1과목.빅데이터 분석 기획

학습 방향 : 1과목 빅데이터 분석 기획은 데이터 전문가라면 기본 소양으로 알고 있어야 하는 내용이 많으며 출제 역시 전반적인 개념과 사례에 대해 물어보는 형태가 주로 출제되고 있다. 각종 절차와 내용, 기술 용어 등의 이해와 함께 빅데이터로 인한 개인정보침해 등 사회적 문제에 대해서도 출제가 자주 되고 있으니 대비 해야 한다.

시험에 관한 안내

01 필기 응시 자격 조건

- 대학 졸업자 및 졸업예정자(전공 무관)
- 그 외 응시자격은 시행처의 자격안내 확인

02 필기 원서 접수하기

- www.dataq.or.kr에서 접수
- 연 2회 시행

03 필기 시험

- 신분증, 컴퓨터용 사인펜, 수험표 지참
- 120분 동안 진행

04 필기 합격자 발표

- www.dataq.or.kr에서 합격자 발표

시험에 관한 안내

1. 빅데이터 분석기사 정의

; 빅데이터 이해를 기반으로 빅데이터 분석 기획, 빅데이터 수집·저장·처리, 빅데이터 분석 및 시각화를 수행하는 실무자를 말한다.

● 대용량의 데이터 집합으로부터 유용한 정보를 찾고 결과를 예측하기 위해 목적에 따라 분석 기술과 방법론을 기반으로 정형/비정형 대용량 데이터를 구축, 탐색, 분석하고 시각화를 수행하는 업무를 수행한다.

시험에 관한 안내

2. 검정 요강

필기과목명	문제수	주요항목
빅데이터 분석기획	20	빅데이터의 이해
		데이터 분석 계획
		데이터 수집 및 저장 계획
빅데이터 탐색	20	데이터 전처리
		데이터 탐색
		통계기법 이해
빅데이터 모델링	20	분석모형 설계
		분석기법 적용
빅데이터 결과 해석	20	분석모형 평가 및 개선
		분석결과 해석 및 활용

3. 합격 기준

과목당 100점 만점으로

① 각 과목별 40점 이상

② 전 과목 평균 60점 이상

빅데이터 분석 기사(1과목. 빅데이터 분석 기획)

CHAPTER 1. 빅데이터의 이해
CHAPTER 2. 데이터 분석 계획
CHAPTER 3. 데이터 수집 및 저장 계획

빅데이터 이해

빅데이터 이해 챕터는 총 2개의 작은 섹션으로 구성된다.

1. 빅데이터 개요 및 활용

2. 빅데이터 기술 및 제도

1. 빅데이터의 이해 - 빅데이터 개요 및 활용

01 데이터와 정보

1) 데이터와 정보

; 데이터는 1646년 영국 문헌에 처음 등장하였으며, '주어진 것'이란 의미를 갖는 라틴어 dare(주다, to give)의 과거분사형으로 사용되었다.

- 데이터는 추론과 추정의 근거를 이루는 사실이다.
- 현실 세계에서 관찰하거나 측정하여 수집한 사실이다.
- 단순한 객체로도 가치가 있으며 다른 객체와의 상호관계 속에서 더 큰 가치를 갖는다.
- 객관적 사실이라는 존재적 특성을 갖는다.
- 추론, 추정, 예측, 전망을 위한 근거로써 당위적 특성을 갖는다.

2) 데이터의 구분

- 정량적 데이터(Quantitative Data) : 주로 숫자로 이루어진 데이터이다.
  - 예) 2020년, 100km/h 등
- 정성적 데이터(Qualitative Data) : 문자와 같은 텍스트로 구성되며 함축적 의미를 지니고 있는 데이터이다.
  - 예) 철수가 시험에 합격하였다.

1. 빅데이터의 이해 - 빅데이터 개요 및 활용

2) 데이터의 구분

정량적 데이터와 정성적 데이터의 비교

	정량적 데이터	정성적 데이터
유형	정형 데이터, 반정형 데이터	비정형 데이터
특징	여러 요소의 결합으로 의미 부여	객체 하나가 함축된 의미 내포
관점	주로 객관적 내용	주로 주관적 내용
구성	수치나 기호 등	문자나 언어 등
형태	데이터베이스, 스프레드시트 등	웹 로그, 텍스트 파일 등
위치	DBMS, 로컬 시스템 등 내부	웹사이트, 모바일 플랫폼 등 외부
분석	통계 분석 시 용이	통계 분석 시 어려움

1. 빅데이터의 이해 - 빅데이터 개요 및 활용

4) 데이터의 유형

① 정형 데이터(Structured Data) : 정해진 형식과 구조에 맞게 저장되도록 구성된 데이터이며, 연산이 가능하다.

예) 관계형 데이터베이스의 테이블에 저장되는 데이터 등

② 반정형 데이터(Semi-structured Data) : 데이터의 형식과 구조가 비교적 유연하고, 스키마 정보를 데이터와 함께 제공하는 파일 형식의 데이터이며, 연산이 불가능하다.

예) JSON, XML, RDF, HTML 등

③ 비정형 데이터(Unstructured Data) : 구조가 정해지지 않은 대부분의 데이터이며, 연산이 불가능하다.

예) 동영상, 이미지, 음성, 문서, 메일 등

스키마 : 자료의 구조, 표현 방법
JSON(JavaScript Object Notation) : 데이터 오브젝트를 전달하기 위해 인간이 읽을 수 있는 텍스트를 사용하는 개방형 표준 포맷
XML(eXtensible Markup Language) : 여러 특수 목적을 갖는 마크업 언어를 만드는 용도로 권장되는 다목적 마크업 언어

1. 빅데이터의 이해 - 빅데이터 개요 및 활용

5) 데이터의 기능

; 과학적 발견은 개인의 암묵적 지식에 기초하는 경우가 많으며, 이를 활용하려면 데이터를 기반으로 한 암묵지와 형식지의 상호작용이 중요하다.

① 암묵지 : 어떠한 시행착오나 다양하고 오랜 경험을 통해 개인에게 체계화되어 있으며, 외부에 표출되지 않은 무형의 지식으로 그 전달과 공유가 어렵다.

② 형식지 : 형상화된 유형의 지식으로 그 전달과 공유가 쉽다.

암묵지 : 학습과 경험을 통하여 개인에게 체화되어 있지만 겉으로 드러나지 않는 지식
형식지 : 명시적으로 알 수 있는 형태, 형식을 갖추어 표현되고 공유가 가능한 지식

1. 빅데이터의 이해 - 빅데이터 개요 및 활용

6) 지식창조 매커니즘

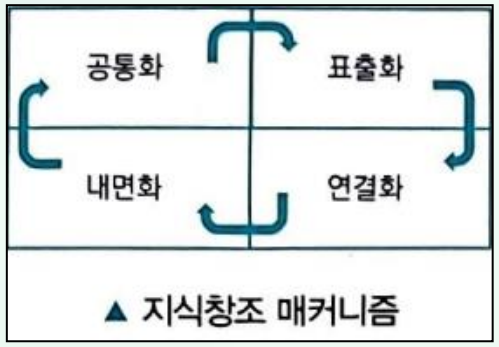
; 암묵지와 형식지 간 상호작용을 위한 일본의 경영학자 노나카 이쿠지로의 지식창조 매커니즘은 다음의 4단계로 구성된다.

① 공통화(Socialization) : 서로의 경험이나 인식을 공유하며 한 차원 높은 암묵지로 발전시킨다.

② 표출화(Externalization) : 암묵지가 구체화되어 외부(형식지)로 표현된다.

③ 연결화(Combination) : 형식지를 재분류하여 체계화한다.

④ 내면화(Internalization) : 전달받은 형식지를 다시 개인의 것으로 만든다.



▲ 지식창조 매커니즘

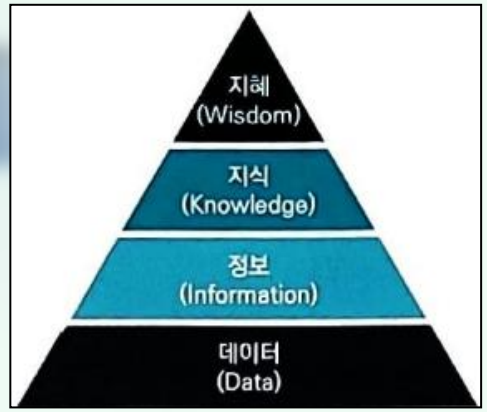
암묵지 : 학습과 경험을 통하여 개인에게 체화되어 있지만 겉으로 드러나지 않는 지식
형식지 : 명시적으로 알 수 있는 형태, 형식을 갖추어 표현되고 공유가 가능한 지식

1. 빅데이터의 이해 - 빅데이터 개요 및 활용

7) 데이터, 정보, 지식, 지혜

; 데이터, 정보, 지식, 지혜는 인간의 사회활동 속에서 가치창출을 위한 일련의 프로세스로 연결되어 기능한다.

▶ 지식의 피라미드(가치창출 프로세스)



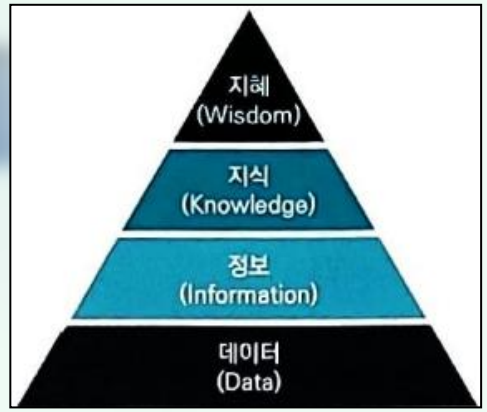
지혜 (Wisdom)	축적된 지식을 통해 근본적인 원리를 이해하고 아이디어를 결합하여 도출한 창의적 산물이다. 예 다른 상품들도 온라인 쇼핑 시 오프라인 상점보다 저렴할 것이다.
지식 (Knowledge)	상호 연결된 정보를 구조화하여 유의미한 정보를 분류하고 개인적인 경험을 결합시켜 내재화한 고유의 결과물이다. 예 오프라인 상점보다 저렴한 온라인 쇼핑으로 노트북을 구매할 것이다.
정보 (Information)	데이터를 가공하거나 처리하여 데이터 간 관계를 분석하고 그 속에서 도출된 의미를 말하며, 항상 유용한 것은 아니다. 예 오프라인 상점보다 온라인 쇼핑 시 노트북 가격이 더 저렴하다.
데이터 (Data)	현실 세계에서 관찰하거나 측정하여 수집한 사실이나 값으로 개별 데이터로는 그 의미가 중요하지 않은 객관적인 사실이다. 예 온라인 쇼핑 시 노트북 가격은 100만 원이며, 오프라인 상점의 노트북 가격은 150만 원이다.

1. 빅데이터의 이해 - 빅데이터 개요 및 활용

7) 데이터, 정보, 지식, 지혜

; 데이터, 정보, 지식, 지혜는 인간의 사회활동 속에서 가치창출을 위한 일련의 프로세스로 연결되어 기능한다.

▶ 지식의 피라미드(가치창출 프로세스)



지혜 (Wisdom)	축적된 지식을 통해 근본적인 원리를 이해하고 아이디어를 결합하여 도출한 창의적 산물이다. 예 다른 상품들도 온라인 쇼핑 시 오프라인 상점보다 저렴할 것이다.
지식 (Knowledge)	상호 연결된 정보를 구조화하여 유의미한 정보를 분류하고 개인적인 경험을 결합시켜 내재화한 고유의 결과물이다. 예 오프라인 상점보다 저렴한 온라인 쇼핑으로 노트북을 구매할 것이다.
정보 (Information)	데이터를 가공하거나 처리하여 데이터 간 관계를 분석하고 그 속에서 도출된 의미를 말하며, 항상 유용한 것은 아니다. 예 오프라인 상점보다 온라인 쇼핑 시 노트북 가격이 더 저렴하다.
데이터 (Data)	현실 세계에서 관찰하거나 측정하여 수집한 사실이나 값으로 개별 데이터로는 그 의미가 중요하지 않은 객관적인 사실이다. 예 온라인 쇼핑 시 노트북 가격은 100만 원이며, 오프라인 상점의 노트북 가격은 150만 원이다.

1. 빅데이터의 이해 - 빅데이터 개요 및 활용

개념 체크

1. 다음 중 데이터에 대한 설명으로 옳은 것은?

① 데이터는 추론과 추정의 근거를 이루는 사실로, 단순한 객체로는 가치가 없고 다른 객체와의 상호관계 속에서 가치를 갖는다.

② 정량적 데이터는 정형 데이터이고, 정성적 데이터는 반정형 데이터와 비정형 데이터이다.

③ 데이터는 추론과 추정의 근거를 이루는 사실로, 정량적 데이터와 정성적 데이터 모두 객관적 내용을 내포하고 있다.

④ 데이터의 유형을 유연성을 기준으로 나열했을 때 비정형 데이터가 가장 유연하고, 정형 데이터는 유연성이 부족하다.

데이터는 단순한 객체로도 가치가 있으며 다른 객체와의 상호관계 속에서는 더 큰 가치를 갖는다.

정량적 데이터는 정형 데이터와 반정형 데이터이고, 정성적 데이터에는 비정형 데이터이다.

정량적 데이터는 주로 객관적 내용을 나타내고, 정성적 데이터는 주관적인 내용을 내포하고 있다.

정형 데이터는 정해진 형식과 구조에 맞게 저장하여야 하지만,

1. 빅데이터의 이해 - 빅데이터 개요 및 활용

02 데이터베이스

; 데이터베이스(DataBase)라는 용어는 1963년 6월에 컴퓨터 중심의 데이터베이스 개발과 관리라는 주제로 미국 SDC(System Development Corporation)가 개최한 심포지엄에서 공식적으로 사용되었다.

1) 데이터베이스의 정의

; 체계적이거나 조직적으로 정리되고 전자식 또는 기타 수단으로 개별적으로 접근할 수 있는 독립된 저작물, 데이터 또는 기타 소재의 수집물이다.

- 데이터베이스는 소재를 체계적으로 배열 또는 구성한 편집물로서 개별적으로 그 소재에 접근하거나 그 소재를 검색할 수 있도록 한 것이다.(저작권법)
- 동시에 복수의 적용 업무를 지원할 수 있도록 복수 이용자의 요구에 대응해서 데이터를 받아들이고 저장, 공급하기 위하여 일정한 구조에 따라서 편성된 데이터의 집합이다.
- 문자, 기호, 음성, 화상, 영상 등 상호 관련된 다수의 콘텐츠를 정보 처리 및 정보 통신 기기에 의하여 체계적으로 수집, 축적하여 다양한 용도와 방법으로 이용할 수 있도록 정리한 정보의 집합체이다.

1. 빅데이터의 이해 - 빅데이터 개요 및 활용

2) 데이터베이스관리시스템(DBMS: DataBase Management System)

; 데이터베이스를 관리하며 응용 프로그램들이 데이터베이스를 공유하며 사용할 수 있는 환경을 제공하는 소프트웨어이다.

● 데이터베이스 관리 시스템의 종류

종류	설명
관계형 DBMS	데이터를 열과 행을 이루는 테이블로 표현하는 모델이다.
객체지향 DBMS	정보를 객체 형태로 표현하는 모델이다.
네트워크 DBMS	그래프 구조를 기반으로 하는 모델이다.
계층형 DBMS	트리 구조를 기반으로 하는 모델이다.

● SQL(Structured Query Language)

- 데이터베이스에 접근할 때 사용하는 언어이다.
- 단순한 질의 기능 뿐만 아니라 데이터 정의와 조작기능을 갖추고 있다.
- 테이블 단위로 연산을 수행하며 초보자들도 비교적 쉽게 사용 가능하다.

1. 빅데이터의 이해 - 빅데이터 개요 및 활용

3) 데이터베이스의 특징

① 통합된 데이터(Integrated Data)

; 동일한 데이터가 중복되어 저장되지 않음을 의미한다.

- 데이터의 중복은 관리상 복잡하고 다양한 문제를 초래한다.

② 저장된 데이터(Stored Data)

; 컴퓨터가 접근할 수 있는 저장매체에 데이터를 저장한다.

③ 공용 데이터(Shared Data)

; 여러 사용자가 서로 다른 목적으로 데이터를 함께 이용한다.

- 일반적으로 대용량화 되어 있고 구조가 복잡하다.

④ 변화되는 데이터(Changed Data)

; 데이터는 현시점의 상태를 나타내며 지속적으로 갱신된다.

- 갱신으로 변화하면서도 현재의 정확한 데이터를 유지해야 한다.

1. 빅데이터의 이해 - 빅데이터 개요 및 활용

3) 데이터베이스의 특징

▶ 데이터베이스의 장단점

장점	단점
•데이터 중복 최소화 •실시간 접근 가능 •데이터 보안강화 •논리적 및 물리적 독립성 제공 •데이터 일관성 제공 •데이터 무결성 보장 •데이터 공유용이	•구축과 유지에 따른 비용 발생 •백업과 복구 등 관리 필요

1. 빅데이터의 이해 - 빅데이터 개요 및 활용

4) 데이터베이스의 활용

① OLTP(OnLine Transaction Processing)

; 호스트 컴퓨터와 온라인으로 접속된 여러 단말 간 처리 형태의 하나로 데이터베이스의 데이터를 수시로 갱신하는 프로세싱을 의미한다.

▶ 여러 단말에서 보내온 메시지에 따라 호스트 컴퓨터가 데이터베이스를 액세스 하고, 바로 처리 결과를 돌려보내는 형태를 말한다.

▶ 현재 시점의 데이터만을 데이터베이스가 관리한다는 개념이다.

- 이미 발생된 트랜잭션에 대해서는 데이터 값이 과거의 데이터로 다른 디스크나 테이프 등에 보관될 수 있다.

② OLAP(OnLine Analytical Processing)

; 정보 위주의 분석 처리를 하는 것으로, OLTP에서 처리된 트랜잭션 데이터를 분석해 제품의 판매 추이, 구매 성향 파악, 재무 회계 분석 등을 프로세싱 하는 것을 의미한다.

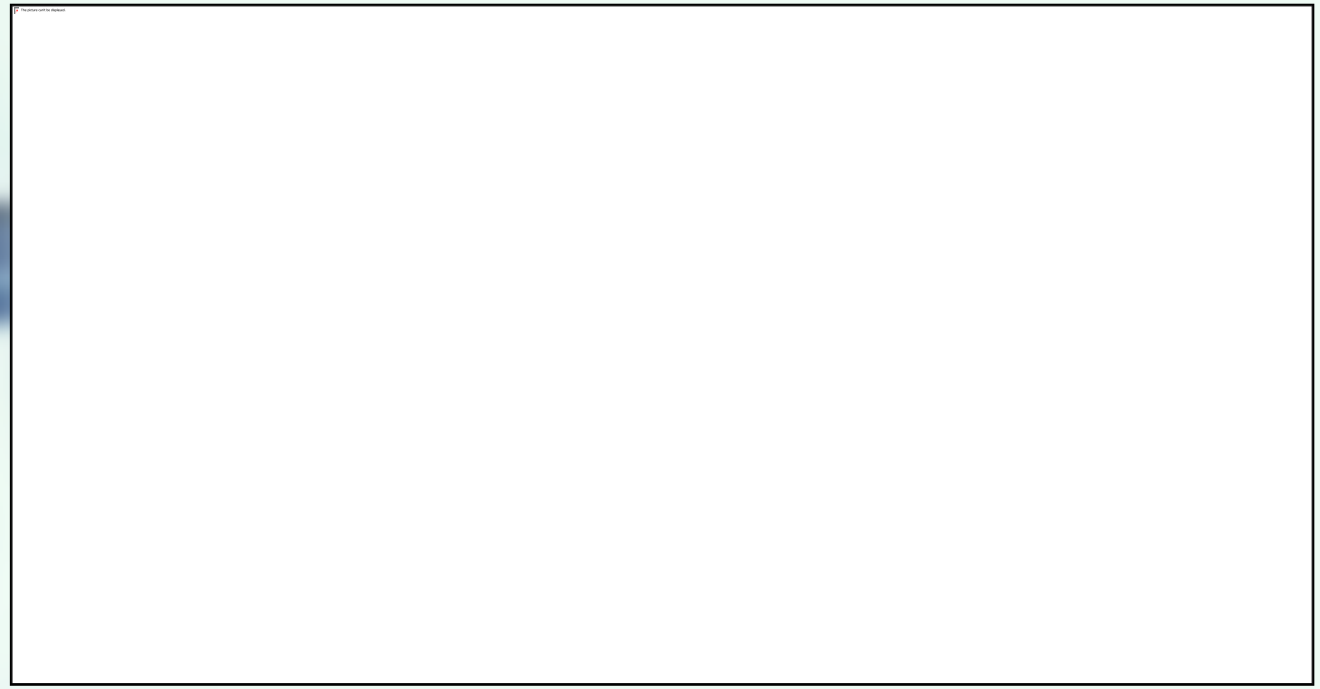
▶ 다양한 비즈니스 관점에서 쉽고 빠르게 다차원적인 데이터에 접근하여 의사결정에 활용할 수 있는 정보를 얻을 수 있게 하는 기술이다.

OLTP가 데이터 갱신 위주라면, OLAP는 데이터 조회 위주라고 할 수 있다.

1. 빅데이터의 이해 - 빅데이터 개요 및 활용

4) 데이터베이스의 활용

▶ OLTP와 OLAP의 비교



1. 빅데이터의 이해 - 빅데이터 개요 및 활용

5) 데이터 웨어하우스(DW: Data Warehouse)

; 사용자의 의사결정에 도움을 주기 위하여 기관시스템의 데이터베이스에 축적된 데이터를 공통의 형식으로 변환해서 관리하는 데이터베이스이다. 데이터 웨어하우스는 일정한 시간 동안의 데이터를 축적하고 의사결정을 위한 다양한 분석 작업을 수행한다.

▶ 데이터 웨어하우스의 특징

특징	내용
주제지향성(Subject-orientation)	고객 제품 등과 같은 중요한 주제를 중심으로 그 주제와 관련된 데이터들로 구성된다.
통합성(Integration)	데이터가 데이터 웨어하우스에 입력될 때는 일관된 형태로 변환되며, 전사적인 관점에서 통합된다.
시계열성(Time-variant)	데이터 웨어하우스의 데이터는 일정 기간 동안 시점 별로 이어진다.
비휘발성(Non-volatilization)	데이터 웨어하우스에 일단 데이터가 적재되면 일괄 처리 작업에 의한 갱신 이외에는 변경이 수행되지 않는다.

* 데이터 웨어하우스 : 데이터만이 아닌 분석 방법까지도 포함하여 조직 내 의사결정을 지원하는 정보 관리 시스템

1. 빅데이터의 이해 - 빅데이터 개요 및 활용

5) 데이터 웨어하우스(DW: Data Warehouse)

▶ 데이터 웨어하우스의 구성

구성 요소	내용
데이터 모델(Data Model)	주제 중심적으로 구성된 다차원의 개체-관계형(Entity Relation) 모델로 설계 된다.
ETL(Extract, Transform, Load)	기업의 내부 또는 외부로부터 데이터를 추출, 정제 및 가공하여 데이터 웨어 하우스에 적재한다.
ODS(Operational Data Store)	다양한 DBMS 시스템에서 추출한 데이터를 통합적으로 관리한다.
DW 메타데이터	데이터 모델에 대한 스키마 정보와 비즈니스 측면에서 활용되는 정보를 제공 한다.
OLAP(Online Analytical Processing)	사용자가 직접 다차원의 데이터를 확인할 수 있는 솔루션이다.
데이터마이닝(Data Mining)	대용량의 데이터로부터 인사이트를 도출할 수 있는 방법론이다.
분석 도구	데이터마이닝을 활용하여 데이터 웨어하우스에 적재된 데이터를 분석할 수 있는 도구이다.
경영기반 솔루션	KMS, DSS, BI와 같은 경영의사결정을 지원하기 위한 솔루션이다.

KMS(Knowledge Management System) : 지식 관리 시스템
DSS(Decision Support System) : 의사 결정 지원 시스템
BI(Business Intelligence) : 데이터를 분석해 기업의 의사결정에 활용하는 일련의 프로세스

1. 빅데이터의 이해 - 빅데이터 개요 및 활용

개념 체크

1. 다음 중 데이터베이스에 대한 설명으로 틀린 것은?

① 데이터베이스는 관련된 레코드의 집합이며, 이를 위한 소프트웨어를 데이터베이스 관리 시스템이라 한다.

② SQL은 데이터베이스에 접근할 때 사용하는 언어이며, 데이터 정의와 조작이 가능하다.

③ 데이터베이스는 통합된 데이터, 저장된 데이터, 공용 데이터 그리고 변화되는 데이터이다.

④ OLTP는 조회 중심의 데이터베이스로 현재 시점의 데이터만을 관리한다는 개념이다.

데이터베이스는 동시에 복수의 적용 업무를 지원할 수 있도록 복수 이용자의 요구에 대응해서 데이터를 받아들이 고 저장, 공급하기 위하여 일정한 구조에 따라서 편성된 데이터의 집합체이며, 이를 효율적으로 관리하기 위한 시스템이 바로 DBMS이다. SQL은 단순한 질의뿐만 아니라 데이터 정의와 조작이 가능하며, 초보자들도 비교적 쉽게 사용할 수 있는 언어이다.

데이터베이스는 통합된 데이터(Integrated Data), 저장된

2. 다음 중 데이터 웨어하우스에 대한 설명으로 틀린 것은?

① 데이터만이 아닌 분석 방법까지도 포함하여 조직 내 의사결정을 지원하는 정보 관리 시스템이다.

② 주제 중심적이고, 각 주제별로 분리되어 있으며, 시계열 형태의 비휘발성 데이터이다.

③ 데이터 웨어하우스를 구성하는 ETL은 Extract, Transform, Load의 약어이다.

④ DW 메타데이터는 데이터 모델에 대한 스키마 정보 등을 제공한다.

데이터 웨어하우스는 사용자의 의사결정에 도움을 주기 위하여 기관 시스템의 데이터베이스에 축적된 데이터를 공통의 형식으로 변환해서 관리하는 데이터베이스이다. 주제지향성, 통합성, 시계열성, 비휘발성이라는 특징을 갖고 있다.

ETL은 기업의 내부 또는 외부로부터 데이터를 추출(Extract) , 변환, 정제(Transform), 데이터 웨어하우스에 적재(Load) 하는 과정이다.

DW 메타데이터는 데이터 모델에 대한 스키마 정보와

1. 빅데이터의 이해 - 빅데이터 개요 및 활용

03 빅데이터 개요

; 빅데이터는 기존 데이터보다 너무 방대하여 기존의 방법이나 도구로 수집/저장/분석 등이 어려운 정형 및 비정형 데이터들을 의미한다.

- 빅데이터는 일반적인 데이터베이스 소프트웨어로 저장, 관리, 분석할 수 있는 범위를 초과하는 규모의 데이터이다.
- 빅데이터는 다양한 종류의 대규모 데이터로부터 저렴한 비용으로 가치를 추출하고 데이터의 초고속 수집, 발굴, 분석을 지원하도록 고안된 차세대 기술 및 아키텍처이다.
- 빅데이터는 대용량 데이터를 활용해 작은 용량에서는 얻을 수 없었던 새로운 통찰이나 가치를 추출해 내며, 나아가 이를 활용해 시장과 기업 및 시민과 정부의 관계 등 많은 분야에 변화를 가져오는 것이다.

빅데이터에 대한 인식은 데이터 규모와 기술 측면에서 시작했지만, 빅데이터의 가치와 효과 측면으로 최근 그 의미가 확대되고 있다.

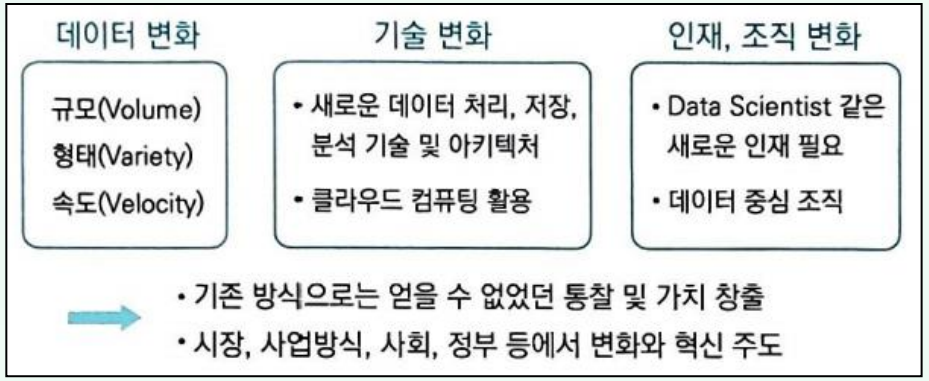
1. 빅데이터의 이해 - 빅데이터 개요 및 활용

1) 빅데이터의 등장과 변화

① 빅데이터의 등장 배경

디지털화, 저장 기술, 인터넷 보급, 모바일 혁명, 클라우드 컴퓨팅 등 관련 기술이 빠르게 발전하고 있다.

- 기업에서는 온·오프라인 고객 데이터가 많이 축적되면서 데이터에 숨어 있는 가치를 발굴해 새로운 성장 동력으로 활용하고 있다.
- 학계에서는 인간 게놈 프로젝트, 기후 관찰 등 거대 데이터를 다루는 학문 분야가 확산되면서 필요한 기술 아키텍처 및 분석 기법들이 발전하고 있다.



인간 게놈 프로젝트(Human Genome Project)는 인간 게놈을 구성하는 30억쌍의 염기서열 전체를 밝히고 유전자지도를 완성하고자 하는 초거대 프로젝트이다. 말하자면 호모사피엔스의 '생명의 책'을 해독하는 작업인 것이다.

1. 빅데이터의 이해 - 빅데이터 개요 및 활용

1) 빅데이터의 등장과 변화

② 빅데이터의 등장으로 인한 변화

● 데이터 처리 시점이 사전 처리(pre-processing)에서 사후 처리(post-processing)로 이동하였다.

- 기존에 필요한 정보만 수집하는 시스템에서 가능한 한 많은 데이터를 모으고 다양한 방식으로 조합하여 숨은 정보를 얻는 방식으로 변화

● 데이터 처리 범주가 표본조사에서 전수조사로 확대되었다.

- 기술 발전으로 인한 데이터 처리비용 감소로 표본조사가 아닌 전수조사를 통해 샘플링이 주지 못하는 패턴이나 정보를 발견하는 방식으로 변화

● 데이터의 가치 판단기준이 질(quality)보다 양(quantity)으로 그 중요도가 달라졌다.

- 데이터의 지속적 추가는 양질의 정보가 오류 정보보다 많아 전체적으로 좋은 결과를 산출하는 데 긍정적인 영향을 미친다는 추론을 바탕으로 변화

● 데이터를 분석하는 방향이 이론적 인과관계 중심에서 단순한 상관관계로 변화되는 경향이 있다.

- 데이터 기반의 상관관계 분석으로 특정 현상의 발생 가능성을 포착하여 대응하는 방식으로 변화

1. 빅데이터의 이해 - 빅데이터 개요 및 활용

2) 빅데이터의 특징

; 빅데이터 용어가 사용된 초기에 가트너(Gartner) 그룹은 3V(규모, 유형, 속도)로 빅데이터의 특징을 설명했으며, 최근에는 빅데이터 분석을 통해 얻을 수 있는 가치와 데이터에 대한 품질의 중요성이 강조되고 있다.

▶ 빅데이터의 특징

광의	협의	특징	내용
5V	3V	규모(Volume)	• 데이터 양이 급격하게 증가(대용량화) • 기존 데이터 관리 시스템의 성능적 한계 도달
		유형(Variety)	• 데이터의 종류와 근원 확대(다양화) • 정형 데이터 외 반정형 및 비정형 데이터로 확장
		속도(Velocity)	• 데이터 수집과 처리 속도의 변화(고속화) • 대용량 데이터의 신속하고 즉각적인 분석 요구
	+2V	품질(Veracity)	• 데이터의 신뢰성, 정확성, 타당성 보장이 필수 • 고품질의 데이터에서 고수준 인사이트 도출 가능
		가치(Value)	• 대용량의 데이터 안에 숨겨진 가치 발굴이 중요 • 다른 데이터들과 연계 시 가치가 배로 증대

가트너 주식회사(Gartner, Inc.)는 미국의 정보 기술 연구 및 자문 회사이다. 1979년 창립되었다.

1. 빅데이터의 이해 - 빅데이터 개요 및 활용

2) 빅데이터의 특징

▶ 전통적 데이터와 빅데이터 비교

	전통적 데이터	빅데이터
규모	기가바이트(GB) 이하	테라바이트(TB) 이상
처리단위	시간 또는 일 단위 처리	실시간 처리
유형	정형 데이터	정형+반정형, 비정형 데이터
처리방식	중앙집중식 처리	분산 처리
시스템	Relational DBMS	Hadoop, HDFS, Hbase, NoSQL 등

▶ 데이터 크기의 발전 과정

킬로바이트(KB)	2^10	10^3
메가바이트(MB)	2^20	10^6
기가바이트(GB)	2^30	10^9
테라바이트(TB)	2^40	10^12
페타바이트(PB)	2^50	10^15
엑사바이트(EB)	2^60	10^18
제타바이트(ZB)	2^70	10^21
요타바이트(YB)	2^80	10^24

바이트(byte)는 컴퓨터가 조작하는 정보의 최소 처리 단위이다.

1. 빅데이터의 이해 - 빅데이터 개요 및 활용

3) 빅데이터의 활용

▶ 빅데이터의 활용을 위한 3요소

구성 요소	내용
자원(Resource)[빅데이터]	•정형, 반정형, 비정형 데이터를 실시간으로 수집한다. •수집된 데이터를 전처리 과정을 통해 품질을 향상시킨다.
기술(Technology) [빅데이터플랫폼, AI]	•분산 파일 시스템을 통해 대용량 데이터를 분산 처리한다. •데이터마이닝 등을 통해 데이터를 분석 및 시각화한다. •데이터를 스스로 학습, 처리할 수 있는AI 기술을 활용한다.
인력(People) [알고리즈미스트, 데이터사이언티스트]	•통계학, 수학, 컴퓨터공학, 경영학분야전문지식을 갖춘다. •도메인 지식을 습득하여 데이터 분석 및 결과를 해석한다.

빅데이터 활용을 위한 3대 요소로는 자원, 기술 인력이 있다.

1. 빅데이터의 이해 - 빅데이터 개요 및 활용

3) 빅데이터의 활용

▶ 빅데이터의 활용을 위한 기본 테크닉

테크닉	설명	예시
연관규칙학습	변인들 간 주목할 만한 상관관계가 있는지 찾아내는 방법	도시락을 구매하는 사람이 음료수를 더 많이 구매하는가?
유형분석	문서를 분류하거나 조직을 그룹화할 때 사용	이것은 어떤 특성을 가진 집단에 속하는가?
유전 알고리즘	최적화가 필요한 문제를 생물 진화의 과정을 모방하여 점진적으로 해결책을 찾는 방법	시청률을 최고치로 하기 위해 어떤 프로그램을 어떤 시간에 방송해야 하는가?
기계학습	데이터로부터 학습한 알려진 특성을 활용하여 예측	시청 기록을 바탕으로 어떤 영화를 가장 보고 싶어하는가?
회귀분석	독립변수가 종속변수에 미치는 영향을 분석할 때 사용	경력과 학력이 연봉에 미치는 영향은?
감정분석	특정 주제에 대해 말을 하거나 글을 쓴 사람의 감정을 분석	새로운 할인 정책에 대한 고객의 평은 어떤가?
소셜네트워크(사회관계망)분석	특정인과 다른 사람의 관계를 파악하고 영향력 있는 사람을 분석할 때 사용	고객들 간 관계망은 어떻게 구성되는가?

1. 빅데이터의 이해 - 빅데이터 개요 및 활용

개념 체크

1. 다음 중 빅데이터에 대한 설명으로 틀린 것은?

① 데이터 처리 범주가 표본조사에서 전수조사로 확대 되었다.

② 가트너 그룹은 3V(규모, 유형, 속도)로 빅데이터의 특징을 설명하였다.

③ 빅데이터를 활용하기 위한 3요소로 자원, 기술, 인력이 필요하며, 그 중 자원은 빅데이터와 플랫폼으로 구성된다.

④ 빅데이터 활용을 위한 방법들로는 연관규칙분석, 기계학습, 회귀분석 등이 존재한다.

기술 발전으로 인한 데이터 처리비용 감소로 표본조사가 아닌 전수조사를 통해 패턴이나 정보를 발견하는 방식으로 변화되었다.

빅데이터는 규모(Volume)면에서 대용량화, 유형(Variety)면에서는 다양화, 속도(Velocity)면에서는 고속화된 특징을 갖고 있다.

빅데이터 활용을 위한 3요소로는 자원(빅데이터), 기술(빅데이터 플랫폼, AI), 인력(알고리즈미스트, 데이터 사이언티스트)가 필요하다.

1. 빅데이터의 이해 - 빅데이터 개요 및 활용

04 빅데이터의 가치

▶ 빅데이터 활용을 통해 얻는 가치

기관명	경제적 효과
Economist(2010)	데이터는 자본이나 노동력과 거의 동등한 레벨의 경제적 투입자본으로 비즈니스의 새로운 원자재 역할을 한다.
MIT Sloan(2010)	데이터 분석을 잘 활용하는 조직일수록 차별적 경쟁력을 갖추고 높은 성과를 창출한다.
Gartner(2011)	데이터는 21세기의 원유이며 미래 경쟁 우위를 결정한다. 기업은 다가올 데이터 경제시대를 이해하고 정보고립을 경계해야 한다.
McKinsey(2011)	빅데이터는 혁신, 경쟁력, 생산성의 핵심요소이다.

빅데이터의 역할
- 4차 산업혁명시대의 석탄이나 철, 원유와 같은 역할
- 사실관계를 상세하게 들여다볼 수 있는 렌즈 역할
- 다양한 개발자들에게 사업 기회를 주는 플랫폼 역할

1. 빅데이터의 이해 - 빅데이터 개요 및 활용

1) 빅데이터의 기능과 효과

- 빅데이터는 이를 활용하는 기존 사업자에게 경쟁 우위를 제공한다.
  - 새롭게 시장에 진입하려는 잠재적 경쟁자에게는 진입장벽과도 같다.
  - 고객 세분화와 맞춤형 개인화 서비스를 제공할 수 있다.
  - 시뮬레이션을 통한 수요 포착과 변수 탐색으로 경쟁력을 강화하고, 비즈니스 모델이나 제품 또는 서비스의 혁신을 가져온다.
- 빅데이터는 알고리즘 기반으로 의사결정을 지원하거나 이를 대신한다.
- 빅데이터는 투명성을 높여 R&D 및 관리 효율성을 제고한다.

1. 빅데이터의 이해 - 빅데이터 개요 및 활용

2) 빅데이터의 가치 측정의 어려움

; 특정 데이터의 가치는 그 데이터의 활용 및 가치 창출 방식과 분석 기술의 발전여부 등에 따라 달라질 수 있어 이를 측정하고 판단하는 것은 쉽지 않다.

① 데이터 활용 방식 : 데이터를 재사용하거나 재결합, 다목적용 데이터 개발 등이 일반화되면서 특정 데이터를 누가, 언제, 어디서 활용할지 알 수 없기에 그 가치를 측정하기 어렵다.

② 가치 창출 방식 : 데이터는 어떠한 목적을 갖고서 어떻게 가공하는가에 따라 기존에 없던 가치를 창출할 수도 있어 사전에 그 가치를 측정하기 어렵다.

③ 분석 기술 발전 : 데이터는 지금의 기술 상황에서는 가치가 없어 보일지라도 새로운 분석 기법이 등장할 경우 큰 가치를 찾아낼 수 있으므로 당장 그 가치를 측정하기 어렵다.

④ 데이터 수집 원가 : 데이터는 달성하려는 목적에 따라 수집하거나 가공하는 비용이 상황에 따라 달라질 수 있어 그 가치를 측정하기 어렵다.

대용량 데이터에 맞는 자료 관리 기술과 자료 분석 기술이 필요해졌다.

1. 빅데이터의 이해 - 빅데이터 개요 및 활용

3) 빅데이터의 영향

- 기업에게 혁신과 경쟁력 강화, 생산성 향상의 근간이 된다.
  - 빅데이터를 활용해 소비자의 행동을 분석하고 시장 변동을 예측해 비즈니스 모델을 혁신하거나 신사업을 발굴한다.
- 정부에게 환경 탐색과 상황 분석, 미래 대응 수단을 제공한다.
  - 기상, 인구이동, 각종 통계, 법제 데이터 등을 수집해 사회 변화를 추정하여 관련 정보를 추출한다.
- 개인에게 활용 목적에 따라 스마트화를 통해 영향을 준다.
  - 빅데이터를 서비스하는 기업이 많아지고 데이터 분석 비용은 지속적으로 하락하여 활용이 용이해졌다.

1. 빅데이터의 이해 - 빅데이터 개요 및 활용

개념 체크

1. 다음 중 빅데이터의 가치에 대한 설명으로 틀린 것은?

① 빅데이터는 이를 활용하는 사업자에게 경쟁우위를 제공한다.

② 빅데이터는 다양한 개발자들에게 사업 기회를 주는 플랫폼의 역할을 한다.

③ 빅데이터는 가치를 측정하는데 어려움이 있다.

④ 빅데이터는 기업이나 정부에게는 영향을 미치지만 개인에게는 별다른 영향을 미치지 않는다.

빅데이터는 잠재적 경쟁자에게 진입장벽과도 같다.

빅데이터는 4차 산업혁명시대의 석탄이나, 철, 원유와 같은 원자재 같은 역할을 한다.

특정 데이터의 가치는 그 데이터의 활용 및 가치 창출 방식과 분석 기술의 발전 여부 등에 따라 달라질 수 있어 이를 측정하고 판단하는 것은 쉽지 않다.

빅데이터는 정부, 기업, 개인한테 사회 전반적인 모든 곳에 영향을 끼친다. 특히 개인에게는 빅데이터로 활용 목적에 따라 스마트화를 통해서 영향을 준다.

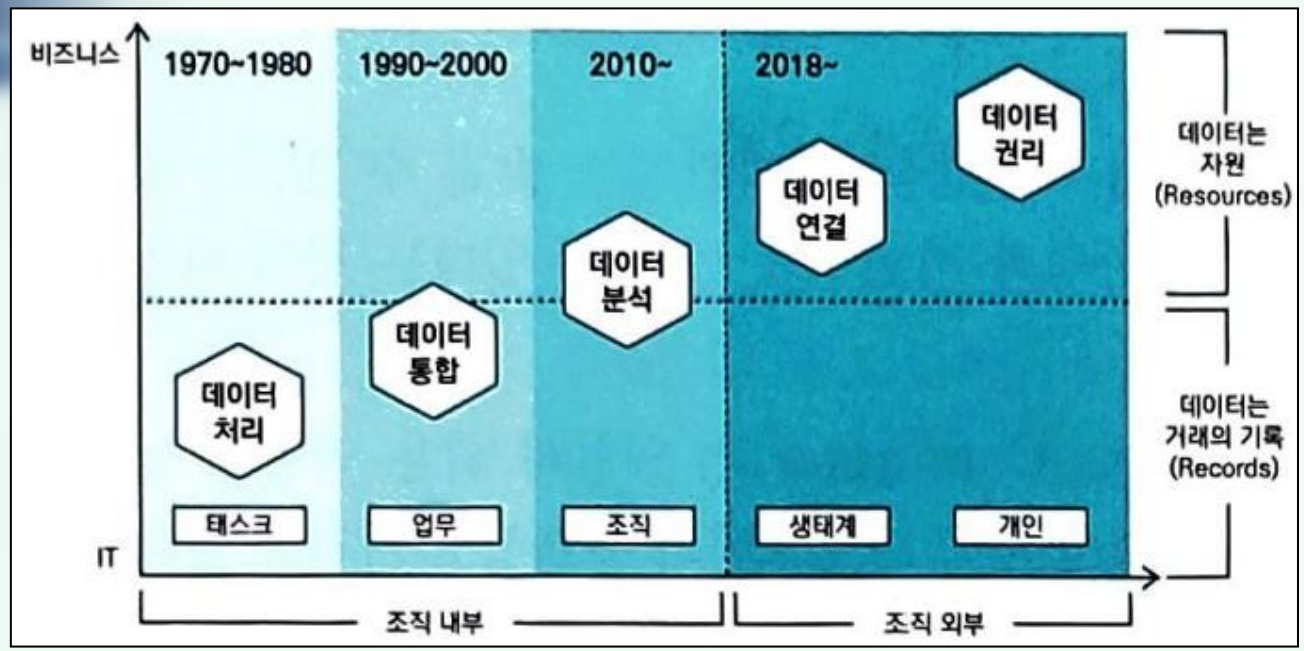
1. 빅데이터의 이해 - 빅데이터 개요 및 활용

05 데이터 산업의 이해

1) 데이터 산업의 진화

; 데이터 산업은 데이터 처리 - 통합 - 분석 - 연결 - 권리 시대로 진화하고 있다.

● 데이터 통합 시대까지 데이터의 역할은 거래를 정확하게 기록하고 거래의 자동화를 지원하는 것이었다. 데이터 분석 수준이 향상되면서 데이터의 자원 활용이 가능해졌다.



1. 빅데이터의 이해 - 빅데이터 개요 및 활용

1) 데이터 산업의 진화

① 데이터 처리 시대

▶ 컴퓨터 프로그래밍 언어를 이용하여 대규모 데이터를 빠르고 정확하게 처리할 수 있게 되었으며 결과는 파일 형태로 보관되었다.

▶ 기업들은 EDPS(Electronic Data Processing System)를 도입하여 급여계산, 회계 전표 처리 등의 업무에 적용하였다.

▶ 데이터는 업무 처리의 대상으로 새로운 가치를 제공하지는 않았다.

1. 빅데이터의 이해 - 빅데이터 개요 및 활용

1) 데이터 산업의 진화

② 데이터 통합시대

- ▶ 데이터 처리가 여러 업무에 적용되기 시작하면서 데이터가 쌓이기 시작했고 전사적으로 데이터 일관성을 확보하기가 어려워졌다.
- ▶ 데이터 모델링과 데이터베이스 관리 시스템이 등장했다.
- ▶ 데이터 조회와 보고서 산출, 원인 분석 등을 위해 데이터 웨어하우스가 도입되었다.

③ 데이터 분석 시대

- ▶ 대부분 업무에 정보기술이 적용되고, 모바일 기기 보급, 공정센서 확대, 소셜 네트워크 이용 확산 등으로 인해 데이터가 폭발적으로 증가했다.
- ▶ 대규모 데이터를 보관하고 관리할 수 있는 하둡, 스파크 등의 빅데이터 기술이 등장했다.
- ▶ 데이터를 학습하여 전문가보다도 정확한 의사결정을 빠르게 내릴 수 있는 인공지능 기술도 상용화되었다.
- ▶ 데이터 소비자(Data Consumer)의 역할과 활용 역량을 높이기 위한 데이터 리터러시(Data Literacy) 프로그램의 중요성도 커지고 있다.

데이터 모델링(Data Modeling) : 통합된 데이터를 일관성 있게 관리하기 위한 데이터베이스 설계 기법
하둡 : 일반 상용 서버로 구성된 클러스터에서 사용할 수 있는 분산 파일 시스템과 대량의 자료를 처리하기 위한 분산 처리 시스템을 제공하는 오픈 소스 프레임워크
데이터 리터러시 : 데이터를 읽고 그 의미를 파악하는 해독 능력

1. 빅데이터의 이해 - 빅데이터 개요 및 활용

1) 데이터 산업의 진화

④ 데이터 연결 시대

▶ 기업 또는 기관, 사람, 사물 등 모든 것이 항상 그리고 동시에 둘 이상의 방식으로 연결되어 데이터를 주고 받는다.

▶ 연결은 네트워크를 만들고, 네트워크는 새로운 비즈니스 모델을 탄생시킨다.

▶ 디지털 경제의 주축 세력인 디지털 원주민은 융합된 서비스를 원한다.

- 융합된 서비스를 제공하기 위해서는 다양한 기업들의 서비스 연결이 필요하고, 이는 기업 간 데이터로 연결되어야 한다.

▶ 데이터 경제의 데이터 연결을 강조하는 의미에서, 오픈 API 경제라는 용어가 사용되기도 한다.

또한, 오픈 API 제공 수 및 접속 수, 오픈 API로 연결된 외부 실체 수 등이 기업의 지속 가능성과 성장성을 확인할 수 있는 지표가 되기도 한다.

▶ 현재 오픈 API를 제공하는 것은 해당 기업의 자율적 판단에 달려 있지만, 점차 의무화 되는 추세이다.

플랫폼 비즈니스 : 네트워크 효과를 이용한 비즈니스 모델
Open API : 특정 서비스를 제공하는 업체가 자신들의 서비스에 접근할 수 있도록 그 방법을 외부에 공개한 것

1. 빅데이터의 이해 - 빅데이터 개요 및 활용

1) 데이터 산업의 진화

⑤ 데이터 권리 시대

▶ 개인이 자신의 데이터를 자신을 위해서 사용한다.

- 데이터의 원래 소유자인 개인이 자신의 데이터에 대한 권리를 보유하고 있으며 스스로 행사할 수 있어야 한다는 마이데이터(My Data)가 등장하였다.

▶ 데이터 권리를 개인이 갖게 된다는 것은 산업이 데이터를 중심으로 재편될 수 있다는 뜻이다.

- 데이터는 기본적으로 거래 행위의 부산물이었다. 기업들은 개인과 거래를 하는 과정에서 개인의 데이터가 있어야 했고, 이를 확보하였지만 몇 가지 문제(유출, 미동의 활용, 데이터의 산재)를 일으켰다.
- 개인의 데이터를 관리해 줄 수 있는 서비스와 필요한 수요자에게 데이터를 팔아주는 서비스가 나타날수 있다.
- 개인은 스스로 데이터를 만들고 자신이 만든 데이터를 기반으로 하는 비즈니스 모델을 구상할 수 있다.

▶ 데이터의 공정한 사용이 보장되어야 하며, 데이터 독점이 유발할 수 있는 경제 독점이 방지되어야 한다.

마이데이터 : 개인 데이터의 활용처와 활용범위 등에 대한 정보주체의 능동적인 의사결정을 지원, 개인 정보 자기 결정권 보장

1. 빅데이터의 이해 - 빅데이터 개요 및 활용

개념 체크

1. 다음 중 데이터 산업의 진화과정에 대한 설명으로 틀린 것은?

① 데이터 처리 시대에는 데이터가 업무 처리의 대상으로 새로운 가치를 제공하였다.

② 데이터 통합 시대에는 데이터 모델링과 데이터베이스 관리 시스템이 등장하였다.

③ 데이터 분석 시대에는 대규모 데이터를 보관하고 관리할 수 있는 빅데이터 기술이 등장하였다.

④ 데이터 권리 시대에는 개인이 자신의 데이터를 자신을 위해서 사용하게 되었다.

데이터 처리 시대의 업무 처리의 대상으로 새로운 가치를 제공하지는 않았다.

데이터 통합 시대에는 전사적으로 데이터 일관성을 확보 하기 위해 데이터 모델링과 DBMS를 도입하기 시작하였다.

데이터 분석 시대에는 하둡, 스파크 등 빅데이터 기술과 의사결정을 빠르고 정확하게 내릴 수 있는 인공지능 기술이 상용화 되었다.

1. 빅데이터의 이해 - 빅데이터 개요 및 활용

06 빅데이터 조직 및 인력

; 기업의 경쟁력 확보를 위해 비즈니스 질문을 도출하고, 이를 충족하기 위한 가치를 발굴하며, 비즈니스를 최적화하기 위하여 빅데이터 조직 및 인력 구성 방안을 수립한다.

1) 필요성

- 빅데이터와 관련된 기술적인 문제들은 기술의 발전으로 어느 정도 해소되었다.
- 데이터 분석 및 활용을 위한 조직체계나 분석 전문가 확보에 어려움이 있다.
- 데이터 분석 관점의 컨트롤 타워에 대한 필요성이 제기되고 있다.

2) 조직의 역할

- 전사 및 부서의 분석 업무를 발굴한다.
- 전문적인 분석 기법과 도구를 활용하여 빅데이터 속에서 인사이트를 찾아낸다.
- 발견한 인사이트를 전파하고 이를 실행한다.

데이터 분석 활용을 통한 성과 창출을 위해서는 조직 역량의 개발, 인력의 영입 등과 같은 전사 관점의 전략이 필요하다.
데이터 인사이트 : 엔터프라이즈 조직의 정보를 깊이 있게 이해하여 얻을 수 있는 직접적인 이점입니다. 특정 문제 및 패턴에 대한 분석은 데이터 인사이트로 이어지며, 조직은 이를 의사 결정에 활용하여 ROI(Return On Investment, 투자 대비 이익율) 개선, 시장에 대한 이해도 강화, 조직 자체 및 클라이언트 기반에서 얻는 이점 증가 등의 효과를 얻을 수 있습니다.

1. 빅데이터의 이해 - 빅데이터 개요 및 활용

3) 조직의 구성

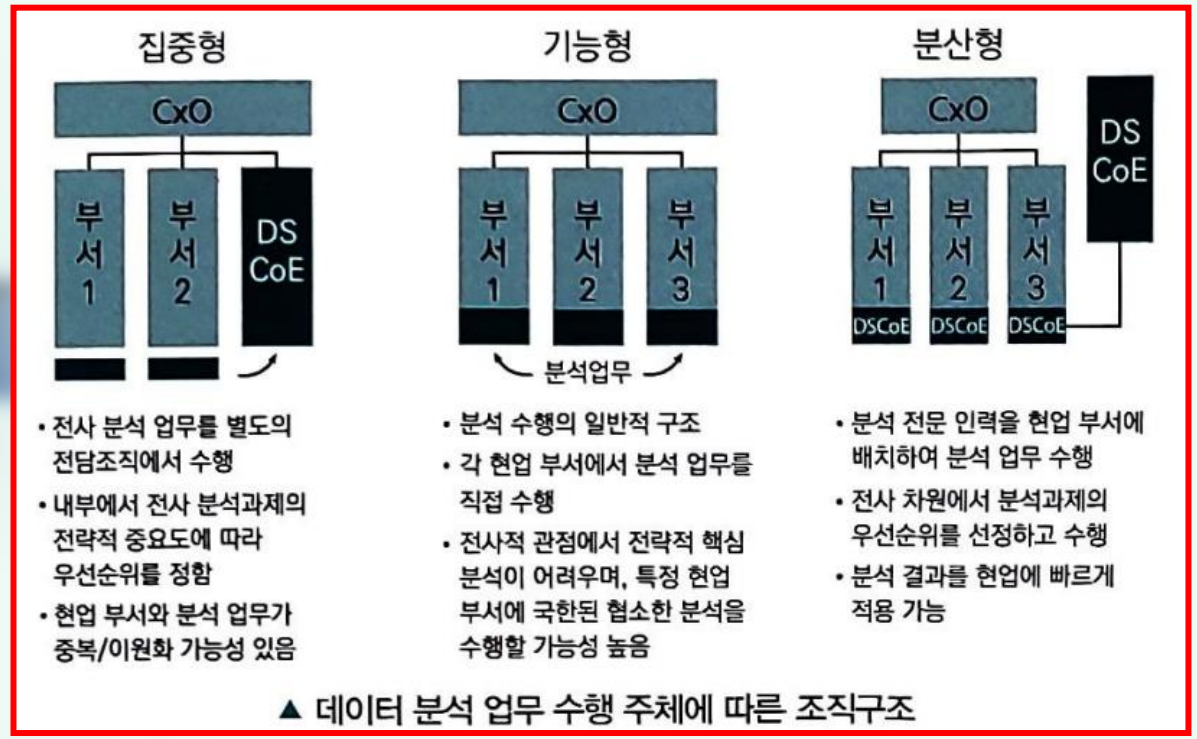
통계학이나 분석 방법에 대한 지식과 분석 경험이 있는 전문인력을 중심으로 전사 또는 특정 부서 내 조직으로 구성하여 운영한다.

① 조직 구성을 위한 체크리스트

- 비즈니스 질문을 선제적으로 찾아낼 수 있는 구조인가?
- 분석 전담조직과 타 부서 간 유기적인 협조와 지원이 원활한 구조인가?
- 효율적인 분석 업무를 수행하기 위한 분석 조직의 내부조직구조는?
- 전사 및 단위부서가 필요 시 접촉하며 지원할 수 있는 구조인가?
- 어떤 형태의 조직(집중형, 기능형, 분산형)으로 구성하는 것이 효율적인가?

1. 빅데이터의 이해 - 빅데이터 개요 및 활용

3) 조직의 구성



▲ 데이터 분석 업무 수행 주체에 따른 조직구조

CxO(Chief Experience Officer) : 최고 경영 책임자
DSCoE(Data Science Center of Excellence) : 데이터 분석 조직

1. 빅데이터의 이해 - 빅데이터 개요 및 활용

3) 조직의 구성

② 인력 구성을 위한 체크리스트

- 비즈니스 및 IT 전문가의 조합으로 구성 되어야 하는가?
- 어떤 경험과 어떤 스킬을 갖춘 사람으로 구성해야 하는가?
- 통계적 기법 및 분석 모델링 전문인력을 별도로 구성해야 하는가?
- 전사 비즈니스를 커버하는 인력이 없다면?
- 전사 분석업무에 대한 적합한 인력 규모는 어느 정도인가?

③ 구성 인력과 필요 역량

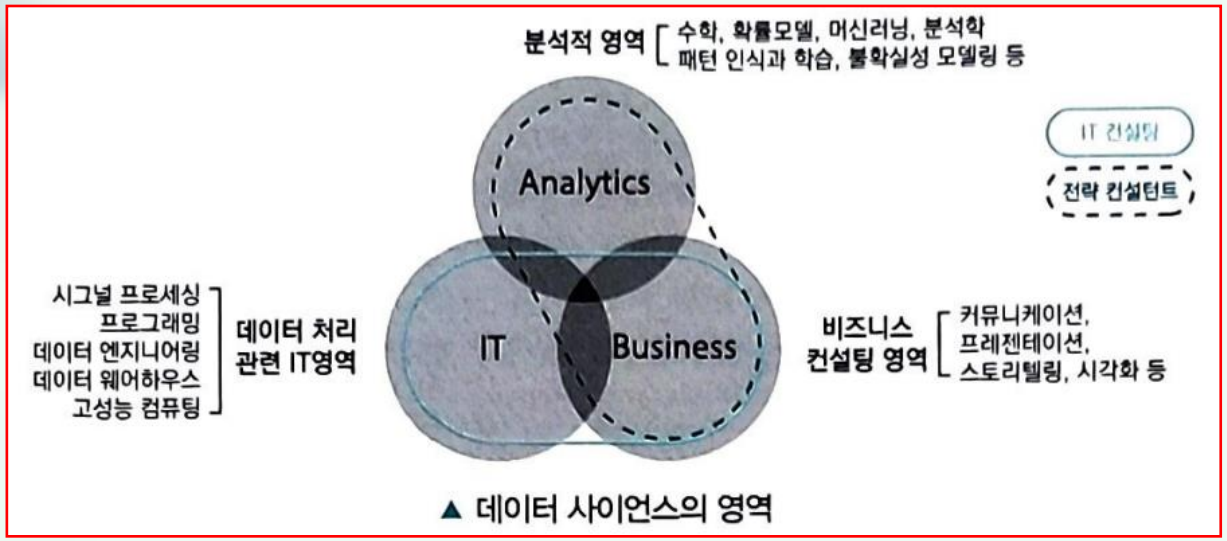
- 비즈니스를 이해하고 있는 인력
- 분석에 필요한 컴퓨터 공학적인 기술을 이해하고 있는 인력
- 통계를 이용한 다양한 분석기법을 활용할 수 있는 분석 지식을 갖춘 인력
- 조직 내 분석 문화확산을 위한 변화 관리 인력
- 분석조직뿐 아니라 관련 부서 조직원의 분석 역량 향상을 위한 교육담당 인력

1. 빅데이터의 이해 - 빅데이터 개요 및 활용

4) 데이터 사이언스 역량

데이터 사이언스는 정형, 비정형 형태를 포함한 다양한 데이터로부터 지식과 인사이트를 추출하는 데 과학적 방법론, 프로세스, 알고리즘, 시스템을 동원하는 융합 분야이다.

● 데이터 사이언스는 데이터를 통해 실제 현상을 이해하고 분석하는 데 필요한 통계학, 데이터분석, 기계학습과 연관된 방법론을 통합하는 개념으로 정의되기도 한다.



▲ 데이터 사이언스의 영역

① 데이터 사이언스의 기능

▶ 비즈니스 성과를 좌우하는 핵심이슈에 답할 수 있다.

▶ 사업의 성과를 견인해 나갈 수 있다.

1. 빅데이터의 이해 - 빅데이터 개요 및 활용

4) 데이터 사이언스 역량

② 데이터 사이언스 실현을 위한 인문학적 요소

- ▶ 스토리텔링 능력
- ▶ 커뮤니케이션 능력
- ▶ 창의력과 직관력
- ▶ 비판적 시각과 열정

③ 데이터 사이언스의 한계

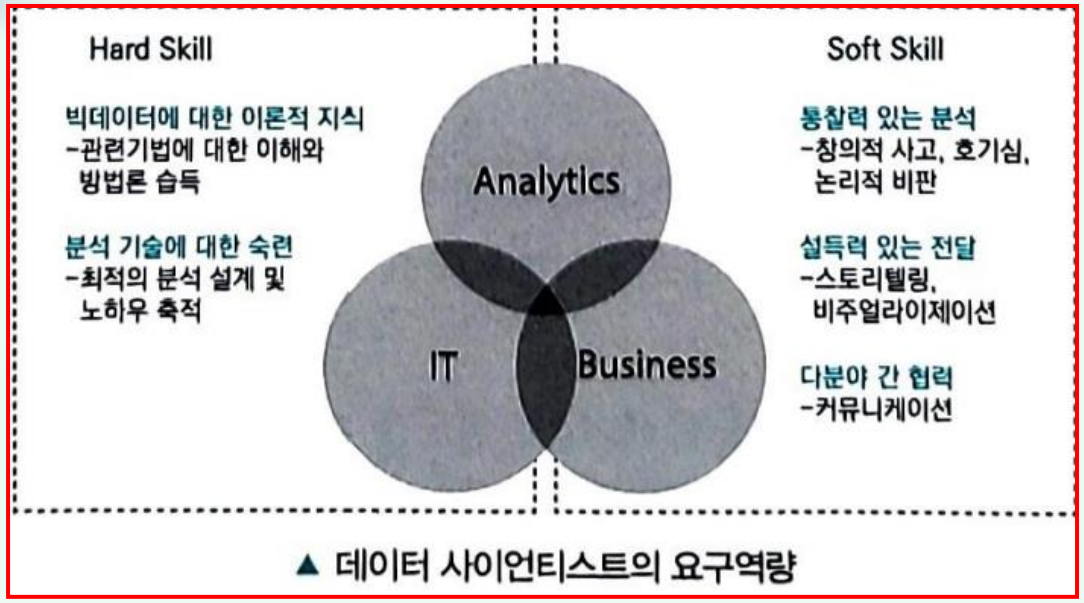
- ▶ 분석 과정에서 가정과 같이 인간의 해석이 개입되는 단계가 불가피하다.
- ▶ 분석 결과를 바라보는 사람에 따라 서로 다른 해석과 결론을 내릴 수 있다.
- ▶ 아무리 정량적인 분석이라 할지라도 모든 분석은 가정에 근거한다.

1. 빅데이터의 이해 - 빅데이터 개요 및 활용

5) 데이터 사이언티스트

; 데이터에 대한 이론적 지식과 숙련된 분석 기술을 바탕으로 통찰력과 전달력 및 협업 능력을 갖춘 데이터 분야 전문가이다.

- 데이터의 다각적 분석을 통해 인사이트를 도출하고 이를 조직의 전략 방향 제시에 활용할 수 있는 기획자이기도 하다.
- 문제를 집중적으로 파고들어 질문을 찾고, 검증 가능한 가설을 세워야 한다.



▲ 데이터 사이언티스트의 요구역량

데이터 사이언티스트 : 데이터의 근원을 찾고 대용량의 복잡한 데이터를 구조화하며 서로 연결하는 역할

1. 빅데이터의 이해 - 빅데이터 개요 및 활용

6) 빅데이터 관련 직업군

① 데이터 분석가(Data Analyst)

- ▶ 데이터 분석을 기반으로 비즈니스에서 최적의 의사결정을 내릴 수 있는 인사이트를 분석하여 제공하는 업무
- ▶ 유관 부서들과 업무적 연계를 통한 데이터 수집 및 분석 업무
- ▶ 비즈니스 도메인 지식, 데이터 시각화 능력, 데이터 분석을 위한 언어(Python, R 등) 활용 능력 및 통계 지식 능력, SQL 활용 능력 필요

② 데이터 사이언티스트(Data Scientist)

- ▶ 통계, 머신러닝, AI에 대한 지식을 활용하여 데이터 내의 인사이트를 발견하는 업무
- ▶ 예측 모델링, 추천 시스템 등을 개발하여 비즈니스 의사결정에 대한 유의미한 결정을 제공
- ▶ 머신러닝 모델 구축을 위한 기본적인 코딩 스킬, 데이터 분석을 위한 통계적 지식, SQL 활용 능력 필요

1. 빅데이터의 이해 - 빅데이터 개요 및 활용

6) 빅데이터 관련 직업군

③ 데이터 엔지니어(Data Engineer)

- ▶ 데이터 플랫폼과 파이프라인(데이터가 적절한 시기에 적절한 방법으로 직원들에게 흘러갈 수 있도록 해주는 시스템) 아키텍쳐를 개발하고 운영하는 업무
- ▶ 비즈니스를 이해하고 대량의 데이터 세트를 가공하고 대용량 데이터 분산 처리 시스템을 개발
- ▶ 코딩 스킬, 빅데이터 분산처리 시스템 구조에 대한 이해 능력 필요

④ 데이터 아키텍트(Data Architect)

- ▶ 전사 데이터 관리 시스템을 위한 데이터 구조 및 관리 체계를 설계하는 업무
- ▶ 회사의 잠재적인 데이터 소스(내부 및 외부)를 평가한 후 통합하고 중앙 집중화하며, 보호 및 관리하는 계획을 설계하는 업무
- ▶ 데이터 요건분석, 데이터 표준화, 데이터 모델링, 데이터베이스 설계와 이용에 대한 전문지식 및 실무적 수행 능력 필요

1. 빅데이터의 이해 - 빅데이터 개요 및 활용

6) 빅데이터 관련 직업군

⑤ 머신러닝 엔지니어(Machine Learning Engineer)

▶ 머신러닝 세부 기술(음성 인식, 영상 인식, 자연어 처리)과 서비스를 개발하는 업무

▶ 데이터를 활용하여 현실 문제를 머신러닝 모델로 해결하는 업무

▶ 수학 및 통계분석 능력, 주요 언어를 활용한 프로그램 구현 능력, 딥러닝 알고리즘 이해 및 구현 능력 필요

7) 데이터 거버넌스(Data Governance)

● 기업에서 사용하는 데이터의 가용성, 유용성, 통합성, 보안성을 관리하기 위한 정책과 프로세스를 다루며 프라이버시 보안성, 데이터 품질, 관리 규정 준수를 강조하는 모델을 의미한다.

● 데이터 거버넌스의 구성 요소는 원칙, 조직, 프로세스이다.

- 원칙(Principle) : 데이터를 관리하기 위한 규칙
- 조직(Organization) : 데이터를 관리할 수 있는 조직의 역할과 책임
- 프로세스(Process) : 데이터 관리를 위한 활동 과정

1. 빅데이터의 이해 - 빅데이터 개요 및 활용

8) 데이터 분석 수준 진단 결과

분석 준비도와 분석 성숙도 진단에 따른 데이터 분석 수준 진단 결과는 준비형, 정착형, 도입형, 확산형의 4가지로 나뉜다.

▶ 준비형

- 낮은 준비도, 낮은 성숙도, 사전 준비 필요

▶ 정착형

- 낮은 준비도, 높은 성숙도, 분석의 정착 필요

▶ 도입형

- 높은 준비도, 높은 성숙도, 데이터 분석 도입 가능

▶ 확산형

- 높은 준비도, 높은 성숙도, 지속적 확산 가능

1. 빅데이터의 이해 - 빅데이터 개요 및 활용

개념 체크

1. 다음 중 빅데이터 조직 구성에 대한 설명으로 틀린 것은?

① 전문인력을 중심으로 전사 또는 특정 부서 내 조직으로 구성한다.

② 집중형, 기능형, 분산형 형태의 조직으로 구성할 수 있다.

③ 조직원의 분석 역량 향상을 위한 교육담당 인력도 필요하다.

④ 기능형 조직의 경우 분석전담조직이 필요하지만, 분산형 조직은 필요치 않다.

통계학이나 분석 방법에 대한 지식과 분석 경험이 있는 분석인력을 중심으로 전사 또는 특정 부서 내 조직으로 구성하여 운영한다.

전사 분석 업무를 별도의 전담조직에서 수행하는 집중형, 각 현업 부서에서 분석 업무를 직접 수행하는 기능형, 분석 전문 인력을 현업 부서에 배치하여 분석 업무를 수행하는 분산형 조직으로 구성할 수도 있다.

분석조직 뿐만 아니라 관련 부서 조직원의 분석 역량 향상 을 위한 교육 담당 인력도 필요하다.

2. 다음 중 데이터 사이언스의 한계에 대한 설명으로 잘못된 것은?

① 아무리 정량적인 분석이라 할지라도 모든 분석은 가정에 근거한다.

② 분석 과정에서 가정과 같이 인간의 해석이 개입되는 단계가 불가피하다.

③ 분석 결과를 바라보는 사람에 따라 서로 다른 해석과 결론을 내릴 수 있다.

④ 데이터에서 상관관계를 발견하더라도 인과관계를 찾아내지 못한다면 신뢰할 수 없다.

연관성 분석과 같이 데이터 분석을 통해서 발견되는 패턴만으로도 충분히 의미를 갖는 경우도 있으며, 반드시 인과관계를 찾아내야지만 데이터 분석 결과를 신뢰할 수 있는 그런 시대는 지났다.

1. 빅데이터의 이해 - 빅데이터 개요 및 활용

개념 체크

3. 다음 중 통계, 머신러닝, 최적화 등 다양한 기술을 활용하여 데이터에 기반한 서비스를 개발하거나 수익 향상을 위한 의사결정을 돕는 전문가를 뜻하는 것은?

① 데이터 사이언티스트

② 데이터 엔지니어

③ 데이터 분석가

④ 프로그래머

데이터 사이언티스트(Data Scientist)

통계, 머신러닝, 최적화, AI 등 다양한 기술을 활용하여 데이터 기반에 서비스를 개발하거나 수익 향상을 위한 의사결정을 돕는 직군

데이터 엔지니어

데이터 플랫폼 및 데이터 파이프라인 구조를 개발하고 운영하는 직군

데이터 분석가 : 데이터 분석 보고서 및 시각화 자료를 통해 인사이트를 도출하고, 비즈니스 결정을 돕는 직군

프로그래머 : 프로그램을 개발하는 직군

4. 데이터 거버넌스의 구성 요소에 해당하지 않는 것은?

① 원칙 ② 조직

③ 프로세스 ④ 분석

데이터 거버넌스(Data Governance)

기업에서 사용하는 데이터의 가용성, 유용성, 통합성, 보안성을 관리하기 위한 정책과 프로세스를 다루면 프라이버시 보안성, 데이터 품질, 관리 규정 준수를 강조하는 모델을 의미한다.

데이터 거버넌스의 구성 요소는 원칙, 조직, 프로세스이다.

1. 빅데이터의 이해 - 빅데이터 개요 및 활용 예상문제

01 다음 중 데이터에 대한 설명으로 틀린 것은?

① 데이터는 일반적으로 정형, 비정형, 반정형 데이터로 구분된다.

② 비정형 데이터는 텍스트, 음성, 영상 등 특수한 데이터 이다.

③ 정형 데이터는 흔히 볼 수 있는 주로 숫자로 구성된 데이터이다.

④ 정형 데이터는 비정형 데이터보다 품질이 우수하며 다양한 분석이 가능하다.

정형, 비정형, 반정형 데이터의 구분은 품질과는 무관하며, 정형 데이터보다 비정형 데이터가 일반적으로 다양한 분석을 시도하기에 유리하다.

02 다음 중 정성적 데이터로 옳은 것은?

① 대통령에 대한 국민들의 인식

② 서울에서 제주까지 비행시간

③ 한국인의 평균 수명

④ 국내 인구 증가율

대통령에 대한 국민들의 인식은 세부 분야별로 나누어 5점 척도나 7점 척도로 측정할 수 있지만 일반적으로 사람에

03 다음 중 반정형 데이터가 아닌 것은?

① XML File ② JSON File

③ TEXT File ④ HTML File

XML이나 JSON, HTML File은 기본 형식은 유지하면서 담고 있는 내용에 대해서는 유연성을 허용하는 반정형 데이터 이며, TEXT 파일의 경우 일정한 형식을 요하지 않는 비정형 데이터에 해당한다.

04 다음 중 비정형 데이터가 아닌 것은?

① 동영상 ② 이미지

③ 음성 ④ 전화번호

전화번호는 일반적으로 숫자로 구성이 되면, 이는 정형 데이터에 해당한다.

05 다음 중 정보의 특징이 아닌 것은?

① 적정성 ② 일관성

③ 관련성 ④ 적시성

정보는 정확성, 적시성, 적정성, 관련성, 적당성의 특징을 갖는다.

1. 빅데이터의 이해 - 빅데이터 개요 및 활용 예상문제

06 다음 중 지식의 피라미드를 순서대로 나열한 것은?

① 데이터 → 정보 → 지식 → 지혜

② 데이터 → 정보 → 지혜 → 지식

③ 데이터 → 지혜 → 정보 → 지식

④ 데이터 → 지식 → 지혜 → 정보

DIKW 피라미드(지식의 피라미드)는 최하위 데이터 단계부터 데이터(Data) -> 정보(Information) -> 지식(Knowledge) ->지혜(Wisdom)의 순서를 따른다.

07 다음 중 지식창조 매커니즘의 단계가 아닌 것은?

① 표출화(Externalization)

② 내면화(Internalization)

③ 통합화(Integration)

④ 공통화(Socialization)

지식창조 메커니즘은 공통화, 표출화, 연결화, 내면화 총 4단계로 구성되어 있다.

08 다음 중 데이터 웨어하우스의 특징이 아닌 것은?

① 주제지향성(Subject-orientation)

② 휘발성(Volatilization)

③ 통합성(Integration)

④ 시계열성(Time-variant)

데이터 웨어하우스의 특정의 주제지향성, 비휘발성, 통합성, 시계열성 이다.

09 다음 중 데이터 웨어하우스의 구성요소가 아닌 것은?

① 데이터 모델(Data Model)

② 데이터 전처리(Data Pre-processing)

③ ETL(Extract, Transform, Load)

④ ODS(Operational Data Store)

데이터 웨어하우스는 데이터 모델, ETL, ODS, DW 메타 데이터, OLAP, 데이터마이닝, 분석 도구, 경영기반 솔루션으로 구성된다.

10 다음 중 빅데이터의 주요 특징으로 틀린 것은?

① 다양성　② 대용량성

③ 신속성　④ 일관성

1. 빅데이터의 이해 - 빅데이터 개요 및 활용 예상문제

11 다음 중 빅데이터를 활용할 때 얻을 수 있는 가치가 아닌 것은?

① 마케팅 효과 극대화

② 제품 생산 비용 절감

③ 비즈니스 의사결정의 고도화

④ 고객 개인정보 활용을 통한 통제

빅데이터 활용 시 고객의 개인정보는 보호되어야 하며, 고객을 통제하는 수단으로 사용하는 것은 부적합하다.

12 다음 중 빅데이터 활용에 필요한 3요소로 옳은 것은?

① 자원, 인력, 프로세스

② 자원, 기술, 인력

③ 기술, 인력, 프로세스

④ 자원, 기술, 프로세스

빅데이터 활용에 필요한 3요소는 자원(데이터), 기술, 인력이다.

13 다음 중 빅데이터가 만들어 낸 변화로 틀린 것은?

① 사전처리에서 사후처리로 변화

② 인과관계에서 상관관계로 변화

③ 전수조사에서 표본조사로 변화

④ 데이터의 질보다 양의 중요도 증가

빅데이터의 출현으로 인해 기존의 표본조사(샘플링)를 하던 방식이 전수조사를 하는 방식으로 변화되고 있다.

14 다음 중 마이데이터가 등장한 시점으로 옳은 것은?

① 데이터 통합 시대

② 데이터 분석 시대

③ 데이터 연결 시대

④ 데이터 권리 시대

마이데이터는 개인이 자신의 데이터를 자신을 위해서 사용한다는 사상을 담은 것으로 데이터 권리 시대에 해당한다.

15 다음 데이터 사이언티스트에 대한 요구역량 중 Soft Skill이 아닌 것은?

① 분석 기술에 대한 숙련

② 설득력 있는 전달

③ 통찰력 있는 분석

④ 다분야 간 협력

감사합니다.