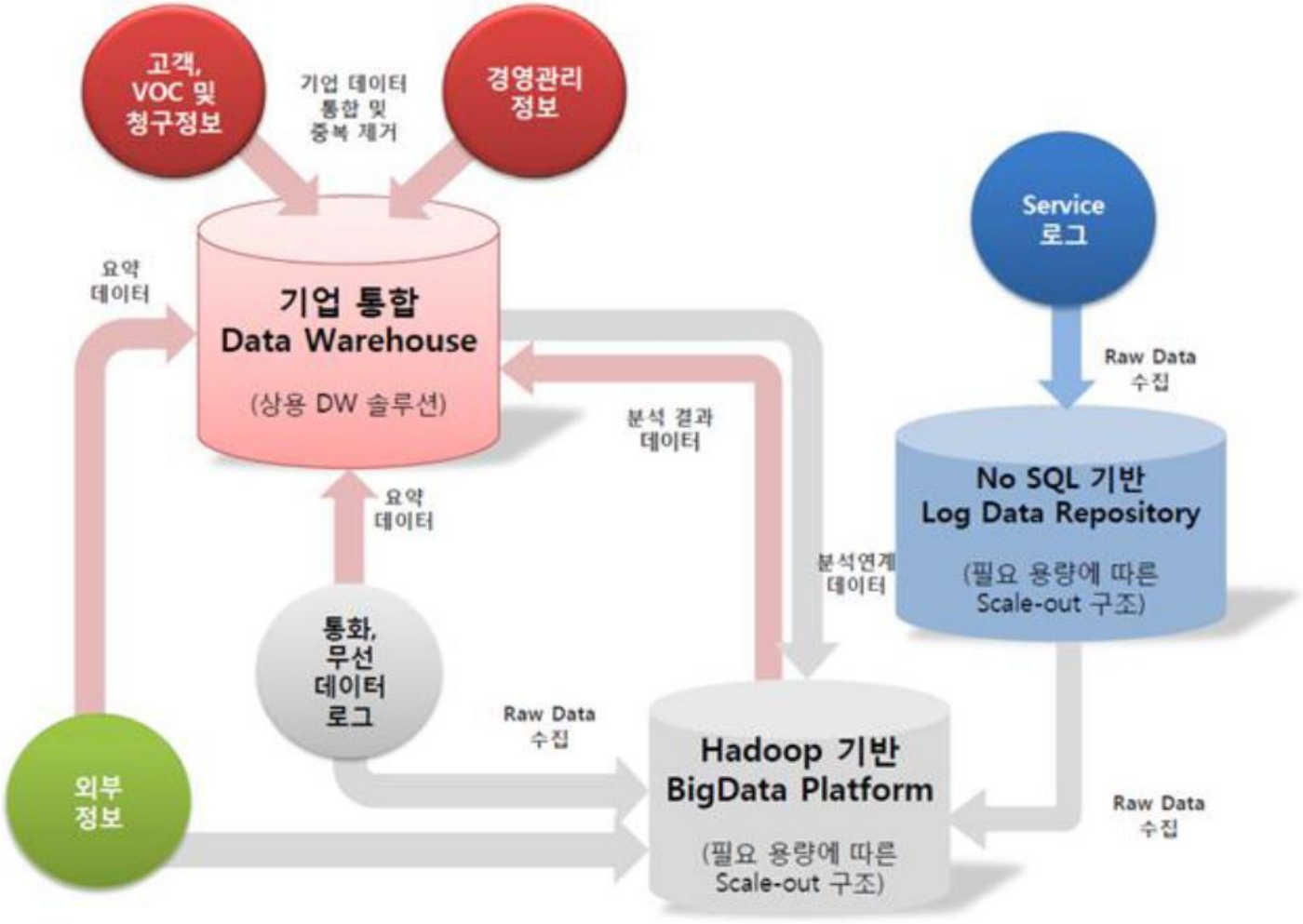
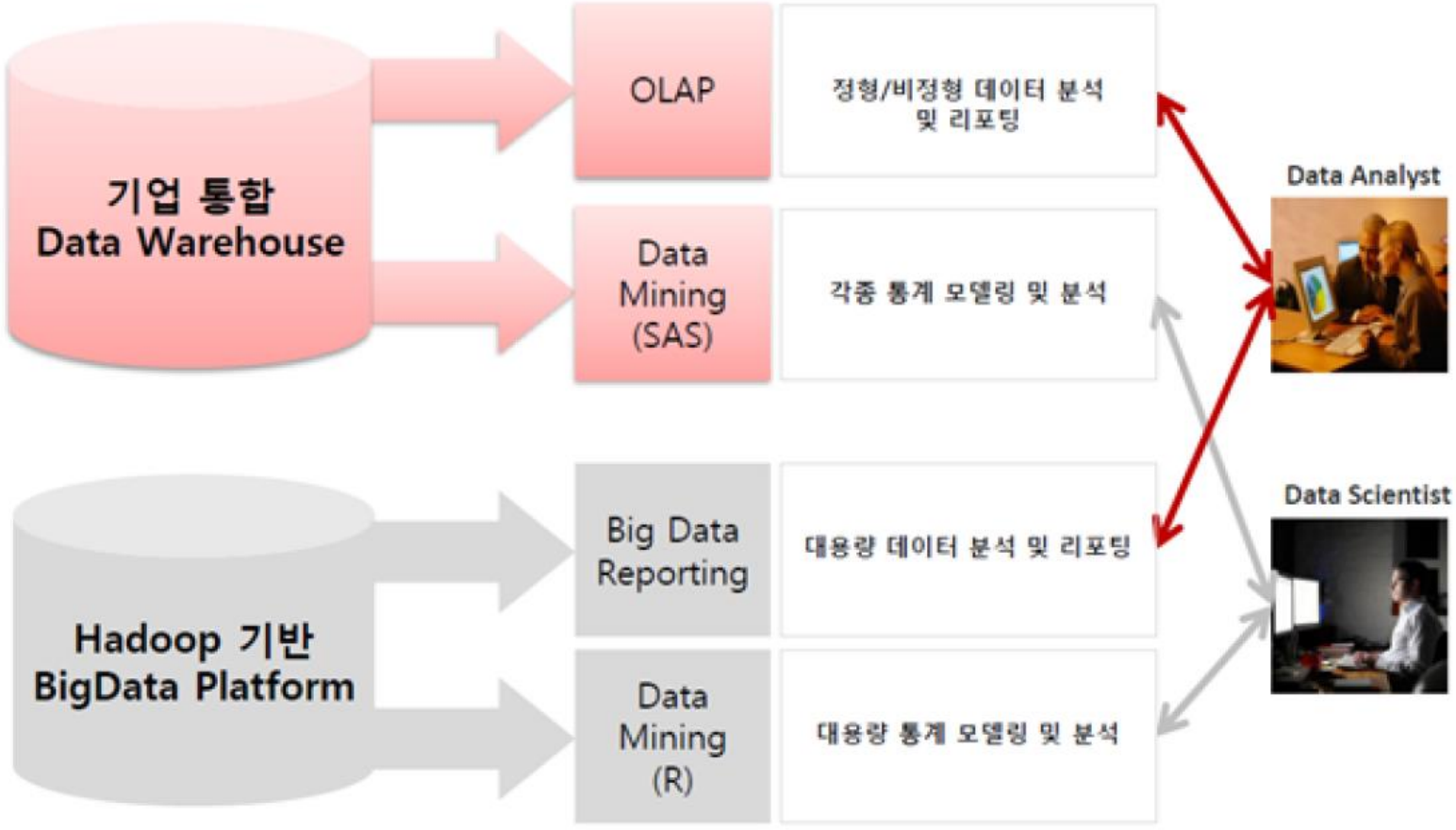
국내외 하둡적용사례

Table of Contents

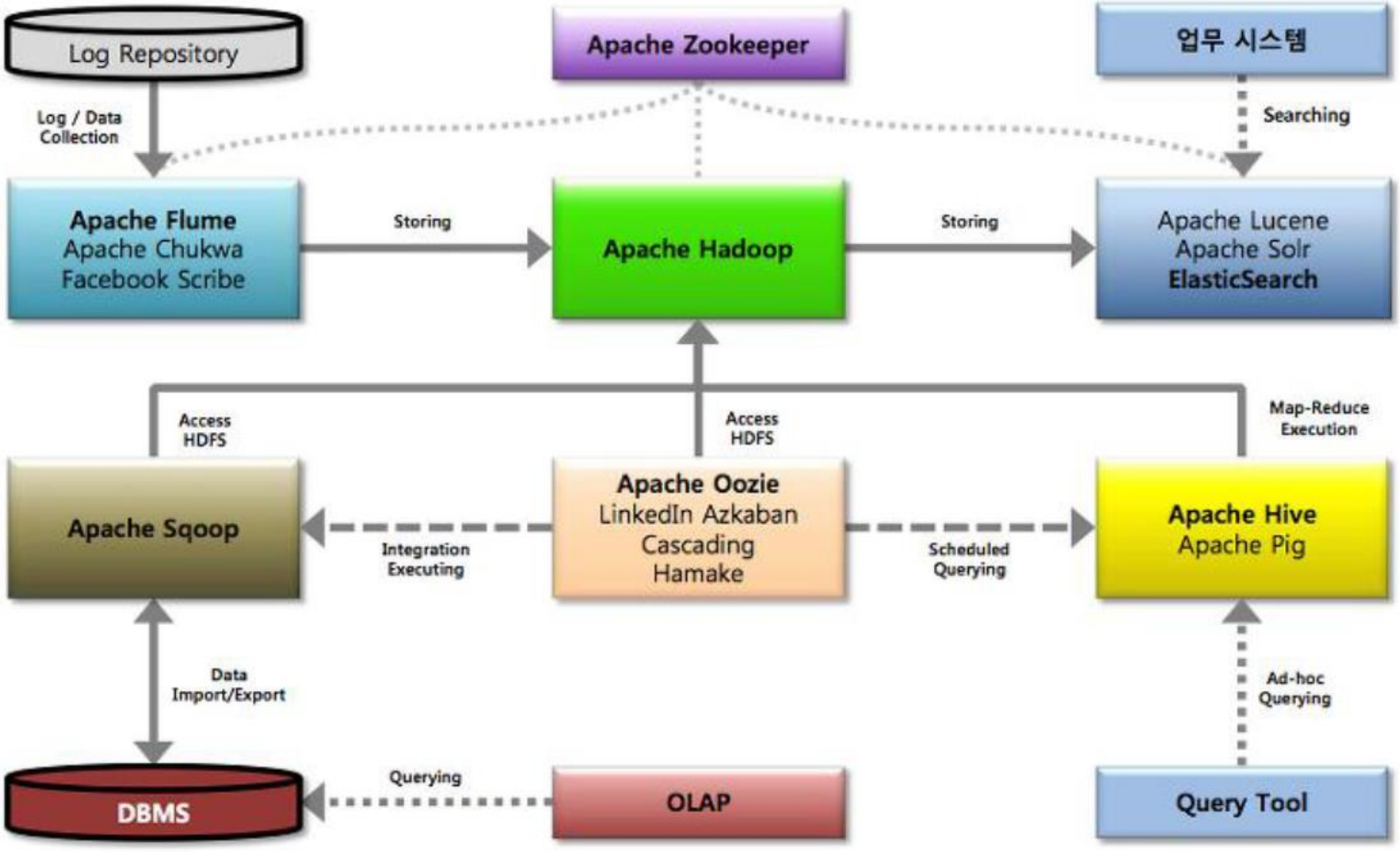
KT, 3대 기준 Data Repository



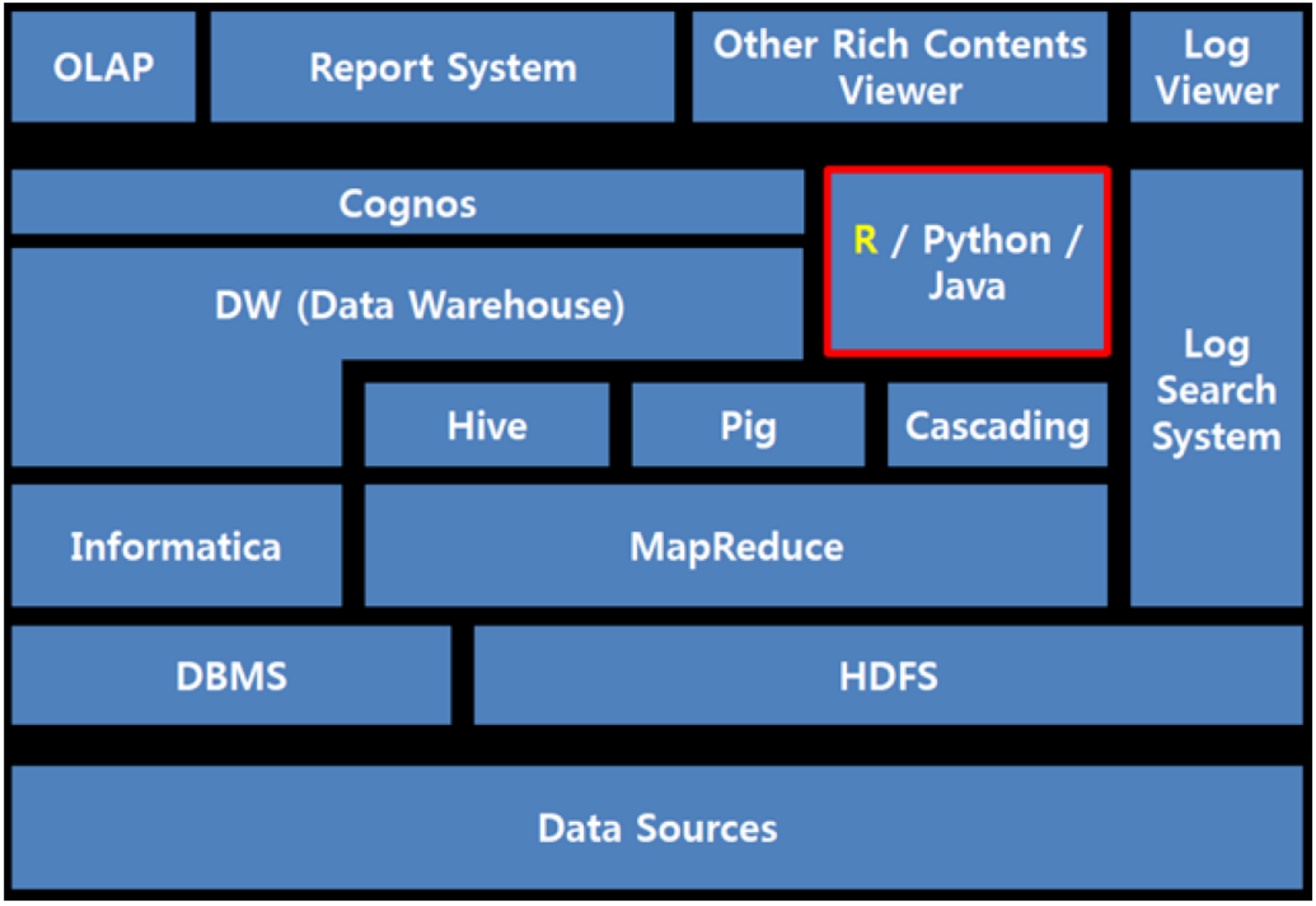
KT-데이터 분석



KT-현업 업무 아키텍처



NCSOFT 게임데이터분석

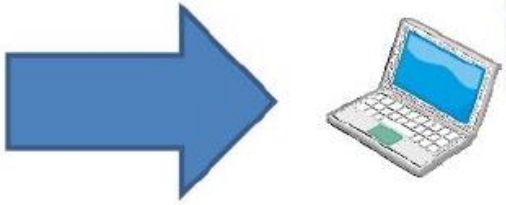


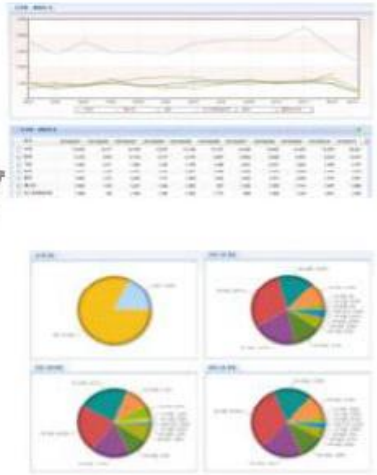
Daum 전사데이터로그분석 – 2008

access.log

1 127.0.0.1 - - [21/Feb/2007:13:47:00 -0500] "GET /console/ HTTP/1.1" 302 -
2 127.0.0.1 - - [21/Feb/2007:13:47:00 -0500] "GET /console/portal/welcome HTTP/1.1" 200 6174
3 12 [21 07:1 0500 /console/main.css HTTP/1.1" 200 8694
4 12 [21 7:1 0500 onsole/favicon.ico HTTP/1.1" 200 3638

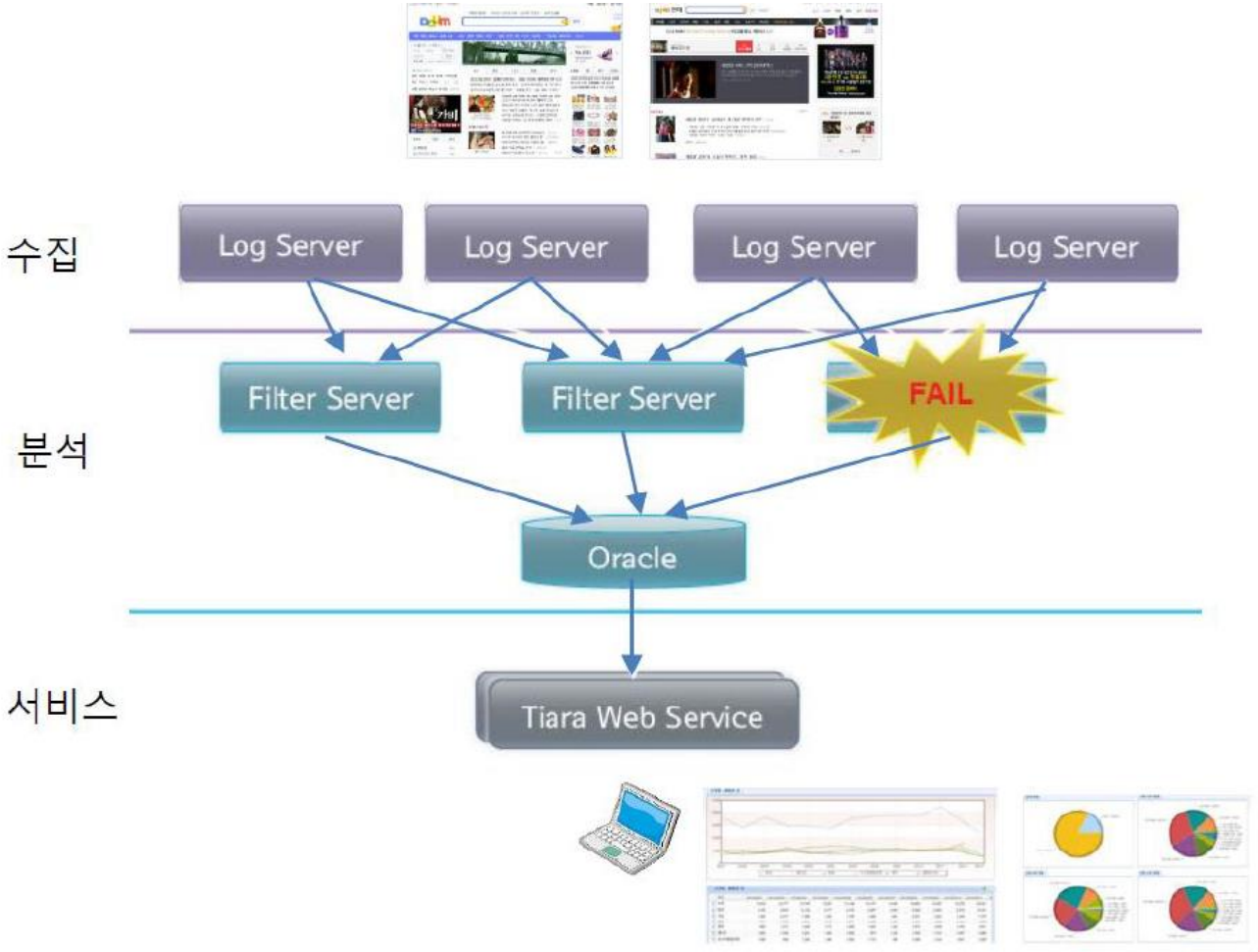
전사 서버 로그 수집



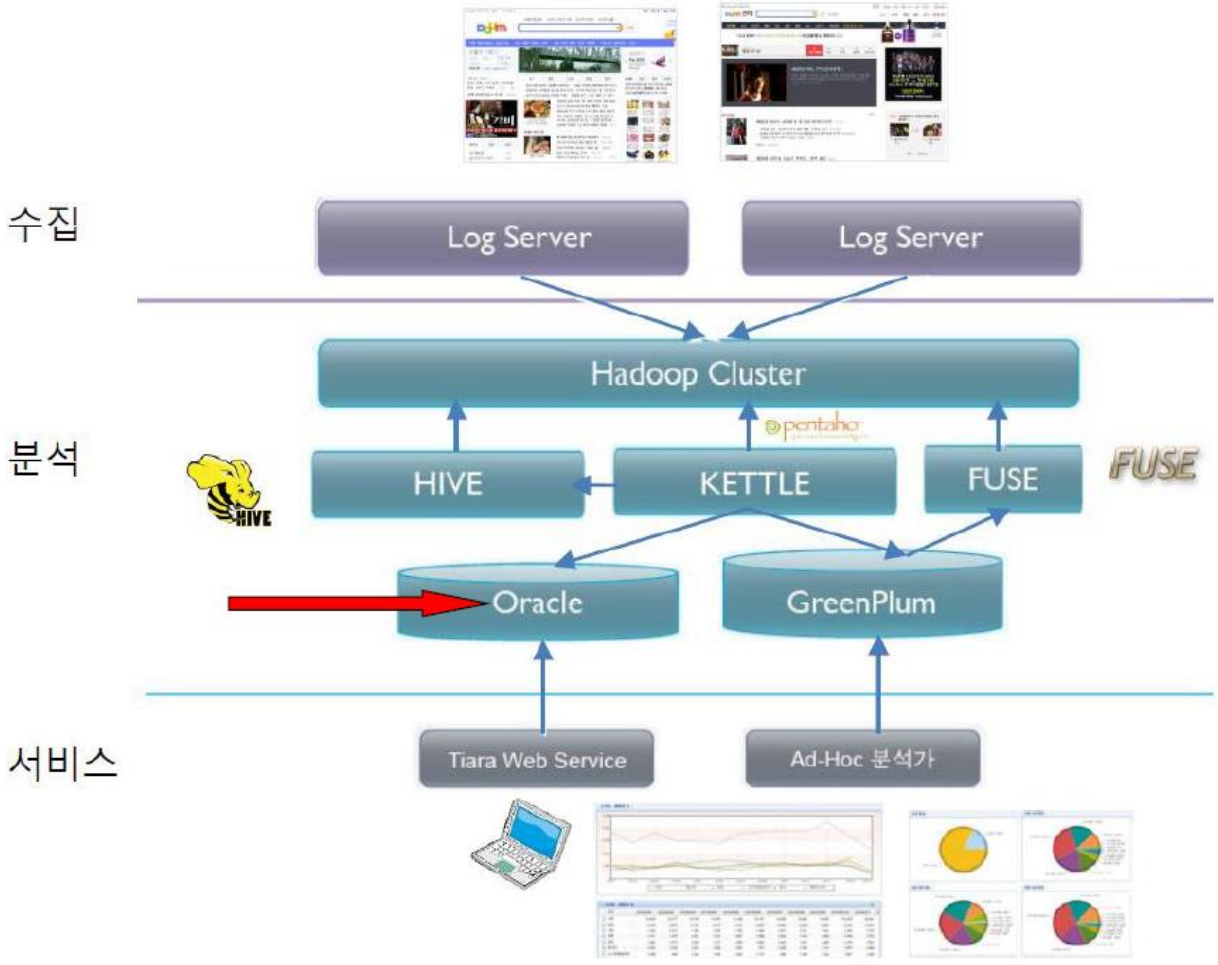


24시간 이후
분석 결과 제공

Daum 전사데이터로그분석 - 2010

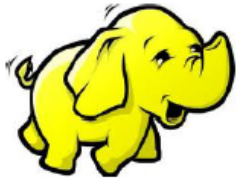


Daum 전사데이터로그분석 - 2010



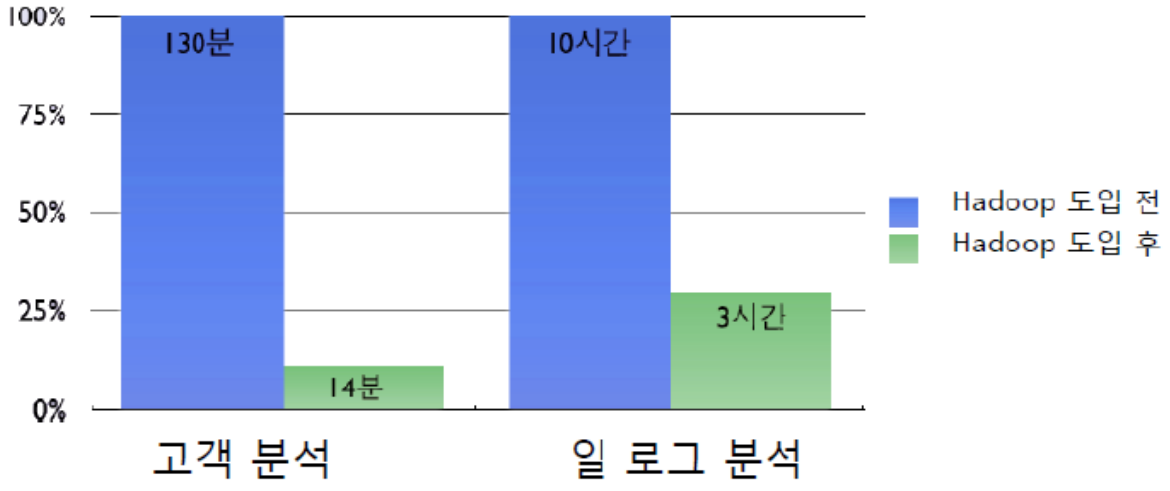
Daum 전사데이터로그분석 – Tiara 시스템

- 일 로그 사이즈 70TB → 전 처리 및 압축
- Daum 서비스 내 발생하는 모든 트래픽을 수집하여 분석 및 리포팅
  - 주요 분석 데이터: Pageview, Clickstream, User Analysis
- 데이터 처리 스택
  - Hadoop: 데이터 전처리
  - Hive (UDF, M/R): SQL 기반 데이터 분석
  - Pentaho Kettle (ETL): 데이터 저장
  - Greenplum: 병렬 데이터베이스
- 기존방식에 비해 데이터 처리 속도 향상 및 데이터 적재기간 증가



Daum 전사데이터로그분석 – Tiara 시스템

- 더 빠른 분석 (10분 단위 실시간 로그 확인 가능)

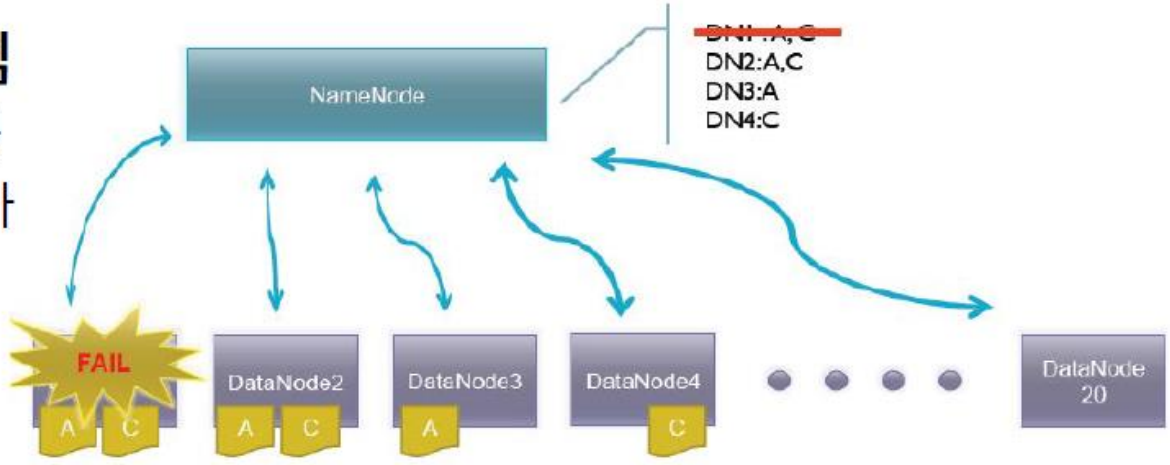


- 더 쉬운 분석 (Hive)


```
selelct serviceId, count(distinct uuid)
from web_log
where dt='20120101' group by serviceId
```

Daum 전사데이터로그분석 – Tiara 시스템

- **더 안정된 시스템**
  - 분산처리로 인한 작업 안정성 증가



- **고려사항**
  - 추가 증설 시 Hadoop 세팅 및 애플리케이션 배포 이슈
  - CPU/Memory intensive job을 해결하기 위한 클러스터 구성 및 관리
  - 네트웍 부하로 인한 10g 구성 비용 증가
  - 스케줄링에 따른 Job tracker를 통한 작업 분산의 어려움

- **Hadoop 기반 클라우드 컴퓨팅 스택의 확산 필요**

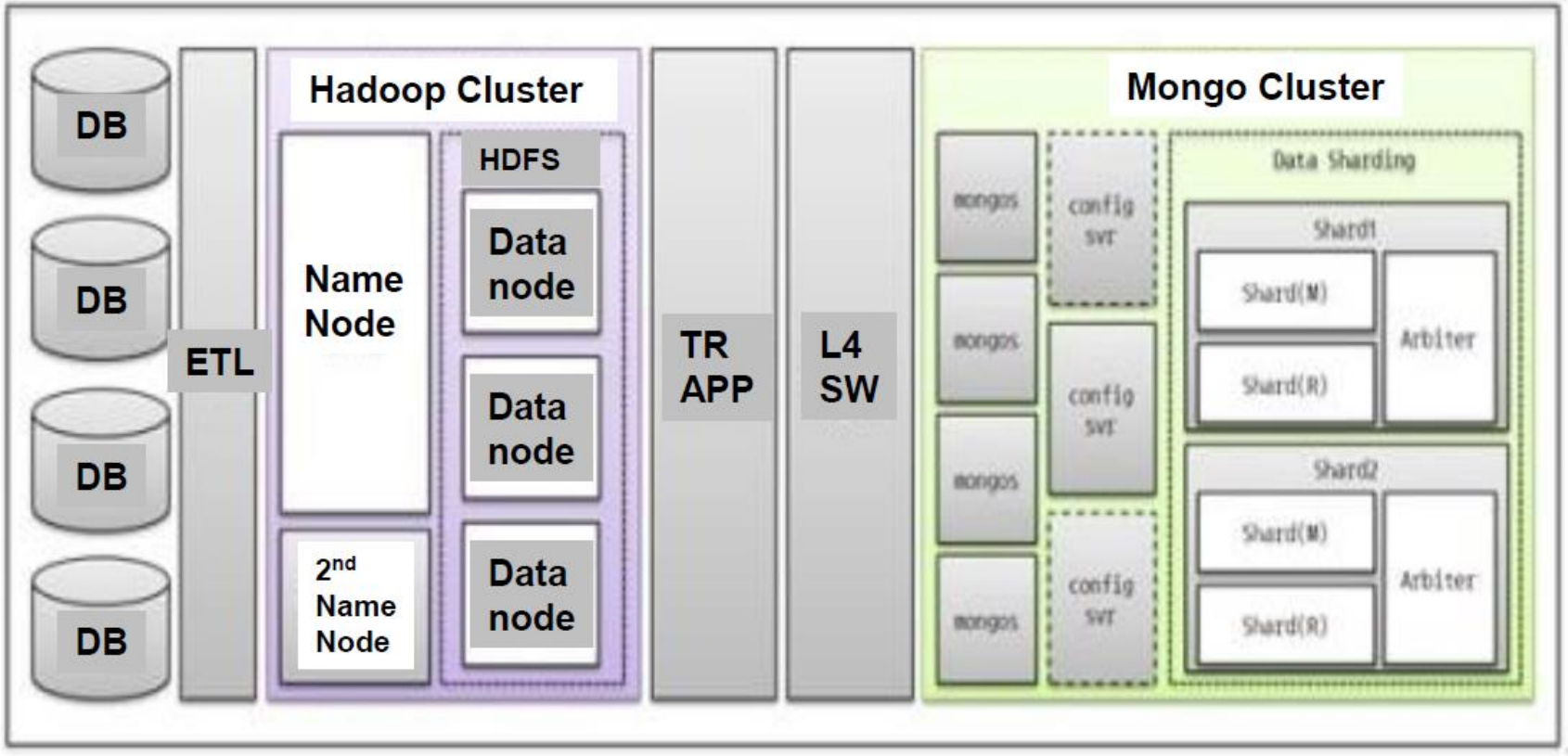
Daum 광고로그 분석시스템

- 광고 로그 및 통계 처리, 매체 토픽 분류 및 과거 로그 데이터를 기반으로 광고 집행 타켓팅 분석

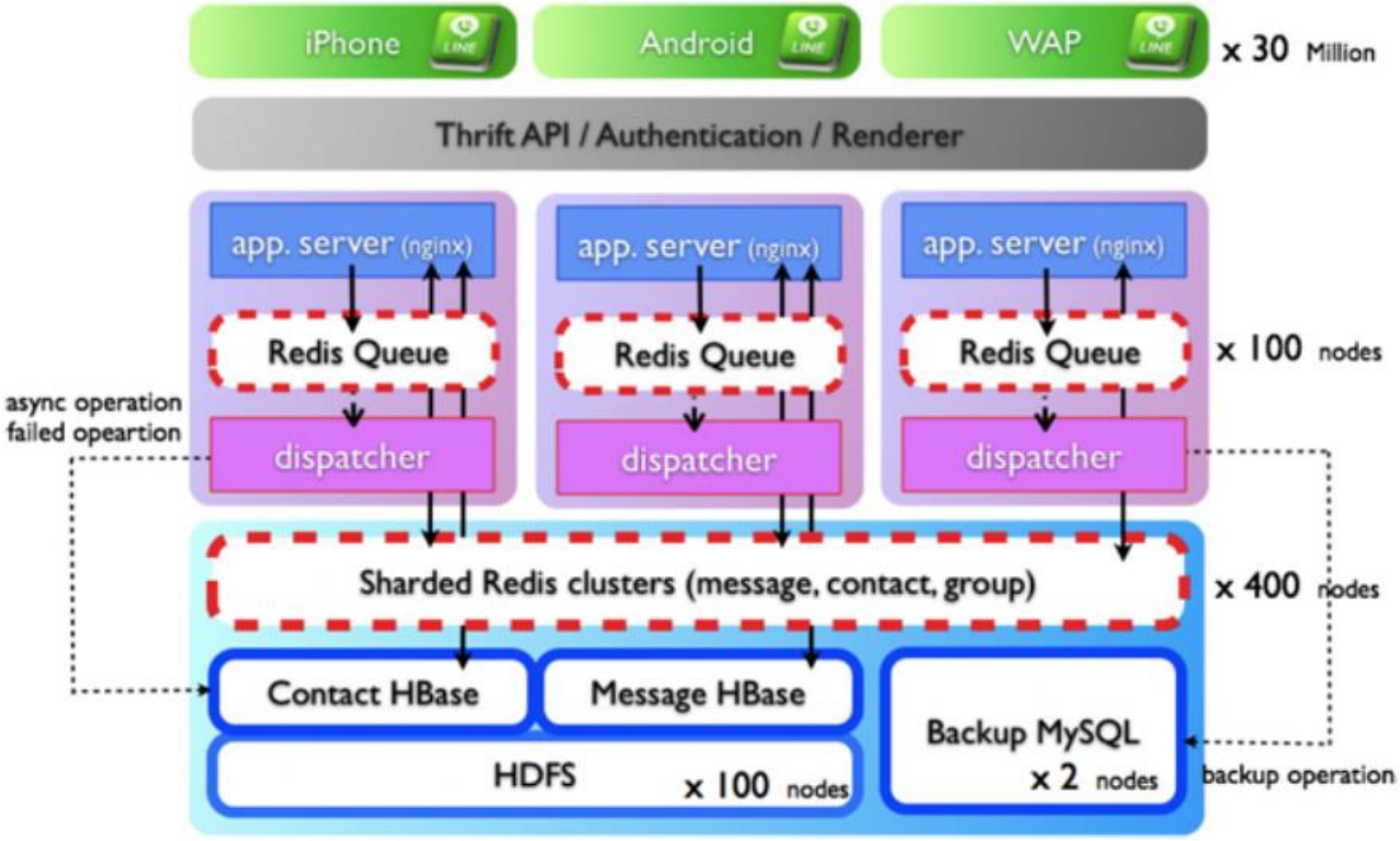
- 광고 데이터 분석용 Hadoop 클러스터 구성
  - 2.40GHz(듀얼 4코어)/ 메모리24GB: 서버 50여대 클러스터 구성
    - input: 과거 집행(노출, 클릭) 로그 데이터 (필요에 따라 일, 주, 월 단위 로그 사용)
    - output 광고에 대한 사용자별 노출 내역 통계 처리
  - 10분에서, 시간당, 일 단위로 다양한 데이터 산출

NHN 로그 분석 시스템

- Hadoop + MongoDB



NHN LINE Storage



Facebook 사용자 활동분석