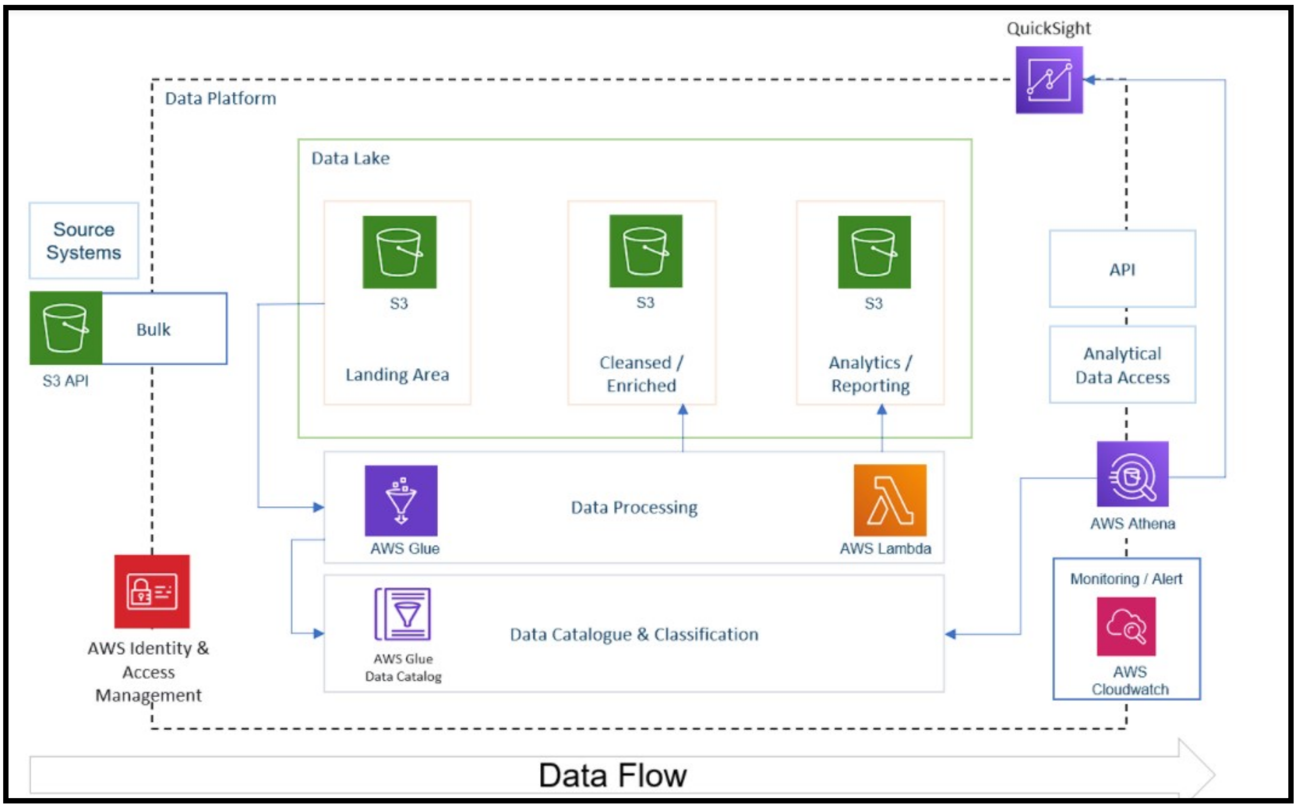
Project Introduction

This project aim to build a cloud-based safe data lake solution which classify data into several storage phases, such as raw, cleansed, and analytical. Then perform analysis on the structured and semi-structured YouTube videos data based on the video categories and the trending metrics.

Tech Stack

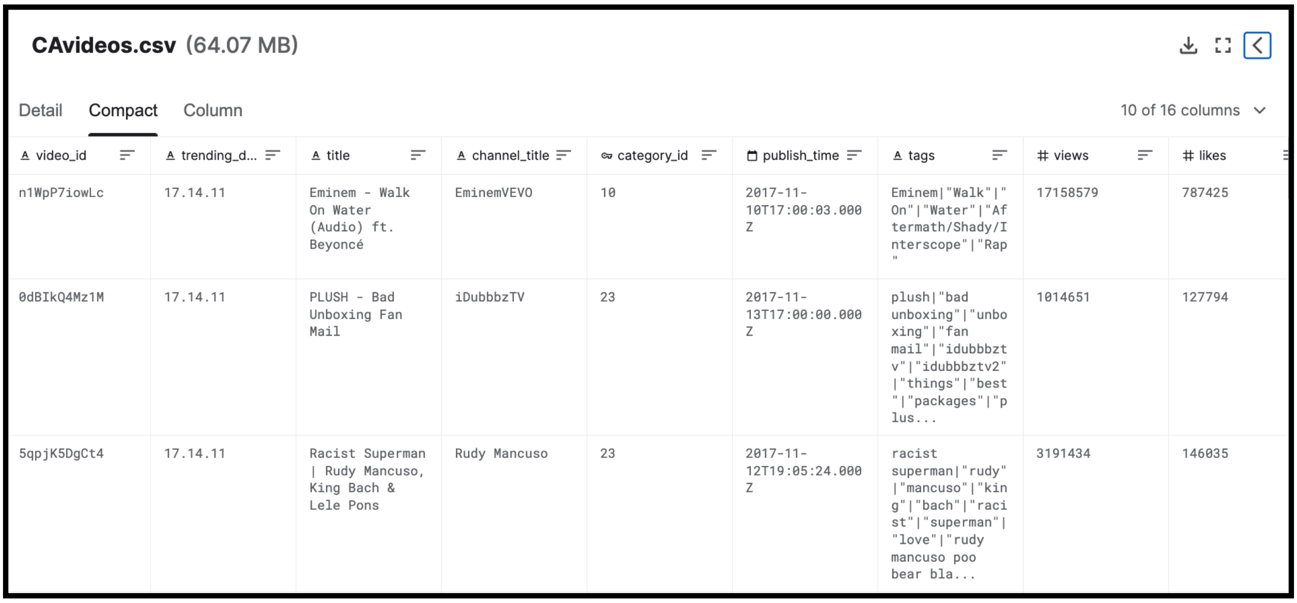
Language: Python, SQL
AWS Cloud Service: AWS S3, AWS Glue, QuickSight, AWS Lambda, AWS Athena, AWS IAM

Diagram



Dataset Description

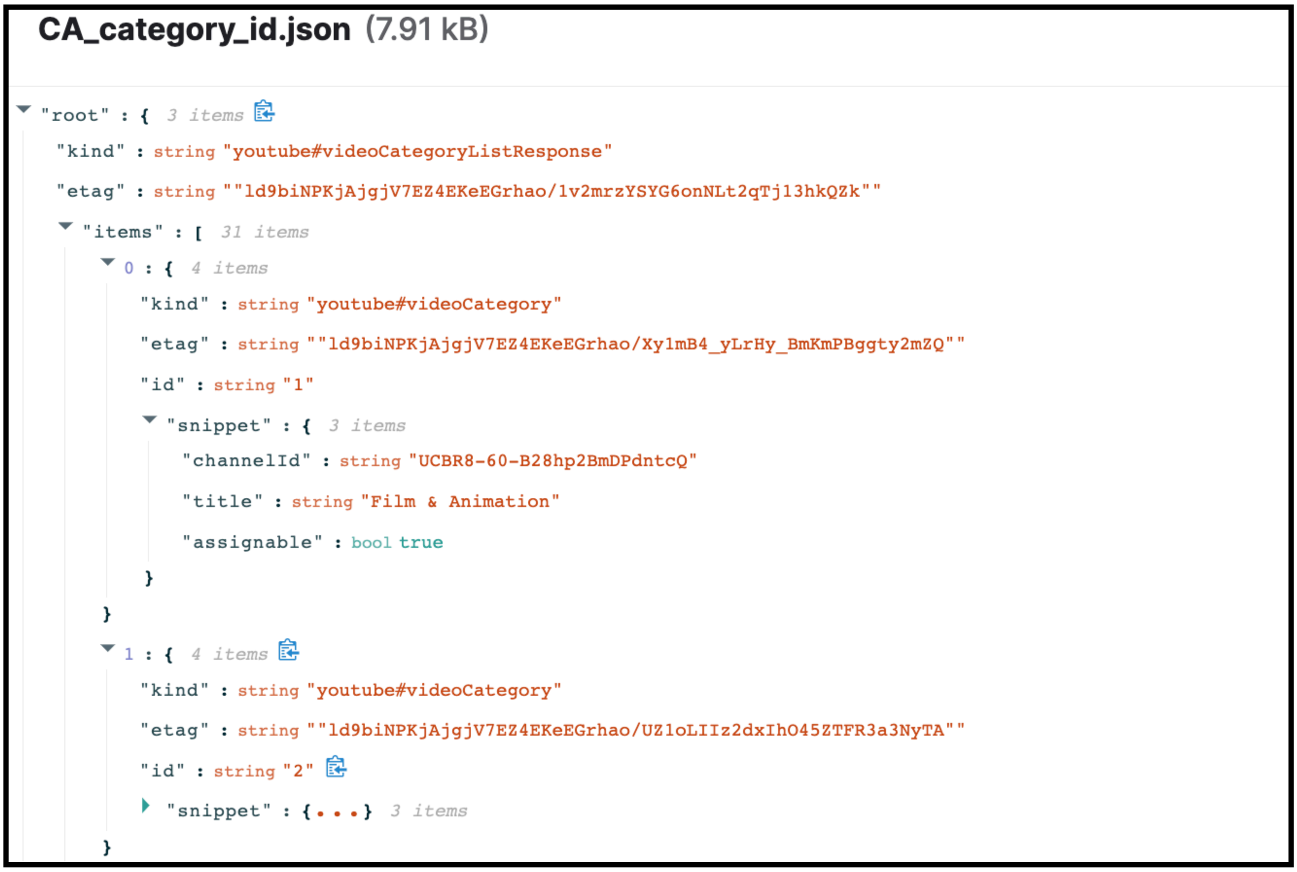
This Kaggle dataset contains statistics (CSV files) on daily popular YouTube videos over the course of many months. There are up to 200 trending videos published every day for many locations. The data for each region is in its own file. The video title, channel title, publication time, tags, views, likes and dislikes, description, and comment count are among the items included in the data. A category_id field, which differs by area, is also included in the JSON file linked to the region.



CAvideos.csv (64.07 MB)

Detail Compact Column 10 of 16 columns

video_id	trending_d...	title	channel_title	category_id	publish_time	tags	views	likes
n1WpP7iowLc	17.14.11	Eminem - Walk On Water (Audio) ft. Beyoncé	EminemVEVO	10	2017-11-10T17:00:03.000Z	Eminem|"Walk"|"On"|"Water"|"Aftermath/Shady/Interscope"|"Rap"	17158579	787425
0dBIkQ4Mz1M	17.14.11	PLUSH - Bad Unboxing Fan Mail	iDubbbzTV	23	2017-11-13T17:00:00.000Z	plush|"bad unboxing"|"unboxing"|"fan mail"|"idubbbztv"|"idubbbztv2"|"things"|"best"|"packages"|"plus...	1014651	127794
5qpjK5DgCt4	17.14.11	Racist Superman | Rudy Mancuso, King Bach & Lele Pons	Rudy Mancuso	23	2017-11-12T19:05:24.000Z	racist superman|"rudy"|"mancuso"|"king"|"bach"|"racist"|"superman"|"love"|"rudy mancuso poo bear bla...	3191434	146035



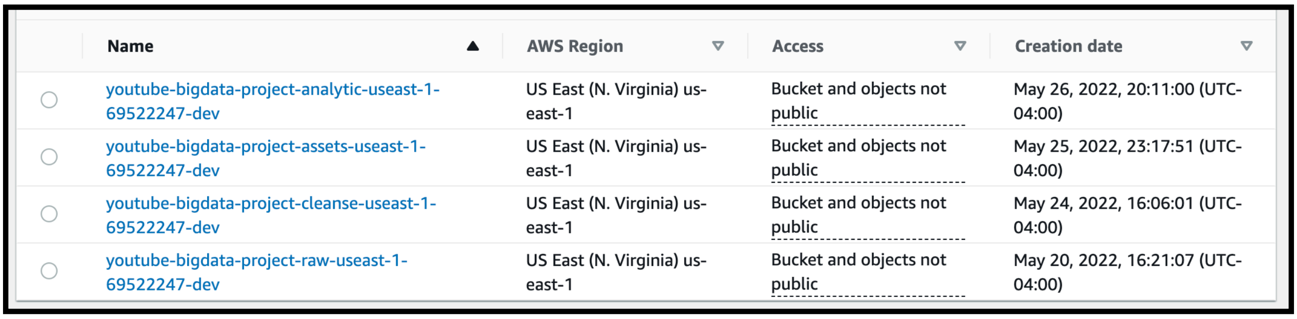
CA_category_id.json (7.91 kB)

```
▼ "root" : { 3 items
   "kind" : string "youtube#videoCategoryListResponse"
   "etag" : string ""ld9biNPKjAjgjV7EZ4EKeEGrhao/1v2mrzYSYG6onNLt2qTj13hkQZk""
  ▼ "items" : [ 31 items
     ▼ 0 : { 4 items
         "kind" : string "youtube#videoCategory"
         "etag" : string ""ld9biNPKjAjgjV7EZ4EKeEGrhao/Xy1mB4_yLrHy_BmKmPBggty2mZQ""
         "id" : string "1"
       ▼ "snippet" : { 3 items
           "channelId" : string "UCBR8-60-B28hp2BmDPdntcQ"
           "title" : string "Film & Animation"
           "assignable" : bool true
         }
     }
     ▼ 1 : { 4 items
         "kind" : string "youtube#videoCategory"
         "etag" : string ""ld9biNPKjAjgjV7EZ4EKeEGrhao/UZ1oLIIz2dxIhO45ZTFR3a3NyTA""
         "id" : string "2"
       ▶ "snippet" : {...} 3 items
     }
```

As you can see the **csv** files contain more video informations, but they only have category_id which is not convenient to do analysis report. On the other side, the **json** file contain less information but they have key title for different categories which are helpful for analysis. The two different file format also have the same column(category_id, id) that can be used to join together.

1. Create Data Lake with AWS S3 buckets

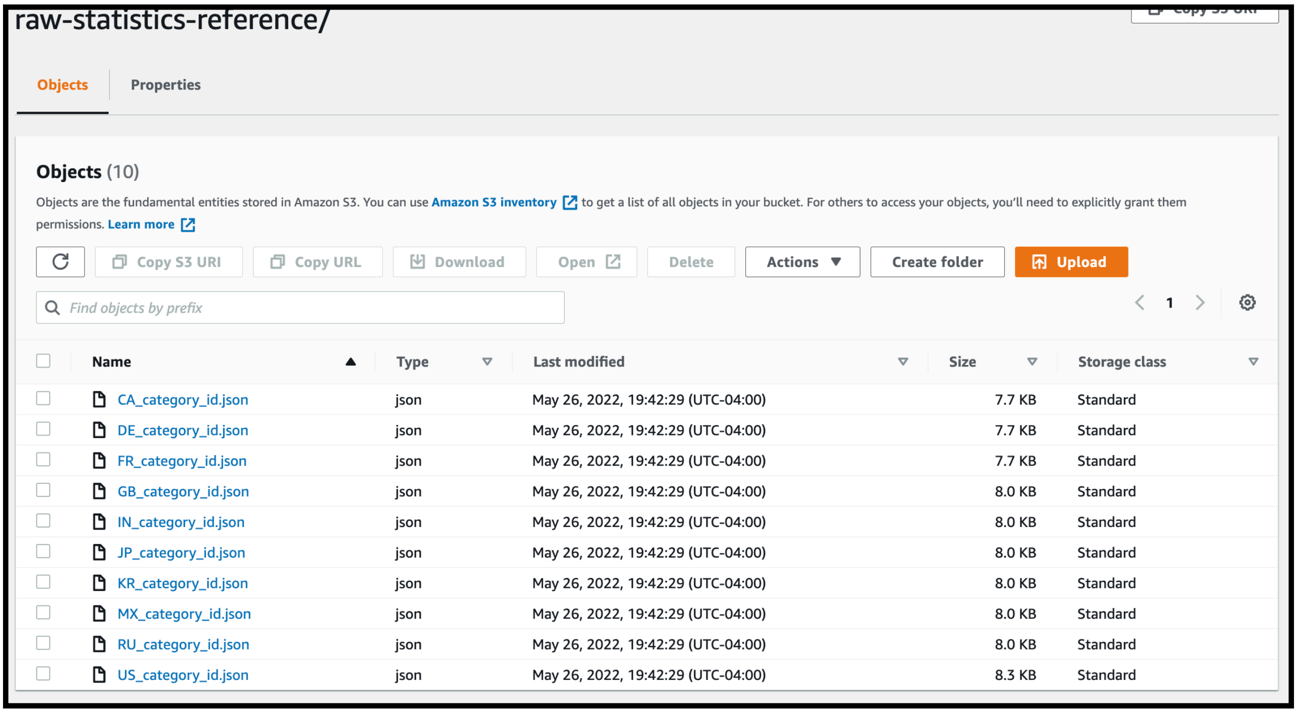
Here we create 3 buckets(raw, cleanse, analytics) for different phases of data and 1 bucket (assets) for storing ETL scripts.



	Name ▲	AWS Region ▽	Access ▽	Creation date ▽
○	youtube-bigdata-project-analytic-useast-1-69522247-dev	US East (N. Virginia) us-east-1	Bucket and objects not public	May 26, 2022, 20:11:00 (UTC-04:00)
○	youtube-bigdata-project-assets-useast-1-69522247-dev	US East (N. Virginia) us-east-1	Bucket and objects not public	May 25, 2022, 23:17:51 (UTC-04:00)
○	youtube-bigdata-project-cleanse-useast-1-69522247-dev	US East (N. Virginia) us-east-1	Bucket and objects not public	May 24, 2022, 16:06:01 (UTC-04:00)
○	youtube-bigdata-project-raw-useast-1-69522247-dev	US East (N. Virginia) us-east-1	Bucket and objects not public	May 20, 2022, 16:21:07 (UTC-04:00)

2. Data lake design in layers, partitioned for cost-performance

First, we upload the json files to the bucket for raw data under the *raw-statistics-reference/* prefix.



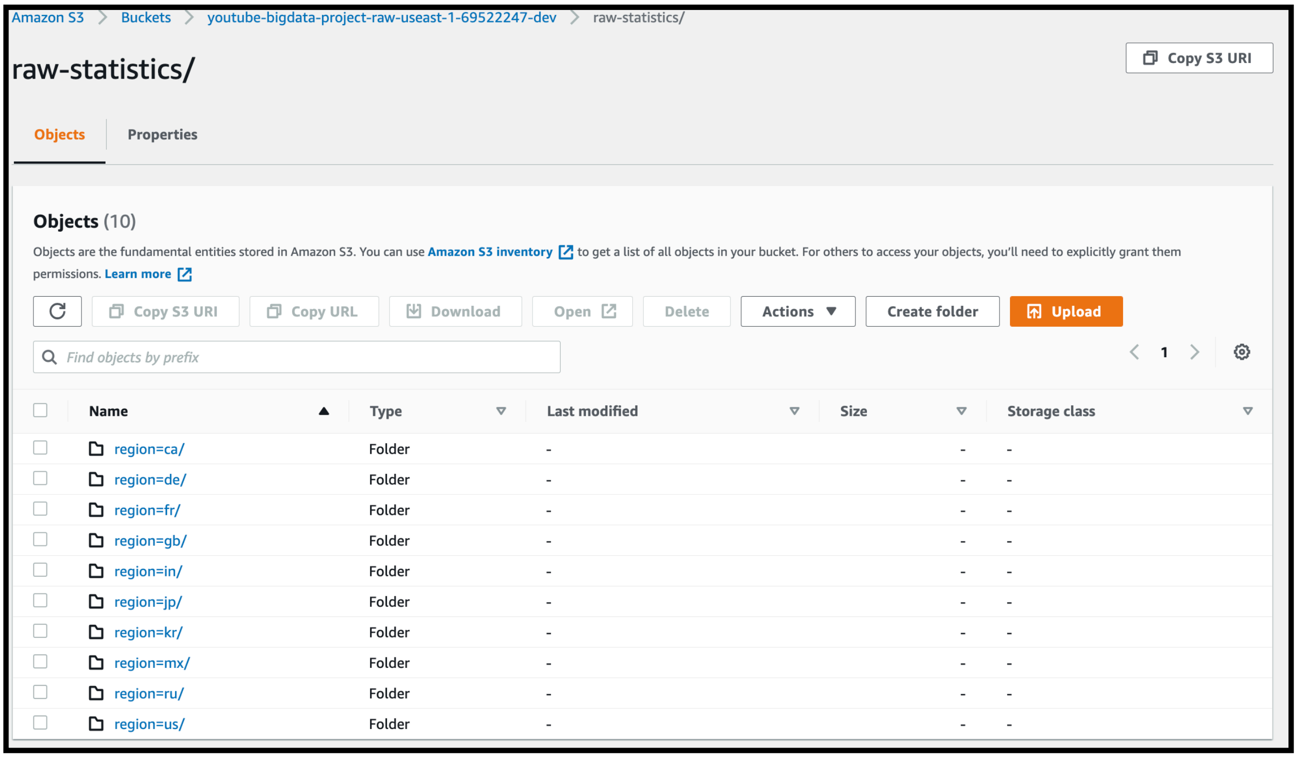
raw-statistics-reference/

Objects (10)

Objects are the fundamental entities stored in Amazon S3. You can use **Amazon S3 inventory** to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. **Learn more**

□	Name ▲	Type ▽	Last modified ▽	Size ▽	Storage class ▽
□	CA_category_id.json	json	May 26, 2022, 19:42:29 (UTC-04:00)	7.7 KB	Standard
□	DE_category_id.json	json	May 26, 2022, 19:42:29 (UTC-04:00)	7.7 KB	Standard
□	FR_category_id.json	json	May 26, 2022, 19:42:29 (UTC-04:00)	7.7 KB	Standard
□	GB_category_id.json	json	May 26, 2022, 19:42:29 (UTC-04:00)	8.0 KB	Standard
□	IN_category_id.json	json	May 26, 2022, 19:42:29 (UTC-04:00)	8.0 KB	Standard
□	JP_category_id.json	json	May 26, 2022, 19:42:29 (UTC-04:00)	8.0 KB	Standard
□	KR_category_id.json	json	May 26, 2022, 19:42:29 (UTC-04:00)	8.0 KB	Standard
□	MX_category_id.json	json	May 26, 2022, 19:42:29 (UTC-04:00)	8.0 KB	Standard
□	RU_category_id.json	json	May 26, 2022, 19:42:29 (UTC-04:00)	8.0 KB	Standard
□	US_category_id.json	json	May 26, 2022, 19:42:29 (UTC-04:00)	8.3 KB	Standard

Second, we upload csv files to the bucket for raw data under the *raw-statistics/* prefix, but this time we will use hive style pattern to partitioned them with different region.

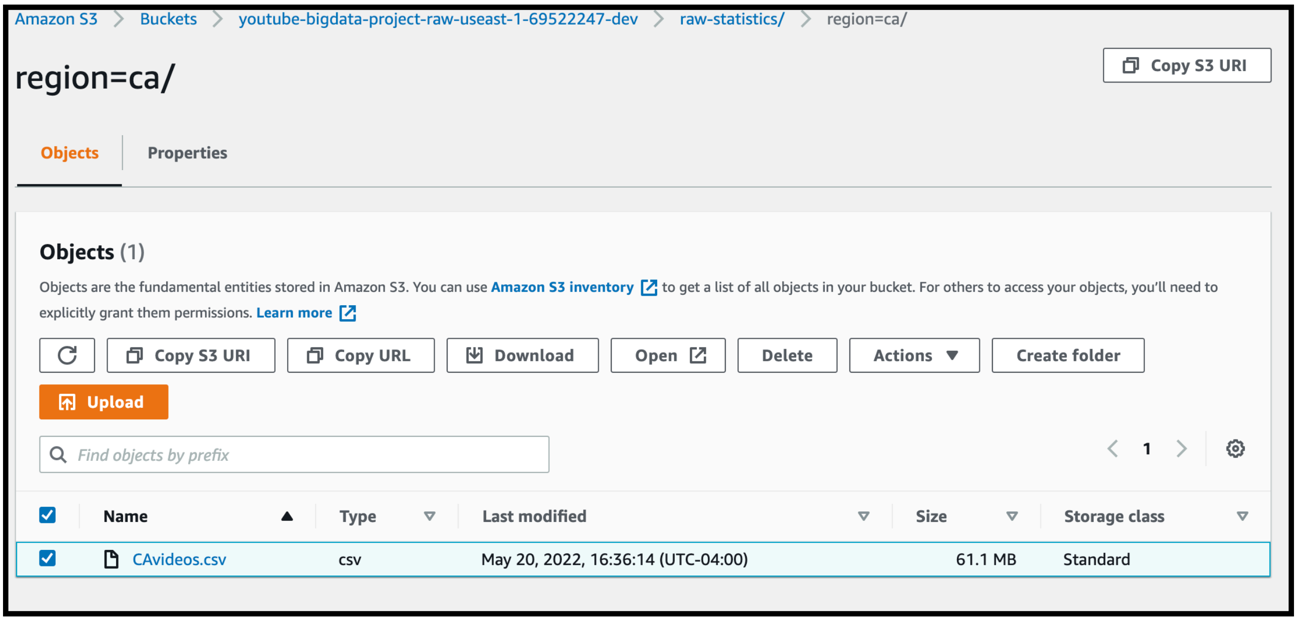


raw-statistics/

Objects (10)

Objects are the fundamental entities stored in Amazon S3. You can use **Amazon S3 inventory** to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. **Learn more**

□	Name ▲	Type	Last modified	Size	Storage class
□	region=ca/	Folder	-	-	-
□	region=de/	Folder	-	-	-
□	region=fr/	Folder	-	-	-
□	region=gb/	Folder	-	-	-
□	region=in/	Folder	-	-	-
□	region=jp/	Folder	-	-	-
□	region=kr/	Folder	-	-	-
□	region=mx/	Folder	-	-	-
□	region=ru/	Folder	-	-	-
□	region=us/	Folder	-	-	-



region=ca/

Objects (1)

Objects are the fundamental entities stored in Amazon S3. You can use **Amazon S3 inventory** to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. **Learn more**

☑	Name ▲	Type	Last modified	Size	Storage class
☑	CAvideos.csv	csv	May 20, 2022, 16:36:14 (UTC-04:00)	61.1 MB	Standard

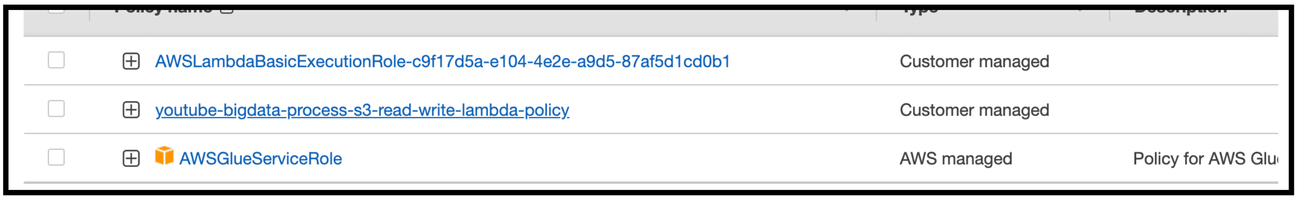
3. Create AWS Lambda functions to transform JSON file to parquet

Ideally, parquet format is more time and cost efficient to do analysis query because it is column base format.

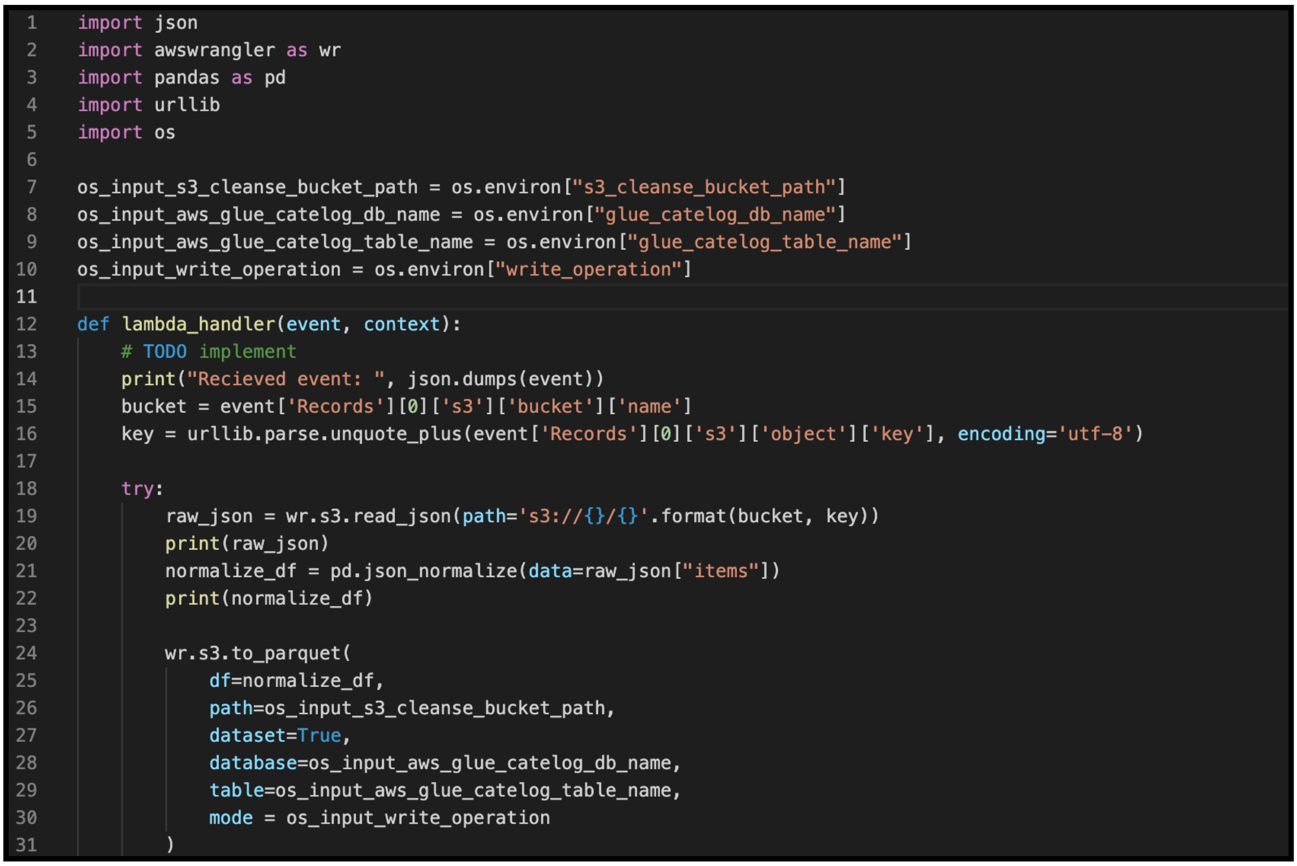
I used AWS wrangler library in lambda function to do the job. In the following function, it parse the bucket name and bucket key from the trigger event. Extract the JSON file base on the bucket name and key, and then transform it to parquet format. Meanwhile, load the data back to S3 cleanse bucket that we created earlier and AWS Glue catalog. Here we use environmental variable to make this AWS Lambda function more flexible, so we can set the target s3 path, AWS Glue catalog database name, table and mode from the console.

IAM

We create a new role with basic Lambda permissions and attach 2 extra policies to let the lambda function access the resource from S3 and GlueService.



□	Policy name	Type	Description
□	⊞ AWSLambdaBasicExecutionRole-c9f17d5a-e104-4e2e-a9d5-87af5d1cd0b1	Customer managed	
□	⊞ youtube-bigdata-process-s3-read-write-lambda-policy	Customer managed	
□	⊞ AWSGlueServiceRole	AWS managed	Policy for AWS Glu



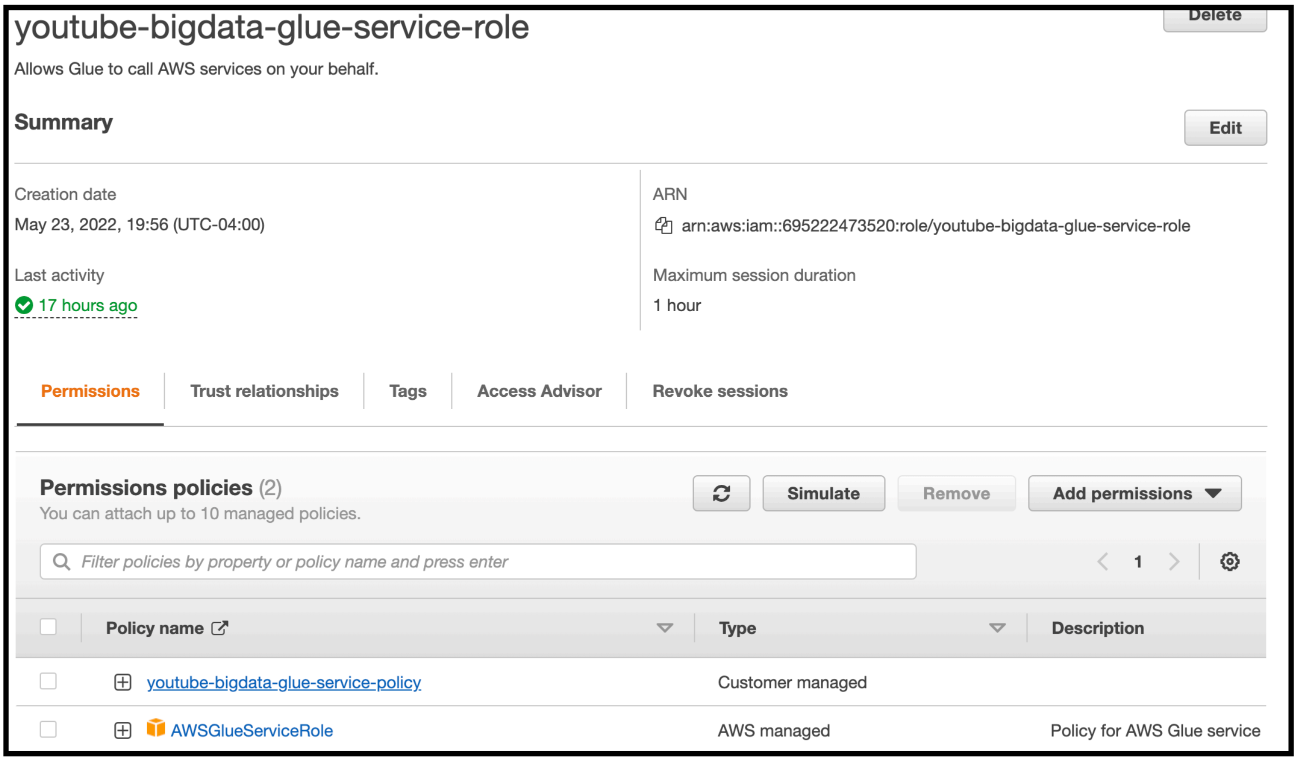
```python
import json
import awswrangler as wr
import pandas as pd
import urllib
import os

os_input_s3_cleanse_bucket_path = os.environ["s3_cleanse_bucket_path"]
os_input_aws_glue_catelog_db_name = os.environ["glue_catelog_db_name"]
os_input_aws_glue_catelog_table_name = os.environ["glue_catelog_table_name"]
os_input_write_operation = os.environ["write_operation"]

def lambda_handler(event, context):
    # TODO implement
    print("Recieved event: ", json.dumps(event))
    bucket = event['Records'][0]['s3']['bucket']['name']
    key = urllib.parse.unquote_plus(event['Records'][0]['s3']['object']['key'], encoding='utf-8')

    try:
        raw_json = wr.s3.read_json(path='s3://{}/{}'.format(bucket, key))
        print(raw_json)
        normalize_df = pd.json_normalize(data=raw_json["items"])
        print(normalize_df)

        wr.s3.to_parquet(
            df=normalize_df,
            path=os_input_s3_cleanse_bucket_path,
            dataset=True,
            database=os_input_aws_glue_catelog_db_name,
            table=os_input_aws_glue_catelog_table_name,
            mode = os_input_write_operation
        )
```

4. Create AWS Spark job through AWS Glue Studio to transform csv file to parquet

Go to AWS Glue Studio and a create job.

IAM

In order to let Glue jobs to have permission to access other service such as S3 and Glue Catalog, We need to go to IAM create a role and attach the suitable policies to it. Here, we need the Glue job has permission to access Glue Catalog and S3 to do the ETL.



youtube-bigdata-glue-service-role

Allows Glue to call AWS services on your behalf.

Summary

Creation date: May 23, 2022, 19:56 (UTC-04:00)

ARN: arn:aws:iam::695222473520:role/youtube-bigdata-glue-service-role

Last activity: 17 hours ago

Maximum session duration: 1 hour

Permissions policies (2)
You can attach up to 10 managed policies.

Policy name	Type	Description
youtube-bigdata-glue-service-policy	Customer managed	
AWSGlueServiceRole	AWS managed	Policy for AWS Glue service

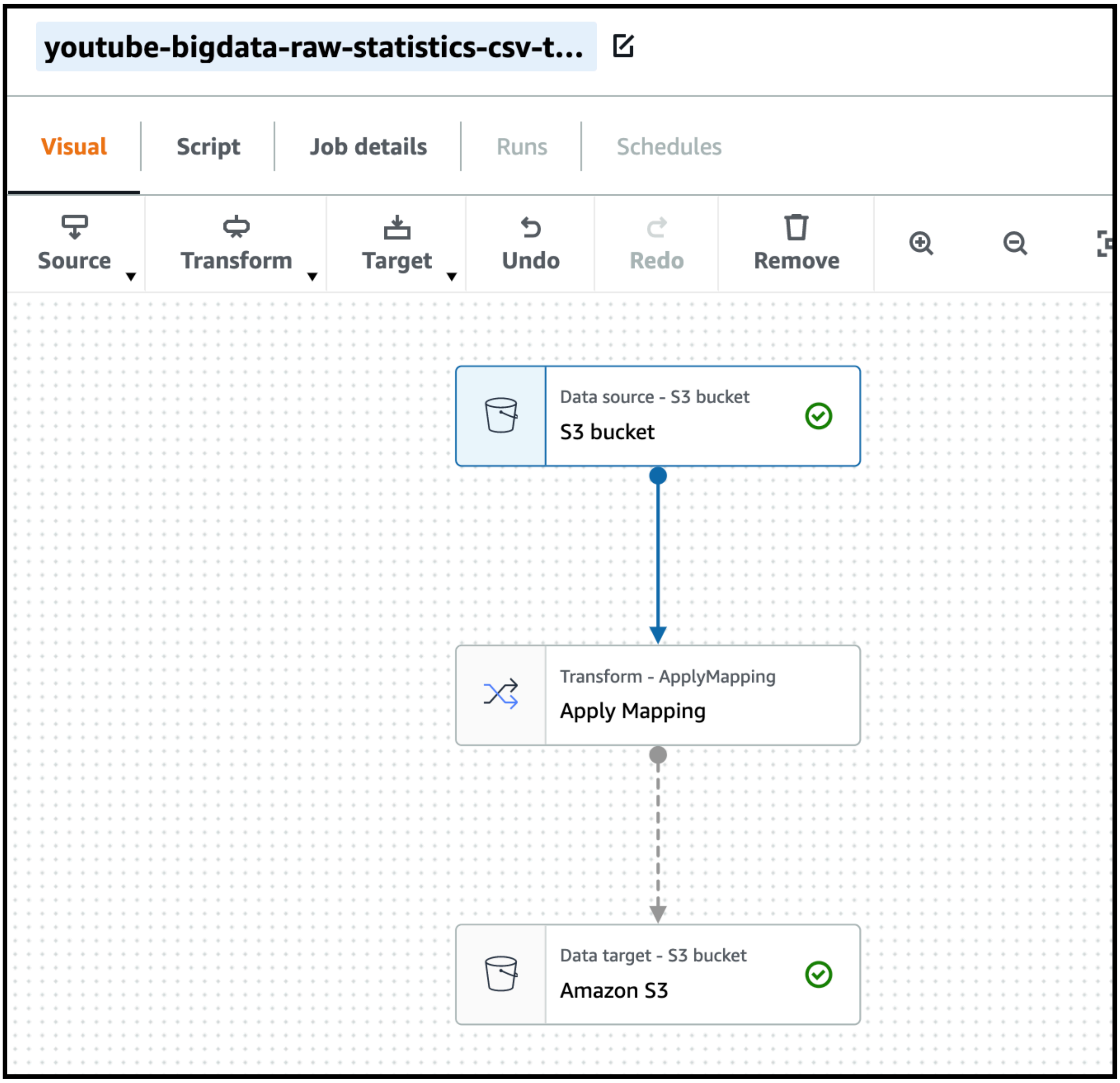
Job
There is a new feature that you can create the script with more visual way.

First select your data source which is the S3 path we store the csv file.

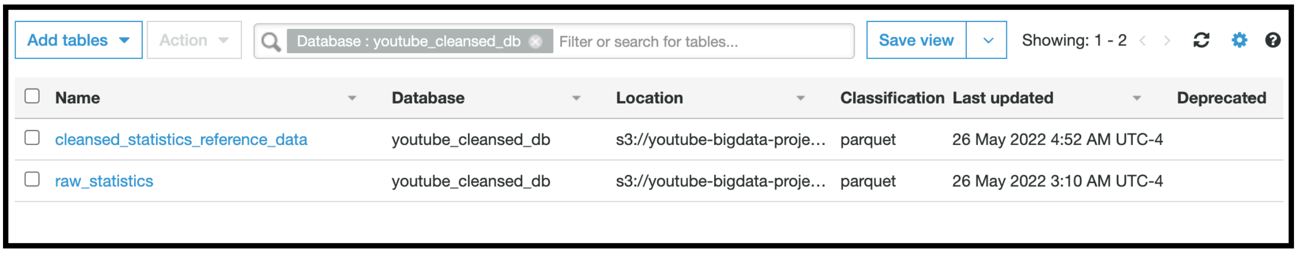
Second, select where you want to store the ETL script, (here we store the script in the S3 bucket we create for scripts before).

Finally, select where you want to output the data.(We output the data to S3 cleanse bucket). Don’t forget to change your output to parquet.

After all the settings, we now run the job.



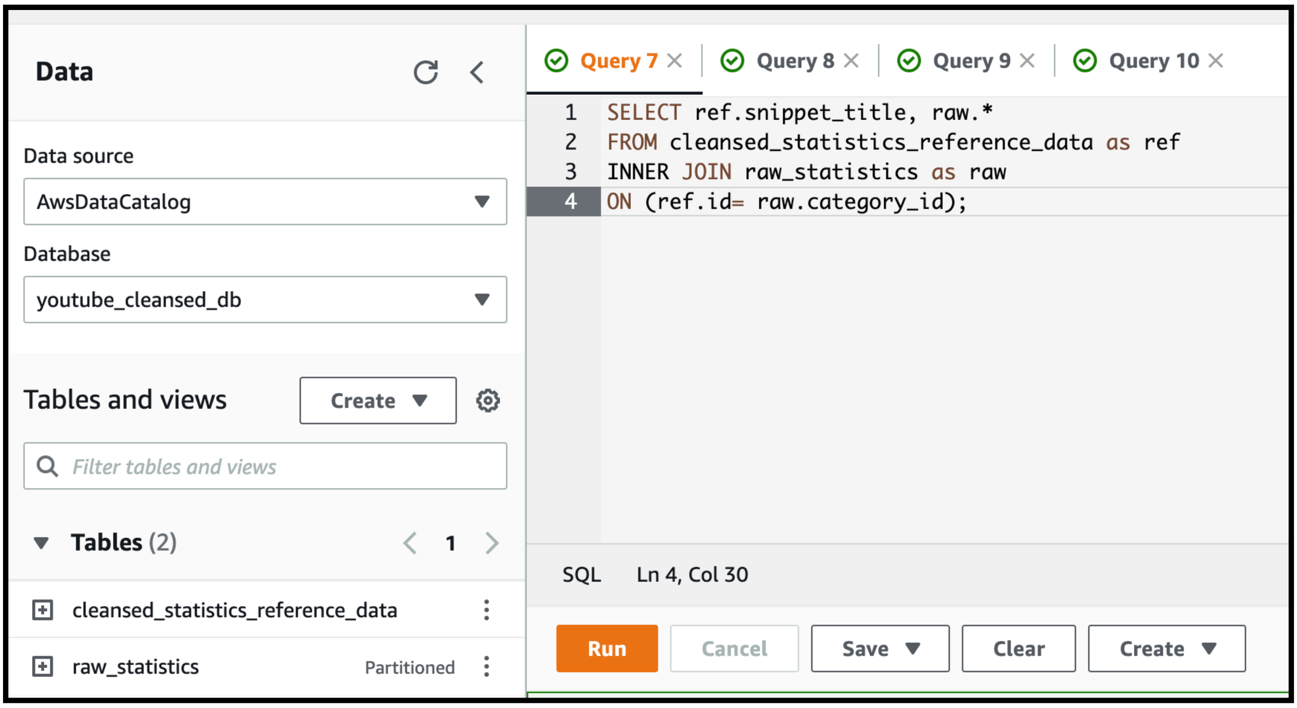
5. Use SQL to join to table from AWS Athena.


Add tables
Action
Database : youtube_cleansed_db
Filter or search for tables...
Save view
Showing: 1 - 2

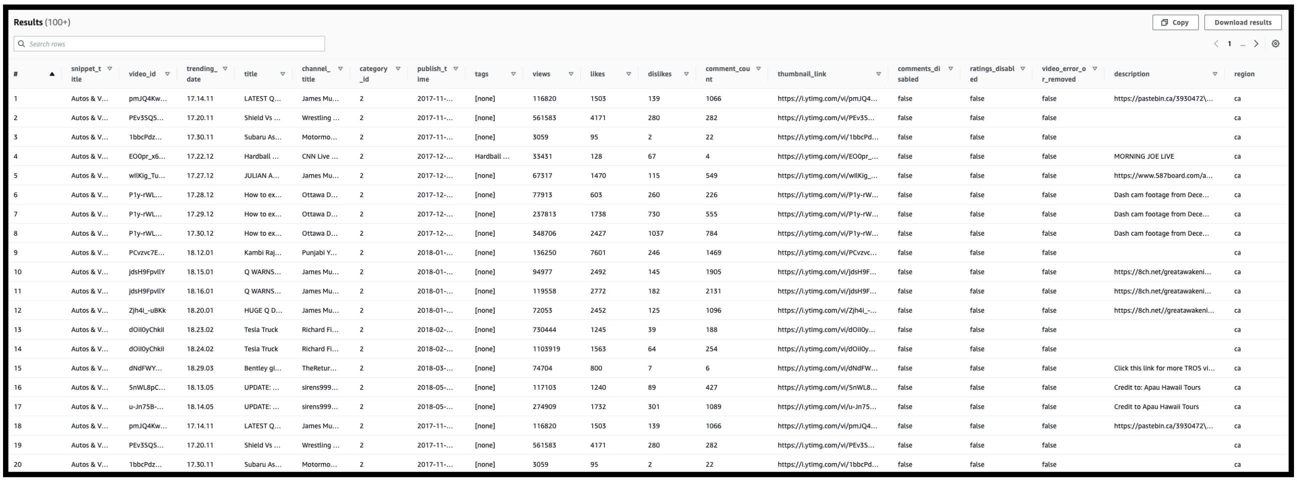
□ Name	Database	Location	Classification	Last updated	Deprecated
□ cleansed_statistics_reference_data	youtube_cleansed_db	s3://youtube-bigdata-proje...	parquet	26 May 2022 4:52 AM UTC-4	
□ raw_statistics	youtube_cleansed_db	s3://youtube-bigdata-proje...	parquet	26 May 2022 3:10 AM UTC-4	

After the AWS lambda function and AWS Glue job pre-processing the data, we should see two table under youtube_cleansed_db in AWS Catalog.

The cleansed_statistic_reference_data table is the data from JSON, the raw_statistics is from CSV. Now they are all in the same format which is parquet, we can use SQL to join to table through AWS Athena.



```sql
SELECT ref.snippet_title, raw.*
FROM cleansed_statistics_reference_data as ref
INNER JOIN raw_statistics as raw
ON (ref.id= raw.category_id);
```

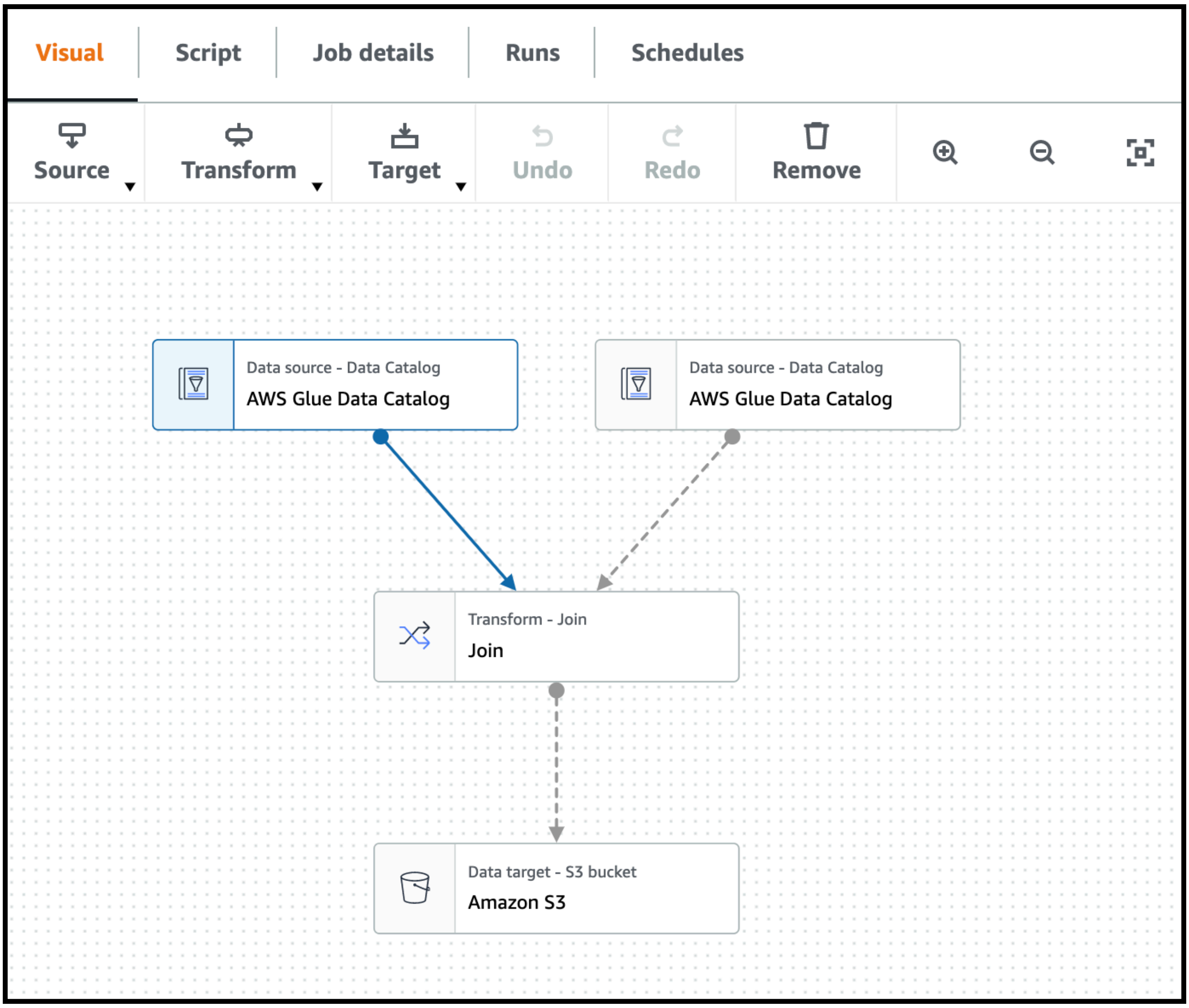


#	snippet_title	video_id	trending_date	title	channel_title	category_id	publish_time	tags	views	likes	dislikes	comment_count	thumbnail_link	comments_disabled	ratings_disabled	video_error_or_removed	description	region
1	Autos & V...	pmJQ4Kw...	17.14.11	LATEST Q...	James Mu...	2	2017-11-...	[none]	116820	1503	139	1066	https://i.ytimg.com/vi/pmJQ4...	false	false	false	https://pastebin.ca/3930472\...	ca
2	Autos & V...	PEv3SQS...	17.20.11	Shield Vs ...	Wrestling ...	2	2017-11-...	[none]	561583	4171	280	282	https://i.ytimg.com/vi/PEv3S...	false	false	false		ca
3	Autos & V...	1bbcPdz...	17.30.11	Subaru As...	Motormo...	2	2017-11-...	[none]	3059	95	2	22	https://i.ytimg.com/vi/1bbcPd...	false	false	false		ca
4	Autos & V...	EO0pr_x6...	17.22.12	Hardball ...	CNN Live ...	2	2017-12-...	Hardball ...	33431	128	67	4	https://i.ytimg.com/vi/EO0pr_...	false	false	false	MORNING JOE LIVE	ca
5	Autos & V...	wlIKig_Tu...	17.27.12	JULIAN A...	James Mu...	2	2017-12-...	[none]	67317	1470	115	549	https://i.ytimg.com/vi/wlIKig_...	false	false	false	https://www.587board.com/a...	ca
6	Autos & V...	P1y-rWL...	17.28.12	How to ex...	Ottawa D...	2	2017-12-...	[none]	77913	603	260	226	https://i.ytimg.com/vi/P1y-rW...	false	false	false	Dash cam footage from Dece...	ca
7	Autos & V...	P1y-rWL...	17.29.12	How to ex...	Ottawa D...	2	2017-12-...	[none]	237813	1738	730	555	https://i.ytimg.com/vi/P1y-rW...	false	false	false	Dash cam footage from Dece...	ca
8	Autos & V...	P1y-rWL...	17.30.12	How to ex...	Ottawa D...	2	2017-12-...	[none]	348706	2427	1037	784	https://i.ytimg.com/vi/P1y-rW...	false	false	false	Dash cam footage from Dece...	ca
9	Autos & V...	PCvzvc7E...	18.12.01	Kambi Raj...	Punjabi Y...	2	2018-01-...	[none]	136250	7601	246	1469	https://i.ytimg.com/vi/PCvzvc...	false	false	false		ca
10	Autos & V...	jdsH9FpvllY	18.15.01	Q WARNS...	James Mu...	2	2018-01-...	[none]	94977	2492	145	1905	https://i.ytimg.com/vi/jdsH9F...	false	false	false	https://8ch.net/greatawakeni...	ca
11	Autos & V...	jdsH9FpvllY	18.16.01	Q WARNS...	James Mu...	2	2018-01-...	[none]	119558	2772	182	2131	https://i.ytimg.com/vi/jdsH9F...	false	false	false	https://8ch.net/greatawakeni...	ca
12	Autos & V...	Zjh4i_-uBKk	18.20.01	HUGE Q D...	James Mu...	2	2018-01-...	[none]	72053	2452	125	1096	https://i.ytimg.com/vi/Zjh4i_-...	false	false	false	https://8ch.net//greatawakeni...	ca
13	Autos & V...	dOiI0yChkiI	18.23.02	Tesla Truck	Richard Fi...	2	2018-02-...	[none]	730444	1245	39	188	https://i.ytimg.com/vi/dOiI0y...	false	false	false		ca
14	Autos & V...	dOiI0yChkiI	18.24.02	Tesla Truck	Richard Fi...	2	2018-02-...	[none]	1103919	1563	64	254	https://i.ytimg.com/vi/dOiI0y...	false	false	false		ca
15	Autos & V...	dNdFWY...	18.29.03	Bentley gi...	TheRetur...	2	2018-03-...	[none]	74704	800	7	6	https://i.ytimg.com/vi/dNdFW...	false	false	false	Click this link for more TROS vi...	ca
16	Autos & V...	5nWL8pC...	18.13.05	UPDATE: ...	sirens999...	2	2018-05-...	[none]	117103	1240	89	427	https://i.ytimg.com/vi/5nWL8...	false	false	false	Credit to: Apau Hawaii Tours	ca
17	Autos & V...	u-Jn75B-...	18.14.05	UPDATE: ...	sirens999...	2	2018-05-...	[none]	274909	1732	301	1089	https://i.ytimg.com/vi/u-Jn75...	false	false	false	Credit to Apau Hawaii Tours	ca
18	Autos & V...	pmJQ4Kw...	17.14.11	LATEST Q...	James Mu...	2	2017-11-...	[none]	116820	1503	139	1066	https://i.ytimg.com/vi/pmJQ4...	false	false	false	https://pastebin.ca/3930472\...	ca
19	Autos & V...	PEv3SQS...	17.20.11	Shield Vs ...	Wrestling ...	2	2017-11-...	[none]	561583	4171	280	282	https://i.ytimg.com/vi/PEv3S...	false	false	false		ca
20	Autos & V...	1bbcPdz...	17.30.11	Subaru As...	Motormo...	2	2017-11-...	[none]	3059	95	2	22	https://i.ytimg.com/vi/1bbcPd...	false	false	false		ca

Now we join two table together, so we can see the title of each category_id, it is much more easier to do analytic report now.

6. Materialize the data using AWS Glue Studio

Since we can use SQL to join two tables for our final result. Now we want to optimize the process. We are going to use AWS Glue Spark job to join two tables and push the joined table to the final layer(S3 analytic bucket, and AWS Glue Catalog analytic database, table).
As a result, we can build our dashboard through the joined table, so we can save time to join tables together to get our final result.



As you can see above, we select two data data source, one is the table for raw statistic which is transform from csv file, another one is the table for reference which is transform from json file.

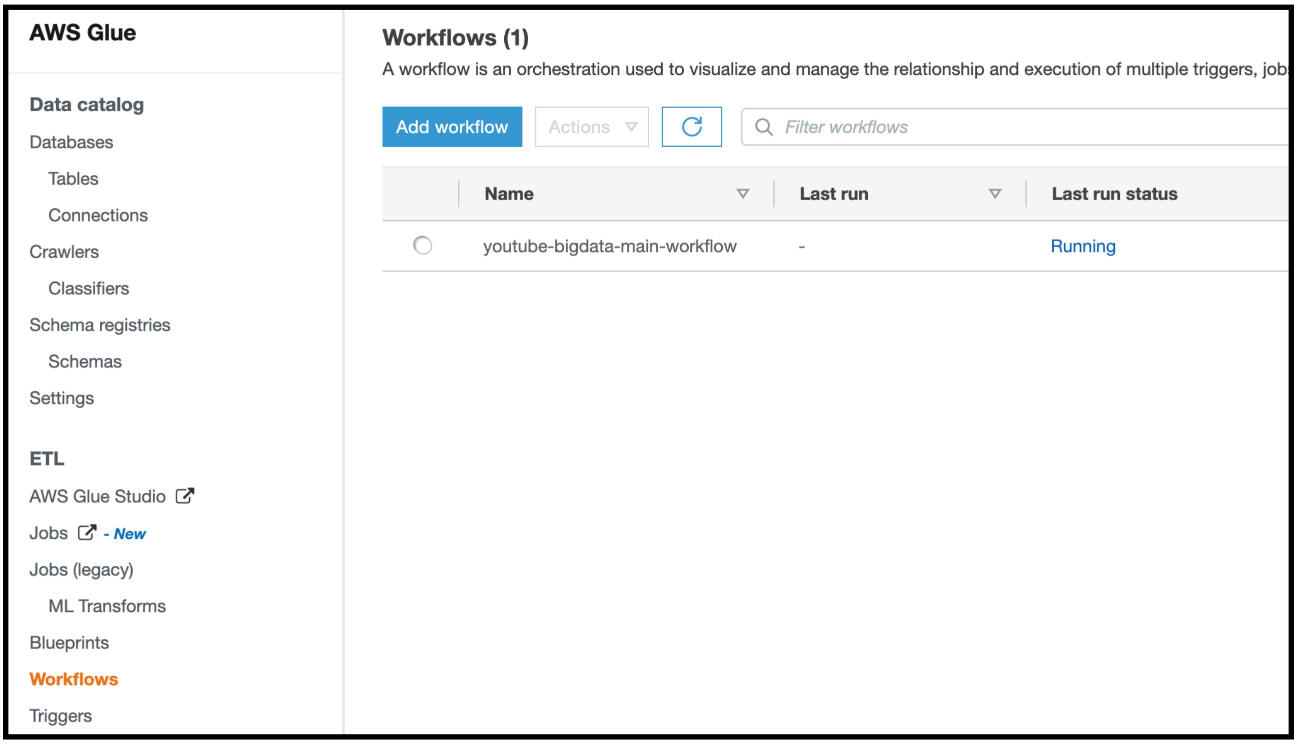
As Transform method, we select Join. Select the columns we are going to join which is the ***category_id***

Finally, we will output the result to S3 analytic bucket and AWS Glue Catalog analytic database, table.

7. Setup Automate Glue Job

The Glue jobs we created before is only run on demand. In order to build a robust data pipeline that can ingest data continuously we need to automate the entire ETL workflow.

In AWS Glue Service, we can add workflows and schedule them once a while base on your need.



Workflows (1)

A workflow is an orchestration used to visualize and manage the relationship and execution of multiple triggers, job

	Name	Last run	Last run status
○	youtube-bigdata-main-workflow	-	Running

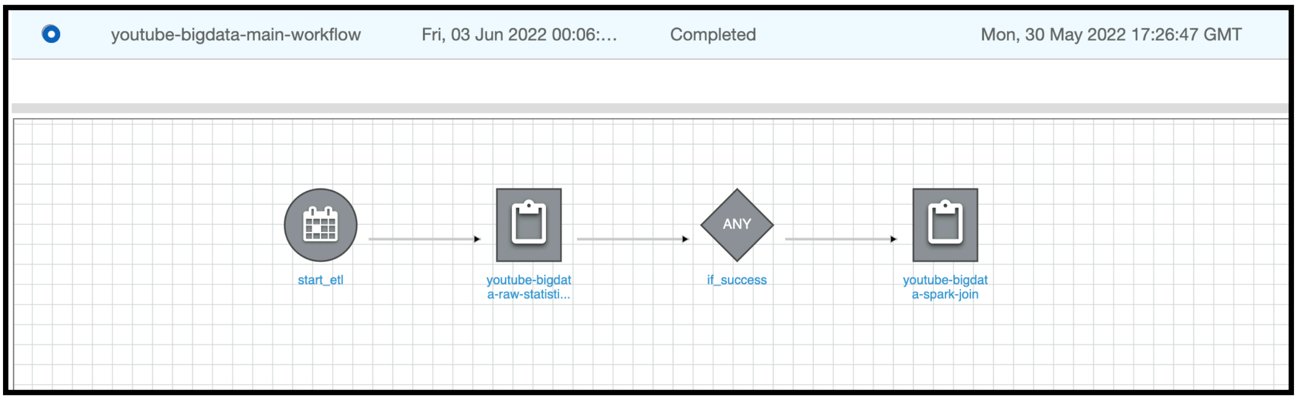
First, we add trigger and select trigger type as schedule. Now this workflow will start every hour or everyday base on your choice.

Second, add a node after the trigger, select the glue job we want to do first which is the one that transform csv file to parquet.

Third, add another trigger after the first glue job which monitor if the previous glue job is success.

Finally, add the glue job that join to table together in the next node.

Now, we have complete a simple workflow with AWS Glue that will process the data base on the schedule.



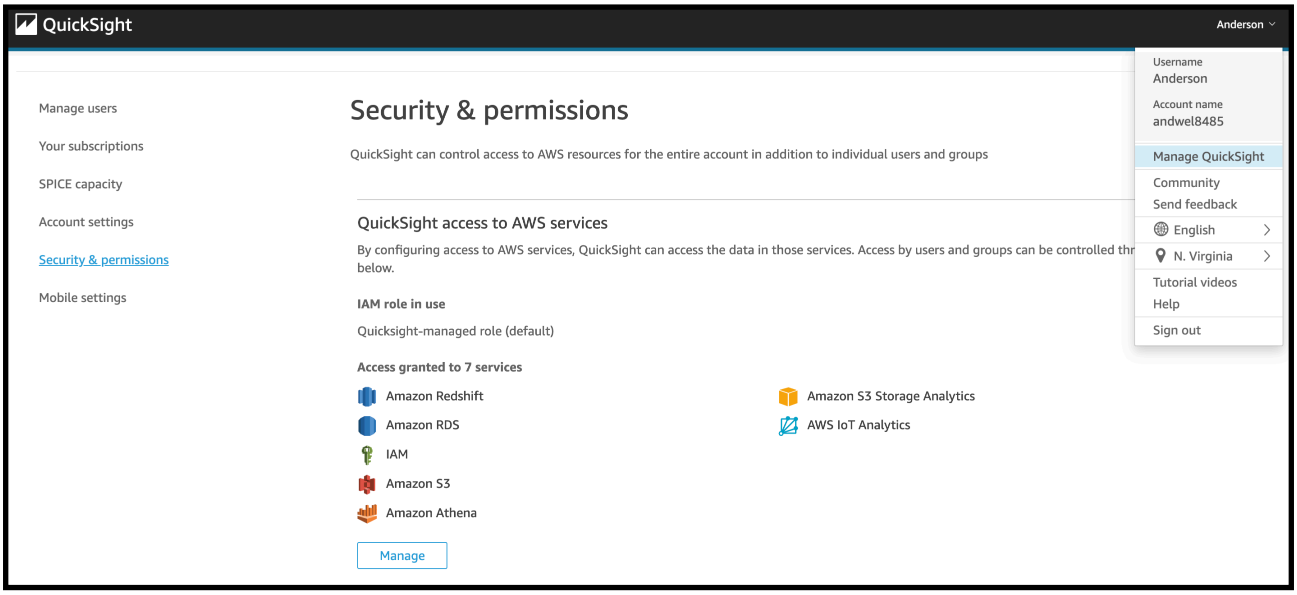
●	youtube-bigdata-main-workflow	Fri, 03 Jun 2022 00:06:…	Completed	Mon, 30 May 2022 17:26:47 GMT

8. Build Dashboard with QuickSight

After you create and log in to your QuickSight Account, we need to grand permission to QuickSight so it can access the data source.

Select Manage QuickSight and go to the security& permissions tab.

Make sure that you grant QuickSight permission to access s3 (it's better specify the bucket that you are going to grant the permission) and Athena.



Security & permissions

QuickSight can control access to AWS resources for the entire account in addition to individual users and groups

QuickSight access to AWS services

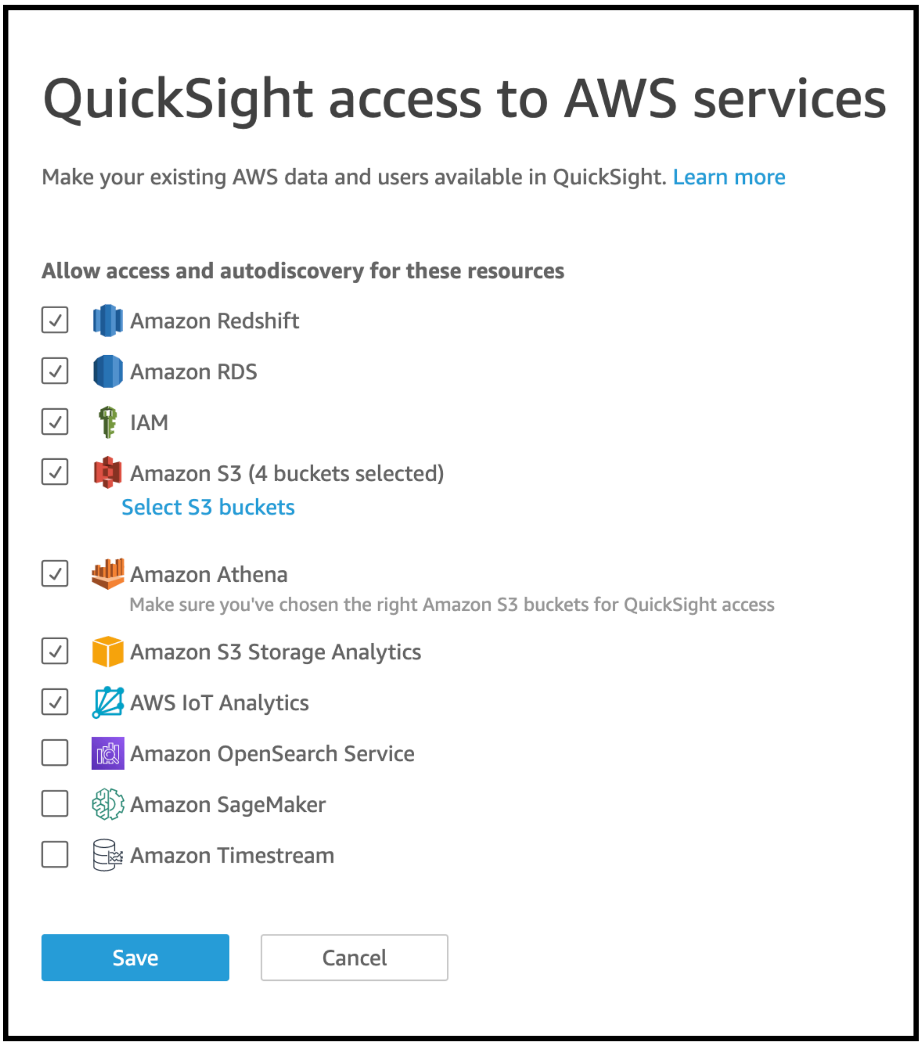
By configuring access to AWS services, QuickSight can access the data in those services. Access by users and groups can be controlled thr below.

IAM role in use

Quicksight-managed role (default)

Access granted to 7 services

- Amazon Redshift
- Amazon RDS
- IAM
- Amazon S3
- Amazon Athena
- Amazon S3 Storage Analytics
- AWS IoT Analytics



QuickSight access to AWS services

Make your existing AWS data and users available in QuickSight. Learn more

Allow access and autodiscovery for these resources

☑ Amazon Redshift

☑ Amazon RDS

☑ IAM

☑ Amazon S3 (4 buckets selected)
Select S3 buckets

☑ Amazon Athena
Make sure you've chosen the right Amazon S3 buckets for QuickSight access

☑ Amazon S3 Storage Analytics

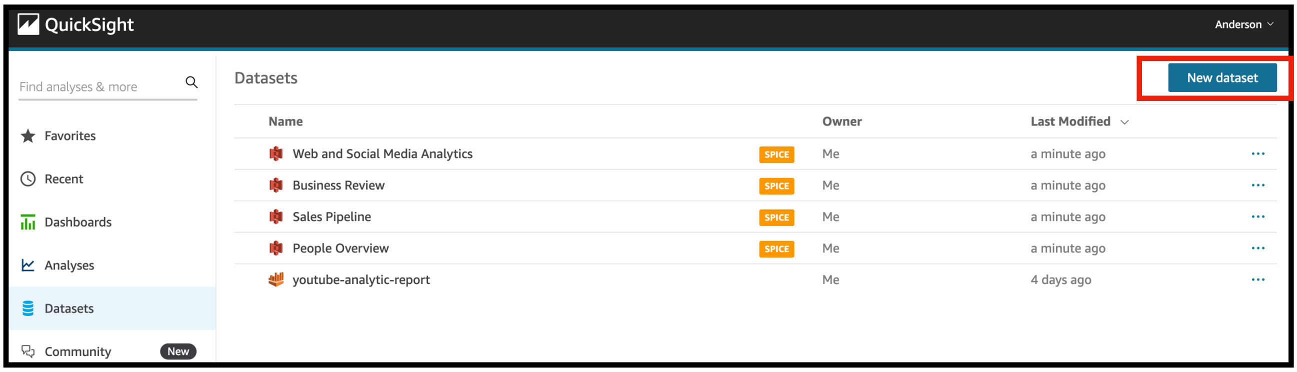
☑ AWS IoT Analytics

□ Amazon OpenSearch Service

□ Amazon SageMaker

□ Amazon Timestream

After the permission settings, we can select add new dataset from Dataset tab.

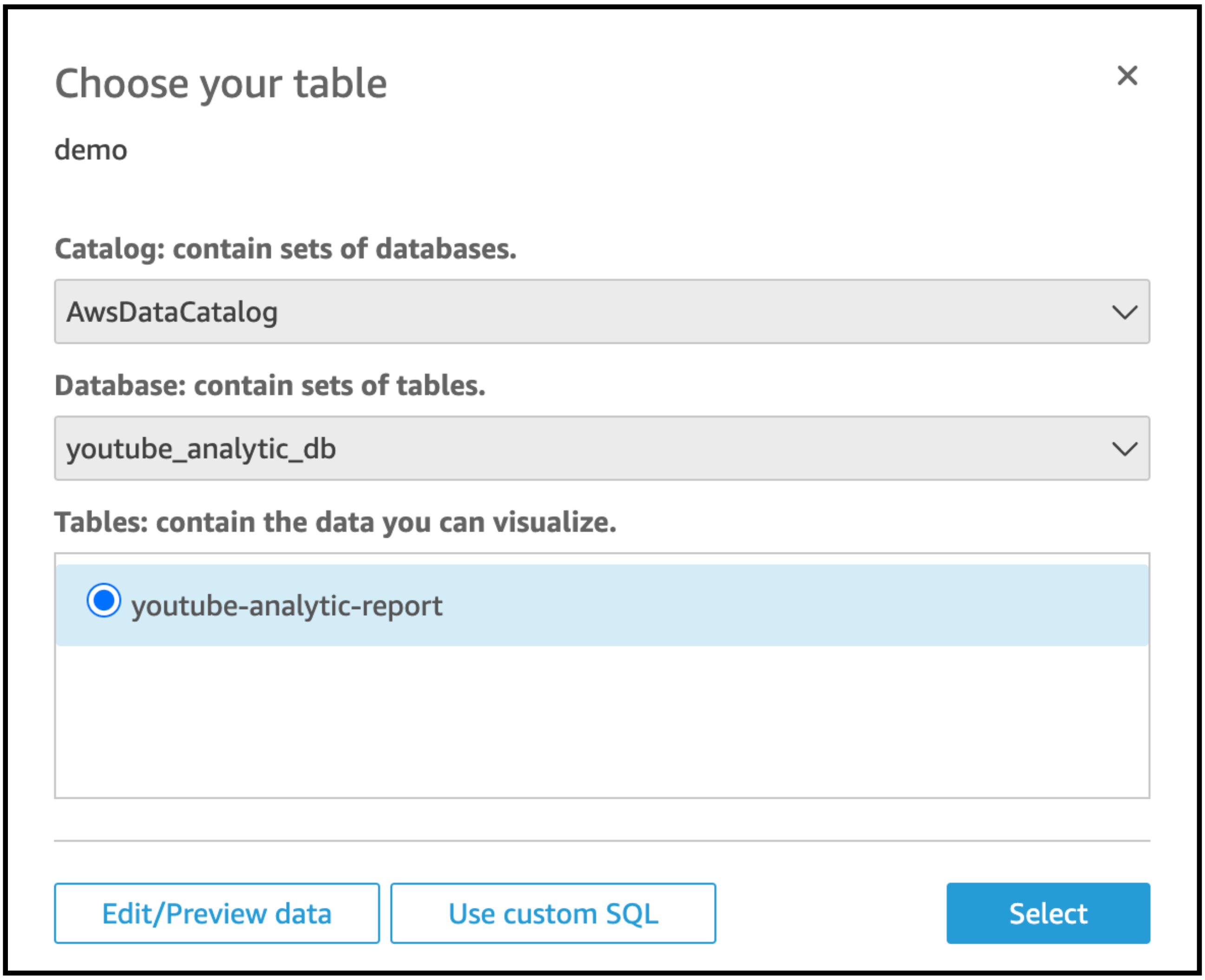


Datasets

New dataset

Name		Owner	Last Modified
Web and Social Media Analytics	SPICE	Me	a minute ago
Business Review	SPICE	Me	a minute ago
Sales Pipeline	SPICE	Me	a minute ago
People Overview	SPICE	Me	a minute ago
youtube-analytic-report		Me	4 days ago

We are going to choose Athena as our data source and after you name your data source, we select the analytic report table in analytic database as your visualize.



Choose your table

demo

Catalog: contain sets of databases.

AwsDataCatalog

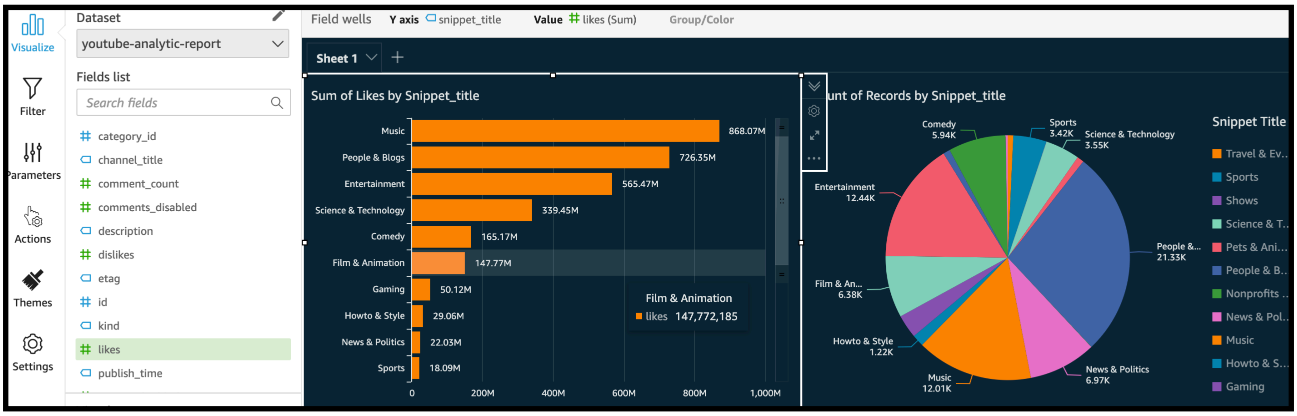
Database: contain sets of tables.

youtube_analytic_db

Tables: contain the data you can visualize.

● youtube-analytic-report

Once you setup your data source you can start to design your dashboard.



Dataset

youtube-analytic-report

Field wells Y axis snippet_title Value likes (Sum) Group/Color

Reference: ProjectPro