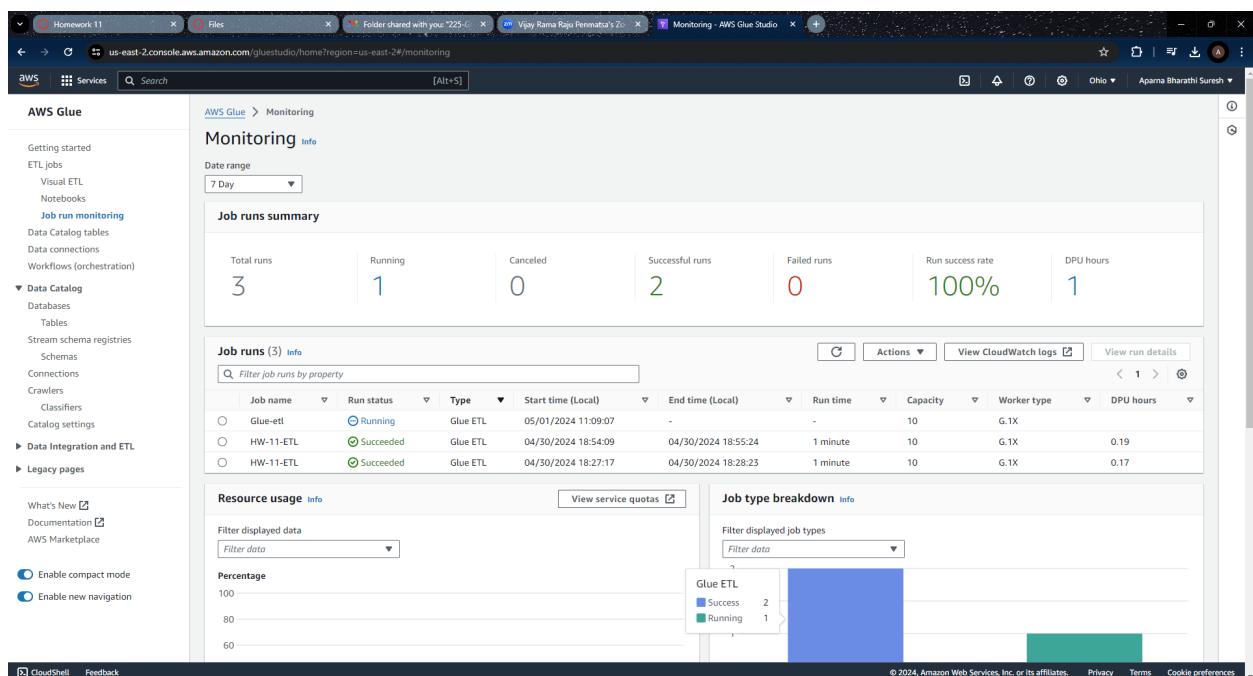
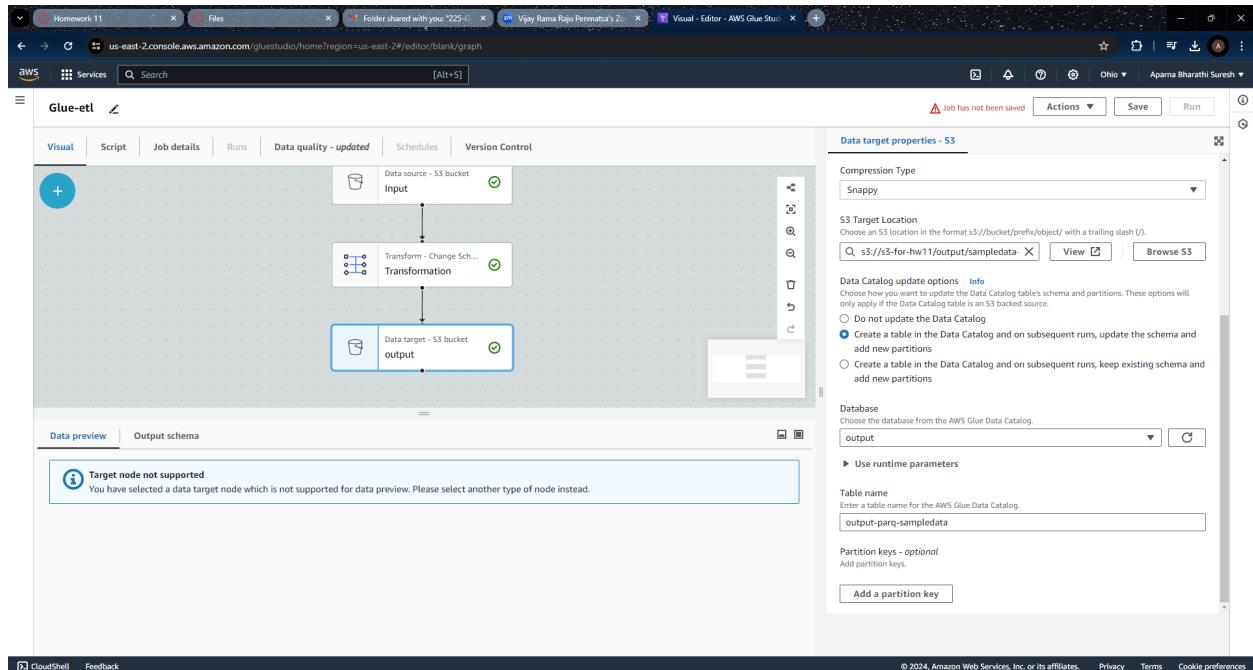
Name	Location	Connection	Deprecated	Last updated
sampledata_csv	s3://s3-for-hw11/input/sampledata-csv/sampledata.csv	-	-	May 1, 2024 at 17:46:08

**Schema**

#	Column name	Data type	Partition key	Comment
1	col0	string	-	-
2	col1	bigint	-	-

### 3. Get the output file in an S3 bucket in parquet format by Glue Job.





Screenshot of AWS Glue Studio Monitoring page:

**Job runs summary**

Total runs	Running	Canceled	Successful runs	Failed runs	Run success rate	DPU hours
3	0	0	3	0	100%	1

**Job runs (3) Info**

Job name	Run status	Type	Start time (Local)	End time (Local)	Run time	Capacity	Worker type	DPU hours
Glue-etl	Succeeded	Glue ETL	05/01/2024 11:09:07	05/01/2024 11:10:05	1 minute	10	G.1X	0.17
HW-11-ETL	Succeeded	Glue ETL	04/30/2024 18:54:09	04/30/2024 18:55:24	1 minute	10	G.1X	0.19
HW-11-ETL	Succeeded	Glue ETL	04/30/2024 18:27:17	04/30/2024 18:28:23	1 minute	10	G.1X	0.17

**Resource usage**

Filter displayed data: Percentage

Percentage: 100, 80, 60

**Job type breakdown**

© 2024, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences

Screenshot of AWS Glue Data Catalog Databases page:

**output**

**Database properties**

Name	Description	Location	Created on (UTC)
output	-	-	May 1, 2024 at 17:35:02

**Tables (1)**

Last updated (UTC): May 1, 2024 at 18:46:26

Name	Database	Location	Classification	Deprecated	View data	Data quality
output-parq-sampled	output	s3://s3-for-hw11/output/s/	Parquet	-	Table data	View data quality

© 2024, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences

The screenshot shows the AWS S3 console interface. On the left, there is a sidebar with various navigation links such as Buckets, Access Grants, 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 displays a list of objects in a bucket named 's3-for-hw11'. The path in the top navigation bar is 'Amazon S3 > Buckets > s3-for-hw11 > output/ > sampledata-parq/'. The list shows one object: 'run-1714586977045-part-block-0-r-00000-snappy.parquet' which is a Parquet file. The table includes columns for Name, Type, Last modified, Size, and Storage class. At the bottom of the page, there are links for CloudShell, Feedback, Privacy, Terms, and Cookie preferences.

Name	Type	Last modified	Size	Storage class
run-1714586977045-part-block-0-r-00000-snappy.parquet	parquet	May 1, 2024, 11:09:54 (UTC-07:00)	1.3 KB	Standard

#### 4. Load the output table to Athena and run 1 query using group by.

The screenshot shows the AWS Athena Query Editor interface. A message at the top indicates "Workgroup query engine upgrade complete". The main area displays a single query named "Query 1" with the following SQL statement:

```
1 SELECT * FROM "output"."output-parq-sampleddata";
```

The results section shows 76 rows of data. The columns are labeled "#", "country", and "number\_of\_visitors". The data is as follows:

#	country	number_of_visitors
1	Albania	1
2	Angola	1
3	Australia	8
4	Austria	6

This screenshot shows the same AWS Athena Query Editor interface as the previous one, but the results section is now expanded to show the full dataset. The columns "#", "country", and "number\_of\_visitors" are visible, and the data rows are identical to those shown in the first screenshot.

```

SELECT country, AVG(number_of_visitors) AS average_visitors
FROM "output"."output-parq-sampleddata"
GROUP BY country ORDER BY average_visitors DESC LIMIT 10;

```

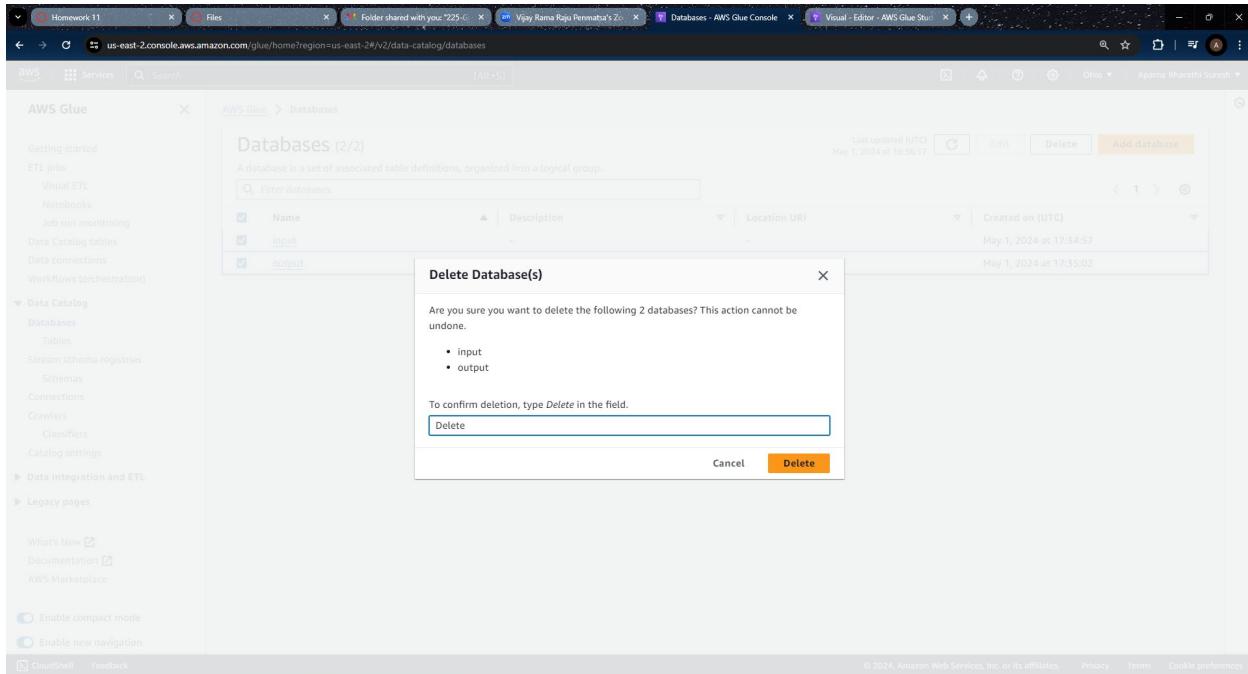
The screenshot shows the AWS Athena Query Editor interface. The query has been executed and completed. The results table displays 10 rows, each containing a country name and its average visitors. The data is as follows:

#	country	average_visitors
1	Switzerland	10.0
2	Cote d'Ivoire	10.0
3	Costa Rica	10.0
4	Libya	10.0
5	Sao Tome and Principe	9.0
6	Sierra Leone	9.0
7	Macedonia	9.0
8	Nicaragua	9.0
9	Niger	9.0
10	Laos	9.0

The screenshot shows the AWS Athena Query Editor interface. The query has been executed and completed. The results table displays 10 rows, each containing a country name and its average visitors. The data is as follows:

#	country	average_visitors
1	Switzerland	10.0
2	Cote d'Ivoire	10.0
3	Costa Rica	10.0
4	Libya	10.0
5	Sao Tome and Principe	9.0
6	Sierra Leone	9.0
7	Macedonia	9.0
8	Nicaragua	9.0
9	Niger	9.0
10	Laos	9.0

## 5. Delete the Glue jobs, IAM role, data in S3, and Crawler



Databases (2/2)

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

Name	Description	Location URI	Created on (UTC)
input			May 1, 2024 at 17:34:57
output			May 1, 2024 at 17:35:02

Delete Database(s)

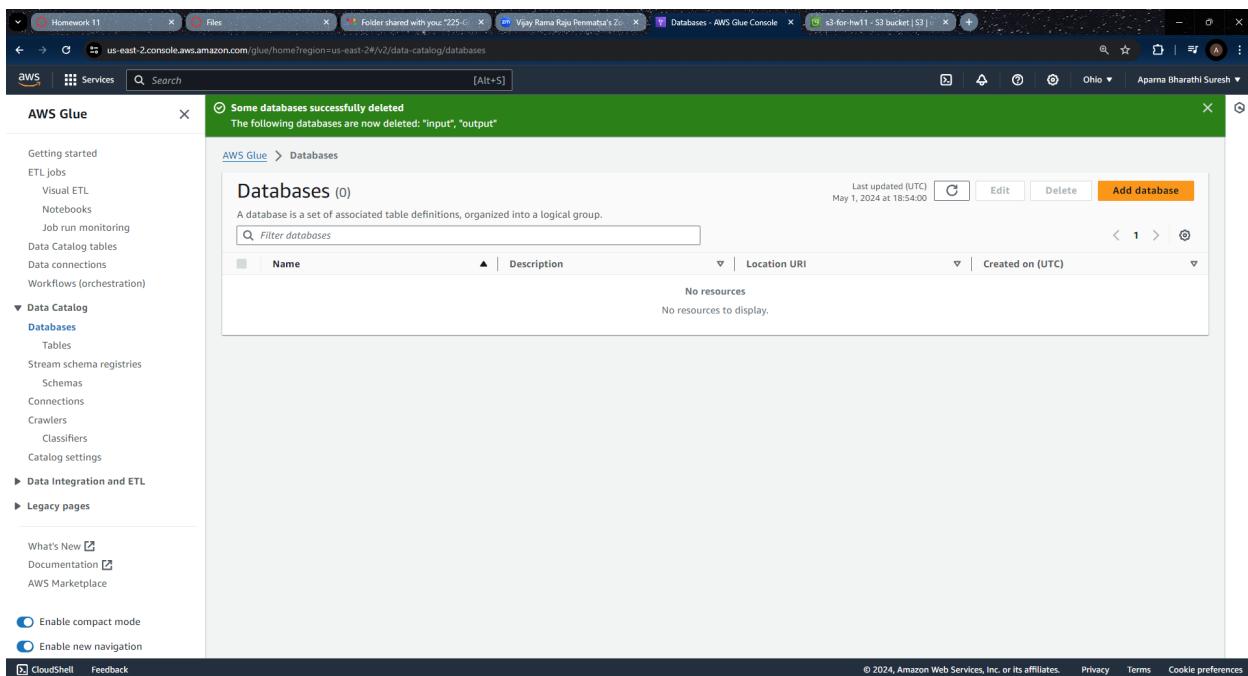
Are you sure you want to delete the following 2 databases? This action cannot be undone.

- input
- output

To confirm deletion, type **Delete** in the field.

Delete

Cancel Delete



Some databases successfully deleted

The following databases are now deleted: "input", "output"

Databases (0)

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

Name	Description	Location URI	Created on (UTC)
No resources			
No resources to display.			

The screenshot shows the AWS Glue Studio interface. On the left, there's a sidebar with various navigation options like 'Getting started', 'ETL jobs', 'Visual ETL', 'Notebooks', etc. The main area is titled 'AWS Glue Studio' and has a 'Create job' section with two options: 'Visual KPL' (selected) and 'Script editor'. Below this is a 'Example jobs' section and a 'Your jobs (1)' section where 'Glue-etl' is listed. A modal window titled 'Delete job(s)' is open, asking if you want to delete 'Glue-etl'. It includes 'Cancel' and 'Delete' buttons.

The screenshot shows the AWS S3 console. The user is in the 'Buckets' section, specifically under 's3-for-hw-11 > Delete bucket'. A modal window titled 'Delete bucket' is open, containing a warning message about the不可逆性 of deleting buckets. It asks the user to confirm the deletion by entering the bucket name 's3-for-hw-11' into a text input field. The modal has 'Cancel' and 'Delete bucket' buttons.

The screenshot shows the AWS Glue console interface. On the left, there's a navigation sidebar with sections like 'Getting started', 'ETL jobs', 'Visual ETL', 'Notebooks', 'Job run monitoring', 'Data Catalog tables', 'Data connections', 'Workflows (orchestration)', 'Data Catalog', 'Databases', 'Tables', 'Stream schema registries', 'Schemas', 'Connections', 'Crawlers', 'Classifiers', and 'Catalog settings'. Under 'Crawlers', a sub-section 'Crawlers' is selected. The main content area is titled 'Crawlers' and contains a table with one row. The table has columns for 'Name' (glue-for-hw11), 'State' (Ready), 'Schedule' (None), 'Last run' (May 1, 2024 at 17:45:09), 'Last run timestamp' (May 1, 2024 at 17:45:09), 'Log' (View log), and 'Table changes from last run' (1 created). At the top of the table, there are buttons for 'Action', 'Run', and 'Create crawler'. A modal dialog box is overlaid on the page, titled 'Delete crawler'. It contains the message 'Are you sure you want to delete the following crawler? This action cannot be undone.' followed by a list item 'glue-for-hw11'. At the bottom of the modal are 'Cancel' and 'Delete' buttons, with 'Delete' being the active one.

This screenshot shows the same AWS Glue console interface as the previous one, but with a green banner at the top stating 'One crawler successfully deleted' and 'The following crawler is now deleted: "glue-for-hw11"'. The main content area is identical to the first screenshot, showing the 'Crawlers' table with one row. The table has columns for 'Name' (glue-for-hw11), 'State' (Ready), 'Schedule' (None), 'Last run' (May 1, 2024 at 21:23:23), 'Last run timestamp' (May 1, 2024 at 21:23:23), 'Log' (View log), and 'Table changes from last run' (No resources to display). The 'Delete' button in the modal dialog is no longer active.

The screenshot shows the AWS IAM Roles page. On the left, the navigation menu includes sections like Identity and Access Management (IAM), Access management, User groups, Users, Roles, Policies, Identity providers, Account settings, Access reports, Access Analyzer, External access, Unified access, Analyzer settings, Confidential report, Organization activity, Service control policies, and Related consoles. The 'Roles' section is selected.

The main content area displays a list of roles with one selected: 'role-for-hw11'. A modal dialog titled 'Delete role-for-hw11?' is open, asking if the user wants to permanently delete the role. It states: 'Delete role-for-hw11 permanently? This will also delete all its inline policies and any attached instance profiles.' Below this, it shows the role name 'role-for-hw11' and last activity '2 hours ago'. A note at the bottom says: 'Note: Recent activity usually appears within 4 hours. Data is stored for a maximum of 365 days, depending when your region began supporting this feature. Learn more'.

The modal has a text input field containing 'role-for-hw11' for confirmation. At the bottom right are 'Cancel' and 'Delete' buttons. To the right of the modal, there's a 'Temporary credentials' section with a link to 'Manage'.