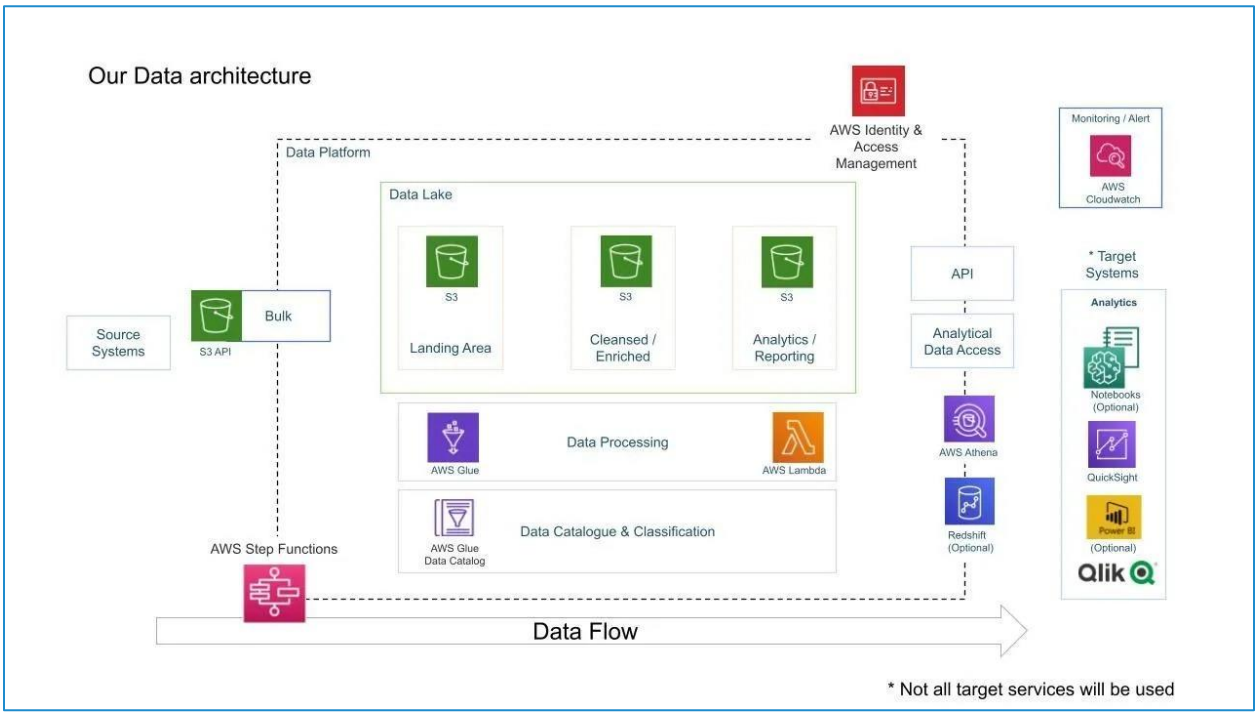
YOU TUBE DATA ANALYSIS - END TO END DATA ENGINEERING PROJECT

Our project's objectives are to safely organize, optimize, and analyze structured and semi-structured YouTube video data using trending metrics and video categories.



The Architecture diagram shows that we load the dataset onto an AWS S3 bucket first. It would be necessary for us to perform specific data transformations to correctly process the provided data. We would use AWS Lambda and Glue ETL to accomplish this. We will store the data in a different bucket after it has been cleaned and converted. To generate the final table, we would connect the two distinct tables that were produced by the previous phases. After that, we would use AWS QuickSight to visualize the data that was created.

Project Goals:

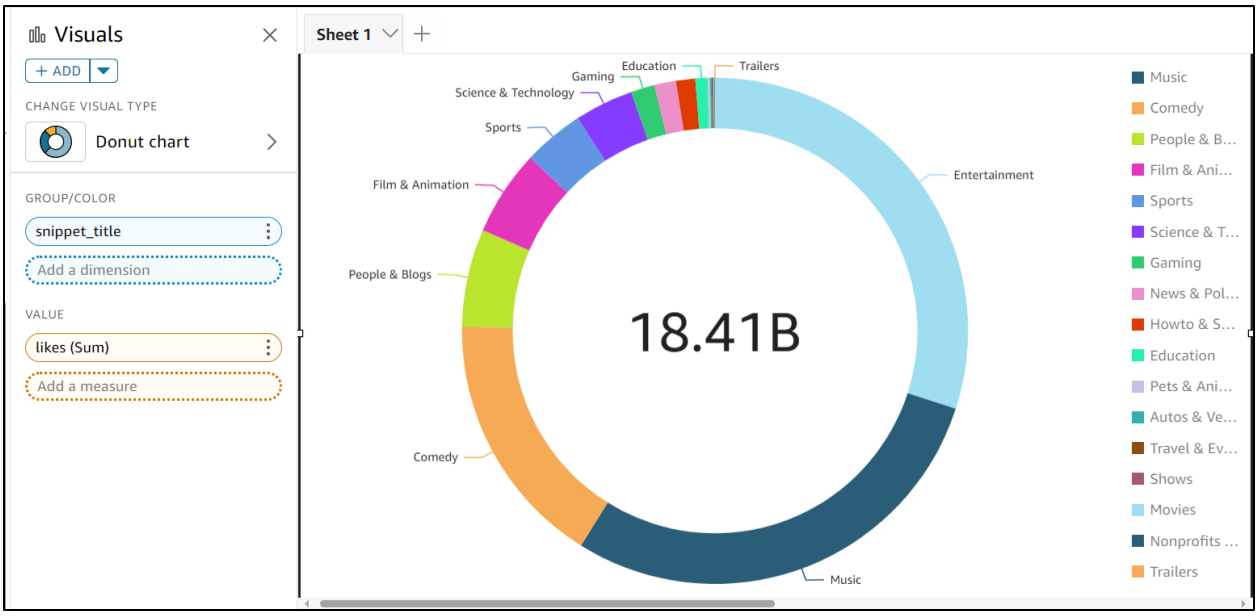
1. Data Ingestion -Building a mechanism to ingest data from different sources.
2. ETL System - Transforming raw data into a proper structured format.

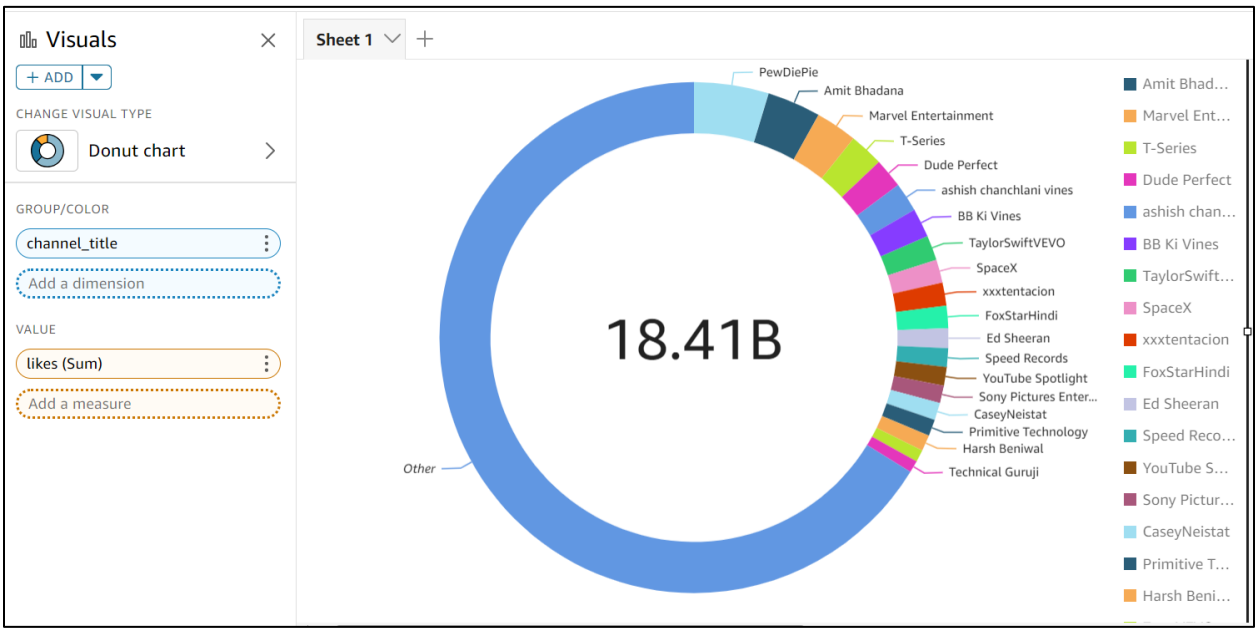
3. Data lake - We will be getting data from multiple sources, so we need a centralized repo to store it.
4. Scalability - As the size of our data increases, we need to make sure our system scales with it.
5. Cloud - We can’t process vast amounts of data on our local computer so we need to use the cloud, in this case, we will use AWS.
6. Reporting - Build a dashboard for data analysis

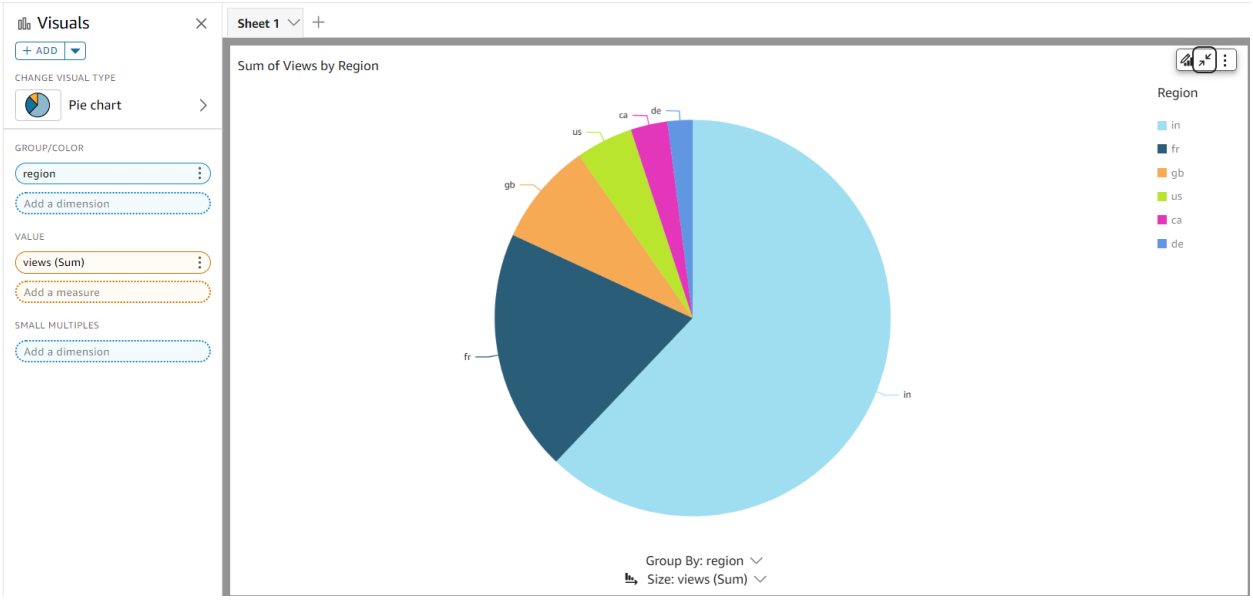
Technologies Used:

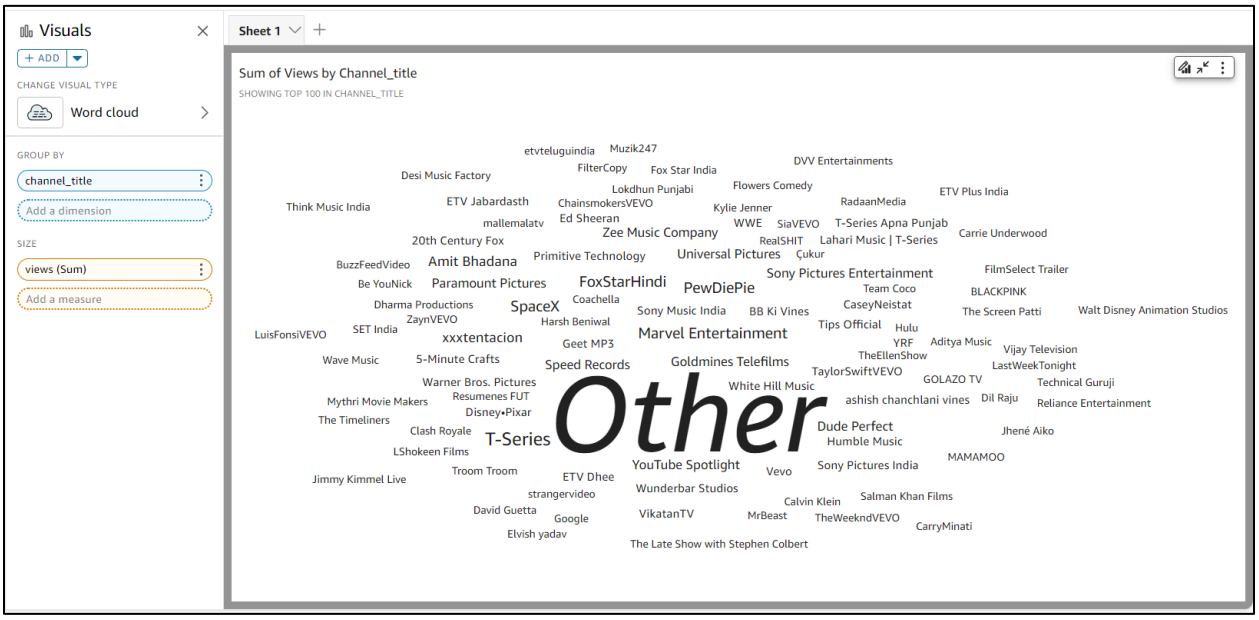
1. Amazon S3: Amazon S3 is an object storage service that provides manufacturing scalability, data availability, security, and performance.
2. AWS IAM: AWS Identity and Access Management (IAM) is a service that manages user identities and their access to AWS resources.
3. QuickSight: Amazon QuickSight is a scalable, serverless, embeddable, machine learning-powered business intelligence (BI) service built for the cloud.
4. AWS Glue: A serverless data integration service that makes it easy to discover, prepare, and combine data for analytics, machine learning, and application development.
5. AWS Lambda: Lambda is a computing service that allows programmers to run code without creating or managing servers.
6. AWS Athena: Athena is an interactive query service for S3 in which there is no need to load data it stays in S3.

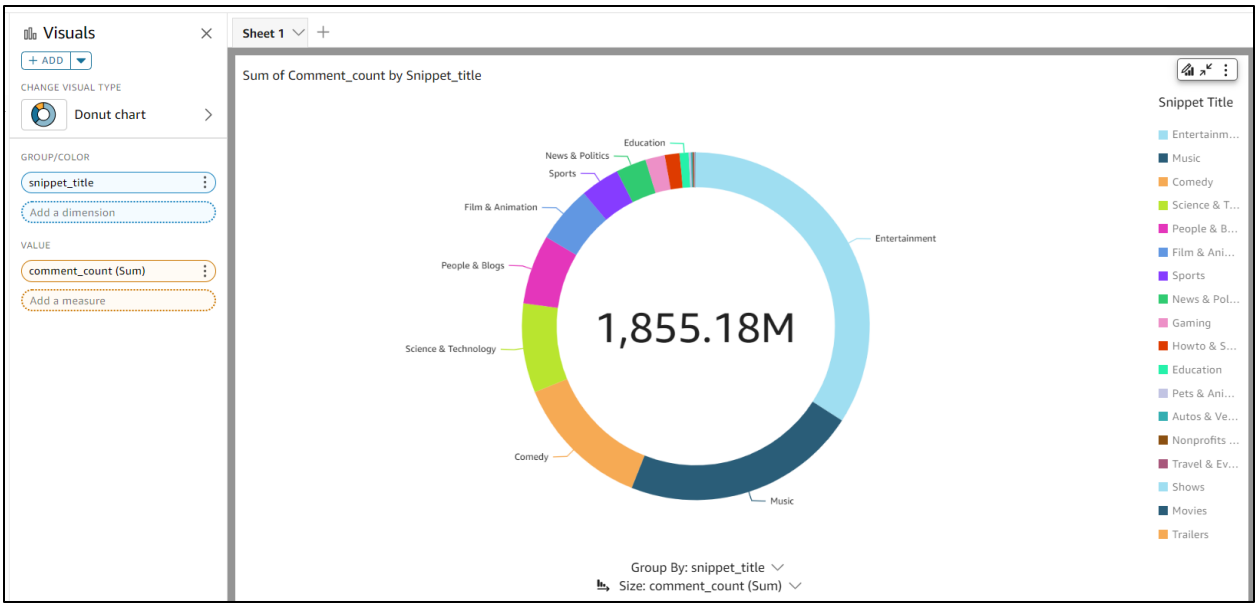
We are taken to the Quicksight website featuring a drawing board after completing all the previously mentioned stages. Quicksight offers a plethora of graphs and visuals around which to build a multitude of analysis. By experimenting with the various visualizations and column names, we can produce insightful and practical knowledge.


Visuals
+ ADD
CHANGE VISUAL TYPE
Donut chart
GROUP/COLOR
snippet_title
Add a dimension
VALUE
likes (Sum)
Add a measure
Sheet 1
Education
Trailers
Gaming
Science & Technology
Sports
Film & Animation
People & Blogs
Comedy
Entertainment
Music
18.41B
Music
Comedy
People & B...
Film & Ani...
Sports
Science & T...
Gaming
News & Pol...
Howto & S...
Education
Pets & Ani...
Autos & Ve...
Travel & Ev...
Shows
Movies
Nonprofits ...
Trailers


Visuals
+ ADD
CHANGE VISUAL TYPE
Donut chart
GROUP/COLOR
channel_title
Add a dimension
VALUE
likes (Sum)
Add a measure
Sheet 1
PewDiePie
Amit Bhadana
Marvel Entertainment
T-Series
Dude Perfect
ashish chanchlani vines
BB Ki Vines
TaylorSwiftVEVO
SpaceX
xxxtentacion
FoxStarHindi
Ed Sheeran
Speed Records
YouTube Spotlight
Sony Pictures Enter...
CaseyNeistat
Primitive Technology
Harsh Beniwal
Technical Guruji
Other
18.41B
Amit Bhad...
Marvel Ent...
T-Series
Dude Perfect
ashish chan...
BB Ki Vines
TaylorSwift...
SpaceX
xxxtentacion
FoxStarHindi
Ed Sheeran
Speed Reco...
YouTube S...
Sony Pictur...
CaseyNeistat
Primitive T...
Harsh Beni...


Visuals
+ ADD
CHANGE VISUAL TYPE
Pie chart
GROUP/COLOR
region
Add a dimension
VALUE
views (Sum)
Add a measure
SMALL MULTIPLES
Add a dimension
Sheet 1
Sum of Views by Region
Region
in
fr
gb
us
ca
de
Group By: region
Size: views (Sum)


Visuals
+ ADD
CHANGE VISUAL TYPE
Word cloud
GROUP BY
channel_title
Add a dimension
SIZE
views (Sum)
Add a measure
Sheet 1
Sum of Views by Channel_title
SHOWING TOP 100 IN CHANNEL_TITLE
Other
T-Series
etvteluguindia
Muzik247
FilterCopy
Fox Star India
DVV Entertainments
Desi Music Factory
Lokdhun Punjabi
Flowers Comedy
ETV Plus India
Think Music India
ETV Jabardasth
ChainsmokersVEVO
Kylie Jenner
RadaanMedia
Ed Sheeran
WWE
SiaVEVO
T-Series Apna Punjab
mallemalatv
Zee Music Company
RealSHIT
Lahari Music | T-Series
Carrie Underwood
20th Century Fox
Primitive Technology
Universal Pictures
Çukur
BuzzFeedVideo
Amit Bhadana
Sony Pictures Entertainment
FilmSelect Trailer
Be YouNick
Paramount Pictures
FoxStarHindi
PewDiePie
Team Coco
BLACKPINK
Dharma Productions
SpaceX
Coachella
CaseyNeistat
ZaynVEVO
Sony Music India
BB Ki Vines
The Screen Patti
Walt Disney Animation Studios
LuisFonsiVEVO
SET India
Harsh Beniwal
Tips Official
Hulu
xxxtentacion
Marvel Entertainment
YRF
Aditya Music
Geet MP3
TheEllenShow
Vijay Television
Wave Music
5-Minute Crafts
Speed Records
Goldmines Telefilms
LastWeekTonight
TaylorSwiftVEVO
Warner Bros. Pictures
White Hill Music
GOLAZO TV
Technical Guruji
Mythri Movie Makers
Resumenes FUT
ashish chanchlani vines
Dil Raju
Reliance Entertainment
Disney•Pixar
The Timeliners
Clash Royale
Dude Perfect
Jhené Aiko
Humble Music
LShokeen Films
MAMAMOO
Troom Troom
YouTube Spotlight
Vevo
Sony Pictures India
Jimmy Kimmel Live
ETV Dhee
Wunderbar Studios
strangervideo
Calvin Klein
Salman Khan Films
David Guetta
VikatanTV
MrBeast
TheWeekndVEVO
Google
CarryMinati
Elvish yadav
The Late Show with Stephen Colbert



Thus, we were able to safely handle, organize, and analyze the structured and semi-structured YouTube video data based on the video categories and trending metrics by utilizing AWS products like Glue, Lambda, Athena, and Quicksight.

By doing so, we ended up creating an ETL pipeline, which extracts raw data from a Kaggle dataset, applies various transformations to it, and finally saves the transformed data in the most efficient manner.