

[Hadoop 개요]

1. Hadoop이란?

- 대량의 자료를 처리할 수 있는 큰 컴퓨터 클러스터에서 동작하는 신뢰성 있고, 확장성 있는, 분산 컴퓨팅을 위한 오픈소스 프레임워크
- 분산 파일 시스템 GFS, 분산 처리 시스템 MapReduce 소프트웨어 구현체

2. Hadoop의 종류

- (1) 아파치 소프트웨어 파운데이션
- (2) 클라우데라(Cloudera)
- (3) 호튼웍스(Hortonworks)
- (4) 마이크로소프트 애저(Azure)
- (5) 아마존(Amazon Web Service)

3. HDFS(Hadoop Distribute File System)

- 어플리케이션 기반 파일 시스템으로, 파일의 분산 저장이 목적
- NameNodes와 DataNodes로 구성
 - 1) Master NameNode: 파일 시스템 이미지(fsimage)와 변경 기록(edits)을 저장
 - 2) Secondary NameNode: Master NameNode의 fsimage파일과 edits 파일의 사본 저장
 - 3) DataNode: 데이터 파일의 블록 저장 (디폴트 블록 크기는 128MB)
- Block(Chunk)단위로 파일 관리: 저장, 복제(안전성, 신뢰성 보장), 삭제