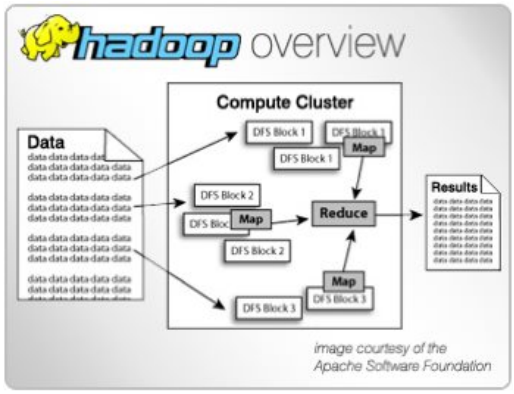
Hadoop Overview

HƯNG LION · THỨ BA, 3 THÁNG 3, 2020 ·

Các bạn đã biết "dữ liệu thế nào gọi là lớn" trong đặc trưng 6V của Big Data, các bạn cũng đã nắm được Apache Hadoop HDFS là một công nghệ được sử dụng phổ biến nhất để implement thành phần Data Storage trong 8 thành phần của Kiến trúc dữ liệu lớn (Big Data Architecture). Trong bài này chúng ta sẽ lướt qua một số thông tin chi tiết hơn về Hadoop cho các bạn cần info nhiều hơn về thành phần này.



Tổng quan về Hadoop

Hadoop là gì?

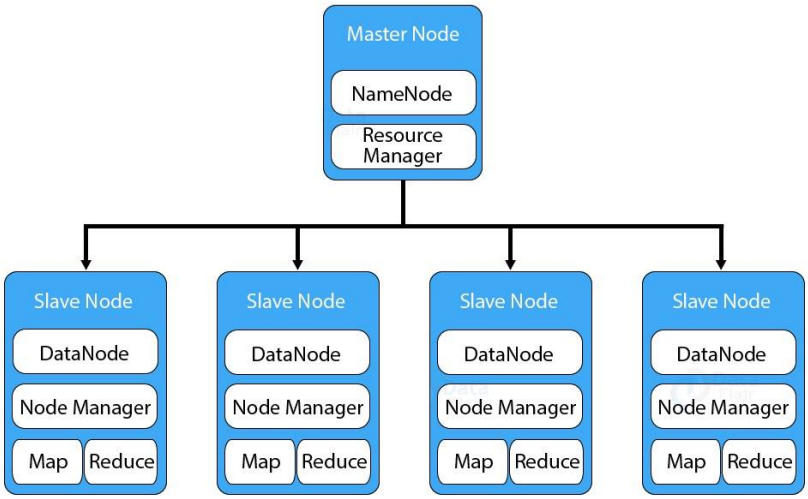
Easy nhỉ 😊, google "Hadoop là gì" cho ra hơn 73K kết quả khác nhau, thậm chí nếu input "What is Hadoop" google còn trả về đến hơn 35 triệu kết quả. Pick lấy 1 cái đơn giản và dễ hiểu nhất (theo mindset của tôi) nhé. ***"Hadoop là một công nghệ để lưu trữ các bộ dữ liệu khổng lồ trên một cụm các máy tính phân tán giá rẻ"***. Hay theo 1 ý khác, Hadoop là một framework phần mềm mã nguồn mở cho phép lưu trữ và xử lý big data trên hệ thống phân tán. Hadoop được phát triển bởi Apache Software Foundation. Hadoop viết bằng Java. Tuy nhiên, nhờ cơ chế streaming, Hadoop cho phép phát triển các ứng dụng phân tán bằng cả java lẫn một số ngôn ngữ lập trình khác như C++, Python, Pearl,...

Một tí lịch sử phát triển của Hadoop:

- 2002: Được bắt đầu với dự án Apache Nutch
- 2007: Yahoo sử dụng hadoop với 1,000 nodes cluster
- 2008: Hadoop trở thành dự án top-level trên Apache
- 2011: Apache released Hadoop version 1.0
- 2012: Apache released Hadoop version 2.0, có chứa Yarn
- 2017: Apache released Hadoop version 3.0
- 2020: Apache released Hadoop version 3.2.1

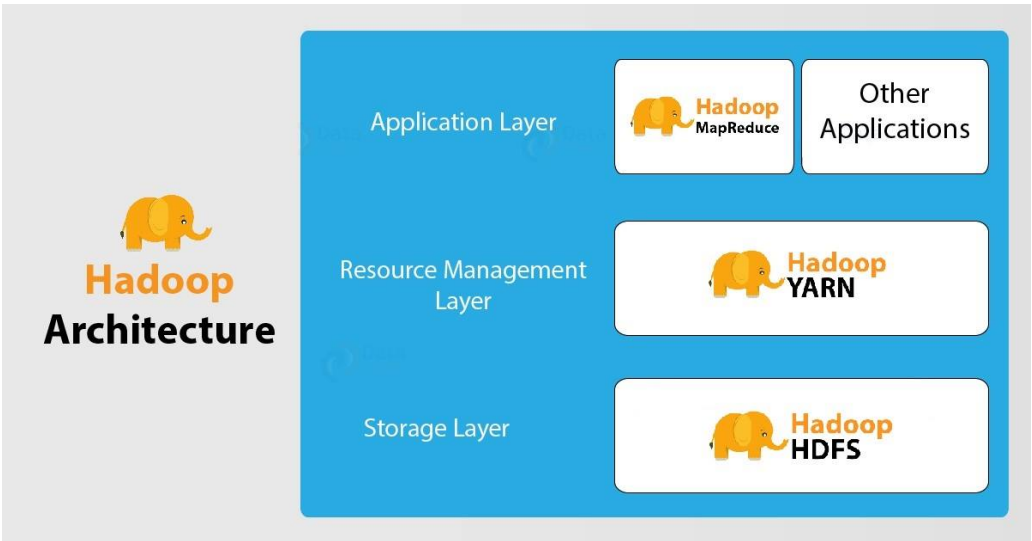
Kiến trúc của Hadoop

Hadoop được cài đặt trên các máy tính trong hệ thống phân tán. Một hệ thống phân tán với các máy cài Hadoop được gọi là một Hadoop Cluster. Hadoop Cluster có kiến trúc Master – Slave hay còn gọi là Name Node – Data Node, chúng ta sẽ có một Name Node và nhiều Data Node, Name Node có chức năng quản lý tài nguyên và phân chia công việc cho các Data Node, mỗi Data Node chính là một máy tính vật lý, dữ liệu thực được lưu trên Data Node, còn trên Name Node lưu metadata (dữ liệu của dữ liệu). Nếu coi hệ thống phân tán là một team, thì Name Node giống như leader, Data Node giống như member của team. Hadoop có thể hoạt động trên một máy (giống như 1 team chỉ có 1 member) hoặc mở rộng tới hàng ngàn máy, với mỗi máy đều có thể sử dụng để lưu trữ hoặc tính toán dữ liệu.



Mô hình Master - Slave

Để đảm bảo tính HA (High Availability), chúng ta còn có Standby Name Node nó là node dự phòng cho Name Node nếu Name Node bị lỗi hoặc không hoạt động.



Kiến trúc Hadoop

Hadoop gồm 3 thành phần chính:

- ***HDFS (Hadoop Distributed File System):*** Là nơi lưu dữ liệu của Hadoop, HDFS chia chia nhỏ dữ liệu thành các đơn vị dữ liệu nhỏ hơn gọi là các blocks và lưu trữ chúng phân tán trong các node của cụm Hadoop. HDFS sử dụng kiến trúc master/slave, trong đó master gồm một Name Node để quản lý hệ thống file metadata và một hay nhiều slave Data Nodes để lưu trữ dữ liệu. Một tập tin với định dạng HDFS được chia thành nhiều khối (blocks) và những khối này được lưu trữ trong một tập các Data Node. Name Node định nghĩa ánh xạ từ các khối đến các Data Node. Các Data Node điều hành các tác vụ đọc và ghi dữ liệu lên hệ thống file. Data Node cũng quản lý việc tạo, huỷ, và nhân rộng các khối thông qua các chỉ thị từ Name Node.
- ***Yarn:*** Quản lý lập lịch các job và quản lý tài nguyên các node.
- ***Map Reduce:*** Là thành phần xử lý dữ liệu của Hadoop, là framework nguồn mở cho phép bạn viết các ứng dụng xử lý dữ liệu. Là nơi để bạn định nghĩa các job.

Một số ưu/ nhược điểm của Hadoop

Ưu điểm: hiển nhiên chắc rất nhiều 😊, vì không phải tự nhiên mà Hadoop là một từ khóa rất hot trong hơn 10 năm qua và được sử dụng trong hầu hết các hệ thống Big Data. Có thể kể đến một số điểm điển hình như:

- Nguồn dữ liệu đa dạng (Varied Data Sources): giải quyết chữ V thứ 3 ttrong đặc trưng 6V của Big Data.
- Giảm chi phí đầu tư (Cost-effective): như đã nói ở phần định nghĩa, Hadoop được thiết kế để chạy "...trên một cụm các máy tính phân tán giá rẻ..." cho nên chi phí để đầu tư các máy tính với cấu hình siêu khủng là không cần thiết nữa rồi.
- Hiệu năng (Performance): như các bạn có thể thấy ở trên, cách thiết kế của Hadoop cho phép các task được thực hiện song song trên các node khác nhau, do đó thời gian truy vấn, lưu trữ giữ liệu sẽ được rút ngắn đến mức tối thiểu.
- Khả năng chịu lỗi (Fault-Tolerant): nghe cái này thì quen rồi, hệ thống phân tán dữ liệu nào cũng sẽ phải đảm bảo việc bảo toàn dữ liệu trong bất cứ tình huống nào.
- Highly Available: điểm chết duy nhất trong cụm Hadoop là Name Node được khắc phục bằng Standby Name Node, thậm chí Hadoop 3 còn support multi Standby Name Node, nên đây ko còn là điểm yếu chí tử của Hadoop nữa
- Khả năng mở rộng (Scalable): nhìn vào kiến trúc phân tán là có thể thấy Hadoop dễ dàng mở rộng capacities bằng cách add thêm các node mới vào cụm rồi
- Open Source: cái này có lẽ là hay ho nhất, còn gì tuyệt vời hơn khi dùng 1 thứ tốt mà free chứ 😊. Nhưng cái hay ho hơn là chúng ta có thể join vào cộng đồng Hadoop để "chọc ngoáy code" cho thỏa đam mê coder, nhỉ.

Nhược điểm: ok, không có cái gì là hoàn hảo, Hadoop cũng ko ngoại lệ. Mặc dù đã được phát triển, nâng cấp qua nhiều phiên bản khác nhau, nhưng Hadoop vẫn đang có những nhược điểm nhất định mà tôi tin là sẽ được khắc phục trong tương lai:

- Vấn đề với số lượng lớn file nhỏ: Hadoop hoạt động rất ổn với lượng nhỏ các file lớn, nhưng khi số lượng file nhiều lên thì việc quản lý của Name Node không hiệu quả nữa. Ở thời điểm hiện tại, HDFS có giới hạn khoảng 350 triệu files và 700 triệu các system objects. Tuy nhiên, những công nghệ tiếp theo như Ozone, đã được đề cập trong VNBDC – Góc tin tức ngày 28/02/2020 có thể giải quyết được vấn đề này.

- Processing Overhead: Hadoop đọc/ ghi dữ liệu thẳng xuống disk nên tốc độ xử lý sẽ phụ thuộc rất nhiều vào tốc độ đọc/ ghi của ổ cứng (cho bạn nào chưa biết thì cost tiêu tốn nhiều nhất khi đọc dữ liệu từ disk là thời gian tìm kiếm đúng block dữ liệu cần đọc, seek time). Hiện tại Hadoop chưa có cơ chế để lưu trữ dữ liệu trên memory, tuy nhiên việc này có thể khắc phục bằng cách dùng các công nghệ khác khi xử lý dữ liệu như Spark chẳng hạn
- Hadoop mới chỉ hỗ trợ xử lý theo lô (Batch Processing): bạn nào chưa biết Batch Processing là gì thì tham khảo lại bài này Batch Processing vs Stream Processing.
- Tính bảo mật (Security): topic có thể gây tranh cãi, nhưng hiện tại Hadoop đang hỗ trợ Kerberos, các bạn có thể gặp khó khăn trong việc quản lý nó cũng như thiếu phần mã hóa ở mức độ storage và network

48 12 bình luận 6 lượt chia sẻ

Thích Bình luận Chia sẻ Lưu

Nhân Đức mình đang quan tâm tới vấn đề security của hadoop cluster, hi vọng có ai đó đã từng tìm hiểu, cài đặt, thử nghiệm về khoản này chia sẻ kiến thức cho anh em 🙂

Thích · Trả lời · 4 tuần 6

Lê Trường đã trả lời · 2 phản hồi

Nguyễn Chí Thanh Klq nhưng năm trc mình đc gặp vs nói chuyện 1 chút chút với bác Doug Cutting, 1 trong 2 co-fouders của Hadoop hiện đang làm việc tại Cloudera

15

Thích · Trả lời · 4 tuần

Đoàn Thanh Tám đã trả lời · 5 phản hồi

Đoàn Thanh Tám Bài này có nhắc đến Standby Name Node. Quick question: Standby Name Node vs Secondary Name Node khác nhau ntn? (Notes: cấm Google)

Thích · Trả lời · 4 tuần 1

Đoàn Thanh Tám đã trả lời · 2 phản hồi

Viết bình luận...