

바닥부터 시작하는 데이터 인프라

- 변성윤 -



- 목차 -

- 발표자 소개
- 본 세션을 듣기 전에 알아야 할 회사 소개
- 오늘 전해드릴 이야기
 - Story 1. 대시보드 만들기
 - Story 2. 데이터 파이프라인 생성
 - Story 3. 음란 사진이 올라온다
- 정리

발표자 소개

변성윤

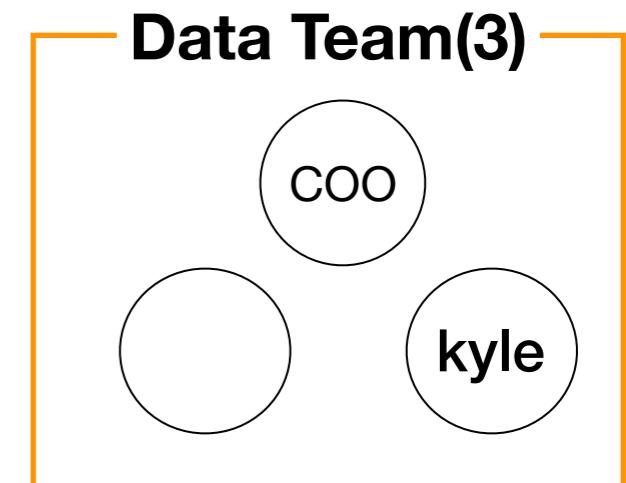
www.instagram.com/data.scientist/

- Team NeighborHood (<https://github.com/Team-Neighborhood>)
- 경영학 전공
- 광고 동아리 (디자인) → 공기업(인천도시공사) → 창업 및 크라우드펀딩 경험 → 빅데이터 동아리 → 패스트캠퍼스 데이터사이언스 스쿨 2기
- 레트리카에서 근무
(데이터 분석가 - 데이터 엔지니어 - 백엔드 엔지니어)
- 현재 퇴사하고 자유롭게 공부중

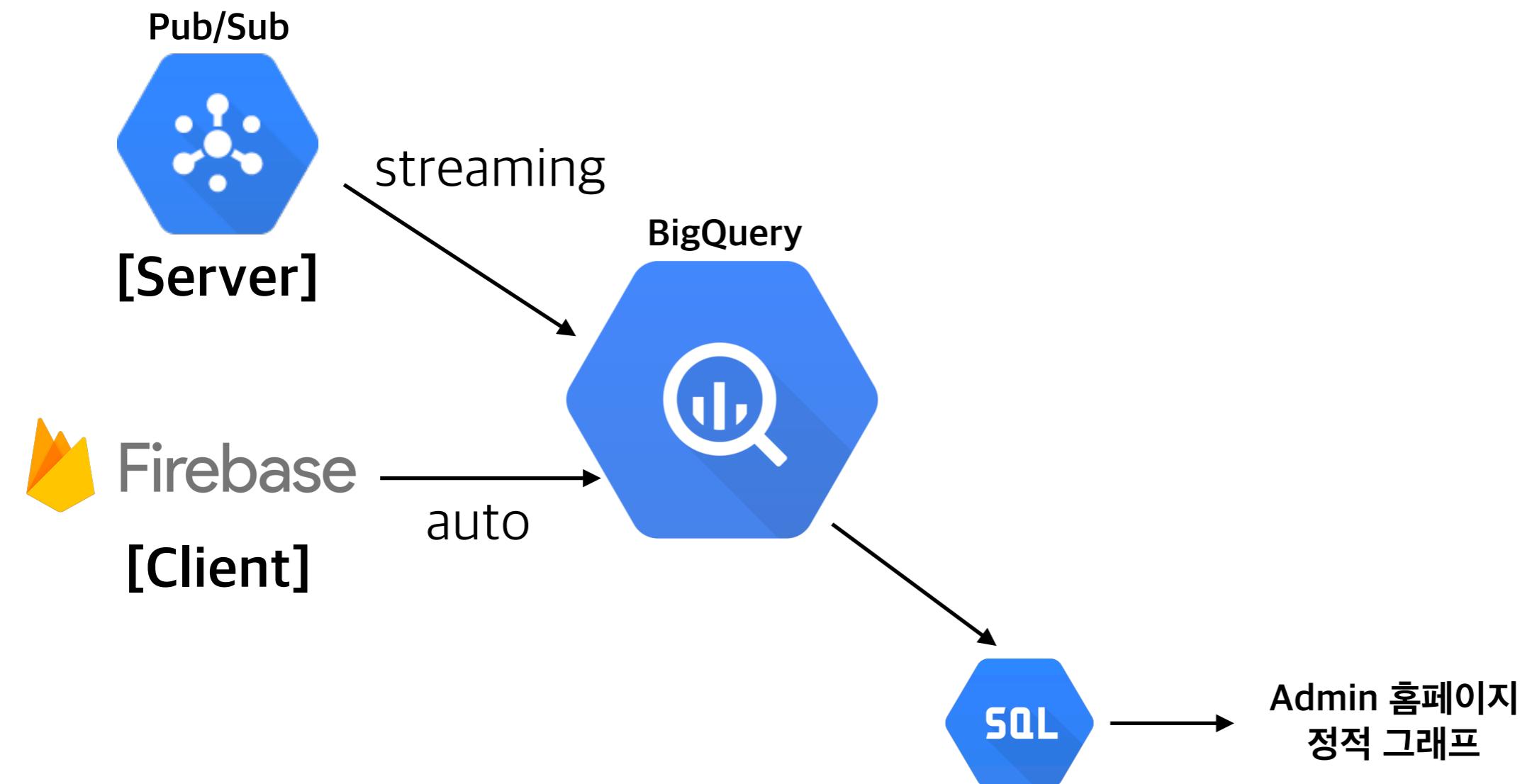
본 세션을 듣기 전에 알아야 할 회사 소개



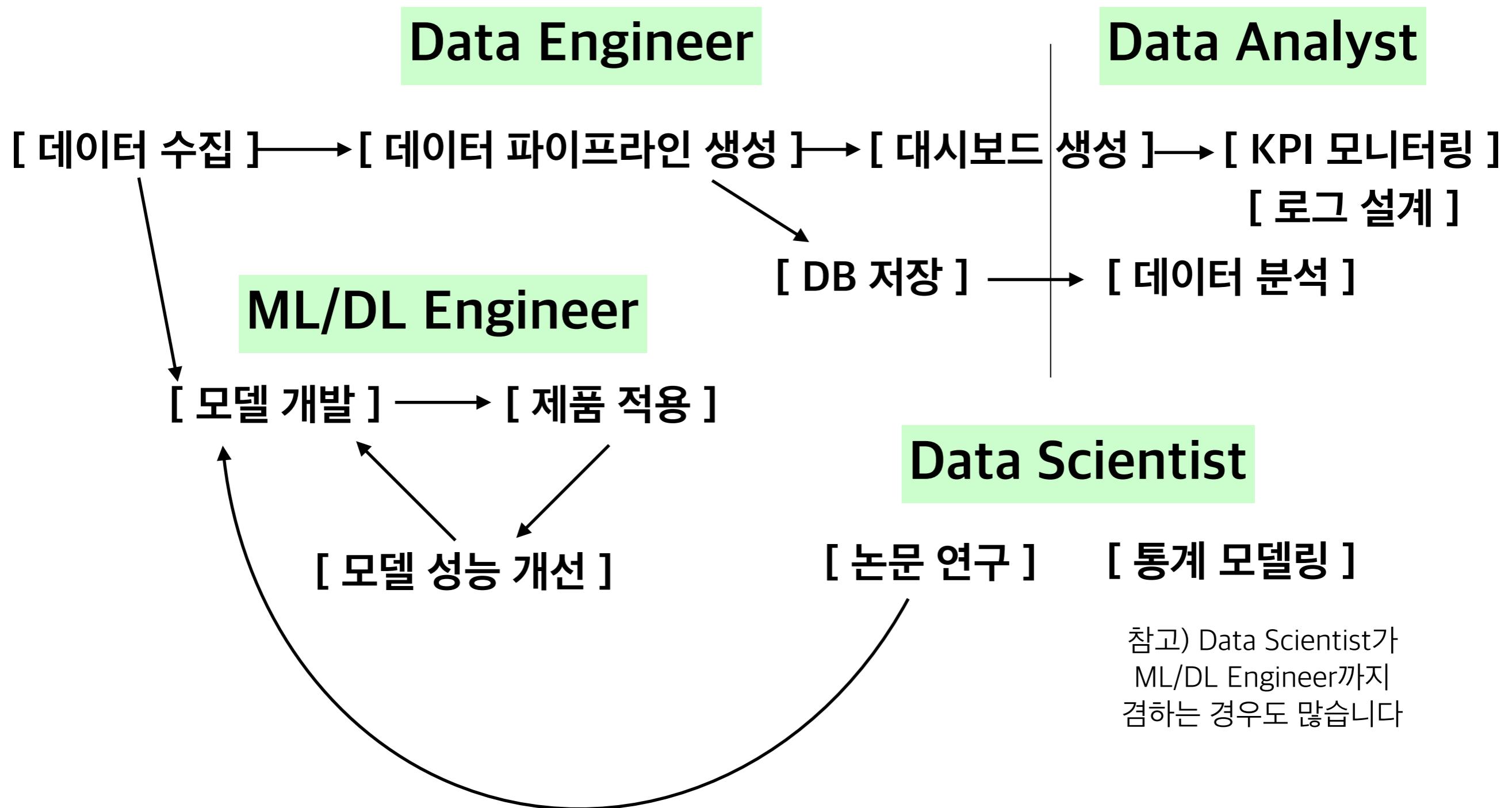
- 카메라 어플 ‘Retrica’
총 누적 다운로드 **3억 이상!**
주간 1400만명의 사용자
하루에 쌓이는 log는 **200g**
- 터키, 인도, 러시아 등에서 많이 사용중
- Google Cloud Platform / Firebase를 사용
- Firebase-BigQuery를 사용하기 때문에 **데이터 수집 비용 : Zero**
- DB는 BigTable, Datastore, **BigQuery, MySQL** 사용



입사 당시 데이터 파이프라인



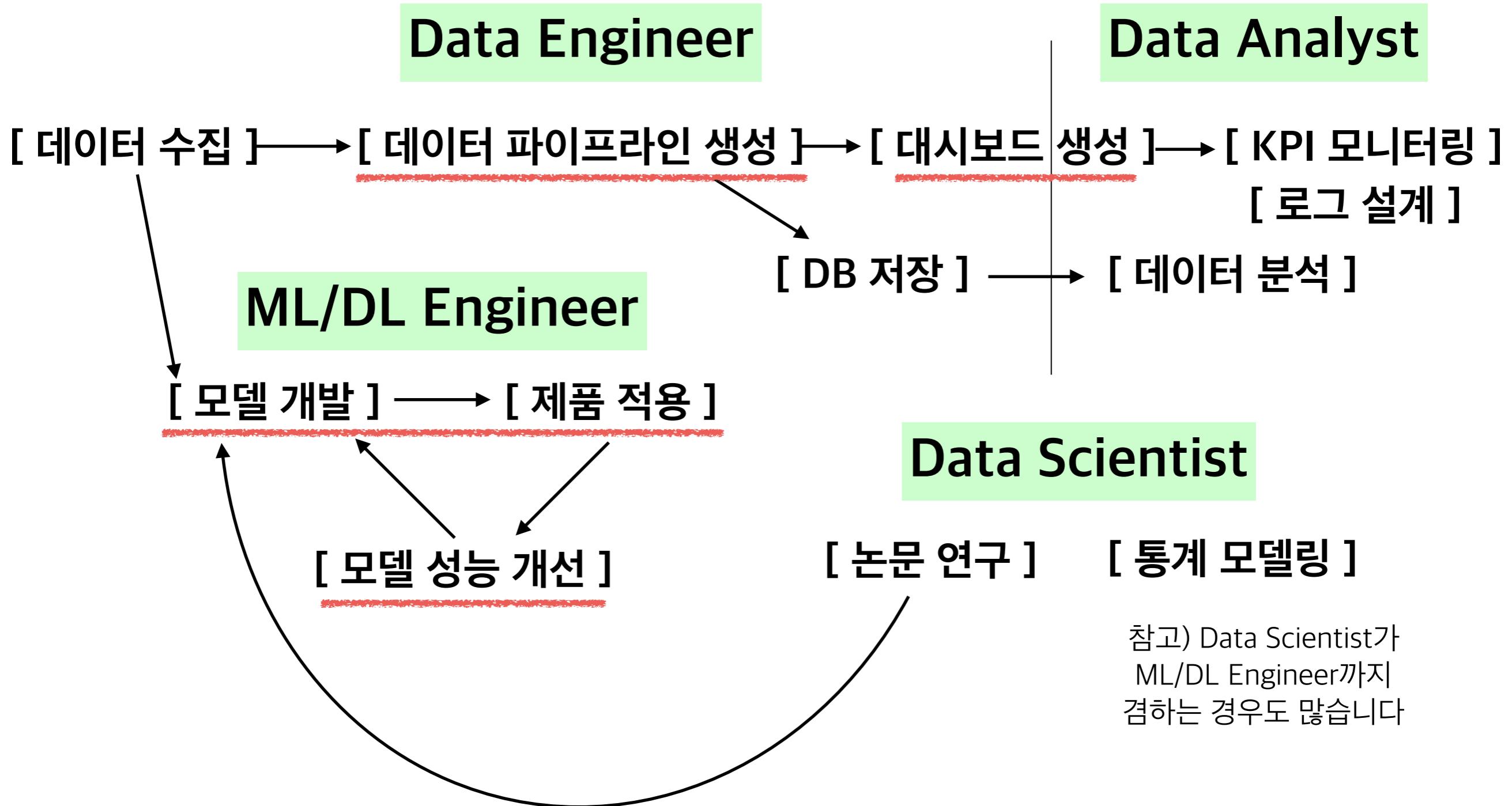
Data Job Overview



참고) Data Scientist가
ML/DL Engineer까지
겸하는 경우도 많습니다

오늘 전해드릴 이야기

[스타트업에서 데이터 관련 인프라를 하나씩 쌓아가며 겪은 삽질]



오늘 전해드릴 이야기

[스타트업에서 데이터 관련 인프라를 하나씩 쌓아가며 겪은 삽질]



제한된 시간 내에 목표 달성을 위해 했던 고민들



빠르게 프로토 타입을 뽑은 후, 발생한 요청/문제를 해결하는 과정

오늘 전해드릴 이야기

[스타트업에서 데이터 관련 인프라를 하나씩 쌓아가며 겪은 삽질]



제한된 시간 내에 목표 달성을 위해 했던 고민들



빠르게 프로토 타입을 뽑은 후, 발생한 요청/문제를 해결하는 과정

[예상 독자]

데이터 인프라를 구축해야 하는 분

스타트업에서 어떤 일을 하는지

궁금하신 분

Firebase-BigQuery를 사용하는 분

[다루지 않는 내용]

기술에 대한 깊은 내용

Hadoop, Spark

Story 1. 대시보드 만들기

입사하고 본 환경

1

[Admin Web Page]

date	country	platform	dau
20170210	IN	android	1,000
20170210	IN	ios	500
20170210	TR	android	800
20170210	TR	ios	200

...

입사하고 본 환경

1

[Admin Web Page]

date	country	platform	dau
20170210	IN	android	1,000
20170210	IN	ios	500
20170210	TR	android	800
20170210	TR	ios	200

...

[Ctrl + C]

[Ctrl + V]



입사하고 본 환경

1

[Admin Web Page]

date	country	platform	dau
20170210	IN	android	1,000
20170210	IN	ios	500
20170210	TR	android	800
20170210	TR	ios	200

...

[Ctrl + C]

[Ctrl + V]



입사하고 본 환경

1

[Admin Web Page]

date	country	platform	dau
20170210	IN	android	1,000
20170210	IN	ios	500
20170210	TR	android	800
20170210	TR	ios	200

...

[Ctrl + C]

[Ctrl + V]



[전사 공유]

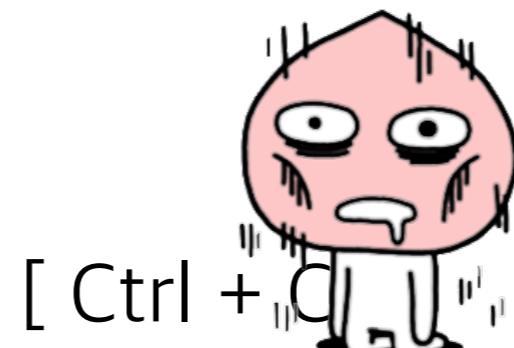


입사하고 본 환경

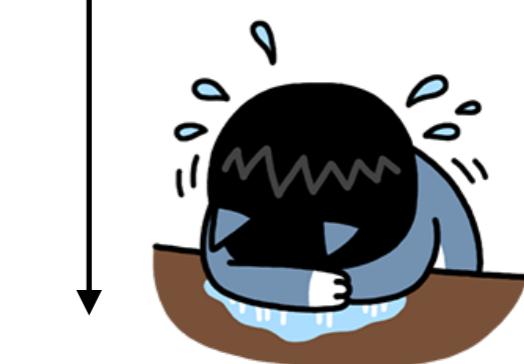
1

[Admin Web Page]

date	country	platform	dau
20170210	IN	android	1,000
20170210	IN	ios	500
20170210	TR	android	800
20170210	TR	ios	200
...			



[Ctrl + V]



[전사 공유]

입사하고 본 환경

2



zzsza 6:09 PM

@data-team 17년 3월 필터 사용량 궁금하네요. 알려주세요-! 😊

입사하고 본 환경

2



zzsza 6:09 PM

@data-team 17년 3월 필터 사용량 궁금하네요. 알려주세요-! 😊



Seongyun 🌴 6:11 PM

@zzsza

요청하신 데이터입니다 😊

https://docs.google.com/spreadsheets/d/1wVt4kfkKf56wwyg98_v25-Omefh-SNTpn4EUOXTqZuw/edit#gid=0

x N번

입사하고 본 환경

2



zzsza 6:09 PM

@data-team 17년 3월 필터 사용량 궁금하네요. 알려주세요! 😊



Seongyun 🌴 6:11 PM

@zzsza

요청하신 데이터입니다 😊

https://docs.google.com/spreadsheets/d/1wVt4kfkKf56wwyg98_v25-Omefh-SNTpn4EUOXTqZuw/edit#gid=0

x N번

중복 요청 발생

Dependency가 생김

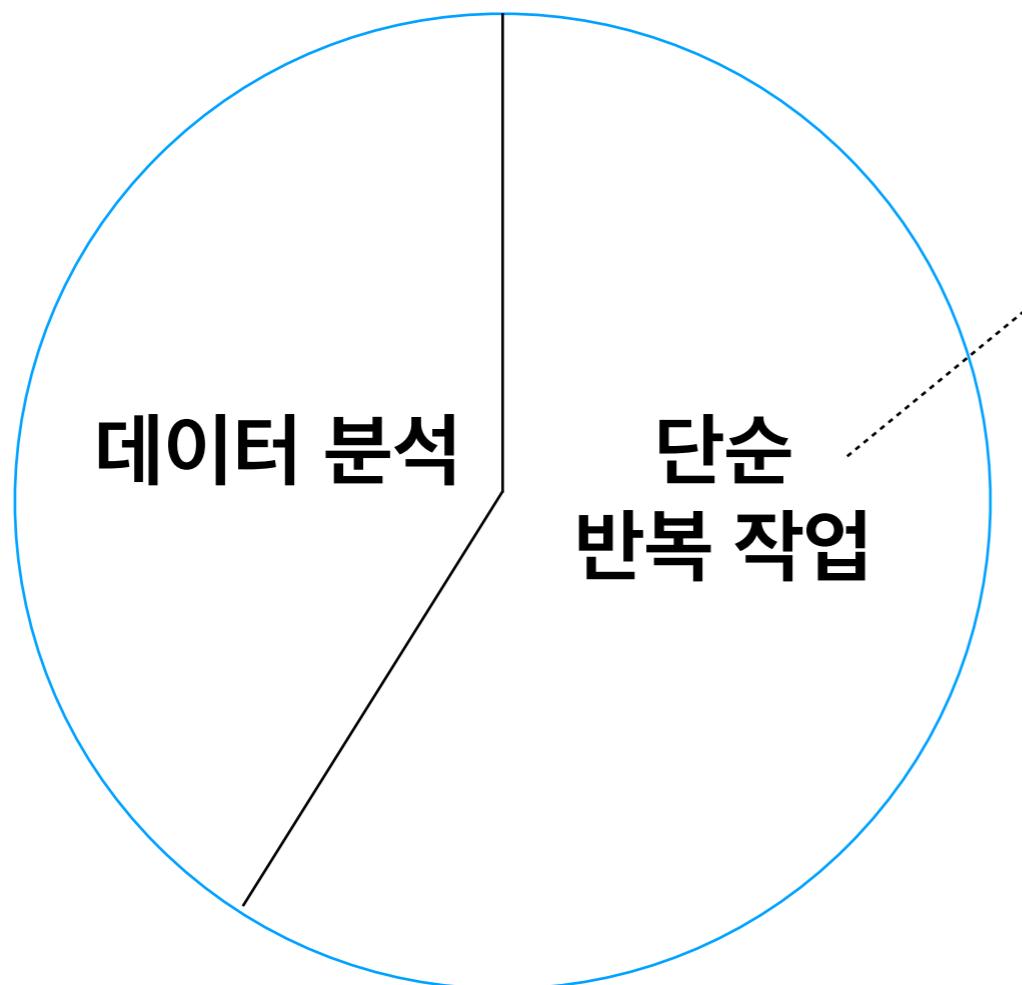
회의할 때 데이터 찾는 경우 많음(=빠르게 답을 줘야하는 경우)

휴가때도 데이터 요청



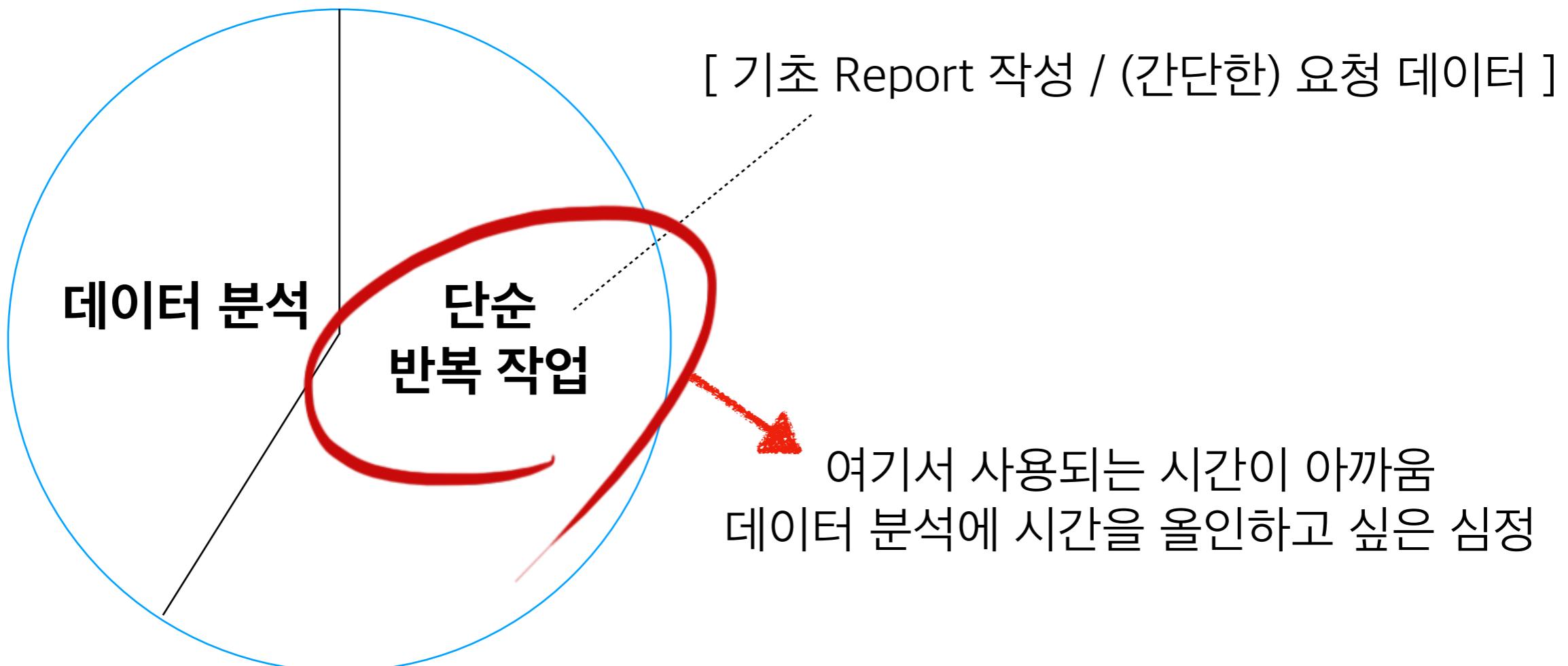
발견한 문제점

[데이터 분석가의 Daily 업무 비율]



발견한 문제점

[데이터 분석가의 Daily 업무 비율]



발견한 문제점

[데이터 분석가의 Daily 업무 비율]



발견한 문제점

[데이터 분석가의 Daily 업무 비율]



제안



캡틴, **대시보드**를
만들겠습니다

제안



캡틴, 대시보드를
만들겠습니다

좋아요.
2주안에 콜?

캡틴

제안



캡틴, **대시보드**를
만들겠습니다

(..!?) 네!
해보겠습니다!

좋아요.
2주안에 콜?

캡틴

제안



캡틴, 대시보드를
만들겠습니다

(..!?) 네!
해보겠습니다!

좋아요.
2주안에 콜?

우선 주요 Report용
Dashboard만 해봐요~

캡틴

제안



캡틴, 대시보드를
만들겠습니다

(..!?) 네!
해보겠습니다!

좋아요.
2주안에 콜?

우선 주요 Report용
Dashboard만 해봐요~

캡틴

사실 저 웹개발의 ○도 모르는
그저 데이터쟁이,,,

열정 넘치던 1주



Flask를 써볼까^^?

	내용	기한
Step 1	<ul style="list-style-type: none">▪ 현재 Daily Report 양식과 유사하게 웹에 노출 (DAU / TPU / Ios / Android)	28 Mar 2017
Step 2	<ul style="list-style-type: none">▪ 시각화를 위한 기본 요소 추가 (숫자에 넘버링 (ex) , 추가), 우측정렬, 증감률에 +일 경우 -일 경우 설정 css로 이쁘게 표현하는것보단, 기본적인 레포트의 요소에 충실히 시각화	02 Apr 2017
Step 3	<ul style="list-style-type: none">▪ Store Chart 데이터 (앱 랭킹) -> SQL	03 Apr 2017
	<ul style="list-style-type: none">▪ dau / wau / mau 데이터 추가 요청	03 Apr 2017
	<ul style="list-style-type: none">▪ wau 삽입 완료	04 Apr 2017
	<ul style="list-style-type: none">▪ dau / wau / mau 추가 데이터를 기반으로 추가 대시보드 생성	04 Apr 2017

후회

1:00 AM



나는 개발을 하러 온 것인가 대시보드를 만들러 온 것인가
데이터 분석을 하러 온 것인가 무엇을 하러 왔기에 나는 여기에 있는가
오늘도 택시비가 5만원 나올 것 같다.
생각해보니 내가 다 만들지 않아도 된다

후회

1:00 AM



나는 개발을 하러 온 것인가 대시보드를 만들러 온 것인가
데이터 분석을 하러 온 것인가 무엇을 하러 왔기에 나는 여기에 있는가
오늘도 택시비가 5만원 나올 것 같다.
생각해보니 내가 다 만들지 않아도 된다

MySQL을 **시각화**해주는 오픈소스들이
세상에 많지 않을까?

환호

[세상은 역시 따뜻해]



Zeppelin



metabase



Tableau



Data Studio



Stack



BI Tool 비교



Base
Code

GUI(단순 Click으로 생성 가능)

Price

오픈소스 사용시 무료
(회사에서 호스팅해주는 경우는 유료)

유료
(Desktop, Server,
Online) 총 3종

무료

DB

MySQL
BigQuery

MySQL

MySQL
BigQuery

Elastic
Search

Github
Star

3,510

8,932

9,520

18,842

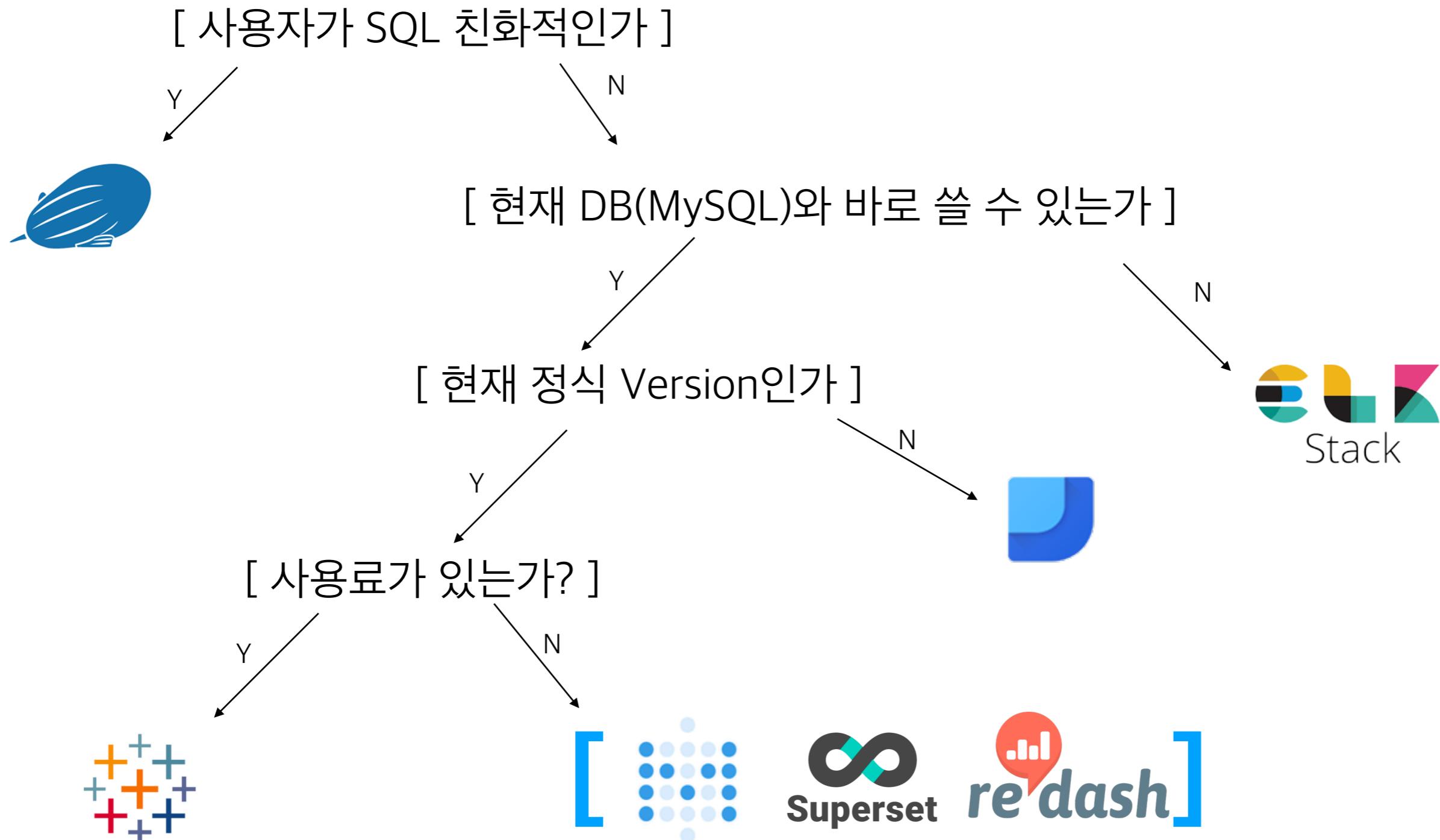
-

-

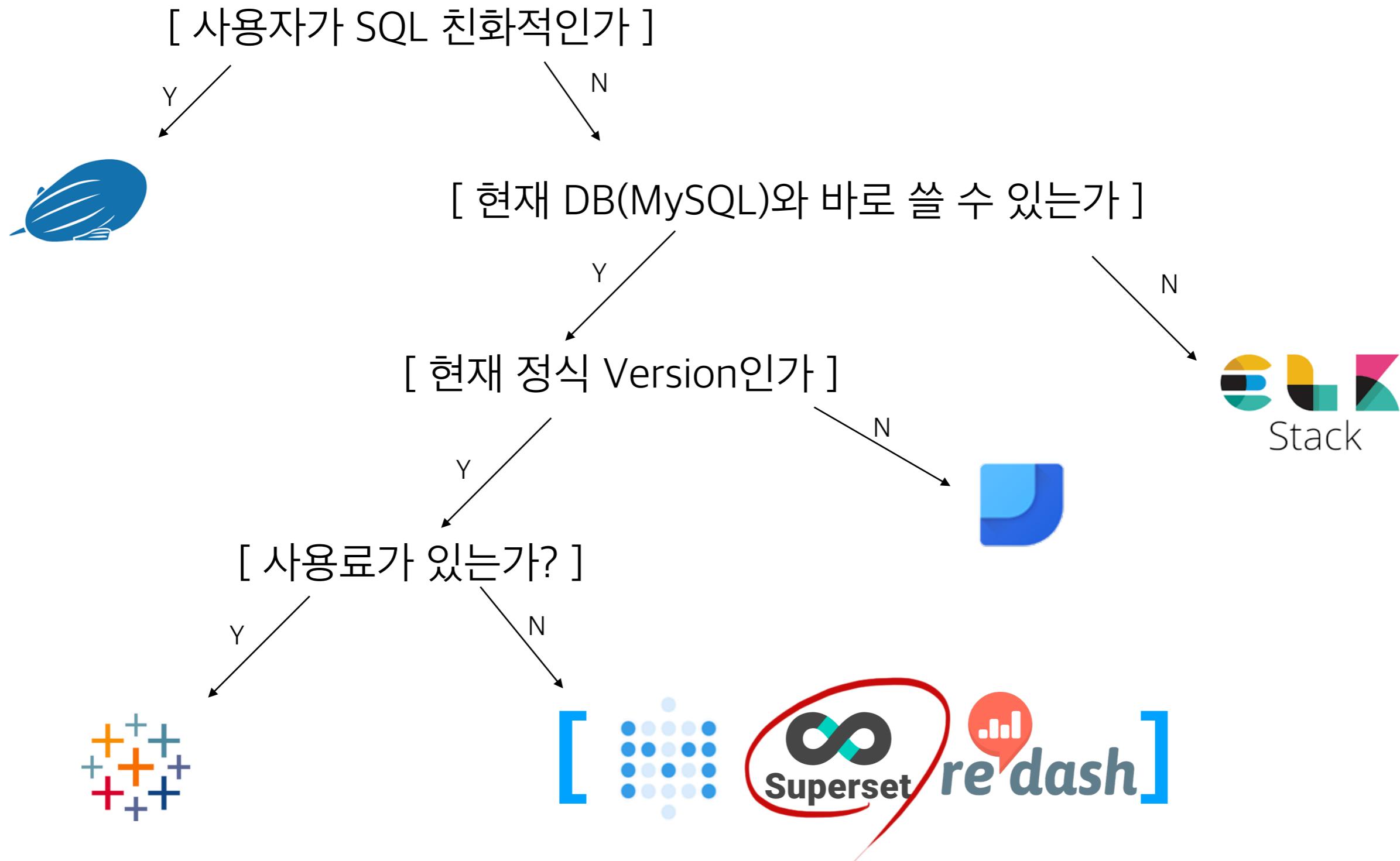
E : 30,374
L : 8,857
K : 9,208

2018년 4월 21일 01:00 기준

Dashboard 고민 과정



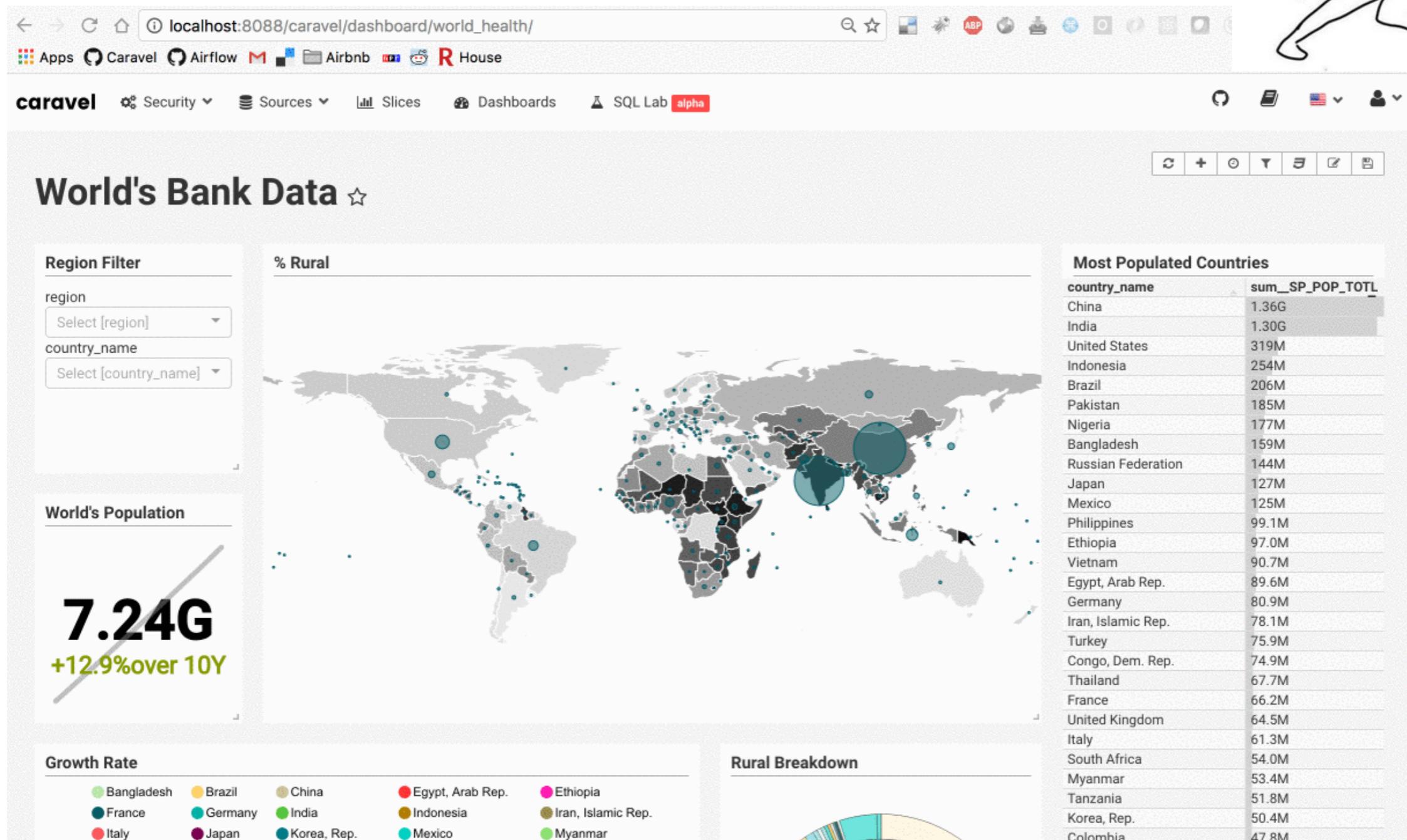
Dashboard 고민 과정



환희



인스턴스 올리고 환경 셋팅하는데 걸린 시간 : 60분



출처 : Superset 공식 문서

Story 1. Summary

문제 상황

Daily 업무 비중이 데이터 분석보다
단순 반복 작업
(Report 작성, 요청 데이터 처리)이
많음

해결 방안

Dashboard 생성
Superset을 활용해 MySQL와 연결한 후,
직접 그래프 생성

의의

단순 반복 작업의 자동화
데이터 분석가들의 시간 효율성 증대
(단순 작업의 비율이 60% -> 10%)

전 직원들이 스스로 데이터 탐색 가능

Story 2. 데이터 파이프라인 생성

대표님의 한마디



헤헷, 네!

이제 주요 지표를
빠르게 볼 수 있겠네요!

캡틴

대표님의 한마디



헤헷, 네!

(ㅎㅎㅎㅎ;;) 네.

이제 주요 지표를
빠르게 볼 수 있겠네요!

그럼 이젠
모든 이벤트
Dashboard 해볼까요?

캡틴

대표님의 한마디



헤헷, 네!

(ㅎㅎㅎㅎ;;) 네.

이제 주요 지표를
빠르게 볼 수 있겠네요!

그럼 이젠
모든 이벤트
Dashboard 해볼까요?

이번에도 2주안에~

캡틴

대표님의 한마디



헤헷, 네!

(ㅎㅎㅎㅎ;;) 네.

이제 주요 지표를
빠르게 볼 수 있겠네요!

그럼 이젠
모든 이벤트
Dashboard 해볼까요?

이번에도 2주안에~

아 그리고 BigQuery
비용이 많이 나가는데
확인 부탁해요!

캡틴

정리해봅시다

목표 1 : 이벤트 레벨까지 데이터를 조회할 수 있는 Dashboard

필요한 정보 : 현재 이벤트 로그는 정리되어 있는가? Table 형태가 아니라면 못씀

정리해봅시다

목표 1 : 이벤트 레벨까지 데이터를 조회할 수 있는 Dashboard

필요한 정보 : 현재 이벤트 로그는 정리되어 있는가? Table 형태가 아니라면 못씀

BigQuery는 아래와 같은 형태로 **한 테이블**에 user_dim / event_dim **2개의 차원**이 존재

event_dim.date	event_dim.name	event_dim.params.key	event_dim.params.value.string_value
20170914	Save_Name	name	firebase
		firebase_event_origin	app
20170914	Save_Name	name	bigquery
		firebase_event_origin	app
20170914	Save_Name	firebase_event_origin	app
		name	firebase
20170914	Save_Name	name	seongyun
		firebase_event_origin	app

정리해봅시다

목표 1 : 이벤트 레벨까지 데이터를 조회할 수 있는 Dashboard

필요한 정보 : 현재 이벤트 로그는 정리되어 있는가? Table 형태가 아니라면 못씀

BigQuery는 아래와 같은 형태로 **한 테이블**에 user_dim / event_dim **2개의 차원**이 존재

Table 형태로 가공만 되면 바로 시각화 가능!

event_dim.date	event_dim.name	event_dim.params.key	event_dim.params.value.string_value
20170914	Save_Name	name	firebase
		firebase_event_origin	app
20170914	Save_Name	name	bigquery
		firebase_event_origin	app
20170914	Save_Name	firebase_event_origin	app
		name	firebase
20170914	Save_Name	name	seongyun
		firebase_event_origin	app

정리해봅시다

목표 2 : BigQuery 비용 절감

필요한 정보 : 현재 BigQuery에 소요되는 비용은? 그리고 그 원인은?

정리해봅시다

목표 2 : BigQuery 비용 절감

필요한 정보 : 현재 BigQuery에 소요되는 비용은? 그리고 그 원인은?

5월 BigQuery Analytics 비용 : **\$5,000** 이상
Analytics 비용(=Query)은 1TB에 **\$5**

정리해봅시다

목표 2 : BigQuery 비용 절감

필요한 정보 : 현재 BigQuery에 소요되는 비용은? 그리고 그 원인은?

5월 BigQuery Analytics 비용 : **\$5,000** 이상
Analytics 비용(=Query)은 1TB에 **\$5**

Query로 1,000TB????



선인의 지혜를 찾아 삼만리

BigQuery에 대한 정보는 많이 없는 상황 + 회사에 도와줄 분이 없음(=사수의 부재)

선인의 지혜를 찾기위해 Slideshare 검색

3 people clipped this slide

N-BT

대용량 로그분석, BigQuery로 간단히 사용하기

SK T아카데미 (2017.02.15)

(주)엔비티 / Devops / 이재광
http://facebook.com/openstacks

1 of 66

1,264 views

대용량 로그분석 Bigquery로 간단히 사용하기
(20170215 T아카데미)

<https://www.slideshare.net/openstacks/bigquery-20170215-t>

선인 이재광님(NBT)과의 대화

빅쿼리에 모든 데이터를 1차로 저장하는 것도 나쁘지 않은 방법이에요.

일단 disk나 성능 문제가 전혀 없어서 무조건 밀어넣기만 하면 되거든요. 관리가 필요없습니다.

그리고 말씀하신 것 처럼 조회 데이터가 너무 많을 경우

이에 대한 사전 처리를 빅쿼리에서 한번 1차 가공한 데이터를 가지고 임시 또는 신규 테이블을 만들어서 데이터 분석용으로 사용하시면 좋아요.

빅쿼리가 조회한 데이터를 기반으로 신규 테이블을 만들기가 워낙 간단해서 이런 작업은 수월하게 하실 수 있을걸 겁니다.

그리고 데이터를 파티션 테이블 형태로 만드셔서 기간에 대한 range 검색을 하시는 것도 좋은 방법이구요.



저희도 빅쿼리로 데이터를 모두 퍼담고 이 데이터를 가지고
용도의 데이터로 각각 가공해서 별도의 테이블에 다시 저장해서 사용중입니다.

각각 봐야하는 지표들이 다들 달라서 굳이 전체 데이터를 대상으로 매번 조회하는게 부담되더라구요.

넵! 저희는 구글 빅쿼리 외에도 카프카를 data lake로 사용하는 플랫폼도 있어요. 혹시 나중에 ELK를 사용하시게 되면 카프카를 한번 꼭 고민해보시고 궁금하신건 언제든 질문주세요.



선인 이재광님(NBT)과의 대화

빅쿼리에 모든 데이터를 1차로 저장하는 것도 나쁘지 않은 방법이에요.

일단 disk나 성능 문제가 전혀 없어서 무조건 밀어넣기만 하면 되거든요. 관리가 필요없습니다.

그리고 말씀하신 것 처럼 조회 데이터가 너무 많을 경우

이에 대한 사전 처리를 빅쿼리에서 한번 1차 가공한 데이터를 가지고 임시 또는 신규 테이블을 만들어서 데이터 분석용으로 사용하시면 좋아요.

빅쿼리가 조회한 데이터를 기반으로 신규 테이블을 만들기가 워낙 간단해서 이런 작업은 수월하게 하실 수 있을걸 겁니다.

그리고 데이터를 파티션 테이블 형태로 만드셔서 기간에 대한 range 검색을 하시는 것도 좋은 방법이구요.



BigQuery는 쿼리하는 만큼
(데이터 사이즈에 비례해)
비용이 부과됩니다

저희도 빅쿼리로 데이터를 모두 퍼담고 이 데이터를 가지고
용도의 데이터로 각각 가공해서 별도의 테이블에 다시 저장해서 사용중입니다.

각각 봐야하는 지표들이 다들 달라서 굳이 전체 데이터를 대상으로 매번 조회하는게 부담되더라구요.

넵! 저희는 구글 빅쿼리 외에도 카프카를 data lake로 사용하는 플랫폼도 있어요. 혹시 나중에 ELK를 사용하시게 되면 카프카를 한번 꼭 고민해보시고 궁금하신건 언제든 질문주세요.



데이터 가공의 시작

event_dim.date	event_dim.name	event_dim.params.key	event_dim.params.value.string_value
20170914	Save_Name	name	firebase
		firebase_event_origin	app
20170914	Save_Name	name	bigquery
		firebase_event_origin	app
20170914	Save_Name	firebase_event_origin	app
		name	firebase
20170914	Save_Name	name	seongyun
		firebase_event_origin	app

[Flatten]



event_dim.date	event_dim.name	name	firebase_event_origin
20170914	Save_Name	firebase	app
20170914	Save_Name	bigquery	app
20170914	Save_Name	firebase	app
20170914	Save_Name	seongyun	app

데이터 가공의 시작

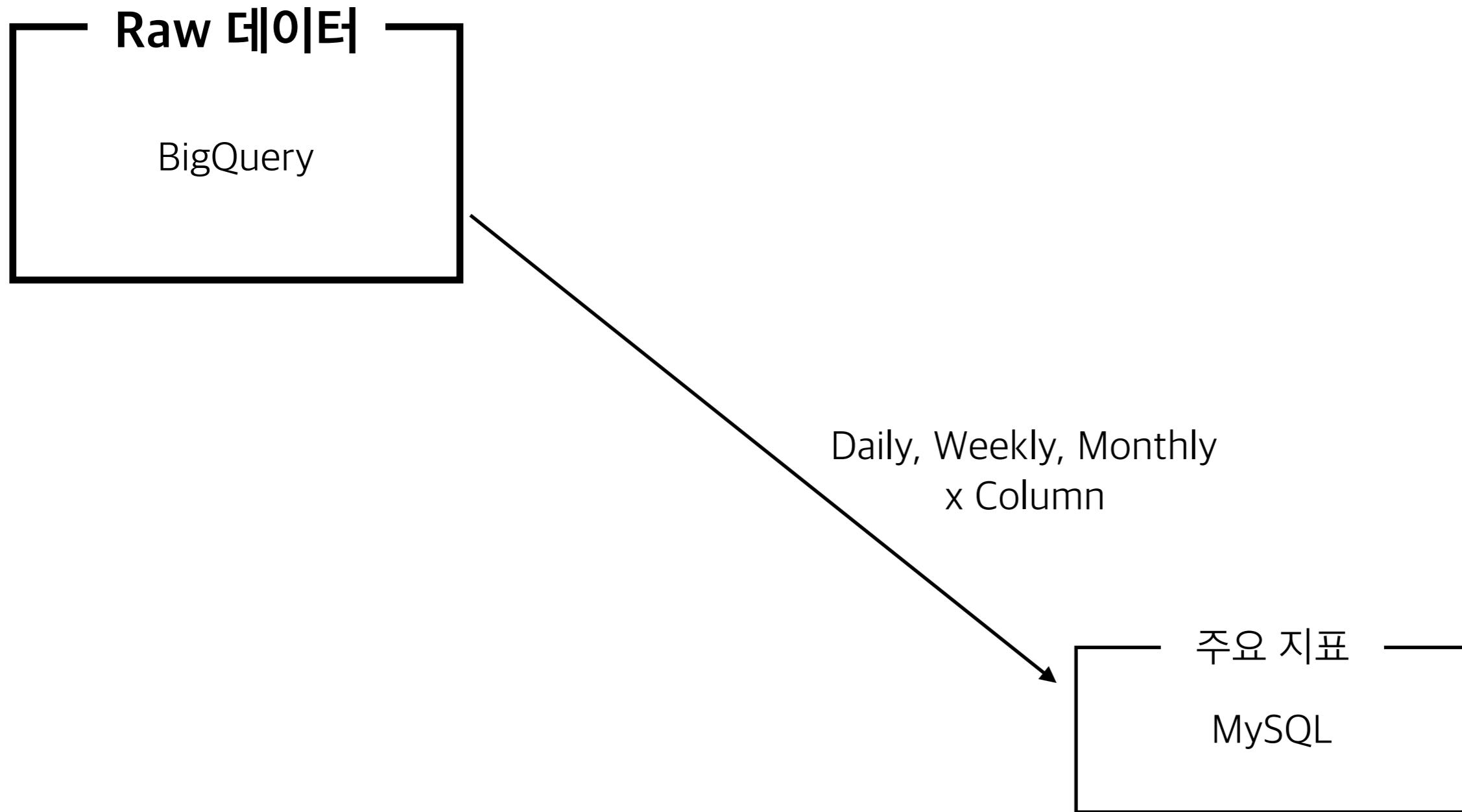
[목적에 맞도록 데이터 구성 : Cohort Retention을 위한 Dataset]

id	첫 사용일	사용일	diff_date
a	2018-03-02	2018-03-02	0
a	2018-03-02	2018-03-03	1
b	2018-01-05	2018-03-03	57
c	2018-04-02	2018-04-04	2

[Cohort Retention Query]

```
SELECT  
    first_use_date,  
    diff_date,  
    count(distinct id) -> pivot하면 Cohort  
FROM TABLE
```

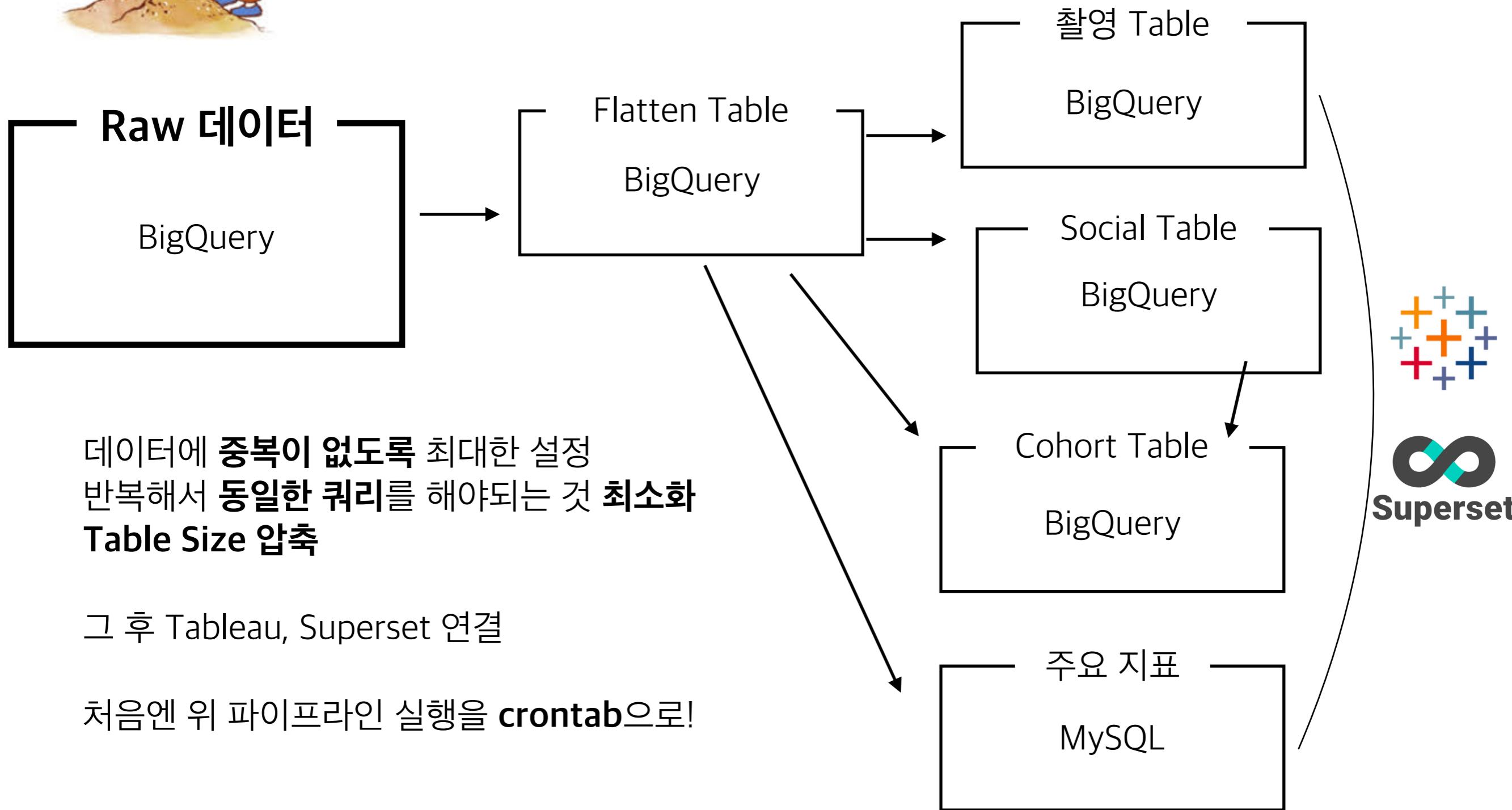
입사 당시 데이터 파이프라인



또 삽질이구나!



개선한 데이터 파이프라인



또 삽질이구나!



파이프라인 구축 방법

	Google Cloud BigQuery	Google Cloud Dataflow
과정	Query	Data Load - Transform - Write
속도	단일 과정이라 상대적으로 빠른 편	Data Load시 BigQuery Raw를 가져오는데 시간이 오래 걸림 Transform, Write는 시간이 덜 소요
비용	쿼리 비용만 부과 (\$ 5/1TB)	vCPU, RAM, 저장소 사용량에 비례해 비용 부과

<https://cloud.google.com/solutions/performing-etl-from-relational-database-into-bigquery>



파이프라인 구축 방법

**Google Cloud
BigQuery**

**Google Cloud
Dataflow**

[선택 이유]

현재 진행하려고 하는 것은 데이터를 flat하게 만들고,
특정 목적별로 이벤트를 따로 관리하도록 설계하는 것!

Dataflow는 BigQuery 데이터를 기반으로 CloudML을
이용해 학습시키는 경우 사용했습니다

또한 BigQuery의 Raw 데이터를 바로 Dataflow로 사용하기보다,
BigQuery로 한번 가공한 후, 사용하는 것이 Load 시간을
줄일 수 있어서 좋습니다

또 삽질이구나!



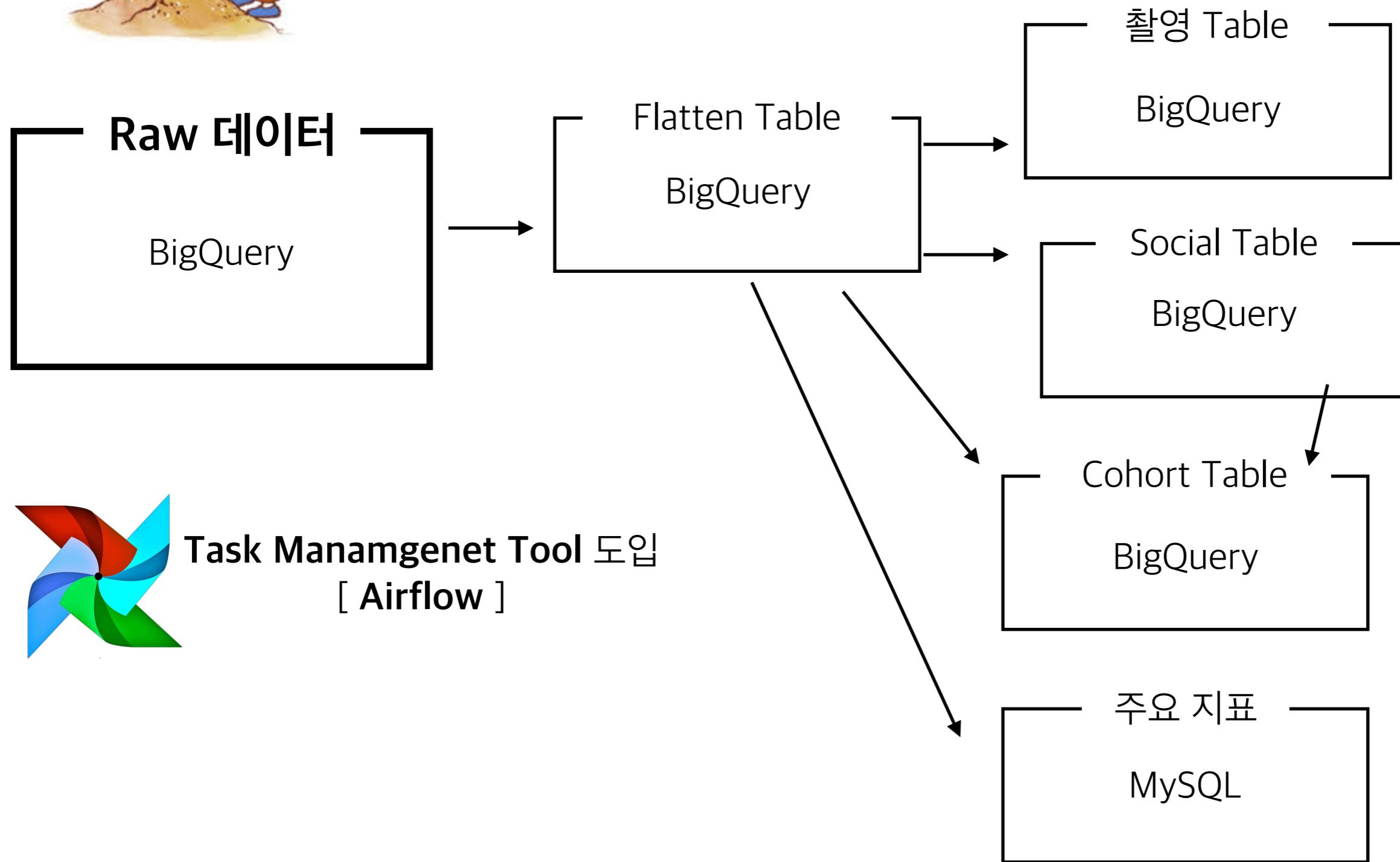
BigQuery Query Example

```
#legacySQL
SELECT
    event_dim.name AS event,
    user_dim.app_info.app_instance_id AS instance_id,
    user_dim.app_info.app_platform AS platform,
    user_dim.app_info.app_version AS app_version,
    user_dim.geo_info.country AS country,
    event_dim.timestamp_micros AS event_time,
    MAX(IF(event_dim.params.key = 'Key Name1', event_dim.params.value.string_value, NULL)) AS Key1,
    MAX(IF(event_dim.params.key = 'Key Name2', event_dim.params.value.string_value, NULL)) AS Key2
FROM TABLE
WHERE
    event_dim.name NOT IN ('user_engagement', 'session_start', 'os_update', 'firebase_campaign',
    'app_exception', 'error','notification_receive', 'notification_dismiss')
GROUP BY event, instance_id, platform, app_version, country, event_time
```

또 삽질이구나!

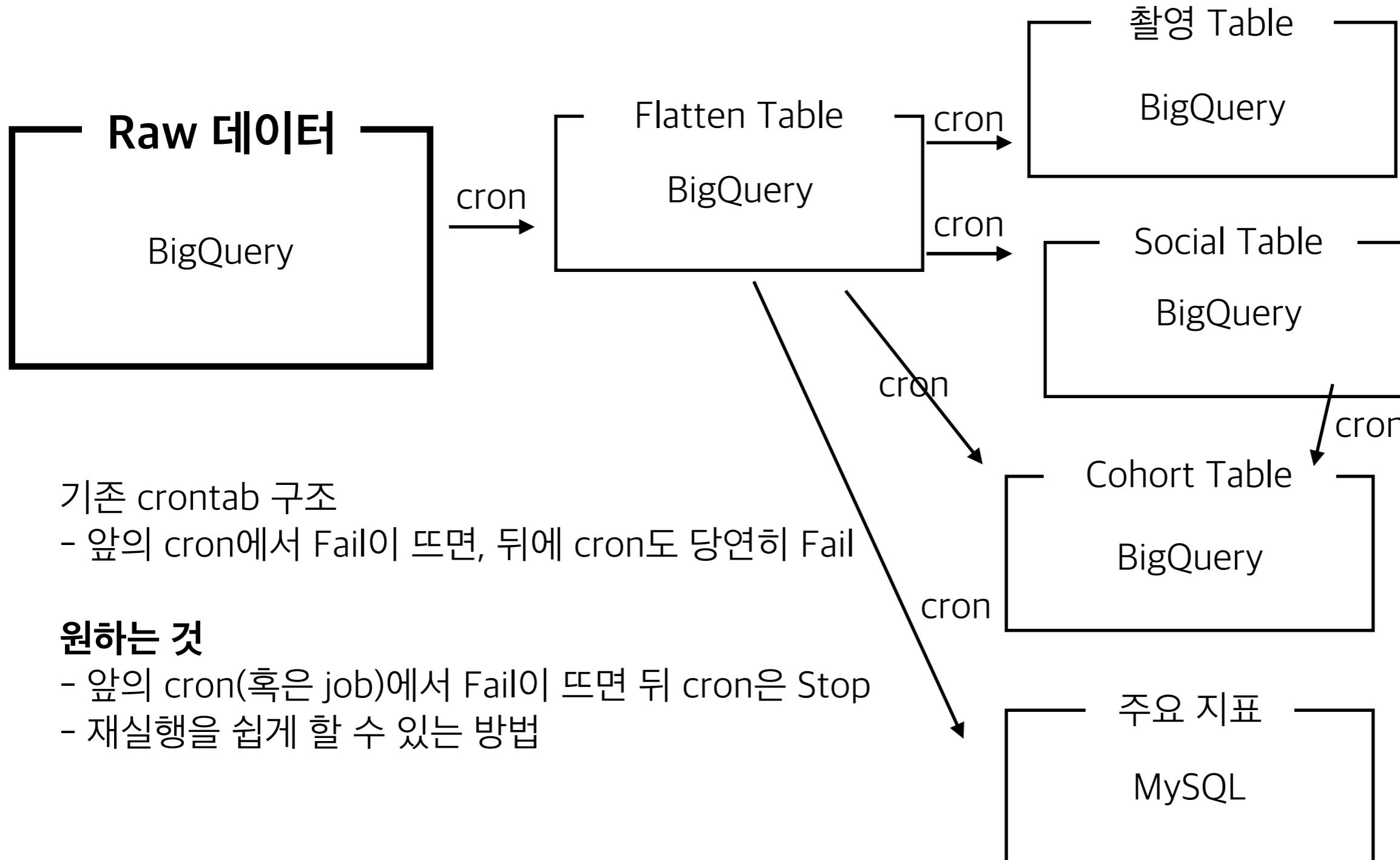


개선한 데이터 파이프라인





Task Management Tool (Airflow) 도입





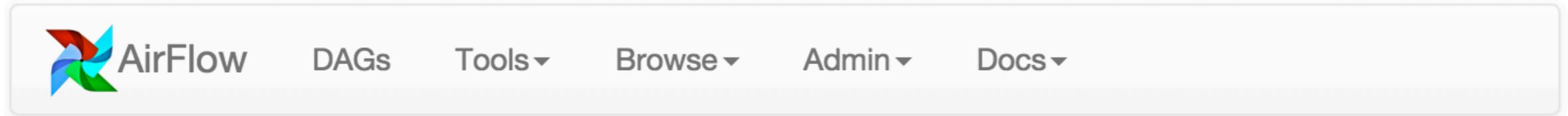
Task Management Tool (Airflow) 도입

DAGs

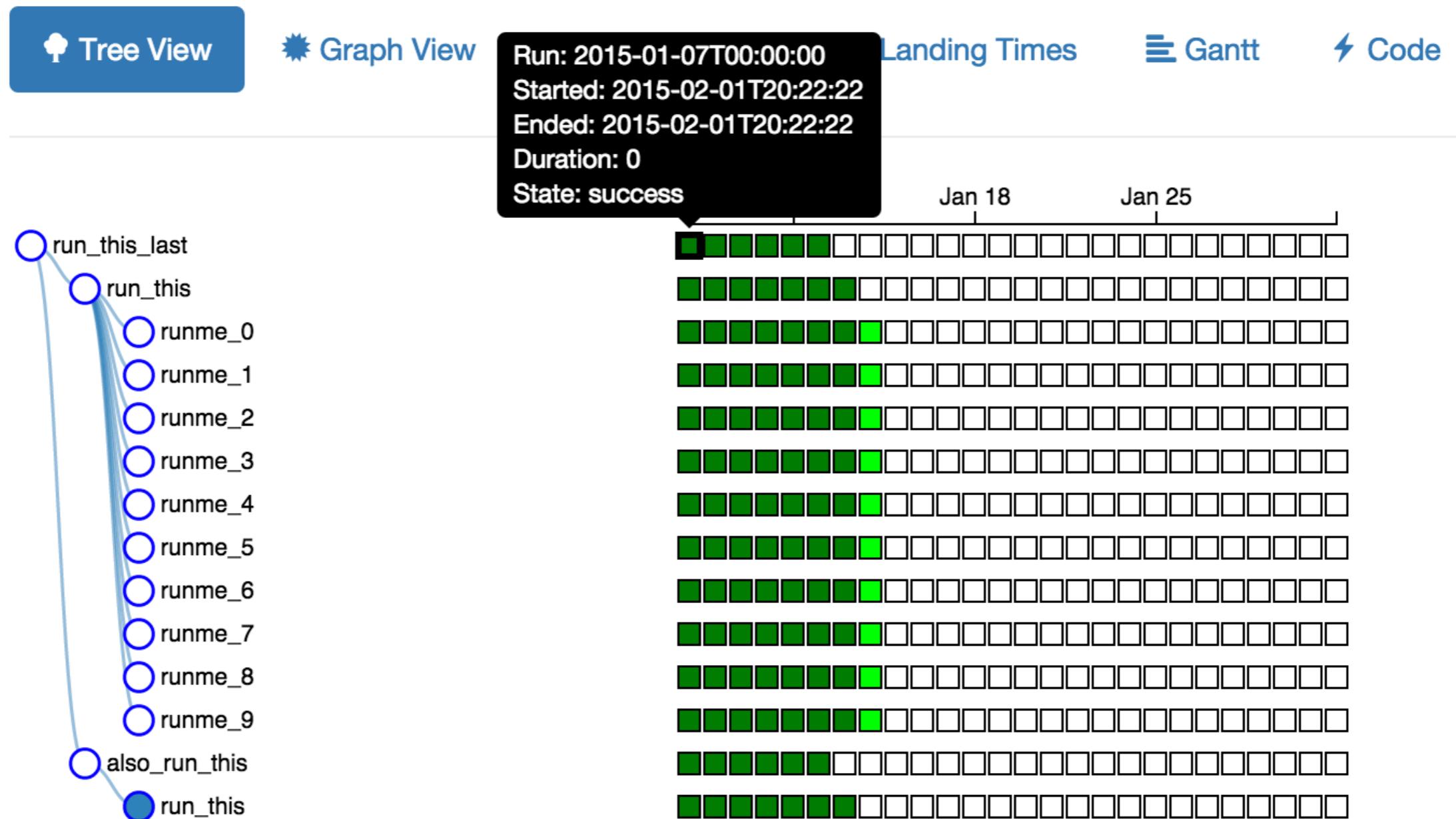
DAG	Filepath	Owner	Task by State	Links
example1	example_dags/example1.py	airflow	80 1 0	
example2	example_dags/example2.py	airflow	128 10 0	
example3	example_dags/example3.py	airflow	138 5 0	



Task Management Tool (Airflow) 도입

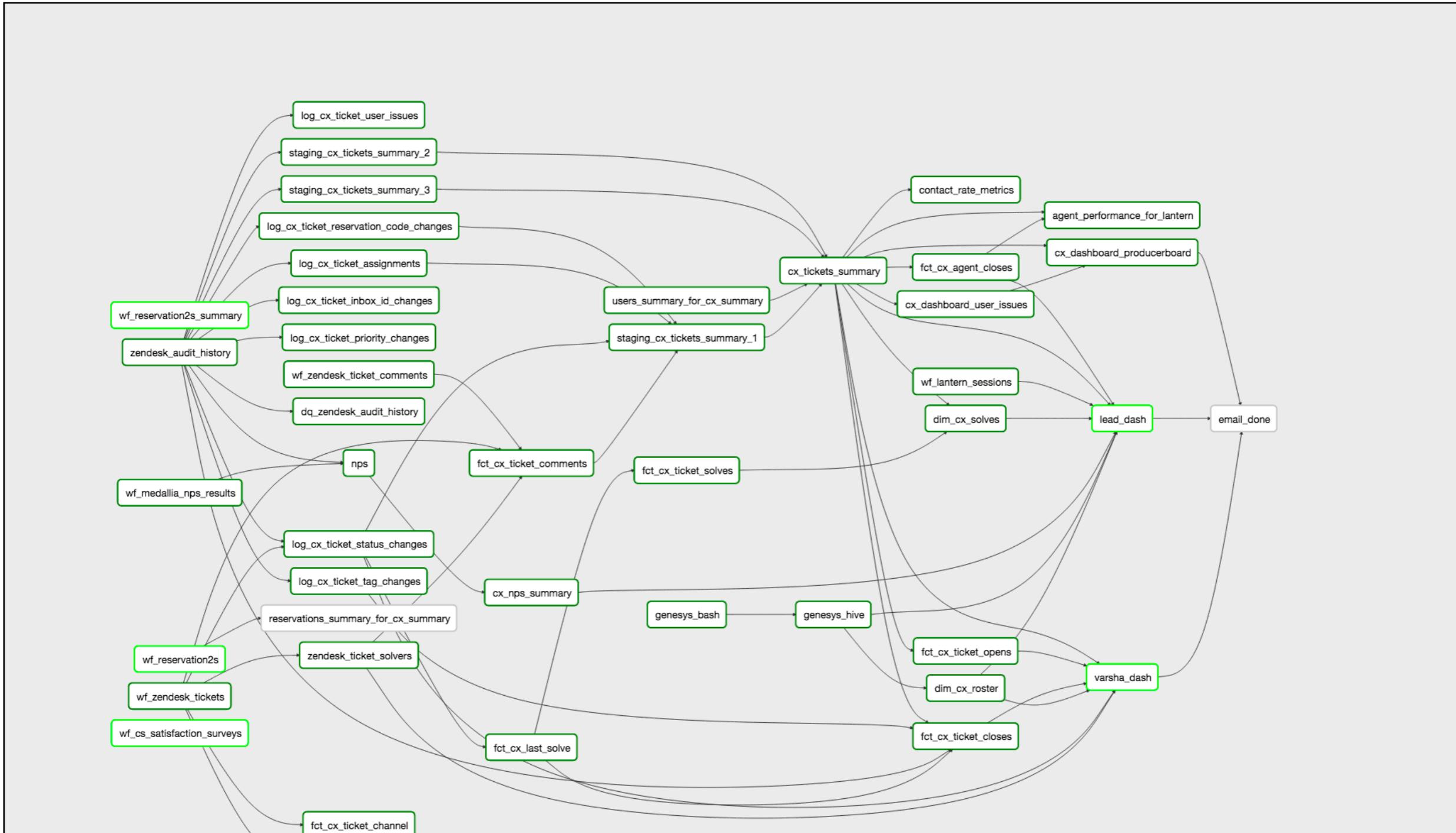


DAG: example2





Task Management Tool (Airflow) 도입

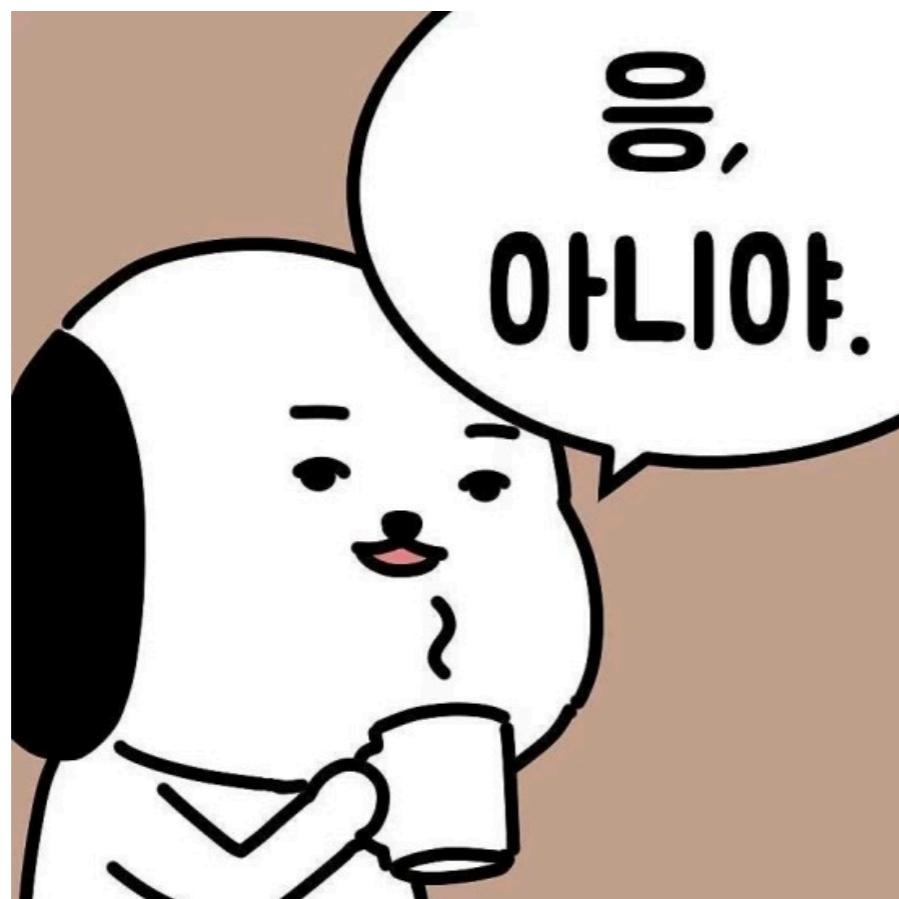


BigQuery 비용 문제

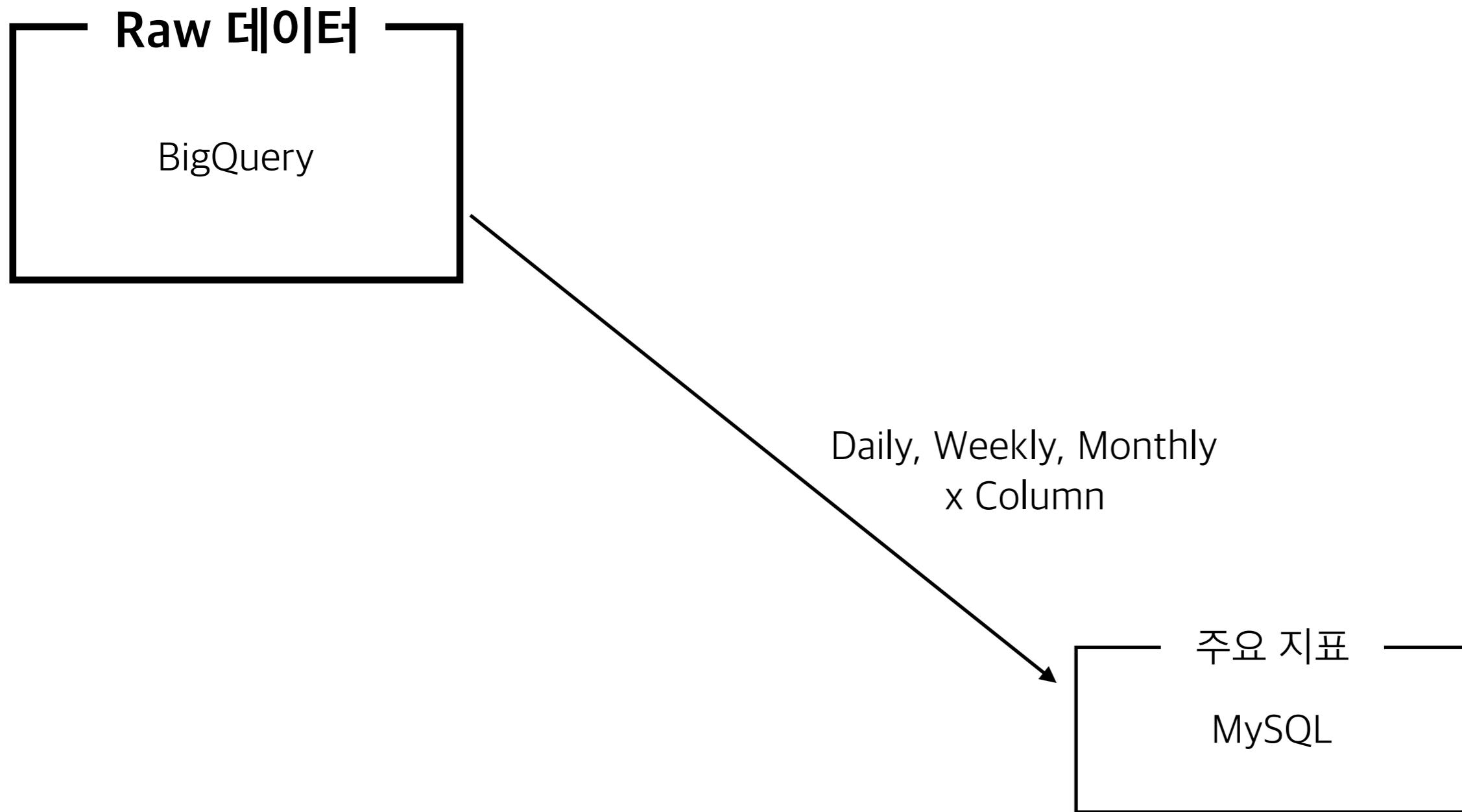
[1달에 1,000 TB 쿼리를 날릴만큼 내가 그렇게 열심히 일을 하고 있나..?]

BigQuery 비용 문제

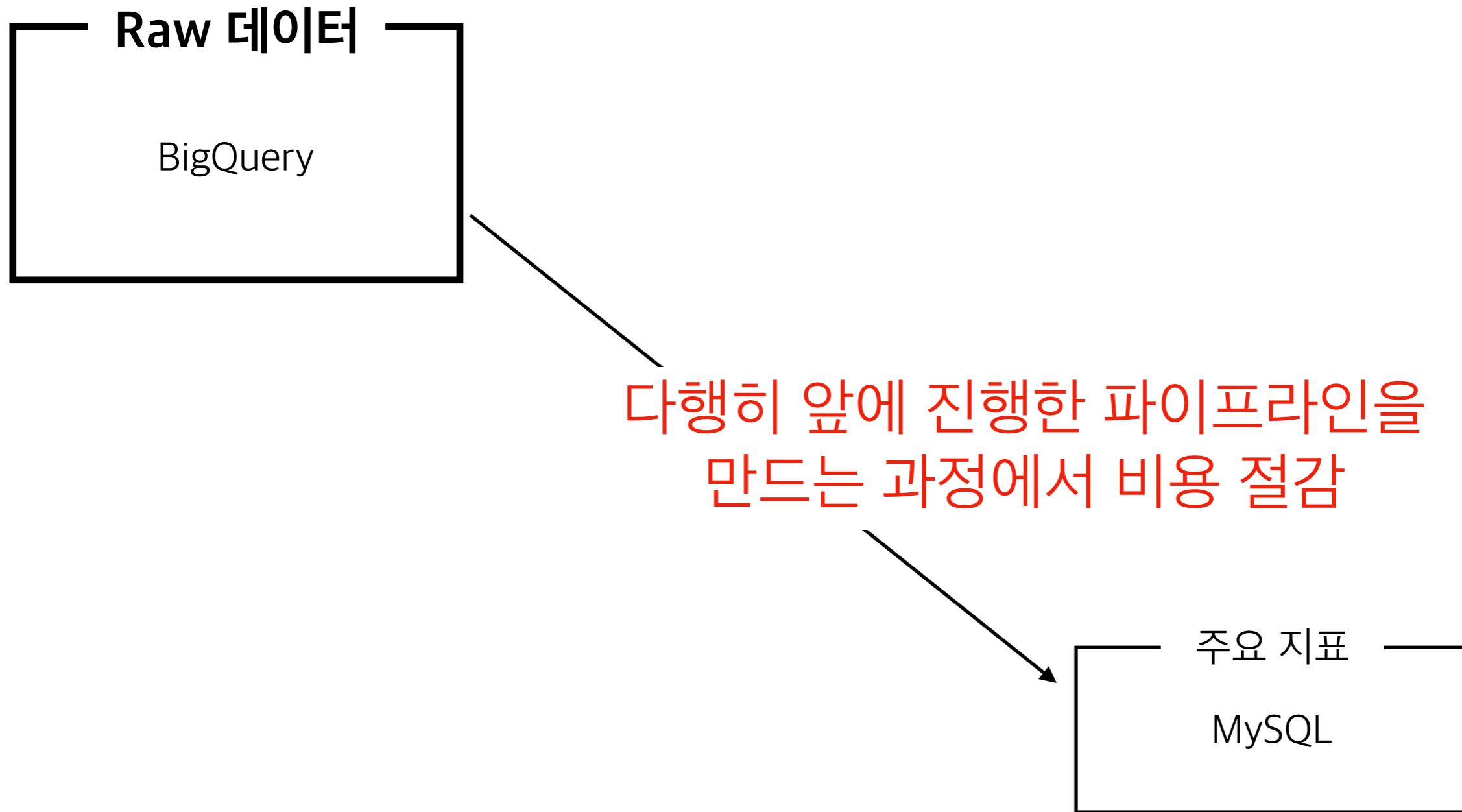
[1달에 1,000 TB 쿼리를 날릴만큼 내가 그렇게 열심히 일을 하고 있나..?]



입사 당시 데이터 파이프라인



입사 당시 데이터 파이프라인



Data Studio



Kyle Byun <kyle@retrica.co>

Terry에게 ▾

병욱님 안녕하세요!
레트리카 변성윤입니다

17. 6. 8. ☆



데이터 스튜디오에서 새로고침을 하니 어제 보내드린 쿼리 로그가 찍히는 것을 발견했습니다.

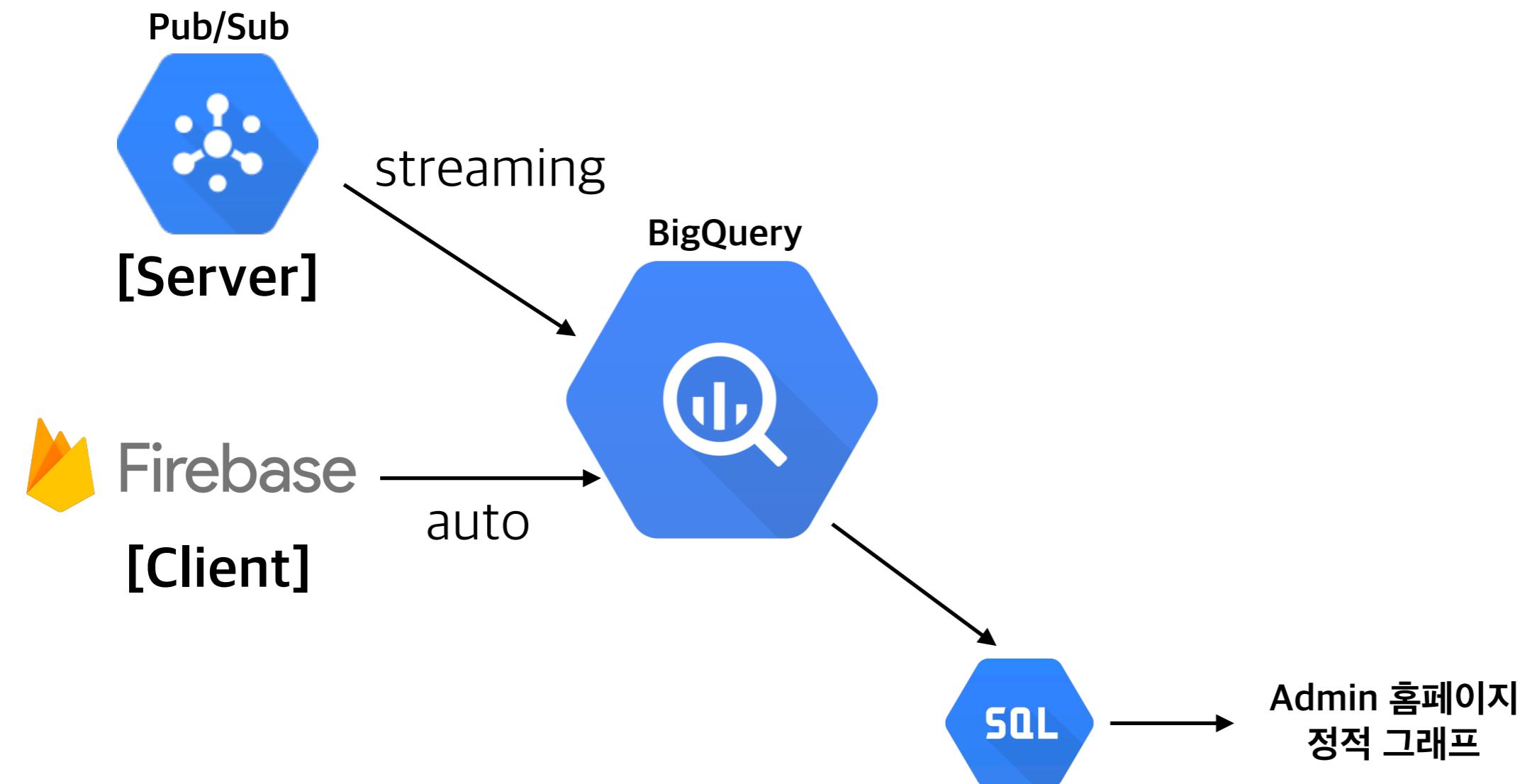
그래서 데이터 스튜디오의 레포트와 데이터 소스를 모두 제거하니, 해당 시간대에 돌던 쿼리가 발생하지 않았습니다

아마 제 생각으론 데이터 스튜디오에서 레포트를 만들어두면, 데이터 스튜디오를 키지 않아도 자동으로 빅쿼리에서 데이터를 가지고 가는 것 같습니다

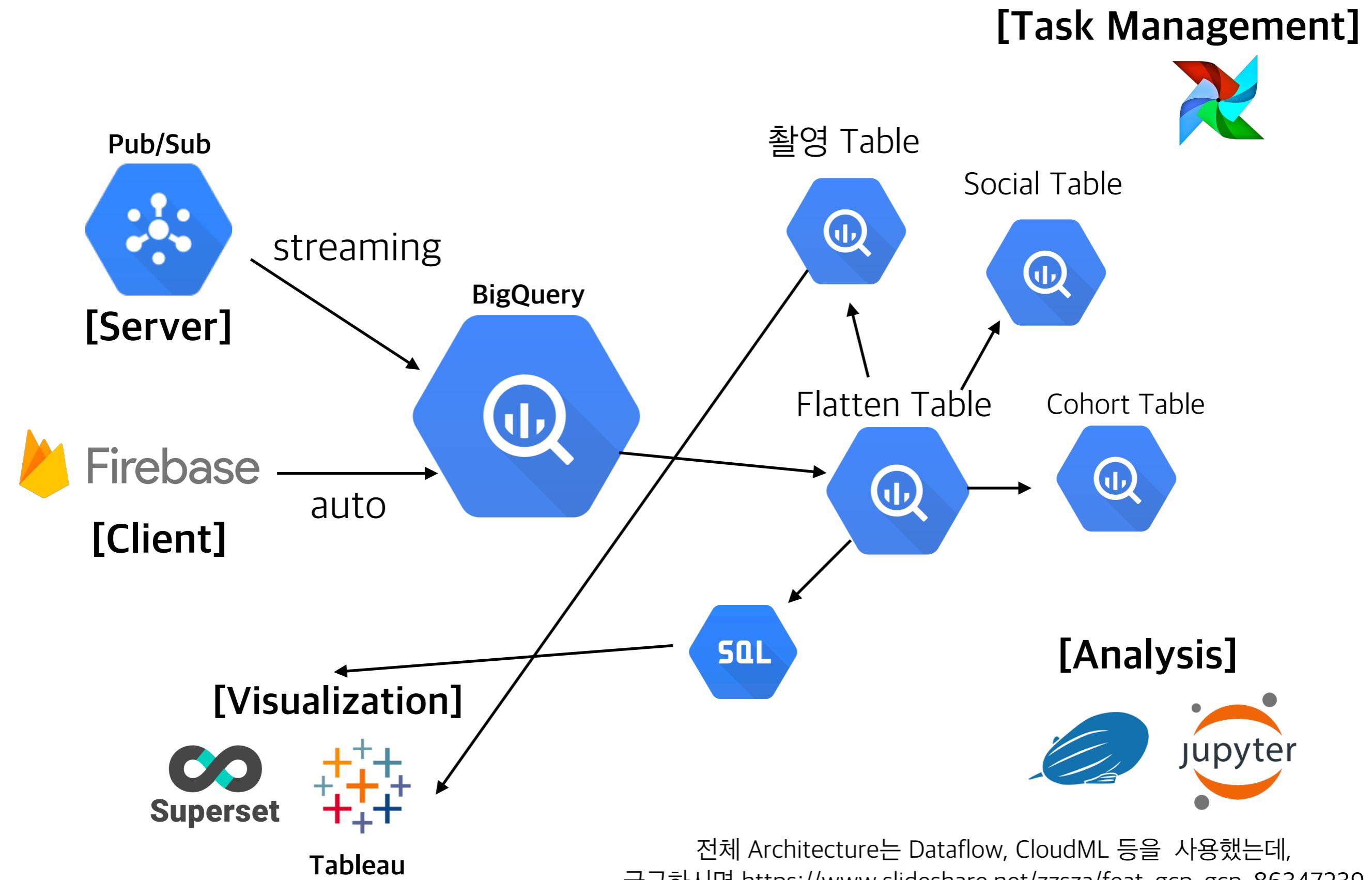
구글 데이터 스튜디오 문서에 자동으로 refresh된다는 글을 못봤는데, 주기를 설정할 수 있는 옵션을 만들거나 주기적으로 빅쿼리에서 데이터를 가지고 온다고 명시해야 좋을 것 같습니다-!!!

Dashboard를 만들기 위해 테스트했던 파일이
자동으로 모든 데이터를 refresh (하루 2번)
(구글님들 문서 좀 친절하게 만들어주세요...)

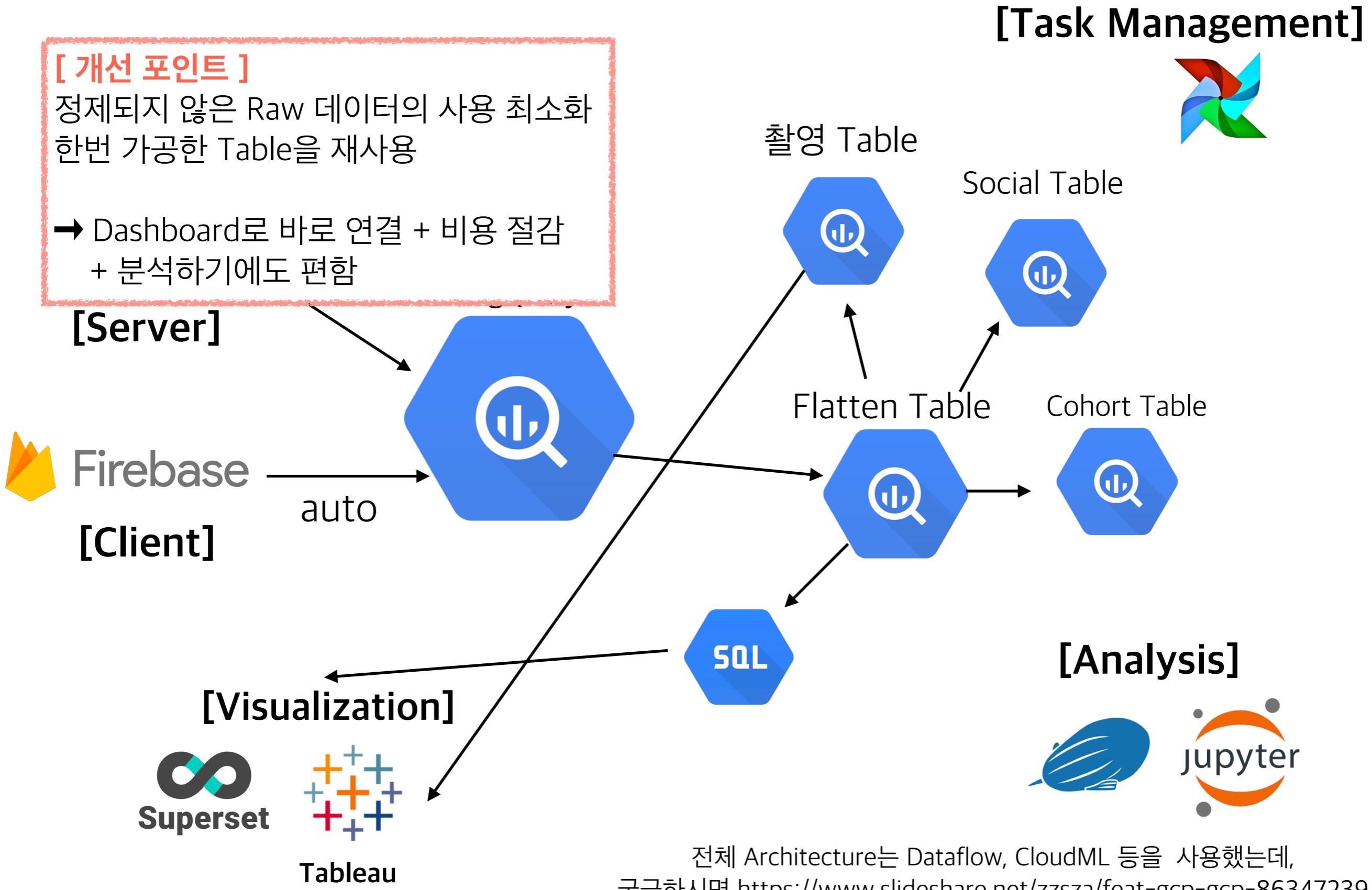
입사 당시 데이터 파이프라인



개선한 데이터 파이프라인



개선한 Data 파이프라인



Story 2. Summary

문제 상황

이벤트 Dashboard 필요
BigQuery 비용이 이상함

해결 방안

Dashboard 생성을 위해 BigQuery Data 정규화
위 과정에서 파이프라인 생성
비용의 누수를 찾아 삭제

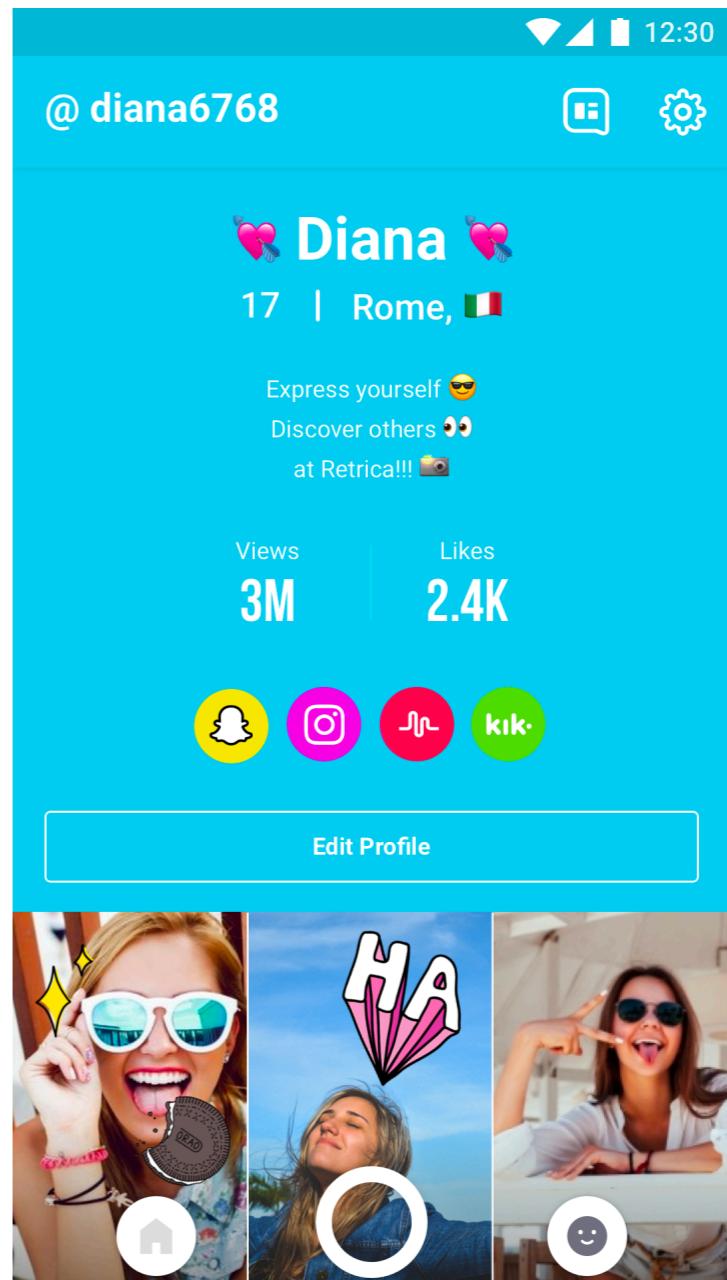
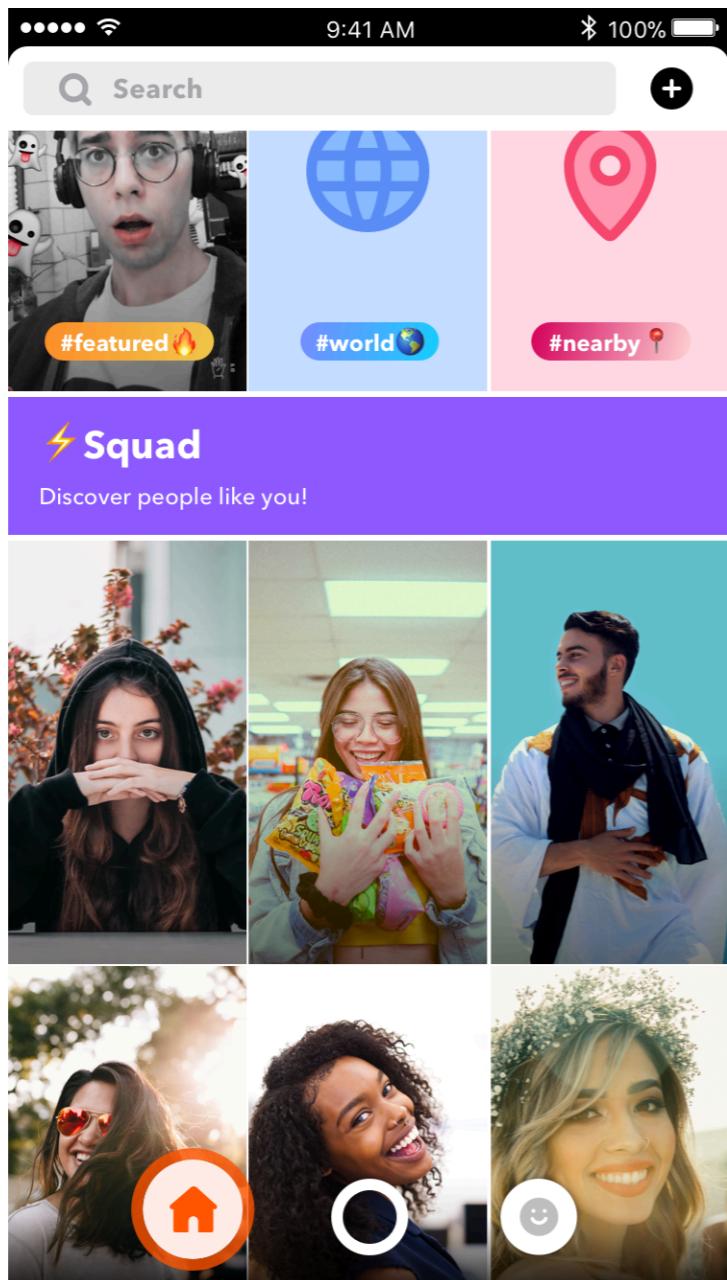
의의

이벤트 Dashboard 생성
BigQuery 파이프라인 생성
필요한 데이터만 자세히 볼 수 있음
Task Management Tool 도입
불필요하게 돌고 있던 Data Studio 삭제
비용 : \$5,000 이상 → \$1,000 이하

Story 3. 음란 사진이 올라온다

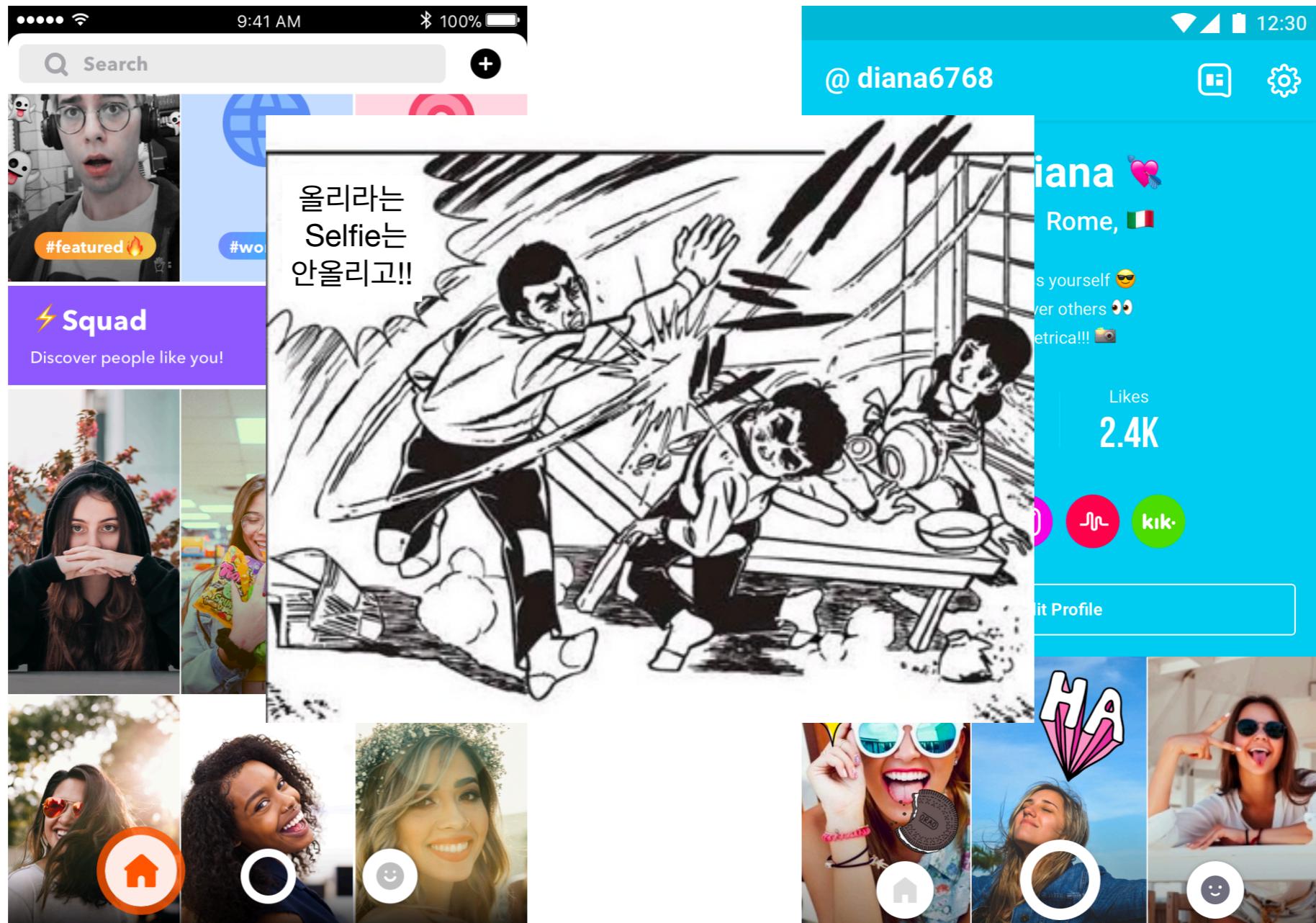
레트리카

[Social]

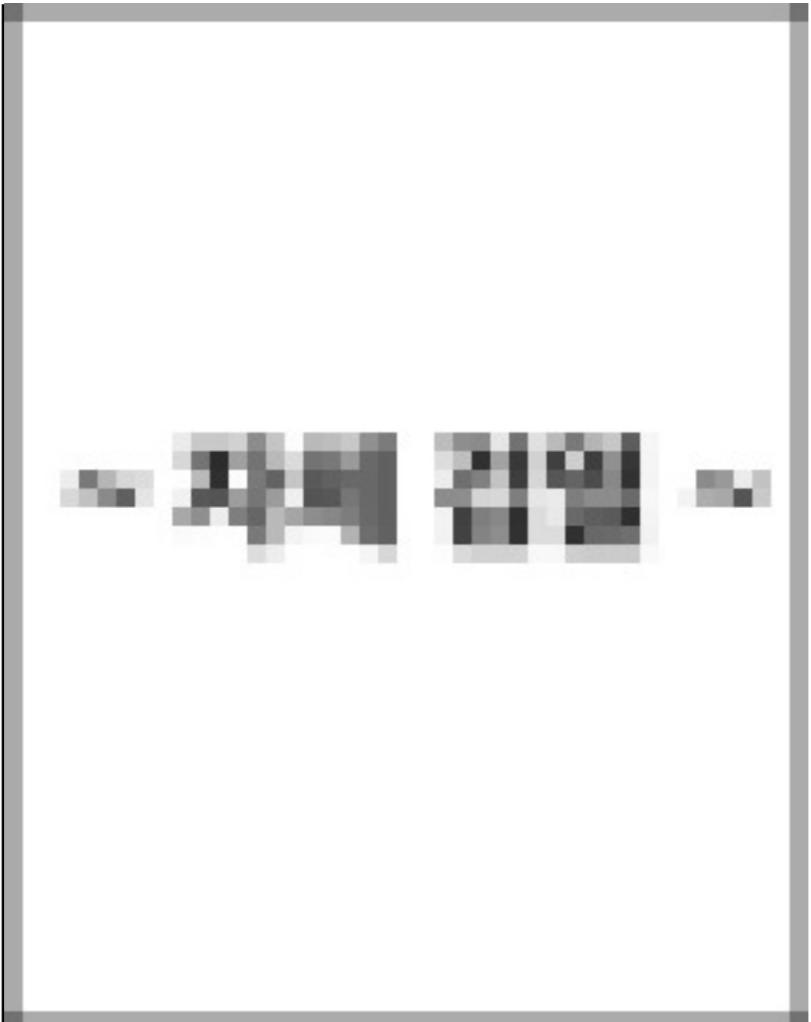


레트리카

[Selfie만 올려줘…]



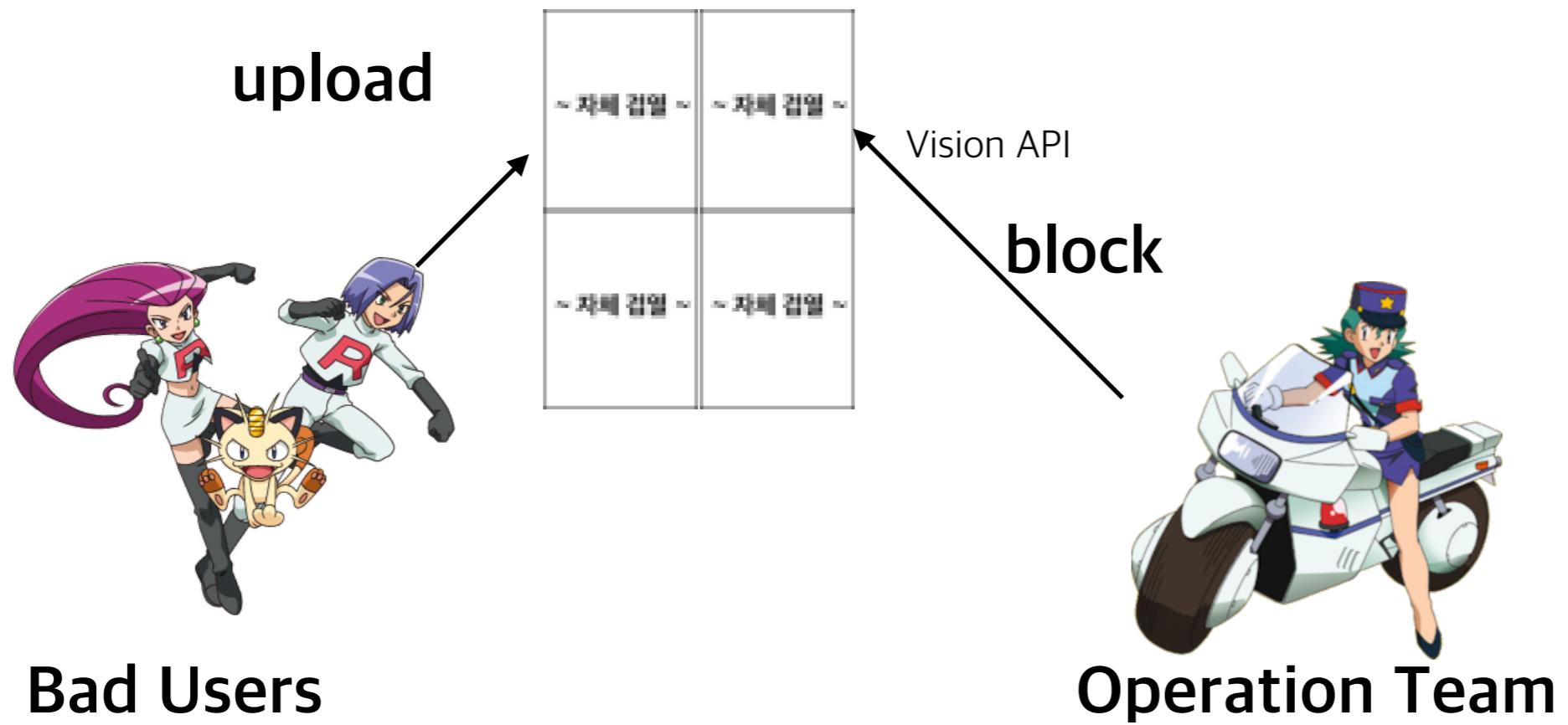
올라오는 사진들



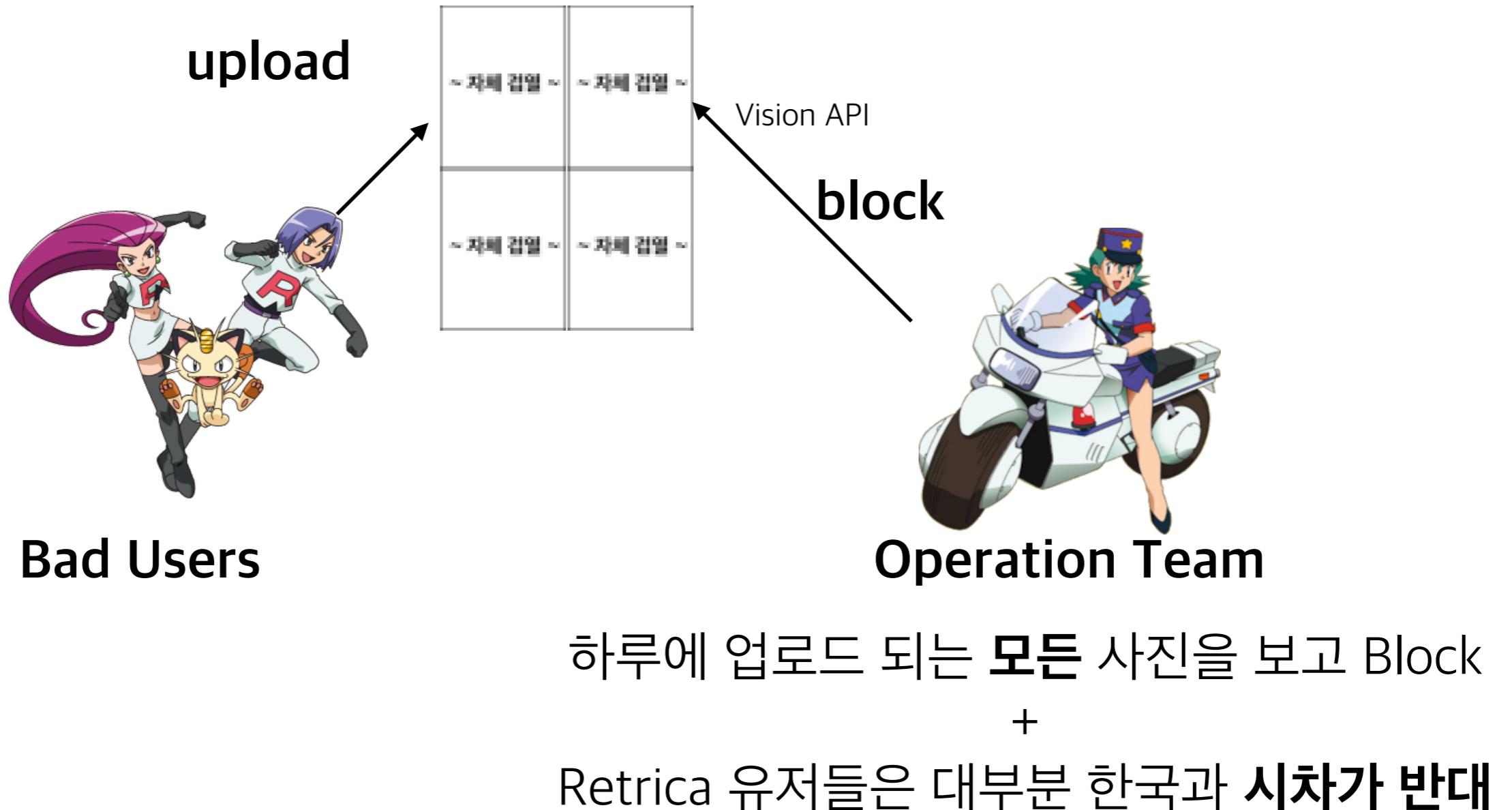
남자 알몸,
남자 성기,
여자 가슴,
여자 성기,
성행위 자세,
야한 짤,
등등..

여러분들의 눈을 위해
자체 검열했습니다

Operation Team의 고생



Operation Team의 고생



대표님이 말하신다



넵!

야한 사진을 검출할 수
있는 모델을 만들어보죠!

캡틴

대표님이 말하신다



넵!

야한 사진을 검출할 수
있는 모델을 만들어보죠!

이번엔 1주일?

캡틴

대표님이 말하신다



넵!

(..) 네~!

야한 사진을 검출할 수
있는 모델을 만들어보죠!

이번엔 1주일?

캡틴

삽질의 연속

[1주안에..?]

문제 정의 → 데이터 수집 및 전처리 → 모델 선정 및 학습 → 성능 개선

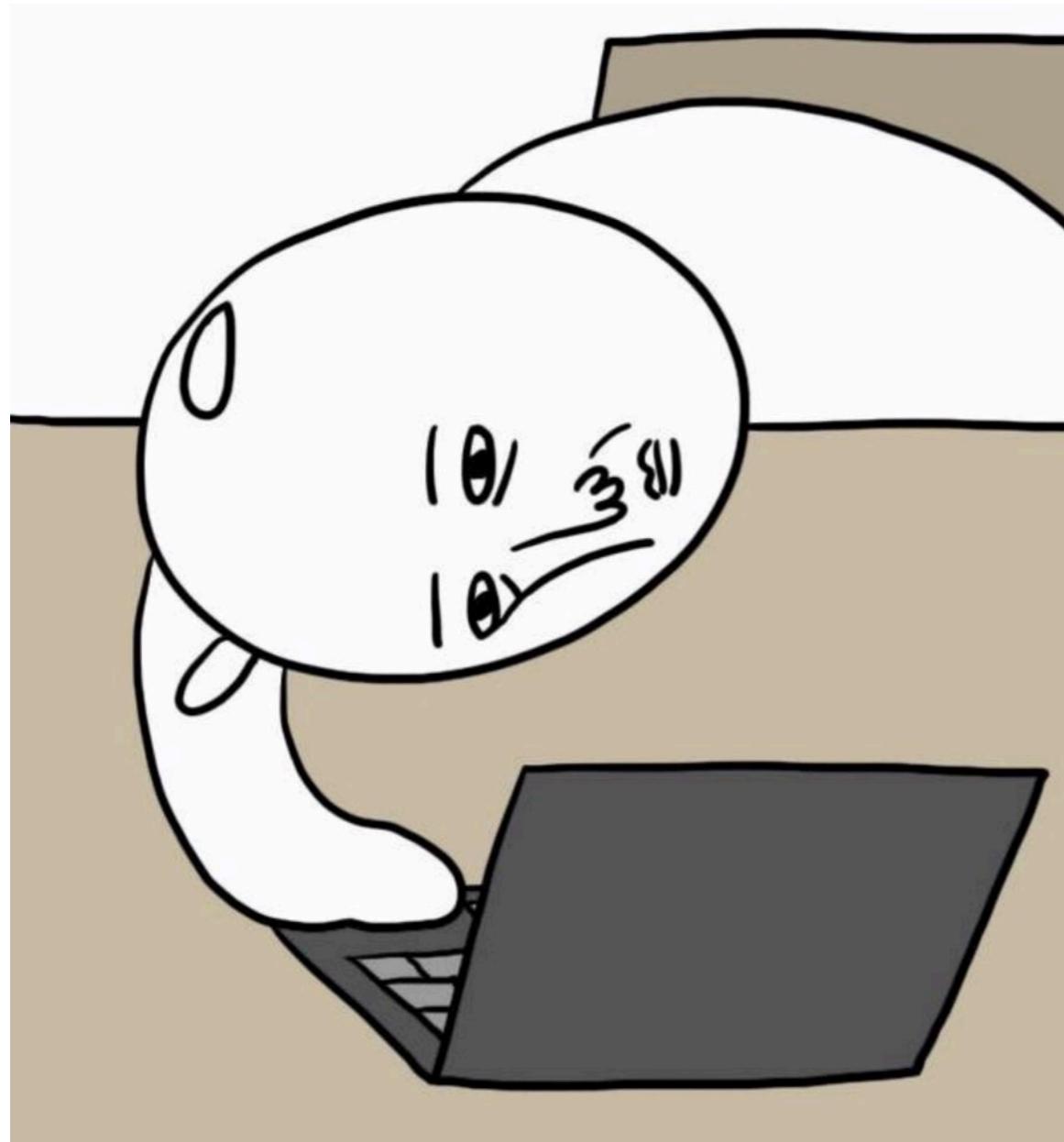
레트리카에서 발생한 야한 사진으로
학습에 터무니 없는 양

… 크롤링 시도
구글 / 인스타 / 야후 / Reddit

삽질의 연속

[뒷자리 직원분들이 뭐 보냐고 수근거림]

문제 정의 → 데이터 수집 및 전처리 → 모델 선정 및 학습 → 성능 개선



내가 왜 이걸.. 하고.. 있지
심지어 생각보다 인터넷엔
좋은 데이터셋이 없었다.

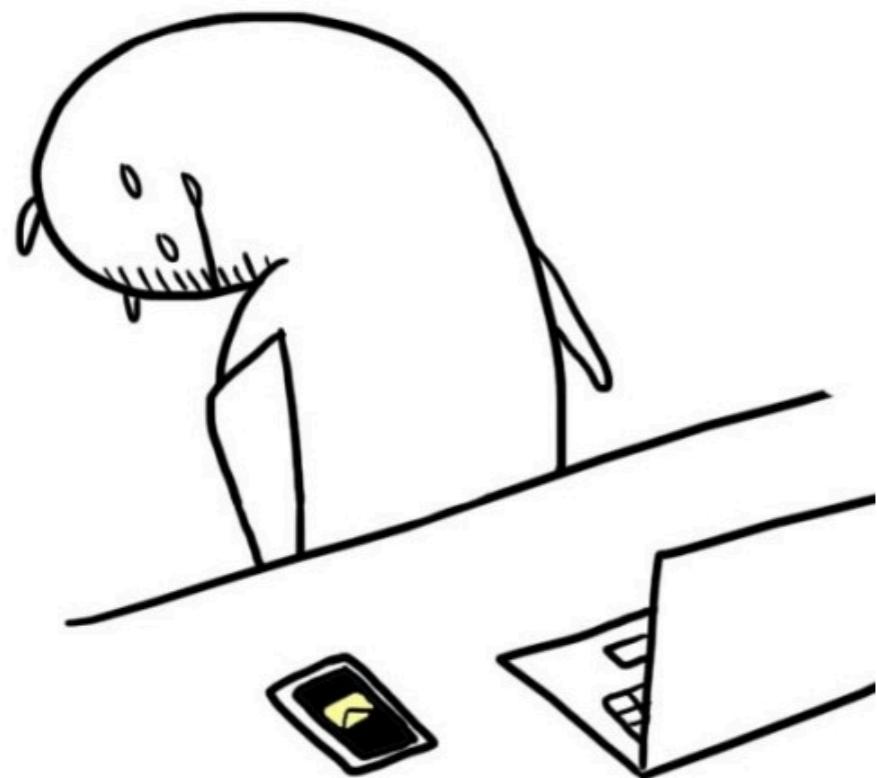
삽질의 연속

【 하지만 불가능은 없다 】

문제 정의 → 데이터 수집 및 전처리 → 모델 선정 및 학습 → 성능 개선



다 했다....



그래도 열심히 모아서 라벨별 1,000개
총 4개의 라벨 = 4,000

빠르게 구현 가능한 vgg16 모델링

삽질의 연속

[내가 잘못했나]

문제 정의 → 데이터 수집 및 전처리 → 모델 선정 및 학습 → 성능 개선



약 67%의 accuracy
그러나 무언가 잘못되었다..
Residual Network로 가보자..



삽질의 연속

[살려주세요]

문제 정의 → 데이터 수집 및 전처리 → 모델 선정 및 학습 → 성능 개선
[Residual Network 역시 이상해… => 데이터가 문제]



삽질의 연속

[대안 찾아 삼만리]

문제 정의 → 데이터 수집 및 전처리 → 모델 선정 및 학습 → 성능 개선



pretrain Model을 찾아봄

https://github.com/yahoo/open_nsfw

input : image
output : nsfw score(0~1)

Yahoo NSFW

This repo contains code for running Not Suitable for Work (NSFW) classification deep neural network Caffe models. Please refer our [blog](#) post which describes this work and experiments in more detail.

Not suitable for work classifier

Detecting offensive / adult images is an important problem which researchers have tackled for decades. With the evolution of computer vision and deep learning the algorithms have matured and we are now able to classify an image as not suitable for work with greater precision.

Defining NSFW material is subjective and the task of identifying these images is non-trivial. Moreover, what may be objectionable in one context can be suitable in another. For this reason, the model we describe below focuses only on one type of NSFW content: pornographic images. The identification of NSFW sketches, cartoons, text, images of graphic violence, or other types of unsuitable content is not addressed with this model.

Since images and user generated content dominate the internet today, filtering nudity and other not suitable for work images becomes an important problem. In this repository we opensource a Caffe deep neural network for preliminary filtering of NSFW images.



Usage

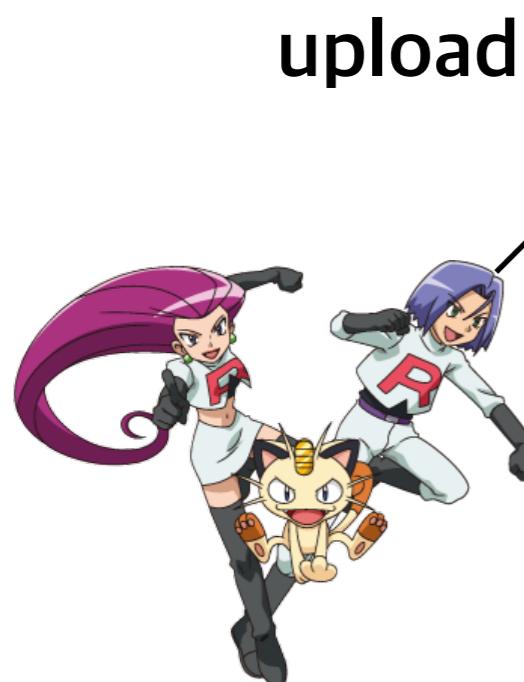
- The network takes in an image and gives output a probability (score between 0-1) which can be used to filter not suitable for work images. Scores < 0.2 indicate that the image is likely to be safe with high probability. Scores > 0.8 indicate that the image is highly probable to be NSFW. Scores in middle range may be binned for different NSFW levels.
- Depending on the dataset, usecase and types of images, we advise developers to choose suitable thresholds. Due to difficult nature of problem, there will be errors, which depend on use-cases / definition / tolerance of NSFW. Ideally developers should create an evaluation set according to the definition of what is safe for their application, then fit a [ROC](#) curve to choose a suitable threshold if they are using the model as it is.
- **Results can be improved by [fine-tuning](#)** the model for your dataset/ usecase / definition of NSFW. We do not provide any guarantees of accuracy of results. Please read the disclaimer below.
- Using human moderation for edge cases in combination with the machine learned solution will help improve performance.

제품 적용

문제 정의 → 데이터 수집 및 전처리 → 모델 선정 및 학습 → 성능 개선

Yahoo NSFW 모델을 Product에 적용
docker 이미지로 말아 API화

[기존 로직]



Vision API

block



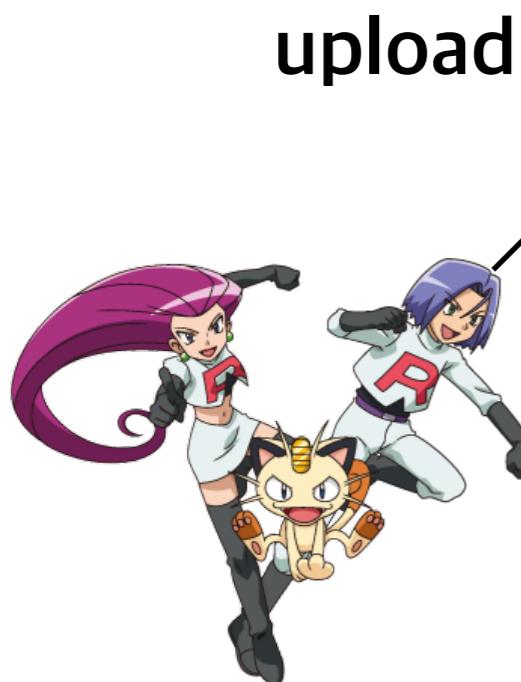
Bad Users

제품 적용

문제 정의 → 데이터 수집 및 전처리 → 모델 선정 및 학습 → 성능 개선

Yahoo NSFW 모델을 Product에 적용
docker 이미지로 말아 API화

[**개선** 로직]



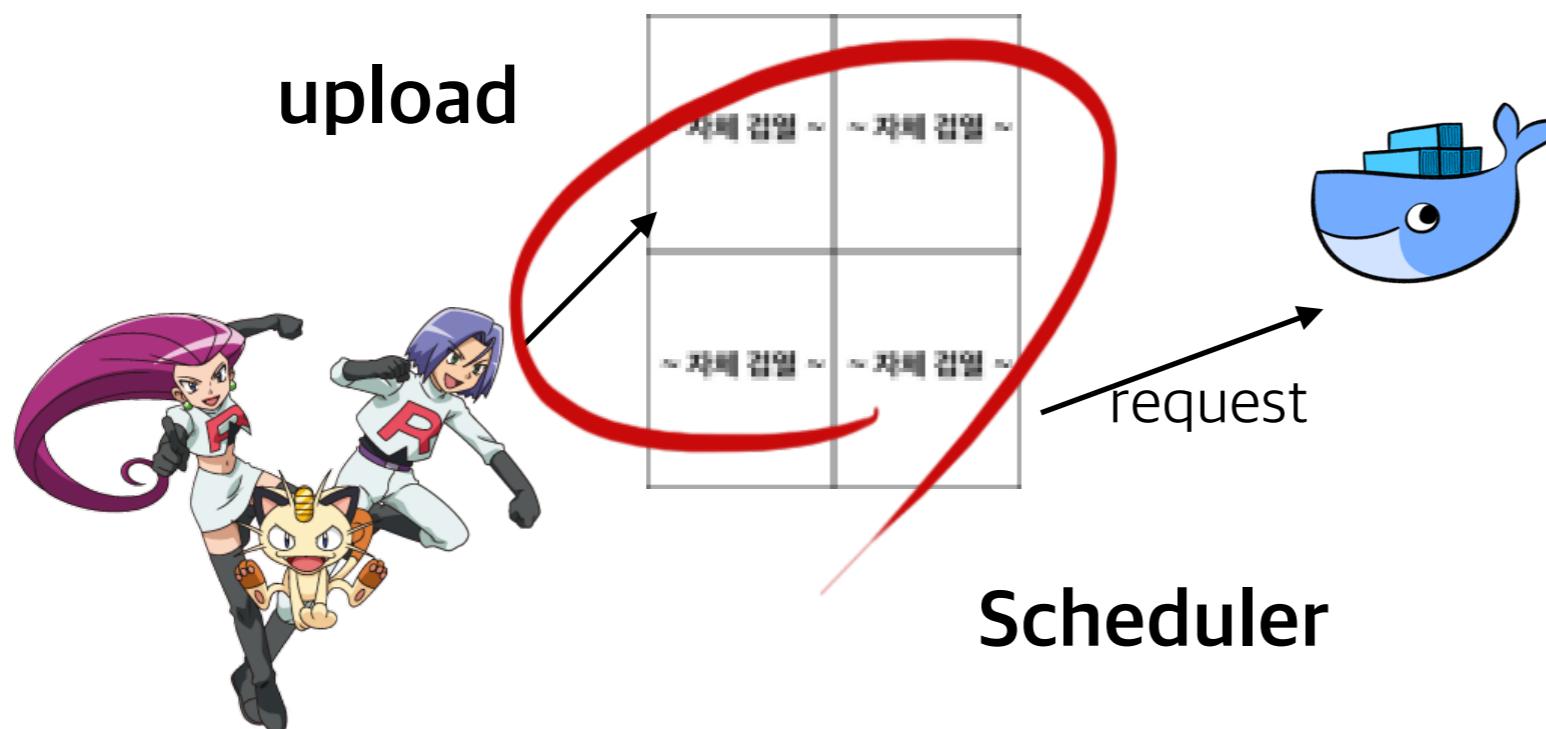
Bad Users

제품 적용

문제 정의 → 데이터 수집 및 전처리 → 모델 선정 및 학습 → 성능 개선

Yahoo NSFW 모델을 Product에 적용
docker 이미지로 말아 API화

[**개선** 로직]



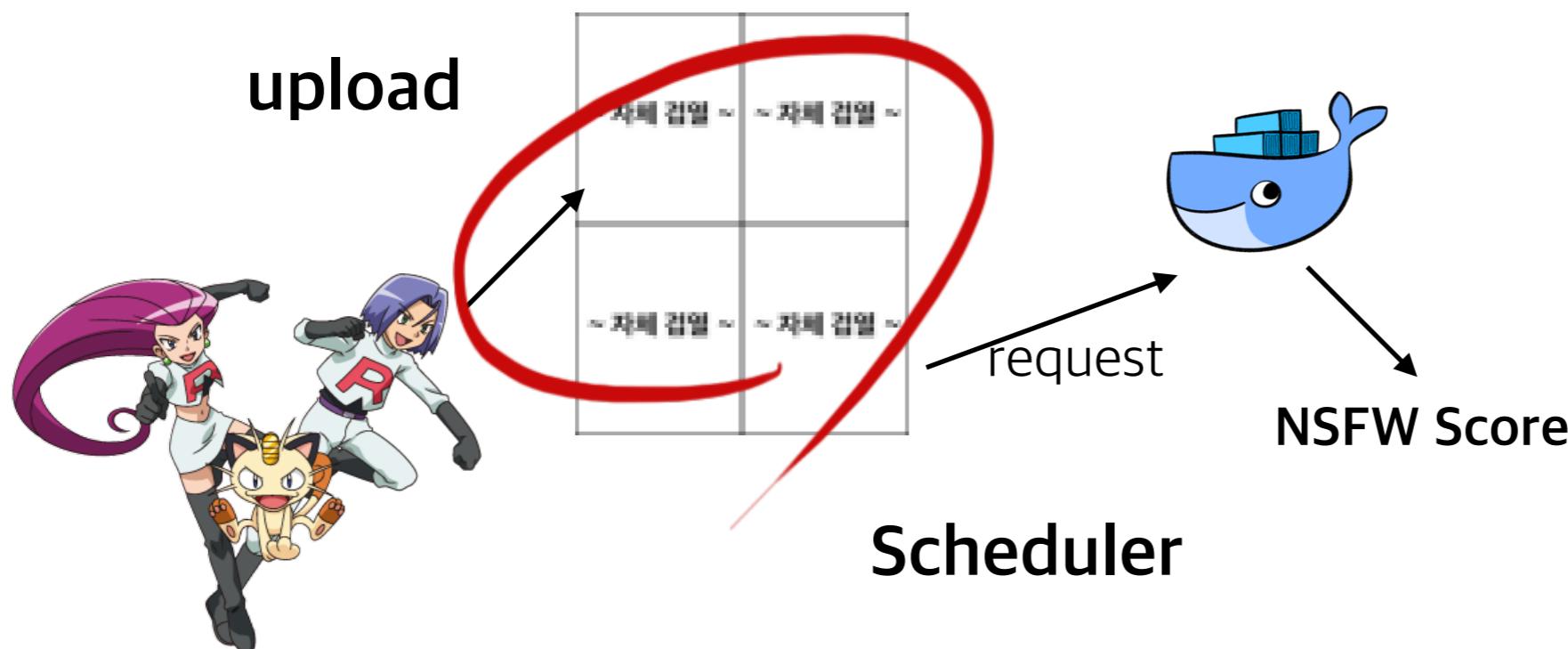
Bad Users

제품 적용

문제 정의 → 데이터 수집 및 전처리 → 모델 선정 및 학습 → 성능 개선

Yahoo NSFW 모델을 Product에 적용
docker 이미지로 말아 API화

[**개선** 로직]

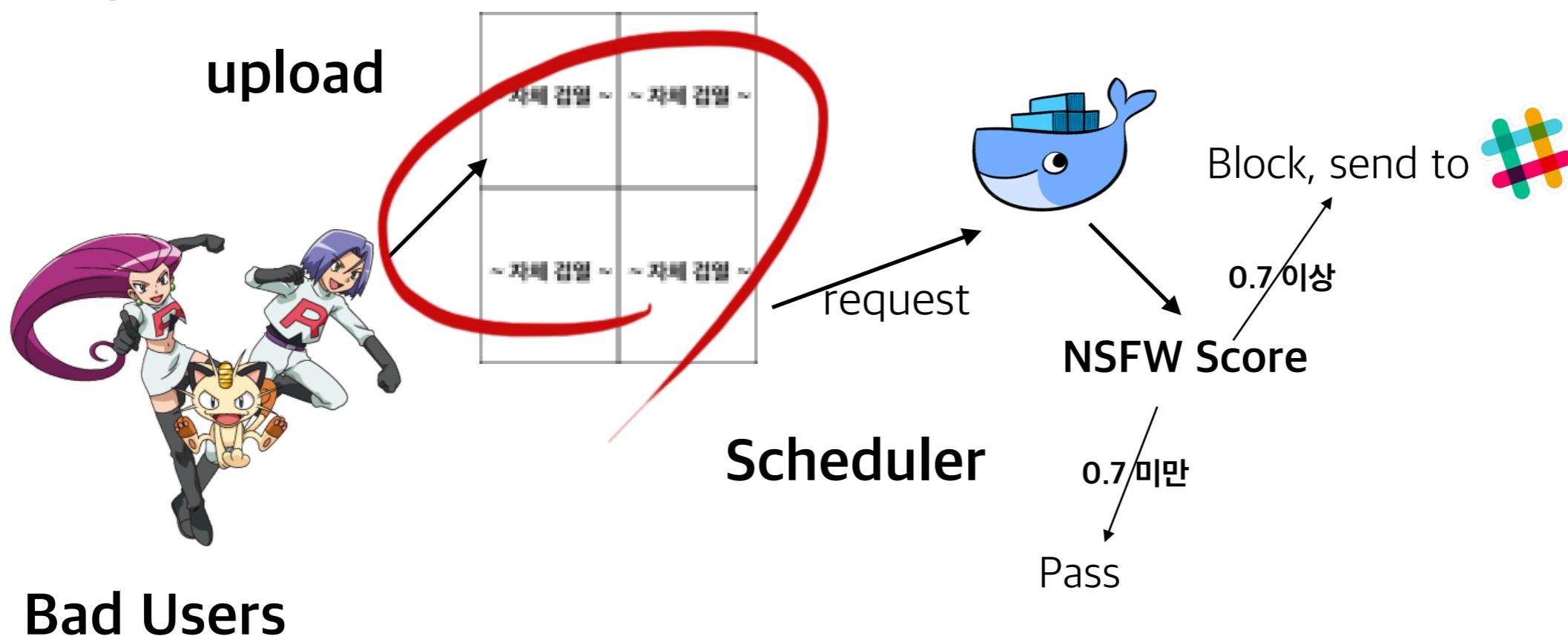


제품 적용

문제 정의 → 데이터 수집 및 전처리 → 모델 선정 및 학습 → 성능 개선

Yahoo NSFW 모델을 Product에 적용
docker 이미지로 말아 API화

[**개선** 로직]



제품 적용

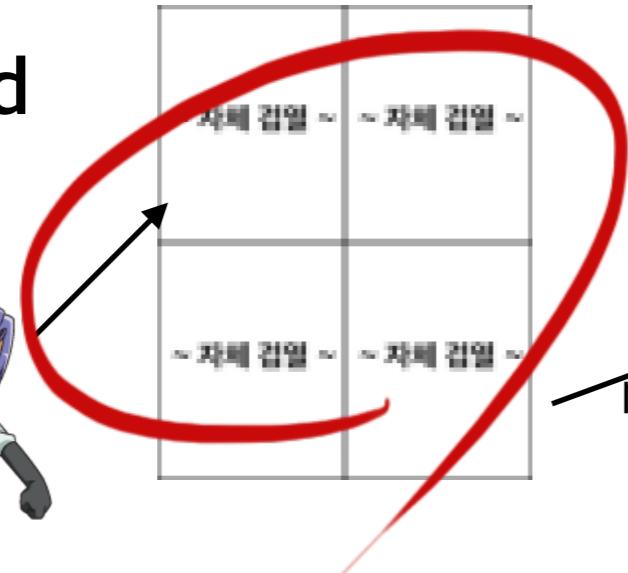
문제 정의 → 데이터 수집 및 전처리 → 모델 선정 및 학습 → 성능 개선

Yahoo NSFW 모델을 Product에 적용
docker 이미지로 말아 API화



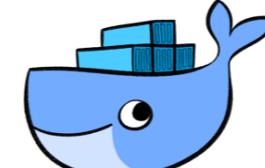
[**개선** 로직]

upload



Scheduler

request



NSFW Score

0.7 이상

Pass

0.7 미만

Block, send to



Bad Users

Operation Team



문제점

[Case #1. 현재 모델은 "이미지"만 처리 가능]

문제 정의 → 데이터 수집 및 전처리 → 모델 선정 및 학습 → 성능 개선



문제점

[Case #1. 현재 모델은 "이미지"만 처리 가능]

문제 정의 → 데이터 수집 및 전처리 → 모델 선정 및 학습 → 성능 개선

Gif 또는 Video Type일 경우 **OpenCV**로
첫 프레임을 잘라서 그 프레임을 Input으로!

문제점

[Case #1. 현재 모델은 "이미지"만 처리 가능]

문제 정의 → 데이터 수집 및 전처리 → 모델 선정 및 학습 → 성능 개선

Gif 또는 Video Type일 경우 **OpenCV**로
첫 프레임을 잘라서 그 프레임을 Input으로!

단순히 자르기 위해서 OpenCV를 빌드하기엔 부담스러움

문제점

[Case #1. 현재 모델은 "이미지"만 처리 가능]

문제 정의 → 데이터 수집 및 전처리 → 모델 선정 및 학습 → 성능 개선

Gif 또는 Video Type일 경우 **OpenCV**로
첫 프레임을 잘라서 그 프레임을 Input으로!

단순히 자르기 위해서 OpenCV를 빌드하기엔 부담스러움

Client에서 Server로 이미지 업로드시, **Webp** 형식 사용

문제점

[Case #1. 현재 모델은 "이미지"만 처리 가능]

문제 정의 → 데이터 수집 및 전처리 → 모델 선정 및 학습 → 성능 개선

Gif 또는 Video Type일 경우 **OpenCV**로
첫 프레임을 잘라서 그 프레임을 Input으로!

단순히 자르기 위해서 OpenCV를 빌드하기엔 부담스러움

Client에서 Server로 이미지 업로드시, **Webp** 형식 사용

Webp to jpg로 변환한 후, 첫 프레임을 Input으로!
=> OpenCV보다 훨씬 가볍게 동작

문제점

[Case #2. Collage Type이 상대적으로 NSFW 점수가 높다]

문제 정의 → 데이터 수집 및 전처리 → 모델 선정 및 학습 → 성능 개선



문제점

[Case #2. Collage Type이 상대적으로 NSFW 점수가 높다]

문제 정의 → 데이터 수집 및 전처리 → 모델 선정 및 학습 → 성능 개선

[모델이 사람의 얼굴이 나온 횟수에 민감하다는 사실을 발견]



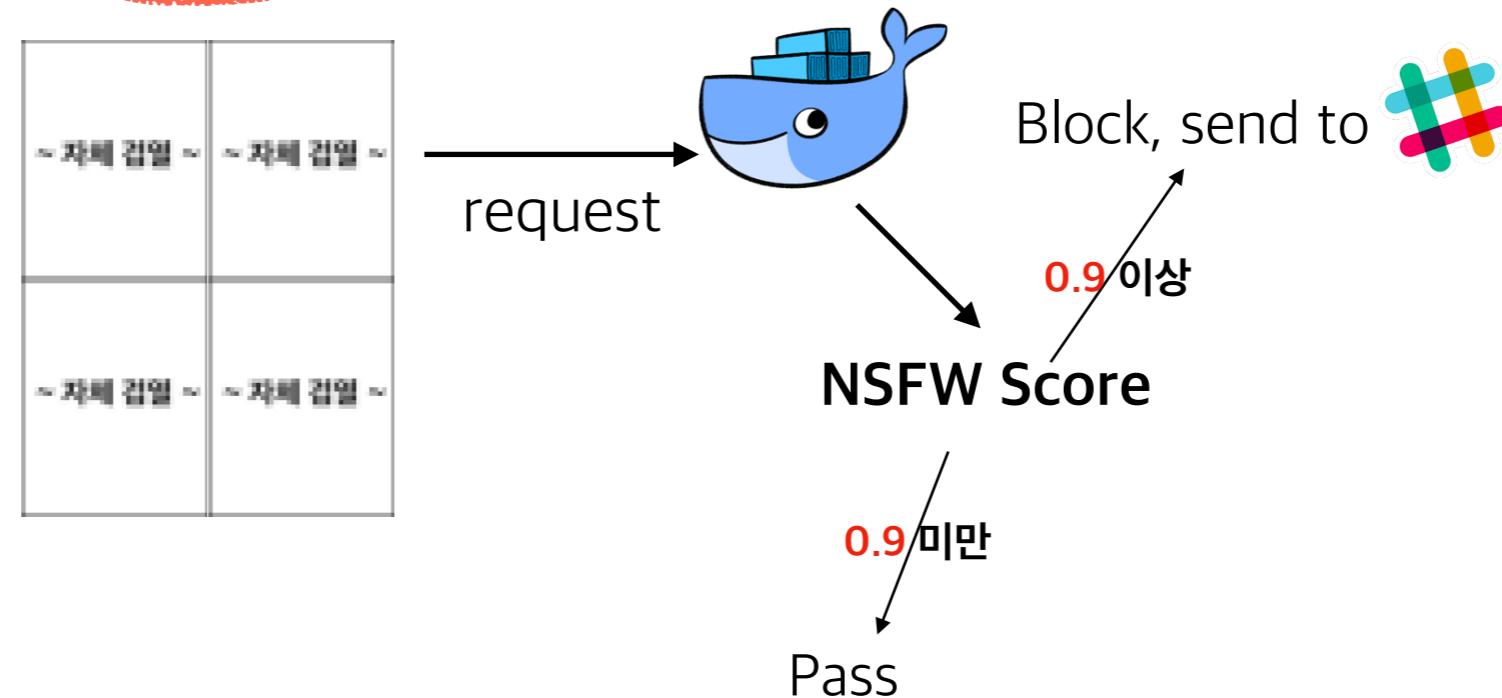
문제점

[Case #2. Collage Type이 상대적으로 NSFW 점수가 높다]

문제 정의 → 데이터 수집 및 전처리 → 모델 선정 및 학습 → 성능 개선

[모델이 사람의 얼굴이 나온 횟수에 민감하다는 사실을 발견]

content_type이 'Collage' 일 경우



문제점

[Case #3. 한번 올리는 유저가 계속 올린다]

문제 정의 → 데이터 수집 및 전처리 → 모델 선정 및 학습 → 성능 개선

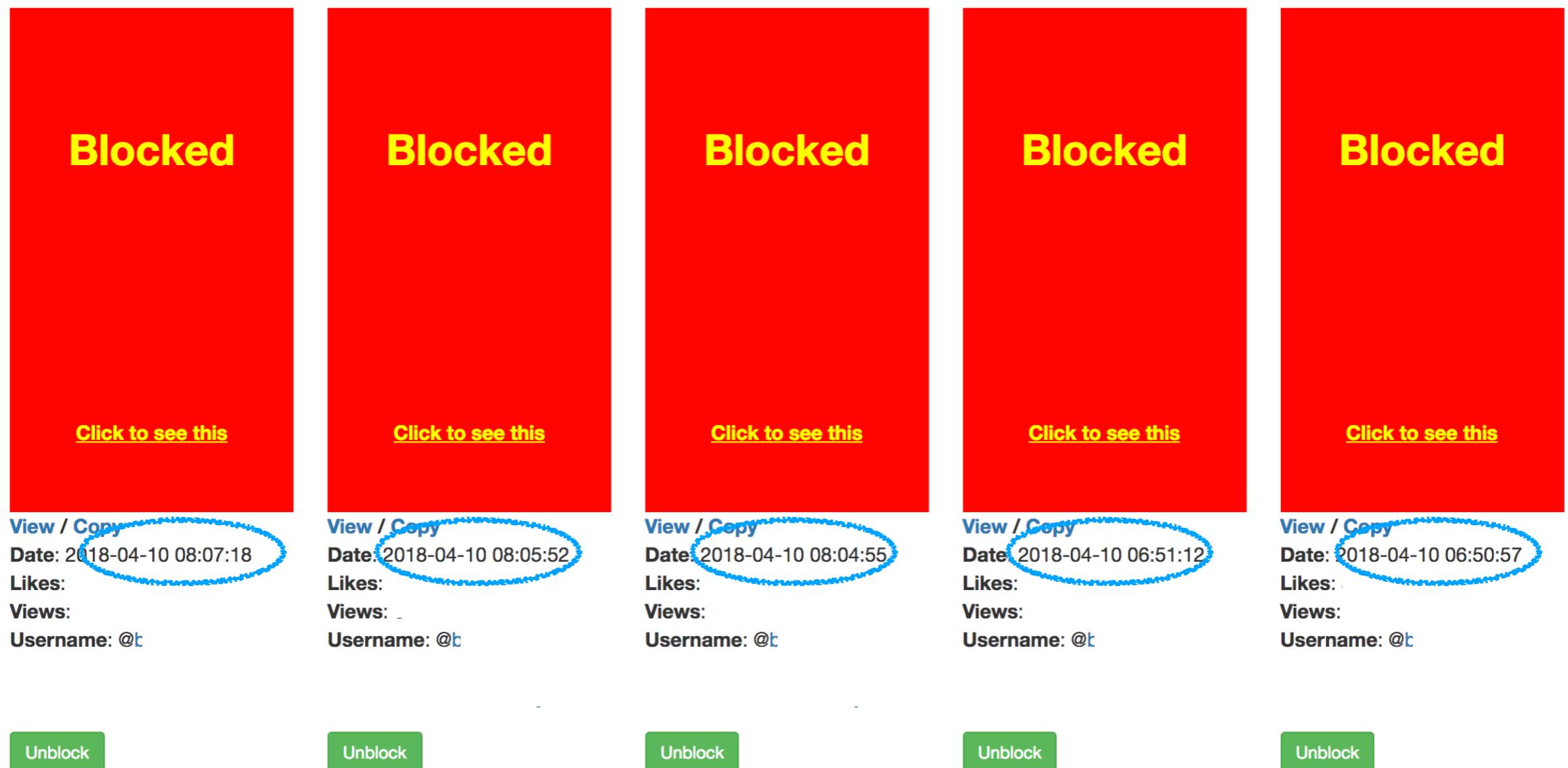
[자신의 신체를 뿐내는(?) 사람들은, 계속 뿐낸다..]

문제점

[Case #3. 한번 올리는 유저가 계속 올린다]

문제 정의 → 데이터 수집 및 전처리 → 모델 선정 및 학습 → 성능 개선

[자신의 신체를 뽐내는(?) 사람들은, 계속 뽐낸다..]



문제점

[Case #3. 한번 올리는 유저가 계속 올린다]

문제 정의 → 데이터 수집 및 전처리 → 모델 선정 및 학습 → 성능 개선

[자신의 신체를 뽐내는(?) 사람들은, 계속 뽐낸다..]

과거 로그를 찾아

자주 야한 사진을 올린다면($n \geq 3$) NSFW Score가 낮아도(score ≥ 0.5) Block

Story 3. Summary

문제 상황	해결 방안
음란 사진이 업로드된다 업로드되는 사진을 Operation Team이 감당하기 힘듬	직접 vgg16, ResNet 구현 (실패) Yahoo NSFW 사용
추가 문제 상황	추가 해결 방안
Case 1. 이미지만 처리 가능 Case 2. Collage Type에 대한 오판 Case 3. 한번 올리는 유저들에 대한 대처	Webp의 첫 프레임 활용 Collage Type일 경우 NSFW Threshold 조절 과거 로그를 탐색해 유저별 NSFW Threshold 조절

의의

Operation Team의 시간 절약
Retrica 유저 환경 개선
Google Vision API 비용 절감(약 \$8,00)

All Summary

Story 1. 문제 상황

Daily 업무 비중이 데이터 분석보다
단순 반복 작업
(Report 작성, 요청 데이터 처리)이
많음

Story 2. 문제 상황

이벤트 Dashboard 필요
BigQuery 비용이 이상함

Story 3. 문제 상황

음란 사진이 업로드된다
업로드되는 사진을 Operation Team이
감당하기 힘듬

Story 3. 추가 문제 상황

Case 1. 이미지만 처리 가능
Case 2. Collage Type에 대한 오판
Case 3. 한번 올리는 유저들에 대한 대처

All Summary

Story 1. 문제 상황

Dashboard

Story 2. 문제 상황

파이프라인

Story 3. 문제 상황

딥러닝 모델 사용

Story 3. 추가 문제 상황

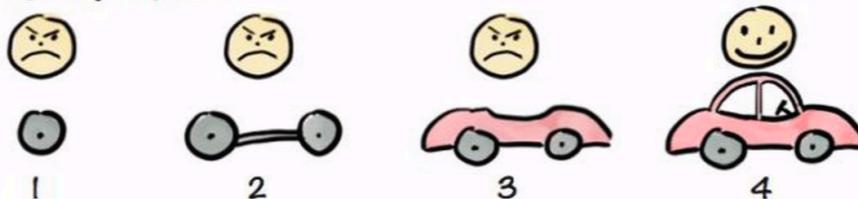
모델 환경 개선

느낀 것들

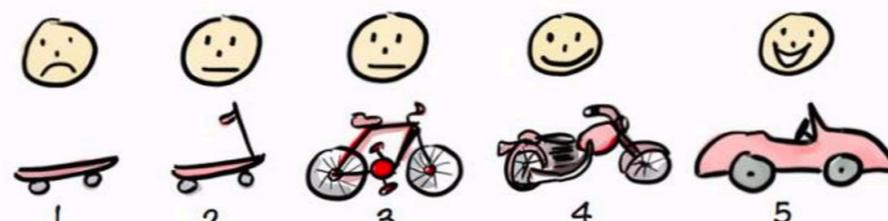
[모든 것을 직접 만들 필요는 없어요]

필요할 경우엔 오픈소스, 모델들을
프로토타입으로 사용해볼 수 있어요.
주어진 시간에 최대한의 Performance

Not like this....

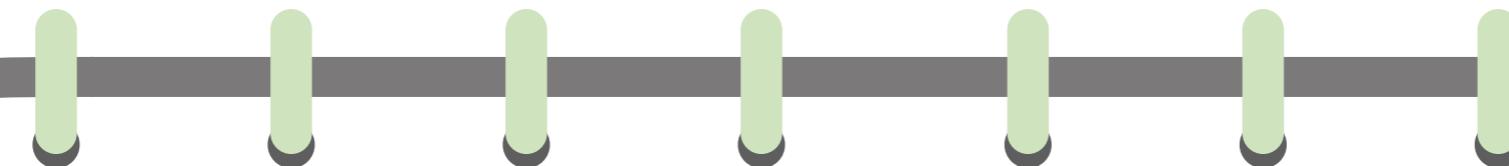


Like this!



Henrik Kniberg

느낀 것들



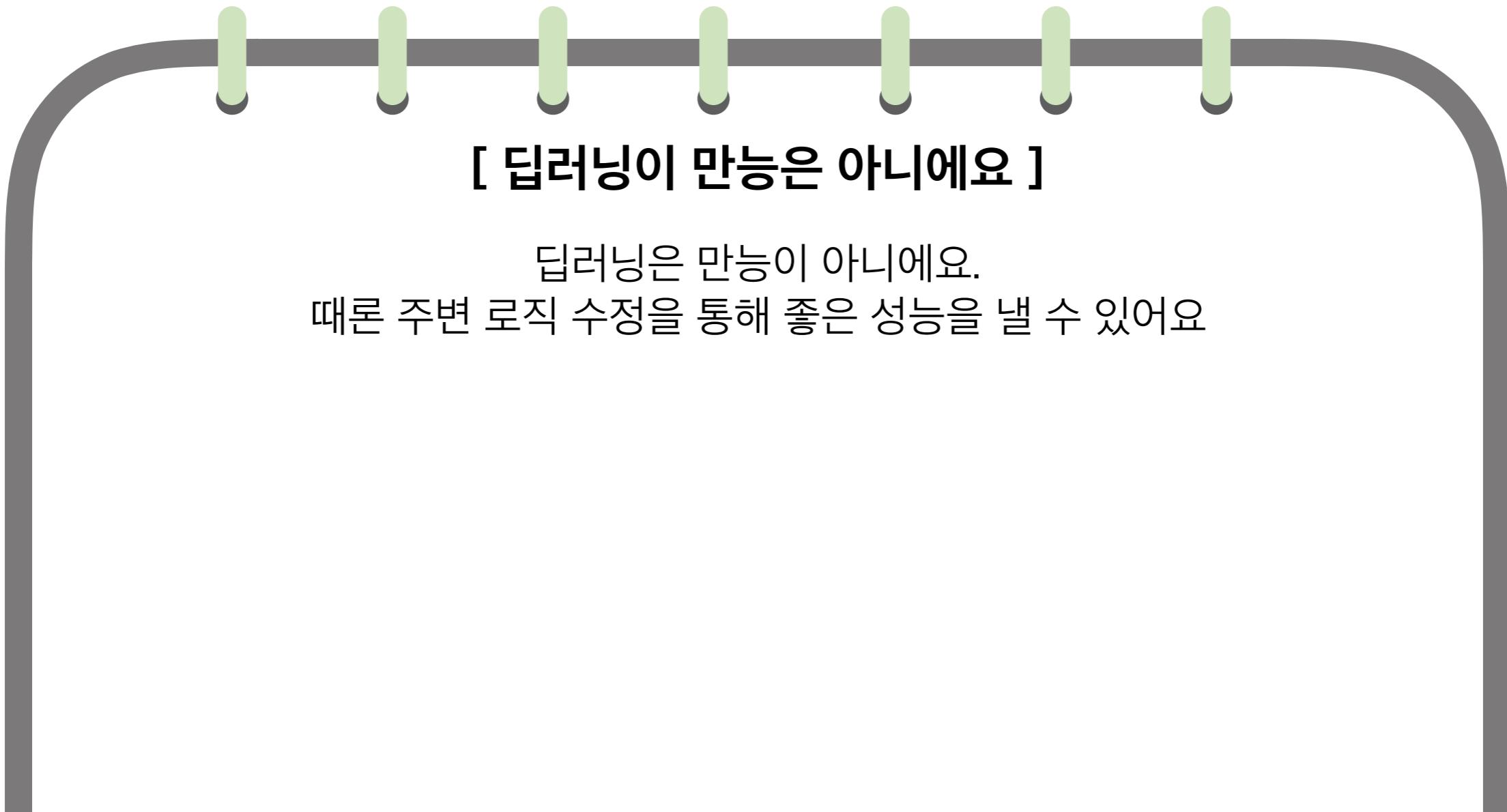
[선인의 지혜를 빌리면 좋아요]

예의있게 여쭤보면, 많은 분들이 도와주실거에요
나중엔 여러분들도 베풀어 주세요

{ Share
your
knowledge



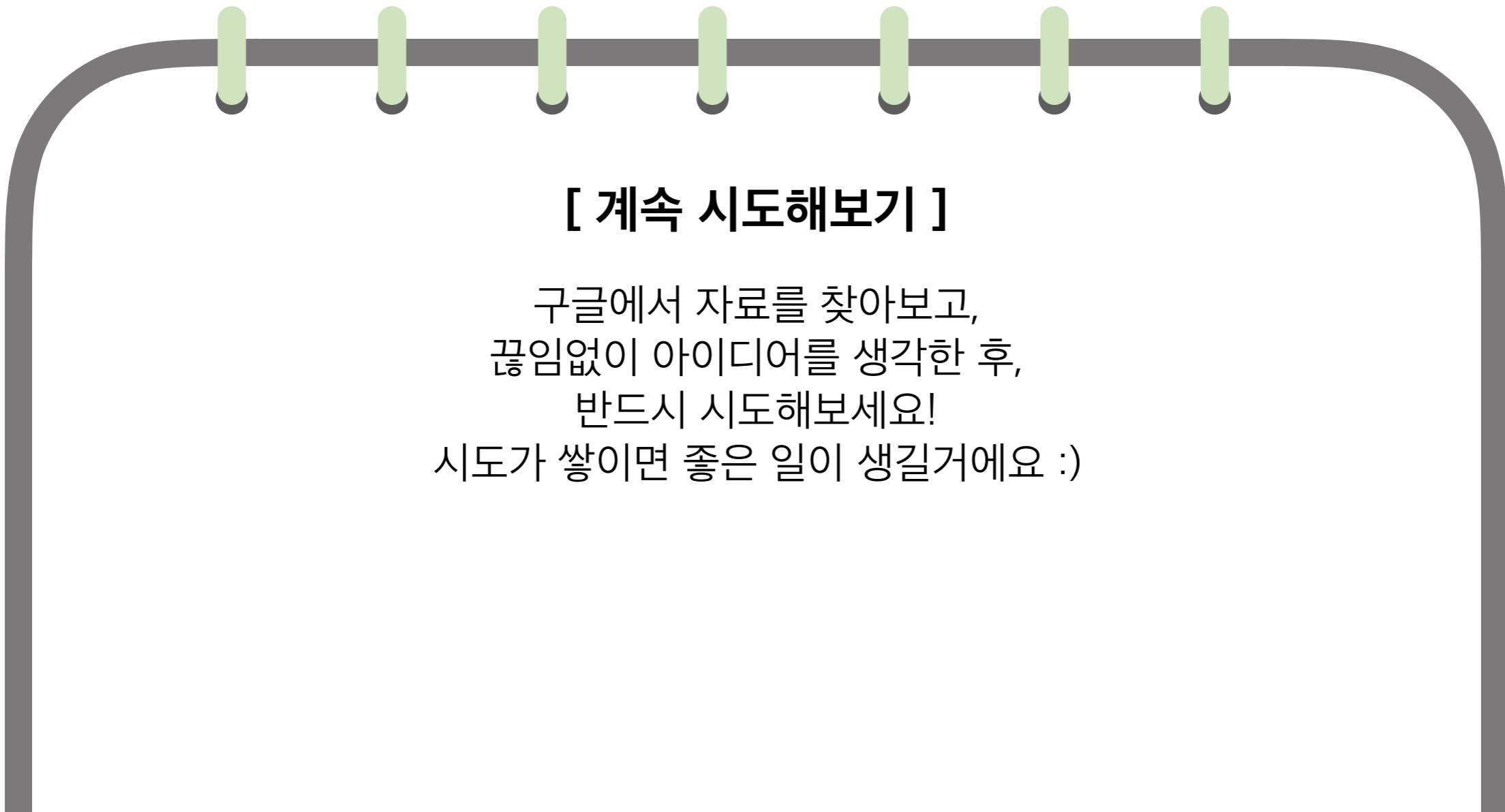
느낀 것들



[딥러닝이 만능은 아니에요]

딥러닝은 만능이 아니에요.
때론 주변 로직 수정을 통해 좋은 성능을 낼 수 있어요

느낀 것들



[계속 시도해보기]

구글에서 자료를 찾아보고,
끊임없이 아이디어를 생각한 후,
반드시 시도해보세요!

시도가 쌓이면 좋은 일이 생길거에요 :)

Thanks to

발표를 들어주신 여러분들
발표를 허락해주신 Retrica 박상원 대표님
발표를 제안주신 진유림님
발표를 후원해주신 Zepl
일러스트를 그려주신 김나연님
BigQuery를 친절하게 알려주셨던 이재광님
GCP 관련 문의를 많이해도 답변주신 조대현님
웹툰 대학일기 자까님

감사합니다

변성윤

메일 : zzsza@naver.com

페이스북 : [@naver.com](https://www.facebook.com/zzsza)

인스타그램 : data.scientist