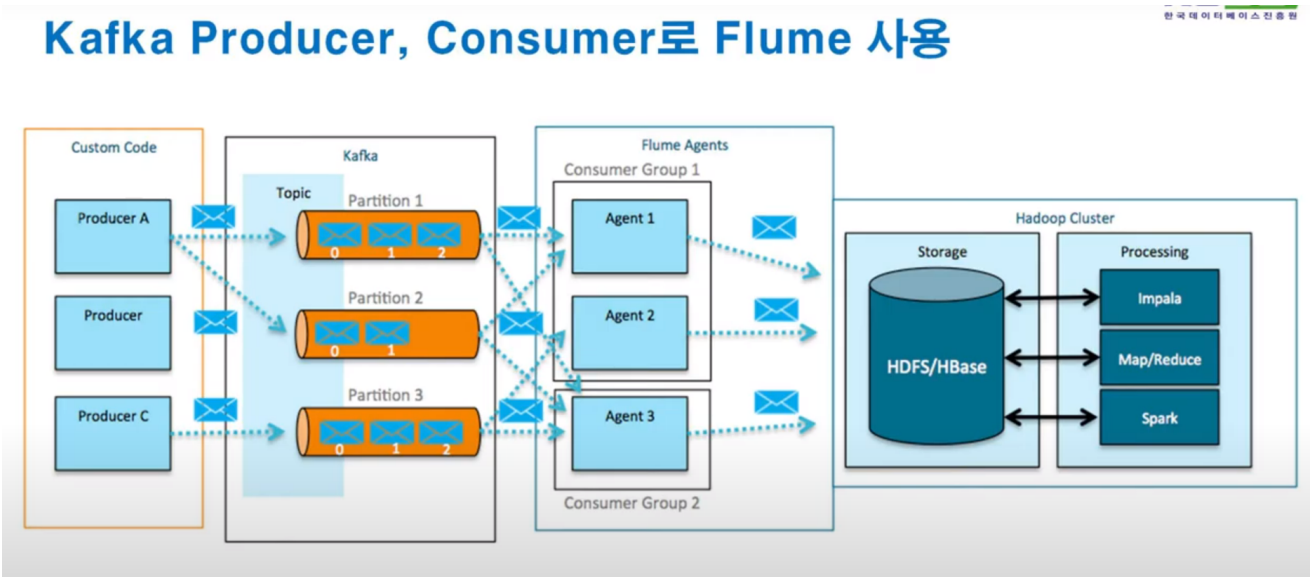
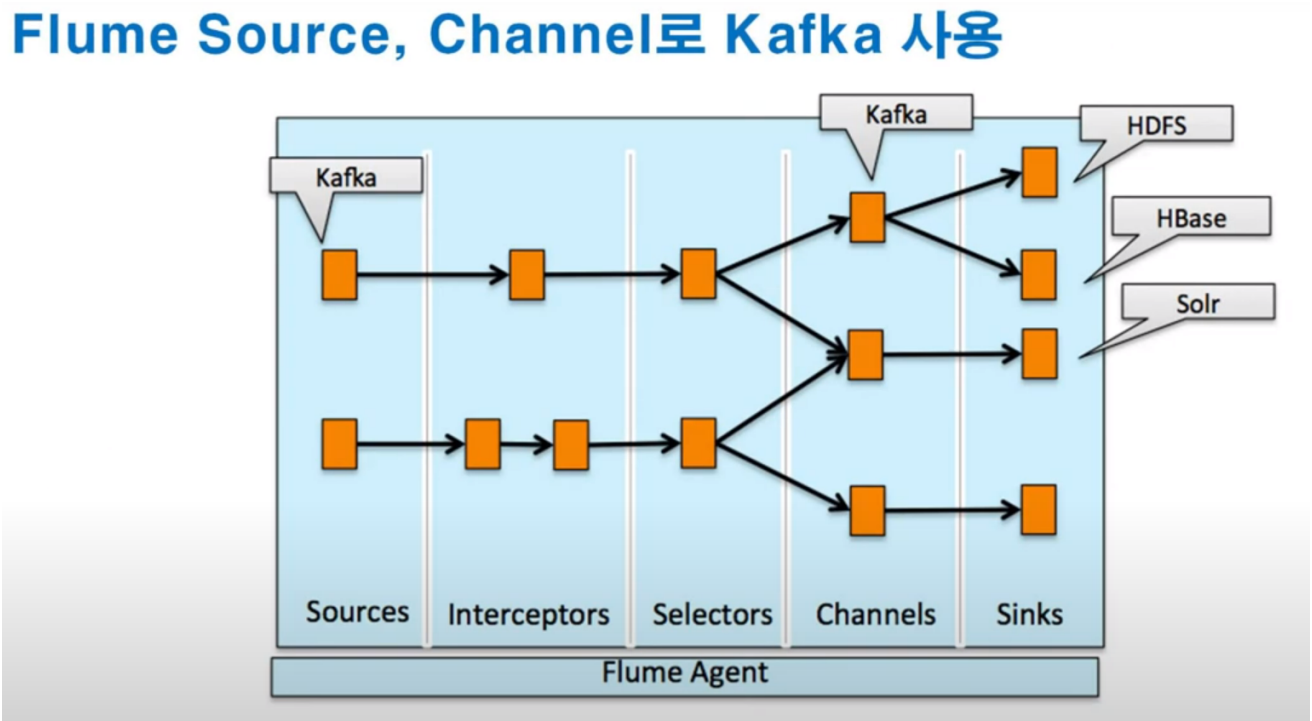
0. 데이터수집_Flume, Kafka 연동

- 여러 회사나 시스템에서 데이터를 수집하기 위한 앞 단계
- Flume(앞)+Kafka(뒤) 식으로 붙이는 경우도 많음



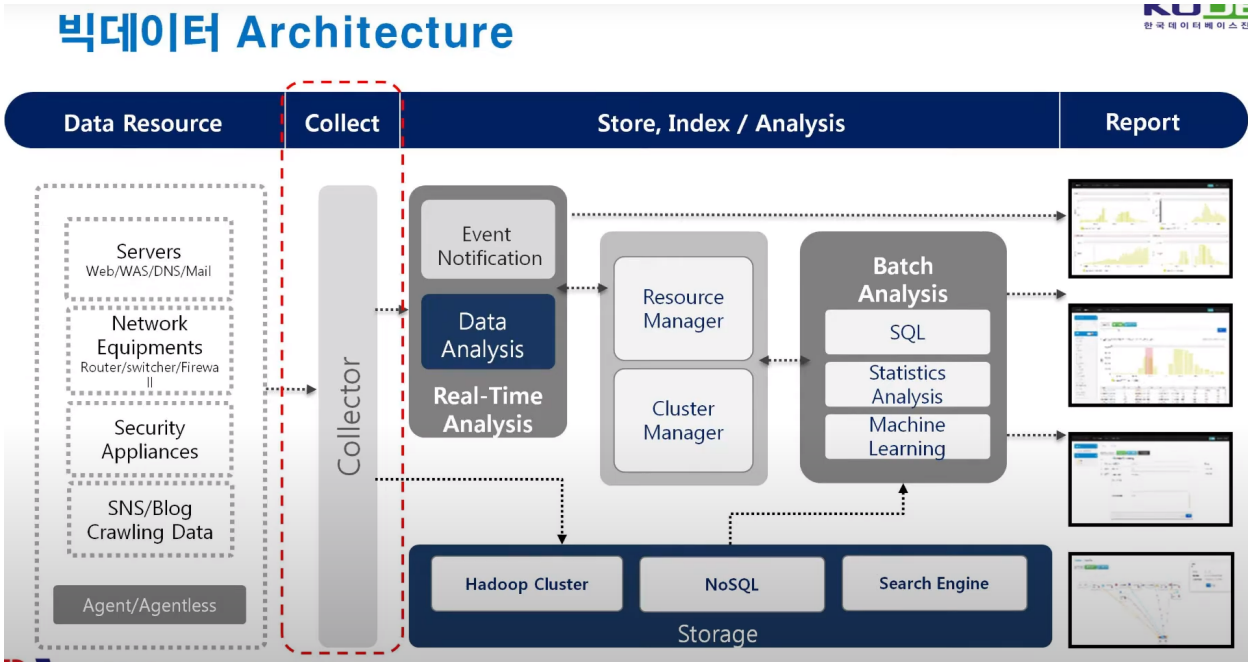
-〉 Flume의 source는 Kafka, sink는 HDFS



:: 목차

- 데이터수집 개요
- 정형 데이터 수집(Sqoop)
- 로그 수집(Flume)
- 분산 메세징 시스템(Kafka)
- Flume, Kafka 연동
- 비정형 데이터 수집(Web Crawling)
- SNS 수집

1. 데이터 수집 개요



- 정형/반정형/비정형 데이터 수집 -> 실시간 데이터 처리 or 배치 처리 -> 보고서(Report)
- 해당 과목에서는 수집을 해서
  i) to 실시간 데이터 처리 엔진
  ii) to 배치 처리 엔진
  iii) to DB나 클러스트

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 싯스템에 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