Mini Project Report *on*

"YouTube Data Analysis"

Submitted by

Sumedh Kamble	S1032180007	PC01
Shreyan Yoge	S1032180058	PC05
Shriya Padhi	S1032181174	PD25
Taha Bohra	S1032181349	PC58

Under the Guidance of

Prof. Vaishali Suryawanshi

At

MIT-WPU
॥ विश्वशान्तिर्ध्रुवं ध्रुवा ॥

Dr. Vishwanath Karad
MIT WORLD PEACE UNIVERSITY | PUNE
TECHNOLOGY, RESEARCH, SOCIAL INNOVATION & PARTNERSHIPS

School of Computer Engineering and Technology

Contents

Abstract ……..….. I

List of Abbreviations……...…... II

List of Figures…….. III

List of Tables…… IV

1. Introduction..……..……..……. 1

2. Motivation. ……..……..…… 12

3. Problem Definition ……..…… 12

4. Problem Statement.……..…… 12

5. Objectives ……....……..…… 12

6. **Tools used ..……..……..……. 12**
6.1 Hadoop Distributed File System ……..…….. 12
6.1 Hive ...…….…….. 12
6.3 Power BI …..…….…….. 12

7. Dataset description……..……. 12

8. System architecture...……..……. 12

9. Data analysis……..……..……. 12

10. Output……..……..……..……. 12

11. Visualization screenshots ……..……. 12

12. Conclusion..……..……..……. 12

13. References..……..……..……. 12

Abstract

YouTube is a free video sharing website that makes it easy to watch online videos. You can even create and upload your own videos to share with others. We live today in a digital world a tremendous amount of data is generated by each digital service we use. This vast amount of data generated is called Big Data. YouTube is one of the best examples of services that produce a massive amount of data in a brief period. We demonstrate how we can extract insightful information from YouTube dataset using Big Data Analytics. Data extraction of a significant amount of data is done using Hadoop and MapReduce to measure performance. Hadoop is a system that offers consistent memory. Storage is provided by HDFS (Hadoop Distributed File System) and MapReduce analysis. MapReduce is a programming model and a corresponding implementation for processing large data sets. Apache Hive is a data warehouse software project built on top of Apache Hadoop for providing data query and analysis. Hive gives an SQL-like interface to query data stored in various databases and file systems that integrate with Hadoop. Power BI is a business analytics service by Microsoft. It aims to provide interactive visualizations and business intelligence capabilities with an interface simple enough for end users to create their own reports and dashboards. It is part of the Microsoft Power Platform. These are the technologies used for the same.

List of Abbreviations

BDA Big Data Analytics

BI Business Intelligence

HDFS Hadoop Distributed File System

POSIX Portable Operating System Interface

API Application Programming Interface

SQL Structured Query Language

FINRA Financial Industry Regulatory Authority

UI User Interface

CSV Comma Separated Values

List of Figures

Figure 1: Hadoop Distributed File System Architecture……..….. I
Figure 2: Hive System Architecture…….....….. I
Figure 3: Power BI System Architecture…….....….. I
Figure 4: YouTube Dataset……..... I
Figure 5: System Architecture for YouTube Dataset Analysis …..……..….. I
Figure 6: Hive Query: Create Table……..….. I
Figure 7: Hive Query: Display count of total videos……....….. I
Figure 8: Hive Query: Display count of total videos with ratings more than 4.7 .I
Figure 9: Hive Query: Analyze video length……..……..... I
Figure 10: Hive Query: Maximum views based on category……..……. I
Figure 11: Hive Query: Minimum views based on category...…… I
Figure 12: Hive Query: Analyze video engagement……...….. I
Figure 13: Hive Query: Top 10 video uploaders……..…...….. I
Figure 14: Hive Query: Top 10 latest videos……...….. I
Figure 15: Hive Query: Count of videos with category comedy……..……...….. I
Figure 16: Hive Query: Count of comments on EvilSquirrelPictures channel…. I
Figure 17: Data visualization: Number of comments by category……..….. I
Figure 18: Data visualization: Number of ratings by category ……...….. I
Figure 19: Data visualization: Time interval by category……...….. I
Figure 20: Data visualization: Video length by category …….……...….. I
Figure 21: Data visualization: Number of views by category……..……...….. I
Figure 22: Data visualization: Average of ratings by category…….…...….. I
Figure 23: Data visualization: Number of comments by uploader name…....…... I
Figure 24: Data visualization: Average of ratings by uploader name….....….. I
Figure 25: Data visualization: Time interval by uploader name.....….. I
Figure 26: Data visualization: Average video length by uploader name.....….. I
Figure 27: Data visualization: Number of views by uploader name…….....….. I
Figure 28: Data visualization: Ratings by uploader name ….....….. I

List of Tables

Table 1: Hadoop Distributed File System Pros and Cons.....….. I

Table 2: Hive Pros and Cons….....…… I

Table 3: Power BI Pros and Cons.....….. I

1. Introduction

Entertainment has become a necessity of life for us and there is enough content today to keep us engaged for every moment of the rest of our lives. So one thing we can be sure of is that no one is going to die of boredom in the near foreseeable future. The internet and the rise of online streaming platforms have paved the way for a new golden age of Television. It may come as a surprise, but there is a good possibility that old reliable T.V. could be permanently replaced by the modern marvels of the internet.

These platforms are both proliferating into the global market as well as evolving in terms of technology. To begin looking at what big data has been doing in the entertainment industry we need to first have a look at one of the earliest examples of online mainstream entertainment- YouTube.

2. Motivation

Online Entertainment as we perceive it today was born in 2005 when YouTube came to be. The online journey of mass-produced entertainment thus started. YouTube has always been a free forum for everyone to upload whatever they deemed fit to be called entertainment. Yet, the amount of popularity it has amassed is astounding.

The key here is audience engagement and Big Data is the magic tool that facilitates this. Big Data offers valuable insights into audience's temperaments and their preferences which can further be used to strategize content creation. From marketing strategy to creating the content itself, there are a lot of aspects that are influenced deeply by Big Data.

3. Problem Definition

The YouTube platform is increasingly under competitive pressure to not only acquire customers but also understand their customers' needs to be able to optimize customer experience and develop long standing relationships. By sharing their data and allowing relaxed privacy in its use, customers expect YouTube to know them, form relevant interactions, and provide a seamless experience across all touch points.

Content is the life-blood of this organization. Our role is to recognize trends that drive strategic roadmap for innovation, new features, and services. Being able to react in real time and make the customer feel personally valued is only possible through advanced analytics.

4. Problem Statement:

Our aim is to generate insightful information to be able to anticipate customer needs, deliver relevant products, personalize and optimize customer experiences in YouTube by using Big Data Analytics.

5. Objectives:

1. To extract meaningful insights from the YouTube dataset.
2. To uncover hidden patterns present in YouTube data.
3. To extract unknown correlations between certain parameters.
4. To understand market trends and customer preferences to enhance YouTube customer base.

6. Tools used:

1. **Hadoop distributed file system:**

   The Hadoop distributed file system (HDFS) is a distributed, scalable, and portable file system written in Java for the Hadoop framework. Some consider it to instead be a data store due to its lack of POSIX compliance, but it does provide shell commands and Java application programming interface (API) methods that are similar to other file systems. A Hadoop instance is divided into HDFS and MapReduce. HDFS is used for storing the data and MapReduce is used for processing data.

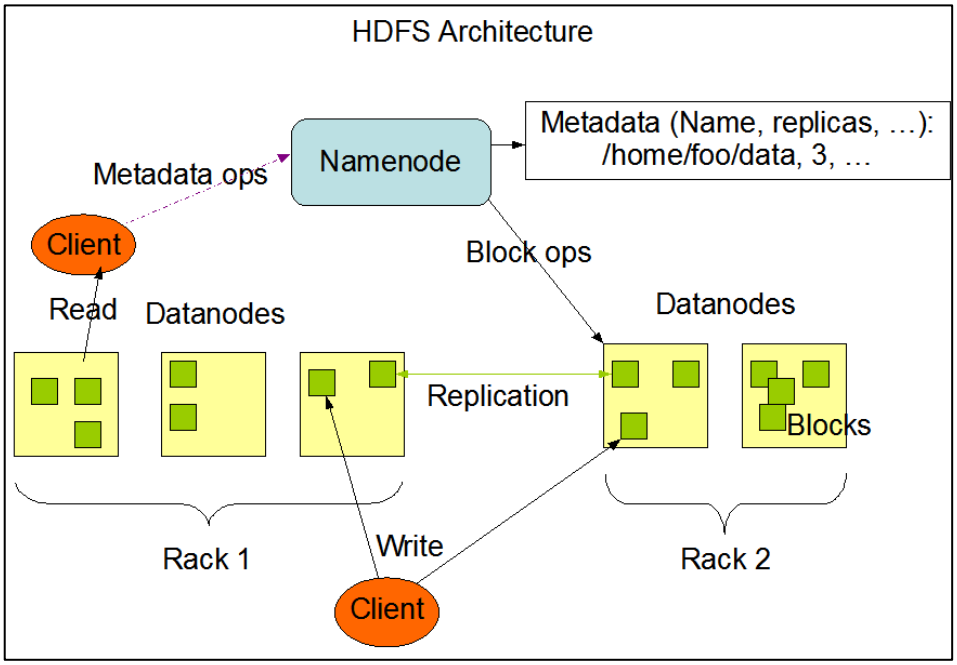


Figure 1: Hadoop Distributed File System Architecture

Pros	Cons
Scalable	Latency
Cost effective	Security
Compatible	Supports Batch Processing only
Easy to use	Latency
Varied data sources	No Real Time Data Processing

Table 1: Hadoop Distributed File System Pros and Cons

2. **Hive:**

Apache Hive is a data warehouse software project built on top of Apache Hadoop for providing data query and analysis. Hive gives an SQL-like interface to query data stored in various databases and file systems that integrate with Hadoop. Traditional SQL queries must be implemented in the MapReduce Java API to execute SQL applications and queries over distributed data. Hive provides the necessary SQL abstraction to integrate SQL-like queries (HiveQL) into the underlying Java without the need to implement queries in the low-level Java API. Since most data warehousing applications work with SQL-based querying languages, Hive aids portability of SQL-based applications to Hadoop. While initially developed by Facebook, Apache Hive is used and developed by other companies such as Netflix and the Financial Industry Regulatory Authority (FINRA). Amazon maintains a software fork of Apache Hive included in Amazon Elastic MapReduce on Amazon Web Services.

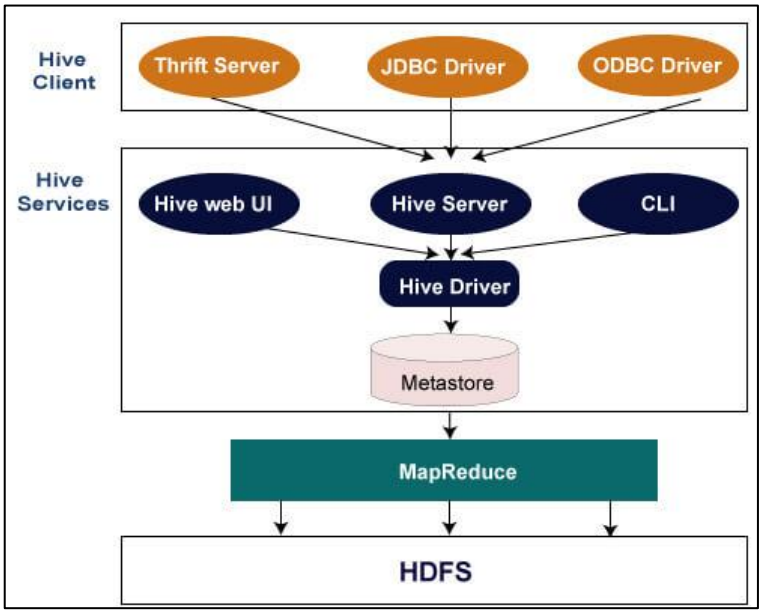


Figure 2: Hive System Architecture

Pros	Cons
Supports Hadoop	Not all standard SQL supported
MapReduce is easy	Limited built-in functions

Table 2: Hive Pros and Cons

3. **Power BI:**

Power BI is a business analytics service by Microsoft. It aims to provide interactive visualizations and business intelligence capabilities with an interface simple enough for end users to create their own reports and dashboards. It is part of the Microsoft Power Platform.

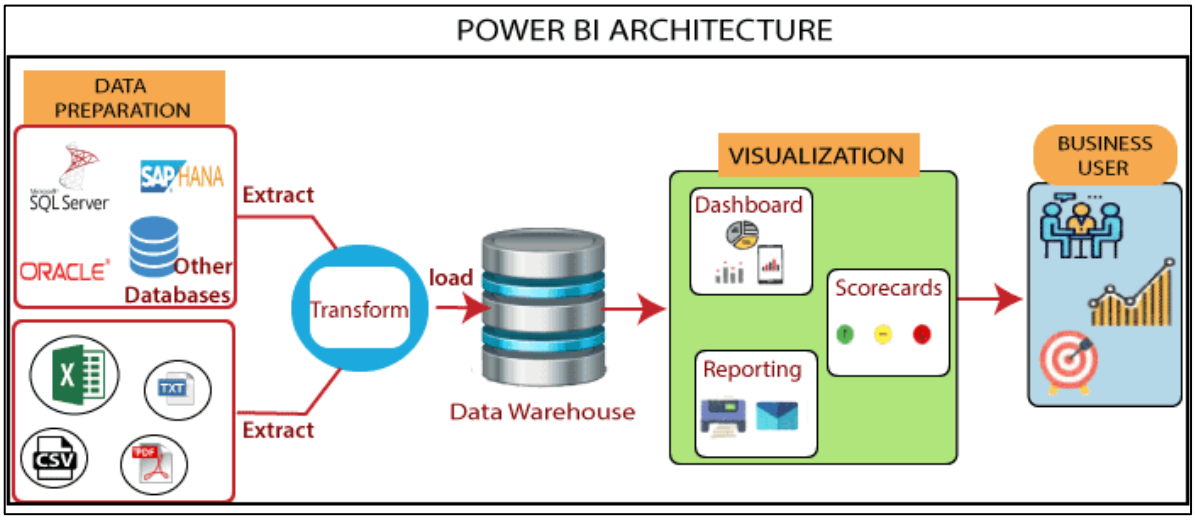


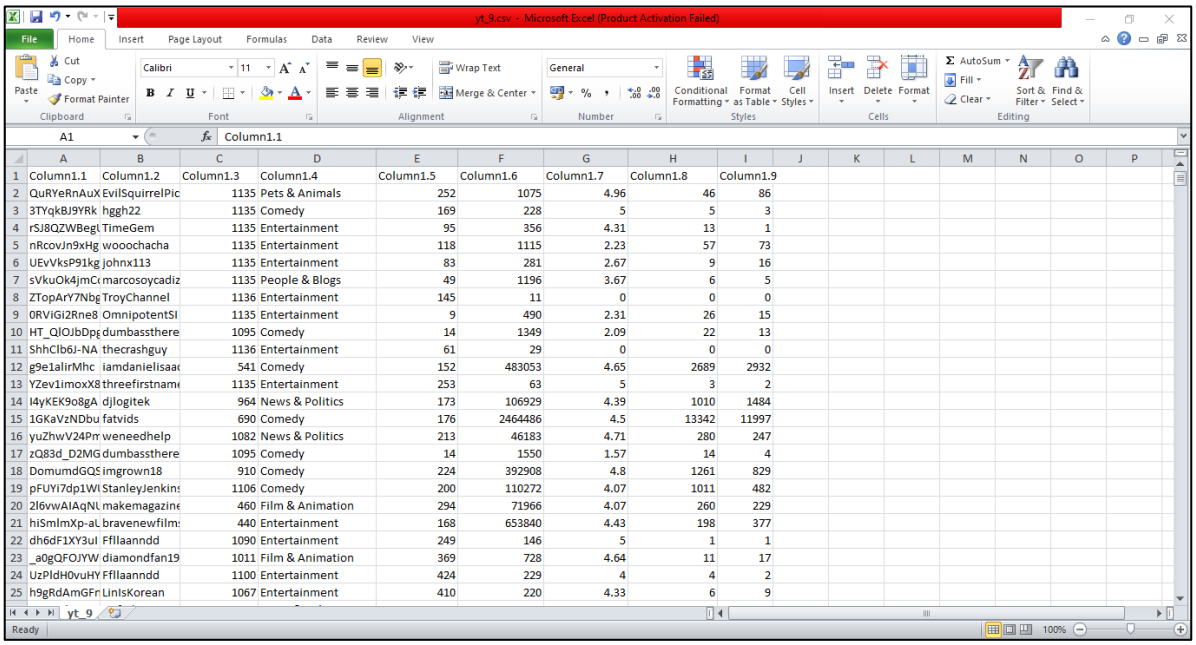
Figure 3: Power BI System Architecture

Pros	Cons
Range of custom visualizations	Difficult to master
Cost effective	Crowded UI
High Data connectivity	Rigid Formulas
Connected with Excel	Limited Data Handling in free version

Table 3: Power BI Pros and Cons

7. Dataset description:

The dataset is named as yt9.csv and is a YouTube dataset. It consists of 9 columns and 4100 rows. The attributes are video id, channel name, time interval when video was uploaded, video category, video length, number of views, ratings, number of ratings and number of comments. These attributes contribute to carry out analytics so as to obtain insightful information.



	A	B	C	D	E	F	G	H	I
1	Column1.1	Column1.2	Column1.3	Column1.4	Column1.5	Column1.6	Column1.7	Column1.8	Column1.9
2	QuRYeRnAuX	EvilSquirrelPic	1135	Pets & Animals	252	1075	4.96	46	86
3	3TYqkBJ9YRk	hggh22	1135	Comedy	169	228	5	5	3
4	rSJ8QZWBegU	TimeGem	1135	Entertainment	95	356	4.31	13	1
5	nRcovJn9xHg	wooochacha	1135	Entertainment	118	1115	2.23	57	73
6	UEvVksP91kg	johnx113	1135	Entertainment	83	281	2.67	9	16
7	sVkuOk4jmC	marcosoycadiz	1135	People & Blogs	49	1196	3.67	6	5
8	ZTopArY7Nbg	TroyChannel	1136	Entertainment	145	11	0	0	0
9	0RViGi2Rne8	OmnipotentSI	1135	Entertainment	9	490	2.31	26	15
10	HT_QlOJbDp	dumbassthere	1095	Comedy	14	1349	2.09	22	13
11	ShhClb6J-NA	thecrashguy	1136	Entertainment	61	29	0	0	0
12	g9e1alirMhc	iamdanielisaac	541	Comedy	152	483053	4.65	2689	2932
13	YZev1imoxX8	threefirstnam	1135	Entertainment	253	63	5	3	2
14	I4yKEK9o8gA	djlogitek	964	News & Politics	173	106929	4.39	1010	1484
15	1GKaVzNDbu	fatvids	690	Comedy	176	2464486	4.5	13342	11997
16	yuZhwV24Pm	weneedhelp	1082	News & Politics	213	46183	4.71	280	247
17	zQ83d_D2MG	dumbassthere	1095	Comedy	14	1550	1.57	14	4
18	DomumdGQS	imgrown18	910	Comedy	224	392908	4.8	1261	829
19	pFUYi7dp1W	StanleyJenkins	1106	Comedy	200	110272	4.07	1011	482
20	2l6vwAIAqN	makemagazine	460	Film & Animation	294	71966	4.07	260	229
21	hiSmlmXp-a	bravenewfilms	440	Entertainment	168	653840	4.43	198	377
22	dh6dF1XY3uI	Ffllaanndd	1090	Entertainment	249	146	5	1	1
23	_a0gQFOJYW	diamondfan19	1011	Film & Animation	369	728	4.64	11	17
24	UzPldH0vuHY	Ffllaanndd	1100	Entertainment	424	229	4	4	2
25	h9gRdAmGFr	LinIsKorean	1067	Entertainment	410	220	4.33	6	9

Figure 4: YouTube Dataset

8. System Architecture:

The CSV file containing the YouTube dataset is uploaded on the Hadoop Distributed File System by running commands on Cloudera Command Prompt. Hive is used to run queries on the data to analyze it. For the same it uses Map Reduce technology. Lastly, the CSV file is used to generate data visualization using Power BI.

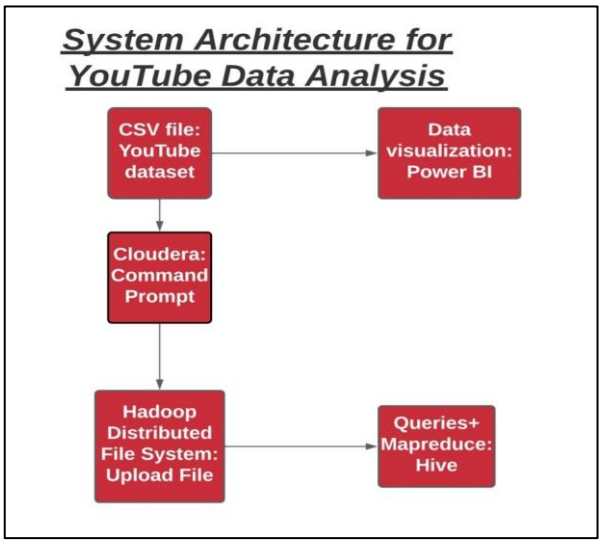


Figure 5: System Architecture for YouTube Dataset Analysis

9. Data analysis (Hive queries)

1. **Create table.**

```
create table ytdata(vid_id string, uploader_name int,
interval_fromytcreated_toupload int, catagory string, video_length
int, no_of_views int, ratings float, no_of_ratings int, no_of_comments
int) row format delimited fields terminated by ',' stored as textfile
location '/user/cloudera/ytdata';
```

2. **Display the count of total videos in database.**

```
select count(*) from ytdata;
```

3. **Display the count of videos with ratings greater than 4.7.**

```
select count(*) from ytdata where ratings > 4.7;
```

4. **Analyse category wise popularity of videos based on ratings.**

```
select category,avg(ratings) as average from ytdata group by category order by
average;
```

5. **Check if longer videos are more popular than shorter ones.**

```
select "longer_videos_rating",avg(ratings) as longer_videos_avg_ratings from
ytdata  where video_length > 200;
select "Shorter Videos Ratings",avg(ratings) as longer_videos_avg_ratings from
ytdata  where video_length < 200;
```

6. **Display category wise maximum views count.**

```
select  category ,max(no_of_views) as max_views  from ytdata group  by
category;
```

7. **Display category wise minimum views count.**

```
select  category ,min(no_of_views) as min_views  from ytdata group  by
category;
```

8. **Display top 3 engaging videos. It is measured as-**
   number of ratings + number of comments /number of views

```
select vid_id,( (no_of_ratings+no_of_comments)/no_of_views) as
engagement_value  from ytdata order by engagement_value desc limit 3 ;
```

9. **Display top 10 video uploaders.**

```
select uploader_name,count(*)as count from ytdata group by uploader_name
order by count desc limit 10;
```

10. Display 10 oldest videos uploaded

```
select vid_id,interval_fromytcreated_toupload from ytdata where
interval_fromytcreated_toupload > 0 order by interval_fromytcreated_toupload
limit 10;
```

11. Display 10 latest videos uploaded

```
select vid_id,interval_fromytcreated_toupload from ytdata where
interval_fromytcreated_toupload > 0 order by interval_fromytcreated_toupload
desc limit 10;
```

12. Display number of videos in comedy category.

```
select count(*) from ytdata where category='Comedy';
```

13. Display total comments on EvilSquirrelPictures channel.

```
select sum(no_of_comments) from ytdata where uploader_name =
'EvilSquirrelPictures';
```

10. Output

1. Create table.

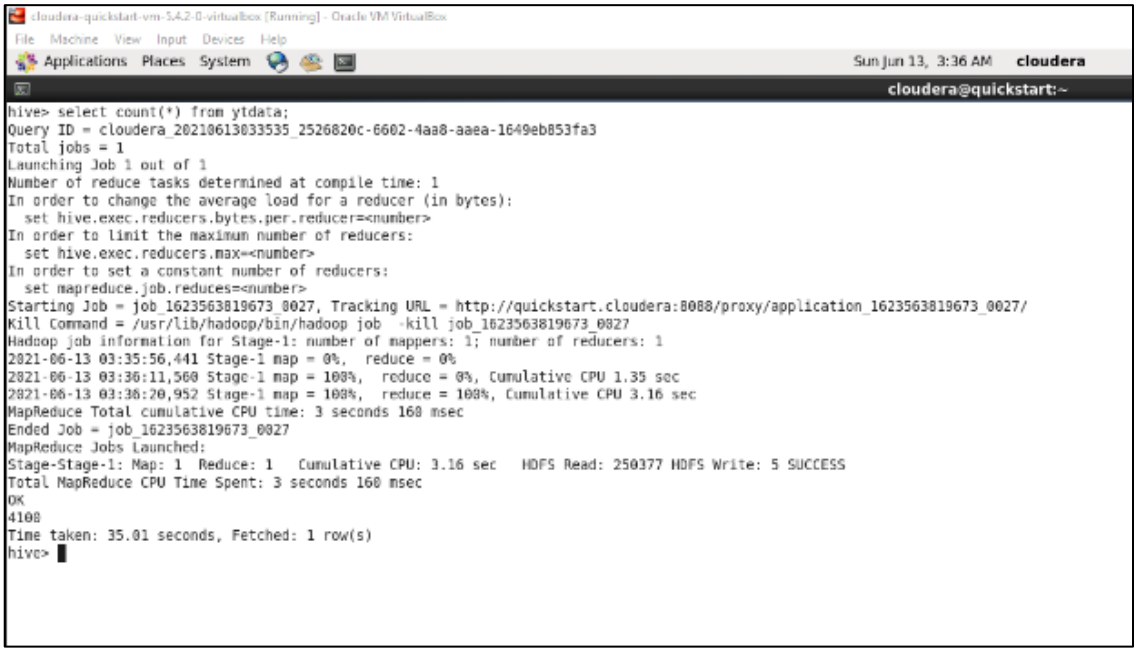


Figure 6: Hive Query: Create Table

2. Display the count of total videos in database.

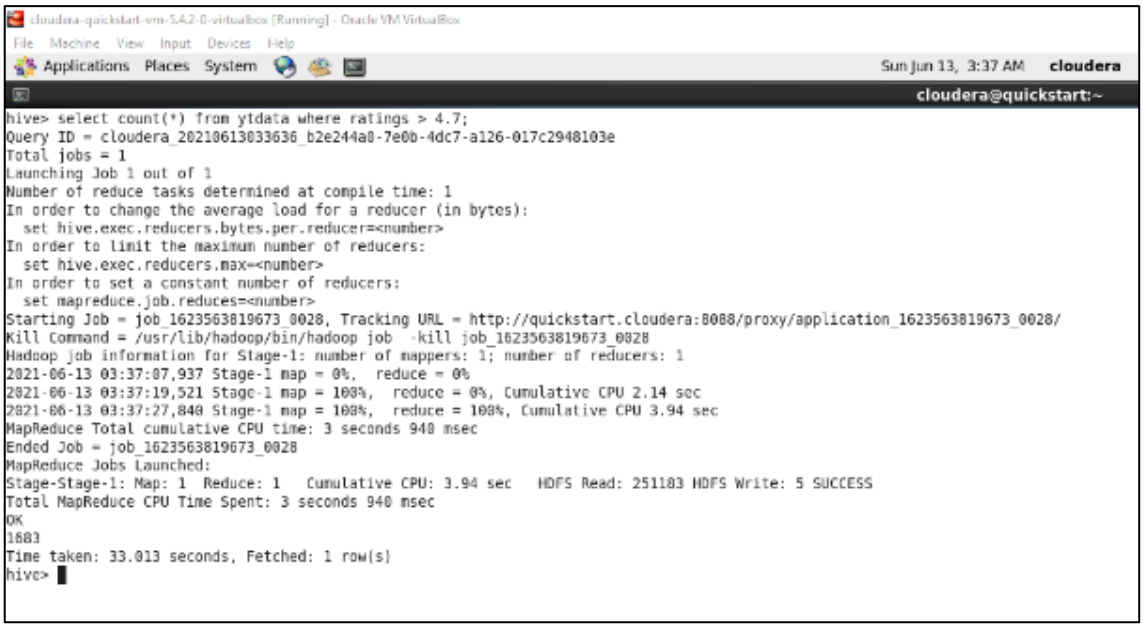


Figure 7: Hive Query: Display count of total videos

3. Display the count of videos with ratings greater than 4.7.

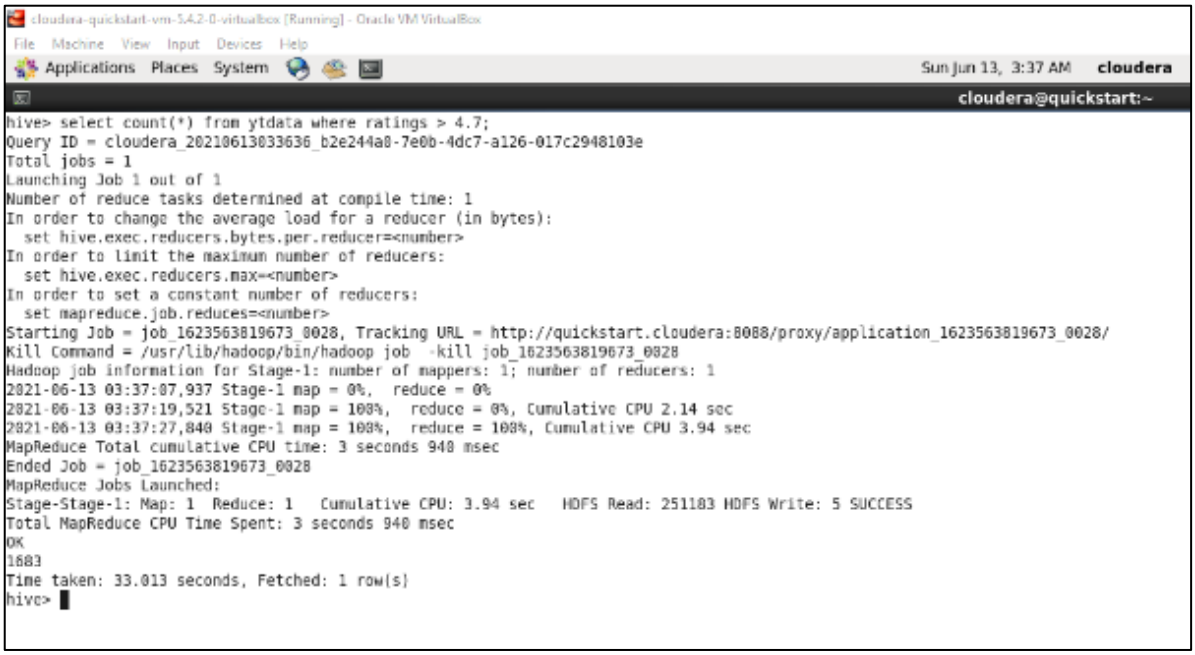


Figure 8: Hive Query: Display count of total videos with ratings more than 4.7

4. Analyze category wise popularity of videos based on ratings.

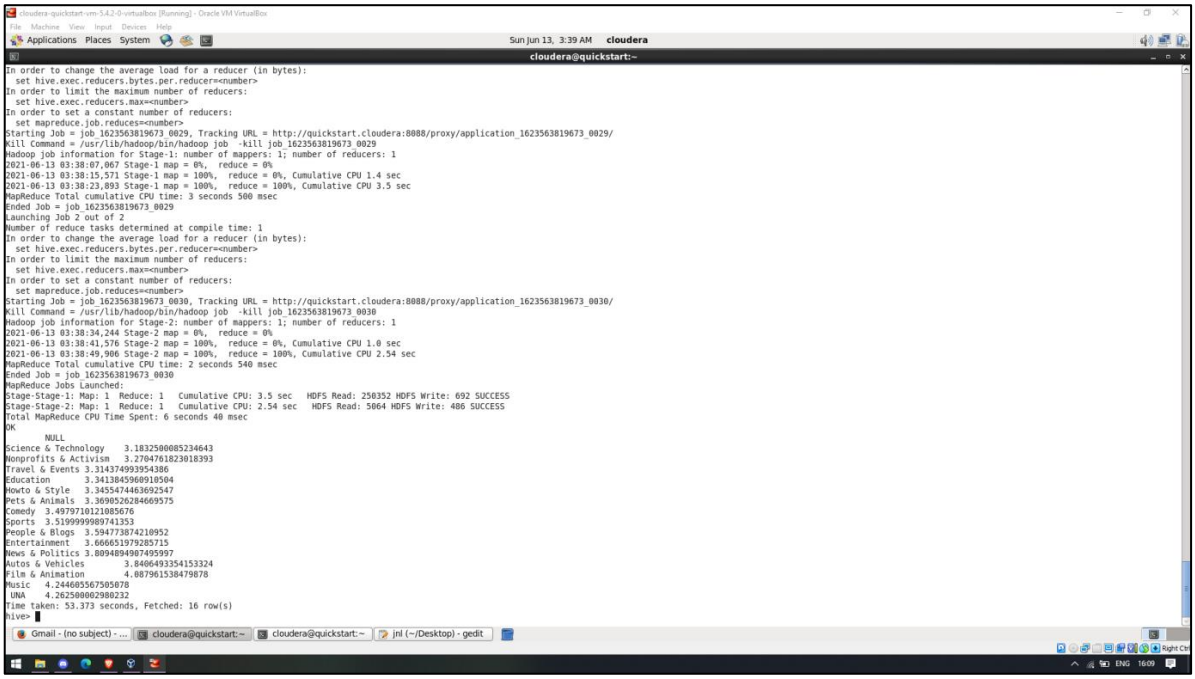


Figure 8: Hive Query: Analyze category wise video popularity

5. Check if longer videos are more popular than shorter ones.

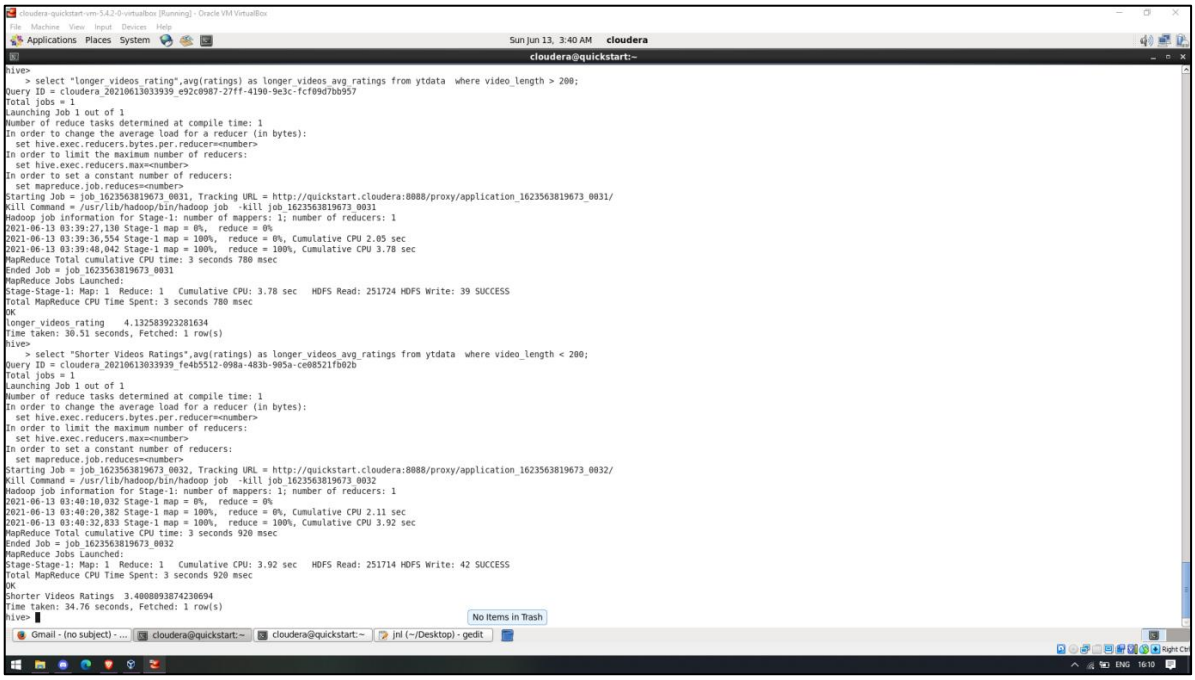


Figure 9: Hive Query: Analyze video length

6. Display category wise maximum views count.

```
hive> select  category ,max(no of views) as max views  from ytdata group  by category;
Query ID = cloudera_20210613034242_2e00c8ba-b039-4e20-99c3-1f140eea86e6
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1623563819673_0034, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1623563819673_0034/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1623563819673_0034
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2021-06-13 03:42:46,122 Stage-1 map = 0%,  reduce = 0%
2021-06-13 03:43:01,996 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.51 sec
2021-06-13 03:43:12,464 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 3.53 sec
MapReduce Total cumulative CPU time: 3 seconds 530 msec
Ended Job = job_1623563819673_0034
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 3.53 sec   HDFS Read: 250690 HDFS Write: 329 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 530 msec
OK
        NULL
 UNA    711384
Autos & Vehicles        1762369
Comedy  11807201
Education       1369522
Entertainment   18235463
Film & Animation        85341925
Howto & Style   6101232
Music   33754615
News & Politics 6824441
Nonprofits & Activism   80967
People & Blogs  5766247
Pets & Animals  27721690
Science & Technology    3234852
Sports  5360384
Travel & Events 3818296
Time taken: 47.407 seconds, Fetched: 16 row(s)
hive>
```

Figure 10: Hive Query: Maximum views based on category

7. **Display category wise minimum views count.**

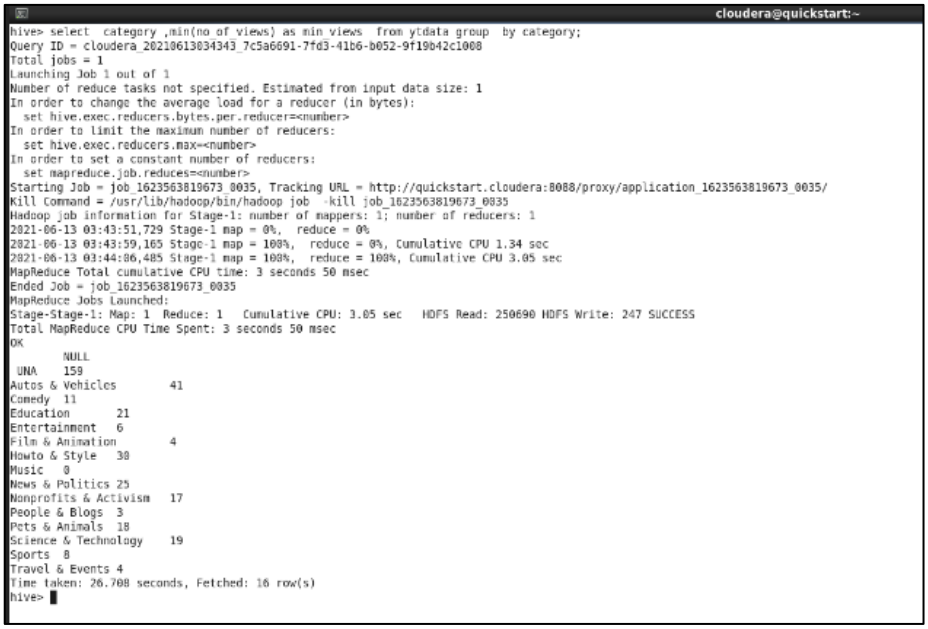


Figure 11: Hive Query: Minimum views based on category

8. **Display top 3 engaging videos. It is measured as-**
   number of ratings + number of comments /number of views

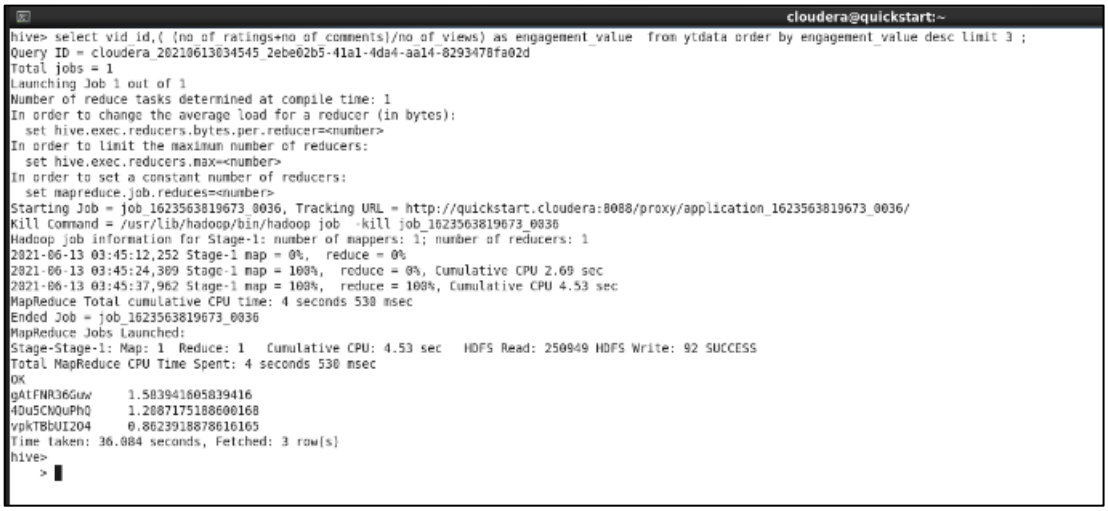


Figure 12: Hive Query: Analyze video engagement

9. Display top 10 video uploaders.

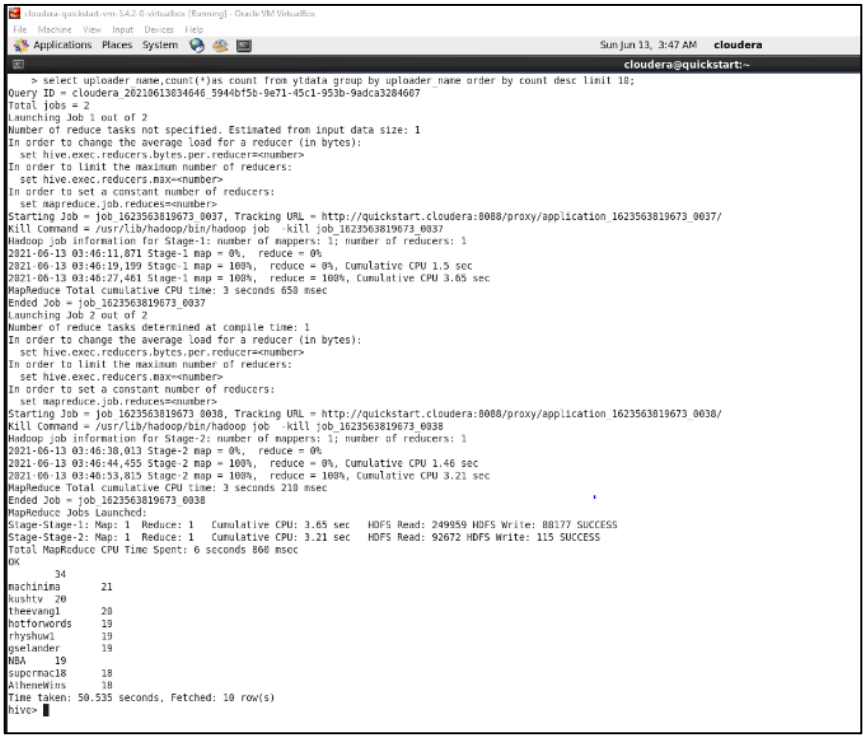


Figure 13: Hive Query: Top 10 video uploaders

10. Display 10 latest videos uploaded

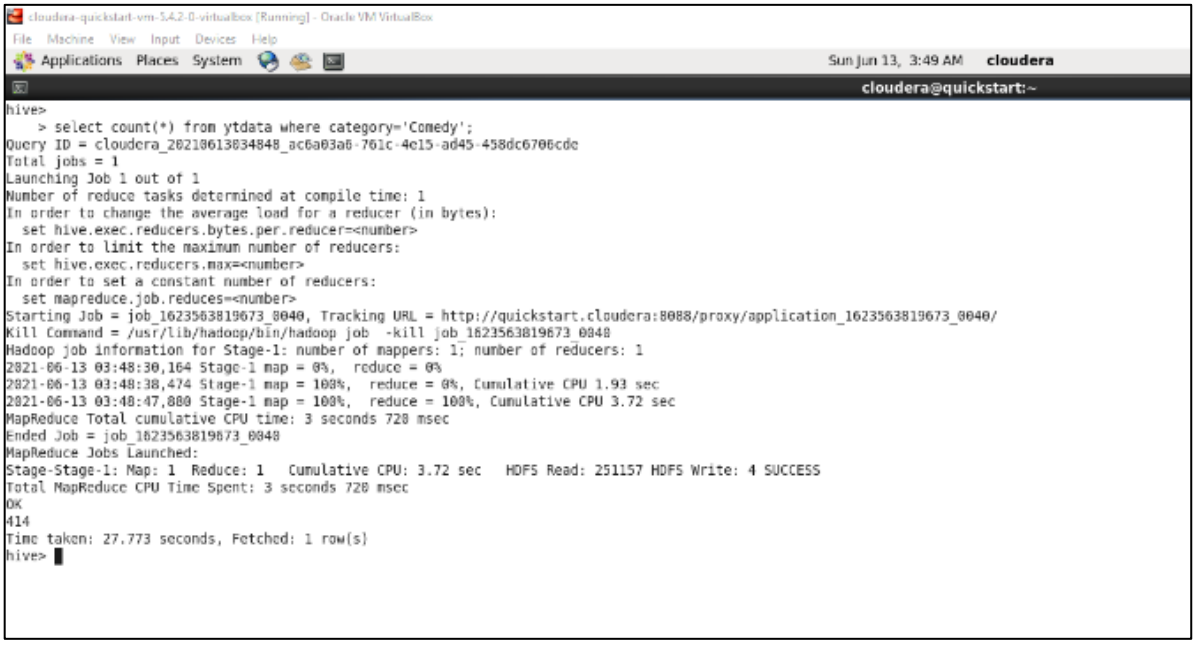


Figure 14: Hive Query: Top 10 latest videos

11. Display number of videos in comedy category.

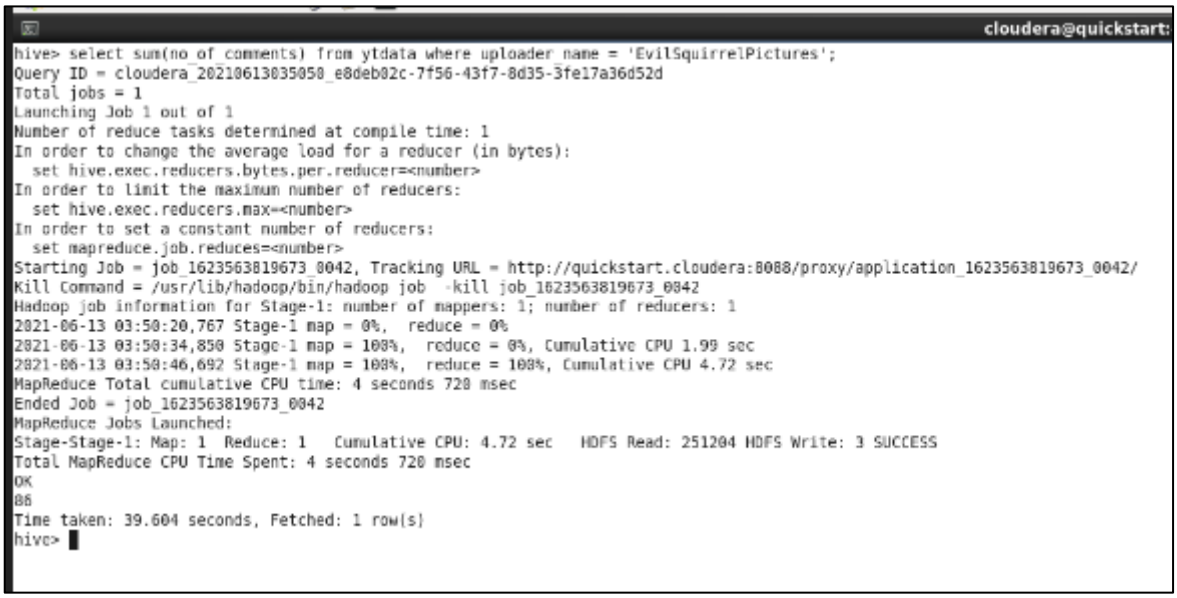


Figure 15: Hive Query: Count of videos with category comedy

12. Display total comments on EvilSquirrelPictures channel.

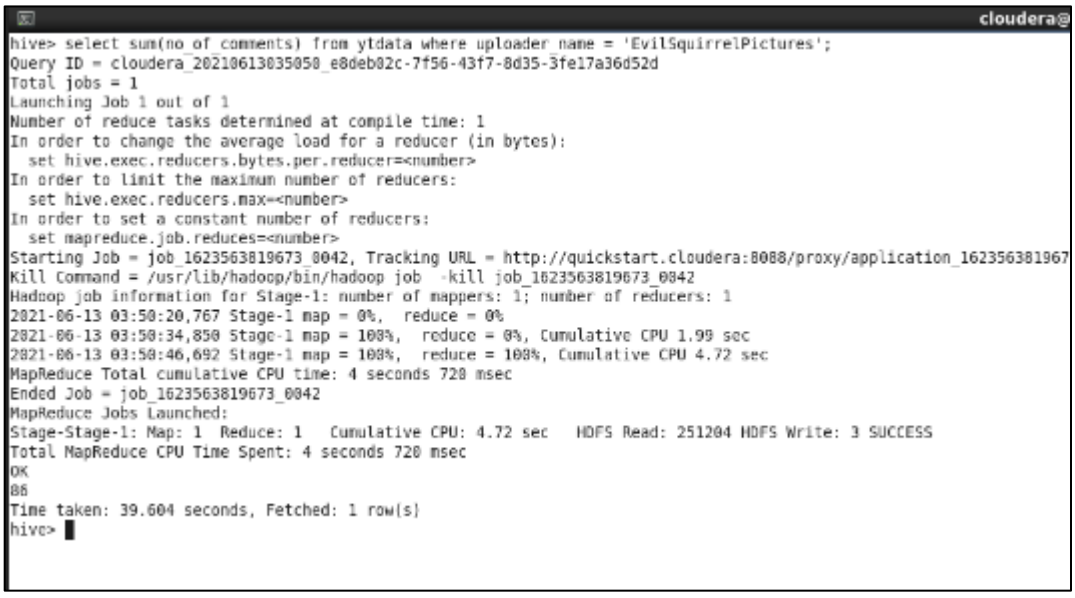


Figure 16: Hive Query: Count of comments on EvilSquirrelPictures channel

12. Visualization Screenshots

1. **Number of comments grouped by category.**

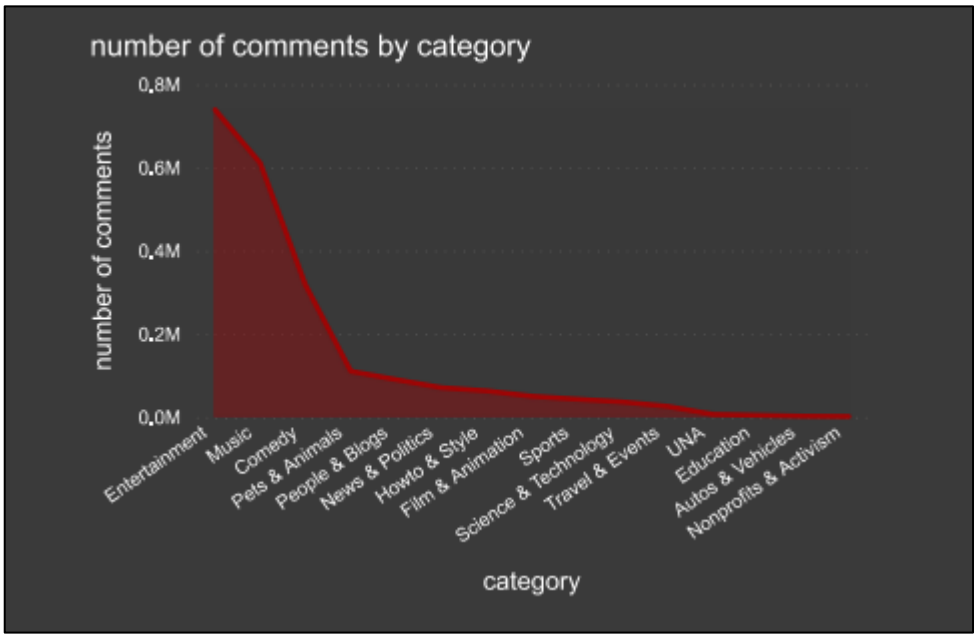


Figure 17: Data visualization: Number of comments by category

2. **Number of ratings grouped by category.**

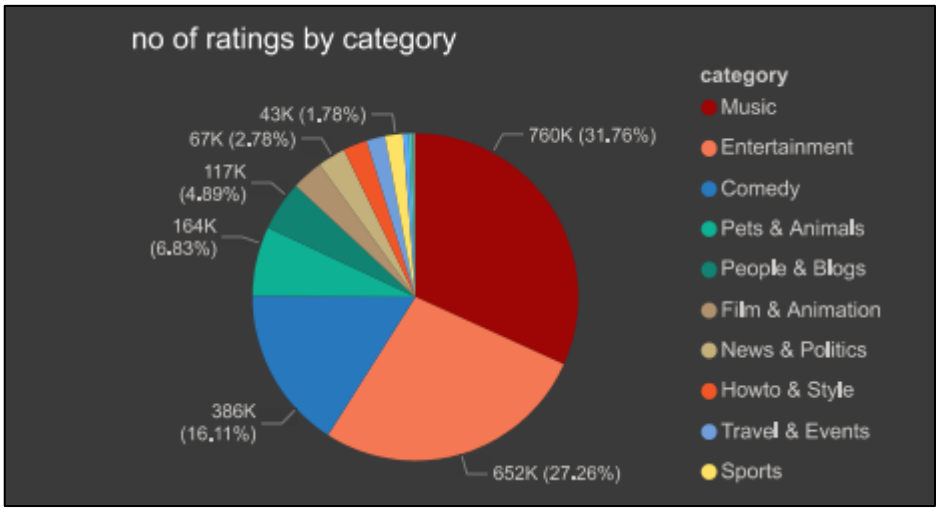


Figure 18: Data visualization: Number of ratings by category

3. **Time interval from when YouTube was created to when the video was uploaded grouped by category.**

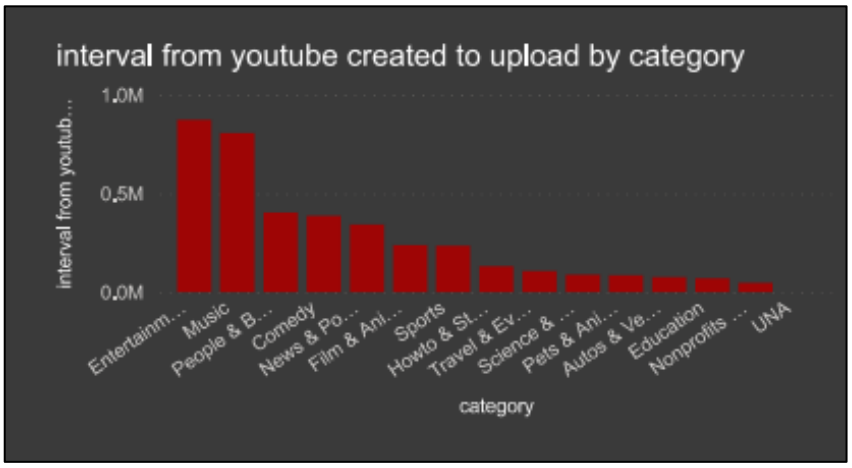


Figure 19: Data visualization: Time interval by category

4. **Video length grouped by category.**

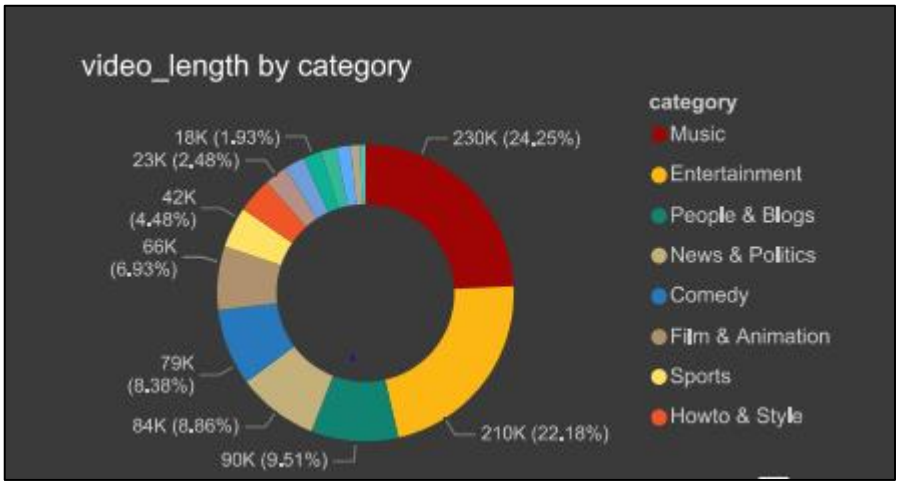


Figure 20: Data visualization: Video length by category

5. **Number of views grouped by category.**

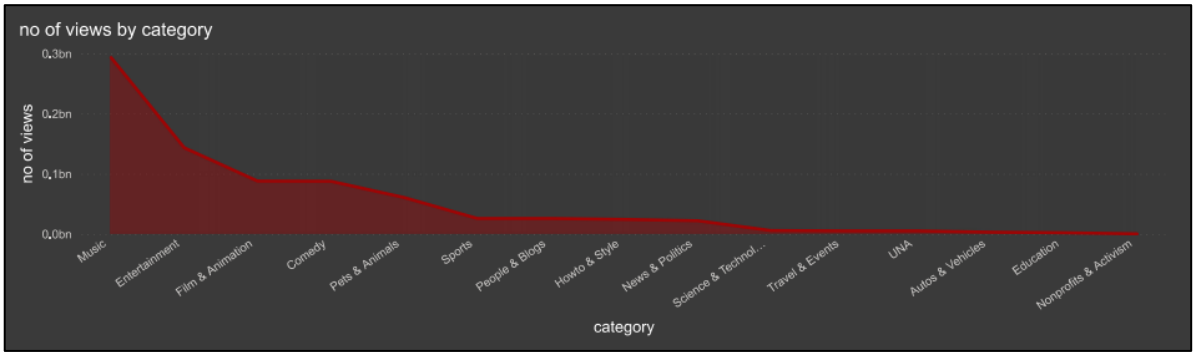


Figure 21: Data visualization: Number of views by category

6. **Average of ratings grouped by category.**

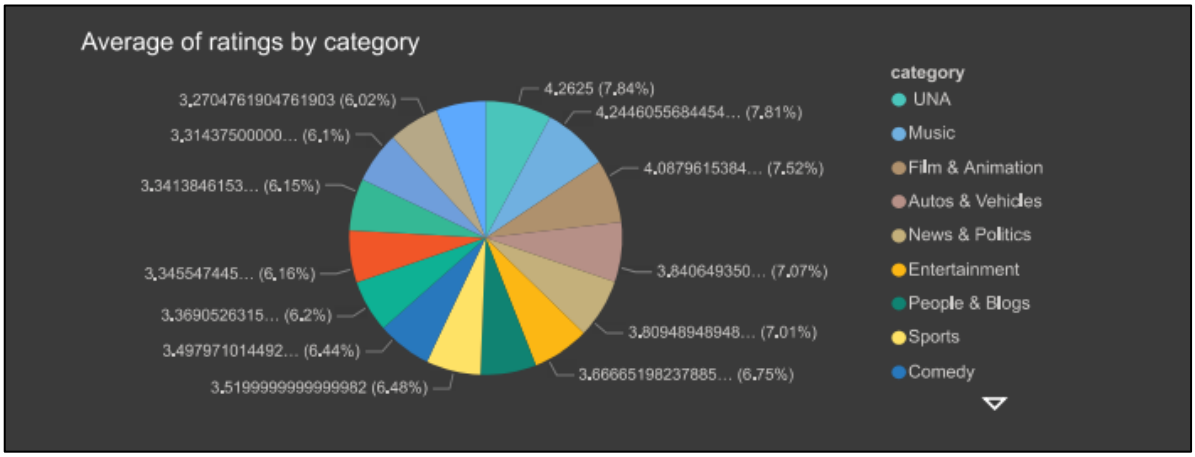


Figure 22: Data visualization: Average of ratings by category

7. **Number of comments grouped by uploader name.**

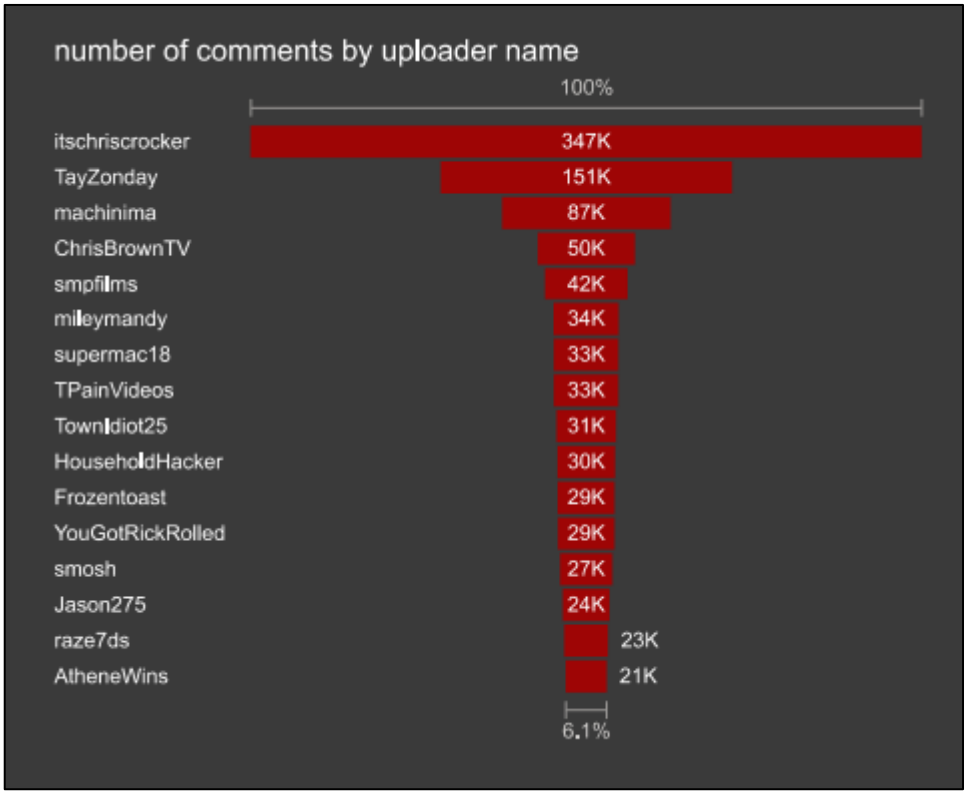


Figure 23: Data visualization: Number of comments by uploader name

8. **Average of ratings grouped by uploader name.**

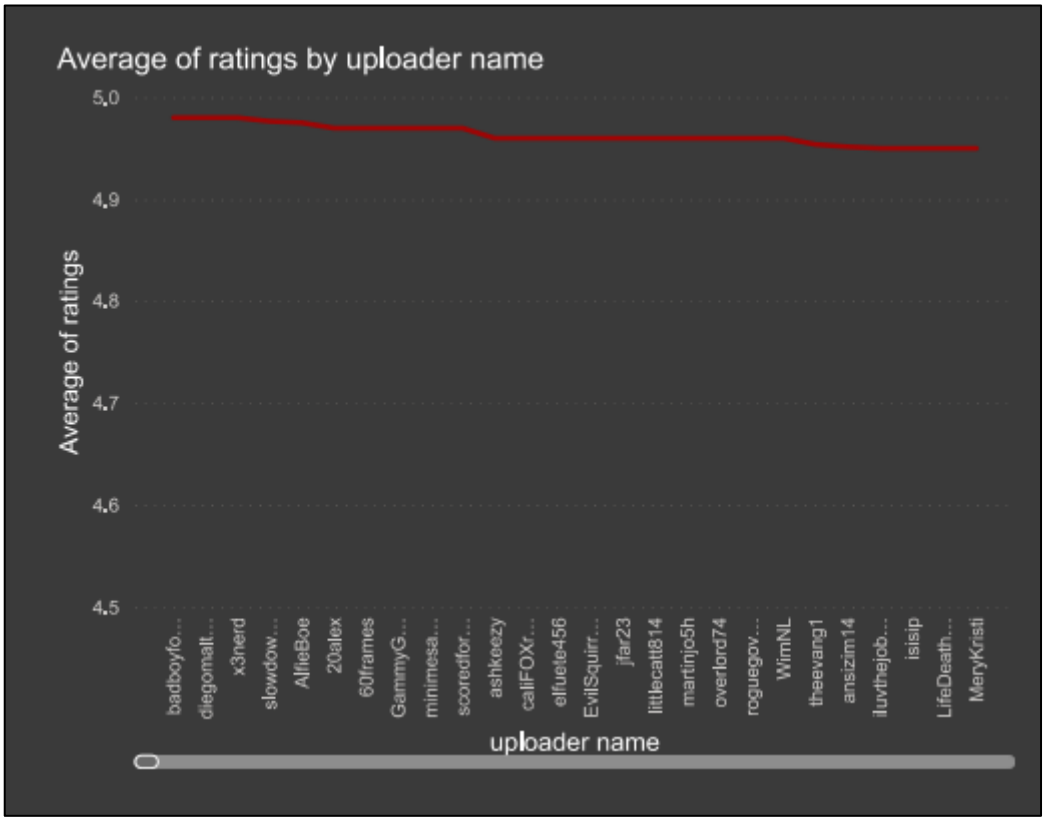


Figure 24: Data visualization: Average of ratings by uploader name

9. **Time interval from when YouTube was created to when the video was uploaded grouped by uploader name.**

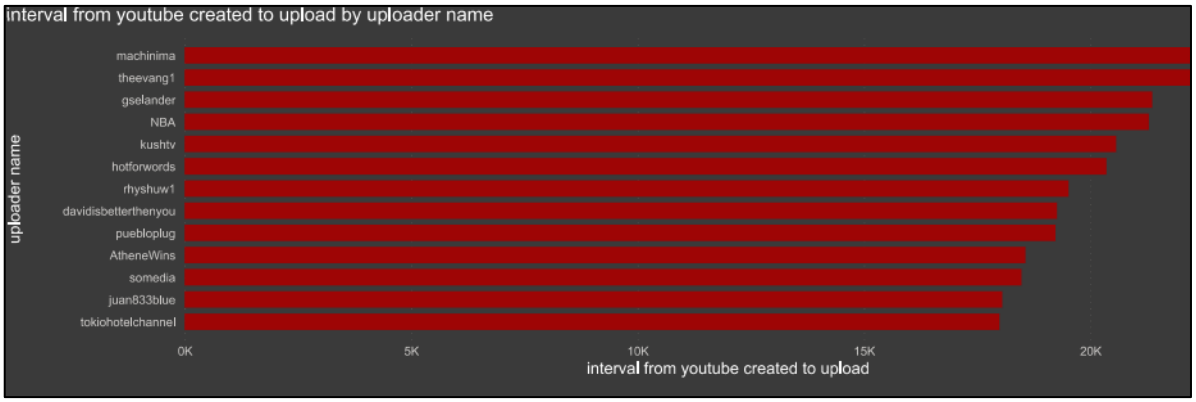


Figure 25: Data visualization: Time interval by uploader name

10. Average of video length grouped by uploader name.

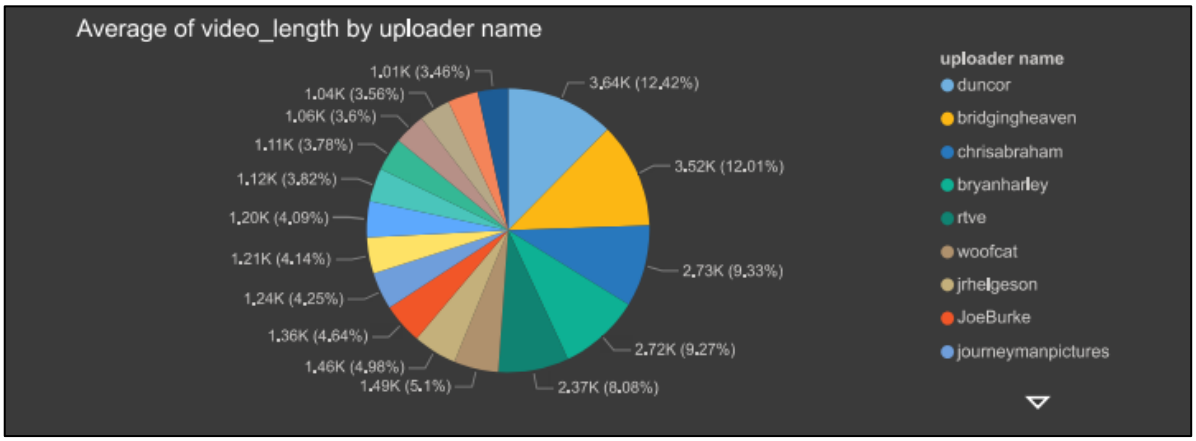


Figure 26: Data visualization: Average video length by uploader name

11. Number of views grouped by uploader name.

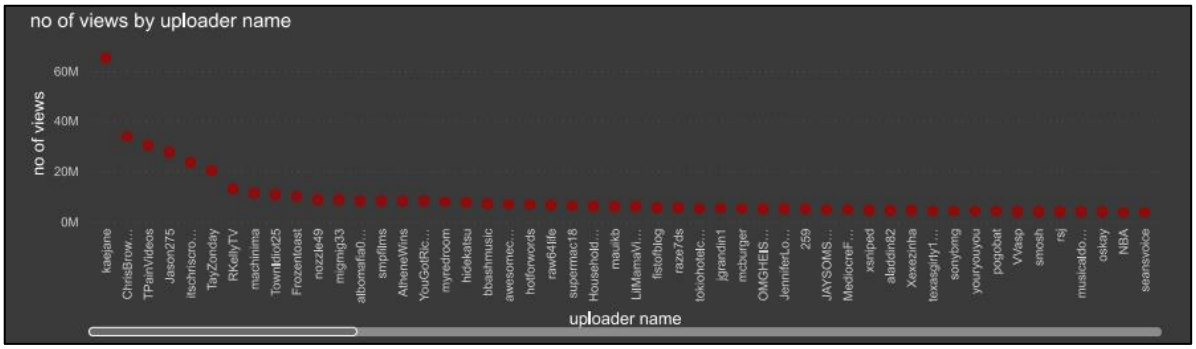


Figure 27: Data visualization: Number of views by uploader name

12. Ratings grouped by uploader name.

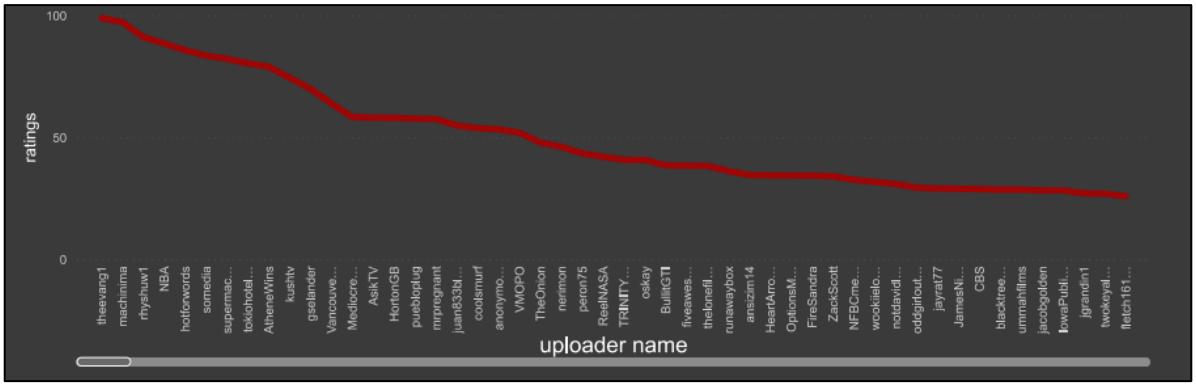


Figure 28: Data visualization: Ratings by uploader name

13. Conclusion

Entertainment has become a necessity of life for us and there is enough content today to keep us engaged for every moment of the rest of our lives. YouTube is one of the best examples of services that produce a massive amount of data in a brief period. We have demonstrated how we can extract insightful information from YouTube dataset using Big Data Analytics. For the same purpose, we have used technologies like Hadoop, Hive and Power BI.

14. References

[1] F. Shaikh, D. Pawaskar, A. Siddiqui and U. Khan, "YouTube Data Analysis using MapReduce on Hadoop," 2018 3rd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT), 2018, pp. 2037-2041, doi: 10.1109/RTEICT42901.2018.9012635.

[2] P. Merla and Y. Liang, "Data analysis using hadoop MapReduce environment," 2017 IEEE International Conference on Big Data (Big Data), 2017, pp. 4783-4785, doi: 10.1109/BigData.2017.8258541.

[3] R. Harsh, G. Acharya and S. Chaudhary, "Big Data Hysteria, Cognizance and Scope," 2018 4th International Conference for Convergence in Technology (I2CT), 2018, pp. 1-8, doi: 10.1109/I2CT42659.2018.9057878.

[4] J. Sang and C. Xu, "On Analyzing the 'Variety' of Big Social Multimedia," 2015 IEEE International Conference on Multimedia Big Data, 2015, pp. 5-8, doi: 10.1109/BigMM.2015.60.