오 픈 소 스 기 반
리 얼 타 임 BIg Data
분 석 시 스 템

한국자바개발자협의회

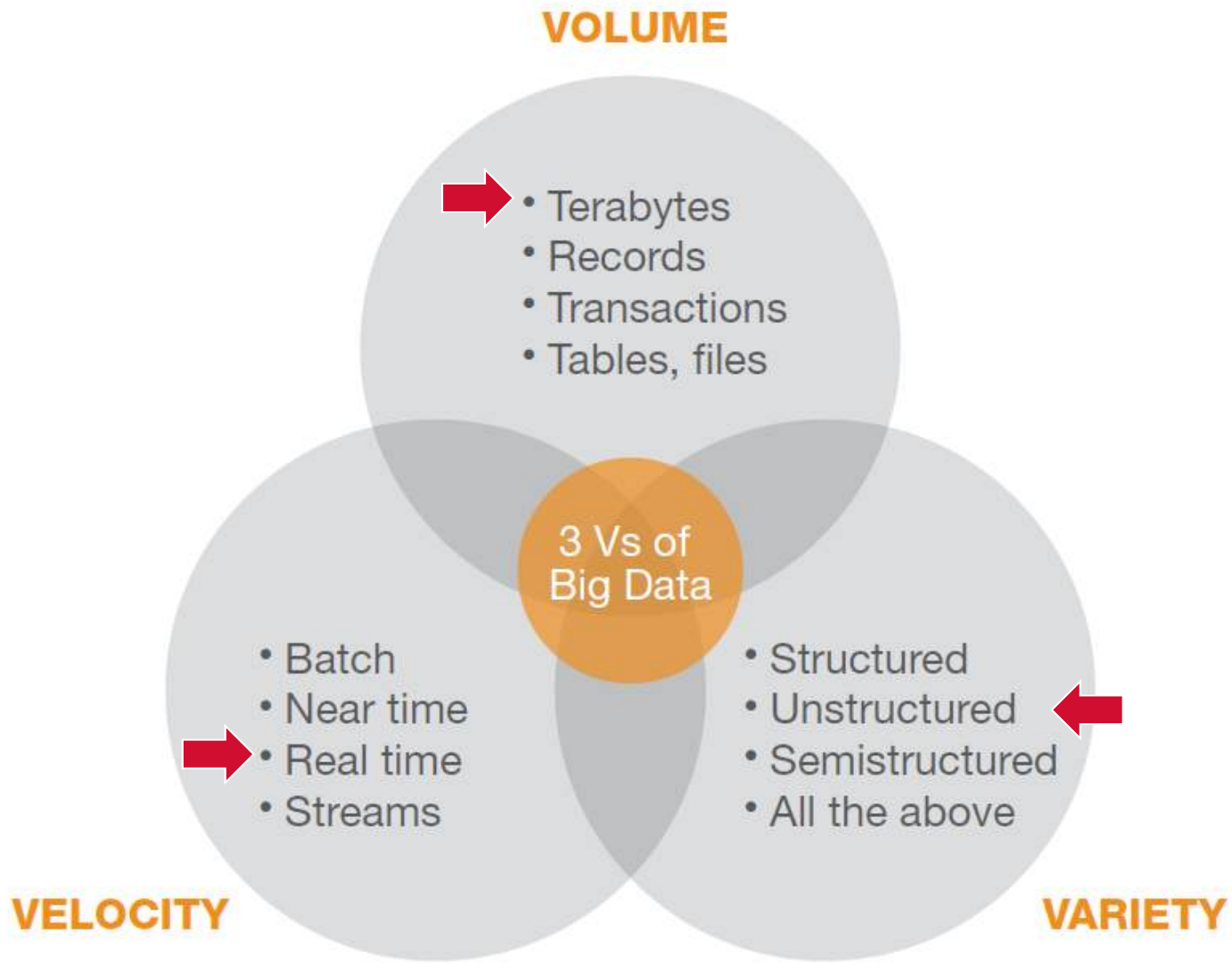
회장 김병곤

jco6th@gmail.com

소개

- (주)클라우다인 대표이사
- 한국자바개발자협의회(JCO) 회장
- JBoss User Group 대표
- 한국스마트개발자협회 부회장
- 지경부/NIPA 소프트웨어 마에스트로 멘토
- 한국IT전문가협회 정회원
- 대용량 분산 컴퓨팅 Technical Architect
- 오프라인 Hadoop 교육 및 온라인 Java EE 교육
- 오픈 소스 Open Flamingo 설립(http://www.openflamingo.org)
- Java Application Performance Tuning 전문가
- 다수 책 집필 및 번역
  - JBoss Application Server5, Enterprise JavaBeans 2/3

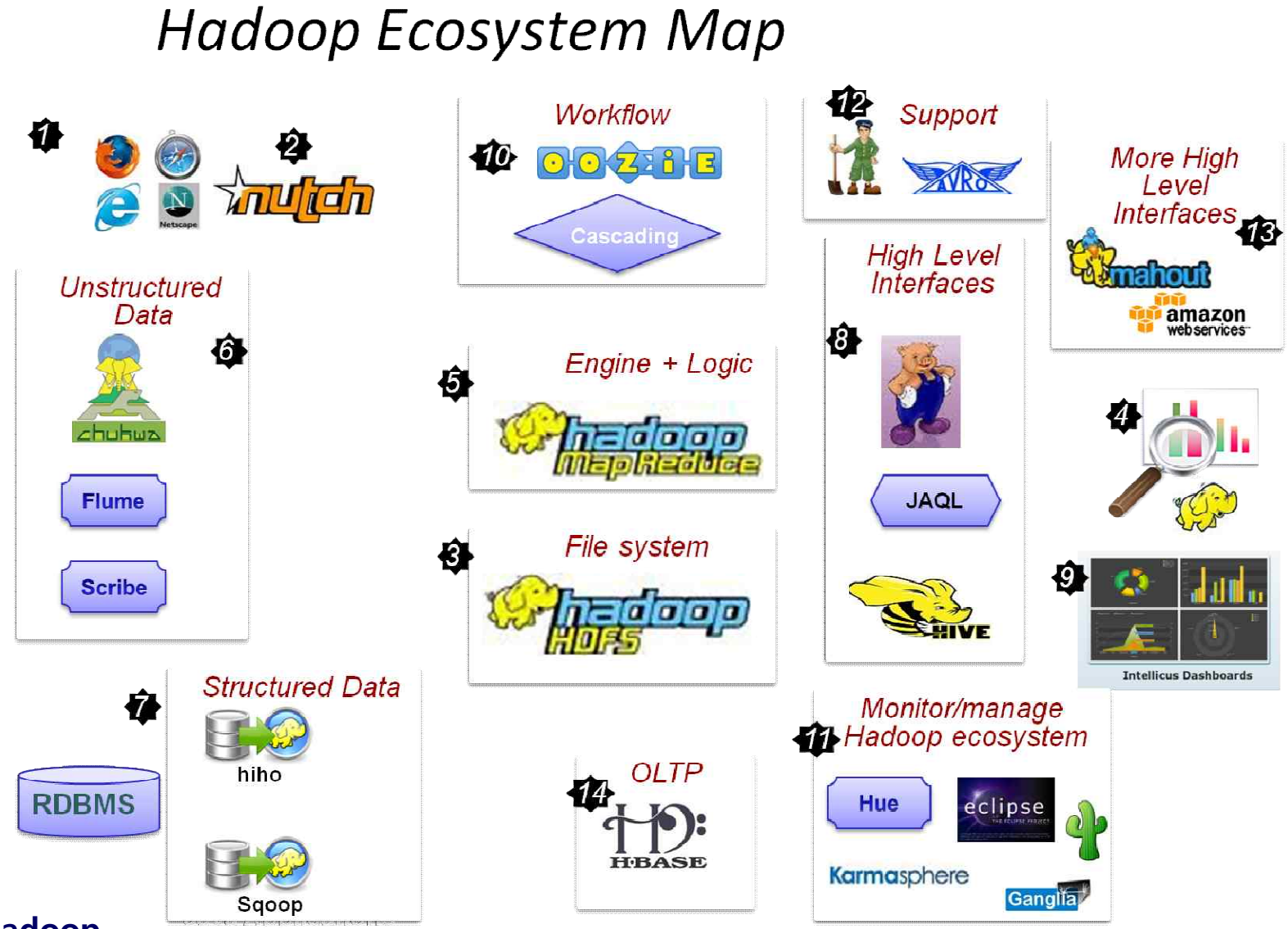
Big Data의 세 가지 속성



Big Data 활용 분야

적용분야	활 용
공공	- **U-City, USN 데이터의** 수집, 분석, 활용 - 환경, 방재, 국방, 기상 등 대용량 데이터 분석 기반의 시스템
금융/통신	- SNS 및 관련 서비스 - **N-Screen Service** - Card사의 결제정보, 로그정보 기반 **개인화 마케팅**
제조 및 일반기업	- Smart TV/Mobile AppStore 등 **제품 기반 B2C 서비스** - 제조 장비 운전 데이터 수집, **분석(SPC), 제어, 모니터링 시스템** - **대용량 EAI, B2Bi 구축** - **BI 2.0**, CRM, ERM, ERP 등의 **의사결정 지원도구 시스템** - 통계 데이터 기반 각종 시뮬레이션 및 예측 시스템
기타	- 인터넷 쇼핑몰의 사용자 패턴 정보 분석 및 타겟 마케팅 - 온라인 게임, 연말정산 등의 일시적 G2C 서비스 등등

Big Data Technology & Hadoop Echosystem



Real Time Big Data 서비스 요건

- 쇼핑몰 사이트의 사용자 클릭 스트림을 통해 실시간 개인화
- 대용량 이메일 서버의 스팸 탐지 및 필터링
- 위치 정보 기반 광고 서비스
- 사용자 및 시스템 이벤트를 이용한 실시간 보안 감시
- 시스템 정보 수집을 통한 장비 고장 예측

Real Time Big Data 구현을 위한 기술

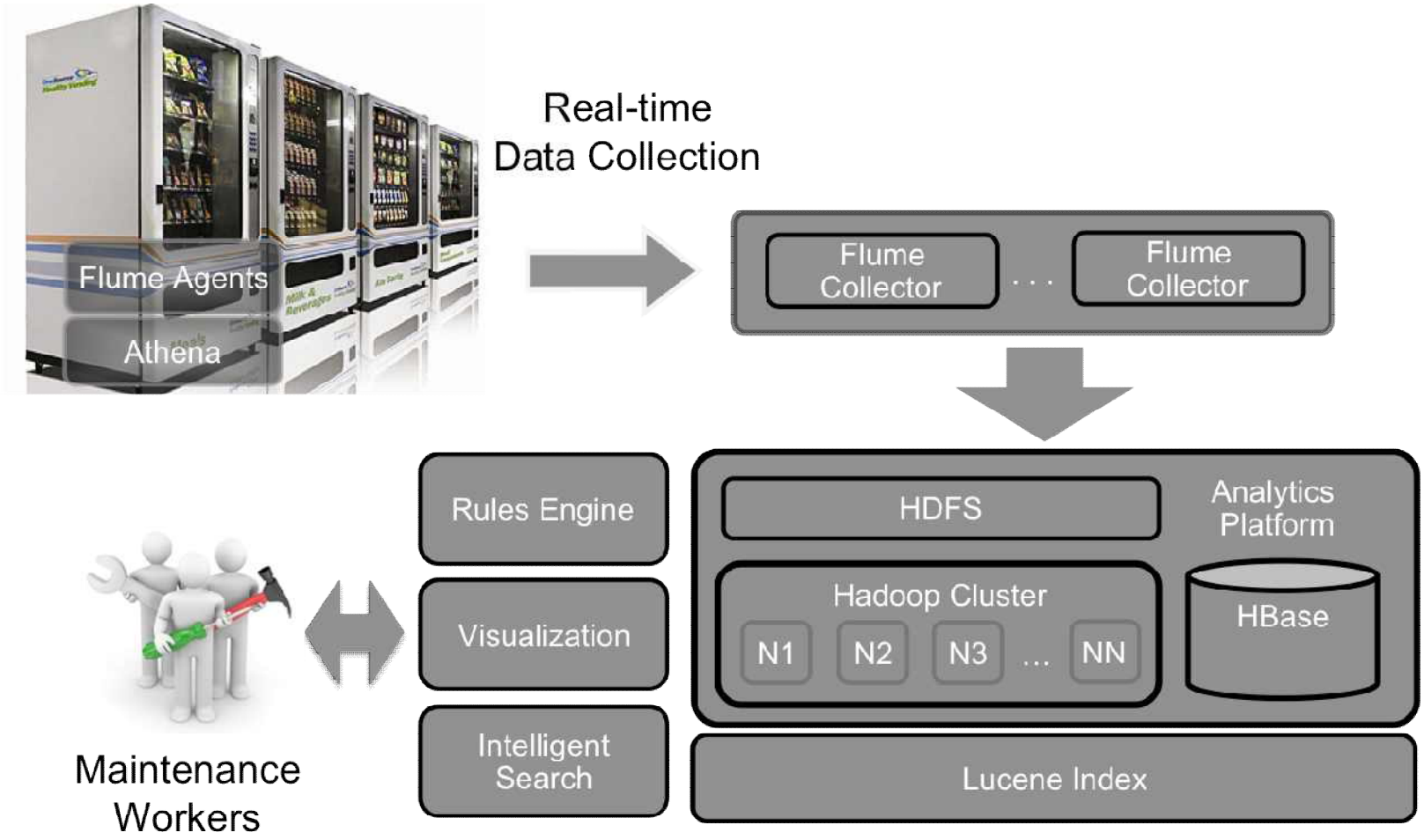
실시간 이벤트 수집 및 처리 기술

로그 수집 및 분배 기술

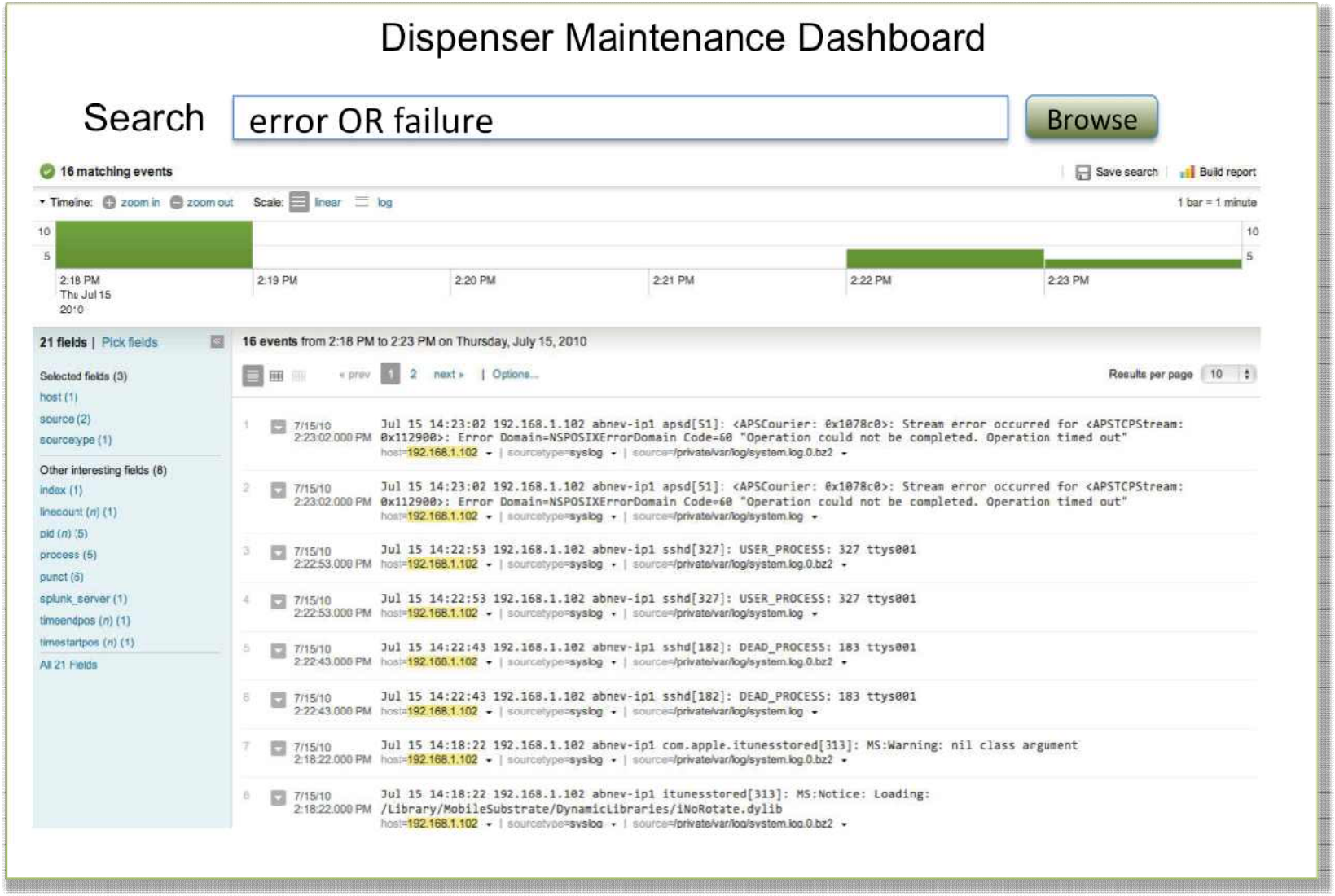
대용량 데이터의 배치 처리 및 분석 기술

통합 기술

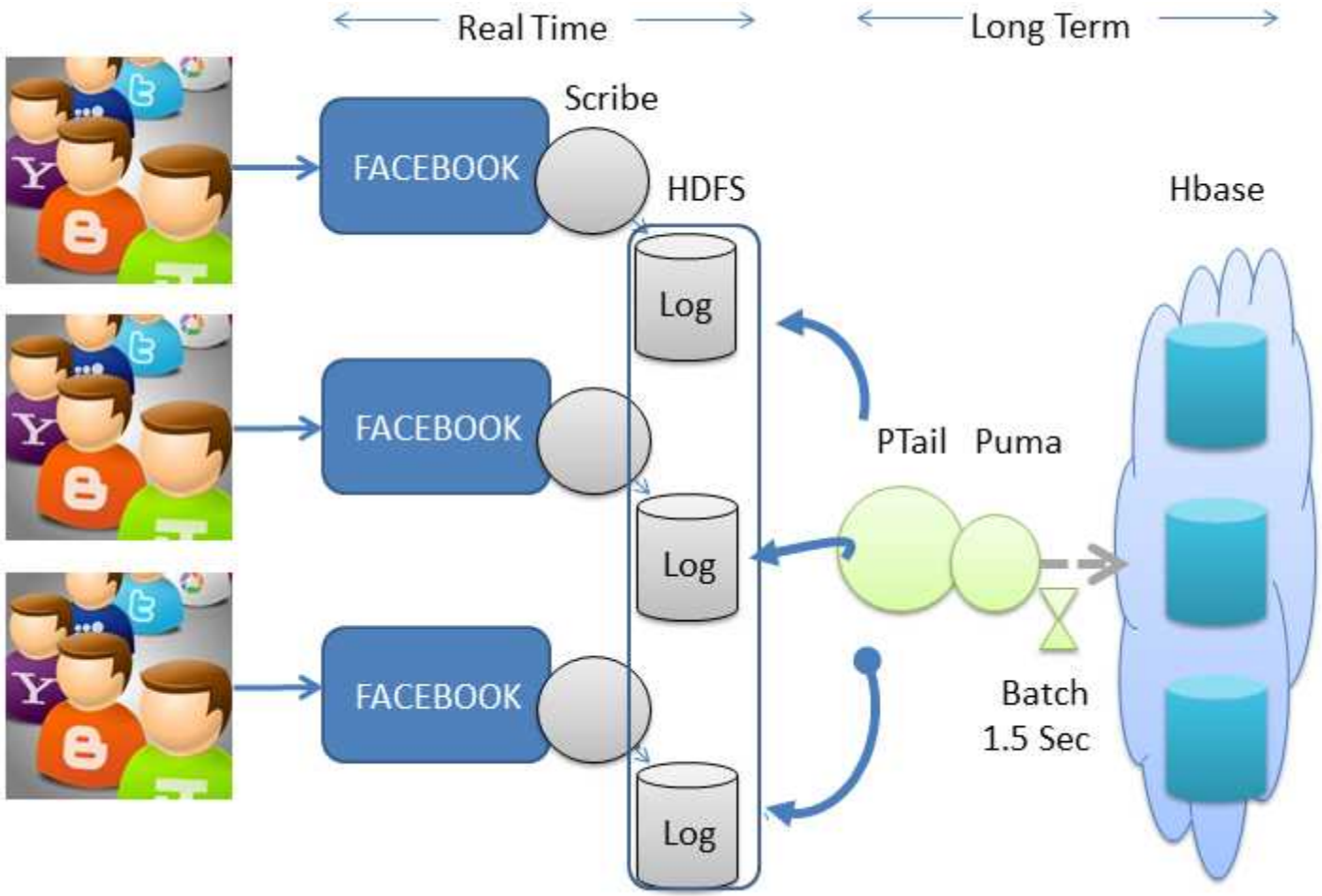
Use-Case: Dispenser



Use-Case: Dispenser



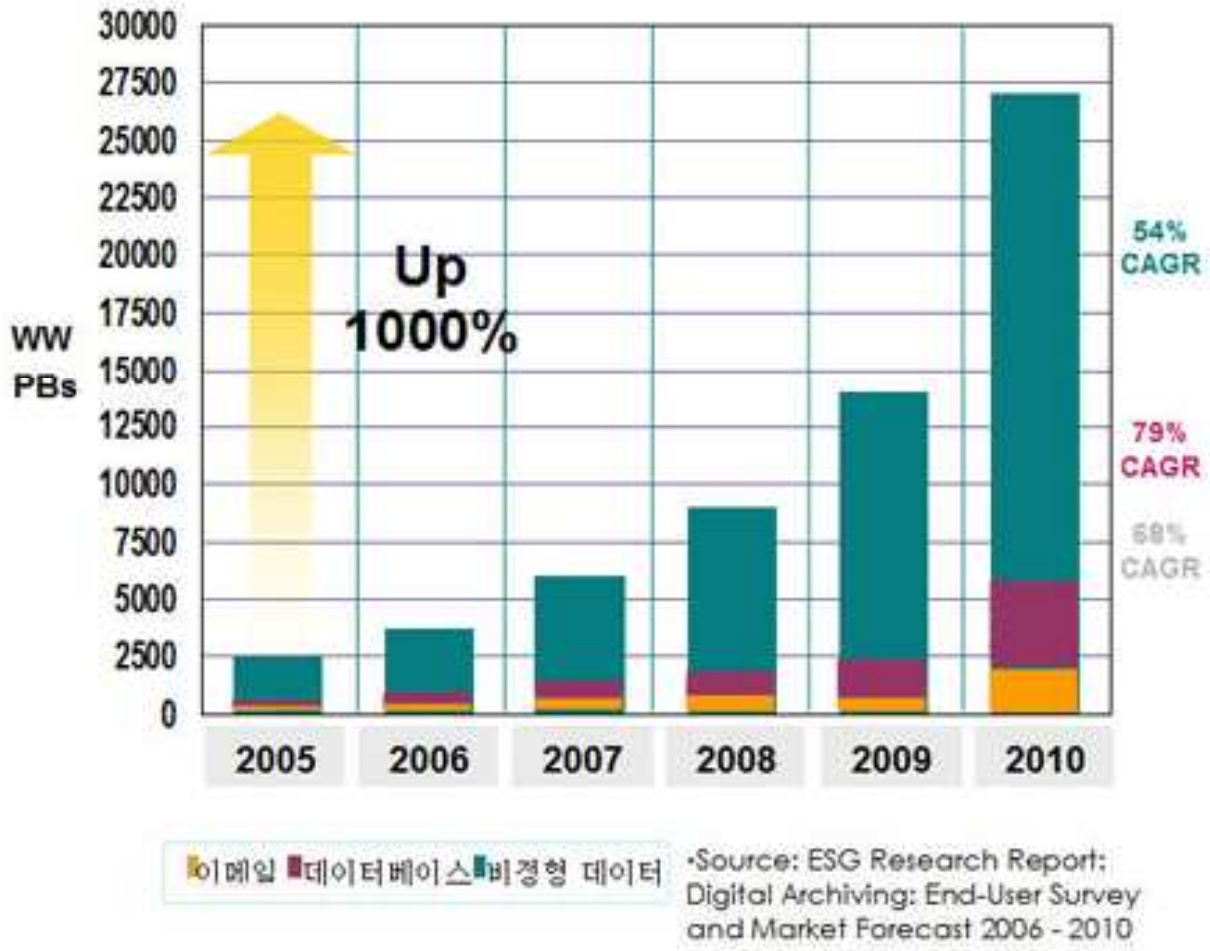
Facebook Real Time Analytics System



Unstructured Data

- 데이터는 유형, 특성 등으로 잘 저장해야 잘 사용할 수 있다!
- 데이터의 약 20%는 정형 데이터, 80%는 비정형 데이터
- Blog, SNS, Mobile 등을 통해 비정형 데이터는 더욱더 빠르게 증가 추세
- 80%의 비정형 데이터를 어떻게 분석할 것인가?
- 비정형 데이터를 이용하여 의미 있는 정보를 찾아내는 것은 정형 데이터를 분석하는 것보다 훨씬 복잡

Unstructured Data



•Source: ESG Research Report: Digital Archiving: End-User Survey and Market Forecast 2006 - 2010

Unstructured Data - NMON Log

NMON Log

하나의 정보가 구조화되지 않은 형태로 불규칙적으로 생성

```
TOP,+PID,Time,%CPU,%Usr,%Sys,Size,ResSet,ResText,ResData,ShdLib,MinorFau
lt,MajorFault,Command
BBBP,000,/etc/release
BBBP,001,/etc/release,"CentOS release 5.7 (Final)"
BBBP,002,lsb_release
BBBP,003,lsb_release,"LSB Version:  :core-4.0-amd64:core-4.0-
ia32:core-4.0-noarch:graphics-4.0-amd64:graphics-4.0-ia32:graphics-4.0-
noarch:printing-4.0-amd64:printing-4.0-ia32:printing-4.0-noarch"
BBBP,004,lsb_release,"Distributor ID:        CentOS"
BBBP,005,lsb_release,"Description:  CentOS release 5.7 (Final)"
BBBP,006,lsb_release,"Release:      5.7"
BBBP,007,lsb_release,"Codename:     Final"
...
```

다수의 라인으로 구성된 불규칙 로그를

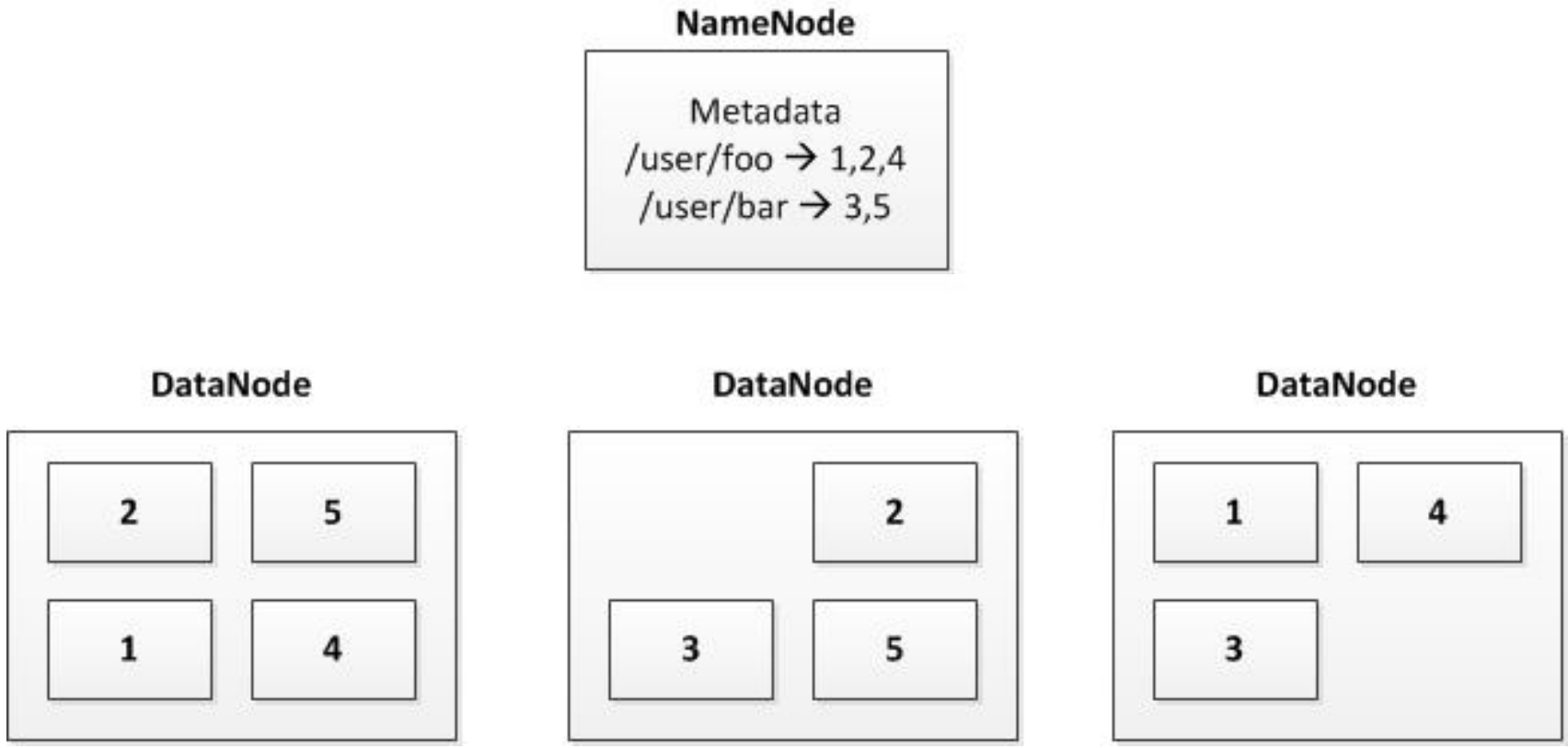
정형화된 로그로 변형하는 것이 관건

MapReduce & DFS: Apache Hadoop

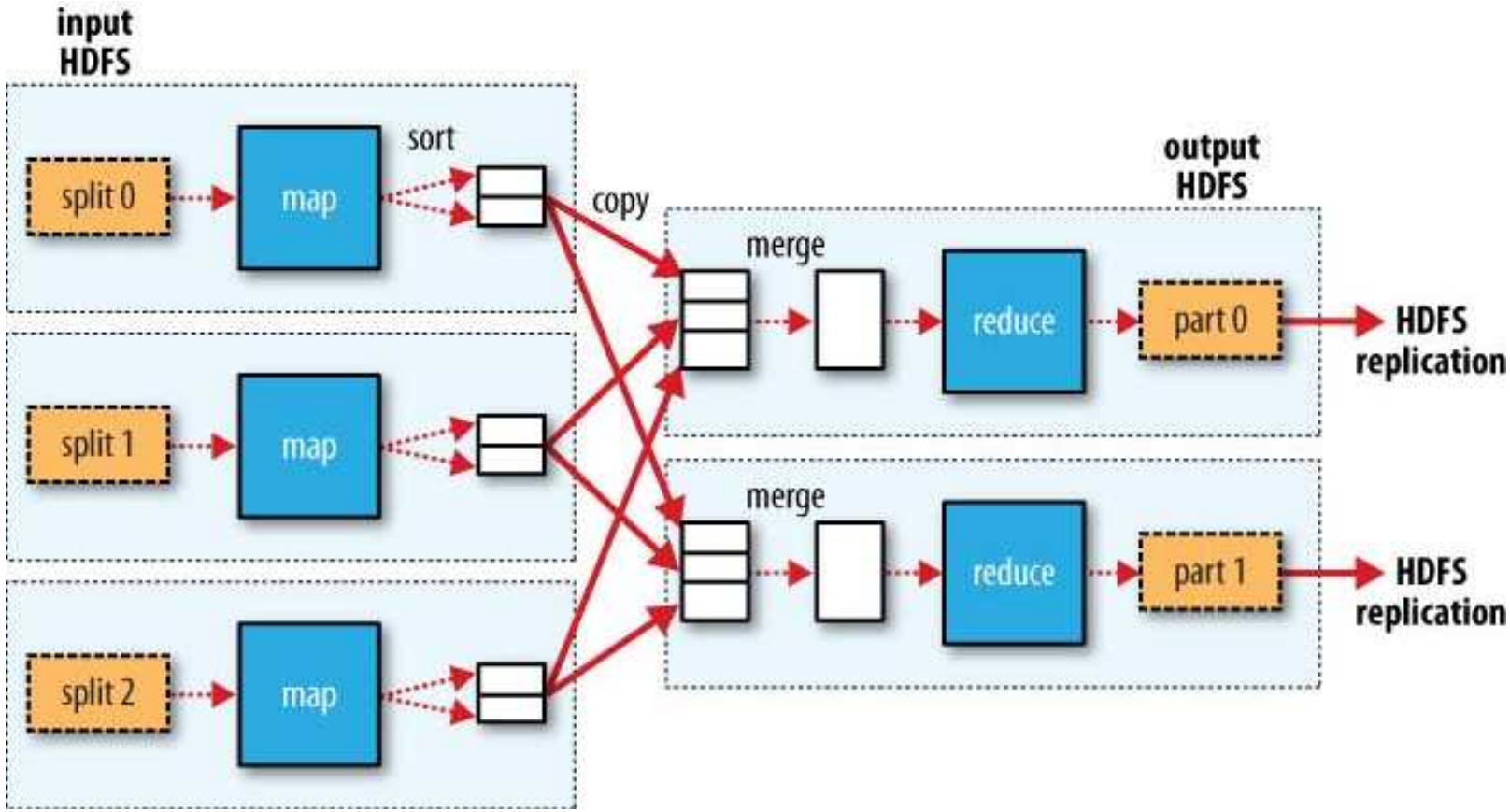
- File System : HDFS(Hadoop Distributed File System)
  - 파일을 64M 단위로 나누어 장비에 나누어서 저장하는 방식
  - 사용자는 하나의 파일로 보이나 실제로는 나누어져 있음
  - 2003년 Google이 논문으로 Google File System을 발표

- 프로그래밍 모델(MapReduce) (2004년 Google이 논문 발표)
  - HDFS의 파일을 이용하여 처리하는 방법을 제공
  - Parallelization, Distribution, Fault-Tolerance …

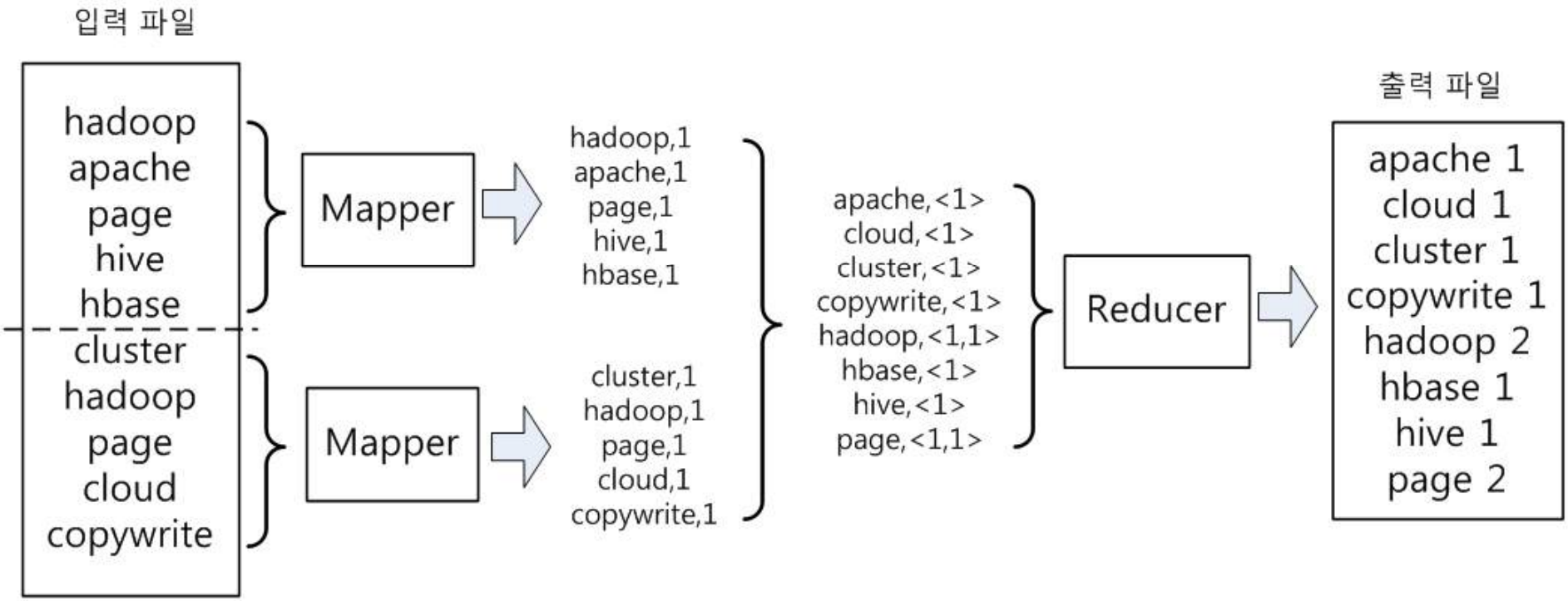
HDFS


NameNode
Metadata
/user/foo → 1,2,4
/user/bar → 3,5
DataNode
2
5
1
4
DataNode
2
3
5
DataNode
1
4
3

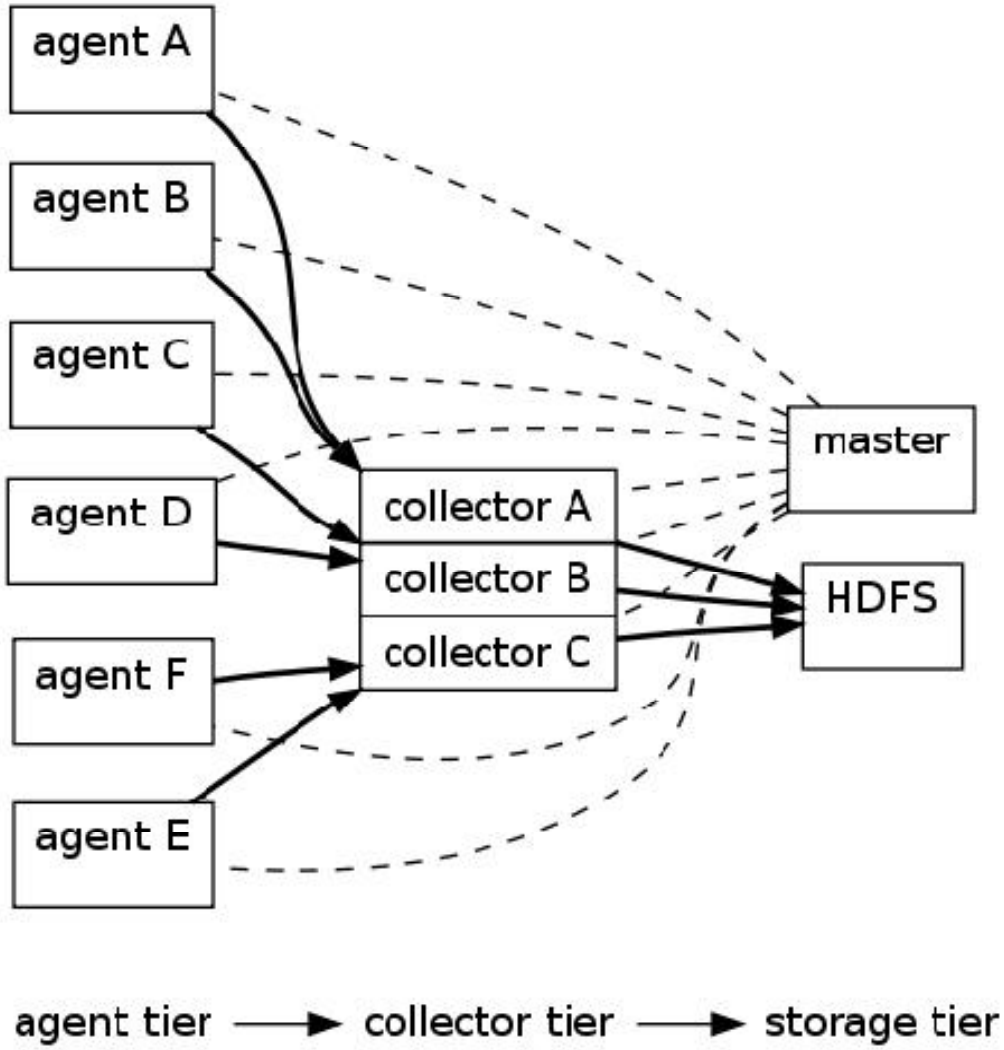
- HDFS의 파일을 처리하기 위한 프로그래밍 모델



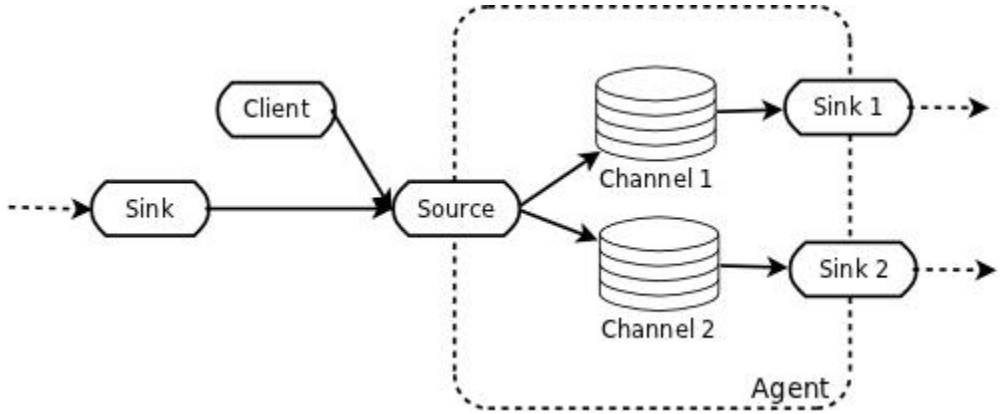
MapReduce: WordCount

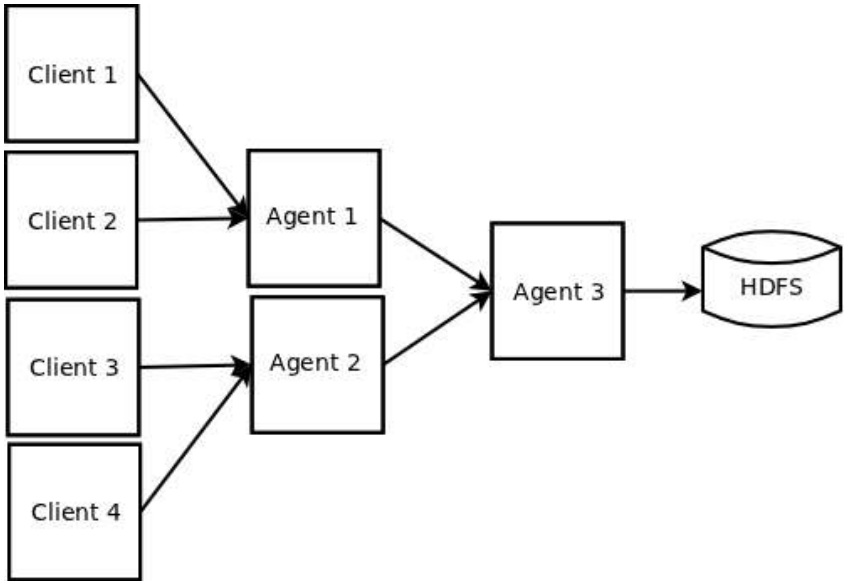


Log Aggregator – Apache Flume



Log Aggregator – Apache Flume





Complex Event Processing?

- CEP의 전통적인 영역
  - prediction, observation, dissemination
  - behavior, active diagnostic

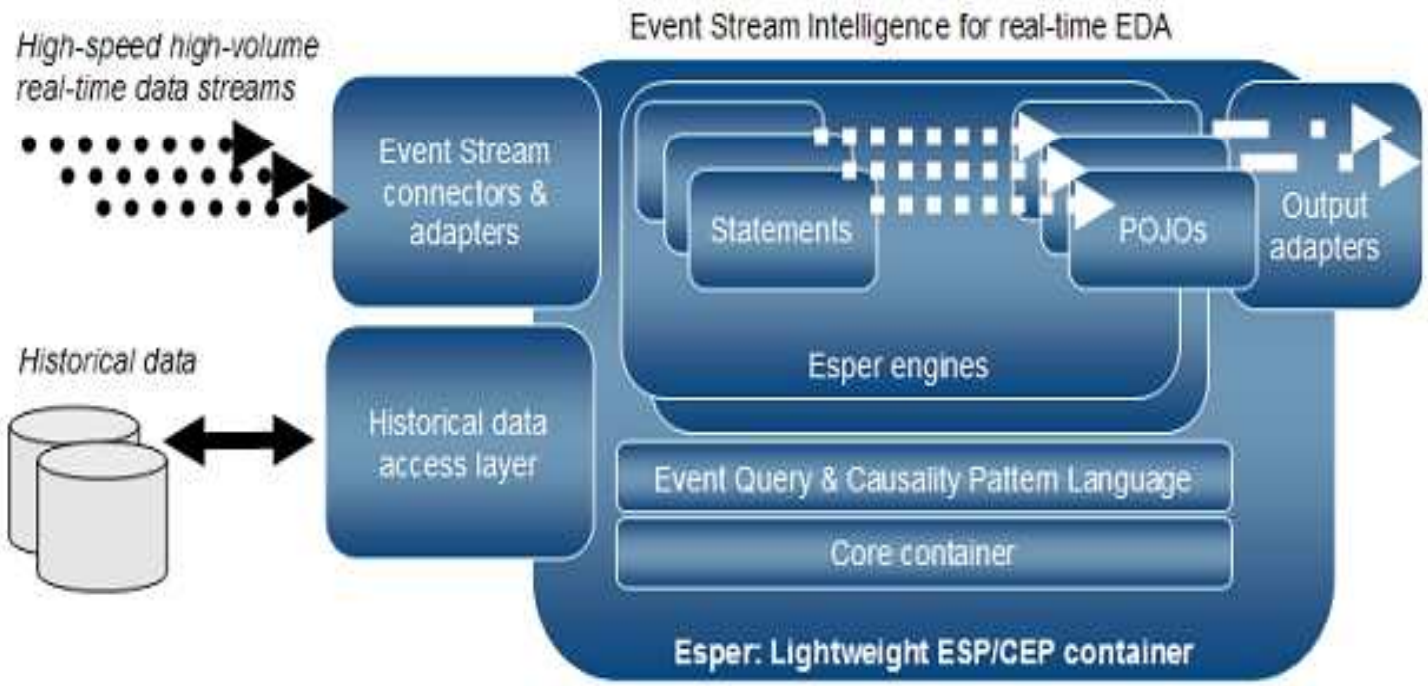
용어	정의
이벤트	실제로 발생한 사건, 일, 메시지 상태의 변경 특정한 액션 또는 상태의 변화를 통해 발생하는 변경 이 불가능한 과거의 기록
이벤트 스트림	시간의 순서대로 연속되는 이벤트의 흐름 시작과 끝이 없는 이벤트의 연속된 흐름
실시간의 특징	현저하게 낮은 수준의 지연 일정한 응답속도 예측 가능한 성능

Complex Event Processing의 특징

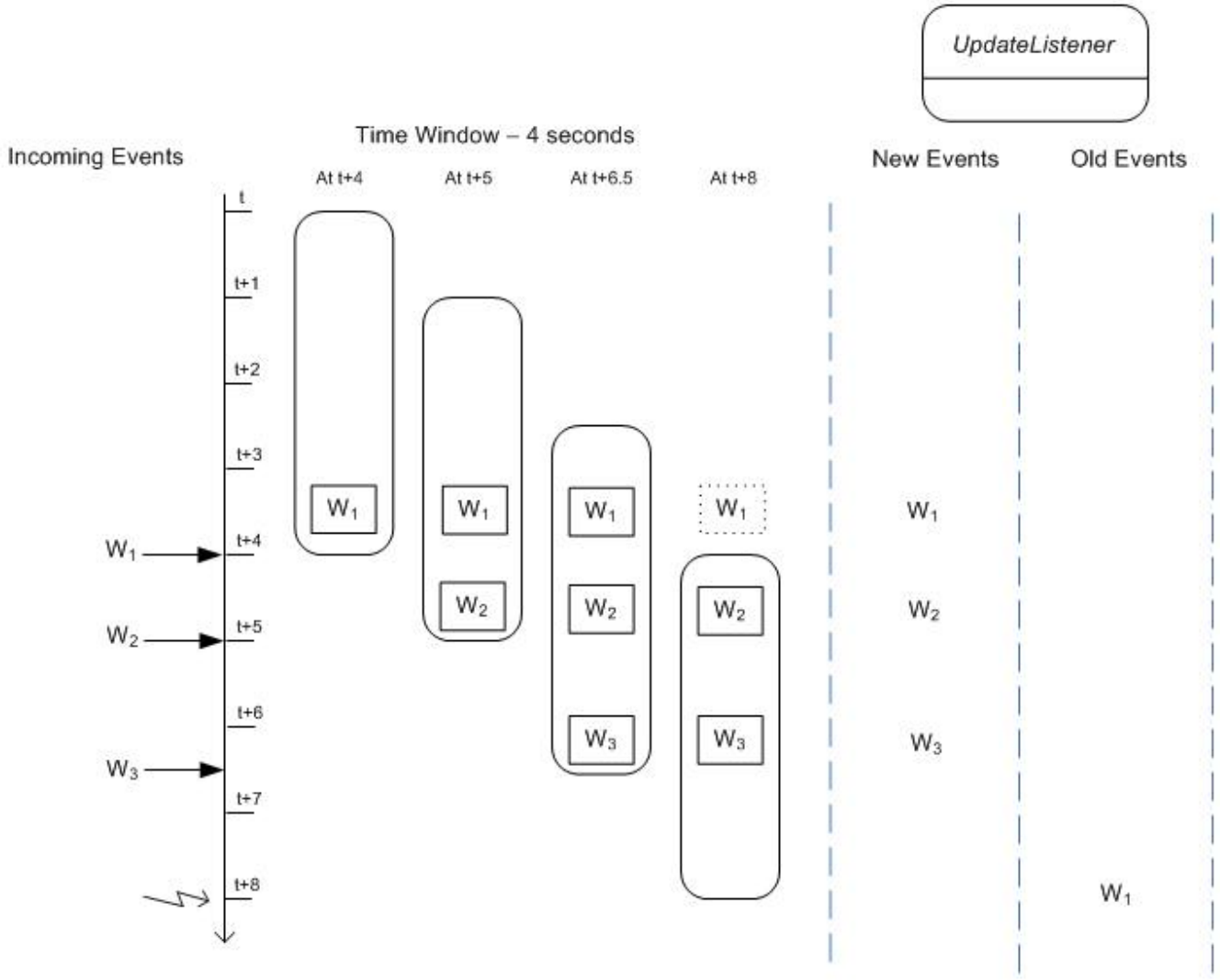
- 데이터의 이동을 일정한 시간 동안 유지
  - 길이가 Nm인 수도관에 물을 흘러가는 것과 같은 이치
- 데이터의 이동(스트림)을 조회 또는 질의
  - 직장이 강남인 30~35세 여성 중 강남역에서 10분 동안 있었던 사람만
- 수집한 데이터 스트림을 처리
- 데이터 스토어(예; Database)와 스트리밍 데이터 결합

OpenSource CEP : Esper

- GPL 라이선스, Oracle CEP의 모태로 알려져 있음
- 경량의 Complex Event Processing Implementation



OpenSource CEP : Esper



OpenSource CEP : Esper

```
// Java Object
public static class StockTick {
        String symbol;
        Double price;
        Date timeStamp;
}

// Esper Event Query (EPL)
// Apple의 Tick이 평균 60이상, 2건 발생한 경우
select * from StockTick(symbol='AAPL').win:length(2) having
avg(price) > 6.0
```

OpenSource CEP : Esper

```
EPServiceProvider epService = EPServiceProviderManager.getDefaultProvider();
String expression = "select avg(price) from
                    org.myapp.event.OrderEvent.win:time(30 sec)";
EPStatement statement = epService.getEPAdministrator().createEPL(expression);

public class MyListener implements UpdateListener {
    public void update(EventBean[] newEvents, EventBean[] oldEvents) {
        EventBean event = newEvents[0];
        System.out.println("avg=" + event.get("avg(price)"));
    }
}

MyListener listener = new MyListener();
statement.addListener(listener);
```

OpenSource CEP : Esper

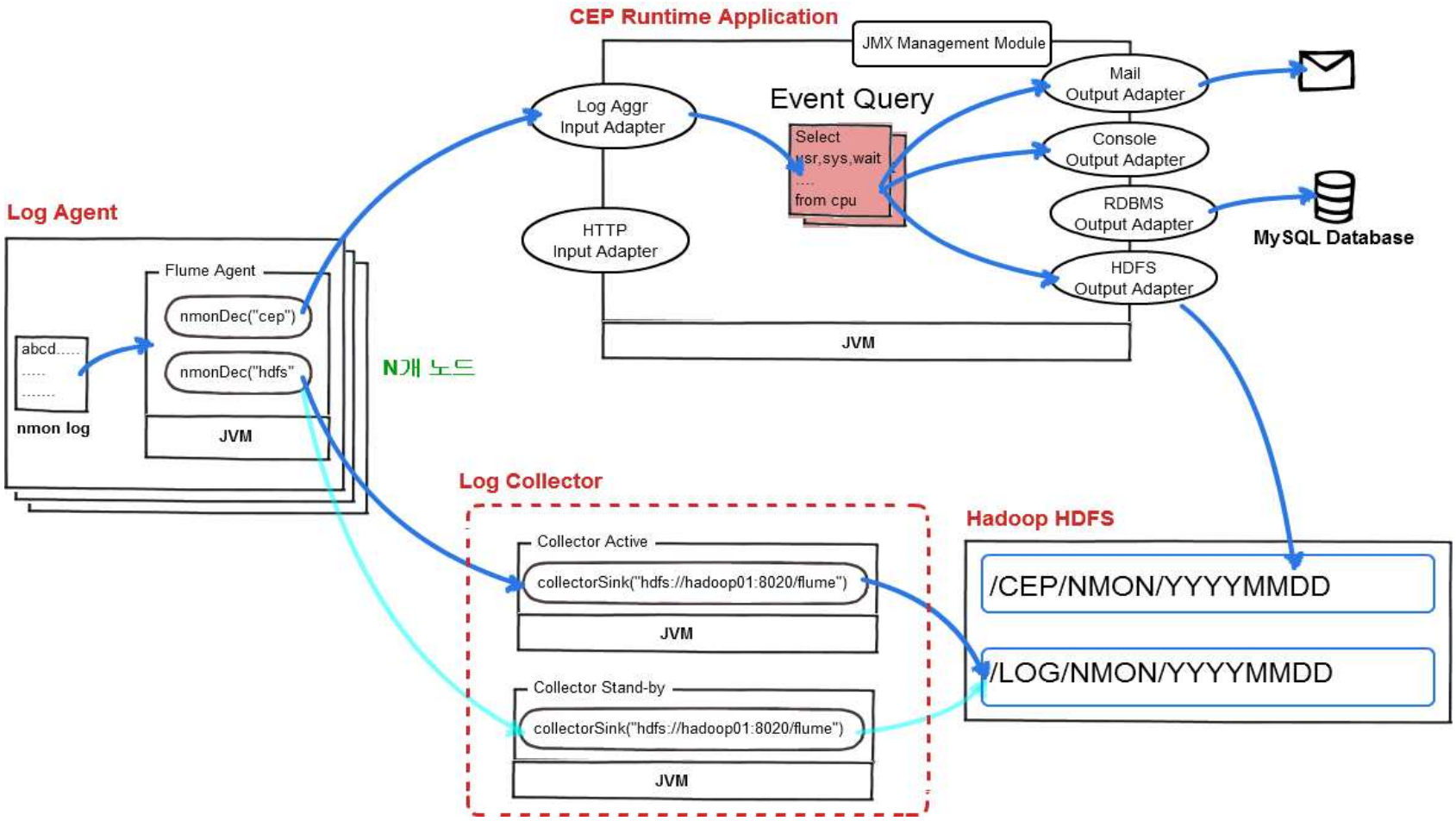
- **최근 20분내 시청 근처에 있었던 급여가 20M 이상이며 나이가 30~35세 이상이고 집이 강남이면서 취미가 쇼핑인 여성**

```
select * from customer(age=' 30~35' ,gender='male',salary>20M,
        location='city hall',
       home='gangnam',hobby='shopping').win:time(20 min)
```

Volume : Analytics
(Apache Hadoop)

Velocity : Real-Time
(CEP; Esper)

로그 수집기와 CEP를 이용한 실시간 처리 아키텍처



Flume + Esper + RHQ Demo

Summary

- Real-Time Big Data는
  - Real-Time과 Analytics의 Convergence
  - High Technology

- 아직까지 Real-Time 이벤트 처리 기술인 CEP에 대한 이해 부족으로 인하여 시장에서 적용 사례 부족

- 향후 Big Data 시장에서 강력한 폭풍이 될 것

References

- Big Data Use-Case: Real-time Dispenser Maintenance
  - http://jameskaskade.com/?p=2177
- Real Time analytics for Big Data: Facebook's New Realtime Analytics System
  - http://tinyurl.com/3cgg6yr
- Esper Documentation
  - http://esper.codehaus.org/esper-4.5.0/doc/reference/en/html_single/index.html