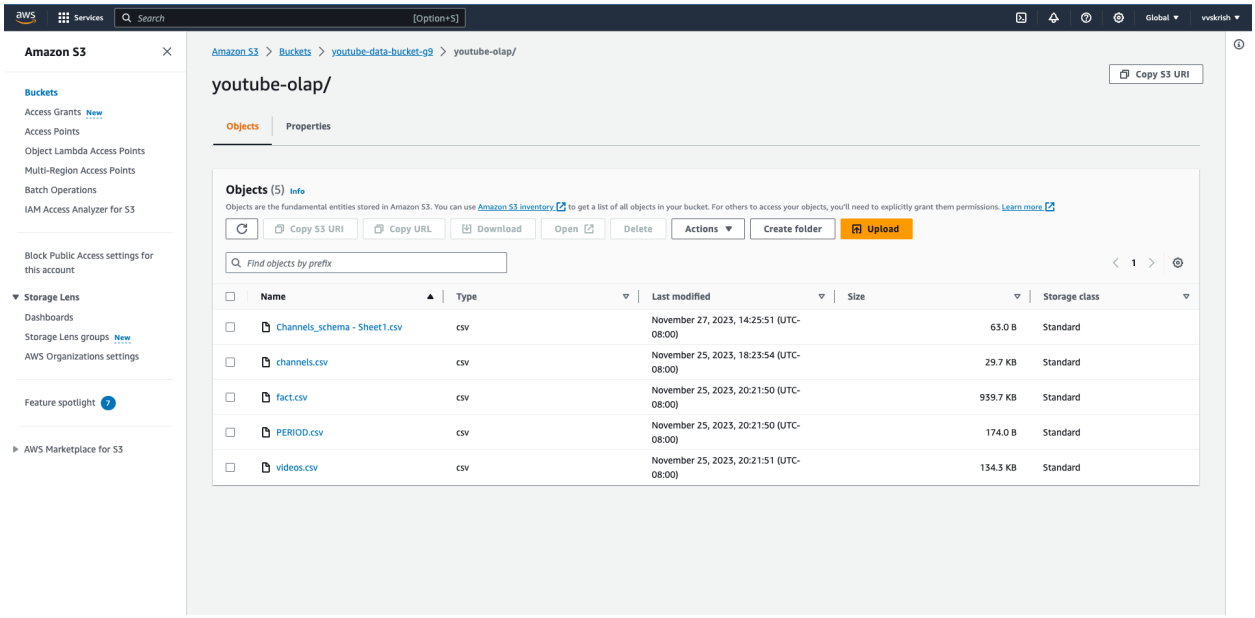
ETL

Amazon S3 Bucket: Link - https://youtube-data-bucket-g9.s3.amazonaws.com/youtube-olap/
An S3 bucket in Amazon S3 is a storage container used to hold various types of data, including CSV files. These CSV files are raw data that typically undergo Extract, Transform, and Load (ETL) operations within AWS Glue or similar services.

During ETL, the CSV files are extracted from the S3 bucket, transformed into a structured format suitable for analysis or other purposes, and then loaded back into another location, often another S3 bucket or a database, ready for use. This process helps in organizing, cleaning, and preparing the data for analytics, reporting, or further processing.



IAM Roles:
In Amazon Web Services (AWS), Identity and Access Management (IAM) roles serve as fundamental tools for securely delegating permissions within AWS resources. These roles define a set of permissions that dictate actions authorized for entities, be it an AWS service or a user, on AWS resources. They eliminate the necessity for long-term credentials like usernames and passwords when granting access to resources.

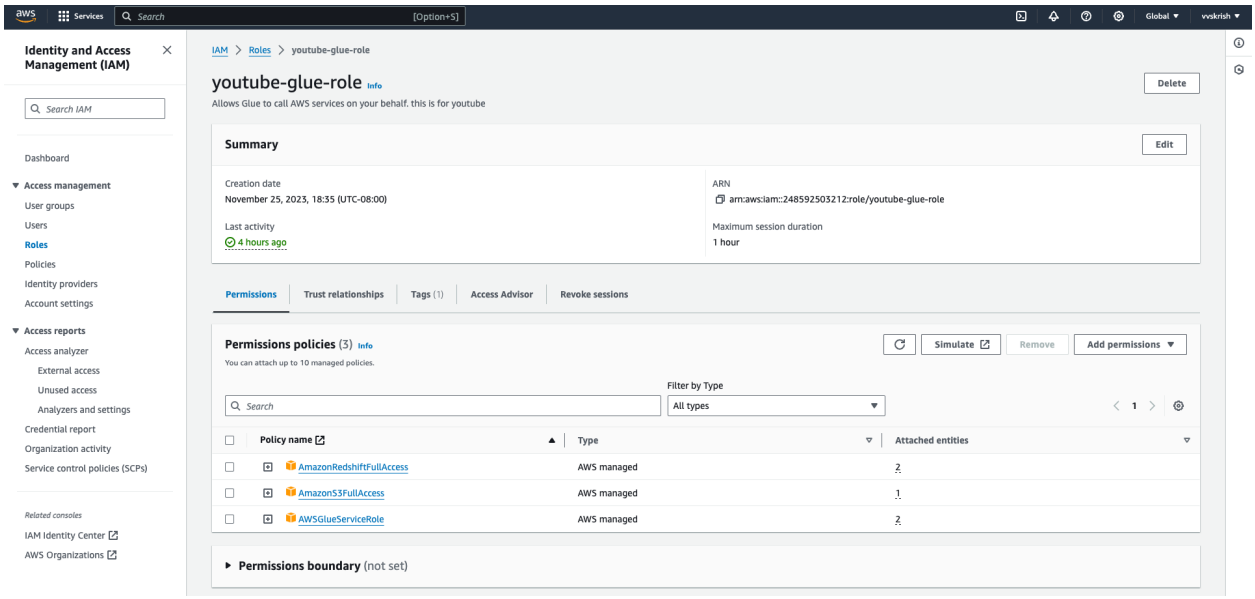
The IAM role in the image below is configured with specific permissions:

- AmazonS3FullAccess: Grants complete access to Amazon S3 (Simple Storage Service), enabling any action on S3 buckets within the AWS account.
- AmazonRedshiftFullAccess: Provides full access to Amazon Redshift, empowering the IAM role to manage Redshift clusters comprehensively. This includes tasks such as creating, modifying, deleting clusters, and handling administrative functions.

- AWSGlueServiceRole: Grants the essential permissions for executing actions within AWS Glue, a fully managed ETL service facilitating data preparation and loading for analytics.

By consolidating these permissions within a single IAM role, entities assuming this role gain the combined access and capabilities across Amazon S3, Amazon Redshift, and AWS Glue. This practice aligns with security best practices, adhering to the principle of least privilege by granting only the necessary permissions for specific tasks.

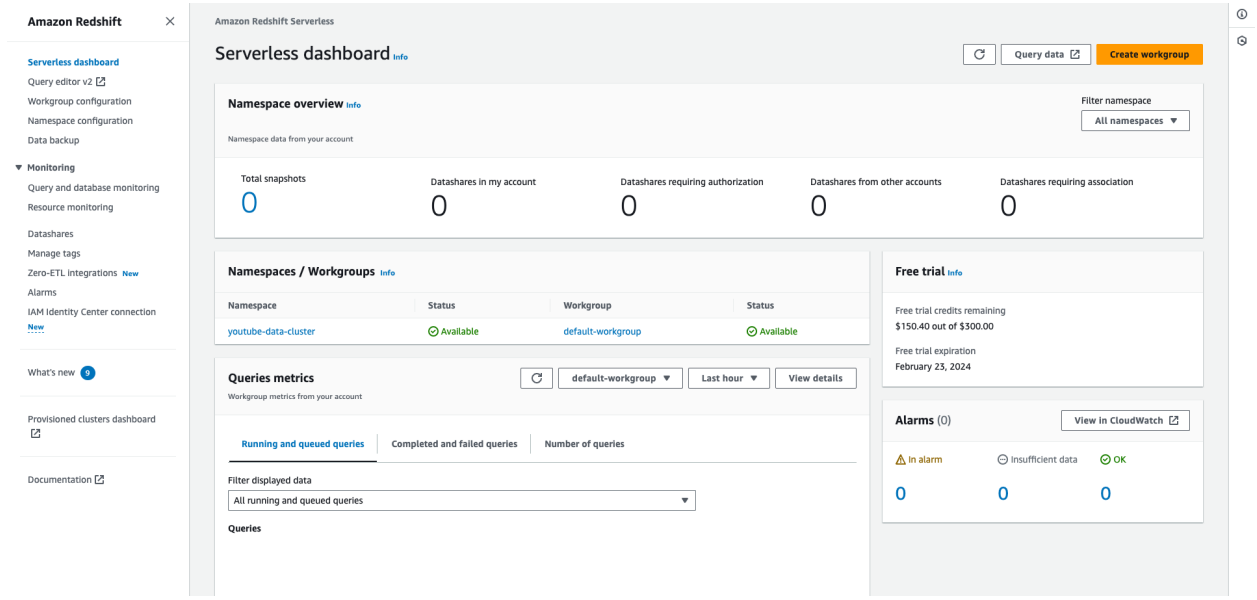
IAM roles offer remarkable flexibility, allowing different entities such as AWS services, applications, or other AWS accounts to assume these roles based on predefined trust policies. Assigning roles to entities enables centralized permission management, upholding security measures, and reducing reliance on persistent credentials, thereby enhancing the overall security posture of AWS resources.



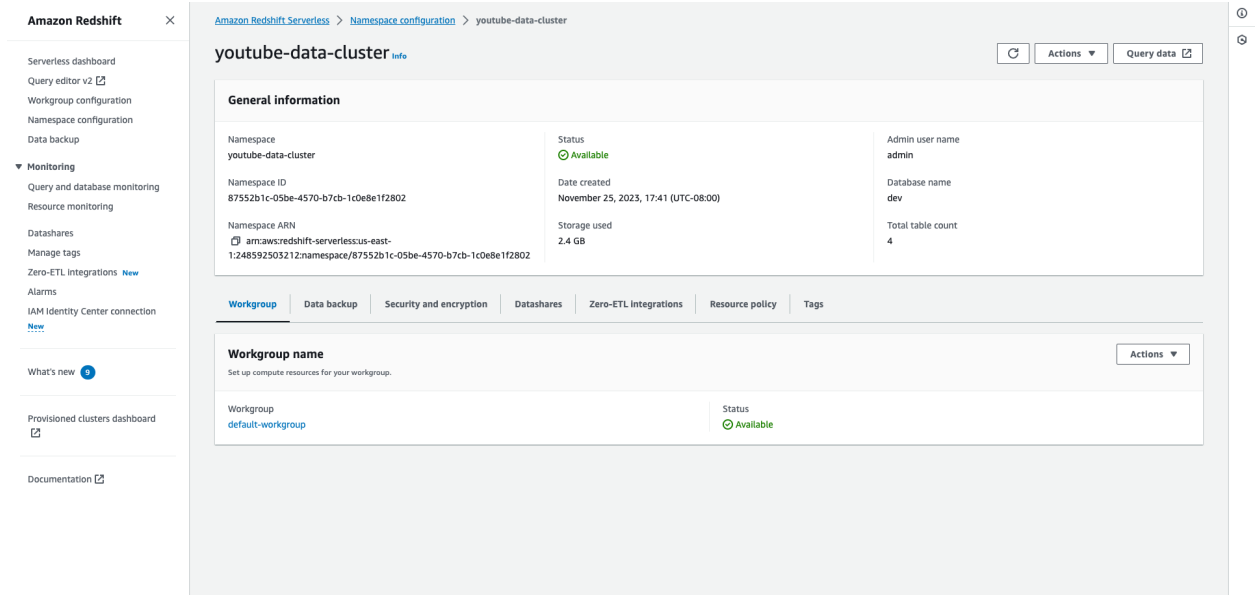
Red Shift Cluster:

Amazon Redshift is a data warehouse service in AWS used for analyzing large datasets. A Redshift cluster is a collection of nodes that work together to handle queries and data storage. After the ETL process, the transformed data from CSV files is often loaded into a Redshift cluster, typically organized into tables.

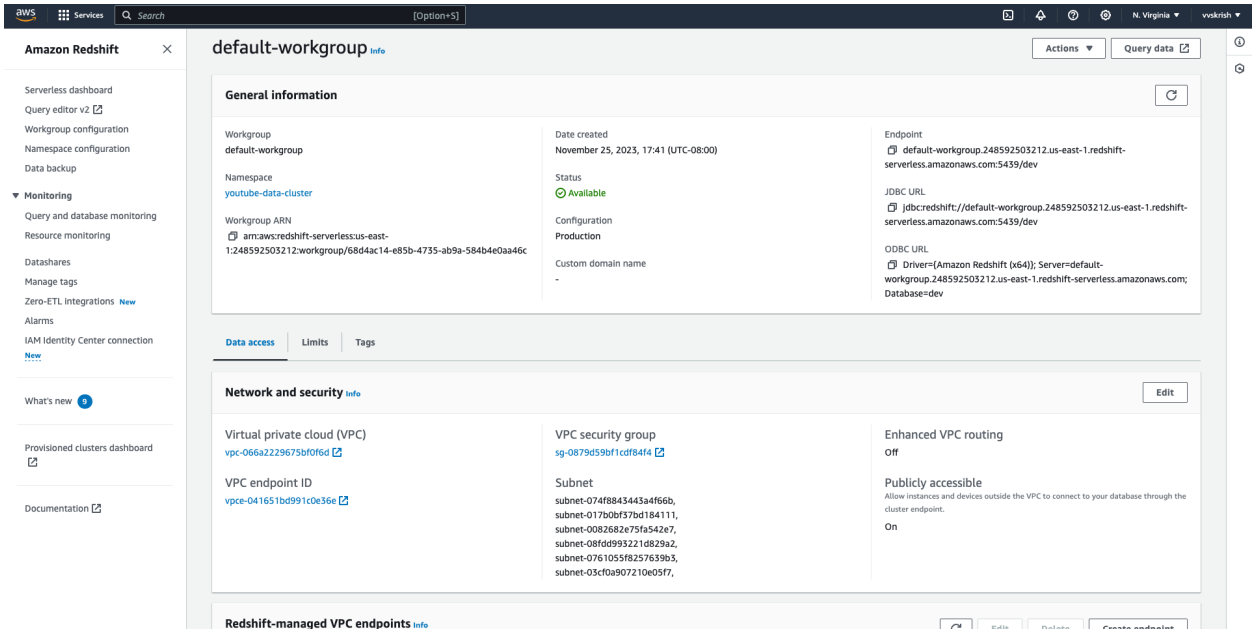
These tables store structured data and are designed to support efficient querying and analysis. In your case, there are four tables within the Redshift cluster, each containing specific datasets that have undergone transformation. These tables serve as organized repositories of data, allowing users to run complex queries and analytics to derive insights.



Details of Redshift cluster:



Connection details of Redshift Cluster:



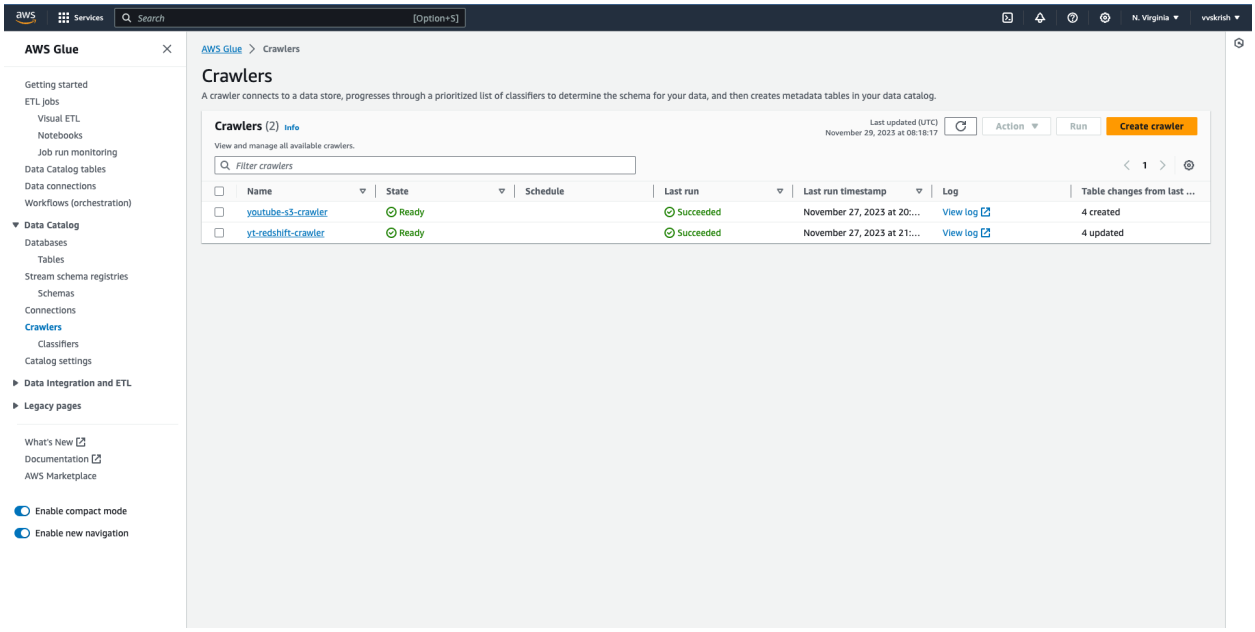
Crawlers:

In AWS Glue, crawlers serve as automated processes used for discovering and cataloging metadata from diverse data sources. These processes analyze data structure and schema, facilitating streamlined data processing within the ETL (Extract, Transform, Load) workflow.

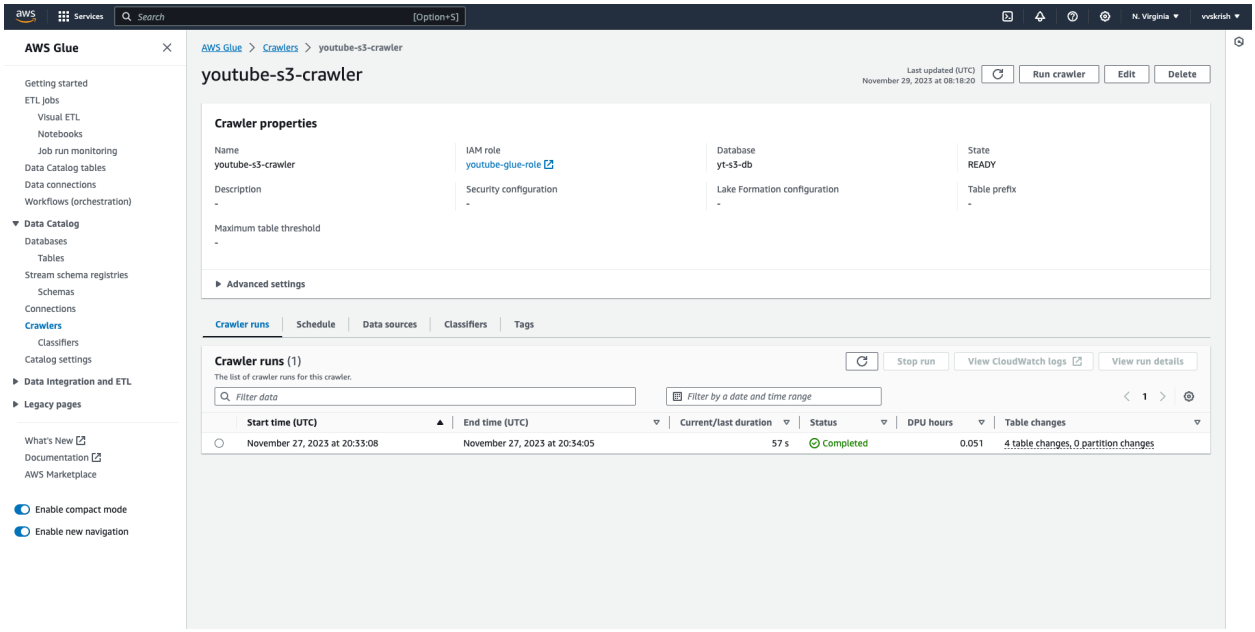
The S3 crawler within AWS Glue inspects data stored in Amazon S3 buckets. It identifies file formats and schema, organizing the data for subsequent use in the ETL process. By scanning S3 data, this crawler comprehends its structure, aiding in the preparation of data for transformation.

Similarly, the Redshift crawler in AWS Glue targets Redshift clusters. It assesses the data within Redshift tables, capturing their structure and schema. This acquired information becomes crucial for mapping and efficiently transforming data during ETL operations. It ensures consistency and accuracy in data processing within the Redshift environment.

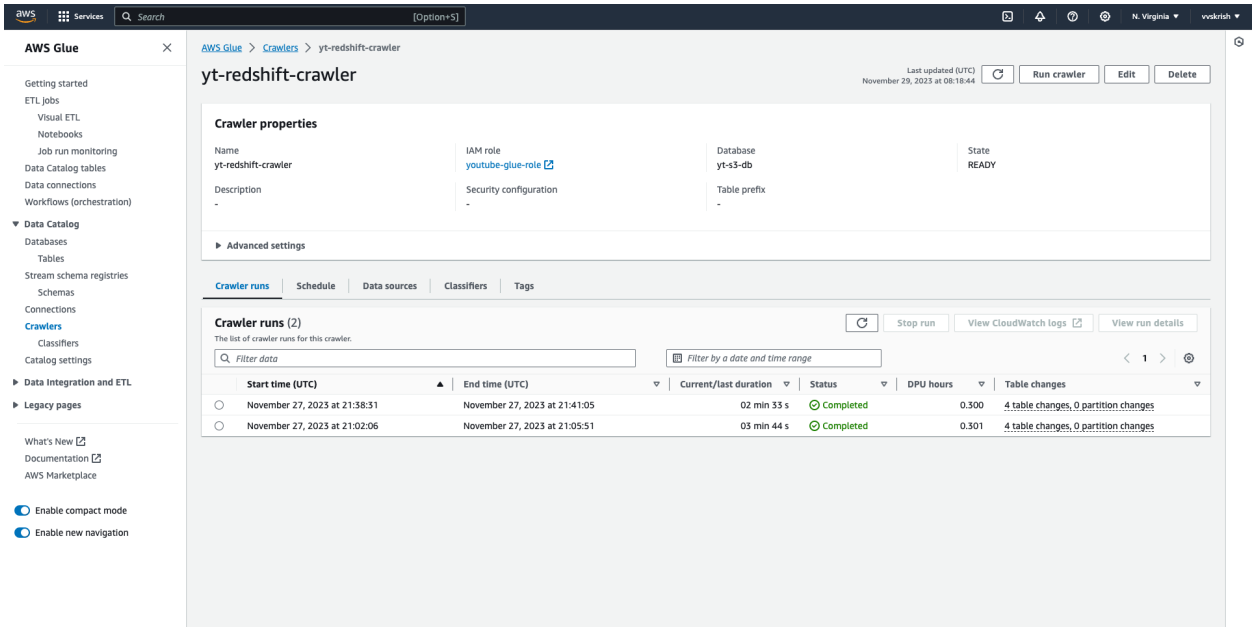
Both crawlers assume a pivotal role in the ETL process by automatically discovering and organizing metadata. Their function streamlines the transformation and loading of data from its raw state into structured datasets, primed for analysis or other intended uses.



Youtube S3 Crawler:

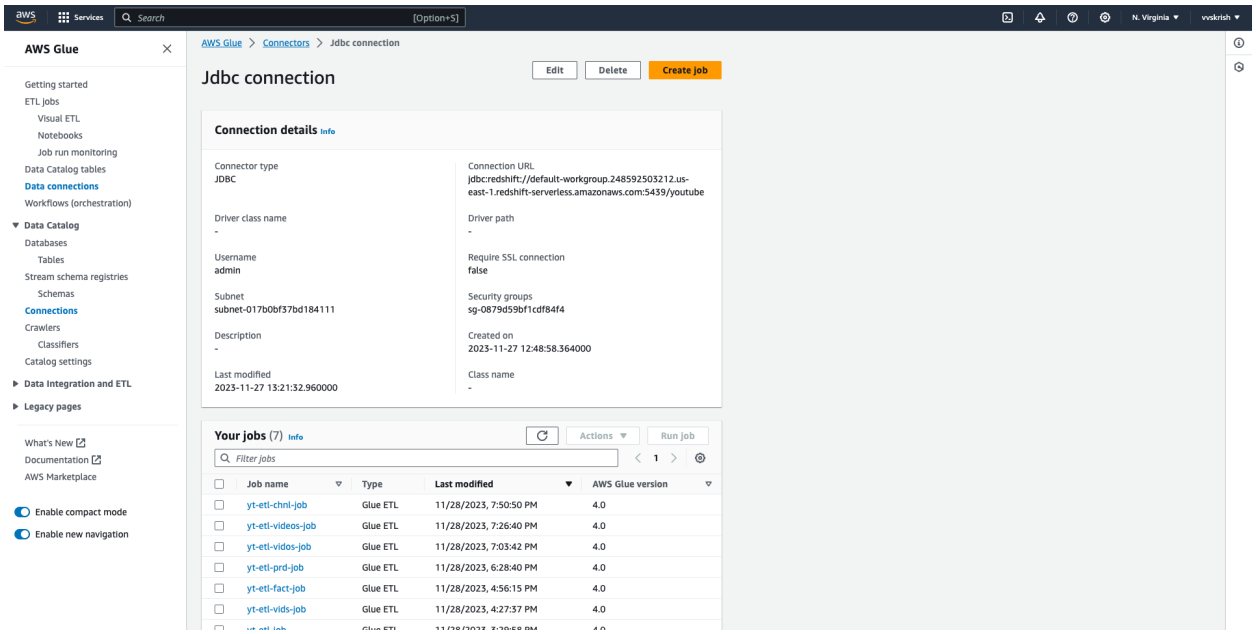


Redshift Crawler:

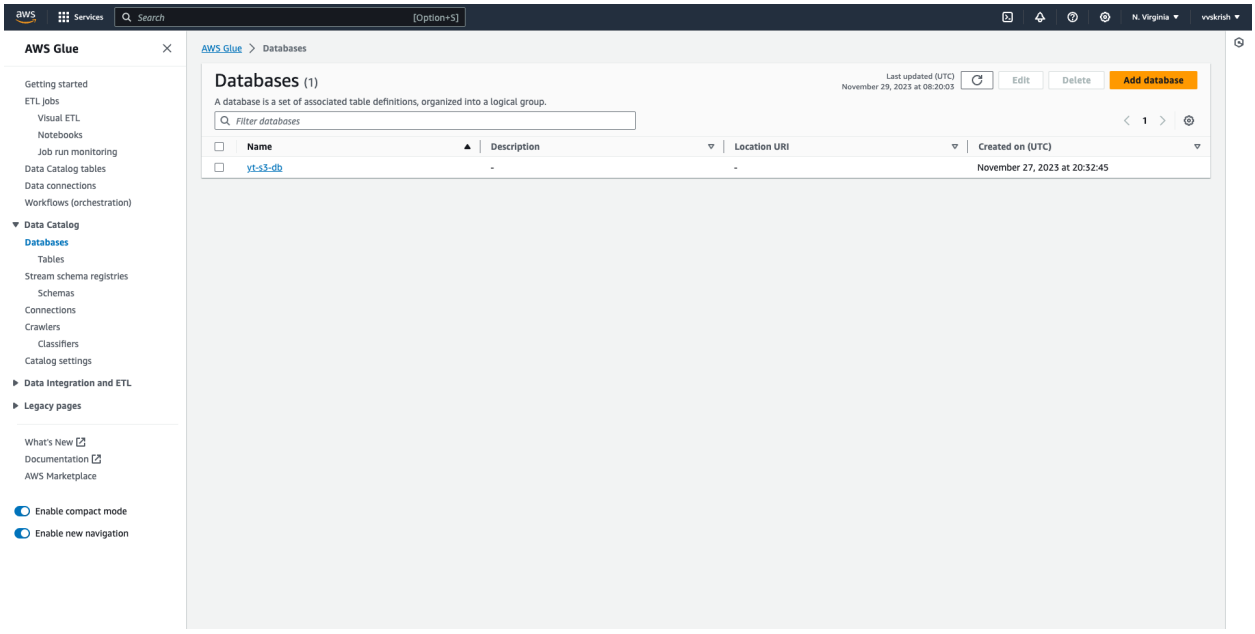


JDBC Connection to Temporary Database:

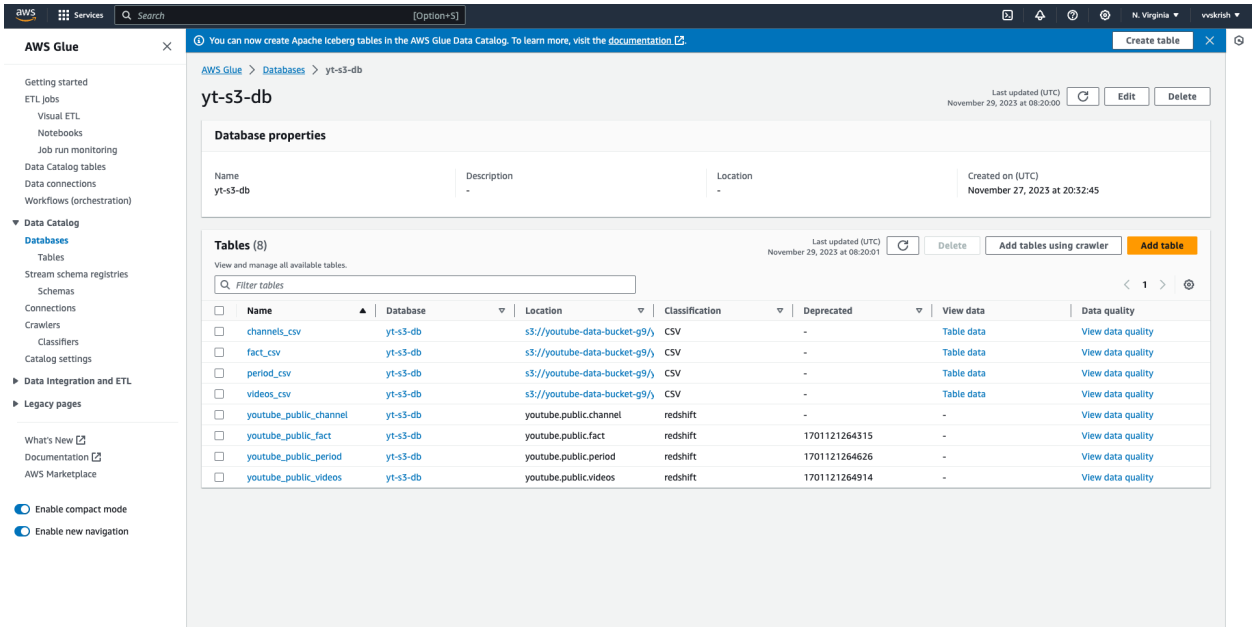
In AWS Glue, a JDBC connection establishes access to a temporary database, essential for the ETL process. JDBC, as a Java Database Connectivity tool, enables data interaction and transfer between the Glue environment and external databases. This connectivity supports AWS Glue in extracting, transforming, and loading data across diverse databases, ensuring smooth data processing throughout the ETL operations.



Temporary Database:



Contents of Temporary Database:



ETL Jobs:
In AWS Glue, an ETL job is created for each of the four tables to manage data flow within the ETL process.

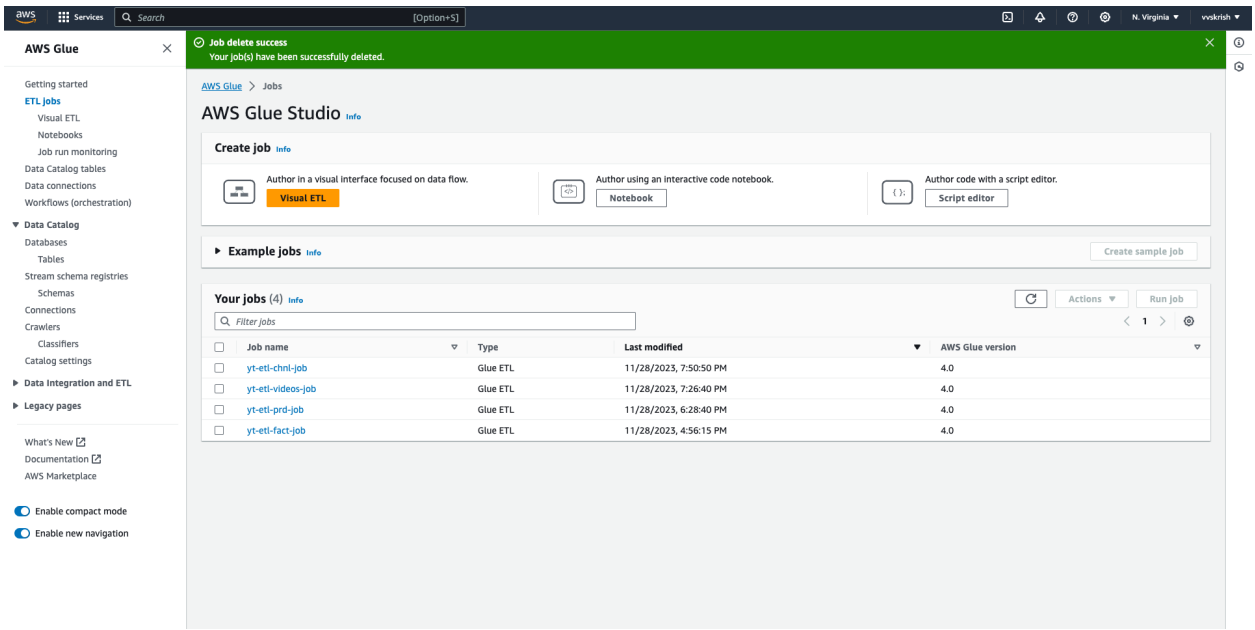
These jobs define Extract, Transform, and Load operations tailored for each table. Here's the breakdown:

1. Extraction (Extract): Data is extracted from respective sources like Amazon S3 buckets or Redshift databases.

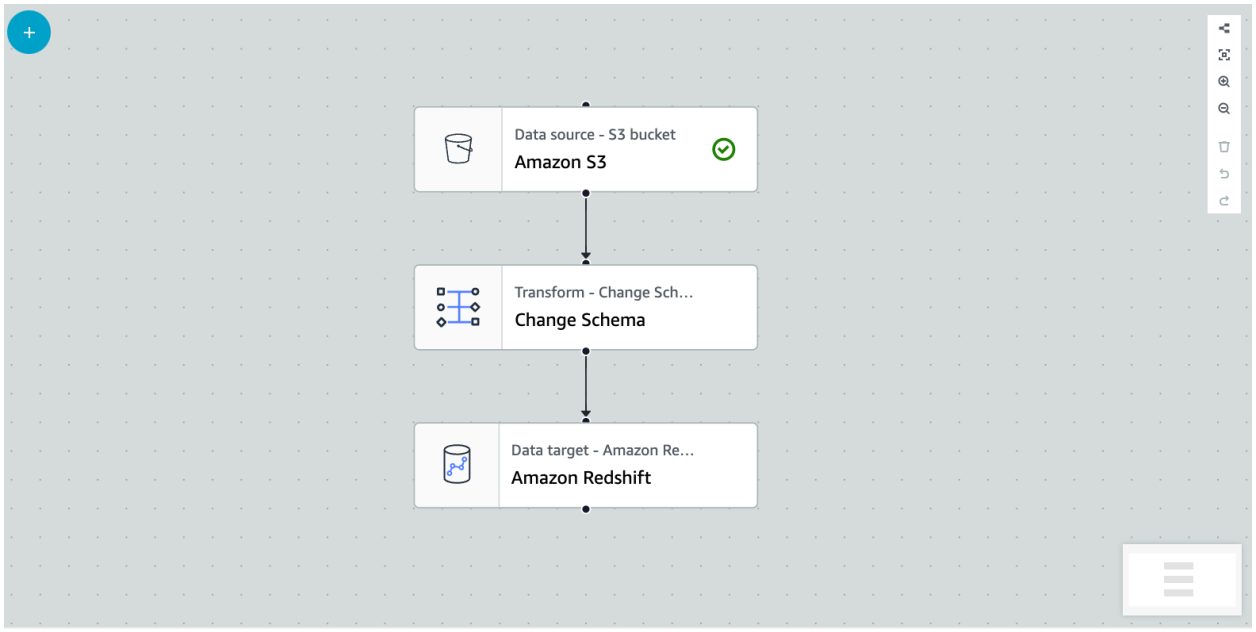
2. Transformation (Transform): Extracted data undergoes predefined logic-based transformations, such as cleaning, restructuring, aggregating, or enriching for analysis.

3. Loading (Load): Transformed data is loaded into the target destination, typically the respective Redshift table. This step ensures structured data storage, ready for efficient querying and analysis.

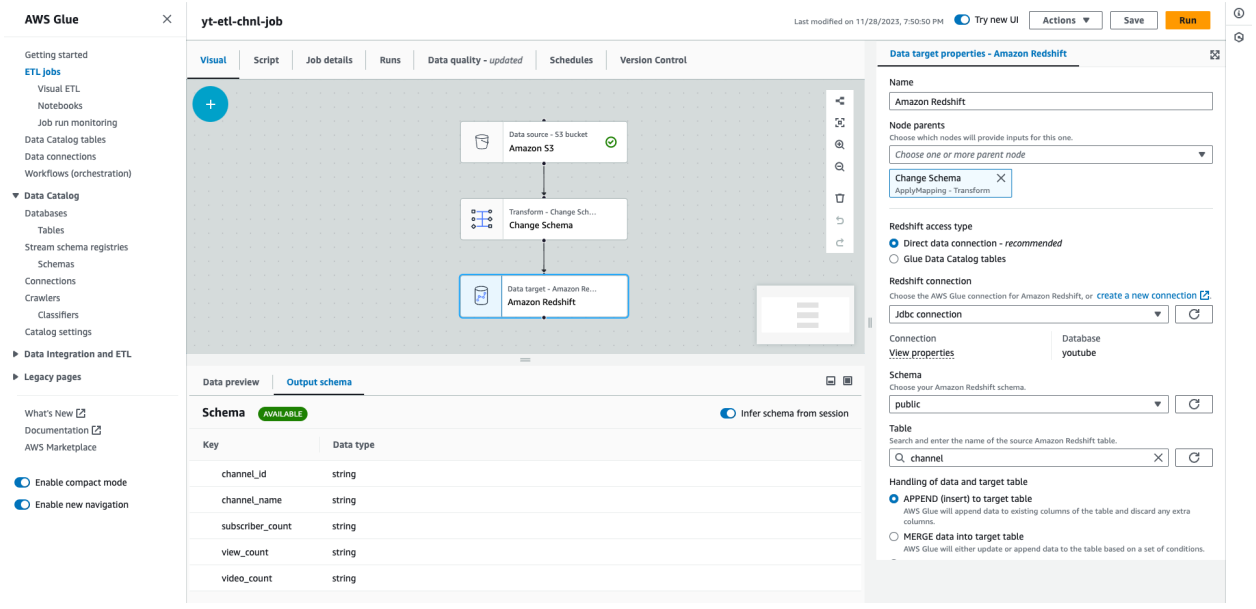
Each ETL job in AWS Glue addresses unique requirements and transformations specific to its table. Creating separate jobs for each table enables better organization, management, and optimization of the ETL process, aligning with the data characteristics within each table.



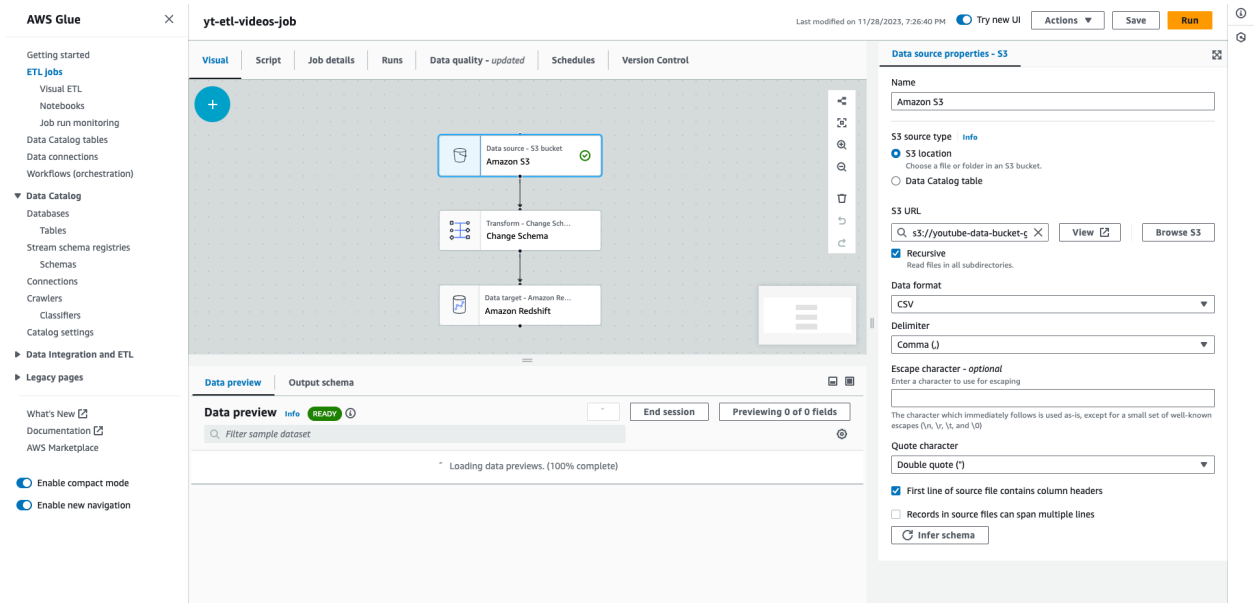
ETL Workflow:



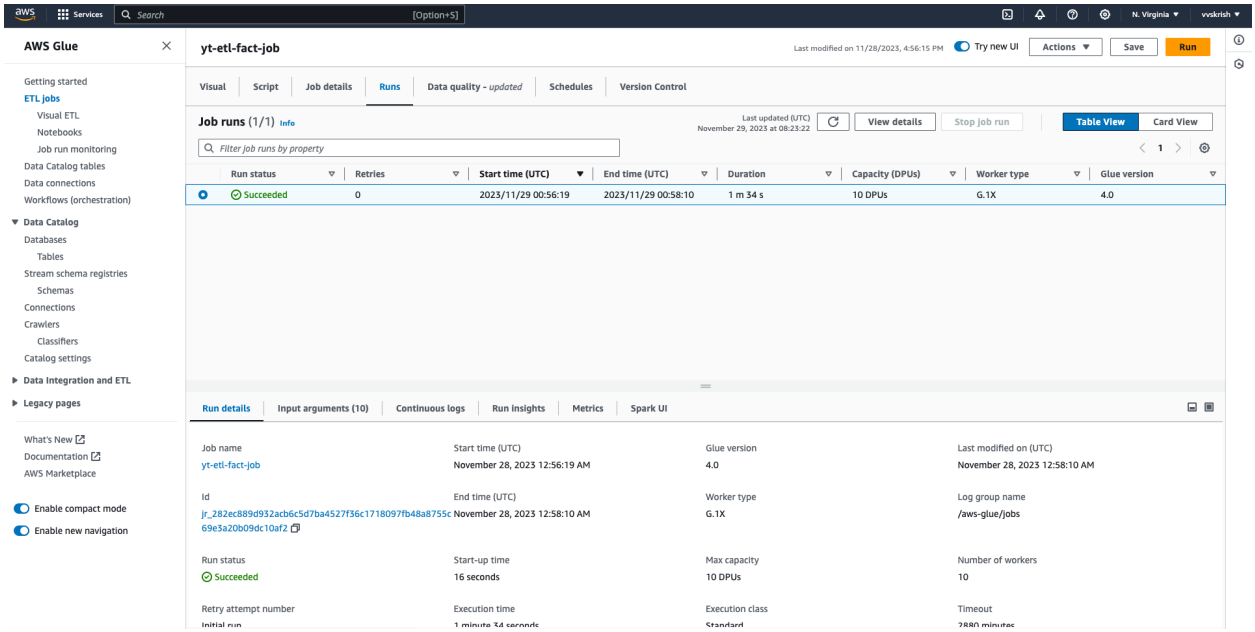
Channel ETL:



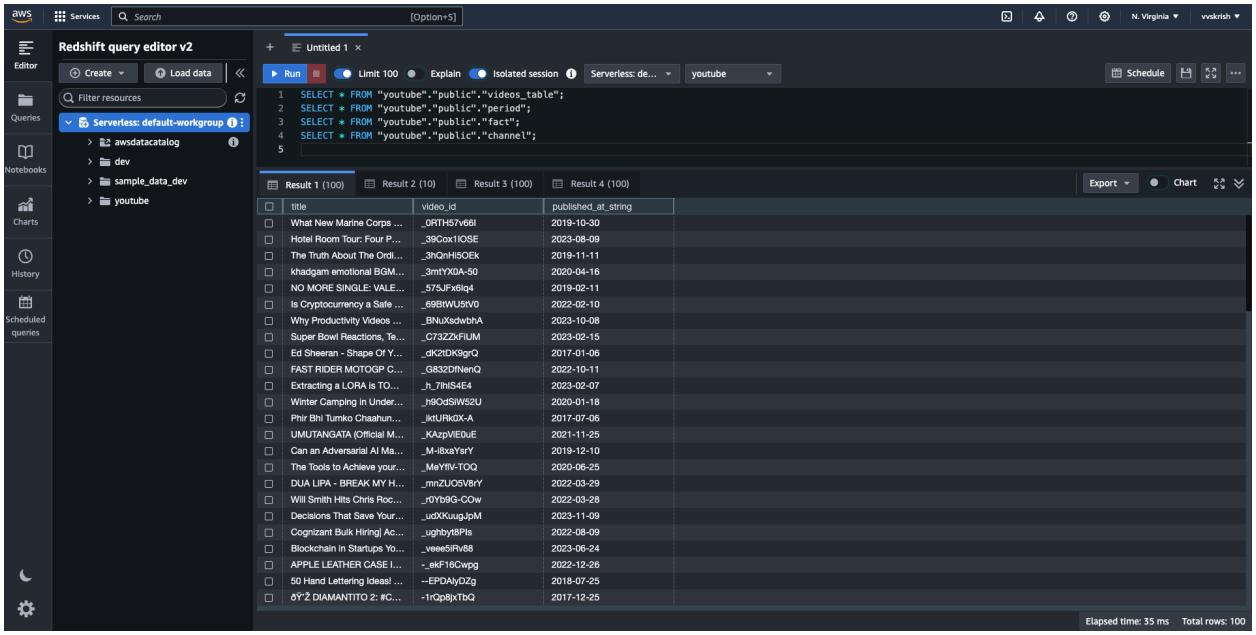
Videos ETL:



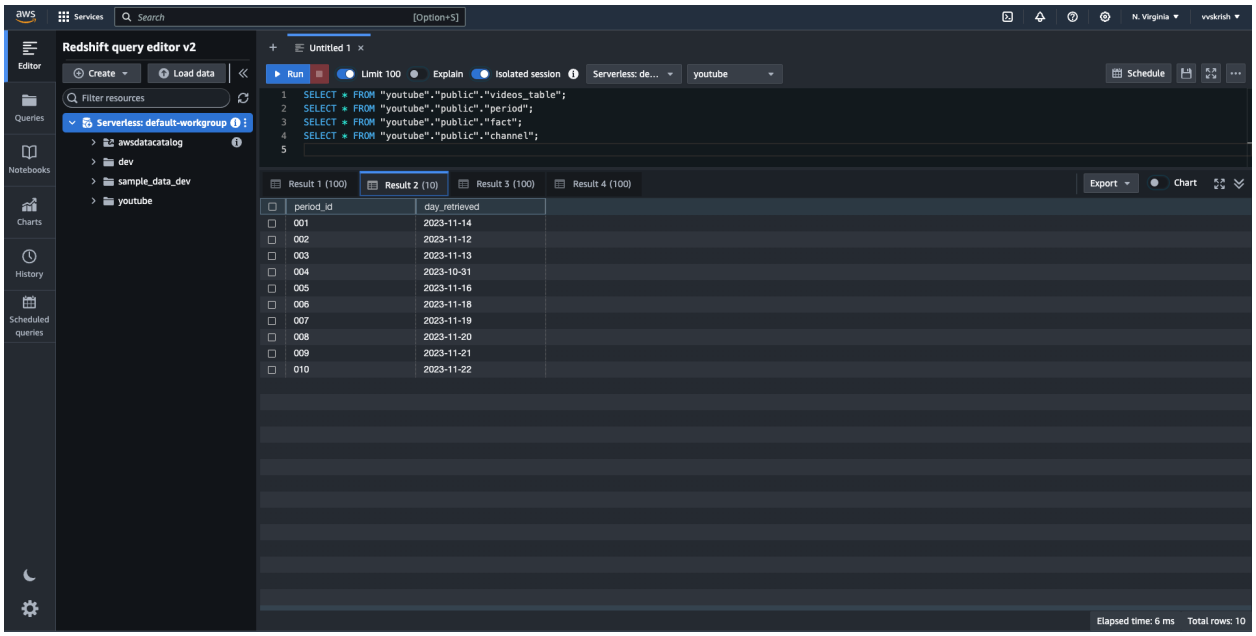
Execution of ETL Job:



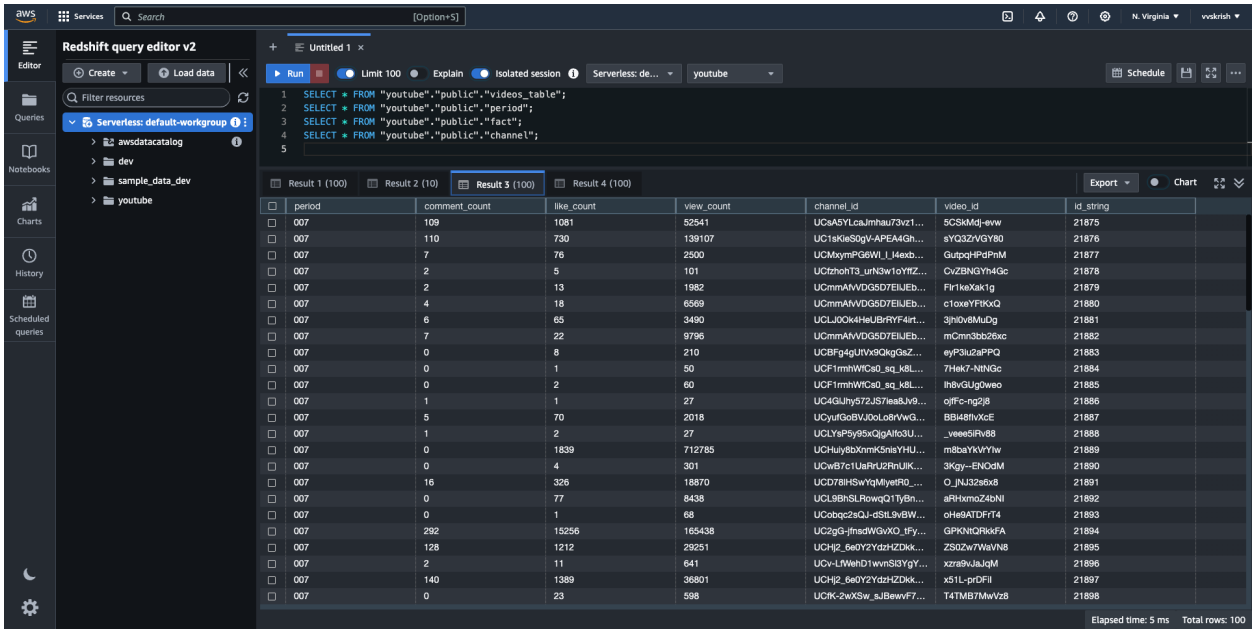
Videos data loaded into the Redshift cluster:



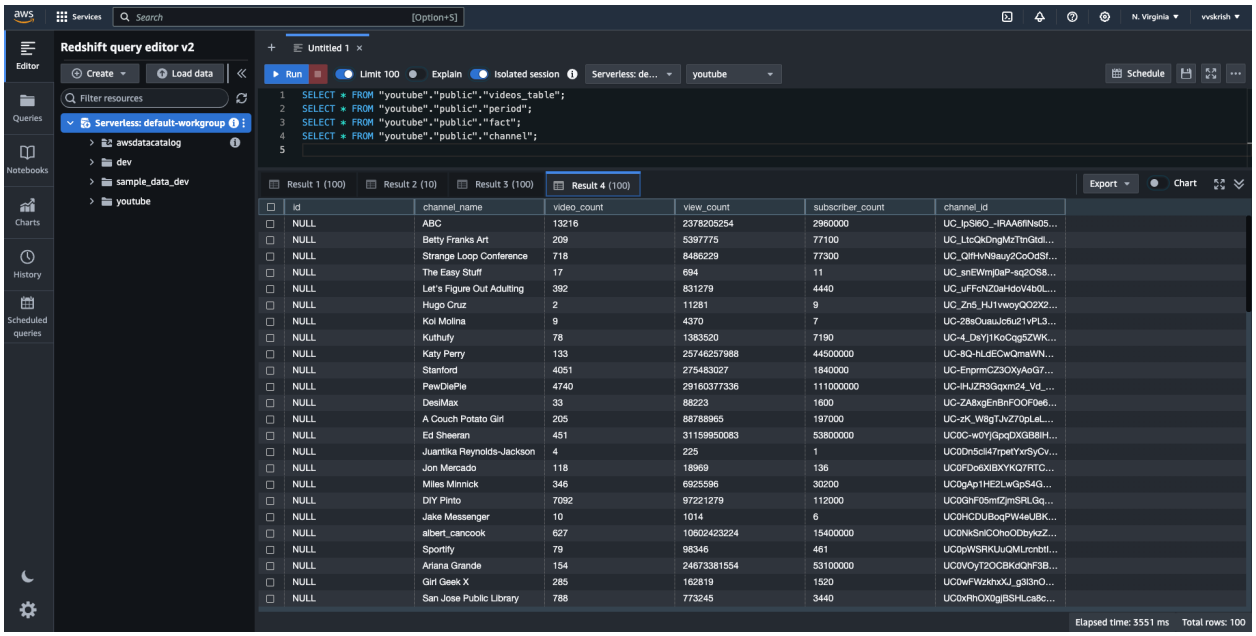
Period data loaded into the Redshift cluster:



Fact Table loaded into the Redshift cluster:

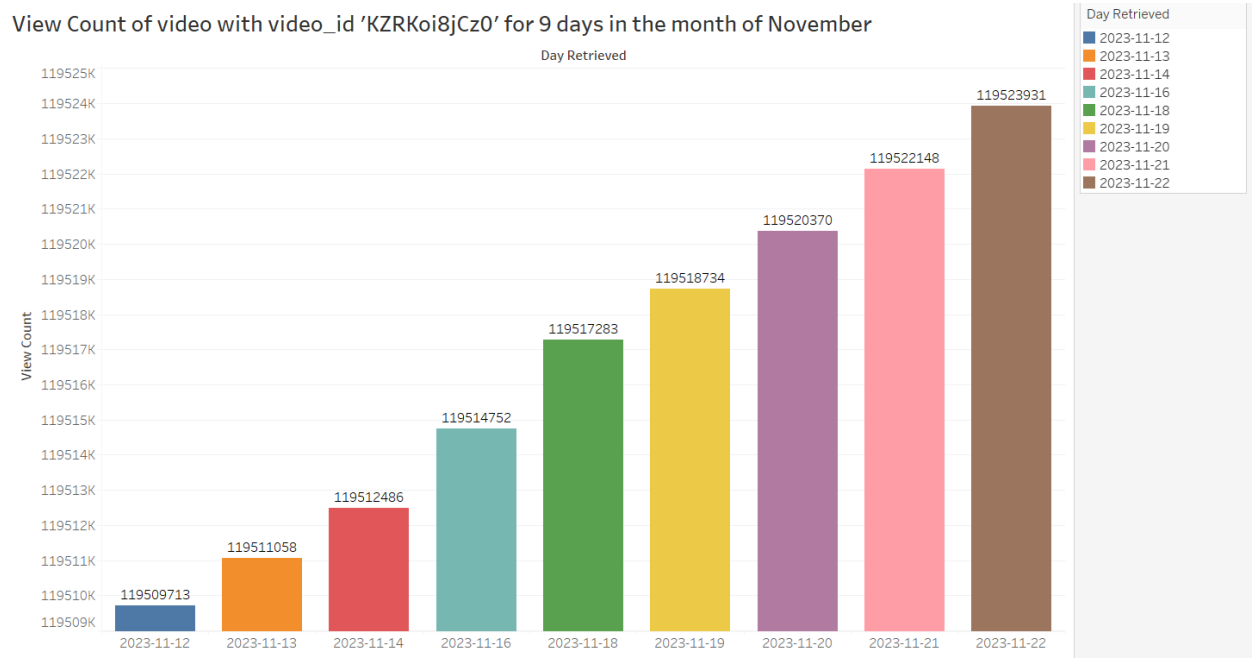


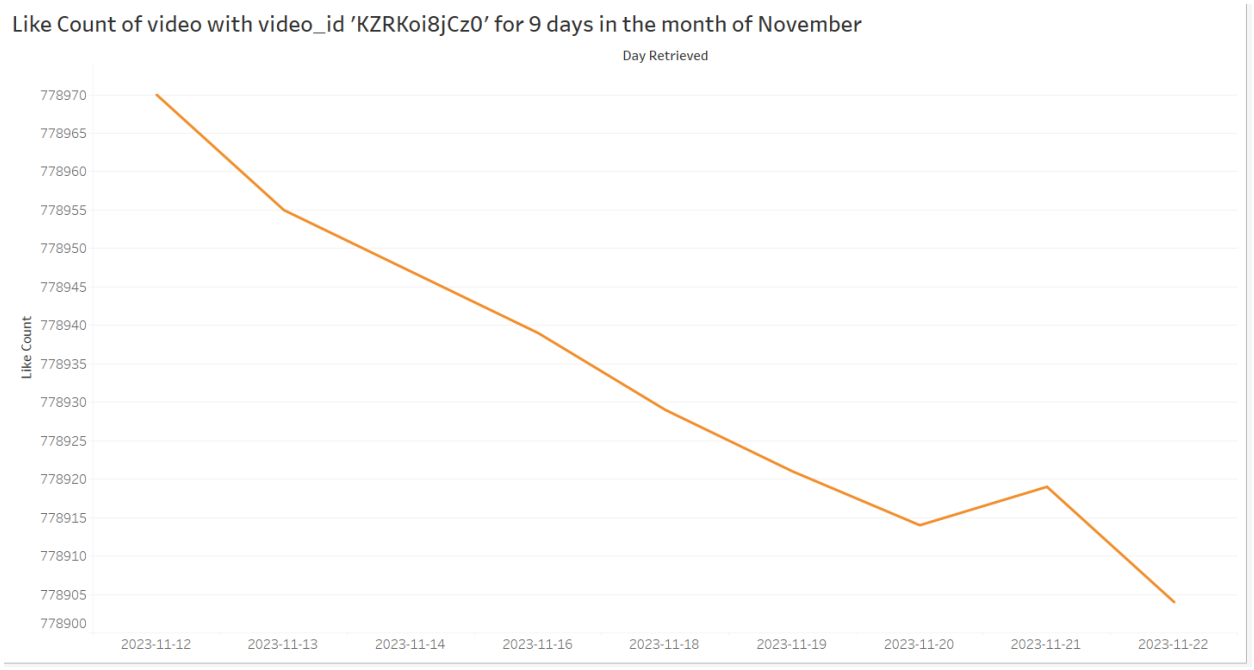
Channel Table loaded into the Redshift cluster:

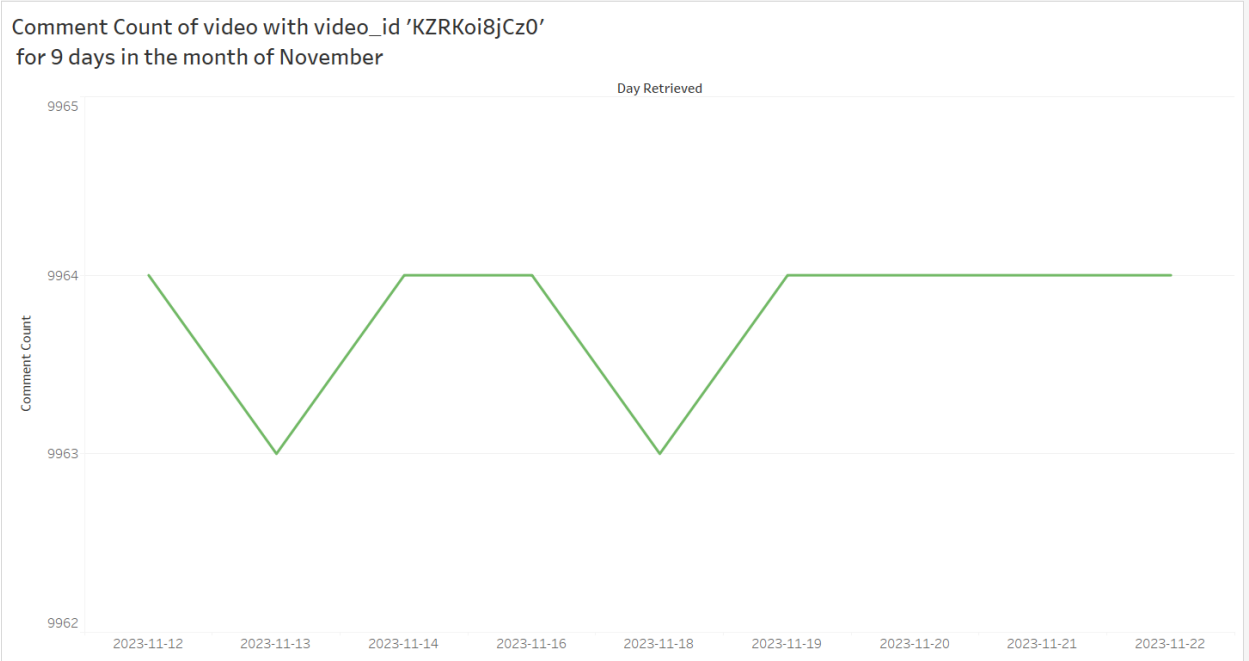


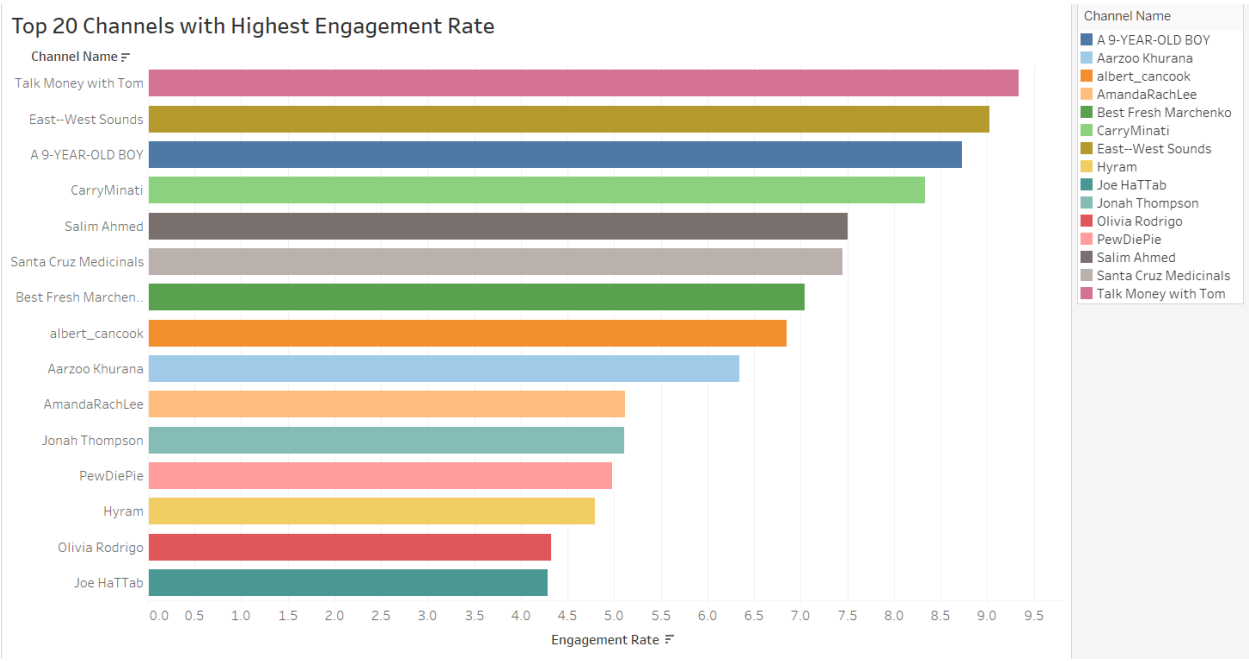
Visualizations (in Tableau):

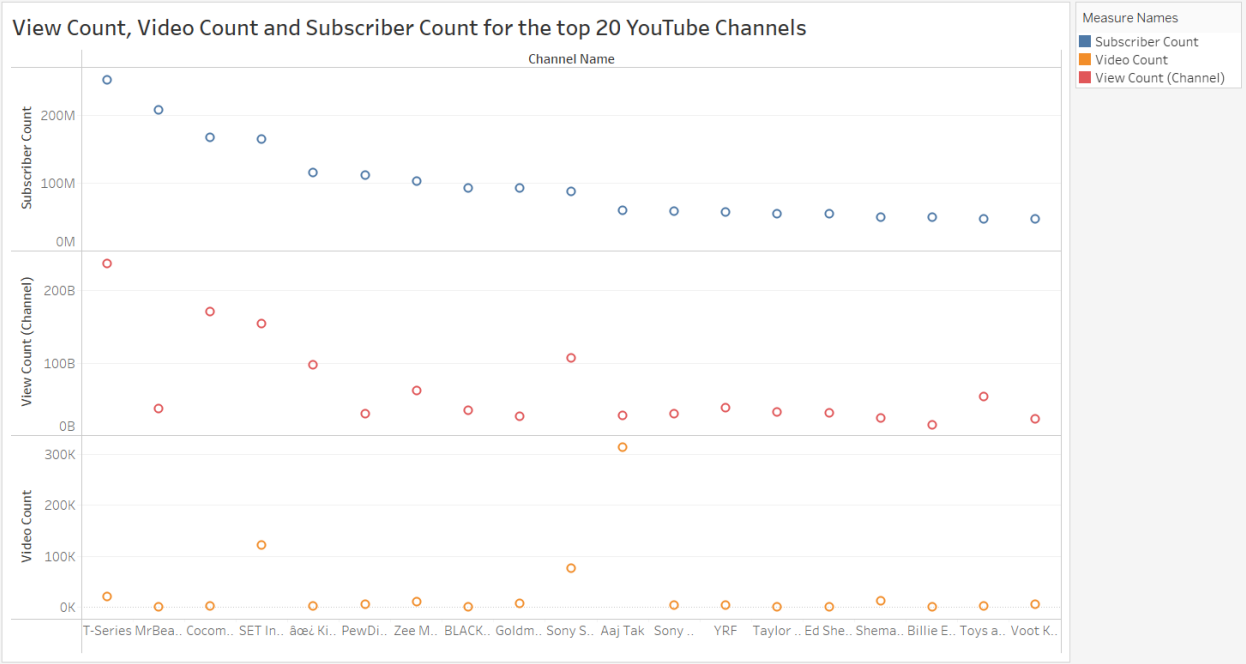
We connected Amazon Redshift with Tableau and performed visualizations.

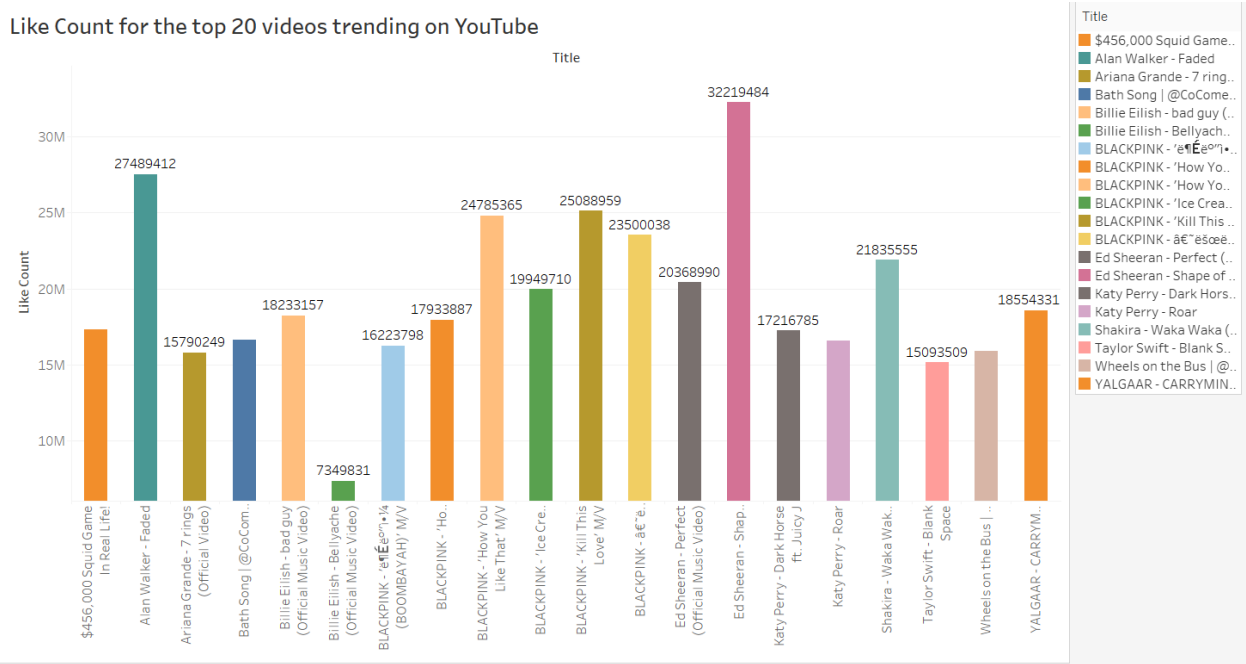


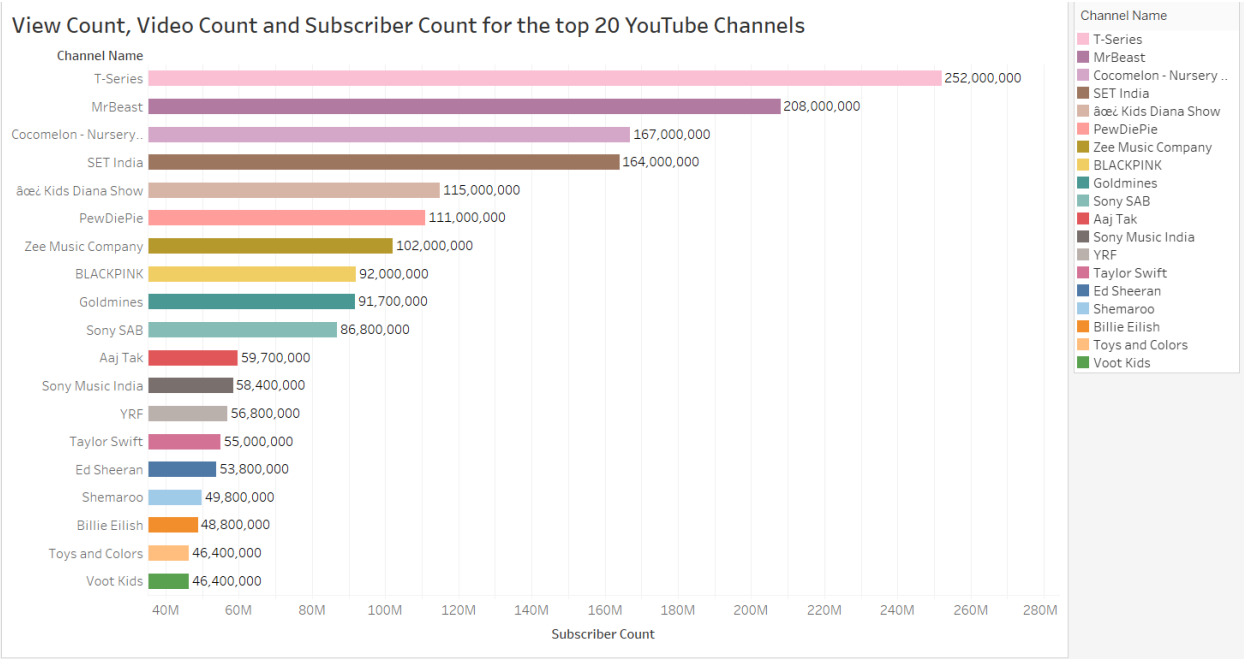



Comment Count of video with video_id 'KZRKoi8jCz0'
for 9 days in the month of November
Day Retrieved
Comment Count
9965
9964
9963
9962
2023-11-12
2023-11-13
2023-11-14
2023-11-16
2023-11-18
2023-11-19
2023-11-20
2023-11-21
2023-11-22


Top 20 Channels with Highest Engagement Rate
Channel Name
Talk Money with Tom
East--West Sounds
A 9-YEAR-OLD BOY
CarryMinati
Salim Ahmed
Santa Cruz Medicinals
Best Fresh Marchen...
albert_cancook
Aarzoo Khurana
AmandaRachLee
Jonah Thompson
PewDiePie
Hyram
Olivia Rodrigo
Joe HaTTab
0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5 4.0 4.5 5.0 5.5 6.0 6.5 7.0 7.5 8.0 8.5 9.0 9.5
Engagement Rate
Channel Name
A 9-YEAR-OLD BOY
Aarzoo Khurana
albert_cancook
AmandaRachLee
Best Fresh Marchenko
CarryMinati
East--West Sounds
Hyram
Joe HaTTab
Jonah Thompson
Olivia Rodrigo
PewDiePie
Salim Ahmed
Santa Cruz Medicinals
Talk Money with Tom


View Count, Video Count and Subscriber Count for the top 20 YouTube Channels
Channel Name
Subscriber Count
200M
100M
0M
View Count (Channel)
200B
100B
0B
Video Count
300K
200K
100K
0K
T-Series
MrBea..
Cocom..
SET In..
âœ¿ Ki..
PewDi..
Zee M..
BLACK..
Goldm..
Sony S..
Aaj Tak
Sony ..
YRF
Taylor ..
Ed She..
Shema..
Billie E..
Toys a..
Voot K..
Measure Names
Subscriber Count
Video Count
View Count (Channel)


Like Count for the top 20 videos trending on YouTube
Title
Like Count
30M
25M
20M
15M
10M
27489412
15790249
18233157
7349831
16223798
17933887
24785365
19949710
25088959
23500038
20368990
32219484
17216785
21835555
15093509
18554331
$456,000 Squid Game In Real Life!
Alan Walker - Faded
Ariana Grande - 7 rings (Official Video)
Bath Song | @CoCom..
Billie Eilish - bad guy (Official Music Video)
Billie Eilish - Bellyache (Official Music Video)
BLACKPINK - ë¶ÉëºŋÂ¼ (BOOMBAYAH)' M/V
BLACKPINK - 'Ho..
BLACKPINK - 'How You Like That' M/V
BLACKPINK - 'Ice Cre..
BLACKPINK - 'Kill This Love' M/V
BLACKPINK - â€˜è..
Ed Sheeran - Perfect (Official Music Video)
Ed Sheeran - Shap..
Katy Perry - Dark Horse ft. Juicy J
Katy Perry - Roar
Shakira - Waka Wak..
Taylor Swift - Blank Space
Wheels on the Bus | ...
YALGAAR - CARRYM..
Title
$456,000 Squid Game..
Alan Walker - Faded
Ariana Grande - 7 ring..
Bath Song | @CoCome..
Billie Eilish - bad guy (..
Billie Eilish - Bellyach..
BLACKPINK - ë¶ÉëºŋÂ¼..
BLACKPINK - 'How Yo..
BLACKPINK - 'How Yo..
BLACKPINK - 'Ice Crea..
BLACKPINK - 'Kill This ..
BLACKPINK - â€˜èšœè..
Ed Sheeran - Perfect (..
Ed Sheeran - Shape of ..
Katy Perry - Dark Hors..
Katy Perry - Roar
Shakira - Waka Waka (..
Taylor Swift - Blank S..
Wheels on the Bus | @..
YALGAAR - CARRYMIN..


View Count, Video Count and Subscriber Count for the top 20 YouTube Channels
Channel Name
T-Series 252,000,000
MrBeast 208,000,000
Cocomelon - Nursery.. 167,000,000
SET India 164,000,000
âœ¿ Kids Diana Show 115,000,000
PewDiePie 111,000,000
Zee Music Company 102,000,000
BLACKPINK 92,000,000
Goldmines 91,700,000
Sony SAB 86,800,000
Aaj Tak 59,700,000
Sony Music India 58,400,000
YRF 56,800,000
Taylor Swift 55,000,000
Ed Sheeran 53,800,000
Shemaroo 49,800,000
Billie Eilish 48,800,000
Toys and Colors 46,400,000
Voot Kids 46,400,000
40M 60M 80M 100M 120M 140M 160M 180M 200M 220M 240M 260M 280M
Subscriber Count
Channel Name
T-Series
MrBeast
Cocomelon - Nursery ..
SET India
âœ¿ Kids Diana Show
PewDiePie
Zee Music Company
BLACKPINK
Goldmines
Sony SAB
Aaj Tak
Sony Music India
YRF
Taylor Swift
Ed Sheeran
Shemaroo
Billie Eilish
Toys and Colors
Voot Kids