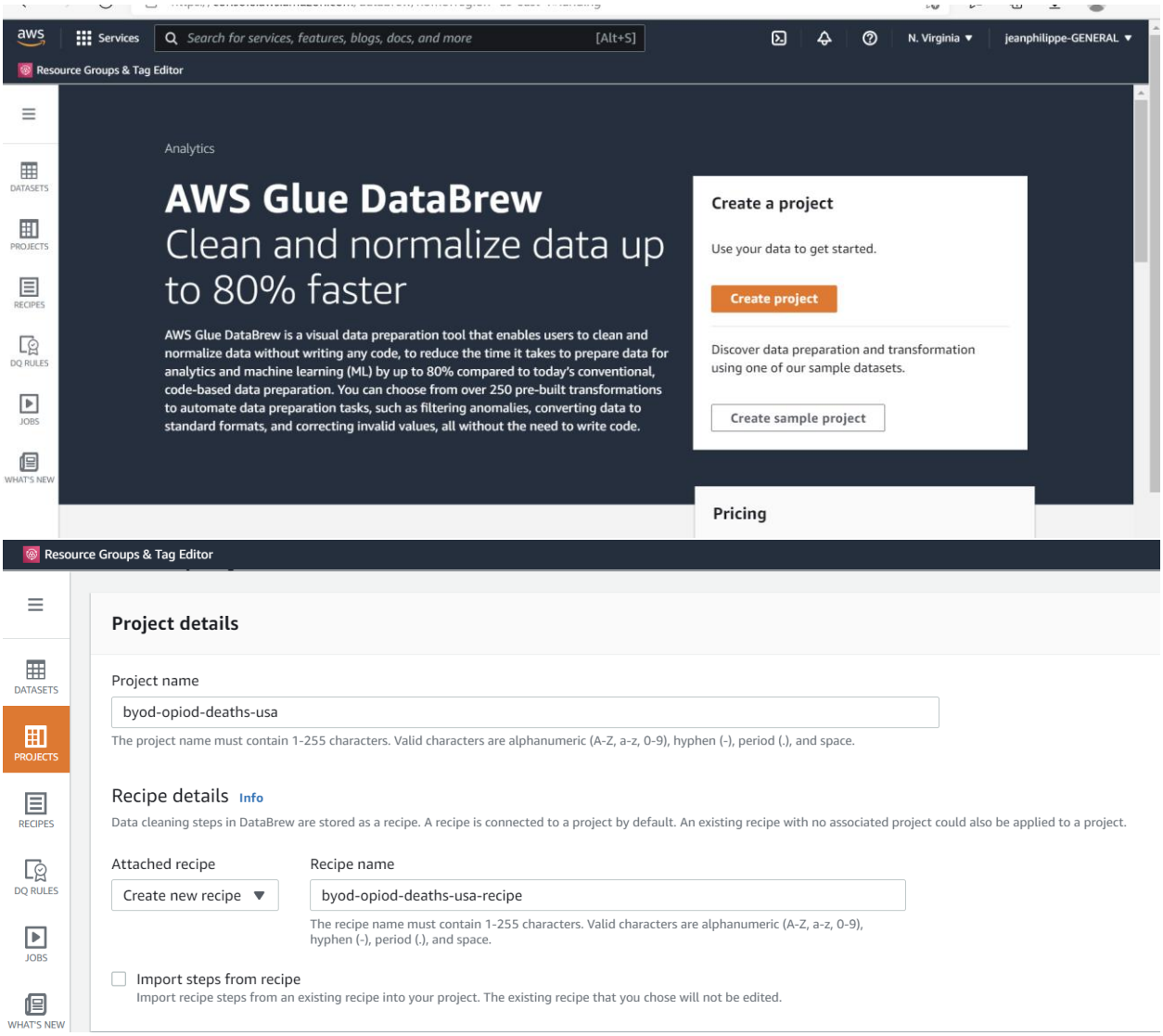
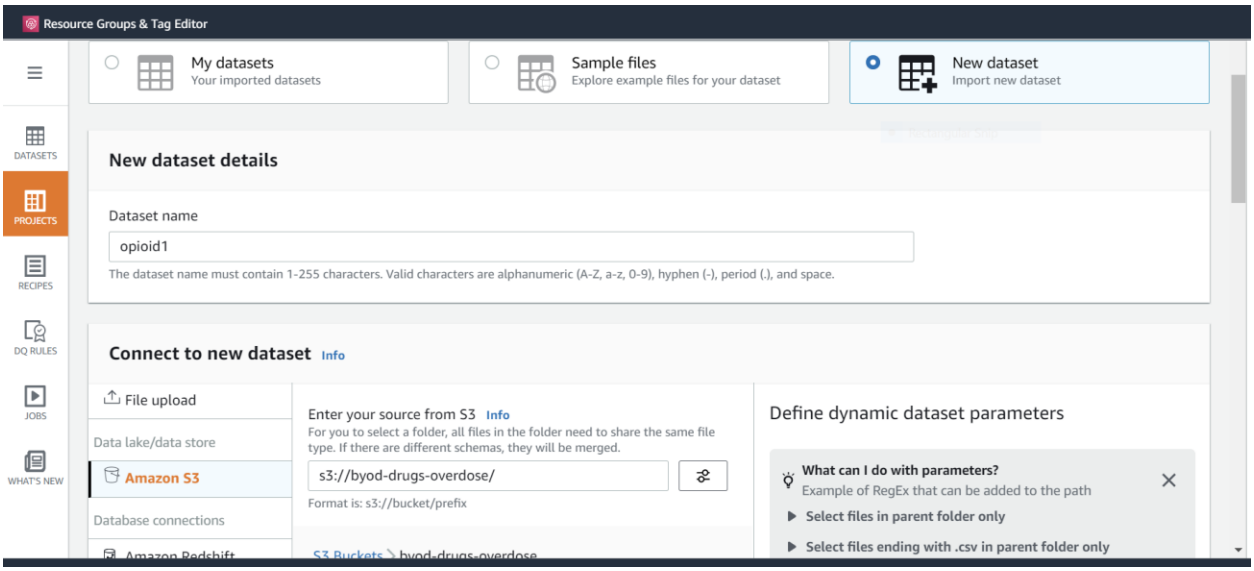
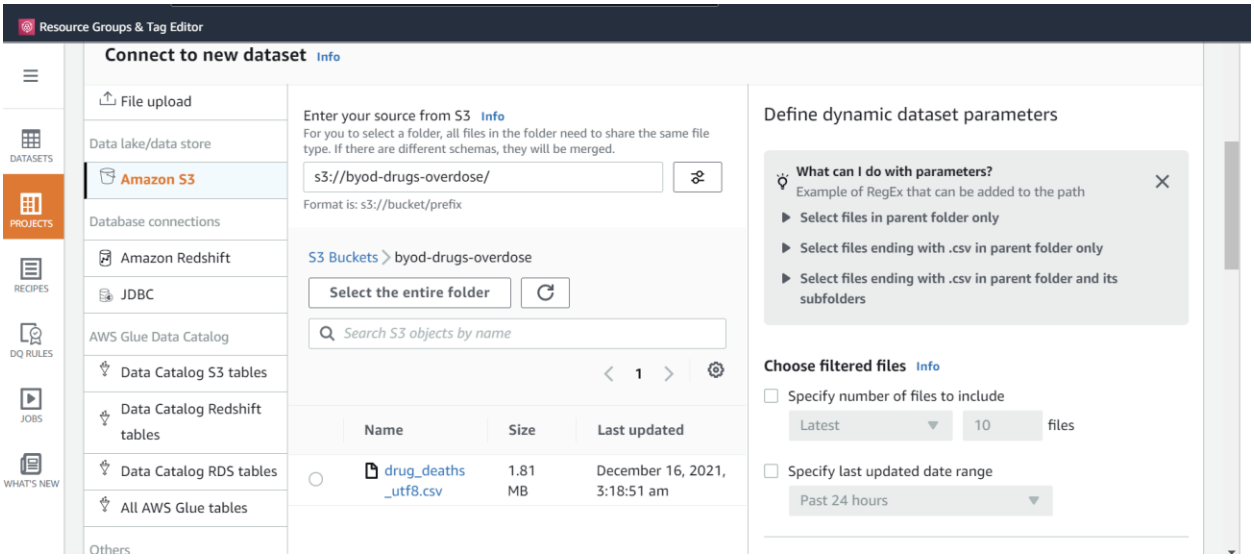
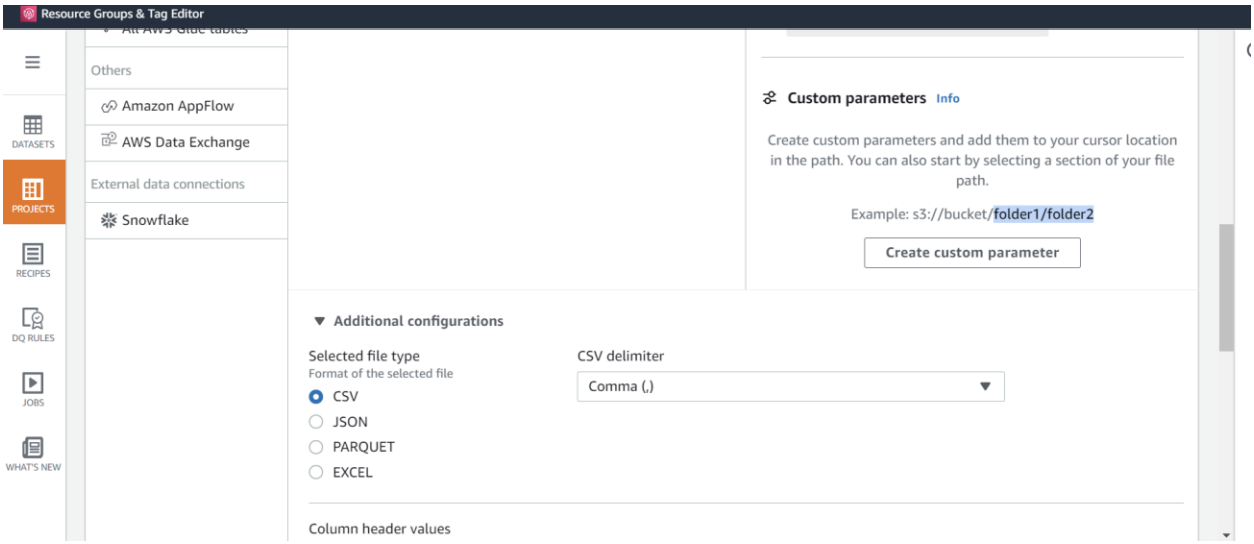
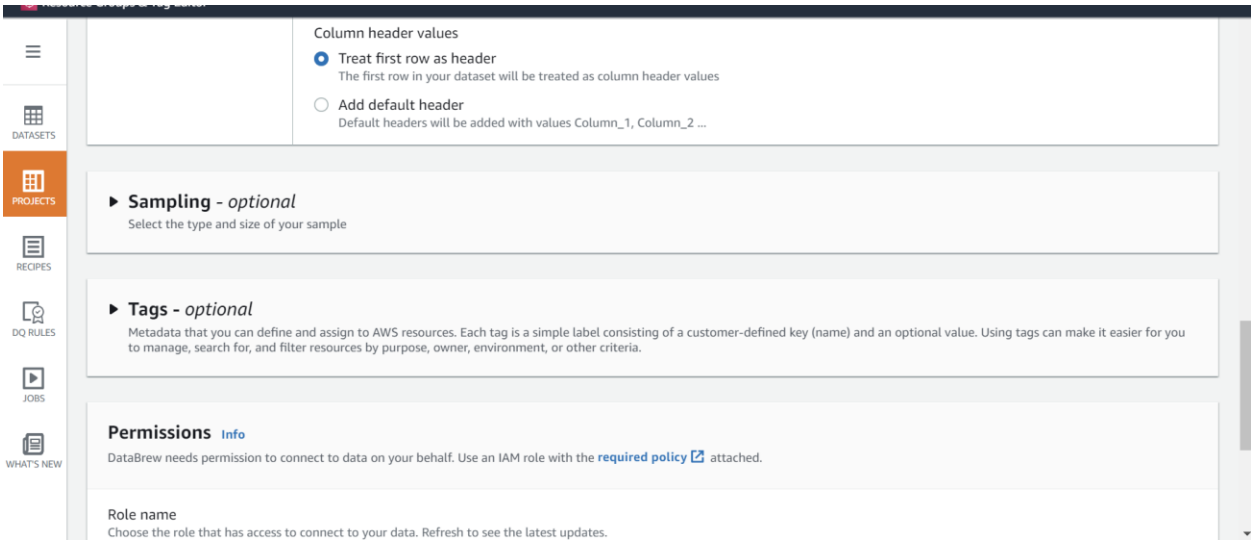

aws
Services
Search for services, features, blogs, docs, and more
[Alt+S]
N. Virginia
jeanphilippe-GENERAL
Resource Groups & Tag Editor
DATASETS
PROJECTS
RECIPES
DQ RULES
JOBS
WHAT'S NEW
Analytics
AWS Glue DataBrew
Clean and normalize data up to 80% faster
AWS Glue DataBrew is a visual data preparation tool that enables users to clean and normalize data without writing any code, to reduce the time it takes to prepare data for analytics and machine learning (ML) by up to 80% compared to today's conventional, code-based data preparation. You can choose from over 250 pre-built transformations to automate data preparation tasks, such as filtering anomalies, converting data to standard formats, and correcting invalid values, all without the need to write code.
Create a project
Use your data to get started.
Create project
Discover data preparation and transformation using one of our sample datasets.
Create sample project
Pricing

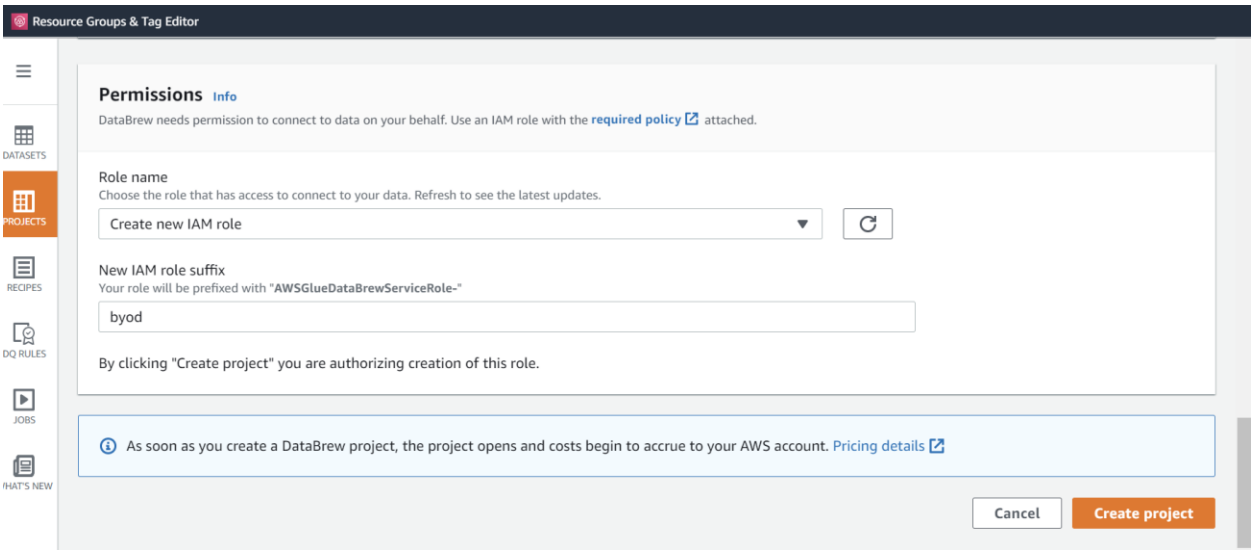
Resource Groups & Tag Editor
DATASETS
PROJECTS
RECIPES
DQ RULES
JOBS
WHAT'S NEW
Project details
Project name
byod-opiod-deaths-usa
The project name must contain 1-255 characters. Valid characters are alphanumeric (A-Z, a-z, 0-9), hyphen (-), period (.), and space.
Recipe details Info
Data cleaning steps 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
byod-opiod-deaths-usa-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 recipe
Import recipe steps from an existing recipe into your project. The existing recipe that you chose will not be edited.


Resource Groups & Tag Editor
My datasets
Your imported datasets
Sample files
Explore example files for your dataset
New dataset
Import new dataset
DATASETS
PROJECTS
RECIPES
DQ RULES
JOBS
WHAT'S NEW
New dataset details
Dataset name
opioid1
The dataset name must contain 1-255 characters. Valid characters are alphanumeric (A-Z, a-z, 0-9), hyphen (-), period (.), and space.
Connect to new dataset Info
File upload
Data lake/data store
Amazon S3
Database connections
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.
s3://byod-drugs-overdose/
Format is: s3://bucket/prefix
Define dynamic dataset parameters
What can I do with parameters?
Example of RegEx that can be added to the path
Select files in parent folder only
Select files ending with .csv in parent folder only

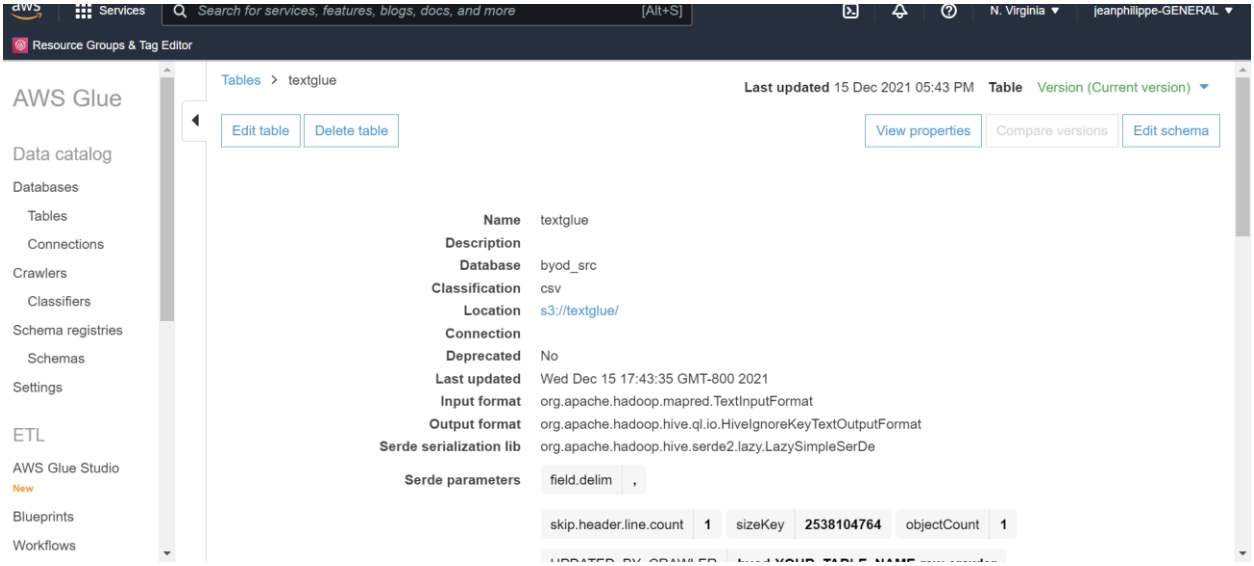

Resource Groups & Tag Editor
Connect to new dataset Info
DATASETS
PROJECTS
RECIPES
DQ RULES
JOBS
WHAT'S NEW
File upload
Data lake/data store
Amazon S3
Database connections
Amazon Redshift
JDBC
AWS Glue Data Catalog
Data Catalog S3 tables
Data Catalog Redshift tables
Data Catalog RDS tables
All AWS Glue tables
Others
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.
s3://byod-drugs-overdose/
Format is: s3://bucket/prefix
S3 Buckets > byod-drugs-overdose
Select the entire folder
Search S3 objects by name
1
Name
Size
Last updated
drug_deaths_utf8.csv
1.81 MB
December 16, 2021, 3:18:51 am
Define dynamic dataset parameters
What can I do with parameters?
Example of RegEx that can be added to the path
Select files in parent folder only
Select files ending with .csv in parent folder only
Select files ending with .csv in parent folder and its subfolders
Choose filtered files Info
Specify number of files to include
Latest
10
files
Specify last updated date range
Past 24 hours


Resource Groups & Tag Editor
DATASETS
PROJECTS
RECIPES
DQ RULES
JOBS
WHAT'S NEW
Others
Amazon AppFlow
AWS Data Exchange
External data connections
Snowflake
Custom parameters Info
Create custom parameters and add them to your cursor location in the path. You can also start by selecting a section of your file path.
Example: s3://bucket/folder1/folder2
Create custom parameter
Additional configurations
Selected file type
Format of the selected file
CSV
JSON
PARQUET
EXCEL
CSV delimiter
Comma (,)
Column header values


Column header values
Treat first row as header
The first row in your dataset will be treated as column header values
Add default header
Default headers will be added with values Column_1, Column_2 ...
Sampling - optional
Select the type and size of your sample
Tags - optional
Metadata that you can define and assign to AWS resources. Each tag is a simple label consisting of a customer-defined key (name) and an optional value. Using tags can make it easier for you to manage, search for, and filter resources by purpose, owner, environment, or other criteria.
Permissions Info
DataBrew needs permission to connect to data on your behalf. Use an IAM role with the required policy attached.
Role name
Choose the role that has access to connect to your data. Refresh to see the latest updates.


Resource Groups & Tag Editor
Permissions Info
DataBrew needs permission to connect to data on your behalf. Use an IAM role with the required policy attached.
Role name
Choose the role that has access to connect to your data. Refresh to see the latest updates.
Create new IAM role
New IAM role suffix
Your role will be prefixed with "AWSGlueDataBrewServiceRole-"
byod
By clicking "Create project" you are authorizing creation of this role.
As soon as you create a DataBrew project, the project opens and costs begin to accrue to your AWS account. Pricing details
Cancel
Create project

AWS GLUE


Services
Search for services, features, blogs, docs, and more
[Alt+S]
N. Virginia
jeanphilippe-GENERAL
Resource Groups & Tag Editor
AWS Glue
Data catalog
Databases
Tables
Connections
Crawlers
Classifiers
Schema registries
Schemas
Settings
ETL
AWS Glue Studio
New
Blueprints
Workflows
Tables > textglue
Last updated 15 Dec 2021 05:43 PM
Table
Version (Current version)
Edit table
Delete table
View properties
Compare versions
Edit schema
Name textglue
Description
Database byod_src
Classification csv
Location s3://textglue/
Connection
Deprecated No
Last updated Wed Dec 15 17:43:35 GMT-800 2021
Input format org.apache.hadoop.mapred.TextInputFormat
Output format org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat
Serde serialization lib org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe
Serde parameters field.delim ,
skip.header.line.count 1
sizeKey 2538104764
objectCount 1