Data Engineering 2: Big Data Architectures

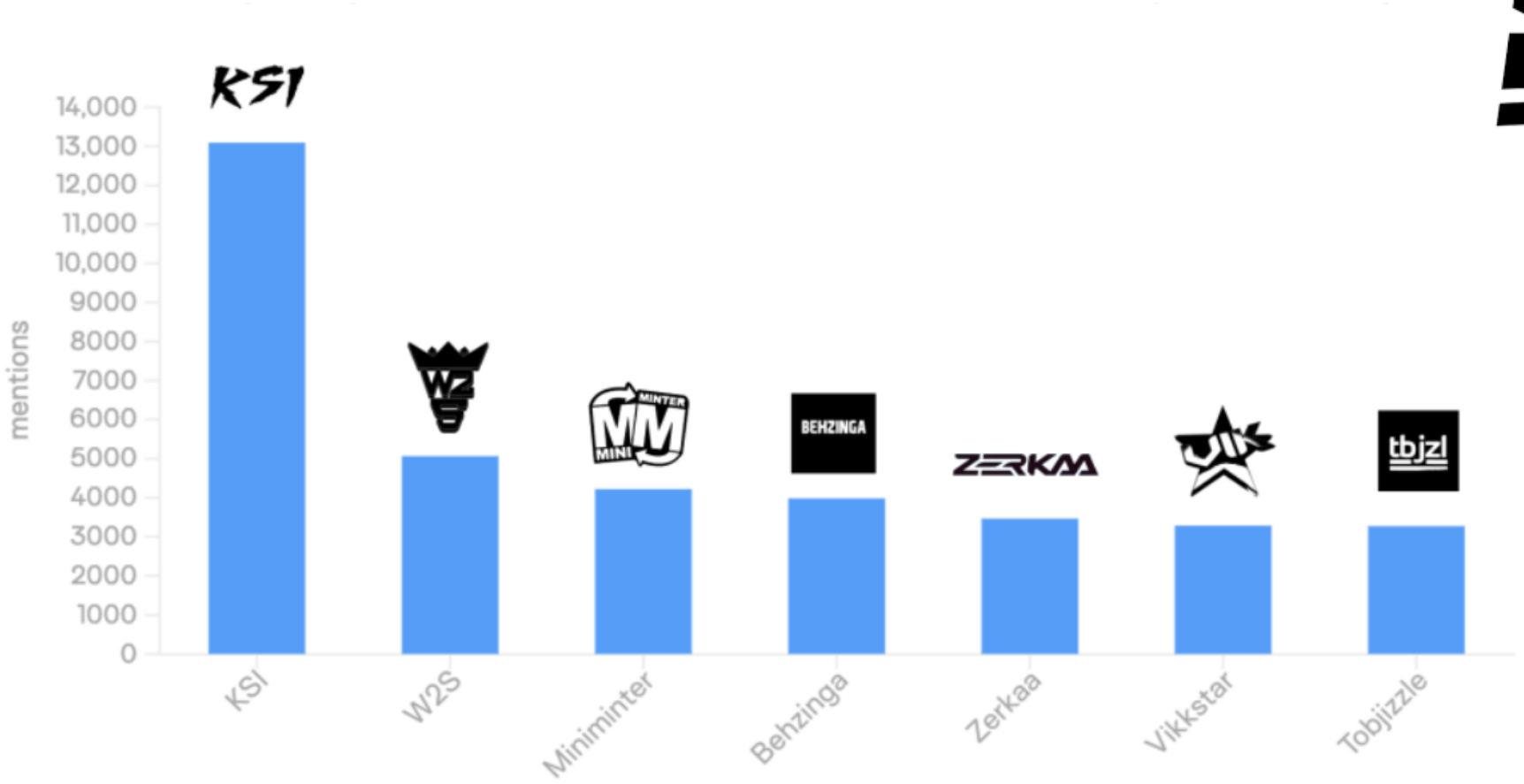
YouTube Data Pipeline

Dharshan Dhanashekar

Harshita Jamadade

Navneeth Krishna Aravind

Introduction : Sidemen



source : https://www.pulsarplatform.com/

Who are they ?

The Sidemen are a group of seven friends from the UK who create fun YouTube videos about games, challenges, and reactions, and have become super popular worldwide.

What Do they Have on Market?

SIDES

Food Ventures

SDMN

Sidemen Clothing

Alcohol

Prime Energy Drink

Introduction to the Data Source

Data Source: YouTube Data API v3

Purpose: Analyze performance of Sidemen's YouTube channels

Channel : Sidemen , Sidemen Reacts , More Sidemen ,Sidemen Shorts

Channel Statistics

subscriber count, total views, video count, and description.

Endpoint
youtube/v3/channels

Uploads Playlist

video IDs from the upload's playlist.

Endpoint
youtube/v3/playlistItems

Video Details

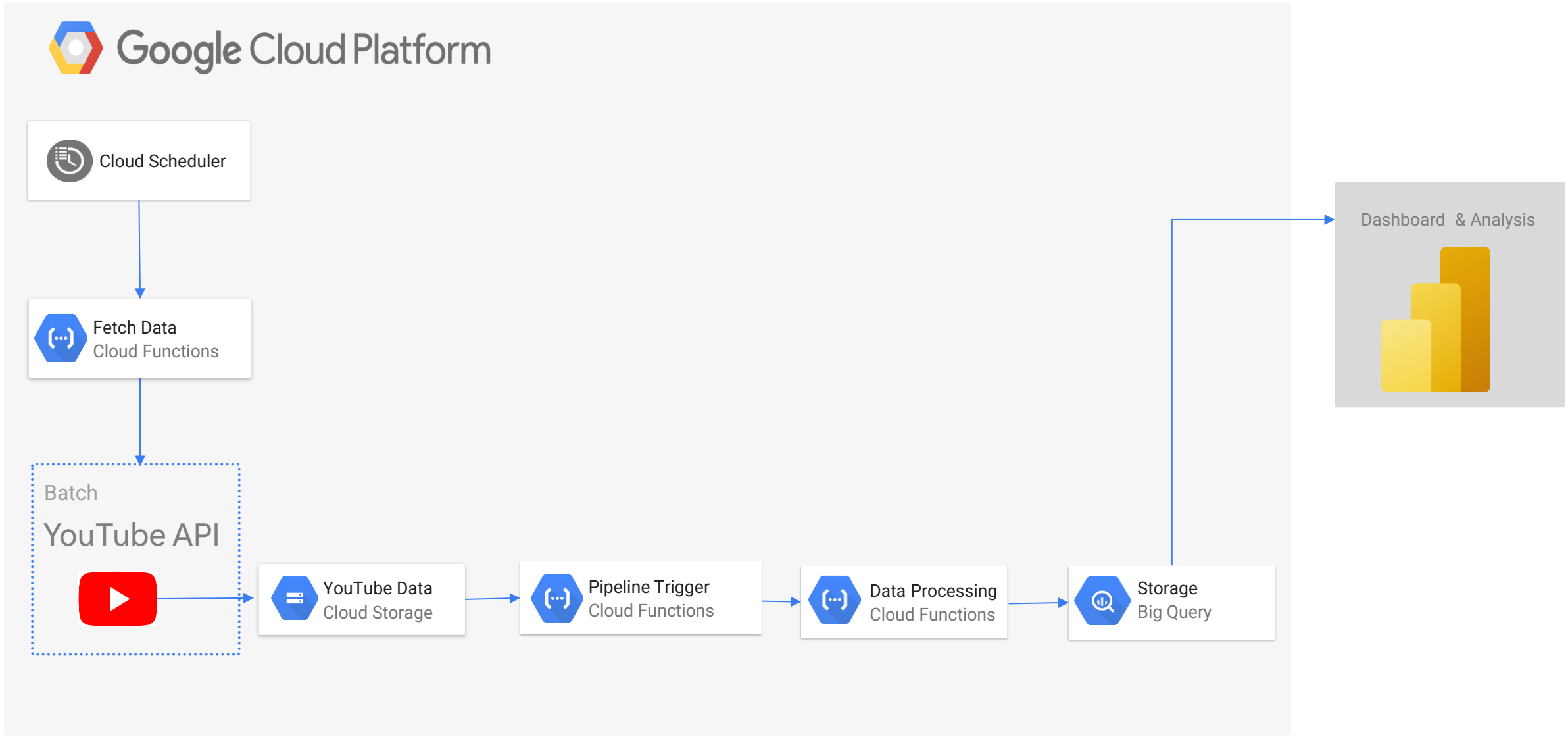
views, likes, comments, tags,length and upload dates.

Endpoint
youtube/v3/videos

Data Storage : Google Cloud Storage (GCS)

Format : JSON

Architechture



Data Processing

Read Raw Data

Data is ingested from JSON files stored in Google Cloud Storage.

Parse and Extract Records

Transforms the JSON array into individual records for further processing.

Transform Data
Custom transformations

- Calculate metrics (engagement rate, average views).
- Format timestamps and video durations.
- Identify viral videos

Load to BigQuery:
Data is written to BigQuery tables using schemas for:

- **Video Stats**
- **Channel Stats**

Challenges

Data Fetching and API Usage

Encountered "Too Many Requests" errors when fetching large-scale global data from the YouTube API.

> **Solution**: Implemented batching, added delays between requests, and optimized API usage to reduce requests.

Runtime and Timeout Issues

Function exceeded the default timeout of 60 seconds when processing large datasets, when handling data for multiple videos.

> **Solution**: Extended the **Cloud Function timeout limit** to the maximum allowable duration of **540 seconds** (9 minutes), enabling the function to process larger datasets.

Challenges

Handling JSON File Conversion in Apache Beam

Parsing large JSON files within an Apache Beam pipeline while maintaining performance.

> **Solution**: Implemented custom pipeline options to handle file paths, input formats, and processing parameters.

Worker Pool and Port Issues

> **Solution :** Debugged and resolved worker-related TypeErrors using insights from community forums.

Handling infinite trigger executions

> **Solution :** Added a timeout check.

Challenges

Data Loading Issues Across Systems

Encountered difficulties in loading and configuring data sources on a teammate's system when working with Tableau.

Solution: Shifted to Microsoft Power BI for visualization due to its compatibility.

Designing Dashboards

Struggled to create a visually appealing and user-friendly layout. Encountered issues with resizing dashboards.

Solution: Avoided overloading the dashboards with unnecessary visuals and opted for a fixed-size dashboard layout.

User Story

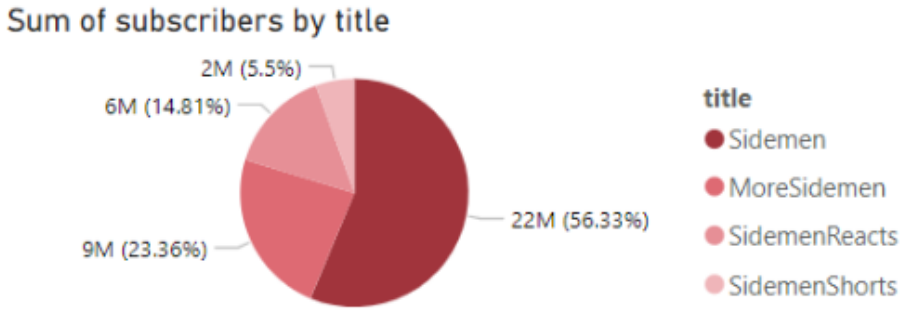
"As a Sidemen content manager, I want to analyze video performance metrics (views, likes, comments) across all Sidemen YouTube channels so that I can identify trends and create more engaging content."

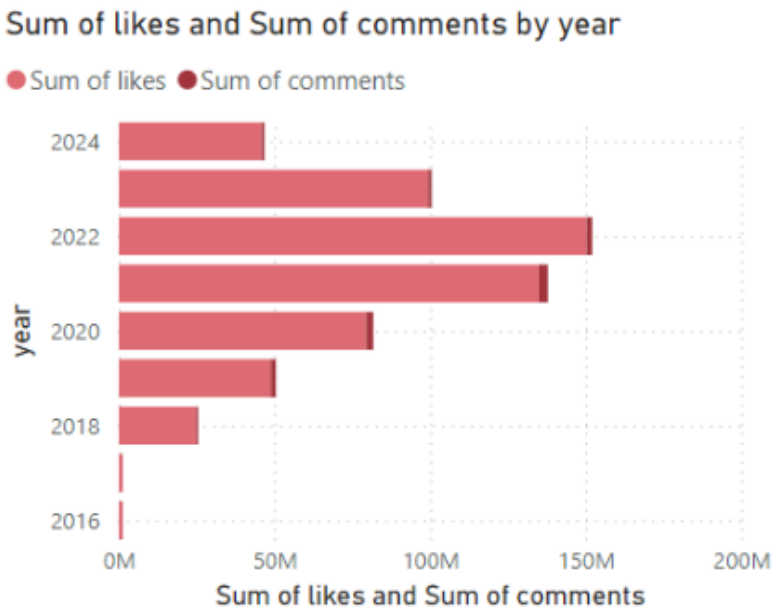
SIDEMEN INSIGHTS

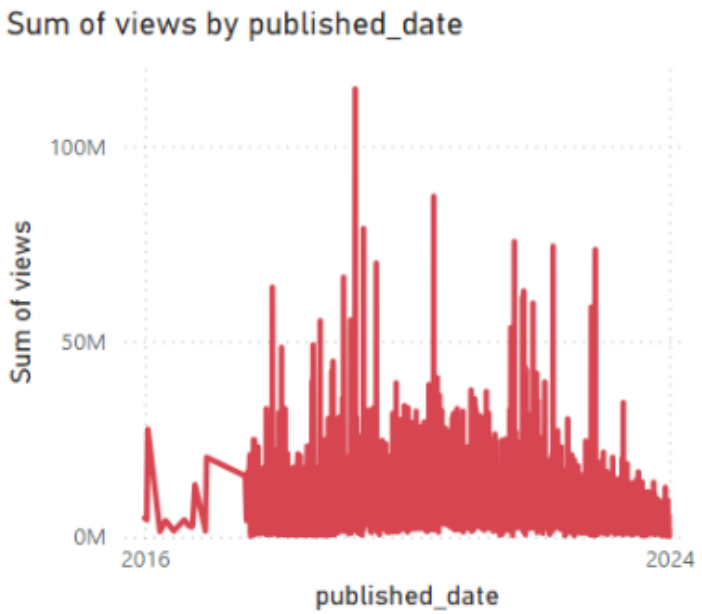
39M
Sum of subscribers

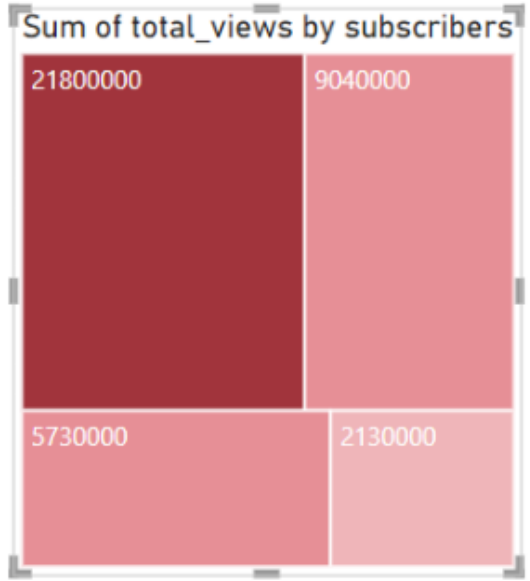
3.59
Engagement_rate

3.38M
Average Views Per ...









weekday

All

Pipeline video

YouTube-Data-Pipeline

fin.