(5) 환경분야 빅데이터 수집방법연구 [한국진]

- 데이터 중심 연구 패러다임 대응을 위한 빅데이터 수집방법 연구

  ※ 데이터 목록은 환경분야 빅데이터 수집방법 연구(BA2017-06)에서 정리 예정

- 공공데이터 포털 활용신청 순위 기준 기상청, 한국환경공단 사례

  1) (공공) 기상청 동네예보정보조회서비스(최근 24시간)(통합 코드 4건+)

  2) (공공) 한국환경공단 대기오염정보 조회 서비스(통합코드 11건+)

  ※ 기상자료공개포털(코드 3건+), 에어코리아(코드 20건+)

  ※ 원내 데이터 활용사례조사(비슷한 결과) 및 연구자 면담 실시(실익 없음)

- 사례 공개(Github), 시범운영 서버 반영 : 2017년 限

(5.1) 진행상황 및 추가사항

- 진행상황

진 행 내 역	1분기	2분기	3분기	9월	10월	11월	12월
연구방향 설정 / 검토							
빅데이터 서비스 검토							
대상 선정(설문조사 등)							
서비스 / 데이터 분석							
사례 공개(Github) 등							
시범운영 서버 반영 / 공개							

- 추가사항 : 시범운영 서버 자동화 사례 구축

※ 자동화 사례 : 수집-저장-전처리-DB 또는 검색엔진-분석SW 연동 또는 시각화

(5.2) 중간보고 내용 및 추가사항

- 원내 설문조사 등 수요(활용) 환경 빅데이터 (9월) / 심층면담(8월)
- 수집된 데이터 DB연계 방안 마련(8월) / 구축(년내)
- Python(검토)과 Elastic Stack(자동화) 등 오픈소스 제약사항 극복(종료 시까지) : 연구자 DB 접근 또는 검색엔진(시각화) 활용
- ~~OpenAPI 활용시 설정부하(코딩) 절감방안 마련(9월 예정)~~
- ~~환경 빅데이터 분석플랫폼 서버 도입시 탑재(종료시까지)~~
- 가능한 범위 내 자동화 코드 공개 / 사례 구축(종료 시까지)
- 환경 빅데이터 분석플랫폼 구성방안 마련(10월) / ISP 준비(~18년)

(5.3) 환경분야 ㅂ

- (공데 포털) 기상청 동네

1. 초단기실황조회	2. 초단계여

데이터 식별 / 데이터 서비스 탐색

메타데이터 확인 / 오픈API 분석

자동화 코드 작성

- 최근 1일, 1달, 3달만 조회됨
  (실시간 확인의 의미)
- 특정 항목 추출 곤란
  (제공하는 모든 항목을 가져와 전처
- 좌표 체계 활용 곤란
  (5km x 5km 해상도 전국을 분할)

- 파이선 : BeautifulSoup,

```
#contents_html
```

http://newsky2.kma.go.kr/service/SecndSrtpdFrcstInfoService2/ForecastGrib?serviceKey=pOlubQDAbC67WYJ1c3bc9Vc2I8xxRwmQ%2FY2mEkL5brwPCEiLYaf3nduU8J5s%2F5ufCTVh3r270f3O5mvOKoVUnQ%3D%3D&base_date=20171017&base_time=0100&nx=66&ny=103&_type=json

In [8]:
```
json.loads(contents_source)
```

Out[8]:
```
{'response': {'body': {'items': {'item': [{'baseDate': 20171017,
        'baseTime': '0100',
        'category': 'LGT',
        'nx': 66,
        'ny': 103,
        'obsrValue': 0},
       {'baseDate': 20171017,
        'baseTime': '0100',
        'category': 'PTY',
        'nx': 66,
        'ny': 103,
        'obsrValue': 0},
       {'baseDate': 20171017,
        'baseTime': '0100',
        'category': 'REH',
        'nx': 66,
        'ny': 103,
        'obsrValue': 98},
       {'baseDate': 20171017,
        'baseTime': '0100',
        'category': 'RN1',
        'nx': 66,
        'ny': 103,
        'obsrValue': 0},
       {'baseDate': 20171017,
        'baseTime': '0100',
        'category': 'SKY',
        'nx': 66,
        'ny': 103,
        'obsrValue': 1},
       {'baseDate': 20171017,
```

(5.4) 환경분야 빅데이터 수집사례(2/4)

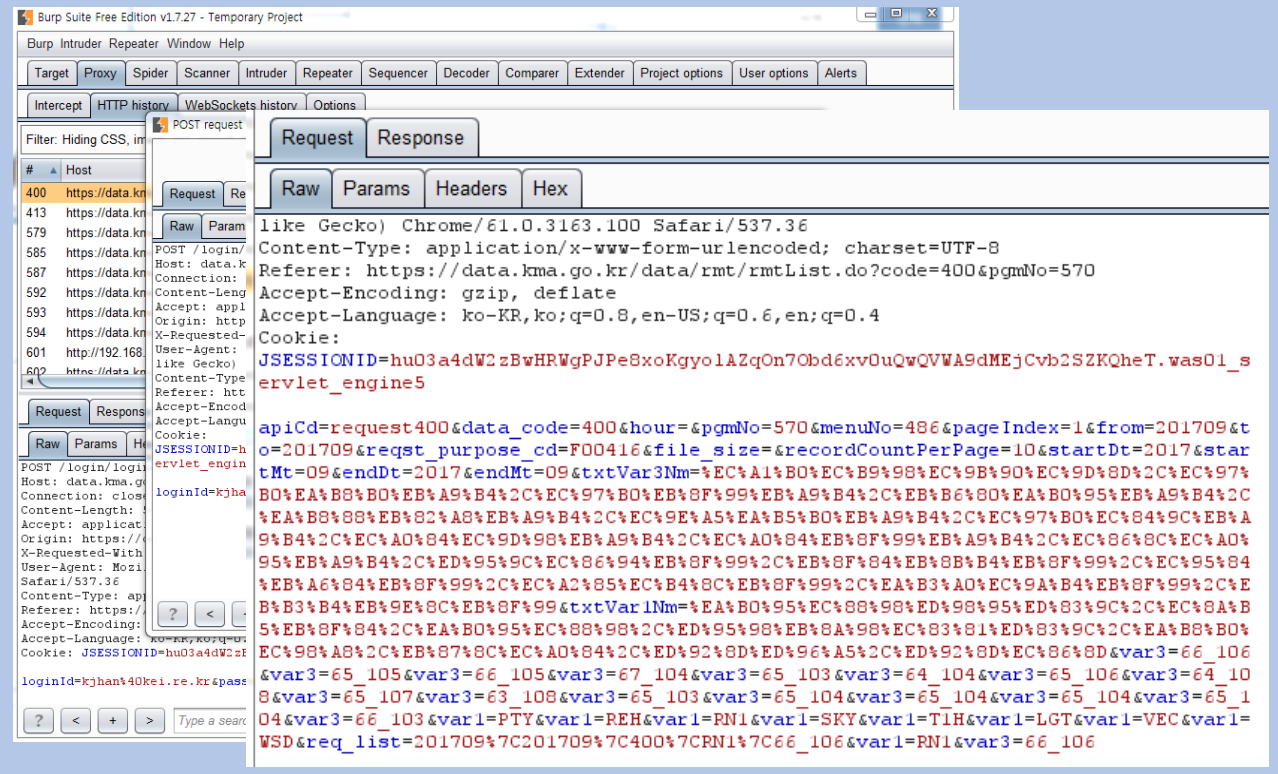
- (기상자료개방포털) 데이터 > 날씨예보 | 압축파일 다운로드

1. 실황분석자료	2. 초단기예보	3. 단기예보	4. ~~예보버전조화~~(없음)

데이터 식별 / 데이터 서비스 탐색

웹페이지 분석(전문SW 필요)

반자동화 코드 작성 + 전문SW 병행



- 과거 데이터 조회됨(시간 많이 소요)
  (실시간 확인의 의미)
- 로그인해야만 데이터 다운로드 가능
  (자동화할 수 있지만 불편함)
- 좌표 체계 활용 곤란
  (5km x 5km 해상도 전국을 분할)

- 파이선 : BeautifulSoup, pandas

(5.3) 환경분야 빅데이터 수집사례(3/4)

- (공데 포털) 한국환경공단 대기오염정보 조회 서비스 | 오픈API(XML)

1. 측정소별 실시간 측정정보	2. 통합대기환경지수 나쁨이상	3. 시도별 실시간 측정정보
4. 미세먼지/오전 예보통보	5. 시도별 실시간 평균정보	6. 시군구별 실시간 평균정보

데이터 식별 / 데이터 서비스 탐색

메타데이터 확인 / 오픈API 분석

자동화 코드 작성

- 로그인은 하지 않아도 되지만 전처리 필요 (그래도 양호)
- 형태는 XML이지만, 데이터 속성이 JSON (RDBMS보다 검색엔진 또는 NoSQL 우수)

```
list_station = []

for station in contents_html.find_all('item'):
    if str(station.find('stationname')) != 'None':
        list_station.append({"stationname": station.find('stationname').text,
                             "addr": station.find('addr').text,
                             "year": station.find('year').text,
                             "oper": station.find('oper').text,
                             "mangname": station.find('mangname').text,
                             "item": station.find('item').text,
                             "dmx": station.find('dmx').text,
                             "dmy": station.find('dmy').text})

cols = ["stationname","addr","year","oper","mangname","item","dmx","dmy"]
#cols = ["측정소명","주소","설치년도","관리기관","측정망","측정항목","위도","경도"]
df_station = pd.DataFrame(list_station, columns=cols)

df_station
```

Out[15]:

	stationname	addr	year	oper	mangname	item	dmx	dmy
0	반송로	경남 창원시 의창구 원이대로 450(시설관리공단 실내수영장 앞)	2008	경상남도보건환경연구원	도로변대기	SO2, CO, O3, NO2, PM10	35.232222	128.671389
1	사파동	경남 창원시 성산구 창이대로 706번길 16-23(사파민원센터)	2009	경상남도보건환경연구원	도시대기	SO2, CO, O3, NO2, PM10	35.221729	128.69825
2	경화동	경남 창원시 진해구 경화로16번길 31(병암동주민센터)	1994	경상남도보건환경연구원	도시대기	SO2, CO, O3, NO2, PM10, PM2.5	35.154972	128.689578

- 파이선 : BeautifulSoup, pandas

(5.3) 환경분야 빅데이터

- (에어코리아) 한국환경공단 대기오염

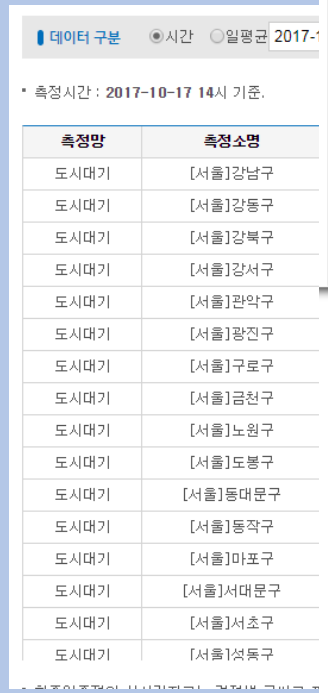
1. (실시간)우리동네대기 정보	2. (실시간)시도별 대
4. 측정소별 확정자료	5. 측정망·항목별 확정

데이터 식별 / 데이터 서비스 탐색

➡ 웹페이지 분석

➡ 자동화 코드 작성

- 분석 만하면 바로 저장이 가능함 (가장 좋은 형태)
- 데이터 분리 오류 발생 가능성 (측정소 코드체계 미공개(추정) 등)

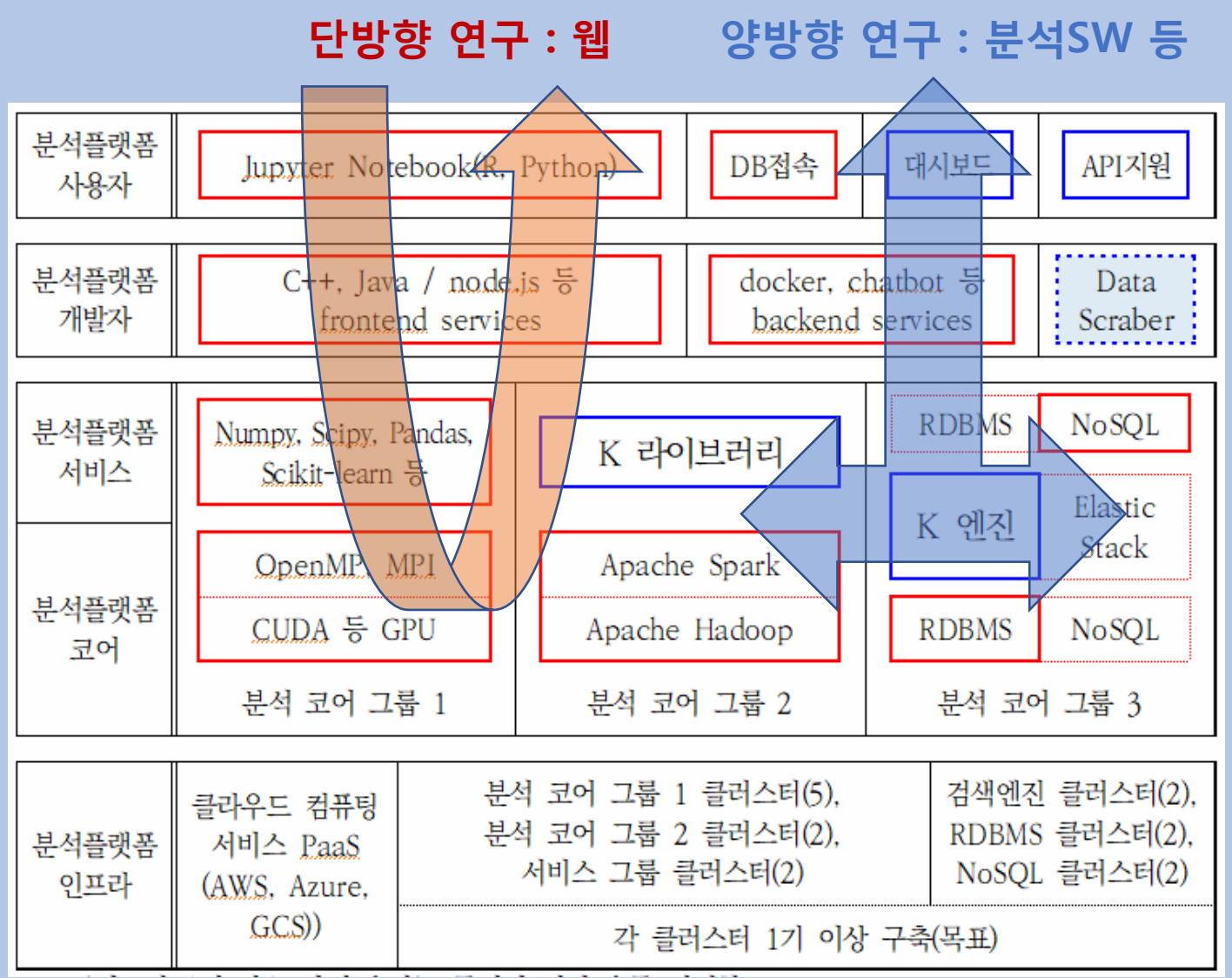



```python
import urllib.parse, urllib.request

#url = r"http://www.airkorea.or.kr/pmRelay?itemCode=10008"
strDateDiv = "strDateDiv=1"
searchDate = "searchDate=2006-01-01"
district = "district=02"
itemCode = "itemCode=10008"
#searchDate_f = "searchDate_f=201708"

#일자별 지역별 PM2.5 실시간 자료조회
url = r"http://www.airkorea.or.kr/pmRelaySub?" ₩
        + strDateDiv + r"&" + searchDate + r"&" + district ₩
        + r"&itemCode=10007"

df = pd.read_html(url, encoding="utf-8")
#df
df[0]

# contents_source = urllib.request.urlopen(url).read().decode('utf-8')
# contents_html = BeautifulSoup(contents_source, 'html.parser')
# contents_html

#URL 작성시

# strDateDiv : 1(시간) 2(일평균)
# searchDate_yyyy : 2008 ~ 2016 / 실제 2006 이후로 데이터 존재
# searchDate_mm : 01 ~ 12
# district : 02(서울) 031(경기) 032(인천) 033(강원) 041(충남) 042(대전) 043(충북) 044(
#            051(부산) 052(울산) 053(대구) 054(경북) 055(경남) 061(전남) 062(광주) 063(

#http://www.airkorea.or.kr/data_pmRelay_search_pop?strDateDiv=2&searchDate=2017-08-30&d
```

	측정망	측정소명	1시	2시	3시	4시	5시	6시	7시	8시	...	15시	16시	17시	18시	19시	20시	21시	22시	23시	24시
0	도시대기	[서울]강남구	55	46	38	48	46	58	73	72	...	66	53	55	56	68	61	65	59	70	77
1	도시대기	[서울]강동구	58	57	58	45	51	58	61	84	...	61	62	44	67	68	61	69	76	78	68
2	도시대기	[서울]강북구	47	55	54	51	72	72	54	50	...	48	46	41	50	55	48	46	46	60	39

- 파이선 : BeautifulSoup, pandas

(?.?) 환경 빅데이터 분석플랫폼 구성방안



- 5계층 / 역할 기준
- 색 구분
  - KEI 최적화(붉은색)
  - KEI 특화(파란색)
- 단계별 중점 전략
  - 1단계 : 데이터 수집기
  - 2단계 : 인프라-코어
  - 3단계 : 서비스-사용자

※ 개발자 : 2단계 이후

8