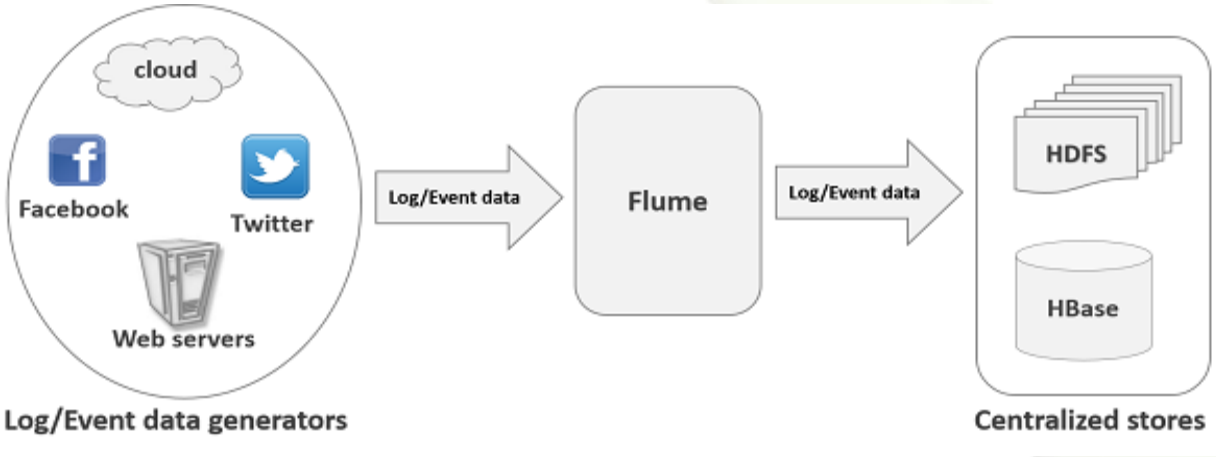
Introduction to Flume

Flume?

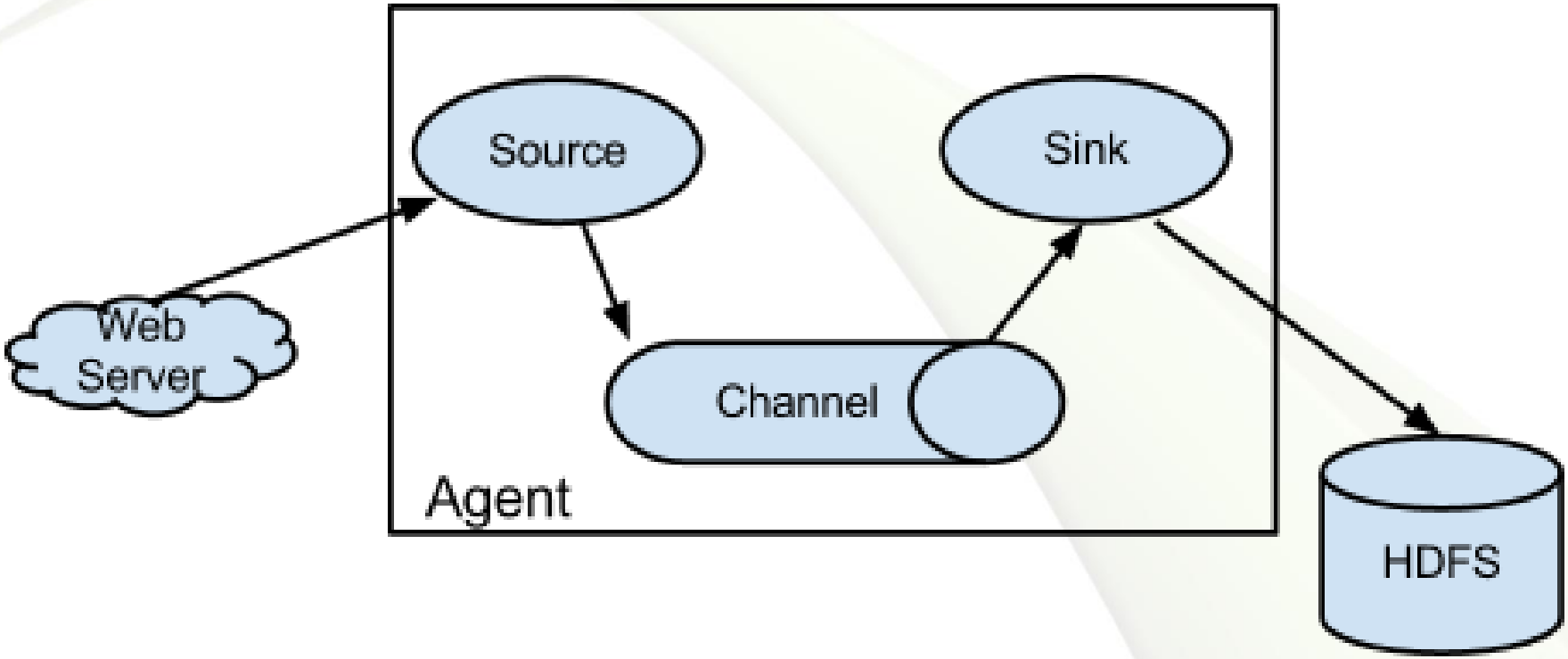
- 대용량 데이터를 하둡으로 수집하기 위한 도구
  - 다양한 수집 요구사항 적용 가능
  - 이벤트 기반



- 클라우데라에서 개발 → 아파치 프로젝트로 등록

- Flume-OG (0.9.x) → Flume-NG (1.x)

Flume Architecture

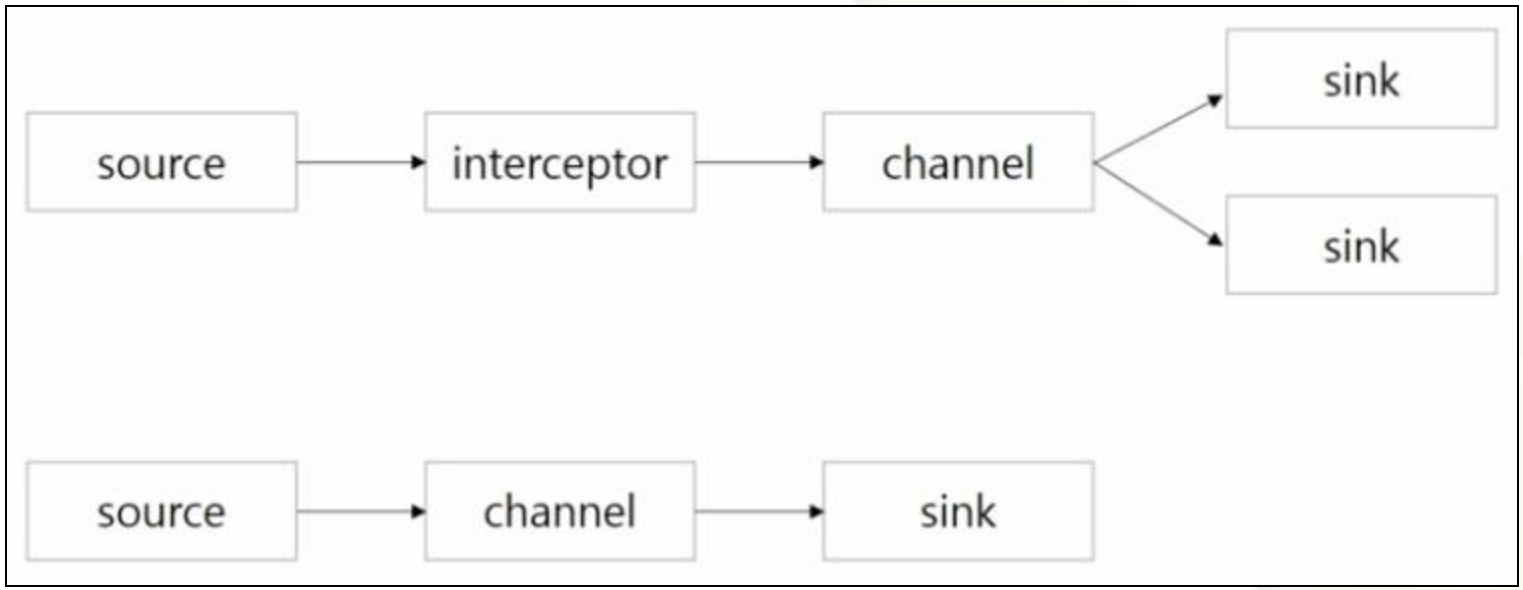


Flume 구성 요소

구성 요소	설명
Source	▪ 다양한 원천 시스템의 데이터를 수집하기 위해 Avro, Thrift, JMS, Spool Directory, Kafka 등의 컴포넌트 제공 ▪ 수집한 데이터를 Channel로 전달
Sink	▪ 수집한 데이터를 Channel로부터 전달 받아 최종 목적지에 저장하기 위한 컴포넌트 ▪ HDFS, Hive, Logger, Avro, ElasticSearch, Thrift 등의 컴포넌트 제공
Channel	▪ Source와 Channel을 연결하고 데이터 버퍼링을 수행하는 컴포넌트 ▪ 메모리, 파일, 데이터베이스를 채널의 저장소로 사용
Interceptor	▪ Source와 Channel 사이에서 데이터 필터링 및 가공하는 컴포넌트 ▪ Timestamp, Host, Regex Filtering 등을 기본 제공 ▪ 사용자 정의 Interceptor 추가 가능
Agent	▪ Source → (Interceptor) → Channel → Sink 순으로 구성된 작업 단위 ▪ 독립된 인스턴스로 생성

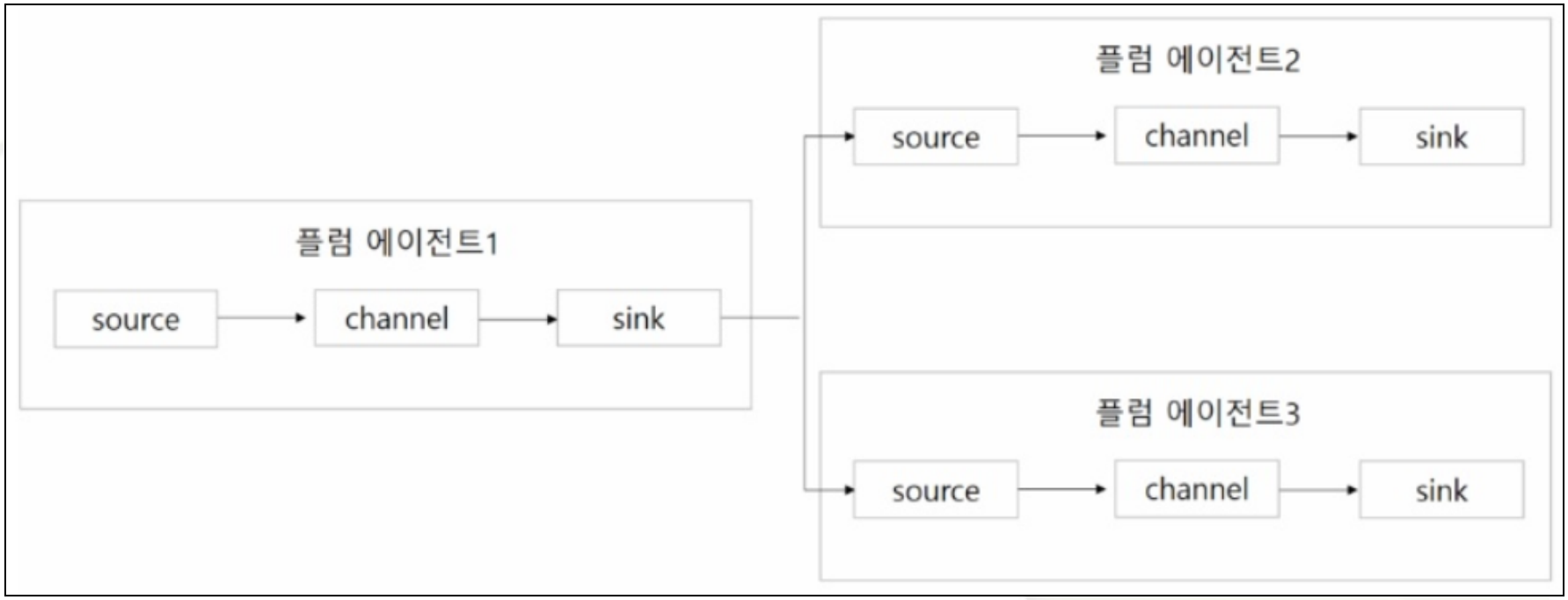
구성 유형

- 원천 데이터를 요구 사항 없이 단순 적재

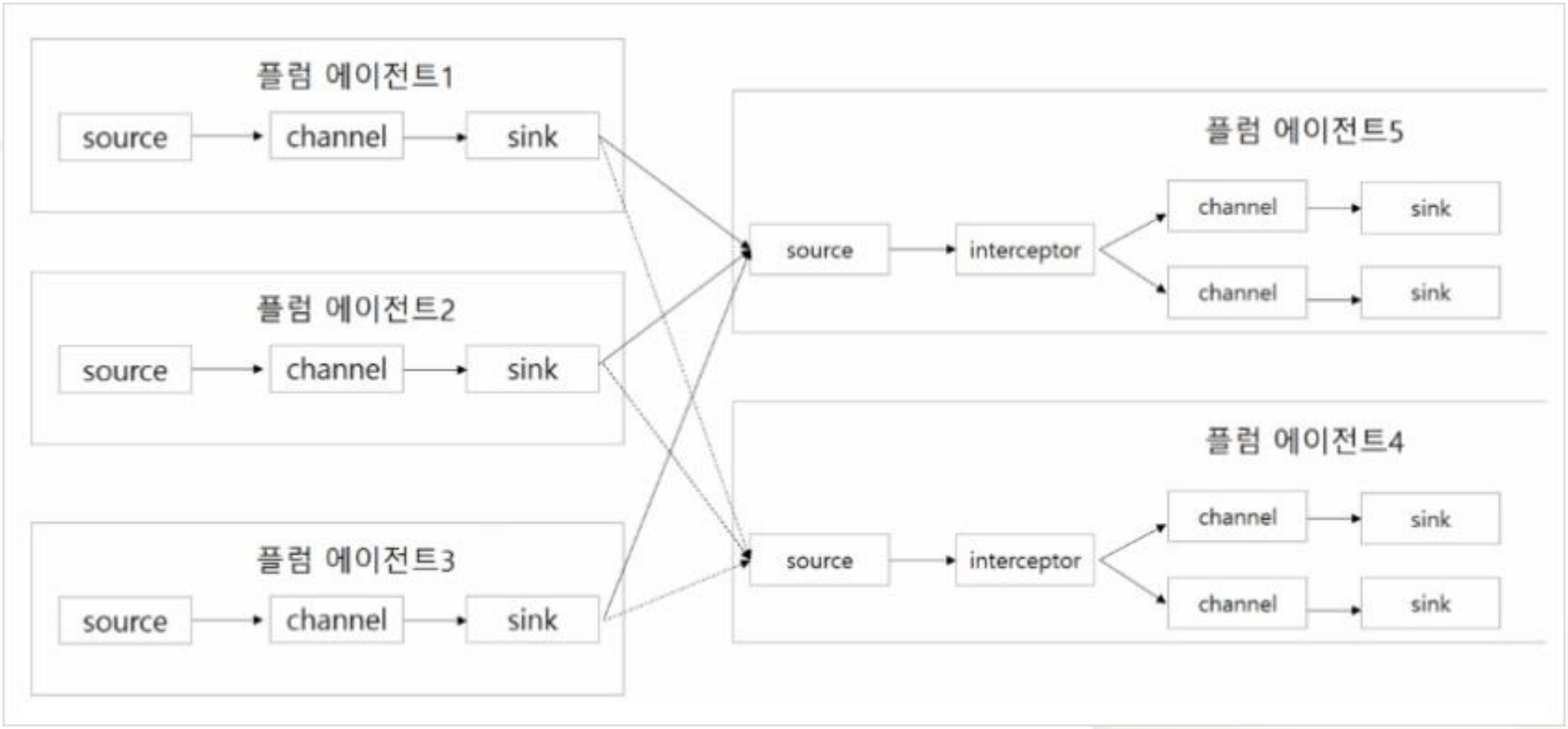


- 인터셉터를 추가해 데이터 가공
- 데이터 특성에 따라 다수의 Sink로 데이터 라우팅
- 하나의 에이전트에 다수의 Source → Channel → Sink 구성

구성 유형



- 한 에이전트에서 수집한 데이터를 다른 에이전트에 분배
- load_balance, failover 등의 기능을 선택적으로 수행



- 원천 시스템에서 다양한 대규모 데이터가 유입되는 상황에 적합한 분산 아키텍처

Introduction to Zookeeper

Zookeeper?

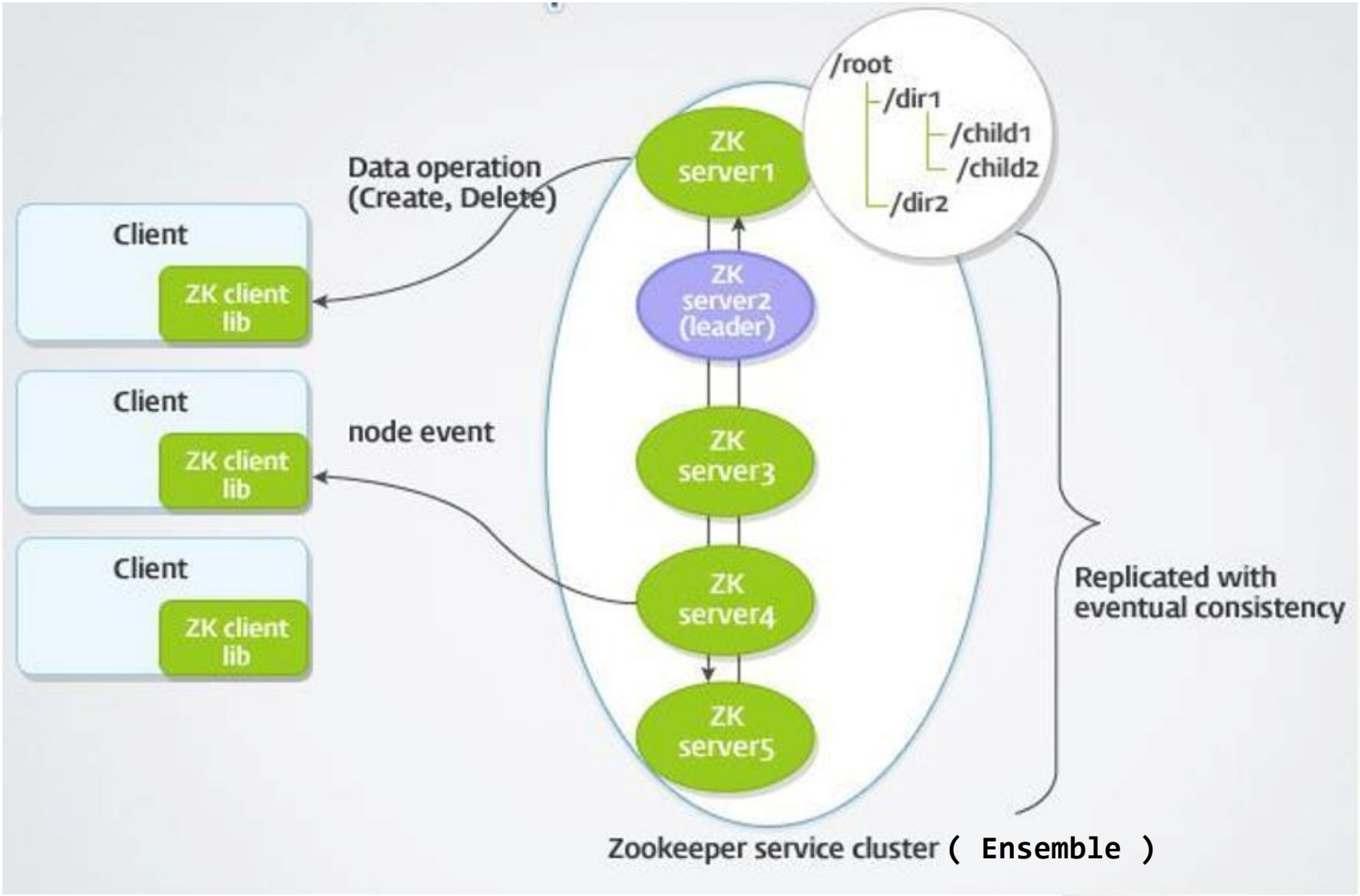
- 많은 수의 서버로 구성된 빅데이터 분산 환경을 효율적으로 관리하기 위해 서버 간의 정보를 쉽고 안전하게 공유하는 요구사항 발생

- Zookeeper는 공유된 정보를 기반으로 서버 간의 중요한 이벤트 관리 및 상호 작용을 조율하는 코디네이터 시스템

- Hadoop, Kafka, HBase 등의 분산 노드 관리에 사용

- Hadoop Sub Project → Apache 최상위 프로젝트로 승격

Zookeeper Architecture



Zookeeper 구성 요소

구성 요소	설명
Client	▪ Zookeeper Znode에 저장된 데이터에 대한 읽기, 쓰기, 삭제 등의 작업을 요청하는 클라이언트
Znode	▪ Zookeeper 서버에 생성되는 파일 시스템의 디렉터리 ▪ 클라이언트의 요청 정보를 계층적으로 관리 ▪ 버전, 접근권한, 상태 등의 정보 관리
Ensemble	▪ 3대 이상의 Zookeeper 서버를 하나의 클러스터로 구성한 HA 아키텍처
Leader	▪ Ensemble에 포함된 서버 중 한 대의 서버를 리더 서버로 선출 ▪ 클라이언트의 요청을 받은 서버는 정보를 리더 서버에 전달 ▪ Leader 서버는 전달 받은 정보를 모든 Follower 서버에 전달 되도록 보장
Follower	▪ Ensemble에 포함된 서버 중 Leader 서버를 제외한 나머지 서버 ▪ Leader 서버와 메시지를 주고 받으면서 ZNode의 데이터 동기화 ▪ Leader 서버에 장애가 발생하면 새로운 Leader 서버 선출

Introduction to Kafka

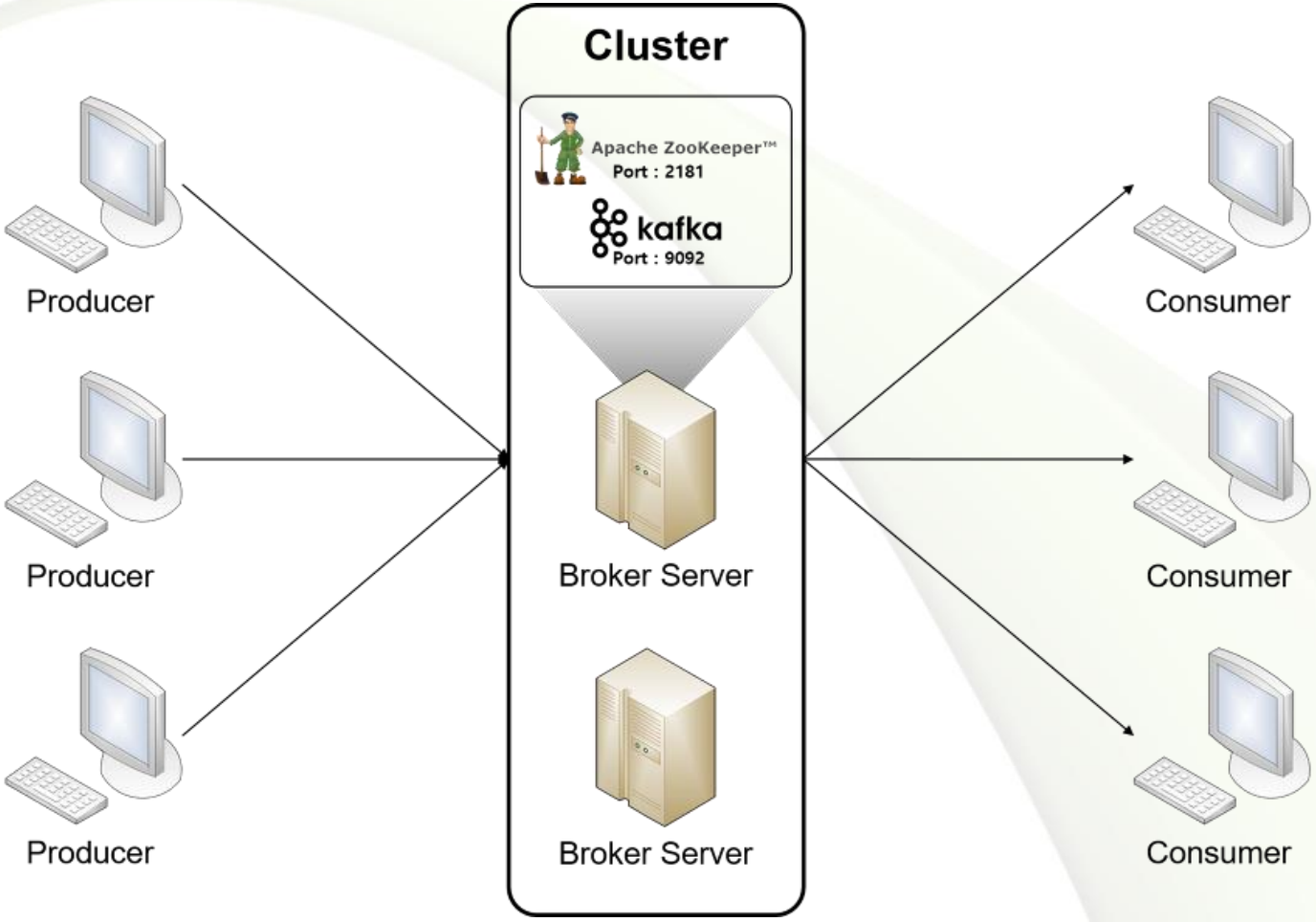
Kafka?

- Message Oriented Middleware 소프트웨어

- 대규모로 발생하는 메시지성 데이터를 비동기 방식으로 중계하는 역할
  - 원천 시스켐으로부터 대규모 트랜잭션 데이터가 발생했을 때 중간에 데이터를 버퍼링하면서 안정적으로 타깃 시스템 데이터를 전송하는 중간 시스템

- 링크드인에서 개발 → 아파치 프로젝트로 등록

Kafka Architecture



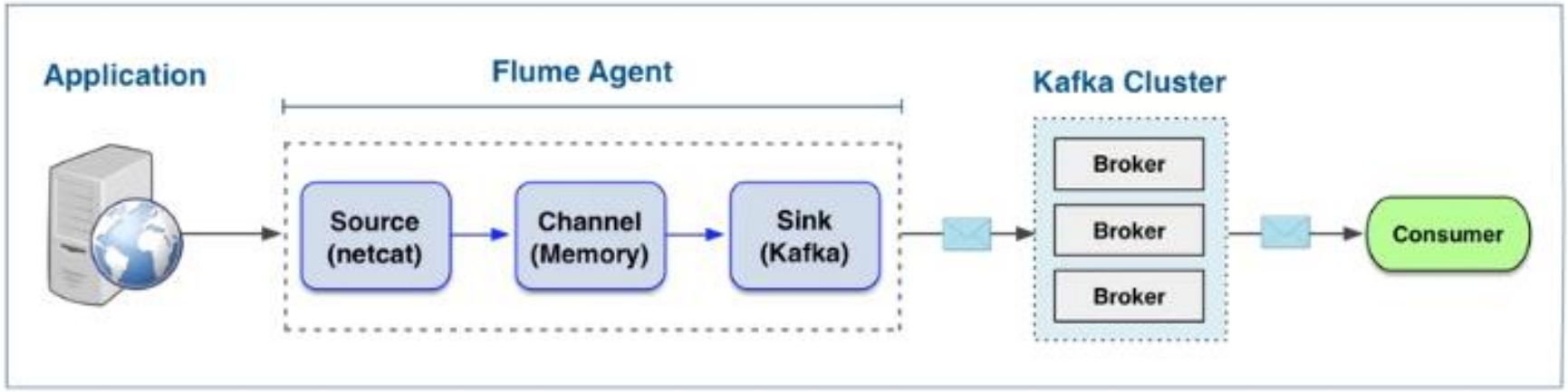
Kafka 구성 요소

구성 요소	설명
Broker	▪ Kafka의 서비스 인스턴스 ▪ 다수의 Broker를 클러스터로 구성하고 Topic이 생성되는 물리적 서버
Topic	▪ Broker에서 데이터의 [발행 / 소비] 처리를 위한 저장소
Provider	▪ Broker의 특정 Topic에 데이터를 전송(발행)하는 역할 수행 ▪ 애플리케이션에서 카프카 라이브러리를 사용해서 구현
Consumer	▪ Broker의 특정 Topic에서 데이터를 수신(소비)하는 역할 수행 ▪ 애플리케이션에서 카프카 라이브러리를 사용해서 구현

Key	Apache Kafka	Flume
Basic	실시간 스트리밍 데이터 처리에 최적화된 분산 데이터 저장소	다양한 소스에서 발생한 대규모 로그 데이터를 효과적으로 수집, 병합하고 목적지로 전달하는 시스템
Saclability	확장 쉬움	Kafka에 비해 상대적으로 확장성 낮음
Push /Pull	Pull model 지원	Push model 지원
Recovery	높은 수준의 장애 복구 지원	Flume Agent 장애 발생 시 데이터 손실 가능
Flexibility	범용 목적의 발행-구독 기반 메시징 시스템	Hadoop을 주 타겟으로 하는 시스템

Flume & Kafka 결합

- 빠르게 발생하는 대량의 데이터를 안정적으로 목적지로 전송하기 위한 구성



- Flume이 빠르게 발생하는 데이터를 실시간으로 수집하고 이를 최종 목적지로 전달하기 전에 중간에서 안정적인 버퍼링 처리

- 목적지에 장애가 발생해도 중간에서 데이터를 저장해 두었다고 장애가 복구되면 데이터 전달

- 비동기 전송 방식으로 수집 속도 개선