

① Big data

- big data là tập hợp dữ liệu có liên quan đến tài sản số ra ~~men~~ nhanh và các tour du lịch trên mạng xã hội
- cu quàn lý trạng - thay đổi - hệ thống và xử lý hiện giờ
- Điện thoại (3V nó tăng lên 5V)

- Volume (thể tích)
- Velocity (tốc độ)
- Variety (đa dạng)
- Veracity (độ tin cậy)
- Value (giá trị)

VD: dữ liệu từ Facebook, camera giao hàng

② Virtualization

- Công nghệ cho phép tạo ra phiên bản ảo của tài nguyên phần cứng như máy chủ, hệ điều hành, hay mạng
- Điện thoại
- tách biệt lớp vật lý và lớp logic
- Cho phép nhả hệ điều hành chạy ngay trên ứng dụng
- này tạo ra môi trường ảo

VD: VMware, Virtual Box, Hyper-V

(3) cloud

Mô hình ứng dụng tài nguyên tính toán (nguồn dữ liệu) đã lưu trữ và xử lý qua internet

- Đặc điểm:

+ theo yêu cầu

+ khả năng mở rộng

+ trả phí theo mức sử dụng

+ truy cập mọi nơi

- Mô hình dịch vụ

+ IaaS : hạ tầng

+ PaaS : nền tảng

+ SaaS : phần mềm

(4)

Data warehouse

là hệ thống tập trung dữ liệu từ nhiều nguồn khác nhau để phân tích và hỗ trợ quyết định

- Đặc điểm

+ Tích hợp

+ Cập nhật

+ Bảo mật

VD: Amazon, Redshift, Snowflake

Học thầy không tày học bạn.



www.huongtrang.net

□□□□
horizontal scale ngang

vertical scale dọc

⑤ Dadaap

Framework mã nguồn mở do Apache phát triển, dùng để lưu trữ dữ liệu lớn theo mô hình phân tán

Thủ phần chủ

- + HDFS : lưu trữ dữ liệu
- + Map reduce : xử lý dữ liệu song song
- + Xaman quản lý tài nguyên
- VD : xử lý log web, phân tích dữ liệu thất bại

⑥ HDFS (Hadoop Distributed File System)

Hệ thống phân tán dữ liệu lớn, lưu trữ dữ liệu dưới dạng block, và nhận biết an toàn

Đặc điểm

- + block mặc định 128MB/256MB
- + Dữ liệu được replicate trên nhiều node (3)
- + Khả năng mở rộng theo chiều ngang (scale-out)
- VD : Twitter, Log file, etc



⑧ Map reduce

Việc xử lý phân tán gần như bất kỳ

- Map: dùng vào dữ liệu và xử lý song song
- Reduce: tổng hợp và trả kết quả

Đặc điểm

- + Xử lý theo batch process
- + Xử lý phân tán và xử lý tập

VD: Đếm số lần xuất hiện từ trong tập dữ liệu và báo

⑨ Hadoop Xerus

Quản lý tài nguyên và lập lịch trong Hadoop

+ Đặc điểm

- + Quản lý CPU, RAM trên toàn cluster
- + Hỗ trợ kỹ thuật vận hành

VD: chạy Spark trên Hadoop thay vì YARN

⑨ HBase

Cơ sở dữ liệu NoSQL, chạy trên HDFS, dựa trên mô hình

- Mô hình:
- ~~Key~~ Mô hình Key - value, schema-less
 - Truy vấn theo giá trị
 - Tích hợp dữ liệu phân tán

Học thầy không tày học bạn. FB duy message



www.huongtrung.net

HBase

(10)

Hive

- Hệ thống data warehouse xây dựng trên Hadoop,
- hỗ trợ ngôn ngữ HiveQL (tương tự SQL)
- + Dễ dùng cho người quen SQL
- + ứng dụng vào SQL như job MapReduce
- thực tế dữ liệu clickstream web.

(11)

Pig

- Nền tảng xử lý dữ liệu liên kết với ngôn ngữ Pig Latin
- + Dễ dàng
- + Đơn giản hơn khi xử lý MapReduce
- + Hỗ trợ các thao tác ETL
- VD: lưu trữ dữ liệu log web

(12)

Sqoop

(Công cụ chuyển dữ liệu giữa Hadoop và cơ sở dữ liệu quan hệ RDBMS)

Dễ dùng

- + Import dữ liệu từ MySQL/Oracle vào Hadoop
- + Export ngược lại RDBMS
- VD: import dữ liệu khách hàng từ MySQL sang HDFS phân tích

Học ăn học nói, học gói học mở.



www.buongtrung.net



13) ZooKeeper

Dữ liệu phân tán, theo dõi trạng thái của các ứng dụng
phân tán

Đặc điểm

- + cơ chế dữ liệu phân tán
- + phân bổ node (tên, mã, thời gian, địa chỉ)
- + ~~Đảm~~ bảo tính nhất quán trong các ứng dụng

VĐ như trong Kafka, HBase để theo dõi cluster



