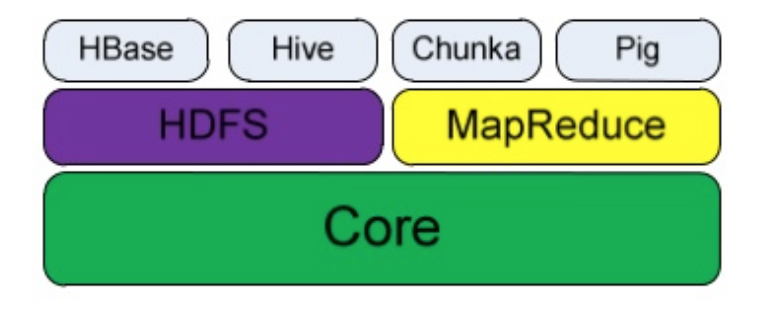
1. **Tổng quan về Hadoop**

Thư viện phần mềm Hadoop là framework cho phép xử lý phân tán một lượng dữ liệu lớn được tiến hành đồng thời qua nhiều máy tính, với một mô hình lập trình đơn giản. Hệ thống được thiết kế để có thể mở từ một máy chủ đơn tới hàn nghìn máy tính mà trong đó mỗi máy tính sẽ cung cấp khả năng tính toán và lưu trữ cục bộ. Ngoài ra, hệ thống cũng dựa vào phần cứng để cấp khả năng khả năng khả dụng và chịu lỗi cao. Bản thân thư viện được thiết kế để tự phát hiện và xử lý các sự cố ở tầng ứng dụngm từ đó cung cấp một dịch vụ có tính khả dụng cao trên nền một cụm máy chủ. Mỗi máy tính trong cụm có thể bị lỗi, nhưng cả cụm máy chủ sẽ không bị ảnh hưởng.

Dự án Hadoop bao gồm một số mô-đun thành phần như sau:

* Hadoop Common: Hệ thống dịch vụ chung hỗ trợ các mô-đun khác.
* Hadôp Distributed File System (HDFS): hệ thống tập tin phân tán cung cấp truy cập thông lượng cao tới dữ liệu ứng dụng.
* Hadoop YARN: Một framework cho việc quản lý lên lịch công việc và quản lý tài nguyên cụm máy chủ.
* Hadoop MapReduce: Một hệ thống dựa trên YARN dành cho việc xử lý song song lượng lớn các tập dữ liệu.

1. **Kiến trúc của Hadoop**

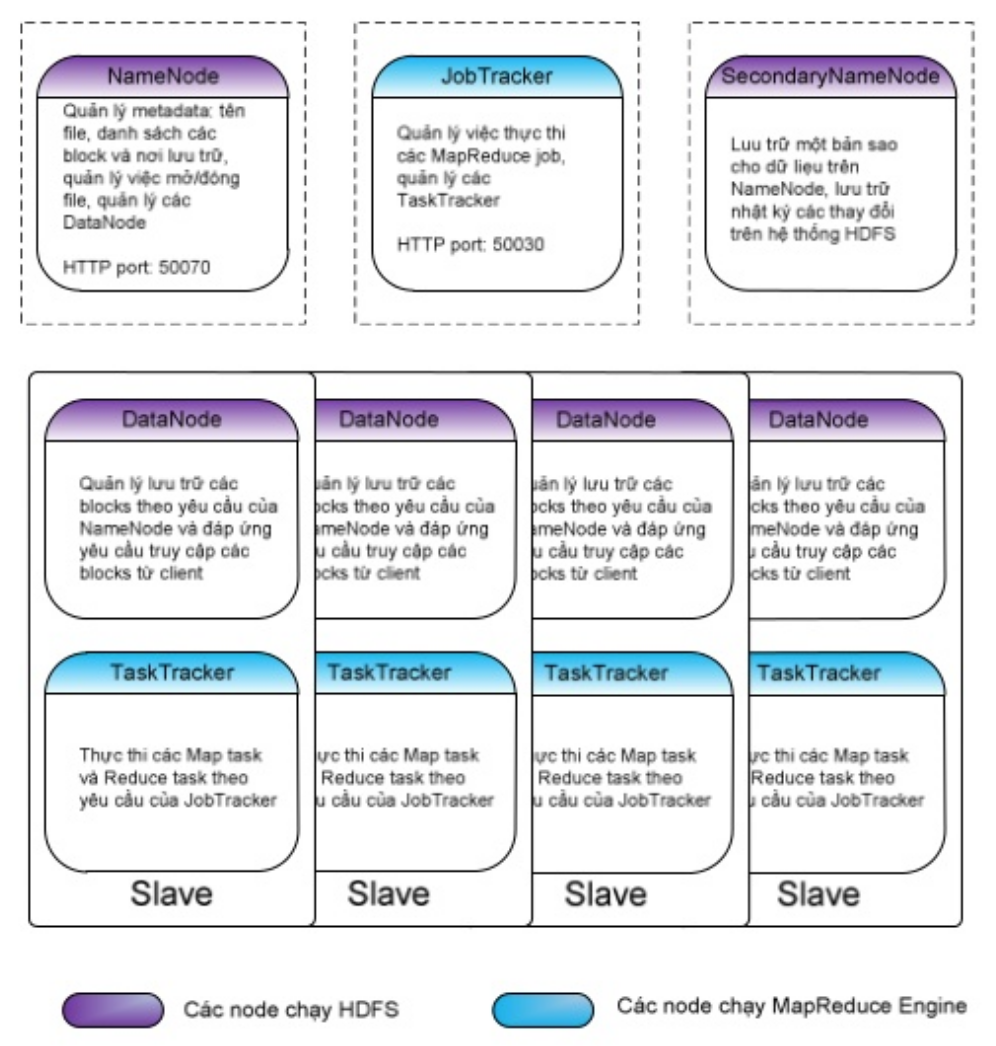


*Cấu trúc của các thành phần của Hadoop*

Hadoop được chia ra làm 3 thành phần chính:

* Core: cung cấp các công cụ và giao diện cho hệ thống phân tán và các tiện ích I/O.Đây là phần lõi để xây dựng nên HDFS và MapReduce.
* MapReduce: một framework giúp phát triển các ứng dụng phân tán theo mô hình MapReduce một cách dễ dàng và mạnh mẽ, ứng dụng phân tán Mapreduce có thể chạy trên một cluster lớn với nhiều node.
* HDFS (Hadoop Distributed Files System): 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.
* Ứng dụng: là các ứng dụng được xây dựng dựa trên nền Hadoop như: Hbase – một CSDL cấu trúc phi quan hệ (NoSQL) dạng cột, Hive – data warehouse cung cấp khả năng truy vấn dữ liệu lưu trữ bằng Hadoop dạng SQL, …

1. Tổng quan Hadoop cluster



*Tổng quan Hadoop Cluster*

Về cơ bản thì Hadoop có kiến trúc master – slave. Với Hadoop, master là node chạy NameNode, Job Tracker. Slave sẽ đóng vai trò Datanode và Task Tracker.

Trên một clustẻ, có duy nhất một node chạy NameNode, một node chạy Job Tracker (NameNode và JobTracker có thể nằm trên cùng một máy vật lý nhưng nên tách riêng thành các máy vật lý khác nhau). ScendaryNameNode là node lưu trữ bản sao của NameNode, lưu trữ nhật ký thay đổi trên hệ thống HDFS.

1. Một số ứng dụng của Hadoop