

검색로그 시스템 with Python

Pycon 2016

카카오 검색로그셀 김동문

소개

Daum + Kakao 에서 발생하는
검색에 대한 로그

취합 / 정제 / (배포) 하는 시스템
구현

대상

여러 대의 서버에서 로그를 취합하여,
하둡 or NoSQL에 Load 하시려는 분

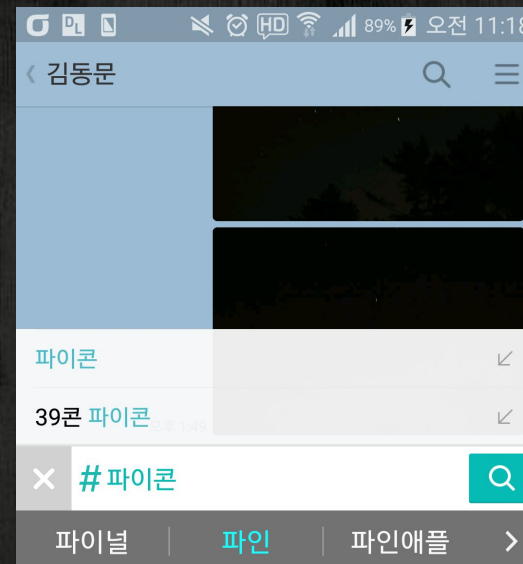
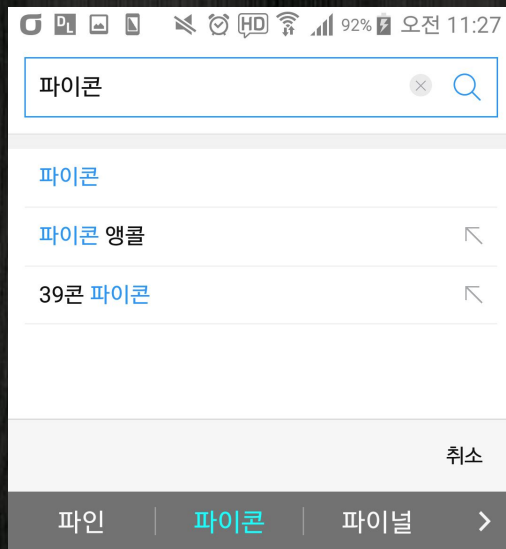
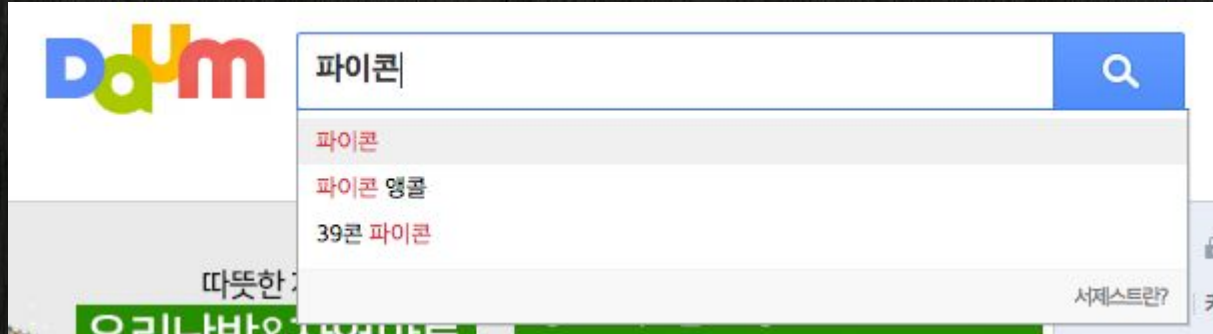
목차

1. Overview
2. Problem
3. System Flow
4. System Implementation
5. Why Python
6. 그 외 사례

1. Overview

Overview

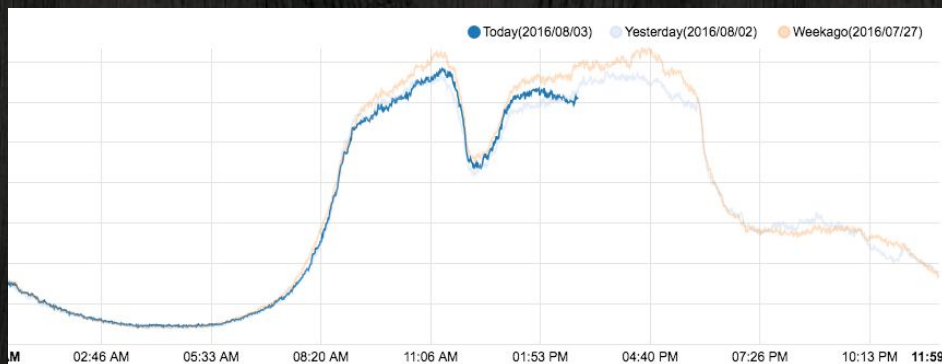
From:



Overview

To :

실시간 이슈 검색어		08.01 11:33
1	하연수 댓글	↑ 39
2	김연지	↑ 201
3	최필강	NEW
4	해운대 교통사고	↑ 77
5	이지은	↑ 55
6	달의 연인	↑ 49
7	김현성	↑ 49
8	김민희 화장품 광고	NEW
9	김태원	↑ 40
10	서태지	↑ 42



관련 검색어

수륙양용차 파이썬

파이썬 이란

파이썬 예제

파이썬 책 추천

파이썬 프로그래밍

파이썬 설치

파이썬 게임

파이썬 용도

파이썬 강좌

파이썬 책

파이썬 맵

스타 맵 파이썬

우분투 파이썬

파이썬 프레임워크

파이썬 개발툴

Overview

“

검색 이력을 중개해주는 시스템

2. Problem

문제점이 될 게 있나요?

Problem

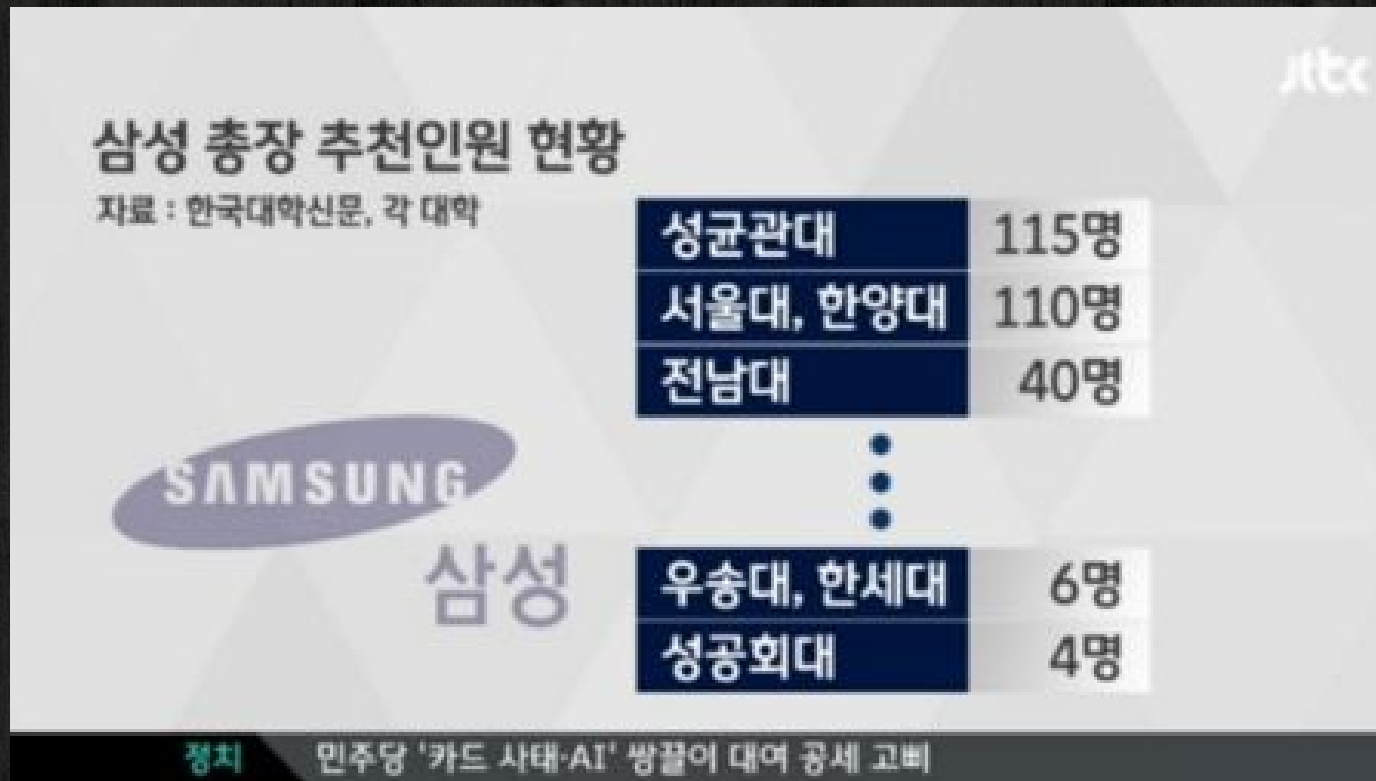
1. 작업은 모두 1분 내

- 밀리기 시작하면, 다음 프로세스에

영향

Problem

이 날이었어요 (2014. 1. 27.)



Problem

이렇게 되었어요



Problem

2. 이슈로 인한 트래픽 증가

- 단기 폭증
 - 천재지변
 - TV 프로그램
- 꾸준히 증가폭 유지
 - 연예
 - 정치
 - 사회

Problem




Problem

뉴스 **연예** 스포츠 자동차 라이프 TV

LIVE 강정호 2루타

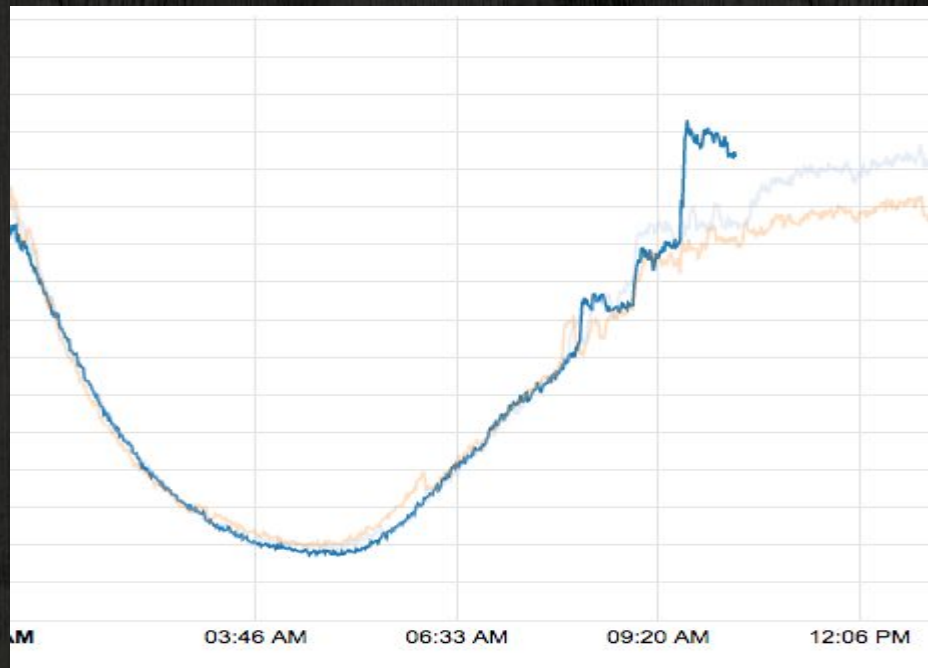
실시간 이슈 ▲ 코스닥 702.26 ▼



'불참' 김국진♥강수지
달달했던 모습

김국진♥강수지, 진짜 '연인' 됐다 "열애 인정"
김국진 "좋은 감정 갖고있다..양가 인사는 아직"
'진짜사나이' PD "이시영, 체력왕..男 다 이기더라"
하석진·김지석·이기우.. 'tvN 공무원 배우'들 속사정
'W' 6분같은 60분, 고구마+클리셰 따윈 취급 안해

1	김국진 강수지	▲ 300
2	김희정	▲ 96
3	불타는 청춘	▲ 67
4	솔비	▲ 39
5	정윤희	▲ 34



Problem

3. 어뷰저 필터링

- 장 / 단기의 데이터를 분석
- 반복적인 행위를 보이는 유저를 제거

Problem

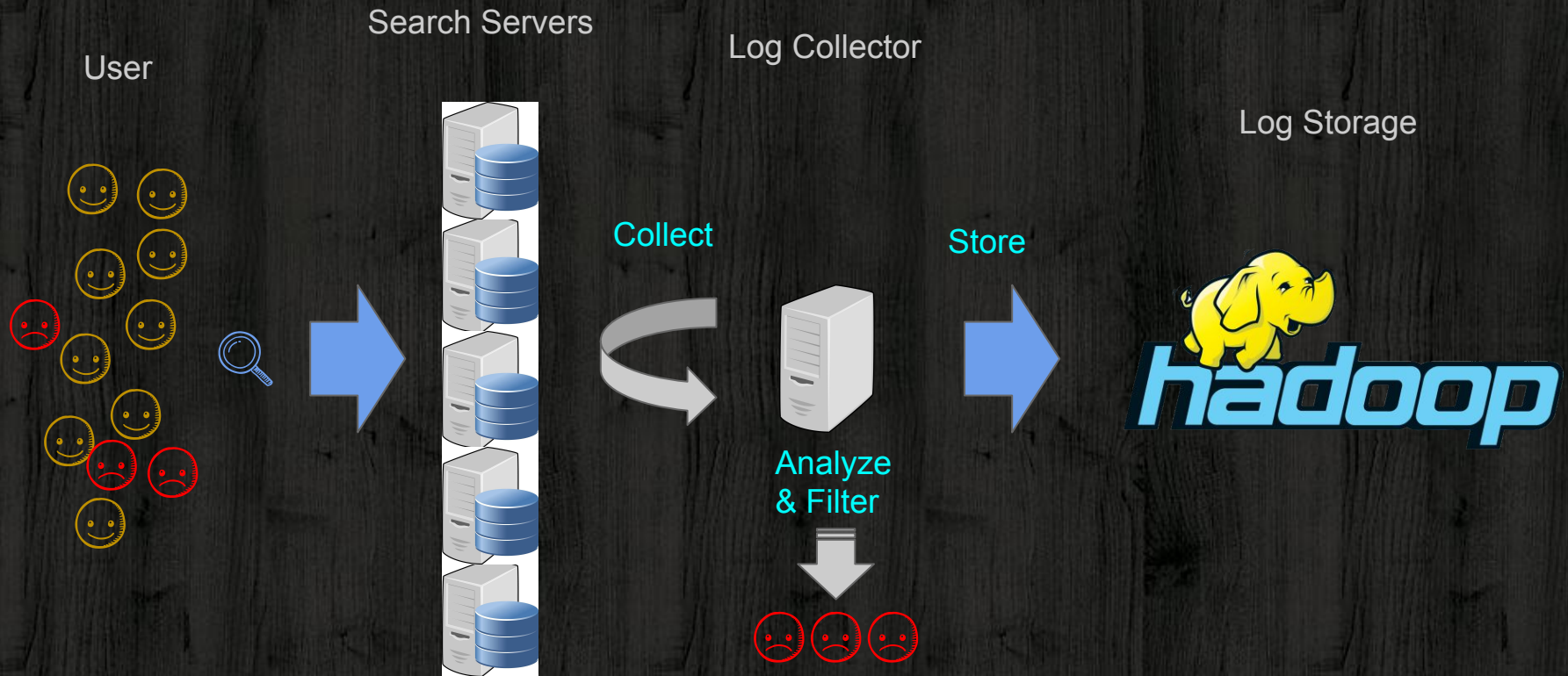
“

Time Attack : 1분
때때로 예상치 못한 트래픽
어뷰저 거르기

3. System Flow

이렇게 할꺼예요

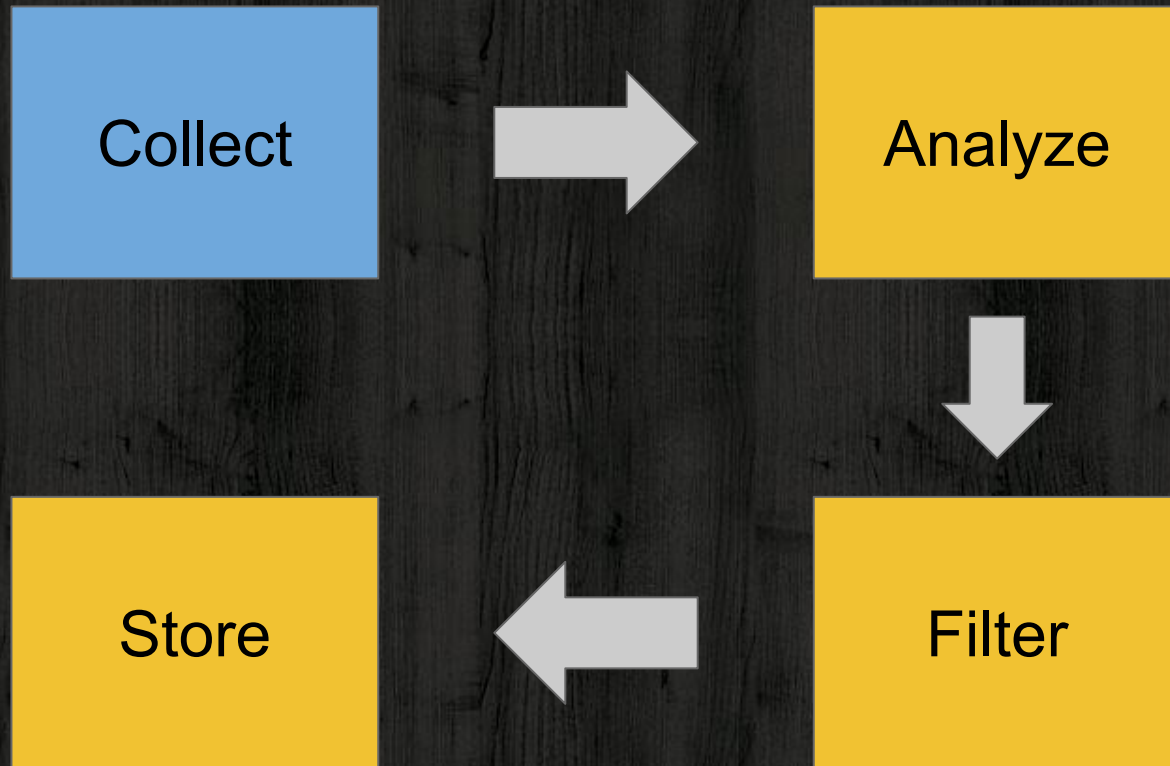
System Flow



4. System Implementation

*Step*별로 구현해볼까요

System Implementation



Collect

- 할 일

- 수십대의 서버에서
- 수메가 바이트의 로그를
- scp 를 통해 PULL

! scp 채택

- 제일 빠르고 안전.
- 솔루션의 결함을 의심할 필요 없음
- 장애 발생시 복원 작업이 수월

Collect

- 가정

- 서버 : 30대
- File Size : 300 MByte
- File Row : 10만
- 수집 서버 CPU Core : 24개

Collect

- Code

- pull 은 생략

```
SERVERS = ["search-server-dn%d" % index for index in xrange(1, 31)]  
[pull(server) for server in SERVERS]
```


Collect

- 수행 시간
 - 7초

“

조금 느린 것 같아요

Collect

- 속도 개선

- 하드웨어

- 세상에서 제일 싼 건 서버 비용
 - 세상에서 제일 비싼 건 당신 연봉

- 소프트웨어

- Pycon이니까 이 방법으로 해결해야 함

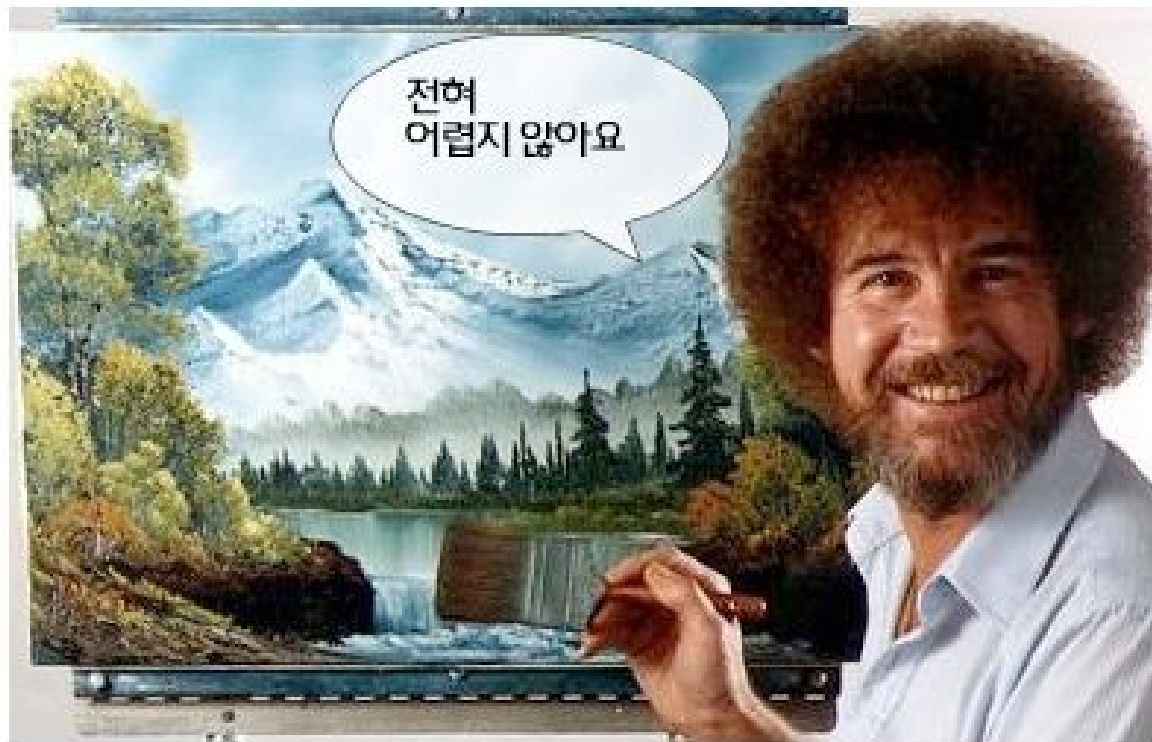
Collect

- 속도 개선
 - multiprocessing 도입

```
[pull(server) for server in SERVERS]
```



```
import multiprocessing
process_list = [multiprocessing.Process(target=pull, args=(server, )) for
server in SERVERS]
[process.start() for process in process_list]
[process.join() for process in process_list]
```

어때요 ? 참 쉽죠 ?

Collect

- Java로 했다면..

```
public static void main(String[]args){
    ArrayList<Thread>threads=new ArrayList<Thread>();
    for(int i=0;i<30;i++){
        Thread t=new Thread(new Pull(i));
        t.start();
        threads.add(t);
    }

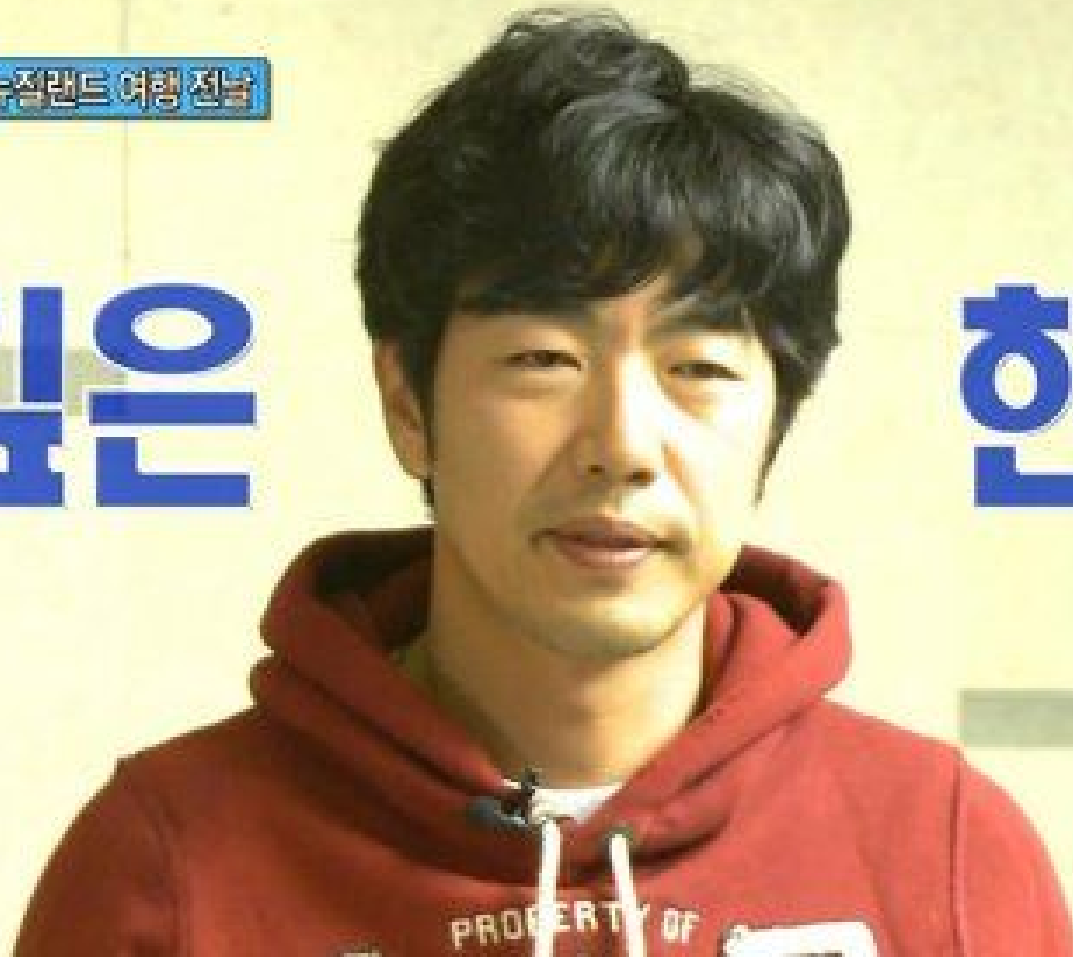
    for(int i=0;i<threads.size();i++){
        Thread t=threads.get(i);
        try{
            t.join();
        }catch(Exception e){
        }
    }
}
```



MBC
일일판 대중 치매 상담
060-700-1122

깊은

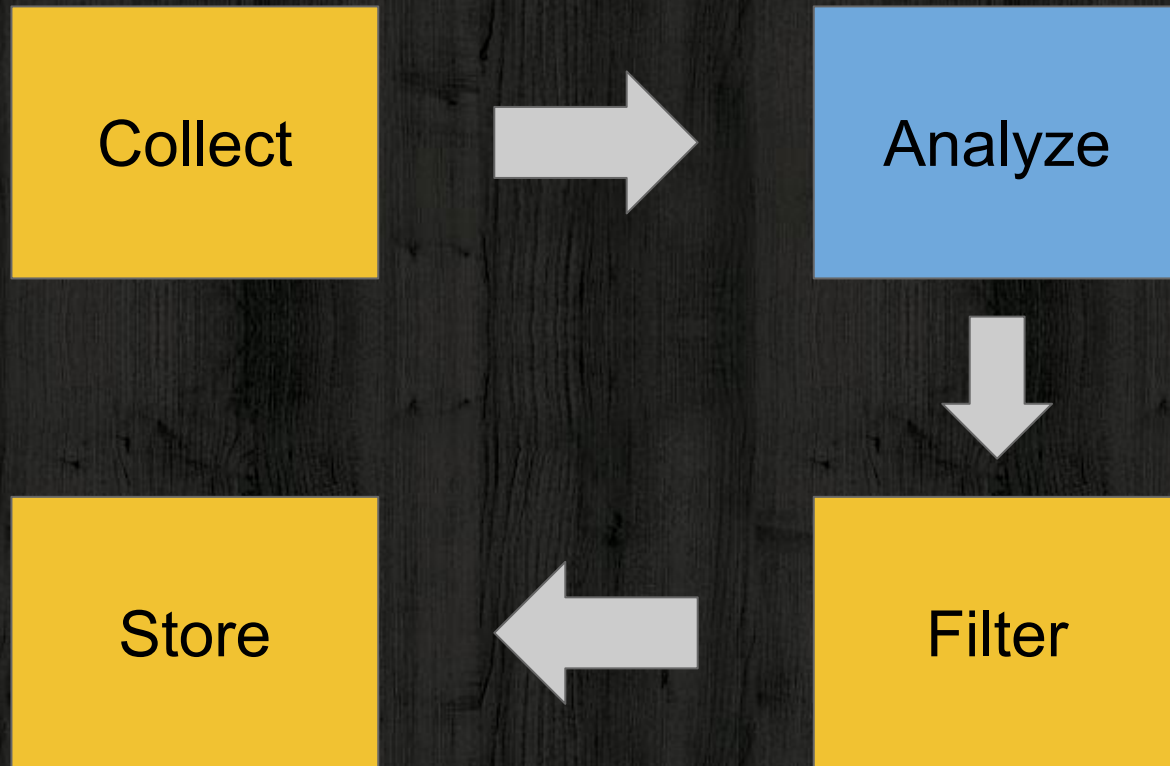
한숨



Collect

- 수행 시간
 - 7초 -> 1.6초
- 남은 시간
 - 58.4초

System Implemetation



Analyze

- 할 일
 - 어뷰저 탐색
 - 상위에 Rating 되는 검색어를
극소수의 사용자가 검색
 - Bot에 의한 Crawling

Analyze

- 자세한 건



Analyze

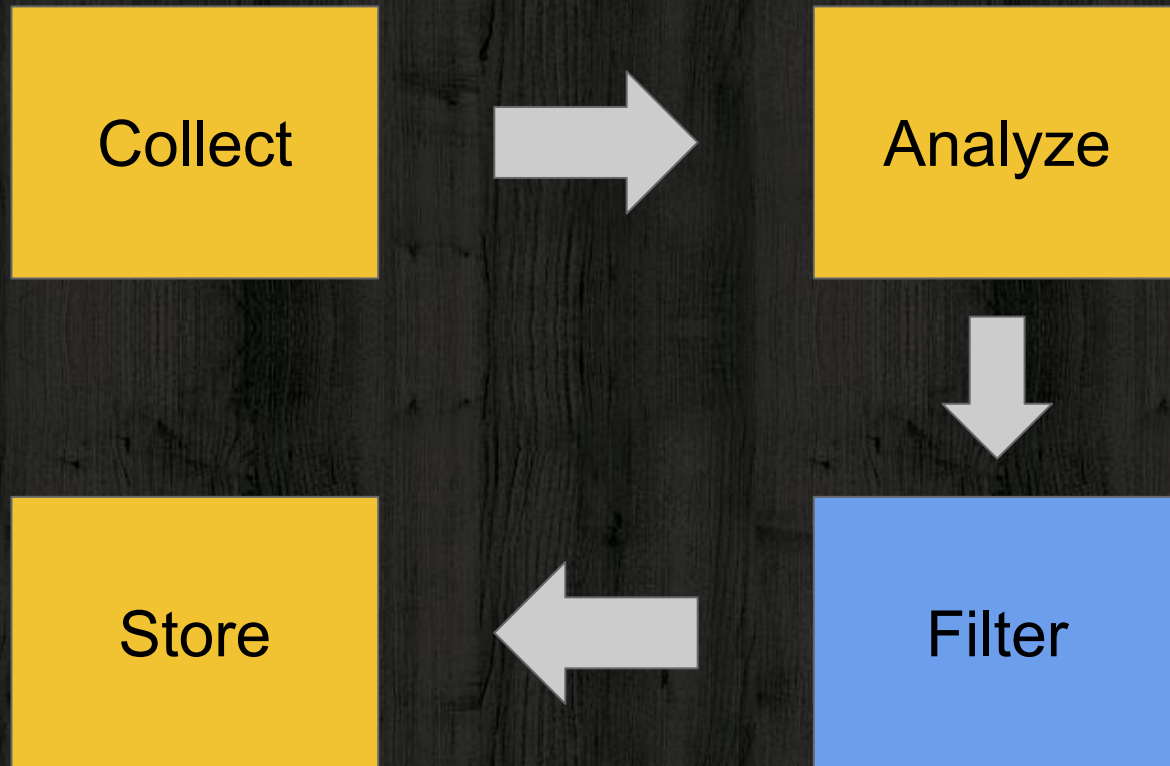
- 수행 시간

- 10초

- 남은 시간

- 48.4초

System Implemetation



Filter

- 할 일

- 앞서 조사한 어뷰저를 제거
- 정제된 로그는 압축 (gzip)
 - Network Traffic 최소화

Filter

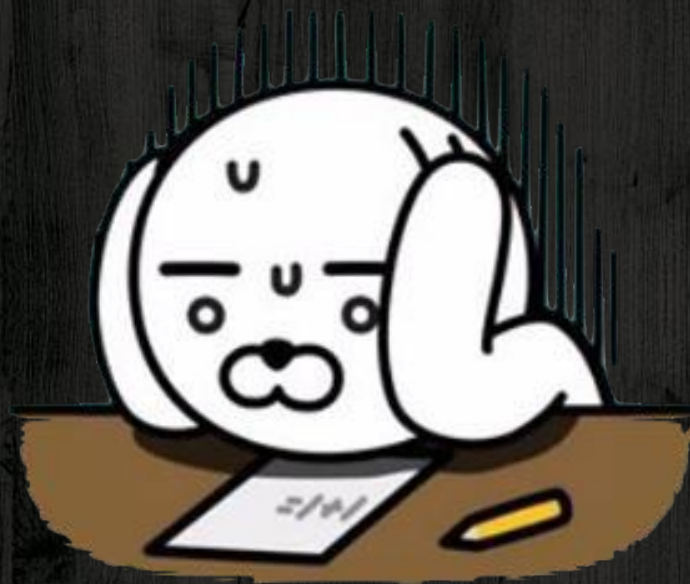
- Code

```
def purify(raw_file_name):  
    log_file = gzip.open(raw_file_name + ".gz", "w")  
    abuse_cnt = 0  
    success_cnt = 0  
  
    with open(raw_file_name, "r") as raw:  
        for line in raw.read().splitlines():  
            if line in abuse:  
                abuse_cnt += 1  
            else:  
                success_cnt += 1  
                log_file.write(line)  
  
    print "Success : %d, Abuse : %d" % (success_cnt, abuse_cnt)
```

Filter

- 수행 시간
 - 37초

Filter



Filter

- 속도 개선
 - multiprocessing 적용
 - 서버별로 받아서, 나뉘어져 있던 파일에 각각 Filtering 적용
 - gzip의 merge하기 쉬운 장점을 이용
 - 1 process -> 30 process

Filter

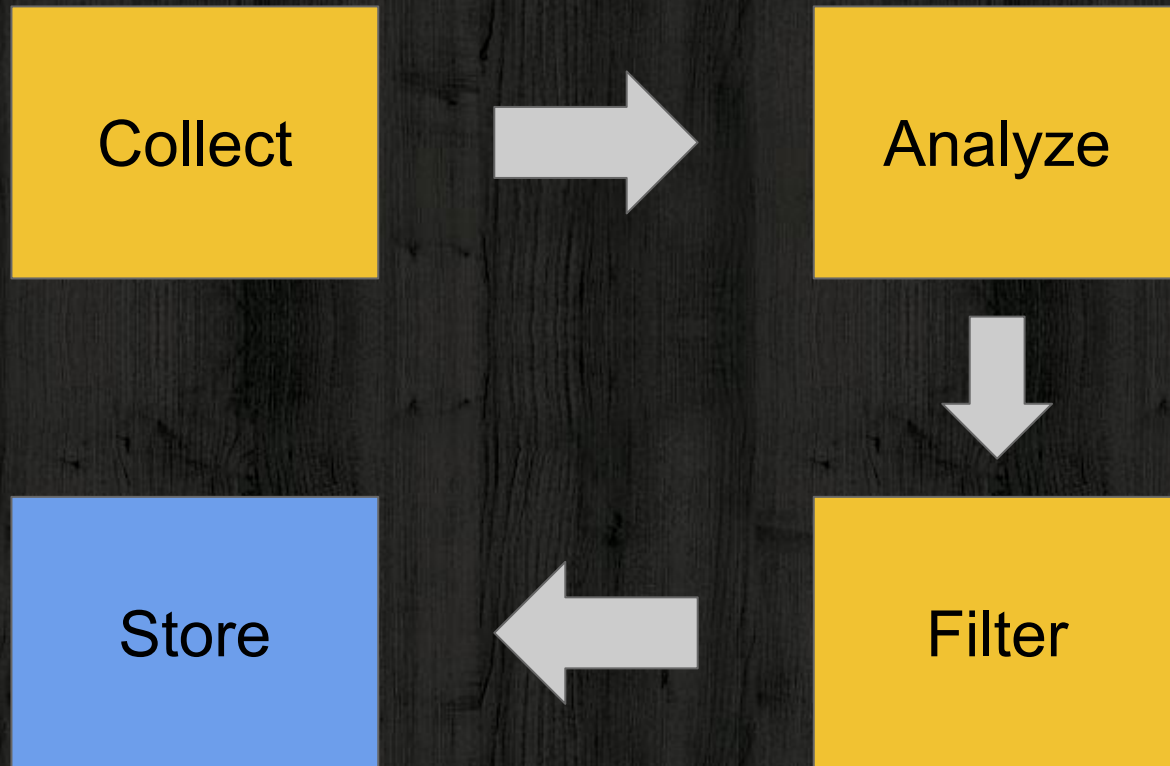
- Code

```
process_list =  
[multiprocessing.Process(target=purify,  
args=("/pycon/incoming/min_log.%s" % server, ))  
for server in SERVERS]  
[process.start() for process in process_list]  
[process.join() for process in process_list]
```

Filter

- 수행 시간
 - 37초 -> 2초
- 남은 시간
 - 46.4초

System Implemetation



Analyze

- 할 일

- 압축한 파일을 모아서 Hadoop에 Upload

Analyze

- 수행 시간

- 6초

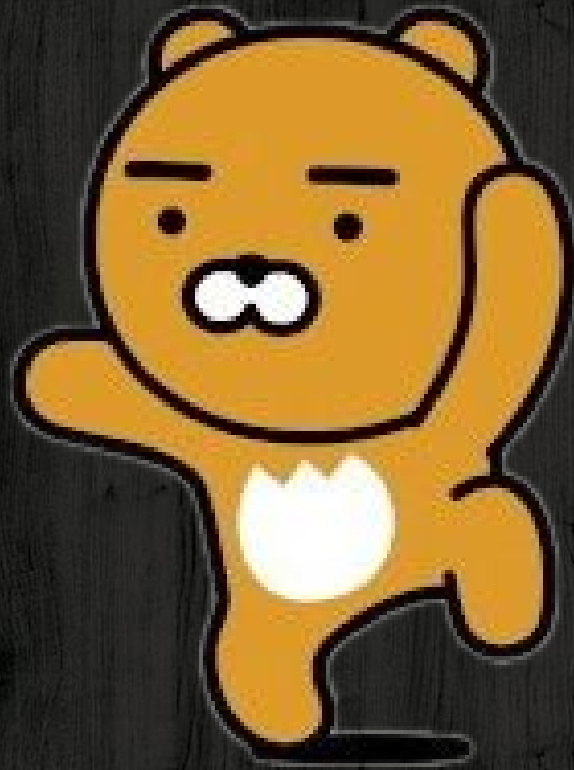
- 남은 시간

- 40.4초

“

serial : 60 s

multiprocessing : 19.6 s



20초만에 끝



부하 테스트

트래픽이 급증하는 경우를 대비하여
부하량을 올려보겠습니다.

부하 테스트

소요시간

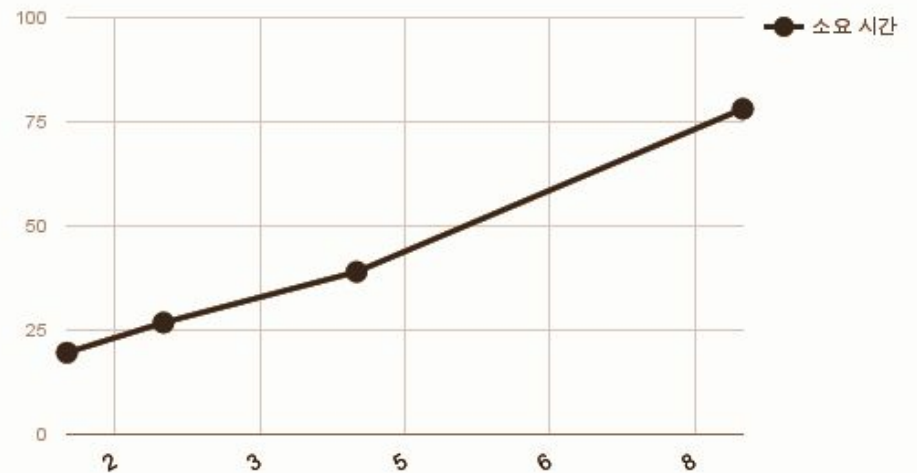
1 19.6

2 26.8

4 39

8 78.1

부하테스트



60초가 넘으면

- 로그 정제

- 사용하지 않는 데이터 제거
- 사용처에 따라 로그를 분할

- 서버 Upgrade

- SSD 설치해달라고 조르기
- 담당할 서버를 나눠서 병렬 처리

5. Why Python

Why Python

- 빠른 구현
- 풍부한 라이브러리
- 시스템 커스터마이징
 - 자동 재시도 기능
 - 장애 알림 기능
 - 장애 복구 기능

Impala
1.0.1 (Prod) 출시 : 2013. 6

Impyla
0.7 출시 : 2013. 5

Why Python

- vs LogStash, Fluentd
 - 단점
 - Streaming 안함
 - Visualization 안함
 - PetaByte 처리 못함

Why Python

- vs LogStash, Fluentd
 - 장점
 - 안정성
 - Tool의 bug 걱정 없음
 - Pull 방식의 중앙 집중화된 시스템
 - 기능 확장
 - RealTime 어뷰징 분석 / 적용
 - 각종 Customizing 적용 용이
 - 풍부한 Python의 Library를 사용 가능

Why Python

- 장애 감소
 - 이전 Version
 - Java로 구성
 - 월 1-2회 장애 발생 (Traffic)
 - 이후 Version
 - Python 으로 구성
 - 연 2-4회 장애 발생 (개발자 실수)

“

로그 시스템 소개를 마치고,
다른 사례도 소개해 드리려고 합니다.

6. 그 외 사례

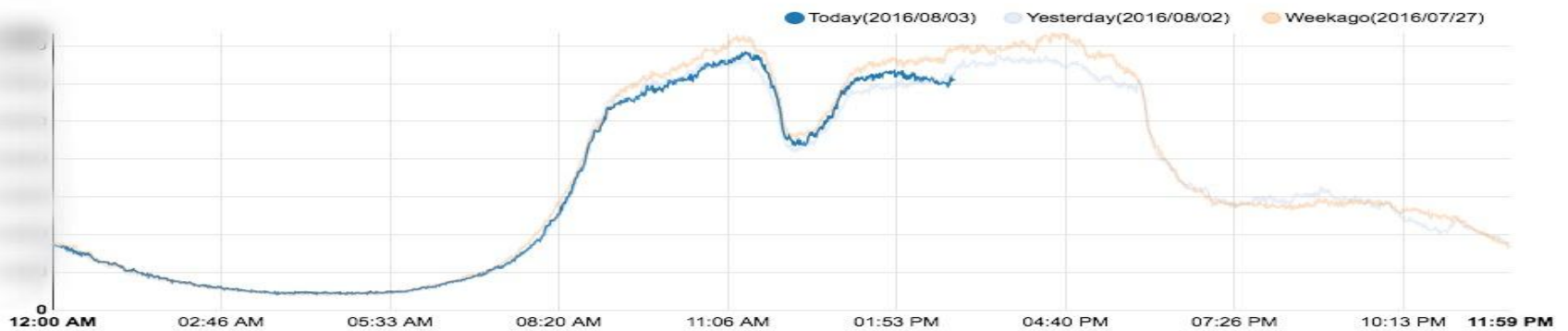
팀에서 *Python*을 이용하여
개발한 다른 사례도 보여 드릴게요
(+ 개발 기간)

실시간 유입 현황


Mobile Query Graph

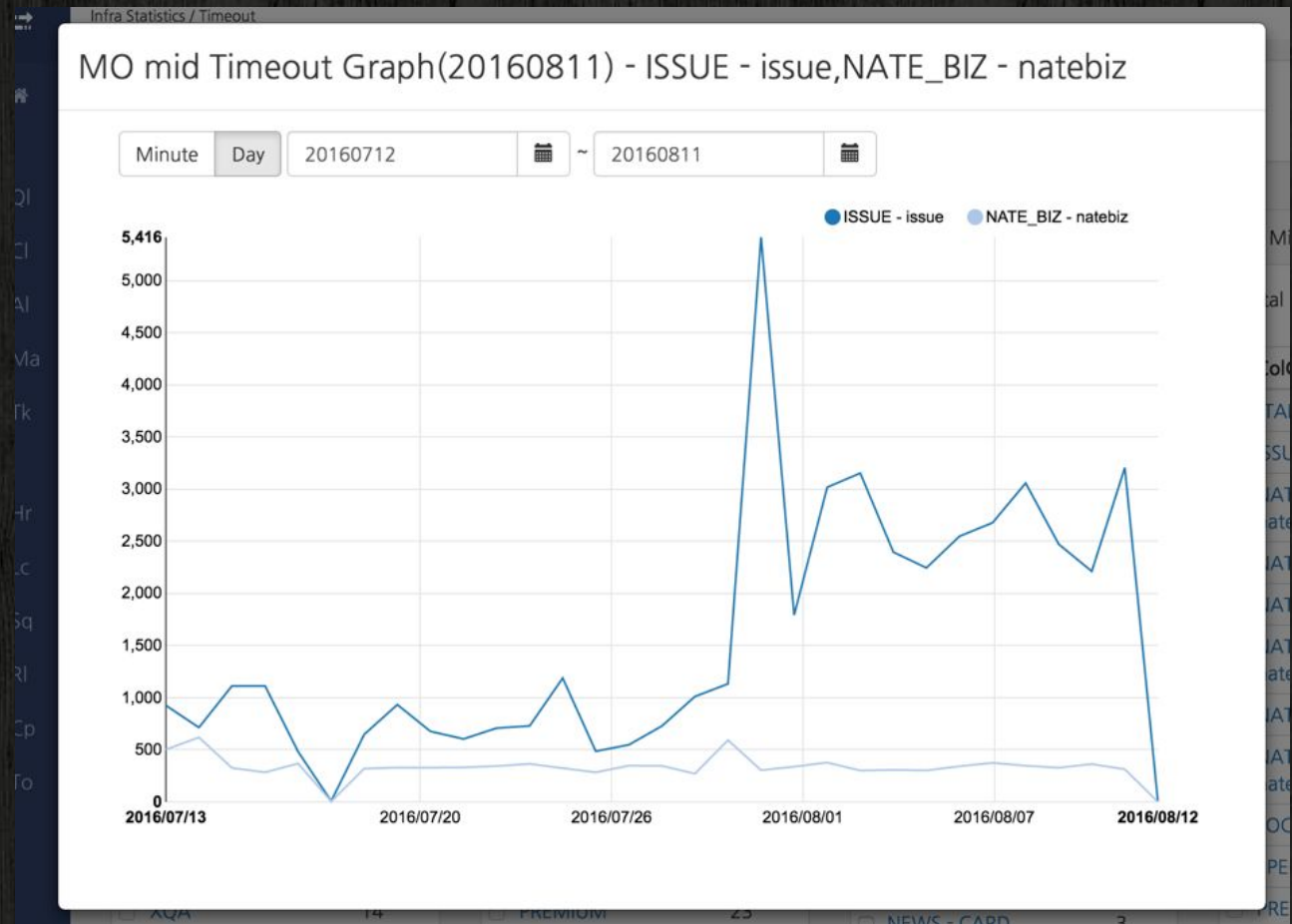


PC Query Graph



서버 로그 분석 / 타임 아웃 현황 조사

Mobile Middle	
Total : 4,601	
	
ColGroup	Count
<input type="checkbox"/> ISSUE - issue	3,206
<input type="checkbox"/> NATE_BIZ - natebiz	313
<input type="checkbox"/> NATE_BRAND - natebiz	313
<input type="checkbox"/> NATE_BRAND - natebrand	301
<input type="checkbox"/> NATE_BIZ - natebrand	301
<input type="checkbox"/> LOCAL - localbiz	72
<input type="checkbox"/> FUSION - web	27
<input type="checkbox"/> SHOPPING - shop	27
<input type="checkbox"/> TWITTER - swBoard	10
<input type="checkbox"/> PREMIUM_GRP - premium	6




“

+ 네트워크 현황 / 캐시율 / 블러킹

실시간 분석 툴 개발

2명 / 8주

시스템 모니터링

 Watchtower

[David++]
[2016/07/31 18:35] 모바일 쿼리가 급증하였습니다.
변동폭 : -> (23.05 %)

Rank	Query	Count	CTR
1 -	북면가왕 휘발유		78.28
2 -	씨야 김연지		67.70
3 ▲3	북면가왕		82.70
4 ▼1	한동근		202.99
5 new	불광동 휘발유 정체		105.89
6 ▲29	김연지		104.81
7 -	김천 물놀이 사고		94.25
8 ▲2	네이버		107.39
9 -	하연수 인성논란		111.17
10 ▲1	차예련		102.81
11 ▲5	불광동 휘발유		83.57
12 ▲1	로또 당첨 번호		86.55
13 ▼8	로이킴		107.57
14 -	도겸		65.21
15 -	날씨		273.43
16 new	북면가왕 휘발유 정체		89.43
17 ▲2	슈퍼맨이 돌아왔다		111.58
18 new	세븐틴		103.29
19 ▲6	하연수		120.62
20 ▼8	로맨틱 흑기사		85.96

순위 변동폭은 집계 시점보다 5분전 기준입니다

“

검색 유입 급증 / 급감 *Alarm Plugin*
1명 / 0.5 일

검색 트렌드 분석

2016/30주차

1페이지

[다음페이지](#)

keyword	discode	category	최근 1년간 포함회수	4주간 쿼리볼륨	4주간 쿼리변화율 ↓
사이판지진	tot	기상/날씨 - 기상정보 - 해외기상		30,062	125308 %
서남표	tot	인물 - 인물 - 인명		7,388	15500 %
김성주종교	tot	- -		268	6700 %
서장원포천시장	tot	인물 - 인명+기타 - 인명+정보성		35,227	6314 %
대전트램노선도	tot	뉴스/잡지 - 뉴스 - 종합		855	5408 %
접도	tot	여행/국가 - 자연명소 - 섬		5,615	4750 %
승건	tot	인물 - 인물 - 인명		365	4612 %
유인나키	tot	인물 - 인명+기타 - 인명+정보성		866	4366 %
김남희아나운서	tot	인물 - 인명+기타 - 인명+정보성	1	49,851	3996 %
순대국수	tot	- -		363	3835 %
천안함합장최원일	tot	- -		144	3600 %
KIASK	tot	- -	1	4,536	3442 %
불광동휘발유	tot	방송/엔터 - 프로그램 - 프로그램 출연진/캐릭터	1	9,465	3373 %
서울포항	tot	스포츠/레저/자동차 - 스포츠경기 - 축구		1,749	3165 %
경찰청장내정자	tot	뉴스/잡지 - 뉴스 - 종합		1,045	3026 %

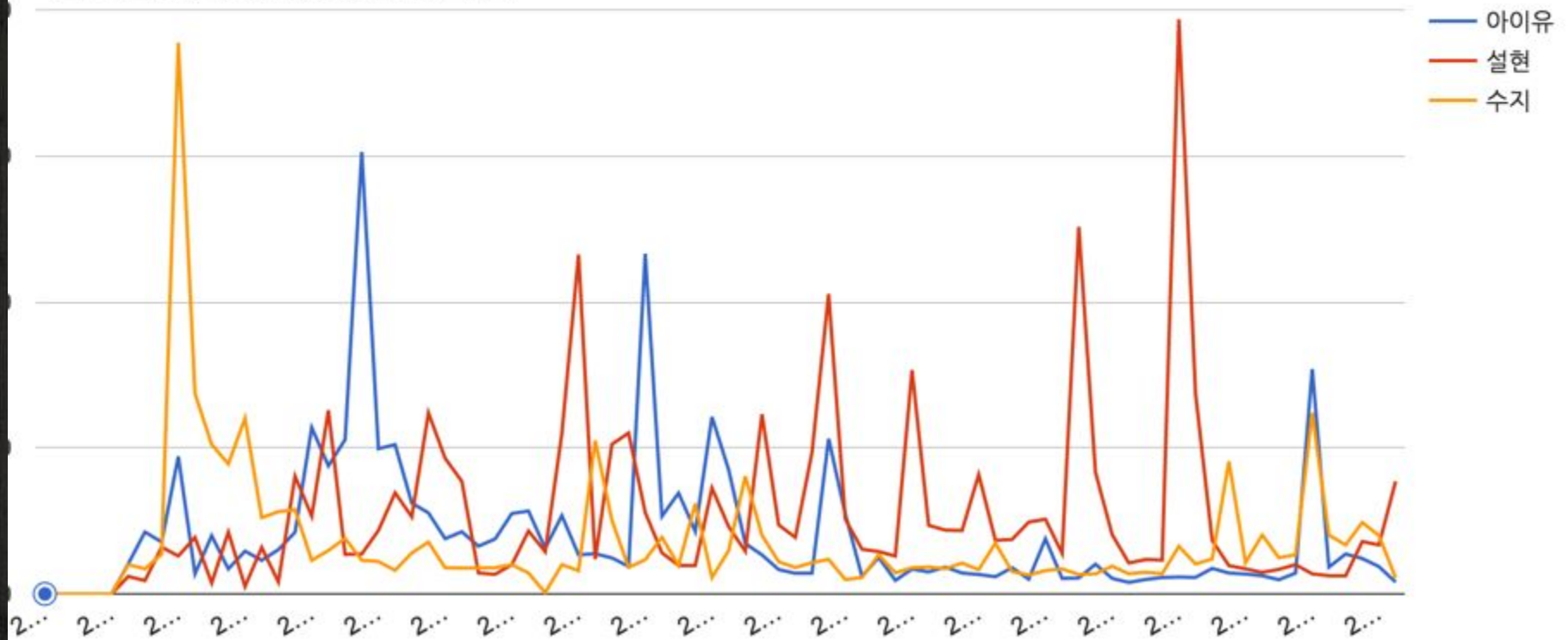
검색 트렌드 분석



설현

수지

비로그인포함 전체, Query Count 주간 그래프



“

+ 보여 드릴 수 없는 데이터

검색 트렌드 분석

1명 / 4개월

(private) 예매 가능 확인

- 잔여석 조회 API가 노출된 일부 사이트
- API를 호출, 응답 내용을 파싱
- 잔여석이 1 이상일 때 메시지를 통해 알림
- 명절 때 요긴

“

예약 어뷰징 시스템
1명 / 0.5 일

Project Prototyping

- 프로젝트 발주 전, 가치 판단을 위해 사용
- 아이디어 구상 -> 프로토타이핑
 - Under 1 week
- 하기 싫은 프로젝트 빠르게 포기 가능

“

이상 사례 발표를 마칩니다

“

*Python*은 사용하기 쉽고, *Reference*가 많습니다

python × 611492

a dynamic and strongly typed programming language designed to emphasize usability. Two similar but incompatible versions of Python are

652 asked today, 3724 this week

ruby × 165237

a multi-platform open-source, dynamic object-oriented interpreted language, created by Yukihiro Matsumoto (Matz) in 1995.

107 asked today, 565 this week

scala × 54191

a general purpose programming language principally targeting the Java Virtual Machine. Designed to express common programming

73 asked today, 304 this week

java × 1114366

Java (not to be confused with JavaScript) is a general-purpose object-oriented programming language designed to be used in conjunction

822 asked today, 4735 this week

c × 224040

a general-purpose computer programming language used for operating systems, libraries, games and other high performance work. It is

123 asked today, 689 this week

StackOverflow 의 Tag 수
2016. 8. 4. 10:15

“

다음, 카카오톡 검색 많이
사용해주세요

Naver나 Google 말고

트래픽 더 받을 수 있어요



감사합니다!
폭삭 속았수다

contact :
dm.k@kakaocorp.com
