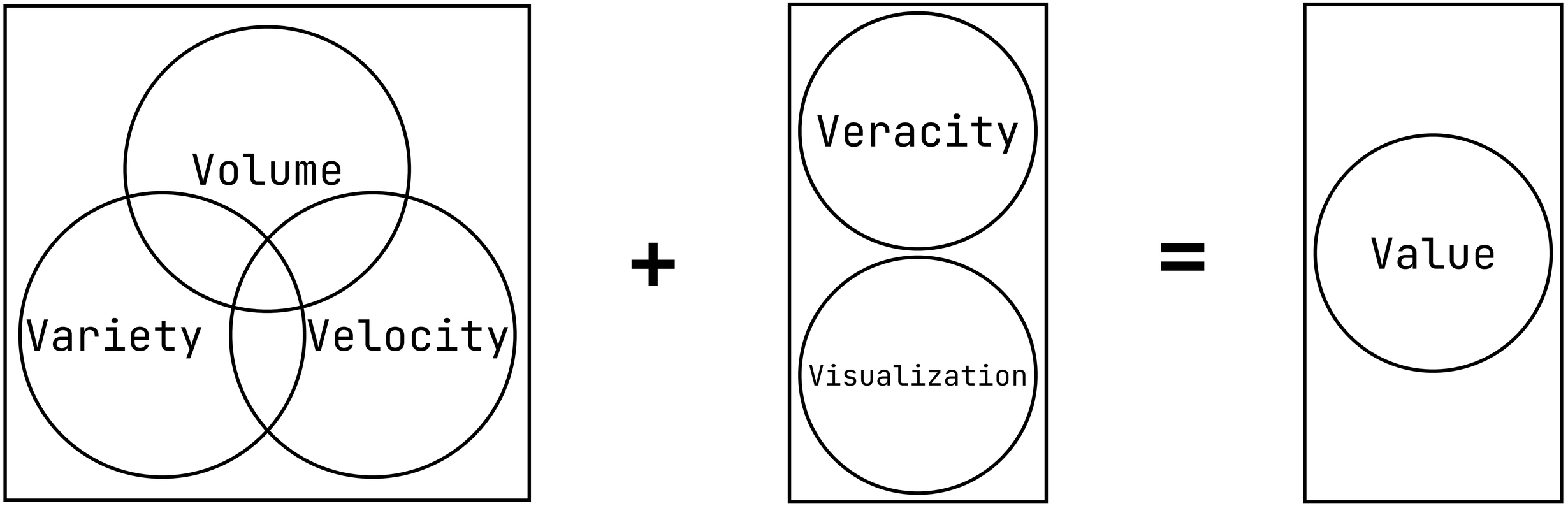
BigData

0.BigData

정의



0.BigData

정의

Velocity (속도) : 생산, 처리, 분석 속도

batch → periodic → near RealTime → realTime

Variety (형태) : 다양한 데이터

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

Volume (규모) : 데이터 크기

MegaByte → GigaByte → TeraByte → PetaByte → ExaByte

0.BigData

정의

Veracity (정확성) : 데이터 품질, 값의 신뢰성

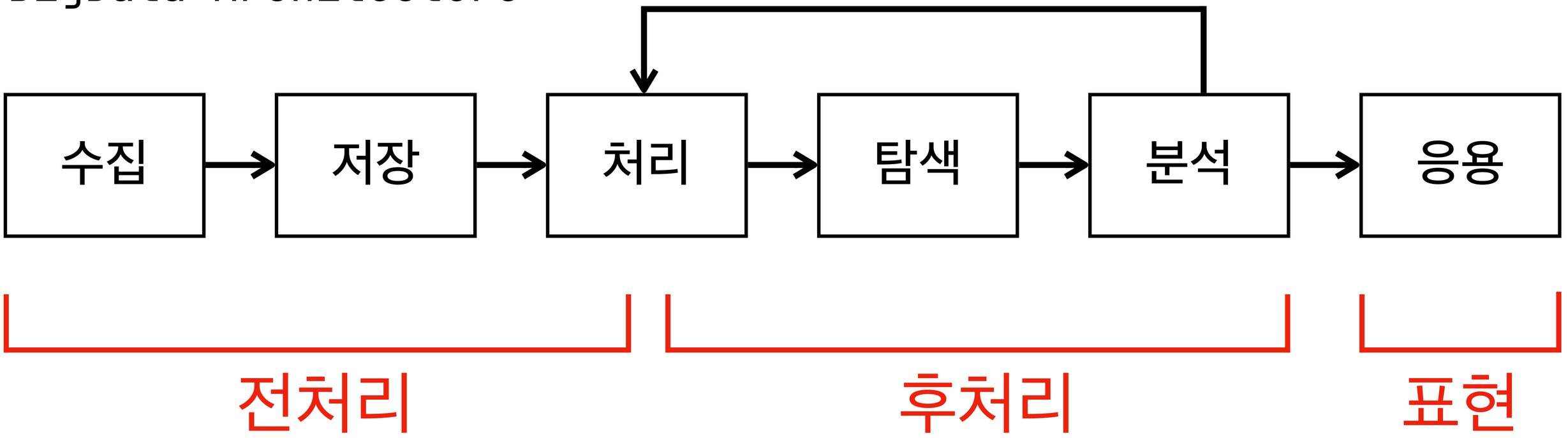
Visualization (시각화) : 복잡한 대규모 데이터를 시각적으로 표현

Value (가치) : 데이터를 통한 가치 창출

0.BigData

아키텍처

BigData Architecture



* 비정형 데이터는 구조화(결측치, 이상치 등 정제) 필요

0.BigData

아키텍처

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

전처리

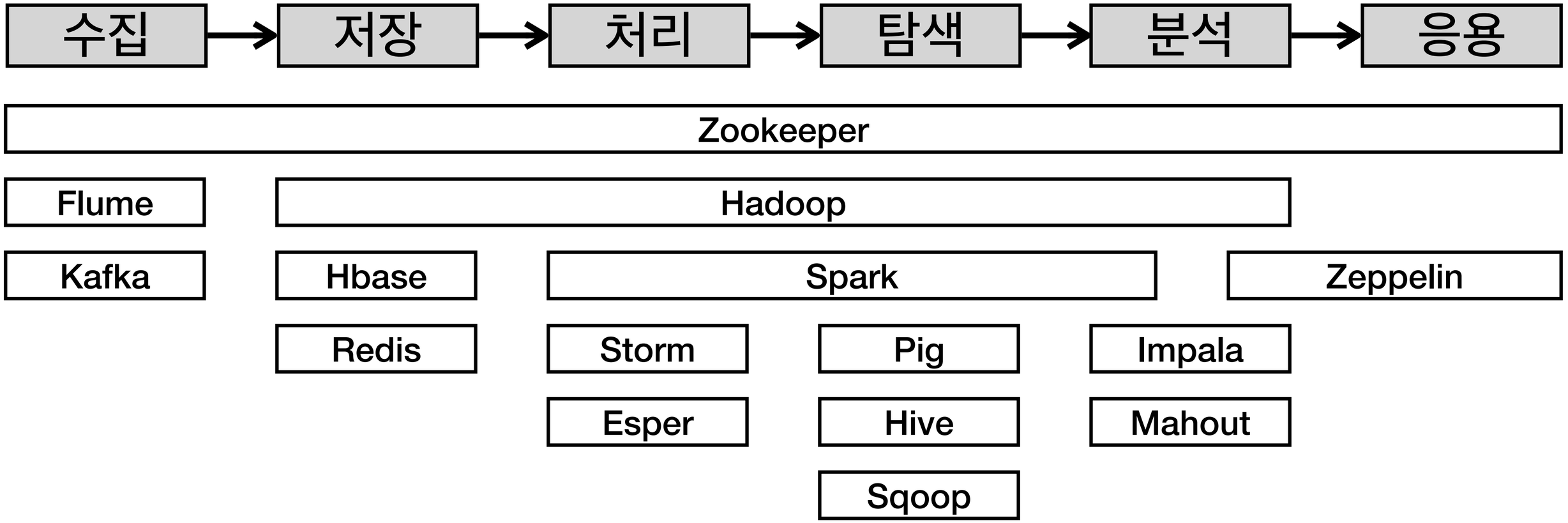
후처리

표현

0.BigData

hadoop ecosystem

Hadoop Ecosystem Architecture



0.BigData

hadoop ecosystem

Zookeeper	분산 환경에서 서버 간의 안정적인 분산 조정
Flume	대용량 로그데이터 수집
Kafka	데이터(메시지) 분산 스트리밍
Hadoop	대용량 데이터 분산 저장/처리 프레임워크
Hbase	google BigTable 기반 분산저장 DataBase

0.BigData

hadoop ecosystem

Redis	in memory DataBase
Spark	대용량 데이터 분산 처리/분석
Storm	실시간 데이터 연산
Esper	복잡한 이벤트 처리 (Complex event Processing)
Pig	복잡한 맵리듀스 프로그램 생성/병렬 분석 언어

0.BigData

hadoop ecosystem

Hive	분산 데이터 SQL을 처리하는 데이터 웨어하우스
Sqoop	데이터를 데이터저장소(RDBMS, NoSql 등) 에 신속 전송
Impala	실시간 SQL 병렬 처리
Mahout	분산 선형 대수 프레임워크 (MachineLearning library)
Zeppelin	데이터 분석/시각화를 위한 Web 기반 notebook

1.Hadoop

정의

대용량 데이터를 분산 처리할 수 있는 자바 기반의 오픈소스 프레임워크

- 구글이 논문으로 발표한 Google File System 및 MapReduce를 구현
- 여러 대의 서버에 데이터를 분산 저장
- 저장되어 있는 각 서버의 데이터를 동시 처리
- 데이터의 복제본을 저장하기에 데이터 유실 시 복구 용이

1.Hadoop

module

Hadoop Modules

- commons　　다른 hadoop module 을 지원하는 공통 모듈
- HDFS　　대용량 데이터 분산 파일 시스템
- MapReduce　　데이터 셋 병렬 처리
- Yarn　　작업 예약 및 리소스 관리

1.Hadoop

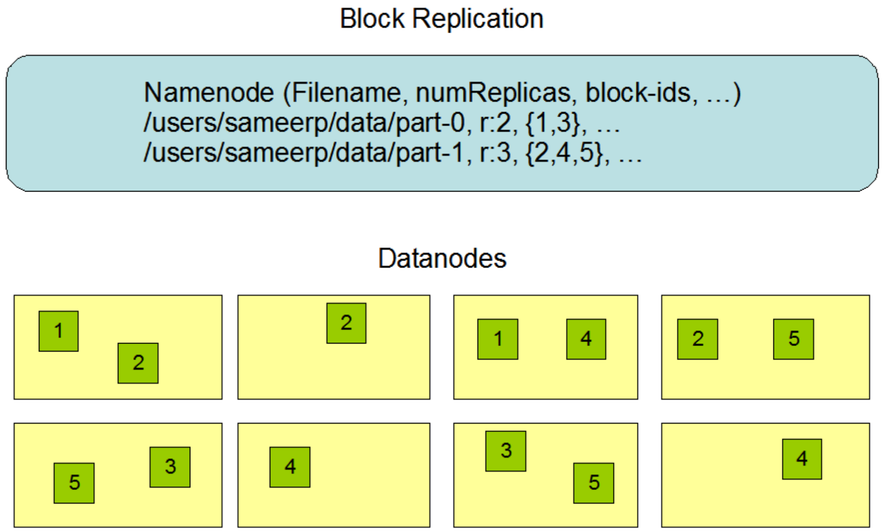
hdfs

HDFS (Hadoop Distributed File System)

- Google File System 을 기반으로 만든 대용량 분산 저장/처리 파일시스템

- NameNode 와 DataNode 를 가지는 Master-Slave Architecture

- Block 구조 파일 시스템



1.Hadoop

hdfs

NameNode

- 메타데이터 관리
- 데이터노드 모니터링
- 블록 관리
- 클라이언트 요청 접수

Secondary NameNode

- 체크포인트 노드
  fsimage + edit
- 네임스페이스 동기화

DataNode

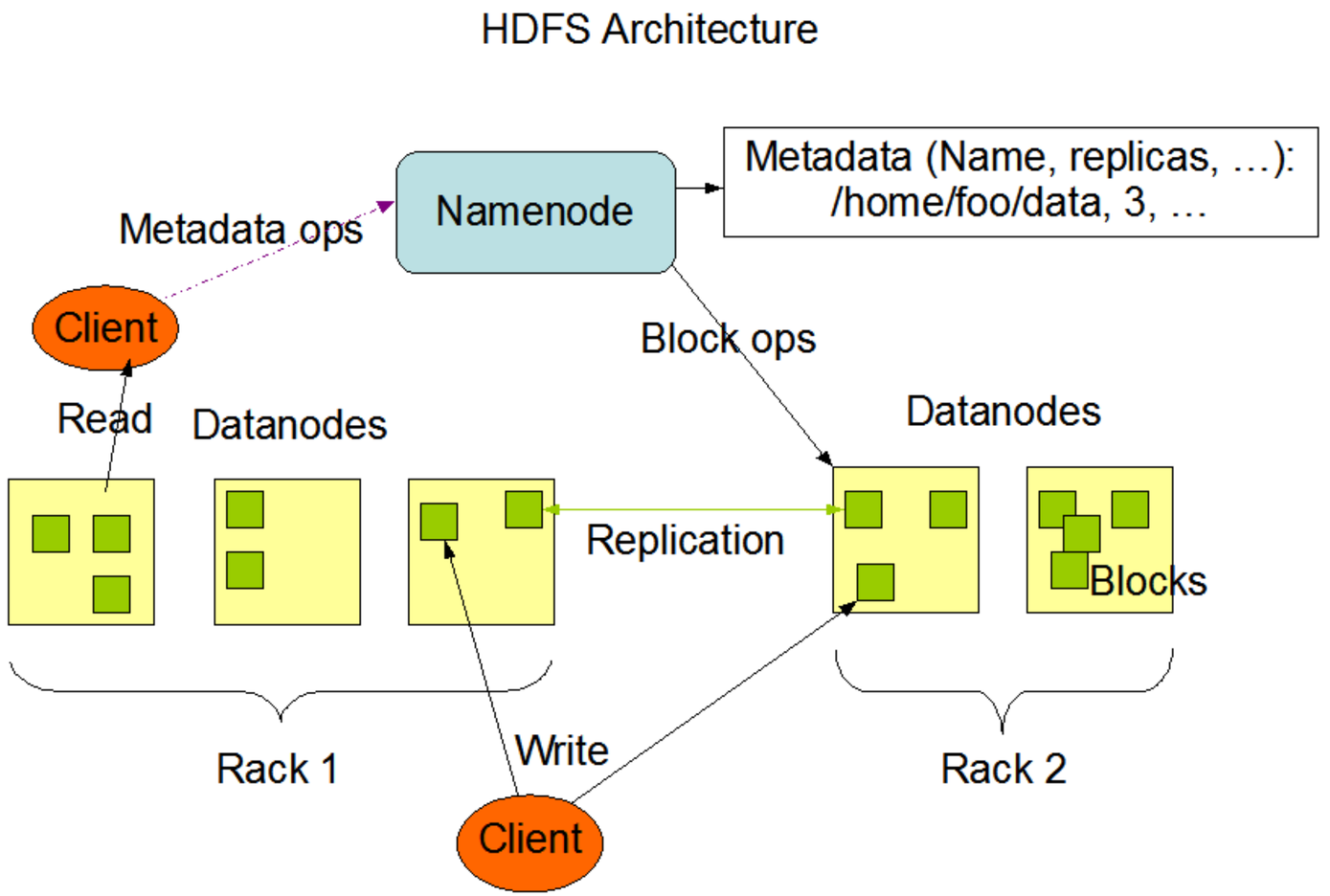
- 데이터 저장

1.Hadoop

hdfs

목적

- 장애 복구
- 스트리밍 방식의 데이터 접근
- 대용량 데이터 저장
- 데이터 무결성



1.Hadoop

mapreduce

Hadoop MapReduce

- 대량의 데이터를 병렬로 분석, 분산 처리 지원

- 함수형 프로그래밍 + 분산 컴퓨팅

- 데이터 전송, 분산 병렬 처리 등은 MapReduce Framework 가 자동 처리
  개발자는 MapReduce 알고리즘에 맞게 분석프로그램 개발

1.Hadoop

mapreduce

순서 : Map → Shuffle & Sort → Reduce

- Map : 데이터 변형 (transformation)
  입력 데이터를 split → key/value 해석 → 레코드를 map이 받아서 처리

- Shuffle : Map Task에서 처리된 데이터를 정렬 후 Reducer 로 전달

- Reduce : 데이터 집계 (aggregation)
  Mapper가 처리한 결과 집계

1.Hadoop

mapreduce

MapReduce Framework

- JobClient Hadoop MapReduce API

- JobTracker Job scheduling. TaskTracker에 Job 할당
  일반적으로 NameNode 에서 실행

- TaskTracker 요청된 Job을 받아 MapReduce(Task) 실행
  JobTracker에게 Heartbeat 전송
  DataNode 에서 실행

1.Hadoop

yarn

Hadoop 1.x MapReduce의 문제점

- JobTracker가 스케줄링 + 태스크 관리 기능을 수행
  작업 관리와 자원 분배의 비효율성

- JobTracker 미 실행시 TaskTracker 실행해도 MapReduce 불가

- MapReduce API 로 개발된 애플리케이션만 실행 가능

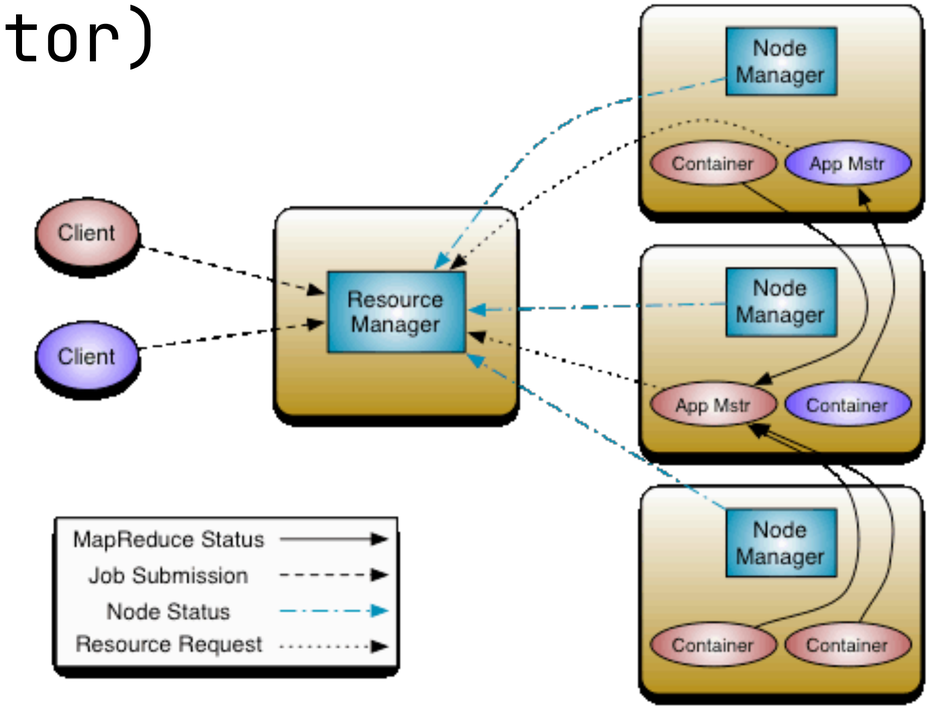
1.Hadoop

yarn

YARN (Yet Another Resource Negotiator)

- 리소스 관리 및 작업 예약/모니터링

- JobTracker를 분리
  ResourceManager : 리소스 관리
  ApplicationMater : 스케줄링
  JobHistoryServer : 이력관리



- TaskTracker 의 단일 계산 리소스 관리 부분은 NodeManager 가 담당

1.Hadoop

yarn

- ResourceManager YARN application 시작, DataNode 리소스 할당
- ApplicationMater Task scheduling 및 실행 관리
- JobHistoryServer 모든 Job의 메타데이터 관리
- NodeManager container 단위로 단일 서버 리소스 관리