# Flume Ecosystem



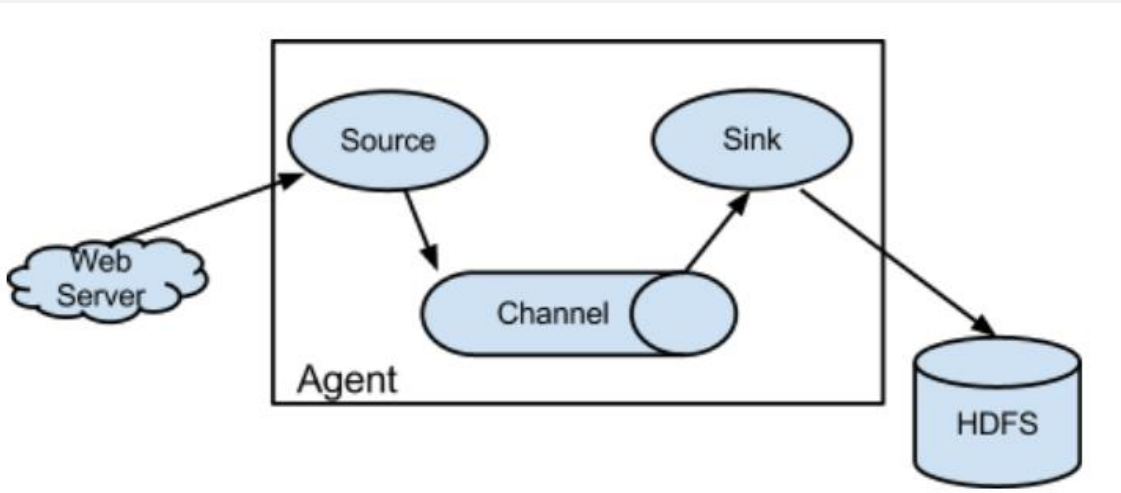
Kim Hye Kyung

topickim@naver.com

# Flume

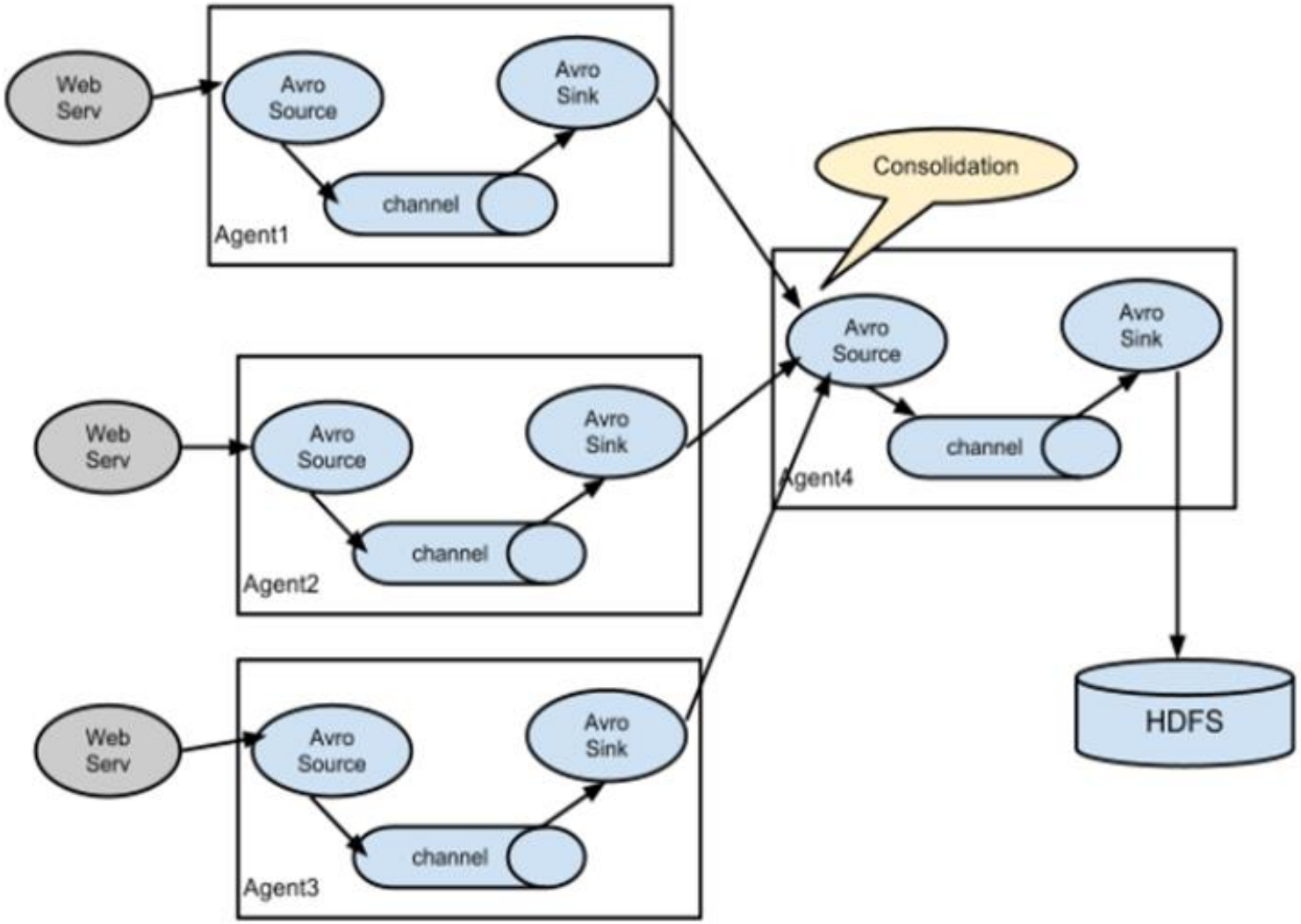
# Flume 개요

- 빅데이터를 수집할 때 다양한 수집 요구 사항들을 해결하기 위한 기능으로 구성된 소프트웨어
  - https://flume.apache.org/
  - 연속적으로 생성되는 데이터 스트림(로그파일, 소셜 미디어 데이터, 이메일 메신저 등)을 수집하여 HDFS에 저장할 수 있는 도구
  - 분산환경에서 대량의 로그 데이터를 효과적으로 수집하여, 합친 후 다른곳으로 전송할 수 있는 신뢰할 수 있는 서비스




Flume 아키텍처

# Flume : 멀티에이전트 통합구성



# Flume 기본 요소

| 주요 구성 요소 | 특징 |
|---|---|
| Source | 다양한 원천 시스템의 데이터를 수집하기 위해 Avro, Thrift, JMS등<br>여러 주요 컴포넌트를 제공하며, 수집한 데이터를 Channel로 전달 |
| Sink | 수집한 데이터를 Channel로 부터 전달 받아 최종 목적지에 저장하기 위한 기능으로<br>HDFS, Hive, Logger, Avro, ElasticSearch등 제공 |
| Channel | Source와 Sink를 연결하며, 데이터를 버퍼링하는 컴포넌트로 메모리, 파일,<br>데이터베이스를 채널의 저장소로 활용 |
| Interceptor | Source와 Channel 사이에서 데이터 필터링 및 가공하는 컴포넌트로서 Timestamp,<br>Host, Regex Filtering 등을 기본 제공하며, 필요 시 사용자 정의 Interceptor 추가 |
| Agent | Source -> Channel -> Sink 컴포넌트 순으로 구성된 작업 단위로 독립된<br>인스턴스로 생성 |

# Flume 기본 요소 - Source

| 주요 구성 요소 | 특징 |
|---|---|
| avro | Avro 클라이언트에서 전송하는 이벤트를 입력으로 사용,<br>Agent와 Agent를 연결해줄 때 유용 |
| exec | System Command를 수행하고 출력 내용을 수집 |
| spooldir | WAS의 Log 파일 위치의 디렉토리를 Spooling하여 파일이 만들어졌을<br>때를 모니터링하여 수집 |
| thirft | WAS에서 로그를 파일로 별도로 남기지 않고, Thrift 통신으로 직접<br>Agent에 발송하여 로그 수집 |
| jms | JMS 메시지 수집 |

# Flume 기본 요소 - Channel

| 주요 구성 요소 | 특징 |
| --- | --- |
| memory | Source에서 받은 이벤트를 Memory에 가지고 있는 구조로, 간편하고 빠른 고성능(High Throughput)을 제공하지만 이벤트 유실 가능성이 있음<br>즉, 프로세스가 비정상적으로 죽을 경우 데이터가 유실될 수 있음 |
| jdbc | JDBC로 저장 |
| file | JDBC와 마찬가지로 속도는 Memory기반에 비해 느리지만, 프로세스가 비정상적으로 죽더라도 transactional하게 프로세스를 재시작하여 재처리하며 이벤트 유실이 없는 것이 장점 |

# Flume 기본 요소 – Sink

| 주요 구성 요소 | 특징 |
|---|---|
| logger | 테스트 또는 디버깅을 위한 로깅 |
| avro | 다른 Avro 서버(Avro Source)로 이벤트 전달 |
| hdfs | HDFS에 저장. Hive,Pig,R,Mahout 등으로 배치 분석,기계어 학습 등에 활용 |
| elasticsearch | 이벤트를 변환해서 ElasticSearch에 저장. Kibana와 같은 로그 통계,모니터링을 통해 서비스 가능 |
| … | |