데이터 분류

데이터 분류

❑ 정형 데이터

- 일반적으로 관계형 데이터베이스(RDBMS)에 저장되는 스프레드시트 형 데이터
- 데이터 스키마와 실제 정보를 가지는 파일로 구성

❑ 반정형 데이터

- 데이터 내부에 정형데이터의 스키마에 해당되는 메타데이터를 가짐
- 일반적으로 파일 형태로 저장

❑ 비정형 데이터

- 데이터세트가 아닌 하나의 데이터가 수집 데이터로 객체화
- 텍스트 데이터, 이미지, 동영상 같은 멀티미디어 데이터
- HTML은 반정형 데이터 또는 비정형 데이터로도 볼 수 있음

데이터의 일반적인 특징

❑ 데이터 구분

구분	정성적 데이터	정량적 데이터
형태	비정형 데이터	정형, 반정형 데이터
특징	객체 하나의 함의된 정보를 가짐	속성이 모여 객체를 이룸
구성	언어, 문자	수치, 도형, 기호
저장형태	파일, 웹	데이터 베이스, 스프레드시트
소스	외부 시스템	내부 시스템(RDBMS, Legacy)

❑ 데이터 종류

- 레코드기반 데이터 : Data Matrix, Document Data, Transaction Data
- 그래프기반 데이터 : World Wide Web, Molecular Structure
- 서열형 데이터 : Spatial Data, Temporal Data, Sequential Data

데이터 수집 위치

❑ 내부 데이터

- 원천 데이터의 저장소가 내부 시스템에 있는 데이터
- 단순한 물리적 위치에 대한 구분이 아닌 데이터 제공자에 대한 구분
- 수집에 대한 기술적 제약이나 보안에 대한 문제가 적음

❑ 외부 데이터

- 원천 데이터가 외부 시스템에 위치
- 데이터 수집시 데이터 제공자와의 협의 필요
- 수집 주기, 수집 방법, 보안 등 고려해야 할 요소 증가

데이터 수집 절차

❑ 수집 절차

❑ 고려 사항

- 수집 가능성 : 원천 데이터 제공 여부, 수집 주기, 전후처리 고려
- 보안 : 개인정보보호 정책, 저작권 문제 고려
- 정확성 : 수집하는 데이터가 서비스의 활용목적에 사용할 수 있는 데이터 인지 검토
- 수집 난이도 : 수집에 대한 기술적인 문제 고려
- 수집 비용 : 데이터 구입 비용, 수집 시스템 구축 비용, 전후처리 비용

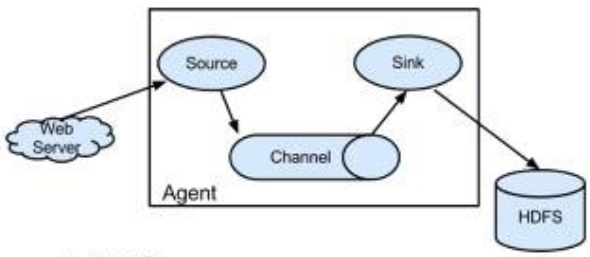
Flume

로그 데이터

- ❑ 다양한 데이터 소스로 부터 정형, 반정형, 비정형 데이터 생산
- ❑ IoT device을 발달과 함께 효용성이 날로 증가
- ❑ 대부분 기계에 의해 발생하는 데이터이므로 bias가 적음
- ❑ 수집 시 고려 사항
  - 확장성 : 수집 대상 시스템이 얼마나 늘어날 것인가?
  - 안정성 : 수집되는 데이터가 손실되지 않고 안정적으로 저장 가능한가?
  - 유연성 : 다양한 데이터의 형식과 접속 프로토콜을 지원하는가?
  - 주기성 : 수집 데이터가 실시간으로 반영돼야 하는가 혹은 배치처리를 해도 가능한가?

Apache Flume

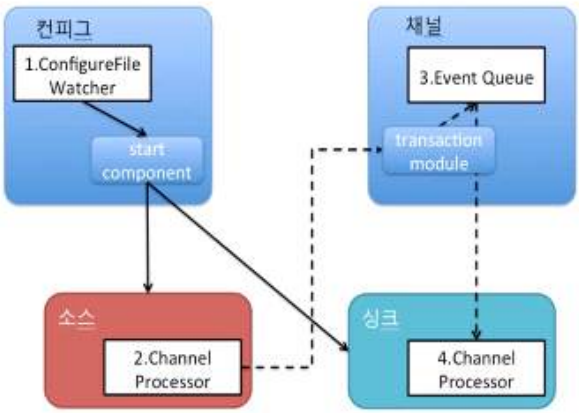
- ❑ 대규모의 로그데이터를 효율적으로 수집, 통합, 전송하기 위한 서비스



- ❑ Source, Sink, Channel 구성
  - Source : 데이터의 유입 지점
  - Sink : 데이터를 내보내는 곳
  - Channel : Source와 Sink 사이의 Queue

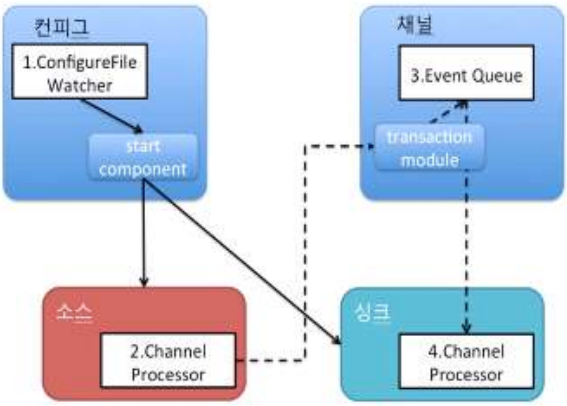
Flume 아키텍처

❑ Configuration



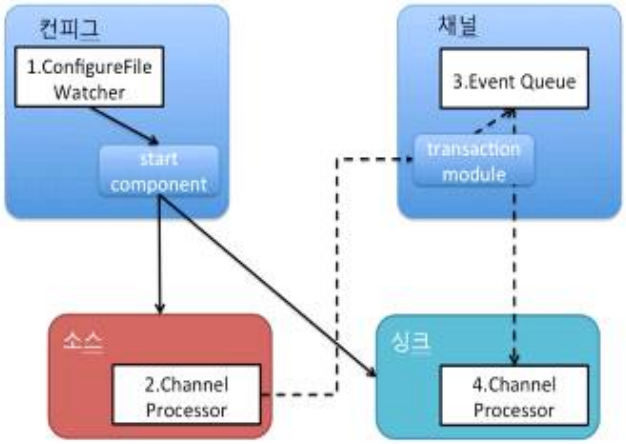
- 설정 파일을 읽어서 채널, 소스, 싱크 설정
- 30초마다 파일 재로드
- 재시작 하지 않아도 재설정 가능

❑ Source



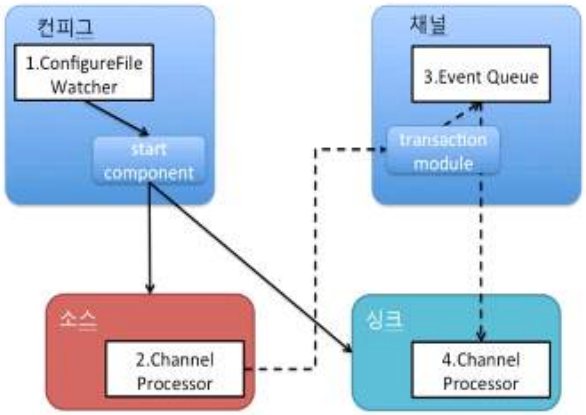
- 설정을 읽은 후에 Source runne가 실행되어 데이터를 채널로 전송
- PoolableSourceRunner
- EventDrivenSourceRunner

❑ Channel



- 이벤트의 트랜잭션 관리
- 이벤트를 Queue에 저장

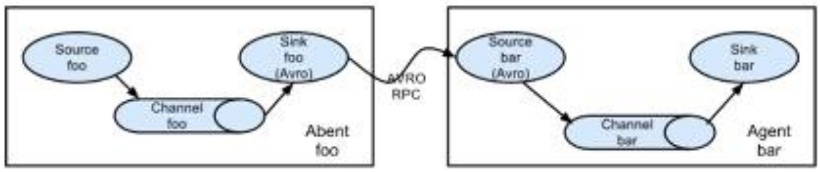
❑ Sink



- 설정파일에서 채널의 리스트를 가져옴
- 채널에서 이벤트를 꺼냄

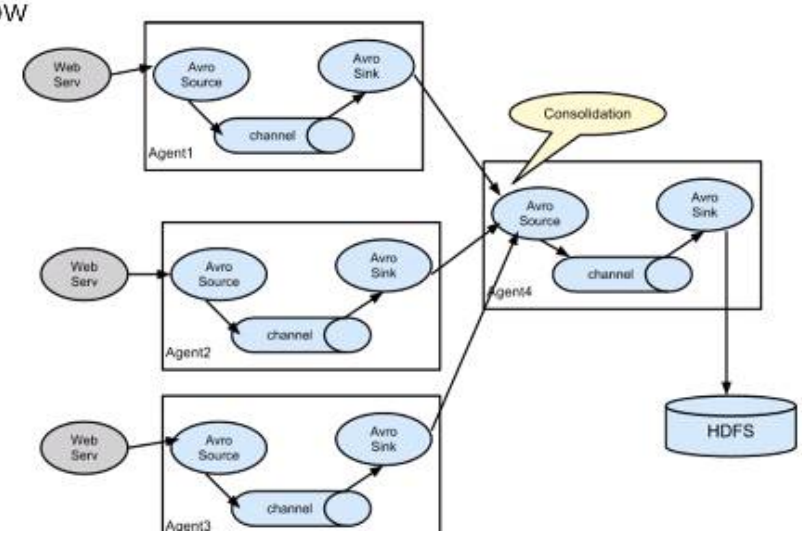
Data flow 구성

❑ Multi-Flow



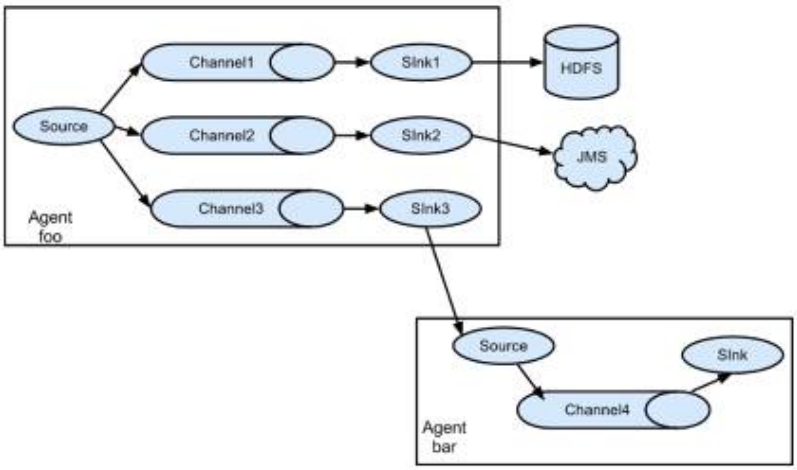
Data flow 구성

❑ Consolidation Flow



Data flow 구성

❑ MultiFlexing Flow



❑ 설정 파일

```
agent.sources = seqGenSrc
agent.channels = memoryChannel
agent.sinks = loggerSink
```
이름 정의

```
agent.sources.seqGenSrc.type = seq
agent.sources.seqGenSrc.channels = memoryChannel
```
Source 설정

```
agent.sinks.loggerSink.type = logger
agent.sinks.loggerSink.channel = memoryChannel
```
Sink 설정

```
agent.channels.memoryChannel.type = memory
agent.channels.memoryChannel.capacity = 100
```
Channel 설정

❑ 실행 : bin/flume-ng <command> [options] ...

종류	명령어/옵션	설명
command	help	도움말 표시
	agent	Flume agent 실행
	avro-client	Avro Flume 클라이언트 실행
	version	버전 정보 표시
global option	--conf, -c <conf>	<conf> 디렉토리의 config 사용
	--classpath, -C <cp>	추가적인 클래스패스 지정
	--dryrun, -d	실제 Flume을 기동하지 않고 명령어만 출력
	-Dproperty=value	Java system 속성 설정
	-Xproperty=value	Java –X 옵션 설정

❑ 실행 : bin/flume-ng <command> [options] ...

종류	명령어/옵션	설명
agent option	--name, -n <name>	agent 이름 지정(필수)
	--conf-file, -f <file>	설정 파일 지정(zookeeper를 사용하지 않을 경우 필수)
	--zkConnString, -z <str>	사용할 zookeeper 연결 스트링 지정(설정 파일을 사용하지 않을 경우 필수)
	--zkBasePath, -p <path>	Zookeeper 기본 경로 지정
	--no-reload-conf	설정 파일이 바뀌어도 다시 로딩하지 않음
	--help, -h	도움말 표시

Flume Sources

❑ Exec

속성	기본값	설명
channels		사용할 channel 지정
type		exec
bind		실행할 명령어
shell		명령어를 실행할 때 사용할 shell, 예 : /bin/sh -c
batchSize	20	한번에 읽어서 채널로 전송할 최대 라인 수
batchTimeout	3000	버퍼가 다 차기 전에 기다리는 시간

❖ exec source는 asynchronous, 모든 비동기 소스는 채널에 이벤트를 보내는 것을 실패해도 client는 알 수가 없음(데이터 유실)

❑ Spooling Directory

- Flume이 재 시작 되거나 강제 종료 되어도 데이터는 유실 되지 않는다
- spooling 디렉토리에 있는 파일을 수정하면 Flume은 에러를 로그 파일에 기록하고 종료
- 동일한 이름의 파일이 들어오면 Flume은 에러를 로그 파일에 기록하고 종료
- 파일이름에 timestamp 같은 식별자를 붙여 사용하는 것을 권장

속성	기본값	설명
channels		사용할 channel 지정
type		spooldir
spoolDir		파일을 읽어 올 디렉토리

❑ Syslog TCP

속성	기본값	설명
channels		사용할 channel 지정
type		syslogtpc
host		연결할 hostname(ip address)
port		연결할 port

Sink

❑ HDFS

- 이벤트를 HDFS에 저장, text, sequence type 지원, 압축 지원

속성	기본값	설명
channels		사용할 channel 지정
type		hdfs
hdfs.path		저장할 hdfs path(hdfs://namenode/flume/webdata)

- escape sequences

alias	설명	alias	설명
%{host}	이벤트 헤더의 호스트 정보로 대치	%A	요일명 전체(Monday, Tuesday, ...)
%t	Milisenconds 로 표시된 유닉스 시간	%b	월 이름 요약형(Jan, Feb, ...)
%a	요일명 요약형(Mon, Tue, ..)	%B	월 이름 전체(January, February, ...)

❑ File Roll

속성	기본값	설명
channels		사용할 channel 지정
type		file_roll
sink.directory		저장할 디렉토리
sink.rollinterval	30	매 30초마다 파일이 rolling 됨. 0으로 지정하면 파일 롤링이 일어나지 않아서 모든 이벤트가 하나의 파일에 저장