



AWS Glue DataBrew

AWS Glue DataBrew is a visual data preparation tool that helps users clean and normalize data without writing any code. It simplifies the data preparation process, enabling data analysts, data scientists, and data engineers to work with raw data directly in a visual interface. Here's an overview of AWS Glue DataBrew:

Key Features:

1. **Data Profiling:** DataBrew automatically profiles data to detect data quality issues, such as missing values, duplicates, outliers, and inconsistent formats.
2. **Visual Interface:** It provides a no-code, point-and-click interface where users can perform over 250 built-in transformations, such as filtering, normalization, and data validation.
3. **Data Cleaning and Transformation:** DataBrew enables users to clean and prepare data for analysis or machine learning by applying rules and transformations to datasets.
4. **Recipe-Based Data Processing:** Users can create reusable "recipes" that define a series of data preparation steps. These recipes can be reused across different datasets.
5. **Integration with AWS Services:** It integrates seamlessly with other AWS services, such as Amazon S3, Amazon Redshift, Amazon RDS, and AWS Glue, for easy data processing and storage.
6. **Scheduled Jobs:** You can automate data preparation tasks by scheduling DataBrew jobs to run at specified intervals.
7. **Data Sources:** It works with structured and semi-structured data from sources like CSV, JSON, Parquet, and more.

Use Cases:

- Cleaning and enriching raw data before analysis or machine learning model development.
- Preparing data for reporting, dashboards, and business intelligence tools.
- Performing exploratory data analysis with quick visualizations and profiling.

AWS Glue DataBrew is beneficial for those who want to streamline the data preparation process without the need for manual coding or scripting.

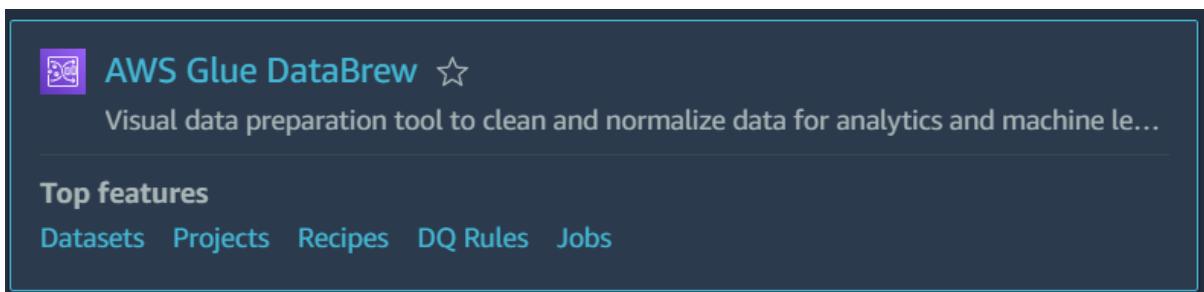
In this example, we use AWS Glue DataBrew to transform a CSV file by converting all data in a specific column to uppercase. The process begins by creating a sample project in DataBrew and selecting the "Popular names for babies in 2020" dataset. After creating the project and setting up the IAM role, we load the data into the DataBrew interface.

Once the data is ready, we choose the "Name" column and apply a formatting transformation to change all values to uppercase. After applying this transformation, we can see the updated data. We then have the option to export the processed data as a CSV file, giving it a custom name before saving it.

The end goal is to visually clean and prepare data, especially when dealing with large, complex datasets. AWS Glue DataBrew helps automate the data transformation process, making it easy to manipulate data for various use cases without the need for coding.

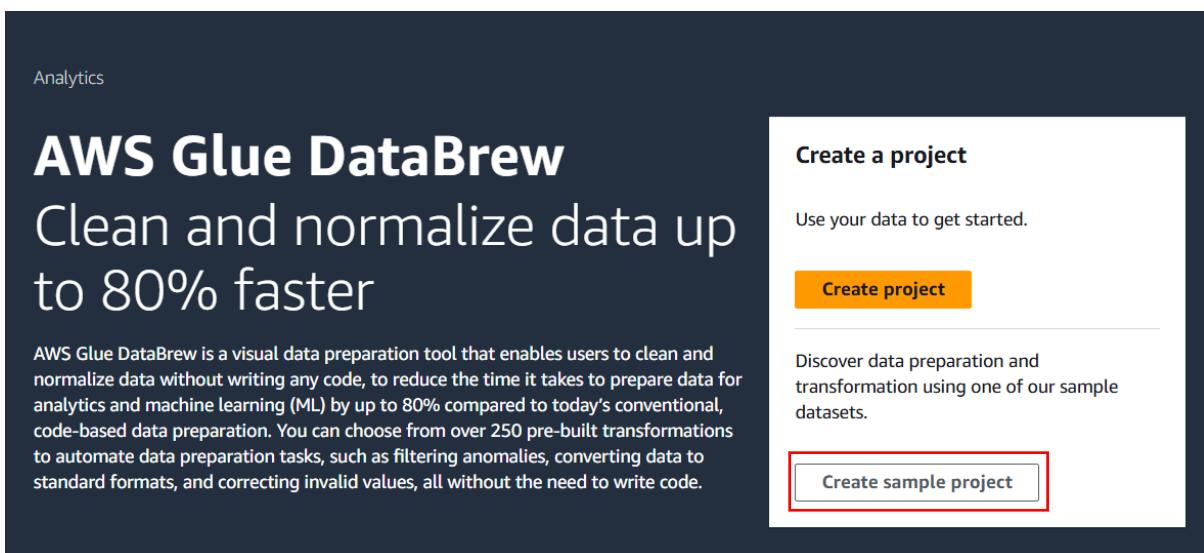
To begin with the Lab:

1. Now the situation is we've got a file that is in a CSV format, and we want to convert, let's say, all the columns into uppercase before we process the data.
2. So that's like an example data we will do now. So, we want to load some data, and we want to perform some transformation over there.
3. First, in your AWS Console search for AWS Glue DataBrew and choose the service accordingly.



The screenshot shows the AWS Glue DataBrew service page. At the top, there is a purple header bar with the service name and a star icon. Below the header, the page title is "AWS Glue DataBrew" followed by a star icon. A sub-header reads "Visual data preparation tool to clean and normalize data for analytics and machine le...". Underneath, there is a section titled "Top features" with links to "Datasets", "Projects", "Recipes", "DQ Rules", and "Jobs".

4. From its dashboard, you need to click on Create sample project.



The screenshot shows the AWS Glue DataBrew dashboard. On the left, there is a dark sidebar with the word "Analytics". The main area has a dark background with white text. The title "AWS Glue DataBrew" is prominently displayed in large white font, followed by the tagline "Clean and normalize data up to 80% faster". Below the title, there is a paragraph describing the service. On the right side, there is a white call-to-action box with the heading "Create a project". Inside the box, there is a sub-headline "Use your data to get started.", a large orange "Create project" button, and a smaller text block "Discover data preparation and transformation using one of our sample datasets." At the bottom of the box, there is another button labeled "Create sample project" which is outlined in red.

5. Then you need to choose a sample project (Popular names for babies in 2020) as shown below and then choose to create a new IAM role and give it a name then click on create project.

Create sample project

Popular names for babies in 2020
dataset-national-baby-names.json | JSON file | 3.7 MiB
Popular baby names in 2020 in the United States with each record tracking name, sex and number of occurrences of the name.

ChEMBL drug discovery data
chembl-27.parquet | Parquet file | 2.2 MiB
ChEMBL is a manually curated database of bioactive molecules with drug-like properties. It brings together chemical, bioactivity and genomic data to aid the translation of genomic information into effective new drugs.

Famous chess game moves
chess-games.xlsx | Microsoft Excel file | 4.4 MiB
All of the information available from 20,000 chess games and how much meta (out-of-game) factors affect a game.

Role name
Choose the role that has access to connect to your data. Refresh to see the latest updates.

New IAM role suffix
Your role will be prefixed with "AWSGlueDataBrewServiceRole-"

6. Now you have to wait until your session gets ready. Once you data is loaded it would look like this.

The screenshot shows the AWS Glue DataBrew interface. On the left, there's a sidebar with 'PROJECTS' selected, showing a list of datasets, DQ rules, jobs, and what's new. The main area displays a 'Sample project - 1' with a dataset named 'dataset-national-baby-names'. A sample of 500 rows is shown in a grid view. The first few rows are:

7065	F	1	Mary
2604	F	2	Anna
2003	F	3	Emma
1939	F	4	Elizabeth
1746	F	5	Minnie
1578	F	6	Margaret
1472	F	7	Ida
1414	F	8	Alice
1320	F	9	Bertha
1288	F	10	Sarah

To the right, there's a 'Recipe (0)' section titled 'Sample recipe - 1' with a note 'Working version'. It shows a 'Build your recipe' section with a 'Add step' button.

7. You need to select the Name column and then click on Format, choose to change to uppercase.

Sample project - 1

Dataset: dataset-national-baby-names

Sample: First n sample (500 rows)

The screenshot shows the Alteryx Designer interface with a workflow titled "Sample project - 1". The left pane displays a table with 500 rows and 5 columns, showing a histogram for the "id" column. A context menu is open over the first column, with the option "Change to uppercase" highlighted. The right pane shows a sample of the data with summary statistics for the "year" column.

#	Name	Year
1	Margaret	1880
2	Ida	1880
3	Alice	1880
4	Bertha	1880
5	Sarah	1880
6		
7		
8		
9		
10		

8. Then from the right side you will have the ability to apply this format.

Format column

Format column to

Uppercase

Example

THE QUICK BROWN FOX JUMPED OVER THE FENCE

Apply transform to

All rows (500 rows)
Transformation will be applied to all rows in the dataset

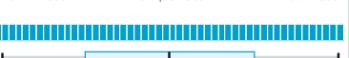
Filtered rows - 0 filters applied (500/500 rows)
Transformation will be applied to filtered rows in the grid

 Preview changes

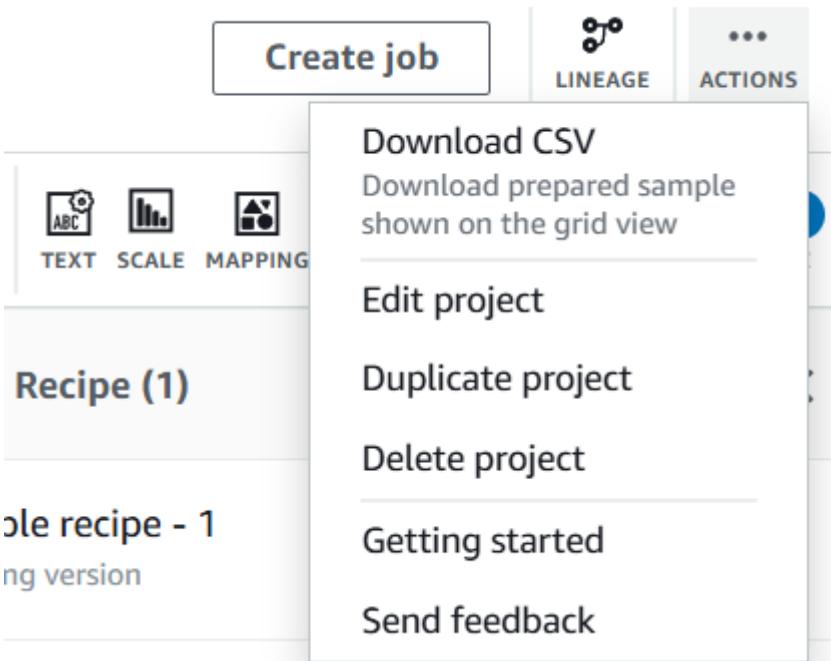
Cancel

Apply

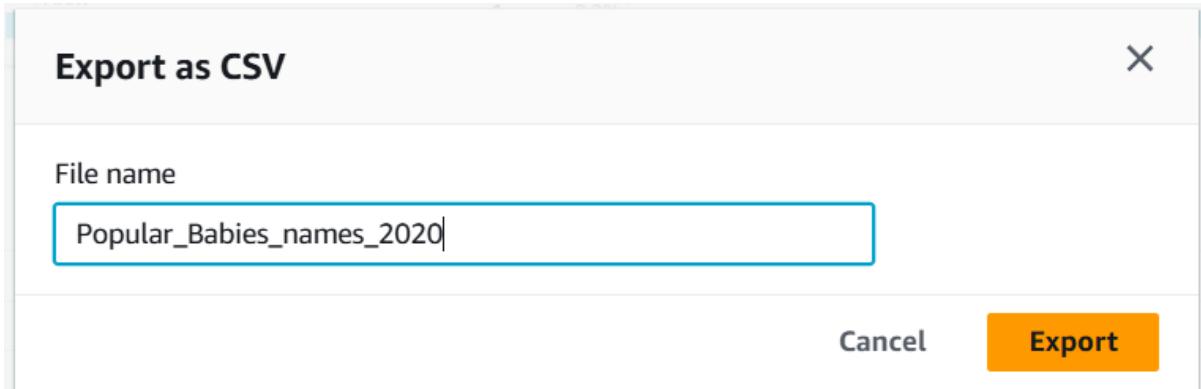
9. Below you can see that your name column is now in Uppercase.

SAMPLE					GRID	SCHEMA	PROFILE
	#	id	ABC	name	#	year	
Total	500	Distinct 500	Unique 500	Total 500	Distinct 500	Unique 500	Total 500
500	100%						
	1			MARY	1	0.2%	
	2			ANNA	1	0.2%	
	3			EMMA	1	0.2%	
	4			All other values	497	99.4%	
	5			MARY		1880	
	6			ANNA		1880	
	7			EMMA		1880	
	8			ELIZABETH		1880	
	9			MINNIE		1880	
	10			MARGARET		1880	
				IDA		1880	
				ALICE		1880	
				BERTHA		1880	
				SARAH		1880	

- From the top right corner if you click on actions, you will see that you have the ability to download the CSV file onto your laptop.



- You can also change the name of the file as per your requirement and click on export.



- Now, you might be wondering where we use it and what is the use case of this AWS Databrew. Now when we are dealing with the data, especially when we are dealing with large amounts of data, which is heterogeneous data sources, in such scenarios, we will be required to manipulate the data so that we can use it on the use case that we are interested in.

- Also, Databrew is visual and you can perform so many things on your data.



- So, we generally use this glue Databrew in the web browser to visually design the recipe job and to process the data as well and we can also preview the results as well. And by

doing this we can automate the data processing workflow, which means whatever the transformation that we are doing.

15. If you click on Create Job, you can save your data in your S3 bucket.
16. Once you are done testing it just delete your job and recipe.