😊 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.

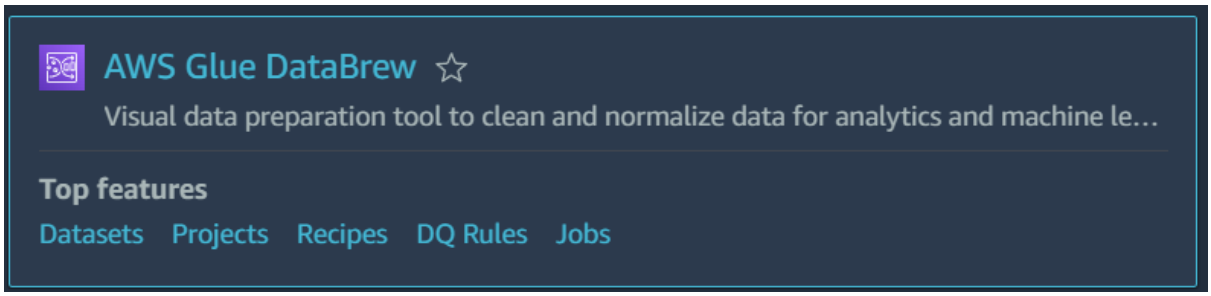
In this example, we use AWS Glue DataBrew to transform a CSV file by converting all data in a specific column to uppercase. The process begins by creating a sample project in DataBrew and selecting the "Popular names for babies in 2020" dataset. After creating the project and setting up the IAM role,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

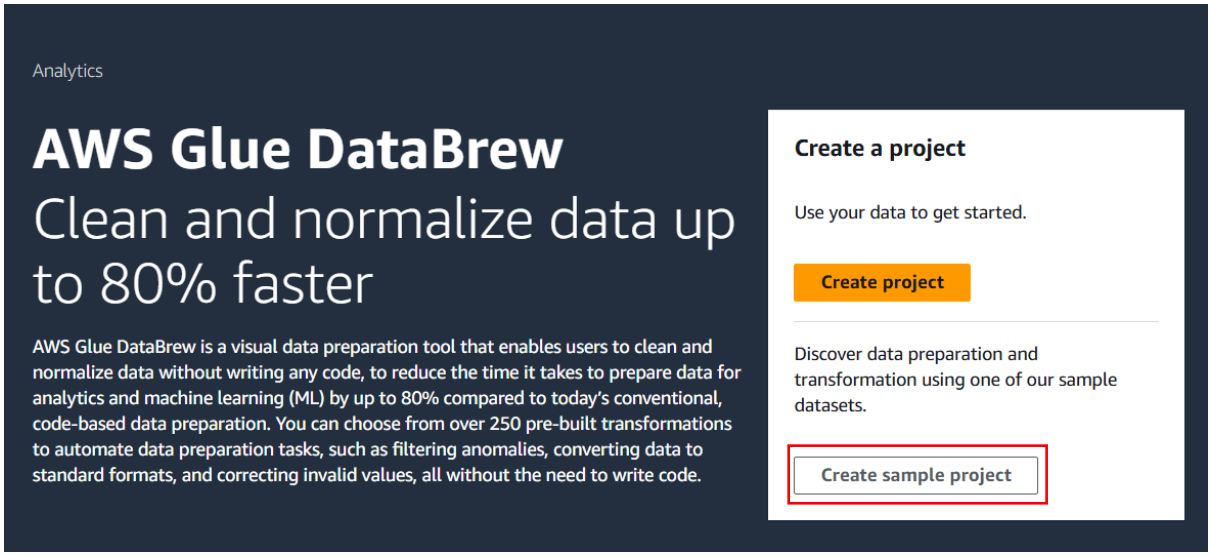
The end goal is to visually clean and prepare data, especially when dealing with large, complex datasets. AWS Glue DataBrew helps automate the data transformation process, making it easy to manipulate data for various use cases without the need for coding.

😄 To begin with the Lab:

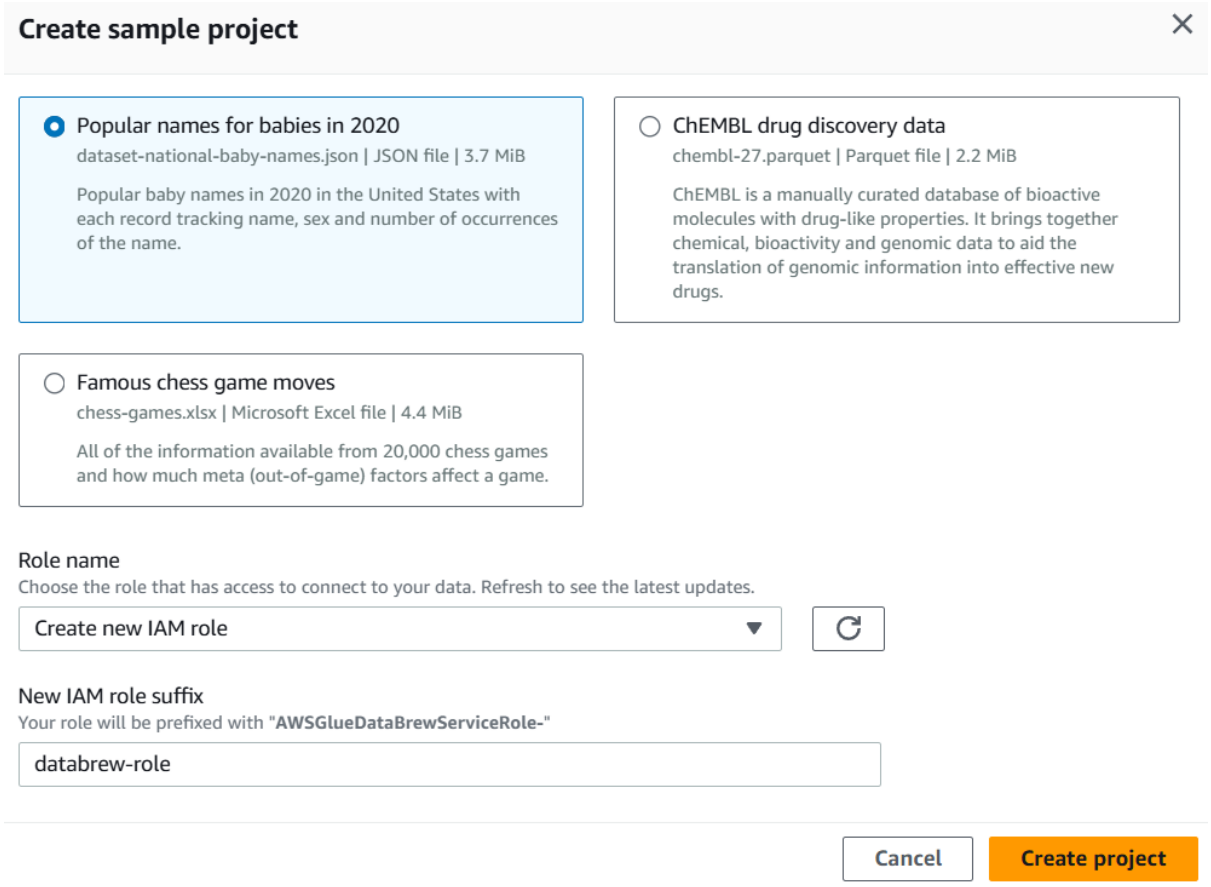
1. Now the situation is we've got a file that is in a CSV format, and we want to convert, let's say, all the columns into uppercase before we process the data.
2. So that's like an example data we will do now. So, we want to load some data, and we want to perform some transformation over there.
3. First,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



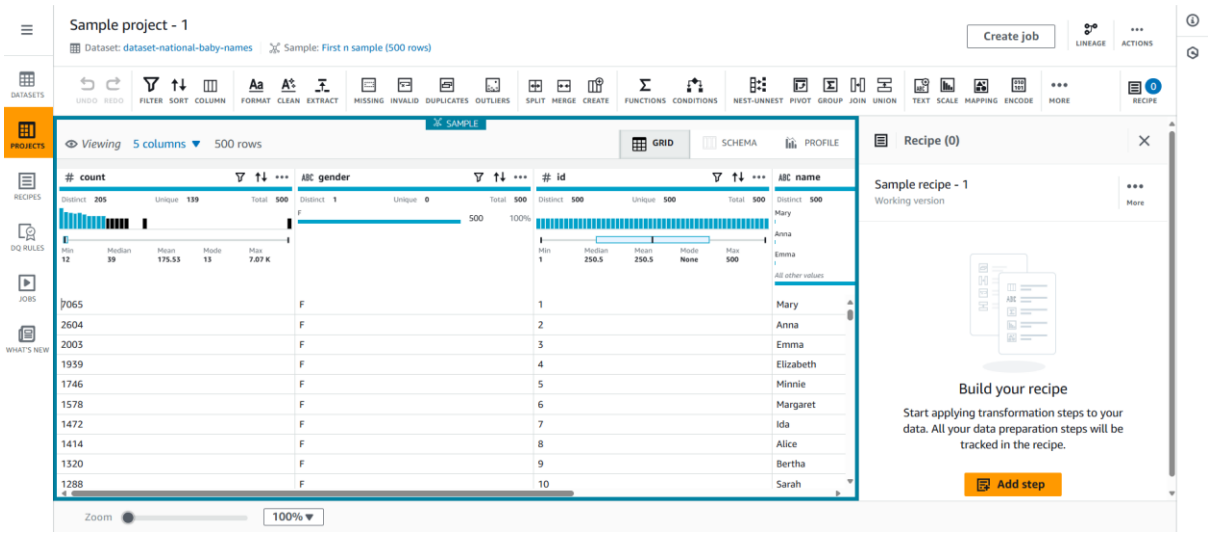
4. From its dashboard, you need to click on Create sample project.



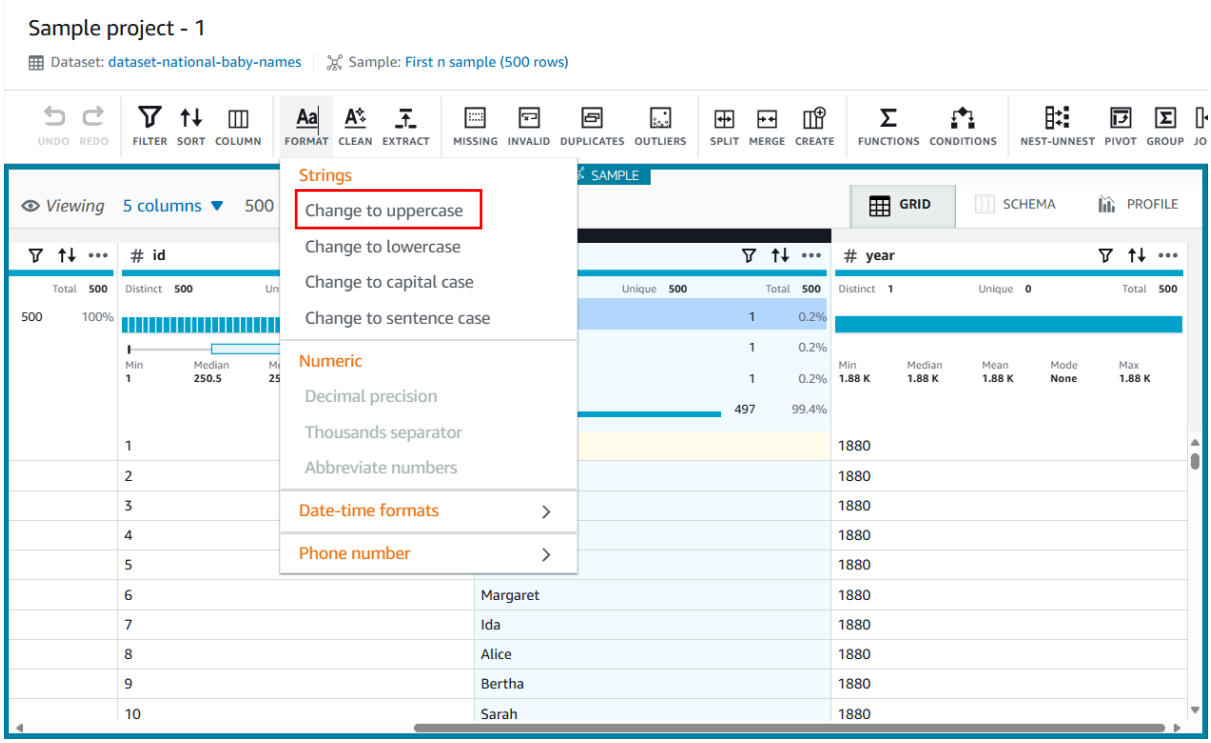
5. Then you need to choose a sample project (Popular names for babies in 2020) as shown below and then choose to create a new IAM role and give it a name then click on create project.



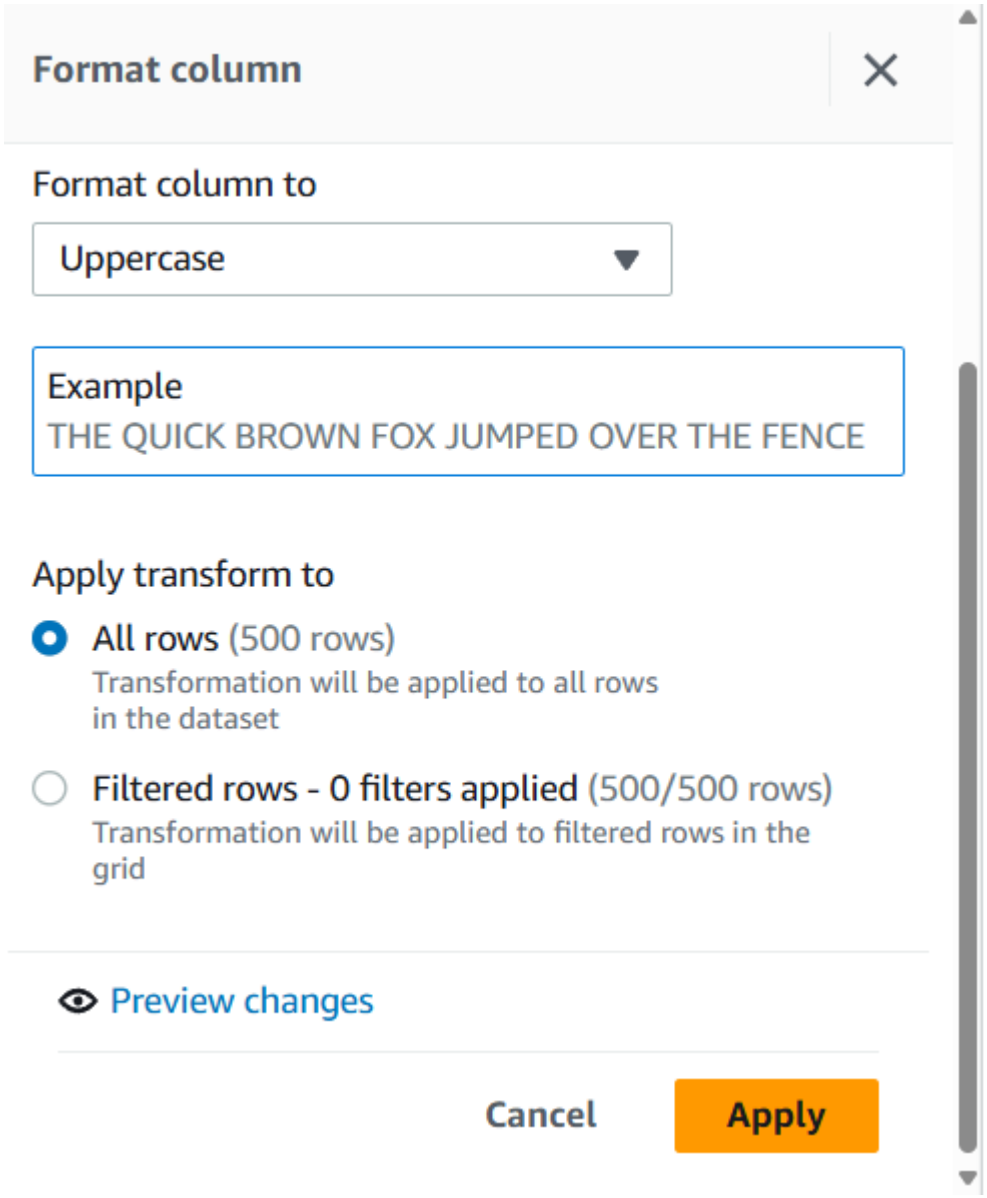
6. Now you have to wait until your session gets ready. Once you data is loaded it would look like this.



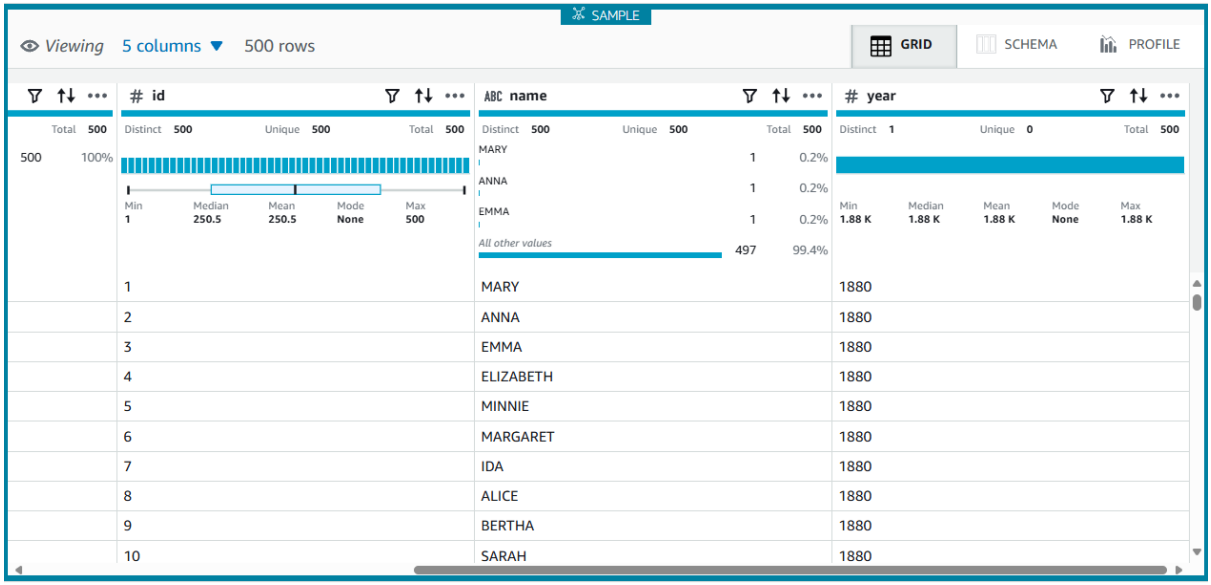
7. You need to select the Name column and then click on Format, choose to change to uppercase.



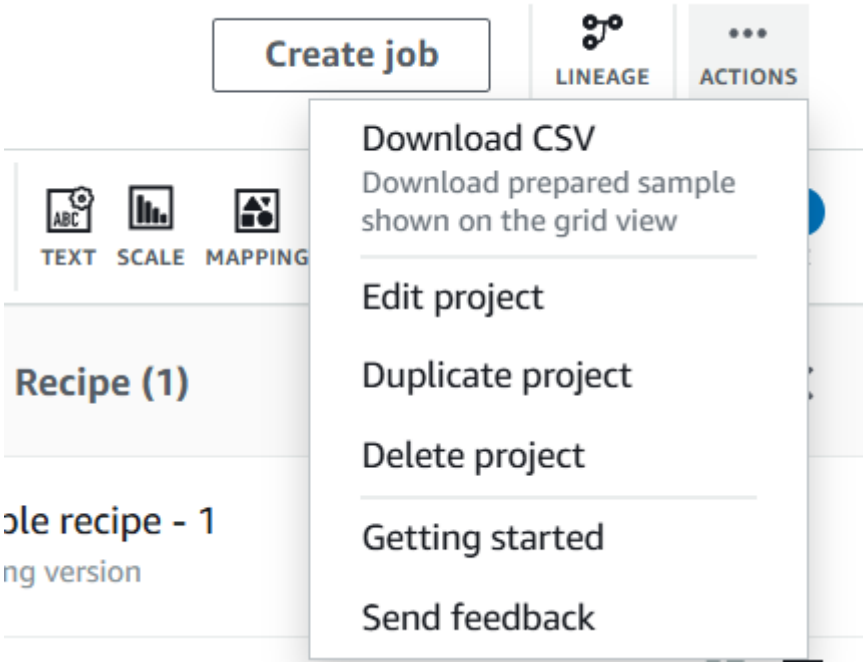
8. Then from the right side you will have the ability to apply this format.



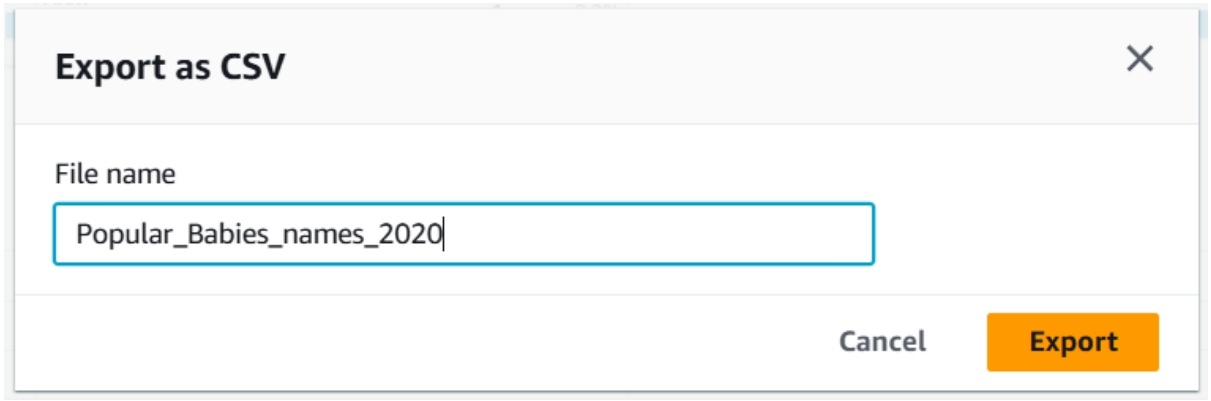
9. Below you can see that your name column is now in Uppercase.



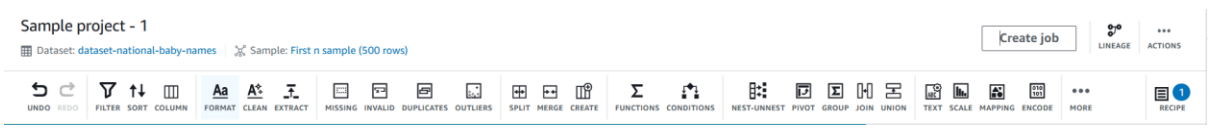
10. From the top right corner if you click on actions, you will see that you have the ability to download the CSV file onto your laptop.



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



12. 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.
13. Also, Databrew is visual and you can perform so many things on your data.



14. 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.