# Data Loading from AWS(S3)

Data Scientist
안건이

# 목차

- Python 가상 환경 Setting
- AWS(S3)
- Multi-processing
- 실습 진행

# Python

- Python
  전세계 가장 많이 쓰는 언어 Top 3위에 랭크

# 가상환경

- 가상환경 Setting

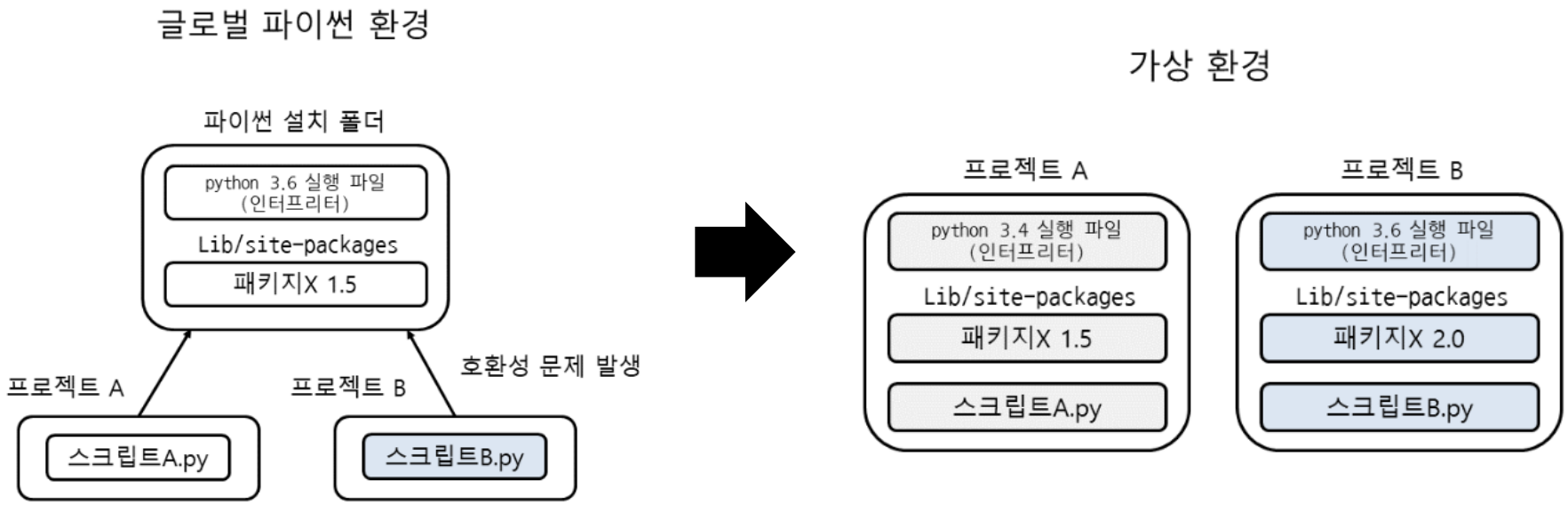
- **가상 환경 확인**
  **[root@ vm envs]# conda info --envs**
- **가상 환경 삭제**
  **[root@ vm envs]# conda remove --name gonie --all**

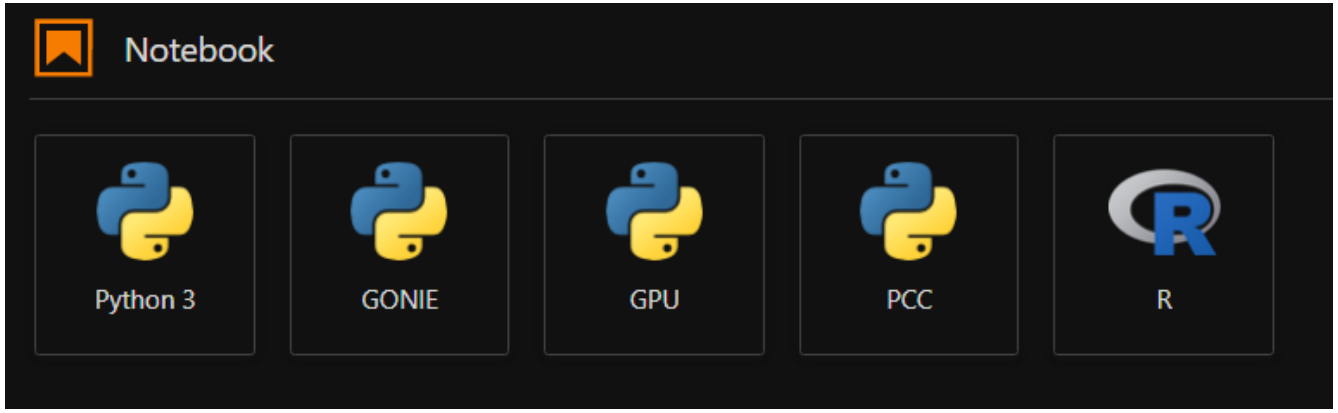
- **가상 환경 생성**
  - conda create -n gonie python=3.7 anaconda (conda create -n gonie python=3.7 pip), pip는 기본 Package 설치 안되어 있음...
  - conda activate gonie
  - conda install ipykernel
  - python -m ipykernel install --user --name gonie --display-name gonie (kernel 추가) // jupyter kernelspec uninstall gonie (kernel 삭제)

- **gonie만 원하는 이름으로 바꾸기**

# 가상환경

- 분석환경(anaconda)내 가상환경 관리는 무조건 필수
  - 100번 말씀 드려도 아무도 잘 안 하심
  - 진짜 나중에 눈물 흘릴 수가 있음 → 명심!!!

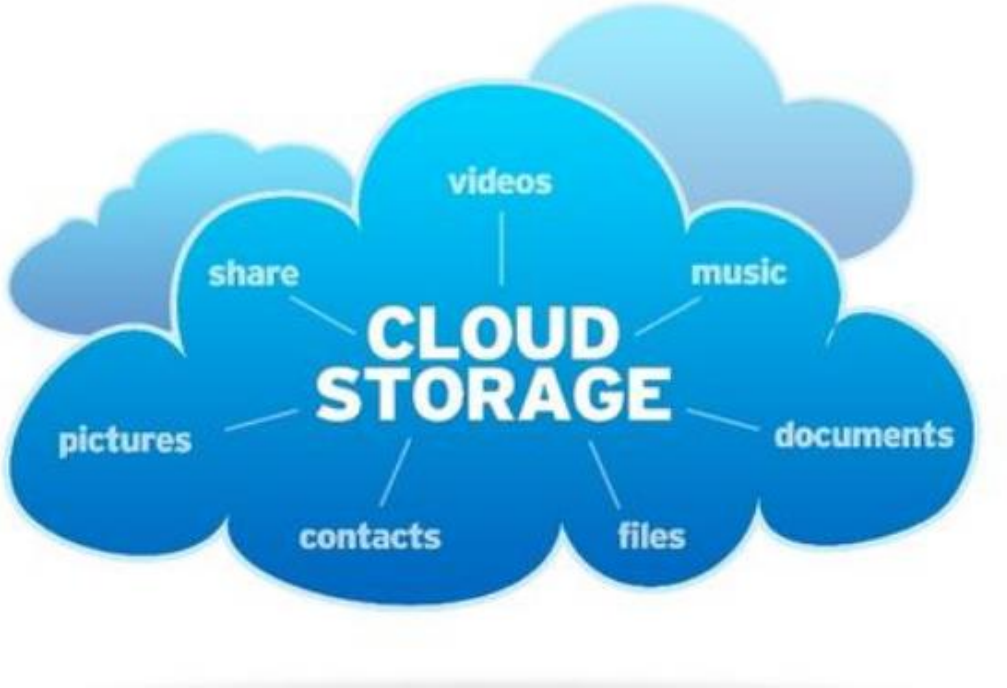


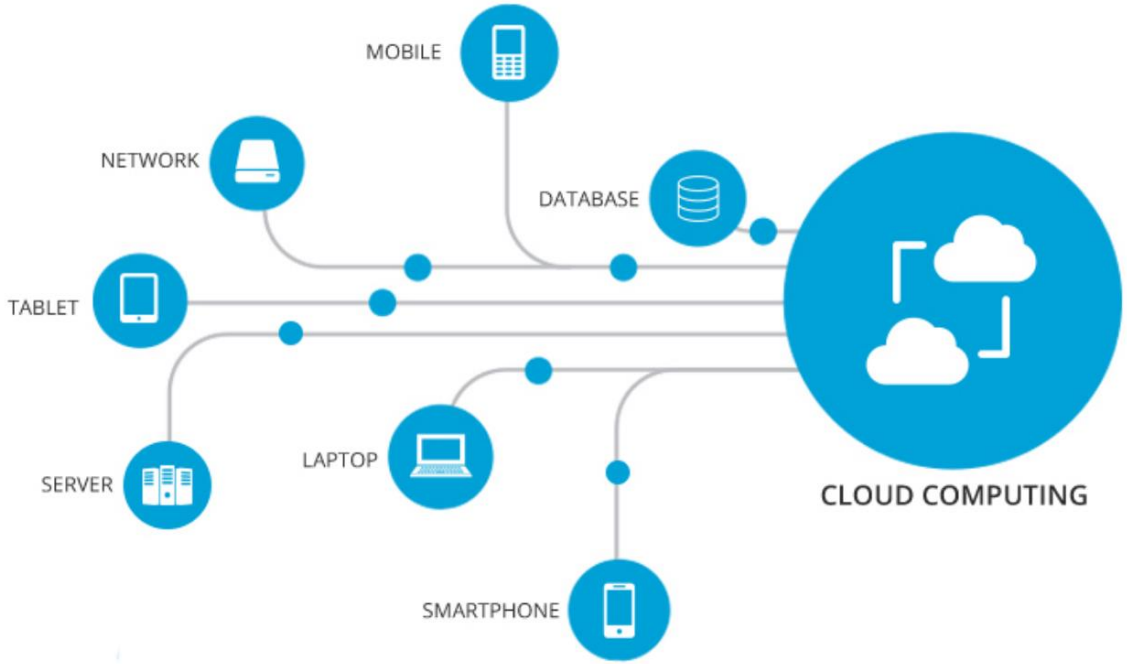
# Cloud Data – AWS (S3)

- S3: AWS에서 제공하는 Cloud 저장소
  - NAVER Cloud, GOOGLE Cloud Platform과 같은 개념

- 언제 어디서든 전세계 데이터를 받아 볼 수 있음
  - 5G 상용화 → 6G가 된다면?
  - 데이터는 더더욱 중요해지고 다양한 분석들이 일어날 수 있음




Just Cloud



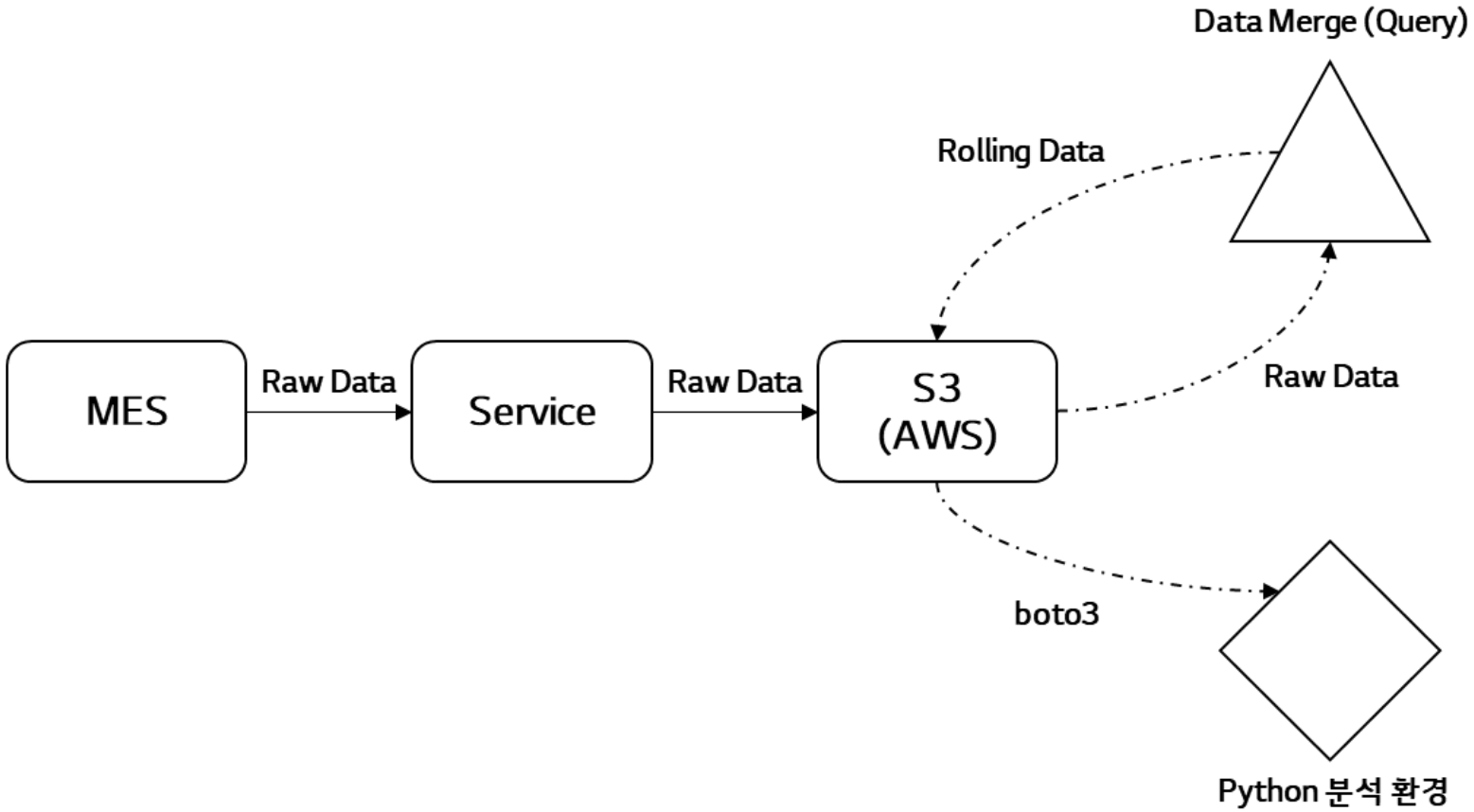

원격 제어

# Cloud Data – AWS (S3)

- S3: AWS에서 제공하는 Cloud 저장소
  - NAVER Cloud, GOOGLE Cloud Platform과 같은 개념

- boto3 : S3에서 Python 분석 환경으로 데이터를 바로 Loading 할 수 있는 Package

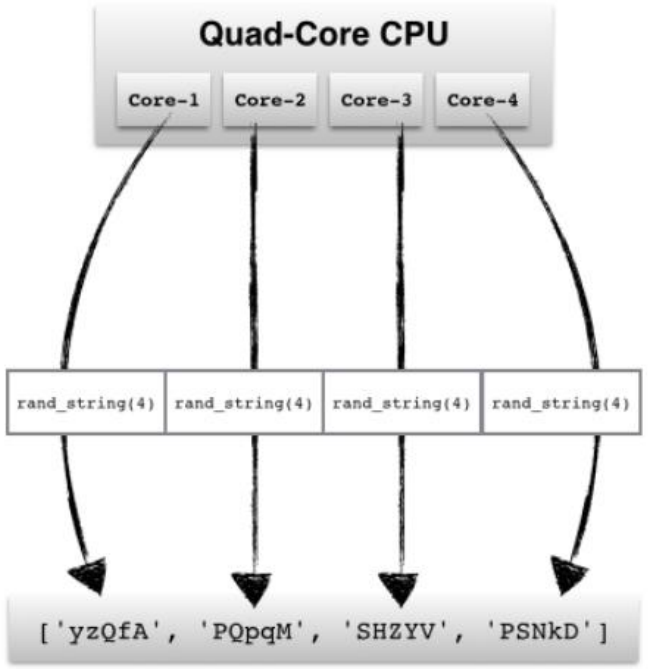
- 대부분의 데이터는 Table로 쪼개져 있음 → Key 값을 통해 분석할 수 있는 데이터로 Rolling이 필요함



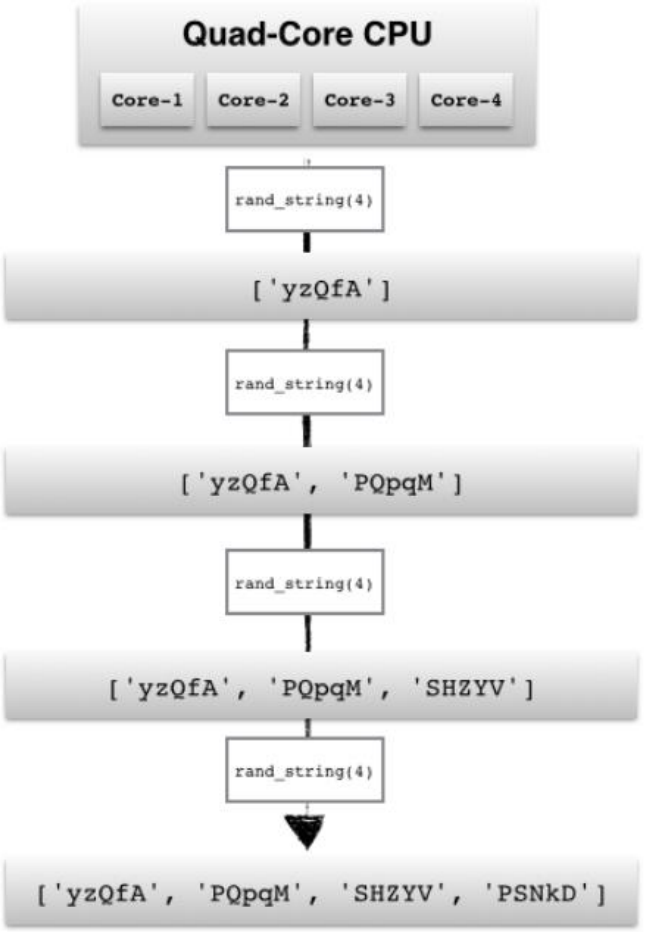
# Multi-processing

- Multi-Processing
  - 항상 적용할 수 있는가? → No !
  - 알고리즘 단에서는 Package 자체에 Multi-Processing이 거의 탑재되어 있는 경우가 많음(n_jobs)

[parallel processing]

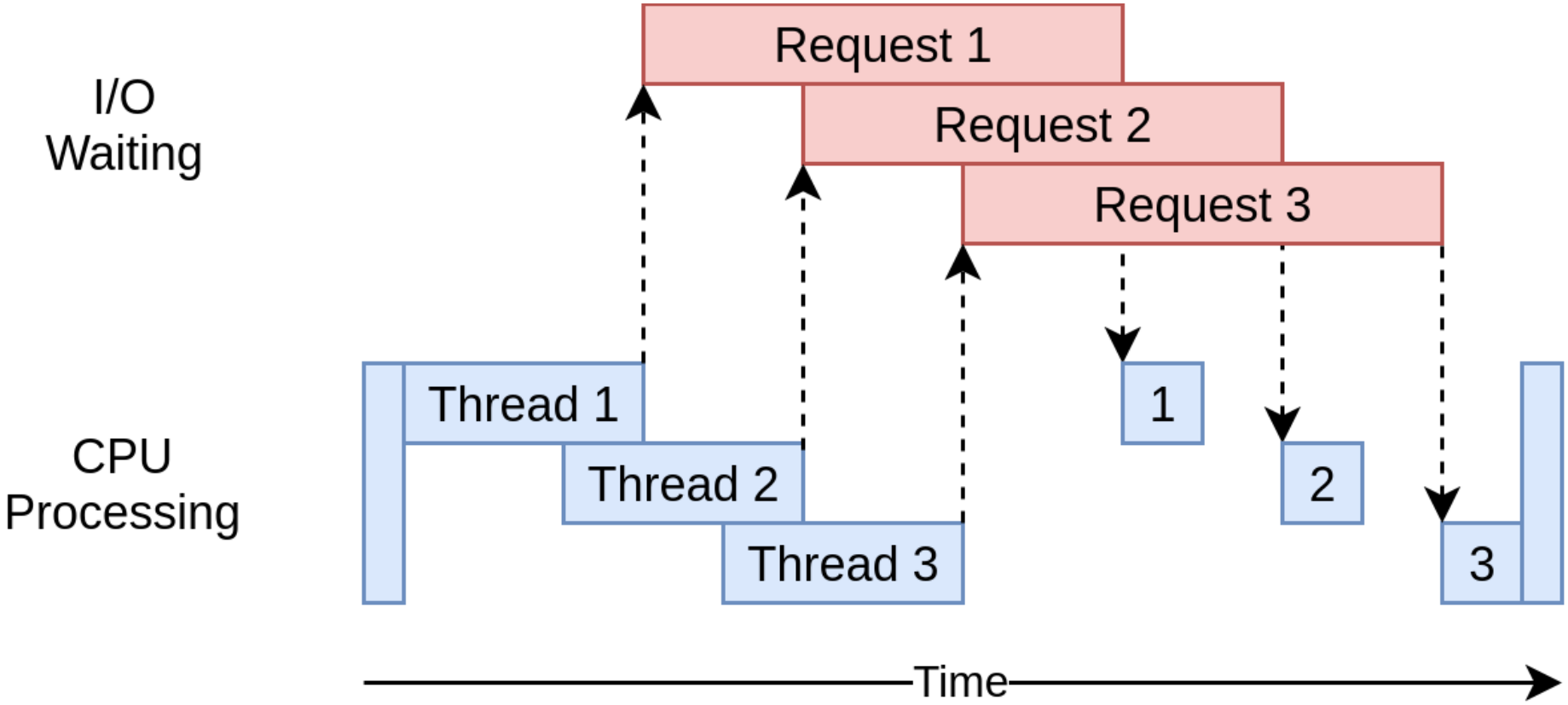



[serial processing]



# Multi-processing

- Multi-Processing
  - 항상 적용할 수 있는가? → No !
  - 알고리즘 단에서는 Package 자체에 Multi-Processing이 거의 탑재되어 있는 경우가 많음(n_jobs)
  - Core가 72개 있다고 72개의 Multi-Processing을 하면 안됨!
    - I/O문제가 발생할 수 있음
    - 적당한 실험을 통해 지정해야함 → 하나의 Core가 처리하는데 걸리는 시간이 어느정도 보장되어야 효과가 있음



# Q & A