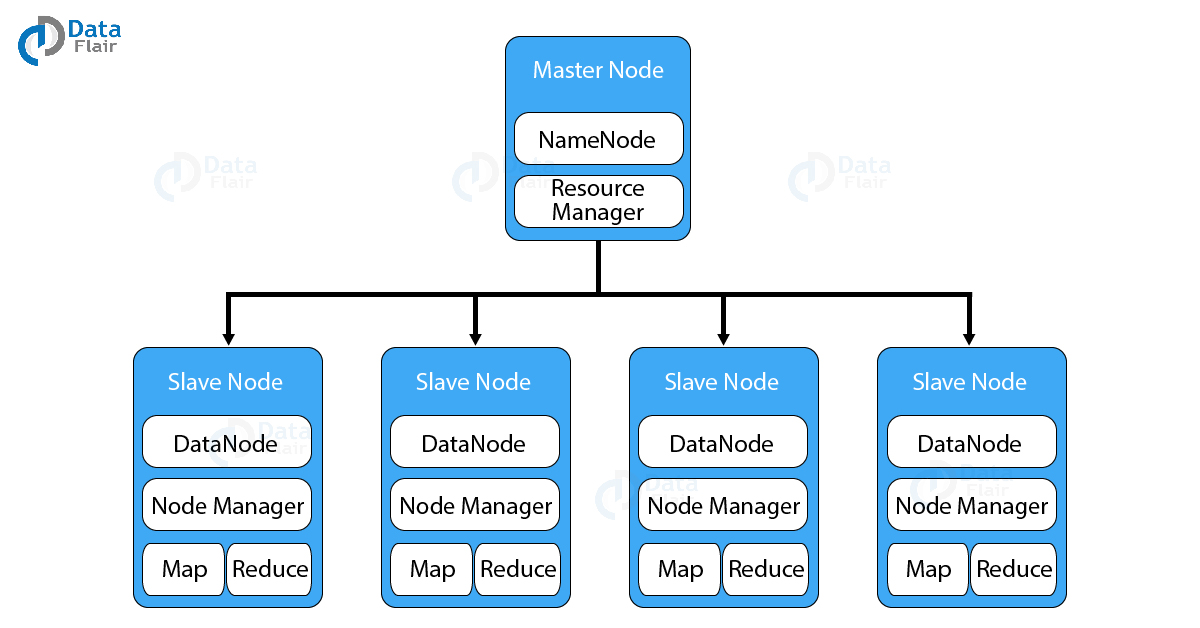
# 1:Kiến trúc HDFS

-HDFS(Hadoop Distributed File System) :là hệ thống lưu trữ file phân tán của hadoop

-HDFS có kiến trúc master-slave .Có 2 daemon NameNode và DataNode lần lượt chạy trân node master và node slave

* NameNode chịu trách nghiệm quản lý,sửa đổi namespace và quyền truy cập của các máy khách tới các node
* DataNode chịu trách nghiệm lưu trữ dữ liệu ,phục vụ yêu cầu đọc ghi



-Trong master và slave node còn có 2 daemon nữa là resource manager và node manager thuộc YARN để quản lý tài nguyên

# 2:YARN

-Một số khái niệm :

* Resource Container (RC) là một tập các tài nguyên vật lý
* Application thì tượng trưng cho 1 tập các tác vụ .Trong YARN thì Application tương đương với MapReduce's job

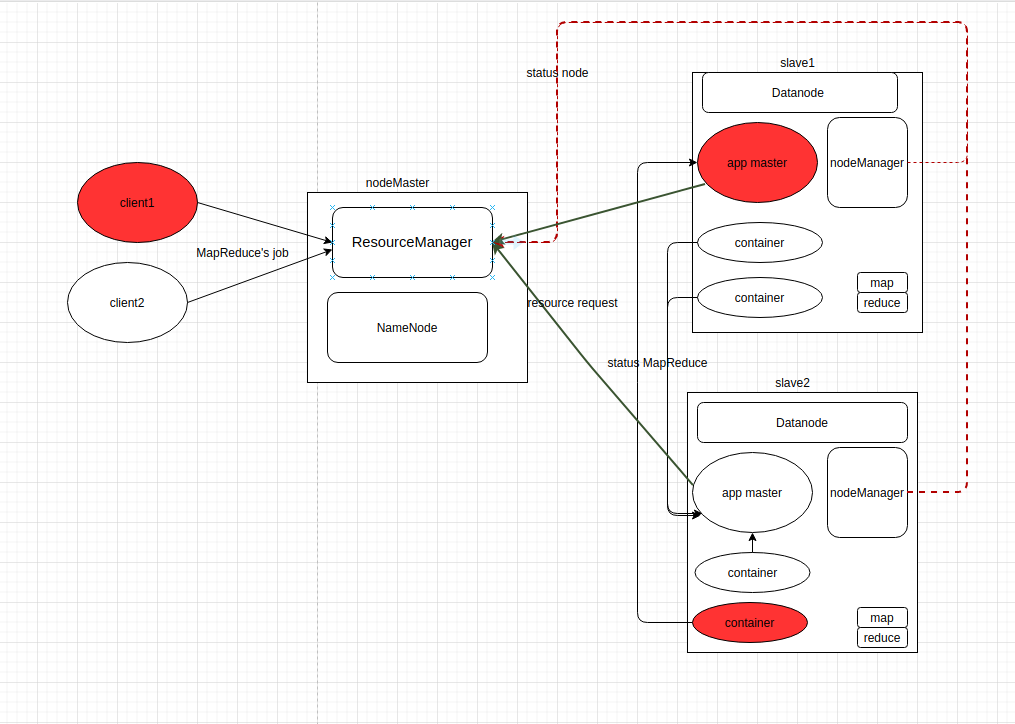
-YARN(Yet Another Resource Negotiator) là lớp quản lý tài nguyên trong Hadoop .Ngoài quản lý tài nguyên thì YARN cũng thực hiện lập lịch công việc

-YARN có 3 thành phần chính là resource manager, node manager và ApplicationMaster

* RM(resource manager)(một trên mỗi cụm ) : Là nơi quản lý và phân bổ tài nguyên .Quan trọng nhất trong đó là Scheduler
  + Scheduler ở trong RM thực hiện việc lập lịch cho các ứng dụng đang chạy nhưng không giám sát theo dõi trạng thái của các tiến trình trong ứng dụng
  + bên cạnh đó còn Application manager: chịu trách nghiệm khởi động ApplicationMaster giám sát và khởi động lại chúng trên các node nếu bị lỗi
* NodeManager(một trên mỗi node) : quản lý việc sử dụng tài nguyên như ram ,cpu,bộ nhớ,.. của các Container và báo cáo lại cho resource manager
* ApplicationMaster(một trên application) :mỗi application trên Hadoop đều có một ApplicationMaster cho riêng nó .ApplicationMaster cứ định kỳ sẽ gửi trạng thái của application về ResourceManager,cũng như yêu cầu thêm về tài nguyên nếu có ApplicationMaster sẽ theo dõi việc thực thi của một application trong suốt thời gian tồn tại của nó

-Quy trình thực thi application với YARN

1. client gửi MapReduce application cho ResourceManager, đồng thời gửi thông tin để chạy ApplicationMaster của application đó
2. ResourceManager trao đổi một container cho ApplicationMaster và khởi chạy ApplicationMaster.
3. ApplicationMaster khởi động và đăng ký với ResourceManager, cho phép client gọi ban đầu giao tiếp trực tiếp với ApplicationMaster.
4. ApplicationMaster trao đổi tài nguyên(resource containers) cho client application
5. ApplicationMaster cung cấp thông số kỹ thuật khởi chạy container cho NodeManager,NodeManager đó sẽ khởi chạy container cho application.
6. Trong quá trình thực thi ,client sẽ xem trạng thái và tiến trình của application thông qua ApplicationMaster
7. Sau khi hoàn thành, ApplicationMaster hủy đăng ký với ResourceManager và tắt, trả các container của nó về resource pool.



# 3: Một số lệnh với HDFS

Xem các mục có trong path: hdfs dfs -ls +path :

Tạo thư mục: hdfs dfs -mkdir /user/dirName

Tạo file : hdfs dfs -touch /user/fileName

Tạo file có nội dung : hdfs dfs -echo "abc" >fileName

Xóa thư mục : hdfs dfs -rm -r -f /user/dirName

Trao quyền : hdfs dfs -chown -R userName /user/dirName

Đẩy 1 file từ ngoài vào hdfs: hdfs dfs -put myFile2 /user/fileName

Các lệnh sử dụng với quyền admin (super user)

Format lại hệ thống file : hadoop namenode -format

Bật tắt hệ thống hdsf : sbin/start-all.sh sbin/stop-all.sh

Status hệ thống hdfs : hadoop fsck /

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

kiểm tra tất cả hệ thống (fsck : filesystem check) : hdfs fsck /

đưa ra thông tin hệ thống : hdfs dfsadmin

phân bổ các block 1 cách đồng đều : start-balancer.sh