­­­ĐẠI HỌC BÁCH KHOA HÀ NỘI

TRƯỜNG CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

──────── \* ───────

Ảnh có chứa văn bản, áp phích, Phông chữ, Đồ họa

Mô tả được tạo tự động

**BÁO CÁO**

HỌC PHẦN: LƯU TRỮ VÀ XỬ LÝ DỮ LIỆU LỚN

*(Mã học phần: IT4931)*

***ĐỀ TÀI:*** Lưu trữ và xử lý dữ liệu giá cổ phiếu tập đoàn Apple

GVHD: **TS. Trần Việt Trung**

Mã lớp: 144943

Nhóm: 21

Danh sách sinh viên thực hiện:

|  |  |  |
| --- | --- | --- |
| STT | Họ và tên | MSSV |
| 1 | Đào Tường Vinh | 20204705 |
| 2 | Nguyễn Tuấn Anh | 20204808 |
| 3 | Trương Văn Hiển | 20194276 |
| 4 | Bùi Trần Phương | 20204848 |
| 5 | Nguyễn Như Phước | 20204597 |

***Hà Nội, tháng 12 năm 2023***

Nội dung

[I. Giới thiệu đề tài 3](#_Toc154659729)

[II. Công nghệ sử dụng 4](#_Toc154659730)

[1. Hadoop 4](#_Toc154659731)

[2. Apache Spark 5](#_Toc154659732)

[3. Elasticsearch và Kibana 6](#_Toc154659733)

[III. Các bước tiến hành 8](#_Toc154659734)

[IV. Đánh giá và hướng phát triển 23](#_Toc154659735)

[1. Đánh giá 23](#_Toc154659736)

[2. Hướng phát triển 23](#_Toc154659737)

# Giới thiệu đề tài

Trước đây, khi mạng Internet còn chưa phát triển, lượng dữ liệu con người sinh ra khá nhỏ giọt và thưa thớt, nhìn chung, lượng dữ liệu này vẫn nằm trong khả năng xử lý của con người dù bằng tay hay bằng máy tính. Tuy nhiên trong kỷ nguyên số, khi mà sự bùng nổ công nghệ truyền thông đã dẫn tới sự bùng nổ dữ liệu người dùng, lượng dữ liệu được tạo ra vô cùng lớn và đa dạng, đòi hỏi một hệ thống đủ mạnh để phân tích và xử lý những dữ liệu đó.

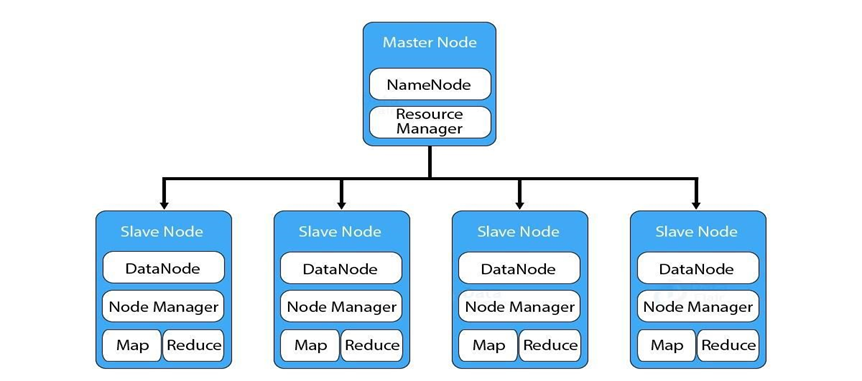
Khái niệm Big Data đề cập tới dữ liệu lớn theo 3 khía canh khác nhau, thứ nhất là tốc độ sinh dữ liệu (velocity), thứ hai là lượng dữ liệu (volume) và thứ ba là độ đa dạng (variety). Lượng dữ liệu này có thể đến từ nhiều nguồn khác nhau như các nền tảng truyền thông Google, Facebook, Twitter, … hay thông số thu thập từ các cảm biến, thiết bị IoT trong đời sống, … Và một sự thật rằng doanh nghiệp nào có thể kiểm soát và tạo ra tri thức từ những dữ liệu này sẽ tạo ra một tiềm lực rất lớn để cạnh tranh với những doanh nghiệp khác. Có thể nói rằng dữ liệu là sức mạnh của kỷ nguyên số cũng không hề ngoa một chút nào.

Để tiếp cận với lĩnh vực này, nhóm chúng em quyết định chọn một loại dữ liệu đủ lớn trong khả năng để tiến hành phân tích và lưu trữ. Dữ liệu của nhóm em là thông tin về giá cổ phiếu hàng ngày của Apple (được lấy trực tiếp từ trang Yahoo Finance) từ 15/12/1980 đến 30/10/2023, bao gồm các trường như Date (ngày giao dịch), Open (giá mở cửa), Close(giá đóng cửa), High(giá cao nhất), Low(giá thấp nhất), Volume (tổng số cổ phiếu được giao dịch)

# Công nghệ sử dụng

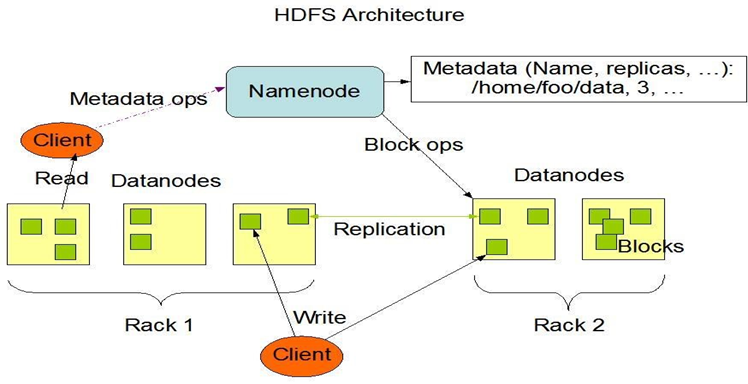
## Hadoop

Hadoop Cluster là hệ thống file phân tán, cung cấp khả năng lưu trữ dữ liệu khổng lồ và tính năng tối ưu hoá việc sử dụng băng thông giữa các node.



Hadoop được cài đặt trên các máy tính trong hệ thống phân tán theo kiến trúc master – slave. Hadoop có thể hoạt động trên một máy (giống như 1 team chỉ có 1 member) hoặc mở rộng tới hàng ngàn máy, với mỗi máy đều có thể sử dụng để lưu trữ hoặc tính toán dữ liệu. Khi lưu trữ trên Hadoop, file dữ liệu được chia thành các chunk và được lưu thành nhiều bản sao, giúp cho cụm Hadoop có khả năng chịu lỗi.

HDFS là nơi lưu dữ liệu của Hadoop, HDFS chia nhỏ dữ liệu thành các đơn vị dữ liệu nhỏ hơn gọi là các blocks và lưu trữ chúng phân tán trong các node của cụm Hadoop. HDFS sử dụng kiến trúc master-slave, trong đó master gồm một Name Node để quản lý hệ thống file metadata v và một hay nhiều slave Data Nodes để lưu trữ dữ liệu.



Khi lượng dữ liệu tăng lên, kiến trúc này có thể mở rộng thêm bằng cách bổ sung các Datanode để tăng cường dung lượng lưu trữ của hệ thống.

## Apache Spark

Apache Spark là một framework xử lý dữ liệu mã nguồn mở trên quy mô lớn. Spark cung cấp một giao diện để lập trình các cụm tính toán song song với khả năng chịu lỗi.

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 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.

Những điểm nổi bật của Spark:

- Xử lý dữ liệu: Spark xử lý dữ liệu theo lô và theo thời gian thực.

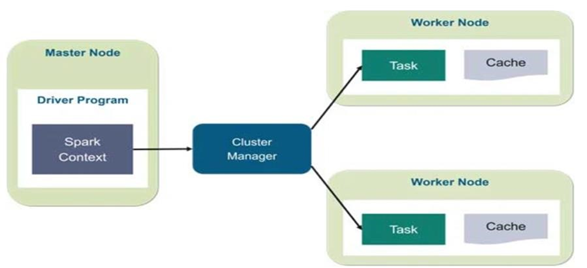
- Tính tương thích: Có thể tích hợp với tất cả nguồn dữ liệu và định dạng tệp được hỗ trợ bởi cụm Hadoop.

- Hỗ trợ ngôn ngữ: Java, Python, Scala, R.

- Phân tích thời gian thực.

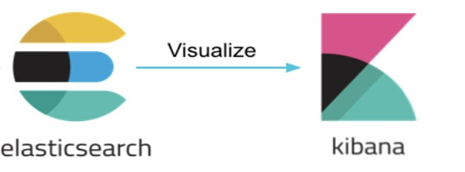
Kiến trúc của Spark bao gồm hai thành phần chính: trình điều khiển (driver) và trình thực thi (executors). Trình điều khiển dùng để chuyển đổi mã của người dùng thành nhiều tác vụ (tasks) có thể được phân phối trên các nút xử lý (worker nodes). Khi thực thi, trình điều khiển Driver tạo ra 1 SparkContext, sau đó giao tiếp với Cluster Manager để tính toán tài nguyên và phân chia các tác vụ đến cho các worker nodes.

Apache Spark xây dựng các lệnh xử lý dữ liệu của người dùng thành Đồ thị vòng có hướng hoặc DAG. DAG là lớp lập lịch của Apache Spark; nó xác định những tác vụ nào được thực thi trên những nút nào và theo trình tự nào.



## Elasticsearch và Kibana

Dữ liệu sau khi được làm sạch bởi Spark cần được biểu diễn dưới dạng bảng biểu, đồ thị để mang đến cho người dùng góc nhìn trực quan nhất. Elasticsearch và Kibana là những ứng dụng phù hợp để đảm nhận vai trò này. Là một công cụ tìm kiếm (với tốc độ gần thời gian thực) và phân tích dữ liệu phân tán, Elasticsearch có thể lưu trữ và phân tích nhiều loại dữ liệu khác nhau như: dữ liệu có cấu trúc, giữ liệu phi cấu trúc, giữ liệu số, dữ liệu về không gian địa lý, đánh chỉ mục dữ liệu một cách hiệu quả nhằm hỗ trợ quá trình tìm kiếm được thực hiện nhanh chóng. Các truy vấn trên Elasticsearch được thực hiện thông qua API, curl, python, hoặc qua Kibana. Kibana cung cấp giao diện đồ hoạ để người dùng dễ dàng hơn trong việc khai phá, biểu diễn trực quan dữ liệu được lưu trên Elasticsearch.



# Các bước tiến hành

***Luồng xử lý chính:***

**Ảnh có chứa ảnh chụp màn hình, biểu đồ, văn bản, Phông chữ

Mô tả được tạo tự động**

* Dữ liệu giá cổ phiếu theo ngày (Từ ngày 15/12/1980 đến ngày 30/10/2023) của tập đoàn Apple được lấy trực tiếp từ trang Yahoo Finance dưới dạng file csv gồm các trường:

+ Date: Ngày giao dịch

+ Open: Giá mở cửa trong ngày

+ High: Giá cao nhất trong ngày

+ Low: Giá thấp nhất trong ngày

+ Close: Giá đóng cửa trong ngày

+ Volume: Tổng số cổ phiếu được giao dịch

* Sau đó lưu dữ liệu vào HDFS.
* Spark lấy dữ liệu từ HDFS để xử lý và trực quan hóa dữ liệu.
* Sau khi xử lý xong, dữ liệu đó được lưu vào Elasticsearch để tìm kiếm.
* Kibana lấy dữ liệu từ Elasticsearch để phân tích.

***Các bước tiến hành:***

* ***Khởi động hệ thống HDFS***

cd hadoop-3.3.4/sbin

./start-all.sh

./stop-all.sh

***Ảnh có chứa văn bản, ảnh chụp màn hình, phần mềm, Phần mềm đa phương tiện

Mô tả được tạo tự động***

***Ảnh có chứa ảnh chụp màn hình, văn bản, phần mềm, Biểu tượng máy tính

Mô tả được tạo tự động***

* ***Up file data lên HDFS để lưu trữ***

***Ảnh có chứa văn bản, phần mềm, Biểu tượng máy tính, số

Mô tả được tạo tự động***

* ***Spark lấy dữ liệu từ HDFS để xử lý***

***Ảnh có chứa văn bản, ảnh chụp màn hình, Phông chữ

Mô tả được tạo tự động***

***Ảnh có chứa văn bản, ảnh chụp màn hình, Phông chữ

Mô tả được tạo tự động***

* ***Trực quan hóa dữ liệu với Spark***

***Đồ thị biểu diễn giá Open, High, Low, Close theo ngày***

***Ảnh có chứa văn bản, ảnh chụp màn hình, hàng, Sơ đồ

Mô tả được tạo tự động***

***Đồ thị biểu diễn tổng số cổ phiếu giao dịch Volume theo ngày***

***Ảnh có chứa Sơ đồ, hàng, ảnh chụp màn hình

Mô tả được tạo tự động***

***Đồ thị biểu diễn tỉ lệ thay đổi giữa giá Open và Close theo ngày***

***Ảnh có chứa Sơ đồ, hàng, ảnh chụp màn hình

Mô tả được tạo tự động***

***Đồ thị biểu diễn min của Open, High, Low, Close theo năm***

***Ảnh có chứa văn bản, ảnh chụp màn hình, Sơ đồ, hàng

Mô tả được tạo tự động***

***Đồ thị biểu diễn max của Open, High, Low, Close theo năm***

***Ảnh có chứa văn bản, ảnh chụp màn hình, Sơ đồ, hàng

Mô tả được tạo tự động***

***Đồ thị biểu diễn mean của Open, High, Low, Close theo năm***

***Ảnh có chứa hàng, Sơ đồ, văn bản, ảnh chụp màn hình

Mô tả được tạo tự động***

***Đồ thị biểu diễn min của Volume theo năm***

***Ảnh có chứa ảnh chụp màn hình, Sơ đồ, văn bản, hàng

Mô tả được tạo tự động***

***Đồ thị biểu diễn max của Volume theo năm***

***Ảnh có chứa văn bản, ảnh chụp màn hình, Sơ đồ, hàng

Mô tả được tạo tự động***

***Đồ thị biểu diễn mean của Volume theo năm***

***Ảnh có chứa ảnh chụp màn hình, văn bản, Sơ đồ, hàng

Mô tả được tạo tự động***

***Đồ thị biểu diễn min của Change theo năm***

***Ảnh có chứa hàng, ảnh chụp màn hình, Sơ đồ, Song song

Mô tả được tạo tự động***

***Đồ thị biểu diễn max của Change theo năm***

***Ảnh có chứa ảnh chụp màn hình, Sơ đồ, hàng, văn bản

Mô tả được tạo tự động***

***Đồ thị biểu diễn mean của Change theo năm***

***Ảnh có chứa văn bản, Sơ đồ, hàng, Phông chữ

Mô tả được tạo tự động***

* ***Khởi động Elasticsearch***

Sudo systemctl start elasticsearch

Sudo systemctl status elasticsearch

Sudo systemctl stop elasticsearch

***Ảnh có chứa văn bản, phần mềm, Phần mềm đa phương tiện, Website

Mô tả được tạo tự động***

***Ảnh có chứa văn bản, ảnh chụp màn hình, phần mềm, Biểu tượng máy tính

Mô tả được tạo tự động***

* ***Lưu dữ liệu sau khi xử lý vào Elasticsearch để tìm kiếm***

***Ảnh có chứa văn bản, ảnh chụp màn hình, phần mềm, Phần mềm đa phương tiện

Mô tả được tạo tự động***

***Ảnh có chứa văn bản, ảnh chụp màn hình, phần mềm, Phần mềm đa phương tiện

Mô tả được tạo tự động***

* ***Tìm kiếm với Elasticsearch***

***Tìm kiếm theo ngày***

***Ảnh có chứa văn bản, ảnh chụp màn hình, phần mềm, Phần mềm đa phương tiện

Mô tả được tạo tự động***

***Tìm kiếm top 10 ngày có tỉ lệ tăng nhiều nhất***

***Ảnh có chứa văn bản, ảnh chụp màn hình, phần mềm, Phần mềm đa phương tiện

Mô tả được tạo tự động***

***Tìm kiếm top 10 ngày có tỉ lệ giảm nhiều nhất***

***Ảnh có chứa văn bản, ảnh chụp màn hình, Phông chữ

Mô tả được tạo tự động***

***Tìm kiếm top 10 ngày giao dịch nhiều nhất***

***Ảnh có chứa văn bản, Phần mềm đa phương tiện, phần mềm, ảnh chụp màn hình

Mô tả được tạo tự động***

***Tìm kiếm top 10 ngày giao dịch ít nhất***

***Ảnh có chứa văn bản, ảnh chụp màn hình

Mô tả được tạo tự động***

* ***Khởi động Kibana***

Sudo systemctl start kibana

Sudo systemctl status kibana

Sudo systemctl stop kibana

Ảnh có chứa văn bản, ảnh chụp màn hình, phần mềm, Phông chữ

Mô tả được tạo tự động

Ảnh có chứa văn bản, ảnh chụp màn hình, phần mềm, Trang web

Mô tả được tạo tự động

* ***Trực quan hóa dữ liệu với Kibana***

***Ảnh có chứa văn bản, phần mềm, Biểu tượng máy tính, Hệ điều hành

Mô tả được tạo tự động***

***Đồ thị biểu diễn min của Open, High, Low, Close theo năm***

***Ảnh có chứa văn bản, ảnh chụp màn hình, phần mềm, Biểu tượng máy tính

Mô tả được tạo tự động***

***Đồ thị biểu diễn max của Open, High, Low, Close theo năm***

***Ảnh có chứa ảnh chụp màn hình, văn bản, phần mềm

Mô tả được tạo tự động***

***Đồ thị biểu diễn mean của Open, High, Low, Close theo năm***

***Ảnh có chứa văn bản, phần mềm, ảnh chụp màn hình, Biểu tượng máy tính

Mô tả được tạo tự động***

***Đồ thị biểu diễn min của Volume theo năm***

***Ảnh có chứa văn bản, phần mềm, biểu đồ, Phần mềm đồ họa

Mô tả được tạo tự động***

***Đồ thị biểu diễn max của Volume theo năm***

***Ảnh có chứa văn bản, phần mềm, biểu đồ, Sơ đồ

Mô tả được tạo tự động***

***Đồ thị biểu diễn mean của Volume theo năm***

***Ảnh có chứa ảnh chụp màn hình, văn bản, biểu đồ, phần mềm

Mô tả được tạo tự động***

***Đồ thị biểu diễn min của Change theo năm***

***Ảnh có chứa văn bản, ảnh chụp màn hình, biểu đồ, thiết kế

Mô tả được tạo tự động***

***Đồ thị biểu diễn max của Change theo năm***

***Ảnh có chứa ảnh chụp màn hình, văn bản, phần mềm, biểu đồ

Mô tả được tạo tự động***

***Đồ thị biểu diễn mean của Change theo năm***

***Ảnh có chứa văn bản, ảnh chụp màn hình, phần mềm, biểu đồ

Mô tả được tạo tự động***

***Đồ thị biểu diễn sum của Volume theo năm***

***Ảnh có chứa ảnh chụp màn hình, văn bản, biểu đồ, thiết kế

Mô tả được tạo tự động***

# Đánh giá và hướng phát triển

## Đánh giá

* Sau quá trình tìm hiểu, nghiên cứu và thực hiện đồ án, nhóm em đã hoàn thành được các mục tiêu đề ra bao gồm việc tìm hiểu các kiến thức về công nghệ được học và áp dụng vào xây dựng hệ thống lưu trữ, xử lý và tra cứu dữ liệu chứng khoán.
* Đồ án đã tạo điều kiện cho nhóm tìm hiểu và thực hành các công nghệ được sử dụng hiện nay, mang tính thực tiễn cao, thúc đẩy bản thân chúng em không ngừng tìm kiếm, học hỏi thêm nhiều kiến thức mới. Đó là những nền tảng vững chắc để em có thể phát triển bản thân trên con đường sự nghiệp sau này.
* Bên cạnh đó chúng em cũng gặp không ít khó khăn trong việc triển khai hệ thống, đặc biệt trong việc triển khai cụm Hadoop nhiều node. Vì thời gian học phần có hạn nên trong đồ án lần này chúng em chỉ sử dụng cụm Hadoop với 1 node duy nhất.

## Hướng phát triển

* Triển khai cụm Hadoop nhiều node
* Tiếp tục tìm hiểu, trau dồi kiến thức về các công nghệ mới để phát triển, mở rộng thêm trong đồ án này.