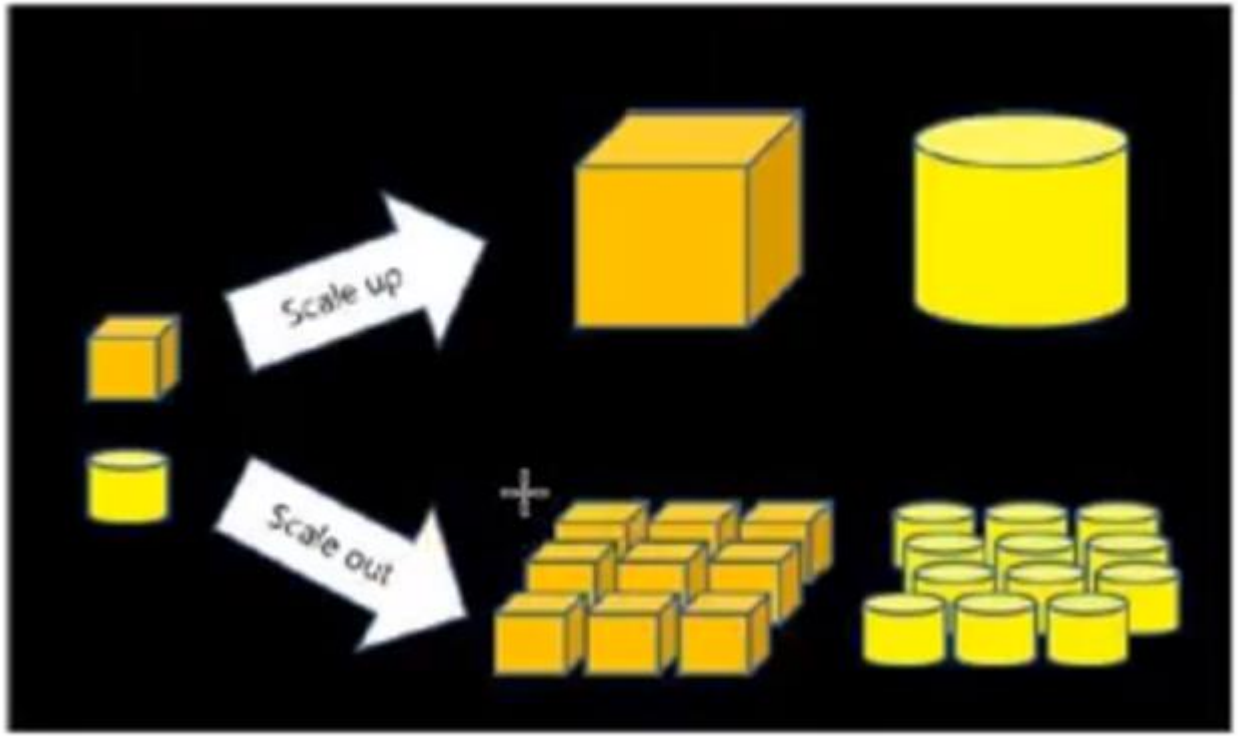
Hadoop

1. 빅데이터와 하둡

1. 빅데이터와 하둡

❖ BigData의 정의

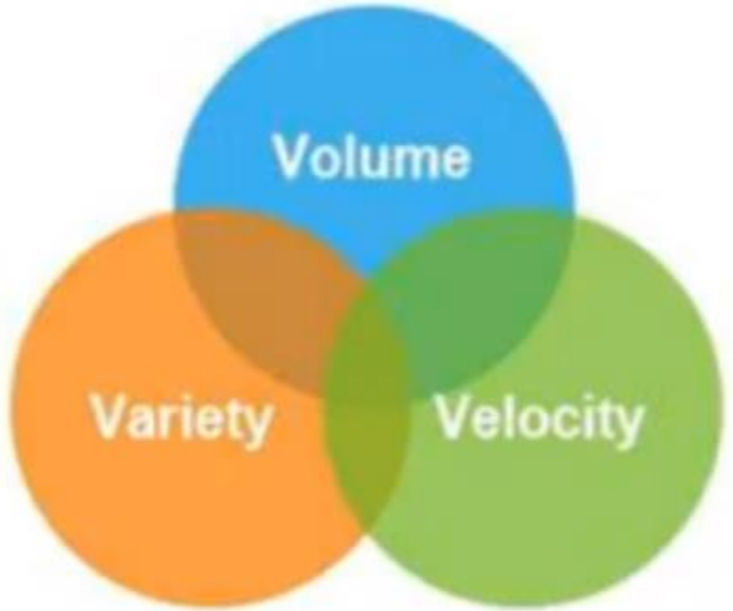
- 서버 한대로 처리할 수 없는 규모의 데이터(2012, John Rauser, 아마존 수석 엔지니어)
- 기본의 소프트웨어(RBMS 등)로 처리할 수 없는 규모의 데이터



1. 빅데이터와 하둡

❖ BigData 정의

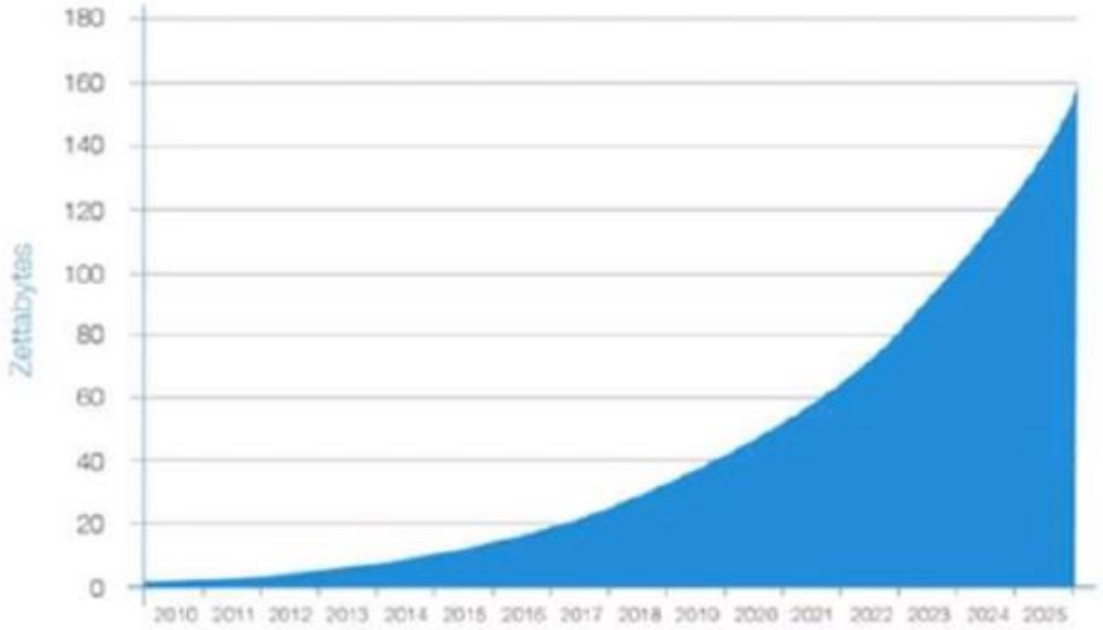
- 3Vs of BigData
  - Volume : 데이터의 크기(Tera Bythe, Peta Byte 단위)
    - 1 Tera Byte : 1024 GB
    - 1Peta Byte : 약 100만 GB(6GB DVD영화를 17만 4000편 저장)
  - Velocity : 데이터 생성 속도
  - Variety : 데이터의 다양성
    - 구조화, 비구조화 된 데이터를 모두 포함



1. 빅데이터와 하둡

❖ BigData의 예

- 웹 검색엔진 데이터
- 웹 페이지 데이터
  - 구글의 경우 수 조개의 웹페이지들의 정보를 수집하여 인덱스를 생성 함
  - 페이지당 4KB X 1조 페이지 = 4PB(Peta Byte)
- 다양한 Device에서 생성되는 데이터
  - 스마트폰, 스마트TV 등의 다양한 디바이스에서 생성되는 데이터가 2025년 163ZB예상
- 소셜미디어 데이터



[출처] 씨게이트가 IDC(IT 시장조사 기관)에 의뢰해 발행한 Data Age 2025 백서

KB < MB < GB <TB < PB(페타바이트) < EB(엑사바이트) < ZB(제타바이트) <YB(요타바이트)

1. 빅데이터와 하둡

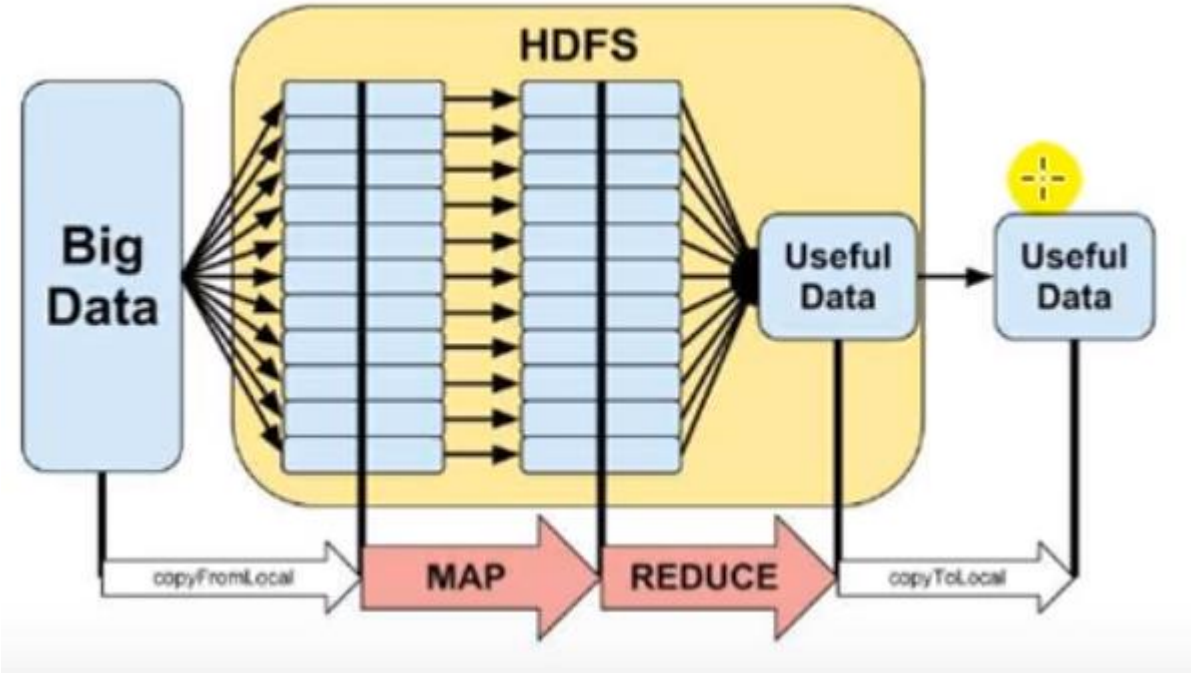
❖ Hadoop이란?

- 대용량 데이터를 분산처리 할 수 있는 자바 기반의 오픈소스 Framework
- 더그 커팅이 구글 논문(2003년, “The Google File system”, 2004년 “MapReduce : Simplified Data Processing on Large Cluster”)을 참조하여 구현
- 더그 커팅의 아들이 노란 코끼리 장난감 인형을 hadoop이라고 부른 것을 듣고 명명
- 공식사이트 : http://Hadoop.apache.org

1. 빅데이터와 하둡

❖ Hadoop 관련 용어 정리

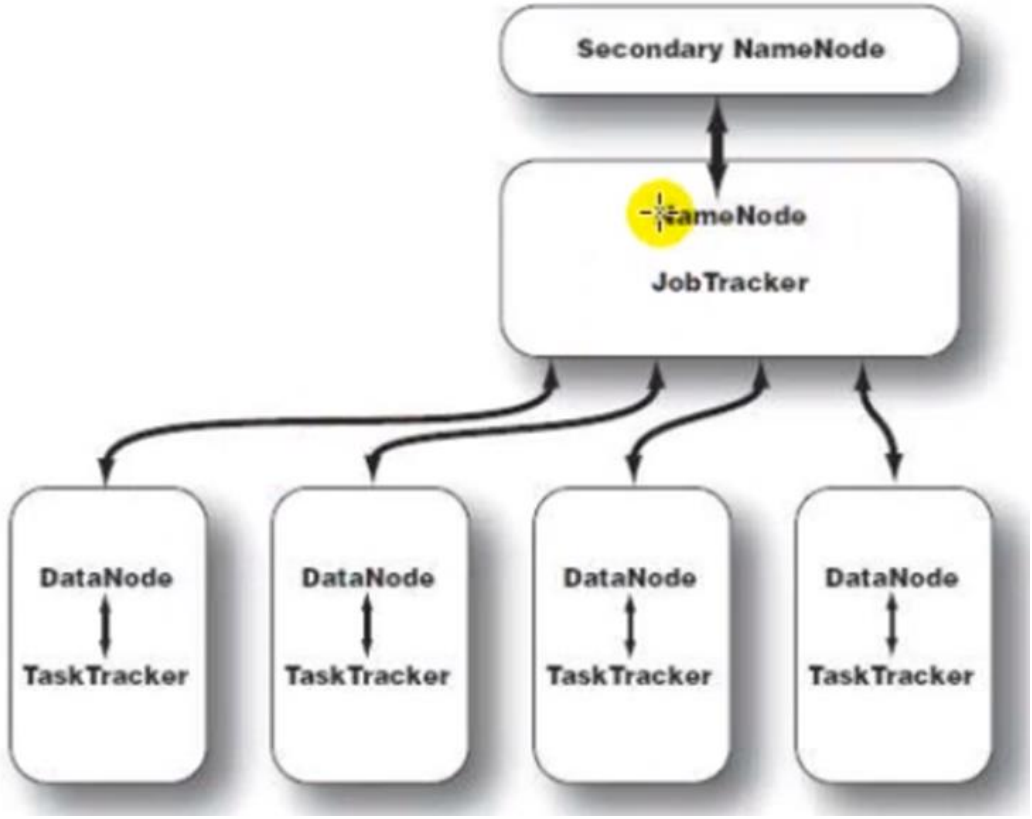
- HDFS(Hadoop Distribute File System)
  - 대용량 파일을 분산된 서버에 설치하고 많은 클라이언트가 저장된 데이터를 빠르게 처리할 수 있게 설계된 파일 시스템

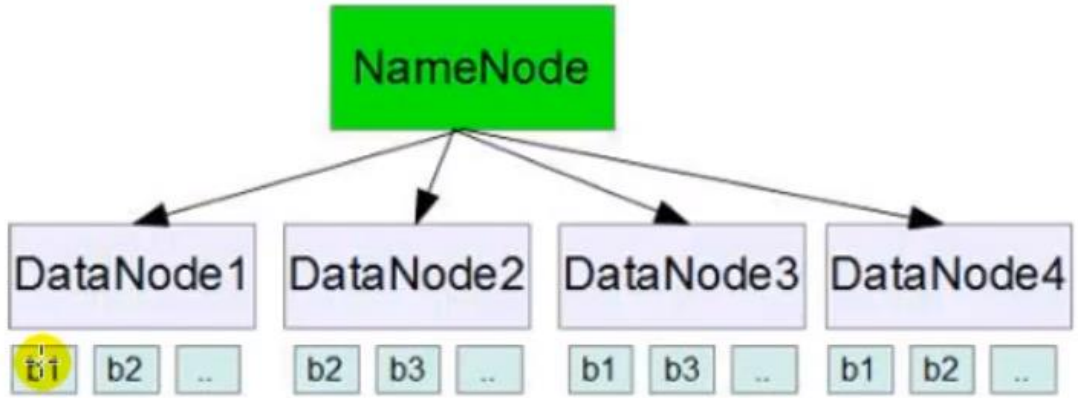


1. 빅데이터와 하둡

❖ Hadoop 관련 용어 정리

- NameNode
  - HDFS의 모든 메타데이터를 관리하고 클라이언트가 HDFS에 저장된 파일에 접근 할 수 있도록 처리하는 노드
- DataNode
  - HDFS에서 데이터를 입력하면 입력 데이터는 32MB의 블록으로 나눠져서 여러 대의 데이터 노드에 분산되어 저장
- Secondary NameNode
  - 주기적으로 네임노드의 파일 시스템 이미지 파일을 갱신하는 역할을 수행하는 노드

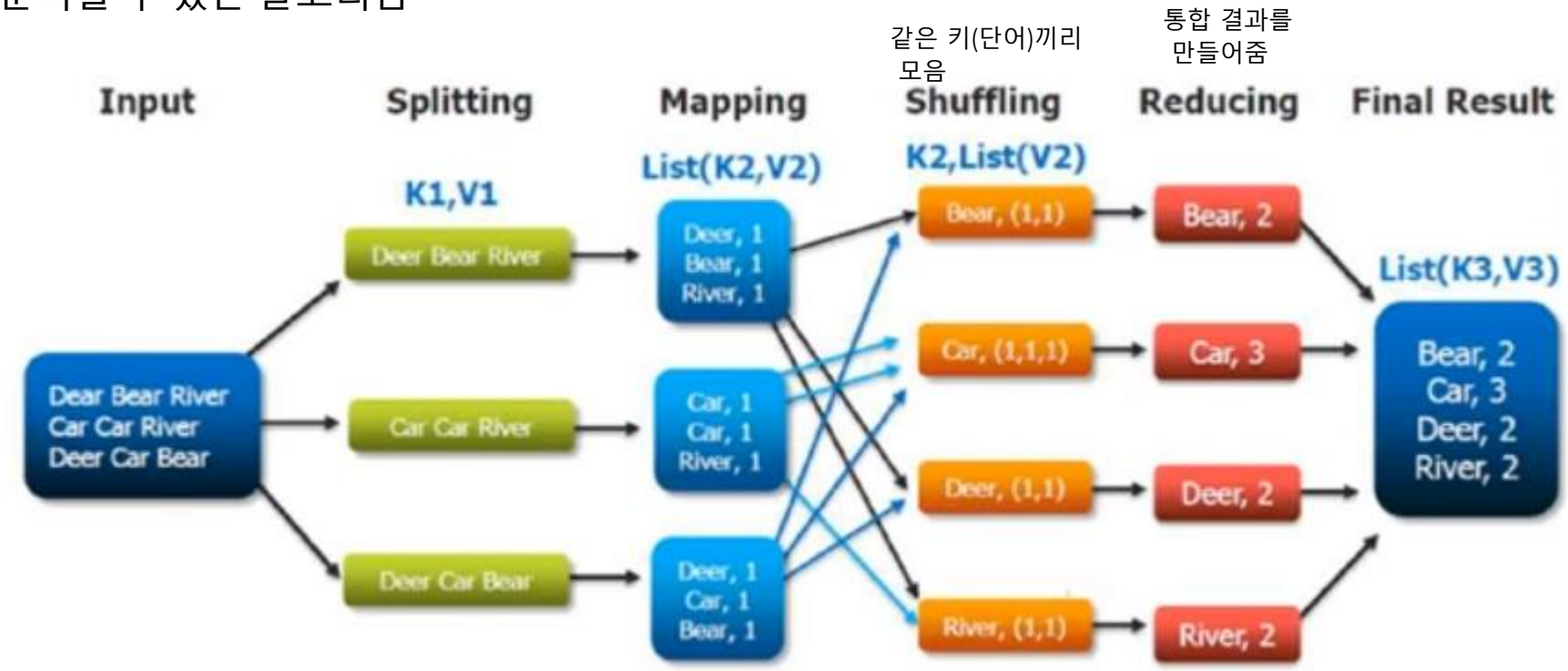




1. 빅데이터와 하둡

❖ Hadoop 관련 용어 정리

- MapReduce
  - Map과 reduce라는 두 개의 method로 구성된 대규모 컴퓨팅 혹은 단일 컴퓨팅 환경에서 대량의 데이터를 병렬로 분석할 수 있는 알고리즘



1. 빅데이터와 하둡

❖ Hadoop 관련 용어 정리

- JobTracker
  - 하둡 클러스터에 등록된 전체 job의 스케줄링을 관리하고 모니터링하는 노드
  - 전체 하둡 클러스터에서 하나의 JobTracker가 실행됨.
  - 보통 하둡 클러스터의 네임노드(마스터)에서 실행됨
- TaskTracker
  - JobTracker의 작업을 요청받고 JobTracker가 요청한 맵과 리듀스 개수만큼 Map Task와 Reduce Task를 생성함
  - 하둡 클러스터의 데이터노드에서 실행됨
- Mapper
  - 맵리듀스 프로그래밍 모델에서 map metho의 역할을 수행하는 클래스 키와 값으로 구성된 입력 데이터를 전달받아 이 데이터를 가공하고 분류해서 새로운 데이터를 생성함
- Reducer
  - 맵리듀스 프로그래밍 모델에서 reduce methid의 역할을 수행하는 클래스. map task의 출력 데ㅣ이터로 전달받아 집게 연산을 수행

1. 빅데이터와 하둡

❖ Hadoop 관련 용어

- YARN(Yet Another Resource Negotiator)
  - 맵리듀스의 차세대 기술, 맵리듀스의 확장성과 속도 문제를 해소하기 위해 개발된 프로젝트
- SSH(Secure Shell)
  - 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 포트 사용
  - 하둡에서 SSH 프로토콜을 이용하여 하둡 클러스터 간의 내부 통신을 수행. 이때 SSH를 이용할 수 없다면 하둡을 실행할 수 없게 됨. 따라서 네임노드에서 SSH 공개키를 설정하고 전체 서버에 복사하는 작업을 진행

1. 빅데이터와 하둡

❖ Hadoop 관련 용어

- NoSQL(Not Only SQL)
  - 관계형 데이터 모델과 SQL문을 사용하지 않는 데이터베이스 시스템 혹은 데이터 저장소
  - 기존 RDBMS가 분산 환경에 적합하지 않기 때문에 이를 극복하기 위해 고안됨
  - row 단위가 아닌 집합 형태로 저장함
  - 기존 RDBMS처럼 완벽한 데이터 무결성을 제공하지 않음
  - 기업의 핵심 데이터는 RDBMS를 이용하고 핵심은 아니지만 데이터를 보관하고 처리를 해야하는 겨우 NoSQL 이용
  - MongoDB, HBase 등 다양한 솔루션 제공