TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

VIỆN TRÍ TUỆ NHÂN TẠO

-----***-----

BÁO CÁO MÔN HỌC KỸ THUẬT VÀ CÔNG NGHỆ DỮ LIỆU LỚN

ĐỀ TÀI

YOUTUBE DATA ANALYSIS

Giảng viên hướng dẫn: TS. Trần Hồng Việt

CN. Đỗ Thu Uyên

Nhóm sinh viên thực hiện:

1. Trịnh Minh Hiếu 22022536
2. Trần Hồng Đăng 22022646
3. Trần Kim Thành 22022532
4. Nguyễn Trường Huy 22022509

Hà Nội - 2024.

MỞ ĐẦU

Ngày hôm nay, chúng ta đang sống trong một thời kỳ cách mạng về khoa học và kĩ thuật, thời điểm mà đổi mới và áp dụng công nghệ trở thành cuộc đua cho những doanh nghiệp với mục tiêu thành công và phát triển lâu dài.

Big Data là một lĩnh vực vô cùng quan trọng, giúp cho AI và IoT phát triển nhanh và mạnh vì nó có thể tác động sâu rộng vào hoạt động kinh doanh, sản xuất của doanh nghiệp và cả đời sống con người thông qua khả năng thu thập, phân tích và lưu trữ lượng lớn dữ liệu, từ đó nâng cao hiệu quả kinh doanh và chất lượng cuộc sống, thấu hiểu tâm lý khách hàng và nắm bắt xu hướng thị trường.

Là một trong những nền tảng đa phương tiện lớn nhất trên thế giới, YouTube là nơi mà chúng ta có thể tìm thấy mọi nội dung mà chính sách của họ cho phép. Theo thống kê vào tháng 3/2024, mỗi tháng YouTube có 2.49 tỷ người dùng thường xuyên, nghĩa là lượng dữ liệu mà YouTube xử lý và lưu trữ là vô cùng khổng lồ, tạo ra tiềm năng phân tích và hiểu sâu hơn về trải nghiệm cá nhân của mỗi người dùng, hoặc các hiện tượng xã hội khác. Chính vì vậy, nhóm chúng em đã lựa chọn chủ đề **YouTube Data Analysis** để làm báo kết thúc môn học của mình.

Báo cáo bao gồm 6 chương:

1. Tổng quan về Big Data.
2. Dữ liệu và tiền xử lý dữ liệu.
3. Phân tích dữ liệu theo chủ đề (Category).
4. Phân tích dữ liệu theo thời gian (Time).
5. Phân tích dữ liệu theo tương tác (Views, Likes, Dislikes).
6. Kết luận và hướng phát triển.

Link Github của nhóm: https://github.com/huynt119/Youtube-Data-Analysis-by-PySpark

MỤC LỤC

MỞ ĐẦU 1

CHƯƠNG 1 - TỔNG QUAN VỀ DỮ LIỆU LỚN. 3

1.1 Định nghĩa. 3

1.2 Đặc trưng. 3

1.3 Nền tảng Apache Spark. 4

CHƯƠNG 2 - DỮ LIỆU VÀ TIỀN XỬ LÝ DỮ LIỆU. 5

2.1 Tổng quan bộ dữ liệu. 5

2.2 Hướng tiếp cận. 5

2.3 Tiền xử lý dữ liệu. 5

CHƯƠNG 3 - PHÂN TÍCH DỮ LIỆU THEO CHỦ ĐỀ (CATEGORY). 8

3.1 Tổng quát về chủ đề. 8

3.2 Phân tích chủ đề nào có xu hướng lên top training nhiều nhất. 8

3.3 Chủ đề có tổng lượt xem cao nhất. 10

3.4 Chủ đề có nhiều kênh YouTube làm nhất. 11

CHƯƠNG 4 - PHÂN TÍCH DỮ LIỆU THEO THỜI GIAN (TIME). 13

4.1 Tổng quát về thời gian. 13

4.2 Số lượt xem khi bắt đầu trending. 13

4.3 Số video trending ngay trong ngày đăng. 15

4.4 Giờ đăng tải có số video trending nhiều nhất. 16

4.5 Giờ đăng tải để video trending lâu nhất. 17

CHƯƠNG 5 - PHÂN TÍCH THEO TƯƠNG TÁC (VIEWS, LIKES, DISLIKES). 18

5.1 Tổng quan về tương tác. 18

5.2 Phân tích về tương tác. 19

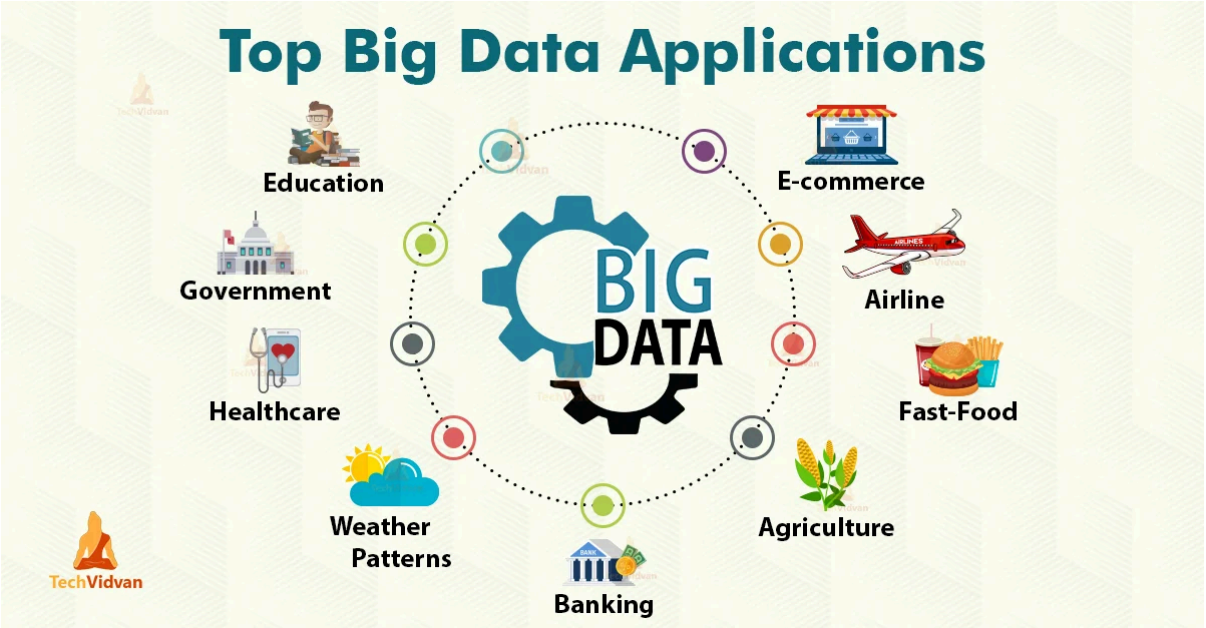
CHƯƠNG 6 - KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN. 23

NHIỆM VỤ CỦA CÁC THÀNH VIÊN. 24

CHƯƠNG 1 - TỔNG QUAN VỀ DỮ LIỆU LỚN.

1) Định nghĩa.

- Dữ liệu lớn là một thuật ngữ cho việc xử lý một tập hợp dữ liệu rất lớn và phức tạp mà các ứng dụng xử lý dữ liệu truyền thống không xử lý được. Dữ liệu lớn bao gồm các thách thức như phân tích, thu thập, giám sát dữ liệu, tìm kiếm, chia sẻ, lưu trữ, truyền nhận, trực quan, truy vấn và tính riêng tư.
- Theo Gartner, dữ liệu lớn là "khối lượng lớn, tốc độ cao và/hoặc loại hình thông tin rất đa dạng mà yêu cầu phương thức xử lý mới để cho phép tăng cường ra quyết định, khám phá bên trong và xử lý tối ưu" (2012).
- Big data được áp dụng trong mọi mặt của cuộc sống hiện nay.



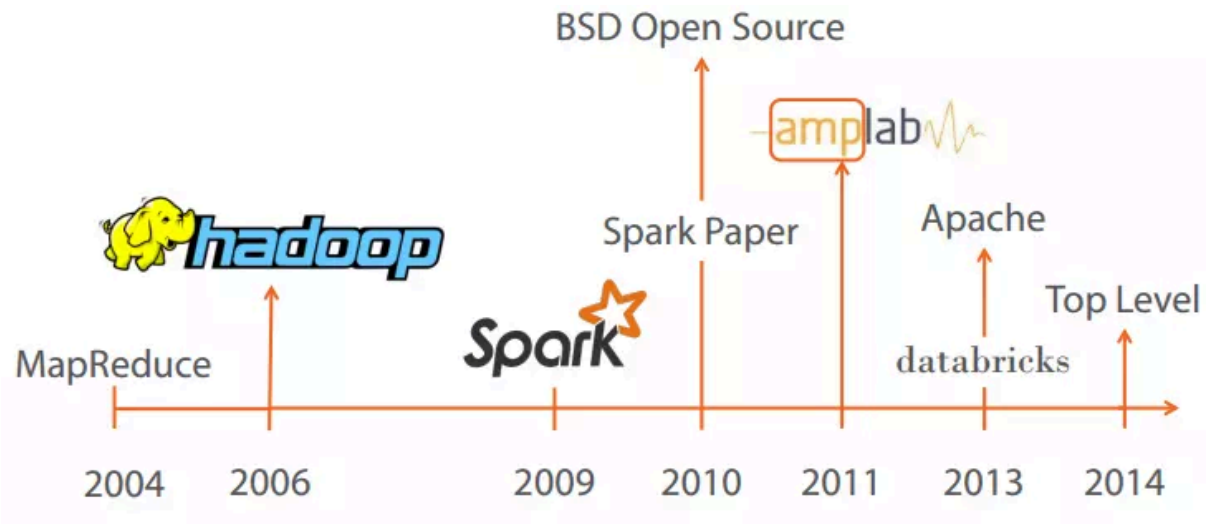
Các lĩnh vực ứng dụng big data.

2) Đặc trưng.

- Big Data được mô tả bởi những đặc trưng sau (5 Vs):
  + **Volume** (Dung lượng): Kích thước của dữ liệu sẽ được đánh giá là có giá trị, tiềm năng, và liệu nó có được coi là dữ liệu lớn không. Năm 2024, các doanh nghiệp lớn phải xử lý đến hàng **petabyte** dữ liệu.
  + **Variety** (Đa dạng): Các dạng và kiểu của dữ liệu. Dữ liệu được thu thập từ nhiều nguồn khác nhau và các kiểu dữ liệu cũng có rất nhiều cấu trúc khác nhau.
  + **Velocity** (Tốc độ): Nói tới tốc độ các dữ liệu được tạo ra và xử lý để đáp ứng các nhu cầu và thách thức trên con đường tăng trưởng và phát triển.

+ **Veracity** (Xác thực): Độ chính xác của dữ liệu thu được có thể phân hóa nhiều, ảnh hưởng đến kết quả phân tích.
+ **Value** (Giá trị): Ý nghĩa của dữ liệu, mang lại giá trị nhất định cho doanh nghiệp thông qua phân tích.

3) Nền tảng Apache Spark.



- **Apache Spark** là một framework mã nguồn mở tính toán cụm, được phát triển sơ khởi vào năm 2009 bởi AMPLab. Sau này, Spark đã được trao cho Apache Software Foundation vào năm 2013 và được phát triển cho đến nay.
- Apache Spark gồm có 5 thành phần chính : **Spark Core**, **Spark Streaming**, **Spark SQL**, **MLlib** và **GraphX**.
- Tốc độ xử lý của Spark là do tính toán được thực hiện cùng lúc trên nhiều máy khác nhau, được thực hiện ở bộ nhớ trong (**in-memories**) hoặc thực hiện hoàn toàn trên **RAM**.
- Spark hỗ trợ xử lý dữ liệu theo thời gian thực hoặc theo lô với dữ liệu từ các hệ thống file tương thích như: HDFS, Cassandra, S3,… và Spark hỗ trợ nhiều kiểu định dạng file khác nhau (text, csv, json…).
- Hỗ trợ ngôn ngữ: Java, Scala, Python và R.

CHƯƠNG 2 - DỮ LIỆU VÀ TIỀN XỬ LÝ DỮ LIỆU

1) Tổng quan bộ dữ liệu.

- Nguồn từ link Github tham khảo: https://github.com/SarahAyaz/YouTube_Data_Analysis/tree/master
- Kích thước khoảng 53MB, bao gồm 16 trường thông tin từ 43295 video (có trùng lặp) trong top trending kéo dài từ đầu 2018 đến 31/05/2018.
- Mỗi điểm dữ liệu trùng lặp tương ứng với 1 ngày trên top trending của video đó.

2) Hướng tiếp cận.

- Việc tận dụng lượng dữ liệu khổng lồ được sinh ra từ trải nghiệm người dùng YouTube được coi là đặc biệt quan trọng trong các lĩnh vực như truyền thông, tài chính và marketing. Từ những phân tích đó, doanh nghiệp có thể hiểu về nhu cầu của số đông, từ đó tạo ra các chiến dịch quảng cáo hợp thời, có hiệu quả cao. Để phân tích dữ liệu Youtube hiệu quả, nhóm sẽ phân tích 3 phần chính:
    + Phân tích theo hạng mục: Việc phân tích theo hạng mục giúp ta chọn lọc được hạng mục thu hút sự chú ý người dùng và đầu tư nhiều tài nguyên hơn cho hạng mục đó.
    + Phân tích theo thời gian: Xu hướng là điều luôn luôn thay đổi và phát triển. Nếu tận dụng được dữ liệu thời gian, doanh nghiệp sẽ có cái nhìn sâu hơn về những chuyển biến không ngừng.
    + Phân tích theo lượng tương tác: Sẽ giúp chọn được kênh thông tin phù hợp nhất với nhu cầu của từng người dùng.
- Tuy nhiên, các thư viện Python truyền thống như Numpy, Pandas,... không thật sự phù hợp để phân tích khối lượng dữ liệu khổng lồ như vậy, đặc biệt là khi cần phải phân tích luồng dữ liệu trực tiếp. Cùng với những đặc điểm nổi trội của Spark, nhóm em quyết định sử dụng **Spark** cho **YouTube Data Analysis**.

3) Tiền xử lý dữ liệu.

- Các bước thực hiện:
    + Bước 1: Khởi tạo Spark Session và đọc file .csv thành dataframe Spark. Điều này cho phép sử dụng các hàm xử lý của framework lên bộ dữ liệu đã đọc.
    + Bước 2: Thống kê sơ bộ bộ dữ liệu gốc bằng hàm **.printSchema()**.

```
----------------------------------------
DataFrame thông tin:
Số dòng: 43295
Số cột: 16
----------------------------------------
Schema:
root
 |-- video_id: string (nullable = true)
 |-- trending_date: string (nullable = true)
 |-- title: string (nullable = true)
 |-- channel_title: string (nullable = true)
 |-- category_id: string (nullable = true)
 |-- publish_time: string (nullable = true)
 |-- tags: string (nullable = true)
 |-- views: string (nullable = true)
 |-- likes: string (nullable = true)
 |-- dislikes: string (nullable = true)
 |-- comment_count: string (nullable = true)
 |-- thumbnail_link: string (nullable = true)
 |-- comments_disabled: string (nullable = true)
 |-- ratings_disabled: string (nullable = true)
 |-- video_error_or_removed: string (nullable = true)
 |-- description: string (nullable = true)
```

+ Bước 3: Tiến hành xóa các cột không cần thiết để giảm kích thước lưu trữ sử dụng hàm **.drop()**.

```
----------------------------------------
DataFrame thông tin:
Số dòng: 43295
Số cột: 12
----------------------------------------
Schema:
root
 |-- video_id: string (nullable = true)
 |-- trending_date: string (nullable = true)
 |-- title: string (nullable = true)
 |-- channel_title: string (nullable = true)
 |-- category_id: string (nullable = true)
 |-- publish_time: string (nullable = true)
 |-- tags: string (nullable = true)
 |-- views: string (nullable = true)
 |-- likes: string (nullable = true)
 |-- dislikes: string (nullable = true)
 |-- comment_count: string (nullable = true)
 |-- description: string (nullable = true)
```

+ Bước 4: Sử dụng hàm **.filter()** xóa tất cả các hàng chứa trường dữ liệu null, xử lý cột “trending_date” có chứa nhiều định dạng lỗi khiến cho các giá trị khác trong hàng có nhiều giá trị null .

```
Số giá trị null trong mỗi cột:
+--------+-------------+-----+-------------+-----------+------------+----+-----+-----+--------+-------------+-----------+
|video_id|trending_date|title|channel_title|category_id|publish_time|tags|views|likes|dislikes|comment_count|description|
+--------+-------------+-----+-------------+-----------+------------+----+-----+-----+--------+-------------+-----------+
|     110|         3755| 3872|         3937|       4093|        4134|4352| 4365| 4365|    4400|         4400|       5015|
+--------+-------------+-----+-------------+-----------+------------+----+-----+-----+--------+-------------+-----------+
```

```
Số giá trị null trong mỗi cột:
+--------+-------------+-----+-------------+-----------+------------+----+-----+-----+--------+-------------+-----------+
|video_id|trending_date|title|channel_title|category_id|publish_time|tags|views|likes|dislikes|comment_count|description|
+--------+-------------+-----+-------------+-----------+------------+----+-----+-----+--------+-------------+-----------+
|       0|         3645| 3762|         3827|       3983|        4024|4242| 4255| 4255|    4290|         4290|       4905|
+--------+-------------+-----+-------------+-----------+------------+----+-----+-----+--------+-------------+-----------+
```

```
Số giá trị null trong mỗi cột:
+--------+-------------+-----+-------------+-----------+------------+----+-----+-----+--------+-------------+-----------+
|video_id|trending_date|title|channel_title|category_id|publish_time|tags|views|likes|dislikes|comment_count|description|
+--------+-------------+-----+-------------+-----------+------------+----+-----+-----+--------+-------------+-----------+
|       0|            0|    0|            0|          0|           0|   0|    0|    0|       0|            0|        560|
+--------+-------------+-----+-------------+-----------+------------+----+-----+-----+--------+-------------+-----------+
```

+ Bước 5: Thêm các xâu rỗng vào cột “description” bằng **.fillna()** do một số video không viết phần mô tả, không phải do lỗi.
+ Bước 6: Chuẩn hóa các cột dữ liệu thời gian dùng **.withColumn()** để chọn cột cụ thể và kết hợp với **to_timestamp**. Cột “trending_date” không có dữ liệu giờ, thì mặc định là 00:00:00.
+ Bước 7: Lưu file đã tiền xử lý dưới định dạng .csv.

CHƯƠNG 3 - PHÂN TÍCH DỮ LIỆU THEO CHỦ ĐỀ (CATEGORY)

1) Tổng quát về chủ đề.

a) Chuẩn bị.

- Bước 1: Khởi tạo Spark Session và đọc file dữ liệu đã tiền xử lý, định dạng .csv thành dataframe Spark.
- Bước 2: Sử dụng chuỗi hàm:
  + **select(col_name)**: Chọn ra một cột trong DataFrame. Ở đây là cột "category_id".
  + **distinct([numPartitions]))**: Trả về một tập dữ liệu mới chứa các phần tử riêng biệt của tập dữ liệu nguồn. Ở đây lấy ra các giá trị riêng biệt trong cột "category_id".
  + **show()**: In tập dữ liệu.
  + **count()**: Trả về số lượng phần tử trong tập dữ liệu.

  => Các câu lệnh lấy ra các giá trị category riêng biệt và đếm số lượng category.

b) Phân tích.

- Dữ liệu top trending chứa 14 category: *Education*, *Gaming*, *Entertainment*, *Travel & Events*, *Science & Technology*, *Sports*, *Howto & Style*, *Film & Animation*, *People & Blogs*, *News & Politics*, *Pets & Animals*, *Autos & Vehicles*, *Music*, *Comedy* cho thấy bộ dữ liệu rất đa dạng các thể loại video, nói lên rằng sở thích của người dùng YouTube phân hóa đều trên mọi lĩnh vực.

2) Phân tích chủ đề nào có xu hướng lên top trending nhiều nhất (tính theo ngày).

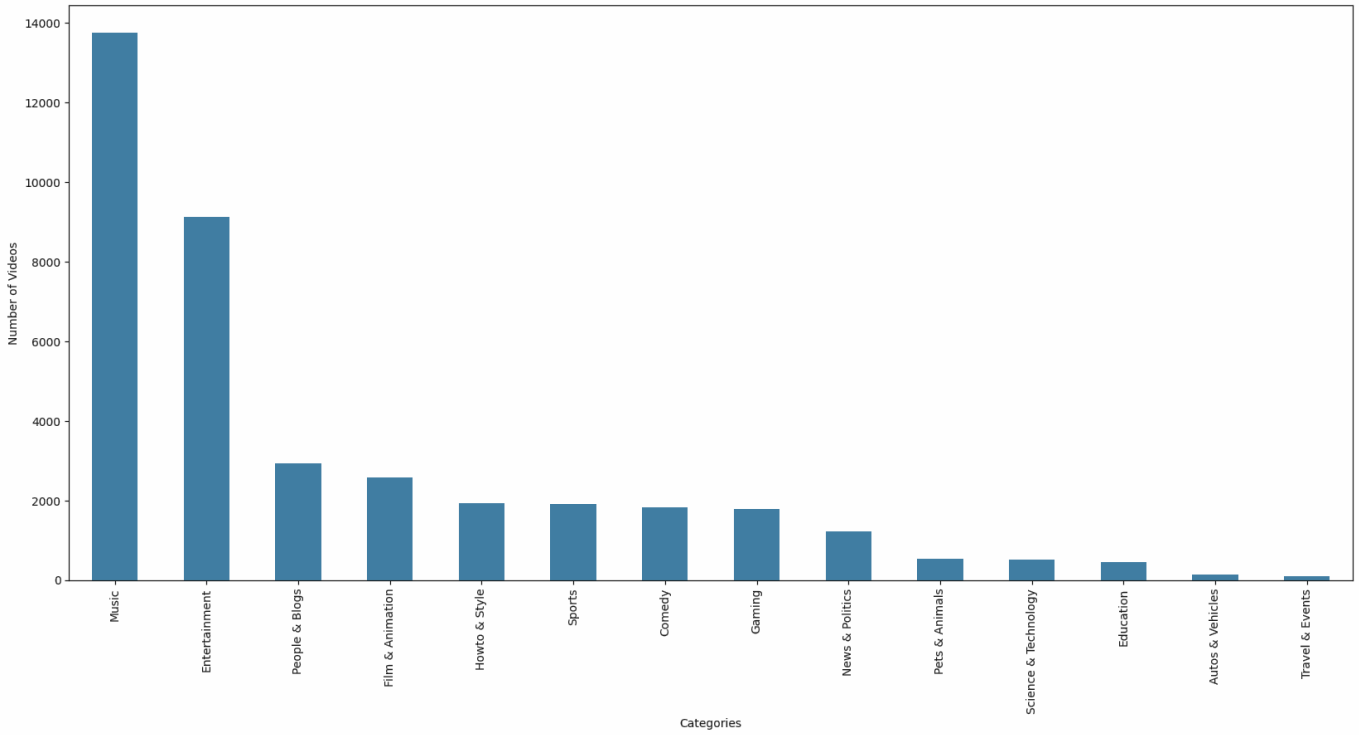
a) Code.

- Sử dụng chuỗi các hàm:
  + **groupBy(col_name)**: Nhóm các dòng trong DataFrame theo các cột được chỉ định để có thể chạy tổng hợp trên chúng. Ở đây, tất cả các dòng có cùng giá trị trong cột "category_id" sẽ được gom thành một nhóm.
  + **count()**: Trả về số lượng dòng trong mỗi nhóm.
  + **orderBy(col_name, ascending)**: Trả về một dataframe mới được sắp xếp theo các cột được chỉ định. Ở đây sắp xếp cột "count" theo thứ tự giảm dần.

  => Câu lệnh đếm số lần xuất hiện của mỗi category riêng biệt trong tập dữ liệu

b) Phân tích.

category_id	count
Music	13754
Entertainment	9124
People & Blogs	2926
Film & Animation	2577
Howto & Style	1928
Sports	1907
Comedy	1828
Gaming	1788
News & Politics	1225
Pets & Animals	534
Science & Technology	518
Education	457
Autos & Vehicles	144
Travel & Events	96



- Các category thuộc lĩnh vực mang tính **giải trí** như *Music*, *Entertainment* có xu hướng lên top trending nhiều nhất với tổng số ngày trending lần lượt là **13745 ngày và 9124 ngày** do phù hợp với hầu hết mọi đối tượng.
- Các category tập trung vào một lĩnh vực **chuyên môn** như *Autos & Vehicles*, *Education* ít lên trending do chỉ phù hợp với một số đối tượng nhất định.

- Các lĩnh vực còn lại có số liệu không chênh lệch quá nhiều, và *Music* có số ngày trend **gấp 1.5** lần *Entertainment* cho thấy các video âm nhạc có khả năng cao được xem lại (nghe lại) nhiều hoặc số lượng video *Music* nhiều hơn *Entertainment*.

3) Chủ đề có tổng số lượt xem cao nhất.

a) Chuẩn bị.

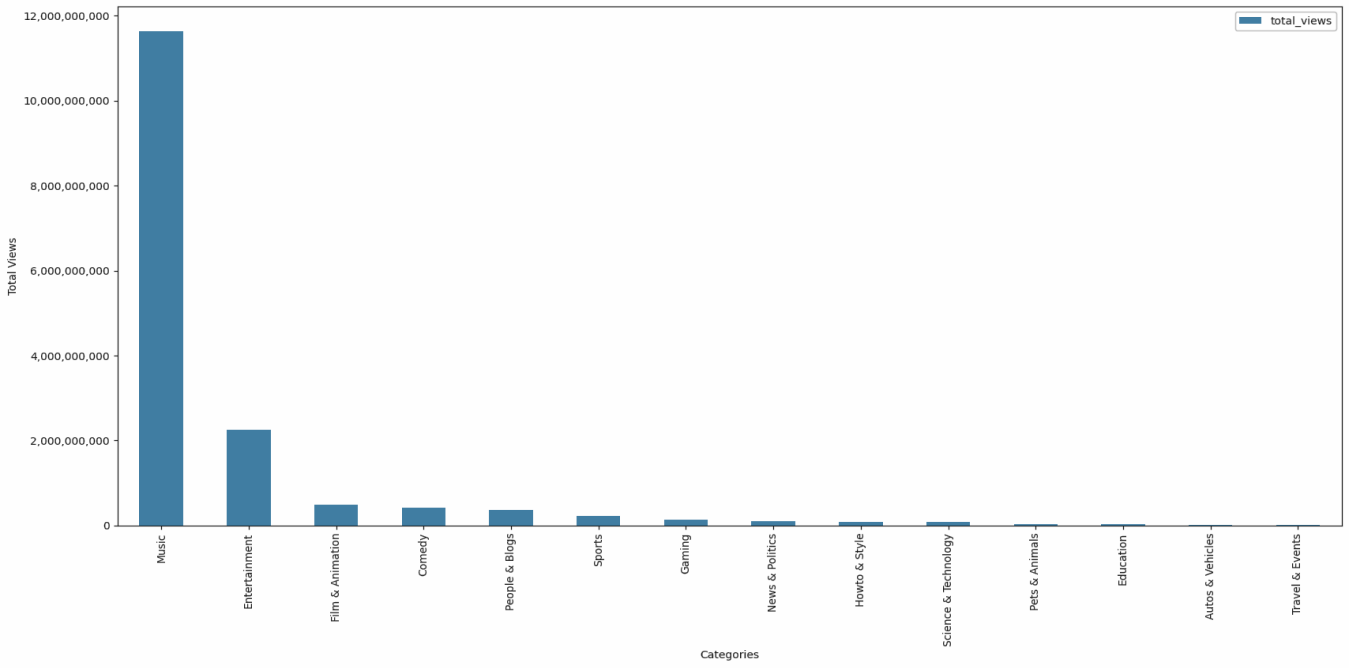
- Vì bộ dữ liệu là một số video nhất định trên top trending trong một thời gian nhất định nên sẽ có hiện tượng trùng lặp. Sử dụng chuỗi các hàm sau để sắp xếp các video riêng biệt theo thứ tự view và lấy ra số view cao nhất của mỗi video.
    + **Window.partitionBy(col_name)**: Tạo một WindowSpec với phân vùng được xác định. Ở đây nó chia dữ liệu thành các nhóm con theo title của video.
    + **orderBy(column.desc())**: Sắp xếp các giá trị giảm dần theo cột "views".
    + **withColumn(col_name, column)**: Trả về một dataframe mới bằng cách thêm một cột hoặc thay thế cột hiện có có cùng tên. Ở đây thêm cột "view_rank".
    + **row_number().over(Window function)**: Trả về một cột số tuần tự bắt đầu từ 1 trong phân vùng cửa sổ.
    + **filter(condition)**: Lọc các hàng bằng điều kiện đã cho. Ở đây lấy ra các hàng có views cao nhất (rank == 1).

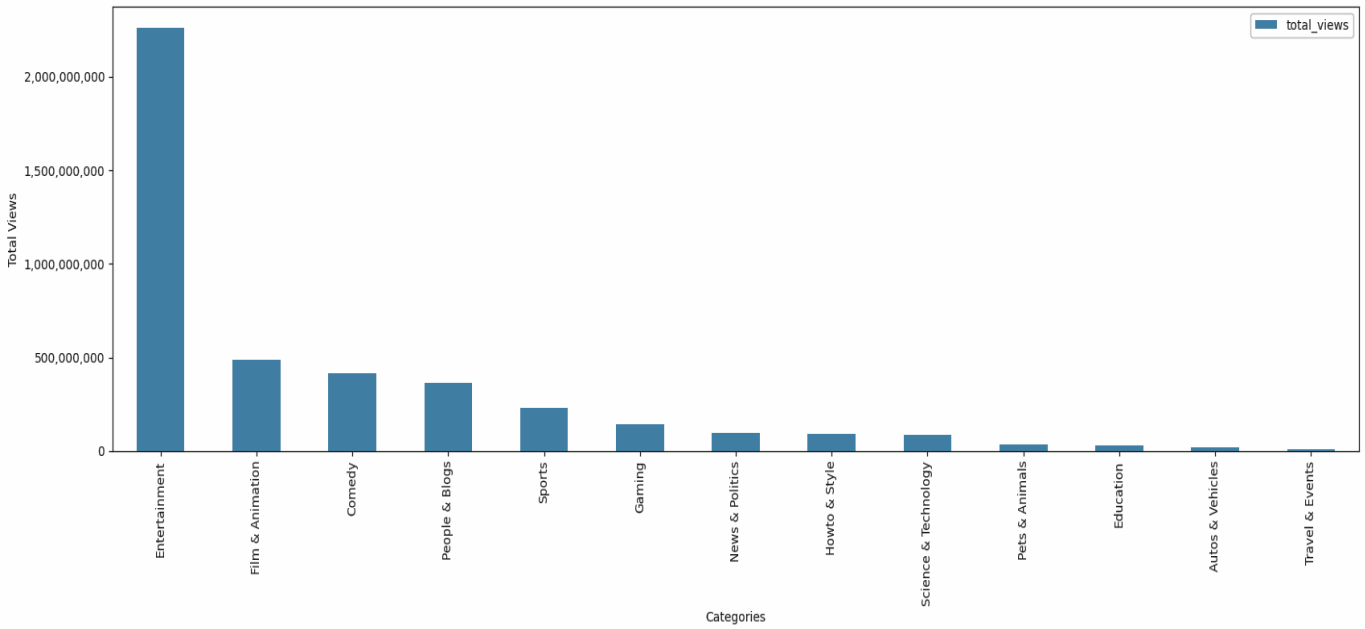
```
+-----+-----------+-------------+--------------------+--------------------+--------------+--------------------+--------------------+--------+------+--------+-------------+--------------------
|  _c0|   video_id|trending_date|               title|       channel_title|   category_id|        publish_time|                tags|   views| likes|dislikes|comment_count|                desc
+-----+-----------+-------------+--------------------+--------------------+--------------+--------------------+--------------------+--------+------+--------+-------------+--------------------
|  567|NLCfRHdi6Ig|   2017-11-16|#21 How to go FAS...|          Ben Cathro|        Sports|2017-11-05 20:10:16|"""Ben Cathro|"""...|   16074|   689|       8|          142|How flipping
| 1791|1cgK-BIrXes|   2017-11-22|#VeteransDay: Tha...|   YouTube Spotlight|Entertainment|2017-11-10 15:07:13|"""vets|""""veter...|  916104| 27405|    2153|         5292|This #Veteran
|19540|aydY1eiGjgs|   2018-02-22|'The Greatest Sho...|        TheEllenShow|Entertainment|2018-02-21 14:01:29|"""Ellen|""""dege...|  167263|  8357|      53|          365|The Greatest
|  701|d6YkGejS6IM|   2017-11-17|(ORIGINAL) I choo...|           Demon Inu|        Gaming|2017-11-08 05:32:12|              [none]|  681434|  6923|     234|         2482|I don't owned
| 7963|s0qeYQZ7mEo|   2017-12-23|12/11/17: White H...|     The White House|News & Politics|2017-12-11 20:09:02|              [none]|   42538|   896|     102|          969|      The Whit
|38334|yNOAs3poxMw|   2018-06-12|13 Reasons Why: S...|             Netflix|Entertainment|2018-06-06 14:59:47|"""Netflix|""""Tr...|  933253| 32177|    4589|         5448|13 Reasons Wh
|12965|31mjfjIkWiM|   2018-01-19|2017 BEAUTY FAVOU...|       Estée Lalonde|People & Blogs|2018-01-07 17:00:01|"""Estée Lalonde|...|  200232|  6011|     156|          416|WATCH MY 2016
|38118|pojXEFDFMw8|   2018-06-10|2018 FIFA World C...|                 ITV|Entertainment|2018-05-11 11:00:09|"""tv|""""televis...|  369011|  4782|     483|            0|Join us this
|28357|LPTlvQ1Zet0|   2018-04-07|21 Savage, Offset...|        21SavageVEVO|         Music|2018-03-01 14:00:02|"""21 savage|""""...|69907195|835963|   30861|        40903|Without Warni
|34848|M84oouivF7k|   2018-05-21|24 Hours With Cam...|               Vogue|Entertainment|2018-04-18 17:13:16|"""camilla cabell...|  880824| 35154|     598|         1416|We're all suc
| 9371|lw_KFCgcvA4|   2017-12-30|3 EVERYDAY LOOKS ...|      Samantha Maria|Howto & Style|2017-12-13 17:00:04|"""samantha|""""m...|   72266|  4522|     105|          175|3 Autumn/wint
| 3583|MfLcjWq6m9o|   2017-12-01|360° Norwegian Ke...|           BBC Earth|Pets & Animals|2017-11-22 10:00:01|"""BBC|""""BBC Wo...|   34425|   904|      38|           50|Join us under
|38357|kPEzyLTR-a4|   2018-06-12|500 Days of Ameri...|The Opposition w/...|         Comedy|2018-06-05 01:55:22|"""The Opposition...|  374013|  3460|     267|          501|Jordan celebr
|10176|iiWm8AxvWtw|   2018-01-03|73 Questions With...|               Vogue|Entertainment|2017-12-14 13:00:54|"""73 qs|""""73 q...|  431006| 12007|     256|          874|Rapper, produ
| 4584|WkzmaFnPKOs|   2017-12-06|81. I'm not good ...|     Will Darbyshire|Film & Animation|2017-11-26 20:10:33|"""willdarbyshire...|  128972| 12444|      40|          790|Cheers to Ola
|16307|6ZwXRnueOMg|   2018-02-05|9 SITUATIONS EVER...|      smoothiethecat|Pets & Animals|2018-01-26 17:38:30|"""9 Situations E...|   68307|  3322|      43|          303|Cats all have
| 5582|nVrkQyLS4oE|   2017-12-11|90's Toy Mystery ...|        Tyler Oakley|Entertainment|2017-11-28 20:01:26|"""tyler oakley|"""...|  438173| 30622|     383|         1645|Check out thi
| 3791|97IfPfjSaDg|   2017-12-02|   A Cuphead Cartoon|     hotdiggedydemon|Film & Animation|2017-11-18 22:03:41|"""CUPHEAD|""""MU...| 3683523|113231|    5488|        17462|Cuphead learn
|26388|pHvXZLREt9E|   2018-03-28|A Dying Toys R Us...|WorldClassBullshi...|Entertainment|2018-03-11 06:06:57|"""toys r us|""""...|  201287|  4402|     719|         2601|Toys R Us is
| 4650|H83cVi2sCJM|   2017-12-07|A FAST GRWM! OUTF...|      Samantha Maria|Howto & Style|2017-12-03 09:00:00|"""beauty|""""GRW...|   65368|  4594|     152|          173|When i'm out
+-----+-----------+-------------+--------------------+--------------------+--------------+--------------------+--------------------+--------+------+--------+-------------+--------------------
only showing top 20 rows
```

- Sau khi đã có một dataframe không trùng lặp, tiến hành các câu lệnh tính tổng số views của các video thuộc về từng loại category riêng biệt.
    + **agg(Union)**: Tổng hợp trên toàn bộ DataFrame. Ở đây áp dụng tính tổng lượng views **(sum(col_name))** cho mỗi category riêng trong tập dữ liệu
    + **alias(col_name)**: Trả về cột này được đặt tên mới.
    + **cast(type)**: Đổi định dạng cột. Ở đây chuyển thành kiểu long number.

b) **Phân tích.**

- Các video về lĩnh vực giải trí như: *Music*, *Entertainment* vẫn có lượng views vượt trội hơn so với lĩnh vực còn lại với *Music* trội hơn rất nhiều *Entertainment* (**gần 12 tỷ và gần 3 tỷ**). Điều này có thể hiểu được từ phân tích thời gian trending lâu nhất.



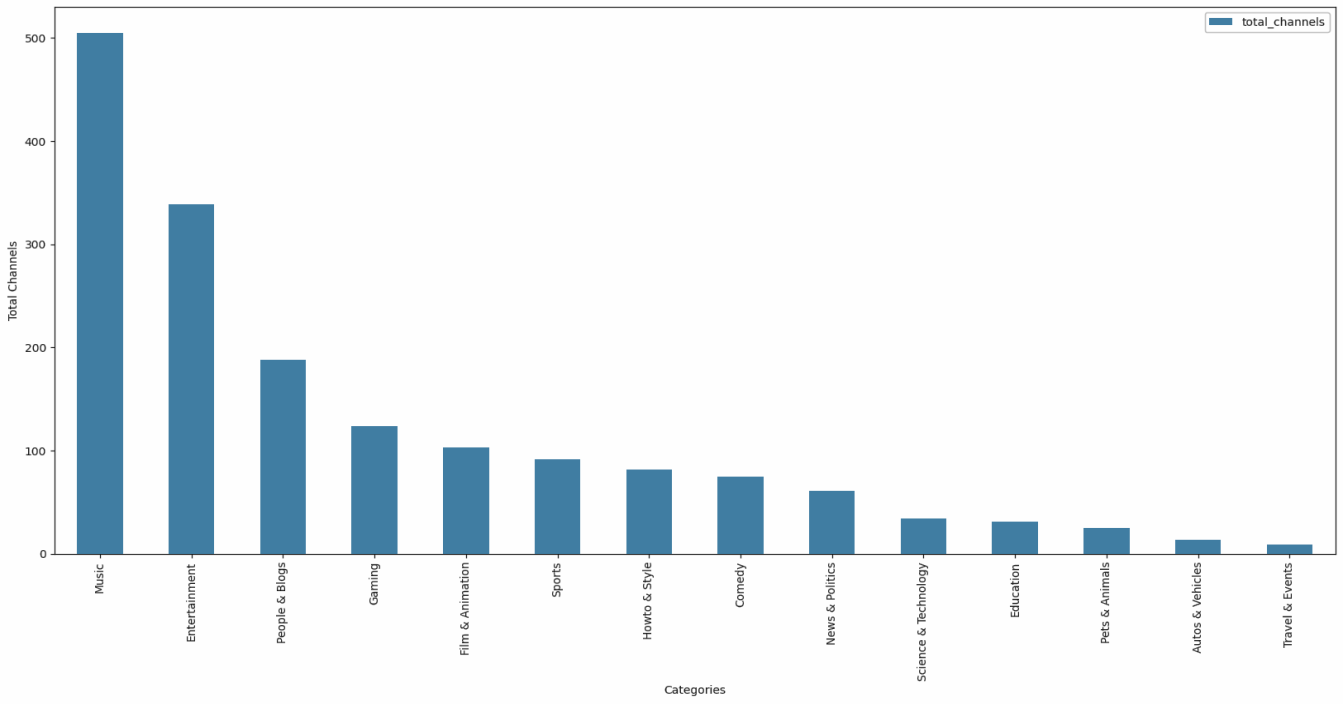


4) **Chủ đề có nhiều kênh YouTube làm nhất.**

a) Code.

- **count_distinct(col_name)**: Trả về cột mới đếm số lượng giá trị riêng biệt của một cột. Ở đây đếm số lượng kênh riêng biệt. Ở đây áp dụng lên dataframe đã lọc trùng lặp ở bước phân tích trước.

b) **Phân tích.**



- Các kênh làm về lĩnh vực giải trí vẫn có xu hướng lên top trending nhiều nhất. Với *Music* thống trị mọi số liệu, theo sau luôn là *Entertainment* cho thấy YouTube là nơi đa số người dùng nghe nhạc và xem video giải trí.

CHƯƠNG 4 - PHÂN TÍCH DỮ LIỆU THEO THỜI GIAN (TIME)

1) Tổng quát về time.

- Bước 1: Khởi tạo Spark Session và đọc file dữ liệu đã tiền xử lý, định dạng .csv thành dataframe Spark.
- Bước 2: Sử dụng chuỗi các hàm:
  + **.withColumn()**: Trả về một dataframe mới bằng cách thêm một cột hoặc thay thế cột hiện có có cùng tên.
  + **.to_date()**: Đổi sang dữ liệu chỉ bao gồm ngày, tháng, năm.
  + **.date_format()**: Đổi sang dữ liệu chỉ bao gồm giờ, phút, giây.
  + **.select()**: Tạo một dataframe mới từ các cột của dataframe gốc.

  => Format lại ngày giờ để dễ dàng phân tích, và sẽ không xem xét múi giờ do không có dữ liệu.

```
+-----------+------------+--------+------+-------------+-------------+-----------------+
|   video_id|publish_date|   views| likes|comment_count|trending_date|publish_time_only|
+-----------+------------+--------+------+-------------+-------------+-----------------+
|Jw1Y-zhQURU|  2017-11-10| 7224515| 55681|         9479|   2017-11-14|         07:38:29|
|3s1rvMFUweQ|  2017-11-12| 1053632| 25561|         2757|   2017-11-14|         06:24:44|
|n1WpP7iowLc|  2017-11-10|17158579|787420|       125882|   2017-11-14|         17:00:03|
|PUTEiSjKwJU|  2017-11-13|   27833|   193|           37|   2017-11-14|         02:30:38|
|rHwDegptbI4|  2017-11-13|    9815|    30|           30|   2017-11-14|         01:45:13|
+-----------+------------+--------+------+-------------+-------------+-----------------+
only showing top 5 rows
```

2) Lượng view khi bắt đầu trending.

a) Code.

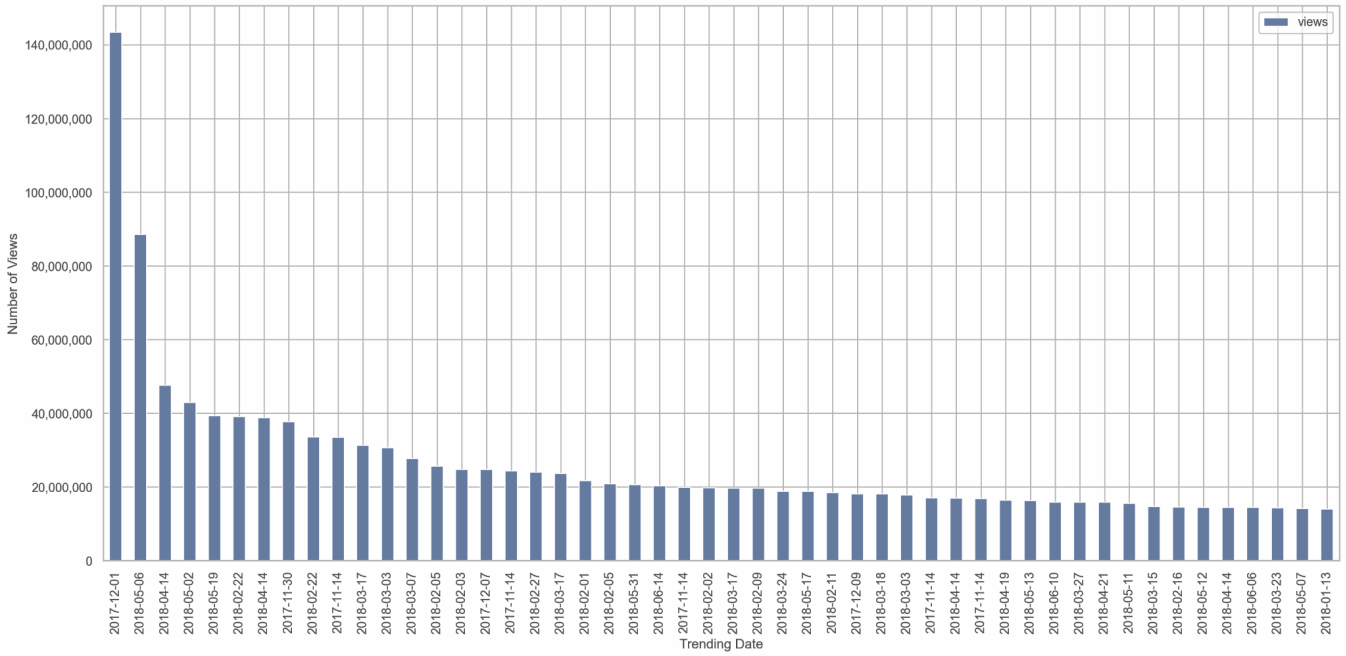
- Sử dụng chuỗi các hàm:
  + **Window.partitionBy()**: tạo một cửa sổ để nhóm theo “video_id”.
  + **withColumn(“rank”, F.row_number().over(window_spec))**: thêm cột rank để xác định hàng đầu tiên cho mỗi “video_id”.
  + **filter(F.col(“rank”) == 1).drop(“rank”)**: Lọc chỉ giữ lại hàng đầu tiên của mỗi “video_id”.
  + **orderBy(F.desc(“views”)).limit(100)**: xếp hạng giảm dần số view của các video vừa lấy và chỉ lấy 100 cái đầu tiên.

b) Phân tích.

- Nhìn chung, các video khi bắt đầu trending có lượng view không chênh lệch nhiều, chỉ có sự khác biệt lớn khi so sánh video có nhiều view nhất và ít view nhất trước khi trending.

video_id	publish_date	views	likes	comment_count	trending_date	publish_time_only
TyHvyGVs42U	2017-11-17	143408235	2686169	144217	2017-12-01	05:00:01
zEf423kYfqk	2018-04-20	88568646	1185357	70242	2018-05-06	10:40:51
WtE011iVx1Q	2018-03-30	47669287	396337	15955	2018-04-14	04:00:02
Ck4xHocysLw	2018-04-26	42923278	495422	18091	2018-05-02	17:52:13
7C2z4GqqS5E	2018-05-18	39349927	3880074	692311	2018-05-19	09:00:02
VTzD0jNdrmo	2018-02-09	39118664	383030	13358	2018-02-22	14:06:54
i0p1bmr0EmE	2018-04-09	38873543	1111595	206639	2018-04-14	08:59:51
6ZfuNTqbHE8	2017-11-29	37736281	1735902	241237	2017-11-30	13:26:24
xpVfcZ0ZcFM	2018-02-17	33591858	2152150	140512	2018-02-22	05:00:01
2Vv-BfVoq4g	2017-11-09	33523622	1634124	85067	2017-11-14	11:04:14
LPTlvQ1Zet0	2018-03-01	31301612	594755	32225	2018-03-17	14:00:02
_I_D_8Z4sJE	2018-03-02	30686233	304506	13903	2018-03-03	05:00:19
0XElmYomloA	2018-02-22	27801810	104931	2396	2018-03-07	22:08:34
J6-8DQALGt4	2018-01-31	25703405	62460	4428	2018-02-05	13:59:16
sD9_l3oDOag	2018-01-26	24843733	372300	14316	2018-02-03	05:00:01
FlsCjmMhFmw	2017-12-06	24784863	1149190	462108	2017-12-07	17:58:51
4GFAZBKZVJY	2017-11-03	24412837	248684	0	2017-11-14	12:07:57
JNkTNAknE4I	2018-02-21	24034926	125062	3857	2018-02-27	23:49:59
pgN-vvVVxMA	2018-03-02	23782607	780002	66159	2018-03-17	05:05:53
wfWkmURBNv8	2018-01-30	21800407	567061	31460	2018-02-01	15:00:05
BhIEIO0vaBE	2018-02-04	20921796	0	0	2018-02-05	20:27:38
xTlNMmZKwpA	2018-05-29	20723565	1018786	68790	2018-05-31	14:05:10
V15BYnSr0P8	2018-06-08	20409647	568680	64390	2018-06-14	11:58:38
V54CEElTF_U	2017-11-03	19971162	694768	50164	2017-11-14	04:00:01
qtTM2YV3bI8	2018-01-28	19878085	874521	237473	2018-02-02	23:41:31
QwievZ1Tx-8	2018-03-16	19716689	975725	127045	2018-03-17	13:02:41
6OcOO1k-vGE	2018-01-29	19701117	409728	24240	2018-02-09	09:49:15
ToY6sjSV8h8	2018-03-05	18857011	310205	14524	2018-03-24	05:00:00
6YNZlXfW6Ho	2018-04-26	18825555	311827	11193	2018-05-17	17:00:01
9TRjE7i0ERY	2018-02-09	18541827	184210	6538	2018-02-11	11:00:01

only showing top 30 rows



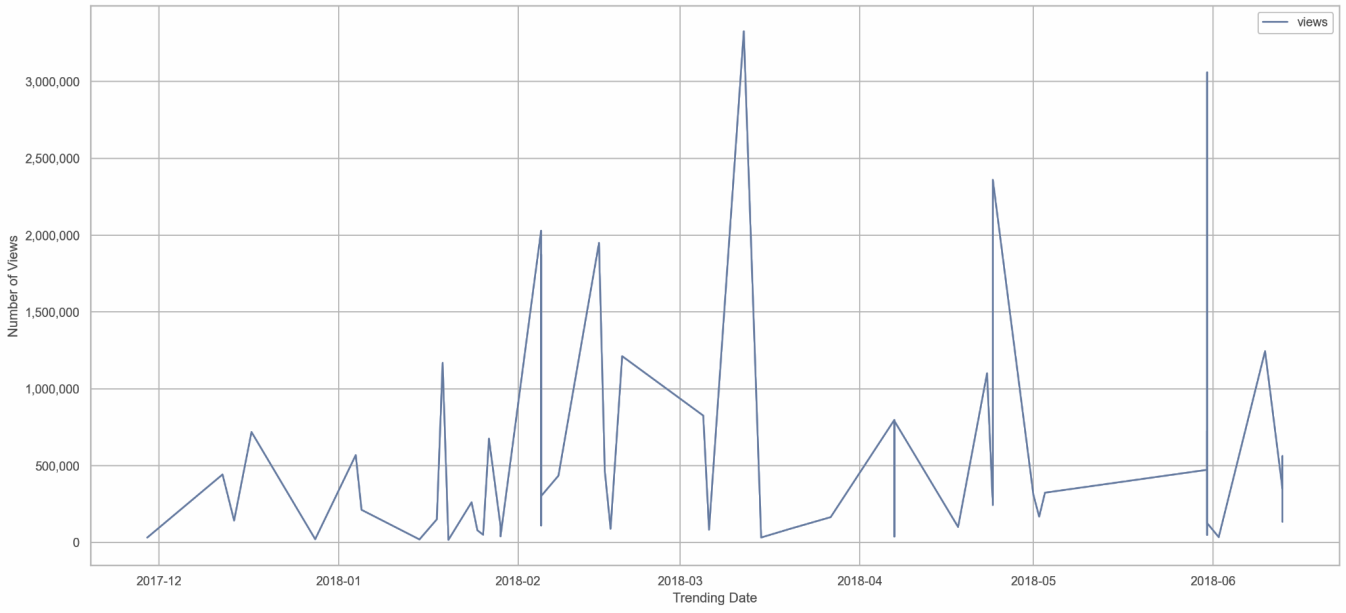
3) Số video trending ngay trong ngày đăng.

a) Code.

- **.withColumn("days_since_upload",F.datediff(F.col("trending_date"), F.col("publish_date")))**: Tạo thêm cột "days_since_upload", tính khoảng cách giữa ngày đăng và ngày trending.
- **.filter(F.col("days_since_upload") == 0).count()**: Lọc những hàng có giá trị cột "days_since_upload" = 0 và đếm số lượng hàng.
- **filter(F.col("days_since_upload") == 0).orderBy(F.asc("trending_date"))**: Lọc những hàng có giá trị cột days_since_upload = 0 và xếp tăng dần theo cột "trending_date".

```
+-----------+------------+-------+------+-------------+-------------+-----------------+-----------------+
|   video_id|publish_date|  views| likes|comment_count|trending_date|publish_time_only|days_since_upload|
+-----------+------------+-------+------+-------------+-------------+-----------------+-----------------+
|     #NAME?|  2017-11-10|1164201| 57309|          624|   2017-11-14|         19:19:43|                4|
|-3VBPAZPTQI|  2017-12-29| 209192| 11119|          740|   2017-12-31|         15:11:23|                2|
|-43MBOJnVks|  2018-02-13| 544638|  7985|         1038|   2018-02-14|         17:00:07|                1|
|-5WBCrazSfg|  2017-11-27| 172090|  4354|          587|   2017-11-29|         19:44:49|                2|
|-5aaJJQFvOg|  2018-02-21| 509454| 82757|         4535|   2018-02-24|         22:01:06|                3|
|-7tSTUR7FG0|  2018-02-27|1092530|152536|         8272|   2018-02-28|         12:00:00|                1|
|-8X32zNup1o|  2018-05-21| 513127|  6414|         3369|   2018-05-23|         22:13:07|                2|
|0-VAnh7r-_8|  2018-06-12|1100530| 33729|        18027|   2018-06-13|         01:30:02|                1|
|03RQLmDNIEw|  2017-12-03| 243420| 10374|          387|   2017-12-05|         15:00:01|                2|
|07w3u8iLa-s|  2018-04-07| 284180| 11281|         2562|   2018-04-14|         21:38:21|                7|
+-----------+------------+-------+------+-------------+-------------+-----------------+-----------------+
only showing top 10 rows
```

b) **Phân tích.**



- Lượng view khi bắt đầu trending của các video có sự phân hóa mạnh và không đồng đều, nói lên rằng video trending không nhất thiết phải có lượng view khổng lồ.

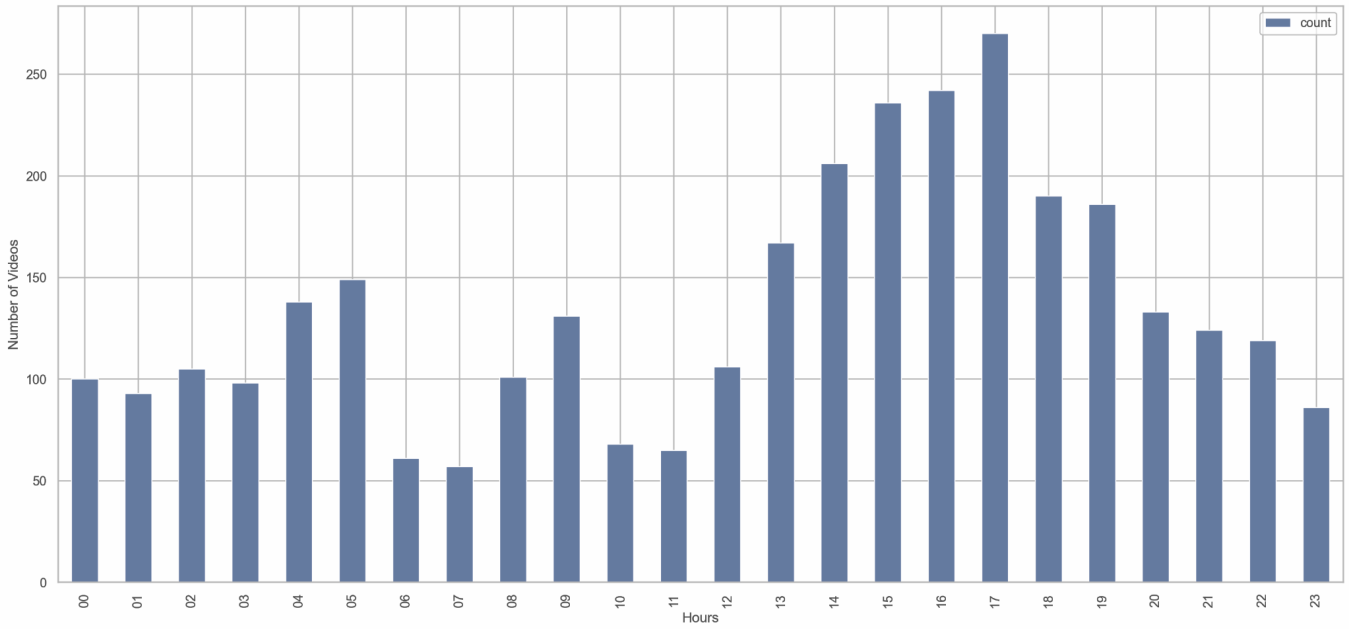
4) Giờ đăng tải có số video trending nhiều nhất.

a) Code.

- **.withColumn("hour", substring(col("publish_time_only"), 1, 2))**: Thêm cột hour, dùng substring để lấy 2 ký tự đầu là giờ trong cột "publish_time_only".
- **.groupBy("hour").count().orderBy(F.asc("hour"))**: Nhóm theo "hour" để đếm số lượng và sắp xếp tăng dần theo "hour".

```
+-----------+------------+-------+-----+-------------+-------------+-----------------+-----------------+----+
|   video_id|publish_date|  views|likes|comment_count|trending_date|publish_time_only|days_since_upload|hour|
+-----------+------------+-------+-----+-------------+-------------+-----------------+-----------------+----+
|     #NAME?|  2017-11-10|1164201|57309|          624|   2017-11-14|         19:19:43|                4|  19|
|-3VBPAZPTQI|  2017-12-29| 209192|11119|          740|   2017-12-31|         15:11:23|                2|  15|
|-43MBOJnVks|  2018-02-13| 544638| 7985|         1038|   2018-02-14|         17:00:07|                1|  17|
|-5WBCrazSfg|  2017-11-27| 172090| 4354|          587|   2017-11-29|         19:44:49|                2|  19|
|-5aaJJQFvOg|  2018-02-21| 509454|82757|         4535|   2018-02-24|         22:01:06|                3|  22|
+-----------+------------+-------+-----+-------------+-------------+-----------------+-----------------+----+
only showing top 5 rows
```

b) Phân tích.



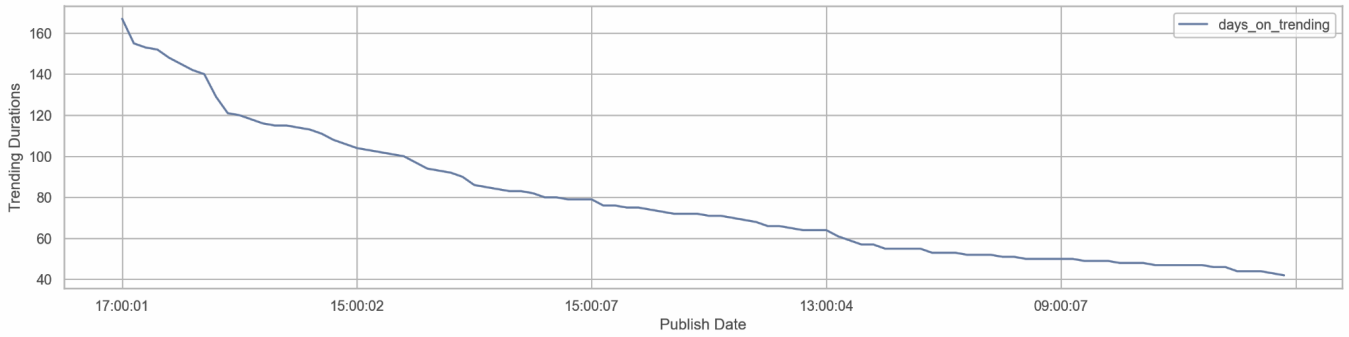
- Những video được đăng tải vào chiều hoặc tối thì sẽ có lượng view cao hơn do chiều là khi mà một phần người dùng kết thúc ngày làm việc, và YouTube là nơi họ có thể nghe nhạc, giải trí (thấy rõ tại phân tích category). Hơn nữa, 17 giờ là giờ tan ca của phần lớn người dùng, nên họ có thời gian lên YouTube. Trang chủ YouTube thường đẩy những video mới, phù hợp hoặc của kênh nổi tiếng lên đầu đề xuất. Chính vì vậy, video sẽ trend rất nhanh, và rất nhiều (**200 video tại 5 giờ chiều**).
- Sau đó video trend giảm dần về tối và đêm có thể là người dùng bận những công việc buổi tối, hoặc dành thời gian cho gia đình, còn ban ngày thấp nhất do họ phải đi làm, đi học.

5) Giờ đăng tải để video trending lâu nhất (hơn 50 ngày).

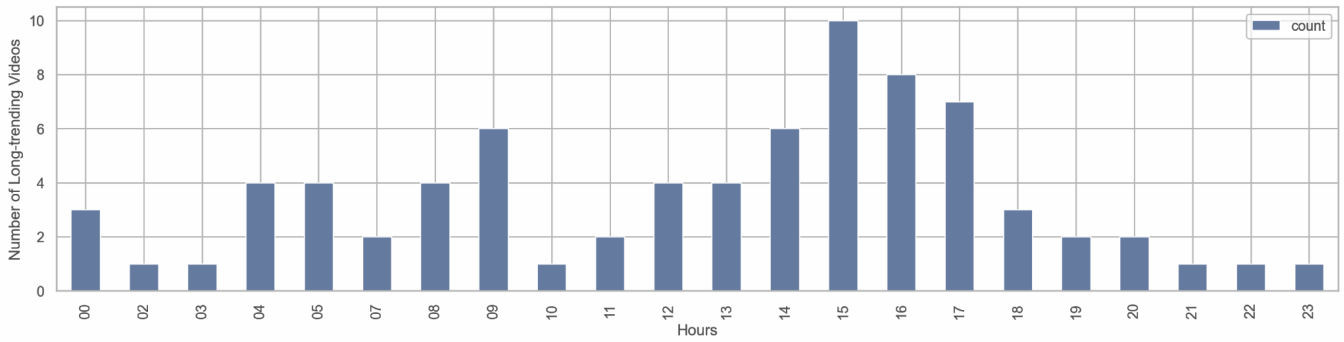
a) Code.

- **.groupBy("publish_time_only").agg(countDistinct("trending_date").alias("days_on_trending")**: Nhóm dữ liệu theo cột "publish_time_only", đếm số lượng những ngày khác nhau mà video xuất hiện trending và thêm cột kết quả "days_on_trending".
- **.orderBy(F.desc("days_on_trending")).limit(100)**: Sắp xếp giảm dần theo số ngày trên trending và lấy 100 hàng đầu.
- **.withColumn("hour", substring(col("publish_time_only"), 1, 2))**: Thêm cột "hour" bao gồm 2 ký tự đầu của cột "publish_time_only" là "hour".
- **.filter(F.col("days_on_trending") > 50).groupBy("hour").count()**: Lọc những hàng có giá trị cột "days_on_trending" > 50 và nhóm theo "hour" và đếm số lượng.

b) Phân tích.



- Trong khoảng từ **15h đến 17h**, các video có xu hướng trending lâu hơn, với thời gian lâu nhất là **hơn 160 ngày**, đăng tải lúc 17 giờ.



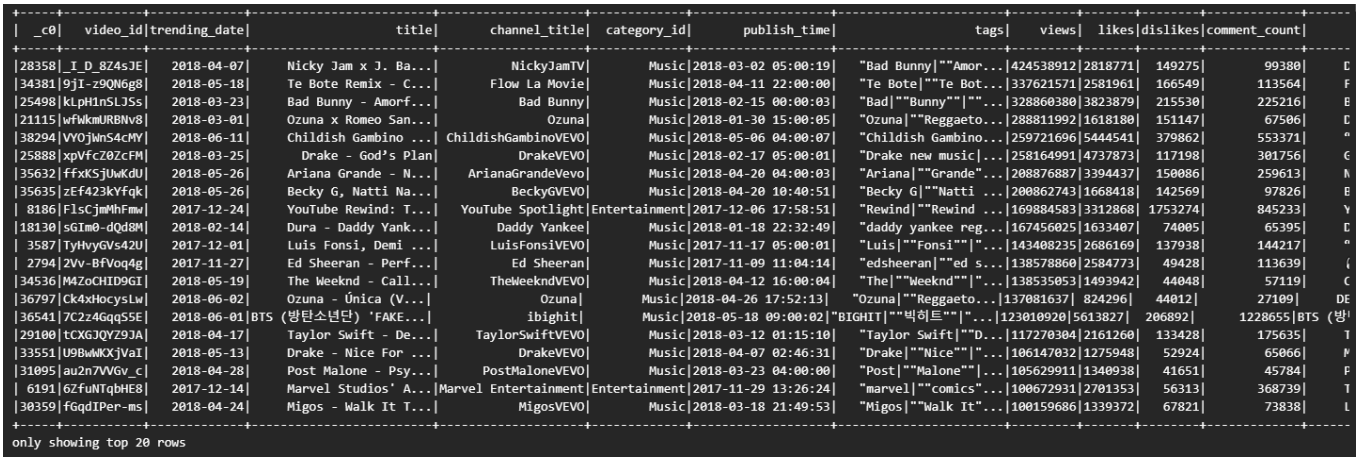
- Ở biểu đồ cột, khung giờ từ **15 giờ đến 17 giờ** có số video trending hơn 50 ngày nhiều nhất, được ghi nhận là **10, 8 và 7 video**. Khởi đầu bằng cách đăng đúng giờ đông người truy cập làm cho trending nhanh, cộng với nội dung hấp dẫn đã khiến cho một số video giữ top trending trong một thời gian dài.
- Hơn nữa, với số lượng lớn video là âm nhạc và giải trí, như đã nói, tỷ lệ xem lại (nghe lại) khá cao, làm cho video trụ lại top trending lâu hơn.

CHƯƠNG 5 - PHÂN TÍCH THEO TƯƠNG TÁC (VIEWS, LIKES, DISLIKES)

1) Tổng quan về tương tác.

a) Chuẩn bị.

- Bước 1: Khởi tạo Spark Session và đọc file dữ liệu đã tiền xử lý, định dạng .csv thành dataframe Spark.
- Bước 2: Lọc những video trùng lặp để lấy ra danh sách chuẩn như những phân tích trước.



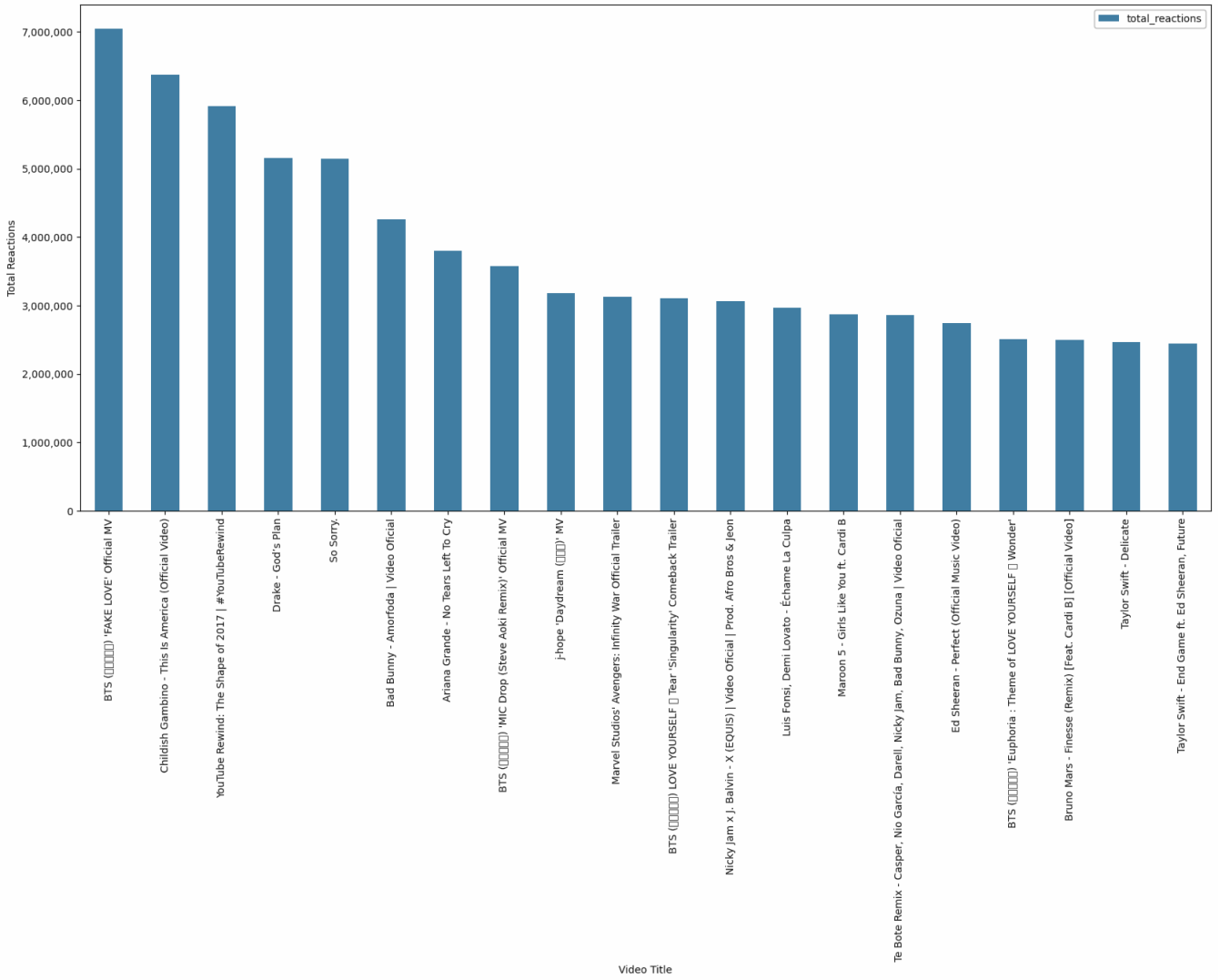
```
+-----+-----------+-------------+--------------------+--------------------+-------------+-------------------+--------------------+--------+-------+--------+-------------+------
|  _c0|   video_id|trending_date|               title|       channel_title|  category_id|       publish_time|                tags|   views|  likes|dislikes|comment_count|
+-----+-----------+-------------+--------------------+--------------------+-------------+-------------------+--------------------+--------+-------+--------+-------------+------
|28358|_I_D_8Z4sJE|   2018-04-07|     Nicky Jam x J. Ba...|          NickyJamTV|        Music|2018-03-02 05:00:19|   "Bad Bunny|""Amor...|424538912|2818771|  149275|        99380|    D
|34381|9jI-z9QN6g8|   2018-05-18|     Te Bote Remix - C...|       Flow La Movie|        Music|2018-04-11 22:00:00|   "Te Bote|""Te Bot...|337621571|2581961|  166549|       113564|    F
|25498|kLpH1nSLJSs|   2018-03-23|     Bad Bunny - Amorf...|           Bad Bunny|        Music|2018-02-15 00:00:03|   "Bad|""Bunny""|""...|328860380|3823879|  215530|       225216|    B
|21115|wfWkmURBNv8|   2018-03-01|     Ozuna x Romeo San...|               Ozuna|        Music|2018-01-30 15:00:05|   "Ozuna|""Reggaeto...|288811992|1618180|  151147|        67506|    D
|38294|VYOjWnS4cMY|   2018-06-11|     Childish Gambino ...| ChildishGambinoVEVO|        Music|2018-05-06 04:00:07|   "Childish Gambino...|259721696|5444541|  379862|       553371|    "
|25888|xpVfcZ0ZcFM|   2018-03-25|       Drake - God’s Plan|          DrakeVEVO|        Music|2018-02-17 05:00:01|   "Drake new music|...|258164991|4737873|  117198|       301756|    6
|35632|ffxKSjUwKdU|   2018-05-26|     Ariana Grande - N...|    ArianaGrandeVevo|        Music|2018-04-20 04:00:03|   "Ariana|""Grande"...|208876887|3394437|  150086|       259613|    N
|35635|zEf423kYfqk|   2018-05-26|     Becky G, Natti Na...|          BeckyGVEVO|        Music|2018-04-20 10:40:51|   "Becky G|""Natti ...|200862743|1668418|  142569|        97826|    B
| 8186|FlsCjmMhFmw|   2017-12-24|     YouTube Rewind: T...|   YouTube Spotlight|Entertainment|2017-12-06 17:58:51|   "Rewind|""Rewind ...|169884583|3312868| 1753274|       845233|    Y
|18130|sGIm0-dQd8M|   2018-02-14|     Dura - Daddy Yank...|       Daddy Yankee|        Music|2018-01-18 22:32:49|   "daddy yankee reg...|167456025|1633407|   74005|        65395|    D
| 3587|TyHvyGVs42U|   2017-12-01|     Luis Fonsi, Demi ...|       LuisFonsiVEVO|        Music|2017-11-17 05:00:01|   "Luis|""Fonsi""|"...|143408235|2686169|  137938|       144217|    "
| 2794|2Vv-BfVoq4g|   2017-11-27|     Ed Sheeran - Perf...|          Ed Sheeran|        Music|2017-11-09 11:04:14|   "edsheeran|""ed s...|138578860|2584773|   49428|       113639|    ¿
|34536|M4ZoCHID9GI|   2018-05-19|     The Weeknd - Call...|       TheWeekndVEVO|        Music|2018-04-12 16:00:04|   "The|""Weeknd""|"...|138535053|1493942|   44048|        57119|    C
|36797|Ck4xHocysLw|   2018-06-02|     Ozuna - Única (V...|               Ozuna|        Music|2018-04-26 17:52:13|   "Ozuna|""Reggaeto...|137081637| 824296|   44012|        27109|    DE
|36541|7C2z4GqqS5E|   2018-06-01|BTS (방탄소년단) 'FAKE...|             ibighit|        Music|2018-05-18 09:00:02|"BIGHIT|""빅히트""|"...|123010920|5613827|  206892|      1228655|BTS (방
|29100|tCXGJQYZ9JA|   2018-04-17|     Taylor Swift - De...|     TaylorSwiftVEVO|        Music|2018-03-12 01:15:10|   "Taylor Swift|""D...|117270304|2161260|  133428|       175635|    T
|33551|U9BwWKXjVaI|   2018-05-13|     Drake - Nice For ...|          DrakeVEVO|        Music|2018-04-07 02:46:31|   "Drake|""Nice""|"...|106147032|1275948|   52924|        65066|    M
|31095|au2n7VVGv_c|   2018-04-28|     Post Malone - Psy...|      PostMaloneVEVO|        Music|2018-03-23 04:00:00|   "Post|""Malone""|...|105629911|1340938|   41651|        45784|    P
| 6191|6ZfuNTqbHE8|   2017-12-14|     Marvel Studios' A...|Marvel Entertainment|Entertainment|2017-11-29 13:26:24|   "marvel|""comics"...|100672931|2701353|   56313|       368739|    T
|30359|fGqdIPer-ms|   2018-04-24|     Migos - Walk It T...|           MigosVEVO|        Music|2018-03-18 21:49:53|   "Migos|""Walk It"...|100159686|1339372|   67821|        73838|    L
+-----+-----------+-------------+--------------------+--------------------+-------------+-------------------+--------------------+--------+-------+--------+-------------+------
only showing top 20 rows
```

b) Triển khai phân tích.

- Bước 1: Sắp xếp lại dữ liệu theo thứ tự giảm dần về số lượng xem. Sau đó in ra 20 video có lượng người xem cao nhất và vẽ biểu đồ.
- Bước 2: Tạo thêm cột ‘total_reactions = “likes” + “dislikes” + “comment_count” để xem tổng số lượng tương tác. Sau đó, tạo một bảng dữ liệu mới để dễ dàng theo dõi 20 videos có tổng số lượt tương tác cao nhất.
- Bước 3, 4, 5: Thực hiện tạo bảng dữ liệu mới tương ứng với 3 hạng mục: “likes”, “dislikes” và “total_comments”, in và tạo đồ thị cho 20 hạng mục dữ liệu có số liệu cao nhất của từng hạng mục.

2) Phân tích.

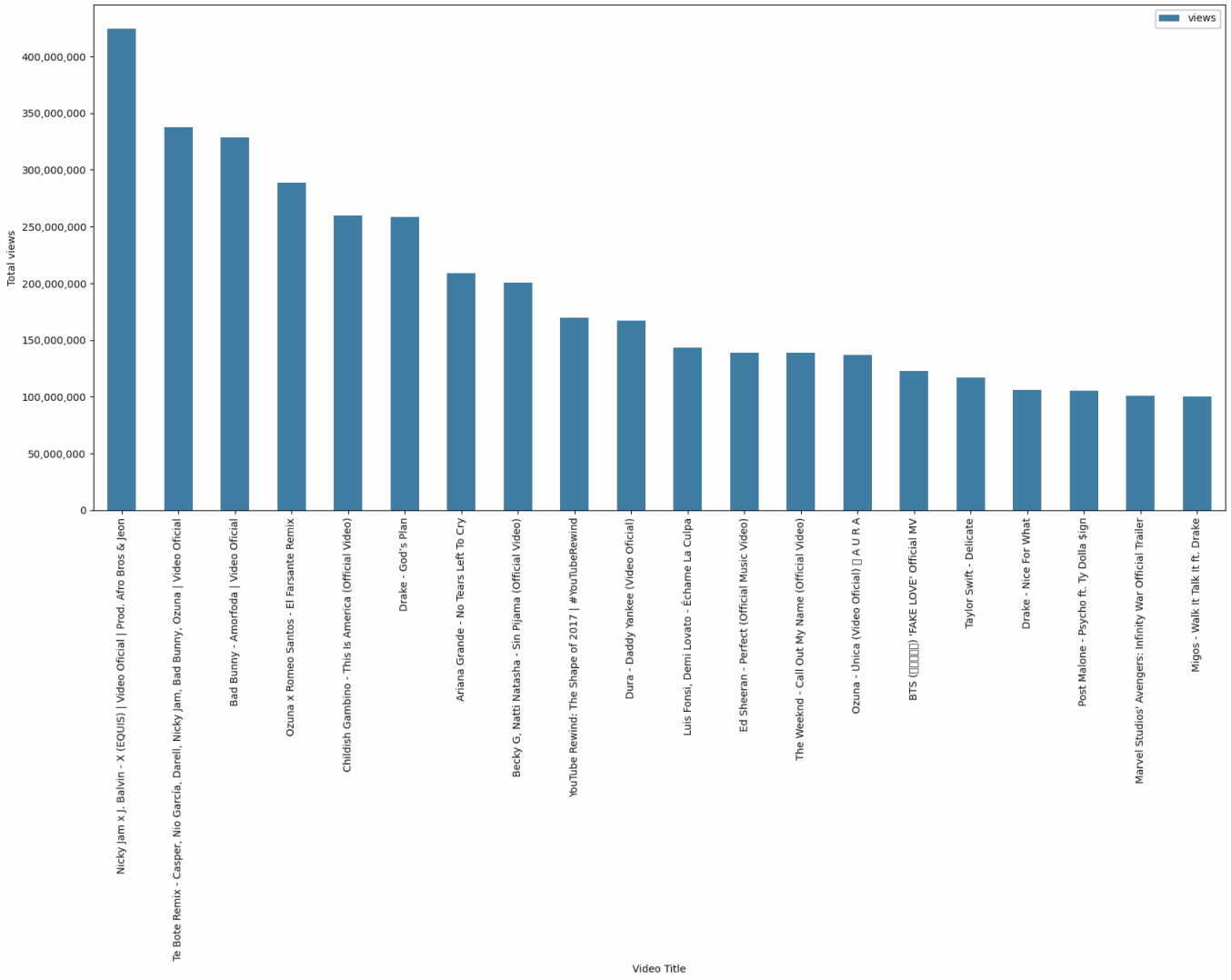
a) Video có nhiều lượt tương tác nhất.



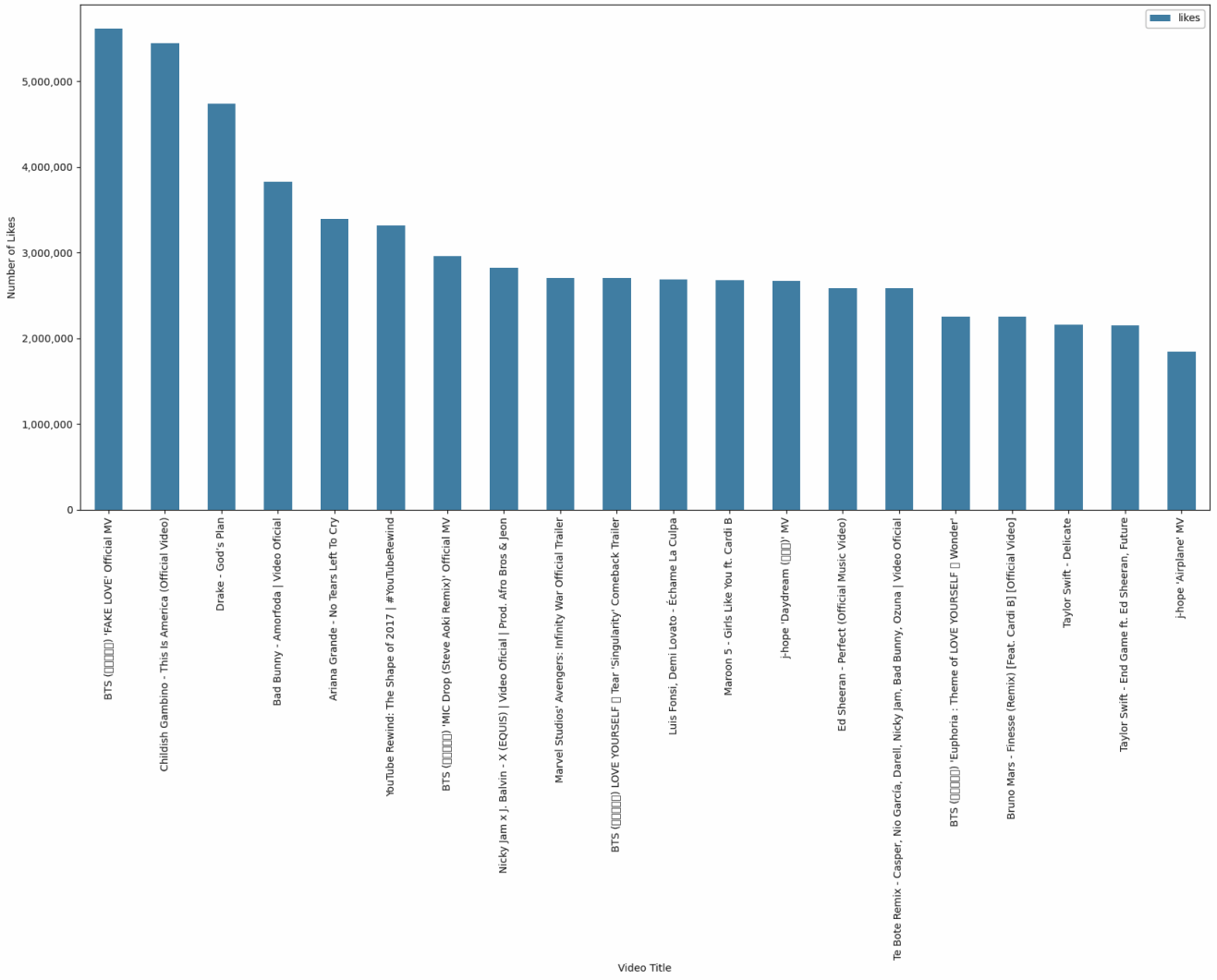
- Những video có lượng tương tác lớn nhất thuộc về những kênh của người, ban nhạc, tổ chức nổi tiếng như BTS, Childish Gambino, YouTube, Drake,... Đây là điều dễ hiểu khi lượng fan của những kênh này rất lớn, với kênh nhỏ nhất là 6.8 triệu đăng ký.

b) Video có nhiều views, likes và dislikes nhất.

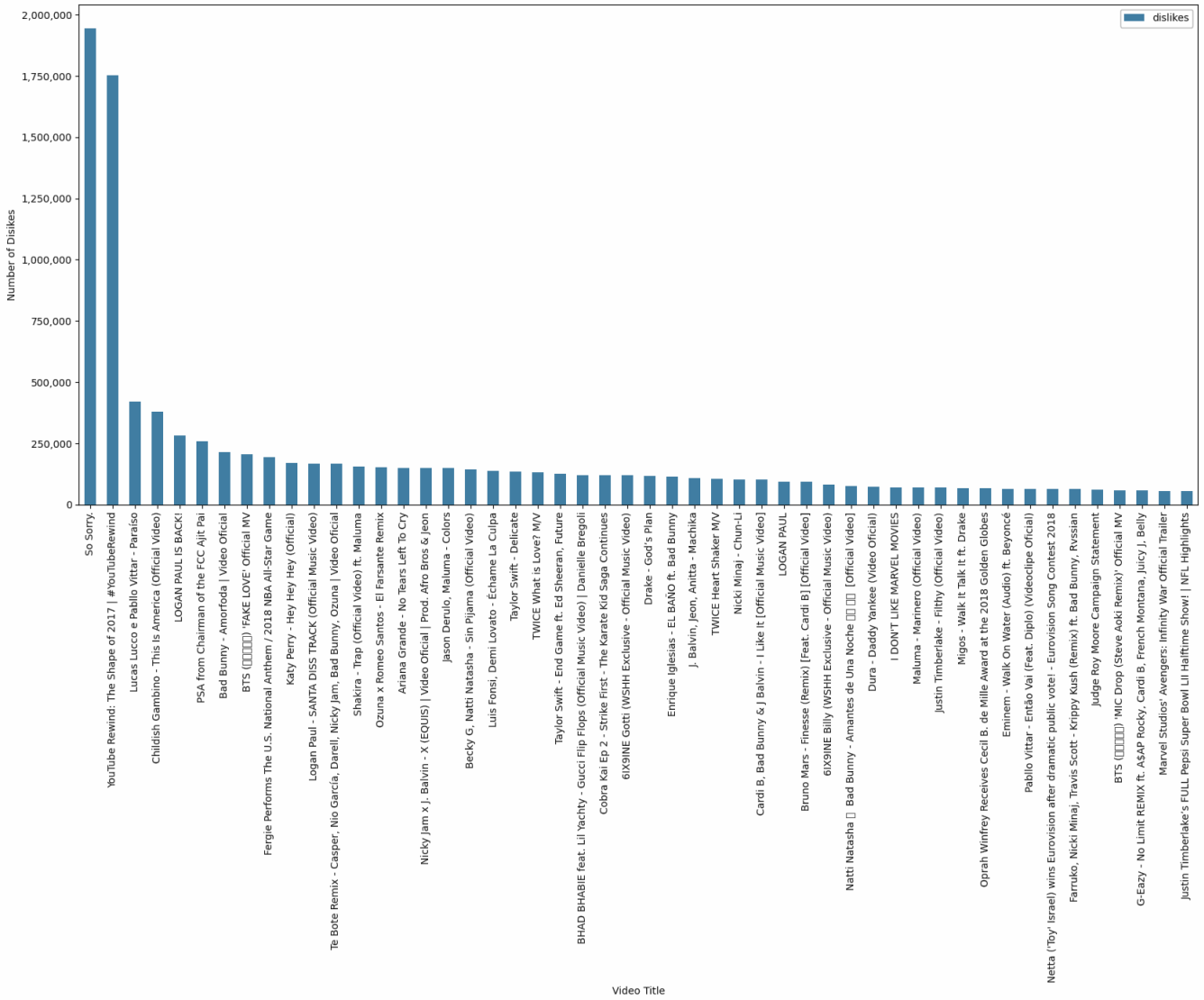
- Views.


views
Total views
400,000,000
350,000,000
300,000,000
250,000,000
200,000,000
150,000,000
100,000,000
50,000,000
0
Nicky Jam x J. Balvin - X (EQUIS) | Video Oficial | Prod. Afro Bros & Jeon
Te Bote Remix - Casper, Nio Garcia, Darell, Nicky Jam, Bad Bunny, Ozuna | Video Oficial
Bad Bunny - Amorfoda | Video Oficial
Ozuna x Romeo Santos - El Farsante Remix
Childish Gambino - This Is America (Official Video)
Drake - God's Plan
Ariana Grande - No Tears Left To Cry
Becky G, Natti Natasha - Sin Pijama (Official Video)
YouTube Rewind: The Shape of 2017 | #YouTubeRewind
Dura - Daddy Yankee (Video Oficial)
Luis Fonsi, Demi Lovato - Échame La Culpa
Ed Sheeran - Perfect (Official Music Video)
The Weeknd - Call Out My Name (Official Video)
Ozuna - Única (Video Oficial) A U R A
BTS ( ) 'FAKE LOVE' Official MV
Taylor Swift - Delicate
Drake - Nice For What
Post Malone - Psycho ft. Ty Dolla $ign
Marvel Studios' Avengers: Infinity War Official Trailer
Migos - Walk It Talk It ft. Drake
Video Title

- Likes.


likes
Number of Likes
5,000,000
4,000,000
3,000,000
2,000,000
1,000,000
0
BTS (□□□□□) 'FAKE LOVE' Official MV
Childish Gambino - This Is America (Official Video)
Drake - God's Plan
Bad Bunny - Amorfoda | Video Oficial
Ariana Grande - No Tears Left To Cry
YouTube Rewind: The Shape of 2017 | #YouTubeRewind
BTS (□□□□□) 'MIC Drop (Steve Aoki Remix)' Official MV
Nicky Jam x J. Balvin - X (EQUIS) | Video Oficial | Prod. Afro Bros & Jeon
Marvel Studios' Avengers: Infinity War Official Trailer
BTS (□□□□□) 'LOVE YOURSELF □ Tear 'Singularity' Comeback Trailer
Luis Fonsi, Demi Lovato - Échame La Culpa
Maroon 5 - Girls Like You ft. Cardi B
j-hope 'Daydream (□□)' MV
Ed Sheeran - Perfect (Official Music Video)
Te Bote Remix - Casper, Nio García, Darell, Nicky Jam, Bad Bunny, Ozuna | Video Oficial
BTS (□□□□□) 'Euphoria : Theme of LOVE YOURSELF □ Wonder'
Bruno Mars - Finesse (Remix) [Feat. Cardi B] [Official Video]
Taylor Swift - Delicate
Taylor Swift - End Game ft. Ed Sheeran, Future
j-hope 'Airplane' MV
Video Title

- Dislikes.



- Top 5 những video nhiều likes và views nhất đều là âm nhạc, và trong top 5 video nhiều dislike nhất thì 2 là âm nhạc, và 3 là entertainment. Điều này hợp lý với phân tích *Music* và *Entertainment* là hai chủ đề có nhiều video trend và nhiều kênh YouTube làm nhất ở trên. Nó cũng đồng thời chứng minh cho nhận định người dùng YouTube nghe nhạc và xem video giải trí nhiều nhất, và việc nó trụ lại top trending lâu như vậy là do nó có rất nhiều người xem và tương tác, khiến cho YouTube đề xuất video tới nhiều người dùng hơn.

CHƯƠNG 6 - KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

1) Kết luận.

- Big Data là thách thức được đặt ra cho các tổ chức, doanh nghiệp trong thời đại số. Một khi làm chủ được dữ liệu lớn thì họ sẽ có cơ hội thành công và dẫn đầu trong bối cảnh cạnh tranh ngày nay, thế giới thì sẽ được hưởng lợi hơn từ việc trích xuất thông tin một cách chính xác hơn, hữu ích hơn với chi phí thấp hơn. Đối với đề tài **YouTube Data Analysis**, nhóm đã sử dụng Apache Spark để xử lý và phân tích dữ liệu từ YouTube, tập trung vào các yếu tố như chủ đề, lượt xem, lượt thích, bình luận, và thời gian đăng tải. Kết quả từ phân tích đã cho một cái nhìn tổng quan về hành vi người dùng, giúp dự đoán các xu hướng phổ biến và hỗ trợ cho chiến lược tối ưu hóa nội dung.

2) Hướng phát triển.

- Với bộ dữ liệu lớn, đa dạng thông tin như đã được phân tích trong đề tài, nhóm nhận thấy tiềm năng áp dụng công nghệ học máy nhằm đưa ra các dự đoán, hỗ trợ các nhà sáng tạo nội dung và người dùng nâng cao trải nghiệm và năng suất.
- Ví dụ như hệ thống gợi ý video, chức năng gợi ý gán nhãn (tags) dựa vào mô tả và tiêu đề video, mô hình dự đoán lượt view và ngày trending, mô hình phân cụm cảm xúc của video dựa vào comments và mô tả.

NHIỆM VỤ CỦA CÁC THÀNH VIÊN

Họ tên và MSSV	Công việc
Trịnh Minh Hiếu - 22022536	● Tìm hiểu tổng quan, đặc trưng, ứng dụng của big data. ● Tìm hiểu tổng quan về Spark. ● Chỉnh sửa, đồng bộ code. ● Tổng hợp báo cáo riêng và viết báo cáo chung. ● Đưa ra kết luận. ● Thuyết trình.
Trần Hồng Đăng - 22022646	● Code phân tích dữ liệu theo thời gian (Time). ● Viết báo cáo riêng và phân tích dữ liệu thời gian. ● Làm slide.
Nguyễn Trường Huy - 22022509	● Tiền xử lý dữ liệu. ● Code phân tích dữ liệu theo chủ đề (Category). ● Viết báo cáo riêng và phân tích dữ liệu chủ đề. ● Làm slide.
Trần Kim Thành - 22022532	● Code phân tích dữ liệu theo tương tác (Views, Likes, Dislikes). ● Viết báo cáo riêng và phân tích dữ liệu lượt xem. ● Làm slide.