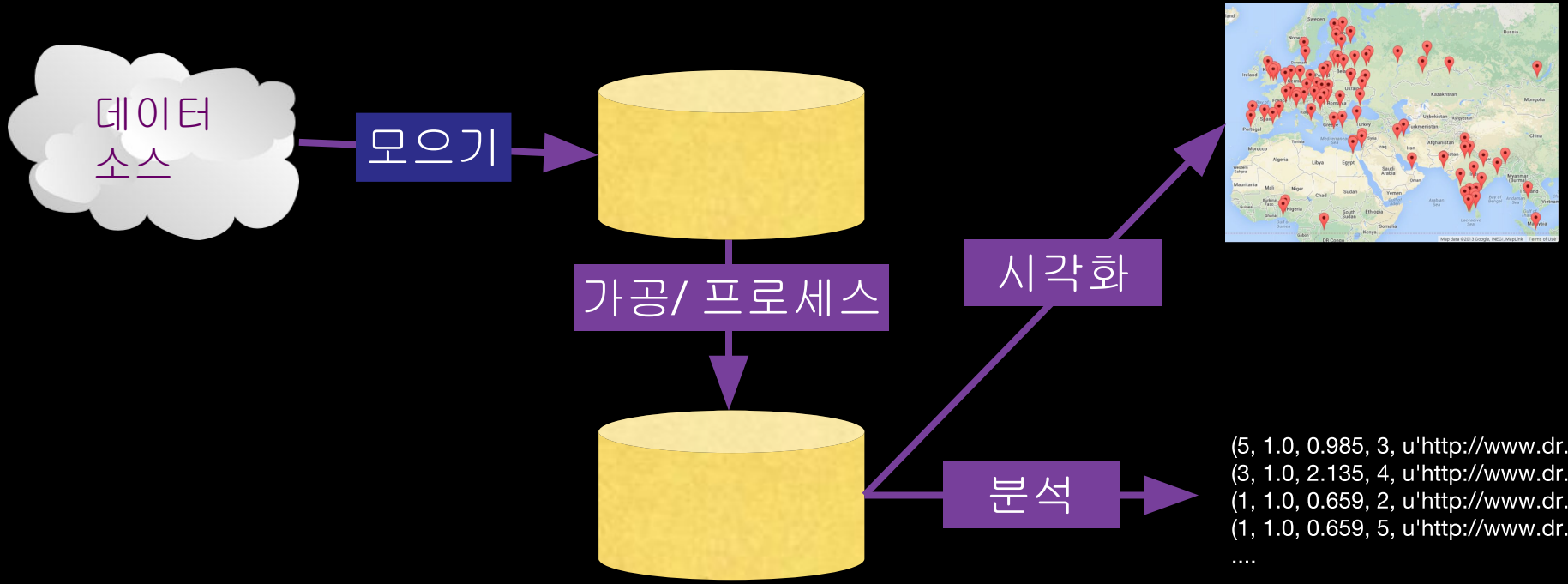
데이터 검색과 시각화

Charles Severance

모두를 위한 파이썬

www.py4e.com

다단계 데이터 분석


데이터
소스
모으기
가공/ 프로세스
시각화
분석
(5, 1.0, 0.985, 3, u'http://www.dr..')
(3, 1.0, 2.135, 4, u'http://www.dr..')
(1, 1.0, 0.659, 2, u'http://www.dr..')
(1, 1.0, 0.659, 5, u'http://www.dr..')
....

다양한 데이터 마이닝 기술

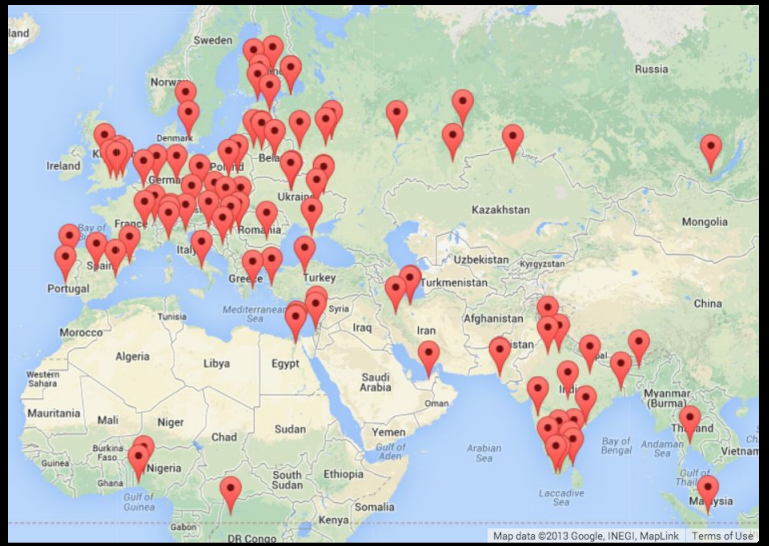
- https://hadoop.apache.org/
- http://spark.apache.org/
- https://aws.amazon.com/redshift/
- http://community.pentaho.com/
- ....

“개인 데이터 마이닝"

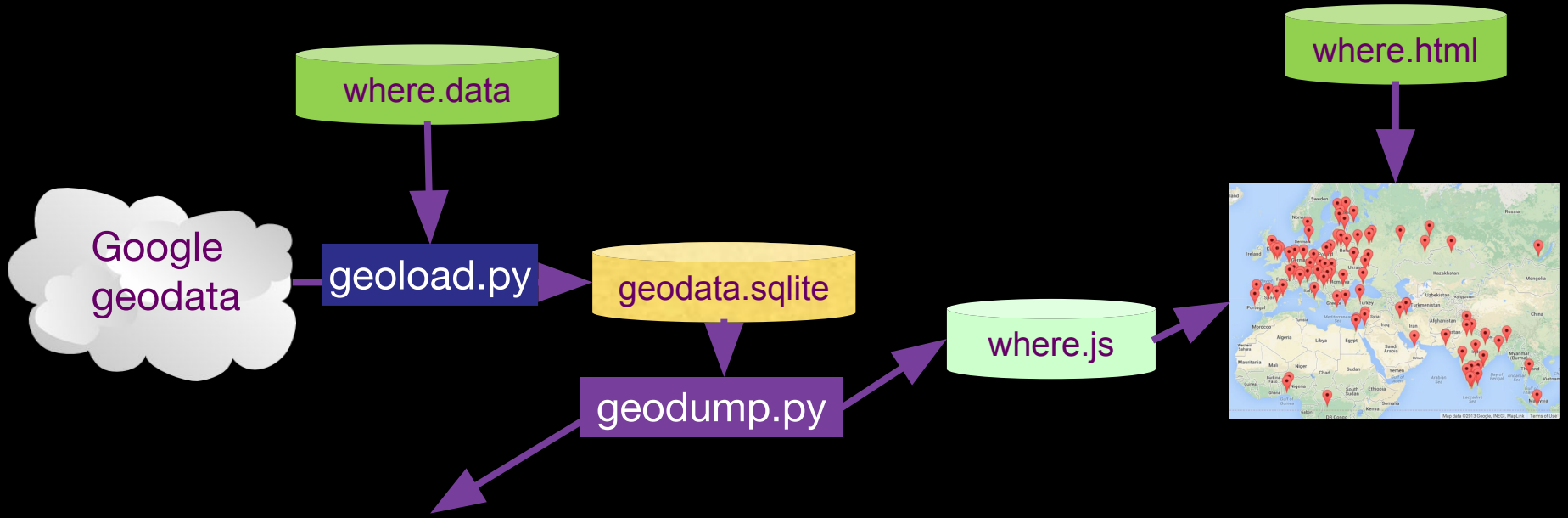
여러분을 더 좋은 프로그래머로 만드는 것이 우리의 목표 – 데이터 마이닝 전문가로 만드는 것은 아님

GeoData

- 유저가 입력한 데이터로 구글 맵을 만듦
- Google Geodata API를 사용
- 속도 제한을 피하고 재시작을 위해 데이터베이스에 저장
- Google Maps API을 사용해 브라우저에 시각화



http://www.py4e.com/code3/geodata.zip

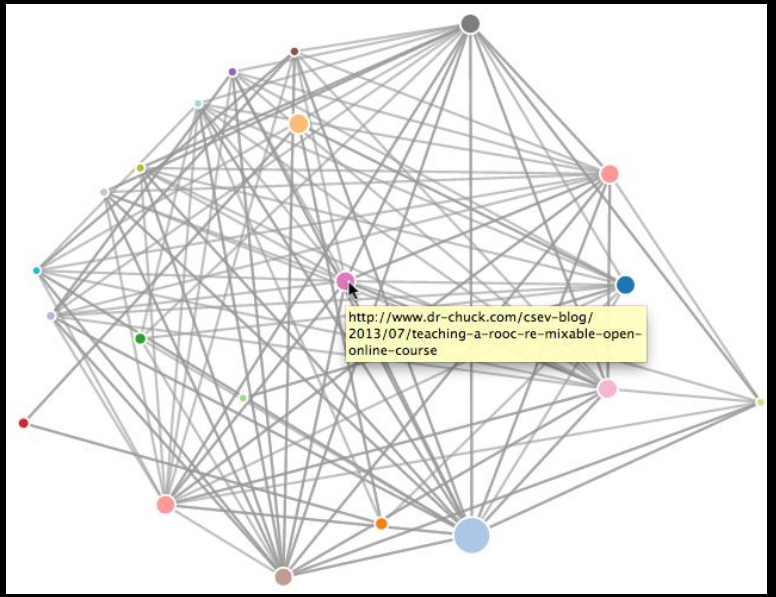



```
Northeastern University, ... Boston, MA 02115, USA 42.3396998 -71.08975
Bradley University, 1501 ... Peoria, IL 61625, USA 40.6963857 -89.6160811
...
Technion, Viazman 87, Kesalsaba, 32000, Israel 32.7775 35.0216667
Monash University Clayton ... VIC 3800, Australia -37.9152113 145.134682
Kokshetau, Kazakhstan 53.2833333 69.3833333
...
12 records written to where.js
Open where.html to view the data in a browser
```

http://www.py4e.com/code3/geodata.zip

Page Rank

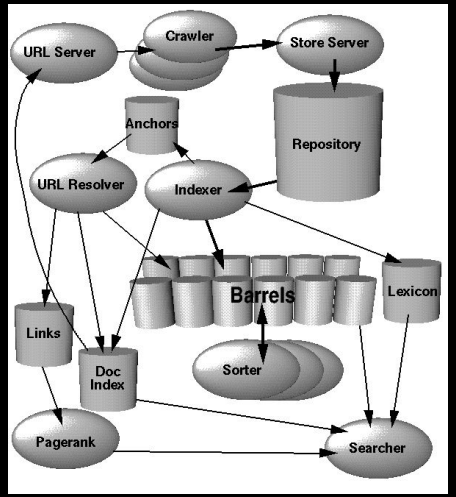
- 간단한 웹 페이지 크롤러 작성
- Google's Page Rank
  알고리즘의 간단한 버전을 계산
- 결과 네트워크를 시각화



http://www.py4e.com/code3/pagerank.zip

검색 엔진 아키텍쳐

- 웹 크롤링
- 색인 구축
- 검색



http://infolab.stanford.edu/~backrub/google.html

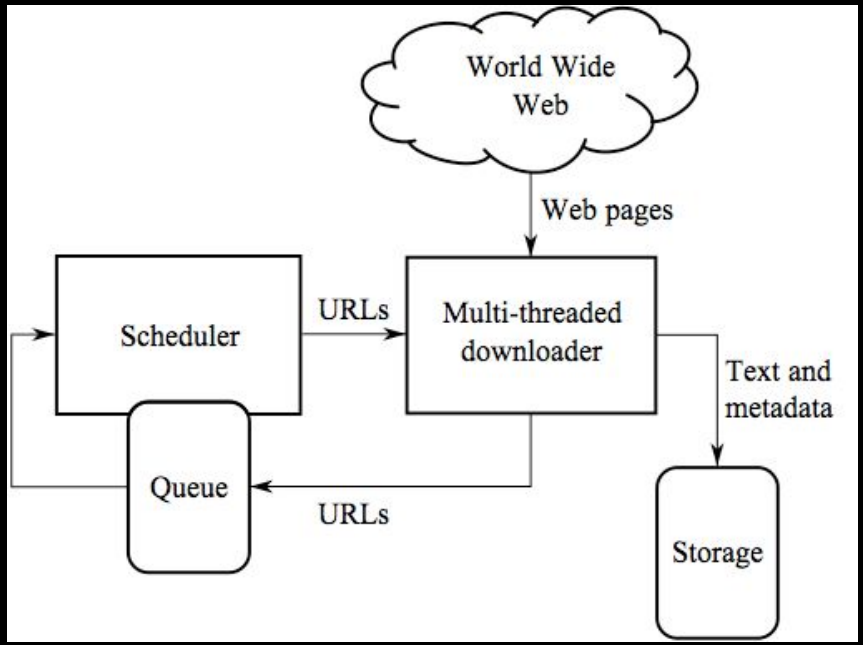
웹 크롤러

웹 크롤러는 체계적이고 자동적인 방식으로
World Wide Web을 열람하는 컴퓨터 프로그램.
웹 크롤러는 주로 빠른 검색을 제공하기 위해 다운로드된
페이지를 색인하는 검색 엔진에서 쓰이는데, 모든 방문한
페이지의 복사본을 생성하게 됨.

http://en.wikipedia.org/wiki/Web_crawler

Web Crawler

- 페이지 검색
- 페이지의 링크를 찾음
- 앞으로 검색될 사이트 목록에 링크 추가
- 반복...



http://en.wikipedia.org/wiki/Web_crawler

웹 크롤링 정책

- selection policy는 어떤 페이지를 다운로드할지,
- re-visit policy는 페이지의 변화를 언제 확인할지,
- politeness policy는 웹사이트 과부하를 어떻게 피할지,
- parallelization policy는 분산된 웹 크롤러를 어떻게 조정할지를 제시한다

robots.txt

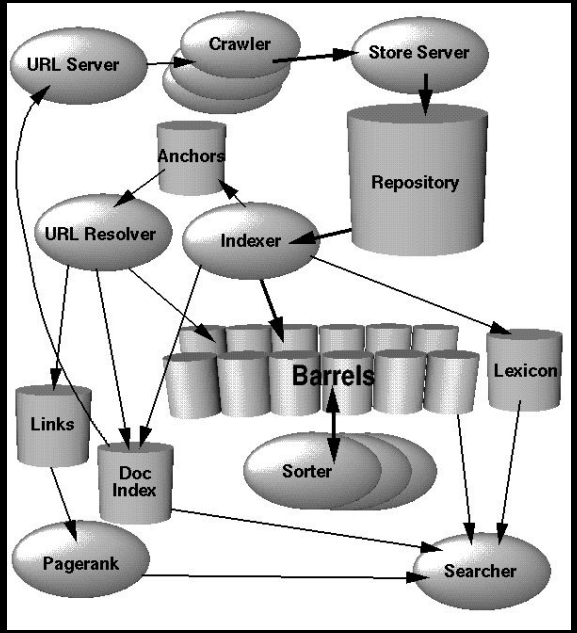
- 웹사이트가 웹 크롤러와 소통하는 방법
- 비공식적이고 자발적인 표준
- 때때로 사람들은 “나쁜" 거미를 잡기 위해 “거미 덫"을 놓음

```
User-agent: *
Disallow: /cgi-bin/
Disallow: /images/
Disallow: /tmp/
Disallow: /private/
```

http://en.wikipedia.org/wiki/Robots_Exclusion_Standard
http://en.wikipedia.org/wiki/Spider_trap

Google Architecture

- 웹 크롤링
- 색인 구축
- 검색

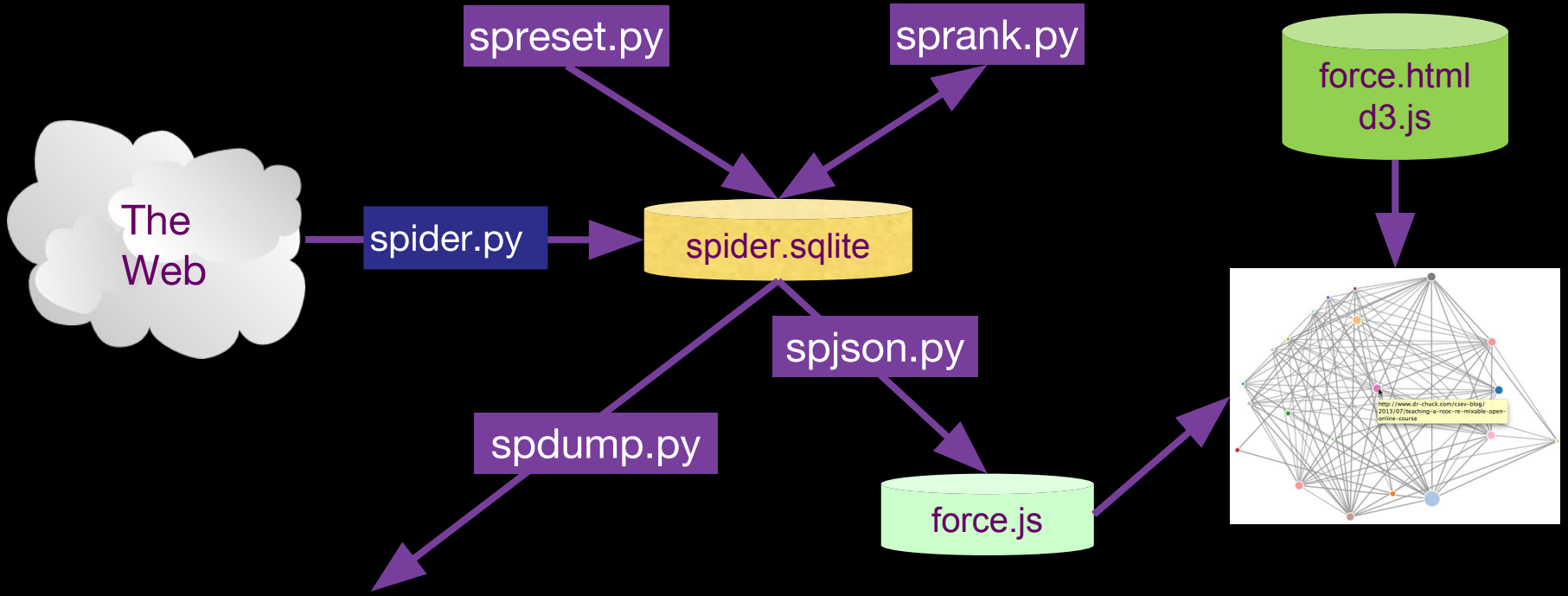


http://infolab.stanford.edu/~backrub/google.html

검색 색인

검색 엔진 색인은 빠르고 정확한 정보의 검색을 위해 데이터를 모으고, 파싱하고, 저장. 색인을 저장하는 목적은 검색 쿼리에 대해서 연관 있는 자료를 찾기 위한 속도와 성능을 최적화하기 위함. 색인이 없다면 검색 엔진은
군집의 모든 자료를 스캔할텐데, 이는 상당한 시간과 컴퓨팅 파워가 필요.

http://en.wikipedia.org/wiki/Index_(search_engine)



(5, None, 1.0, 3, u'http://www.dr-chuck.com/csev-blog')
(3, None, 1.0, 4, u'http://www.dr-chuck.com/dr-chuck/resume/speaking.htm')
(1, None, 1.0, 2, u'http://www.dr-chuck.com/csev-blog/')
(1, None, 1.0, 5, u'http://www.dr-chuck.com/dr-chuck/resume/index.htm')
4 rows.

http://www.py4e.com/code3/pagerank.zip

메일링 리스트 - Gmane

- 메일링 리스트의 아카이브를 크롤링
- 몇 가지 분석과 가공
- 단어 구름과 라인으로 시각화

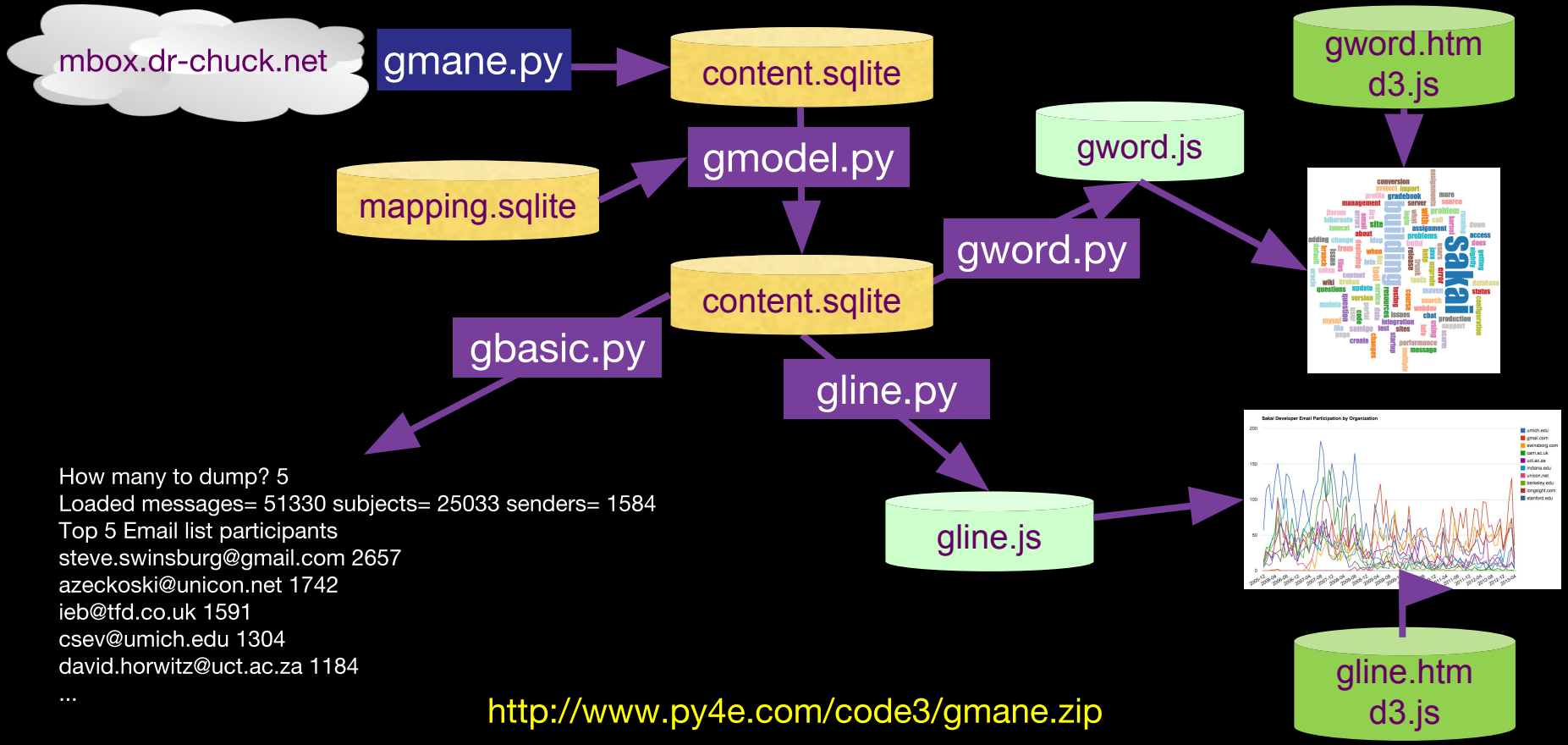
http://www.py4e.com/code3/gmane.zip

경고: 데이터 집합 > 1GB

- gmane.org 로 프로그램을 지정하고 실행하지 마시오
- 속도 제한이 없음 - 좋은 사람들임.

테스트할 때 이것을 사용:

http://mbox.dr-chuck.net/sakai.devel/4/5


mbox.dr-chuck.net
gmane.py
content.sqlite
gword.htm
d3.js
gword.js
gmodel.py
mapping.sqlite
gword.py
content.sqlite
gbasic.py
gline.py
gline.js
gline.htm
d3.js
How many to dump? 5
Loaded messages= 51330 subjects= 25033 senders= 1584
Top 5 Email list participants
steve.swinsburg@gmail.com 2657
azeckoski@unicon.net 1742
ieb@tfd.co.uk 1591
csev@umich.edu 1304
david.horwitz@uct.ac.za 1184
...
http://www.py4e.com/code3/gmane.zip

Acknowledgements / Contributions

Initial Development: Charles Severance, University of Michigan School of Information

Contributor:

- Seung-June Lee (plusjune@gmail.com)
- Connect Foundation

Translator:

- Jaeyi Hong
- Jeungmin Oh (tangza@gmail.com)

...