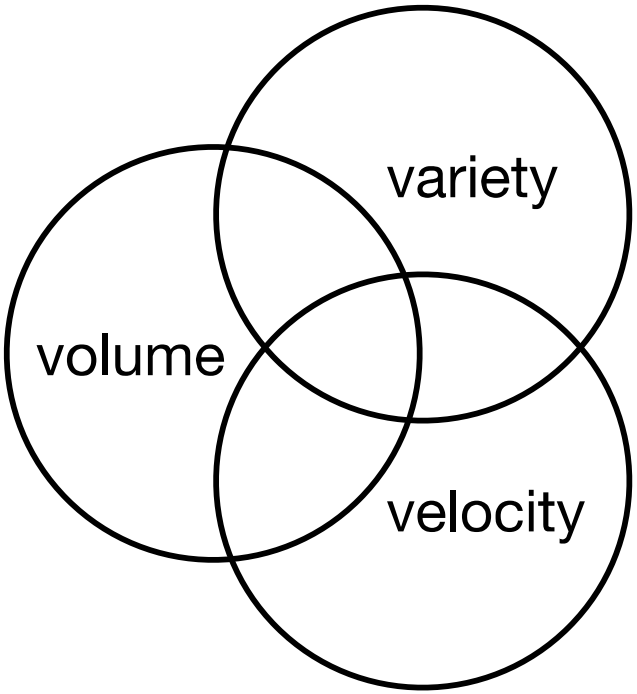
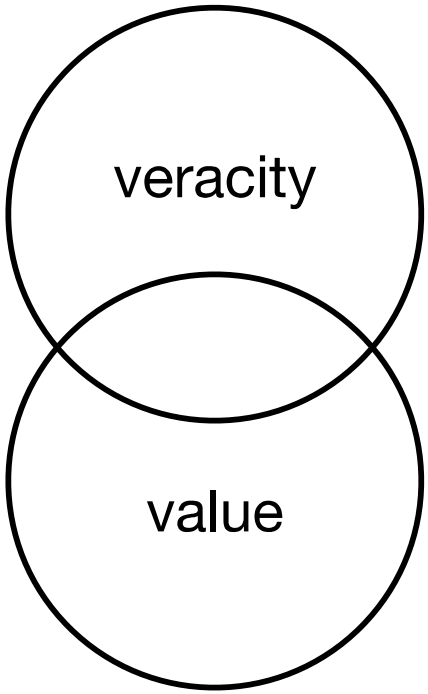
BigData

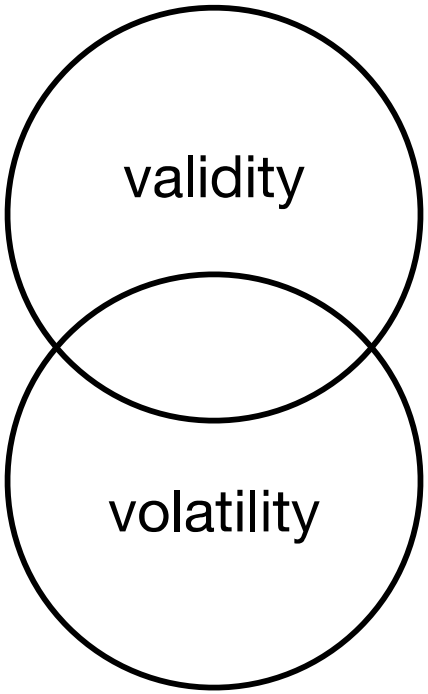
1.bigdata

3v + 2v + 2v

visualization







1.bigdata

3v

volume (규모)　　데이터의 크기

MegaByte → GigaByte → TeraByte → PetaByte → ExaByte → ZettaByte

variety (형태)　　다양한 데이터

database → Web, Photo, Audio → Social, Video, unStructured

velocity (속도)　　생산, 처리, 분석 속도

batch → periodic → near RealTime → realTime

1.bigdata

5v

veracity (정확성)　　데이터의 품질, 값의 신뢰성

데이터의 결측치, 이상치 등

value (가치)　　데이터를 통한 가치 창출

비즈니스, 연구에 도움이 되는 데이터

1.bigdata

7v

validity (타당성)　목표에 일치하는 데이터

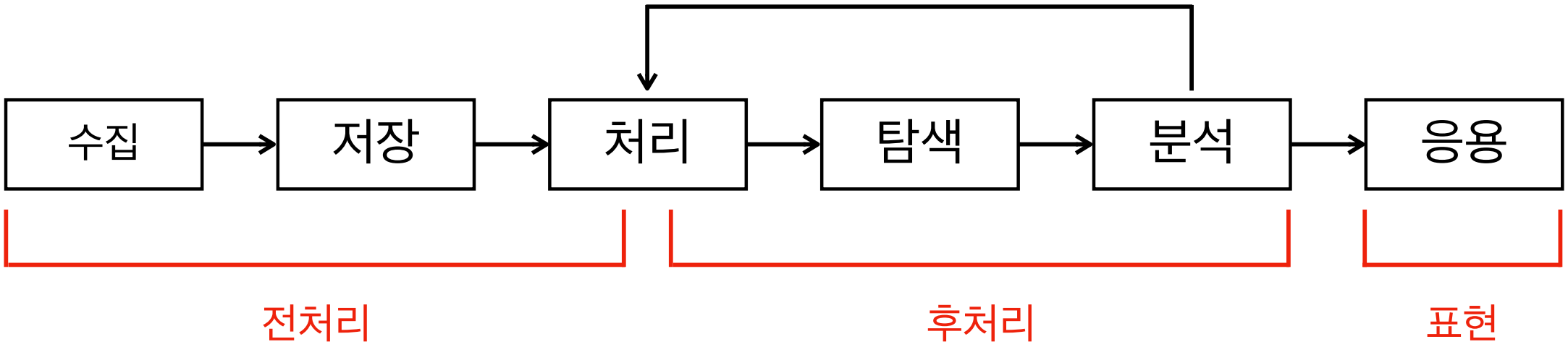
얼마나 정확하게 데이터를 가져왔는지

volatility (휘발성)　데이터의 유효기간

언제까지 사용할 수 있는지

1.bigdata

architecture



* 비정형 데이터는 구조화(결측치, 이상치 등 정제) 필요

1.bigdata

architecture

수집	내 / 외부 데이터 연동 및 통합
저장	대용량 / 실시간 데이터 분산 저장
처리	데이터 선택, 변환, 통합, 축소
탐색	데이터 질의
분석	통계 분석
응용	시각화

2.hadoop

정의

대용량 데이터를 분산 저장 및 처리할 수 있는 자바 기반의 오픈소스 프레임워크

구글이 논문으로 발표한 Google FileSystem 및 MapReduce 구현

여러 대의 서버에 데이터를 분산 저장

각 서버에 분산되어 있는 데이터를 동시 처리

데이터를 복제하여 저장 (데이터 유실 시 복구 용이)

2.hadoop

modules

Commons	다른 모듈을 연결 및 지원하는 기본 모듈
HDFS	대용량 데이터 분산 파일 시스템
MapReduce	데이터셋 병렬 처리
Yarn	작업 예약 및 리소스 관리

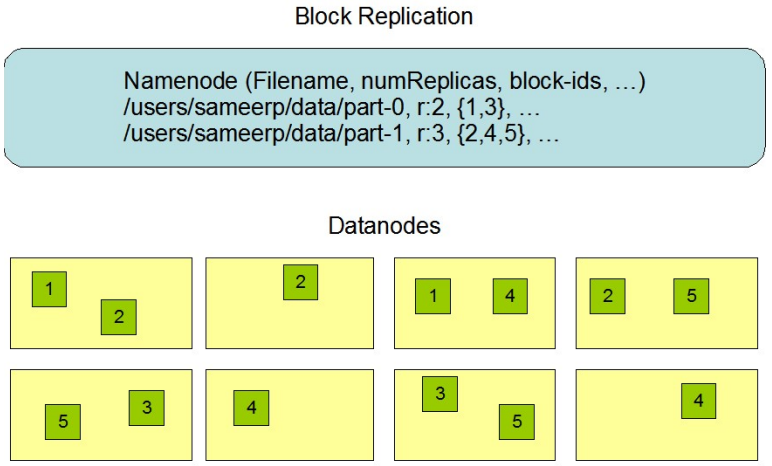
2.hadoop

Hadoop Distributed File System

google file system 을 기반으로 만든 대용량 분산 저장 / 처리 파일 시스템

NameNode 와 DataNode 를 가지는 Master - Slaver Architecture

Block 구조 파일 시스템



2.hadoop

NameNode	메타데이터 관리	데이터노드 모니터링
	블록 관리	클라이언트 요청 접수
Secondary NameNode	체크포인트 노드 (fsimage + edit)	
	네임스페이스 동기화	
DataNode	데이터 저장	

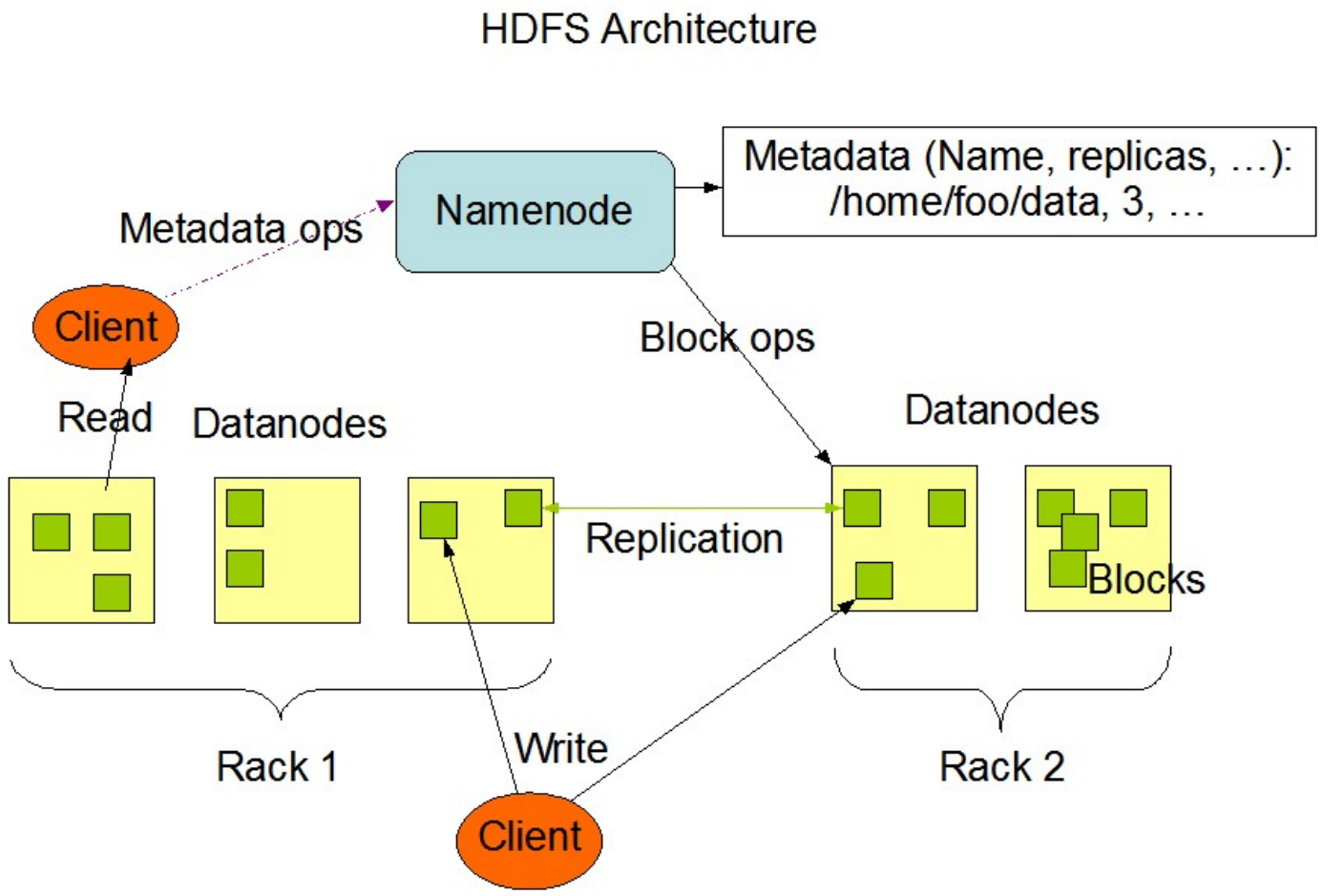
2.hadoop

장애 복구

스트리밍 방식의 데이터 접근

대용량 데이터 저장

데이터 무결성



2.hadoop

map reduce

대량의 데이터를 병렬로 분석 → 분산 처리 지원

함수형 프로그래밍 + 분산 컴퓨팅

데이터 전송, 분산 병렬 처리 등은 MapReduce Framework 가 자동으로 처리
→ 개발자는 MapReduce 알고리즘에 맞게 분석프로그램 개발