**AWS Data Brew Jobs**

AWS Data Brew Jobs allows us to create a data quality checks that can be used for checking our dataset such as park at and CSV another file formats to maintain the data quality across the datasets with the data quality DQ (Data Quality) checks you can define custom rules for a data, and you can specify custom validation based on our data quality condition

Below diagram shows how data brew helps us to customize and validate our data based on needed condition and we can schedule the jobs for quality checks at frequent times

Diagram

Description automatically generated

Initial step of our process is to create a dataset that need to add data quality validations to create a dataset we have tables flash views that can be created using crawlers we going to use athena data set for creating a DQ dataset we can use custom query to create a deque data set that we going to validate

Ruleset its basically allows you to specify the rules that to be applied to the column to check the data quality you can apply multiple rules for the single column

Rules can be applied with dataset level column level and column values level

Dataset level checks includes we can validate

1. Number of rows
2. Number of columns
3. Number of duplicate rows within the dataset

In our case we can check a duplicate value with adgroupCFID and date columns such as year month and day

Column statistic check is not useful in our case

Column value-based data quality check is useful in our case for validating a data quality such as

1. adgroupCFID should not be null / empty
2. date should not be null / empty
3. deaID should not be null /empty
4. region should not be null / empty

On creation of DQ rules we can create profile jobs to run the data quality checks using glue jobs the profile jobs can be created using full dataset and partial data set using number of rows in our case we’re going to create our daily jobs that looks for data for last one day and a weekly job that runs for entire dataset

we can create scheduler for scheduling the profile glue jobs to run at convenient time in our case we going to create two scheduler that runs every day and every week everyday scheduler that looks for dataset based on date month and year column in r data set for last one day and second scheduler that runs for every week without having any filter on the dataset

**Alerts / Email on Failed Data Quality**

You can create alerts / emails on failed data quality check using a AWS event bridge, even bridge allows to listen for a trigger that can be triggered with data brew ruleset validation result

**Sample Error Validation**

**{  
 "version"**: "0",  
 **"id"**: "abcdef00-1234-5678-9abc-def012345678",  
 **"detail-type"**: "DataBrew Ruleset Validation Result",  
 **"source"**: "aws.databrew",  
 **"account"**: "123456789012",  
 **"time"**: "2017-09-07T18:57:21Z",  
 **"region"**: "us-west-2",  
 **"resources"**: **[]**,  
 **"detail"**: **{  
 "jobName"**: "MyJob",  
 **"rulesetArn"**: "arn:aws:databrew:us-west-2:123456789012:ruleset/MyRuleset",  
 **"datasetName"**: "MyDataset",  
 **"validationReportS3Location"**: "s3://myBucket/profile-job-output/MyDataset\_abcdef0123456789abcdef0123456789abcdef0123456789abcdef0123456789\_dq-validation-report.json",  
 **"rulesetName"**: "MyRuleset",  
 **"jobRunId"**: "db\_abcdef0123456789abcdef0123456789abcdef0123456789abcdef0123456789",  
 **"validationState"**: "ERROR"  
 **}  
}**

**Advantages of Data Brew**

* You can run multiple jobs at the same time
* Logging can be done automatically
* Internally uses [Deequ](https://github.com/awslabs/deequ) for data quality checks
* Highly scalable for large datasets
* Allows to do quality checks on partial and full data sets
* You can automate for creating rules and dataset using CDK

**Dis-Advantages of Data Brew**

* Jobs takes more time for large datasets