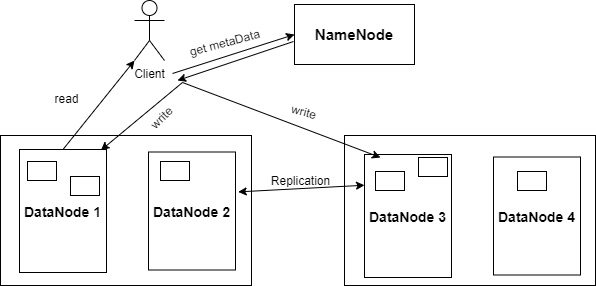
**Hadoop trong xử lý bigdata**

1. **Kiến trúc HDFS**



NameNode : là node cho quản lí và điều khiển các dataNode của HDFS. Cũng được nói như master Node. Là đầu mối truy cập thực hiện các file từ client. Luôn theo dõi các datanode để đảm bảo chúng còn sống.

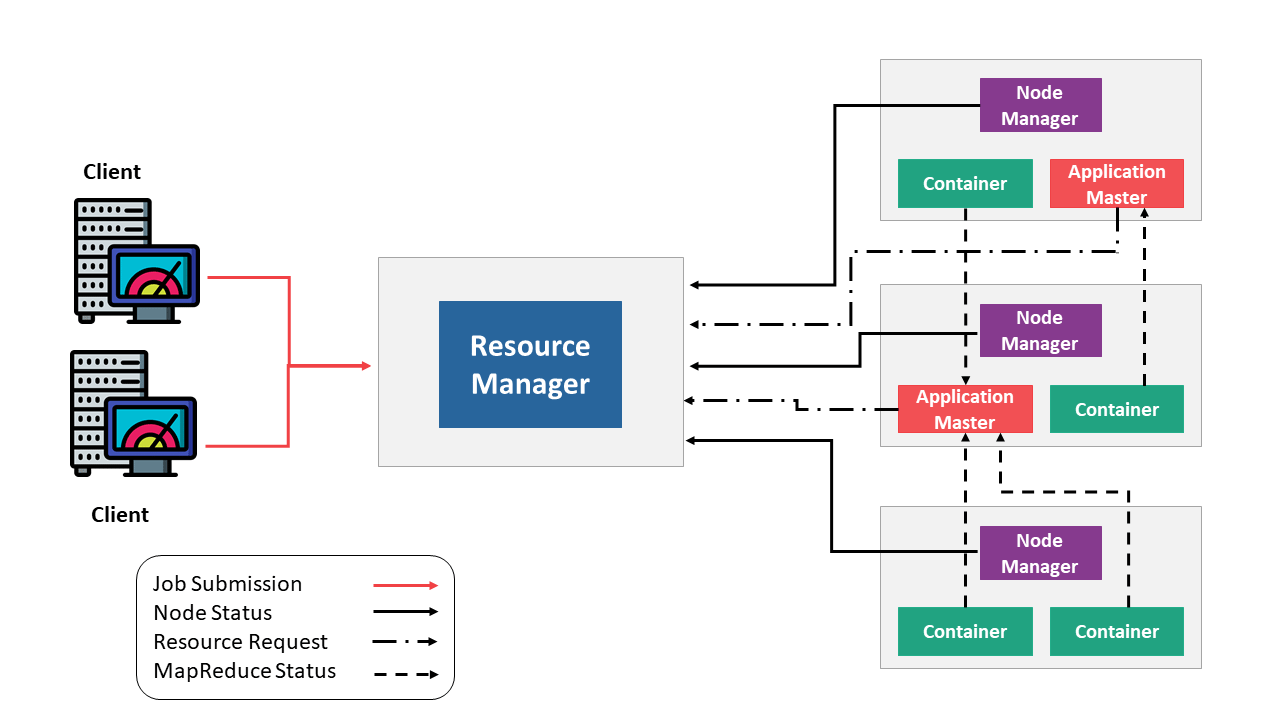
DataNode : nơi lưu trữ các file dữ liệu của DataNode. Là nơi chạy các tiến trình dữ liệu và gwuir giữ liệu định kì tình trạng sức khỏe chính nó. Các dataNode có cơ chế sao lưu dữ liệu cho nhau để phục hồi dữ liệu khi có một Node nào đó bị hỏng hoặc bị mất.

Khi truy vấn dữ liệu trong dataNode , đầu tiên thì ta phải truy vấn các meta data trong NameNode như dữ liệu nằm ở Node nào , bắt đầu từ địa chỉ nào , bao nhiêu bộ nhớ,...

Khi viết dữ liệu thì ta viết đồng thời vào dữ liệu các Node.

1. **Kiến trúc về Yarn**

Đây là một framework cho việc cho việc lập lịch job và quản lí tài nguyên cho các cụm.

 Resource Manager : Client giao một task cho resource Manager , resource Manager sẽ giao sẽ tính toán các tài nguyên cần thiết theo yêu cầu của ứng dụng với application master được chuyển đến chạy một nút tính tóan.

Resource Manager có hai thành phần chính là Scheduler và Application Manager.

* Scheduler : có nhiệm vụ cung cấp tài nguyên của các ứng dụng. Không có nhiệm vụ quản lý task hay theo dõi tình trạng của ứng dụng hoặc như việc restart các task bị fail. Việc cung cấp tài nguyên dựa vào yêu cầu ứng dụng,các tài nguyên, ram,..
* ApplicationManager : Có nhiệm vụ tiếp nhận các job từ client, cung cấp tài nguyên để bắt đầu application master cho từng ứng dụng.

Node Manager được cài trên mỗi node có nhiệm vụ quản lý tài nguyên và job trên node đó. Node manager gửi heardbeart tới Resource manager để thông báo tình trạng. mục đích chính của node manager là quản lý các container mà resource manager giao cho nó

Application master là một container được khởi tạo bởi RM, tương ứng với mỗi job submit lên Resource Manager là một app master. App master có nhiệm vụ điều phối các hoạt động của app trong cụm

Containner là tiến trình yarn , tập hợp các tài nguyên vật lý như core và ram. Dối tượng này lưu trữ các thông tin như biến môi trường, dependencies, security token và chỉ thị để tạo tiến trình.

1. **Các câu lệnh trong hệ thống file HDFS**

Câu lênh tổng quát của HDFS : hadoop fs <args>

**mkdir** tạo thư mục : hadoop fs -mkdir <name>

**appendtoFile** nối chuỗi vào file : hadoop fs -appendtoFile <local res> <des>

**cat** copy dữ liệu đường dẫn ra stout: hadoop fs -cat <src>

**chmod** thay đổi quyền cho phép thao táccủa file. Cùng -R là đệ quy vào trong thư mục file hadoop fs -chmod -R <src>

**chown** thay đổi người sở hữu file. Người sử dụng phải là super-user:

hadoop fs -chown – R <src>

**cp** copy file source tới destination: hadoop fs -cp <src> <des>

**put** copy file từ source tới hệ thống file hadoop: hadoop fs -put <src> <des>

**touch** tạo file để viết : hadoop fs -touch <name>

**tail** hiển thị kilobyte cuối cùng của file: hadoop fs -tail <src>

**…**