



데이터 처리 과정



학습내용

- 1 데이터 처리 과정
- 2 실 세계 데이터셋의 처리 과정 예시

학습목표

- 데이터 처리 과정을 확인할 수 있다.
- 본 과정에서 다루는 실 세계 데이터셋의 기본적인 처리 과정을 확인할 수 있다.



Data Up

데이터 수집

데이터 가공

데이터 저장

데이터 수집





데이터 수집

API 호출

센서 신호 읽기

파일 읽기

데이터베이스 접근



데이터 소스에 따라
데이터 수집 방법 상이



데이터 수집



개발자들에게
API를 통해 데이터를 제공함

데이터 수집

그래프 API 탐색기

액세스 토큰:

→

Edge: me/
☒ email
☒ id
☒ name

```
{  
  "email": "  
  "id": "  
  "name": "  
}
```

본인의 email,id,name 획득


/me? fields=email,id,name + 액세스 토큰

그래프 API 구문에 대해 더 알아보기

SNAP (Stanford Network Analysis Project)

50개 이상의 네트워크 데이터셋 제공

By Jure Leskovec
STANFORD UNIVERSITY



Stanford Large Network Dataset Collection

- Social networks** : online social networks, edges represent interactions between people
- Networks with ground-truth communities** : ground-truth network communities in social and information networks
- Communication networks** : email communication networks with edges representing communication
- Citation networks** : nodes represent papers, edges represent citations
- Collaboration networks** : nodes represent scientists, edges represent collaborations (co-authoring a paper)
- Web graphs** : nodes represent webpages and edges are hyperlinks
- Amazon networks** : nodes represent products and edges link commonly co-purchased products
- Internet networks** : nodes represent computers and edges communication
- Road networks** : nodes represent intersections and edges roads connecting the intersections
- Autonomous systems** : graphs of the internet
- Signed networks** : networks with positive and negative edges (friend/foe, trust/distrust)
- Location-based online social networks** : social networks with geographic check-ins
- Wikipedia networks, articles, and metadata** : talk, editing, voting, and article data from Wikipedia
- Temporal networks** : networks where edges have timestamps
- Twitter and Memetracker** : memetracker phrases, links and 467 million Tweets
- Online communities** : data from online communities such as Reddit and Flickr
- Online reviews** : data from online review systems such as BeerAdvocate and Amazon
- User actions** : actions of users on social platforms.
- Face-to-face communication networks** : networks of face-to-face (non-online) interactions
- Graph classification datasets** : disjoint graphs from different classes

SNAP networks are also available from [SuiteSparse Matrix Collection](#) by Tim Davis.

- SNAP for C++
- SNAP for Python
- SNAP Datasets
- BIOSNAP Datasets
- What's new
- People
- Papers
- Projects
- Citing SNAP
- Links
- About
- Contact us

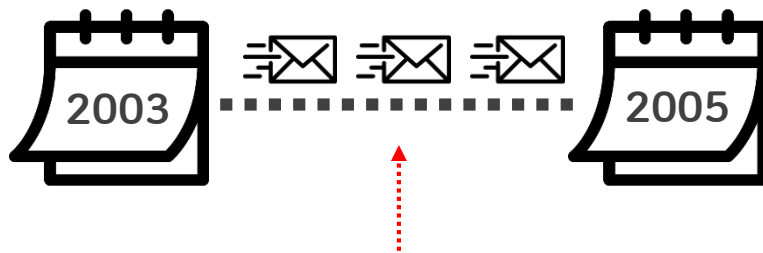
Open positions

Open research positions in SNAP group are

데이터 수집

이메일 데이터

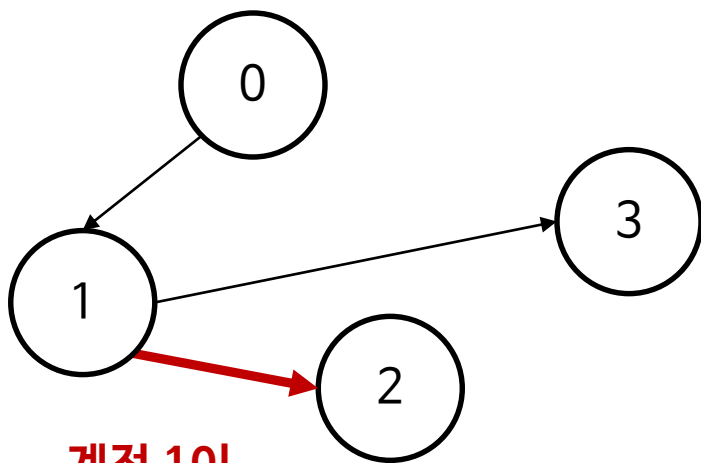
<http://snap.stanford.edu/data/email-EuAll.html>



이메일을 보낸 사실에 대한 기록 획득

데이터 수집

이메일 데이터



계정 1이
계정 2에게
이메일을 보냄



# 데이터 명세	
# 데이터 명세	
# ...	
0	1
1	2
1	3
...	

데이터 가공

데이터 저장을 위한 형태로 데이터 처리



추상화

파싱

검증

정제

데이터 가공

# 데이터 명세	
# 데이터 명세	
# ...	
0	1
0	1
0	
*1912L	5
0	8
0	11
0	20
0	48
0	430
...	

----- 데이터 명세 제거

----- 중복 데이터 제거

----- 누락 데이터 제거

----- 결함 데이터 제거

----- 처리할 데이터

"0" "20"

----- 토큰화 -----

"0"
"20"

----- 정수변환 -----> 0
20



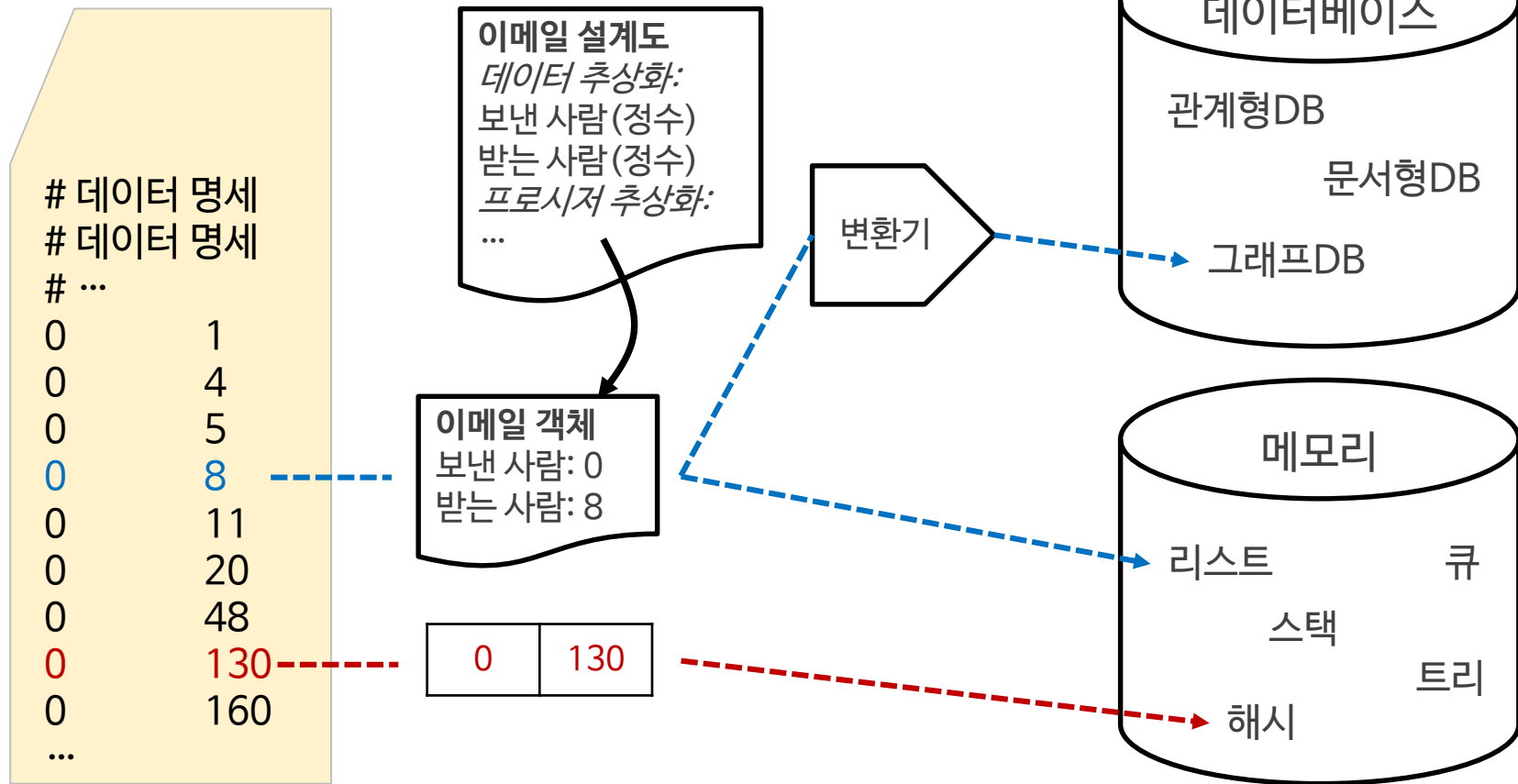
데이터 저장

데이터를 특정 목적의 서비스를
만들기 적절한 형태로 유지

메모리

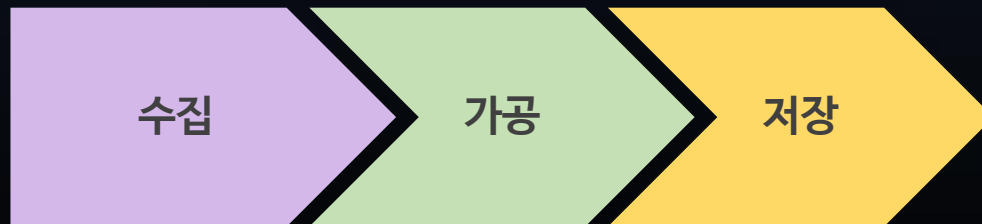
데이터
베이스

데이터 저장



Remind

데이터 처리 과정



#01 위키피디아, 2021, URL : <https://en.wikipedia.org/wiki/Facebook>

#02 developers, 2021, URL : <https://developers.facebook.com/tools/explorer>

#03 snap, 2021, URL : <http://snap.stanford.edu/data/email-EuAll.html>

#04 flaticon, 2021, URL : https://www.flaticon.com/free-icon/calendar_2838779

#05 flaticon, 2021, URL : https://www.flaticon.com/free-icon/mail-send_91848