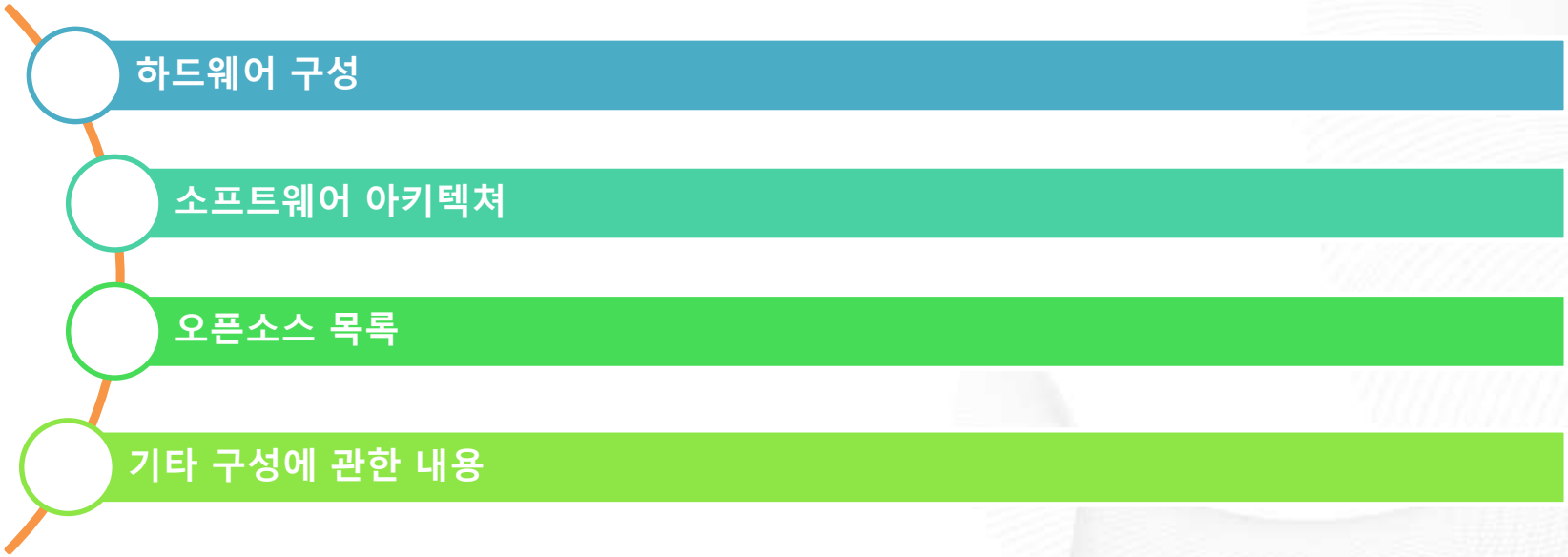
빅데이터 구축 사례

오태현
paranmul@gmail.com
paranmul@inbrein.com
INBREIN

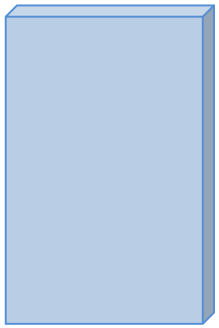
❖개요

▪빅데이터 구축 사례 소개

❖주요 진행 내용

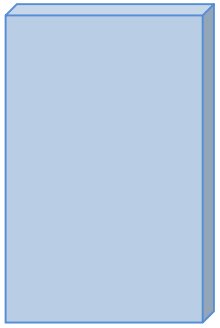


하드웨어 구성



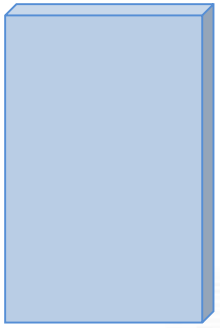
x 3

Master server



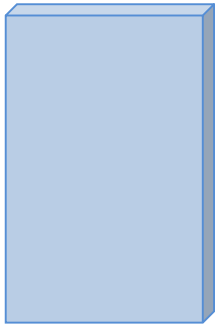
x 7

Hadoop
Hbase



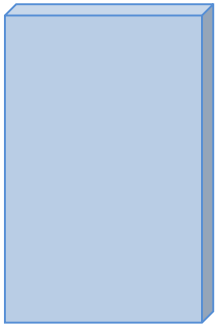
x 2

Kafka



x 4

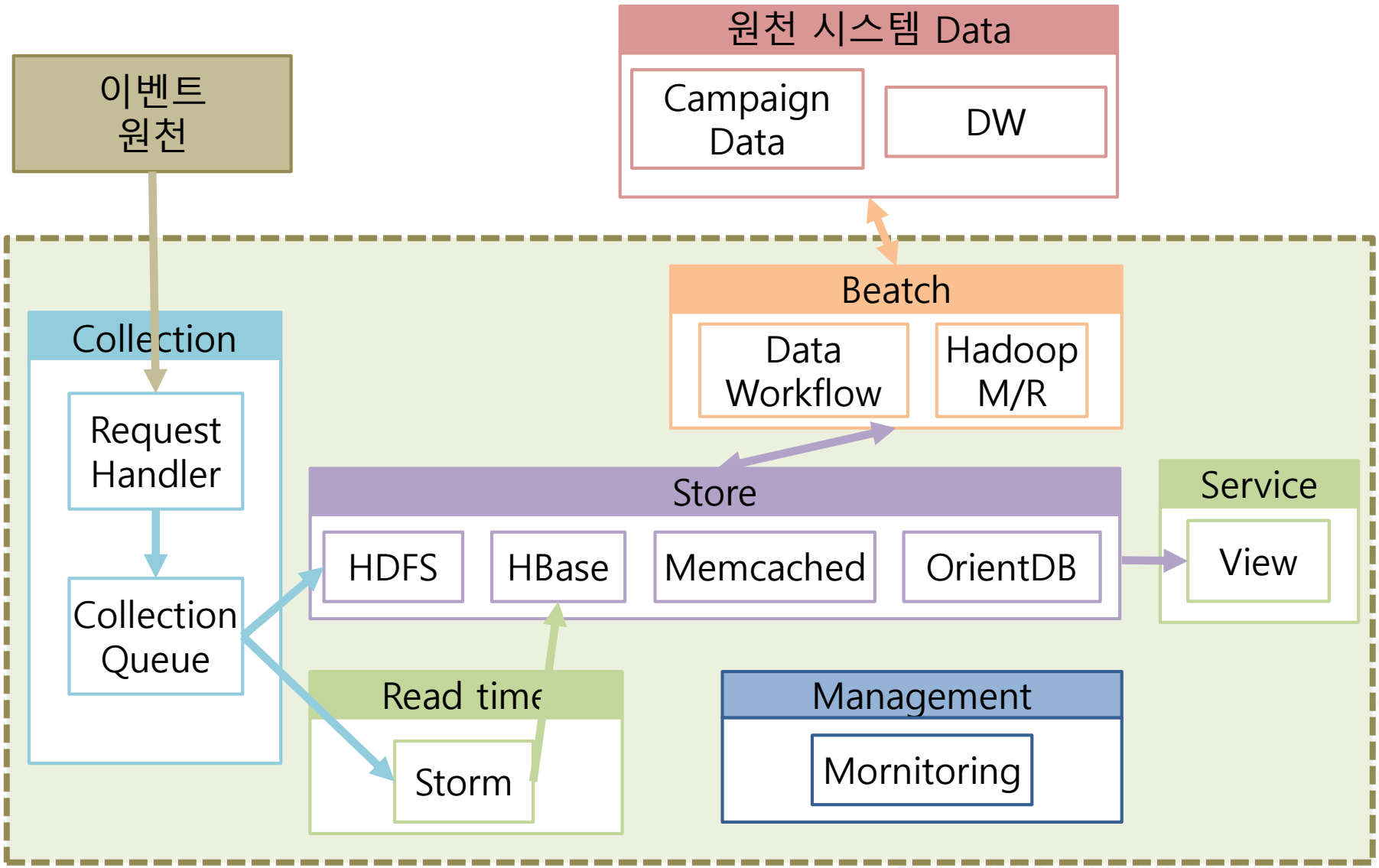
Storm



x 2

Backup node

소프트웨어 아키텍쳐


원천 시스템 Data
Campaign Data
DW
이벤트 원천
Beatch
Data Workflow
Hadoop M/R
Collection
Request Handler
Collection Queue
Store
HDFS
HBase
Memcached
OrientDB
Service
View
Read time
Storm
Management
Mornitoring

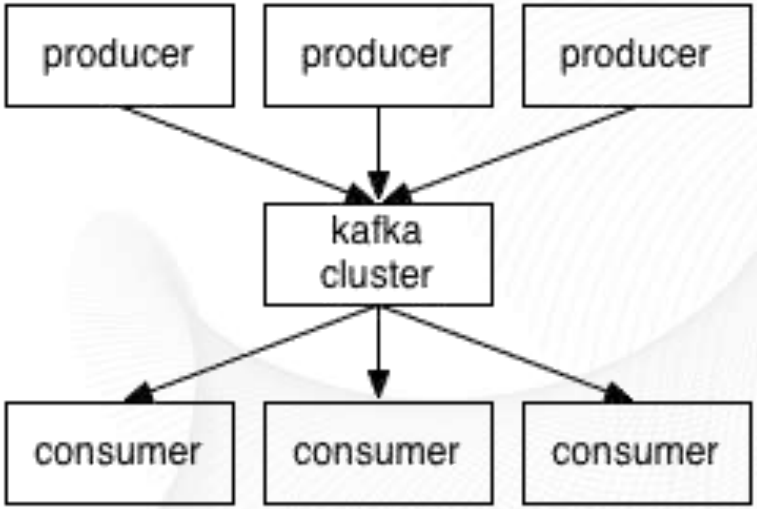
사용된 오픈소스 목록

- Verte.x
- Flume
- Kafka *
- Camus
- Storm *
- Hive
- Hadoop
- jCascalog *
- Sqoop *
- Hbase
- Azkaban *
- Graphite *
- Ganglia *
- Memcached
- MariaDB

Kafka

- 데이터를 분산, 파티션, 복제 커밋 로그서버
- Topic
- 초당 18000개, 로그 하나 크기 1k,
- 1초에 17M(한 서버당),1분에 840M, 1시간에 50G

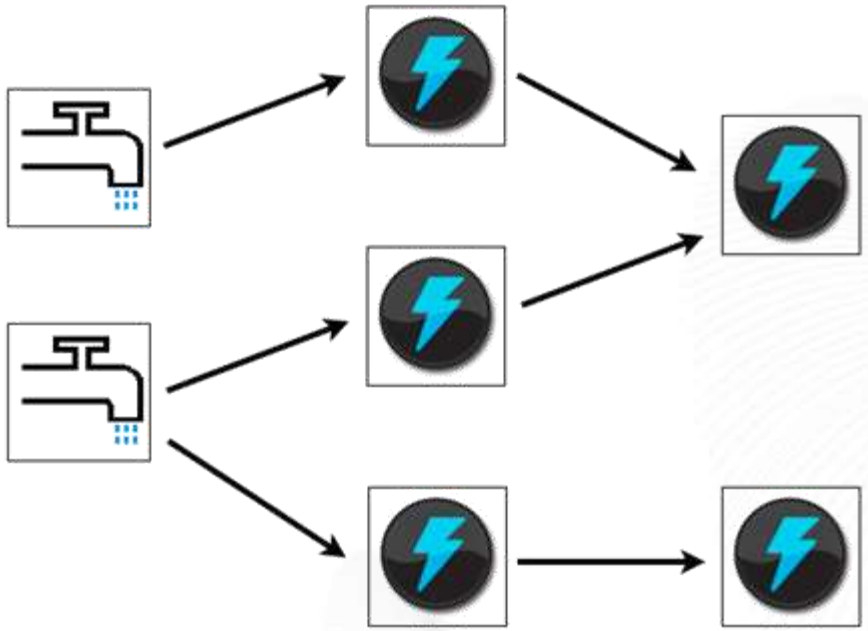
- 두대의 Kafka server를 사용



Storm

- Realtime

- Nimbus
- Topology
- Spout
- Bolt
- -> Hbase



jCascalog

- Hadoop 의 mapreduce 코딩을 쉽게 하는 오픈소스
- Cascading을 java로 사용할 수 있게함

mbrid	pcid
101	001
102	002
103	003
104	004
105	005
106	006

```
Tap inputTab = new Hfs(new TextDelimited(false, "\t"), inputPath);

Tap outputTap = new Hfs(new TextDelimited(false, "\t" , "\""), outputPath);

Subquery base = new Subquery("!mbr-id-long", "!pcid-cnt")
        .predicate(inputTab, "!mbr-id", "!pcid-cnt")
        .predicate(new ConvertData(),"!mbr-id").out("!mbr-id-long");

Api.execute(outputTap, base);
```

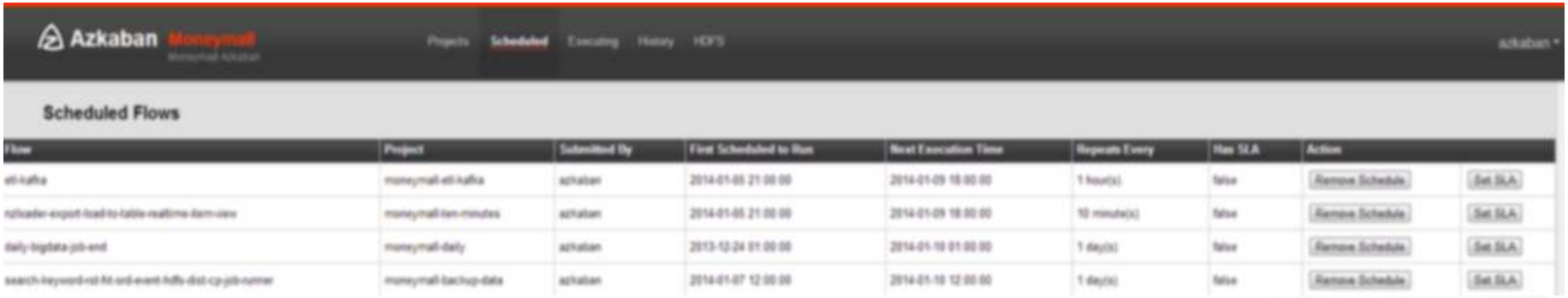

Sqoop

- RDB 와 HDFS사이에 데이터를 전달 하는 오픈소스
- pwd, mssql, mariaDB, Oracle, netezza, mysql, hbase

```
import
-D      transaction.isolation.level=READ_UNCOMMITTED
--num-mappers    1
--driver         com.microsoft.sqlserver.jdbc.SQLServerDriver
--connect        jdbc:sqlserver://ipaddress:17001;databaseName=dbname;
--username       user
--password       user
--query
[
SELECT * FROM
(
SELECT    *     FROM    tablename
) B WHERE $CONDITIONS
]
--split-by       ID
--fields-terminated-by  \t
--target-dir     /data/db
```

Azkaban

- 스케줄러

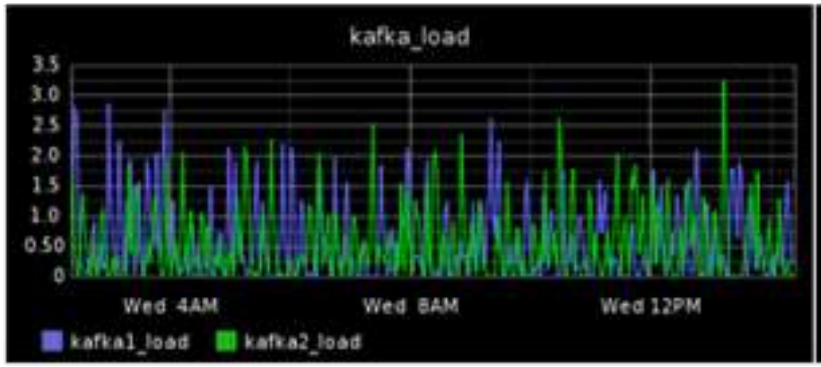


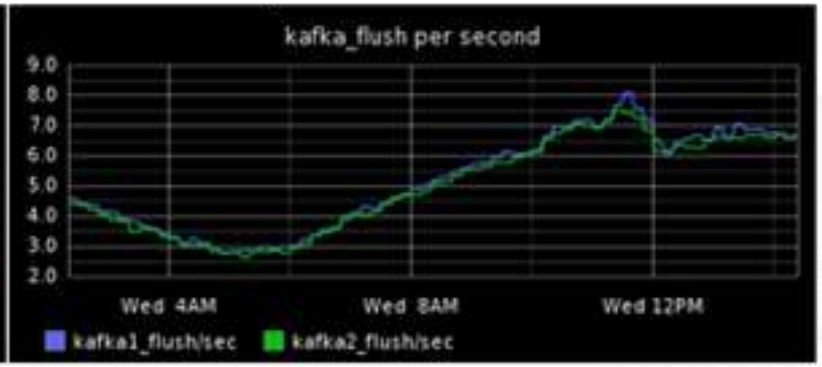
Azkaban

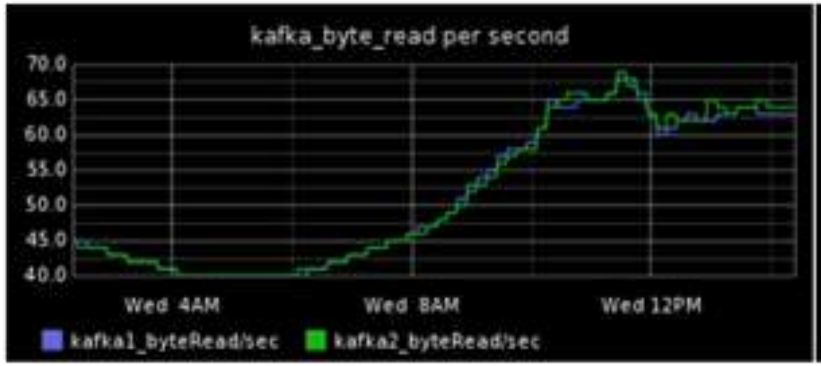
- 스케줄러

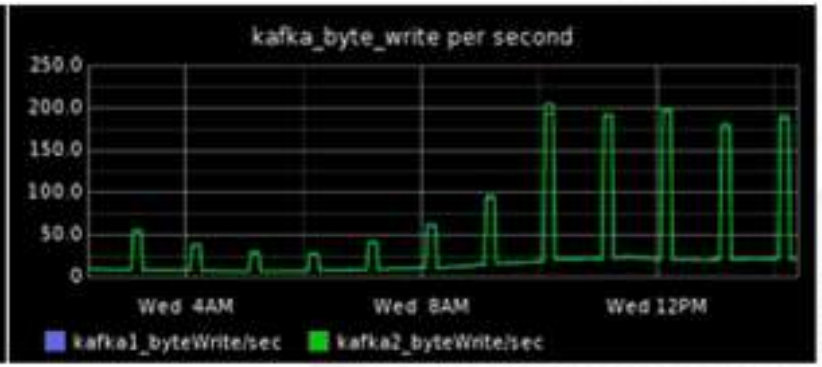
Graphite

- Python 기반
- Java 프로세스 모니터링



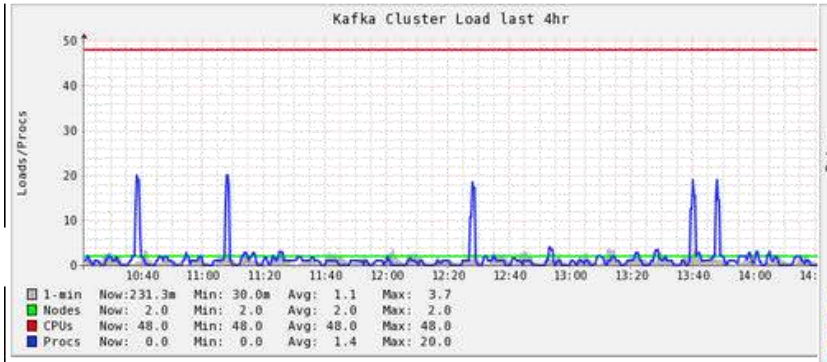


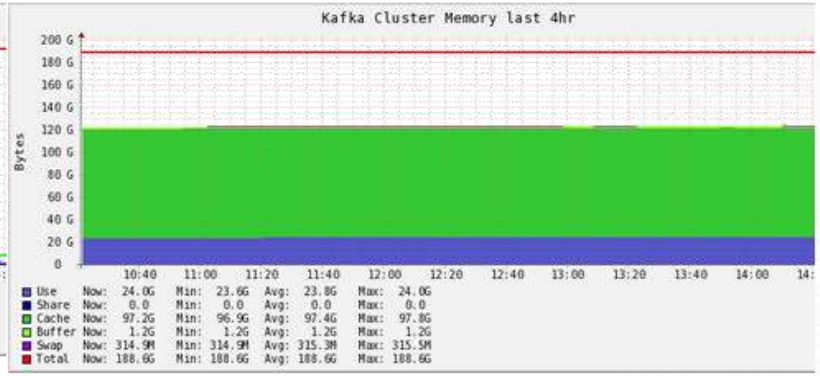


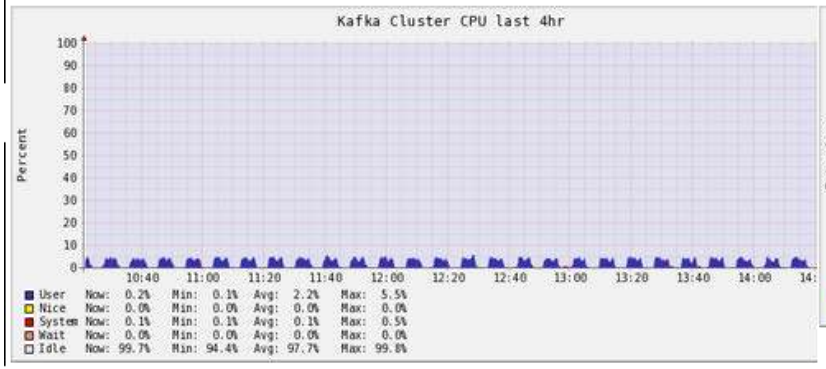


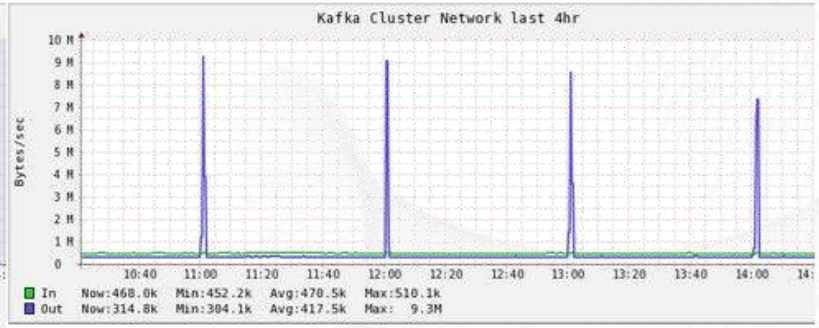
Ganglia

- 서버 자원 상황 모니터링
- 네트워크, 메모리, CPU









환경 구성에 관해서

- 기간, 인원
- 구성 중 문제점들
  - Hortonworks
  - Oozie 관리 페이지 로딩
  - Zookeeper 사용하는 오픈소스 증가 – 분리
  - 한 서버에 다양한 오픈소스 공생
  - 많은 수의 포트 오픈
  - MR job 증가
  - 개발서버 부재
  - 스톰 트라이던트 & 카프카
  - Hbase 키 설계

감사합니다