[전사원 UP!] 빅데이터로 관통하는 4차산업의 미래
학습 운영 목차

[전사원 UP!] 빅데이터로 관통하는 4차산업의 미래 학습요약 자료집

<01_쉽게 이해하는 빅데이터의 개요>

● 빅데이터의 처리
패러다임이 지능화되고 개인화된 시대를 '빅데이터 시대'라고 합니다. 빅데이터의 개념이 등장하며 자연스럽게 데이터에 관심이 증가하게 되고, 정보통신 기술의 발달로 데이터의 규모와 유형 및 특성도 따라서 변화하게 됩니다.

● 빅데이터의 개념과 속성
가트너(Gartner)는 현재 가장 널리 사용하는 빅데이터의 속성을 3V, 즉 대규모(Volume), 다양성(Variety), 속도(Velocity) 등의 세 가지로 정의했습니다. IBM은 여기에 정확성(Veracity) 요소를 더해 4V라 정의했고, 최근에는 가치(Value)를 포함하여 5V라고 정의하기도 합니다.

● 빅데이터의 속성과 처리의 특징
빅데이터는 하드웨어부터 소프트웨어까지, 컴퓨터 공학에서 인간 공학, 심지어 뇌 과학과 언어학까지 총망라한 기술이 모두 적용된 분야입니다. 따라서 통계학, 경제학, 정보기술, 수학 등의 포괄적인 학문에 대한 이해가 필요합니다. 또 학문적인 지식 외에 통합적 사고와 직관력 등도 요구됩니다.

● 플랫폼 기술
-데이터의 자가 증식과 수집 및 정제 기술
-다양한 응용 패턴의 통합 지원 기술
-멀티모델 데이터의 통합, 고신뢰 데이터 관리 및 다각도 분석 기술
-초연결 데이터 관리 및 협업 기술
-플랫폼의 빅데이터 처리와 저장 및 관리 기술

● 플랫폼 분석 기술
-지능형 예측 분석 기술
-이종소스 심층 융합 분석 기술
-엣지 분석 및 협업 분석 기술
-모사현실 모델링 프레임워크

<02_디지털 뉴딜로 바라보는 빅데이터의 미래 - 1>

● 한국판 뉴딜 정책과 방향

한국판 뉴딜 정책은 크게 디지털(Digital) 뉴딜과 그린(Green) 뉴딜이라는 양대 축을 중심에 두고, 안전망 강화를 포함하는 3대 프로젝트와 10대 중점 추진 과제로 구성되었습니다. 디지털 뉴딜 정책은 DNA(Data-Network-AI) 생태계를 강화해 전 산업에 걸쳐 데이터와 디지털 혁신을 이루고자 합니다.

● 디지털 뉴딜 4개 분야
-DNA 생태계의 강화
-교육 인프라의 디지털 전환
-비대면 산업의 육성
-사회간접자본(SOC: Social Overhead Capital)의 디지털화

●
한국판 뉴딜 10대 중점 추진 과제
-데이터 댐
-지능형(AI) 정부
-스마트 의료 인프라
-그린 스마트 스쿨
-디지털 트윈
-국민 안전 SOC 디지털화
-스마트 그린 산단
-그린 리모델링
-그린 에너지
-친환경 미래 모빌리티

● 데이터 댐
많은 양의 데이터로 고도화한 인공지능 융합 서비스를 확산하는 사업입니다. 또 모든 산업에서 5G 이동통신망을 기반으로 데이터를 수집, 가공, 결합, 거래, 활용할 수 있습니다. 이를 구체화하기 위해 공공데이터를 개방하고, 데이터의 수집과 활용을 확대하며, 데이터의 거래를 활성화하고, 인공지능 학습용 데이터의 구축을 적극 추진합니다.

● 스마트 의료 인프라
-디지털 기반의 스마트 병원을 구축해 새로운 의료 모델을 도입합니다.
-감염병 대응을 위한 호흡기 전담 클리닉도 별도로 신설해 운영합니다.
-의원급 의료기관에 비대면 영상진료 장비를 지원해 안전한 진료 환경을 조성합니다.
-의료 데이터의 품질과 의료 서비스의 질을 향상하기 위해 전자의무기록(EMR: Electronic Medical Record)도 표준화합니다.

● 사회간접자본(SOC)의 디지털화
안전하고 효율적인 교통망의 구축을 위해 도로, 철도, 공항 등의 사회기반시설에 인공지능 기술과 디지털 관리 체계를 도입합니다. 사물인터넷 기반의 조기 위험 경보 시스템을 설치하고 단계적으로 자율주행의 기준을 마련하고 규제 완화를 추진하여 디지털 안전 체계를 확보합니

다.

<03_디지털 뉴딜로 바라보는 빅데이터의 미래 - 2>

● 데이터 3법
2020년에 「개인정보보호법」, 「정보통신망 이용촉진 및 정보보호 등에 관한 법률」, 「신용정보의 이용 및 보호에 관한 법률」 등이 공포되었습니다. 이와 같은 데이터 3법에 근거하여 데이터의 가치가 높아지고 활용 가능한 데이터의 종류가 다양해지고 있습니다. 이에 따라 새로운 혁신 서비스의 창출이 가능해질 것으로 전망되고 있습니다.

● 데이터 3법 개정법은 기존의 개인정보 보호 관련 법률과 그 모습을 달리했습니다. 개인정보 관련 체계를 개인정보, 가명정보, 익명정보로 나누어 명시하고 가명정보의 처리를 위한 상세한 규정을 신설했습니다. 또 개인정보보호위원회를 감독기구로 설치했습니다. 개정된 「개인정보보호법」에는 정보통신 서비스 제공자들의 개인정보 보고 관련 조항이 이관되었습니다.

● 데이터 3법 개정법의 주요 변화
-마이데이터(MyData) 산업이 본격적으로 진행됩니다.
-데이터 거래의 활성화를 바탕으로 다양한 산업의 생태계가 출현할 것입니다.
-신용평가 시장이 확대될 것입니다.
-기존의 금융 서비스 전달 체계의 변화와 핀테크 산업 영역의 확대가 예상됩니다.

● DNA(Data-Network-AI)
디지털 뉴딜의 키워드는 DNA이며, 데이터(Data), 네트워크(Network), 인공지능(AI)을 통해 생태계를 강화하고자 합니다. 이를 위해서는 데이터의 구축과 개발 및 활용을 통해 새로운 혁신 산업이 창출되어야 합니다.

● 디지털 뉴딜의 예상 성장 산업
- 클라우드 서비스 산업 [인프라 서비스(IaaS: Infrastructure as a Service), 플랫폼 서비스(PaaS: Platform as a Service), 소프트웨어 서비스(SaaS: Software as a Service) 등]
- 원격진료 등의 디지털 헬스케어 산업
- 스마트 스토어 등의 지급결제 산업
- 디지털 콘텐츠 산업
- 정보 보호 산업

● 데이터 거래소
데이터의 공급자와 수요자를 매칭하여 비식별 정보나 기업 정보 등의 데이터를 거래할 수 있는 중개 시스템이 '데이터 거래소'입니다. 데이터의 유통에 해당하는 검색, 계약, 결제, 분석 등을 원스톱 방식으로 지원합니다. 거래소 시스템을 통해 데이터의 모든 거래 절차를 진행할 수 있습니다.

<04_빅데이터 서비스를 위한 워밍업 - 목표 수립>

● 빅데이터 기반 서비스의 항목 발굴
빅데이터 서비스의 목표를 수립하기 위해서는 적용하려는 산업군의 서비스 발굴, 빅데이터 기술의 분석 이해, 적용 이해의 분석, 빅데이터 및 시스템 차원의 표준화 분석이 필요합니다. 이러한 분석과 해당 데이터를 통해 빅데이터 서비스의 목표를 수립하고 요구 사항을 식별하여 적용하려는 서비스를 기획합니다.

● 빅데이터 서비스의 요구 사항에 대한 식별 단계 프로세스
① 다양한 요구 사항을 청취한 뒤에 취합하는 단계
② 취합된 요구 사항과 그 해결을 위한 데이터 세트를 수집하는 단계
③ 데이터 세트를 통한 각각 이해 당사자 간의 협의체 구성 단계
④ 요구 사항을 도출하는 프로세스 단계

● 빅데이터 서비스 요구 사항의 도출
① 빅데이터 서비스의 요구 사항을 도출하기 위해 목록을 작성합니다.
② 목록 작성은 요구 사항을 도출하기 위한 단계로서 이러한 자가진단 체크리스트를 통해 다양한 요구 사항을 도출할 수 있습니다.
③ 중요도는 빅데이터 서비스의 요구 사항을 분석할 때 중요한 요소로서 자가진단이 잘 되어야 부족한 부분을 사전에 예방할 수 있습니다.

● BSR의 개념
BSR(Bigdata Service Request)는 다양한 빅데이터 서비스의 요구 사항을 통해 분석 및 활용에 대한 평가 기준을 나타낸 것입니다. 이는 현재의 서비스를 고도화할 수 있는 부분과 함께 지속적인 서비스 개선에 필요한 핵심 과정입니다.

● BSR 목적
- 모든 핵심적 서비스 단위를 선정하여 우선순위를 부여합니다.
- 구분의 산정은 빅데이터 서비스의 요구 사항 분석을 통해 우선적으로 모델을 수립해야 할 절차를 산정하는 것입니다.
- 기준과 결과에 따른 세부 기준을 통해 서비스 요구 사항의 결과로 산출하게 됩니다.
- 빅데이터 관점에서 살펴보면 서비스 관점과 데이터 관점으로 나눌 수 있는데 이 두가지 관점을 통합해서 다양한 분석 결과가 활용되는 효과적인 빅데이터 서비스 모델의 구축이 가능합니다.

● BSR 핵심 도출 사항 간의 관계 및 상세 내용
- BSR 핵심 도출 사항 간의 관계
- BSR 핵심 도출 사항 간의 우선순위 및 선·후행 실행 순위의 설정 필요
- BSR 핵심 도출 사항의 상세 내용
- 핵심 사항 도출에 의한 분류 기준의 우선순위 결정

● 빅데이터 서비스 분석 항목의 목표 및 기대 효과
① 목표: 서비스의 목표를 달성하기 위한 단계입니다.
② 기대 효과: 요구 사항을 통합적으로 수집 및 취합하면 서비스의 창출을 원활하게 진행하기 위한 서비스의 전반적인 일관성 유지 및 효율성이 증가합니다. 또 빅데이터 서비스를 담당하는 조직 구성원들의 역량 및 지식을 통해 적절하게 분배하게 되고, 따라서 효율의 증대가 가능해집니다.
지속적인 요구 사항을 수렴하여 개선함으로써 능동적(proactive)이고 예측 가능(predictive)한 서비스를 제공합니다. 이에 따라 전반적인 서비스 수준의 향상, 데이터에 대한 과도한 수집 및 관련 부대비용의 절감과 같은 기대 효과를 달성합니다.

<05_빅데이터 서비스를 위한 워밍업 - 모델 수립>

● 빅데이터 서비스 참조 모델
① 정보 저장형: 다양한 빅데이터를 기반으로 정보기술 자원을 효율적으로 저장하여 활용하는 모델입니다.
② 정보 제공형: 데이터를 융합하여 다양한 업무 및 서비스에 제공하는 모델입니다.
③ 데이터의 이상 탐지 제공형: 데이터의 이상 탐지를 기반으로 보안적인 업무와 이상 탐지를 활용하는 모델입니다.

● 분석 전략 계획의 수립 단계
① 자가진단을 통한 분석 준비 및 성숙도를 진단
② 분석에 요구되는 필요 사항의 발굴
③ 지식의 내재화
④ 사고의 내재화
⑤ 페르소나 분석

● 빅데이터의 서비스 활용에 따른 관리적 측면의 이해
비정형 데이터의 활용 측면에서 각종 분석 알고리즘 및 통계적 분석과의 차이점을 이해해야 합니다. 이는 서비스 모델을 수립하기 전에 어떤 종류의 데이터를 가지고 있는지, 어떤 종류의 추론인지, 어떤 종류의 분석 기법을 이용해 서비스 모델을 수립할 것인지를 정의하는 기본 정보로서 고려 사항에 반드시 포함되어야 합니다.

● 모델 활용의 구분
원가를 절감하기 위한 서비스 모델인지의 여부는 기업 내에 존재하는 각종 공정을 효율적으로 배치하는 것이 활용 가능한지의 여부와 관련이 있습니다. 이 부분은 제조 공정에서 빅데이터를 활용한 다양한 접근 방법에 대한 시도이며, 효율적인 공정을 설계한다면 다양한 부분에서 활용이 가능합니다.

● 비즈니스의 창출

데이터의 융합을 통해 혁신적인 비즈니스를 창출하는 부문입니다. 이 부문은 데이터를 통해 융합을 한다면 새로운 제품을 기획하거나 서비스를 기획할 경우 종래의 영역에서는 보지 못했던 다양한 접근이 가능합니다. 이와 같이 기존의 산업군이 다양한 융합을 거치면서 새로운 부문으로 진화하고 있기 때문에 융합이라는 키워드가 매우 중요해지고 있습니다.

● 빅데이터 서비스 모델의 사용자 요구 사항의 이해와 관련한 정보 예측 :
빅데이터에서는 예측이라는 영역이 더해지게 되며, 예측이라는 측면은 산업마다 고유한 특성을 가지고 있습니다. 정보에 기반을 둔 서비스별 특성이 더해져서 예측의 정확도가 높아집니다.
-데이터, 정보, 지식은 데이터의 영역이고 빅데이터를 중심으로 한 인공지능의 영역인 지혜의 영역까지 보조해 주는 중요한 영역입니다.
-지혜는 빅데이터를 기반으로 하는 다양한 서비스가 딥러닝이나 머신 러닝 등의 학습을 통해 원천 데이터가 지혜의 영역, 즉 인공지능의 영역으로 발전하게 되는 것입니다.

● 빅데이터 서비스 모델의 생성 과정
정보의 원천은 이미지, 텍스트, 보이스 등의 다양한 비정형 데이터 형태의 원천 데이터를 기반으로 가공됩니다. 비정형 데이터는 빅데이터를 통해 향후 인공지능으로 진전하는 데에 가장 중요한 기반 구조가 되며, 이러한 비정형 데이터의 효율적인 수집 및 가공으로 인해 패턴을 가진 의미 있는 데이터의 구조를 형성하게 됩니다.
이러한 정보는 데이터에 관한 일반적인 공식이나 경험을 통해 지식으로 발전하게 됩니다. 지식은 흔히 통계적인 기반을 중심으로 만들어진 공식과 이를 통한 경험치가 축적되어 발전한 모델을 의미합니다.

● 빅데이터 서비스의 참조 모델
빅데이터를 기반으로 하는 서비스의 산업별 또는 업무별로 다양한 모델들이 있습니다.
① 정보 저장형: 다양한 빅데이터를 기반으로 정보기술 자원을 효율적으로 저장하여 활용하는 모델입니다.
② 정보 제공형: 데이터를 융합하여 다양한 업무 및 서비스에 제공하는 모델입니다.
③ 데이터의 이상 탐지 제공형: 데이터의 이상 탐지를 기반으로 보안적인 업무와 이상 탐지를 활용하는 모델입니다.

<06_빅데이터 서비스를 위한 워밍업 - 전략과 정책>

● 4C 적용모델의 기본적 요소
콘텐츠(Contents), 커뮤니케이션(Communication), 커머스(Commerce), 커뮤니티(Community)를 지칭하는 말이며, 서비스 측면에서 갖추어야 할 기본적 요소를 말합니다.

● 빅데이터적 관점에서의 서비스 단계의 이해
빅데이터 서비스를 실행하기 위해서는 빅데이터적 관점에서의 서비스 실행 공정도의 5단계를 실행합니다. 이러한 5단계에 포함되는 부분은 도입과 운영에 따른 다양한 요소가 포함되며,

단계적인 로드맵에 따라 수행됩니다. 분석적 측면, IT 인프라적 측면, 활용의 측면, 성과에 대한 관리 및 전략적 측면이 포함되어 있습니다.

● 데이터 서비스 실행 5단계
-계획 단계: 대표적인 프로젝트 수행의 계획 영역입니다.
-분석 단계: 요구 사항의 정의, 요구 사항의 분석, 현행 운영 시스템의 분석, 업무 기능의 분석, 개념 데이터의 모델링과 분석의 검토가 포함됩니다.
-설계 단계: 업무적인 부분도 포함되어 있지만, 실제 이 단계는 데이터를 저장하는 논리적·물리적인 단계를 나타냅니다.
-구현 단계: 물리적인 데이터의 구현, 데이터의 수집에서 추출, 변형, 적재 등이 실제 구현되는 단계입니다.
-전개 단계: 빅데이터의 서비스에 기반을 둔 물리적인 시스템 환경을 시범적으로 운영하는 테스트의 단계입니다.

● 원천 데이터의 영역
빅데이터는 생성하는 과정에서 외부 또는 내부 데이터, 정형 또는 비정형 데이터로 나누어집니다. 빅데이터는 처음에는 원천 데이터에서 생성이 되는데, 인위적으로 생성되는 데이터뿐만 아니라 센서 등의 사물인터넷에서 생성되기도 합니다.

● 변환 정제의 단계
생성된 데이터는 원천 시스템 및 요건을 파악한 뒤에 서비스에 적용할 데이터의 용량 및 현행 시스템의 성능을 파악합니다. 이를 통해 수집된 데이터의 주제별 전략을 수립하고 자동화 방안 등을 수립합니다.

● 통합 영역의 적재 단계
임시 영역에 저장된 데이터는 빅데이터 시스템 및 목적과 특성에 따라 사용이 가능한 NoSQL을 선정합니다. 그리고 전처리를 거친 다음에 실제 빅데이터 영역에 저장합니다. NoSQL은 비관계형 데이터베이스로서 대용량 처리에 적합한 데이터베이스의 특징을 가지고 있습니다.

● 빅데이터 서비스의 위험 요인
빅데이터 서비스의 정책을 실행하기 위해서는 빅데이터의 관점에서 리스크 요인을 살펴보고 세밀한 검토를 거쳐 수립되어야 합니다. 단계적인 로드맵에 따라 빅데이터 서비스의 정책을 실행할 때는 분석적 측면, IT 인프라적 측면, 활용 측면 등도 함께 살펴보아야 합니다.
• 빅데이터를 서비스할 때의 위험 요인
-데이터의 보안: 데이터의 해킹, 비밀 유지, 유실 방지 등
-데이터의 품질: 대용량 데이터에 따른 품질 체계의 유지 등
-개인정보 보호: 개인정보에 대한 보호 및 유지 등
-데이터의 소유: 데이터의 소유에 대한 침해 방지 등

<07_빅데이터 핵심! 빅데이터 생산과 수집 기술>

● 데이터의 입력 방식
- 정형 데이터를 이용하는 방식: 조직 내에서 자체적으로 개발한 업무용 애플리케이션을 통해 생산합니다.
- 비정형 데이터를 이용하는 방식: SNS나 블로그 등과 같이 개인이 생산합니다.

● 데이터의 자동 생산 방식
사물인터넷 기술의 확산은 데이터 생산 방식의 다양화를 가져왔습니다. 그리고 이러한 배경에서 생산되는 데이터의 저장 및 관리 기술은 계속해서 중요해지고 있습니다.

● 데이터의 수집
데이터의 형태와 종류에 구애받지 않는 기술과 수집 방법이 필요합니다. 더불어 이와 같은 방식으로 수집한 데이터를 가공하고 재생산할 수 있어야 합니다. 수집 방법에는 데이터를 가공하고 재생산하는 과정이 포함됩니다. 서비스는 가시적인 형태로 제공할 수 있어야 하며, 효율적인 데이터 사용을 위한 저장 방법이 필요합니다. 또 정형 데이터와 비정형 데이터의 형태에 따라 데이터를 저장하는 기능이 제공되어야 합니다.

● ESB의 구현 방법
ESB는 플랫폼 내의 구성 요소를 사용하는 등 다양한 방법으로 구현될 수 있습니다. 구성 요소에는 기본적인 메시징, EAI, 중개(라우팅) 기술을 사용하거나 J2EE 시스템의 서비스 통합 버스 등이 포함됩니다. ESB는 구현 방식이 전반적인 아키텍처에 영향을 미치지 않는다는 전제 하에 EAI와 애플리케이션 서버 기술을 함께 조합하여 구현이 가능합니다.

● EAI(Enterprise Application Integration)
EAI는 기업 내·외부의 서로 다른 시스템을 통합하기 위해 사용하는 기법입니다. 에이전트를 이용하여 개별 애플리케이션을 중앙 허브와 연결하고 중앙 허브를 통해 상호 데이터를 수집합니다.

● EAI의 지원 기능
EAI 영역에서 처리하는 리소스 형태는 TCP 서비스, 웹 서비스, 데이터베이스, 파일, 애플리케이션 등 그 종류가 다양합니다. 인터페이스 아키텍처는 과거에 사용했던 피어 투 피어(Peer to Peer) 방식을 버리고, 허브 앤 스포크(Hub & Spoke) 방식, 버스 방식 등의 구성 방안의 다양성을 제공하여 고객맞춤형 연계 시스템을 구축합니다.

● API(Application Programming Interface) 플랫폼
데이터 소유 주체가 웹 개발자나 사용자를 위해 정보/데이터를 정해진 방식으로 공개하는 기술을 오픈 API라고 합니다. HTTP 프로토콜을 기반으로 하는 웹 서비스이며, SOAP(Simple Object Access Protocol)와 REST를 기반으로 하는 웹 서비스 기술을 제공합니다.

<08_빅데이터 핵심! 빅데이터 관리와 분석 기술>

● 데이터 거버넌스

데이터 거버넌스란, 전사적으로 보유하고 있는 데이터에 대한 관리 체계를 의미합니다. 데이터에 대한 관리나 정책, 지침, 표준, 전략 및 방향 수립을 포함하며, 데이터를 관리할 수 있는 조직이나 서비스의 정의도 포함하는 의미입니다.

조직 내에 거버넌스가 확립되지 못하면 품질이 낮은 데이터를 사용하게 되고 이에 따라 오류가 생성되거나 규제에 직면할 수 있습니다. 또 개인정보 관련 데이터가 유출되는 사태가 발생하여 한순간에 고객의 신뢰를 잃을 수도 있습니다.

데이터 거버넌스는 고품질의 데이터를 확보하고 적극적인 활용을 통해 조직의 가치 창출에 지속적으로 기여하는 것을 목표로 합니다. 데이터를 통해 위험을 예측하고, 관리 비용을 최적화하며, 데이터의 활용이 촉진됨으로써 데이터의 가치가 향상됩니다. 이는 곧 비즈니스 목적에 부합하는 서비스가 지속될 수 있는 힘을 가지게 합니다.

● 데이터 세트에 대한 수명 주기 및 자원 관리 요구 기능

- 비즈니스 사용자 관점의 메타데이터 관리
- 샌드박스에서 테스트한 뒤에 분석 결과를 공유하기 위해 운영 환경으로 옮기기 위한 검토 프로세스
- 조직의 역할과 같은 새로운 데이터 거버넌스의 절차와 기능

● 데이터 거버넌스의 중요성

데이터 레이크와 셀프 분석 환경의 확대는 더 많은 데이터 사용자를 만들어 냈지만 인사이트의 획득에는 아직 어려움이 있습니다. 데이터의 활용이 강조되고 이에 따라 데이터의 유형이 다양화되면 데이터 거버넌스의 중요성은 그에 비례하여 부각될 것입니다.

데이터는 실시간으로 분석될 수 있어야 하며, 이를 통해 예측할 수 있어야 합니다. 의사결정에 활용되는 데이터는 품질이 보장되어 있어야 하며, 프라이버시의 보호와 데이터의 수명 관리, 데이터의 소유와 관리 권한의 명확화 등이 함께 이루어져야 하기 때문입니다.

데이터의 활용을 촉진하고 동시에 위험을 감소하기 위한 사람과 조직 및 기술의 총합이 바로 데이터 거버넌스인 것입니다. 따라서 데이터 거버넌스 시스템의 도입은 한 번에 이루어질 수 없습니다. 관련자 모두의 의지와 노력을 통해 문화로서 정착할 수 있어야 합니다.

● 데이터 분석 기술

크게 전체 데이터 분석의 일부분인 단순 집계와, 전문적인 도구가 필요한 고급 분석으로 구분됩니다. 단순 집계는 중요성이 매우 높아 고급 분석에도 단순 집계가 전후 단계에서 함께 수행되면서 시너지 효과를 내고 있습니다. 많은 소프트웨어 도구들이 신축적인 단순 집계가 가능하도록 발전하고 있으며, 융합이 더욱 손쉽게 가능하도록 고급 분석을 지원하는 기능들이 점점 더 다양하게 제공되고 있습니다.

● 고급 분석

통계적 분석과 머신 러닝을 포함합니다. 전문적인 도구가 필요하고 보급 자체가 한정적이며 유용성과 활용 방법 부분에서도 한정적이어서 본격적인 활용이 지연되어 왔습니다. 빅데이터의 개념이 전파되면서 머신 러닝과 인공지능에 대한 일반인들의 이해가 높아졌고 오픈소스 소

프트웨어가 보편화되기 시작했습니다. 이러한 발달의 과정은 고급 분석이 도입되고 실무 활용이 본격화되는 기반이 되었습니다.

<09_빅데이터 분석 - 데이터 확보 기획>

● 분석 데이터의 확보
우선으로 고려해야 할 사항은 수집 대상 데이터의 유형입니다. 분석 요건의 정의를 통해 목표를 도출하고, 도출된 목표를 어떤 데이터를 가지고 수행할 것인지에 대한 분석 기법을 결정합니다. 이와 같이 수립한 계획에 따라 데이터 유형을 선택하고 분석 변수를 정의합니다.

● 분석 데이터의 유형
- 정형 데이터는 관계형 데이터베이스 시스템의 테이블과 같이 고정된 필드에 저장되어 활용되는 구조화된(Structured) 데이터입니다.
- 반정형 데이터는 데이터 내부에 정형 데이터의 스키마에 해당하는 메타데이터(Metadata)를 포함한 반구조적(semi-structured) 형태를 가지는 데이터입니다.
- 비정형 데이터는 고정된 필드가 아닌 구조화되지 않은(Unstructured) 데이터입니다.

● 빅데이터 분석 데이터에 대한 이해
효과적 빅데이터 분석을 위해서는 요구 정의에 의해 도출된 활용 시나리오에 적용할 수 있는 다양한 분석 데이터 세트를 수집하여 분석에 활용해야 합니다. 이러한 과정을 통해 의미 있는 분석 결과를 도출할 수 있도록 준비합니다.

● 분석 데이터 세트의 활용
분석 결과의 일반화에 따르는 오류를 예방하기 위해 교차검증(Cross Validation) 기법을 적용하여 테스트 세트(Test Set), 밸리데이션 세트(Validation Set), 트레이닝 세트(Training Set)를 혼합하여 빅데이터 분석 변수로 활용합니다.

● 분석 데이터를 확보할 때의 유의 사항
수집되는 많은 데이터에는 산업 기밀이나 개인정보 등 비밀이 보장되어야 하는 데이터가 다수 포함되어 있습니다. 이러한 데이터는 사전에 비식별 조치를 거쳐 정보의 유출을 방지할 수 있도록 계획해야 합니다.

● 빅데이터의 수집 기법
-크롤링: SNS, 웹, 뉴스 정보 등의 인터넷상에 제공하는 웹 문서 정보를 수집합니다.
-스크래핑: 인터넷 웹 사이트에 보이는 내용 중에서 특정 정보만을 추출하고 모든 동작을 자동으로 수행합니다.
-FTP: TCP/IP 프로토콜을 이용하여 인터넷 서버로부터 각종 파일을 송수신합니다.
-오픈 API: 서비스, 정보, 데이터 등의 개방된 정보로부터 API를 통해 데이터를 수집합니다.
-RSS: 웹상의 최신 정보를 공유하기 위한 XML 기반의 콘텐츠 배급 프로토콜입니다.
-스트리밍(Streaming): 인터넷에서 음성, 오디오, 비디오 등의 멀티미디어 데이터를 송수신하

는 기술입니다.
-로그 Aggregator: 웹 서버 로그, 웹 로그, 트랜잭션 로그, 데이터베이스 로그 등의 각종 서비스 로그를 수집하는 오픈소스 기술입니다.
-RDB Aggregator: 관계형 데이터베이스에서 정형 데이터를 수집하여 하둡 분산 파일 시스템(HDFS)이나 HBase 등의 NoSQL에 저장하는 오픈소스 기술입니다.

<10_빅데이터 분석 - 데이터 탐색 기획>

● 탐색적 데이터 분석(EDA: Exploratory Data Analysis)
데이터를 가공하지 않고 있는 그대로 보여 주는 것을 핵심으로 삼아 데이터를 분석하는 기법입니다. 주로 있는 그대로의 현실 상황을 나타내기 위해 빅데이터 분석에서 자주 사용하는 데이터 분석 기법입니다.

● 탐색적 데이터 분석의 필요성
① 잠재적 문제의 발견
- 데이터의 분포 및 값을 검토함으로써 데이터가 표현하는 현상을 이해합니다.
- 데이터의 잠재적 문제를 발견할 수 있습니다.
- 본격적 분석에 앞서 수집에 대한 의사결정을 지원합니다.

② 가설의 설정
- 다양한 각도의 분석 과정에서 문제의 정의 단계에서 미처 발견하지 못한 패턴의 발견이 가능합니다.
- 기존의 가설을 수정하거나 새로운 가설을 설정할 수 있습니다.

● 탐색적 데이터 분석의 절차
문제의 정의 단계에서 설정한 분석의 주제와 가설을 바탕으로 분석 계획을 세웁니다. 이후의 분석 계획에는 어떤 속성 및 속성 간의 관계를 집중적으로 관찰해야 하는지를, 이를 위한 최적의 방법은 무엇인지를 도출합니다.

● 머신 러닝 기반의 데이터 탐색 기법
클러스터링(Clustering), 클래시피케이션(Classification) 등을 이용해 분석 데이터를 확보합니다. 머신 러닝 학습의 알고리즘 지도나 비지도 또는 강화 학습을 통해 데이터를 탐색할 수 있습니다.

● 머신 러닝 학습의 탐색 유형에 의한 분류
• 지도 학습: 유형을 구분 짓는 속성을 가지는 주어진 데이터 집합으로부터 함수적 모델을 찾아 데이터를 분석하는 기술입니다.
• 비지도 학습: 유형을 구분 짓는 속성을 가지지 않는 데이터 집합으로부터 데이터 자체의 상호 유사성을 분석하는 기법입니다.
• 강화 학습: 데이터의 상태를 인식하고, 이에 반응한 행위에 대해 환경으로부터 받는 리워드를 학습하며 분석하는 기술입니다.

● 빅데이터 분석 기법을 위한 데이터 마이닝의 이해
-데이터 마이닝에 대한 개념: 대용량 데이터 간의 관계, 패턴, 추세를 발견하고, 이를 의미 있는 정보로 변환하여 기업의 의사결정에 활용하는 기술입니다.
-데이터 마이닝 프로세스: 데이터 분석에 필요한 변수를 식별하고 데이터 전처리 과정을 통해 다양한 모형을 도출하여 결과를 분류하고 예측하는 기법입니다.
-데이터 마이닝의 상세 수행 절차: 데이터를 선택하고 정제하여 데이터를 보완한 뒤에 변환과 모델링 과정을 거쳐 마이닝을 수행합니다.

● 데이터 분석 방법 설명
• 기술 통계학(Descriptive Statistics): 데이터가 가진 정보를 데이터 탐색만으로 얻는 방법입니다.
• 결론 유추 분석 방법: 기술 통계량(tools), 그래프, 데이터 분석 경험(know-how)을 이용합니다.
• 데이터 분석: 데이터로부터 정보를 도출하기 위한 다양한 방법을 시도합니다.
• 정보의 정확도: 탐색적 분석을 통해 얻은 정보를 이용해 통계적 가설 또는 모형을 선정하여 연구합니다.
• 페이퍼 펜슬 방법(Paper-Pencil Method): 수학적 그래프 또는 통계량을 직접 계산할 수 있는 방법입니다.
• 데이터 마이닝(Data Mining): 대용량 데이터의 패턴(pattern)이나 규칙(rules)을 발견하는 방법입니다.

● 빅데이터 분석 주요 기법의 상세 유형
- 텍스트 마이닝(Text Mining) :
텍스트 마이닝은 비정형 및 반정형 텍스트 데이터에서 자연어 처리 기술을 기반으로 유용한 정보를 추출하고 가공하는 것을 목적으로 하는 기술입니다.
- 소셜 분석(SNA: Social Network Analysis) :
소셜 네트워크 분석은 수학의 그래프 이론에 기반을 두고 있는 방법입니다. 소셜 네트워크의 연결 구조 및 연결 강도 등을 바탕으로 사용자의 명성이나 영향력을 측정합니다.
- 클러스터 분석(Cluster Analysis) :
군집 분석은 비슷한 특성을 가진 개체를 합하면서 최종적으로 유사한 특성의 군집을 발굴하는 데에 사용됩니다.
- 분류 분석(Classification) :
다수의 속성 또는 변수를 가지는 객체를 사전에 정해진 그룹 또는 범주(Class, Category) 중의 하나로 분류하는 분석 기법입니다.

<11_빅데이터 분석 - 모델링 기획과 결과 적용>

● 데이터 모델링의 방법
-개념적 데이터 모델링: 조직의 업무 요건의 충족을 위해 주제 영역과 핵심 데이터 집합 간의

관계를 정의하고 설계하는 모델링 기법입니다.
-논리적 데이터 모델링: 업무의 모습을 모델링 표기법으로 형상화하여 사람이 이해하기 쉽게 표현하는 절차입니다.
-물리적 데이터 모델링: 일괄 전환, 구조 조정, 성능 향상의 작업을 위해 논리 데이터 모델을 특정 DBMS의 특성에 맞게 성능을 고려하여 물리적 스키마를 만드는 일련의 과정입니다.

● 빅데이터 모델링
현실 상황의 과정과 결괏값이 완성된 단계 이후에 복잡한 데이터를 새롭게 정의합니다. 그리고 기업이나 공공기관에서 새로운 서비스나 신사업의 전략적 의사결정에 중요한 가치를 창출하기 위한 전략적 모델링 과정입니다. 빅데이터를 근거로 새로운 사업과 서비스의 가치 모형을 만들어 내는 의사결정 구조의 설계 과정입니다.

● 계층적 프로세스 모델
-최상위 계층(Phase): 각 단계별 산출물의 생성과 단계별 기준선을 설정하여 관리합니다. 버전 관리 등을 통해 통제합니다.
-태스크(Task): 각 단계를 구성하는 단위 활동입니다. 물리적 또는 논리적 단위로 품질을 검토합니다.
-마지막 계층(Step): WBS의 워크 패키지와 같은 계층입니다. 입력 자료, 처리 및 도구, 출력 자료로 구성된 단위 프로세스입니다.

● 빅데이터 시각화 분석 기법
① 확률 기반
-커널 밀도 추정: 모수적 밀도 추정, 비모수적 밀도 추정
-탐색적 자료 분석: 데이터양과 정밀도 및 연관 패턴의 분석
② 시간과 공간
-선적분 회선 분석: 텍스처 기반의 유동 시각화, 2D Flow 패턴의 시각화
-스펙트럼 큐브: 2차원 지형 공간의 이벤트 정보, 이벤트의 시간 흐름 분석

● 빅데이터 큐레이션 분야
빅데이터 큐레이션의 목적은 예측하고 요구 사항을 발견하며 고객맞춤형 서비스를 제공하여 비즈니스를 지원하고 발전시키는 것입니다. 이를 위해 인과관계의 분석, 데이터의 숨은 패턴을 발견하기 위한 마이닝과 데이터의 시각화, 맞춤형 서비스를 위한 우선순위 선별 작업 등을 수행합니다.

● 빅데이터의 주요 특징에는 대규모(Volume), 다양성(Variety), 속도(Velocity)가 있어 처리 속도에 대한 한계가 존재합니다. 이에 따라 기업 또는 현장에서 개선을 요구하는 목소리가 증가하게 되었으며, 이를 해결하기 위해 실시간으로 빅데이터의 수집과 처리가 가능한 아키텍처를 개발하게 되었습니다.

<12_그래프 데이터베이스와 빅데이터 시각화 기술>

● 시각화 분석의 접근 방식
시각화 분석의 기본 접근은 '재질의가 가능한가'에 달려 있습니다. 데이터의 시각화에서 질의의 형태는 시각적인 형상화와 대화형의 인터페이스여야 합니다. 또 시각화의 설계 요소는 하나의 분석에 다양한 관점을 가지게 하는 유연한 모습으로 제공되어야 합니다.

● 기존의 관계형 데이터베이스와의 차이점
기존의 관계형 데이터베이스는 제정이 매우 복잡하고, 저장하더라도 다수의 테이블 간의 조인 등에 의해 질의 처리가 비효율적이라는 단점이 있습니다. 그래프 형태의 데이터를 사용하는 이유는 효율적인 저장과 효율적인 질의의 지원이 가능하기 때문입니다.

● 그래프 데이터 구성 요소
그래프는 다양한 개체의 복잡한 관계를 표현하는 데 매우 유용하게 사용되는 데이터 모델입니다.
-노드: 데이터에 포함된 각각의 개체입니다.
-간선: 관계(Relationship)라고도 하며, 이들 개체 간의 관계를 나타냅니다.
-속성: 각 노드와 간선에는 다양한 속성과 함께할 수 있으며 해당 노드와 간선에 대한 자세한 정보를 표현하는 역할을 합니다.

● 그래프 데이터베이스의 특징
기존의 관계형 데이터베이스에 비해 속도가 빠르고, 데이터의 표현이 쉽습니다. 그래프 데이터베이스는 그래프 형태의 데이터를 저장과 질의 및 관리하는 데에 최적화되어 있습니다. 특히, 많은 개체 수의 다양한 관계 정보를 관리하는 데 적합합니다.

● Neo4j
고유의 그래프 모델을 제안하고 구현한 최초의 그래프 데이터베이스 중의 하나입니다. 현존하는 그래프 데이터베이스 중에서 가장 유명하며, 오픈소스로 제공되는 그래프 데이터베이스입니다.

● 그래프 데이터베이스의 대표적인 기능
① 그래프 데이터에 대한 CRUD(Create, Read, Update, Delete) 연산을 제공합니다.
② 대부분 그래프 연산에 대한 ACID(Atomicity, Consistency, Isolation, Durability) 트랜잭션을 제공합니다.
③ 노드 간의 연결 상태를 저장하고 있는 색인 또는 인접 정보를 유지하여 연결된 노드들을 빠르게 탐색할 수 있습니다.
④ 색인 또는 인접 정보를 사용하여 그래프에 대한 질의를 빠르게 처리합니다.

<13_빅데이터 아키텍처 기획하기>

● 아키텍처 요구 사항의 정의
- 기능 요구 사항: 정보 시스템의 구축을 통해 적용되는 업무 프로세스를 파악하여 현행 프로

세스의 문제점과 불편함을 개선합니다. 이후 효율적인 업무가 가능하도록 지원하기 위해 필요한 기능을 정의한 것입니다.
- 비기능 요구 사항: 기능 요구 사항을 제외한 요구 사항의 통칭입니다. 성능, 보안, 사용자 편의성 등 기능 요구 사항의 업무 처리를 위한 프로세스처럼 가시적으로 나타나지는 않지만, 구축 시스템의 품질과 밀접한 관련이 있습니다.

● 아키텍처의 정의
요구 사항을 구현하기 위한 기반 기술을 정의하는 과정입니다. 요구 사항을 반영하여 하드웨어 아키텍처와 소프트웨어 아키텍처를 정의합니다. 이는 정보 시스템의 개발, 테스트, 이관을 위한 기술적 기반이 됩니다.

● 인프라의 설계
아키텍처 요구 사항이 확정되면 해당 요구 사항을 반영한 인프라의 상세 설계를 수행합니다. 인프라를 설계할 때는 장비의 규격, 수량, 용량 등을 고려하여 구성합니다. 인프라를 설계할 때 하드웨어 아키텍처와 소프트웨어 아키텍처의 구성은 상호 제약 사항으로 작용할 수 있습니다. 그러므로 소프트웨어와 하드웨어 구성에 대한 지속적인 검토를 수행하고 점진적으로 설계에 반영할 수 있도록 합니다.

● 클라우드 컴퓨터 서비스의 유형
-SaaS(Software as a Service): 사용자가 소프트웨어 서비스를 이용한 시간 또는 제공된 용량에 따라 과금이 되는 클라우드 서비스입니다.
-PaaS(Platform as a Service): 응용 시스템을 개발 및 실행할 수 있도록 운영 환경을 제공하는 클라우드 서비스입니다.
-IaaS(Infrastructure as a Service): 네트워크 장비, 서버 등의 하드웨어 인프라를 가상화 시스템으로 제공하는 서비스입니다.

● 스케일 업이란, 제품 내에서 프로세서, 메모리, I/O Device 등 컴퓨팅 리소스의 동적인 확장을 통한 성능 향상을 뜻합니다. 스케일 아웃이란, 저렴한 노드를 병렬로 증설하여 유연한 확장성과 성능을 확보하는 것을 뜻합니다.

● ‘스케일 업(Scale-up)’과 ‘스케일 아웃(Scale-out)’ 요건의 검토
시스템의 확장성을 고려하여 적정 용량 및 증설 전략을 검토합니다. 빅데이터 플랫폼의 경우에는 다수의 노드를 통한 분산 처리 시스템으로 구성하는 것이 일반적입니다. 개별 노드에 대해서는 스케일 업(Scale-up) 방안을, 전체 시스템에 대해서는 스케일 아웃(Scale-out) 방안을 검토합니다.

<14_빅데이터 수집 기술 기획하기>

● 크롤링(Crawling)
웹(WWW) URL에 존재하는 HTML 문서에 접근하여 해당 내용을 추출합니다. 문서에 포함된

하이퍼링크(hyperlink)를 통해 재귀적으로 다른 문서에 접근하여 콘텐츠 수집을 반복하는 기술입니다.

● EAI
기업에서 운영하는 이기종 간의 애플리케이션은 네트워크 프로토콜, DB, 운영 체제와 같은 소프트웨어가 있습니다.

● ESB(Enterprise Service Bus)
버스(분산 구조) 형태의 느슨한 결합을 이용한 서비스 중심의 시스템 연계 방식입니다.

● CEP(Complex Event Processing)
복합 이벤트 프로세싱은 실시간으로 생성되는 다수의 원천 데이터로부터 발생한 이벤트를 대상으로 의미 있는 데이터를 실시간으로 추출하고 처리하는 기술입니다.

● RDBMS 테이블
관계형 데이터베이스의 테이블은 행(ROW)과 열(COLUMN)로 구성된 대표적인 정형 데이터입니다. 추출 및 가공이 쉬우며, 표준을 따르고 있는 DBMS의 경우에는 조인이나 합병 등의 데이터 병합을 통한 결과를 출력할 수 있습니다.

● XML(Extensible Markup Language)
XML은 W3C 표준의 다목적 마크업 언어입니다. SGML의 일종으로 데이터를 기술하는 데 사용할 수 있습니다. XML은 이기종 시스템 간의 데이터를 손쉽게 교환하기 위해 고안되었습니다. 또 데이터의 구조와 값을 함께 기술합니다. XML을 기반으로 한 데이터 표준은 RDF, RSS, 아톰(Atom), XHTML, SVG 등이 있습니다.

● 수집 대상 데이터 유형의 식별
데이터 수집 대상의 시스템이 식별되었으면 수집하고자 하는 데이터의 유형을 확인해야 합니다. 엑셀(XLS) 파일과 같은 정형 데이터, 멀티미디어 콘텐츠와 같은 비정형 데이터, XML이나 제이슨(JSON) 형태의 반정형 데이터 등 데이터 유형에 따라 데이터 수집 방식이 달라질 수 있습니다.

● 제이슨(JSON: JavaScript Object Notation)
제이슨은 키-값(Key-Value)으로 이루어진 데이터 오브젝트를 교환하기 위한 개방형 표준 포맷입니다. 브라우저 비동기 통신(AJAX)에 주로 사용되며, XML을 대체하는 주요 데이터 포맷입니다. 데이터의 유형에 큰 제한이 없으며, 경량의 통신에 적합합니다.
XML과 달리 데이터에 대한 구조 및 형식(Schema)을 따로 기술하지 않습니다. 그리고 제이슨 데이터(Value) 내부에 배열 형태로 데이터를 저장할 수 있으므로 계층적 데이터의 표현이 가능합니다.
XML은 규격화된 데이터의 구조와 포맷을 정의하고 데이터를 전송하는 데 주로 사용합니다. 반면에 제이슨은 키-값 형태의 단순 데이터 전송에 적합합니다.

<15_빅데이터 저장 기술 기획하기>

● 하둡의 분산 노드
HDFS는 네임 노드(Name Node)와 데이터 노드(Data Node)로 구성되어 있습니다. 네임 노드는 마스터 역할을 수행하며, 데이터 노드를 관리합니다. HDFS에 대한 메타데이터, 블록, 헬스 체크 기능 등을 관리합니다. 데이터 노드는 실제로 정보가 저장되는 노드이며, 각 블록은 3개의 노드에 중복 저장됩니다.

● 맵리듀스의 절차
- 스플리트(Split): 입력 데이터는 스플리트 단위로 쪼개져 각 스플리트마다 하나의 맵 태스크가 할당됩니다.
- 맵퍼(Maper): 맵 태스크(Map Task)는 각각의 스플리트 데이터를 레코드 (Record) 단위로 읽어서 처리하고, 그 결과를 Key와 Value의 쌍으로 만들어 로컬에 저장합니다.
- 셔플 앤 소트(Shuffle & Sort): 리듀스 태스크에서 개별적으로 처리하는 파티션(Partition) 데이터를 로드하여 병합(Shuffle) 정렬(Sort)을 수행합니다.
- 리듀서(Reducer): Key, Value List 정렬이 완료된 다음에 쌍으로 묶어 리듀스 태스크에서 처리합니다.
- 결과 저장: 리듀스 작업까지 완료된 최종 결과는 사전에 설정된 결과 디렉터리에 적재됩니다.

● NoSQL
데이터베이스는 일반적으로 RDBMS에서 사용하는 표준 SQL을 질의어로 사용하지 않는 DBMS로 고정된 테이블 구조를 사용하지 않습니다. 또 트랜잭션 관리도 엄격하게 적용하지 않습니다. RDBMS와 비교해서 데이터의 일관성보다는 성능이나 확장성에 중점을 두고 읽고 쓰는 작업에 최적화되어 있어 빅데이터 저장에 적합합니다.

● CAP(Consistency, Availability, Partition tolerance) 이론
데이터 저장 시스템이 가지는 세 가지 특성인 일관성(Consistency), 가용성(Availability), 분할 내성(Partition tolerance)에 대한 이론입니다. CAP 이론에서 위 세 가지 특성은 상호 트레이드오프(trade off) 관계가 있으므로 최대 두 가지만 충족할 수 있고, 세 가지 특성을 모두 충족하는 것은 불가능합니다.

● 샤딩(Sharding)
샤딩은 테라바이트(TB) 단위의 대량 데이터를 처리하기 위해 데이터를 파티셔닝하는 일종의 수평 분할 기술입니다. 샤딩은 데이터를 여러 서버에 분할하므로 데이터베이스 서버의 스케일업을 하지 않고도 더 많은 데이터를 저장하고 처리할 수 있습니다.

<16_빅데이터 처리 기술 기획하기>

● 데이터의 시각화

분석된 데이터를 시각화하여 표현하는 것이며, 시각화된 데이터를 통해 분석자가 데이터의 패턴(pattern), 미분(differentiation), 이상점(outlier)을 발견할 수 있습니다. 분석자는 데이터에 내재되어 있는 인사이트를 얻기 위한 목적으로 이 방법을 사용합니다.

● ETL(Extract, Transform, Load)
데이터 원천으로부터 필요한 데이터를 추출하여 목표 시스템의 데이터 마트나 데이터 웨어하우스에 적재하는 기능을 제공하는 기법입니다. 데이터에 대한 품질과 신뢰성 보장을 위한 솔루션입니다. 데이터의 추출, 재구성, 정제, 통합, 변형 등을 포함하고 있으며, 시스템에서 필요한 데이터가 개별 시스템에 상이한 형태로 분산되어 있는 문제를 해결합니다.

● 데이터 분석 기법
데이터베이스로부터 아직 알려지지 않은 유용하고 업무 요구 사항에 적합한 정보를 도출하기 위한 지식의 발견 과정을 의미합니다. 주어진 데이터를 설명하는 패턴을 탐색하거나 주어진 데이터에 근거한 모델을 만들고, 이를 이용하여 새로운 데이터에 대한 예측을 목적으로 데이터 분석을 수행합니다.

● 실시간 분석 방안의 정의
데이터를 보관하지 않은 상태에서 분석이 필요한 업무이거나 이동 중인 데이터에서 가치 있는 인사이트를 찾기 위한 업무에 사용되는 분석입니다. 지속적으로 입력되는 대량의 데이터나 순서에 의미가 있는 데이터를 분석합니다.

● 사용자 대시 보드의 구성
사용자 대시 보드의 구성 데이터 처리 결과를 사용자에게 전달하기 위한 수단으로 대시 보드를 구현할 수 있습니다. 대시 보드의 구축 자체로 끝이 아니라 대시 보드는 단지 사용자에게 겉으로 보이는 모습만을 나타냅니다. 그러므로 내부적으로 수집 데이터를 기준에 따라 체계적으로 분류하고 이상 없이 활용할 수 있도록 데이터의 시각화나 사용자 접근성까지 고려되어야 합니다. 대시 보드의 구성 요소에는 정보 리포팅, 사용자 인터페이스, 페이지 레이아웃이 있습니다.
① 리포팅 기능을 정의합니다.
② 사용자 인터페이스를 설계합니다.
③ 편의성을 고려한 페이지 레이아웃을 정의합니다.

● 대표적인 데이터 분석 기법
-예측: 데이터 분석을 통해 아직 발생하지 않은 사건에 대한 예측을 수행하는 기법입니다.
-분류: 사전에 알려진 기준에 따라 분류된 데이터군을 학습시켜 새로운 데이터가 추가될 경우 적합한 데이터군을 찾는 기법입니다.
-클러스터링(Clustering): 유사한 특징이 있는 데이터들을 차례로 합병하면서 유사 특성군으로 분류하는 방법입니다.
-데이터 마이닝(Data Mining): 데이터베이스의 대용량 데이터로부터 패턴 인식, 통계 분석 등을 이용하여 데이터 간의 연관성 분석 및 유용한 정보를 추출하는 기법입니다.
-텍스트 마이닝(Text Mining): 텍스트 기반의 데이터로부터 새로운 정보를 발견할 수 있도록

정보의 검색, 추출, 체계화, 분석을 모두 아우르는 텍스트 프로세싱(Text-Processing) 기술 및 처리 과정입니다.

<17_산업에서 활용되는 빅데이터 비즈니스 - 1>

● 금융 빅데이터 개방 시스템(CreDB)
한국신용정보원은 '금융 빅데이터 개방 시스템(CreDB)'을 통해 일반신용정보, 기술신용정보, 보험신용정보를 세 차례에 걸쳐 차례대로 공개했습니다. 이는 중소형 금융사나 핀테크사, 연구 기관에 상품의 개발, 시장 분석, 연구 등을 수행할 수 있는 데이터의 기반을 마련합니다.

● 신용정보관리업(MyData)
-개인의 신용 정보 관리와 소비 패턴을 분석한 신용 관리, 자산 관리 서비스를 제공합니다.
-신용 정보 조회를 무료로, 무제한 조회할 수 있도록 제공합니다.
-여러 금융사에 흩어져 있는 자산 정보를 한눈에 볼 수 있는 서비스를 제공합니다.

● 오픈뱅킹
은행이 보유하고 있는 고객의 금융 정보에 다른 서비스 제공자가 오픈 API 등의 방식으로 안전하게 접근하도록 허용하는 정부의 정책 또는 은행의 자발적 행동을 뜻합니다. 이때 고객의 명시적 동의를 전제로 하며, 다른 서비스 제공자는 은행이 될 수도 있고 제3의 서비스 제공자가 될 수도 있습니다.

● 규제 샌드박스(Sandbox)
놀이터에서 완충 역할을 하는 모래사장처럼 새로운 서비스를 출시할 때 일정 기간 기존의 규제를 면제하거나 유예해 주는 제도를 말합니다. 금융, ICT, 산업, 지역 등의 4대 분야에서 시행하고 있습니다. 금융 분야에서 혁신금융 서비스라고 지정받으면 최대 4년 동안 관련 금융법상의 규제 적용에 대한 유예를 받을 수 있습니다.

● 클라우드 컴퓨팅
클라우드 컴퓨팅은 서버 등의 IT 인프라, 애플리케이션, 개발 도구를 인터넷 상의 서버를 이용해 필요한 만큼 IT 자원에 대해 비용을 지불하고 이용하는 서비스입니다. 기업 소유의 빅데이터가 많아지면서 이를 저장하고 처리한 후 분석하는 데 소요되는 비용이 증가하여 클라우드로의 전환이 대안으로 떠오르고 있습니다.

● 클라우드 컴퓨팅의 특징
-IT 기반 시설의 구축과 유지비용의 감소, 그리고 신기술을 개발하는 데 필요한 IT 자원을 유동적으로 이용할 수 있습니다.
-최근 전자금융감독 규정의 개정에 따라 클라우드 도입이 가능한 정보가 비중요 정보에서 개인신용 정보, 고유식별 정보를 포함한 중요 정보로 확대되었습니다.
-중요 정보를 클라우드에서 다루는 경우는 금융감독원에 의무적으로 보고해야 합니다.
-클라우드 제공 기업은 클라우드를 이용할 때 발생할 수 있는 장애나 재해가 발생했을 때 복

구할 수 있도록 서버 분산과 백업 등으로 이에 대비해야 합니다.

<18_산업에서 활용되는 빅데이터 비즈니스 - 2>

● 상품의 비즈니스 효율성 평가
-개발된 서비스가 필요한 사람이 누구인지를 생각합니다.
-생명에 직결되는 질병인지, 건강증진 등의 웰니스(Wellness)인지를 구별합니다.
-병원 안에서 사용하는 것인지 병원 바깥에서 사용하는 것인지를 구분합니다.

● 임상 연구 및 진료 데이터
임상 데이터는 의료기관 관련자가 사람을 대상으로 하는 진료 과정 중에서 생성하는 데이터를 의미합니다. 진단 기술이 발전함에 따라 직접 입력하지 않고 의료 장비로부터 생산되어 저장되는 데이터가 늘어나고 있습니다. 이러한 데이터는 전자문서 형태인 전자의무 기록과 처방전달 시스템에 기록됩니다. 또 엑스레이나 MRI와 같은 영상 데이터는 영상저장전송 시스템(PACS)에 저장하여 관리합니다.

● 오믹스(Omics)데이터
인체로부터 수집된 인체 구성물의 DNA와 RNA 시퀀스 정보 및 미세결실검사(microarray) 분석 정보를 뜻합니다. 최근에는 국내 병원에서도 암 유전체를 분석한 맞춤형 의료 서비스가 시작되었으며, 유전체 분석 전문 기업들로 인해 사업화가 빠르게 진행되고 있습니다.

● 개인건강 데이터 및 라이프로그
병원에서 진료가 이루어지는 동안에 지속적으로 수집하고 제공되는 개인의 일상 헬스케어 데이터를 개인건강 데이터(PHR: Personal Health Record)라고 합니다. 이는 개인의 기록, 웨어러블 기기, 스마트폰 등을 통해 수집이 가능합니다.

● 라이프로그 및 개인이 생산하는 건강 데이터 관련 비즈니스
웨어러블(Wearable) 기기로 측정되는 신호를 라이프로그(Lifelog)라고 합니다. 이를 확장하는 용어로 의료 사물인터넷(IoMT: Internet of Medical Things)을 사용하기도 합니다. 하지만 기기가 측정할 수 있는 변수가 많지 않고 질병의 예측과 치료 효과에 대한 증명이 부족합니다. 이러한 단점을 극복하기 위해 질병의 치료와 웰니스에 모두 적용이 가능한 구체적인 수익 모델이 개발되어야 합니다.
반면에 24시간 내내 측정이 가능하다는 장점이 있습니다. 아직은 기업별로 수집하는 데이터가 상이하고 표준안에 대한 논의가 부족하여 데이터의 수집과 정제를 따로 해야 한다는 과제가 남아 있습니다.

● 유전체 데이터 관련 비즈니스
병원에서 유전자 검사가 필요하다고 판단되는 경우 유전체기업협의회에 검사를 의뢰할 수 있습니다. 기업 검사실에서 분석된 결과에 따라 의사가 결과를 해석하여 치료에 반영하게 됩니다. 소비자 직접 의뢰 방식(DTC: Direct to Consumer)을 통해 소비자가 비용을 지불하고 유전자 검사를 받을 수도 있습니다.

● 오믹스(Omics)데이터
인체로부터 수집된 인체 구성물의 DNA와 RNA 시퀀스 정보 및 미세결실검사(microarray) 분석 정보를 뜻합니다. 최근에는 국내 병원에서도 암 유전체를 분석한 맞춤형 의료 서비스가 시작되었으며, 유전체 분석 전문 기업들로 인해 사업화가 빠르게 진행되고 있습니다.

<19_산업에서 활용되는 빅데이터 비즈니스 - 3>

● 스마트 팩토리(Smart Factory)
4차 산업혁명의 기반 기술 중에서 빅데이터, 클라우드, 인공지능, 산업인터넷과 같은 최신 기술을 활용해 제조 기업의 고도화를 이끄는 것이 '스마트 팩토리'입니다.

● 스마트 팩토리의 활용
데이터 마이닝, 머신 러닝, BI(Business Intelligence) 등의 데이터 분석 기술이 고도화됨에 따라 데이터 분석에 대한 관심이 커지고 있습니다. 기업이나 공장에서는 축적된 데이터를 활용하여 변화의 기회를 만들어 내거나 위험 요소를 줄이며 자산의 효율성을 높입니다.

● 클라우드(Cloud)를 활용하는 스마트 팩토리
제조 공장이 인공지능을 도입하여 성과를 내기 위해서는 ICT 분야의 전문가들과 협업하는 것이 중요한데, 대부분의 제조 공정은 전문 인력이 없거나 부족한 상태입니다. 이러한 문제를 해결하기 위해 클라우드 스마트 팩토리의 구현을 선택할 수 있습니다. 클라우드를 통해 서버의 인프라, 플랫폼, 분석 인프라를 제공받을 수 있어 스마트 팩토리 도입의 진입 장벽을 낮출 수 있습니다.

● 스마트 팩토리의 추진 이슈
-커넥티드 팩토리(Connected Factory)의 구축이 요구됩니다.
-스마트 팩토리의 도입을 위한 이니셔티브를 확립해야 합니다.
-통합 미들웨어(Middleware)와 표준화를 위한 협의가 필요합니다.

● 인공지능과 접목한 제조업
제조 데이터는 직관적인 해석이 어려워 기술의 도입에 제약이 많습니다. 제조업 특성상 수율과 안전이 중요하여 인공지능 기술의 운영이 쉽지 않습니다. 하지만 인공지능 기술을 도입하는 제조 기업들이 늘어나면서 다양한 적용 분야가 생겼고 급속한 기술 발달로 시장 규모도 커지고 있습니다.

<20_산업에서 활용되는 빅데이터 비즈니스 - 4>

● 말단 배송 서비스의 중요성
말단 배송은 택배 서비스의 전체 소요 시간과 비용의 측면에서 가장 높은 비중을 차지하고 있

습니다. 또 말단 배송은 고객과의 최접점에 위치하고 있기 때문에 서비스의 만족도를 결정하는 중요한 요인입니다. 말단 배송에 대한 중요성을 인식하면서 편의점의 픽업이나 인홈(In Home) 배송과 같은 다양한 방법을 통해 비용은 절감하고 서비스의 수준은 향상되고 있습니다.

● 유통·물류 분야 데이터 활용 비즈니스 현황
4차 산업혁명과 코로나19의 확산은 사회 전반에 '언택트(Untact)' 문화를 가져왔고 이에 따라 배송해야 할 물건의 양이 폭발적으로 증가했습니다. 반면에 배송에 요구되는 시간은 더 줄어들었습니다. 유통·물류 분야의 빅데이터 적용은 이러한 문제에 대응할 수 있도록 도와줍니다. 과거에 서비스를 제공하는 과정에서 축적해 놓은 엄청난 양의 데이터가 효율적이며 정확한 물류 서비스의 운영을 가능하게 합니다. 유통·물류 분야에서 물류 서비스의 혁신은 데이터 기반의 의사결정이 이루어진다는 것입니다. 온라인 소비 트렌드의 분석, 소비자 니즈의 파악, 가격 전략의 수립 등의 많은 서비스에 적용이 가능하며 이를 토대로 다양한 상품을 빠른 시간 내에 배송할 수 있습니다.

● 데이터 기반 풀필먼트 센터(Fulfillment Center) 운영 시스템
풀필먼트 센터는 고객의 요구 사항을 가장 가까운 곳에서 만족시킨다는 의미를 가지고 있습니다. 쿠팡과 마켓컬리의 물류센터는 자동화 시스템의 효율성보다는 빅데이터를 기반으로 하는 정확한 수요 예측에 집중합니다. 또 이를 통해 상품을 전달하는 배송 과정의 효율성을 높이는 것을 목표로 합니다.

● 데이터 기반의 유통 및 물류 산업 혁신의 가속화
유통·물류 산업의 한계를 극복하는 방법은 정보의 디지털 전환으로부터 구체화 되었습니다. 코로나19의 확산으로 디지털 전환이 가속화됨에 따라 많은 사람들이 빠른 속도로 온라인 서비스를 이용하게 되었습니다. 이러한 사회 변화로 인해 더 많은 데이터가 확보될 수 있는 배경이 되었습니다.

● 디지털운행기록계(DTG: Digital Tacho Graph)
운송 차량의 운행 기록을 데이터화하여 저장합니다. 자동차 운행과 관련된 정보를 기록하는 장치이며, 이러한 정보를 실시간으로 저장하여 업무 효율을 높이는 데에 활용합니다. 교통안전공단에서는 DTG의 데이터를 분석하여 운전자의 운전 습관을 파악하고 실증적인 운전 관리를 수행합니다.

● 랜덤스토(Random Stow) 방식
재고 관리에 빅데이터를 활용하여 같은 제품을 한 곳에 몰아 놓지 않고 소량을 진열대 곳곳에 배치하는 방식을 말합니다. 데이터 분석을 통해 제품의 보관 위치를 정하고 주문이 들어오면 출고 담당 직원에게 가장 가까운 진열대를 알려주어 동선을 최소화합니다.

● 큐레이션 커머스
큐레이터가 작품 등을 전시하고 기획하는 것처럼 전문가들이 독창적이거나 품질이 좋은 제품을 판매하는 것을 큐레이션 커머스(Curation Commerce)라고 합니다. 실시간으로 데이터의

수집이 가능하다는 것은 단 한 명의 고객을 위한 개인화 서비스의 등장을 암시합니다. 기존의 단순한 상품 추천 서비스를 넘어 패키지 형태의 상품 구매와 연관된 서비스들이 결합되어 하나의 상품으로 제공되는 것입니다. 이런 큐레이션 커머스(Curation Commerce)는 빠른 시일 내에 등장할 것으로 전망됩니다.