한국환경정책·평가연구원

빅데이터와 환경

01. 빅데이터란?

1. 빅데이터의 개념과 배경

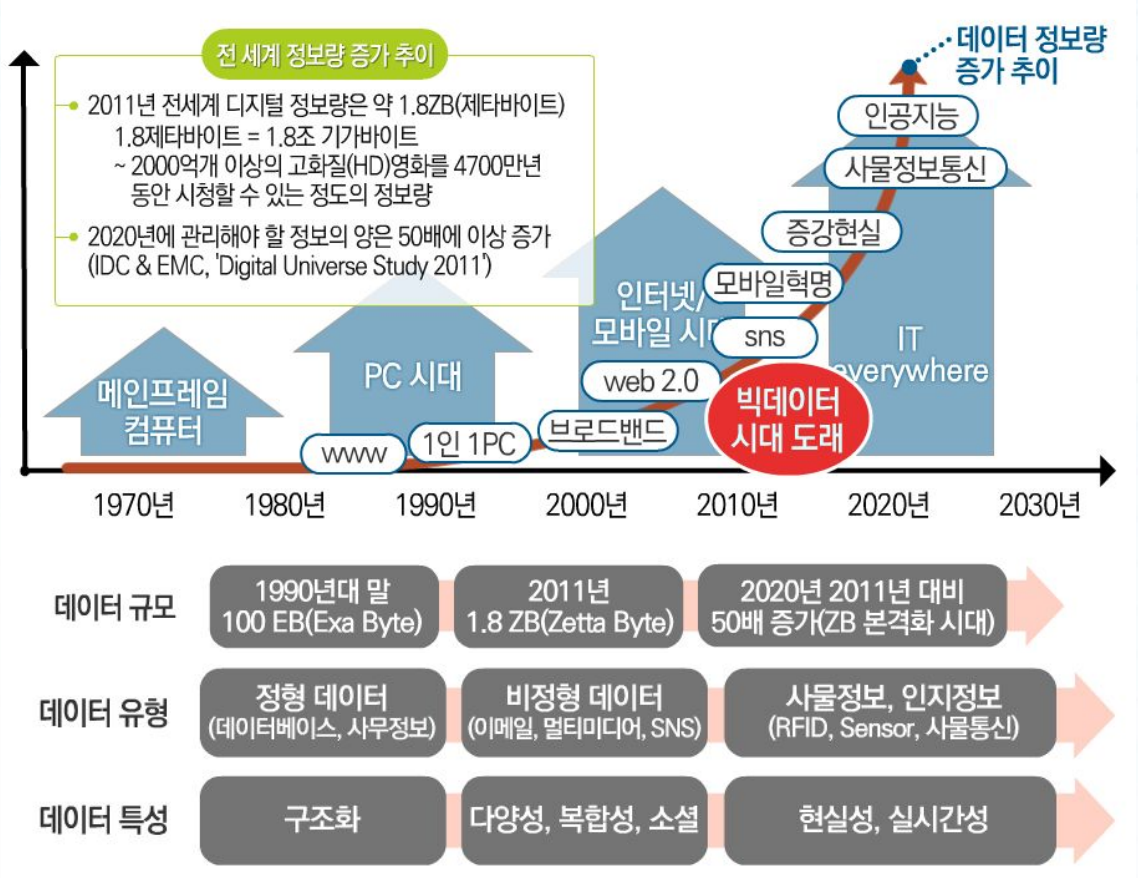
(1) 빅데이터 시대의 도래

우리는 최근 몇 년간 정보통신 기술의 비약적인 발전에 힘입어 이전에 경험해보지 못한 급격한 삶의 변화를 맞이하였다. 특히 정보저장 기술의 발전과 더불어 스마트 기기가 널리 보급되면서 일상생활의 모든 정보가 데이터로 축적되고 있다. 데이터 생성이 기하급수적으로 증가하면서 기존의 데이터와 구분되는 빅데이터(Big Data) 개념이 등장하였다[그림 1]. 정부와 기업은 방대한 데이터 속에서 필요한 정보만을 추출 및 분석하기 위해 많은 노력을 기울이고 있다. 즉, 주위에 널려 있는 수많은 데이터 정보로부터 유용한 통찰력과 지식을 누가 더 빨리 찾아내는지가 중요한 문제가 되었다.

빅데이터의 개념이 등장한 초기에는 데이터 대홍수, 데이터 범람, 대량의 데이터 등으로 지칭되었으나, 이후 다양한 분야에서 활발하게 논의 및 활용되었다. 빅데이터는 일반적으로 좁은 의미에서 기존 기술로 처리하기 어려운 방대한 양의 데이터를 지칭하며, 넓게는 이를 분석·관리하는 능력이나 기술, 도구 등을 포괄적으로 의미한다(함유근·채승병, 2012). 여기서 데이터를 분석·관리하는 기술이란 대규모 데이터를 분산 처리하는 '하둡(Hadoop)'과 확장성이 뛰어난 NoSQL 데이터베이스, 그리고 기계학습이나 통계 분석 등을 의미한다. Gartner (2012)는 빅데이터를 "고급 통찰력 및 의사결정을 위해 비용효과가 높은 혁신적인 정보처리 과정을 필요로 하면서 대량이며 급격하게 늘어나고 다양한 정보 자산"으로 정의하였다.

2008년 세계적인 과학저널 Nature[네이처]에서는 인터넷 이후 세상을 바꿀 가장 중요한 기술로 빅데이터를 언급하였다. 2010년 글로벌 컨설팅기업 McKinsey[맥킨지]는 비즈니스 지형을 바꿀 10가지 기술 트랜드 중 하나로 빅데이터를 선정하였으며, 빅데이터를 토대로 새로운 정보를 찾아내는 것이 가치창출 효과를 가져 온다고 분석하였다. 2011년 Gartner[가트너] 역시 주목해야할 기술로 빅데이터를 소개하였으며, 데이터는 21세기 '원유'이자 새로운 가치와 경쟁력의 원천임을 주장하였다.

[그림 1] 빅데이터 시대의 도래



자료: 김현곤(2012), 빅데이터 기반 선진 국정운영의 비전과 전략, 빅데이터 미래전략 세미나 발표자료.

(2) 빅데이터의 확산 요인

데이터의 양적 측면에서는 이전부터 빅데이터가 존재했다고 할 수 있지만, 빅데이터가 특수한 영역(우주, 항공, 과학 등)이 아니라 일상생활과 밀접한 환경에서 생성되고 있다는 점이 가장 큰 변화이다. 기존에는 전문가 및 특정 주체에 의해서 이루어지던 데이터 생산의 범위가 모바일 스마트 기기의 보급으로 일반대중으로까지 확대되었으며, 그 결과 누구나 데이터 생산의 주체가 될 수 있다. 페이스북(Facebook) 이나 트위터(Twitter) 등 소셜미디어(Social Media)의 텍스트 데이터가 대표적이며, 스마트폰의 보급도 빅데이터의 대중화 및 확산에 기여하였다.

<데이터의 대폭발>

- ○ 2010년~2015년까지 전세계 모바일 데이터 트래픽은 연평균 92%, 인터넷 트래픽은 연평균 34% 증가할 것으로 예상
- ○ 트위터는 전 세계 1억 명의 이용자들에 의해 하루 평균 2억 개의 트윗을 발생
  - 우리나라는 월 평균 100만여 명의 20~40대 이용자가 매월 1억 개의 트윗으로 의견을 표현
- ○ 오늘날 11억 인구가 소셜 네트워크 서비스(Social Network Service)를 이용하고 있고, 2억 5천만 명이 매일 페이스북에 사진을 업로드하고 있음

자료: 한국정보화진흥원(2013), 새로운 미래를 여는 빅데이터 시대.

더불어 컴퓨터의 가격성능비는 향상하고 데이터 저장매체의 가격은 하락하면서, 그동안 수집되지 않았던 방대한 양의 데이터를 본격적으로 축적할 수 있게 되었다. 그 결과, 데이터 수집 단위가 과거 메가바이트(megabyte, MB), 기가바이트(gigabyte, GB)에서 현재 테라바이트(terabyte, TB), 페타바이트(petabyte, PB), 제타바이트(zettabyte, ZB)로 크게 증가하였다.

<데이터의 단위>

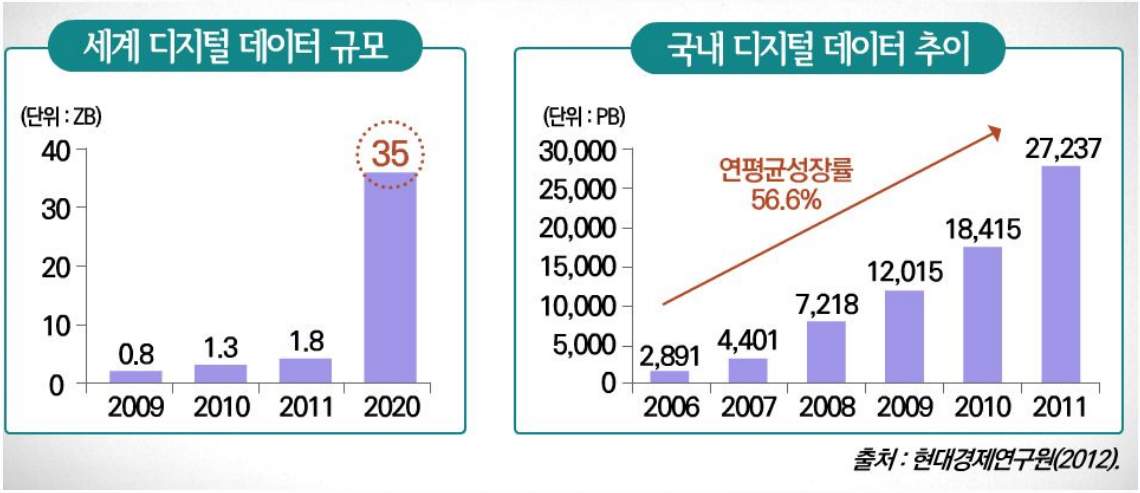
- ○ KB(kilobyte) = 10^3 byte
- ○ MB(megabyte) = 10^6 byte
- ○ GB(gigabyte) = 10^9 byte
- ○ TB(terabyte) = 10^{12} byte
- ○ PB(petabyte) = 10^{15} byte
- ○ ZB(zettabyte) = 10^{21} byte

빅데이터는 분 또는 초 단위 이하로 빠르게 생성 및 유통되기 때문에, 데이터 생성주기가 크게 단축되어 실시간으로 정보를 분석하고 활용하는 것이 가능해졌다. 무엇보다도 이전에는 쓸모없이 폐기되던 정보들에서 의미를 도출하고 새로운 가치를 창출할 수 있게 된 것이 빅데이터 분석의 가장 큰 특징이라고 할 수 있다. 이러한 대규모 데이터를 고속으로 처리할 수 있는 소프트웨어 기술 '하둡(Hadoop)'의 등장과 클라우드 컴퓨팅(cloud computing)의 보급으로 빅데이터를 축적·처리하는 기술이 보편화되었다.

(3) 빅데이터의 시장

빅데이터 시장은 매년 꾸준히 증가하고 있으며, 향후 증가속도가 더욱 빨라질 것으로 전망된다. 전 세계적으로 생성되는 디지털 데이터의 양은 2010년부터 매 2년마다 2배 이상 성장하여 2020년에는 약 35~40ZB 규모가 될 것으로 예측되며(IDC, 2012), 국내 디지털 데이터 생성량도 2006년 2,891PB에서 2011년 27,237 PB로 10배 가까이 급증하였다[그림 2].

[그림 2] 세계 디지털 데이터 규모 전망과 국내 디지털 데이터 추이



자료: 현대경제연구원(2012), 현안과 과제: 빅데이터의 생성과 새로운 사업 기회 창출.

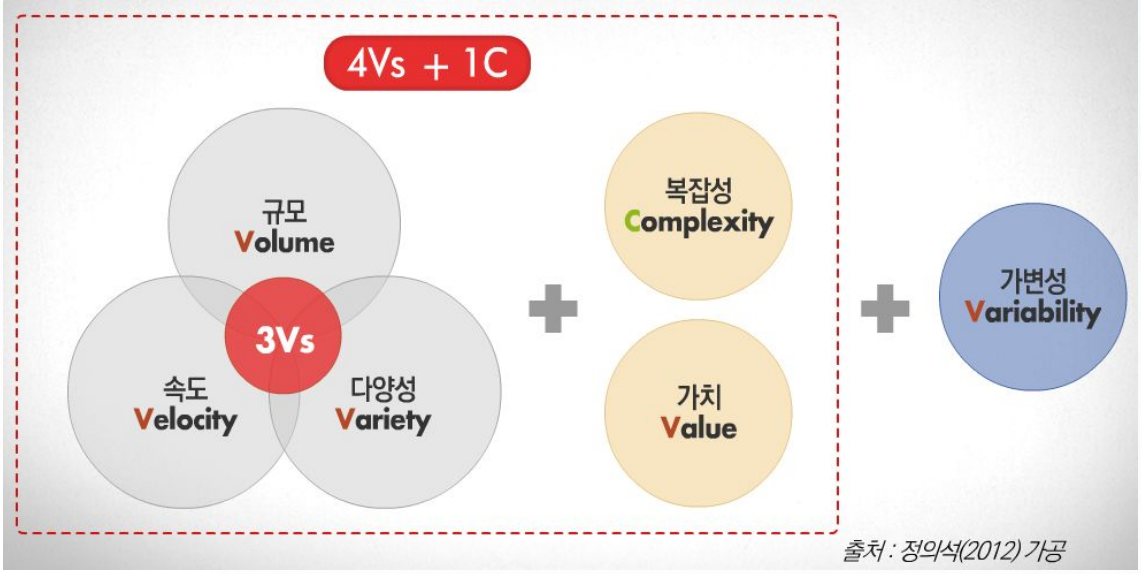
전 세계 빅데이터 산업의 시장규모는 2011년 74억 달러에서 2017년 405억 달러로 연평균 32.9%씩 성장할 것으로 전망된다[그림 3]. 특히 IDC는 2013년 보고서에서 빅데이터 산업의 성장속도가 ICT 산업 전체의 성장속도에 비해 6배 빠르게 성장할 것으로 전망한 바 있다(한국전자통신연구원, 2014). 또한 IT 하드웨어, 소프트웨어 등 정보통신 산업 전반에 대한 투자규모도 2020년까지 2012년 대비 40% 정도 꾸준히 증가함에 따라, 기가바이트당 투자비용이 현재 2달러 수준에서 2020년 0.2 달러까지 감소하여 빅데이터 산업이 더욱 활성화될 것으로 보인다(IDC, 2012).

2. 빅데이터의 특성과 유형

(1) 빅데이터의 특성

일반적으로 기존 데이터와 구분되는 빅데이터의 특성은 규모(Volume), 속도(Velocity), 다양성(Variety)을 의미하는 3Vs로 정의되며, 연구자에 따라 복잡성(Complexity)과 가치(Value)를 추가적으로 강조한 4Vs+1C로 확대하기도 한다[그림 3]. 이러한 특성 외에도 가변성(Variability)을 의미하는 V가 추가되기도 한다. 즉, 빅데이터는 데이터 규모와 기술 측면에서 출발했지만, 점차 가치와 활용의 측면으로 의미가 확대되는 추세이다.

[그림 3] 빅데이터의 주요 특성



자료: 정의석(2012) 가공, 이미숙 외(2014) 재인용.

규모(Volume)의 증가

'빅데이터'라고 하면 일반적으로 대량의 데이터를 생각하게 된다. 정보통신 기술의 발전으로 디지털 정보량이 기하급수적으로 증가함에 따라 제타바이트 시대로 진입하였으며, 특정 규모 이상을 빅데이터로 지칭하기보다는 기존 기술로는 관리할 수 없는 데이터양을 상대적으로 지칭한다고 이해할 수 있다.

다양성(Variety)의 증가

기업의 판매 데이터나 재고 데이터, 웹 로그 기록, 트위터나 페이스북과 같은 소셜 미디어 안의 텍스트 데이터, GPS 위치정보, 콜센터 통화 이력 등 다양한 종류의 데이터가 생성되고 있다. 또한 텍스트 데이터 외에 이미지, 동영상, 멀티미디어 등

수집하고 분석해야 할 데이터의 종류는 매우 다양하다.

속도(Velocity)의 증가

속도는 데이터의 변화와 축적, 분석 속도를 의미한다. 실시간성 정보가 증가하고 이로 인한 데이터 생성 및 이동 속도가 증가함에 따라, 대규모 데이터를 처리하고 가치 있는 정보를 활용하기 위해서는 데이터 처리 및 분석 속도가 중요하다.

복잡성(Complexity)의 증가

정형화된 텍스트 데이터에서 구조화되지 않은 비정형 데이터로 데이터의 종류가 확대되고, 외부 데이터의 활용으로 관리대상이 증가하며, 데이터 저장방식의 차이나 중복 등으로 인해 데이터의 관리 및 처리가 복잡해진다. 복잡한 데이터를 처리하고 분석하기 위해서는 기술적으로 새로운 기법을 필요로 한다.

가치(Value)의 창출

규모, 다양성, 속도의 3Vs (또는 를 가진 새로운 유형의 빅데이터로부터 과거에는 답할 수 없던 인사이트(Insight)나 새로운 가치(Value)를 도출할 수 있다는 측면에서 가치(Value)의 창출은 빅데이터의 중요한 특성으로 언급된다.

가변성(Variability)의 증가

데이터는 고정된 축적되지만 많은 옵션과 변수에 의해 분석과 해석이 쉽지 않다. 가변성은 많은 옵션과 다양한 변수로 인해 일정한 데이터로 분석되고 해석되지 않는 상황을 말한다. 전통적으로 제시되는 빅데이터의 특성 외에 추가적인 특성이다.

(2) 빅데이터의 유형

빅데이터 자원의 종류는 생성주체와 유형에 따라 다양하게 구분되며[표 1], 구조화 정도의 수준에 따라 크게 구조화된 데이터와 비구조화된 데이터로 분류할 수 있다.

[표 1] 빅데이터 자원의 종류

생성 주체	컴퓨터 생산 데이터	사람 생산 데이터	관계 데이터
	• 애플리케이션 서버 로그(웹사이트, 게임 등) • 센서 데이터(날씨, 물, 스마트 그리드 등) • 이미지, 비디오(트래픽, 보안 카메라 등)	트위터, 블로그, 뉴스, 게시판 글, 이메일, 사진 등	페이스북, 링크드인 등
유형	정형	반정형	비정형
	DB에 저장된 구조적 데이터	웹 문서, 메타 데이터, 센서 데이터, 공정 컨트롤 데이터, 콜 상세 데이터 등	소셜 데이터, 문서, 오디오, 비디오, 동영상, 이미지 등

자료: 이미숙 외(2014), 빅데이터를 활용한 환경분야 정책수요 분석, 한국환경정책평가연구원.

구조화된 데이터는 데이터 형식이 정형화되어 있으며, 데이터 간에 연계성을 바탕으로 데이터 정렬과 분석을 쉽고 빠르게 할 수 있다는 특징이 있다. 반면 비구조화된 데이터는 구조화가 되지 않았거나 구조화할 수 없는 데이터로서 다소 일관성이 없는 데이터 형식이다. 예를 들어 웹사이트에서 사용자에 의해 발생되는 클릭 스트림 데이터가 이에 해당된다.

최근 동영상, 음악, 위치정보 등의 미디어 콘텐츠에 대한 사용자 접근성이 용이해지면서 비구조화된 데이터의 양이 증가하였다. 실제로 2007년에 생성된 디지털 데이터의 약 95% 이상이 비정형 데이터이다(IDC, 2007). 비정형 데이터는 규격화된 데이터 필드에 저장되지 않는 데이터를 통칭하며, 텍스트 문서 외에도 이미지, 동영상, 음성 데이터 등을 포함한다(박대현·송동현, 2014).

3. 빅데이터의 가치와 활용

빅데이터의 활용은 불확실한 미래 사회에서 새로운 기회를 창출하고 위험요인을 해소하며 사회 발전을 위한 엔진 역할을 수행할 수 있다. 미래 사회의 발전 속도가 빨라지고 위험요인과 복잡성이 증가할수록 신속하게 사회적 변화를 감지하고 대응할 필요가 있다. 이러한 측면에서 빅데이터 분석은 미래 사회에서의 통찰력, 대응력, 경쟁력, 창조력을 향상시킬 수 있는 잠재력을 보유하고 있다[그림 4].

세계 각 국의 정부와 기업들은 빅데이터가 향후 기업의 성패를 가늠할 새로운 경제적 가치의 원천이 될 것으로 기대하고 있다. 예를 들어 McKinsey(2011)는 빅데이터의 활용으로 인해 의료, 공공행정, 소매, 제조, 개인정보 부문에서 추가적으로 1%의 생산성 향상이 가능할 것으로 기대하였으며, 산업 부문별로

상당한 경제적 가치를 창출할 것으로 예상하였다[그림 5].

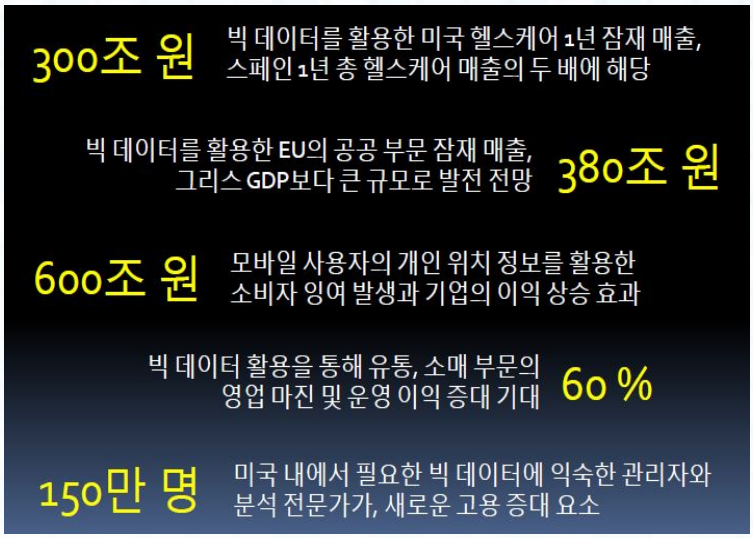
현재까지 빅데이터 활용의 선두 주자는 기업이다. 기업과 같은 민간부문에서는 생산력 제고, 효율 향상, 고객수요 파악, 브랜드 홍보 등 시장경쟁력 확보 차원에서 빅데이터 분석을 활용한다. 이 경우 기업 내부 데이터 외에도 페이스북, 트위터와 같은 소셜미디어를 통해 제품에 대한 고객의 반응을 실시간으로 모니터링하고 새로운 시장을 개척하는 데 필요한 정보를 확보하는 작업은 매우 중요하다. 뿐만 아니라 공공 부문에서도 위험관리시스템, 탈세 등 부정행위방지, 공공데이터 공개 정책 등 빅데이터를 활용하기 위해 다양한 노력을 기울이고 있다[그림 6].

[그림 4] 미래사회의 특성과 빅데이터의 역할

미래사회 특성		빅데이터의 역할	
불확실성	➡	통찰력	• 사회현상, 현실세계의 데이터를 기반으로 한 패턴분석과 미래전망 • 여러 가지 가능성에 대한 시나리오 시뮬레이션 • 다각적인 상황이 고려된 통찰력을 제시 • 다수의 시나리오로 상황 변화에 유연하게 대처
리스크	➡	대응력	• 환경, 소셜, 모니터링 정보의 패턴 분석을 통한 위험징후, 이상 신호 포착 • 이슈를 사전에 인지 · 분석하고, 빠른 의사결정과 실시간 대응 지원 • 기업과 국가 경영의 투명성 제고 및 낭비요소 절감
스마트	➡	경쟁력	• 대규모 데이터 분석을 통한 상황인지, 인공지능 서비스 등 가능 • 개인화, 지능화 서비스 제공 확대 • 소셜(니즈)분석, 평가, 신용, 평판 분석을 통해 최적의 선택 지원 • 트렌드 변화 분석을 통한 제품 경쟁력 확보
융 합	➡	창조력	• 타분야와의 결합을 통한 새로운 가치창출(의료정보, 자동차 정보, 건물정보, 환경정보 등) • 인과관계, 상관관계가 복잡한 컨버전스 분야의 데이터 분석으로 안전성 향상, 시행착오 최소화 • 방대한 데이터 활용을 통한 새로운 융합시장 창출

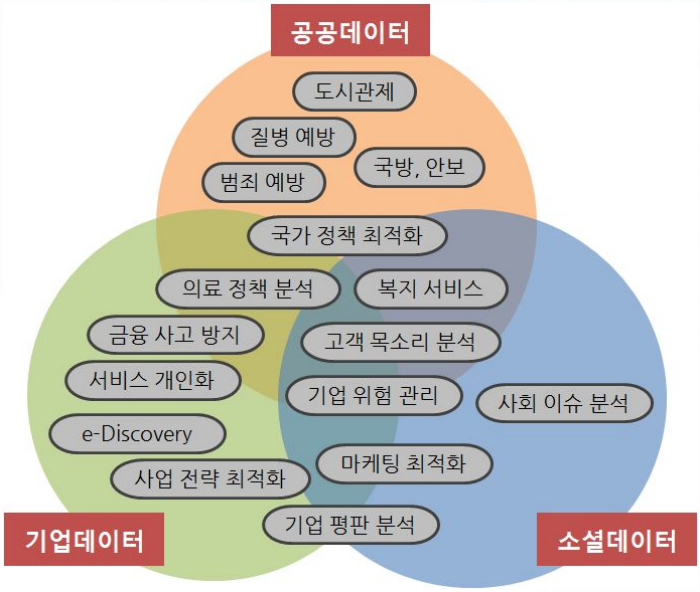
자료: 한국정보화진흥원(2013), 새로운 미래를 여는 빅데이터 시대.

[그림 5] 빅데이터의 활용 가치(McKinsey, 2011)



자료: 이경일(2012), 빅데이터와 신 가치창출, 빅데이터 미래전략 세미나 발표자료.

[그림 6] 빅데이터 분석 응용 사례



자료: 이경일(2012), 빅데이터와 신 가치창출, 빅데이터 미래전략 세미나 발표자료.

[정리하기]

1. 빅데이터의 개념과 배경

■ 정보저장 및 통신 기술의 비약적인 발전으로 누구나 손쉽게 정보를 생성하고 공유할 수 있는 기반이 마련됨에 따라 많은 양의 데이터가 빠르게 축적되기 시작하였고, 기존의 데이터와 구분되는 빅데이터라는 개념이 등장하게 되었다.

■ 빅데이터는 좁은 의미에서 기존 기술로 처리하기 어려운 방대한 양의 데이터를

지칭하며, 넓게는 이를 분석·관리하는 능력이나 기술, 도구 등을 포괄적으로 의미한다.

2. 빅데이터의 특성과 유형

- 일반적으로 기존 데이터와 구분되는 빅데이터의 3대 특성은 규모(Volume), 속도(Velocity), 다양성(Variety)을 의미하는 3Vs로 정의되며, 경우에 따라 복잡성(Complexity), 가치(Value), 가변성(Variability) 등의 특성이 추가되기도 한다.
- 빅데이터 자원의 종류는 생성주체와 유형에 따라 다양하게 구분될 수 있으며, 기존의 통계자료와 같은 구조화된 데이터뿐만 아니라 이미지, 동영상, 음성 데이터 등 비구조화된 데이터를 모두 포함한다.
- 최근 동영상, 음악, 위치정보 등의 미디어 콘텐츠에 대한 사용자 접근성이 용이해지고 스마트폰과 소셜 미디어의 확산으로 인해 비구조화된 데이터의 양이 급격히 증가함에 따라 빅데이터 분석 비중이 높아지고 있다.

3. 빅데이터의 가치와 활용

- 빅데이터는 미래 사회의 불확실성과 위험을 줄이고 급변하는 상황에 신속하게 대응하며 새로운 기회와 가치를 창출하기 위해 활용될 수 있다.
- 빅데이터의 분석 주체는 민간기업에서 공공기관에 이르기까지 다양하며, 데이터의 유형과 특성, 분석 목적에 따라 다양한 분야에 적용될 수 있다.

[참고문헌]

- 남광수(2004) 폐탄광 광해방지사업 추진실적 및 향후 계획, 광해방지와 환경복원 심포지움, pp. 17-39.
- Stumm and Morgan (1970)「Aquatic Chemistry」. Wiley-Interscience
- Nordstrom and Southdam (1997) Geomicrobiology of sulfide mineral oxidation. In: Geomicrobilogy: Interactions between Microbes and Minerals. JF Banfield and KH Nealson (eds), Reviews in Mineralogy 35. Mineralogical Society of America.
- Mason, B and Berry LG (1968)「Elements of Mineralogy」. Freeman.
- 김현곤(2012), 빅데이터 기반 선진 국정운영의 비전과 전략, 빅데이터 미래전략 세미나 발표자료.
- 박대현, 송동현(2014), 비정형 데이터 활성화의 정치, 경제, 문화적 함의,「Internet & Security Focus」, 한국인터넷진흥원.
- 이경일(2012), 빅데이터와 신 가치창출, 빅데이터 미래전략 세미나 발표자료.

- 이미숙 외(2014), 빅데이터를 활용한 환경분야 정책수요 분석, 한국환경정책평가연구원.
- 정의석(2012), 스마트 교육환경과 빅데이터 활용전략, 제2차 빅데이터 국가전략포럼 발표자료.
- 한국전자통신연구원(2014), 특허분석을 통한 빅데이터 기술개발 동향, 「전자통신동향분석」 제29권 제2호.
- 한국정보화진흥원(2013), 새로운 미래를 여는 빅데이터 시대.
- 함유근, 채승병(2012), 빅데이터, 경영을 바꾸다, 삼성경제연구소.
- 현대경제연구원(2012), 현안과 과제: 빅데이터의 생성과 새로운 사업 기회 창출.
- 황승구(2012), 빅데이터 기술현황 및 전망, 빅데이터 미래전략 세미나 발표자료.
- Gartner(2011), Hype Cycle for Emerging Technologies.
- Gartner(2012), The Importance of 'Big Data': A Definition.
- IDC(2007), The Expanding Digital Universe, An IDC White Paper - Sponsored by EMC.
- IDC(2012), IDC Digital Universe Study: Big Data, Bigger Digital Shadows and Biggest Growth in the Far East, Sponsored by EMC. ppt file.
- McKinsey&Company(2010), Clouds, big data, and smart assets: Ten tech-enabled business trends to watch.