TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

VIỆN TRÍ TUỆ NHÂN TẠO

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

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

ĐỀ TÀI

PYSPARK TRONG PHÂN TÍCH DỮ LIỆU

Nhóm sinh viên thực hiện:

1. Đinh Duy Bách - 22022531
2. Lê Hữu Đức - 22022535

Giảng viên hướng dẫn: TS. Trần Hồng Việt

ThS.Ngô Minh Hương

HÀ NỘI, 12/2024

MỞ ĐẦU

Trong thời đại công nghệ số hiện nay, dữ liệu đang trở thành tài sản quý giá nhất đối với các doanh nghiệp, tổ chức, và thậm chí cả cá nhân. Hằng ngày, lượng dữ liệu khổng lồ được tạo ra từ các hoạt động trực tuyến như giao dịch thương mại điện tử, tương tác trên mạng xã hội, cảm biến từ các thiết bị IoT, và nhiều nguồn khác. Từ dữ liệu y tế hỗ trợ chẩn đoán bệnh đến dữ liệu giao thông giúp quản lý đô thị, sự bùng nổ của thông tin đang thay đổi cách chúng ta sống, làm việc và ra quyết định.

Tuy nhiên, với khối lượng, tốc độ và sự đa dạng của dữ liệu hiện nay, các công cụ và phương pháp truyền thống không còn đủ khả năng để quản lý và khai thác hiệu quả. Chính vì vậy, khái niệm Big Data đã xuất hiện như một bước tiến mới trong công nghệ xử lý và phân tích thông tin. Big Data không chỉ mang lại cơ hội để khám phá những hiểu biết sâu sắc, mà còn giúp các tổ chức đạt được lợi thế cạnh tranh và tối ưu hóa hoạt động.

Bài báo cáo này tập trung vào việc giới thiệu khái niệm Big Data, các đặc điểm nổi bật, và vai trò của nó trong nhiều lĩnh vực khác nhau. Đặc biệt, báo cáo sẽ đi sâu vào ứng dụng của công nghệ Apache Spark và PySpark trong xử lý dữ liệu lớn, nhằm minh họa cách Big Data có thể được khai thác một cách hiệu quả và mang lại giá trị thực tiễn.

Báo cáo gồm 3 chương:

Chương 1: Tổng quan về dữ liệu lớn.

Chương 2: Ứng dụng PySpark trong phân tích dữ liệu

Chương 3: Kết luận và hướng phát triển.

MỤC LỤC

CHƯƠNG 1: TỔNG QUAN VỀ DỮ LIỆU LỚN

1.1 Định nghĩa.

Dữ liệu lớn là thuật ngữ chỉ một lượng dữ liệu khổng lồ và phức tạp mà các công cụ quản lý dữ liệu truyền thống không thể xử lý hiệu quả. Dữ liệu lớn thường đến từ nhiều nguồn khác nhau và được đặc trưng bởi các yếu tố 5V:

- Volume (Khối lượng): Lượng dữ liệu rất lớn, có thể từ terabyte (TB) đến petabyte (PB).
- Velocity (Tốc độ): Dữ liệu được tạo ra và cập nhật liên tục với tốc độ nhanh.
- Variety (Đa dạng): Dữ liệu đến từ nhiều nguồn với các định dạng khác nhau, như văn bản, hình ảnh, video, âm thanh, và dữ liệu phi cấu trúc.
- Veracity (Độ tin cậy): Chất lượng và tính chính xác của dữ liệu có thể không đồng nhất.
- Value (Giá trị): Giá trị thông tin mang lại

1.2 Nguồn gốc:

Dữ liệu lớn đến từ nhiều nguồn khác nhau trong cuộc sống hàng ngày và hoạt động kinh doanh, bao gồm:

- Truyền thông xã hội: Facebook, Twitter, Instagram, YouTube, v.v.
- Internet of Things (IoT): Cảm biến, thiết bị thông minh, hệ thống GPS.
- Dữ liệu giao dịch: Hoá đơn bán lẻ, giao dịch ngân hàng, thương mại điện tử.
- Dữ liệu khoa học: Dữ liệu y tế, dữ liệu nghiên cứu, thí nghiệm khoa học.
- Dữ liệu hình ảnh và âm thanh: Camera an ninh, video giám sát, hình ảnh vệ tinh.

1.3 Công nghệ hỗ trợ dữ liệu lớn:

- Hệ thống lưu trữ phân tán: Hadoop, Apache Spark.
- Cơ sở dữ liệu NoSQL: MongoDB, Cassandra.
- Phân tích dữ liệu: Apache Hive, Tableau, Power BI.
- Trí tuệ nhân tạo và máy học: TensorFlow, PyTorch, Scikit-learn

1.4 Tổng quan về Apache Spark

Apache Spark là một nền tảng xử lý dữ liệu mã nguồn mở, được thiết kế để xử lý dữ liệu lớn (Big Data) với tốc độ cao và khả năng phân tán. Spark nổi bật nhờ khả năng xử lý dữ liệu trong bộ nhớ (in-memory computing), giúp tăng tốc độ so với các hệ thống xử lý truyền thống như Hadoop MapReduce.

Các thành phần chính của Apache Spark:
Apache Spark có một hệ sinh thái hoàn chỉnh bao gồm các thành phần sau:

- Spark Core: Thành phần lõi, chịu trách nhiệm xử lý dữ liệu phân tán, quản lý bộ nhớ và lịch trình công việc.
- Spark SQL: Cung cấp giao diện xử lý dữ liệu dạng bảng và truy vấn SQL. Hỗ trợ tích hợp với các hệ thống dữ liệu như Hive, Cassandra, và MySQL.
- Spark Streaming: Hỗ trợ xử lý dữ liệu thời gian thực, lý tưởng cho các ứng dụng phân tích luồng dữ liệu trực tiếp.
- MLlib (Machine Learning Library): Thư viện học máy phân tán, cung cấp các thuật toán

như phân cụm, hồi quy, và phân loại.

- ❖ GraphX: Công cụ xử lý đồ thị, hỗ trợ phân tích dữ liệu dạng mạng như mạng xã hội hoặc đồ thị đường dẫn.

1.5 Tổng quan về PySpark

PySpark là một giao diện lập trình Python của Apache Spark, giúp các nhà phát triển và nhà khoa học dữ liệu dễ dàng sử dụng Spark để xử lý và phân tích dữ liệu lớn. PySpark tận dụng sức mạnh của Spark trong việc xử lý dữ liệu phân tán và tính toán nhanh chóng, đồng thời mang lại sự linh hoạt và thân thiện nhờ ngôn ngữ Python.

PySpark được phát triển để tích hợp với các thư viện Python phổ biến như Pandas, NumPy, Matplotlib, và Scikit-learn, cho phép người dùng dễ dàng kết hợp Spark với các công cụ khoa học dữ liệu quen thuộc.

Cấu trúc của PySpark

PySpark được xây dựng dựa trên các thành phần cốt lõi của Apache Spark, cung cấp các API Python để làm việc với:

- ❖ RDD (Resilient Distributed Dataset): Mô hình dữ liệu phân tán chịu lỗi, nền tảng của Spark.
- ❖ DataFrame: Một cấu trúc dữ liệu phổ biến trong PySpark, tương tự như DataFrame trong Pandas, hỗ trợ thao tác dữ liệu dạng bảng.
- ❖ SQL API: Cho phép viết các truy vấn SQL để phân tích dữ liệu lớn.
- ❖ Machine Learning API: Hỗ trợ các thuật toán học máy phân tán qua thư viện MLlib.

1.6 Các ứng dụng của PySpark trên thực tế

- Xử lý dữ liệu thời gian thực: Một trong những tính năng mạnh mẽ nhất của Spark là khả năng xử lý luồng dữ liệu thời gian thực bằng Spark Streaming. Khả năng này được sử dụng rộng rãi trong các ứng dụng như phát hiện gian lận trong giao dịch tài chính, giám sát dữ liệu cảm biến để bảo trì dự đoán trong sản xuất và phân tích luồng phương tiện truyền thông xã hội cho bài toán sentiment analysis.

- Học máy và AI: Spark cung cấp một thư viện tích hợp có tên là MLlib, giúp đơn giản hóa việc triển khai các thuật toán học máy như phân loại, hồi quy, phân cụm và lọc cộng tác. Điều này hữu ích trong nhiều ngành công nghiệp, từ hệ thống đề xuất trong thương mại điện tử (như Amazon hoặc Netflix) đến phân tích dự đoán trong chăm sóc sức khỏe. Trong thương mại điện tử, các mô hình học máy được xây dựng bằng Spark có thể đề xuất các sản phẩm được cá nhân hóa dựa trên hành vi trước đây của khách hàng.

- Phân tích dữ liệu lớn: Khả năng xử lý và phân tích các tập dữ liệu lớn nhanh chóng của Spark đã cách mạng hóa các ngành công nghiệp dựa vào dữ liệu lớn. Ví dụ, các nhà bán lẻ phân tích các mô hình mua hàng của khách hàng để tối ưu hóa hàng tồn kho và cá nhân hóa các chiến lược tiếp thị. Trong ngành viễn thông, Spark được sử dụng để phân tích hồ sơ cuộc gọi, lưu lượng mạng và hành vi của khách hàng để cải thiện dịch vụ cung cấp và dự đoán tỷ lệ khách hàng hủy dịch vụ. Khả năng kết hợp dữ liệu từ nhiều nguồn khác nhau, chẳng hạn như tệp nhật ký, cơ sở dữ liệu quan hệ và nguồn cấp dữ liệu thời gian thực, khiến Spark trở thành một công cụ thiết yếu để phân tích dữ liệu lớn.

- Kho dữ liệu và quy trình ETL: Spark thường được sử dụng như một giải pháp thay thế nhanh hơn cho các công cụ ETL (Trích xuất, Chuyển đổi, Tải) truyền thống để xử lý khối lượng dữ liệu lớn. Công cụ này rất hiệu quả trong việc dọn dẹp, chuyển đổi và tích hợp dữ liệu từ nhiều nguồn, sau đó có thể được lưu trữ trong kho dữ liệu để phân tích thêm. Các doanh nghiệp sử dụng Spark để xử lý và phân tích dữ liệu có cấu trúc, bán cấu trúc và không có cấu trúc, cho phép họ khám phá ra những thông tin chi tiết có giá trị nhanh hơn.

- Xử lý đồ thị: Thư viện GraphX của Spark được sử dụng cho các ứng dụng xử lý đồ thị, liên quan đến việc phân tích mạng và các mối quan hệ. Điều này đặc biệt có lợi trong phân tích

mạng xã hội, phát hiện gian lận và hệ thống đề xuất. Ví dụ: Spark có thể phân tích dữ liệu phương tiện truyền thông xã hội để tìm cộng đồng hoặc xác định những người có ảnh hưởng và trong viễn thông, nó có thể được sử dụng để phát hiện gian lận bằng cách phân tích biểu đồ cuộc gọi.

- Phân tích không gian địa lý: Khả năng xử lý dữ liệu không gian địa lý của Spark cung cấp một công cụ tuyệt vời cho các dịch vụ dựa trên vị trí, chẳng hạn như trong các ứng dụng chia sẻ chuyến đi, thành phố thông minh,…. Spark có thể xử lý các tập dữ liệu lớn về thông tin địa lý, chẳng hạn như tuyến đường giao hàng hoặc dữ liệu thời tiết, để tối ưu hóa các tuyến đường cho giao thông hoặc quy hoạch đô thị.

- Dịch vụ tài chính: Trong lĩnh vực tài chính, Spark được sử dụng để phân tích rủi ro, phát hiện gian lận, giao dịch bằng thuật toán và phân tích khách hàng. Ví dụ: các công ty đầu tư có thể sử dụng Spark để giao dịch với tần suất cao, xử lý dữ liệu thị trường theo thời gian thực để xác định các cơ hội sinh lời. Khả năng scale của Spark cũng giúp các tổ chức tài chính phân tích khối lượng giao dịch lớn một cách nhanh chóng để xác định các hoạt động đáng ngờ.

- Chăm sóc sức khỏe và y tế: Lượng dữ liệu từ ngành y tế đang tăng lên nhanh chóng nhờ vào sự xuất hiện của các thiết bị và công nghệ tân tiến, như các máy đo chỉ số sinh trắc thời gian thực, máy MRI hay các thiết bị đeo phổ thông. Phân tích lượng dữ liệu này sẽ giúp các bác sĩ và các nhà nghiên cứu phát hiện ra những thông tin hữu ích cho việc chăm sóc bệnh nhân. Trong chăm sóc sức khỏe, Spark có thể được sử dụng để phân tích dữ liệu bệnh nhân nhằm dự đoán các đợt bùng phát dịch bệnh hoặc hiệu quả của một số phương pháp điều trị nhất định.

- Giải trí và phương tiện truyền thông: Spark cung cấp năng lượng cho các công cụ đề xuất cho các nền tảng như Netflix và Spotify, xử lý lượng lớn dữ liệu tương tác của người dùng để đề xuất nội dung được cá nhân hóa. Spark cũng được sử dụng để phân tích sở thích và hành vi của người dùng nhằm cải thiện chiến lược phân phối nội dung và tương tác. Đối với các công ty truyền thông, Spark cho phép phân phối nội dung nhanh hơn và hiểu biết sâu sắc về tương tác của khán giả, cho phép họ điều chỉnh các dịch vụ của mình hiệu quả hơn.

CHƯƠNG 2: Ứng dụng PySpark trong phân tích dữ liệu

2.1 Xử lý dữ liệu.

Ánh xạ categoryID vào bảng bằng cách sử dụng tệp JSON.

```
[ ]  json_path = '/content/drive/MyDrive/bigdata/data/youtube_data/US_category_id.json'
     with open(json_path, 'r') as file:
         category_data = json.load(file)

     categories = pd.json_normalize(category_data['items'])
     categories = categories[['id', 'snippet.title']]
     categories.columns = ['category_Id', 'category_name']
     categories['category_Id'] = categories['category_Id'].astype(int)

     df = df.merge(categories, on='category_Id', how='left')

     df[['video_id', 'category_Id', 'category_name']].head()
```

	video_id	category_Id	category_name
0	3C66w5Z0ixs	22	People & Blogs
1	M9Pmf9AB4Mo	20	Gaming
2	J78aPJ3VyNs	24	Entertainment
3	kXLn3HkpjaA	10	Music
4	VIUo6yapDbc	26	Howto & Style

Bỏ những cột không cần thiết cho việc phân tích dữ liệu

```
[9] df_pyspark=df_pyspark.drop('video_id','channel_Id','thumbnail_link','comments_disabled','ratings_disabled')
    df_pyspark.show()
```

Thêm 2 cột ngày và giờ riêng biệt từ cột 'publishAt'

```
[10] from pyspark.sql.functions import split, regexp_extract

     df_pyspark = df_pyspark.withColumn("publishedAt_date", regexp_extract("publishedAt", "(.*?) ", 1)) \
                            .withColumn("publishedAt_time", regexp_extract("publishedAt", " (.*)", 1))

     df_pyspark=df_pyspark.drop("publishedAt")
```

Chỉnh lại kiểu dữ liệu

```
[32] from pyspark.sql.types import IntegerType
     df_pyspark = df_pyspark.withColumn("category_Id", df_pyspark["category_Id"].cast(IntegerType()))
     df_pyspark = df_pyspark.withColumn("view_count", df_pyspark["view_count"].cast(IntegerType()))
     df_pyspark = df_pyspark.withColumn("likes", df_pyspark["likes"].cast(IntegerType()))
     df_pyspark = df_pyspark.withColumn("dislikes", df_pyspark["dislikes"].cast(IntegerType()))
     df_pyspark = df_pyspark.withColumn("comment_count", df_pyspark["comment_count"].cast(IntegerType()))
```

```
[33] df_pyspark.printSchema()

root
 |-- title: string (nullable = true)
 |-- channel_Title: string (nullable = true)
 |-- category_Id: integer (nullable = true)
 |-- trending_date: string (nullable = true)
 |-- tags: string (nullable = true)
 |-- view_count: integer (nullable = true)
 |-- likes: integer (nullable = true)
 |-- dislikes: integer (nullable = true)
 |-- comment_count: integer (nullable = true)
 |-- description: string (nullable = true)
 |-- category_name: string (nullable = true)
 |-- publishedAt_date: date (nullable = true)
 |-- publishedAt_time: string (nullable = true)
 |-- year: integer (nullable = true)
 |-- month: integer (nullable = true)
```

2.2 Làm sạch dữ liệu

Loại bỏ giá trị NULL

```
[12] df_pyspark = df_pyspark.na.drop()
```

Kết quả cuối cùng

title	channel_Title	category_Id	trending_date	tags	view_count	likes	dislikes	comment_count	description	category_name	publishedAt_date	publishedAt_time	year	month
Apex Legends | St...	Apex Legends	20	2020-08-12	Apex Legends|Apex...	2381688	146739	2794	16549	While running her...	Gaming	2020-08-11	17:00:10	2020	8
I left youtube fo...	jacksepticeye	24	2020-08-12	jacksepticeye|fun...	2038853	353787	2628	40221	I left youtube fo...	Entertainment	2020-08-11	16:34:06	2020	8
Ultimate DIY Home...	Mr. Kate	26	2020-08-12	The LaBrant Famil...	1123889	45802	964	2196	Transforming The ...	Howto & Style	2020-08-11	15:10:05	2020	8
I Haven't Been Ho...	Professor Live	24	2020-08-12	Professor injury|...	949491	77487	746	7506	Subscribe To My C...	Entertainment	2020-08-11	20:00:04	2020	8
OUR FIRST FAMILY ...	Les Do Makeup	26	2020-08-12	[None]	470446	47990	440	4558	Hi babygirls! Th...	Howto & Style	2020-08-12	00:17:41	2020	8
CGP Grey was WRONG	CGP Grey	27	2020-08-12	cgpgrey|education...	1050143	89190	854	6455	- What Was TEKOI:...	Education	2020-08-11	17:15:11	2020	8
SURPRISING MY DAD...	Louie's Life	24	2020-08-12	surprising|dad|fa...	1402687	95694	2158	6613	Since I was littl...	Entertainment	2020-08-10	22:26:59	2020	8
i don't know what...	CaseyNeistat	22	2020-08-12	[None]	940036	87111	1860	7052	ssend love to my ...	People & Blogs	2020-08-11	20:24:34	2020	8
Try Not To Laugh ...	Smosh Pit	22	2020-08-12	smosh|smosh pit|s...	591837	44168	409	2652	You know what tim...	People & Blogs	2020-08-11	17:00:31	2020	8
Rainbow Six Siege...	Ubisoft North Ame...	20	2020-08-12	R6|R6S|Siege|New ...	320872	14288	774	2085	"Prepare. Execute...	Gaming	2020-08-11	17:13:53	2020	8
Lil Yachty & Futu...	LilYachtyVEVO	10	2020-08-12	Lil Yachty|Lil Bo...	413372	26440	293	1495	Watch the officia...	Music	2020-08-11	19:00:10	2020	8
When Our Generati...	Kyle Exum	23	2020-08-12	When Our Generati...	921261	124183	1678	16460	500,000 Likes and...	Comedy	2020-08-10	22:33:48	2020	8
Ten Minutes with ...	Tyler Cameron	22	2020-08-12	the bachelor|the ...	105955	4511	69	673	Come hang out me ...	People & Blogs	2020-08-11	22:00:05	2020	8
Kylie Jenner Reac...	HollywoodLife	24	2020-08-12	kylie jenner|kend...	1007540	10102	7932	2763	Kylie Jenner diss...	Entertainment	2020-08-10	18:41:19	2020	8
Our Farm Got Dest...	Cole The Cornstar	22	2020-08-12	farming|family fa...	277338	37533	197	3666	Wind storm, rain,...	People & Blogs	2020-08-11	23:00:06	2020	8
Time to Talk..	Chloe Ting	26	2020-08-12	chloe ting|chloet...	1648441	130147	1425	15773	I talked about so...	Howto & Style	2020-08-11	12:04:40	2020	8
ITZY "Not Shy" M/...	JYP Entertainment	10	2020-08-12	JYP Entertainment...	5999732	714287	15174	31039	ITZY Not Shy M/V[...	Music	2020-08-11	15:00:13	2020	8
Shark Attack Test...	Mark Rober	28	2020-08-12	sharks|sharkweek|...	14684474	544038	15818	33507	I personally got ...	Science & Technology	2020-08-09	16:00:11	2020	8
Honest Trailers |...	Screen Junkies	1	2020-08-12	screenjunkies|scr...	833369	50181	1120	4634	▶▶Subscribe to Sc...	Film & Animation	2020-08-11	17:03:59	2020	8
EXTREME Game of H...	FaZe Rug	24	2020-08-12	faze rug|rug|rugf...	3061467	206840	2646	14934	THIS WAS SO MUCH ...	Entertainment	2020-08-10	17:09:53	2020	8

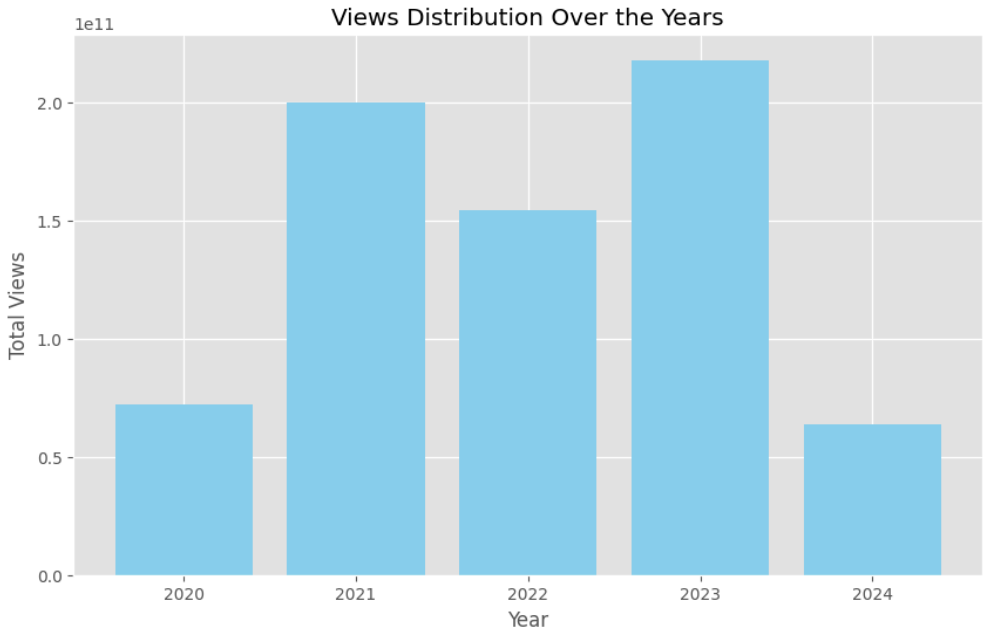
only showing top 20 rows

2.2 Phân tích dữ liệu với Apache Pyspark

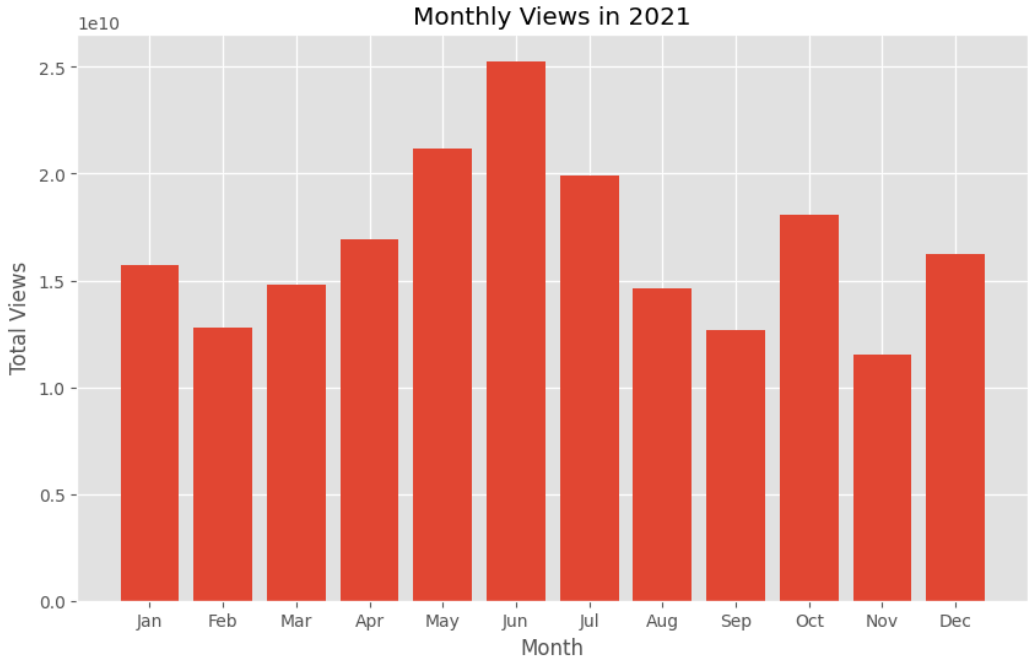
Ở mục này, ta sẽ thực hiện việc phân tích dữ liệu cho bộ dữ liệu về thông số tương tác và một vài thông tin liên quan như mô tả, kênh sản xuất nội dung của các video trên YouTube. Bộ dữ liệu được thu thập trong khoảng từ tháng 8/2020 đến tháng 6/2024. Bộ dữ liệu này không có đầy đủ thông tin của tất cả các video đã được đăng lên YouTube trong khoảng thời gian nói trên do lượng video này là khổng lồ và khó có thể được xử lý hết trong phạm vi dự án này. Bộ dữ liệu được sử dụng chứa thông tin của khoảng gần 300 nghìn video. Do đó, việc phân tích bộ dữ liệu này không mang lại đầy đủ thông tin tuy nhiên cũng phần nào mang lại góc nhìn thống kê về các video được đăng lên YouTube trong khoảng thời gian nói trên.

Apache Spark và các thư viện liên quan được sử dụng cho việc phân tích bộ dữ liệu.

Lượng lượt xem qua từng năm được thống kê với biểu đồ dưới đây. Với bộ dữ liệu này, năm 2023 có lượng lượt xem lớn nhất với hơn 217 tỷ lượt xem.

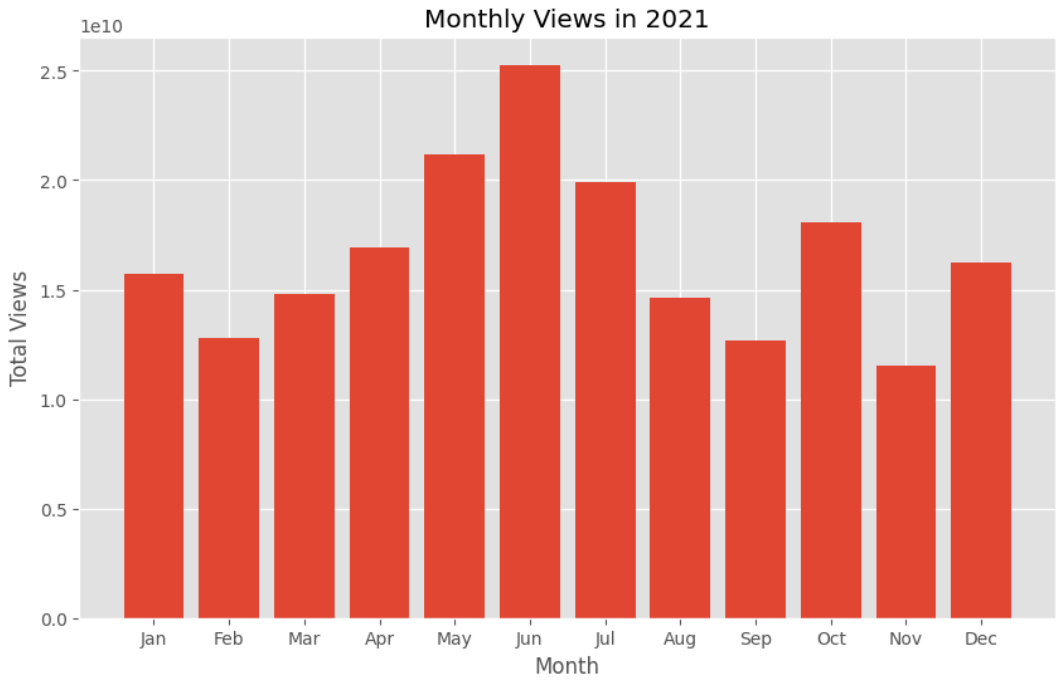


Ta sẽ xem xét cụ thể phân bố lượt xem trong vài năm cụ thể. Trong năm 2021

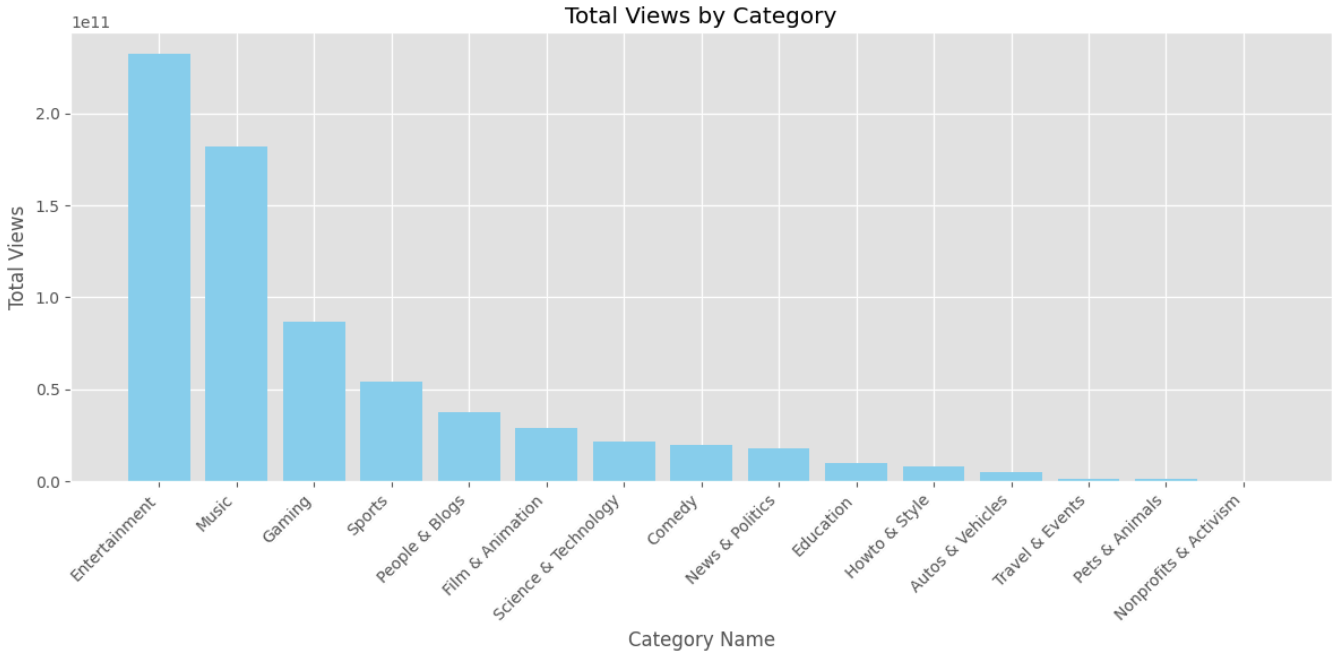


Lượt xem nhiều nhất nằm ở mùa hè và dịp cuối năm. Điều này có thể được giải thích bởi việc trẻ em, lực lượng tiêu thụ nội dung trên YouTube đông đảo, đang trong kỳ nghỉ hè, dẫn đến việc xem YouTube nhiều hơn. Lượng lượt xem tăng vào dịp cuối năm có thể được lý giải bằng việc đây là mùa lễ hội.

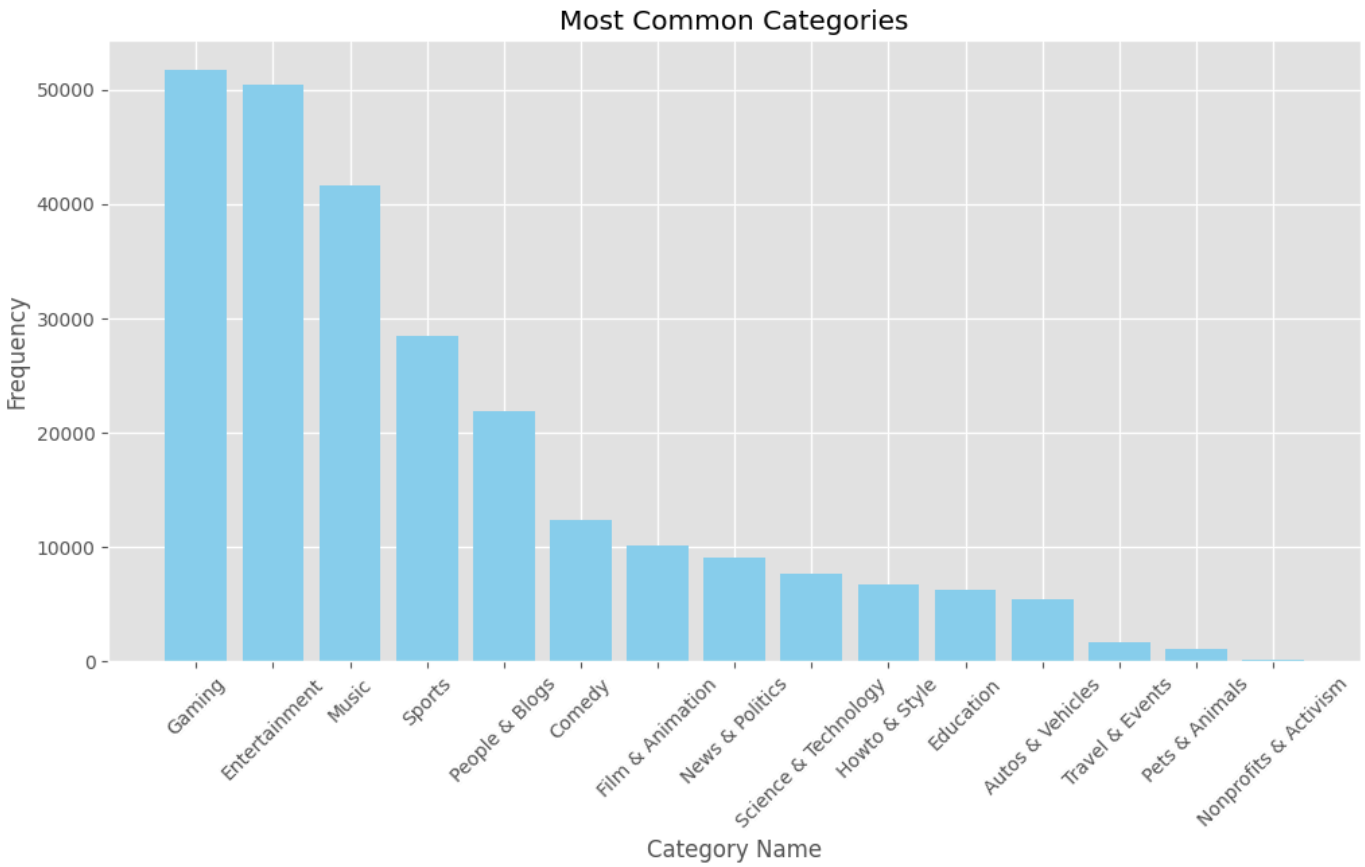
Năm 2022 có lượng lượt xem giữa các tháng đều nhau hơn và cũng không có xu hướng tăng vào mùa hè và mùa lễ hội như năm 2021.



Tiếp theo ta xem xét đến độ phổ biến của các hạng mục nội dung. Thể hiện ở biểu đồ độ phổ biến của loại video với tổng lượt xem

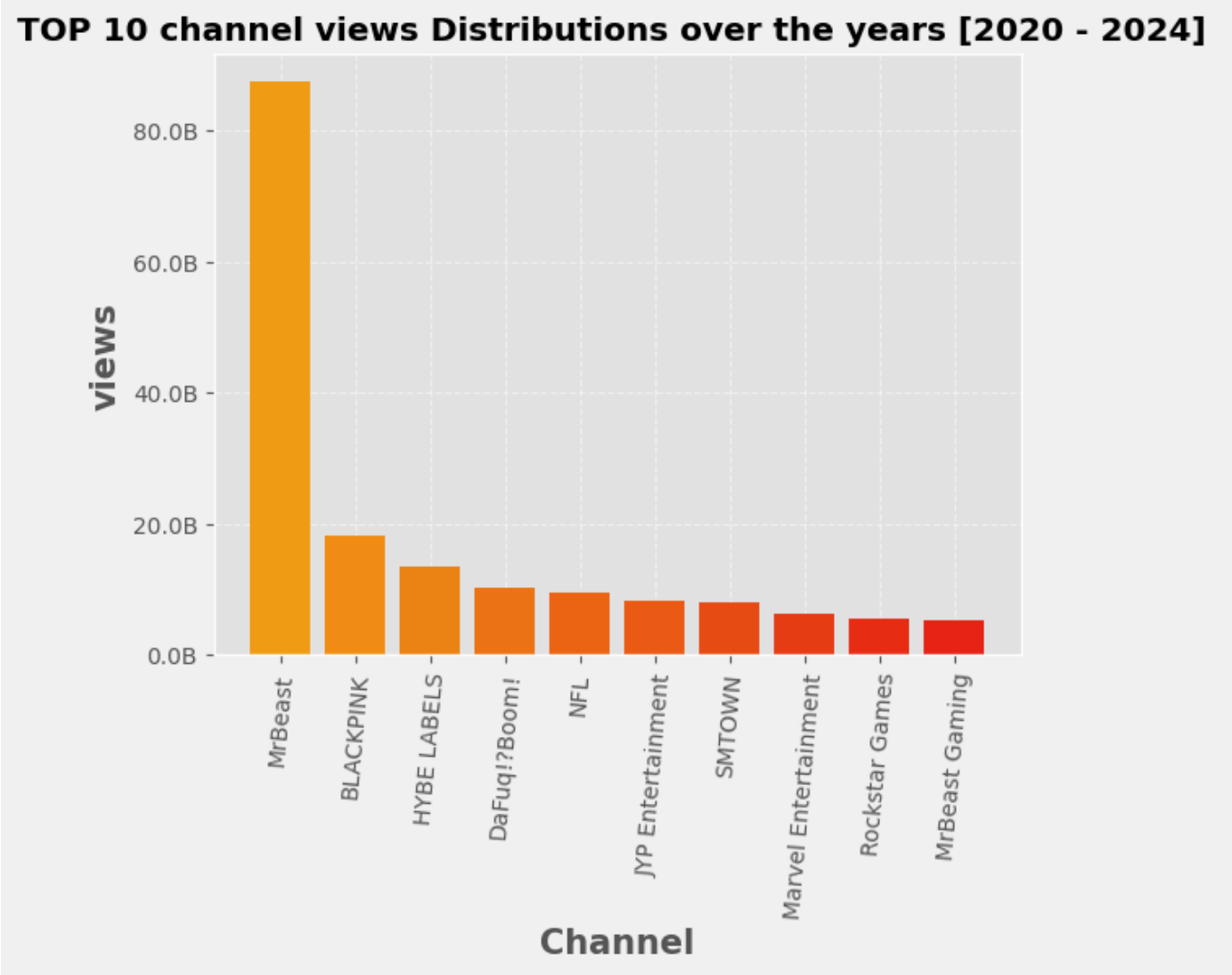


và với tổng lượng video của mỗi thể loại



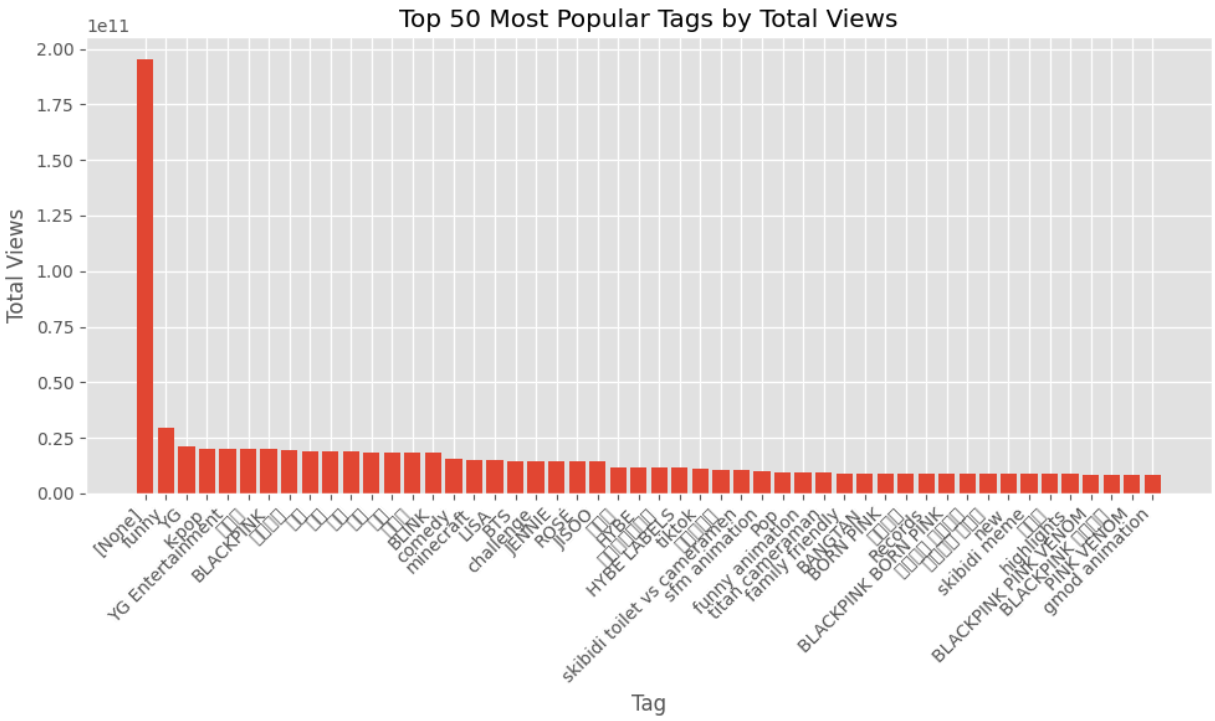
Ta có thể thấy 3 thể loại video là âm nhạc, giải trí và game có mức độ phổ biến lớn nhất, về cả lượng video và lượt xem. Các thể loại video khác như thể thao hay tin tức thời sự cũng có độ phổ biến lớn.

Ta xem xét đến các kênh YouTube phổ biến nhất.



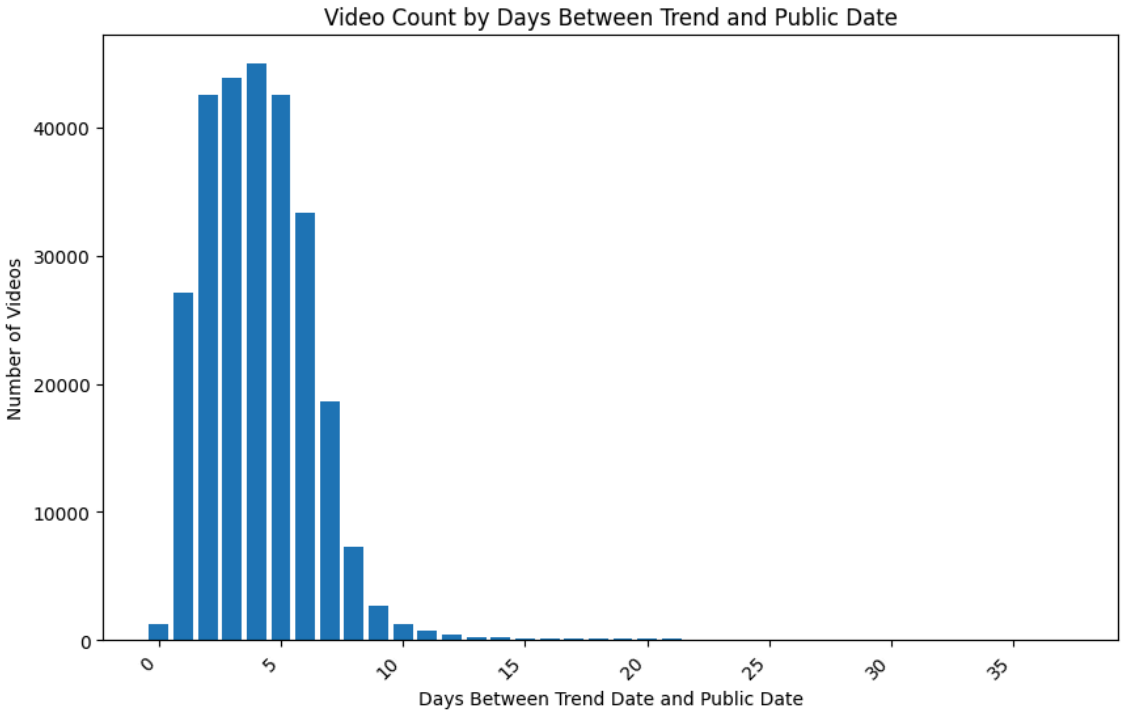
Các kênh này đều có tổng lượng lượt xem trong giai đoạn quan tâm trên 5 tỷ. Đặc biệt có kênh "MrBeast" thu được gần 90 tỷ lượt xem trong giai đoạn này.

Các tag của video cũng dự đoán độ phổ biến của video đó.



Nhờ vào biểu đồ trên, ta có thể thấy các nội dung Kpop có lượng theo dõi rất lớn.

Cuối dùng là phân tích về khoảng thời gian cần thiết để một video có thể trở nên thịnh hành. Ta có thể thấy phần lớn các video trở nên thịnh hành trong khoảng thời gian 1 tuần kể từ ngày đăng.



CHƯƠNG 3: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

3.1 Kết luận.

Big data đã thách thức đặt ra cho cá tổ chức, doanh nghiệp nhiều cơ hội, thách thức và tài sản quý giá. PySpark giúp xử lý dữ liệu lớn trên các cụm máy tính (cluster), hỗ trợ phân chia và tính toán song song, giúp giảm thời gian xử lý so với các công cụ truyền thống.
Đề tài này đã áp dụng PySpark là API ngôn ngữ python của Apache Spark để báo cáo phân tích dữ liệu của Youtube từ năm 2020-2024.

3.2 Hướng phát triển.

Áp dụng kiến thức về Big data, apache spark để cải tiến và xây dựng ứng dụng phân tích dữ liệu lớn hơn , vào nhiều lĩnh vực khác.

Trong quá trình hoàn thành bài tập lớn, nhóm em đã cố gắng tìm hiểu và tham khảo các tài liệu liên quan. Tuy nhiên thời gian có hạn nên nhóm em không tránh khỏi những thiếu sót, rất mong nhận được sự đóng góp ý kiến từ thầy và cô gián để kỹ năng của chúng em ngày càng được hoàn thiện hơn.

NHIỆM VỤ CỦA CÁC THÀNH VIÊN

Họ và tên	**Công việc**
Đinh Duy Bách	+ Làm báo cáo + Tìm hiểu tổng quan về Apache Spark + Phân tích và trực quan hóa dữ liệu
Lê Hữu Đức	+ Làm slide + Tìm hiểu tổng quan về dữ liệu lớn + Thu thập dữ liệu, xử lý dữ liệu và phân tích dữ liệu