

## XML 이론 및 실습



데이터분석 방법 및 실습

# 주요 IT 기술 변화

- 모든 과학기술 분야에서의 데이터 홍수(Data Deluge)
- 멀티코어(Multicore) 에 따른 병렬(적시) 처리의 중요성
  - 클러속도에서 추가 코어를 통한 성능 개선
  - GPU 을 사용한 처리 능력 및 빠른 계산
- 클라우드
  - 상업적 데이터 센터 모델
  - 소유에서 대여로, 서비스 계산자원 이용
- 경량 클라이언트
  - 센서, 스마트 폰 및 테플릿
  - 백엔드 서비스에 대한 필요성
- 기업 주도의 기술 혁신과 기술 도입

# 데이터 과학자

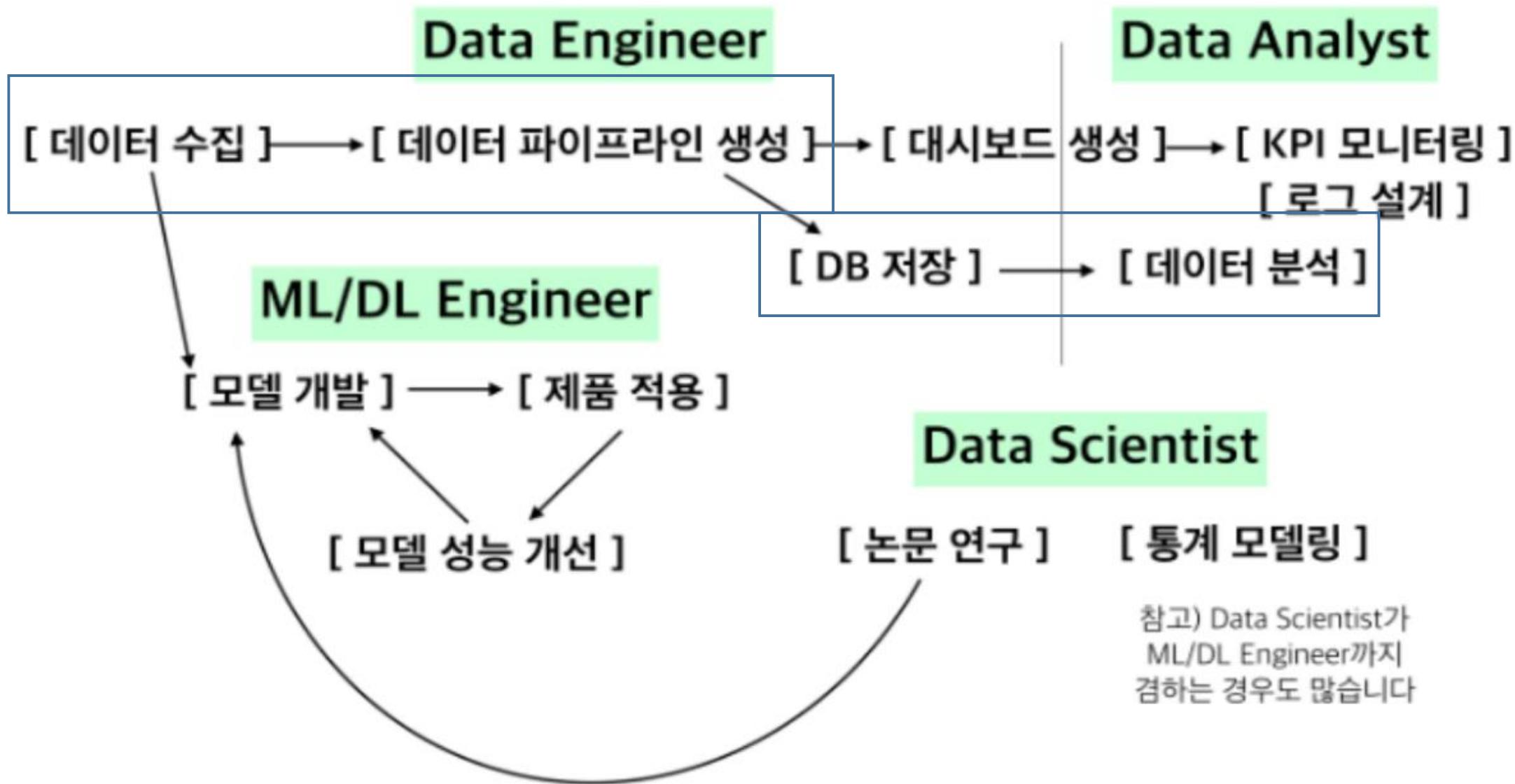
데이터 과학(data science)이란, 데이터 마이닝(Data Mining)과 유사하게 정형, 비정형 형태를 포함한 다양한 데이터로부터 지식과 인사이트를 추출하는데 과학적 방법론, 프로세스, 알고리즘, 시스템을 동원하는 융합분야다.<sup>[1]</sup>

데이터 과학은 데이터를 통해 실제 현상을 이해하고 분석하는데 통계학, 데이터 분석, 기계학습과 연관된 방법론을 통합하는 개념으로 정의되기도 한다.

데이터의 구체적인 내용이 아닌 서로 다른 성질의 내용이나 형식의 데이터에 공통으로 존재하는 성질, 또는 그것들을 다루기 위한 기술의 개발에 착안점을 둔다는 특징을 가진다. 사용되는 기술은 여러분야에 걸쳐있으며 수학, 통계학, 계산기과학, 정보공학, 패턴인식, 기계학습, 데이터마이닝, 데이터베이스 등과 관련이 있다. 데이터 과학을 연구하는 사람을 데이터 과학자라고 한다.

튜링상을 수상한 짐그레이(Jim Gray) 박사는 데이터 과학은 과학의 네번째 패러다임으로 정의하고 과학(경험, 이론, 계산, 그리고 이제 데이터)에 관한 모든 것이 바뀌고 있는데 이유는 정보 기술과 데이터 범람(data deluge) 때문이라고 주장했다.

데이터 과학은 생물학, 의학, 공학, 사회학, 인문과학 등의 여러 분야에 응용되고 있다.





데이터 추출 및 저장  
(Ingest)

# 주차별 수업 내용

- 수업내용
  - 시작에서 적용까지

파이썬 언어 이해	데이터 준비를 위한 자료 다루기	데이터 분석 프로젝트
자료형 및 제어문	CSV JSON XML	공공데이터 살펴보기
함수 및 예외처리		텀 프로젝트 계획서 작성하기
클래스와 객체 모듈	오픈 API HTTP 및 REST	프로젝트 목표 정의 데이터 수집 데이터 처리 결과 저장 및 시각화
	자료처리 패키지: numpy, pandas, matplotlib	

# 주차별 수업 내용

주차	수업주제 및 내용	주차 목표
1	수업 목표 및 수업 내용 소개	파이썬 환경 구축
2	파이썬 자료형 - 숫자, 문자열자료형 - 컬렉션 자료형	<ul style="list-style-type: none"><li>• IDLE 사용 자료를 변수에 저장한 후 확인함</li><li>• 공공데이터자료 소개 후 프로젝트 주제 선정 준비</li></ul>
3	파이썬 제어문 - 조건문 - 반복문	활용 자료 탐색 및 이해하기
4	파이썬 함수 및 예외처리 - 파이썬 함수 정의 및 호출 - 예외처리	활용 자료 형식 및 특징이해하기
5	파이썬 클래스 및 객체 활용 - 클래스와 객체 - 모듈	텀프로젝트 계획서 작성
6	파이썬 자료 활용 - CSV 개요 및 활용하기	오픈 API 소개
7	파이썬 자료 활용 - JSON 개요 및 활용하기	HTTP 및 REST 이해하기 (공공데이터 접근: JSON)

# 주차별 수업 내용

주차	수업주제 및 내용	주차 목표
8	파이썬 자료 활용 - XML 개요 및 활용하기	HTTP 및 REST 이해하기 (공공데이터 접근: XML)
9	빅데이터 자료처리 - numpy 패키지	숫자데이터 활용 통계처리
10	빅데이터 자료처리 - pandas 패키지	접근 자료를 데이터프레임으로 구성
11	빅데이터 자료처리 - matplotlib 패키지	데이터 시간화
12	웹 자료크롤링(crawling) 및 스크래핑(scraping)	HTTP 요청/응답으로 HTML 문서 다운로드 및 자료추출
13	지도데이터 처리	데이터 융합 및 데이터 저장
14	팀별 프로젝트 결과 발표	팀별 발표



# 데이터 분석 필요와 수업의 이해

← → ↺

github.com/iceman67/Web-Scraping

Apps

★ Bookmarks

🌐 새 탭

📄 Science Cloud Sum...

📁 k

📄 yunheekanghomep...

📄 ICECECE 2013 - Int...

CS CS545: Machine Le...

🌐

🌐

📁

## Web Scraping with Python

[Manage topics](#)

🕒 7 commits

🌿 1 branch


📦 0 packages








🔍

Branch: master ▼

New pull request

Create new

 iceman67 Update README

 <a href="#">Getting started Web scraping.ipynb</a>	Add files via upload
 <a href="#">README</a>	Update README
 <a href="#">Web Scraping.ipynb</a>	Add files via upload
 <a href="#">Worldcup 2018 player.ipynb</a>	Add files via upload
 <a href="#">books_and_authors.csv</a>	Add files via upload
 <a href="#">data visualization with python.ipynb</a>	Add files via upload
 <a href="#">simple_web_scraper.py</a>	Add files via upload

# 데이터 분석 필요와 수업의 이해

[←](#) [→](#) [↻](#) [github.com/iceman67/Public\\_OpenAPI](#)

[Apps](#) [★ Bookmarks](#) [🌐 새 탭](#) [Science Cloud Sum...](#) [k](#) [yunheekanghomep...](#) [ICECECE 2013 - Int...](#) [CS CS545: Machine Le...](#) [🌐](#) [🌐](#)

[↔ Code](#) [! Issues 0](#) [🔗 Pull requests 0](#) [▶ Actions](#) [📊 Projects 0](#) [📖 Wiki](#) [🛡️ Secu](#)

*No description, website, or topics provided.*

[Manage topics](#)

🔒 5 commits

🔗 1 branch

📦 0 packages

Branch: master ▼

New pull request

Create new

🚧 iceman67 Update README

📄 [02-1-XML-응용-openAPI.pdf](#)

Add files via upload

📄 [README](#)

Update README

📄 [WeatherSightSeeing.py](#)

Add files via upload

📄 [weather\\_app.py](#)

Add files via upload

# 데이터 분석 필요와 수업의 이해

Branch: master ▾

Data-Visualization / Geo Starbucks.ipynb



iceman67 Add files via upload

1 contributor

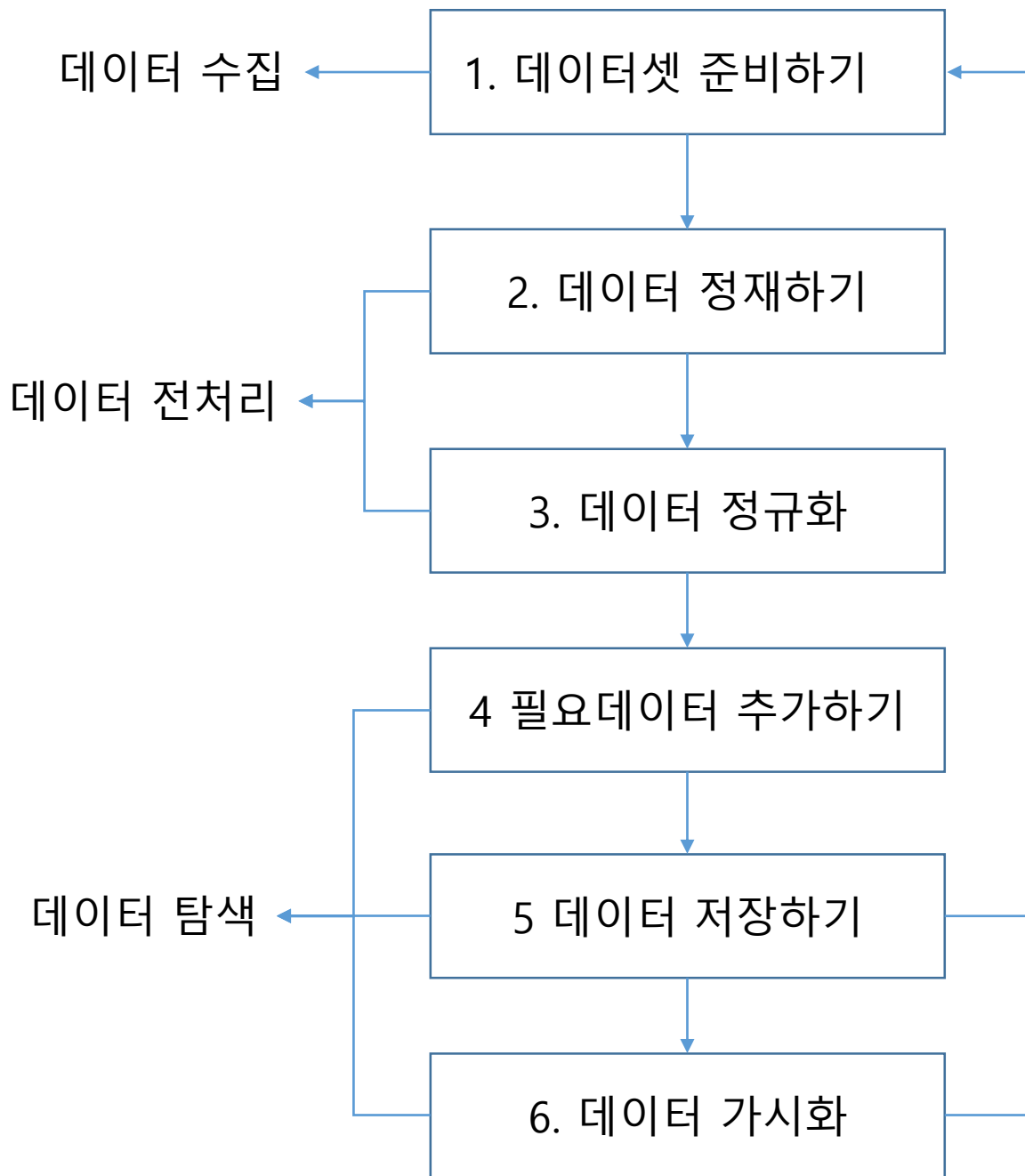
401 lines (401 sloc) | 61.4 KB

```
In [1]: import requests
import json
import pandas as pd
from pandas.io.json import json_normalize
```

```
In [2]: data = {
    'ins_lat': '37.56682', # 지정한 위도와 경도에서 가까운 순으로 나열
    'ins_lng': '126.97865',
    'p_sido_cd': '01', # 01=서울시, 08=경기 ... 16=제주
    'p_gugun_cd': '', # 세부지역 (지정하지 않으면 시/도 전체)
    'in_biz_cd': '',
    'set_date': '',
    'ins_cd': '1000'
```

## 데이터 파이프라인

## 재사용 가능한 코드 구축



In [7]: `df=pd.read_csv('season.csv')`

`# This is our dataframe`  
`df`

Out[7]:

	dates	day	temp	wind-speed
0	2/1/2012	sunny	45.0	12
1	3/1/2012	rainy	46.0	34
2	4/1/2012	hot	47.0	45
3	5/1/2012	NaN	NaN	56
4	6/1/2012	hot	49.0	Not available
5	7/1/2012	NaN	NaN	Not available
6	8/1/2012	hot	12.0	45
7	9/1/2012	rainy	23.0	41
8	10/1/2012	NaN	NaN	NaN
9	11/1/2012	NaN	NaN	NaN

수집한 데이터가 들어왔을 때, 이를 다양한 각도에서 관찰하고 이해하는 과정

데이터 수집

1. 데이터셋 준비하기



## 원시데이터 수집방법

- 웹(데이터) 스크래핑(scraping)
- 파일
- API
- DBMS

## 원시데이터 종류 별 특징

- 자연어로 작성된 비정형 텍스트 (한국어, 영어, 중국어...)
- 정형데이터
  - CSV
  - JSON
  - HTML/XML 마이크업 자료
  - 데이터베이스에 저장된 테이블 자료

# 파이썬 개발환경 설치 및 시작하기

백석대학교 강윤희

# Python 버전 차이 예

- Python3:

```
>>> 10 / 3  
3.3333333333333335
```

- Python 2.x:

```
>>> 10 / 3  
3
```

- Python3:

```
>>> print ("hello Python")
```

- Python2:

```
>>> print "hello Python"
```

Python3에서는 long 형 자료형  
이 int형으로 통일됨

# IDLE(Integrated DeveLopment Environment)

- 통합 개발 환경
- 세개의 닫는 꺾쇠(>>>) 프롬프트에 코드 입력



```
Python 3.7.2 Shell
File Edit Shell Debug Options Window Help
Python 3.7.2 (tags/v3.7.2:9a3ffc0492, Dec 23 2018, 22:20:52) [MSC v.1916 32 bit
(Intel)] on win32
Type "help", "copyright", "credits" or "license()" for more information.
>>> |
>>> print("Hello World")
Hello World
```



# Python vs C 언어

```
def factorial(x):  
    if x == 0:  
        return 1  
    else:  
        return x * factorial(x - 1)
```

```
int factorial(int x)  
{  
    if (x == 0)  
    {  
        return 1;  
    }  
    else  
    {  
        return x * factorial(x - 1);  
    }  
}
```

# Python vs C 언어

 factorial.py - C:\Users\Yunhee\Desktop\LINC플러스특강\python\algorithm... —

File Edit Format Run Options Window Help

---

```
def factorial(x):  
    if x == 0:  
        return 1  
    else:  
        return x * factorial(x - 1)  
  
if __name__ == '__main__':  
    print(factorial(5))
```

단순한 문법  
짧은 코드, 높은 생산성  
높은 가독성

# 변수 및 자료형

백석대학교 강윤희

# 변수 사용하기

- 변수 명(식별자)는 자료형을 지정하지 않고 값을 변수에 저장할 수 있음 (by 배정문)
- C 언어와 동일하게 식별자로 변수명을 사용함

```
name = "Bob"
```

```
age = 15
```

```
age = 15
```

```
age += 1 # increment age by 1
```

```
print(age)
```

# 자료형

- 숫자형(Number)
  - 정수형, 실수형, 복소수
  - 8진수와 16진수
- 문자열(String)
  - 문자, 단어 등으로 구성된 문자들의 집합
- 불(bool)
  - 참(True)과 거짓(False)을 나타내는 자료형
- 리스트(List)
  - 값의 생성, 삭제, 수정이 가능한 가변 배열
- 튜플(Tuple)
  - 값을 변경할 수 없는 리스트
- 딕셔너리(Dictionary)
  - 키를 사용하여 값을 접근하는 연관 배열(Associative array)
- 집합(Set)
  - 중복을 허용하지 않고 순서가 없는 리스트

리스트, 튜플, 딕셔너리, 집합을  
**컬렉션** 자료형으로 구분함

# 자료형

- 숫자형

```
x = 1    # int
```

```
y = 2.8  # float
```

```
z = 1j    # complex
```

```
print(type(x))
```

```
print(type(y))
```

```
print(type(z))
```

# 문제풀이

1. 다음 중 변수명으로 적합한 것은?

① 컴퓨터 ② 63building ③ file\_name ④ font&

2. 다음 중 변수명으로 적합하지 않은 것은?

① eng\_score ② font1 ③ studentName ④ file name

3. 변수 carname 에 값으로 Volvo를 저장하도록 문장을 완성하십시오.

# 문제풀이

4. 다음 파이썬 프로그램의 수행결과를 보이시오

```
x = "awesome"  
print("Python is " + x)
```

5. 코드의 실행 결과를 예상하라

```
a = "Life is too short"  
a.split()  
a = "Art is long. Life is too short"  
a.split('.')
```