
FROM A BLUE
M O O N

자기 소개

- 이름
- 자신을 기억하도록 만드는 특징?
- 쉴 때 뭐해요? (뭐하고 쉬어요?)

시간표

모든 주차					
시간/요일	월	화	수	목	금
10:00 ~ 10:30	월요일 그룹 활동	데일리 스크럼			
10:30 ~ 11:00					주간 수업
11:00 ~ 11:30					
11:30 ~ 12:30	주간 수업	개발 활동	개발 활동	개발 활동	주간 피드백
12:30 ~ 14:00	점심 시간				
14:00 ~ 14:30		그룹세션	그룹세션	그룹세션	개발 활동 및 데모(프로젝트기간)
14:30 ~ 15:00	주간 수업				
15:00 ~ 16:00					
16:00 ~ 17:00					스쿼드 세션
17:00 ~ 18:00	개발 활동	개발 활동	개발 활동	개발 활동	개발 활동
18:00 ~ 18:30	성장노트 작성 + 그룹 회고				

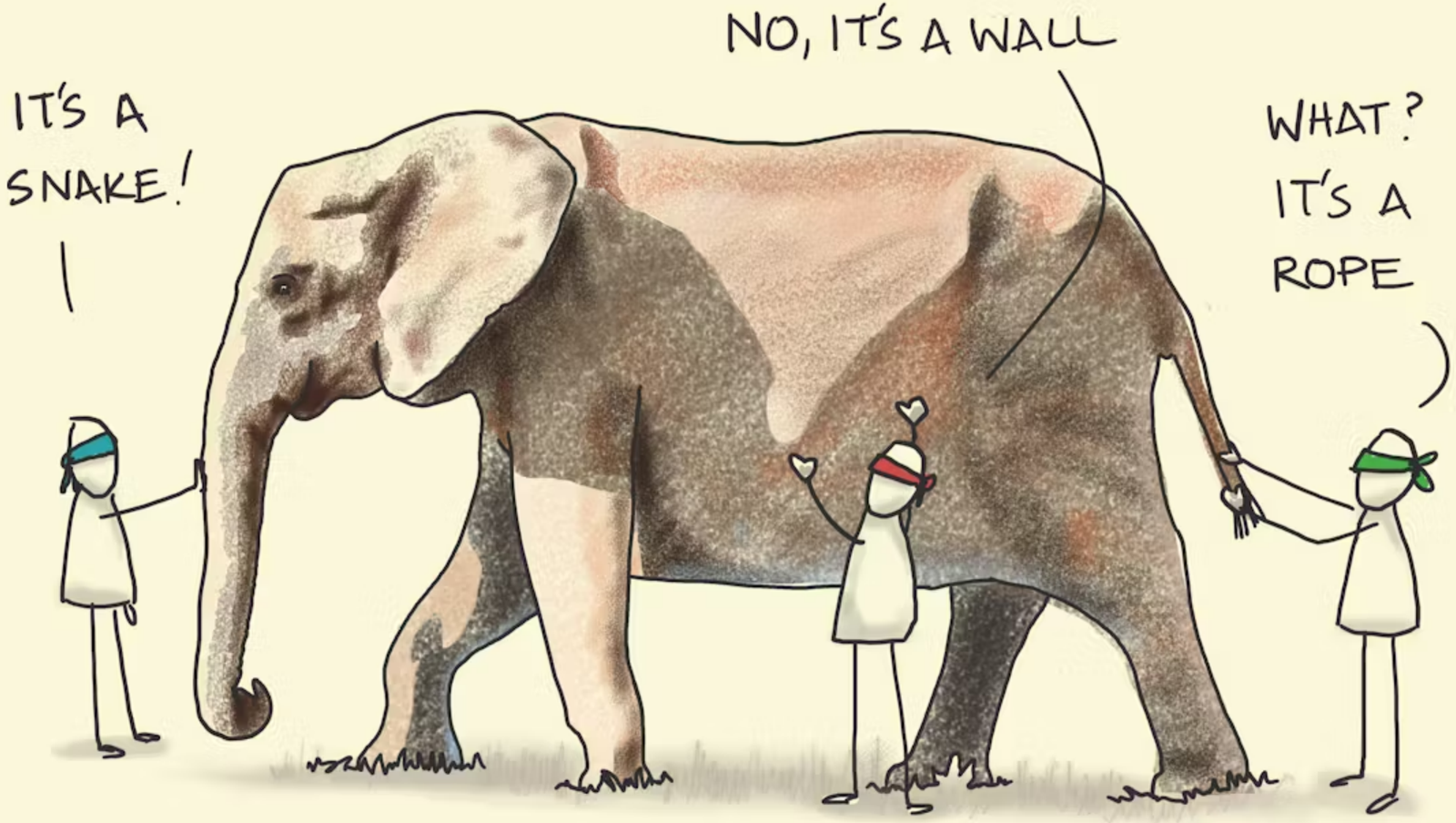
Honor Code

- 어떤 경우에도 동료를 존중하고, 함부로 대하지 않는다.
- 혼자가 아닌 함께 성장하려고 노력한다.
- 지식을 적극적으로 나누고, 서로를 돕는다.

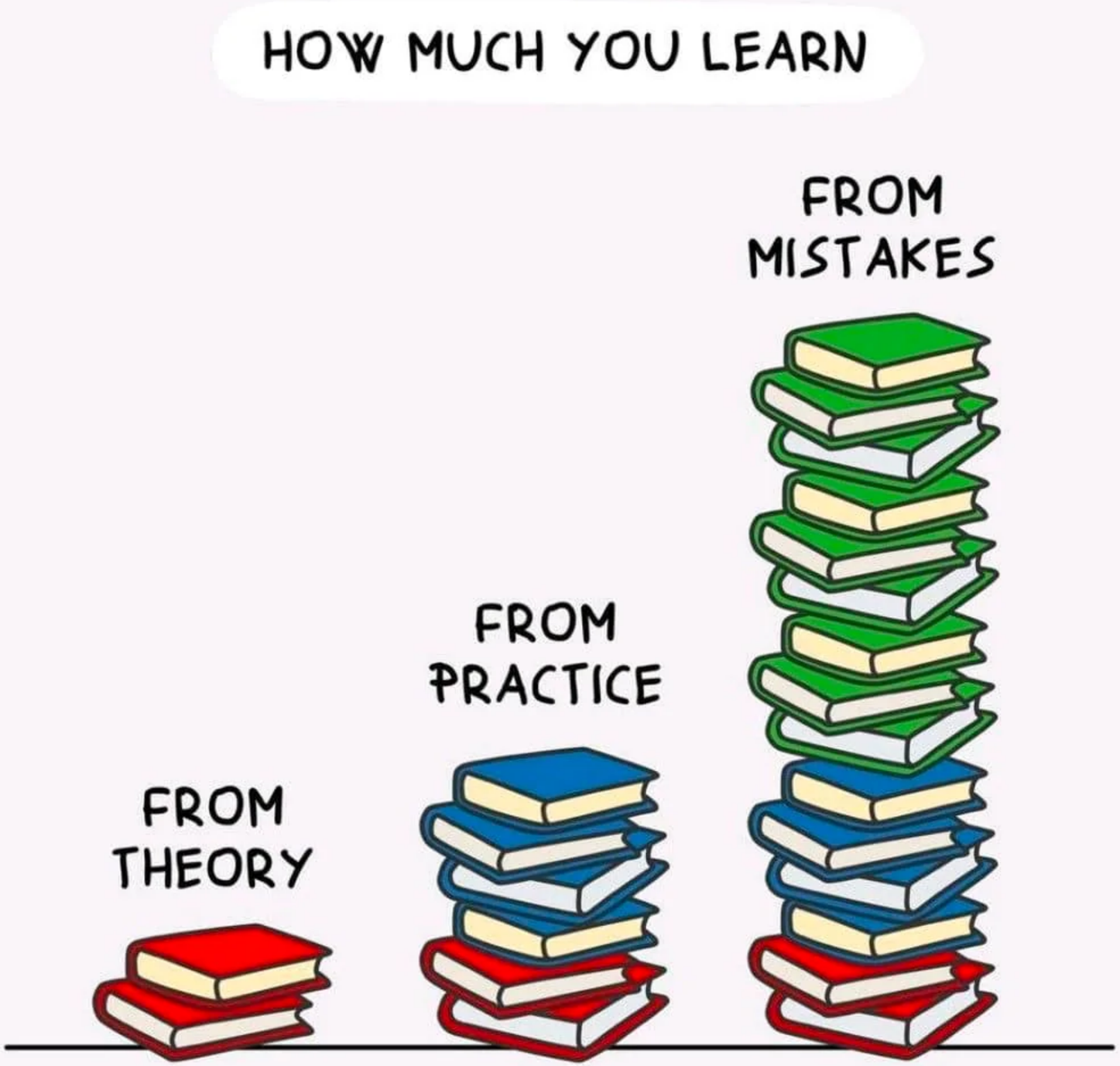
현업에서 학습을 한다는 것은?

THE BLIND AND THE ELEPHANT

OUR OWN EXPERIENCE IS RARELY THE WHOLE TRUTH

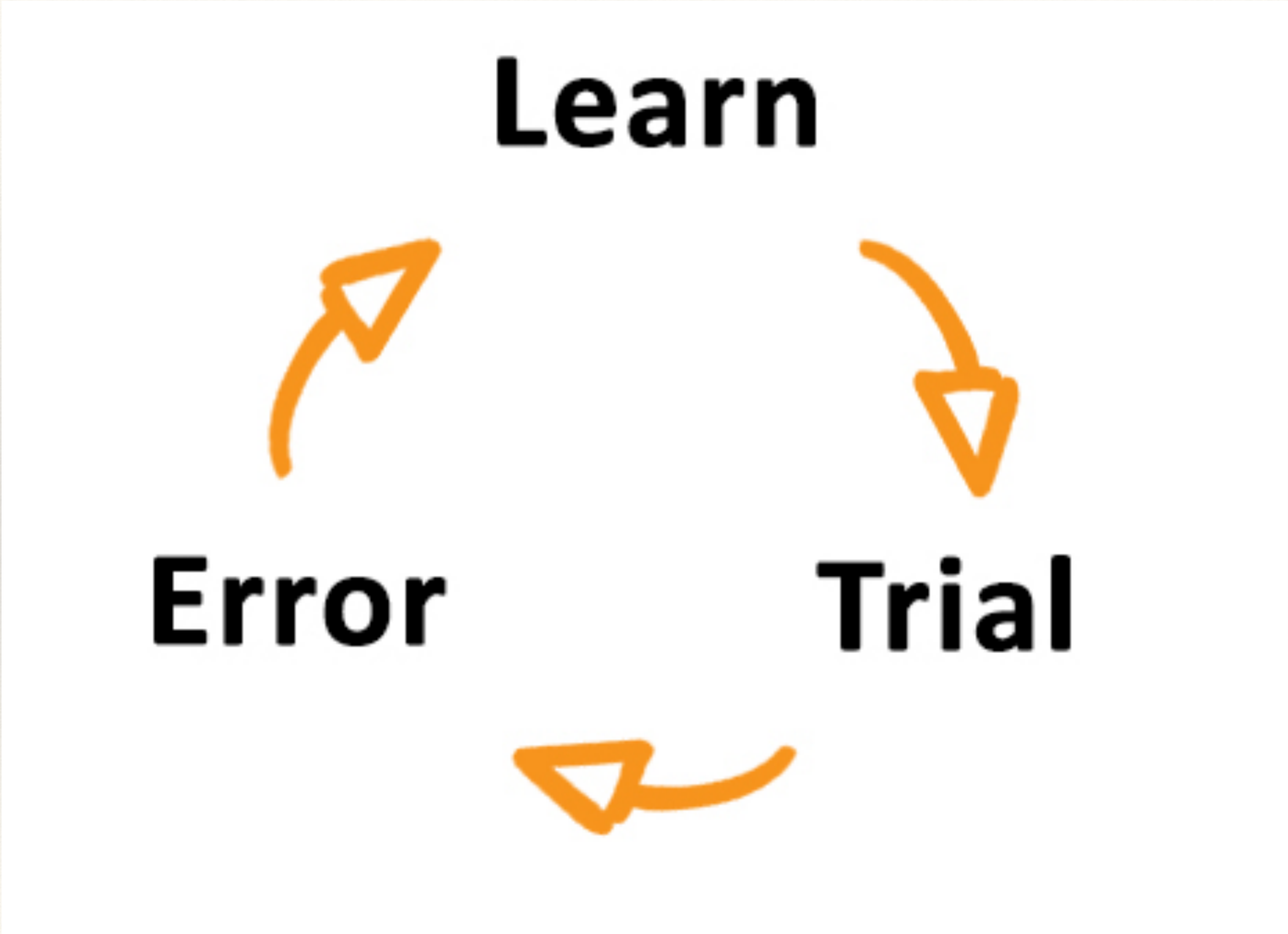


sketchplanations



source: reddit

Trial and Error



HMG 소프티어 부트캠프 6기

Introduction to Data Engineering

Dano Lee

Modern Data Ecosystem


DATA

Why use Data?

Monetary Value

- Data를 금전적 가치가 있는 정보로 바꾼다.
- 처리비용이 비싸다는 건 '그 비용보다 훨씬 더 가치있는 정보가 나와야 한다'는 것.

Problem - Solution

Monetary Value

Data + Technology

“If I were given one hour to save the planet,
I would spend 59 minutes defining the problem
and one minute resolving it,”

Why Data Matters

We often rely on **heuristic** decision-making — fast, intuitive, but prone to bias.

But we need to shift toward **evidence-based** decisions — grounded in facts, not just instincts.

Data provides the evidence we need to make informed, reliable, and scalable decisions.

Actionable Data

information that can be acted upon or information that gives enough insight into the future that the actions that should be taken become clear for decision makers.

source: coursera

Question

- CPI - Acquisition Funnel

Data Professionals

- Data Engineers
- Data Analysts
- Data Scientists
- Business Analysts
- Business Intelligence Analysts

source: coursera

Data Professionals

- Data Engineering converts raw data into usable data
- Data Analytics uses data to generate insights
- Data Scientists use Data Analytics and Data Engineering to predict the future using data from the past
- Business Analysts and Business Intelligence Analysts use these insight and predictions to drive decisions that benefit and grow their business

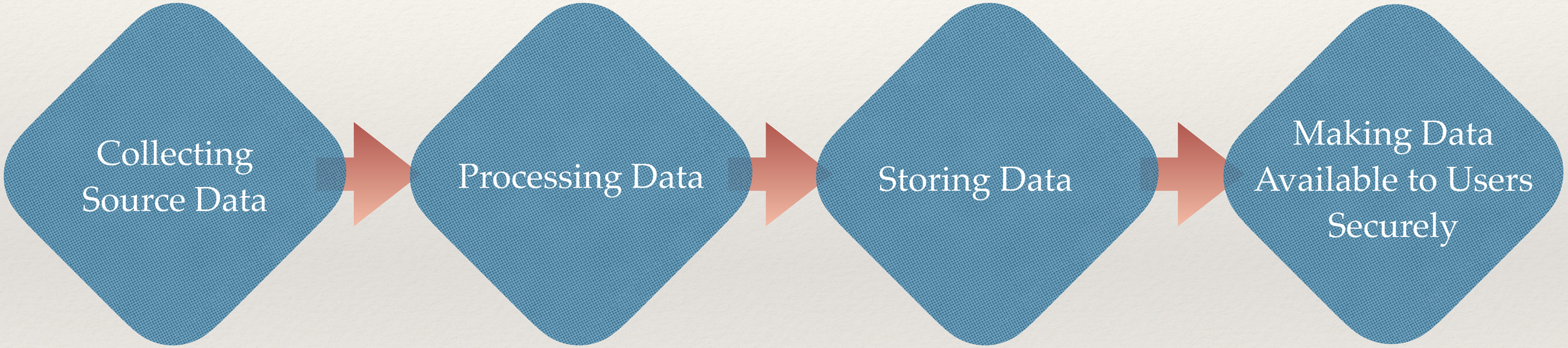
source: coursera

Data Engineering

The goal of Data Engineering is to make quality data available for analytics and decision-making. And it does this by collecting raw source data, processing data so it becomes usable, storing data, and making quality data available to users securely.

source: coursera

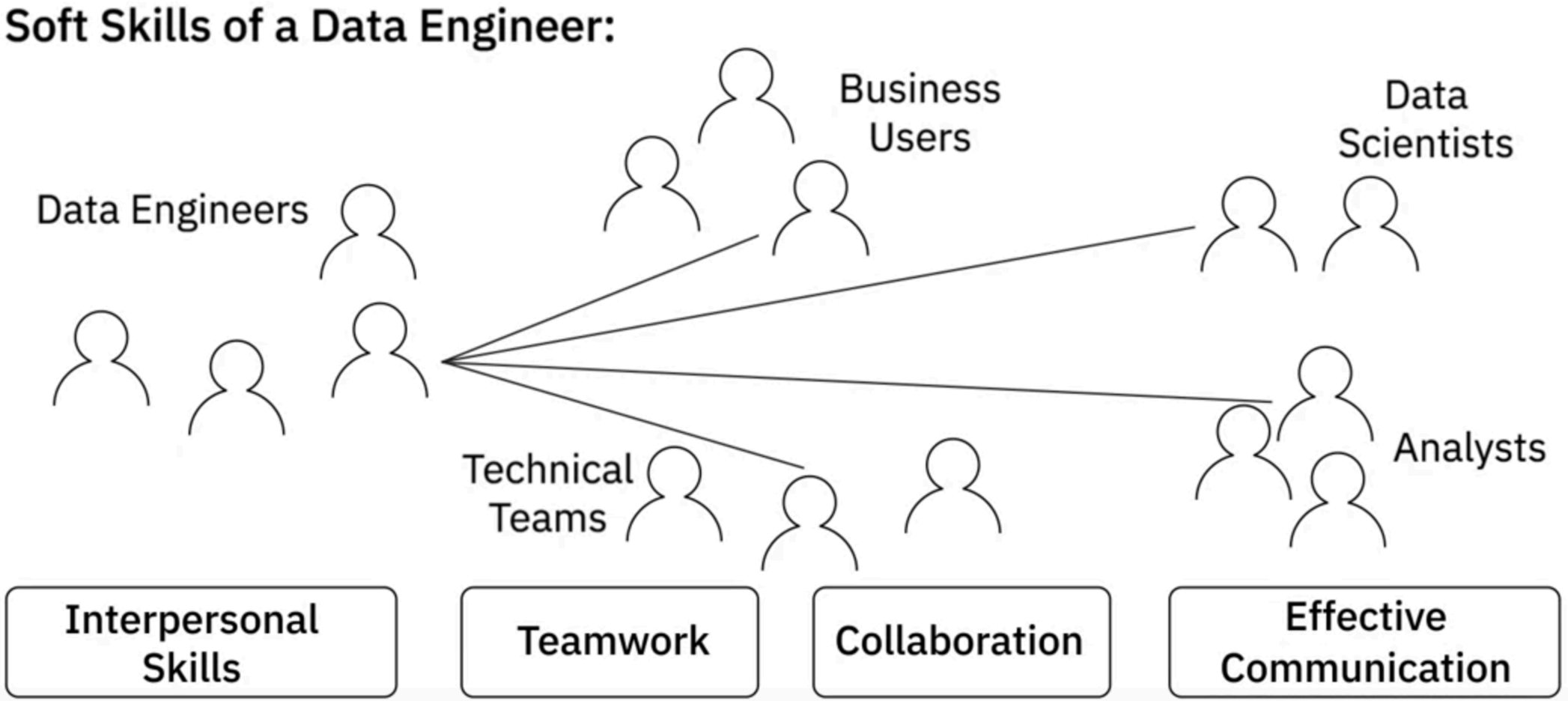
Data Engineering



Data Engineering is a team sport

source: coursera

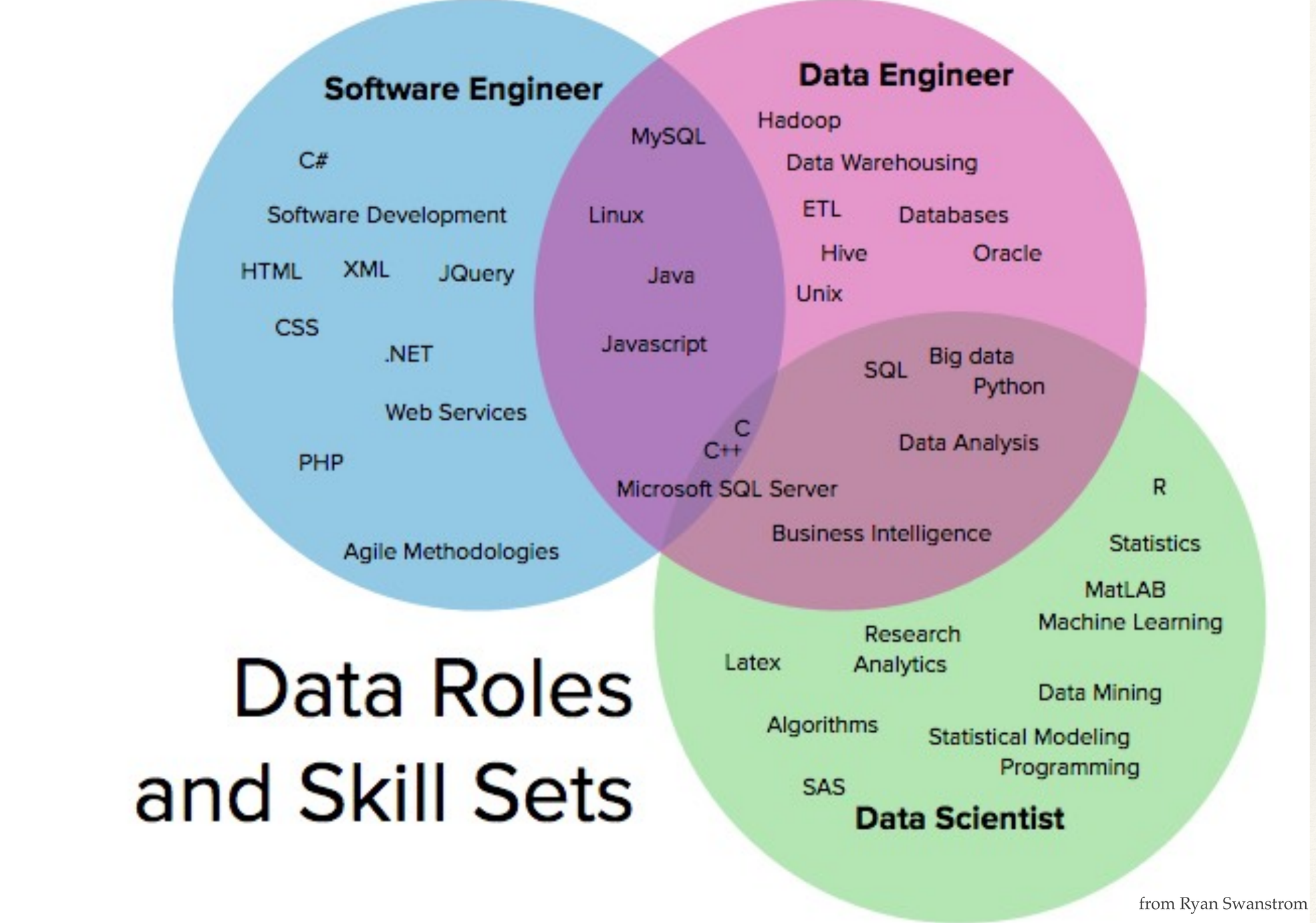
Soft Skills



source: coursera

What’s the moral of the story?

What if you are…?



from Ryan Swanstrom

Technical Skills

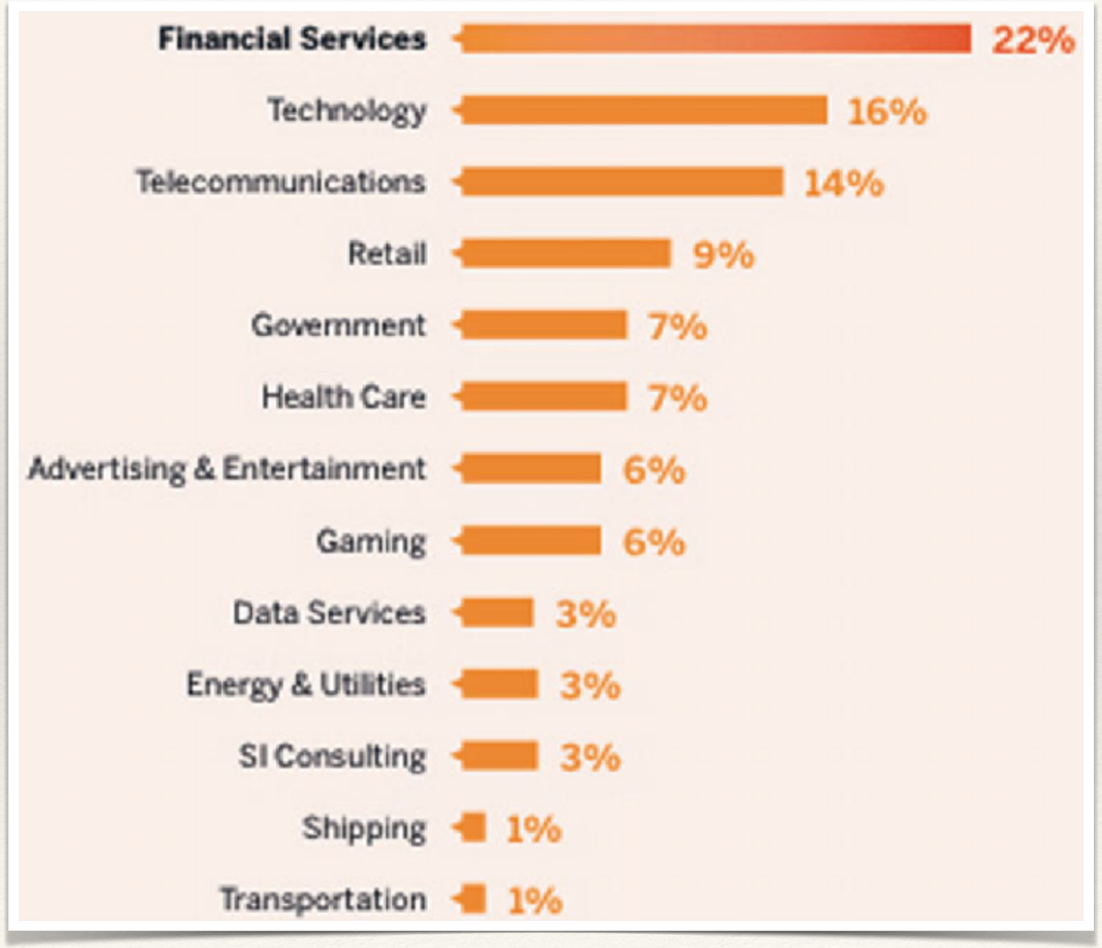
- Working with different operating systems and infrastructure components such as virtual machines, networks, and application services.
- It also includes working with databases and data warehouses, data pipelines, ETL tools, big data processing tools, and languages for querying, manipulating, and processing data.

Functional Skills

- Convert business requirements into technical specifications
- Work with the complete software development lifecycle
  - Ideation -> Architecture -> Design -> Prototyping -> Testing -> Deployment -> Monitoring -> Optimization
- Understand data's potential application in business
- Understand risks of poor data management
  - Data quality, Data privacy, Security, and Compliance

Big Data Usage by Industry

- Where do you want to go?



source: Utilizing Big Data for Health Care Automation: Obligations, Fitness and Challenges

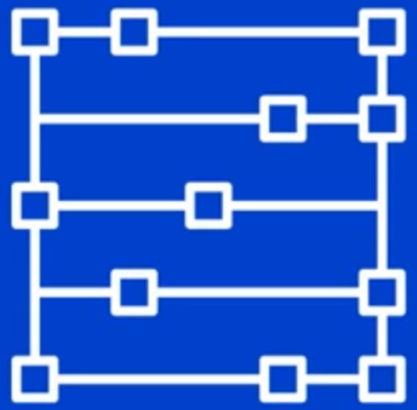
Data Types

Types of Data

- Structured Data
- Semi-structured Data
- Unstructured Data

source: coursera

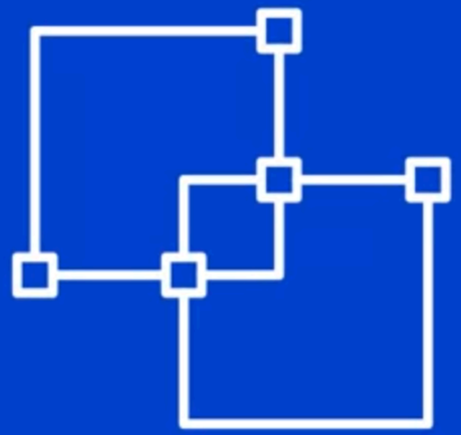
Structured Data



- SQL Databases
- Online Transaction Processing
- Spreadsheets
- Online forms
- Sensors GPS and RFID
- Network and Web server logs

source: coursera

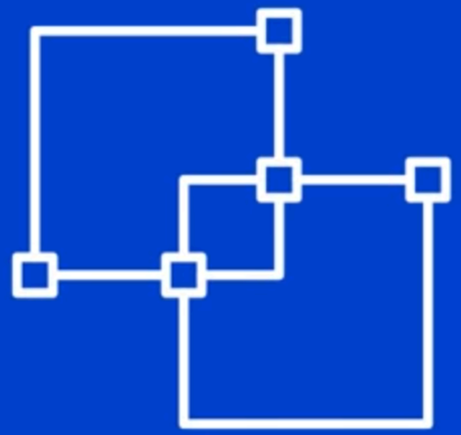
Semi-Structured Data



- E-mails
- XML and other markup languages
- Binary executables
- TCP/IP packets
- Zipped files
- Integration of data

source: coursera

Semi-Structured Data



XML

JSON

Allow users to

Define Tags

Attributes

To store data

source: coursera

Unstructured Data

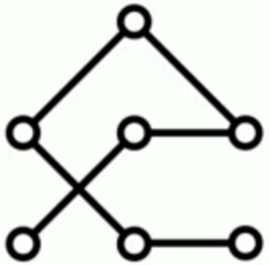
- Web pages
- Social media feeds
- Images in varied file formats
- Video and Audio files
- Documents and PDF files
- PowerPoint presentations

source: coursera

Unstructured Data

Files and Docs

Manual Analysis



NoSQL

Analysis Tools

source: coursera

Home Address?

- String?
- Structured data?
- Semi-structured data?

결국 중요한 정보가 어떤 형태의 데이터냐?

- 동영상 화일: 동영상의 내용 vs 동영상의 metadata
- Email: Email의 body vs Email의 header
- 테이블 (표): cell의 값 vs ????

Question

- Table: Person
- columns: last name, first name, identification number, address, etc
- address?
- etc column에 json 형식의 데이터가 들어있다.

Python Programming

PEP 8 – Style Guide for Python Code

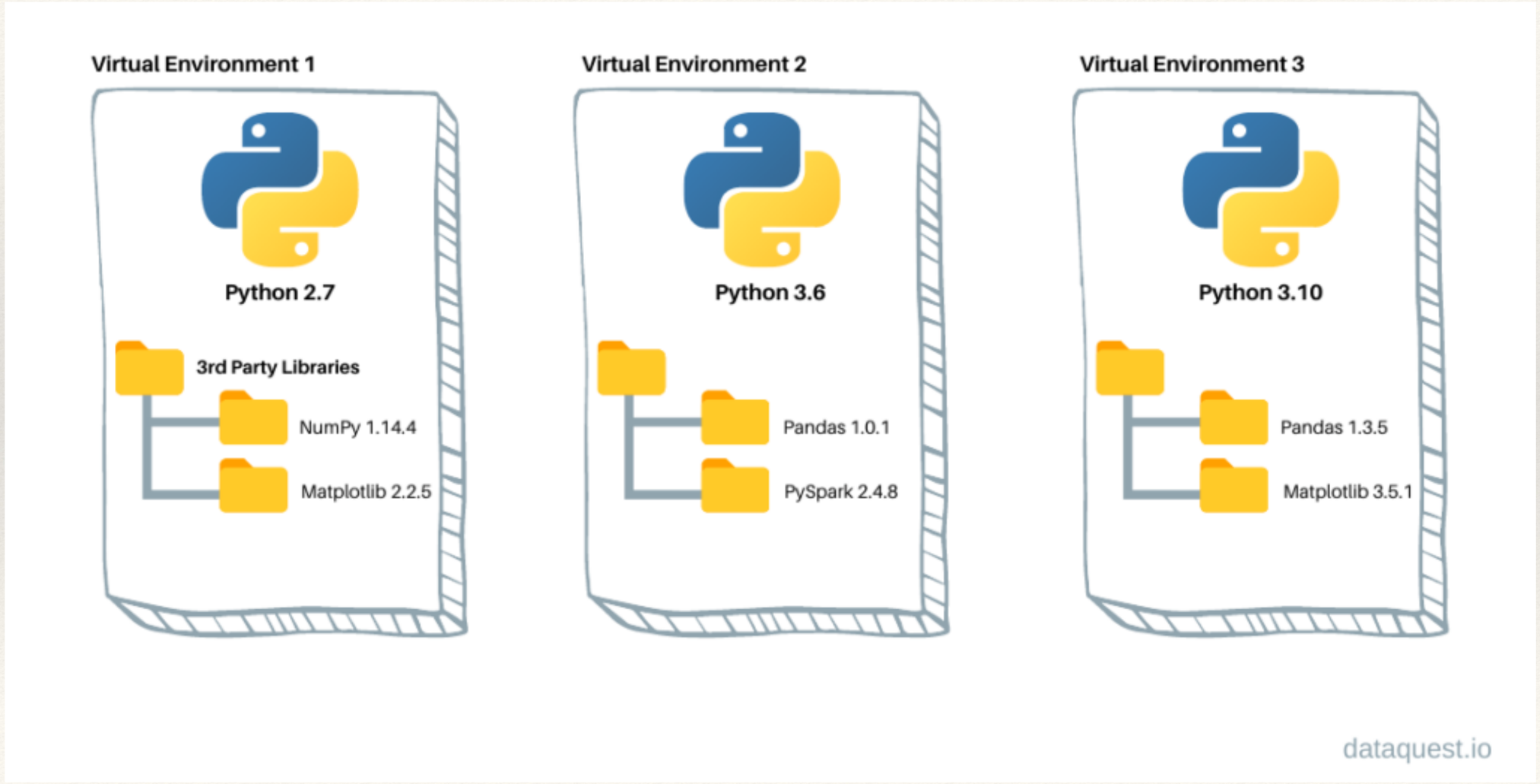
- https://peps.python.org/pep-0008/
  - Indentation: Use 4 spaces per indentation level
  - Tabs or Spaces?
    - Python disallows mixing tabs and spaces for indentation.

- Pylint: a static code analyser for Python. Pylint analyses your code without actually running it. It checks for errors, enforces a coding standard, looks for code smells, and can make suggestions about how the code could be refactored.

pyenv

- Simple Python Version Management: pyenv
- https://github.com/pyenv/pyenv

venv


Virtual Environment 1
Python 2.7
3rd Party Libraries
NumPy 1.14.4
Matplotlib 2.2.5
Virtual Environment 2
Python 3.6
Pandas 1.0.1
PySpark 2.4.8
Virtual Environment 3
Python 3.10
Pandas 1.3.5
Matplotlib 3.5.1
dataquest.io

GitHub

- GitHub
  - personal repositories: wiki에 모든 리서치한 내용들을 담으세요.
    - missions 폴더 아래에 주차별 폴더를 만들고 과제 화일을 담으세요. ex)W1, 1주차

- git
  - .gitignore: code로 관리하지 않아도 되는 것들. ex) .venv

- GitHub desktop (optional)

Tools for Data Analysis

JupyterLab and Jupyter Notebook

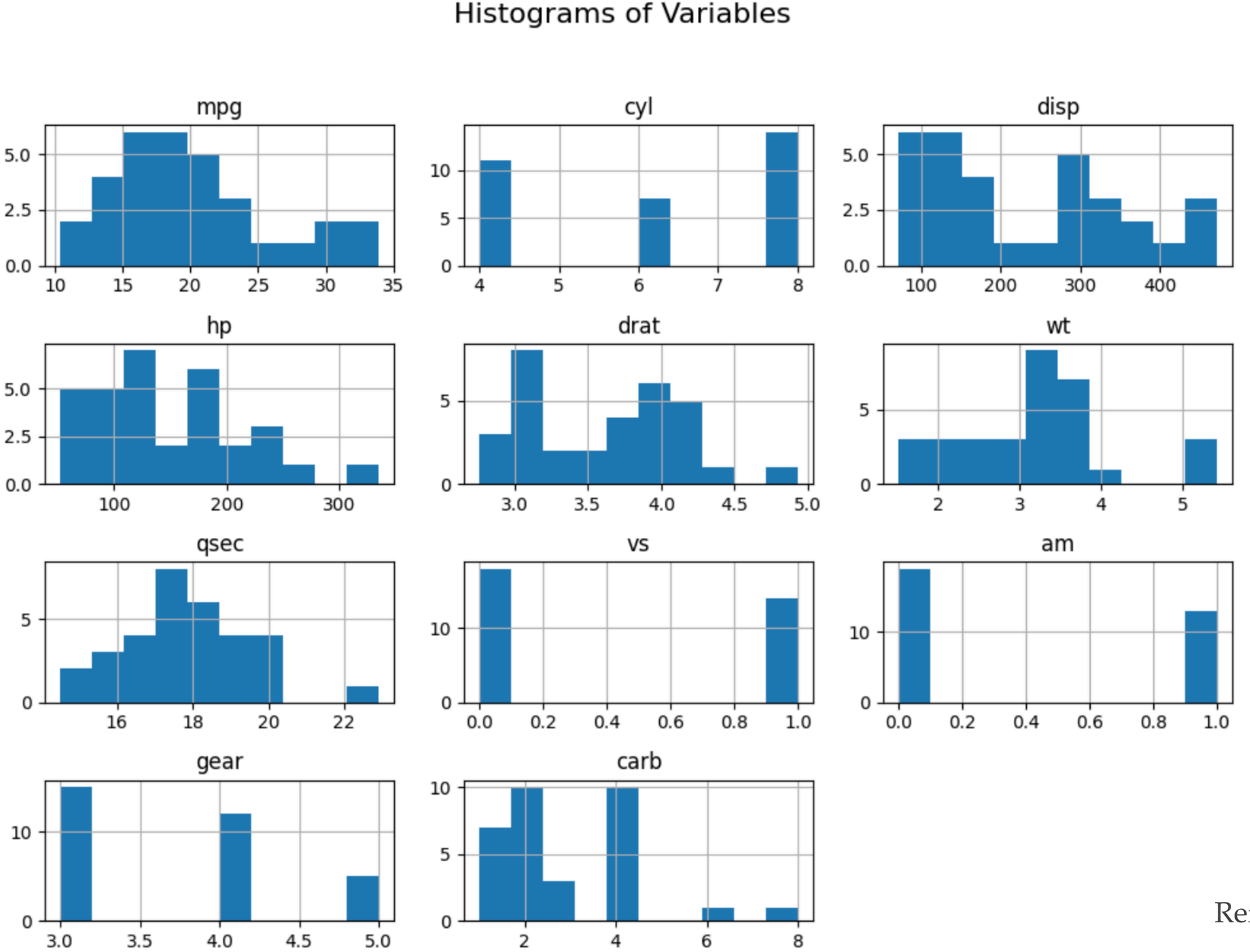
- https://jupyter.org/
- **JupyterLab** is the latest web-based interactive development environment for notebooks, code, and data. Its flexible interface allows users to configure and arrange workflows in data science, scientific computing, computational journalism, and machine learning. A modular design invites extensions to expand and enrich functionality.
- The **Jupyter Notebook** is the original web application for creating and sharing computational documents. It offers a simple, streamlined, document-centric experience.

Pandas

- pandas is a fast, powerful, flexible and easy to use open source data analysis and manipulation tool, built on top of the Python programming language.
- refer to API reference

Matplotlib

- Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python. Matplotlib makes easy things easy and hard things possible.
- refer to API reference



Ref: matplotlib

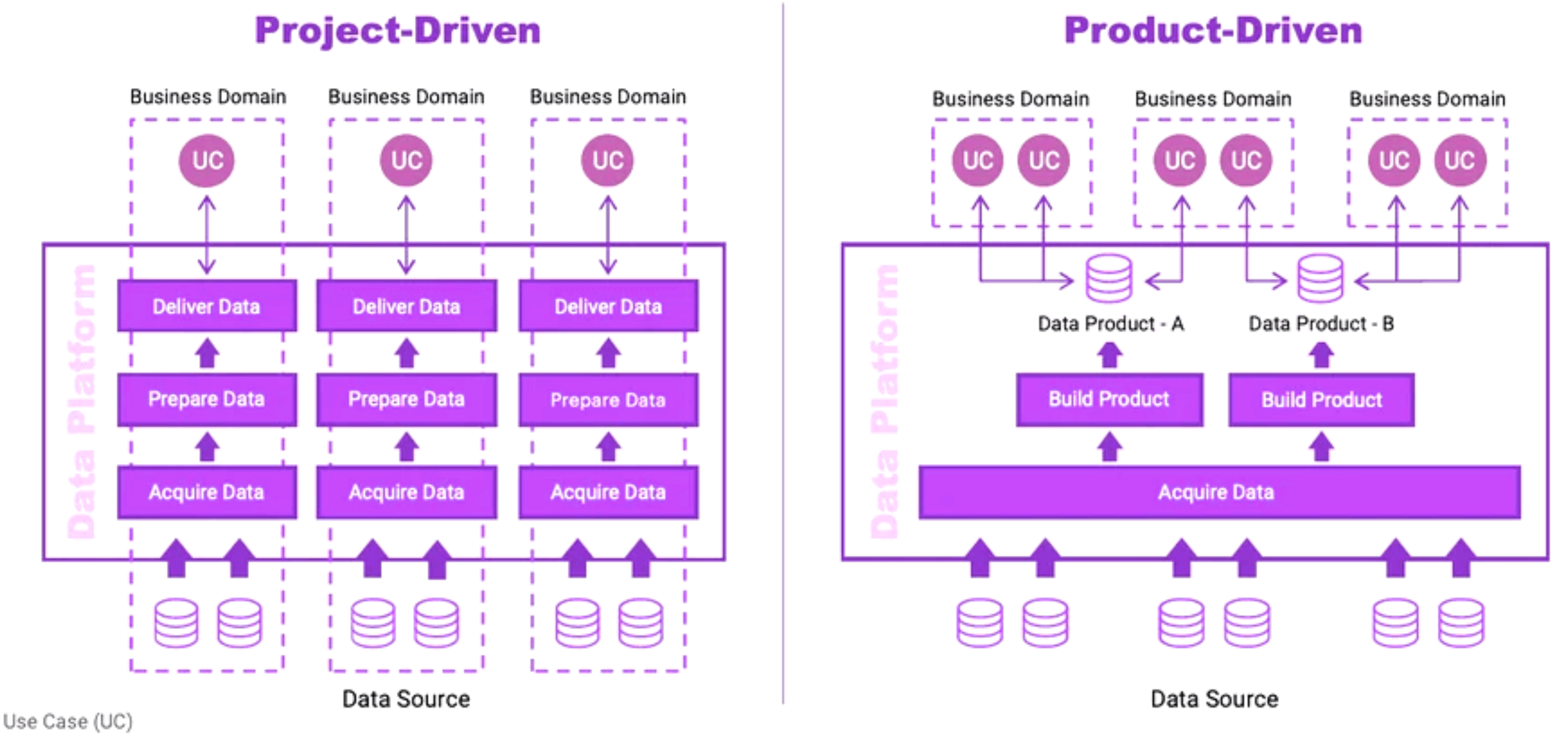
Let’s Pause Here

Data Product

Data Product

- A data product is a reusable data asset that bundles data together with everything needed to make it independently usable by authorized consumers.

source: k2view.com



source: k2view.com

우리는 이미 몇개의 Data Product을 만들어 봤다?

(사실은 Prototype이지만...)

Why Data Product?

McKinsey says,

Data-driven companies are
23x more likely to acquire customers
19x more likely to be profitable

source: k2view.com

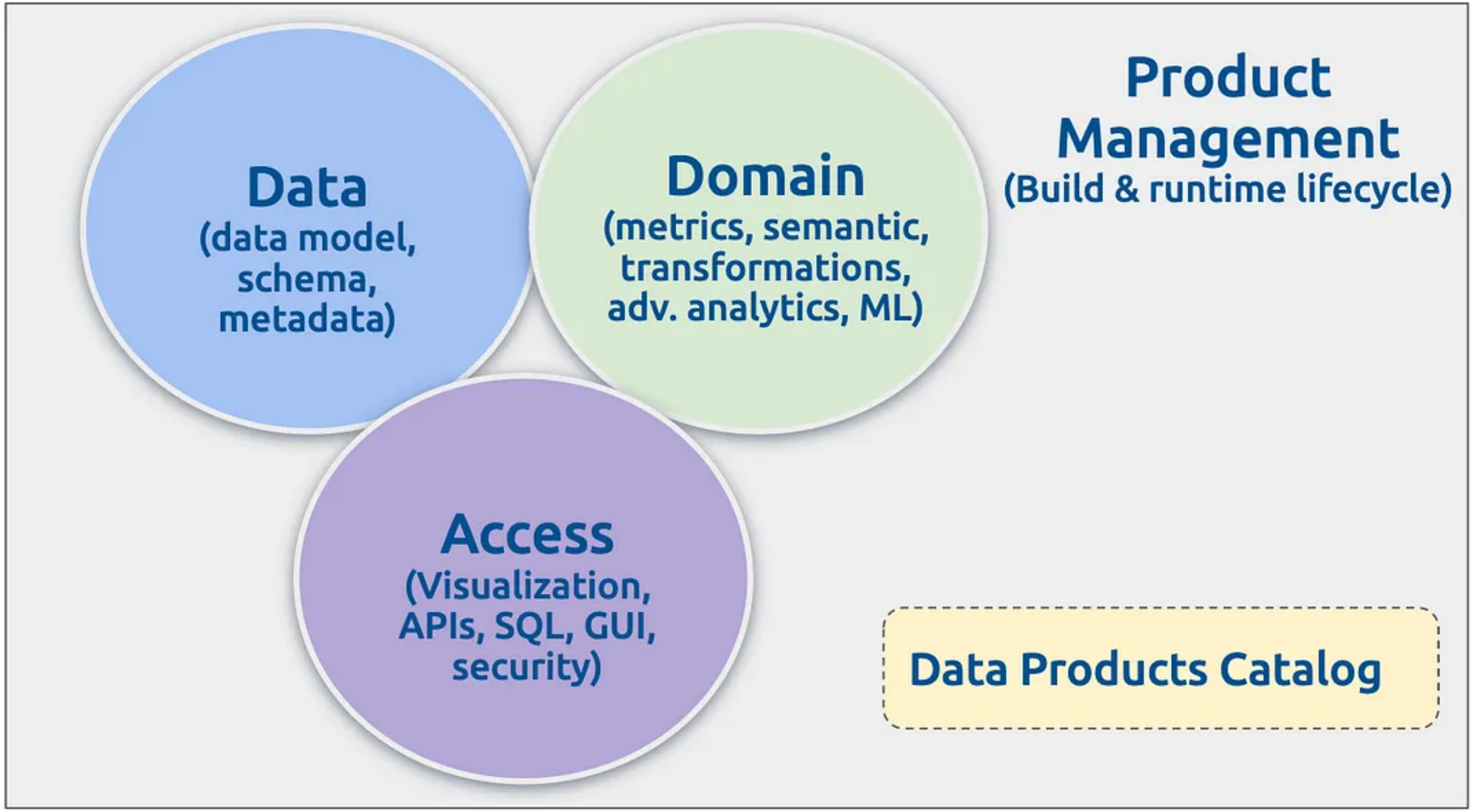
Why Data Product?

Dark data:

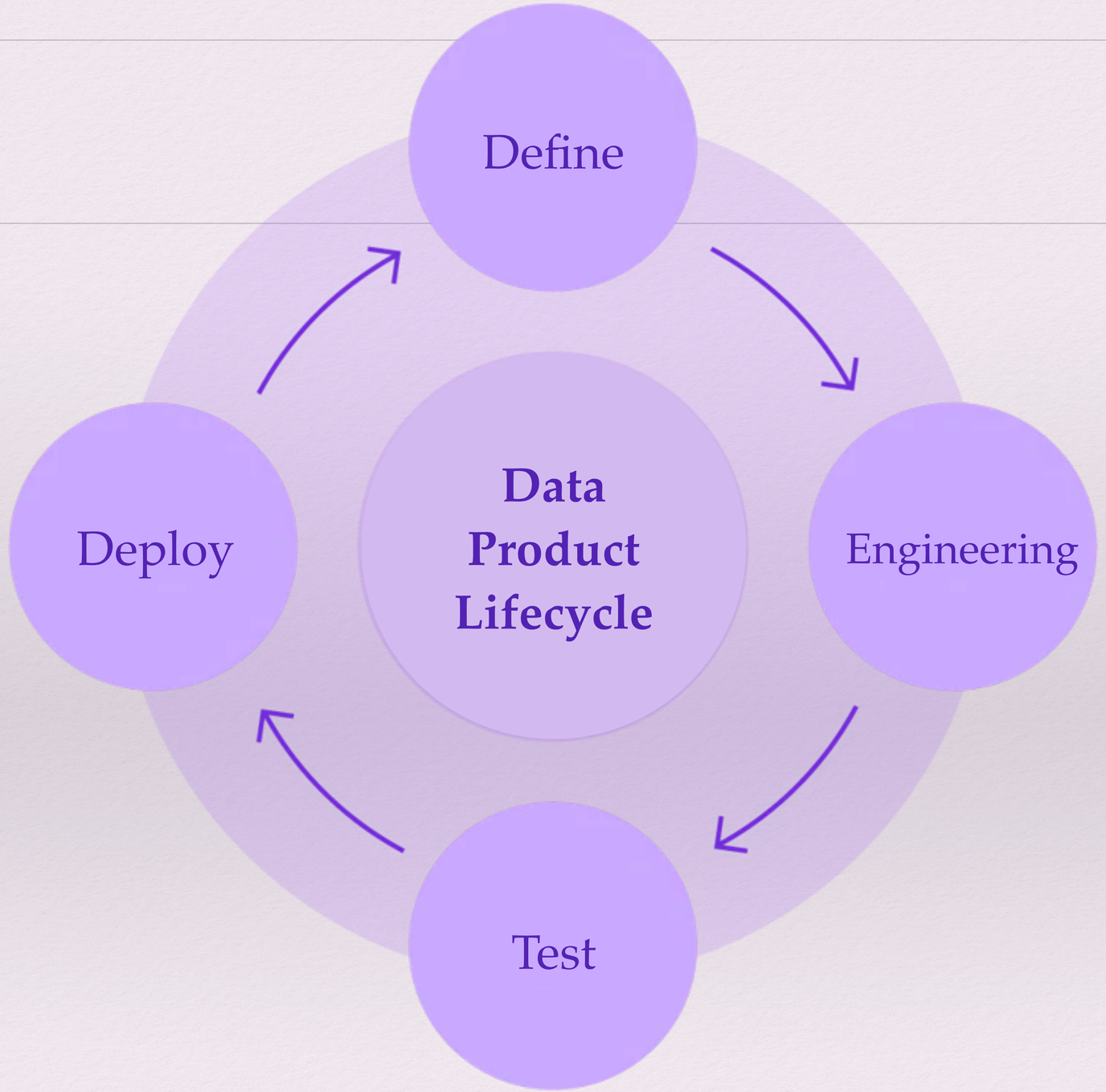
Over 80% of enterprise data is “in the dark”, in the sense that it's inaccessible and not being used – to drive business decisions or to improve customer experiences or operational efficiencies. It's only weighing companies down.

source: k2view.com

Data Products Overview



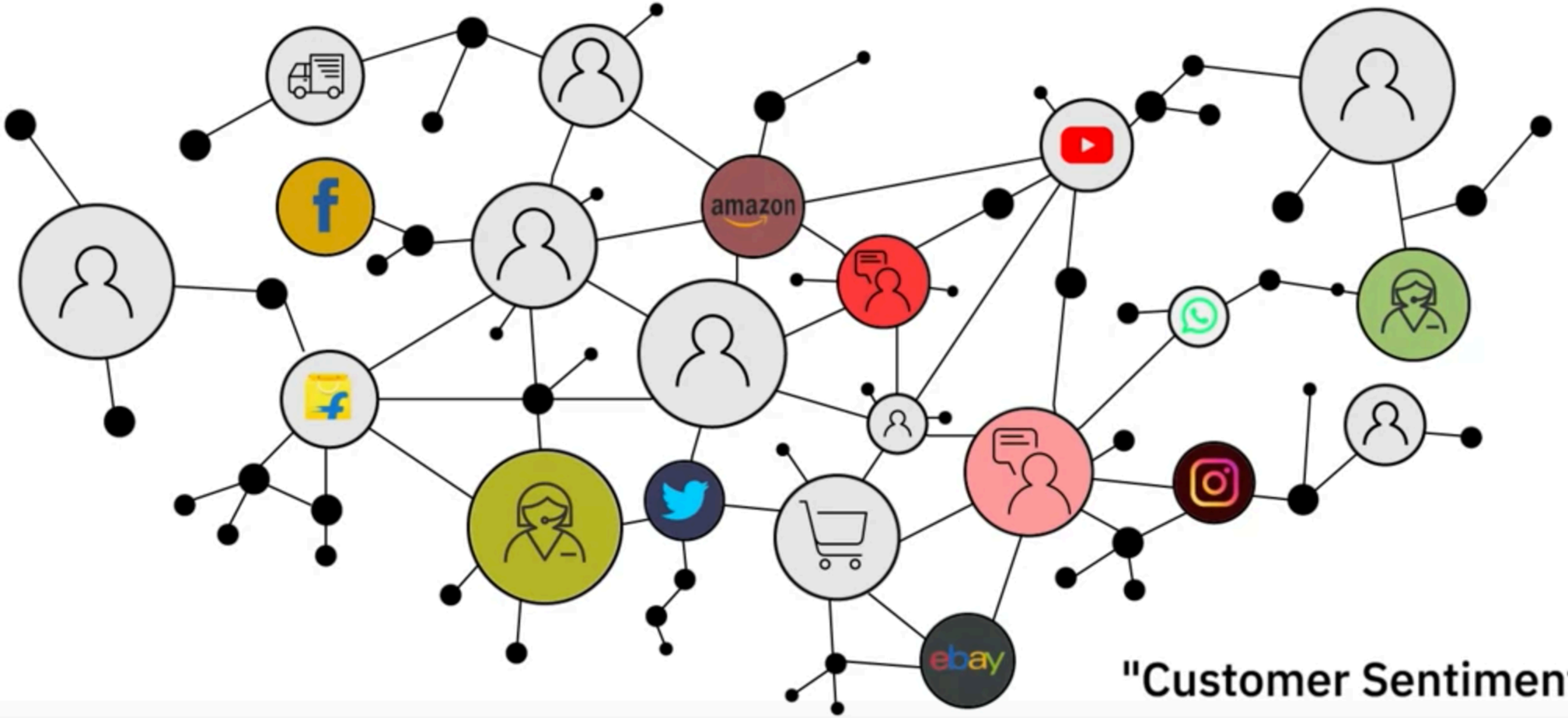
source



source: k2view.com

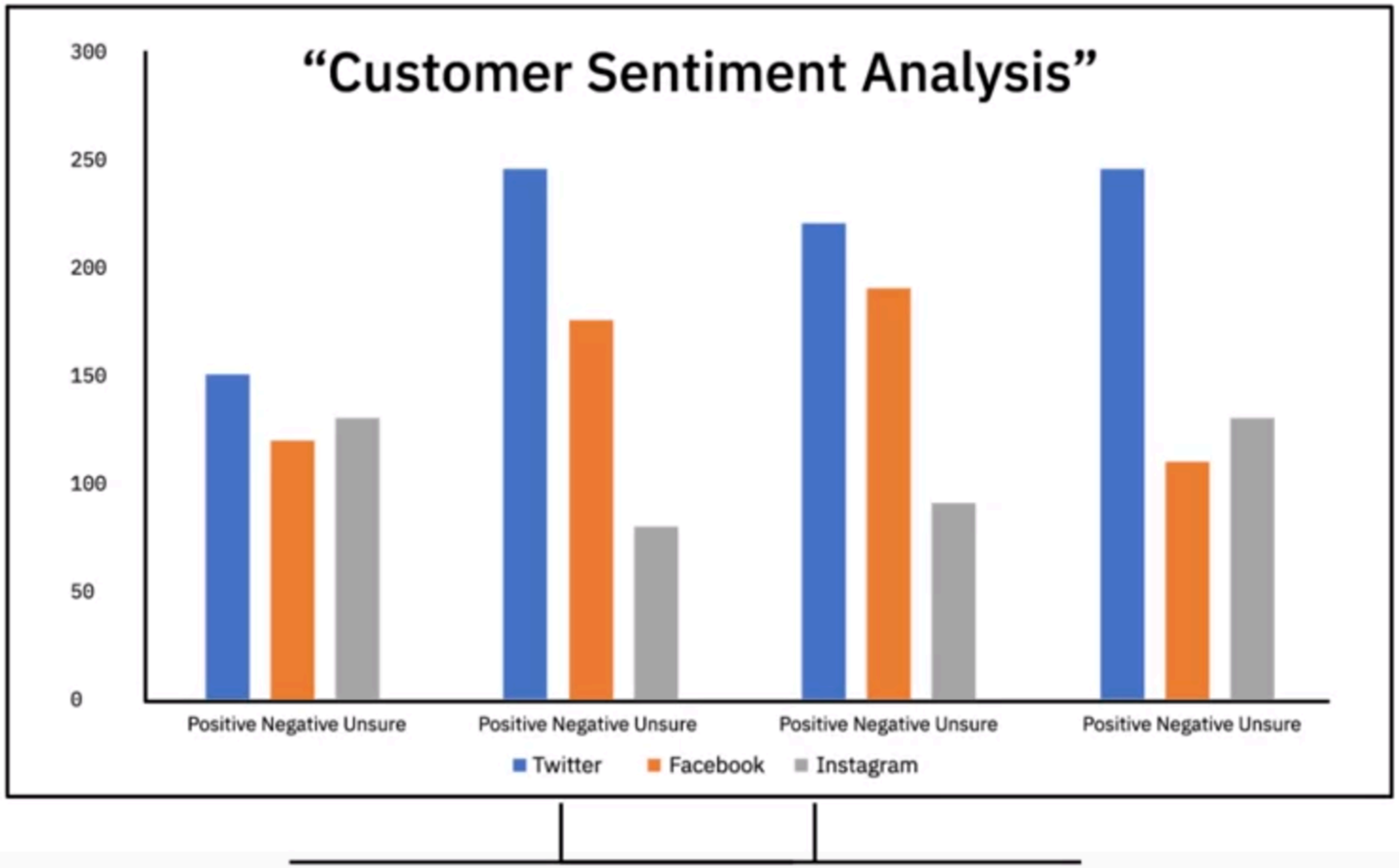
Scenario: Sentiment Analysis

The Need



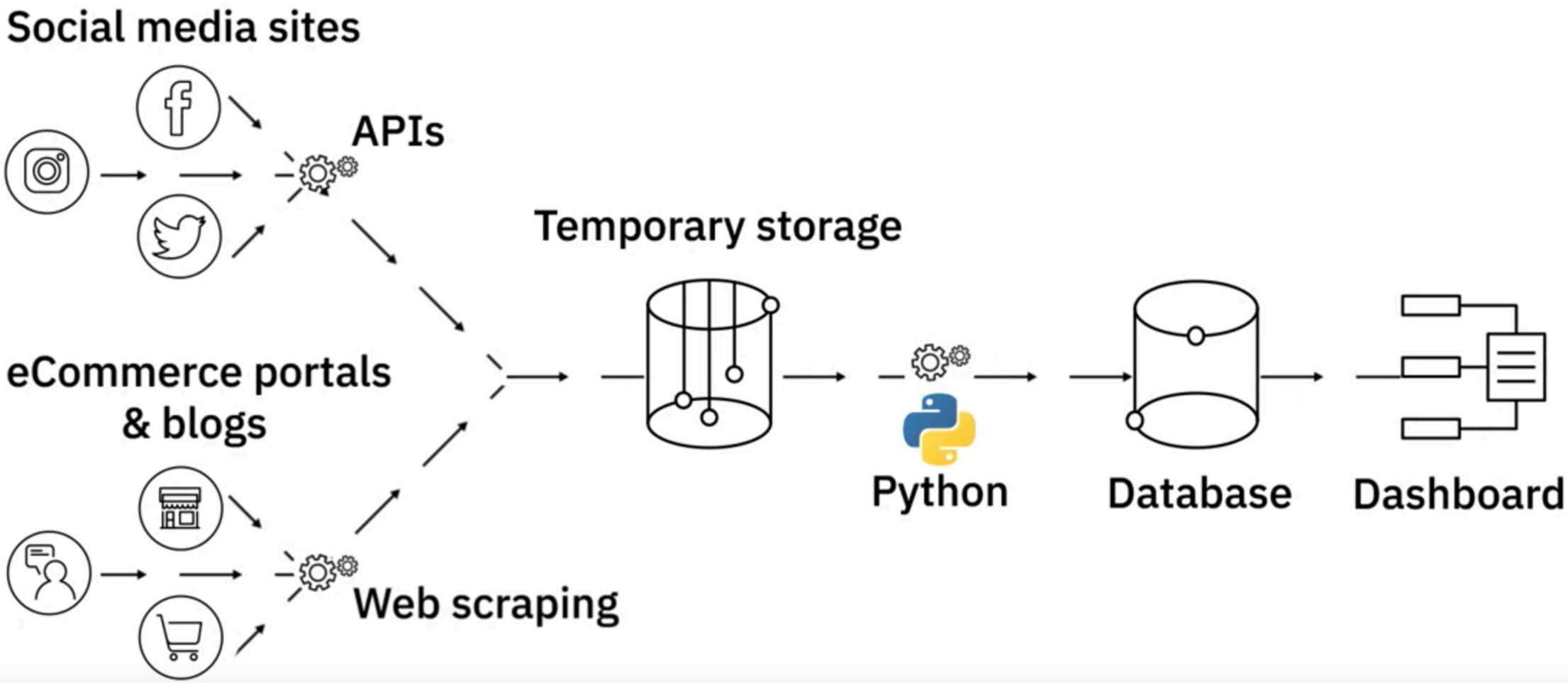
source: coursera

The Goal



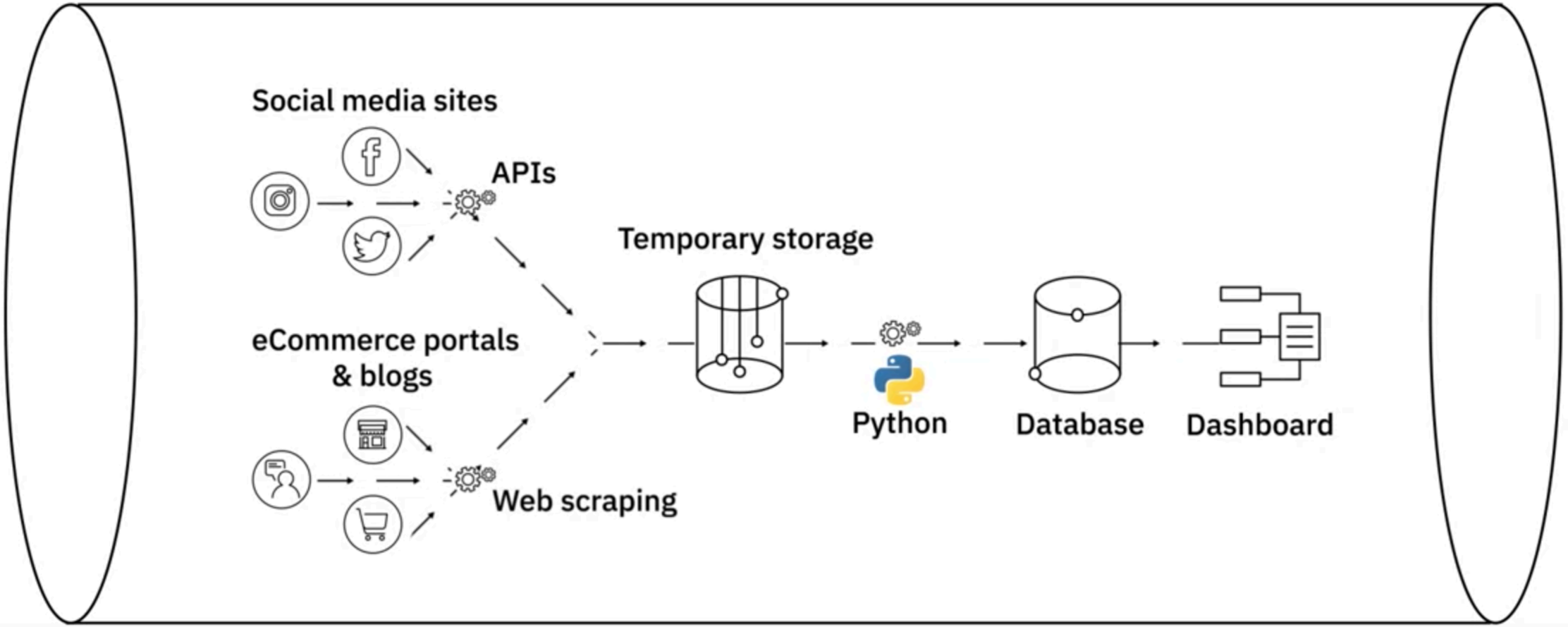
source: coursera

The Solution

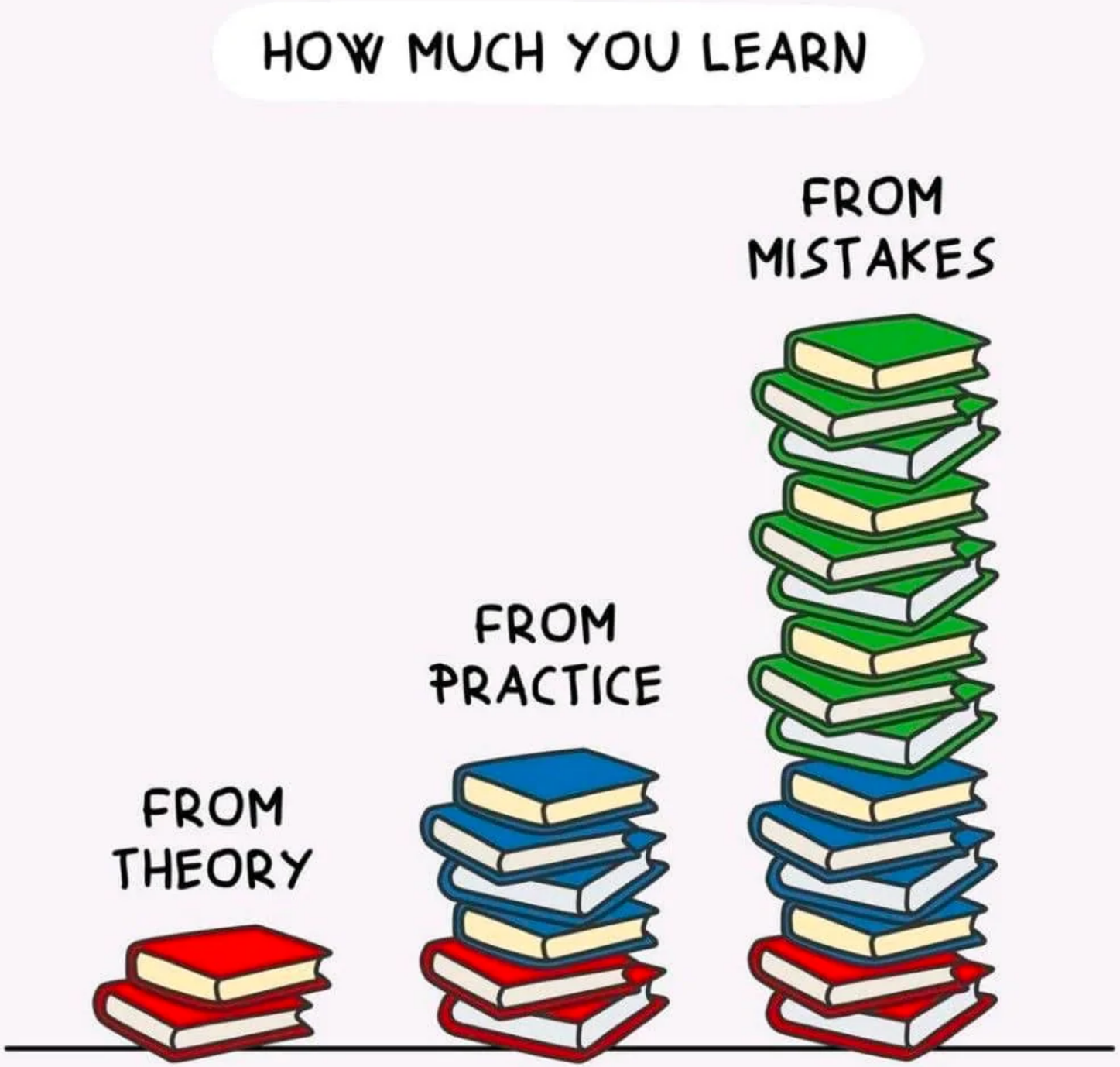


source: coursera

Next Steps



source: coursera



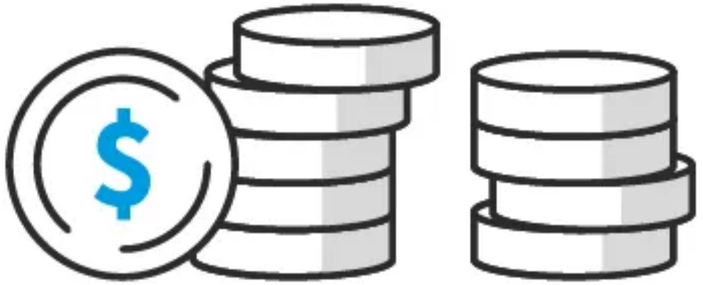
source: reddit

Prototyping

The 1-10-100 Rule: How Early Prototyping Prevents Costly Errors in Advance

Prevention Cost: $1

E.g., evaluating usability through early paper prototypes



Correction Cost: $10

E.g., fixing usability errors discovered through usability tests with hi-fidelity prototypes

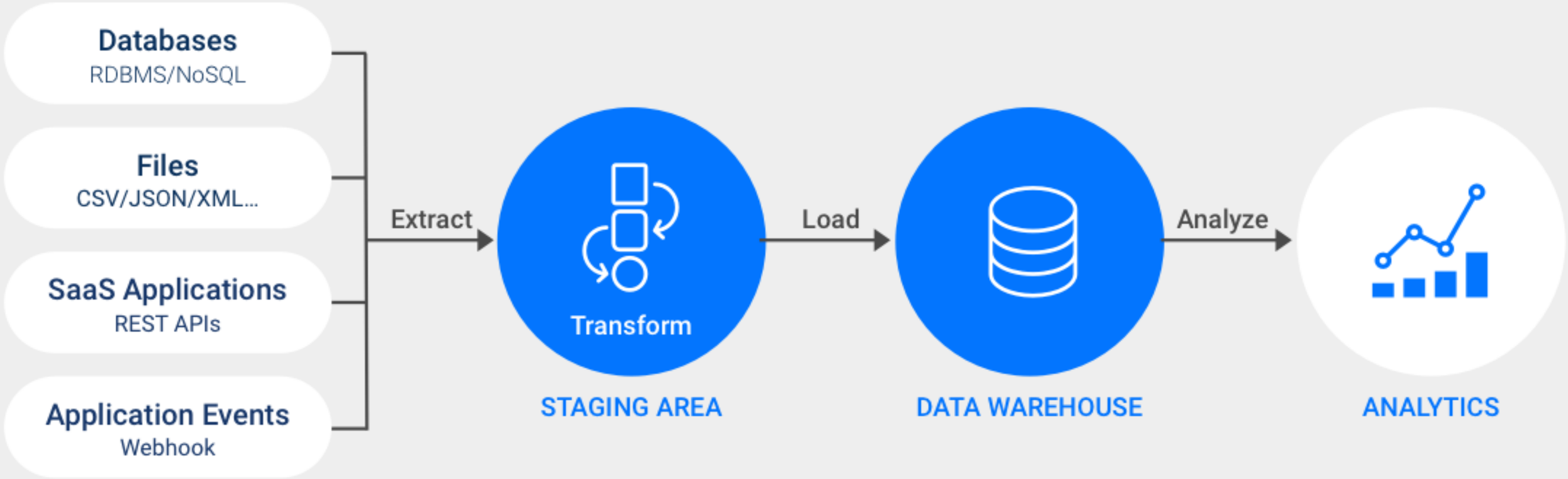
Correction Cost: $100

E.g., fixing the code and lost revenue from an error in the final product

Interaction Design Foundation
interaction-design.org

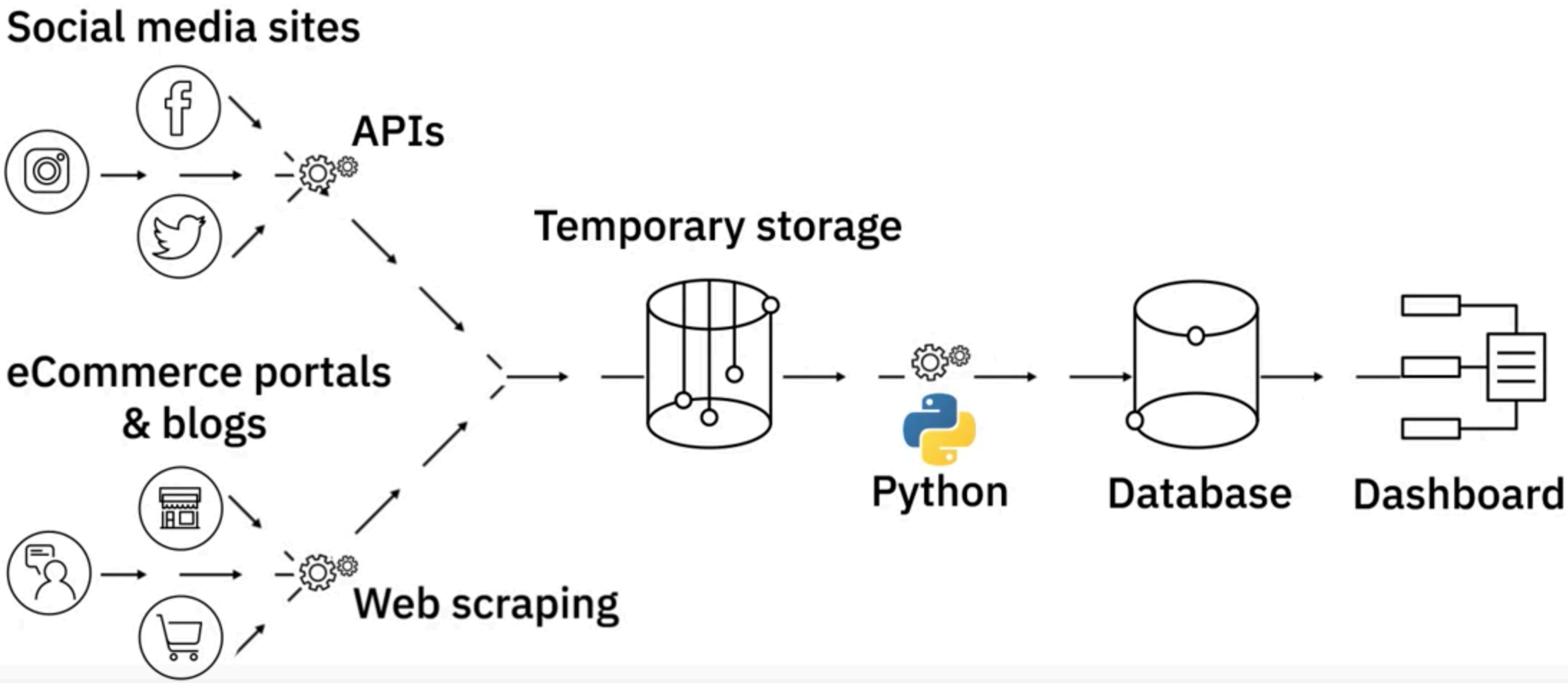
Extract, Transform & Load

ETL PROCESS



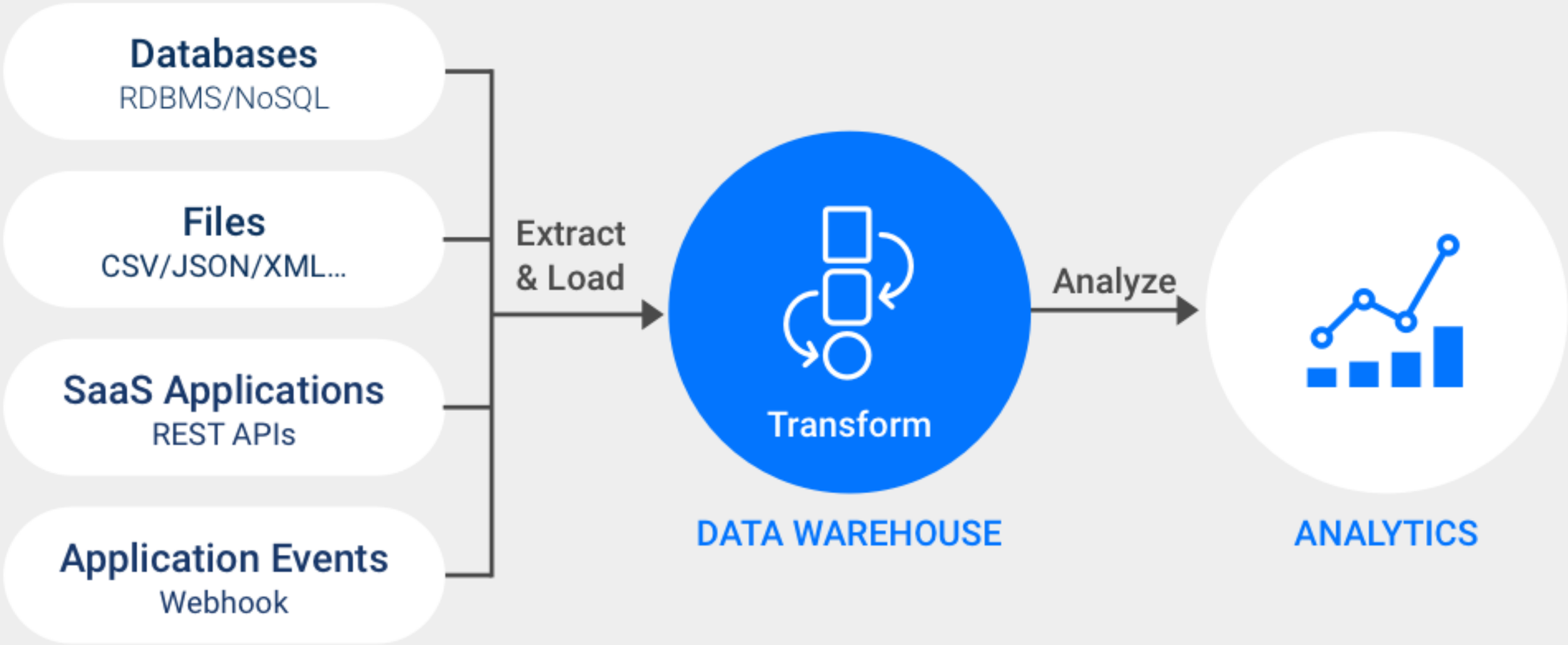
source: revery.io

The Solution



source: coursera

ELT PROCESS



source: revery.io

Question

- Region 정보를 어디서 어떻게 구할 것인가?
- Data, Code, config의 분리
- ETL code는 반복 사용이 가능해야
- parallel processing이 가능한가?