Amazon Web Services
Data Engineering Immersion Day

Lab 6. Bonus Lab: Glue DataBrew
July 2021

Table of Contents

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

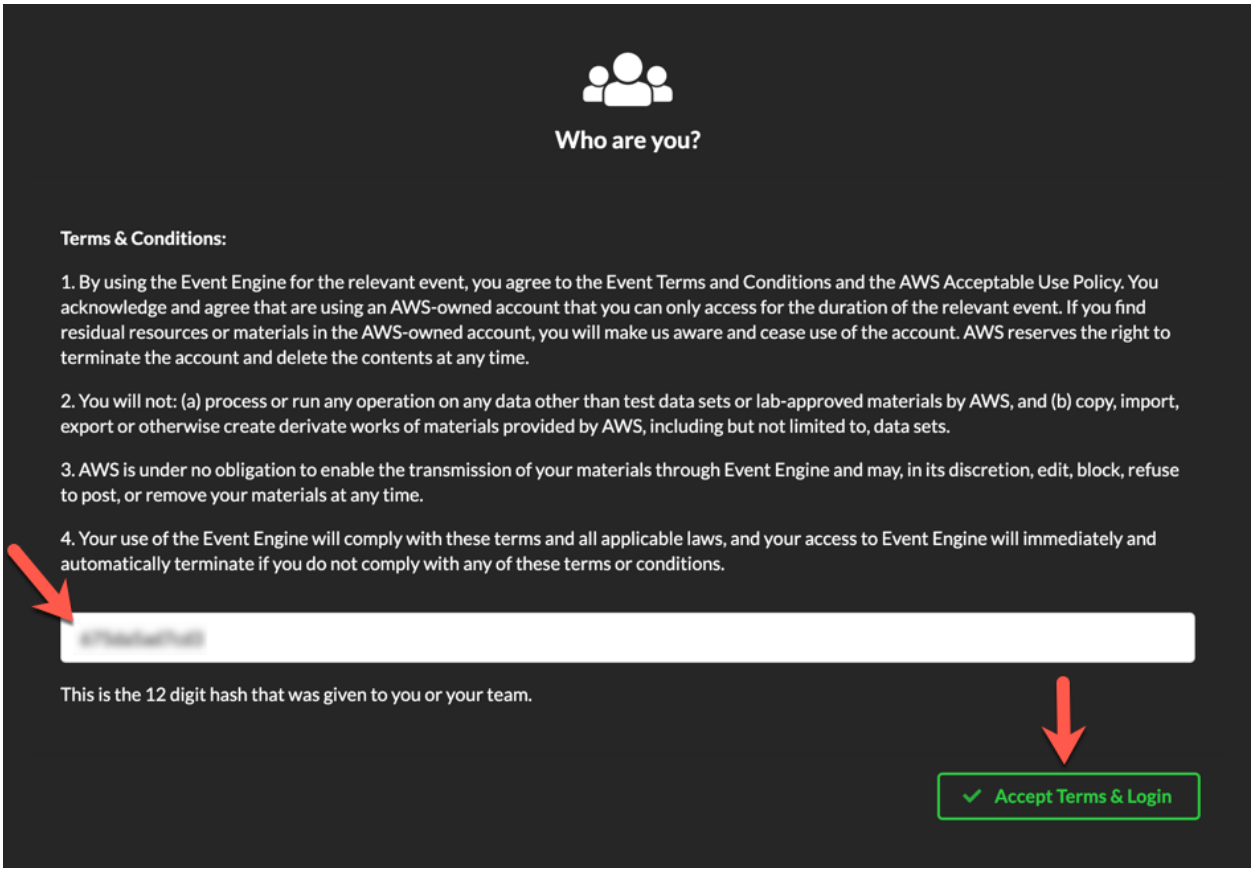
Get Started Using the Lab Environment

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.

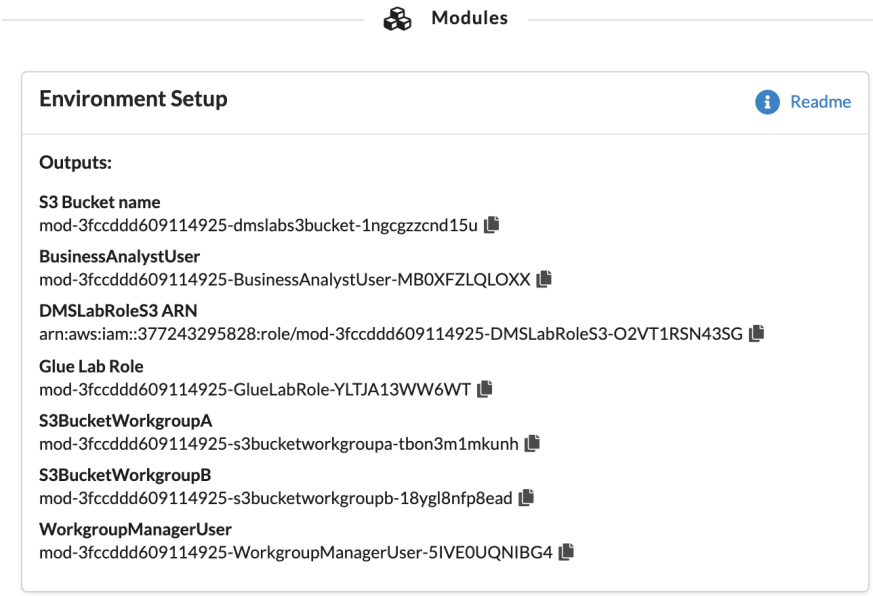
A 12-character access code (or 'hash') is the access code that grants you permission to use a dedicated AWS account for the purposes of this workshop.

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

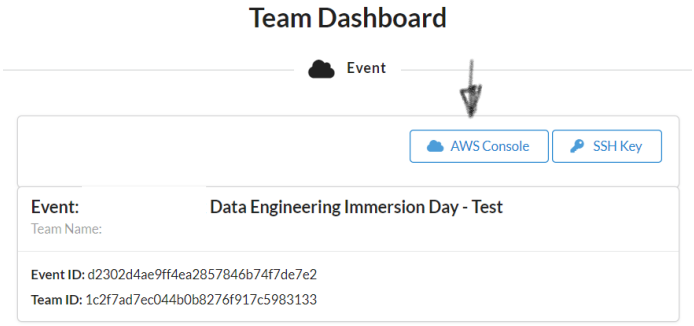


2. On the Team Dashboard web page you will see a set of parameters that you will need during the labs. Best to save them to a text file locally, alternatively you can always go to this page to review them. Replace the parameters with the corresponding values from here where indicated in subsequent labs:

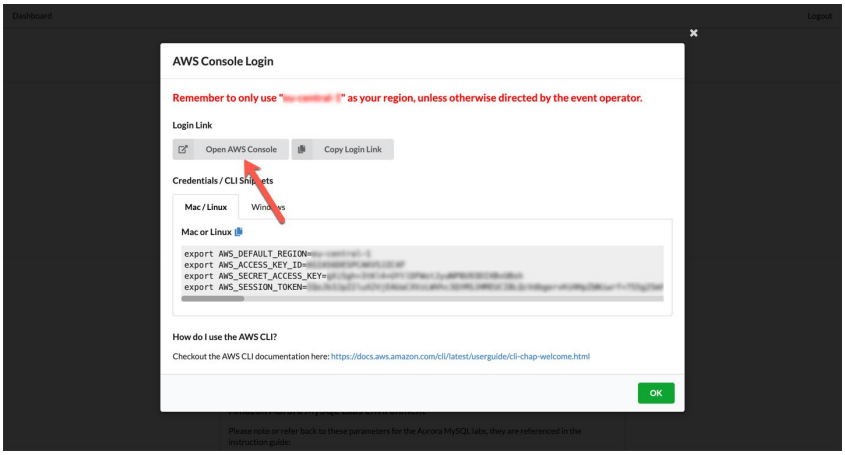
Because you're at a formal event, some AWS resources have been pre-deployed for your convenience, for example you can see a list of parameters on your event dashboard:



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



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



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

DataBrew - Pre-Lab Setup

** Skip this if you are attending an AWS event. As it's already deployed for you **

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.

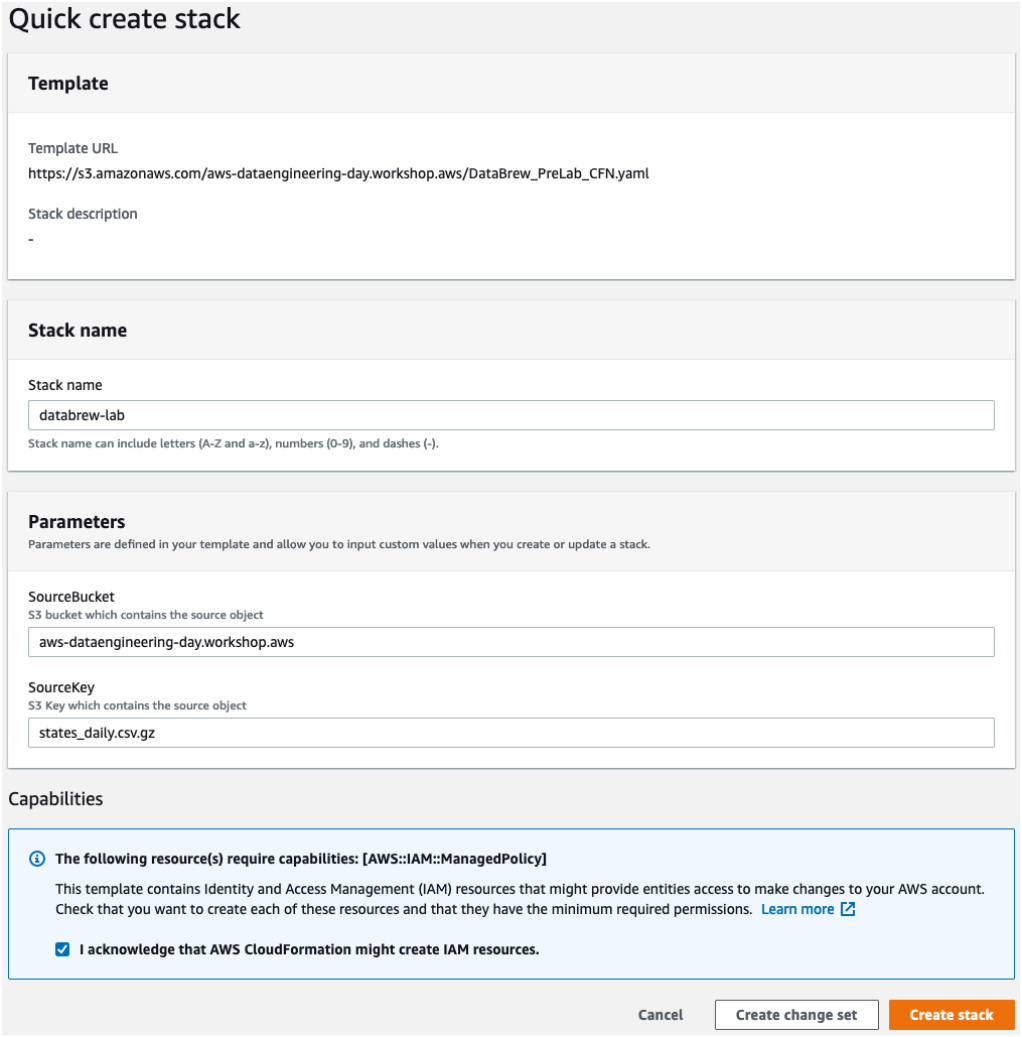
To do this, we will first set up an IAM role to use in DataBrew, and an S3 bucket for the results from the DataBrew jobs.

CloudFormation Stack Deployment

Choose the same region as where you are running the whole workshop

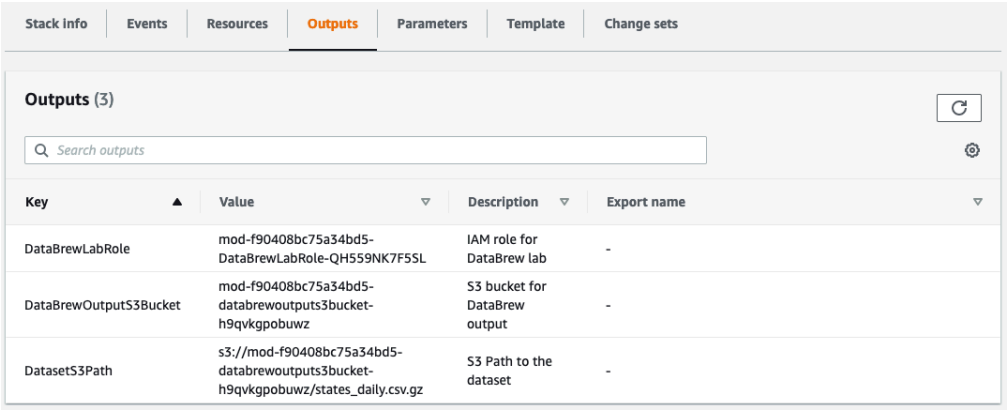
1. Click the **Deploy to AWS** icon below to create the AWS resources for the lab.

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

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



Note the values for **DatasetS3Path**, **DataBrewLabRole** and **DataBrewOutputS3Bucket** which will be used in the lab.

Congratulations! You are all done with the CloudFormation deployment.

Data preparation with Glue DataBrew

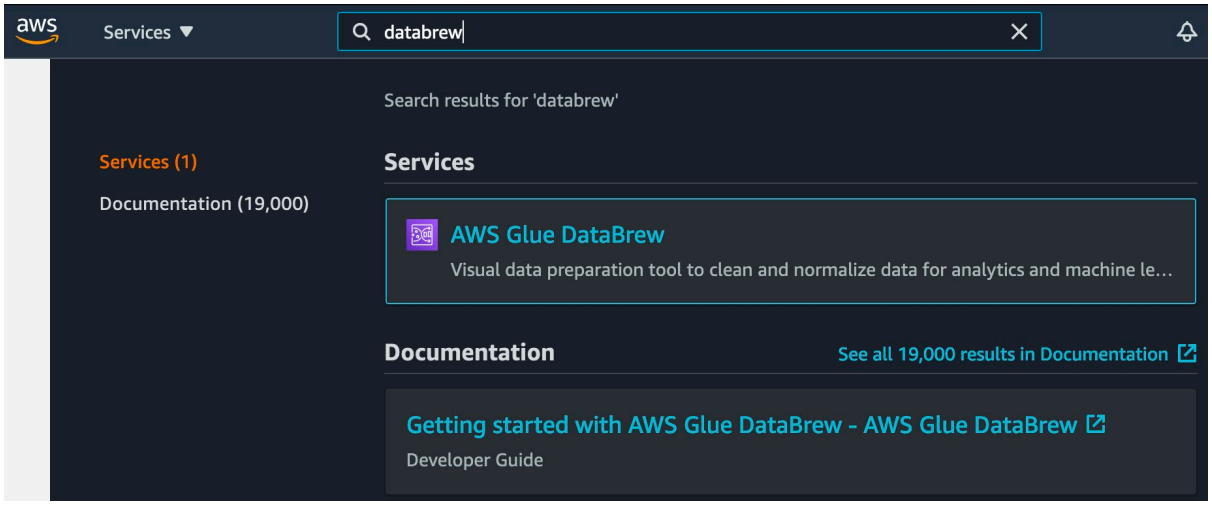
In this lab you will be completing the following tasks.

Tasks Completed in this Lab:

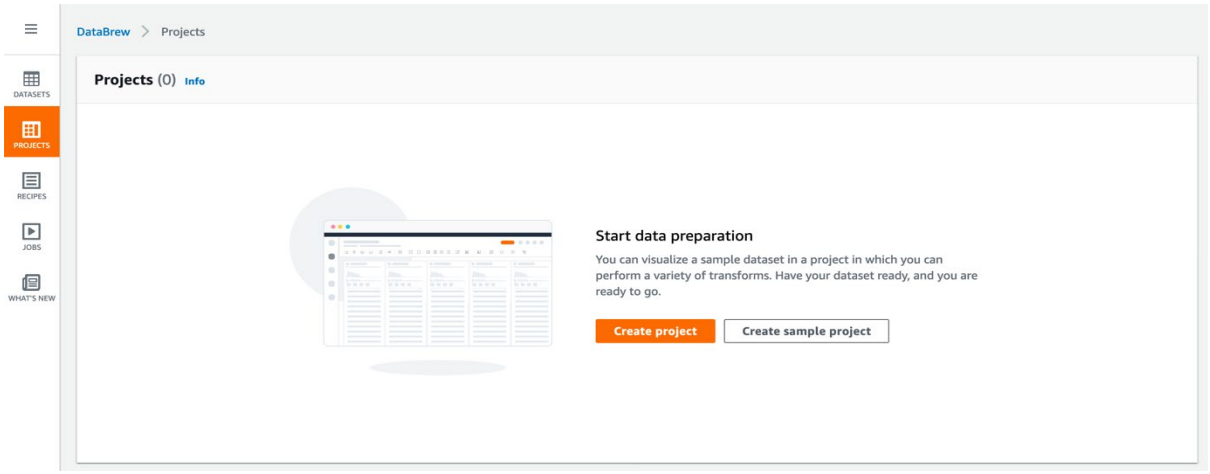
- Create a Glue DataBrew project to explore a dataset
- Connect a sample dataset from S3
- Explore the dataset in Glue DataBrew
- Generate a rich data profile for the dataset
- Create a recipe and job to clean and normalize data in the dataset

Creating a project

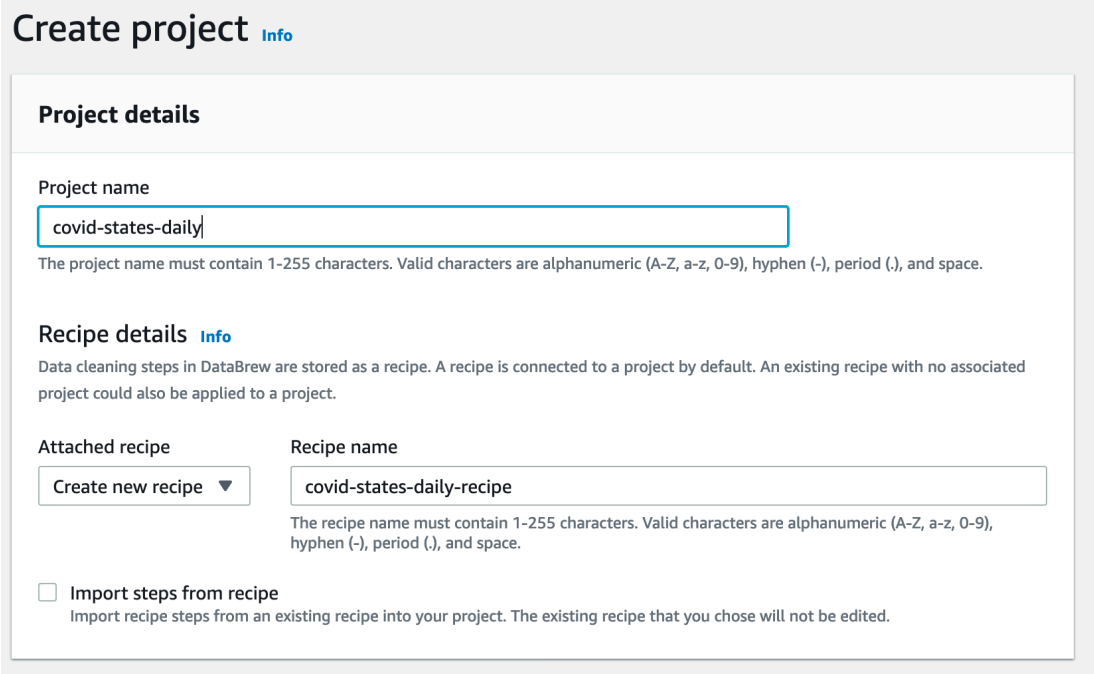
1. Navigate to the AWS Glue DataBrew service



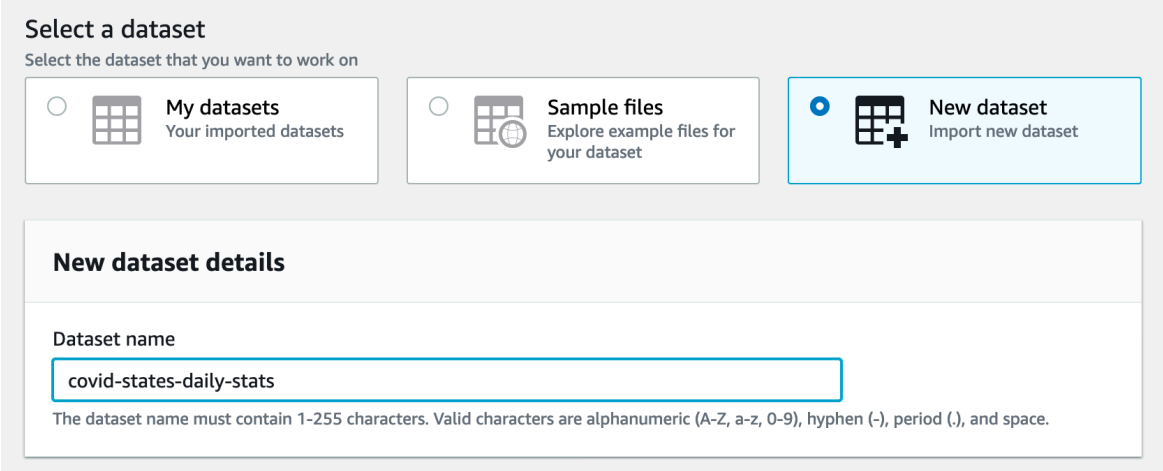
2. On the DataBrew console, select Projects



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

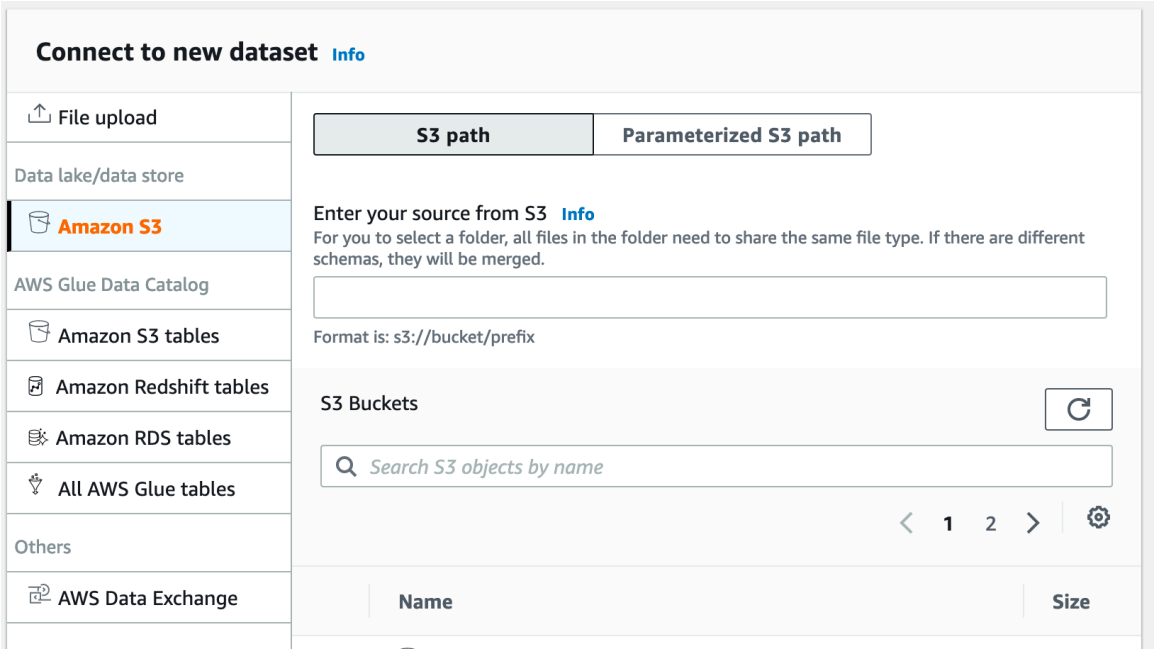


5. In the **Select a dataset** section, select **New dataset** and enter covid-states-daily-stats

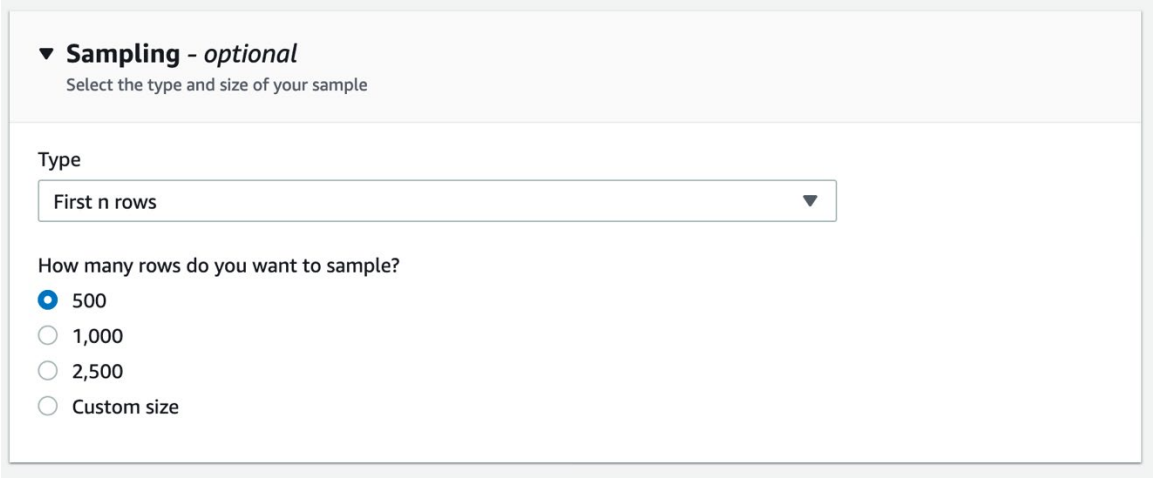


6. In the **Connect to a new dataset** section, select **Amazon S3** under "Data lake/data store"

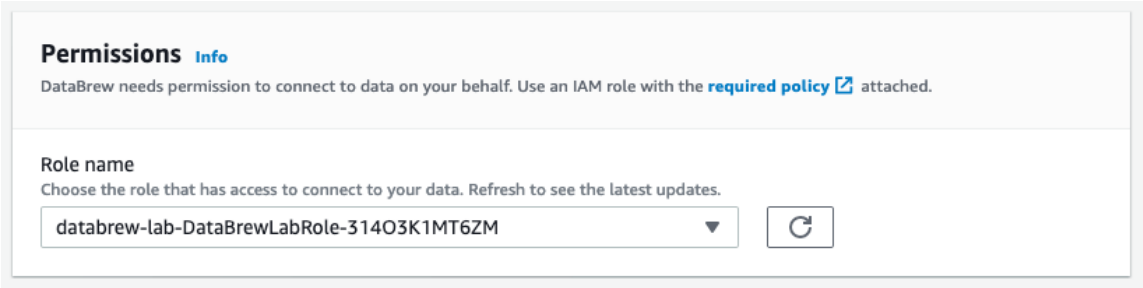
   Enter the **DatasetS3Path** that is available in Event Engine Team Dashboard or outputs section of your CloudFormation stack



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

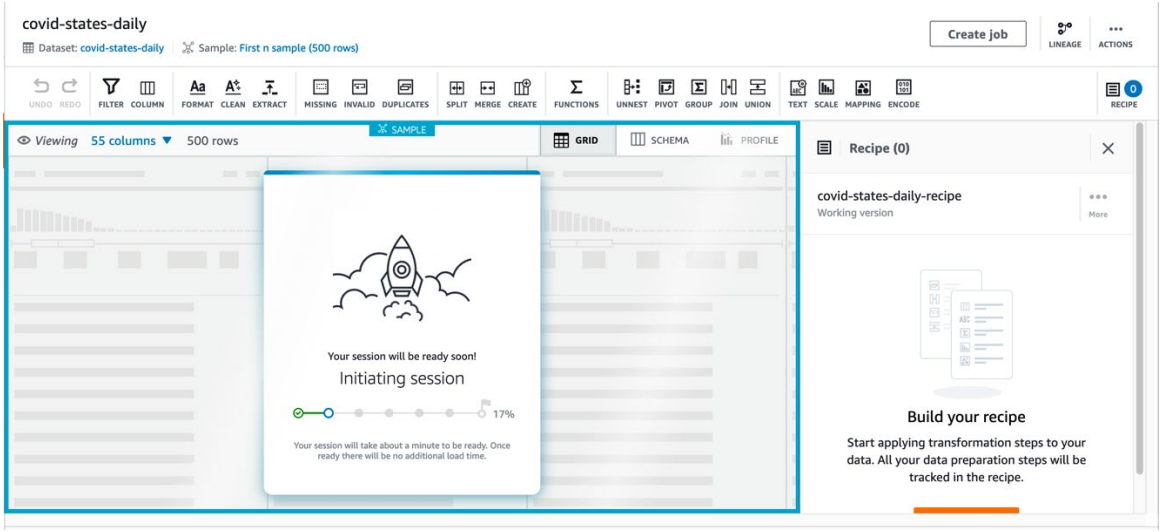


8. In the **Permissions** section, select the role xxxx-DataBrewLabRole-xxxxx from the drop-down list



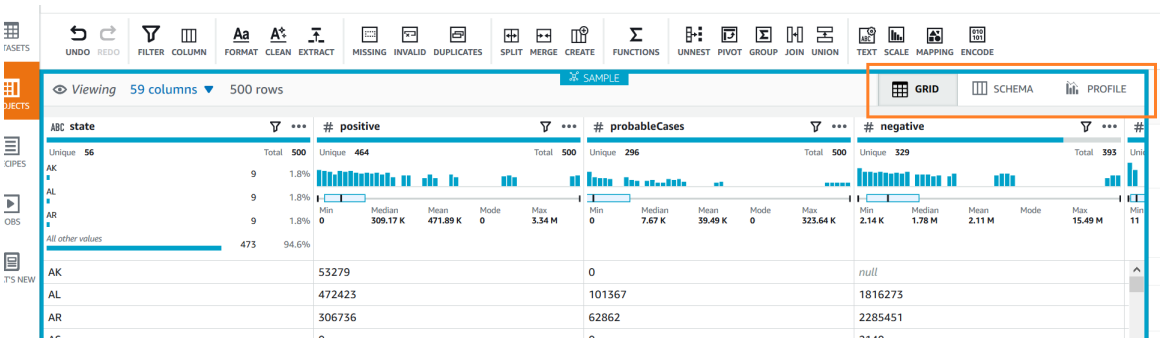
9. Click **Create project**

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



Exploring the dataset

10. 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

11. 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

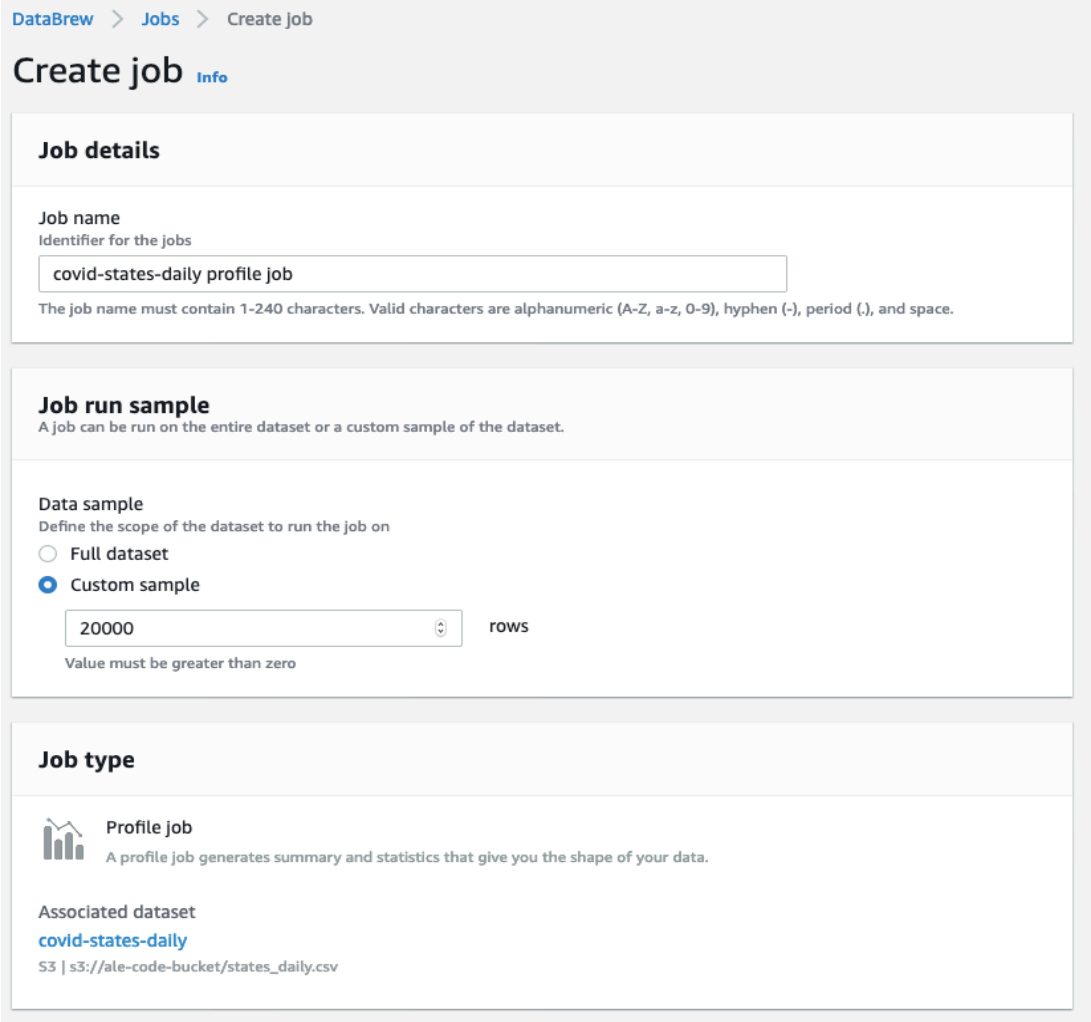
- Select the checkbox next to a column to view the summary of statistics for the column values
- Show/Hide columns
- Rename columns
- Change the data type of columns
- Rearrange the column order by dragging and dropping the columns

12. Click on the **Profile** tab

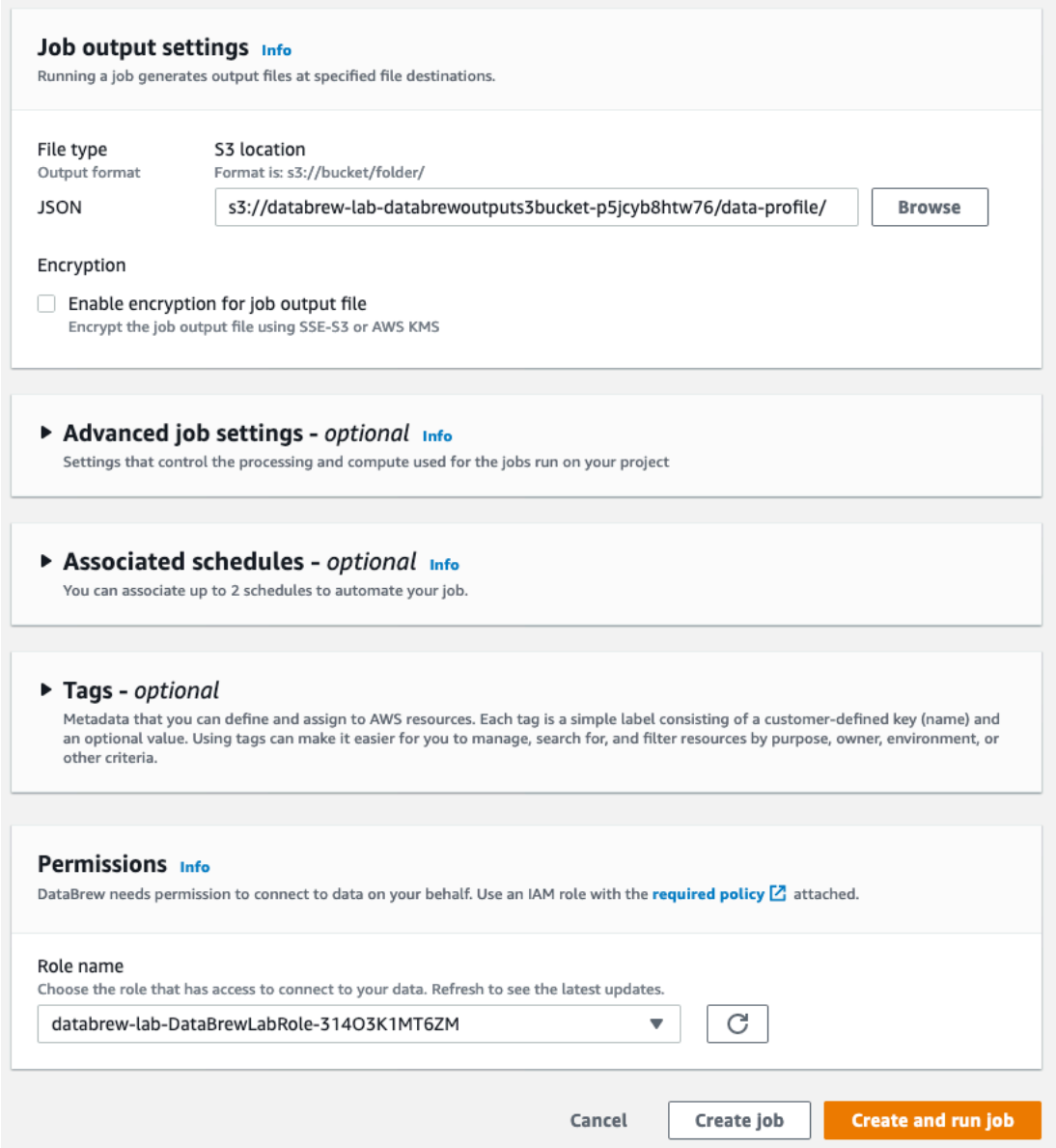
In the Profile view, you can run a data profile job to examine and collect statistical summaries about the data. A data profile is an assessment in terms of structure, content, relationships, and derivation.

Click on **Run data profile.**

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



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)



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

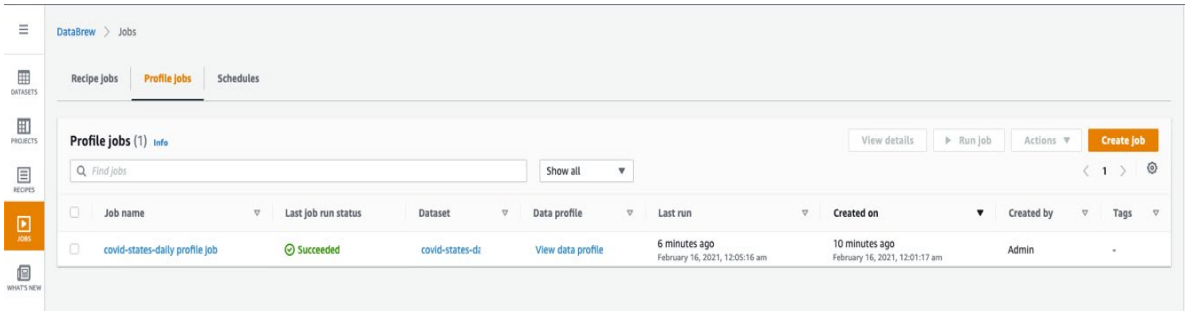
e) Click **Create and run job**

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.

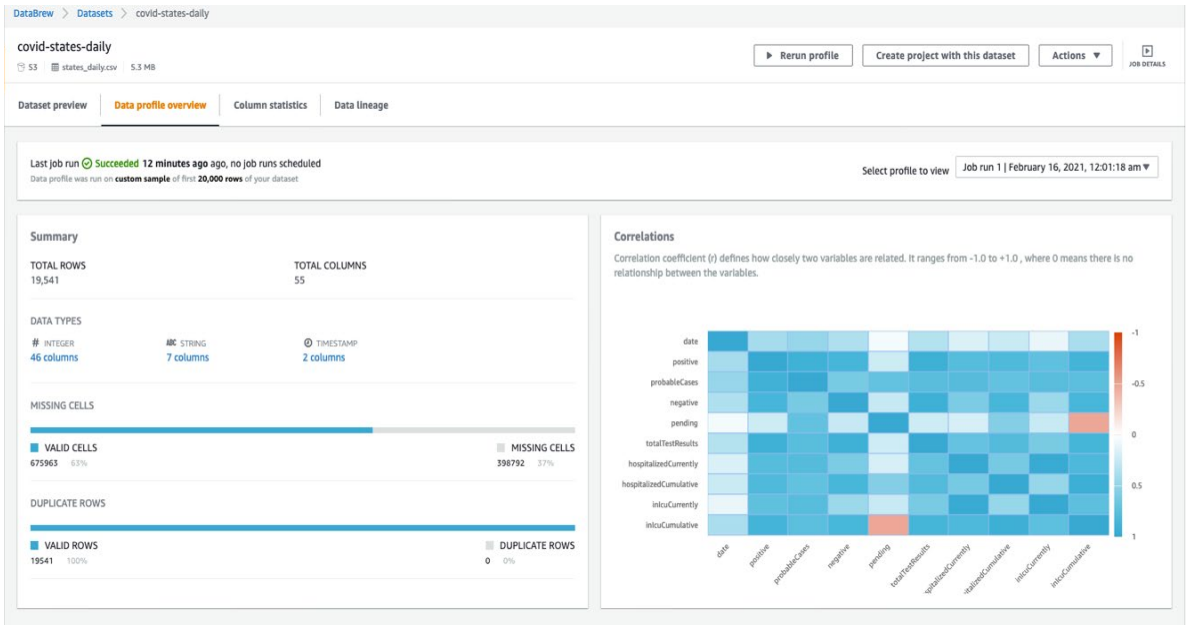
13. Click on **Jobs** from the menu on the left-hand side of the DataBrew console.

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

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



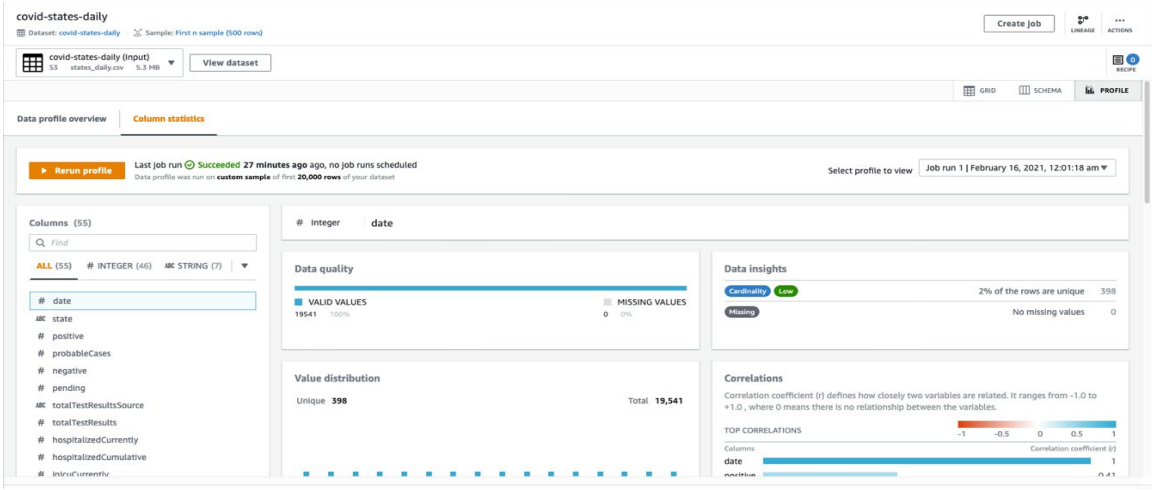
When the profile job has successfully completed, click on **View data profile.**



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.



Preparing the dataset

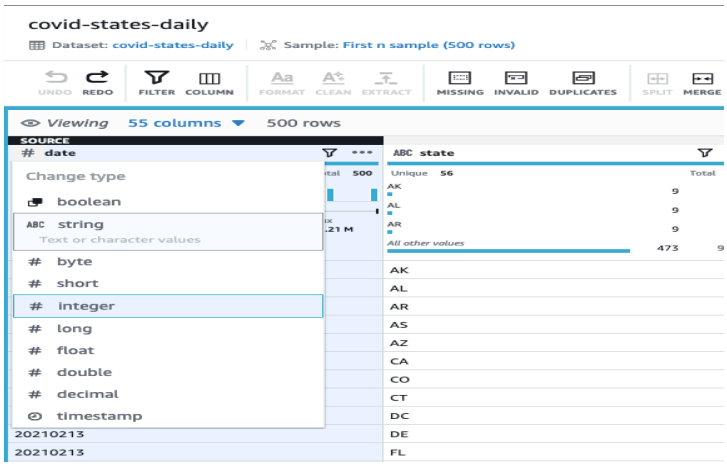
In this section, we will apply the following transformations to the dataset.

- Convert the date column to 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

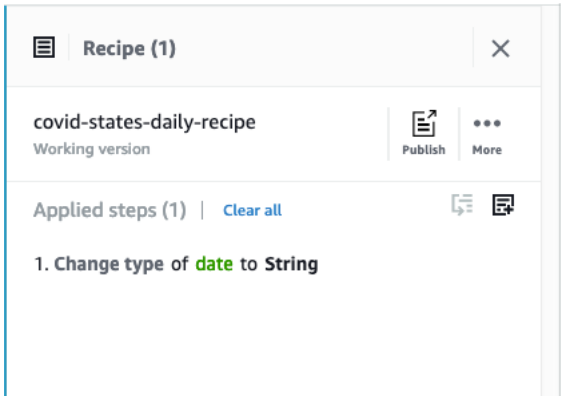
1. Navigate back to the covid-states-daily project grid view.

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**

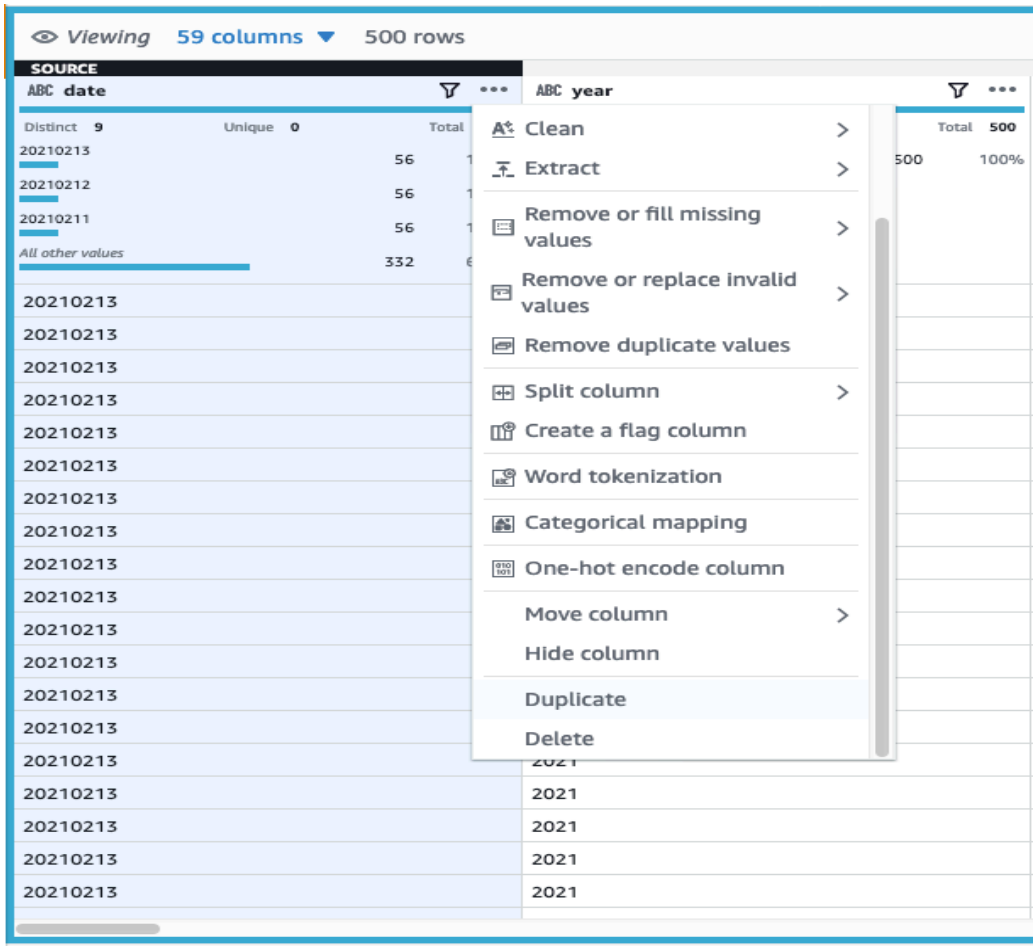


   Note that the transformation is added to the recipe at the right.

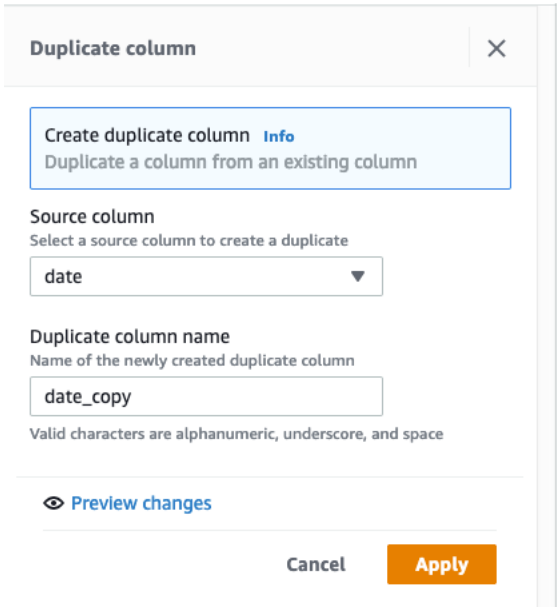


4. We will duplicate the date column first before splitting it into year, month, day columns, as the original column will be deleted by this transformation.

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



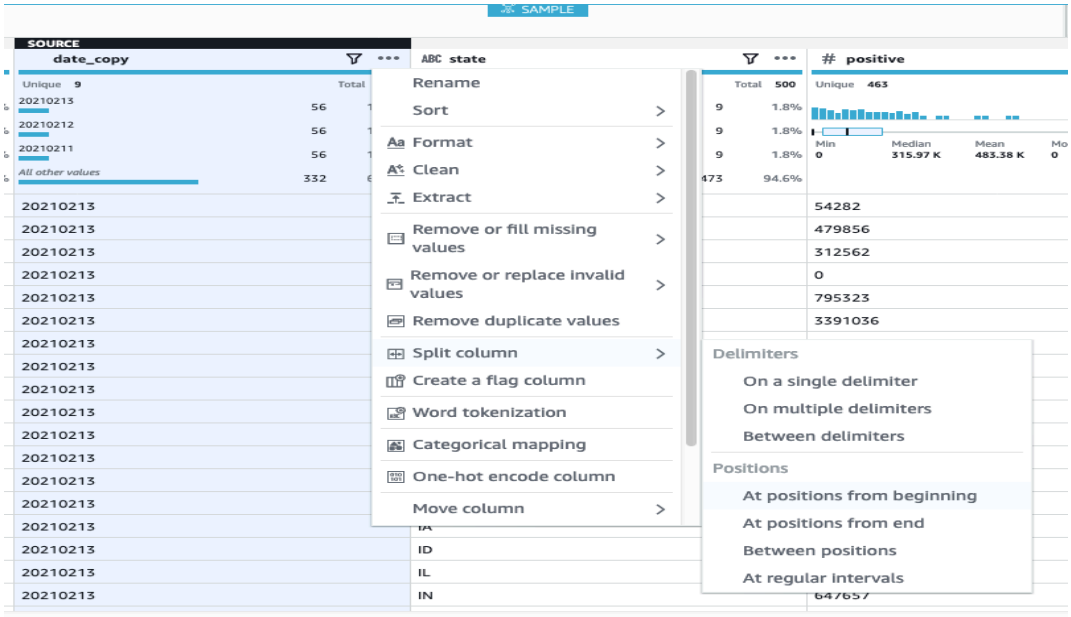
5. Leave the default settings in the **Duplicate column** dialog, click **Apply**



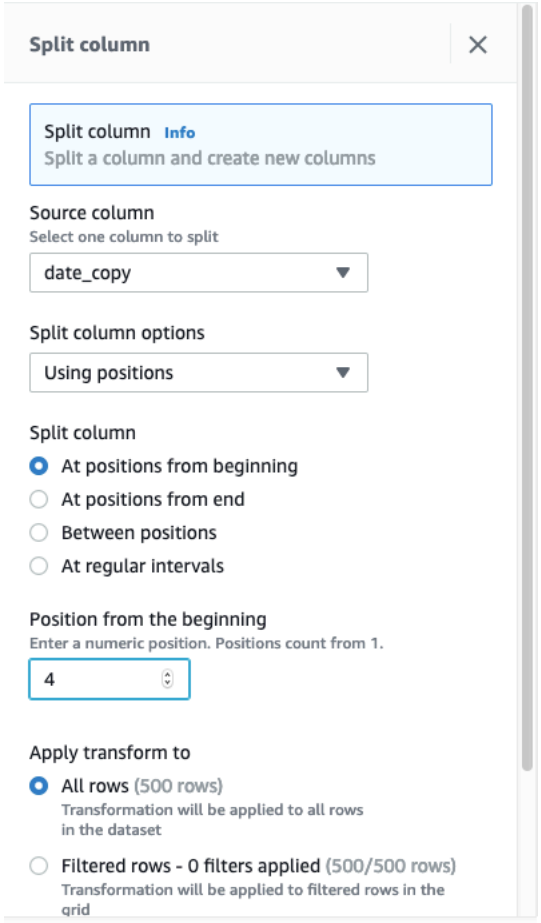
A copy of the date column is created with the name date_copy. Note that the **duplicate column** transformation is added as a step to the recipe at the right.

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

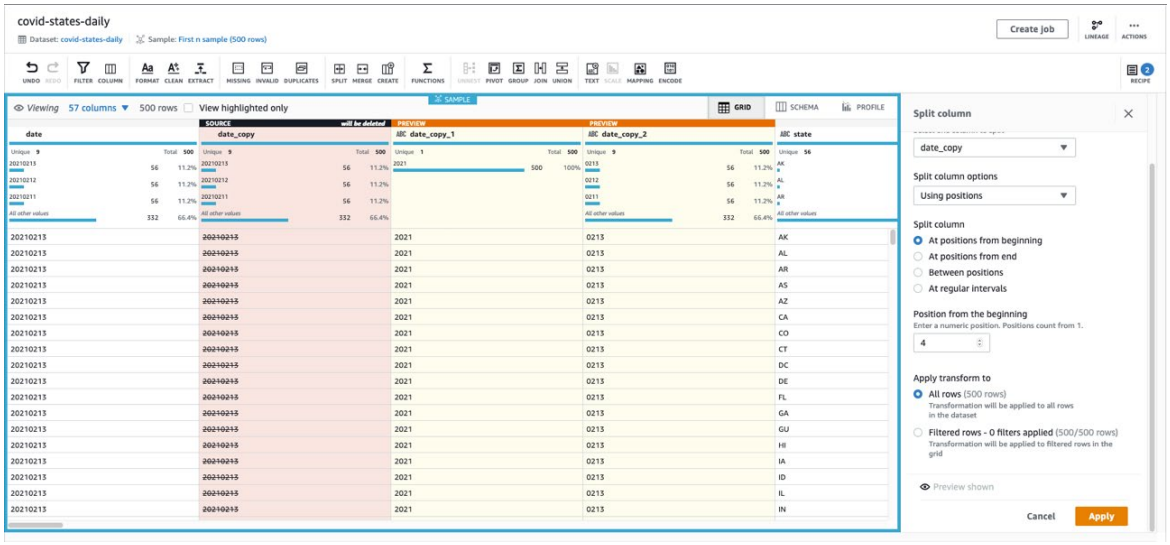
   Select the **...** at the top of the date_copy column. Select **Split column** / **At positions from beginning**



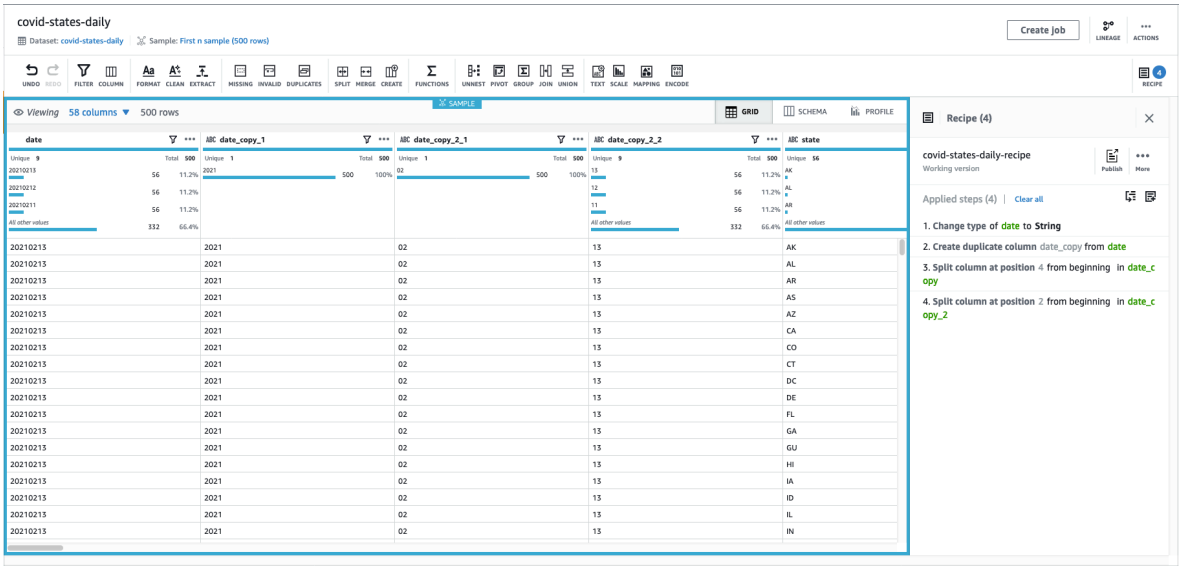
7. In the **Split column** dialog, enter 4 for **Position from the beginning** to split out the year. Leave all other default settings.



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.**

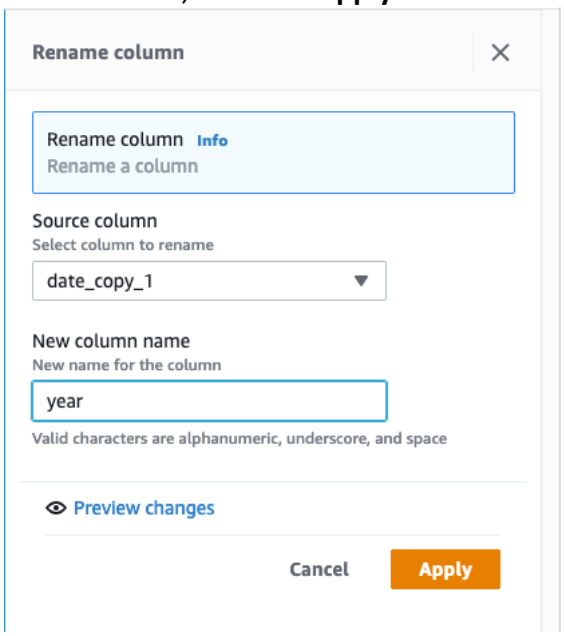


9. Next, **split** the **date_copy_2** column into month and day. The result should look like the screenshot below.



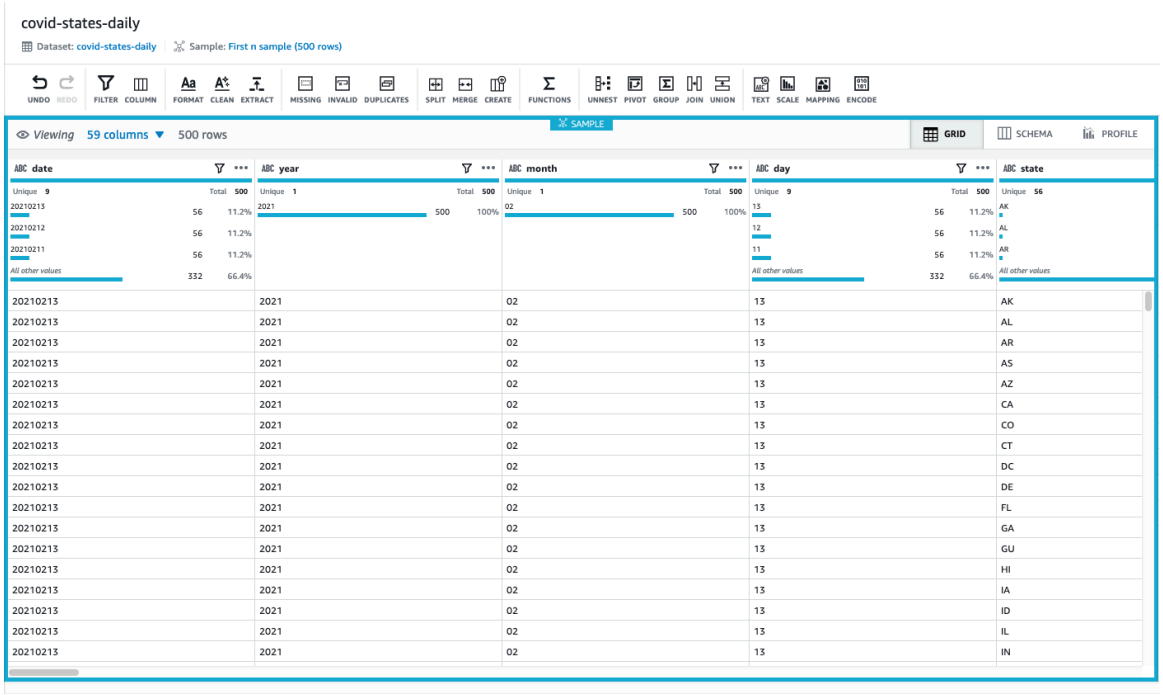
10. Let's **rename** the new columns to year, month, day.

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



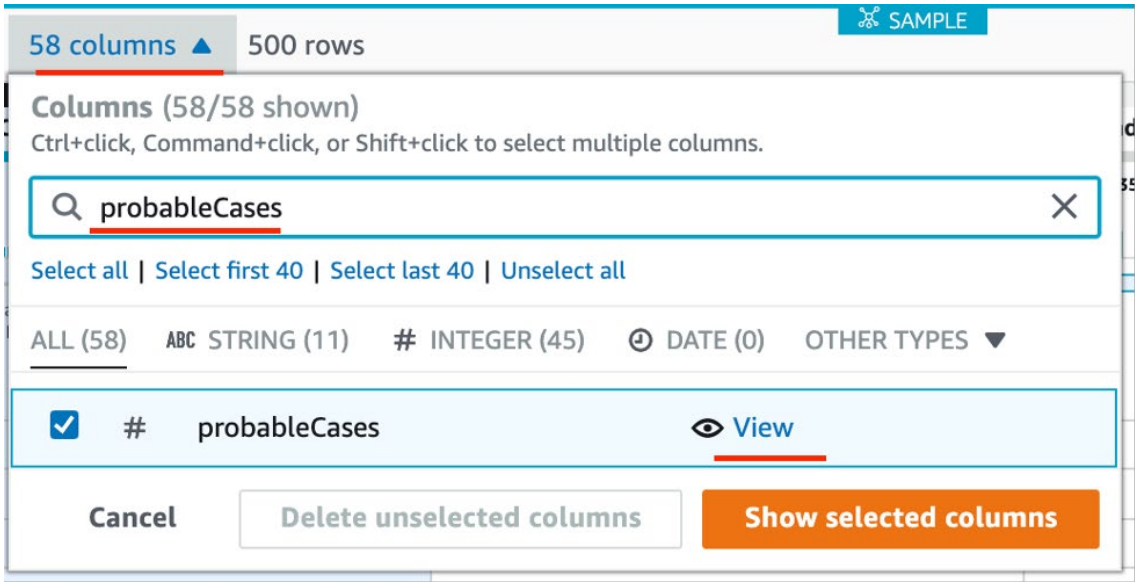
    Rename the other two new columns - date_copy_2_1 and date_copy_2_2 - to **month** and **day** respectively.

    The result should look like the following.

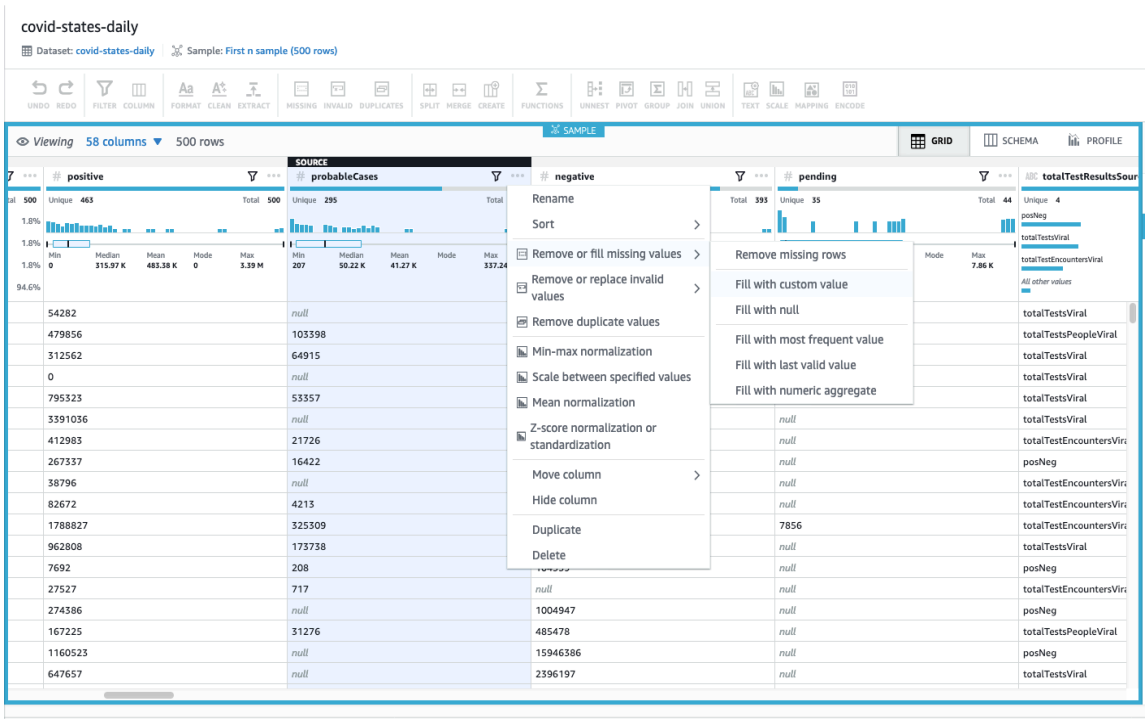


11. The **probableCases** column has some missing values. We will set these missing values to **0**.

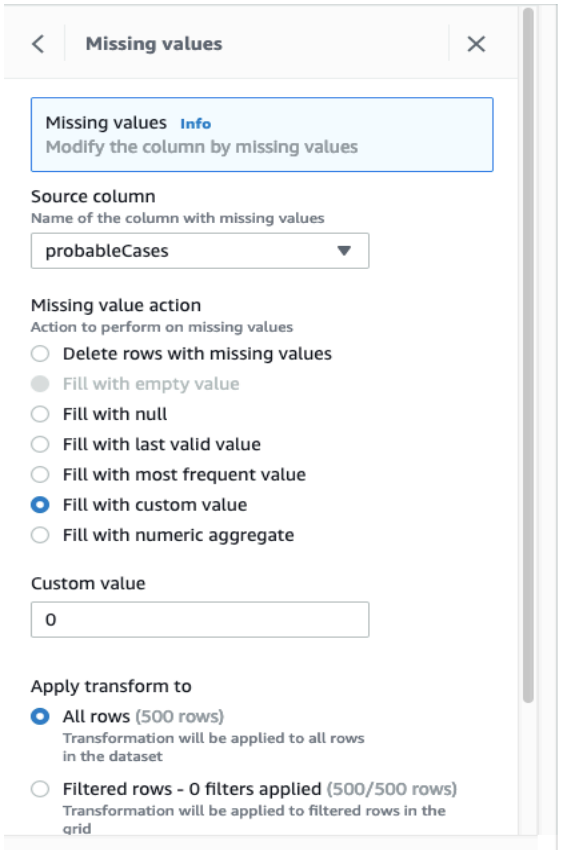
    To navigate to the probableCases column, click on the **columns** drop-down list at the top, enter **probableCases** in the search field and click **View**.



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

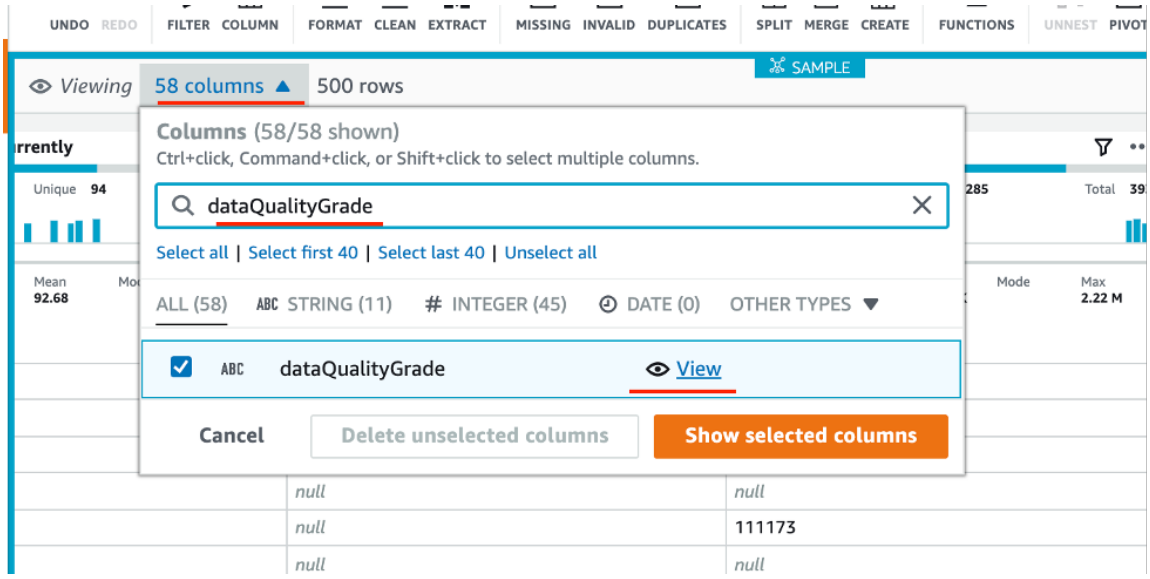


Enter 0 as the **Custom value** and click **Apply**

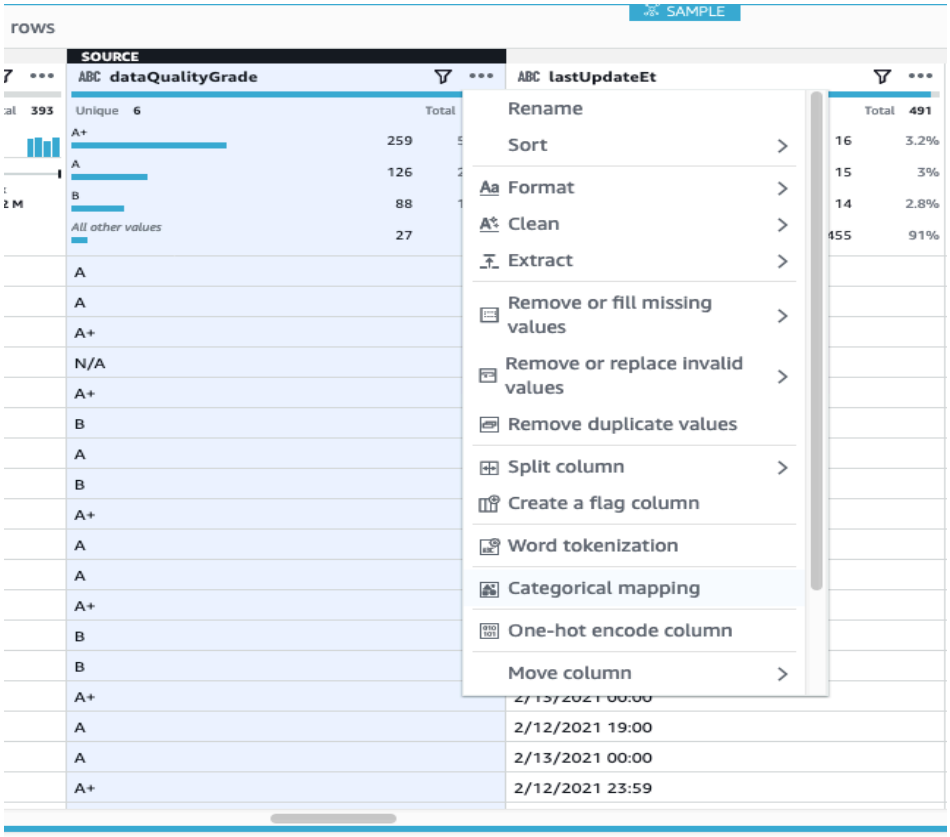


12. Map the values of the **dataQualityGrade** column to **numerical** values.

    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**.



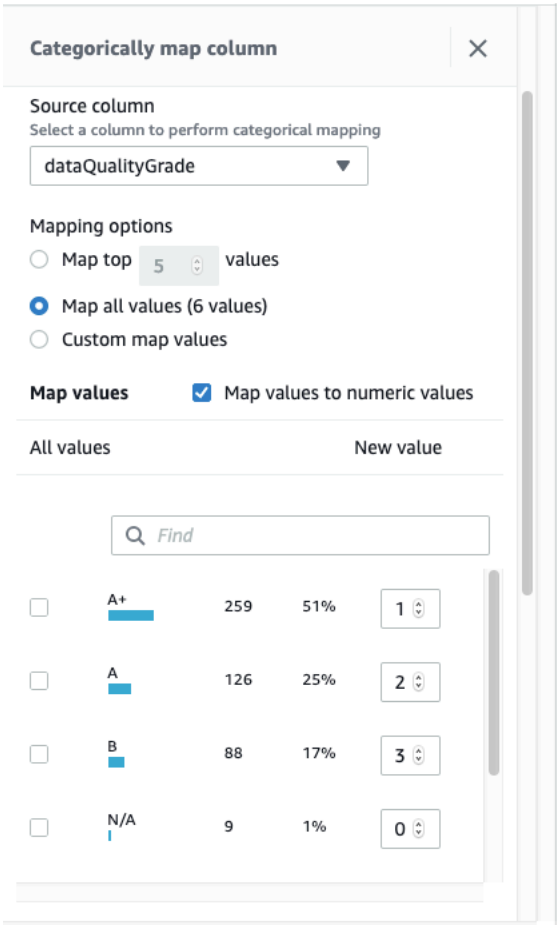
    Click on the **dataQualityGrade** column and select **Categorical mapping**



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

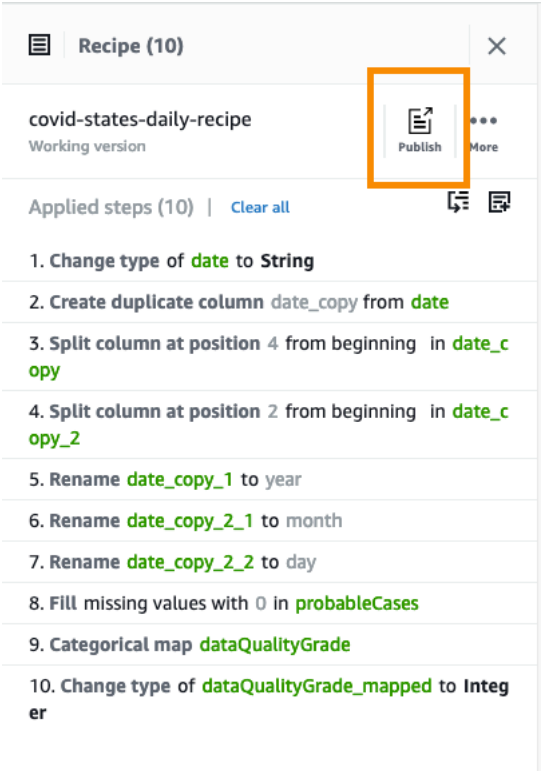
dataQualityGrade	New value
N/A	0
A+	1
A	2
B	3
C	4
D	5



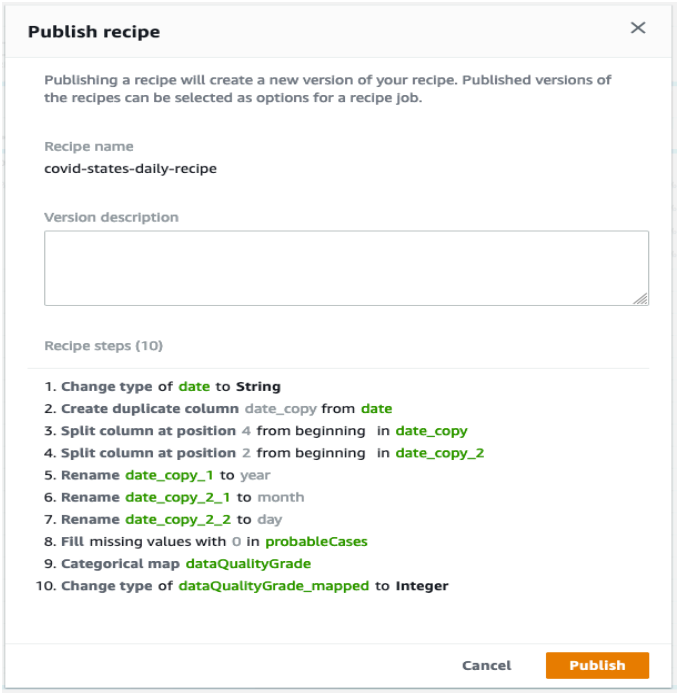
Leave all other settings as default. Click **Apply**

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.

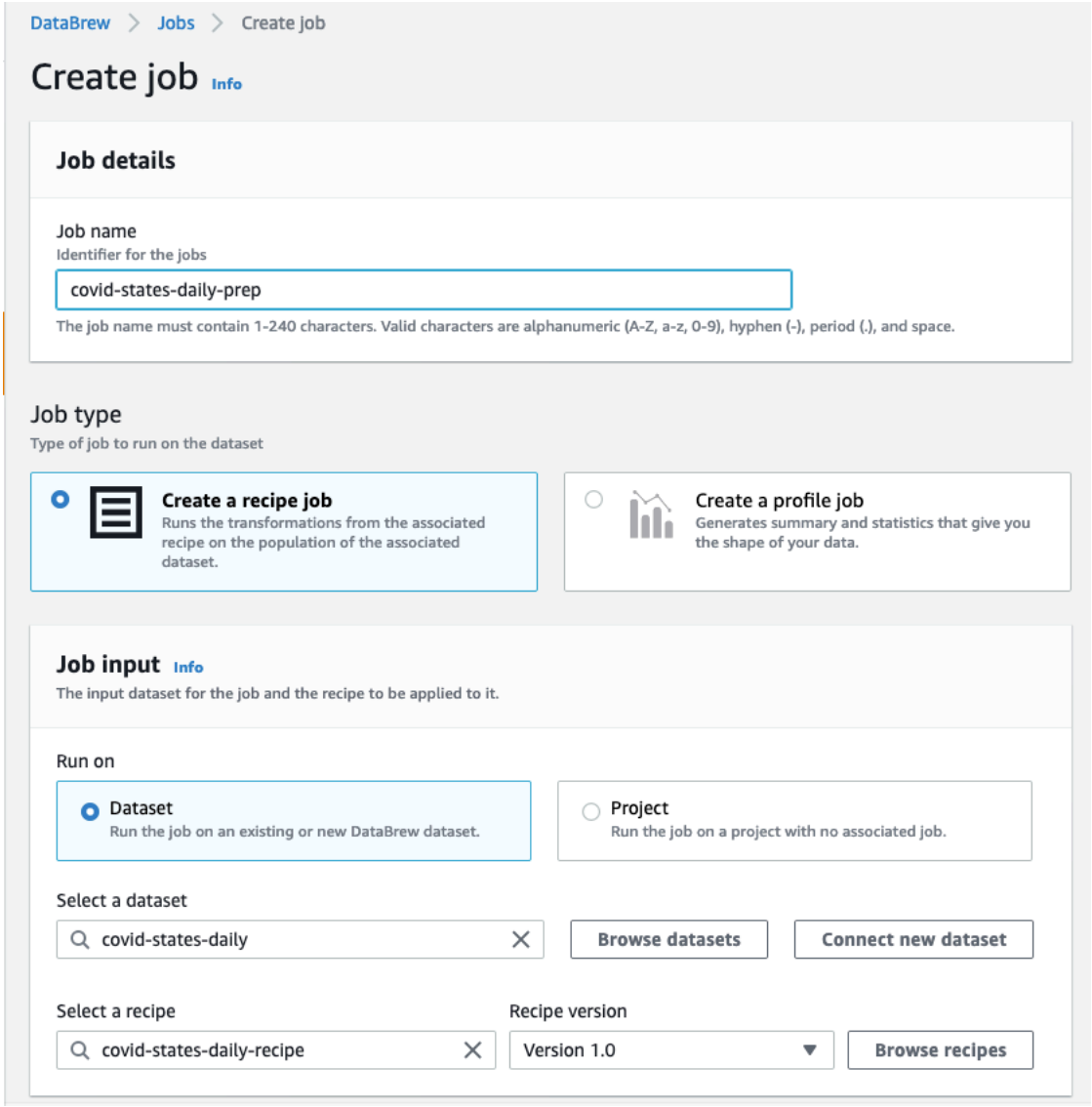


Creating a DataBrew job

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

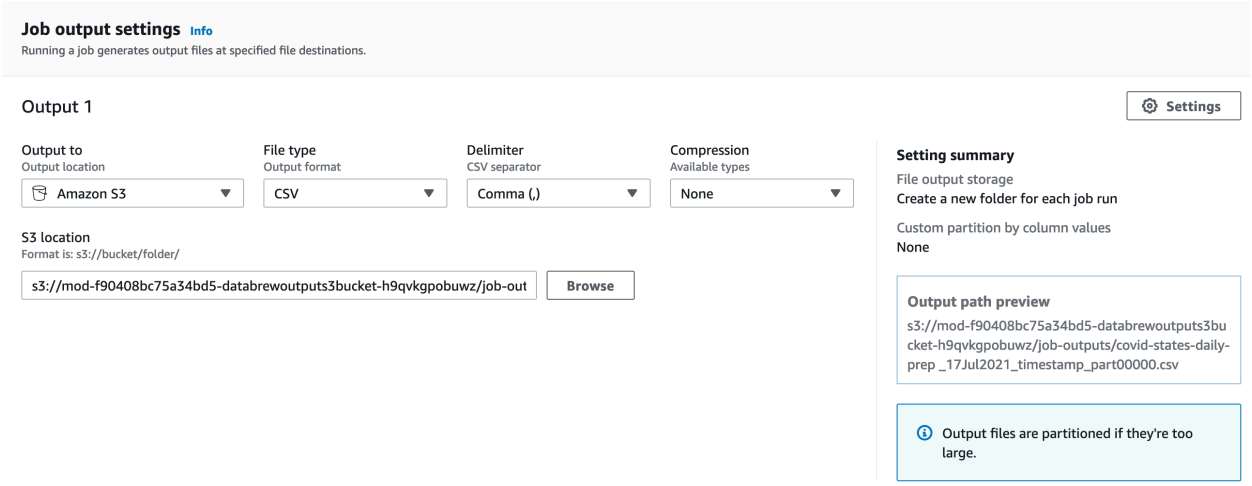
2. On the **Recipe jobs** tab, click on **Create job**. Enter **covid-states-daily-prep** for the job name.

3. Select **Create a recipe job**. Choose the **covid-states-daily** dataset and select the '**covid-states-daily-recipe'.**

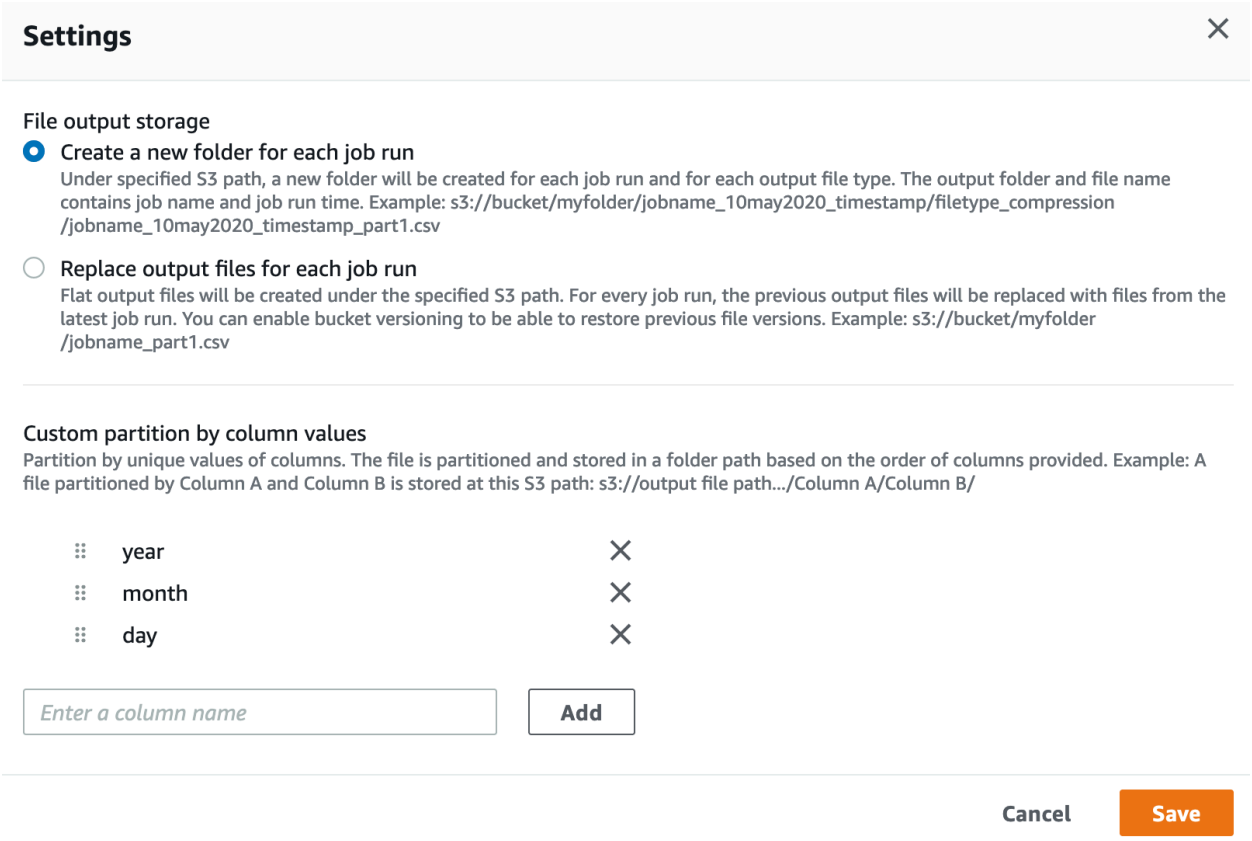


4. In the **Job output settings** section, enter the S3 location **s3://databrew-lab-databrewoutputs3bucket-xxxxx/job-outputs/**.

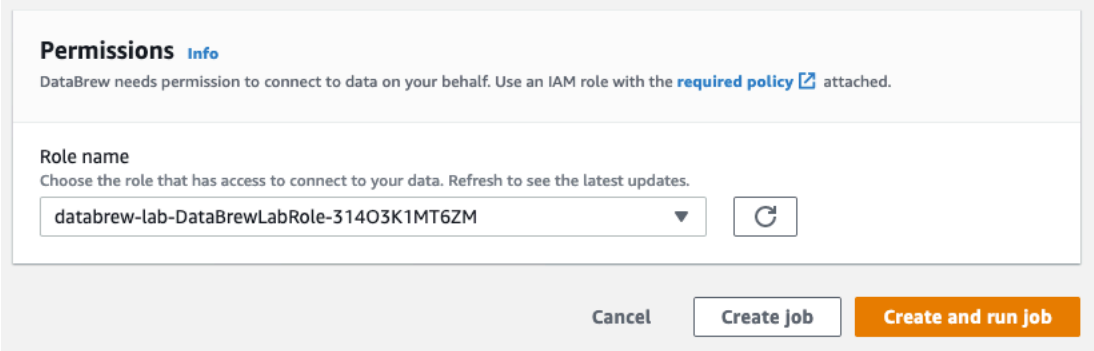
   Click **Settings** under **Job Output Settings**



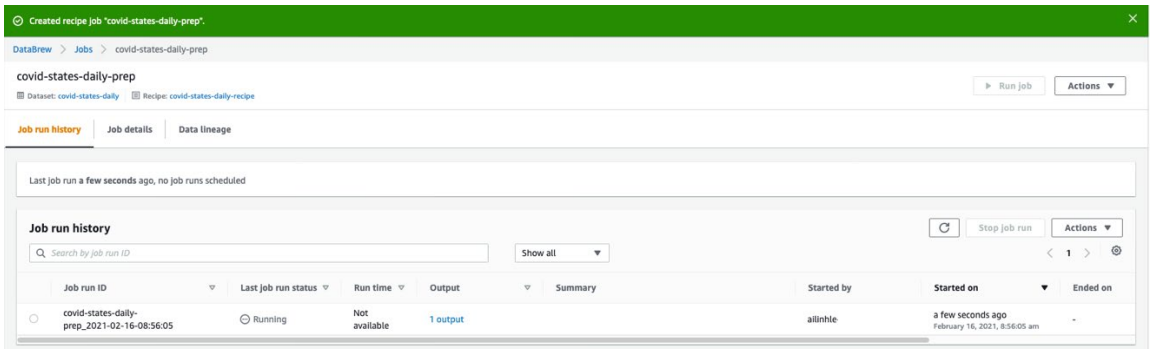
Under **Custom partition by column values** add year, month and day columns. This will partition the data in the output folder by year, month and day, select **Save**.



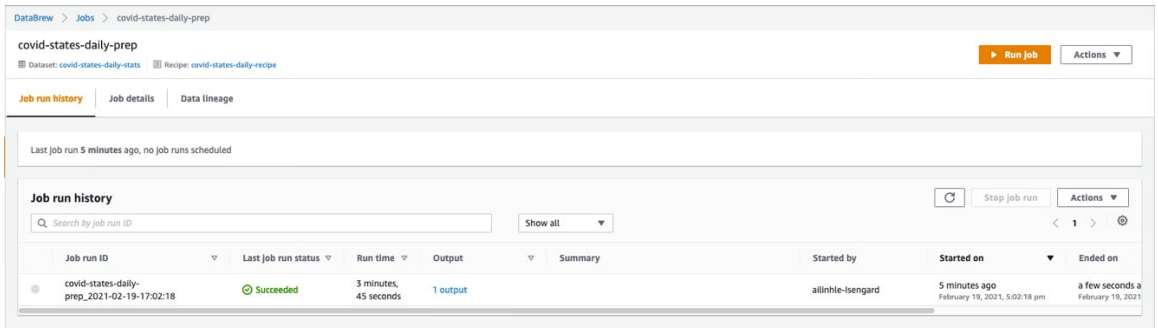
5. In the **Permissions** section, select the role **databrew-lab-DataBrewLabRole-xxxxx.** Click **Create and run job.**



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



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



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

Viewing data lineage

1. 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.