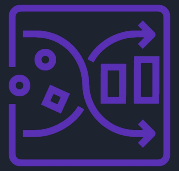
aws
COMMUNITY DAY
SOUTH ASIA
VIRTUAL EDITION 2021



AWS Glue DataBrew

aws
COMMUNITY DAY
SOUTH ASIA
VIRTUAL EDITION 2021

Agenda

Introductions

Overview on DataBrew

Demo

Q&A

aws
COMMUNITY DAY
SOUTH ASIA
VIRTUAL EDITION 2021

Speaker

Sanchit Jain

Lead Architect - AWS at Quantiphi
AWS APN Ambassador

FOLLOW ME

What is 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 50% 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

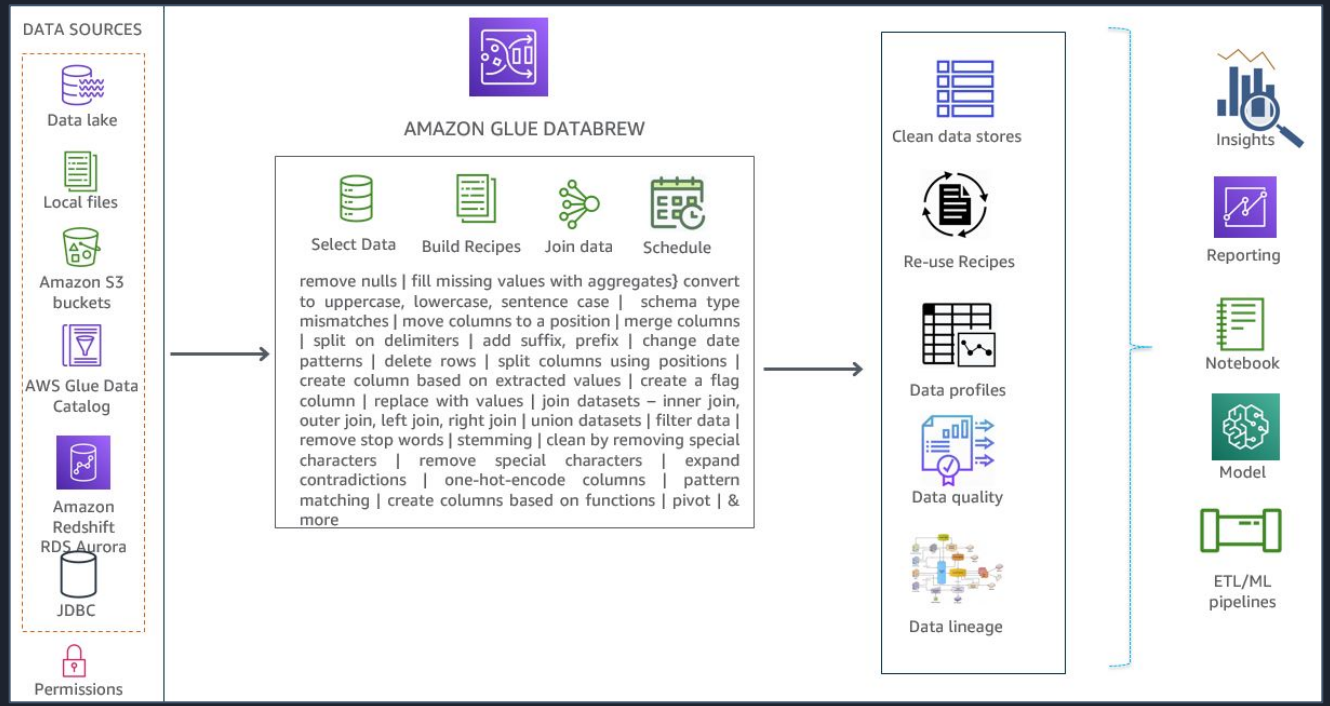
- Needs a lot of code-based heavy lifting to work at scale
- Hard to operationalize & build repeatable workflows

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.

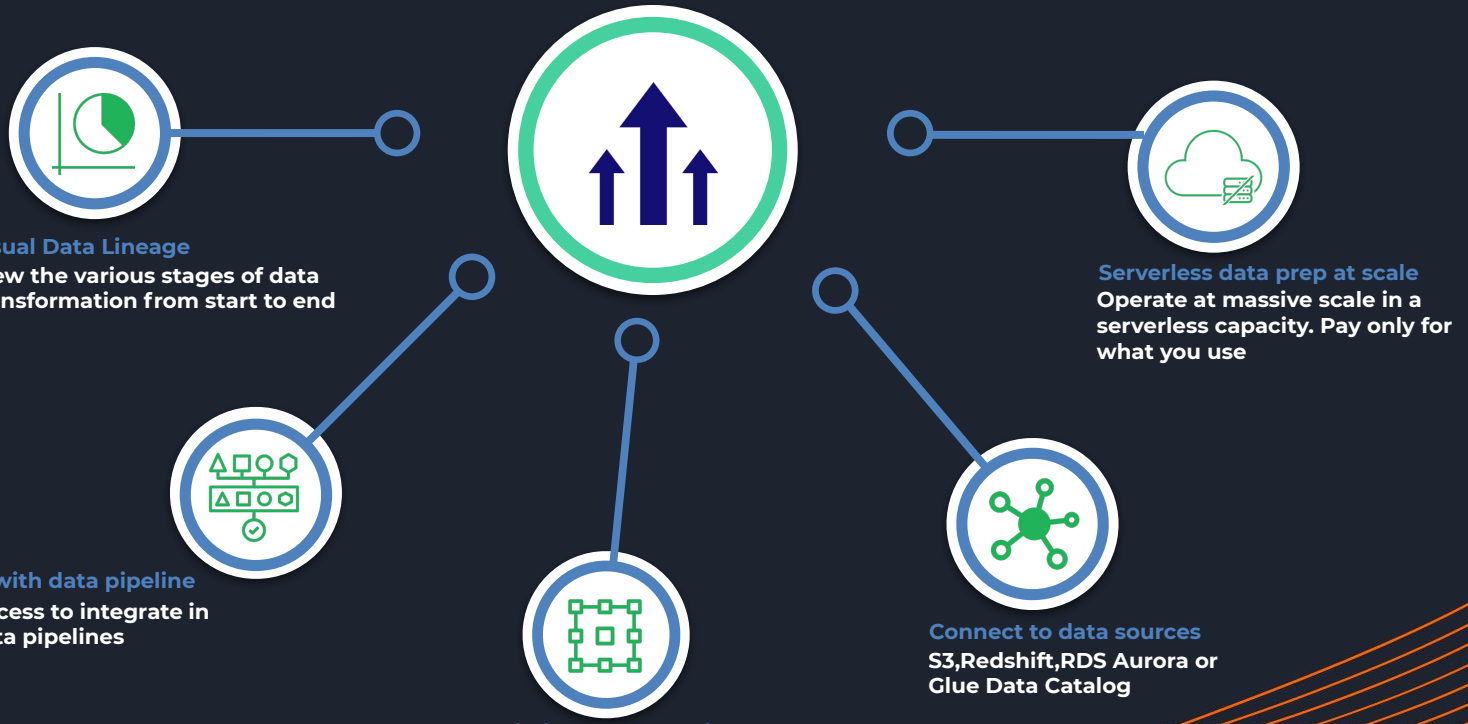
For input data, AWS Glue DataBrew supports commonly used file formats, such as comma-separated values (.csv), JSON and nested JSON, Apache Parquet and nested Apache Parquet, and Excel sheets. For output data, AWS Glue DataBrew supports comma-separated values (.csv), JSON, Apache Parquet, Apache Avro,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

Overview of DataBrew



aws COMMUNITY DAY SOUTH ASIA VIRTUAL EDITION 2021

Advantages of DataBrew



Visual Data Lineage
View the various stages of data transformation from start to end

Integrates with data pipeline
SDK/API access to integrate in existing data pipelines

250+ built in transformations
Join,tokenize,split,merge,extract, remove, Group,pivot, normalize, label encode or more

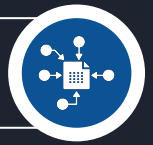
Connect to data sources
S3,Redshift,RDS Aurora or Glue Data Catalog

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

Core Concepts

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

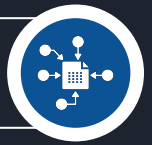
Project - The interactive data preparation workspace in DataBrew



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

aws COMMUNITY DAY SOUTH ASIA VIRTUAL EDITION 2021

Demo

Prerequisites

Create Project & Dataset

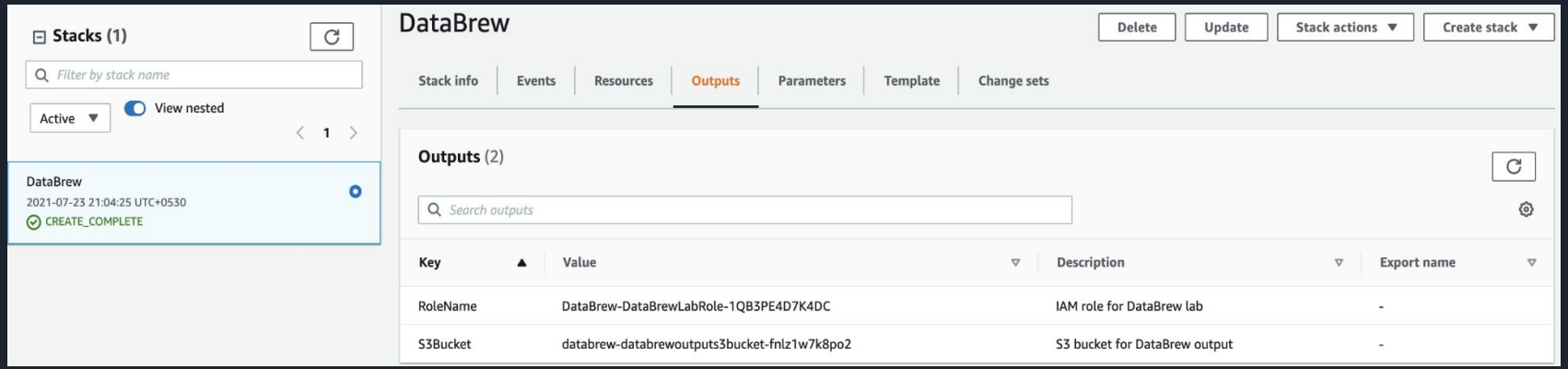
Exploring & Preparing Dataset

Creating a DataBrew job

Viewing data lineage

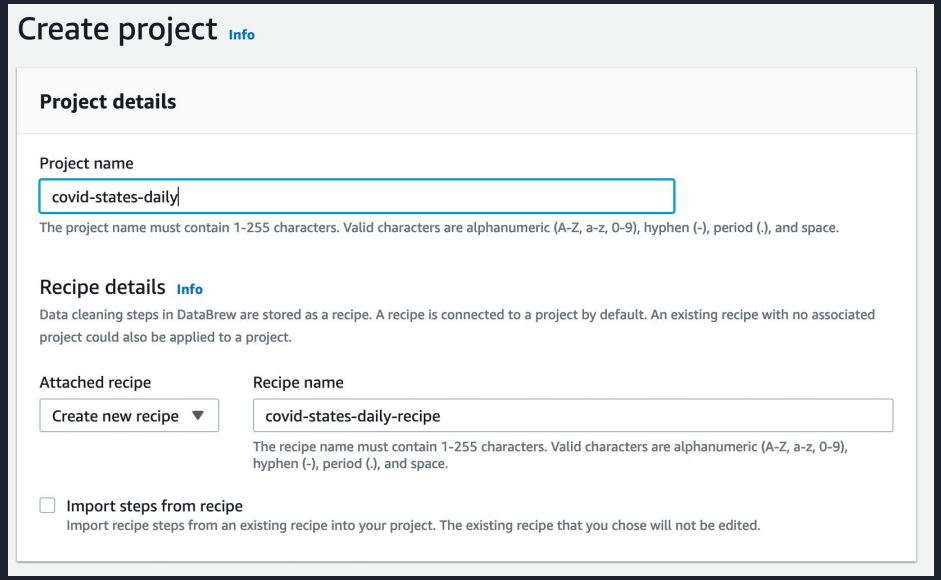
Prerequisites

- 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



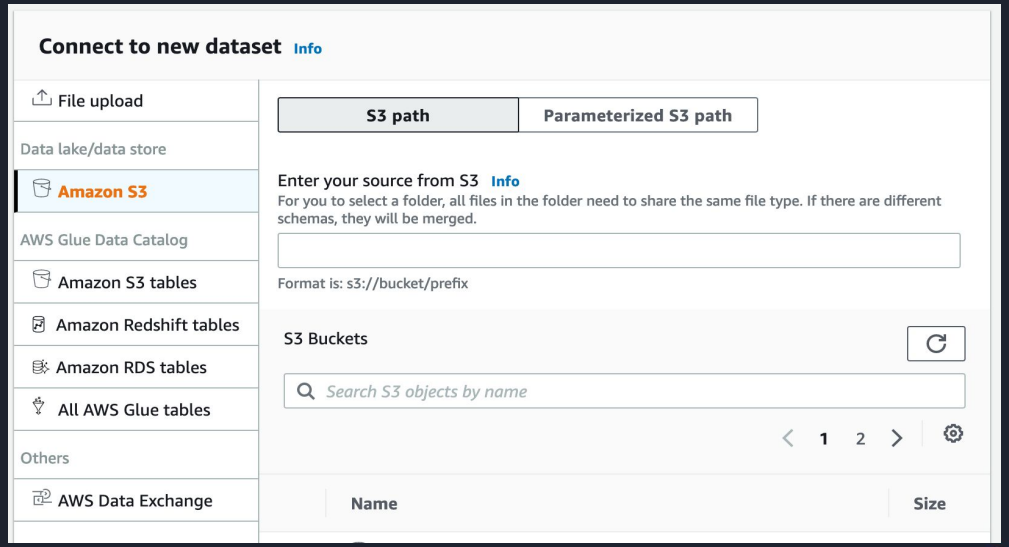
Creating a project

- 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



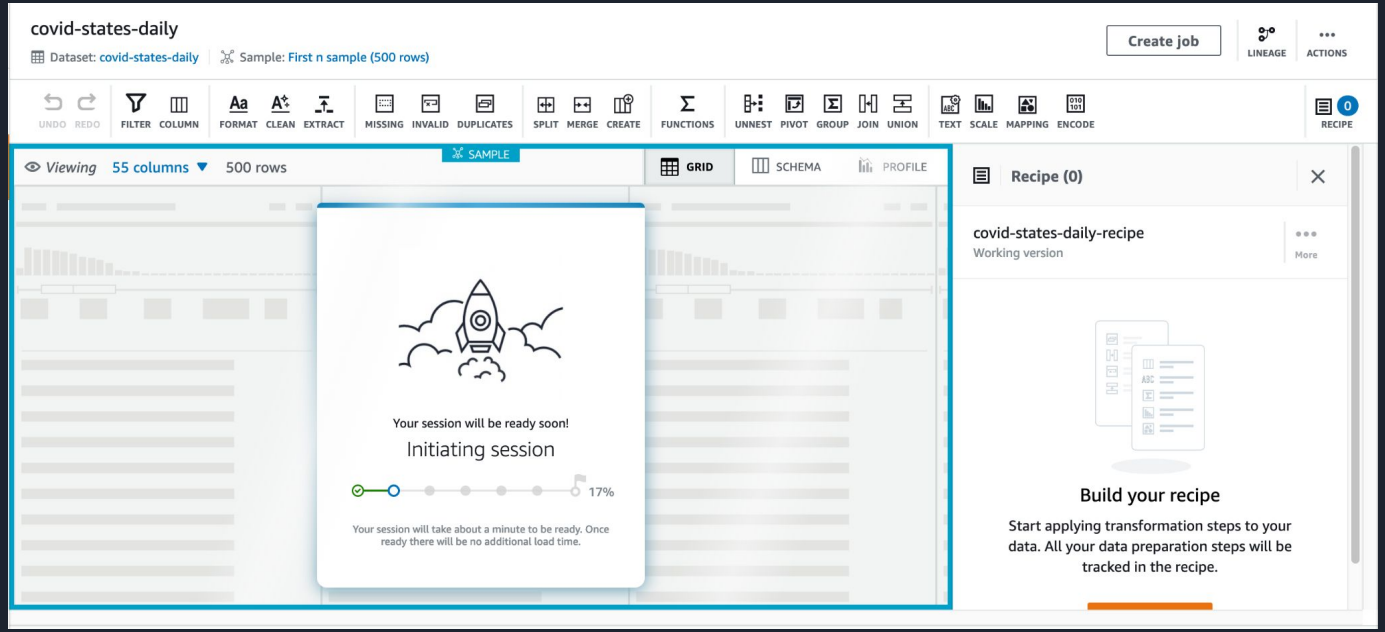
Creating a dataset

- 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



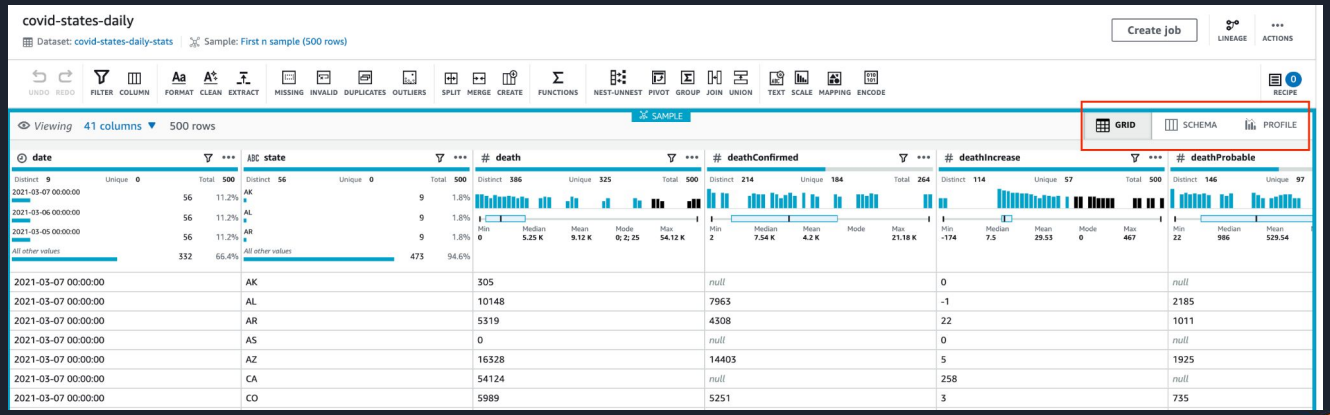
Creating a dataset

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



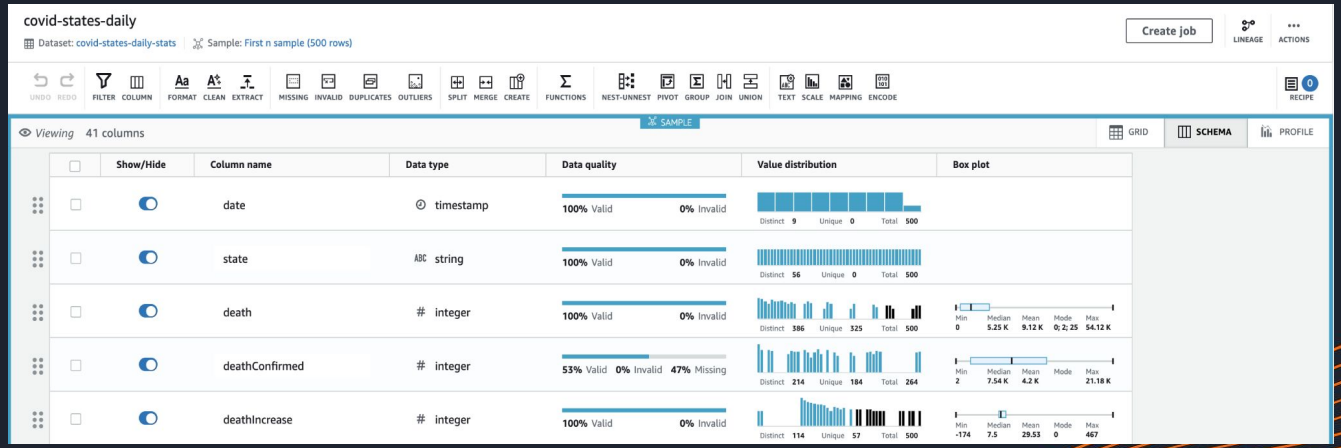
Exploring the dataset

- 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

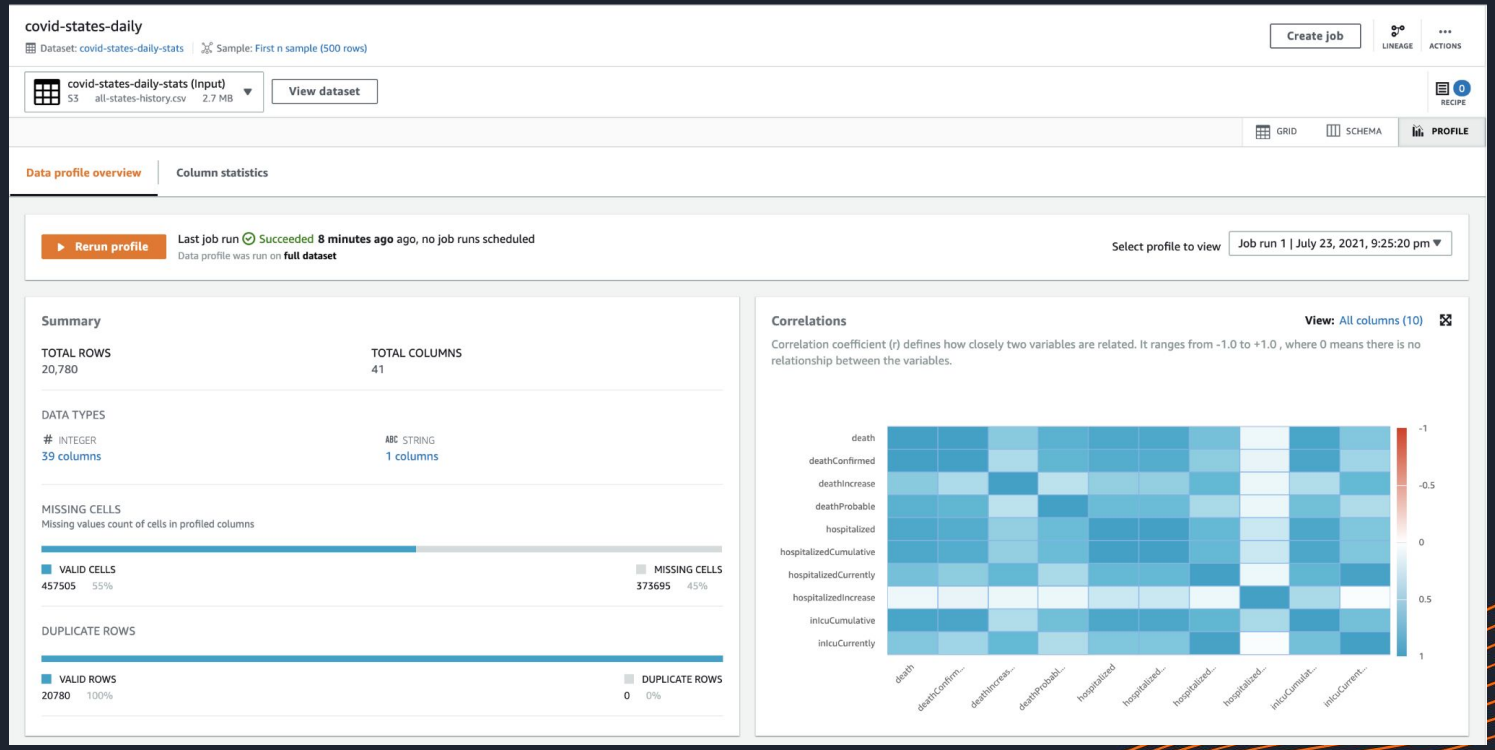


Exploring the dataset

- 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

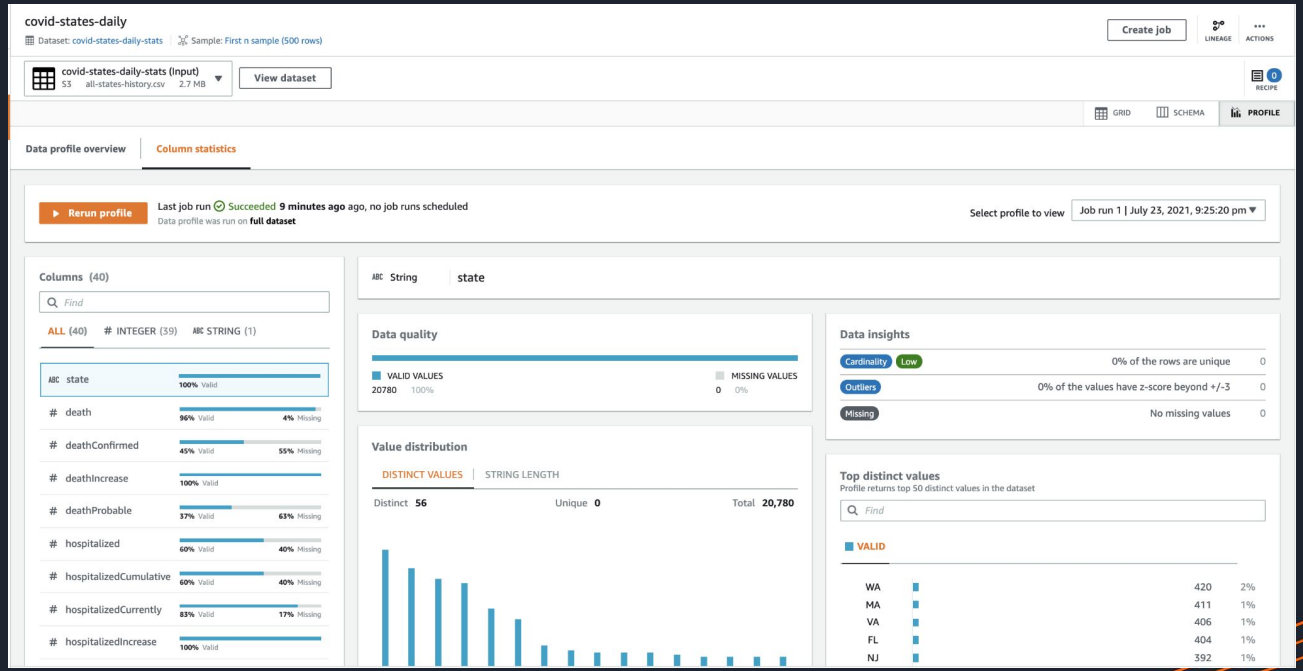
Exploring the dataset

- Profiling Output



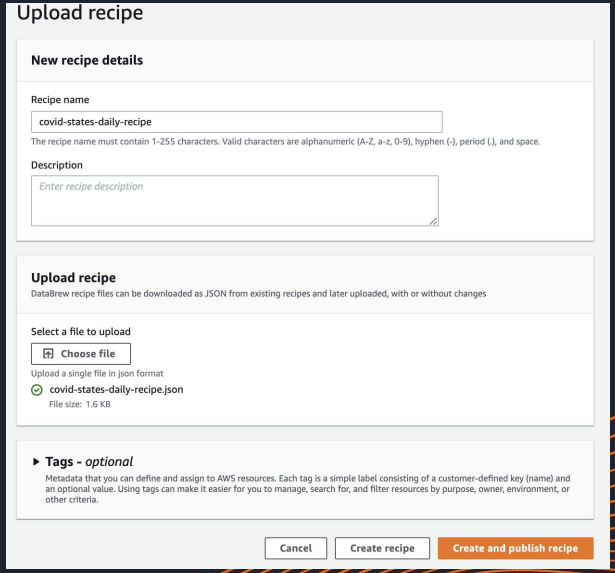
Exploring the dataset

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



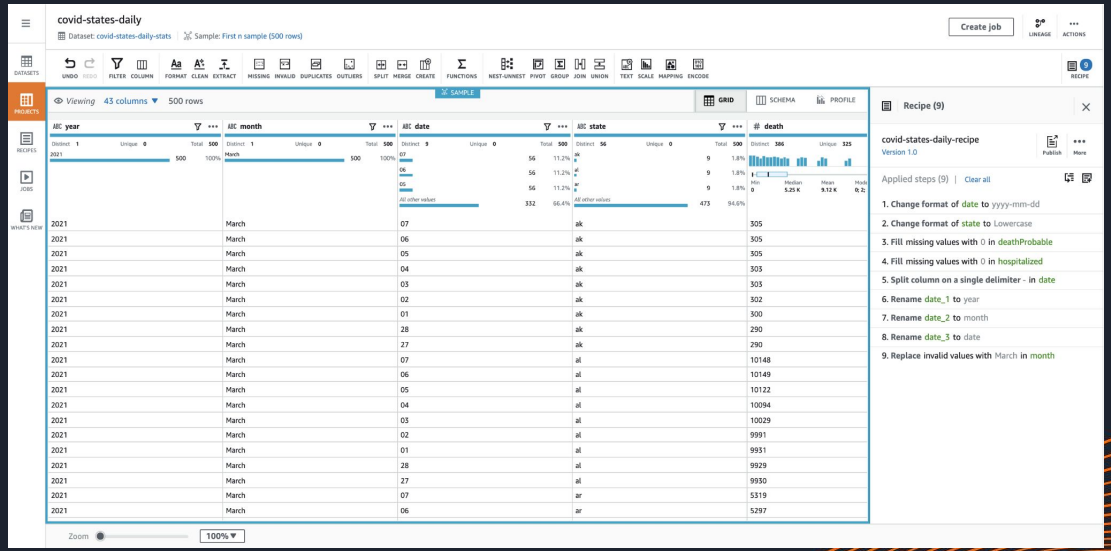
Preparing the dataset

- 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
- Provide the below details -
  - Recipe Name
  - Upload Recipe json script downloaded under Step 1
  - Select 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

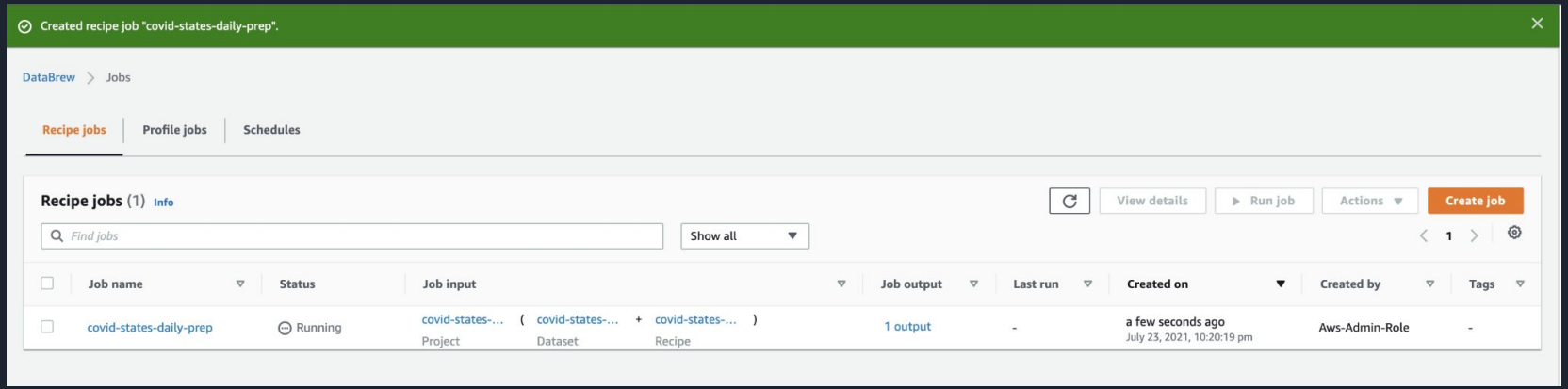


Creating a DataBrew job

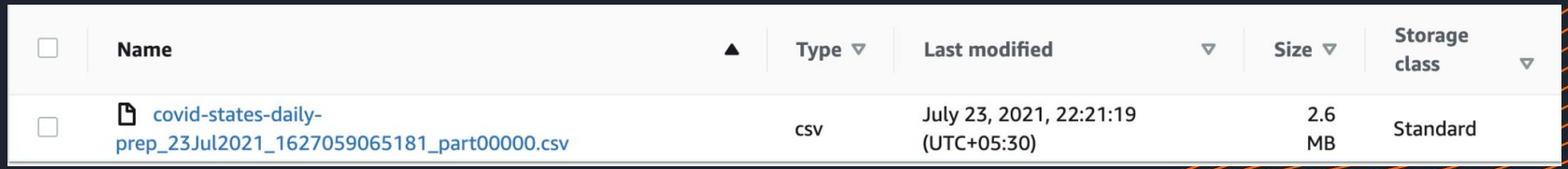
- Click on Jobs from the menu on the left hand side of the DataBrew console
  - On the Recipe jobs tab, click on Create job
  - Enter covid-states-daily-prep for the job name
  - Select Create a recipe job
  - Select the covid-states-daily dataset
  - Select the ‘covid-states-daily-recipe’
  - In the 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

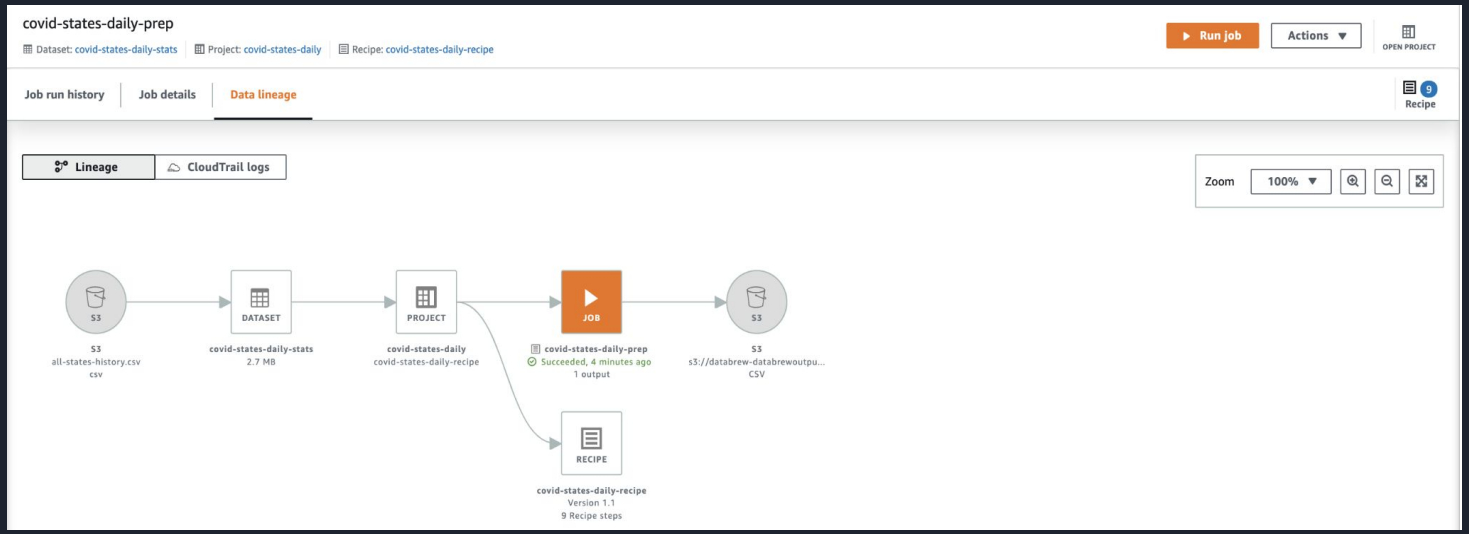


- Click on the link to the job output, and verify that the output files in the S3 bucket



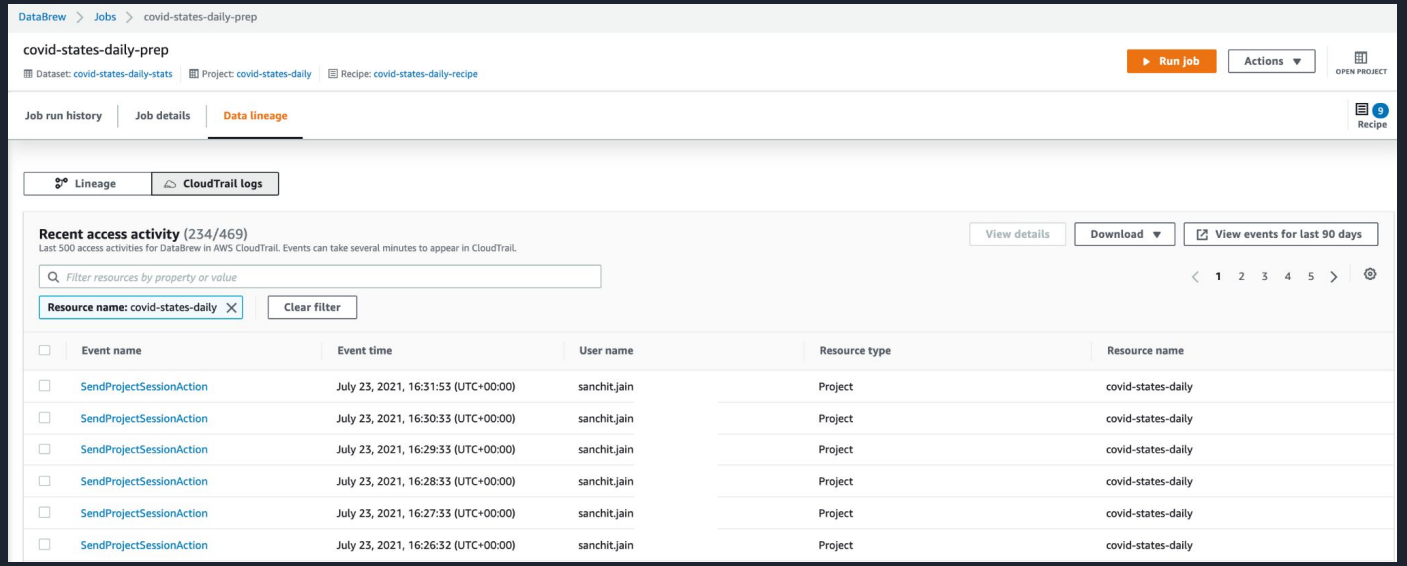
Viewing data lineage

- 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.



THANK YOU