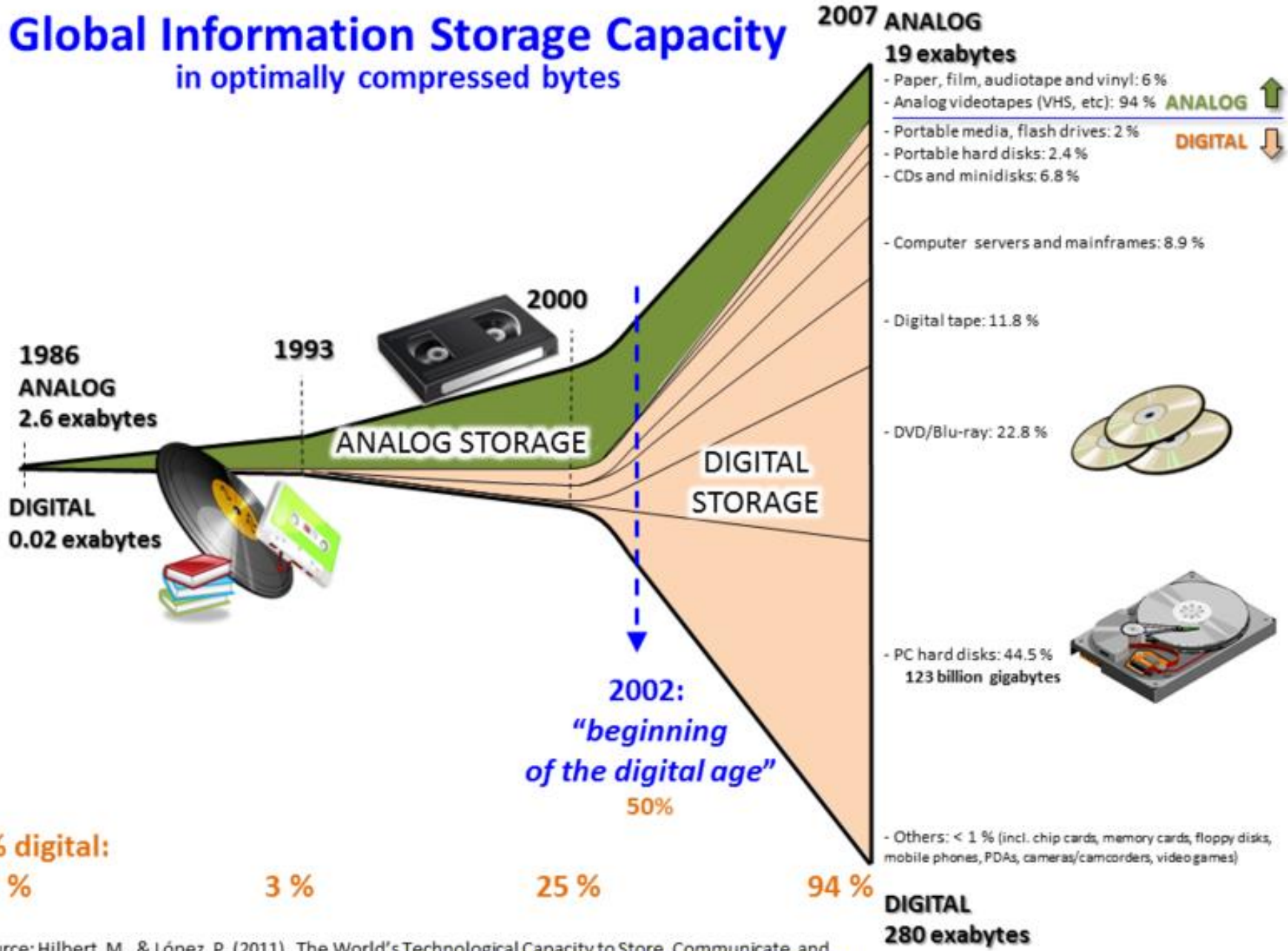
Dựa theo bài báo: Dữ liệu xã hội – Những Thành tựu và thách thức gần đây

Speaker: Lương Văn Quý

Global Information Storage Capacity

in optimally compressed bytes



2007 **ANALOG**
19 exabytes

- Paper, film, audiotape and vinyl: 6 %
- Analog videotapes (VHS, etc): 94 % **ANALOG**

DIGITAL

- Portable media, flash drives: 2 %
- Portable hard disks: 2.4 %
- CDs and minidisks: 6.8 %
- Computer servers and mainframes: 8.9 %
- Digital tape: 11.8 %
- DVD/Blu-ray: 22.8 %

- PC hard disks: 44.5 %
  123 billion gigabytes

- Others: < 1 % (incl. chip cards, memory cards, floppy disks, mobile phones, PDAs, cameras/camcorders, video games)

DIGITAL
280 exabytes

Source: Hilbert, M., & López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*, 332(6025), 60 –65. http://www.martinhilbert.net/WorldInfoCapacity.html

1. Tổng quan về dữ liệu lớn

2. Giới thiệu phương pháp phân tích dữ liệu lớn

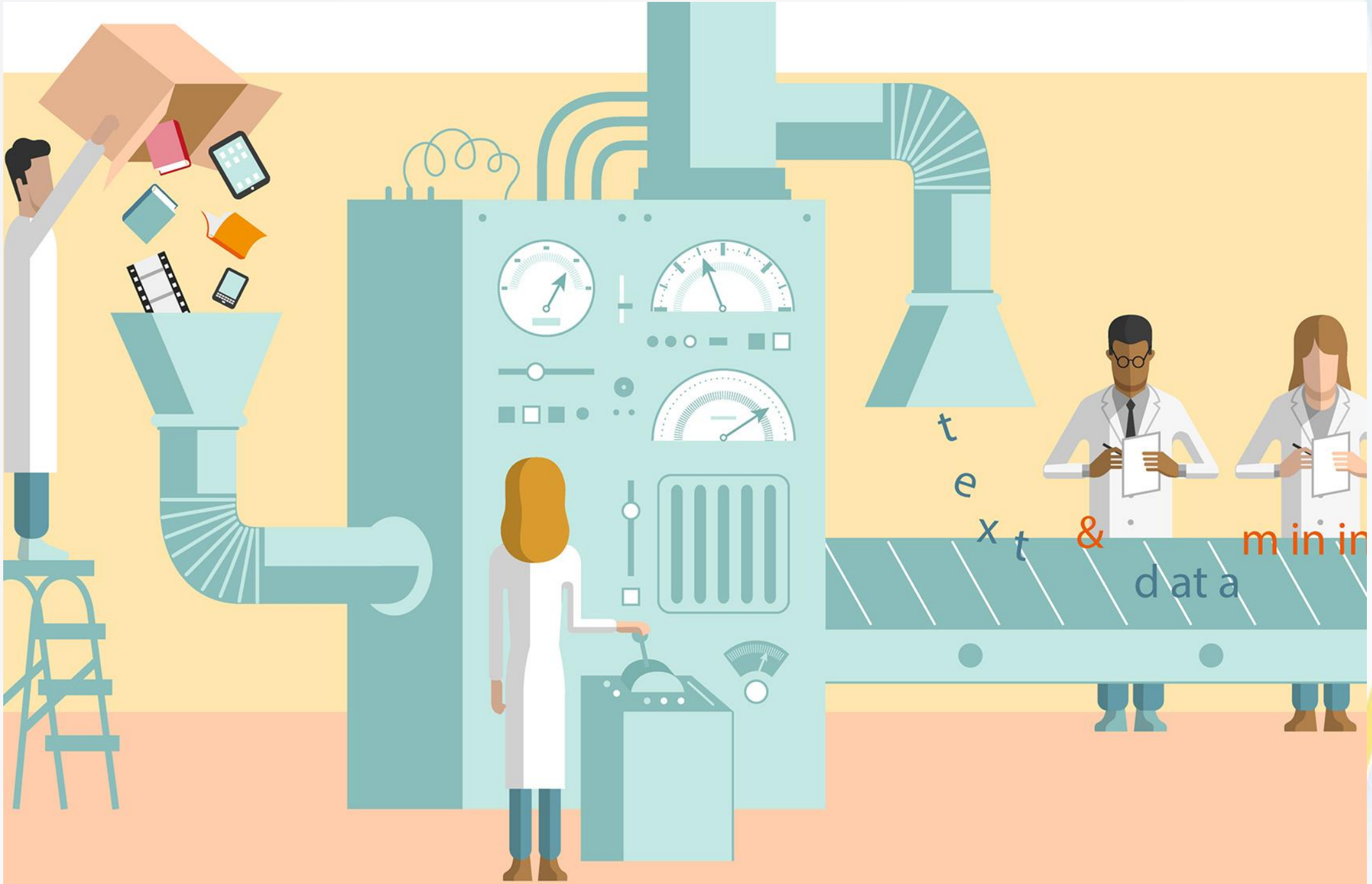
3. Những thành tựu đã đạt được

4. Những vấn đề mở

Khái niệm dữ liệu lớn

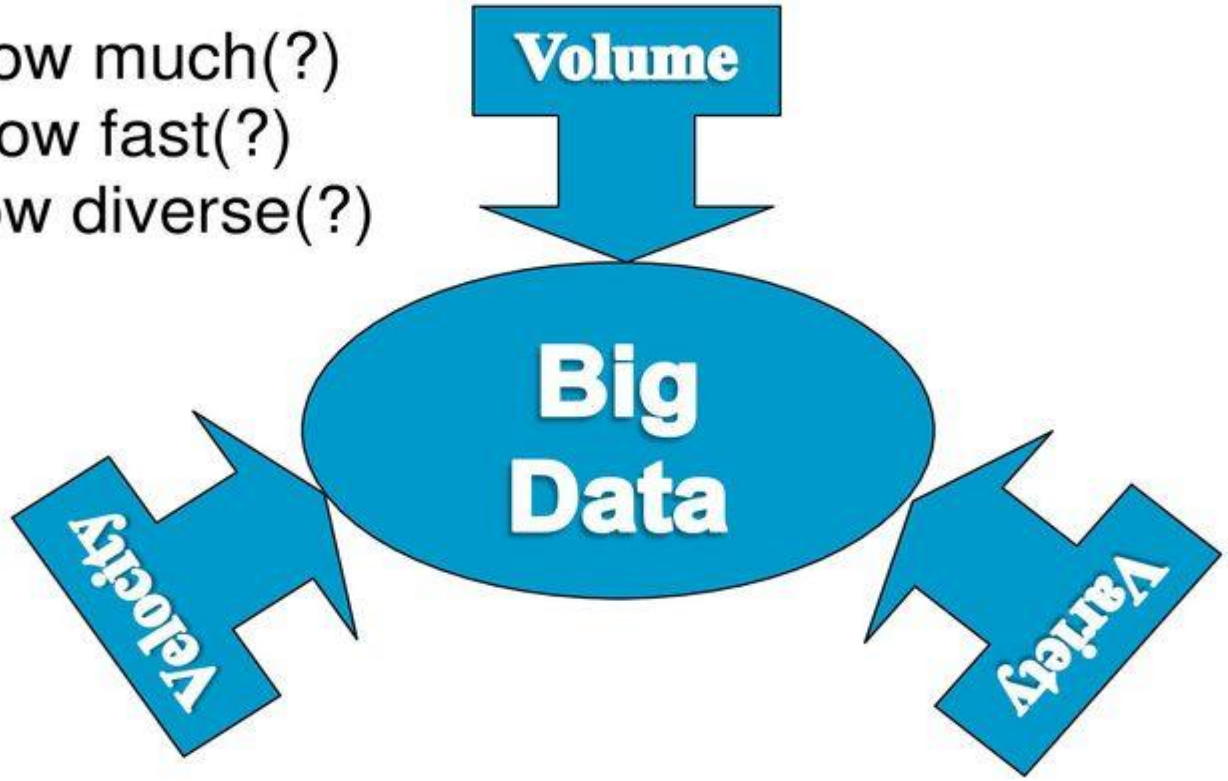
- Theo Wikipedia:

Dữ liệu lớn (Tiếng Anh: **Big data**) là một thuật ngữ cho việc xử lý một tập hợp dữ liệu rất lớn và phức tạp mà các ứng dụng xử lý dữ liệu truyền thống không xử lý được.


t
e
x t
&
m in in
d at a

The 3V's of Big Data

Volume – how much(?)
Velocity – how fast(?)
Variety – how diverse(?)



The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015

4.4 MILLION IT JOBS

will be created globally to support big data, with 1.9 million in the United States

Volume

SCALE OF DATA

40 ZETTABYTES

[43 TRILLION GIGABYTES]

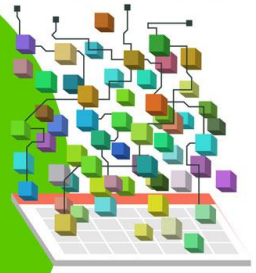
of data will be created by 2020, an increase of 300 times from 2005

It's estimated that

2.5 QUINTILLION BYTES

[2.3 TRILLION GIGABYTES]

of data are created each day



6 BILLION PEOPLE

have cell phones

WORLD POPULATION: 7 BILLION

Most companies in the U.S. have at least

100 TERABYTES

[100,000 GIGABYTES]

of data stored

Variety

DIFFERENT FORMS OF DATA

As of 2011, the global size of data in healthcare was estimated to be

150 EXABYTES

[161 BILLION GIGABYTES]

30 BILLION PIECES OF CONTENT

are shared on Facebook every month

By 2014, it's anticipated there will be

420 MILLION WEARABLE, WIRELESS HEALTH MONITORS

4 BILLION+ HOURS OF VIDEO

are watched on YouTube each month

400 MILLION TWEETS

are sent per day by about 200 million monthly active users

Velocity

ANALYSIS OF STREAMING DATA

The New York Stock Exchange captures

1 TB OF TRADE INFORMATION

during each trading session

Modern cars have close to

100 SENSORS

that monitor items such as fuel level and tire pressure

By 2016, it is projected there will be

18.9 BILLION NETWORK CONNECTIONS

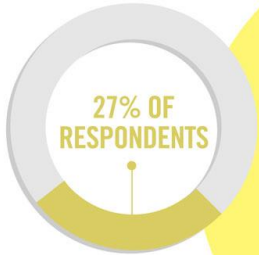
– almost 2.5 connections per person on earth

Veracity

UNCERTAINTY OF DATA

1 IN 3 BUSINESS LEADERS

don't trust the information they use to make decisions



in one survey were unsure of how much of their data was inaccurate

Poor data quality costs the US economy around

$3.1 TRILLION A YEAR

Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

IBM

Minh họa MapReduce bằng bài toán đếm từ

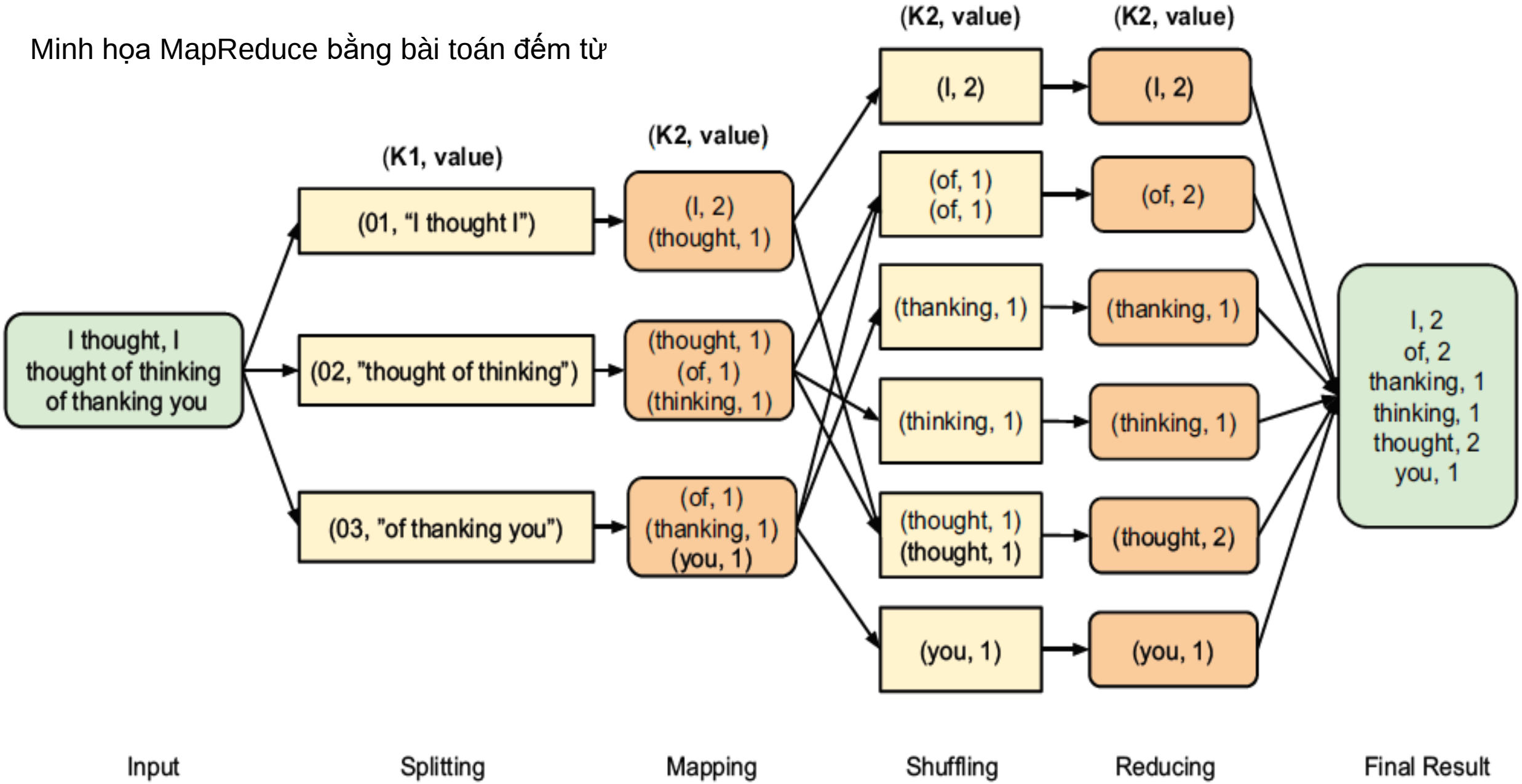
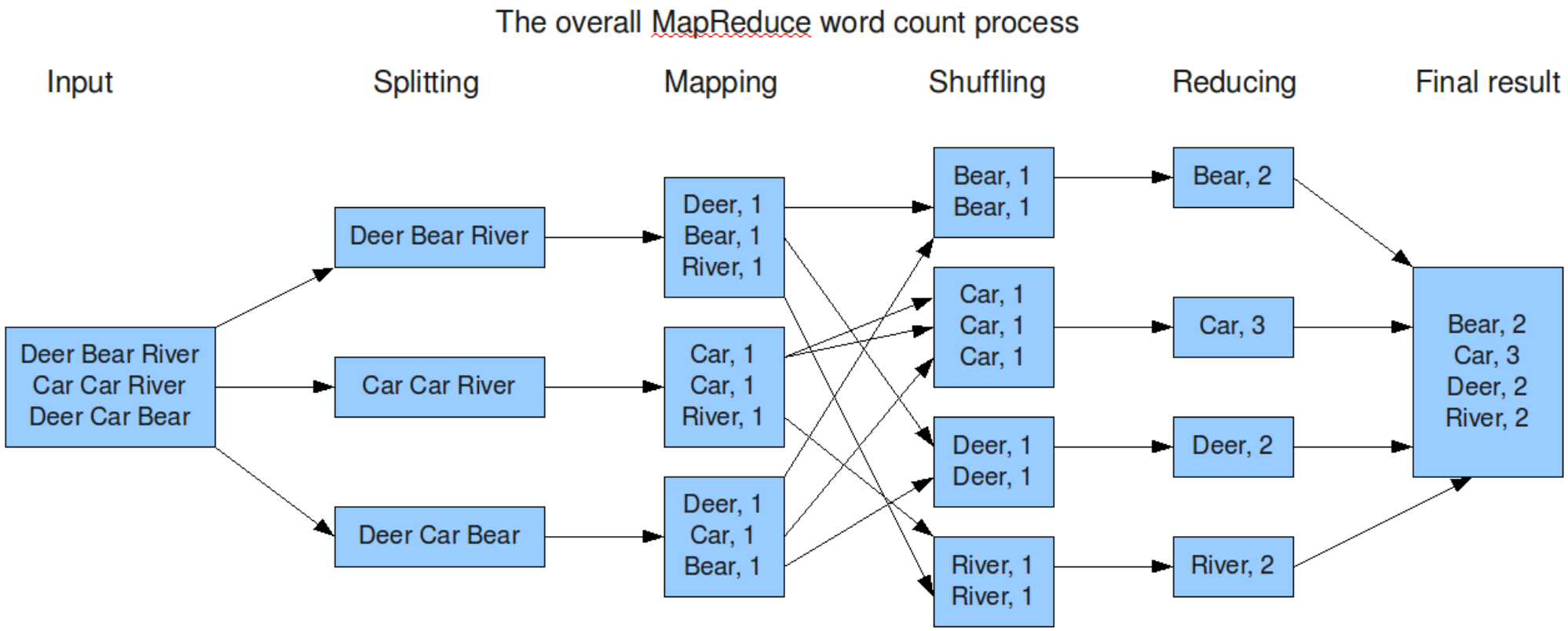
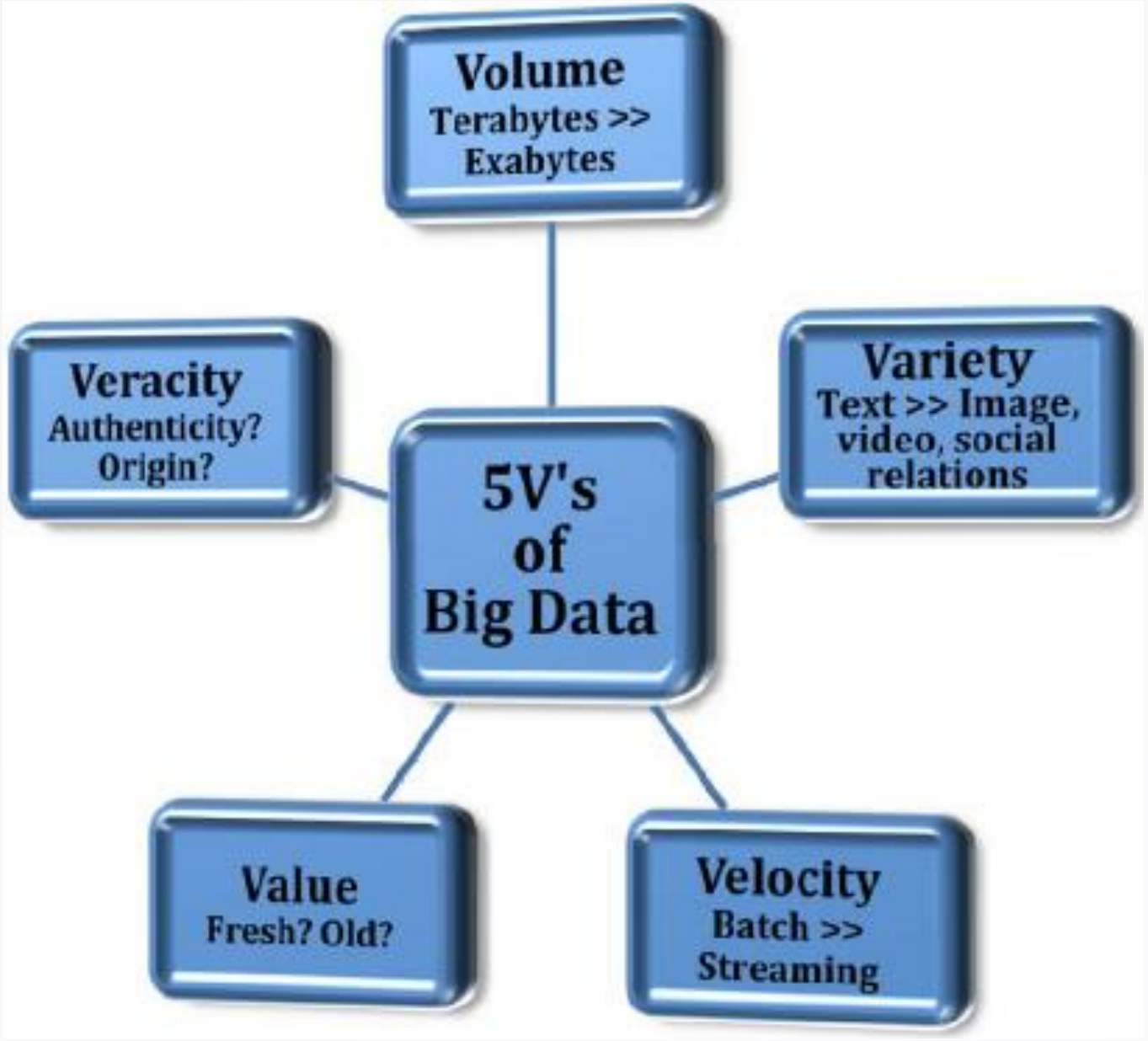


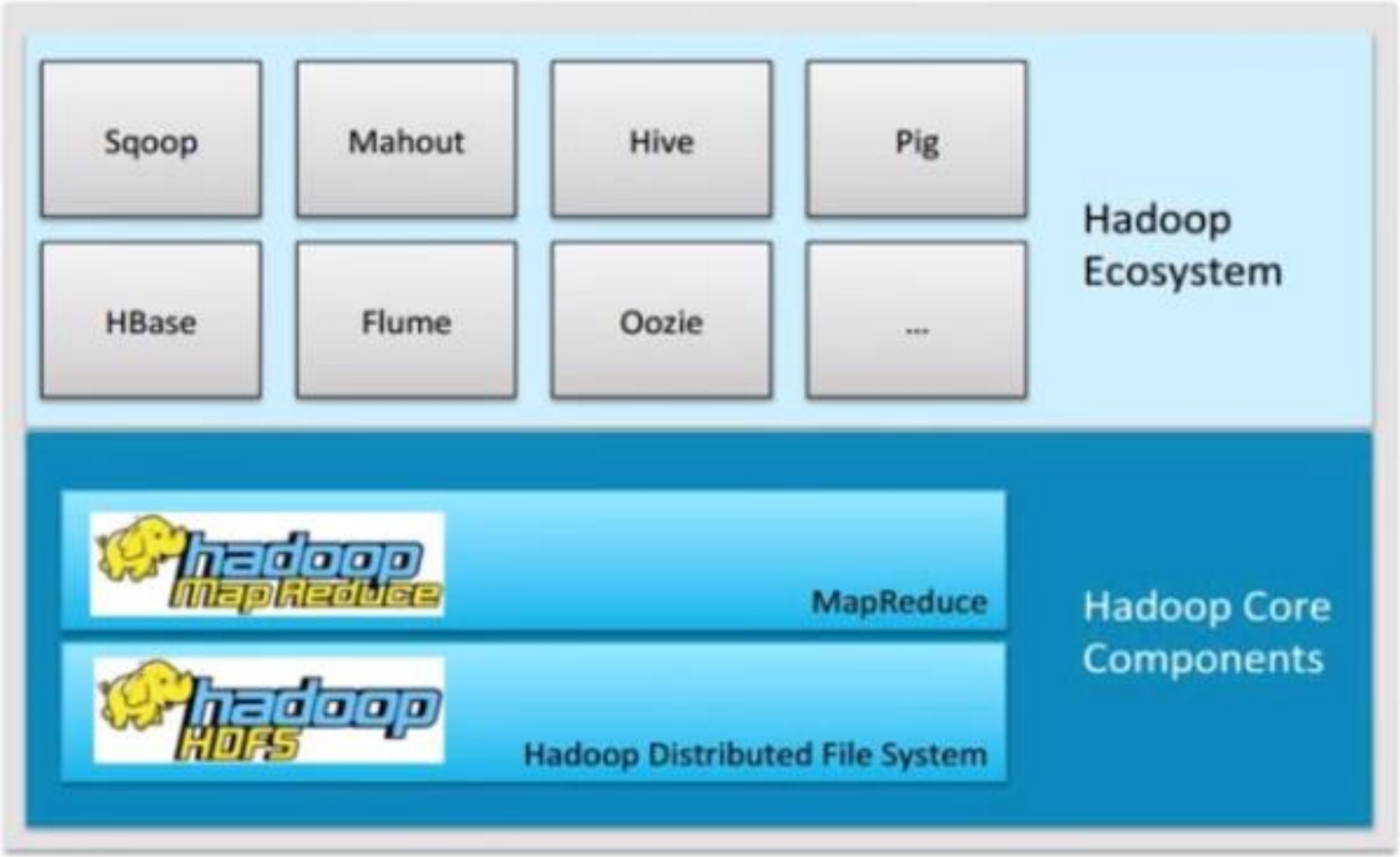
Fig. 2. The MapReduce processes for counting words in a text.

MapReduce


The overall MapReduce word count process
Input
Splitting
Mapping
Shuffling
Reducing
Final result
Deer Bear River
Car Car River
Deer Car Bear
Deer Bear River
Car Car River
Deer Car Bear
Deer, 1
Bear, 1
River, 1
Car, 1
Car, 1
River, 1
Deer, 1
Car, 1
Bear, 1
Bear, 1
Bear, 1
Car, 1
Car, 1
Car, 1
Deer, 1
Deer, 1
River, 1
River, 1
Bear, 2
Car, 3
Deer, 2
River, 2
Bear, 2
Car, 3
Deer, 2
River, 2


Volume
Terabytes >>
Exabytes
Veracity
Authenticity?
Origin?
5V's
of
Big Data
Variety
Text >> Image,
video, social
relations
Value
Fresh? Old?
Velocity
Batch >>
Streaming

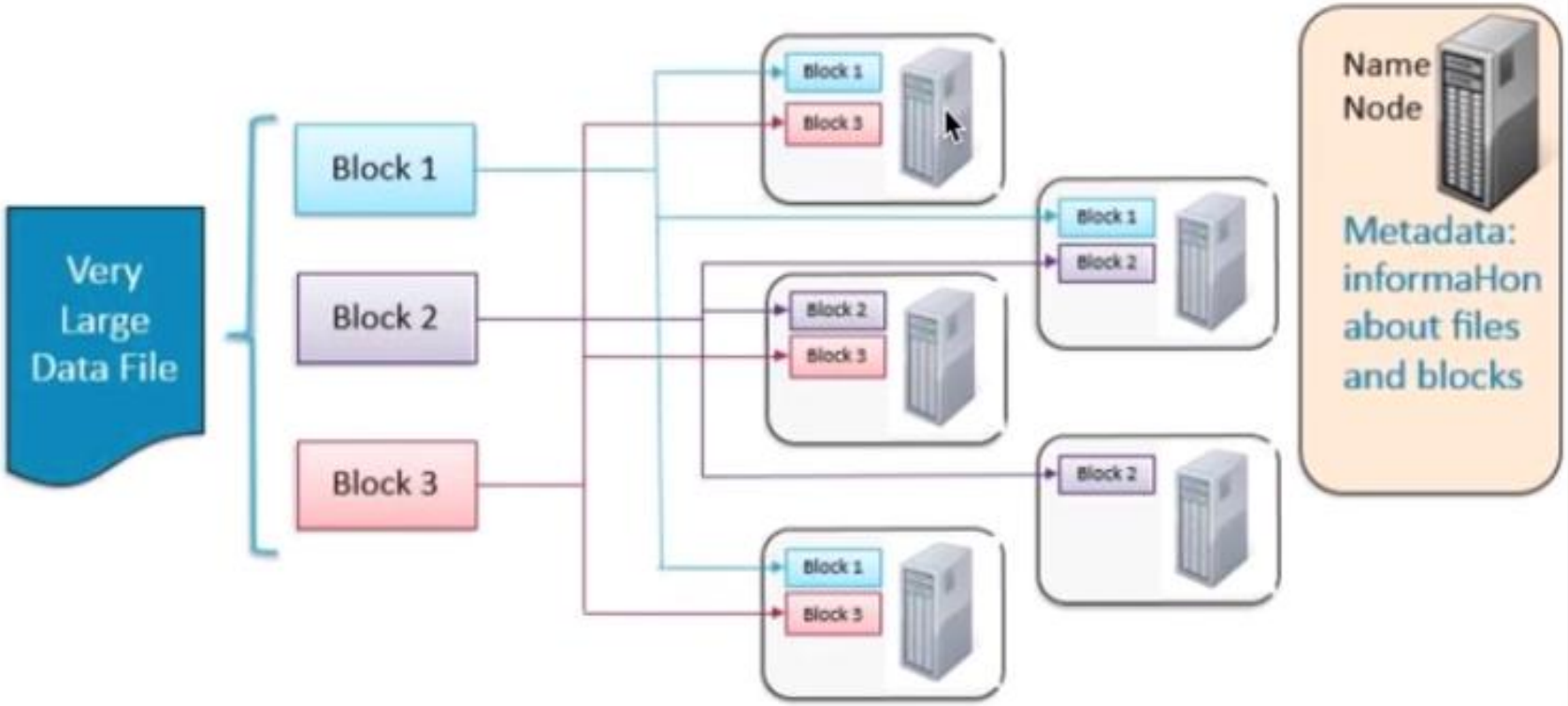
Hadoop


Sqoop
Mahout
Hive
Pig
HBase
Flume
Oozie
...
Hadoop Ecosystem
hadoop MapReduce
MapReduce
hadoop HDFS
Hadoop Distributed File System
Hadoop Core Components

Core Components: HDFS and MapReduce

- **HDFS (Hadoop Distributed File System)**
  - Stores data on the cluster
- **MapReduce**
  - Processes data on the cluster

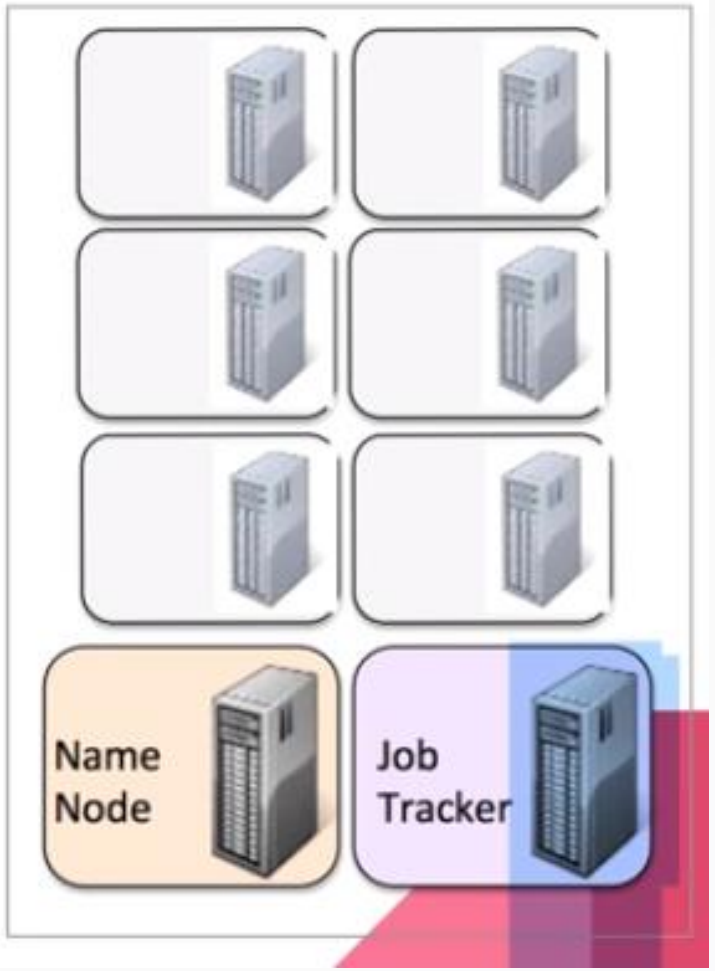
- Data files are split into blocks and distributed at load time
- Each block is replicated on multiple data nodes (default 3x)
- NameNode stores metadata



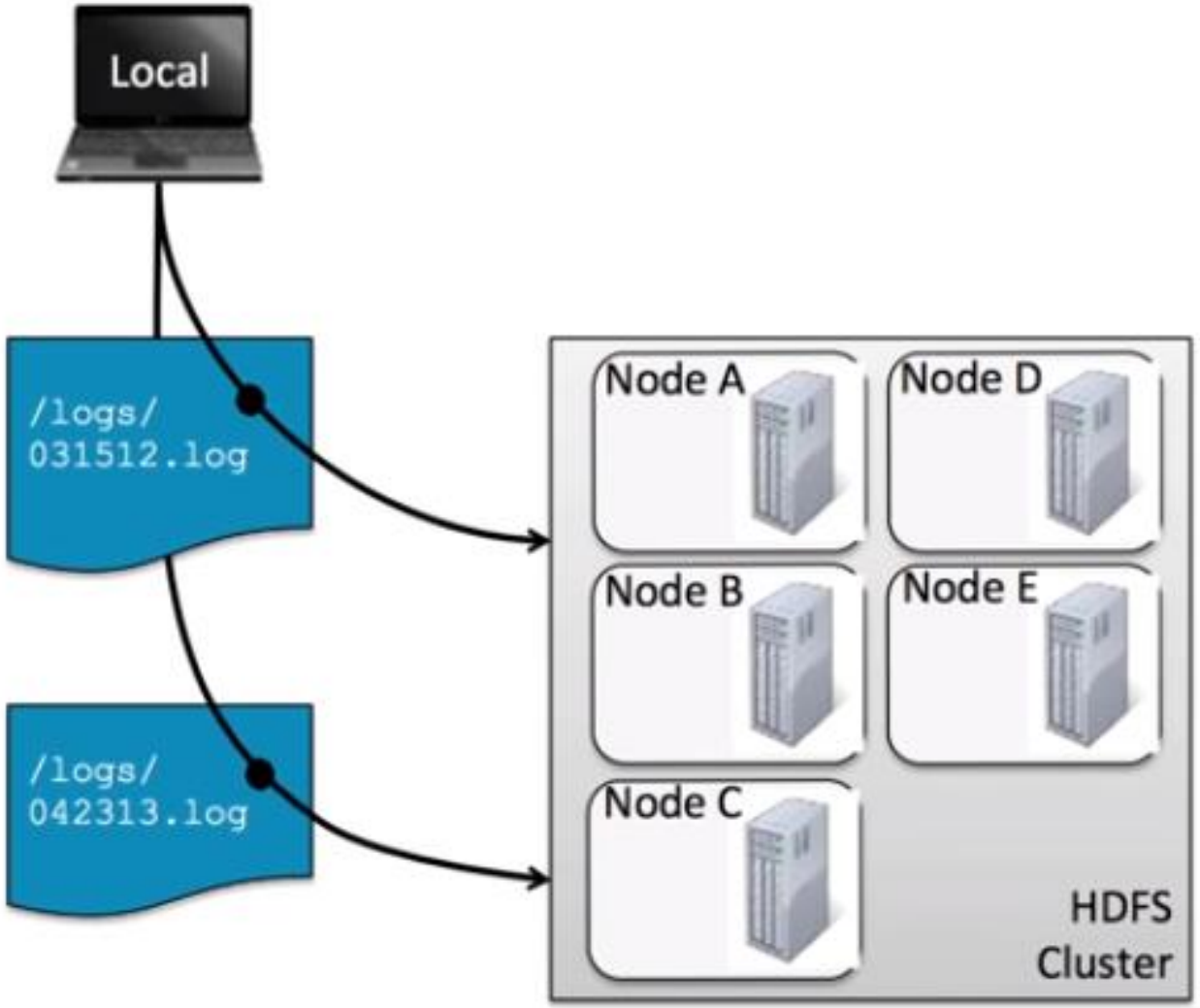
A Simple Hadoop Cluster

- **A Hadoop cluster: a group of machines working together to store and process data**

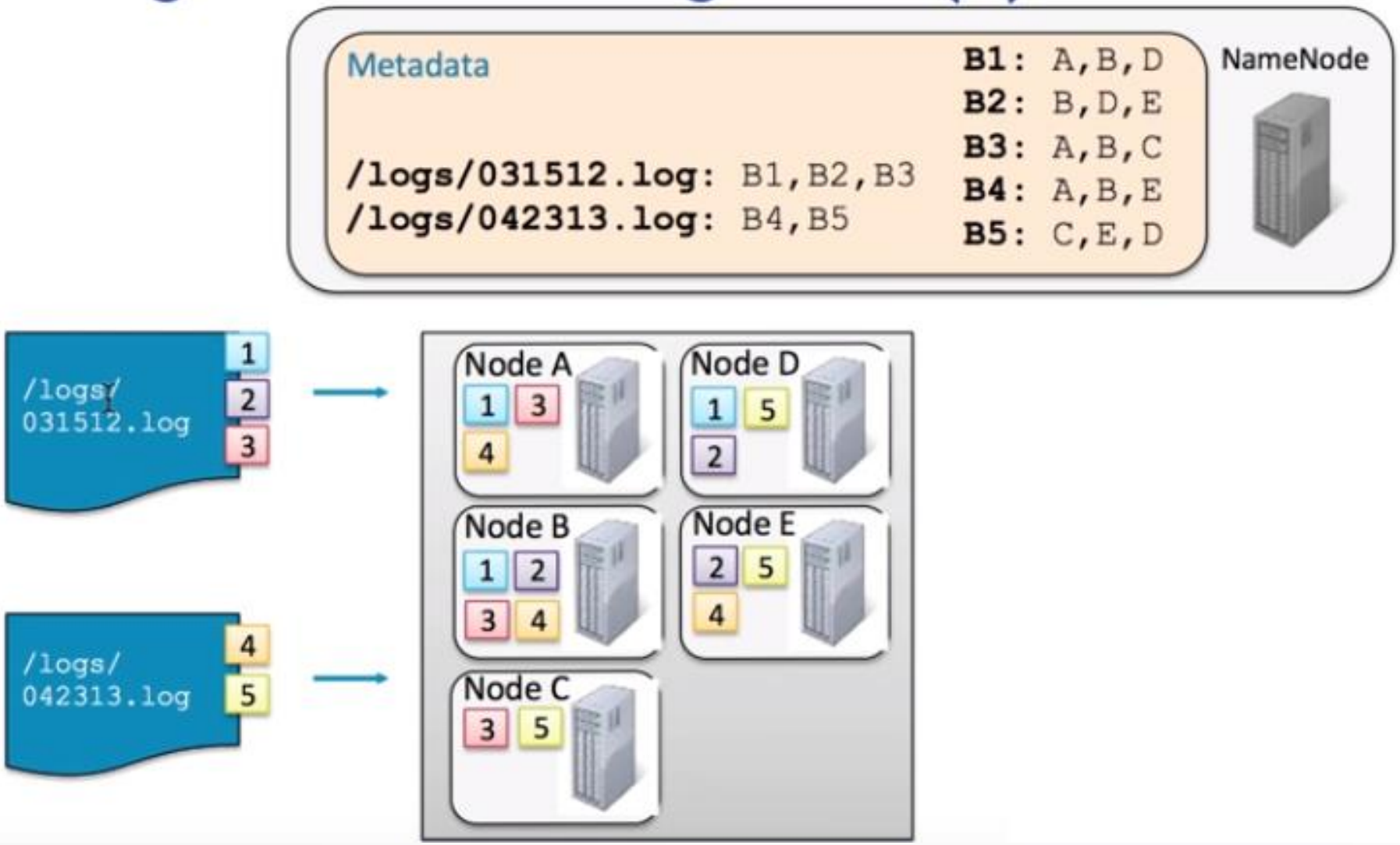
- **Any number of 'slave' or 'worker' nodes:**
  HDFS to store data
  MapReduce to process data



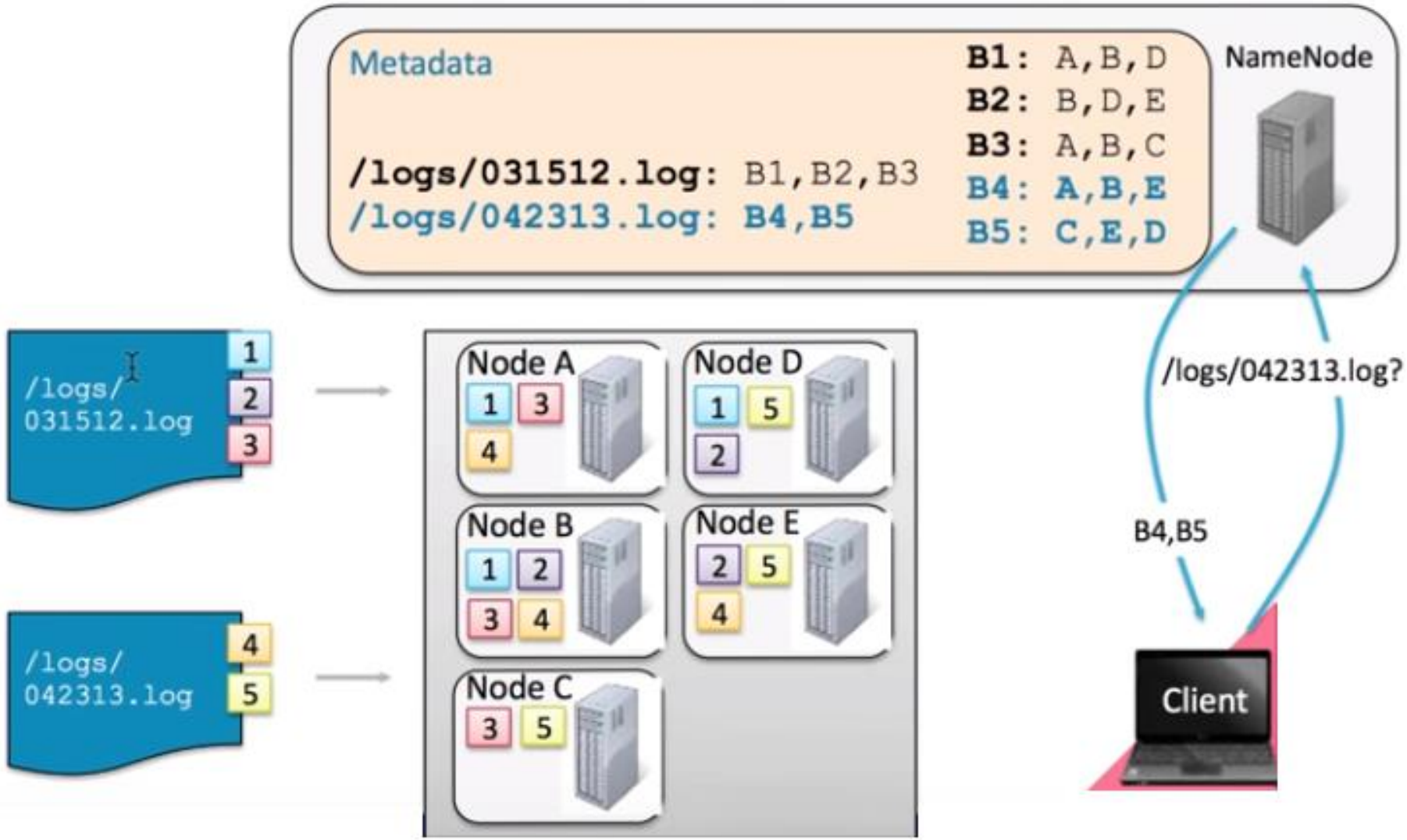
Example: Storing and Retrieving Files (1)



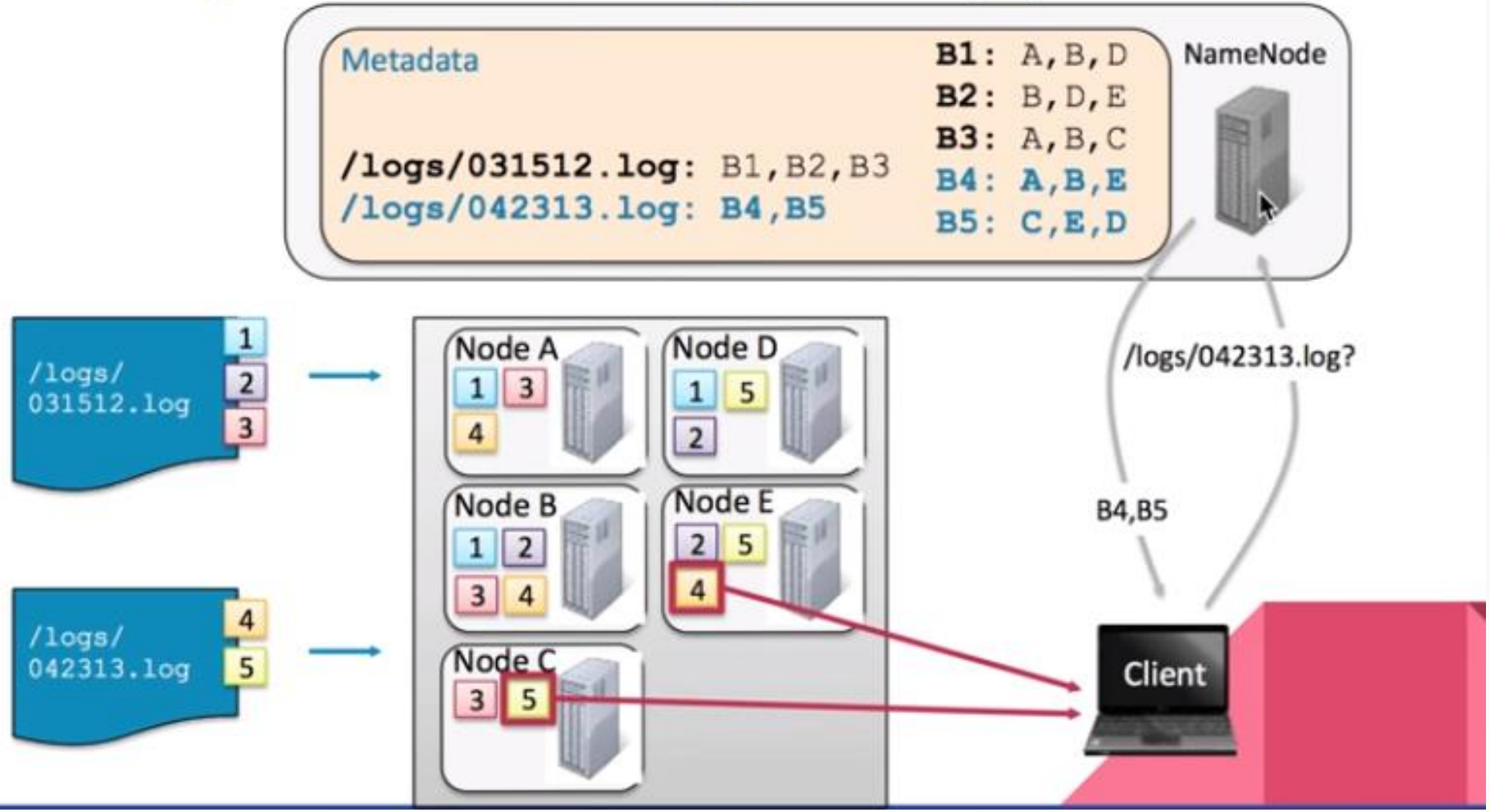
Ex: Storing and Retrieving Files (2)



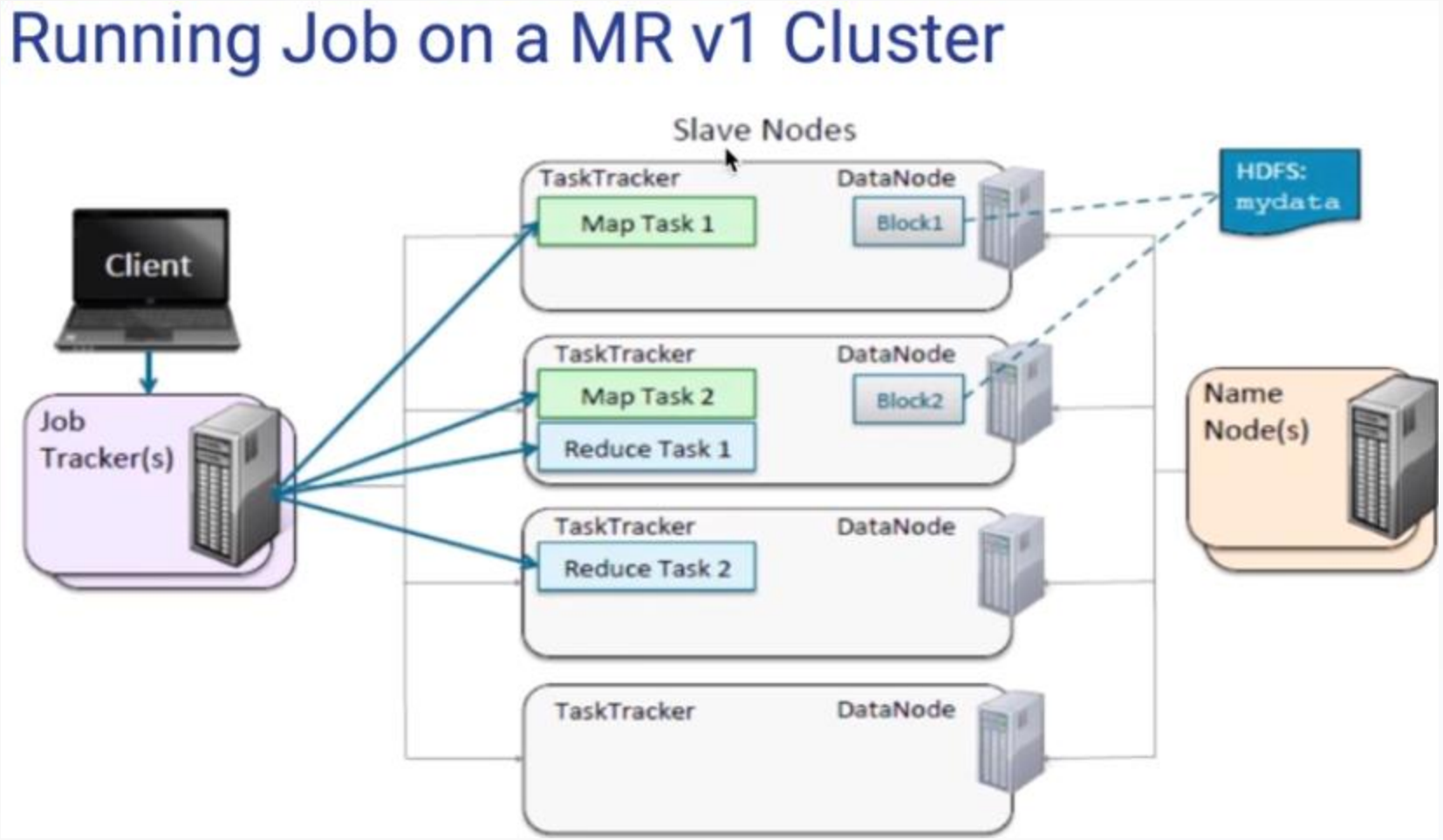
Ex: Storing and Retrieving Files (3)



Ex: Storing and Retrieving Files (4)



Running Job on a MR v1 Cluster



Comparison To Hadoop

MapReduce

YARN

HDFS

Hadoop: combined compute, resource management + storage

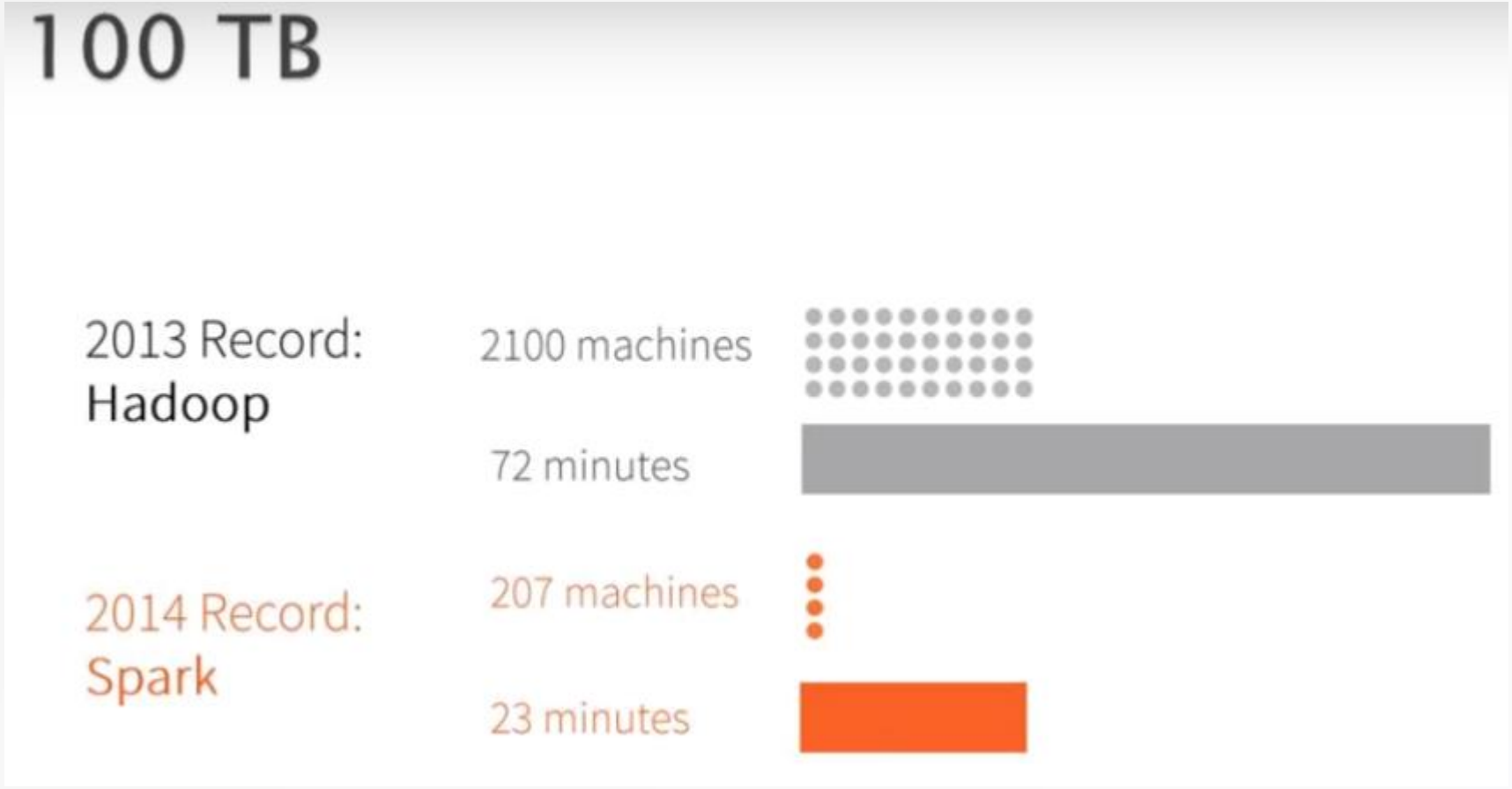
Spark

Spark

YARN | Standalone | Mesos | ..

HDFS | NoSQL | SQL | ..

Spark: independent of resource management + storage layers


100 TB
2013 Record:
Hadoop
2100 machines
72 minutes
2014 Record:
Spark
207 machines
23 minutes

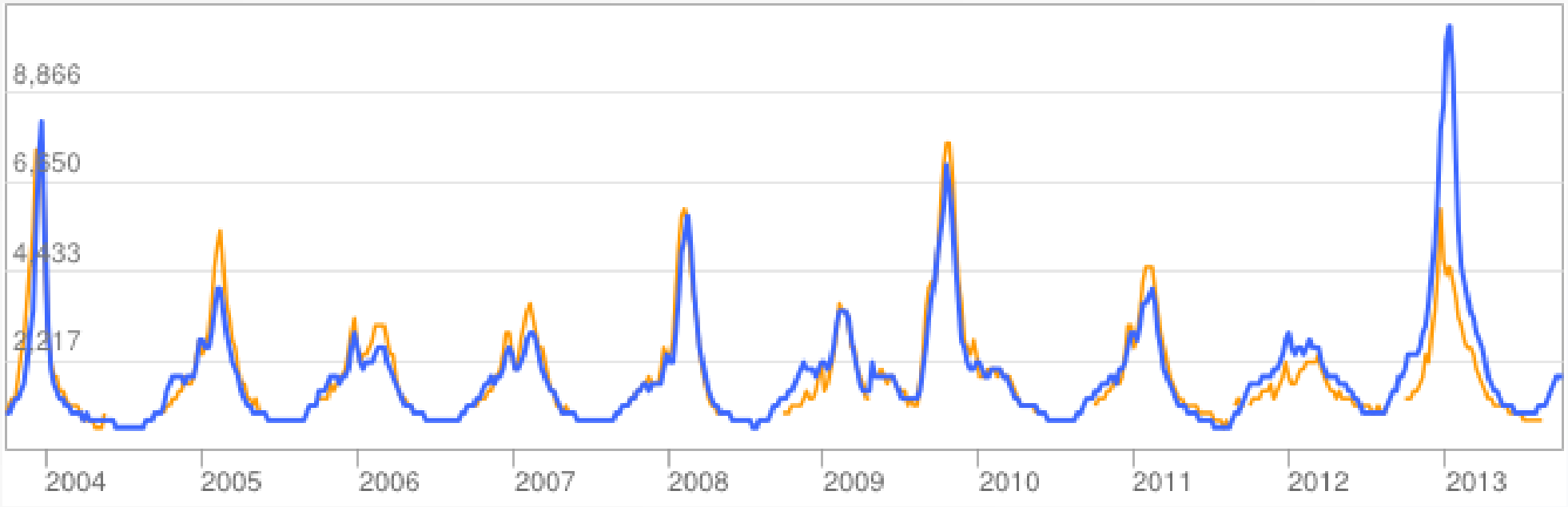
Các ứng dụng

- Tiếp thị
- Phân tích tội phạm
- Dự đoán dịch bệnh
- Hiển thị dựa trên trải nghiệm người dùng

Dự liệu lớn với tiếp thị

- Các nhà nghiên cứu tiếp thị tin rằng việc phân tích dữ liệu lớn sẽ đem lại cơ hội có một không hai cho doanh nghiệp để hiểu khách hàng và có những chiến lược tiếp thị hiệu quả.

Dự báo dịch bệnh



Đường màu xanh là dự đoán của Google Flu Trends dựa trên số từ khóa tìm kiếm liên quan đến các dịch cúm, màu vàng là dữ liệu do cơ quan phòng chống dịch của Mỹ đưa ra.

Các vấn đề mở

- Vấn đề bảo mật
- Tốc độ/tính chuyển động liên tục của dữ liệu
- Tính chính xác và tin cậy của dữ liệu

Tính bảo mật

- Trong các dữ liệu có được từ các phương tiện truyền thông hay mạng xã hội các thông tin cá nhân của nhiều người thường có liên quan đến nhau và dễ dàng bị “đào xới” bởi các ứng dụng khai phá dữ liệu.
- Thách thức với những người làm quản lý và các nhà nghiên cứu là vừa phải có những chính sách đúng đắn và phương pháp tiếp cận để quản lý việc chia sẻ dữ liệu cá nhân trong khi vẫn tạo điều kiện cho các hoạt động khai phá dữ liệu hợp pháp.

Tốc độ/tính chuyển động liên tục của dữ liệu

- Các kĩ thuật khai phá dữ liệu lớn phải có khả năng truy cập dữ liệu nhanh chóng.
- Hệ thống xử lý dữ liệu phải hoàn thành việc xử lý/khai phá dòng dữ liệu đó trong một thời gian nhất định.
- Tốc độ khai phá dữ liệu phụ thuộc vào hai yếu tố chính: thời gian truy cập dữ liệu (được xác định chủ yếu bởi hệ thống lưu trữ dữ liệu) và hiệu quả của các thuật toán khai phá dữ liệu.

Tính chính xác và tin cậy

- Dữ liệu có thể đến từ nhiều nguồn khác nhau, có thể từ nguồn không tin cậy và không thể kiểm chứng.
- Do dữ liệu lớn có tính động (dynamic) cao nên hệ thống phân tích và quản lý dữ liệu lớn cũng phải cho phép các dữ liệu được quản lý trong đó thay đổi và phát triển.
- Khi dữ liệu có sự thay đổi, phát triển thì các độ đo độ tin cậy cần được thay đổi hoặc cập nhật.