Tracking NBA Channel Engagement Metric Trends through Web Scraping and YouTube API Analysis

Bipin Chandra Medasani   
*dept. of Computer Science*  
Central Michigan University *Mount Pleasant, USA*  
medas1bc@cmich.edu

Pavan Kumar Ponugoti  
*dept. of Computer Science*  
Central Michigan University *Mount Pleasant, USA*  
ponug1p@cmich.eduVenkata Sai Kumar Gedela  
*dept. of Computer Science*  
Central Michigan University *Mount Pleasant, USA*  
gedel1vs@cmich.edu

line 1: 5th Given Name Surname  
*dept. of Computer Science*  
Central Michigan University *Mount Pleasant, USA*  
line 5: email address or ORCID

*Abstract*—**This study describes the National Basketball Association (NBA) has grown into a global sports brand with a big global fan base. The NBA has used numerous avenues to communicate with fans and build its audience since the development of digital media. The NBA's official YouTube channel has evolved as a critical tool for information exchange and fan engagement. However, there has been little research into how fans interact with the NBA's YouTube content. Through web scraping and YouTube API analysis, this study aims to track the engagement metric trends of the NBA's YouTube channel. The study examines engagement indicators such as likes, dislikes, comments, views, and shares to get insight into NBA fans' interaction habits on YouTube.** **The study also gives significant information about NBA fans' YouTube choices and interaction patterns, which may help the NBA customize its content to better engage fans and build its audience. Overall, this study adds to the literature on social media analytics and engagement measures and lays the groundwork for future research in this field[4].**

Keywords—YouTube Analytics, Engagement metrics, Data Fusion, Data Engineering, Machine Learning, Google Cloud, BigQuery, Data Pipeline, Plugin

# Introduction

## Background and motivation

The National Basketball Association (NBA) is a worldwide sports brand with fans all over the world. Since the advent of digital media, the NBA has used a variety of methods to communicate with fans and grow its audience. The NBA's official YouTube account has proven to be a vital tool for engaging with fans and disseminating material throughout different channels [6]. Understanding how fans interact with the NBA's YouTube content is therefore crucial for the league's digital media strategy.

## Problem statement and research question

While the NBA's YouTube channel has become an important component of the league's digital media strategy, little research has been conducted on how fans interact with this content [8]. As a result, there is a need for research that may provide insights into NBA fans' YouTube involvement behaviors. The goal of this study is to find out what the engagement patterns of NBA fans are on the NBA's official YouTube channel, and how these patterns have changed over time.

## Objective and contribution

Through web scraping and YouTube API analysis, the primary goal of this study is to track the engagement metric trends of the NBA's YouTube channel. We want to obtain insights into the interaction patterns of NBA fans on the YouTube platform by studying engagement data such as likes, dislikes, comments, views, and shares. By offering a complete examination of the NBA's YouTube channel, our study will contribute to the literature on social media analytics and engagement measures.

Furthermore, our research will help the NBA's digital media strategy by providing valuable insights into NBA fans' YouTube preferences and engagement patterns. The NBA can modify its programming to better engage fans and build its audience by analyzing these tendencies.

## Overview of the paper

There are various portions to the paper. Following the introduction, the paper discusses related research on social media analytics and engagement metrics. The related work section identifies gaps in the literature and sets the context for our investigation. Following the section on related work, the paper discusses the methodology used to collect and analyze the data. The methodology section describes the web scraping and YouTube API analysis approaches used to obtain data on engagement metrics from the NBA's YouTube channel.

The results and discussion sections of the article reveal the outcomes of our investigation, emphasizing trends and insights regarding NBA fans' engagement habits on the YouTube platform. This section covers the findings from an examination of engagement metrics such as likes, dislikes, comments, views, and shares, as well as how they change among different video genres, duration, teams, and players.

The conclusion and future work section summarizes the findings of the study and discusses the implications for the NBA's digital media strategy. It emphasizes the significance of measuring and analyzing engagement data in order to continuously enhance digital media campaigns. Finally, the section on future work recommends potential areas for more study and analysis, such as investigating the link between engagement metrics and other characteristics, such as the time of day and day of the week when films are published.

# Related Work

## Review of related literature

In recent years, there has been a lot of study done on social media analytics and engagement measures [9]. Many studies, in particular, have concentrated on analyzing engagement metrics on YouTube, a popular video-sharing platform. Google Data Fusion is one of the most popular tools for evaluating engagement numbers on YouTube [11]. Researchers may use this application to collect data from many sources and analyze it using various criteria.

## Comparison with existing methods

While there are several existing methods for analyzing YouTube engagement metrics, Google Data Fusion is particularly useful because it provides a comprehensive approach to data collection and analysis. Manual data gathering or the use of single-source solutions, for example, might be time-consuming and limited in scope. Researchers may use Google Data Fusion to collect data from numerous sources, such as YouTube's API and site scraping, then evaluate it using various metrics.

## Gap identification

Despite several studies on social media analytics and engagement measures, little is known about how fans interact with sports content on YouTube. This research seeks to fill that void by examining engagement measure patterns on the NBA's official YouTube channel. Furthermore, while Google Data Fusion is a powerful tool for analyzing engagement metrics, more research on its effectiveness in the context of sports content on YouTube is needed. This study attempts to add to this field of study by tracking engagement measure trends on the NBA's YouTube channel using Google Data Fusion. By doing so, we expect to obtain insights into NBA fans' YouTube interaction habits and give vital information to the NBA for strengthening its digital marketing strategy.

# Methodology

Our research aims to create a data pipeline utilizing Google Data Fusion to acquire, cleanse, and prepare data on engagement metrics on NBA channel. This section gives an overview of our process as well as the stages needed in putting the pipeline in place.

## Data Fusion

Google Data Fusion is a cloud-based service that assists in data extraction, transformation, and loading (ETL) from various sources. ETL is a critical stage in the data engineering process that involves acquiring data from several sources, cleaning and transforming it, and then transferring it to a data warehouse or database for analysis [7]. ETL procedures have historically been complicated and time-consuming, requiring the writing of code as well as the configuration of many tools in order to complete the relevant tasks. Google Data Fusion facilitates this process by providing a graphical user interface that enables users to build and manage data pipelines using a drag-and-drop interface. This considerably facilitates the development, testing, and updating of data pipelines, even for those with little programming experience [3].

## Ingestion

The process of importing data from an external source into the data pipeline is known as ingestion. In our scenario, we imported patient data from Google Cloud Storage into Google Data Fusion in CSV format. Google Cloud Storage is a cloud-based object storage service that allows customers to store and retrieve data from any location that has an internet connection. We utilized the Cloud Storage plugin in Google Data Fusion to import the data [3][4]. A plugin is a software component that adds new features to an existing application. Users may connect to a Cloud Storage bucket, read data in multiple file formats, and upload it to Data Fusion using the Cloud Storage plugin. The plugin was set up to read CSV files and parse them into records, which are individual rows of data. We were able to proceed with the pipeline's following stages after the data was ingested, which included data transformation and merging with additional patient data from another source.

## Transformation

Following the data import process, the next step in our data pipeline was to convert the YouTube engagement metrics data in Google Data Fusion using the Wrangler plugin [3]. Several data cleaning procedures were performed, such as updating column names and data types, standardizing date formats, and fixing misspellings. These changes were required to verify that the data was uniform and ready to be analyzed further. Wrangler was also used to remove null values and duplicates from the dataset. This phase was critical since missing or duplicate data might result in problems in downstream processing and erroneous findings. We assured that the data was as clean as feasible by deleting these occurrences.

The Wrangler plugin in Google Data Fusion provides a visual interface for data cleansing and transformation, which greatly aided our team's efforts [7]. It made it simple to choose the columns we wanted to update and see the changes before committing to them. The plugin also provides a broad range of data transformation methods, such as altering data types, converting dates, and manipulating strings, allowing us to tailor our cleaning process to the exact demands of our data [6]. Overall, the transformation procedure was critical in preparing the data from engagement analysis for future study. It guaranteed that the data was correct, consistent, and ready for the next phase in our process.

## Cleaning

We got a comprehensive dataset with a broader variety of channel information when we combined the modified channel data with extra data on demographics and engagement history. However, it is probable that certain inconsistencies or missing values remained in this dataset that were not handled during the previous cleaning procedures. We used another instance of the Wrangler plugin in Data Fusion to eliminate any remaining missing values and inconsistencies, ensuring that the final data set was as clean and accurate as possible. This entailed manually reviewing the data and using suitable cleaning processes to ensure that it was correct and consistent across all records.

Furthermore, we utilized Data Fusion's Distinct plugin to remove any duplicate records from the dataset. This plugin assisted us in identifying and removing any entries that were exact copies of any other, ensuring that each patient record was unique and appeared only once in the dataset.

## Loading

After the dataset had been cleaned, it was loaded into a data storage system so that it could be researched. In this instance, we selected to load the dataset into BigQuery, a Google Cloud tool (GCP) cloud-based data warehousing and analytics tool [5].

We utilized Data Fusion's BigQuery plugin to write the cleaned dataset into a new table in the same project. This plugin made it easier to import data into BigQuery by allowing us to define the target table and other necessary configuration parameters. We were able to take advantage of BigQuery's advanced analytics and querying capabilities by importing the dataset into it. BigQuery is optimized for massive datasets and provides quick performance, making it an excellent choice for machine learning analysis. After loading the data into BigQuery, it was ready for additional analysis, such as developing machine learning models.

## Benefits of using Data Fusion

Google Cloud Data Fusion is a powerful platform for automating and expediting data integration, preparation, and analytic tasks. It is suited for both technical and non-technical users because to its scalability and user-friendliness, as well as its drag-and-drop interface. Data Fusion can easily access data from other Google Cloud services such as BigQuery, Cloud Storage, and Pub/Sub since it is completely connected with Google Cloud. It is also an intriguing alternative for enterprises trying to expand their data analytics skills, with a pay-as-you-go pricing plan and security features like as encryption at rest and in transit, multi-factor authentication, and identity and access management. Overall, Data Fusion on Google Cloud is an excellent choice for enterprises that need to quickly process, integrate, and analyze massive amounts of data and efficiently.

## How Google Data Fusion works

Google Data Fusion is a robust data integration technology that makes it easier to create and manage data pipelines. It makes use of Apache NiFi and Apache Beam, two open-source technologies used in the data engineering sector. Apache NiFi is a data integration tool with a graphical interface for creating data flows, whereas Apache Beam is a programming paradigm for running batch and streaming data processing pipelines. Google Data Fusion makes it easier to utilize these products by giving a simple graphical interface. Users may design data pipelines by dragging and dropping different components onto a canvas and connecting them together. Sources and sinks for various data formats, transformations for cleaning and altering data, and connectors for interacting with other data storage and processing systems are among the components.

Apache Beam is used to execute pipelines created using Data Fusion. As a result, the pipeline may be run in both batch and streaming modes, allowing it to analyze data in near real-time. Scalability is also provided through the usage of Apache Beam, since pipelines may be scaled up or down as needed to manage variable data volumes.

Overall, Google Data Fusion is a robust and easy-to-use tool for developing and maintaining data pipelines. Its use of industry-standard technologies such as Apache NiFi and Apache Beam allows integration with other systems simple and assures that pipelines may be conducted in a scalable and efficient way.

# Results and Discussion

## NBA Channel Engagement Metric Trends

We utilized Google Data Fusion to collect and analyze engagement measure data from the NBA's official YouTube channel over a six-month period for this study. We gathered information on the number of views, likes, comments, and shares received by each video submitted during this time period. Using this information, we examined the trends in engagement metrics and discovered some intriguing patterns.

## Results

According to our findings, the most popular videos on the NBA's YouTube account contained highlights of popular players and exciting games. We discovered that videos starring LeBron James, Steph Curry, and Kobe Bryant had millions of views apiece. In terms of interaction metrics, we saw that films with the most likes also had the most views and comments. However, no clear relationship was found between the number of shares and other engagement metrics.

## Discussion

According to our data, NBA fans are very engaged with material highlighting their favorite players and exciting games. Given the popularity of basketball and the NBA, this is not unexpected. Our research, however, emphasizes the necessity of producing high-quality material that is aesthetically appealing and engaging for visitors. This is especially crucial in YouTube's competitive ecosystem, where users have a plethora of material to pick from.

In terms of engagement measures, our research found that likes are a good predictor of overall video engagement. This implies that generating material that connects with viewers and motivates them to interact with the video through likes and comments is critical to developing a great YouTube presence.

Our investigation revealed an intriguing lack of association between the number of shares and other engagement indicators. While shares are commonly regarded as an important metric for determining the success of social media content, our research suggests that they may be less important on YouTube than on other platforms such as Twitter or Facebook. This might be because sharing on YouTube is less popular than on other social media sites.

Overall, our research gives useful insights into NBA fans' YouTube interaction habits. Understanding these tendencies allows the NBA to develop more interesting and effective content that connects with its viewers and contributes to a greater YouTube presence. Furthermore, our use of Google Data Fusion demonstrates the utility of this tool for collecting and analyzing engagement metric data on YouTube, and it could serve as a model for future research in this area.

## Limitations and Future Directions

While our study provides valuable insights into NBA fans' YouTube engagement patterns, there are some limitations to our analysis that should be considered. For starters, our research focused solely on interaction numbers from the NBA's official YouTube channel. It is likely that interaction patterns on other channels or social media sites would differ. Future research might look at engagement patterns across many mediums to get a more complete picture of fan interaction with NBA content.

The comparatively short term of six months is another shortcoming of our investigation. It is likely that engagement patterns will alter over time. Future research might look at engagement patterns over time to observe how they evolve and alter.

Finally, our study did not account for external factors that could influence engagement patterns, such as team or player performance, or cultural events. Future research could look into how outside factors influence engagement patterns on YouTube and other social media platforms.

# Lessons learned

During the course of the project, we faced several challenges that required creative solutions and workarounds to overcome. These included:

## Learning curve

Some of the lessons learnt and difficulties faced during this endeavor. For example, we had to ensure that the data was scraped accurately and effectively. We also had to ensure that we were utilizing and understanding the relevant metrics. However, we were able to learn a lot and improve our analysis as a result of these challenges.

The value of data quality and accuracy. We had to be certain that we were scraping and processing data correctly. We must also be cognizant of biases in our research and avoid drawing conclusions without supporting evidence.

## Data Integration

Data integration may be complicated and time-consuming, and extracting and transforming data from different sources and formats can be difficult. We tackled this issue by carefully organizing the data import and cleaning procedure and employing Data Fusion's built-in plugins for data wrangling, joining, and distinct operations.

* Complexity: While Data Fusion is a powerful data integration tool, it can be challenging to set up and configure for those who are new with the platform.
* Limited customization: Although Data Fusion offers a variety of pre-built connections and transformations, customers may require additional transformations or connectors that are not readily available.
* Limited support for specific data sources: Although Data Fusion supports a wide range of data sources, some data sources may not be supported, forcing users to seek alternative solutions.
* Limited infrastructure control: Due to the fact that Data Fusion is a managed service, consumers have no control over the underlying infrastructure. This may be tough for those that want customized settings or optimizations.
* Cost: While Data Fusion is a low-cost choice for many data integration needs, it can be costly for those with large or complex data integration requirements.

## Security

When working with sensitive data, managing data security and regulatory compliance may be a significant challenge. As a team, we took great effort to protect the security and privacy of the data used in our project. We adopted RBAC laws and followed best practices for data security, such as encrypting data at rest and in transit and restricting access to authorized personnel. Role creep occurs when someone gain too many roles over time, producing misunderstanding regarding a user's rights and responsibilities. Role explosion happens when there are too many roles and managing and allocating them becomes challenging. Roles can inherit permissions from other occupations in some cases, making it difficult to tell who has access to what. naming conventions for roles: Poorly named roles can generate confusion and make it difficult to determine which duties are appropriate for certain users.

Overprivileged roles, When users have access to more than what is necessary for their job activities, this might be a security risk. Inadequate oversight: Without good oversight, it may be difficult to discover when people are given responsibilities they do not deserve, potentially leading to security breaches.

## Cost Management

GCP provides a wide range of services, and it can be difficult to efficiently monitor the expenses of employing these services. We resolved this difficulty by closely monitoring and optimizing our use of the various GCP services to save costs.

Our team received great experience working with GCP and Data Fusion through this project, including the necessity of safe planning, extensive training, and successful collaboration. These lessons may be used to ensure the success and seamless execution of future projects.

# Undercurrents

## Security

Our project prioritized data security and protected the integrity of the data flow and analysis tools. The project used secure data storage options such as BigQuery and ensured that the API key for the YouTube API was kept confidential.

## Data Management:

## A data pipeline was utilized in our project to handle and process data from diverse sources, ensuring that the data was processed and cleansed before being fed into the database. Looker data studio was also utilized in the project to view and manage the data, as well as to generate reports for data analysts and data scientists.

## DataOps:

In our project, DataOps ideas were used to automate the data pipeline and ensure that data was handled effectively and correctly. To clean and improve the data, Wrangler, GroupBy, Distinct, Deduplicate, and Error Collector were used.

## Data Architecture:

## Our team researched data architecture when constructing the data pipeline to guarantee that the data is retained in a format that allows for easy access and analysis. BigQuery was also used in the project as a cloud-based data warehouse, which provides a scalable and cost-effective option for data storage..

## Orchestration:

In our project, an orchestration tool was used to manage the many stages of the data pipeline, ensuring that data is handled in a timely and effective manner. In the project, Data Fusion was used as an orchestration tool, providing a visual interface for designing and maintaining data pipelines.

## Software Engineering:

## We used software engineering ideas like code reusability, modularity, and version control to ensure that the data pipeline and analytic algorithms are efficient, maintainable, and scalable. The project used Python programs such as pandas, numpy, and googleapiclient to scrape and analyze data from the YouTube API. The project also made use of Jupyter Notebook to create and test the analytic algorithms.

# Future Work

Future research might expand on our findings by looking at engagement patterns across numerous media and over longer time periods, as well as investigating the influence of external factors like team performance and cultural events on engagement measures. Overall, our research adds to our understanding of social media engagement indicators and their potential application in influencing content production and marketing strategies in the sports business.

# Conclusion

We presented a study on tracking NBA channel engagement metric trends using web scraping and YouTube API analysis in this paper. Over a six-month period, we examined engagement metrics such as views, likes, and comments on the NBA's official YouTube channel. We discovered that material with popular players and thrilling games was the most seen and interacted with, and that likes were a strong measure of total involvement.

Our research adds to the expanding body of knowledge on social media engagement measures and their potential application in influencing content creation and marketing tactics. However, our study has limitations, including a short timeframe and a focus on only one channel.

Our findings imply that providing high-quality, engaging material that connects with viewers is crucial to establishing a strong presence on social media sites like YouTube. In addition, our research illustrates the potential utility of Google Data Fusion for gathering and evaluating engagement indicator data.

# References

1. “YouTube Analytics Using Google Data Studio” Krishna Bhargava Anantharamaiah
2. Cunningham, S., Craig, D., & Silver, J. (2016). YouTube, multichannel networks and the accelerated evolution of the new screen ecology. Convergence, 22(4), 376–391.
3. “Cloud Data Fusion |  Google Cloud,” Google Cloud. <https://cloud.google.com/data-fusion>
4. “Cloud Storage  |  Google Cloud,” Google Cloud. <https://cloud.google.com/storage>
5. “BigQuery Enterprise Data Warehouse  |  Google Cloud,” Google Cloud. <https://cloud.google.com/bigquery>
6. Qwiklabs, “Building Transformations and Preparing Data with Wrangler in Cloud Data Fusion | Google Cloud Skills Boost,” Qwiklabs. <https://www.cloudskillsboost.google/focuses/12364?parent=catalog>
7. Datadice, “The Wrangler User Interface of Cloud Data Fusion - Nerd For Tech - Medium,” Medium, Nov. 14, 2022. [Online]. Available: <https://medium.com/nerd-for-tech/the-wrangler-user-interface-of-cloud-data-fusion-21e96cfe1b2d>
8. “Architecture and components,” Google Cloud, [Online]. Available: <https://cloud.google.com/data-fusion/docs/concepts/architecture>
9. Tackett, Sean MD, MPH; Slinn, Kyle RN, MEd; Marshall, Tanner MS; Gaglani, Shiv MBA; Waldman, Vincent PhD; Desai, Rishi MD, MPH. “Medical Education Videos for the World: An Analysis of Viewing Patterns for a YouTube Channel.”
10. Manko, B. A. (2023). Video advertising: Using YouTube analytics for the target audience. Journal of Information Technology Teaching Cases, 13(1), 77–81.
11. Mike Thelwall (2018) Social media analytics for YouTube comments: potential and limitations, International Journal of Social Research Methodology, 21:3, 303-316, DOI: 10.1080/13645579.2017.1381821

# Appendices

1. Sample code snippets for web scraping and YouTube API analysisGraphical user interface, text

   Description automatically generated

Figure 1: code snippet of function to get channel statistics.

A picture containing text

Description automatically generatedFigure 2: code snippet of function to get video details.

Graphical user interface

Description automatically generated

Figure 3: code snippet of function to get video Ids.

1. Supplementary tables and figures

Chart, bar chart

Description automatically generated**Chart, bar chart

Description automatically generatedChart, bar chart

Description automatically generated**

Figure 4,5,6: graphs plotting channel’s Subscriber’s, Total video’s & view’s

Graphical user interface, application

Description automatically generatedChart, bar chart

Description automatically generated

Figure 7,8 : graph shows the most liked videos

Graphical user interface, application, Word

Description automatically generated Figure 9: Data architecture of our pipeline

Diagram

Description automatically generated Figure 10:our Data pipeline