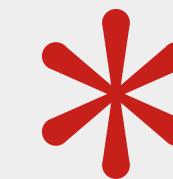
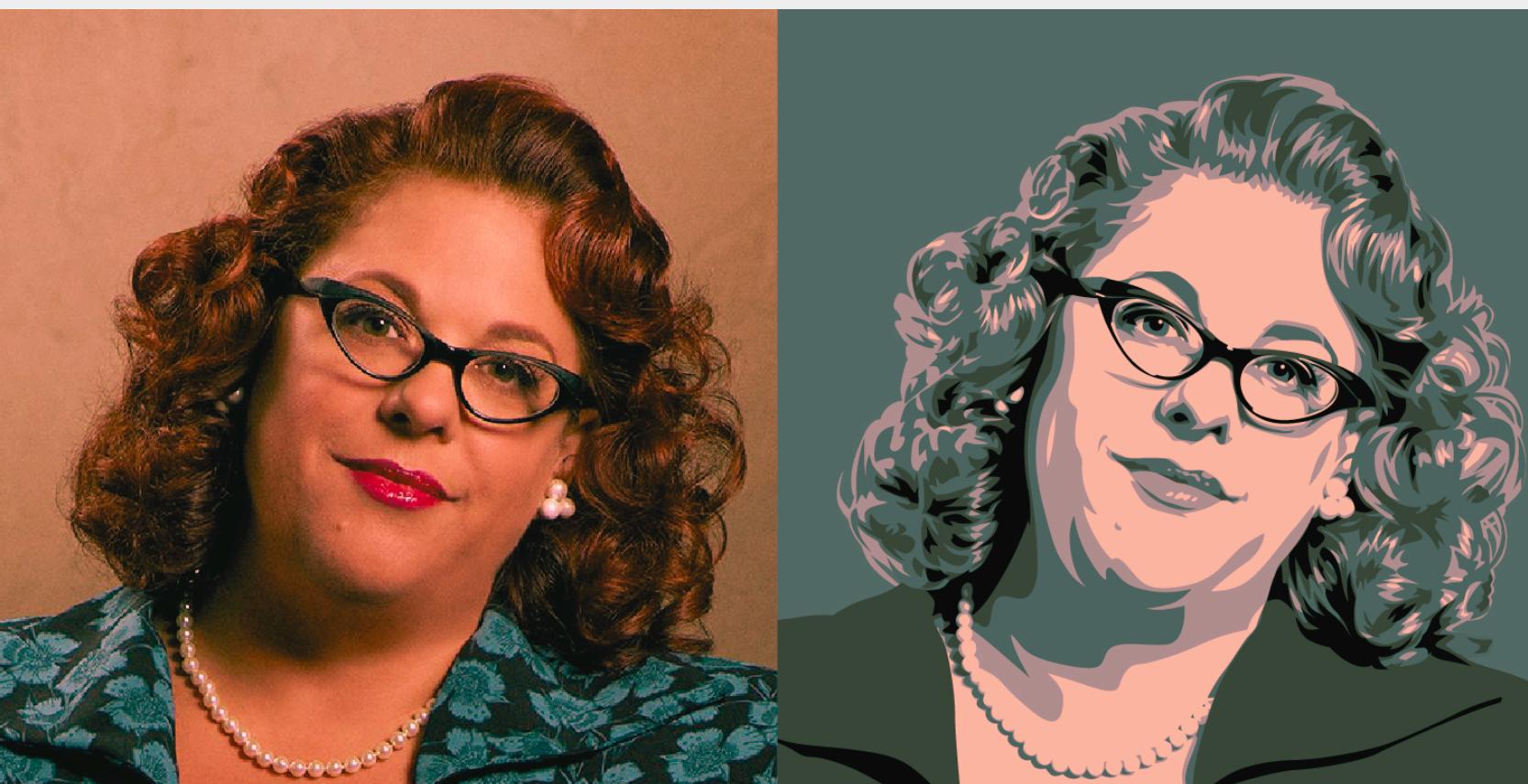


# PHÂN CỤM HÌNH ẢNH

## SỬ DỤNG KMEANS MAPREDUCE



### NHÓM VITAMIN TRI THỨC

Nguyễn Viết Vũ  
Nguyễn Tông Quân  
Phạm Đăng Phong  
Nguyễn Xuân Trình

- 22022632  
- 22022632  
- 22022632  
- 22022632

# mục lục

Dữ liệu lớn

---

Thuật toán K-Means

---

K-Means MapReduce trong phân cụm ảnh

---

Cải tiến thuật toán & ứng dụng

---

Demo

# dữ liệu lớn

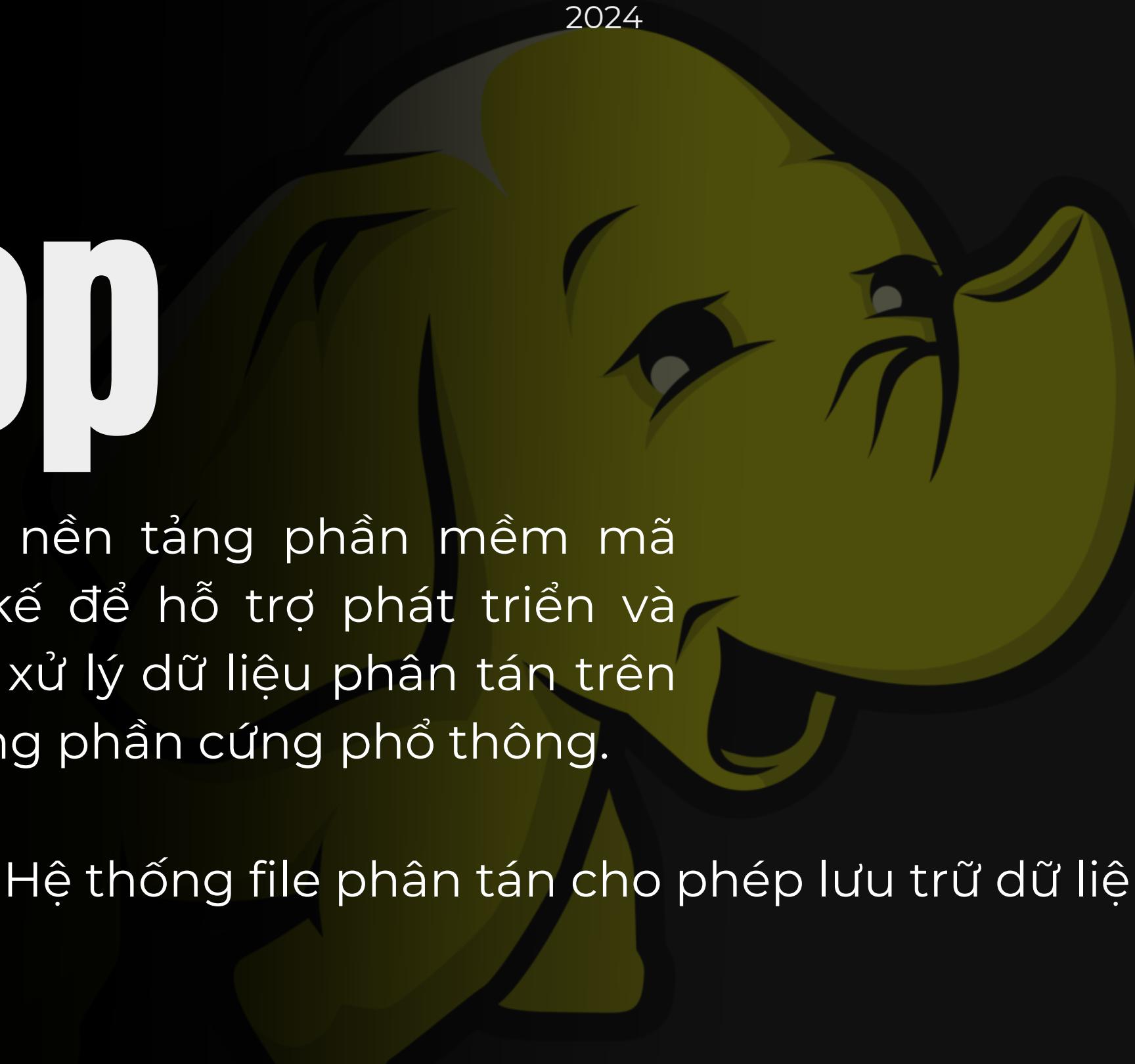


Dữ liệu lớn là thuật ngữ dùng để mô tả các bộ dữ liệu có kích thước hoặc độ phức tạp vượt quá khả năng xử lý của các phương pháp truyền thống.

- \* **khối lượng lớn**
- \* **tốc độ xử lý cao**
- \* **định dạng đa dạng**

Giao dịch trực tuyến, mạng xã hội, cảm biến IoT, hoặc các tài liệu, hình ảnh được chia sẻ trên internet...

# hadoop



Apache Hadoop là một nền tảng phần mềm mã nguồn mở, được thiết kế để hỗ trợ phát triển và triển khai các ứng dụng xử lý dữ liệu phân tán trên các cụm máy tính sử dụng phần cứng phổ thông.

## \* **HDFS**

Hệ thống file phân tán cho phép lưu trữ dữ liệu trên nhiều node

## \* **mapreduce**

Mô hình lập trình hỗ trợ xử lý dữ liệu bằng cách chia nhỏ công việc

# hadoop

- \* **HDFS**
- \* **mapreduce**

**mapreduce**

Xử lý dữ liệu

**others**

Xử lý dữ liệu

**YARN**

Quản lý tài nguyên cụm

**HDFS**

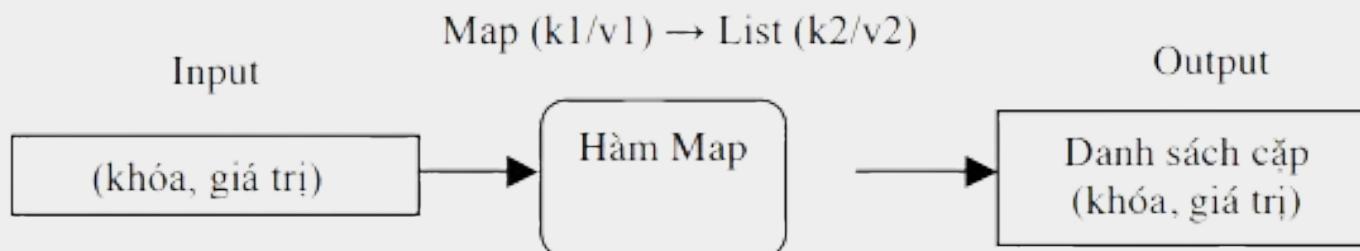
Lưu trữ dữ liệu

# mapreduce

MapReduce là một mô hình lập trình được thiết kế để xử lý tính toán song song và phân tán trên các hệ thống phân tán.

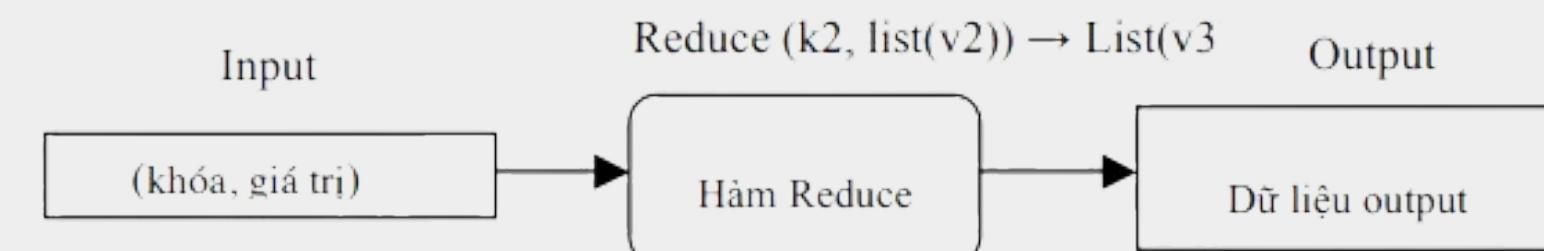
## \* map

Ánh xạ dữ liệu đầu vào thành các cặp khóa/giá trị trung gian.



## \* reduce

Tập hợp các giá trị có cùng khóa từ kết quả của hàm Map và tạo ra kết quả cuối cùng.



Thuật toán K-Means được sử dụng phổ biến trong bài toán phân cụm (Clustering). Đây là một thuật toán học không giám sát, có mục tiêu nhóm dữ liệu thành k cụm dựa trên sự tương đồng giữa các điểm dữ liệu.

---

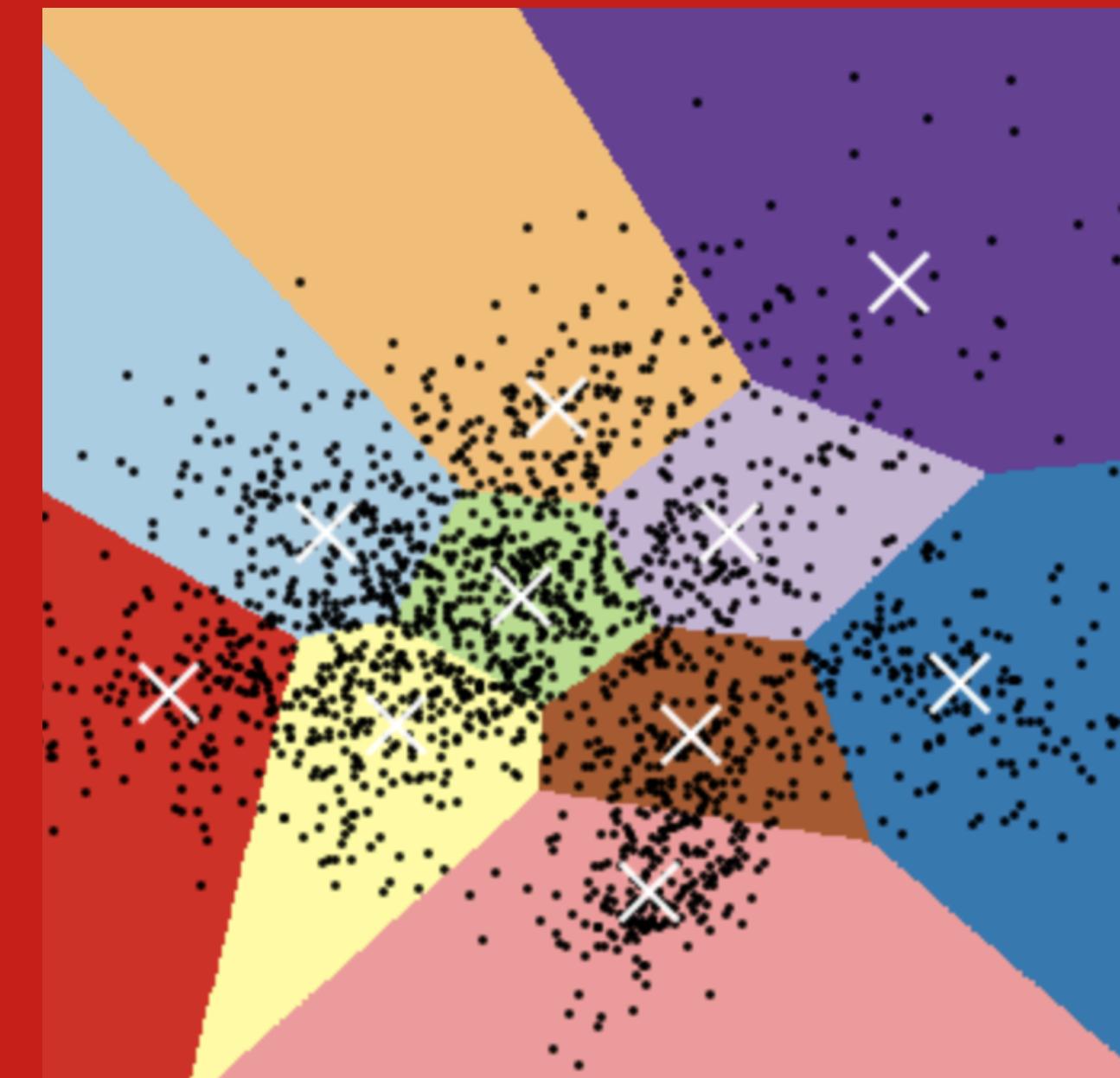
### \* **thực hiện**

Gán điểm dữ liệu cho khoảng cách đến tâm cụm là ngắn nhất.

Cập nhật tâm cụm = trung bình của các điểm dữ liệu trong cụm.

Lặp lại đến khi tâm cụm không còn thay đổi hoặc hội tụ.

# kmeans



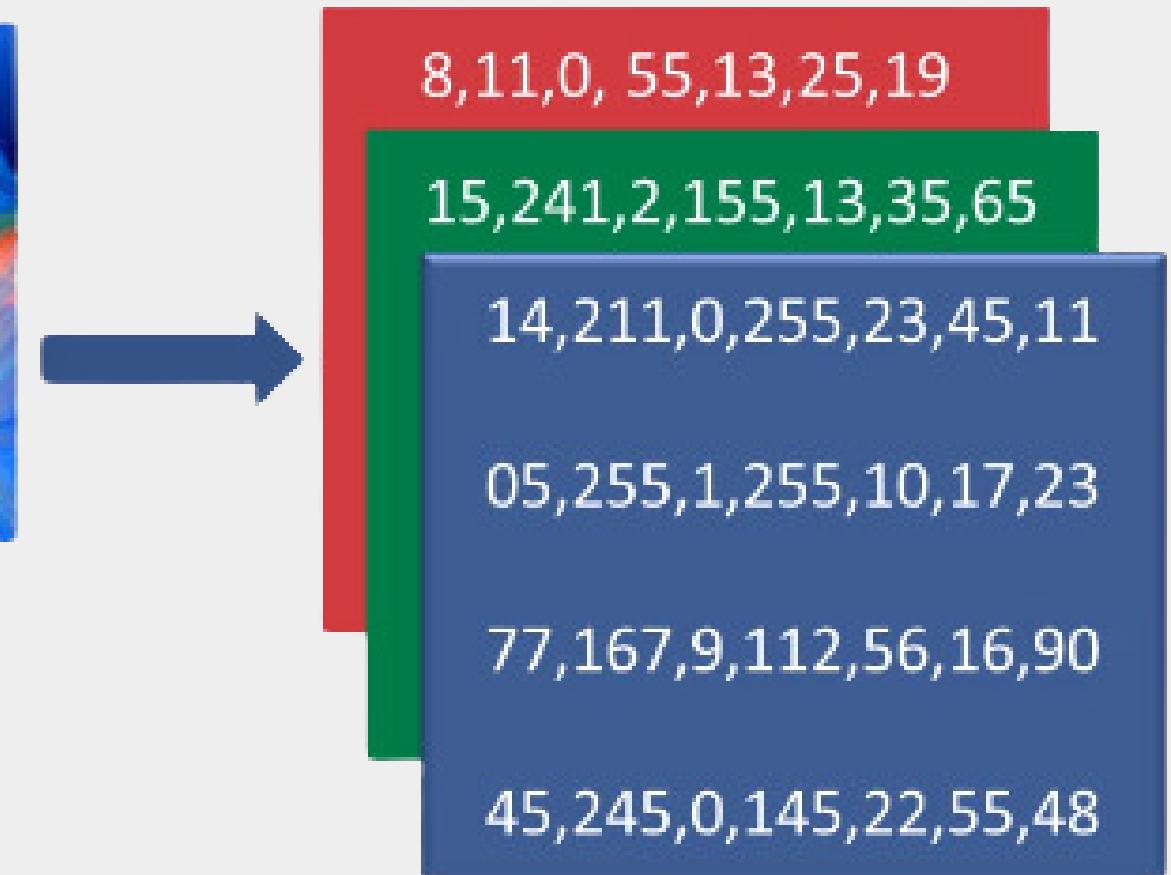
# phân cụm hình ảnh

## sử dụng kmeans mapreduce

### \* Ảnh

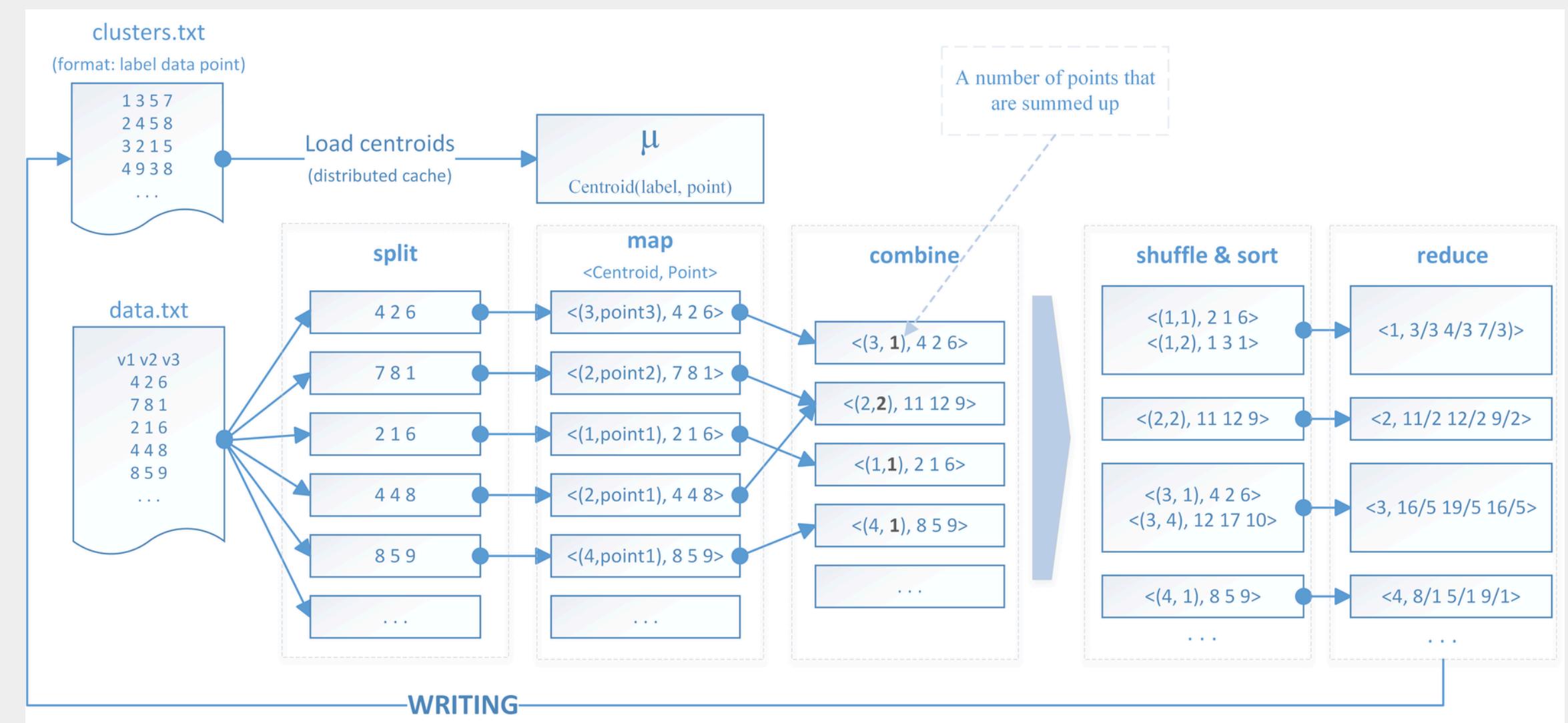
Ảnh là một ma trận điểm ảnh (R, G, B).

Mỗi điểm ảnh là một vector đặc trưng  
[R, G, B]



# phân cụm hình ảnh

- Chuyển đổi ảnh thành tập hợp điểm
- Khởi tạo tâm cụm
- Thực hiện MapReduce:
  - Mapper: tính khoảng cách và gán điểm ứng với cụm gần nhất
  - Reduce: tập hợp các điểm thuộc mỗi cụm, cập nhật tâm cụm mới
- Lặp lại quá trình tới khi tâm cụm hội tụ



# cải tiến

hàm khoảng cách

thuật toán Kmeans++

thử nghiệm với HOG và hệ màu CMYK

# cải tiến

## hàm khoảng cách

Ban đầu: Khoảng cách Euclidean

- Thời gian tính toán kéo dài
- Phân nhóm không tối ưu nếu cụm màu tương tự nhau

Bổ sung: Khoảng cách Cosine

- Giảm nhiễu nếu dữ liệu có độ chênh lệch lớn về không gian

# cải tiến

## thuật toán Kmeans++

Ban đầu: Tâm cụm được khởi tạo ngẫu nhiên

- Dẫn đến hội tụ chậm hoặc kém hiệu quả
- Mất một số tâm sau khi hội tụ

Cải tiến: Thuật toán Kmeans++

- Trích xuất được các tâm phản ảnh màu chính của ảnh
- Số vòng lặp để hội tụ giảm dẫn đến hội tụ nhanh hơn

# cải tiến

## thuật toán Kmeans++

- Một tâm cụm được chọn ngẫu nhiên trong tập khởi tạo
- Tính khoảng cách đến centroid đã chọn
- Chọn centroid tiếp theo: việc chọn này không ngẫu nhiên, mà có xác suất tỷ lệ bình phương khoảng cách từ điểm ảnh đến centroid gần nhất
- Lặp lại bước 2 và bước 3 đến khi đủ số lượng centroid

# cải tiến

## thử nghiệm với HOG và hệ màu CMYK

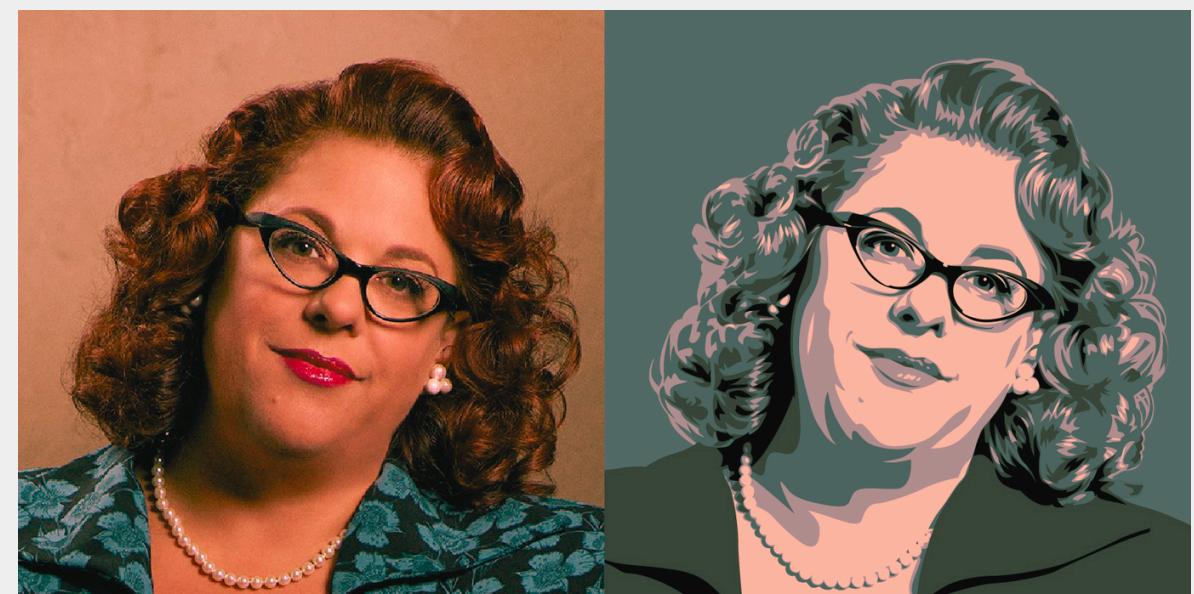
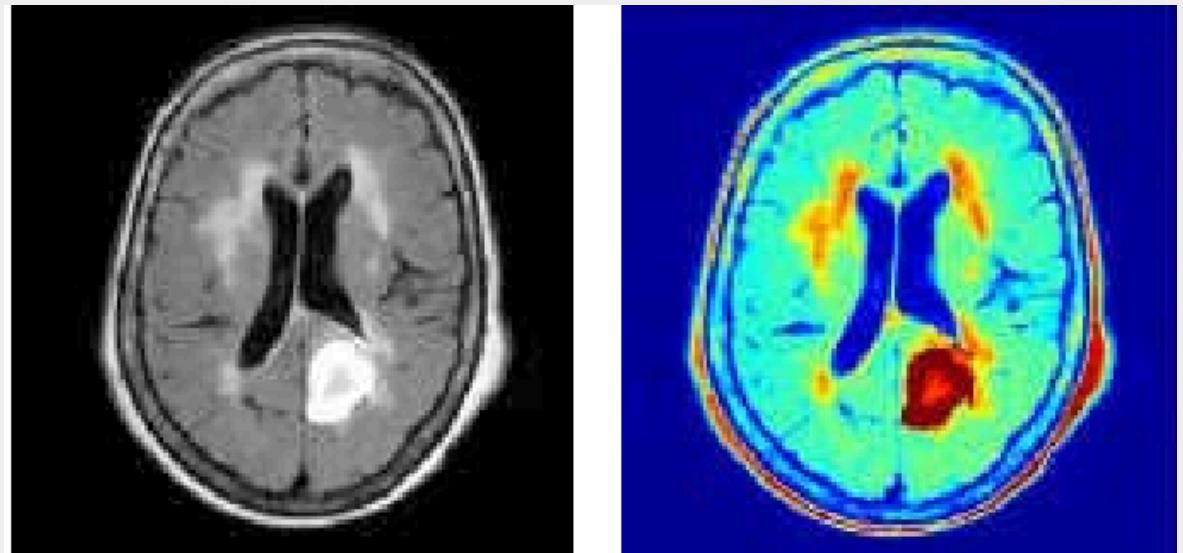
- Việc sử dụng HOG có thể giúp giảm kích thước ảnh và giúp giảm độ phức tạp tính toán.
- Thử nghiệm với hệ màu CMYK là vector 4 chiều thay vì 3 chiều như RGB



# Ứng dụng



- Phân đoạn ảnh: Tách vùng có đặc trưng giống nhau (y tế, thị giác máy).
- Nén ảnh: Giảm số lượng màu sắc, tối ưu lưu trữ.
- Nhận diện đối tượng: Nhóm đặc trưng để nhận dạng vật thể.
- Phân loại ảnh: Nhóm ảnh theo cảnh quan (biển, rừng, đô thị).
- Tạo bảng màu: Trích xuất màu chính từ ảnh (đồ họa, thời trang).
- Xử lý video: Phân cụm khung hình tương đồng, tối ưu dữ liệu.



**DEMO**

trinhdz@DESKTOP-2EP69KG:~/bigdata/kmeans\_mapreduce-master\$ start-all.sh  
WARNING: Attempting to start all Apache Hadoop daemons as trinhdz in 10 seconds.  
WARNING: This is not a recommended production deployment configuration.  
WARNING: Use CTRL-C to abort.  
Starting namenodes on [0.0.0.0]  
Starting datanodes  
Starting secondary namenodes [DESKTOP-2EP69KG]  
Starting resourcemanager  
Starting nodemanagers  
trinhdz@DESKTOP-2EP69KG:~/bigdata/kmeans\_mapreduce-master\$ jps  
2032 ResourceManager  
1433 DataNode  
2505 Jps  
1691 SecondaryNameNode  
1213 NameNode  
2190 NodeManager  
trinhdz@DESKTOP-2EP69KG:~/bigdata/kmeans\_mapreduce-master\$

Hadoop Overview Datanodes Datanode Volume Failures Snapshot Startup Progress Utilities ▾

## Browse Directory

/ Go! 📁 ⤴ 📄 🔗

Show 25 entries Search:

<input type="checkbox"/>	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name	<span>trash</span>
<input type="checkbox"/>	drwxr-xr-x	trinhdz	supergroup	0 B	Nov 23 10:48	0	0 B	KMeans	<span>trash</span>
<input type="checkbox"/>	drwxr-xr-x	trinhdz	supergroup	0 B	Nov 23 10:49	0	0 B	tmp	<span>trash</span>
<input type="checkbox"/>	drwxr-xr-x	trinhdz	supergroup	0 B	Dec 02 07:55	0	0 B	user	<span>trash</span>

Showing 1 to 3 of 3 entries Previous 1 Next

(venv) trinhdz@DESKTOP-2EP69KG:~/bigdata/kmeans\_mapreduce-master\$ python3 data\_prep\_scripts/data\_prep.py  
Points saved in: /home/trinhdz/bigdata/kmeans\_mapreduce-master/Resources/Input/points.txt  
Centroids saved in: /home/trinhdz/bigdata/kmeans\_mapreduce-master/Resources/Input/clusters.txt





**thanks for watching**