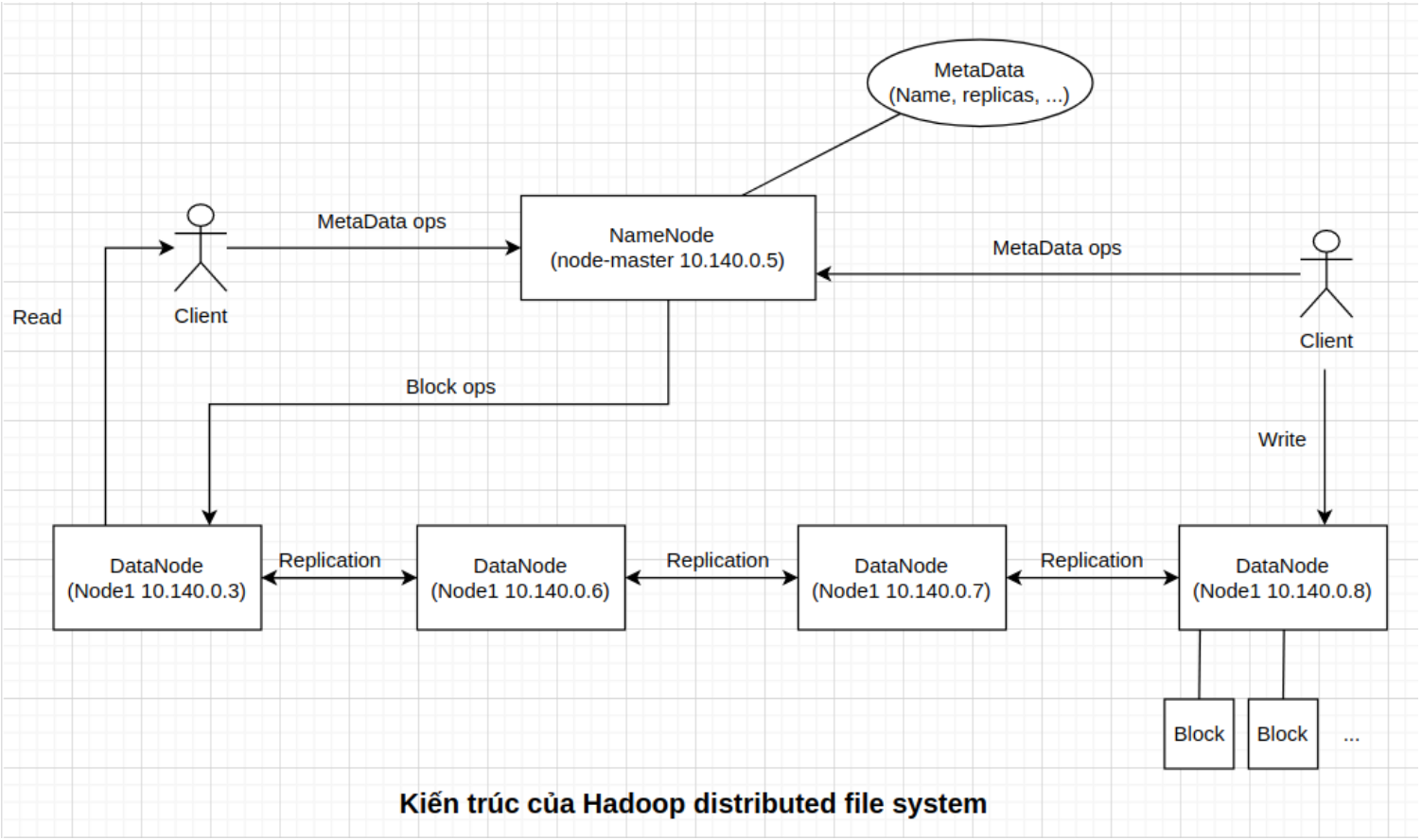
Bài tập tuần 4 - Hadoop - Chương trình masterDev mùa 3
Họ và tên: Phạm Văn Ngọc - ngocpv22

1. Kiến trúc Hadoop distributed file system (HDFS Architecture)



Kiến trúc của Hadoop distributed file system

Mô tả các thành phần trong kiến trúc HDFS:

1.NameNode:

NameNode là nơi lưu trữ và cập nhật các meta data như địa chỉ của block đặt tại datanode, quyền truy cập của client tới các block, … => Nó kiểm soát quyền truy cập từ phía client.

NameNode có nhiệm vụ chính là quản lí và duy trì các dataNode, kiểm soát thông tin tình trạng của các dataNode, luôn lắng nghe để đảm bảo các dataNode còn sống. Nó điều khiển sự cân bằng về lưu lượng lưu trữ và lượng truy cập tới các dataNode.

Nếu NameNode chết thì cả cụm hadoop sẽ chết theo.

2.DataNode

DataNode có vai trò:

Là nơi lưu trữ các dữ liệu khi dữ liệu được đưa vào HDFS
Là nơi chạy các tiến trình xử lý dữ liệu (Khi các job được chia cho các dataNode)
Gửi tình trạng health về cho nameNode thường xuyên (khoảng 3s một lần để NameNode biết tình trạng của nó).
Trong quá trình write từ phía client, DataStream sẽ chuyển packet vào dataNode phù hợp nhất, sau đó dataNode này sẽ chuyển tiếp đến các dataNode khác (Quá trình replication)

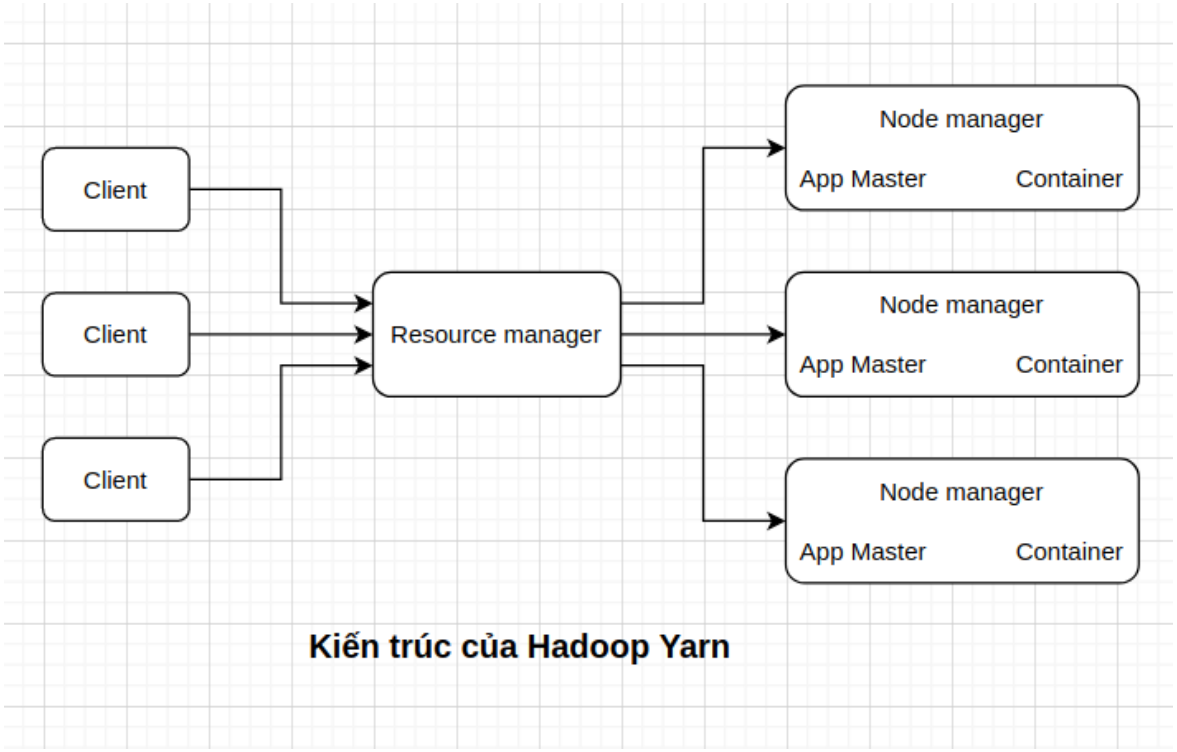
3.Block

Dữ liệu trong HDFS sẽ được chia nhỏ thành nhiều mảnh, mỗi mảnh như vậy được gọi là block, các block này nằm trong các dataNode. Thông tin về dữ liệu được chia thành bao nhiêu block, block nằm ở dataNode nào sẽ được lưu trong metaData của nameNode.
Các block được nhân bản (quá trình replication) và lưu tại các dataNode, để đảm bảo không mất mác dữ liệu khi một dataNode nào đó gặp sự cố hoặc chết, đây là một trong những đặc trưng của HDFS về độ tin cậy và khả năng phục hồi sau lỗi tốt.

2. Kiến trúc Yarn

Yarn (“Yet-Another-Resource-Negotiator”) là framework hỗ trợ phát triển ứng dụng phân tán, được sử dụng làm hệ thống quản lý resource trong Hadoop.

Kiến trúc của Yarn để quản lý tài nguyên và quản lý job thành chia thành 2 components riêng biệt: Resource manager và Application Master.



Kiến trúc của Hadoop Yarn

Mô tả các thành phần trong kiến trúc của Yarn:
Resource manager: Quản lý toàn bộ tài nguyên tính toán của toàn bộ cụm hadoop.

Node manager: Mỗi Node manager quản lý các job trên mỗi node. Các node tính toán trong cluster bây giờ sẽ chạy NodeManager để quản lý các tiến trình chạy trên máy đó.

Application Master: Quản lý vòng đời của job trên các DataNode. Nó hoạt động cùng với Node Manager và giám sát việc thực thi các task.

Container: Quản lý tài nguyên vật lý cho các node như RAM, CPU, ...

Client: Đóng vai trò đưa các job map-reduce cho hệ thống xử lý.

3. Tóm tắt các lệnh cơ bản thao tác với HDFS

*Thao tác với hệ thống file HDFS:

Xem các mục có trong path:	hdfs dfs -ls +path :
Tạo thư mục:	hdfs dfs -mkdir /user/ngocpv22
Tạo file:	hdfs dfs -touch /user/ngocpv22/myFile
Tạo file có nội dung:	hdfs dfs -echo "abc" >myFile
Xóa thư mục:	hdfs dfs -rm -r -f /user/ngocpv22
Trao quyền:	hdfs dfs -chown -R ngocpv22 /user/ngocpv22
Đẩy 1 file từ ngoài vào hdfs:	hdfs dfs -put myFile2 /user/ngocpv22

Các lệnh sử dụng với quyền admin (super user)

Format lại hệ thống file:	hadoop namenode -format
Khởi động/Dừng hệ thống hdsf:	sbin/start-all.sh sbin/stop-all.sh

Kiểm tra tình trạng hệj thống hdfs: hadoop fsck /

Báo cáo và thống kê chi tiết tình trạng HDFS: hadoop dfsadmin -report