

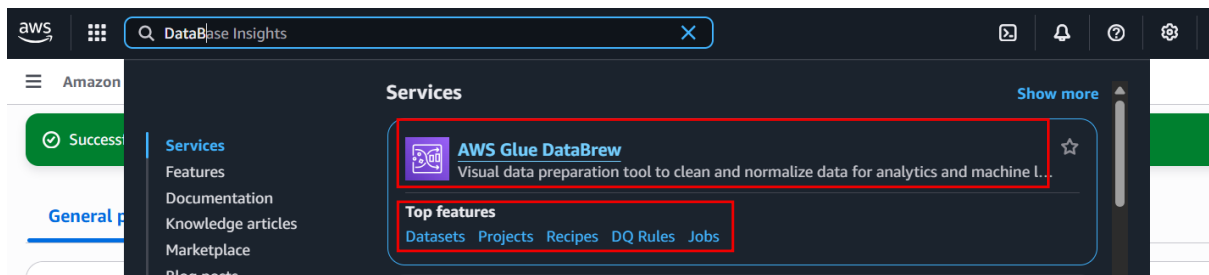
AWS Glue DataBrew

To Begin with the Lab

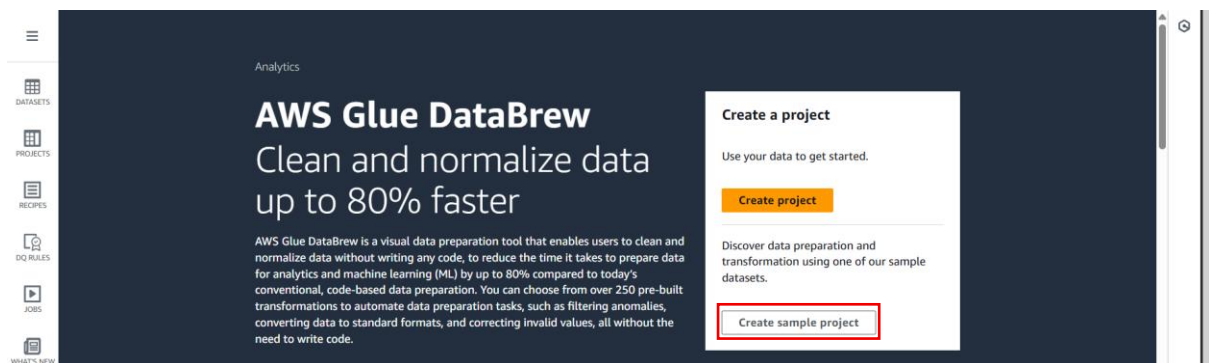
Summary of the Lab

In this lab, you explore **AWS Glue DataBrew** to visually clean, transform, and prepare data for analytics or machine learning. You start by creating a **sample project** using a built-in dataset (e.g., Chess game moves) and assigning an IAM role for S3 access. Once the dataset loads, you can interactively view data, analyze column details, and apply transformations like converting text to uppercase or deleting columns. Each transformation is saved as a **recipe step**. Finally, you create and run a **DataBrew job**, specifying output details such as S3 location, file format, and compression, to generate the processed dataset.

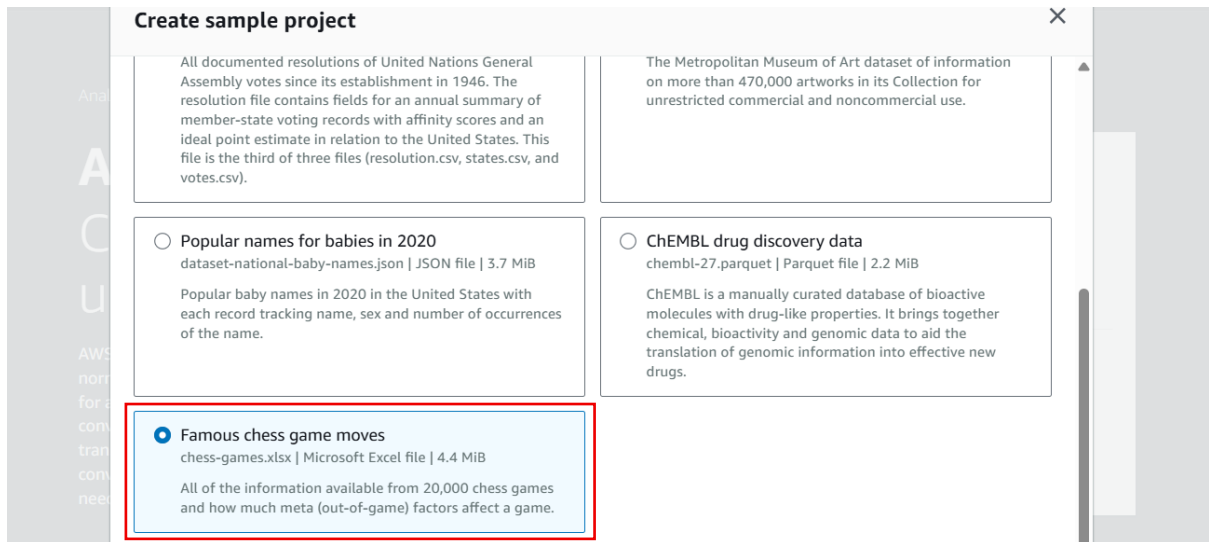
- Sign in to the **AWS Management Console**.
- In the search bar, type **DataBrew** and select **AWS Glue DataBrew**.
- You'll see options like **Projects**, **Datasets**, **Recipes**, and **Jobs**.



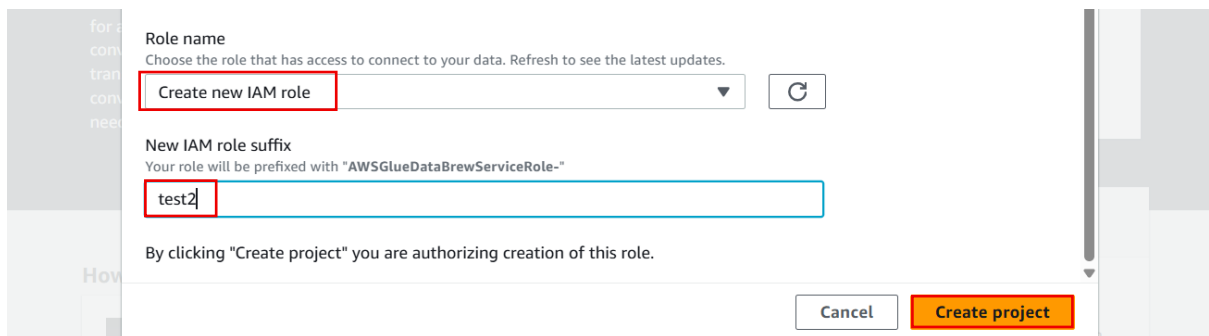
- Click on **Create sample project**.



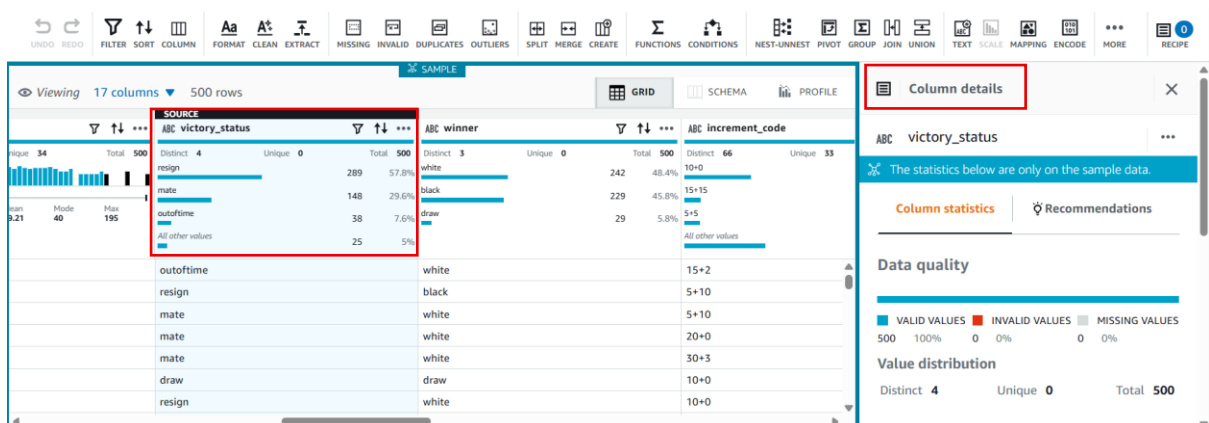
- Choose a **sample dataset** or upload your own.
 - Example: Select the *“Chess game moves”* dataset.



- Under **IAM Role**, select an existing role or click **Create new role**.
 - DataBrew will automatically assign permissions to access S3.
- Click **Create project**.
- Wait for a minute while the interactive session and dataset load.

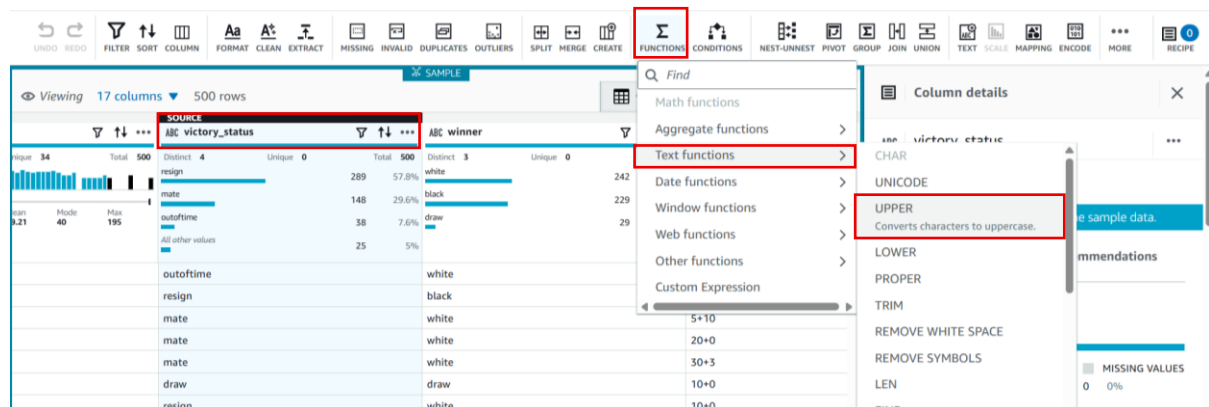


- View the dataset preview in the **Data grid** area.
- When you select the column, you can check the **Column details** section to see:
 - Data types
- Optionally, click **Profile** to generate a **data profile report** (shows distributions and quality).



- Select a **column** from your dataset.

- Choose a transformation from the top menu, such as:
- **Text functions** → convert to **uppercase**.
- Click **Preview changes** to see results before applying.



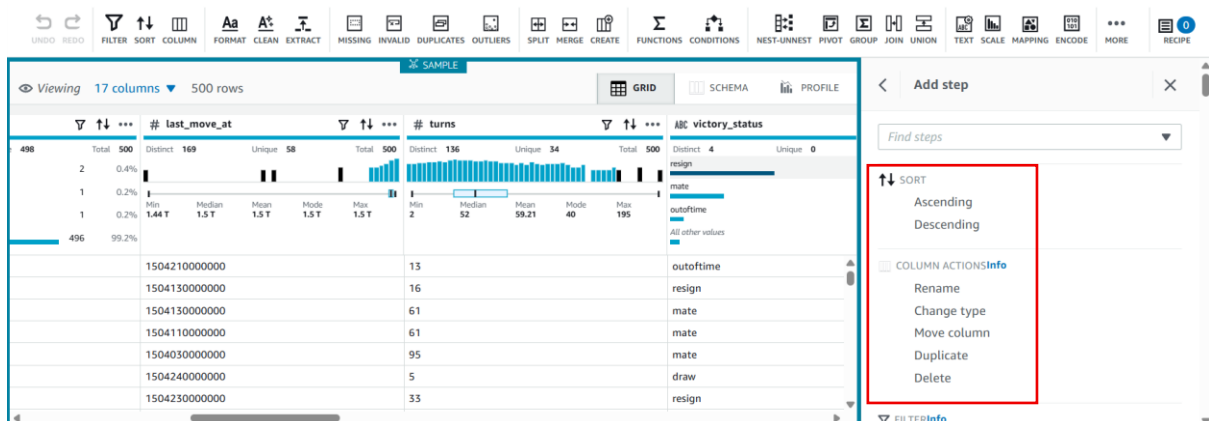
- If you click on the SCHEMA, you get an overview of all the columns.

Show/Hide	Column name	Data type	Data quality	Value distribution	Box plot
<input type="checkbox"/>	id	ABC string	100% Valid 0% Invalid	Distinct 500 Unique 500 Total 500	
<input type="checkbox"/>	rated	boolean	100% Valid 0% Invalid	Distinct 2 Unique 0 Total 500	
<input type="checkbox"/>	created_at	# long	100% Valid 0% Invalid	Distinct 172 Unique 61 Total 500	Min 1.44 T Median 1.5 T Mean 1.5 T Max 1.5 T
<input type="checkbox"/>	creation_date	timestamp	100% Valid 0% Invalid	Distinct 499 Unique 498 Total 500	
<input type="checkbox"/>	last_move_at	# long	100% Valid 0% Invalid		Min Median Mean Max

- When you click on the Add Steps under Recipe

id	rated	created_at	creation_date	last_move_at
1504210000000	13	outoftime		
1504130000000	16	resign		
1504130000000	61	mate		
1504110000000	61	mate		
1504030000000	95	mate		
1504240000000	5	draw		

- You can perform a lot of action such as delete, move columns etc.



- Let's try to delete and preview the column.

- Here the preview is shown for the ‘created_at’ column.
- After click on the Apply.

The screenshot shows a data table with columns: # created_at, creation_date, and # last_move_at. The 'Delete column' dialog box is open on the right, showing the 'Delete column info' section with the source column name '# created_at' selected. The 'Preview shown' section is empty. The 'Apply' button is highlighted in orange.

- All transformations are saved as steps in **Recipe**.

The screenshot shows the same data table as before, but with the 'Recipe' panel open on the right. The 'Recipe' panel shows a list of applied steps, with the first step being '1. Delete column created_at'. The 'Delete' button for this step is highlighted in orange.

- In your project, click **Create job**.

The screenshot shows a project page titled 'Sample project - 2'. The 'Create job' button is highlighted with a red box. The 'Dataset: chess-games' and 'Sample: First n sample (500 rows)' are also visible.

- Enter a **Job name** (e.g., ml-prep-job1).

DataBrew > Jobs > Create job

Create job Info


Job details

Job name
Identifier for the job

ml-prep-job1

The job name must contain 1-240 characters. Valid characters are alphanumeric (A-Z, a-z, 0-9), hyphen (-), period (.), and space.

Job type


Recipe job
A recipe job runs the transformation from the associated recipe on the population of the associated dataset.

Associated dataset
[chess-games](#)
S3 | s3://databrew-public-datasets-us-east-1/chess-games.xlsx


Associated recipe
[Sample recipe - 2](#)
Working version

- Configure the **Output settings**:
 - Choose an **S3 bucket** to store the processed output.
 - Select a **file format** (e.g., CSV or Parquet).
 - Add compression if needed.
- After all the configurations, you can click on Create and run job

Job output settings Info

Running a job generates output files at specified file destinations.

Output 1

Output to Output location	File type Output format	Delimiter CSV separator	Compression Available types
 Amazon S3	CSV	Comma (,)	None

S3 bucket owner's AWS account

☒ **Current AWS account**
463646775279

☐ Another AWS account

S3 location
Format is: s3://bucket/folder/

Browse

Add another output

► Additional configuration - optional

Setting summary

File output storage
Create a new folder

File output
Autogenerate files

Custom partition by
Disabled