



**Amazon Web Services
Data Engineering Immersion Day**

Lab 5. Bonus Lab: Glue DataBrew
March 2021

Table of Contents

Introduction	2
Get Started Using the Lab Environment	3
DataBrew - Pre-Lab Setup	5
Data preparation with Glue DataBrew	7
Creating a project.....	7
Exploring the dataset.....	11
Preparing the dataset	15
Creating a DataBrew job.....	25
Viewing data lineage.....	27

Introduction

In [DataBrew](#) lab you will use a different dataset than event ticket dataset, which has data anomalies. It will help you to learn about DataBrew which makes it easy for data analysts and data scientists to clean and normalize data to prepare it for analytics and machine learning.

Below is a list of the steps for this lab:

- [DataBrew Pre-Lab Setup](#)
- [Data preparation with Glue DataBrew](#)

Today, you are attending a formal AWS event, so we provide an AWS account to you. If in the future you might want to perform these labs in your own AWS environment by yourself, suggest you to save this file to your computer for the future reuse.

Alternatively, run the lab again by following the online instruction here - <https://aws-dataengineering-day.workshop.aws/1300.html>

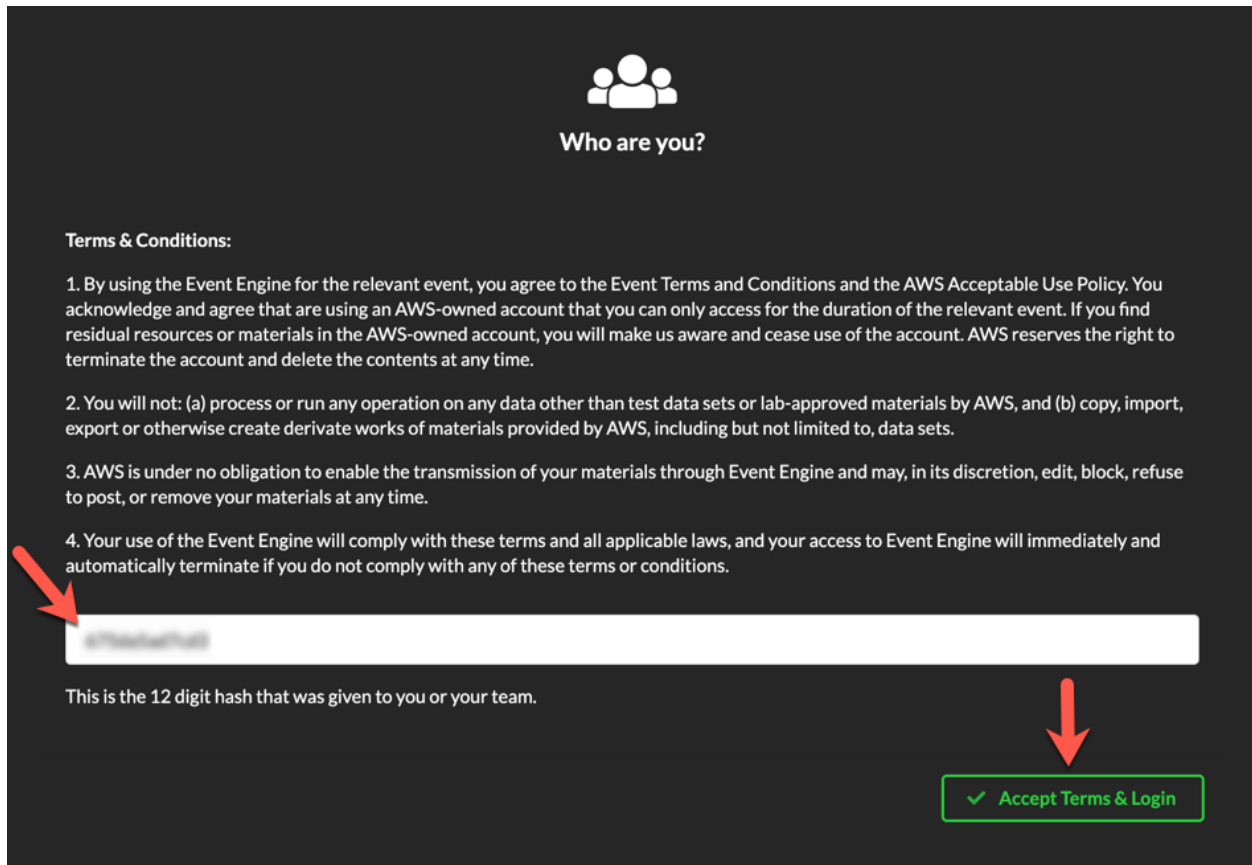
Get Started Using the Lab Environment

Please skip this section if you are running the lab on your own AWS account.

Today, you are attending an AWS event and you will have been sent your access details beforehand. If in the future you might want to perform these labs in your own AWS environment by yourself, you can follow instructions on GitHub - <https://github.com/aws-samples/data-engineering-for-aws-immersion-day>.

A 12-character access code (or 'hash') is the access code that grants you permission to use a dedicated AWS account for the purposes of this workshop.

1. Go to <https://dashboard.eventengine.run/>, enter the access code and click Proceed:



Who are you?

Terms & Conditions:

1. By using the Event Engine for the relevant event, you agree to the Event Terms and Conditions and the AWS Acceptable Use Policy. You acknowledge and agree that are using an AWS-owned account that you can only access for the duration of the relevant event. If you find residual resources or materials in the AWS-owned account, you will make us aware and cease use of the account. AWS reserves the right to terminate the account and delete the contents at any time.
2. You will not: (a) process or run any operation on any data other than test data sets or lab-approved materials by AWS, and (b) copy, import, export or otherwise create derivate works of materials provided by AWS, including but not limited to, data sets.
3. AWS is under no obligation to enable the transmission of your materials through Event Engine and may, in its discretion, edit, block, refuse to post, or remove your materials at any time.
4. Your use of the Event Engine will comply with these terms and all applicable laws, and your access to Event Engine will immediately and automatically terminate if you do not comply with any of these terms or conditions.


This is the 12 digit hash that was given to you or your team.

✓ Accept Terms & Login

2. On the Team Dashboard web page you will see a set of parameters that you will need during the labs. Best to save them to a text file locally, alternatively you can always go to this page to review them. Replace the parameters with the corresponding values from here where indicated in subsequent labs:

Lab 5. Bonus Lab: Glue DataBrew


Because you're at a formal event, some AWS resources have been pre-deployed for your convenience, for example you can see a list of parameters on your event dashboard:


 Modules


Environment Setup


[Readme](#)


Outputs:


S3 Bucket name
mod-3fccddd609114925-dmslabs3bucket-1ngcgzzcnd15u 


BusinessAnalystUser
mod-3fccddd609114925-BusinessAnalystUser-MBOXFZLQLOXX 

DMSLabRoleS3 ARN
arn:aws:iam::377243295828:role/mod-3fccddd609114925-DMSLabRoleS3-O2VT1RSN43SG 

Glue Lab Role
mod-3fccddd609114925-GlueLabRole-YLTJA13WW6WT 

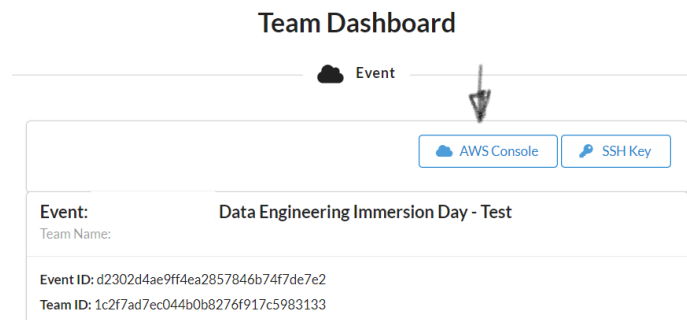
S3BucketWorkgroupA
mod-3fccddd609114925-s3bucketworkgroupa-tbon3m1mkunh 

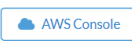

S3BucketWorkgroupB
mod-3fccddd609114925-s3bucketworkgroupb-18ygl8nfp8ead 

WorkgroupManagerUser
mod-3fccddd609114925-WorkgroupManagerUser-5IVE0UQNIBG4 

- On the Team Dashboard, please click AWS Console to log into the AWS Management Console:

Team Dashboard

 Event

Event: Data Engineering Immersion Day - Test

Team Name:

Event ID: d2302d4ae9ff4ea2857846b74f7de7e2



Team ID: 1c2f7ad7ec044b0b8276f917c5983133

- Click Open Console. For the purposes of this workshop, you will not need to use command line and API access credentials

AWS Console Login

Remember to only use "us-east-1" as your region, unless otherwise directed by the event operator.

Login Link

Credentials / CLI Snippets

Mac / Linux Windows

Mac or Linux

```
export AWS_DEFAULT_REGION=us-east-1
export AWS_ACCESS_KEY_ID=mod-3fccddd609114925-BusinessAnalystUser-MBOXFZLQLOXX
export AWS_SECRET_ACCESS_KEY=mod-3fccddd609114925-BusinessAnalystUser-MBOXFZLQLOXX
export AWS_SESSION_TOKEN=mod-3fccddd609114925-BusinessAnalystUser-MBOXFZLQLOXX
```

How do I use the AWS CLI?

Checkout the AWS CLI documentation here: <https://docs.aws.amazon.com/cli/latest/userguide/cli-welcome.html>

OK

Once you have completed these steps, you can continue with the rest of this lab.

DataBrew - Pre-Lab Setup

Steps

- Introduction
- CloudFormation Stack Deployment

Introduction


In this lab, we will be using AWS Glue DataBrew to explore a dataset in S3, and to clean and prepare the data.

To do this, we will first set up an IAM role to use in DataBrew, and an S3 bucket for the results from the DataBrew jobs.

CloudFormation Stack Deployment

Make sure you are in the us-east-1 (N. Virginia) region

1. Click the **Deploy to AWS** icon below to create the AWS resources for the lab.

Launch Quick Create Template	Region
	N.Virginia (us-east-1)

2. Check the box "I acknowledge that ...", then click on "Create Stack" to create the stack.

Lab 5. Bonus Lab: Glue DataBrew

CloudFormation > Stacks > Create stack

Quick create stack

Template

Template URL
https://s3.us-east-1.amazonaws.com/aws-dataengineering-day/workshop.aws/DataBrew_PreLab_CFN.yaml

Stack description
-

Stack name

Stack name
databrew-lab
Stack name can include letters (A-Z and a-z), numbers (0-9), and dashes (-).

Parameters

Parameters are defined in your template and allow you to input custom values when you create or update a stack.

No parameters
There are no parameters defined in your template.

Capabilities

The following resource(s) require capabilities: [AWS::IAM::ManagedPolicy]
This template contains Identity and Access Management (IAM) resources that might provide entities access to make changes to your AWS account. Check that you want to create each of these resources and that they have the minimum required permissions. [Learn more](#)

☒ I acknowledge that AWS CloudFormation might create IAM resources.

Cancel Create change set **Create stack**

In case you aren't able to launch the quick create stack, you can download the [template file](#) and then follow the steps to [create stack](#) manually.

- Once your stack is deployed, click the **Outputs** tab to view more information

databrew-lab

Delete Update Stack actions Create stack

Stack Info Events Resources **Outputs** Parameters Template Change sets

Outputs (2)

Search outputs

Key	Value	Description	Export name
DataBrewLabRole	databrew-lab-DataBrewLabRole-314O3K1MT6ZM	IAM role for DataBrew lab	-
DataBrewOutputS3Bucket	databrew-lab-databrewoutputs3bucket-p5jcyb8htw76	S3 bucket for DataBrew output	-

Note the values for **DataBrewLabRole** and **DataBrewOutputS3Bucket** which will be used in the lab.

Congratulations! You are all done with the CloudFormation deployment.

Data preparation with Glue DataBrew

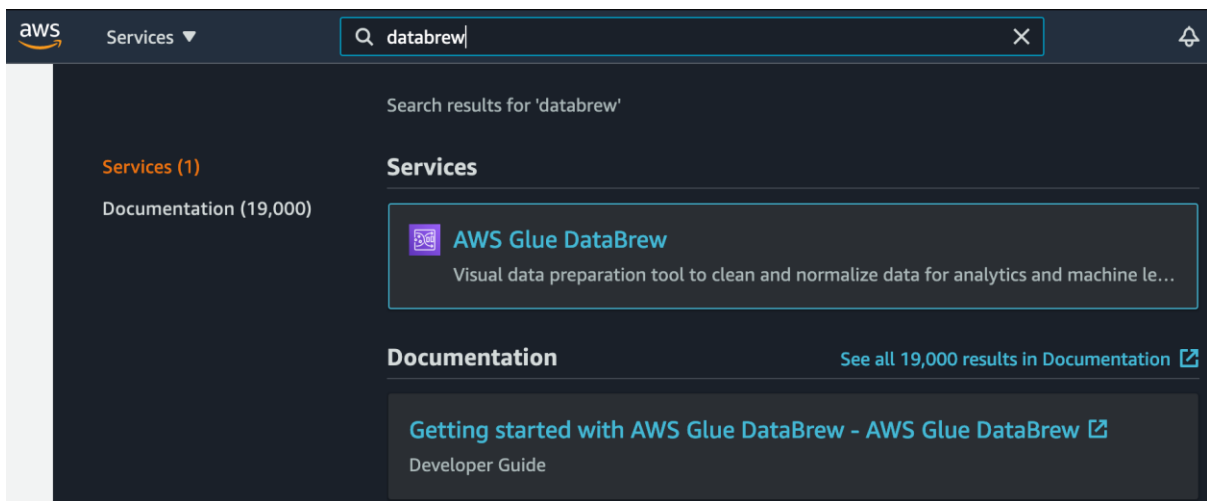
In this lab you will be completing the following tasks.

Tasks Completed in this Lab:

- Create a Glue DataBrew project to explore a dataset
- Connect a sample dataset from S3
- Explore the dataset in Glue DataBrew
- Generate a rich data profile for the dataset
- Create a recipe and job to clean and normalize data in the dataset

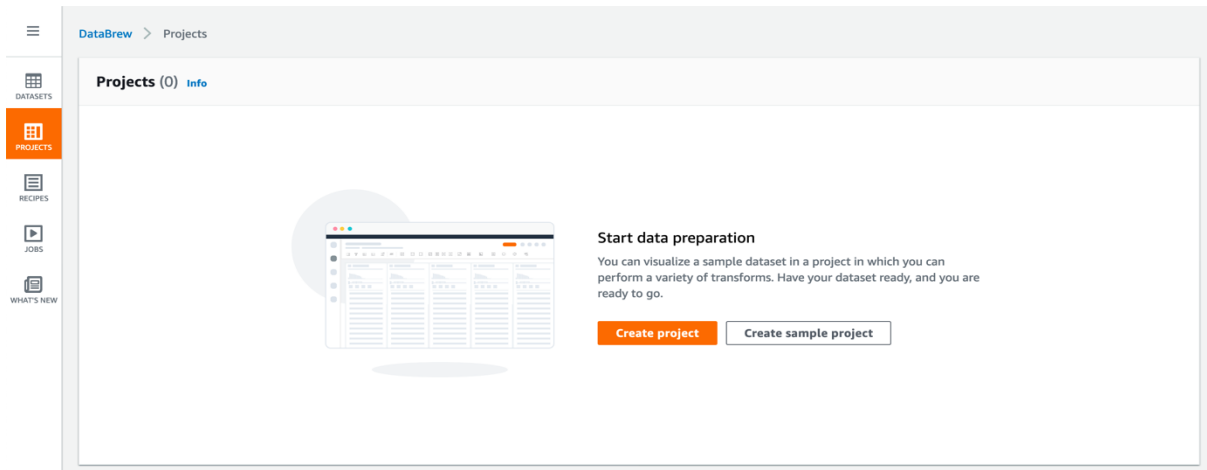
Creating a project

1. Navigate to the AWS Glue DataBrew service



2. On the DataBrew console, select [Projects](#)

Lab 5. Bonus Lab: Glue DataBrew



3. Click **Create project**
4. In the **Project details** section, enter covid-states-daily as the project name

Create project [Info](#)

Project details

Project name

The project name must contain 1-255 characters. Valid characters are alphanumeric (A-Z, a-z, 0-9), hyphen (-), period (.), and space.

Recipe details [Info](#)

Data cleaning steps in DataBrew are stored as a recipe. A recipe is connected to a project by default. An existing recipe with no associated project could also be applied to a project.

Attached recipe

Create new recipe ▼

Recipe name

The recipe name must contain 1-255 characters. Valid characters are alphanumeric (A-Z, a-z, 0-9), hyphen (-), period (.), and space.

☐ **Import steps from recipe**


Import recipe steps from an existing recipe into your project. The existing recipe that you chose will not be edited.


5. In the **Select a dataset** section, select **New dataset** and enter covid-states-daily-stats


Lab 5. Bonus Lab: Glue DataBrew

Select a dataset

Select the dataset that you want to work on

☐  **My datasets**
Your imported datasets

☐  **Sample files**
Explore example files for your dataset

☒  **New dataset**
Import new dataset

New dataset details


Dataset name

The dataset name must contain 1-255 characters. Valid characters are alphanumeric (A-Z, a-z, 0-9), hyphen (-), period (.), and space.


6. In the **Connect to a new dataset** section, select **Amazon S3** under “Data lake/data store”

Enter the S3 path `s3://aws-dataengineering-day.workshop.aws/states_daily.csv.gz`


Connect to new dataset [Info](#)


 File upload


Data lake/data store


 **Amazon S3**

AWS Glue Data Catalog


 Amazon S3 tables

 Amazon Redshift tables

 Amazon RDS tables

 All AWS Glue tables

Others

 AWS Data Exchange


S3 path

Parameterized S3 path


Enter your source from S3 [Info](#)

For you to select a folder, all files in the folder need to share the same file type. If there are different schemas, they will be merged.

Format is: `s3://bucket/prefix`

S3 Buckets 

< 1 2 >



Name	Size
-	

7. In the **Sampling** section, leave the default configuration values

9

Lab 5. Bonus Lab: Glue DataBrew

▼ **Sampling - optional**
Select the type and size of your sample

Type

First n rows ▼

How many rows do you want to sample?

☒ 500

☐ 1,000

☐ 2,500

☐ Custom size


8. In the **Permissions** section, select the role databrew-lab-DataBrewLabRole-xxxxx from the drop-down list

Permissions [Info](#)

DataBrew needs permission to connect to data on your behalf. Use an IAM role with the [required policy](#) attached.

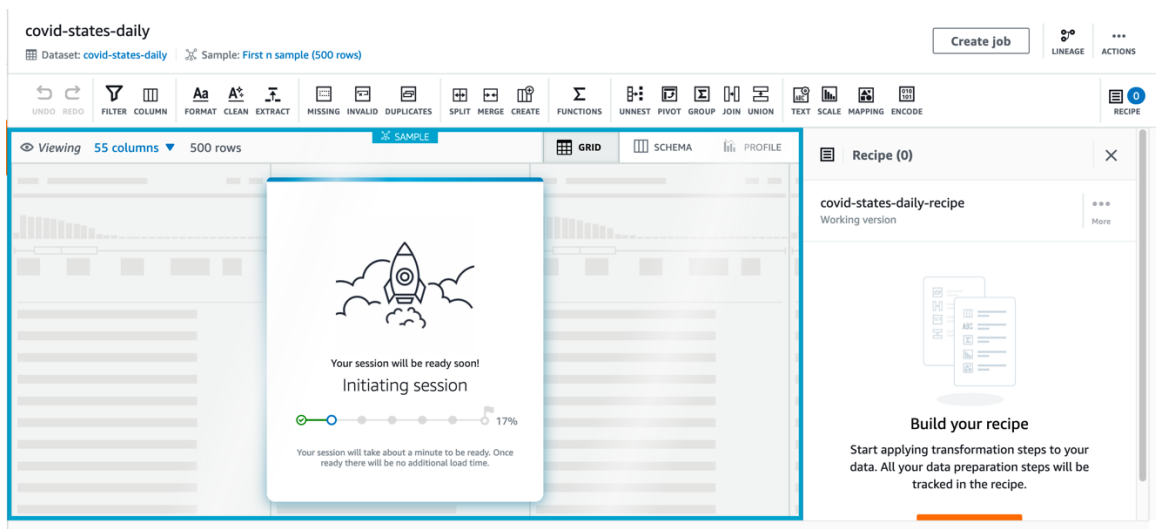
Role name

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

databrew-lab-DataBrewLabRole-314O3K1MT6ZM ▼ 

9. Click **Create project**

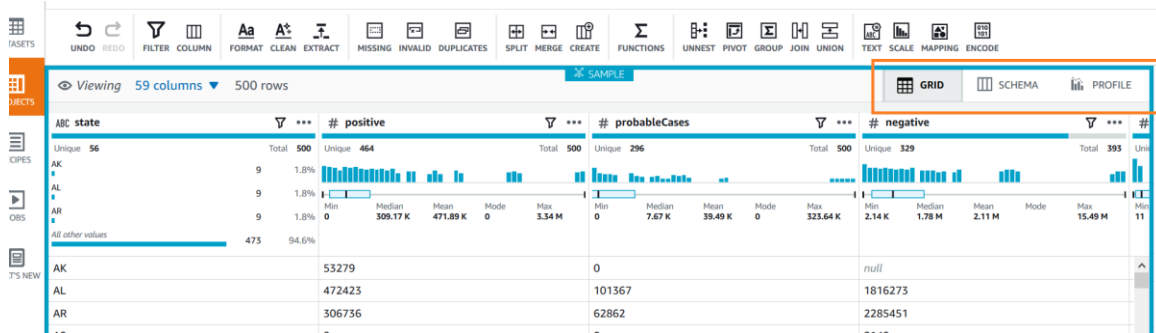
Glue DataBrew will create the project, this may take a few minutes.



The screenshot shows the Glue DataBrew interface. At the top, the dataset 'covid-states-daily' is selected, with a sample of 'First n sample (500 rows)'. The 'Create Job' button is visible. Below the dataset, a toolbar contains various transformation and filtering options. The main workspace is divided into two panes: 'Viewing' on the left and 'Recipe (0)' on the right. A modal window is centered over the workspace, displaying a rocket icon and the text 'Your session will be ready soon! Initiating session'. A progress bar below the text shows 17% completion. The 'Recipe (0)' pane on the right shows a 'covid-states-daily-recipe' with a 'Working version' and a 'Build your recipe' section.

Exploring the dataset

- When the project has been created, you will be presented with the **Grid** view. This is the default view, where a sample of the data is shown in tabular format.



The Grid view shows

- Columns in the dataset
- Data type of each column
- Summary of the range of values that have been found
- Statistical distribution for numerical columns

- Click on the **Schema** tab

The Schema view shows the schema that has been inferred from the dataset. In schema view, you can see statistics about the data values in each column.

In the Schema view, you can

- Select the checkbox next to a column to view the summary of statistics for the column values
- Show/Hide columns
- Rename columns
- Change the data type of columns
- Rearrange the column order by dragging and dropping the columns

- Click on the **Profile** tab

In the Profile view, you can run a data profile job to examine and collect statistical summaries about the data. A data profile is an assessment in terms of structure, content, relationships, and derivation.

Click on **Run data profile**.

- In the **job details** and **job run sample** panels, leave the default values.

Lab 5. Bonus Lab: Glue DataBrew

DataBrew > Jobs > Create job

Create job [Info](#)

Job details

Job name
Identifier for the jobs

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

Job run sample

A job can be run on the entire dataset or a custom sample of the dataset.

Data sample
Define the scope of the dataset to run the job on


☐ Full dataset

☒ Custom sample

rows

Value must be greater than zero

Job type



Profile job
A profile job generates summary and statistics that give you the shape of your data.

Associated dataset
[covid-states-daily](#)
S3 | s3://ale-code-bucket/states_daily.csv

- b) In the **Job output settings** section, select the S3 bucket with the name databrew-lab-databrewoutputs3bucket-xxxxx and a folder name (eg. data-profile)

Lab 5. Bonus Lab: Glue DataBrew

Job output settings [Info](#)

Running a job generates output files at specified file destinations.

File type	S3 location
Output format	Format is: s3://bucket/folder/
JSON	<input type="text" value="s3://databrew-lab-databrewoutputs3bucket-p5jcyb8htw76/data-profile/"/> <input type="button" value="Browse"/>

Encryption

☐ Enable encryption for job output file
Encrypt the job output file using SSE-S3 or AWS KMS

► **Advanced job settings - optional** [Info](#)

Settings that control the processing and compute used for the jobs run on your project

► **Associated schedules - optional** [Info](#)

You can associate up to 2 schedules to automate your job.

► **Tags - optional**

Metadata that you can define and assign to AWS resources. Each tag is a simple label consisting of a customer-defined key (name) and an optional value. Using tags can make it easier for you to manage, search for, and filter resources by purpose, owner, environment, or other criteria.

Permissions [Info](#)

DataBrew needs permission to connect to data on your behalf. Use an IAM role with the [required policy](#) attached.

Role name

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

- c) In the **Permissions** section, select the IAM role with the name databrew-lab-DataBrewLabRole-xxxxx
- d) Leave all other settings as the default values
- e) Click **Create and run job**

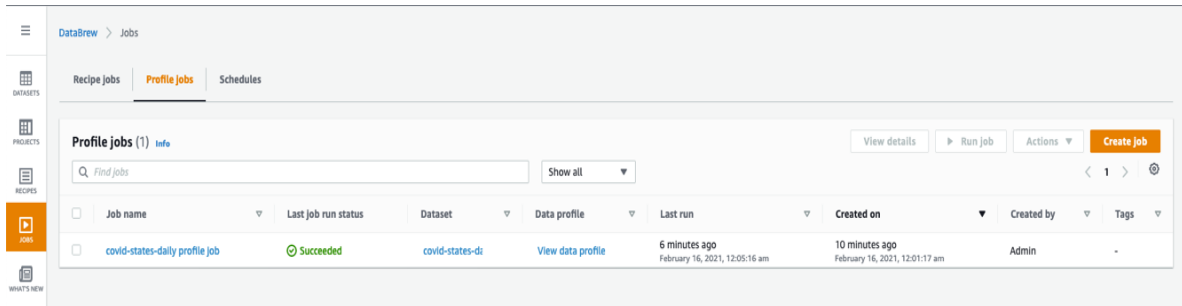
The data profile job takes approximately 5 minutes complete. You can continue with the rest of the labs from step 15 below while you wait, and return to the following steps to examine the profile of the dataset.

13. Click on **Jobs** from the menu on the left-hand side of the DataBrew console.

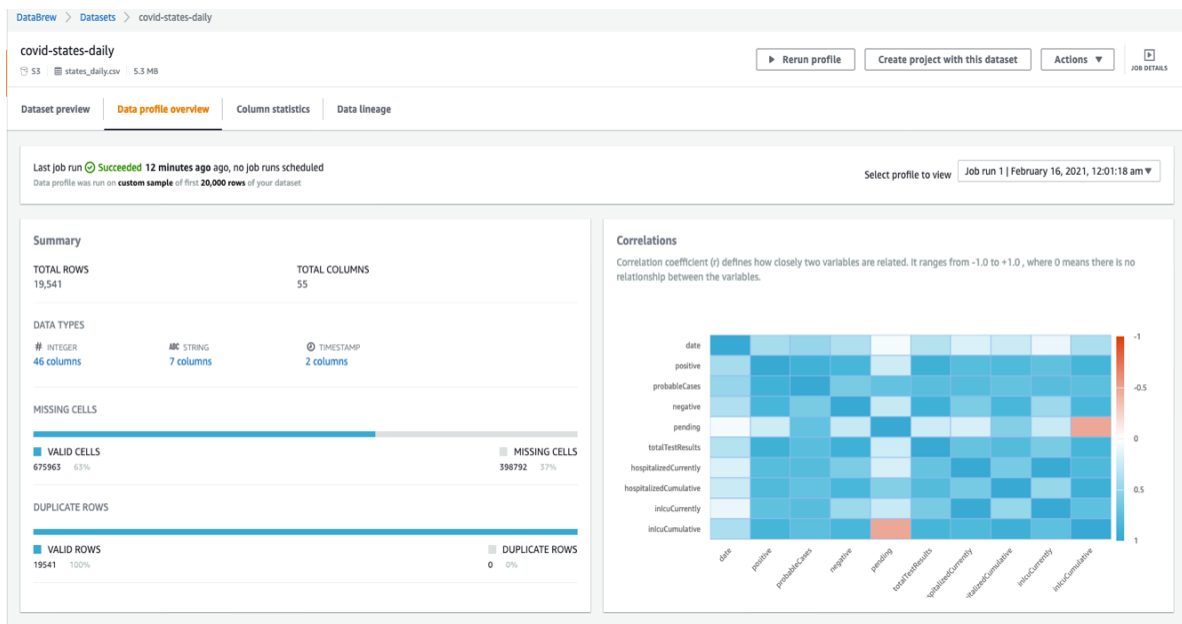
Lab 5. Bonus Lab: Glue DataBrew

Click on **Profile jobs** tab to view a list of profile jobs.

You can see the status of your profile job on this screen.



When the profile job has successfully completed, click on **View data profile**.

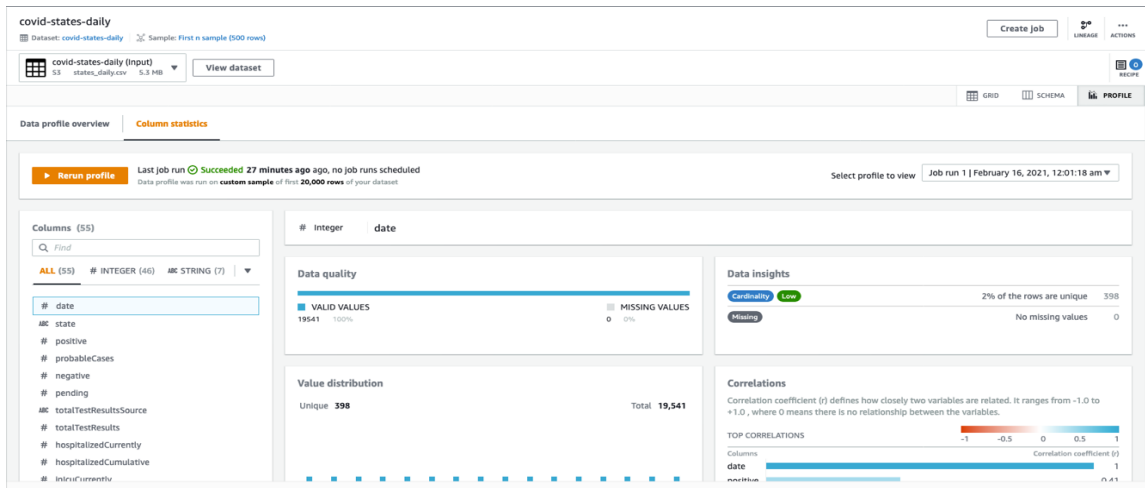


You can also access the data profile from the **Profile** tab in the project.

The data profile shows a summary of the rows and columns in the dataset, how many columns and rows are valid, and correlations between columns.

14. Click on the **Column statistics** tab to view a column-by-column breakdown of the data values.

Lab 5. Bonus Lab: Glue DataBrew



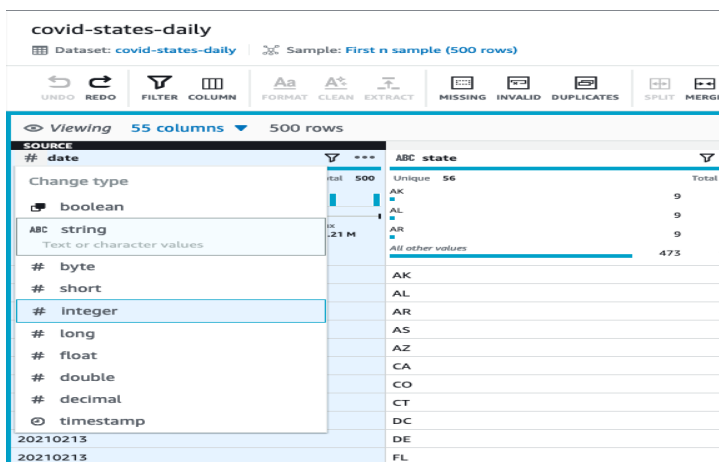
Preparing the dataset

In this section, we will apply the following transformations to the dataset.

- Convert the date column from integer to string
- Split the date column into three new columns (year, month, day) to partition the data by these columns
- Fill the missing values in the probableCases column with 0
- Map the values of the dataQualityGrade column to a numerical value

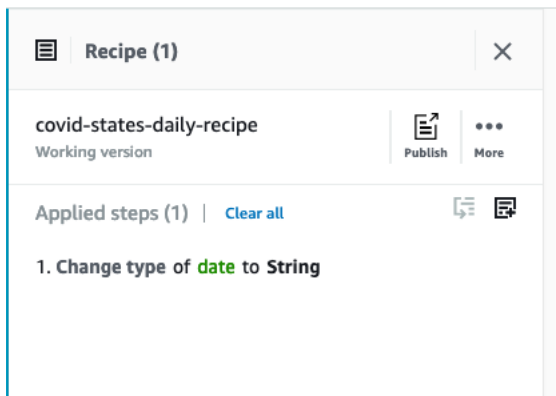
1. Navigate back to the covid-states-daily project grid view.
2. DataBrew has inferred data type of the date column as integer. We will convert the data type of the date column to string.

Click on the **#** icon next to the date column name and select **string**



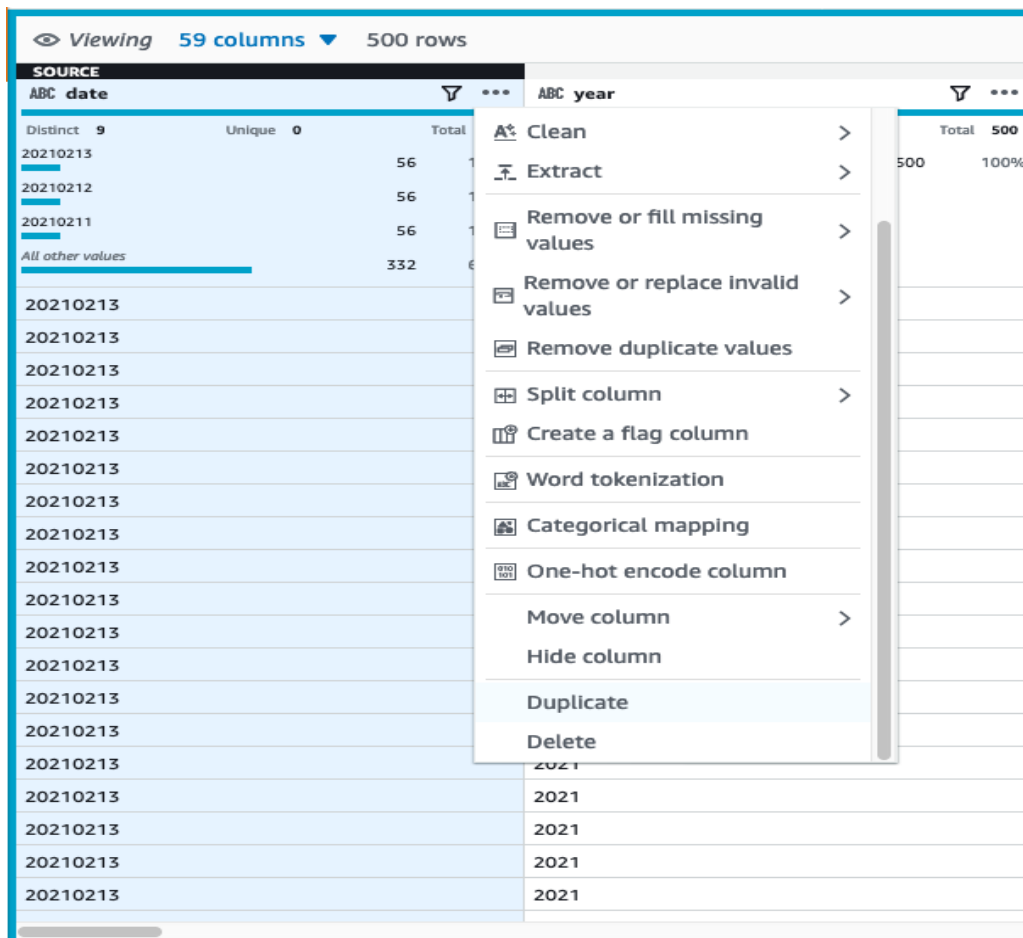
Note that the transformation is added to the recipe at the right.

Lab 5. Bonus Lab: Glue DataBrew



4. We will duplicate the date column first before splitting it into year, month, day columns, as the original column will be deleted by this transformation.

Select the ... at the top of the date column. From the pop-up menu, scroll to the bottom and select **Duplicate**



5. Leave the default settings in the **Duplicate column** dialog, click **Apply**

Lab 5. Bonus Lab: Glue DataBrew

Duplicate column

Create duplicate column [Info](#)

Duplicate a column from an existing column

Source column

Select a source column to create a duplicate

date

Duplicate column name

Name of the newly created duplicate column

date_copy

Valid characters are alphanumeric, underscore, and space

[Preview changes](#)

Cancel

Apply

A copy of the date column is created with the name date_copy. Note that the **duplicate column** transformation is added as a step to the recipe at the right.

6. Let's split the date_copy column into year, month, day columns.

Select the ... at the top of the date_copy column. Select **Split column / At positions from beginning**

The screenshot shows the Glue DataBrew interface. On the left, a table with the column 'date_copy' is displayed. The table has 9 unique values and a total of 500 rows. The values are '20210213', '20210212', '20210211', and 'All other values'. A context menu is open over the 'date_copy' column, showing various transformations. The 'Split column' option is selected, and a sub-menu is open showing the 'At positions from beginning' option. On the right, a summary panel for the '# positive' column is visible, showing a total of 500, unique values of 463, and a distribution of values.

Source	date_copy	Total
Unique	9	
20210213		56
20210212		56
20210211		56
All other values		332

Target	# positive	Total	Unique
9	1.8%	500	463
9	1.8%		
9	1.8%		
473	94.6%		
			54282
			479856
			312562
			0
			795323
			3391036

Delimiters

- On a single delimiter
- On multiple delimiters
- Between delimiters

Positions

- At positions from beginning
- At positions from end
- Between positions
- At regular intervals

Lab 5. Bonus Lab: Glue DataBrew

7. In the **Split column** dialog, enter 4 for **Position from the beginning** to split out the year. Leave all other default settings.

Split column

×

Split column [Info](#)

Split a column and create new columns

Source column

Select one column to split

date_copy ▼

Split column options

Using positions ▼

Split column

☒ At positions from beginning

☐ At positions from end

☐ Between positions

☐ At regular intervals

Position from the beginning

Enter a numeric position. Positions count from 1.

4

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

8. In the **Split column** dialog, scroll down and click **Preview changes** to see how the column is split. Note that the **date_copy** column is marked for **deletion**. Click **Apply**.

covid-states-daily

Dataset: covid-states-daily Sample: First n sample (500 rows)

UNDO REDO FILTER COLUMN FORMAT CLEAN EXTRACT MISSING INVALID DUPLICATES SPLIT MERGE CREATE FUNCTIONS SUMMER PIVOT GROUP JOIN UNION TEXT SCALE MAPPIING ENCODE

Viewing 57 columns ▼ 500 rows ▢ View highlighted only

SOURCE		will be deleted		TARGET		TARGET		AIC state	
date_copy		AIC date_copy_1		AIC date_copy_2					
Unique 9	Total: 369	Unique 9	Total: 369	Unique 9	Total: 369	Unique 9	Total: 369	Unique 9	Total: 369
20210211	\$6 11.3%	20210211	\$6 11.3%	20210211	\$6 11.3%	AK		AK	
20210212	\$6 11.3%	20210212	\$6 11.3%	20210212	\$6 11.3%	AL		AL	
20210211	\$6 11.3%	20210211	\$6 11.3%	20210211	\$6 11.3%	AR		AR	
All other values	\$52 66.4%	All other values	\$52 66.4%	All other values	\$52 66.4%	All other values		All other values	
20210213	20210213	20210213	20210213	20210213	20210213	AK		AK	
20210213	20210213	20210213	20210213	20210213	20210213	AL		AL	
20210213	20210213	20210213	20210213	20210213	20210213	AR		AR	
20210213	20210213	20210213	20210213	20210213	20210213	AS		AS	
20210213	20210213	20210213	20210213	20210213	20210213	AZ		AZ	
20210213	20210213	20210213	20210213	20210213	20210213	CA		CA	
20210213	20210213	20210213	20210213	20210213	20210213	CO		CO	
20210213	20210213	20210213	20210213	20210213	20210213	CT		CT	
20210213	20210213	20210213	20210213	20210213	20210213	DC		DC	
20210213	20210213	20210213	20210213	20210213	20210213	DE		DE	
20210213	20210213	20210213	20210213	20210213	20210213	FL		FL	
20210213	20210213	20210213	20210213	20210213	20210213	GA		GA	
20210213	20210213	20210213	20210213	20210213	20210213	GU		GU	
20210213	20210213	20210213	20210213	20210213	20210213	IA		IA	
20210213	20210213	20210213	20210213	20210213	20210213	ID		ID	
20210213	20210213	20210213	20210213	20210213	20210213	IL		IL	
20210213	20210213	20210213	20210213	20210213	20210213	IN		IN	

Split column

date_copy ▼

Split column options
Using positions ▼

Split column

- ☒ At positions from beginning
- ☐ At positions from end
- ☐ Between positions
- ☐ At regular intervals

Position from the beginning
Enter a numeric position. Positions count from 1.

4

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 shown

Cancel Apply

Lab 5. Bonus Lab: Glue DataBrew

9. Next, **split** the **date_copy_2** column into month and day. The result should look like the screenshot below.

date	date_copy_1	date_copy_2	state
20210215	2021	02	AK
20210215	2021	02	AL
20210215	2021	02	AR
20210215	2021	02	AS
20210215	2021	02	AZ
20210215	2021	02	CA
20210215	2021	02	CO
20210215	2021	02	CT
20210215	2021	02	DC
20210215	2021	02	DE
20210215	2021	02	FL
20210215	2021	02	GA
20210215	2021	02	GU
20210215	2021	02	HI
20210215	2021	02	IA
20210215	2021	02	ID
20210215	2021	02	IL
20210215	2021	02	IN

10. Let's **rename** the new columns to year, month, day.

Click on the **date_copy_1** column and select **Rename** from the menu. Enter **year** as the new column name, and click **Apply**

Rename column

Rename column

Info

Rename a column

Source column

Select column to rename

date_copy_1

New column name

New name for the column

year

Valid characters are alphanumeric, underscore, and space

Preview changes

Cancel

Apply

Rename the other two new columns - date_copy_2_1 and date_copy_2_2 - to month and day respectively.

The result should look like the following.

Lab 5. Bonus Lab: Glue DataBrew

covid-states-daily
Dataset: covid-states-daily Sample: First n sample (500 rows)

UNDO REDO FILTER COLUMN FORMAT CLEAN EXTRACT MISSING INVALID DUPLICATES SPLIT MERGE CREATE FUNCTIONS UNNEST PIVOT GROUP JOIN UNION TEXT SCALE MAPPING ENCODE

Viewing 59 columns 500 rows

ABC date	ABC year	ABC month	ABC day	ABC state
20210213	2021	02	13	AK
20210213	2021	02	13	AL
20210213	2021	02	13	AR
20210213	2021	02	13	AS
20210213	2021	02	13	AZ
20210213	2021	02	13	CA
20210213	2021	02	13	CO
20210213	2021	02	13	CT
20210213	2021	02	13	DC
20210213	2021	02	13	DE
20210213	2021	02	13	FL
20210213	2021	02	13	GA
20210213	2021	02	13	GU
20210213	2021	02	13	HI
20210213	2021	02	13	IA
20210213	2021	02	13	ID
20210213	2021	02	13	IL
20210213	2021	02	13	IN

11. The **probableCases** column has some missing values. We will set these missing values to 0.

To navigate to the **probableCases** column, click on the **columns** drop-down list at the top, enter **probableCases** in the search field and click **View**.

58 columns 500 rows

Columns (58/58 shown)
Ctrl+click, Command+click, or Shift+click to select multiple columns.

Q probableCases

Select all | Select first 40 | Select last 40 | Unselect all

ALL (58) ABC STRING (11) # INTEGER (45) DATE (0) OTHER TYPES

☒ # probableCases View

Cancel Delete unselected columns Show selected columns

Click on the **probableCases** column and select **Remove or fill missing values / Fill with custom value**

Lab 5. Bonus Lab: Glue DataBrew

covid-states-daily
Dataset: covid-states-daily | Sample: First n sample (500 rows)

UNDO REDO FILTER COLUMN FORMAT CLEAN EXTRACT MISSING INVALID DUPLICATES SPLIT MERGE CREATE FUNCTIONS UNNEST PIVOT GROUP JOIN UNION TEXT SCALE MAPPING ENCODE

Viewing 58 columns 500 rows

SOURCE

# positive	# probableCases	# negative	# pending	totalTestResultsSour
500	295	393	35	44
Unique 463	Unique 295	Unique 35	Unique 4	Unique 4
Total 500	Total 295	Total 393	Total 35	Total 44
Min 0	Min 207	Min 0	Min 0	Min 0
Median 315.97 K	Median 50.22 K	Median 0	Median 0	Median 0
Mean 483.38 K	Mean 41.27 K	Mean 0	Mean 0	Mean 0
Mode 0	Mode 0	Mode 0	Mode 0	Mode 0
Max 3.39 M	Max 337.24	Max 0	Max 0	Max 0
94.6%				
54282	null			totalTestsViral
479856	103398			totalTestsPeopleViral
312562	64915			totalTestsViral
0	null			totalTestsViral
795323	53357			totalTestsViral
3391036	null			totalTestsViral
412983	21726			totalTestEncountersViri
267337	16422			posNeg
38796	null			totalTestEncountersViri
82672	4213			totalTestEncountersViri
1788827	325309			totalTestEncountersViri
962808	173738			totalTestsViral
7692	208			posNeg
27527	717			totalTestEncountersViri
274386	null			posNeg
167225	31276			totalTestsPeopleViral
1160523	null			posNeg
647657	null			totalTestsViral

Context menu options:

- Rename
- Sort
- Remove or fill missing values
- Remove or replace invalid values
- Remove duplicate values
- Min-max normalization
- Scale between specified values
- Mean normalization
- Z-score normalization or standardization
- Move column
- Hide column
- Duplicate
- Delete

Fill options:

- Remove missing rows
- Fill with custom value
- Fill with null
- Fill with most frequent value
- Fill with last valid value
- Fill with numeric aggregate

Enter 0 as the **Custom value** and click **Apply**

Missing values

Missing values [Info](#)
Modify the column by missing values

Source column
Name of the column with missing values
probableCases

Missing value action
Action to perform on missing values

☐ Delete rows with missing values

☒ Fill with empty value

☐ Fill with null

☐ Fill with last valid value

☐ Fill with most frequent value

☒ Fill with custom value

☐ Fill with numeric aggregate

Custom value
0

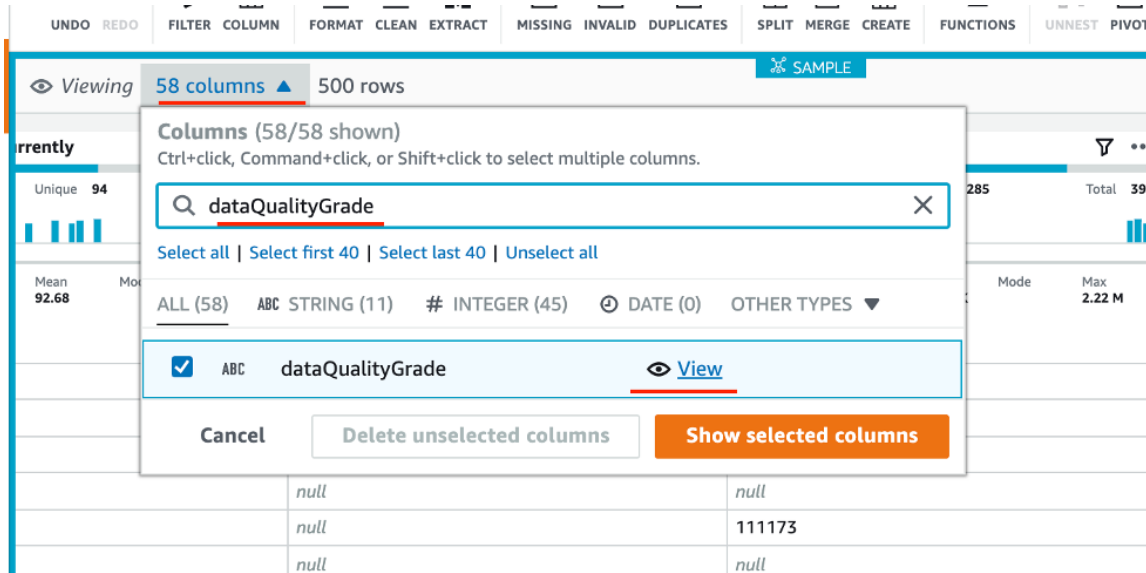
Apply transform to

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

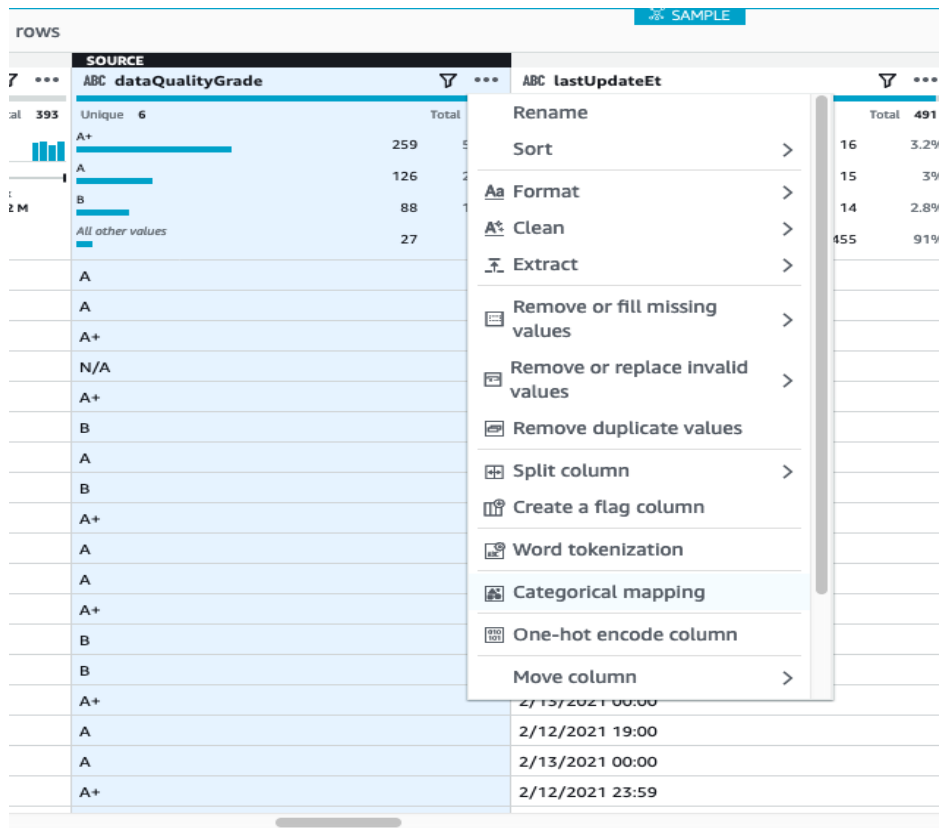
☐ Filtered rows - 0 filters applied (500/500 rows)
Transformation will be applied to filtered rows in the grid

12. Map the values of the **dataQualityGrade** column to **numerical** values.

To navigate to the **dataQualityGrade** column, click on the **columns** drop-down list at the top, enter **dataQualityGrade** in the search field and click **View**.



Click on the **dataQualityGrade** column and select **Categorical mapping**



In the **Categorically map column** dialog

- Select the option **Map all values**
- Enable **Map values to numeric values**
- Map the current dataQualityGrade value to the new value as follows

dataQualityGrade	New value
N/A	0
A+	1
A	2
B	3
C	4
D	5

Categorically map column [X]

Source column
Select a column to perform categorical mapping
dataQualityGrade ▼

Mapping options
☐ Map top 5 values
☒ Map all values (6 values)
☐ Custom map values

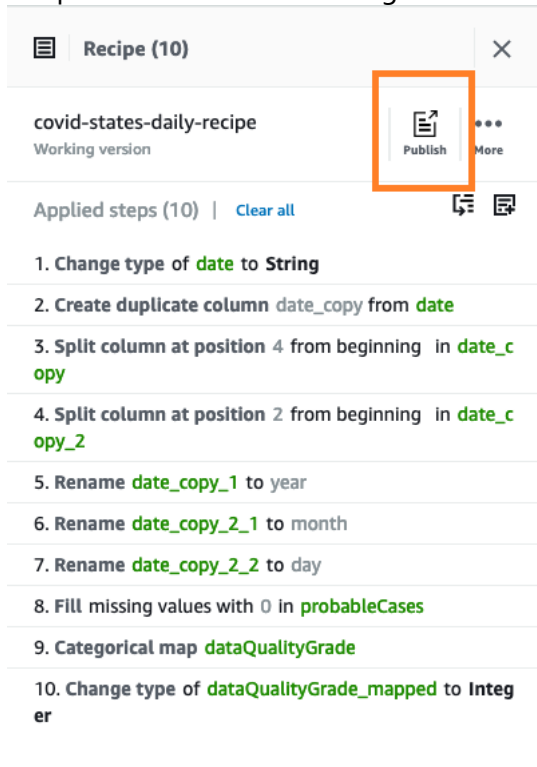
Map values ☒ Map values to numeric values

All values	New value
<input type="checkbox"/> A+ 259 51%	<input type="text" value="1"/>
<input type="checkbox"/> A 126 25%	<input type="text" value="2"/>
<input type="checkbox"/> B 88 17%	<input type="text" value="3"/>
<input type="checkbox"/> N/A 9 1%	<input type="text" value="0"/>

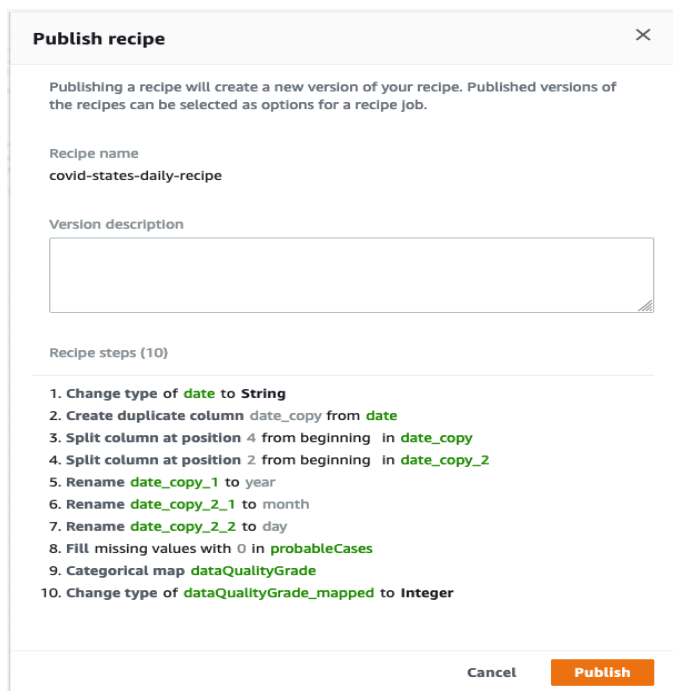
Leave all other settings as default. Click **Apply**

13. After this transform, the new column **dataQualityGrade_mapped** is of type **double**, convert this column to integer.

14. You are now ready to publish the recipe so that it can be used in DataBrew jobs. The final recipe looks like the following.



15. Click on the **Publish** button at the top of the recipe. Optionally enter a version description, and click **Publish**. The recipe is published as Version 1.0. DataBrew applies a version number when a recipe is published.



Creating a DataBrew job

1. Click on **Jobs** from the menu on the left-hand side of the DataBrew console.
2. On the **Recipe jobs** tab, click on **Create job**. Enter **covid-states-daily-prep** for the job name.
3. Select **Create a recipe job**. Choose the **covid-states-daily** dataset and select the 'covid-states-daily-recipe'.

DataBrew > Jobs > Create job

Create job [Info](#)

Job details

Job name
Identifier for the jobs

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

Job type

Type of job to run on the dataset

☒ **Create a recipe job**
Runs the transformations from the associated recipe on the population of the associated dataset.

☐ **Create a profile job**
Generates summary and statistics that give you the shape of your data.

Job input [Info](#)

The input dataset for the job and the recipe to be applied to it.

Run on

☒ **Dataset**
Run the job on an existing or new DataBrew dataset.

☐ **Project**
Run the job on a project with no associated job.

Select a dataset

 ×

Select a recipe

 ×

Recipe version

 ▼

4. In the **Job output settings** section, enter the S3 location **s3://databrew-lab-databrewoutputs3bucket-xxxxx/job-outputs/**.

Expand the **Additional Configuration - optional** panel.

Under **Custom partition by column values** add year, month and day columns. This will partition the data in the output folder by year, month and day.

Job output settings [Info](#)
Running a job generates output files at specified file destinations.

File type
Output format: CSV ▼

S3 location
Format is: s3://bucket/folder/
s3://databrew-lab-databrew [Browse](#)

Compression
Available types: None ▼

Delimiter
CSV separator: Comm... ▼

[Add another output](#)

Additional configuration - optional

Custom partition by column values
Partition by unique values of columns. The file is partitioned and stored in a folder path based on the order of columns provided.
Example: A file partitioned by Column A and Column B is stored at this S3 path: s3://output file path.../Column A/Column B/

- year X
- month X
- day X

[Enter a column name](#) [Add](#)

- In the **Permissions** section, select the role **databrew-lab-DataBrewLabRole-xxxxxx**. Click **Create and run job**.

Permissions [Info](#)
DataBrew needs permission to connect to data on your behalf. Use an IAM role with the [required policy](#) attached.

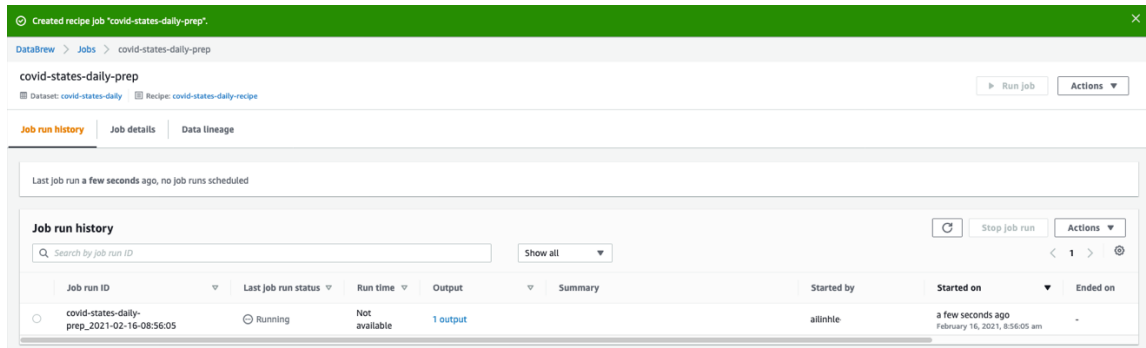
Role name
Choose the role that has access to connect to your data. Refresh to see the latest updates.

databrew-lab-DataBrewLabRole-31403K1MT6ZM [Refresh](#)

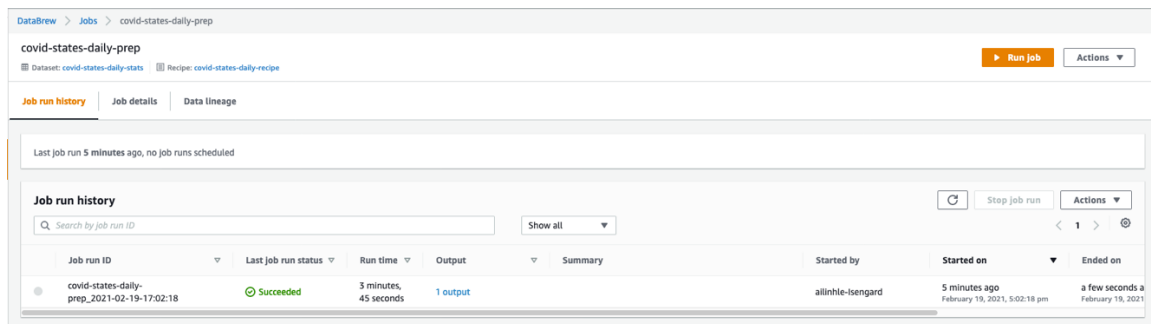
[Cancel](#) [Create job](#) [Create and run job](#)

- The DataBrew job is created and the job **status** is **Running**.

Lab 5. Bonus Lab: Glue DataBrew



- Wait until the job has completed successfully (approx. 4 minutes)



- Click on the link to the job output, and verify that the output files are partitioned in the S3 bucket

Viewing data lineage

- In DataBrew, navigate back to the **covid-states-daily** project. Click on **Lineage** at the top right.

This view shows the origin of the data and the transformation steps that the data has been through.



Congratulations, you have completed the DataBrew lab. If you haven't already done so, you can return to step 13 to examine the data profile.