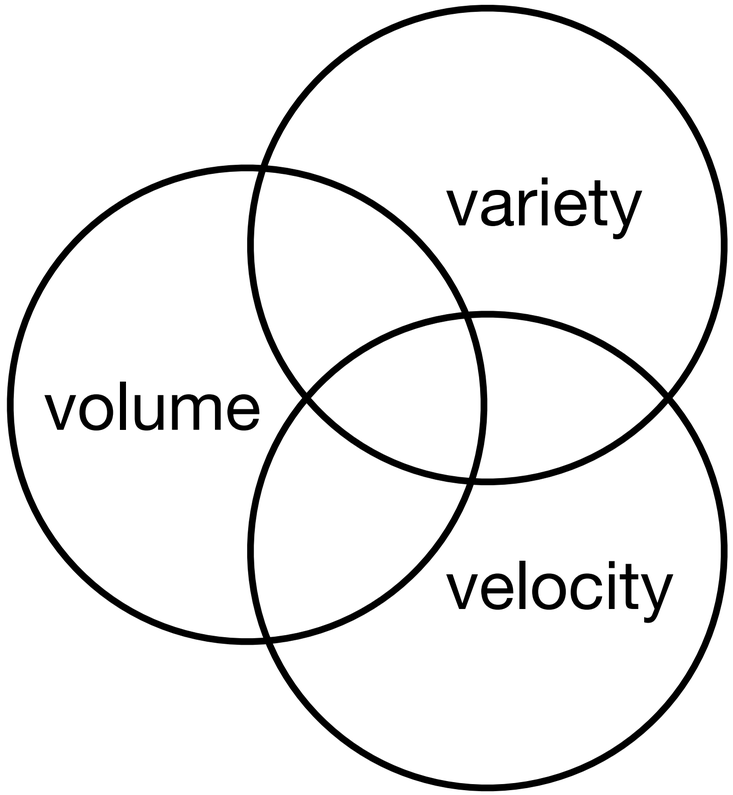
BigData

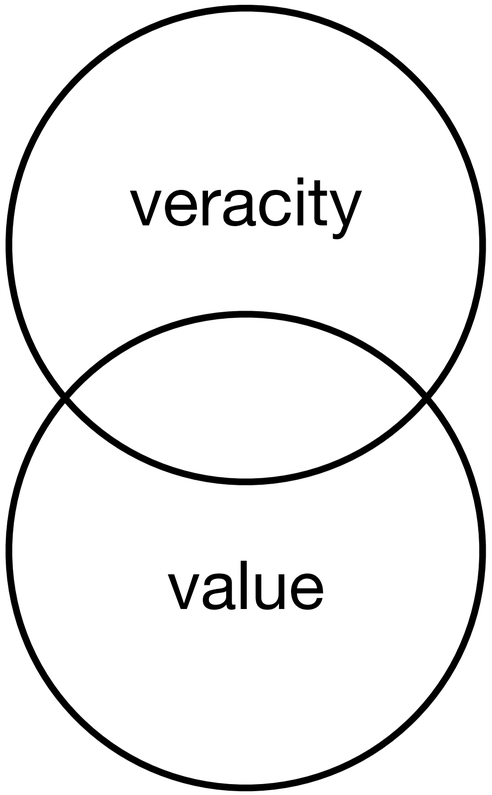
Big data primarily refers to data sets that are too large or complex to be dealt with by traditional data-processing application software.

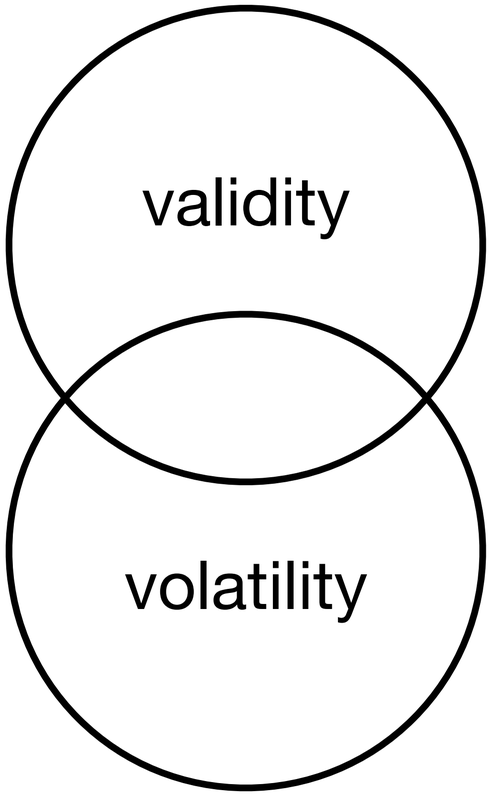
1.bigdata

3v + 2v + 2v

visualization







1.bigdata

3v + 2v + 2v

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

3v + 2v + 2v

veracity (정확성)

데이터의 품질, 값의 신뢰성

데이터의 결측치, 이상치 등

value (가치)

데이터를 통한 가치 창출

비즈니스, 연구에 도움이 되는 데이터

1.bigdata

3v + 2v + 2v

validity (타당성)

목표에 일치하는 데이터

얼마나 정확하게 데이터를 가져왔는지

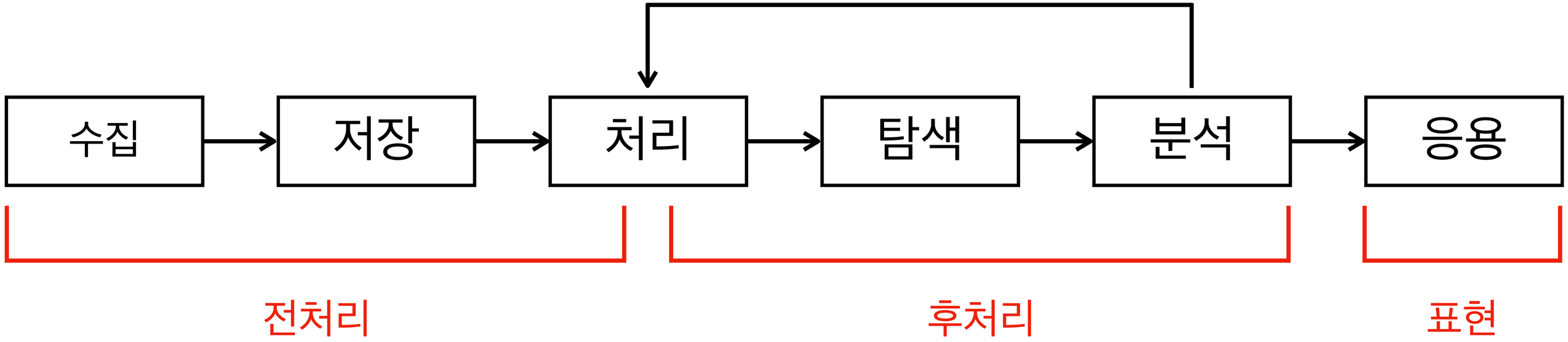
volatility (휘발성)

데이터의 유효기간

언제까지 사용할 수 있는지

1.bigdata

architecture



* 비정형 데이터는 구조화(결측치, 이상치 등 정제) 필요

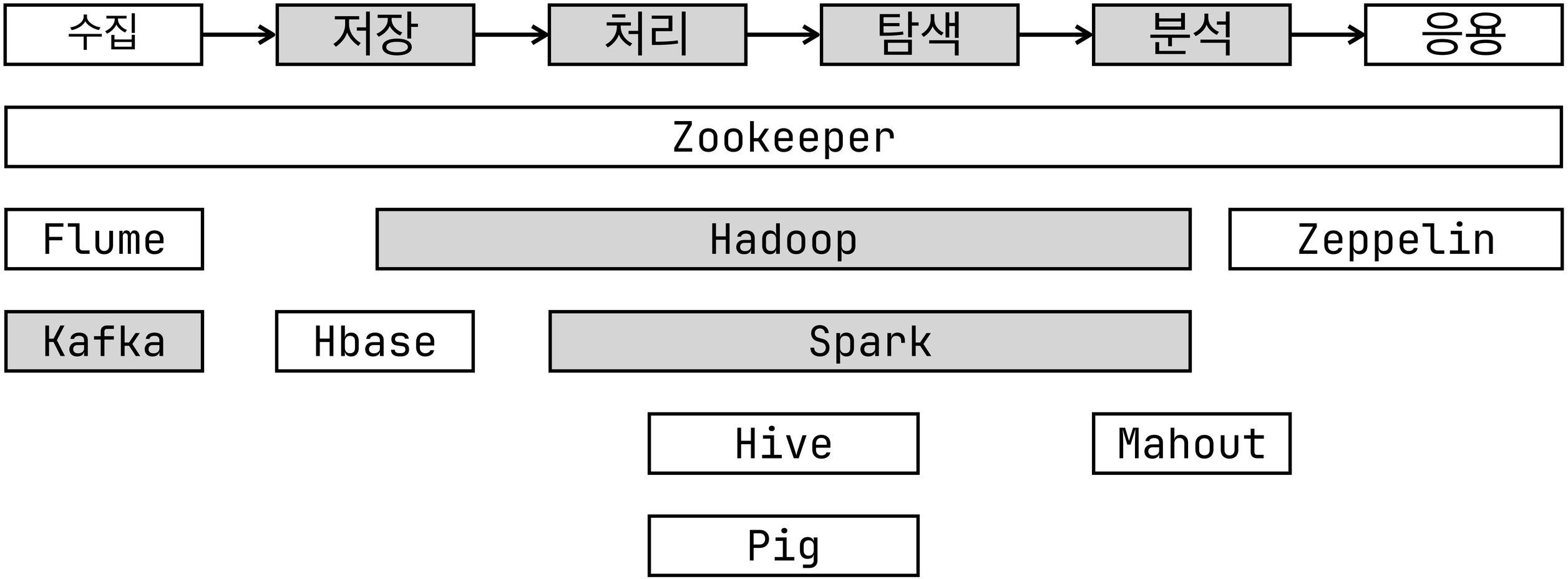
1.bigdata

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

1.bigdata

hadoop ecosystem



hadoop ecosystem : hadoop + subprojects

1.bigdata

hadoop ecosystem

Zookeeper	분산 환경에서 서버 간 안정적인 분산 조정
Flume	대용량 로그데이터 수집
Kafka	실시간 메시지 분산 스트리밍
Hbase	google Bigtable 기반 분산저장 DataBase
Hive	분산 데이터 처리 SQL을 지원하는 데이터 웨어하우스
Pig	복잡한 맵리듀스 프로그램 생성 및 병렬 분석 언어
Mahout	분산 선형대수 프레임워크 (머신러닝 라이브러리)
zeppelin	데이터 분석 및 시각화를 위한 웹 기반 노트북

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

module

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

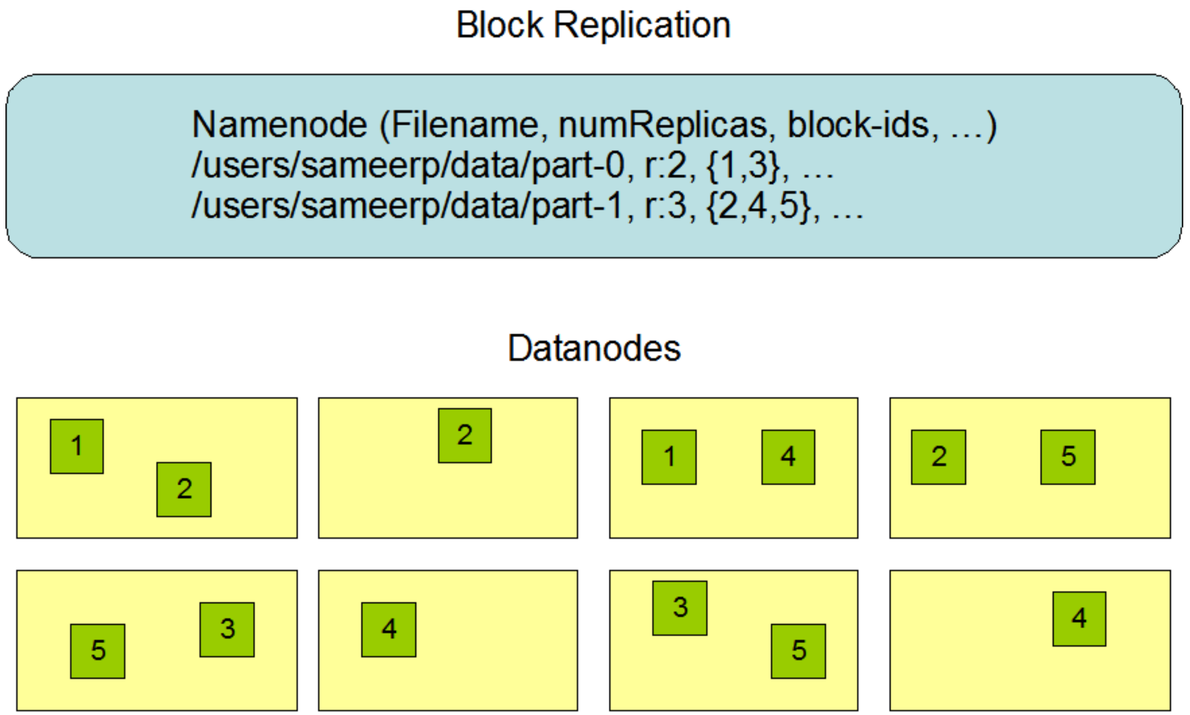
hdfs

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

hdfs

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

2.hadoop

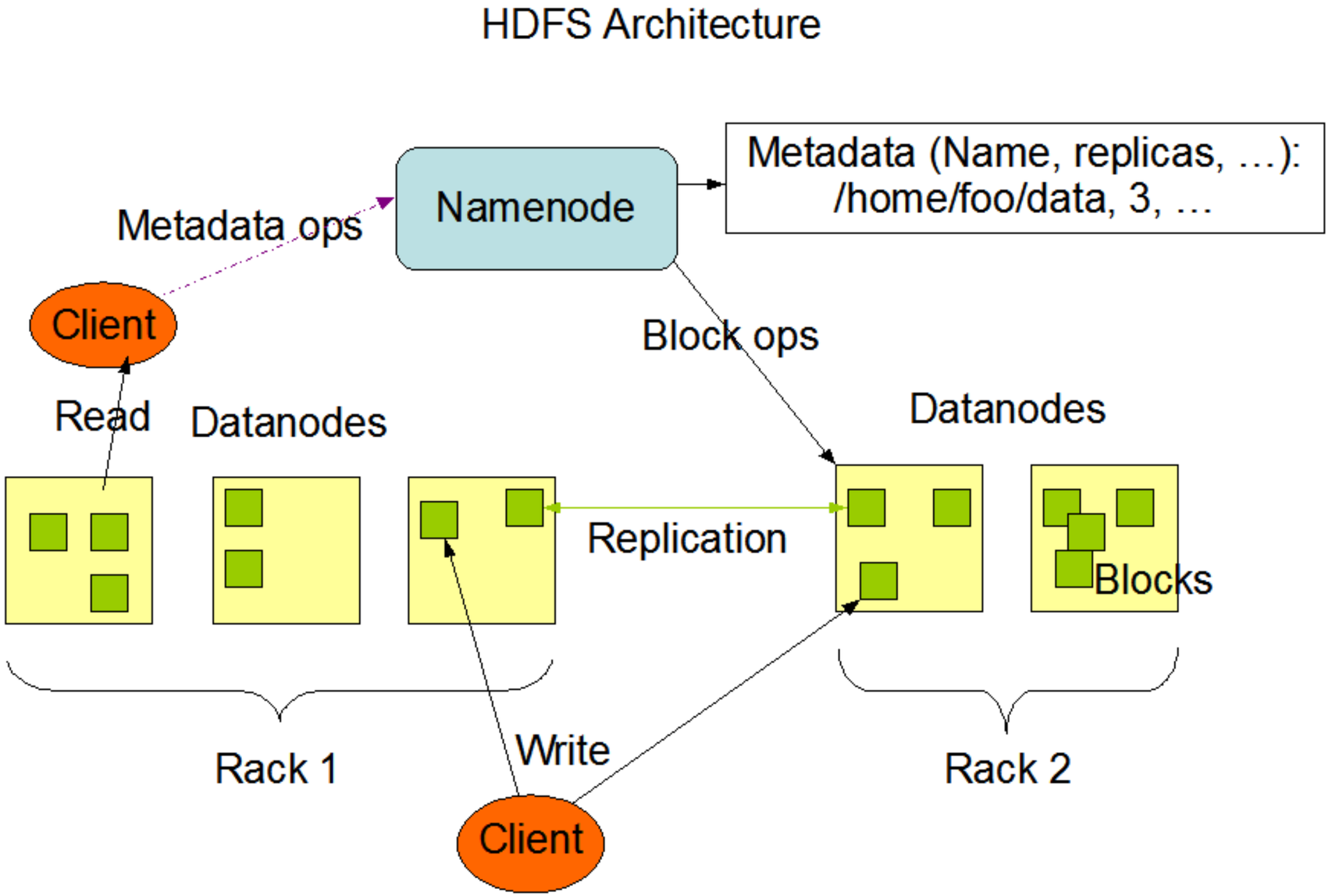
hdfs

장애 복구

스트리밍 방식의 데이터 접근

대용량 데이터 저장

데이터 무결성



2.hadoop

mapreduce

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

2.hadoop

mapreduce

Map → Suffle & Sort → Reduce

Map (Data Transformation)
입력 데이터를 split → key / value 해석 → 레코드를 map이 받아서 처리

Shuffle & Sort
Map Task 에서 처리된 데이터를 섞어서 정렬 후 Reducer로 전달

Reduce (Data Aggregation)
처리되어 전달된 결과 집계

2.hadoop

mapreduce framework

JobClient	Hadoop MapReduce API
JobTracker	JobScheduling 및 TaskTracker 에 Job 할당 (일반적으로 NameNode 에서 실행)
TaskTracker	요청된 Job 을 받아 MapReduce(Task) 실행 JobTracker 에게 HeartBeat 전송 DataNode 에서 실행

2.hadoop

hadoop 1.x 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

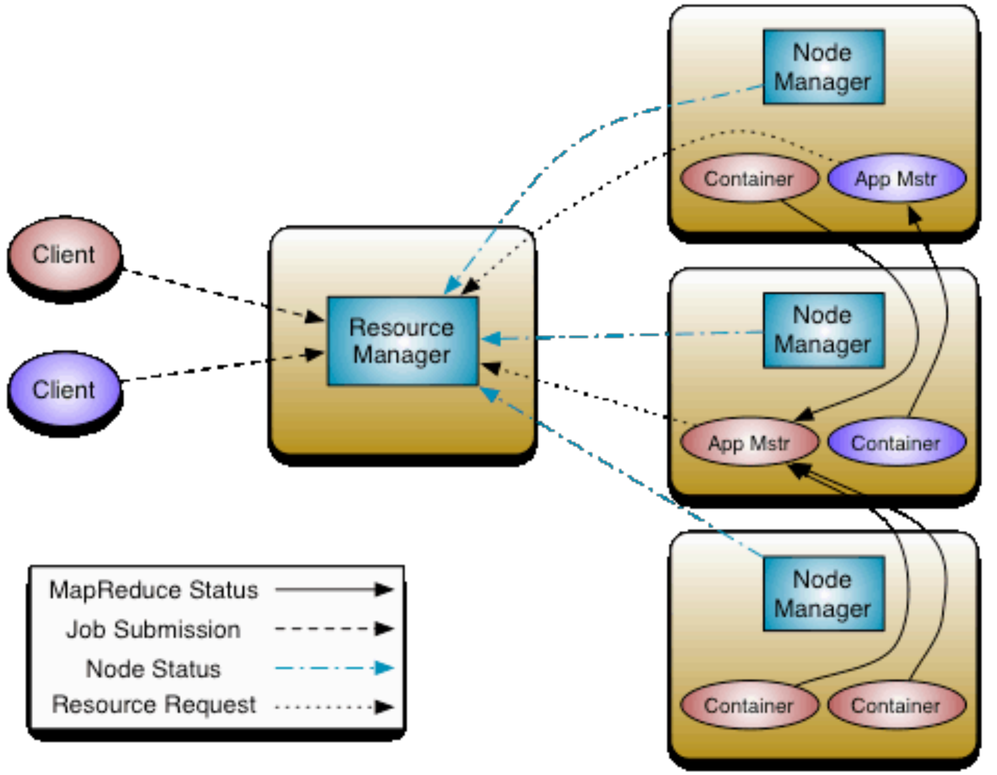
2.hadoop

yarn

Yet Another Resource Negotiator

리소스 관리 및 작업 예약/모니터링

JobTracker를 분리 ResourceManager
ApplicationMater
JobHistoryServer



TaskTracker 의 단일 계산 리소스 관리 부분은 NodeManager 가 담당

2.hadoop

yarn

ResourceManager	Yarn Application 시작 / DataNode 리소스 할당
ApplicationManager	Task Scheduling 및 실행 관리
JobHistoryServer	모든 Job 에 대한 metadata 관리
NodeManager	Container 단위로 단일 서버 리소스 관리

3.spark

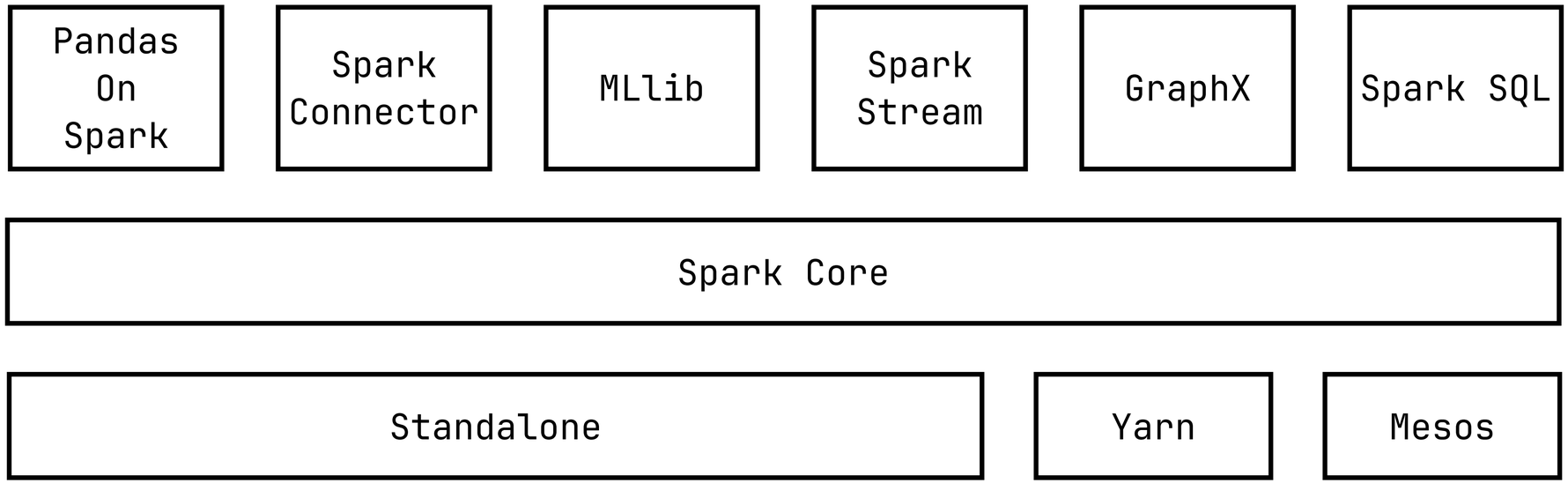
정의

apache spark : 빅데이터 처리를 위한 오픈소스 분석엔진

성능과 편의성 모두 고려

in-memory 방식으로 인한 빠른 속도

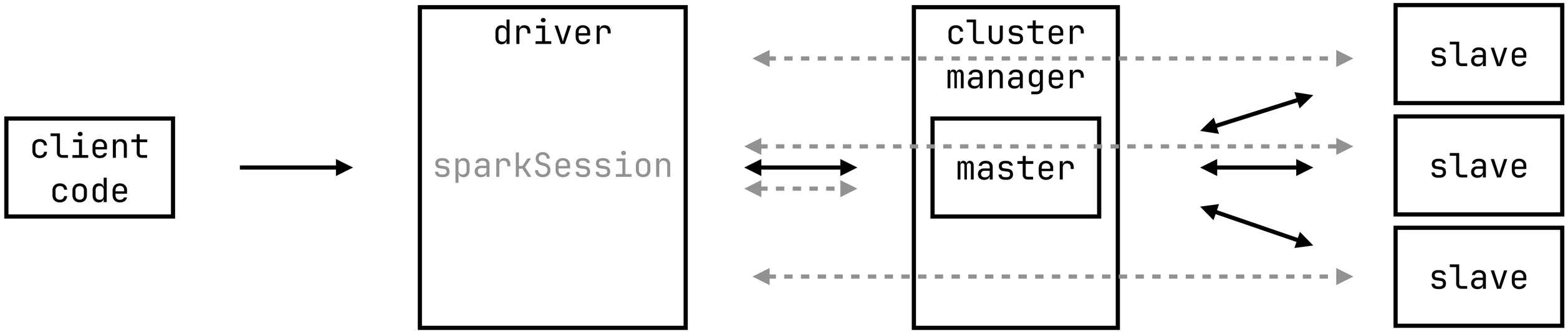
lazy-evaluation (lazy-execution)



3.spark

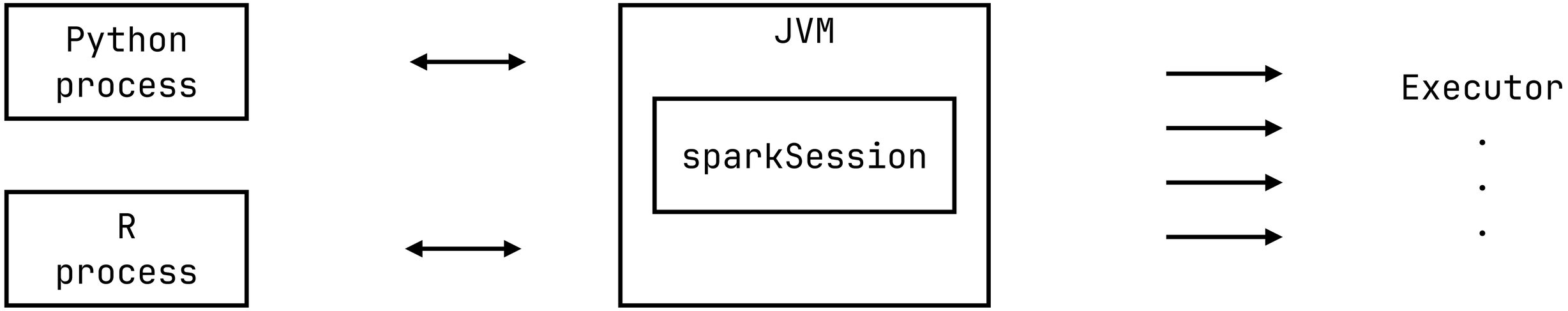
구성요소

driver	sparkSession 생성, main 호출, 프로그램 스케쥴링, 분석 및 배포 등
cluster manager	executor(master, slave) 관리
executor	프로세스가 할당한 job 수행
sparkSession	client의 명령을 cluster(executor)에서 실행



3.spark

구성요소



* spark 는 기본적으로 scala (jvm 기반)

3.spark

data structure

RDD — Resilient Distributed Dataset (내결함성 분산 데이터셋)

DataFrame — row 와 column의 개념을 가지는 자료구조
DataSet[Row] (Row 타입으로 구성된 Dataset)
비타입형 api - 런타임 시 데이터의 타입을 확인

DataSet — java와 scala만 사용 가능한 자료구조
타입형 api - 컴파일 시 데이터의 타입을 확인

3.spark

spark core

main component (기본 기능)

배치 단위의 RDD (Resilient Distributed Dataset) 연산 및 처리

* RDD (내 결함성 분산 데이터셋) : 장애 복구에 용이

3.spark

rdd

transformation	filter	reduceByKey
	map	sortBy
	flatmap	glom
	zip	...
action	collect	reduce
	first	fold
	stats	aggregate
	take	...

3.spark

dataframe

row + column 형식을 가지는 분산 테이블 형태의 컬렉션

schema	dataframe 의 column, dataType 등을 정의
execution ploan	연산이 데이터에 적용되는 순서 정의
ansi sql	ansi 표준 sql 을 사용하여 데이터 질의 가능
	spark.sql(“ansi sql”)

3.spark

spark sql

query 작업을 가능하게 해주는 라이브러리

Ansi 표준 sql, Hive sql 등의 문법을 사용하여 데이터 관리

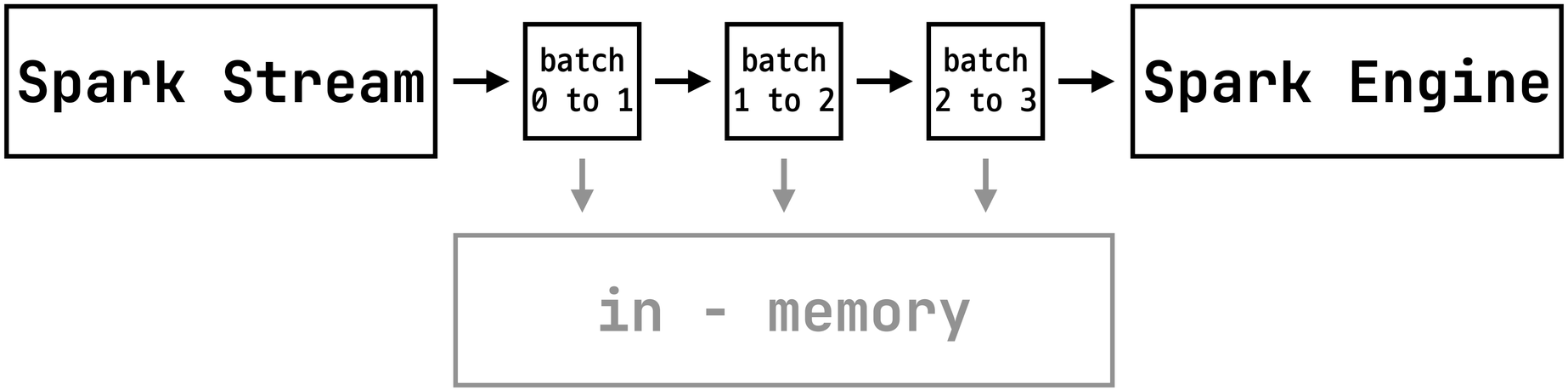
RDD-Like type 사용 (DataFrame, DataSet)

3.spark

spark stream

스트림 데이터를 mini-batch size로 분리 → RDD와 유사하게 처리

DStream(Discretized Stream) : 데이터 추상화
(데이터의 연속적인 흐름)



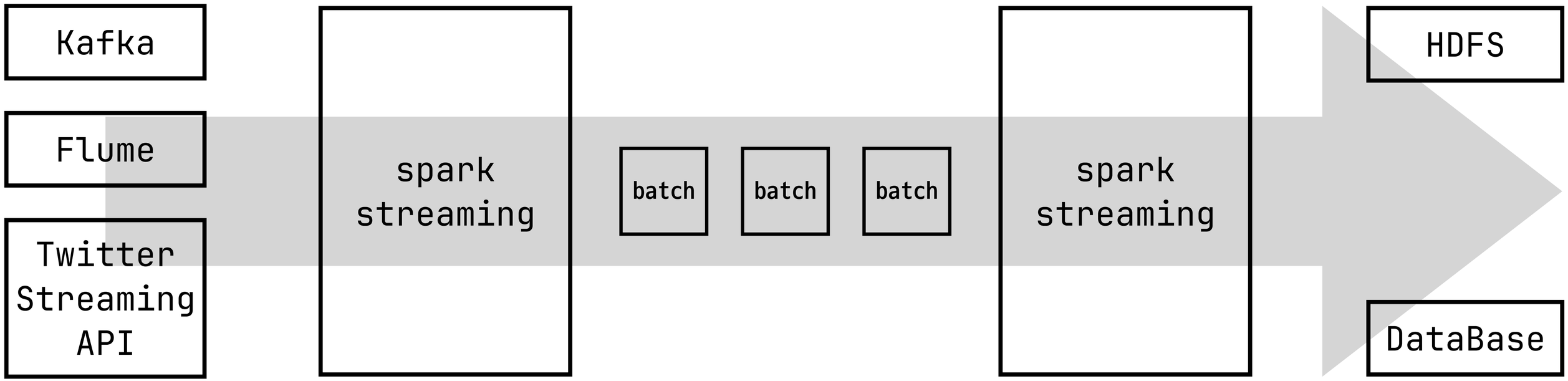
3.spark

dstream

연속적인 데이터 스트림에서 생성된 연속적인 RDD 시퀀스

소켓, 메시징 시스템, 스트리밍 api 등의 데이터 소스를 받아서 생성

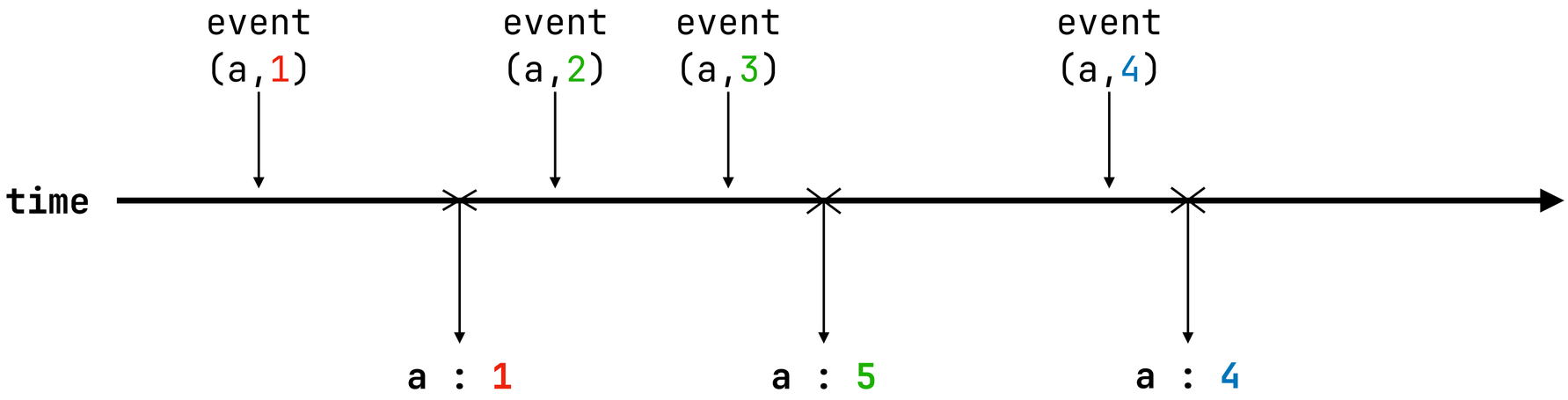
RDD가 작성되는 것과 같은 방식으로 입력 데이터 저장 (내부적으로 RDD)



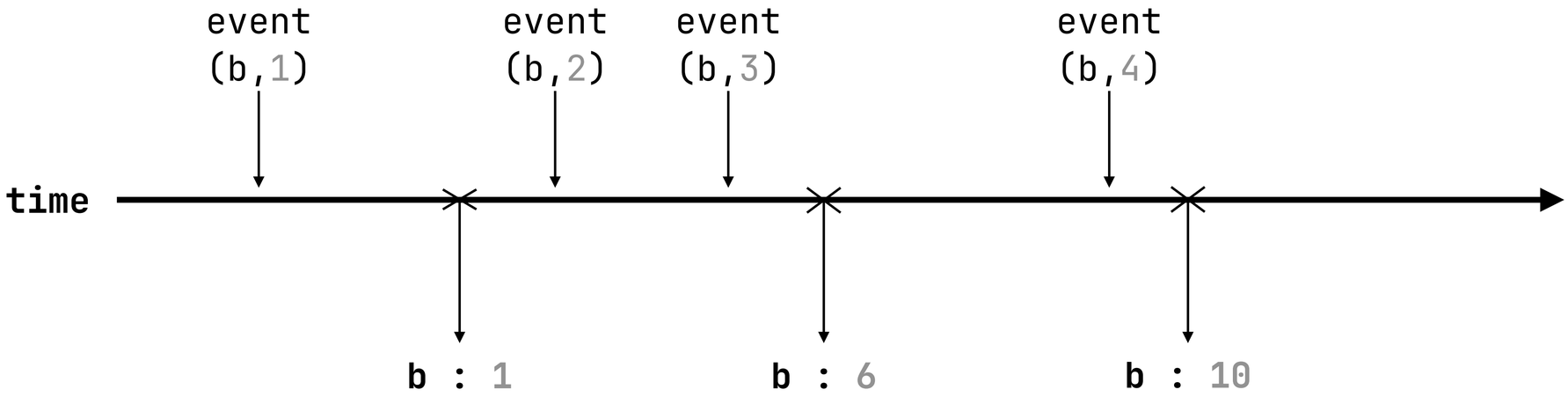
3.spark

dstream

tumbling window



sliding window



4.kafka

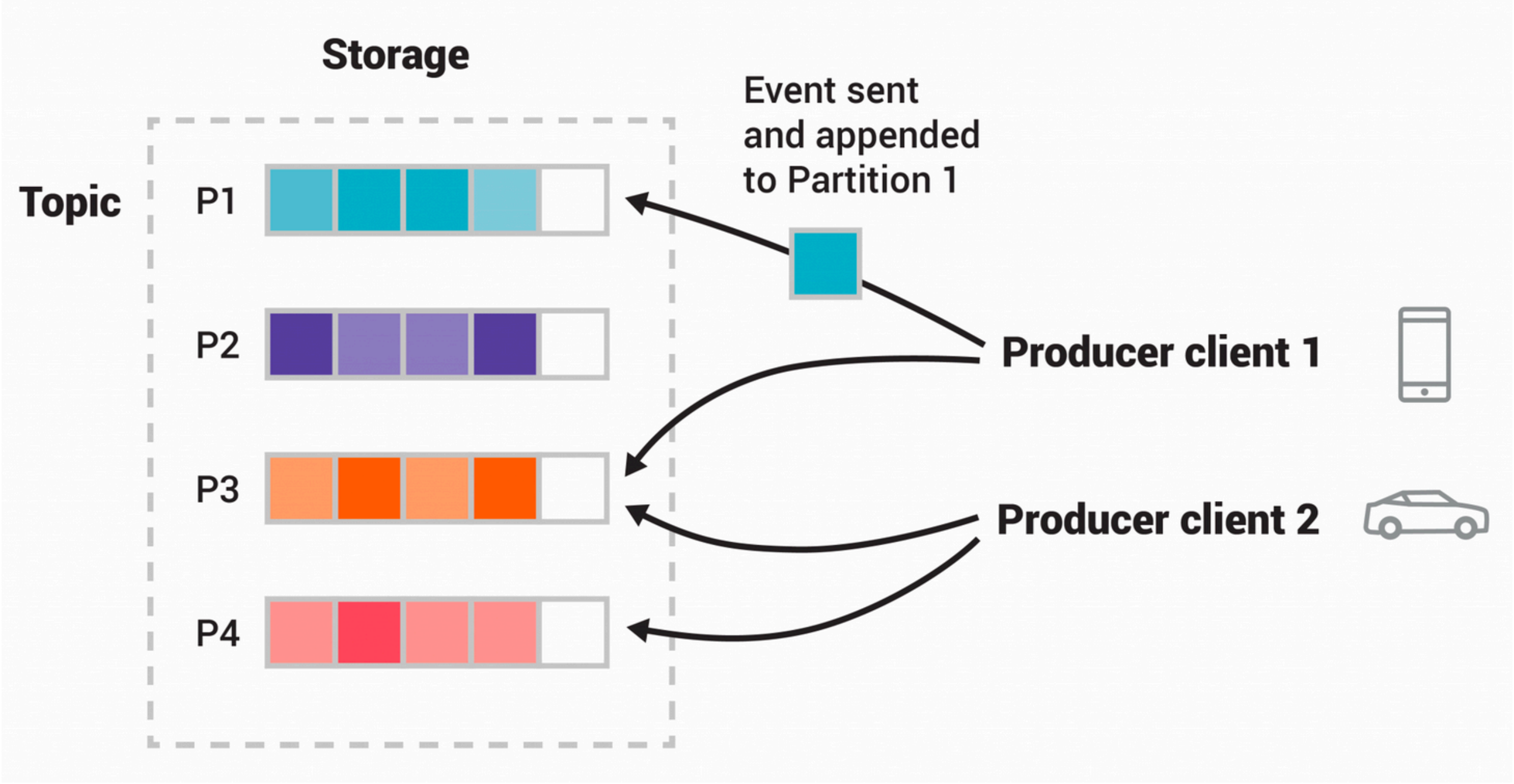
개념

apache kafka : event streaming platform

event streaming
이벤트 소스에서 이벤트 스트림의 형태로 데이터를 실시간으로 캡처, 저장, 처리

pub/sub model
publisher가 topic을 발행하면 subscriber가 구독하는 형태

4.kafka


Storage
Topic
P1
P2
P3
P4
Event sent
and appended
to Partition 1
Producer client 1
Producer client 2

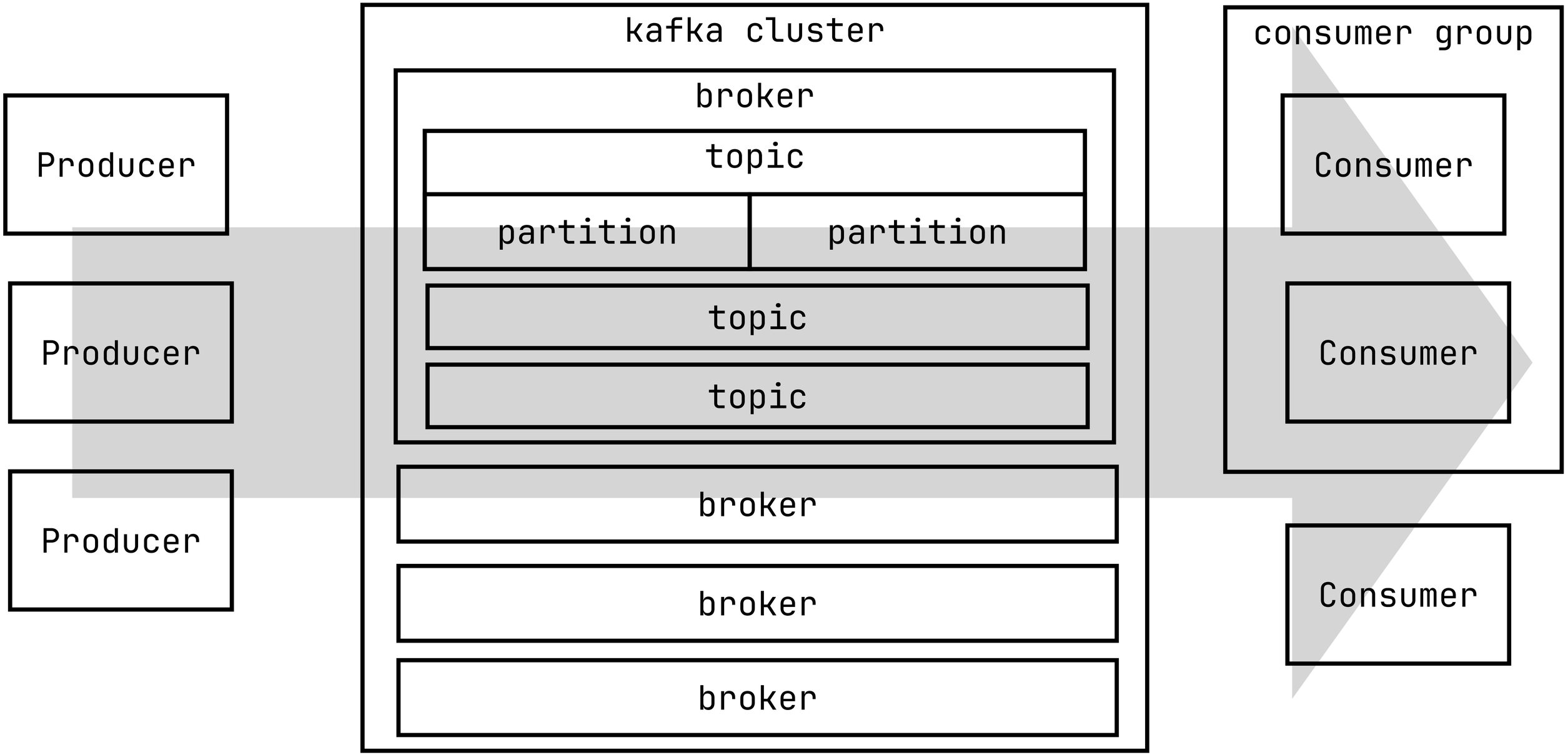
4.kafka

용어

event	(=record, message) 실시간 데이터
topic	데이터를 저장하는 논리적 이름 (partition 단위로 구성)
broker	(message broker) Kafka server를 의미. topic 저장
producer	event를 생성하고, topic을 발행
consumer	topic을 구독하여 event를 소비

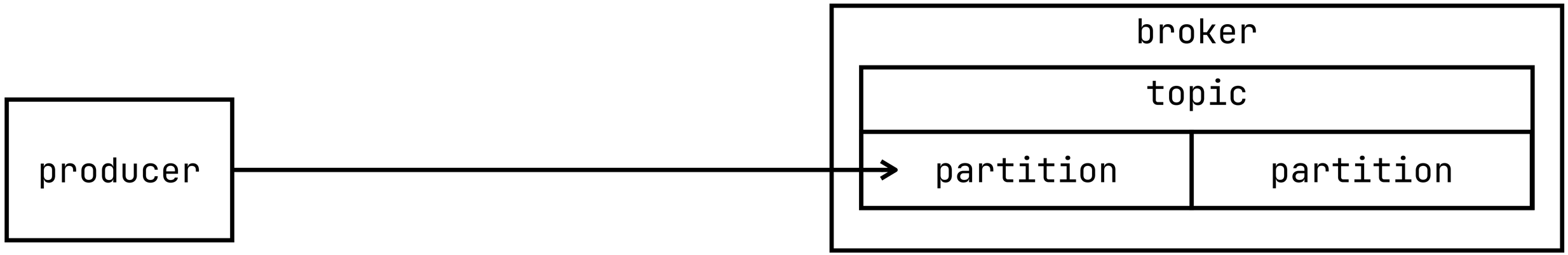
4.kafka

용어


kafka cluster
broker
topic
partition
partition
topic
topic
broker
broker
broker
Producer
Producer
Producer
consumer group
Consumer
Consumer
Consumer

4.kafka

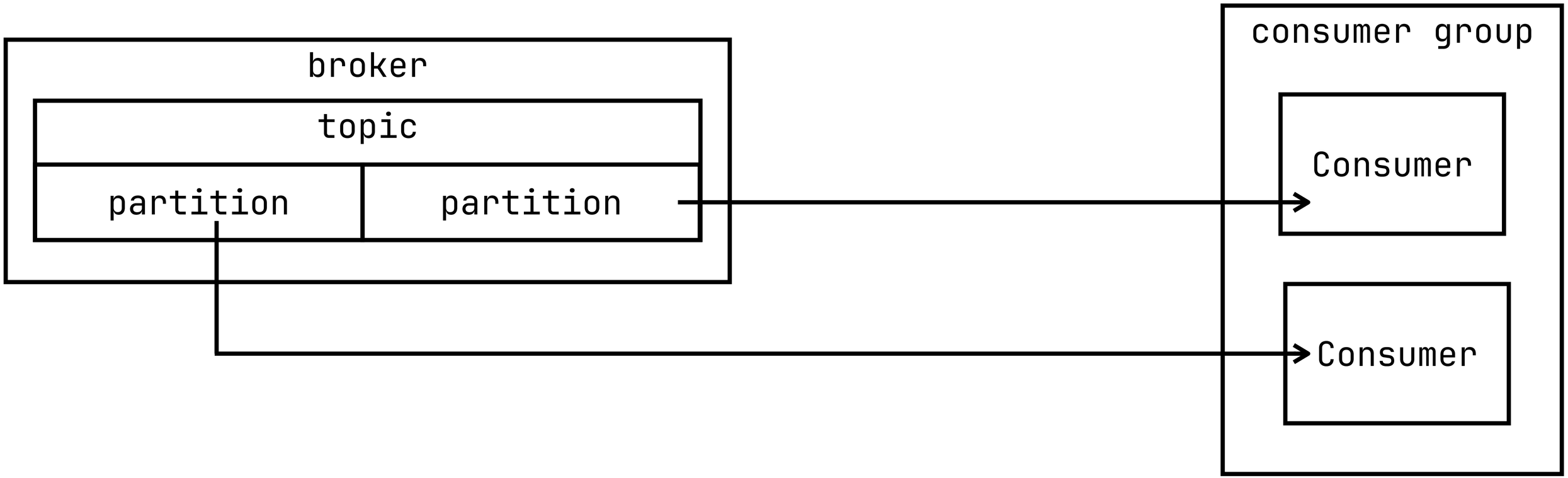
publish



producer가 event stream을 특정 topic의 partition에 publish

4.kafka

subscribe



consumer가 특정 topic을 subscribe 하여 데이터 소비

4.kafka

내장 프레임워크

kafka connect	데이터 수집과 관련된 외부 시스템과의 연결 라이브러리 모음
kafka streams	스트림 처리 라이브러리 모음
kraft	zookeeper 를 사용하지 않고 kafka가 동작할 수 있도록 하는 모드

5.elk

개념

Elasticsearch : 분산 검색엔진

Logstash : 이벤트 수집, 정제

Kibana : 시각화, 관리

+

Beats : 경량 수집 도구

Elastic Stack 으로 부르기도 한다.

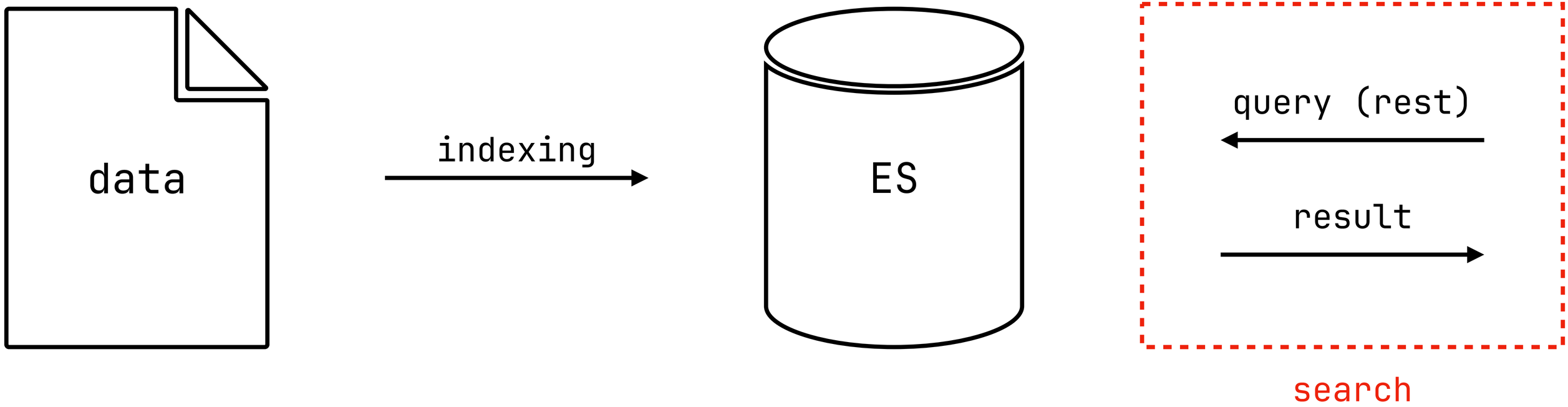
5.elk

elasticsearch

lucene 기반

nosql (vector db) 처럼 사용 → 검색엔진 (full text search engine)

rest api 기반



5.elk

elasticsearch

cluster	여러 대의 컴퓨터를 병렬로 연결해 하나의 시스템을 구성
index	도큐먼트를 저장하는 논리적 구분자
document	실제 데이터를 저장하는 단위
field	{field: value} 의 json 형태로 저장
mapping	json 형태의 데이터를 루신이 이해할 수 있도록 바꿔주는 작업

rdb	es
table	index
record	document
column	field
schema	mapping

5.elk

elasticsearch

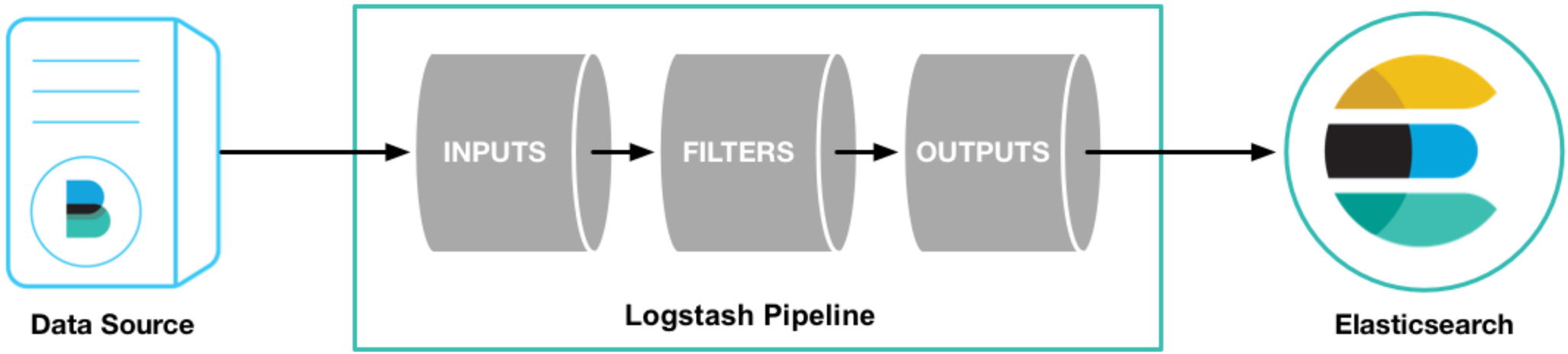
cluster	여러 대의 컴퓨터를 병렬로 연결해 하나의 시스템을 구성
node	cluster를 구성하는 요소 (하나의 컴퓨터에 하나의 node) - master node - data node
shard	데이터 분산 저장 단위 - primary shard - replica shard
segment	실제 데이터 (토큰화된 역인덱스 데이터 + 소스 데이터)

5.elk

logstash

실시간 파이프라인을 사용한 이벤트 수집 / 정제

플러그인 기반



5.elk

logstash

filename.conf

```
input {
}

filter {
}

output {
}
```

입력 플러그인 : 데이터 수집

필터 플러그인 : 데이터 구조화

출력 플러그인 : 데이터 전송

5.elk

kibana

1) Add data 데이터 추가

2) Explore 필터링

3) Visualize 시각화

4) Model data behavior 분석

5) Share 공유