데이터분석 w/ 파이썬

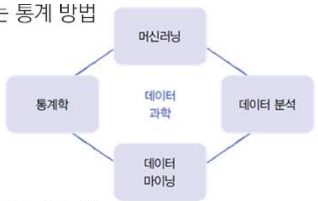
데이터분석 기본지식

한국폴리텍대학 성남캠퍼스 인공지능소프트웨어과
이혜정 교수

1

데이터 분석

- **올바른 의사 결정을 돕기 위한 통찰(insight)을 제공**
  - 기술통계 : 관측이나 실험을 통해 수집한 데이터를 정량화하거나 요약하는 기법
  - 탐색적 데이터 분석 : 데이터를 시각적으로 표현하여 주요 특징을 찾고 분석하는 방법
  - 가설검정 : 주어진 데이터를 기반으로 특정 가정이 합당한지 평가하는 통계 방법

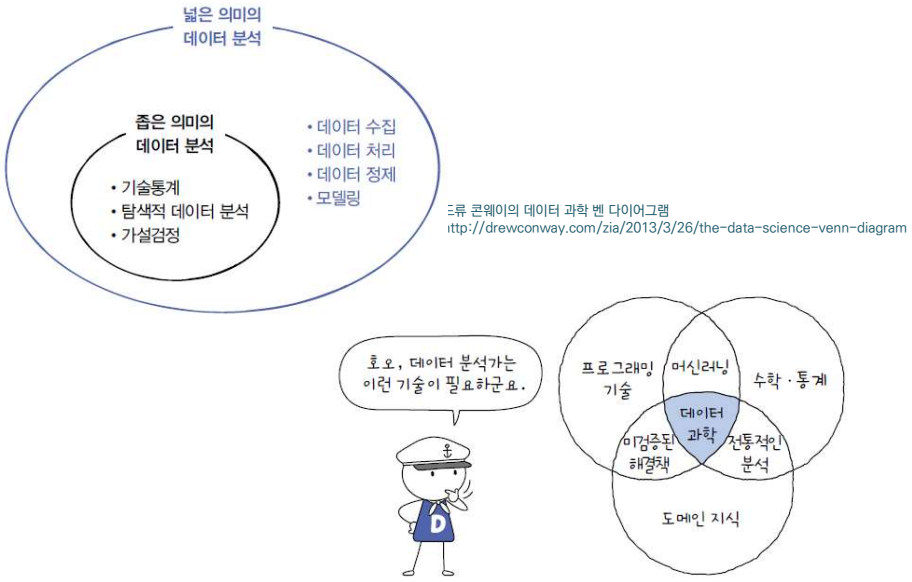


- **데이터 과학**
  - 한 걸음 더 나아가 문제 해결을 위한 최선의 솔루션(solution)을 만드는 데 초점
  - 데이터 과학은 통계학(statistics), 데이터 분석, 머신러닝(machine learning), 데이터 마이닝(data mining) 등을 아우르는 큰 개념

특징	데이터 분석	데이터 과학
범주	비교적 소규모	대규모
목표	의사 결정을 돕기 위한 통찰을 제공하는 일	문제 해결을 위해 최선의 솔루션을 만드는 일
주요 기술	컴퓨터 과학, 통계학, 시각화 등	컴퓨터 과학, 통계학, 머신러닝, 인공지능 등
빅데이터	사용	사용

2

데이터 분석가



3

Data Mining, Machine Learning

- **데이터 마이닝**
  - 데이터에서 패턴 혹은 지식을 추출하는 작업
    - 머신러닝, 통계학, 데이터 베이스 시스템(database system)과 관련이 많음
    - 패턴과 지식은 사람이 의사 결정을 내리기 위해 활용

- **머신러닝**
  - 데이터에서 자동으로 규칙을 학습하여 문제를 해결하는 소프트웨어를 만드는 기술
    - 딥러닝(deep learning)도 머신러닝 알고리즘의 한 종류
    - 패턴을 사용하는 주체가 사람이 아닌 컴퓨터
    - 모델(model) – 머신러닝으로 학습한 소프트웨어 객체

4

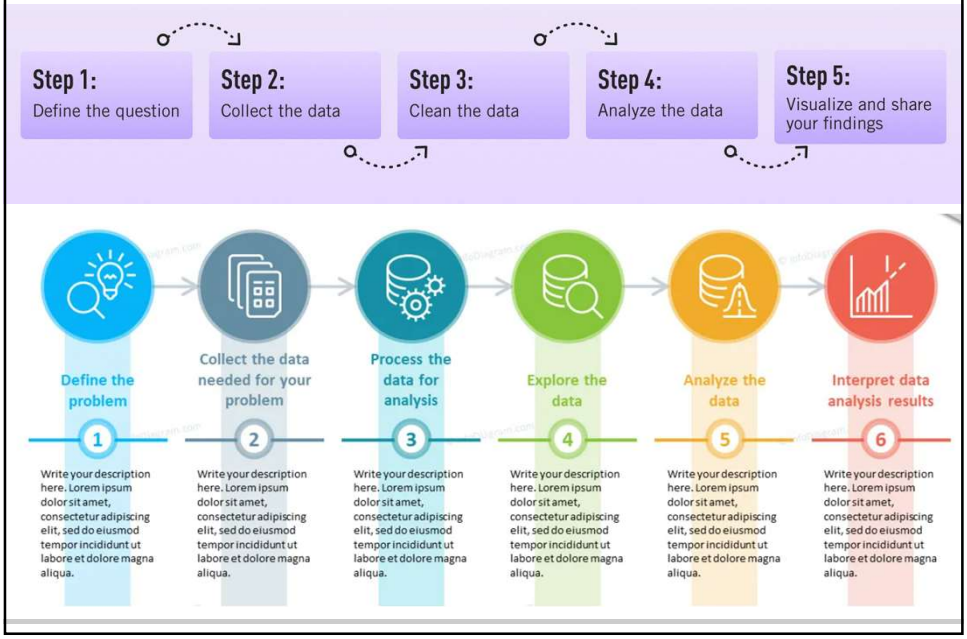
데이터 분석을 위한 도구

- **프로그래밍 언어: 파이썬과 R**
  - 파이썬 – 귀도 반 로섬Guido van Rossum이 1991년에 만든 범용 프로그래밍 언어
  - R – 1995년에 통계 계산을 위해 개발된 언어, 범용 프로그래밍 언어는 아님
- **데이터 분석을 위한 파이썬 패키지**
  - Numpy, Pandas, MatPlotLib
  - SciPy, SKLearn
- **프로그래밍 환경: 구글 코랩, 주피터 노트북**

- 데이터 분석은 데이터를 수집, 처리, 정제, 분석, 모델링하여 의사 결정 지원
- 데이터 과학은 데이터 분석, 머신러닝을 아울러서 문제해결 솔루션 도출
- 파이썬은 데이터 분석, 데이터 과학, 머신러닝에 사용할 수 있는 인기 프로그래밍 언어

5

데이터 분석 절차



6

데이터 엔지니어

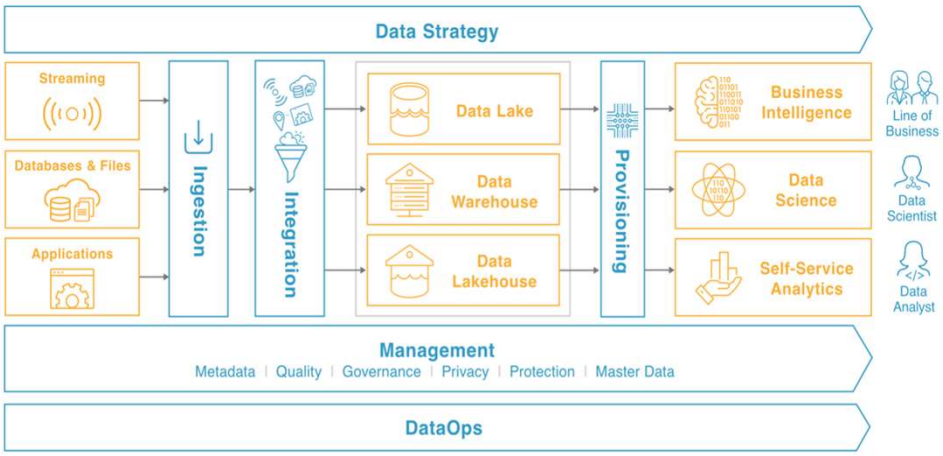
- **데이터 수집, 저장, 처리, 전송 등의 기술적인 측면에서 전문가**
  - 데이터 플랫폼과 데이터 파이프라인 아키텍처를 설계, 구축하고 데이터를 수집, 전처리하고, 저장소에 저장하는 등의 작업을 수행
  - 분산시스템, 클라우드 컴퓨터, 데이터베이스 등의 기술을 활용

- **주요 역할과 책임**
  - 데이터 아키텍처 설계, 데이터 파이프라인 구축 및 유지 보수
    - 아키텍처를 설계하여 데이터의 흐름과 처리를 구조화
    - 빅 데이터 기술과 도구를 활용하여 대용량 데이터를 처리하고 저장 (Hadoop, Spark, Kafka 등)
    - 구축한 데이터 파이프라인의 성능을 모니터링, 필요한 경우 성능을 최적화 또는 스케일 업/아웃
  - 데이터 처리 : 수집 및 추출 → 변환 및 전처리 → 저장, 품질 관리, 실시간 데이터 처리
  - 데이터 보안과 개인정보 보호, 규정 등을 고려하여 데이터를 처리하고 저장
  - 데이터 과학자, 비즈니스 분석가, 소프트웨어 개발자 등과 협업 → 데이터 파이프라인 개선 및 의미 있는 결과 도출

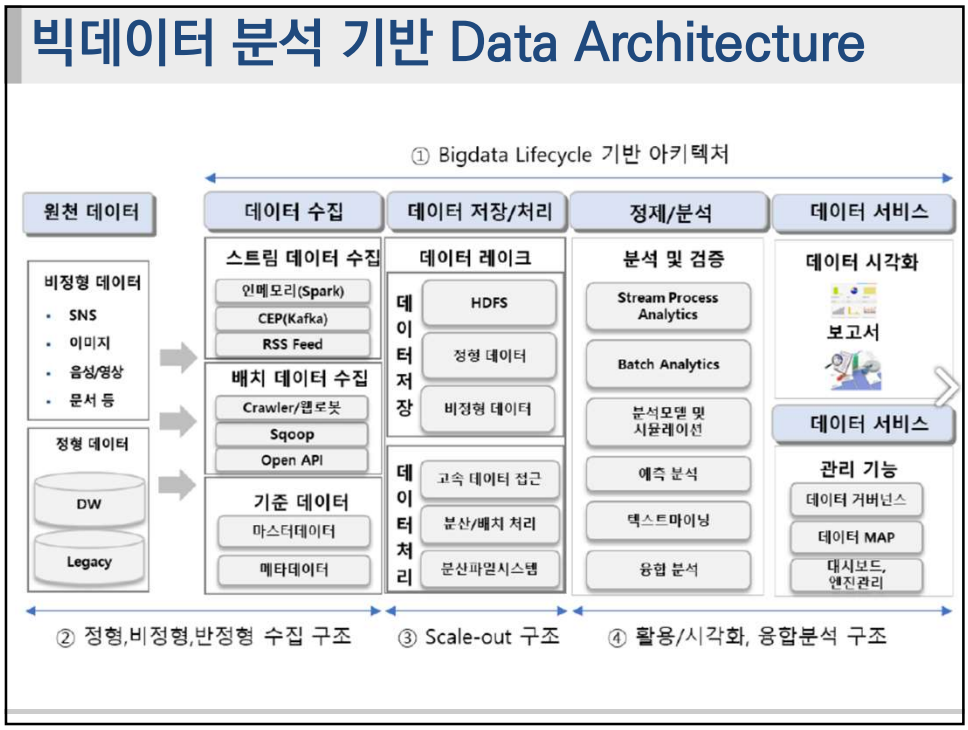
7

Data Pipeline

- **데이터 원본에서부터 데이터 사용까지의 전체 데이터 흐름을 처리하고 관리**



8


빅데이터 분석 기반 Data Architecture
① Bigdata Lifecycle 기반 아키텍처
원천 데이터
데이터 수집
데이터 저장/처리
정제/분석
데이터 서비스
비정형 데이터
SNS
이미지
음성/영상
문서 등
정형 데이터
DW
Legacy
스트림 데이터 수집
인메모리(Spark)
CEP(Kafka)
RSS Feed
배치 데이터 수집
Crawler/웹로봇
Sqoop
Open API
기준 데이터
마스터데이터
메타데이터
데이터 레이크
데이터 저장
HDFS
정형 데이터
비정형 데이터
데이터 처리
고속 데이터 접근
분산/배치 처리
분산파일시스템
분석 및 검증
Stream Process Analytics
Batch Analytics
분석모델 및 시뮬레이션
예측 분석
텍스트마이닝
융합 분석
데이터 시각화
보고서
데이터 서비스
관리 기능
데이터 거버넌스
데이터 MAP
대시보드, 엔진관리
② 정형,비정형,반정형 수집 구조
③ Scale-out 구조
④ 활용/시각화, 융합분석 구조

9