Các công nghệ liên quan

1. Kafka

2. Zookeeper

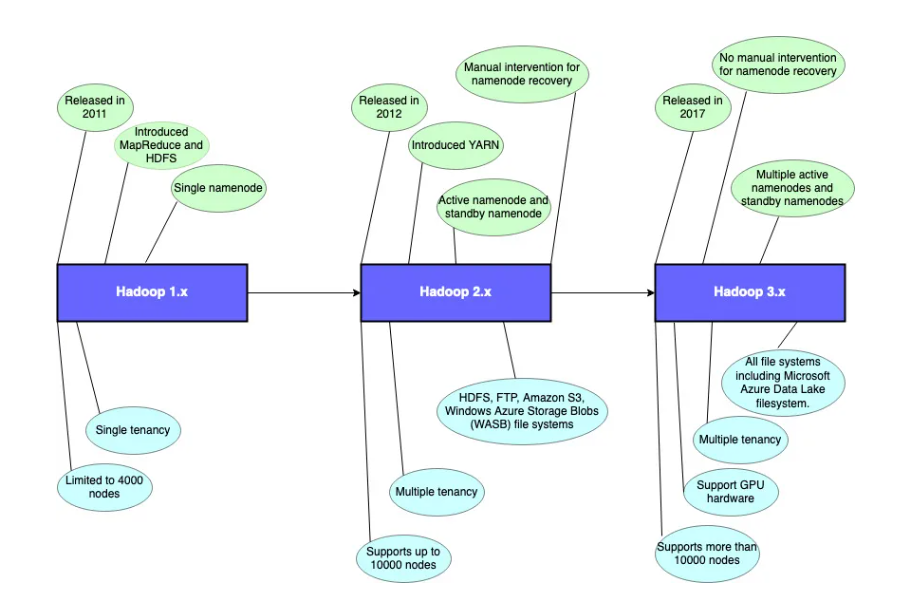
3. Hbase

4. Nifi

5. Airflow

Sự khác nhau giữa các phiên bản Hadoop

|  |  |  |
| --- | --- | --- |
| Hadoop 1.x | Hadoop 2.x | Hadoop 3.x |
| Phát hành năm 2011 | Phát hành 2012 | Phát hành 2017 |
| Giới thiệu MapReduce và HDFS. MapReduce được sử dụng để xử lý dữ liệu và quản lý tài nguyên | Yarn được thêm vào để cải thiện việc quản lý tài nguyên. Nó cho phép đa người dùng sử dụng cùng một cụm máy chủ bởi cả MapReduce và một số quy trình khác sử dụng YARN. | Mô hình tài nguyên YARN được tổng quát hóa để hỗ trợ các loại tài nguyên do người dùng xác định khác CPU và bộ nhớ. |
| Hỗ trợ đơn người dùng độc lập | Hỗ trợ đa người dùng thông qua Yarn | Hỗ trợ đa người dùng cho phép nhiều người dùng chung một cụm máy chủ |
| Sử dụng kiến trúc Master-Slave với một master đơn và nhiều slaves. Do đó, nếu master node gặp sự cố, cả cụm máy chủ sẽ trở nên không khả dụng | Kiến trúc Master-Slave bao gồm nhiều master bao gồm active namenode và standby namenode. | Hỗ trợ nhiều active namenode |
| Giới hạn 4000 node mỗi cụm | Hỗ trợ lên đến 10000 node mỗi cụm | Khả năng mở rộng được cải thiện, có thể hỗ trợ hơn 10000 node mỗi cụm |
|  | Cần can thiệp thủ công để phục hồi namenode | Không cần can thiệp thủ công để phục hồi namenode |
|  | Hỗ trợ tối thiểu Java 7 | Hỗ trợ tối thiểu Java 8 |
|  | Hỗ trợ các hệ thống tệp HDFS (mặc định), FTP, Amazon S3 và Windows Azure Storage Blobs (WASB). | Tất cả các hệ thống tệp bao gồm hệ thống tệp Microsoft Azure Data Lake đều tương thích với Hadoop 3.x. |
|  | Phương pháp nhân bản 3x với chi phí phụ 200% | Sử dụng mã hóa eraser trong HDFS giúp giảm chi phí lưu trữ chỉ còn 50% dung lượng lưu trữ phụ. |
|  |  | Bổ sung hỗ trợ cho phần cứng GPU có thể được sử dụng để thực thi các thuật toán nghiêng sâu trên cụm Hadoop |



PySpark

