|  |  |
| --- | --- |
| **bỘ GIÁO DỤC VÀ ĐÀO TẠO** | **BỘ NÔNG NGHIỆP VÀ PTNT** |

**TRƯỜNG ĐẠI HỌC THỦY LỢI**



**BÁO CÁO BÀI TẬP LỚN**

**HỌC PHẦN PHÂN TÍCH DỮ LIỆU LỚN**

ĐỀ TÀI: Sử dụng thuật toán K-Mean kết hợp Hadoop

cho bài toán phân cụm chất lượng sữa.

Giáo viên hướng dẫn: TS. Tạ Quang Chiểu

Nhóm sinh viên thực hiện: Nhóm 04

1. Nguyễn Thị Mây, lớp 62TH1

2. Nguyễn Quang Minh, lớp 62TH1

3. Nguyễn Hồng Thương, lớp 62TH1

Hà Nội, ngày 19 tháng 8 năm 2023

**MỤC LỤC**

[PHÂN CÔNG NHIỆM VỤ 3](#_Toc143456731)

[DANH MỤC CÁC HÌNH ẢNH 4](#_Toc143456732)

[Chương 1: Giới thiệu bài toán, giới thiệu tổng quan 5](#_Toc143456733)

[Chương 2: Trình bày lý thuyết 5](#_Toc143456734)

[Chương 3: Kết quả và đánh giá mô hình 5](#_Toc143456735)

[Kết luận 5](#_Toc143456736)

[CHƯƠNG 1 GIỚI THIỆU BÀI TOÁN 1](#_Toc143456737)

[1.1 Mô tả bài toán 1](#_Toc143456738)

[1.2 Phần nội dung. 1](#_Toc143456739)

[CHƯƠNG 2 TRÌNH BÀY LÝ THUYẾT 3](#_Toc143456740)

[2.1 Giới thiệu về học máy 3](#_Toc143456741)

[1.1.1 Lịch sử và vai trò của machine learning 3](#_Toc143456742)

[1.1.2 Các phương pháp của machine learning 3](#_Toc143456743)

[1.2 Thuật toán K-mean 4](#_Toc143456744)

[1.2.1 Cách thực hiện 4](#_Toc143456745)

[1.3 Giới thiệu về Bigdata 5](#_Toc143456746)

[1.3.1 Định nghĩa 5](#_Toc143456747)

[1.3.2 Đặc trưng. 6](#_Toc143456748)

[1.4 Giới thiệu về Hadoop 6](#_Toc143456749)

[1.4.1 Hadoop là gì? 6](#_Toc143456750)

[1.4.2 Các thành phần của Hadoop 8](#_Toc143456751)

[1.4.3 HDFS là gì? 8](#_Toc143456752)

[1.4.4 MapReduce là gì? 11](#_Toc143456753)

[CHƯƠNG 2 KẾT QUẢ VÀ ĐÁNH GIÁ MÔ HÌNH 11](#_Toc143456754)

[2.1 Phân tích chi tiết bài toán 11](#_Toc143456755)

[2.2 Xây dựng lưu đồ 14](#_Toc143456756)

[2.3 Kết quả. 15](#_Toc143456757)

[2.4 Đánh giá mô hình 16](#_Toc143456758)

[2.5 Sử dụng mô hình học được để phân cụm 1 mẫu dữ liệu sữa. 17](#_Toc143456759)

[KẾT LUẬN 18](#_Toc143456760)

# PHÂN CÔNG NHIỆM VỤ

|  |  |
| --- | --- |
| **Sinh viên** | **Công việc** |
| Nguyễn Thị Mây – 2051063448 | * Tìm hiểu code thuật toán k-mean mapreduce hóa. * Thiết kế giao diện phân cụm trên 1 mẫu dữ liệu sữa. * Code đánh giá mô hình sử dụng python. |
| Nguyễn Quang Minh - 2051060616 | * Tìm hiểu nội dung, thuyết trình, làm PowerPoint. |
| Nguyễn Hồng Thương - 2051063623 | * Tìm hiểu nội dung, viết báo cáo. |

# DANH MỤC CÁC HÌNH ẢNH

[Hình 2.3.1 Minh hoạ nguồn gốc của dữ liệu 5](#_heading=h.3j2qqm3)

[Hình 2.4.1 Biểu tượng của Hadoop 6](#_heading=h.3whwml4)

[Hình 2-1.2 Thành phần của Hadoop 7](#_heading=h.qsh70q)

[Hình 2.4.3.1 Kiến trúc HDFS 8](#_heading=h.1pxezwc)

[Hình 2.4.3.2 Quá trình đọc file 8](#_heading=h.49x2ik5)

[Hình 2.4.3.3 Quá trình tạo và ghi dữ liệu lên file HSFS 9](#_heading=h.2p2csry)

[Hình 2.4.4.1 Mô hình thực thi MapReduce 10](#_heading=h.3o7alnk)

[Hình 3.2.1 Lưu đồ MapReduce của thuật toán K-Mean 12](#_heading=h.1hmsyys)

[Hình 3.3.1 kết quả tọa độ k tâm cụm 13](#_heading=h.2grqrue)

[Hình 3.3.2 Tập dữ liệu đầu đã được phân cụm 13](#_heading=h.vx1227)

[Hình 3.4.1 Công thức silhouette 14](#_heading=h.1v1yuxt)

[Hình 3.4.2 Thông số đánh giá 14](#_heading=h.4f1mdlm)

[Hình 3.5.1 Chương trình chạy phân cụm 15](#_heading=h.19c6y18)

**CẤU TRÚC CỦA BÀI TẬP LỚN**

## Chương 1: Giới thiệu bài toán, giới thiệu tổng quan

## Chương 2: Trình bày lý thuyết

* Lịch sử và vai trò của học máy
* Thuật toán K-Mean
* Giới thiệu về Big Data
* Giới thiệu về Hadoop

## Chương 3: Kết quả và đánh giá mô hình

* Phân tích chi tiết bài toán
* Xây dựng lưu đồ
* Kết quả thực thi
* Đánh giá mô hình

## Kết luận

# GIỚI THIỆU BÀI TOÁN

Công nghệ Big data đã đạt đến đỉnh cao trong việc thực hiện các chức năng của nó. Trong những năm gần đây Big data đã vượt ra khỏi bảng xếp hạng những công nghệ mới nổi Cycle Hype của Gartner và tạo ra một tiếng vang lớn cho xu hướng công nghệ của thế giới. Big data chứa trong mình rất nhiều thông tin quý giá mà nếu trích xuất thành công, nó sẽ giúp rất nhiều trong nhiều lĩnh vực như y tế, giao thông, giáo dục, …

Chính vì thế những framework giúp việc xử lý BIGDATA cũng đang ngày càng được xử lý và phát triển mạnh. Một trong những công nghệ cốt lõi cho việc lưu trữ và truy cập số lượng lớn dữ liệu là Hadoop - một framework giúp lưu trữ và xử lý Big data áp dụng MapReduce.

Từ đó, em đã chọn đề tài: "Sử dụng thuật toán k-mean kết hợp Hadoop để phân cụm chất lượng sữa" để làm báo kết thúc môn học.

## Mô tả bài toán

* Tên bài toán: Sử dụng thuật toán K-mean kết hợp Hadoop để phân cụm chất lượng sữa.
* Mục đích của bài toán: Phân cụm chất lượng sữa dựa vào các thuộc tính: pH, Nhiệt độ, Hương vị, mùi, chất béo, Độ trong, Màu sữa.
* Input: Giá trị các thuộc tính của sữa (pH, Nhiệt độ, Hương vị, mùi, chất béo, Độ trong của sữa, Màu sữa).
* Output: k tâm cụm

## Phần nội dung.

* Sử dụng tập dữ liệu data.txt để xây dựng mô hình và phân cụm bằng thuật toán k-mean.
* Sử dụng hadoop chia tập dữ liệu thành nhiều tập con và tính toán song song.
* Sử dụng thư viện sklearn với độ silhouette\_score để đánh giá mô hình.
* Thiết kế giao diện người dùng: phân cụm 1 mẫu sữa khi đã biết các thông tin thuộc tính (pH, Nhiệt độ, Hương vị, mùi, chất béo, Độ trong của sữa, Màu sữa).

# TRÌNH BÀY LÝ THUYẾT

## Giới thiệu về học máy

### Lịch sử và vai trò của machine learning

* Machine learning là 1 ứng dụng của trí tuệ nhân tạo cung cấp cho các hệ thống khả năng tự học dựa trên dữ liệu mà không cần lập trình rõ ràng, máy có thể tự dự đoán hoặc đưa ra quyết định mà không cần lập trình.
* Các ứng dụng phổ biến của machine learning:
  + Tự động phân loại
  + Ứng dụng trong các mạng xã hội.
  + Nhận diện hình ảnh.
  + Chăm sóc sức khỏe.
  + Tài chính.

### Các phương pháp của machine learning

* Machine learning được chia thành ba loại chính đó là học có giám sát, học không giám sát, học tăng cường

#### Học có giám sát

* Ưu điểm
  + Cho phép thu thập dữ liệu và tạo ra dữ liệu đầu ra từ những kinh nghiệm trước đó
  + Thuật toán học có giám sát giúp dự đoán chính xác hơn và cung cấp kết quả dự đoán đáng tin cậy hơn.
  + Nó cũng cho phép tìm hiểu các mối quan hệ giữa các đặc trưng và đầu ra
* Nhược điểm
  + Phụ thuộc vào chất lượng và độ phong phú của dữ liệu huấn luyện
  + Nếu dữ liệu không đại diện cho dữ liệu thực tế hoặc không đủ để đào tạo mô hình, kết quả dự đoán có thể không chính xác.

#### Học không giám sát

* Ưu điểm
  + Học không giám sát cho phép tìm kiếm các cấu trúc và mối quan hệ ẩn trong dữ liệu mà không yêu cầu nhãn.
  + Điều này cho phép phát hiện các mô hình phức tạp hơn và phát triển các giải pháp đa dạng.
* Nhược điểm
  + Một nhược điểm của học không giám sát là không có phương pháp đánh giá hiệu quả như học có giám sát.
  + Nó cũng có thể tạo ra kết quả dự đoán không chính xác khi dữ liệu không đủ để đào tạo mô hình hoặc không đại diện cho dữ liệu thực tế.

#### Học tăng cường

* Ưu điểm
  + Học tăng cường có thể giải quyết các vấn đề phức tạp và đưa ra các quyết định tốt hơn bằng cách tối đa hóa lợi ích dài hạn.
  + Nó cũng cho phép mô hình tương tác với môi trường để đưa ra quyết định tối ưu.
* Nhược điểm
  + Một nhược điểm của học tăng cường là nó có thể tạo ra các quyết định không chính xác khi mô hình chưa đủ thông tin để đưa ra quyết định tối ưu.
  + Nó cũng khó huấn luyện và tốn nhiều thời gian để điều chỉnh các thông số mô hình.

## Thuật toán K-mean

* Thuật toán K-Mean là thuật toán học không giám sát.
* Input: tập dữ liệu không có nhãn và số cụm cần tìm k.
* Output: Các cụm Ci(i=1…k) sao cho hàm tiêu chuẩn E đạt giái trị tối thiểu.

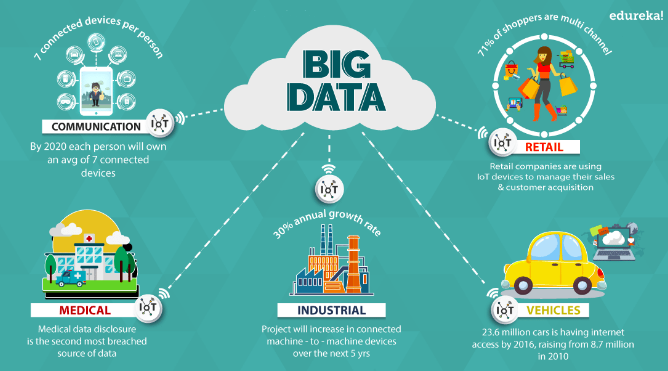
### Cách thực hiện

* Bước 1: Khởi tạo tâm cụm:
  + chọn k đối tượng làm trọng tâm ban đầu của k cụm (chọn ngẫu nhiên, theo kinh nghiệm).
* Bước 2: Tính toán khoảng cách và gán cụm:
  + với mỗi điểm tính khoảng cách đến tâm cụm sau đó tìm trọng tâm gần nhất với mỗi điểm (tức là khoảng cách của mỗi điểm tới các tâm cụm, khoảng cách nào ngắn nhất thì ta gán nó vào cụm đó).
* Bước 3: Tính toán và điều chỉnh tâm các cụm:
  + với mỗi cụm cập nhật trọng tâm bằng cách xác định trung bình cộng các vector đối tượng dữ liệu (tức là tính trung bình cộng tọa độ của các điểm thuộc cụm).
* Bước 4: Kiểm tra điều kiện dừng:
  + lặp lại các bước 2 và 3 cho đến khi tâm cụm không thay đổi.

## Giới thiệu về Bigdata

### Định nghĩa

* Theo wikipedia:Dữ liệu lớn là một thuật ngữ chỉ bộ dữ liệu lớn hoặc phức tạp mà các phương pháp truyền thống không đủ các ứng dụng để xử lý dữ liệu này.
* Theo Gartner : Dữ liệu lớn là những nguồn thông tin có đặc điểm chung khối lượng lớn, tốc độ nhanh và dữ liệu định dạng dưới nhiều hình thức khác nhau,do đó muốn khai thác được phải đòi hỏi phải có hình thức mới để đưa ra quyết định khám phá và tối ưu hóa quy trình.
* Dữ liệu đến từ rất nhiều nguồn khác nhau:



*Hình 2.3.1 Minh hoạ nguồn gốc của dữ liệu*

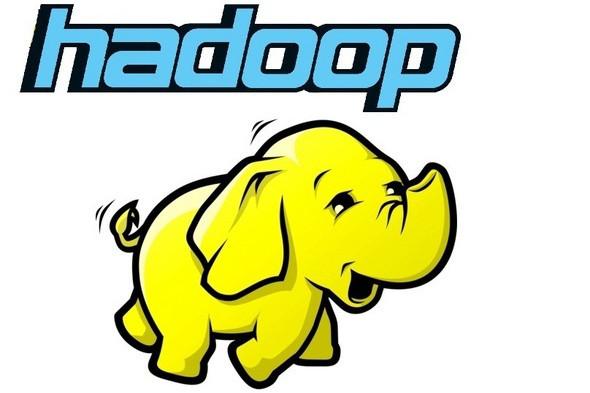
### Đặc trưng.

* *Khối lượng lớn (Volume)*: Khối lượng dữ liệu rất lớn và đang ngày càng tăng lên, tính đến 2014 thì có thể  trong khoảng vài trăm terabyte.
* *Tốc độ (Velocity)*: Khối lượng dữ liệu gia tăng rất nhanh.
* Đa dạng (Variety): Ngày nay hơn 80% dữ liệu được sinh ra là phi cấu trúc( tài liệu, blog, hình ảnh,...)
* Độ tin cậy/chính xác(Veracity): Bài toán phân tích và loại bỏ dữ liệu thiếu chính xác và nhiễu đang là tính chất quan trọng của bigdata.
* Giá trị(Value): Giá trị thông tin mang lại.

## Giới thiệu về Hadoop

### Hadoop là gì?

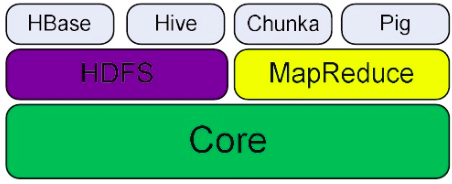
* Theo Apache Hadoop định nghĩa thì “Apache Hadoop là một framework dùng để chạy những ứng dụng trên 1 cluster lớn được xây dựng trên những phần cứng thông thường. Hadoop hiện thực mô hình Map/Reduce, đây là mô hình mà ứng dụng sẽ được chia nhỏ ra thành nhiều phân đoạn khác nhau, và các phần này sẽ được chạy song song trên nhiều node khác nhau. Thêm vào đó, Hadoop cung cấp 1 hệ thống file phân tán (HDFS) cho phép lưu trữ dữ liệu lên trên nhiều node. Cả Map/Reduce và HDFS đều được thiết kế sao cho framework sẽ tự động quản lý được các lỗi, các hư hỏng về phần cứng của các node.”
* Theo Wiki thì “Hadoop là một framework nguồn mở viết bằng Java cho phép phát triển các ứng dụng phân tán có cường độ dữ liệu lớn một cách miễn phí. Nó cho phép các ứng dụng có thể làm việc với hàng ngàn node khác nhau và hàng petabyte dữ liệu. Hadoop lấy được phát triển dựa trên ý tưởng từ các công bố của Google về mô hình MapReduce và hệ thống file phân tán Google File System (GFS).”
* Vậy ta có thể kết luận về Hadoop như sau:
* Hadoop là một framework cho phép phát triển các ứng dụng phân tán.
* Hadoop viết bằng Java.
* Hadoop cung cấp một phương tiện lưu trữ dữ liệu phân tán trên nhiều node, hỗ trợ tối ưu hoá lưu lượng mạng, đó là HDFS. HDSF che giấu tất cả các thành phần phân tán, các nhà phát triển ứng dụng phân tán sẽ chỉ nhìn thấy HDFS như một hệ thống file cục bộ bình thường.
* Hadoop giúp các nhà phát triển ứng dụng phân tán tập trung tối đa vào phần logic của ứng dụng, bỏ qua được một số phần chi tiết kỹ thuật phân tán bên dưới (phần này do Hadoop tự động quản lý).



*Hình 2.4.1 Biểu tượng của Hadoop*

### Các thành phần của Hadoop

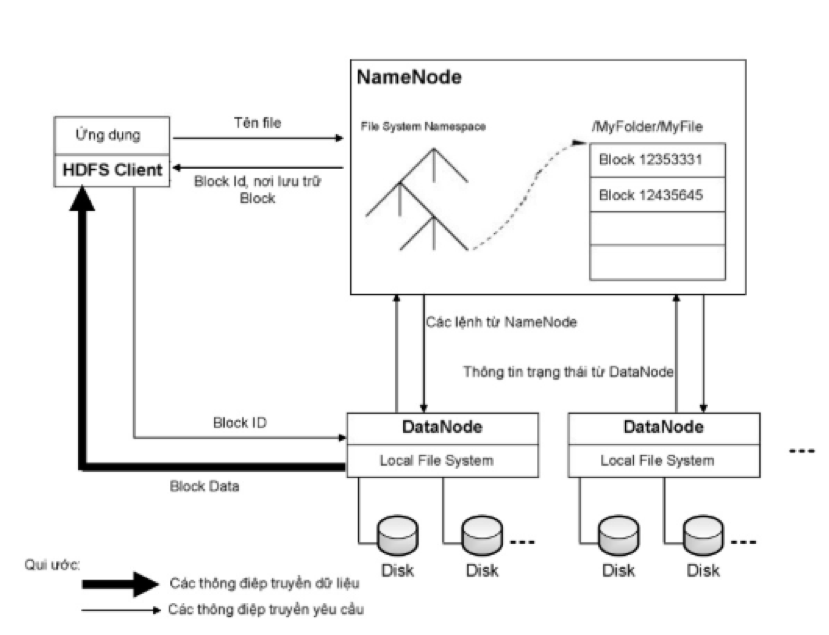
Hiện nay hadoop bao gồm các thành phần như: Core, MapReduce, HDFS, HBase, Hive, Pig, Chukwa,.. Tuy nhiên ta chỉ tập chung vào 2 thành phần quan trọng nhất: HDFS(lưu trữ) và MapReduce(xử lý).



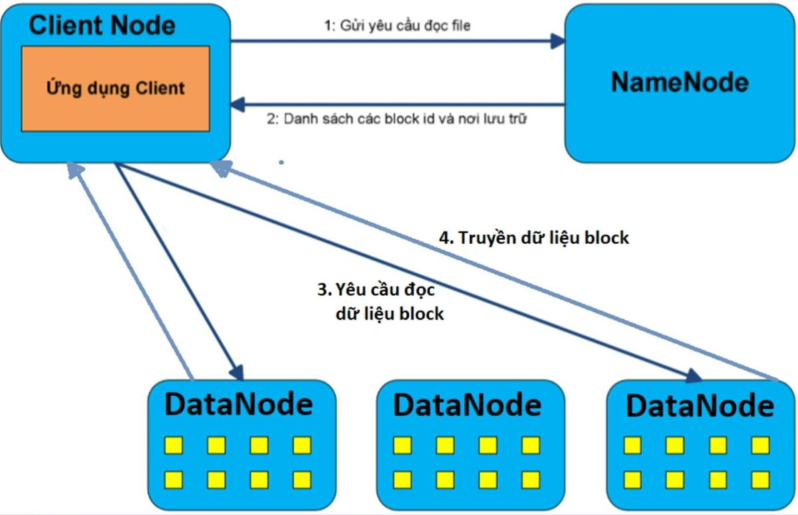
*Hình 2-1.2 Thành phần của Hadoop*

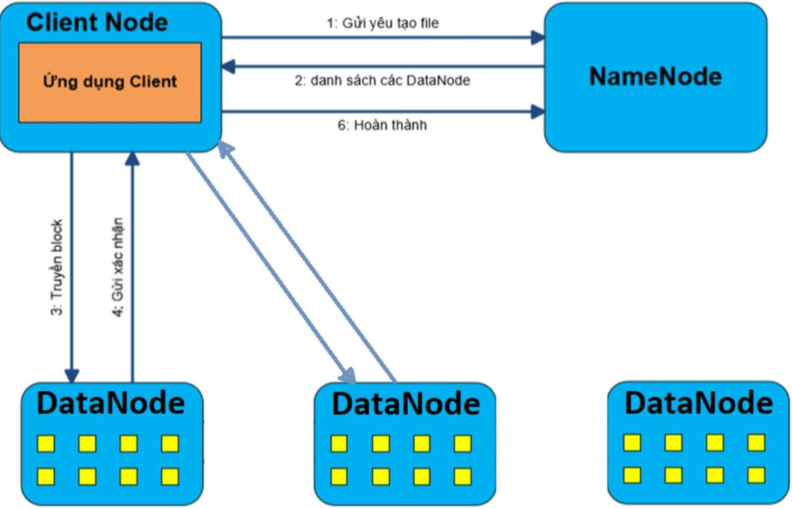
### HDFS là gì?

* Hệ thống file phân tán, cung cấp khả năng lưu trữ dữ liệu khổng lồ và tính năng tối ưu hóa việc sử dụng băng thông giữa các node.
* Có thể được sử dụng để chạy trên 1 cluster lớn với hàng chục ngàn node



*Hình 2.4.3.1 Kiến trúc HDFS*

 *Hình 2.4.3.2 Quá trình đọc file*

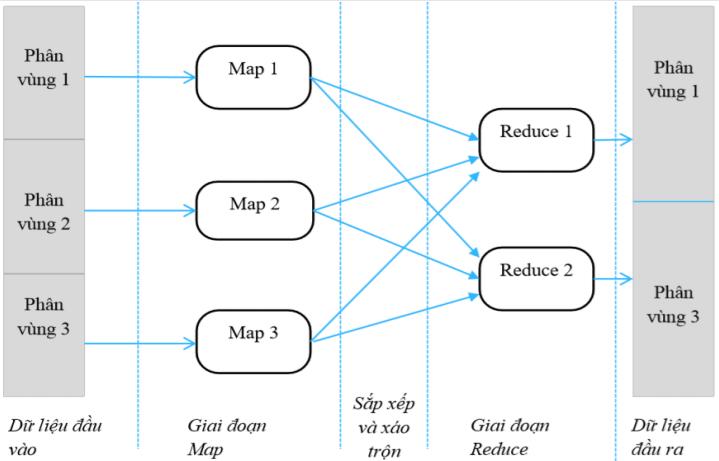


*Hình 2.4.3.3 Quá trình tạo và ghi dữ liệu lên file HDFS*

* NameNode: Là nút quản lý trung tâm của HDFS, duy trì thông tin về các tệp và thư mục trong hệ thống. Nó lưu trữ siêu dữ liệu (metadata) về các tệp dữ liệu, bao gồm thông tin về địa chỉ vật lý của các DataNode lưu trữ các phần của các tệp, quản lý quá trình đọc/ghi dữ liệu, phân phối tệp dữ liệu trên nhiều DataNode, và xử lý các yêu cầu của người dùng và ứng dụng.
* DataNode: Là các nút lưu trữ dữ liệu thực tế trong HDFS. Các DataNode lưu trữ các phần của các tệp dữ liệu, chịu trách nhiệm quản lý và bảo vệ các phần dữ liệu này trên đĩa cứng của máy chủ của nó, và báo cáo tình trạng của nó về các phần dữ liệu đó đến NameNode. Các DataNode cũng thực hiện các tác vụ như đọc và ghi dữ liệu từ/đến đĩa cứng của nó và trả lại dữ liệu cho người dùng và ứng dụng khi được yêu cầu.
* Block: HDFS lưu trữ các tập tin dưới dạng các khối có kích thước cố định. Các khối này có thể được lưu trữ trên nhiều DataNode khác nhau trong hệ thống.

### MapReduce là gì?

* MapReduce là mô hình dùng cho xử lý tính toán song song và phân tán trên hệ thống phân tán
* Map sử dụng để xử lý và trích xuất dữ liệu từ tập dữ liệu đầu vào và tạo ra các cặp khóa-giá trị tạm thời. Hệ thống thực hiện bước trung gian để trộn và sắp xếp lại kết quả.
* Reduce sử dụng để tổng hợp các cặp khóa-giá trị tạm thời và tạo ra kết quả đầu ra cuối cùng.



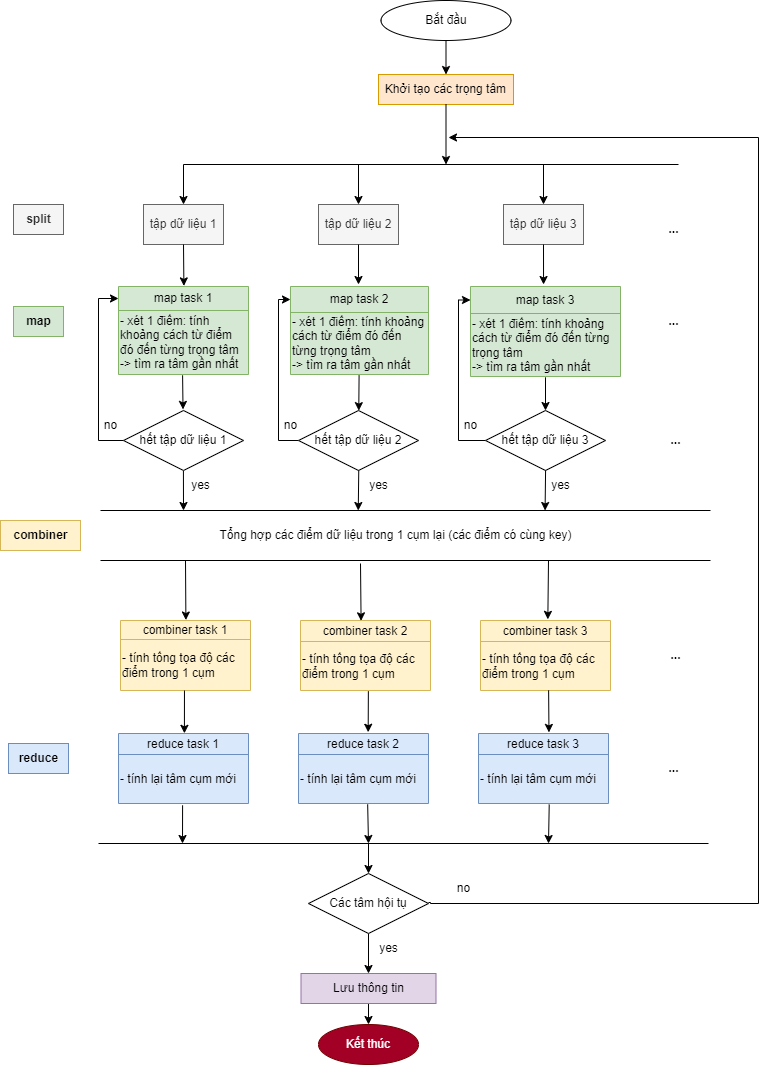
*Hình 2.4.4.1 Mô hình thực thi MapReduce*

# KẾT QUẢ VÀ ĐÁNH GIÁ MÔ HÌNH

## Phân tích chi tiết bài toán

* Sau khi tìm hiểu thuật toán K-means ta nhận thấy việc tính toán khoảng cách từ các điểm đến các tâm cụm tốn rất nhiều thời gian => cần Mapreduce hóa K-Means để có thể thực hiện tính toán song song, đồng thời giúp giảm thời gian thực hiện thuật toán và tăng hiệu quả.
* Giải pháp MapReduce hóa K-Mean:
  + Chia dữ liệu thành các phần nhỏ, khởi tạo k trọng tâm ngẫu nhiên.
  + Map(class KMapper.java):
    - Phân bổ từng phần dữ liệu nhỏ vào từng map task, thực thi công việc tính toán song song trên các map task, mỗi map task sẽ thực hiện công việc sau:
      * Intput: cặp (key,value): key giá trị byte offset, value: tập giá trị thuộc tính của 1 mẫu dữ liệu.
      * Công việc: tính toán khoảng cách từ 1 điểm đến từng tâm cụm, lưu lại id tâm cụm có khoảng cách nhỏ nhất đến điểm đang xét.
      * Output: cặp (key,value): key:id tâm cụm, value: điểm dữ liệu đang xét.
  + Combiner(class KCombiner.java):
    - Tổng hợp các cặp (key,value) có cùng key từ kết quả đầu ra của bước Map.
      * Intput: cặp (key,value): key id tâm cụm, value: danh sách các điểm dữ liệu có cùng key(các điểm dữ liệu cùng 1 cụm).
      * Công việc: tính tổng tọa độ các điểm trong 1 cụm dữ liệu.
      * Output: cặp (key,value): key:id tâm cụm, value: điểm dữ liệu có tọa độ = tổng tọa độ các điểm dữ liệu trong 1 cụm.
  + Reduce(class KReduce.java):
    - Nhận các cặp (key,value) từ bước KCombiner.
      * Intput: cặp (key,value): key id tâm cụm, value: danh sách các điểm dữ liệu có cùng key(các điểm dữ liệu cùng 1 cụm).
      * Công việc: tính tổng tọa độ các điểm trong 1 cụm dữ liệu, sau đó tính tạo độ trung bình của các điểm dữ liệu.
      * Output: cặp (key,value): key:id tâm cụm, value: tâm cụm mới.
* Quá trình mapreduce kết thúc khi tâm mới tạo ra không thay đổi so với tâm cũ trước đó.

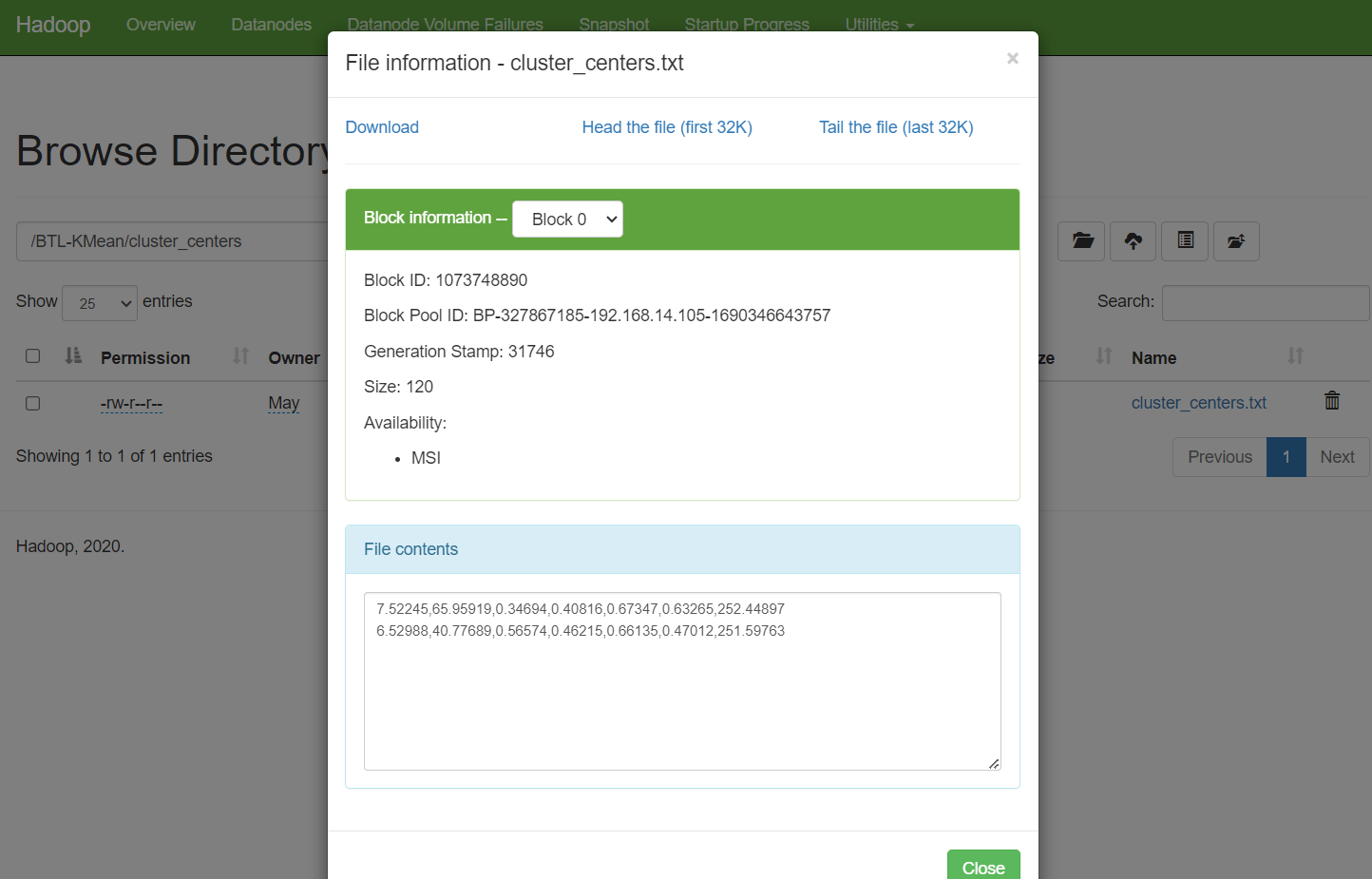
## Xây dựng lưu đồ



*Hình 3.2.1 Lưu đồ MapReduce của thuật toán K-Mean*

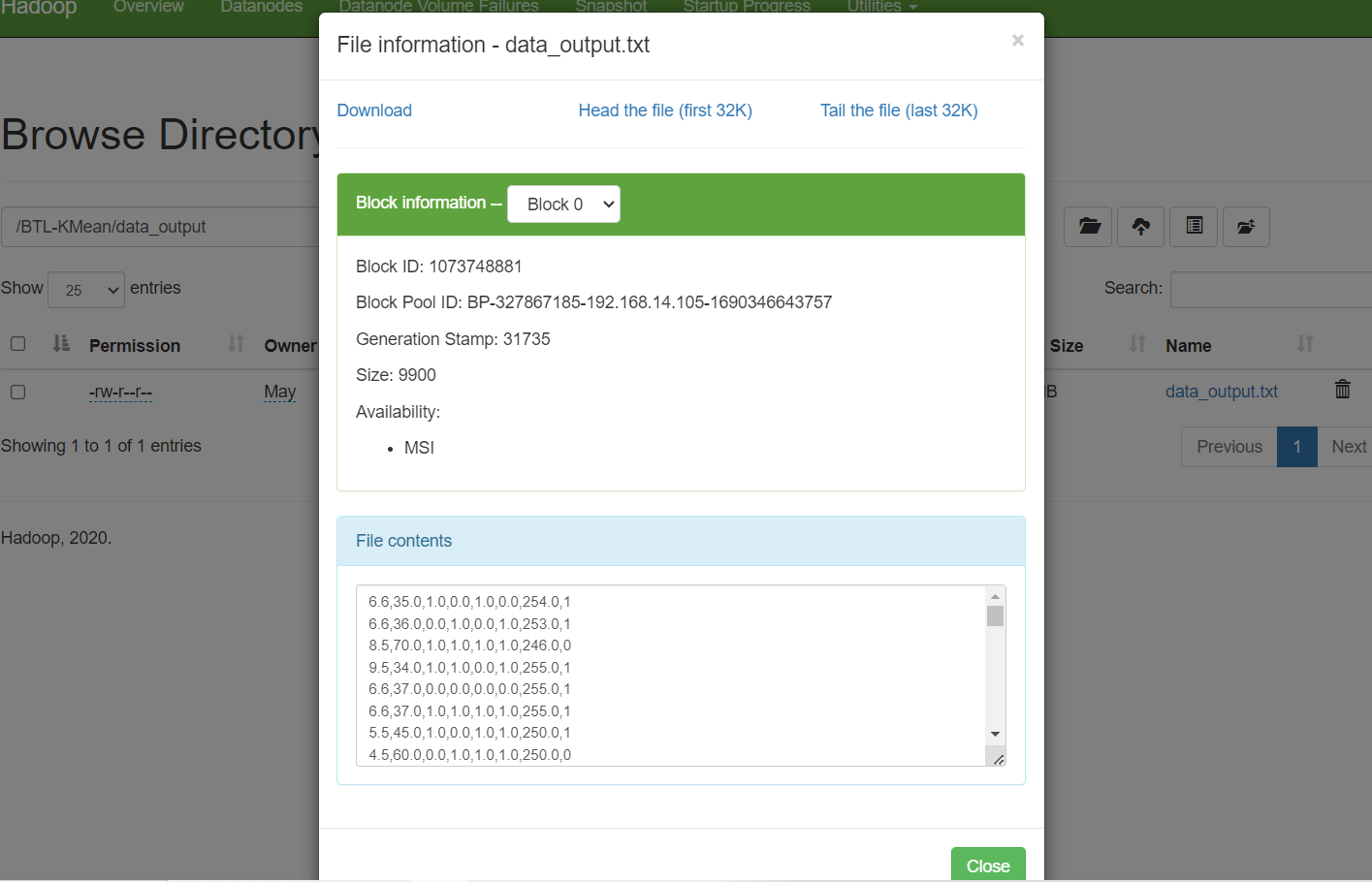
## Kết quả.

* Tọa độ k tâm cụm:



*Hình 3.3.1 kết quả tọa độ k tâm cụm*

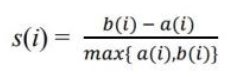
* Tập dữ liệu đã được phân cụm:



*Hình 3.3.2 Tập dữ liệu đầu đã được phân cụm*

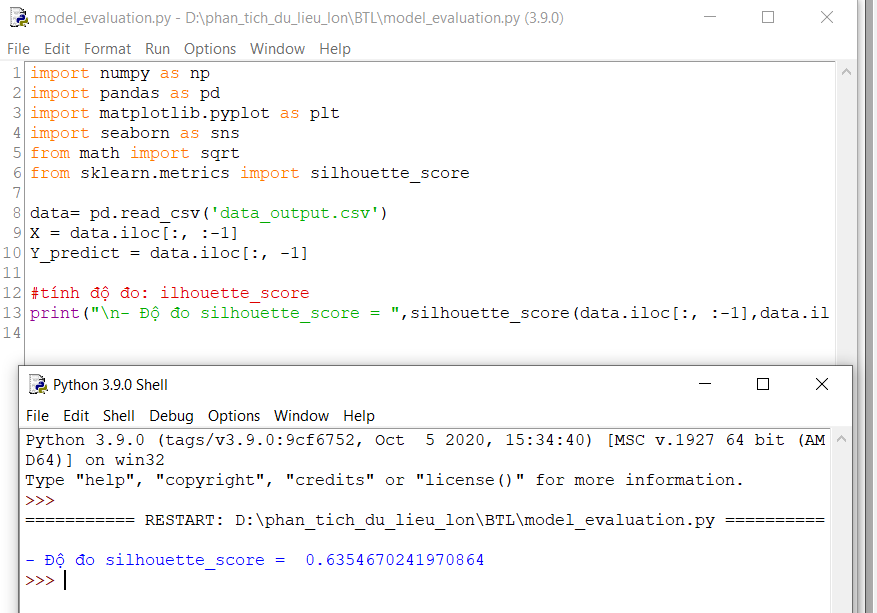
## Đánh giá mô hình

* Độ bóng (silhouette) có giá trị [-1,1]: giá trị càng tiến đến 1 thì mô hình càng tốt.
* Giả có k cụm, t có công thức:
* a(i): khoảng cách trung bình từ điểm i tới tất cả các điểm dữ liệu có chung cụm với i.
* b(i): khoảng cách trung bình ngắn nhất từ i tới bất kì cụm nào không chứa i. Cụm tương ứng với b(i) này được gọi là cụm hàng xóm của i.



*Hình 3.4.1 Công thức silhouette*

* Độ đo silhouette trong bài toán:

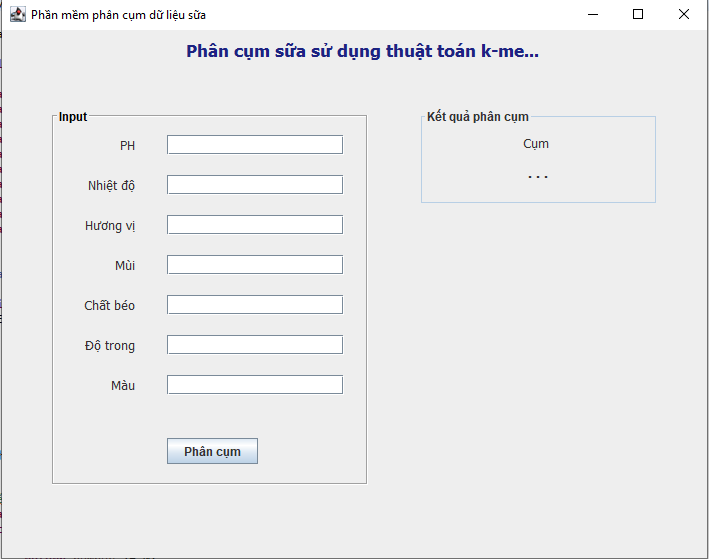


*Hình 3.4.2 Thông số đánh giá*

* Với độ chính xác, độ đo silhouette = 0.64 mô hình học máy này chưa thực sự tốt để phân cụm sữa.

## Sử dụng mô hình học được để phân cụm 1 mẫu dữ liệu sữa.

* Chức năng của chương trình: phân cụm sữa dựa trên các thông tin do người dùng nhập vào gồm 7 thông tin: pH, Nhiệt độ, Hương vị, mùi, chất béo, Độ trong của sữa, Màu sữa.



*Hình 3.5.1 Chương trình chạy phân cụm*

# KẾT LUẬN

Big data đã đặt ra thách thức cho các tổ chức, doanh nghiệp cũng như nhiều cơ hội. MapReduce chia việc xử lý thành nhiều khối công việc nhỏ, phân tán khắp các nút tính toán, rồi thu thập lại kết quả. Đề tài này đã áp dụng mô hình MapReduce trên một framework nguồn mở Hadoop để phân tích phân cụm dữ liệu sử dụng giải thuật K-Mean. Hoàn thành đề tài “Sử dụng thuật toán K-Mean kết hợp Hadoop phân cụm sữa”.

Em đã đạt được một số kết quả như sau:

* Hiểu tổng quan về BigData, Hadoop, MapReduce.
* Biết được cách kết hợp MapReduce với thuật toán phân cụm K-Mean.
* Xây dựng được chương trình MapReduce kết hợp K-Mean.
* Đánh giá được mô hình và chương trình.

Trong quá trình hoàn thành bài tập lớn, em có tìm hiểu và tham khảo các tài liệu liên quan. Tuy nhiên, thời gian có hạn nên chúng em không tránh được thiếu sót, rất mong sự đóng góp ý kiến của thầy để báo cáo và kỹ năng của em càng ngày càng hoàn thiện hơn. Em xin chân thành cảm ơn.

**TÀI LIỆU THAM KHẢO**

|  |  |
| --- | --- |
| [1] | H. S. Hùng. [Online]. Available: https://techmaster.vn/posts/33923/lich-su-phat-trien-machine-learning. |
| [2] | [Online]. Available: https://hadoop.apache.org/. |
| [3] | "wikipedia," [Online]. Available: https://vi.wikipedia.org/wiki/Apache\_Hadoop. |

[4] <https://www.kaggle.com/datasets/yrohit199/milk-quality>