03_디지털 융합의 중심! 빅데이터의 이해

#1
1. 빅데이터(Big Data)
가. 등장 배경
① 패러다임의 변화
1990년 이후 인터넷이 확산되면서 정형화된 형태의 데이터와 비정형화된 형태의 데이터가 무수히 생성되어 '정보 홍수'의 개념이 등장합니다. 이것이 오늘날 빅데이터 개념으로 이어지게 되었습니다. 이후 데이터양이 엄청나게 증가하여 기존 데이터의 저장, 관리, 분석 기법만으로는 쏟아지는 데이터를 처리하는데 한계가 있음이 드러납니다. 그로 인해 정보기술의 패러다임도 바뀌게 됩니다. 이때 '빅데이터'라는 용어가 등장합니다.

② 빅데이터 시대의 시작
패러다임이 지능화되고 개인화된 시대를 '빅데이터 시대'라고 합니다. 빅데이터의 개념이 등장하며 자연스럽게 데이터에 관심이 증가하게 되고, 정보통신기술의 발달로 데이터의 규모와 유형 및 특성도 따라서 변화하게 됩니다.

#2
나. 개인화 서비스의 측면의 빅데이터
① 개인화 서비스의 시초
개인화 서비스란, 고객의 성향이나 수입, 규모, 소비의 형태 등을 바탕으로 하는 서비스를 말합니다. 신상품이 들어오면 고객의 취향에 맞추어 해당 상품의 정보를 팸플릿이나 소책자 또는 휴대폰 문자 메시지로 고객에게 제공하는 것이 초기 형태의 빅데이터 서비스입니다.

② 빅데이터 활용의 변화
이후의 빅데이터에는 스마트 기기 사용자가 본 영화, 들은 음악, 찍은 사진, 촬영한 동영상, 쇼핑한 물건, 저녁을 먹은 레스토랑 등 노출되는 모든 활동이 포함됩니다. 이러한 비정형 데이터를 분석하기 시작하며 개개인의 생각과 행동, 경향과 패턴을 파악할 수 있게 되었습니다. 또 패턴 분석을 통해 대중의 변화를 예측하고 개인에게 최적화된 맞춤형 서비스까지 가능해졌습니다.

#3

다. 차세대 이슈로서의 빅데이터
① 정보통신 기술의 주도권이 데이터로 이동
모바일, 클라우드, 소셜 네트워크 서비스 등이 등장하면서 정보통신 기술의 주도권이 인프라와 기술 등에서 데이터로 이전되고 있습니다. 이에 데이터의 폭발적인 증가에 대응하고 데이터를 분석하는 방법이 정보통신 기술의 가장 중요한 이슈로 부각되었습니다.

#4
② 공간, 시간, 관계, 세상 등을 담은 빅데이터
스마트 기기의 확산으로 사용자가 자발적으로 참여하고 정보를 생성하는 '소셜 데이터 혁명'이 일어납니다. '소셜 데이터 혁명'은 정보의 생성자, 규모, 파급 효과 등에서 1990년대 기업이 고객의 정보를 축적했던 '정보 혁명'과는 구분됩니다.
- 소셜 네트워크 서비스의 이용 확산과 소통 방식의 변화는 데이터의 변혁을 가져오는 가장 중요한 요인이 되었습니다.
- 소셜 네트워크 서비스로 제공되는 정보는 지식 정보와 함께 정서적인 공감에 바탕을 둔 감성적 정보가 큰 비중을 차지합니다.
- 소셜 네트워크 서비스에서는 개인의 취향이 더욱 직접적으로 반영되기 때문에 진실성, 진정성, 관련성이 증가되어 데이터로서의 가치가 매우 높게 평가됩니다.

#5
③ 미래의 경쟁력과 가치 창출의 원천
빅데이터에는 잠재적 가치와 위험이 공존하는데, 사회적 또는 경제적으로 성패를 좌우하는 핵심 원천이 될 것으로 평가됩니다. 이에 세계 각국의 정부와 기업은 빅데이터가 향후 기업 경영의 성패를 가늠할 새로운 경제적 가치가 될 것이라고 예상합니다.
데이터가 폭발적으로 증가하면서 빅데이터가 등장했지만, 방대한 양의 데이터 중에서 의미 있는 데이터는 소수에 불과합니다. 따라서 의미 있는 데이터를 찾아내려면 빅데이터를 효과적으로 처리할 수 있는 기술이 필요합니다.

#6
2. 빅데이터의 개념과 속성
가트너(Gartner)는 현재 가장 널리 사용하는 빅데이터의 속성을 3V, 즉 양

(Volume), 다양성(Variety), 속도(Velocity) 등의 세 가지로 정의했습니다. 여기에 IBM은 정확성(Veracity)을 더했고, 오라클(Oracle)은 가치(Value)를 추가하였습니다.

#7
가. 5V
① 양(Volume)
양(Volume)은 '규모' 또는 '용량'을 뜻합니다. 미디어나 위치 정보, 동영상 등과 같이 다루어야 할 데이터의 크기를 말하는 것입니다. 물리적인 크기뿐만 아니라 현재의 기술로 처리가 가능한 양인지, 불가능한 양인지에 따라 빅데이터를 판단합니다.

② 다양성(Variety)
다양성이란, 다양한 종류의 데이터를 수용하는 속성을 말합니다. 빅데이터에는 형식이 정해져 있는 정형 데이터뿐만 아니라 형식이 정해지지 않은 다양한 비정형 데이터도 있습니다. 감시 카메라에서 생성되는 동영상, 개인이 디지털카메라로 생성하여 웹 사이트에 올리는 사진, 소셜 네트워크 서비스로 전달되는 메시지, 물건에 부착되거나 주변에 설치된 센서에서 발생하는 RFID 태그나 센서값 등이 비정형 데이터에 해당합니다.

③ 속도(Velocity)
속도란, 대용량의 데이터를 빠르게 처리하고 분석할 수 있는 속성을 말합니다. 데이터를 자동으로 생성하는 센서나 스마트폰 등의 데이터 생성 및 유통 채널이 다변화하면서 데이터의 생성 속도가 빨라집니다. 이는 처리 속도의 가속화를 요구합니다.

#8
④ 정확성(Veracity)
정확성이란, 데이터에 부여할 수 있는 신뢰 수준을 말합니다. 높은 데이터 품질을 유지하는 것은 빅데이터의 중요한 요구 사항이자 어려운 과제입니다. 하지만 최상의 데이터 정제(Data Cleansing) 기법을 사용하더라도 날씨나 경제, 고객의 미래 구매 결정과 같은 일부 데이터의 본질적인 불확실성은 제거할 수 없습니다. 소셜 네트워크와 같은 인간 환경에서 생산되는 데이터는 신뢰하기가 어렵고, 미래는 예측하기가 쉽지 않습니다. 사람과 자연, 보이지 않는 시장의

힘 등이 빅데이터의 다양한 불확실성의 형태로 나타납니다.

⑤ 가치(Value)
가치란, 빅데이터를 저장하려고 IT 인프라 구조 시스템을 구현하는 비용을 말합니다. 빅데이터의 규모는 엄청나게 크며 대부분은 비정형적인 텍스트와 이미지 등으로 구성되어 있습니다. 이 데이터들은 시간이 지남에 따라 빠르게 전파되면서 변화하므로 그 전체를 파악하고 일정한 패턴을 발견하기가 쉽지 않아 가치의 중요성이 강조됩니다.

#9
3. 빅데이터의 종류
가. 정형 데이터
고정된 필드에 저장된 데이터를 뜻합니다. 관계형 데이터베이스나 스프레드시트가 여기에 해당합니다. 정형 데이터는 일정한 규칙에 따라 체계적으로 정리한 데이터입니다. 이러한 데이터는 정형화되어 있어 그 자체로도 의미 해석이 가능하며 바로 활용이 가능합니다.

나. 반정형 데이터
고정된 필드에 저장되어 있지는 않지만 메타데이터나 스키마 등을 포함하는 데이터입니다. XML, HTML, 텍스트가 여기에 해당합니다. 페이스북, 트위터, 카카오톡 등의 소셜 네트워크 서비스 사용자가 생성하는 데이터들이 여기에 해당합니다.

다. 비정형 데이터
고정된 필드에 저장되어 있지 않은 데이터를 뜻합니다. 텍스트 분석이 가능한 텍스트 문서, 이미지나 동영상, 음성 데이터가 여기에 해당합니다. 비정형 데이터의 증가 속도는 누구도 예측할 수 없을 정도입니다. 비교적 선형적으로 증가하던 정형 데이터조차 연간 40~60%에 이르는 증가세를 보이기 때문입니다. 스마트 기기로 생성하는 소셜 데이터 외에도 이메일, 동영상 등의 비정형 데이터가 향후 10년 동안 생성하는 양은 전체 데이터의 90%에 달할 것으로 전망됩니다.

#10
질문자: 빅데이터의 처리를 위해서는 무엇을 갖추어야 하나요?

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

#11
4. 기술의 정의 및 범위
혁신성장 동력 시행 계획에 제시된 빅데이터 기술의 범위는 데이터의 수집·저장·처리 등의 플랫폼 기술, 이와 연계된 분석 기술, 새로운 통찰력과 비즈니스 가치를 창출하는 활용 기술을 포괄하고 있습니다.

#12
가. 플랫폼
① 데이터의 자가 증식과 수집 및 정제 기술: 데이터의 양적인 확대를 위해 알고리즘을 활용하여 자가 증식을 하거나, 유효하지 않은 데이터를 필터링하거나 샘플링하고, 수집 및 정제하는 기술입니다.

② 다양한 응용 패턴의 통합 지원 기술: 데이터가 실제 사용되는 시점에 데이터의 사용 목적에 따른 데이터 모델에 맞추어 실시간으로 데이터를 구성하여 제공합니다. 또 다양한 응용 패턴(배치, 대화형, 스트림 등)을 통합하여 동시 수행을 지원하는 멀티타입 빅데이터 처리 프레임워크를 의미합니다.

③ 멀티모델 데이터의 통합, 고신뢰 데이터 관리 및 다각도 분석 기술: 분석 목적에 맞게 다양한 모델의 데이터를 통합하여 데이터의 신뢰성을 확보합니다. 통계적으로 중요도를 가지는 결과를 자동 탐색하며 실시간으로 다각도로 분석을 하는 기술입니다.

④ 초연결 데이터 관리 및 협업 기술: 초연결 인공지능 구현을 위해 물리적인 데이터의 위치나 종류와는 무관하게 데이터를 제공할 수 있는 초연결 데이터를 관리하는 기술입니다. 그리고 최적의 분석 결과를 도출하기 위한 다수의 다양한 지능을 가지는 객체 간의 집단 협업지능 플랫폼 기술입니다.

#13
나. 분석

① 지능형 예측 분석 기술: 데이터에 숨겨진 패턴을 찾아 과거와 현재의 상황에 대한 이해를 바탕으로 미래 상황을 예측함으로써 선제적인 의사결정을 지원합니다.

② 이종소스 심층 융합 분석 기술: 비정형 텍스트, 관계형 DB 저장 데이터와 더불어 이미지나 비디오 및 IoT 스트림 데이터 등의 복합형 데이터를 대상으로 통합하고 분석합니다.

③ 엣지 분석 및 협업 분석 기술: 초연결 시대에 발생하는 패스트 데이터에 대한 엣지 분석과 영역별로 산재하는 다수의 엣지 분석 플랫폼들이 연계하여 하나의 글로벌 문제를 분석하고 해결하는 분산형 및 협업형 데이터 분석 기술입니다.
④ 모사현실 모델링 프레임워크: 복잡한 실제 세계를 모사현실로 구현하는 대규모 개방형 모델링 프레임워크 및 최적화 기술입니다.

#14
다. 활용
① 빅데이터의 유통 플랫폼 기술: 공공과 민간의 자유롭고 편리한 데이터의 등록과 검색 및 활용을 지원하는 플랫폼과, 데이터의 익명화와 같은 개인정보 보안성을 제공하는 빅데이터 유통 인프라를 구축합니다.

② 워크플로우 기반 적용 시나리오 구현 기술: 빅데이터 플랫폼의 적용 범위 확대를 위해 응용 분야별로 특화된 적용 시나리오를 워크플로우 기반으로 제공함으로써 빅데이터 플랫폼 및 분석 기술의 활용성을 제고합니다.

③ 데이터 품질의 정량화 및 최적화 기술: 데이터의 가치를 향상하기 위한 데이터의 체계적 축적 및 지속적 관리 체계를 구축하는 데이터의 라이프 사이클 관리 기술과 데이터의 품질 진단 및 개선 기술입니다.

④ 빅데이터 응용 및 서비스 기술: 누적된 데이터 또는 실시간으로 데이터를 발생하는 다양한 산업 분야(의료·건강, 소비·거래, 에너지, 재난안전 분야 등)의 도메인 지식(Domain Knowledge)과 융합하여 빅데이터 플랫폼 및 분석 기술을 적용하고 활용하는 응용서비스 기술입니다.