



2021

BOAZ TWEETVIZ

DATA ENGINEERING

CONTENTS

01 주제 선정 배경

- TweetViz 주제 선정 배경
- 기존 TweetDeck의 문제점

03 데이터 파이프라인

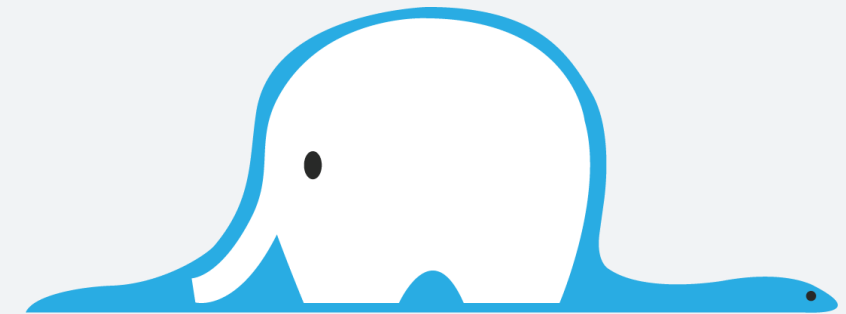
- 파이프라인 구축 시행착오
- 프로젝트 데이터 파이프라인 소개
- 데이터 파이프라인 추가 구현

02 데이터 설명

- 활용 데이터 소개

04 시연 및 한계점

- 프로젝트 시연
- 개선해야 할 점과 아쉬운 점





TweetViz

팀원소개



15기 분석
김서영



15기 시각화
김정연



15기 분석
남상대



12기 분석
박서호

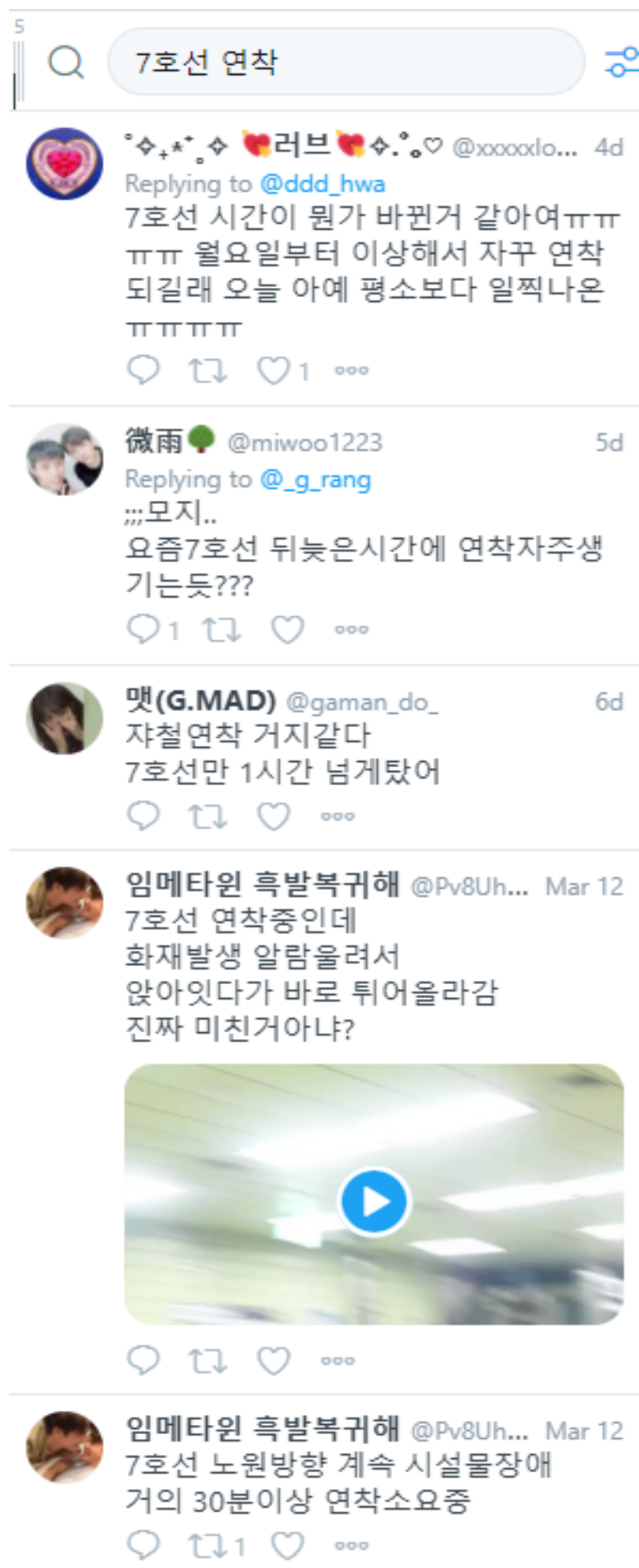


13기 분석
조수연

01

주제 선정 배경

주제 선정 배경



사회 > 사건/사고

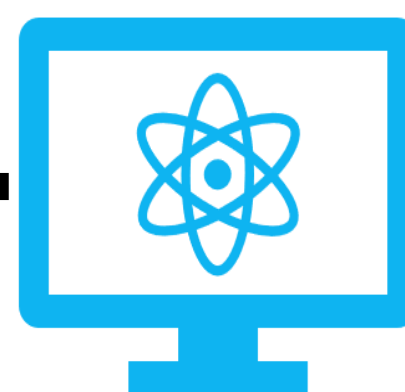
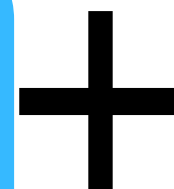
언론보다 빠른 SNS..."이젠 SNS를 언론이 보도하는 시대"

뉴스 > 정책뉴스

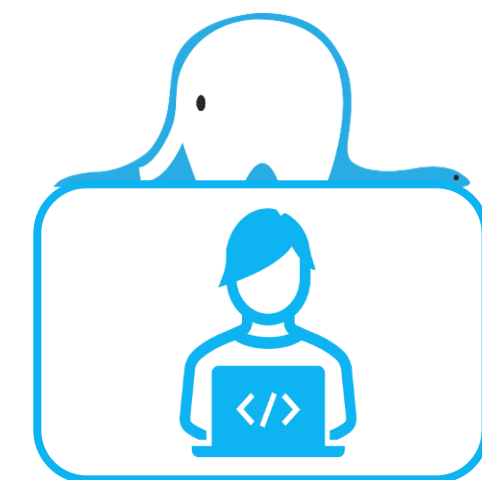
태풍보다 빠른 SNS "피해 줄였다"



트위터 실시간 데이터



Visualization



지하철 연착, 지진 등의 키워드의
실시간 뉴스를 받아오는 서비스



주제 선정 배경



배경



데이터



파이프라인



결과



< 기존 TWEETDECK >



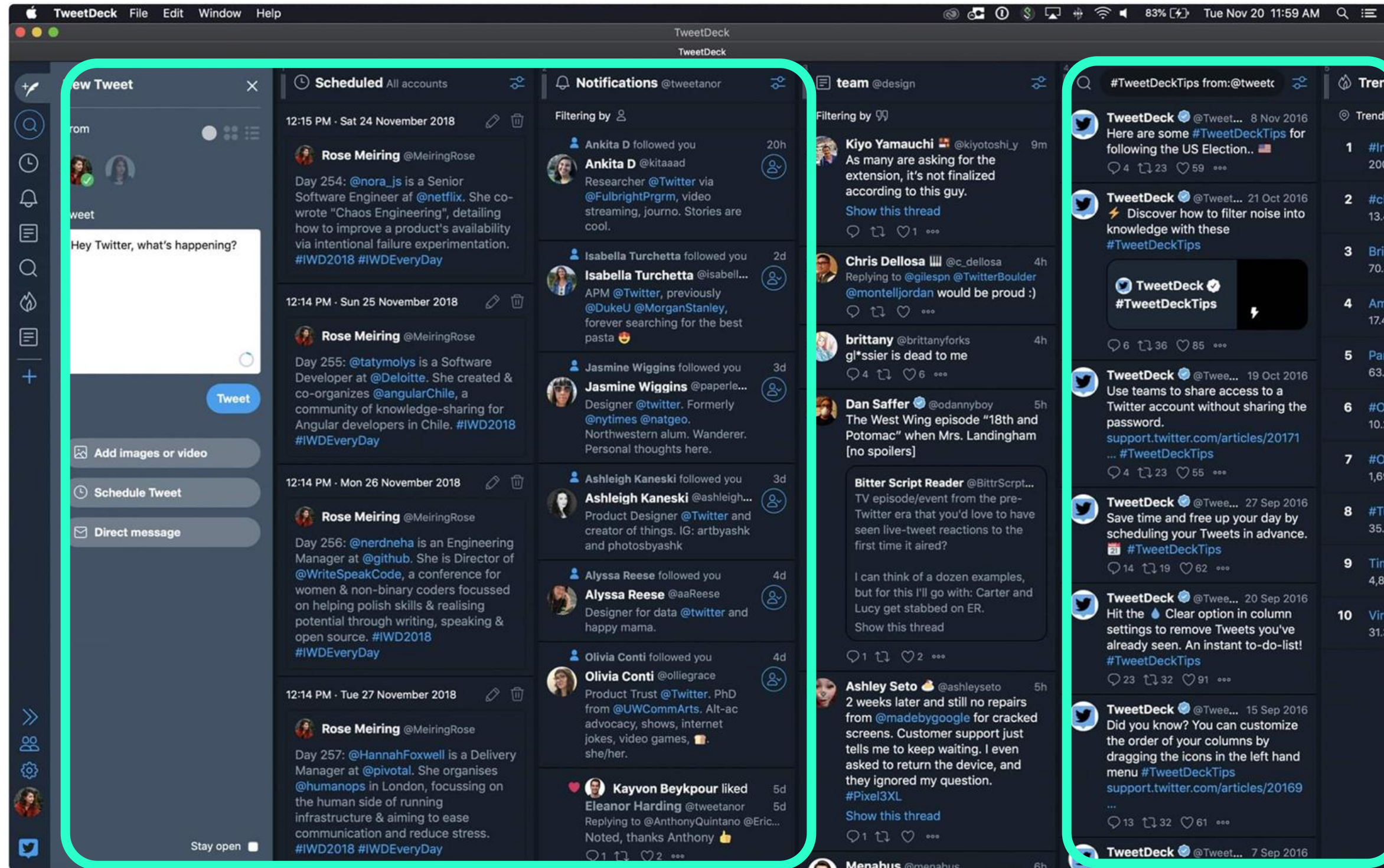
데이터



파이프라인



결과



개인 계정에 대한 활동

키워드 검색 칸



문제점

- 트윗 발생 날짜에 대한 필터링 불가
- 발생 지역을 알 수 없음(지리정보X)
- 사용자는 관여할 수 없는 자동 추천
- 알고리즘을 통한 트윗 노출
- 불필요한 칼럼 구성



주제 선정 배경



배경



데이터

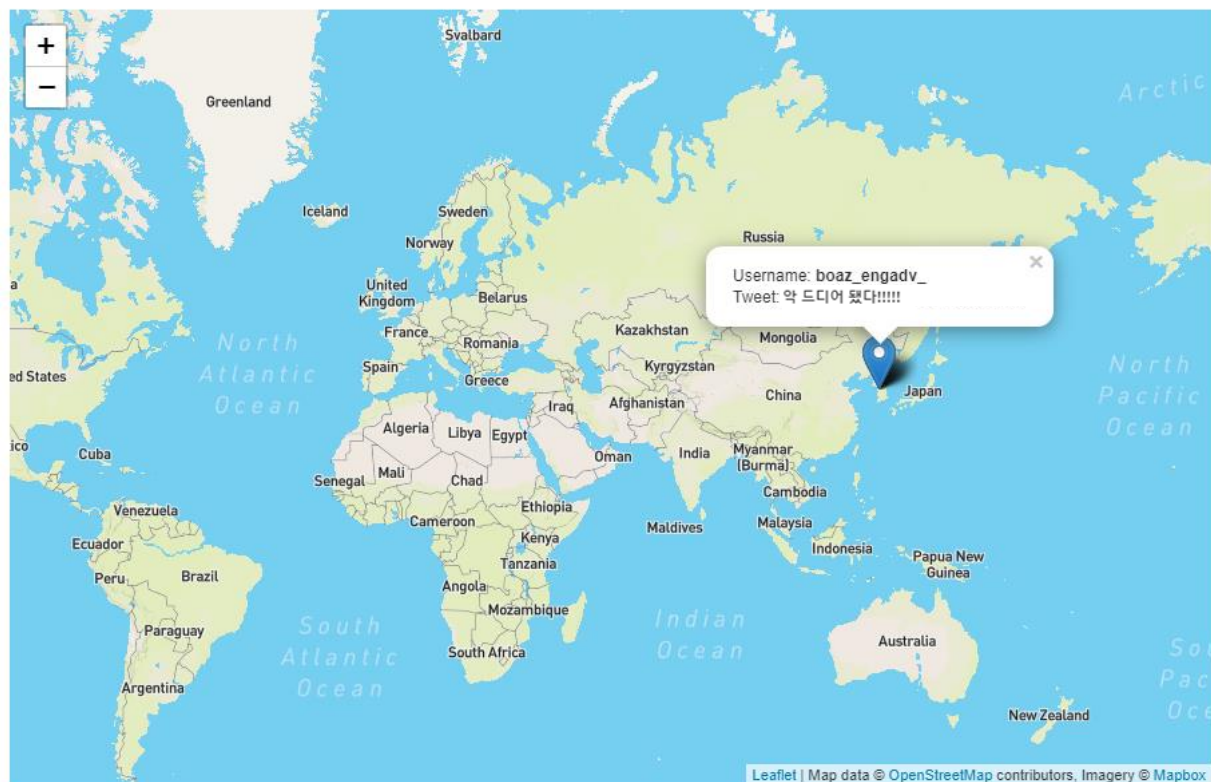


파이프라인

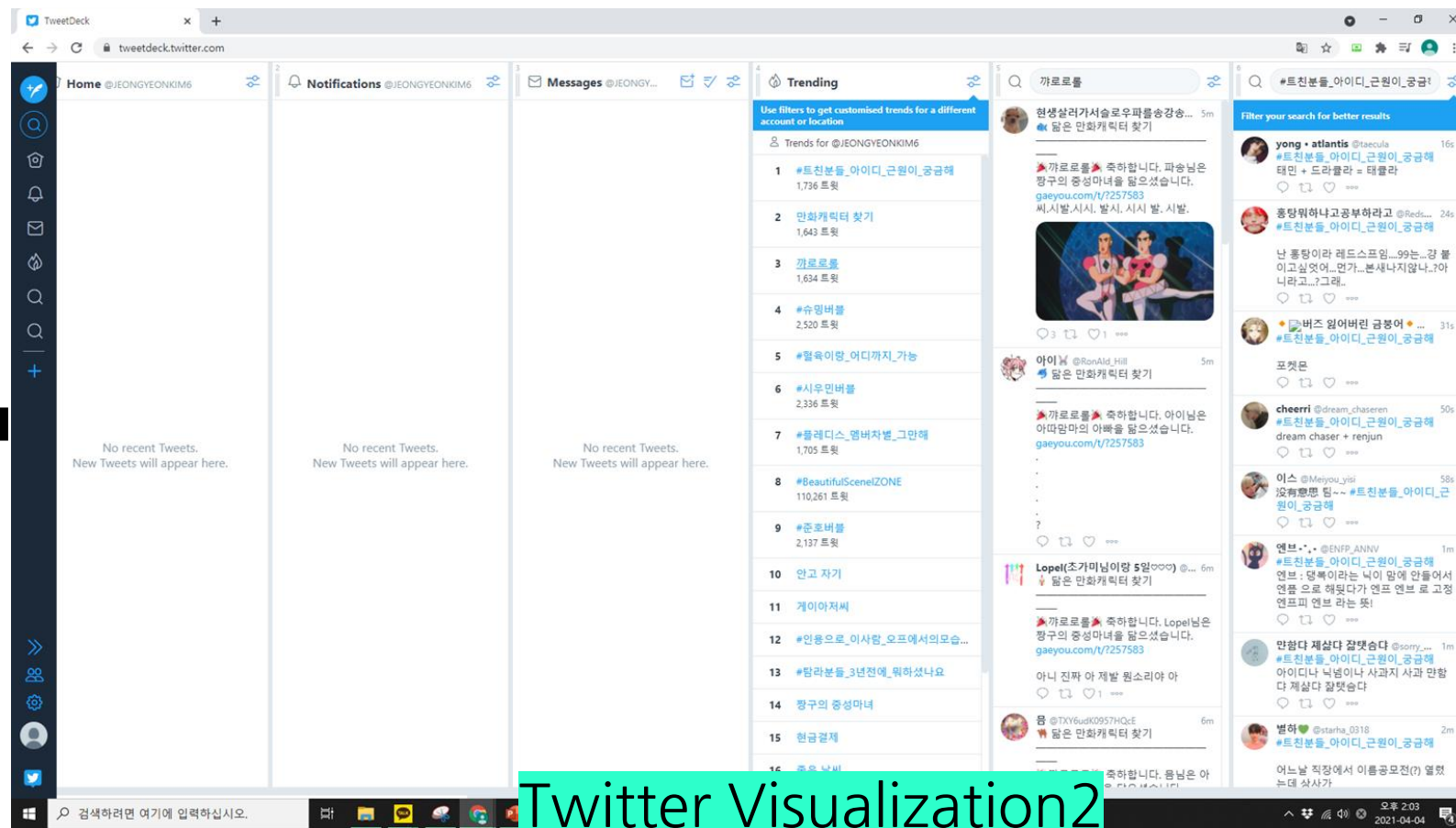
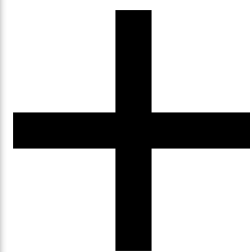


결과

Twitter World Map



: Tweets based on Map



Twitter Visualization2
: TweetDeck



02

데이터 설명

데이터 설명



트위터 데이터

- id
- created_at (트윗 발생 시간)
- text (내용)
- retweet_count
- favorite_count
- author_name
- author_screen_name
- profile_image_url
- topic(topic name 추가)
- Geo (위치 정보)
 - lat
 - long



Original Tweet



Re-Tweet

- Retweeted status를 통해 origin data에 접근
- Quoted 데이터의 경우 인용할 때 추가로 작성된 글은 제외
- Retweet 또는 Like 개수가 늘어난 origin data를 업데이트



배경



데이터



파이프라인



결과

Origin Data

```
_id: ObjectId("60e159086ea61dbe7b1b2c74")
created_at: "Sun Jul 04 06:45:25 +0000 2021"
id: 1411576865148395520
id_str: "1411576865148395520"
text:
  "장맛비 소강에 우산 접는 시민들..내일 다시 비
  날하했던 #장마 전선은 내일 다시 복상할 것으로 예보돼 날해안 지역부터 다..."
source: "<a href='http://twitter.com/download/android' rel='nofollow'>Twitter f..."
truncated: true
in_reply_to_status_id: null
in_reply_to_status_id_str: null
in_reply_to_user_id: null
in_reply_to_user_id_str: null
in_reply_to_screen_name: null
> user: Object
  geo: null
  coordinates: null
  place: null
  contributors: null
  is_quote_status: false
> extended_tweet: Object
  quote_count: 0
  reply_count: 0
  retweet_count: 0
  favorite_count: 0
> entities: Object
  favorited: false
  retweeted: false
  possibly_sensitive: false
  filter_level: "low"
  lang: "ko"
  timestamp_ms: "1625381125138"
```

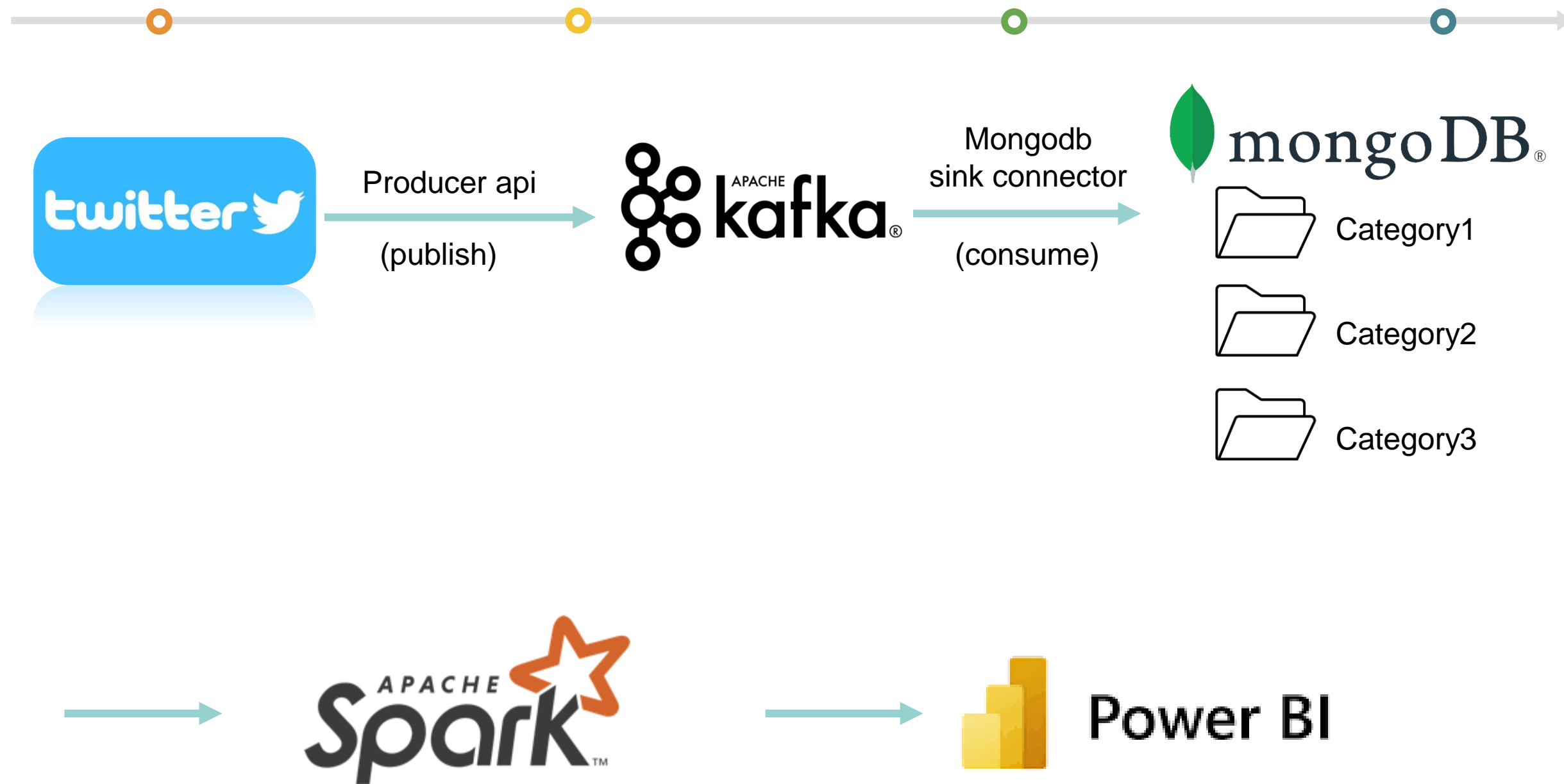
```
contributors: null
> retweeted_status: Object
  created_at: "Sun Jul 04 06:16:18 +0000 2021"
  id: 1411569537695338497
  id_str: "1411569537695338497"
  text:
    "Korea (Seoul) 📍
    #korean #seoul #장마 @북촌한옥형 https://t.co/D1s..."
  source: "<a href='http://instagram.com' rel='nofollow'>Instagram</a>"
  truncated: false
  in_reply_to_status_id: null
  in_reply_to_status_id_str: null
  in_reply_to_user_id: null
  in_reply_to_user_id_str: null
  in_reply_to_screen_name: null
> user: Object
  geo: null
  coordinates: null
  place: null
  contributors: null
  is_quote_status: false
  quote_count: 0
  reply_count: 2
  retweet_count: 7
  favorite_count: 20
> entities: Object
  favorited: false
  retweeted: false
  possibly_sensitive: false
  filter_level: "low"
  lang: "ko"
is_quote_status: false
quote_count: 0
reply_count: 0
```

Retweeted Data

03

파이프라인

프로젝트 파이프라인 (시행착오)





■ 시행 착오 Spark - Kafka - MongoDB



배경



데이터



파이프라인



결과

Twitter - Kafka - MongoDB



Kafka Connect

■ 시행 착오 Spark - Kafka - MongoDB



id	created_at	text	retweet_count	favorite_count	name	screen_name	profile_image_url	RT_id
1412602037892567041	Wed Jul 07 02:39:...	RT @iukkxrddfhan:...	0	0	rpzlee	Rapunzelx31	http://pbs.twimg....	null
1412602038236680193	Wed Jul 07 02:39:...	RT @nct_menfess: ...	0	0	Kasihh.tryanaa	KasihhTryana	http://pbs.twimg....	1412440500527796228
1412657043043852289	Wed Jul 07 06:17:...	RT @NCT127STRMth:...	0	0	🌈🌻🌱🙌😊	knick_jj	http://pbs.twimg....	null
1412602051884851207	Wed Jul 07 02:39:...	RT @markeu_yayy: ...	0	0	kyy ☺ 3ldwm 📌	markyourss	http://pbs.twimg....	1412588863088848900
1412616883702927364	Wed Jul 07 03:38:...	RT @HanteoNews: #...	0	0	syaa	myleejeno00	http://pbs.twimg....	1412564786848534533

[Spark DataFrame]

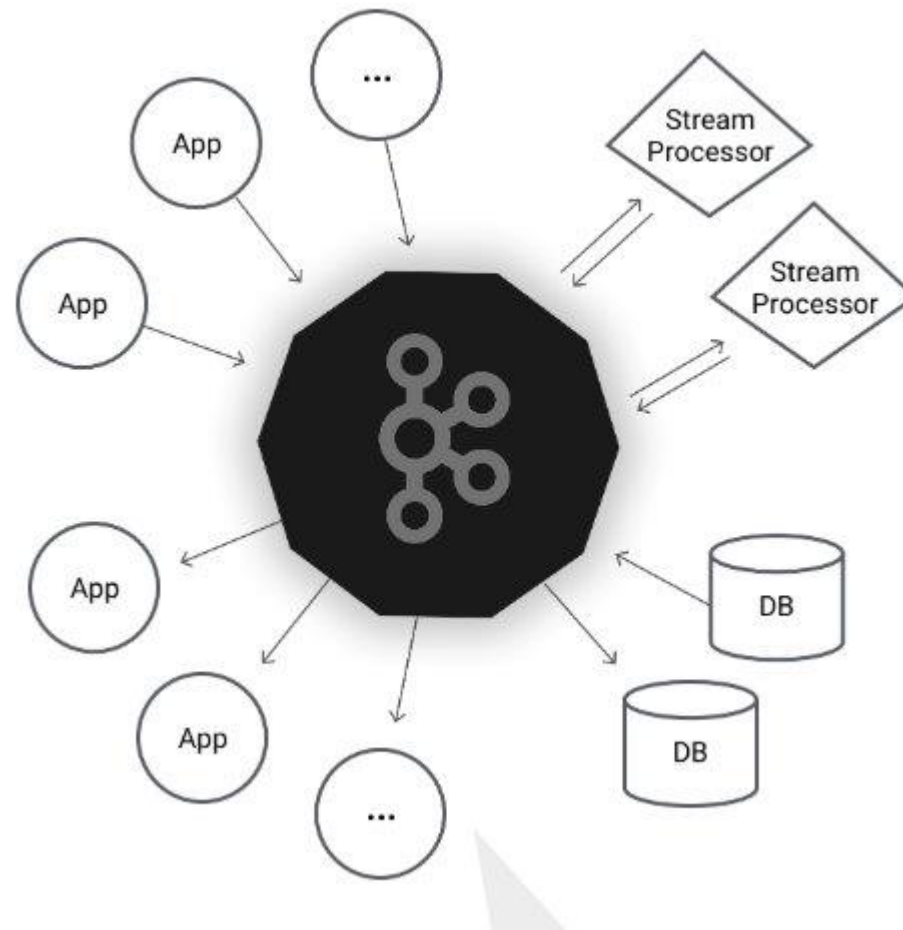
프로젝트 파이프라인



프로젝트 프레임워크 - Kafka



Why Kafka?



SCALABLE



HIGH THROUGHPUT

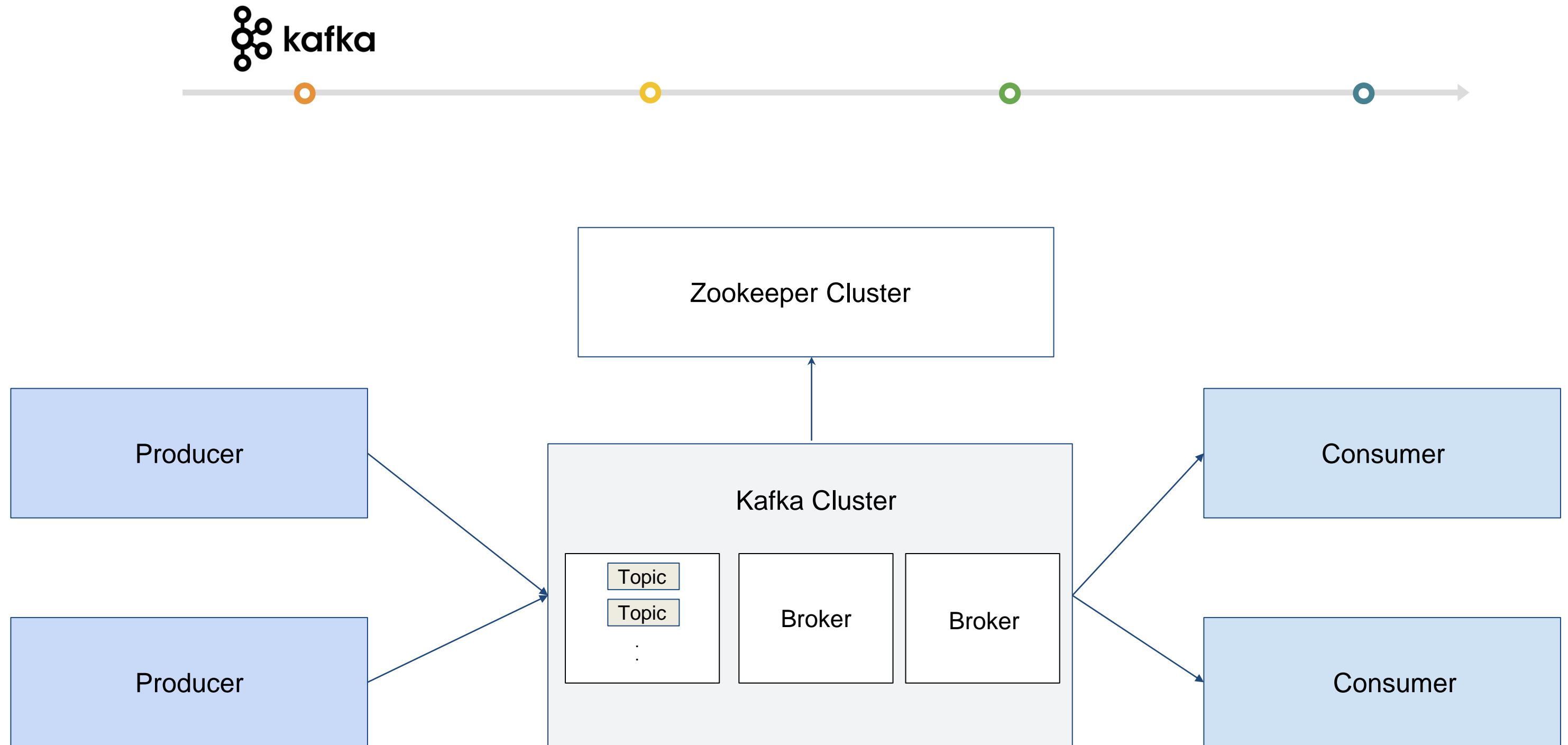


PERMANENT STORAGE

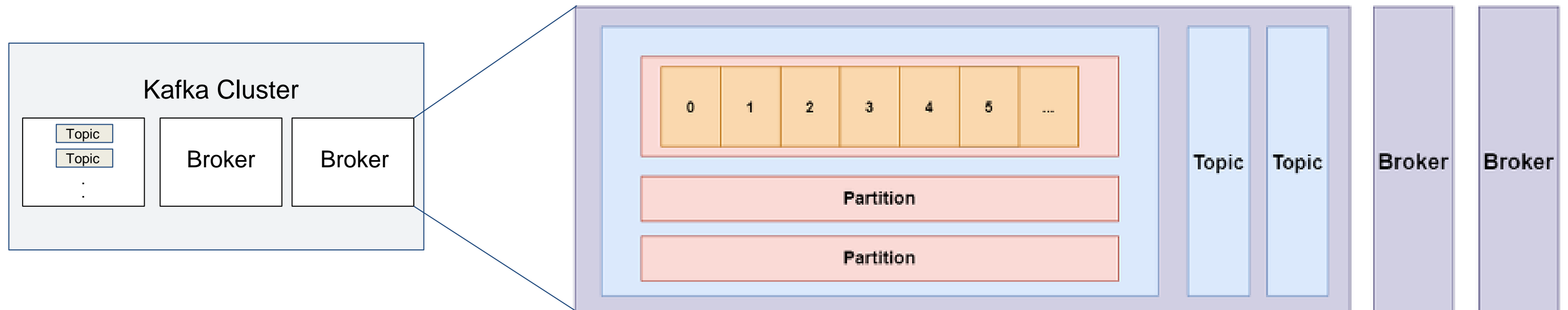


HIGH AVAILABILITY

프로젝트 프레임워크 - Kafka



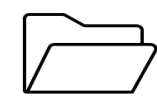


프로젝트 프레임워크 - Kafka



프로젝트 프레임워크 - Kafka

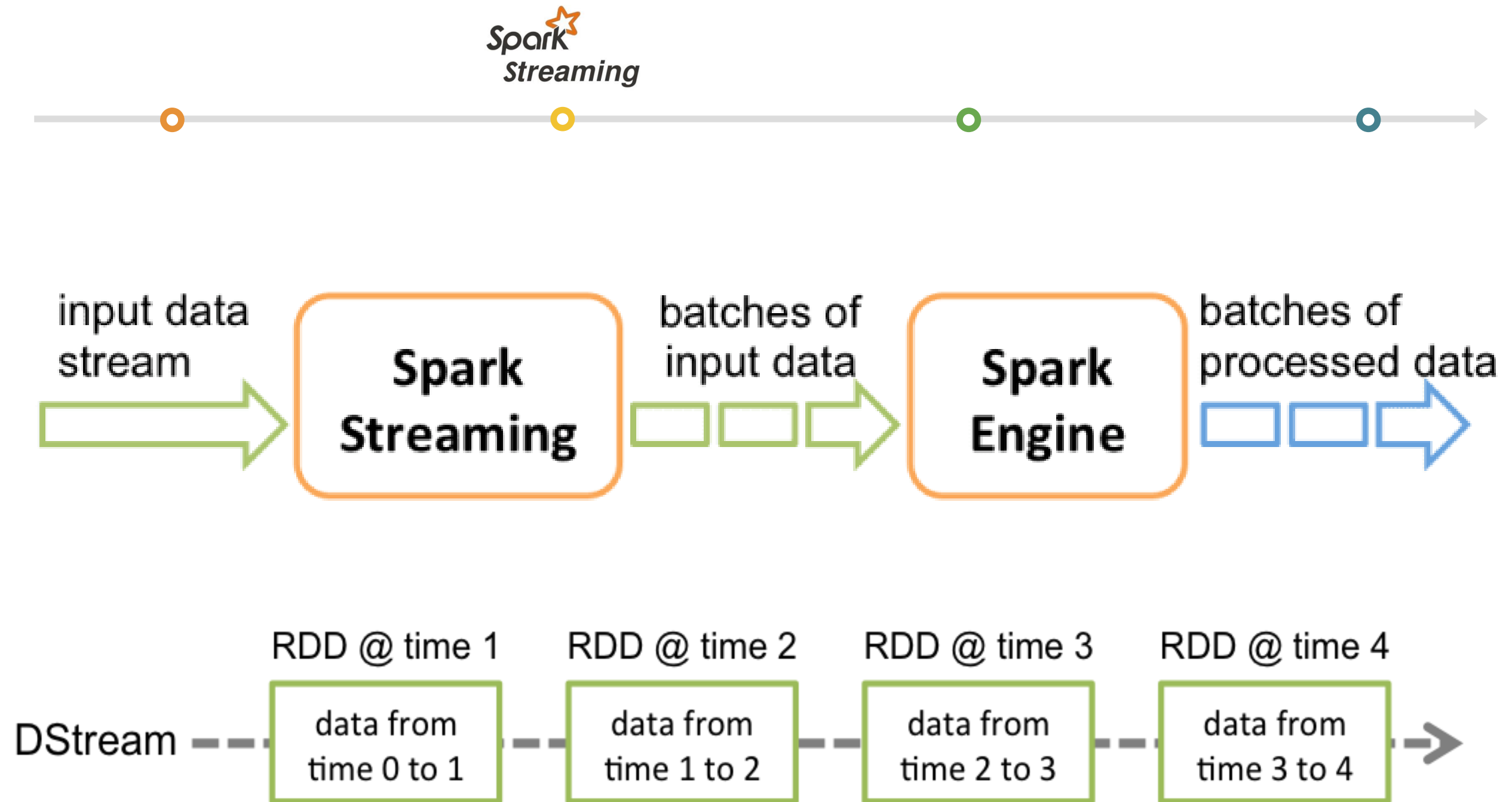


tweet 데이터를 키워드의 카테고리로 Topic 구분

-  1. 코로나 -> COVID 19
-  2. 태풍, 장마, 홍수 -> disaster
-  3. 폭염, 더위, 무더위, 더워 -> temperature

=> 각 Topic 별로 Partition 3개. 총 topic 3, partition 3개로 구성

프로젝트 프레임워크- Spark Streaming



프로젝트 프레임워크- Spark Streaming



프로젝트 프레임워크 - BigQuery & Visualization



서버리스 데이터 웨어하우스

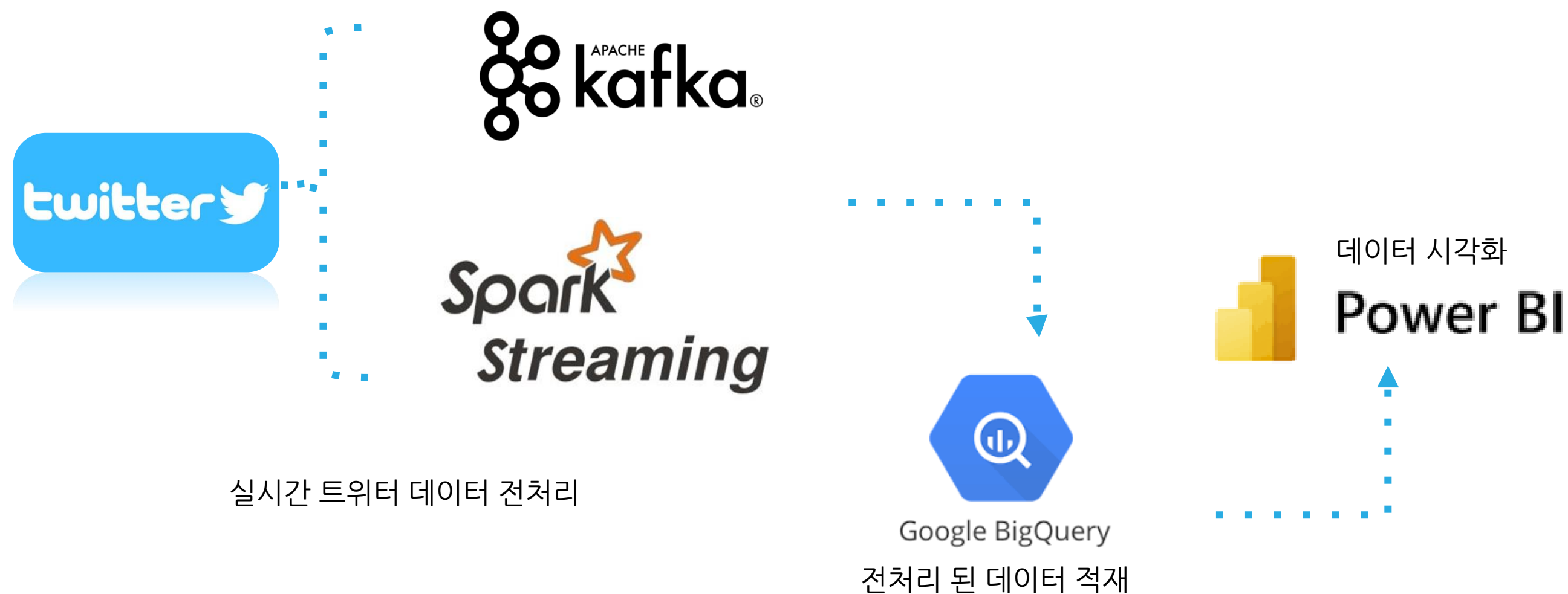
객체 스토리지, 스프레드시트의 데이터,
관리형 열 형식 스토리지를 통해 논리적 데이터
웨어하우스를 생성하여 모든 배치와 연속적으로
생성되는 스트림이 데이터를 분석



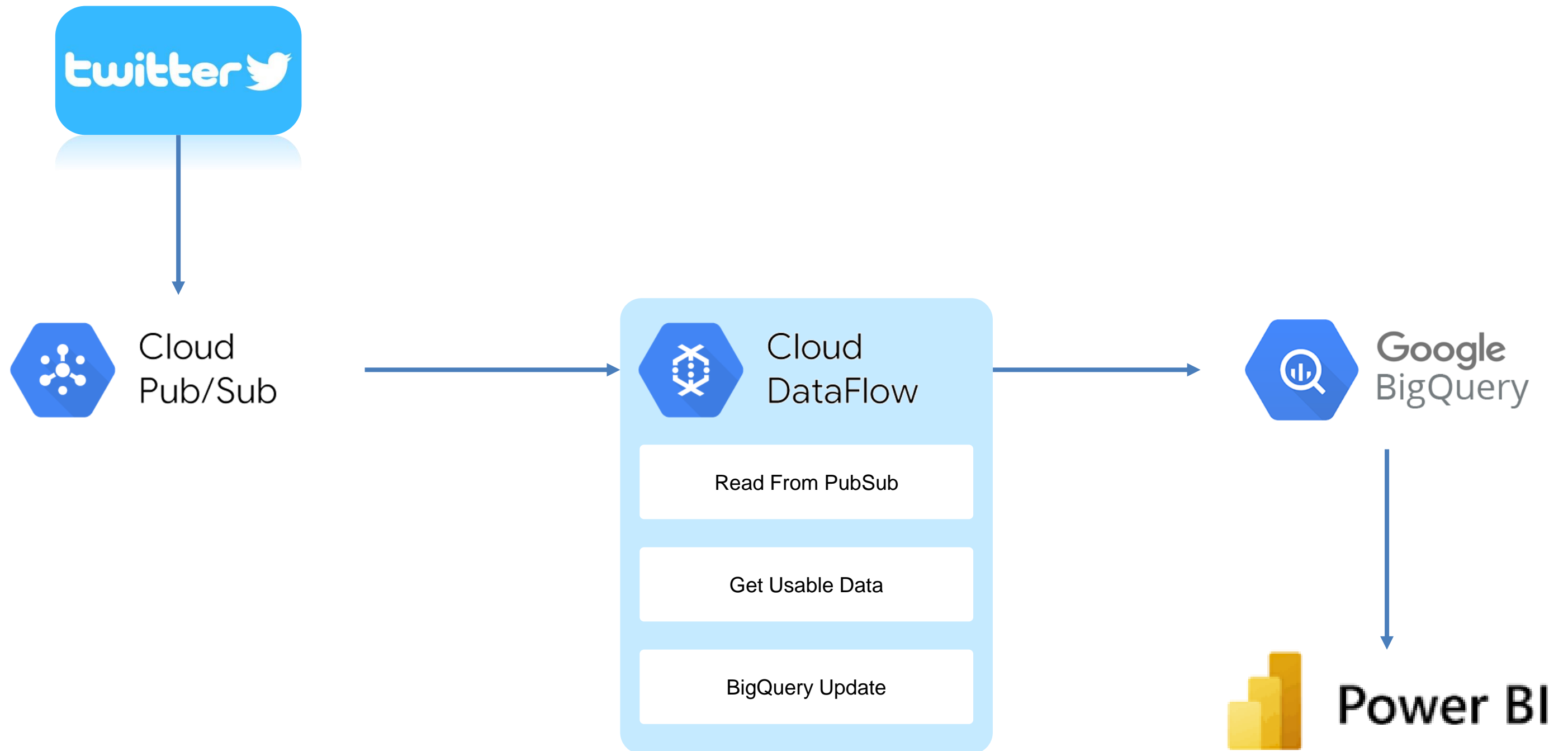
데이터 시각화 툴

마이크로소프트에서 만든
데이터 시각화 자동화 툴

프로젝트 전체 정리



■ GCP를 활용한 파이프라인



04

시연 및 한계점



시연 영상



배경



데이터



파이프라인

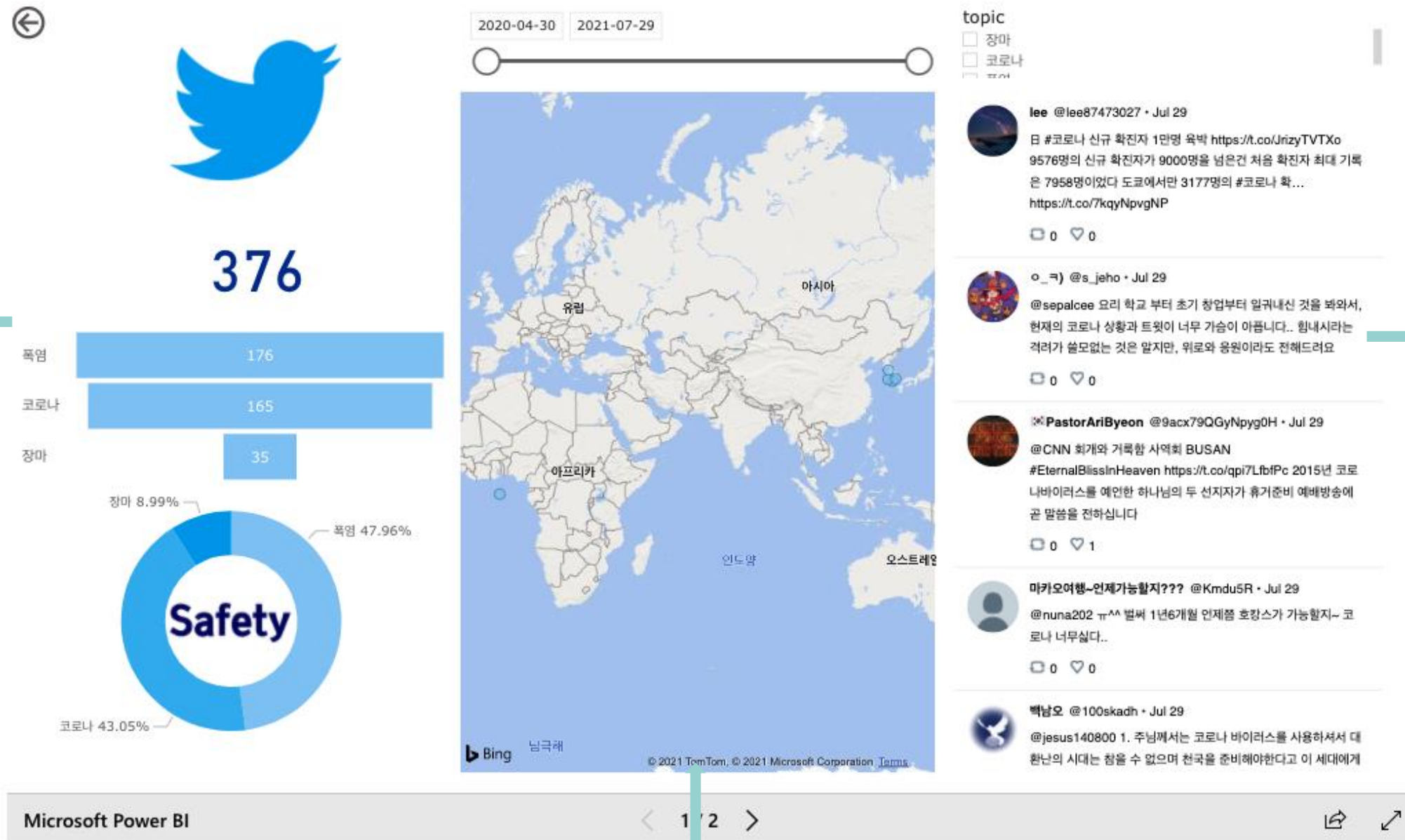


결과

TWEETVIZ Dashboard Filtering Slider

해당 대시보드는 PowerBI를 통해 시각화 처리 되었습니다. #장마 #코로나 #폭염에 대한 트윗을 확인해보세요.

Twitter Data



Tweet List

Tweet Location



배경



데이터



파이프라인



결과

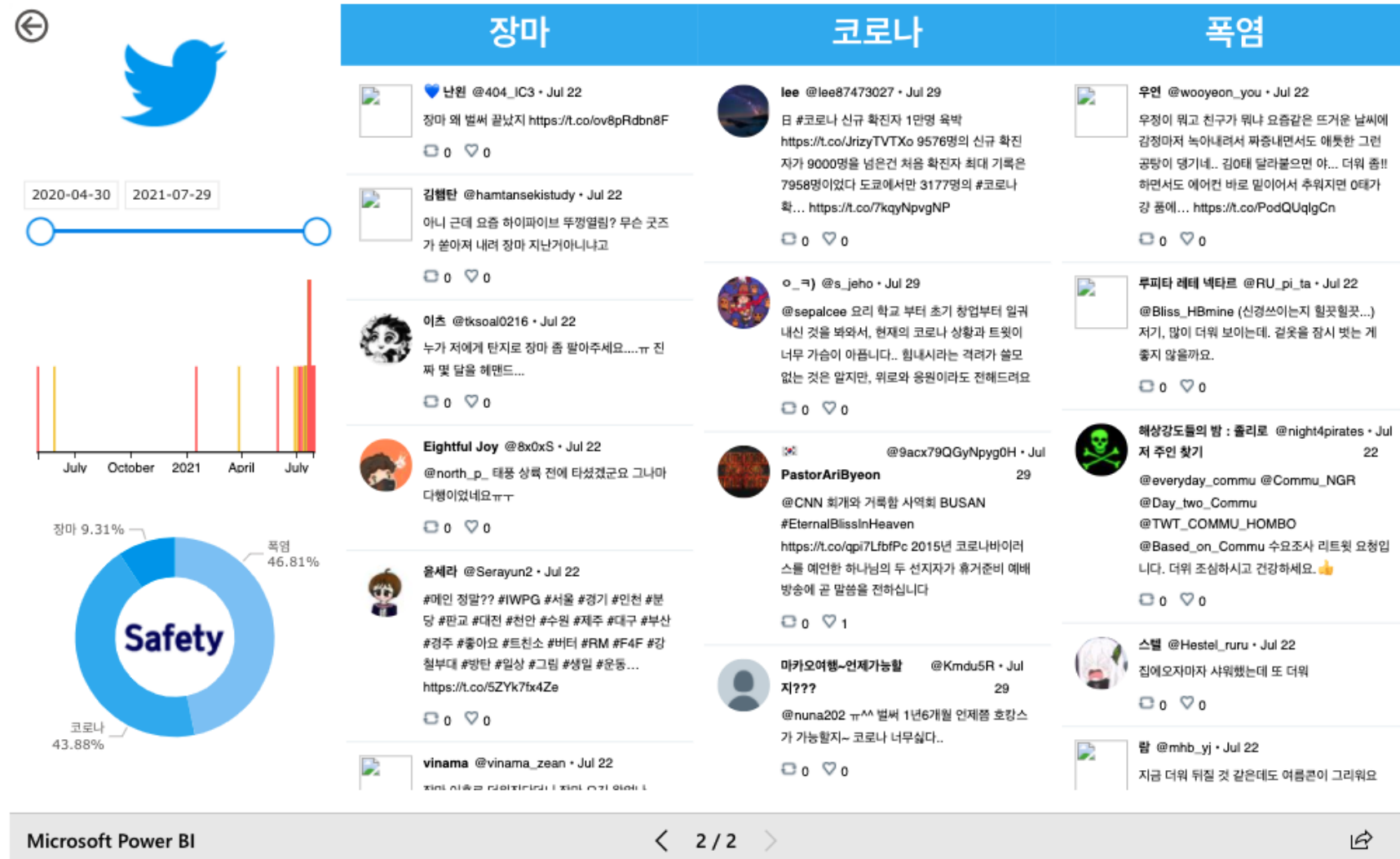
시연 영상

TWEETVIZ DASHBOARD

해당 대시보드는 PowerBI를 통해 시각화 처리 되었습니다. #장마 #코로나 #폭염에 대한 트윗을 확인해보세요.

Filtering
Slider

Twitter
Data



Tweet
List



배경



데이터

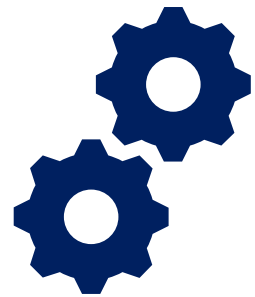


파이프라인



결과

■ 개선할 점 & 아쉬운 점



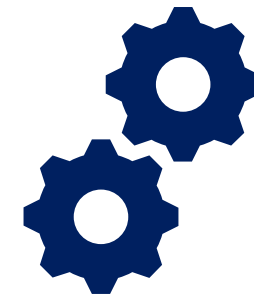
환경구축의 어려움

너무 다양한 엔지니어링 툴
시간 및 정보 부족



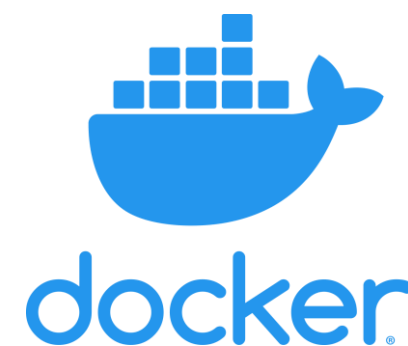
분석의 부족

시행착오가 많아 다양한 것들을
배웠으나 그만큼 분석에 미흡
실시간 분석이나 배치 처리를 통한 분석 부족



시각화 툴의 한계

시각화 툴에 대한 다양한 이슈가 발생하여
다채로운 시각화 진행이 어려웠음



도커

도커를 활용한 작업을 진행하려고
기초적인 실습을 모두 진행했으나
활용하지 못한 아쉬움



THANK YOU

<https://boaz-tweetviz.github.io/>

2021
BOAZ TWEETVIZ
DATA ENGINEERING