

PART

01

데이터 이해

1.1

1.2

1.3

1.4

1.5

1.6

1.7

1.8

1.9

1.10

1.11

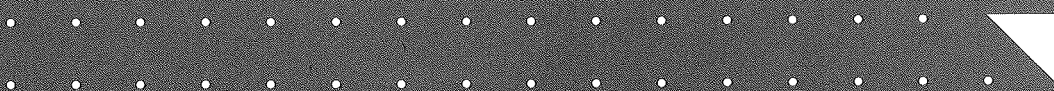
1.12

1.13

1장 데이터의 이해

2장 데이터의 가치와 미래

3장 가치창조를 위한 데이터 사이언스와
전략 인사이트



1장. 데이터의 이해

데이터 정보, 데이터베이스 정의와 특징, 데이터베이스 활용에 대해 살펴본다.

2장. 데이터의 가치와 미래

빅데이터의 이해와 가치 그리고 그 영향, 그리고 빅데이터가 활용되는 비즈니스 모델과 미래의 빅데이터에 대해 살펴본다.

3장. 가치창조를 위한 데이터 사이언스와 전략 인사이트

빅데이터 분석이 사회와 기업에 미치는 영향, 빅데이터를 통한 기업의 전략적 인사이트의 이해 그리고 빅데이터의 발전 방향에 대해 살펴본다.

Learning Map

어떤 것을 학습하게 될지 살펴보자!

1장

데이터의 이해

- 데이터와 정보
- 데이터베이스 정의와 특징
- 데이터베이스 활용

2장

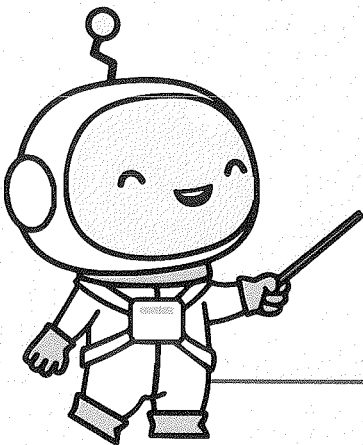
데이터의 가치와 미래

- 빅데이터의 이해
- 빅데이터의 가치와 영향
- 비즈니스 모델
- 위기요인과 통제방안
- 미래의 빅데이터

3장

가치창조를 위한 데이터 사이언스와 전략 인사이트

- 빅데이터 분석과 전략 인사이트
- 전략 인사이트 도출을 위한 필요 역량
- 빅데이터 그리고 데이터 사이언스의 미래



데이터의 이해

1 DAY



출제 포인트

제1장은 데이터 분석 전문가가 알아야 하는 기본소양에 대한 내용입니다!
중요한 용어의 의미와 쓰임에 대해 이해해야 하는 것이 포인트!



○ 학습 목표

- 데이터 정의에 대해 이해한다.
- 데이터베이스 정의와 특징을 이해한다.
- 데이터베이스 활용에 대해 이해한다.

○ 눈높이 체크

✓ 데이터의 정의를 알고계신가요?

데이터라는 단어를 한 번도 못 들어본 분은 없을 것입니다. 옥스퍼드 대사전에서 데이터는 “추론과 추정의 근거를 이루는 사실”이라고 정의하고 있습니다. 1940년대 이후 컴퓨터시대가 시작되면서 자연과학 뿐만 아니라 경영학, 통계학 등 다양한 사회과학이 진일보하면서 데이터의 의미는 과거의 관념적이고 추상적인 개념에서 기술적이고 사실적인 의미로 변화되고 있습니다.

✓ 데이터와 정보 그리고 지식의 관계는 어떻게 이루어질까요?

데이터 → 정보 → 지식 → 지혜로 발전하면서 데이터는 추론·예측·전망·추정을 위한 근거가 됩니다.

✓ 데이터베이스를 다루어 본 적이 있으신가요?

현재 우리는 다양한 데이터베이스 시스템을 활용하고 있습니다. 간단하게는 Access, MSSQL, mySQL, 오라클 등을 통해 데이터베이스를 접해보셨다면 본 강의를 더욱 쉽게 이해하실 것입니다. 혹시 데이터베이스를 사용해 보지 못하셨더라도 엑셀을 잘 사용하신다면 이해하실 수 있을 것입니다.

1절

1장 데이터의 이해

데이터와 정보



출 제
포인트

정성적 데이터와 정량적 데이터의 차이점을 묻는
문제가 간혹 출제되니 꼭 짚고 넘어가세요~



1

데이터의 정의와 특성

예시

수요조사나 실험, 검사, 측정 등을 통해 데이터를 수집, 축적하고 다양한 방법으로 분석하여 간단한 마케팅 리포트부터 심도있는 논문, 미래 예측을 위한 경영전략 또는 정책을 수립하는 일련의 가치 창출 과정에서 가장 기초를 이루는 것을 데이터라 한다.

가. 데이터의 정의

- 1) 데이터(Data)라는 용어는 1646년 영국 문헌에 처음 등장하였으며 라틴어인 Dare(주다)의 과거분사형으로 '주어진 것'이란 의미로 사용되었다.
- 2) 1940년대 이후 컴퓨터 시대 시작과 함께 자연과학뿐만 아니라 경영학, 통계학 등 다양한 사회 과학이 진일보하며, 데이터의 의미는 과거의 관념적이고 추상적인 개념에서 기술적이고 사실적인 의미로 변화되었다.
- 3) 데이터는 추론과 추정의 근거를 이루는 사실이다.(옥스퍼드 대사전)
- 4) 데이터는 단순한 객체로서의 가치뿐만 아니라 다른 객체와의 상호관계 속에서 가치를 갖는 것으로 설명되고 있다.

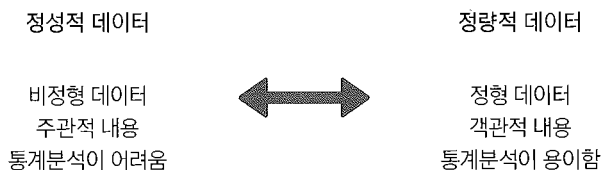
나. 데이터의 특성

구분	특성
존재적 특성	객관적 사실(Fact, Raw Material)
당위적 특성	추론·예측·전망·추정을 위한 근거(Basis)

2

데이터의 유형

구분	형태	예	특징
정성적 데이터 (Qualitative Data)	언어, 문자 등	회사 매출이 증가함 등	저장·검색·분석에 많은 비용이 소모됨
정량적 데이터 (Quantitative Data)	수치, 도형, 기호 등	나이, 몸무게, 주가 등	정형화된 데이터로 비용 소모가 적음



- 데이터는 지식경영의 핵심 이슈인 암묵지(暗黙知, Tacit Knowledge)와 형식지(型式知, Explicit Knowledge)의 상호작용에 있어 중요한 역할을 한다.(Polany, 1966)

지식경영의 핵심 이슈

구분	의미	예	특징	상호작용
암묵지	학습과 경험을 통해 개인에게 체화되어 있지만 걸어로 드러나지 않는 지식	김장김치 담그기, 자전거 타기	사회적으로 중요하지 만 다른 사람에게 공유되기 어려움	공통화, 내면화
형식지	문서나 매뉴얼처럼 형상화된 지식	교과서, 비디오, DB	전달과 공유가 용이함	표출화, 연결화

암묵지와 형식지의 상호작용관계

- 1단계 : 공통화
(암묵지를 타인에게 알려주기)
- 2단계 : 표출화
(암묵지를 책 등 형식지로 만들기)
- 3단계 : 연결화
(책 등에 자신이 아는 새로운 지
식 추가하기)
- 4단계 : 내면화
(책 등을 보고 타인들이 암묵적
지식 습득)

참고

- 암묵지 :
개인에게 축적된 내면화(Internalization)된 지식 → 조직의 지식으로 공통화(Socialization)
- 형식지 :
언어, 기호, 숫자로 표출화(Externalization)된 지식 → 개인의 지식으로 연결화(Combination)



가. DIKW의 정의

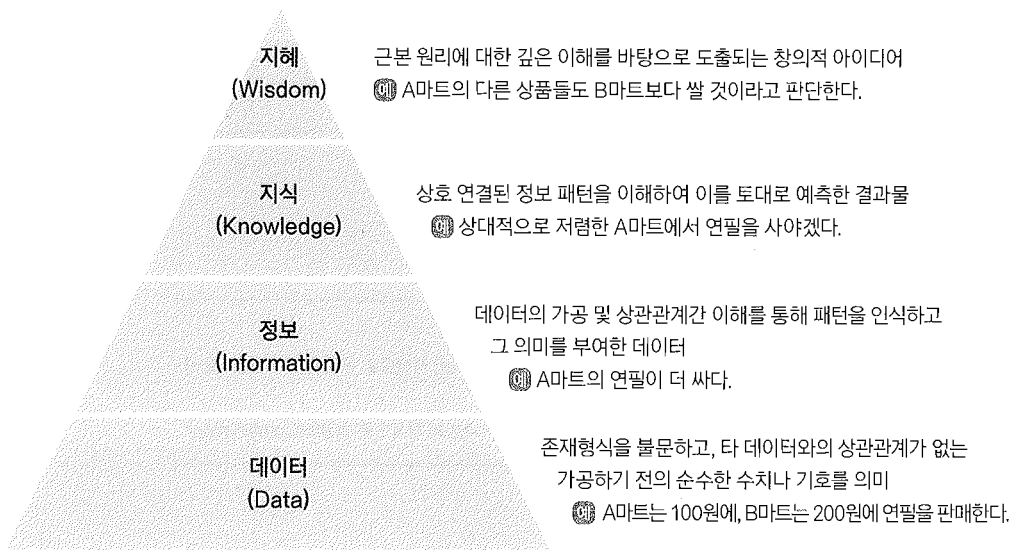
구분	특성
데이터(Data)	개별 데이터 자체로는 의미가 중요하지 않은 객관적인 사실
정보(Information)	데이터의 가공, 처리와 데이터간 연관관계 속에서 의미가 도출된 것
지식(Knowledge)	데이터를 통해 도출된 다양한 정보를 구조화하여 유의미한 정보를 분류하고 개인적인 경험을 결합시켜 고유의 지식으로 내재화된 것
지혜(Wisdom)	지식의 축적과 아이디어가 결합된 창의적인 산물

나. DIKW 피라미드

- DIKW 피라미드에서는 데이터, 정보, 지식을 통해 최종적으로 지혜를 얻어내는 과정을 계층구조로 설명하고 있다.

출 제
포인트

DIKW 각각의 정의를 묻는 문제, 예시에 대한 문제가 자주 출제되니 꼼꼼히 체크합니다.



2절

1장 데이터의 이해

데이터베이스 정의와 특징

1

용어의 연혁

출 처	내 용
1950년대	미국에서 군대의 군비상황을 집중 관리하기 위하여 컴퓨터 도서관을 설립하면서 데이터(Data)의 기지(Base)라는 뜻의 데이터베이스가 탄생
1963년 6월	미국 'SDC'가 개최한 심포지엄에서 데이터베이스라는 용어 공식사용 데이터베이스 초기 개념인 '대량의 데이터를 축적하는 기지'
1963년	GE의 C.바크만은 데이터베이스 관리 시스템인 IDS 개발
1965년	2차 심포지엄에서 시스템을 통한 체계적 관리와 저장 등의 의미를 담은 '데이터베이스 시스템'이라는 용어 등장
1970년대 초반	유럽에서 '데이터베이스'라는 단일어가 일반화 됨
1975년	미국의 CAC가 KORSTIC을 통해 서비스되면서 우리나라에서 데이터베이스 이용이 이루어짐
1980년	KORSTIC이 해외 전문 데이터베이스를 확충하여 'TECHNOLINE' 이라는 온라인 정보검색 서비스를 개시하여 본격적인 데이터베이스 서비스 시대를 맞이함
1980년대 중반	국내의 데이터베이스 관련 기술의 연구, 개발



출 제 포인트

용어와 연결하여 내용을 굳이 다 외울 필요는 없습니다. 데이터베이스가 어떻게 정의 되는지 흐름을 파악하는 것이 중요합니다.



2

데이터베이스의 정의

구분	특성
1차 개념확대 정형데이터 관리	EU <데이터베이스의 법적 보호에 관한 지침> 국내 '저작권법'
2차 개념확대 빅데이터의 출현으로 비정형데이터 포함	체계적이거나 조직적으로 정리되고 전자식 또는 기타 수단으로 개별적으로 접근할 수 있는 독립된 저작물, 데이터 또는 기타 소 재의 수집물 소재를 체계적으로 배열 또는 구성한 편집 물로서 개별적으로 그 소재에 접근하거나 그 소재를 검색할 수 있도록 한 것 국내 '컴퓨터 용어사전' 동시에 복수의 적용 업무를 지원할 수 있도 록 복수 이용자의 요구에 대응해서 데이터 를 받아들이고 저장, 공급하기 위하여 일정 한 구조에 따라서 편성된 데이터의 집합 국내 'Wikipedia' 관련된 레코드의 집합, 소프트웨어로는 데 이터베이스관리시스템(DBMS)을 의미 국내 '데이터분석 전문가 가이드' 문자, 기호, 음성, 화상, 영상 등 상호 관련 된 다수의 콘텐츠를 정보 처리 및 정보통신 기기에 의하여 체계적으로 수집·축적하여 다양한 용도와 방법으로 이용할 수 있도록 정리한 정보의 집합체



출 제 포인트

데이터베이스의 일반적인 특징은 자주 출제가 되는 부분이니 반
드시 체크해주세요. (특징 중 맞거나 틀린 것을 선택하는 문항)



3

데이터베이스의 특징

가. 데이터베이스의 일반적인 특징

데이터베이스 특징	설 명
통합된 데이터 Integrated Data	동일한 내용의 데이터가 중복되어 있지 않다는 것을 의미. 데이터 중복은 관리상의 복잡한 부작용을 초래
저장된 데이터 Stored Data	자기 디스크나 자기 테이프 등과 같이 컴퓨터가 접근할 수 있는 저장 매체에 저장되는 것을 의미. 데이터베이스는 기본적으로 컴퓨터 기술을 바탕으로 한 것
공용 데이터 Shared Data	여러 사용자가 서로 다른 목적으로 데이터를 공동으로 이용한다는 것을 의미. 대용량화되고 구조가 복잡한 것이 보통
변화되는 데이터 Changable Data	데이터베이스에 저장된 내용은 곧 데이터베이스의 현 시점에서의 상태를 나타냄. 다만 이 상태는 새로운 데이터의 삽입, 기존 데이터 의 삭제, 갱신으로 항상 변화하면서도 항상 현재의 정확한 데이터를 유지해야 함

나. 데이터베이스의 다양한 측면에서의 특징

측면	특성
정보의 축적 및 전달 측면	<ul style="list-style-type: none"> • 기계 가독성: 일정한 형식에 따라 컴퓨터 등의 정보처리기가 읽을 수 있음 • 검색 가독성: 다양한 방법으로 필요한 정보를 검색 • 원격 조작성: 정보통신망을 통하여 원거리에서도 즉시 온라인을 이용
정보 이용 측면	<ul style="list-style-type: none"> • 이용자의 정보 요구에 따라 다양한 정보를 신속하게 획득 • 원하는 정보를 정확하고 경제적으로 찾아낼 수 있다는 특성
정보 관리 측면	<ul style="list-style-type: none"> • 정보를 일정한 질서와 구조에 따라 정리, 저장, 검색, 관리 할 수 있도록 하여 방대한 양의 정보를 체계적으로 축적하고 새로운 내용의 추가나 갱신이 용이
정보기술 발전 측면	<ul style="list-style-type: none"> • 데이터 베이스는 정보처리, 검색·관리 소프트웨어, 관련 하드웨어, 정보 전송을 위한 네트워크 기술의 발전을 견인할 수 있음
경제·산업 측면	<ul style="list-style-type: none"> • 다양한 정보를 필요에 따라 신속하게 제공·이용할 수 있는 인프라라는 특성을 가지고 있어 경제, 산업, 사회 활동의 효율성을 제고하고 국민의 편의를 증진하는 수단으로서 의미를 가짐

MEMO

데이터베이스의 활용



출 제
포인트

약어의 의미를 잘못 설명한 보기를 찾는 문제가
출제되니 기본적인 내용은 숙지합니다.



1 기업내부 데이터베이스

정보통신망 구축이 가속화되면서 1990년대의 기업내부 데이터베이스는 기업 경영 전반에 관한 인사, 조직, 생산, 영업 활동 등을 포함한 모든 자료를 연계하여 일관된 체계로 구축, 운영하는 경영 활동의 기반이 되는 전사 시스템으로 확대되었다.

가. 1980년대 기업내부 데이터베이스

- OLTP(On-Line Transaction Processing) : 호스트 컴퓨터와 온라인으로 접속된 여러 단말 간의 처리 형태의 하나이다. 여러 단말에서 보내온 메시지에 따라 호스트 컴퓨터가 데이터베이스를 액세스하고, 바로 처리 결과를 돌려보내는 형태를 말한다. 즉, 데이터베이스의 데이터를 수시로 갱신하는 프로세싱을 의미한다. 주문입력시스템, 재고관리시스템 등 현업의 거의 모든 업무는 이와 같은 성격을 띄고 있다. <참조 : (컴퓨터인터넷IT용어대사전, 2011.1.20, 일진사)>
- OLAP(On-Line Analytical Processing) : 정보 위주의 분석 처리를 의미하며, 다양한 비즈니스 관점에서 쉽고 빠르게 다차원적인 데이터에 접근하여 의사 결정에 활용할 수 있는 정보를 얻을 수 있게 해주는 기술이다. OLTP에서 처리된 트랜잭션 데이터를 분석해 제품의 판매 추이, 구매 성향 파악, 재무 회계 분석 등을 프로세싱하는 것을 의미한다. OLTP가 데이터 갱신 위주라면, OLAP는 데이터 조회 위주라고 할 수 있다. <참조 : (컴퓨터인터넷IT용어대사전, 2011.1.20, 일진사)>



출 제 포인트

데이터 분석 준전문가 과정에서 비교하는 부분은 잘 출제되지 않으나, 꼭 이해하고 넘어가세요.



OLTP와 OLAP의 비교

구 분	OLTP	OLAP
데이터 구조	복잡	단순
데이터 갱신	동적으로 순간적	정적으로 주기적
응답 시간	수 초 이내	수 초에서 몇 분 사이
데이터 범위	수 십일 전후	오랜 기간 저장
데이터 성격	정규적인 핵심 데이터	비정규적인 읽기 전용 데이터
데이터 크기	수 기가 바이트	수 테라 바이트
데이터 내용	현재 데이터	요약된 데이터
데이터 특성	트랜잭션 중심	주제 중심
데이터 액세스 빈도	높음	보통
질의 결과 예측	주기적이며 예측 가능	예측하기가 어렵다

참조 : 소설처럼 읽는 DB 모델링 이야기 (영진닷컴)

나. 2000년대 기업내부 데이터베이스

- CRM(Customer Relationship Management) : ‘고객관계관리’라고 하며, 기업이 고객과 관련된 내·외부 자료를 분석·통합해 고객 중심 자원을 극대화하고, 이를 토대로 고객특성에 맞게 마케팅 활동을 계획·지원·평가하는 과정이다. CRM은 최근에 등장한 데이터베이스 마케팅(DB marketing)의 일대일 마케팅(One-to-One marketing), 관계 마케팅(Relationship marketing)에서 진화한 요소들을 기반으로 등장하게 되었다.

〈참조 : CRM (시사상식사전, 박문각)〉

- SCM(Supply Chain Management) : ‘공급망 관리’를 뜻하는 말로, 기업에서 원재료의 생산·유통 등 모든 공급망 단계를 최적화해 수요자가 원하는 제품을 원하는 시간과 장소에 제공하는 것이다. SCM은 부품 공급업체와 생산업체 그리고 고객에 이르기까지 거래관계에 있는 기업들 간 IT를 이용한 실시간 정보공유를 통해 시장이나 수요자들의 요구에 기민하게 대응토록 지원하는 것이다. 〈참고 : SCM [Supply Chain Management](시사 상식사전, (박문각))

다. 각 분야별 내부 데이터베이스

1) 분야별 데이터베이스 개념

가) 제조부문

- 제조업의 데이터베이스 기술 적용은 2000년을 기점으로 전환되었다.
- 클라이언트/서버 기반의 내부 정보시스템에서 웹기반의 데이터베이스로 전환되고 있다.
- 대기업을 위주로 ERP에서 CRM으로 발전하고 있다.
- 대기업은 중소기업과의 협업으로 인해 중소기업에 투자를 확대할 필요성을 인지하고 있으며 RTE를 통한 협업적 IT화의 비중을 확대하고 있다.

나) 금융부문

- 1998년 IMF 이후, 금융부문은 업무 프로세스 효율화나 통합시스템 구축으로 확산되었다.
- 2000년대 초반, EAI, ERP, e-CRM을 통한 정보 공유 및 통합, 그리고 고객 정보의 전략적 활용이 시작되었다.
- 2000년대 중반, DW(Data Warehouse) 도입을 통한 DB활용 마케팅이 강화되었고, DW를 위한 최적화와 BI 기반의 시스템 구축이 급속도로 퍼지게 되었다.
- 바젤2 등의 대형 프로젝트가 마무리 되면서 향후 EDW(Enterprise Data Warehouse)의 확장이 DB 시장 확장에 기여하고 있다.

다) 유통부문

- 2000년 이후, IT 환경 변화에 따라 CRM과 SCM의 구축이 활발하게 진행되고 있다.
- 상거래를 위한 인프라와 KMS를 위한 백업시스템 구축도 함께 진행되었다.
- RFID의 등장으로 유비쿼터스 시대에 접어들었다.

2) 분야별 데이터베이스 소개

분 야	내 용
제조분야	<ul style="list-style-type: none"> • ERP(Enterprise Resource Planning) : 인사·재무·생산 등 기업의 전 부문에 걸쳐 독립적으로 운영되던 각종 관리 시스템의 경영자원을 하나의 통합 시스템으로 재구축함으로써 생산성을 극대화하려는 경영혁신기법을 의미한다. • BI(Business Intelligence) : 비즈니스 인텔리전스(Business Intelligence, BI)란 기업이 보유하고 있는 수많은 데이터를 정리하고 분석해 기업의 의사결정에 활용하는 일련의 프로세스를 말한다.

분야	내용
제조분야	<ul style="list-style-type: none"> • CRM(Customer Relationship Management) : '고객관계관리'라고 한다. 기업이 고객과 관련된 내외부 자료를 분석·통합해 고객 중심 자원을 극대화하고 이를 토대로 고객특성에 맞게 마케팅 활동을 계획·지원·평가하는 과정이다. • RTE(Real-Time Enterprise) : 회사의 주요 경영정보를 통합관리하는 실시간 기업의 새로운 기업경영시스템이다. 전사적 자원관리(ERP), 판매망관리(SCM), 고객관리(CRM) 등 부분별 전산화에서 한발 나아가 회사 전 부문의 정보를 하나로 통합함으로써 경영자의 빠른 의사결정을 이끌어 내려는 목적에서 만들어졌으며 기업활동이 글로벌화되고 기술의 발전으로 제품 수명이 짧아지는 현실에 대응되고 있다.
금융부문	<ul style="list-style-type: none"> • EAI(Enterprise Application Integration) : 기업 내 상호 연관된 모든 애플리케이션을 유기적으로 연동하여 필요한 정보를 중앙 집중적으로 통합, 관리, 사용할 수 있는 환경을 구현하는 것으로 e-비즈니스를 위한 기본 인프라이다. • EDW(Enterprise Data Warehouse) : 기존 DW(Data Warehouse)를 전사적으로 확장한 모델로 BPR과 CRM, BSC 같은 다양한 분석 애플리케이션들을 위한 원천이 된다. 따라서 EDW를 구축하는 것은 단순히 정보를 빠르게 전달하는 대형 시스템을 도입한다는 의미가 아니라 기업 리소스의 유기적 통합, 다원화된 관리 체계 정비, 데이터의 중복 방지 등을 위해 시스템을 재설계 하는 것을 나타낸다.
유통부문	<ul style="list-style-type: none"> • KMS(Knowledge Management System) : 지식관리시스템을 의미하며, 기업의 환경이 물품을 주로 생산하던 산업 사회에서 지적 재산의 중요성이 커지는 지식사회로 급격히 이동함에 따라, 기업 경영을 지식이라는 관점에서 새롭게 조명하는 접근방식이다. • RFID(Radio Frequency, RF) : 주파수를 이용해 ID를 식별하는 System으로 일명 전자태그로 불린다. 전파를 이용해 먼 거리에서 정보를 인식하는 기술로 적용대상에 RFID 칩을 부착한 후 리더기를 통해 정보를 인식한다.

라. 사회기반구조로서의 데이터베이스

1) 개념

1990년대 사회 각 부문의 정보화가 본격화되면서 데이터베이스 구축이 활발하게 추진되었다. 정부를 중심으로 무역, 통관, 물류, 조세, 국세, 조달 등 사회간접자본(SOC) 차원에서 EDI를 활용하여 부가가치통신망(VAN)을 통해 정보망이 구축되기 시작하였다. 1990년대 후반에는 지리, 교통부문의 데이터베이스가 구축되기 시작했고, 2000년대 들어서 더욱 고도화 되어 일반 국민들도 가정에서 손쉽게 생활에 필요한 정보를 습득하고 있다.

2) 종류

가) EDI(Electronic Data Interchange) : 주문서, 납품서, 청구서 등 무역에 필요한 각종 서류를 표준화된 양식을 통해 전자적 신호로 바뀐 컴퓨터통신망을 이용하여, 거래처에 전송하는 시스템이다.

나) VAN(Value Added Network) : 부가가치통신망, 공중 전기통신사업자(예컨대 한국전기통신공사)로부터 통신회선을 차용하여 독자적인 네트워크를 형성하는 것이다. 독자적인 네트워크로 각종 정보를 부호, 영상, 음성 등으로 교환하거나 정보를 축적하거나 또는 복수로 해서 전송하는 등 단순한 통신이 아니라 부가가치가 높은 서비스를 하는 것이다.

다) CALS(Commerce At Light Speed) : 전자상거래 구축을 위해 기업 내에서 비용 절감과 생산성 향상을 추구할 목적으로 시작된, 제품의 설계·개발·생산에서 유통·폐기에 이르기까지 제품의 라이프 사이클(Life Cycle) 전반에 관련된 데이터를 통합하고 공유·교환할 수 있도록 한 경영통합정보시스템을 말한다. 1980년대 초, 미 국방성 제품의 전 생산·유통 과정에서 컴퓨터를 활용한 자동화시스템을 구축해 효율적인 군수 조달을 위해 개발된 시스템이다.

3) 분야별 사회기반 구조의 데이터베이스

분야	솔루션
물류부문	<ul style="list-style-type: none"> • CVO(Commercial Vehicle Operation System, 화물운송정보) • PORT-MIS(항만운영정보시스템) • KROIS(철도운영정보시스템)
지리/교통부문	<ul style="list-style-type: none"> • GIS(Geographic Information System, 지리정보시스템) • RS(Remote Sensing, 원격탐사) • GPS(Global Positioning System, 범지구위치결정시스템) • ITS(Intelligent Transport System, 지능형교통시스템) • LBS(Location Based Service, 위치기반서비스) • SIM(Spatial Information Management, 공간정보관리)
의료부문	<ul style="list-style-type: none"> • PACS(Picture Archiving and Communication System) • U헬스(Ubiquitous-Health)
교육부문	<ul style="list-style-type: none"> • NEIS(National Education Information System, 교육행정정보시스템)

데이터의 가치와 미래

○ 학습 목표

- 빅데이터의 정의와 출현배경을 이해한다.
- 빅데이터의 기능과 빅데이터가 만들어 내는 본질적인 변화를 이해한다.
- 빅데이터의 가치와 영향을 이해한다.
- 빅데이터를 통한 위기 요인과 통제 방안을 이해한다.
- 빅데이터의 미래를 예상할 수 있다.

○ 눈높이 체크

✓ 빅데이터의 정의를 알고계신가요?

빅데이터는 말 그대로 큰 데이터를 의미합니다. 단순히 용량만 방대한 것이 아니라 복잡성도 증가해서 기존의 데이터 처리 톨로 다루기 어려운 데이터 셋을 지칭하기도 합니다. 이번 강의에서는 빅데이터의 정확한 정의와 출현배경, 기능에 대해 알아보고 빅데이터가 만들어 내는 본질적인 변화를 학습해 보도록 하겠습니다.

✓ 빅데이터가 우리 생활을 어떻게 바꾸어 가는지 알고 계신가요?

2012년 미국의 44대 대통령 선거에 당선된 오바마의 빅데이터를 통한 선거운동, 2013년 서울의 심야버스인 올빼미 버스의 빅데이터를 통한 노선변경 등 우리 생활에 빅데이터를 통한 적용사례는 점점 많아지고 있습니다. 빅데이터의 가치와 영향을 학습하도록 하겠습니다.

✓ 빅데이터가 발전함에 따라 위기 요인은 어떤 것들이 있을까요?

빅데이터의 활용을 통해 우리의 생활이 편리해지고 있지만 그와 반대로 사생활 침해 등의 문제도 증가함으로 인해 우리의 개인적인 삶이 노출되어 범죄에 악용될 수도 있습니다. 또한 범죄를 미리 예측해서 관리하고자 할 때 자칫 책임원칙 주의가 훼손될 수 있습니다. 더불어 데이터의 오남용으로 잘못된 미래 예측이 더 큰 피해를 불러올 수도 있습니다.

✓ 빅데이터 시대가 진행되면서 부각되는 어두운 면은 어떤 것이 있을까요?

빅데이터로 인해 부각되는 사생활 침해, 책임원칙 훼손, 데이터 오용 등은 빅데이터 시대의 부작용이라고 할 수 있습니다. 이러한 부작용을 자세히 학습하고 이러한 위기를 통제할 수 있는 방안을 논의해 보도록 하겠습니다.

✓ 미래의 빅데이터 시대는 어떻게 변할까요? 또 무엇을 준비해야 할까요?

초고속 인터넷 시대에서 모바일 광대역 네트워크시대를 살고 있는 지금 모든 물건에 센서를 연결하는 사물인터넷(IoT)시대가 도래하고 있습니다. 또 스마트폰 시장은 웨어러블 단말 시장으로 변하고 있습니다. 이러한 기술의 발전은 더욱 더 많은 데이터를 생산할 것이고 이러한 정형/비정형 데이터들을 활용한 빅데이터를 통해 우리의 삶은 더욱 편리하고 빠른 의사결정을 할 수 있도록 변화할 것입니다. 이런 미래의 빅데이터 시대에 요구되는 데이터, 기술, 인력에 대해 학습해 보도록 하겠습니다.

빅데이터의 이해

1

빅데이터의 이해

가. 빅데이터의 정의

1) 관점에 따른 정의

- 빅데이터의 정의는 빅데이터를 보는 관점에 따라 3가지로 정의한다.

첫째, 3V로 요약되는 데이터 자체의 특성 변화에 초점을 맞춘 좁은 범위의 정의가 있다.

둘째, 데이터 자체뿐 아니라 처리, 분석 기술적 변화까지 포함되는 중간 범위의 정의가 있다.

셋째, 인재, 조직 변화까지 포함한 넓은 관점에서의 빅데이터에 대한 정의가 있다.

출제
포인트

3V의 용어와 정의를 정확히 이해하고 넘어갑시다.



[가트너 그룹(Gartner Group)의 더그 래니(Doug Laney)의 3V]

퍼스널 빅데이터

사용자의 모든 행동을 복합적으로 축적한 데이터로 이동, 구매, 식사 같은 실생활 패턴 외에 웹이나 소셜 로그 같은 온라인 활동도 포함

3V			4V
양(Volume)	다양성(Variety)	속도(Velocity)	가치(Value) 진실성(Veracity) 정확성(Validity) 휘발성(Volatility)
↓	↓	↓	
데이터의 규모 측면	데이터의 유형과 소스 측면	데이터의 수집과 처리 측면	
센싱데이터, 비정형데이터	정형, 비정형데이터 (영상, 사진)	원하는 데이터의 추출 및 분석속도	

※ 3V에 가치를 추가하면 4V, 진실성, 정확성, 휘발성을 추가하면 7V의 개념이 생성되고 있음.

참고

- 맥킨지, 2011 : 빅데이터는 일반적인 데이터베이스 소프트웨어로 저장, 관리, 분석할 수 있는 범위를 초과하는 규모의 데이터를 의미한다. → 데이터 규모에 중점을 둔 정의
- IDC, 2011 : 빅데이터는 다양한 종류의 대규모 데이터로부터 저렴한 비용으로 가치를 추출하고 데이터의 초고속 수집·발굴·분석을 지원하도록 고안된 차세대 기술 및 아키텍처이다. → 분석 비용 및 기술에 초점을 맞춘 정의
- 메이아-윈베르거와퀴커, 2013 : 빅데이터란 대용량 데이터를 활용해 작은 용량에서는 얻을 수 없었던 새로운 통찰이나 가치를 추출해내는 일이다. 나아가 이를 활용해 시장, 기업 및 시민과 정부의 관계 등 많은 분야에 변화를 가져오는 일이다.

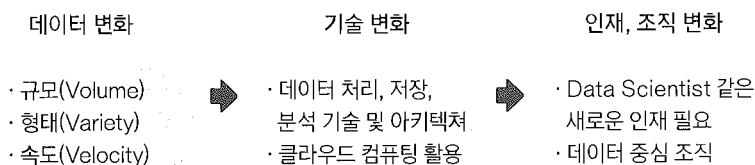


출 제
포인트

빅데이터의 범주가 '데이터의 변화 → 기술변화 → 인재, 조직의 변화'로 점점 확대되고 있음을 알고 그 내용이 무엇인지 알아야 합니다. 객관식뿐만 아니라 주관식으로도 출제될 수 있으니 정확히 숙지합니다.



2) 빅데이터 정의의 범주 및 효과



- * 기존 방식으로는 얻을 수 없는 통찰 및 가치 창출
- * 사업방식, 시장, 사회, 정부 등에서 변화와 혁신 주도

가. 출현 배경

- 빅데이터 현상은 없었던 것이 새로 등장한 것이 아니라 기존의 데이터, 처리 방식, 다루는 사람과 조직 차원에서 일어나는 '변화'를 말한다.

1) 3가지 출현 배경

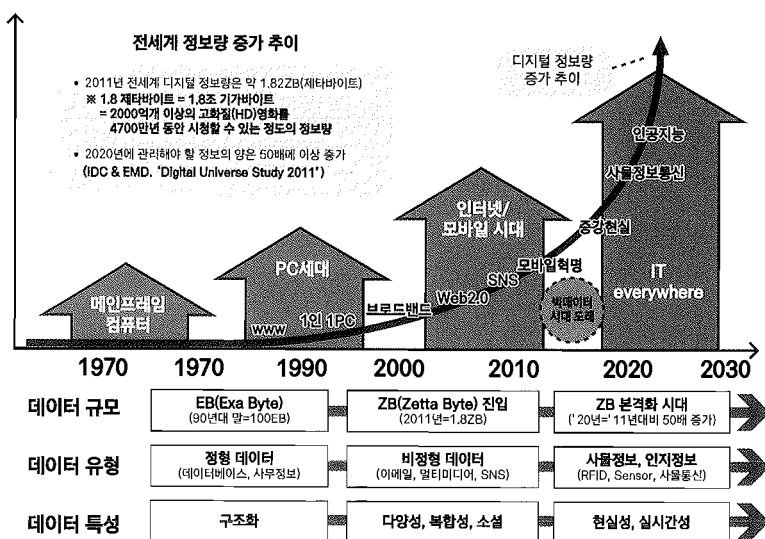
	출현배경	내 용
산업계	고객 데이터 축적	고객 데이터를 축적하여 보유함으로써 데이터에 숨어있는 가치를 발굴해 새로운 성장동력원으로서의 기술 확보
학 계	거대 데이터 활용, 과학 확산	거대 데이터를 다루는 학문 분야가 늘어나면서 필요한 기술 아키텍처 및 통계 도구들이 발전
기술발전	관련기술의 발달	디지털화, 저장 기술의 발달, 인터넷 보급, 모바일 혁명, 클라우드 컴퓨팅

예시

산업계	미국 테스코의 경우 매달 15억 건 이상의 고객데이터를 수집하고 있으며, 액시엄의 경우 전세계 5억명, 미국인 96%에 관련된 데이터를 보관하고 있다.
학 계	인간 게놈프로젝트를 통해 인간 유전자 정보를 해석, NASA의 기후 예측 시뮬레이션 센터에서는 약 32페타바이트의 기후관찰 정보를 활용하고 있다.
기술발전	아날로그의 디지털화는 데이터의 생산·유통·저장의 편리성을 개선하였으며, 저장 기술의 발달로 비용절감, 인터넷, 모바일의 발달을 통해 기술이 발전하고 있다.

2
출현 배경과
변화

2) ICT의 발전과 빅데이터의 출현



< 출처 : NIA(한국지능정보사회진흥원) - 새로운 미래를 여는 빅데이터 시대(2013) >



출 제 포인트



빅데이터의 비유 및 기능을 묻는 문제가 출제되고
있으므로 각 비유별 내용을 숙지 할 수 있도록 합니다.



3

빅데이터의 기능

가. 빅데이터에 거는 기대를 표현한 비유

산업혁명의 석탄, 철	제조업 뿐만 아니라 서비스 분야의 생산성을 획기적으로 끌어올려 사회·경제·문화·생활 전반에 혁명적 변화를 가져올 것으로 기대됨
21세기의 원유	경제 성장에 필요한 정보를 제공함으로써 산업 전반의 생산성을 한 단계 향상시키고, 기존에 없던 새로운 범주의 산업을 만들어낼 것으로 전망됨
렌즈	렌즈를 통해 현미경이 생물학 발전에 미쳤던 영향만큼이나 데이터 가 산업 발전에 영향을 미칠 것으로 기대됨  Ngram Viewer
플랫폼	'공동 활용의 목적으로 구축된 유무형의 구조물'으로써의 다양한 서 드파티 비즈니스에 활용되면서 플랫폼 역할을 할 것으로 전망됨  카카오톡, 페이스북 등



출 제 포인트

빅데이터에서 중요시 여기는 부분이 과거에서 현재로 어떻게 변화되었는지
헛갈리지 않게 체크합시다. 예제와 연결하여 이해하면 더 도움이 될 것입니다.



가. 과거에서 현재로의 변화

사전처리



사후처리

필요한 정보만 수집하고 필요하지 않은 정보를 버리는 시스템에서 가능한 한 많은 데이
터를 모으고 그 데이터를 다양한 방식으로 조합해 숨은 정보를 찾아낸다.

표본조사



전수조사

데이터 수집 비용의 감소와 클라우드 컴퓨팅 기술의 발전으로 데이터 처리비용이 감소하
게 되었다. 이로 인해 표본을 조사하는 기존의 지식발견 방식에서 전수조사를 통해 샘플
링이 주지 못하는 패턴이나 정보를 발견하는 방식으로 데이터 활용방법이 변화되었다.

질



양

데이터가 지속적으로 추가될 경우 양질의 정보가 오류 정보보다 많아 전체적으로 좋은
결과 산출에 긍정적인 영향을 미친다는 추론에 바탕을 둔 변화가 나타나고 있다.

인과관계



상관관계

상관관계를 통해 특정 현상의 발생 가능성이 포착되고, 그에 상응하는 행동을 하도록 추천
되는 일이 점점 늘어나고 있다. 이처럼 데이터 기반의 상관관계 분석이 주는 인사이트가 인
과관계에 의한 미래 예측을 점점 더 압도해 가는 시대가 도래하게 될 것으로 전망된다.

빅데이터가
만들어 내는
본질적인 변화

빅데이터의 가치와 영향

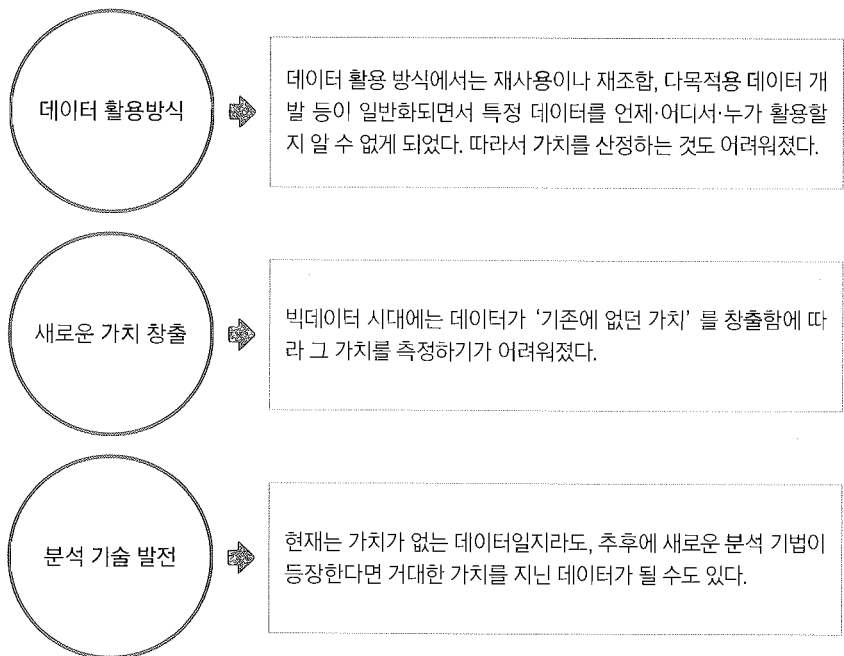
1

빅데이터의
가치

가. 빅데이터 가치 산정이 어려운 이유

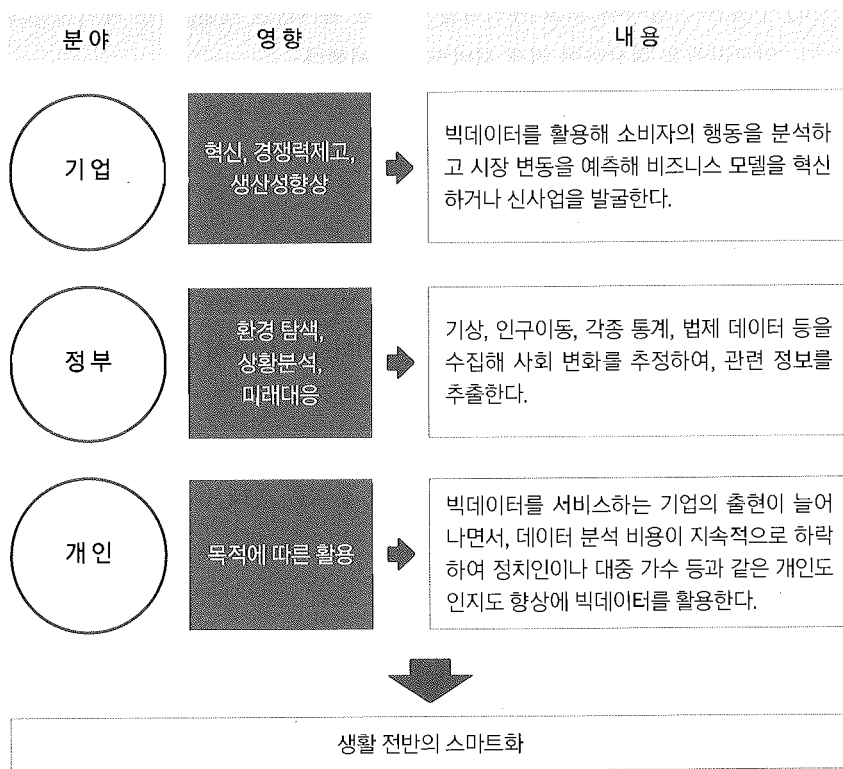
- 여러 가지 변수로 인해 빅데이터 시대에서는 가치를 측정하는 것이 쉽지 않다.

빅데이터 가치 산정이 어려운 이유



- 다양한 시장 주체들이 빅데이터를 활용함에 따라 소비자이면서 국민인 일반인들은 맞춤형 서비스를 저렴한 비용으로 이용하게 되고, 적시에 필요한 정보를 얻음으로써 다양한 형태로 기회비용을 절약할 수 있어 사람들의 생활이 점점 스마트해지고 있다.

빅데이터가 미치는 영향



참고

맥킨지가 언급한 빅데이터가 가치를 만들어 내는 다섯가지 방식

- ① 투명성 제고로 연구개발 및 관리 효율성 제고
- ② 시뮬레이션을 통한 수요 포착 및 주요 변수 탐색으로 경쟁력 강화
- ③ 고객 세분화 및 맞춤 서비스 제공
- ④ 알고리즘을 활용한 의사결정 보조 혹은 대체
- ⑤ 비즈니스 모델과 제품, 서비스의 혁신



3절

2장 데이터의 가치와 미래

비즈니스 모델

1

빅데이터 활용 사례

빅데이터 활용사례

- 사용자 로그데이터를 분석해서 기존의 페이지 링크 알고리즘을 개선(구글)
- 고객의 구매패턴을 분석해서 상품 진열을 바꿈(월마트)
- 실시간 자동 번역시스템을 통해 의사소통의 불편을 해소(페이스북, 구글 등)
- 전자책 관련 데이터를 분석하여 저자에게 독서 패턴 정보 제공(아마존)

가. 기업

- 1) 구글은 사용자의 로그 데이터를 활용한 검색엔진 개발, 기존 페이지랭크 알고리즘을 혁신하여 검색 서비스를 개선했다.
- 2) 월마트는 고객의 구매패턴을 분석해 상품진열에 활용했다.

나. 정부

- 1) 정부는 실시간 교통정보 수집, 기후 정보, 각종 지질 활동, 소방 서비스 등 다양한 국가 안전 확보 활동을 위해 실시간 모니터링을 활용한다. 이 밖에도 미래 의제인 의료와 교육 개선을 위해 빅데이터를 활용해 해결책을 모색한다.

다. 개인

- 1) 정치인은 선거 승리를 위해 사회관계망 분석을 통해 유세 지역을 선정하고, 해당 지역의 유권자에게 영향을 줄 수 있는 내용을 선정해 효과적인 선거 활동을 펼친다.
- 2) 가수는 팬들의 음악 청취 기록 분석을 통해 실제 공연에서 부를 노래 순서를 짜는데 활용한다.



출 제 포인트

빅데이터 활용 기본 테크닉 7가지를 달달 외울 필요는 없지만 각각의 테크닉이 어떤 기술인지, 어떻게 활용되고 있는지는 반드시 숙지해야 합니다.



가. 빅데이터를 활용한 기본 테크닉

테크닉	내용	예시
 연관규칙학습	변인들 간에 주목할 만한 상관관계가 있는지를 찾아내는 방법	커피를 구매하는 사람이 탄산음료를 더 많이 사는가?
 유형분석	문서를 분류하거나 조직을 그룹으로 나눌 때, 또는 온라인 수강생들을 특성에 따라 분류할 때 사용	이 사용자는 어떤 특성을 가진 집단에 속하는가?
 유전자 알고리즘	최적화가 필요한 문제의 해결책을 자연선택, 돌연변이 등과 같은 메커니즘을 통해 점진적으로 진화(Evolve)시켜 나가는 방법	최대의 시청률을 얻으려면 어떤 프로그램을 어떤 시간대에 방송해야 하는가?
 기계학습	훈련 데이터로부터 학습한 알려진 특성을 활용해 예측하는 방법	기존의 시청 기록을 바탕으로 시청자가 현재 보유한 영화 중에서 어떤 것을 가장 보고 싶어할까?
 회귀분석	독립변수를 조작함에 따라, 종속변수가 어떻게 변하는지를 보면서 두 변인의 관계를 파악할 때 사용	구매자의 나이가 구매 차량의 타입에 어떤 영향을 미치는가?
 감정분석	특정 주제에 대해 말하거나 글을 쓴 사람의 감정을 분석	새로운 환불 정책에 대한 고객의 평가는 어떤가?
 소셜네트워크분석 (=사회관계망분석)	특정인과 다른 사람이 몇 촌 정도의 관계인가를 파악할 때 사용하고, 영향력있는 사람을 찾아낼 때 사용	고객들 간 관계망은 어떻게 구성되어 있나?

빅데이터 활용 기본 테크닉

예측적 분석 (Predictive Analysis)

미래의 불확실한 사실을 사전에 예측하거나 알려지지 않은 결과의 가능성을 파악하기 위하여 사용하는 분석 방법



출 제 포인트

빅데이터가 등장하기 이전엔 정형데이터를 주로 이용했습니다. (연관규칙학습, 유형분석, 유전자 알고리즘, 기계학습, 회귀분석) 하지만 최근 SNS가 발달함에 따라 비정형화된 데이터를 많이 이용하고 있습니다. (감정분석) 뒤에서 더 자세히 학습합니다.



4절

2장 데이터의 가치와 미래

위기 요인과 통제 방안

출제
포인트

빅데이터 시대의 위기 요인과 예시, 그리고 통제 방안에 대해서 시험이 자주 출제되므로 정확히 숙지하여야 합니다.



1

빅데이터
시대의
위기 요인

가. 사생활 침해

내용	개인정보가 포함된 데이터를 목적 외에 활용할 경우 사생활 침해를 넘어 사회·경제적 위협으로 변형될 수 있다.
예시	여행 사실을 트위터 한 사람의 집을 강도가 노리는 고전적 사례 발생 → 익명화(Anonymization) 기술 발전이 필요하다.

나. 책임 원칙 훼손

분산 서비스 거부(DDoS) 공격
특정 서버를 대상으로 지속적
이고 많은 양의 트래픽을 유발
시켜, 정상적인 서비스 제공이
불가능하도록 만드는 해킹 기법

내용	빅데이터 기본분석과 예측기술이 발전하면서 정확도가 증가한 만큼, 분석대상이 되는 사람들은 예측 알고리즘의 희생양이 될 가능성도 증가한다. 민주주의 국가에서는 잠재적 위협이 아닌 명확한 결과에 대한 책임을 묻고 있어 이에 따른 원리를 훼손할 가능성이 있다.
예시	영화 “마이너리티 리포트”에 나오는 것처럼 범죄 예측 프로그램에 의해 범행을 저지르기 전에 체포, 자신의 신용도와 무관하고 부당하게 대출이 거절되었다. → 민주주의 국가의 형사 처벌은 잠재적 위협이 아닌 명확하게 행동한 결과에 대해 책임을 묻고 있다.

다. 데이터 오용

내용	빅데이터는 일어난 일에 대한 데이터에 의존하기 때문에 이를 바탕으로 미래를 예측하는 것은 적지 않은 정확도를 가질 수 있지만 항상 맞을 수는 없다. 또한 잘못된 지표를 사용하는 것도 빅데이터의 폐해가 될 수 있다.
예시	베트남 전쟁 때, 맥나마라 장군은 적군 사망자 수를 전쟁의 진척상황을 나타내는 지표로 활용했고 그 결과 적군 사망자 수는 과장돼 보고되는 경향을 보여 결과적으로 전쟁 상황을 오보하는 결과를 일으켰다.

2

위기 요인에 따른 통제 방안

가. 동의에서 책임으로

내 용	빅데이터에 의한 사생활침해 문제를 해결하기에는 부족한 측면이 많고 매번 개인정보 제공 동의를 하는 비효율적인 단계를 줄이고자 개인정보를 사용하는 사용자의 '책임'으로 해결하는 방안을 제시하였다. (‘개인정보 제공자의 동의’ → ‘개인정보 사용자의 책임’)
기대효과	개인정보 유출 및 사용으로 발생하는 피해에 대해 사용자가 책임을 지게됨으로 사용자체의 적극적인 보호장치를 강구할 수 있다.

참고

소비자 프라이버시 보호 3대 권고사항



- ① 기업은 상품 개발 단계에서부터 소비자 프라이버시 보호 방안을 적용하라.
- ② 기업은 소비자에게 공유정보 선택 옵션을 제공하라.
- ③ 소비자에게 수집된 정보 내용 공개 및 접근권을 부여하라.

나. 결과 기반 책임 원칙 고수

내 용	책임원칙 훼손 위기요인에 대한 통제 방안으로 기존의 원칙을 좀 더 보강하고 강화할 필요가 있으며, 예측 자료에 의한 불이익을 당할 가능성을 최소화하는 장치를 마련하는 것이 필요하다.
기대효과	잘못된 예측 알고리즘을 통한 판단을 근거로 불이익을 줄 수 없으며, 이에 따른 피해 최소화 장치를 마련해야 한다.

다. 알고리즘 접근 허용

내 용	데이터 오용의 위기요소에 대한 대응책으로 ‘알고리즘에 대한 접근권’을 제공하여 예측 알고리즘의 부당함을 반증할 수 있는 방법을 명시해 공개할 것을 주문한다.
기대효과	불이익을 당한 사람들을 대변할 전문가(알고리즘미스트)가 필요하게 되었다.

미래의 빅데이터

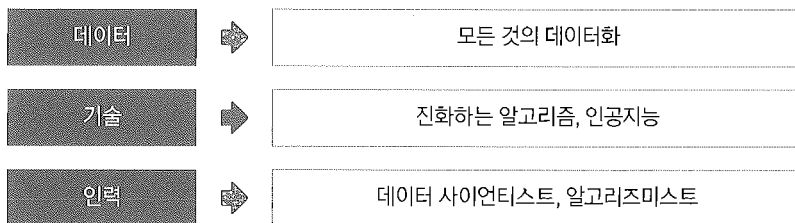
1

빅데이터 활용의 3요소

모든 것의 데이터화
(Datafication)

사물인터넷(Internet of Things, IoT) 시대에 웨어러블(Wearable) 단말의 발전으로 대화기록, 음악청취 기록 등이 저장되어 사물인터넷 시대가 되어 훨씬 더 많은 정보가 생산, 공유됨

가. 기본 3요소



1) 데이터

- 모든 것을 데이터화(Datafication) 하는 현 추세로 특정 목적없이 축적된 데이터를 통한 창의적인 분석이 가능해져, 새로운 가치로 부상하고 있다.

2) 기술

- 대용량의 데이터를 빠르게 처리하기 위한 알고리즘의 진화와 함께 스스로 학습하고 데이터를 처리할 수 있는 인공지능 기술이 출현하였다.

3) 인력

- 빅데이터를 처리하기 위한 데이터 사이언티스트와 알고리즘미스트의 역할을 통해 빅데이터의 다각적 분석을 통한 인사이트 도출이 중요해지고 있다.

참고

- 데이터 사이언티스트 :**
빅데이터에 대한 이론적 지식과 숙련된 분석 기술을 바탕으로 통찰력, 전달력, 협업 능력을 두루 갖춘 전문인력으로써 빅데이터의 다각적 분석을 통해 인사이트를 도출하고 이를 조직의 전략 방향제시에 활용할 줄 아는 기획자
- 알고리즘미스트 :**
데이터 사이언티스트가 한 일로 인해 부당하게 피해가 발생하는 것을 막는 역할을 하며 알고리즘 코딩 해석을 통해 빅데이터 알고리즘에 의해 부당하게 피해를 입은 사람을 구제하는 전문인력



가치 창조를 위한 데이터 사이언스와 전략 인사이트

○ 학습 목표

- 빅데이터 회의론의 원인과 해소 방안을 이해한다.
- 일차원적 분석과 전략도출을 위한 가치 기반 분석의 차이를 이해한다.
- 데이터 사이언스와 데이터 사이언티스트를 이해한다.
- 빅데이터 시대의 가치 패러다임의 변화를 이해한다.

○ 눈높이 체크

✓ 빅데이터의 회의론과 우려의 목소리를 들어보셨나요?

최근 빅데이터에 관한 회의론과 우려의 목소리가 나오고 있습니다. 과거의 CRM과 같은 경영시스템을 도입하기 위해 하드웨어와 소프트웨어를 도입하고도 성과를 충분히 내지 못했던 기업들이 많았습니다. 이런 기업들의 실패 경험들이 빅데이터 시스템의 도입도 머뭇거리고 있습니다. 기업들의 우려 섞인 목소리의 원인과 이러한 의구심을 불식시키기 위한 전략적 발전 방향을 살펴보도록 합시다.

✓ 데이터 사이언스와 데이터 사이언티스트에 대해 들어 보셨나요?

빅데이터 시대를 이끌어 나가기 위해서는 데이터 사이언스라는 융합 학문이 필요합니다. 기존의 통계학과 컴퓨터공학, 그리고 경영학과 인문학을 아우르는 학문적 소양을 배우고 빅데이터 시대를 이끌어 나갈 데이터 사이언티스트를 양산함으로써 기업과 우리 생활을 변화시킬 수 있는 전략적 가치를 만들 수 있습니다.

1절

3장 가치창조를 위한 데이터 사이언스와 전략 인사이트

빅데이터 분석과 전략 인사이트

①

빅데이터 열풍과 회의론

빅데이터의 열풍은 ‘빨리 끓어 오른 냄비가 빨리 식는다’는 일종의 거품현상을 우려하는 시선도 없지 않다. 그래서 벌써부터 빅데이터 회의론이 심심찮게 흘러 나오기까지 하여, 자칫 이런 회의론이 갖는 문제는 실제 우리가 빅데이터 분석에서 찾을 수 있는 수많은 가치들을 제대로 발굴해 보기도 전에 그 활용 자체를 사전에 차단해 버릴 수 있다.

②

빅데이터 회의론의 원인 및 진단

가. 투자효과를 거두지 못했던 부정적 학습효과 → 과거의 고객관계관리(CRM)

- 과거 CRM의 부정적 학습효과
 - 공포 마케팅이 잘 통하는 영역 : 도입만 하면 모든 문제를 한번에 해소할 것처럼 강조
 - 막상 거액을 투자하여 하드웨어와 솔루션을 도입해도 어떻게 활용하고 어떻게 가치를 뽑아내야 할지 난감해 함

나. 빅데이터 성공사례가 기존 분석 프로젝트를 포함해 놓은 것이 많다.

- 굳이 빅데이터가 필요 없는 경우(우수고객, 이탈예측, 구매패턴 분석 등)
- 국내 빅데이터 업체들이 CRM 분석 성과를 빅데이터 분석으로 과대포장

빅데이터 분석도 기존의 분석과 마찬가지로, 데이터에서 가치, 즉 통찰을 끌어내 성과를 창출하는 것이 관건이며, 단순히 빅데이터에 포커스를 두지 말고 분석을 통해 가치를 만드는 것에 집중해야 한다.

왜 싸이월드는 페이스북이 되지 못했나.



■ 싸이월드

- 2004년 경 세계 최대의 소셜 네트워크 서비스(SNS)

■ 싸이월드 퇴보 원인

- OLAP과 같은 분석 인프라가 존재하였으나 중요한 의사결정이 데이터 분석에 기초하지 못했다.
- '웹로그 분석을 통한 일차원적 분석 ⇒ 사업 상황 확인'을 위한 협소한 문제에 집중되었다.
- 2004년 당시 비즈니스의 핵심 가치와 관련된 어떤 심도있는 분석도 수행되지 않았다.
- 소셜 네트워킹 서비스지만 회원들의 소셜 네트워킹 활동 특성과 관련된 분석을 위한 프레임워크나 평가 지표조차 없었다.
- 트렌드 변화가 사업모델에 미치는 영향을 적시에 알아차리지 못했다.

■ 전략적 분석을 통해 놀라운 성과를 올린 하라스엔터테인먼트의 회장 러브먼이 언급한 분석 기반 경영이 도입되지 못하는 이유

- 기존 관행을 그냥 따를 뿐 중요한 시도를 하지 않는다.
- 경영진의 의사결정이 정확성이나 공정한 분석을 필요로 하지 않으며, 오히려 정반대로 직관적 결정을 귀한 재능으로 칭송받는 경향이 있다.
- 분석적 실험을 갈망하거나 능숙하게 해내는 사람이 거의 없어, 적절한 방법조차 제대로 익히지 못한 사람들에게 분석 업무가 주어진다.
- 사람들은 아이디어 자체보다는 아이디어를 낸 사람이 누구인지 관심을 두는 경향이 있다.



- 전략적 분석은 치열한 시장에서 기업 생존을 좌우할 정도로 중요할 수 있다.

로직오류와 프로세스 오류

- 로직 오류 : 의도치 않은, 바라지 않은 결과 유발
- 프로세스 오류 : 작동에 문제가 발생한 오류

가. 빅데이터에 대한 관심 증대

- 데이터 기반의 통찰의 중요성에 대한 공감대 상승과 동시에 긍정적 효과를 기대한다.

나. 빅데이터 프로젝트에 거는 기대

- 기존 프로세스의 자동화를 우선 시행한 후 점차적으로 거시적이고, 전략적인 가치를 이끌어 낼 수 있을 것으로 기대한다.

③

빅데이터 분석,
'Big'이
핵심 아니다.

다. 빅데이터 분석의 가치

- 데이터는 크기의 이슈가 아니라, 거기에서 어떤 시각과 통찰을 얻을 수 있느냐의 문제가 중요하다. 무작정 ‘빅’한 데이터를 찾을 것이 아니라, 비즈니스의 핵심에 대해 보다 객관적이고 종합적인 통찰을 줄 수 있는 데이터를 찾는 것이 그 무엇보다 중요하다.
- 전략과 비즈니스의 핵심 가치에 집중하고 이와 관련된 분석 평가지표를 개발하고 이를 통해 효과적으로 시장과 고객 변화에 대응할 수 있을 때 빅데이터 분석은 가치를 줄 수 있다.

참고

- 조슈아 보거 박사는 “직관에 기초한 의사결정보다 데이터에 기초한 의사결정이 그 만큼 중요하다”고 말했으며 이는 데이터 자체의 중요성을 강조한 것이다.
- 빅데이터 프로젝트 초기 단계에 자주 나오는 질문
 “빅데이터를 가장 효과적으로 소비하는 것은 인간인가 기계인가?”
 “고객 데이터와 운영 데이터 중 어느 것이 더 중요한가?”
 “새로운 데이터가 새로운 인사이트 도출을 촉진하는가, 아니면 단순히 기존 가설을 입증할 뿐인가?”



4 전략적 통찰이 없는 분석의 함정

- 단순히 분석을 많이 사용하는 것이 곧바로 경쟁우위를 가져다 주지는 않는다.
- 분석이 경쟁의 본질을 제대로 바라보지 못할 때에는 쓸모없는 분석 결과들만 잔뜩 쏟아내게 된다. 이를 예방하기 위해서는 전략적인 통찰력을 가지고 분석하고 핵심적인 비즈니스 이슈에 집중하여 데이터를 분석하고 차별적인 전략으로 기업을 운영하여야 한다.

참고

아메리칸항공	사우스웨스트항공
수익관리, 가격 최적화의 분석접근법 적용 → 3년만에 14억 달러 수익	단순 최적화 모델을 통해 가격 책정과 운영
비행경로와 승무원들의 일정을 최적화 → 12기종, 250개 목적지, 매일 3,400회 비행	한가지 기종의 비행기로 단순화 ↓
↓ 타 경쟁사들이 비슷한 분석역량과 수익관리 능력을 갖추었으므로 경쟁우위 하락 → 수익 절감	단순 최적화 모델로 좌석 가격 책정 및 운영 결과 경쟁우위 상승 → 36년 연속 흑자, 미국 항공사들의 시장가 치를 합친 것 보다 높은 시장가치 확보

- 위의 결과를 통해 분석을 보다 전략적으로 사용하기 위해 노력하지 않으면 차별화가 어려움을 판단할 수 있으며 비즈니스 모델을 뒷받침하는 분석의 한계를 아메리칸항공이 나타내고 있다.





출제
포인트

분석 애플리케이션이 어느 산업에서 활용되는 애플리케이션인지에 대해 자주 시험문제가 출제되므로 꼭 숙지하고 넘어가세요.



가. 산업별 분석 애플리케이션

산업	일차원적 분석 애플리케이션
금융 서비스	신용점수 산정, 사기 탐지, 가격 책정, 프로그램 트레이딩, 클레임 분석, 고객 수익성 분석
소매업	판매, 매대 관리, 수요 예측, 재고 보충, 가격 및 제조 최적화
제조업	공급사슬 최적화, 수요 예측, 재고 보충, 보증서 분석, 맞춤형 상품 개발, 신상품 개발
운송업	일정 관리, 노선 배정, 수익 관리
헬스케어	약품 거래, 예비 진단, 질병 관리
병원	가격 책정, 고객 로열티, 수익 관리
에너지	트레이딩, 공급/수요 예측
커뮤니케이션	가격 계획 최적화, 고객 보유, 수요 예측, 생산능력 계획, 네트워크 최적화, 고객 수익성 관리
서비스	콜센터 직원 관리, 서비스-수익 사슬 관리
정부	사기 탐지, 사례 관리, 범죄 방지, 수익 최적화
온라인	웹 매트릭스, 사이트 설계, 고객 추천
모든사업	성과 관리

5
일차원적인 분석
vs 전략도출
위한
가치기반 분석



출제
포인트

일차적인 분석과 전략도출 가치기반과 관련하여 잘못된 설명을 묻는 문항이 출제되고 있으니 숙지하고 넘어가시기 바랍니다.



나. 일차적인 분석의 문제점

- 일차적인 분석을 통해서도 해당 부서나 업무 영역에서는 상당한 효과를 얻을 수 있지만 일차적인 분석만으로는 환경변화와 같은 큰 변화에 제대로 대응하거나 고객 환경의 변화를 파악하고 새로운 기회를 포착하기 어렵다. 특히, 급변하는 환경에서는 분석을 일차적 차원에서 점증적, 진술적으로 사용하면 성과는 미미할 수 있다.

전략 인사이트 도출을 위한 필요 역량



출 제 포인트

데이터 사이언스에 대해 묻는 문제가 출제될 수 있으니 꼭 숙지하고 넘어가세요.



가. 의미

- 데이터 사이언스란 데이터 공학, 수학, 통계학, 컴퓨터공학, 시각화, 해커의 사고방식, 해당 분야의 전문지식을 종합한 학문이다. 데이터로부터 의미있는 정보를 추출해내는 학문으로 정형 또는 비정형을 막론하고 인터넷, 휴대전화, 감시용 카메라 등에서 생성되는 숫자와 문자, 영상 정보 등 다양한 유형의 데이터를 대상으로 분석뿐 아니라 이를 효과적으로 구현하고 전달하는 과정까지를 포함한 포괄적 개념이다.

나. 역할

- 데이터 사이언티스트는 비즈니스의 성과를 좌우하는 핵심이슈에 답을 하고, 사업의 성과를 견인해 나갈 수 있어야 한다. 이는 데이터 사이언스의 중요한 역량 중 하나인 소통력이 필요한 이유이다.

데이터 사이언스의 의미와 역할

참고

- 링크드인(Linkedln) : 비즈니스 네트워킹 서비스
- 골드만(스탠퍼드 물리학 박사 출신의 데이터 사이언티스트)
→ 당신이 알 수도 있는 사람들(People You May Know) 이라는 배너를 추가해 백만 개의 새로운 뷰를 창출





출제 포인트

데이터 사이언스의 구성요소와 그 내용에 대한 객관식 문제가 출제될 가능성이 있으니 확실히 정리해봅시다.



2

데이터 사이언스의 구성요소

가. 데이터 사이언스의 영역

분석적 영역

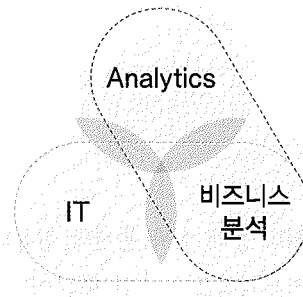
수학, 확률모델, 머신러닝, 분석학,
패턴 인식과 학습, 불확실성 모델링 등

IT 컨설팅

전략 컨설팅

데이터 처리와 관련된 IT 영역

시그널 프로세싱, 프로그래밍,
데이터 엔지니어링,
데이터 웨어하우스,
고성능 컴퓨팅



비즈니스 컨설팅 영역

커뮤니케이션, 프레젠테이션,
스토리텔링, 시각화 등

나. 데이터 사이언티스트의 역할

- 데이터 사이언티스트는 데이터 홍수 속에서 해답을 찾고, 데이터 소스를 찾고, 복잡한 대용량 데이터를 구조화, 불완전한 데이터를 서로 연결해야 한다.
- 데이터 사이언티스트가 갖춰야 할 역량 중 한 가지는 '강력한 호기심'이다. 호기심이란 문제의 이면을 파고들고, 질문들을 찾고, 검증 가능한 가설을 세우는 능력을 의미한다.
- 데이터 사이언티스트는 스토리텔링, 커뮤니케이션, 창의력, 열정, 직관력, 비판적 시각, 글쓰기 능력, 대화능력 등을 갖춰야 한다.



출제 포인트

데이터 사이언티스트에 요구되는 Hard Skill과 Soft Skill의 내용에 대해 해당하거나 해당하지 않는 것들을 고르는 문제가 자주 출제되니 이해하고 넘어갑시다.



3

데이터 사이언티스트의 요구 역량

● Hard Skill

- ① 빅데이터에 대한 이론적 지식
: 관련 기법에 대한 이해와
방법론 습득
- ② 분석 기술에 대한 숙련
: 최적의 분석 설계 및
노하우 축적

Analytics

IT 전문성

비즈니스
분석

● Soft Skill

- ③ 통찰력 있는 분석
: 창의적 사고, 호기심,
논리적 비판
- ④ 설득력 있는 전달
: 스토리텔링, 비주얼라
이제이션
- ⑤ 다분야 간 협력
: 커뮤니케이션

④

데이터 사이언스 : 과학과 인문의 교차로

⑤

전략적 통찰력과 인문학의 부활

- 분석기술보다 더 중요한 것은 소프트 스킬로 전략적 통찰을 주는 분석은 단순 통계나 데이터 처리와 관련된 지식 외에도 스토리텔링, 커뮤니케이션, 창의력, 열정, 직관력, 비판적 시각, 대화능력 등 인문학적 요소가 필요하다.

가. 통찰력있는 분석

- 직관과 전략, 경영 프레임워크 경험의 혼합을 통해 통찰력있는 분석을 수행할 수 있어야 한다.
- 본인 회사 뿐 아니라 전체 업계의 방향과 고객이 무엇을 중시하는지에 대한 이해가 필요하다.
- 좁은 시각으로 나무만 보는 것이 아니라 넓은 시각으로 숲을 볼 수 있어야 한다.

나. 인문학의 열풍

- 우리는 지금 기존 사고의 틀을 벗어나 문제를 바라보고 해결하는 능력, 비즈니스의 핵심가치를 이해하고 고객과 지원의 내면적 요구를 이해하는 능력 등 인문학에서 배울 수 있는 역량이 점점 더 절실히 요구되는 시대를 맞이하고 있다.

외부 환경적 측면에서 본 인문학 열풍의 이유

외부환경의 변화	내용	예시
컨버전스 ↓ 디버전스	단순세계화에서 복잡한 세계화로의 변화	규모의 경제, 세계화, 표준화, 이성화 → 복잡한 세계, 다양성, 관계, 연결성, 창조성
생산 ↓ 서비스	비즈니스 중심이 제품생산에서 서비스로 이동	고장나지 않는 제품의 생산 → 뛰어난 서비스로 응대
생산 ↓ 시장창조	공급자 중심의 기술경쟁에서 무형자산의 경쟁으로 변화	생산에 관련된 기술 중심, 기술 중심의 대규모 투자 → 현재 패러다임에 근거한 시장창조 현저 사회와 문화에 관한 지식

빅데이터 그리고 데이터 사이언스의 미래

1

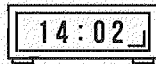
빅데이터의
시대

- 디지털 환경의 진전과 더불어 실로 엄청난 ‘빅’ 데이터가 생성되고 있다.
(2011년 전 세계에서 생성되는 디지털 정보량은 1.8 제타바이트)
- 빅데이터 분석은 선거결과에 결정적인 영향을 미칠 수도 있다. 기업의 측면에서는 비용 절감, 시간 절약, 매출 증대, 고객서비스 향상, 신규 비즈니스 창출, 내부 의사결정 지원 등에 있어 상당한 가치를 발휘하고 있다.

2

빅데이터
회의론을 넘어
가치 패러다임의
변화

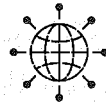
서비타이제이션
(Servitization)
제품과 서비스의 결합, 서비스
의 상품화, 그리고 기존 서비
스와 신규 서비스의 결합 현상
을 포괄하는 개념



Digitalization

과거

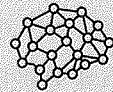
- 아날로그 세상을 어떻게 효과적으로 디지털화하는지가 과거의 가치 창출 원천



Connection

현재

- 디지털화된 정보와 대상들은 서로 연결 시작
- 연결을 더 효과적이고 효율적으로 제공하는가가 성공요인



Agency

미래

- 복잡한 연결을 얼마나 효과적이고 믿을 수 있게 관리하는가의 이슈

가. 데이터 사이언스의 한계

- 분석과정에서는 가정 등 인간의 해석이 개입되는 단계를 반드시 거친다.
- 분석결과가 의미하는 바는 사람에 따라 전혀 다른 해석과 결론을 내릴 수 있다.
- 아무리 정량적인 분석이라도 모든 분석은 가정에 근거한다는 사실이다.

나. 데이터 사이언스와 인문학

- 인문학을 이용하여 빅데이터와 데이터 사이언스가 데이터에 묻혀 있는 잠재력을 풀어내고, 새로운 기회를 찾고, 누구도 보지 못한 창조의 밑그림을 그릴 수 있는 힘을 발휘하게 될 것이다.

미래사회의 특성과
빅데이터의 역할

불확실성·통찰력, 리스크·대응
력, 스마트·경쟁력, 융합·창조력

MEMO

최신 빅데이터 상식

①

DBMS와 SQL

가. DBMS

1) DBMS란 무엇인가

- DBMS는 Data Base Management System의 약자로서 데이터베이스를 관리하여 응용 프로그램들이 데이터베이스를 공유하며 사용할 수 있는 환경을 제공하는 소프트웨어다.
- 데이터베이스를 구축하는 틀을 제공하며, 효율적인 데이터 검색, 저장 기능 등을 제공한다.
- 대표적인 데이터베이스 관리시스템에는 오라클, 인포믹스, 액세스 등이 있다.

2) 데이터베이스 관리시스템 종류

가) 관계형 DBMS

- 이 모델은 데이터를 컬럼(Column)과 로우(Row)를 이루는 하나 이상의 테이블(또는 관계)로 정리하며, 고유키(Primary Key)가 각 로우를 식별한다. 로우는 레코드나 튜플로 부르며, 일반적으로 각 테이블은 하나의 엔티티 타입(고객이나 제품과 같은)을 대표한다. 로우는 그 엔티티 종류의 인스턴스(예 : “Lee” 등)를 대표하며 컬럼은 그 인스턴스의 속성이 되는 값들(예 : 주소나 가격)을 대표한다.

나) 객체지향 DBMS

- 객체지향DB는 일반적으로 사용되는 테이블 기반의 관계형DB와 다르게 정보를 ‘객체’ 형태로 표현하는 데이터베이스 모델이다.

다) 네트워크 DBMS

- 레코드들이 노드로, 레코드들 사이의 관계가 간선으로 표현되는 그 래프를 기반으로 하는 데이터베이스 모델이다.



라) 계층형 DBMS

- 트리 구조를 기반으로 하는 계층 데이터베이스 모델이다.

3) 데이터베이스의 설계절차

- 요구사항 분석 → 개념적 설계 → 논리적 설계 → 물리적 설계 → 구현

4) Relationship

- 관계(Relationship)란 관리하고자 하는 업무 영역 내의 특정한 두 개의 엔터티 사이에 존재하는 많은 관계 중 특별히 관리하고자 하는 직접적인 관계(업무적 연관성)를 의미한다.
- 관계의 형태는 크게 1:1, m:1, m:n의 세 가지로 나눌 수 있다.

가) 1:1 관계

- 어느 쪽 당사자의 입장에서 상대를 보더라도 반드시 단 하나씩과 관계를 가지는 것을 말한다.
- 현실에서 매우 드물게 나타나며, 업무의 흐름에 따라 데이터가 설계된 형태에서 많이 나타난다.

나) m:1 관계

- 가장 흔하게 나타나는 매우 일반적인 형태이며, 한쪽은 m(many)이고 다른 한쪽은 1(one)인 것을 말한다. 부모와 자식 관계라고 생각하면 부모는 자식을 1명 이상 낳을 수 있지만, 자식은 부모를 하나만 가질 수 있다.

다) m:n 관계

- 서로가 서로를 1:N관계로 보는 것으로 쇼핑몰에서 회원과 상품이 관계를 생각해보면, 한 회원은 쇼핑몰의 여러 상품들을 가질 수 있으며, 반대로 한 티셔츠도 여러 회원들을 가질 수 있다.

5) 데이터 웨어하우스와 ETL

- 데이터 웨어하우스(Data Warehouse) : 방대한 조직 내에서 분산 운영되는 각각의 데이터 베이스 관리 시스템들을 효율적으로 통합하여 조정·관리하는 역할을 하여 효율적인 의사 결정 시스템을 위한 기초를 제공하는 실무적인 활용 방법론이 제공되고 있다.
- 특징은 아래와 같다.

특징	설명
주제지향성(Subject Oriented)	업무 중심이 아닌 주제 중심
통합성(Integrated)	혼재한 DB로부터의 데이터 통합
시계열성(Time Variant)	시간에 따른 변경 정보를 나타냄
비휘발성(Non-Volatile)	데이터 변경 없이 리포팅을 위한 read only 사용

- ETL(Extract, Transform, Load)이란 데이터 웨어하우스 구축 시 데이터를 운영 시스템에서 추출하여 가공(변환, 정제)한 후 데이터 웨어하우스(DW)에 적재하는 과정을 말한다.

6) NoSQL

- 데이터의 폭발적인 증가에 대응하기 위해 빅데이터 분산처리 및 저장기술과 함께 발달된 분산 데이터베이스 기술로 확장성, 가용성 높은 성능을 제공한다. 비관계형 데이터베이스 관리 시스템으로, SQL 계열 쿼리언어를 사용할 수 있다는 사실을 강조한다는 면에서 'Not only SQL'로 불리기도 한다.
- Key와 Value의 형태로 자료를 저장하고, 대용량 데이터 처리와 대규모의 수평적 확장성을 제공한다. 대부분 오픈소스이며, MongoDB, Hbase, Redis, Cassandra 등이 있다.

나. SQL

1) SQL이란 무엇인가?

- SQL은 Structured Query Language의 약자로, 데이터베이스를 사용할 때 데이터베이스에 접근할 수 있는 데이터베이스의 하부 언어로, 단순한 질의 기능 뿐만 아니라 완전한 데이터의 정의와 조작 기능을 갖추고 있다.
- 테이블을 단위로 연산을 수행하며, 영어 문장과 비슷한 구문으로 초보자들도 비교적 쉽게 사용할 수 있다.

2) SQL의 분류

가) DDL(Data Definition Language, 데이터 정의어)

- 데이터베이스를 정의하는 언어를 말하며, 데이터를 생성, 수정, 삭제 등 데이터의 전체 골격을 결정하는 역할의 언어이다. CREATE, ALTER, DROP, TRUNCATE가 있다.

나) DML(Data Manipulation Language, 데이터 조작어)

- 정의된 데이터베이스에 입력된 레코드를 조회, 수정, 삭제하는 등 역할을 하는 언어이다. SELECT, INSERT, UPDATE, DELETE가 있다.

다) DCL(Data Control Language, 데이터 제어어)

- 데이터베이스에 접근하거나 객체에 권한을 주는 등의 역할을 하는 언어이다. 데이터의 보안, 무결성, 회복 등을 정의하는데 사용한다. GRANT, REVOKE, COMMIT, ROLLBACK가 있다.

3) SQL 집계함수

함수명	설 명	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

- 인덱스(Index)
데이터베이스의 테이블에 대한 검색 속도를 향상시켜 주는 자료구조
- 트리거(Trigger)
테이블에 대한 이벤트에 반응해 자동으로 실행되는 작업

- AVG, SUM, STDDEV는 각 열은 수치 데이터만 포함이 가능하고, COUNT는 어떠한 데이터 타입에서도 사용 가능하다.

4) SQL 주요 구문

쿼리명	설 명
WHERE	· SELECT, UPDATE, DELETE문 등에서 특정 레코드에 대한 조건을 설정할 때 사용되는 구문
ORDER BY	· 데이터를 지정된 컬럼으로 정렬하기 위한 구문으로 기본적으로 오름차순으로 정렬하며, desc는 내림차순 정렬을 의미함
GROUP BY	· 데이터를 그룹별로 나눠 합계, 평균 등의 연산을 할 경우 사용하는 구문
HAVING	· GROUP BY를 통해 그룹별 연산 함수들의 결과값에 조건식을 달기 위해 사용하는 구문. 독립적으로 사용될 수 있지만, GROUP BY와 함께 사용되는 경우가 많음 · 연산 함수들의 결과값은 직접 WHERE절에서 조건식으로 사용될 수 없으며, WHERE은 ROW 레벨 필터링을 제공하는 반면, HAVING은 GROUP 레벨 필터링을 제공

5) 간단한 SQL 문장 해석

```
SELECT NAME, GENDER, SALARY
FROM CUSTOMERS
WHERE AGE BETWEEN 20 AND 39
```

- 첫 번째 줄의 SELECT는 하나 또는 그 이상의 테이블에서 데이터를 추출하는 명령어이다.
NAME, GENDER, SALARY는 추출하고자하는 데이터명이다.
- FROM은 테이블을 지정해주는 명령어로서 CUSTOMERS라는 테이블을 지정하고 있다.
- WHERE는 데이터를 추출하는 선택 조건식을 지정하는 명령어이다.
AGE가 20과 39 사이의 데이터를 추출하는 것을 뜻한다.

```
SELECT CUSTOMER_NAME, 고객명, CUSTOMER_ENAME, 고객영문명
FROM CUSTOMER
WHERE CUSTOMER_ENAME LIKE '_A%'
```

- 위의 예제와 동일한 형태의 SELECT, FROM, WHERE구문이 활용되었으며, 새로 등장한 LIKE 구문에 대해 확인해보자.
- LIKE 연산자는 문자열의 패턴을 검색하는데 사용하며, %는 모든 문자, _는 한 글자를 의미한다. '_A%'는 맨 앞에 한 글자 뒤에 'A' 글자가 있는 ROW를 출력한다.

가. 개인정보 비식별 기술

- 비식별 기술이란 데이터 셋에서 개인을 식별할 수 있는 요소를 전부 또는 일부를 삭제하거나 다른값으로 대체하는 등의 방법으로 개인을 알아볼 수 없도록 하는 기술을 일컫는다.

비식별 기술의 종류와 예

비식별 기술	내 용	예 시
데이터 마스킹	데이터의 길이, 유형, 형식과 같은 속성을 유지한 채, 새롭고 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, 서울거주, 국내대 재학
총계처리	데이터의 총합 값을 보임으로서 개별 데이터의 값을 보이지 않도록 함. 단, 특정 속성을 지닌 개인으로 구성된 단체의 속성 정보를 공개하는 것은 개인 정보를 공개하는 것과 마찬가지로 주의해야 함	임꺽정 180cm, 홍길동 170cm, 이콩쥐 160cm, 김팔쥐 150cm → 물리학과 학생 키 합: 660cm, 평균키 165cm
데이터값 삭제	데이터 공유, 개방 목적에 따라 데이터 셋에 구성된 값 중에 필요 없는 값 또는 개인 식별에 중요한 값을 삭제. 개인과 관련된 날짜 정보(자격취득일자, 합격일 등)은 연단위로 처리	홍길동, 35세, 서울 거주, 한국대 졸업 → 35세, 서울 거주, 주민등록 번호 901206-1234567 → 90년대 생, 남자
데이터 범주화	데이터의 값을 범주의 값으로 변환하여 값을 숨김	홍길동, 35세 → 홍씨, 30~40세

2

Data에 관련된 기술

- 난수화
데이터를 특정한 순서나 규칙을 가지지 않는 무작위 숫자로 변환
- 익명화
데이터에 포함된 개인 식별 정보를 삭제하거나 알아볼 수 없는 형태로 변환

나. 무결성과 레이크

1) 데이터 무결성(Data Integrity)

- 데이터베이스 내의 데이터에 대한 정확한 일관성, 유효성, 신뢰성을 보장하기 위해 데이터 변경/수정 시 여러 가지 제한을 두어 데이터의 정확성을 보증하는 것을 말한다. 무결성제한의 유형은 개체 무결성(Entity Integrity), 참조 무결성(Referential Integrity), 범위 무결성(Domain Integrity)이 있다.

2) 데이터 레이크(Data Lake)

- 수 많은 정보 속에서 의미있는 내용을 찾기 위해 방식에 상관없이 데이터를 저장하는 시스템으로, 대용량의 정형 및 비정형 데이터를 저장할 뿐만 아니라 접근도 쉽게 할 수 있는 대규모의 저장소를 의미한다. Apache Hadoop, Teradata Integrated Big Data Platform 1700 등과 같은 플랫폼으로 구성된 솔루션을 제공하고 있다.

가. 하둡(Hadoop)

- 하둡은 여러 개의 컴퓨터를 하나인 것처럼 묶어 대용량 데이터를 처리하는 기술이다. 분산파일 시스템(HDFS)을 통해 수 천대의 장비에 대용량 파일을 저장할 수 있는 기능을 제공하고 맵리듀스(Map Reduce)로 HDFS에 저장된 대용량의 데이터들을 대상으로 SQL을 이용해 사용자의 질의를 실시간으로 처리하는 기술로 이루어져 있다.
- 하둡의 부족한 기능을 서로 보완하는 ‘하둡 에코시스템’이 등장하여 다양한 솔루션을 제공한다.

나. Apache Spark

- 실시간 분산형 컴퓨팅 플랫폼으로써 스칼라로 작성이 되어 있지만 스칼라, 자바, R, 파이썬, API를 지원한다. In-Memory 방식으로 처리를 하기 때문에 하둡에 비해 처리속도가 빠른 것이 특징이다.

다. Smart Factory

- 공장 내 설비와 기계에 사물인터넷(IoT)이 설치되어, 공정 데이터가 실시간으로 수집되고 데이터에 기반한 의사결정이 이뤄짐으로써 생산성을 극대화할 수 있는 기술이다.

라. Machine Learning & Deep Learning

- 머신 러닝은 인공지능의 연구 분야 중 하나로, 인간의 학습 능력과 같은 기능을 컴퓨터에서 실현하고자하는 기술 및 기법이다.
- 딥 러닝은 컴퓨터가 많은 데이터를 이용해 사람처럼 스스로 학습할 수 있게 하기 위하여 인공 신경망(Artificial Neural Network, ANN) 등의 기술을 기반으로 구축한 기계 학습 기술 중 하나이다.
- 대표적인 딥러닝 기법으로는 DNN, CNN, RNN, LSTM, Autoencoder, RBM 등이 있으며, 음성, 영상인식, 자연어처리 등의 여러 분야에서 활용되고 있다.
- 이러한 딥러닝을 구현하기 위한 소프트웨어 라이브러리로는 Tensorflow, Caffe, Torch, Theano, Gensim 등이 있다.

가. 데이터양의 단위

단 위	데이터량	단 위	데이터량
바이트(B)	1byte, 2^0B	페타바이트(PB)	1024TB, 2^{50}B
킬로바이트(KB)	1024B, 2^{10}B	엑사바이트(EB)	1024PB, 2^{60}B
메가바이트(MB)	1024KB, 2^{20}B	제타바이트(ZB)	1024EB, 2^{70}B
기가바이트(GB)	1024MB, 2^{30}B	요타바이트(YB)	1024ZB, 2^{80}B
테라바이트(TB)	1024GB, 2^{40}B		

나. B2B와 B2C

1) B2B

- 기업과 기업 사이의 거래를 기반으로 한 비즈니스 모델을 의미하며, 기업이 필요로 하는 장비, 재료나 공사입찰 등이 있다.

2) B2C

- 기업과 고객 사이의 거래를 기반으로 한 비즈니스 모델을 의미하며, 이동통신사, 여행회사, 신용카드회사, 옥션, 지마켓 등이 있다.

다. 블록체인

- 블록체인(Block Chain) : 거래정보를 하나의 덩어리로 보고 이를 차례로 연결한 거래장부다.
- 기존 금융회사의 경우 중앙 집중형 서버에 거래 기록을 보관하는 반면, 블록체인은 거래에 참여하는 모든 사용자에게 거래 내역을 보내 주며 거래 때마다 이를 대조해 데이터 위조를 막는 방식을 사용한다.

라. 데이터의 유형

- 메타데이터(Meta Data)
데이터에 관한 구조화된 데이터로, 다른 데이터를 설명해주는 데이터
- 스키마(Schema)
데이터베이스의 구조와 제약 조건에 관한 전반적인 명세를 기술한 메타데이터의 집합

유형	내용	예시
정형 데이터	<ul style="list-style-type: none"> • 형태(고정된 필드)가 있으며, 연산이 가능함. 주로 관계형 데이터베이스(RDBMS)에 저장됨 • 데이터 수집 난이도가 낮고 형식이 정해져 있어 처리가 쉬운 편 	관계형 데이터베이스, 스프레드시트, CSV 등
반정형 데이터	<ul style="list-style-type: none"> • 형태(스키마, 메타데이터)가 있으며, 연산이 불가능. 주로 파일로 저장됨 • 데이터 수집 난이도가 중간. 보통 API 형태로 제공되기 때문에 데이터처리 기술(파싱)이 요구됨 	XML, HTML, JSON, 로그형태 (웹로그, 센서데이터) 등
비정형 데이터	<ul style="list-style-type: none"> • 형태가 없으며, 연산이 불가능. 주로 NoSQL에 저장됨 • 데이터 수집 난이도가 높으며 텍스트 마이닝 혹은 파일일 경우 파일을 데이터 형태로 파싱해야 하기 때문에 수집 데이터 처리가 어려움 	소셜데이터(트위터, 페이스북), 영상, 이미지, 음성, 텍스트(word, PDF 등) 등

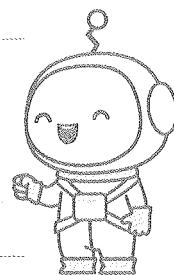
참고

※ XML이란?

- Extensible Markup Language의 약자로 다목적 마크업 언어(태그를 이용한 언어)이다.
- 인터넷에 연결된 시스템끼리 데이터를 쉽게 주고받을 수 있게 하여 HTML의 한계를 극복할 목적으로 만들어졌다.
- XML 기반 언어는 XHTML, SVG 등이 있다.



A small, friendly-looking cartoon robot is positioned in the bottom right corner of the page. It has a round head with a small antenna on top, a smiling face with two eyes and a wide mouth, and a simple body with a rectangular torso and two legs. It appears to be waving or gesturing with its right hand.



데이터의 이해

✓ 15회 기출

01 데이터는 그 형태에 따라 정성 데이터와 정량 데이터로 구분된다. 다음 중 정성 데이터에 속하는 것은?

- ① 풍향 ② 습도 ③ 기상특보 ④ 1시간 강수량

✓ 23회 기출

02 암묵지와 형식지의 상호작용 관계를 가장 적절하게 표현한 것은 무엇인가?

- ① 내면화 → 연결화 → 표출화 → 공통화
 ② 표출화 → 공통화 → 내면화 → 연결화
 ③ 공통화 → 표출화 → 연결화 → 내면화
 ④ 연결화 → 내면화 → 표출화 → 공통화

✓ 15회 기출

03 SQL은 다양한 집계함수를 제공하는데, 다음 집계함수 중 어떠한 데이터의 타입에도 사용이 가능한 것은?

- ① AVG ② COUNT ③ SUM ④ STDDEV

✓ 21회 기출

04 다음 중 개인정보 비식별화 기법을 설명한 것으로 부적절한 것은?

- ① 가명처리 - 개인 식별이 가능한 데이터에 대하여 직접적으로 식별 할 수 없는 다른 값으로 대체
 ② 범주화 - 단일 식별 정보를 해당 그룹의 대표 값으로 변환
 ③ 데이터마스킹 - 개인 정보 식별이 가능한 특정 데이터 값 삭제 처리
 ④ 총계처리 - 개별 데이터 값을 총합 또는 평균값으로 대체하는 것

✓ 12회 기출

05 다음 중 데이터에 대한 설명으로 가장 적절하지 않는 것은 무엇인가?

- ① 양질의 데이터를 확보하지 못하면 잘못된 분석 결과를 얻음
 ② 창의적인 데이터 매시업(Mashup)은 기존에 풀기 어려웠던 문제 해결에 도움
 ③ 비정형 데이터는 데이터 내부에 메타 데이터를 갖고 있으며 일반적으로 파일 형태로 저장
 ④ 공공부문에서 개방하고 있는 대표적인 데이터는 교통 데이터, 물가 데이터, 의료 데이터이다.

✓18회 기출

06 개인에게 내재된 경험을 객관적인 데이터로 문서나 매체에 저장, 가공, 분석하는 과정은?

- ① 연결화
- ② 내면화
- ③ 표출화
- ④ 공통화

✓22회 기출

07 다음 중 그 자체로는 의미가 중요하지 않은 객관적인 사실인 데이터를 가공 및 처리하여 얻을 수 있는 것으로 부적절한 것은 ?

- ① 정보
- ② 지혜
- ③ 지식
- ④ 기호

✓15회 기출

08 다음 중 지식(Knowledge)에 대한 예시로 가장 적절한 것은?

- ① A사이트보다 B사이트가 다른 물건도 비싸게 팔 것이다.
- ② B사이트보다 가격이 상대적으로 저렴한 A사이트에서 USB를 사야겠다.
- ③ A사이트는 10,000원에, B사이트는 15,000원에 USB를 팔고 있다.
- ④ B사이트의 USB 판매가격이 A사이트보다 더 비싸다.

✓14회 기출

09 다음 중 글로벌 기업의 빅데이터 활용사례로 그 연결이 부적절한 것은?

- ① 구글 - 실시간 자동 번역시스템을 통한 의사소통의 불편 해소
- ② 라쿠텐 - 이용자의 콘텐츠 기호를 파악하여 새로운 영화를 추천해주는 Cinematch 시스템 운영
- ③ 월마트 - 소셜 미디어를 통해 고객 소비 패턴을 분석하는 월마트랩(Walmart Labs) 운영
- ④ 자라 - 일일 판매량을 실시간 데이터 분석으로 상품 수요를 예측

✓25회 기출

10 다음은 데이터베이스의 구성요소들을 설명한 것이다. 각 설명에 해당하는 구성요소를 가장 적절하게 나열한 것은?

(A) 데이터에 관한 구조화된 데이터로, 다른 데이터를 설명해 주는 데이터
(B) 데이터베이스 내의 데이터를 신속하게 정렬하고 탐색하게 해주는 구조

- ① (A) - 메타데이터, (B) - 인덱스
- ② (A) - 데이터모델, (B) - 트리거
- ③ (A) - 백업데이터, (B) - 저장된 절차
- ④ (A) - 스키마구조, (B) - 데이터 마트

✓23회 기출

11 데이터웨어하우스는 기업 내의 의사결정지원 애플리케이션에 정보 기반을 제공하는 하나의 통합된 데이터 저장 공간을 말한다. 다음 중 데이터웨어하우스의 고유한 특성이 아닌 것은?

- ① 데이터웨어하우스에서는 데이터의 지속적 갱신에 따른 무결성 유지가 무엇보다 중요하다.
- ② 데이터웨어하우스의 데이터들은 전사적 차원에서 일괄된 형식으로 정의된다.
- ③ 데이터웨어하우스에서 관리하는 데이터들은 시간의 흐름에 따라 변화하는 값을 저장한다.
- ④ 데이터웨어하우스에서는 특정 주제에 따라 데이터들이 분류, 저장, 관리된다.

✓15회 기출

12 다음 중 주요 데이터 분석 기술에 대한 설명으로 가장 부적절한 것은?

- ① OLAP - 다차원의 데이터를 대화식으로 분석하기 위한 기술
- ② Business Intelligence - 데이터 기반 의사결정을 지원하기 위한 리포트 중심의 도구
- ③ Business Analytics - 의사결정을 위한 통계적이고 수학적인 분석에 초점을 둔 기법
- ④ Deep Learning - 대용량 데이터에서 의미있는 정보를 추출하여 의사결정에 활용하는 기술

✓16회 기출

13 아래는 특정산업의 일차원적 분석 사례를 나열한 것이다. 다음 중 특정산업으로 적절한 것은?

트레이딩, 공급, 수요예측

- ① 소매업 ② 에너지 ③ 운송업 ④ 금융서비스

✓ 12회 기출

14 다음 중 기업내부 데이터베이스인 고객관계관리(CRM)에 대한 설명으로 적절한 것은 무엇인가?

- ① 부품의 설계, 제조, 유통 등의 공정 포함
- ② 외부 공급업체와의 정보시스템 통합으로 시간과 비용 최적화
- ③ 기업의 내부 고객들만을 대상으로 한 정보시스템
- ④ 단순한 정보의 수집에서 탈피, 분석 중심의 시스템 구축 지향

✓ 19회 기출

15 아래는 데이터베이스를 기반으로 기업 내 구축되는 주요 정보시스템 중 하나를 설명한 것이다. 보기에서 가장 적합한 것을 고르시오.

기업 전체를 경영자원의 효과적 이용이라는 관점에서 통합적으로 관리하고 경영의 효율화를 기하기 위한 시스템

- ① ERP
- ② CRM
- ③ SCM
- ④ KMS

✓ 13회 기출

16 다음 중 사회기반 구조로서의 데이터베이스에 대한 설명으로 가장 부적절한 것은?

- ① 물류, 무역, 조세 등 사회간접자본 차원에서 정보망을 통해 유통, 이용된 정보가 데이터베이스로 구축
- ② 지리, 교통 부문에서 데이터베이스가 보다 고도화되어 데이터베이스를 구축
- ③ 인터넷의 보편화로 데이터베이스가 사회 전반의 인프라로 자리매김
- ④ 의료, 교육, 행정 부문에서는 데이터베이스 구축과 활용이 활성화되지 못함

✓ 18회 기출

17 러셀 L. 액오프가 1989년에 이야기한 DIKW Hierarchy 는 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 예상된다.
- ④ 8월 A상품 구매 고객의 80%가 40대 여성 고객으로 대부분 회사원이다.

✓22회 기출

18 다음 중 일반적으로 통용되고 있는 빅데이터의 정의와 거리가 가장 먼 것은?

- ① 빅데이터는 일반적인 데이터베이스 소프트웨어로 저장·관리·분석할 수 있는 범위를 초과하는 규모의 데이터다.
- ② 빅데이터는 다양한 종류의 대규모 데이터로부터 저렴한 비용으로 가치를 추출하고 데이터의 초고속 수집·발굴·분석을 지원하도록 고안된 차세대 기술 및 아키텍처이다.
- ③ 빅데이터는 데이터의 양(Volume), 데이터 유형과 소스 측면의 다양성(Variety), 데이터 수집과 처리 측면에서 속도(Velocity)가 급격히 증가하면서 나타난 현상이다.
- ④ 빅데이터는 기존의 작은 데이터 처리 분석으로는 얻을 수 없었던 통찰과 가치를 하둡(Hadoop)을 기반으로 하는 대용량의 분산처리 기술을 통해 창출하는 새로운 방식이다.

✓16회 기출

19 빅데이터 활용에 필요한 기본적인 3요소로 가장 적절한 것은?

- ① 데이터, 기술, 인력
- ② 데이터, 기술, 프로세스
- ③ 기술, 인력, 프로세스
- ④ 데이터, 인력, 프로세스

✓21회 기출

20 다음 중 빅데이터 현상이 출현하게 된 배경과 가장 거리가 먼 것은?

- ① 의료정보 등 공공데이터의 개방 가속화
- ② M2M, Iot와 같은 통신 기술의 발전
- ③ 하둡 등 분산처리 기술의 발전
- ④ 트위터, 페이스북 등 SNS의 급격한 확산

✓13회 기출

21 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

✓ 12회 기출

22 빅데이터의 기능 중 '공동 활용의 목적으로 구축된 유무형의 구조물 역할을 수행한다.'라는 것에 해당하는 내용은 무엇인가?

- ① 산업혁명 시대의 석탄, 철 ② 21세기의 원유
- ③ 렌즈 ④ 플랫폼

✓ 21회 기출

23 다음 중 빅데이터가 만들어 내는 변화와 가장 거리가 먼 것은?

- ① 가치가 있을 것이라고 예상되는 특정한 정보만 모아서 처리하는 것이 아니라 가능한 한 많은 데이터를 모으고 그 데이터를 다양한 방식으로 조합해 숨은 정보를 찾아내는 방식이 중요해진다.
- ② 데이터의 규모가 증가함에 따라 사소한 몇 개의 오류 데이터는 분석결과에 영향을 미치지 않기 때문에 데이터세트에 포함하여 분석해도 상관없는 경우가 많아진다.
- ③ 데이터의 양이 증가하고 유형이 복잡해짐에 따라 수많은 데이터 중에서 분석에 필요한 데이터를 선정하기 위해 정교한 표본조사 기법의 중요성이 대두되고 있다.
- ④ 인과관계의 규명 없이 상관관계 분석 결과만으로도 인사이트를 얻고 이를 바탕으로 수익을 창출할 수 있는 기회가 점차 늘어나고 있다.

✓ 20회 기출

24 다음 중 상품, 서비스, 기술 등의 기반 위에 다른 이해관계자들이 보완적인 상품, 서비스, 기술을 제공하는 생태계 구축을 목표로 하는 비즈니스 모델은?

- ① 플랫폼형 비즈니스 모델
- ② 가치사슬형 비즈니스 모델
- ③ 사회적 가치 기반형 비즈니스 모델
- ④ 고객 중심형 비즈니스 모델

✓ 12회 기출

25 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.

✓17회 기출

26 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

✓23회 기출

27 다음 중 감성 분석(Sentimental Analysis)에 대한 설명으로 가장 부적절한 것은?

- ① 특정 주제에 대한 사용자의 긍정·부정 의견을 분석한다.
- ② 주로 온라인 쇼핑몰에서 사용자의 상품평에 대한 분석이 대표적 사례이다.
- ③ 사용자간의 소셜 관계를 알아내고자 할 때 이용한다.
- ④ 사용자가 사용한 문장이나 단어가 분석 대상이 된다.

✓20회 기출

28 다음 중 비즈니스 모델에서 빅데이터 분석 방법과 사례를 연결한 것으로 부적절한 것은?

- ① 맥주를 사는 사람은 콜라도 같이 구매하는 경우가 많은가? - 연관규칙학습
- ② 택배차량을 어떻게 배치하는 것이 가장 비용 효율적인가? - 유형분석
- ③ 친분관계가 승진에 어떤 영향을 미치는가? - 소셜네트워크분석
- ④ 고객의 만족도가 충성도에 어떤 영향을 미치는가? - 회귀분석

✓17회 기출

29 아래 빅데이터 활용을 위한 기본 테크닉 중 어떤 사례에 해당하는가?

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

- ① 회귀분석
- ② 연관성분석
- ③ 유형분석
- ④ 구문분석

✓14회 기출

30 다음 핀테크 분야에서 빅데이터 활용이 가장 핵심적인 분야인 것은?

- ① 크라우드 펀딩(Crowd Funding)
- ② 신용평가(Credit Rating)
- ③ 간편결제(Simple Payment)
- ④ 블록체인(Block Chain)

✓25회 기출

31 다음 중 딥러닝과 가장 관련 없는 분석 기법은?

- ① CNN
- ② LSTM
- ③ SVM
- ④ Autoencoder

✓13회 기출

32 최근에 딥러닝(Deep Learning)에 대한 관심이 전 세계적으로 높아지며, 딥러닝을 활용하기 위해 다양한 오픈소스가 개발되어 제공되고 있다. 다음 중 이와 가장 관련이 없는 것은?

- ① Caffe ② Tensorflow ③ Anaconda ④ Theano

✓14회 기출

33 다음 중 빅데이터 시대에 발생할 수 있는 위기 요인으로 가장 부적절한 것은?

- ① 재산권 침해 ② 데이터 오용 ③ 책임원칙 훼손 ④ 사생활 침해

✓25회 기출

34 다음 중 빅데이터 시대에 발생할 수 있는 위기 요인 중 사생활 침해 문제를 해결하기 위한 방법으로 가장 적절한 것은?

- ① 알고리즘 접근 허용
- ② 결과기반 책임 원칙 고수
- ③ 데이터 오용 방지
- ④ 정보 사용자 책임제로 변환

✓23회 기출

35 아래에서 빅데이터 시대의 위기와 통제에 대한 설명으로 가장 타당한 것끼리 묶은 것은?

- 가) 데이터 익명화(Anonymization)는 사생활 침해에 대한 근본요인을 차단할 수 있어 빠른 기술발전이 필요하다.
 나) 빅데이터 분석은 일어난 일에 대한 데이터에 의존하므로 예측의 정확도는 높지만 항상 맞을 수는 없어 데이터 오용의 피해가 발생할 수 있다.
 다) 개인정보 사용자의 정보사용에 대한 무한책임의 한계로 개인정보 사용 책임제 보다 동의제를 더욱 강화시켜야 한다.
 라) 민주주의에서 '행동결과'에 따른 처벌의 모순을 교훈삼아 빅데이터 사전 '성향' 분석을 통한 통제가 강화될 필요가 있다.
 마) 빅데이터가 발생시키는 문제를 중간자 입장에서 중재하며 해결해 주는 알고리즘리스트(Algorithmist)도 새로운 직업으로 부상하게 될 것이다.

- ① 가, 다 ② 나, 다 ③ 가, 라 ④ 나, 마

✓14회 기출

36 다음 중 데이터화(Datafication) 현상에 큰 영향을 미치는 기술로 적절한 것은?

- ① 사물인터넷(Internet of Things)
 ② 인공지능(Artificial Intelligence)
 ③ 가상현실(Virtual Reality)
 ④ 3D 프린팅(3D-Printing)

✓17회 기출

37 다음 중 사용자 정의 데이터 및 멀티미디어 데이터 등 복잡한 데이터 구조를 표현, 관리할 수 있는 데이터베이스 관리 시스템은 무엇인가?

- ① 관계형 DBMS ② 객체지향 DBMS
 ③ 네트워크 DBMS ④ 계층형 DBMS

✓22회 기출

38 데이터 사이언스는 데이터 처리와 관련된 IT 영역, 분석적 영역, 그리고 비즈니스 컨설팅 영역을 포괄하고 있다. 다음 중 세 개의 영역과 다른 영역에 속하는 하나는?

- ① 데이터 시각화
 ② 데이터 웨어하우징
 ③ 분산 컴퓨팅
 ④ 파이썬 프로그래밍

✓18회 기출

39 데이터 사이언스에서 인문학적 사고는 반드시 필요한 요소이다. 다음 중 인문학 열풍을 가져오게 한 외부 환경 요소로 가장 부적절한 것은?

- ① 디버전스 동역학이 작용하는 복잡한 세계화
- ② 비즈니스 중심이 제품생산에서 체험 경제를 기초로 한 서비스로 이동
- ③ 경제의 논리가 생산에서 최근 패러다임인 시장 창조로 변화
- ④ 빅데이터 분석 기법의 이해와 분석 방법론 확대

✓22회 기출

40 빅데이터를 다각적으로 분석하여 인사이트를 도출하는 데이터 사이언티스트의 필요 역량이 아닌 것은?

- ① 통찰력 있는 분석 능력
- ② 다분야 간 커뮤니케이션 능력
- ③ 뉴럴네트워크 최적화 능력
- ④ 설득력 있는 스토리텔링 능력

✓17회 기출

41 데이터 사이언스에 대한 설명으로 가장 부적절한 것은?

- ① 데이터 사이언스는 데이터로부터 의미있는 정보를 추출하는 학문이다.
- ② 주로 분석의 정확성에 초점을 두고 진행한다.
- ③ 정형데이터 뿐만 아니라 다양한 데이터를 대상으로 한다.
- ④ 기존의 통계학과는 달리 총체적 접근법을 사용한다.

✓15회 기출

42 고객테이블(CUSTOMERS)로부터 나이(AGE)가 20~30대인 고객정보(NAME, GENDER, SALARY)를 추출하기 위해 아래와 같은 SQL문을 작성하려고 한다. 다음 (가) 안에 들어갈 적절한 구문을 채워 쓰시오.

```
SELECT NAME, GENDER, SALARY
FROM CUSTOMERS
WHERE AGE ( 가 ) 20 AND 39
```

()

✓15회 기출

43 아래에서 설명하고 있는 (가)와 (나) 적절한 용어를 쓰시오.

데이터 사이언티스트가 갖춰야 할 역량은 빅데이터의 처리 및 분석에 필요한 이론적 지식과 기술적 숙련에 관련된 능력인 (가) skill 과 데이터 속에 숨겨진 가치를 발견하고 새로운 발전 기회를 만들어 내기 위한 능력인 (나) skill 로 나누어진다.

()

✓25회 기출

44 아래 (가) 안에 들어갈 용어를 기입하십시오.

(가)는 데이터의 가공 및 상관관계 간 이해를 통해 패턴을 인식하고 그 의미를 부여한 것이며, 지식을 도출하기 위한 재료가 된다.

()

✓14회 기출

45 아래 데이터 분석과 관련된 기술을 설명한 것이다. (가)에 들어갈 용어를 기입하십시오.

기업의 의사결정 과정을 지원하기 위한 주제 중심으로 통합적이며 시간성을 가지는 비휘발성 데이터의 집합을 (가)라고 한다.

()

✓13회 기출

46 아래에서 설명하고 있는 (㉠)은 무엇인가?

지난 몇 년간 여러 사일로 대신 하나의 데이터 소스를 추구하는 경향이 생겼다. 전사적으로 쉽게 인사이트를 공유하는 데 도움이 되기 때문이다. 다시 말해 별도로 정제되지 않은 자연스러운 상태의 아주 큰 데이터 세트인 (㉠)을/를 기업들이 구현하는 것은 2017년 새롭게 등장한 트렌드가 아니다. 그러나 2017년은 이를 적절히 관리해 운영하는 첫해가 될 전망이다.

()

✓16회 기출

47 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 DBMS에 저장, 검색, 분석하여 활용하기가 용이하다.

()

✓22회 기출

48 아래는 기업 내부에서 활용되는 데이터베이스의 활용에 대한 설명이다. (가)에 들어갈 말로 적절한 것은 무엇인가?

(가)은 기업이 외부 공급업체 또는 제휴업체와 통합된 정보시스템으로 연계하여 시간과 비용을 최적화시키기 위한 것으로, 자재 구매, 생산, 제고, 유통, 판매, 고객 데이터로 구성된다.

()

✓18회 기출

49 아래에서 빈칸에 공통적으로 들어갈 용어는?

가) 페이스북은 2006년 F8 행사를 기점으로 자신들의 소셜 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.

()

✓22회 기출

50 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 x 를 찾기 위해, 진화를 모방한(Simulated Evolution) 탐색 알고리즘이라고 말할 수 있다.

()

정답 및 해설

01	③	11	①	21	①	31	③	41	②
02	③	12	④	22	④	32	③	42	BETWEEN
03	②	13	②	23	③	33	①	43	가: 하드, 나: 소프트
04	③	14	④	24	①	34	④	44	정보
05	③	15	①	25	②	35	④	45	데이터 웨어하우스(Data Warehouse)
06	③	16	④	26	②	36	①	46	데이터 레이크
07	④	17	③	27	③	37	②	47	정량적 데이터
08	②	18	④	28	②	38	①	48	SCM(Supply Chain Management)
09	②	19	①	29	②	39	④	49	플랫폼(Platform)
10	①	20	①	30	②	40	③	50	유전자 알고리즘(Genetic Algorithms)

01. 정량적 데이터의 형태는 수치, 도형, 기호 등으로 기술이 되며, 정성 데이터의 형태는 언어, 문자 등으로 기술된다. (정답 : ③)

02. 암목지와 형식지의 상호작용 관계는 '공통화 → 표출화 → 연결화 → 내면화' 이다. (정답 : ③)

03. 보기의 SQL 집계함수를 정리하면 아래와 같다. (정답 : ②)

함수명	설 명	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

04. 데이터 마스킹은 식과 같은 속성을 유지한 채, 새롭고 읽기 쉬운 데이터를 익명으로 생성하는 기술이다. (정답 : ③)

05. 데이터 내부에 메타 데이터를 갖고 있으며 일반적으로 파일형태로 저장되는 것은 반정형 데이터이다. (정답 : ③)

06. 표출화는 형식지 요소 중 하나로 개인에게 내재된 경험을 객관적인 데이터로 문서나 매체에 저장, 가공, 분석하는 과정이다. (정답 : ③)

07. DIKW 피라미드에서 개별 데이터 자체로는 의미가 중요하지 않은 객관적 사실에서 데이터를 가공 및 처리하여 정보, 지식, 지혜를 얻을 수 있다. (정답: ④)
08. 지식은 상호 연결된 정보 패턴을 이해하여 이를 토대로 예측한 결과물이다. (정답: ②)
09. Cinematch 시스템은 넷플릭스(Netflix)에서 개발한 영화 추천 알고리즘이다. (정답: ②)
10. 메타데이터는 데이터에 대한 데이터로서 하위레벨의 데이터를 설명/기술하려는 것이며, 인덱스는 데이터베이스의 테이블에서 고속의 검색동작뿐만 아니라 레코드 접근과 관련 효율적인 순서 매김 동작에 대한 기초를 제공한다. (정답: ①)
11. 데이터웨어하우스는 데이터의 주제 지향성, 데이터 통합, 데이터의 시계열성, 데이터의 비휘발성이라는 4가지 특성을 갖는다. (정답: ①)
12. 데이터 마이닝(Data Mining)은 대용량 데이터에서 의미있는 정보를 추출하여 의사결정에 활용하는 기술이다. 딥 러닝(Deep Learning)은 다층구조 형태의 신경망을 바탕으로 하는 머신 러닝의 한 분야이다. (정답: ④)
13. 산업별 분석 애플리케이션에서 분석 사례 중 에너지는 트레이딩, 공급/수요 예측 등이 있다. (정답: ②)
14. CRM은 데이터베이스를 기초로 고객을 세부적으로 분류하여 효과적이고 효율적인 마케팅 전략을 개발한다. (정답: ④)
15. ERP(Enterprise Resource Planning)는 인사·재무·생산 등 기업의 전 부문에 걸쳐 독립적으로 운영되던 각종 관리 시스템의 경영자원을 하나의 통합 시스템으로 재구축함으로써 생산성을 극대화하려는 경영혁신기법을 의미한다. (정답: ①)
16. 사회기반 구조로서의 데이터베이스는 물류, 지리/교통, 의료, 교육 등 부문에서 구축되었으며 활용이 되고 있다. (정답: ④)
17. ③은 지식에 해당하며 나머지 항목들은 정보에 해당하는 내용이다. (정답: ③)
18. 빅데이터란 대용량 데이터를 활용해 작은 용량에서는 얻을 수 없었던 새로운 통찰이나 가치를 추출해내는 일이다. 하둡은 빅데이터 플랫폼 환경 구축을 위해 사용할 뿐 빅데이터가 하둡을 기반으로 하는 것은 아니다. (정답: ④)
19. 빅데이터 활용의 기본 3요소는 데이터, 기술, 인력이다. (정답: ①)
20. 빅데이터 출현 배경에는 고객데이터의 축적과 거대 데이터의 활용이 늘어남으로 필요한 기술 아키텍처 및 통계 도구들의 발전, 모바일 혁명 등의 관련기술의 발달을 들 수 있다. (정답: ①)
21. 빅데이터의 수집, 구축, 분석의 최종 목적은 기존 방식으로는 얻을 수 없었던 통찰 및 가치 창출, 사업방식, 시장, 사회, 정부 등에서 변화와 혁신 주도이다. (정답: ①)
22. 플랫폼이란 비즈니스 측면에서는 일반적으로 '공동 활용의 목적으로 구축된 유무형의 구조물'을 의미하며 빅데이터가 최근에는 다양한 서드파티 비즈니스에 활용되면서 플랫폼 역할을 할 것으로 전망된다. (정답: ④)

23. 빅데이터의 등장으로 데이터 수집비용의 감소와 클라우드 컴퓨팅 기술의 발전으로 데이터 처리비용이 감소하게 되었다. 이로 인해 표본을 조사하는 기존의 지식발견 방식에서 전수조사를 통해 샘플링이 주지 못하는 패턴이나 정보를 발견하는 방식으로 데이터 활용방법이 변화되었다. (정답 : ③)
24. 플랫폼형 비즈니스 모델은 서비스, 기술 등의 기반 위에 다른 이해관계자들이 보완적인 상품, 서비스, 기술을 제공하는 생태계 구축을 목표로 하는 모델이다. (정답 : ①)
25. 데이터의 가치를 측정하기 어려운 이유는 다음과 같다. (정답 : ②)
- 데이터 활용 방식 : 재사용, 재조합(Mashup), 다목적용 개발
 - 새로운 가치 창출
 - 분석 기술 발전
26. 데이터 난수화를 사용하면 고객의 과거 구매기록이나 나이, 수입, 건강정보와 같은 데이터가 해독이 불가능한 난수화를 통해 변경된 채로 기업에 전송된다. (정답 : ②)
27. 사용자간의 소셜 관계를 알아내고자 할 때 이용하는 분석은 소셜 네트워크 분석(Social Network Analysis)이다. (정답 : ③)
28. 유형분석은 문서를 분류하거나 조직을 그룹으로 나눌 때 또는 온라인 수강생들을 특성에 따라 분류할 때 사용하는 기법으로 사용자가 어떤 특성을 가진 집단에 속하는지 알아볼 때 사용한다. (정답 : ②)
29. 연관성분석은 기업의 데이터베이스에서 상품의 구매, 서비스 등 일련의 거래 또는 사건들 간의 규칙을 발견하기 위한 분석으로 흔히 장바구니 분석 등이 있다. (정답 : ②)
30. 신용평가(Credit Rating)은 투자자 보호를 위하여 금융상품 및 신용공여 등에 대하여 그 원리금이 상환될 가능성과 기업·법인 및 간접투자기구 등의 신용도를 평가하는 행위이며 핀테크 분야에서 빅데이터 활용이 활발하게 이루어지고 있다. (정답 : ②)
31. SVM은 분류분석의 기법 중 하나로 딥러닝과 관련 없는 분석 기법이다. (정답 : ③)
32. Anaconda는 Python프로그램의 Machine Learning기능을 강화해 주는 소프트웨어이다. Caffe, Tensorflow, Theano는 Deep Learning 소프트웨어이다. (정답 : ③)
33. 빅데이터 시대에 발생할 수 있는 위기 요인은 사생활 침해, 책임 원칙 훼손, 데이터 오용이 있다. (정답 : ①)
34. 사생활 침해 문제를 해결하기 위해서는 동의에서 책임으로 변화되어야 한다. 1번은 데이터 오용, 2번은 책임 원칙 훼손에 대한 해결 방법이다. (정답 : ④)
35. 다) 개인정보 사용자의 정보사용에 대한 무한책임의 한계로 개인정보 사용 동의제보다 책임제로 더욱 강화시켜야 한다. 라) 민주주의 국가의 형사 처벌과 같이 잠재적 위험이 아닌 명확하게 행동한 결과에 대해 책임을 묻기 때문에 빅데이터 사전 성향 분석을 실시한다면 책임 원칙을 훼손한다. (정답 : ④)

36. 사물인터넷(Internet of Things)은 인터넷을 기반으로 모든 사물을 연결해 사람과 사물, 사물과 사물 간의 정보를 상호 소통하는 지능형 기술 및 서비스이며, 사물에서 생성되는 Data를 활용한 분석을 통해 마케팅 등에 활용할 수 있다.
(정답: ①)
37. 객체지향DB는 일반적으로 사용되는 테이블 기반의 관계형DB와 다르게 정보를 '객체' 형태로 표현하는 데이터베이스 모델로 멀티미디어 등 복잡한 데이터 구조를 관리하는 DBMS이다. (정답: ②)
38. 데이터시각화는 비즈니스 컨설팅 영역이며 나머지 3개는 데이터처리와 관련된 IT영역이다. (정답: ①)
39. 컨버전스에서 디버전스로의 변화, 생산에서 서비스로의 변화, 생산에서 시장창조로의 변화가 인문학 열풍을 가져오게 한 외부환경 요소이다. (정답: ④)
40. 데이터사이언티스트에 요구되는 역량으로는 빅데이터에 대한 이론적 지식, 분석 기술에 대한 숙련, 통찰력 있는 분석, 설득력 있는 전달, 다분야간 협력이 있다. (정답: ③)
41. 데이터 사이언스는 통찰력 있는 분석에 초점을 두고 진행한다. (정답: ②)
42. 정답: BETWEEN
43. 정답: 가 - 하드, 나 - 소프트
44. 정답: 정보
45. 정답: 데이터웨어하우스(Data Warehouse)
46. 정답: 데이터 레이크
47. 정답: 정량적 데이터
48. 정답: SCM(Supply Chain Management)
49. 정답: 플랫폼(Platform)
50. 정답: 유전자 알고리즘(Genetic Algorithms)