빅데이터 처리

-Big Data Processing

1주차

오늘 수업 순서

- **실습실 안전 교육**
- **교과목 개요**
- **빅데이터란?**
  - 정의
  - 분야
  - 플랫폼/에코 시스템
- **취업 분야**
- **데이터 분석 공모전**
- **빅데이터 분석 예**
- **프로젝트 계획 수립**

실습실 안전 교육

Microsoft
erPoint 프레젠테

교과목 개요

- **목표**
  - 빅데이터 분야의 개발자가 되기 위한 역량을 갖춘다. 빅데이터의 개념과 필요성을 이해하고 빅테이터의 **수집, 가공, 분석** 및 **시각화** 기법을 습득한다.
- **평가**
  - 중간 평가 : 서술형 시험, 30%,
  - 과제 : 25%,
  - 기말 평가 : 데이터 분석 프로젝트 (개인), 35% , github 포트폴리오
- **사용 언어, 라이브러리**
  - Python, Pandas, Matplotlib, Seaborn, Scikit_learn 등
- **실습 환경**
  - 구글 colab, Anaconda

교과목 개요

- **주교재**
  - 1) 파이썬 머신러닝 판다스 데이터 분석, 정보 문화사
- **부교재**
  - 2) 데이터 과학 기반의 파이썬 빅데이터 분석 , 한빛 아카데미
  - 3) 파이썬 데이터 과학 통계 학습 , 정보 문화사

교과목 개요

- **주 차 별 수업 내용**
  - 1주차 : 교과목 개요 – 교재 2
  - 2주차 : 데이터 수집 (크롤링 등) - 교재 2
  - 3 -4 주차 : 데이터 가공, 정제 (Pandas) - 교재 1
  - 5주차 : 데이터 시각화 (Matplotlib, Seaborn) - 교재 1
  - 6-7 주차 : 데이터 분석 알고리즘 (Scikit-learn) - 교재 3
  - 8 주차 : 중간 평가
  - 9주차: 데이터 분석 (Scikit-learn) - 교재 3
  - 10주차 : 텍스트 마이닝 - 교재 2
  - 11주차 : 지리 정보 분석 - 교재 2

교과목 개요

- **주 차 별 수업 내용**
  - 12주차 : 시계열 데이터 분석
  - 13-14주차 : 데이터 분석 예제 실습 - 교재 2
  - 15 주차 : 프로젝트 결과 공유

교과목 개요

- **LINC 3.0**

교과목 개요

▪ 보강 계획

- A반
  - 9월 12일 (월) -> 9월 16일(금) 16:20~ , 동영상 강의 (과제)
  - 10월 3일 (월) -> 10월 8일(금) 16:20~ , 동영상 강의 (과제)
  - 10월 10일(월)-> 보강 주
- B반
  - 9월 12일 (월) -> 9월 14일(수) 16:20~ , 동영상 강의 (과제)
  - 10월 3일 (월) -> 10월 5일(수) 16:20~ , 동영상 강의 (과제)
  - 10월 10일(월)-> 보강 주

빅데이터

- **빅데이터의 정의**
  - **디지털** 환경에서 발생하는 **대량의 모든 데이터**
  - 대규모의 데이터를 **저장·관리·분석**할 수 있는 하드웨어 및 소프트웨어 기술, 데이터를 유통·활용하는 모든 프로세스를 포함
  - 빅데이터 플랫폼을 구성하는 하드웨어, 소프트웨어, 애플리케이션 간의 유기적 순환에 의해 **가치**를 창출

출처: 데이터 과학 기반의 파이썬 빅데이터 분석 , 한빛 아카데미

빅데이터

▪ 빅데이터의 출현

- 기술의 발달과 비용 저하, 소셜 네트워크 서비스 발달, 그림자 정보와 사물 정보 증가 등의 ICT 패러다임의 변화
- 빅데이터에 전문 역량과 기술을 더하여 전략적으로 활용할 방법이 주목됨
- 경제적 가치 창출, 사회 문제 해결, 새로운 ICT 패러다임 견인이라는 신가치 창출

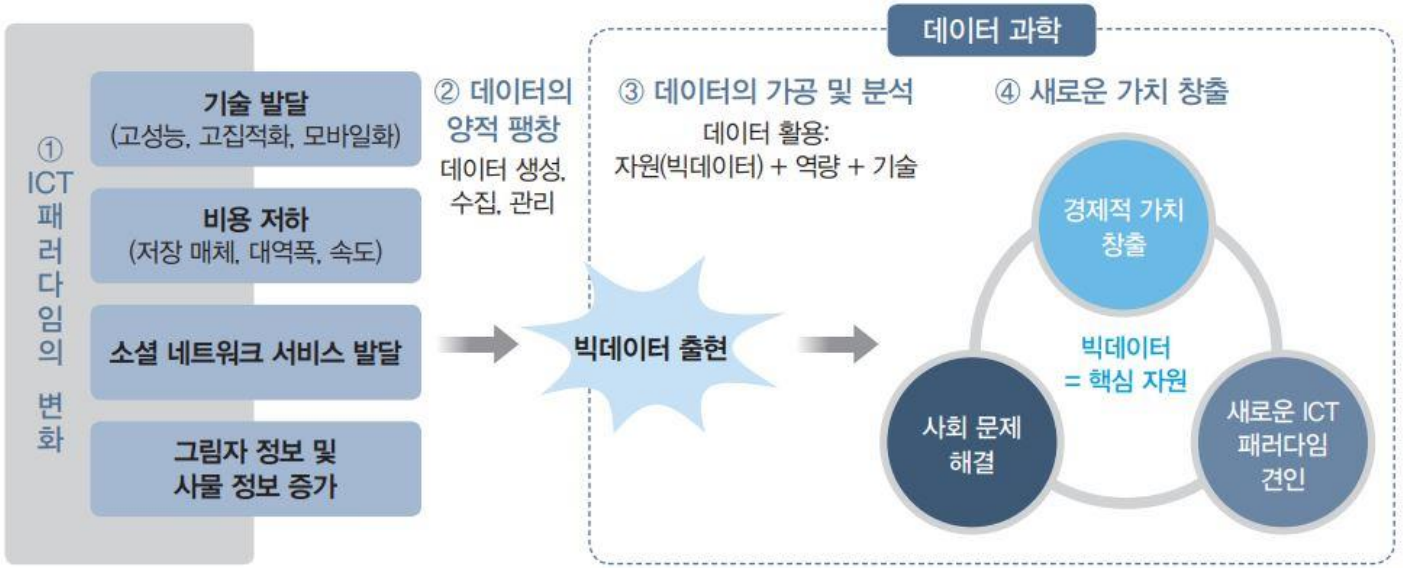


그림 2-2 빅데이터의 출현과 신가치 창출의 흐름

출처: 데이터 과학 기반의 파이썬 빅데이터 분석 , 한빛 아카데미

▪ 빅데이터의 특징 -데이터

표 2-3 빅데이터의 특징 – 데이터 측면

구분	특징	설명
1차 특징	규모	• ICT 기술의 발전으로 디지털 정보량이 기하급수적으로 폭증하여 제타바이트 시대로 진입
	다양성	• 데이터의 종류 증가: 로그 기록, 소셜/위치/소비/현실 데이터 등 • 데이터의 유형 다양화: 텍스트 외에 멀티미디어 등의 비정형 데이터 증가
	속도	• 센서, 모니터링 등의 사물 정보와 스트리밍 등의 실시간 정보가 증가하면서 데이터의 생성 및 이동(유통) 속도 증가 • 대규모 데이터를 처리하고 가치 있는 정보를 활용하기 위한 데이터 처리 및 분석 속도 증가
추가 특징	정확성	• 방대한 데이터를 기반으로 분석을 수행하므로 정확성 향상
	가치	• 빅데이터 분석으로 도출된 최종 결과물이 문제 해결을 위한 통찰력을 제공하므로 새로운 가치 창출 가능

출처: 데이터 과학 기반의 파이썬 빅데이터 분석 , 한빛 아카데미

빅데이터

▪ 빅데이터의 특징 –처리

- 빅데이터는 기존 데이터베이스 관리 시스템(DBMS)으로 처리하던 것에 비해 100배 이상 많은 정형, 비정형 데이터를 처리

표 2-5 빅데이터의 특징 – 처리 방식 측면

구분	이전의 데이터 처리 방식	빅데이터 처리 방식
데이터 트래픽	• 테라바이트 수준	• 페타바이트 수준: 최소 100테라바이트 이상 • 정보의 장기간 수집 및 분석 • 방대한 처리량
데이터 유형	• 정형 데이터 중심	• 비정형 데이터 비중이 높음: SNS 데이터, 로그 파일, 클릭스트림 데이터, 콜센터 로그 통신, CDR 로그 등 • 처리 복잡성 증대
프로세스 및 기술	• 단순한 프로세스 및 기술 • 정형화된 처리 및 분석 결과 • 원인 및 결과 규명 중심	• 다양한 데이터 소스와 복잡한 로직 처리 • 처리 복잡도가 높아 분산 처리 기술 필요 • 새롭고 다양한 처리 방법 필요: 정의된 데이터 모델/상관관계/절차 등이 없음 • 상관관계 규명 중심 • 하둡, NoSQL 등 개방형 소프트웨어 사용

출처: 데이터 과학 기반의 파이썬 빅데이터 분석 , 한빛 아카데미

빅데이터

▪ 빅데이터의 가치

- 빅데이터 분석이 제공하는 스마트 서비스는 기존 비즈니스에 **효율**화, **개인**화, 그리고 **미래 예측력을** 통한 혁신을 제공
- 단순히 새로운 기술이나 비즈니스 모델이 아니라 새로운 패러다임으로의 변화를 의미
- 빅데이터 자체부터 이를 활용한 사용자 애플리케이션까지 광범위하여 빅데이터 플랫폼과 에코시스템으로 확장

표 2-6 빅데이터 분석을 통한 비즈니스 혁신 방향

방향	설명
효율화	• 빅데이터를 이용해 과거 및 현재의 현상을 파악할 수 있다. • 물류, 재무, 기획, 마케팅 등 경영 전반의 데이터를 실시간으로 분석한 후 최선의 의사결정을 할 수 있다.
개인화	• 온라인 이용자의 활동 정보와 SNS 등으로 축적된 개인 정보를 결합하여 사용자에게 특화된 서비스를 제공할 수 있다. • 현재 개인 정보는 광고 분야에 활용 중인데 이를 넘어 의료, 교육 등 모든 분야로 확대가 가능하다.
미래 예측력	• 과거 및 실시간 데이터를 분석하여 축적한 개인 정보로 개인 또는 조직 전체의 행동 및 의사결정 패턴을 도출할 수 있다. • 미래에 적용 가능한 시나리오를 제시하고 예측 가능한 행동 및 발생 가능한 문제점을 사전에 방지하는 서비스가 가능하다.

출처: 데이터 과학 기반의 파이썬 빅데이터 분석 , 한빛 아카데미

▪ 빅데이터의 역할

- 미래 사회의 특성은 불확실성, 리스크, 스마트, 융합으로 대변됨
- 빅데이터를 활용해 여러 가지 가능성에 대한 시나리오 시뮬레이션을 하면 불확실한 상황 변화에 유연하게 대처 가능
- 빅데이터에 기반한 정보 패턴 분석으로 리스크에 대응할 수 있음
- 개인화 및 지능화된 서비스를 제공하여 삶의 질을 향상시킴

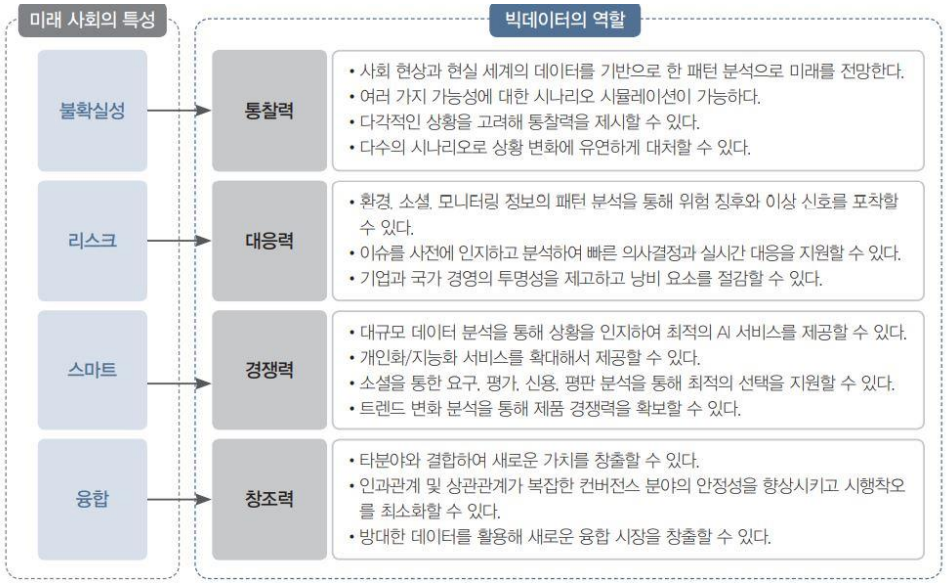


그림 2-3 미래 사회의 특성과 대응되는 빅데이터의 역할

출처: 데이터 과학 기반의 파이썬 빅데이터 분석 , 한빛 아카데미

빅데이터

- **빅데이터 처리 단계**

표 2-10 빅데이터의 처리 단계별 기술 영역

단계	기술 영역	내용
데이터 소스	내부 데이터	데이터베이스, 파일 관리 시스템
	외부 데이터	파일, 멀티미디어, 스트리밍
수집	크롤링crawling	검색 엔진 로봇을 이용한 데이터 수집
	ETL: 추출Extraction, 변환Transformation, 적재Loading	소스 데이터의 추출, 전송, 변환, 적재
저장	데이터 관리: NoSQL	비정형 데이터 관리
	저장소	빅데이터 저장
	서버	초경량 서버
처리	맵리듀스mapReduce	데이터 추출
	작업 처리	다중 작업 처리
분석	신경 언어 프로그래밍NLP, Neuro Linguistic Programming	자연어 처리
	머신러닝	데이터 패턴 발견
	직렬화serialization	데이터 간 순서화
표현	시각화visualization	데이터를 도표나 그래픽으로 표현
	획득acquisition	데이터의 획득 및 재해석

출처: 데이터 과학 기반의 파이썬 빅데이터 분석 , 한빛 아카데미

빅데이터

▪ 데이터를 근거로한 의사 결정

- KPI (Key Performance Indicator)

KPI란 프로젝트의 현황을 파악하기 위한 숫자로 업계마다 중요한 지표.

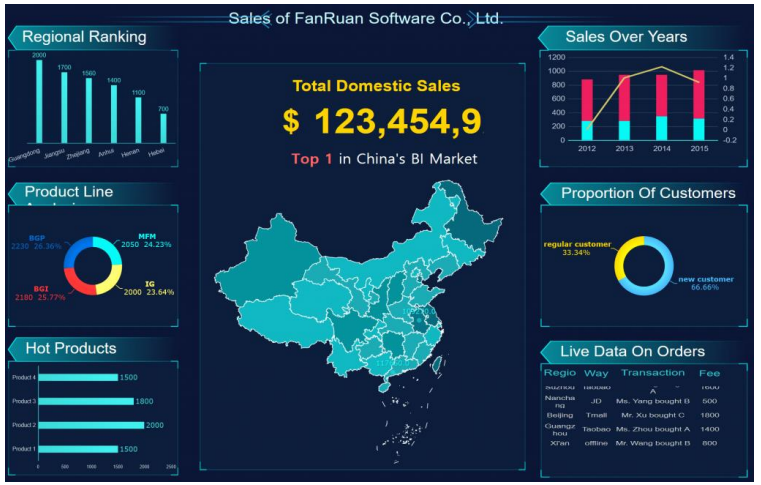
예를 들어, 서비스를 이용한 1일 또는 한달 유저수인 DAU(Dailly Activity User) ,MAU(Monthly Activity User) 등이 있다.

결과에 따라 자신의 다음행동이 결정될지 여부를 알고자 사용한다.

목표와 결과가 다르면 행동을 해야한다. 자신의 행동을 결정할때 직감이 아니라 객관적인 데이터에 근거하여 판단한다.

- KPI 예시

  : 광고 클릭 수, 동영상 조회 수, 쇼핑몰 매출



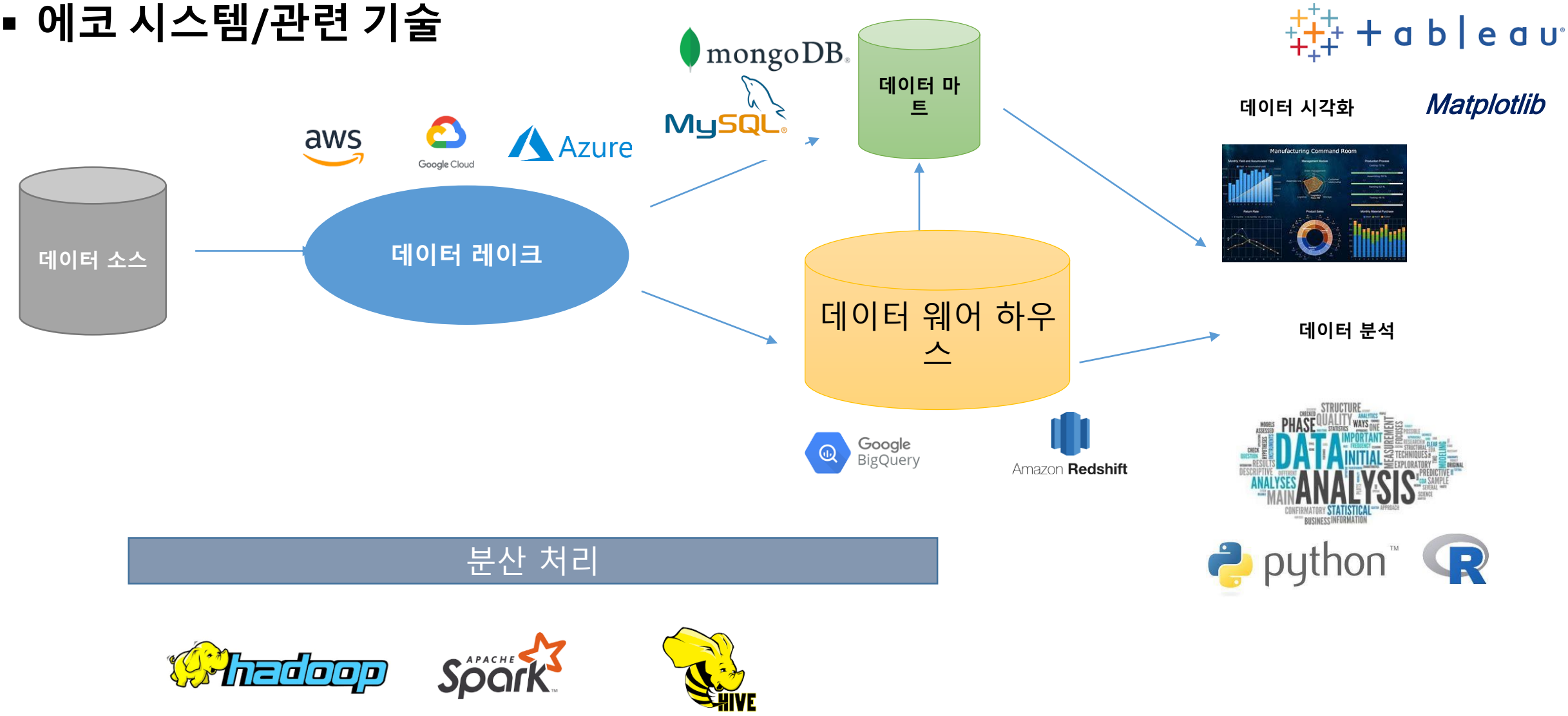
출처 : 빅데이터를 지탱하는 기술

빅데이터

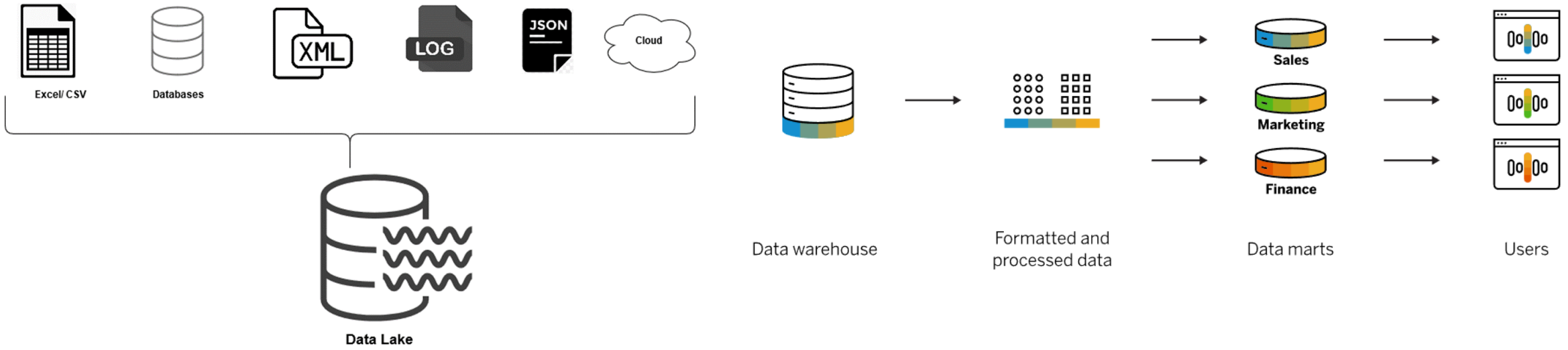
데이터 특징	**대량**의 데이터 **다양**한 데이터
필요 기술	분산 처리 병렬 처리 비 정형 데이터 처리 (No SQL) 대용량 저장 분석
사용 되는 기술 기술	Hadoop: 다수의 컴퓨터에서 대량의 데이터 처리 Hive : sql을 hadoop에서 사용하기 위한 기술 Mongo DB: no SQL 지원 DB Python , scikit learn, R, Pandas

빅데이터

- 에코 시스템/관련 기술



빅데이터

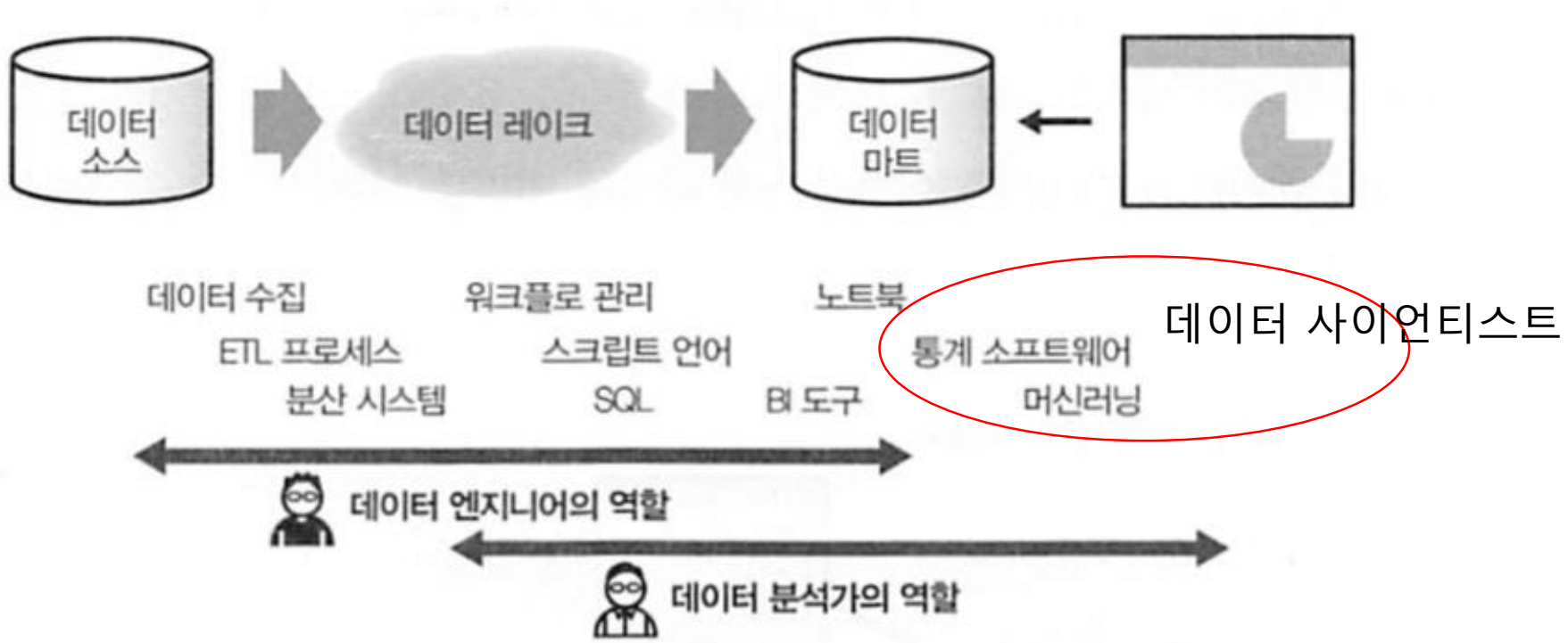


출처: 데이터 웨어하우스란? | 정의, 구성 요소, 아키텍처 | SAP Insights

데이터 레이크 : 다양한 형테의 데이터를 저장. 스토리지 개념
데이터 웨어 하우스 : 데이터 분석에 적합한 정규화된 데이터를 적재하고 관리. 대량의 데이터 장기 보존 .
DB 형태
데이터 마트 : 분석이 필요한 데이터나 프로젝트/업무 단위의 상세한 데이터 관리. 시각과 툴과 연동되어 사용.
DB 형태

직무 분야

-데이터 엔지니어
-데이터 분석가
-데이터 사이언티스트



데이터 사이언티스트

출처 : 빅데이터를 지탱하는 기술

채용 공고

AI – Python developer
수집하고 가공한 데이터를 관리합니다.
Data Pipeline 유지 보수

Python에 능숙하신 분 (또는 다른 언어에 능숙하고, Python을 다룰 줄 아시는 분)

[기술 우대사항]
Tensorflow, PyTorch, Keras, Pandas, Numpy를 사용해본 경험이 있으신 분
REST API 서버를 개발하고 운영해보신 분
Python 기반 시스템을 구현하여 패키지화를 해보신 분

데이터분석 (모집인원 : ○명)

담당업무

- 비즈니스 KPI 설계, 분석 및 보고서 작성
- 정형/비정형 데이터 분석을 통한 비즈니스 인사이트 도출
- 정형/비정형 데이터 기반 실험설계 및 분석
- 정형/비정형 데이터 전처리, 가공, 분석, 시각화, 예측 및 레포팅

[사용언어 및 기술]

SQL, Python, R, Java, pandas, scikit-learn, matplotlib, seaborn, Tensorflow, keras 등

채용 공고

Data Engineer

Python 또는 javascript를 활용해 AI 모델을 개발한 경험이 있으신 분
- 머신러닝 관련 지식을 가지고 계신 분
- 통계학, 컴퓨터공학 복수전공자
- dacon과 같은 AI 해커톤 플랫폼에 참여하거나 수상 경력이 있으신 분

채용 공고

인하공전 컴퓨터 정보 과

데이터 분석 및 사업 기획 모집 (신입 / 경력) AI 팀 0명

▶ 자격요건

* Big Data, 통계, 분석, 모델링, 시각화

• 데이터 분석을 위한 프로그래밍 역량 (Python, R, SQL 등)

• 데이터 크롤링, EDA(Exploratory Data Analysis), 전처리 업무 숙련자

• **Pandas 및 Numpy** 등 라이브러리 활용하여 데이터 **전처리** 및 가공 가능하신 분

• 텍스트마이닝을 위한 라이브러리 사용 가능자

▶ 우대사항

• 정부 또는 기관에서 진행하는 빅데이터 분석과정을 수료하거나 관련 **공모전에서 수상한 경험이** 있는 분

• **데이터분석 관련 논문** 또는 관련 교육 기관에서 진행한 **포트폴리오**가 있으신 분
• Pytorch, Keras, Tensorflow 등 라이브러리 활용 해 보신분

• 분석 툴 사용 및 보고서 기획/작성 경험이 있는 분
• 시각화 툴 사용 및 개발 경험이 있는 분

• 데이터 분석 관련 **자격증** 보유하신 분

채용 공고

데이터 분석가
(신입/경력 무관)

- 클라우드 서비스를 이용하는 다양한 고객사의 AI/ML 비지니스 지원
- 자사 전략 사업 분야에 대한 데이터 분석/모델링 및 솔루션화

[자격요건]

- 컴퓨터 공학 전공 혹은 수학, 통계학과 전공자
- 정형/비정형 데이터에 대한 기초 가공 작업 가능자
- Python/R 언어 중에 하나 이상 경험

[우대사항]

- 데이터 분석 관련 수행한 프로젝트의 포트폴리오 (PDF)
- AWS 클라우드 사용 경험

Data Specialist

모델링, 데이터 이행, ETL 개발 구축 경험 및 다양한 RDBMS 활용 기술 역량

Big Data 관련 AWS 클라우드 및 Hadoop Eco (Hadoop, Spark, Kafka, Hive, NoSQL 등) 기반 프로젝트 수행 역량

개발에 필요한 오픈 소스 (Java, Python 등) 사용 능력 및 개발 역량

채용 공고

데이터 엔지니어 모집

SQL 문법의 이해 및 사용 경험이 있으신 분

AWS Aurora 혹은 Athena 사용 경험이 있으신 분
Apache Hive, Google BigQuery 사용 경험이 있으신 분
Python 혹은 R스크립트에 대한 이해가 있으신 분
DBMS에 대한 탁월한 이해가 있으신 분
통계, 데이터 분석 능력이 있으신 분

데이터 엔지니어

- 1년 이상 ETL 파이프라인 설계 및 개발 실무 경험이 있으신 분
- Python, SQL, Spark 등을 능숙하게 사용 가능하신 분
- 데이터 모델링, Data Lake 구축 및 운영 경험이 있으신 분
- 분산처리 프레임워크를 통한 실시간/비실시간 데이터 처리 경험이 있으신 분
- Kafka / Kinesis / Hadoop / Hive / Athena / Presto 등을 활용한 경험이 있으신 분

채용 공고

데이터 엔지니어

•데이터 레이크 및 데이터 파이프라인 시스템 설계, 개발 및 운영
•데이터 분석 환경 제공을 위한 도구 적용 및 관련 인프라 설계, 개발 및 운영

자격 조건

•데이터 가공 및 분산 처리 기술에 대한 이해
•알고리즘, 데이터구조, OS, 데이터베이스 등 기본적인 전산 지식에 대한 이해
•기본적인 SQL 및 프로그래밍 능력

우대사항

•Hadoop MR, Hive, Spark 등 분산 처리 기술 관련 개발 경험
•대용량/실시간 데이터 분산 처리 시스템 설계 및 운영 경험
•AWS, GCP등 클라우드 환경에서의 데이터 파이프라인 구축 및 운영 경험
•GA, Tableau 등 외부 분석/시각화 도구에 대한 지식
•Python, Java, Scala 중 하나 이상을 이용한 개발 경험

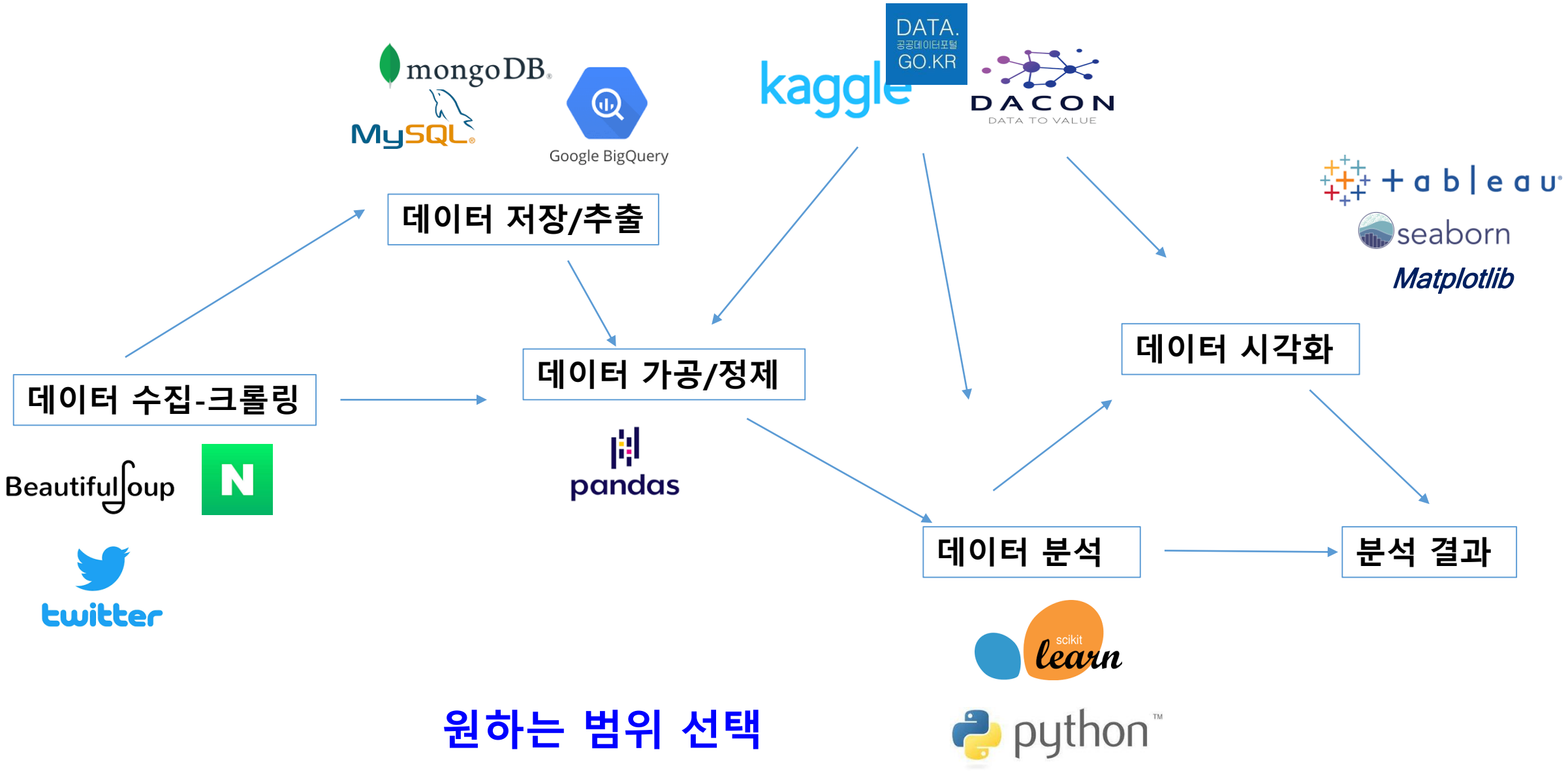
공모전

kaggle

DACON
DATA TO VALUE

공공 빅데이터 분석 공모전

프로젝트 범위



원하는 범위 선택

프로젝트 상세 내용

- **단계별 산출 물**
  - 4주차
    - 프로젝트 계획
      - » 사용 데이터, 사용 기술, 프로젝트 범위
  - 10주차
    - 중간 결과
  - 15주차
    - 최종 결과. Git hub 포트폴리오 형식 (보고서)

데이터 분석 예제

- 프로야구 정규시즌 팀 승률, 타율 및 방어율 예측
- KBO 정규시즌 팀별 승률, 타율 및 방어율(평균자책점) 예측
- 항공운항 데이터를 활용한 "항공편별 지연 여부 예측"
- 개봉(예정) 영화에 대한 관객 수 예측
- 게임 내 활동데이터를 활용한 '게임 유저 이탈 예측 모형' 개발
- 보험, 통신, 신용평가사(개인정보 비식별) 결합데이터를 활용한 대출상환 예측 알고리즘 개발
- 퇴근 시간 버스 승차 인원 예측
- 상점 신용카드 매출 예측

[Kaggle] Nexfilx movies and TV shows 데이터 분석 3 (EDA / 시각화 / Review) (tistory.com)

관련 자료

- 빅데이터 커리어 가이드북 ,길벗
- 빅데이터를 지탱 하는 기술, 제이펍
- 데이콘 경진 대회 1등 솔루션, 위키 북스
- 파이썬으로 캐글 뽀개기,비제이 퍼블릭

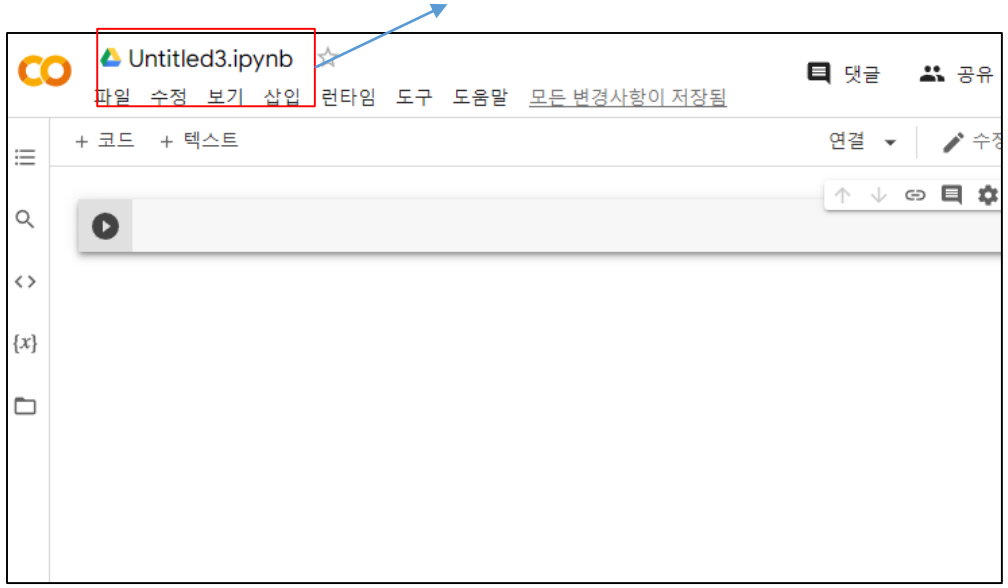
수고하셨습니다

jhmin@inhatc.ac.kr

구글 Colab 실습

https://colab.research.google.com/ 구글 계정이 없으면 계정 생성

파일->새노트



+코드 버튼 을 누른 후 다음 내용 입력

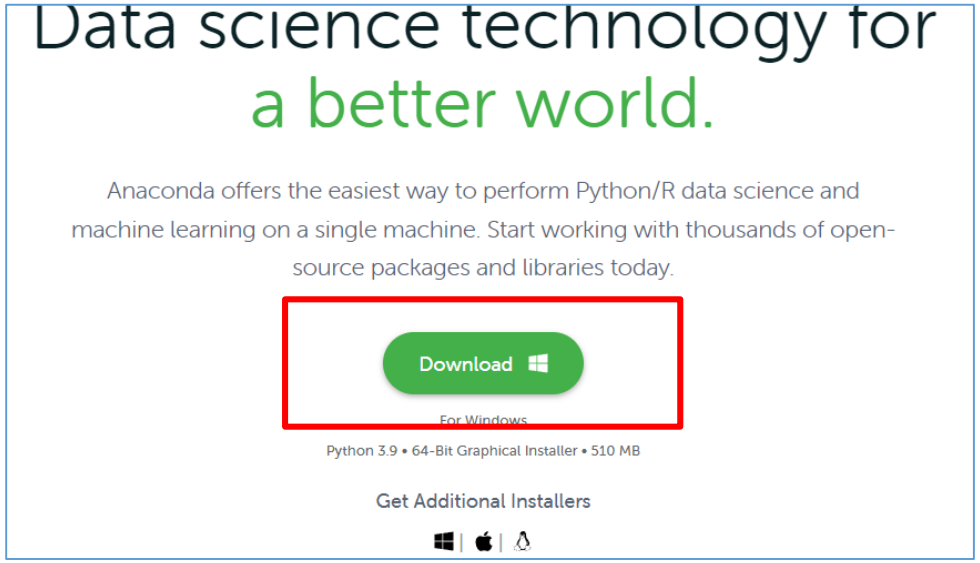
```
import tensorflow as tf
print(tf.__version__)
```

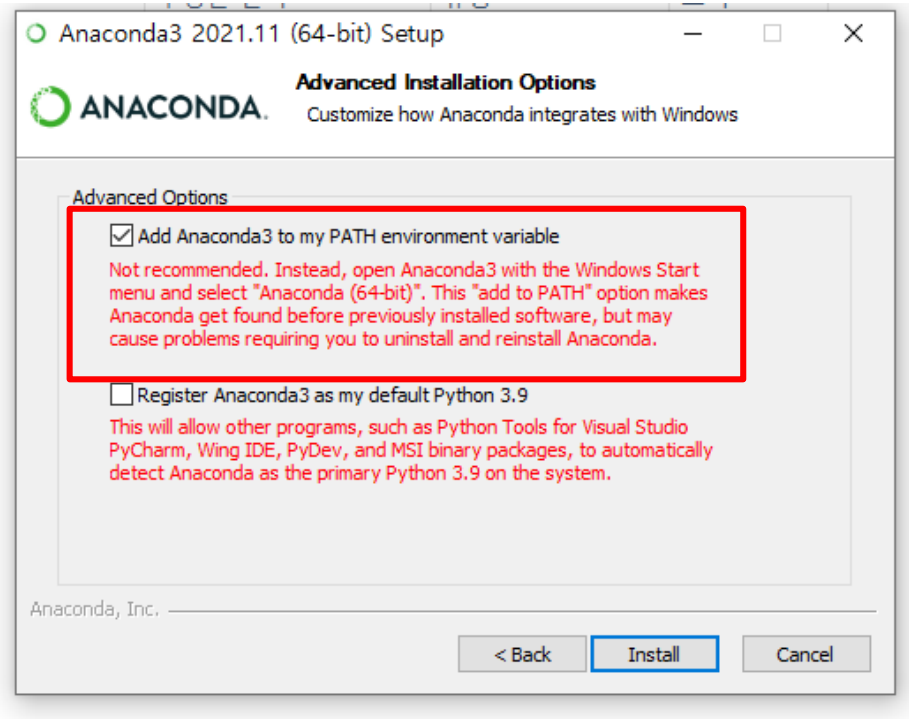
재생 버튼 클릭 후 결과 확인

아나콘다&주피터 설치 하기

인하공전 컴퓨터 정보 과

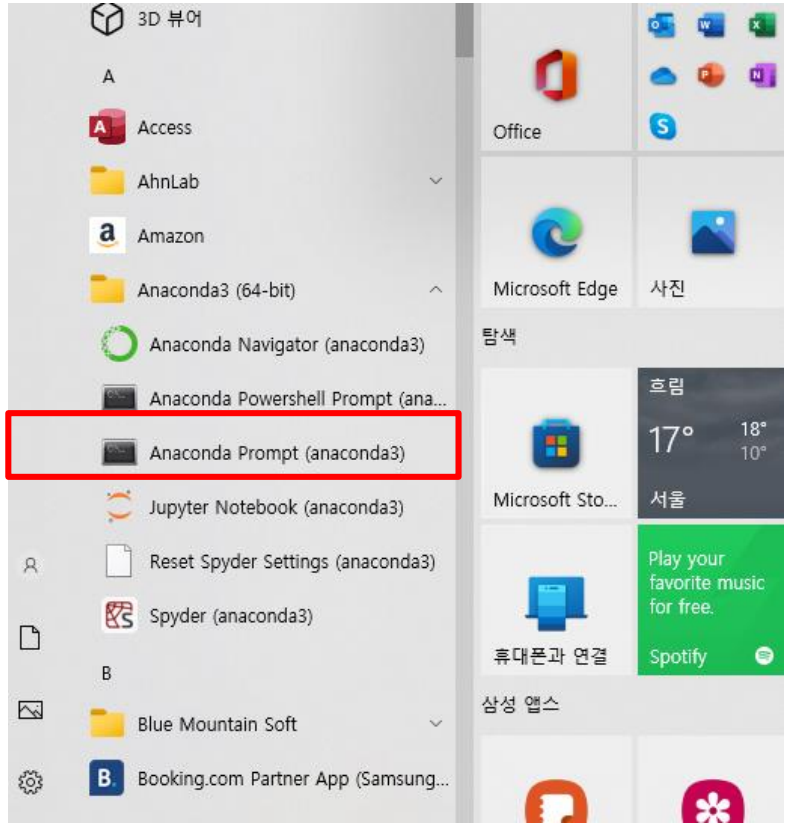
https://www.anaconda.com/





아나콘다&주피터 설치 하기

https://www.anaconda.com/




```
base>>conda create –n py37 python=3.7
base>>conda activate py37

py37>>pip install tensorflow==2.0.0
py37>>pip install keras==2.3
py37>>pip install scikit-learn
py37>>pip install matplitlib
py37>>pip install opencv-python
py37>>pip install pandas
py37>>pip install seaborn

py37>>conda install nb_conda
```

아나콘다&주피터 설치 하기

인하공전 컴퓨터 정보 과