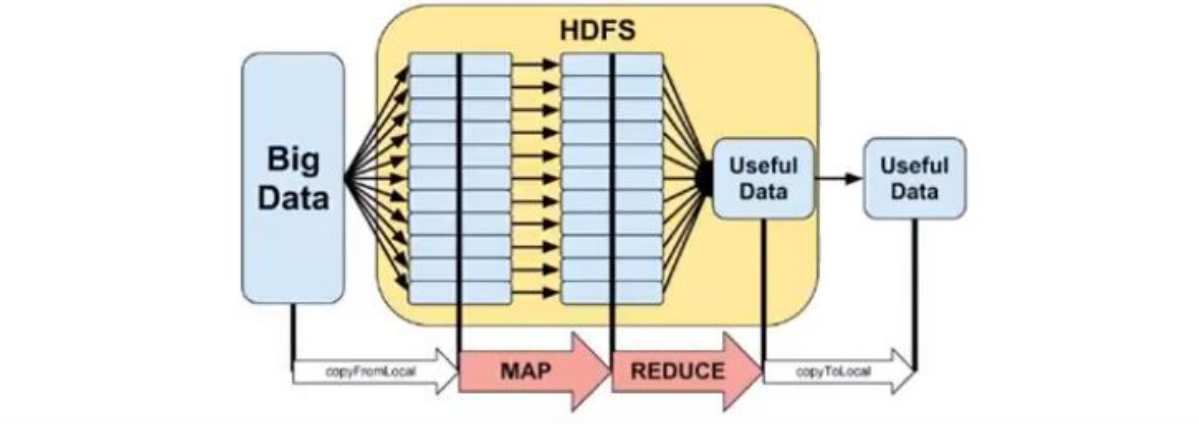
하둡이란?

1.대용량 데이터를 분산처리 할 수 있는 자바 기반의 오픈소스 Framework
2.더그 커팅이 구글의 논문(2003년, “The Google File system”, 2004년, “MapReduce: Simplified Data Processing on Large Cluster”)을 참조하여 구현함
3.더그 커팅의 아들이 노란 코끼리 장난감 인형을 hadoop이라고 부르는 것을 듣고 명명함
4.공식사이트: http://hadoop.apache.org

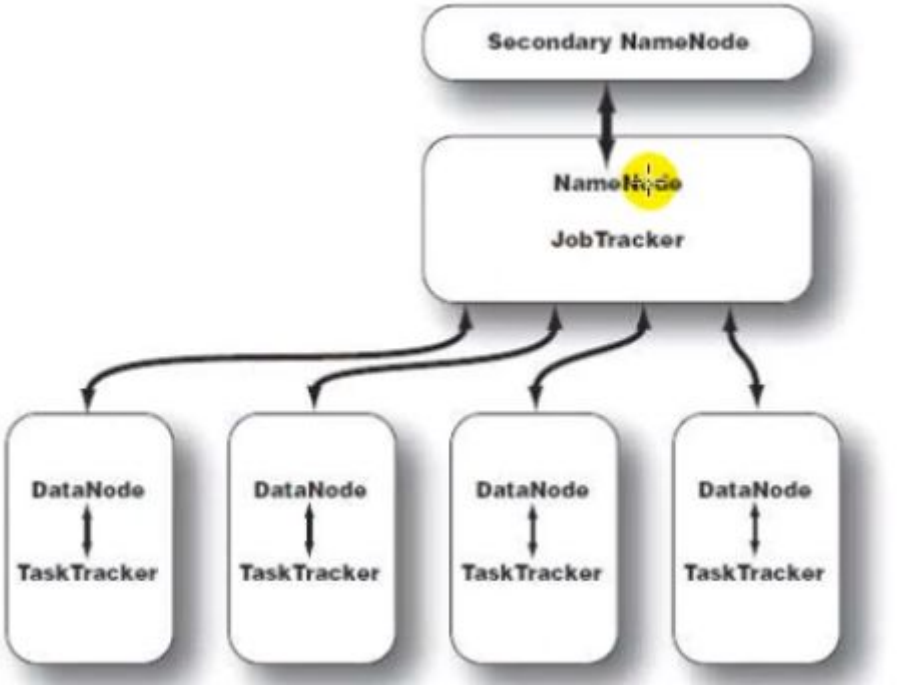
Hapdoop 관련 용어 정리

1)HDFS(Hadoop Distributed FIle System)

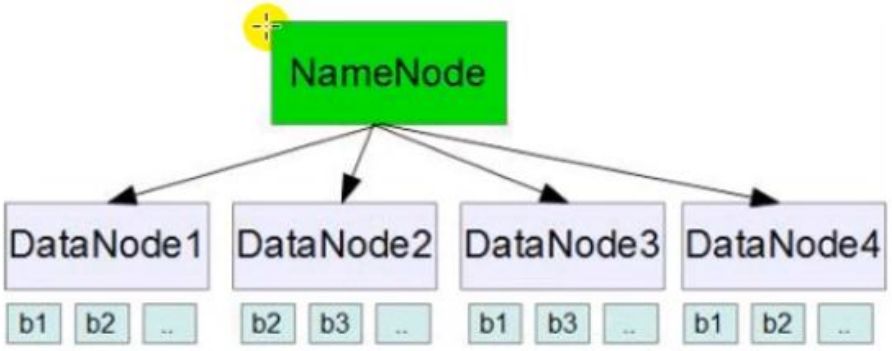
대용량 파일을 분산된 서버에 설치하고 많은 클라이언트가 저장된 데이터를 빠르게 처리할 수 있게 설계된 파일 시스템



큰 데이터가 있을 때, HDFS에 넣어주게되면 MAP과정을 통해 분산처리를 하고, Reduce과정을 통해 취합한다.



2)NameNode:가장 중요한(main)역할을 하는 노드



HDFS의 모든 메타데이터를 관리하고 클라이언트가 HDFS에 저장된 파일에 접근할 수 있도록 처리하는 노드

3)Secondary NameNode : NameNode의 보조 노드

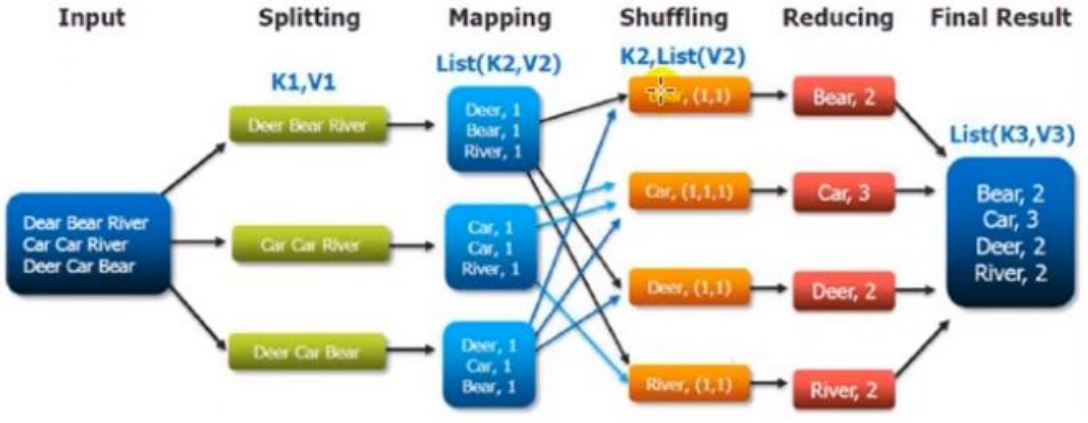
주기적으로 네임노드의 파일 시스템 이미지 파일을 갱신하는 역할을 수행하는 노드

4)DataNode:실제로 데이터를 분산해서 처리하는 노드

HDFS에 데이터를 입력하면 입력 데이터는 32MB의 블록으로 나눠져서 여러대의 데이터노드에 분산되어 저장된다.

5)MapReduce

map과 reduce라는 두 개의 method로 구성됨. 대규모 분산 컴퓨팅 혹은 단일 컴퓨팅 환경에서 대량의 데이터를 병렬로 분석할 수 있는 알고리즘



6)JobTracker

하둡 클러스터에 등록된 전체 job의 스케줄링을 관리하고 모니터링하는 노드
전체 하둡 클러스터에서 하나의 JobTracker가 실행됨. 보통 하둡 클러스터의 네임노드(마스터)에서 실행됨

7)TaskTracker

JobTracker의 작업을 요청받고 JobTracker가 요청한 맵과 리듀스 개수만큼 Map Task와 Reduce Task를 생성함. 하둡 클러스터의 데이터노드에서 실행됨.

8)Mapper

맵리듀스 프로그래밍 모델에서 map method의 역할을 수행하는 클래스. 키와 값으로 구성된 입력 데이터를 전달받아 이 데이터를 가공하고 분류해서 새로운 데이터를 생성함

9)Reducer

맵리듀스 프로그래밍 모델에서 reduce method의 역할을 수행하는 클래스. map task의 출력 데이터를 입력 데이터로 전달받아 집계 연산을 수행

10)YARN(Yet Another Resouce Negotiator)

맵리듀스의 차세대 기술, 맵리듀스의 확장성과 속도 문제를 해소하기 위해 개발된 프로젝트

11)SSH(Secure Shell)

: NameNode는 하나지만, DataNode를 여러개이다. 따라서 노드간 안전한 통신이 되어야한다. 여러대의 PC들이 안전하게 데이터를 주고받기 위해서는 접속을 해야한다.
접속하려면, Login을 해야하고, 통신간 보안이 유지되어야 하기 때문에, SSH를 이용한다.

네트워크상의 다른 컴퓨터에 로그인하거나 원격 시스템에서 명령을 실행하고 다른 시스템으로 파일을 복사할 수 있게 해주는 응용 프로토콜, 기존의 telnet을 대체하기 위해 설계되었으며 암호화 기법을 사용하여 강력한 인증 방법 및 안전하지 못한 네트워크에서 안전하게 통신할 수 있는 기능을 제공함, 기본적으로 22번 포트를 사용함.

하둡에서는 SSH 프로토콜을 이용하여 하둡 클러스터 간의 내부 통신을 수행한다. 이 때 SSH를 이용할 수 없다면 하둡을 실행할 수 없게 된다. 따라서 네임노드에서 SSH 공개키를 설정하고 이 공개키를 전체 서버에 복사하는 작업을 진행한다.

12)NoSQL(Not Only SQL)

관계형 데이터 모델과 SQL 문을 사용하지 않는 데이터베이스 시스템 혹은 데이터 저장소. 기존 RDBMS가 분산 환경에 적합하지 않기 때문에 이를 극복하기 위해 고안됨. row 단위가 아닌 집합 형태로 저장됨. 또한 Sharding(샤딩)이라는 기능이 있어서 데이터를 여러 서버에 분산하여 저장함. 기존 RDBMS처럼 완벽한 데이터 무결성을 제공하지는 않음. 기업의 핵심 데이터는 RDBMS를 이용하고 핵심은 아니지만 데이터를 보관하고 처리를 해야 하는 경우 NoSQL 이용.

MongoDB, HBase 등 다양한 솔루션이 있음.

9분까지 수강