Trường đại học tài nguyên và môi trường Thành phố Hồ Chí Minh

Báo cáo cuối kỳ

***Chủ đề: Thống kê sự quan tâm của người dùng Facebook đối với một số sản phẩm*.**

Môn học: Công nghệ dữ liệu lớn
Giảng viên hướng dẫn: Phạm Trọng Huynh

Lớp: 10_ĐH_CNTT3
Thành viên nhóm 2:
Võ Minh Hân: 1050080096
Trà Ngọc Thông: 1050080120
Lâm Thị Phương Thảo (Nhóm trưởng): 1050080118

NỘI DUNG

I. TỔNG QUAN BIG DATA

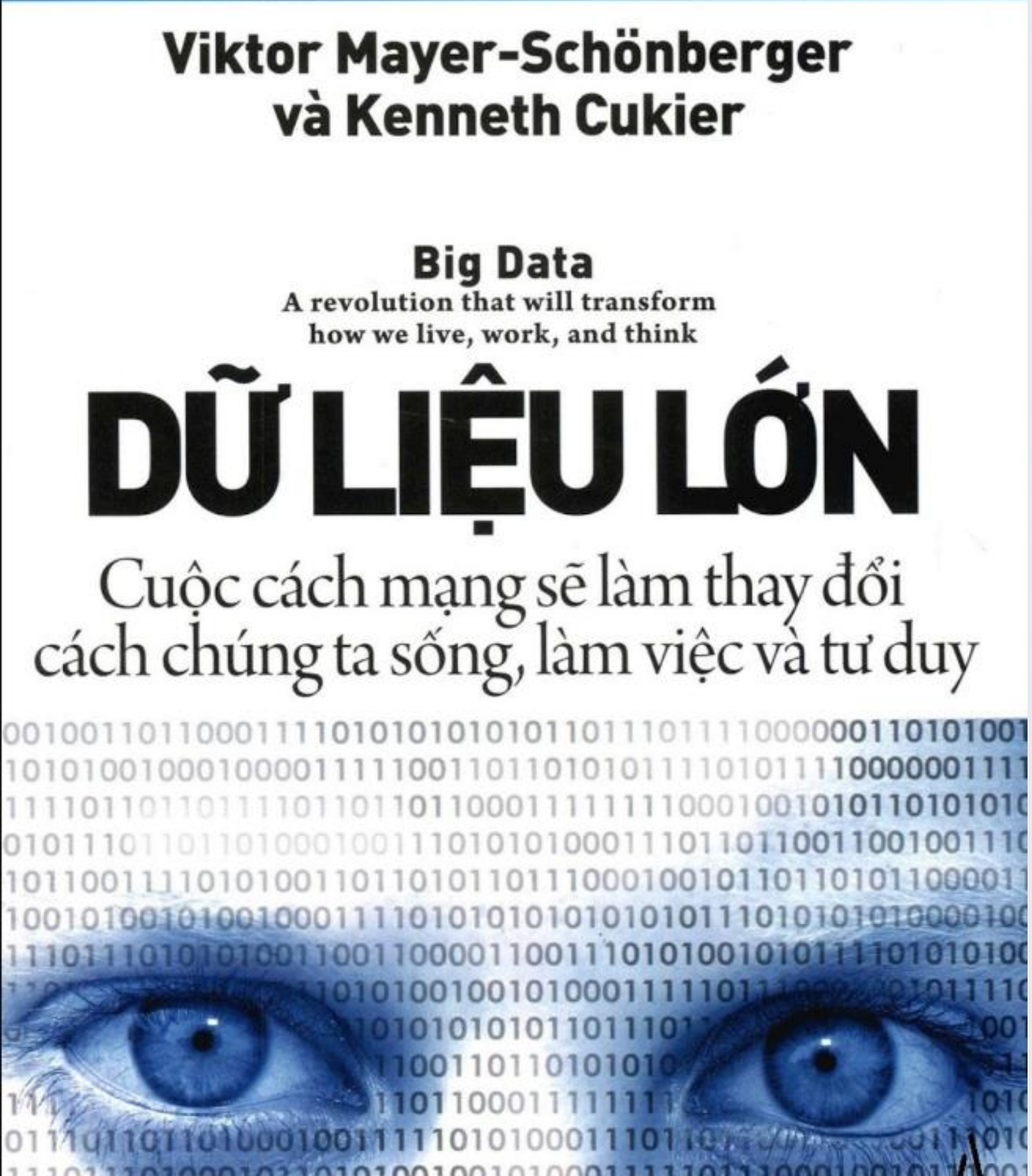
II. HỆ SINH THÁI HADOOP VÀ SPARK

III. BÀI TOÁN THỰC HIỆN

IV. PHƯƠNG PHÁP VÀ MÔ HÌNH

V. ƯU ĐIỂM VÀ NHƯỢC ĐIỂM CỦA MÔ HÌNH

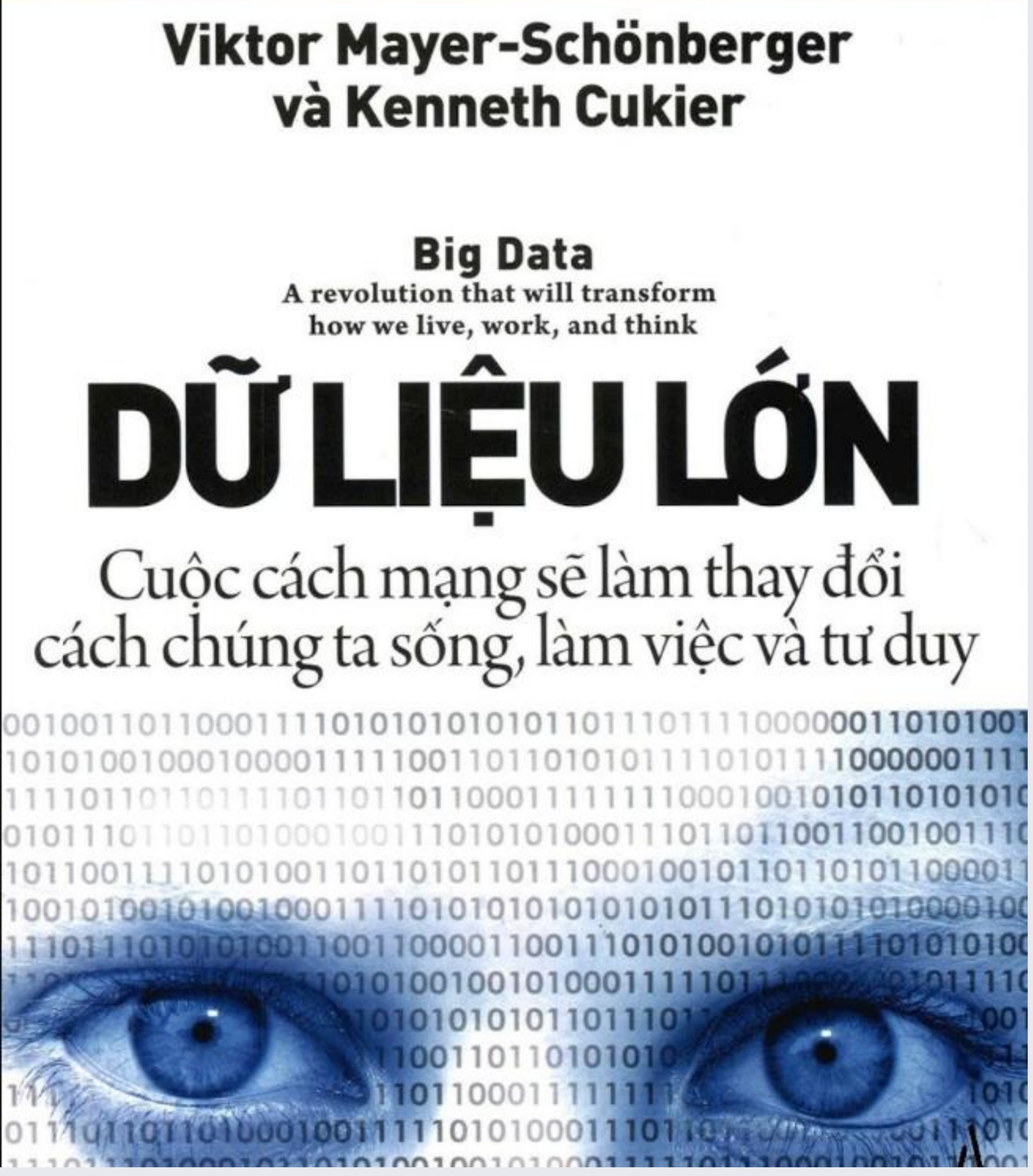
VI. PHÁT TRIỂN



I. Tổng quan Big Data

BIG DATA LÀ GÌ ?

Big Data không chỉ đánh dấu sự thay đổi trong cách chúng ta làm việc và tương tác với thông tin, mà còn mở ra một loạt cơ hội và thách thức cho nhiều lĩnh vực khác nhau, bao gồm công nghiệp, khoa học, chính trị, và xã hội học. Trong luận văn này, chúng ta sẽ khám phá sâu hơn về khái niệm Big Data, ý nghĩa của nó, và những ảnh hưởng to lớn mà nó mang lại.

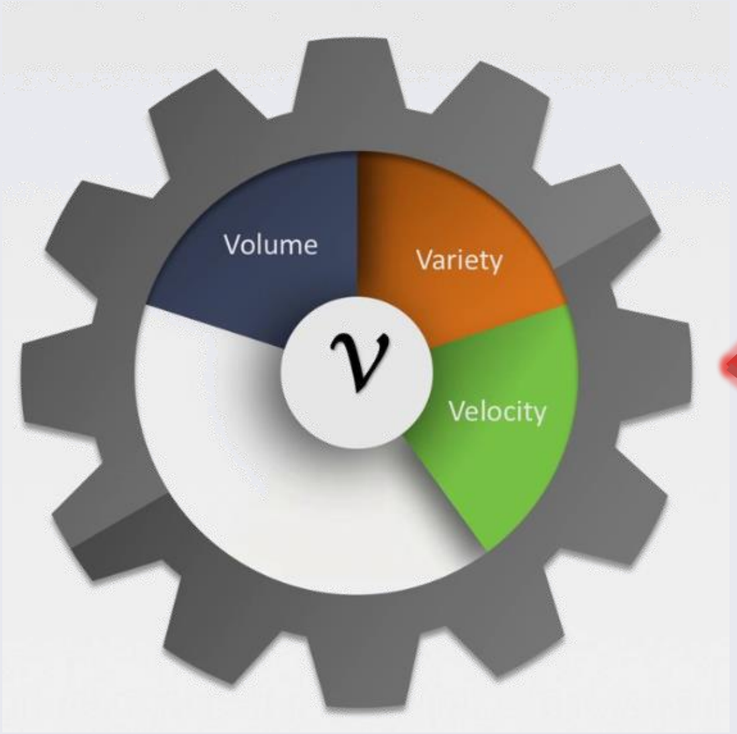


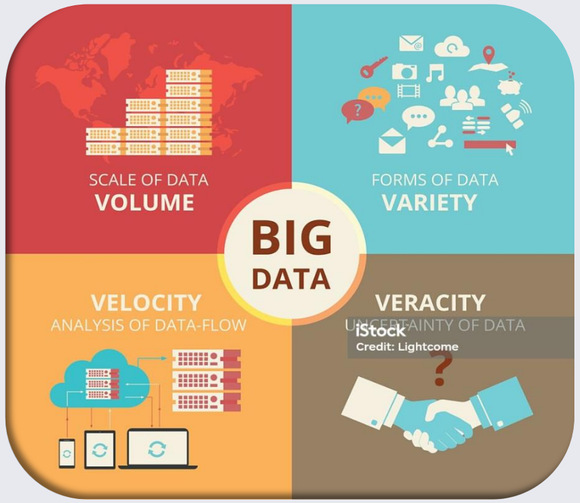
I. Tổng quan Big Data

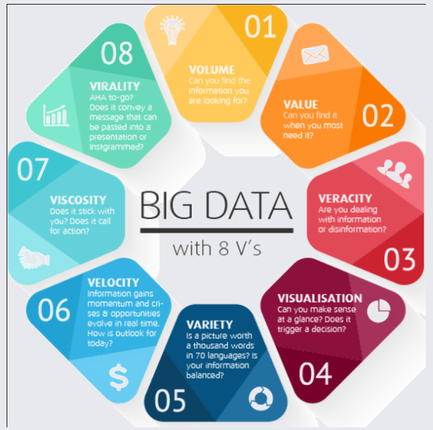
LỊCH SỬ BIG DATA

Dự án dữ liệu lớn đầu tiên bắt đầu vào năm 1937 theo lệnh của Chính quyền Franklin D. Roosevelt tại Hoa Kỳ. Máy xử lý dữ liệu đầu tiên là Colossus, phát triển bởi người Anh vào năm 1943 để giải mã mật mã Đức Quốc xã trong Chiến tranh Thế giới II. Từ những năm 90, dữ liệu lớn trở nên phổ biến khi nhiều thiết bị được kết nối với Internet.

Đặc trưng

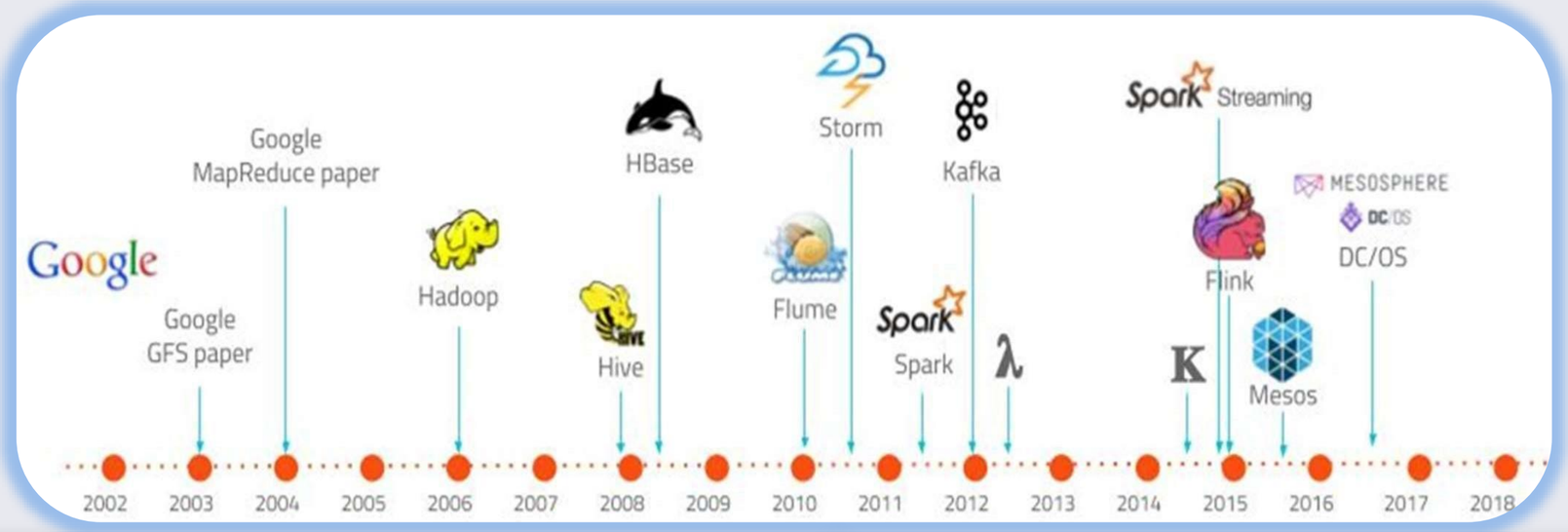

Volume
Variety
Velocity
V

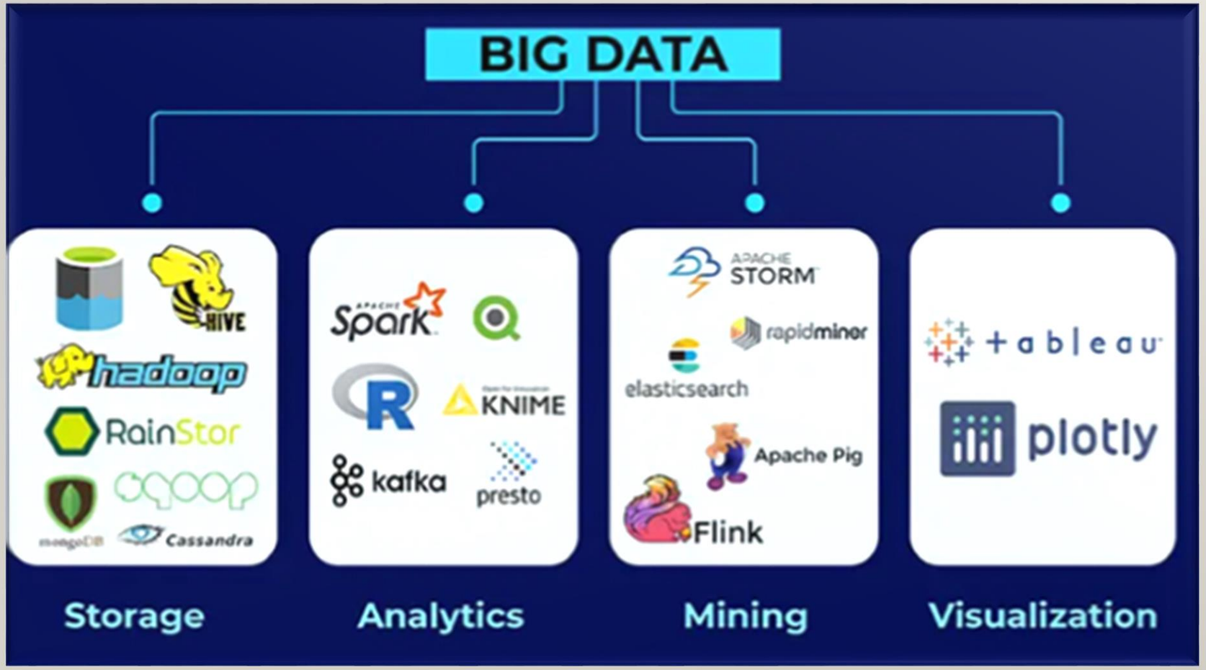

SCALE OF DATA
VOLUME
FORMS OF DATA
VARIETY
BIG
DATA
VELOCITY
ANALYSIS OF DATA-FLOW
VERACITY
UNCERTAINTY OF DATA
iStock
Credit: Lightcome


01
VOLUME
Can you find the information you are looking for?
02
VALUE
Can you find it when you most need it?
03
VERACITY
Are you dealing with information or disinformation?
04
VISUALISATION
Can you make sense at a glance? Does it trigger a decision?
05
VARIETY
Is a picture worth a thousand words in 70 languages? Is your information balanced?
06
VELOCITY
Information gains momentum and ci- ses & opportunities evolve in real time. How is outlook for today?
07
VISCOSITY
Does it stick with you? Does it call for action?
08
VIRALITY
AHA to-go? Does it convey a message that can be passed into a presentation or instagrammed?
BIG DATA
with 8 V's


1 Volume
2 Velocity
3 Variety
4 Veracity
5 Value

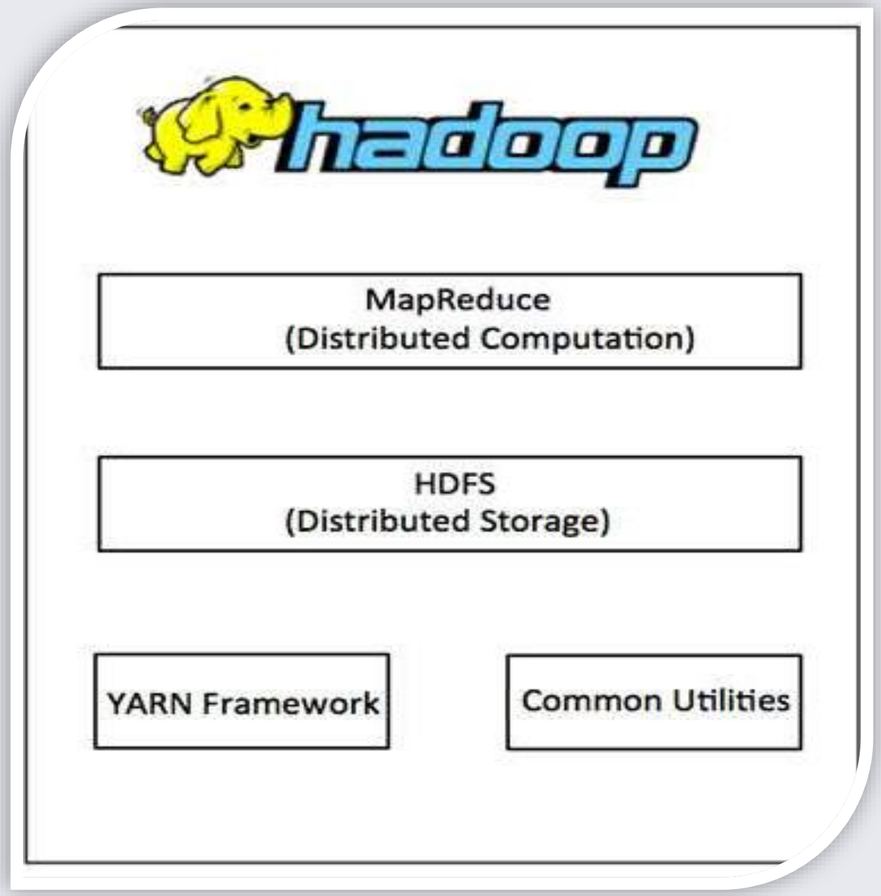
Công cụ được sử dụng trong big data





I. Hệ sinh thái Hadoop và Spark

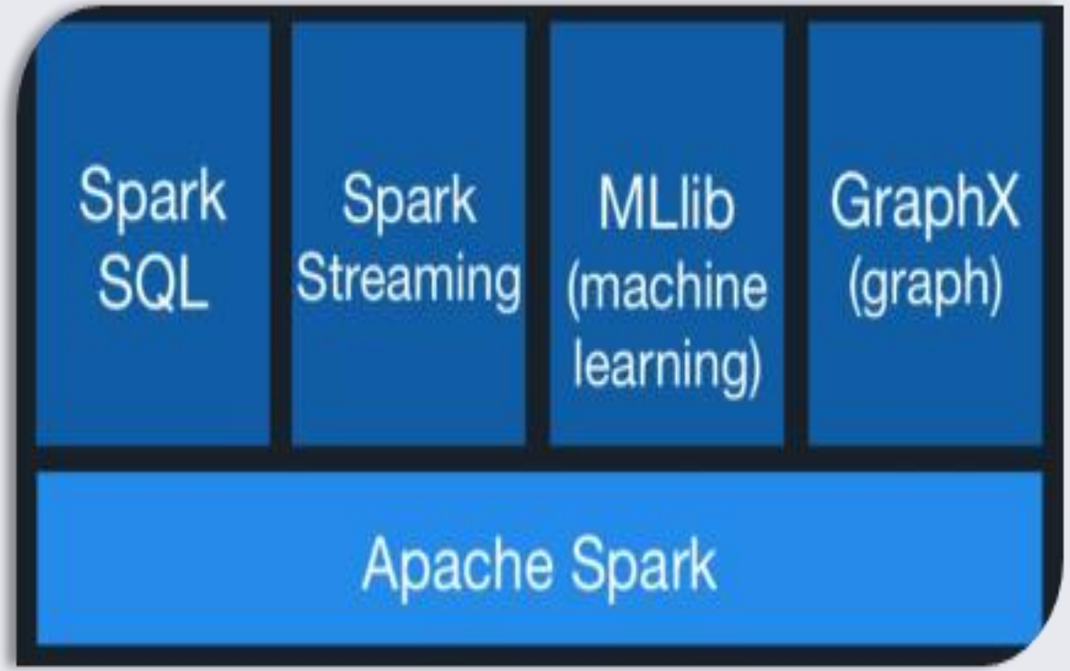
Hadoop là một framework mã nguồn mở quan trọng cho lưu trữ và xử lý dữ liệu lớn. Nó bao gồm Hadoop Distributed File System (HDFS) cho việc lưu trữ dữ liệu và MapReduce cho xử lý dữ liệu song song.



- Có thể thêm node mới và thay đổi chúng khi cần.
- Không cần phần cứng đặc biệt để chạy Hadoop.
- Hadoop được xây dựng với tiêu chí xử lý dữ liệu có cấu trúc và không cấu trúc.
- Khi 1 node lỗi, nền tảng Hadoop tự động chuyển sang node khác.

I. Hệ sinh thái Hadoop và Spark

Apache Spark bắt đầu như một dự án nghiên cứu tại UC Berkeley AMPLab vào năm 2009 và có nguồn mở vào đầu năm 2010. Nhiều ý tưởng đằng sau hệ thống đã được trình bày trong nhiều tài liệu nguyên cứu khác nhau trong nhiều năm.

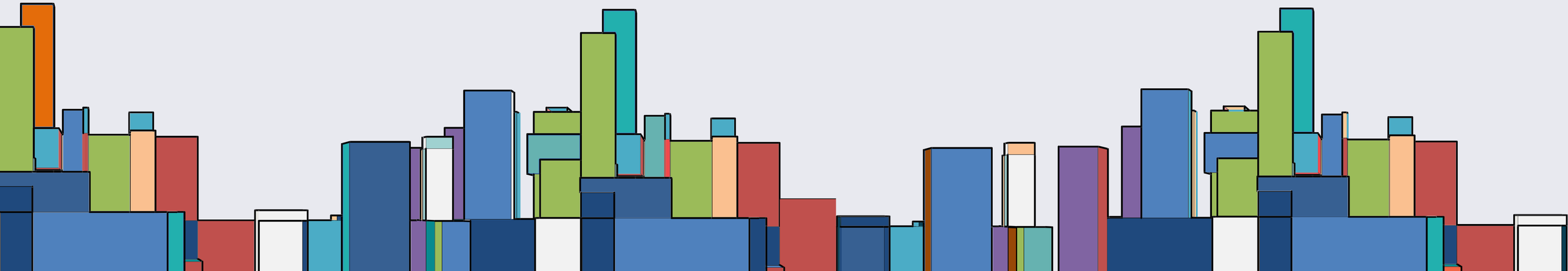


- Khả năng xử lý dữ liệu với tốc độ cao
- Khả năng tương thích cao
- Hỗ trợ nhiều ngôn ngữ
- Điều chỉnh được độ trễ:

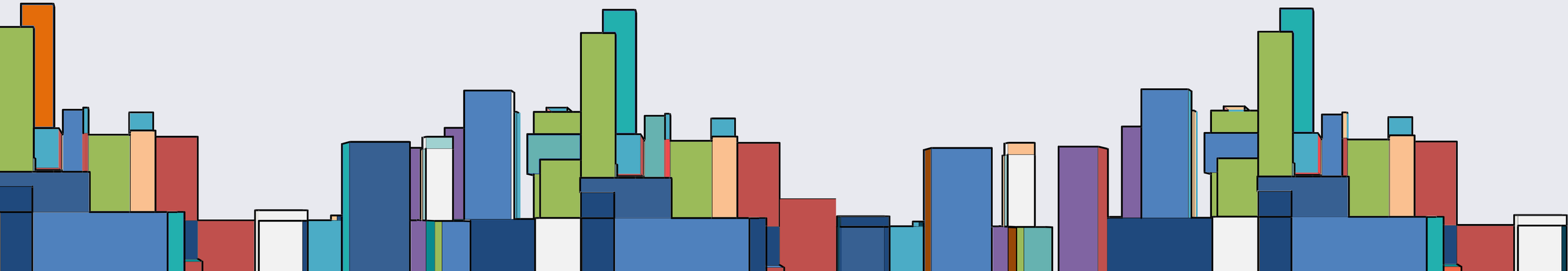
III. BÀI TOÁN THỰC HIỆN

THỐNG KÊ SỰ QUAN TÂM CỦA NGƯỜI DÙNG FACEBOOK ĐỐI VỚI MỘT SỐ SẢN PHẨM.

Trong thời đại số hóa hiện nay, việc thu thập và phân tích dữ liệu về sự quan tâm của người dùng Facebook đối với các sản phẩm trở thành một phần không thể thiếu đối với mọi doanh nghiệp và người buôn bán. Dữ liệu này, nếu không được phân loại và xử lý một cách hiệu quả, có thể gây ra nhiều khó khăn và thách thức đối với cả người tìm kiếm thông tin và người cung cấp sản phẩm hoặc dịch vụ.



IV. PHƯƠNG PHÁP VÀ MÔ HÌNH



Phương pháp

Apache Spark

- Sử dụng Spark để kết nối với API của Facebook hoặc các nguồn dữ liệu khác để thu thập thông tin

- Spark có khả năng xử lý dữ liệu theo thời gian thực, cho phép bạn theo dõi sự quan tâm của người dùng ngay lập tức và tổ chức dữ liệu một cách hiệu quả.

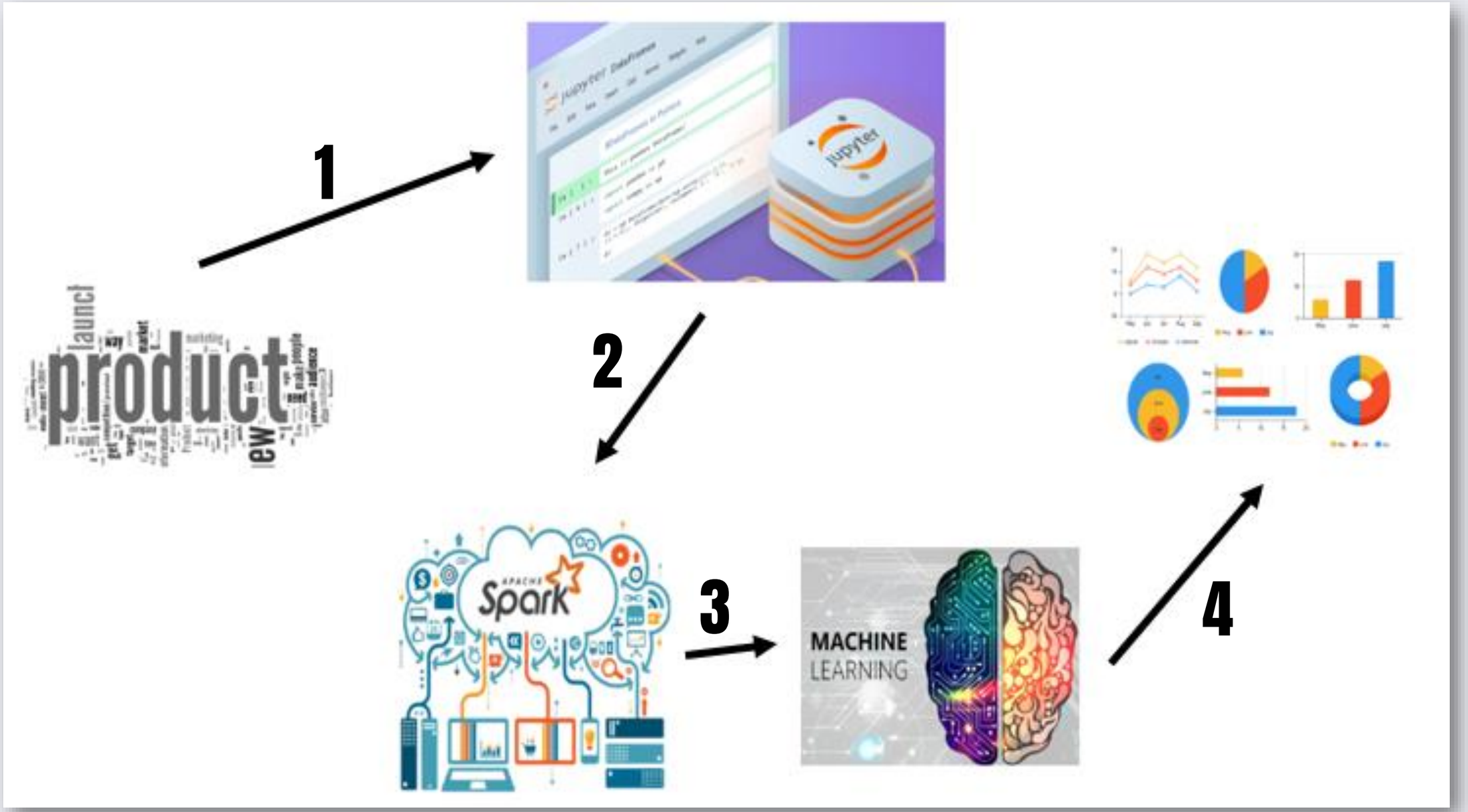
Phương pháp

Machine Learning

Dùng phương pháp LMlib Regression (hồi quy tuyến tính) vào bài toán

- ***Phương thức làm việc:*** sử dụng hàm tuyến tính để xác định mối quan hệ giữa biến đầu vào và đầu ra.
- ***Loại mô hình:*** Mô hình này giả định mối quan hệ tuyến tính giữa biến đầu vào và đầu ra.
- ***Phân loại và hồi quy:*** thích hợp cho nhiệm vụ hồi quy (dự đoán giá trị số liên tục).
- ***Tính khả thi với dữ liệu phi tuyến tính:*** hiệu quả khi mối quan hệ giữa biến đầu vào và đầu ra là tuyến tính. Khó khăn khi xử lý dữ liệu phi tuyến tính.
- ***Xử lý đặc trưng và nhiễu:*** đòi hỏi sự lựa chọn đặc trưng và xử lý nhiễu để đảm bảo tính ổn định của mô hình.

MÔ HÌNH



I. ƯU ĐIỂM VÀ NHƯỢC ĐIỂM CỦA MÔ HÌNH

1. ƯU ĐIỂM:

Hiệu quả cao: Apache Spark là một framework xử lý dữ liệu lớn, cho phép thu thập và xử lý dữ liệu từ Facebook một cách hiệu quả.

Tùy chỉnh nội dung: Dựa trên dữ liệu về sự quan tâm của người dùng, bạn có thể tạo nội dung tùy chỉnh và quảng cáo phù hợp với từng người dùng.

Chi phí không cao: Là hình thức Marketing hoàn toàn miễn phí, hiệu quả cao, chi phí thấp.

I. ƯU ĐIỂM VÀ NHƯỢC ĐIỂM CỦA MÔ HÌNH

2. NHƯỢC ĐIỂM:

Độ chính xác: Có thể có những sai sót trong quá trình thu thập và xử lý dữ liệu.

Phụ thuộc vào API của Facebook: Việc thu thập dữ liệu phụ thuộc vào API của Facebook, nếu có bất kỳ thay đổi nào từ phía Facebook có thể ảnh hưởng đến việc thu thập dữ liệu.

I. ƯU ĐIỂM VÀ NHƯỢC ĐIỂM CỦA MÔ HÌNH

2. NHƯỢC ĐIỂM:

Độ chính xác: Có thể có những sai sót trong quá trình thu thập và xử lý dữ liệu.

Phụ thuộc vào API của Facebook: Việc thu thập dữ liệu phụ thuộc vào API của Facebook, nếu có bất kỳ thay đổi nào từ phía Facebook có thể ảnh hưởng đến việc thu thập dữ liệu.

VI. PHÁT TRIỂN

THỨ NHẤT

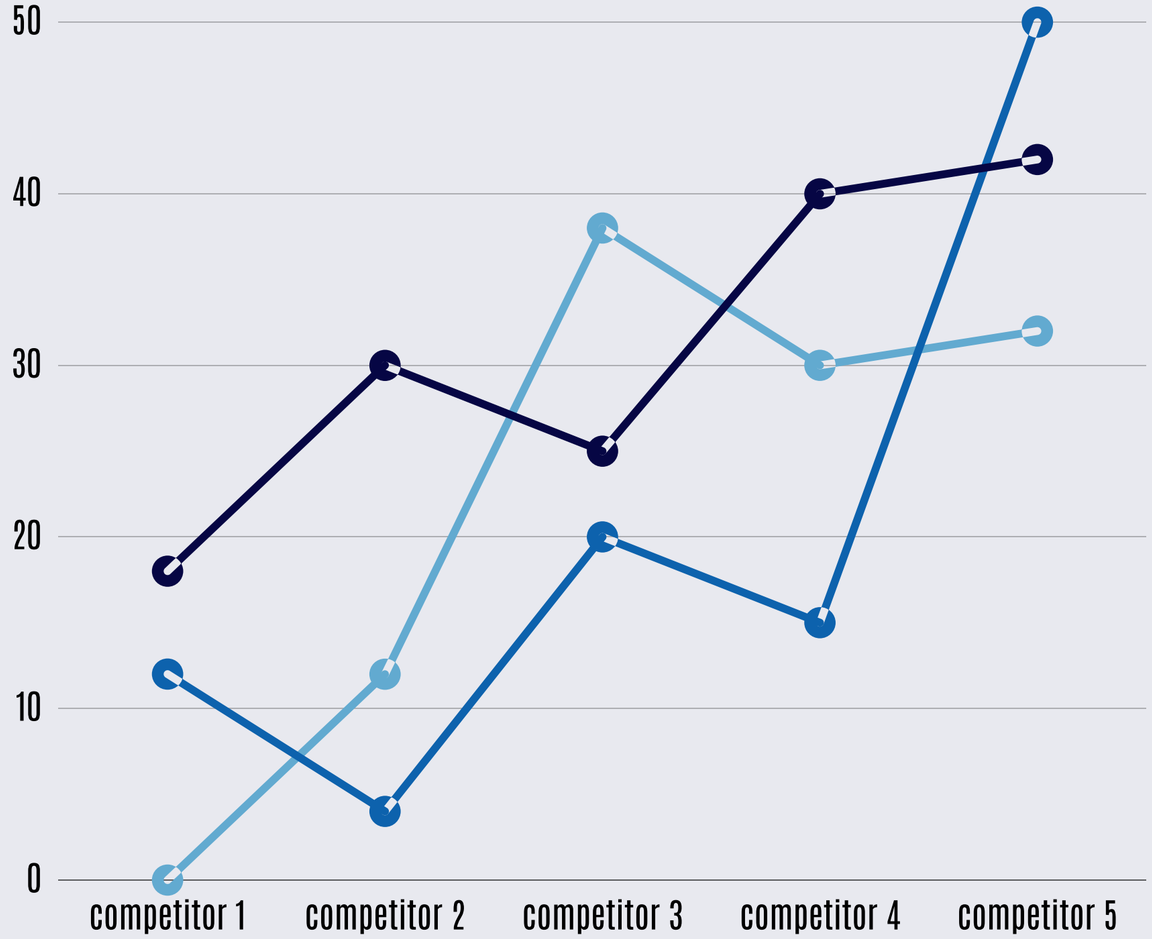
- Cải tiến công nghệ: Cải tiến công nghệ để tăng cường khả năng thu thập và xử lý dữ liệu.

THỨ HAI

- Phân tích sâu hơn: Sử dụng các công cụ phân tích sâu hơn để hiểu rõ hơn về sự quan tâm của người dùng.

THỨ 3

- Tối ưu hóa nội dung: Dựa trên thông tin thu được, tối ưu hóa nội dung để phù hợp với từng người dùng.



BIG DATA

THANK YOU!

FOR YOUR ATTENTION