

# 데이터랭글링 개론

# 데이터 랭글링

## ➤ Data Wrangling

- ✓ 데이터 랭글링(Data Wrangling) 혹은 데이터 먼징(Data Munging)은 원자료(raw data)를 또다른 형태로 수작업으로 전환하거나 매핑하는 과정
  - wikipedia
- ✓ 반자동화 도구의 도움 필요 -> 파이썬 + 라이브러리

## ➤ 새로운 분야 공부 Tip1

- ✓ History와 관련 용어로 개념정리

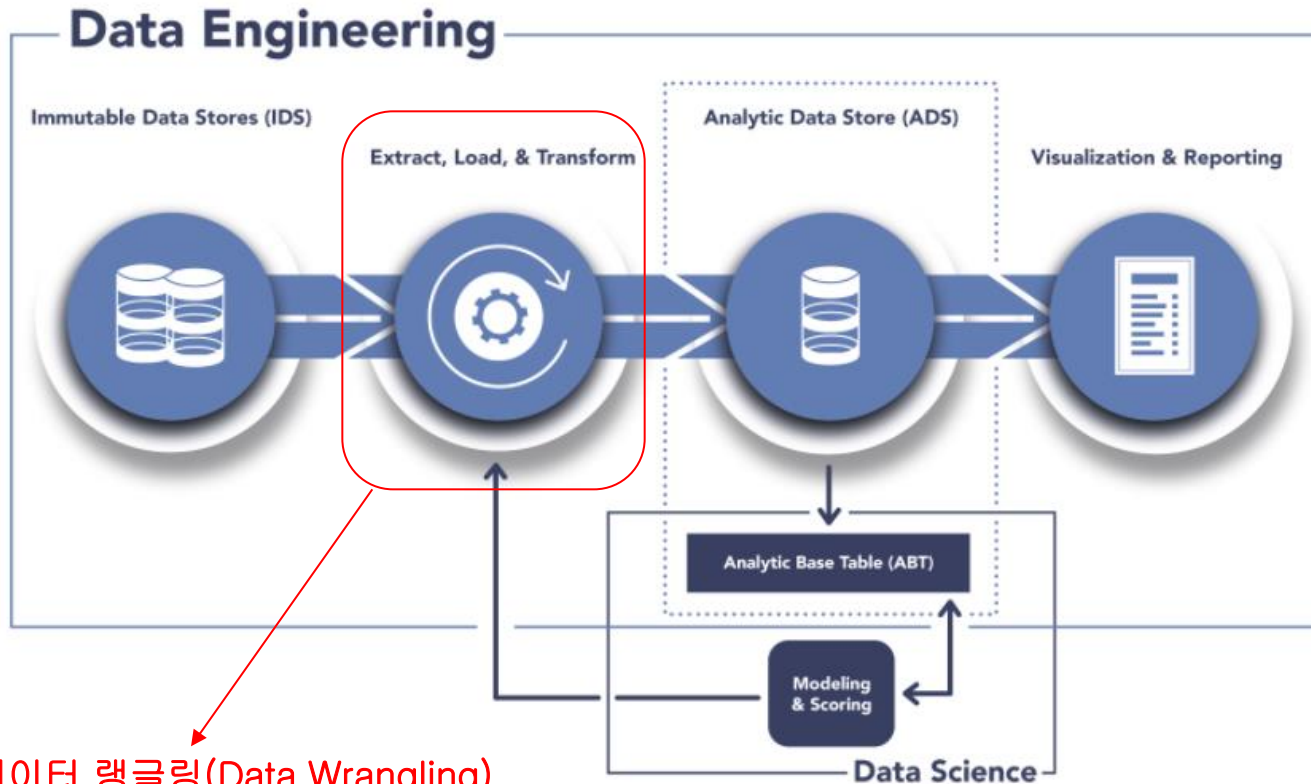
# (사례) API vs. Library

## ➤ 자바API - 위키백과

- ✓ 자바 **API**(Java API)는 자바를 사용하여 쉽게 구현할 수 있도록 한 클래스 **라이브러리**의 집합이다. 즉, 자바라는 언어를 사용하여 사용자의 부담을 최소화하는 반면에 입출력, 화면 구성, 이미지, 네트워크와 같이 복잡하지만 필요한 클래스들을 미리 구현하여 사용자가 쉽게 구현하도록 하는 API이다.

- ✓ cf. Framework

# 데이터 엔지니어링이란?

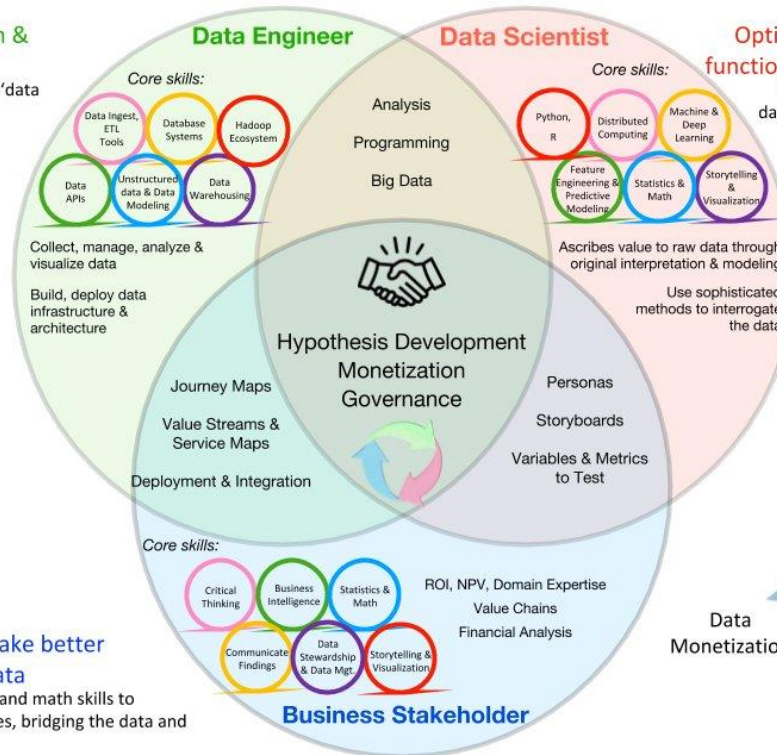


# 데이터 엔지니어란?

## Data Science Roles & How They Interact

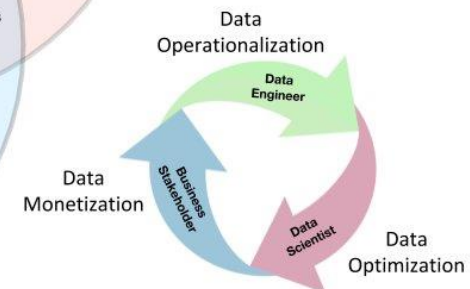
### Enable data access & utilization & enable value capture

Builds and supports the infrastructure or 'data pipe' and all associated SW engineering infrastructure tasks.



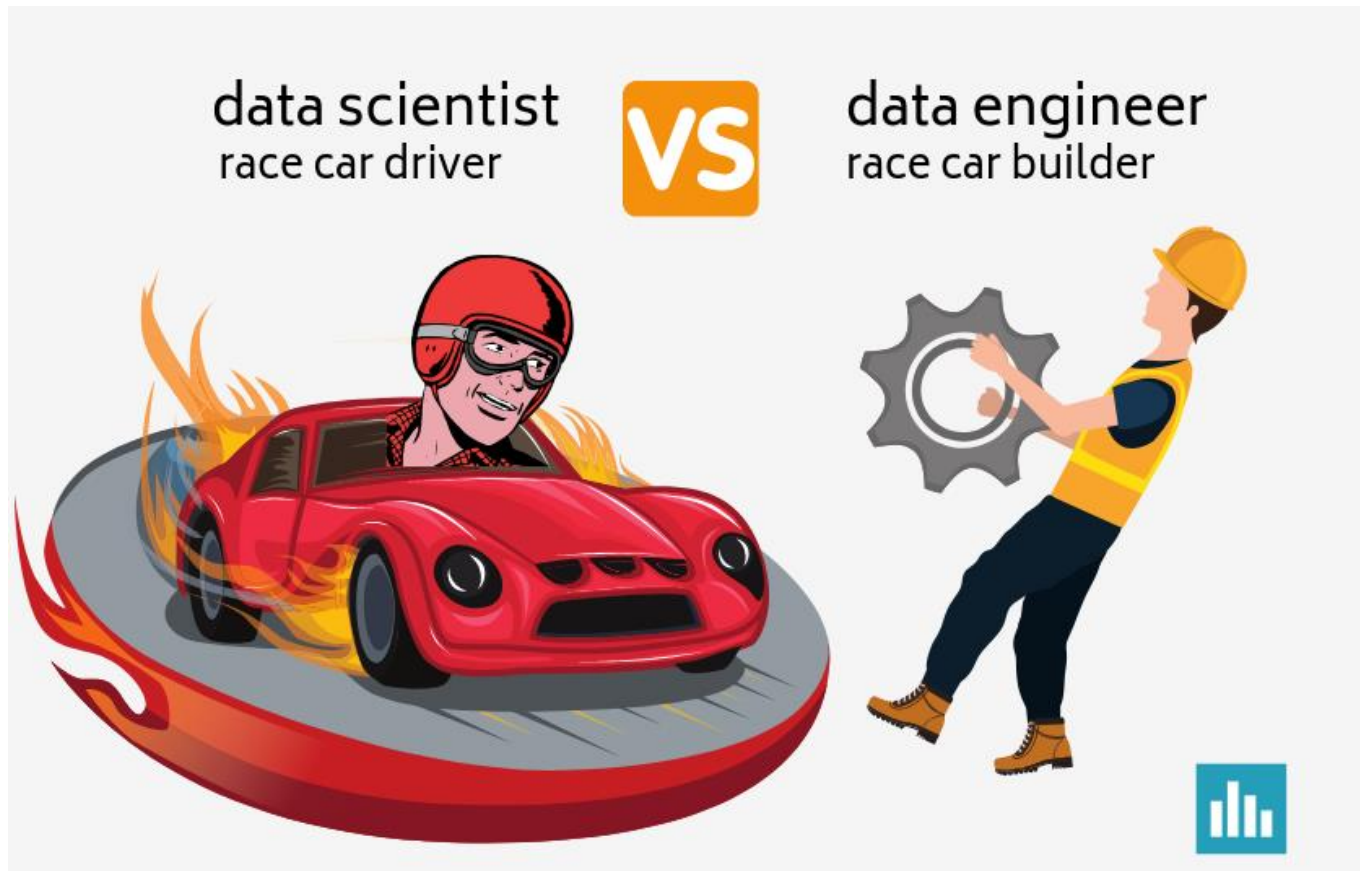
### Help the business make better decisions through data

Blend of business, analytic and math skills to explore and solve challenges, bridging the data and business communities.



<https://buff.ly/2NXhazc>

# Data Scientist vs. Data Engineer



# ETL vs ELT

## ➤ ETL이란

- ✓ ETL is a process that extracts the data from different source systems, then transforms the data (like applying calculations, etc.) and finally loads the data into the **Data Warehouse** system.

## ➤ 새로운 분야 공부 Tip2

- ✓ 좋은 Essay와 논문 찾아 읽기

<https://www.trifacta.com/blog/is-etl-dead/>

( ETL에서 Data Wrangling의 시대로? )

# ETL to ELT

<https://www.trifacta.com/blog/is-etl-dead/> Essay중에서

## ETL vs ELT: Decoupling ETL with ELT

Traditional ETL might be considered a bottleneck, but that doesn't mean it's invaluable. What is ELT? The same basic challenges that ETL tools and processes were designed to solve still exist, even if many of the surrounding factors have changed. For example, at a fundamental level, organizations still need to extract (E) data from legacy systems and load (L) it into their data lake. And they still need to transform (T) that data for use in analytics projects. "ETL" work needs to get done—but what can change is the order in which it is achieved and new technologies that can support this work.

Instead of an ETL pipeline, many organizations are taking an "ELT" approach. So what is ETL? ETL is a traditional type of data integration, and it stands for extract, transform, load. Data is extracted from its source, converted into a usable format, and loaded into a system for analysis. Most often analysts use this process to build data warehouses. ETL became popular in the 1970s and remained popular through the 90s, but with modern data innovations, including the cloud, its prevalence has diminished. With some of these innovations, companies have also adapted a similar process called ELT. ELT stands for extract, load, transform. We'll examine why it's often better to use ELT tools and load before transforming.



# Is ETL dead?

<https://www.trifacta.com/blog/is-etl-dead/> Essay중에서

## ELT Tools and Data Wrangling in the Cloud

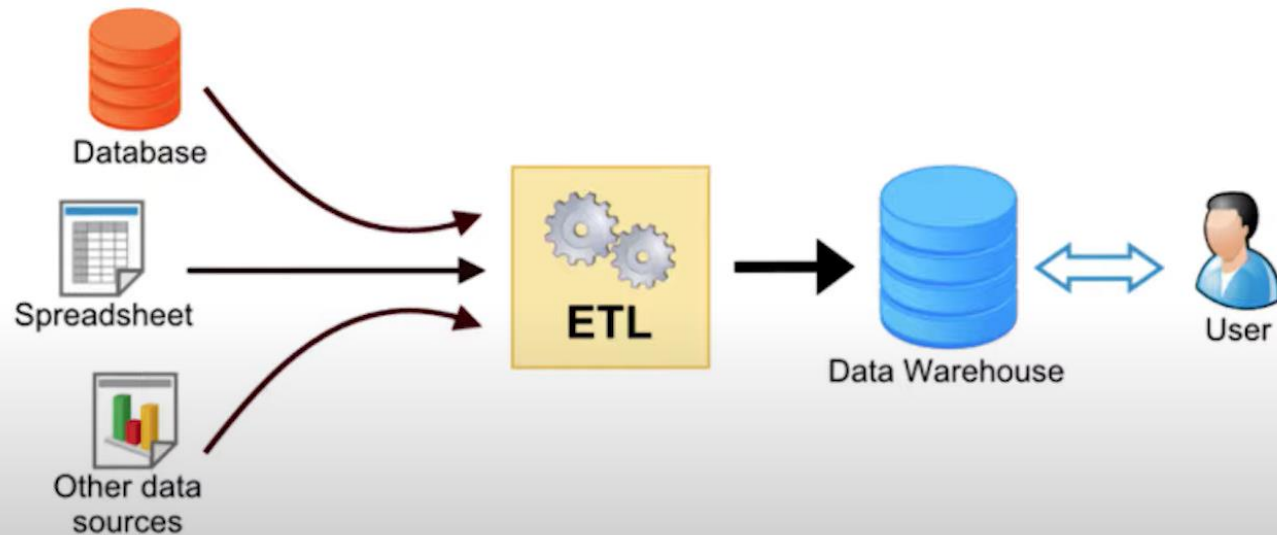


Will Davis · December 21, 2020

Is ETL dead? What is the difference between ETL vs ELT? Did ELT take over or is something new taking its place? It's a question that has come up a lot in recent years as organizations modernize their analytics infrastructure. Huge shifts are underfoot in the analytics landscape and it isn't always clear where this change leaves ETL. The short answer? No, ETL is not dead. But the ETL pipeline looks different today than it did a few decades ago. Organizations might not need to ditch ETL entirely, but they do need to closely evaluate its current role and understand how it could be better utilized to fit within a modern analytics landscape.

# ETL

## ETL: Extract - Transform - Load



**Skillcurb**  
Learn Cutting-Edge Technology from Experts

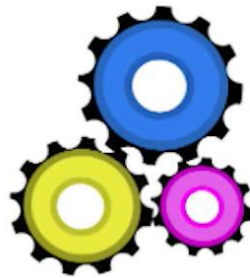
# ETL

## Extract - Transform - Load



### Extract

- From a source
- Passed to staging
- Structured data
- Unstructured data



### Transform

- Data Cleaning/Organizing
- Single System Format
- Improving Data Quality

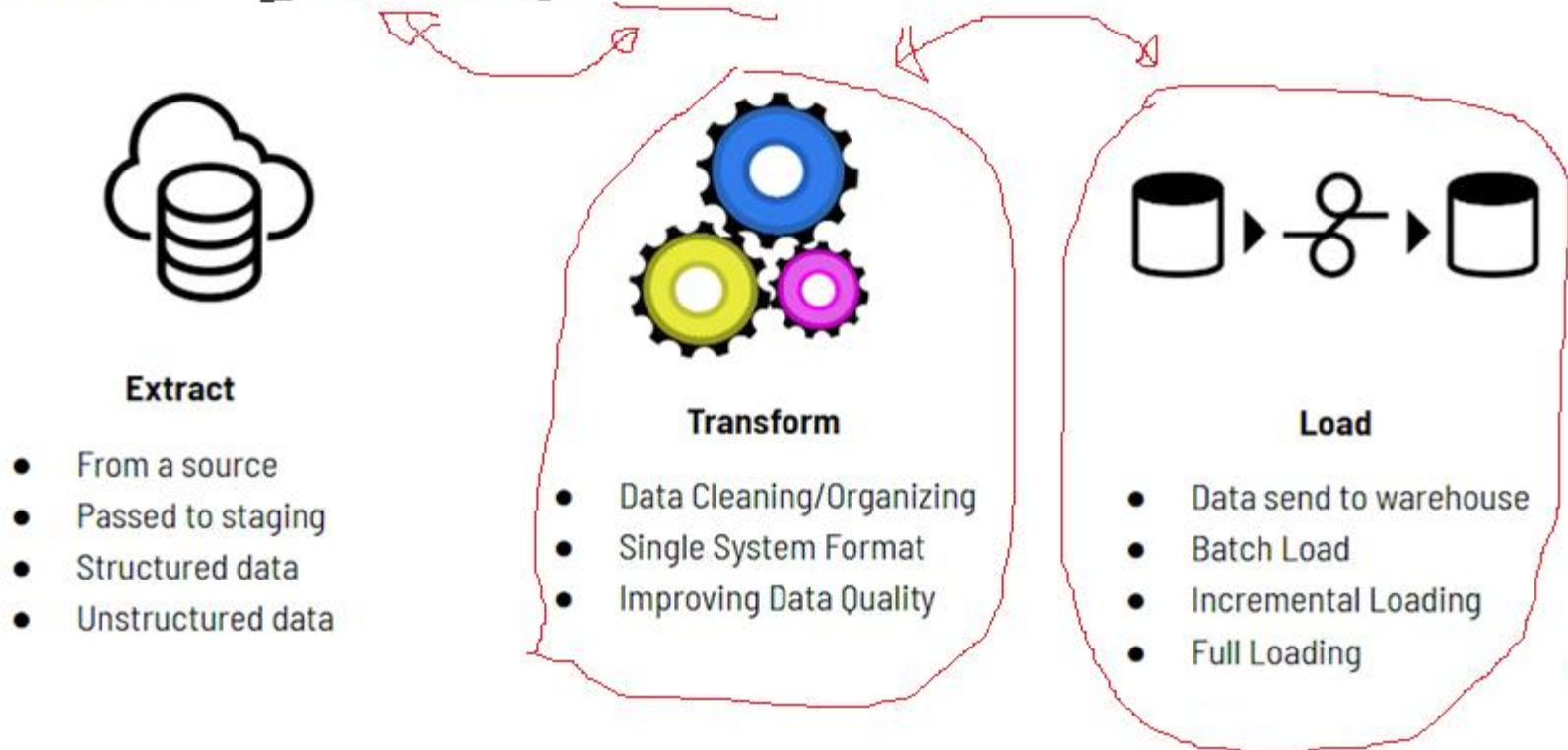


### Load

- Data send to warehouse
- Batch Load
- Incremental Loading
- Full Loading

# cf. ELT

## Extract - Transform - Load



※ Extract와 Load를 자동화하기 위해

# Data Lake

## ➤ Data Lake vs. Data Warehouse

### ✓ Data structure: raw vs. processed (by ETL)

<https://www.talend.com/resources/data-lake-vs-data-warehouse/>

**talend**

Products ▾ Solutions ▾ Pricing

## Data Lake vs Data Warehouse

[Knowledge center](#) » [Data integration](#) » [What is a Data Warehouse and Why Does It...](#) » [Data Lake vs Data Warehouse](#)

[Data lakes](#) and [data warehouses](#) are both widely used for storing [big data](#), but they are not interchangeable terms. A data lake is a vast pool of raw data, the purpose for which is not yet defined. A data warehouse is a repository for structured, filtered data that has already been processed for a specific purpose. There is even an emerging data management architecture trend of the [data lakehouse](#), which combines the flexibility of a data lake with the data management capabilities of a data warehouse.

# Data Lake vs. Data Warehouse

## ➤ Data Warehouse

- ✓ 정형데이터만 저장(RDBMS 사용) => ETL 작업 필요.
- ✓ ETL -> Data Warehouse
- ✓ 전통적 저장 시스템

## ➤ Data Lake

- ✓ 정형, 비정형 데이터 모두 저장. raw data 형태 저장.
- ✓ Data Lake -> Data Wrangling
- ✓ Bigdata 시대의 저장 시스템

# Data Lake 사례1





# Data Lake 활용





# Data Lake 사례2



# 데이터는 21세기의 천연자원



# 데이터 시대의 청바지?

136년 만에 금광에서 발견 된 세계에서 가장 오래된  
리바이스 청바지 (2021.06)



※ 청바지는 서부시대에 텐트업자가 남는 원단을 가지고 만든 게 시초



# 데이터 사이언티스트는

Big Data News portal인 datanami의 조사에 따르면



언어 선택 | ▼

[Translation Disclaimer](#)

Search this site

Search

About Resources [Subscribe](#)

Follow Datanami: [f](#) [t](#) [in](#) [rss](#)

HOME

COVID-19

FEATURES ▼

SECTORS ▼

APPLICATIONS ▼

TECHNOLOGIES ▼

VENDORS

JOB BANK

EVENTS ▼

July 6, 2020

## Data Prep Still Dominates Data Scientists' Time, Survey Finds

Alex Woodie



(BEST-BACKGROUNDS/Shutterstock)

Data scientists spend about 45% of their time on data preparation tasks, including loading and cleaning data, according to a survey of data scientists conducted by Anaconda. The company also analyzed the gap between what data scientists learn as students, and what the enterprises demand.

Data cleansing – fixing or discarding anomalous or wrong numbers and otherwise ensuring the data is an accurate representation of the phenomenon it is meant to measure — accounts for more than a quarter of average day for data scientists, followed by 19% for data loading (the “L” in ETL), according to [Anaconda's](#) annual survey.

Data visualization tasks occupied for about 21% of their time, while model selection, model training and scoring, and model deployment each consume 11% to 12% of the day, the survey found.

THIS JUST IN

MOST READ

August 6, 2021

- ▶ [Teradata Reports Second Quarter 2021 Financial Results](#)
- ▶ [Elastic Announces Availability of Elastic Agent, Support for Azure Private Link](#)
- ▶ [Teradata Cloud Momentum Continues with New Customers in First-half 2021](#)
- ▶ [BrainChip Demonstrates that Intelligent AI is Everywhere at AI Hardware Summit 2021](#)
- ▶ [Confluent Announces Second Quarter 2021 Financial Results](#)
- ▶ [AWS Announces Amazon Transcribe Call Analytics for Customer Conversation Insight Extraction](#)

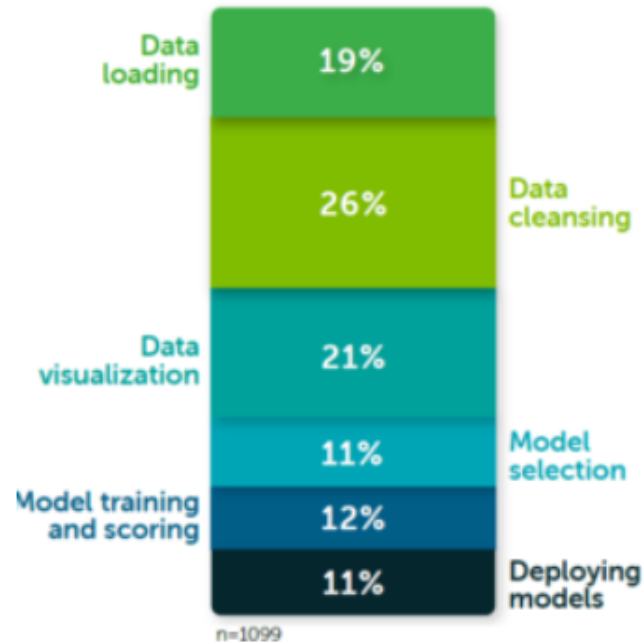
August 5, 2021

- ▶ [New IDC Forecast for IaaS, PaaS Workloads](#)

# 45%의 시간을

데이터 로딩과 데이터  
클리닝에 할애하고  
있다.

그리고 나머지 시간  
중에도 21%를  
데이터시각화에



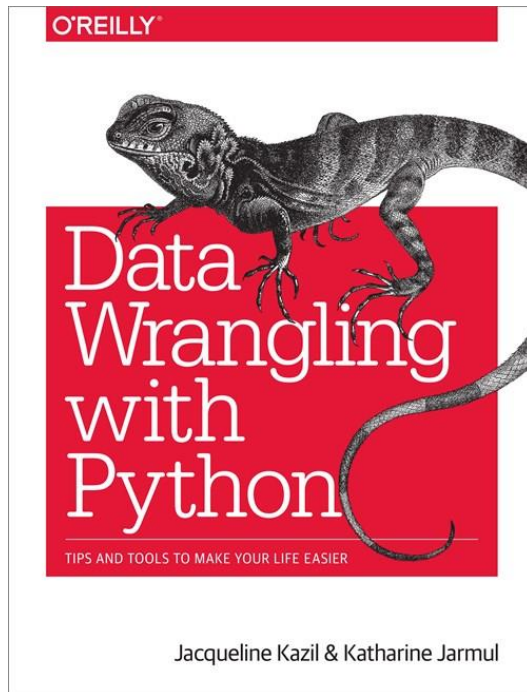
*How data scientists spend their time (Image courtesy  
Anaconda [“2020 State of Data Science: Moving From  
Hype Toward Maturity.”](#))*

<http://nirvacana.com/thoughts/2013/07/08/becoming-a-data-scientist/>



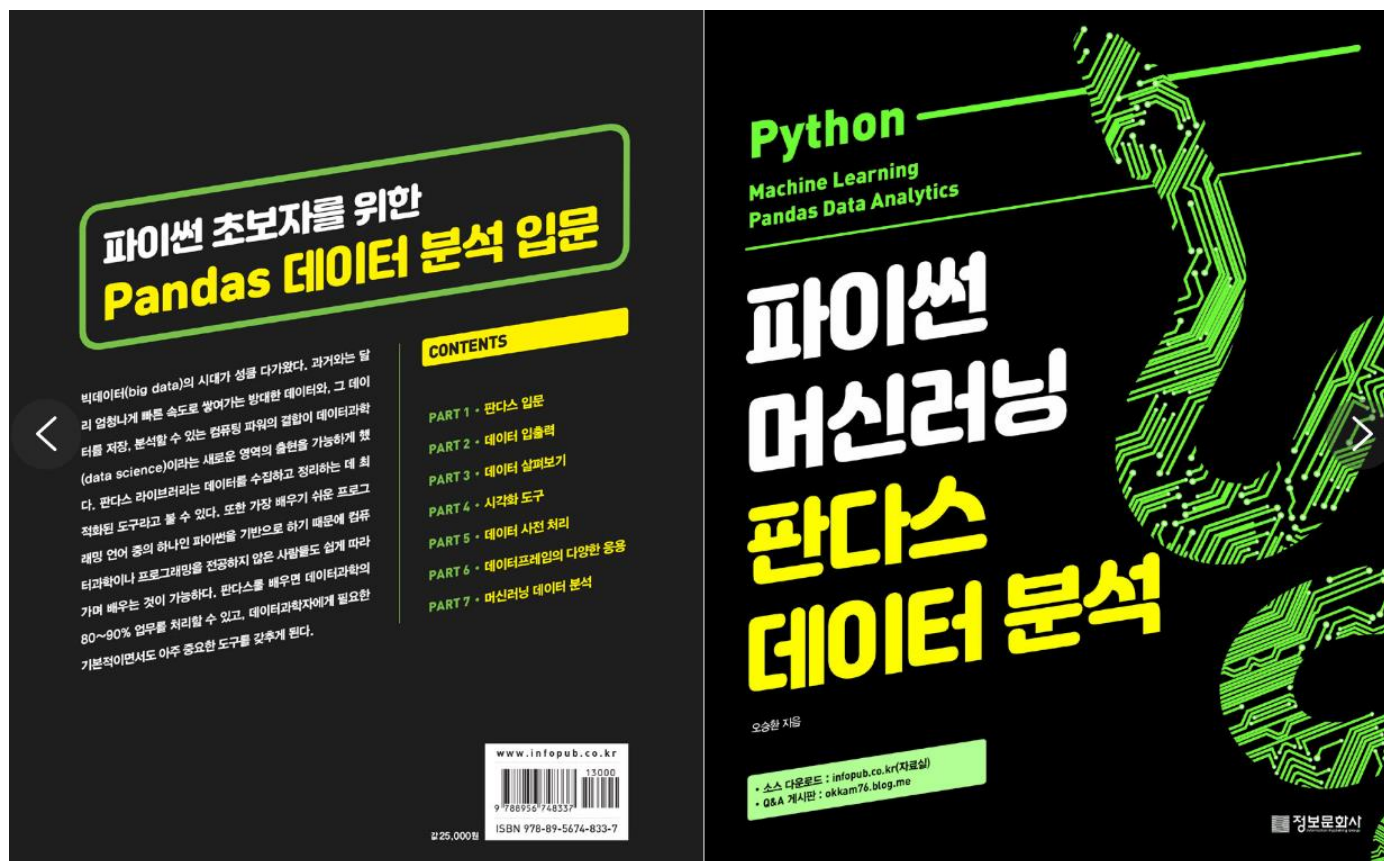
# 강의교재1

- Jacqueline Kazil & Katharine Jarmul, "Data Wrangling with Python", 2016, O'Reilly
- 번역판: 인사이트



## 강의교재2

- 파이썬 머신러닝 판다스 데이터분석
- 정보문화사 오승환 지음







# 교재 둘러보기

## ➤ 새로운 분야 공부 Tip3

✓ 관련 교재의 introduction과 summary읽기

## ➤ 새로운 분야 공부 Tip4

✓ 관련 교재의 다른 환경 설정해보기

① 교재1 : Python + pip + sublime (code editor)

② 교재2 : Anaconda + Spider IDE

③ 교재3 : Anaconda + Jupyter Notebook

# 파이썬 개발환경

## ➤ Python Interpreter

<https://www.python.org/downloads/>

✓ ver3.8 vs. ver2.7 (old ver. – 교재 예제)

✓ 기본설치 IDE – 파이썬IDLE ( & Learning )

## ➤ 통합개발환경 (IDE)

✓ PyCharm – JetBrains사

<https://www.jetbrains.com/pycharm/download>

✓ Jupyter Notebook – 웹IDE

# Python Package Manager

## ➤ pip

- ✓ default package manager for Python Library
- ✓ general use, python specific

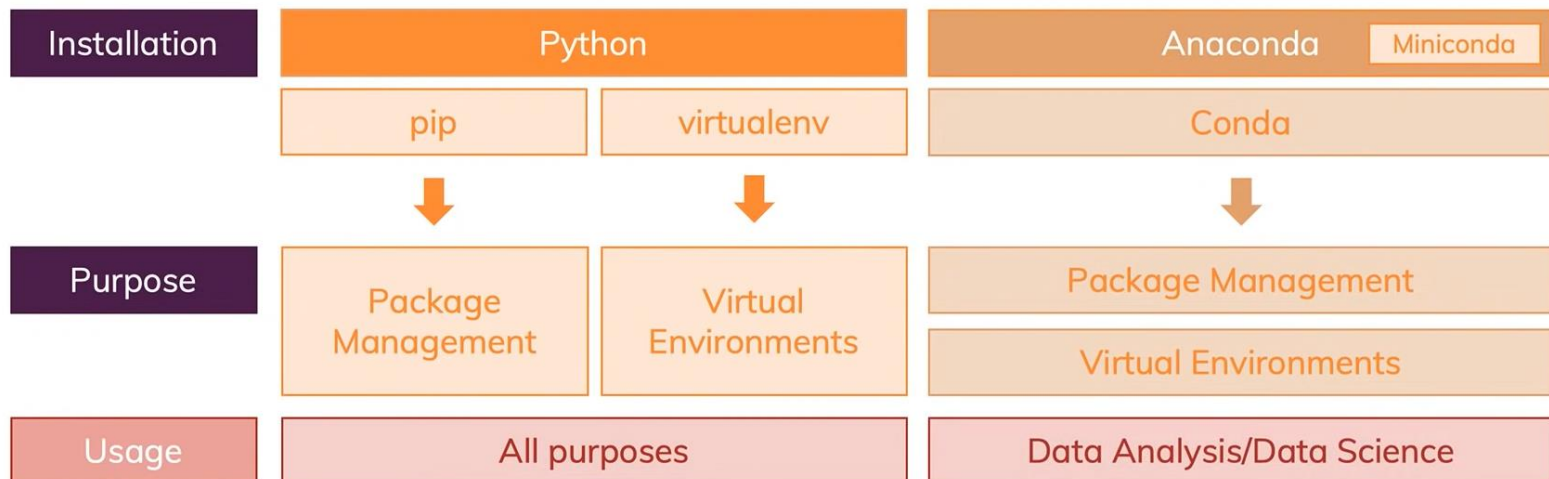
## ➤ conda

- ✓ package manager for data science
- ✓ R, C 등 non-python 패키지도 설치 가능
- ✓ from the **Anaconda** repository & cloud

# 개발 환경 선택



## Installing Python



# 우리의 개발환경

## ➤ Anaconda Individual Edition 설치

<https://www.anaconda.com/>

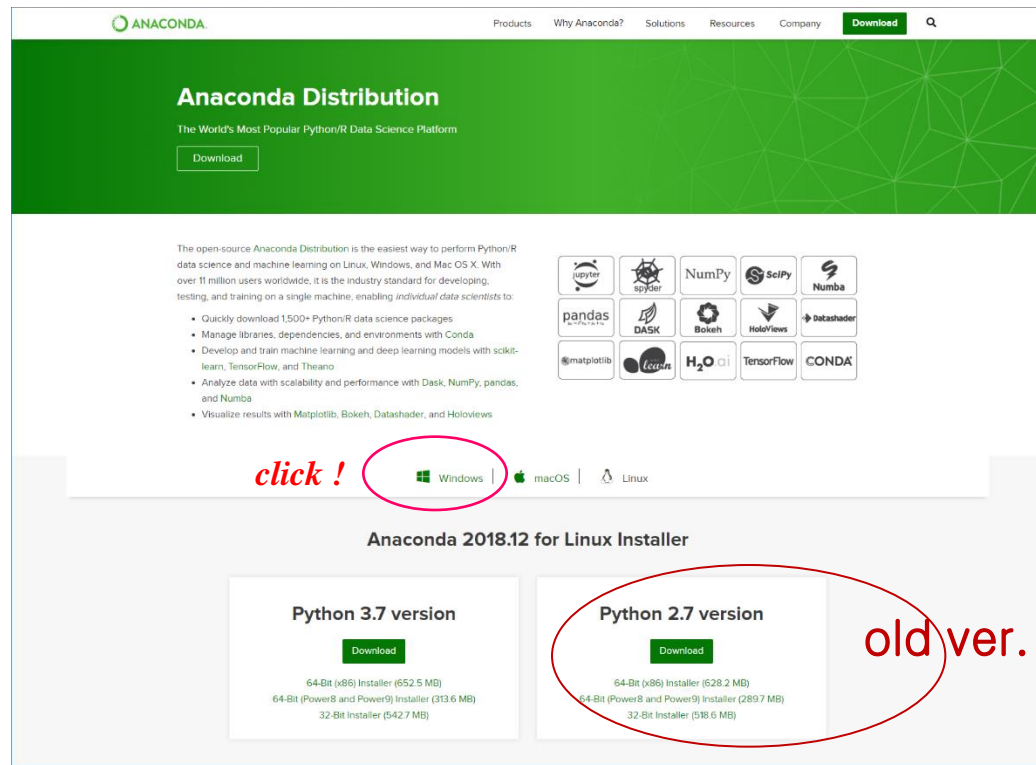
- ✓ Data science toolkit : open source
- ✓ 파이썬 Interpreter 최신 버전
- ✓ Conda Package
- ✓ Jupyter Notebook

## ➤ Anaconda 2019타입 설치예제 참고

- ✓ Next Page

# 윈도우 용 Anaconda Python 설치 (2019 old style)

- 다운로드 사이트: <https://www.anaconda.com/distribution/#windows>
- 사용중인 Windows 가 64-bit 버전인지 혹은 32-bit 버전인지 미리 확인 (64-bit 권장)



# 파이썬 설치

- Windows 용 Anaconda Python 3.7 version 다운로드 및 설치 파일 실행
  - 64-Bit Graphical Installer 권장

The screenshot shows the Anaconda Distribution website. The main heading is "Anaconda Distribution" with the subtitle "The World's Most Popular Python/R Data Science Platform". Below this is a "Download" button. The page lists various data science packages supported by Anaconda, including Jupyter, NumPy, SciPy, Numba, pandas, Dask, Bokeh, HoloViews, and Datashader. At the bottom, there are links for Windows, macOS, and Linux. The "Anaconda 2019.10 for Windows Installer" section shows two options: "Python 3.7 version" and "Python 2.7 version". Under "Python 3.7 version", the "64-Bit Graphical Installer (462 MB)" is highlighted with a red circle and a red arrow pointing to it from the text "click ! 64-Bit Graphical Installer".

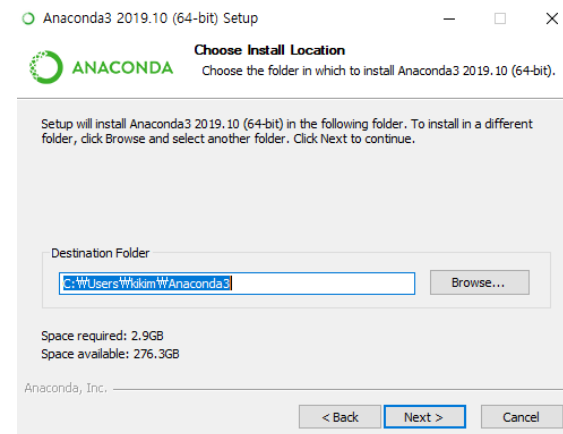
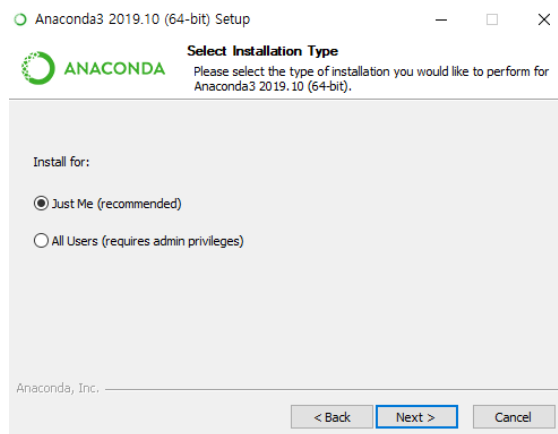
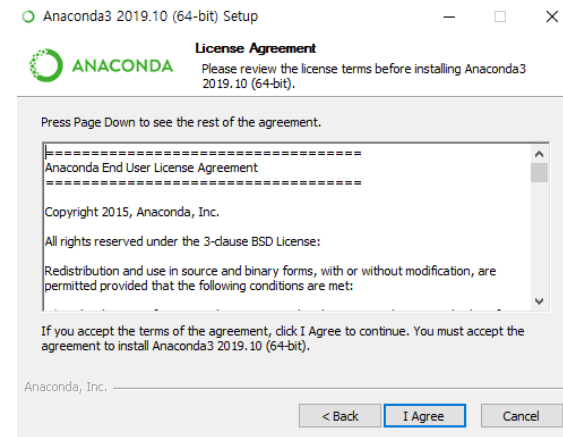
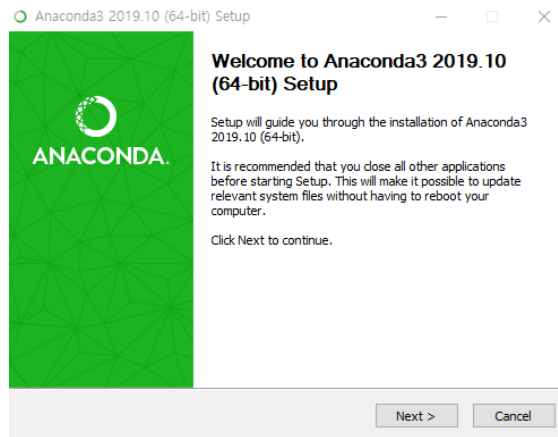
**click ! 64-Bit Graphical Installer**

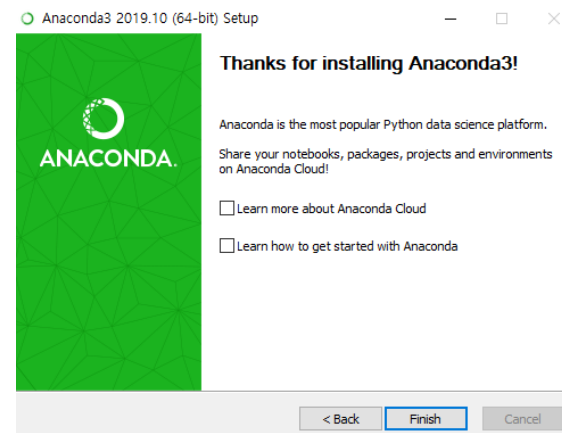
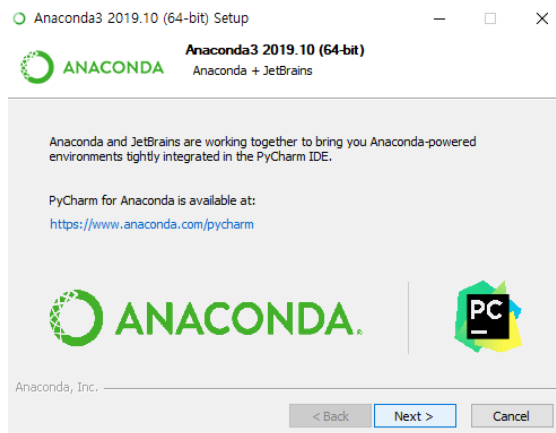
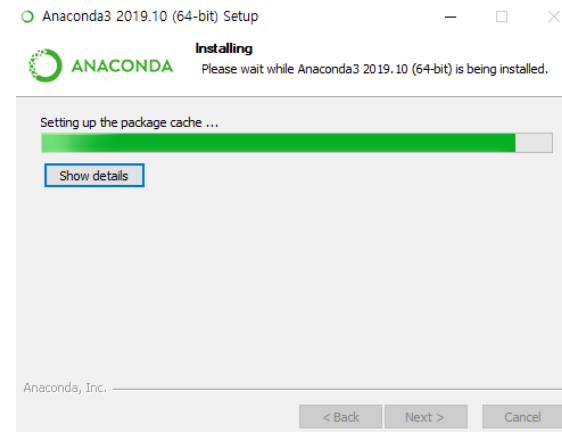
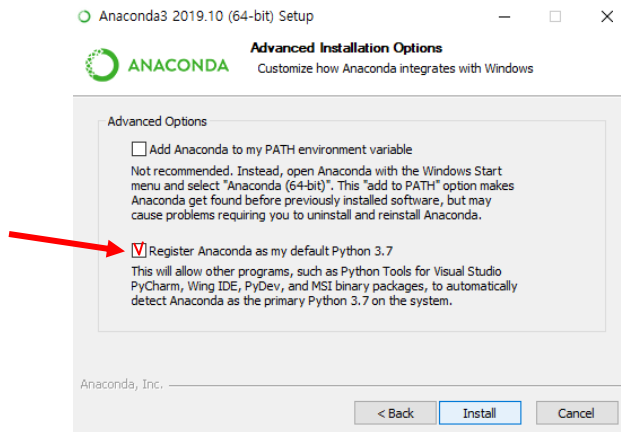
**Anaconda 2019.10 for Windows Installer**

Python 3.7 version	Python 2.7 version
<a href="#">Download</a>	<a href="#">Download</a>
64-Bit Graphical Installer (462 MB)	64-Bit Graphical Installer (413 MB)
32-Bit Graphical Installer (410 MB)	32-Bit Graphical Installer (356 MB)



- 설치 과정

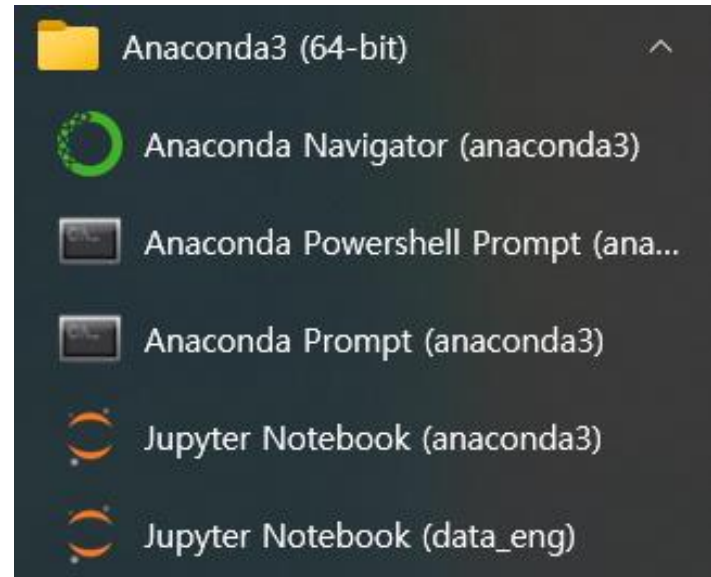




# Anaconda 둘러보기

## ➤ 윈도우 시작 메뉴

- ✓ Anaconda Prompt
- ✓ Anaconda Navigator  
: GUI
- ✓ Jupyter Notebook  
(가상환경)



# Anaconda 가상환경

## ➤ 가상환경을 사용하는 이유

- ✓ python 버전 관리가 용이하고 패키지의 충돌을 방지할 수 있기 때문

## ➤ 가상환경 생성 -> 콘솔(console)로


- ✓ CUI : Anaconda (Powershell) Prompt에서 명령어로
- ✓ GUI : Anaconda Navigator로
  - Environments 메뉴에서 생성

# Anaconda 가상환경 폴더구조

Anaconda Prompt (anaconda3)

```
(base) C:\Users\hwoon>conda env list
# conda environments:
#
base                * C:\Users\hwoon\anaconda3
class               C:\Users\hwoon\anaconda3\envs\class
data_eng            C:\Users\hwoon\anaconda3\envs\data_eng
test                C:\Users\hwoon\anaconda3\envs\test
```

로컬 디스크 (C:) > 사용자 > hwoon > anaconda3 > envs >

 ^	<input type="checkbox"/> 이름	수정한 날짜
	 class	2020-09-05 오후 4:03
	 data_eng	2021-12-25 오후 2:31
	 test	2022-11-23 오후 4:10

# Anaconda 가상환경 Dir

## ➤ 가상환경마다 별도의 폴더로 관리

- ✓ python 명령어를 실행하면 해당 폴더(가상환경)에 설치된 버전의 python 실행

```
(base) C:\Users\whwoon>conda activate data_eng
```

```
(data_eng) C:\Users\whwoon>python  
Python 3.7.9 (default, Aug 31 2020, 17:10:11) [MSC v.1916 64 bit (AMD64)] :: Anaconda, Inc. on win32  
Type "help", "copyright", "credits" or "license" for more information.  
>>> print("Hello")  
Hello  
>>> _
```

# 실습: 주피터 노트북

## ➤ Jupyter Notebook 특징

- ✓ Programming in the web browser (Chrome)
- ✓ Displayed "in-line" / code와 story 기술  
(활용사례) <https://github.com/ndb796>  
(사용법) <https://www.dataquest.io/blog/jupyter-notebook-tutorial/>

## ➤ Jupyter Notebook 설치

- ✓ (가상환경) conda install jupyter notebook

# 선수과정 : 파이썬

## ➤ 기본 문법

- ✓ Data Type, 제어문-조건문, 반복문
- ✓ 함수/Method 등

## ➤ 자료구조

- ✓ List, Dictionary 등

## ➤ Library

- ✓ Pandas, Numpy 등



End