1. Data from kaggle.com , downloaded and uploaded on s3
2. Analysed data using glue crawler.
3. We had errors in the data we transformed data from JSON to the table using AWS Lambda function, stored data in a new s3 bucket called having the clean data.
4. Need to assess CSV files and transform if needed.

The CSV files get saved as partitioned data in a new table but in the same database where the JSON data was saved. The data is partitioned according to the region.,

"de-youtube-raw"."raw\_statistics" joined to

db-youtube-cleaned"."cleaned\_statistics\_reference\_data"

SELECT \* FROM "de-youtube-raw"."raw\_statistics" t1

inner join "db-youtube-cleaned"."cleaned\_statistics\_reference\_data" t2

on t1.category\_id = cast (t2.id as int);

s3://de-on-youtube-dev/youtube/raw\_statistics\_reference\_data/

AWS Region : Physical location around the world

Loaded data on S3 using the AWS CLI

Glue can be used to crawl through that data catalog. Glue can extract the metadata from your data.

Json data 🡪 S3(raw data) 🡪 AWS Lambda(to clean the data) 🡪 S3(Clean Data) 🡪 Glue Crawler

LAMBDA CODE: