

Amazon Web Services Data Engineering Immersion Day

Lab 6. Bonus Lab: Glue DataBrew July 2021

Table of Contents

Introduction	
Get Started Using the Lab Environment	
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	
Viewing data lineage	

Introduction

In <u>DataBrew</u> 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 -<u>https://aws-dataengineering-day.workshop.aws/1300.html</u>

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 - <u>https://github.com/aws-samples/data-engineering-for-aws-immersion-day.</u>

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.			
statutut			
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:

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:

Environment Setup	i Readme
Outputs:	
S3 Bucket name mod-3fccddd609114925-dmslabs3bucket-1ngcgzzcnd15u 順	
BusinessAnalystUser mod-3fccddd609114925-BusinessAnalystUser-MB0XFZLQLOXX 旧	
DMSLabRoleS3 ARN arn:aws:iam::377243295828:role/mod-3fccddd609114925-DMSLabRoleS3-O2V	T1RSN43SG 🌓
Glue Lab Role mod-3fccddd609114925-GlueLabRole-YLTJA13WW6WT 🌓	
S3BucketWorkgroupA mod-3fccddd609114925-s3bucketworkgroupa-tbon3m1mkunh 🏢	
S3BucketWorkgroupB mod-3fccddd609114925-s3bucketworkgroupb-18ygl8nfp8ead 🏢	
WorkgroupManagerUser mod-3fccddd609114925-WorkgroupManagerUser-5IVF0U0NIRG4 🎼	

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

Team Dashboard



4. 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 " "as your region, unless otherwise directed by the event operator. Login Link	
Credentials / CLI Shin ets Mac/Linux Wind op Mac or Linux (#)	
euport MS_DEFAULT_RECOM- euport MS_SECRET_ACCESS_KEY- euport MS_SECRET_ACCESS_KEY- euport MS_SESSION_TOKEN-	
How do I use the AWS CLI? Checkout the AWS CLI documentation here: https://docs.aws.amazon.com/cli/intest/userguide/cli-chap-welcome.html	
OK	

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.

Quick create stack			
Template			
Template URL https://s3.amazonaws.com/aws-dataengineering-day.workshop.aws/DataBrew_PreLab_CFN.y Stack description -	aml		
Stack name			
Stack name			
databrew-lab			
Stack name can include letters (A-Z and a-z), numbers (0-9), and dashes (-).			
SourceBucket S3 bucket which contains the source object			
aws-dataengineering-day.workshop.aws			
SourceKey S3 Key which contains the source object			
states_daily.csv.gz			
pabilities			
The following resource(s) require capabilities: [AWS::IAM::ManagedPolicy] This template contains identity and Access Management (IAM) resources that might provide the contains identity and Access Management (IAM) resources that might provide the contains in the create each of these resources and that they have the minimum received that you want to create each of these resources and that they have the minimum received that AWS CloudFormation might create IAM resources.	de entities acc quired permis	tess to make changes to yessions. Learn more 💈	our AWS account.

In case you aren't able to launch the quick create stack, you can download the <u>template file</u> and then follow the steps to <u>create stack</u> manually.

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

Stack info Events	Resources Outputs	Parameters Template	Change sets	
Outputs (3)				C
Q Search outputs				۲
Key	Value	\bigtriangledown Description \bigtriangledown	Export name	$\overline{\nabla}$
DataBrewLabRole	mod-f90408bc75a34bd DataBrewLabRole-QH55	5- IAM role for 9NK7F5SL DataBrew lab	-	
DataBrewOutputS3Bucke	mod-f90408bc75a34bd databrewoutputs3bucke h9qvkgpobuwz	5- S3 bucket for t- DataBrew output	-	
DatasetS3Path	s3://mod-f90408bc75a3 databrewoutputs3bucke h9gykgpobuwz/states d	4bd5- S3 Path to the t- dataset		

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

≡	DataBrew > Projects
DATASETS	Projects (0) Info
PROJECTS	
RECIPES	
JOBS	Start data preparation You can visualize a sample dataset in a project in which you can
WHAT'S NEW	perform a variety of transforms. Have your dataset ready, and you are ready to go.
	Create project Create sample project

- 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	
covid-states-daily	
The project name must contain	n 1-255 characters. Valid characters are alphanumeric (A-Z, a-z, 0-9), hyphen (-), period (.), and space.
project could also be applied t Attached recipe	o a project. Recipe name
Create new recipe 🔻	covid-states-daily-recipe
	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 recipient of the steps from a steps	pe 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

elect a dataset elect the dataset that you want to work on	
O My datasets Your imported datasets	Sample files Explore example files for your dataset
New dataset details	
Dataset name	
covid-states-daily-stats	
The dataset name must contain 1-255 chara	ters. 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 **DatasetS3Path** that is available in Event Engine Team Dashboard or outputs section of your CloudFormation stack

Connect to new data	set Info		
🕂 File upload	S3 path	Parameterized S3 path	
Data lake/data store			
Amazon S3	Enter your source from S3 In For you to select a folder, all files	nfo in the folder need to share the same file t	ype. If there are different
AWS Glue Data Catalog	Schemas, they will be merged.		
Amazon S3 tables	Format is: s3://bucket/prefix		
🖻 Amazon Redshift tables	S3 Buckets		C
🗟 Amazon RDS tables			
🕅 All AWS Glue tables	Q Search S3 objects by no	nme	
Others			< 1 2 > @
P AWS Data Exchange	Name		Size
	_		

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

Select the type and size of your	sample		
Туре			
First n rows		•	
	a annula?		
How many rows do you want t	to sample?		
How many rows do you want t • 500	to sample?		
How many rows do you want t 500 1,000	to sample?		
How many rows do you want t 500 1,000 2,500	to sample?		

8. In the **Permissions** section, select the role xxxx-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 requir	ed 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	C

9. Click Create project

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

covid-states-daily The data set: covid-states-daily & Sample: First n s	ample (500 rows)			Create job	NEAGE ACTIONS
UNDO REDO FILTER COLUMN FORMAT CLEAN EXTR	ACT MISSING INVALID DUPLICATES SPLIT MERGE CREATE	D FUNCTIONS	UNNEST PIVOT GROUP JOIN UNION	TEXT SCALE MAPPING ENCODE	
	X SAMPLE	🖽 GRID	SCHEMA	E Recipe (0)	×
	Vor session will be ready soon! Initiating session Initiating session <tr< th=""><th></th><th></th><th>covid-states-daily-recipe Working version</th><th>eve Mare</th></tr<>			covid-states-daily-recipe Working version	eve Mare

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.

TASETS		Aa A* .	TRACT MISSING INVALID DUPLICATES SP	LIT MERGE CRE	P Σ B: D Σ H Ξ FUNCTIONS UNNEST PIVOT GROUP JOIN UNION	TEXT SCALE MAPPING ENCODE	
Ⅱ)лестѕ		500 rows		×	SAMPLE	GRID III SCHEMA	🛍 PROFILE
_	ABC state	ን …	# positive	ን …	# probableCases ∇ •••	# negative	ዎ … #
	Unique 56	Total 500	Unique 464	Total 500	Unique 296 Total 500	Unique 329	Total 393 Unic
CIPES	AK	9 1.8%	Indiana a a a		http://www.control.com	littlittlittere atte	ահե
Þ	AL	9 1.8%	H I		1		(CE
OBS	AR	9 1.8%	Min Median Mean Mode 0 309.17 K 471.89 K 0	Max 3.34 M	Min Median Mean Mode Max 0 7.67 K 39.49 K 0 323.64 K	Min Median Mean Mode 2.14 K 1.78 M 2.11 M	Max Min 15.49 M 11
_	All other values	473 94.6%					
LE NEW	AK		53279		0	null	^
	AL		472423		101367	1816273	
	AR		306736		62862	2285451	
			•		•	24.40	

The Grid view shows

- Columns in the dataset
- o Data type of each column
- o Summary of the range of values that have been found
- o 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
- o 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.

DataBrew > Jobs > Create job
Create job Info
Job details
Job name Identifier for the jobs
covid-states-daily profile job
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 O Custom sample 20000 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-labdatabrewoutputs3bucket-xxxx and a folder name (eg. data-profile)

Job output se Running a job gener	ates output files at specified file destinations.
File type Output format	S3 location Format is: s3://bucket/folder/
JSON	s3://databrew-lab-databrewoutputs3bucket-p5jcyb8htw76/data-profile/ Browse
Encryption Enable encrypt Encrypt the job	otion for job output file output file using SSE-S3 or AWS KMS
Advanced Settings that core	job settings - optional Info ntrol the processing and compute used for the jobs run on your project
 Associated You can associat 	d schedules - <i>optional</i> Info e up to 2 schedules to automate your job.
▶ Tags - option Metadata that yu an optional valuu other criteria.	Dnal Du can define and assign to AWS resources. Each tag is a simple label consisting of a customer-defined key (name) and e. Using tags can make it easier for you to manage, search for, and filter resources by purpose, owner, environment, or
Permissions DataBrew needs per	Info mission 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-D	ataBrewLabRole-314O3K1MT6ZM
	Cancel Create job Create and run job

- 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.

Ξ	DataBrew > Jobs					
DATASETS	Recipe Jobs Profile Jobs Schedules					
PROJECTS	Profile jobs (1) Info			View details 🕨 Run Job	Actions 🔻	Create job
RECIPES	Q Find jobs	Show all 🛛 🔻			<	(1) ◎
Þ	□ Job name	Data profile 🛛 🗢	Last run 🗢	Created on v	Created by	' Tags ⊽
1085	Covid-states-daily profile job	View data profile	6 minutes ago February 16, 2021, 12:05:16 am	10 minutes ago February 16, 2021, 12:01:17 am	Admin	
WHAT'S NEW						

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

DataBrew > Datasets > covid-states-daily		
covid-states-daily [©] S3		► Rerun profile Create project with this dataset Actions ▼ P
Dataset preview Data profile overview Column statistics Data lineage		
Last Job run Succeeded 12 minutes ago ago, no Job runs scheduled Data profile was run on custom sample of fini 20,000 rows of your dataset.		Select profile to view Job run 1 February 16, 2021, 12:01:18 am 🔻
Summary TOTAL ROWS TOTAL COLUMNS 19,541 55		Correlations Correlation coefficient (r) defines how dosely two variables are related. It ranges from -1.0 to +1.0, where 0 means there is no relationship between the variables.
DATA TYPES # INTEGER ARC STEING © TIMESTAMP 46 columns 7 columns 2 columns		data and a second secon
VALID CELLS	MISSING CELLS	regarise and a second s
675963 63% DUPLICATE ROWS	398792 37%	hoptalizeGumetry Image: Comparison of the second
VALD ROWS 19541 100%	DUPLICATE ROWS	

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.

covid-states-daily Dataset: covid-states-daily Sample: First n sample (500 rows)				Create Job UNRAGE ACTIONS
covid-states-daily (Input) S3 states_daily.csv S.3 MB View dataset]			
				GRID III SCHEMA
Data profile overview Column statistics				
Rerun profile Last Job run Succeeded 27 min Data profile was run on custom sample	hutes ago ago, no job runs scheduled of fini 20,000 rows of your dataset		Select profile to view Job	run 1 February 16, 2021, 12:01:18 am 🔻
Columns (55)	# Integer date			
Q Find				
ALL (55) # INTEGER (46) MC STRING (7) ▼	Data quality		Data insights	
# date	VALID VALUES	MISSING VALUES	Cardinality Low	2% of the rows are unique 398
ABC state	19541 100%	O 0%	Missing	No missing values 0
# positive				
# probableCases				
# negative	Value distribution		Correlations	
# pending			Correlation coefficient (r) defines how closely two varia	ables are related. It ranges from -1.0 to
MC totalTestResultsSource	Unique 398	Total 19,541	+1.0, where 0 means there is no relationship between	the variables.
# totalTestResults			TOP CORRELATIONS	
# hospitalizedCurrently			Columns	-1 -0.5 0 0.5 1
# hospitalizedCumulative			date	1
# InkuCurrently			and the second se	0.41

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.

covid-states-daily 😹 Sample: First n sample (500 rows) I Dataset: covid-states-daily Aa A* ____ ++ Solumns ▼ 500 rows SOURCE # date ል ... ABC state V 500 Change type Un AK 56 🖝 boolean AL ABC string .21 M All of # byte AK # short AL AR # integer AS # long AZ # float CA # double co # decimal ст O timestamp DC DE 20210213 20210213

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.

×
Publish More
G 🗗

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**

	00 rows						
ABC date	V	•••	ABC year			V	•••
♦ Viewing 59 columns 5 SOURCE ABC date 0 Distinct 9 Unique 0 20210213 20210213 0 0 20210211 All other values 0 0 0 20210213 20210213 0 0 0 0 20210213 20210213 0 <t< td=""><td>56 1 56 1 56 1 332 6</td><td></td><td>ABC year Clean Extract Remove or fill missing values Remove or replace invalid values Remove duplicate values Split column Create a flag column Word tokenization Categorical mapping One-hot encode column Hide column</td><td>> > > ></td><td>500</td><td>V</td><td>500 100%</td></t<>	56 1 56 1 56 1 332 6		ABC year Clean Extract Remove or fill missing values Remove or replace invalid values Remove duplicate values Split column Create a flag column Word tokenization Categorical mapping One-hot encode column Hide column	> > > >	500	V	500 100%
20210213 20210213 20210213 20210213 20210213 20210213 20210213			Duplicate Delete 2021 2021 2021 2021 2021 2021				

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

Duplicate column		×
Create duplicate column Ir Duplicate a column from ar	1 fo 1 existing col	umn
Source column Select a source column to create a	a duplicate	
date	•	
Duplicate column name Name of the newly created duplic date_copy	ate column]
Valid characters are alphanumeric	, underscore, a	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**

			ぶ SAMPLE			
SOURCE						
date_copy	ア	•••	ABC state		ም	# positive
Unique 9	Total		Rename		Total 500	Unique 463
20210213	56 1		Sort	>	9 1.8%	
20210212	56 1	-		<u> </u>	9 1.8%	
20210211	56 1	Aa	Format	>	9 1.8%	Min Median Mean Mo 0 315.97 K 483.38 K 0
All other values	332 6	<u>A</u> *	Clean	>	473 94.6%	
20210213		<u>_</u> †_	Extract	>		54282
20210213		_	Remove or fill missing			479856
20210213			values	>		312562
20210213			Remove or replace invalid			0
20210213			values	1		795323
20210213		ð	Remove duplicate values			3391036
20210213			Split column	>	Delimiters	
20210213				-	0	
20210213		Ша	Create a flag column		On a si	ngle delimiter
20210213		es.	Word tokenization		On mu	ltiple delimiters
20210213		45	Categorical mapping		Betwee	en delimiters
20210213					Positions	
20210213		101	One-hot encode column		At posi	tions from beginning
20210213			Move column	>	At posi	cions nom beginning
20210213			IA	_	At posi	tions from end
20210213			ID		Betwee	n positions
20210213			IL		At regu	lar intervals
20210213			IN			647657

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 X
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
O 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 r	n sample (500 rows)				Create job	ACTIONS
	Aa A*	TRACT MISSING INVALID DUPLICATES SPUT MERGE CRE		TEXT SCALE MAPPING ENCODE			RECIP
> Viewing 57 columns ▼	500 rows 🗌	View highlighted only	a sample	GRID GRID	SCHEMA	Split column	×
		SOURCE will be deleted	PREVIEW	PREVIEW		spin country	~
date		date_copy	ABC date_copy_1	ARC date_copy_2	ARC state	date conv	
ique 9 10213 10212	Total 500 56 11.2% 56 11.2%	Unique 9 Total 500 20210213 56 11.2% 20210212 56 11.2% 20210211 56	Unique 1 Total 500 2021 500 100%	Unique 9 Total 500 0215 56 11.2% 0212 56 11.2% 0211	Unique 56 AK AL AR	Split column options Using positions	
other values	56 11.2% 332 66.4%	All other values 332 66.4%	-	56 11.2% All other values 332 66.4%	All other values	Split column	
210213		20210213	2021	0213	AK	 At positions from beginning 	
210213		20210215	2021	0215	AL AD	 At positions from end 	
210213		20210213	2021	0213	20	 Between positions 	
210213		30310313	2021	0213 47		Position from the beginning	
210213		20210213	2021	0213 CA			
210213		20210213	2021	0213	co	Enter a numeric position. Positions count from 1.	
210213		20210213	2021	0213	cτ	4 0	
210213		20210213	2021	0213	DC		
210213		20210213	2021	0213	DE	Apply transform to	
210213		20210213	2021	0213	FL.	 All rows (500 rows) 	
210213		20210213	2021	0213	GA	iransformation will be applied to all rows in the dataset	
210213		20210213	2021	0213	GU	 Filtered rows - 0 filters applied (500/500 rows) 	rs)
210213		20210213	2021	0213	HE	Transformation will be applied to filtered rows in the	
210213		20210213	2021	0213	IA	2 million	
210213		20210213	2021	0213	ID		
210213		20210213	2021	0213	IL.	Preview snown	
210213		20210213	2021	0213	IN		

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

covid-states-daily Dataset: covid-states-daily	ζ Sample: First	n sample (500 rows)								Create Job	eage actions
	Aa A*	TRACT MISSING INVALID DUPLICATES	SPLIT MERGE CREA	FUNCTIONS	UNNEST PIVOT GROUP JOIN UNION	TEXT SCALE MAPPING ENCODE					
⊗ Viewing 58 columns ▼ 5	500 rows			Xs	IMPLE		GRID	III SCHEMA	ÎÎÎ PROFILE	E Recipe (4)	×
date	¥	ABC date_copy_1	γ	ABC date_copy_	2_1 \\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\	ABC date_copy_2_2	y	ABC state			
Unique 9	Total 500	Unique 1	Total 500	Unique 1	Total 50	Unique 9	Total 500	Unique 56		covid-states-daily-recipe	E ***
20210213	56 11.2%	2021	500 100%	02	500 100	6 13	56 11.29	6 *		Total greater	auton Prore
20210212	56 11.2%					12	56 11.29	AL AR		Applied steps (4) Clear all	17 B
All other volues	332 66.4%					All other volues	332 66.49	All other volves		1. Change type of date to String	
20210213		2021		02		13		AK		2. Create duplicate column date_copy from	date
20210213		2021		02		13		AL		3 Split column at position 4 from beginnin	n in date c
20210213		2021		02		13		AR		ору	g in dute_t
20210213		2021		02		13		AS		4. Split column at position 2 from beginnin	g in date c
20210213		2021		02		13		AZ		opy_2	
20210213		2021		02		13		CA			
20210213		2021		02		13		со			
20210213		2021		02		13		ст			
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		н			
20210213		2021		02		13		IA			
20210213		2021		02		13		ID			
20210213		2021		02		13		IL.			
20210213		2021		02		13		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 Info		
Rename a column		
Source column		
Select column to rename		
date copy 1	•	
New column name		
New column name New name for the column year		
New column name New name for the column year ralid characters are alphanu	meric, underscore, and s	pace

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.

covid-states-daily III Dataset: covid-states-daily							
UNDO REDO FILTER COLUMN AA A*		FUNCTIONS UNNEST PIVOT GROUP JOIN UNION	TEXT SCALE MAPPING ENCODE				
Solumns ▼ 500 rows		₩ SAMPLE	GRID	🛄 SCHEMA 🛛 🕅 PROFILE			
ABC date 🛛 🏹 👓	ABC year 🛛 🏹 …	ABC month 🛛 🖓 …	ABC day ♀ …	ABC state			
Unique 9 Total 500 2020213 56 11.2% 2020213 56 11.2% 2020213 56 11.2% Al other values 332 66.4% 20210213 20210213 20210213 20210213 20210213 20210213 20210213 20210213 20210213 20210213	Unique 1 Teal 500 2021 500 100% 2021 2021 2021 2021 2021 2021 2021 2021 2021 2021 2021 2021	Unique 1 Total 500 22 500 100% 02 0	Unique 9 Total 500 13 56 11.2% 14 56 11.2% 15 61 11.2% 13 56 11.2% 14 other valuer 332 66.4% 13 13 13 13 13 13 13 13 13 13 13 13 13 1	AK AK AL AL AL AL AL AL AL AL AL AL AL AL AL			
20210213	2021	02	13	FL			
20210213 20210213 20210213	2021 2021 2021	02 02 02	13	GU HI			
20210213 20210213 20210213	2021 2021	02	13 13	IA ID			
20210213 20210213	2021	02	13	IL 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		ぷ SAMPLE	
Columns (58/5 Ctrl+click, Commar	8 shown) nd+click, or Shift+click to select m	ultiple columns.		
Q probableCa	ases			×
Select all Select f	irst 40 Select last 40 Unselect	all		
ALL (58) ABC ST	RING (11) # INTEGER (45)	② DATE (0)	OTHER TYPES ▼	
🗹 # pro	bableCases		v	-
Cancel	Delete unselected colur	nns Sh	ow selected colun	nns

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

CO\ ⊞ □	covid-states-daily Dataset: covid-states-daily 🐇 Sample: First n sample (500 rows)								
UNI	Image: Decision The addition Addit Addition Addition <								
⊙ V	iewing 58 columns 🔻 500 rows		X SAMPLE		I	GRID	III SCH	iema 🕅 profile	
7	# positive ♀ …	SOURCE # probableCases 7	••• # negative	γ	# pending	٦	7	ABC totalTestResultsSour	
tal 500	Unique 463 Total 500	Unique 295 Total	Rename	Total 393	Unique 35	т	otal 44	Unique 4	
1.8%	linii	ha an	Sort >		lt i rinnt			posNeg	
1.8%	Min Median Mean Mode Max	Min Median Mean Mode Max	Remove or fill missing values >	Remov	e missing rows	Mode Ma	×	totalTestEncountersViral	
94.6%	0 312/3/K 482/30K 0 2/33 M	20/ 30.22 K 41.2/ K 337.24	Remove or replace invalid >	Fill wit	h custom value	7.8	10 K	All other values	
	54282	null	values	Fill wit	h null			totalTestsViral	
	479856	103398	Remove duplicate values	Fill wit	h most frequent value			totalTestsPeopleViral	
	312562	64915	Min-max normalization		hand and the second second			totalTestsViral	
	0	null	Scale between specified values	Fill Wit	n last valid value			totalTestsViral	
	795323	53357	Mean normalization	Fill wit	h numeric aggregate			totalTestsViral	
	3391036	null	7 come normalization or		null			totalTestsViral	
	412983	21726	standardization		null			totalTestEncountersVira	
	267337	16422			null			posNeg	
	38796	null	Move column >		null			totalTestEncountersVira	
	82672	4213	Hide column		null			totalTestEncountersVira	
	1788827	325309	Duplicate		7856			totalTestEncountersVira	
	962808	173738	Delete		null			totalTestsViral	
	7692	208	Delete		null			posNeg	
	27527	717	null		null			totalTestEncountersVira	
	274386	null	1004947		null			posNeg	
	167225	31276	485478		null			totalTestsPeopleViral	
	1160523	null	15946386		null			posNeg	
	647657	null	2396197		null			totalTestsViral	

Lab 5. Bonus Lab: Glue DataBrew

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**.

UNDO REDO	FILTER COLUMN FORMAT CLEAN EXTRACT MISSING INVALID DUPLICATES SPLIT MER	GE CREATE FUNCTION	UNNEST PIVO				
	58 columns 🔺 500 rows	LE					
rrently	Columns (58/58 shown) Ctrl+click, Command+click, or Shift+click to select multiple columns.		ሻ ••				
Unique 94	94 Q dataQualityGrade X						
Mean Mor	Select all Select first 40 Select last 40 Unselect all						
92.68	ALL (58) ABC STRING (11) # INTEGER (45) ④ DATE (0) OTHER TYP	ES 🔻	2.22 M				
	ABC dataQualityGrade <u>View</u>						
	Cancel Delete unselected columns Show selected	columns					
,	null null						
	null 111173						
	null null						

rows SOURCE √ ••• ABC lastUpdateEt 7 ••• ABC dataQualityGrade ን … Rename Unique 6 Total 491 al 393 Total A+ 259 16 3.2% Sort > A 126 15 3% -11 в Aa Format > 2 M 14 88 2.8% At Clean > All other values 27 455 91% ★ Extract > А А Remove or fill missing > values A+ Remove or replace invalid values N/A > A+ в Remove duplicate values А 🕀 Split column > в 🖫 Create a flag column A+ Word tokenization А А 🔊 Categorical mapping A+ 📟 One-hot encode column в в Move column > 2/15/2021 00:00 A+ А 2/12/2021 19:00 2/13/2021 00:00 А A+ 2/12/2021 23:59

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
А	2
В	3
С	4
D	5

Categorically map column						
Source Select	e column a column to per	rform catego	orical mapp	ping		
data	QualityGrade	9	•			
Mapping options						
 Map all values (6 values) Custom map values 						
Мар \	alues	🗹 Map v	alues to n	umeric values		
All va	lues			New value		
	Q Find	1				
	A+	259	51%	1 3		
	A	126	25%	2 🕄		
	В	88	17%	3 🕄		
	N/A I	9	1%	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.

Publish recipe	×
Publishing a recipe will create a new version of your recipe. Published versions of the recipes can be selected as options for a recipe job.	
Recipe name	
covid-states-daily-recipe	
Version description	
	11.
Recipe steps (10)	
1. Change type of date to String	
2. Create duplicate column date_copy from date	
3. Split column at position 4 from beginning in date_copy	
4. Split column at position 2 from beginning in date_copy_2	
6 Rename date_copy_1 to year	
7 Rename date copy 2 2 to day	
8. Fill missing values with 0 in probableCases	
9. Categorical map dataQualityGrade	
10. Change type of dataQualityGrade_mapped to Integer	
Cancel Publish	

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
covid-states-daily-prep
The job name must contain 1-240 characters. Valid characters are alphanumeric (A-Z, a-z, 0-9), hyphen (-), period (.), and space.
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.
Job input Info The input dataset for the job and the recipe to be applied to it.
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
Q covid-states-daily X Browse datasets Connect new dataset
Select a recipe Recipe version
Q covid-states-daily-recipe X Version 1.0 ▼ Browse recipes

4. In the Job output settings section, enter the S3 location s3://databrew-labdatabrewoutputs3bucket-xxxxx/job-outputs/.

Click Settings under Job Output Settings

Job output settings Info Running a job generates output files at sp	pecified file destinations.			
Output 1				Settings
Output to Output location	File type Output format	Delimiter CSV separator	Compression Available types	Setting summary File output storage
S3 location Format is: s3://bucket/folder/				Create a new rolder for each job run Custom partition by column values None
s3://mod-f90408bc75a34bd5-da	tabrewoutputs3bucket-h9qvkgpol	ouwz/job-out Browse		Output path preview s3://mod-f90408bc75a34bd5-databrewoutputs3bu cket-h9qvkgpobuwz/job-outputs/covid-states-daily-prep_17Jul2021_timestamp_part00000.csv ③ Output files are partitioned if they're too large.

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**.

e outr	ut storage	
Crea	te a new folder for each j	b run
Unde conta /jobn	r specified S3 path, a new fol- ins job name and job run tim- ame_10may2020_timestamp	er will be created for each job run and for each output file type. The output folder and file name Example: s3://bucket/myfolder/jobname_10may2020_timestamp/filetype_compression part1.csv
Repl	ace output files for each j	b run er the specified S3 path. For every job run, the previous output files will be replaced with files fro
latest /jobn	job run. You can enable buck ame_part1.csv	t versioning to be able to restore previous file versions. Example: s3://bucket/myfolder
stom rtition l partiti	job run. You can enable buck ame_part1.csv partition by column value by unique values of columns. ioned by Column A and Colun year	t versioning to be able to restore previous file versions. Example: s3://bucket/myfolder
riat o latest /jobn rition l partiti ::	job run. You can enable buck ame_part1.csv partition by column value oy unique values of columns. oned by Column A and Colun year month	t versioning to be able to restore previous file versions. Example: s3://bucket/myfolder

Cancel

Save

5. In the **Permissions** section, select the role **databrew-lab-DataBrewLabRole-xxxxx**. 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	e the latest updates.
databrew-lab-DataBrewLabRole-314O3K1MT6ZM	• C

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

⊘ Crea	ed recipe job *covid-states-daily-p	rep".						
DataBrev	V > Jobs > covid-states-daily-	-prep						
covid-	states-daily-prep et: covid-states-daily 🔲 Recipe: covid	-states-dai	ly-recipe					▶ Run job
Job run	history Job details Da	ata lineag	e					
Last	job run a few seconds ago, no job r	runs scher	duled					
lob	run history							C. Stop job run Actions 🔻
Q	Search by job run ID					Show all 🔻		< 1 > @
	Job run ID		Last job run status 👳	Run time 🗢	Output	▽ Summary	Started by	Started on Ended on
	covid-states-daily- prep_2021-02-16-08:56:05) Running	Not available	1 output		ailinhle	a few seconds ago - February 16, 2021, 8:56:05 am

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

ovid-states-daily-prep	► Run job Actions ▼
Dataset: covid-states-daily-stats 🔠 Recipe: covid-states-daily-recipe	
ob run history Job details Data lineage	
Last job run 5 minutes ago, no job runs scheduled	
Last Job run § minutes ago, no Job runs scheduled	
Last job run § minutes ago, no job runs scheduled Job run history	C Stop job run Actions V
Last job run § minutes ago, no job runs scheduled Job run history Q. Stearch by jeb run ID Stow all	C Step jeb run Actions ▼ ▼ < 1 > ●
Last job run 5 minutes age, no job runs scheduled Job run history Q. Search by job run ID Show all	C Step jeb run Actions ▼ ▼ < 1 > ●
Last job run 5 minutes ago, no job runs scheduled Job run history Show all Q. Search by job run ID The status The st	C Step job run Actions ▼ ▼ < 1 > ● ● y Started by Started on ▼ Ended on
Last job run 5 minutes ago, no job runs scheduled Job run history C. Search by jeb run 1D Job run 1D V Last job run status V Run time V Output V Summar covid-states-daily- O Surrended 3 minutes, 1 putrunt	Y Started by Started on V Ended on allipsig-tensor 5 minutes ago a few second

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.

=	DataBrew $>$ Projects $>$ covid-states-daily $>$ Project data lineage	
DATASETS	Data lineage	OPEN PROJECT
PROJECTS	2° Lineage ○ CloudTrall logs	Zoom 100% ¥ @ @ 53
RECIPES		
JOBS WHAT'S NEW	33 State and COV COV S1 Balance S1 Bal	

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.