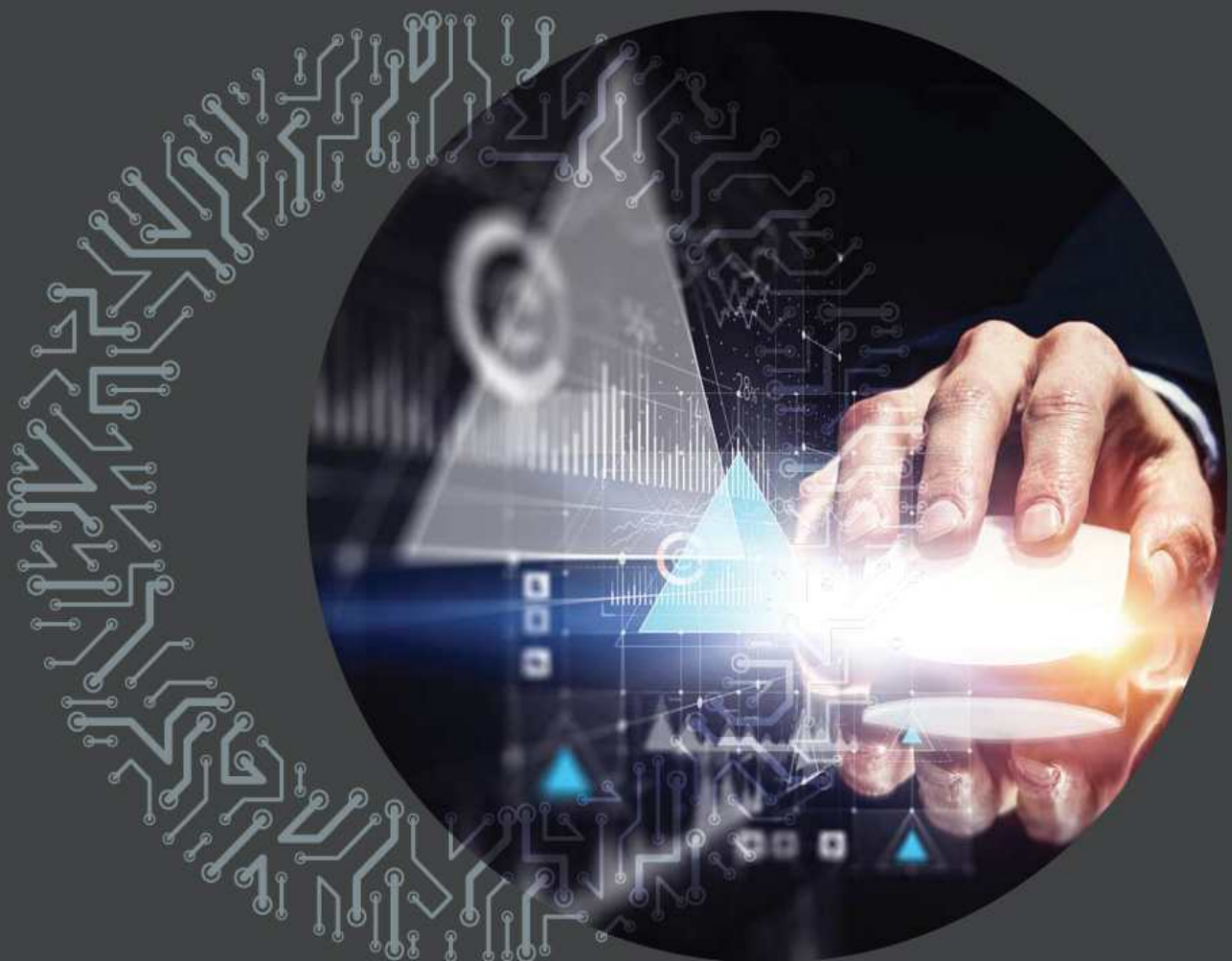


Insight Report

빅데이터 플랫폼 현황 및 이슈 분석



※ 본 보고서의 내용은 필자의 개인적인 견해이며, 한국전자통신연구원의 공식 견해가 아님을 알려드립니다.

본 문서에서 음영처리된 부분은 () 정보공개법 제9조의 비공개대상정보와 저작권법 및 그 밖의 다른 법령에서 보호하고 있는 제3자의 권리가 포함된 저작물로 공개대상에서 제외되었습니다.



본 저작물은 공공누리 제4유형: 출처표시+상업적이용 금지+변경금지 조건에 따라 이용할 수 있습니다.



요 약	1
I. 빅데이터 플랫폼 출현	4
II. 빅데이터 플랫폼 현황	9
1. 빅데이터 플랫폼의 구성	9
2. 빅데이터 및 플랫폼 산업 생태계 현황	17
3. 빅데이터 플랫폼 기업 현황	22
III. 빅데이터 플랫폼 이슈 분석	41
IV. 시사점 및 향후 제언	56
참고문헌	59



요 약

빅데이터와 플랫폼 출현

- 인류는 데이터를 기반으로 중요 의사결정을 수행하며 현재의 문명을 건설
 - 모든 학문은 과학적 방법론을 중시하며 실험과 경험에 기반한 원리 도출에 연구를 통해 발전하였으며, 데이터는 의사결정 행위에 핵심적 도구로 활용
- ICT의 발전을 통해 주변 모든 데이터가 디지털화 되었다면, 현재는 디지털화된 데이터가 인간 삶의 질 향상을 위해 활용되기 시작하는 단계
 - 데이터 이해도 증가, 대규모 데이터 기반 추론을 가능하게 하는 ICT 기술의 발전으로 새로운 부가가치가 창출되는 영역들의 출현 증가
- 현재 생성되는 데이터는 규모·생성속도 측면에서 기존 데이터와 차별하여 빅데이터로 불리고 있으며, 가치 창출을 위한 핵심 기반으로 평가
 - 기존에는 데이터를 생성하는 주체가 인간/소비자였지만, 이외에 무인자동차, IoT 등의 신기술을 통해 기계가 빅데이터를 실시간으로 무수히 많이 생산
 - 빅데이터는 활용, 검증의 성격을 넘어 가치 창출의 기반 엔진으로서 역할을 할 전망이며, 특히 Digital Transformation 실현을 위한 핵심 영역으로 인식

빅데이터 플랫폼 현황

- 빅데이터 플랫폼은 빅데이터에서 가치를 추출하기 위한 일련의 과정을 지원하기 위한 프로세스를 규격화한 기술·서비스 모임
 - 빅데이터 플랫폼은 데이터의 수집 → 저장 → 처리 → 분석 → 시각화 등을 통해 원시데이터(Raw Data)로 부터 Insight 및 가치 추출
 - 전통적인 통계적 처리와 빅데이터 플랫폼의 차별성은 큰 틀에서는 데이터의 수집, 저장, 처리 영역에서 발생하는 것으로 평가됨
- (꾸준한 성장세의 글로벌 시장) 빅데이터 시장은 '20년까지 꾸준한 성장세를 보일 전망이며, Digital Transformation이 본격화되는 시점('20년 예상, IDC) 부터는 큰 변화가 예상됨

- 국내 빅데이터 시장은 '16년 3,440억원 시장 추산(전년대비 31% 성장)
 - 국내 기업의 빅데이터 시스템 도입율은 '15년 4.3%에서 '16년 5.8%로 상승
- (산업생태계 및 경쟁상황) 빅데이터 인프라 시장은 글로벌 사업자들의 선점이 상당히 진행되었으며, 국내기업은 빅데이터 활용을 위한 기반 마련 중
 - 데이터의 저장과 Query, 분석, 시각화 등의 모든 영역은 이미 글로벌 기업들이 범용성 있는 제품·서비스·SW 출시 중
 - '16년 국내 빅데이터 기술은 최고기술 대비 65.7%, 약 3.1년의 격차 추산
- (SW·애플리케이션 중요성 부상) 인프라 구축 진행으로 저장·처리 영역보다는 분석과 가치 창출 지원 SW와 애플리케이션 영역의 중요성 증대
 - 저장과 전달에 치중되었던 빅데이터 기술 개발은 실시간 가치창출을 위한 인메모리 처리, 하둡 개선, 사용자 중심 기술 등으로 R&D 핵심이 이동

빅데이터 플랫폼 이슈 분석

- (빅데이터의 부재) 실질적으로 활용할 수 있는 빅데이터는 미흡한 구축 현황
 - ▶ 빅데이터를 도입하지 않은 국내기업들이 지적한 가장 큰 요인은 빅데이터라고 할 만한 데이터가 없어, 빅데이터 구축 기반 형성이 어렵다는 의견
 - 빅데이터 구축이 활발한 정부나 공공기관의 공공데이터 공개는 투명한 국정 운영, 국민의 알권리 제고, 경제적 효과가 있는 것으로 평가
 - 미국, 영국 등의 선진국에서는 공공정보 공유와 정보 활용을 위한 웹사이트를 구축하여 제공하는 중
- (빅데이터의 품질) 다양한 형태(시장, SNS, IoT)와 유형(정형·비정형)의 빅데이터 증가는 데이터 품질 및 데이터 자체의 편의(bias) 문제 유발
 - ▶ IBM은 2016년 낮은 품질의 데이터로 인한 미국의 지출 비용을 3.1조(Trillion) 달러로 추정(HBR, 2016.9.22.)
- (개인권리 침해) 구축된 빅데이터를 활용하여 개인 정보의 확보나 추정이 가능해져 개인 권리를 침해하는 사례도 급증
 - 국내에서는 법령상 '빅데이터 개인정보보호 가이드라인'('14.12월)을 통해 빅데이터의 비식별화 조치를 할 경우 적극적 이용이 가능

- (데이터의 과용·오용) 빅데이터를 활용한 가치 창출 노력이 과하여 아무 의미 없는 데이터에 의미를 부여하고 연관짓는 빅데이터 오용 사례도 발생
 - 데이터 기반 추론이 과하게 진행되어 Apophenia 현상의 발생 가능성 존재
- (오픈소스) 빅데이터 플랫폼을 구성하는 하둡, 맵리듀스, 스파크, 하이브 등 핵심 기술들은 모두 오픈소스로 개발 중
 - 글로벌 기업들의 오픈소스 채택 및 활용은 꾸준한 증가세에 있음
 - ▶ ‘16년 오픈소스 사용하는 기업은 65%(Black Duck SW, 2016). ‘15년의 60%
 - ▶ 기업들이 오픈소스를 사용하는 이유로는 솔루션 품질, 경쟁적 특징과 기술적 역량, 커스터마이징과 수정 역량 등을 지적
 - ▶ 국내의 오픈소스 개발자는 11,000명, 커미터는 약 500명 정도로 저조한 수준

시사점 및 향후 제언

- 빅데이터와 빅데이터 플랫폼은 아직은 인프라 구축과 데이터를 효율적으로 축적하는 기술 발전에 머무르는 상황으로 평가됨
 - 빅데이터의 장점인 비정형 데이터, 실시간 분석 영역은 현재 기술 개발이나 구축이 진행 중이나, 본격적인 활용에는 미흡
- 오픈소스 기반의 플랫폼 기술 성장으로 전통적인 클라우드 기업뿐만 아니라 수요자 중심의 커스터마이징화된 SW를 제공하는 중소벤처기업들도 시장에서 성장할 수 있는 경쟁 구도
- (인공지능과의 연계 강화) 왓슨, 알파고를 시작으로 산업 측면에서 인공지능의 활용성에 대한 기대가 높아지는 상황
- 데이터 강국으로 진입하기 위한 공개와 공유, 수집 기반 조성
- 행정, 의료, 재난, 환경, 교육 등의 정보를 효율적으로 저장·분석 가능한 빅데이터 플랫폼으로 전환 구축
 - 국내 공공·복지서비스 제공에 필요한 국내 공공 정형·비정형데이터를 통합관리할 수 있는 통합플랫폼의 개발 및 Customized된 서비스 구현
 - 일반 사용자가 쉽게 활용할 수 있는 오픈소스 기반 빅데이터 분석 Tool도 한국형으로 개발하고 이를 기반으로 공공 빅데이터 플랫폼의 역량 제고

I 빅데이터 플랫폼 출현

인류 문명 발전의 원동력은 데이터에 대한 이해

- 인류는 데이터를 기반으로 중요한 의사결정을 수행하며 현재의 문명을 건설
 - 중세시대 이후 물리학, 천문학 등의 발달로 인해 인류는 자연 현상의 관찰을 통해 추상화, 모델링함으로써 현상의 기반을 형성하는 법칙을 탐구
 - 산업혁명이후에는 자연과학뿐만 아니라, 경제학, 사회과학 등 모든 학문 영역에서 과학적 연구 방법론을 도입하며 실험과 경험에 기반한 원리 도출에 많은 연구를 진행하였으며, 데이터는 연구 방법론의 발전에 핵심적 도구 역할
- 현대의 복잡한 사회, 산업에 대한 이해를 위해서 철학적 요소보다는 객관성을 높이기 위한 데이터 수집에 집중하며 현상을 연구
 - 사회과학에서는 인간의 행태를 설명하기 위해 인간의 행위를 데이터 이해하려고 하였으며 이에 기반한 합리적 이해와 추론, 비즈니스 의사결정을 추진
 - ▶ 90년대 마케팅에서는 기업 목표를 달성을 위해 통제 가능한 마케팅 변수로서 제품(Product), 가격(Price), 유통(Place), 촉진(Promotion) 등 소위 4P를 기반으로 의사결정을 진행했으며, 이를 위해 기반 데이터로서 소비자 성향 조사를 활용
 - 공학, 의학, 자연과학 등의 모든 부문에서도 객관적 설명을 위해 추상화, 모델링, 시뮬레이션을 기본적으로 수행하고 있으며, empirical study는 필요한 과정으로서 자리 잡음
 - ▶ 이론의 검증을 위해 컴퓨터를 이용한 모의실험이 필수적인 단계가 되었지만, 오히려 이론의 발전 속도를 컴퓨팅 파워가 쫓아오지 못해 다차원 모델링, 최적화 등에서 한계를 보이기도 함
 - ▶ 예로서, 신경망 이론은 1990년대 부분적으로 이론적인 성숙을 보였으나, 컴퓨팅 파워의 한계로 인해 실시간에 기반한 현실적 활용에는 어려움 존재
- 모바일 인터넷, SNS, IoT, 음성비서, 영상인식 등의 ICT기반 기술·서비스의 급격한 발전으로 압도적인 볼륨과 종류의 데이터가 만들어지는 상황
 - 이와 더불어, 클라우드 구축, 엣지 파워의 증대로 인해 실시간으로 처리할 수 있는 데이터의 양은 지속적으로 증가하는 상황
 - 컴퓨팅 파워의 증대는 기존에 불가능했던 이론 기반의 모의실험을 가능하게

하고, 특히 실시간으로 결과를 얻게 됨에 따라 기기와 연동하며 빠른 대응이 필요한 서비스의 제공을 가능케 함

- ▶ 알파고는 분산형 컴퓨팅과 혁신적인 알고리즘을 이용하여 실질적으로 바둑 명인과 대국이 가능할 정도로 성능이 향상됨
- ▶ IBM 왓슨(Watson)은 퀴즈쇼에서 인간을 굴복시키고, 현재는 의학 등에서 영상인식을 통해 경험많은 의사보다 높은 질의 진료 행위를 수행 중
- ▶ 자율주행차가 실시간으로 처리하는 데이터는 현재 시간당 GB라고 하지만, 향후에는 초당 1GB 데이터를 처리해야 시점도 올 수도 있다는 전망

● 데이터에 대한 이해도 증가와 데이터를 처리할 수 있도록 지원하는 ICT 기술의 발전은 새로운 부가가치를 만들어 낼 수 있는 영역들을 생성시키고 있으나, 이 영역들이 대부분의 인간의 행위와 중복되며 우려를 발생시킴

- 인공지능은 데이터를 중심으로 발전하는 기술·서비스 영역에서 실제로 체감할 수 있는 분야
- 데이터가 알고리즘 등과 연동되며 인간의 아날로그적·감성적 영역을 제외한 여러 부문에서 인간 행위와의 대체성이 높아질 전망이며, 실질적으로 인간의 행위 영역을 무차별적으로 침범 중
 - ▶ 제46차 세계경제포럼(2016.1)에서는 AI, 로봇공학, IoT 등의 발전으로 2020년까지 전세계 일자리 710만개가 사라지고, 200만개가 새로 생길 것으로 전망
 - ▶ 최근 연구에 의하면 경기와 고용은 정의 관계성을 갖는 것으로 인식되지만, 90년 이후로는 연관성이 낮아지고 있다는 점을 강조. 이는 기업이 생산량이 늘이기 위해 사람보다는 기계를 사용하는 것이 원인으로 설명¹⁾
 - ▶ BBC에서는 향후 40년 이내에 AI가 다음 부문에서는 인간보다는 월등해질 것으로 예측. translating languages (2024년), writing school essays (2026년), driving a lorry (2027년), working in retail (2031년), writing a best-selling book (2049년), working as a surgeon (2053년)²⁾
 - ▶ 현재까지 인공지능이 강점을 보이는 작업 성격은 빠른 데이터 처리와 분류(Classification)로서 자율자동차, X-ray 판독, 음성 인식 등이 예
- 그럼에도 인간의 고유한 의사결정과 감정적인 요소, 창조성 등은 인공지능이 흉내낼 수 있는 범주는 아닌 것으로 평가됨에 따라 이러한 업무에 대한 인간의 기회는 모든 데이터에 기반한 결정이 진행되더라도 존재함이 지적됨

1) 에릭 브린올프슨·앤드루 맥아피. (2014). 제2의 기계시대(The Second Machine Age, 2014)

2) BBC. (2017.6.20.). AI may take your job - in 120 years

- ▶ 혹자는 인공지능이 가능한 업무는 절차적인 요소가 강한 닫힌 시스템의 업무이며, 외부적 매개가 필요한 영역은 여전히 인간의 몫으로 남을 것으로 판단

빅데이터는 디지털 전환(Digital Transformation)의 기반 영역

- 네트워크 구축 기반의 ICT 발전이 사회의 디지털화를 위한 기반을 제공하였다면, 사회의 디지털화 성과가 인간을 위해 피드백 되기 시작

- 2000년대 초반까지 IT진화의 중심이었던 네트워크 구축, 정보화, 휴대폰(피쳐폰 중심)등을 통해 사람들이 정보를 생성하는 주체로서 변화됨

- ▶ 이 시점에도 많은 양의 데이터가 생성되었으나, 데이터에 내재된 가치에 대한 인식 수준은 높지 않아 매우 제한적 자료만 공유되거나 대부분의 경우 버려지는 상황. 필요시 기업들은 자료를 별도로 재생산하는 구도

- 2000년대 후반 스마트폰의 출현은 유선이 중심적 역할을 하고, 무선·모바일은 보조였던 소비자의 행동 및 데이터 생산·소비 양식을 일거에 모바일 중심으로 완벽히 변화 시키며 연관된 모바일 기반 비즈니스도 크게 확장

- ▶ 위치기반 서비스, 소셜 네트워크 서비스(SNS) 등의 모바일 서비스는 소비자 스스로 다양한 종류의 데이터를 생성할 수 있는 상황을 마련

- ▶ 특히, 이 시점에는 개인 소비를 위한 데이터 뿐만 아니라, 선거, 공공 등 다양한 분야에서 영리 목적으로 데이터가 사용되기 시작

- ▶ 2008년 미국 대통령 선거에서 버락 오바마는 유권자 데이터를 기반으로 유권자별로 차별화된 맞춤형 전략 추진하여, 당선에 도움이 되었다는 평가

- ▶ 검색을 중심으로 성장한 Google, SNS인 Facebook 등은 검색자료나 친구 관계를 통한 막대한 데이터를 축적하고 있으며, 데이터 주도권을 확보

- 2010년대 중반부터는 인공지능, 클라우드 등 컴퓨팅의 빠른 발전과 더불어 무인자동차, IoT 등 다량의 데이터를 생성하거나 소비하는 영역이 발전

- ▶ 기존에는 데이터를 생성하는 주체가 인간·소비자였지만, 무인자동차, IoT 등은 통해 기계가 데이터를 실시간으로 무수히 많이 생산

- ▶ 소비자가 생성하는 데이터는 인구통계학적, 취미, 기호 등 DB화가 가능한 형태였으나, SNS, 모바일이 생성하는 데이터는 형태가 매우 다양해지고, 일률적인 수집에도 어려움 발생

- ▶ 이들 영역은 데이터를 능동적으로 생성, 저장하고 빠른 분산처리를 통해 실시간으로 결과를 도출하여 최적의 대응을 가능하게 할 전망

- ▶ 인공지능은 그동안 암흑기를 거친 후에 힌튼 교수의 연구결과와 알파고라는 대중적인 사건을 통해 발전 가능성이 논의되고 있음
- 현재 생성되는 데이터는 산업, 인간 속성을 모두 반영되며 다양한 가치를 함의하고 있으나, 인간이 주도적으로 처리하기에는 규모가 방대하고 다차원적 요소 포함
 - 현재의 데이터는 빅데이터(Big Data)로 명명되고 있으며, 잠재된 가치를 실시간으로 채굴하는 국가, 기업이 성장 주도권을 갖는 것으로 인식됨
 - ▶ 빅데이터는 일반적으로 크기(Volume), 다양성(Variety), 생성 속도(Velocity) 등의 3V로서 그 특징을 설명
 - 데이터 관점에서 가장 주목해야할 관점은 Variety로서 데이터의 생성처가 다양해지며 일관된 구조를 가지지 않은 SNS, 댓글, 로그 데이터 등 소스에 따라 특성화된 비구조화된(Unstructured) 데이터가 확대되고 있음
 - ▶ 분석시 다양한 데이터 소스로 부터의 다양한 관점을 파악할 수 있어 의사결정의 효율성·효과성을 높일 수 있다는 측면이 있음
- 디지털 전환을 위해서는 필연적으로 빅데이터 구축이 필요하며, 구축된 빅데이터가 다시 생성자로 환원되며 가치를 발생할 때 전환은 완성될 전망
 - 이들 데이터는 인간이 이성으로 인지하며 구축하는 데이터 범위를 넘어서며, 전체 현실의 세계를 디지털화하여 데이터로 투영한 것임
 - Digital Transformation의 완성에서는 디지털화된 데이터들을 기반으로 가상세계에서 사회를 구성하고 가상세계 속의 의사결정을 위한 최적의 경로를 찾아가며 다시 실제 사회로 피드백되는 구도로 평가됨
 - 실제 세상의 모든 데이터가 Digital로 변환된다는 것은 컴퓨터가 이를 직접적으로 활용할 수 있기 때문에, 기존 인간이 데이터를 활용할 수 있게 컴퓨터를 조종하는 단계에서 궁극적으로 인간이 배제될 수 있는 환경에 놓이게 됨
 - 컴퓨터와 인간의 연결 고리가 끊어지는 시점에서는 컴퓨터의 활용 체계 및 역할, 인간의 역할, 인간 생활과 비즈니스가 현재와는 완전히 다른 체계로 변화될 전망되며, 파괴적 혁신 이상의 진화가 일어날 가능성 존재

<참고> 디지털 전환(Digital Transformation)의 정의

- 디지털 전환은 통일된 정의는 없으나, 공통적으로 기존의 모든 것이 Digital 기반으로 재정의되고, 이에 기반하여 전체적인 시스템이 디지털을 기반으로 변화·혁신·진화되는 것
- Bain & Company, PWC 등 다양한 기관에서 디지털 전환에 대한 정의를 하고 있으나, 기반이 되는 정의는 디지털 기반으로 새로운 산업과 비즈니스를 형성하는 것에 주목
 - ▶ Bain & Company : 디지털 엔터프라이즈 산업을 디지털 기반으로 재정의하고 게임의 법칙을 근본적으로 뒤집음으로써 변화를 일으키는 것
 - ▶ PWC : 기업경영에서 디지털 소비자 및 에코시스템이 기대하는 것들을 비즈니스 모델 및 운영에 적용시키는 일련의 과정

● 디지털전환은 위험성을 가진 면이 있지만 기업들에게는 완벽한 신규 환경이 주어진다는 관점에서 자사가 보유한 역량을 기반으로 빅데이터와 관련된 일련의 프로세스 역량을 확보하기 위해 다양한 전략을 추진

- 현재 빅데이터를 통한 가치 창출에 대한 의문점은 여전히 해소되지 않은 상황으로서, 정부·기업 등은 효율성 제고, 신규 서비스 발굴, 비용 절감 등의 관점에서 소극적인 활용과 준비를 진행
- 초기 빅데이터 성장시에는 국가별로 데이터 수집이 활발한 정부·공공기관에서 가치를 창출하기에 알맞은 영역으로 특히 의료, 제조, 재난·공공 등의 분야에서 축적된 데이터를 통해 활용사례가 증가 중
 - ▶ AI 확산 조기대응, GE의 지능형 항공운행 서비스 등 실질적인 성과 도출(미래부·NIA, 2015년 빅데이터 글로벌 사례집)

Ⅱ 빅데이터 플랫폼 현황

1. 빅데이터 플랫폼의 구성

빅데이터의 처리 과정

- 빅데이터 플랫폼은 빅데이터에서 가치를 추출하기 위한 일련의 과정을 지원하기 위한 프로세스를 규격화한 기술·서비스 모임
 - 빅데이터로 부터 가치 추출을 위한 프로세스는 경영학이나 경제학, 사회과학, 전산학 등에서 진행 중인 전통적인 자료 분석과 유사
 - 그럼에도 빅데이터의 특성인 Volume과 비정형, 실시간성으로 인해 데이터의 수집과 저장, 분석 등의 모든 영역에서 학문적인 측면보다는 서비스 중심적인 개념으로 표현하는 경향을 보임
- 빅데이터 플랫폼은 데이터에 대해 수집 → 저장 → 처리 → 분석 → 시각화 등을 통해 원시데이터(Raw Data)로 부터 Insight 및 가치(Value) 추출
 - 전통적인 통계적 처리와 빅데이터 플랫폼의 차별성은 큰 틀에서는 데이터의 수집, 저장, 처리 영역에서 발생하는 것으로 평가됨
 - 분석, 시각화 측면에서는 가치 추출을 위해서 전통적인 통계 분석에서는 인과관계³⁾를 최종적인 결과로서 제시한다면, 빅데이터에서는 연관관계·상관관계⁴⁾를 중심으로 시사점을 도출하려는 경향을 보임

표 1 빅데이터 플랫폼의 구성 요소별 주요 기능들

수집	저장	처리	분석	표현
비정형 수집 정형 수집 ETL Web Robot EAI, ESB, FTP 등 Open API	원본 데이터 NoSQL 메모리 검색엔진 데이터보안	배치처리 실시간처리(CEP)	텍스트분석 Machine Learning 통계 Data Mining SNS 분석 예측분석(알고리즘)	시각화

3) 인과관계(Causality) : causation 혹은 cause and effect라고도 함. 선행하는 한 변인(cause)이 후행하는 다른 변인(effect)의 원인에 책임이 있으며, 후행 변인은 선행 변인에 부분적으로 의존하는 현상

4) 상관관계(Correlation) : 한 통계적 변인이 다른 여러 통계적 변인들과 공변하는 관계. 회귀분석, 상관계수 등을 활용하여 측정

- 빅데이터가 처리되는 절차를 중심으로 조망하는 방법은 결국에는 절차 구현을 위한 기술 중심으로 분류하는 관점
 - ▶ 기술과 이에 기반한 제품, 주된 활동 영역 등이 혼재되어 있어 이해하기 힘들 수 있으나, 가장 쉬운 접근법
- 현재 빅데이터는 기술적인 관점에서 기업, 정부 등이 고민없이 사용하기 위해서는 확장성, 동시성, 보안성 등을 만족시켜야 함
 - ▶ 하둡은 현재 가장 많이 활용되는 분산처리 시스템이지만, 디스크 기반이며 사용법을 익히기 어려워 엔지니어가 필요하다는 단점이 있다는 평가
 - ▶ 동시성 측면에서 하둡 이외에 스파크(Spark)가 우위에 서있는 것으로 평가받지만, 여전히 메모리 기반하에서 생기는 확장성과 보안성에 대한 연구가 진행 중

● 빅데이터의 생성 · 수집

- 빅데이터는 주관성이 개인되는 정형 데이터(Structured Data), 인터넷과 연계되거나 IoT 등에 의해 자동적으로 생성되는 비정형 데이터로 크게 분류
 - ▶ 정형데이터(Structured data)는 조사할 방향성을 미리 결정하거나 혹은 자동 분류가 가능하여 고정된 필드 설정이 가능한 데이터로서, 센싱, 마케팅 정보, 선호도 조사, 과학 데이터 등이 해당
 - ▶ 비정형데이터(Unstructured data)는 고정된 필드 형태로 항상 저장 가능하지 않은 데이터로서 SNS, Web 등에서 생성되는 실시간 데이터가 다수 점유
- 데이터의 폭증은 디지털서비스의 발전에 기인한 것으로 SNS, Web 로그 등 인터넷 데이터를 기반으로, 공장, 환경, 자동차 등에서 발생하는 계측·센싱 등의 기계적으로 실시간성, 비정형성을 가진 데이터를 양산하고 있는 상황
- 데이터 수집처로서는 정부 영역, 위치기관 서비스, 금융 및 경제 영역, IoT, 의료 등의 부문에서 대표적으로 발생하고 있으며, 스마트폰과 웨어러블 기기, 자율자동차 등 향후 인간과 연계된 모든 영역에서 데이터가 생산될 전망
 - ▶ 정부 부문에서는 미국의 Data.gov, 한국의 Data.go.kr 등이 대표적이며, 위치기관 서비스로는 Foursquare, Navigation, 재무 데이터로는 블룸버그, 다우존스, IoT 영역에서는 GE, Health에서는 Jawbone 등 다양한 데이터 소스 존재
- 데이터를 수집하는 주요 기술로는 ETL(Extraction, Transformation, Loading), Crawling Engine, 로그 수집기, Web Robot, IoT 센싱 등 있음

● 빅데이터의 저장

- 빅데이터의 3V 특성을 반영하여 데이터를 저장하는 기술로서, 단순히 저장만을 목적으로 하는 것이 아니라, 저장된 데이터를 분석할 수 있는 형태로 다시 로드하고 공유하는 측면이 충분히 고려되어야 함
- 정형화된 프레임을 갖춘 데이터베이스에 순차적으로 데이터를 저장하는 작업은 불가능하여, 정형화된 틀을 적용시키지 않기 위한 다양한 기술들이 개발되고, 특히 추후 Searching 혹은 Query를 위한 사전 작업이 반영됨
- 주요 기술로는 분산파일시스템, NoSQL, 비디스크 기반 DBMS 등이 있음
 - ▶ 분산파일시스템에는 Hadoop 시스템, 구글 파일 시스템, 아마존 S3 파일시스템 등이 대표적
 - ▶ NoSQL의 대표적 사례로는 키 값 모델 기반(Dynamo, Membase), 열 기반(Bigtable, Hbase, Cassandra), 문서 기반(CouchDB, MongoDB) 등이 있음
 - ▶ 분산 메인 메모리 기반 DBMS로는 SAP HANA와 VoltDB가 대표적이며, 플래시 메모리 활용 DBMS는 Oracle Smart Flash Cache 등이 대표적

● 빅데이터의 처리

- 저장된 빅데이터에 대해 수요발생시 적시에 가공하고 분석을 지원하는 기술
- 데이터 처리를 위해서는 실시간 처리, 분산병렬 처리, 인-메모리 처리, 인-데이터베이스 처리 등의 방법이 있음
- 주요 기술로는 구글의 맵리듀스(Map Reduce), Hadoop의 맵리듀스, 마이크로소프트의 Dryad 등이 있으며, 처리 프로그래밍 기술로는 구글의 Sawzall, Hadoop의 Pig, Apach Hive 등의 있음

● 빅데이터의 분석

- 저장되어 있는 빅데이터에서 잠재된 가치와 통찰력을 추출하기 위한 작업
- 전통적인 통계적 분석을 중심으로, 변량들의 특징, 연관관계, 인과관계 등을 도출하고, 종합적으로 사고할 수 있는 틀을 제공
- 분석의 수준은 분석을 진행하는 연구자에 의해 크게 좌우되며, 분석의 적절한 방향 설정이 알맞게 되어야 명료한 분석 결과 도출이 가능
- 주요 방법론으로 전통적 통계 분석, 데이터 마이닝, 텍스트 마이닝, 연관 분석, SNS 분석 등이 있으며, 지속적으로 급격한 발전을 이루는 분야

● 빅데이터의 시각화

- 주관적인 판단의 제거, 왜곡없는 객관성 확보, 부차적인 인지과정 없이 추론하는 것이 데이터 분석 기반하에 추구하는 형태
 - 일반적으로 최초 분석시 데이터의 표현 방법은 수치나 키워드, 그래프 구조로 제시되지만, 정보 의미를 가장 빠르고 정확하게 전달하는 효율적인 그림 혹은 이미지로 제공하는 방법
 - ▶ Frieman⁵⁾(2008) : 데이터 시각화의 주요 목적은 그래픽 수단을 통해 명확하고 효율적으로 정보를 교류하는 것으로 정의. 지루하거나 극단적인 아름다움을 추구하는 것이 아니라, sparse하거나 복잡한 데이터를 직관적 관점을 가지고, 미적 형태와 기능성 두가지를 모두 가져야 함을 지적
 - 최근에는 Computing 능력의 확대, 그래픽 도구 및 기능 향상, 정보 전달 기법의 발전, 모바일 인터넷의 실생활 확산, 데이터에 시간, 공간 정보 추가 데이터 표현 다변화
 - 시각적으로 표현된 자료는 감각적으로 인식하면서 정보에 대한 빠른 수용과 이해도 제고가 가능. 특히, 시계열적인 요소가 있는 데이터에 대해 시간에 따른 데이터의 변화를 보여줄 경우 매우 유용한 정보의 획득 가능
 - 디자인을 통해 데이터를 표현하는 인포그래픽스 영역 성장 중
 - 시각화는 정확한 분석 방향을 가지지 못할 경우 탐색적인 연구를 위해 지속적으로 시각적으로 데이터를 표현하여 분석의 방향성을 탐색함
 - 시각화의 주요 분야로서 전통적인 통계 결과의 그래픽 처리 뿐만 아니라, Web log 시각화, SNS 시각화, 지도 시각화 등이 있음
 - 데이터 시각화의 주요 도구로서는 Timeline, Modestmaps, Leafletjs, Wolframalpha, Visualizefree, d3js 등이 있음
- (산업별 분류 체계) 빅데이터 시장에 대한 시장 조사 자료에서는 전통적인 분류 방식에 따라 빅데이터를 구현하는 제품별로 시장을 분류
- 주요 부문으로 Server, Storage, Network, Software, Service 등⁶⁾

5) Vitaly Friedman (2008) "Data Visualization and Infographics"

6) 한국정보화진흥원(2016.1), "2015년 BIGDATA 시장현황 조사"

하둡 생태계

● 빅데이터의 이해는 결국 하둡(Hadoop) 생태계의 이해

- 빅데이터를 다루기 위해서는 분산처리 기술의 활용성이 가장 높으며, 이를 대표하는 기술은 하둡(Hadoop)

▶ Hadoop : High-Availability Distributed Object-Oriented Platform

- 하둡은 자바기반 프레임워크로서 대용량의 데이터를 여러개의 저장소에서 분산 처리하는 방식

▶ 너치의 분산처리를 지원하기 위해 개발되었으며, 아파치 루씬의 하부 프로젝트

▶ 분산처리 시스템인 구글 파일 시스템을 대체할 수 있는 하둡 분산 파일 시스템(HDFS: Hadoop Distributed File System)과 맵리듀스를 구현한 것

● 분산처리를 위한 Hadoop의 초기 솔루션으로서 HDFS와 MapReduce의 모듈로 구성

- 상기 두 모듈은 오픈소스로서 비전문가들이 빅데이터를 활용하기에는 어려움이 많았으며, 이의 해결을 위해서 다양한 주변 모듈들이 구현되어 패키징됨
- 데이터의 통합, 이동, 애플리케이션 매니지먼트, 시스템 매니지먼트 등을 위한 지원 SW가 하둡 프로젝트의 일환으로 개발됨
- 하둡 지원 서비스 프로그램은 빅데이터의 수집, 저장·활용, 처리, 관리 등을 데이터 처리와 관련된 모든 영역을 대상으로 개발 진행 중

하둡의 기능을 보완하는 서비스 오픈소스 프로그램들

빅데이터 구분	주요 기술	기술별 주요 기능
스트리밍 데이터 수집	Flume	• 비정형 데이터 수집 • 클라우드데라에서 개발됐으며, 현재 아파치 인큐베이션에 포함됨
	Scribe	• 비정형 데이터 수집 플랫폼 • 중앙 집중 서버로 전송하는 방식으로, 페이스북에서 개발
	Chuckwa	• 비정형 데이터 수집 플랫폼으로 HDFS에 분산데이터를 저장
정형 데이터 수집	Sqoop	• 정형 데이터 수집 • 관계형 DB로부터 데이터 가져오기 • HDFS, NoSQL 등 다양한 저장소로의 전송 지원
	Hiho	• 대용량 정형 데이터 수집 및 전송 솔루션
분산 데이터 베이스	Hbase	• 분산 데이터베이스 • HDFS기반의 컬럼 기반 NoSQL 데이터베이스, 구글의 BigTable 논문을 기반으로 개발됨 • 야후, 트위터 등이 사용하며, 국내 NHN도 라인에 적용
	Cassandra	• 오픈소스 분산 데이터베이스 관리 시스템 • 컬럼 중심 DB와 행 중심 DB의 복합형

빅데이터 구분	주요 기술	기술별 주요 기능
실시간 SQL 질의	Impala	- NoSQL의 하나 - 하둡 기반의 실시간 SQL 질의 시스템 - 클라우드데라에서 개발 - 맵리듀스로 처리하지 않고, 자체 개발한 엔진 사용
	Tajo	- 국내 대학 학생들이 주도해 개발한 하둡 기반의 DW 시스템 2013년 아파치 재단의 인큐베이션 프로젝트로 선정
메타 데이터 관리	HCatalog	빅데이터 메타 정보 관리
데이터 분석	Hive	- 하둡 기반의 데이터웨어하우징용 솔루션 - 유사 SQL 기반 빅데이터 처리 - 페이스북에서 개발하여, 오픈 소스로 공개
	Pig	- 데이터 분석 - MapReduce 대신 자체 언어 Pig Latin 제공
인메모리 처리	Spark	- 오픈소스 클러스터 컴퓨팅 프레임워크 - UC 버클리 AMPLab에서 처음 개발함
데이터 마이닝	Mahout	- 데이터 마이닝 - 하둡 기반의 오픈 소스
워크 플로우 관리	Oozie	- 빅데이터 처리 과정 관리 - 하둡 작업을 관리
분산 코디네이터	Zookeeper	- 빅데이터 서버 시스템 관리 - 분산 환경 서버들간의 상호 조정 서비스
직렬화	Avro	RPC(Remote Procedure Call)과 데이터 직렬화를 지원하는 프레임워크
리소스 매니저	YARN	- 리소스 관리 플랫폼 - 분산 컴퓨팅 환경 제공 - 클러스터내 컴퓨팅 자원 관리 - 사용자의 애플리케이션 스케줄링 사용 관리
자료 : http://over153cm.tistory.com/entry/하둡-에코시스템HadoopEcosystem이란 & http://iamdaisy.tistory.com/11 , 제타위키 내용을 기반으로 서술		

빅데이터 플랫폼 기술 발전 전망

- (SW·애플리케이션 중요성 부상) 인프라 구축이 진행됨에 따라 저장/처리 영역보다는 분석과 가치 창출을 지원하는 SW와 애플리케이션 영역의 기술 개발과 제품 출시 중요성이 확대 중
 - 하둡 기반의 저장과 전달에 치중되었던 빅데이터 기술 개발은 실시간 가치창출을 위한 인메모리 처리, 하둡의 개선 등으로 R&D 핵심이 변화되는 중
 - ▶ 맵리듀스를 이용한 하둡 기반의 배치처리 구축은 소기의 성과를 보이고 있으며, 스파크 등 실시간 데이터 처리 플랫폼에 대한 신규 제품 등이 출현
 - ▶ 다양한 종류의 데이터를 추가적인 처리없이 데이터로서 활용가능한 분산 처리에 대한 수요가 매우 높아, 이에 대응되는 저장, 처리 기술의 출현 가속화
 - 데이터로서 부터 의미있는 가치를 쉽게 도출하기 위해 분석자 중심으로 분석 SW와 시각화 툴이 지속적으로 발전하고 있는 상황

- (플랫폼의 다양화) 데이터의 수집부터 시각화까지의 모든 단계를 포괄하는 범용적인 플랫폼의 개발 뿐만 아니라, 특화된 데이터에 대한 순차적 처리가 가능한 플랫폼에 대한 수요도 확대될 전망
 - 기업에서는 데이터 엔지니어, 분석가 뿐만 아니라, 일반 실무자들도 데이터를 다루고 가치를 도출할 수 있는 범용 플랫폼의 개발 수요 급증 전망
 - ▶ 하둡을 기반으로 데이터 처리를 하기 위해서는 상당한 수준의 데이터 처리 역량이 필요하며, 또한, 가치 도출에 있어서도 시각화 툴 등을 해석할 수 있는 능력이 필요하여, 실무자들의 경험이 빅데이터 분석에는 직접적으로 활용되기 어려운 상황
 - 의료 기록 분석, 환경 모니터링, 범죄 현황 분석, 무인자동차 자료 분석 등과 같이 특정 영역에서 발생하는 빅데이터에 대해 특화된 분석을 지원하는 플랫폼이 발전 추세
 - 또한, 특화된 영역의 빅데이터 플랫폼은 신규 데이터와의 연계성 확대, 기계 학습 등의 자동화 기술과 연계되면서 분석의 다양성이 확대될 것으로 전망
 - ▶ 빅데이터 영역 중 복잡한 데이터를 가장 많이 보유한 의료 영역에서는 빅데이터 플랫폼 도입이 활발하면서도, IBM 왓슨을 도입하여 데이터 분석과 의사결정을 위한 협력 관계를 확대하고 있음. 보험사 웰포인트는 의료진의 치료계획안이 적절했는지를 왓슨을 활용하여 판단함
 - ▶ 국내에서는 효성 ITX의 스마트팩토리 구현을 위한 빅데이터 플랫폼, 수원시와 부산시의 맞춤형·시민개방형 빅데이터 시스템(공공데이터 및 국민연금, 의료, 신용·부채 등의 분석, 시각화 지원)
- 향후 빅데이터 기술의 성공은 분석 단계에서의 기계학습, 인공지능과의 연계하여 실시간으로 의사결정 지원 등의 가치를 창출할 수 있는 기술 역량이 가장 중요성 성공요인으로 등장할 전망
 - 데이터의 입력, 탐색과 데이터의 분석, 표현, 출력 등을 인공지능 영역이 담당하면서 핵심적인 인간의 역할 없이 사용자를 이해하고 최적 맞춤형 가치의 창출·제공이 가능해진다는 점이 큰 차별성 발생
 - ▶ 무인자동차는 자체 센서나 IoT를 통한 외부 데이터로부터 수집되는 빅데이터를 분류, 선택하고 이에 대한 최적의 판단을 하는 알고리즘을 탑재
 - ▶ 영상의학에서는 인체 사진에 대한 패턴을 기존에 구축된 의학 빅데이터를 통해 학습하고, 신규 사진에 대한 패턴 인식을 통해 질병 진단

- 향후 데이터의 수집과 분류, 패턴 인식 등 반복적인 업무 영역에서는 빅데이터를 위한 분석이 모두 인공지능으로 이루어질 전망이며, 비즈니스나 산업의 생산성은 크게 높아지고, 인간의 참여 정도가 지속적으로 낮아질 전망
 - ▶ 이미 빅데이터 플랫폼에는 IBM의 Watson, Apple의 Siri, Google의 Knowledge Graph 등이 접목되면서 빅데이터 활용이 진화되고 있음
 - ▶ 그럼에도 가치라는 영역은 최적과는 다른 개념이므로, 인간 역할은 축소되지만, 인간에게 영향을 주는 부문에 대해서는 지속적인 참여가 예상됨
 - ▶ 인공지능 플랫폼은 빅데이터 플랫폼에 비해 강조하는 영역이 사용자와의 접점으로 음성인식(아마존 알렉사, 시리), 시각인식(Nvidia의 DrivePx), 자연어처리 등을 핵심적 기술로 제시. 사용자로 부터 인지한 내용은 결국 빅데이터 플랫폼 내에서 처리된 후 피드백되는 형식을 취함
- 과학기술, 기초과학 영역에서 발생하는 빅데이터 분석은 데이터의 방대함과 함께 컴퓨팅 역량, 분석 알고리즘 고도화를 위해 HPC, 컴퓨터 사이언스, 기초과학과의 연계가 필요
 - ▶ 향후 빅데이터 산업은 종합과학의 성격을 가짐에 따라 활용에 치중한 산업발전이 지속될 경우 선진국과의 격차는 줄지 않을 전망
 - ▶ 환경, 의료, 재난 등에서 발생하는 빅데이터는 이론 연구, 알고리즘, 시뮬레이션의 종합적인 측면이 필요함에 따라 장기적인 정부투자 필요성 확대
 - ▶ 유전체 분석을 위한 연구 등이 ETRI에서 진행되었으나, 단기성에 그침

2. 빅데이터 및 플랫폼 산업 생태계 현황

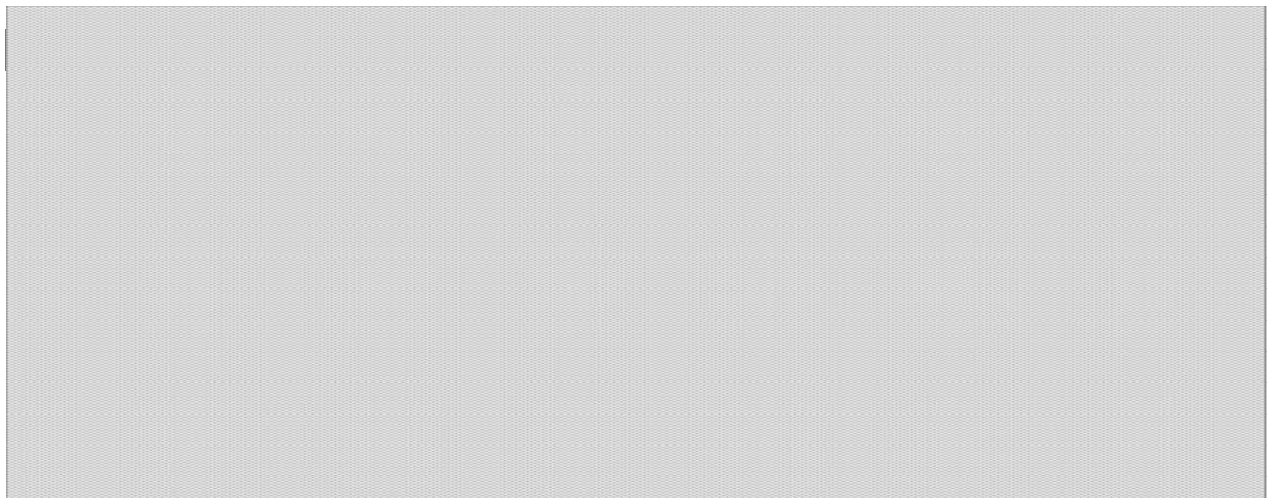
글로벌 비즈니스는 플랫폼 기반 비즈니스로 변화

- 빅데이터와 관련된 일련의 프로세스는 빅데이터 플랫폼이라는 용어로서 표현되고 있으며, 글로벌 기업들은 자사의 역량을 기반으로 빅데이터 플랫폼을 구축하고 파생되는 비즈니스의 주도권을 잡기 위해 전략 추진 중
 - 플랫폼의 주도권을 확보하기 위해서는 기기, SW, 콘텐츠의 주도권을 확보하는 것도 중요하지만, 현재는 이를 이루는 데이터의 주도권을 확보하는 것이 매우 중요한 전략
- 현재 글로벌 ICT 비즈니스에 가장 큰 영향을 주는 플랫폼 혹은 기업은 Google, Apple, Amazon, Microsoft 등이 가장 많이 언급되고 있으며, 여러 타 업종에서도 데이터를 이용한 비즈니스로 전환 중
 - 이들은 모두 자사의 역량을 기반으로 비즈니스 생태계를 자사 중심으로 변화시켰으며, 데이터에 대한 주도권을 확보하고 있다는 특징을 보임
 - ▶ Google : 검색, Apple : 모바일을 통한 소비자 행태 데이터, 아마존 : 상품 및 클라우드, Microsoft : 데이터 분석 툴
 - 특히, 상기 기업들은 인공지능 기술을 적용하여 자율주행자동차, 로봇, 챗봇, 인공지능 등의 신규 기술을 빠르게 완성하고 있어 플랫폼 주도권 확보를 위한 비즈니스 범위도 크게 확대될 전망
- 전통적인 제조업, 의약계 등에서도 제조공정이나 진료 행위 중에 발생하는 빅데이터를 기반으로 신규 부가가치 창출을 추진하는 양상
 - ▶ 오르론 : 제조 공정상 빅데이터 수집·분석 플랫폼을 도입하여 생산성 향상
 - ▶ 볼보 : 자동차 부품에 센서를 탑재하여 차량 데이터를 수집하는 등의 차량 데이터 분석 플랫폼을 구현하여 차량의 운행 관리와 효율성 제공

빅데이터 관련 시장 전망 및 산업생태계 현황

- (꾸준한 성장세의 글로벌 시장) 빅데이터 시장은 '20년까지 꾸준한 성장세를 보일 전망이며, 디지털 전환이 본격화되는 시점부터는 큰 변화가 예상됨
 - 빅데이터는 '16년 264억불 시장에서 '20년 589억불 시장으로 성장할 전망

- 빅데이터 인프라 영역이 가장 큰 비중을 점할 것으로 예상되며, '16년 136억 불 시장에서 '20년 277억불로 연평균 20.3%의 성장을 보일 전망
 - ▶ 향후 빅데이터 인프라 영역은 소프트웨어, 서비스 영역에 비해 성장세는 다소 주춤할 것으로 예상되며, 빅데이터의 가치 창출에 대한 의문점이 해소되지 않는다면 빅데이터의 인프라 부문은 성장 정체를 겪을 전망
- 빅데이터SW는 '20년까지 159억불로 성장하여 서비스 시장과 유사할 전망
- 서비스 부문도 24%(15~20)의 성장세를 보이며 SW와 유사한 규모인 152억 불의 시장을 형성할 전망
- SW에서 주목할 점은 탐색 및 분석 SW가 연평균 27.2%로 빠른 성장 전망
 - ▶ 이는 현재 빅데이터 플랫폼을 구성하는 인프라 영역과 함께 분석을 지원하는 SW 영역의 수요가 크게 성장할 전망이므로, 이를 통해 빅데이터 플랫폼의 구성이 완성되어갈 것으로 전망됨



- 국내 빅데이터 시장⁷⁾은 '16년 3,440억원 시장 추산(전년대비 31% 성장)
 - ▶ '13년 1,643억원 → '14년 2,013억원(전년대비 22%) → '15년 2,623억원(전년대비 30%) → '16년 3,440억원(전년대비 31%)
- 시장 영역별로는 정부·공공 시장이 998.6억원(29%), 민간시장인 2,441억원(71%)으로 추산
 - ▶ 특히 정부·공공 시장은 공공기관에의 빅데이터 도입·활용에 대한 적극적인 권고로 전년대비 43.1%로 크게 성장
- 영역별로는 스토리지 시장이 전체 시장의 26.5%인 911.3억원으로 가장 크

7) 한국정보화진흥원. (2017.3). 2016년 빅데이터 시장현황 조사

며, SW 809.3억원(23.5%), 서버 736.5억원(21.4%), 서비스 715.1억원(20.8%), 네트워크 267.4억원(7.8%)로 구성됨

- ▶ 서비스 비중은 '14년 11.3%에서 '16년 20.8%로 성장하여, 빅데이터 분석의 중요성이 증대되는 것으로 평가
- ▶ 국내 빅데이터 시장에서 스토리지 비중이 26.5%인 점을 고려할 때 국내 시장은 아직은 빅데이터 인프라 구축에 치중하고 있으며, 빅데이터 활용 측면에서는 시간이 걸릴 것으로 판단됨
- 국내 일반기업의 빅데이터 시스템 도입율은 '15년 4.3%에서 '16년 5.8%로 상승하였으며, 특히 중견·대기업(종업원 수 1천명, 매출액 1천억 이상)은 '15년 9.6%에서 '16년 12.3%로 증가
 - ▶ 산업별로는 응답자 기준으로 금융 14%, 공공 6.9%, 통신미디어 6.5%, 의료 6.3%, 유통·서비스 5.5%, 제조 4.2% 순으로 데이터가 많이 구축되는 분야는 빅데이터의 도입에 상대적으로 적극적인 것으로 판단됨
 - ▶ 빅데이터 시스템 미도입 기업의 61.6%는 도입에 대해 논의한 적이 없다고 응답하였으며, 도입 논의된 기업들 중 26.3%는 전산부서 차원에서만 논의 진행
- 빅데이터 전문인력은 '16년 9,300명이 있는 것으로 추산되며, '19년에는 15,772명으로 전망되어 약 6,500명의 추가 인력이 필요할 전망
 - 전문인력 중 빅데이터 개발자가 '16년 2,703명으로 29%로 가장 많이 종사하는 직종이며, 엔지니어, 사이언티스트, 컨설턴트 등이 17% 정도 유사한 비중
 - 빅데이터 분석가는 '16년 11.3%(1,052명)을 점유하여 낮은 수준이나, '19년에는 14.5%(2,280명)으로 1,200여명의 신규 수요가 발생할 전망
 - 빅데이터 개발자는 '19년에도 여전히 가장 높은 비중을 점할 전망

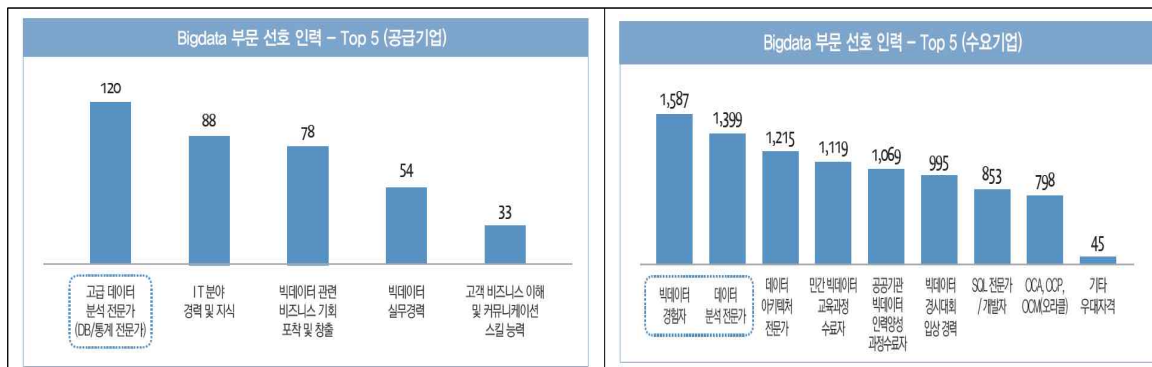
표 3 직무별 빅데이터 전문인력 현황 및 전망(전체 산업대상 추정치) (단위 : 명)

구분	2016년		2019년		16년 대비 '19년 필요인력 증가	
	인력 수	비중	인력 수	비중	인력 수	비중
빅데이터 개발자	2,703	29.0%	5,841	37.0%	3,138	116%
빅데이터 엔지니어(하둡, NoSQL)	1,602	17.2%	2,253	14.3%	651	41%
빅데이터 분석가	1,052	11.3%	2,280	14.5%	1,228	117%
데이터 사이언티스트	1,662	17.8%	2,218	14.1%	556	33%
빅데이터 컨설턴트	1,606	17.2%	2,142	13.6%	536	33%
빅데이터 기획, 마케팅	696	7.5%	1,038	6.6%	342	49%
총계	9,321	100.0%	15,772	100.0%	6,451	69%

자료 : 한국데이터진흥원. (2017.7). 2017 데이터산업 백서

- 국내 빅데이터 공급기업이 원하는 빅데이터 전문인력은 고급 데이터 분석 전문가(DB/통계전문가)이며, 국내 빅데이터 수요기업이 원하는 전문인력은 빅데이터 경험자, 데이터 분석 전문가로 조사됨

국내 빅데이터 공급 및 수요 기업의 전문인력 요건

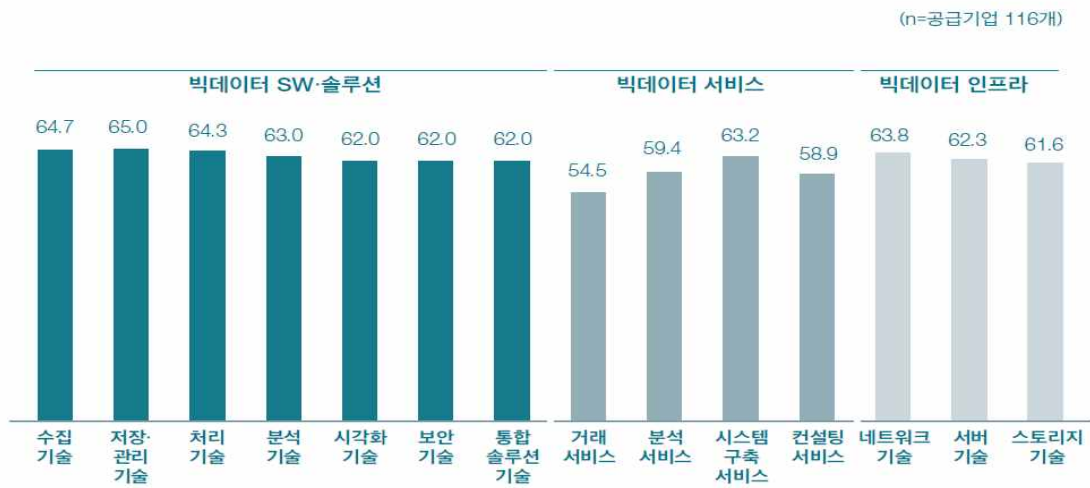


자료 : 한국정보화진흥원. (2016.2). 2015년 빅데이터 시장현황 조사.

● (플랫폼 산업생태계 및 경쟁상황) 빅데이터 인프라 시장은 글로벌 사업자들의 선점이 상당히 진행되었으나, 국내의 기술수준은 미흡한 실정

- 데이터의 저장과 Query, 분석, 시각화 등의 모든 영역은 이미 글로벌 기업들이 범용성 있는 제품·서비스·SW 출시 중
 - ▶ '15년 Business Analytics SW 시장 점유율은 Oracle(16.4%), SAP(13.1%), IBM(10.3%) 순(IDC, 2016)
 - ▶ '15년 EMEA 지역의 Big Data Storage Vendor 점유율은 HP(21.2%), Dell(17.4%), ODM Direct(15.4%), EMC(14.6%) 순으로 조사됨(IDC, 2016)
- (국내 기술 수준 현황) '16년 국내 기술 수준은 최고기술 대비 65.7% 수준이며, 약 3.1년의 격차를 보이는 것으로 추산됨
 - ▶ '15년의 국내기술 수준인 62.6%, 3.3년의 격차에 비해 소폭 축소
 - ▶ 공급기업 분야별로는 전체적으로 65%이하이며 매우 열위한 상태. 이 중 거래, 분석, 컨설팅이 포함된 빅데이터 서비스 영역은 더욱 낮은 수준

분야별 국내 기술수준 평가(공급기업)



자료 : 한국데이터진흥원. (2017.7). 2017 데이터산업 백서

3. 빅데이터 플랫폼 기업 현황

빅데이터 플랫폼 기업 특성

- 빅데이터 플랫폼 기업들은 빅데이터가 극복해야 하는 실시간 처리, 메모리 기반 처리, 쿼리의 용이성, 다양한 파일 시스템 접근성 등의 문제점을 해결하는 전략을 추진하며 기업 생태계를 형성
 - 아마존 등 클라우드 역량을 확보하였던 기업들은 자사의 클라우드와 오픈소스 기반 빅데이터 솔루션을 연계하여 차별화된 플랫폼 구축 노력 진행
 - 기존 기업정보화를 지향하던 IBM, Microsoft 등은 빅데이터로 전략을 수정하여 기존 역량을 활용
 - 구글, 애플 등 모바일 기반의 신형 기업들은 자사 서비스 제공을 위해 구축된 서버와 사용자 데이터 정보를 기반으로 빅데이터 분석 체계 구축
- 빅데이터 플랫폼을 구현하는 기업은 토털솔루션을 제공하는 경우와 특정 영역의 솔루션 중심으로 제공하는 경우로 구분 가능
 - 아마존, IBM, 구글 등은 빅데이터 분석의 전과정을 지원하는 자체적인 플랫폼을 구성하고, 이를 기반으로 기업에게 컨설팅을 제공하여 수익 창출
 - 테라데이터, 굿데이터 등과 같이 구축된 빅데이터에 대해 분석, 시각화 등의 기능을 구현하는 솔루션 기업들도 있으며 차별화된 성능을 구축하고자 노력

표 4 | 빅데이터 플랫폼 기업과 주요 특징

기업명	강점 플랫폼	주요 특징
아마존	AWS (Amazon Web Service) 아마존 EC2, 아마존 S3	- 클라우드 컴퓨팅 서비스 - 인공지능 비서 알렉사를 AWS와 연계
IBM	InfoSpheres BigInsights Stream 등	- 기업용 솔루션의 전통적인 강자 - 왓슨을 기반으로 빅데이터와 인공지능을 결합
마이크로소프트	Azure	- 클라우드 컴퓨팅 플랫폼 및 인프라스트럭처 서비스 - Azure IoT Suite를 통해 IoT를 손쉽게 구현
오라클	Oracle R Advanced Analytics for Hadoop, Data Visualization 클라우드 서비스	- 가장 성공적인 데이터베이스 기업
SAP	HANA 클라우드 플랫폼 빅데이터 허브	- ERP, CRM 등의 전통적 강자 - 사물인터넷 분야의 선점 전략 추진 - 인메모리 컴퓨팅 플랫폼(SAP S/4 HANA) 제공 중

기업명	강점 플랫폼	주요 특징
구글	구글 클라우드 플랫폼	- 구글의 데이터 센터를 기반으로 Compute, 스토리지, 빅데이터, 머신러닝 기능 제공 - 클라우드를 저렴하게 사용할 수 있는 세분화된 가격 정책을 도입
인텔	Xeon 프로세스	- 인메모리 지원을 강화한 서버용 프로세서 Xeon Process E7 등 제품 개발
시스코	CPA (Common Platform Architecture)	- 하둡 기반 빅데이터 어플라이언스 플랫폼 제공 - 하둡 솔루션은 고객이 선택 가능한 개방적 전략
GE	Predix	- 산업인터넷용 어플리케이션을 적시에 개발할 수 있도록 지원하는 소프트웨어 서비스 세트 - 현재 프레딕스 생태계에는 약 2만명의 서드파티 개발자들이 활동 중
맵알 (MapR)	Converged Data Platform	- 하둡 상용 솔루션 전문 기업 - 엣지(Edge) 플랫폼 강화 전략 추진
클라우데라 (Cloudera)	DSWB (Data Science WorkBench)	- 하둡 상용 솔루션 전문 기업 - 교육 및 기술지원 서비스에 주력
호튼웍스 (Hortonworks)	HDP (Hortonworks Data Platform)	- 하둡 상용 솔루션 전문 기업
DataStax	DataStax Studio	- 아파치 카산드라 상용화 버전 개발
Informatica	PowerCenter	- 데이터 통합·관리 SW업체 - 모든 유형의 트랜잭션 데이터에 액세스 가능
VMware	VMware Cloud Management Platform (vSphere, vRealize 등)	- x86 호환 컴퓨터를 위한 가상화 SW 공급 - 크로스 클라우드 아키텍처 개발 - 이기종 하이브리드 클라우드 관리 솔루션
Kognitio	Analytical Platform	- In-memory DB 개발 기업

표 5 빅데이터 어널리틱스 중심 기업과 주요 특징

기업명	주요 제품	주요 특징
Teradata	AoTAs(Analytics of Things Accelerators), Aster Analytics, Teradata Everywhere, Borderless Analytics	- 빅데이터 및 DW 관련 어널리틱스 솔루션과 컨설팅 서비스 제공(1979년 설립) - 아키텍처 컨설팅 등의 전략 제시
GoodData	Enterprise Insights Platform-as-a-Service	- 클라우드 컴퓨팅 기반의 빅데이터 어널리틱스와 BI 지향 - 기업용 BI 플랫폼-as-a-서비스 시장에서 강점
Datameer	DAS (Datameer Analytics Solutions)	- 빅데이터 분석 및 시각화 회사 - DAS는 스프레드 시트 인터페이스 제공 - DAS는 주요 하둡 배포판에서 사용 가능
1010data	Insight Platform	- 빅데이터 분석기관 - 다양한 종류의 데이터 처리와 분석을 위한 플랫폼 제공 - 머신러닝 등 개선된 빌트인 분석 기능 포함
구글	BigQuery	- 페타 바이트급의 데이터 저장 및 분석용 클라우드 서비스 - 하둡, 스파크보다 상대적으로 사용이 용이
Splunk	SIEM (통합로그분석시스템)	- '통합로그분석시스템(SIEM)' SW 시장에서 두각을 보이는 비정형 데이터 분석 솔루션 기업 - 로그 데이터, 실시간 이벤트 데이터, 장비 데이터를 수집하고 다룰 수 있는 엔진을 제공

기업명	주요 제품	주요 특징
Mu Sigma	muOBI, muPDNA, muDSC 등	- 세계적으로 유명한 데이터 분석 전문 회사 - 기업들에게 전문적으로 '빅 데이터' 분석을 의뢰 받아 의사 결정 지원 시스템을 제공
Opera Solutions	Signal Hub	- 특정 시장에서의 문제 해결을 도와줄 수 있는 시그널을 수집해 모은 후 데이터 분석에 활용
Amdocs	Amdocs Real-Time Data Management Platform	- 고객경험 SW 솔루션 및 서비스 분야 기업 - 글로벌 통신사, 엔터테인먼트 및 미디어 서비스 공급사들을 대상으로 영업
Syncsort	DMX/DMPExpress, DMX-h 등	- 빅데이터의 데이터 통합, 처리, 정렬 SW 기업 - 하둡, MS 윈도우, UNIX, Linux에 저장된 데이터 모두 지원
Attivio	Active Intelligence Engine	- 모든 데이터를 단일화된 인덱싱을 통해 검색, 분석이 가능한 정보 액세스 플랫폼
FICO	Data Orchestrator, Decision Management Platform 등	- 데이터 어널리틱스 기업
Flytxt	NEON-dX	- 통신서비스 사업자에게 빅데이터 분석 솔루션 공급 - 인메모리 DB 사용
SGI	SGI UV	- 인메모리 고성능 데이터 분석 기술이 강점 - HPE가 인수(2016년)

토털솔루션 제공 Big Data Platform 기업

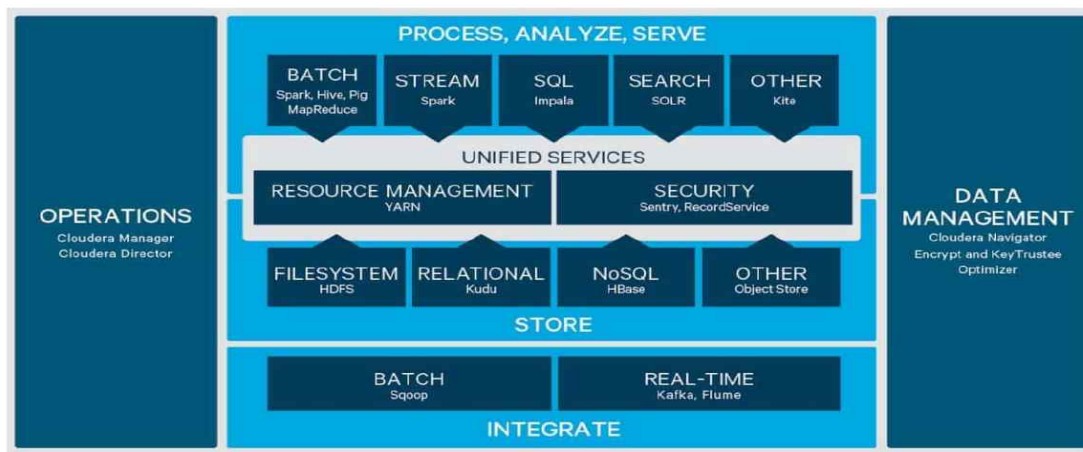
Amazon Web Service(AWS)

- 아마존닷컴이 제공하는 원격 클라우드 컴퓨팅 서비스(웹서비스)
- AWS는 스타트업, 엔터프라이즈, 공공영역 등에서 2백여 국가가 활용 중
 - ▶ 국내에서는 MBC, SK텔레콤, SK플래닛, SM엔터테인먼트, 르노삼성자동차, 직방 등 다수의 사업자가 AWS 활용
- 클라우드 초기 시장 성장시 AWS는 컴퓨팅, 스토리지, DB, 네트워크 및 콘텐츠 전송 등의 모든 영역에서 광범위한 제품을 보유하고 서비스 제공
 - ▶ 컴퓨팅 제품은 아마존 EC2, 아마존 Lightsail 등, 스토리지 제품은 아마존 S3, 아마존 Elastic File System 등과 DB에는 아마존 Aurora, 아마존 RDS 등
 - ▶ AWS 솔루션도 웹사이트, 백업 및 복구, 보안, 빅데이터, 금융서비스 고성능 컴퓨팅, 모바일 서비스, 사물인터넷, 게임, 과학 컴퓨팅 등 ICT의 전영역에 걸침
- 최근에는 음성인식 기반의 인공지능 비서 알렉사 서비스를 제공하며, 빅데이터, 유통, 인공지능 등의 전체 영역에서 자체적인 생태계를 구축
- 현재 아마존은 알렉사를 통해 유통매출을 확대하고, AWS를 연계해 고객들이 다양한 상품을 구매할 수 있는 환경을 제공하는 등 연계 전략 확대 중

● IBM Big Data 플랫폼

- IBM은 기업용 솔루션을 제공하는 업체로서 전통적인 강자이며, 빅데이터 플랫폼 뿐만 아니라 인공지능 왓슨을 발전시키고 향후 높은 역량이 기대됨
- IBM 빅데이터 플랫폼은 Data 관리 및 웨어하우스, Hadoop System, 스트림 컴퓨팅, 콘텐츠 관리, 정보 통합 및 거버넌스 등의 세부 플랫폼으로 구성
 - ▶ 세부 솔루션으로서 InfoSpheres BigInsights(텍스트 분석 모듈 및 데이터 탐색), Stream(다양한 유형의 데이터에 대한 실시간 분석), Netezza(빠른 데이터 처리), Smart Analytics System(분석 소프트웨어 등) 등을 개발 제공
- 그럼에도 IBM은 빅데이터 플랫폼에 의지하기 보다는, 자사가 개발 보유한 왓슨(Watson)을 기반으로 빅데이터와 인공지능을 결합한 빅데이터 분석을 표방하며 타사와는 다른 전략적 포지셔닝을 구축 중
 - ▶ 특히, 최근에는 IBM이 인공지능 왓슨의 대화, 번역, 성격 분석 등 6가지 기능을 무료로 기간제한없이 개방하여 표준 선점을 위한 노력 진행 중⁸⁾
 - ▶ IBM은 '16년 서로 다른 부서였던 왓슨 사업부와 클라우드 사업부를 통합

클라우드라 Big Data 플랫폼의 구성도



자료 : 클라우드라 홈페이지

● 마이크로소프트 애저(Microsoft Azure)

- Microsoft(MS)가 만든 클라우드 컴퓨팅 플랫폼 및 인프라스트럭처 서비스

8) 서울경제(2017.10.26.), “‘왓슨 무료개방’ IBM, 표준선점 승부수”

- ▶ 기업용 공용클라우드 시장은 IBM, 오라클, HP 등 IT업계의 다양한 거대 기업이 시장 도전을 하였으나, AWS의 역량에 못 미친 분야
- ▶ MS의 클라우드 사업은 '14년 취임한 사티아 나델라 최고경영자(CEO)가 '모바일 퍼스트, 클라우드 퍼스트'를 지향하며, 클라우드 중심의 사업 전략을 진행 중
- ▶ 애저 클라우드의 국내 고객으로 삼성전자, 서울아산병원, LG유플러스, 레드사하라 (게임업체), 영림원소프트랩(ERP 솔루션) 등 다양한 부문에서 활용 중
- 다양한 오픈소스를 지원하며, 데이터 어널리틱스, 기계학습 등 기능 탑재. 특히 Azure IoT Suite를 통해 IoT를 손쉽게 구현할 수 있는 것으로 평가
- ▶ MS는 R기반 분석 솔루션업체 'Revolution Analytics'를 인수했고 최근에는 'SQL on Linux'의 퍼블릭 프리뷰를 발표하는 등 빅데이터 영역에 대한 투자 확대 중
- ▶ MS는 리눅스 재단 가입 이전에도 Node.js, 오픈데이라이트, 오픈 컨테이너 이니셔티브, R 컨소시엄 등 리눅스 재단이 관리하는 프로젝트에 적극적으로 참여
- IaaS 퍼블릭 클라우드 시장에서 아마존이 '16년 약 44%의 점유율로 부동의 1위를 점하며, MS, 구글, 오라클, IBM, 알리바바 등이 추격하는 형세

● Oracle Bigdata 클라우드 서비스

- 가장 성공적인 데이터베이스 기업으로서, 자사의 역량에 기반하여 빅데이터 분석 지원을 위한 클라우드 플랫폼 개발 제공 중
- 오라클 빅데이터 플랫폼은 클라우드 서비스, 빅데이터 SQL 클라우드 서비스, 오라클 빅데이터 어널리틱스 솔루션 등 데이터의 수집부터 분석, 시각화까지 지원
- ▶ 어널리틱스 솔루션에는 Oracle R Advanced Analytics for Hadoop, Data Visualization 클라우드 서비스 등이 있음
- Oracle Data Visualization 클라우드 서비스는 한달에 75달러

● SAP HANA 클라우드 Platform

- 1972년 독일에서 설립된 세계 최대의 기업용 SW업체로서 주력 서비스는 ERP, CRM 등으로 차기 주력시장으로 빅데이터 분석을 지목
- '17년 빅데이터 분석 기술인 '빅데이터 허브(big data hub)'를 공개하고, 사물인터넷 분야의 선점을 위해 '레오나르도 센터' 3곳도 설립

- ▶ SAP은 '22년까지 사물인터넷 분야에 22억달러 이상을 투자할 계획이며, '16년 클라우드 분야에서만 30억유로의 매출을 기록해 1년 전보다 30% 이상 성장
- ▶ SAP은 빅데이터 기업으로 가기 위한 중간 단계로서 IoT를 선택하였으며, IoT 플랫폼인 레오나르도에는 머신러닝, 블록체인, 데이터인텔리전스, 빅데이터, 사물인터넷, 애널리틱스, 대화형AI 등 7가지 구성요소를 포함해서 구성
- SAP은 기업 변신을 위해 과감한 M&A전략과 개방형 혁신을 적극 활용
 - ▶ '11년 클라우드 기반 SW 업체인 석세스팩터스를 34억달러에 인수했고, '14년에는 비용 분석 소프트웨어 업체인 콘커를 83억달러에 인수하는 등 매년 많은 규모의 M&A 시행
- 현재 빅데이터 및 사물인터넷, 머신러닝을 활용할 수 있는 SAP 클라우드 플랫폼을 제공 중
- 클라우드 플랫폼은 S/4 HANA, SAP SuccessFactors, SAP Ariba, SAP Fieldglass 및 Concur 등의 SAP 핵심 제품군을 지원
- 특히 SAP S/4 HANA는 인메모리 컴퓨팅 플랫폼을 제공 중
 - ▶ SAP의 인메모리 데이터베이스 플랫폼기술인 '하나(HANA)'는 서울대 차상균 교수 연구팀이 개발한 원천기술을 활용해 '11.6월 출시

● Google Bigdata - Google Cloud Platform

- Google Cloud Platform은 Google의 데이터센터 인프라를 기반으로 컴퓨팅, 스토리지, 네트워킹, 빅데이터, 머신 러닝 등의 서비스를 제공하는 글로벌 클라우드
- Google Cloud의 특징으로는 빅데이터와 머신 러닝 서비스, 구글 전용 네트워크를 이용한 글로벌 커버리지, 저렴한 가격 모델 등임
 - ▶ 구글 클라우드는 타사에 비해 세부화되는 가격 정책으로 도입하여 분 단위 과금, 사용량에 따른 할인을 증가 등 적용

● Intel Bigdata

- 자사의 HW인 Xeon 프로세서, 10 Gigabit server adapters, SSD 등을 결합하고, SAS, Sap, 클라우데라 등과 함께 협력하여 빅데이터 솔루션 제공
 - ▶ 인텔은 자사의 x86 프로세서를 최적화시킨 자체 하둡 배포판을 개발 하였으나, 이를 중단하고 클라우데라 하둡 배포판을 지원

- '16년에는 인메모리 지원을 강화한 서버용 프로세서인 Xeon Process E7의 최신 제품을 선보이며, 실시간 분석에 필요한 성능을 제공하는 전략 추진

● Cisco CPA(Common Platform Architecture))

- Hadoop 기반의 빅데이터 어플라이언스 플랫폼으로서, 최적 시스템 설계와 솔루션도 제공
- CPA는 Cisco Unified Computing System(UCS) x86서버에 Hadoop을 통합한 어플라이언스로 Cloudera, Hortonworks, MapR, Intel Hadoop 등을 고객이 선택할 수 있도록 함
 - ▶ 여러 글로벌 IT업체들이 특정 하둡을 중심으로 어플라이언스를 제공하는 상황과 비교하면 CISCO는 상대적으로 개방적인 접근 선택
 - ▶뱅크오브아메리카(BOA)에서 CPA를 1천대 규모로 도입하기로 한 것으로 예상

● GE Industrial Internet(IoT 플랫폼 Predix)

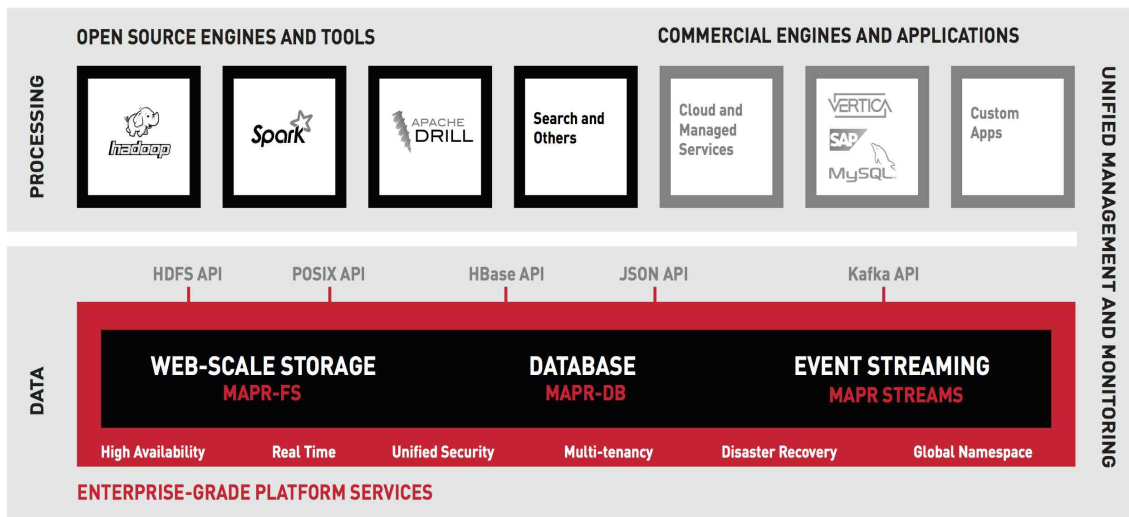
- Predix는 산업인터넷용 응용애플리케이션의 개발을 지원하는 분산형 테크놀로지 플랫폼
 - ▶ Predix는 센서를 통해 얻은 기기의 데이터를 모니터링하고, 분석하여 인간이 효율적으로 기기를 관리할 수 있도록 지원
- 현재 프레딕스 생태계에는 약 2만 명의 서드파티 개발자들이 포함되어 있으며, GE에 따르면 프레딕스 및 GE Digital 관련 소프트웨어 매출은 2016년에 20% 증가한 60억 달러에 달할 것으로 추정

● MapR Converged Data Platform

- MapR Data Technologies는 실리콘밸리의 엔터프라이즈급 SW 회사로서, 일괄처리, 대화형 또는 실시간 응용프로그램을 지원하는 광범위한 Apache 소스 생태계 프로젝트 패키지를 제공
 - ▶ MapR(맵알)은 클라우데라, 호튼웍스 등과 함께 글로벌 3대 하둡 전문기업으로 평가받는 회사
- 맵알의 Converged Data 플랫폼은 하둡(스파크 지원), NoSQL, 스트리밍 기능을 통합하여 구현한 것으로 기업 자체 평가

- 맵알은 IoT 시장에 대응하기 위해 엣지(Edge) 플랫폼의 강화 전략 추진
 - ▶ 엣지 플랫폼은 엣지단에서 생성되는 모든 데이터를 클라우드로 보내는 대신 로컬에서 컴퓨팅을 하는 개념으로서, 처리 결과만 센서에 저장해 보내는 개념. 즉, 엣지단에서 처리를 하고 중앙에서 관리만 하는 형태
 - ▶ 이를 위해 소형 데이터 플랫폼 ‘맵알 엣지(Edge)’를 기존 컨버지드 데이터 플랫폼에 추가(‘17.11)

MapR의 Converged Data Platform 구현 모델



자료 : MapR 홈페이지

● Cloudera Enterprise Bigdata

- 빅데이터 분석 전문 기업으로 클라우드 컴퓨팅 용어를 처음 만든 비시글리아를 중심으로 오라클 등의 전문가들이 설립
- 머신러닝과 첨단 분석 플랫폼을 제공하는 업체로, 교육 서비스와 기술지원 서비스에 주력하는 오픈소스 기반 빅데이터 플랫폼 업체로 포지셔닝
 - ▶ 자사의 강점으로는 머신러닝을 데이터 분석에 활용하는 측면에 차별성 강조
 - ▶ 금융회사뱅크오브아메리카는 클라우데라 하둡을 이용해 금융거래 내역 수십 억건을 분석, 금융 사기 방지 프로그램을 도입
- 주요 특징으로 Data Engineering, 분석 및 운영 데이터베이스, 보안 등이 있으며, Data Science를 위한 DSWB(Data Science WorkBench)를 보유

- ▶ Impala 기술을 통해 SQL을 하둡 환경에도 쓸 수 있도록 만들었으며, 이를 기반으로 DW나 비즈니스분석 시장에 침투
- ▶ DSWB는 데이터 과학자들이 하둡 클러스터 위에서 머닝러닝을 포함한 여러 분석을 수행할 수 있는 플랫폼

● Hortonworks Data Platform(HDP)

- '11년 Yahoo! 하둡 개발 및 운영 팀 엔지니어를 주축으로 설립된 하둡 상용 솔루션 업체
 - ▶ 삼성전자 모바일솔루션센터(MSC)에서 빅데이터 분석을 위해 호튼웍스 하둡 배포판을 사용하면서 국내에서 유명세를 얻음
- HDP는 호튼웍스(Hortonworks)가 분산인프라용 데이터 처리 플랫폼으로 소개한 것으로, 아파치 하둡의 HDFS, 맵리듀스, 피그, 하이브, HBase, 주키퍼 등을 포함하는 오픈소스 기반 플랫폼

● DataStax Bigdata

- 클라우드 애플리케이션용 데이터베이스 SW 전문업체로서, 오픈소스 SW인 아파치 카산드라의 상용화 버전을 개발하는 회사

● Informatica PowerCenter Big Data Edition

- 데이터 통합 · 관리 SW업체로서, 시각적 개발 환경을 사용하여 Hadoop에서 기본적으로 실행되는 ETL 데이터를 구축하는 기업용 데이터 통합 SW

● VMware

- EMC 코퍼레이션의 공식 자회사이며, VM웨어 워크스테이션과 프리웨어 제품인 VM웨어 서버, VM웨어 플레이어를 포함한, x86 호환 컴퓨터를 위한 가상화 소프트웨어를 공급하고 있음. 1998년 설립
- VM웨어는 하이브리드 클라우드 전략을 세워, 고객 기업이 AWS나 MS의 애저, IBM 클라우드 등의 특정 클라우드에 종속되지 않고 종류와 상관없이 쓸 수 있도록 할 수 있는 크로스 클라우드 아키텍처를 개발
- VMware Cloud Management Platform은 이기종 하이브리드 클라우드를 관리하기 위한 솔루션을 제공

- ▶ vSphere는 가상화 계층으로서 서버 HW를 가상화하고 다종의 가상화 머신이 공유할 수 있게끔 지원
- ▶ vRealize 클라우드 관리 플랫폼은 하이브리드 환경에서 애플리케이션에서부터 전체 인프라 스트럭처(컴퓨팅, 스토리지 및 네트워크)까지 지능형 운영 제공

● Kognitio Analytical Platform

- Kognitio는 데이터 과학 및 비즈니스 인텔리전스를 위한 고성능, 확장 가능한 In-memory DB 개발 기업
- ▶ 대기 시간이 짧은 대용량 데이터로드 및 처리량이 많은 복잡한 분석 작업 부하에 맞게 최적화된 스케일 아웃(scale-out) 메모리 내 대량 병렬 처리(MPP) 소프트웨어 기술

어널리틱스 중심의 빅데이터 플랫폼

● Teradata Bigdata Analytics 플랫폼

- 전세계적으로 빅데이터 및 데이터 웨어하우스 관련 Analytic 솔루션과 컨설팅 서비스를 제공하는 글로벌 기업
- ▶ 1979년 최초 설립. 데이터웨어하우징 및 기업 인텔리전스 분야, DW/CRM 관련 플랫폼과 산업별 특화된 솔루션과 포괄적 전문 서비스에 역량 보유
- ▶ '11년 애스터(Aster), Revelytix(2014), Hadapt(2014) 등의 인수합병을 진행하며 빅데이터 역량 강화 중
- 최근에는 새로운 포트폴리오로 비즈니스 솔루션, 아키텍처 컨설팅, 기술 솔루션으로 구성된 새로운 전략 제시
- ▶ 주요 솔루션으로는 AoTAs(Analytics of Things Accelerators), Aster Analytics, Teradata Everywhere, Borderless Analytics 등을 선보임

● GoodData

- 클라우드 컴퓨팅 기반의 빅데이터 어널리틱스와 BI를 지향하는 플랫폼 개발하는 기업으로 2007년 설립
- 자료의 수집, 저장, 연계, 분석, 시각화 등 비즈니스 인텔리전스(BI)의 모든 핵심 솔루션을 만들 수 있는 툴, API, 프레임워크 툴로 구성
- Forester에 의하면 GoodData는 기업용 BI 플랫폼 시장에서 10위 이내의 역량있는 기업으로 평가

● Datameer

- 빅데이터 분석 및 시각화 회사로서 빅데이터 어널리틱스 플랫폼을 제공
- 이 회사의 제품인 DAS(Datameer Analytics Solution)은 데이터 소스 통합, 200개가 넘는 분석 기능을 갖춘 스프레드 시트 인터페이스가 제공되며, 아파치, 클라우데라, 맵알, 아마존 등 주요 하둡 배포판에서 사용 가능

● 1010data

- 뉴욕에 기반한 빅데이터 분석기관
- 정형 및 비정형 등 다양한 종류의 데이터와 분석을 위한 툴을 하나의 플랫폼으로 제공
- 플랫폼에는 개선된 빌트인 분석 기능을 포함
 - ▶ 예로서 단순통계량 계산(분포 분석, 상관계수, 분산), 예측 모델링 및 Predictive modeling and forecasting (선형 및 다변량 회귀분석, 로지스틱 회귀분석), 머신러닝(클러스터링 분석, 몬테카를로 마코프 체인, 주성분 분석) 등을 포함

● Google BigQuery

- 페타 바이트급의 데이터를 저장하고 분석하는 클라우드 서비스
 - ▶ 구글의 빅쿼리는 Dremel이라는 논문을 근간으로 구현
- BigQuery의 특징은 클라우드 서비스이며(NoOps), 기존의 RDBMS에서 쓰는 SQL과 유사한 언어를 사용하고, 대용량 지원과 빠른 성능, 데이터 복제를 통한 안정성, 배치와 스트리밍 모두 지원, 저렴한 가격임
 - ▶ 8,800개의 CPU와 3,600개의 디스크를 사용하는 대규모 인프라를 활용하여 1,000억 개의 레코드에 대한 질의를 30초 정도에 수행
 - ▶ Hadoop, Spark 등은 Map&Reduce 로직이나 SparkSQL을 사용하기 위해선 일정 수준 이상의 전문성이 필요하지만, BigQuery는 상대적으로 실질적 사용과 유지보수 등이 매우 간편하다는 장점을 가지는 것으로 평가

● Splunk

- 2006년 스타트업으로 시작하였으며, '통합로그분석시스템(SIEM)' SW 시장에서 두각을 보이는 비정형 데이터 분석 솔루션 기업

- ▶ 로그 데이터, 이벤트 데이터 등의 기계가 생성한 빅데이터를 웹 기반 인터페이스를 통해 모니터링 및 분석
- 나스닥 IPO 이후 국내에도 지사를 설치해 안랩·SK인포섹 등을 포함하여 글로벌한 사업을 진행 중
 - ▶ '17년 2분기 매출은 작년 같은 기간보다 30% 늘어나고 올 한해 12억달러(1조 3500억원)의 매출을 기대
 - ▶ 시만텍, 시스코, IBM 등이 인수를 고려하거나 고려한 바 있음

● Mu Sigma⁹⁾

- '05년 설립되었으며, 세계적으로 가장 유명한 데이터 분석 전문 회사로서 “데이터 계의 마스터(Master of data universe)”라고 불림
 - ▶ 2016년 매출은 1.65억달러로서, 2015년의 1.84억달러에 비해 다소 감소
- 뮤 시그마는 기업들에게 전문적으로 ‘빅 데이터’ 분석을 의뢰 받아 의사 결정 지원 시스템을 제공
 - ▶ 마이크로소프트는 뮤 시그마의 첫 고객으로 2005년에 실험 연구에 들어가 MSN 메신저와 검색엔진 Bing의 개발에 직접적인 정보를 제공
- 2011년 12월 28일, 세계주식 투자기업 제너럴 아틀란틱(general atlantic)과 IT전문 벤처펀드 세쿼이어캐피탈(Sequoia capital)은 뮤 시그마에게 업계 사상 최고 금액인 1억 800만 달러를 (1,200억원)을 투자

● Opera Solutions Bigdata

- 데이터 분석 전문 기업으로서, 분석 플랫폼인 ‘Signal Hub’ 등 보유
 - ▶ 시그널에 대해 “문제를 해결할 수 있는 유의미한 패턴을 빅데이터로 부터 찾아내는 것을 의미”한다고 정의¹⁰⁾
 - ▶ Signal Hub는 insight를 추출하기 위해 기계 학습을 활용

● Amdocs Insight

- 고객경험 SW 솔루션·서비스 분야 기업으로 통신사와 엔터테인먼트, 미디어 서비스 공급사들을 고객으로 비즈니스 진행 중

9) 빅데이터 해외동향(<https://sites.google.com/a/chosunbiz.com/bigdata/news?offset=10>) 자료를 발췌 정리

10) <https://blog.naver.com/kosena21/40186323485>

- ▶ 앰독스는 약 2만4천여 소속직원을 보유하고 있으며, NASDAQ Global Select Market에 상장했으며 2015 회계연도에 36억 달러의 수익을 기록

● Syncsort

- Syncsort는 Big Data, 초고속 정렬 제품, 데이터 통합과 서비스를 제공하는 SW 회사로서, Hadoop, MS windows, UNIX, Linux를 모두 지원
- Hadoop, cloud 및 Splunk와 같은 차세대 분석 플랫폼을 모두 지원하며, 각 플랫폼에 특화된 솔루션을 제공

● Attivio Active Intelligence Engine

- Attivio의 Active Intelligence Engine은 Hadoop, Big Data, Big Content를 결합하는 단일화된 정보 액세스 플랫폼으로서, Universal indexing과 자동 ad hoc JOIN을 통한 정보의 자동 연결, 시각화 툴을 보유

● FICO Big Data Analyzer

- Big Data Analytics SW 솔루션인 Predictive Analytics and Business Intelligence 툴을 제공
- 툴에는 Data Orchestrator, Decision Management Platform, Decision Optimizer, Model Builder, Model Central Solution, Predictive Analytics, Solution Stack으로 구성하여 데이터의 이해, 쿼리, 탐색, 시각화 등을 제공

● Flytxt

- Flytxt는 전 세계 CSP와 모바일 업체들을 위한 빅 데이터 분석 솔루션을 제공하는 업체. 2008년 설립. 본사는 네덜란드
 - ▶ 2015년 상반기에는 다음 성장 단계에 필요한 자금을 지원할 1천만 유로 규모의 자금조달 라운딩을 마친 것으로 발표
- Flytxt의 Big Data Analytics platform은 전통적인 RDBMS를 메타데이터로 하여 Hadoop에서 활용되며, 실시간 거래 데이터의 처리를 위해 인메모리 DB를 사용

● SGI Bigdata

- SGI는 최초 1982년 훗날 넷스케이프를 설립한 짐 클록이 창업
 - ▶ 본래 Silicon Graphics라고 불렸지만 파산 후 2009년에 Rackable Systems에 인수되면서 이름을 변경
 - ▶ SGI는 전 세계적으로 약 1,100명의 직원을 두고 있으며, 2015년 회계연도 실적은 매출 5억 3,300만 달러, 순손실 1,100달러를 기록
 - ▶ 2016년 HPE가 SGI를 2억 7,500만 달러에 인수
 - ▶ HPE는 빅데이터 분석 사업 성장 뿐만 아니라, 공공과 생명과학, 고등교육과 연구, 제조, 슈퍼컴퓨팅 등 업종별 시장에서도 HPE의 영향력을 강화 예상
- SGI UV는 실시간 분석을 위한 공유 메모리 플랫폼으로서, SGI는 특히 인메모리 고성능 데이터 분석 기술이 높은 것으로 알려짐

<참고> 성공하지 못한 플랫폼

- 모바일 서비스, 소셜네트워크서비스, 앱 환경이 성장하면서 플랫폼의 가장 중요한 요소로서 모든 영역의 플레이어 참여가 매우 중요한 전략적 요소로 부상
 - 플랫폼 등장 초기에는 마이크로소프트사의 Dos, Window 등과 같이 OS SW를 플랫폼이라고 인식됨
 - 개방·공유·참여의 키워드 아래 Web2.0의 개념이 발전하고, 특히 모바일 서비스의 발전과 App 생태계는 플랫폼의 개념을 발전시킴
 - ▶ Web 2.0 : 개방, 참여, 공유의 정신을 바탕으로 사용자가 직접 정보를 생산하여 쌍방향으로 소통하는 웹 기술(위키백과)
- 플랫폼에 기반한 생태계나 산업·경쟁 구조는 이미 현재에는 정착이 된 모델로서, 구글, 애플, 페이스북, 아마존 등이 글로벌 시장을 선도
- 인터넷 보급이 진행되던 초반에 국내외적으로 성공하거나 비즈니스 모범 사례로 분석되던 서비스들이 국내에서도 다수 있었으며, 특히 세계에서 가장 먼저 진행했던 서비스도 존재
 - 이들 서비스는 이미 대규모의 사용자 기반을 보유한 플랫폼 형태로서 존재하였음에도, 부침을 겪으며 실패하거나 명맥만 유지하는 상황이 발생
 - Cyworld, Myspace, WIPI 등은 많은 가입자 기반이 있었음에도 추가적인 성장을 하지 못하고 시장에서 소멸
 - 이외에 소셜네트워크 서비스 기반으로 마케팅을 전개하고자 한 소셜 분석 플랫폼은 많은 경우 비용대비 성과 확보에 실패한 것으로 평가됨
- 싸이월드(Cyworld)¹¹⁾ - 비운의 국내 신화
 - 아이폰과 페이스북 이전에 2000년대를 상징하던 한국판 SNS로서 사용자가 3,200만에 이를 만큼 큰 성장을 기록
 - ▶ 하나의 서비스에 국민의 절반 이상이 가입하였다는 측면에서 현재의 플랫폼 기반보다 훨씬 큰 영향력을 가졌을 것으로 판단
 - ▶ 특히 사용자들의 행태를 이해할 수 있는 다양한 데이터의 축적이 가능했을

것으로 판단됨

- 국내에서 싸이월드의 성공요인에 대해서 시뮬레이션하는 한국 특유의 緣 문화, 도토리, 초고속인터넷 보급 등을 근본 요인으로 제시¹²⁾

- ▶ ‘聯’을 중시하는 한국 특유의 문화를 기반으로 ‘일촌’을 도입한 것이 핵심 요인. ‘일촌’ 시스템은 한국의 ‘촌수’를 흉내 낸 것을 인터넷 상에 구현
- ▶ ‘도토리’를 통하여 개성있는 홈페이지 구현이 가능
- ▶ 서비스 당시 초고속 인터넷과 디지털 카메라 보급으로 사진 업로드가 빠르게 가능
- ▶ 싸이월드 브랜드는 ‘싸이’ = ‘사이버’ = ‘사이’라는 의미를 전달. 브랜드명도 ‘사이 좋은 월드’라는 의미로 SNS를 아주 잘 설명해주고 있음

- 그럼에도 싸이월드는 플랫폼의 모바일 전환 실패, 해외 현지화 전략 실패, 플랫폼의 폐쇄성 측면에서 전략적 실패를 겪으며 소멸 초래

- ▶ ‘04년 모바일로 플랫폼의 체질 변화를 시도하였으나, PC상의 UX를 모바일로 단순 전환하고, 꾸미기 기능도 모바일에서 제대로 구현을 하지 못함
- ▶ ‘06년부터 중국과 일본에 서비스를 확장하였으나, 현지화에는 실패. 국내의 성공 요인이 오히려 외국에서는 실패요인으로 작용. 즉, 희박한 일촌 개념, 다소 부정적인 Cyber(=싸이) 언어, 불완전한 지불/결제 시스템, 느린 인터넷 등이 원인
- ▶ ‘03년 SK컴즈에 인수되며, 도토리, 아바타, 음원 등의 모든 영역을 SK컴즈가 주도하는 폐쇄적인 플랫폼 운영으로 시장에서 외면을 받음
- ▶ 또한, 해외 진출시 미국은 페이스북, 마이스페이스, 일본은 mixi 등 이미 현지 SNS 서비스들이 자국내 시장을 선점하였으며, 도토리는 사치로 인식됨

- 싸이월드의 실패로 인해 소위 데이터 이동성에 대한 논의가 촉발

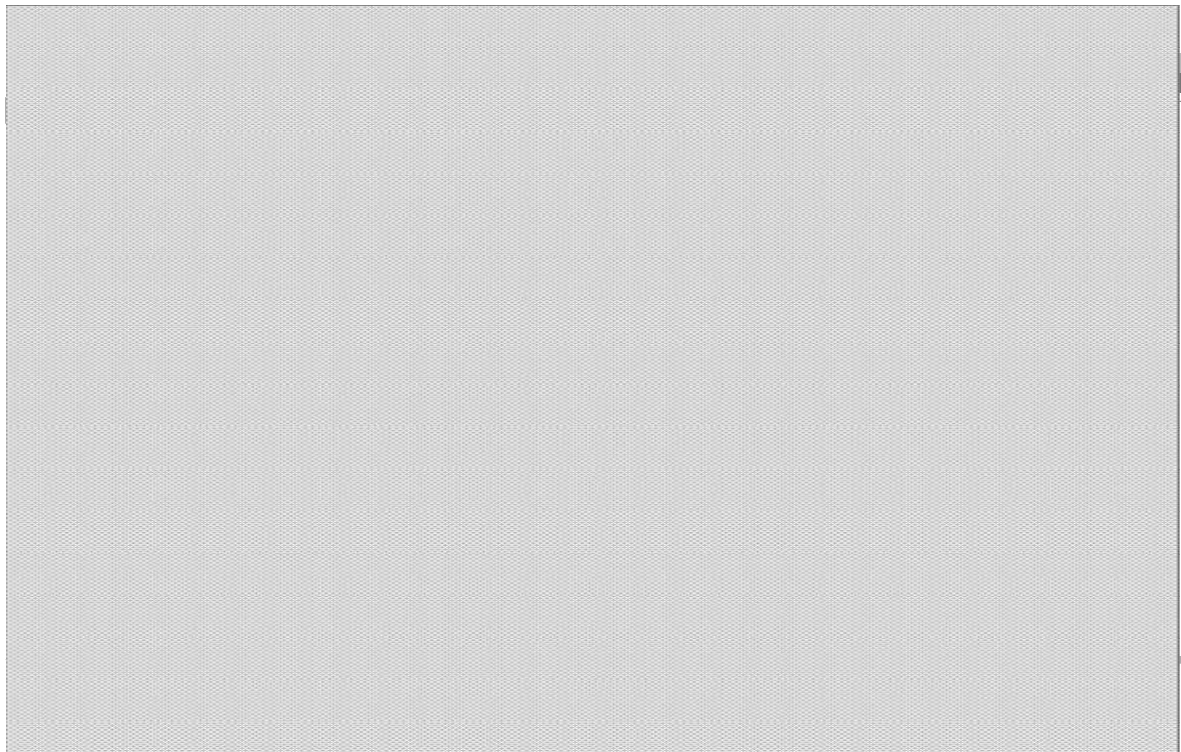
● 마이스페이스(MySpace)¹³⁾ : 소셜네트워크 서비스의 할아버지

- ‘03년 미국 캘리포니아에서 시작되어 ‘08년까지 2억 명 이상의 사용자를 확보하였으며, 페이스북을 한때 압도한 SNS 플랫폼으로서 음악 등 콘텐츠와 결합된 SNS라는 장점을 가짐

- ▶ ‘08년 4월 SNS 선두 자리를 페이스북에 내줌

- 페이스북에 뒤쳐진 주요 원인으로 ‘05년 뉴스코퍼레이션에 인수됨에 따라 서비스가 안정화되기 전에 스타트업으로서의 정신으로 유지하지 못한 점이 지적됨

- ▶ 페이스북은 인간 본연의 관계를 엮는 SNS로 차별화하였으며, 마이스페이스에 비해 빠른 UX와 오픈 생태계를 적극 활용
- ▶ '05년 루퍼트 머독이 소유한 뉴스코퍼레이션이 5억8,000만 달러에 마이스페이스를 인수. 마이스페이스 직원들은 매출 목표액 달성에 집중하는 상황 발생. 스타트업의 정신이 관료주의에 의해 변형되었다는 평가
- 전략 측면에서 마이스페이스는 경쟁사들의 빠른 움직임을 따라가지 못했으며, 수익 중심 비즈니스에만 너무 몰두하는 경향이 보임
- ▶ 스마트폰이 확산되는 시점에서 마이스페이스는 주로 UI 디자인의 파격적인 교체, 무거운 배너 광고 등에만 집중하면서 모바일로의 전환에 실패
- ▶ 이에 반해 페이스북은 모바일과 경쟁사에 대한 빠른 기술 대응과 최소한의 광고, 팜빌가 같은 소셜 게임을 제공하는 등 고객의 행동패턴과 개방성을 강조하며 자사의 서비스에 반영



- 단, 콘텐츠-사람간 VS 사람-사람간의 연결성 차이가 두 SNS의 성패를 좌우했는지에 대해서는 명확한 결론을 내지 못하고 있음

● 아이러브스쿨 - 서비스는 성공, 협력은 실패

- 학연으로 기반으로 1999년 서비스되면서 1년 만에 회원 500만명을 확보
- '친구 찾기 열풍'을 선도하며 급격한 회원 수 증가 경험

- 야후코리아에서 500억원에 인수 제의도 받았으나, 결국 타 기업 매각에 실패하며 소멸
- 아이러브스쿨은 국내 실정에 적합한 '학연'을 기반으로 빠른 성장세를 보였으나, 전문성과 장기 전략성, 나아가 시장에서 플랫폼을 갖춘 기업과의 연계가 중요함을 지적하는 사례로 판단됨
- 데이터 혹은 서비스 측면에서는 서버 확충, 수익 모델 등에 대한 고민을 면밀히 하지 못한 상태에서 양질의 서비스 제공을 위한 투자에 집중
- 아이러브스쿨은 당시 오프라인의 학연 데이터를 온라인으로 변화시키고 관리해 줄 수 있는 관계 혹은 데이터 중심의 서비스였지만, 서비스의 중심을 사업 자체에 집중하며 데이터 기반 전략성 파악에 실패하였다는 평가

● Quirky - 수익 창출 구도가 미약

- 마음에 드는 제품 또는 서비스 아이디어들을 투표로 올려 이 회사가 직접 제품으로 만드는 서비스를 제공하는 발명품 플랫폼
- Andreessen Horowitz, GE, RRE 벤처 등으로부터 총 1억8,500만 달러를 투자 받으며 유망한 스타트업으로 주목받음
- 그렇지만, 많은 Quirky 제품들의 마진 폭이 너무 작아 수익 창출에 실패하고 2015년 9월 부도 신청
 - ▶ Quirky가 개발했던 블루투스 스피커는 40만 달러의 비용이 들어갔는데 단 28대만 팔려나감
 - ▶ Wink 스마트홈 시스템도 보안 업데이트가 잘 안되면서 리콜을 부르면서 사업을 접고, 150만 달러에 매각

● 노키아

- 피쳐폰 시장에서 세계 1위를 오랜동안 점유하고 있었으나, 스마트폰으로 휴대폰 시장이 변화되며 소위 플랫폼 경제에 편입되지 못하고 쇠퇴
- 스마트폰용 OS인 심비안과 미고를 개발하기 시작하였으나, 이미 iOS와 안드로이드가 모바일폰용 OS 시장을 선점
- MS와 협력하여 윈도우OS를 채택하는 전략을 추진하였으나 여전히 성장이 지연되며 '13년 MS에 노키아의 단말기 사업 부문을 MS에 매각

11) 이코노믹리뷰, “살아남는 자가가 강한자다”.

- 노키아의 스마트폰 플랫폼 구축 실패의 원인은 미고(MeeGo)라는 기술적으로 우수한 OS가 있었지만, 이의 확산 전략에서 의사결정 지연, 경쟁 구도의 냉철한 분석 미흡, 협력 관계 구축의 불확실성에서 기인

● WIPI(wireless internet platform for interoperability)

- 우리나라의 표준 모바일 플랫폼. 모바일 플랫폼을 표준화하여, 하나의 콘텐츠를 여러 통신사에 서비스 할 수 있도록 제정됨
 - ▶ 이때 해외 시장은 이미 BREW와 J2ME 플랫폼을 기반으로 서비스 제공되었으며, 국내에서는 이동통신사 주관의 다수 플랫폼이 활용되던 중
- 위피는 국내시장 보호라는 순기능도 있었으나, 외산 휴대폰의 국내 시장 진입에의 무역장벽으로 작용하여 소비자 선택권을 침해하였다는 평가도 있음
 - ▶ 이통 3사간 콘텐츠 호환은 여전히 10% 수준으로 미진했으며, WIPI가 썬마이크로시스템즈의 특허를 상당부분 침해하여 로열티를 지불하는 상황 발생
- '05.4월 국내 판매 이동통신 휴대 단말에는 모두 WIPI가 의무적으로 탑재하기로 했으나, '09.4월까지 시행된 후 해제
- WIPI는 천만대 이상의 국내 휴대 단말에 탑재된 플랫폼이었으나, 결국 CP와의 연계를 플랫폼 생태계로 발전시키지 못함

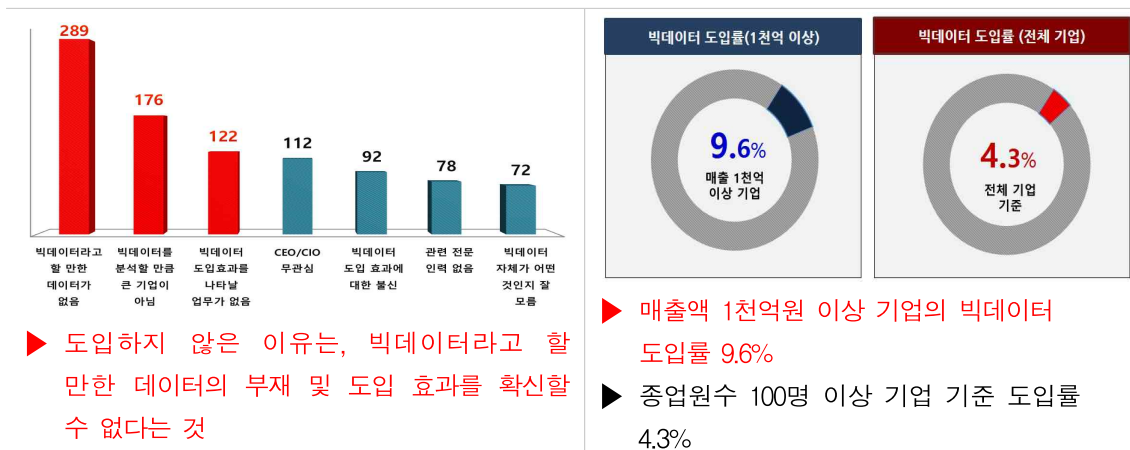
<https://wonning.wordpress.com/2015/12/02/싸이월드의-세계화-진출-실패-1탄> , & 2탄,
http://www.ohmynews.com/NWS_Web/View/at_pg.aspx?CNTN_CD=A0002154415 등 참조하여 구성
 12) 시몽뵈로(2010), “글로벌 비즈니스 마인드 세트”
 13) 이코노믹리뷰, “살아남는 자가가 강한자다”, <http://jeremy68.tistory.com/304> 의 내용을 참조하여 정리

III 빅데이터 플랫폼 이슈 분석

빅데이터 - 신뢰성 있는 데이터의 확보

- (빅데이터의 부재) 가장 큰 난제로는 빅데이터가 없다는 측면으로 활용하고자 하지만, 활용할 수 있는 데이터가 없다는 점이 가장 많이 지적됨
 - 데이터가 없다는 점은 기술적인 측면만으로 극복하기 힘든 사항이며, 아울러 데이터에 기반한 의사결정 자체가 이루어지지 못하는 상황 연출
 - ▶ 빅데이터를 도입하지 않은 국내기업들이 지적한 가장 큰 요인은 빅데이터라고 할 만한 데이터가 없어, 빅데이터 구축 기반의 형성이 어렵다는 의견
 - ▶ 국내 전체기업(종업원 100인 이상)의 빅데이터 도입률은 2014년 3.9%에서 2015년 4.3%로 상승. 특히, 매출액 1천억원 이상 기업의 도입률은 2014년 8.1%에서 2015년 9.6%로 확장 추세
 - ▶ 빅데이터 비도입 이유는 빅데이터의 부재, 작은 기업 규모, 도입 효과의 확산 등으로 조사됨

국내 기업의 빅데이터 활용 현황



자료 : 과학기술정보통신부(2016.1), “2015년 빅데이터 시장현황 조사결과 발표”

- 빅데이터 혹은 빅데이터 플랫폼이 의미를 가지기 위해서는 데이터의 수집과 저장, 공유는 우선적으로 진행되어야 할 절차

● (빅데이터의 품질) 다양한 형태와 유형의 데이터 증가는 빅데이터의 품질 문제 및 데이터 자체의 편의 문제 유발

- 음성, 이미지, 텍스트, 동영상, 댓글, 웹로그데이터 등의 비정형 데이터가 증가되면서 분석할 수 없거나 힘든 데이터가 구축될 가능성 존재
- 빅데이터를 포함하여 불량한 데이터는 기업들에게 의미없는 지출을 발생시키며, 잘못된 결정으로 미래 비즈니스 전략에 영향을 주기도 함
 - ▶ TDWI¹⁴⁾의 2002년 리포트에는 데이터의 품질에 대한 인식 수준과 실질적인 품질에는 커다란 gap이 존재하며, 데이터 품질 문제로 인해 미국 비즈니스는 1년에 6,000억불 이상의 비용이 발생하는 것으로 분석. 본 보고서가 2002년 자료를 고려했을 때, 현재는 관련 비용을 더욱 더 커졌을 것으로 판단됨
 - ▶ IBM은 2016년 낮은 품질의 데이터로 인한 미국의 지출 비용을 3.1조(Trillion) 달러로 추정¹⁵⁾
- 특히, 소셜 데이터는 일반인의 평균적인 정보가 아닌 Heavy User에 의한 대가성 참여로 인한 평가 및 리뷰 왜곡, 불명확한 출처를 가진 데이터로 인해 분석의 결과도 편향되게 만들 가능성 존재

● 국가별로 데이터를 가장 많이 수집하고 저장하는 장소는 정부나 공공기관

- 공공데이터는 공개될 경우 투명한 국정 운영, 국민의 알권리 제고, 경제적 효과가 있는 것으로 평가되면서 공개에 대한 국민적 요구와 정책적 노력이 지속적으로 진행 중
 - ▶ 공공데이터는 정부가 생산하고 보유하고 관리하는 모든 데이터를 의미
 - ▶ 오픈데이터 정책은 공공데이터를 누구든지 활용하고 가치와 수익을 창출할 수 있도록 개방하는 것
 - ▶ World Bank의 보고서에 따르면 정부 및 공공기관에서 제공하는 오픈데이터가 개인 및 비영리 부문에 대한 경제적 잠재력을 갖고 있다고 제시
- 산업적으로는 공공부문에서 발생하는 수많은 데이터를 기업, 시민들이 접근하고 이용하면서 새로운 서비스도 생산
 - ▶ '09년 국내 한 고등학생이 각 지역의 버스 데이터를 활용하여 버스 도착시간을 실시간으로 알려주는 앱 서비스를 만들어 큰 인기를 끌. 그러나, 정부에서는 공공데이터를 무단 활용했다는 사유로 운영을 잠시 중단시킨 사례가 있음

14) TDWI : Data Quality and the Bottom Line: Achieving Business Success through a Commitment to High Quality Data)

15) HBR(2016.9.22.), "Bad Data Costs the U.S. \$3 Trillion Per Year")

- 미국, 영국 등의 선진국에서는 공공정보 공유와 정보 활용을 위한 웹사이트를 구축하여 제공하는 중
- 국내에서도 2013년 소위 “공공데이터법”을 제정하여 공공데이터 개방을 확대하고 있는 추세
- 세계 여러 나라의 공공데이터 개방 현황에 평가 지표인 ODB(Open Data Barometer)에 의하면 2016년 한국은 호주와 함께 평가국 중 5위
 - ▶ ODB는 월드와이드웹 재단(World Wide Web Foundation)에서 조사하여 발표
 - ▶ ODB 보고서에서는 오픈데이터는 접근성(accessible), 기계 판독성(machine readable), 재활용성(re-usable) 등의 특징을 가진 것을 지칭
 - ▶ ODB 지표는 준비도(Readiness), 이행도(Implementation), 영향력(Impacts)로 구성되며, 세 지표의 점수를 통해 국가 순위 도출
 - ▶ 국내 점수는 준비도와 영향력에서는 95, 100으로 매우 높으나, 이행도 측면에서는 매우 낮은 수준

ODB의 국가별 순위

Open Data Barometer (2016)

0 25 50 75 100

The Open Data Barometer
A global measure of how governments are publishing and using open data for accountability, innovation and social impact.

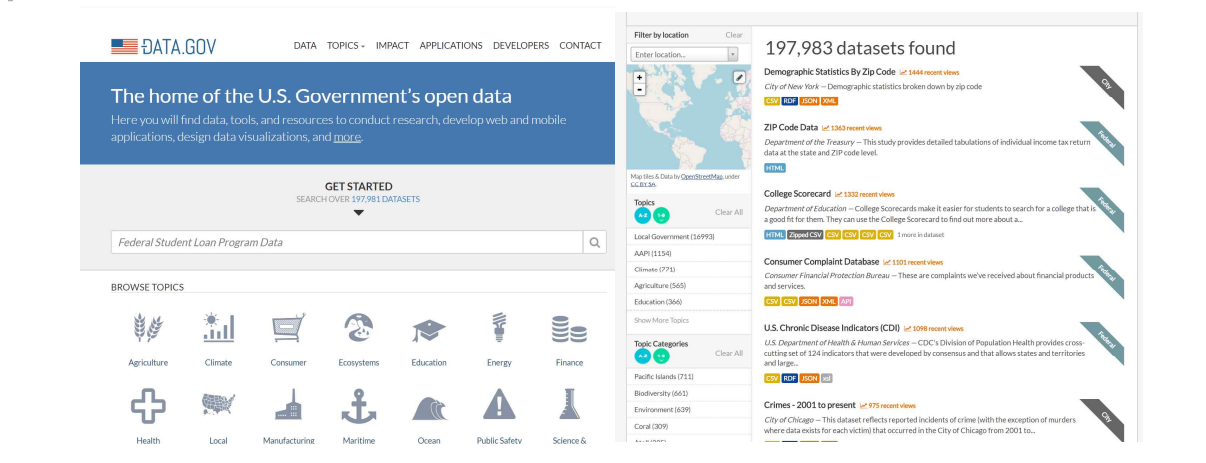
Country	Rank ?	Score ? OUT OF 100	Change ?	Score Trend ? OVER PAST EDITIONS	Readiness ? OUT OF 100	Implementation ? OUT OF 100	Emerging Impact ? OUT OF 100
United Kingdom See details	1	100	0 -		99	100	94
Canada See details	2	90	2 ▲		96	87	82
France See details	3	85	-1 ▼		100	71	88
United States of America See details	4	82	-2 ▼		96	71	80
Korea See details	5	81	3 ▲		95	59	100
Australia See details	5	81	5 ▲		85	78	78
New Zealand See details	7	79	-1 ▼		92	58	99
Japan See details	8	75	5 ▲		84	60	89
Netherlands See details	8	75	-1 ▼		94	64	68
Norway See details	10	74	7 ▲		77	71	73

출처 : http://opendatabarometer.org/?_year=2016&indicator=ODB

● 미국의 공공데이터

- 2009년 “열린정부 이니셔티브”를 기반으로 공공데이터 포털을 구축
 - ▶ data.gov로서 ‘17년 현재 20만여 개의 데이터셋을 제공하고 있으며, 데이터셋의 31.9%가 CSV, XML 등의 오픈 포맷의 데이터를 제공
 - ▶ 2014년 5월 “오픈데이터 액션 플랜”을 마련하고, 미국 각 주의 데이터 개방 수준 평가(1위 하와이, 2위 일리노이 순)
- Data.gov의 백엔드 시스템으로서 ‘오픈 거버먼트 플랫폼’(OGPL)을 오픈소스로 공개('13.1)
 - ▶ OGP의 핵심가치는 정부의 투명성확보와 시민의 참여
 - ▶ OGPL 소스코드는 GitHub에서 공개되었으며, 인도, 가나 등에서는 OGPL을 이용하여 자국의 공공 데이터플랫폼을 구축 중

미국의 공공데이터포털인 Data.gov

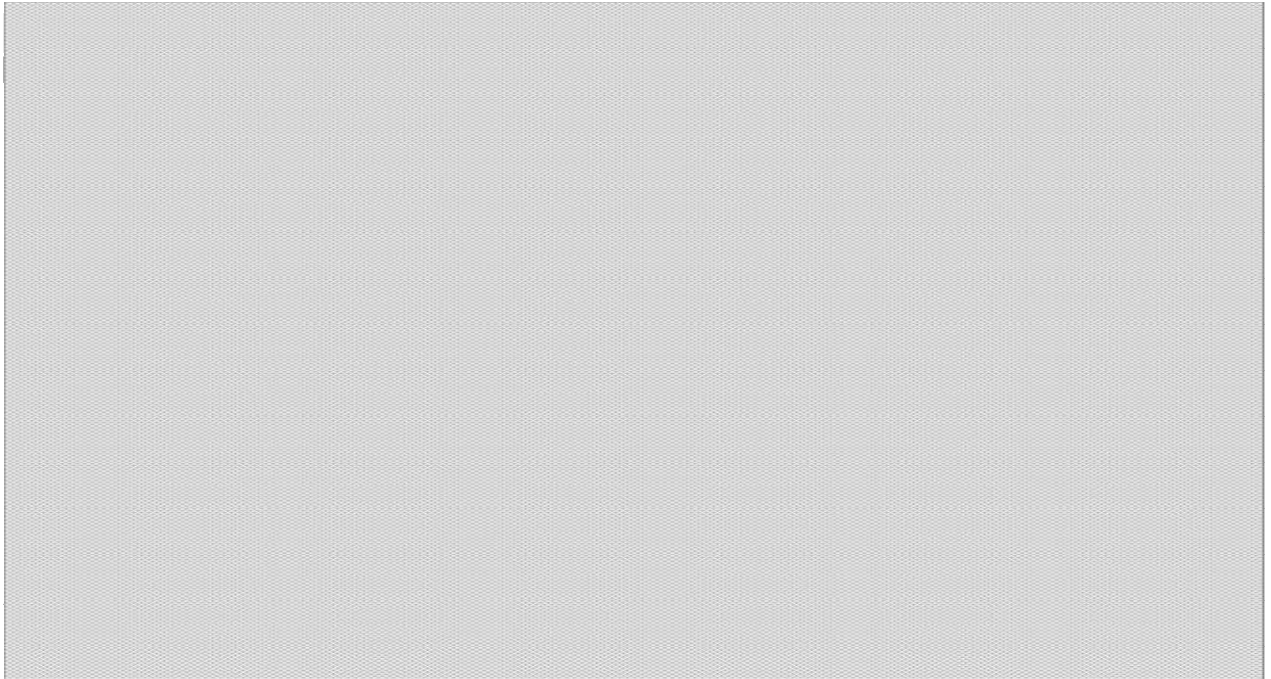


자료 : www.data.gov 홈페이지

● 영국의 공공데이터

- 2000년 11월 ‘정보자유법(Freedom of Information Act 2000)’을 제정하고 2005년 1월 이를 발효
 - ▶ 본 Act는 공공기관의 기록 정보에 대한 국민의 접근법을 보장
- 2010년 1월에는 data.gov.uk를 통해 공공 데이터 개방
 - ▶ 2014년부터 1,171개 기관에서 14,020개의 데이터셋을 개방

- ▶ 데이터셋의 44.8%가 CSV, XML, RDF 등의 오픈포맷 데이터를 제공
- 2010년에는 영국의 다양한 데이터를 보관하는 오픈데이터 관리 플랫폼인 'CKAN'을 오픈소스로 공개
- ▶ 2013년 5월 10일 Open Knowledge Foundation은 CKAN 2.0 버전을 출시



● 이외에 여러 국가, 도시, 기업·개인 등도 구축 데이터의 개방 확대

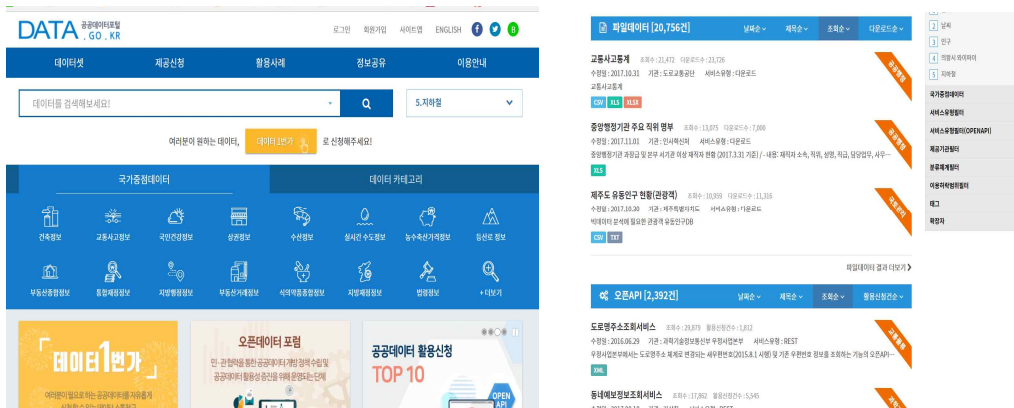
- ▶ 주 혹은 연방의 데이터 포털 : 캐나다 도시(data.edmonton.ca), 샌프란시스코(datasf.org), 런던(data.london.gov.uk) 등
- ▶ 기업·개인이 구축 : 오스트리아(gov.opendata.at), 벨기에(openbelgium.be), 월드뱅크(<https://data.worldbank.org/data-catalog>) 등

● 국내의 공공 데이터

- 공공데이터의 공개에 대한 선진국의 동향에 부응하여, 국내에서도 여러 공공기관 등이 웹사이트 구축을 통해 공공데이터 개방 중
- ▶ 13년 '공공데이터의 제공 및 이용 활성화에 관한 법률'이 제정된 후 공개 범위를 확대 중
- ▶ 법의 주요 내용은 공공데이터 국민 이용권 보장을 보장하는 것이며 이를 본격 추진하기 위해 국무총리 소속의 공공데이터전략위원회, 공공데이터제공분쟁조정위원회, 공공데이터활용지원센터 등 추진체계를 마련

- 정부·지자체·공공기관이 보유하고 있는 공공데이터를 개방하고 이를 이용할 수 있도록 공유자원 포털인 공공데이터포털(data.go.kr), 서울열린데이터광장 등이 구축됨

data.go.kr 홈페이지 및 공개 자료들



자료 : www.data.go.kr 홈페이지

- 2014년 9월에는 “공공데이터 개방 발전전략”을 발표
 - ▶ 주요 추진과제로 고가치·고수요 데이터 우선 개방, 공공데이터 품질보장 강화, 민간-공공 상생의 데이터 생태계 조성, 기반확충 및 지자체 확산 등을 제시
- 그럼에도 국내 공공정보 웹사이트는 로그인 절차, 약관 동의 등의 별도 과정이 필요하다는 등 데이터 공유보다는 공시에 가깝다는 평가
- 국내에서는 여전히 공공데이터 이용을 위해서는 로그인 등의 절차를 강제하고 있다는 단점이 있으나, 추후 유연한 정책이 추진될 예정
 - ▶ 미국과 영국의 공공정보 웹사이트는 별도의 로그인 절차 없이 공공정보에 접근해 데이터 다운로드가 가능
 - ▶ 미국과 영국 등은 해외 사용자들도 자국의 데이터를 이용할 수 있도록 하며, 특히 데이터 플랫폼을 오픈소스로 개방하는 등 영향력 확대를 추진

빅데이터 - 과활용의 문제와 데이터 관리

- (개인권리 침해) 기업이 내외부의 데이터를 구축하면서 구체적인 개인의 정보를 확보하거나 추정이 가능해지며 개인 정보와 이를 기반으로 개인 권리를 침해하는 사례도 급증
 - SNS, 모바일 기기, 웨어러블 기기 등을 통해 방대한 개인 데이터가 생성되고 저장되면서 기본적으로 개인정보 노출은 가능한 시대
 - ▶ 현재 국내에서 벌어지는 통신사, 카드사, 의료기관 등의 데이터 활용은 긍정적인 측면도 있으나, 기업들이 개인 맞춤형 서비스라는 미명 아래 모바일 마케팅 전화 급증, 개인의 약점을 파고드는 상품 권유, 보이스피싱 등이 점차 증가일로
 - 이러한 개인정보를 텔레마케터, 금융기관, 고용주 또는 정부가 활용할 경우 소비 패턴, 구직, 금융활동 등을 상업적인 용도로 활용할 경우 개인 권리 침해가 발생할 가능성이 매우 높음
 - ▶ 유명한 예로서 미국의 Target은 가족에게도 알리지 않은 청소년의 임신 사실을 미리 분석을 통해 예측하고 육아 용품 광고물을 보낸 일화가 있음
 - ▶ 미국의 전자 사생활 정보센터(EPIC)는 “공공 부문 및 사설 기관에서의 예측 분석(predictive analytics) 이용은 ... 정부와 기업이 개인의 비행 가능 여부, 구직, 입/출국, 신용카드 사용 가능 여부 등을 좌우할 수 있는 결과를 초래할 것이다. 또한 예측 분석을 통해 개인에게 부정적 영향을 미칠 수 있는 의사 결정을 내리는 것은, 결사의 자유를 직접적으로 침해하는 행위이다.”라는 논평을 미국 정부에 제시¹⁶⁾
 - 국내에서는 현재 법령상 개인정보를 수집·이용하고자 할 때는 이용자의 사전동의를 반드시 얻어야 하지만, ‘빅데이터 개인정보보호 가이드라인’(‘14.12월)을 통해 비식별화 조치를 할 경우 적극적 이용이 가능
 - ▶ 비식별화 조치를 진행한 개인정보에 대해서는 개인 동의 없이 활용 가능하며, 이용자가 거부 의사를 표현하면 활용은 불가
 - 그럼에도 다양해지는 데이터 소스, 분석 기법의 발달 등으로 공개를 원하지 않는 개인정보에 대한 유추가 점차 가능해지는 실정¹⁷⁾
 - ▶ 페이스북의 인맥 링크를 활용하여 개인의 민감한 정보도 추론이 가능하다는 예와 같이 공개된 정보의 조합은 위험한 정보로 변형 가능

16) <http://www.itworld.co.kr/news/91004>

17) LGERI(215.3), “ 빅데이터의 현실, 기대와 큰 격차 “

- ▶ 미국의 일부 헬스케어 기업들은 개인의 약품 구매 정보를 통해 의료기록을 추론하여 정보로서 활용하는 부작용 발생
- 미국, 유럽 등에서는 개인정보보호를 위한 규정을 2018년부터 시행하여, 자국민들의 보호를 강력히 추진할 전망
 - ▶ 유럽연합(EU)의 2018년 5월 GDPR(General Data Protection Regulation, 개인정보보호규정)을 시행할 예정. GDPR은 기업의 자유로운 데이터 활용을 허용하되 정보 주체의 개인정보에 대한 자기결정 권한을 대폭 확대한다는 취지. 이 법안은 EU 회원국 전체에 법적 구속력을 지니며, EU에 사업장을 보유했거나 유럽 거주 시민의 개인정보를 처리하는 모든 기업에 적용. GDPR 위반 범칙금도 2천만 유로(한화 270억 원) 혹은 글로벌 매출의 2% 중 큰 금액으로 부과되며, 매출 기준은 모회사 혹은 자회사의 매출을 모두 포함한 금액
 - ▶ 미국은 기업들을 대상으로 사베인 옥슬리(Sarbanes-Oxley, SOX), GLBA, HIPAA 등의 법규 준수를 요구 중
 - ▶ 사베인-옥슬리 법(Sarbanes-Oxley Act): 기업 관리자가 금융 데이터의 잘못된 해석에 대한 개인적인 책임을 지도록 규정한다. 이로 인해 많은 기업들이 회계관리, 내부통제시스템 도입 및 경영인증시스템을 구축했다.
 - ▶ 이에 따라 기업들은 IT Compliance에 대한 중요성을 인지해야만 하며, 단순히 데이터를 축적하는 것 이외에 기업내 시스템의 정비가 필요해짐

● (데이터의 과용·오용) 빅데이터를 활용한 가치 창출 노력이 과하여 아무 의미없는 데이터에 의미를 부여하고 연관짓는 빅데이터 오용 사례도 발생

- 데이터 기반 추론이 과하게 진행되어 Apophenia 현상의 발생 가능성 존재
 - ▶ Apophenia : 서로 무관한 현상들 간의 관련과 의미를 자동적으로 인식하는 일을 의미하며 클라우스 콘라트(Klaus Conrad)가 1958년에 처음 만든 용어
- 데이터 분석 결과의 성급한 일반화를 통해 Health, 범죄 예방 등에서 개인의 권리를 침해받는 경우(행동 스코어링)가 다수 발생¹⁸⁾. 이는 데이터 기반하에 유사 특징을 가지는 소비자들을 동일시하며 기업의 Risk 회피 전략
 - ▶ 보험사들은 대형 사이즈 의류를 구매한 기록이 있는 고객들을 추적해 그들에게 체중 감량 프로그램 참여를 유도
 - ▶ 의료 보험사들이 전문의약품 구매기록에 근거해 일부 환자들의 보험가입 거부
 - ▶ 특정 상점에서 상품을 구매했는데, 분석 결과 해당 상점의 다른 구매 고객들의

18) “분석없는 빅데이터, 황당하고도 위험할수도”의 사례정리(<http://www.itworld.co.kr/news/83548?page=0.1>)

상환 기록이 나쁘다면, 카드사는 현재 이용자의 신용 한도도 낮춤

- ▶ 80년대 말~90년대 초반 I-95 활주로에서 중고가 차량을 모는 20~27세 흑인, 히스패닉 남성은 약물 복용자일 확률이 높다는 데이터가 존재. 경찰들은 유사 차량을 검문하는 상황이 자주 발생
- ▶ 국내의 사례로서는 금융거래실적이 충분하지 않은 고객에 대해 개인의 행동 패턴, 관심사 분석 등으로 고객 성향 분석을 통해 적절한 신용평가를 진행

● (운용과 가치 도출의 가능성 문제) 빅데이터 전략을 위해 필수적인 빅데이터 운용(수집·저장·처리 등)과 분석을 위한 양질의 인력이 부족

- 빅데이터 전문가란 데이터의 저장, 처리, 분석, 시각화 등을 기존 소프트웨어를 활용하여 니즈에 부합되는 수준으로 데이터를 다룰 수 있는 인력임에도 실질적으로 필요한 인력을 찾거나 단기간에 양성하기 매우 어려운 상황
- 현재 빅데이터의 운용을 위한 소프트웨어가 많이 사용자 지향적으로 변했다고 하지만, 여전히 Hadoop, Spark 등을 사용하기 위해서는 공개소프트웨어에 대한 이해와 다양한 프로그래밍 기술이 필요
- 더욱 중요한 점은 가치를 도출하기 위한 중급이상의 빅데이터 분석가, 혹은 데이터 과학자가 많이 필요하나, 특히 국내에서는 매우 부족한 현실
- ▶ 정부 발표에 의하면(2016.1), 국내 공급기업에서 필요로 하는 인력은 고급데이터 분석전문가가 가장 많았으며, 수요기업에서는 빅데이터 경험자 및 데이터 분석 전문가를 필요로 함을 지적

공급기업과 수요기업의 빅데이터 선호 인력



- ▶ 고급 데이터 분석가에 대한 수요 급증
- ▶ IT분야, 통계, 비즈니스 마인드를 동시에 요구



- ▶ 실제 경험자 우대 및 선호
- ▶ 분석 부문 경험자, 숙련자를 우선 선호

자료 : 과학기술정보통신부(2016.1), “2015년 빅데이터 시장현황 조사결과 발표”

- ▶ 빅데이터 전문인력 전망(2017.7)에 의하면, '16년에 대비 '19년 필요인력수는 빅데이터 개발자가 '16년 2,703명에서 '19년 5,841명(116%p 증가), 빅데이터 분석가는 1,052명에서 2,280명(117%p) 증가할 것으로 추정¹⁹⁾

19) 한국데이터진흥원(2017.7), “2017 데이터산업 백서”

● (빅데이터 분석의 오용) 데이터 기반 분석이 데이터 중심 분석으로 변형되며 분석은 성공적이나 분석 과정과 결과의 편의성 발생으로 신뢰할 수 없는 사례를 발생시키며 분석자 혹은 기업의 신뢰성을 저하시킴

- 빅데이터 분석 과정 중 임의적인 자료의 활용, 수리적 모델에 집착하는 등 명확한 방향성 없이 기계적인 접근에 의존하는 경우 결과는 얻지만, 최적의 방향성을 설정하지 못하며 경제적·사회적 위험성 발생

▶ 케이시 오닐²⁰⁾은 본인의 저서에서 수학의 오용 사례로 주택담보대출 파생금융 상품이 빅데이터 분석에 쓰이는 수학적 알고리즘, 분석의 오용 사례로 교사 평가, 미국내 24개 주에서 쓰이는 재범 확률 예측 기법 등을 지적

<참고 사례> 구글 독감 트렌드 서비스

● 구글 검색 데이터의 축적을 높이 평가하거나 혹은 위험성을 경고한 사례

- 구글에서는 2008년 논문을 통해 2004년부터 구글의 독감 관련 검색 데이터와 미국 질병통제센터(CDC²¹⁾)의 독감의심환자(ILI²²⁾) 비율이 0.85의 상관관계를 보인다고 결과를 제시
- ▶ 이는 독감이 유행하는 지역에서는 독감 관련 사안들의 검색 빈도가 다른 지역보다는 높다는 점에 착안

20) Cathy O'Neil. 2016.9월 Weapons of Math Destruction : How Big Data Increases Inequality and Threatens Democracy를 저술

21) CDC : Centers for Disease Control and Prevention

- 구글은 상기 논문을 기반으로 검색기반하에 ILI를 추정하는 웹서비스인 Google Flu Trends(GFT)를 선보임
 - ▶ 25개의 국가에서 서비스를 진행하였으며, 높은 예측력을 보인 것으로 평가
 - ▶ 본 서비스는 정보제공이 미국 질병통제센터(CDC) 등이 병원에 접수된 독감 의심환자 숫자에 기반하여 독감주의보를 낸다는 점을 고려할때 일주일 이상 빠르게 독감에 대한 주의보 발령이 가능
- 그럼에도 2011년 8월을 기점으로 108주 중에서 100주 동안 실제 독감발병보다 높은 추정치를 예측, GFT는 2014년 이후 예측 서비스를 중단함

- 이에 대해서는 “구글트렌드가 준 교훈: 빅데이터 분석의 함정(The Parable of Google Flu: Traps in Big Data Analysis)” 등의 논문을 통해 빅데이터의 위험성을 비판하는 등의 사례로 활용
 - ▶ 그럼에도, 데이터 관점에서는 독감 트렌드와 CDC의 결과 등을 조합할 경우 신뢰성이 높은 시스템으로 활용가능함을 고려할 때, 매우 의미있는 연구

22) ILI : influenza-like illness

<참고> 빅데이터에 대한 반성 스몰 데이터

- (스몰 데이터) 아직 정립된 용어는 아니지만, 일반적으로 빅데이터와는 달리 특정 영역이나 주제에 대한 상세한 탐색을 주 목적으로 생성되거나 저장된 데이터를 일컬음
 - 예로 기존 통계학, 데이터 마이닝 등에서 활용되던 정형화되어 있으며, 전통적인 방법론이나 Computing으로 처리할 수 있는 데이터를 지칭하거나,
 - 혹은 개인의 취향이나 니즈, Health, 생활양식 등의 기존 자료의 트렌드와는 결과가 거리를 둘 수 있는 데이터를 포괄
 - ▶ 빅데이터는 거대한 데이터로 부터 실시간 혹은 배치 처리를 통해 상관·연관관계, 트렌드 등의 결과 도출이 목적
 - ▶ 아마존의 구매 추천 서비스처럼 빅데이터는 개개인의 특성에 맞는 맞춤형 서비스를 가능케 했고, 텍스트, 오디오, 비디오 등 구조화되지 않은 비정형 데이터들로부터 전체적인 구조나 패턴을 도출하는 데 도움을 줌에 따라, 상관관계를 유추해 미래를 예측해야 할 때는 빅데이터가 효과적
- 스몰데이터는 빅데이터의 특징인 3V 중 1V가 정도가 적용되어 구체적이며 실질적인 가치를 추출할 수 있다는 차별성을 가짐
 - (LEGO 사례) LEGO는 2000년대 초 파산 위기에서 레고마니아를 통한 스몰데이터 분석을 통해 위기 극복
 - ▶ 수차례 빅데이터를 분석을 통해 디지털세대를 겨냥하여 블록의 크기를 키우고 제품을 단순화한 제품을 출시하였으나 여전한 시장 실패. 레고 마니아를 통해 본인의 선택과 성취 제품에 대한 애착을 발견하여, 블록의 크기를 줄이고 조립설명서도 상세하게 만든 제품을 통해 파산 위기 극복
 - (스포츠 의류 브랜드 언더아머 사례) 빅데이터 및 스몰데이터 분석을 통해 소비자 타깃을 재조정하고 사업을 추진하는 것으로 알려짐
 - ▶ 언더아머는 자사 고객층과 흡사한 피트니스 트레이너 구매자들의 빅데이터를 분석해 연령대별 세그먼트 특징을 도출한 후, 각 세그먼트별로 심층 인터뷰를 진행. 즉, 스몰데이터 분석을 통해 이른바 감정적 구매동기를 확인
 - ▶ 빅데이터와 스몰데이터 결과에 따라 연령대별로 차별화된 스포츠 의류 전략을 세워 사업을 추진

- ▶ 언더아머는 최근 고객의 피트니스 정보와 웨어러블 기술을 연결하여, 데이터 기반의 고객 맞춤서비스와 실시간 서비스를 제공하는 디지털 비즈니스 모델 혁신으로서 스마트 기업으로 변신 중
- (에어비앤비) 초기 사업 부진 문제를 해결하기 위해 전통적인 통계기법을 활용하고, 이를 통해 서비스 개선에 반영
 - ▶ 현지 집주인과의 인터뷰를 통해 숙박시설 사진 품질의 중요성을 인식한 후, 사진 품질과 예약율의 상관관계와 연관성을 검증하는 회귀분석 기법을 적용하고, 예약율이 높아짐에 따라 전문 사진작가 촬영지원 서비스를 본격적으로 제공 시작
 - ▶ 또한, 숙박비의 가격 범위를 넓혔을 때와 그렇지 않았을 때, 즉시 예약 기능을 넣었을 때와 그렇지 않을 때 등 의사결정을 위한 가설을 설정한 후 소규모 데이터의 A/B 테스트와 파일럿 실험을 진행
- 스몰 데이터는 빅데이터와의 Correlation과 달리 Causality가 주요 주제로서 원인을 파악하는 경향에 따라 일대일 고객 맞춤형 서비스의 도입을 위해서는 필수적인 데이터로 평가됨
 - 소규모이지만 의미가 있는 스몰데이터의 분석 결과는 빅데이터의 추세와 패턴에 명확한 이유를 부여하여 실용적인 결과를 도출
 - 또한, 개인의 취향이나 필요, 건강 상태, 생활양식 등 사소한 행동에서 나오는 스몰데이터는 쉽고 간단해 빠르게 분석할 수 있고, 관리가 쉽고 효율적이라는 장점도 가짐
- 스몰데이터는 전통적인 통계학의 영역으로 간주될 수 있으며, 실제로 데이터 자체로는 전통적인 통계학의 영역과 차별성이 크지 않음

오픈소스 - 매출에 대한 부담감

- 빅데이터 플랫폼을 구성하는 하둡, 맵리듀스, 스파크, 하이브 등 핵심 기술들은 모두 오픈소스로 개발 중
 - 단순히 오픈소스를 인용하여 빅데이터를 처리하기에는 코드 레벨 수준의 제어가 필요하여, 비전문가들이 실질적으로 적용하기에는 어려움 존재
 - ▶ 클라우데라, 맵알, 호튼웍스 등에서는 수요자들이 쉽게 분산처리를 할 수 있도록 아파치 재단의 하둡 오픈소스를 활용하여 제품화한 사례
 - 따라서, 빅데이터 플랫폼 관련 기술 역량의 제고를 위해서는 공개된 오픈소스를 단순히 활용하는 것보다는 커뮤니티나 재단에 적극적으로 참여하고 협력관계를 유지하는 것이 필요
- 글로벌 기업들의 오픈소스 채택 및 활용은 꾸준한 증가세에 있음
 - ‘16년 오픈소스를 사용하는 기업은 65%²³⁾로서 ‘15년의 60%보다 상승
 - ▶ 기업들이 오픈소스를 사용하는 이유로는 솔루션 품질, 경쟁적 특징과 기술적 역량, 커스터마이징과 수정 역량 등을 지적
 - 국내에서도 삼성, LG 등이 오픈소스의 중요성을 깨닫고 오픈소스 커뮤니티나 재단에 참여하는 것으로 알려져 있으나, 전반적으로 기여도 등에서 아직 미진한 것으로 평가 받음
 - ▶ 국내의 오픈소스 개발자는 11,000명, 커미터는 약 500명 뿐으로 저조한 수준

- ▶ 현재 국내 오픈소스SW 시장²⁴⁾은 2016년 1,602억으로 추산하고 있으며, 2020년에는 2,862억원으로 성장할 것으로 전망(NIPA)

23) Black Duck Software(2016), “2016 Future of Open Source Survey Results”

24) 국내에서는 오픈소스를 공개소프트웨어(OSS, Open Source Software)로 사용. NIPA는 “소프트웨어의 내용을 프로그래밍 언어로 나타낸 ‘소스코드’를 공개하여 누구나 개량·재배포할 수 있는 소프트웨어”로 정의

● 국내 SW 개발환경에서는 지속적으로 오픈소스 참여를 하기 위해서는 장기적인 안목으로 R&D 투자를 진행하는 것이 매우 중요

- 오픈소스 업무 진행시 단기적으로는 인력투입 대비 성과를 얻기 어렵기 때문에 기업이나 출연연 등도 정규적인 업무로 인정하기 힘든 상황
- 정부에서는 오픈소스SW 역량 강화 및 수요 창출을 위해 공개SW 역량 프라자 운영, 온라인 교육 포털 서비스, 오픈소스 리더 양성을 위한 공개SW개발자Lab인 KossLab 사업, 글로벌 오픈프론티어 사업 등 다양한 지원 진행 중
- 현재 국내 오픈소스 생태계는 양적으로는 성장하고 있으나, 성장 자체가 활용 측면에 치우쳐 있으며, 공급 측면에서는 아직은 시장 초기로 판단
- 글로벌 기업들이 자사의 SW를 모두 오픈소스화하거나 오픈 커뮤니티의 스폰서로 활동하는 현 상황에서 국내에서도 SW개발자의 자율적 참여에 의존하기 보다는 오픈소스에 대한 인식 변화 필요
 - ▶ 오픈소스는 무료로 쓰는 프로그램이 아니라 오픈소스를 통해 생태계에 들어갈 수 있는 가능성이 높아짐을 인정해 주는 것이 필요
- 오픈소스의 성공을 위한 요인으로서 국내 대기업과 중소벤처기업 간의 SW 개발 관계성에 대해 개선이 필요
 - ▶ 오픈소스에 기반하여 개발한 스타트업, 벤처 기업의 솔루션에 대해 대기업 등이 협력 관계나 M&A를 하지 않고 아이디어 도용이나 인력 빼가기 등의 행위를 하지 않도록 건전한 관계성 조성을 독려
 - ▶ 이러한 관계성은 정부의 제도적 지원으로는 한계가 있으며, SW 혹은 Knowledge 등의 무형자산에 대한 민간의 자발적인 참여와 실천유도가 바람직

● 아파치 타조(Tajo) 사례

- 타조는 아파치(Apache) 하둡(Hadoop) 기반의 분산 데이터 웨어하우스 프로젝트로서 한국의 그루터가 주도하여 개발
 - ▶ 타조는 하둡 빅데이터 분석시 맵리듀스를 사용하지 않고 SQL을 사용하여 HDFS파일을 바로 읽어 내는 기술
- 타조는 아파치 소프트웨어 재단으로부터 인큐베이션 프로젝트로 승인 받음
- 그루터는 국내 S사와 기술협력을 맺은 후, 기술지원 문제 등으로 관계 파기
 - ▶ 언론에서는 오픈소스에 대한 몰이해, 국내 대기업과 중소기업간의 파트너 등의 인식 접근에서 문제가 발생한 것으로 설명²⁵⁾

IV 시사점 및 향후 제언

여전히 차별화되지 못한 빅데이터 & 빅데이터 플랫폼

● 빅데이터에 대한 효용성에 대한 기대는 아직도 진행 중

- 플랫폼 중심의 시장·산업 구조로 변화되면서 이러한 구조를 더욱 공고히 하고 기업의 위상 정립을 위해서 빅데이터는 소홀히 할 수 없는 영역
- 플랫폼 기반 비즈니스를 진행하며 자연적으로 발생하는 사용자·소비자 데이터를 활용하여 기존과 동일한 매출·가치가 발생한다면, 데이터의 집합체로서 빅데이터는 지속적 투자가 발생하는 전략적 투자 대상 전망
- 특히, 데이터 수집과 저장을 위한 IoT, 클라우드, 모바일 기술의 발전으로 기업들은 인프라에 대한 무리한 투자없이, 기업은 사용한 만큼 지불함에 따라 빅데이터 플랫폼에 대한 수요는 지속적으로 발생할 전망

● 빅데이터와 빅데이터 플랫폼은 아직은 인프라 구축과 데이터를 효율적으로 축적하는 기술 발전에 머무르는 상황으로 평가됨

- 하둡기반의 빅데이터 인프라가 여전히 사용자 범용성을 갖지 못하고, 엔지니어의 역할이 중요한 비중을 점유
 - ▶ 빅데이터를 대표하는 기술인 하둡은 현재 하둡 생태계를 이룰 정도로 유관 기술 개발이 적극적으로 진행되고 있으나, 비엔지니어들이 하둡 자체를 활용하기에 기술적 난이도가 있어 활용에 어려움 존재
- 빅데이터의 장점인 비정형 데이터, 실시간 분석 영역은 현재 기술 개발이나 구축이 진행 중이나, 본격적인 활용에는 미흡
 - ▶ 비정형 데이터의 활용은 여전히 그 비중이 높지 않으며, 사전 지식을 가진 경우에만 쿼리와 분석이 용이하다는 측면이 있음
 - ▶ 발생하는 데이터에 대한 실시간 분석은 탐색적인 가치 창출에 활용되지 않고, 주어진 모델에 대한 이상치의 발견에 활용되는 것으로 판단됨
 - ▶ 컬럼형DBMS에서 파생된 솔루션으로는 파스트림, 아이리스DB, 인피니플렉스 등이 있으며, 로그처리시스템에서 발전된 솔루션으로는 스플렁크, 테라스트림 바스, 로그프레소, D2 등이 있음

25) <http://www.pressian.com/news/article.html?no=128341>

- 실시간 분석은 하둡과 같은 범용의 환경 적용보다는 의료, 보안, 물류, 교통, 환경 등 특정 영역을 선택한 후, 범용 기술을 선택 영역에 적합한 형태로 데이터 수집과 저장, 분석을 모델링에 기반하여 정형화할 것으로 예상됨
 - ▶ 실시간 데이터는 제조, 물류 뿐만 아니라, IoT, 자율형 기기, 의료 센서 등에서 발생하는 이상치를 빠르게 판별하여 즉시 대응하는 상황에 부합되는 기술로서 많은 경우 기계가 생성하는(machine-generated) 데이터가 대상인 경우가 많음
 - ▶ 이러한 측면에서 ‘엣지 애널리틱스(Edge Analytics)’와 보안 영역이 실시간 분석 관점에서 크게 성장할 전망(가트너)
- 가치 창출을 위해 실질적으로 사용되는 분석 툴이나 시각화 영역은 빅데이터 인프라 투자에 비해 미흡한 것으로 평가됨

강자가 많지만, 약자도 없는 빅데이터 플랫폼 생태계

- 오픈소스 기반의 플랫폼 기술 성장으로 전통적인 클라우드 기업뿐만 아니라 수요자 중심의 커스터마이징화된 SW를 제공하는 중소벤처기업들도 시장에서 성장할 수 있는 경쟁 구도
 - 초기 빅데이터 시장은 오픈소스 기반하에 클라우드 구축과 같은 정보시스템 사업 경험을 가진 기업들 중심으로 형성되는 파일럿 시장으로 평가됨
 - ▶ 아파치 재단의 하둡과 맵리듀스를 시스템화하는 사업 중심으로 전개되는 양상
 - ▶ 초기 빅데이터 시장은 클라우드 시장으로 생각되며, 아마존, IBM, 마이크로소프트, 오라클, 구글 등이 선점
 - 이후 하둡을 중심으로 다양한 지원 프로젝트가 진행되며, 수요자 중심으로 패키징화하고 적시 서비스를 진행하는 기업들이 점차 중심에 포진
 - ▶ 클라우데라, 맵알, 호튼웍스 등이 사용자 중심의 오픈소스기반 상용 SW와 카운셀링 서비스를 제공하며 하둡의 활용을 주도
 - ▶ 그럼에도, 아마존 등의 기존 기업들이 사업을 철수하는 것이 아닌 자사의 서비스와 발전된 오픈소스를 결합하여 서비스의 질적 향상 도모
- (인공지능과의 연계 강화) 왓슨, 알파고를 시작으로 산업 측면에서 인공지능의 활용성에 대한 기대가 높아지는 상황
 - 영상의학 판독, 무인자동차, 음성 인식 등 인간의 반복적인 인지행위를 모방하는 영역에서는 인공지능의 학습을 위해 빅데이터는 필수적

- 전통적인 클라우드 기업들은 자사의 인프라와 인공지능을 연계하여 소비자를 겨냥하며, 경쟁업체 간에도 협력을 통한 새로운 확장 생태계 구현 중
 - ▶ 아마존은 AWS로 기반으로 인공지능 비서 알렉사를 활용하여 유통 사업에서 아마존 주도의 생태계를 구현하고자 노력 중
 - ▶ 특히 아마존은 알렉사 플랫폼을 Best Buy 등의 오프라인 유통업체에 개방하는 전략을 통해 연계 체계를 확대할 계획인 것으로 알려졌다²⁶⁾
 - ▶ 아마존은 마이크로소프트와 협력을 기반으로 AWS와 애저가 호환되는 오픈소스 SW 글루온을 선보이고, AI 스피커 에코에서 MS 코타나를 실행할 수 있도록 상호 연동 시스템 구축에 대한 파트너십을 체결하는 등 협력

국내 빅데이터 플랫폼 경쟁력 확보를 위한 제언

● 데이터 강국으로 진입하기 위한 공개와 공유, 수집 기반 조성

- (생활 데이터의 공개 확대) 날씨, 교통, 부동산 등의 국민 생활과 연계된 데이터에 대해서는 실시간으로 활용할 수 있도록 지속적으로 공개를 확대하고, 활용 측면에서도 최소한의 제한만을 두는 수요자 중심 형태로 운영
- (ICT 데이터의 공개 강화) 통신사, 제조사, IoT 등에서 휴대폰, 센서 등을 통해 발생하는 데이터에 대해서는 공개 최대한 확대
- (데이터 수집 창고 개설) 필요 데이터에 대한 국민 수요 등의 수집을 통해 수집 데이터에 대한 점검과 수정 진행

● 오픈소스 역량 확보를 위한 사회적 분위기 마련

- 최근 SW 개발 동향인 오픈소스 커뮤니티 중심의 소스 개발에 부합되게, 오픈소스 커뮤니티 활동 확대 및 역량 확대 집중

● 행정, 의료, 재난, 환경, 교육 등의 정보를 효율적으로 저장·분석 가능한 빅데이터 플랫폼으로 전환 구축

- 국내 공공·복지서비스 제공에 필요한 국내 공공 정형·비정형데이터를 통합 관리할 수 있는 통합플랫폼의 개발 및 Customized된 서비스 구현
- 일반 사용자가 쉽게 활용할 수 있는 오픈소스 기반 빅데이터 분석 Tool도 한 국형으로 개발하고 이를 기반으로 공공 빅데이터 플랫폼의 역량 제고

26) <http://www.kidd.co.kr/news/197665>

※ | 참고문헌

- 과학기술정보통신부. (2016.1). 2015년 빅데이터 시장현황 조사결과 발표
- 김민희. (2015.3.10.). 빅데이터의 현실, 기대와 큰 격차. LG경제연구원.
- 미래부 · NIA. 2015년 빅데이터 글로벌 사례집.
- 분석없는 빅데이터, ‘황당하고도 위험할수도’. (2013.08.30.). Retrieved from <http://www.itworld.co.kr/news/83548?page=0,1>
- 빅데이터 해외동향. (2013.3.31.). 조선비즈. Retrieved from <https://sites.google.com/a/chosunbiz.com/bigdata/news?offset=10>
- 빅데이터 플랫폼으로 ‘삶의 질’ 높이는 시카고시. (2017.6.27.). Retrieved from <http://www.ciokorea.com/news/34662>
- 빅데이터와 사생활 침해, 5가지 시나리오. (2014.12.12.). Retrieved from <http://www.itworld.co.kr/news/91004>
- 빅데이터엔 의미있는 반드시 패턴이 숨어있다… 시그널 분석이란?. (2013.4.4.). 디지털데일리. Retrieved from <http://www.ddaily.co.kr/news/article.html?no=103039>
- 싸이월드의 세계화 진출 실패 1탄. (2015.12.2.), Retrieved from <https://woning.wordpress.com/2015/12/02/싸이월드의-세계화-진출-실패-1탄>
- 시몽 뷔로. (2010.1.10.). 글로벌 비즈니스 마인드 세트. 청림출판.
- 아마존 알렉사(Amazon alexa), 타 유통업체와 합종연횡 노리나. (2017.11.13.). Retrieved from <http://www.kidd.co.kr/news/197665>
- 에릭 브린올프슨 · 앤드루 맥아피. (2014). *제2의 기계시대(The Second Machine Age)*. 청림출판.
- 위키백과 홈 페이지. (2017.11.19.). 몽고DB. Retrieved from <https://ko.wikipedia.org/wiki/몽고DB>
- “왓슨 무료개방” IBM, 표준선점 승부수. (2017.10.26.). *서울경제*.
- 왜 지금 디지털트랜스포메이션(Digital Transformation)인가?(1편). (2016.2.28.).

- Retrieved from
<http://digitalretail.co.kr/왜-지금-디지털트랜스포메이션digital-transformation인가1편//>
- IaaS 1위 AWS, 2위 마이크로소프트, 3위 알리바바. (2017.10.18.). CIO KOREA.
 Retrieved from <http://www.ciokorea.com/news/35917>
- 조재홍. (2017.5.24.). 한국의 공개SW 시장동향과 주요 활성화 정책방향.
 Retrieved from <http://www.bloter.net/archives/280604>
- 조대협 외. (2016.8). *빠르게 훑어보는 구글 클라우드 플랫폼*. 한빛미디어.
- 최진홍. (2017.1.2.). 살아남는 자가가 강한자다. 한국정보화진흥원 : 케이콘텐츠
- 하둡 에코시스템(Hadoop-Ecosystem)이란?. (2014.4.4.). Retrieved from
[http://over153cm.tistory.com/entry/하둡-에코시스템HadoopEcosystem이](http://over153cm.tistory.com/entry/하둡-에코시스템HadoopEcosystem이란)
 란
- 하둡 이해. (2015.12.09.). Retrieved from <http://iamdaisy.tistory.com/11>
- 한국데이터진흥원. (2017.7). 2017 데이터산업 백서
- 한국정보화진흥원. (2016.1). 2015년 BIGDATA 시장현황 조사.
- 한국정보화진흥원. (2017.3). 2016년 빅데이터 시장현황 조사
- BBC. (2017.6.20.). AI may take your job - in 120 years.
- Black Duck Software. (2016). 2016 Future of Open Source Survey
 Results. Retrieved from
<https://www.blackducksoftware.com/2016-future-of-open-source>
- Cathy O'Neil. (2016.9.6.). Weapons of Math Destruction: How Big Data
 Increases Inequality and Threatens Democracy. Penguin UK.
- Dave McClure. (2007.12.12.). Social Platform Wars. Retrieved from
<https://www.slideshare.net/dmc500hats/social-graph-platform-wars>
- Lazer, D., Kennedy, R., King, G., Vespignani, A. (2014.3.14.). The Parable
 of Google Flu: Traps in Big Data Analysis. Science. Vol. 343, Issue
 6176, pp. 1203-1205.
- IDC. (2016.12). Worldwide Big Data Technology and Services Forecast,
 2016-2020

- Imanuel. (2017). TOP 50 Bigdata Platforms and Bigdata analytics software. Retrieved from <https://www.predictiveanalyticstoday.com/bigdata-platforms-bigdata-analytics-software/#content-anchor>
- Matt Turck. (2017.4.5.). Firing on All Cylinders: The 2017 Big Data Landscape. Retrieved from <http://matrturck.com/bigdata2017/>
- Tableau Software. (2017.3.27.). 2017년도 상위 10가지 빅데이터 동향. Retrieved from <https://www.tableau.com/ko-kr/resource/top-10-big-data-trends-2017>.
- Thomas C. Redman. (2016.9.22.). Bad Data Costs the U.S. \$3 Trillion Per Year. Havard Business Review. Retrieved from <https://hbr.org/2016/09/bad-data-costs-the-u-s-3-trillion-per-year>
- Vitaly Friedman. (2008). Data Visualization and Infographics

저자소개

안춘모 ETRI 미래전략연구소 기술경제연구본부 기술경제연구그룹 책임연구원
e-mail: cmahn@etri.re.kr Tel. 042-860-5790

빅데이터 플랫폼 현황 및 이슈 분석

발행인 : 한 성 수

발행처 : 한국전자통신연구원 미래전략연구소 기술경제연구본부

발행일 : 2017년 12월

ETRI 한국전자통신연구원
미 래 전 략 연 구 소

305-700 대전광역시 유성구 가정로 218
전화 : (042) 860-1182, 팩스 : (042) 860-6504

* 주의 : 본서의 일부 또는 전부를 무단으로 전재하거나 복사하는 것은
저작권 및 출판권을 침해하게 되오니 유의하시기 바랍니다.

