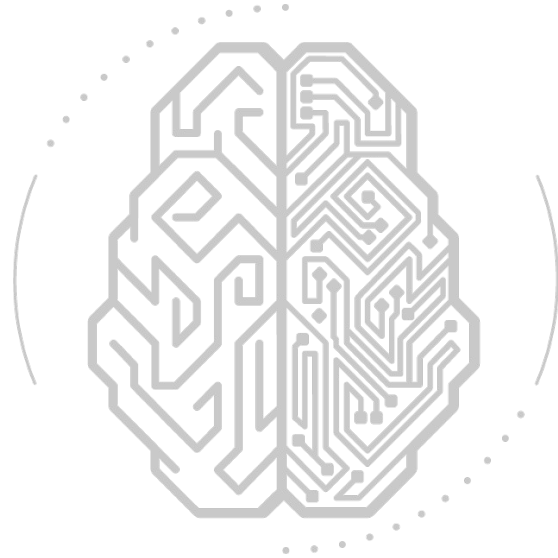




Transform Data With AWS Glue DataBrew



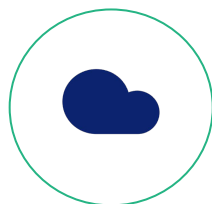
Agenda



Introductions



Overview on DataBrew



Demo



Q&A



Sanchit Jain

Lead Architect - AWS at Quantiphi
AWS APN Ambassador

Why we need DataBrew?

- We all are fascinated with various analytical stuff like fancy visualization, BI report, Machine learning output
- But do we know, 50% of the time is just spend in cleansing, understanding & exploring your data
- We all like the elegant outcome but really no one want really to invest the time required to clean the data due to combustion process, multiple rounds of back and forth, and time consuming, etc.
- With this growing ask from the industry, AWS launched Glue DataBrew service in Nov 2020 with an idea to purely focus on business use-case rather than preparation work required for the same.



What is AWS Glue DataBrew ?

AWS Glue DataBrew is a visual data preparation tool that makes it easy for data analysts and data scientists to prepare data with an interactive, point-and-click visual interface without writing code.



CAPABILITIES OF GLUE DATABREW

Profile

Clean and Normalize

Map Data Lineage

Automate



NEED FOR DATABREW

"Upto 80% of data analysis time is spent on preparing data"

Time Consuming

- Multi-step process to extract, clean, normalize & load data at scale
- The right tools for the right persona must be integrated

Expensive

- Costly user licenses & siloed tools that cause rework
- Often requires moving large amount of data into silos

Manual

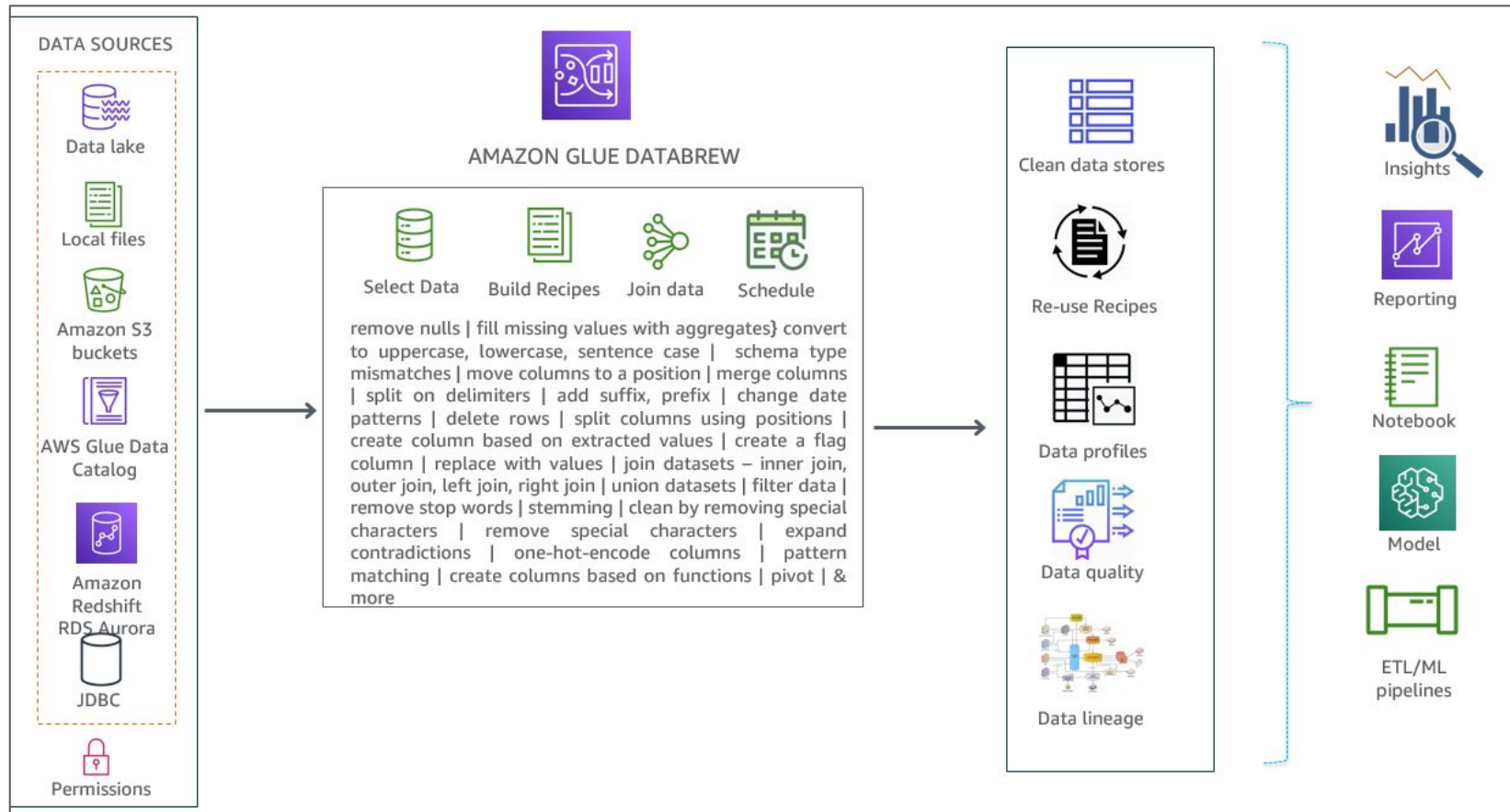
- Needs a lot of code-based heavy lifting to work at scale
- Hard to operationalize & build repeatable workflows

How does DataBrew Work ?

- You can choose from over 250 built-in transformations to combine, pivot, and transpose the data without writing code.
- AWS Glue DataBrew also automatically recommends transformations such as filtering anomalies, correcting invalid, incorrectly classified, or duplicate data, normalizing data to standard date and time values, or generating aggregates for analyses.
- For complex transformations, such as converting words to a common base or root word, DataBrew provides transformations that use advanced machine learning techniques such as Natural Language Processing (NLP).
- You can group multiple transformations together, save them as recipes, and apply the recipes directly to newly incoming data.

For input data, AWS Glue DataBrew supports commonly used file formats, such as comma-separated values (.csv), JSON and nested JSON, Apache Parquet and nested Apache Parquet, and Excel sheets. For output data, AWS Glue DataBrew supports comma-separated values (.csv), JSON, Apache Parquet, Apache Avro, Apache ORC and XML.

Overview of DataBrew



Advantages of DataBrew

ADVANTAGES



Visual Data Lineage

View the various stages of data transformation from start to end



Serverless data preparation at scale

Operate at massive scale in a serverless capacity. Pay only for what you use



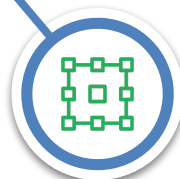
Integrates with data pipeline

SDK/API access to integrate in existing data pipelines



Connect to data sources

S3, Redshift, RDS Aurora or Glue Data Catalog



250+ built in transformations

Join, tokenize, split, merge, extract, remove, Group, pivot, normalize, label encode or more

Core Concepts

Dataset - a set of data—rows or records that are divided into columns or fields



Project - The interactive data preparation workspace in DataBrew



Recipe - is a set of instructions or steps for data that you want DataBrew to act on



Job - The process of running these instructions when you make the recipe is called a job.



Data Lineage - DataBrew tracks your data in a visual interface to determine its origin, called a data lineage.



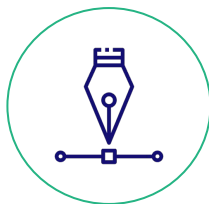
Data Profile - A report which summarizes your existing shape of your data



Demo



Prerequisites



Create Project
& Dataset



Exploring & Preparing
Dataset



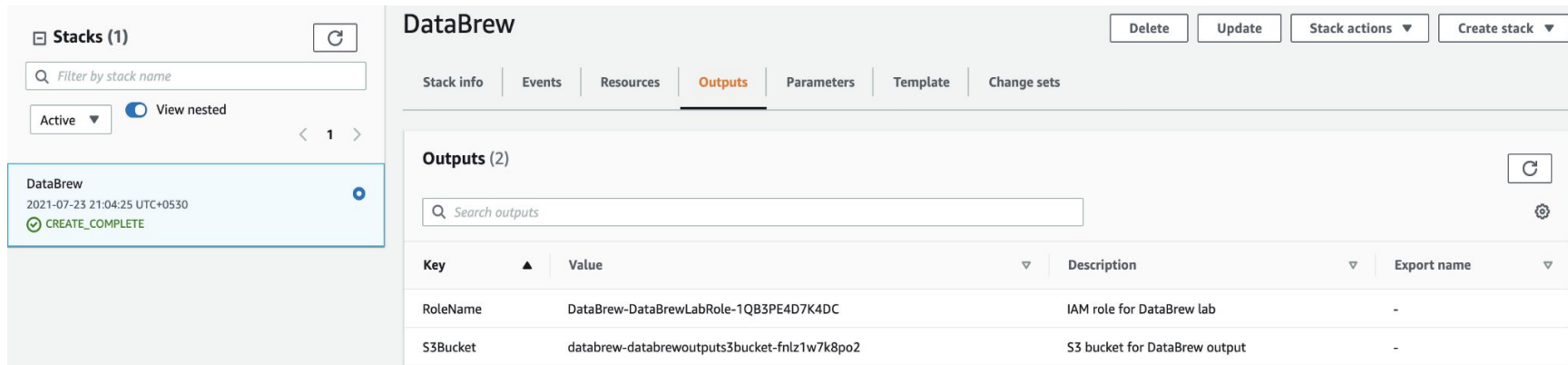
Creating a
DataBrew job



Viewing data
lineage

Prerequisites

- Download the [dataset](#) from this link and upload it to the S3 bucket
- Download the [cloudformation template](#) from this link and Deploy it
- Once the Cloudformation stack is deployed successfully please capture the values for RoleName and S3Bucket details



The screenshot displays the AWS CloudFormation console interface. On the left, the 'Stacks (1)' sidebar shows a single stack named 'DataBrew' with a status of 'CREATE_COMPLETE'. The main panel is titled 'DataBrew' and features tabs for 'Stack info', 'Events', 'Resources', 'Outputs', 'Parameters', 'Template', and 'Change sets'. The 'Outputs' tab is selected, showing a table of two outputs: 'RoleName' and 'S3Bucket'. The 'RoleName' output has a value of 'DataBrew-DataBrewLabRole-1QB3PE4D7K4DC' and a description of 'IAM role for DataBrew lab'. The 'S3Bucket' output has a value of 'databrew-databrewoutputs3bucket-fnlz1w7k8po2' and a description of 'S3 bucket for DataBrew output'.

Stacks (1)

Filter by stack name

Active View nested

DataBrew
2021-07-23 21:04:25 UTC+0530
CREATE_COMPLETE

DataBrew

Delete Update Stack actions Create stack

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

Outputs (2)

Search outputs

Key	Value	Description	Export name
RoleName	DataBrew-DataBrewLabRole-1QB3PE4D7K4DC	IAM role for DataBrew lab	-
S3Bucket	databrew-databrewoutputs3bucket-fnlz1w7k8po2	S3 bucket for DataBrew output	-

Creating a project

- Navigate to the AWS Glue DataBrew service
- On the DataBrew console, select Projects
- Click Create project
- 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.

Creating a dataset

- In the Select a dataset section, select New dataset and enter covid-states-daily-stats
- In the Connect to a new dataset section, select Amazon S3 under “Data lake/data store” and Enter the S3 path. Leave the default configuration values
- In the Permissions section, select the role DataBrew-DataBrewLabRole--xxxxx from the drop-down list
- Click Create project

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

Search S3 objects by name

< 1 2 >

Name

Size

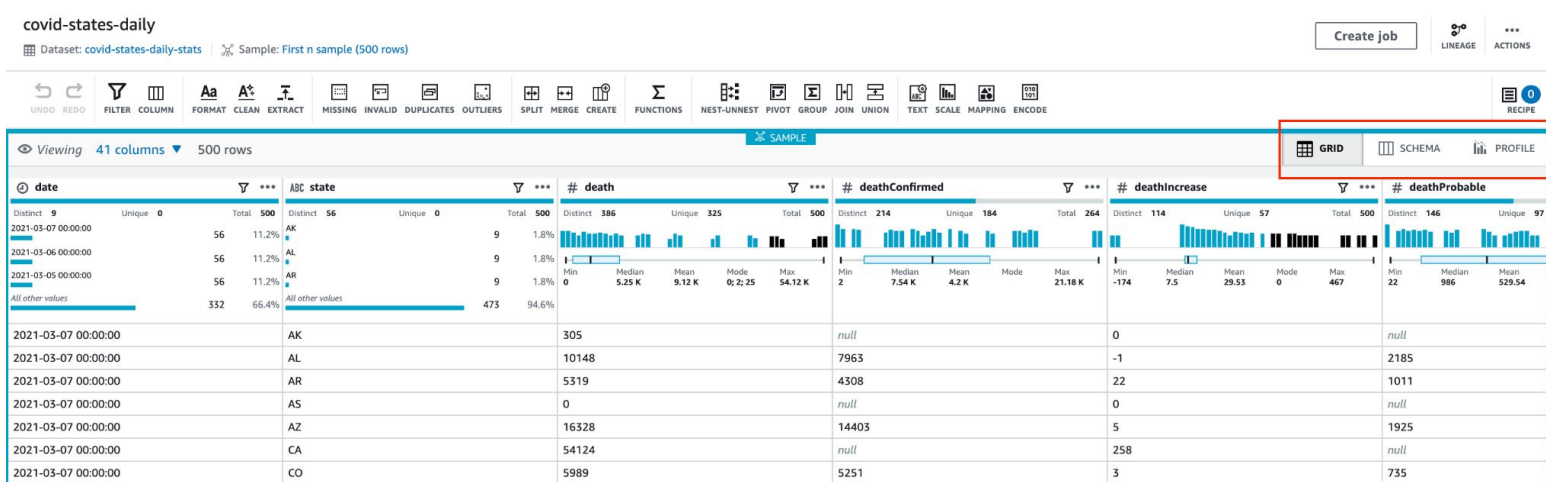
Creating a dataset

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

The screenshot displays the Glue DataBrew interface for a dataset named 'covid-states-daily'. The top navigation bar includes a 'Create job' button, 'LINEAGE', and 'ACTIONS' menu. Below this is a toolbar with various data transformation tools such as UNDO, REDO, FILTER, COLUMN, FORMAT, CLEAN, EXTRACT, MISSING, INVALID, DUPLICATES, SPLIT, MERGE, CREATE, FUNCTIONS, UNNEST, PIVOT, GROUP, JOIN, UNION, TEXT, SCALE, MAPPING, and ENCODE. The main workspace shows the dataset details: 'Dataset: covid-states-daily' and 'Sample: First n sample (500 rows)'. A modal window is open in the center, titled 'Initiating session', with a rocket icon and a progress bar at 17%. The modal text states: 'Your session will be ready soon! Initiating session' and 'Your session will take about a minute to be ready. Once ready there will be no additional load time.' On the right side, there is a 'Recipe (0)' panel showing a 'covid-states-daily-recipe' as the 'Working version'. Below this, there is a section titled 'Build your recipe' with the text: 'Start applying transformation steps to your data. All your data preparation steps will be tracked in the recipe.'

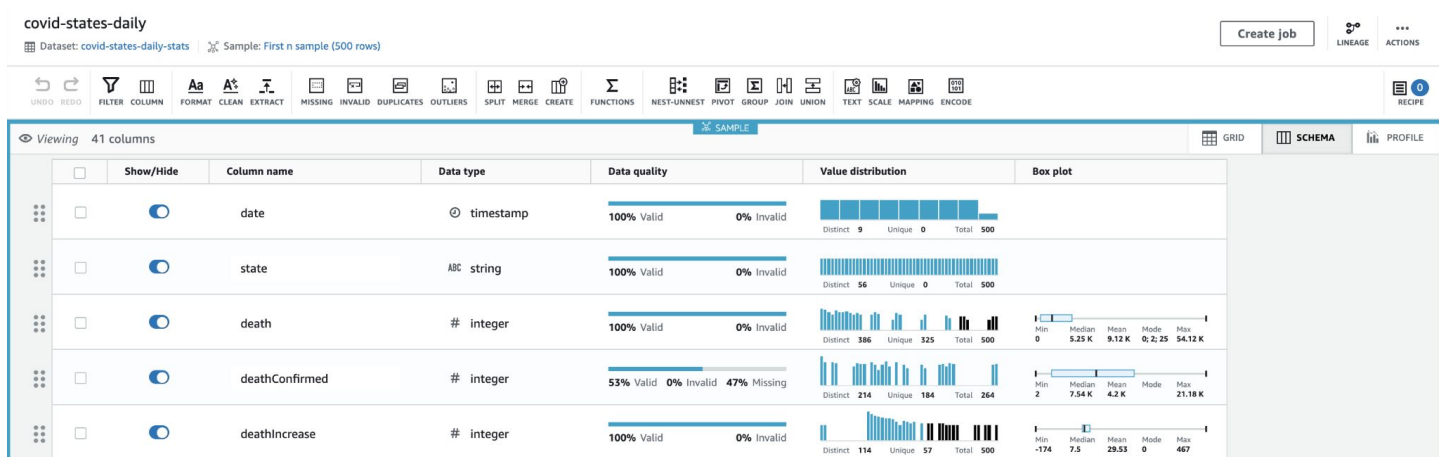
Exploring the dataset

- Grid view - 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



Exploring the dataset

- Schema view - 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

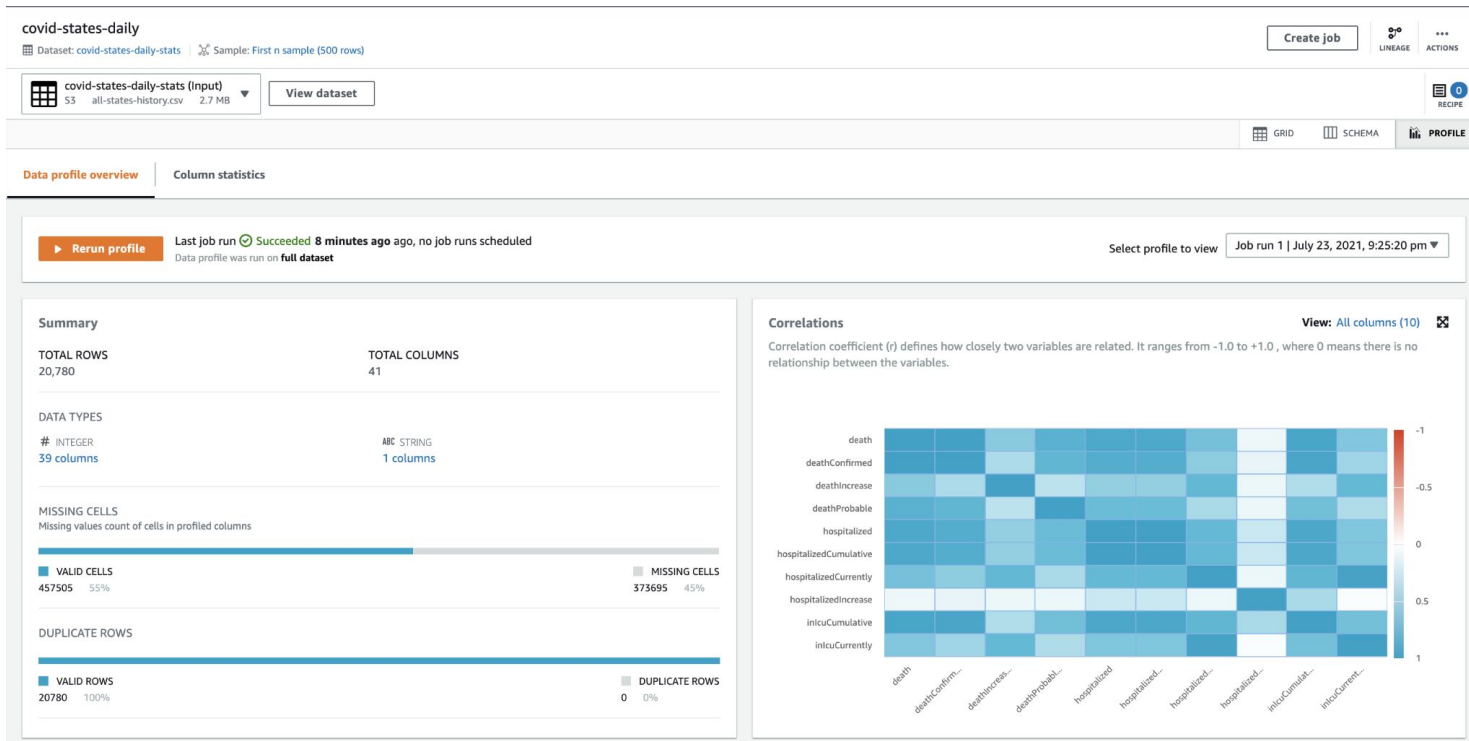


Exploring the dataset

- Profile view - 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.
 - Job Name
 - Click on Run data profile
 - In the job details and job run sample panels, leave the default values
 - Job Output Setting
 - In the Job output settings section, select the S3 bucket with the name DataBrew-DataBrewLabRole--xxxxx and a folder name (eg. data-profile)
 - In the Permissions section, select the IAM role with the name databrew-lab-DataBrewLabRole-xxxxx
 - Leave all other settings as the default values
 - Click Create and run job
 - When the profile job has successfully completed, click on View data profile under Jobs from the menu on the left hand side of the DataBrew console

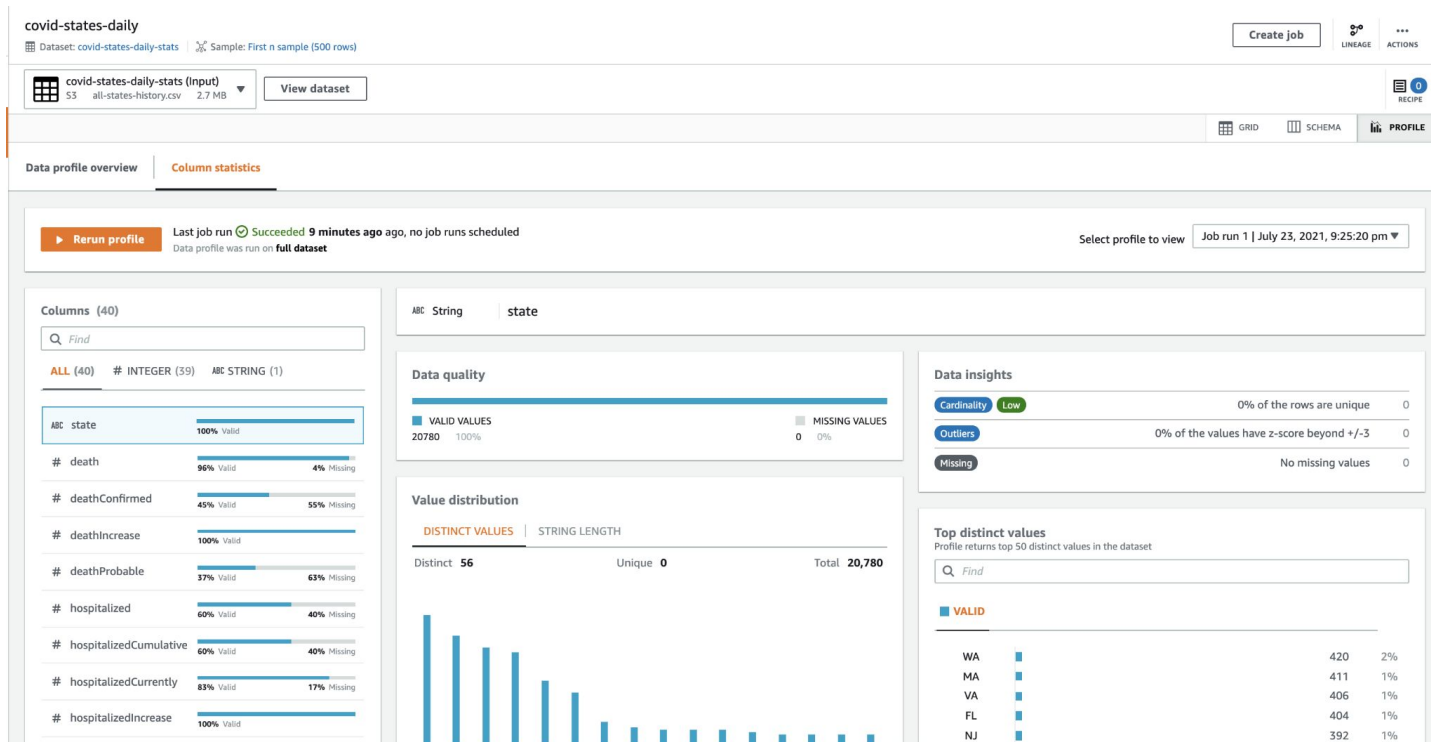
Exploring the dataset

- Profiling Output



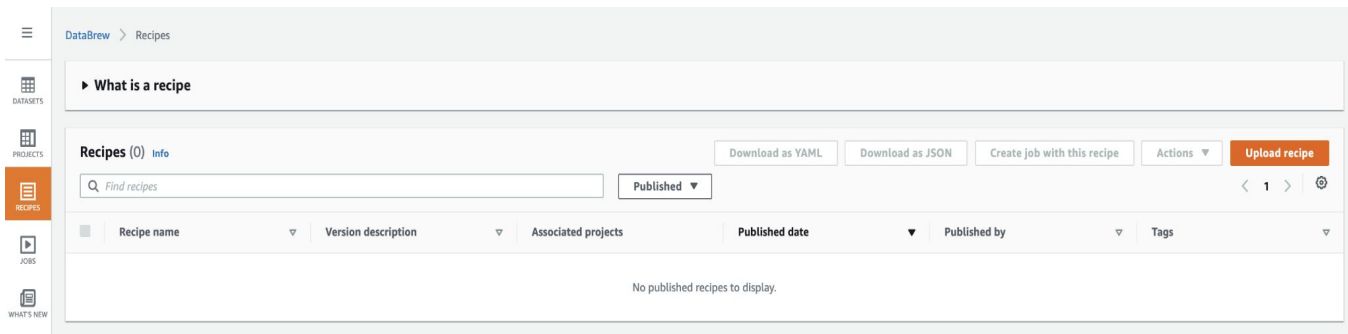
Exploring the dataset

- 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 different transformations to the dataset.
 - Rename columns
 - Change the data type of columns
 - Filled with the most frequent value
- Download the recipe from this [link](#)
- Select on Recipe from the menu on the left-hand side of the DataBrew console and click Upload Recipe



Preparing the dataset

- Provide below details
 - Recipe Name
 - Upload Recipe json script downloaded under Step 1
 - Select Create and publish recipe

Upload recipe

New recipe details

Recipe name

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

Description

Enter recipe description

Upload recipe

DataBrew recipe files can be downloaded as JSON from existing recipes and later uploaded, with or without changes

Select a file to upload

Upload a single file in json format

☒

 covid-states-daily-recipe.json
File size: 1.6 KB

► **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.

Cancel

Create recipe

Create and publish recipe

Preparing the dataset

- Now let's apply this recipe, click the project we configured now and the right side, click Import recipe
- Under Import recipe, select the recipe we configured and click Next
- Select the Append option from the right side and click Next
- Now let's validate the recipe and wait for all validation to be successful
- Once the validation is successful, click Import
- Now we will be back to the project screen and with the recipe implied on the dataset

The screenshot displays the Alteryx interface for a project named 'covid-states-daily'. The main workspace shows a data grid with columns: 'a/c year', 'a/c month', 'a/c date', 'a/c state', and '# death'. The data is filtered to show rows for the year 2021 and month of March. The right-hand pane shows the 'Recipe (9)' configuration, which includes a list of steps: 1. Change format of date to yyyy-mm-dd, 2. Change format of state to lowercase, 3. Fill missing values with 0 in deathProbable, 4. Fill missing values with 0 in hospitalized, 5. Split column on a single delimiter - in date, 6. Rename date_1 to year, 7. Rename date_2 to month, 8. Rename date_3 to date, and 9. Replace invalid values with March in month.

a/c year	a/c month	a/c date	a/c state	# death
2021	March	07	ak	305
2021	March	06	ak	305
2021	March	05	ak	305
2021	March	04	ak	305
2021	March	03	ak	305
2021	March	02	ak	302
2021	March	01	ak	300
2021	March	28	ak	290
2021	March	27	ak	290
2021	March	07	al	10148
2021	March	06	al	10149
2021	March	05	al	10122
2021	March	04	al	10094
2021	March	03	al	10029
2021	March	02	al	9991
2021	March	01	al	9931
2021	March	28	al	9929
2021	March	27	al	9930
2021	March	07	ar	5319
2021	March	06	ar	5297

Creating a DataBrew job

- Click on Jobs from the menu on the left hand side of the DataBrew console
 - On the Recipe jobs tab, click on Create job
 - Enter covid-states-daily-prep for the job name
 - Select Create a recipe job
 - Select the covid-states-daily dataset
 - Select the 'covid-states-daily-recipe'
 - In the Job output settings section, enter the S3 location s3://bucket-name/job-outputs/.
 - In the Permissions section, select the role DataBrew-DataBrewLabRole--xxxxx
 - Click Create and run job

Creating a DataBrew job

- DataBrew job is created and the job status is Running

Created recipe job "covid-states-daily-prep".

DataBrew > Jobs

Recipe jobs | Profile jobs | Schedules

Recipe jobs (1) Info

Find jobs Show all

View details Run job Actions Create job

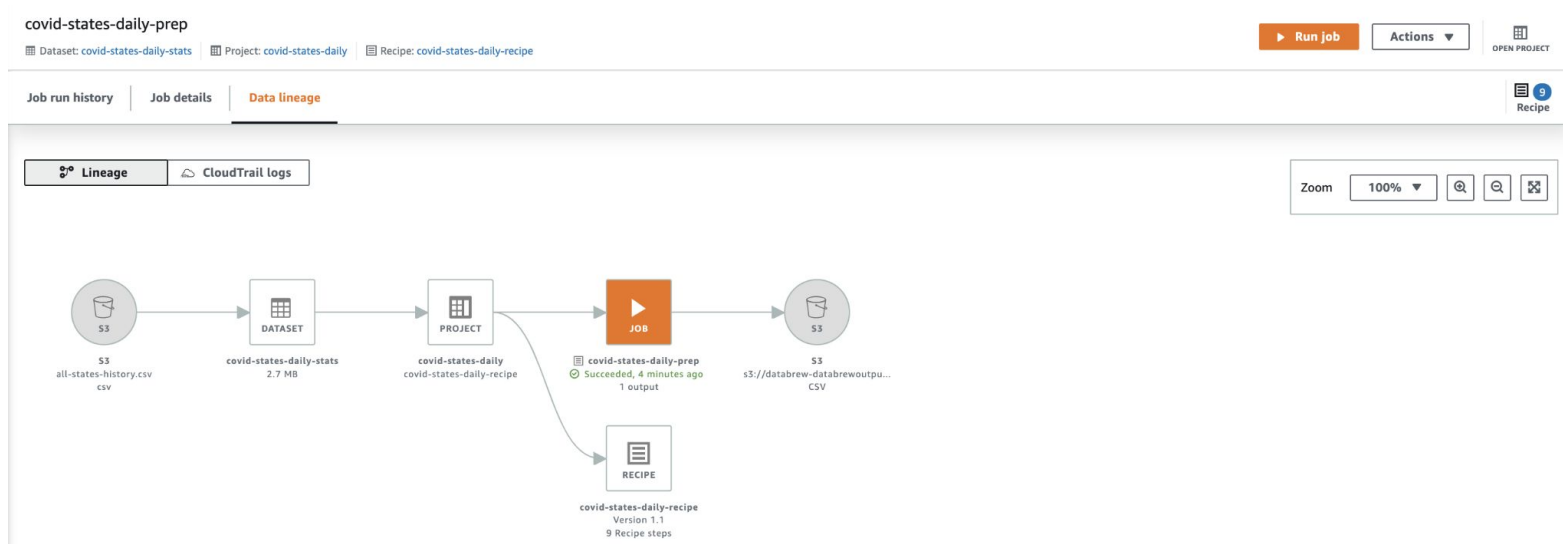
<input type="checkbox"/>	Job name	Status	Job input	Job output	Last run	Created on	Created by	Tags
<input type="checkbox"/>	covid-states-daily-prep	Running	covid-states-... (covid-states-... + covid-states-...) Project Dataset Recipe	1 output	-	a few seconds ago July 23, 2021, 10:20:19 pm	Aws-Admin-Role	-

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

<input type="checkbox"/>	Name	Type	Last modified	Size	Storage class
<input type="checkbox"/>	covid-states-daily-prep_23Jul2021_1627059065181_part00000.csv	csv	July 23, 2021, 22:21:19 (UTC+05:30)	2.6 MB	Standard

Viewing data lineage

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



Viewing data lineage

- Select CloudTrail logs to view all the action on this dataset.

DataBrew > Jobs > covid-states-daily-prep

covid-states-daily-prep

Dataset: covid-states-daily-stats | Project: covid-states-daily | Recipe: covid-states-daily-recipe

[Run job](#) [Actions](#) [OPEN PROJECT](#)

[Job run history](#) | [Job details](#) | [Data lineage](#) 9 Recipe

[Lineage](#) [CloudTrail logs](#)

Recent access activity (234/469)
Last 500 access activities for DataBrew in AWS CloudTrail. Events can take several minutes to appear in CloudTrail.

[View details](#) [Download](#) [View events for last 90 days](#)

Filter resources by property or value

Resource name: covid-states-daily [Clear filter](#)

<input type="checkbox"/>	Event name	Event time	User name	Resource type	Resource name
<input type="checkbox"/>	SendProjectSessionAction	July 23, 2021, 16:31:53 (UTC+00:00)	sanchit.jain	Project	covid-states-daily
<input type="checkbox"/>	SendProjectSessionAction	July 23, 2021, 16:30:33 (UTC+00:00)	sanchit.jain	Project	covid-states-daily
<input type="checkbox"/>	SendProjectSessionAction	July 23, 2021, 16:29:33 (UTC+00:00)	sanchit.jain	Project	covid-states-daily
<input type="checkbox"/>	SendProjectSessionAction	July 23, 2021, 16:28:33 (UTC+00:00)	sanchit.jain	Project	covid-states-daily
<input type="checkbox"/>	SendProjectSessionAction	July 23, 2021, 16:27:33 (UTC+00:00)	sanchit.jain	Project	covid-states-daily
<input type="checkbox"/>	SendProjectSessionAction	July 23, 2021, 16:26:32 (UTC+00:00)	sanchit.jain	Project	covid-states-daily

THANK YOU