**TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI**

**Viện Công nghệ thông tin và Truyền thông**



**Báo cáo môn học**

**Lưu trữ và xử lý dữ liệu lớn**

***Đề tài: Phân tích dữ liệu phục vụ kinh doanh***

**GVHD: Trần Việt Trung**

***Nhóm 14***: Giang Công Nguyên -20173289

Phạm Thanh Hải -20173092

Trần Tuấn Anh -20172944

Lương Thế Huy -20173172

Hà Nội ,tháng 12/2020

# Mục lục

[Mục lục 2](#_Toc60518416)

[1.Giới thiệu 3](#_Toc60518417)

[1.1 Giới thiệu đề tài 3](#_Toc60518418)

[1.2 Sàn thương mại điện tử Shopee 3](#_Toc60518419)

[1.3 Mục tiêu 3](#_Toc60518420)

[2. Cơ sở lí thuyết 4](#_Toc60518421)

[2.1 Hadoop 4](#_Toc60518422)

[2.2 Spark 5](#_Toc60518423)

[3.Triển khai 7](#_Toc60518424)

[3.1 Crawl dữ liệu 7](#_Toc60518425)

[3.2 Lưu trữ 8](#_Toc60518426)

[3.3 Xử lý 8](#_Toc60518427)

[3.4 Một số kết quả 9](#_Toc60518428)

[4.Kết quả thu được 10](#_Toc60518429)

[5.Kết luận 11](#_Toc60518430)

# 1.Giới thiệu

## 1.1 Giới thiệu đề tài

+Đặt vấn đề :Ngày nay, việc mua sắm online trên các sàn thương mại điện tử ngày càng trở nên phổ biến và thông dụng với chúng ta. Từ nhu cầu đó, rất nhiều người bắt đầu khởi nghiệp bằng cách kinh doanh online trên các sàn thương mại như Shopee, Sendo, Tiki…Bạn đang muốn kinh doanh một mặt hàng với hình thức online mà không biết lên lựa chọn mặt hàng nào ,xu hướng người mua sẽ thích các sản phẩm trong tầm giá nào ,bên cạnh đó còn là cách bán sản phẩm đó thể nào để tối ưu hóa lượng người mua…

=>Vì vậy, nhóm chúng em đã tiến hành nghiên cứu và thực hiện đề tài :*Phân tích dữ liệu phục vụ kinh doanh* sử dụng dữ liệu trực tiếp từ các sàn thương mại điện tử và các công nghệ khoa học dữ liệu cơ bản với mục đích giải quyết được một phần của vấn đề nêu trên.

## 1.2 Sàn thương mại điện tử Shopee

+Shopee là nền tảng thương mại điện tử hàng đầu ở Đông Nam Á và Đài Loan. Đây là một nền tảng phù hợp với khu vực, cung cấp cho cả người mua và người bán trải nghiệm mua sắm trực tuyến dễ dàng, an toàn và nhanh chóng thông qua thanh toán mạnh mẽ và hỗ trợ hậu cần.

Vì vậy trong đề tài này, chúng em sử dụng trực tiếp dữ liệu về các mặt hàng có trên Shopee để tiến hành xử lí và phân tích.

## 1.3 Mục tiêu

* Mục tiêu đưa ra được xu hướng người mua sẽ thích các sản phẩm trong

tầm giá nào,mức giá nào có lượng người mua nhiều để từ đó người bán có thể chọn các loại sản phẩm trong tầm giá đó để bán…

* Bên cạch đó còn là việc tối ưu hóa để làm sao bán hiệu quả nhất
* Người dùng thích mức giảm giá bao nhiêu.
* Độ dài video quảng cáo tối ưu là bao nhiêu
* Các sản phẩm bán chạy thường sẽ có những tính chất gì:số option,

thời lượng video(nếu có),giảm giá…

# 2. Cơ sở lí thuyết

## 2.1 Hadoop

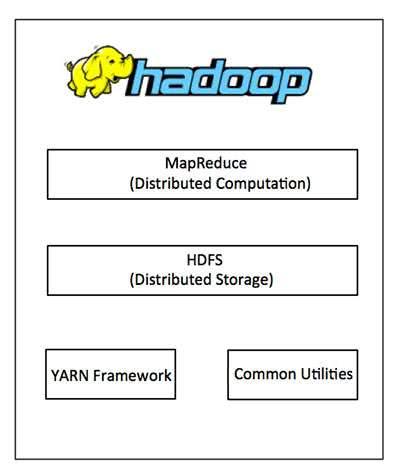
Hadoop là một Apache framework mã nguồn mở cho phép phát triển các ứng dụng phân tán (distributed processing) để lưu trữ và quản lý các tập dữ liệu lớn. Hadoop hiện thực mô hình MapReduce, mô hình mà ứng dụng sẽ được chia nhỏ ra thành nhiều phân đoạn khác nhau được chạy song song trên nhiều node khác nhau. Hadoop được viết bằng Java tuy nhiên vẫn hỗ trợ C++, Python, Perl bằng cơ chế streaming.

Hadoop giải quyết vấn đề gì?

* Xử lý và làm việc khối lượng dữ liệu khổng lồ tính bằng Petabyte.
* Xử lý trong môi trường phân tán, dữ liệu lưu trữ ở nhiều phần cứng khác nhau, yêu cầu xử lý đồng bộ
* Các lỗi xuất hiện thường xuyên.
* Băng thông giữa các phần cứng vật lý chứa dữ liệu phân tán có giới hạn.

Kiến trúc Hadoop là gì?

Một cụm Hadoop nhỏ gồm 1 master node và nhiều worker/slave node. Toàn bộ cụm chứa 2 lớp, một lớp MapReduce Layer và lớp kia là HDFS Layer. Mỗi lớp có các thành phần liên quan riêng. Master node gồm JobTracker, TaskTracker, NameNode, và DataNode. Slave/worker node gồm DataNode, and TaskTracker. Cũng có thể slave/worker node chỉ là dữ liệu hoặc node để tính toán.



Các điểm thuận lợi khi dùng Hadoop:

* Robust and Scalable – Có thể thêm node mới và thay đổi chúng khi cần.
* Affordable and Cost Effective – Không cần phần cứng đặc biệt để chạy Hadoop.
* Adaptive and Flexible – Hadoop được xây dựng với tiêu chí xử lý dữ liệu có cấu trúc và không cấu trúc.
* Highly Available and Fault Tolerant – Khi 1 node lỗi, nền tảng Hadoop tự động chuyển sang node khác.

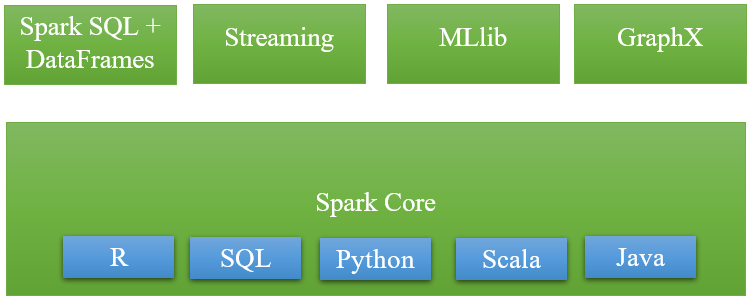
## 2.2 Spark

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.

Tốc độ xử lý của Spark có được do việc tính toán được thực hiện cùng lúc trên nhiều máy khác nhau. Đồng thời việc tính toán được thực hiện ở bộ nhớ trong (in-memories) hay thực hiện hoàn toàn trên RAM.

Spark cho phép xử lý dữ liệu theo thời gian thực, vừa nhận dữ liệu từ các nguồn khác nhau đồng thời thực hiện ngay việc xử lý trên dữ liệu vừa nhận được (Spark Streaming).

Spark không có hệ thống file của riêng mình, nó sử dụng hệ thống file khác như: HDFS, Cassandra, S3 …. Spark hỗ trợ nhiều kiểu định dạng file khác nhau (text, csv, json…) đồng thời nó hoàn toàn không phụ thuộc vào bất cứ một hệ thống file nào.



Apache Spark gồm có 5 thành phần chính: Spark Core, Spark Streaming, Spark SQL, MLlib và GraphX, trong đó:

Những điểm nổi bật của Spark

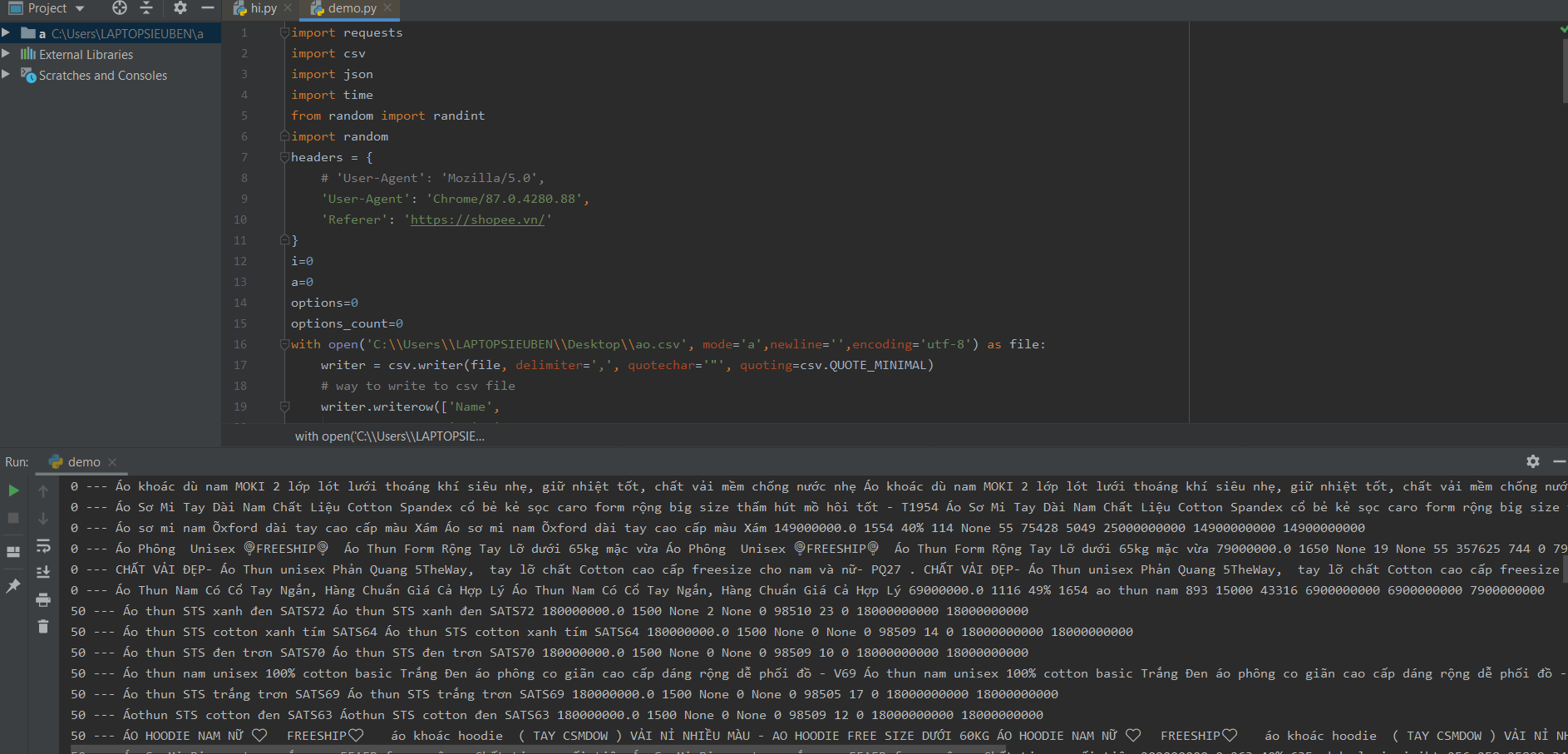
* Xử lý dữ liệu: Spark xử lý dữ liệu theo lô và thời gian thực
* Tính tương thích: Có thể tích hợp với tất cả các nguồn dữ liệu và định dạng tệp được hỗ trợ bởi cụm Hadoop.
* Hỗ trợ ngôn ngữ: hỗ trợ Java, Scala, Python và R.
* Phân tích thời gian thực:
  + Apache Spark có thể xử lý dữ liệu thời gian thực tức là dữ liệu đến từ các luồng sự kiện thời gian thực với tốc độ hàng triệu sự kiện mỗi giây. Ví dụ: Data Twitter chẳng hạn hoặc lượt chia sẻ, đăng bài trên Facebook. Sức mạnh Spark là khả năng xử lý luồng trực tiếp hiệu quả.
  + Apache Spark có thể được sử dụng để xử lý phát hiện gian lận trong khi thực hiện các giao dịch ngân hàng. Đó là bởi vì, tất cả các khoản thanh toán trực tuyến được thực hiện trong thời gian thực và chúng ta cần ngừng giao dịch gian lận trong khi quá trình thanh toán đang diễn ra.

=>Vì vậy trong trong bài có sử dụng Hadoop để lưu trữ dữ liệu và Pyspark để xử lý dữ liệu.

# 3.Triển khai

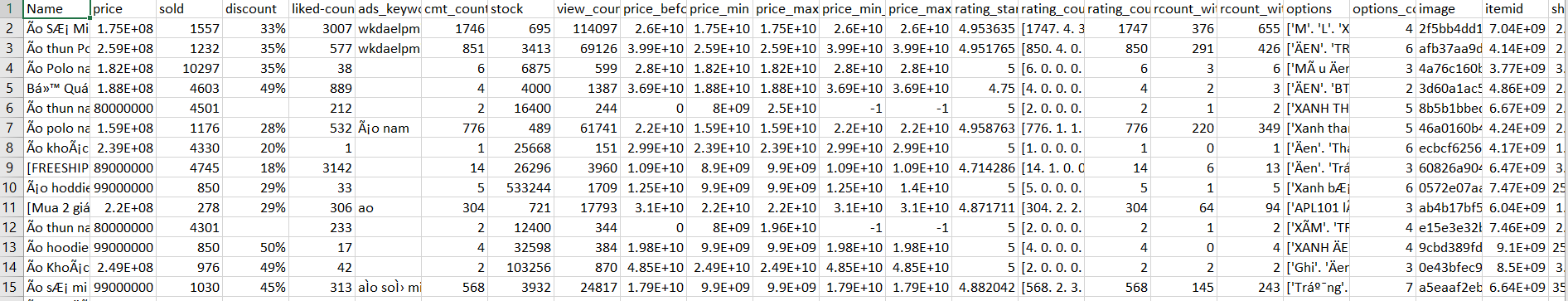
## 3.1 Crawl dữ liệu

* Crawl dữ liệu từ: Shopee
* Kết quả thu được:
* Có 3 loại dữ liệu phổ biến
* Gồm 27 trường
* Mỗi loại dữ liệu có khoảng gần 10.000 bản ghi



Màn hình crawl dữ liệu

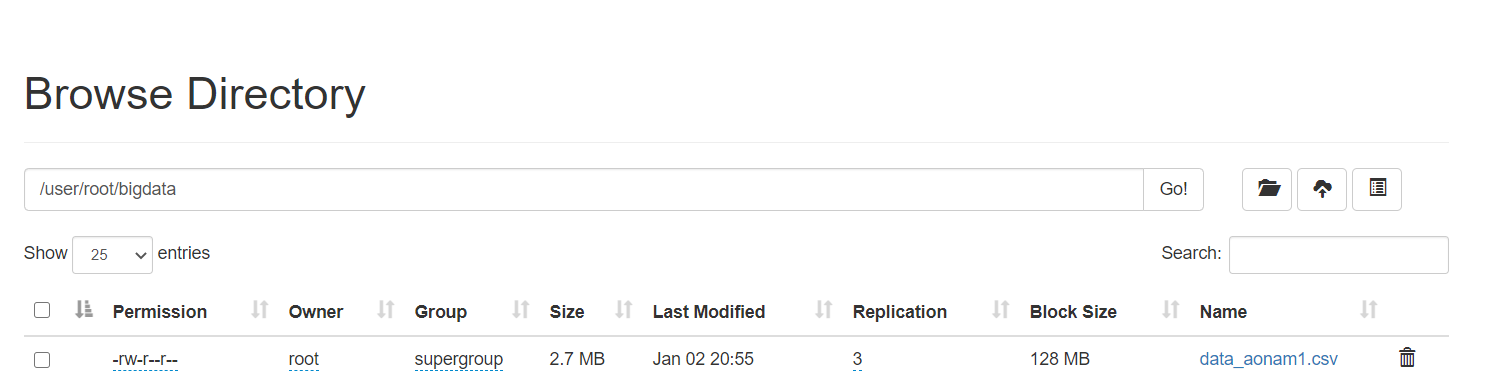
* Dữ liệu thu được



Dữ liệu thu được .csv

## 3.2 Lưu trữ

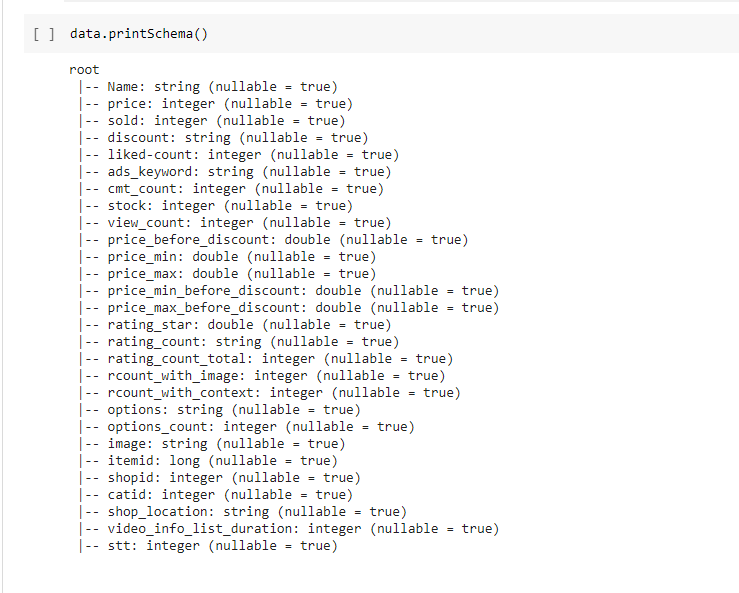
+Dữ liệu được lưu trữ bằng việc sử dụng Hadoop để lưu dữ liệu phân tán



## 3.3 Xử lý

+Dữ liệu được xử lý cho loại sản phẩm áo nam(data\_aonam1.csv).

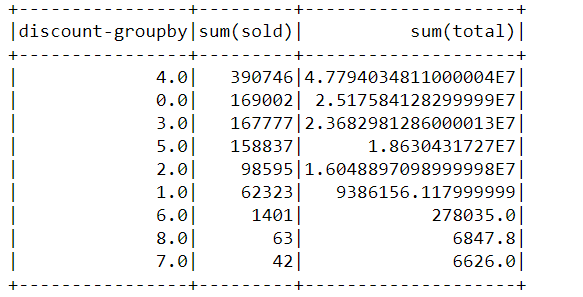
+Chi tiết phần xử lý nhóm em để trong file Demo/demo1a.ipynp.



* Dữ liệu trước khi đem xử lý sẽ được làm sạch.

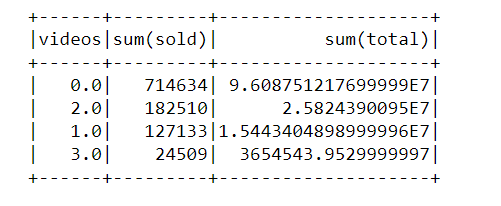
## 3.4 Một số kết quả

* Mức khuyến mại được ưu chuộng(40%-50%)



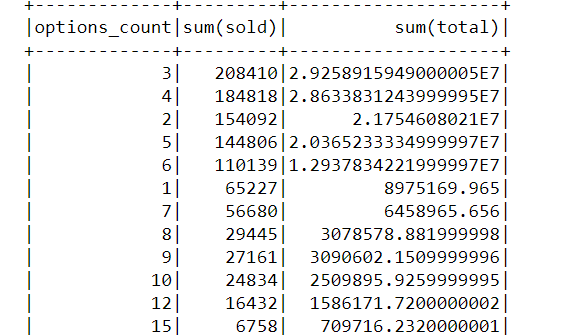
bucketBorders = [0, 10, 20, 30,40,50, 60, 70, 80,90]

* Thời lượng video tối ưu nếu có

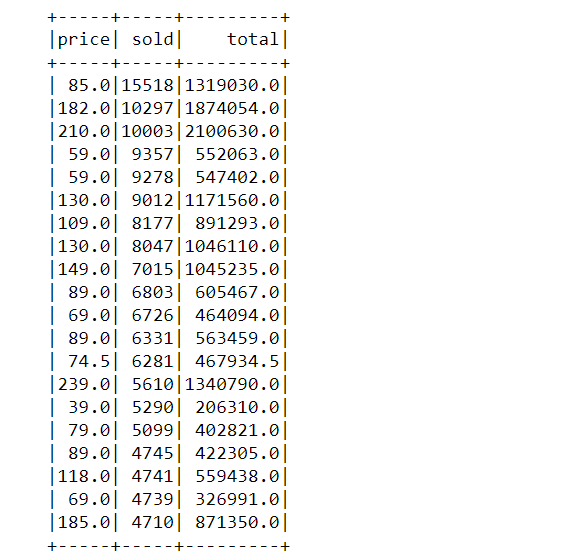


bucketBorders = [0, 1, 30,60,1000]

* Số option tối ưu



* Người mua thích các loại sản phẩm trong tầm giá nào



# 4.Kết quả thu được

+Sau khi thực hiện các quá trình đã nêu trên, chương trình cho kết quả thu được như hình dưới đây.



+Lưu ý : Kết quả mục Price là để giúp người bán biết với sản phẩm “Áo nam” thì người mua thích mua ,có xu hướng chọn những sản phẩm có giá như trên,do đó người bán có thể chọn những loại sản phẩm có thể bán với nhứng giá trên để bán để tối ưu lượng mua.

# 5.Kết luận

* Qua phân tích đã nêu được phần nào xu hướng,sở thích của người mua do đó giúp ích cho người bán việc chọn sản phẩm để bán và một số cách để tối ưu hóa lượng người mua.
* Do dữ liệu chưa được nhiều(chỉ từ shoppe) lên kết quả có thể chưa sát nhất với thực tế,việc xử lý dữ liệu có thể còn nhiều sai sót.

Hết