PART 1. 데이터의 이해

데이터 준전문가

ADSP, Advanced Data Analytics semi-Professional

류영표 강사
ryp1662@gmail.com

강사 소개

류영표

Youngpyo Ryu

동국대학교 수학과/응용수학 석사수료

現 Upstage AI X 네이버 부스트 캠프 AI tech 1~6기 멘토

前 Innovation on Quantum & CT(IQCT) 이사

前 한국파스퇴르연구소 Image Mining 인턴(Deep learning)

前 ㈜셈웨어(수학컨텐츠, 데이터 분석 개발 및 연구인턴)

강의 경력

- 현대자동차 연구원 강의 (인공지능/머신러닝/딥러닝/강화학습)
- ㈜ 모두의연구소 Aiffel 1기 퍼실리테이터(인공지능 교육)
- 인공지능 자연어처리(NLP) 기업데이터 분석 전문가 양성과정 멘토
- 공공데이터 청년 인턴 / SW공개개발자대회 멘토
- 이젠 종로 아카데미(파이썬, ADSP 강사)
- 최적화된 도구(R/파이썬)을 활용한 애널리스트 양성과정(국비과정) 강사
- 한화, 하나금융사, 한전 KDN 교육
- 인공지능 신뢰성 확보를 위한 실무 전문가 자문 활동
- 인공지능 학습용 데이터 구축 사업 품질검증 전문가 자문 활동
- 보건 · 바이오 AI활용 S/W개발 및 응용전문가 양성과정 강사
- Upstage AI X KT 융합기술원 기업교육 모델최적화 담당 조교

주요 프로젝트 및 기타사항

- 개인 맞춤형 당뇨병 예방·관리 인공지능 시스템 개발 및 고도화(안정화)
- 폐플라스틱 이미지 객체 검출 경진대회 3위
- 인공지능(AI)기반 데이터 사이언티스트 전문가 양성과정 1기 수료
- 제 1회 산업 수학 스터디 그룹 (질병에 영향을 미치는 유전자 정보 분석)
- 제 4,5회 산업 수학 스터디 그룹 (피부암, 유방암 분류)
- 빅데이터 여름학교 참석 (혼잡도를 최소화하는 새로운 노선 건설 위치의 최적화 문제)

ADSP, Advanced Data Analytics semi-professional

- 데이터 이해에 대한 기본 지식을 바탕으로 데이터 분석 기획, 데이터 분석등의 직무를 수행하는 실무자.
- 국가 공인 자격증
- 입문, 기초 수준의 시험으로 실기 없이 필기시험으로 구성됨.
- 응시료는 1회 5만원

ADSP

응시자격

응시자격
제한 없음

합격기준

합격기준	과락기준
총점 60점 이상	과목별 40% 미만 취득

ADP

응시자격

데이터분석 전문가 자격검정 시험의 응시자격은 아래와 같으며 응시자격은 필기시험일 기준 시험일 이전에 응시자격 요건이 충족되어야 한다. 경력/학력기준 또는 자격보유기준 중 한가지의 요건이 충족될 경우 응시자격이 부여된다

응시자격	
학력 및 경력 기준	박사학위를 취득한자
	석사학위를 취득하고 해당 분야의 실무경력 1년 이상인자
	학사학위를 취득하고 해당 분야의 실무경력 3년 이상인자
	전문대학 졸업후 해당 분야의 실무경력 6년 이상인자
	고등학교 졸업후 해당 분야의 실무경력 9년 이상인자
자격보유 기준	데이터분석 준전문가 자격을 취득한 자
실기	필기시험 합격자 발표일로부터 2년의 유효기간을 가지며, 2년 이내에 시행되는(시행일 기준) 실기검정에 응시 가능 (다만, 필기시험 합격일로부터 2년간 검정이 2회 미만으로 시행된 경우 그 다음에 이어지는 해당 필기시험 1회를 면제한다.)

합격기준

데이터분석 전문가 자격검정 시험의 합격기준은 아래와 같으며 실기시험 합격자는 응시자격 증빙서류를 제출하여야 한다.

합격기준		과락기준
필기합격	총점 100점 기준 70점 이상	과목별 40% 미만 취득
실기합격	실기 총점 100점 기준 75점 이상	
최종합격	응시자격심의 서류 통과자	

ADSP, Advanced Data Analytics semi-professional

□ 변경 내용

o 검정과목 : 전문가(ADP) 필기

- 5과목 데이터 시각화

변경 전		⇒	변경 후	
주요항목	**세부항목**		**주요항목**	**세부항목**
시각화 구현	• 분석 도구를 이용한 시각화 구현 : R		시각화 구현	• 분석 도구를 이용한 시각화 구현

o 검정방법 : 준전문가(ADsP)

변경 전	⇒	변경 후
선택형 40문항, 단답형 10문항		선택형 50문항

□ 적용 시기

o 2024년 제32회 전문가(ADP) 필기 및 제40회 준전문가(ADsP)부터 적용

ADSP, Advanced Data Analytics semi-professional

과목별 세부 내용

시험과목	과목별 세부 항목
데이터 이해	데이터의 이해
	데이터의 가치와 미래
	가치 창조를 위한 데이터 사이언스와 전략 인사이트
데이터분석 기획	데이터분석 기획의 이해
	분석 마스터 플랜
데이터분석	R기초와 데이터 마트
	통계분석
	정형 데이터 마이닝

데이터 이해	데이터의 이해	데이터와 정보	데이터분석	R기초와 데이터 마트	R기초
		데이터베이스의 정의와 특징			데이터 마트
		데이터베이스 활용			결측값 처리와 이상값 검색
	데이터의 가치와 미래	빅데이터의 이해		통계분석	통계학 개론
		빅데이터의 가치와 영향			기초 통계분석
		비즈니스 모델			다변량 분석
		위기 요인과 통제 방안			시계열 예측
		미래의 빅데이터			
	가치 창조를 위한 데이터 사이언스와 전략 인사이트	빅데이터분석과 전략 인사이트		정형 데이터 마이닝	데이터 마이닝 개요
		전략 인사이트 도출을 위한 필요 역량			분류분석(Classification)
		빅데이터 그리고 데이터 사이언스의 미래			군집분석(Clustering)
					연관분석(Association Analysis)
데이터분석 기획	데이터분석 기획의 이해	분석 기획 방향성 도출			
		분석 방법론			
		분석 과제 발굴			
		분석 프로젝트 관리 방안			
	분석 마스터 플랜	마스터 플랜 수립			
		분석 거버넌스 체계 수립			

ADSP, Advanced Data Analytics semi-professional

[2024년도 데이터 전문가 자격시험 일정]

구분	회차		접수기간	시험일	점수공개	결과발표	서류제출
빅데이터 분석기사	제8회	필기	3.4~8	4.6(토)	4.19~23	4.26	4.29~5.9
		실기	5.20~24	6.22(토)	7.5~9	7.12	–
	제9회	필기	8.5~9	9.7(토)	9.20~24	9.27	9.30~10.10
		실기	10.28~11.1	11.30(토)	12.13~17	12.20	–
데이터분석 전문가 ADP	제32회	필기	1.22~28	2.24(토)	3.15~19	3.22	–
		실기	3.22~29	4.27(토)	5.17~21	5.24	5.24~3.1
	제33회	필기	7.1~5	8.10(토)	8.30~9.3	9.6	–
		실기	9.9~13	10.12(토)	11.1~5	11.8	11.8~15
데이터분석 준전문가 ADsP	제40회	–	1.22~28	2.24(토)	3.15~19	3.22	–
	제41회	–	4.8~12	5.11(토)	5.31~6.4	6.7	–
	제42회	–	7.1~5	8.10(토)	8.30~9.3	9.6	–
	제43회	–	9.30~10.4	11.3(일)	11.22~26	11.29	–
SQL 전문가 SQLP	제50회	–	1.29~2.2	3.9(토)	3.29~4.2	4.5	4.5~12
	제51회	–	7.22~26	8.24(토)	9.6~10	9.20	9.20~27
SQL 개발자 SQLD	제52회	–	1.29~2.2	3.9(토)	3.29~4.2	4.5	–
	제53회	–	4.22~26	5.25(토)	6.14~18	6.21	–
	제54회	–	7.22~26	8.24(토)	9.6~10	9.20	–
	제55회	–	10.14~18	11.17(일)	12.6~10	12.13	–
데이터아키텍처 전문가 DAP	제61회	–	2.12~16	3.16(토)	4.5~9	4.12	4.12~19
	제62회	–	8.26~30	9.28(토)	10.18~22	10.25	10.25~11.1
데이터아키텍처 준전문가 DAeP	제56회	–	2.12~16	3.16(토)	4.5~9	4.12	–
	제57회	–	8.26~30	9.28(토)	10.18~22	10.25	–
경영정보시각화능력 (BIS)	1회	필기	3.18~3.24, 4.17~4.23	5.18(토)	–	6.18	–
	1회	실기	8.28~9.3	9.28(토)	–	11.18	–
	2회	필기	9.30~10.6, 10.30~11.5	11.30(토)	–	12.31	–

출처 : https://cafe.naver.com/sqlpd

데이터 정의

- 데이터(data)라는 용어는 1646년 영국 문헌에 처음 등장하였으며,
  라틴어인 dare(주다, to give)의 과거분사형으로 '주어진 것'이란 의미로 사용 됨.
- 데이터는 추론과 추정의 근거를 이루는 사실
- 사실을 나타내는 수치 / 위키피디아
- 데이터는 객관적 사실(fact, raw material)이라는 존재적 특성을 가짐.
  → 데이터는 개별 데이터 자체로는 의미가 중요하지 않은 객관적 사실을 말함.
- 데이터는 추론, 예측 전망, 추정을 위한 근거(basis)로 기능하는 당위적 특성을 가짐.
  → 다른 객체와 상호 관계 속에서 가치를 가짐.
- 사물, 현상, 사건 인간관계 등에 관한 관찰 기록.

❖ 존재적 : 그것의 의미를 해석하려면 존재하는 것의 존재. 그 기초적인 존재방식을 명확히 하여야 함.

❖ 당위적 : 마땅히 그렇게 하거나 되어야 하는

데이터 유형

구분	**정성적 데이터** (Qualitative data)	**정량적 데이터**(Quantitative data)
형태	언어,문자 등 비정형 데이터 자료의 성질, 특징을 설명 또는 요약	수치, 도형, 기호 등 (정형 데이터) 자료를 수치화
예시	회사 매출이 증가함	30(나이), 48.4(kg)
특징	상대적으로 많은 비용과 기술적 투자가 수반됨. 주관적인 내용 통계분석이 어려움	저장, 검색,분석 활용에 용이하다 객관적 내용 통계분석이 용이함.

지식경영의 핵심 이슈

구분	암묵지(tacit knowledge)	형식지(explicit knowledge)
형태	학습과 경험을 통해 개인에게 습득되어 있지만, 겉으로 드러내지 않은 지식	문서나 메뉴얼처럼 형상화된 지식
특징	사회적으로 중요하지만 다른 사람에게 공유되기 어려움	전달과 공유가 용이
상호작용	내면화(Internalization), 공통화(socialization)	표출화(Externalization), 연결화(Combination)
예	김장김치 담그기, 자전거 타기	교과서,비디오, DB(Data base)

DIKW 피라미드(Data, Information, Knowledge, Wisdom)

근본 원리에 대한 깊은 이해를 바탕으로 도출되는 창의적 아이디어

예) A마트의 다른 상품들도 B마트보다 쌀 것이라 판단한다

상호 연결된 정보 패턴을 이해하여 이를 토대로 예측한 결과물

예) 상대적으로 저렴한 A마트에서 연필을 사야겠다

데이터의 가공 및 상관관계간 이해를 통해 패턴을 인식하고 그 의미를 부여한 데이터

예) A마트의 연필이 더 싸다

존재형식을 불문하고, 타 데이터와 상관관계가 없는 가공하기전의 순수한 수치나 기호를 의미

예) A마트는 100원에, B마트는 200원에 연필을 판매한다.

데이터베이스의 정의

- 대량의 데이터를 축적하는 기지(base)라고 불림.
- 동시에 복수의 적용 업무를 지원할 수 있도록 복수 이용자의 요구에 대응해서
  데이터를 받아들이고 저장, 공급하기 위하여 일정한 구조에 따라서 편성된 데이터의 집합

- 데이터베이스의 개념 : 체계적이고 정렬된 데이터 집합

- DBMS(Database Management System) 는 이용자가 쉽게 데이터베이스를 구축하고 유지할 수 있도록 하는 소프트웨어로서 데이터베이스와 구분되며, 일반적으로 데이터베이스와 DBMS를 함께 데이터베이스 시스템이라고 한다.

  -〉 데이터 베이스 : 문자, 기호, 음성, 화상, 영상 등 상호 연관된 다수의 컨텐츠를 정보 처리 및 정보통신 기기에 의하여 체계적으로 수집, 축적하여 다양한 용도와 방법으로 이용할 수 있도록 정리한 정보의 집합체.

데이터베이스의 특징

- 통합된 데이터(Integrated data) : 데이터베이스에서 동일한 내용의 데이터가 중복되어 있지 않음

  데이터 중복은 관리상의 복잡한 부작용을 초래

- 저장된 데이터(Stored data) : 컴퓨터가 접근할 수 있는 **저장 매체**에 저장되는 것

- 공용 데이터(Shared data) : 여러 사용자가 서로 다른 목적으로 **데이터베이스의 데이터를 공동 이용**

- 변화되는 데이터(Changeable data) : 새로운 데이터의 추가, 기존 데이터의 삭제, 갱신으로 항상 변화하면서도

  항상 현재의 정확한 데이터를 유지해야 한다

데이터베이스의 특성

- **정보의 축적 및 전달 측면**
  - 기계 가독성 : 대량의 정보를 정보처리기기가 읽고 쓸 수 있음
  - 검색 가능성 : 필요한 정보를 검색할 수 있음
  - 원격 조작성 : 정보통신망을 이용하여 원거리에서도 온라인으로 이용할 수 있음
- **정보이용 측면** 이용자의 정보 요구에 따라 다양한 정보를 신속하게 획득하고 원하는 정보를 경제적으로 찾아 낼 수 있음
- **정보관리 측면** 방대한 양의 정보를 체계적으로 축적하고 새로운 내용 추가나 갱신이 용이함
- **정보기술발전 측면** 데이터베이스는 정보처리, 검색, 관리 소프트웨어 등 네트워크 발전기술을 견인할 수 있음
- **경제, 산업적 측면** 데이터베이스는 인프라로서 특성을 가지고 있어 경제, 산업, 사회 활동의 효율성을 제고하고 국민의 편의를 증진하는 수단으로 의미를 가짐

데이터베이스의 활용

1990년대 기업내부 데이터베이스

OLTP (On-Line Transaction Processing)	OLAP (On-Line Analytical Processing)
온라인 **거래** 처리 (운영자)	온라인 **분석** 처리 (분석가)
업무 처리 기반	**데이터 분석 기반**
• 호스트 컴퓨터가 데이터베이스를 엑세스하고, 바로 **처리** 결과로 돌려보내는 형태 • 다양한 과정의 연산이 하나의 단위 프로세스로 실행되도록 하는 단순 자동화에 치우쳐 있는 시스템 • 예) 주문입력시스템, 재고관리 시스템 등 현업의 거의 모든 업무가 이와 같은 성격을 띰	• 모든 유형의 비즈니스 활동을 다양한 차원으로 나타냄 • 다차원의 데이터를 대화식으로 분석하기 위한 기술 • 의사 결정에 활용할 수 있는 통계적인 요약 정보를 제공 • 예) OLTP에서 처리된 트랜잭션 데이터를 분석해 제품의 판매 추이, 구매 성향 파알, 재무 회계 분석 등을 프로세싱
호스트 + 여러단말 -> 호스트가 요구처리해서 결과 내보냄 => 데이터 **갱신** 위주	다양한 데이터에 접근하여 '분석'을 통해 정보 도출 => 데이터 **조회** 위주

❖ 트랜잭션 : 데이터 베이스의 일관성(데이터 베이스의 상태가 일관되어야 한다는 성질)을 보전하는 프로그램의 실행단위.

데이터베이스의 활용

2000년대 기업 내부 데이터 베이스

CRM (Customer Relationship Management)	SCM (Supply Chain Management)
고객관계관리	공급망 관리
• 기업이 고객과 관련된 내·외부 자료를 분석·통합해 고객 중심 자원을 극대화 • 고객특성에 맞게 마케팅 활동을 계획·지원·평가하는 과정 • 고객에 대한 정보를 이해하고 그로 인해 고객이 원하는 제품과 서비스를 제공하므로 인해 고객과의 관계를 장기적으로 구축하는 고객관계관리 프로세스	• 기업에서 원재료의 생산·유통 등 모든 공급망 단계를 최적화 • 수요자가 원하는 제품을 원하는 시간과 장소에 제공 • 거래관계에 있는 기업들간 IT를 이용한 실시간 정보공유를 통해 시장이나 수요자들의 요구에 기민하게 대응토록 지원

데이터베이스의 활용

분야별 데이터베이스 소개

분야	데이터베이스	내용
제조분야	ERP (Enterprise Resource Planning)	• **기업 자원 관리** • 기업의 전 부문에 걸쳐 독립적으로 운영되던 각종 관리 시스템의 경영자원을 하나의 통합 시스템으로 재구축함으로써 생산성을 극대화 • 기업 전체의 자원을 효과적이고 통합적으로 관리하여 경영의 효율화를 기함
	CRM (Customer Relationship Management)	• **고객관계관리** • 기업이 고객과 관련된 내·외부 자료를 분석·통합해 고객 중심 자원을 극대화하고 이를 토대로 고객 특성에 맞게 마케팅 활동을 계획·지원·평가하는 과정
	RTE (Real-Time Enterprise)	• 회사의 주요 경영정보를 통합관리하는 **실시간 기업**의 새로운 기업경영 시스템 • 회사 전 부문의 정보를 하나로 통합함으로써 경영자의 빠른 의사결정을 이끌어냄 • 가트너는 '최신 정보를 사용해 자사의 핵심 비즈니스 프로세스들의 관리와 실행 과정에서 생기는 지연 사태를 지속적으로 제거함으로써 경쟁하는 기업'으로 정의

❖ BI(Business Intelligence) : 데이터를 통합/분석하여 기업 활동에 연관된 의사결정을 돕는 프로세스

데이터베이스의 활용

분야별 데이터베이스 소개

분야	데이터베이스	내용
금융부문	EAI (Enterprise Application Integration)	• **기업 애플리케이션 통합** • 기업 내 상호 연관된 모든 애플리케이션을 유기적으로 연동 • 필요한 정보를 중앙 집중적으로 통합, 관리, 사용할 수 있는 환경을 구현 • e-비즈니스를 위한 기본 인프라
	EDW (Enterprise Data Warehouse)	• 기존 DW를 전사적으로 확장한 모델로 BPR, CRM, BSC 같은 다양한 분석 애플리케이션들을 위한 원천이 됨 • 기업 리소스의 유기적 통합, 다원화된 관리체계 정비, 데이터의 중복 방지 등을 위해 시스템을 재설계 • 업무 시스템별로 흩어져 있는 업무정보를 최종 사용자가 쉽게 활용할 수 있도록 함

- BPR(Business Process Reengineering) : 현재 하고 있는 일을 개선하는 것이 아니라 처음부터 다시 시작하는 혁명적인 개념에서 출발
- BSC(Balanced Scorecard(균형 성과표)) : 조직의 목표와 전략을 효율적으로 실행하고 관리하기 위한 일종의 경영관리 기법.

데이터베이스의 활용

분야별 데이터베이스 소개

분야	데이터베이스	내용
유통부문	**KMS** (Knowledge Management System)	• **지식 관리 시스템** • 지적 재산의 중요성이 커짐에 따라 기업 경영을 지식이라는 관점에서 관리하는 접근방식
	RFID (Radio Frequency; RF)	• 주파수를 이용해 ID를 식별하는 시스템으로 전자태그로 불림 • 전파를 이용해 먼거리에서 정보를 인식

빅데이터 (Big Data)

“빅 데이터는 미래 경쟁력을 좌우하는 21세기의 원유이다.”

– Gartner

“빅 데이터는 혁신, 경쟁과 생산성에 있어서 차세대 첨단 주자다.”

– McKinsey

“빅 데이터는 화폐나 금처럼 새로운 자산이 될 것이다 ”

– Davos Forum

빅데이터 (Big Date) 란?

빅데이터 ≠ 대용량 자료

- “빅데이터는 일반적인 데이터베이스 소프트웨어로 저장, 관리, 분석할 수 있는 범위를 초과하는 규모의 데이터다.”(McKinsey, 2011)
- 조직의 내 외부에 존재하는 다양한 형태의 데이터를 수집, 처리, 저장
- 목적에 맞게 분석함으로써 해당 분야의 필요 지식을 추출하고
- 이를 조직의 전략적 의사결정에 활용하거나 시스템화하여 상시적으로 생산성 향상에 활용하거나 새로운 비즈니스 모델의 창출에 활용하고자 하는 패러다임.

지식을 탐색하는 방법들

연역

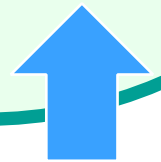
이미 알고 있는 일반적인 지식, 법칙,
원리로부터 논리적인 규칙,
즉 합리적 추론에 따라 필연적 결론
을 이끌어 내는 것

합리적인 방법

귀납

개별 사례들에 대한 관찰을 통해
일반적인 결론을 이끌어 내는 접근방법
즉, 과거의 사례나 축적된 데이터를 분석
하여 현재의 문제를 해결하는 접근법

경험주의 방법



빅데이터는 전형적인 귀납적 지식 탐색 접근법에 해당

빅데이터의 출현 배경?

정보 환경의 변화

- 정보 기술의 발전에 따른 컴퓨터 활용 확대
- 소셜 미디어의 급격한 확산
- IOT 확산에 따른 센서 데이터 증대
- 멀티미디어 콘텐츠와 콘텐츠 사용에 관한 정보 증가
- 데이터 저장 및 처리기술의 발전
- 클라우드 컴퓨팅 기술의 발전

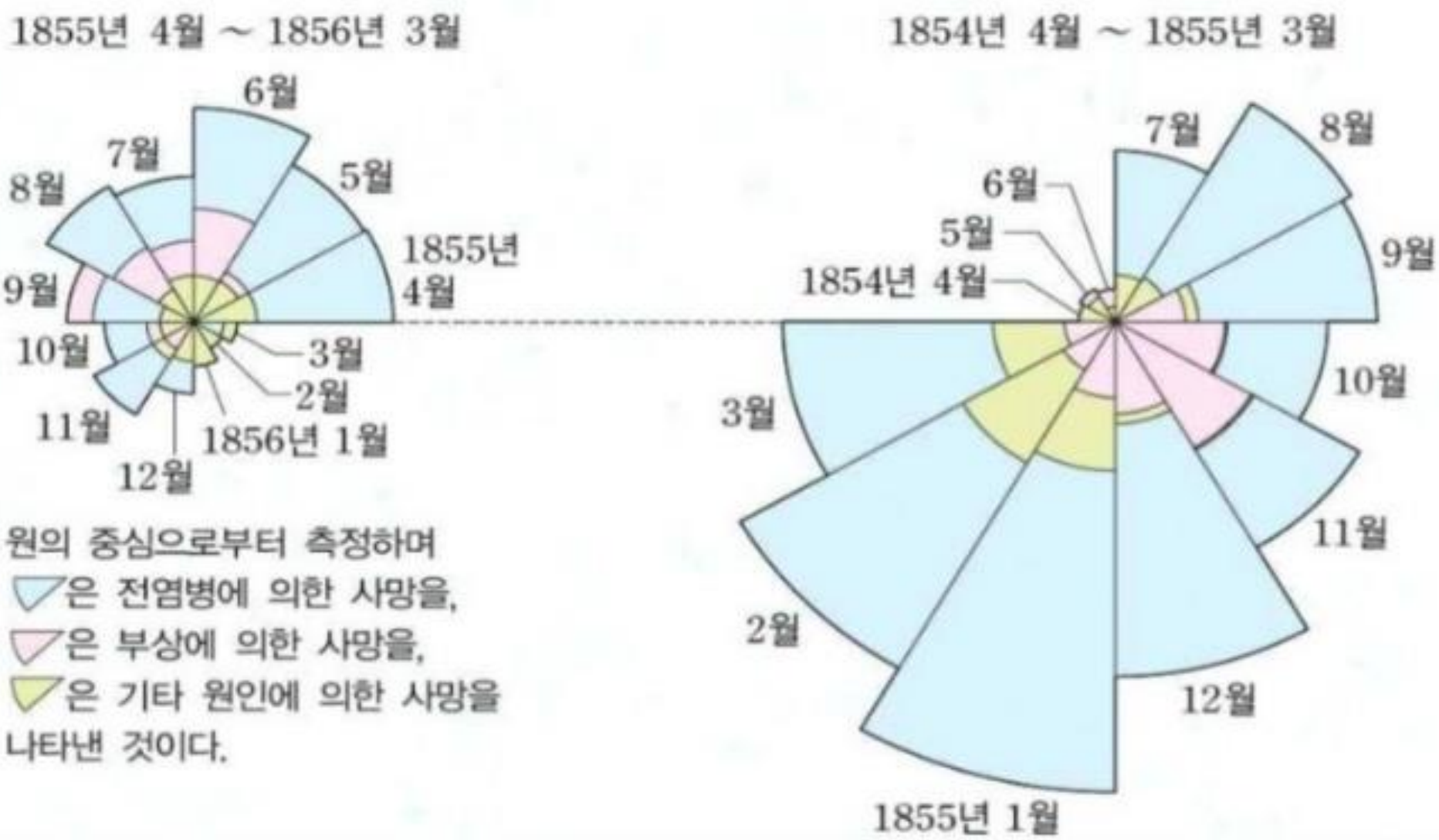
데이터 증가와
수집 비용의 감소

오래된 역사 (빅데이터)

나이팅게일도 사용한 데이터 분석

- 백의의 천사로 알려진 나이팅게일은 통계를 적극적으로 활용
- 1854년 러시아와 연합국 간에 '크림 전쟁 ' 발발
- 당시 런던에서 간호사로 일하고 있던 나이팅게일은 병원 내 주사망원인을 데이터에 기반으로 분석하고, 전투로 인한 사망보다 전염병으로 인한 사망자 수가 많다는 사실을 밝혀 냄
- 이 정보는 병원 위생상태 개선이 매우 급선무임을 인식하게 하여 많은 생명을 구할 수 있었음.

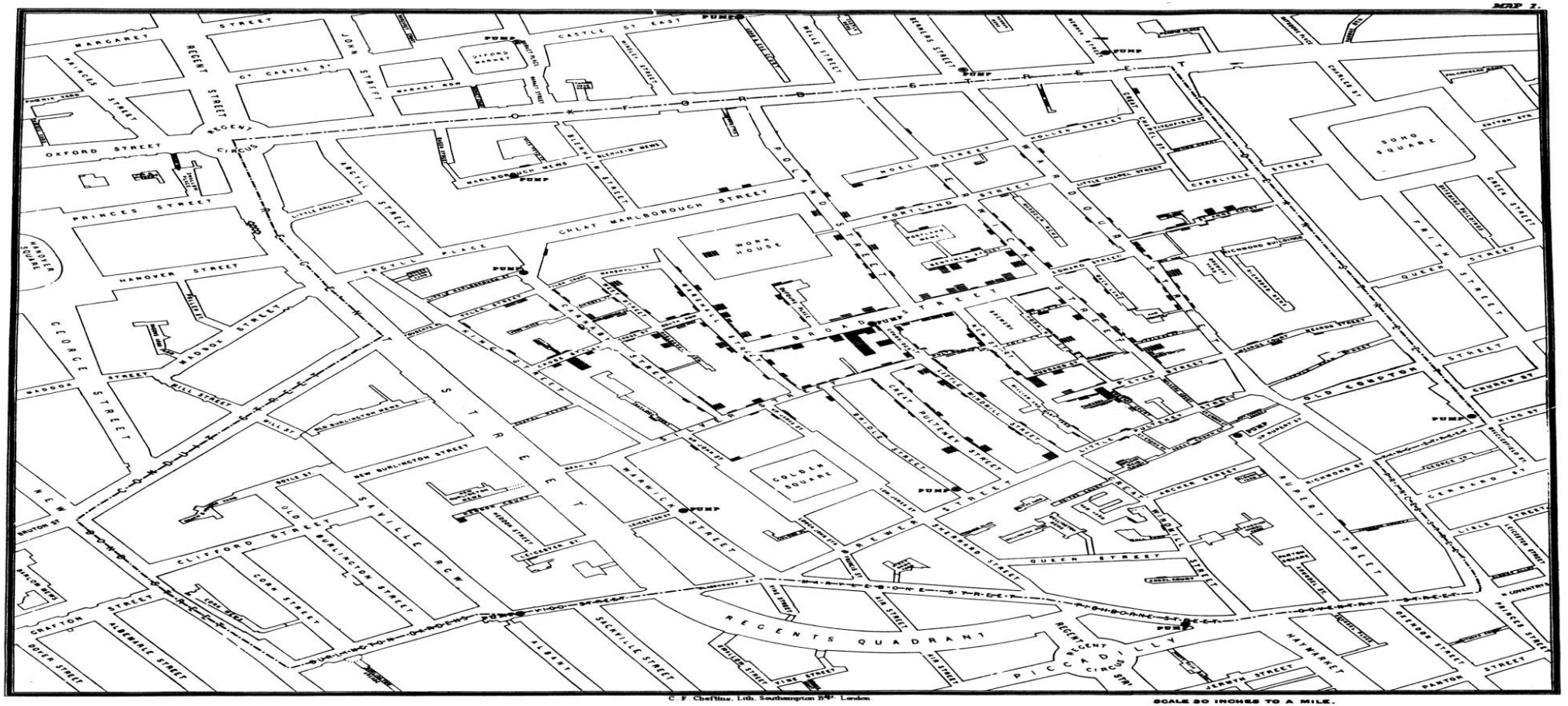
오래된 역사 (빅데이터)



출처 : www.pinterest.co.kr/pin/822399581945452823

오래된 역사 (빅데이터)

전염병 시초였던 흑사병



데이터 분석과 비즈니스 인텔리전스(Business intelligence)

- 기업이 보유하고 있는 대량의 데이터를 수집, 정리, 분석해
  이를 통해 도출된 정보를 기업의 의사결정에 활용하는 일련의 프로세스

- 기업의 사용자가 더 좋은 의사결정을 하도록 데이터를
  수집, 저장, 분석, 접근을 지원하는 기술이자 응용시스템

- Data Mining 과 같은 데이터 분석 기술

- 빅데이터란 대용량 데이터를 활용해 작은 용량에서는 얻을 수 없었던 새로운 통찰이나 가치를 추출해 내는 일이다. 나아가 이를

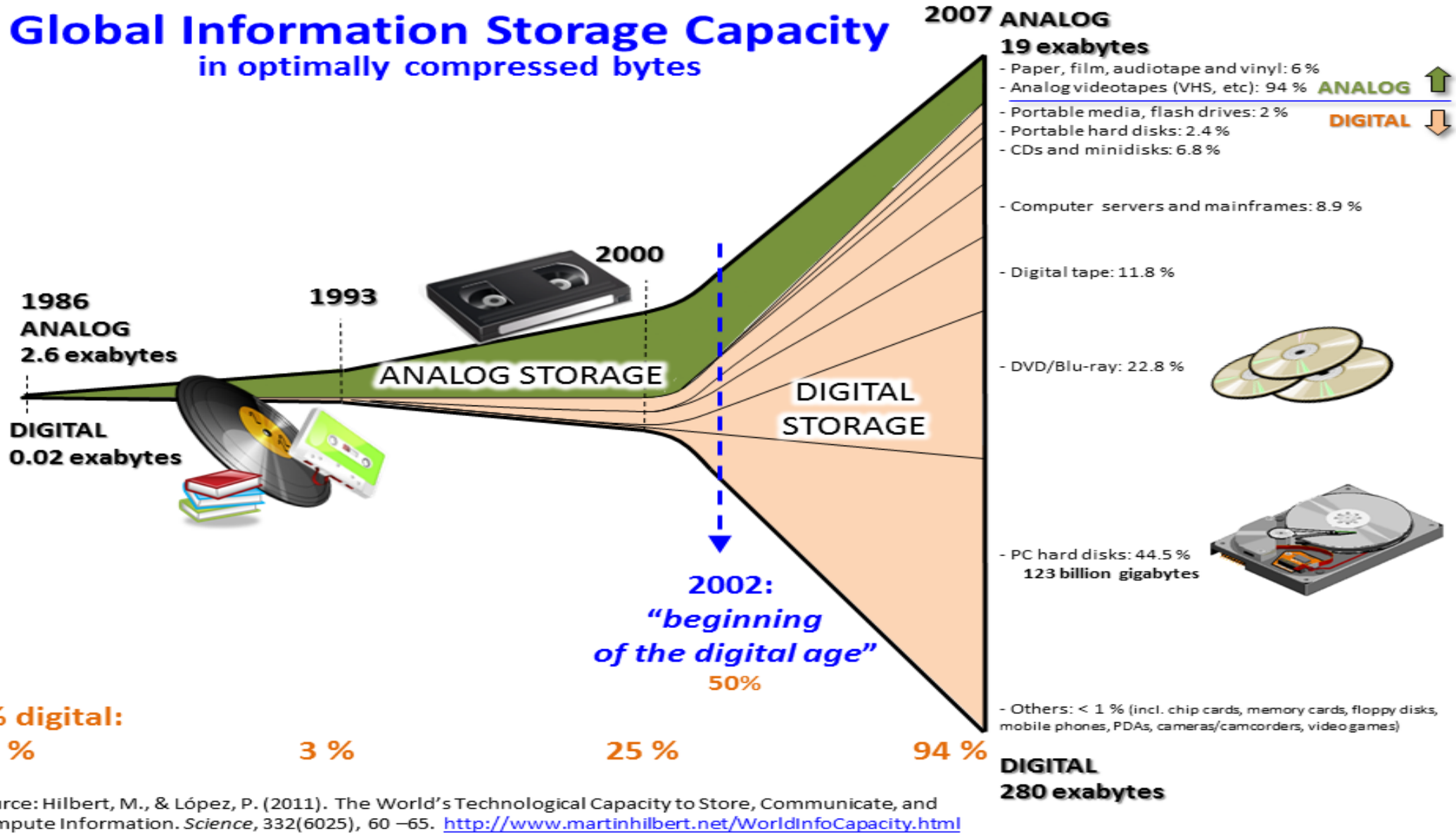
활용해 시장, 기업 및 시민과 정부의 관계 등 많은 분야에 변화를 가져오는 일이다.(Mayer Schanberget&Cukier,2013)

전통적인 데이터 분석과 빅데이터 분석의 차이

출처 : 삼성경제연구소 "빅데이터: 산업 지각변동의 진원"

출처 : m.blog.naver.com/PostView.nhn?blogId=blue91822&logNo=220253774118&proxyReferer=https:%2F%2Fwww.google.com%2F

전통적인 데이터 분석과 빅데이터 분석의 차이



Source: Hilbert, M., & López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*, 332(6025), 60 –65. http://www.martinhilbert.net/WorldInfoCapacity.html

출처 : upload.wikimedia.org/wikipedia/commons/7/7c/Hilbert_InfoGrowth.png

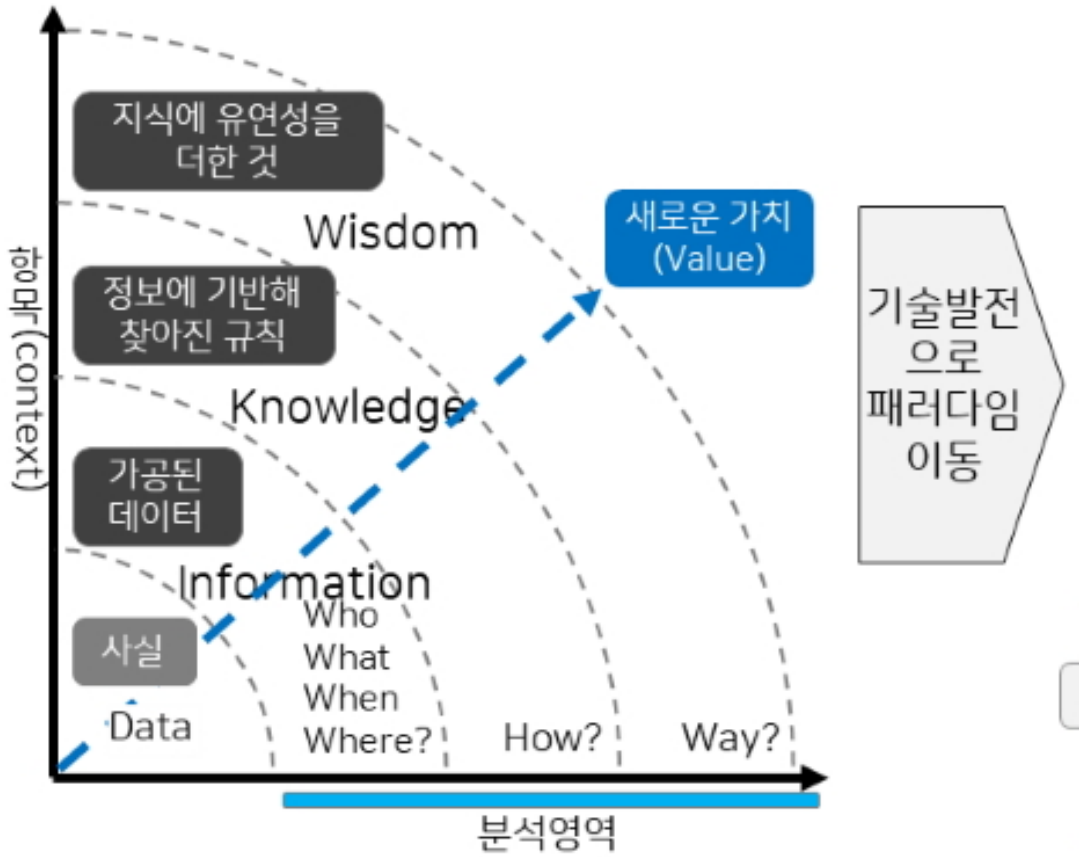
전통적인 데이터 분석과 빅데이터 분석의 차이

데이터 기억용량 단위		
바이트	1Byte	8bit
킬로바이트	1KiloByte	1024Byte
메가바이트	1MegaByte	1024KB
기가바이트	1GigaByte	1024MB
테라바이트	1TeraByte	1024GB
페타파이트	1PetaByte	1024TB
엑사바이트	1ExaByte	1024PB
제타바이트	1ZetaByte	1024EB
요타바이트	1YottaByte	1024ZB

출 처 : m.blog.naver.com/PostView.naver?isHttpsRedirect=true&blogId=history11109&logNo=221212391519

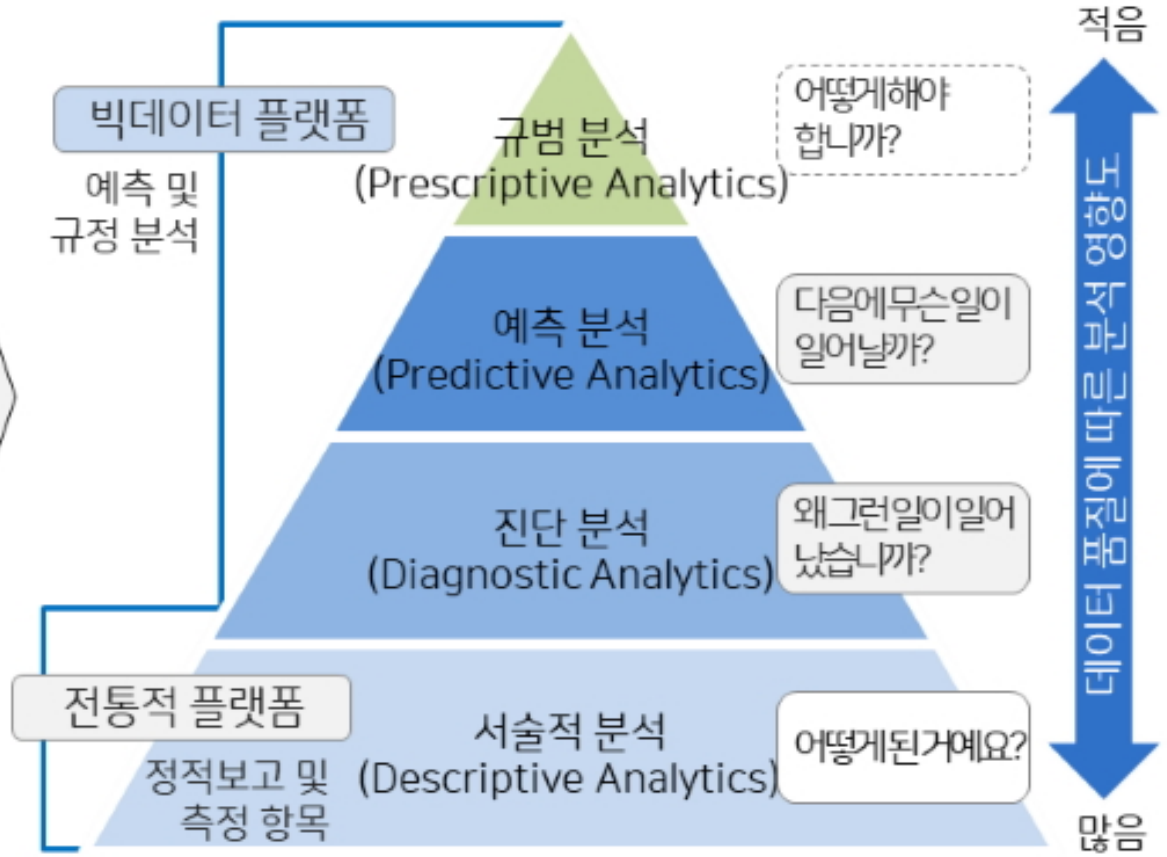
전통적인 데이터 분석과 빅데이터 분석의 차이

개념적 지식화 모델(DIKW 모델)



출처 : 로리 제니퍼 (2007). "지혜 계층 : DIKW 계층의 표현 " 자료 재구성

빅데이터 분석 모델



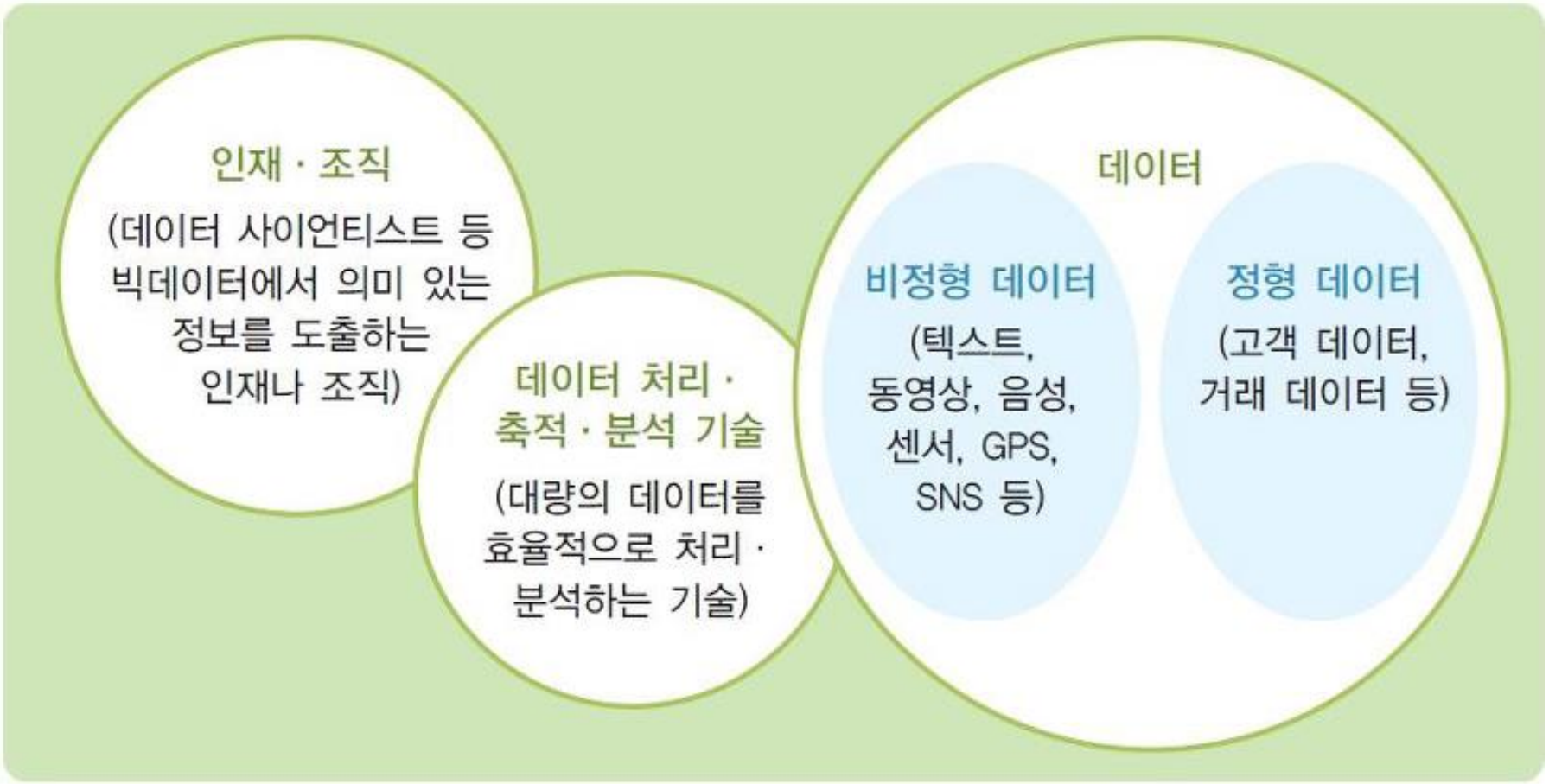
출처 : Gartner 자료 재구성

출처 : www.cisp.or.kr/archives/22193

전통적인 데이터 분석과 빅데이터 분석의 차이

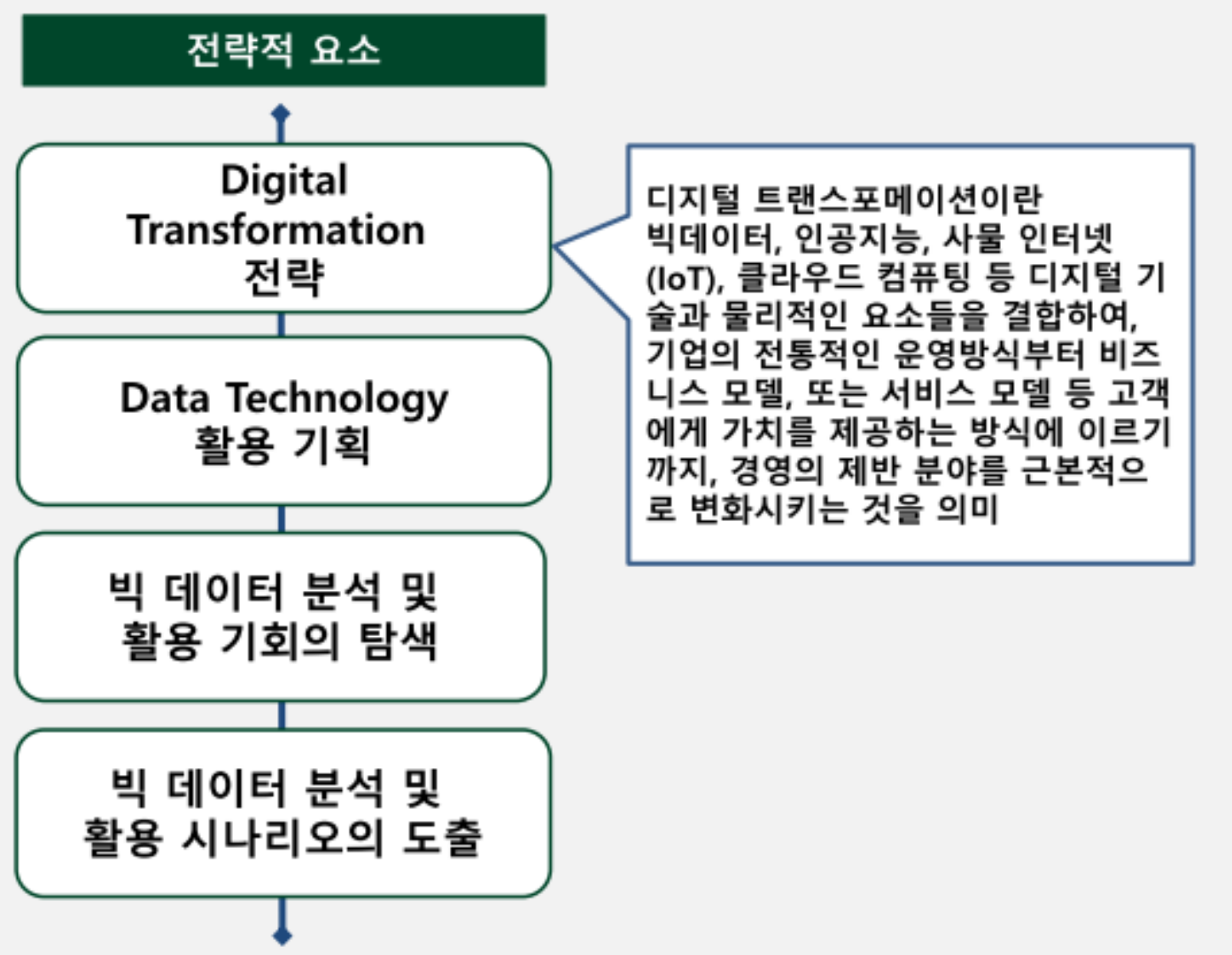
출처 : wikidocs.net/22652

빅데이터의 주요 요소

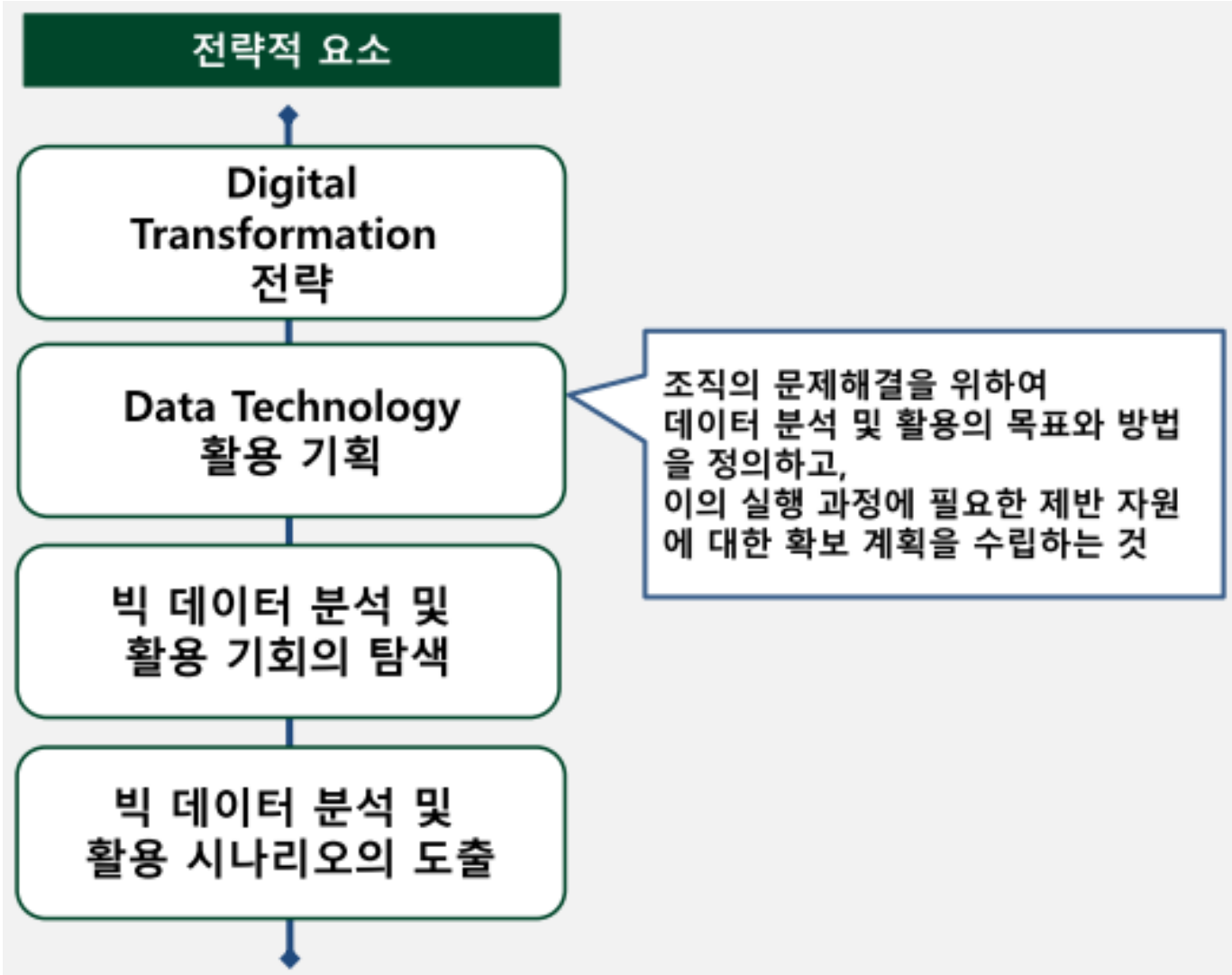


출처 : ikkison.tistory.com/66

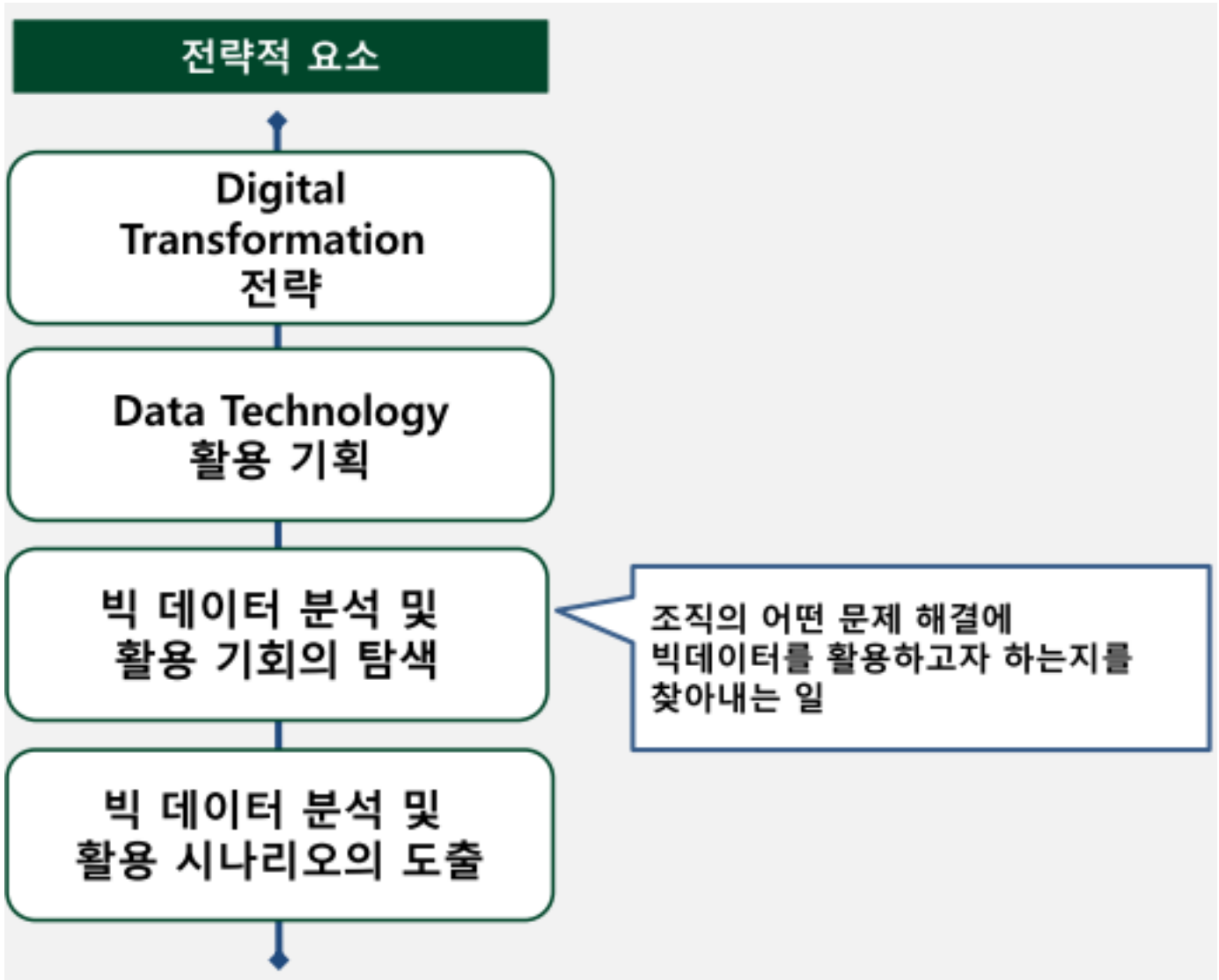
빅데이터의 주요 요소


전략적 요소
Digital Transformation 전략
Data Technology 활용 기획
빅 데이터 분석 및 활용 기회의 탐색
빅 데이터 분석 및 활용 시나리오의 도출
디지털 트랜스포메이션이란 빅데이터, 인공지능, 사물 인터넷(IoT), 클라우드 컴퓨팅 등 디지털 기술과 물리적인 요소들을 결합하여, 기업의 전통적인 운영방식부터 비즈니스 모델, 또는 서비스 모델 등 고객에게 가치를 제공하는 방식에 이르기까지, 경영의 제반 분야를 근본적으로 변화시키는 것을 의미

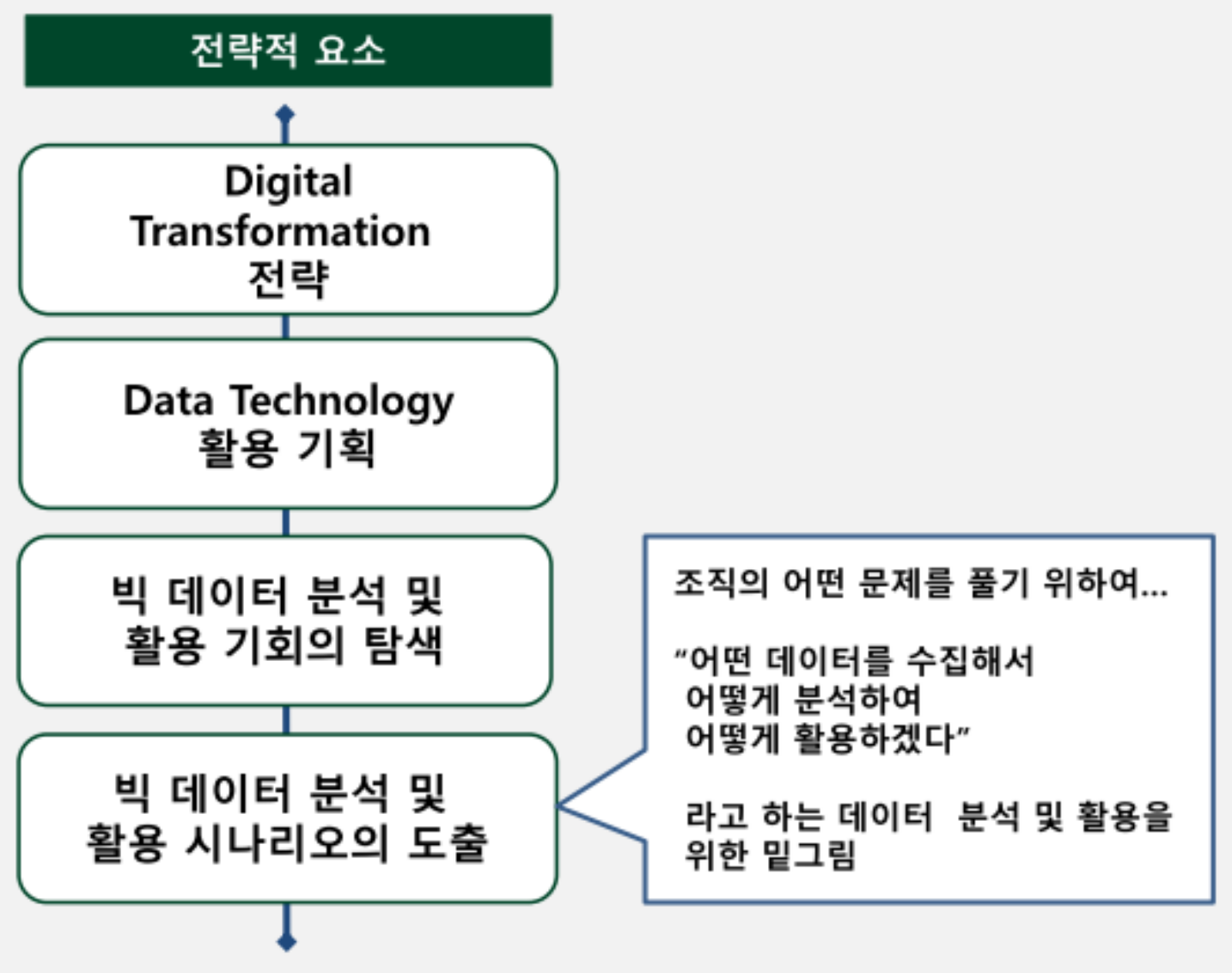
빅데이터의 주요 요소


전략적 요소
Digital Transformation 전략
Data Technology 활용 기획
빅 데이터 분석 및 활용 기회의 탐색
빅 데이터 분석 및 활용 시나리오의 도출
조직의 문제해결을 위하여 데이터 분석 및 활용의 목표와 방법을 정의하고, 이의 실행 과정에 필요한 제반 자원에 대한 확보 계획을 수립하는 것

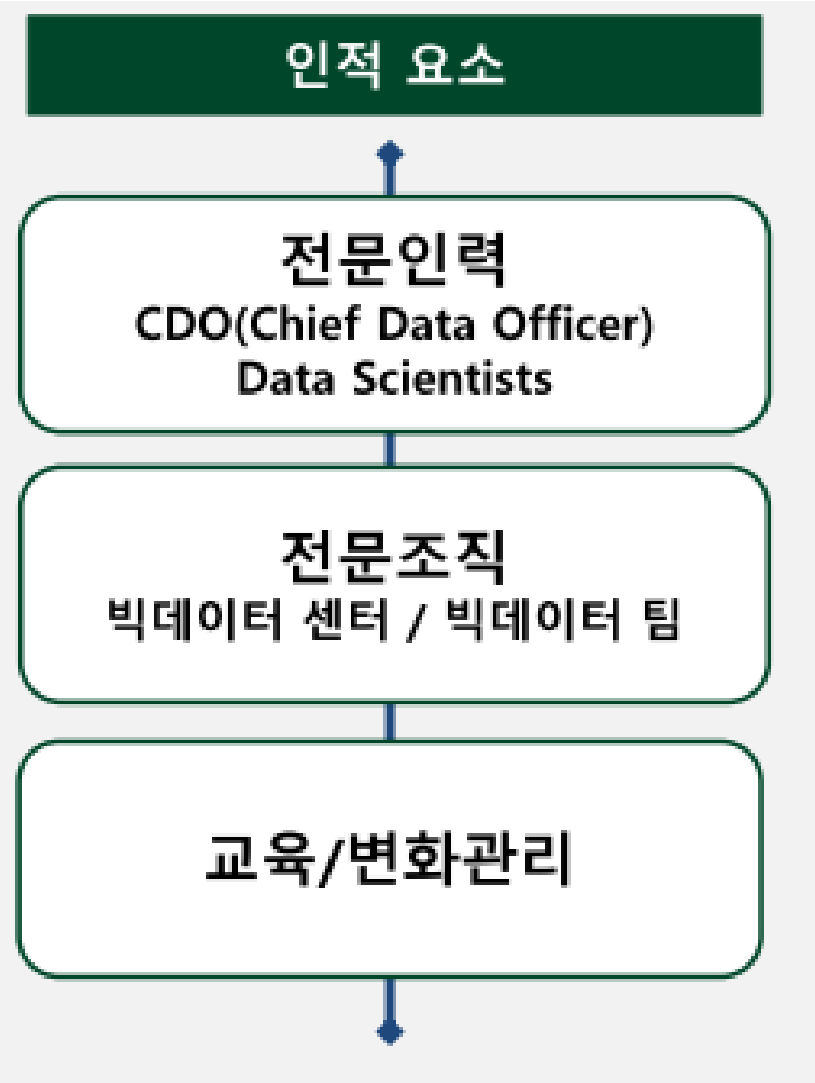
빅데이터의 주요 요소


전략적 요소
Digital Transformation 전략
Data Technology 활용 기획
빅 데이터 분석 및 활용 기회의 탐색
빅 데이터 분석 및 활용 시나리오의 도출
조직의 어떤 문제 해결에 빅데이터를 활용하고자 하는지를 찾아내는 일

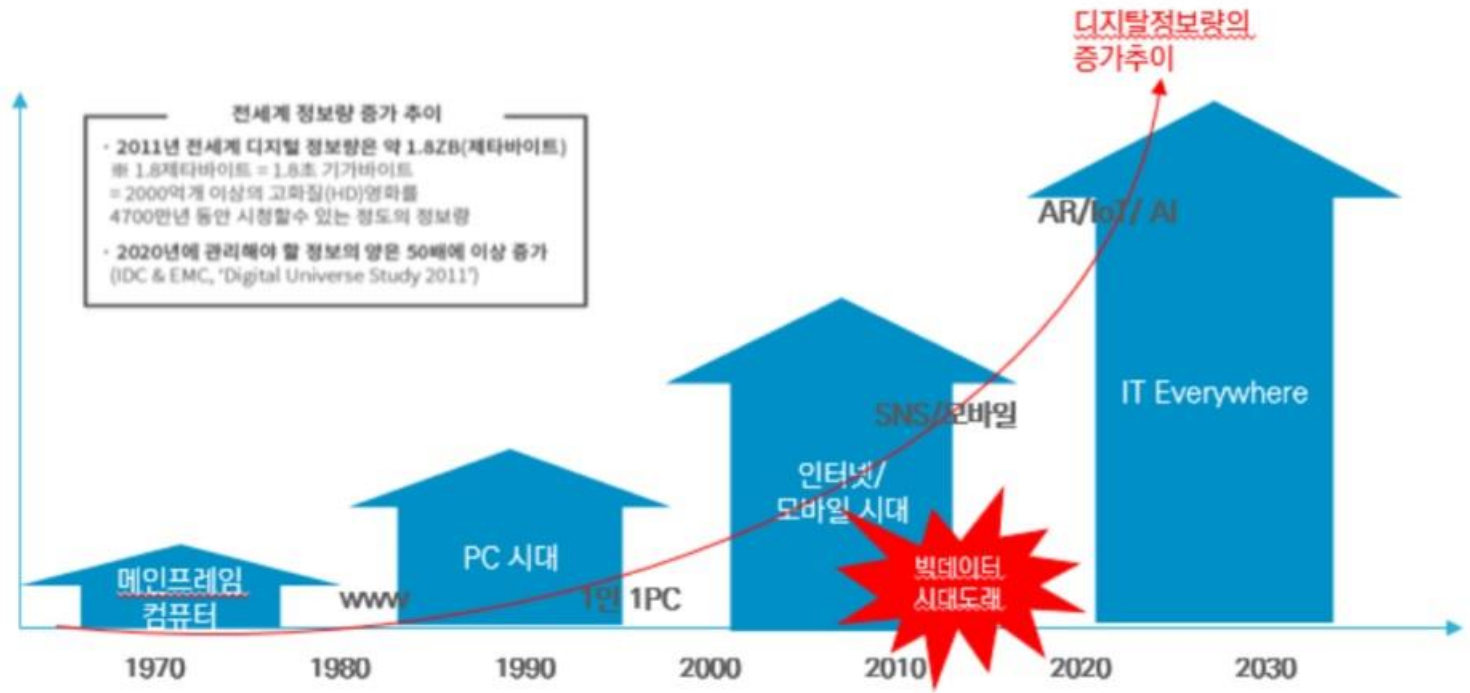
빅데이터의 주요 요소


전략적 요소
Digital Transformation 전략
Data Technology 활용 기획
빅 데이터 분석 및 활용 기회의 탐색
빅 데이터 분석 및 활용 시나리오의 도출
조직의 어떤 문제를 풀기 위하여...
"어떤 데이터를 수집해서 어떻게 분석하여 어떻게 활용하겠다"
라고 하는 데이터 분석 및 활용을 위한 밑그림

빅데이터의 주요 요소


인적 요소
전문인력
CDO(Chief Data Officer)
Data Scientists
전문조직
빅데이터 센터 / 빅데이터 팀
교육/변화관리

빅데이터 와 인공지능



데이터 중심 사회

데이터 규모	EB(Exa Byte) 90년대 말 100EB	ZB(Zeta Byte) 진입 2011년 1.8ZB	ZB(Zeta Byte) 본격화 2020년 은 2011년의 5배
데이터 유형	정형 데이터 (데이터베이스, 사무정보)	비정형 데이터 (이메일,멀티미디어,SNS)	사물정보,인지정보 (RFID,Sensor,IoT)
데이터 특징	구조화	다양성,복합성,소셜	현실성, 실시간성

출 처 : www.comworld.co.kr/news/articleView.html?idxno=49531

빅데이터의 기능

- 빅데이터 거는 기대를 표현한 잘 표현한 비유

산업혁명의 석탄/철	제조업 뿐 아니라 서비스 분야의 생산성을 획기적으로 끌어올려 사회/경제/문화 전반에 혁명적 변화를 가져올 것으로 기대
21세기 원유	경제 성장에 필요한 정보를 제공함으로써 생산성을 한 단계 향상시키고 기존에 없던 새로운 범주의 산업을 만들어 낼 것으로 전망
렌즈	렌즈를 통해 현미경이 생물학에 미쳤던 영향만큼 데이터가 산업 발전에 영향을 미칠 것
플랫폼	다양한 서드파티 비즈니스에 활용되면서 플랫폼 역할을 할 것으로 전망 ex) kakao , facebook

빅데이터가 만들어 내는 본질적인 변화

- **사전처리에서 사후처리 시대로 :** 필요한 정보만 수집하고 필요하지 않은 정보는 버리는 시스템에서 가능한 한 많은 데이터를 모으고 그 데이터를 다양한 방식으로 조합해 숨은 정보를 찾아낸다.

- **표본조사에서 전수조사로** : 표본을 조사하는 기존의 지식 발견 방식이 데이터 수집 비용의 감소와 클라우드 컴퓨팅 기술의 발전으로 인해 전수조사로 변화하게 된다. 이에 따라 샘플링이 주지 못하는 패턴이나 정보를 찾을 수 있게 된다. ★

- **질보다 양으로** : 데이터가 지속적으로 추가될 때 양질의 정보가 오류보다 많아져 전체적으로 좋은 결과 산출에 긍정적인 영향을 미친다는 추론에 바탕을 두고 변화된다.

- **인과관계에서 상관관계로** : 상관관계를 통해 특정 현상의 발생 가능성이 포착되고, 그에 상응하는 행동을 하도록 추천되는 일이 점점 늘어나 데이터 기반의 상관관계 분석이 주는 인사이트가 인과관계에 의해 미래 예측을 점점 더 압도해 가는 시대가 도래하게 될 것으로 전망된다.

빅데이터의 가치

- **데이터 활용 방식** : 빅데이터의 재사용이나 재조합, 다목적용 데이터 개발 등이 일반화되면서 특정 데이터를 누가, 언제,어떻게, 어디서 활용하는지 알 수 없게 되었기 때문에 가치 선정이 어려움.

- **가치 창출 방식** : 기존에 없던 새로운 가치를 창출함에 따라 그 가치를 선정하기 어려움.

- **분석 기술의 발전** : 데이터 분석 기술의 발전으로 가치 있는 데이터와 가치 없는 데이터의 경계를 나누기 어려움. 오늘의 가치 없는 데이터가 내일은 가치 있는 데이터가 될 수 있기에 빅데이터의 가치 산정이 어려움.

빅데이터를 활용한 기본 테크닉

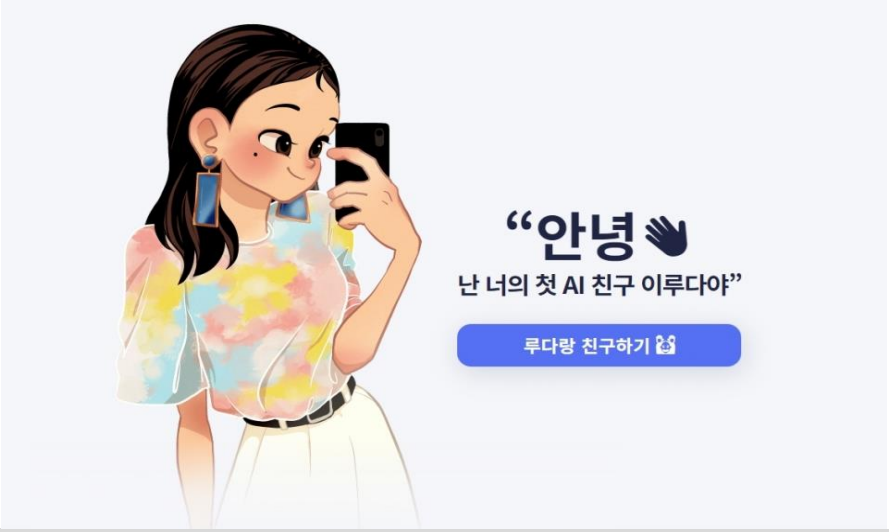
- **연관 규칙 학습** : 변인들 간에 주목할 만한 상관관계 있는지를 찾아내는 방법
- **유형 분석** : 문서를 분류하거나 조직을 그룹으로 나눌 때, 또는 온라인 수강생들을 특성에 따라 분류할 때 사용
- **유전자 알고리즘** :최적화가 필요한 문제의 해결책을 자연선택, 돌연변이 등과 같은 메커니즘을 통해 점진적으로 진화(evolve) 시켜 나가는 방법
- **기계학습** : 훈련 데이터로 부터 학습한 알려진 특성을 활용해 예측하는 방법
- **감정 분석** : 특정 주제에 대해 말하거나 글을 쓴 사람의 감정을 분석
- **소셜 네트워크 분석(=사회관계망분석)** : 특정인과 다른 사람이 몇 촌 정도의 관계인가를 파악할 때 사용하고, 영향력 있는 사람을 찾아낼 때 사용

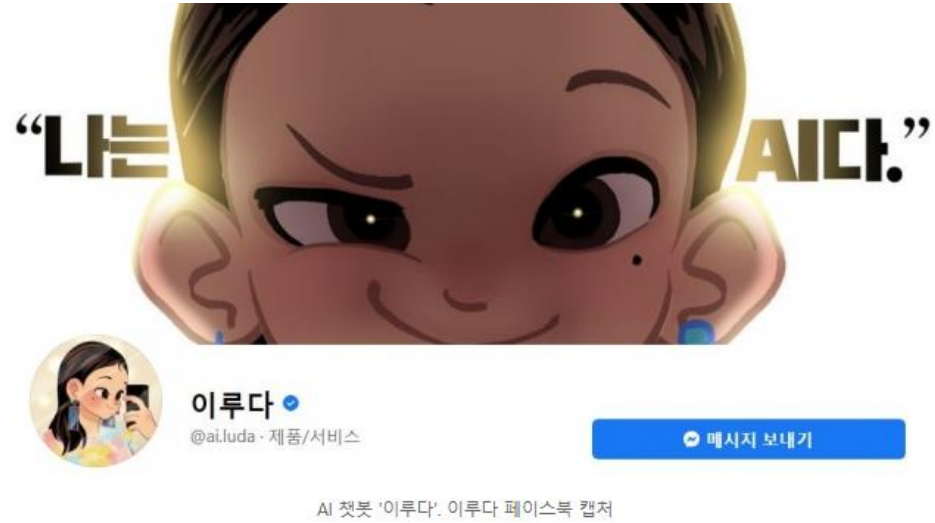
빅 데이터 시대의 위기요인

위기요인 – 빅데이터의 시대가 진행되면서 사생활 침해, 책임원칙 훼손, 데이터 오용 등의 어두운 면 있음

가. 사생활 침해

- M2M(Machine to Machine), IoT(Internet of Things) 시대가 본격화 되면서 정보 수집 센서들의 수가 늘어나고 있음
- 개인 정보의 가치 증대로 많은 기업이 개인정보 습득에 많은 자원 투자
- 사생활 침해 방지를 위해 익명화(anonymization) 기술이 발전하나 충분하지 못하는 의견 다수
- 2013년 정치 스캔들인 미국 NSA(National Security Agency)의 이메일, 전화통화, 문자메시지 등을 수집, 저장한 사건은 대표적인 정부의 사생활 침해 사건





AI 챗봇 '이루다'. 이루다 페이스북 캡처

출 처 : needjarvis.tistory.com/473

빅 데이터 시대의 위기요인

나. 책임 원칙의 훼손

- 빅데이터의 분석 정확도가 증가한 만큼, 분석 대상의 사람들이 예측 알고리즘의 희생양이 될 가능성 존재
- 범죄를 저지르지 않았는데, 저지를 가능성으로 체포를 하는 '마이너리티 리포트' 와 같은 사건이 발생할 수 있음
- 대출 거절, 직원 해고, 환자 수술 거절, 이혼 등 행위 결과가 아닌 아직 일어나지 않는 예측으로 피해를 볼 가능성 존재

출 처 : needjarvis.tistory.com/473

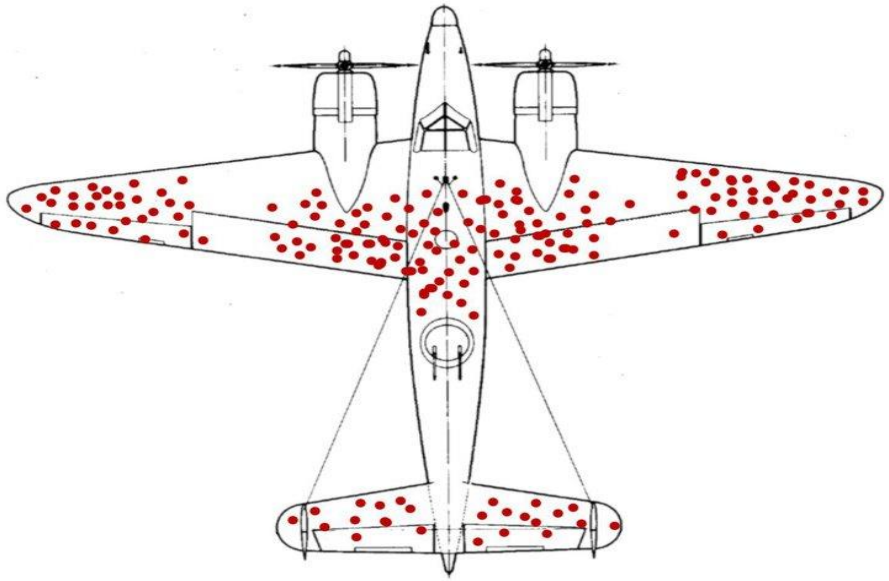
빅 데이터 시대의 위기요인

다. 데이터 오용

- 빅데이터 활용자가 데이터를 과신할 때 문제 발생
- 빅데이터는 과거를 분석하는 것이기 때문에 창조적인 미래를 분석하기 힘듦.

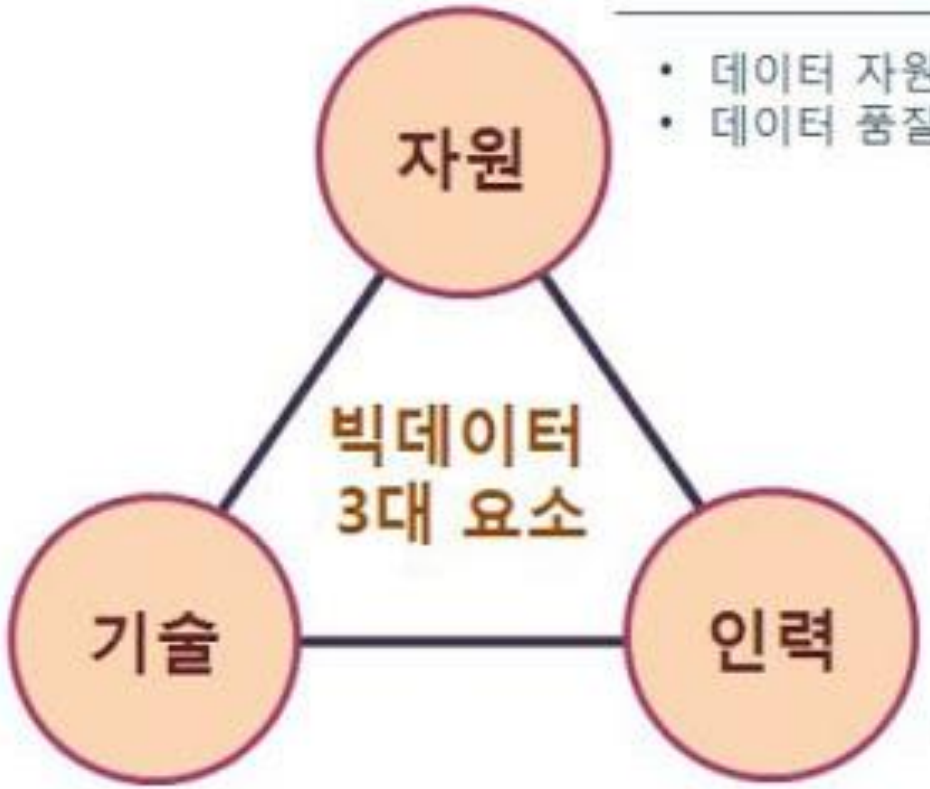
  스티브 잡스(Steve Jobs)는 새로운 제품을 개발할 경우에는 사람들의 의견을 묻지 않음
- 잘못된 지표를 사용하는 것도 빅데이터의 폐해가 될 수 있음.

  맥나마라(Robert McNamara) 장군은 베트남 전쟁 때 적군 사망자 수를 지표로 삼아 전쟁 상황을 오보하는 결과를 가져옴



출 처 : needjarvis.tistory.com/473

빅데이터의 주요 요소 3가지



빅 데이터(Big Data)

- 데이터 자원 확보
- 데이터 품질 관리

빅데이터 플랫폼 (Big Data Platform)

- 데이터 저장, 관리 기술(NoSQL, ETL..)
- 대용량 데이터 처리 (Hadoop, MapReduce..)
- 빅데이터 분석(자연어 처리, 의미분석, 데이터 마이닝..)
- 시각화(Visualization)

데이터 사이언티스트 (Data Scientist)

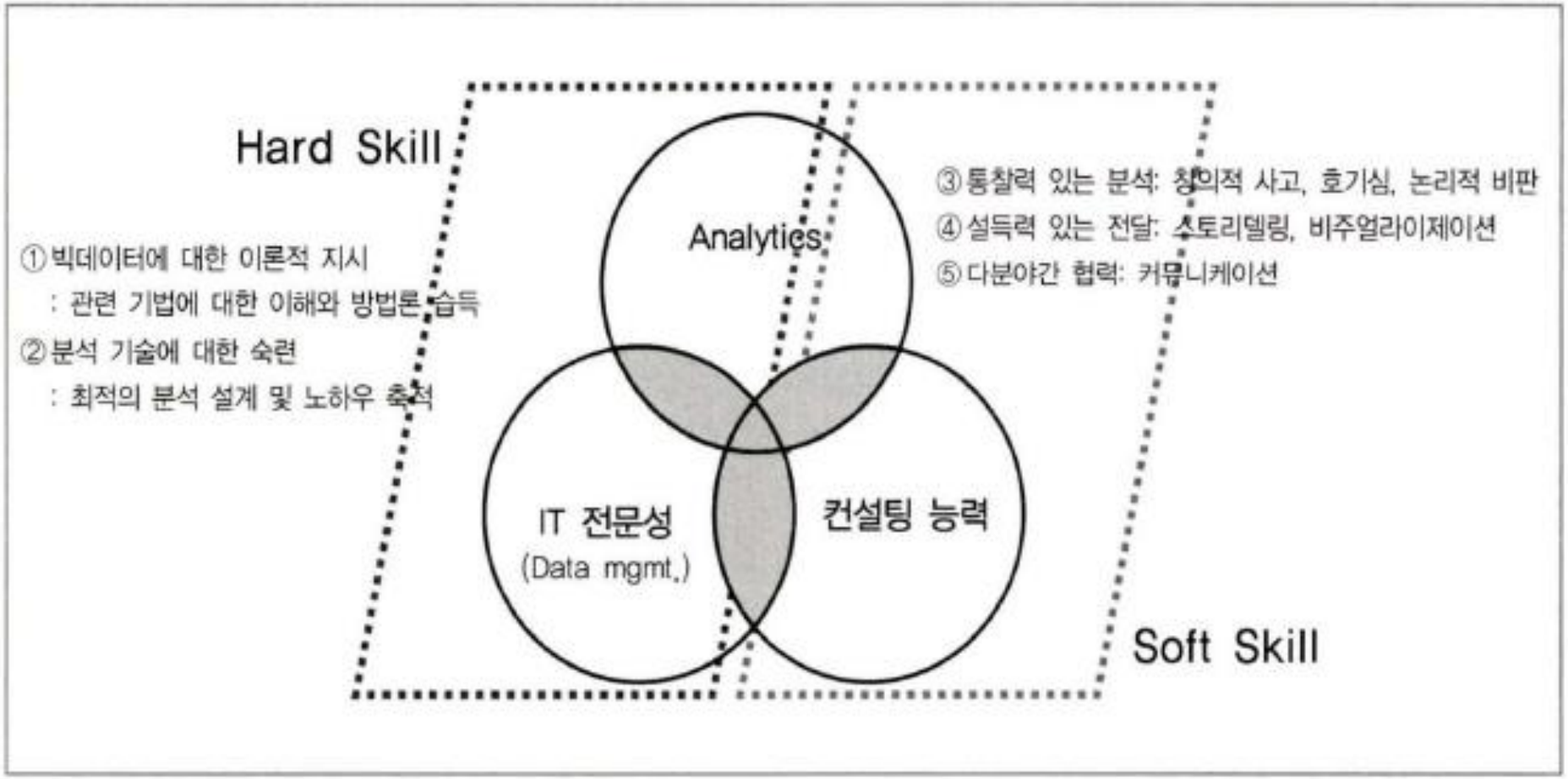
- 수학, 공학(IT기술과 엔지니어링) 능력
- 경제학, 통계학, 심리학 등 다문학적 이해
- 비판적 시각과 커뮤니케이션 능력
- 스토리텔링 등 시각화 능력

출처 : i-bada.blogspot.com/2012/04/blog-post_3023.html

데이터 과학이란?

출처 : news.joins.com/article/23112392

데이터 사이언티스트



[그림 Ⅰ-3-7] 데이터 사이언티스트의 요구역량

출처 : https://m.blog.naver.com/PostView.naver?isHttpsRedirect=true&blogId=jdhpuppy&logNo=221342976923

인공지능의 장점과 폐해

- 인공지능의 기술은 장점으로 작동을 많이 하고 있지만, 음의 방향으로 쓰이기도 한다.
- 최근 인공지능 기술은 데이터 편향성, 알고리즘 차별성, 기술 오남용, 개인 정보 침해, AI 윤리 문제 등 문제점이 지속적으로 나타나고 있다.

컴퓨터 비전의 활용 사례

위기 요인에 따른 통제 방안

미국 연방거래위원회(FTC)는 사생활 침해와 관련해 '**소비자 프라이버시 보호 3대 권고 사항**'을 발표했음

- 기업은 상품 개발 단계에서부터 소비자 프라이버시 보호 방안을 적용(Private by Design)
- 기업은 소비자에게 공유 정보 선택 옵션 제공(Simplified Choice for Business and Consumers)
- 비자에게 수집된 정보 내용 공개 및 접근권 부여(Greater Transparency)
- 일종의 가이드라인으로 사생활 침해를 해결하기에 부족한 측면 존재*

동의에서 책임으로 – 동의제를 책임제로 전환

결과 기반 책임 원칙 고수 – 민주주의 사회에서는 성향에 따라 처벌하는 것이 아니라 '행동 결과'를 보고 처벌했음

알고리즘 접근 허용 – 데이터 오용의 위기 요소의 대응책으로 알고리즘에 대한 접근권 제공이 중요한 이슈로 부상
알고리즘을 통해 불이익을 당한 사람들을 대변할 알고리즈미스트라는 전문가 필요.

*알고리즈미스트 : 알고리즘코딩 해석을 통해 빅데이터 알고리즘에 의해 부당하게 피해를 입은 사람을 구제하고, 데이터 사이언티스트가 한 일로 인한 부당 피해를 막는 역할을 하는 전문 인력

출 처 : needjarvis.tistory.com/473

DBMS

가. DBMS

가) DBMS 는 Data Base Management System의 약자로서 데이터베이스를 관리하여 응용 프로그램들이 데이터베이스를 공유하며 사용할 수 있는 환경을 제공하는 소프트웨어

나. 데이터베이스 관리 시스템의 종류

가) 관계형 DBMS

- 이 모델을 데이터를 열(Column)과 행(Row)를 이루는 하나 이상의 테이블을 정리하여, 고유키(Primary Key)가 각 로우를 식별
- 행은 레코드나 튜플로 부르며, 일반적으로 각 테이블/ 관계는 하나의 엔티티 타입(고객이나 제품)을 대표함

나) 객체지향 DBMS

- 객체 지향 DB는 일반적으로 사용되는 테이블 기반의 관계형 DB와 다르게 정보를 '객체'형태로 표현하는 데이터베이스 모델

SQL

가. SQL

Structured Query Language의 약자로, 데이터베이스를 사용할 때 데이터베이스에 접근할 수 있는

데이터 베이스의 하부 언어로, 단순한 질의 기능 뿐만 아니라 완전한 데이터의 정의와 조작 기능을 갖추고 있다.

나. SQL 집계함수

함수명	설명	유형별 가능 여부
AVG	지정한 열의 평균 값을 반환	수치형
COUNT	테이블의 특정 조건이 맞는 것의 개수를 반환	수치형, 문자형
SUM	지정한 열의 총합을 반환	수치형
STDDEV	지정한 열의 분산을 반환	수치형
MIN	지정한 열의 가장 작은 값을 반환	수치형
MAX	지정한 열의 가장 큰 값을 반환	수치형

데이터에 관련한 기술

개인정보 비식별 기술

비식별 기술	내용	예시
데이터 마스킹	데이터의 길이, 유형, 형식과 같은 속성을 유지한 채, 새롭게 읽기 쉬운 데이터를 익명으로 생성하는 기술	홍길동 35세, 서울 거주, 한국대 재학 -〉 홍**,35세,서울 거주, **대학 재학
가명처리	개인정보 주체의 이름을 다른 이름으로 변경하는 기술, 다른 값으로 대체할 시 일정한 규칙이 노출되지 않도록 주의해야 함	홍길동,35세, 서울거주, 한국대 재학 -〉 임꺽정,30대, 서울 거주, 국내대 재학
총계 처리	데이터의 총합 값을 보임으로서 개별 데이터의 값을 보이지 않도록 함	임꺽정 180cm, 홍길동 170cm, 이콩쥐 160cm, 김팥쥐 150cm -〉물리학과 학생 키 합 :660cm, 평균 키 :165cm
데이터값 삭제	데이터 공유, 개방 목적에 따라 데이터 셋에 구성된 값 중에 필요 없는 값 또는 개인식별에 중요한 값을 삭제	홍길동 35세, 서울 거주, 한국대 졸업 -〉35세, 서울 거주
데이터 범주화	데이터의 값을 범주의 값으로 변환하여 값을 숨김	홍길동,35세 -〉 홍씨,30~40대

데이터에 관련한 기술

개인정보 비식별 기술

Data에 관련한 기술

	내용	유형
데이터 무결성 (Data integrity)	• 데이터 베이스 내의 데이터에 대한 정확한 일관성, 유효성, 신뢰성을 보장하기 위해 데이터 변경/수정 시 여러 가지 제한을 두어 데이터의 정확성을 보증하는 것을 의미.	개체 무결성(Entity integrity) 참조 무결성(Referential integrity) 범위 무결성(Domain in-tegrity)
데이터 레이크 (Data Lake)	• 수 많은 정보 속에서 의미 있는 내용을 찾기 위해 방식에 상관없이 데이터를 저장하는 시스템으로, 대용량의 정형 및 비정형 데이터를 저장할 뿐만 아니라 접근도 쉽게 할 수 있는 대규모의 저장소를 의미.	Apache Hadoop, Teradata Integrated Big Data Platform 1700 등과 같은 플랫폼으로 구성된 솔루션으로 제공.

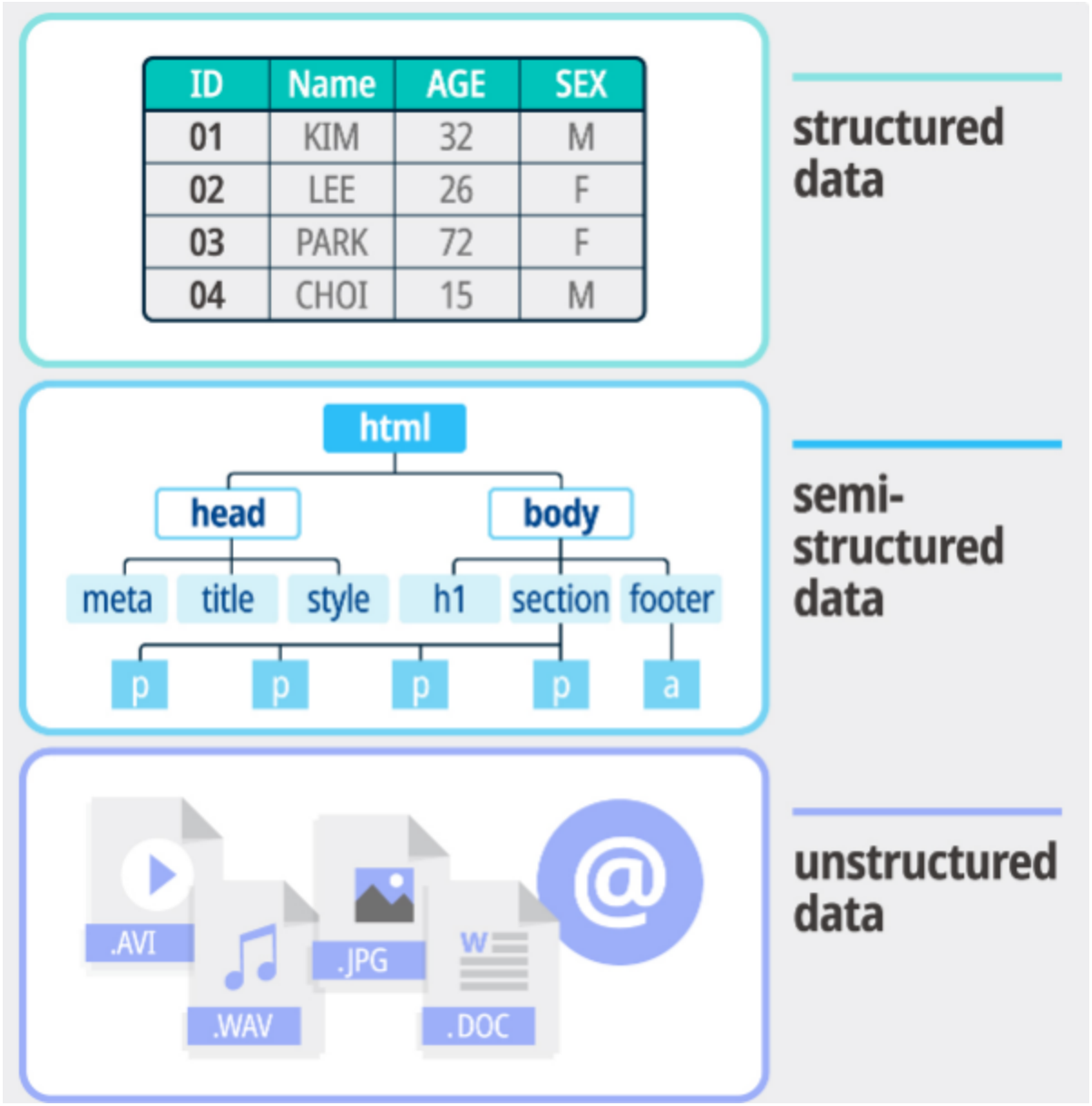
데이터의 유형 복습

유형	내용	예시
정형 데이터	• 정형화된 스키마 구조. 주로 관계형 데이터 (RDBMS)에 저장됨 • 데이터 수집 난이도가 낮고, 형식이 정해져 있어 처리가 쉬운 편	관계형 데이터 베이스, 스프레드 시트, CSV 등
반정형 데이터	• 데이터 내부의 데이터 구조에 대한 메타 정보가 포함된 구조 • 고정된 필드에 저장되어 있지는 않지만, 메타데이터나 데이터 스키마 정보를 포함하는 데이터	XML, HTML, JSON, 로그 형태 (웹 로그, 센서 데이터) 등
비정형 데이터	• 고정 필드 및 메타데이터(스키마 포함)가 정의되지 않음. • 데이터 수집 난이도가 높으며, 텍스트 마이닝 혹은 파일일 경우 파일을 데이터 형태로 파싱해야 하기 때문에 수집 데이터 처리가 어려움	소셜 데이터(트위터, 페이스북), 영상, 이미지, 음성, 텍스트(word, pdf …) 등

*스키마(Schema) : 데이터베이스에서 자료의 구조, 자료의 표현 방법,. 자료 간의 관계를 형식 언어로 정의한 구조.

*메타데이터(Metadata) : 데이터에 관한 구조화된 데이터로, 다른 데이터를 설명해 주는 데이터를 말함.

데이터의 유형 복습




```
{
  "이름" : "오형준",
  "나이" : 23,
  "성별" : "남"
}
```

(a) JSON

```
<친구정보>
  <이름> 오형준 </이름>
  <나이> 23 </나이>
  <성별> 남 </성별>
</친구정보>
```

(b) XML

그림 1-8 반정형 데이터의 예

출처 : https://do-my-best.tistory.com/정형-반정형-비정형-데이터란

데이터의 유형 복습

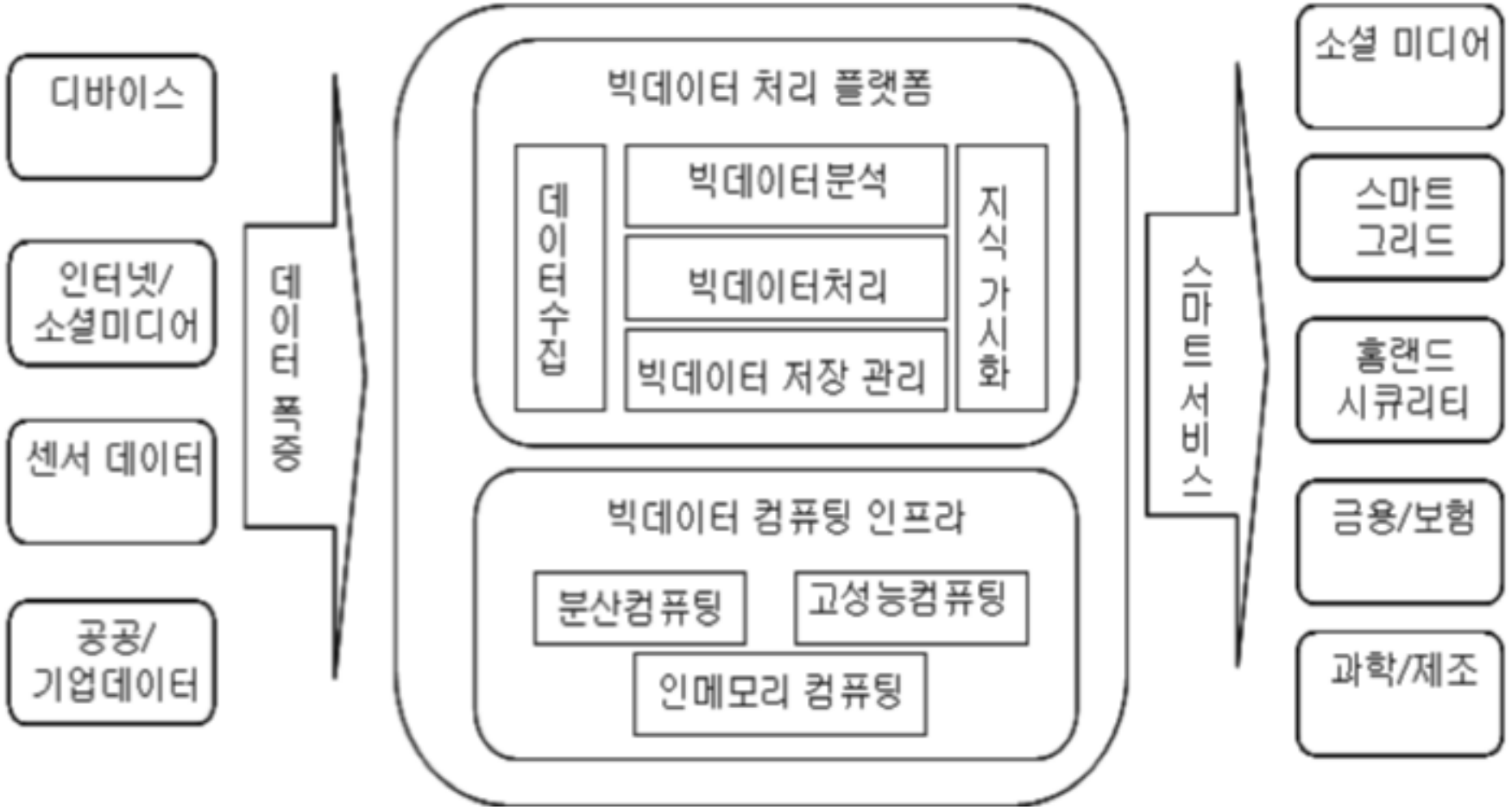
가치	설명
경제적 자산	• 새로운 가치를 창출하고, 위험을 해결하여 사회 및 경제 발전의 엔진 역할을 수행
불확실성 제거	• 사회현상, 현실 세계의 데이터를 기반으로 한 패턴 분석과 미래 전망 • 여러 가지 가능성에 대한 시나리오 시뮬레이션
리스크 감소	• 환경, 소셜, 모니터링 정보의 패턴 분석을 통해 위험 징후 및 이상 신호 포착 • 이슈를 사전에 인지 및 분석하고 빠른 의사 결정과 실시간 대응
스마트한 경쟁력	• 대규모 데이터 분석을 통한 상황 인지, 인공지능 서비스 기능 • 개인화, 지능화 서비스 제공 확대 • 트렌드 변화 분석을 통한 제품 경쟁력 확보
타 분야 융합	• 타 분야와의 융합을 통한 새로운 가치 창출 • 방대한 데이터 활용을 통한 새로운 융합시장 창출

빅데이터 가치 선정이 어려운 이유

가치	설명
데이터 활용 방식의 다양화	• 데이터의 재사용, 데이터의 재조합, 다목적용 데이터 개발 등이 일반화되면서 특정 데이터를 언제/어디서/누가 활용할지 알 수 없어서 가치 선정이 어려움 • 데이터의 창의적 조합으로 인해 기존에 풀 수 없는 문제를 해결하는 데 도움을 주기 때문에 가치 선정이 어려움 • 예) 구글이 검색 결과를 낼 때마다 구글은 클라우드에 저장된 웹 사이트 정보를 매번 사용.
새로운 가치 창출	• 빅데이터 시대에 데이터가 기존에 없던 가치를 창출하여 가치를 선정이 어려움. • 예) 고객의 선향을 분석하여 고객 맞춤 서비스 제공
분석기술의 급속한 발전	• 비용 문제로 인해 분석할 수 없었던 것을 저렴한 비용으로 분석하면서 활용도가 증가하여 가치 선정이 어려움. • 예) 텍스트 마이닝을 통한 SNS 분석

빅데이터 플랫폼

- 빅데이터에서 가치를 추출하기 위해 일련의 과정(수집→저장 →분석 →시각화)을 규격화한 기술.
- 특화된 분석(의료, 환경, 범죄, 자동차 등)을 지원하는 빅데이터 플랫폼이 발전하는 추세이다.



연습문제

1. 데이터에 대한 설명으로 가장 부적절한 것은 무엇인가?

① 데이터를 단순한 객체로서 가치 뿐만 아니라, 다른 객체와의 상호관계 속에서 가치를 갖는 것으로 설명할 수 있다.

② 데이터는 그 형태에 따라 언어,문자 등으로 기술되는 정량적 데이터와 수치, 기호, 도형으로 표시되는 정성적 데이터로 구분된다.

③ 설문조사와 주관식 응답, 트위터나 페이스북, 블로그 등에 올린 글 등과 같은 정성 데이터의 경우 그 형태와 형식이 정해져 있지 않아, 비정형 데이터라고 한다.

④ 지역별 온도, 풍속, 강수량과 같이 수치로 명확하게 표현되는 데이터를 정량 데이터라고 한다.

연습문제

2. 다음 중 구조 관점의 데이터 유형 중 아래에서 설명하는 것은 무엇인가?

스키마(형태) 구조 형태를 가지고 XML, HTML, 웹 로그, 시스템 로그, 알람 등과 같이 메타데이터를 포함하며, 값과 형식에서 일관성을 가지지 않는 데이터

① 정형데이터

② 비정형데이터

③ 반정형 데이터

④ 스트림 데이터

연습문제

3. 다양한 데이터 유형 중 정형 데이터–반정형 데이터–비정형데이터 순서로 가장 알맞은 것은?

① 인스타그램 게시물 - 기상철 날씨 데이터 - 웹 로그 데이터

② 물류창고제고 데이터 - XML - 이메일 전송 데이터

③ CRM 데이터 - 카카오톡 대화 데이터 - Twitter 상태 메세지

④ RFID - IOT 센서 데이터 - 동영상 데이터

연습문제

4. 다음은 DIKW(Data, Information, Knowledge, Wisdom hierarchy)에 대한 설명이다.

가장 적절한 설명은 무엇인가?

① 지식(Knowledge)은 근본 원리에 대한 깊은 이해를 바탕으로 도출되는 창의적 아이디어로 설명할 수 있다.

② 정보(Information)는 상호 연결된 정보 패턴을 이해하고, 이를 토대로 예측한 결과물이라 할 수 있다.

③ 지혜(Wisdom)는 데이터의 가공 및 상관관계 간 이해를 통해 패턴을 인식하고, 그 의미를 부여한 데이터라고 할 수 있다.

④ 데이터(Data)는 존재형식을 불문하고 타 데이터와의 상관관계가 없는 가공하기 전의 순수한 수치나 기호를 의미한다고 할 수 있다.

연습문제

5. 러셀 L.액오프가 1989년에 이야기 한 DIKW Hierarchy는 데이터가 어떻게 진화하는 지를 단계적으로 설명하였다. 다음 DIKW 단계를 설명하는 것 중 다른 하나는 무엇인가?

① 지난 1년 매출액의 50%는 8월에 집중되어 있다.

② 지난 1년 매출은 1월에서 8월까지 증가하였고, 12월까지 다시 증가하였다.

③ 날씨가 따뜻해지고, 지점을 확장하여 올 8월 매출액은 3000만원으로 예상한다.

④ 8월 A상품 구매 고객의 80%가 40대 여성 고객으로 대부분 회사원이다.

연습문제

6. 다음은 데이터베이스의 일반적인 특징에 관한 설명이다. 가장 부적절한 것은 무엇인가?

① 데이터베이스는 통합된 데이터(integrated data)이다.

② 데이터베이스는 저장된 데이터(stored data)이다.

③ 데이터베이스는 공용 데이터(shared data)이다.

④ 데이터베이스는 변화되지 않는 데이터(unchanged data)이다.

연습문제

7. 다음은 데이터베이스의 특성에 관한 설명이다. 가장 부적절한 것을 모두 고르시오.

① 정보의 축적 및 전달 측면에서 대량의 정보를 일정한 형식에 따라 정보처리기기가 읽고 쓰고 검색할 수 있도록 하는 기계가독성과 검색가능성, 그리고 정보 통신망을 통하여 원거리에서도 즉시 온라인으로 이용할 수 있는 원격 조작성을 갖는다.

② 정보 관리 측면에서는 이용자의 정보 요구에 따라 다양한 정보를 신속하게 획득할 수 있고, 원하는 정보를 정확하고 경제적으로 찾아낼 수 있다는 특성을 지닌다.

③ 정보 이용 측면에서는 정보를 일정한 질서와 구조에 따라 정리, 저장하고 검색, 관리할 수 있도록 하여 정보를 체계적으로 축적하고 새로운 내용의 추가나 갱신이 용이하다.

④ 정보기술 발전의 측면에서 정보처리, 검색, 관리 소프트웨어, 관련 하드웨어, 정보 전송을 위한 네트워크 기술 등의 발전을 견인 할 수 있다.

연습문제

8. 다음 중 개인정보 비식별화 기법을 설명한 것으로 가장 부적절한 것은?

① 총계 처리– 데이터의 총합 값을 보임으로써 개별 데이터의 값을 보이지 않도록 함.

② 데이터 마스킹 – 개인 식별에 중요한 데이터 값을 삭제

③ 가명 처리 – 개인 식별에 중요한 데이터를 식별 할 수 없는 다른 값으로 변경

④ 범주화 – 데이터의 값을 범주의 값으로 변환하여 값을 추출

연습문제

9. 다음 중 빅데이터가 기업에게 주는 가치가 아닌 것은 무엇인가?

① 혁신 수단 제공

② 경쟁력 강화

③ 생산성 제고

④ 환경 탐색

연습문제

10. 다음 중 빅데이터가 만들어 내는 변화가 아닌 것은?

① 데이터의 질보다 양에 비중을 둠

② 데이터의 사전 처리보다 사후 처리에 비중을 둠

③ 새로운 것에 대한 발견법으로 상관관계보다 인과관계에 비중을 둠

④ 조사 방법으로서 표본조사보다 전수조사에 비중을 둠

연습문제

11. 빅데이터 출현 배경 중 거대한 데이터의 분석 비용 문제를 해결해 준 것은 무엇인가?

① 디지털 기술

② 클라우드 컴퓨팅 기술

③ 하드 드라이브 가격의 하락

④ SNS 확산

연습문제

12. 커피를 사는 사람들이 탄산음료도 많이 구매하는 지를 알아보기 위해 사용되는 분석은?

① 회귀 분석(Regression Analysis)

② 기계 학습(Machine learning)

③ 유전 알고리즘(Genetic algorithm)

④ 연관 규칙 학습(association rule learning)

연습문제

13. 빅데이터 활용에 필요한 기본적인 3요소로 가장 적절한 것은?

① 데이터, 기술, 인력

② 데이터, 기술, 프로세스

③ 기술, 인력, 프로세스

④ 데이터, 인력, 프로세스

연습문제

14. 다음 중 빅데이터 출현 배경에 관한 설명으로 부적절한 것은?

① 개별 기업의 데이터 축적 및 데이터 활용에 대한 니즈 증가

② 데이터 저장 기술 발전과 저장 비용 감소

③ 인터넷, SNS와 사물네트워크의 확산으로 데이터 생산량 증가

④ 수집 관리 및 분석에 용이한 형태로 데이터 구조의 정형화

연습문제

15. 다음 중 빅데이터의 수집, 구축, 분석의 최종 목적으로 가장 적절한 것은?

① 새로운 통찰과 가치를 창출

② 데이터 중심 조직 구성

③ 초고속 데이터 처리 기술 개발

④ 데이터 관리 비용 절감

연습문제

16. 다음 중 데이터의 가치 측정이 어려운 이유로 적절하지 않은 것은 무엇인가?

① 데이터 재사용의 일반화로 특정 데이터를 언제 누가 사용했는지 알기 힘들기 때문이다.

② 빅데이터 전문 인력의 증가로 다양한 곳에서 빅데이터가 활용되고 있기 때문이다.

③ 분석기술의 발전으로 과거에 분석이 불가능했던 데이터를 분석할 수 있게 되었기 때문이다

④ 빅데이터는 기존에 존재하지 않던 새로운 가치를 창출하기 때문이다.

연습문제

17. 다음 중 빅데이터가 만들어 내는 변화로 가장 부적절한 것은?

① 사전처리에서 사후처리 시대로의 변화

② 대면조사에서 표본조사로의 변화

③ 데이터의 질보다 양의 중요도 증가

④ 인과관계에서 상관관계의 중요도 증가

연습문제

18. 인터넷의 진화는 수많은 센서들이 인터넷으로 연결되는 사물인터넷(IOT) 시대로 나아가고 있다. 미래의 빅데이터 관점으로 볼 때 사물인터넷과 가장 관련이 큰 것은?

① 인공지능(AI)

② 데이터화(Datafication)

③ 스마트 데이터(Smart Data)

④ 지능적 서비스(Intelligence Service)

연습문제

19. 빅데이터 출현 배경과 거리가 먼 것은?

① 소셜 미디어, 영상 등 비정형 데이터의 급격한 확산

② 데이터 처리 기술의 발전

③ 학계의 거대 데이터 활용 과학 확산

④ 정부의 공공데이터 개방 확산

연습문제

20. 아래에서 빅데이터 시대의 위기와 통제에 대한 설명으로 가장 올바르게 묶인 것은?

A) 데이터 익명화(Anonymization)은 사생활 침해에 대한 근본요인을 차단할 수 있어 빠른 기술의 발전이 필요하다.

B) 개인정보 사용자의 정보사용에 대한 책임의 한계로 개인정보 사용 책임제도보다 동의제도를 더욱 강화해야 한다.

C) 민주주의에서 '행동 결과'에 따른 처벌의 모순을 교훈삼아 빅데이터 사전 '성향'분석을 통한 통제의 강화가 필요하다.

D) 빅데이터 분석은 실제 일어난 일에 대한 데이터에 의존하기 때문에 이를 바탕으로 미래를 예측하는 것은 언제나 맞을 수 없는 오류가 존재한다.

E) 알고리즘을 통해 불이익을 당한 사람들을 대변할 알고리즈미스트라는 전문가가 필요하다.

① A,C

② B,C

③ D,E

④ B,E

연습문제

21. 다음 중 사생활 침해를 막기 위해 개인정보를 무작위 처리하는 등 데이터가 본래 목적 외에 가공되고 처리되는 것을 방지하는 기술은 무엇인가?

① 정규화

② 난수화

③ 익명화

④ 일반화

연습문제

22. 아래 빅데이터 활용을 위한 기본 테크닉 중 어떤 사례를 활용하는가?

A 마트는 금요일 저녁에 맥주를 사는 사람은 기저귀도 함께 구매했다는 사실을 발견하고, 두가지 상품을 가까운 곳에 진열하기로 결정했다.

① 회귀분석

② 연관성분석

③ 유형분석

④ 구문분석

연습문제

23. 다음 중 감성 분석에 대한 설명으로 부적절한 것은?

① 특정 주제에 대해 말하거나 글을 쓴 사람의 감정을 분석한다.

② 소셜 미디어에 나타난 의견을 바탕으로 고객이 원하는 것을 찾아낼 때 활용한다.

③ 고객들 간 소셜 네트워크 관계를 파악할 수 있다.

④ 호텔에서 고객의 논평을 받아 서비스를 개선하기 위해 활용한다.

연습문제

24. 다음 설명 중 맞는 것은 무엇 인가?

① 빅데이터 분석에서 가치 창출은 데이터의 크기에 좌우된다.

② 빅데이터 과제의 주된 걸림돌은 비용이 아니라 분석적 방법에 대한 이해 부족이다.

③ 성과가 높은 기업들은 모두 폭넓은 가치 분석적 통찰력을 갖추고 있다.

④ 분석을 다방면에 많이 사용하는 것이 경쟁 우위를 가져다 주는 첫 번째 요소이다.

연습문제

25. 아래 보기에서 데이터의 양의 표시 단위를 작은것부터 큰 순서로 나열하시오.

a. 엑사바이트(EB)

b. 페타바이트(PB)

c. 요타바이트(YB)

d. 제타바이트(ZB)

연습문제

26. 조직이나 기업의 인적 자원이 축적되고 있는 개별적인 지식을 체계화하여 공유함으로써 경쟁력을 향상시키기 위한 기업 정보시스템을 무엇이라 하는가?

연습문제

27. 인터넷으로 연결된 기계마다 통신 장치를 있는 환경에서 사람 또는 기계끼리 자동으로 통신하는 기술로써 사물과 사람, 사물과 사물 간의 상호 소통하는 방식을 무엇이라 하는가?

연습문제

28. 아래 데이터 분석과 관련된 기술을 설명한 것이다. (가)에 들어갈 용어를 기입하시오.

기업의 의사 결정 과정을 지원하기 위한 주제 중심적으로 통합적이며 시간성을 가지는 비휘발성 데이터의 집합을 (가)라고 한다.

연습문제

29. 아래는 (가)라는 데이터의 유형을 설명한 것이다. 데이터 (가)는 무엇인가?

(가) 데이터는 지역별 매출액, 영업이익률, 판매량과 같이 수치로 명확하게 표현되는 데이터로, 그 양이 크게 증가하더라도, 이를 DBMS 에 저장, 검색, 분석하여 활용하기가 용이하다.

연습문제

30. 아래에서 빈칸에 공통적으로 들어갈 용어는?

가) 페이스북은 2006년 F8 행사를 기점으로 자신들의 소설 그래프 자산을 외부 개발자들에게 공개하고 서드파티 개발자들이 페이스북 위에서 작동하는 앱을 만들기 시작하면서 ( )역할을 하기 시작했다.

나) 하둡은 대규모 분산 병렬 처리의 업계 표준으로 맵리듀스 시스템과 분산 파일 시스템인 HDFS로 구성된 ( ) 기술이며, 선형적인 성능과 용량 확장성, 고장 감내성을 가지고 있다. 아마존(Amazon)은 S3와 BC2 환경을 제공함으로써 ( )을(를) 위한 클라우드 서비스를 최초로 실현하였다.

연습문제

31. 아래에서 설명하고 있는 빅데이터 활용 기본 테크닉은 무엇인가?

가) 생명의 진화를 모방하여 최적해(Optimal Solution)를 구하는 알고리즘으로 존 홀랜드(John Holland)가 1975년에 개발하였다.

나) '최대의 시청률을 얻으려면 어떤 시간대에 방송해야 하는가? ' 와 같은 문제를 해결할 때 사용된다.

다) 어떤 미지의 함수 Y=f(x)를 최적화하는 해 x를 찾기 위해, 진화를 모방한(Simulated evolution) 탐색 알고리즘이라고 말할 수 있다.

연습문제

32. 아래의 보기는 데이터의 이용과 분석에 대한 시대별 용어와 의미를 서로 연결한 것이다.

빈칸에 알맞은 용어는?

OLAP - 다차원의 데이터를 대화식으로 분석하기 위한 소프트웨어

( ) - 데이터 기반의 의사결정을 지원하기 위한 리포트 중심의 도구

BA - 의사결정을 위한 통계적이고 수학적인 분석에 초점을 둔 기법

연습문제

33. 아래 빈 칸에 알맞은 데이터베이스 용어는?

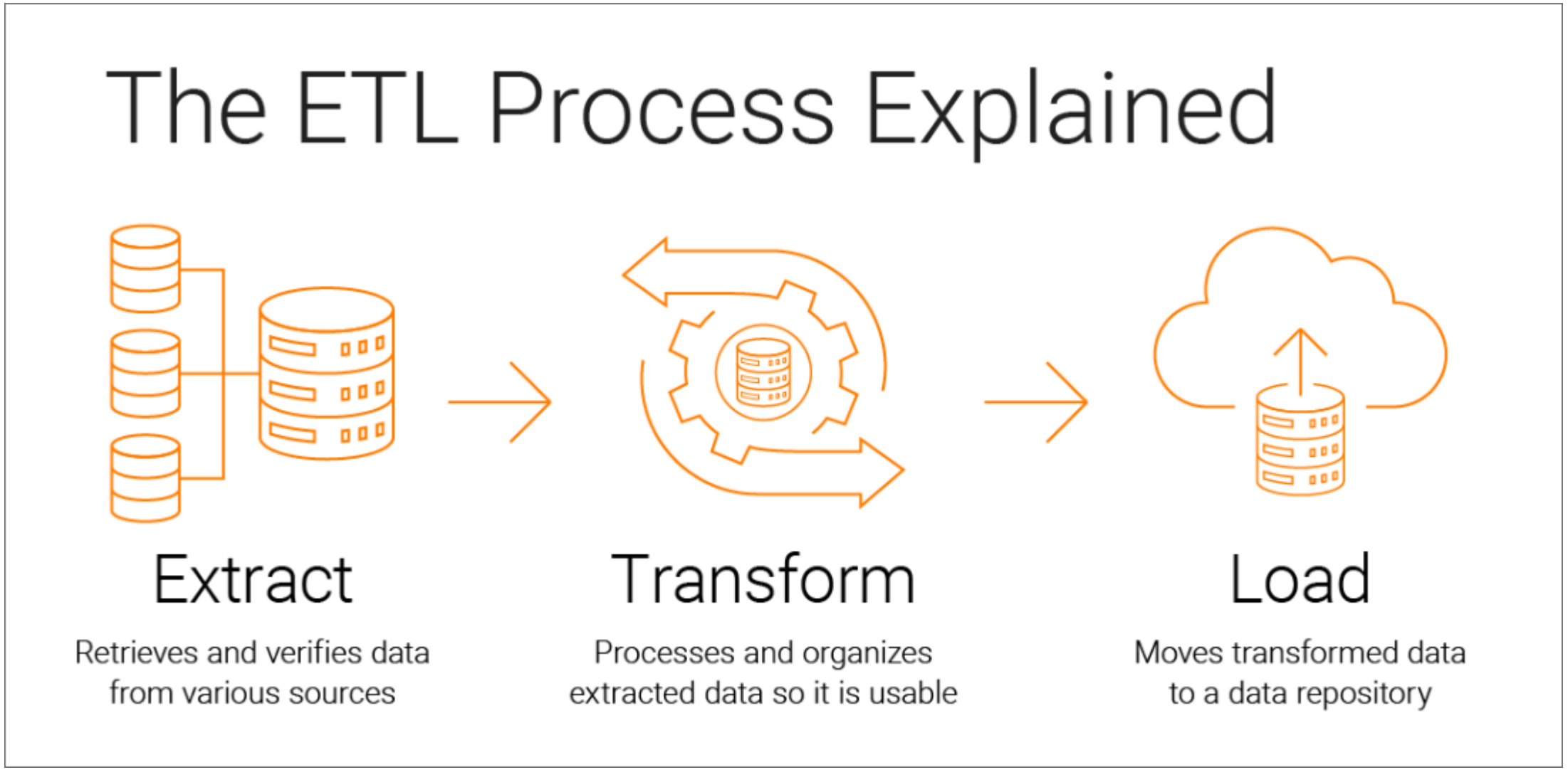
데이터는 ( )는 데이터베이스 내의 데이터에 대한 정확성, 일관성, 유효성, 신뢰성을 보장하기 위해 데이터 변경 혹은 수정시 여러 가지 제한을 두어 데이터의 정확성을 보증하는 것을 말한다.

연습문제

34. 아래 빈 칸에 알맞은 용어는?

( ) 프로세스는 기업 시스템 내의 축적된 데이터 중 A 시스템에 데이터를 추출(Extract), 데이터웨어하우스 등에서 이용하기 쉬운 형태로 변환(Transform)해, 대상이 되는 B 시스템에 적재(Load) 또는 이들 일련의 처리 지원하는 이렇게 처리된 데이터는 시각화(Visualization) 작업을 통해 기업 의사결정권자들을 위한 비즈니스 리포로 만들어 진다.

Extract, Transform and Load(ETL)



출처 : https://www.airops.com/blog/etl-vs-elt


Thank you.
ADSP/ 류영표 강사
ryp1662@gmail.com