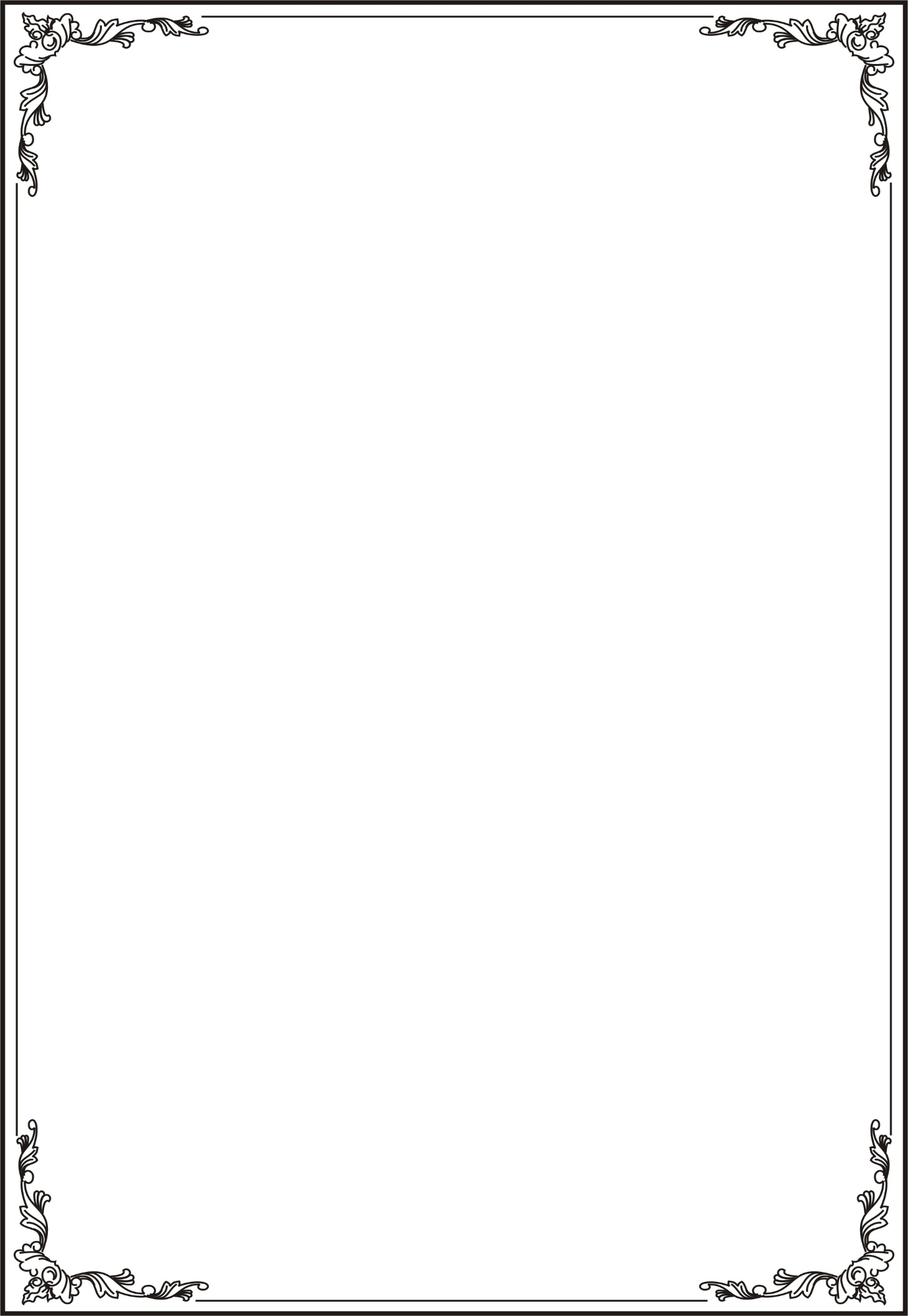
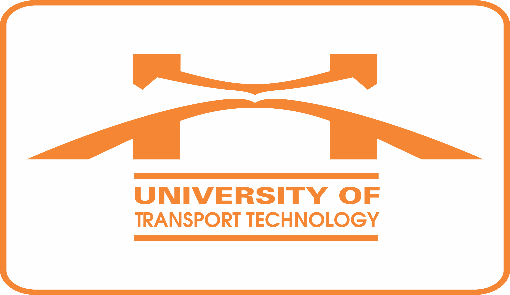
**ĐẠI HỌC CÔNG NGHỆ GIAO THÔNG VẬN TẢI**

**KHOA CÔNG NGHỆ THÔNG TIN**

**\*\*\*\*\*\*\*\*\***

****

**BÁO CÁO** **BÀI TẬP LỚN MÔN HỌC**

**BIG DATA**

***Đề tài: Số 4***

**Nhóm sinh viên thực hiện: Nhóm 1**

**Thành viên nhóm:** Nguyễn Văn Chung

Nguyễn Văn Tú

Nguyễn Văn Đạt

Dương Minh Đạt

Nguyễn Trọng Tấn

**Giảng viên hướng dẫn:** Phạm Thị Thùy Liên

*Vĩnh Phúc, năm 2024*

**MỤC LỤC**

[**LỜI CẢM ƠN** 1](#_Toc155887472)

[**LỜI NÓI ĐẦU** 2](#_Toc155887473)

[**CHƯƠNG 1 : BÀI TẬP THỰC HÀNH** 3](#_Toc155887474)

[Câu 1. So sánh Spark và Map Reduce? 3](#_Toc155887475)

[Câu 2. Trình bày về DStream? (Khái niệm, cách thức hoạt động) 5](#_Toc155887476)

[Câu 3. Cho tập dữ liệu về các phỏng vấn một số khách hàng của ngân hàng có nội dung như bên dưới (File đính kèm), sử dụng Spark RDD hoặc Spark SQL để trả lời một số câu hỏi sau: 6](#_Toc155887477)

[3. Code mở đầu 6](#_Toc155887478)

[3.1 Có bao nhiêu loại nghề nghiệp trong thống kê? 6](#_Toc155887479)

[3.2 Liệt kê danh sách và số lượng các loại công việc trong thống kê? 7](#_Toc155887480)

[3.3 Có bao nhiêu người đã kết hôn và độc thân trong thống kê? 8](#_Toc155887481)

[3.4 Có bao nhiêu loại tình trạng giáo dục trong thống kê và tương ứng số lượng người 8](#_Toc155887482)

[3.5 Tính độ tuổi trung bình của các khách hàng trong danh sách? 9](#_Toc155887483)

[3.6 Khách hàng có độ tuổi lớn nhất, nhỏ nhất là bao nhiêu? 9](#_Toc155887484)

[3.7 Liệt kê danh sách các khách hàng còn độc thân, đồng thời có khoản vay ngân hàng và được giáo dục ở mức “secondary” 9](#_Toc155887485)

[3.8 Liệt kê danh sách các khách hàng có tuổi lớn hơn độ tuổi trung bình, đồng thời không có khoản nợ ngân hàng và chưa vay mua nhà? 10](#_Toc155887486)

[3.9 Tạo một cột mới nếu: Age < 25 điền là Y, 25 <= Age <= 30 điền là T, 30 < Age điền là O. 11](#_Toc155887487)

[3.10 Tạo một cột mới nếu giá trị tại cột “Housing” là Yes và cột “Loan” là Yes điền là “Type1”, cột “Housing” là Yes và cột “Loan” là No điền là “Type2”, cột “Housing” là No và cột “Loan” là No điền là “Type3”, còn lại điền là “Type4” 12](#_Toc155887488)

[Câu 4. Sử dụng số liệu thống kê số lần click vào các đường dẫn phim trên NetFlix trong năm 2017, 2018, 2019 (file đính kèm) để thực hành streaming kết quả 3 bộ phim được quan tâm nhất theo năm. 18](#_Toc155887489)

[**KẾT LUẬN** 19](#_Toc155887490)

[**TÀI LIỆU THAM KHẢO** 20](#_Toc155887491)

**BẢNG PHÂN CÔNG NHIỆM VỤ TỪNG THÀNH VIÊN**

|  |  |
| --- | --- |
| **Thành Viên** | **Nhiệm Vụ** |
| Nguyễn Văn Chung | 1) Câu 1 lý thuyết  2) Câu 2 lý thuyết  3) Câu 3 thực hành |
| Nguyễn Văn Tú | 1) Câu 1 lý thuyết  2) Câu 2 lý thuyết  3) Câu 3 thực hành |
| Nguyễn Văn Đạt | 1) Câu 1 lý thuyết  2) Câu 2 lý thuyết  3) Câu 3 thực hành |
| Dương Minh Đạt | 1) Câu 1 lý thuyết  2) Câu 2 lý thuyết  3) Câu 3 thực hành |
| Nguyễn Trọng Tấn | 1) Câu 1 lý thuyết  2) Câu 2 lý thuyết  3) Câu 3 thực hành |

# **LỜI CẢM ƠN**

Trước hết chúng em xin bày tỏ lòng biết ơn sâu sắc nhất tới giảng viên hướng dẫn cô **Phạm Thị Thùy Liên** đã tận tình giúp đỡ, hướng dẫn chúng em rất nhiều trong suốt quá trình tìm hiểu nghiên cứu và hoàn thành bài tập lớn.

Chúng em cũng muốn gửi lời cảm ơn đặc biệt đến các thành viên trong nhóm, bởi sự cống hiến, hợp tác và nỗ lực trong suốt quá trình làm việc. Sự hợp tác tốt đẹp của chúng ta đã tạo nên một sản phẩm hoàn thiện và đáng tự hào.

Cuối cùng, chúng em rất biết ơn về cơ hội được tham gia vào dự án này. Quá trình học tập và áp dụng kiến thức thực tế trong bài tập lớn này đã mang lại cho chúng em rất nhiều kinh nghiệm quý báu.

Vì thời gian có hạn, trình độ hiểu biết của chúng em còn nhiều hạn chế. Cho nên trong bài tập lớn không tránh khỏi những thiếu sót, chúng em rất mong nhận được sự đóng góp ý kiến của cô giáo để bài báo cáo cuối kì này của chúng em được từng bước hoàn thiện hơn.

Chúng Em xin chân thành cảm ơn!

# **LỜI NÓI ĐẦU**

Big Data - dữ liệu lớn và phức tạp - đã trở thành một yếu tố cực kỳ quan trọng trong cuộc sống và công việc hiện đại. Với sự phát triển nhanh chóng của công nghệ thông tin và kết nối mạng, chúng ta đã chứng kiến một sự gia tăng vượt bậc về khối lượng và tốc độ thu thập dữ liệu. Các nguồn dữ liệu đa dạng như hệ thống thông tin, mạng xã hội, thiết bị di động, cảm biến và Internet of Things đã tạo ra một nguồn cung dồi dào và đa chiều của dữ liệu. Điều này đã mở ra những cơ hội mới và đồng thời đặt ra những thách thức khó khăn mà chúng ta cần đối mặt.Về cơ bản, Big Data mang lại những cơ hội vô cùng hứa hẹn. Dữ liệu lớn được coi là nguồn tài nguyên quý giá, có khả năng cung cấp thông tin phong phú và chi tiết về hành vi con người, xu hướng thị trường, dự báo về tương lai, và nhiều khía cạnh khác của thế giới xung quanh chúng ta. Bằng cách phân tích và sử dụng Big Data một cách thông minh, chúng ta có thể tìm thấy các mẫu, quy luật và thông tin giá trị từ dữ liệu để hỗ trợ ra quyết định, phát triển sản phẩm và dịch vụ mới, tăng cường hiệu suất và sáng tạo trong công việc, tuy nhiên việc làm việc với Big Data cũng đặt ra nhiều thách thức.

Bài tập lớn về đề tài “Số 4” này là một cơ hội để chúng em áp dụng những kiến thức đã học và nghiên cứu sâu hơn về Big Data. Với mong muốn vận dụng những kiến thức đã học giúp bản thân chúng em củng cố lại những kiến thức của mình trong quá trình cùng nhau làm và hoàn thiện đề tài. Mặc dù đã rất cố gắng để hoàn thành công việc, nhưng do thời gian có hạn và thiếu kinh nghiệm cũng như kỹ năng chưa cao nên còn nhiều thiếu sót, kính mong cô góp ý, bổ sung để chúng em hoàn thiện tốt hơn nữa.

# **CHƯƠNG 1 : BÀI TẬP THỰC HÀNH**

Câu 1. So sánh Spark và Map Reduce?

Spark và MapReduce là hai công nghệ phân tán phổ biến được sử dụng để xử lý dữ liệu lớn. Đây là một số điểm giống và khác nhau giữa Spark và MapReduce:

Giống nhau:

* Mã nguồn mở: Cả Spark và MapReduce đều được phát triển và phân phối dưới dạng mã nguồn mở. Điều này có nghĩa là bất kỳ ai cũng có thể truy cập vào mã nguồn của chúng, tạo ra các bản vá lỗi, cải tiến tính năng, và thậm chí tạo ra các phiên bản tùy chỉnh phù hợp với nhu cầu cụ thể.
* Khả năng chịu lỗi: Cả Spark và MapReduce đều có khả năng chịu lỗi và chịu sự cố. Cả hai công nghệ này đều hỗ trợ khả năng sao lưu dữ liệu và khả năng khôi phục lại quá trình xử lý trong trường hợp xảy ra sự cố.
* Xử lý dữ liệu phân tán: Cả Spark và MapReduce đều được thiết kế để xử lý dữ liệu phân tán trên một cụm máy tính. Cả hai công nghệ này có khả năng chia nhỏ dữ liệu thành các công việc nhỏ hơn và thực hiện xử lý song song trên các nút của cụm.
* Phân tích dữ liệu lớn: Cả Spark và MapReduce được sử dụng rộng rãi trong các tác vụ phân tích dữ liệu lớn. Cả hai công nghệ đều cung cấp khả năng xử lý và phân tích dữ liệu lớn thông qua các thư viện và API phong phú.

Khác nhau:

| **Điểm so sánh** | **Apache Spark** | **MapReduce** |
| --- | --- | --- |
| Hiệu suất | Nhanh hơn so với MapReduce | Thường chậm hơn Spark vì cơ chế đọc/ghi dữ liệu trên đĩa |
| Bộ nhớ | Sử dụng bộ nhớ RAM để lưu trữ dữ liệu tạm thời | Sử dụng đĩa để lưu trữ dữ liệu tạm thời |
| Xử lý đồng thời | Hỗ trợ xử lý đồng thời (in-memory computing) | Không hỗ trợ xử lý đồng thời (in-memory computing) |
| Hỗ trợ ngôn ngữ | Hỗ trợ nhiều ngôn ngữ bao gồm Scala, Java, Python, R... | Chủ yếu hỗ trợ ngôn ngữ Java |
| Công cụ phân tích dữ liệu | Cung cấp các công cụ phân tích dữ liệu tích hợp sẵn | Không cung cấp các công cụ phân tích dữ liệu tích hợp sẵn |
| Tính linh hoạt | Cung cấp các API linh hoạt cho xử lý dữ liệu | Cung cấp một số ít API cơ bản cho xử lý dữ liệu |
| Thíc h hợp | Phù hợp với các ứng dụng xử lý dữ liệu phức tạp | Phù hợp với các ứng dụng xử lý dữ liệu truyền thống |

Câu 2. Trình bày về DStream? (Khái niệm, cách thức hoạt động)

DStream là một khái niệm trong Apache Spark Streaming, đại diện cho một chuỗi các RDD (Resilient Distributed Datasets). DStream được tạo ra bằng cách chia dữ liệu dòng thành các micro-batches và xử lý chúng độc lập. DStream cung cấp các phương thức để thực hiện các phép biến đổi và phân tích dữ liệu dòng. Điều này giúp Spark Streaming xử lý dữ liệu gần thời gian thực và tận dụng tính mạnh mẽ của Spark.

Cách thức hoạt động của DStream như sau:

DStream hoạt động dựa trên hai điểm chính

* Phép biến đổi trên RDD: DStream thực hiện các phép biến đổi trên RDD (Resilient Distributed Datasets) thông qua Spark Core. Các phép biến đổi này bao gồm map, filter, reduce, join và các phép biến đổi khác để xử lý dữ liệu trong mỗi batch.
* Chuyển đổi hoạt động từ DStream sang RDD: Các hoạt động trên DStream được chuyển đổi thành các hoạt động cơ bản trên RDD. Điều này cho phép DStream tận dụng các tính năng và khả năng phân phối của Spark Core để xử lý dữ liệu streaming một cách song song trên một cluster.

Tóm lại, DStream sử dụng Spark Core để thực hiện các phép biến đổi trên RDD và chuyển đổi các hoạt động trên DStream thành các hoạt động cơ bản trên RDD để xử lý dữ liệu streaming.

## Câu 3. Cho tập dữ liệu về các phỏng vấn một số khách hàng của ngân hàng có nội dung như bên dưới (File đính kèm), sử dụng Spark RDD hoặc Spark SQL để trả lời một số câu hỏi sau:

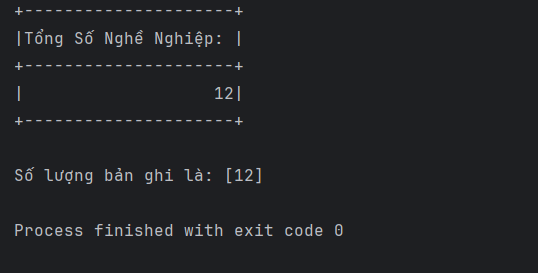
### 3. Code mở đầu

*//khởi tạo spark session*val spark = SparkSession.*builder*.master("local[\*]").appName("BigExerciseTeam1").getOrCreate()  
spark.sparkContext.setLogLevel("ERROR")  
  
*//Tạo dataframe bằng cách đọc từ file csv*val dataFrame : DataFrame = *//tạo 1 data frame từ file csv* spark.read *//đọc dữ liệu* .option("header", "true") *//giữ nguyên các dòng đầu tiên trong dataset* .option("inferSchema", "true") *//tự động ép kiểu dữ liệu cho các cột trong dataset* .csv("D:\\DeTai4\_BaiTapLon\\DeTai4\_BigEx\\Spark\_Banking\_Marketing.csv") *//path dataset  
  
//Tạo bảng tạm từ (temp view) từ Data Frame có tên là duLieuKhachHang, nó sẽ chứa dữ liệu file csv để ta có thể truy vấn*dataFrame.createOrReplaceTempView("duLieuKhachHang")

### 3.1 Có bao nhiêu loại nghề nghiệp trong thống kê?

*/\*Câu 1: Tính xem có bao nhiêu nghề nghiệp trong thống kê  
val c1 = spark.sql("SELECT COUNT (DISTINCT job) AS `Tổng Số Nghề Nghiệp: ` FROM duLieuKhachHang")  
c1.show()  
val count = c1.head()  
println("Số lượng bản ghi là: " + count)\*/*

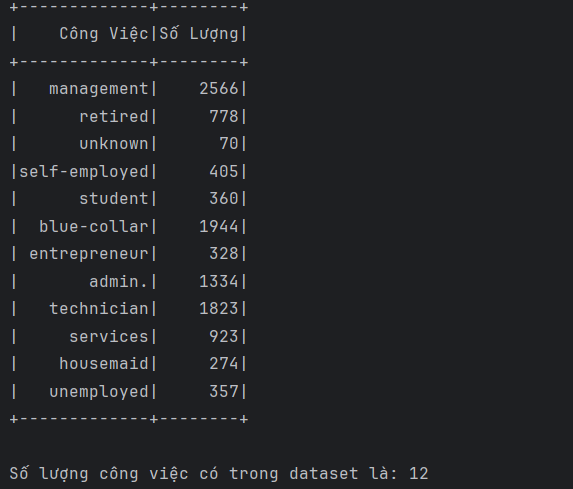
Kết quả: đếm được 12 loại nghề.



### 3.2 Liệt kê danh sách và số lượng các loại công việc trong thống kê?

*/\*Câu 2: Liệt kê danh sách và số lượng các loại công việc trong thống kê?  
val c2 = spark.sql("SELECT job AS `Công Việc`, COUNT(job) AS `Số Lượng` FROM duLieuKhachHang GROUP BY job")  
c2.show()  
val count = c2.count  
println("Số lượng công việc có trong dataset là: " + count)\*/*

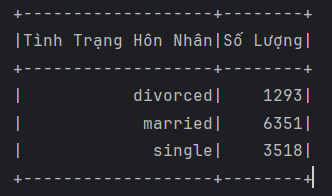
Kết quả:



### 3.3 Có bao nhiêu người đã kết hôn và độc thân trong thống kê?

*/\*Câu 3: Có bao nhiêu người đã kết hôn và độc thân trong thống kê?  
val c3 = spark.  
 sql (  
 "SELECT marital AS `Tình Trạng Hôn Nhân`, COUNT(\*) AS `Số Lượng` " +  
 "FROM duLieuKhachHang " +  
 "GROUP BY marital ).show()\*/*

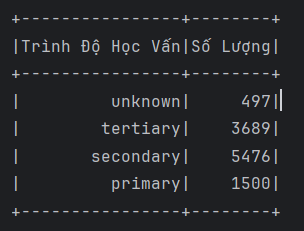
Kết quả: In ra tình trạng hôn nhân và số người thuộc các tình trạng hôn nhân trong đó.



### 3.4 Có bao nhiêu loại tình trạng giáo dục trong thống kê và tương ứng số lượng người

*/\*Câu 4: Có bao nhiêu loại tình trạng giáo dục trong thống kê và tương ứng số lượng người  
val c4 =  
 spark.sql(  
 "SELECT education AS `Trình Độ Học Vấn`, COUNT(\*) AS `Số Lượng` " +  
 "FROM duLieuKhachHang " +  
 "GROUP BY education").show()*

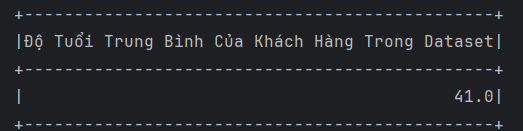
Kết quả:



### 3.5 Tính độ tuổi trung bình của các khách hàng trong danh sách?

*/\*Câu 5: Tính độ tuổi trung bình của các khách hàng trong danh sách?  
val c5 = spark.sql("SELECT ROUND(AVG(age)) AS `Độ Tuổi Trung Bình Của Khách Hàng Trong Dataset` FROM duLieuKhachHang").show()*

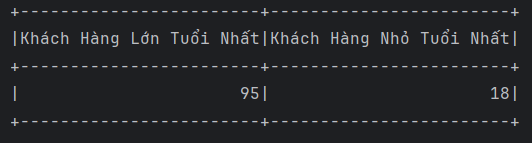
Kết quả:



### 3.6 Khách hàng có độ tuổi lớn nhất, nhỏ nhất là bao nhiêu?

*/\*Câu 6: Khách hàng có độ tuổi lớn nhất, nhỏ nhất là bao nhiêu?  
val c6 = spark.  
 sql("SELECT MAX(age) AS `Khách Hàng Lớn Tuổi Nhất`, MIN(age) AS `Khách Hàng Nhỏ Tuổi Nhất`" +  
 "FROM duLieuKhachHang").show()*

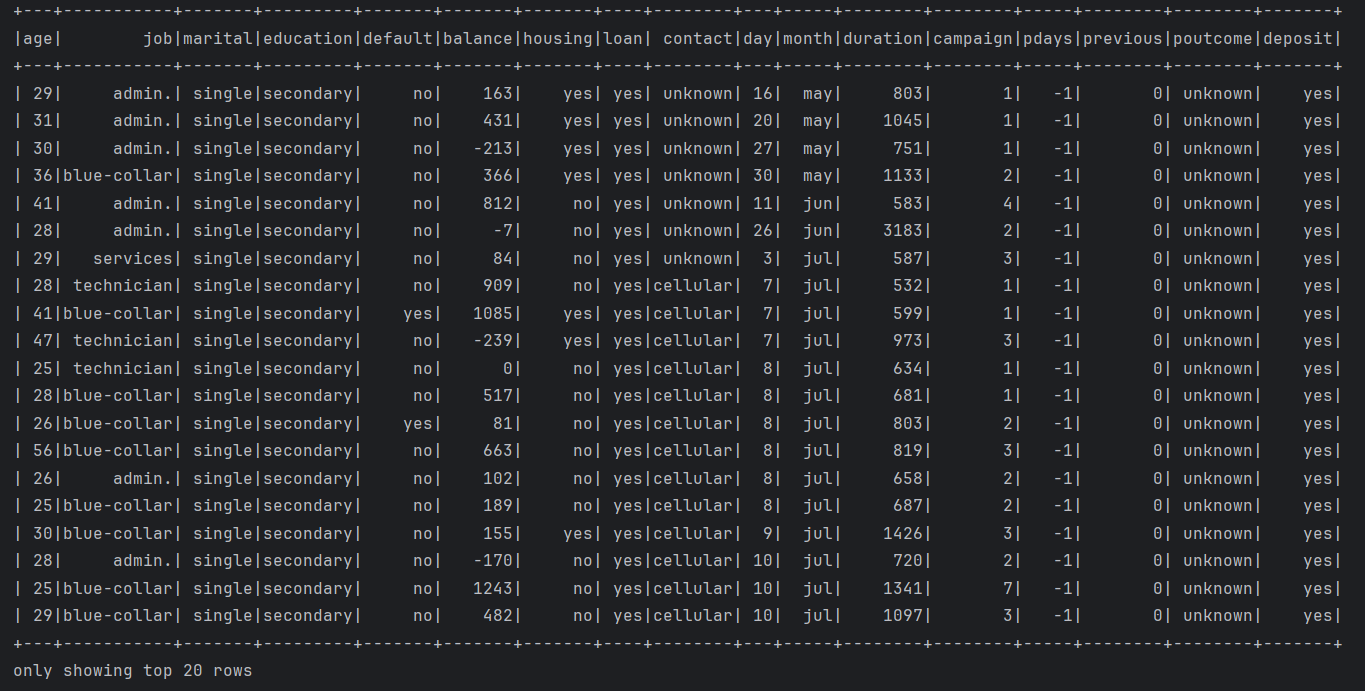
Kết quả:



### 3.7 Liệt kê danh sách các khách hàng còn độc thân, đồng thời có khoản vay ngân hàng và được giáo dục ở mức “secondary”

*/\*Câu 7: Liệt kê danh sách các khách hàng còn độc thân, đồng thời có khoản vay ngân hàng và được giáo dục ở mức “secondary”  
val c7 = spark.sql("SELECT \* FROM duLieuKhachHang WHERE marital = 'single' AND loan = 'yes' AND education = 'secondary'").show()*

Kết quả:

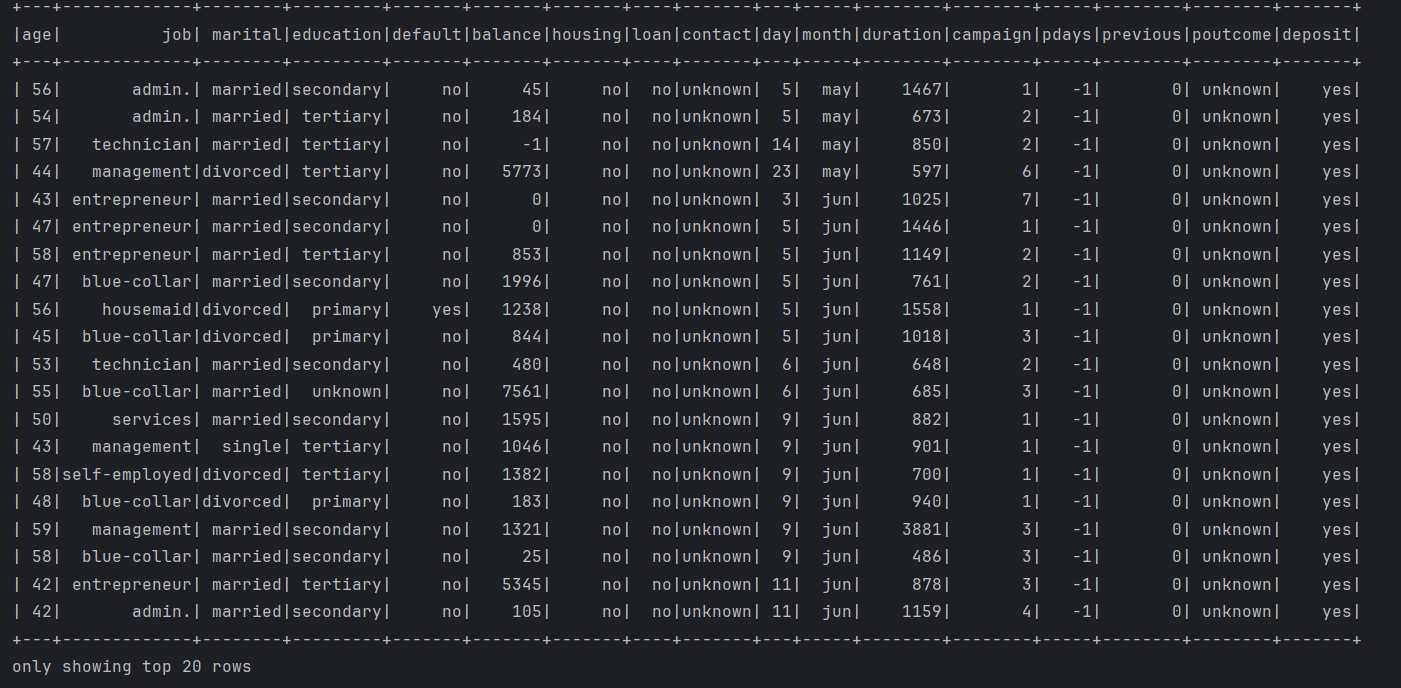


### 3.8 Liệt kê danh sách các khách hàng có tuổi lớn hơn độ tuổi trung bình, đồng thời không có khoản nợ ngân hàng và chưa vay mua nhà?

*/\*Câu 8: Liệt kê danh sách các khách hàng có tuổi lớn hơn độ tuổi trung bình, đồng thời không có khoản nợ ngân hàng và chưa vay mua nhà?  
val c8 = spark.sql( "SELECT \* FROM duLieuKhachHang " +  
 "WHERE age > (SELECT AVG(age) FROM duLieuKhachHang) " +  
 "AND loan = 'no' " +*

*"AND housing = 'no'").show()*

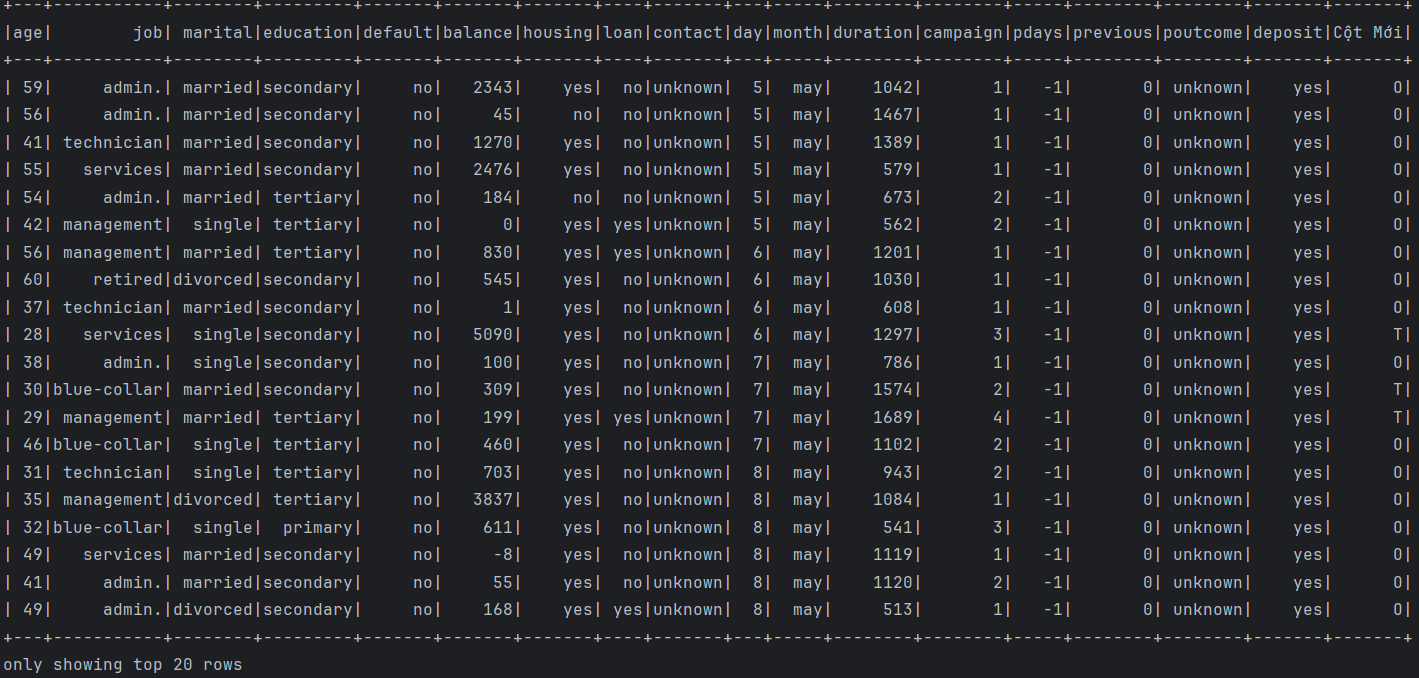
Kết quả: In ra những khách hàng có độ tuổi lớn hơn độ tuổi trung bình (41 tuổi) và không có khoản vay ngân hàng (loan = ‘no’) và chưa vay mua nhà (housing = ‘no’)



### 3.9 Tạo một cột mới nếu: Age < 25 điền là Y, 25 <= Age <= 30 điền là T, 30 < Age điền là O.

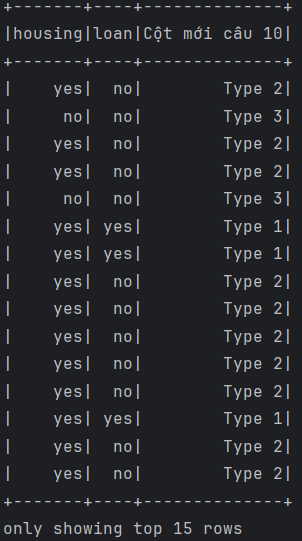
*val themCotMoi : DataFrame = spark.sql("""SELECT \*,  
 | CASE WHEN age < 25 THEN 'Y'  
 | WHEN age >= 25 AND age <= 30 THEN 'T'  
 | WHEN age > 30 THEN 'O'  
 | ELSE NULL  
 | END AS `Cột Mới`  
 |FROM duLieuKhachHang""".stripMargin)  
themCotMoi.show()*

Kết quả:



### 3.10 Tạo một cột mới nếu giá trị tại cột “Housing” là Yes và cột “Loan” là Yes điền là “Type1”, cột “Housing” là Yes và cột “Loan” là No điền là “Type2”, cột “Housing” là No và cột “Loan” là No điền là “Type3”, còn lại điền là “Type4”

*val addNewColumn : DataFrame = spark.sql(  
 """SELECT housing, loan,  
 | CASE WHEN housing = 'yes' AND loan = 'yes' THEN 'Type 1'  
 | WHEN housing = 'yes' AND loan = 'no' THEN 'Type 2'  
 | WHEN housing = 'no' AND loan = 'no' THEN 'Type 3'  
 | ELSE 'Type 4'  
 | END AS `Cột mới câu 10`  
 | FROM duLieuKhachHang""".stripMargin)  
addNewColumn.show(15)*

Kết quả: In ra 15 dòng và chỉ lấy cột “housing” và “loan” kèm cột mới  


SOURCE CODE CÂU 3:

*package nhom1\_DeTai4\_BigData  
  
import org.apache.spark.sql.\_  
import org.apache.spark.sql.functions.\_  
  
object happyEnding {  
 def main(args: Array[String]) {  
 //khởi tạo spark session  
 val spark = SparkSession.builder.master("local[\*]").appName("BigExerciseTeam1").getOrCreate()  
 spark.sparkContext.setLogLevel("ERROR")  
  
 //Tạo dataframe bằng cách đọc từ file csv  
 val dataFrame : DataFrame = //tạo 1 data frame từ file csv  
 spark.read //đọc dữ liệu  
 .option("header", "true") //giữ nguyên các dòng đầu tiên trong dataset  
 .option("inferSchema", "true") //tự động ép kiểu dữ liệu cho các cột trong dataset  
 .csv("D:\\DeTai4\_BaiTapLon\\DeTai4\_BigEx\\Spark\_Banking\_Marketing.csv") //path dataset  
  
 //Tạo bảng tạm từ (temp view) từ Data Frame có tên là duLieuKhachHang, nó sẽ chứa dữ liệu file csv để ta có thể truy vấn  
 dataFrame.createOrReplaceTempView("duLieuKhachHang")  
  
 /\*Câu 1: Tính xem có bao nhiêu nghề nghiệp trong thống kê  
 val c1 = spark.sql("SELECT COUNT (DISTINCT job) AS `Tổng Số Nghề Nghiệp: ` FROM duLieuKhachHang")  
 c1.show()  
 val count = c1.head()  
 println("Số lượng bản ghi là: " + count)\*/  
  
 /\*Câu 2: Liệt kê danh sách và số lượng các loại công việc trong thống kê?  
 val c2 = spark.sql("SELECT job AS `Công Việc`, COUNT(job) AS `Số Lượng` FROM duLieuKhachHang GROUP BY job")  
 c2.show()  
 val count = c2.count  
 println("Số lượng công việc có trong dataset là: " + count)\*/  
  
 /\*Câu 3: Có bao nhiêu người đã kết hôn và độc thân trong thống kê?  
 val c3 = spark.  
 sql (  
 "SELECT marital AS `Tình Trạng Hôn Nhân`, COUNT(\*) AS `Số Lượng` " +  
 "FROM duLieuKhachHang " +  
 "GROUP BY marital"  
 ).show()\*/  
  
 /\*Câu 4: Có bao nhiêu loại tình trạng giáo dục trong thống kê và tương ứng số lượng người  
 val c4 =  
 spark.sql(  
 "SELECT education AS `Trình Độ Học Vấn`, COUNT(\*) AS `Số Lượng` " +  
 "FROM duLieuKhachHang " +  
 "GROUP BY education").show()  
  
 \*/  
  
 /\*Câu 5: Tính độ tuổi trung bình của các khách hàng trong danh sách?  
 val c5 = spark.sql("SELECT ROUND(AVG(age)) AS `Độ Tuổi Trung Bình Của Khách Hàng Trong Dataset` FROM duLieuKhachHang").show()  
 \*/  
  
 /\*Câu 6: Khách hàng có độ tuổi lớn nhất, nhỏ nhất là bao nhiêu?  
 val c6 = spark.  
 sql("SELECT MAX(age) AS `Khách Hàng Lớn Tuổi Nhất`, MIN(age) AS `Khách Hàng Nhỏ Tuổi Nhất`" +  
 "FROM duLieuKhachHang").show()  
 \*/  
  
 /\*Câu 7: Liệt kê danh sách các khách hàng còn độc thân, đồng thời có khoản vay ngân hàng và được giáo dục ở mức “secondary”  
 val c7 = spark.sql("SELECT \* FROM duLieuKhachHang WHERE marital = 'single' AND loan = 'yes' AND education = 'secondary'").show()  
 \*/  
  
 /\*Câu 8: Liệt kê danh sách các khách hàng có tuổi lớn hơn độ tuổi trung bình, đồng thời không có khoản nợ ngân hàng và chưa vay mua nhà?  
 val c8 = spark.sql( "SELECT \* FROM duLieuKhachHang " +  
 "WHERE age > (SELECT AVG(age) FROM duLieuKhachHang) " +  
 "AND loan = 'no' " +  
 "AND housing = 'no'").show()  
 \*/  
  
 /\*Câu 9: Tạo một cột mới nếu: Age < 25 điền là Y, 25 <= Age <= 30 điền là T, 30 < Age điền là O.  
 val themCotMoi : DataFrame = spark.sql("""SELECT \*,  
 | CASE WHEN age < 25 THEN 'Y'  
 | WHEN age >= 25 AND age <= 30 THEN 'T'  
 | WHEN age > 30 THEN 'O'  
 | ELSE NULL  
 | END AS `Cột Mới`  
 |FROM duLieuKhachHang""".stripMargin)  
 themCotMoi.show()  
 Nếu muốn chỉ hiển thị cột tuổi (age) và cột mới thì ta chỉ cần SELECT age thôi(khỏi \*) :))))  
 \*/  
  
 /\*Câu 10: Tạo một cột mới nếu giá trị tại cột “Housing” là Yes và cột “Loan” là Yes điền là  
 \*“Type1”, cột “Housing” là Yes và cột “Loan” là No điền là “Type2”, cột  
 “Housing” là No và cột “Loan” là No điền là “Type3”, còn lại điền là “Type4”  
 val addNewColumn : DataFrame = spark.sql(  
 """SELECT housing, loan,  
 | CASE WHEN housing = 'yes' AND loan = 'yes' THEN 'Type 1'  
 | WHEN housing = 'yes' AND loan = 'no' THEN 'Type 2'  
 | WHEN housing = 'no' AND loan = 'no' THEN 'Type 3'  
 | ELSE 'Type 4'  
 | END AS `Cột mới câu 10`  
 | FROM duLieuKhachHang""".stripMargin)  
 addNewColumn.show(15)  
 \*/  
  
 spark.stop()//dừng spark context(hay spark session). Tóm lại là kết thúc phiên làm việc  
 }  
}*

Câu 4. Sử dụng số liệu thống kê số lần click vào các đường dẫn phim trên NetFlix trong năm 2017, 2018, 2019 (file đính kèm) để thực hành streaming kết quả 3 bộ phim được quan tâm nhất theo năm.

# **KẾT LUẬN**

Trong thời đại 4.0 hiện nay, Big Data đóng vai trò quan trọng và có ảnh hưởng sâu sắc đến nhiều lĩnh vực trong xã hội và kinh tế, là thuật ngữ dùng để mô tả lượng dữ liệu lớn và phức tạp mà các tổ chức và công ty thu thập và xử lý từ nhiều nguồn khác nhau. Dữ liệu này được tạo ra từ các nguồn như các trang web, thiết bị di động, cảm biến, mạng xã hội, và các hệ thống máy móc khác. Big Data có tiềm năng mang lại nhiều lợi ích đáng kể. Đầu tiên, Big Data cung cấp thông tin chi tiết và đa dạng về hành vi và xu hướng của người dùng. Điều này giúp các tổ chức và doanh nghiệp hiểu rõ hơn về khách hàng và tạo ra các sản phẩm và dịch vụ tốt hơn, từ đó tăng cường trải nghiệm người dùng và tăng cường hiệu quả kinh doanh. Thứ hai, Big Data cung cấp cơ sở cho việc phát triển và áp dụng trí tuệ nhân tạo và học máy, giúp tạo ra các ứng dụng thông minh và tự động hóa các quy trình và quyết định. Thứ ba, Big Data cung cấp thông tin quan trọng cho các lĩnh vực như y tế, nông nghiệp, quản lý đô thị và giao thông, giúp cải thiện chất lượng cuộc sống và tạo ra các giải pháp thông minh cho các vấn đề phức tạp.

Tuy nhiên, cùng với những lợi ích này, Big Data cũng đối mặt với một số bất cập. Một trong số đó là vấn đề về bảo mật và riêng tư dữ liệu. Với khối lượng dữ liệu lớn và phức tạp, việc bảo vệ thông tin cá nhân và đảm bảo an toàn dữ liệu trở nên khó khăn hơn. Ngoài ra, việc thu thập và xử lý Big Data cũng đặt ra thách thức về khả năng lưu trữ và xử lý dữ liệu. Cần có cơ sở hạ tầng phù hợp và các công nghệ xử lý dữ liệu hiệu quả để khai thác toàn bộ tiềm năng của Big Data.

Ngoài ra, Big Data cũng đặt ra vấn đề về đạo đức và trách nhiệm. Việc sử dụng dữ liệu một cách có ý thức và đảm bảo tính công bằng, đồng thời tránh các hành vi xâm phạm đến quyền riêng tư và phân biệt đối xử là một thách thức quan trọng. Cần có các quy định và quy tắc rõ ràng để đảm bảo việc sử dụng và chia sẻ dữ liệu một cách đáng tin cậy và đúng đắn.

# **TÀI LIỆU THAM KHẢO**

1. Trang web “Paceedu.vn”, BIG DATA LÀ GÌ? ĐẶC ĐIỂM & ỨNG DỤNG BIG DATA VÀO CÁC NGÀNH.

Link trang web: <https://www.pace.edu.vn/tin-kho-tri-thuc/big-data-la-gi>

2. Trang web “brandsvietnam.com”, Big Data là gì? Tất tần tật những điều bạn nên biết về Big Data.

Link trang web: <https://www.brandsvietnam.com/congdong/topic/320226-Big-Data-la-gi-Tat-tan-tat-nhung-dieu-ban-nen-biet-ve-Big-Data>

3. Trang web “vdigital.vn”, Đặc trưng của Big Data – Sự khác biệt giữa “rừng” công nghệ mới thời đại 4.0

Link trang web: <https://vdigital.vn/dac-trung-cua-big-data/>

4. Trang web “maytinhcincoze.com”, Big data – Cốt lõi của nền công nghiệp 4.0, đăng ngày 05/04/2021

Link trang web: <https://maytinhcincoze.com/big-data-cot-loi-cua-nen-cong-nghiep-4-0/>

5. Trang web “grad.hcmut.edu.vn”, Đề cương môn học Big Data Trường Đại học Bách Khoa ĐHQG Tp.HCM

Link trang web: <https://grad.hcmut.edu.vn/gv/front.php/capnhatdecuong/public/printdecuongcdio?option=tv&mamh=055153&manganh=8440214>