week 2) hadoop Dockerfile

Hadoop이란?

여러 대의 컴퓨터를 클러스터화해서, 대규모 데이터를 클러스터에서 분산 처리하는 자바기반 오픈소스 프레임워크

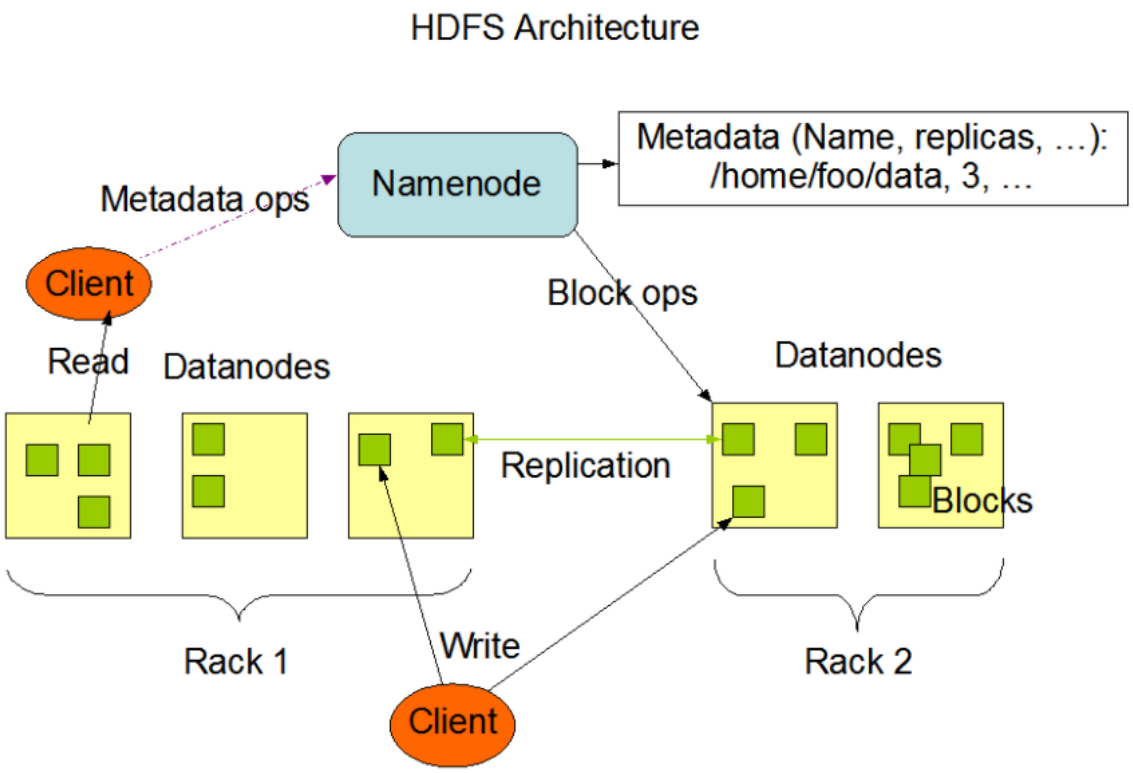
개요

기본 Hadoop 프레임워크는 다음과 같은 4개의 모듈로 구성되며, 이러한 모듈이 함께 작동하여 hadoop 생태계를 형성합니다.

- **Hadoop 분산 파일 시스템 (HDFS)**
  - 분산 저장을 처리하기 위한 모듈
  - 여러 개의 서버를 하나의 서버처럼 묶어서 데이터를 저장
- **맵리듀스 (MapReduce)**
  - 분산되어 저장된 데이터를 병렬 처리할 수 있게 도와주는 분산 처리 모듈
- **Yet Another Resource Negotiator(YARN)**
  - 병렬 처리를 위한 클러스터 자원관리 및 스케줄링 담당
- **Hadoop Common**
  - 하둡의 다른 모듈을 지원하기 위한 공통 커포넌트 모듈

HDFS (하둡 분산 파일 시스템)

- HDFS는 여러 컴퓨터에 대용량 파일들을 나눠서 저장한다
- 마스터 슬레이브 구조 = 하나의 네임노드 + 여러 개의 데이터 노드



1) 네임노드

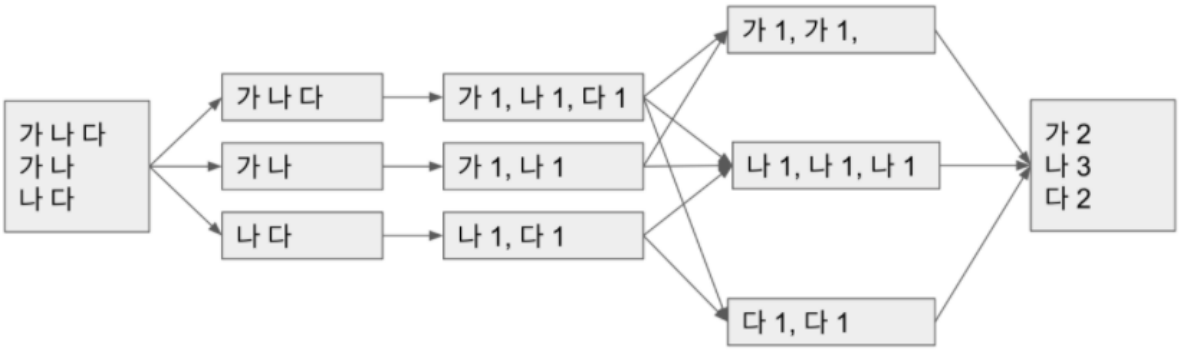
- 메타데이터 관리
- 데이터 노드 관리

2) 데이터 노드

- 파일을 블록 단위로 저장
- 주기적으로 네임노드에 하트비트와 블록 리포트 전달
  - 하트 비트 : 네임노드는 하트비트를 기반으로 데이터 노드의 동작 여부 판단.
  - 블록 리포트 : 블록의 변경사항 체크, 네임노드의 메타데이터 갱신

맵리듀스

- 분산되어 저장된 데이터를 병렬 처리할 수 있게 도와주는 분산 처리 모듈
- 간단한 단위 작업을 반복하여 처리
  - Map 작업 : 간단한 단위 작업을 처리
  - Reduce : 맵 작업의 결과물을 모아서 집계
- 맵, 리듀스 작업은 병렬로 처리가능 → 처리 속도 증가



참고 사이트

https://wikidocs.net/22827

https://ko.wikipedia.org/wiki/아파치_하둡