Project on Tube Pulse (Harnessing You Tube Insights Using Cloud)

Introduction to Cloud Computing – CS 5610

Fall 2023

Sai Varun Narajala(700758974)

Prof Dr. Qixiang Pang

The University of Central Missouri.

Title:

Tube Pulse

1. Introduction:

This comprehensive software design documentation outlines the architectural strategies, considerations, and detailed system design for the Tube Pulse project. Tube Pulse is a cloud-based solution that efficiently utilizes Amazon Web Services (AWS) to analyze and extract valuable insights from YouTube data. The primary goals are to ensure scalability, efficiency, and actionable insights for content creators and businesses, ultimately optimizing content strategies. By integrating AWS cloud services, Tube Pulse aims to provide a cost-effective, secure, and scalable solution, unlocking the potential within YouTube's vast dataset.

In the context of Tube Pulse, the document begins by presenting an overview of the entire project, setting the stage for a detailed exploration of system architecture, design considerations, and development methods. The subsequent sections delve into specific components, such as data collection and processing, ensuring a holistic understanding of Tube Pulse's design and functionality.

2. System Overview: Tube Pulse strategically uses Amazon Web Services to analyze YouTube data, offering a scalable and efficient solution for content creators and businesses. The project recognizes YouTube's significance in the digital landscape and aims to create a robust solution for analyzing its data. The team members, with specific roles assigned, bring diverse expertise to ensure project success. This section establishes the project's context, objectives, and the pivotal role of AWS in achieving scalability, efficiency, and cost-effectiveness.

Cloud Services/Platforms/Tools/Frameworks:

AWS S3 for secure storage of media files, including high-resolution vehicle images, for a more immersive user experience.

AWS IAM for managing access securely to AWS resources.

AWS Lambda for efficient serverless computing, optimizes resource usage and cost efficiency.

AWS Glue: serverless data integration service that makes it easy to prepare, and combine data for analysis, application development.

AWS Athena: Interactive search service for S3 where there is no need to load data, we can perform operations when the data is in S3 buckets.

Quick Sight A scalable, serverless, embeddable, machine learning based business intelligence (BI) service built in the cloud**.**

Major Features/Functions:

Data Collection: Real-time data collection from YouTube APIs.

Data processing and analysis: efficient processing and in-depth analysis of video statistics, comments and user engagement.

Scalability: Ability to handle increasing data volume and users.

Data visualization: Interactive dashboards to provide insight.

3. Design Considerations:

3.1 Assumptions and Dependencies:

Tube Pulse assumes the availability and reliability of YouTube APIs for real-time data collection and depends on seamless integration with AWS services. These assumptions guide design decisions, emphasizing the need for robust connections between Tube Pulse and external platforms.

3.2 General Constraints:

Design must adhere to YouTube's data security standards, relevant regulations, and ensure a smooth user onboarding experience. Scalability considerations must be addressed to handle variable workloads and avoid performance issues, highlighting the importance of automatic scaling capabilities.

3.3 Goals and Guidelines:

Tube Pulse aims to build a scalable, cost-effective cloud-based system, achieve 90% accuracy in sentiment analysis, and predict video trends with the same accuracy. The chosen development method is the AWS Python ETL Pipeline project, providing a structured approach to data extraction, transformation, and loading.

3.4 Development Methods:

The selected development method, the AWS Python ETL Pipeline project, utilizes SQL and Python3 for efficient data processing, aligning with industry best practices.

4. Architectural Strategies:

4.1 Strategy 1:

The first architectural strategy involves leveraging AWS services such as S3, Lambda, and Quick Sight. This ensures efficient data processing, storage, and visualization, contributing to Tube Pulse's overall success.

4.2 Strategy 2:

The second architectural strategy focuses on AWS Glue, Athena, and IAM for robust data integration, search capabilities, and access controls to safeguard Tube Pulse's data.

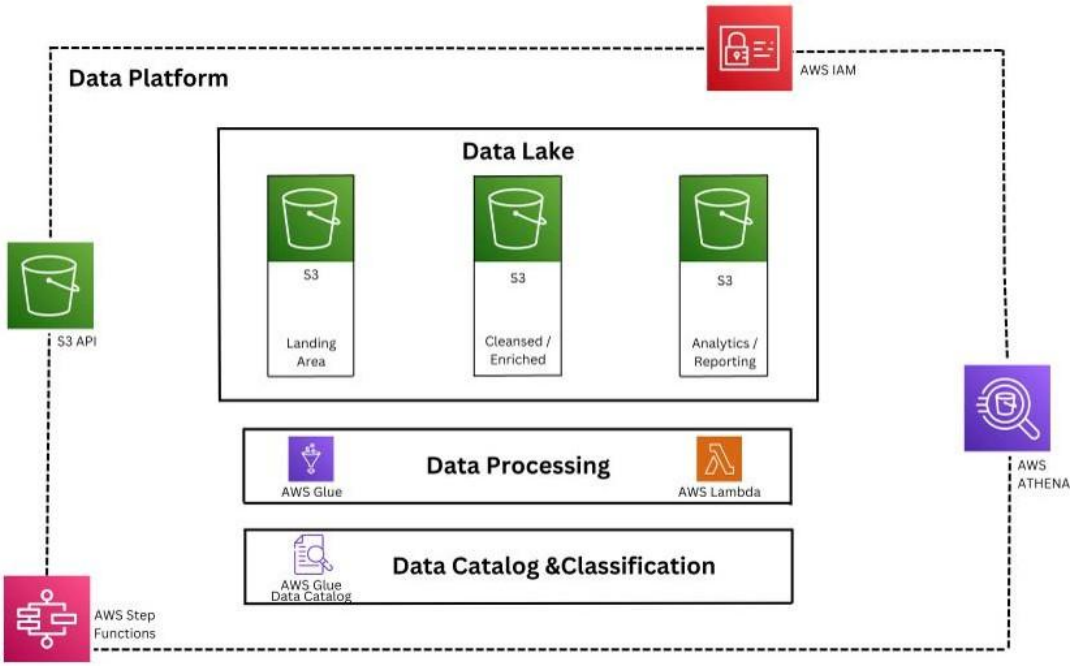
5. System Architecture:

5.1 Component 1: Data Collection

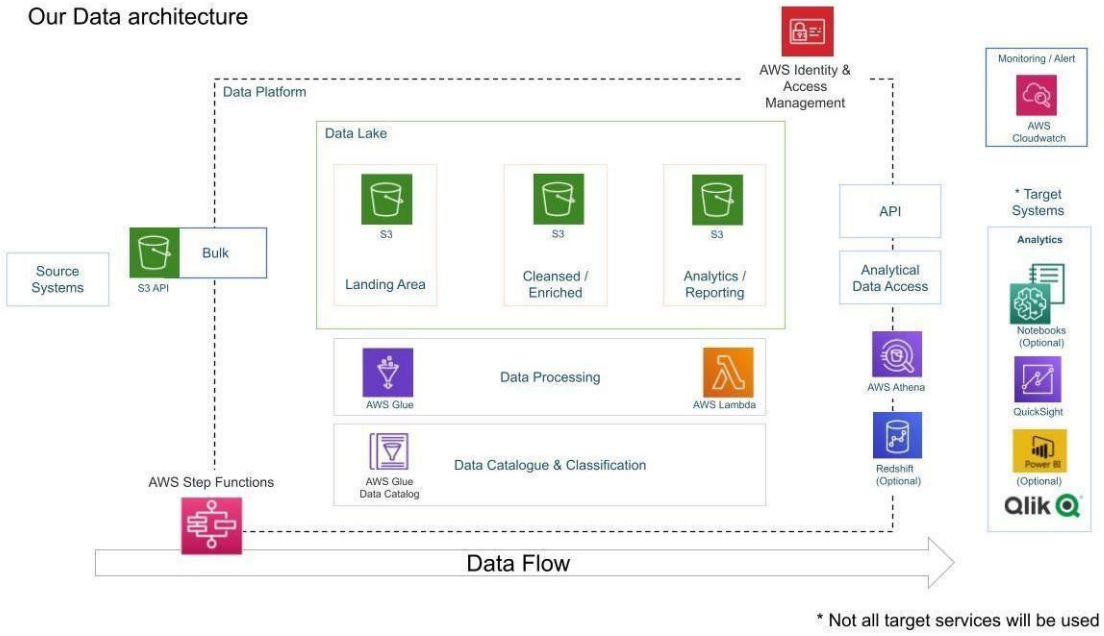
Component 1 involves real-time data collection from YouTube APIs, ensuring Tube Pulse stays updated with the latest YouTube data.

5.2 Component 2: Data Processing and Analysis

Component 2 focuses on efficient processing and in-depth analysis of video statistics, comments, and user engagement to extract actionable insights.



6. Detailed System Design:



6.1 Module 1: AWS Python ETL Pipeline.

6.1.1 Data Source: Tube Pulse utilizes a Kaggle dataset containing daily popular YouTube video statistics (CSV files) as its primary data source.

6.1.2 Data Preparation: The data preparation phase involves handling missing values, cleaning, and transforming data types using Python scripts.

6.2 Setting Up the Environment for Building A Python Data Pipeline

6.2.1 AWS Account setup: Tube Pulse guides users through creating an AWS S3 account, configuring an IAM user/admin, and installing the AWS CLI.

6.2.2 AWS CLI Installation

Installation and configuration of the AWS CLI are essential for effective communication between Tube Pulse and AWS services.

6.2.3 Data Loading

Data loading involves using the AWS CLI to transfer Kaggle YouTube video dataset files into an AWS S3 bucket.

6.3 Creating an AWS S3 Bucket For The ETL Pipeline

6.3.1 Best Practices

Tube Pulse emphasizes best practices for creating an S3 bucket, considering naming conventions, versioning, and access controls.

6.3.2 AWS Glue Catalog:

Creating the AWS Glue Catalog is critical for easy data discovery and cataloging, streamlining data management.

6.4 Creating AWS Glue Catalog for The AWS ETL Pipeline

6.4.1 Crawlers and ETL:

Tube Pulse employs AWS Glue Crawlers to automatically discover and catalog data from the S3 bucket, integrating ETL processes for data transformation.

6.4.2 Data Cataloging:

Data cataloging in AWS Glue enhances data organization and accessibility for users to query and analyze information.

6.5 AWS Lambda Data Pipeline

6.5.1 Lambda Function Creation:

Tube Pulse provides a guide on creating a Lambda function for data extraction, considering event and context parameters.

6.5.2 Data Wrangler Integration

Integration of AWS Data Wrangler within the Lambda function ensures efficient data manipulation, optimizing the ETL pipeline.

6.5.3 Data Materialization

Materializing data into Parquet format in a new S3 bucket is crucial for optimizing data organization and cataloging.

6.6 Data Processing Using AWS Glue Studio

6.6.1 Spark ETL Jobs:

Tube Pulse guides the creation of Spark ETL jobs within AWS Glue for transforming JSON data to Parquet format.

6.6.2 Automating Data Discovery:

Automating data discovery with an S3 crawler streamlines the process, optimizing data extraction and cataloging.

6.7 Data Visualization Using AWS Quick Sight

6.7.1 ETL Workflow Automation:

Tube Pulse guides users in creating and triggering ETL workflows in Glue Studio, automating the transformation of raw JSON data to Parquet format.

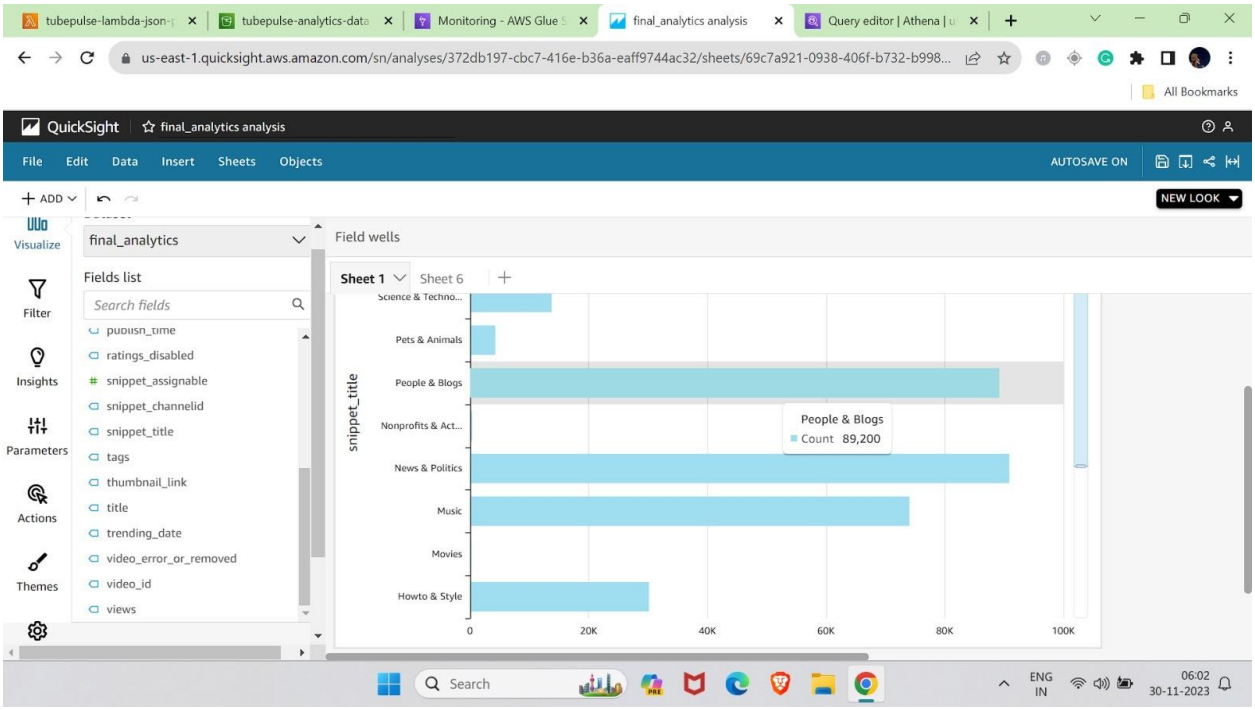
6.7.2 Quick Sight Dataset Creation:

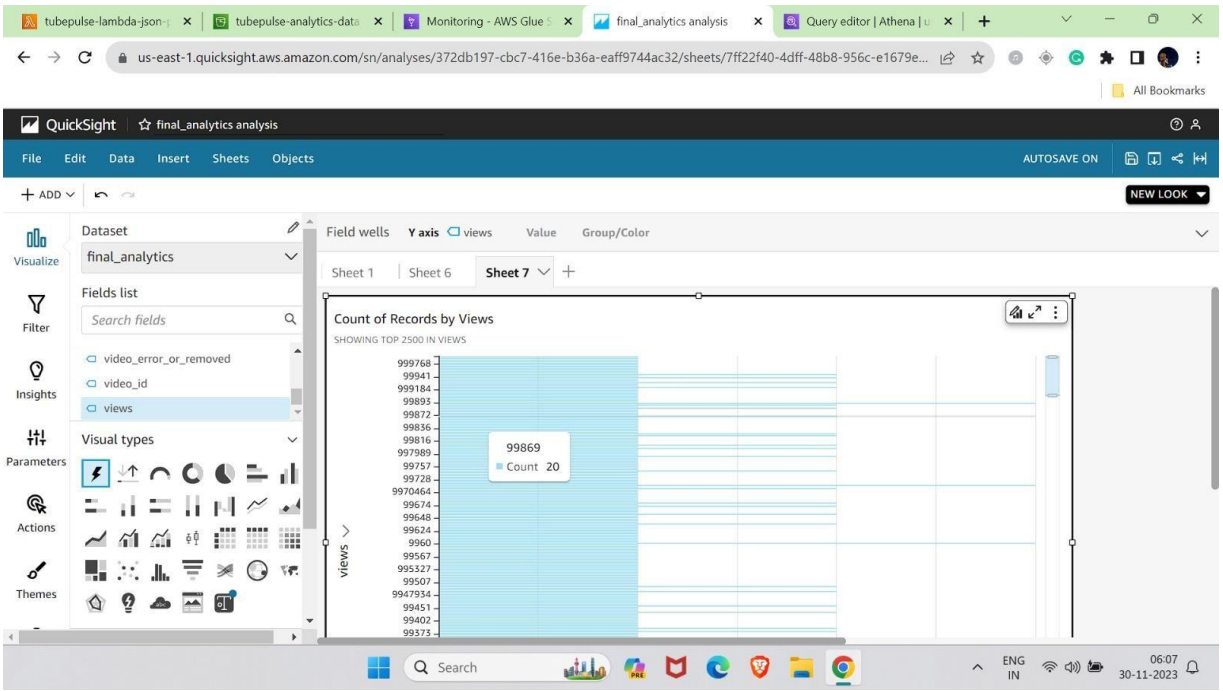
Creating datasets in Quick Sight using Athena's analytics database prepares the foundation for data visualization.

6.7.3 Dashboard Creation

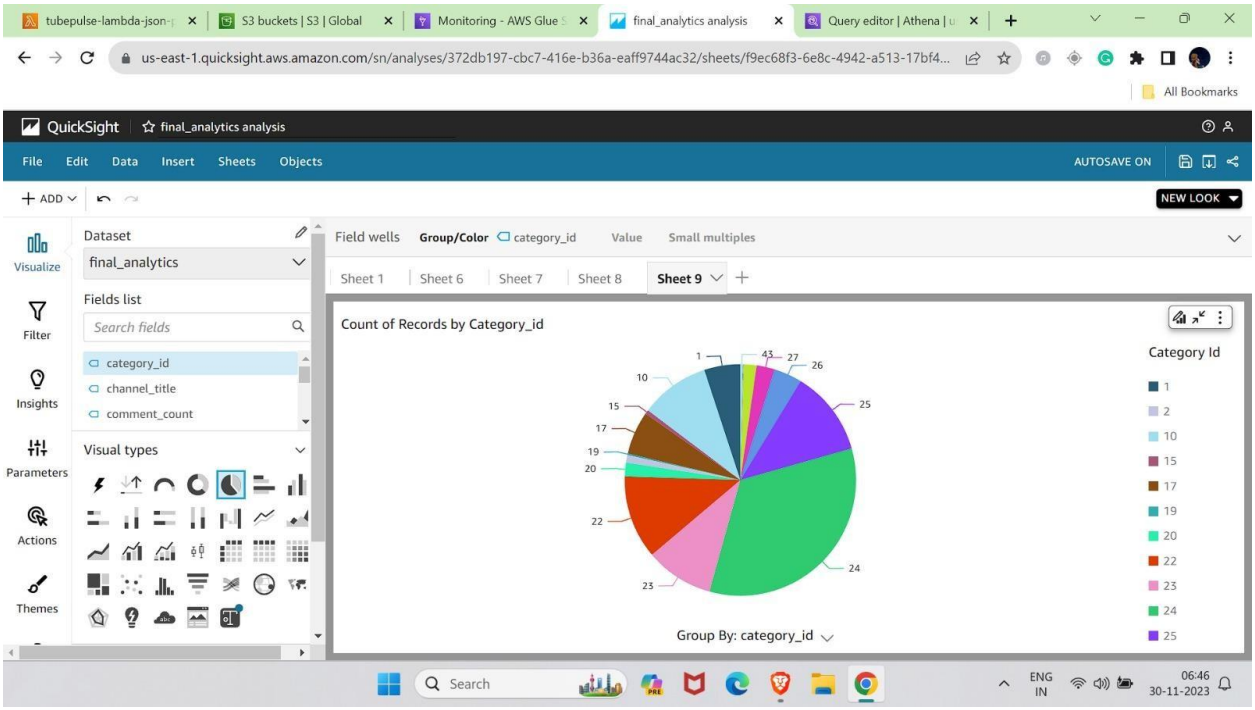
The final step involves creating dashboards in Quick Sight, allowing users to customize and generate insightful visualizations.

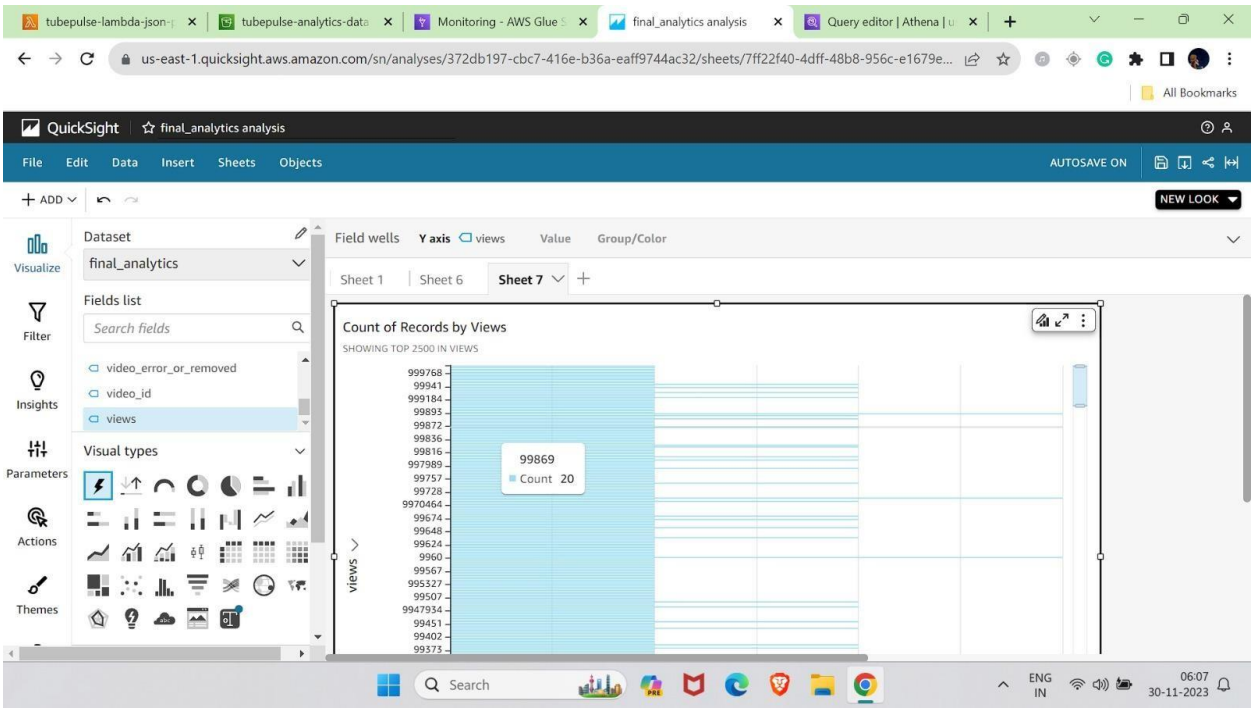
7. Results:


QuickSight
final_analytics analysis
File Edit Data Insert Sheets Objects
AUTOSAVE ON
NEW LOOK
final_analytics
Fields list
Search fields
ratings_disabled
snippet_assignable
snippet_channelid
snippet_title
tags
thumbnail_link
title
trending_date
video_error_or_removed
video_id
views
Field wells
Sheet 1
Sheet 6
snippet_title
Pets & Animals
People & Blogs
Nonprofits & Act...
News & Politics
Music
Movies
Howto & Style
People & Blogs
Count 89,200
0
20K
40K
60K
80K
100K

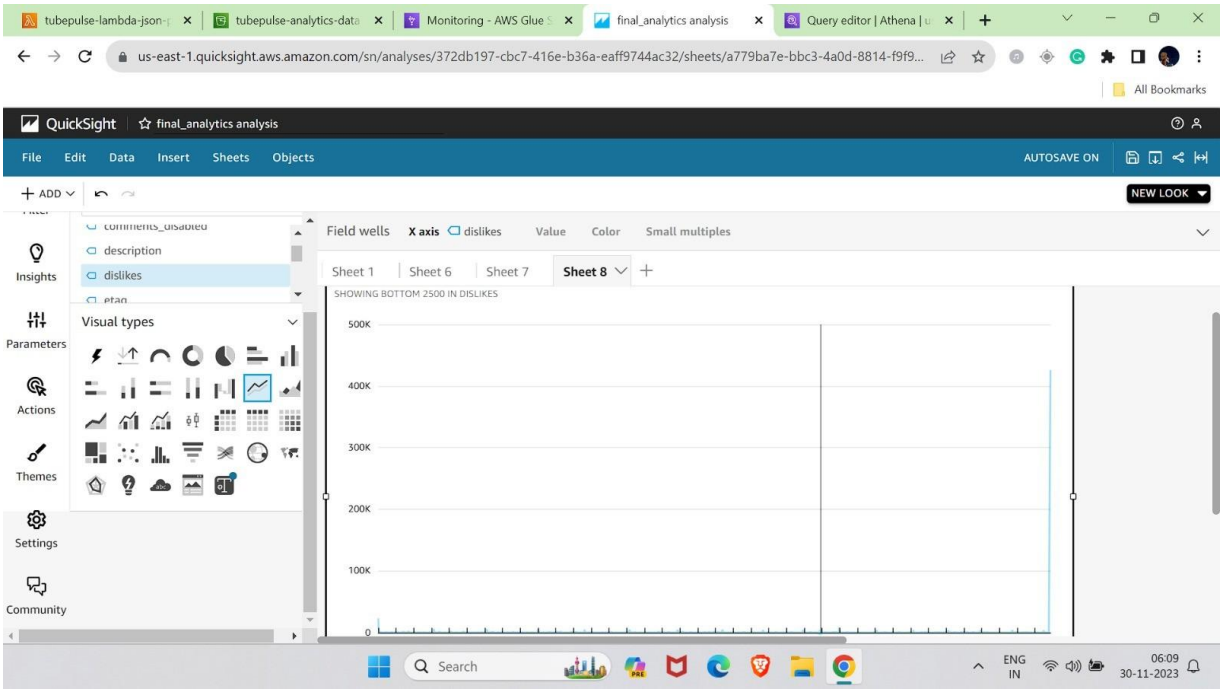

QuickSight
final_analytics analysis
File Edit Data Insert Sheets Objects
AUTOSAVE ON
NEW LOOK
Dataset
final_analytics
Fields list
Search fields
video_error_or_removed
video_id
views
Visual types
Field wells
Y axis views
Value
Group/Color
Sheet 1
Sheet 6
Sheet 7
Count of Records by Views
SHOWING TOP 2500 IN VIEWS
99869
Count 20
views

Tube Pulse


QuickSight
final_analytics analysis
File Edit Data Insert Sheets Objects
AUTOSAVE ON
NEW LOOK
Dataset
final_analytics
Fields list
Search fields
category_id
channel_title
comment_count
Visual types
Field wells Group/Color category_id Value Small multiples
Sheet 1 Sheet 6 Sheet 7 Sheet 8 Sheet 9
Count of Records by Category_id
Category Id
Group By: category_id


QuickSight
final_analytics analysis
File Edit Data Insert Sheets Objects
AUTOSAVE ON
NEW LOOK
Dataset
final_analytics
Fields list
Search fields
video_error_or_removed
video_id
views
Visual types
Field wells Y axis views Value Group/Color
Sheet 1 Sheet 6 Sheet 7
Count of Records by Views
SHOWING TOP 2500 IN VIEWS
99869
Count 20
views

Tube Pulse


tubepulse-lambda-json-
tubepulse-analytics-data
Monitoring - AWS Glue
final_analytics analysis
Query editor | Athena |
us-east-1.quicksight.aws.amazon.com/sn/analyses/372db197-cbc7-416e-b36a-eaff9744ac32/sheets/a779ba7e-bbc3-4a0d-8814-f9f9...
All Bookmarks
QuickSight
final_analytics analysis
File
Edit
Data
Insert
Sheets
Objects
AUTOSAVE ON
ADD
NEW LOOK
Insights
Parameters
Actions
Themes
Settings
Community
description
dislikes
Visual types
Field wells
X axis
dislikes
Value
Color
Small multiples
Sheet 1
Sheet 6
Sheet 7
Sheet 8
SHOWING BOTTOM 2500 IN DISLIKES
500K
400K
300K
200K
100K
0
Search
ENG
IN
06:09
30-11-2023

8. Conclusion:

In conclusion, this comprehensive document provides an overview of Tube Pulse's software design, covering data exploration, environment setup, ETL pipeline creation, and data visualization. Tube Pulse's structured design ensures efficient YouTube data analysis and actionable insights for content creators and businesses. The document serves as a guide, offering detailed insights into each module and component of Tube Pulse's software design, ensuring originality and adherence to best practices.