

Bộ môn Chuyên đề chọn lọc trong Hệ thống thông tin

Hệ thống thông tin



SEMINAR

Bộ môn Chuyên đề chọn lọc trong Hệ thống thông tin

Hadoop



Hệ thống thông tin

*Sinh viên thực hiện: Nguyễn Hoàng Phúc – 1312440

Apache Hadoop là gì?

Apache™ Hadoop® là một dự án phát triển phần mềm mã nguồn mở với tiêu chí đáng tin cậy, khả năng mở rộng, tính toán phân tán.

Thư viện phần mềm Apache Hadoop là một nền tảng được thiết kế xử lý phân tán các dữ liệu lớn trên các cụm máy tính sử dụng công nghệ lập trình đơn giản. Việc này được thiết kế để mở rộng quy mô từ các máy server đơn giản đến hàng ngàn thiết bị. Thay vì dựa vào việc nâng cấp phần cứng để có tính sẵn sàng cao, thư viện được thiết kế để phát hiện và xử lý vấn đề ở lớp ứng dụng, do đó việc cung cấp một dịch vụ cao có sẵn trên một cụm máy tính mà mỗi máy trong đó có thể bị lỗi thì hữu ích.

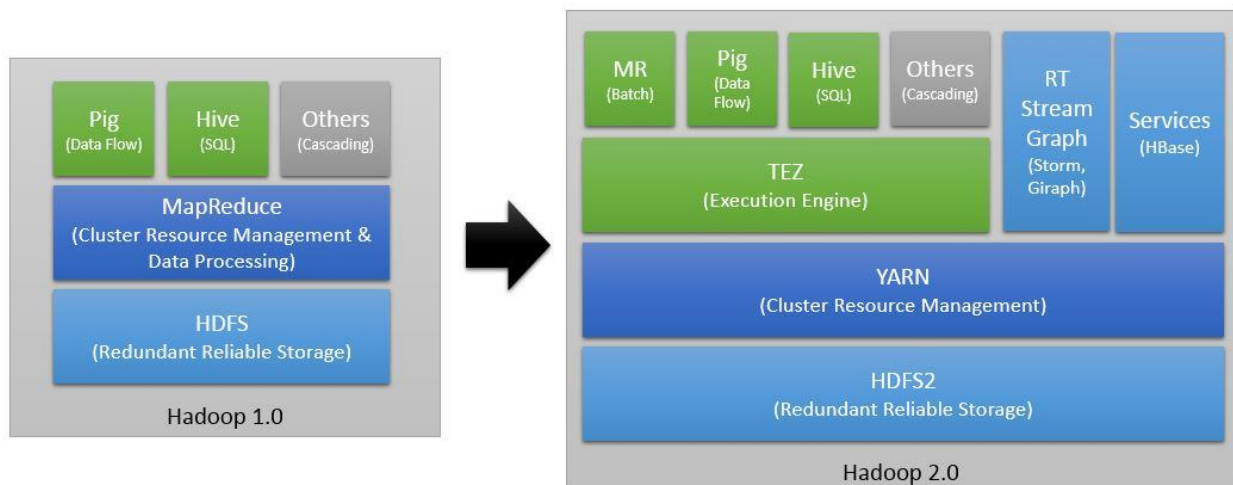
Dự án bao gồm các phân hệ:

- **Hadoop Common**: Các tiện ích phổ biến mà hỗ trợ các module hadoop khác.
- **Hadoop Distributed File System (HDFS™)**: Một hệ thống tập tin phân phối để cung cấp truy cập cao và liên quan (high-throughput access) đến dữ liệu ứng dụng.
- **Hadoop YARN**: Một nền tảng cho việc lập kế hoạch và quản lý tài nguyên cluster.
- **Hadoop MapReduce**: Một hệ thống YARN-based cho việc truy cập song song của dữ liệu cài đặt lớn.

Nguồn tham khảo: <http://hadoop.apache.org/>

Phiên bản và Download

Hadoop đến nay đã trải qua nhiều lần cập nhật và phát triển và mỗi lần phát triển có một chút thay đổi.



Chúng ta có thể download các phiên bản của Hadoop tại

<http://hadoop.apache.org/releases.html> .

Bộ môn Chuyên đề chọn lọc trong Hệ thống thông tin

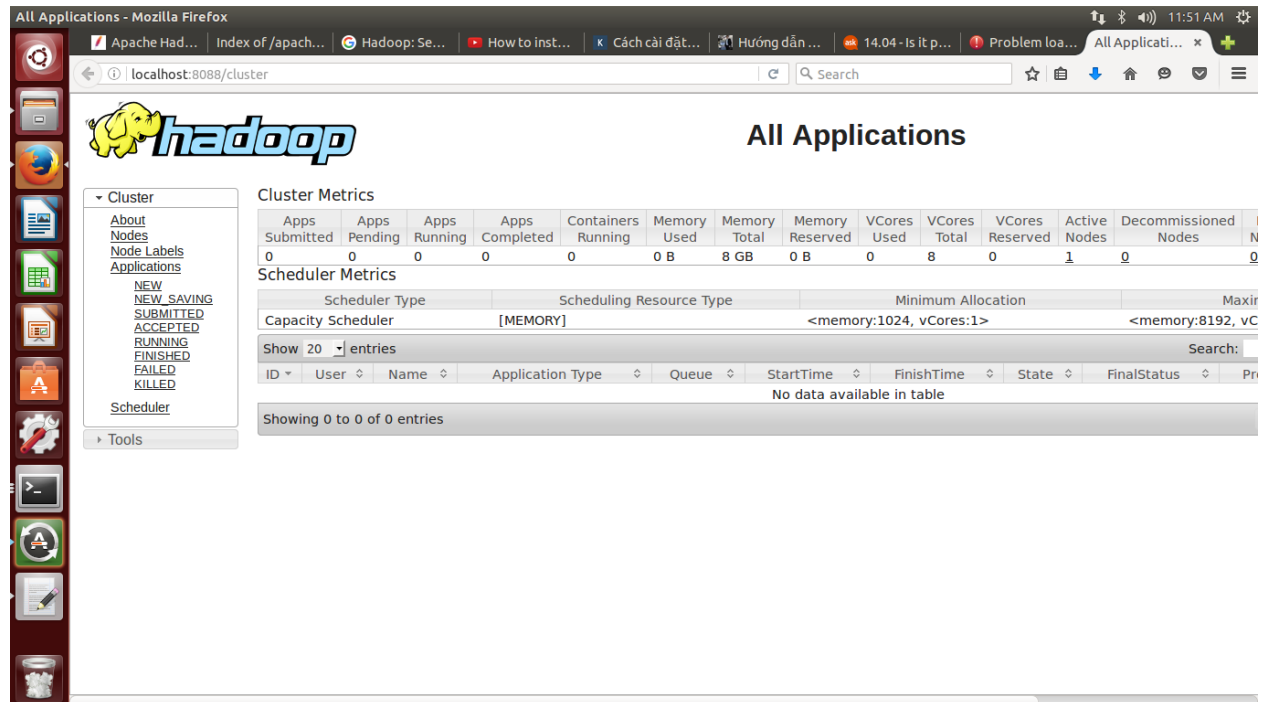
Hệ thống thông tin

Demo chạy hadoop trên máy đơn Linux (Ubuntu).

Các bước thực hiện (nguồn: <https://hadoop.apache.org/docs/stable/hadoop-project-dist/hadoop-common/SingleCluster.html>)

Sản phẩm sau khi thực hiện:

Localhost:8088



The screenshot shows the Hadoop All Applications page in a web browser. The page title is "All Applications". On the left, there is a sidebar with a "Cluster" menu and a "Tools" button. The main content area displays "Cluster Metrics" and "Scheduler Metrics".

Cluster Metrics

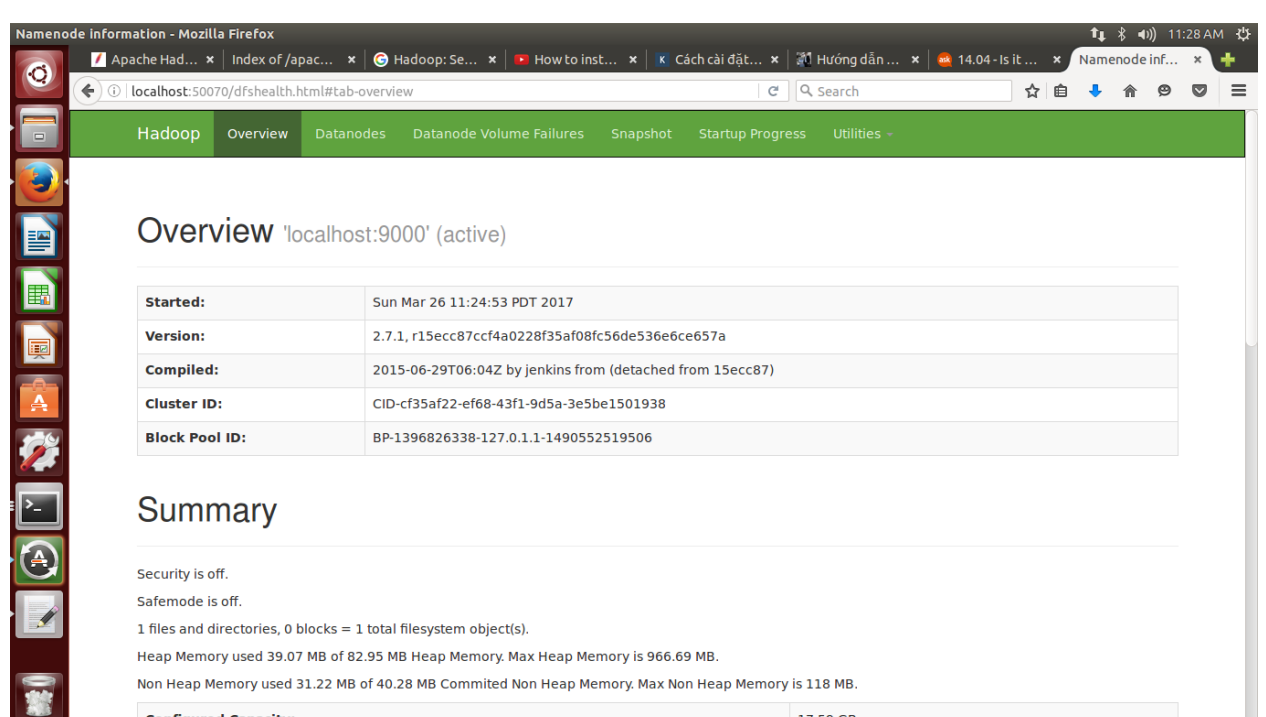
Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memory Total	Memory Reserved	VCores Used	VCores Total	VCores Reserved	Active Nodes	Decommissioned Nodes
0	0	0	0	0	0 B	8 GB	0 B	0	8	0	1	0

Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation
Capacity Scheduler	[MEMORY]	<memory:1024, vCores:1>	<memory:8192, vCores:1>

Showing 0 to 0 of 0 entries

localhost: 50070



The screenshot shows the Hadoop Overview page in a web browser. The page title is "Overview 'localhost:9000' (active)". The page displays various metrics and information about the Hadoop cluster.

Started: Sun Mar 26 11:24:53 PDT 2017

Version: 2.7.1, r15ecc87ccf4a0228f35af08fc56de536e6ce657a

Compiled: 2015-06-29T06:04Z by jenkins from (detached from 15ecc87)

Cluster ID: CID-cf35af22-ef68-43f1-9d5a-3e5be1501938

Block Pool ID: BP-1396826338-127.0.1.1-1490552519506

Summary

Security is off.
Safemode is off.
1 files and directories, 0 blocks = 1 total filesystem object(s).
Heap Memory used 39.07 MB of 82.95 MB Heap Memory. Max Heap Memory is 966.69 MB.
Non Heap Memory used 31.22 MB of 40.28 MB Committed Non Heap Memory. Max Non Heap Memory is 118 MB.

Configured Capacity: 17.59 GB