

CONNECTS LAB_DE

발화 분석 리포트 생성
파이프라인 구축
Team : #88한아이

2023 November 15
Team Connects LAB_DE



Our Team



대화를 할 때 내가 어떻게 말했는지에 대해 잘 모를 때가 많았고 이를 분석하는 프로젝트에 관심을 가지게 되었습니다. 또한 인스타를 통해 얻을 수 있는 데이터를 활용하여 얻을 수 있는 인사이트가 궁금해서 선택했습니다.

최준용



인스타그램의 월간 활성 사용자 수는 20억 명 정도라고 합니다. 많은 사람들이 사용하므로 많은 데이터가 수집될 것입니다. 사람들이 어떤 생각과 대화를 하는지 분석하고 시각화 해보고 싶습니다.

강성구



데이터 시대에 살고 있는 우리, 쏟아져 나오기 때문에 때로는 사소해 보이는 일상을 기록한 SNS 속에서 데이터의 가치를 찾아 건강한 가정을 만드는데 도움을 줄 수 있다는 것에 관심이 생겨 참여하게 되었습니다.

이상호



이제 한 살이 된 조카가 있어서 부모와 아이에 대한 발화 분석은 관심이 가는 주제이고 인스타그램은 사용자가 엄청나게 많기 때문에 다양한 정보를 얻을 수 있다고 생각했습니다.

곽준목

Overview

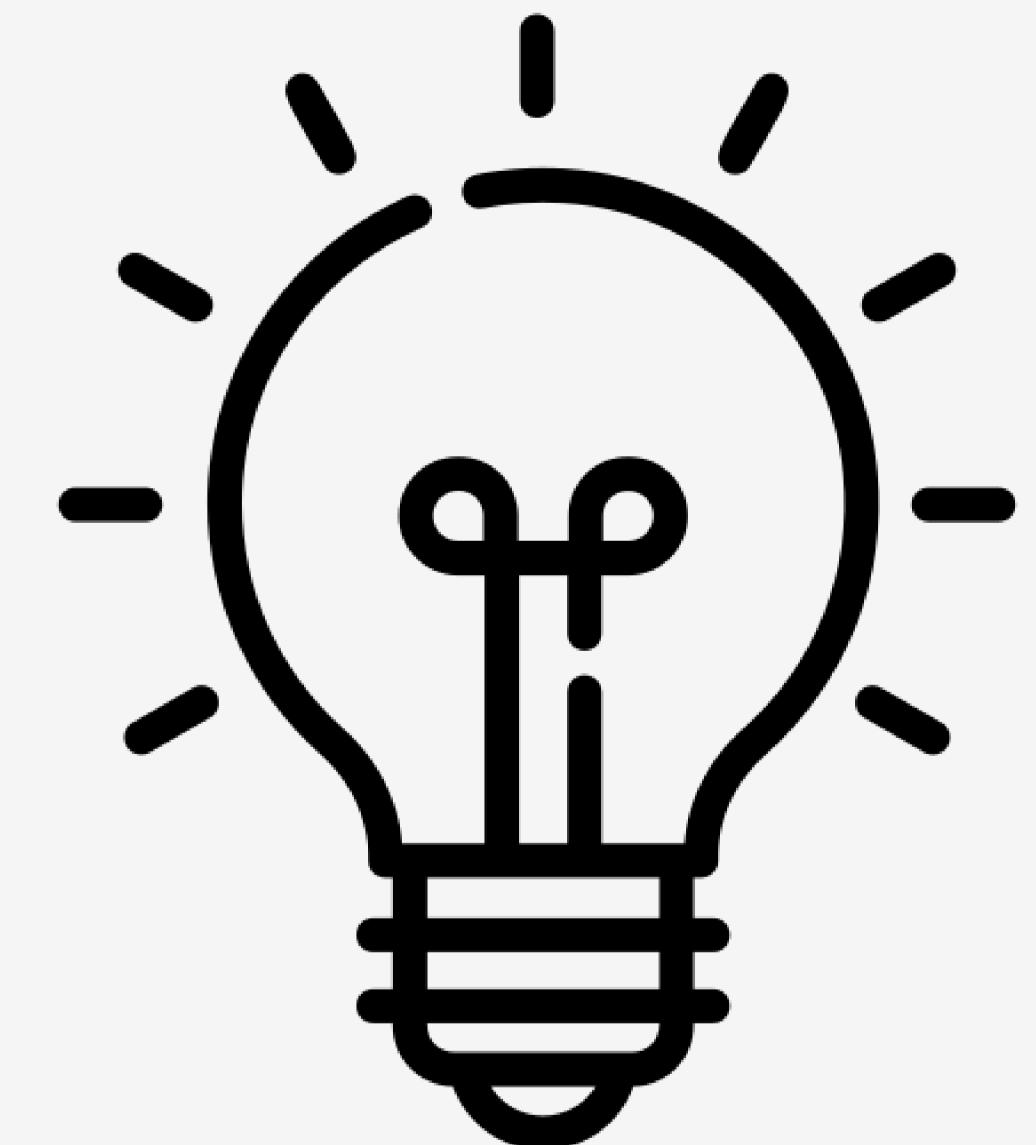
왜 아이와의 대화가 중요할까요?

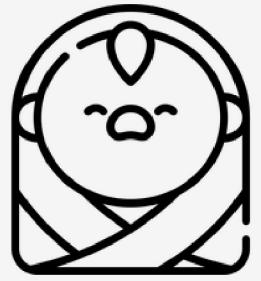
태어나는 순간부터 성장기 내내

부모와의 **소통**을 통해 아이는 말을 배웁니다.

또한 엄마 아빠와 대화를 통해 **사고하는 능력**과
사회적인 가치관을 습득하게 됩니다.

특히 생애초기의 아이들에게 전해지는 '**부모의 말**'은 아이
의 정서와 인지 발달에 중요한 양분으로 작용 합니다.





태아기

- 청각 발달
- 외부 자극 반응



영아기

- 한 두 단어
- 주로 명사
- 부모말 흉내

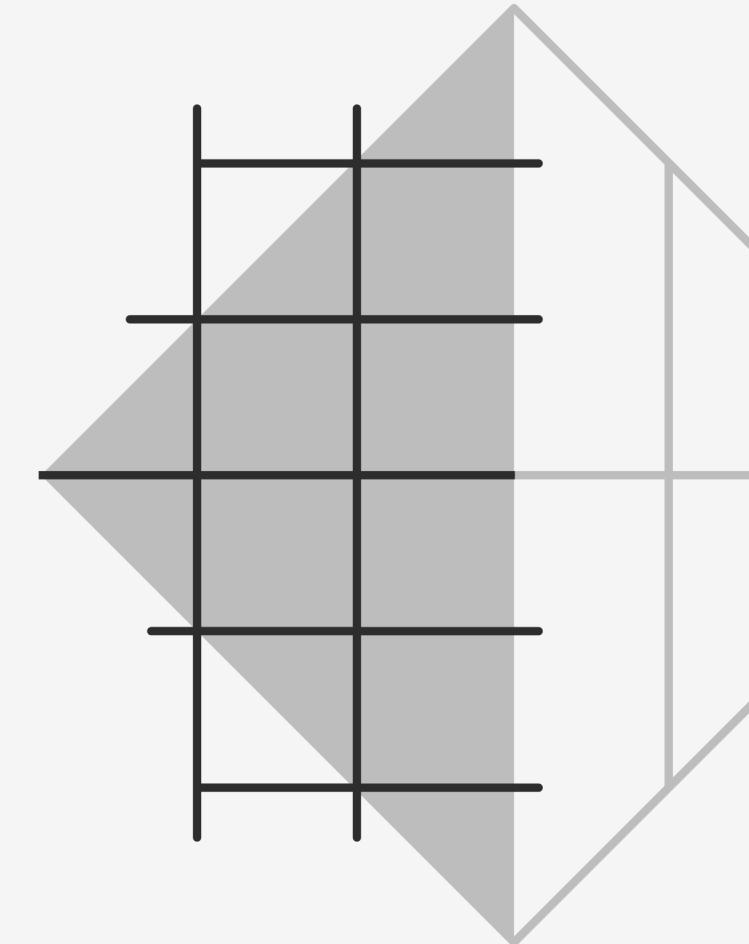


유아기

- 낱말 조합
- 문장 구조 형성
- 담화구조 이해



- 어휘력 발달
- 의사소통수단:
대화로 전환





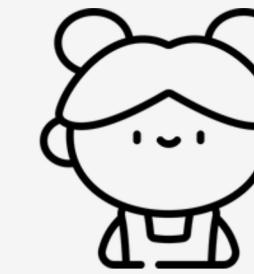
태아기

- 청각 발달
- 외부 자극 반응



영아기

- 한 두 단어
- 주로 명사
- 부모말 흉내

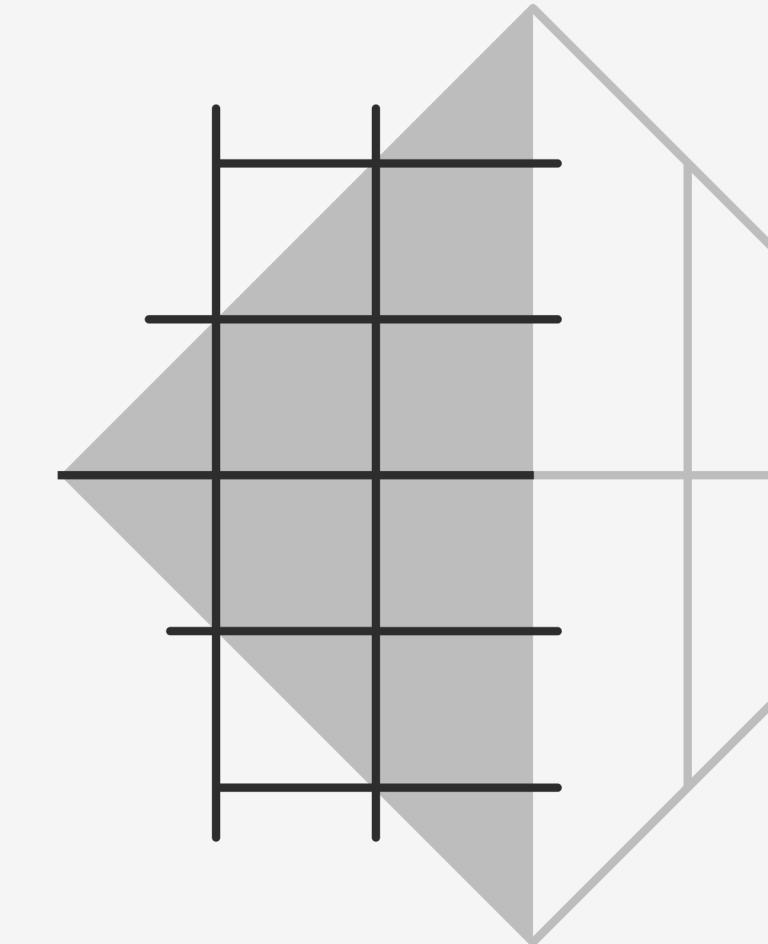


유아기

- 낱말 조합
- 문장 구조 형성
- 담화구조 이해



- 어휘력 발달
- 의사소통수단:
대화로 전환



두뇌 발달의 잔인한 진실

뇌의 **생물학적 성장은 4세가 되면 대부분 끝난다.** 아동의 학습 나이도와 인생 전체의 설계는 이 기간 동안 어떤 일이 일어나는지에 따라 크게 달라진다. 이것이 잔인한 진실인 이유는 무엇일까?

...(중략)...

태어나서 3년동안 적절한 음식을 충분히 먹지 못한 아기는 살아남을 수는 있겠지만 결코 원래 컸어야 하는 만큼 자라지 못한다. 마찬가지로 두뇌발달에 필요한 적절한 말을 충분히 듣지 못한 아기는 살아남기는 하겠지만, **배움에 커다란 어려움을 겪고 결코 자신의 지적 잠재력을 온전히 실현하지 못하기 마련이다.**

부모의 말, 아이의 뇌

두뇌 발달과
학습 능력을 결정짓는
3천만 단어의 힘



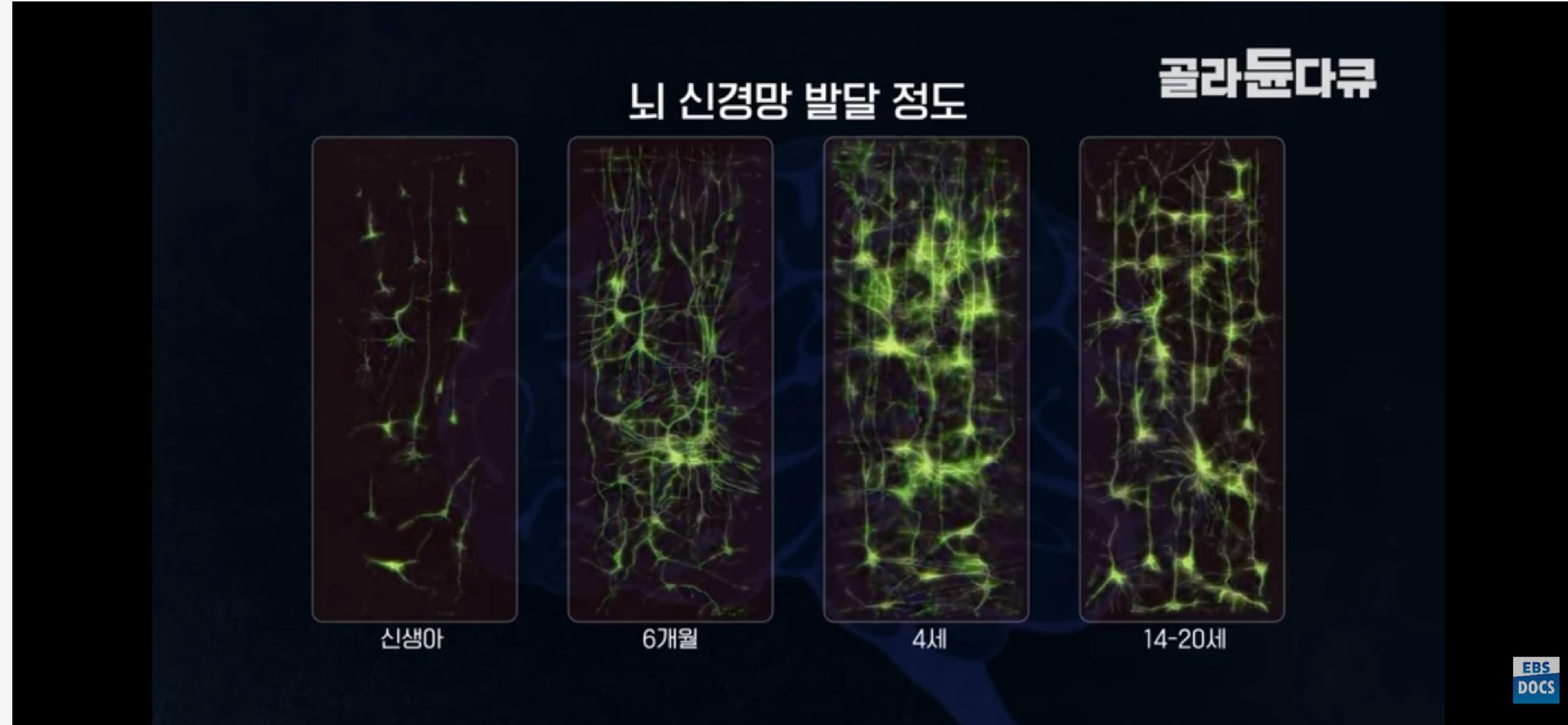
**“내 아이의 인생을 바꾸는 것은
IQ, 재능, 돈이 아니라 ‘부모의 말’이다!”**

EBS 육아 멘토, 『엄마의 말 공부』 이임숙 소장
과학 팩트 육아 채널 〈베싸TV〉 박정은 대표 강력 추천!

데이나 서비스 키드
최다인 옮김

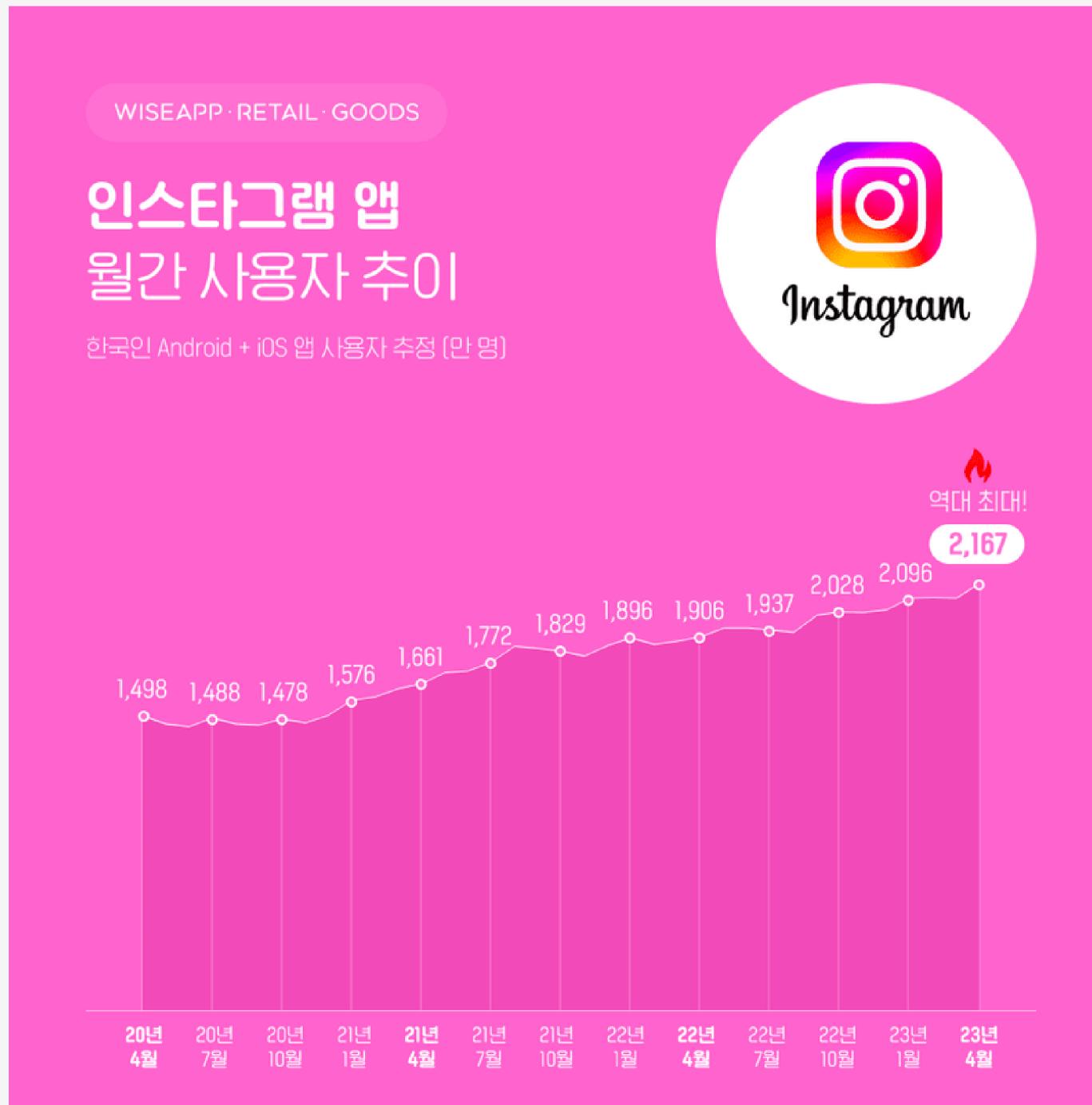
자자인
보호
자료

부·피



- 생애초기(0-4세)는 뇌 신경망의 발달이 폭발적으로 일어나는 시기
- 이 시기에 다양한 외부 자극을 접하고 소화하며 아이들의 두뇌가 성장
- 가장 중요한 외부 자극 중 하나가 '부모의 말'

Why Instagram?



아이와의 대화 비교군은 어디서 구할 수 있을까?

데이터가 범람하는 시대이지만 한 가정에서 아이가 성장하는 동안의

대화내용을 담은 음성파일 데이터를 구하기 쉽지 않다.

얻고자 하는 데이터와 관련성이 있는 유사데이터를 이용하여 데이터의

양적 분석을 통한 질적 분석의 퀄리티 향상을 도모 부모의 육아일기가

아이들과 대화 과정에서 한 아이와 부모간 생긴 일을 담는 경우가 많다.

일기는 어디서 구하는게 좋을까?

#육아일기 태그 관련글 1487만글이 있는 인스타그램

—da_in_ · 팔로우

—da_in_ #육아일기 #일상

대뜸 툭 던졌던 말에..

“다인아”

“응?”

“사랑해”

“그럴줄 알았어”

“어떻게 알았어?”

“나도 사랑하니깐 다 알지”

진짜 이아이 마음속엔 뭐가 들었을까요 ♥♥

#너때문에행복해 #5세아기

#그래도잠은일찍자줄래

2주

ann_iberriss ❤️😊😊😊😊😊

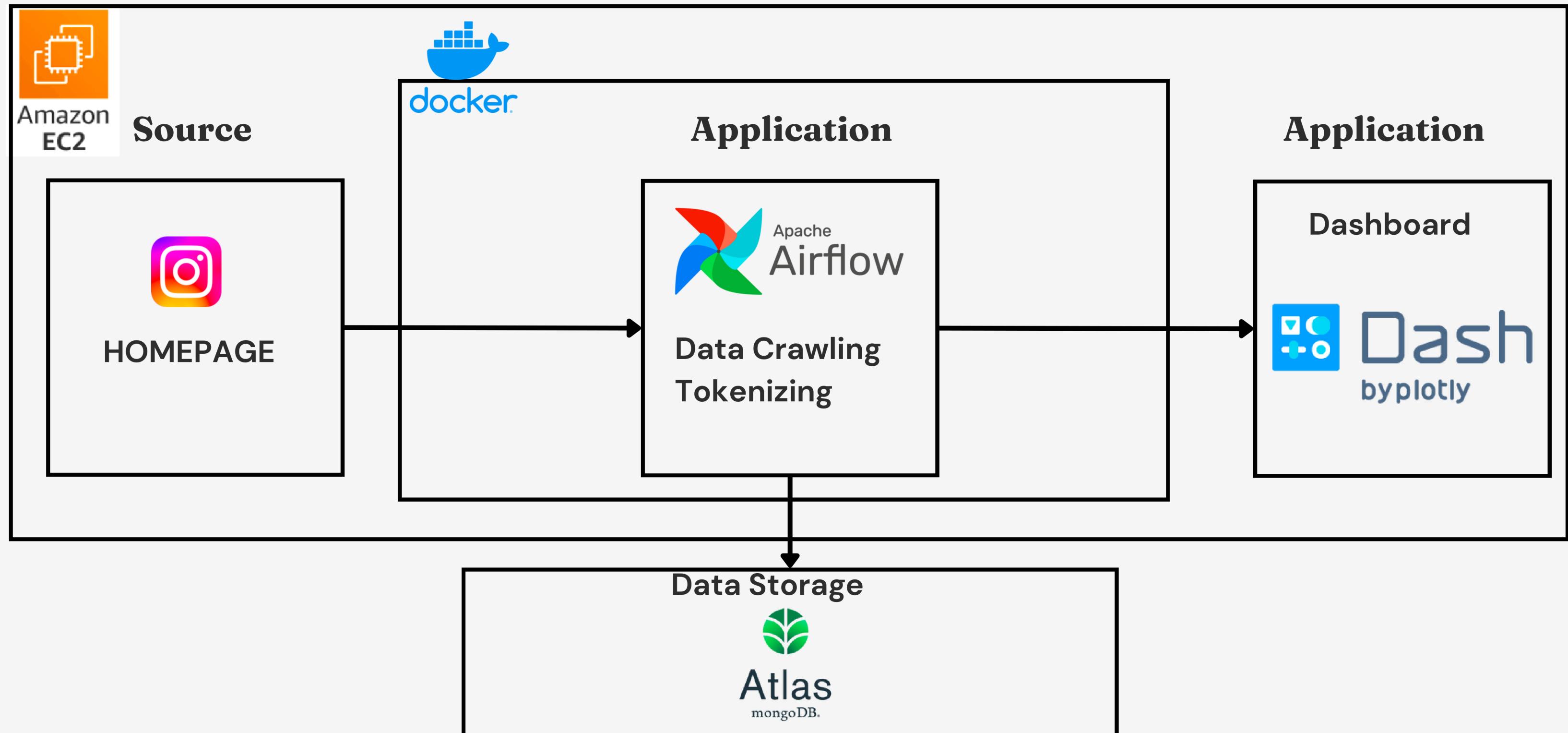
2주 답글 달기

srsccsw ❤️

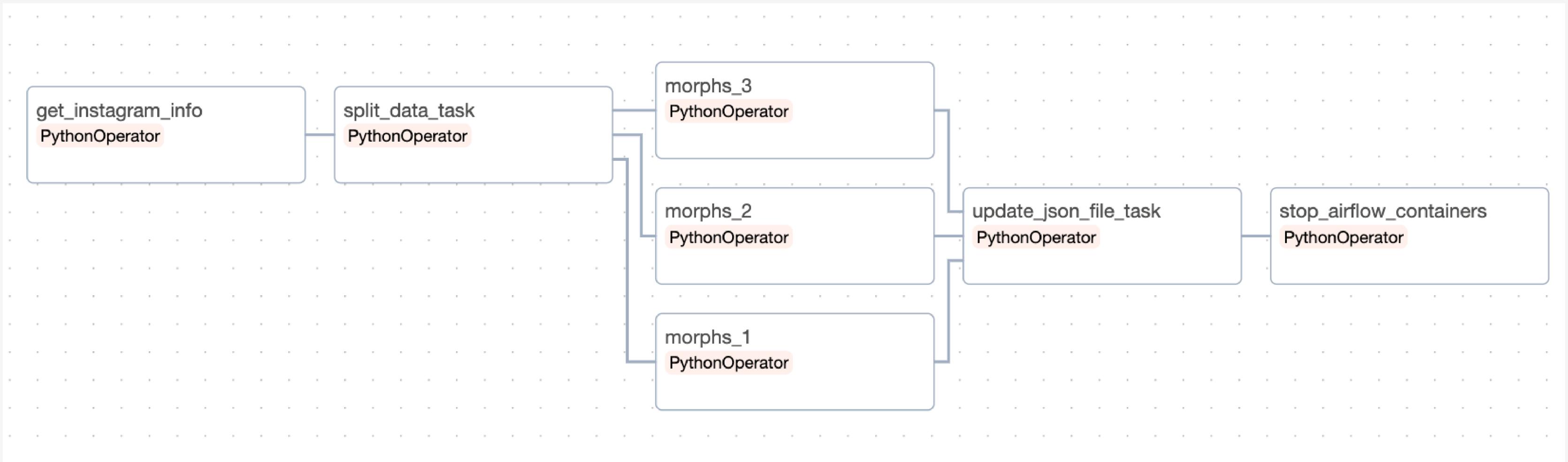
2주 답글 달기

'#그림일기 #나글놀이학교 \n#에듀블룸 #육아일기 #커피',
'요즘 서은이 갓베이비때 사진\n보면서 지금은 작아져서 못입는\ne복들 재고 찾아서 다시 입혀 보고 있다😊\n.'
'2023.10.07.토\n우리집 두찌 처음으로 이 뽑은날^^\n언니는 7살때 처음 뽑았는데...\n년.. 월 잘 안씹어서
'승후의 첫 크루즈\U0001fa75\n\n뚱한 표정이 트레이드 마크인 김뚱후\U0001fae5\n승후야 좋은거 맞지?😊\n'
'오공베이킹DAY❤\n빵킬러 오공이 집에서 거리가 좀잇는 빵집이라,,\n집에서만들어뿌기😍\n어서와,, 찰깨빵? 깨찰빵
. \n엄마 배 위에서 하는 터미타임은 꿀잼이징★\n노느라 정신팔려 잠도 이기고 맘마도 대충 먹는 너\n아주 잘 드
'D+66\n타이니모빌이 고장나서 디즈니 모빌에.. \n\nto끼띠아들 #토끼띠 #육아 #육아스타그램 #토끼띠맘소통
'나의수요일에는\n이른 점심\xa0청담\n우니파스타\xa0해물떡볶이\xa0청담맛집\n우니를 뽀시는게 아니었어~😊\n'*\n한걸음 먼저 사회탐방 \n세상에 태어나 나를 알고 가족을 알고\n그 다음 알아가고 함께 살아가는 ‘사회’\n"으\n핸드폰 바꾼지\n6개월만에\n사진정리가\n끝나간다 ,,\n너무 이뻤던\n2020년\n우래기들 '.'\n'코감기를 오래 앓아서인지 키가 커서인지\n요몇일 허벅지랑 종아리가 눈에 띄게 얇아져서 살짝 아쉬웠는데 뱃깥은
'감기 시작 😊\n기관지가 약한 울애킹,, 쪼꼼만 아프고 지나가자-!!',
'#20231101 #49일\n맨바닥에서 하는 터 미 타 임~~~\n집중하는 눈\ne\n힘들어지면 침이 뚝뚝 떨어진다.\n' #육아용품 #한샘샘키즈 #내돈내산 \U0001fa77\n진짜 엄마들 대란템인건 다 이유가 있더라구요..\n'밥알파티 시~~~~~작((\u263a\u263a\u263a))\n#육아소통 #육아스타그램 #육아기록 #육아맘 #육아맞팔 #육아일상 #육아 #
'*\n오늘은 마미북클럽 날 \n11월 한 달은 오은영 박사님의\n‘어떻게 말해줘야 할까’ 책을 함께 읽어요
'#생후105일 #육아일기\n아이고오~~ 이제는 밥먹고 소화도 안됐는데 역방구에 눕히기만하면 뒤집으려하는 우리 솔
'2시간 푹 자고 일어나서 기분이 매우 죄음\ne\n귀여워....훗😊',
'미용실 브금에 몸을 맡기기\ne',
'\ud83d\udcbb+278\n도레미파솔라시도\ud83d\udcbb',
'앞니가 흔들려 불편하다하여 밤에 뽑는데 아랫니보단 아팠는데 뽑고 왕왕😭\nㅋㅋㅋㅋㅋ그치만 금새 회복하고 웃어보
'엄마가 로션발라주니까\n잠이 솔솔와요😊😊\nD+53',
'날 위한 필라테스\ud83d\udcbb\n아기 낳으면 난 절대 아기꺼는 사고 내꺼는 안사고 이리지 말아야지 했다. 사고\n'2023.11.15 + 100days \n세상에서 제일 소중한 내 아가 \n도하야 엄마아빠에게 와줘서 고마워\n너의 모든
'Week 6\nEvelyn went to church for the first time. It was such a hassle getting ready for church because Evelyn is still so young and doesn't understand what's happening. But we tried our best to make it a positive experience for her.'

Pipeline Architecture



Airflow Graph



crontab- 8 * * * /usr/local/bin/docker-compose -f /home/ubuntu/airflow-docker/docker-compose.yml up -d

Error

- 자동입력방지

안전을 위해 자동 입력 방지 문자를
입력해 주세요



VF9m2tu

- 계정 차단

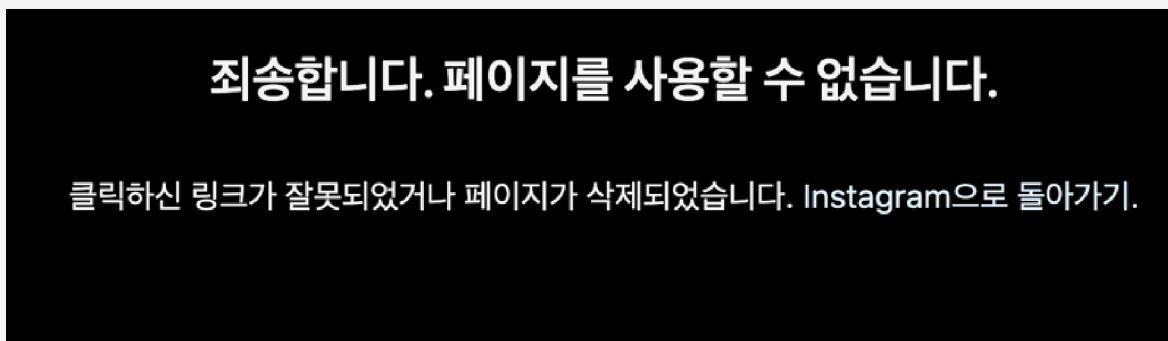
2022년 10월 31일에 회원님의
계정을 일시 차단했습니다

이 결정에 이의를 제기할 수 있는 기간이 30일
남았습니다.

team08 앱 오후 1:34
date : 2023-12-14
alert :
Fail!
task id : morphs_info,
dag id : get_instagram_info,
log url : http://43.201.124.210:8080/log?execution_date=2023-12-14T04%3A31%3A45.057548%2E
date : 2023-12-14
alert :
Success!
task id : morphs_info,
dag id : get_instagram_info,
log url : http://43.201.124.210:8080/log?execution_date=2023-12-14T04%3A36%3A13.568055%2E

Error

- ID 비활성화



- 상단고정



```
try:  
    select_First(driver)  
    driver.implicitly_wait(15)  
    data = get_content(driver)  
    date = data['upload_date']  
  
except:  
    delete_id(user_id)  
    print(f"{user_id} does not exist.")  
    continue  
  
for post in range(3):  
    if date == before_one_day:  
        results.append(data)  
    move_next(driver)  
    driver.implicitly_wait(15)  
    data = get_content(driver)  
    date = data['upload_date']  
  
  
while date >= before_one_day:  
    if date == before_one_day:  
        results.append(data)  
    move_next(driver)  
    driver.implicitly_wait(15)  
    data = get_content(driver)  
    date = data['upload_date']
```

Problem

```
def update_graph(selected_pos):
    morphs_type = ['NNG', 'NNP', 'NP', 'Verb', 'Adjective', 'Adverb', 'Conjunction', 'Josa', 'Number']
    # MongoDB에서 선택한 데이터 가져오기

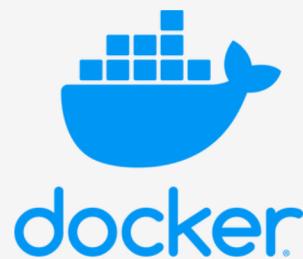
    data = list(collection.find({}, {"selected_pos": 1, "_id": 0, "user_name": 1}).limit(100))
```

- 기존 DB가 Update 되면 바로 적용이 되게 하기 위해 dash코드에서 함수별로 DB를 바라보고 있게 만듦
- 데이터를 조회하고 찾아오는 시간이 너무 오래걸림

```
#Json 데이터 로드 함수 정의
def load_data():
    with open('/home/ubuntu/airflow-docker/dags/services/insta_crawling_morphs.json', 'r', encoding='utf-8') as f:
        data = json.load(f)
    return data
```

- Airflow에 daily_data를 기존 .json 파일에 추가하는 task 추가.
- .json 파일을 불러와서 처리하는 방식으로 바꿈

Why



AWS 인스턴스를 개발환경으로 하기엔 비용적인 부분이 비효율적이다.
로컬환경과 인스턴스의 환경을 동일하게 세팅하고 설치하기에는 큰 번거로움이 있기에
도커 컨테이너를 사용함으로 환경의 일관성을 유지하여 애플리케이션 배포를 용이하게 함.



Task 구성은 Task1 : Crawling과 Task2: Tokenizing 으로 구성되어 있으며 Crawling이 완료를
하면 Tockenizing을 하는 Task들 간의 의존성과 Task를 병렬처리를 구성하기 유용하고,
Task를 처리한 후 DB를에서 다시 조회를 하지않고 작업 간에 데이터를 Xcom으로 불러와
처리시간을 줄임.



Atals는 MongoDB의 관리형 클라우드 서비스로, MongoDB 데이터베이스의 전체적 관리
(배포, 운영, 백업 및 보안 등)을 맡아주므로 데이터베이스 운영에 대한 부담을 덜어
데이터 분석에 개발자가 더 집중할 수 있게 도움

Database(MongoDB)

The screenshot shows the MongoDB Atlas interface for the 'Team08' project. The top navigation bar includes 'CNLAB > PROJECT 0 > DATABASES', the project name 'Team08', and system status 'VERSION 6.0.12 REGION AWS Seoul (ap-northeast-2)'. Below this are tabs for 'Overview', 'Real Time', 'Metrics', 'Collections' (which is selected), 'Search', 'Profiler', 'Performance Advisor', 'Online Archive', and 'Cmd Line Tools'. A 'CREATE COLLECTION' button is visible on the right.

The 'Collections' tab displays the 'ConnectsLab' database with two collections: 'insta_crawling' and 'insta_crawling_morphs'. The 'insta_crawling' collection has 29582 documents, logical data size of 15.62MB, and storage size of 9.32MB. The 'insta_crawling_morphs' collection has 29582 documents, logical data size of 28.32MB, and storage size of 12.54MB.

The 'Data Services' tab shows the 'ConnectsLab.insta_crawling' service with 29582 total documents. It includes tabs for 'Find', 'Indexes', 'Schema Anti-Patterns', 'Aggregation', and 'Search Indexes'. A 'Type a query' input field contains the query '{ field: 'value' }'. The results section displays 'QUERY RESULTS: 1-20 OF MANY' with two document snippets:

```
_id: ObjectId('6566c962a01b97c6e2cb63c7')
user_name: "to.gangnang"
like_count: "315"
caption: "낭이 루돌프가 파는 성냥 사두세여~🙏🎄"
upload_date: "2023-11-17"

_id: ObjectId('6566c962a01b97c6e2cb63c8')
user_name: "to.gangnang"
like_count: "93"
caption: "#힙한 강남이 백일상 사진을 셀프 활영만 해서 아쉬웠었는데디어제이스튜디오에서 좋은 기회를 주셔서 멋진 백일상 사진을 남기게 되~"
upload_date: "2023-11-22"

_id: ObjectId('6566c962a01b97c6e2cb63c9')
user_name: "to.gangnang"
```

Navigation buttons at the bottom include 'PREVIOUS', 'NEXT', and '1-20 of many results'.

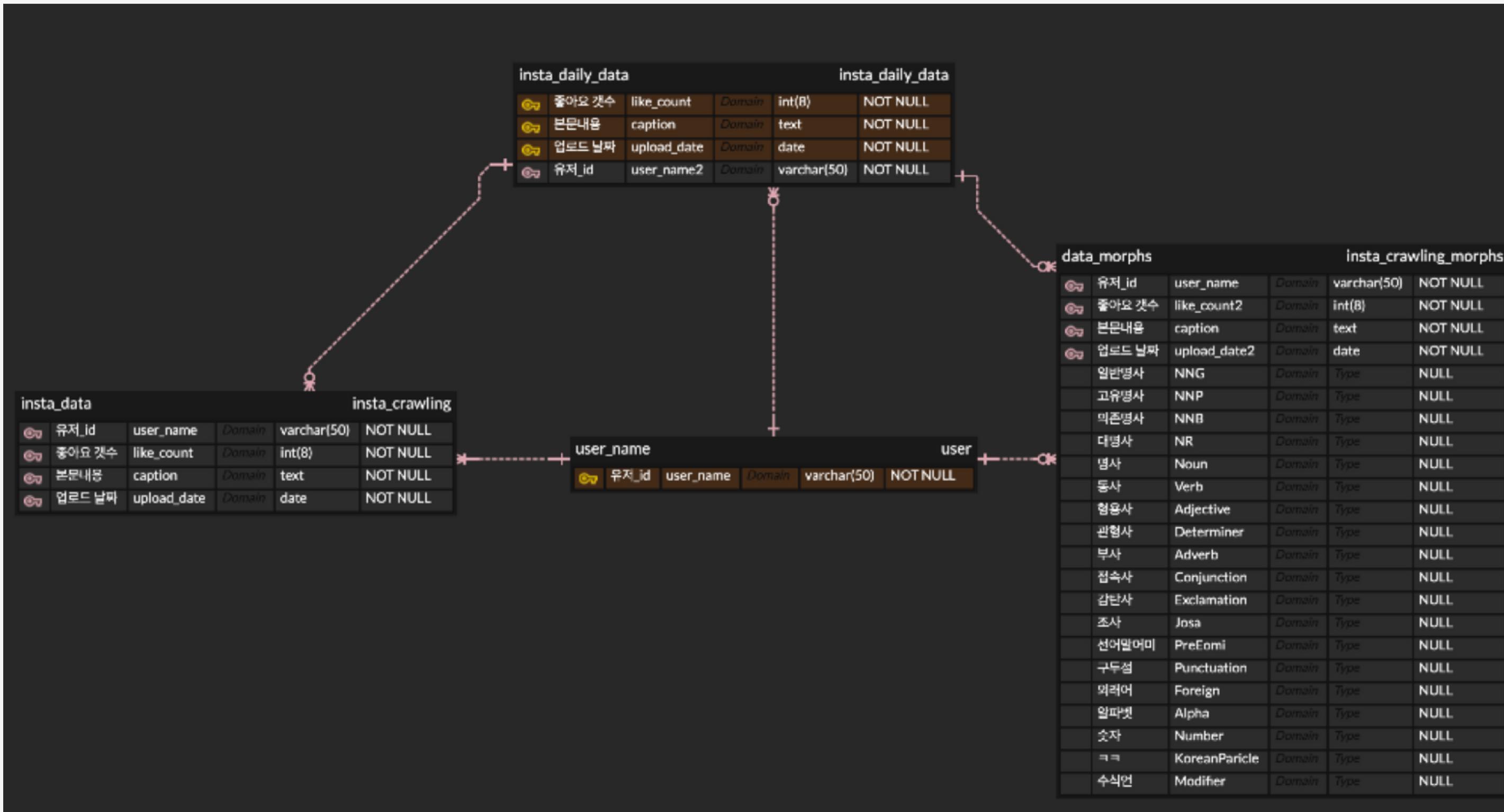
QUERY RESULTS: 1-20 OF MANY

```
_id: ObjectId('6566c962a01b97c6e2cb63c7')
user_id: "to.gangnang"
content: "낭이 루돌프가 파는 성냥 사두세여~🙏🎄"
date: "2023-11-17"
like: "315"
```

```
_id: ObjectId('6566c7faa01b97c6e2caf039')
user_id: "to.gangnang"
```

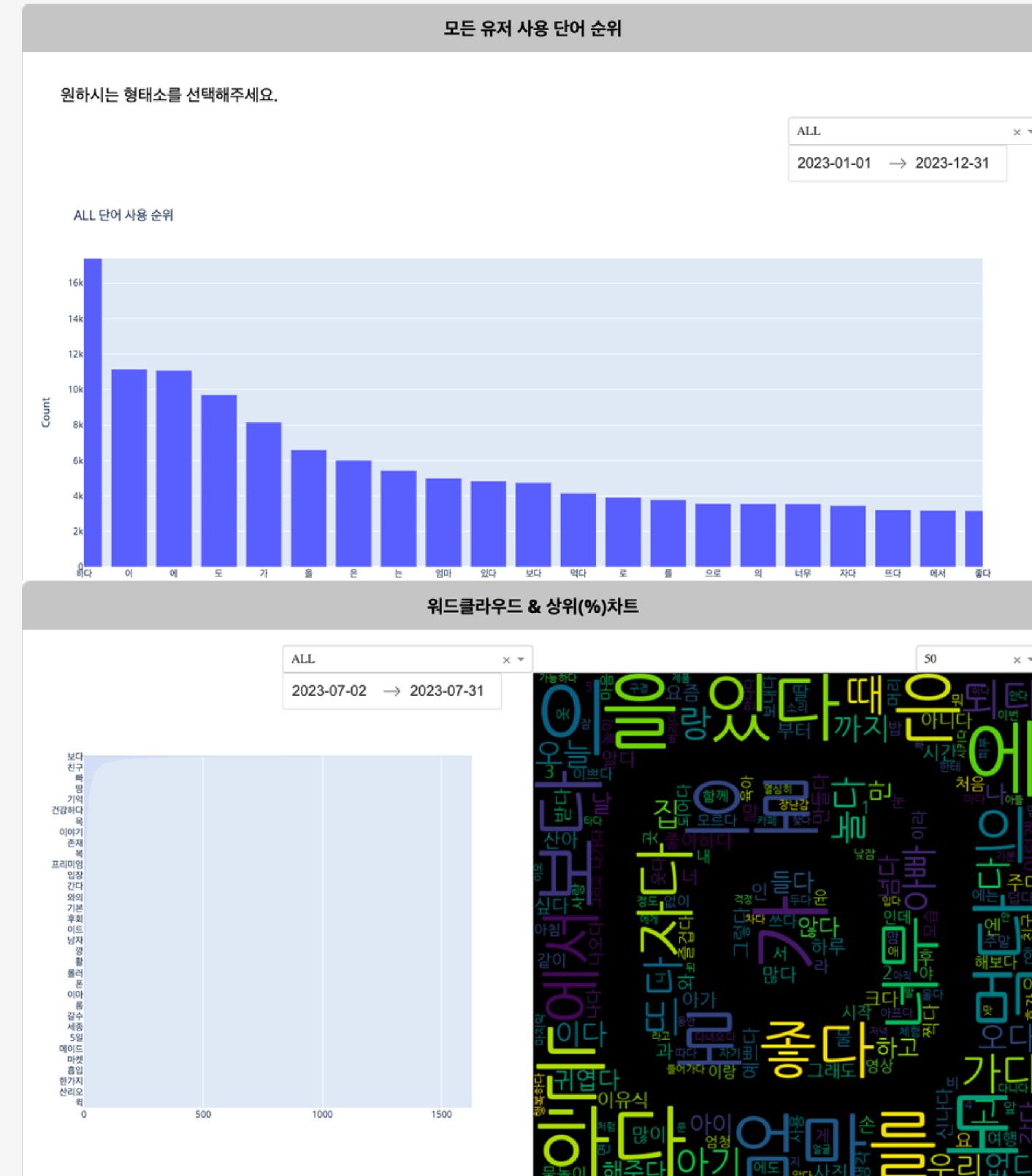
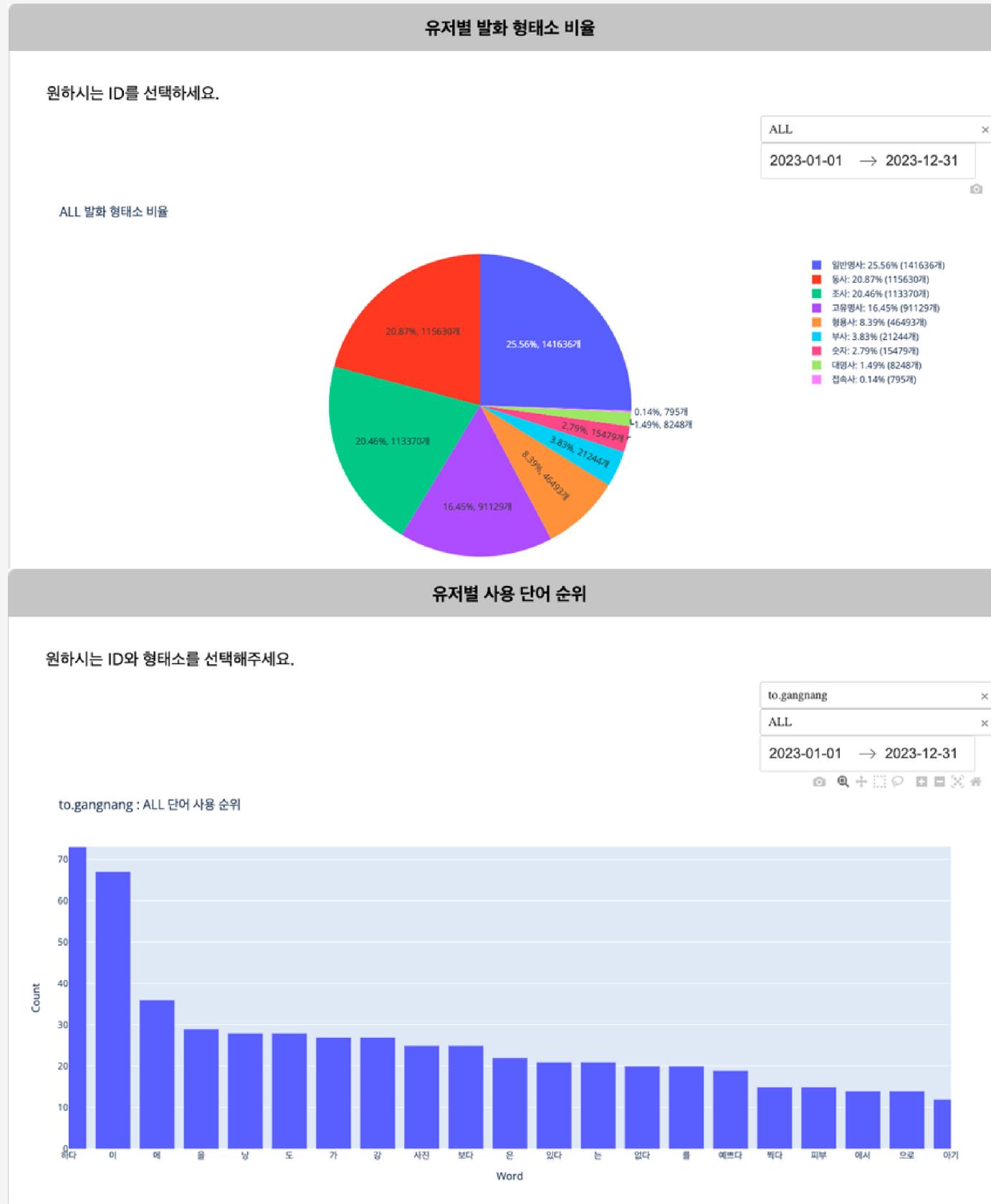
- ▶ Noun: Array (4)
 - ▶ Verb: Array (1)
 - ▶ Adjective: Array (empty)
 - ▶ Determiner: Array (empty)
 - ▶ Adverb: Array (empty)
 - ▶ Conjunction: Array (empty)
 - ▶ Exclamation: Array (empty)
 - ▶ Josa: Array (2)
 - ▶ PreEomi: Array (empty)
 - ▶ Punctuation: Array (empty)
 - ▶ Foreign: Array (empty)
 - ▶ Alpha: Array (empty)
 - ▶ Number: Array (empty)
 - ▶ KoreanParticle: Array (empty)
 - ▶ Modifier: Array (2)
- date: "2023-11-17"
like: "315"

Database



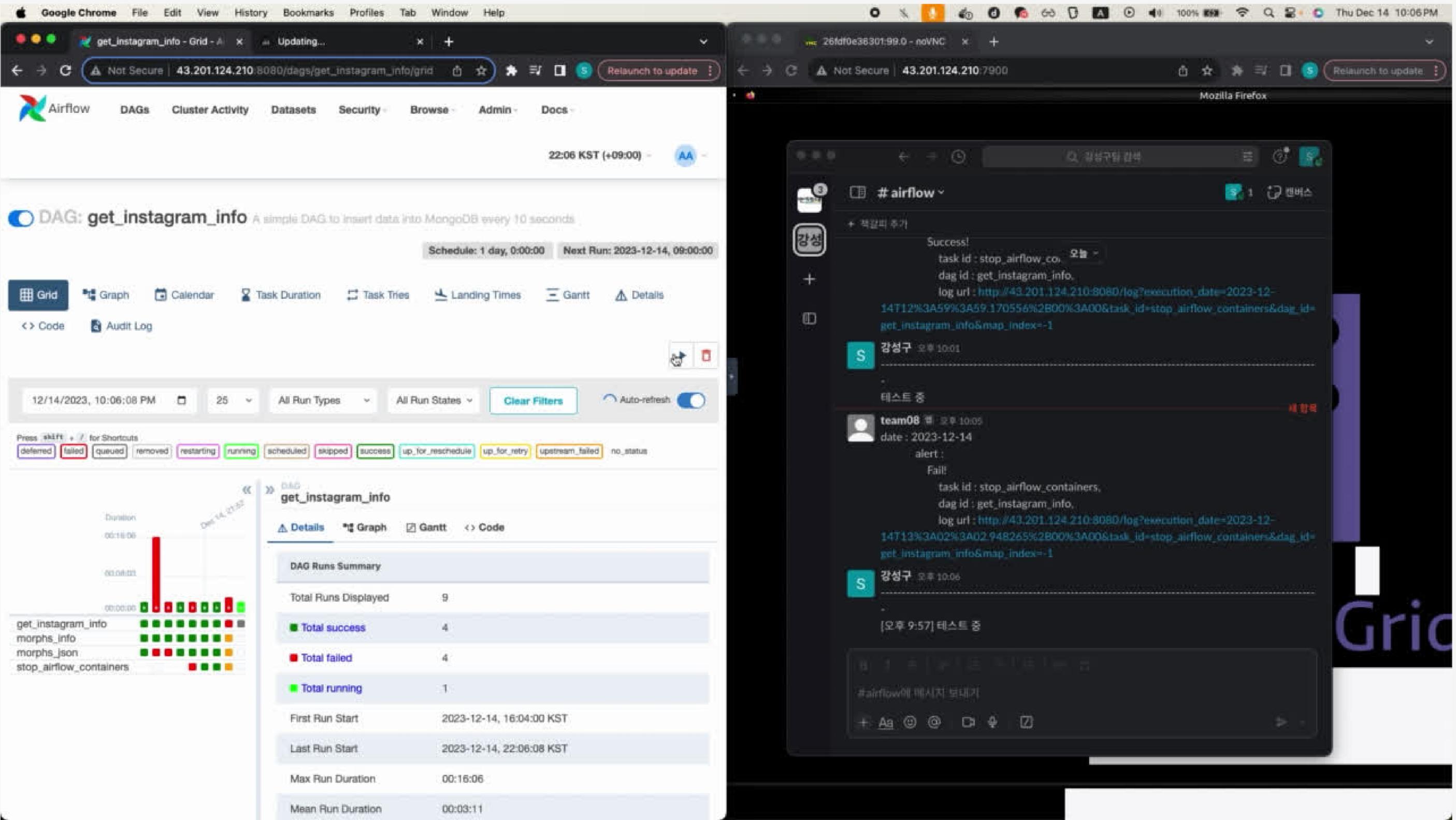
- Mongo DB(NOSQL) 라서 관계형 DB가 아니지만 MongoDB 구조를 시각적으로 보여주기 위한 ERD

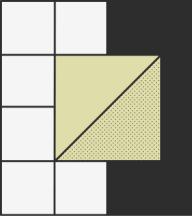
dashboard



dashboard 주소 : <http://43.201.124.210:8061/>

Demo





Thank you
for listening

