Big Data

GV: TS Võ Đình Hiếu

Thực hiện:

Phạm Công Thiên Lý

Dương Bà Cường

Nguyễn Khắc Chung

Đinh Anh Thái

Nội dung

- Giới thiệu Big Data
- Các thành phần Big Data
- Tổ chức lưu trữ dữ liệu BigData
- Giải pháp Big data của Oracle

Giới thiệu BIG DATA

BIG DATA ?

- Là những số lượng khủng về các hồ sơ khách hàng, âm thanh, hình ảnh, văn bản…

BIG DATA ?

- Dữ liệu có số lượng lớn cần được lưu trữ như
  - Truyền thống: thông tin khách hàng, giao dịch…
  - Thu thập tự động qua cảm biến: thời tiết, nhật ký…
  - Mạng xã hội: comment trên facebook, twitter…
- Đặc trưng
  - Số lượng
  - Tốc độ
  - Đa dạng
  - Giá trị

Big Data


Variety
Velocity
Volume
Structured & Unstructured
Batch
Structured
Big Data
Streaming Data
Zettabytes
Terabytes

Dung lượng

- Nhu cầu lưu trữ ngày càng tăng
  - 2000: 800000 (PB) lưu trữ trên thế giới(*)
  - 2020: 35 ZB trên toàn thế giới?(*)

➔Làm thế nào để quản lý?

- Dữ liệu càng lớn thì:
  - Khả năng xử lý giảm?
  - Phân tích dữ liệu giảm
  - Truy xuất chậm

(*)Số liệu từ IBM
1ZB = 10^{21} bytes
1PB = 10^{15} bytes

Đa dạng

- Dữ liệu đến từ nhiều nguồn:
  - Cảm biến
  - Smart device
  - Mạng xã hội
  - Tin tức
  - …
- Dữ liệu phức tạp
  - Truyền thống và không truyền thống
  - Có cấu trúc, bán cấu trúc, không cấu trúc…

Tốc độ

- Khối lượng dữ liệu là rất lớn

  →tốc độ truy xuất chậm
- Yêu cầu từ người sử dụng:
  - Nhanh
  - Ổn định
  - Chính xác

Tầm quan trọng Big Data

- Mang tới sự hiểu biết sâu sắc hơn cho doanh nghiệp
- Là sự tồn tại của doanh nghiệp
- Mang tới sự hiểu biết mới

Các thành phần Big Data

Các thành phần

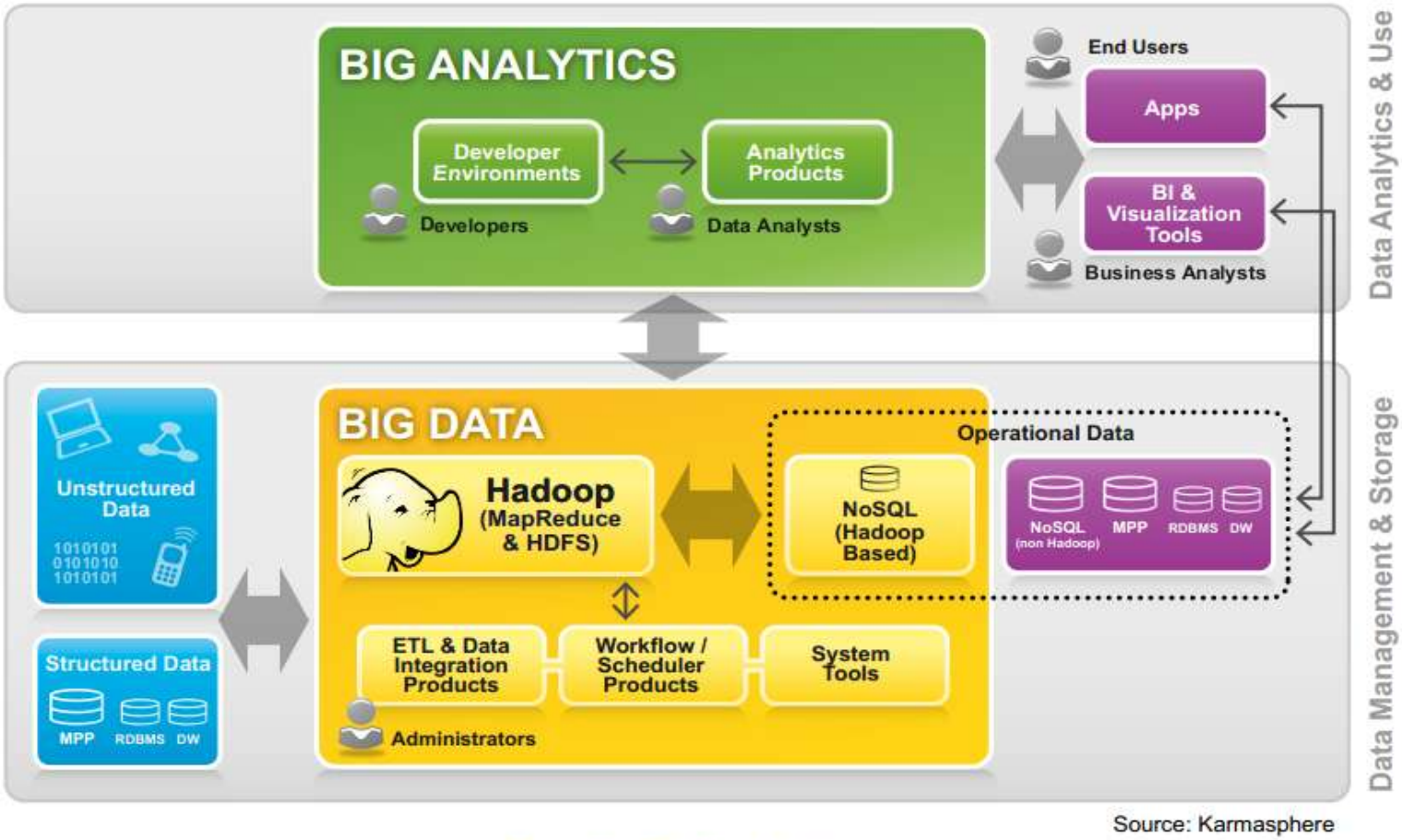


Figure 5 – Big Data Architecture

Các thành phần

- Quản lý dữ liệu: cơ sở hạ tầng lưu trữ dữ liệu, và nguồn để thao tác nó.
- Phân tích dữ liệu: công nghệ và các công cụ để phân tích các dữ liệu và thu thập hiểu biết sâu sắc từ nó
- Sử dụng dữ liệu: đưa dữ liệu lớn đã phân tích để phục vụ trong Kinh doanh thông minh và các ứng dụng của người dùng cuối

Quản lý dữ liệu

- Hệ dữ liệu có cấu trúc
  - Hệ thống quản lý cơ sở dữ liệu quan hệ(RDBMS): để lưu trữ và thao tác dữ liệu có cấu trúc.
  - Hệ thống MPP: tập hợp dữ liệu đồ sộ ngày càng lớn thêm và tăng cường dữ liệu tăng trưởng.
  - Kho dữ liệu: tập hợp và lưu trữ dữ liệu cho các báo cáo sau này.
  - Hạn chế
    - Khó mở rộng, hiệu suất chậm lại.
    - Biểu diễn dữ liệu

Quản lý dữ liệu

- Hệ dữ liệu không cấu trúc: phù hợp cho việc lưu trữ dữ liệu có cấu trúc phức tạp và dễ dàng mở rộng
- Dữ liệu
  - Dữ liệu có cấu trúc và không có cấu trúc
  - Lấy từ nhiều nguồn với kích cỡ khác nhau
  - Dữ liệu thường rất lớn, yêu cầu tốc độ xử lý cao

  → Yêu cầu tổ chức dữ liệu để đáp ứng: Apache Hadoop

Phân tích dữ liệu

- Là nơi mà các công ty bắt đầu trích xuất giá trị dữ liệu lớn.
- Liên quan tới việc phát triển các ứng dụng và sử dụng các ứng dụng để đạt được cái nhìn sâu sắc vào dữ liệu lớn.
- Xây dựng các tool phân tích dữ liệu

Sử dụng dữ liệu

- Là các hoạt đông trên dữ liệu được phân tích

Tổ chức lưu trữ dữ liệu BigData

Hadoop

- Giới thiệu về Hadoop
- Các thành phần của Hadoop
- HDFS (Hadoop Distributed file System)

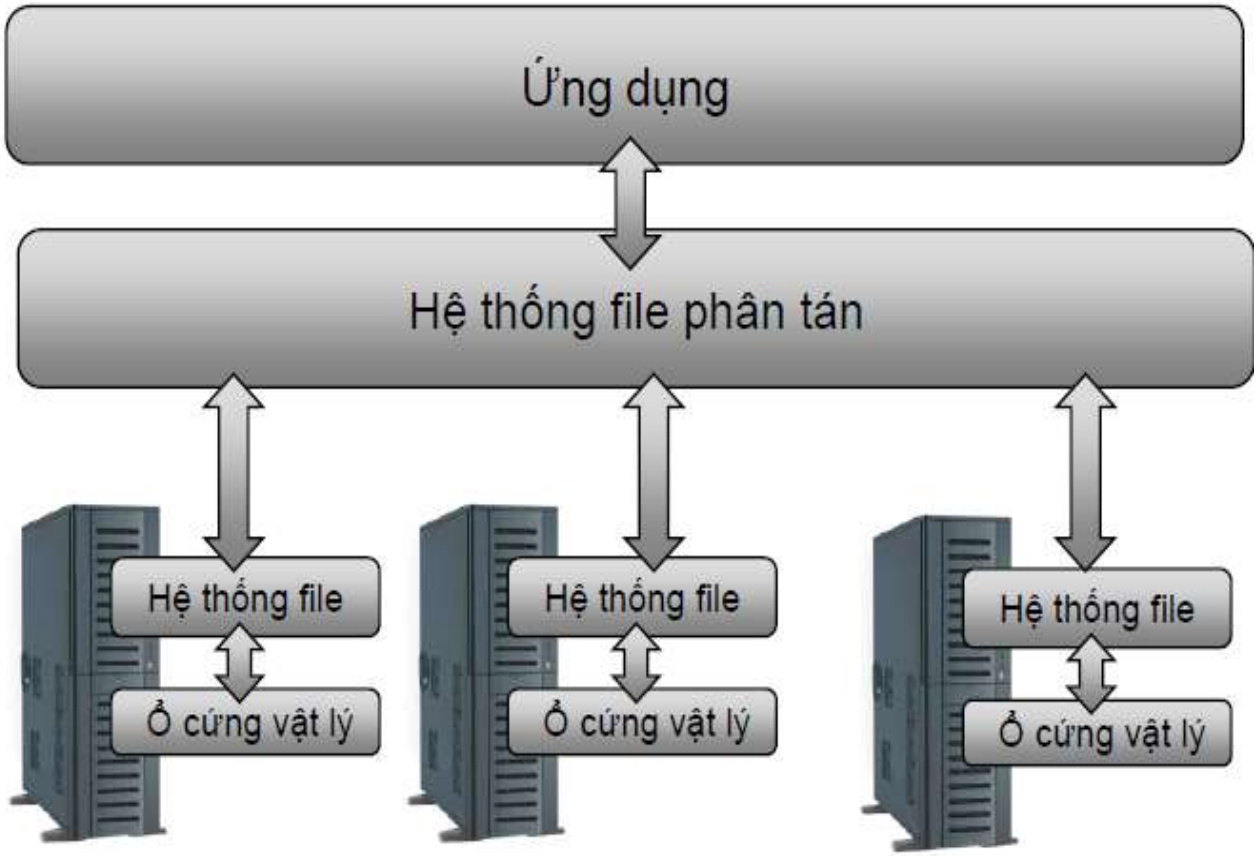
Hadoop là gì?

- Một nền tảng ứng dụng hỗ trợ các ứng dụng phân tán với dữ liệu rất lớn
  - Hàng terabyte
  - Hàng ngàn node
- Cung cấp phương tiện lưu trữ dữ liệu trên nhiều node, hỗ trợ tối ưu hóa lưu lượng mạng.

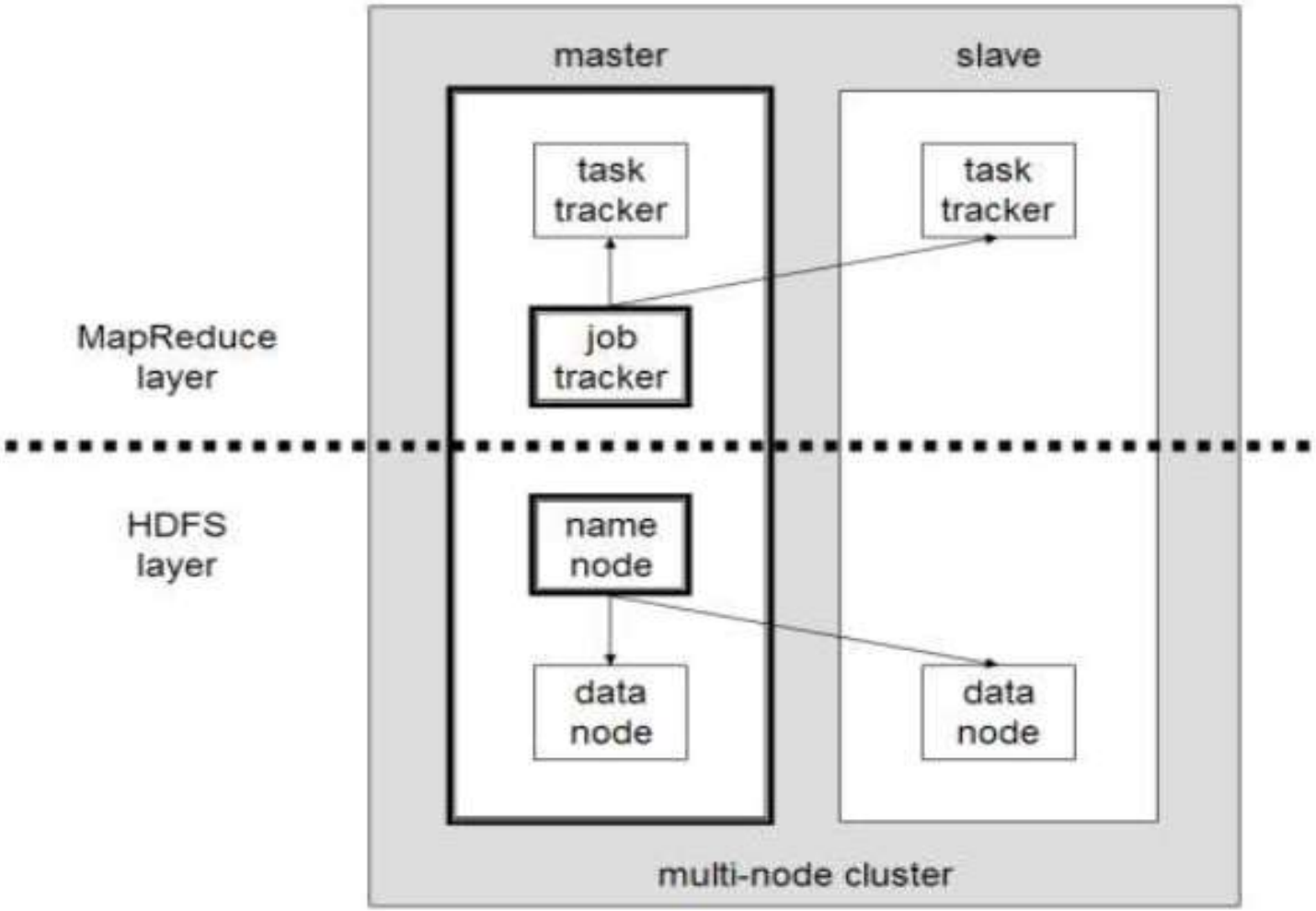
Thành phần của Hadoop

- Xử lý (MapReduce): một framework giúp phát triển các ứng dụng phân tán theo mô hình MapReduce một cách dễ dàng và mạnh mẽ.
- Lưu trữ (HDFS): 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.

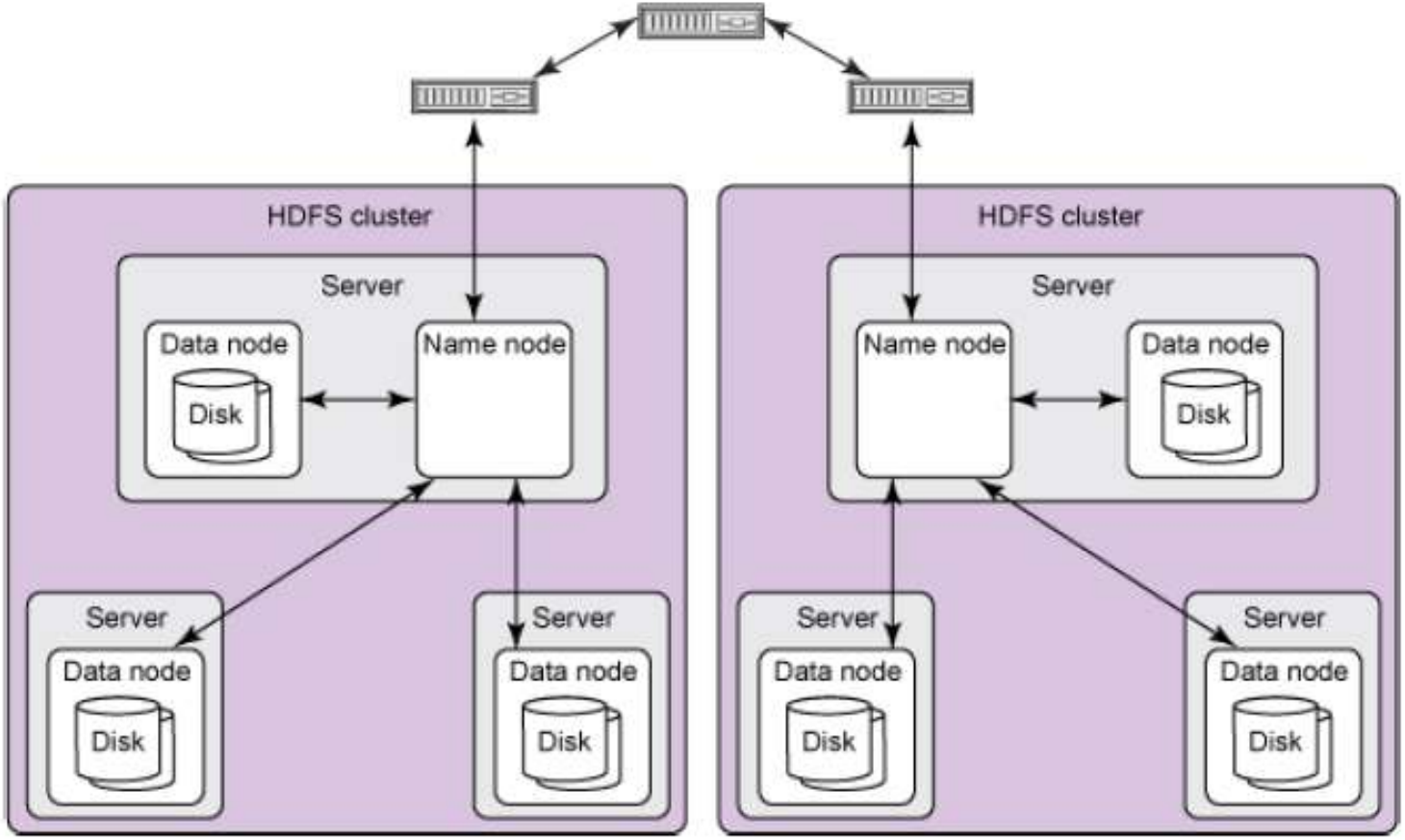
Hadoop Distributed file System



Hadoop Distributed file System



Kiến trúc của HDFS



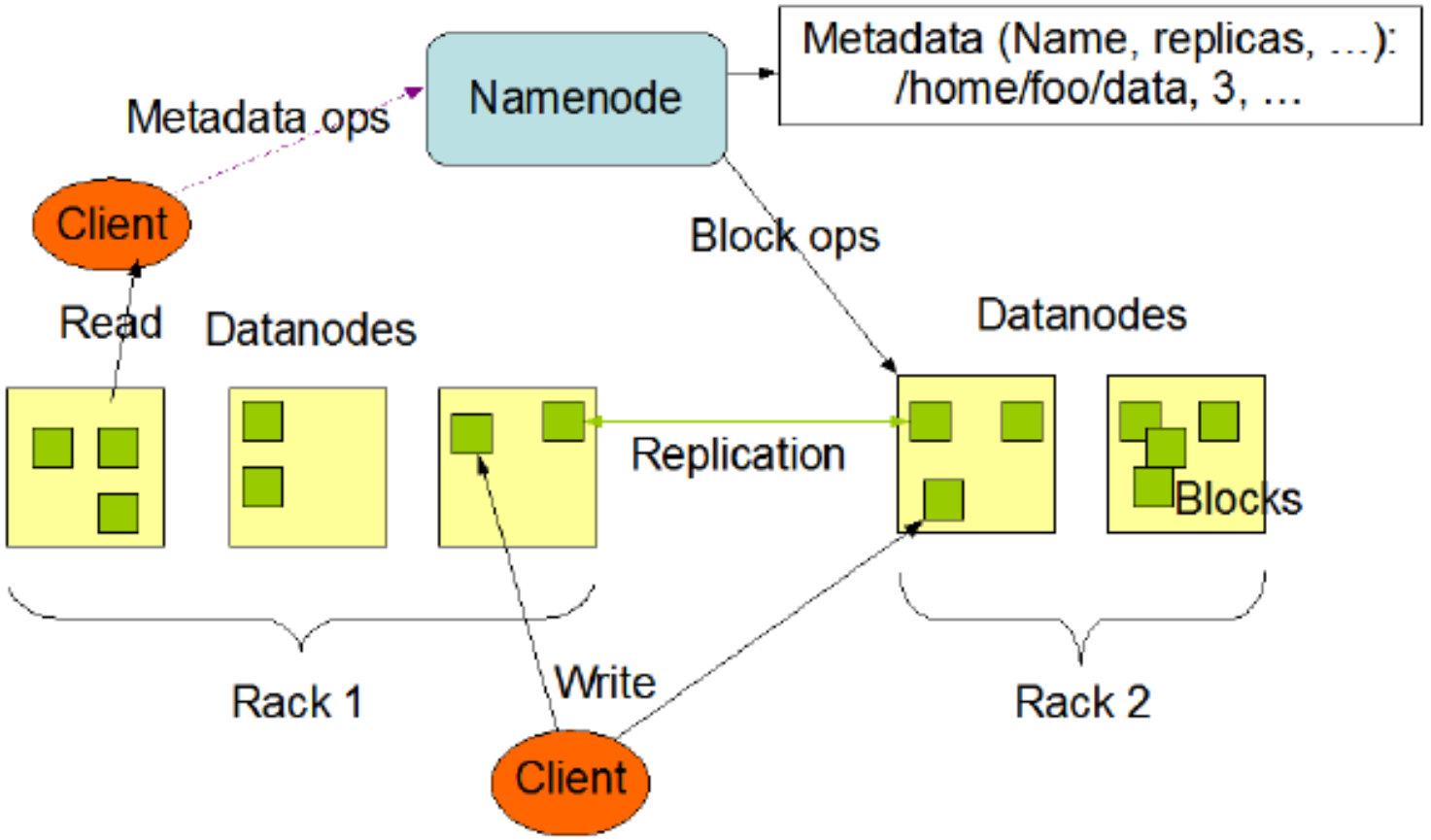
Kiến trúc của HDFS

- Name node: Đóng vai trò là master của hệ thống HDFS, quản lý thông tin các file, block id tương ứng cho từng file
- Block: đơn vị lưu trữ dữ liệu nhỏ nhất
  - Hadoop dùng mặc định 64MB/block
  - Một file chia làm nhiều block
  - Các block chứa ở bất kỳ node nào trong cluster
- DataNode: Chứa các block

Kiến trúc của HDFS

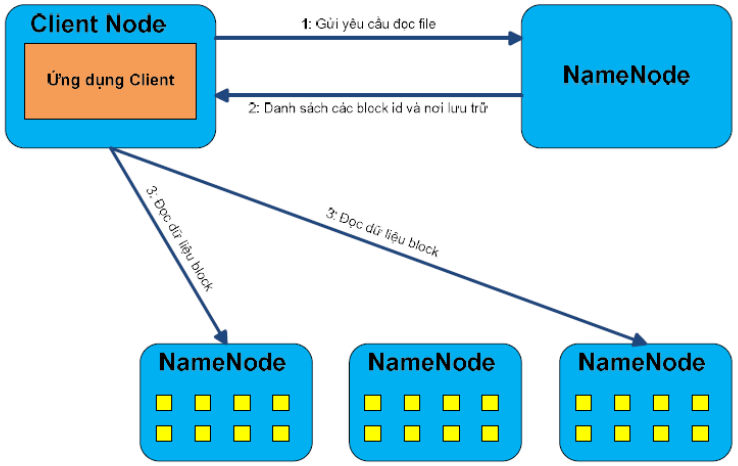
- JobTracker: tiếp nhận các yêu cầu thực thi các MapReduce job.
  - Phân chia job và giao task cho task tracker
  - Quản lý tình trạng của từng node
- TaskTracker:
  - Nhận các task từ jobTracker và thực hiện task

Cơ chế hoạt động HDFS



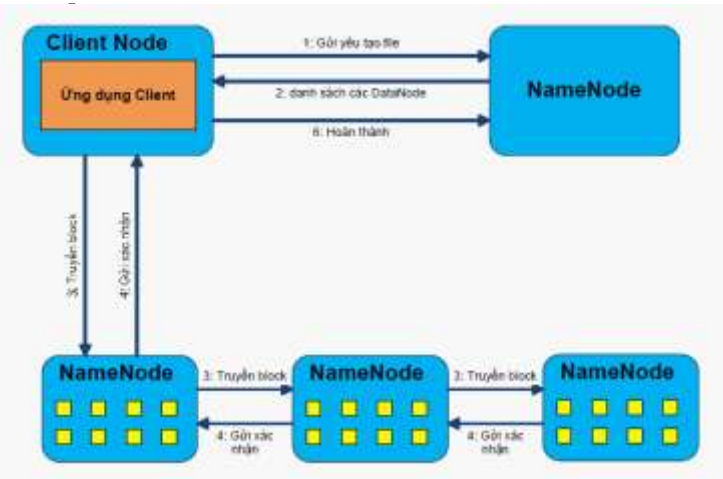
Cơ chế hoạt động HDFS

- Đọc
  - client yêu cầu đọc dữ liệu từ Name Node, namenode trả về vị trí các block của dữ liệu
  - Chương trình trực tiếp yêu cầu dữ liệu tại các node



Cơ chế hoạt động HDFS

- Ghi
  - Ghi theo dạng đường ống (pipeline)
  - client yêu cầu thao tác ghi ở Name Node
  - Namenode kiểm tra quyền ghi và đảm bảo file không tồn tại
  - Các bản sao của block tạo thành đường ống để dữ liệu tuần tự được ghi



Hadoop Distributed file System

- Ưu điểm
  - Lưu trữ được lượng file rất lớn
  - Truy cập dữ liệu theo dòng
  - Liên kết dữ liệu đơn giản
  - Phần cứng phổ thông, đa dạng
  - Tự động phát hiện lỗi, phục hồi dữ liệu nhanh
- Nhược điểm
  - Có độ trễ truy cập
  - Không thể lưu trữ quá nhiều file trên cùng 1 cluster

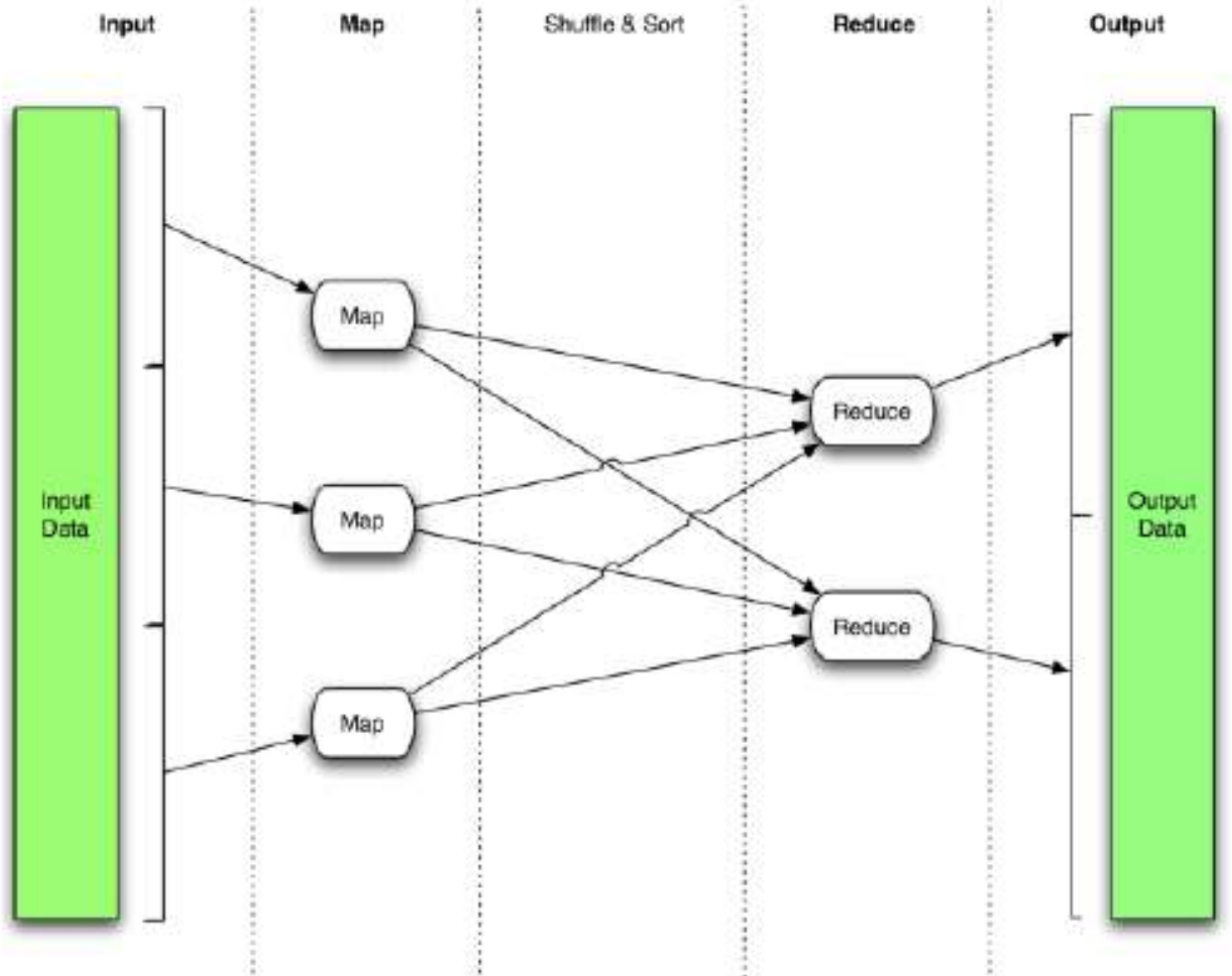
Hadoop Common

- Tập hợp các thư viện hỗ trợ cho Hadoop
- Bao gồm tập các lệnh
  - *Cat* copy file tới bộ ra chuẩn(stdout)
  - *Chmod* chuyển quyền đọc và ghi cho một file
  - *Chown* chuyển quyền sở hữu của một file hoặc 1 tập hợp file
  - *……*

MapReduce

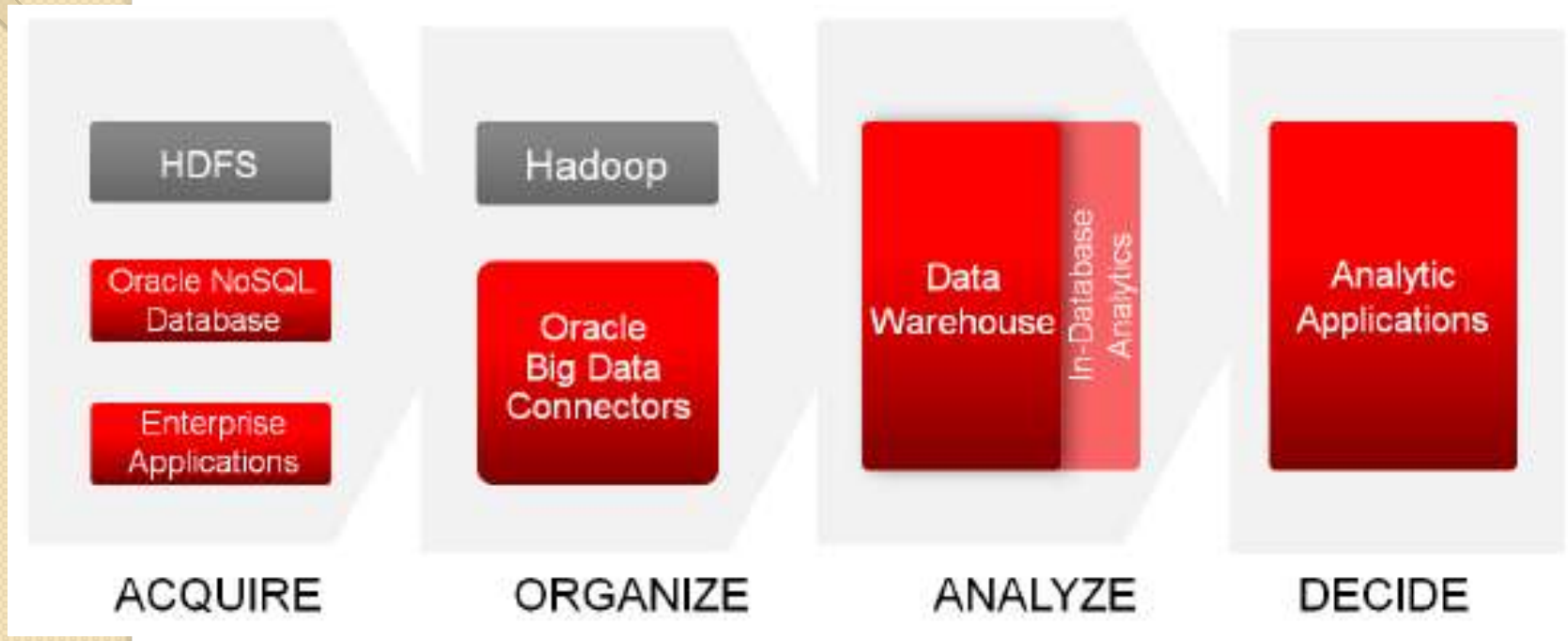
- Quản lý tiến trình song song, phân tán, sắp xếp lịch trình I/O
- *Quản lý trạng thái dữ liệu*
- *Quản lý số lượng lớn dữ liệu có quan hệ phụ thuộc nhau*
- *Xử lý lỗi*
- *Trừu tượng hóa với lập trình viên*

MapReduce


Input
Map
Shuffle & Sort
Reduce
Output
Input Data
Map
Map
Map
Reduce
Reduce
Output Data

Oracle Big Data

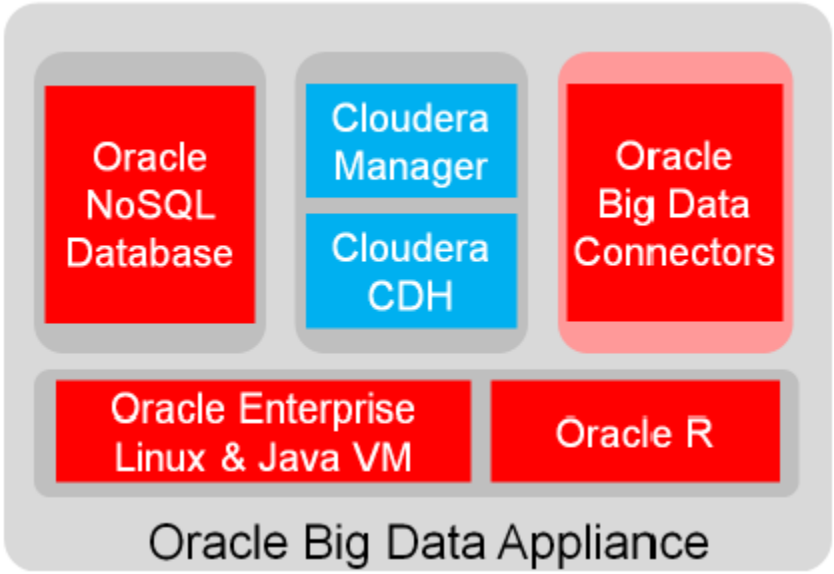
Tổng quan


HDFS
Oracle NoSQL Database
Enterprise Applications
Hadoop
Oracle Big Data Connectors
Data Warehouse
In-Database Analytics
Analytic Applications
ACQUIRE
ORGANIZE
ANALYZE
DECIDE

Oracle Big data

- Là sự kết hợp cả phần cứng và phần mềm
- Phần cứng:
  - 18 server Sun
  - Dung lượng 648TB
  - 2CPU/server, 6 nhân/CPU → 216 nhân
  - 48GB RAM

Oracle Big data



- Phần mềm
  - Bản đầy đủ của Cloudera's Distribution(bao gồm cả Apache Hadoop) (CDH)
  - Cloudera manager: để quản trị Cloudera CDH
  - Gói R là một mã nguồn mở cho việc phân tích dữ liệu chưa được xử lý trên Oracle Big Data
  - Oracle NoSQL database
  - Hệ điều hành Oracle Enterprise Linux cùng với Oracle Java VM

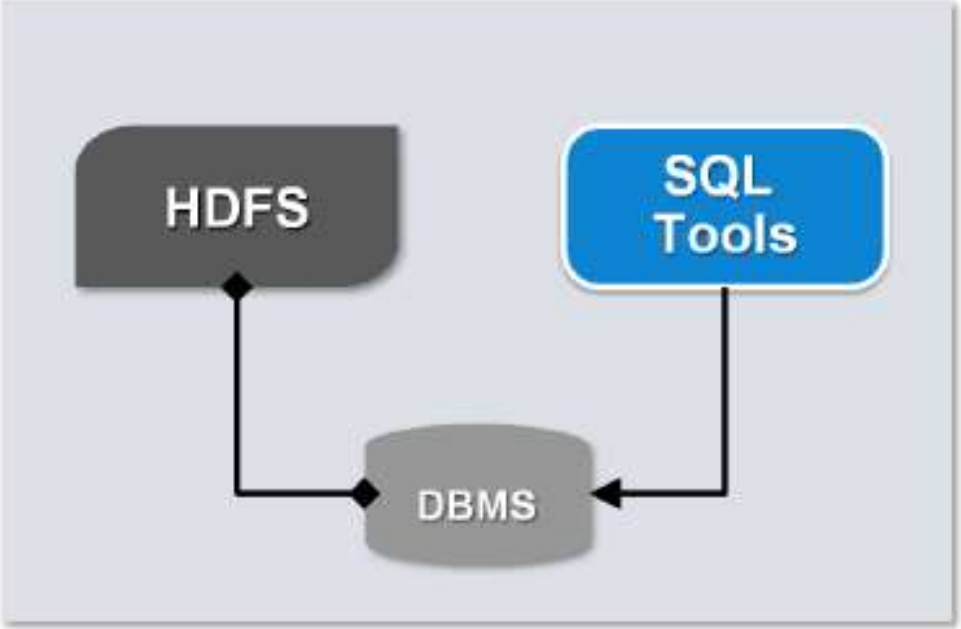
Oracle Big data

- Các thành phần chính
  - CDH và Cloudera Manager
  - Oracle Big data connectors
  - Oracle Loader cho Hadoop
  - Oracle Direct Connector for Hadoop Distributed file system
  - Oracle data intergator application adapter cho Hadoop
  - Oracle R connector for Hadoop
  - Oracle NoSQL database

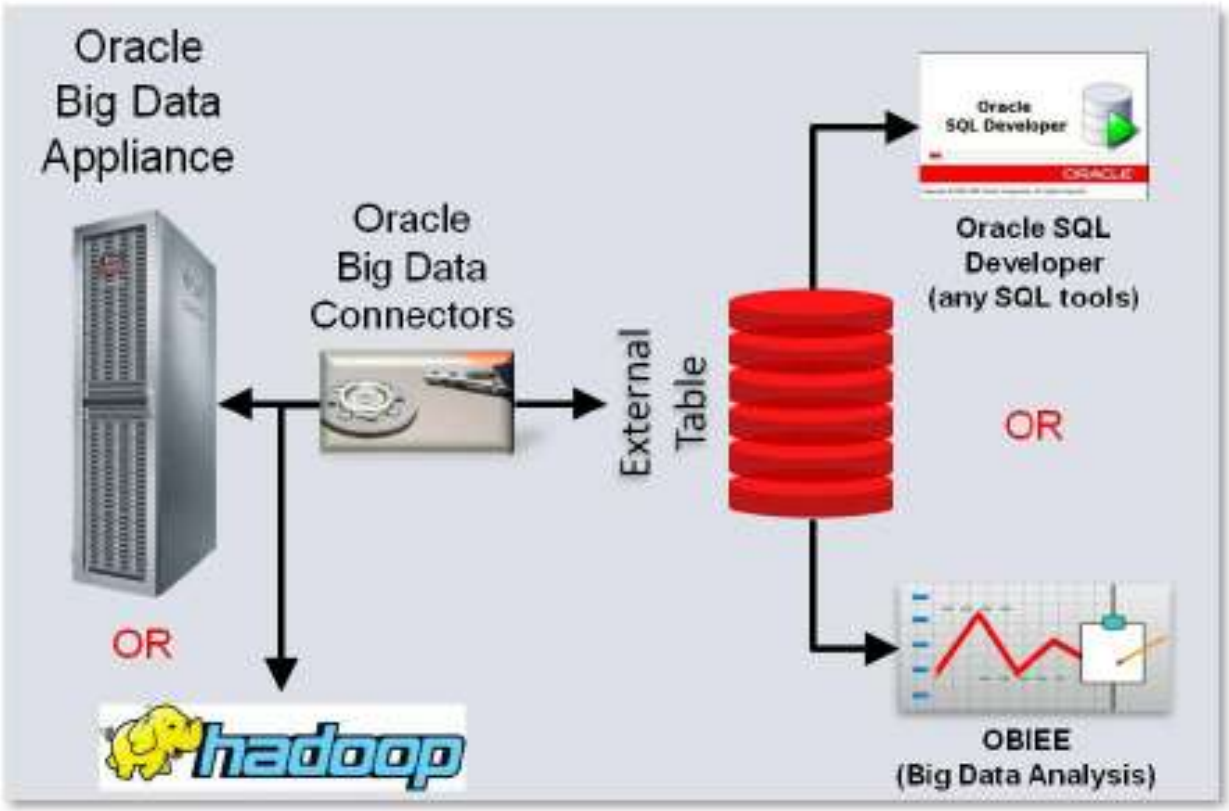
Phân tích dữ liệu

- Ví dụ:
  - Hệ thống bán hàng online

→ các đối tượng được xác định rõ ràng

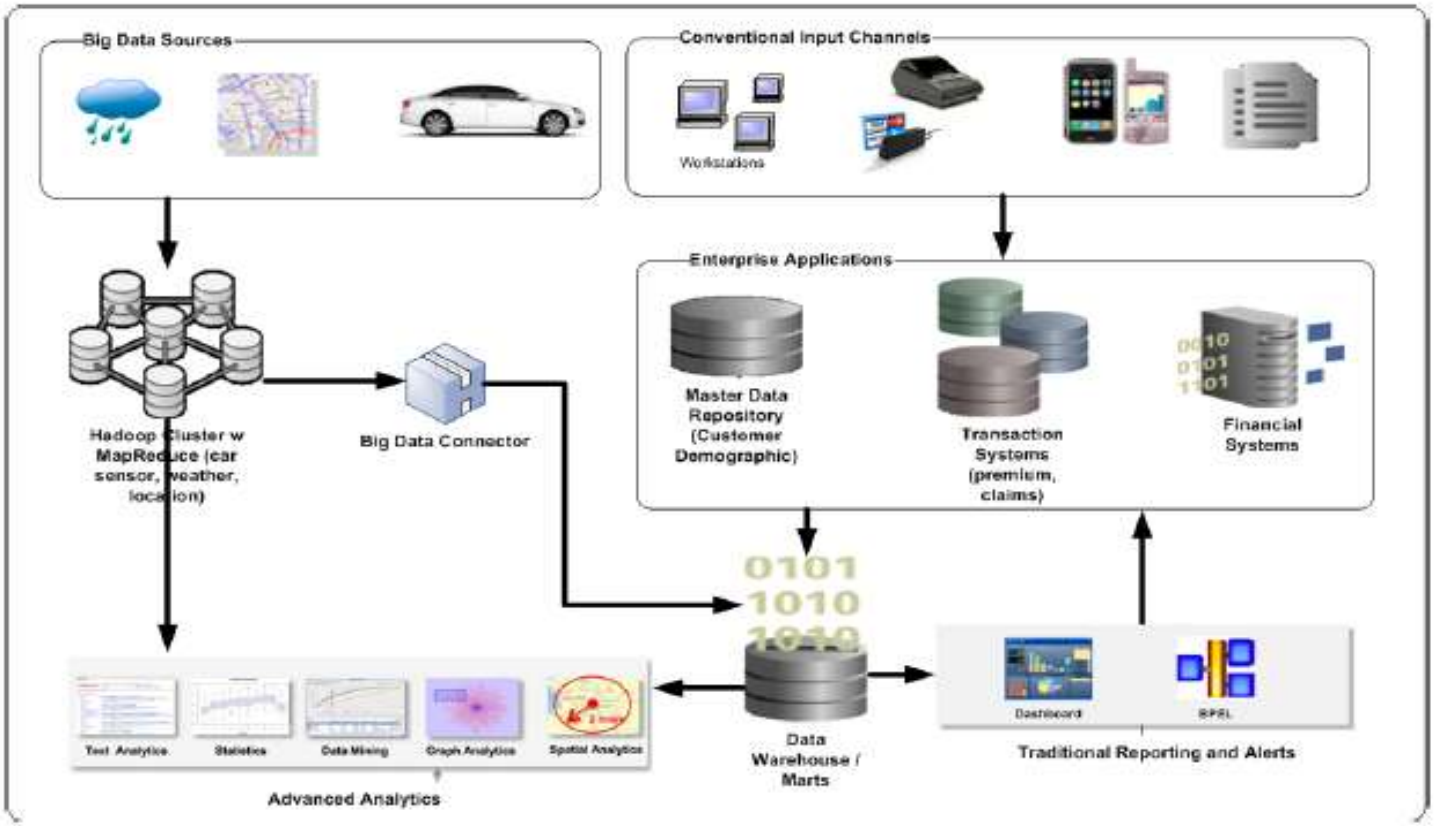


Phân tích dữ liệu

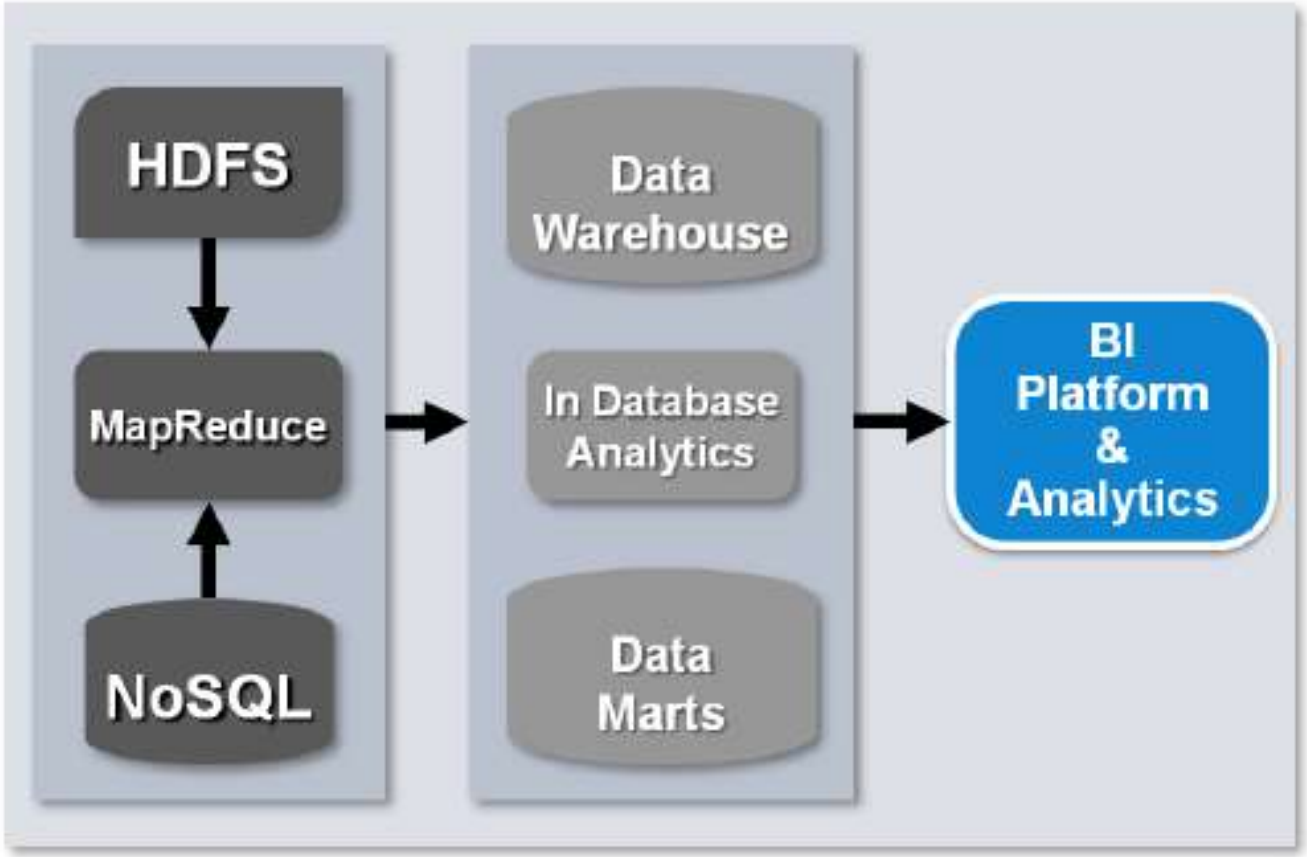


Phân tích dữ liệu

- Ví dụ:
  - Dữ liệu được thu thập từ nhiều nguồn, ko có cấu trúc



Phân tích dữ liệu



Tài liệu tham khảo

- Big-Data Computing: Creating revolutionary breakthroughs in commerce, science, and society (Randal E. Bryant Carnegie Mellon University, Randy H. Katz University of California, Berkeley, Edward D. Lazowska University of Washington)
- Understanding the Elements of Big Data: More than a Hadoop Distribution(Martin Hall, Founder, Karmasphere)
- Big Data The power and possibilities of Big Data
- Basic Data Analysis Tutorial
- Oracle: Big Data for the enterprise