빅데이터 컴퓨팅 ::

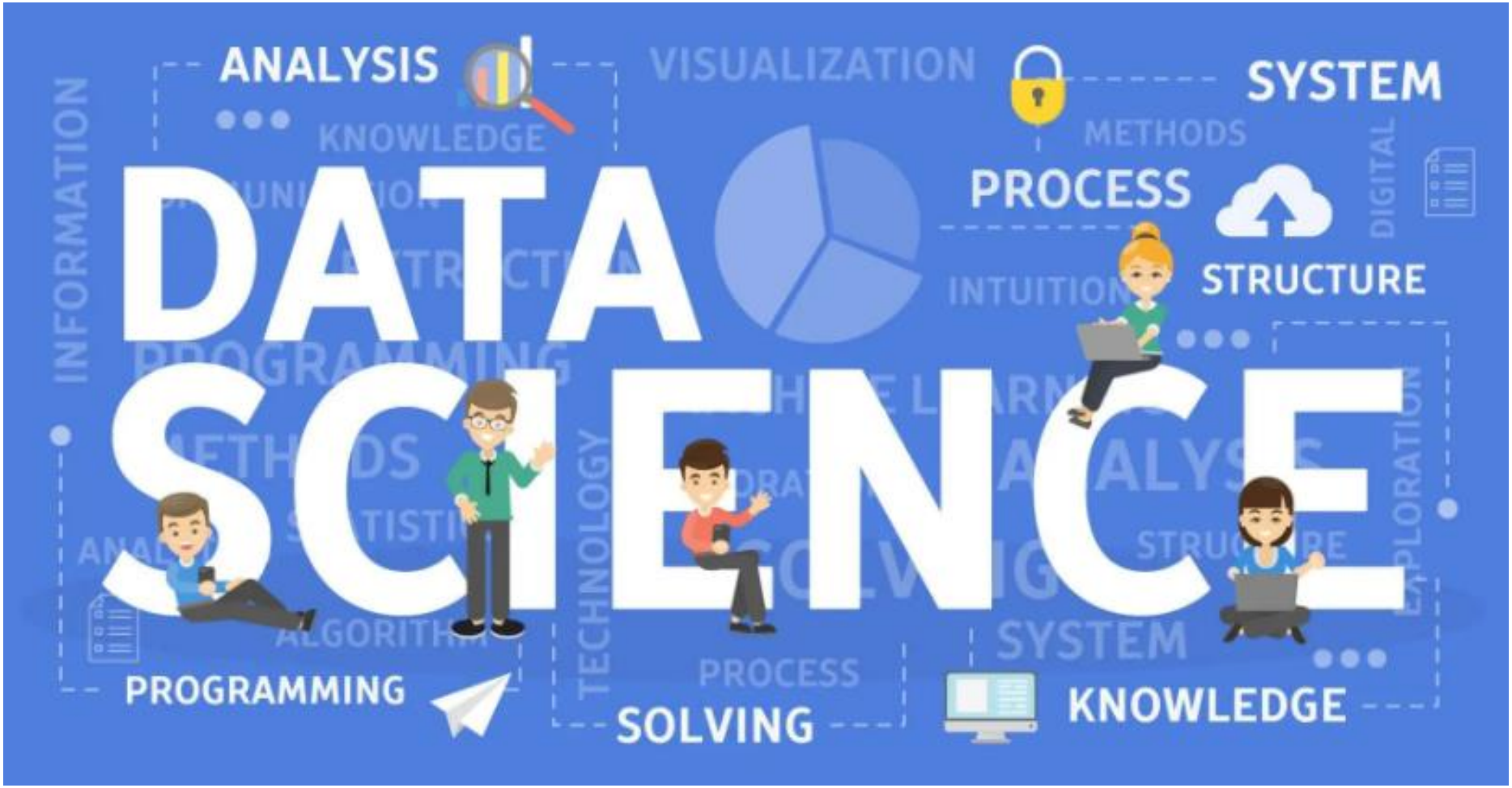
빅데이터 분석

서강대학교
SOGANG UNIVERSITY

데이터사이언스의 정의

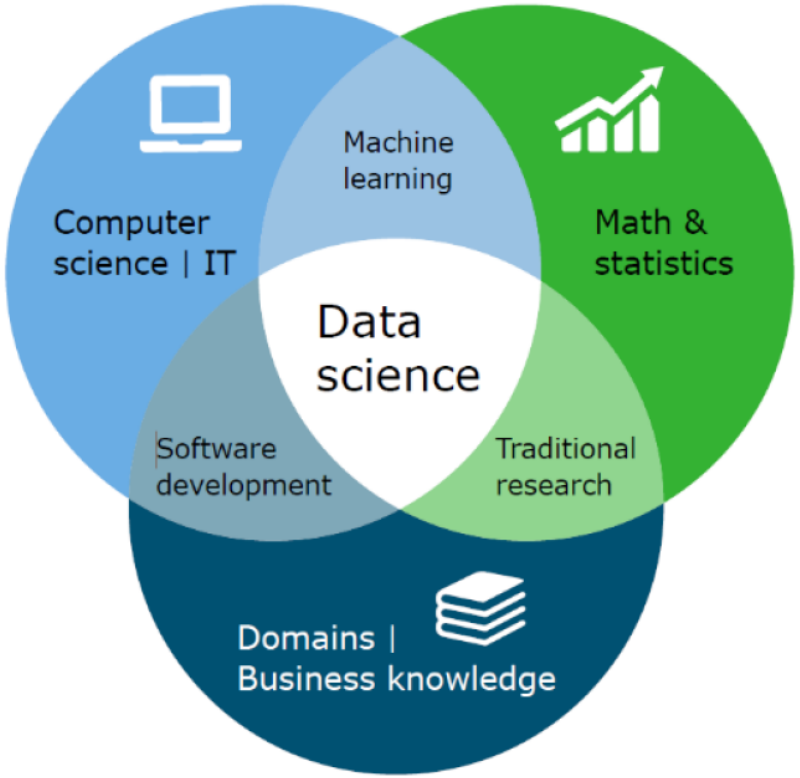
- 데이터사이언스의 기원
  - 데이터사이언스란 용어는 1974년 덴마크의 컴퓨터 과학자 페테르 나우르가 쓴 Concise Survey of Computer Methods에서 언급. 나우르는 데이터 과학을 데이터 수집, 저장, 조작 등 데이터를 컴퓨터로 처리하는 과학이라고 정의함
- 위키피디아
  - 데이터 과학이란 정형, 비정형 형태를 포함한 다양한 데이터로부터 지식과 인사이트를 추출하는데 과학적 방법론, 프로세스, 알고리즘, 시스템을 동원하는 융합분야
  - 데이터 과학은 데이터를 통해 실제 현상을 이해하고 분석하는데 통계학, 데이터 분석, 기계학습과 연관된 방법론을 통합하는 개념으로 정의

데이터사이언스의 정의


ANALYSIS
VISUALIZATION
SYSTEM
KNOWLEDGE
METHODS
PROCESS
STRUCTURE
INFORMATION
DIGITAL
INTUITION
DATA
SCIENCE
PROGRAMMING
TECHNOLOGY
EXPLORATION
ALGORITHM
PROCESS
SYSTEM
PROGRAMMING
SOLVING
KNOWLEDGE

데이터사이언스의 특징

- 데이터 사이언스는 다양한 성질의 내용이나 형식의 데이터들 간에 공통적으로 존재하는 성질과 그것들을 다루기 위한 기술
- 사용되는 기술은 컴퓨터, 수학, 경영 분야에 걸쳐 있음



ANATOMY OF A
DATA SCIENTIST

SALARY
Average salary of data scientists is **$120,000/year**

EDUCATION
- **88%** of all data scientists have at least a Master's degree
- **46%** of data scientists have a PhD

BENEFITS
- Harvard Business Review called data science the **"Sexiest Job of the 21st Century"**
- One of the fastest growing careers in the United States
- **94%** of data science graduates have found jobs since 2011

SKILLS
- Programming languages (R, Python, SQL, Hive, etc.)
- Statistics
- Multivariable calculus and linear algebra
- Machine learning
- Software engineering
- Wrangle, visualize, and communicate data to management

RESPONSIBILITIES
- Conduct research
- Extract, clean, and analyze data from varied sources
- Solve problems
- Build automation tools
- Communicate findings to management

CAREER POSSIBILITIES
- The majority of data scientists work in the **technology industry.**
- Other options include marketing, consulting, healthcare and pharmaceuticals, finance, government, gaming, and many more.

RESOURCES:
https://insidebigdata.com/2017/08/05/benefits-data-scientist-career/
https://www.glassdoor.com/Salaries/us-data-scientist-salary-SRCH_IL.0,2_IN1_KO3,17.htm
https://blog.udacity.com/2014/11/data-science-job-skills.html
https://online.rutgers.edu/resources/infographics/what-can-you-do-with-a-career-in-data-science/?program=ms

THE COMPUTER MERCHANT, LTD.
THE IT STAFFING COMPANY

The Life of a
DATA SCIENTIST

Data scientists extract knowledge, insights, or solutions from big data. Here's a look at the life of a data scientist, based on two surveys of data scientists around the globe.

Data Science Methods[1]
64% Logistic regression
50% Decision trees
46% Random forests
38% Neural networks

AVERAGE AGE[1]
30 years old

Coding Languages Used[1]
76% Python
59% R
54% SQL

Average Salary - Senior Data Scientist (in USD)[2]
$138,068 USA
$85,247 Europe
$71,086 APAC

Biggest Barriers to Great Data Science Work[1]
49% Dirty data
42% Lack data science talent
37% Lack management/ financial support
30% Lack clear question to answer

Type of Data Used at Work[1]
66% Relational data
53% Text data
18% Image data

Highest Education Level Attained[1]
42% Master's degree
32% Bachelor's degree
16% Doctoral degree

Satisfaction with Job[2]
52% to 65% are unhappy in their current jobs
46% would leave due to higher earning potential elsewhere

CloudFactory provides curated teams that can scale fast and structure data accurately on virtually any platform.

Sources:
1 - *2017 State of Data Science & Machine Learning*, Kaggle
2 - *2018 Data Science Salary Report*, Big Cloud

cloudfactory.com

cloudfactory

MODERN DATA SCIENTIST

Data Scientist, the sexiest job of 21th century requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS
- Machine learning
- Statistical modeling
- Experiment design
- Bayesian inference
- Supervised learning: decision trees, random forests, logistic regression
- Unsupervised learning: clustering, dimensionality reduction
- Optimization: gradient descent and variants

PROGRAMMING & DATABASE
- Computer science fundamentals
- Scripting language e.g. Python
- Statistical computing package e.g. R
- Databases SQL and NoSQL
- Relational algebra
- Parallel databases and parallel query processing
- MapReduce concepts
- Hadoop and Hive/Pig
- Custom reducers
- Experience with xaaS like AWS

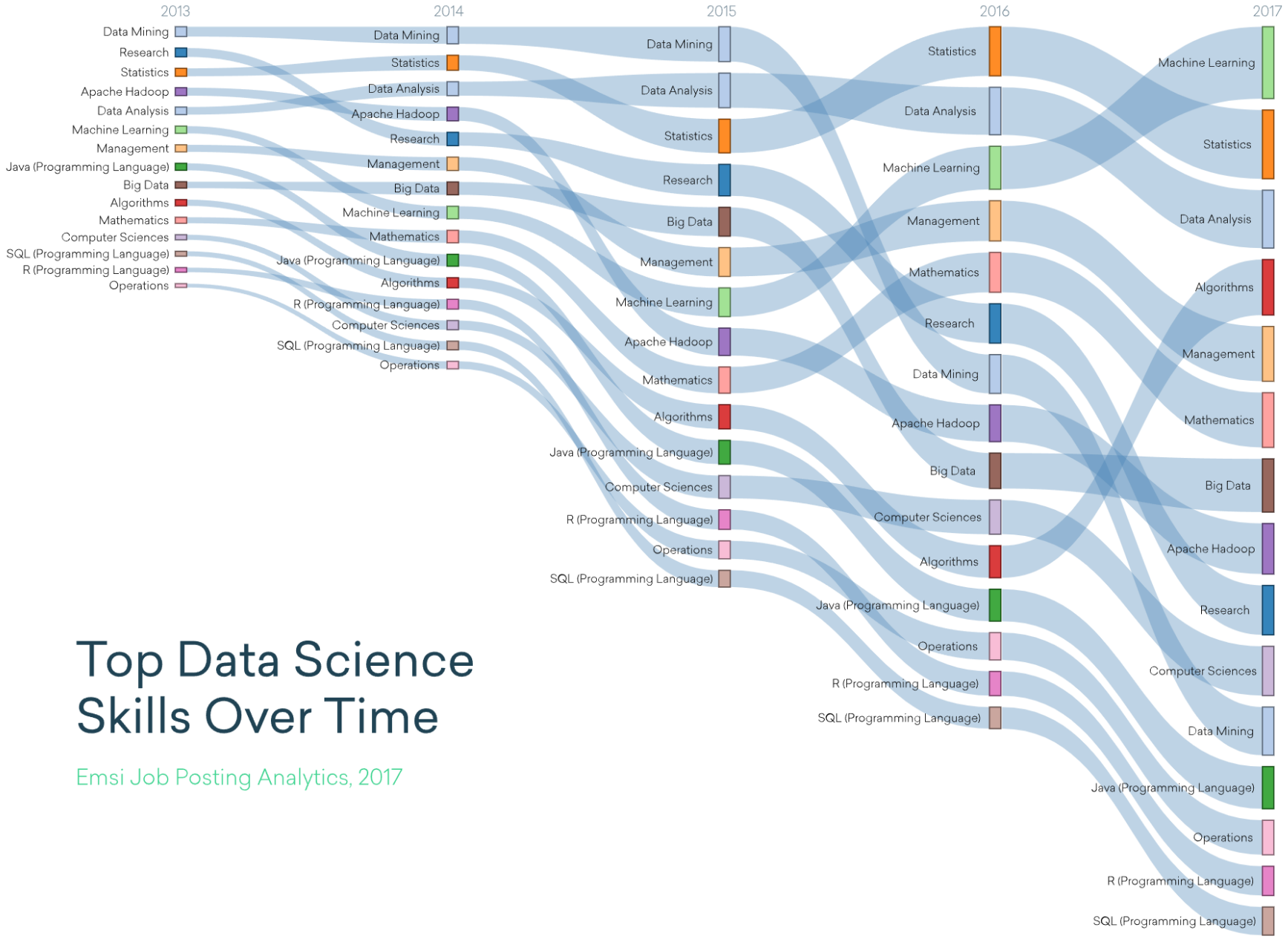
DOMAIN KNOWLEDGE & SOFT SKILLS
- Passionate about the business
- Curious about data
- Influence without authority
- Hacker mindset
- Problem solver
- Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION
- Able to engage with senior management
- Story telling skills
- Translate data-driven insights into decisions and actions
- Visual art design
- R packages like ggplot or lattice
- Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

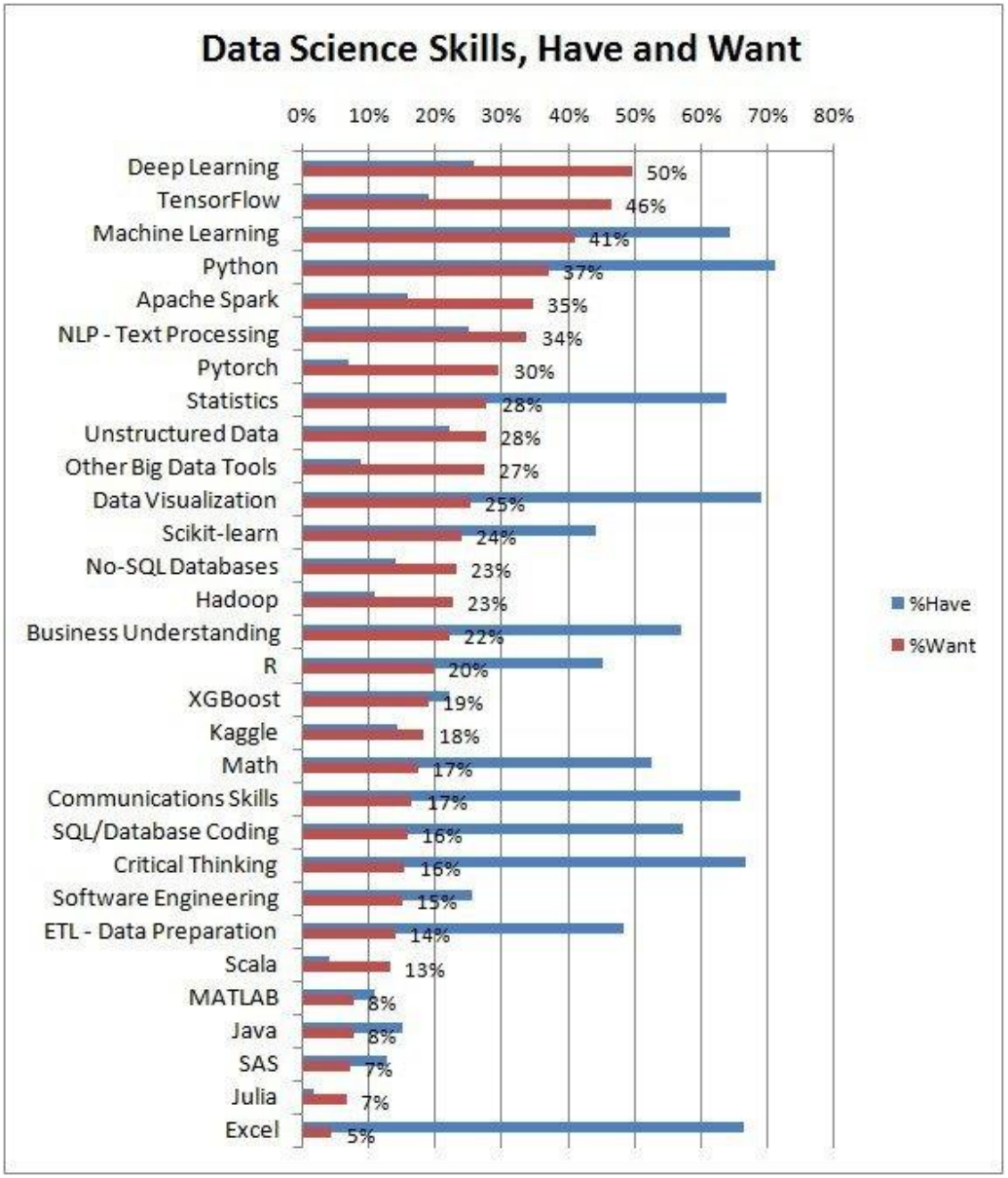
MarketingDistillery.com is a group of practitioners in the area of e-commerce marketing. Our fields of expertise include: marketing strategy and optimization: customer tracking and on-site analytics: predictive analytics and econometrics: data warehousing and big data systems: marketing channel insights in Paid Search, SEO, Social, CRM and brand.

Marketing DISTILLERY

데이터사이언스의 스킬 수요



데이터사이언스의 스킬 수요



* kdnuggets Poll, 2019

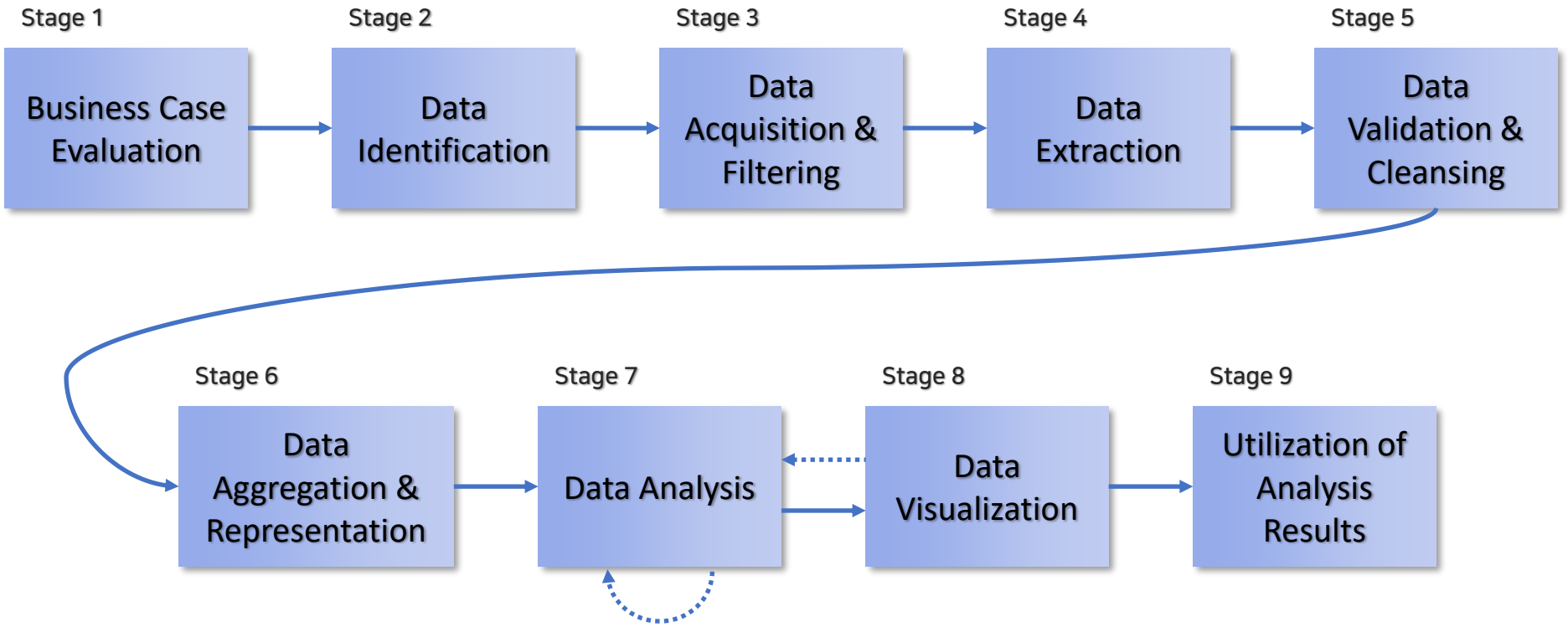
빅데이터 분석 프로젝트의 주요 역할

- Business User
  - 프로젝트의 업무를 이해하는 사람
    - 의료 데이터 분석의 경우 의사나 간호사
    - 금융 데이터 분석의 경우 은행원
  - 프로젝트 작업 팀에 세부 요구사항 및 조언을 제공
- Project Sponsor
  - 프로젝트 시작 책임자, 발주자
  - 자금을 제공하는 사람
  - 주요 관심사를 소개하고 원하는 결과를 요구함
  - 팀의 최종 결과물에서 가치를 평가
- Project Manager
  - 목표 달성을 위해 이정표를 확인하고 회의를 주관하는 사람

빅데이터 분석 프로젝트의 주요 역할

- Business Intelligence Analyst
  - 프로젝트 팀 내의 업무 분야 전문가
  - 보고서를 만들고 데이터와 소스에 대해서 알고 있는 사람
- Database Administrator (DBA)
  - 팀의 요구사항을 만족시키는 데이터베이스 환경을 제공
  - 테이블에 대한 권한, 보안 단계, 데이터의 저장소 위치 등을 책임
- Data Engineer
  - 데이터 관리, 추출, 분석 샌드박스 등을 지원하기 위한 높은 수준의 기술 제공
- Data Scientist
  - 분석 기술, 데이터 모델링에 대한 전문 지식 지원
  - 전반적인 데이터 분석 목표가 달성되었는지 확인
  - 프로젝트에 분석 모델을 제시하고 데이터가 유효한지 확인

빅데이터 분석 수명 주기



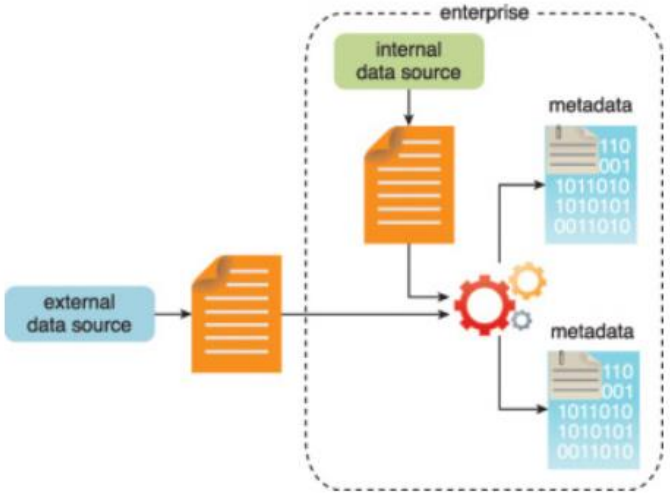
* Big Data Fundamentals

빅데이터 분석 수명 주기

- 1단계: 비지니스 사례 평가 (Business Case Evaluation)
  - 프로젝트의 목표와 비지니스 요구사항 정의
  - 핵심성과지표(KPI)나 SMART(Specific, Measurable, Attainable, Relevant, Timely)한 목표 설정 필요
  - 이 과정을 통해 어떤 자원이 필요할 지, 어떤 문제를 분석해야 할 지 이해하는데 도움을 줌
  - 빅데이터 도입에 필요한 예산 결정에 참고 역할
- 2단계: 데이터 확인 (Data Identification)
  - 신뢰성 있는 데이터 출처인지 확인 - 신뢰성 있는 입력이어야 출력도 신뢰성이 확보됨
  - 프로젝트와 관련된 가능한 많은 데이터 확보는 높은 통찰력을 제공 – 패턴이나 상관관계를 발견할 확률이 상승

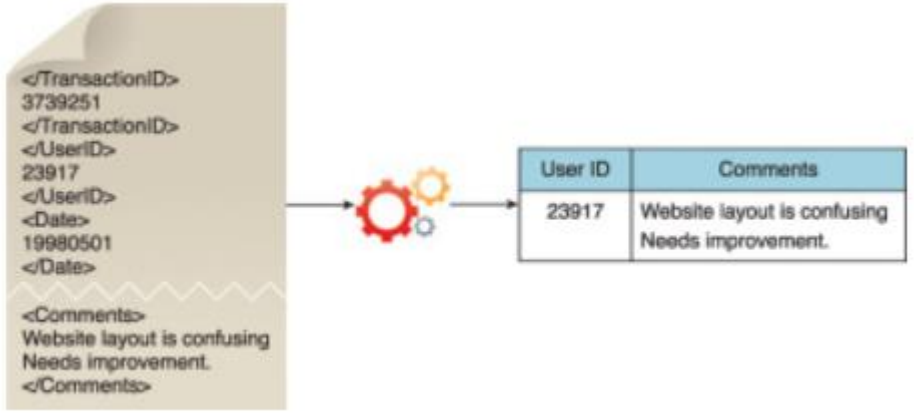
빅데이터 분석 수명 주기

- 3단계: 데이터 습득과 여과 (Data Acquisition and Filtering)
  - 확인된 다양한 데이터 출처(외부나 내부)에서 데이터를 획득
  - 외부의 데이터들은 노이즈가 있을 수 있으므로, 여과해야 한다. (removal)
    - 정제된 데이터들이 다른 유형의 분석에는 유용할 수 있으므로 사본을 유지
  - 메타데이터 추가
    - 분류나 쿼리의 성능을 향상
    - 메타데이터: 데이터셋의 크기, 구조, 출처, 생성시간, 수집시간, 언어 등의 정보



빅데이터 분석 수명 주기

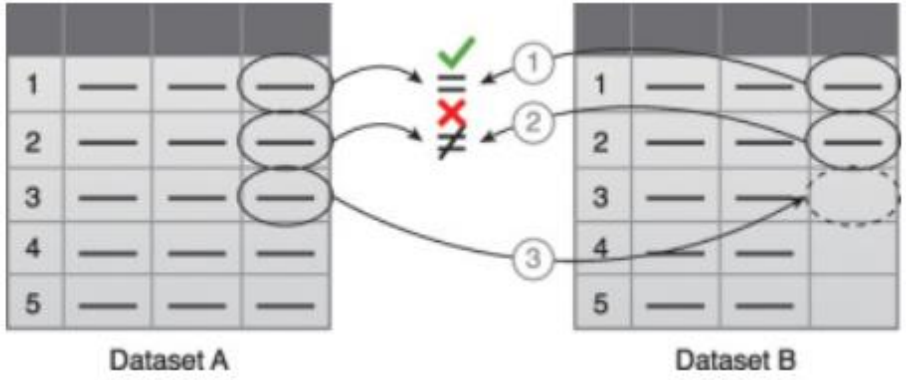
- 4단계: 데이터 추출 (Data Extraction)
  - 데이터를 솔루션에서 필요한 형식에 맞게 추출
  - xml로부터 추출



  - JSON으로부터 추출

빅데이터 분석 수명 주기

- 5단계: 데이터 검증 및 정제 (Data Validation and Cleansing)
  - 빅데이터는 유효하지 않은 값을 가질 수 있음 (잘못된 데이터는 잘못된 결과를 야기)
  - 업무 전문가들이 정한 규칙으로 검증하여 데이터 값의 유효성 여부를 판단 (valid, invalid)
  - 검증하여 유효한 데이터 확보



1행: A와 B에서 3열의 데이터 값이 같은 경우 앞의 데이터들로 검증했더니 3열의 데이터는 유효

2행: A와 B에서 3열의 데이터 값이 같은 경우 앞의 데이터들로 검증했더니 3열의 데이터는 무효 -> 삭제

3행: 앞의 데이터들로 검증했더니 A의 값을 B에 복사해서 사용 가능

빅데이터 분석 수명 주기

- 6단계: 데이터 통합 및 표현 (Data Aggregation and Representation)
  - 여러 데이터는 여러 데이터 셋에 분산되어 존재 하므로 하나의 데이터셋으로 통합하여 표현
  - 데이터 통합에 의해 빅데이터 분석 과정의 속도가 증가
  - 데이터 구조 (Data Structure)의 차이

  - 데이터 의미(Semantics)의 차이
    - 데이터 항목에 surname, lastname일 경우
    - 프로그램에서 동의어 사전을 제공
  - 이 단계는 매우 복잡

빅데이터 분석 수명 주기

- 7단계: 데이터 분석 (Data Analysis)
  - 코드나 알고리즘의 실행으로 실제 결과를 이끌어내는 단계
  - 여러 통계분석이나 데이터마이닝 기법을 사용하여 패턴을 발견하거나 수학적인 모델을 생성하는 작업
  - 결과를 이끌어낼 때까지 반복적인 수행
  - 확증적 데이터 분석 (Confirmatory Data Analysis)
    - 연역적인 방법
    - 가설을 제안하고 데이터 분석으로 가설을 입증(반증)
  - 탐색적 데이터 분석 (Exploratory Data Analysis)
    - 귀납적인 방법
    - 데이터마이닝과 관련된 방법
    - 가설을 미리 만들지 않음
    - 패턴이나 이상 현상을 쉽게 발견할 수 있는 일반적인 방향 제공

빅데이터 분석 수명 주기

- 8단계: 데이터 시각화 (Data Visualization)
  - 분석을 잘하는 것과 모두에게 이해시키는 것은 별개
  - 분석된 결과를 시각화하여 비지니스 사용자들의 해석을 용이하게 함
  - 데이터 시각화로 비지니스 사용자들은 다양한 데이터 분석을 요구 가능(feedback)

빅데이터 분석 수명 주기

- 9단계: 분석 결과 활용 (Utilization of Analysis Results)
  - 분석 결과를 제공한 뒤 생성된 모델을 어떻게, 어디에 활용할 것인가
  - 모델은 수식, 규칙 등 다양한 형태로 존재
  - 모델은 다음과 같이 사용될 수 있다.
    - 결과가 자동으로 시스템에 적용되어 기업 운영 최적화, 실적 향상에 활용
    - 패턴, 상관관계, 분석으로 발견된 이상치로 비지니스 프로세스 개선
    - 기존 시스템에 반영되거나 새로운 시스템이나 소프트웨어 개발에 반영