|  |
| --- |
| **TRƯỜNG ĐẠI HỌC SƯ PHẠM HÀ NỘI**  **KHOA CÔNG NGHỆ THÔNG TIN**  **---\*\*\*---** A blue and white logo  AI-generated content may be incorrect. |
| **DỮ LIỆU LỚN (BIG DATA)**  **ĐỀ TÀI:**  ***ỨNG DỤNG BIG DATA TRONG PHÂN TÍCH HÀNH VI KHÁCH HÀNG VÀ ĐỀ XUẤT CHIẾN LƯỢC THƯƠNG MẠI ĐIỆN TỬ.***  GVHD: TS. Giang Thành Trung  Nhóm: 4 Thành viên  Sinh viên thực hiện : Lê Thị Mỹ Duyên -725105055  Vũ Thùy Dương - 725105049  Trần Thị Huệ - 7251050  Phan Thị Uyên - 715105262 |
| ***Hà Nội, 2025*** |

MỤC LỤC

[**LỜI NÓI ĐẦU** 4](#_Toc197425402)

[**CHƯƠNG I: QUY TRÌNH THỰC HIỆN VÀ KẾT QUẢ** 5](#_Toc197425403)

[**1. Các công cụ được sử dụng** 5](#_Toc197425404)

[1.1. Hadoop 5](#_Toc197425405)

[*1.1.2. Ứng dụng của hadoop trong đề tài* 5](#_Toc197425406)

[1.2. Apache Spark 6](#_Toc197425407)

[1.3. KNIME 7](#_Toc197425408)

[1.4. Power BI 7](#_Toc197425409)

[2. Mô tả dữ liệu 8](#_Toc197425410)

[2.1. Thông tin về dữ liệu 8](#_Toc197425411)

[2.2 Chi tiết về dữ liệu 9](#_Toc197425412)

[3. Tiền xử lý dữ liệu. 9](#_Toc197425413)

[3.1 Đọc dữ liệu 10](#_Toc197425414)

[3.2 Kiểm tra và chuẩn hóa dữ liệu 11](#_Toc197425415)

[3.3 Xử lý giá trị thiếu(Missing values) 12](#_Toc197425416)

[3.4 Ghi dữ liệu rồi làm sạch ra file 14](#_Toc197425417)

[4. Lưu trữ dữ liệu lên HDFS 14](#_Toc197425418)

[5. Truy vấn và phân tích dữ liệu với PySpark 15](#_Toc197425419)

[5.1. Khởi tạo Spark Session và đọc dữ liệu từ HDFS 15](#_Toc197425420)

[5.2. Truy vấn dữ liệu 16](#_Toc197425421)

[6. Trực quan hóa kết quả bằng công cụ Power BI 25](#_Toc197425422)

[**CHƯƠNG II. KÊT LUẬN** 34](#_Toc197425423)

# **LỜI NÓI ĐẦU**

  Trong bối cảnh công nghệ số phát triển mạnh mẽ, thương mại điện tử đang trở thành lĩnh vực có tốc độ tăng trưởng dữ liệu vượt bậc. Các hành vi người dùng như lượt truy cập, thêm sản phẩm vào giỏ hàng, hoặc hoàn tất giao dịch không chỉ là các thao tác đơn lẻ, mà còn là những dấu vết số có giá trị chiến lược, giúp doanh nghiệp hiểu rõ hơn về khách hàng, tối ưu hóa hoạt động kinh doanh và gia tăng lợi nhuận.

Với sự phát triển của lưu trữ dữ liệu và xử lý dữ liệu lớn, các nền tảng như **Hadoop** và **Apache Spark** đóng vai trò then chốt trong việc thu thập, lưu trữ, và phân tích khối lượng dữ liệu khổng lồ đến từ thương mại điện tử.

Đề tài **"** **ỨNG DỤNG BIG DATA TRONG PHÂN TÍCH HÀNH VI KHÁCH HÀNG VÀ ĐỀ XUẤT CHIẾN LƯỢC THƯƠNG MẠI ĐIỆN TỬ."** nhằm xây dựng một hệ thống xử lý dữ liệu sử dụng Hadoop và Spark, thực hiện truy vấn và phân tích để khám phá sâu các xu hướng tiêu dùng, tỷ lệ chuyển đổi, ngành hàng phổ biến,... từ đó đưa ra các gợi ý chiến lược hỗ trợ ra quyết định kinh doanh.

Lý do nhóm lựa chọn đề tài này xuất phát từ nhu cầu thực tiễn và tính ứng dụng cao của phân tích dữ liệu lớn trong lĩnh vực thương mại điện tử. Trong thời đại số hóa, việc hiểu rõ hành vi người tiêu dùng thông qua dữ liệu không còn là lợi thế mà đã trở thành yêu cầu thiết yếu để doanh nghiệp có thể cạnh tranh và phát triển. Với bộ dữ liệu hành vi người dùng có quy mô hơn 50 triệu dòng, đề tài mang lại cơ hội tiếp cận và khai thác nguồn thông tin phong phú để khám phá các xu hướng mua sắm, tỷ lệ chuyển đổi và mức độ phổ biến của từng nhóm sản phẩm.

Bên cạnh đó, việc ứng dụng các công nghệ hiện đại như Hadoop và Apache Spark giúp nhóm tiếp cận với quy trình xử lý và phân tích dữ liệu lớn, từ khâu lưu trữ đến phân tích và trực quan hóa kết quả. Đây là những công cụ phổ biến trong ngành công nghiệp dữ liệu hiện nay, góp phần nâng cao năng lực chuyên môn và kỹ năng công nghệ cần thiết cho định hướng nghề nghiệp trong lĩnh vực khoa học dữ liệu và phân tích kinh doanh. Vì vậy, đề tài không chỉ mang tính học thuật mà còn là bước chuẩn bị quan trọng cho nhu cầu công việc trong tương lai.

# **CHƯƠNG I: QUY TRÌNH THỰC HIỆN VÀ KẾT QUẢ**

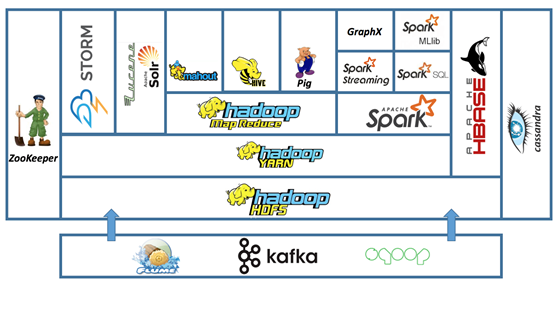
## **1. Các công cụ được sử dụng**

### 1.1. Hadoop

#### 1.1.1. Hadoop là gì

Hadoop là một công nghệ phân tán và mã nguồn mở được sử dụng phổ biến để xử lý và lưu trữ khối dữ liệu lớn trên các cụm máy tính phân tán, được thiết kế để xử lý và lưu trữ dữ liệu lớn một cách hiệu quả.

Cung cấp khả năng xử lý dữ liệu lớn, mở rộng linh hoạt và chi phí thấp, làm cho nó trở thành một công nghệ không thể thiếu trong các hệ thống xử lý dữ liệu hiện đại. Nó được phát triển để giải quyết các thách thức trong lĩnh vực Big Data mà các công nghệ cũ không thể đáp ứng.



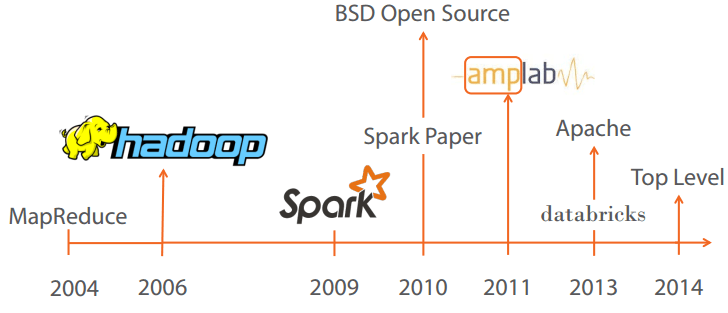
#### 1.1.2. Ứng dụng của hadoop trong đề tài

Hadoop được sử dụng để lưu trữ dữ liệu hành vi người dùng với quy mô lớn thông qua hệ thống HDFS. Nhờ khả năng phân tán và chịu lỗi cao, HDFS giúp lưu trữ dữ liệu an toàn, hiệu quả và dễ mở rộng. Dữ liệu sau khi được đưa vào HDFS sẽ phục vụ cho các truy vấn phân tích trên Apache Spark, đảm bảo quá trình xử lý nhanh chóng và ổn định.

### 1.2. Apache Spark

#### 1.2.1. Spark là gì

Apache Spark là một nền tảng điện toán cụm mạnh mẽ được thiết kế để xử lý dữ liệu lớn với tốc độ nhanh và phục vụ mục đích chung.



Về mặt tốc độ, Spark mở rộng mô hình MapReduce phổ biến để hỗ trợ hiệu quả nhiều loại tính toán hơn, bao gồm truy vấn tương tác và xử lý luồng. Tốc độ rất quan trọng trong việc xử lý các tập dữ liệu lớn, vì nó có nghĩa là sự khác biệt giữa việc khám phá dữ liệu một cách tương tác và chờ đợi nhiều phút hoặc nhiều giờ. Một trong những tính năng chính mà Spark cung cấp về tốc độ là khả năng chạy các phép tính trong bộ nhớ, nhưng hệ thống cũng hiệu quả hơn MapReduce đối với các ứng dụng phức tạp chạy trên đĩa.

Apache Spark được thiết kế để hỗ trợ nhiều khối lượng công việc khác nhau, bao gồm xử lý dữ liệu hàng loạt, thuật toán lặp, truy vấn tương tác và phát trực tuyến. Bằng cách tích hợp các loại xử lý này vào một công cụ duy nhất, Spark giúp giảm chi phí và đơn giản hóa quy trình phân tích dữ liệu.

Spark cung cấp các API dễ sử dụng cho Python, Java, Scala và SQL, đồng thời tích hợp chặt chẽ với hệ sinh thái Dữ liệu lớn như Hadoop và Cassandra.

#### 1.2.2. Ứng dụng của Spark trong đề tài

Trong đề tài sẽ sử dụng 1 thành phần của spark là PySpark: API của Spark cho Python, cho phép nhóm sử dụng Python để làm việc với Spark. Và PySpark sẽ được sử dụng để tiền xử lý dữ liệu và đẩy dữ liệu lên một database phù hợp.

### 1.3. KNIME

#### 1.3.1. KNIME là gì

KNIME (Konstanz Information Miner) là một nền tảng phân tích dữ liệu mã nguồn mở, được thiết kế để trực quan hóa quy trình xử lý dữ liệu mà không yêu cầu nhiều kiến thức lập trình. KNIME hỗ trợ tích hợp nhiều nguồn dữ liệu khác nhau, thực hiện các thao tác tiền xử lý, khai phá dữ liệu, phân tích thống kê, học máy, và trực quan hóa kết quả thông qua giao diện kéo – thả (drag and drop) thân thiện với người dùng. Với khả năng mở rộng cao và thư viện node phong phú, KNIME ngày càng được sử dụng rộng rãi trong các dự án phân tích dữ liệu thực tế.

#### 1.3.2. KNIME ứng dụng cho đề tài.

KNIME được sử dụng ở giai đoạn đầu để **tiền xử lý dữ liệu** trước khi đưa vào hệ thống lưu trữ và phân tích lớn như Hadoop và Spark. Cụ thể, KNIME hỗ trợ đọc tệp CSV lớn, lọc các trường không cần thiết, xử lý định dạng ngày giờ, phát hiện và loại bỏ các bản ghi rỗng hoặc không hợp lệ.

Quy trình sử dụng KNIME trong dự án gồm các bước:

1. **Đọc dữ liệu thô:** KNIME đọc file dữ liệu hành vi người dùng từ nền tảng thương mại điện tử (hơn 50 triệu dòng).
2. **Làm sạch và chuẩn hóa:** Sử dụng các node để lọc dữ liệu thiếu, chuyển đổi định dạng ngày tháng và chuẩn hóa cột.
3. **Kiểm tra thống kê sơ bộ:** Tính nhanh số lượng bản ghi, tỷ lệ các loại hành vi (xem, thêm giỏ, mua hàng,...) để đánh giá sơ bộ chất lượng dữ liệu.
4. **Xuất dữ liệu đã xử lý:** Dữ liệu sạch sẽ được lưu dưới dạng file hoặc đẩy tiếp vào pipeline để nhập vào hệ thống Hadoop.

Việc sử dụng KNIME giúp nhóm tiết kiệm thời gian xử lý thủ công, giảm lỗi, đồng thời đảm bảo chất lượng dữ liệu đầu vào cho các bước phân tích tiếp theo.

### 1.4. Power BI

#### 1.4.1. Power BI là gì.

Power BI là một công cụ phân tích và trực quan hóa dữ liệu do Microsoft phát triển, cho phép người dùng kết nối đến nhiều nguồn dữ liệu khác nhau, xử lý dữ liệu và tạo ra các báo cáo, dashboard tương tác một cách dễ dàng. Với giao diện kéo – thả thân thiện, cùng khả năng trực quan hóa mạnh mẽ, Power BI giúp người dùng không chuyên về lập trình vẫn có thể trình bày dữ liệu một cách trực quan và chuyên nghiệp.

#### 1.4.2. Ứng dụng Power BI trong đề tài.

Power BI được sử dụng ở **giai đoạn cuối của quy trình phân tích**, nhằm trực quan hóa các kết quả phân tích hành vi người dùng đã được xử lý và tổng hợp từ Spark hoặc chuyển đổi qua Python. Đây là bước quan trọng để trình bày insight một cách dễ hiểu và sinh động.

Cụ thể, quy trình sử dụng Power BI trong đề tài gồm:

1. **Kết nối dữ liệu:** Power BI được kết nối với các file dữ liệu đầu ra từ Spark hoặc Python (thường là CSV).
2. **Xây dựng báo cáo và dashboard:** Tạo các biểu đồ như biểu đồ cột, biểu đồ tròn, biểu đồ đường thể hiện
3. **Tùy chỉnh và tương tác:** Sử dụng slicers, filters và drill-down giúp người xem tương tác với báo cáo linh hoạt hơn.
4. **Chia sẻ kết quả:** Dashboard có thể được xuất bản lên Power BI Service để chia sẻ nội bộ hoặc trình bày trước hội đồng/kết luận dự án.

Việc sử dụng Power BI giúp nhóm không chỉ trình bày kết quả rõ ràng mà còn dễ dàng tạo các báo cáo chuyên nghiệp phục vụ cho việc ra quyết định chiến lược.

## 2. Mô tả dữ liệu

### 2.1. Thông tin về dữ liệu

- Bộ dữ liệu được lấy từ nền tảng **Kaggle**, thuộc dự án **eCommerce Behavior Data from Multi Category Store**, ghi nhận hành vi người dùng trong môi trường thương mại điện tử.

- Đường dẫn bộ dữ liệu: <https://www.kaggle.com/datasets/mkechinov/ecommerce-behavior-data-from-multi-category-store>

- Bộ dữ liệu được lưu trữ dưới dạng **tệp CSV** có tên 2019-Nov.csv, ghi nhận hơn **50 triệu dòng dữ liệu hành vi người dùng** từ một cửa hàng thương mại điện tử trong tháng 11 năm 2019.

- Tổng dung lượng **toàn bộ dataset gốc** là khoảng **14GB**. Tuy nhiên, trong dự án này, **nhóm sẽ sử dụng tệp 2019-Nov.csv** có dung lượng khoảng **9.01GB** để phân tích và xử lý, phù hợp với các bài toán dữ liệu lớn (Big Data) sử dụng **Hadoop và Spark**.

### 2.2 Chi tiết về dữ liệu

Bộ dữ liệu có **5 thuộc tính chính**, mỗi dòng là một hành vi của người dùng đối với một sản phẩm cụ thể:

|  |  |  |  |
| --- | --- | --- | --- |
| STT | Thuộc tính | Định nghĩa | Định dạng |
| 1 | event\_time | Thời gian diễn ra hành vi của người dùng | yyyy-mm-dd hh:mm:ss |
| 2 | event\_type | Loại hành vi của người dùng: view, cart, remove\_from\_cart, purchase | String |
| 3 | product\_id | ID sản phẩm người dùng tương tác | Integer |
| 4 | category\_id | ID của danh mục chứa sản phẩm | Float |
| 5 | category\_code | Tên danh mục sản phẩm dạng phân cấp (ví dụ: electronics.smartphone) | String |
| 6 | brand | Tên thương hiệu của sản phẩm | String |
| 7 | price | Giá của sản phẩm tại thời điểm tương tác | Float (USD) |
| 8 | user\_id | ID người dùng thực hiện hành vi | Integer |
| 9 | user\_session | Mã phiên truy cập của người dùng | String |

Với dữ liệu gốc có dung lượng 14GB và tệp 2019-Nov.csv sử dụng khoảng 9.01GB, việc sử dụng các công cụ như Hadoop và Spark sẽ giúp bạn xử lý hiệu quả dữ liệu lớn này để đưa ra các phân tích sâu sắc về hành vi người dùng.

## 3. Tiền xử lý dữ liệu.

Trước khi tiến hành phân tích, dữ liệu thu thập từ tệp CSV cần được tiền xử lý nhằm đảm bảo chất lượng và định dạng phù hợp với các thao tác xử lý tiếp theo trong hệ thống. Dữ liệu ban đầu bao gồm các thông tin liên quan đến sự kiện người dùng (user behavior), như thời gian, loại sự kiện, sản phẩm, danh mục, thương hiệu, giá cả, v.v.

Công cụ sử dụng: KNIME

Sử dụng công cụ KNIME để thực hiện xử lý:

* Loại bỏ các bản ghi có giá trị thiếu (price, event\_type, category\_code…)

• Lọc ra các sự kiện có ý nghĩa phân tích: view, cart, purchase

* Chuẩn hóa định dạng thời gian (event\_time) về định dạng chuẩn ISO 8601

• Chuyển đổi cột category\_code thành dạng có thể nhóm được (ví dụ lấy phần chính: electronics, appliances,…)

* Sau khi làm sạch, dữ liệu sẽ được lưu lại thành tệp .csv

Các bước tiền xử lý được thực hiện như sau:

### 3.1 Đọc dữ liệu

Node dùng: CSV Reader

Mục tiêu: Tải dữ liệu vào KNIME để xử lý

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

### 3.2 Kiểm tra và chuẩn hóa dữ liệu

Node dùng:String to Date

Mục tiêu: Chuyển đổi các cột về đúng kiểu (Số, ngày tháng, chuỗi)

A screenshot of a computer

AI-generated content may be incorrect.

Ví dụ: Đổi cột event\_type sang kiểu dữ liệu  là: yyyy-mm-dd hh:mm:cc utc

A screenshot of a computer

AI-generated content may be incorrect.

### 3.3 Xử lý giá trị thiếu(Missing values)

Node dùng: Missing Value

A yellow pencil and a yellow pencil

AI-generated content may be incorrect.

Cách xử lý:

* Với dữ liệu dạng text: điền "Unknown" hoặc "None"
* Với dữ liệu dạng số: điền 0, mean, median, hoặc loại bỏ dòng

A screenshot of a computer

AI-generated content may be incorrect.

Kết quả sau khi xử lý:

A screenshot of a computer

AI-generated content may be incorrect.

### 3.4 Ghi dữ liệu rồi làm sạch ra file

Node dùng: CSV Writer

File mới: dataoutput.csv

A screenshot of a computer

AI-generated content may be incorrect.

## 4. Lưu trữ dữ liệu lên HDFS

- Tệp dataoutput.csv là tệp làm sạch được chuyển lên hệ hệ thống HDFS bằng dòng lệnh

**hdfs dfs -put /usr/local/dataoutput.csv /input/**

A computer screen with white text

AI-generated content may be incorrect.

-Dữ liệu đã được đẩy lên

A screenshot of a computer

AI-generated content may be incorrect.

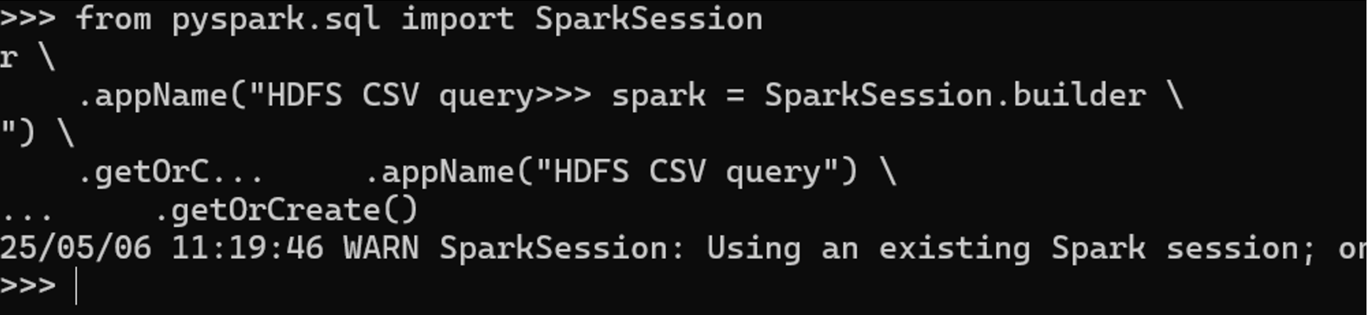
## 5. Truy vấn và phân tích dữ liệu với PySpark

### 5.1. Khởi tạo Spark Session và đọc dữ liệu từ HDFS

**spark = SparkSession.builder \**

**.appName("HDFS CSV query") \**

**.getOrCreate()**



### 5.2. Truy vấn dữ liệu

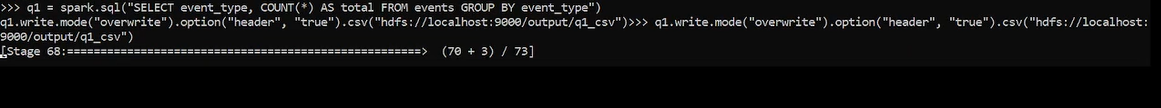
##### 5.2.1 Tổng số sự kiện theo loại

q1 = spark.sql("SELECT event\_type, COUNT(\*) AS total FROM events GROUP BY event\_type")

* Lưu file vào hdfs

q1.write.mode("overwrite").option("header", "true").csv("hdfs://localhost:9000/output/q1\_csv")

.



* Phân tích: Dữ liệu cho thấy phần lớn người dùng chỉ dừng ở mức xem sản phẩm, ít thêm vào giỏ và mua hàng. Cần tối ưu giao diện, khuyến mãi và quy trình thanh toán để tăng tỷ lệ chuyển đổi từ xem → mua.
* Chiến lược phát triển:
* Thúc đẩy hành động thêm vào giỏ  
  Dùng pop-up khuyến mãi khi người dùng lướt lâu hoặc cuộn đến cuối trang sản phẩm để kích thích hành động.
* Giảm bỏ giỏ bằng cam kết rõ ràng  
  Hiển thị “miễn phí vận chuyển” hoặc “đổi trả dễ dàng” ngay cạnh nút mua, giảm tâm lý do dự.
* Tăng mua hàng bằng cảm giác gấp rút  
  Thêm đồng hồ đếm ngược cho ưu đãi giới hạn hoặc thông báo “chỉ còn X sản phẩm” để tạo động lực mua nhanh

##### 5.2.2. Giá trung bình theo brand

q2 = spark.sql("SELECT brand, AVG(price) AS avg\_price FROM events GROUP BY brand")

q2.write.mode("overwrite").option("header", "true").csv("hdfs://localhost:9000/output/q2\_csv")

- Phân tích: Dữ liệu cho thấy mỗi thương hiệu có mức giá trung bình khác nhau, phản ánh định vị sản phẩm trên thị trường. Một số brand có giá trung bình cao nhưng có thể không tương xứng với mức độ quan tâm hoặc mua hàng. Doanh nghiệp cần rà soát lại chiến lược giá, đồng thời phân nhóm thương hiệu theo giá để xây dựng chiến lược marketing phù hợp. Với nhóm brand giá cao, có thể cần tăng trải nghiệm, nội dung mô tả giá trị sản phẩm; còn với nhóm giá thấp, cần đẩy mạnh khuyến mãi và quảng bá số lượng lớn để tăng doanh số.

- Chiến lược:

* Phân khúc thị trường: Sử dụng giá trung bình để phân loại các thương hiệu thành các phân khúc như cao cấp, trung cấp và giá rẻ. Điều này giúp xác định đối tượng khách hàng mục tiêu cho từng phân khúc.
* Chiến lược định giá: Phân tích giá trung bình giúp điều chỉnh chiến lược định giá để cạnh tranh hiệu quả hơn. Ví dụ, nếu một thương hiệu có giá trung bình cao nhưng doanh số thấp, có thể cần xem xét lại chiến lược định giá hoặc giá trị cung cấp.
* Tối ưu hóa danh mục sản phẩm: Hiểu rõ giá trung bình theo thương hiệu giúp quyết định loại sản phẩm nào nên được ưu tiên trong danh mục, đặc biệt là trong các chiến dịch tiếp thị hoặc khuyến mãi.
* Phân tích hành vi tiêu dùng: Kết hợp dữ liệu giá trung bình với các loại sự kiện như purchase để hiểu rõ hơn về hành vi mua sắm của khách hàng đối với các thương hiệu khác nhau.

##### 5.2.3. Sản phẩm theo category

q3 = spark.sql("SELECT category\_code, COUNT(DISTINCT product\_id) AS product\_count FROM events GROUP BY category\_code")

q3.write.mode("overwrite").option("header", "true").csv("hdfs://localhost:9000/output/q3\_csv")

- Phân tích: Phân phối số lượng sản phẩm theo từng danh mục (category) cho thấy sự chênh lệch rõ rệt giữa các ngành hàng. Những danh mục có số lượng sản phẩm lớn thường gắn liền với thị trường tiêu dùng phổ biến như điện tử, thời trang, hoặc đồ gia dụng – đồng thời cũng là những khu vực có mức độ cạnh tranh cao. Ngược lại, các danh mục ít sản phẩm có thể phản ánh khoảng trống trong nguồn cung hoặc nhu cầu thấp từ phía khách hàng.

- Chiến lược đề xuất:

* Tăng cường tối ưu hóa nội dung và trải nghiệm tìm kiếm cho các danh mục đông sản phẩm để giảm tỷ lệ thoát.
* Đối với danh mục ít sản phẩm: rà soát hiệu suất kinh doanh và hành vi tìm kiếm để quyết định giữ lại, mở rộng hoặc loại bỏ.
* Phân bổ ngân sách quảng cáo có trọng điểm theo độ phủ và tiềm năng doanh thu từng danh mục.

##### 5.2.4. Top 5 sản phẩm đắt nhất

q4 = spark.sql("SELECT product\_id, price FROM events ORDER BY price DESC LIMIT 5")

q4.write.mode("overwrite").option("header", "true").csv("hdfs://localhost:9000/output/q4\_csv")

- Phân tích: Top 5 sản phẩm có mức giá cao nhất đại diện cho nhóm hàng có giá trị cao, thường mang yếu tố độc quyền hoặc công nghệ cao. Dù đóng góp đáng kể vào tổng doanh thu trên mỗi đơn hàng, nhóm sản phẩm này cũng đòi hỏi chiến lược bán hàng đặc thù do rào cản giá và hành vi cân nhắc kỹ lưỡng từ phía người tiêu dùng.

- Chiến lược:

* Tập trung vào truyền thông giá trị, nhấn mạnh yếu tố chất lượng, thương hiệu, hoặc tính năng nổi bật.
* Áp dụng chiến lược tiếp cận cá nhân hóa, ưu tiên remarketing và targeting theo nhóm khách hàng có thu nhập cao.
* Nâng cao dịch vụ hậu mãi, tư vấn chuyên sâu, chính sách đổi trả minh bạch nhằm giảm rào cản tâm lý khi mua hàng giá cao.

##### 5.2.5. Doanh thu theo ngày

q5 = spark.sql("SELECT DATE(event\_time) AS day, SUM(price) AS total\_revenue FROM events GROUP BY day ORDER BY day")

q5.write.mode("overwrite").option("header", "true").csv("hdfs://localhost:9000/output/q5\_csv")

- Phân tích: Doanh thu theo ngày thể hiện xu hướng dao động mang tính chu kỳ, có thể liên quan đến hành vi tiêu dùng theo ngày trong tuần hoặc tác động từ các chương trình khuyến mãi. Các đỉnh doanh thu thường trùng với cuối tuần hoặc ngày khuyến mãi, trong khi các đáy rơi vào đầu tuần hoặc ngày thường không có chiến dịch kích cầu rõ rệt.

- Chiến lược:

* Thiết lập mô hình dự báo doanh thu theo thời gian để chủ động hoạch định khuyến mãi và điều phối hàng tồn kho.
* Đẩy mạnh các chương trình thúc đẩy tiêu dùng trong ngày thấp điểm, ví dụ: miễn phí vận chuyển, flash sale ngắn hạn.
* Phân tích tương quan giữa doanh thu – lượt truy cập – tỷ lệ chuyển đổi theo ngày để hiểu rõ nguyên nhân biến động và tối ưu hiệu suất hoạt động theo chu kỳ.

##### 5.2.6. Sự kiện theo giờ

q6 = spark.sql("SELECT HOUR(event\_time) AS hour, COUNT(\*) AS event\_count FROM events GROUP BY hour ORDER BY hour")

q6.write.mode("overwrite").option("header", "true").csv("hdfs://localhost:9000/output/q6\_csv")

- Phân tích: Dựa trên kết quả truy vấn số lượng sự kiện theo giờ trong ngày, ta quan sát thấy hoạt động người dùng đạt đỉnh vào buổi tối (14-17h), chiếm tỷ trọng lớn nhất trong tổng số hành vi truy cập. Trong khi đó, từ 21h tối đến 1h sáng hệ thống ghi nhận rất ít sự kiện – cho thấy đây là khoảng thời gian "ngủ yên" của nền tảng. Mô hình hành vi này phản ánh rõ rệt lối sống hiện đại, khi người dùng có xu hướng truy cập và mua sắm trực tuyến sau giờ làm việc

- Chiến lược:

* Tăng cường quảng cáo trong khung giờ cao điểm (19h–22h):Phân bổ ngân sách quảng cáo cao hơn cho Google Ads, Facebook, TikTok trong khung giờ người dùng hoạt động mạnh nhất.
* Tổ chức Flash Sale theo khung giờ vàng:Triển khai các chương trình giảm giá ngắn hạn, giới hạn số lượng từ 19h–21h để thúc đẩy hành vi mua hàng tức thời.
* Gửi thông báo tiếp thị (Push Notification / Email) đúng thời điểm:Gửi các thông báo nhắc nhở hoặc ưu đãi vào khoảng 17h–18h, để đón đầu thói quen người dùng truy cập buổi tối.
* Tự động mở rộng tài nguyên hệ thống (autoscaling):Cấu hình hạ tầng để tăng băng thông, CPU hoặc container xử lý vào các giờ cao điểm nhằm đảm bảo hiệu năng.
* Tận dụng khung giờ thấp điểm để chạy batch jobs hoặc bảo trì:Lên lịch xử lý dữ liệu lớn bằng Spark, ETL hoặc bảo trì hệ thống vào khoảng 1h–5h sáng khi lưu lượng thấp.
* Phân tích nâng cao hành vi theo giờ và loại sự kiện:Kết hợp phân tích theo event\_type (view, cart, purchase) để xác định chính xác điểm rơi của chuyển đổi (conversion) và cải thiện phễu bán hàng theo giờ.

##### 5.2.7. User duy nhất mỗi ngày

q7 = spark.sql("SELECT DATE(event\_time) AS day, COUNT(DISTINCT user\_id) AS unique\_users FROM events GROUP BY day")

q7.write.mode("overwrite").option("header", "true").csv("hdfs://localhost:9000/output/q7\_csv")

- Phân tích: Truy vấn q7 cho thấy số lượng người dùng duy nhất theo ngày, giúp đánh giá sức hút và khả năng duy trì lượng truy cập của nền tảng thương mại điện tử. Nếu biểu đồ thể hiện sự tăng trưởng ổn định hoặc có đỉnh điểm vào cuối tuần và dịp lễ, đây là tín hiệu tốt cho hoạt động marketing và chiến dịch ưu đãi. Ngược lại, nếu xu hướng người dùng giảm dần theo thời gian, điều này có thể phản ánh sự suy giảm về hiệu quả tiếp thị, chất lượng trải nghiệm hoặc mức độ giữ chân người dùng.

- Chiến lược:

* Tăng cường chiến dịch thu hút người dùng vào cuối tuần hoặc ngày cao điểm:Đẩy mạnh quảng cáo và khuyến mãi vào các ngày có lưu lượng user tự nhiên cao để tối ưu chuyển đổi.
* Theo dõi và phân tích các đợt tăng trưởng đột biến:Gắn kết dữ liệu với lịch chiến dịch hoặc kênh marketing để xác định nguyên nhân thành công và áp dụng lại.
* Xây dựng mô hình dự báo số người dùng truy cập theo ngày:Dùng ARIMA, Prophet hoặc các mô hình học máy để dự đoán lưu lượng giúp chủ động lên kế hoạch vận hành.
* Giám sát xu hướng giảm và phản ứng nhanh:Nếu lượng user giảm, cần kiểm tra lại trải nghiệm người dùng, chính sách giá, nội dung sản phẩm và tối ưu hóa lại các kênh thu hút.
* Kết hợp chỉ số này với dữ liệu lịch sử để tính retention rate:Đánh giá khả năng giữ chân người dùng và triển khai các chương trình khách hàng thân thiết nếu tỷ lệ quay lại thấp.

##### 5.2.8. Top 10 sessions hoạt động nhiều nhất

q8 = spark.sql("SELECT user\_session, COUNT(\*) AS event\_count FROM events GROUP BY user\_session ORDER BY event\_count DESC LIMIT 10")

q8.write.mode("overwrite").option("header", "true").csv("hdfs://localhost:9000/output/q8\_csv")

##### 5.2.9. Tỷ lệ loại sự kiện

q9 = spark.sql("""

    SELECT event\_type,

           ROUND(COUNT(\*) \* 100.0 / (SELECT COUNT(\*) FROM events), 2) AS percentage

    FROM events GROUP BY event\_type

""")

q9.write.mode("overwrite").option("header", "true").csv("hdfs://localhost:9000/output/q9\_csv")

- Phân tích: Truy vấn q8 xác định 10 phiên truy cập (user\_session) có số lượng sự kiện nhiều nhất, giúp phát hiện những phiên người dùng hoạt động mạnh, có tương tác sâu với hệ thống. Những phiên này có thể đại diện cho hành vi mua sắm nghiêm túc, tìm kiếm sản phẩm kỹ lưỡng hoặc thậm chí là hành vi bất thường (spam/bot). Phân tích sâu các phiên này sẽ giúp hiểu rõ hơn về các dạng hành vi tiêu biểu: từ người mua tiềm năng đến khách hàng trung thành hoặc người dùng mất chuyển đổi (drop-off sau nhiều tương tác).

- Chiến lược:

* Phân tích chi tiết hành vi của các phiên truy cập top đầu:Truy vết loại sự kiện trong từng session (view, cart, purchase) để hiểu rõ hành trình mua sắm và tối ưu phễu chuyển đổi.
* Tận dụng dữ liệu phiên tương tác cao để huấn luyện mô hình recommendation:Các session có nhiều tương tác thường chứa nhiều tín hiệu sở thích, phù hợp để huấn luyện hệ thống gợi ý sản phẩm.
* Theo dõi dấu hiệu bất thường để phát hiện bot hoặc lạm dụng:Nếu một số session có lượng tương tác bất thường mà không mua hàng, có thể cần kiểm tra chống spam/click fraud.
* Tạo phân khúc người dùng theo độ sâu tương tác:Sử dụng số lượng sự kiện trong session để phân nhóm người dùng tiềm năng (high engagement) và triển khai remarketing hiệu quả hơn.
* Cải thiện UX dựa trên hành vi tương tác nhiều:Dựa vào các bước phổ biến trong session dài, cải thiện giao diện hoặc rút ngắn quy trình mua hàng để giảm bỏ cuộc giữa chừng.

##### 5.2.10. Giá trung bình theo ngày 30-11-2019 của nhãn hàng

q10 = spark.sql("""

    SELECT DATE(event\_time) AS day, brand, AVG(price) AS avg\_price

    FROM events

    WHERE brand IS NOT NULL AND DATE(event\_time)=’2019-11-30’

    GROUP BY day

""")

q10.write.mode("overwrite").option("header", "true").csv("hdfs://localhost:9000/output/q10\_csv")

- Phân tích: Truy vấn q10 cung cấp cái nhìn theo thời gian về giá trung bình của từng thương hiệu (brand) ngày 30-11-2019, giúp theo dõi sự biến động giá của sản phẩm từ nhiều hãng khác nhau. Qua đó, doanh nghiệp có thể phát hiện các đợt tăng/giảm giá theo mùa, chiến dịch hoặc cạnh tranh. Những thương hiệu có giá ổn định thể hiện tính nhất quán và định vị thị trường rõ ràng, trong khi thương hiệu có giá biến động nhiều có thể đang điều chỉnh chính sách giá linh hoạt để phù hợp nhu cầu hoặc cạnh tranh.

- Chiến lược:

* Theo dõi biến động giá để định hướng chiến lược giá linh hoạt (dynamic pricing):Điều chỉnh giá bán theo xu hướng thị trường và hành vi người dùng, giúp tối ưu doanh thu mà vẫn giữ tính cạnh tranh.
* Tận dụng thời điểm giá thấp để chạy chiến dịch marketing:Khi giá trung bình của brand giảm, có thể đẩy mạnh truyền thông như “Giá tốt nhất tháng này từ thương hiệu X”.
* So sánh giữa các thương hiệu cùng phân khúc:Phân tích giá trung bình theo thời gian giúp so sánh cạnh tranh trực tiếp giữa các thương hiệu và định vị lại sản phẩm nếu cần.
* Phát hiện sớm lỗi dữ liệu hoặc định giá sai:Những biến động giá bất thường có thể chỉ ra lỗi nhập liệu, sai lệch giá bán hoặc sự cố từ phía nhà cung cấp.
* Kết hợp dữ liệu này để phân tích mối quan hệ giữa giá và hành vi mua:Gắn dữ liệu giá với tỷ lệ purchase để xác định ngưỡng giá lý tưởng cho từng thương hiệu.

A screenshot of a computer program

AI-generated content may be incorrect.

A screenshot of a computer program

AI-generated content may be incorrect.

Các truy vấn sau khi được thực hiện và lưu lên HDFS

A screenshot of a computer

AI-generated content may be incorrect.

## 6. Trực quan hóa kết quả bằng công cụ Power BI

Dữ liệu kết quả được tải xuống từ HDFS về máy tính cục bộ hoặc đọc trực tiếp từ Spark

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

 • Import file .csv vào Power BI.

A screenshot of a computer

AI-generated content may be incorrect.

 Các biểu đồ được xây dựng:

|  |  |  |
| --- | --- | --- |
| Truy vấn | Mục đích | Biểu đồ |
| 1 | Tỉ lệ các loại sự kiện | Hình tròn |
| 2 | Giá trị trung bình theo thương hiệu | Cột ngang |
| 3 | Tổng số sản phẩm theo danh mục | Bảng |
| 4 | Top 5 sản phẩm đắt nhất | Đường |
| 5 | Danh thu theo ngày | Đường |
| 6 | Sự kiện theo giờ trong ngày | Biểu đồ cột |
| 7 | Người dùng duy nhất mỗi ngày | Vùng |
| 8 | 10 session hoạt động nhiều nhất | Cột |
| 9 | Tỷ lệ các loại sự kiện | Vành |
| 10 | Giá trung bình theo ngày và brand | Cột |

#### 6.1. Số lượng từng loại sự kiện

A purple circle with a blue and orange bar

AI-generated content may be incorrect.

#### 6.2 Giá trị trung bình theo thương hiệu

A graph with blue lines and black text

AI-generated content may be incorrect.

#### 6.3 Tổng số sản phẩm theo danh mục

Nhiều nhất:

A screenshot of a computer

AI-generated content may be incorrect.

Ít nhất:

A screenshot of a computer

AI-generated content may be incorrect.

#### 6.5 Doanh thu theo ngày

A graph showing a line of a graph

AI-generated content may be incorrect.

#### 6.6 Sự kiện theo giờ trong ngày

A graph of blue bars

AI-generated content may be incorrect.

#### 6.7 Người dùng duy nhất mỗi ngày

A blue line graph with white text

AI-generated content may be incorrect.

#### 6.8 10 session hoạt động nhiều nhất

A graph of blue squares

AI-generated content may be incorrect.

#### 6.9 Tỷ lệ các loại sự kiện

A blue circle with a number of numbers and a number of text

AI-generated content may be incorrect.

#### 6.10 Giá trị trung bình theo ngày 30-11-2019 của thương hiệu

A blue and white bar graph

AI-generated content may be incorrect.

# **CHƯƠNG II. KÊT LUẬN**

Trong thời đại dữ liệu lớn bùng nổ, việc khai thác và phân tích hành vi người tiêu dùng thông qua các nền tảng như Hadoop và Spark không chỉ là một lợi thế cạnh tranh mà còn là yêu cầu thiết yếu đối với doanh nghiệp thương mại điện tử. Qua quá trình thực hiện đề tài, nhóm đã tiếp cận một cách toàn diện quy trình phân tích dữ liệu lớn từ khâu thu thập, xử lý đến trực quan hóa.

Kết quả phân tích cho thấy các xu hướng đáng chú ý trong hành vi tiêu dùng, như: tỷ lệ lớn người dùng dừng lại ở bước xem sản phẩm mà không thực hiện mua hàng, chênh lệch giá giữa các thương hiệu, và sự phân bố sản phẩm giữa các ngành hàng. Các truy vấn Spark được sử dụng để làm sáng tỏ những khía cạnh cụ thể như giá trung bình theo thương hiệu, số lượng sản phẩm theo danh mục, sản phẩm đắt nhất và doanh thu theo ngày. Từ đó, nhóm đã đưa ra được các gợi ý chiến lược mang tính thực tiễn nhằm nâng cao hiệu quả kinh doanh, tối ưu tỷ lệ chuyển đổi và cải thiện trải nghiệm khách hàng.

Không chỉ dừng lại ở kết quả, đề tài còn mang đến cơ hội cọ xát với các công nghệ cốt lõi trong lĩnh vực dữ liệu lớn như Spark SQL, HDFS và mô hình xử lý phân tán. Đây là nền tảng quan trọng giúp nhóm xây dựng năng lực chuyên môn vững chắc, phục vụ cho các bài toán thực tế trong tương lai.