줌인터넷(주)
빅데이터 활용사례

김우승

소개

- 줌인터넷(주) 연구소장
- 이력
  - 줌인터넷
  - SK 플래닛
  - SK 텔레콤
  - 삼성전자

- http://kimws.wordpress.com
- @kimws

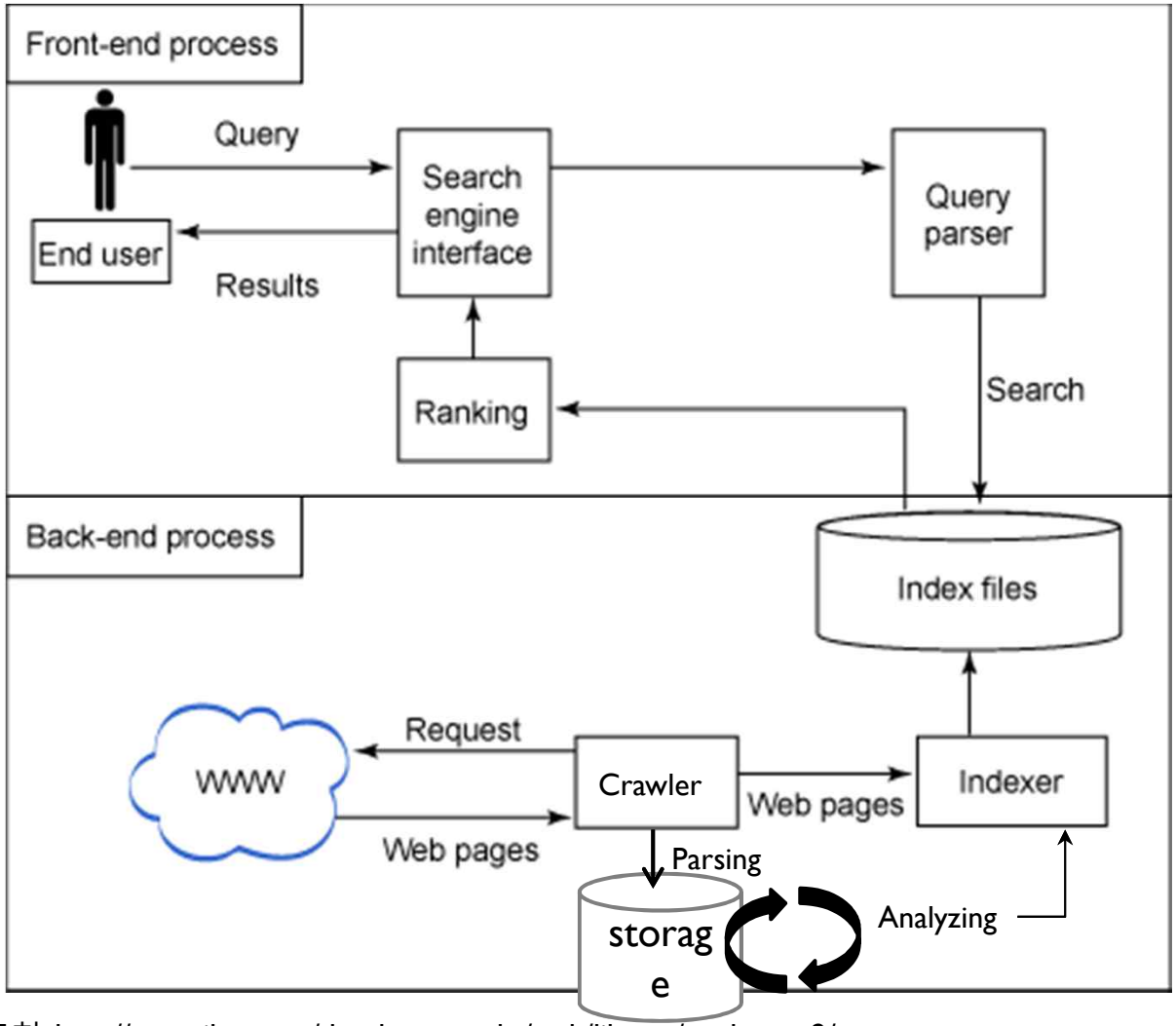
회사 소개 : zum.com

검색 포털 회사

더 편리한 인터넷을 위한 도전,
zum internet

zum internet

검색 회사(검색 서비스)가 하는 일 …



Serving

Real-time

Analyzing

Batch

Crawling

출처: http://www.ibm.com/developerworks/web/library/wa-lucene2/

Crawl

문서를 크롤하면 (> 10억건)

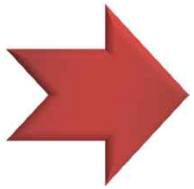
메타데이터 저장 → Structured Data

HTML 저장 → Semi-structured Data

제목, 본문 추출 → Unstructured Data

이미지 추출

- 썸네일 생성 → Multimedia Data
- 원본 이미지 저장



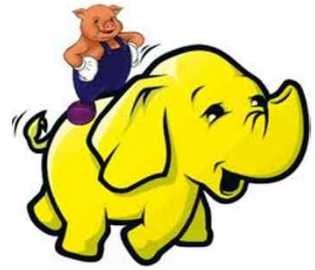
Crawler 시스템은 분산 처리를 하기 위해서 MapReduce + Downloader 등으로 구현

Analyzer

- 분석 데이터의 종류와 범위에 따라서 다양한 주기로 설정된 스케줄로 프로세스들이 실행
- 프로세스의 우선 순위에 따라 Pig & MR Job에 우선순위와 리소스를 상이하게 할당

- 중복 문서 제거
- 스팸 문서
- 성인 필터링
- 검색 랭킹 계산
- 문서 클러스터링
- 수십여가지 분석프로세스

(Pig + UDF) + Python + Shell



Java Map-Reduce program

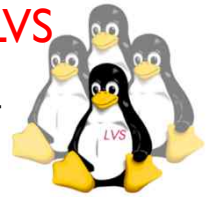
Hadoop & HBase ➔ Legacy System

Service

- Daily 수천만 PV 이상을 처리하기 위해서 …

LVS

Web & Application & DB Clustering 필수



Apache HTTP SERVER PROJECT

Tomcat

각 Layer 별로 Caching System 적용

Squid

Memcached

인덱싱 & 검색 처리

Apache Solr

수십/수백대로 구성된 서버팜

100% Linux Server

시스템 장애 감지와 통합 모니터링 시스템 역시 필수

오픈 소스 및 자체 개발 시스템을 통합해서 활용

장애 대응을 위한 HA 구성

Pacemaker

그 밖에 오픈소스들

But

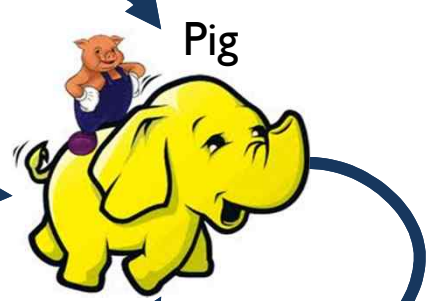
서비스와 데이터 요구사항에 맞게 코어가 되는
검색엔진, 미디어 서버, KeyValue 시스템등을 자체 개발

ZUM Data Platform

- 로그 수집 체계
  - 로그 포맷 표준화
  - 중앙 데이터 저장소 구축
  - 로그 데이터 수집 프레임워크 개발 (Flume-ng, ~~Fluentd~~)
    - Access log
    - ZUM service log
    - Application log
- 분석시스템 개발
  - **Hive 가 메인 도구**
  - **Pig 는 Hive Table 을 생성하는 전처리(ETL) 용 스크립트**
  - **Job Scheduler**

Log Analysis

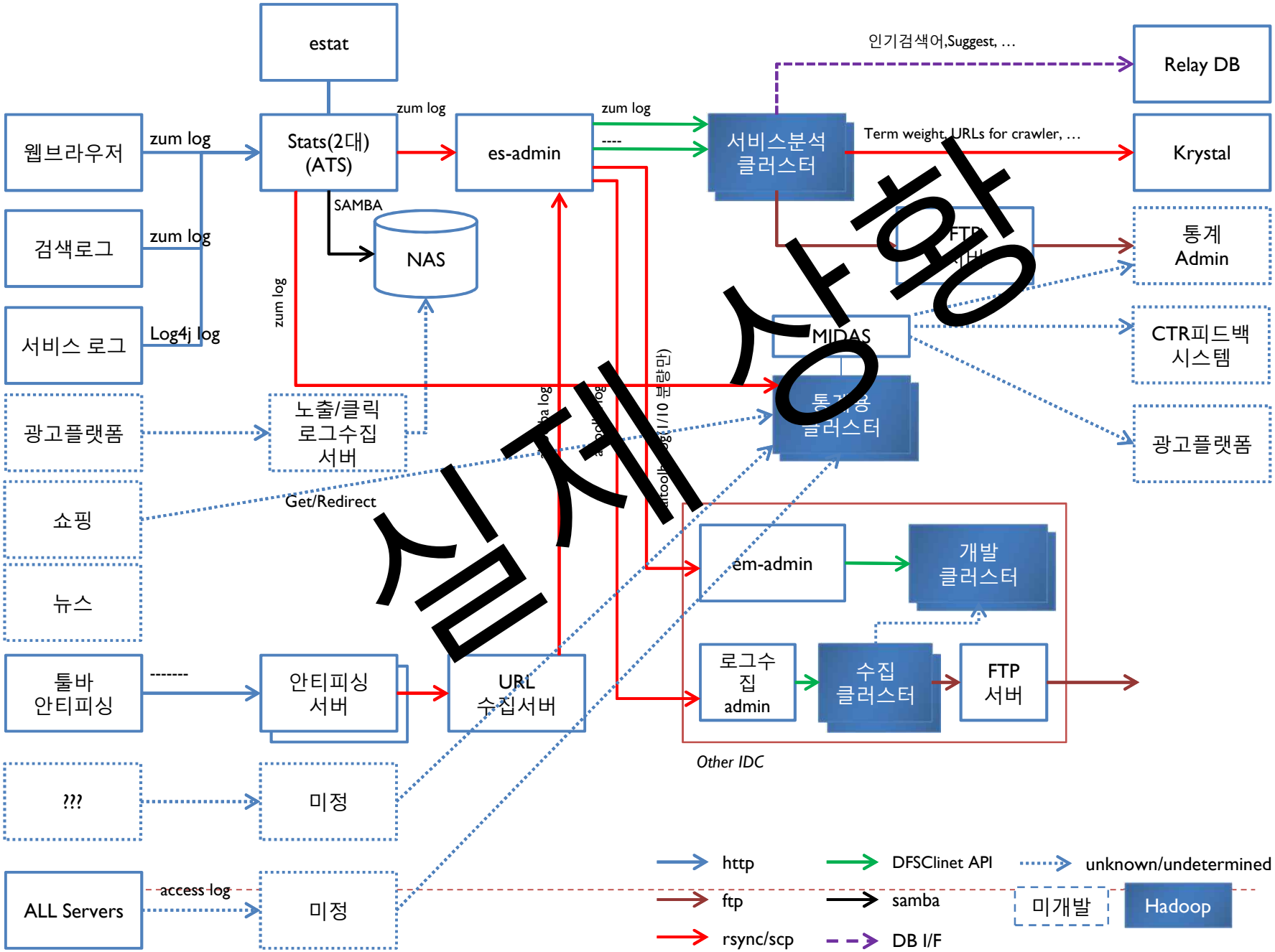
- 각종 Service Log → JSON
- Server Access Log → Apache Log Format
- Application Log → Log4J



개발자들이 Apache Pig Script 로 분석

기획자들이 직접 데이터 통계나 분석을 할려면?

Log Data Flow Diagram



통계 지표 관리

- 고정 지표
  - 상시 분석해서 결과를 파악해야 하는 지표
  - 공통 지표 정의
  - 서비스별 지표 정의
  - 배치성 지표
  - 실시간성 지표

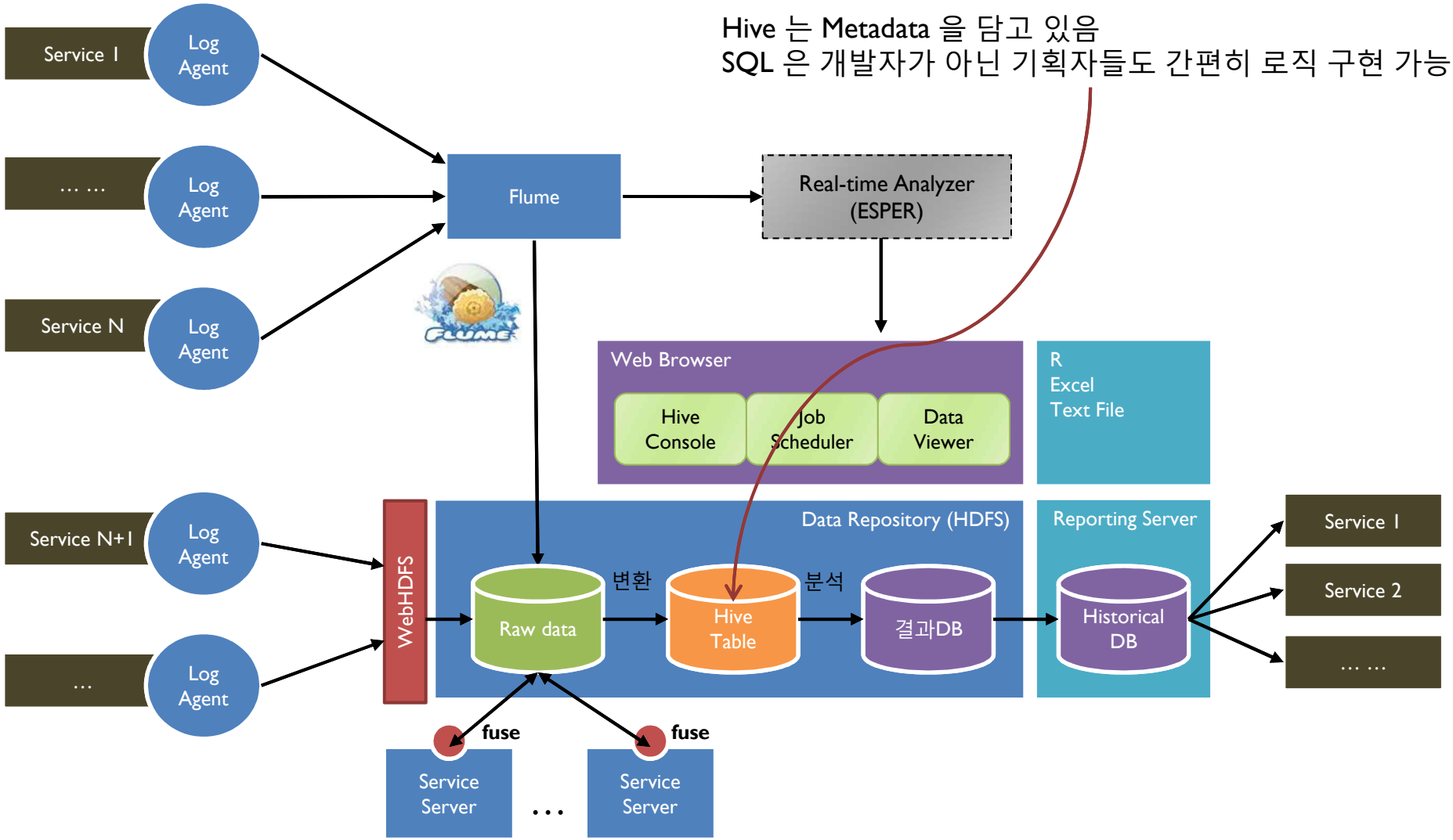
→ 자동화의 대상

- 유동 지표
  - 서비스별로 필요한 경우에 따라서 파악할 지표
  - 중요도에 따라 고정지표로 전환

→ Ad-hoc 업무

데이터 분석 업무가 기존 개발자에서
기획자, 데이터 분석가의 손에서 다루어질 수 있도록

ZUM Data Platform



요즘 드는 생각 …

Apaceh Hadoop / HBase 가 이미 Legacy System 되어간다면 …

결국 점점 복잡해지고, 무거워지고, 이해하기 힘들어지고 …

다른 이가 만든 기술에 대한 의존은 여전 & 한계!

Don't reinvent the wheel?

Core 에 대한 깊은 이해 없이

대충 이해하고 응용 어플리케이션 / 서비스 만들기도 바쁘다

그래도?

직접 만들어 볼 필요가 있다!!!

Q&A

감사합니다