시간별 크롤량 그래프 표현

DEMO Version

목적 :

DB를 이용하여 시간당 크롤량 시각화

고찰 :

1) Dash 그래프 표현은 Local 에서 가능하고 DB 쿼리는 mysql 에서
속도가 느려서 지정한 날짜 사이 데이터의 개수 Count 불가능

2) 많은 양의 데이터를 어떻게 실시간으로 표현할 수 있는가

3) 크롤량이 비정상적인 경우 어떠한 기준으로 비정상적인 크롤량을 정의하고 감지

4) 사이트별 크롤량을 시각화 하기 위해서 어떻게 해야하는가
(사이트 정보는 mysql 에 존재하지만 연산가능한 nosql에는 사이트 정보 없음)

고찰:

1) Dash 그래프 표현은 Local 에서 가능하고 DB 쿼리는 mysql 에서
속도가 느려서 지정한 날짜 사이 데이터의 개수 Count 불가능
-> NoSql 에서 연산후 PQ 파일로 저장, local 에서 읽어서 그래프로 표현

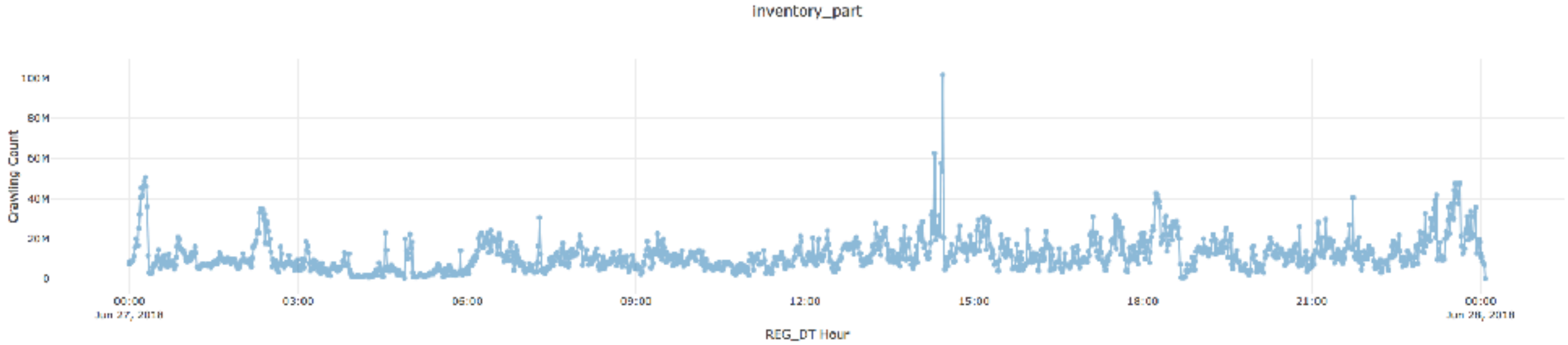
2) 많은 양의 데이터를 어떻게 실시간으로 표현할 수 있는가
-> 현재 해결하지 못함. RDMS는 느려서 NoSql DB 사용 예정 (NoSql 에서의 Dash 불가능한 이슈)

3) 크롤량이 비정상적인 경우 어떠한 기준으로 비정상적인 크롤량을 정의하고 감지
-> 창배님의 도움을 받아 차분을 적용할 예정.

4) 사이트별 크롤량을 시각화 하기 위해서 어떻게 해야하는가
-> 현재 해결 하지 못함. (Presto NoSql 에서의 RDMS DB Join 이 불가능한 이슈)
각 site 정보는 RDMS에 있음

inventory_part

Submit




```
# PERIOD              : 2018-06-27 ~ 2018-06-28
# TOTAL   CRAWL AMOUNT : 135184866316
# TOTAL   CRAWL COUNT  : 1416
# AVERAGE CRAWL AMOUNT : 131555245.41355433
```

시간은 1분 기준
Presto NoSql에 Limit 걸지 않음
미리 연산한 PQ파일을 읽어서 표현
이상데이터 감지 하지 못하고 시각화만 가능
사이트별이 아닌 1분을 기준으로 모든 사이트에서의 크롤량 표현

해결해야할 이슈

1) 실시간 불가능

2) 사이트별로 크롤량 표현

3) 비정상적인 크롤량 감지

4) 비정상적인 상황 감지 후 텔레그램봇 알림 기능

5) 기타 버그 테스트 및 수정

시간별 크롤량 그래프 표현

Version1.0

추가 작업 내용 :

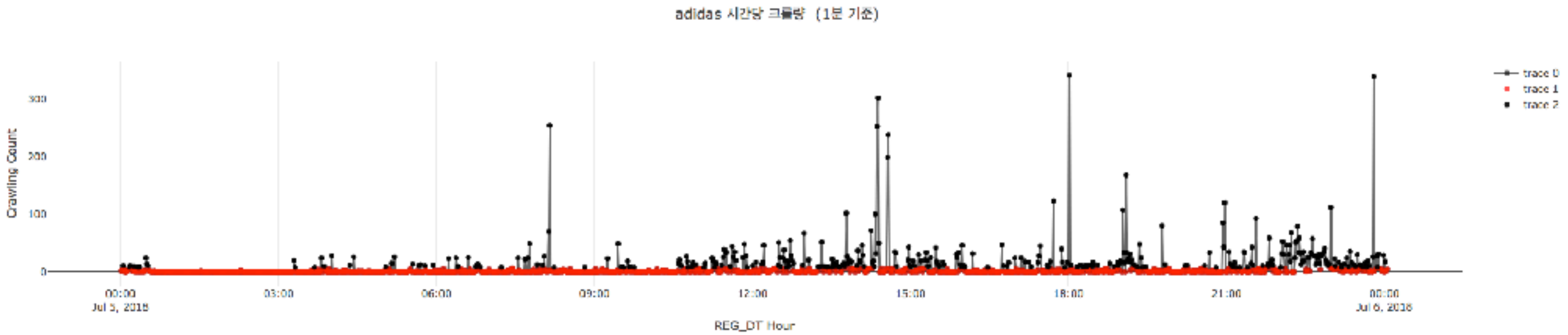
1) 사이트 별로 크롤량 표현 가능

2) http://133.186.159.246:7180 서버에서 Dash 시각화

3) 1분당, 15분당, 1시간 당 그래프 표현 가능

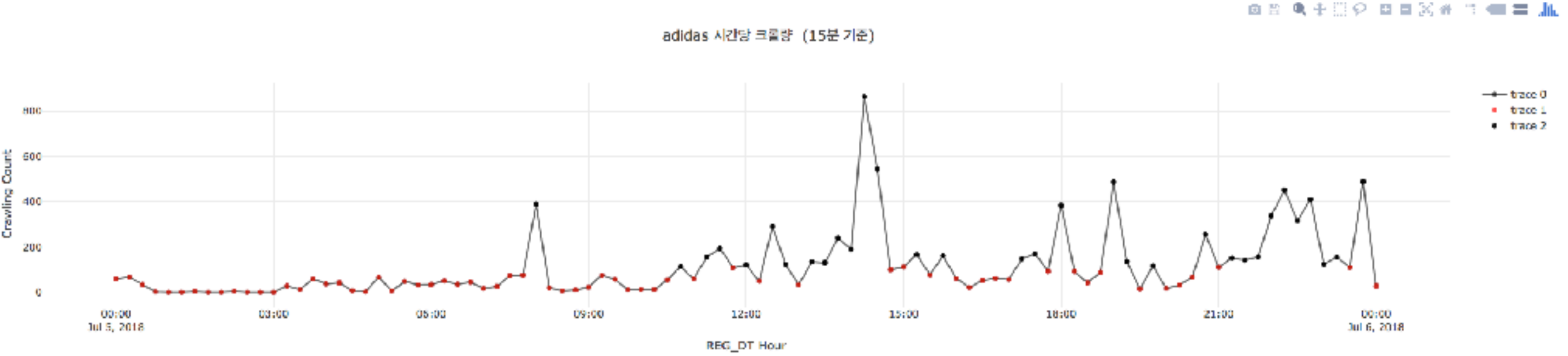
4) Parallel 을 이용하여 CPU 병렬처리로 빠른 속도로 parquet 파일 생성

ventory_part
idas
ubmit

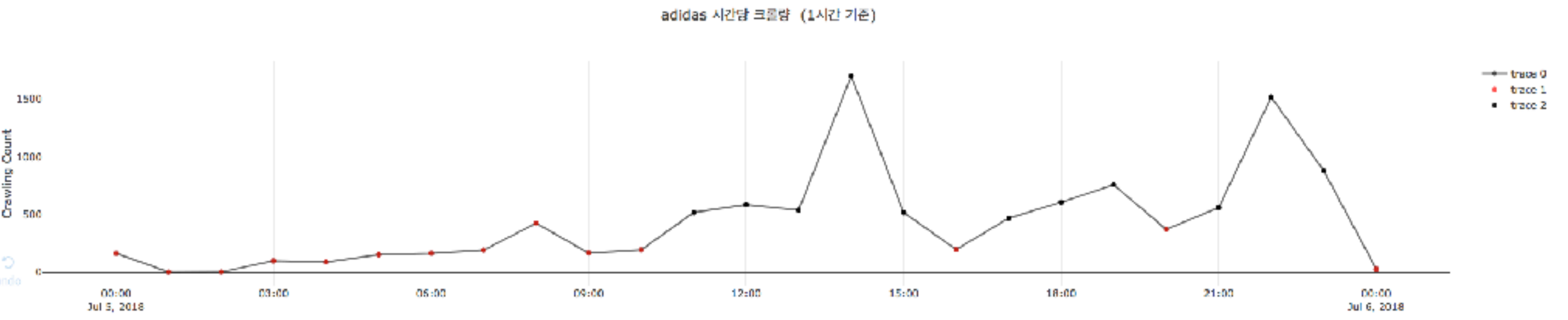


PERIOD : 2018-07-05 ~ 2018-07-06
TOTAL GRANT AMOUNT : 10903
TOTAL GRANT COUNT : 1444
AVERAGE GRANT AMOUNT : 7.550554016620499

1분 기준, Adidas




```
# PERIOD              : 2018-07-05 ~ 2018-07-06
# TOTAL   CRAWL AMOUNT : 10903
# TOTAL   CRAWL COUNT  : 97
# AVERAGE CRAWL AMOUNT : 112.40206185567101
```



15분, 1시간 기준, Adidas

빨간 점 : 평균 이하 값 (차분으로 변경 예정)

개선 사항 :

1) 빠른 속도를 위해 주차 별로 parquet 파일 생성 (현재 월별)

2) pq 파일이 있는 경우 pq 파일을 읽고, pq 파일이 없는 경우 hql 쿼리 실행

3) item_part 추가 (기존에는 inventory_part 이용함. item_part 구현 중)

4) Data 기록 모듈, Dash 시각화 모듈 분리 후 airflow 진행 예정

시간별 크롤량 그래프 표현

Version 2.0

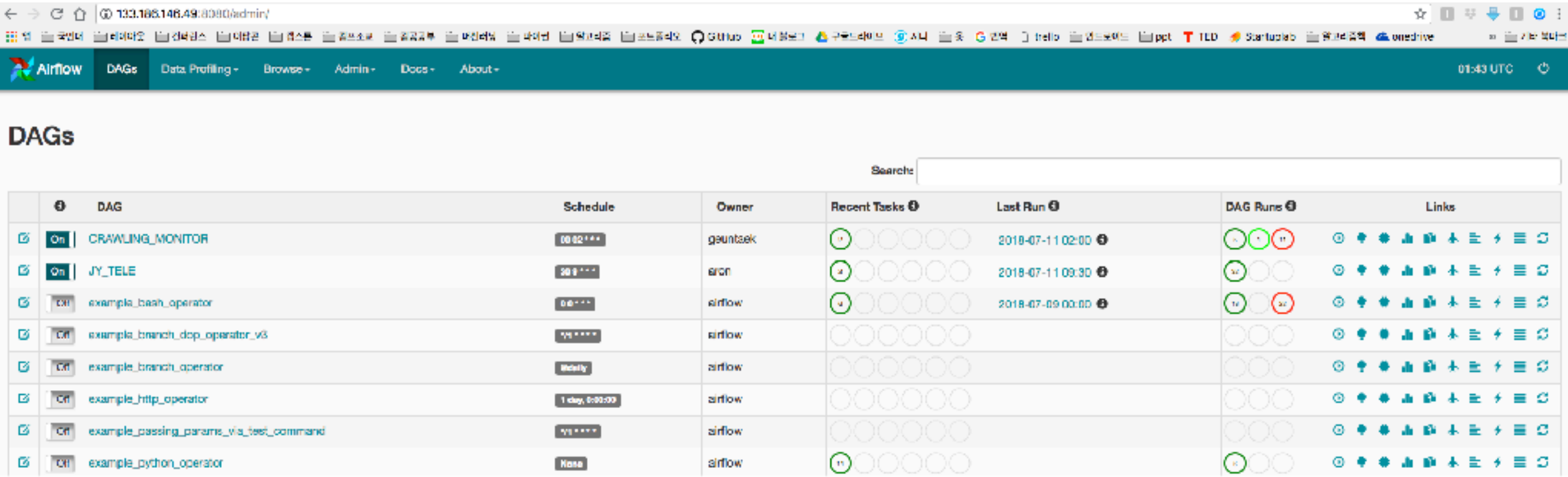
< 데이터 모듈과 시각화 모듈 분리 >

데이터 모듈 : 최근 업데이트 날짜 이후의 Presto 데이터를 가져온 후
사이트, 날짜 별 크롤량을 계산. 그 후 146.142 DB 저장

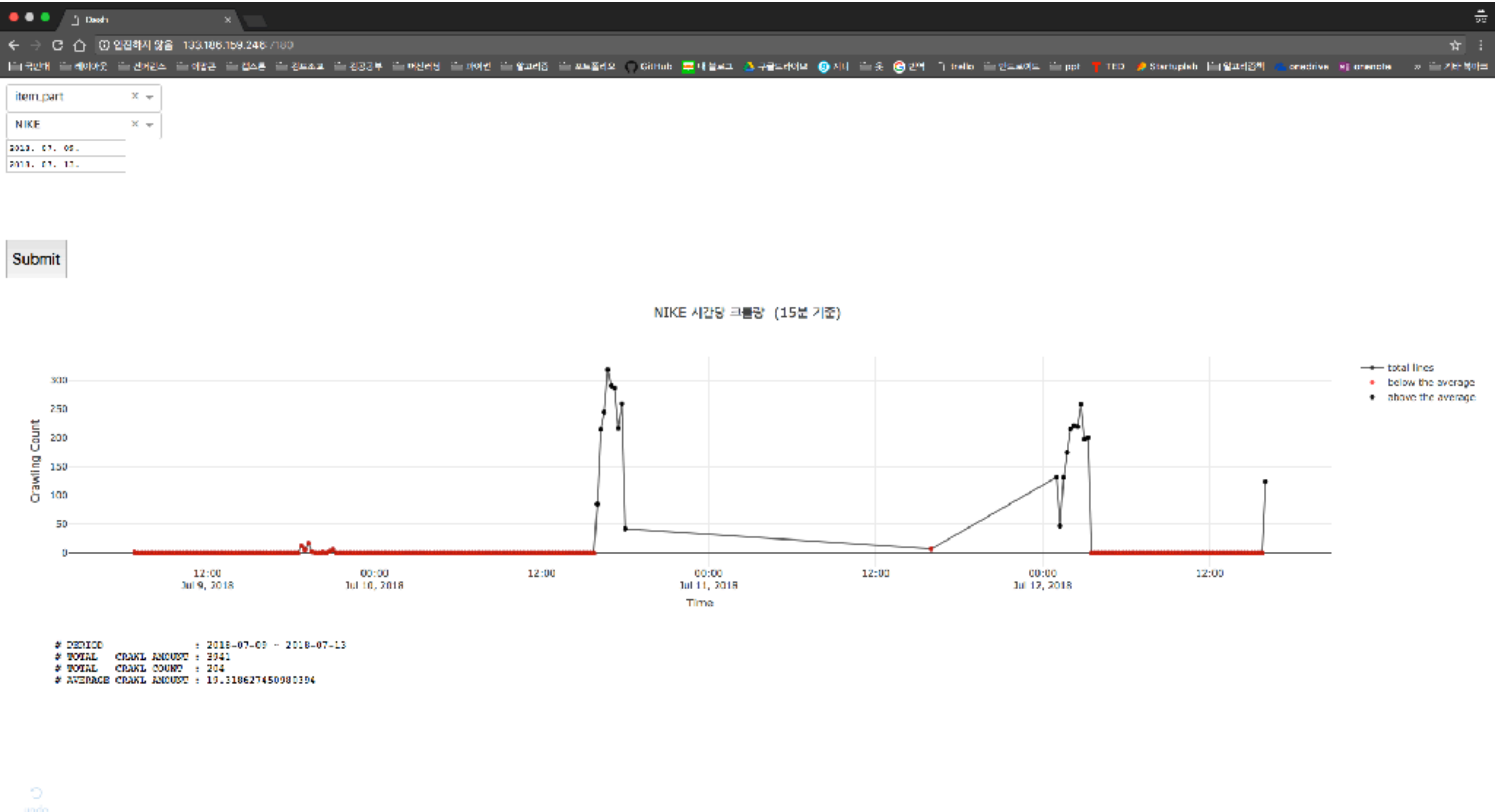
시각화 모듈 : 146.142 DB 의 정보를 시각화

DAGs : 새벽 2시에 주기적으로 데이터 모듈 실행 (정확한 시간은 아님)

1분, 1시간 기준은 제외하고 15분 기준으로만 표현



Airflow 에서 새벽 2시마다 실행
(서버 상황에 따라 실행 시간은 변경 될 수도 있음)


Dash
item_part
NIKE
Submit
NIKE 시간당 크롤량 (15분 기준)
Crawling Count
Time
total lines
below the average
above the average
12:00 Jul 9, 2018
00:00 Jul 10, 2018
00:00 Jul 11, 2018
00:00 Jul 12, 2018
PERIOD : 2018-07-09 ~ 2018-07-13
TOTAL CRAWL AMOUNT : 3941
TOTAL CRAWL COUNT : 204
AVERAGE CRAWL AMOUNT : 19.318627450980394
undo

Dash

안전하지 않음 133.186.159.246:7100

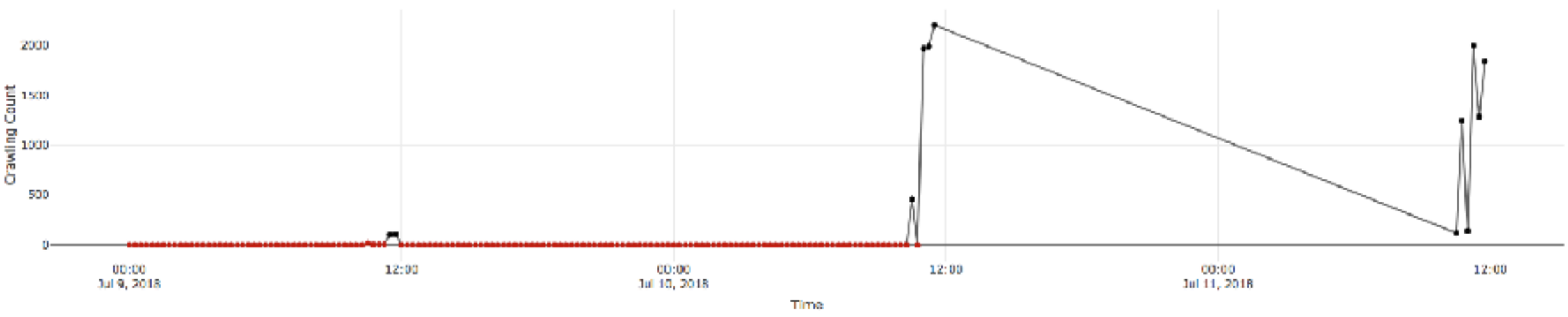
item_part

OLIVEYOUNG

2018. 07. 09.

2018. 07. 13.

Submit




```
# DERIOD              : 2018-07-09 ~ 2018-07-13
# TOTAL   CRAWL AMOUNT : 13495
# TOTAL   CRAWL COUNT  : 140
# AVERAGE CRAWL AMOUNT : 90.35046979865771
```

undo

http://133.186.159.246:7180/

한번 써보세요..