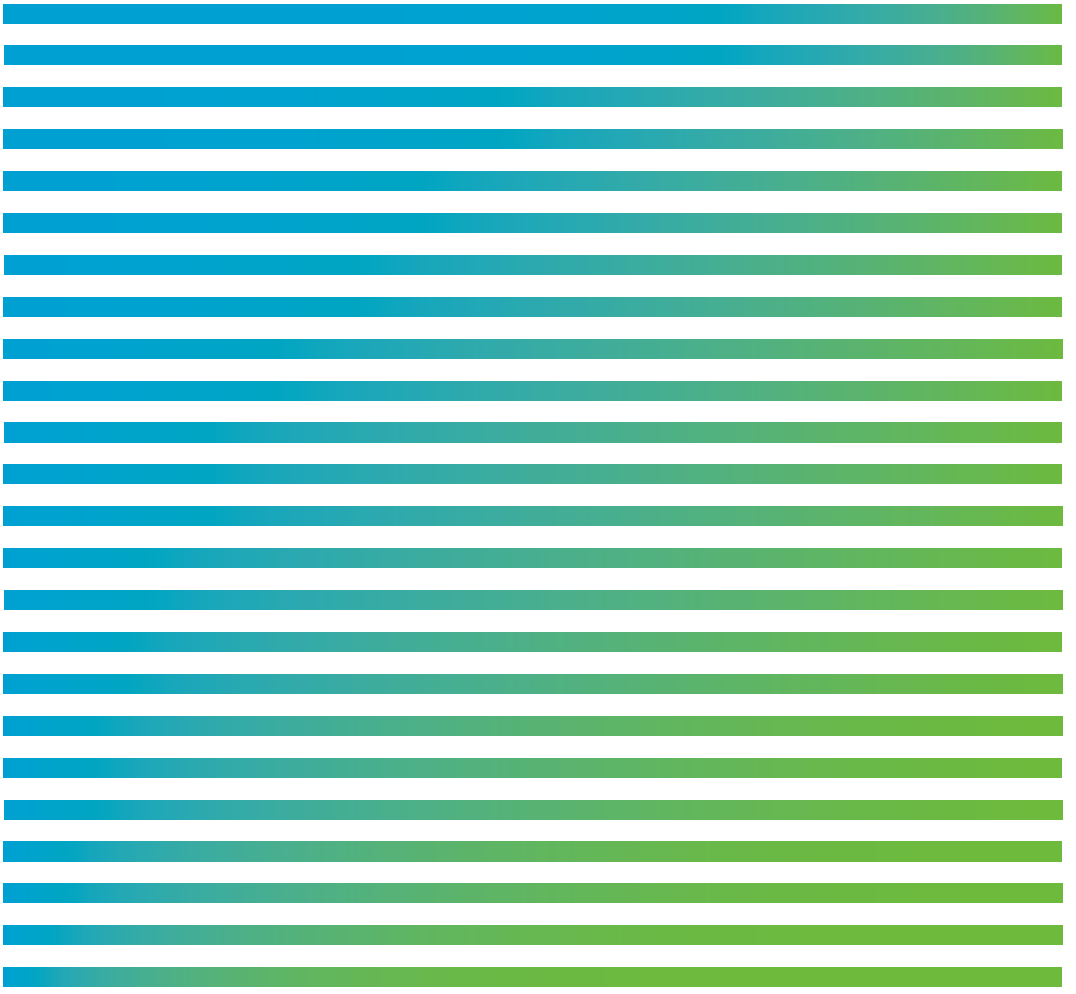
2017 데이터산업 백서

2017 Data Industry White Paper



Kdata 한국데이터진흥원

2017
데이터산업 백서

발간사

데이터 신가치 창출을 위한 데이터 혁신 필요

지능정보사회 도래와 함께 데이터의 중요성이 강조되고 있습니다. 데이터가 경제활동의 중심으로 대두되면서 미래의 국가 경쟁력 또한 데이터를 어떻게 활용하느냐에 좌우되고 있습니다. 데이터가 비즈니스를 주도하는 시대가 된 것입니다.

정보화 시대를 선도한 우리나라가 지능정보사회에서도 경쟁력을 갖추기 위해서는 데이터 신가치 창출을 위한 새로운 프레임워크가 필요합니다. 각종 데이터를 수집하고 실시간으로 전달해 효율적으로 분석하기 위해서는 ICT 기술을 바탕으로 스스로 데이터를 확보할 수 있는 생태계를 구축하고 이를 활용할 수 있는 알고리즘을 보유해야 합니다. 나아가 경제 발전의 패러다임도 데이터를 적극적으로 활용하여 혁신을 일으키는 데이터 주도 경제(Data Driven Business)로 거듭나야 합니다. 이러한 차원에서 한국데이터진흥원은 지능정보사회의 원천 자원인 데이터산업을 육성하는 데 앞장서는 데이터 혁신 드라이버의 역할을 수행하고 있습니다.

올해로 스무 번째 발간된 「2017 데이터산업 백서」는 제4차산업혁명으로 대표되는 지능정보사회의 원천 자원인 데이터와 이와 관련된 내용들을 정리한 것입니다. 데이터산업의 분야별 동향뿐 아니라 데이터 관련 최신 기술 트렌드 및 국내외 데이터 정책을 수록하여 데이터 관련 동향을 조망할 수 있도록 구성하였습니다. 특히 '2018 데이터산업 이슈 TOP 10'은 국내 데이터 업계·학계 전문가를 대상으로 설문조사해 분석·시각화한 결과입니다. 이를 통해 국내 데이터산업의 미래를 예측하고, 급변하는 ICT 환경 변화에 대한 대응방안을 모색해 볼 수 있을 것입니다.

본 백서가 학자나 연구자는 물론 데이터산업계 기업인들과 정책을 입안하는 정부 관계자들 모두에게 유용한 자료로 활용되어 데이터 중심의 지능정보사회를 준비하는 밑거름이 되기를 바랍니다. 아울러 「2017 데이터산업 백서」가 발간되기까지 많은 도움을 주신 편찬위원님들과 집필진, 편집진 여러분께 감사의 말씀을 드립니다.

감사합니다.

한국데이터진흥원장 **이영덕**

추천사

데이터산업이 나아가야 할 길 제시

데이터는 제4차산업혁명의 키워드인 '연결(Connect)'의 도구로 인식되고 있습니다. 데이터로 산업과 산업을 연결하고 각 산업의 정보가 공유되어 상호 협력 작용을 가능하게 합니다. 방대한 데이터를 활용하여 다양한 비즈니스 모델을 만들 수 있으며, 더 많은 부가가치와 일자리 창출에 크게 기여할 수 있습니다. 글로벌 경기침체와 새로운 성장동력 창출이 절실하게 필요한 현 상황에서 데이터산업의 역할은 매우 중요합니다. 이러한 시대의 요구에 따라 데이터산업 또한 빅데이터와 인공지능이 결합된 형태로 급속히 성장할 것입니다.

인공지능에서의 학습 영역을 구체화한 기술이 머신러닝으로, 과거의 데이터에서 숨겨진 패턴을 읽어내어 기계가 학습한 후 미래를 예측하는 기술입니다. 머신러닝의 적용 분야는 음성과 영상을 인식하는 원천 기술 분야와 고객 분류, 이상 징후 예측, 예측 정비, 큐레이션 추천 등의 인공지능 비즈니스 애플리케이션으로 나눌 수 있습니다. 인공지능 소프트웨어(SW)는 지금 한창 시장을 지배하는 원천기술 분야와 더불어 앞으로는 머신러닝을 비즈니스에 적용한 애플리케이션이 크게 성장할 것입니다. 머신러닝 기반의 비즈니스 애플리케이션은 산업별 도메인 지식, 데이터 전처리와 알고리즘이 어우러진 다양한 형태의 SW와 서비스로 발전될 것입니다.

제4차산업혁명에서 데이터는 원유 이상의 중요성을 가지고 있지만 데이터가 정확하지 못하고, 표준화되어 있지 않아 활용에 많은 문제를 갖고 있습니다. 또한 데이터 양이 늘어날수록 데이터 융합이 가속화될수록 데이터 품질은 떨어질 수밖에 없습니다. 그래서 인공지능 사업에 데이터 품질이 보증되지 않은 데이터를 사용하는 경우 데이터 전처리 작업의 비중이 전체 사업 소요 노력의 70~80%에 이를 수도 있다고 많은 전문가들이 경고하고 있습니다. 부정확한 데이터를 정제하고, 데이터를 표준화하는 노력이 많이 들기 때문입니다. 따라서 지속적인 데이터 정비를 통해 공학적으로 데이터 품질을 유지시키는 노력이 꼭 필요합니다.

한국데이터진흥원의 「2017 데이터산업 백서」 발간은 도래하고 있는 데이터 사회에 대한 이야기와 변화된 환경 하에서 데이터산업의 나아갈 방향을 제시하는 매우 의미있는 작업이라고 생각합니다. 또한 국내외 데이터산업에 대한 정확한 통계 제공, 관련 분야 동향, 최신 트렌드 등을 충실히 수록하여 관련 업계 종사자에게 뿐만 아니라 학술적으로도 매우 유용한 자료로 활용될 것으로 기대합니다.

한국데이터산업협의회 회장 **김종현**

추천사

가치창출 활동이 축적돼 완성될 제4차산업혁명

1·2·3차에 걸친 산업혁명이 생산성 혁명을 통해 생산, 소비, 조직, 고용 등 경제 및 산업 전반에 근본적인 변화를 가져왔습니다. 제4차산업혁명 또한 훨씬 빠른 확산 속도로 우리 삶에 새로운 패러다임을 정착시킬 것입니다. 제4차산업혁명은 꿈꾸는 리더들의 수많은 도전과 실패, 시행착오와 성공, 다양한 가치 창출 활동들이 축적되어 완성될 것입니다. 마찬가지로 데이터산업에서도 이전의 틀을 벗어나는 창의적인 선도자들이 더욱 많아져서 한국의 글로벌 경쟁력이 우뚝 서게 될 것이라고 기대합니다.

세계 경제가 글로벌화로 통합되고, 정보통신기술로 긴밀하게 연결되면서 경제력의 집중, 승자독식, 부익부 빈익빈의 경향이 강화되는 추세 속에 데이터를 제대로 다루는 사고능력과 기술능력은 더욱 중요해지고 있습니다. 한국데이터진흥원의 이번 백서는 데이터산업의 과거와 현재의 통계자료, 동향, 그리고 각계 전문가들이 전망하는 미래까지를 망라하여 전략적 가치를 제공함으로써 국내 데이터산업의 방향을 제시하고 있습니다.

후발 추격자 시절에 우리는 글로벌 선두 기업들을 목표로 벤치마킹하고 계획을 수립하여 신속하게 밀어붙이는 추진력이 탁월하였기에 지금의 위치에 왔습니다. 그러나 이제 리더군의 일원으로서 우리는 실패를 수용하고, 새로운 결과물을 글로벌 표준화하는 한편, 기회를 기다릴 줄 아는 태도가 필요합니다. 특히 ICT 산업에서 공동체 정신과 개방적 조직문화를 가져야만 환경변화에 다이내믹하게 대응하고 주도해나갈 수 있을 것입니다. 상생의 사회 실현은 궁극적으로 일자리와 소득 창출에 달려 있다고 봅니다. 이를 위해 학계도 미래의 노동에 맞는 교육체제로 변화해야 하며, 디자인 씽킹·컴퓨테이셔널 씽킹 등 새로운 사고와 다양성을 장려하는 교육·환경 변화에 선제적으로 대응하는 능력을 배양하는 훈련을 제공해야 할 것입니다. 이 백서는 학계의 연구와 교육에도 큰 도움을 줄 것입니다.

「2017 데이터산업 백서」 발간을 진심으로 축하하며, 데이터산업에 종사하는 모든 학자와 연구자, 기업인, 정책 입안자들에게 현재의 위상과 미래 의사결정 방향을 파악하고 연구하는 데 중요한 자료로 활용되어 국내 데이터산업 성장에 일조하기를 기원합니다.

(사)한국데이터베이스학회 회장 **황재훈**

추천사

빅데이터 시대를 위한 고언

빅데이터는 팩트입니다. 정보의 생성, 저장, 분석, 활용에 있어서 빅데이터가 대세이지요. 제품 원재료와 생산과정, 설비기기들로부터 품질 점검, 그리고 유통·물류 이동 및 판매시설·고객 활동 정보, 카드 결제나 통신 등의 정보가 이미 빅데이터입니다. 축적된 공공 빅데이터에 API를 빨대처럼 꽂아서 생성되는 2차적 빅데이터도 있습니다. 바이오 및 유전체 정보·의료정보 등도 거대 빅데이터 원천입니다. 개인이 방송사가 되는 시대가 되면서 사진과 동영상을 클라우드에 공유하는 소셜 네트워크도 폭증하고 있습니다. 사물마다 센서를 부착하겠다는 사물인터넷(IoT)에서는 정보 증가가 예측을 초월할 것입니다. 데이터 양을 단적으로 표현하면 역사시대 이래 지난해까지 만들어진 데이터 총량보다 올 한 해에 생성된 데이터 양이 더 많으며, 매년 이런 추세가 되면 이를 지수적 증가라고 합니다.

빅데이터 자체는 원유(Crude Oil)입니다. 첨단 공정과 처리능력을 갖추어야 각종 유익한 정보가 생산될 수 있기 때문입니다. 빅데이터가 없으면 인공지능도 작동하지 않습니다. 알파고가 이세돌 기사를 첫 상대로 선택한 이유 가운데 하나도 10년 이상의 기보가 공개되어 빅데이터로 승부를 예측하기가 쉬웠기 때문입니다. 빅데이터를 저장하는 클라우드와 보안기술, 장비와 머신러닝 및 데이터 마이닝 기술, 빅데이터 활용 서비스와 비즈니스 모델, 사업 기회, 그리고 빅데이터 전문가 양성 교육과정 등 각종 생태계가 확산될 것이 분명합니다. 빅데이터 시대는 세계적으로 초입 단계이며, 한국은 앞선 정보기술과 뛰어난 인프라를 갖고 있으므로 날개를 달고 올라 가리라고 다들 예상하고 있습니다. 그러나 실상은 그렇지 않습니다.

뜻밖의 이유는 보안입니다. 보안에 대한 과다 입법이 계속되고, 공공기관은 개인정보보호를 이유로 정보 공유에 소극적입니다. 전문가 집단은 논문 경쟁에 매몰되어 있고, 기술 이해가 부족한 사람들이 과도한 불안감을 조성하는 모습도 보입니다. 불행히도 빅데이터를 활용한 새로운 사업화와 효율적인 빅데이터 활용은 원천봉쇄되어 한국의 빅데이터 미래를 어둡게 하고 있습니다. 마치 바깥 세상이 두렵다고 아이들을 방에 가두어 놓은 형국입니다. 다른 집 아이들은 학교 다니고 축구를 하는데 언제까지 가두어 놓아야 할까요? 최소한 한국만 있는 법령은 과감히 폐지하고, 유사법령도 통폐합해 빅데이터 활용과 창업의 장애물을 치워야 빅데이터와 제4차산업혁명의 호기를 제대로 활용할 수 있을 것입니다.

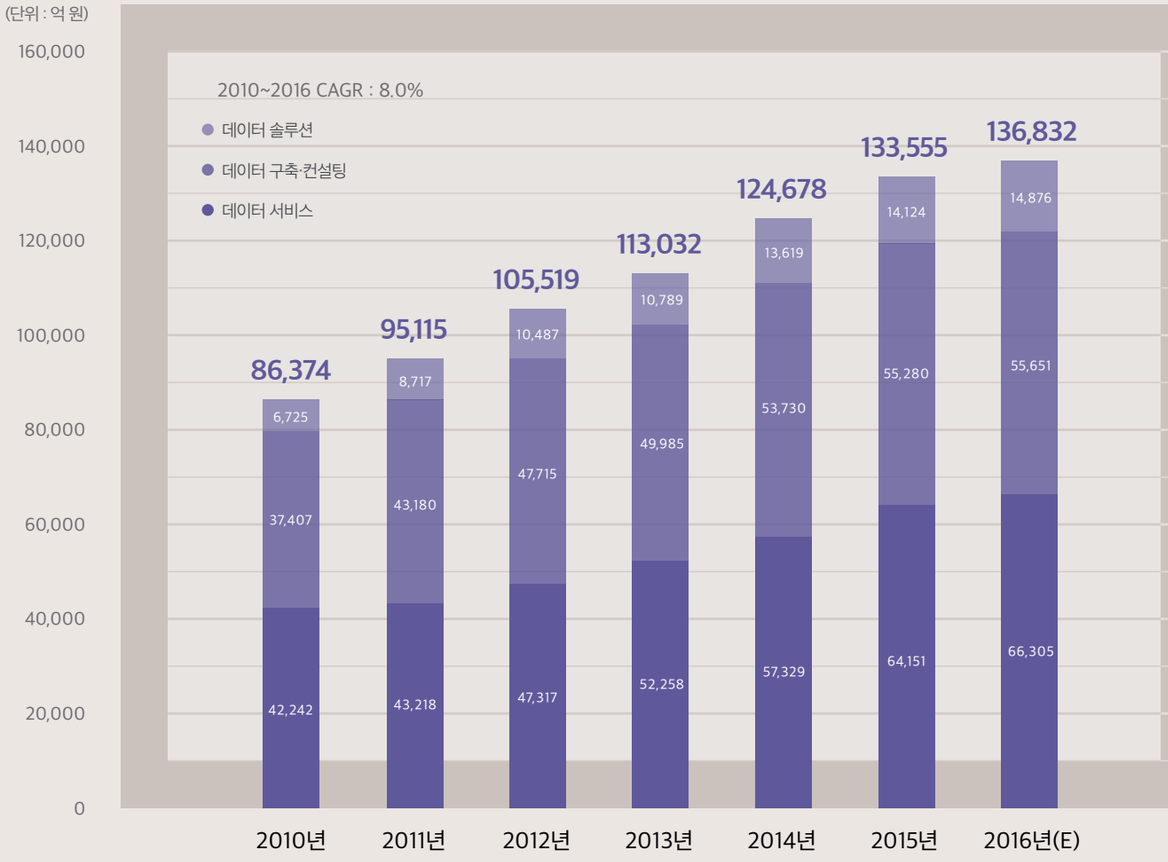
(사)한국정보과학회 데이터베이스 소사이어티 회장 **이우기**

그림으로 보는 데이터산업 동향

2016년 국내 데이터산업 전체 시장 규모는 2015년 13조 3,555억 원에서 2.5% 성장한 13조 6,832억 원으로 형성돼 2010년 이후 연평균 8.0%의 성장세를 유지하고 있는 것으로 조사됐다. 광고 매출 등 데이터와 관련된 간접매출을 제외한 직접매출 규모는 5조 8,803억 원 규모로 형성됐다.

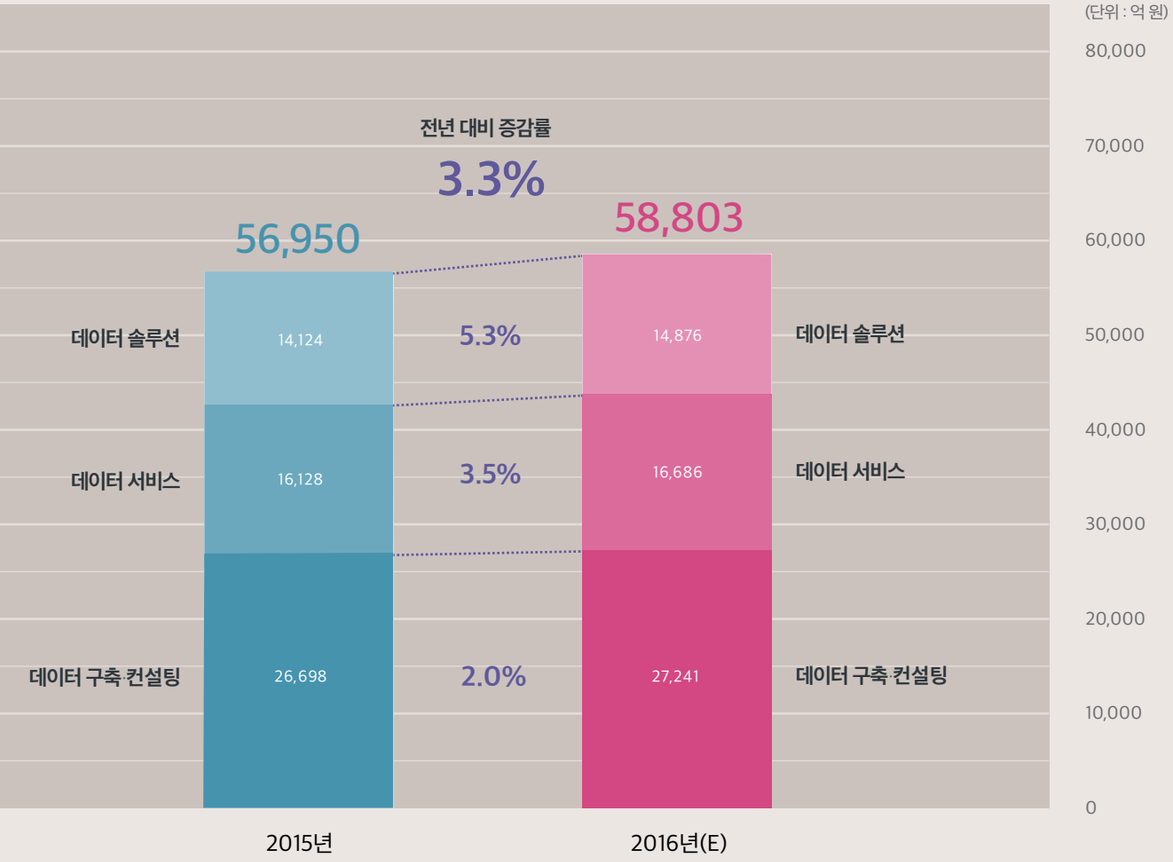
데이터산업 전체 시장 규모

13조 6,832억 원으로 형성된 2016년 데이터산업 전체 시장에서
데이터 서비스가 48.5%로 가장 큰 비중을 차지했으며,
데이터 구축·컨설팅이 40.7%, 데이터 솔루션이 10.9%를 차지했다.



데이터산업 직접매출 시장 규모

2016년 국내 데이터산업의 직접매출 시장 규모는 2015년 5조 6,950억 원에서 3.3% 성장한 5조 8,803억 원으로 나타났다.
직접매출 시장 규모는 데이터를 매개로 하는 광고 매출, DB 시스템 구축 용역 매출 등 DB와 관련된 간접매출을 제외한 결과다.
직접매출 대상 항목들의 성장률 측면에서는 데이터 솔루션 시장이 5.3%로 가장 높은 성장률을 기록했다.



그림으로 보는 데이터산업 동향

2016년 데이터산업계를 이끌고 있는 데이터직무 인력은
데이터 개발자가 38.0%로 가장 높은 비중을 차지했다.
직무별 인력 변동에서는 데이터 분석가가 43.3% 증가해
최근 데이터의 활용·분석에 대한 인기를 실감하게 했다.

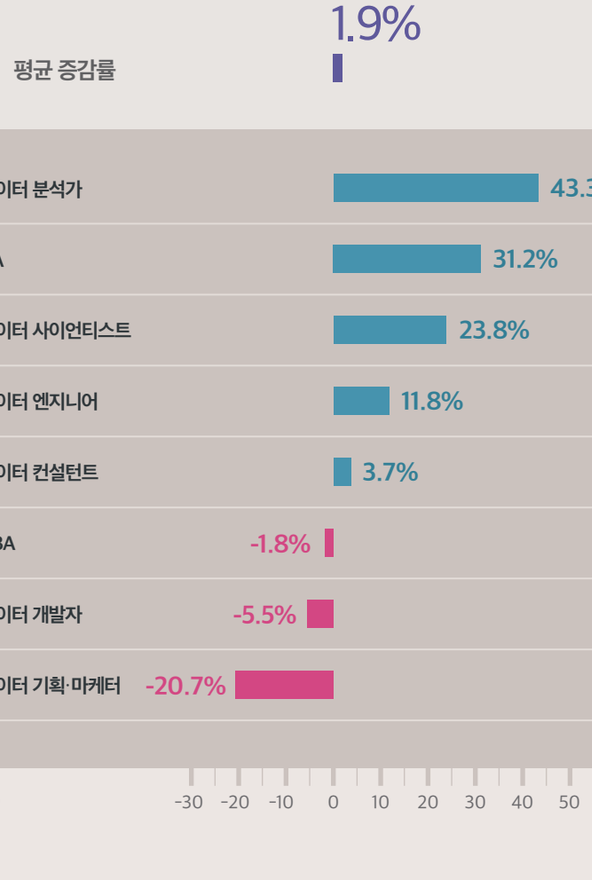
데이터직무 인력 현황

2016년 데이터직무 인력은 10만 2,375명으로 집계됐다. 이 중 71.6%인 7만 3,256명은 데이터산업계에서,
28.4%인 2만 9,119명은 일반산업에서 데이터 관련 업무를 하고 있는 것으로 나타났다.

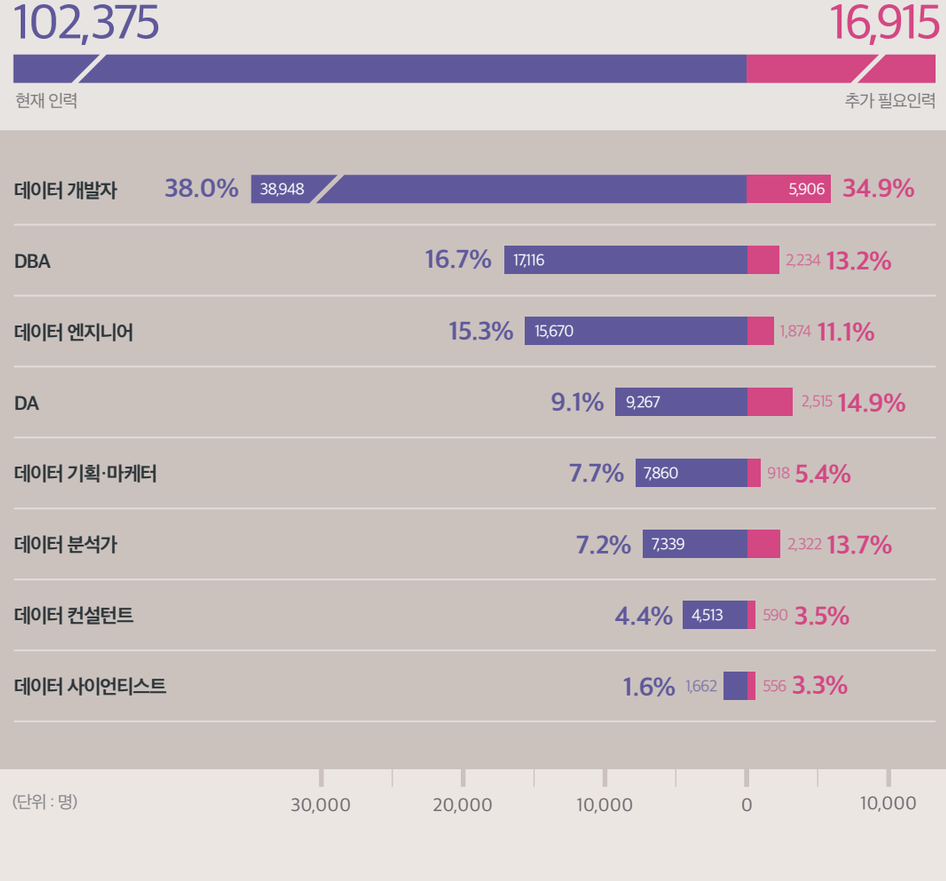
데이터직무 인력 부족률

데이터직무의 인력 부족률은 평균 14.2%로 전체적으로 인력이 부족한 가운데,
데이터 사이언티스트와 데이터 분석가가 각각 25.1%와 24.0%로 나타나 인력난이 가장 심한 것으로 조사됐다.

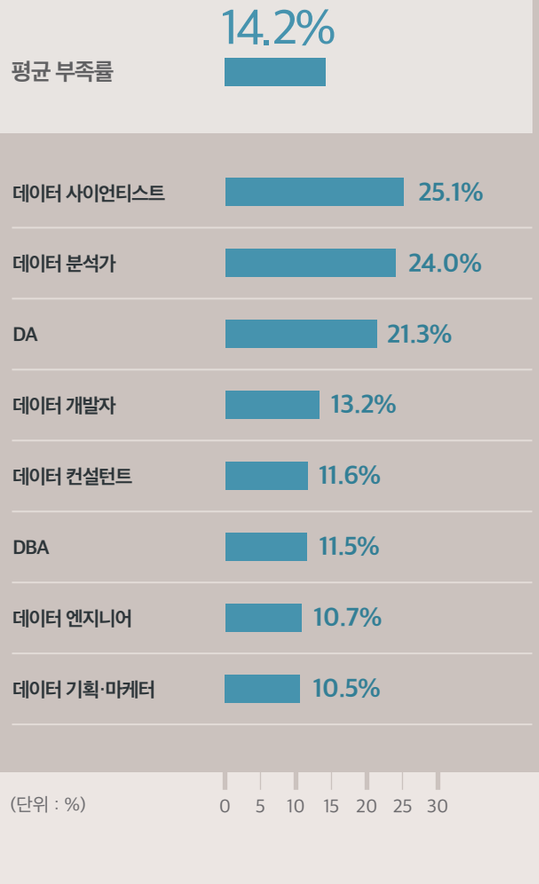
데이터직무 인력 증감률



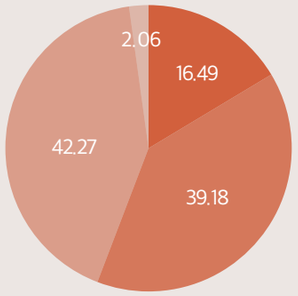
데이터직무 인력 현황



데이터직무 인력 부족률

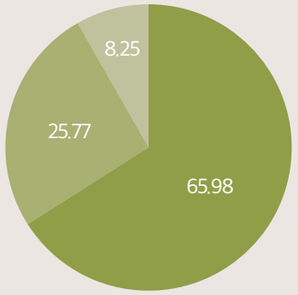


설문 응답자 비율



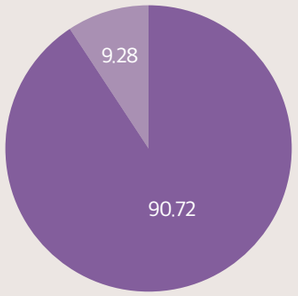
연령별 응답자

● 20~30대 ● 40대
● 50대 ● 60대



직업별 응답자

● 산업계 ● 학계 ● 기타



성별 응답자

● 남성 ● 여성

항목별 비율 | 표본 수 97명 (단위 : %)

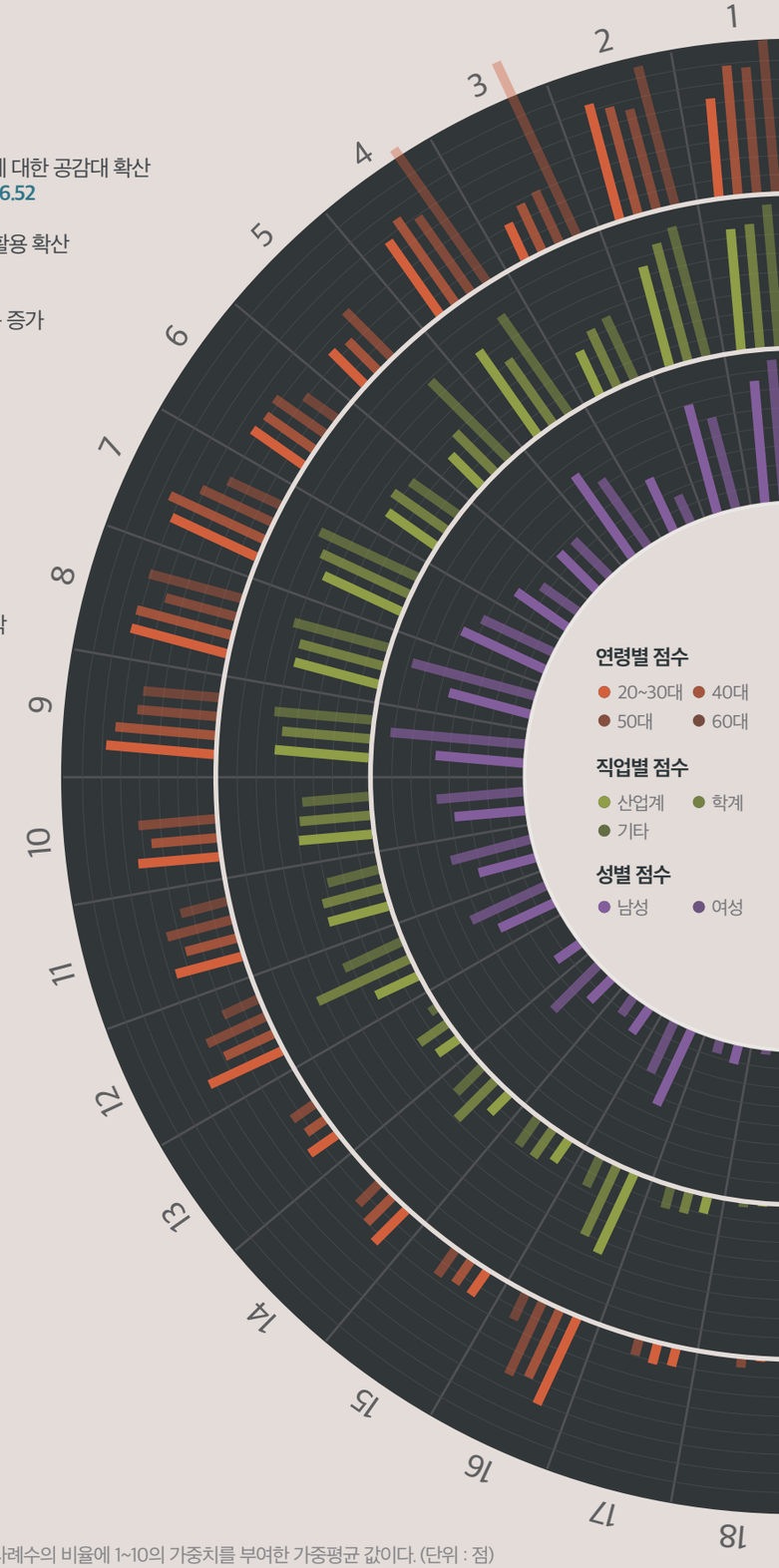
2018 데이터산업 이슈 TOP 10

2018 최고 이슈는 '데이터 자산가치'

백서 편집부는 데이터 업계와 학계 전문가들의 자문을 거쳐 오른쪽 페이지와 같은 총 18개의 설문 항목을 도출하였다. 그리고 사전에 선정한 250여 명의 데이터 업계 및 학계 관계자를 조사 대상으로 2018년도 데이터산업의 이슈 10개를 중요도 순으로 고르도록 했다. 조사는 2017년 4월 17일부터 5월 16일까지 1개월 간 폐쇄형 온라인 설문조사로 진행했다. 접수된 총 106건의 응답 가운데 9건은 무응답이거나 잘못된 응답이라고 판단돼 분석에서 제외했다.

설문 항목

1. 데이터 자산의 필요성 및 자산적 가치 평가에 대한 공감대 확산 — 6.52
2. 사물인터넷(IoT) 등의 생성 데이터 급증 및 활용 확산 — 5.76
3. 데이터 공유 수단으로서 블록체인 기술 활용 증가 — 2.88
4. 데이터 환경으로서 클라우드 도입 확대 — 5.12
5. 퍼스널데이터 시대 도래 — 2.71
6. 오픈 API 플랫폼을 통한 데이터 공유 확대 — 3.16
7. 데이터 유통 활성화를 위한 생태계 조성 시작 — 4.73
8. 데이터 전문가 수요 급증 — 4.61
9. 데이터 서비스의 다양화 및 확대 추세 — 4.88
10. 데이터 품질 향상 노력 확산 — 3.78
11. 데이터 표준화 필요성 증대 — 3.21
12. 개인데이터의 보호와 활용 논쟁 지속 — 3.28
13. 스파크를 필두로 한 실시간 분석 기술 발전 — 1.35
14. 데이터 주권에 대한 이슈 발생 — 1.79
15. 스몰데이터 각광 — 1.53
16. 데이터 기반 의사결정 확대 — 4.18
17. 데이터 기업의 인수합병 활발 — 0.95
18. 기타 — 0.24



문항별 점수 | 항목별 점수는 해당 문항별 순위 사례수의 비율에 1~10의 가중치를 부여한 가중평균 값이다. (단위 : 점)

이슈 TOP 10을 통해 본 데이터산업 트렌드

1 데이터의 자산적 가치에 대한 공감대 확산

지능정보사회에는 지식이나 기술 등 무형자산이 기업 가치의 원천으로 인식된다. 무형자산 중 하나인 데이터는 '21세기 원유'에 비견될 정도로 그 가치를 인정받고 있다. 재무제표 상의 경영 실적에는 드러나지 않지만 무한한 성장 잠재력을 보유한 데이터, 그리고 데이터의 집합인 데이터베이스는 미래의 성장 가치로 적절히 평가돼야 할 것이다.

2 사물인터넷 등의 데이터 급증과 활용 확산

사물인터넷(IoT)의 확산세 속에 기업들은 다양한 기기들을 통해 막대한 데이터를 수집하고 있다. 이러한 기계 데이터(센서·기기 데이터, 기계학습 데이터, AR/VR 생성 데이터 포함)는 매우 다양한 유형과 그 방대함에도 실시간으로 처리되는 추세이며, 지금까지 생각할 수 없었던 새로운 디지털 혁신기업을 탄생시키고 있다.

3 데이터 환경으로서 클라우드 확산

세계 굴지의 클라우드 서비스 제공사가 국내에 데이터센터를 설립하는 등 클라우드가 현실이 되고 있다. 특히 정보 인프라 차원의 클라우드 서비스(IaaS, Infrastructure as a Service)가 확산되자 데이터 기업들은 클라우드형 서비스로 고개를 돌리는 등 이제 데이터산업계에서 클라우드는 피할 수 없는 선택으로 자리잡았다.

4 데이터 서비스 다양화 및 확대 추세

빅데이터, 인공지능 등 기술 발전과 함께 데이터를 활용한 서비스에도 변화가 일고 있다. 최신 데이터 서비스들은 단순한 정보 전달에 그치지 않고, 데이터 활용을 기반으로 자동화한 서비스 추천, 맞춤형 서비스 제공, O2O 서비스 등 데이터 수익화(data monetization)를 기본으로 통찰(insight)과 실천(action)으로 연결될 수 있는 모델들이 부각되고 있다.

5 데이터 유통 활성화를 위한 생태계 조성

데이터의 중요성이 강조되면서 데이터 분석 역량과 함께 양질의 데이터를 쉽게 구할 수 있는 환경이 요구되고 있다. 이에 개인데이터 비식별화, 데이터 유통 플랫폼 조성, 데이터 프리존 운영 등 고부가가치 데이터의 자유로운 생산·수집·융합·활용을 위한 데이터 유통 활성화 노력이 적극 펼쳐지고 있다.

6 데이터 전문가 수요 급증

지능정보사회의 핵심 자산인 데이터를 다룰 수 있는 데이터 전문가의 수요는 점차 증대되고 있다. 특히 데이터 분석의 보편화는 데이터 분석가에 대한 인기를 지속시키고 있고, 기업에서는 관련 부서와 직무를 두면서 CDO(Chief Data Officer)의 역할도 부각되고 있다.

7 데이터 기반 의사결정 확대

데이터 분석의 보편화로 데이터를 자사의 비즈니스 최적화에 활용하는 등 데이터에 근거한 의사결정이 확대되면서 데이터 주도 또는 데이터 기반 비즈니스(data-driven business)가 늘어날 전망이다.

8 데이터 품질 향상 노력 확산

데이터 활용·분석이 활발해지면서 데이터 품질에 대한 관심이 증대되었으나, 기업들은 데이터 품질 진단과 문제점 파악에만 관심이 집중될 뿐, 적극적인 데이터 품질 향상을 위한 노력에는 다소 미흡한 실정이었다. 그러나 데이터 품질 관리 기술도 빅데이터와 인공지능 기술의 영향으로 올해 큰 전환점을 맞이할 것으로 보이면서 데이터 품질 향상을 위한 노력이 본격적으로 확산될 전망이다.

9 개인데이터 보호와 활용 논쟁 지속

OECD 및 세계경제포럼(WEF)에서는 개인데이터를 개인과 관련한 모든 데이터로 정의하고 있다. 빅데이터 기술의 발전으로 개인데이터 분석 수요가 늘어나고 있고, 개인데이터를 통한 가치 창출에 대한 관심도 올라가고 있다. 이에 해외 각국에서 다양한 노력도 진행 중이다. 반면 개인 프라이버시의 보호 가치도 동시에 강조되고 있어서 개인데이터 보호 이슈와 활용 논쟁은 당분간 지속될 전망이다.

10 데이터 표준화 필요성 증대

이종 데이터 연계 등 데이터의 활용 확산을 위해 분야별 표준 용어, 표준 도메인, 표준 코드 등 표준을 개발하고 적용할 수 있는 체계 마련이 시급하다. 더불어 데이터 가공 분석을 위한 데이터 표준 전환, 비식별화, 누락 데이터 보완 등의 전처리 과정, 오픈소스 기반의 데이터 유통 플랫폼(CKAN) 등의 데이터 관리 및 공유 방식 표준화 등의 필요성도 증대되고 있다.

연령별 TOP 10

20~30대

40대

50대

60대 이상

직업별 TOP 10

산업계

학계

기타

목차

제6부 데이터 기술 동향

제7부 데이터산업 정책 동향

표 목차

그림 목차

제 1 부

데이터산업의 미래

제1장

데이터 기반 혁신

제4차산업혁명 정책이 핵심으로 떠올랐지만, 아직 우리나라는 새로운 관점의 기술 기반 혁신을 준비하지 못 하고 있다. 기술 기반 혁신은 산업 생태계 혁신, 개인데이터 기반 혁신, 인공지능 기반 혁신이라는 세 가지로 정리할 수 있는데, 모두 데이터 기반 혁신과 깊은 관련이 있다. 데이터 기반 혁신은 산업 생태계를 변화시키고 기업 비즈니스 모델을 변화시키고 개인생활과 행태를 변화시킬 것이다. 이러한 변화에 적극적으로 대응하지 못 하면 우리나라는 미래 경쟁력을 가질 수 없다.

1. 기술 기반의 3대 혁신

2017년 IT 화두는 단연 제4차산업혁명이다. 2016년 1월, 다보스포럼 이후에 제4차산업혁명에 대한 관심이 폭발적으로 증가하고 있으며, 심지어 우리나라의 2017년 대선에서도 제4차산업혁명 정책이 핵심 공약이 됐다. 이러한 뜨거운 관심에도 우리나라는 미국, 독일, 일본, 중국에 비하면 기술적으로 많이 뒤쳐지고 있다. 더 큰 문제는 아직 우리나라가 새로운 관점의 기술 기반 혁신을 준비하지 못 하고 있다는 점이다.

새로운 관점의 기술 기반 혁신은 크게 세 가지로 정리될 수 있다. 첫 번째는 기업 내부 혁신에서 산업 생태계 혁신이다. 두 번째는 조직데이터 기반 혁신에서 개인데이터 기반 혁신이다. 세 번째는 현업 분석 기반 혁신에서 인공지능 기반 혁신이다. 그리고 이러한 세 가지 혁신은 모두 데이터 기반 혁신과 깊은 관련이 있다.

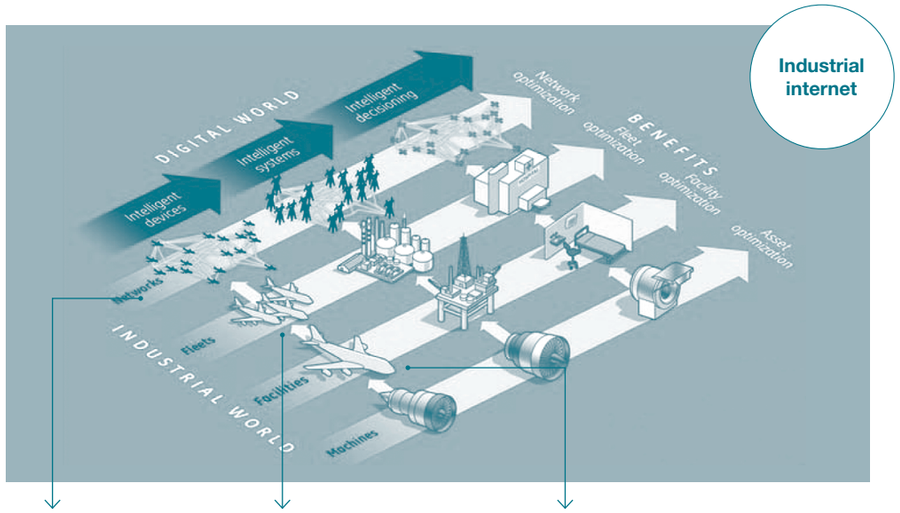
* 필자: 박주석(경희대학교 경영학과 교수)

가. 산업 생태계 혁신

제조업의 경우 제4차산업혁명의 궁극적인 목표는 공급사슬관리의 사이버물리시스템(Cyber Physical Systems, CPS) 구현이다. 단순히 스마트공장을 구축하는 것이 아니라 스마트공장, 스마트유통, 스마트제품, 스마트서비스를 융합하는 것이다. 대표적인 사례가 제너럴일렉트릭(GE) 회사의 지능형 항공운영이다. GE사는 스마트항공기를 생산해 장비 및 부품 정비는 물론이고 항공 운영상의 다양한 프로세스를 개선하고 있다. 또한 고객과 협력업체를 위한 지능화한 스마트서비스를 준비하고 있다. 궁극적으로 부품업체부터 항공사에 이르는 공급사슬관리의 CPS 구현을 목표로 한다.

이러한 공급사슬관리의 CPS 구현을 위한 핵심 인프라는 공급사슬 상의 빅데이터 체계이다. 사물인터넷으로 수집되는 엄청난 데이터를 분석해 각 사물들이 자율적으로 의사결정을 내리고, 지능화된 각 사물들이 생성한 데이터를 다시 통

[그림 1-1-1] GE의 산업인터넷 사례



출처: GE Imagination at Work, 2012

합해 가치사슬 상에서 지능화된 의사결정을 내릴 수 있어야 한다. 이러한 데이터 기반 혁신은 이제 시작이며 결코 우리나라가 늦지 않았다.

앞에서 제조업을 위한 데이터 혁신을 살펴보았지만, 모든 산업에서 데이터 혁신이 일어나고 있다. 금융산업에서는 핀테크(FinTech) 데이터 생태계가 구축되고 있고, 의료산업에서는 웰니스(Wellness) 데이터 생태계가 형성되고 있으며, 교통산업에서는 자율주행차를 위한 데이터 생태계가 논의되고 있다. 공공분야에서는 스마트시티(Smart City) 데이터 생태계가 나타나고 있다.

나. 개인데이터 기반 혁신

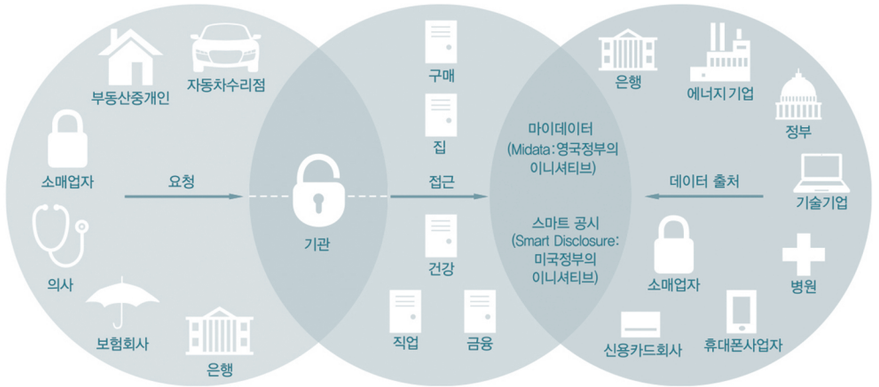
그동안 조직데이터 기반 혁신은 DBMS와 데이터 웨어하우스(Data Warehouse, DW)를 통해 꾸준히 이루어졌다. 더 나아가서 조직데이터의 내부 공유뿐 아니라 조직데이터의 외부 공개를 위한 여러 가지 정책 및 시스템이 추진됐다. 대표적인 사례가 공공데이터 포털인 Data.Gov이다.

최근에 공공데이터와 산업데이터 생태계가 융합되면서 새로운 비즈니스를 창출하고 새로운 가치를 만들어준다. 맥킨지컨설팅(McKinsey Consulting)은 2013년에 오픈데이터로 인한 가치를 연간 3조 달러로 산정했고, 향후 더욱 커질 것으로 예측했다. 반면에 정보기술 발전으로 개인데이터 유출 및 오남용 위험성이 증가하면서 전 세계적으로 개인데이터 보호 강화를 위한 법제도가 일반화됐다. 특히 우리나라에서는 2011년 카드회사의 데이터 유출 사태를 겪으면서 데이터 활용보다는 데이터 보호를 훨씬 더 중요시하게 됐다.

그럼에도 앞으로는 개인데이터 활용의 시대, 즉 퍼스널데이터의 시대가 열릴 것으로 전망된다. 스마트폰을 비롯한 웨어러블 디바이스 등 개인이 보유한 컴퓨팅 장비의 파워가 강력해졌다. 개인의 데이터를 개인이 보유하고 활용할 수 있는 시대가 된 것이다. 개인자산관리 앱은 금융회사에 흩어져 있는 개인 금융데이터를 한 곳에 모아서 보여주고, 이를 분석해 바람직한 투자와 소비를 가이드하기도 한다. 건강관리 앱은 병원과 약국 등의 진료 및 처방 기록을 통합·분석함으로써 질병 치료 및 건강 증진에 도움을 준다.

퍼스널데이터는 데이터 생태계에 흩어져 있는 개인데이터를 개인에게 제공해 개인 스스로 활용할 수 있도록 하는 것이다. 따라서 데이터 생태계의 폭과 깊이가 넓어질수록 거기서 융합된 퍼스널데이터의 가치는 더욱 높아진다. 예를 들

[그림 1-1-2] 퍼스널데이터 생태계



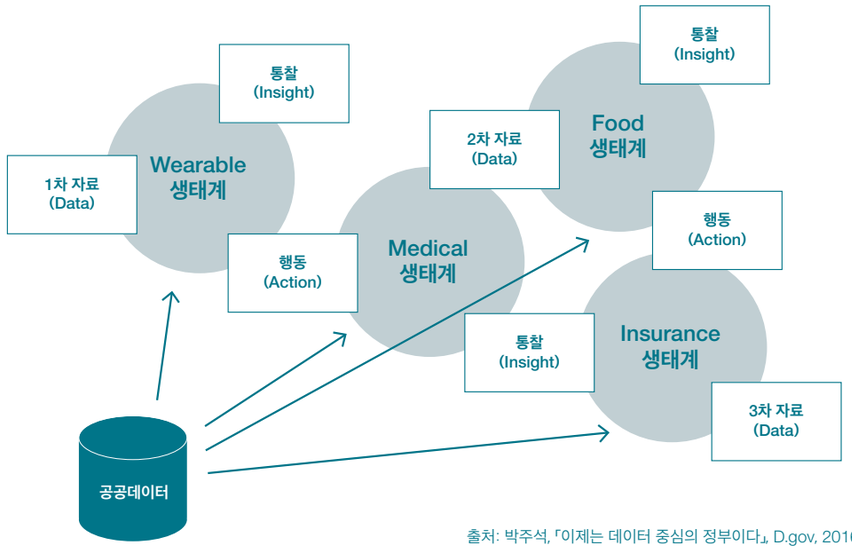
출처: 조엘구린, 「오픈데이터, 지금이 기회다!」, 투이컨설팅 번역, 2016

어 의료데이터 생태계를 확장해 웰니스 데이터 생태계를 구현한다면 융합데이터의 가치는 훨씬 높아진다. 웰니스는 웰빙(well-being)과 건강(fitness)의 합성어로 신체적·정신적·사회적 건강이 조화를 이루는 이상적인 상태를 말한다.

웰니스 데이터 생태계는 전자산업과 통신산업이 주도하는 웨어러블 데이터 생태계, 의료기관이 주도하는 의료 데이터 생태계, 피트니스산업과 건강보조식품이 주도하는 건강 데이터 생태계, 건강보험공단과 보험회사가 주도하는 보험 데이터 생태계가 상호 연계되어 있다. 즉 개인의 웰니스를 위해 다양한 관점의 데이터가 융합되어 서비스된다.

퍼스널데이터는 국민 개인의 생활 편의성과 복리 향상을 가져다 준다. 또한 개인을 위한 서비스를 제공하는 스타트업들의 탄생을 통해 산업 발전을 꾀할 수도 있다. 선진국은 이미 이런 점에 착안해 퍼스널데이터 정책을 펼쳐 왔다. 영국의 마이데이터(Midata) 정책은 개인이 원하면 개인데이터를 보유한 기업 또는 단체는 개인이 지정하는 형식으로 데이터를 제공하도록 하고 있다. 미국은 스마트 공시(Smart Disclosure) 정책으로 민간과 공공 부문의 개인데이터 제공을 확대해 나가고 있다.

[그림 1-1-3] 웰니스 데이터 생태계 사례



출처: 박주석, 「이제는 데이터 중심의 정부이다」, D.gov, 2016

다. 인공지능 기반 혁신

과거 산업혁명이 육체노동의 자동화를 이루었다고 한다면, 제4차산업혁명은 지식노동의 자동화를 이루었다고 한다. 특히 인공지능이 현업의 데이터 분석을 대체하기 시작했다. 로보어드바이저가 투자 자문을 하기 시작했고, 음성 비서가 콜센터 상담을 맡을 가능성이 커지고 있다. 챗봇은 메신저로 고객 문의를 접수하고 적절한 조치를 내릴 수 있다. 영상 인지 기술의 발전으로, 디지털사이니지는 영상으로 고객의 성별과 연령을 알아내기도 한다. 머신러닝은 문제를 해결하기 위한 로직을 만들고 데이터를 적용하는 것이 아니라, 데이터에 숨어 있는 로직을 찾아내서 필요한 코드를 만들어내는 방식이다.

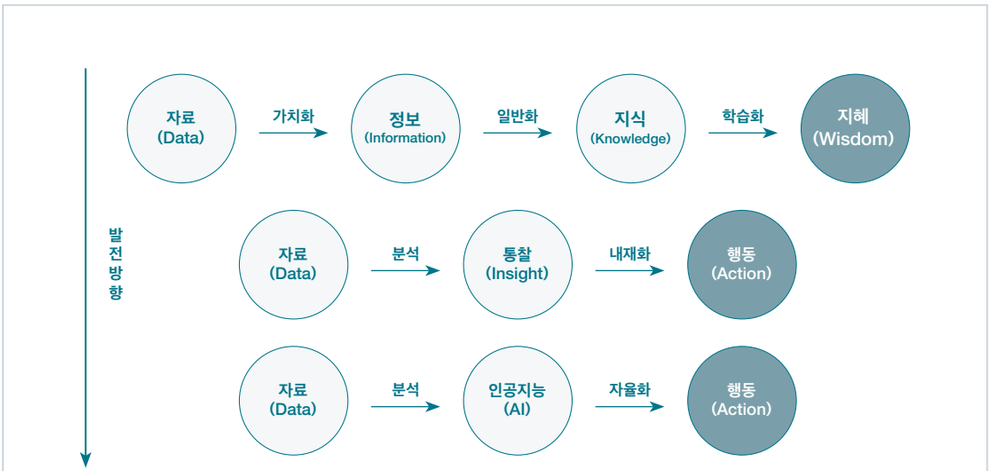
머신러닝의 성과는 알고리즘과 데이터가 좌우한다. 더 진화된 기술인 딥러닝으로 기존 방법보다 탁월한 성능을 보임에 따라 업무에 적용할 수 있는 수준이 됐다. 알고리즘은 오픈소스 생태계를 통해 빠르게 발전하고 또한 공유되고 있다. 인공지능(Artificial Intelligence, AI)이 학습하기 위해서는 방대한 데이터가 필요하다. 데이터의 양과 질이 인공지능의 학습량을 좌우한다. 더 정확하고 많은 콜 데이터를 학습한 콜센터 봇이 더 나은 고객 응대를 할 수 있다.

2. 알고리즘과 데이터의 힘

향후 인공지능은 가장 중요한 혁신 도구로 활용될 것이다. 그리고 인공지능이 혁신 도구로 사용되기 위해서는 데이터 혁신이 필수적이다. 전통적 데이터 생명주기는 DIKW(Data, Information, Knowledge, Wisdom)이다. 데이터를 가공하면 정보가 되고, 정보가 일반화되면 지식이 되며, 지식이 학습되면 지혜가 된다는 것이다. 최근에 회자되는 데이터 생명주기는 DIA(Data, Insight, Action)이다. 데이터를 분석해 통찰을 얻고 통찰에 의해서 바로 행동을 실행한다는 것이다. 중요한 점은 통찰을 얻기 위해 전문가의 도움으로 분석 알고리즘이 도출돼야 한다는 것이다. 향후 데이터 생명주기는 DAA(Data, AI, Action)가 될 것이다. 즉 전문가의 도움 없이 자동적으로 분석이 이루어지고 자율적으로 행동을 실행한다는 것이다. 따라서 기업환경에 맞게 이러한 새로운 데이터 생명주기 체계를 구현하는 것이 중요하다.

결론적으로 위에 언급된 세 가지 관점의 기술 혁신은 데이터 기반의 혁신이다. 데이터 기반의 혁신은 산업 생태계를 변화시키고 기업 비즈니스 모델을 변화시키고 개인생활과 행태를 변화시킬 것이다. 이러한 변화에 적극적으로 대응하지 못하면 우리나라는 미래 경쟁력을 가질 수 없을 것이다.

[그림 1-1-4] 데이터 생명주기의 발전 방향



출처: 박주석·김인현, 「전통적 환경과 빅데이터 환경의 데이터자원관리 비교 연구」, 한국빅데이터학회, 2016

제2장

데이터 기반 비즈니스 모델의 진화

데이터가 축적되고 활용하기 시작하면서 비즈니스 모델을 변화시키고 있다. 기존 기업은 데이터를 비즈니스 최적화에 활용한다. 과거를 분석하고, 미래를 예측하며, 최적화를 통해 성과를 높이고 있다. 최적화를 뛰어 넘어서 새로운 비즈니스를 창출하기도 한다. 스타트업들은 데이터를 수집하고 가공하며 분석하는 다양한 비즈니스 모델 패턴을 도입하고 있다.

1. 데이터 세상과 비즈니스 모델

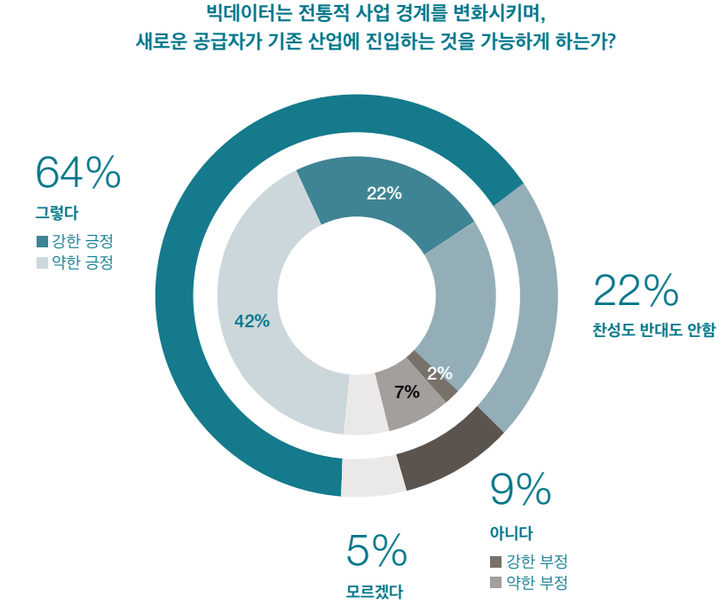
빅데이터가 주목받기 시작한 것은 2012년 무렵이다. 데이터 폭증과 함께 데이터 처리 기술의 혁신이 있었다. 데이터를 비즈니스에 어떻게 활용할 것인가가 주된 관심 사항이었다. 원인을 파악하고, 미래를 예측하고, 최적의 조치를 찾아내는 등 기존 비즈니스 프로세스를 더 잘 수행하기 위한 목적으로 데이터를 바라보았다. 성과가 뛰어난 우량 기업들은 보통 기업들과 비교할 때 데이터 분석을 더 광범위하게 활용한다는 조사 결과가 발표되기도 했다.** 「분석으로 경쟁하라 (Competing on Analytics)」의 저자인 데븐포트(Thomas H. Davenport)는 데이터 분석 능력은 기업의 최후 경쟁무기라고도 했다.*** 데이터의 가치는 기존 프로세스 성과를 높이는 데 있다고 본 것이다.

* 필자: 김인현(투이컨설팅 대표)

** LaValle, Steve, Eric Lesser, Rebecca Shockley, Michael S. Hopkins and Nina Kruschwitz (2011), *Big Data, Analytics and the Path from Insights to value*, MIT Sloan Management Review

*** Davenport, Thomas. H. (2016), *Competing on Analytics*, Harvard Business Review

[그림 1-2-1] 데이터가 산업에 미치는 영향



출처: 캡제미니(Capgemini), 2015

데이터는 기존 프로세스를 강화하는 차원을 넘어 비즈니스 모델을 바꾸거나 새로운 비즈니스 모델을 만들어내는 존재로 진화했다. 우버(Uber)는 자동차를 소유하고 있지 않지만 세계 최대의 택시 서비스 회사다. 에어비앤비(Airbnb)는 건물을 소유하고 있지 않지만 세계 최대의 숙박 서비스 회사가 됐다. 우버와 에어비앤비의 공통점은 데이터를 활용한다는 점이다. 데이터 기반의 비즈니스 모델로 큰 성공을 거두고 있다.* 2015년 캡제미니(Capgemini)**는 최고경영자들을 대상으로 설문 조사를 했다. 응답자의 64%는 데이터가 전통적 기업 경계를 파괴한다고 응답했다. 또한 27%는 인접한 산업의 기업들이 데이터를 활용해 새로운 경쟁자로 떠오르고 있다고 대답했다. 53%는 데이터 기반의 스타트업들이 경쟁자로 등장하는 상황에 이미 직면했다고 한다. 데이터는 비즈니스 프로세스를 강화하는 차

* Fersht, Phil (2016), *Data is eating the services industry!*, Horses for Sources http://www.horsesforsources.com/data-is-eating-services_011116

** Capgemini (2015), *Big & Fast data: The rise of insight-driven business*

원을 뛰어넘어서 비즈니스 모델을 변화시키거나, 새로운 비즈니스 모델을 탄생시키는 역할을 하고 있는 것이다.

[표 1-2-1] 데이터 활용 비즈니스의 기회

	데이터 수집	데이터 분석	지능 서비스
오픈데이터	데이터수집 데이터 필터링 데이터 정제	개인 및 기업 대상 공공데이터 분석 예) 기상 분석	공공기관 대상 대민 서비스 지능화
기업데이터	센서 데이터 수집 로그 데이터 정제 소셜 데이터 수집	마케팅 분석 공정 최적화 미래 전략 수립	데이터 기반 새로운 비즈니스 모델 도입
퍼스널데이터	개인을 위한 데이터베이스 구축 (본인 동의 필요)	에너지 소비 분석 카드 상품 비교 영화·음악 추천	개인 재무 관리 개인생활서비스 인공지능 비서

데이터 기반 비즈니스 기회는 점점 많아지고 있다. 데이터의 원천으로 구분하면 오픈데이터, 기업데이터, 퍼스널데이터 등을 활용할 수 있다. 오픈데이터는 정부 등 공공기관이 민간에게 공개하는 데이터다. 데이터 민주화 운동에 따라 일부 대기업들도 보유 데이터를 스타트업 등을 위해 공개하는 경우도 있다. 기업데이터는 빅데이터의 협의의 뜻에 해당한다고 할 수 있다. 기업이 생성하거나 확보하는 데이터로 기업 목적을 위해 사용된다. 퍼스널데이터는 개인 또는 기업 등이 자신을 위해 사용할 수 있는 데이터를 의미한다. 병원데이터, 금융데이터, 교육훈련데이터 등이 퍼스널데이터에 속한다. 퍼스널데이터는 데이터 주체가 원하면 보관하고 있는 기업 또는 기관은 데이터 주체에게 의무적으로 제공하는 방향으로 제도화가 이루어지고 있다. 데이터 원천 관점에서 활용할 수 있는 데이터는 더욱 빠르게 늘어날 전망이다.*

데이터 활동은 수집, 분석, 지능 서비스 등으로 구분할 수 있다. 데이터 활동은 데이터 관련 제도 및 규제에 크게 영향을 받는다. 우리나라는 데이터 규제가

* Filippov, Sergey (2014), *Data-driven business models: Powering startups in the digital age*, European Digital Forum

강력한 편이다. 따라서 아직은 주로 데이터 수집 영역에 그치고 있다. 앞으로 데이터 관련 제도화가 성숙되면 데이터를 활용하는 활동 폭도 넓어질 것으로 전망된다. 데이터 기반 비즈니스 모델에 직접적으로 영향을 미치는 것은 데이터를 어디까지 활용할 수 있는가이다. 데이터를 기반으로 기존 비즈니스 모델 혁신이 이루어지거나 새로운 비즈니스가 등장하기 위해서는 데이터 분석, 데이터를 이용한 지능 서비스 개발 등에 관련된 제약이 해소돼야 한다.

2. 기존 기업의 데이터 활용 진화

기존 기업들은 핵심 사업을 수행하는 과정에서 생산된 데이터 활용 방식을 변화시켜 왔다. 초기에는 데이터 웨어하우스(Data Warehouse, DW) 구축을 통해 비즈니스 상황을 파악하고 원인을 분석하는 데 활용했다. 이는 데이터 소비 확대로 이어졌다. 운영 데이터가 광범위하게 축적됨에 따라 분석(애널리틱스, analytics)을 도입해 프로세스 수행에 필요한 의사결정을 지원했다. 데이터 분석이 비즈니스 프로세스의 효율성을 향상시켰으며, 데이터 활용 수준이 프로세스 성과를 결정하게 됐다.* 데이터 기반 경영이 정착하게 되면 데이터는 비즈니스 모델의 핵심 질문을 해결함으로써, 기업의 가치를 높이는 데 기여하게 된다. 비구조적 데이터를 포함해 데이터 축적, 정제, 가공 등의 능력이 발전하면 데이터 자체가 새로운 수익 원천이 된다.** 데이터가 비즈니스 모델을 변화시키거나 또는 새로운 비즈니스를 창출하는 역할을 하게 된다.***

가. 소비 중심 데이터 활용

전사적으로 데이터를 더 많이 활용하는 것을 목표로 한다. 데이터 소비가 늘어나

* W.Wootton, Danny (2014), *Enabling data driven business models*, https://www.cgi-group.co.uk/blog/enabling-data-driven-business-models

** EIP-AGRI Seminar (2016), *'Data revolution: emerging new data-driven business models in the agri-food sector'*

*** Lokitz, Justin (2015), *Exploring big data business models & the winning value propositions behind them*, http://www.businessmodelsinc.com/exploring-big-data-business-models-the-winning-value-propositions-behind-them/

면 사실에 근거한 의사결정이 늘어나고 이는 기업 성과로 이어질 것으로 기대할 수 있다. 기업은 데이터 소비를 확대하기 위해, 데이터 사전 개발, 현업 사용자를 위한 데이터 활용 기반 구축 등에 투자한다. 데이터 소비를 확대하기 위한 교육과 훈련에도 노력한다. 소비 중심 데이터 활용(data consumption oriented)은 투자에 따른 성과가 직접적이지 않다. 데이터를 활용하는 사람의 능력에 따라 차이가 발생한다. 또한 데이터 가용 시점부터 성과 발생 시점까지 시간이 오래 걸리는 단점도 있다.

나. 성과 중심 데이터 활용

데이터를 활용해 프로세스 생산성을 높이는 것을 목표로 한다. 고객 서비스 처리율을 향상시키거나, 생산 활동 효율성을 제고하기 위한 데이터 활용 방안을 도입한다. 데이터 분석을 적용해 고객 서비스 요구를 사전에 분류하고 빠르게 처리하는 시스템을 활용한다. 또는 생산 공정 데이터를 수집하고 모니터링해 생산 활동이 극대화될 수 있도록 생산 라인을 조정한다. 성과 중심 데이터 활용(process performance oriented)을 통해 프로세스 효율을 높일 수 있다. 프로세스 효율이 높아진다고 해서 기업 가치 향상에 반드시 도움이 되는 것은 아니다. 제품을 싸고 빠르게 많이 생산한다고 해서 기업의 수익이 증대되는 것은 아니기 때문이다.

●○
기업들은 핵심 사업을 수행하는 과정에서 초기에는 데이터 웨어하우스(data warehouse) 구축을 통해 비즈니스 상황을 파악하고 원인을 분석했지만 이후 광범위하게 축적된 운영 데이터를 분석해 비즈니스 프로세스의 효율성을 향상시켰다.

다. 가치 중심 데이터 활용

수익 증대, 비용 절감 또는 고객 기반 강화 등 기업 가치를 향상시키는 데 목표를 둔다. 가치 중심(value driven) 데이터 활용은 비즈니스 모델의 핵심 가정을 찾아내고 이를 확인하고 최적화할 수 있는 데이터 활용 기반을 도입한다. 기업의 매출은 가격 결정에 크게 영향을 받는다. 최적 가격은 수요 곡선과 공급 곡선의 교차점에서 이루어지는 가격 수준이다. 이보다 높은 가격을 결정하면 기회 손실이 발생한다. 낮은 가격을 결정하면 수익을 최대화할 수 없다. 수요 곡선은 시장 상황에 따라 유동적이다. 데이터 분석에 의해 동적 가격결정을 할 수 있게 되면 이는 직접적으로 기업 가치 향상에 도움이 된다. 데이터 활용이 비즈니스 모델의 핵심 위치에 자리잡게 된다. 데이터 자산이 실물 자산을 대체하는 현상이 발생할 수도 있다.

라. 결과 중심 데이터 활용

데이터를 새로운 비즈니스 수단으로 이용하는 데 목표를 둔다(outcome based). 데이터는 기존 비즈니스 결과로 생성됐지만, 활용은 기존 프로세스를 위해서가 아니라 새로운 상품이나 서비스를 만들어내는 데 활용된다. 또는 데이터 그 자체가 새로운 비즈니스의 원천이 되기도 한다. 예를 들면, 금융회사는 고객 신용등급 판정 능력을 이용해 온라인 상거래 비즈니스를 할 수도 있다. 또는 축적된 광범위한 고객의 행동 습관, 거래 방식 데이터 등을 활용해 다른 상거래 기업에게 마케팅 서비스를 제공할 수도 있다. 또는 고객 데이터를 직접 판매하는 비즈니스 모델을 도입하는 것도 가능하다. 데이터는 산업의 원유(crude oil)가 되는 것이다.

3. 기존 기업의 데이터 활용 비즈니스 모델 패턴

기존 기업의 데이터 활용 수준이 가치 중심 및 결과 중심 수준이 되면 비즈니스 모델의 변화가 발생한다. 데이터를 활용하는 패턴은 다섯 가지 유형으로 구분할 수 있다.*

패턴 1: 데이터 판매

기존 비즈니스를 통해 축적된 데이터를 외부에 판매하는 비즈니스 패턴이다. 통신회사가 고객의 통신 요금 납부 데이터를 금융회사에 판매하면, 금융회사는 이를 여신 심사 데이터로 활용할 수 있다. 데이터 판매는 개인데이터 보호 규제를 준수해야 한다. 데이터 판매는 세가지 유형이 있다. 데이터 서비스(DaaS, Data as a Service), 정보 서비스(IaaS, Information as a Service), 분석 서비스(AaaS, Analytics as a Service) 등이다. 데이터 판매 서비스를 하기 위해서는 익명화 등 개인데이터 보호 준수 능력, 데이터 품질 관리 체계, 데이터 분석 역량 등이 필요하다.

패턴 2: 상품 혁신

기존 서비스와 상품 판매 및 운영을 통해서 생산된 데이터를 활용해 기존 상품을

* Hofman, Ralph and Arent van't Spijker (2013), *Patterns in data driven strategy*, Blink Paper

혁신하거나 새로운 서비스 또는 상품을 개발하는 데 적용하는 패턴이다. 보험회사는 고객 데이터를 분석해 고객에게 가장 적합한 보험상품 추천 역량을 발전시킬 수 있다. 금융회사는 부정 사용 데이터를 분석해 일정한 규칙을 발견하고 이를 프로세스에 적용시킴으로써, 부정 사용에 대응하는 능력을 획기적으로 개선할 수 있다. 통신회사는 고객의 통화 사용 패턴을 분석해 특정 고객 집단에 맞춤형 상품이나 서비스를 출시할 수 있다. 상품 혁신을 이루기 위해서는 데이터 사건들의 연관성을 추적할 수 있는 분석 능력을 갖추어야 한다.

패턴 3: 상품 스왑

기존에 보유한 고객에게 기존 상품이 아닌 다른 상품을 판매하는 패턴이다. 자동차 내비게이션 소프트웨어를 판매하는 회사는 운전자의 운전 습관 데이터를 토대로 적절한 보험 상품을 판매하는 서비스를 시도할 수 있다. 또는 기상정보를 서비스하는 회사는 고객의 생명주기 데이터와 상품을 연계해 레저 추천 서비스를 추가할 수 있다. 상품 혁신 서비스를 하기 위해서는 대상 고객 데이터를 고객 중심으로 통합 관리하는 능력과 매칭 및 추천 등의 데이터 분석 능력이 필요하다.

패턴 4: 가치사슬 통합

두 개 이상의 비즈니스 모델이 데이터를 공유함으로써 각자의 가치사슬을 강화하는 패턴이다. 대형 할인 판매점은 상품의 월별·일별 판매실적 데이터와 판매 예측 데이터를 상품 공급회사와 공유함으로써 재고자산 규모를 줄이고 필요한 상품 부족 현상을 방지할 수 있다. 병원이 고객의 건강데이터를 보험회사와 통합하면, 병원은 고객의 병원비 지급 능력을 미리 확보할 수 있고 보험회사는 건강보험 인수 심사 적중 수준을 향상시킬 수 있다. 데이터 규제를 준수하기 위해 직접 데이터 교환보다는 오픈API를 통한 분석 결과의 공유가 선호되는 방식이다. 가치사슬 통합은 자신의 필요에 의해 시도하기도 하지만 중요 거래처의 의지로 인해 불가피하게 추진해야 하는 경우도 있다. 내부적으로는 프로세스가 자동화돼 있어야 하며, 외부적으로는 오픈API 플랫폼을 연계해 접속을 원활하게 추진할 수 있어야 한다.

●○ 장기적으로 모든 비즈니스 모델은 다른 비즈니스 모델의 가치사슬과 통합되는 현상이 일반화될 것이다.

패턴 5: 가치 네트워크 창출

동일한 고객 집단을 대상으로 비즈니스를 운영하는 여러 비즈니스 모델이 고객 데이터를 공유함으로써 가치 네트워크를 창출하는 패턴이다. 고객 관점에서 오퍼링을 최적화함으로써 고객 장악력을 강하게 하고, 고객당 거래 처리 비용을 낮추는 효과를 기대할 수 있다. 특정 대학교 학생들을 대상으로 통신회사가 통신 서비스를 제공하면 앱 서비스사가 학교 근처 식당들과 제휴해 할인 서비스를 연계해 제공하는 방식을 생각해볼 수 있다. 여행 예약 사이트는 항공사, 렌터카, 숙박업체, 관광시설 등과 제휴해 통합된 오퍼링을 제시한다. 가치 네트워크 창출은 동일한 고객 집단을 대상으로 서비스를 제공하는 여러 비즈니스 모델이 존재하는 것을 조건으로 한다. 고객 접점에서 서비스를 제공하는 회사만이 가치 네트워크를 주도하는 것은 아니다. 고객 데이터를 고객 관점에서 융합하고, 다양한 서비스들을 통합해 매칭시킬 수 있는 능력이 핵심 역량이다.

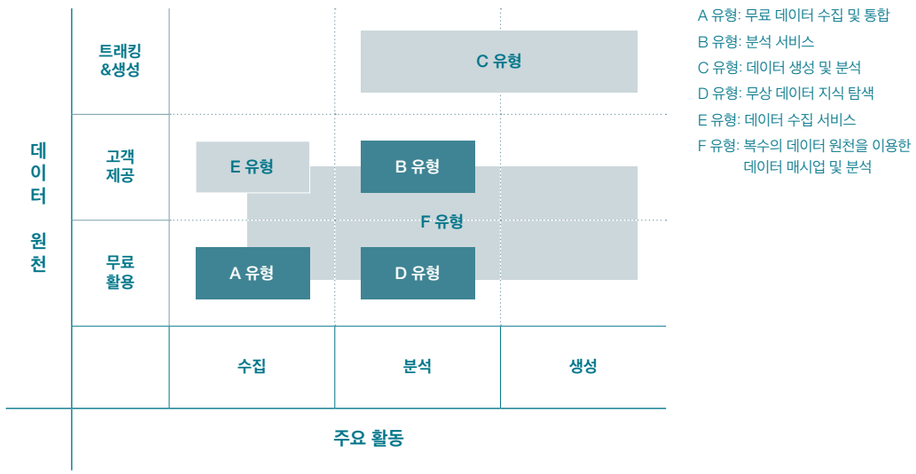
4. 데이터 기반 스타트업의 비즈니스 모델 패턴

독일 칼슈헤 공대(Karlsruhe Institute of Technology)의 하트만 등이 캠브리지대학교(University of Cambridge)의 캠브리지 서비스 얼라이언스(Cambridge Service Alliance)에 데이터 기반 스타트업 100개를 조사해 비즈니스 모델 패턴을 발표했다. 스타트업들의 비즈니스 방식을 데이터 원천, 핵심 활동, 오퍼링, 목표 고객, 수익 모델, 특별한 원가 혜택 등의 관점으로 분류했다. 클러스터링을 통해서 최종적으로 두 가지 축을 기준으로 여섯 가지 비즈니스 모델 패턴을 도출했다.*

핵심 데이터 원천은 무료로 확보 가능한 데이터, 고객이 제공하는 데이터, 추적 및 생성되는 데이터의 세 가지 차원으로 구분했고, 핵심 활동은 데이터 수집, 분석, 데이터 생성 등의 세 가지 차원으로 구분했다. 랜덤하게 선정된 100개 스타트업의 비즈니스 모델을 매트릭스에 매핑한 결과 여섯 개의 비즈니스 모델 패턴이 도출됐다.

* Hartmann, Philipp Max, Mohamed Zaki, Niels Feldmann and Andy Neely (2014), *Big data for big business? A taxonomy of data-driven business models used by start-up firms*, University of Cambridge

[그림 1-2-2] 하트만의 데이터 기반 비즈니스 모델 매트릭스



출처: 「주요국 데이터 주도 비즈니스 현황 및 비즈니스 유형 비교 분석」, 한국데이터진흥원, 2016. 9

패턴 1: 무료 데이터 수집 및 통합

대부분 무료로 획득할 수 있는 데이터들을 수집해 가공한 후 API 또는 웹대시보드를 이용해 유통시키는 비즈니스 모델이다. 데이터 원천은 주로 소셜 데이터(65%), 크롤링과 시각화가 주된 활동으로, 소비자에게 지역의 식당이나 바를 소개하는 등의 비즈니스를 말한다.

패턴 2: 분석 서비스

고객이 제공하는 데이터를 분석해주는 비즈니스 모델로, 부정사용 분석, 상권 분석 등 부분적으로 분석 결과를 API를 통해 유통시키거나 시각화해 제공하기도 한다. 대부분의 분석 서비스(Analytics as a Service) 기업은 기술 분석(descriptive analytics)을 제공하고 있으며, 일부만 예측 분석(predictive analytics) 서비스를 제공한다.

패턴 3: 데이터 생성 및 분석

기존 데이터를 수집해 제공하는 것이 아니라 스스로 데이터를 생성해 공급하는 비즈니스 모델이다. 생성한 데이터를 기반으로 분석 서비스와 함께 제공한다. 크라우드소싱으로 데이터 확보, 웹 분석 기업, 스마트폰 또는 물리적 센서 장비를 이용한 데이터 생성 기업 등의 세 가지 유형이 있다.

패턴 4: 무상 데이터 지식 탐색

무료로 획득할 수 있는 데이터를 활용해 분석 결과를 제공하는 비즈니스 모델이다. 예를 들면 소프트웨어 개발자의 능력 확인을 위해 오픈소스 포탈 등의 사이트 활동을 검색해서 알려주는 서비스를 제공한다.

패턴 5: 데이터 수집 서비스

고객으로부터 데이터를 직접 수집해 제공하는 비즈니스 모델이다(Data aggregation as a Service). 개인 또는 조직으로부터 데이터를 확보해 다양한 형태로 데이터를 유통시키거나 시각화해 제공한다.

패턴 6: 복수의 데이터 원천을 이용해 데이터 매시업 및 분석

무료로 이용 가능한 데이터와 고객으로부터 획득할 수 있는 데이터 등을 통합해 데이터를 제공하거나 분석해 제공하는 비즈니스다. 디지털 마케팅 지원을 위해 기업의 고객 데이터와 웹크롤링 등으로 확보한 외부 데이터를 융합해 제공한다. 타기팅(targeting), 추천, 성과 분석 등의 분석 활동도 수행한다.

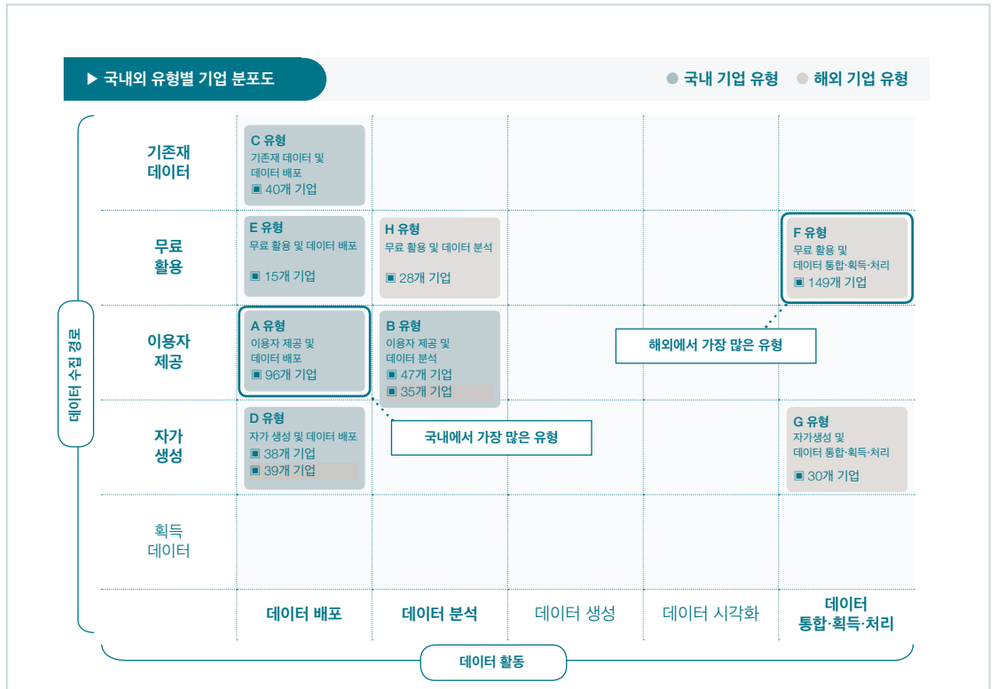
한국데이터진흥원은 2016년에 비석세스에 의뢰해 국내외 데이터 기반 스타트업들의 비즈니스 모델 패턴을 분석했다.* 분석 대상은 국내 기업 310개와 해외 기업 379개이다. 비즈니스 모델 패턴은 하트만 등이 발표한 보고서의 프레임워크를 기준으로 매트릭스를 만들었다. 데이터 소스는 기존 데이터 활용, 무료 데이터 활용, 이용자 제공, 자가 생성, 획득 데이터 등으로 구분했다. 데이터 활동은 데이터 배포·분석·생성·시각화·통합·획득·처리 등으로 구분했다. 국내외 데이터 기

* 한국데이터진흥원, 「국내외 데이터 주도 비즈니스 현황 비교」, Data Issue Report 2017-01 제105호, 2017

반 기업 689개를 매핑한 결과 8개의 비즈니스 모델 패턴을 도출했다. 특이한 점은 국내와 해외의 데이터 기반 비즈니스 방식이 상당 부분 다르다는 점이다. 국내에서 가장 많이 나타나는 유형은 '이용자 제공 데이터+데이터 배포'이며, 해외에서 가장 많이 나타나는 유형은 '무료 데이터+데이터 통합·처리'이다.

국내와 해외의 데이터 기반 스타트업 비즈니스 모델 분포가 차이가 나는 것은 두 가지 원인이 있다. 첫째, 우리나라는 데이터 규제가 세계적으로도 높은 수준이다. 예를 들어 개인데이터를 임의로 수집하는 행위는 금지돼 있으며 데이터 유통도 엄격하게 제한돼 있다. 따라서 개인정보를 획득하기 위해 무료 데이터를 수집해 배포하는 스타트업은 우리나라에서는 등장할 수 없는 여건이다. 둘째, 상대적으로 우리나라는 데이터 활용 산업이 아직 성숙되지 못했다. 데이터가 필요한 기업은 스스로 데이터를 수집해 활용해야 한다. 데이터 분석을 대신해 수행하는 생태계도 갖추어지지 않았다. 그 결과로 제한된 범위에서 데이터 기반 스타트

[그림 1-2-3] 한국데이터진흥원의 데이터 기반 비즈니스 모델 매트릭스



출처: https://www.datastore.or.kr

업들이 활동하고 있다.

앞으로는 데이터 기반 스타트업들이 우리나라에도 빠른 속도로 확산되고 데이터 관련 규제는 정부 주도로 완화될 전망이다. 또한 이용할 수 있는 데이터 범위도 확장되고 있다. 정부의 공공데이터 개방 정책과 데이터 활용 시범 사업 투자 확대는 데이터 기반 스타트업들의 등장을 촉진시킬 것이다. 기존 기업 차원에서도 데이터 기반 스타트업들이 다양하게 등장하는 것은 긍정적 요인이 더 크다. 데이터 생산부터 가공·분석·유통 등의 산업 구조가 정착되면 기존 기업들은 더 쉽게 데이터를 활용할 수 있게 되기 때문이다.

5. 데이터 기반 비즈니스 모델 전망

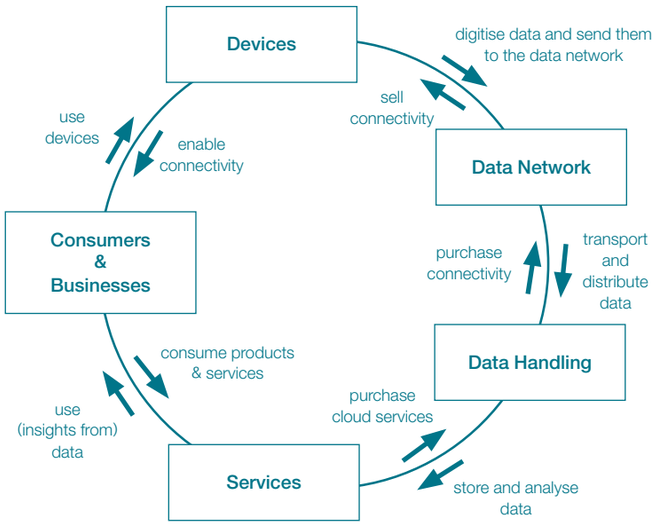
독일의 아놀드(Rene C.G Arnold)는 데이터 가치 순환도(Data Value Circle)를 제시했다. 데이터는 디바이스에서 발생해 네트워크를 통해 수집된다. 데이터 관리를 거쳐서 서비스에 활용되며 결과적으로 고객의 경험과 기업의 프로세스에 긍정적 효과를 미친다. 고객 경험과 프로세스의 발전은 다시 디바이스의 진화로 이어진다.*

데이터 기반 비즈니스 모델은 데이터 가치 순환을 강화시키는 방향으로 발전할 것이다. 디바이스로부터 데이터를 획득하기 위해서는 데이터 네트워크 서비스를 통해 연결해주는 비즈니스가 필요하다. 데이터를 서비스에 활용하기 위해서는 데이터를 보관하고 처리하기 위한 인프라가 필요하다. 데이터 클라우드 서비스는 하나의 예이다. 고객 경험과 비즈니스 프로세스 혁신을 위해 데이터와 분석 서비스 수요가 증가할 것이다. 데이터 기반으로 고객 경험을 혁신하기 위해서는 가상현실, 웨어러블, 드론, 자율주행자동차 등 혁신적인 디바이스 수요를 증대시킬 것이다.

데이터 기반 비즈니스 모델은 개별 기업 만의 이슈는 아니다. 데이터 생태계는 기존 기업의 비즈니스 혁신에 필수 요소이기 때문이다. 또한 데이터 기반 비즈

* Arnold, Rene C.G and Waldburger (2016), *The impact of data on ICT business models*, WIK Wissenschaftliches Institut für Infrastruktur und Kommunikationsdienste GmbH

[그림 1-2-4] 데이터 가치 순환도(The data value circle)



출처: Arnold, Rene C.G and Martin Waldburger(2016), *The impact of data on ICT business models*

니스 모델은 융합을 통해 가치 네트워크를 형성하게 된다. 이는 고객에게 제공하는 서비스 수준의 획기적 향상과 참여 기업의 기업 성과 창출에도 크게 기여할 것이다. 개별 기업은 데이터 경제 성숙에 대응하기 위한, 산업 및 국가 차원에서는 새로운 산업 창출을 위한 필수 요소이다. 또한 데이터 기반 비즈니스 모델은 스타트업은 물론, 기존 기업에도 중요한 미래 과제이다.

제 3 장

지능정보시대의 데이터 전문가 전망

제4차산업혁명에 대한 관심이 세계적으로 올라가고 있다. 2012년을 전후해 국내에 소개될 당시만 해도 빅데이터는 곧 버즈워드로 전락할 것이라는 의견이 있었다. 하지만 우려와는 달리 빅데이터는 제4차산업혁명의 기반으로 확실히 자리매김을 해가고 있다. 본 장에서는 제4차산업혁명의 성공에 중요한 역할을 할 데이터 전문가, 그 중에서도 데이터 분석 전문가를 점검해본다.

1. 인공지능 기술과 데이터 전문가

국내에서 빅데이터 전문가에 대한 수요는 여전히 높다. 한편, 글로벌 ICT 기업들의 인공지능 솔루션을 도입하면 인공지능과 관련한 새로운 이슈들을 해결할 수 있을 것이라는 분위기가 다시 팽배해 지고 있다.

분명한 것은 인공지능(Artificial Intelligence, AI)은 딥러닝과 같은 특정 알고리즘이나 대용량 데이터 처리 등의 테크놀로지 이슈가 전부는 아니라는 점이다. 현재 많은 국내 기업에서 고객과의 전화상담 내용을 텍스트 데이터로 전환·분석해 고객에 대한 이해를 높이려는 노력을 기울이고 있다. 하지만 한국어의 인식률은 영어 등 일부 외국어의 인식 수준인 95% 수준에 이르지 못했다. 필자가 소속된 회사에서도 인공지능 프로젝트를 수행하면서 음성인식 전문가를 찾지 못해 자체 육성하기로 결정했다. 인공지능 기술은 정교한 학습 알고리즘이나 빅데이터 처리 기술 등의 발전과 궤를 함께하지만, 이런 기술들을 이해하고 비즈니스에 접목시켜줄 수 있는 데이터 전문가를 필요로 한다.

* 필자: 이종석(신한카드 빅데이터센터장)

인공지능 기술과 관련한 데이터 전문가들을 육성하기까지 상당 기간이 필요하다. 그럼에도 사회적인 관심은 공공데이터 공개와 개인정보 규제, 데이터 확보 등 법적인 이슈와 원칙론 중심에 머물고 있다. 물론 머신러닝, 인공지능 등의 학습에 필요한 데이터를 확보하는 것은 매우 중요한 일이다. 하지만 많은 데이터를 확보하더라도 이를 기업이 추구하는 비즈니스에 맞게 해석해내거나 필요한 분석 알고리즘을 만들어 내지 못한다면, 궁극적으로 도래할 제4차산업혁명 시대의 국내 기업들은 공급자보다는 소비자 역할에 머물 수밖에 없다.

모바일 기기용 운영체제나 DBMS의 예를 들어보자. 이 영역은 고부가가치 영역이지만 여전히 국내 기업들은 소비자에 가깝다. 국내에서도 운영체제나 시스템 소프트웨어를 국산화하기 위해 기업과 정부 차원에서 노력하고 있지만, 글로벌 선발업체들과 보조를 맞추기가 쉽지 않은 상황이다.

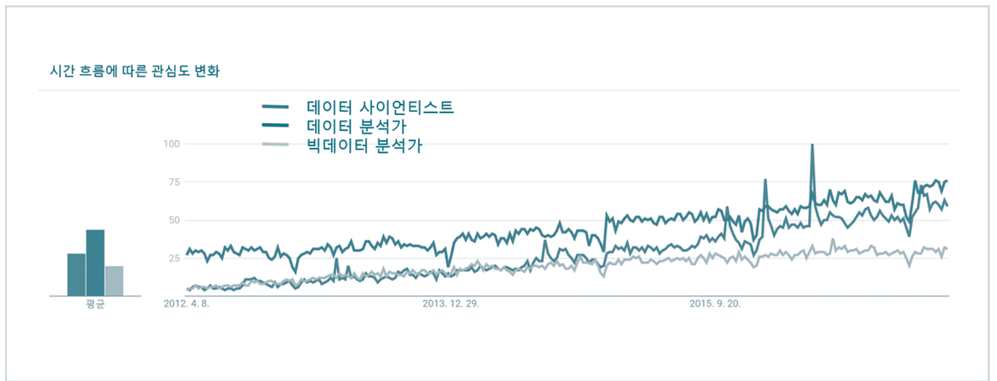
2. 표준이 없는 인공지능 기술

운영체제나 DBMS 영역에서 국내 기업들이 기회를 잡지 못했다고 해서 국내 비즈니스에 큰 영향을 주지는 않았다. ICT 생태계에서는 표준화한 시스템을 판매하는 기업이 별도로 있었기 때문이다. 하지만 제4차산업혁명 시대에는 표준화한 핵심 기술을 판매하는 기업은 아마도 없을 것이다.

지난해 말부터 눈길을 끌고 있는 '아마존고(Amazon Go)'의 예를 들어 보자. 아마존고는 스마트폰, 인간의 시각에 해당하는 비전 센서(vision sensor), 인간의 뇌에 해당하는 인공지능의 세 가지 제4차산업혁명 시대 기술을 활용해 무엇을 구매했는지 기록하고 최종 계산을 해주는 무인점포 시스템이다. 그 성공여부는 지켜봐야겠지만, 업무 흐름 자체를 바꿔 놓을 인공지능 기술은 특정 표준이 아닌 자체 개발할 수밖에 없음을 보여준다는 점에 주목할 필요가 있다. 아마존고가 확실히 성공을 거둔다면, 아마존이 이 기술을 경쟁기업에게 공급할지 의문이다. 경쟁기업들은 동일한 기술 수준을 확보하는 데 상당한 기간이 소요될 것이고 결론적으로 경쟁에서 뒤쳐질 것이다.

[그림 1-3-1]에서 보여주듯이 빅데이터 전문가에 대한 수요는 계속 늘어날 전망이다. 국내 기업들도 곧 이러한 흐름에 동참할 수밖에 없다.

[그림 1-3-1] 데이터 전문가 관련 구글 트렌드 분석(2017.04.03 기준)



3. 데이터 전문가 유형 및 필요 기술

데이터 전문가는 빅데이터와 인공지능의 부각에 따라 그 범위가 확장되고 있다. 한국데이터진흥원의 「데이터 전문인력 양성 방안 연구 결과 보고서(2016. 9)」에 수록된 '데이터 전문인력의 직무 및 역량' 기준을 따르면 국내외의 기관에서 DBA, DB 개발자, 데이터 컨설턴트 등 전통적인 데이터 직무에, 데이터 사이언티스트, 빅데이터 개발자, 빅데이터 엔지니어 등 빅데이터 관련 직무까지 추가하고 있음을 알 수 있다. 국내외를 막론하고 데이터 관련 직무와 그 명칭, 구분 형태도 다르다. 이는 데이터산업이 급격히 확장되면서 나타나는 현상으로 풀이된다. 이를 비교한 결과가 다음 페이지의 [표 1-3-1] 이다.

2017년 현재 자주 회자되는 빅데이터, 머신러닝, 인공지능 등은 모두 데이터 전문가가 필요한 영역으로 상호 연관성이 높다. 이 세 가지 영역이 어떻게 상호작용하는지 먼저 알아본 다음, 데이터 전문가 유형과 필요 기술·경력에 대해서 설명하겠다.

인간은 지능을 다양한 문제 해결을 위해 활용한다. 그러기 위해 인간은 오랜 학습 기간을 필요로 하며 보거나 듣거나 만지거나 냄새를 맡거나 맛을 음미하여 주변의 다양한 상황들을 인지하면서 학습한다. 인공지능도 문제 해결을 위해 인간과 마찬가지로 다음의 세 가지 사항을 기반으로 학습한다. 그 결과 일부 영역에서는 인공지능이 인간을 앞서는 결과를 도출해 내기도 한다.

[표 1-3-1] 국내외 데이터 직무 구분 비교표

구분		한국데이터진흥원	한국정보화진흥원	고용노동부	노규성·박성택·박경혜	DATA FLOQ	SAS
데이터 설계		데이터 아키텍트	하둡/NoSQL /MapReduce /엔지니어	DB 엔지니어	빅데이터 관리 및 기술(지원)	빅데이터 솔루션 아키텍트	빅데이터 아키텍트
데이터 개발		데이터 개발자				빅데이터 엔지니어	빅데이터 개발자
데이터 운영관리		데이터 엔지니어					
데이터 활용	데이터 분석	데이터 분석가/ 데이터 사이언티스트	빅데이터 분석가	빅데이터 분석	빅데이터 분석	빅데이터 분석가/ 빅데이터 사이언티스트	빅데이터 분석가/ 데이터 사이언티스트
	데이터 시각화					빅데이터 비주얼라이저	빅데이터 디자이너
	데이터 연구					빅데이터 연구원	데이터 사이언티스트
데이터 관리		데이터 사이언티스트/ DBA			빅데이터 활용/ 데이터 관리 및 기술(지원)	총괄책임자/ 빅데이터 매니저/ 빅데이터 사이언티스트	프로젝트 매니저/ 데이터 사이언티스트/ 빅데이터 관리자
데이터 판매		데이터 마케터	빅데이터 기술영업/ 마케터	-	-	-	-
데이터 컨설팅		데이터 컨설턴트	빅데이터 컨설턴트	-	빅데이터 기획	빅데이터 컨설턴트	-

출처: 「데이터 전문인력 양성 방안 연구 결과 보고서」, 한국데이터진흥원, 2016

① 인간 오감과 유사한 인지를 할 수 있는 센서로부터 수집한 학습에 필요한 빅데이터

② 이 데이터를 다양한 방법으로 대입하여 학습하는 머신러닝

③ 다양한 학습을 하여 얻은 여러 가지 예측들 중 최적의 안을 결정할 수 있는 의사결정 관련 추론 알고리즘

머신러닝이나 최적의 의사결정을 위한 추론 알고리즘에 필요한 전문가는 해당 알고리즘에 대한 전문지식과 실무 적용 경험이 필수적이다. 이 영역의 데이터 전문가들조차도 관련 학계에서 나온 이론들에 대한 지식을 다 갖추기는 어렵다. 그래서 필자의 현장 경험과 국내 상황을 고려해 대부분의 기업들이 필요로 하는 범용성을 기준으로 4가지 전문가를 구분·선정했다. 데이터 전문가로 활동하려는 사람이 일정 기간의 노력을 기울이면 충분히 해당 분야의 전문가가 될 수 있다고 판단한 유형들이다. 물론 관점에 따라 다른 의견을 제시할 수 있다. 현재 상황에서 향후 3년 정도를 고려한 기준이다. 앞으로 관련 기술이 발전하면 범용적 데이터 전문가들보다는 특정 영역(domain)에서 전문성을 가진 데이터 전문가들이 점점 더 늘어날 것으로 예상된다.

제시할 네 가지 데이터 전문가에 인공지능, 머신러닝, 빅데이터 등 신기술을 적용하기 위해서는 빅데이터 플랫폼이 매우 중요하다. 현재 클라우데라, 호튼웍스, 맵알 등 글로벌 빅데이터 플랫폼 업체들이 이 시장을 리딩하는 모습이다. 국내에 관련 스타트업들이 등장했다가 크게 성공하지 못했다. 이 때문인지 글로벌 상용 빅데이터 플랫폼을 도입한 국내 기업들이 자체 개발 엔지니어 부족으로 필요한 기능을 제대로 활용하지 못하는 모습을 볼 수 있다. 따라서 빅데이터 수집·저장 관련 솔루션 전문가도 유형에 포함했다.

가. 데이터 사이언티스트

데이터 사이언티스트를 단순히 통계 또는 알고리즘 전문가로 오해하는 것은 상당히 안타까운 일이다.

호주의 한 은행이 페이스북에 반려견 사진을 올려놓은 고객들을 찾아 반려견 보험을 타깃 마케팅하는 기사를 본적이 있다. 기업 내 분석인력들은 '그럴 수도 있겠네' 하고 생각하겠지만 정작 자신들은 그러한 데이터가 타기팅에 유효한 데이터가 될 수도 있다는 것을 생각하지도 못하는 경우가 많다.

예를 들면, 종양의 크기에 따른 암 여부에 대한 데이터만 데이터 웨어하우스에 있다고 가정해보자. 기존 데이터만을 활용해 예측한 결과, 정확도가 60% 수준이었다고 하면 정확도 개선을 요구받을 것이다.

데이터 사이언티스트는 '데이터 웨어하우스에 적재된 데이터 중 결합을 통해 새로운 변수를 찾을 수 있는지', '서류상으로 존재하는 데이터들 중 암 진단에 유효한 데이터가 있는지' 등 먼저 종양크기 외의 데이터가 없는지 찾아 나설 것이다. 그런 노력들을 통해 뜻하지 않게 환자의 연령이 종양의 크기에 따른 암 진단 여부에 영향을 줄 수 있다는 새로운 사실을 발견해 낼 수도 있다.

그렇다고 무턱대고 수많은 데이터를 다 찾기는 어렵기 때문에 데이터 사이언티스트는 분석에 영향력이 높은 데이터들을 찾아 낼 수 있는 다양한 차원 축소, 공간 매핑, 시각화, 패턴 인식, 시그널 프로세싱 등 여러 학문 영역에서 데이터 분석에 활용할 수 있는 다양한 기법을 이해하고 있어야 한다.

공간 매핑의 예를 들면, 시계열 데이터 분석을 시간축이 아닌 주파수축에서 그려본다면 시간축에서 일정 주기로 파동을 나타내는 데이터들이 주파수축에서 일정한 주파수를 가진 막대로 표현될 것이다. 즉 시간축에서 여러 주파수의 혼합으로 나타나는 데이터들을 특정 주파수 단위로 구분하면, 문제해결에 좀 더 중요한 변수를 쉽게 발견해 낼 수도 있을 것이다. 혹자는 이러한 차원 축소 기법이나 공간 매핑 기법의 단점으로 실세계(real world)의 데이터가 소실되어 분석 결과가 정확히 나왔더라도 그 의미를 알 수 없는 블랙박스가 되어버릴 수 있다고 주장한다. 하지만 변형된 공간에서 학습된 결과를 다시 실세계의 데이터로 환원시켜주는 다양한 기법들이 존재한다. 이미 20년 전부터 관련 연구는 많았다.

●○
관련 기술이 발전하면 범용적 데이터 전문가들보다는 특정 영역(domain)에서 전문성을 가진 데이터 전문가들이 점점 증가할 것이다.

나. 빅데이터 수집·저장 솔루션 개발 전문가

빅데이터의 수집·저장과 관련해 이미 많은 기술이 오픈소스로 공개돼 있다. 문제는 이러한 오픈소스들이 특정 목적을 위해 만들어지다 보니, 이를 한 데 모아 기업의 목적에 맞게 구성하려면 대단히 어려운 일이 되고 만다. 결국 몇몇 글로벌 기업들이 이 문제를 해결한 상용 솔루션을 내놓았고 도입 사례도 생겼다. 국내에도 이런 기술을 가진 일부 스타트업들이 2012년 초반부터 왕성하게 활동했으나 현재는 그 수가 많이 줄어들었다. 기술에만 매달린 나머지 도입하려는 기업들의 니즈에 대한 이해가 부족했기 때문이다. 이 영역의 솔루션 개발 전문가가 되기 위해서는 현재까지 나와 있는 다양한 빅데이터 오픈소스들을 직접 개발·운영해본 경험과 지식이 필요하다. 관련 기업들은 전문가 커뮤니티 후원 등 상호 소통하는 문화 조성에 관심을 기울일 필요도 있다.

현재는 글로벌 기업들이 다양한 경험을 통해 고객의 니즈가 무엇인지 어느 정도 파악했고 상당한 기술적 발전을 이룬 상태다. 그럼에도 국내 기업들은 이러한 글로벌 상용 솔루션을 도입하고도 제대로 활용을 못하는 실정이다. 그 이유는 도입 기업마다 다를 수 있겠지만 필자의 경험을 따르면, 기업 내부에 오픈소스를 기반으로 하는 분석 툴들을 제대로 활용할 수 있는 인력이 부족하다는 점이다. 따라서 기존의 데이터 웨어하우스의 데이터를 분석하는 데 익숙한 분석가들에게 어떤 식으로 빅데이터 플랫폼을 활용하게 할 것인지 다리 역할을 하는 빅데이터 솔루션 개발 전문가가 필요하다. 기업의 IT 조직에서 일한 경험자가 빅데이터 수집·저장 관련 오픈소스 지식을 갖춘다면 충분히 이 역할을 수행해 낼 수 있다.

다. 음성 인식·이미지 인식 관련 알고리즘 개발 전문가

많은 글로벌 기업들이 음성 인식이나 이미지 인식 등의 기술을 가진 스타트업들을 왜 인수할까? 글로벌 기업들의 인공지능 개발 현황을 보면 인터넷이나 소셜 네트워크에서 획득할 수 있는 데이터만 활용한 경우가 많다.

기업 업무들 가운데 특정 기준이나 규칙을 기준으로 판단하는 일이 상당히 많다. 특정 기준에 의거한 판단은 이미 딥러닝 등 다양한 머신러닝 기법들이 어느 정도까지는 대체 가능하다는 것을 입증하고 있다. 그렇다면 필요한 것은 업무담당자처럼 동일하게 업무에 필요한 데이터를 보고 들을 수만 있으면 된다. 음성 인식과 이미지 인식 영역에 필요한 기술은 시그널 프로세싱, 이미지 프로세싱, 패턴 분석, 최적화, 데이터 압축, 머신러닝 기법(Hidden Markov Model, 딥러닝) 등이다.

라. 음성 인식·이미지 인식 학습 전문가

훌륭한 음성 인식 혹은 이미지 인식 솔루션을 도입하더라도 더 정교한 결과를 내기 위해서는 지속적인 학습이 필요하다. 하지만 이런 학습을 매번 개발업체에 의존할 수만은 없다. 문제는 개발업체는 해당 기업의 비즈니스나 활용 목적에 대한 이해가 부족하다는 점이다. '역량 있는 학생'도 중요하지만 '역량 있는 선생님'이 있어야 그 학생의 숨은 역량을 끌어낼 수 있는 것과 같다. 학습전문가가 되기 위해서는 업무에 대한 이해도가 높으면 높을수록 유리하다. 학습을 위해서는 누가 학습된 내용이 맞는지 틀린지에 대해 가르쳐 주어야하는데 결국 기업 내 해당 업무를 맡고 있는 직원들이 이 역할을 해주어야 한다. 이때 정확한 피드백을 얻기 위해서는 학습 전문가가 담당 직원들에게 정답과 오답을 구분할 수 있는 방법을 이해하기 쉽게 설명해 주는 것이 필수적이다.

인공지능 최종 수요기업들이 전문인력을 확보하지 않은 채로 솔루션 업체에 일임하다 보니 나중에 활용할 때 부족한 부분이 생기고 그때마다 추가 개발을 해야 하는 상황에 놓이게 된다. 기업에서 외부 업체의 솔루션을 도입하기 전에 다양한 음성 혹은 이미지 인식 전문 업체와 파일럿 프로젝트를 추진하면서 기업의 업무를 잘 아는 인력을 투입한다면 학습 전문가는 의외로 쉽게 육성할 수 있다.

이런 경험을 한 인력들은 향후 도입한 시스템의 개선 방향에 대해서도 효율적인 방안을 제시할 수 있을 것이다.

2016 데이터 구루상 수상자

“공유·활용의 작은 본보기를 보여주고 싶다”

이경일 | 솔트룩스 대표

“인공지능은 알고리즘뿐 아니라 데이터가 궁극적 차이를 만드는 게임이다. … 데이터를 연료로 하는 인공지능의 엔진은 충분한 양질의 데이터라는 원료를 요구한다. 우리나라에서 공공데이터의 활용이 낮은 이유도 데이터 품질과 가독성 문제 때문이다. 공개 데이터를 hwp나 pdf 대신에 csv나 rdf로 올려도 사용자들에게는 큰 힘이 된다.”

2016 데이터 구루상을 수상한 소감은.

앞서 수상했던 선배 수상자들만큼 내가 칭찬받을 만한 일을 했는지, 나 스스로에게 물어보게 되고 책임감을 느낀다. ‘데이터산업 발전을 위해 작은 밀알 역할을 했느냐?’에 대한 자기반성과 함께 더 무거운 사명감을 가지게 되었다.

데이터와 어떻게 인연을 맺게 되었는가.

소프트웨어는 데이터와 뗄 수 없는 관계다. 대학생 때, 직접 만든 소프트웨어를 판매해 수익을 냈던 경험을 갖고 있다. 그 경험으로 비교적 일찍 소프트웨어 사업을 했는데, 사업 초기부터 ‘자연어 처리’ 소프트웨어를 개발하고 싶었다. 당시에는 인공지능이라는 말을 앞세우지 못했지만, 지금에 와서 보면 그때 바랐던 길을 걷고 있음을 느낀다.

자연어 처리 방법을 그때와 지금을 비교해 보면.

기호적/비기호적 차이라고 할 수 있다. 창업할 당시에는 기호적 처리, 즉 사전과 규칙을 만들어 그걸 기반으로 자연어를 처리했다. 사람이 하나하나 규칙을 만들어줬으므로 기호적 접근이라고도 한다. 최근 딥러닝을 포함한 머신러닝은 원천데이터(raw data)가 풍부해 이를 재료로 자연어를 처리하는 비기호적 접근을 한다. 머신러닝과 인공신경망(딥러닝)이 이러한 접근 방식이다. 자연어 처리도 최근에는 기호적 접근과 비기호적 접근을 융합하는 형태로 발전되고 있다. 자연어 처리도 구글 ‘알파고’처럼 인공지능 엔진이 스스로를 복제해 셀프 플레이로 (자연어 처리의) 정확도를 높인다.

자연어 처리와 더불어 시맨틱 검색 엔진 사업을 꾸준히 해왔는데, 그것에 대한 공로를 인정받은 것이 아닌가 싶다.

처음에 생각했던 사업 영역에서 벗어나지 않고 지켜온 것을 자랑스럽게 생각한다. 다음 주(2017년 4월 3주차)에 국내외에서 수집한 소셜 데이터 80억 건과 흩어져 있던 국내외 공공데이터를 모아서 공개할 계획이다. 이것이 건강한 데이터 생태계 유지를 위한 작지만 건강한 본보기가 될 것으로 기대한다.

2017년의 데이터 분야 최고 이슈는 무엇이라고 보나.

데이터의 양과 질에 대해 말하고 싶다. 다보스포럼에서도 제4차산업혁명의 핵심을 데이터라고 보고, 국가별로 위키피디어에 공개된 데이터 기준으로 조사한 결과, 한국은 데이터 양과 질에서 25~27위로 평가되었다. 데이터는 문화와 관련돼 있으므로 단기간에 순위를 끌어올리기가 쉽지 않다. 그동안 한국이 사용해온 ‘패스트 팔로우’ 전략은 다른 나라에서 20여 년에 걸쳐 이룩해 놓은 일을 2년 만에 달성하는 기반이 되었지만, 데이터에서는 이 전략이 잘 통하지 않는다. 현재 필요한 것은 공유·활용이라는 협력적 문화다. 하지만 우리 주변을 보면, 내가 만든 유무형의 자산을 누구에게 (무료로) 공개한다는 것에 관대하지 못한 것 같다. 인공지능은 알고리즘뿐 아니라 데이터가 궁극적 차이를 만드는 게임이다. 물리학에서 우주선이 지구의 중력권을 빠져나가려면 충분한 원료를 태워서 강력한 힘을 내야하듯이, 데이터를 연료로 하는 인공지능 엔진은 충분한 양질의 데이터라는 원료를 요구한다. 우리나라에서 공공데이터의 활용이 낮은 이유도 데이터 품질과 가독성 문제 때문이다. 공공데이터는 한마디로 품질이 낮다. 공개 데이터를 hwp나 pdf 대신에 csv(comma-separated values)나 rdf(resource description framework)로 올려도 사용자들에게는 큰 힘이 된다. 정부 차원에서 데이터가 중요하다는 사실을 보여주는 마중물을 부어서 데이터를 지상으로 뽑아 올리는 노력이 필요하다.

인공지능 현상을 어떻게 보는가.

첨단기술의 성과만 보면 과다한 기대를 하게 되는데, 인공지능도 그 기대만큼 큰 실망을 안겨줄 수 있다. 기술의 성숙도를 표현하기 위한 시각적 도구로 자주 인용되는 가트너의 하이프 사이클(Hype Cycle) 주기를 기준으로 보면, 현재 인공지능(AI)은 최고점에 도달해 있는 거 같다. 2~3년 안에 하강 국면을 걷다가 일반인들의 관심권에서 벗어날 때쯤 국내외 시장에서 강력한 AI 주자들이 자리를 잡을 거라고 본다.

데이터 영역의 입문자들에게 한마디 한다면.

데이터 영역의 소프트웨어 엔지니어가 된다면, 인공지능에게 일자리를 빼앗기지 않을 것이다. 다만 데이터 영역에서 뜻하는 성과를 내려면 1~2년의 노력만으로는 어렵다. 중장기적 안목을 갖고 자신에게 맞는 영역을 파고들라고 추천한다. 짧게는 6~7년, 적어도 10년은 데이터 영역에서 일을 해야 기술·경험적으로 인정을 받을 수 있다. 10년에 걸쳐 높은 수준에 이르도록 한우물을 팔 수 있어야 하고, 오랜 기간 버틸 수 있는 정신력도 필요하다. 한국 정부도 ‘데이터 시대’를 과거 ICT 시절처럼 금방 황금알을 낳는 영역으로 보지 않고 길게 보기 시작할 것이다. 가까운 중국도 미국처럼 장기적으로 보고 (데이터 영역에 대한) 투자를 아끼지 않고 있다. 정부가 앞장서서 단기적 성과에 집착하지 않고 뛰어놀 수 있는 데이터 무대를 만들기 위해 노력할 것이라고 기대한다.

제 **2** 부

데이터산업 시장 현황

제1장

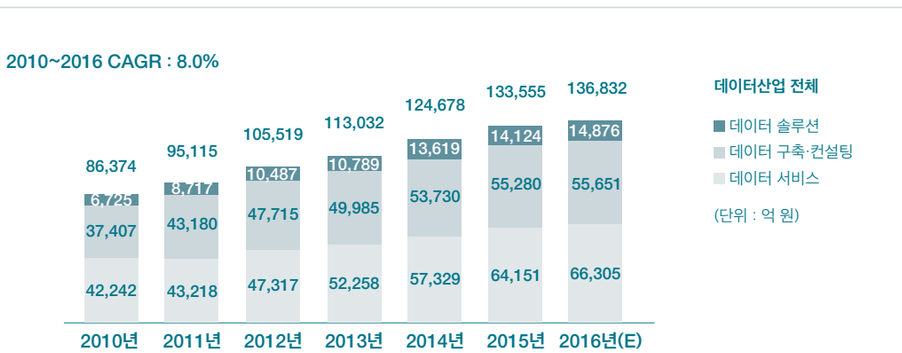
국내 데이터산업 현황

본 장의 국내 데이터산업 현황은 「2016년 데이터산업 현황 조사(미래창조과학부, 한국데이터진흥원, 2016.4)」 결과보고서를 요약·발췌해 작성했다. 데이터산업 현황 조사는 데이터산업을 '데이터의 생산·수집·처리·분석·유통·활용 등을 통해 가치를 창출하는 상품과 서비스를 생산·제공하는 산업'으로 정의하고 있다. 이러한 정의에 따라 데이터산업의 비즈니스 유형을 크게 데이터 관련 제품을 판매하거나 기술을 제공하는 데이터 솔루션, 데이터 구축, 데이터 컨설팅 비즈니스와 데이터를 기반으로 서비스를 제공하는 데이터 서비스 비즈니스로 구분했다. 데이터산업의 시장 규모는 이러한 데이터 비즈니스를 영위하는 사업체들의 데이터 관련 직간접적 매출을 조사해 추정한 결과다. 참고로 데이터산업 현황 조사는 매년 조사 시점 매출액은 예상 매출액으로 산출하고, 차기년도에 매출액 확정치를 조사해 반영하고 있다. 이에 본 백서 상의 통계(2016년 조사결과)와 향후 발표될 2017년 데이터산업 현황 조사 결과가 다를 수 있음을 미리 밝힌다.

1. 국내 데이터산업 시장 규모

2016년 데이터산업 시장 규모는 13조 6,832억 원으로 2015년 대비 2.5% 성장했으며, 2010년 이후 연평균 증가율 8.0%로 매년 꾸준한 성장세를 유지하는 것으로 나타났다.

[그림 2-1-1] 2010~2016년(E) 데이터산업 시장 규모



출처: 「2016년 데이터산업 현황 조사」, 미래창조과학부·한국데이터진흥원, 2016. 4(이하 동일)

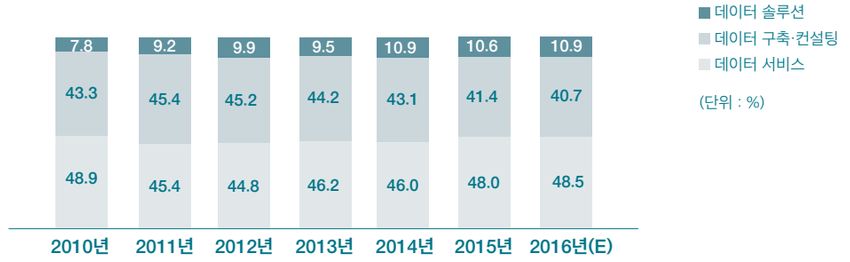
* 필자: 하진희(한국데이터진흥원 정책기획실 선임연구원)

부문별로는 2016년 데이터 솔루션 시장이 1조 4,876억 원, 데이터 구축·컨설팅 시장이 5조 5,651억 원, 그리고 데이터 서비스 시장이 6조 6,305억 원으로 조사됐다. 부문별 비중으로는 2016년 기준으로 데이터 서비스 시장이 48.5%로 가장 큰 비중을 차지했고, 이어서 데이터 구축·컨설팅 시장이 40.7%, 데이터 솔루션 시장이 10.9% 순으로 나타났다.

[표 2-1-1] 2010~2016년(E) 데이터산업 부문별 시장 규모 (단위: 억 원, %)

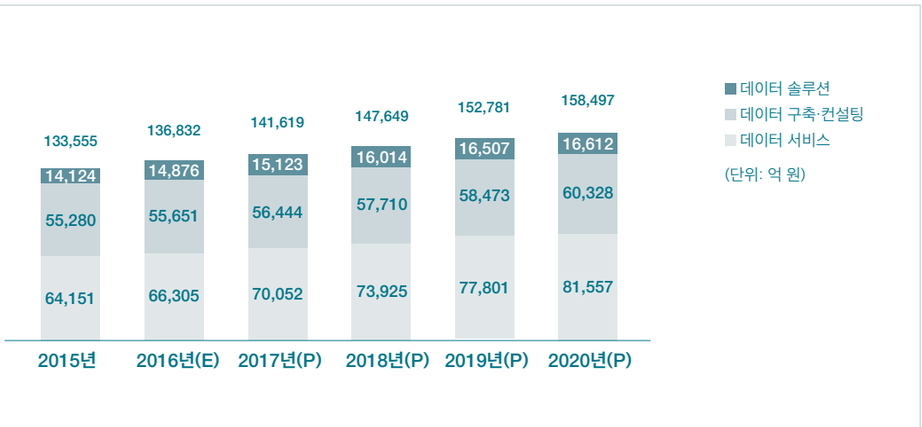
구 분	2010년	2011년	2012년	2013년	2014년	2015년	2016년(E)	증감률 '15~'16	CAGR '10~'16
데이터 솔루션	6,725	8,717	10,487	10,789	13,619	14,124	14,876	5.3	14.1
데이터 구축·컨설팅	37,407	43,180	47,715	49,985	53,730	55,280	55,651	0.7	6.8
데이터 서비스	42,242	43,218	47,317	52,258	57,329	64,151	66,305	3.4	7.8
전체	86,374	95,115	105,519	113,032	124,678	133,555	136,832	2.5	8.0

[그림 2-1-2] 데이터산업 부문별 시장 규모 비중



빅데이터, 사물인터넷(IoT), 인공지능(Artificial Intelligence, AI) 등 ICT 환경 변화에 따른 지능정보사회의 도래는 데이터산업 시장 성장을 이끄는 주요 동력이 될 것이며, 이에 향후 데이터산업은 2020년까지 연평균 증가율 3.5%의 성장세로 16조원 대 시장 진입이 임박할 것으로 전망된다.

[그림 2-1-3] 2015~2020년(P) 데이터산업 시장 전망

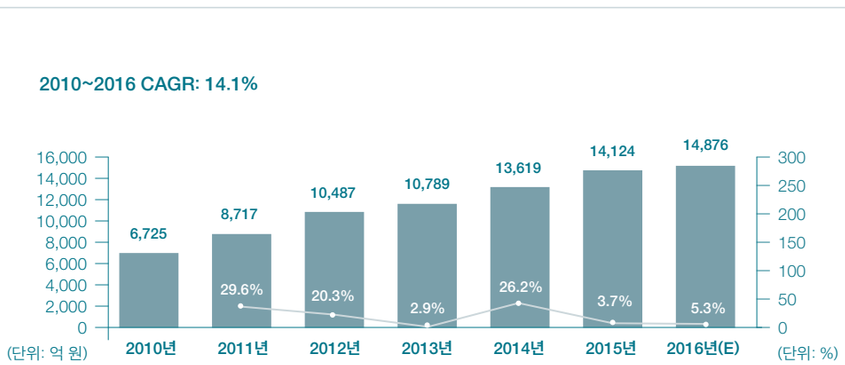


2. 국내 데이터산업 부문별 시장 규모

가. 데이터 솔루션 시장

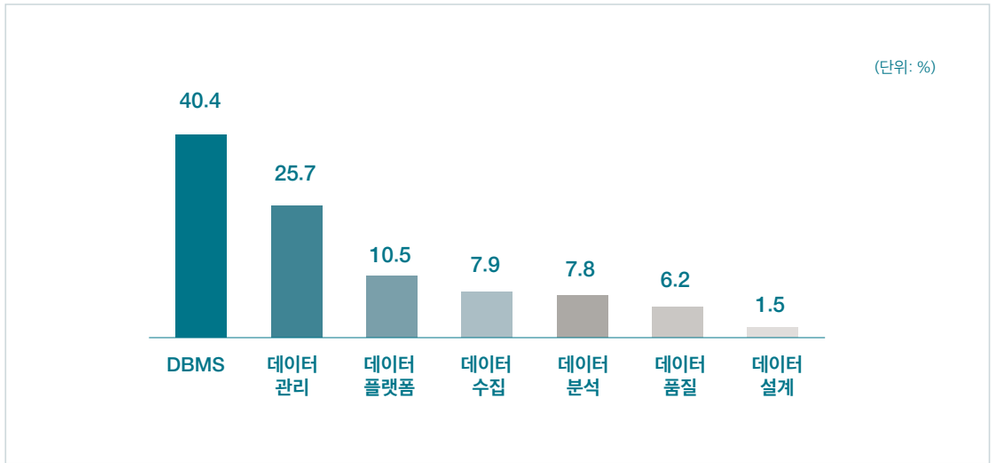
2016년 데이터 솔루션 시장 규모는 2015년 대비 5.3% 성장한 1조 4,876억 원으로 나타났으며, 연평균 성장률(CAGR) 14.1%로 성장하고 있다.

[그림 2-1-4] 데이터 솔루션 시장 규모



데이터 솔루션 시장은 크게 데이터 수집, 데이터 설계, DBMS, 데이터 관리, 데이터 품질, 데이터 분석, 데이터 플랫폼의 7개 중분류로 구분하고 있다.* 이 중 DBMS 분야가 40.4%로 가장 큰 비중을 차지하는 것으로 나타났다.

[그림 2-1-5] 2016년 데이터 솔루션 중분류별 시장 규모 비중



[표 2-1-2] 데이터 솔루션 중분류별 시장 규모 (단위: 억 원, %)

구 분	2010년	2011년	2012년	2013년	2014년	2015년	2016년(E)	증감률 '15~'16	CAGR '10~'16
데이터 수집	448	570	617	622	1,076	1,115	1,178	5.7	17.5
데이터 설계	129	136	151	166	202	207	217	4.8	9.1
DBMS	3,169	4,272	5,301	5,488	5,502	5,727	6,005	4.9	11.2
데이터 관리	1,237	1,516	2,075	2,489	3,446	3,574	3,826	7.1	20.7
데이터 품질	686	949	942	855	883	918	923	0.5	5.1
데이터 분석	1,056	1,274	1,401	1,169	1,121	1,157	1,160	0.3	1.6
데이터 플랫폼	-	-	-	-	1,389	1,426	1,567	9.9	6.2**
데이터 솔루션 전체	6,725	8,717	10,487	10,789	13,619	14,124	14,876	5.3	14.1

* 세부영역별 자세한 내용은 「2016년 데이터산업 현황 조사(미래창조과학부·한국데이터진흥원, 2016. 4)」 상의 데이터산업 분류체계를 참고하기 바람

** 2014년~2016년 CAGR

데이터 솔루션 시장의 매출 영역은 크게 라이선스·개발·유지보수로 구분하며, 데이터 솔루션 중분류별로 매출 영역별 시장 규모는 다음과 같다.

[표 2-1-3] 데이터 솔루션 영역별 시장 규모 (단위: 억 원, %)

구 분		2010년	2011년	2012년	2013년	2014년	2015년	2016년(E)	증감률 '15~'16	CAGR '10~'16
데이터 수집	라이선스	370	433	477	492	316	225	242	7.6	14.0
	개발					507	523	571	9.2	
	유지보수	78	137	140	130	253	367	365	-0.5	29.3
데이터 설계	라이선스	111	117	130	144	25	33	42	27.3	7.0
	개발					141	133	125	-6.0	
	유지보수	18	19	21	22	36	41	50	22.0	18.6
DBMS	라이선스	2,531	3,590	4,718	4,722	2,698	2,712	2,838	4.6	11.8
	개발					1,987	2,101	2,102	0.0	
	유지보수	638	682	583	766	817	914	1,065	16.5	8.9
데이터 관리	라이선스	953	1,212	1,692	1,987	1,395	363	406	11.8	15.6
	개발					1,189	1,796	1,868	4.1	
	유지보수	281	304	383	502	862	1,415	1,552	9.7	33.0
데이터 품질	라이선스	481	662	630	587	436	131	137	3.1	0.7
	개발					243	354	365	2.5	
	유지보수	205	287	312	268	204	433	421	-3.2	12.7
데이터 분석	라이선스	841	1,111	1,319	967	104	116	157	35.3	2.0
	개발					821	846	792	-6.5	
	유지보수	215	163	82	202	196	195	213	9.2	-0.2
데이터 플랫폼	라이선스	-	-	-	-	391	361	353	-2.2	-5.0
	개발	-	-	-	-	746	823	925	9.3	11.4
	유지보수	-	-	-	-	252	242	289	19.4	7.1
데이터 솔루션 전체	라이선스	5,287	7,125	8,966	8,899	5,365	3,941	4,175	5.9	12.9
	개발					5,634	6,576	6,748	2.6	
	유지보수	1,435	1,592	1,521	1,890	2,620	3,607	3,955	9.6	18.4

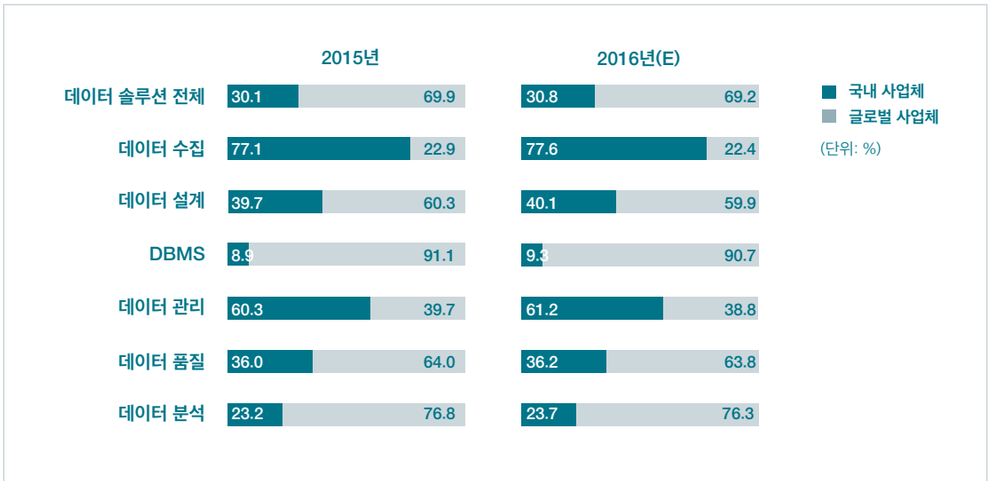
데이터 솔루션 시장의 경우 공공분야가 가장 큰 수요층인 것으로 나타났으며, 공공부문의 투자 증대가 데이터 솔루션 시장 성장의 주요 요인으로 보인다.

[표 2-1-4] 데이터 솔루션 업종별 매출 비중 (단위: %)

구 분		공공	금융	제조·건설	유통·서비스	통신·미디어	의료	전체
데이터 수집	2015년	43.5	9.4	15.6	26.9	2.6	2.1	100.0
	2016년(E)	43.7	10.4	14.8	27.6	2.8	0.8	100.0
데이터 설계	2015년	27.8	14.8	15.5	34.1	4.7	3.3	100.0
	2016년(E)	25.8	12.0	20.5	31.4	6.9	3.3	100.0
DBMS	2015년	46.6	7.8	13.7	23.3	5.0	3.6	100.0
	2016년(E)	49.6	7.5	16.3	17.8	5.0	3.9	100.0
데이터 관리	2015년	44.0	8.5	17.2	21.3	5.5	3.5	100.0
	2016년(E)	42.3	7.7	17.2	24.3	4.4	4.1	100.0
데이터 품질	2015년	22.8	26.7	23.5	15.7	9.6	1.7	100.0
	2016년(E)	22.4	29.4	22.8	14.1	9.6	1.7	100.0
데이터 분석	2015년	39.3	14.4	18.1	16.5	8.9	2.9	100.0
	2016년(E)	41.3	12.5	17.2	17.6	8.2	3.3	100.0
데이터 플랫폼	2015년	51.4	8.2	10.5	14.1	15.5	0.5	100.0
	2016년(E)	48.1	10.8	8.9	16.9	13.9	1.5	100.0
데이터 솔루션 전체	2015년	39.3	12.8	16.3	21.7	7.4	2.5	100.0
	2016년(E)	39.0	12.9	16.8	21.4	7.2	2.7	100.0

데이터 솔루션 시장의 국내 사업체와 글로벌 사업체 간 시장점유율을 보면, 2016년은 전반적으로 국내 사업체 비중이 소폭 상승했다. 특히 데이터 수집, 데이터 관리 분야에서 국내 사업체가 비교적 강세를 보이고 있는 것으로 나타났다. DBMS 시장은 오라클을 중심으로 IBM, 마이크로소프트 등 글로벌 기업이 여전히 강세를 보이고 있으나 티맥스소프트 등 국내 사업체의 점유율이 2015년 8.9%에서 2016년 9.3%로 소폭 증가한 것으로 나타났다.

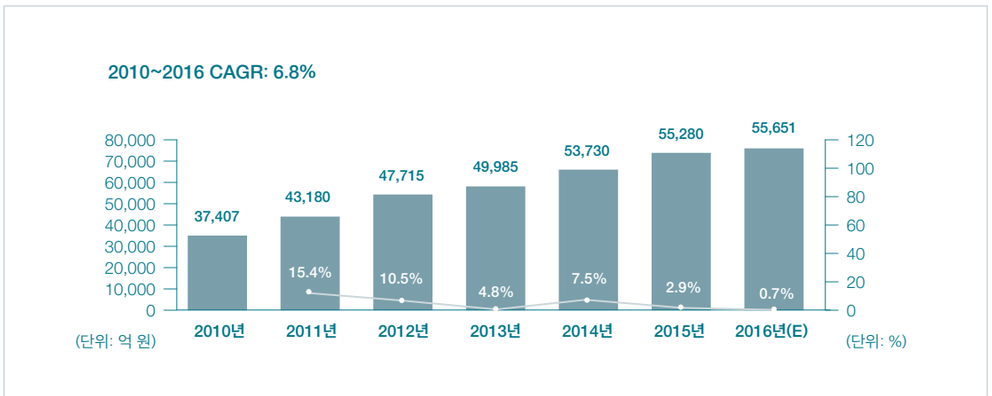
[그림 2-1-6] 국내 데이터 솔루션 시장점유율



나. 데이터 구축·컨설팅 시장

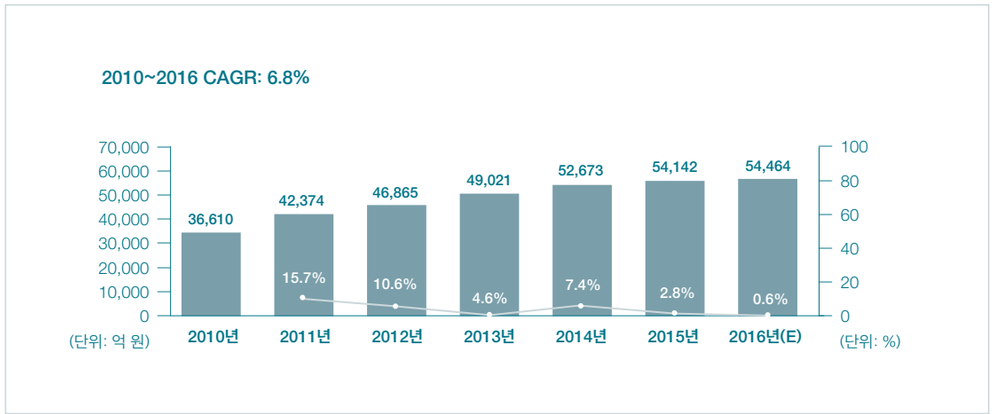
2016년 데이터 구축·컨설팅 시장 규모는 5조 5,651억 원으로 2015년 대비 소폭 상승한 것으로 조사됐으며, 2010년 이후 연평균 증가율은 6.8%로 나타났다.

[그림 2-1-7] 데이터 구축·컨설팅 시장 규모



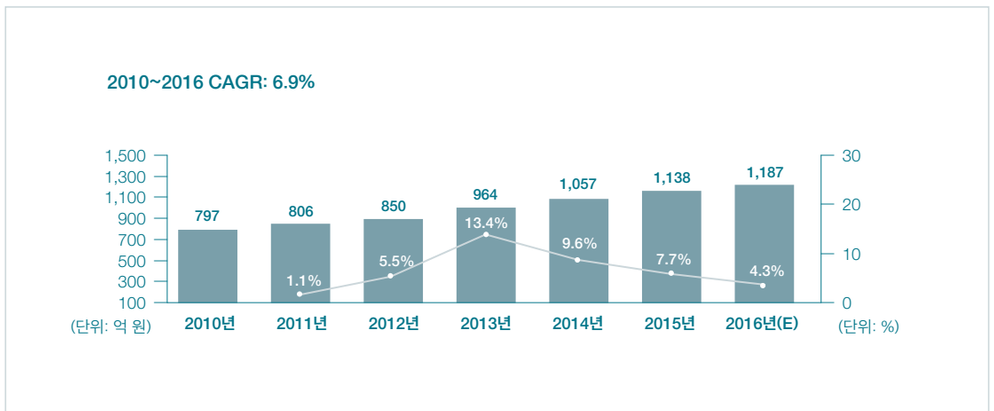
데이터 이행·처리, DB 설계·구축, 기계 처리형 데이터 구축 분야를 포함한 데이터 구축 시장은 2016년 5조 4,464억 원으로 2015년과 비슷한 수준으로 나타났으며, 2010년 이후 연평균 성장률 6.8%로 집계됐다.

[그림 2-1-8] 데이터 구축 시장 규모



데이터 컨설팅 시장은 데이터 설계 컨설팅, 데이터 품질 컨설팅, 데이터 관리 성능개선 컨설팅, 데이터 거버넌스 컨설팅, 데이터 활용 컨설팅 등의 영역을 포함하고 있다. 이러한 데이터 컨설팅 시장은 2016년 1,187억 원으로 2015년 대비 4.3% 성장했으며, 2010년 이후 연평균 성장률 6.9%로 꾸준한 성장세를 유지하는 것으로 나타났다. 업계에 따르면 특히 빅데이터 분석에 대한 관심 및 투자 증대로 데이터 활용 컨설팅 분야와 데이터 품질 컨설팅 분야가 꾸준히 증가하고 있으며, 단순히 데이터 저장보다는 데이터 분석 기법이 늘어나면서 시스템 고도화 프로젝트 수요도 확대되고 있어 사업이 더 활발해질 것이다.

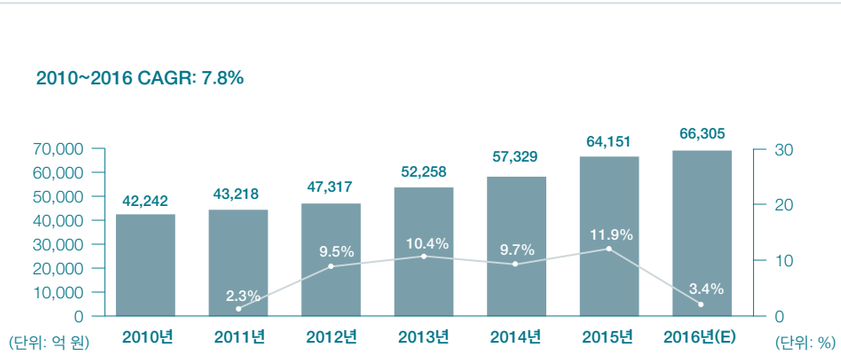
[그림 2-1-9] 데이터 컨설팅 시장 규모



다. 데이터 서비스 시장

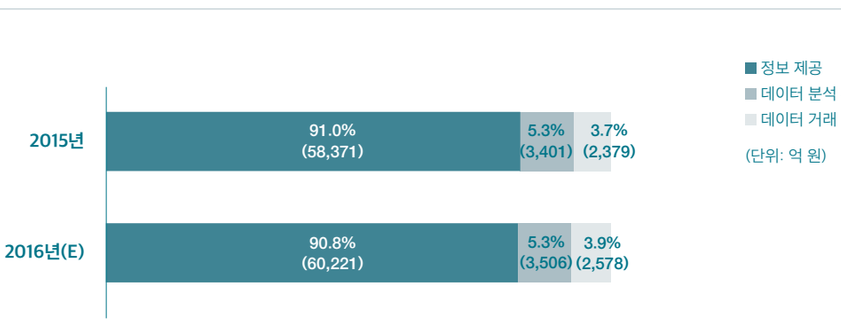
데이터 서비스 시장은 2016년 6조 6,305억 원으로 2015년 대비 3.4% 성장할 것으로 나타났으며, 2010년 이후 연평균 성장률 7.8%로 성장하고 있다.

[그림 2-1-10] 데이터 서비스 시장 규모



데이터 서비스는 크게 데이터나 DB를 기반으로 정보를 제공하는 정보 제공 서비스, 데이터를 직접 판매하거나 중개하는 데이터 거래 서비스, 데이터를 분석해 새로운 가치를 제공하는 데이터 분석 제공 서비스로 구분할 수 있다. 이 중 정보 제공 서비스가 90%로 가장 큰 시장을 형성하며, 이어서 데이터 분석 제공 서비스, 데이터 거래 서비스 순이다.

[그림 2-1-11] 데이터 서비스 중분류별 시장 규모 비중



그러나 최근 데이터 활용과 분석이 늘어남에 따라 데이터 거래와 데이터 분석 제공 분야의 성장률이 두드러지며, 특히 2013년 이후 연평균 성장률을 보면 비교적 높은 성장률을 보이고 있다.

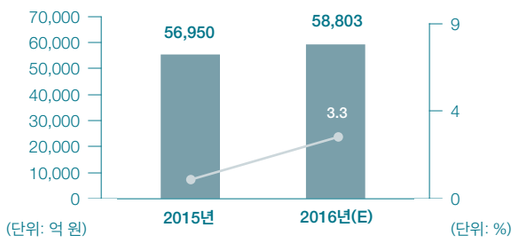
[표 2-1-5] 데이터 서비스 중분류별 시장 규모 (단위: 억 원, %)

구분	2013년		2014년		2015년		2016년(E)		증감률 '15~'16	CAGR '13~'16
	규모	비중	규모	비중	규모	비중	규모	비중		
데이터 거래	1,708	3.3	2,019	3.5	2,379	3.7	2,578	3.9	8.4	14.7
정보 제공	48,284	92.4	51,985	90.7	58,171	90.7	59,921	90.4	3.0	7.5
데이터 분석 제공	2,266	4.3	3,325	5.8	3,601	5.6	3,806	5.7	5.7	18.9
데이터 서비스 전체	52,258	100.0	57,329	100.0	64,151	100.0	66,305	100.0	3.4	8.3

3. 국내 데이터산업 직접매출 시장 규모*

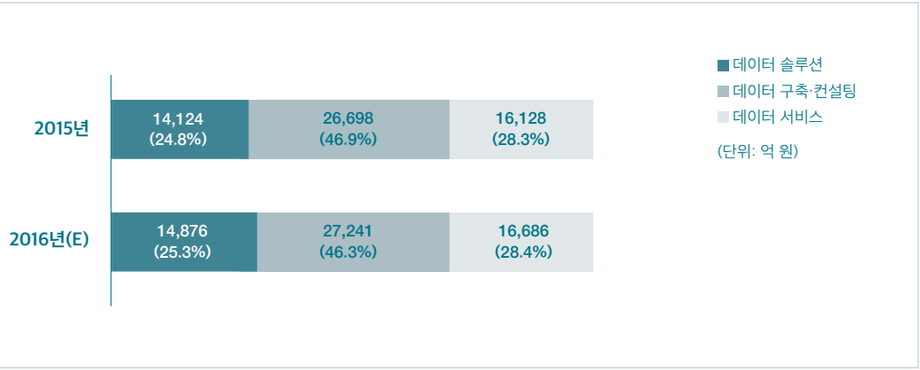
2016년 데이터산업 직접매출 시장 규모는 5조 8,803억 원으로 2015년 대비 3.3% 성장한 것으로 조사됐다.

[그림 2-1-12] 데이터산업 직접매출 시장 규모



* 데이터산업 직접매출 시장 규모는 데이터와 직접적으로 관련이 있는 매출 기준을 적용한 결과로, 전통적인 DB 산업을 아우르는 기존 조사 범위에서 데이터를 매개로 하는 광고매출, DB 시스템 구축 용역 매출 등 DB 관련 간접 매출을 제외하고 산출한 결과임

[그림 2-1-13] 데이터산업 부문별 직접매출 시장 규모 비중



[표 2-1-6] 데이터산업 세부 영역별 직접매출 시장 규모

(단위: 억 원, %)

구분		2015년	2016년(E)	증감률 '15~'16
데이터 솔루션	데이터 수집	1,115	1,178	5.7
	데이터 설계	207	217	4.8
	DBMS	5,727	6,005	4.9
	데이터 관리	3,574	3,826	7.1
	데이터 품질	918	923	0.5
	데이터 분석	1,157	1,160	0.3
	데이터 플랫폼	1,426	1,567	9.9
데이터 솔루션 전체		14,124	14,876	5.3
데이터 구축·컨설팅	데이터 구축	25,560	26,054	1.9
	데이터 컨설팅	1,138	1,187	4.3
데이터 구축·컨설팅 전체		26,698	27,241	2.0
데이터 서비스	데이터 거래	2,350	2,422	3.1
	정보 제공	10,726	11,183	4.3
	데이터 분석 제공	3,052	3,081	1.0
데이터 서비스 전체		16,128	16,686	3.5
전체		56,950	58,803	3.3

2016년 조사에서는 처음으로 기존 부문별 시장 규모에서 특히 데이터 구축과 데이터 서비스 부문에서 데이터를 매개로 하는 광고매출, DB 시스템 구축 용역 매출 등 DB 관련 간접매출을 제외한 데이터와 직접적으로 관련 있는 매출 중심으로 데이터산업 직접매출 시장 규모를 살펴보았다. 그 결과 데이터 구축·컨설팅 시장이 2016년 2조 7,241억 원으로 가장 큰 비중을 차지했고, 이어서 데이터 서비스 시장이 1조 6,686억 원, 데이터 솔루션 시장이 1조 4,876억 원으로 나타났다.

4. 국내 데이터 직무 인력 현황

국내 데이터산업에 종사하고 있는 인력은 총 28만 8,621명으로 전년 대비 3.0% 증가했으며, 이 중 데이터 직무 인력*은 7만 3,256명으로 전년 대비 4.1% 증가한 것으로 나타났다.

[표 2-1-7] 데이터산업 인력 현황

(단위: 명, %)

구 분	2015년		2016년		증감률 '15~'16
	인력 규모	비중	인력 규모	비중	
데이터 직무	70,338	25.1	73,256	25.4	4.1
데이터 직무 외	209,985	74.9	215,365	74.6	2.6
전체	280,323	100.0	288,621	100.0	3.0

데이터 직무 인력의 증가율이 데이터 직무 외 인력의 증가율보다 높게 나타났으며, 이는 데이터 컨설팅 부문에서 데이터 직무 인력 증가가 주된 요인으로 보인다.

* 데이터 직무 인력이란 기업에서 필요한 데이터 관련 기획·개발·분석·운영 및 관리 직무를 수행하는 인력을 말하며, DA(Data Architect)와 데이터 개발자, 데이터 엔지니어, 데이터 분석가, DBA(DB Administrator), 데이터 사이언티스트, 데이터 컨설턴트, 데이터 기획·마케터로 구분한다.

[표 2-1-8] 데이터산업 내 부문별 데이터 직무 인력 현황 (단위: 명, %)

구분	2015년		2016년		증감률 '15~'16
	인력 규모	비중	인력 규모	비중	
데이터 솔루션	8,886	12.6	9,272	12.7	4.3
데이터 구축·컨설팅	34,323	48.8	35,404	48.4	3.1
구축	29,304	41.7	29,941	40.9	2.2
컨설팅	5,019	7.1	5,463	7.5	8.8
데이터 서비스	27,129	38.6	28,580	39.0	5.3
데이터산업 전체	70,338	100.0	73,256	100.0	4.1

※ 이하 데이터 구축·컨설팅 분야 데이터 직무 인력은 편의상 데이터 구축과 데이터 컨설팅으로 구분해 표시함

데이터산업뿐만 아니라 일반산업*까지 포함해 전 산업의 데이터 직무 인력을 보면, 2016년 총 10만 2,375명으로 2015년 대비 1.9% 증가한 것으로 나타났다.

[표 2-1-9] 전 산업 내 데이터 직무 인력 현황 (단위: 명, %)

구분	2015년		2016년		증감률 '15~'16
	인력 규모	비중	인력 규모	비중	
데이터산업	70,338	70.0	73,256	71.6	4.1
일반산업	30,102	30.0	29,119	28.4	-3.3
전 산업	100,440	100.0	102,375	100.0	1.9

데이터 직무별로 보면 데이터 개발자가 38,948명으로 가장 큰 비중을 차지했다. 이어서 DBA가 17,116명, 데이터 엔지니어가 15,670명 순으로 나타났다.

[표 2-1-10] 2016년 전 산업의 데이터 직무별 인력 현황 (단위: 명)

구분	데이터산업					일반산업	전체 (데이터 +일반)
	데이터 솔루션	데이터 구축	데이터 컨설팅	데이터 서비스	데이터산업 전체		
DA	569	3,213	499	455	4,736	4,531	9,267

(계속)

* 여기에서 일반산업이란 금융, 제조, 유통·서비스, 공공, 통신·미디어, 의료의 주요 6대 산업을 말함

구분	데이터산업					일반산업	전체 (데이터 +일반)
	데이터 솔루션	데이터 구축	데이터 컨설팅	데이터 서비스	데이터산업 전체		
데이터 개발자	3,835	12,865	2,106	8,573	27,379	11,569	38,948
데이터 엔지니어	1,380	6,260	626	4,398	12,664	3,006	15,670
데이터 분석가	588	1,513	304	2,072	4,477	2,862	7,339
DBA	648	3,536	333	6,971	11,488	5,628	17,116
데이터 사이언티스트	203	388	150	444	1,185	477	1,662
데이터 컨설턴트	1,446	1,292	1,168	122	4,028	485	4,513
데이터 기획·마케터	603	874	277	5,545	7,299	561	7,860
전체	9,272	29,941	5,463	28,580	73,256	29,119	102,375

2015년 대비 가장 높은 증가율을 나타낸 데이터 직무는 데이터 분석가로서 2016년은 전년 대비 43.3% 증가한 7,339명인 것으로 조사됐다.*

[표 2-1-11] 2015~2016년 전 산업의 데이터 직무별 인력 현황 비교 (단위: 명, %)

구분	2015년		2016년		증감률 '15~'16
	규모	비중	규모	비중	
DA	7,065	7.0	9,267	9.1	31.2
데이터 개발자	41,205	41.0	38,948	38.0	-5.5
데이터 엔지니어	14,013	14.0	15,670	15.3	11.8
데이터 분석가	5,120	5.1	7,339	7.2	43.3
DBA	17,427	17.4	17,116	16.7	-1.8
데이터 사이언티스트	1,343	1.3	1,662	1.6	23.8
데이터 컨설턴트	4,351	4.3	4,513	4.4	3.7
데이터 기획·마케터	9,916	9.9	7,860	7.7	-20.7
전체	100,440	100.0	102,375	100.0	1.9

* 데이터 기획·마케터의 경우, 분석·개발·컨설팅 등의 데이터 관련 업무 수행에서도 필요한 업무로서 해당 직무 인력이 데이터 기획 등의 업무를 병행하는 경향이 늘면서 나타난 결과로 예상됨

전 산업의 데이터 직무 인력 중 빅데이터 관련 인력은 총 9,321명으로, 전체의 9.1%를 차지하고 있다.

[표 2-1-12] 2016년 전 산업 내 데이터 직무 중 빅데이터 관련 인력 현황 (단위: 명, %)

구분	전산업 내 데이터 직무 전체		빅데이터 관련 인력		전 산업 내 데이터 직무 중 빅데이터 인력 비중
	규모	비중	규모	비중	
데이터 개발자	38,948	38.0	2,703	29.0	6.9
데이터 엔지니어	15,670	15.3	1,602	17.2	10.2
데이터 분석가	7,339	7.2	1,052	11.3	14.3
데이터 사이언티스트	1,662	1.6	1,662	17.8	100.0
데이터 컨설턴트	4,513	4.4	1,606	17.2	35.6
데이터 기획·마케터	7,860	7.7	696	7.5	8.9
전체	102,375	100.0	9,321	100.0	9.1

※ 빅데이터 관련 인력: 데이터 직무 인력 중 빅데이터 기술을 보유한 인력으로 데이터 직무 인력 수에 포함됨
※ DA와 DBA는 조사 제외 대상이며, 데이터 사이언티스트 직무는 100% 빅데이터 관련 인력으로 간주
※ 전산업 데이터 직무 전체(102,375)는 DA(9,267)와 DBA(17,116) 인력 수가 포함된 수치임

5. 국내 데이터 직무 인력 수요

향후 3년 내, 일반산업을 포함한 전 산업에서 추가로 필요로 하는 데이터 직무 인력은 총 1만 6,915명으로 조사됐다. 이 중 데이터 개발자 수요가 34.9%로 가장 높게 나타났다. 데이터산업의 경우 데이터 엔지니어 직무 수요, 일반 산업의 경우 데이터 분석가 수요가 비교적 높은 것으로 조사됐다.

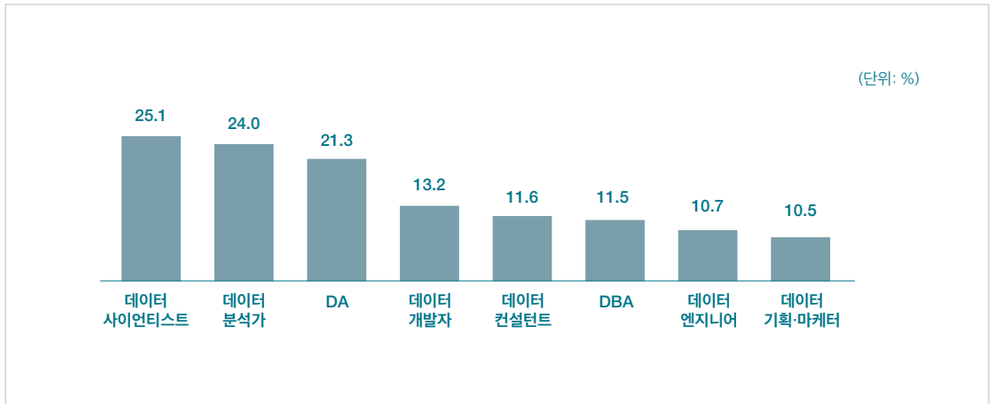
[표 2-1-13] 2016년 전 산업 내 데이터 직무별 필요인력(향후 3년 내) (단위: 명)

구분	데이터산업					일반산업	전 산업 (데이터+일반)
	데이터 솔루션	데이터 구축	데이터 컨설팅	데이터 서비스	데이터산업 전체		
DA	144	110	46	40	340	2,175	2,515
	13.6%	6.4%	6.1%	3.5%	7.2%	17.8%	14.9%

(계속)

구분	데이터산업					일반산업	전 산업 (데이터+일반)
	데이터 솔루션	데이터 구축	데이터 컨설팅	데이터 서비스	데이터산업 전체		
데이터 개발자	487	1,094	345	576	2,502	3,404	5,906
	45.9%	63.4%	45.6%	50.2%	53.3%	27.8%	34.9%
데이터 엔지니어	165	364	119	124	772	1,102	1,874
	15.6%	21.1%	15.7%	10.8%	16.5%	9.0%	11.1%
데이터 분석가	71	27	136	141	375	1,947	2,322
	6.7%	1.6%	18.0%	12.3%	8.0%	15.9%	13.7%
DBA	12	30	14	73	129	2,105	2,234
	1.1%	1.7%	1.8%	6.4%	2.8%	17.2%	13.2%
데이터 사이언티스트	98	40	29	13	180	376	556
	9.2%	2.3%	3.8%	1.1%	3.8%	3.1%	3.3%
데이터 컨설턴트	31	17	47	42	137	453	590
	2.9%	1.0%	6.2%	3.7%	2.9%	3.7%	3.5%
데이터 기획·마케터	53	43	21	138	255	663	918
	5.0%	2.5%	2.8%	12.0%	5.4%	5.4%	5.4%
전체	1,061	1,725	757	1,147	4,690	12,225	16,915
	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

[그림 2-1-14] 2016년 전 산업 내 데이터 직무별 인력 부족률



전체적으로는 데이터 개발자에 대한 수요가 높게 나타나고 있으나, 데이터 직무 인력 부족률*을 보면 데이터 사이언티스트, 데이터 분석가, 데이터 아키텍트(DA) 직무인력이 가장 필요한 것으로 나타났다.

[표 2-1-14] 2016년 전 산업 내 데이터 직무별 인력 부족률 (단위: %)

구분	데이터산업					일반산업	전 산업 (데이터+일반)
	데이터 솔루션	데이터 구축	데이터 컨설팅	데이터 서비스	데이터산업 전체		
DA	20.2	3.3	8.4	8.1	6.7	32.4	21.3
데이터 개발자	11.3	7.8	14.1	6.3	8.4	22.7	13.2
데이터 엔지니어	10.7	5.5	16.0	2.7	5.7	26.8	10.7
데이터 분석가	10.8	1.8	30.9	6.4	7.7	40.5	24.0
DBA	1.8	0.8	4.0	1.0	1.1	27.2	11.5
데이터 사이언티스트	32.6	9.3	16.2	2.8	13.2	44.1	25.1
데이터 컨설턴트	2.1	1.3	3.9	25.6	3.3	48.3	11.6
데이터 기획·마케터	8.1	4.7	7.0	2.4	3.4	54.2	10.5
전체	10.3	5.4	12.2	3.9	6.0	29.6	14.2

※ 인력 부족률 : 필요(부족)인력/(현인력+필요(부족)인력)

이 중 빅데이터 관련 인력 수요는 총 6,451명이다. 빅데이터 관련 데이터 개발자 필요인력이 48.1%로 가장 수요가 높고, 데이터 분석가 수요 비중도 비교적 높게 나타났다.

[표 2-1-15] 2016년 전 산업 내 부문별 데이터 직무 중 빅데이터 관련 필요인력(향후 3년 내) (단위: 명)

구분	전 산업 내 데이터 직무 중 빅데이터 필요인력		
	데이터산업	일반산업	전 산업
데이터 개발자	641	2,497	3,138
	50.9%	48.1%	48.6%

(계속)

* 인력 부족률: 필요(부족)인력/(현인력+필요(부족)인력)

구분	전 산업 내 데이터 직무 중 빅데이터 필요인력		
	데이터산업	일반산업	전 산업
데이터 엔지니어	103	548	651
	8.2%	10.6%	10.1%
데이터 분석가	173	1,055	1,228
	13.7%	20.3%	19.0%
데이터 사이언티스트	180	376	556
	14.3%	7.2%	8.6%
데이터 컨설턴트	83	453	536
	6.6%	8.7%	8.3%
데이터 기획·마케터	79	263	342
	6.3%	5.1%	5.3%
전체	1,259	5,192	6,451
	100.0%	100.0%	100.0%

※ 빅데이터 관련 인력: 데이터 직무 인력 중 빅데이터 기술을 보유한 인력으로 데이터 직무 인력 수에 포함됨
※ DA와 DBA는 조사 제외 대상이며, 데이터 사이언티스트 직무는 100% 빅데이터 관련 인력으로 간주

빅데이터 관련 데이터 직무 인력의 부족률을 살펴보면, 빅데이터 관련 데이터 분석가와 빅데이터 관련 데이터 개발자 인력이 가장 높게 나타났다.

[표 2-1-16] 2016년 전 산업 내 데이터 직무 중 빅데이터 인력 부족률 (단위: %)

구분	전 산업 내 데이터 직무 중 빅데이터 인력 부족률		
	데이터산업	일반산업	전 산업
데이터 개발자	27.0	72.0	53.7
데이터 엔지니어	9.3	47.9	28.9
데이터 분석가	33.7	59.7	53.9
데이터 사이언티스트	13.2	44.1	25.1
데이터 컨설턴트	6.1	57.7	25.0
데이터 기획·마케터	22.6	38.2	32.9
전체	17.8	59.6	40.9

※ 인력 부족률: 필요(부족)인력/(현인력+필요(부족)인력)

6. 마치며

한국은행 '기업경영분석'에 따르면 데이터산업 관련 업종*의 매출 증가율이 연평균 7.9%로 전체 산업 평균, 제조업, 서비스업, 정보통신기술산업보다 높게 나타났다.

2010년에서 2015년까지 데이터산업 관련 업종의 세부 영역별 연평균 증가율은 소프트웨어 개발 및 공급업이 9.8%, 컴퓨터 프로그래밍·시스템 통합 및 관리업이 5.8%, 정보서비스업이 5.9%로 나타났다.

[표 2-1-17] 2010~2015년 국내 주요 산업별 시장 규모 추이 (단위: 억 원, %)

업종 구분	2010년	2011년	2012년	2013년	2014년	2015년	CAGR '10~'15
전산업	29,344,717	32,861,505	34,507,640	35,117,372	35,715,665	35,889,830	4.1
서비스업**	14,542,654	16,039,858	17,002,822	17,752,763	18,455,766	18,927,422	5.4
제조업	14,802,063	16,821,647	17,504,818	17,364,609	17,259,899	16,962,408	2.8
정보통신기술산업	4,746,627	5,259,595	5,714,272	5,983,559	5,659,831	5,627,440	3.5
정보통신기기 제조업***	3,408,036	3,506,645	3,879,050	4,072,918	3,729,152	3,671,747	1.5
정보통신 서비스업****	1,338,591	1,752,950	1,835,222	1,910,641	1,930,679	1,955,693	7.9
데이터산업 관련 업종	459,941	501,197	552,961	592,203	633,971	672,866	7.9
소프트웨어 개발 및 공급업	232,154	249,551	287,988	310,939	340,249	369,741	9.8
컴퓨터 프로그래밍, 시스템 통합 및 관리업	149,040	171,018	179,479	194,431	200,973	198,018	5.8
정보서비스업	78,748	80,628	85,495	86,833	92,749	105,106	5.9
데이터산업	86,374	95,115	105,519	113,032	124,678	133,555	9.1*****
데이터 솔루션	6,725	8,717	10,487	10,789	13,619	14,124	16.0
데이터 구축	36,610	42,374	46,865	49,021	52,673	54,142	8.1
데이터 컨설팅	797	806	850	964	1,057	1,138	7.4
데이터 서비스	42,242	43,218	47,317	52,258	57,329	64,151	8.7

※ 자료1: 기업경영분석, 한국은행 경제통계시스템, 2010~2016
※ 자료2: 데이터산업 시장 규모는 「2016년 데이터산업 현황 조사」 분석 결과임

동일한 기간 내 국내 데이터산업 연평균증가율을 분석한 결과 9.1%라는 비교적 높은 성장세를 보였다. 이는 기존 산업의 정체, 하락 추세에도 빅데이터, 사물인터넷(IoT), 인공지능, 클라우드 등의 ICT 환경 변화가 데이터산업의 성장을 이끄는 주요 동력이 되고 있기 때문으로 보인다.

데이터는 지능정보사회의 원천자원이라고 한다. 이에 데이터산업에 종사하는 데이터 기업들이 지능정보사회의 도래에 따른 시장 환경 변화에 발맞춰 더 적극적으로 투자하고 새로운 사업 기회를 지속적으로 발굴하여 성장을 가속화시킬 수 있도록 한다면 데이터산업의 성장잠재성은 지금보다 더 높을 것으로 전망된다.

* 데이터산업 현황 조사의 조사대상에 해당하는 한국표준산업분류 상의 주된 업종
** 비제조업 중 농업, 어업, 광업, 전기·가스·증기 및 공기조절공급업, 건설업을 제외한 업종임
*** 전자부품·컴퓨터·영상·음향 및 통신장비, 절연선 및 케이블, 측정·시험·항해·제어 및 기타 정밀기기 제조업(광학기기 제외)
**** 정보통신기술 관련 도소매업 및 임대업, 소프트웨어 개발 및 공급업, 통신업, 컴퓨터 프로그래밍·시스템 통합 및 관리업, 정보서비스업
***** 2017년 5월 현재, 한국은행 「2015년 기업경영분석(2016.10)」 보고서까지 공표됐으며, 매출액 비교를 위해 2016년 데이터산업현황조사의 2015년 결과까지 연평균 증가율을 산정했음(「2016년 데이터산업 현황 조사」 보고서에는 2016(E)를 포함해 연평균 증가율을 산출해 8.0%로 수록함)

제2장

국내 빅데이터 시장 현황

국내 빅데이터 시장 현황은 매년 하반기 중에 조사를 실시해 연말 또는 이듬해 초에 보고서로 발간된다. 이에 본 장의 국내 빅데이터 시장 현황은 「2016년 빅데이터 시장현황 조사(한국정보화진흥원, 2017. 3)」를 요약·발췌해 작성했다.

'빅데이터 시장현황 조사'는 빅데이터 관련 HW, SW, 서비스 등을 영위하는 공급기업과 공공, 금융, 유통·서비스, 제조, 의료, 통신·미디어 등 핵심 6개 산업에 해당하는 수요기업(일반기업)으로 구분해 조사를 진행했다. 또한 빅데이터 시장 규모는 공공과 민간시장으로 구분해 산출했다. 정부·공공 시장은 정부·공공기관의 빅데이터 관련 사업 투자액을 기준으로 하고, 민간시장은 수요기업의 빅데이터 시스템 구축 투자액과 공급기업의 빅데이터 서비스분야 매출액을 합산했다.

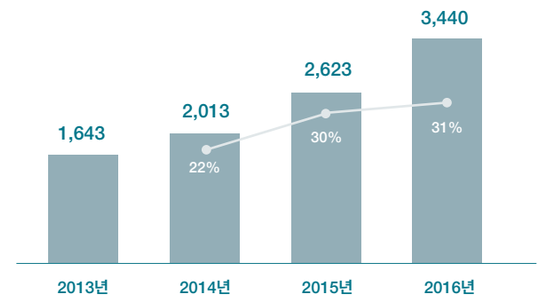
1. 국내 빅데이터 시장 규모

국내 빅데이터 시장은 기업의 빅데이터에 대한 인식 호전과 매출 1,000억 원 이상 중견·대기업의 투자 증가, 정부의 강력한 빅데이터 산업 육성의지에 따라 전년 대비 31.1% 성장한 3,439.6억 원을 기록했다. 특히 정부·공공 투자가 전년 대비 43.1% 성장한 998.6억 원으로 크게 확대되며 시장을 견인했다.

* 필자: 최승우(한국정보화진흥원 빅데이터센터 선임연구원)

[그림 2-2-1] 2013~2016년 국내 빅데이터 시장 규모

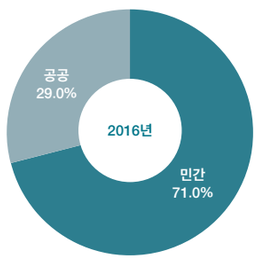
(단위: 억 원)



출처: 「2016년 빅데이터 시장현황 조사」, 한국정보화진흥원, 2017. 3 (이하 동일)

시장 영역별로 정부·공공시장은 빅데이터 산업활성화 법안 등을 기반으로 공공기관에 빅데이터 도입·활용에 대한 적극적인 권고로 인해 43.1% 성장한 998.6억 원을 기록했다. 반면 민간시장은 시장 사이클 관점에서 도입 단계에 해당된다. 성장 잠재력은 높은 편이나 아직 금융, 통신 등의 일부 대기업에 한정된 수요로 인해 공공시장보다는 낮은 26.8% 성장세로 2,441억 원을 기록했다.

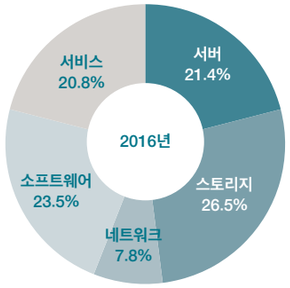
[그림 2-2-2] 2016년 시장 영역별 빅데이터 시장 규모와 비중



시장 구분		'16년 시장 규모 (억 원)	산출 근거
정부·공공 시장		998.6	• 정부, 공공기관의 빅데이터 사업실적 분석 • '16년 조달청 빅데이터 사업실적(223건)
민간 시장	기업의 빅데이터 시스템 투자	2,278.4	• 국내 874개 기업 대상의 빅데이터 투자실태 분석 • 제조, 금융, 유통·서비스, 통신미디어, 의료
	빅데이터 분석 대행 서비스 시장	162.6	• 빅데이터 분석 서비스 기업의 관련 매출 분석
	소계	2,441.0	
합계		3,439.6	

제품별 국내 빅데이터 시장은 서버, 스토리지, 네트워크 등 인프라 투자에 55.7%(1,915.2억 원)가 집중돼 있다. 소프트웨어(23.5%, 809.3억 원), 서비스(20.8%, 715.1억 원) 투자는 선진시장에 비해 비중이 낮은 상황이다. 그러나 그동안 인프라 투자 중심에서 탈피해 빅데이터 분석 시스템 구축이 증가하면서 관련 소프트웨어와 서비스 매출이 각각 34%, 40%*로 빠르게 성장하고 있다.

[그림 2-2-3] 2016년 제품별 국내 빅데이터 시장 규모와 비중

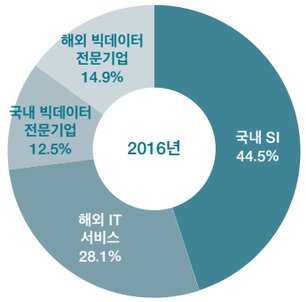


구분	시장 규모(억 원)	비중(%)
서버	736.5	21.4
스토리지	911.3	26.5
네트워크	267.4	7.8
소프트웨어	809.3	23.5
서비스	715.1	20.8
합계	3,439.6	100.0

공급자 유형별 국내 빅데이터 시장은 국내 SI 기업이 44.5%(1,530.6억 원)로 시장을 선도하는 가운데 해외 IT 서비스 기업 28.1%(966.5억 원), 해외 빅데이터 전문기업 14.9%(512.5억 원), 국내 빅데이터 전문기업 12.5%(430억 원) 순으로 집계됐다. 특히 2016년 빅데이터 시장은 공공기관 중심의 투자가 활발하게 전개돼 공공 소프트웨어 구축사업 중심으로 영업을 강화한 국내 기업의 매출 비중이 2015년 48.4%에서 2016년 57.0%로 확대됐다.

* 소프트웨어는 2015년 603.0억 원에서 2016년 809.3억 원으로 34% 증가했고, 서비스는 2015년 512.0억 원에서 2016년 715.1억 원으로 40% 증가했다.

[그림 2-2-4] 2016년 공급자 유형별 국내 빅데이터 시장 규모와 비중



구분	시장 규모(억 원)	비중(%)
국내 SI	1,530.6	44.5
해외 IT 서비스	966.5	28.1
국내 빅데이터 전문기업	430.0	12.5
해외 빅데이터 전문기업	512.5	14.9
합계	3,439.6	100.0

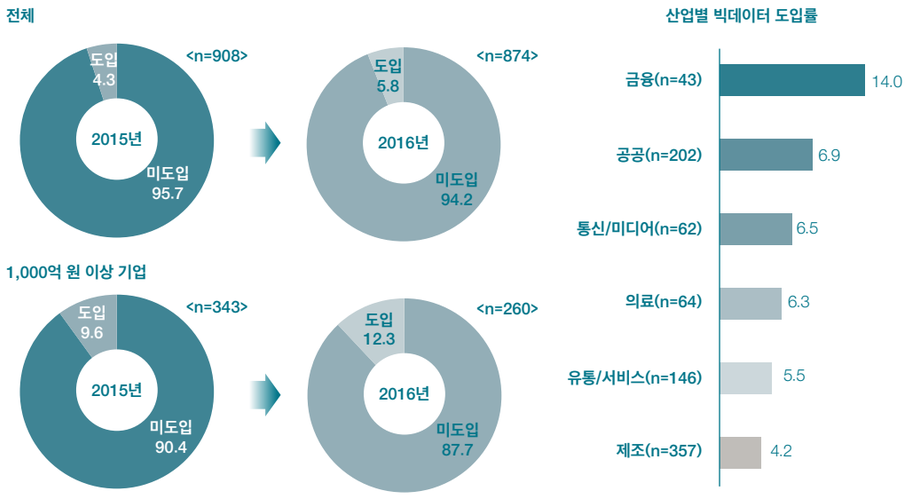
2. 국내 빅데이터 활용 현황

국내 일반기업의 빅데이터 시스템 도입률은 '15년 4.3%에서 '16년 5.8%로 전년 대비 1.5%p 상승한 가운데, 종업원 수 1,000명에 매출액 1,000억 원 이상의 중견·대기업은 '15년 9.6%에서 '16년 12.3%로 2.7% 상승해 금융·공공·통신 분야를 중심으로 중견·대기업의 도입이 활발했다. 이는 빅데이터 도입 및 활용을 원활하게 수행하기 위해서는 일정 수준 이상의 투자 여력이 있어야 하고, 분석할 만한 데이터의 품질과 양이 충분하게 준비돼야 하기 때문인 것으로 보인다.

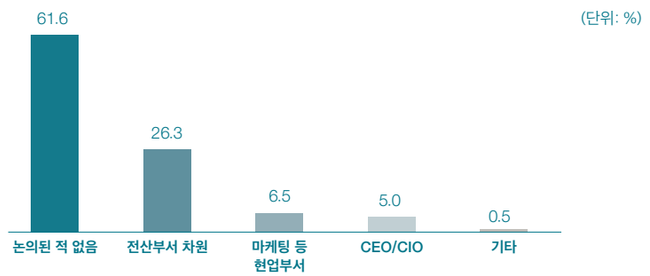
아직 빅데이터를 도입하지 않은 일반기업의 60% 이상은 빅데이터 시스템 도입을 위한 논의가 진행되지 않은 단계라고 답했다. 특히 유통·서비스, 제조, 의료 등 전통적으로 데이터 분석 기반의 의사결정 구조가 약한 산업에서 빅데이터 도입에 대한 관심이 낮은 것으로 나타났다.

반면 빅데이터 도입에 관심이 있는 경우, 주로 전산부서 차원에서 관심을 가지고 있는 것으로 나타나 빅데이터 시스템을 전사적 전략 혹은 마케팅 수단보다는 ICT 시스템의 일부로 인식하고 있는 것으로 분석된다.

[그림 2-2-5] 빅데이터 시스템 도입률 (단위: %)



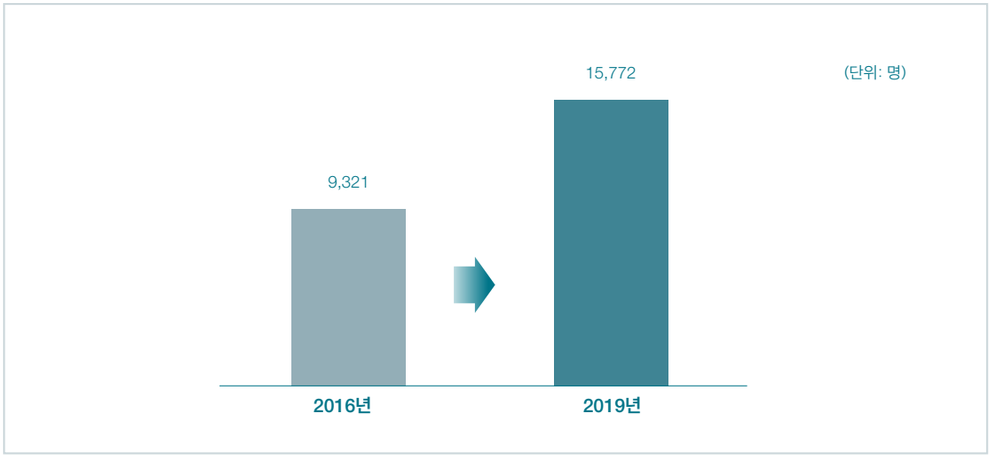
[그림 2-2-6] 미도입 기업의 빅데이터 도입 관심 수준



3. 빅데이터 시장 인력 현황

2016년 국내 공급기업 및 일반기업의 빅데이터 전문인력은 9,321명으로 나타났다. 3년 후인 2019년에는 현재 인력 대비 69.2% 증가한 15,772명 수준의 인력이 필요하다고 조사돼 기업 활동에 빅데이터의 영향력이 점차 증가하면서 관련 전

[그림 2-2-7] 빅데이터 전문인력 현황 및 전망(전체 산업)



문인력의 수요가 크게 증가할 것으로 예상된다.

직무별로는 빅데이터 개발자와 엔지니어가 전체 인력의 4,305명(46.2%)으로 가장 많고, 데이터 사이언티스트 1,662명(17.8%), 빅데이터 컨설턴트 1,606명(17.2%) 등의 순서로 구성돼 있다. 향후 빅데이터 활용과 기술이 고도화되면서 빅데이터 분석가와 개발자의 수요가 크게 증가할 것으로 예상돼 해당 직무에 대한 전문인력 양성 및 시장 공급을 위한 적절한 대책이 요구된다.

[표 2-2-1] 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	성장률
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
빅데이터기획·마케터	696	7.5	1,038	6.6	342	49
총계	9,321	100.0	15,772	100.0	6,451	69

※ 자료: 「2016년 데이터산업 현황 조사(미래창조과학부·한국데이터진흥원, 2016.4)」의 빅데이터 관련 데이터직무 인력 현황 및 수요 조사결과 재구성

4. 국내외 빅데이터 기술수준 비교

공급기업 관점에서 국내 빅데이터 기술수준은 선진국 대비 65.7점, 기술수준 격차는 3.1년, 선진기술 도달시간은 3.4년이 걸릴 것으로 평가됐다. 현재 국내 빅데이터 관련 기술은 선진기업과의 격차가 여전히 크고, 그 격차를 줄이기 위한 시간도 상당 기간 걸릴 것으로 나타났다. 빅데이터 원천 기술을 해외 기업이 주로 선점한 점에서 기술수준 격차가 발생하고 있으며, 그 결과 빅데이터 축적을 위한 인프라(HW)부터 분석 및 활용을 위한 SW까지 대부분 외산제품이 시장을 점유하고 있다. 국내 업체의 기술과 제품은 낮은 안정성과 신뢰성, 유지관리 용이성 등을 이유로 외산제품에 비해 경쟁력이 높지 않은 상황이다.

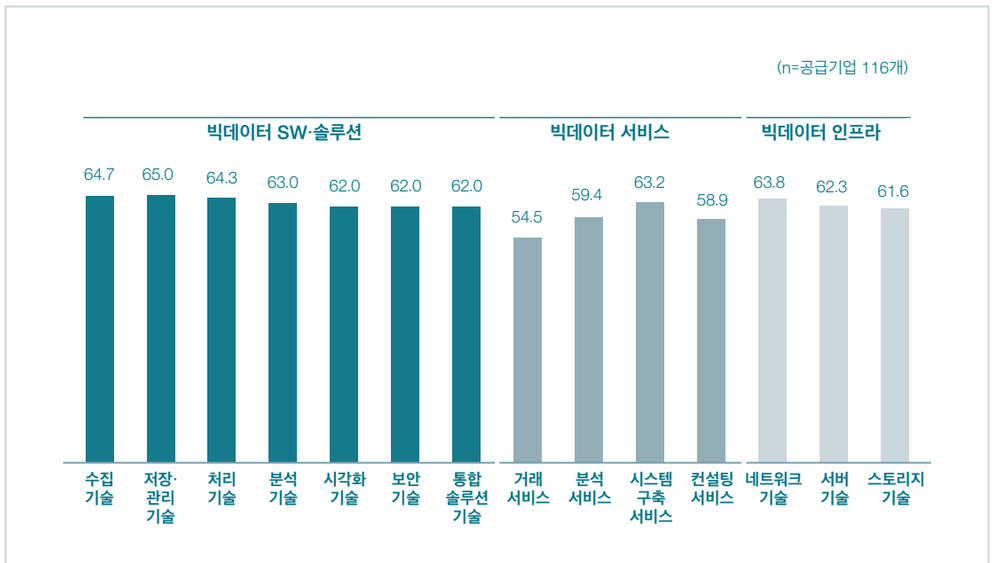
그러나 최근 국내 빅데이터 시장에서 민간기업의 수요와 관련 시장이 빠르게 성장하며 IT 기업의 진출 확대와 선진기술의 도입이 촉진되고 있다. 이로 인해 국내 빅데이터 관련 기술수준은 2015년 대비 3.1점 상승한 65.7점을 기록한 가운데 기술수준 격차는 0.2년 감소한 3.1년, 선진기술 도달 시간은 0.2년 감소한 3.4년으로 동반 개선됐다.

[표 2-2-2] 2015 vs. 2016 빅데이터 공급기업의 국내 기술수준

구분	국내 기술수준 (선진기술 100 기준)	수준 격차(년)	선진기술 도달 시간(년)
2015년	62.6	3.3	3.6
2016년	65.7	3.1	3.4
증감	+3.1	-0.2	-0.2

※ '15년 100개, '16년 114개 응답 기준

[그림 2-2-8] 분야별 국내 기술수준 평가(공급기업)



국내 빅데이터 기술수준은 선진국 대비 열세를 면치 못하고 있으나 분야별로 빅데이터 SW·솔루션과 인프라 분야는 평균 이상을 상회하며 상대적으로 경쟁력을 확보하고 있는 것으로 조사됐다. 반면 서비스 분야는 여전히 취약한 것으로 나타났다. 즉 빅데이터 저장·관리기술, 수집기술, 처리기술 및 관련 인프라 구축 기술력은 상대적으로 높은 경쟁력을 확보하고 있으나 빅데이터 유통체계 부재와 데이터 기반의 응용 서비스 경험 부족에 따라 거래 서비스, 컨설팅 서비스, 분석 서비스 등의 빅데이터 서비스 분야는 경쟁력이 미흡한 것으로 조사됐다.

제3장

해외 데이터산업 현황

데이터산업은 그 범위가 매우 광범위해서 데이터의 정의를 어떻게 하느냐에 따라 시장 규모가 판이하게 나타난다. 이 글에서는 데이터를 가공해 제품 및 서비스로 재생산되는 디지털 데이터 시장으로 그 범위를 한정했다. 물론 이미지 데이터처럼 디지털 기술을 통해 수집, 저장, 가공, 전송되는 멀티미디어 데이터 부문도 포함했다.

세계 데이터산업 시장은 2015년 696억 달러 규모에서 연평균 14% 가량 성장해 2020년에는 1,323억 달러 규모에 도달할 것으로 예상**된다. 또한 데이터산업 시장과는 별도로 데이터를 활용해 온·오프라인 정보 서비스를 제공하는 시장은 '정보 서비스 시장'으로 정의되며, 글로벌 기준 2016년에 1조 5,000억 달러로 추정***되고 있다. 정보 서비스 시장의 규모가 월등히 크다는 것은 데이터를 활용한 정보 서비스가 매우 다양함은 물론, 데이터가 폭넓게 활용되고 있음을 의미한다.

해외 정보 서비스 시장 규모의 절반에 가까운 49.9%는 미국이 차지하고 있다. 미국의 정보 서비스 시장 규모는 2016년에 7,729억 달러 수준인 것으로 추정된다. 미국의 뒤를 이어 유럽이 4,392억 달러로 전체 시장 규모의 28.4%를, 아시아는 2,526억 달러로 16.3%를 차지하고 있다. 따라서 미국과 유럽, 아시아가 전체 시장의 94.6%를 점유하고 있어 정보 서비스 시장에서 만큼은 세 지역에 상당히 편중돼 있다고 볼 수 있다. 정보 서비스 시장이 미국, 유럽, 아시아 지역에 편중·활성화돼 있다는 것은 정보 서비스의 기반이 되는 데이터가 세 지역에서 활발히 활용된다는 것을 의미한다. 데이터산업 시장 역시 세 지역을 중심으로 형성됐을 것으로 추정할 수 있다. 따라서 해외 데이터산업 시장의 현황을 다루는 데 있어서 미국, EU, 아시아 국가 중 일본, 기타 국가 중 브라질을 대상으로 기술했다.* 또한 2013년부터 2016년까지 시장의 추세를 설명했으며, 2016년의 경우 2013년부터 2015년까지의 추세를 바탕으로 추정했다.

1. 데이터산업의 시장 규모

데이터산업 시장 규모는 데이터 관련 기업의 매출액을 기반으로 추정했다. 즉 데이터 기업의 매출은 데이터 관련 제품의 총 가치에 해당하며, 수출을 포함해 해당 국가에 기반을 둔 기업이 생산한 서비스의 매출이다.

미국의 데이터산업 시장은 세계에서 가장 큰 규모를 자랑하고 있으며, 시장의 성장성 또한 비교적 높다. 2013년에 약 1,033억 달러 수준을 기록했으며, 2014년에는 2013년 대비 6.9% 성장한 약 1,104억 달러의 시장 규모를 기록했다. 2015년에는 2014년 대비 무려 11.1% 성장해 급격한 성장세를 보이며 약 1,227억 달러의 규모로 추정됐다.

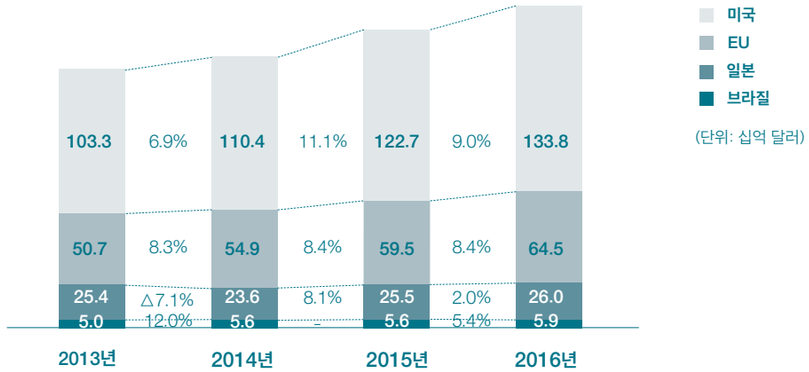
미국의 이러한 빠른 성장 속도는 미국의 데이터산업 시장 규모를 EU보다 2배 가량 크게 만드는 요인으로 작용했고, 2016년에는 약 1,338억 달러 시장 규모를 갖고 있는 것으로 추정됐다. 그러나 미국은 데이터산업 시장이 빠르게 성장하면서 매우 안정적이고 견고해 보이는 외형과 달리 대기업 및 중견기업 중심으로 형성되고, 중소기업들은 현상유지 수준의 낮은 성장률에 머물고 있다. 이 문제는 향후 미국이 풀어야 할 과제다.

EU의 데이터산업 시장 성장 속도는 미국과 견줄 정도로 빠르다. 2013년에 약 507억 달러 수준을 기록했으나, 이후 연평균 성장률 8.0% 내외의 빠른 성장률을 기록해 2016년에는 약 645억 달러의 규모로 추정됐다. 성장률의 큰 변동 없이 꾸준히 높은 성장률을 보인다는 것은 데이터산업이 견고하게 형성돼 있음을 뜻

* 필자: 나영민(날리지리서치그룹 실장)
** 글로벌 데이터산업 시장 규모는 451Research의 데이터산업 시장 보고서 참조
*** 해외 정보 서비스 시장의 규모는 Outsell의 「Information Industry Outlook 2017」 참조

* 본 장의 데이터산업 시장 현황은 「European Data Market(SMART 2013/0063) D8-Second Interim Report(IDC & Open Evidence, 2016.9)」 보고서를 참조

[그림 2-3-1] 2013~2016년 데이터산업 시장 규모



출처: European Data Market Monitoring Tool, 2016 (이하 동일)

한다. 이는 EU 데이터산업 시장의 경우 중소기업 및 벤처기업이 중심이 되고 있어 외부 경제적 요인에 비교적 영향을 덜 받기 때문이다.

일본의 데이터산업 시장 규모는 2013년에 약 254억 달러 수준을 기록했으며, 2014년에는 소폭 하락한 약 236억 달러의 규모로 추정됐다. 그러나 2015년에는 2014년 대비 8.1% 성장해 약 255억 달러의 규모로 추정됐으며, 데이터 기업의 수익률 또한 약 24%로 비교적 견고한 수익구조를 갖고 있는 것으로 판단됐다.

국제통화기금(IMF)에 따르면 일본의 실질 GDP는 2014년과 2015년에 감소했으나, IDC는 일본의 ICT 시장만큼은 2013년에서 2015년 사이에 비교적 완만하게 성장한 것으로 분석했다. 2016년에는 약 260억 달러로 추정되며, 2016년 이후 ICT 투자 환경도 꾸준히 개선될 것으로 예측된다. 또한 투자 환경이 개선됨에 따라 일본의 데이터산업 시장 규모는 지속적인 성장세를 기록할 전망이다.

한편 브라질의 데이터산업 시장 규모는 2013년에 약 50억 달러 수준을 기록했으며, 2014년에는 12.0% 증가해 약 56억 달러의 규모로 추정됐다. 그러나 2015년에는 성장세가 다소 주춤해지면서 2014년의 시장 규모를 유지하는 수준이었다. 브라질 내에서의 ICT 투자 대비 데이터 기업의 수익률을 살펴보면, 2014년 기준 약 4.0%를 기록해 미국의 약 10.0%, EU의 약 8.7%에 비해 수익률이 낮은 것으로 나타났다.

그러나 브라질은 상대적으로 빠른 경제발전 속도를 보여 데이터산업 시장 또한 빠르게 성장할 것으로 예상되며, 프랑스·이탈리아와 같은 나라들과 경쟁할 수 있는 수준으로 올라설 것으로 예상된다. 2016년 브라질 데이터산업 시장 규모는 약 59억 달러로 추정돼 2015년 대비 약 5.4% 성장할 전망이다.

2. 데이터산업 종사자 현황

가. 데이터산업의 종사자 수

데이터산업의 종사자는 데이터 수집·저장·관리·분석을 주요 업무로 하거나 업무 중 상당한 부분을 차지하는 경우로 정의할 수 있다. 데이터산업 종사자는 수많은 데이터와 데이터 기술을 다루며, 의사결정을 위해 정형 또는 비정형 데이터를 시각화하거나 사용하기 용이하게 다듬는 업무를 수행한다.

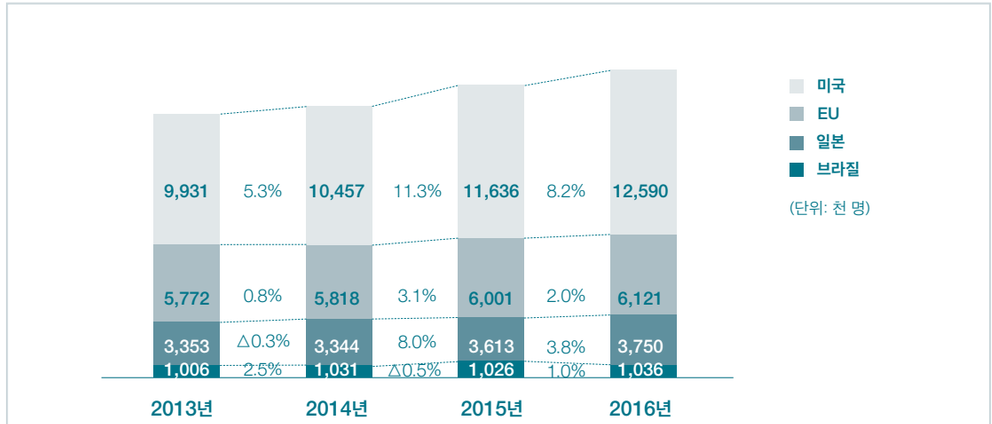
미국, EU, 일본, 브라질의 데이터산업 종사자 수는 2013년부터 꾸준히 증가하고 있다. 특히 2014년부터는 EU, 일본, 브라질의 데이터산업 종사자 수의 합보다 미국의 데이터산업 종사자 수가 많아졌다. 미국의 데이터산업 종사자 수는 2013년 약 993만 명에서 지속적으로 증가해 2016년에는 약 1,259만 명으로 추정됐다. 또한 미국의 데이터산업 종사자 수는 EU의 데이터산업 종사자 수보다 두 배 이상으로 나타나고 있으며, 증가율 역시 EU의 증가율보다 높게 나타났다. 미국의 데이터산업 종사자 수가 많고, 상대적으로 증가율이 높은 이유는 전통적으로 ICT 집약적인 산업구조를 갖고 있어 데이터 이용이 활성화돼 있기 때문이다. 특히 ICT, 금융 서비스, 기타 전문적인 서비스에서 데이터 활용률이 높아 종사자 수가 많이 분포돼 있다.

EU의 데이터산업 종사자 수는 2013년 약 577만 명이었으나 이후 꾸준히 증가해 2016년에는 약 612만 명으로 추정됐다. 미국, 일본보다 데이터산업 종사자 수의 증가율이 상대적으로 낮은 것으로 나타나 고용 시장에서 데이터산업에 대한 매력이 큰 것으로 보이지는 않는다. 그럼에도 꾸준히 증가하고 있다는 것은 데이터산업을 성장시키기 위한 EU 국가들의 노력이 크게 작용한 것으로 볼 수 있다.

일본의 데이터산업 종사자 수는 2013년 약 335만 명이었으나 2016년에는 약 375만 명으로 증가했다. 2014년에는 2013년 대비 소폭 감소했으나 2015년,

2016년 연속으로 증가하는 모습을 보이고 있다. 일본의 경제는 장기적으로 침체에 빠져있어 낮은 GDP 성장률, 소극적인 투자 성향을 보여 왔음에도 불구하고, 데이터산업 종사자 수는 2014년부터 지속적으로 증가하고 있는 것으로 나타났다. 자국 시장에서 데이터산업 시장의 비율이 미국 및 유럽처럼 높아지고 있는 것으로 보이며, 이러한 추세는 향후 지속될 것으로 전망된다.

[그림 2-3-2] 2013~2016년 데이터 종사자 수



한편 브라질의 데이터산업 종사자 수는 2013년 약 100만 명에서 2014년 약 103만 명으로 증가하다가 2015년에 감소하는 모습을 보였다. 또한 2016년에는 약 103만 명으로 다시 증가했는데, 이는 브라질 내에서의 ICT 시장이 상대적으로 작고, ICT 보급률 또한 매우 낮은 수준이라서 성장이 지속적으로 이루어지지 못했기 때문이다.

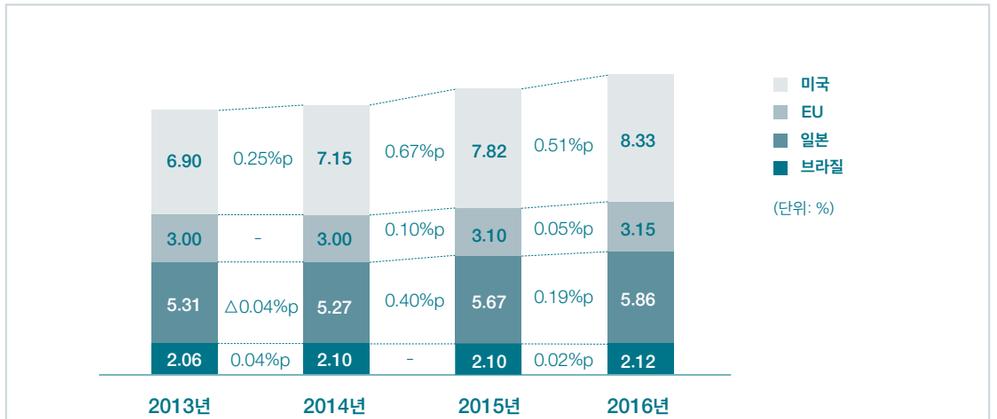
나. 전체 종사자 중 데이터산업 종사자의 비율

전체 산업 종사자 중 데이터산업 종사자의 비율은 데이터산업의 집적 수준을 판단할 수 있는 기준 중 하나다. 즉 데이터산업에 종사하는 노동자의 비율이 높을수록 해당 국가의 산업 구조가 데이터산업에 집적돼 있다는 것을 알 수 있다. 이와 같은 관점에서 미국, EU, 일본, 브라질을 살펴보면 모두 데이터산업 종사자의 비율이 점진적으로 증가하고 있는 것을 알 수 있으며, 이는 데이터 기반의 ICT 산업 집적 구조로 변화되고 있음을 뜻한다.

미국에서 전체 산업의 종사자 중 데이터산업 종사자의 비율은 2013년 약 6.90%에서 2016년 약 8.33%로 증가했으며, 같은 기간 EU·일본·브라질보다 높은 비율을 기록하고 있다. 이는 상대적으로 미국의 산업구조가 ICT 산업 중심으로 형성돼 있다는 것을 보여준다. 미국의 뒤를 이어 일본의 비율이 EU보다 높은 것도 같은 이유이다.

EU에서 전체 산업의 종사자 중 데이터산업 종사자의 비율은 2013년 약 3.0%에서 2016년 약 3.15%로 증가했다. 미국, 일본보다 데이터산업으로 유입되는 종사자 수 증가율뿐만 아니라 종사자 수의 비율 또한 낮은 것으로 조사됐다.

[그림 2-3-3] 2013~2016년 데이터 종사자 수의 비율



일본의 경우 전체 산업의 종사자 중 데이터산업 종사자의 비율은 2013년에 약 5.31%를 기록했으나 2014년에는 5.27%로 0.04%p 감소했다. 이후 2015년, 2016년 연속으로 증가 추세를 보여 2016년에는 약 5.86%로 나타났다.

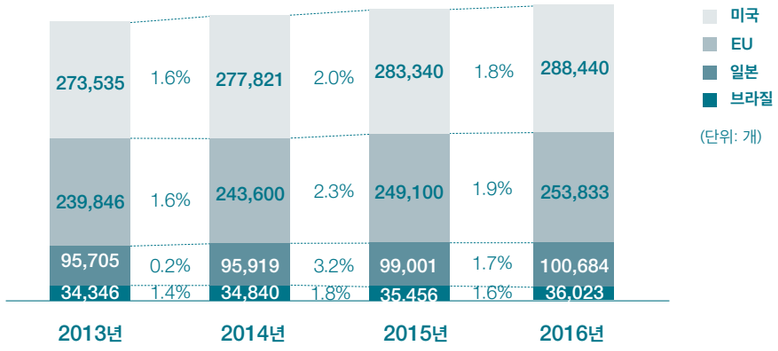
한편 브라질에서 전제 종사자 수 대비 데이터산업 종사자 수의 비율은 2013년 약 2.06%에서 꾸준히 증가해 2016년 약 2.12%의 비중을 차지한 것으로 추정된다. 이는 같은 기간 3.00%에서 3.15%를 기록한 EU의 평균보다 낮은 수준이지만 루마니아, 슬로바키아와 같은 EU 내의 몇몇 신흥 경제국과 같은 수준이다. 향후 브라질의 ICT 보급률이 높아지고, 산업구조가 ICT 집적 구조로 변화할수록 이 비율은 더

높아질 것으로 예측된다.

다. 데이터 기업 수

데이터 기업이란 데이터를 공급하는 기업으로 데이터를 생산 및 전달하는 기업이다. 즉 데이터 관련 제품, 서비스 및 기술에 대한 데이터를 공급하는 기업을 말한다. 데이터 시장이 성장함에 따라 데이터 기업의 수도 지역에 따라 점진적 또는 폭발적으로 증가하고 있다.

[그림 2-3-4] 2013~2016년 주요 국가의 데이터 기업 수



미국, EU, 일본, 브라질의 데이터 관련 기업의 수는 종사자 수와 마찬가지로 2013년부터 꾸준히 증가하고 있다. 이처럼 기업의 수와 종사자 수가 함께 증가하는 것은 자연스런 현상이다. 미국의 데이터 기업 수는 2013년 약 27만 4,000개에서 지속적으로 증가해 2016년에는 약 28만 8,000개로 추정됐다. 미국의 데이터 기업의 수가 가장 많은 것 또한, 미국의 데이터산업 종사자 수가 가장 많은 이유와 같은 맥락에서 이해할 수 있다.

EU의 데이터 기업 수는 2013년 약 24만 개였으나 이후 꾸준히 증가해 2016년에는 약 25만 4,000개 수준이었다. 미국의 데이터 기업 수보다는 작지만 종사자 수의 차이보다는 상대적으로 차이가 작다. 이는 미국의 데이터산업은 대기업 및 중견기업이 중심이 돼 이끌어 가는 것과 달리 EU의 데이터산업은 중소기업, 1인 벤처기업 등 소규모 기업들이 기반이 되고 있는 것을 의미한다.

일본의 데이터 기업 수는 2013년 약 9만 6,000개였으나 2016년에는 약 10만 개로 증가했다. 장기적인 경제 침체를 겪은 일본이지만 데이터 기업 수가 꾸준히 증가했다는 것은 데이터산업에 대한 일본의 기대를 보여준다.

한편 브라질의 데이터 기업 수는 2013년 약 3만 4,000개에서 꾸준히 성장해 2016년 약 3만 6,000개로 증가하는 모습을 보였다. 브라질 내에서의 ICT 시장이 상대적으로 작고, ICT의 보급률 또한 매우 낮은 수준이라서 상대적으로 증가율이 낮게 나타났지만 이후 지속적으로 늘어날 것으로 전망된다.

3. ICT 투자 대비 데이터 분야 투자 비율

데이터산업에 대한 투자는 해당 산업의 발전 가능성을 엿볼 수 있는 지표이다. 특히 산업이 점점 다양해지고 있기 때문에 특정 산업에 대한 투자가 증가한다는 것은 해당 산업에 대한 관심 집중, 자본 유입으로 인한 R&D의 활성화, 노동인구의 유입 등 향후 성장을 기대하게 하는 요인이 되기 때문이다.

미국의 ICT 산업에 대한 투자 중 데이터산업이 차지하는 비율은 2013년 9.8%에서 2016년 13.8%로 증가했다. 데이터산업과 관련된 각종 지표에서 최상위를 차지했던 미국은 투자와 관련해서도 가장 적극적인 움직임을 보이고 있어 향후에도 높은 성장률이 예상된다.

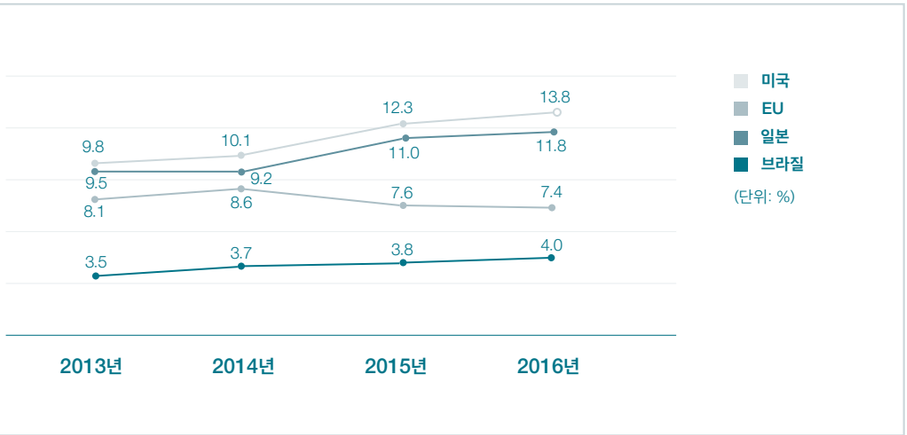
미국의 뒤를 이어 ICT 산업에 대한 투자 중 데이터산업이 차지하는 비율이 가장 높게 나타난 국가는 일본이었다. 2013년 9.5%에서 2016년 11.8%로 증가했는데, 경제적 불황이 심했던 2014년을 제외하면 꾸준히 증가하고 있음을 알 수 있다. 일본은 지금까지 매우 폐쇄적인 개인정보보호 제도를 갖고 있었으나, 개인정보보호와 관련한 법 개정을 통해 데이터의 유통이 활발해질 수 있는 환경을 조성했다. 동시에 빅데이터 등 신기술에 대한 소극적인 투자 기조에서 벗어나 적극적인 투자를 유도하는 등 데이터산업에 대한 높은 관심을 보여주고 있다. 이에 따라 향후 일본 데이터산업은 빠른 성장세를 보일 것으로 예측되고 있다.

브라질 역시 데이터산업에 대한 투자 비율이 증가하고 있는 국가 중 하나이다. 2013년 ICT 산업 투자 중 3.5%를 차지했던 데이터산업의 비율은 점진적으로

증가해 2016년 4.0%에 도달했다. 취약한 ICT 기반으로 인해 성장속도가 더딘 점이 있었으나 적극적인 투자를 통해 ICT 기반을 확대해나가는 한편, 데이터산업의 R&D를 활성화한다면 향후 높은 성장을 기대할 수 있을 것이다.

비교 대상이었던 미국, 일본, 브라질 모두 데이터산업 투자 비율이 증가하고 있는 것과 달리 EU는 2015년, 2016년 연속으로 감소하는 추세였다. 전체 ICT 분야에 대한 투자 중 8.1%를 차지했던 데이터산업은 2016년에 7.4%로 감소했다. 그러나 이는 2015년, 2016년 EU의 불안한 정세로 인한 투자 위축으로 인한 한시적인 현상으로 보이며, IDC는 2020년까지 데이터산업에 대한 투자 비율이 20.0%에 가까운 수준으로 매우 빠르게 증가할 것으로 예상했다.

[그림 2-3-5] 2013~2016년 ICT 투자 중 데이터 투자 비율



4. 데이터산업의 경제적 효과

가. 직접 효과

데이터산업의 경제적 효과는 데이터 시장이 경제 생태계 전체에 미치는 전반적인 영향이 어느 정도인가를 평가하는 분석 방법론을 의미한다. 디지털 기술이 발달함에 따라 데이터의 수집, 저장, 처리, 배포, 분석, 정교화, 전달 및 기타 이용이 용이하게 됐다. 이렇게 데이터의 활용이 용이해짐에 따라 데이터산업이라는 시장이 형성됐고, 이러한 시장은 직간접적으로 경제 전체에 영향을 미치게 됐다.

데이터의 경제적 효과는 직접 효과와 간접 효과로 구분된다. 직접 효과란 데이터산업 자체에 의해 형성된 효과, 즉 데이터 생산과 관련한 모든 비즈니스 활동에 의해 형성된 효과를 의미한다. 정량적인 직접 효과는 판매된 데이터 제품 및 서비스의 수익을 기준으로 측정된다.

미국의 데이터산업 경제적 효과 중 직접 효과는 2013년에 약 997억 달러를 기록한 이후 점차 증가해 2016년에는 약 1,225억 달러로 추정됐다. 즉 데이터에 기반을 둔 제품 및 서비스의 생산에서 파생된 직접적 영향은 EU를 비롯한 일본, 브라질보다 높게 나타났다. 이는 데이터 시장의 규모와 GDP에서의 비중이 상대적으로 크기 때문이다.

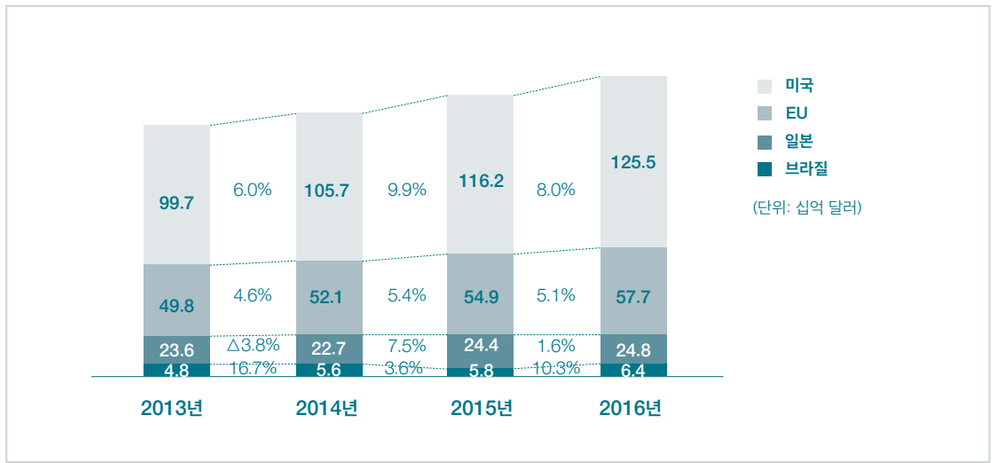
직접 효과는 데이터 사업 자체에 의해 형성되는데, 미국의 경우 데이터 생산에 매우 적극적이라는 것을 알 수 있다. 또한 미국의 데이터산업 직접효과는 EU의 직접 효과의 거의 두 배에 해당하는 수치인데, 이는 미국의 데이터산업이 좀 더 발전돼 있고, 경제적인 확산 속도가 더 빠르다는 것을 의미한다.

EU의 데이터 경제적 효과 중 직접 효과는 2013년에 약 498억 달러를 기록한 이후 점차 증가해 2016년에는 약 577억 달러로 추정됐는데, 연평균 약 5.0% 내외의 꾸준한 증가율을 보였다. 이는 시장 규모의 증가와 궤를 같이 하는데, EU의 데이터산업의 규모가 성장률의 큰 변동 없이 꾸준히 증가한 것처럼 직접 효과 또한 점진적으로 영향력을 높여 나가는 것으로 보인다.

일본의 데이터 경제적 효과 중 직접 효과는 2013년에 약 236억 달러를 기록했으나, 2014년에는 3.8% 하락한 약 227억 달러 규모를 기록했다. 2014년을 기준으로 데이터산업 종사자 수와 시장 규모, 경제적 효과 중 직접 효과 등의 감소 추세를 감안할 때, 일본의 데이터산업에 있어서 2014년은 부진한 해였던 것만큼은 틀림없어 보인다. 일본에서 데이터산업의 경제적 직접 효과는 그 성장률의 변동이 상대적으로 크게 나타나기 때문에 ICT 분야에 대한 투자, 경제 환경의 변화 등 외부적 요인의 영향을 크게 받는 것으로 보인다.

한편 브라질의 데이터산업 경제적 효과 중 직접 효과는 2013년에 약 48억 달러를 기록한 이후 점차 증가해 2016년에는 약 64억 달러로 추정된다. 이는 브라질의 ICT 산업이 빠르게 발전하면서, 그 기반이 되는 데이터 시장 또한 빠르게 성장하고 있기 때문이다. 물론 상대적으로 브라질의 데이터 시장 규모가 작다는 원인도 있으나 직접 효과는 2014년에 약 56억 달러를 기록해 2013년 대비 16.7%

[그림 2-3-6] 2013~2016년 주요 국가 데이터산업의 경제적 효과(직접 효과)



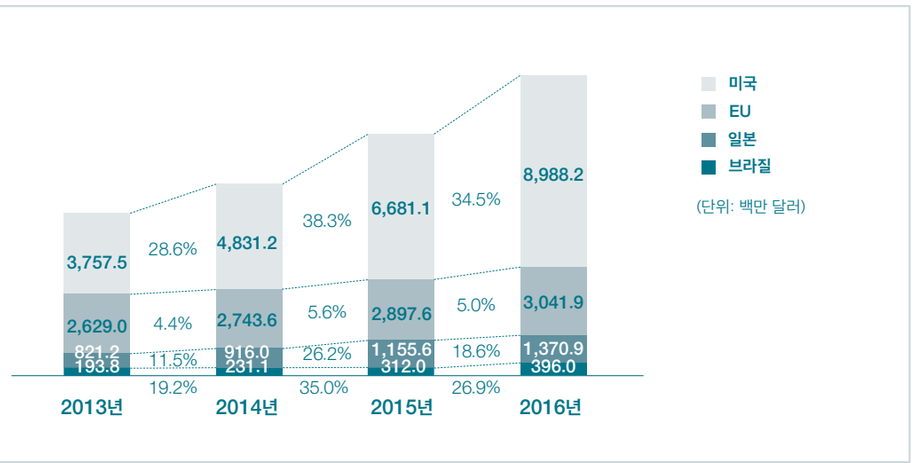
증가한 것으로 나타났다. 세계 경제가 불황에서 아직 벗어나지 못했는데, 브라질은 특히 경제 위기를 더 심하게 겪었다. 그럼에도 불구하고 데이터에 기반을 둔 제품 및 서비스의 영향력이 점차 커지고 있다는 것은 브라질 데이터 시장의 성장을 기대케 하는 요인이다.

나. 간접 후방 효과

직접 효과가 데이터 생산 활동이 데이터산업 내에 미치는 영향을 설명하는 것이라면, 간접 효과는 타 산업에 미치는 영향을 설명하는 것이다. 간접 효과는 전방 효과와 후방 효과로 구분된다. 후방 효과란 데이터산업에 속한 기업이 데이터 기반의 제품 및 서비스를 생산하는 활동을 하는 과정에서, 타 산업으로부터 자원을 제공받을 때에 자원을 제공해준 산업에서 발생되는 효과를 말한다.

반면에 전방 효과는 데이터산업에서 생산된 데이터 기반의 제품 및 서비스를 이용해 타 산업에서 다른 형태의 제품 및 서비스로 가공되는 효과를 말한다. 즉 타 산업에서 데이터를 활용해 제품 및 서비스를 제공함으로써 얻는 수익을 일컫는다. 전방 효과는 세 가지로 구분할 수 있는데, 첫째는 생산 및 공급 프로세스 최적화를 들 수 있다. 이는 데이터에 기반을 두고 프로세스를 통해 생산을 효율화한다. 둘째는 데이터를 활용해 표적 광고 및 개인 맞춤형 광고를 제공함으로써 마케팅 기술을 향상할 수 있는 것이다. 즉 타 산업에서 데이터 분석을 통한 마케팅

[그림 2-3-7] 2013~2016년 주요 국가 데이터산업의 경제적 효과(간접 후방 효과)



을 함으로써 수익이 오르는 효과가 발생한다는 점이다. 마지막으로 데이터 기반의 조직을 구성할 수 있는 점이다. 즉 데이터를 활용해 기존 조직의 관리 방식을 개선함으로써 불필요한 자원 소모를 최소화해 조직의 생산성을 증대할 수 있는 점도 전방 효과라 할 수 있다.

이 글에서는 간접 후방 효과를 중심으로 서술하고자 한다. 데이터산업으로 인한 간접 후방 효과는 매우 빠르게 증가하고 있는데, 특히 미국의 경우 2013년 약 37억 5,750만 달러 규모였으나 2016년에는 두 배 이상 증가한 89억 8,820만 달러를 기록했다. 증가율 또한 매년 약 30.0% 내외의 높은 증가율을 기록했다.

일본의 후방 효과도 빠르게 증가하는 추세다. 2013년에 약 8억 2,120만 달러 규모였던 후방 효과는 지속적으로 증가해 2016년에 약 13억 7,090만 달러를 기록했다. 미국과 마찬가지로 매우 빠른 증가율이 나타났으며, 매년 약 20.0% 내외의 증가율을 보였다. 이러한 현상은 브라질도 마찬가지다. ICT 기반이 상대적으로 취약하고, 규모가 작은 브라질 역시 2013년에는 약 1억 9,380만 달러 규모에 불과했으나 2016년에는 약 3억 9,600만 달러로 대폭 증가했다.

반면 EU의 후방 효과는 상대적으로 증가 추세가 매우 더딘 것으로 나타났다. EU의 경우 공공데이터 개방 정책을 통해 민간데이터보다 공공데이터 중심의 데이터산업이 상당한 비중을 차지하고 있고, 산업내에 1인 기업 및 소규모의 기업들이 많아 고용창출 효과도 상대적으로 낮은 것이 그 원인이라 판단된다. 물론 많

은 국가들이 데이터 시대를 맞이해 공공데이터를 적극적으로 개방하는 추세지만 EU는 특히 더 적극적인 것으로 보인다. 이에 2013년에 약 26억 2,900만 달러 규모였던 후방 효과는 2016년까지 점진적으로 증가해 약 30억 4,190만 달러를 기록했다.

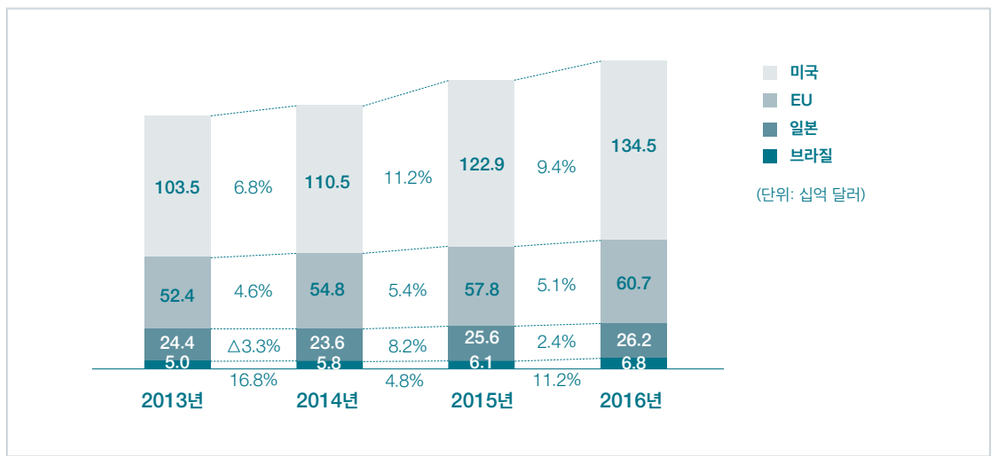
5. GDP 대비 데이터산업의 경제적 효과 비율

앞서 데이터산업의 경제적 효과를 직접 효과와 간접 후방 효과를 중심으로 살펴보았다. 직접 효과와 간접 후방 효과 모두 미국이 가장 앞선 것으로 나타났으며, 두 효과를 합한 경우 2016년 기준 약 1,345억 달러 수준이었다.

시장 규모와 마찬가지로 경제적 효과 또한 미국은 EU의 두 배 이상의 규모를 기록하고 있으며, 그 성장률 또한 더 높은 것으로 나타나 향후 격차는 더 벌어질 것으로 예상된다. 그러나 GDP에서 경제적 효과가 차지하는 비율을 살펴보면 미국은 2013년 0.58%에서 2016년 0.79% 사이의 수준이었으나 EU는 2013년 1.83%에서 2016년 1.89%로 같은 기간의 미국보다 높은 비율을 나타냈다.

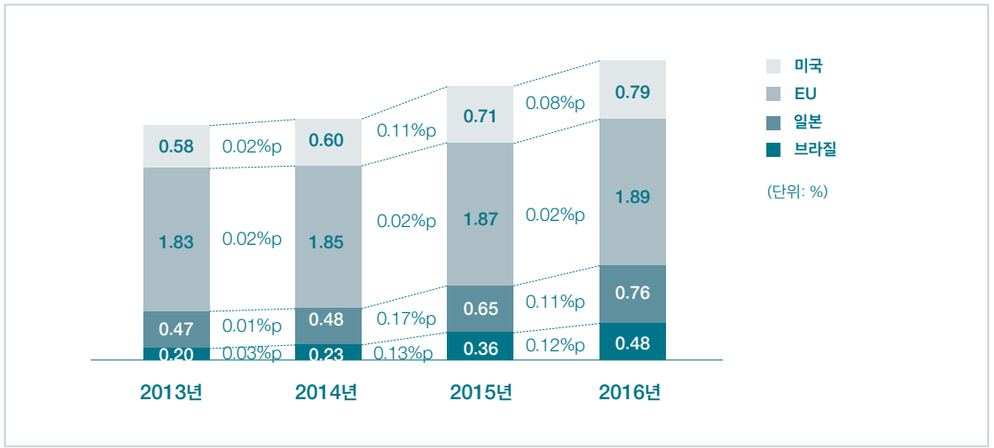
EU 데이터산업의 경제적 효과는 2013년 약 524억 달러 수준에서 점진적으로 증가해 2016년에는 약 607억 달러에 이르렀다. EU의 데이터산업의 특징은 상대적으로 미국·일본·브라질보다 GDP에서 차지하는 비중이 높은 반면, 경제적 효과의 증가율은 높은 편이 아니라는 데에 있다. 이는 높은 공공데이터 활용 비중과 소규모 벤처기업 중심의 산업구조 등이 원인 중 하나로 판단된다.

[그림 2-3-8] 2013~2016년 주요 국가 데이터산업의 경제적 효과(직접 효과 + 간접 후방 효과)



일본의 데이터산업의 경제적 효과는 2013년 약 244억 달러로 추정됐으나, 2014년에는 3.3% 감소한 약 236억 달러를 기록했다. 그러나 GDP에서 차지하는 비중은 소폭 증가했는데, 2014년에 경제적인 불황 및 투자 감소가 GDP 감소 및 타 산업에의 부정적 영향으로 이어져 데이터산업의 경제적 효과가 감소됐음에도 GDP에서 차지하는 비율은 커진 것으로 판단된다.

[그림 2-3-9] 2013~2016년 GDP 대비 데이터산업의 경제적 효과(직접 효과 + 간접 후방 효과)



한편 브라질의 경제적 효과는 2013년 약 50억 달러에서 2016년 약 68억 달러로 증가했으며, GDP에서 차지하는 비율 또한 2013년 0.2%에서 2016년 0.48%로 빠르게 증가한 것으로 추정돼 데이터산업의 빠른 성장을 보여주고 있다. ICT 기반이 취약해 데이터산업의 규모 및 경제적 효과, GDP 대비 경제적 효과 등 모든 것이 미국, EU, 일본보다 낮은 수치이지만 2억이 넘는 인구 수, 거대한 잠재 시장 등으로 인해 향후 발전 가능성은 높다고 판단된다.

2016 데이터 글로벌 이노베이터상 수상자

해외 진출 우선사항은 '현지화'

강용성 | 와이즈넛 대표

"제4차산업혁명의 핵심 동력으로 떠오르는 인공지능 시장에서 국내 우수 데이터 기업이 글로벌 기업과 경쟁하며 해외 시장에 진출할 수 있는 일이 많아지도록 먼저 진출한 선배 회사로서 열심히 길을 닦아야겠다는 생각을 한다."

2016년 데이터 대상 부문 데이터 글로벌 이노베이터상을 수상한 소감은.

먼저 감사드린다. 데이터 글로벌 이노베이터상은 해외시장 개척과 수출 확대에 성과를 낸 공로자에 수여하는 상으로 알고 있다. 데이터 글로벌 이노베이터상을 주신 것은 와이즈넛이 국내뿐만 아니라 해외에서 16년간 쌓은 기술력을 바탕으로 데이터 솔루션 시장 확대에 일조한 공로를 인정받았기 때문이라고 생각한다. 개인뿐 아니라 회사 차원에서도 최고의 영예이다.

빅데이터, 사물인터넷(IoT), 인공지능(AI) 등으로 데이터 산업계가 급변하고 있다. 생존이 아닌 성장을 추구하는 기업에도 영향을 미칠 거 같은데 어떤가.

제4차산업혁명은 산업계 전반의 핵심 화두다. 기존의 산업 분야에 ICT 기술이 융합되면서 새로운 부가가치 영역이 만들어지고, 국내 산업 생태계의 변화로 이어지고 있다. 모든 사물과 사람이 연결되고 통신하는 세상이 도래하는 상황에서 개인화된 인터넷 환경에 빠르게 대응하는 기업들이 변화를 주도할 것이다. 개별 제품과 서비스보다는 '소비자가 원하는 결과를 내놓을 수 있느냐'로 생존이 좌우될 것으로 본다.

와이즈넛같이 일찌감치 해외로 눈을 돌린 기업도 있지만, 해외 진출을 준비중인 곳도 많은 것으로 안다. 데이터 기업의 글로벌 진출과 관련한 이슈는 무엇인가. 이전과 비교했을 때 얼마나 변화했는지 궁금하다.

2004년 동종 업계에서는 처음으로 멕시코 전자정부에 솔루션을 수출한 적이 있다. 우리 회사로서는 글로벌 진출 첫 사례였기에 아무래도 그때와 현재를 비교하게 된다. 최근 1~2년 사이 국내 데이터 기업들의 글로벌 진출을 살펴보면 애로사항이 한둘이 아니다. 가장 큰 장벽은 해당 제품에 대한 글로벌 레퍼런스를 요구한다는 점이다. 이제 막 문을 두드리는 기업에게 이보다 더 큰 장벽은 없다. 예전보다 국내 데이터 기업의 위상이 높아졌다고 하지만 아직도 갈 길이 멀다. 먼저 진출한 선배 기업이 글로벌 시장에서 사업 기회를 확보하고 판로 개척을 열심히 해야겠다는 생각을 하는 요즘이다. 제4차산업혁명의 핵심 동력으로 떠오르는 인공지능 시장에서 국내 우수 데이터 기업이 글로벌 기업과 경쟁하며 해외 시장에 진출할 수 있는 일이 많아지길 바란다.

2016 데이터산업 백서에서는 데이터 기업의 성공적인 해외 진출을 위해서는 '현지 고객 정보 기반 판로 개척(38.5%)'과 '현지 고객의 요구 사항 파악(31.3%)'이 중요한 것으로 조사됐다. 어떻게 생각하나

현지화가 최우선 사항이다. 현지 고객 정보 기반의 판로 개척과 고객의 요구 사항 파악 등도 같은 맥락이라고 본다. 해외 진출 거점이 없는 경우라면 현지 파트너를 물색하고 시장 관련 정보와 인허가 등에 대한 정보를 파악해야 한다. 홍보 마케팅을 위한 적극적인 액션도 필요하다.

데이터산업 전문가가 되고 싶은 젊은이나 데이터 업계 입문자들에게 해주고 싶은 이야기가 있나.

데이터산업 전문가라 할 때, 먼저 데이터가 무엇이고, 어떻게 만들어지고, 어떻게 활용되는 등 데이터와 그 활용에 대한 전반적인 감을 익혀야 한다. 데이터산업 전문가는 학문적, 이론적으로만 만들어지는 것이 아니다. 현업에서 오랜 경험을 쌓아서야 비로소 데이터산업 전문가라고 말할 수 있다. 지름길은 따로 없다. 자신만의 포트폴리오를 만들며 한 단계 한 단계씩 경력을 쌓으면서 인정을 받다 보면 어느새 최고의 자리에 서 있는 자신을 발견하게 될 것이다.

제 **3** 부

데이터 서비스 동향

제1장

지능정보시대를 향한 데이터 서비스

인터넷과 스마트폰이 널리 보급됨에 따라 현대인은 비즈니스에서는 물론 사적 생활의 영역에서도 ICT를 광범위하게 사용하고 있다. 이때 데이터는 산업의 쌀과 같은 존재이며 데이터를 이용한 데이터 서비스는 전 산업의 영역에서 우리의 삶에 큰 영향을 주고 있다. 이 장에서는 데이터 서비스가 무엇인지 살펴보고, 데이터 서비스의 대표적인 활용사례를 개략적으로 살펴보고자 한다.

1. 데이터 서비스의 정의

데이터 서비스는 데이터, 정보, 지식을 활용해 사용자의 요구에 맞게 고안된 서비스를 통칭한다. 최근 인터넷과 모바일 기술의 발달로 현대인의 공적/사적 영역에서 수많은 행위가 데이터로 저장되고 있다. 개인의 검색 기록, 이메일 사용 내역은 포털 서비스에 고스란히 저장돼 사용자의 위치 정보와 결합된 후 개인의 취향을 컴퓨터가 먼저 제안하는 데이터 서비스인 추천시스템(Recommendation System)이 탄생됐다. 전자상거래와 전자정부의 발달로 과거에는 상상할 수 없는 양의 거래 데이터와 공공서비스 데이터가 저장되고 있으며, 기업과 정부는 이 데이터를 활용해 기업의 마케팅과 공공서비스에 대한 품질 만족 등 다양한 데이터 서비스를 개발·제공하고 있다.

데이터는 형태에 따라 정형 데이터, 반정형 데이터, 비정형 데이터로 나눌 수 있다. 데이터 서비스를 위해 정형 데이터는 물론 로그 데이터와 같은 반정형 데이터와 텍스트, 영상, 음성 등의 형태로 제공되는 비정형 데이터까지 포괄적으로 사

* 필자: 이정승(호서대학교 경영학부 교수)

용되고 있다. 이런 데이터를 기반으로 데이터 서비스를 제공하기 위해 전자지도와 같은 위치 정보 기술부터 최근 인간의 학습과정을 모방한 인공신경망(Artificial Neural Network) 등 인공지능(Artificial Intelligence, AI) 기술까지 다양한 기술이 사용되고 있다. 이러한 기술은 한 가지 기술만을 사용한 경우도 있지만 대부분 여러 기술을 융합한 경우가 많다.

2. 유형별 데이터 서비스

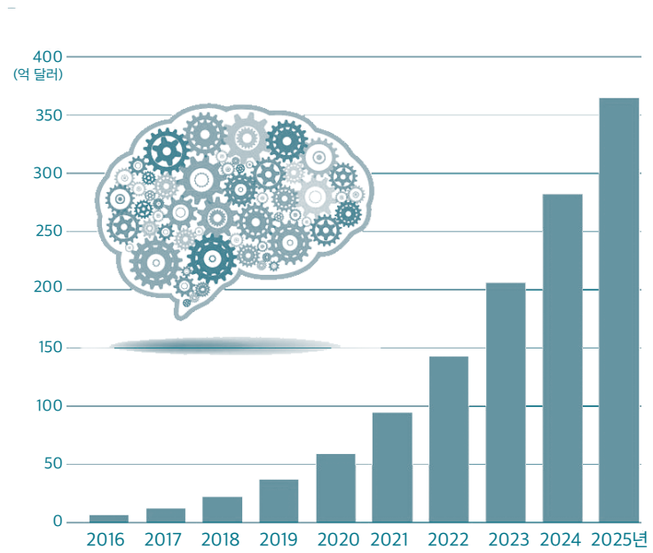
가. 인공지능을 활용한 지능정보 서비스

구글 딥마인드에서 개발한 바둑 인공지능 알파고(AlphaGo)는 최근 바둑 세계랭킹 1위인 커제를 3전 3승으로 제압하고 바둑계에서 은퇴했지만, 고도의 지적 게임인 바둑을 정복함으로써 인공지능의 가능성에 대해 엄청난 반향을 일으켰다. 과거의 기보 데이터를 기반으로 서치 알고리즘(Search Algorithm)과 딥러닝(Deep Learning)으로 학습한 알파고는 고도화된 데이터 서비스의 하나로 볼 수 있으며 퀴즈쇼에서 우승한 IBM 왓슨(Watson), 음성 인식 스마트홈 비서 서비스인 아마존의 에코(Echo), 음성인식 개인비서 서비스인 애플의 시리(Siri)는 모두 텍스트, 영상, 음성과 같은 비정형 데이터를 기반으로 학습한 인공지능형 데이터 서비스임을 알 수 있다. 제4차산업혁명 시대에 인공지능은 혁신을 가져올 ICT 기술의 핵심으로, 시장조사기관 트랙티카(Tractica)의 '글로벌 인공지능 시장 예측' 보고서에 따르면 세계 인공지능 서비스 시장 규모는 2025년까지 급속도로 커질 전망이다.

개인 단위에 스마트홈 서비스가 있다면 공공의 영역에는 도시 단위의 스마트시티(Smart City) 서비스가 있다. 자동차 번호판, 사람, 화재 인식 CCTV와 각종 사물인터넷(IoT) 센서를 통해 취합된 도시 생활 데이터는 스마트폰이나 스마트홈에서 개인이 사용한 서비스 및 그 결과 데이터와 결합해 최적의 교통, 환경, 안전, 에너지 관련 서비스를 해당 도시의 시민에게 맞춤형으로 제공할 수 있다. 이때 무인자동차는 안전 운행을 위해 V2V(Vehicle-to-vehicle)는 물론 V2X(Vehicle-to-everything) 방식으로 데이터를 처리해야 하는데 스마트시티는 이런 서비스를 위한 기반구조라고 볼 수 있다.

기업에서는 사물인터넷 센싱 및 데이터 분석을 통해 공장을 최적화하고 휴머

[그림 3-1-1] 세계 인공지능 서비스 시장 규모와 전망



출처: 트랙티카(2016), IPnomics 재인용

노이드 로봇으로 작업자를 대체하는 스마트 팩토리(Smart Factory)가 점차 퍼져나가고 로보어드바이저로 펀드매니저를 대체하는 인공지능 펀드가 값싼 수수료를 무기 삼아 시장에서 경쟁하고 있다. 아마존과 넷플릭스와 같은 디지털 기업은 고객 구매와 리뷰 데이터를 기반으로 한 데이터 서비스인 추천시스템을 넘어 인공지능 비서 스피커를 팔고, 직접 드라마와 영화를 제작하는 수준에 이르고 있다.

나. 위치정보를 활용한 서비스

위치정보를 활용한 데이터 서비스에서 핵심은 전자지도다. 구글은 최근 구글 검색과 구글 지도를 이용해 OTA(Online Travel Agency) 사업에 직접 뛰어들었다. OTA는 온라인 상에서 각 숙박업소의 예약을 대행해 예약대행의 대가로 수수료를 받는 사업인데 전자지도와 호텔 예약 데이터를 결합한 대표적인 데이터 서비스의 예이다.

각 기업은 이 전자지도에 자사의 데이터를 결합해 다양한 형태의 데이터 서비스를 제공하고 있다. 전자지도에 교통 상황과 최적 경로 알고리즘을 더하면 내

[그림 3-1-2] 전자지도와 호텔 예약 데이터를 결합한 구글 OTA 서비스

출처: https://www.google.co.kr

비게이션 서비스가 되고, 택시나 대리기사의 위치와 소비자를 연결해 주면 콜택시와 대리운전 서비스가 되며, 이 전자지도를 해양에서 날씨 정보와 결합하면 해양 내비게이션 서비스가 된다. 또한 위치정보를 활용한 데이터 서비스는 사용자가 사용하기 쉽도록 증강현실(Augmented Reality, AR)이나 가상현실(Virtual Reality, VR)로 사용자 인터페이스를 현실과 매우 유사하게 만들기도 한다.

다. 포털과 SNS를 활용한 서비스

구글이나 네이버와 같은 포털은 사용자의 검색기록은 물론 이메일, 클라우드, 위치정보를 결합해 다양한 데이터 서비스를 제공하고 있다. 그 중 검색광고가 대표적이며 오늘날의 구글과 네이버를 있게 한 최고의 수익모델이다.

포털은 검색기록과 연관 데이터를 기반으로 인터넷 쇼핑몰, 호텔 및 식당 예약, 부동산, 인터넷 및 모바일 게임 등 다양한 부가서비스를 통해 수익모델을 다양화하고 있다. 구인구직 및 경력 관리와 같은 특정 분야의 경우 일반적인 포털이 아닌 링크드인이나 잡코리아와 같은 특화된 포털이 데이터 서비스를 제공하기도 하며 페이스북, 인스타그램, 트위터와 같은 공개를 전제로 한 SNS의 고객 데이터는 텍스트 마이닝(Text Mining) 기법을 통해 고객의 구매 행태를 분석하고 고객을 고도로 세분화해 맞춤형으로 마케팅하는 데 사용되고 있다.

라. 공공데이터를 활용한 서비스

정부는 정부3.0과 제4차산업혁명 시대를 맞아 공공기관이 보유한 데이터 중 개인정보 식별과 국가 안보에 문제가 없는 경우 이를 민간에 전면 개방해 국민의 알권리를 충족시키고 비즈니스 기회를 제공하고 있다. 대표적으로 미국과 영국의 공공데이터포털을 참조해 공공데이터포털(data.go.kr)을 운영 중인데 파일과 API 형태로 공공데이터를 민간에 제공하고 있다.

[그림 3-1-3] 공공데이터포털 화면

출처: 공공데이터포털(https://www.data.go.kr)

이외에도 정부와 공공기관의 행정 서비스를 전산화함으로써 국민에게는 언제 어디서든 편리한 공공데이터 서비스를 제공하고 있으며, 각종 행정전산망에 축적된 민원 데이터를 분석해 공공서비스에 대한 대국민 품질 만족도를 높이고 있다.

마. 보건의료 데이터를 활용한 서비스

보건의료 분야는 대표적인 데이터 서비스 분야로 양질의 데이터가 풍부하고 활용 분야도 다양하다. 우선 다양한 종류의 웨어러블을 통해 생체 데이터를 직접 생성 및 분석하는 스마트 헬스케어가 있다. 최근 IBM 왓슨(Watson)과 같은 인공지능 기술의 고도화로 의료인의 진단 및 처방에 인공지능의 지원 및 대체 가능성에 대한 논의가 활발하게 진행되고 있다. 의료는 인간의 생명을 다루는 만큼 신중한 접근이 필요하다.

대용량의 데이터를 처리해 DNA를 판독하는 게놈 프로젝트, 전염병의 확산 경로 및 그 심각성을 예측하는 공공의료 분야는 대표적인 데이터 서비스의 예이다.

바. 비즈니스 데이터를 활용한 서비스

기업에서는 각 기업의 비즈니스 경쟁력 확보를 위해 일찍부터 가격 비교, 추천서비스, 고객관리(Customer Relationship Management) 등의 데이터 서비스를 활용하고 있다.

최근 핀테크 기술의 발전과 관련 규제의 완화를 통해 P2P 소액대출, P2P 환전 벤처기업들이 시장에서 치열하게 경쟁하고 있으며, 인터넷 전문은행인 카카오뱅크와 K뱅크는 데이터 분석을 활용해 금융 서비스 영역을 확대하고 있다.

온라인 마켓과 오프라인 소매회사들의 경계는 허물어져서 온라인과 오프라인 통합 환경에서 경쟁하고 있다. 인터넷 쇼핑몰, 할인마트, 백화점, 편의점 등은 개별 기업 혹은 제휴 기업단위로 회원가입과 마일리지 적립을 통해 고객 데이터를 관리하고 신용카드, 모바일카드, 상품권 등 다양한 전자결제 수단을 통해 편리한 서비스를 제공하고 있다.

3. 맺음말

본 장에서는 데이터 서비스를 정의하고 기술별, 응용분야별 활용사례를 살펴보았다. 제4차산업혁명 시대를 맞아 데이터 서비스는 인공지능 기술을 수용해 고도화되고 있으며 포털과 SNS, 공공, 보건의료 분야에서 다양한 비정형 데이터까지 처리해 혁신적인 데이터 서비스를 제공하고 있다.

제2장

선진 데이터 서비스 사례와 우리의 과제

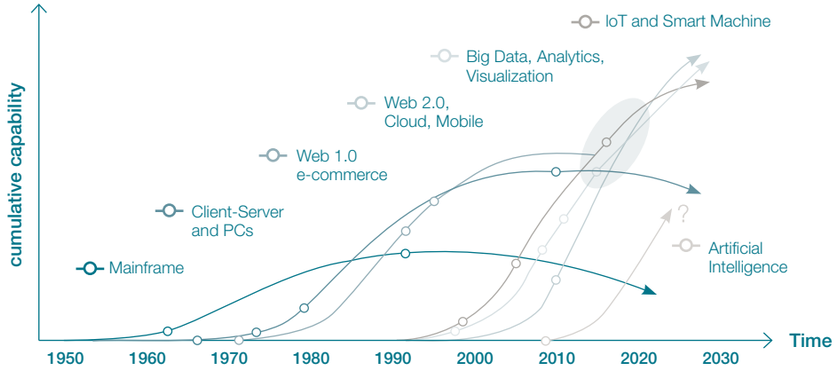
2015년 중국의 알리바바(Alibaba) 그룹의 마윈(Ma Yun) 회장은 "정보기술시대(Information Technology)가 저물고 데이터 기술(Data Technology)에 기반을 둔 시장이 열릴 것"이라고 얘기한 바 있다. 그로부터 2년이 지난 지금 인공지능(Artificial Intelligence, AI), 사물인터넷(IoT), 빅데이터, 클라우드 등의 용어와 함께 데이터는 전 세계를 요동치게 하는 핵심 키워드로 회자되고 있다. 데이터로 시작되는 제4차산업혁명은 과연 어떻게 진행될 것인지, 데이터 서비스의 미래를 위해 무엇을 준비해야 하는지 데이터 기반 비즈니스 현황과 선진 사례를 통해 살펴본다.

1. 데이터 서비스 개요

2016년 다보스포럼(세계경제포럼, WEF)에서 제4차산업의 도래를 알렸고, 올 1월 연차 총회에서는 제4차산업혁명의 주요 기술들이 융합을 통해 공유경제(sharing economy)와 온디맨드 경제(on-demand economy)의 기본이 되는 디지털 플랫폼 기반 기업들을 통해 '제4차산업혁명'이 빠르게 진행될 것이라고 예측했다. 즉 사물인터넷(IoT), 모바일, 인공지능 등 기술이 비약적으로 발전한다고 전망했다. 기술의 결합으로 지수적(exponential) 성장을 달성하면서 기술적 역량이 빠르게 강화될 것으로 전망했다. 이러한 기술 진보는 제품의 가격을 하락시켜 소비자의 다양한 욕구를 충족해줄 것으로 전망된다.

* 필자: 양정식(투이컨설팅 이사)

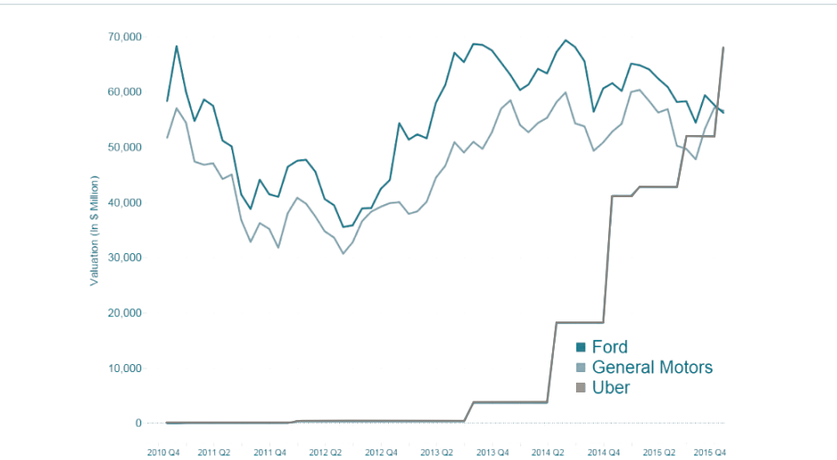
[그림 3-2-1] 기술융합에 따른 변화의 속도



출처: World Economic Forum(2016), *Digital Transformation of Industries-Demystifying Digital and Securing $100 Trillion for Society and Industry by 2025*

기술력과 플랫폼을 갖춘 기업이 등장하면서 기업들간 경쟁은 심화될 것으로 예측된다. 포춘(Fortune)이 선정한 500대 기업은 시가 총액 10억 달러 이상을 달성하는 데 평균적으로 20년이 걸렸지만, 최근 구글(Google)·우버(Uber)·샤오미(Xiaomi) 등 디지털 기술 기반 기업들이 이를 달성하는 데는 불과 8.1년, 4.3년, 1.7년밖에 소요되지 않았다고 한다.

[그림 3-2-2] 우버의 시가총액 10억 달러 도달 연수



출처: Uber's Financials, Forbes, 2016

제4차산업혁명의 핵심은 ICT가 일상의 모든 영역으로 편재돼 사람과 사람, 사람과 기계(시스템), 기계와 기계가 디지털적으로 연결돼 모든 행위가 시스템적으로 제어되고 관리되는 것이라고 볼 수 있다. 디지털적으로 연결되는 사회에서 핵심은 과연 무엇일까? 경희대 박주석 교수는 "정보화 사회의 인프라가 컴퓨터와 통신이었다면 초연결사회와 지능정보사회의 인프라는 데이터"라고 강조하고 있다. 이렇게 급변하는 데이터산업에서 어떻게 방향을 설정하고 대응할지가 매우 중요한 생존전략이 되고 있다.

한국데이터진흥원에 따르면 광의로 바라본 국내 데이터산업 시장 규모는 2016년은 13조 6,832억 원으로 전년 대비 2.5% 성장할 것이라고 한다. 2010년 이후 연평균 증가율은 8.0% 이상으로 꾸준히 증가하는 추세다. 그리고 데이터 서비스 시장도 2016년 6조 6,305억 원 규모를 형성하며 2013년 이래 연평균 성장률 8.3%로 증가 추세다.

[표 3-2-1] 광의의 데이터 서비스 시장 규모 (단위: 억 원, %)

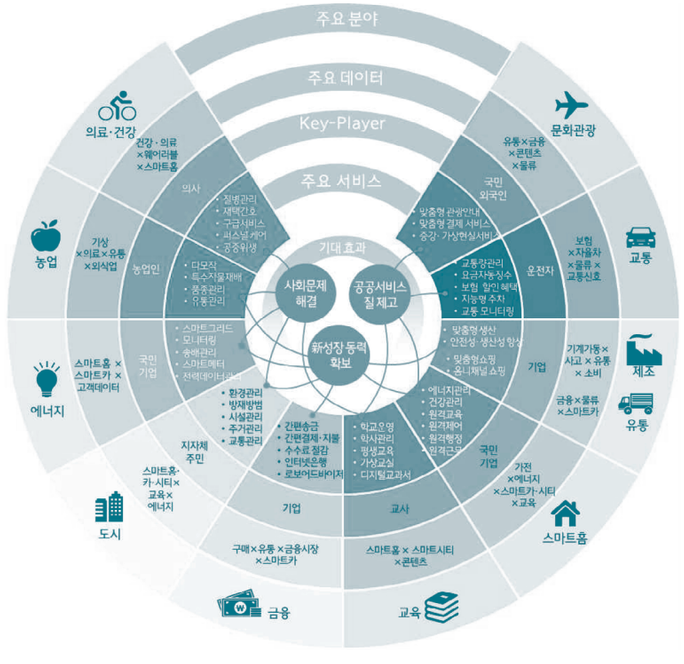
구분	2013년		2014년		2015년		2016년(E)		증감률 '15~'16	CAGR '13~'16
	규모	비중	규모	비중	규모	비중	규모	비중		
데이터 거래	1,078	3.3	2,019	3.5	2,379	3.7	2,578	3.9	8.4	14.7
정보 제공	48,284	92.4	51,985	90.7	58,171	90.7	59,921	90.4	3.0	7.5
데이터 분석 제공	2,266	4.3	3,325	5.8	3,601	5.6	3,806	5.7	5.7	18.9
데이터 서비스 전체	52,258	100.0	57,329	100.0	64,151	100.0	66,305	100.0	3.4	8.3

출처: 「2016년 데이터산업 현황 조사」, 미래창조과학부·한국데이터진흥원, 2017. 4

데이터를 활용한 서비스의 범위가 확대되고 다양해지면서 데이터산업의 규모가 계속 커지고 있다. 데이터 서비스 산업은 정보 서비스, 데이터 거래, 데이터 분석 결과 등을 온오프라인으로 제공해 수익을 창출하는 데이터 기반 비즈니스라 할 수 있다. 제4차산업시대의 도래로 인해 데이터산업은 큰 파도를 넘고 있다.

기존의 데이터가 양적으로 폭증하고 센서 데이터에 음성과 동영상 데이터까지 새로운 사업 환경이 전개되는 가운데, 다양한 분야에서 활용되는 데이터 기반 비즈니스에는 어떤 유형들이 있고 어떻게 서비스되고 있는지 알아보자.

[그림 3-2-3] 주요 분야별 데이터를 활용한 서비스와 비즈니스



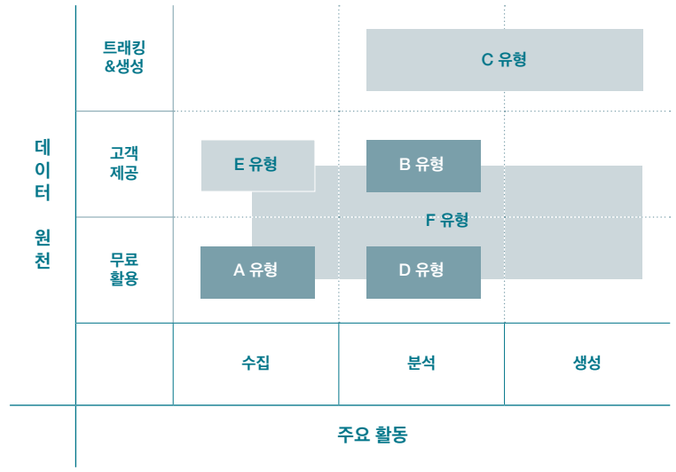
출처: 「지능화 시대의 새로운 생산 3요소: 데이터·AI·알고리즘」, 한국정보화진흥원, 2016. 7

2. 데이터 기반 비즈니스 현황

가. 데이터 기반 비즈니스 모델

데이터 기반(data driven) 비즈니스는 데이터를 자원으로 활용해 수익을 창출하는 기업을 의미한다. 데이터 기반 비즈니스의 유형은 데이터 원천, 주요활동 유형, 고객 구분, 수익 유형에 따라 다양한 형태가 나오게 된다. 캠브리지 모델, 유럽디지털 포럼 모델, TCS 모델 등이 있으며 국내는 2016년 한국데이터진흥원의 연구가 유일하다. 이 중 캠브리지 모델(Cambridge DDBM)은 2014년 100개의 글로벌 스타트업 회사를 추출·분석·분류한 연구결과로 데이터 원천, 주요 활동, 공급 형태, 목표 고객, 수익 모델, 특정 비용측면 이점으로 분석했다. 양적 분석 측면에서 현실적인 모델이라는 평을 받고 있다. 무료 데이터 수집 및 통합 유형(A 유형), 분석 서비스 유형(B 유형), 데이터 생성 및 분석 유형(C 유형), 무료 데이터 정보 디스커버리 유형(D 유형), 데이터 통합 서비스 유형(E 유형), 멀티소스 데이터 통합·분석 유형(F 유형) 총 6가지로 유형화된다.

[그림 3-2-4] 캠브리지 DDBM의 대표적인 6가지 유형



출처: 「주요국 데이터 주도 비즈니스 현황 및 비즈니스 유형 비교분석」, 한국데이터진흥원, 2016. 9

나. 국내외 데이터 기반 비즈니스 현황

한국데이터진흥원은 2016년 캠브리지 모델을 참고해 국내기업 310개 및 해외기업 379개를 분석했다. 캠브리지 DDBM 분류 기준 6가지 항목에서 기존의 특정 비용 측면의 이점을 제외하고, 산업별 분석 대상과 마켓플레이스·금융·마케팅·공공데이터 등 세부 서비스 13개 항목을 추가해 조사 및 분석을 했다.

국내 데이터 기반 비즈니스는 65%가 데이터 서비스의 초기 단계인 데이터 배포를 주로 수행하고, 데이터 분석 약 21%, 데이터 생성 6.5%, 데이터 시각화 및 통합 약 8%를 수행하고 있는 것으로 나타났다. 데이터 원천은 이용자 제공이 51%를 차지하고 있다.

[표 3-2-2] 국내 기업의 데이터 기반 비즈니스 현황 (단위: 개, %)

	데이터 기반 서비스											
	데이터 배포		데이터 분석		데이터 생성		데이터 시각화		데이터 통합·획득·처리		Total	
데이터 원천	개수	비중	개수	비중	개수	비중	개수	비중	개수	비중	개수	비중
기존재 데이터	C 40	12.90%	0	0.00%	0	0.00%	0	0.00%	0	0.00%	40	12.90%
무료 활용	E 15	4.84%	9	2.90%	11	3.55%	0	0.00%	7	2.26%	42	13.55%
이용자 제공	A 96	30.97%	B 47	15.16%	0	0.00%	8	2.58%	6	1.94%	157	50.65%
자가 생성	D 38	12.26%	5	1.61%	7	2.26%	0	0.00%	1	0.32%	51	16.45%
획득 데이터	13	4.19%	3	0.97%	2	0.65%	0	0.00%	2	0.65%	20	6.45%
전체	202	65.16%	64	20.65%	20	6.45%	8	2.58%	16	5.16%	310	100.00%

출처: 「주요 국가의 데이터 주도 비즈니스 현황 및 비즈니스 유형 비교분석」, 한국데이터진흥원, 2016. 9

해외 데이터 기반 비즈니스의 경우는 데이터 통합·획득·처리를 제공하는 서비스가 49%로 가장 큰 비중을 차지하고, 이어서 데이터 분석이 24%, 데이터 배포가 14%, 그 외 데이터 생성 및 시각화가 12%를 차지했다. 데이터 원천의 경우 무료 활용이 54%를 차지했다.

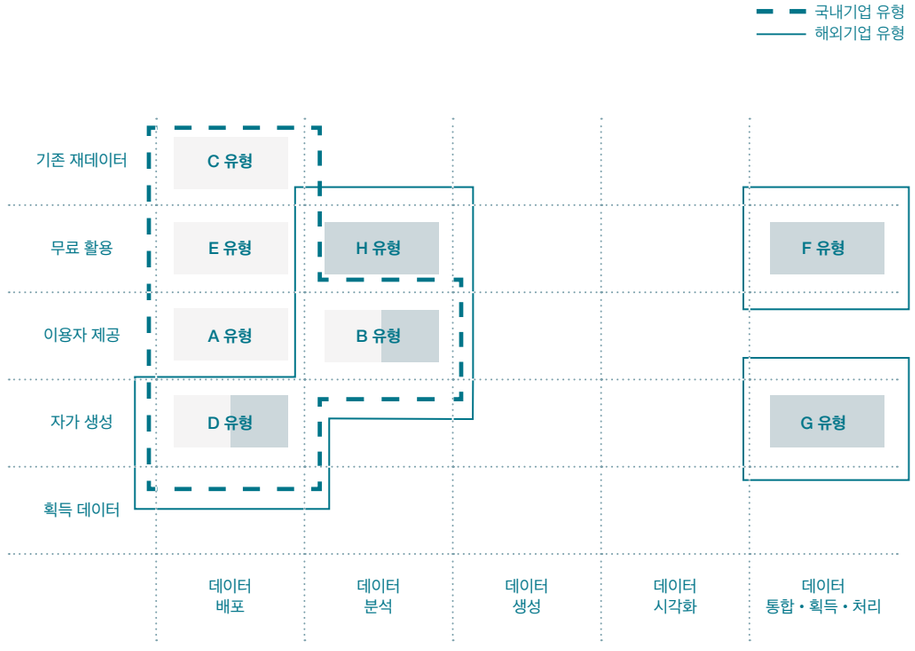
[표 3-2-3] 해외 기업의 데이터 기반 비즈니스 현황 (단위: 개, %)

	데이터 기반 서비스											
	데이터 배포		데이터 분석		데이터 생성		데이터 시각화		데이터 통합·획득·처리		Total	
데이터 원천	개수	비중	개수	비중	개수	비중	개수	비중	개수	비중	개수	비중
기존재 데이터	9	2.37%	0	0.00%	0	0.00%	0	0.00%	0	0.00%	9	2.37%
무료 활용	3	0.79%	H 28	7.39%	10	2.64%	15	3.96%	F 149	39.31%	205	54.09%
이용자 제공	3	0.79%	B 35	9.23%	0	0.00%	3	0.79%	8	2.11%	49	12.93%
자가 생성	D 39	10.29%	26	6.86%	3	0.79%	14	3.69%	G 30	7.92%	112	29.55%
획득 데이터	0	0.00%	3	0.79%	0	0.00%	0	0.00%	1	0.26%	4	1.06%
전체	54	14.25%	92	24.27%	13	3.43%	32	8.44%	188	49.60%	379	100.00%

출처: 「주요 국가의 데이터 주도 비즈니스 현황 및 비즈니스 유형 비교분석」, 한국데이터진흥원, 2016. 9

국내는 해외에 비해 데이터 배포 영역이 많은 비중을 차지하나, 해외는 분석 및 통합·획득·처리 활동 영역이 많은 비중을 차지했다. 특히 오픈 데이터와 같은 무료 활용 데이터 및 트래킹을 통한 자체 생성 데이터 활용이 가장 높게 나타났다. 국내는 이용자 제공 데이터를 분석해 결과를 제공하는 비즈니스 유형에 치중돼 있다. 데이터 기반 비즈니스 유형에서는 우리나라가 이용자 제공 데이터 생산에 치우쳐 있는 것으로 나타난다. 산업 전반에 데이터 활용 환경이 선진국 대비 열악한데 이를 극복하고 다양한 데이터 기반 비즈니스 모델을 발굴하기 위해서는 데이터 생산, 저장, 유통, 분석, 활용 등의 데이터산업 생태계가 조성돼야 한다.

[그림 3-2-5] 국내외 데이터 기반 비즈니스 동향 비교



출처: 「주요 국가의 데이터 주도 비즈니스 현황 및 비즈니스 유형 비교분석」, 한국데이터진흥원, 2016. 9

3. 해외 데이터 기반 서비스 사례

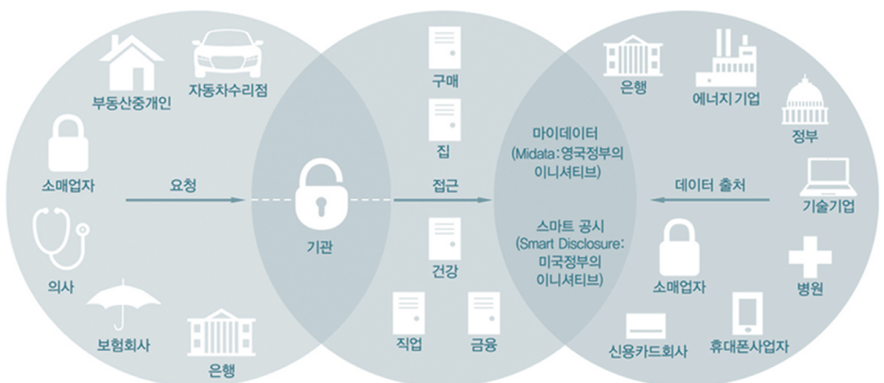
미국 등 해외에서 데이터 기반의 비즈니스를 성공적으로 수행하고 있는 사례와 기업의 분석을 통해 좀 더 의미가 있는 교훈을 찾고자 한다. 국내 스타트업도 활발하게 활동하고 있으나 서비스 영역이 한정적인 편이어서 해외 사례 중심으로 소개하고자 한다.

가. 개인데이터 활용 사례

개인정보와 관련된 데이터 사용은 엄격하게 규제되고 있다. 개인데이터를 활용하면 개인의 편의성 제고는 물론 새로운 산업을 창출할 수 있다는 점으로 인해 선진국에서는 다양한 정책이 시행되고 있다. 따라서 개인데이터 금고를 활용해 개인이 자신의 정보를 저장하고 소비자, 서비스 제공자, 정부로부터 데이터를 얻어 데이터를 원하는 기업에게 선택적으로 배포하는 방안이 추진되고 있다.

세계경제포럼에서는 "데이터를 가진 개인이 데이터를 수집·저장·안전하게 공유하고 그들의 삶에서 데이터로부터 가치를 얻을 수 있도록 기반을 제공해 준다."는 의견이 제기됐다. 소비자가 자신의 데이터를 제어하고 활용함으로써 즐거움과 이익을 얻게 된다는 것으로 영국, 프랑스, 미국 등이 적극적으로 추진하고 있다.

[그림 3-2-6] 개인데이터 금고 활용 개념도

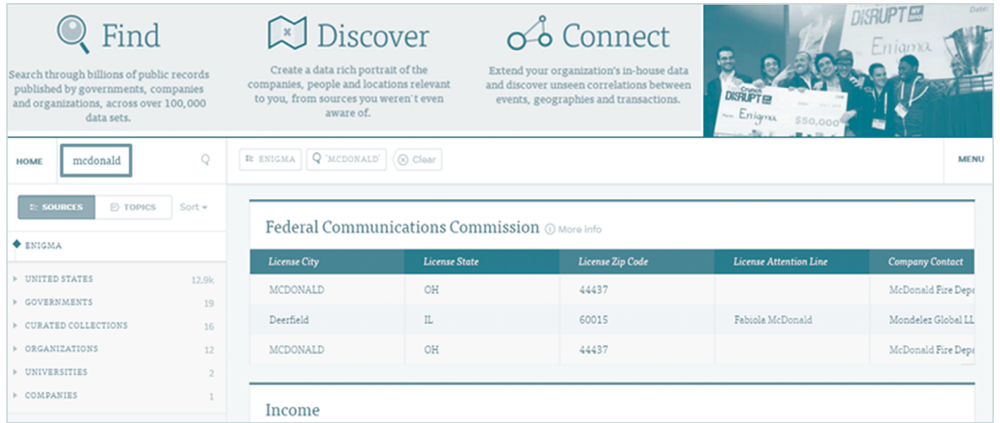


출처: 조엘구린, 「오픈데이터, 지금이 기회다!」, 투이컨설팅 번역, 2016

나. 대규모 데이터 수집·중개 서비스: 에니그마

에니그마(enigma)는 정부의 공공데이터(주 연방기록, 증권위원회자료 등 10만 가지 이상)를 광범위하게 중개해준다. 번거로운 프로세스와 제각각인 포맷(파일, CD, 문서)으로 접근성이 떨어지는 공공데이터를 통일된 포맷으로 공유해 분석 가능한 플랫폼을 제공해 데이터의 접근성을 용이하게 해준다. 더 나아가 데이터의 상관관계를 분석해 필요성을 인식하지 못했던 데이터까지 발견해 제공한다. 데이터 세트에 프로그래밍 방식으로 액세스 가능하도록 API 서비스의 제공 범위를 확장·제공하고 있다.

[그림 3-2-7] 에니그마 서비스 이미지



출처: http://www.enigma.com

다. 데이터 융·복합 서비스: 클라이미트 코퍼레이션

클라이미트 코퍼레이션(The climate corporation)은 미국 국립기상서비스(NWS)의 실시간 지역별 기온, 습도, 강우량 등 기상 데이터와 농무부의 지난 60년간 2평방마일 단위의 수확량과 토양 데이터로 빅데이터 분석을 한다. 이를 통해 세분화한 지역 날씨 조건을 농경지별로 추적·모니터링하고, 작물의 성장 추이와 토양의 습도 조건을 판별해 서비스 가입자에게 당일의 농사일을 가이드 해주는 서비스를 제공한다. 즉 농부에게 오늘 어떤 작업을 해야 할 것인지를 알려줌으로써 20% 이상의 수확량 제고를 목표로 한다. 이 회사는 지난 2013년 몬산토에 약 10억 달러에 매각됐으며 2015년 디지털 농업 플랫폼 서비스를 시작했다.

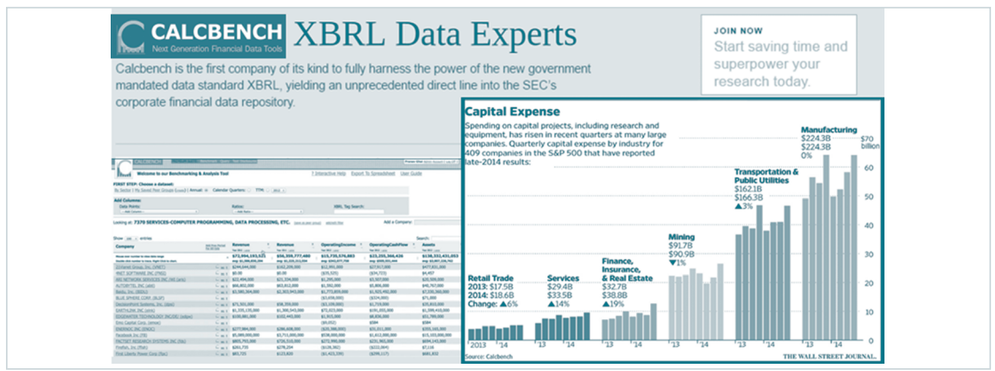
[그림 3-2-8] 클라이미트 코퍼레이션 서비스 이미지

출처: https://climate.com

라. 데이터 가공 서비스: 캘크벤치

캘크벤치(CALCBENCH)는 미국에 상장된 9,000개 이상의 기업에 대한 재무 데이터 조사 및 분석을 수행한다. 미국 증권 거래위원회(SEC)의 기업 재무 데이터 및 기록과 제공된 XBRL(eXtensible Business Reporting Language)을 통해 기계판독 형식으로 표준화해 데이터를 구축하고 재무 분석가, 투자자, 학자, 감사, 경제 전문가 등이 이 데이터를 활용해 산업분야 동향 및 리스크를 파악할 수 있도록 서비스를 제공한다.

[그림 3-2-9] 캘크벤치 서비스 이미지



출처: https://www.calcbench.com

마. 스마트 공시를 위한 선택 엔진 서비스

현재 쏟아지는 수많은 정보는 신중한 쇼핑을 원하는 소비자를 곤경에 빠뜨리고 있다. 스마트 공시(smart disclosure)란 쇼핑, 헬스케어, 금융서비스, 교통, 교육, 의료, 부동산, 여행과 같이 복잡한 영역에서 정보에 입각해 소비자가 올바른 선택을 하도록 돕는 새로운 유형의 데이터 기반 서비스이다. 영국이나 미국에서는 이미 활성화돼 있고 미국의 경우 이용자가 1억 명에 육박하고 있다.

해외 사례를 보면 단순한 데이터 수집 개념이 아니라 수요자가 요구하는 수준에 맞추어 클라우드 등 용이한 환경으로 구축하고 적극적인 융복합을 통해 활용도 높은 서비스를 제공하는 것을 볼 수 있다. 제공되는 데이터의 한계, 즉 데이터 항목·포맷·품질·표준·분석력 등의 문제를 해결하기 위한 적극적인 방안으로 비즈니스 모델을 만들어가는 모습이다. 개인데이터 활용 환경을 갖추기 위해 우리나라도 준비해야 한다. 개인데이터를 활용하면 개인에게 제공되는 서비스는 한층 더 정교해지고 만족도가 올라갈 것이고, 다른 사람의 서비스에도 확실한 해법을 제공할 것으로 예상된다. 분석력과 성과를 높이기 위한 시민과학자의 참여 및 대중의 힘(크라우드 소싱), 전문가의 깊이(오픈 이노베이션)를 서비스 환경으로 전환하기 위해 전문기업의 등장도 참고할 만하다.

국내외 사례를 통해 제4차산업을 성공적으로 맞이하기 위해서는 다음과 같은 도전이 우리에게 예정돼 있다.

첫째, 데이터의 완전성을 달성하기 위한 공공데이터뿐만 아니라 민간데이터를 포함한 오픈데이터 구축이 필요하다. 금융이나 통신 데이터 등의 민간데이터와 각종 시장 및 기업정보, 수요가 높은 공공데이터, 지역기반 데이터에 대한 확충이 시급하다.

둘째, 크롤링 등 데이터 수집에 대한 체계적이고 자동화된 기법 개발과 이를 위한 수집 표준의 수립도 필요하다. 그리고 크라우드소싱을 위한 민관 협력 및 트레킹 데이터의 확보도 필요하다.

셋째, 사용자 수요에 의거한 데이터 통합이 미흡해 데이터 추출방법에 대한 접근 방법도 개선돼야 한다. 그리고 고객에 대한 범위 확대 부분이 중요한다. 즉 B2B 고객의 확대가 관건인데 데이터산업의 생태계와 직접적인 관련이 있다. 위의 여러 가지 문제가 해결되면 어느 정도 개선이 될 것으로 보인다.

4. 향후 전망 및 과제

우버는 택시를 소유하고 있지 않지만 세계에서 가장 큰 택시회사이다. 페이스북(Facebook)은 콘텐츠를 보유하지 않은 세계에서 제일 유명한 미디어 회사이고, 에어비앤비(Airbnb)도 부동산을 소유하지 않은 세계 제일의 숙박업소가 됐다. 소사이어티원(SocietyOne) 또한 화폐가 없는 금융회사이다. 이렇듯 데이터가 서비스 회사를 집어 삼키는 시대가 됐다. 소프트웨어 라이선스에 종속시키지 않고, 대신 자신을 풍성하게 하는 데이터의 가치로 몸값을 보유하는 것이 미래의 락인(lock-in: 고객이 제품 및 서비스에 대해 높은 충성심을 가지게 돼 계속 이용하는 것) 전략이 될 것이다. 지능정보시대의 데이터는 매우 중요해지고 있고 데이터에 기반한 비즈니스는 더욱 분화되고 심화될 것이다. 사람이 아닌 기계가 이해하고 판단할 수 있도록 서비스에 필요 데이터를 필요 시점에 사용할 수 있는 수준의 품질로 확보할 수 있어야 한다.

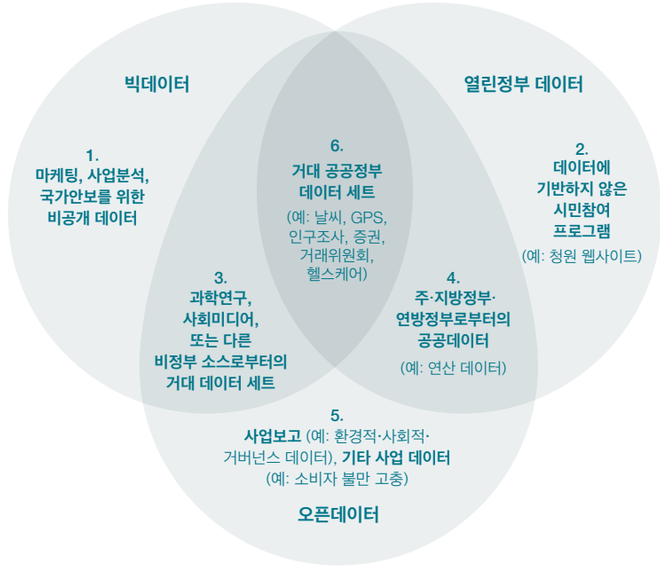
데이터는 활용 주체에 따라 빅데이터, 오픈데이터, 개인데이터로 구분된다. 빅데이터는 기업이 관리하며, 기업의 경영 목적을 위해 사용된다. 개인데이터는 개인에 관한 데이터로 개인 효익을 위해 사용된다. 오픈데이터는 누구나 접근할 수 있는 데이터이다. 오픈데이터는 협의의 오픈데이터와 정부기관이 공개한 열린 정부 데이터로 구분된다. 협의의 오픈데이터는 연구 데이터와 기업공시 데이터 등이 포함된다

기존의 민간데이터와 공공데이터, 사회관계망(SNS) 데이터 등을 포함해 큰 안목으로 접근할 필요가 있다. [그림 3-2-10]의 다이어그램이 3중으로 겹치는 부분의 데이터 세트는 다른 유형에 비해 가장 큰 영향력을 가지고 있다. 정부기관은 매우 거대한 양의 데이터를 수집할 능력과 예산을 갖고 있으며, 이러한 데이터 세트를 공개하는 것은 경제적 효용이 크다. 전국의 날씨 데이터와 GPS 데이터가 가장 자주 인용되는 예이다.

인공지능, 사물인터넷(IoT), 자율주행 등 지능정보사회에서 데이터 기반 비즈니스를 원활히 하기 위해서는 여러 가지 준비와 도전이 필요하다.

첫째, 데이터 수집은 원천 데이터 보유자가 제공하는 데이터의 수동적 수집에서 비즈니스 수요가 있는 데이터를 적극적으로 정의하고 이를 표준화·자동화해 수집·저장·가공해야 한다. 예를 들어 재무 데이터는 XBRL(eXtensible Business Reporting Language) 표준을 준수하도록 데이터 공급자에게 요청하거나 해당 데이터를 자동화한 크롤링 도구를 활용해 수집하고, 수집 데이터에 대해 가치를 증진시키는 활동을 해야 한다. 활용가치에 대한 조사·연구를 통해 통합 제공할 경우 성과를 높이는 데이터를 식별·구축·제공해야 한다.

[그림 3-2-10] 데이터 유형



출처: 조엘구린, 「오픈데이터, 지금이 기회다!」, 투이컨설팅 번역, 2016

둘째, 제공 형태도 고객 비즈니스 환경이나 채널을 검토해 다양하게 제공하는 기반을 갖추어야 한다. CSV, 오픈API, 링크드 오픈 데이터(linked open data) 등 기계 판독성이 높은 형식으로 제공해야 한다. 아울러 제공 플랫폼은 국제 표준을 고려해야 한다. 산업 내에서는 CKAN(Comprehensive Knowledge Archive Network, 오픈소스 데이터 매니지먼트 시스템)과 같은 표준을 따르도록 고려할 필요가 있다.

셋째, 고품질의 데이터 제공 서비스(Data as a Service, DaaS)는 갈수록 비즈니스 가치가 높아질 것이다. 이를 위해 산업군내 기준 데이터나 쉽게 구축하지 못하는 융복합 데이터로 표준을 고려해 구축하고, 이를 클라우드에서 제공하는 시도가 필요하다.

넷째, 인공지능이 활용되기 위해서는 과거 데이터를 통한 기계학습으로 지능을 생성해 학습된 룰 세트(rule set)를 알고리즘에 따라 추가해야 한다. 이의 성공에는 전문인력의 활동이 중요한 포인트가 된다. 따라서 수요 규모에 맞는 데이터 전문가가 필요하나 기업이 요구하는 수준의 데이터 인력과 배출되는 데이터 인력 수준이 일치하지 않는 문제가 발생하고 있어서 국가차원의 해법 모색이 필요하다.

마지막으로, 데이터가 국가 차원에서 만들어지고 활용됨에 따라 공공·민간 데이터 융복합 방법, 정제 기준, 가격 산정 등을 정해 다른 기관과 편리하고 빠르게 공유 및 거래할 수 있는 체계 마련이 필요하다. 아울러 국가 차원에서 중요한 데이터에 대한 메타데이터 관리, 연계 표준 관리가 필요한 시점이다.

최근 영국정부는 'Government Transformation Strategy 2017 to 2020'라는 정책백서를 발표했다. 이 백서는 시민과 국가의 관계를 근본적으로 변화시키기 위해 세계적 수준의 디지털 서비스를 지속적으로 제공하고, 정부가 운영하는 방식을 프론트 엔드에서 백오피스로 현대적이고 효율적인 방식으로 전환하는 것을 골자로 한다. 이 차원에서 데이터와 증거 기반의 서비스 개혁을 강조하고 있다. 이는 결국 시민과 기업에게 많은 혜택이 돌아갈 것이라고 강조하고 있다. 이는 제4차산업의 도래와 함께 각국이 데이터 기반으로 똑똑한(smart) 정부로 변신 준비 중이라는 것을 의미한다. 이에 따라 데이터 기반의 비즈니스는 더욱 발전할 것이므로 기업을 비롯한 경제 주체의 대비가 절실하다.

제3장

이용자 니즈를 이해하는 지능형 데이터 서비스

기술의 발전과 생활방식의 변화에 따라 데이터의 생산과 소비가 더욱 가속화되고 있고, 데이터의 수집 목적은 분석에서 학습으로 이동하고 있다. 데이터 서비스는 개인이 정보를 찾아내기보다는 원할 때 상황과 의도에 적합한 정보가 자동으로 제공되는 형태로 진화할 전망이다.

1. 변화하는 데이터 서비스 요구

자료와 정보를 기록하고 모으고 들여다 보는 것은 인류의 오랜 본능이다. 누구나 스마트폰이라는 고성능 컴퓨터를 들고 다니며 자기 자신도 모르게 데이터를 생산·소비하고 있다. 데이터 자체를 거래하거나, 데이터를 검색해 찾아내는 서비스는 이미 보편적으로 이용하고 있다. 데이터를 기반으로 수많은 서비스가 제공되고 있으며, 끊임없이 데이터를 생산하고 서버와 교류하며 이전에 없던 새로운 가치를 갖는 정보를 창출해 내고 있다.

정보 기록의 집합인 데이터를 기반으로 하는 산업이 자리잡기까지는 몇 가지 큰 사건들이 있었다. PC, 인터넷, 스마트폰 보급과 같은 사건들이 생길 때마다 디지털화한 데이터를 서비스하는 산업이 폭발적으로 성장하는 모습을 볼 수 있었다. 기존의 데이터 서비스는 데이터를 수집·저장하고, 저장한 데이터를 분석 또는 가공하고, 데이터와 그 분석 결과를 송출하거나 검색하는 각 단계가 산업의 성장과 함께 끊임없이 새로이 생겨났다. 데이터 서비스는 점점 더 다양해지고 있다.

* 필자: 안세준(카페인모터큐브 대표)

전 세계에서 가장 공격적으로 검색 서비스를 확대하고 있는 구글이 지난 2008년 발표한 바에 따르면, 당시에 구글은 이미 1조 개 이상의 웹페이지 인덱스를 보유하고 있었다. 그리고 이 무렵부터 검색엔진 서비스들은 인덱스 크기를 경쟁적으로 홍보하는 것은 무의미하다는 견해에 암묵적으로 동의했다. 글로벌 데이터 시장에서의 데이터 양은 이미 그 총량을 가늠하기 어려울 정도로 방대하며, 증가 속도 또한 지수함수적으로 증가하고 있다. 데이터 양이 방대하므로 검색 서비스가 제공하는 검색 결과 역시 너무 많아서 원하는 정보를 검색 서비스에서 찾아내는 것조차 쉽지 않은 일이다. 게다가 회선료와 저장매체 단가의 하락, 소셜 네트워크 서비스(SNS)의 활성화, 웨어러블 기기의 보급 등 기술 혁신과 시장 변화로 인해 새로이 생겨나는 데이터는 일정한 규격없이 제각각의 형태로 빠르게 쌓이고 있다.

다양한 온라인 서비스가 등장하면서 정보를 담은 텍스트 위주의 데이터를 넘어 사용자들이 각 서비스를 사용한 시각, 장소, 콘텐츠에 대해 보인 반응이나 행동, 일련의 행동을 취한 순서, 온라인과 오프라인의 행동 연결 등 예전에는 관심을 갖지 않거나 취급하지 못하던 다양한 형태의 데이터도 생겨나고 있다.

이렇게 막대한 양의 비정형 데이터를 데이터베이스에 쌓아 놓고 분석해 통찰을 얻어내는 것 역시 많은 시간과 자원을 투입해야만 가능하기 때문에 빠른 서비스 응답을 위해 새로운 방식의 처리 기술도 속속 등장하고 있다.

따라서 대량의 비정형 고빈도 데이터를 실시간 또는 실시간에 준하는 속도로 저장·분석·탐색하는 이른바 빅데이터 기술이 널리 각광받고 있다. 이 기술을 기반으로 기발한 아이디어와 고도의 분석 처리 로직을 구현해 개별 사용자가 필요로 할 만한 정보를, 필요한 시기에 애써 찾아보지 않아도 제공해 주는 종류의 지능적인 서비스를 내세운 혁신적인 벤처 창업이 이어지고 있다.

2. 학습을 위한 자원으로서 데이터

대용량의 비정형 데이터를 실시간으로 처리한다는 것은 데이터가 유입되는 시점에 기존에 존재하던 데이터와 새로 유입되는 데이터의 관계와 그 내용에 따라 분석한 결과를 저장할 수 있다는 의미다. 거대한 데이터를 여러 개로 쪼개어 분석한 결과를 병합하는 기법을 통해 이를 가능하게 한다. 막대한 자원과 시간을 투입해야만 분석 결과를 얻을 수 있었던 것과는 달리, 동적인 분석을 할 수 있으므로 효율성에 대한 염려를 줄이고 이전에는 수집하지 않았던 데이터까지도 수집할 수 있다. 기존의 데이터 수집은 분석의 원천 데이터로 사용하는 것이 목적이었던 것에 비해, 빅데이터 서비스에서의 데이터 수집은 데이터를 통한 학습이 주된 목적이 된다.

또한 기존에 따로 떨어져 있어서 연관 짓지 않았던 데이터도 빅데이터 플랫폼에 유입되면서 그 관계가 분석 및 학습되도록 함으로써 이전에는 시도한 적 없던 새로운 분석 결과도 얻을 수 있게 된다. 스스로 구축하거나 구입해 온 데이터를 분석하던 수준을 넘어, 법과 제도에 의해 구축된 공공데이터를 분석에 활용해 더욱 풍성한 분석 결과를 얻어내기도 한다.

●○
빅데이터 기술을 기반으로 기발한 아이디어와 고도의 분석 처리 로직을 구현해 개별 사용자가 필요로 할 만한 정보를 필요한 시기에 애써 찾아보지 않아도 제공해 주는 종류의 지능적인 서비스를 내세운 혁신적인 벤처 창업이 이어지고 있다.

3. 생명주기를 따라가는 지능형 데이터 서비스

먹거리에서도 즐거움과 스타일을 추구하는 문화적 변화와 함께 맛집 정보, 레시피 정보 등의 먹거리 정보 역시 정보 과잉의 상태에 들어서게 됐다. 서비스를 여러 번 이용한 사용자에게는 매 이용시점에 수집한 사용자의 행동 정보를 통해 지금 접속한 사용자가 음식에 대한 평가를 중시하는지, 접근성을 중시하는지, 청결함을 중시하는지에 따라 더 큰 만족을 얻을 수 있는 음식점을 추측해 맞춤형 정보를 보여준다.

음식점의 주인이 바뀌어 음식 맛과 매장의 분위기가 바뀌면 방문한 사람들이 인터넷 상에 생산하는 정보를 감지해 데이터를 축적하는 과정에서 빠르게 해당 음식점의 데이터가 바뀌게 된다. 빅데이터가 없던 시절에는 방문 경험이 있는 사람들의 입소문에 의지하던 정보가 빅데이터 서비스에 의해 자동으로 생성돼 이용할 수 있게 된 것이다.

맛집뿐이 아니다. 신용카드 사용 기록, 현금 인출 기록 등을 단문 메시지나 직접 연동된 카드사로부터 수집하며 영수증은 OCR(Optical Character Reader: 광학식 문자판독기) 등으로 인지가 용이하므로 영수증 사진을 수집하기도 한다. 수집하는 데이터를 소비 패턴과 취향을 분석에 활용해 쿠폰이나 보너스 포인트 혜택, 할인 이벤트 정보 등을 현재 위치에서 사용하도록 알려주는 서비스도 운영 중에 있다. 분석에 활용했던 소비기록을 재정렬해서 보여주면 가계부의 역할을 해낼 수 있고

각종 금융상품과의 연계도 가능하다.

빅데이터 분석을 통한 광고 리타기팅 서비스는 사용자들이 반응했던 이전의 광고 또는 방문했던 사이트, 온라인에서 수행했던 거래의 기록을 분석해 노출할 광고를 결정한다. 각 소비자의 행동 패턴만 분석하는 것이 아니다. 같은 상품을 구입한 소비자 집단의 행동 패턴을 분석하면 광고하고 있는 상품이 꾸준히 재구입해 사용하는 상품인지 한번 구입하면 재구입이 잘 일어나지 않는 상품인지를 알아낼 수 있다. 광고 노출에 어떤 단어를 사용하는 것이 높은 반응을 이끌어 낼 수 있는지도 분석이 가능하다.

광고를 내고 싶은 사업자에게는 높은 성과를 기대할 수 있는 광고 형태를 제안할 수 있으며 소비자에게는 정말 마침 필요한 상품의 광고를 보여줄 수 있으므로 광고를 유용한 정보로 인식하게 한다.

4. 디지털 트랜스포메이션

ICT의 빛을 그다지 크게 보지 못했거나 활용이 제한적이었던 분야에서도 혁신이 일어나고 있다. 주로 대기업의 디지털 트랜스포메이션(Digital Transformation) 전략에 따라 혁신의 결과물이 보급되는 모습을 보이지만, 그 기초가 되는 기술과 서비스는 데이터 서비스를 내세운 벤처 창업에 의해 시작한 경우가 많다.

자동차 정비나 중고차 시장과 같이 소비자들의 불신이 강한 곳에서 빅데이터 기술을 도입해 소비자의 니즈를 파악하고 객관성을 높인 정보를 서비스하는 벤처기업들이 있다. 이들이 취급하는 데이터는 자동차를 정비하거나 수리하는 현장에서 생성되는 이력 데이터뿐 아니라, 법규가 엄격한 자동차 분야에서 제공되는 공공데이터를 기반으로 한다. 고장이나 사고를 유발하는 운전습관을 찾아내고 안전하고 경제적인 차량 운행과 관리를 할 수 있도록 돕는다.

●○
자동차에 이미 장착된 200여 개의 센서로부터 운행 중에 생성되는 정보를 스마트폰이나 전용 통신 모듈을 이용해 수집하면서 분석에 활용하기도 한다.

완성차 제조업계도 이들의 서비스와 제휴·연동하거나 인수 합병하는 적극적인 모습을 보이며 별도의 통신 모듈을 장착하지 않고도 이런 분석이 가능하도록 통신을 포함한 여러 기능을 차량에 내장하는 제품을 계속 연구·개발하고 있다. 특히 전기자동차 분야는 '바퀴 달린 스마트폰'이라 부를 만큼 적극적으로 빅데이터 기술을 도입하고 있다.

최근에는 자동차를 구입해 사용하기보다는 사용료를 내고 계약 기간동안 이용하는 형태의 구매가 활성화되면서 중고차 거래 역시 활발해지고 있다. 거래량이 늘면서 차량의 관리 이력 데이터, 유사 차종의 거래 데이터 등을 토대로 중고차의 가격 시세 정보를 서비스하기도 한다. 전국에서 일어나는 거래정보를 수집하면서 지역, 성별, 연령, 차종, 차령 등에 따른 다양한 시세정보를 제공할 수 있게 됐다.

기업에서는 '구인난'이 심각하다고 느끼고 구직자들은 '구직난'이 심각하다고 하는 아이러니한 상황을 지능적인 데이터 서비스를 통해 해결해 나가기도 한다. 구직자의 이력을 이전 직장정보만이 아닌 온라인에서 인맥을 맺고 교류하는 사람들의 정보와 여러 가지 사회적 활동 기록과 교육 이수 정보도 함께 모아 분석한다. 심지어 거주지와 주요 방문 지역의 정보까지도 활용해 구인중인 업체를 추천해 주는 서비스도 있다. 구인 업체에서는 원하는 유형의 인재를 등록하면 적합한 인재를 추천해 준다. 이렇게 연결된 구인업체와 구직자 간 거래 데이터는 시간의 경과에 따라 점점 더 적합한 인재와 일자리를 추천할 수 있게 된다.

●○
대기업의 디지털 트랜스포메이션 전략에 따라 혁신의 결과물이 보급되는 모습을 보이지만, 그 기초가 되는 기술과 서비스는 데이터 서비스를 내세운 벤처 창업에 의해 시작한 경우가 많다.

5. 상황과 의도에 최적화한 데이터 서비스 예고

빅데이터 기술을 도입한 지능적 데이터 서비스의 혁신은 거의 전 산업분야에 걸쳐 빠르게 일어나고 있다. 기술적 한계와 고정관념으로 인해 기존에는 시도하지 않았던 새로운 관점에서의 데이터 수집과 분석 서비스가 반짝이는 아이디어와 최신의 빅데이터 기술로 무장한 벤처기업들에 의해서 생겨나고 빠르게 생활 방식을 변형시키고 있다.

기술의 발전과 생활방식의 변화에 따라 데이터의 생산과 소비가 더욱 가속화되고 있고 데이터의 수집 목적은 분석에서 학습으로 이동하고 있다. 개인이 정보를 찾아내기보다는 정보를 원할 때 상황과 의도에 적합한 정보가 자동으로 제공되는 형태로 진화해 갈 것이다. 이에 따라 개인이 생산하는 정보를 공개하는 범위를 한정하고 공개하지 않는 범위의 데이터를 학습에는 활용하지만 유출되지 않도록 하는 보안기술의 중요성 또한 강조될 것으로 전망된다.

2016 데이터 서비스 이노베이터상 수상자

“데이터산업계의 핵심 축으로 성장하겠다”

김덕조 | 시냅스엠 대표

“스타트업에서 중견기업으로 성장한 기업들에게 초점을 맞춰 정부의 지원책이 나와야 한다. 현재 투자가 집중된 스타트업이 시장에서 사업을 지속할 수 있는 생태계를 만들기 위해서는 중견기업과의 동반 성장 시나리오도 괜찮다.”

데이터 서비스 부문 이노베이터로 선정되었다. 선정된 배경이 무엇이라고 생각하나.

개인한테 주어지는 상이지만 엄밀하게는 지난 3년 새로운 형태의 서비스를 선보여 의욕적으로 시장을 개척한 점을 인정해서 주는 상이라고 생각한다. 무척 영광스럽다. 우리 회사는 엄밀한 의미에서 정통의 데이터 서비스 회사라고 하기엔 거리가 있다. 비디오 분야에 집중된 클라우드 서비스라고 할 수 있다. 기존 법률이나 날씨 정보를 제공하는 데이터 서비스 기업이 아닌 회사가 수상했다는 점에서 이례적이라는 반응이 있었다.

좀더 자세히 설명을 해달라. 데이터산업계에 새로운 유형의 기업이 수익 모델을 만들어 가고 있다는 점은 고무적인 일이 아닌가.

플랫폼을 빌려주는 PaaS(Platform as a Service, 서비스로서의 플랫폼) 개념이다. 현재 시냅스엠이 인터넷 생중계 전문기업과 진행하는 온라인 생중계 플랫폼 서비스를 예로 들면 이해가 빠를 수 있다. 이 서비스는 우리가 가진 실시간 라이브 방송 플랫폼을 기반으로 생중계 전문 촬영장비와 연출, 송출 장비, 전문 운용인력을 제공해 지상파 중계급 품질의 온라인 생중계 방송을 진행할 수 있는 서비스다. 네트워크 전송량과 통계 데이터는 물론, 개발자를 위한 API도 제공한다. 응용 프로그램을 직접 설치하는 방식이 아니라 인터넷을 이용해 응용 프로그램을 임대·관리해 주는 10여 년 전의 ASP(Application Service Provider) 개념에서 클라우드 서비스 형태로 진화한 개념이다. 인프라 구축에 드는 비용을 절감하고 서비스 요금만 지급하고 사용할 수 있으며, 별도의 유지보수 비용, 관리인력 등이 필요 없다. 이미 해외에서는 활발한 비즈니스가 일어나고 있는데 인건비가 낮고 거리가 협소한 우리나라 구조상 각광받지는 못했다. 그러나 앞으로 크게 기대할 수 있는 시장이라고 생각한다.

‘오픈형 동영상 콘텐츠 플랫폼’이라는 새로운 개념의 동영상 서비스로 사업을 시작한 것으로 안다. 우리나라에서 데이터로서의 동영상에 대한 니즈가 그만큼 많다는 얘기인가.

거의 10여 년 전, 누구나 동영상을 업로드하고 이를 즐길 수 있는 참여형 동영상 미디어 서비스를 표방했지만 10년이 지난 현재 구글의 유투브가 시장을 선점해 버렸다. 초기 동영상 서비스 시장을 만들었던 타 회사도 그렇겠지만 타격이 컸다. 하지만 동영상에 대한 사람들의 니즈는 갈수록 커질 것으로 생각해 동영상 관련 서비스를 내놓을 수 있었다. 누구나 사용할 수 있고, 사용한 만큼 비용을 지불하면 되는 클라우드 개념을 도입했다. 동영상 서비스를 시작할 때 초기 투자비용이 낮아 기업과 단체 등이 찾을 것으로 생각했다. 이번엔 B2C가 아닌 B2B다. 수익모델 측면에서도 바람직한 형태지만 아직까지는 성장을 위해 투자를 더 해야 하는 상황이다. 동남아와 중국 등 해외 시장 역시 차별화한 플랫폼과 콘텐츠를 앞세워 진출했다. 충분히 승산이 있는 시장이라고 본다.

데이터산업계에 허리가 되는 중견기업들이 드문 현실에서 정부에 바라는 점이 있다면.

데이터산업계, 특히 데이터 서비스 분야에서 과감하게 투자할 수 있는 민간기업이 얼마나 될지 의문이다. 수익의 원천이 되는 과금체계 하나를 보더라도 명백히 드러난다. 구글 같은 회사는 소비자들이 쉽게 서비스를 이용할 수 있도록 과금구조를 단순화한다. 소비자들의 생명주기가 변화하는 시점까지 기다릴 수 있는 자금이 뒷받침되어 있어서 가능한 일이다.

그런데 우리나라 같이 인터넷 서비스를 위해 비용을 지불하는 데 익숙해 있지 않은 풍토에서 데이터 서비스 기업은 극한 상황에 몰리게 된다. 과금체계를 섣부르게 결정할 수 없는 이유다.

100억~500억 원대 매출의 중견기업이 사업을 지속할 수 있도록 신경을 써주었으면 한다. 개인적으로 데이터산업 생태계에서 생존해 시장을 독점하는 것은 충분히 납득할 수 있는데, 제도적인 부분이나 상대적인 역차별 등으로 인해 독과점 기업이 생기는 것은 정부의 조율을 바란다. 글로벌 기업과 시장에서 경쟁할 때 특히 역차별이 없었으면 좋겠다. 스타트업에서 중견기업으로 성장한 기업들에게 초점을 맞춰 정부의 지원책이 나와야 한다. 현재 투자가 집중된 스타트업이 시장에서 사업을 지속할 수 있는 생태계를 만들기 위해서는 중견기업과의 동반 성장 시나리오도 괜찮다.

가능성이 큰 데이터산업계에 진출하려는 후배들에게 당부하고 싶은 말은.

기존에 자신이 가지고 있었던 사고체계를 완전히 바꿔야 한다. 아니 지워야 한다. 철저하게 소비자 입장에서 생각했을 때 시장에서 찾는 서비스를 만들 수 있다. 집중해야할 대상은 소비자다. 고객이다.

제 4 부

데이터 솔루션 동향

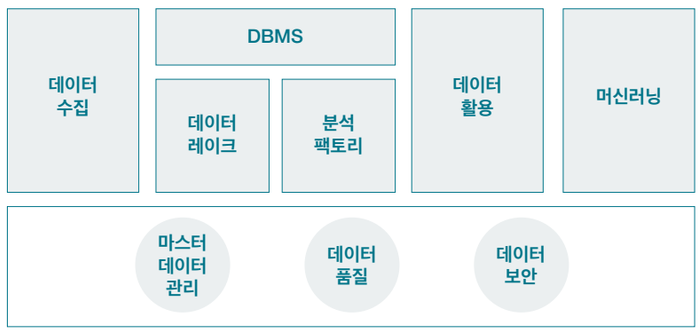
제1장

데이터 솔루션 아키텍처

데이터 솔루션은 데이터 수집, DBMS, 빅데이터의 단일 저장소인 데이터 레이크, 원시 데이터로부터 유용한 정보를 생성하는 분석팩토리, 데이터 활용, 마스터데이터 관리, 데이터 품질, 데이터 보안, 머신러닝으로 구성된다.

제4차산업혁명 시대로 진입하면서 빅데이터를 기반으로 한 머신러닝이 계속 성장할 전망이며, 정보기술 업체들은 기업이 보유한 비즈니스 데이터에 머신러닝을 쉽게 적용할 수 있는 솔루션을 내놓을 것으로 보인다.

[그림 4-1-1] 데이터 솔루션 아키텍처



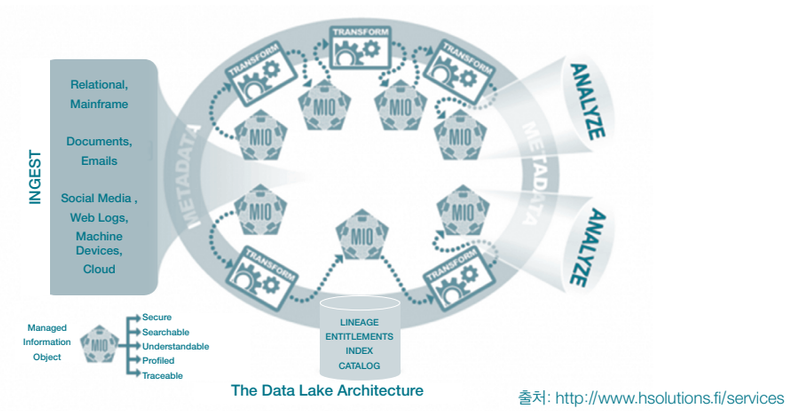
* 필자: 김종현(위세아이텍 대표)

데이터 수집 솔루션은 다양한 형태의 데이터들을 수집하는 데 있어서 어떤 데이터를 어디에서 어떤 방식으로 안정적으로 쉽게 가져올 수 있는가에 초점을 둔다. 데이터를 수집하는 것은 데이터의 형태와 종류에 따라 그 방법이 결정된다. 데이터 형태는 정형 데이터, 반정형 데이터, 비정형 데이터가 있고 데이터 종류에는 관계형 데이터, 이진 파일, 스크립트 파일이 있다. 데이터의 형태, 위치, 주기, 저장형태에 따라 데이터 수집 방법 또한 다양하므로 적절한 솔루션을 선택하는 것이 중요하다.

DBMS(Database Management System)는 응용 프로그램과 데이터의 중재자로서 모든 응용프로그램들이 DB에 접근해 공유할 수 있게 관리해주는 소프트웨어 시스템이다. 데이터를 효율적으로 저장함은 물론 갱신, 삭제, 읽기 등의 조작 기능과 제어 기능을 통해 데이터를 관리하는 프로그램의 집합체이다. DBMS의 장점은 데이터의 접근이 용이하고 데이터 통제와 보안이 강화된다는 점이다. 그러나 빅데이터의 등장에 따라 기존 관계형 DBMS에서 탈피해 NoSQL(Non-Structured Query Language) 기술을 적용하고 있으며, 다수의 기업에서는 관계형 DBMS와 NoSQL을 복합적으로 활용하고 있다.

데이터 레이크 솔루션은 데이터 원천 복제부터 시각화, 다차원 분석, 머신러닝 등의 작업에 사용하기 위한 데이터 변형에 이르기까지 기업의 모든 데이터를 단일 저장소에 저장하는 것이다. 데이터 레이크에는 정형 데이터, 반정형 데이터,

[그림 4-1-2] 데이터 레이크 구성



출처: http://www.hsolutions.fi/services

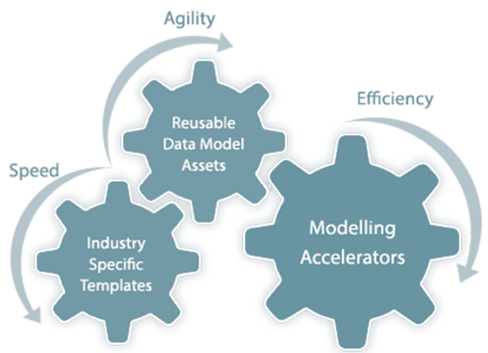
비정형 데이터, 이미지, 오디오 등 모든 형태의 데이터를 수용한다. 최근 빅데이터 활용이 주목 받으면서 다수의 사용자에게 각기 필요한 데이터를 제공하기 위한 데이터 레이크 구축 수요가 꾸준히 증가되고 있다.

분석팩토리는 데이터로부터 비즈니스에 유용한 의사결정을 내릴 수 있는 분석 모델을 만들어내는 기능을 수행한다. 분석팩토리 솔루션은 데이터를 수집하고 분석한 후, 이 결과를 반영한 기반 모델을 생성하는 것까지의 프로세스를 포함한다. 이러한 프로세스에서 비즈니스 규칙을 적용해 데이터에 대한 통찰력을 얻어 모델을 생성하는 것이 중요하다.

데이터 활용 솔루션은 사용자에게 가장 가깝게 다가와 있다. 사용자가 실질적으로 의사결정을 할 수 있도록 직관적이며 명확해야 한다. 때문에 다차원 분석(OLAP) 도구, 데이터 마이닝 도구, 빅데이터 분석에서 많이 활용되고 있는 시각화 도구들이 데이터 활용 솔루션으로 주로 활용되고 있다. 또한 데이터 분석에 대한 관심이 많아지면서 직접적으로 조작을 원하는 사용자의 요구를 충족시키는 추세다. 데이터 활용 솔루션은 빅데이터 분석의 애플리케이션 도구로 활용됨에 따라 앞으로 솔루션의 다양한 기능 강화와 동적 변화가 예상된다.

마스터데이터 관리 솔루션은 비즈니스 수행과 분석의 기반이 되는 마스터데이터를 전사 차원에서 통합적으로 관리함으로써 일관성 있고, 정확하게 이를 생성하고 유지하기 위한 도구이다. 이러한 솔루션에 필요한 요건으로는 데이터 저장·검색, 대용량 데이터 처리, 계층적 분류가 있다.

[그림 4-1-3] 분석팩토리의 3요소



출처: http://www.karvyanalytics.com/Expertise/Analytics-Factory.aspx

데이터 품질 솔루션은 품질 지표를 정의하고, 규칙 기반으로 데이터 품질을 진단해, 오류 현상의 분석 결과를 반영해 품질을 향상시킨다. 이러한 전통적인 데이터 품질 도구는 제4차산업혁명 시대가 도래하면서 획기적인 변화 요구에 직면해 있다. 세계 수준의 데이터 품질 SW는 머신러닝 기반의 이상값 탐지와 다르게 명명된 동일 데이터를 찾아내는 데이터 매칭이 지원되어 데이터 오류를 자동으로 정비하는 기능을 제공한다. 또한 비정형 데이터를 포함한 빅데이터의 품질을 관리하고, 나아가 산업별 빅데이터를 관리할 수 있어야 한다.

데이터 보안 솔루션은 기업 데이터의 기밀성, 무결성, 가용성이라는 목표를 달성하기 위한 솔루션으로 DB 접근 제어, DB 암호화, DB 작업 결재, DB 취약점 분석 기능을 제공한다. 이러한 기능을 기반으로 데이터베이스에 대한 접근 패턴을 학습하고 이상징후 시 이를 차단하도록 하며, 인가된 사용자라 할지라도 기업 DB 내의 개인데이터나 기업의 민감한 데이터를 조회하는 경우에 데이터의 유출 위험을 방지해 데이터의 기밀성을 지킨다. 또한 기존 인프라와 연계된 가상화·클라우드 환경에서도 데이터 무결성과 가용성을 유지한다. 현재 데이터 보안 영역은 보안 정보 이벤트 관리, 침입방지 시스템, 데이터 손실 방지와 같은 다른 보안 영역과의 통합을 통해 앞으로 더욱 성장할 것으로 예상된다.

머신러닝은 과거 데이터에서 숨겨진 패턴을 읽어내어 기계가 학습한 후 미래를 예측하는 기술이다. 머신러닝의 적용 분야는 음성과 영상을 인식하는 원천 기술 분야와 고객 분류, 이상징후 예측, 예측 정비, 큐레이션 등의 인공지능 비즈니스 애플리케이션으로 나눌 수 있다. 인공지능 비즈니스 애플리케이션은 산업별 도메인 지식, 데이터 전처리와 알고리즘 설정이 어우러진 다양한 형태의 도구로 발전될 전망이다.

제2장

데이터 수집 솔루션

최근 정보기술이 발전하면서 대량의 데이터를 분산 저장하고 데이터 생성과 동시에 분석할 수 있는 실시간 분석 플랫폼의 가치가 치솟고 있다. 실시간 처리 기술의 핵심은 데이터가 생성되면 곧바로 처리해야 한다는 것이다. 늘어나는 데이터를 수집·보관·가공·분석하기 위해서는 효율적인 IT 시스템을 구축해야 하며, 이는 빅데이터 수집 기술과 함께 당면한 문제를 해결할 수 있는 첫 관문이다.

1. 개요

점점 빨라지는 데이터 수집 속도가 데이터 비즈니스의 무궁무진한 가능성을 열고 있다. 동시에 출처조차 알 수 없는 정보들이 도처에 넘쳐나는 '정보 과부하 현상'도 심화되고 있다. 정확한 정보도 있지만 왜곡된 정보가 많다는 점이 문제다. 데이터의 폭발적인 증가로 인해 낮은 품질의 데이터 역시 늘고 있다. 이는 기업이 비즈니스 프로세스를 지원하고, 규정을 준수하며 정확한 정보에 입각한 의사결정을 내리는 데 점점 더 위협적인 요소로 대두되고 있다.

디지털 데이터는 어느 주체가 만들었는지, 즉 사람인지 기계인지 또는 어떤 업무 처리 프로세스 과정에서 자연적으로 만들어지는지, 생성 주체와 방법에 따라 형식이 결정된다. 더불어 데이터의 활용 목적에 부합하는 목적론적 특징을 갖고 있다. 이 둘을 명확히 숙지하는 것이 데이터 수집 기술을 정의하는 기본 골격이 될 것이다.

기업 활동의 결과 이루어진 데이터를 포함해 대부분의 데이터는 기업 내 여러 IT 시스템에 각기 다른 유형으로 존재한다. 또한 데이터 거버넌스 관점에서 비즈니스 혜택을 위한 정보 관리 및 이용에 대한 컴플라이언스 대응, 법 제도 범주 내에서의 데이터 공유와 정보 품질, 정보보호 및 정보 생명주기 관리를 포함한다. 탄탄한 데이터 거버넌스를 갖춘 기업들은 의사결정의 개선부터 컴플라이언스 등 규제 강화에 이르기까지 여러 가지 목표를 달성할 수 있다.

다양한 도구와 디바이스를 통해 방대한 양의 데이터가 생성되기 때문에 데이터 분석가는 어디에 어떤 데이터를 분석에 활용할 것인지에 대한 판단을 우선해야 한다. 데이터를 다루는 기술은 수집부터 데이터 준비, 데이터 가공, 정보보호를 위한 변환 등 처리 기술이 요구되며, 이러한 기술들은 상당히 빠른 속도로 발전하고 있다. 늘어나는 데이터를 수집·보관·가공·분석하기 위해서는 효율적인 IT 시스템을 구축해야 하며, 이는 빅데이터 수집 기술과 함께 당면한 문제를 해결할 수 있는 첫 관문이다.

2. 데이터 수집 기능

데이터 수집시 먼저 기업이 보유 중인 데이터가 무엇이고, 어디에 있는지 시스템 간 연관성을 먼저 파악해야 한다. 데이터 소스와 그 소스들을 서로 연계시키는 방법을 먼저 이해하기 위해서다. 결과적으로 이는 아카이빙, 테스트 데이터 관리, 데이터 프라이버시, 데이터 통합, 데이터 취합을 포함한 정보 프로젝트 성공을 위해 즉시 활용된다. 다행히도 전반적인 데이터 검색 프로세스 및 분석은 자동화될 수 있으며 이를 통해 시간을 절약할 뿐 아니라 수작업에 의한 오류도 줄일 수 있다.

가. 데이터 거버넌스

데이터 수집과 데이터 거버넌스 영역은 반드시 검토해야 할 영역이다. 데이터 출처 및 근원정보 관리를 위한 기술을 개발해 빅데이터의 버전 관리, 기관 인증, 출처 추적, 품질 관리, 변화 관리, ETL 작업 등 정형 데이터 가공작업에 대한 영향 분석, 생명주기 관리 등과 같은 데이터 신뢰성을 제고할 수 있는 방법을 검토해야 한다. 또한 분산 다중 사이트(다중 데이터 레이크) 메타데이터의 상호 연동이 가능하도록 구현하며, 빅데이터는 다양한 소스에서 다양한 형식으로 짧은 시간에 대용

* 필자: 신동선(데이터스트림즈 PPC본부 상무)

량 데이터가 쏟아지는 데이터를 잘 활용하기 위해 데이터의 형식을 사전에 정의해야 한다.

나. 데이터 품질 관리

다양한 유형의 빅데이터(스트리밍 데이터, SNS 데이터, 업무용 데이터 등)에 대한 품질 관리 지표에 의한 품질 측정과 모니터링, 다양한 데이터 소스로부터 생성된 데이터의 표준으로 XML 매핑, 일치성과 완전성을 제고할 수 있어야 한다. 빅데이터를 활용한 구조적 정형 데이터의 품질 향상, 사물인터넷(IoT) 응용 등 실시간 분석을 지원하기 위한 스트리밍 데이터 측정, 품질 기준을 정립하고 수집 시 기준에 의한 수집 데이터 품질 측정이 가능해야 한다. 빅데이터 예측 분석의 핵심 잣대인 데이터 충실도, 최신성, 신뢰성 확보를 위한 내외부 데이터 품질 기준을 정의하고 점진적 데이터 수집 및 저장 품질 확보에 중점을 두어야 할 것이다.

다. 메타데이터 관리

빅데이터 분석 효과 제고를 위한 내외부 데이터 내 메타데이터 관리는 비정형 빅데이터에 존재하는 마스터데이터 식별, Rationalization & Data enrichment 등 데이터를 보강하고, 수집된 소스 데이터를 통합·저장하기 위한 물리적 저장소를 포함한다. 또한 이기종 데이터 형식을 하나의 형식으로 변환하기 위한 일련의 통합 규칙과 논리적 연결고리를 다양한 환경에서 유연하게 관리할 수 있도록 지원한다. 데이터 수집 시 또는 수집 이후 통합 분석의 핵심 정보(마스터데이터)들은 싱글뷰를 제공하기 위해 기준정보 데이터 분류에 따른 식별자 속성을 정하고, 식별자에 의해 여러 시스템에 분산돼 있는 기준정보 데이터를 융합(Consolidation)해야 한다. 그 다음 단일 키 기반의 단일화된 버전으로 만들고 관련된 정보를 풍부하게(Enrichment)해 최신화 및 통합화(Rationalization)하는 작업을 수행한다. 이로써 누구나 다 같이 참조하는 기준정보를 하나의 사실(Single Versio of Truth)만 가질 수 있도록 관리할 수 있다.

라. 데이터 통합

통합 대상 정의와 데이터별 추출·가공·적재에 적용할 방법 및 도구·주기 정의, 소스와 타깃 간 매핑 정의서 관리, 실시간 배치 처리를 담당할 통합 엔진, 데이터의 프라이버시와 생명주기 관리 관점, 개인정보·민감정보 등 보안 및 보호가 필요한 속성에 대한 정책 및 관리 기준을 제시해야 한다. 빅데이터 플랫폼에 수집되는 모든 데이터는 수집 시점부터 폐기될 때까지 모니터링해야 한다. 데이터로서 역할을 다 하는 시점에서는 특정한 장소에 적절히 보관되거나 규정에 따라 폐기해야 한다.

앞서 언급한 데이터 거버넌스 요소들은 수집에서 데이터 처리 과정에 이르기까지 전반적인 과정에서 다뤄져야 한다. 이때 수집 단계가 거버넌스의 출발점이 된다.

3. 수집 기술과 아키텍처

데이터 수집 기술은 데이터 플랫폼의 운영을 위한 핵심 기술로, 빅데이터를 수집하는 기술뿐만 아니라 레거시 시스템에서 데이터를 수집하는 기술을 포함한다. 데이터 수집 기술은 데이터 통합을 위해 매우 중요하며, 수집된 데이터를 분석해 의미를 도출하는 기반이 된다.

데이터 수집은 활용에 필요한 데이터를 시스템의 내부 혹은 외부에서 주기적으로 필요한 형태로 모으는 행위로 정의할 수 있다. 서비스 모델을 결정하고 수집할 원천 데이터를 탐색·발견하는 일이다. 데이터를 발견하는 과정에서 수집의 난이도가 고려 대상이지만, 어디에 어떻게 데이터를 활용할지 활용 모델을 감안해 수집을 해야 한다.

[표 4-2-1] 데이터 수집 기술

구분	내용
데이터베이스 기반 데이터 수집	• 20여 종의 다양한 데이터베이스의 데이터 수집 지원 • SQL 기반의 데이터 추출 인터페이스 및 SQL 자동 생성 기능 지원 • 고속의 데이터 수집을 위한 병렬 데이터 추출 및 ROW API 형태의 초고속 추출 방식 적용 • 실시간 변경 데이터 수집 기술을 이용한 변경 데이터 실시간 추출 방식 지원

(계속)

구분	내용
센서 및 머신 데이터 수집	• 센서 및 머신 데이터는 다양한 형태의 인터페이스가 필요한 수집 대상으로, 기본 수집 에이전트를 제공할 뿐만 아니라 커스텀 에이전트를 개발할 수 있는 API 제공 • 다양한 형태의 데이터를 수집하는 에이전트 기본 제공 • CSV, TSV, Delimiter, JSON 등 다양한 형태의 전송 데이터 포맷 지원 • TCP 연결 방식의 데이터 수집 지원
파일 기반 데이터 수집	• 대규모 시스템의 경우 보통 데이터 소스가 분산돼 있어, 수집 에이전트의 원격 배포 및 관리 기능 지원이 필수요소 • 파일 기반 데이터 수집 기술은 에이전트 설정·배포·상태 모니터링·시작·중지 등 관리 기능 제공 • 파일 기반 데이터의 특성인 대량 데이터의 수집 성능을 극대화하기 위해 수집 서버의 분산처리 기능 지원 • 다양한 형태의 데이터를 수집하는 에이전트 기본 제공 • CSV, TSV, Delimiter, JSON 등 다양한 형태의 전송 데이터를 변환하지 않고 그대로 수집
SNS 수집 및 웹 크롤링	• 실시간 SNS(트위터, 페이스북 등) 데이터 수집 • 웹 크롤러 개발을 위한 수집 API 제공
실시간 변경 데이터 수집 (Change Data Capture, CDC)	• DBMS Redo Log로부터 실시간 변경 데이터 감지 • DML 및 DDL의 실시간 변경 인지 • 변경된 데이터에 대한 실시간 추출

가. 로그 데이터 수집

최적의 로그 데이터 수집을 위해 데이터 소스 인터페이스 에이전트를 제공한다. 푸시/풀 서비스를 통한 실시간 데이터 수집, 안전한 저장소 전달과 하둡 기반의 대용량 배치 통합 도구(예: TeraStream for Hadoop)를 통한 배치 데이터 수집 및 전달 처리를 한다. 이와 함께 개인정보에 대한 비식별화 조치로 RSC 모듈에 의한 암호화 작업을 수행해 처리해야 한다. 신규 시스템의 독립성을 위해 1대의 서버에 파티셔닝을 구축해 제공한다.

고속의 수집 상태에 대한 모니터링 기능을 제공하며, 수집 대상 서버, 수집 데이터 Row Count 값, CPU, 메모리 사용량 등 조회 모니터링 기능과 지정 디렉토리 내의 수집 데이터를 한정해 수집 기능을 활용할 수 있도록 제공하고 있다.

로그 파일 추적, 수집 및 자원 리소스 최소화를 위해 수집 에이전트에서 속성(Property) 세팅 시 해당 디렉터리 파일에 대한 자동 추적 기능 사용을 설정할 수 있다. 푸시/풀 기능을 제공해 대상 서버에 에이전트를 설치하는 경우, 병렬 에이전트 수를 조정해 최적의 환경을 구축해야 한다. 또한 안정적인 실시간 데이터 수

집을 위해 에이전트(예: TeraStream BASS BDI)를 제공하며, 배치 수집을 위한 하둡 기반의 대용량 배치 통합 도구를 제공함으로써 더욱 안정적인 데이터 수집 기능을 지원한다.

나. 정형 데이터 수집

기업에서 운영중인 내부 데이터 고객정보, 상품정보, 거래정보 등 정형 데이터 분석 활용을 위한 DBMS를 통해 필요에 따라 운영중인 데이터를 (준)실시간, 배치 형태로 분산 스토리지로 수집하는 기능을 제공한다. 변경 데이터 추출은 DBMS 리두 로그(Redo Log) 및 보관 로그(Archive log)에서 전송된 정보를 이용해 변경 데이터를 추출한다.

SQL 파서는 DBMS에서 전송되는 Hex 값을 아스키 형식으로 변환한다. 메타데이터 형식에 맞지 않는 데이터는 미스매치 파일을 생성한다. DBMS에서 변경된 데이터를 하둡 HDFS에 속성 값(Insert/Update/Delete)을 추가해 데이터를 분산 저장한다. 그리고 HIVE 0.14 버전 이후부터 변경 적재하며, 변경 적재는 하둡 특성상 준실시간 방식으로 지원한다. 수집 대상 데이터의 성격에 따라 CRUD 내역을 추출해 목적 시스템에 반영하거나, 수집 순서에 따른 파티셔닝 형태의 저장 기능을 제공한다. 원천 시스템에서 데이터 수집 상태는 추적(모니터링 포함)이 가능해야 하며, 오류 및 필요에 의한 재가동시에도 직전 수집 이후부터 연속적인 수집 기능을 제공해야 한다.

4. 국내외 데이터 수집 솔루션 동향

과거에는 데이터의 배치 정보 분석, 히스토리 분석에 그쳤으나 최근에는 HW와 SW 기술 발전으로 대량의 데이터를 분산 저장하고 데이터 생성과 동시에 분석할 수 있는 실시간 분석 플랫폼이 매우 중요해지고 있다.

실시간 처리 기술의 핵심은 쌓아 놓고 처리하는 방식이 아니라 데이터가 생성되면 곧바로 처리하는 방식이다. 사람들의 행위나 기타 작용에 의해 생성된 자료는 일종의 이벤트 데이터이므로 이벤트 기반 실시간 처리 기술이라고 한다. 이 방식은 또 실시간으로 생성되는 데이터들이 시냇물처럼 끊임없이 흘러나온다고

해 스트림 처리 기술이라고도 부른다.

사물인터넷 빅데이터 플랫폼은 데이터 수집을 위한 표준 에이전트 방식과 병렬 수집 및 병렬 저장 처리를 위한 구조를 제공하고, 고속의 인덱싱 후 쿼리에 의한 분석 서비스를 제공한다. 다양한 데이터 수집을 위해서는 정형 RDB 이외에 센서 데이터, VOC 데이터, 웹 크롤링 데이터 등 수집 대상 데이터의 다변화가 필요하다.

또한 광통신망 통합 게이트웨이의 패킷 분석을 위해 10G 패킷 전체의 실시간 저장 및 분석 가능한 분산 인메모리 처리 플랫폼에 대한 개발 요구가 있다. SQL 기반의 사용성 및 인터페이스, 실시간 차트 구현 및 개발이 가능한 UI 환경 기반 기술 연구 또한 활발히 전개되고 있다. 분산 인메모리 저장 및 인덱싱 기술, 하이브리드 저장 플랫폼 기반 기술 연구가 다각도로 검토되고 있다.

인메모리 기반의 빠른 데이터 처리 및 모니터링과 히스토리 데이터 분석을 위한 하둡 기반 분석의 두 가지 기능을 제공할 수 있는 하이브리드 기능이 차별화된 기술로 각광받고 있다. 또한 GUI 기반의 편리한 데이터 변환 구성시 별도 코딩 없는 피보팅, 그래프 변환, GUI 환경의 데이터 추출 및 변환, 고객 임원진을 위한 대시보드 구성 등도 편리한 확장성 및 편리성 제공 측면에서 사용자 관점에서 선호하는 기술이다.

업무 증가에 따른 주변 시스템과의 연계와 확장성 있는 하드웨어 인프라를 구성함으로써 다양한 데이터를 쉽게 수집하고, 고속 저장, 실시간 분석하는 사물인터넷 빅데이터 플랫폼으로 구현하는 기술이 빠르게 확산되고 있다.

이러한 기술의 평가 기준은 크게 여덟 가지로 나눌 수 있다. 1) 얼마나 다양한 데이터를 수용할 수 있는지 2) 별도의 코딩 없이 편리하게 수집 가능한지 3) 병렬 수집 및 고속 수집이 가능한지 4) 다양한 데이터를 손쉽게 수집할 수 있는지 5) 얼마나 많은 데이터를 빠르게 인덱싱할 수 있는지 6) 데이터 유실 없이 빠르게 저장할 수 있는지 7) 고객 관점의 서비스 가치를 제공하고 있는지 8) 패키징된 파트너 솔루션이 다양한지 등이 그것이다.

●○
업무 증가에 따른 주변 시스템과의 연계와 확장성 있는 하드웨어 인프라를 구성함으로써 다양한 데이터를 쉽게 수집하고, 고속 저장, 실시간 분석하는 IOT 빅데이터 플랫폼으로 구현하는 기술이 빠르게 확산되고 있다.

5. 향후 추이

데이터 통합 기술은 데이터 플랫폼의 중요 기술 중 하나로 분석에도 사용된다. 특히 데이터 분석 시 발생하는 부하를 줄이기 위해 분석이 쉬운 형태로 데이터를 변환·통합하고, 딥러닝·머신러닝을 위한 학습 데이터 생성 관련 기술력을 확보하고 있어야 한다. 이 기술은 주로 대용량 데이터를 일괄 변환·통합에 사용되지만, 실시간 데이터의 통합에도 동일한 기술을 사용하기 때문에 기술 적용 범위가 넓은 편이다. 또 데이터 통합 시 성능이 매우 중요해 구현을 위한 난이도가 매우 높아 데이터 플랫폼 구현 시 진입 장벽이 되는 기술이기도 하다.

대용량 데이터의 일괄 변환과 통합을 위한 데이터 정렬 및 조인, 변환 규칙에 따른 변환, 빅데이터 플랫폼 연계가 기술의 중요한 기반요소가 된다. 아울러 데이터에 대한 거버넌스 관점에서 기업 내외부에서 수집되는 다양한 데이터를 분류하고(Categorize), 데이터별 오너십과 생성되는 위치, 경로를 정의(Metadata Management)해, 분석에 활용하기 위해 참조하고 있는 데이터를 구성원 누구나 쉽게 찾을 수 있게(Discovery) 한다. 통합 싱글뷰 제공으로 빅데이터 분석 지능을 제공하며, 데이터 생애주기 관점에서 생성부터 활용, 폐기에 이르는 일련의 데이터 처리 흐름 과정을 확인할 수 있도록 통합 데이터 기반 정보 활용의 핵심 아키텍처를 구현해 빅데이터 기반 경영 의사결정을 지원하는 데이터 관리체계 구축으로 발전하는 추세다.

제3장

DBMS 솔루션

18세기 중반부터 19세기 초반까지 영국에서 시작된 산업혁명은 증기기관이라는 파격적인 동력과 에너지, 기술이 결합해 생산력의 비약적인 증대를 가져 왔다. IT에서 최근 가장 큰 화두는 제4차산업혁명이다. 인공지능, 빅데이터, 클라우드, 3D 프린팅, 사물인터넷(IoT) 등 제4차산업혁명으로 거론되는 기술 대부분이 소프트웨어와 관련돼 있으며, 그 중심에 증기기관 역할을 하는 DBMS가 자리잡고 있다.

1. 개요

데이터베이스 관리시스템(DBMS)은 다수의 컴퓨터 사용자들이 컴퓨터에 수록한 수많은 자료를 쉽고 빠르게 추가·수정·삭제할 수 있도록 해주는 소프트웨어다. DBMS는 자료를 축적하고, 축적된 자료를 정의하고 구조화한다. 이렇게 축적되고 구조화된 자료에 대해 조회·추가·변경·정보보안 등의 기능을 제공한다. 보통 DBMS는 애플리케이션 시스템과 데이터를 저장한 파일의 중간 단계에 위치한다. 데이터베이스를 관리하면서 애플리케이션으로부터의 액세스 및 서비스 요구에 대해 효율적이고 적절하게 응답한다. 데이터베이스가 관리하는 데이터에 대해 요청된 서비스를 제공하고 액세스시키는 역할을 담당한다.

DBMS는 계층형·네트워크형·관계형으로 변화돼 왔고, 현재는 관계형 데이터베이스가 널리 사용되고 있다. 관계형 데이터베이스는 키(key)와 값(value)들을 테이블화하고 이런 테이블을 일정한 규칙에 따라 간단하게 관계를 맺어 놓은 것이다. 데이터를 구조화, 시스템화해 가치가 높은 정보를 활용할 목적으로 만든 데이터베이스이다. 관리 기능을 탑재한 데이터베이스 매니지먼트 시스템이 관계형 데이터베이스 관리시스템(Relational Database Management System, RDBMS)이다. RDBMS의 주된 기능으로는 참조 정합성 기능, 배타 제어 기능, 백업 복구 기능, 보안 관리 기능 등이 있다.

NoSQL DBMS는 관계형 구조를 갖지 않은 데이터를 관리하는데 사용된다. 대부분의 빅데이터는 정형화된 구조를 갖지 않기 때문에 관계형DBMS로는 관리할 수 없고, NoSQL DBMS를 적용해야 한다. 조인 연산이 불가능하며, 각 데이터가 독립적으로 설계돼 데이터를 여러 서버에 분산해 작업할 수 있다. 여러 서버에 분산해 처리하는 데이터의 양이 감소하고 대량의 데이터 입출력이 있어도 처리가 간단하다. 많은 장점에도 NoSQL은 개별 애플리케이션을 지원하는 복수의 데이터베이스를 이용해 운영의 복잡도가 증가하며, SQL이 포함한 연관성을 포기해 NoSQL 스토어는 분석 한계 등의 제약이 있다. 기업의 거래 처리를 위한 구조적 데이터는 관계형 DBMS가 적합하다. 센서데이터, 로그데이터 등 구조가 정해져 있지 않으면서 대용량으로 발생하는 데이터는 NoSQL DBMS가 적합하다.

DBMS 기술은 전통적인 RDBMS에서 빅데이터, 클라우드, 모바일, 사물인터넷 등 우리가 직접적으로 느끼지 못하는 곳까지 확산되면서 영향력이 커지고 있다. 이러한 IT 환경의 변화는 DBMS에 더 많은 데이터의 가치를 요구하고 있으며, 네트워크에 연결돼 쏟아지는 데이터 양이 증가함에 따라 DBMS의 중요성은 날로 커지고 있다.

2. DBMS 동향

오라클은 클라우드 사업에 모든 역량을 집중하고 있다. 2013년 7월, 오라클 12c(Oracle 12c)를 클라우드 시장을 타깃으로 발표한 이후, 2017년 4월에는 오라클 엑사데이타 클라우드 머신(Oracle Exadata Cloud Machine)을 국내에 출시했다. 이 솔루션은 자사 데이터베이스를 온프레미스(On-Premise) 환경에 두고자 하는 기업이나 비즈니스 목적이나 규제 등 제약으로 인해 퍼블릭 클라우드로 옮기기 어려운 고객들을 대상으로 한다. 오라클 퍼블릭 클라우드와 동일한 비즈니스 모델, 동일한 솔루션, 오라클 전문가의 인프라 관리 서비스를 제공함으로써 기업 데

* 필자: 이동석(티맥스소프트 상무)

이터센터 내에서 오라클 클라우드를 동일하게 경험할 수 있는 점이 강점이다.

IBM은 2016년 7월 온프레미스 애플리케이션과 클라우드를 더 쉽게 연결해 하이브리드 데이터 아키텍처를 구축할 수 있는 IBM DB2 V11.1을 발표했다. 인메모리 기술을 대량의 병렬처리 아키텍처 전반에 걸쳐 사용할 수 있도록 지원하고 응답시간도 크게 개선했다. 기존 버전인 9.7에서 11.1로 바로 업그레이드 가능하다.

마이크로소프트는 2017년에 SQL 서버(SQL Server) 새 버전을 윈도우와 리눅스용으로 출시할 예정이다. 윈도우 버전은 2016년 많은 부분이 개선된 바 있다. 새 버전에서는 그동안 윈도우 서버에서만 이용할 수 있던 SQL 서버를 리눅스에서도 사용할 수 있게 돼 신규 고객과 기존 고객 모두의 관심을 받게 될 것으로 보인다. SQL 서버를 윈도우, 리눅스에서 모두 사용할 수 있게 돼 리눅스 기반 클라우드에서도 일관된 데이터 플랫폼 서비스를 제공할 수 있게 된다. 또한 엔터프라이즈 수준의 솔루션으로 인식이 전환되면서 SQL 서버의 시장 확대에 좋은 영향을 미칠 것으로 예상된다.

SAP의 HANA 2는 위치에 구애받지 않고 기업용 모델링, 데이터 통합, 향상된 데이터 품질 및 계층형 스토리지 등을 향상해 고객사가 다양한 위치에 저장된 데이터를 활용할 수 있도록 한다. HANA 2에서는 데이터베이스(DB) 관리, 가용성과 보안 등이 업데이트됐다. 또한 SAP는 구글의 기업용 파트너 앱 마켓플레이스인 '구글 클라우드 런처 마켓플레이스(Google Cloud Launcher Martetplace)'를 통해 SAP 하나 익스프레스 에디션(HANA express edition)을 제공하며, 향후엔 서비스형 플랫폼(PaaS) 'SAP 클라우드 플랫폼'을 지원할 예정이다.

티맥스데이터는 2015년도 3월에 Tibero 6를 출시한 이후, 대용량 DB와 기업의 핵심 업무 분야에서의 점유율을 지속적으로 늘려 나가고 있다. Tibero 6는 타 DBMS와의 호환성이 뛰어날 뿐 아니라, DBMS 전환 시 영향 정도를 사전 예측하고 자동 전환하는 T-Up 솔루션을 통해 편의성과 안정성을 강화했다. 티맥스데이터는 2016년 10월 TmaxCloud Day에서 자사의 IaaS와 PaaS 솔루션을 발표한 이후 클라우드 시장으로 발빠르게 움직이고 있다. KT, AWS 등 클라우드 환경에서 Tibero 운영을 지원할 뿐 아니라, 클라우드 환경에서 액티브 클러스터링 기능을 제공하는 Tibero Zeta Edition을 출시해 고성능과 고가용성을 요구하는 기업의 핵심 업무를 클라우드 환경에서도 운영 가능하도록 했다.

선재소프트는 2013년 인메모리 기반 DBMS인 SUNDB를 출시했다. 이 제품은 고성능을 내세워 시장 감시나 증권사의 주문 업무 중심으로 사례를 확대하고 있다. 또한 초저지연(Near-zero latency) 데이터 처리를 가능하게 하는 인메모리 DBM(DataBase Manager) 제품인 Goldilocks를 출시했다. 큐브리드는 국산 오픈소스 DBMS로 포지셔닝되면서 정부통합전산센터 G-Cloud를 비롯한 클라우드 환경에서 상용 DBMS나 다른 오픈소스들과 경쟁하고 있다. 2016년 출시한 Cubrid 10.0은 MVCC(Multi-Version Concurrency Control, 다중 버전 병행 제어) 기반 Snapshot Isolation 기능을 제공하고, 성능 및 확장성 향상, SQL 및 Function 기능을 추가했다.

PostgreSQL은 객체관계형 데이터베이스 관리시스템(ORDBMS)이며, 버클리 소스를 기반으로 확장된 오픈소스다. 표준 SQL 기능을 대부분 지원하며 복합 쿼리, 참조키, 트리거, 업데이트 가능한 뷰, 트랜잭션, MVCC 등의 기능도 제공한다. 최근 AWS의 데이터베이스 엔진인 아마존 오로라(Aurora)에 PostgreSQL 호환 기능이 추가됐다.

마리아DB(MariaDB)는 MySQL과의 90% 이상 호환성을 강조하며 등장한 오픈소스 기반의 DBMS다. 2017년 상반기 중 마리아DB 10.2의 정식 버전이 출시될 예정이다. 마리아DB 10.2에 추가·개선된 기능은 Window Function, Syntax, Information Schema, EXPLAIN, 호환성 Variables, Script 등이다. 빅데이터 분석 시장 공략을 위해 마리아DB 컬럼스토어 1.0(MariaDB ColumnStore 1.0)이 2016년 12월에 발표됐다. 마리아DB 컬럼스토어는 OLTP와 분석을 지원하는 단일 SQL 인터페이스로 트랜잭션 워크로드와 대용량 병렬 분석 워크로드를 동시에 처리하는 스토리지 엔진이다. 이를 통해 마리아DB는 관계형 데이터베이스에서 트랜잭션과 분석 작업을 동시에 처리할 수 있게 됐다.

국내외 DBMS 플레이어들의 가장 큰 관심사항은 클라우드라 할 수 있다. IT 환경은 빠르게 클라우드 환경으로 전환되고 있으며, 전환 과정에서 클라우드 환경을 지원하는 DBMS도 덩달아 성장하고 있다.

3. 국내외 DBMS 기업 동향

최근 국내외 DBMS 플레이어들의 가장 큰 관심사항은 클라우드라 할 수 있다. IT 환경은 빠르게 클라우드 환경으로 전환되고 있으며, 전환 과정에서 클라우드 환경을 잘 지원하는 DBMS 역시 덩달아 성장하고 있다. 이에 따라 많은 DBMS 플

레이어들이 클라우드 환경에 최적화된 DBMS를 시장에 출시하면서 마케팅 활동을 강화하고 있다.

오라클의 기업 혁신 프로그램인 스타트업 클라우드 액셀러레이터 프로그램은 비즈니스 개발과 혁신에 특화돼 있으며, 스타트업 에코시스템을 구축하고 스타트업 기업들이 가장 필요로 하는 실질적인 지원과 글로벌 자원을 제공하고 있다. 심사를 통해 선별된 기업들은 비즈니스 파트너십을 통한 세일즈 기회, 고객 후원(Advocacy), B2B 영업과 마케팅, HR 관련 교육, 글로벌 커뮤니티를 통한 사례 공유 등 차별화된 혜택을 받을 수 있다. 또한 클라우드 시장 진입을 위해 가상화 환경을 그대로 쉽고 빠르게 클라우드 환경으로 옮겨 운영할 수 있는 클라우드 서비스로 IaaS 클라우드 시장을 공략하고 있다. 온프레미스에서 운영되고 있는 네트워크 환경을 그대로 클라우드 상에서 구현해 기업의 다양한 업무 용도로 확장할 수도 있다.

IBM과 레드햇은 프라이빗 클라우드에 레드햇 오픈스택 플랫폼 및 솔루션을 결합하는 파트너십을 체결했다. 이를 통해 IBM은 레드햇 인증 클라우드 및 서비스 제공업체(Red Hat Certified Cloud and Service Provider)로 지정됐다. 새로운 서비스가 공개되면 레드햇 오픈스택 플랫폼(Red Hat OpenStack Platform) 및 레드햇 세프 스토리지(Red Hat Ceph Storage)를 IBM 프라이빗 클라우드에서 사용할 수 있게 된다. 또한 레드햇 클라우드 액세스(Red Hat Cloud Access)를 2017년 2분기까지 IBM 클라우드(IBM Cloud)에서 사용 가능하며, 레드햇 고객들이 미사용 레드햇 엔터프라이즈 리눅스(Red Hat Enterprise Linux) 서브스크립션을 자사의 데이터센터에서 전 세계 IBM 클라우드 데이터센터(IBM Cloud Data Centers)에 있는 퍼블릭 및 가상화 클라우드 환경으로 전환할 수 있도록 지원한다.

마이크로소프트는 윈도우 서버와 SQL 서버에 새로운 라이선스 옵션을 추가했다. 별도 비용을 내면 기존 기술지원 기간에 6년을 추가해 총 16년간 보안 취약점 패치를 제공한다. '프리미엄 보증(Premium Assurance)'이라는 이 서비스는 윈도우 서버 2008과 SQL 서버 2008을 포함한 이후 버전에 적용된다. 프리미엄 보증 서비스가 적용되는 가장 오래된 제품은 윈도우 서버 2008과 2008 R2, SQL 서버 2008과 2008 R2 등이다. 또한 개발자들이 클라우드에 추가된 다양한 기술과 서비스를 제대로 이용할 수 있게 하기 위해 개발자 재교육을 위한 강좌 진행과 행사를 개최하고 있다.

SAP는 최근 미국 샌프란시스코에서 개최된 '구글 클라우드 넥스트 17(Google Cloud Next 17)' 컨퍼런스에서 구글과 클라우드 파트너십을 발표했다. 이번 파트너십으로 구글은 전 세계에서 SAP의 세 번째 퍼블릭 클라우드 파트너가 됐다. 이로써 기업 고객은 미션 크리티컬 애플리케이션과 분석 솔루션을 '구글 클라우드 플랫폼(Google Cloud Platform)' 기반의 SAP HANA에서 구동할 수 있게 됐다. SAP는 구글의 기업용 파트너 애플리케이션 마켓플레이스인 구글 클라우드 런처 마켓플레이스(Google Cloud Launcher Marketplace)를 통해 SAP HANA 익스프레스 에디션(SAP HANA express edition)을 제공하고 개발자 환경을 지원한다. 지난해 AWS와 마이크로소프트와도 퍼블릭 클라우드 파트너십을 맺고 고객들에게 다양한 선택의 폭을 제공하는 SAP는 구글과의 협력을 추진함으로써 클라우드 시장을 안정적으로 확장하는 기반을 확보하게 됐다.

티맥스데이터 역시 여러 CSP(Cloud Service Provider)들과의 제휴를 통해 온프레미스뿐만 아니라 퍼블릭 클라우드 비즈니스를 확대하고 있다. 고객사의 프라이빗 클라우드 구축 프로젝트를 통해 DBMS를 공급하려는 노력도 진행중이다. 또한 온프레미스 환경에서 마이그레이션을 두려워하는 고객들을 위해 이기종 DBMS를 하나의 도구(Tool)로 관리할 수 있는 DBMS 매니저 기술을 제공하는 등 DBMS 영역을 확장하고 있다. 2016년에는 중국, 러시아, 태국, 브라질, 말레이시아 등 해외 매출도 지속적으로 증가하고 있다.

●○ DBMS 업체들간 끊임없는 기술과 가격 경쟁으로 인해 OLTP 중심의 RDBMS 시장에서는 DBMS 다변화로 인한 이기종 DBMS의 운영 또는 다른 DBMS나 클라우드 환경으로의 마이그레이션이 더 활발해질 것으로 예상된다.

알티베이스는 기존보다 성능을 높인 Altibase 6.5를 중국 통신사와 일본 기업에 공급하는 등 해외시장에서 성과를 내고 있다. 선재소프트는 2016년 시스템 중단 없이 서버를 추가하는 SUNDB 3.0 Maxscale을 개발했다. 이 제품은 빅데이터 시대에 시스템 중단 없이 데이터베이스 시스템을 확장하는 스케일아웃 방식이다. 선재소프트도 다른 국산 소프트웨어 업체와 마찬가지로 해외시장 진출을 시도하고 있는데 중국 통신사에 SUNDB를 공급하는 등의 성과를 내기도 했다.

4. 향후 전망

18세기 중반부터 19세기 초반까지 영국에서 시작된 기술 혁신과 이로 인해 일어난 사회, 경제 등의 큰 변혁을 '산업혁명'이라고 한다. 산업혁명은 증기기관이라는,

당시에는 파격적인 새로운 동력과 에너지, 기술이 결합돼 생산력의 비약적인 증대를 가져 왔다. 공장을 세우고 생산력이 크게 확대되면서 인간의 삶은 더욱 풍요로워졌으며, 이는 현대 자본주의 문명과 과학기술의 기초를 다질 수 있게 했다.

최근 IT에서 가장 큰 화두는 제4차산업혁명으로, 각종 매체와 세미나, 컨퍼런스에서 가장 많이 다루는 주제이기도 하다. 제4차산업혁명으로 거론되는 기술은 인공지능, 빅데이터, 클라우드, 3D 프린팅, 사물인터넷(IoT) 등 대부분이 소프트웨어와 관련된다. 이들 기술의 핵심에 DBMS가 자리잡고 있다.

먼저 클라우드 기술을 살펴보면, 클라우드 초기에는 OS 가상화를 통해 하드웨어 자원 관리를 자동화·효율화시키는 데 중점을 뒀으나, 최근 엔터프라이즈의 핵심 업무들이 클라우드 환경에서 운영되기 시작하면서 클라우드에서의 RDBMS의 중요성이 날로 강조되고 있다. 이에 따라 기존 온프레미스에서의 DBMS 성능과 가용성을 보장하면서 클라우드 환경에서 운영 가능한 기술들이 등장하고 있다. 앞으로의 DBMS 시장은 온프레미스 환경을 클라우드 환경으로 쉽게 마이그레이션 할 수 있고, 클라우드 환경에서 관리가 수월한 DBMS가 시장을 주도하게 될 것이다.

빅데이터 기술은 클라우드 기술과 인공지능이 융합된 형태의 기술로 발전하고 있으며, 이 과정에서 대량의 데이터를 낮은 비용으로 처리할 수 있는 스토리지 서버 기술이 매우 중요해질 것이다. 또한 빅데이터 기술이 기존 데이터 분석의 영역을 넘어 CRM(Customer Relationship Management), 보안, 모니터링, 운영관리 등의 서비스를 제공할 수 있는 기술로 발전하면서 정형 데이터와 비정형 데이터를 함께 처리할 수 있는 DBMS가 중요해질 것으로 예상된다.

기업에서는 끊임없이 IT 운영비용 절감을 요구하고 있고, DBMS 업체들간 끊임없는 기술과 가격경쟁으로 인해 OLTP 중심의 RDBMS 시장에도 변화가 예고된다. DBMS 다변화로 인한 이기종 DBMS의 운영 또는 다른 DBMS나 클라우드 환경으로의 마이그레이션이 더 활발해질 것으로 예상된다. DB 어플라이언스의 경우 데이터 분석에 대한 고객의 니즈가 증가함에 따라 대량 데이터의 저장, 처리 및 분석 고도화, DB 통합 등 고객 운영 상황에 특화된 전략에 따라 수요처가 넓어질 것으로 예상되며, 국내 DB 어플라이언스와 글로벌 DB 어플라이언스간 경쟁도 더 심화될 것으로 보인다.

제4장

데이터 레이크 솔루션

데이터 레이크는 빅데이터를 활용할 수 있도록 정제되지 않은 원천 데이터들을 한 곳에 통합 저장해 통찰력을 도출해 내기 위해 생성된 개념이다. 도입과 활용 측면에서 아직 활성화 되지 못한 빅데이터의 발전을 견인할 데이터 레이크 솔루션의 현주소를 알아본다.

1. 개요

데이터 레이크(Data Lake)란 오픈소스 기반의 BI(Business Intelligence) 기업 펜타호(Pentaho)의 창립자이자 CTO인 제임스 딕슨(James Dixon)이 2014년 처음으로 사용한 용어로, 정제되지 않은 다양한 형태의 데이터 저장소를 의미한다. 기존에는 BI 기술을 적용하는 정형 데이터 중심의 데이터 웨어하우스에 비정형 데이터 소스 기반의 빅데이터 기술이 적용됨으로써, 실질적인 빅데이터 활용에 제약이 많았다. 따라서 이런 한계를 해결할 수 있는 방안으로 제시된 것이 데이터 레이크라고 볼 수 있다.

빅데이터에 대한 관심이 전 사회적으로 증가하는 가운데 빅데이터 솔루션은 다양한 사용자의 접근이 용이하도록 성숙해야 하며, 단순히 고급 프로그램의 전용물 이상의 도구가 돼야 할 것이다. 빅데이터 기술의 최신 경향인 데이터 레이크는 당연히 폭넓은 사용자 계층을 포용할 수 있어야 한다.

* 필자: 공상휘(티맥스소프트 상무)

2. 데이터 레이크 솔루션의 주요 기능

첫 번째, 데이터 레이크 정의에 따라 가공되지 않은 상태의 다양한 비정형 포맷의 데이터를 저장하고 가공할 수 있어야 한다. 전통적인 관계형 데이터베이스의 정형 데이터뿐만 아니라 CSV, 단순 로그 텍스트, 하둡 기반의 RCFile(Record Columnar File), ORC(Optimized Row Columnar) 등의 데이터 타입은 물론, 저장공간의 효율적인 사용을 위해 GZip(GNU zip), LZO(Lempel-Ziv-Oberhumer) 등의 압축 방식도 지원해야 한다.

두 번째, 전통적인 기술로 처리하기 어려운 규모의 대용량 데이터를 저장·관리할 수 있어야 하며, 비용을 최소화할 수 있도록 클라우드 기술 기반의 스토리지를 제공할 필요도 있다.

세 번째, 데이터의 입력 및 활용을 위한 표준화된 인터페이스와 활용 대상 데이터의 메타 정보를 제공해야 사용자의 데이터 활용도가 올라갈 수 있다.

네 번째, 데이터를 찾고 가공할 수 있는 도구가 필요하다. 하둡의 경우, 파이썬(Python)이나 피그(Pig)와 같은 스크립트 언어를 이용해 단순한 커맨드 방식의 도구를 제공하는 것도 하나의 예가 될 수 있다. 또한 SQL은 강력한 데이터 가공 도구이자 분석 도구다. 이러한 분석 언어는 최근 웹 기반의 GUI 도구를 통해 제공하는 경우도 다수 존재한다.

●○
데이터 레이크 솔루션은 기존 BI 기술을 적용하는 정형 데이터 중심의 DW가 비정형 데이터 소스를 기반으로 빅데이터 기술을 적용함으로써 빅데이터 활용에 제약이 많다는 한계를 해결할 수 있는 방안으로 제시됐다.

다섯 번째, 사물인터넷(IoT) 데이터가 중요한 데이터 원천이므로 기존의 정적 데이터의 배치 처리 이외에도 동적인 스트리밍 데이터의 동적 분석을 동시 요청하는 경우가 늘고 있다. 일반적으로 CQL(Continuous Query Language)을 통한 스트리밍 데이터의 실시간 분석 기능이 대표적인 예가 될 수 있다.

여섯 번째, 데이터 처리 오케스트레이션(orchestration) 기능이 필요하다. 특정 이벤트에 의해 기동돼 일련의 예정된 처리 플로우를 수행함으로써 데이터의 획득, 가공, 변환, 생성 등의 처리를 효율적으로 수행할 수 있다.

일곱 번째, 데이터 레이크에 저장된 데이터 접근과 변경에 대한 사용자 관점의 보안 정책을 적용할 수 있어야 한다. 데이터 접근을 통제하는 인증(authentication)과 데이터 활용 권한을 통제하는 인가(authorization) 기능, 각종 데이터 관리 및 활용에 대한 감사(audit) 기능이 필요하다.

3. 국내외 주요 제품의 특징 및 업체 동향

데이터 레이크 솔루션을 제공하는 해외 기업은 크게 전통적인 솔루션 벤더과 클라우드 서비스 사업자로 나눌 수 있다.

가. 해외 솔루션 벤더

기존 온프레미스(on-premise) 솔루션 벤더로 가장 두드러진 기업은 테라데이터(TeraData)와 델이엠씨(DellEMC)이다.

테라데이터는 최근 아파치 하둡(Apache Hadoop), 아파치 스파크(Apache Spark), 아파치 나이파이(Apache NiFi)와 같은 최신 오픈소스를 기반으로 하는 데이터 레이크(Data Lake) 관리 소프트웨어 플랫폼인 카일로(Kylo)를 제공한다. 카일로는 테라데이터의 후원 하에 아파치 2.0 라이선스 기반으로 제공되는 오픈소스 프로젝트를 통해 만들어진 제품이다. 다양한 형태의 머신 로그와 이메일, 웹페이지 등의 비정형 데이터를 배치 및 스트리밍 방식으로 획득해 분석한다. 고성능의 파이프라인 방식의 데이터 관리 기능을 제공하며, 프로그램을 모르는 현업 담당자도 사용할 수 있도록 셀프 서비스를 위한 직관적인 사용자 인터페이스를 제공하는 것이 큰 장점이다.

델이엠씨의 경우, EMC는 피보탈(Pivotal), VM웨어(VMware) 등 EMC 패더레이션 기술을 하나로 묶어 빅데이터 관리와 활용을 위한 어플라이언스 형태의 '페더레이션 비즈니스 데이터 레이크(FBDL, Federation Business Data Lake)' 솔루션을 제공한다. 대규모의 정형·비정형 데이터 저장과 하둡, 인메모리, NoSQL, 스케일아웃 MPP 등 다양한 데이터의 관리와 분석, 그리고 실시간으로 분석을 가능하게 한다. 특히 FBDL은 자회사인 VM웨어, 피보탈의 애플리케이션과 스토리지 등 하드웨어가 결합된 어플라이언스 제품으로 데이터 레이크 구축 시간 단축이라는 장점을 가지고 있다. 분석 레이어는 VM웨어 가상화를 기반으로 SQL온하둡(SQL-on-Hadoop) 엔진인 호크(HAWQ)를 포함한 피보탈HD(PivotalHD)로 구성됐다. SAS, 태블로(Tableau)의 분석 툴과 연계가 가능하며 클라우데라(Cloudera), 호튼웍스(Hortonworks) 등의 하둡 배포판과도 호환된다. 또 'EMC V블록(EMC VBlock)'과 'EMC 아이실론(EMC ISILON)' 등 다양한 스토리지 솔루션으로 정형·비정형 데이터를 효율적으로 저장해 데이터 증가 시에도 쉽게 시스템을 확장할 수 있다.

나. 해외 클라우드 서비스 사업자

데이터 레이크 솔루션을 제공하는 대표적인 클라우드 서비스 사업자(CSP, Cloud Service Provider)로는 마이크로소프트 애져(Azure)와 아마존 AWS가 있다.

마이크로소프트 애져는 페타바이트급의 대용량, 다양한 형태의 데이터를 저장할 수 있으며 U-SQL, R, 파이썬, 닷넷(.Net) 등의 프로그램 언어와 플랫폼 기반으로 데이터 처리와 분석을 수행한다. 또한 배치 및 스트리밍, 대화식 분석 기능을 제공해 데이터 수집 및 저장에 따른 복잡성을 최소화하려고 했다.

특히 비주얼 스튜디오(Visual studio)와 이클립스(Eclipse) 등과도 연동돼 레거시 개발자들이 쉽게 데이터 레이크를 구축 및 활용할 수 있게 해준다. 또한 U-SQL과 스파크(Spark), 하이브(Hive), 스톰(Storm)의 처리 작업을 GUI를 통해 제공함으로써 사용자 접근성을 높였다.

아마존의 AWS의 경우, 클라우드 플랫폼에서 제공하는 AWS Cloud Formation 템플릿을 이용해 기본적인 데이터 레이크의 서비스 구성요소를 생성하고, 다양한 포맷의 데이터를 저장한 후 Amazon Dynamo DB를 통해 메타데이터를 관리할 수 있도록 한다. 검색 기반의 엘라스틱 서치(Elastic search)를 통해 분석 기능을 제공한다. 또한 AWS에서 제공하는 암호화, 인증 등의 보안 기능을 제공하고 있다.

데이터 레이크를 구축하기 위한 솔루션들은 하둡 기반의 오픈소스를 대부분 이용하고 있으며, 각 벤더의 특징에 따라 접근 방법과 방향이 상이하다.

테라데이터는 기존 관계형 데이터베이스의 강자로 정형 데이터의 DW와 비정형 데이터를 통합 분석할 수 있는 구조를 갖고 있으며, 이는 국내 유일한 RDBMS 제품을 만드는 티맥스소프트의 AnyMiner도 유사하다. 델이엠씨는 가상화 기술과 스토리지 기술을 접목해 데이터 레이크의 유일한 어플라이언스 제품을 시장에 공급하고 있다는 점에서 다른 경쟁사와 큰 차별성을 보인다. 클라우드 서비스 사업자의 경우는 기존 클라우드 플랫폼에서 제공하는 저렴한 컴퓨팅 파워와 스토리지 자원을 기반으로 오픈소스의 분석 기술을 접목해 사용자의 접근도를 높이는 방향으로 솔루션을 제공한다.

다. 국내 솔루션 벤더

데이터 레이크 제품을 제공하는 국내 기업으로는 티맥스데이터가 있다. 티맥스의 AnyMiner는 다양한 포맷의 정형·비정형 데이터를 대상으로 클라우드 스토리지 기반의 데이터 레이크를 구축할 수 있다. 특히 머신 및 소프트웨어에서 만들어지는 로그와 IoT 기반의 스트리밍 데이터, 관계형 데이터베이스의 정형 데이터까지 배치 및 스트리밍 분석을 병행할 수 있는 기능을 제공한다. 비정형 데이터의 분석을 위해 역색인(Inverted index) 기반의 검색기술을 이용한 고성능 분산 병렬 방식을 사용한다. 다양한 프로그램 언어를 통한 분석 기능을 제공하지 않는 점은 단점이지만, 프로그램 언어에 익숙하지 않은 현업 담당자가 직접 셀프 서비스를 구현할 수 있는 분석 UI를 제공한다는 점은 강점으로 꼽힌다.

4. 데이터 레이크의 미래

데이터 레이크는 빅데이터 산업을 선도해온 미국 및 일부 국가에서 빅데이터를 활용할 수 있도록 정제되지 않은 원천 데이터를 한 곳에 통합 저장함으로써 신선한 아이디어와 통찰력을 도출해내는 기반을 구축하기 위해 만들어진 개념이자 기술이다. 2014년 이후 수 년이 지났지만 여전히 그 효용에 대해서는 우려와 기대가 공존하고 있다.

우리나라의 경우, 기존에 빅데이터 기술이 통신사나 일부 대기업과 연구기관에서 상당한 하드웨어 인프라와 데이터과학자 그룹을 구성해 빅데이터 분석 성공사례를 만들었다. 대부분의 기업 고객은 빅데이터를 반드시 활용해야 할 새로운 동력으로 인식하고 있지만, 도입과 활용 측면에서는 여전히 연구·탐색 수준이라고 볼 수 있다.

빅데이터가 아직 보편화되지 못한 이유로 세 가지를 들 수 있다. 첫째는 프로그램을 모르는 업무 분석자도 쉽게 데이터를 저장·처리·분석할 수 있는 사용자 접근성이 아직 부족하다는 점이다. 두 번째, 실제 대용량의 빅데이터를 저장하고 처리할 컴퓨팅 자원에 대한 비용 부담이 크다는 점이다. 특히 데이터 레이크를 구축하면 더 다양한 형태의 데이터를 처리할 수 있으며, 더 많은 양의 데이터를 관리해야 하는데, 이에 대한 기술적 해결이 반드시 수반돼야 한다. 세 번째, 데이터 품질과 표준화의 한계이다. 데이터 정제(cleansing)와 표준화 작업에는 많은 노력이 들어가며 데이터 거버넌스의 정착이 필수다.

이에 데이터 레이크의 미래는 이러한 한계를 극복하기 위한 기술 발전이 관

●○ 현재는 퍼블릭 클라우드 서비스에서만 데이터 레이크를 서비스 형태로 제공하고 있지만, 기업 고객의 경우 활용성이 높은 데이터는 상당한 보안이 필요하므로 프라이빗 클라우드 환경에 데이터 레이크를 구성하는 것도 현실적인 대안이 될 수 있을 것이다.

건이다. 분석의 용이성을 높이기 위해서는 프로그래밍 중심의 하둡 생태계 시스템에만 의존하기보다는 직관적이고 사용성이 뛰어난 검색 기반의 분석 기술을 적용해야 할 것이다. 또한 대규모 시스템 구성에 따른 고비용 부담을 줄이기 위해 클라우드 기술을 접목해야만 현실적인 적용이 가능하다.

현재는 퍼블릭 클라우드 서비스에서만 해당 데이터 레이크를 서비스 형태로 제공하고 있지만, 기업 고객의 경우 활용성이 높은 데이터는 상당한 보안이 필요하므로 프라이빗 클라우드 환경에 데이터 레이크를 구성하는 것도 현실적인 대안이 될 수 있다. 데이터 품질을 높이기 위해서는 기업 내 데이터 거버넌스가 내재돼야 하며, 아울러 데이터 레이크 솔루션에서 관련 기능도 보완돼야 한다.

제 5 장

데이터 활용 솔루션

비즈니스의 인사이트를 얻기 위해서는 데이터를 활용한 분석이 필요하다. 데이터 활용 솔루션은 빅데이터와 인공지능으로 대표되는 시대 흐름을 담기 위해 시각화 분석 기능과 예측 분석 및 시뮬레이션 기능을 지속적으로 개발 중이며, 고급분석 기능을 자체적으로 쉽게 제공하는 제품으로 발전하고 있다.

1. 개요

기업은 데이터에서 고객 가치를 분석하고 고객에게 개인화·맞춤화 서비스를 제공하며, 기업내 생산성을 향상시킨다. 또 정부는 공공데이터를 개방해 새로운 종류의 서비스 제공과 삶의 질을 개선하고 있다.

데이터를 활용하기 위해서는 먼저 분석이 선행돼야 한다. 데이터를 활용한 분석은 인사이트(통찰)를 얻는 전체 과정을 의미하는데 다양한 과학적, 공학적 기법들을 활용해 데이터를 탐색하고 데이터 속에 숨겨진 패턴을 발견하며 그 패턴을 더 잘 설명할 수 있는 통계적 모델 개발과 시각화 과정을 포함한다.

일반적으로 데이터를 활용한 분석은 BI(Business Intelligence)와 고급분석(Advanced Analytics)으로 구분할 수 있다. 기존의 BI 업체들은 고급분석 기능을 제품화하기 위해 기존 제품에 셀프 서비스(Self-Service) 기능을 강화하고 있다. 더불어 빅데이터와 인공지능으로 대표되는 시대 흐름을 담기 위해 시각화 분석 기능과 예측 분석 및 시뮬레이션 기능을 지속적으로 개발 중이다. 고급분석 기능을 자

* 필자: 박민규(비아이매트릭스 수석연구원)

체적으로 쉽게 제공하는 제품으로 발전하고 있다.

기술 측면에서는 빅데이터 분석을 위해 스파크(Spark) 기반의 하둡 환경으로 빠르게 재편되고 있으며, 오픈소스 R과 파이썬(Python)이 인기를 얻고 있다. 마지막으로 구글 클라우드 플랫폼을 기반으로 한 인공지능과 머신러닝 서비스 출시 등 BDaaS(Big-Data-As-A-Service) 서비스에 대한 관심이 늘어나는 추세다.

[표 4-5-1] BI와 고급분석의 비교

	전통적 BI	고급분석
데이터 소스	구조화된 데이터 (선행 모델링 필요)	구조화·비구조화 데이터 (선행 모델링 불필요)
분석 방법론	보고서(KPIs, 지표) 자동화된 모니터링·알림 대시보드 OLAP(Cube) 애드혹 쿼리	예측 모델·탐색 모델 데이터 마이닝 텍스트 마이닝 통계적·정량적 분석 시뮬레이션·최적화
대상 사용자	IT 부서·권한이 있는 현업 사용자	데이터 과학자·현업 사용자
의사결정 방식	결과 기반 의사결정	선제적 의사결정

2. 제품 동향

데이터 활용 솔루션은 BI, 고급분석, 공공데이터 관련 제품 등으로 구분할 수 있다. BI는 데이터 웨어하우스를 기반으로 정형데이터를 활용한 정보 서비스를 제공한다. 고급분석은 인공지능 기법 등을 활용해 미래를 예측하거나 처방하는 서비스를 제공한다. 공공데이터 관련 제품은 오픈데이터 유통과 활용을 지원한다.

가. BI 솔루션

국내 BI 제품은 비아이매트릭스의 MATRIX Suite, 야인소프트의 Octagon EOS/ERS, 위세아이텍의 WISE OLAP 등이 있다. 이들 제품은 빅데이터 분석을 위한 하둡 연동 커넥터, D3 기반의 시각화 확장 기능, 오픈소스 R 기반의 분석 연계 등을 강점으로 내세우고 있다.

글로벌 제품인 SAP BO와 IBM Cognos 역시 다양한 기능을 강화하고 있지만 전통적인 OLAP 리포팅 기능에 머물러 있다. 또한 전문 시각화 제품으로 두각을 보이는 클릭뷰(Qlikview), 태블로(Tableau) 등이 있고, 비정형 로그 분석에 특화된 스플렁크(Splunk) 등이 주목을 받고 있다.

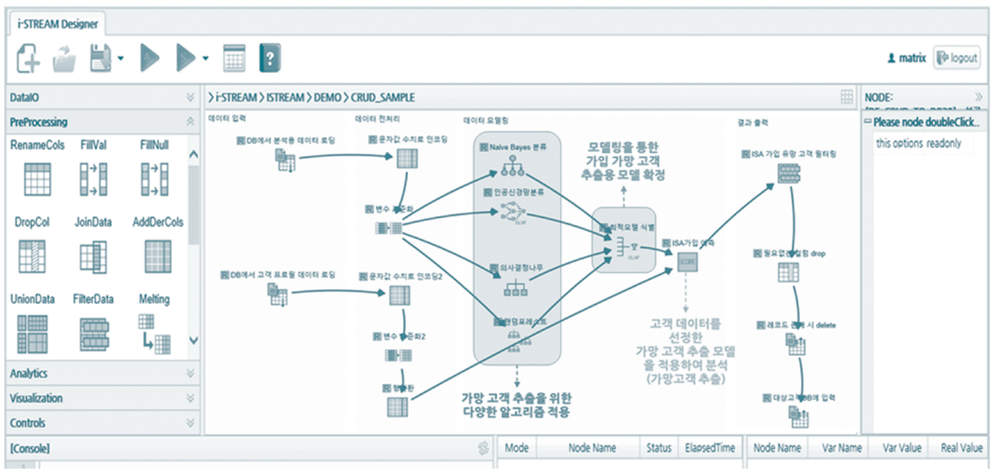
나. 고급분석 솔루션

고급분석 제품은 인공지능 기반의 빅데이터 분석 제품으로 솔트룩스의 BigO가 대표적이다. 또 다른 형태인 클라우드 기반 분석으로는 마이크로소프트 애져(Azure)의 Cortana Intelligence Suite와 구글 클라우드 플랫폼 등이 있는데, 이들 제품은 인공지능·머신러닝 알고리즘을 이용해 빅데이터 예측 분석을 제공하고 있다.

기본 BI 제품에 고급분석 기능을 추가해 기능을 확장한 솔루션도 잇따라 출시되고 있다. 위세아이텍, 비아이매트릭스와 마이크로스트래티지 등은 기본 BI 제품에 인공지능·머신러닝 기반의 데이터 마이닝 기능을 추가해 고급분석 기능을 강화하고 있다.

고급분석을 위해 오픈소스 R 기반의 분석이 확산되는 가운데, 파이썬의 아나콘다(Python Anaconda) 배포판 설치의 용이성과 주피터(Jupyter) 기반의 편리한 개발 환경 때문에 판다스(Pandas), 사이킷런(Scikit-learn) 기반 분석 방법도 인기를 얻고 있다.

[그림 4-5-1] 데이터 활용 개요도



3. 기술 동향

빅데이터 분석(BDA)을 위해 하둡 진영이 급변하고 있다. 기존 맵리듀스(Map Reduce) 기법을 개선한 스파크(Spark)가 등장하면서 하둡 환경이 빠르게 재편되고 있다. 하둡 HDFS에 저장된 데이터에 SQL 질의를 쓸 수 있도록 한 SQL-on-Hadoop 제품인 아파치 타조(Apache Tajo), 클라우데라(Cloudera)의 임팔라(Impala) 등이 등장하면서 빅데이터 분석 시장으로 접근이 가능해졌다. 한편 MongoDB로 대표되는 NoSQL 기술은 복수의 저가 서버들을 클러스터링, 샤딩 등의 방법으로 데이터를 분리·처리하면서 대량의 데이터를 빠르게 처리할 수 있는 기술로 평가받고 있다. 또한 필요에 따라 Key-Value, Documents, Graph Database 등으로 활용할 수 있다는 점이 특징이다.

SAP HANA로 대표되는 인메모리 DBMS(In-Memory DBMS, IMDBMS) 기술은 디스크가 아닌 주 메모리에 모든 데이터를 저장하는 데이터베이스로 디스크 검색보다 자료 접근이 훨씬 빠른 것이 가장 큰 장점이다. 데이터가 빠르게 증가하면서 데이터베이스 응답 속도가 떨어지는 문제를 해결할 수 있는 대안으로 평가받는다. 또 기존의 DW 환경과 기반 솔루션들을 그대로 사용할 수 있다는 점에서 강점이 있다.

글로벌 IT 기업들이 딥러닝 기술을 확보하기 위한 다양한 노력이 진행 중이다. 특히 구글은 머신러닝 기술인 텐서플로(TensorFlow)를 오픈소스로 공개해 인공지능 기술에 대한 영향력을 강화하고 있다. 마이크로소프트 또한 자신들의 DMTK(Distributed Machine Learning Toolkit)라는 개발자용 서버 기반 프레임워크를 공개했다. 여러 개의 서버를 병렬로 연결하여 딥러닝을 수행할 수 있는 장점이 있다. 중국 최대 검색엔진 바이두(Baidu) 역시 구글의 딥러닝 프로젝트를 이끌던 앤드류 응(Andrew Ng) 교수를 영입해 Warp-CTC(Connectionist Temporal Classification)라는 학습 알고리즘을 구현한 소프트웨어를 공개했다.

4. 향후 발전 방향

IDC 보고서에 따르면 빅데이터를 위한 클라우드 플랫폼 도입과 전문 분석 서비스에 대한 관심이 높아지면서, BDaaS와 같은 새로운 빅데이터 공급체계가 주목받고 있다고 한다. 그 대표적인 예로 BigML(https://bigml.com)을 들 수 있다. 빅데이터가 여전히 중요한 요소임이 분명하지만, 앞으로는 데이터 자체보다는 데이터를 통해 가치를 창출해내는 분석 기술과 서비스에 더 집중될 것이다. 고도화되고 예측 가능한 분석 기능들이 인공지능과 통합된 데이터 분석 애플리케이션의 성장으로 이어질 전망이다.

또한 개발자, 데이터 과학자, 데이터 분석가들에게 직접 데이터를 탐사할 수 있는 권한을 부여해주는 거버넌스 기능의 빅데이터 '셀프서비스' 환경으로 진화할 것이다. 기존 분석 기능에 다양한 인공지능·머신러닝 기반 기술의 접목으로 좀 더 정교하고 빠른 예측분석(Predictive analytics)도 가능해질 것으로 전망된다.

제6장

데이터 분석 솔루션

데이터 분석은 OLAP 중심의 BI, 시각화 중심의 리포팅에서 통계 분석, 데이터 마이닝, 머신러닝과 같은 고급분석으로 진화하고 있다. 전통적인 BI 솔루션 업체를 중심으로 기술이 확장되고, 데이터 분석 결과를 적용한 활용 중심의 통합 솔루션과 서비스로 발전하고 있다. 데이터 분석은 데이터 수집, 처리와 유기적인 관계를 가지고 있어 연관 솔루션과의 융합과 협업이 필수다.

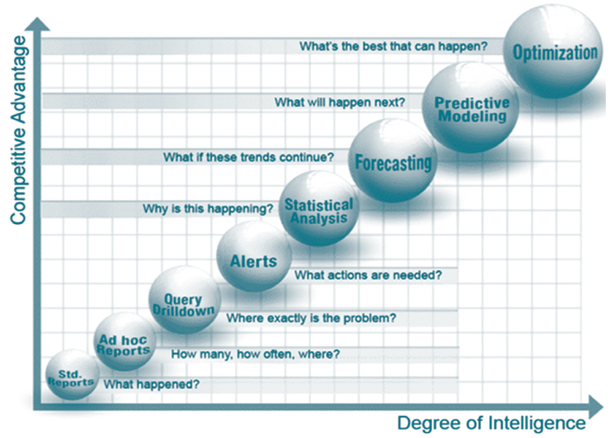
1. 개요

전통적인 데이터 분석은 비즈니스 인텔리전스(Business Intelligence, BI) 솔루션 기업들이 주도해왔다. 초기 BI 솔루션은 정형 DBMS를 기반으로 비정형 보고서, 정형 보고서와 대시보드 서비스를 제공했으나, 빅데이터 시대를 맞이해 통계 분석이 강화되고 시각화 등의 신기술을 적용하면서 계속 진화하고 있다. 데이터 분석 기술은 기존 BI 솔루션에서 다루어 왔던 정형 데이터와 검색 솔루션 업체 중심의 비정형 데이터를 융합해 분석하고 데이터 마이닝, 머신러닝을 적용한 분류와 예측으로 진화하고 있다.

데이터 분석은 OLAP 중심의 BI, 통계분석, 데이터 마이닝과 머신러닝으로 구분한다. 기존 BI 솔루션은 OLAP 중심에서 기본적인 통계 분석이 주요 기능이었으나 빅데이터 기술이 대두되면서 고급 통계분석, 데이터 마이닝과 머신러닝을 연계해 제공한다. 자체적으로 데이터 마이닝과 머신러닝을 개발해 적용하거나 오픈소스 R, 파이썬(Python), 스파크(Spark)를 연계해 제공하기도 한다. 분석 시

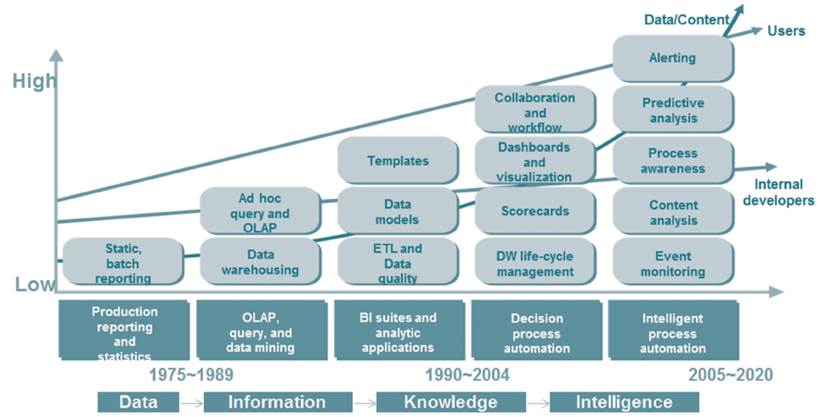
* 필자: 김선영(화수목 대표)

[그림 4-6-1] 데이터 분석의 진화



출처 : IDC

[그림 4-6-2] 데이터 분석 솔루션의 흐름



출처 : IDC

스템 구성은 기존 인프라와 연계해 확장하거나 구축한 빅데이터 플랫폼을 활용하기도 한다. 최근 데이터 분석 솔루션은 통합적으로 구축하기 때문에 자체 연관 솔루션을 연계하거나 분야별 전문회사가 협업해 구축하기도 한다.

2. 제품 동향

데이터 분석 분야의 국내 솔루션은 웹기반의 통합 솔루션인 위세아이텍의 Wise-OLAP과 엑셀 기반으로 시작한 비아이매트릭스의 MATRIX가 주도하고 있다. 오브젠의 DATAPLANET, 야인소프트의 Octagon EOS/ERS도 대표적인 제품이다. 해외 솔루션으로는 BI 중심의 SAP BO, Microstrategy MSTR과 시각화 중심의 태블로소프트웨어(TableauSoftware)의 Tableau, 클릭테크(QlikTech)의 QlikView, 팁코소프트웨어의 TIBCO Spotfire가 대표적이다. 오픈소스 하둡 기반의 맵리듀스, 오픈소스 R 기반의 분석 제품인 Shiny나 ggplot2도 있지만 아직까지는 시범 구축이나 학습용으로 많이 사용되고 있다. 최근 대다수 데이터 분석 솔루션 업체는 빅데이터 기술을 적용해 기존 솔루션을 확장하거나 다른 솔루션과 융합해 고부가가치의 솔루션을 제공하고 있다.

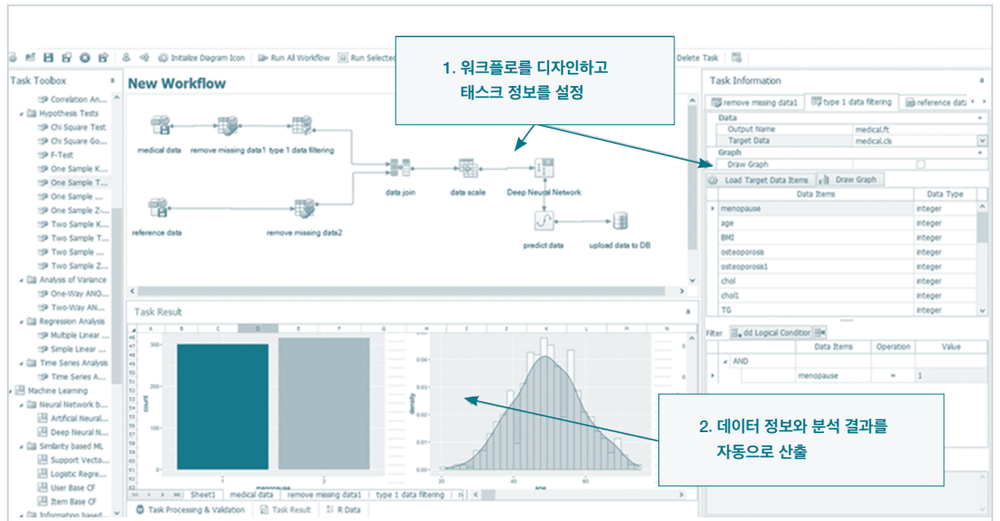
위세아이텍의 WISE OLAP의 경우 머신러닝 솔루션인 WISE Advisor, 시각화 솔루션인 WISE Visual, 캠페인 솔루션인 WISE Campaign과 연계해 다양한 통합 솔루션을 제공한다. 빅데이터 시대를 맞이해 대두된 빅데이터 품질과 관련해 데이터 품질 솔루션인 WISE DQ와도 연계해 서비스를 제공한다. 또한 클라우드 시장이 활성화될 것을 대비해 3년 전부터 고객사를 대상으로 클라우드 서비스를 제공하고 있다.

비아이매트릭스의 MATRIX는 빅데이터 분석 플랫폼인 i-BIG과의 연계로 대용량 데이터 분석 서비스를 제공하며, 오브젠의 DATAPLANET은 오픈소스 기반 기술을 적용해 R, D3, 하둡 기반의 개방형 구조가 특징이다. 야인소프트의 Octagon EOS/ERS에 추가해 인메모리 기반의 Octagon BI Platform, 이종 데이터, R/GIS 연계 분석을 Octagon Advantage, 시각화 중심의 Octagon Visualization을 제공한다.

어니컴의 ankus는 하둡 기반의 맵리듀스(MapReduce)로 설계된 알고리즘을 포함한 오픈소스 빅데이터 마이닝 도구를 제공한다.

엑셈은 오픈소스 기반의 빅데이터 플랫폼에서 스파크를 활용해 분석하며, 화수목은 오프소스 R 기반의 데이터 분석을 위한 전처리, 통계와 머신러닝 기능을 제공한다.

[그림 4-6-3] 오픈소스 R 기반의 데이터 분석 솔루션 사례



출처: '화수목 Analytics'

3. 기술 동향

데이터 분석 기술은 전통적인 정형 데이터의 집계와 보고서 기능 위주에서 통계분석과 데이터 마이닝, 머신러닝을 적용하고 있다. 데이터 분석 솔루션 업체들은 기존 제품의 기능 확장, 유관 제품과의 통합 서비스 제공과 오픈소스 R·파이썬과의 연계를 통해 진화하고 있으며, 오픈소스 스파크를 적용하는 기업 또한 늘고 있다.

데이터 분석은 데이터로부터 숨겨진 패턴과 알려지지 않은 정보 간의 관계를 찾아가는 과정이다. 이미 수많은 분석 방법이 도출돼 왔고, 데이터 마이닝이나 분류, 예측 분석 방법들이 분석을 위해 활용되고 있다. 동일한 데이터를 여러 관점에서 바라보는 다차원 데이터 분석인 OLAP, 데이터 특성을 분석해 유사성을 기초로 그룹화하는 군집분석, 데이터 간의 관련성을 분석하는 상관분석은 기존에도 자주 활용돼 왔다. 머신러닝을 활용한 사례 기반의 추론, 연관 분석, 인공 신경망, 의사결정나무, 유전자 알고리즘 등의 방법을 활용하는 사례가 점점 늘어나고 있다. 이는 데이터 분석 기술뿐만 아니라 데이터의 수집, 처리 기술의 발전으로 가능하게 된 것이다.

최근에는 텍스트 마이닝 기술을 기반으로 한 활용 사례인 챗봇이 활성화 되고 있다. 이외에 다양한 분석 결과를 선택하기 쉽지 않아 여러 가지 제시된 대안 중 최적의 대안을 찾도록 지원하는 최적화 기법이 대두되고 있다. 사회적 이슈와 맞물려 소셜 네트워크 연결 구조와 강도를 분석해 영향력을 판단하는 소셜 네트워크 분석 기술도 사용되고 있다.

4. 시장 동향

데이터 분석 시장은 배치에서 실시간 분석으로, 실적 중심에서 예측 분석으로, 정형 데이터 분석에서 정형과 비정형 데이터의 융합 분석으로 발전되고 있다. 여기에 전통적인 데이터 분석 솔루션에서 종합적인 데이터 분석 솔루션으로 진화되면서 유관 솔루션과의 융합으로 또 다른 부가가치를 창출하는 중이다. 또한 다수의 데이터 분석 솔루션 업체가 클라우드 서비스를 적용하고 있다.

과거에는 대용량 데이터 처리 시간이 많이 소요됐으나 서버·디스크·네트워크 등 인프라의 발전으로 빠른 처리가 가능하게 돼 고객들이 원하던 실시간 결과를 나타낼 수 있게 됐다. 또한 단순 데이터 집계나 통계분석에서 데이터 마이닝, 머신러닝을 적용해 다양한 분류와 예측 분석이 가능하게 됐다. 정형 데이터 위주의 분석에서 정형·비정형 데이터의 융합 분석을 필요로 하게 됐으며, 기존에 OLAP 중심이거나 시각화된 보고서 위주의 데이터 분석 솔루션이 통계분석, 데이터 마이닝, 머신러닝과 같은 고급분석 기능을 확장하거나 유관 솔루션과의 연계를 통해 분석에서 끝나지 않고 활용까지 감안한 통합 솔루션과 서비스를 제공하는 사례가 늘고 있다.

5. 향후 전망

데이터 분석 솔루션은 정형 데이터와 비정형 데이터의 융합 분석, 수집된 데이터의 실시간 분석, 분석 결과를 신속하게 활용하기 위한 유관 솔루션과의 밀접한 연계가 주요 이슈가 될 것이다. 또한 데이터 마이닝과 머신러닝이 쉽고 빠르게 적용되고 다양한 결과를 선택하기 위한 최적화가 대두될 전망이다. 내외부 데이터를 융합해 분석하기 위해 데이터 유통 생태계가 점진적으로 조성되고, 클라우드 환경에서 서비스하는 비중이 늘어날 것으로 보인다.

제4차산업혁명 시대에 들어서면서 근간이 되는 데이터, 특히 데이터 분석이 중요해졌다. 국내 기업들의 빅데이터 도입율은 약 4.3% 수준으로 향후 빅데이터 수요는 전체 기업의 30.2% 수준으로 예상된다. 도입 고려 시기는 2018년(77개)과 2019년(98개) 이후이다.* 데이터 분석이 특정 산업이나 기업에만 적용 가능하고, 기업 내에서도 특정 부서나 담당자만의 문제라는 인식의 전환이 이뤄져야 한다. 데이터 기반의 의사결정과 같은 문화 조성도 요구되며, 의사결정권자의 추진 의지와 관심도 필요하다. 그렇게 되면 사내 부서간 데이터 공유와 활용도 활성화 될 수 있다.

데이터 분석에 대한 인식과 문화가 바뀌고 활용 목적과 활용할 수 있는 데이터에 대한 정확한 파악 및 추가 내외부 데이터 확보, 분석 단계별 필요 인력과 솔루션 확보 등에 대한 노력을 기울일 때 데이터 분석 결과를 활용할 수 있는 길이 활짝 열린다. 이러한 데이터 분석 활성화에 대한 노력이 결실을 맺을 때 국내 데이터산업뿐만이 아닌 대부분의 산업이 제4차산업혁명 시대에 부응할 수 있는 경쟁력을 확보할 수 있다.

* 「2015년 빅데이터 시장현황 조사」, 한국정보화진흥원, 2016

제7장

마스터데이터 관리 솔루션

마스터데이터는 기업과 조직 전반에서 여러 응용 시스템에 걸쳐 공유해야 하는 자산이다. 제4차산업혁명 시대의 스마트 팩토리, 인공지능과 빅데이터 등에 있어서 정확하고 의미있는 실행과 분석을 위해 마스터데이터 관리는 필요하다.

1. 개요

마스터데이터(Master Data)는 ERP와 SCM, MES, MMS와 같은 트랜잭션 시스템이나 BI, EDW와 같은 분석 시스템에서 실행의 기준으로 사용되는 핵심 엔티티에 대한 데이터다. 여기에는 식별 ID, 엔티티를 기술하는 정보, 프로세스 실행에 필요한 제어 정보, 마스터데이터들 간의 관계 정보가 포함된다. 마스터데이터의 예로는 제품, 자재, 고객, 공급사, 위치, 조직, 사원 등을 들 수 있다. 비즈니스 커뮤니케이션의 신속성, 정확성, 경제성을 높이기 위해 사용되는 부호화된 데이터인 업무 코드(Business Code) 또는 표준 참조 코드(Reference Code)도 마스터데이터에 포함될 수 있다.

마스터데이터는 기업과 조직 전반에서 여러 응용 시스템에 걸쳐 공유해야 하는 정보 자산이다. 그러므로 마스터데이터는 데이터 표준체계를 따라야 하고, 정확성과 일관성 등 데이터 품질이 보장되어야 한다. 또한 데이터 원칙과 정책, 권한과 역할, 관리 프로세스 등 데이터 거버넌스가 지켜져야 한다.

* 필자: 오세창(투비웨이 대표)

2. 마스터데이터 관리 솔루션의 정의

마스터데이터 관리 솔루션(Master Data Management Solution, MDM 솔루션)은 마스터데이터에 대한 모델 관리, 거버넌스 관리, 품질 관리, 프로세스 관리, 연계·배포 관리 기능이 통합되어 있는 업무 중심적인 데이터 관리 솔루션이다.

MDM 솔루션을 개별 기능 관점에서 보면 메타관리 솔루션, 데이터 품질 관리(이하 DQ) 솔루션, 데이터 통합(이하 DI) 솔루션 등과 유사한 측면이 많다. ERP나 MES 같은 트랜잭션 시스템의 마스터데이터 관련 생성, 변경, 삭제 메뉴 기능과 유사해 보일 수 있다. 하지만 다른 유형의 데이터 관리 솔루션들이 일반적으로 트랜잭션 데이터나 분석 데이터 등 모든 유형의 데이터를 대상으로 IT 관리자를 위해 특정 분야에 집중한 관리 기능을 제공하는데 반해, MDM 솔루션은 마스터데이터를 전사적 차원에서 다룰 수 있도록 하기 위해 필요한 기능을 업무 사용자에게 통합적으로 제공한다는 점에서 구별된다.

●○ 마스터데이터는 데이터 표준체계를 따라야 하고, 정확성과 일관성 등 데이터 품질이 보장되어야 한다. 또한 데이터 원칙과 정책, 권한과 역할, 관리 프로세스 등 데이터 거버넌스가 지켜져야 한다.

국내외 주요 MDM 솔루션 벤더들은 대부분 자사의 주력 솔루션인 데이터 품질 관리 솔루션, 데이터 통합 솔루션, ERP 시스템과는 별도로 마스터데이터 관리에 특화된 MDM 솔루션을 공급하고 있으며, 자사 다른 솔루션과의 통합을 강조하고 있다.

3. 국내외 MDM 솔루션의 특징

국내 MDM 솔루션 시장에서는 마스터데이터 관리 애플리케이션, ERP, 데이터 관리 플랫폼(Data Management Platform) 분야 국내외 벤더들이 기능, 가격, 구현 등을 무기로 경합하고 있다. 국내 MDM 솔루션 업체로는 투비웨이, 데이터스트림즈, 아이큐엠씨 등이 있으며, 글로벌 MDM 솔루션 업체로는 SAP, IBM, 오라클, 인포매티카, 스티보시스템즈(StiboSystems) 등이 있다(표 4-7-1 참조).

가. 기능 측면

마스터데이터 관리 애플리케이션 벤더의 MDM 솔루션은 데이터 모델의 유연성과 사용자 UI의 편리성이 강점이고, ERP 벤더의 MDM 솔루션은 자사 ERP 제품과의 모델 통합성과 연계의 용이성을 강점으로 내세우고 있다. 데이터 관리 플랫폼 벤더의 MDM 솔루션은 데이터 모델의 범용성과 자사 DQ, DI 제품과의 통합이 강점이다.

나. 가격 측면

가격 관점에서는 다른 양상이 펼쳐진다. 글로벌 MDM 솔루션 업체들은 레코드 건수와 사용자 수에 비례해 가격 라이선스 정책을 유지한다. 마스터데이터 레코드 건수가 많은 경우에는 솔루션 도입 비용이 높고, 상대적으로 높은 유지보수 요율로 인해 연간 솔루션 유지보수 비용이 높다는 점이 약점으로 지적된다. 이에 따라 2000년대 초중반에 글로벌 MDM 솔루션을 도입해 운영해 오던 일부 국내기업들이 최근 들어 국산 MDM 솔루션으로의 교체를 검토하거나, 자체 재개발을 고려하고 있다.

다. 구현 측면

마스터데이터 관리 애플리케이션 벤더의 MDM 솔루션은 데이터 모델링 템플릿을 제공해 모델링 기간을 단축하고, ERP 이외의 실행 시스템에서 필요로 하는 마스터데이터를 위한 기능까지 제공함으로써 마스터데이터에 대한 관리 수준을 높이고 대상 범위를 확대하면서 경쟁력을 강화한다.

ERP 벤더는 MDM 솔루션이 ERP 패키지 솔루션과 동일한 데이터 모델을 사용하고, 마스터데이터 유형별로 특화된 관리 화면을 기본으로 제공하며, ERP 연계 어댑터를 사용해 용이하게 구현할 수 있다는 점에서 경쟁력이 높다고 할 수 있다. ERP 이외의 타 시스템에서 필요로 하는 마스터데이터 유형과 관리 속성까지 마스터데이터 범위를 확장하는 경우에는 커스텀 데이터 유형을 구현하고 이를 연계하는 노력이 필요하다.

데이터 관리 플랫폼 벤더의 MDM 솔루션은 자사의 DQ, DI 제품을 함께 활용함으로써 품질 관리가 가능하고 데이터 연계가 가능하다는 점에서 강점이 있다. 그러나 국내 제조산업 등에서 필수 요구사항인 분류별 속성체계 기능과 일반 사용자용 UI가 미흡하기 때문에 이를 보완하기 위한 모델링 및 추가 UI 개발이 필요한 것은 약점이다.

4. 국내외 MDM 솔루션 현황

데이터스트림즈의 MasterStream은 데이터관리 플랫폼 회사로서의 역량을 활용해 전사 마스터데이터 거버넌스 기능, 마스터데이터 품질 관리(DQM) 기능, 업무 규칙 설정과 적용을 통한 사전/사후 검증 기능, 데이터 값에 의한 비즈니스 프로세스 통제 연동 기능을 제공한다.

아이큐엠씨의 Global MDM은 활용 정의 설정(Configuration Setting) 기능을 제공해 추가 개발이 필요없다. 기존 BI 역량을 활용한 마스터데이터 지표 및 운영 현황 리포트 기능을 제공한다.

투비웨이의 ToBeWAY Enterprise MDM은 분류별 속성 체계를 포함한 유연한 데이터 모델, 인라인 데이터 품질 검증, 데이터 배포 모니터링, 마스터데이터 360뷰 등의 기능을 제공해 국내 하이테크, 화학, 유통, 물류, 건설, 통신, 식품산업 분야에서 다수의 레퍼런스를 확보하고 있다. ToBeWAY Operational MDM은 MES, SCM, WMS, MMS 등에서 필요한 제조 마스터데이터, 계획 마스터데이터, 물류 실행 마스터데이터, 유지보수 마스터데이터 관리 기능을 제공한다. ToBeWAY Standard Reference Code는 표준 참조코드를 관리한다.

인포매티카의 Multi-domain MDM은 2017년 초 발표된 가트너의 Magic Quadrant for MDM에서 리더로 평가받은 솔루션이다. 여러 시스템에서 분산 관리되는 마스터데이터 데이터에 대해 골든 레코드 관리가 용이하고, 자사 DQM 및 ETL 제품 기능과 통합돼 MDM 병합(Consolidation) 시나리오 구현이 용이하다. 데이터 모델링은 범용적이나 분류별 속성체계 지원에 제한이 있어 제품 및 자재 도메인에 MDM을 적용하기에는 약점이 있다. IDD라 불리는 데이터 관리자용 UI 이외에 현업 사용자용 UI가 필요한 경우에는 커스텀 UI 개발이 필요할 수 있다. Product 360 제품은 하일러 소프트웨어(Heiler)의 제품을 인수한 PIM(Product Information Management) 솔루션으로 제품·상품 콘텐츠를 관리하고 싱글뷰와 옴니채널 고객경험 등을 제공하지만 Multi-domain MDM 등 자사의 타 제품들과

●○ 국내외 주요 MDM 솔루션 벤더들은 대부분 자사의 주력 솔루션인 데이터 품질 관리 솔루션, 데이터 통합 솔루션, ERP 시스템과는 별도로 마스터데이터 관리에 특화된 MDM 솔루션을 공급하고 있다.

의 기능 통합은 이뤄지지 않았다.

SAP의 Master Data Governance(이하 MDG)는 재무, 공급사, 고객, 자재 데이터 도메인에 특화된 마스터데이터 관리 애플리케이션이다. 커스텀 데이터 모델도 지원한다. 자사 ERP 시스템의 데이터 모델을 기본으로 사용하며 ERP 시스템에 존재하는 비즈니스 로직과 컨피규레이션을 MDM에서 바로 활용할 수 있어 주로 ERP를 중심으로 한 마스터데이터 중앙관리(Central Management) 시나리오 구현에 용이하다. 최근 데이터 적재, 표준화, 매칭 등 병합 시나리오를 지원하는 제품을 출시한 바 있다. 고유의 개발언어인 ABAP를 사용하기 때문에 SAP ERP를 사용하지 않는 조직은 MDG 사용이 어려우며, ERP 이외의 마스터데이터 데이터를 수용하기 위해서는 모델링 및 구현 노력이 필요하다. MDG는 ERP 솔루션과 함께 번들링되어 제안되는 경우가 많다. MDM 운영 시 발생하는 유지보수 비용과 마스터데이터 데이터 건수가 증가해 라이선스 허용 범위를 초과할 때 발생하는 감사위험(Audit Risk)은 약점으로 지적된다.

[표 4-7-1] 주요 MDM 솔루션 현황

	회사	MDM 회사유형	국가	주요 MDM 제품	MDM 유형
국내	데이터스트림즈	데이터 관리 플랫폼	한국	MasterStream	Multi-domain
	아이큐엠씨	MDM 애플리케이션	한국	Global MDM	Multi-domain
	투비웨이	MDM 애플리케이션	한국	ToBeWAY Enterprise MDM v.8.5	Multi-domain
				ToBeWAY Operational MDM	Operational
				ToBeWAY Standard Reference Code	Reference
해외	EnterWorks	MDM 애플리케이션	미국	EnterWorks Enable v.8.1	Multi-domain
	IBM	데이터 관리 플랫폼	미국	InfoSphere MDM v11.5	Multi-domain
				InfoSphere MDM Reference Data Hub	Reference
해외	Informatica	데이터 관리 플랫폼	미국	Informatica MDM v.10.2	Multi-domain
				Supplier 360 v.10.1, Customer 360	Domain Specific
				Cloud Customer 360 for Salesforce	Cloud
				Product 360 v.8.0.5 (2012년 Heiler PIM 인수)	Domain Specific
	Magnitude Software	MDM 애플리케이션	미국	Kalido MDM v.9.1 SP3	Multi-domain
				Magnitude One	Cloud
	ORACLE	데이터 관리 플랫폼, ERP	미국	Product Hub, Supplier Hub, Site Hub vR12.2	Domain Specific
				Customer Hub vIP 2016	Domain Specific
				Product Cloud Service, Customer Cloud Service vR11	Cloud
				Data Relationship Management vR11.1 (Hyperion)	Multi-domain
	Orchestra Networks	MDM 애플리케이션	프랑스	EBX5 v.5.7	Multi-domain
	Riversand	MDM 애플리케이션	미국	MDMCenter v.7.8	Multi-domain
	SAP	ERP, 데이터 관리 플랫폼	독일	Master Data Governance (MDG) v.9.0	Multi-domain
				SAP NetWeaver MDM 7.1 SP17	Multi-domain
				Hybris Product Content Management (PCM) v.6.2	Domain Specific
	Stibo Systems	MDM 애플리케이션	덴마크	Stibo Enterprise Platform (STEP) Trailblazer v.8.0	Multi-domain
	TIBCO Software	데이터 관리 플랫폼	미국	TIBCO MDM v.9.0	Multi-domain
				TIBCO Cloud MDM v.9.0	Cloud
	Reltio	MDM 애플리케이션	미국	Reltio Cloud	Analytics, Cloud

(계속)

(회사 표기 순서는 가나다, 알파벳 순)

5. 향후 발전방향

제4차산업혁명 시대의 스마트 팩토리, 인공지능과 빅데이터 등에 있어서 데이터 특히 마스터데이터 관리는 정확하고 의미있는 실행과 분석을 위해 필수적이다. 최근 10여 년 동안 해외 MDM 솔루션들은 특정 도메인에 특화된 MDM(Domain-Specific MDM)에서 다중 도메인 지원 MDM(Multi-Domain MDM)으로 발전해 왔다. 현재 대부분 MDM 솔루션 벤더들은 자사 MDM 솔루션을 PIM(제품정보관리)이나 CIM(고객정보관리) 제품으로 한정하기보다는 다중 도메인 지원 솔루션으로 포지셔닝하고 있다. 2016년까지 제품 데이터 솔루션과 고객 데이터 솔루션 2개 분야로 나누어 각각 매직 쿼드런드(Magic Quadrant)를 발표하던 가트너가 이를 하나로 통합해 2017년 'Magic Quadrant for MDM'을 발표한 바 있다. 한 개의 마스터데이터 도메인에서 MDM 구축을 시작하더라도 기 구축한 MDM 솔루션을 기반으로 마스터데이터 도메인을 확대하려는 고객의 요구사항을 반영한 결과다.

최근 가트너는 다중 벡터(Multi-Vector) MDM이란 용어를 사용하기 시작했다. 이는 MDM 복잡성에 영향을 미치는 5가지 벡터, 즉 산업, 데이터 도메인, 사용 케이스(디자인용, 운영용, 분석용), 조직구조(중앙집중형, 연방형, 로컬형), 구현 스타일(레지스터리형, 병합형, 동시존재형, 중앙관리형)을 모두 지원할 수 있는 MDM 솔루션이란 의미로 보인다. 향후 몇 년간 MDM 솔루션은 다중 도메인 지원을 넘어 여러 산업의 요구사항을 수용하며 여러 구현 스타일을 동시에 지원하고, 운영 MDM과 분석 MDM을 동시에 지원하는 방향으로 발전할 것으로 예상된다. 아직 국내에서는 생소하지만 외국 MDM 벤더들이 도입하기 시작한 클라우드 MDM이나 그래프 기반 분석 MDM 솔루션도 점차 늘어날 전망이다.

국내 MDM 솔루션 도입은 빠르게 확장되지는 않을 것이다. 대규모 빅뱅 방식의 GSI(Global Single Instance) ERP 구축 프로젝트 감소와 전반적인 경기침체에 따른 IT 투자 위축의 영향을 받아 대규모 프로젝트 단위로 이루어지기는 어려울 것으로 예상된다. 마스터데이터 공유를 전사적으로 촉진하는 경영혁신 차원에서 마스터데이터를 활용해 현업 실행 업무의 애로사항을 점진적으로 해결하고, 기존 낙후된 데이터 관리 시스템을 부분적으로 대체해 나가는 방식으로 이루어질 것으로 예상된다. 그리고 ERP 이외의 실행시스템을 수행하기 위한 제조 마스터데이터, 물류 마스터데이터, 설비 마스터데이터 등에 대한 관리 요구도 높아질 것이다.

위와 같은 환경 변화에 대응하기 위해 MDM 솔루션은 데이터 표준화 과업을 지원하는 기능을 강화할 필요가 있다. 그리고 신규 시스템과 기존 레거시 시스템 사이에서 MDM은 코드 매핑 등 신·구 표준체계를 동시에 지원하는 완충 역할을 수행해야 한다. 또한 조직의 실행력을 제고할 수 있는 운영 마스터데이터에 대한 특화된 기능이 필요하다.

글로벌 MDM 솔루션들은 자국의 경쟁력이 높은 산업 분야인 유통과 금융 산업 등에서 MDM 솔루션을 고도화해 왔다. 국산 MDM 솔루션도 하이테크 제조나 화학 장치산업 등 우리나라가 경쟁력을 가지고 있는 산업계에서 쌓은 MDM 구현 경험을 적극적으로 MDM 솔루션에 내재화해야 한다. 이를 통해 국내 산업경쟁력 강화에 지속적으로 도움을 제공하고 중국, 베트남 등 제조 중심의 해외 국가에 적극적으로 진출해야 할 것이다.

제8장

데이터 품질 관리 솔루션

데이터의 품질을 유지하기 위한 데이터 품질 관리 솔루션이 빅데이터 시장 성장에 맞춰 재조명을 받고 있다. 국내외 기업들의 데이터 품질 관리 솔루션 간 차이가 있다. 이 중 국산 품질 관리 솔루션은 전통적인 데이터 품질 관리 시장을 벗어나 빅데이터 등 다양한 영역의 데이터 품질 관리 수요를 동시 수용해야 할 필요성이 대두되고 있다.

1. 개요

데이터 품질 관리 솔루션은 데이터의 품질을 유지하기 위한 기능과 프로세스를 지원하는 제품이다. 기본적으로 데이터 품질 지표, 규칙 관리, 품질 측정, 측정 결과 모니터링, 오류 데이터의 개선 활동 관리 기능과 데이터 품질 활동 절차를 지원하기 위한 프로세스 기능으로 구성된다.

빅데이터 물결에 따라 데이터 활용이 증가하면서 데이터 품질의 중요성도 더욱 강조되고 있다. 저품질 데이터로 인해 잘못된 의사결정을 할 때 생길 수 있는 사회적·경제적 손실을 예방하기 위해 정확한 데이터를 유지할 필요가 있다.

국내 데이터 품질 관리 솔루션 시장 상황은 그리 낙관적인 편은 아니다. 글로벌 기업들이 빅데이터 환경이나 다양성을 지원하는 데이터 품질 관리를 강조하면서 국내 시장에 적극적으로 침투하기 시작한 반면, 국내 기업들은 데이터 거버넌스의 구축 또는 관계형 데이터베이스 위주의 데이터 품질 관리 수요시장 공략 단계에 머물러 있기 때문이다.

* 필자: 최용준(위세아이텍 상무)

국내 기업들도 전통적인 데이터 품질 관리 시장을 벗어나 빅데이터 등 다양한 영역의 데이터 품질 관리 수요를 공략할 필요가 있다.

2. 데이터 품질 관리 솔루션 동향

국내 데이터 품질 관리 솔루션은 데이터 거버넌스 구축 수요가 있는 고객군을 위주로 강세를 보이고 있다. 국내 데이터 품질 관리 솔루션을 제공하는 국내 기업으로는 데이터스트림즈, 엔코아, 위세아이텍, 지티원 등이 있으며 여러 기업에서 솔루션 외에도 자체 개발한 데이터 품질 관리 방법론을 보유하고 있다.

[표 4-8-1] 데이터 품질 관리 솔루션 주요 기업의 제품과 특징

구분	기업명	제품명	특징
국내 기업	데이터스트림즈	QualityStream	• 분석 대상 데이터베이스를 프로파일링해 현재 상태의 품질 수준을 분석한 후 비즈니스 룰을 스케줄링해 결과를 분석하고 오류사항에 대한 정비 프로세스를 활용해 관리
	엔코아	DATAWARE DQ#	• 데이터 품질 관리 기준 정의 기능, 프로파일 및 업무규칙 관리 기능, 품질진단 실행 기능, 측정 결과 분석 기능, 오류 원인 개선 관리 기능 제공
	위세아이텍	WISE DQ	• 데이터 프로파일링, 데이터 오류 감시, 데이터 규칙 관리, 스케줄링 기능, 측정 결과의 다양한 분석 현황 제공 • 인공지능 알고리즘을 통해 비표준 데이터의 도메인 자동 판별, 이상 값 탐지, 텍스트 데이터의 개선 데이터 추천 기능 제공
	지티원	DQMiner	• 자동 데이터 프로파일링, 데이터 오디팅, 데이터 룰 관리, 데이터 품질 분석, 데이터 품질 분석 결과 보고 등 제공
글로벌 기업	인포매티카	Informatica Data Quality	• 수동 작업을 위한 워크플로, 예외 처리를 위한 검토, 수정 및 승인 기능 지원 • 데이터 품질 규칙을 다양한 환경에서 재사용, 비즈니스 규칙 가속화를 위한 템플릿을 통해 신속한 데이터 검색과 프로파일링, 구조 및 컨텍스트 검사 기능 제공
	IBM	InfoSphere Information Server	• 소스 데이터 조사, 정보 표준화의 자동화, 비즈니스 규칙 측정, 데이터 분석, 정리, 모니터링 기능 제공. • 데이터 거버넌스 지원, 데이터 유효성 검사 규칙, 시퀀싱 및 영향도 분석 기능 제공
	SAS	SAS Data Quality	• 데이터 프로파일링, 모니터링 및 프로세스의 통합, 비표준적인 기록이나 중복 기록과 알려지지 않은 데이터 유형을 수정할 수 있는 데이터 정리, 테이블, 데이터베이스, 소스 애플리케이션에서 나타나는 관계를 파악할 수 있는 데이터 프로파일링 기능 제공

가. 국내 솔루션

위세아이텍의 WISE DQ

데이터 품질 관리를 위해 인적자원이 많이 투입되는 데이터 사전분석(pre-analysis) 작업을 자동화할 수 있도록 인공지능 알고리즘을 적용하고 있다. 비표준 데이터의 메타 정보를 인공지능 알고리즘을 통해 분석해 데이터 도메인을 자동 판별하는 기능이 있으며, 빅데이터 품질 관리를 겨냥해 정확한 규칙의 적용 없이도 이상 값(Outlier)을 탐지할 수 있는 기능과 텍스트 데이터에 대해 군집 분석 후 개선 데이터를 추천하는 기능을 제공한다.

지티원의 DQMiner

기업이 데이터 품질 목표를 달성할 수 있도록 상시 데이터 품질 모니터링 시스템을 구축할 수 있게 해준다. 또한 자사 애플리케이션 영향도 분석 솔루션과 연계해 오류 데이터의 원인을 쉽게 파악할 수 있으며 애플리케이션 추적 관리, 오류 데이터 정제 관리가 가능하다. 80여 종의 데이터 품질 관련 보고서를 제공한다.

데이터스트림즈의 QualityStream

통합 리포지터리 구성을 통한 프로파일 관리 기능, 비즈니스 룰 관리 기능, 품질 진단에 대한 특정 기준별 결과 검색과 통계정보를 제공하는 검증 결과 관리 기능, 오류 데이터에 대한 정비 프로세스를 지원하는 정비 관리 기능을 제공하며, 자사 CDC 솔루션과의 연계를 통한 품질 관리를 지원한다.

엔코아의 DATAWARE DQ#

기업의 목표 달성을 위한 비즈니스 환경 예측과 성과 지향적 데이터 관리 체계 수립을 위한 설계부터 관리에 대한 토털 데이터 솔루션(DATAWARE DQ#)을 단일 리포지터리 기반으로 제공한다. 이 중 DATAWARE DQ#은 자사 모델링 도구인 DATAWARE DA#, 메타데이터 관리 솔루션인 DATAWARE Meta#과 연계해 데이터 설계를 기반으로 한 데이터 품질 관리 기능을 제공한다.

나. 글로벌 솔루션

글로벌 솔루션으로는 인포매티카(Informatica)의 Informatica DataQuality(IDQ), IBM의 InfoSphere QualityStage, SAS의 SAS Data Quality 등이 있다.

인포매티카의 Informatica DataQuality(IDQ)

10년 연속 가트너 데이터 품질 관리 툴 부문의 리더로 선정된 솔루션이다. 현업 부서 사용자가 IT에 의존하지 않고 비즈니스 규칙을 신속하게 개발할 수 있도록 지원한다. IoT, 대용량 데이터 분석, 데이터 거버넌스 및 콘텐츠 중심의 데이터 분석과 같은 새로운 시나리오를 해결하기 위해 머신러닝(Machine Learning)과 예측 분석 알고리즘을 사용한다.

IBM의 InfoSphere Information Server

데이터를 이해, 정리, 모니터링, 변환 및 제공할 수 있도록 지원하는 IBM의 정보 통합 플랫폼이다. 빅데이터, 충돌점 분석, 비즈니스 인텔리전스, 데이터 웨어하우징, 마스터데이터 관리, 애플리케이션 통합 및 마이그레이션 등 신뢰할 수 있는 정보를 생성 후 유지 관리해 전략적 비즈니스 이니셔티브를 지원한다. 데이터 품질을 클라우드 환경으로 확장하고, 단순한 데이터 구조에서 복잡한 데이터 구조에 이르기까지 대용량 데이터 정제를 지원한다. 제품 전반에 머신러닝 알고리즘을 적용하고, IoT 데이터를 지원한다.

SAS의 SAS Data Quality

다양한 소스의 ETL(Extract, transform and load) 및 ELT(Extract, load and transform) 활동에 데이터 품질을 추가할 수 있고, 기본적인 마스터데이터 관리 뷰를 지원한다. 통합된 웹 기반 콘솔을 통해 데이터 품질 작업을 모니터링하고 데이터 문제 및 거버넌스 활동을 확인할 수 있다.

글로벌 제품의 경우 관계형 데이터베이스(RDB)만을 대상으로 하는 데이터 품질 관리를 넘어서 하둡, 클라우드, 하이브리드 등 다양한 환경을 지원하고, 사물인터넷(IoT) 데이터 등 다양한 유형의 데이터, 즉 빅데이터의 품질 관리로 나아가고 있다.

●○ 빅데이터 물결에 따라 데이터 활용이 증가하면서 데이터 품질의 중요성도 더욱 강조되고 있다. 저품질 데이터로 인해 잘못된 의사결정을 할 때 생길 수 있는 사회적·경제적 손실을 예방하기 위해 정확한 데이터를 유지할 필요가 있다.

글로벌 기업들의 제품은 독립적인 제품으로 구분되기도 하지만, 빅데이터 분석, 관리 플랫폼에 통합되어 있거나 전처리(Big data Preprocessing) 과정에 통합되어 빅데이터 환경을 지원한다. 또한 다양하고 방대한 빅데이터의 특성을 고려해 인공지능 알고리즘을 활용하며, 이를 통해 데이터 품질 관리의 자동화 수준을 향상시킨 것으로 보인다. 하지만, 국내 제품은 아직 다양한 환경에서 다양한 유형의 데이터 품질을 관리할 수 있는 기능을 지원하지 않는다.

국내는 데이터 품질 관리의 거버넌스 구축에 초점이 맞추어져 있다. 데이터 거버넌스의 구축이 매우 중요한 요소인 것은 틀림없다. 하지만 제4차산업혁명으로 데이터 활용 환경이 급변하고 활용 가능한 데이터의 유형이 증가하는 상황에서 다양한 환경에서의 데이터 품질 관리나 다양한 유형의 데이터를 진단할 수 있는 기술은 아직 요원한 상태다.

국내 데이터 품질 관리 솔루션들도 데이터 품질 관리 프로세스, 데이터 품질 진단, 오류 개선 활동의 효율을 높이기 위한 자동화 기능을 개발 또는 고도화하거나 다양한 환경에서 다양한 유형의 데이터 품질을 진단할 수 있도록 기술 개발이 필요하다. 이를 위해 해외와 같이 인공지능 기술도 적극적으로 활용해 볼 만하다.

3. 향후 전망

빅데이터 시장이 폭발적으로 성장하고 있고, 빅데이터의 활용 영역도 확대되고 있다. 해외의 한 설문조사에 따르면 데이터 품질 관리가 가장 필요한 부문 1위에 빅데이터가 랭크되었다. 가트너 매직 쿼드런트(Magic Quadrant) 보고서에는 데이터 품질 관리의 혁신적 기술 요소로 머신러닝을 꼽기도 했다.

IBM 왓슨(Watson)이 환자를 분석하는 데 10초면 충분하다고 한다. 그런데 왓슨에게 제공된 환자의 엑스레이 사진이 좌우가 바뀐다면 어떻게 될까? 또 실시간 분석이 필요한 IoT 데이터의 경우나 SNS 데이터의 품질은 어떻게 관리해야 할까? 제4차산업혁명으로 산업의 패러다임이 바뀌면서 데이터 품질 관리의 패러다임도 바뀔 것이다.

환자의 생체정보 또는 엑스레이와 같은 이미지의 품질, 미세먼지 농도를 측정하는 실시간 센서 데이터의 정확성, 문서의 항목과 내용 일치 여부 또는 개인정보 비식별화 등 빅데이터 활용 영역 확대와 함께 기존 정형 데이터의 데이터 품질 관리 방법으로는 해결할 수 없는 이슈들도 증가하고 있다. 이런 이슈를 해결하기 위한 새로운 데이터 품질 관리 방법의 필요성이 더욱 대두될 것이다. 이로 인해 2~3년 내에는 정형 데이터에 국한되었던 데이터 품질 관리에서 벗어나 빅데이터의 품질 관리를 위한 품질 지표와 방법이 새롭게 정의되고, 다양한 유형의 데이터 측정과 빅데이터 환경을 지원할 수 있는 기술발전이 이루어질 것으로 전망된다.

제9장

데이터 보안 솔루션

많은 데이터가 수집되고 이를 이용한 비즈니스 모델들이 나오면서 데이터 보안이 중요한 과제가 되고 있다. 개인이나 조직이든 모든 사람에게 공유할 수 없는 데이터들이 당연히 존재할 수밖에 없기 때문이다. 데이터 보안에 대한 전반적인 상황과 개요 그리고 관련 솔루션의 트렌드를 살펴본다.

1. 개요

모바일, 클라우드, 사물인터넷(IoT) 등 정보기술이 발전함에 따라 언제 어디서나 원하는 정보에 접속할 수 있게 됐다. 업무에 활용되는 다양한 디바이스와 서비스들을 통해 수많은 중요한 데이터들이 생산·유통되고 있다. 더불어 APT(Advanced Persistent Threat)와 같은 더 지능적인 해킹 방식이 급증하고 내부 사용자에 의한 보안 위협이 늘면서, 네트워크 또는 시스템 레벨에서 차단하는 경계선 기반 보안 솔루션으로는 정보유출 사고를 100% 막을 수 없게 됐다.

멀웨어, APT 등 대부분의 해킹 목적은 결국 중요 정보를 탈취하는 데 있는데, 경계선을 설치하고 방어막을 아무리 높게 쳐도 해커들은 어떻게든 뚫고 들어올 수 있다. 물론 경계선을 지키고 방어막을 높이는 전략도 기본적으로 해야 하지만, 이제는 지켜야 할 핵심 가치인 주요 데이터를 안전하게 보호하는 데 더욱 집중해야 한다.

지금까지 데이터 자체를 보호하는 보안 솔루션으로는 보안의 기능 중 기밀성을 보장하기 위해 데이터를 암호화하는 솔루션이 기본이었다. 단순 암호화 솔루션은 데이터가 저장된 상태에서 권한이 없는 사람이 볼 수 없도록 하는 기능만 있으므로, 데이터의 특성상 여러 사람이 공유하고 이동하는 경우 효율적이지 못했다. 여기에서 더욱 발전한 솔루션으로는 문서보안 솔루션이라고 부르는 DRM 기술 기반의 솔루션이 있다. 하나의 문서 또는 데이터가 암호화된 상태로 유통되므로 기밀성이 유지되고, 권한이 있는 사람만 열어볼 수 있도록 권한 관리가 가능해 데이터 보안을 할 때 필수적으로 사용되는 솔루션이다. 최근 보안 요구가 대폭 강화된 개인데이터를 사용하고 취급할 경우 개인정보보호법을 준수해야 하므로, 법규 준수 상황을 처리하고 확인시켜 주는 기능까지 포함하는 솔루션이 있다.

실제 데이터를 사용하는 환경에서는 이러한 데이터 보안 솔루션들을 어떻게 적용해 사용하느냐가 더 중요한 문제다. 여기서는 다양한 종류의 데이터가 다양한 곳에 저장되고, 또한 다양한 방법으로 유통되는 환경에서 과연 어떻게 적절하게 데이터 보안이 가능할까 하는 측면에서 기술해 보려 한다.

2. 데이터 보안 프레임워크

빠르게 변화하는 IT 환경과 지능화·고도화되는 사이버위협에 대응하기 위해서는 조직에서 전체적인 보안 전략을 수립해야 한다. 많은 기업들이 다양한 보안 솔루션들을 도입해 사용하고 있지만, 개별 솔루션들의 정책과 관리가 통합되지 않고 각자 동작하면서 보안의 빈틈들이 발생하게 된다. 이를 방지하기 위해서는 단일 보안 솔루션에 집중할 것이 아니라, 기업 및 기관 내 필요한 보안 전략을 검토한 후 필요한 솔루션들을 도입하되 상호 연계돼 일관된 정책을 적용할 수 있도록 통합 관리 방안을 마련해야 할 것이다. 이와 함께 보안 사고를 미리 예측해 보안 위협을 최소화할 수 있는 리스크 관리 방안도 함께 마련하는 등 큰 그림의 보안 전략을 고민해야 한다.

변화하는 IT 환경에 최적화된 데이터 보안 프레임워크를 구현해야 하는데, 이는 데이터 자체에 대한 보안을 중심으로 예외 상황을 통해 발생할 수 있는 정보유출의 빈틈을 방지하고 정보유출 리스크를 함께 관리할 수 있어야 한다. 이를 위해서는 데이터 중심의 보안, 사용자 중심의 정책, 멀티계층의 접근을 통해 더 탄

* 필자: 안혜연(파수닷컴 부사장)

탄하고 빈틈없는 보안을 제공해야 한다.

구체적인 솔루션으로 설명을 해 보자. 먼저 중요 데이터들을 보관하려면 우리가 어떤 중요한 정보를 어떤 형태로 어느 곳에 보관하고 있는지 파악해야 한다. 이러한 솔루션을 데이터 거버넌스 제품이라고 부른다. 주요 데이터를 파악하고 나면 어떤 데이터에 대해 어떤 정책과 룰을 가지고 관리할 것인지 정하고, 그 정책에 따라 앞에서 말한 문서보안 솔루션을 적용하면 된다. 이것으로 끝난 것이 아니다. 실제 이러한 데이터들의 사용 현황을 모니터링해 불법이라고 생각되는 형태를 찾아내는 것 또한 중요한 일이다. 이 때문에 보안 리스크 관리 솔루션에 대한 요구가 많아지고 있고, 실제 구축 사례가 늘고 있다.

3단계의 데이터 시큐리티 프레임워크가 적용되면 기업 및 기관이 보유한 모든 데이터를 탐지하고 분류해 사전에 설정된 정책에 따라 DRM 암호화 등의 적절한 통제가 가능하다. 또한 데이터 보안 시스템 로그를 비롯해 조직 내 구축된 다양한 시스템 로그들의 연관관계를 분석해 내부 데이터 유출 리스크를 함께 관리함으로써 어떠한 보안 위협에도 중요한 정보를 안전하게 보호할 수 있다.

3. 데이터 보안 솔루션

DRM 기반 문서보안 솔루션

기업 내의 중요 문서들을 파일 단위로 암호화, 사용자 또는 그룹 별로 중요 문서들에 대한 세분화한 권한 제어 및 사용내역 관리를 통해 더 완벽한 데이터 보안을 적용할 수 있도록 설계된 솔루션이다.

데이터 거버넌스 솔루션

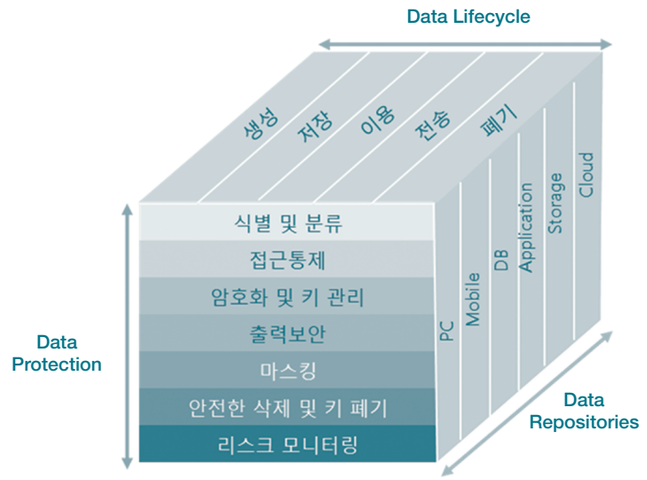
사용자의 PC 및 파일 서버 등에 저장·사용되고 있어 관리하기 어려운 비정형 데이터들을 자동으로 찾아 분류하고 중요 문서들을 지속적으로 보호·관리할 수 있다.

데이터 리스크 관리·모니터링 솔루션

사용자의 문서 사용 이력, 계정사용 로그, 출입 시스템 로그 등 여러 보안 관련 빅데이터를 기반으로 정상 및 비정상 패턴을 분석한다. 필요할 때 보안 담당자 및

LOB(Line Of Business) 관리자들의 확인을 통해 사전에 위협에 대응할 수 있도록 설계된 리스크 관리 솔루션이다.

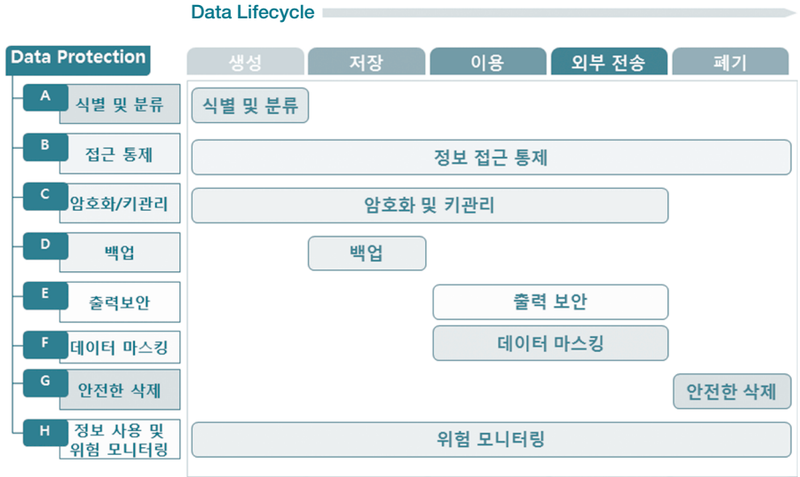
[그림 4-9-1] 데이터 보안을 위한 기본적인 프레임워크



실제 데이터는 PC, 모바일 디바이스, DB, 애플리케이션, 스토리지 등에 존재할 수 있다. 최근에는 클라우드에 존재하기도 한다. 이러한 데이터는 한 곳에 머무르는 것이 아니라 어떤 데이터든 생성에서부터 폐기까지 하나의 주기, 즉 생명주기를 갖고 있다. 이러한 모든 상황에 보안의 주요 기능들, 즉 기밀성, 무결성, 접근제어, 심지어 폐기까지 적절하게 작동을 해야 한다.

[그림 4-9-2]에서는 단계별로 보안 기능을 구현할 수 있는 솔루션의 적용 형태를 보여주고 있다. 예를 들면 데이터를 출력하거나 전송할 경우 보안 정책에 따라 출력을 제지하거나 중요 데이터에 대해 마스킹한 상태로 출력하는 등 보안 기능을 적용해야 한다.

[그림 4-9-2] 보안 정책과 데이터 생명주기에 따른 보안 솔루션 적용 형태



4. 개인정보 비식별화 솔루션

세기의 대결이라 일컬어지며 지난해 가장 큰 이슈가 됐던 이세돌 9단과 알파고의 바둑 경기를 기억하고 있을 것이다. 인공지능 컴퓨터인 알파고는 방대하게 수집된 빅데이터 분석을 기반으로 하고 있다. 이처럼 빅데이터 활용 사례는 기업은 물론 공공, 통신, 금융, 병원, 유통 등 다양한 산업 분야로 빠르게 확산되고 있다. 연구개발 또는 품질 개선을 위해 혹은 새로운 상품 개발과 맞춤형 마케팅을 위해 활용되기도 한다. IT 트렌드를 주도하고 있는 IoT, 핀테크, 인공지능 등의 공통점은 바로 방대하게 생성된 데이터를 핵심으로 한다. 최신 정보기술들이 빠르게 성장할수록 빅데이터 시장은 확대될 수밖에 없다.

하지만 빅데이터 산업의 활성화와 성장을 위해 해결해야 할 과제들이 있다. 바로 개인정보보호에 대한 이슈다. 최근 개인정보보호법, 정보통신망법, 신용정보보호법 등 관련 법규들은 비식별화라는 기술을 이용해 빅데이터 활용을 활성화하는 방향으로 추진되고 있다.

비식별화란 개인정보의 일부 또는 전부를 삭제하거나 대체함으로써, 다른 정보와 결합해도 특정 개인을 식별할 수 없도록 하는 기술적 조치다. 개인정보 비식별화 기술은 PPDM(Privacy Preservation in Data Mining)이란 이름으로 이미 2000년초부터 학문적인 연구가 활발히 이뤄지고 있는 분야이며, 이를 통해 다양한 비식별화 세부 방법론들이 학계에 제안되고 있다. 그 중에 가장 일반적으로 활용되는 방법론으로는 익명화나 암호화 등을 통한 가명처리 기법, 특정 범주나 범위의 값으로 일반화하는 범주화 기법, 임의 잡음 추가 등의 노이즈 기법들이 있다. 이러한 비식별화 기술들을 적절히 사용할 때 재식별화에 대한 걱정이나 개인정보 유출에 대한 우려없이 가치 있는 빅데이터 분석 결과를 도출해 낼 수 있다.

이미 업계에는 이러한 빅데이터 분석을 위한 다양한 개인정보 비식별화 솔루션들이 존재한다. 하지만 재식별 방지를 위해 지나친 비식별 기술을 적용할 경우 데이터의 활용도가 떨어질 수 있다. 데이터의 활용도를 높이기 위해 비식별 범위를 최소화하면 개인정보 재식별 위험이 높아질 우려가 있다. 아직까지는 적절한 분석 방안을 제공하는 솔루션이나 컨설팅 서비스가 많이 부족한 것이 현실이다.

●○ 빅데이터 산업의 활성화와 성장을 위해 개인정보보호법, 정보통신망법, 신용정보보호법 등 관련 법규들은 비식별화라는 기술을 이용하고 있다.

빅데이터를 활용한 시장을 만들고, 글로벌 시장에서 앞서 나가기 위해서는 개인정보 비식별화 솔루션을 제공하는 벤더와 빅데이터 활용 기업의 공동 노력이 필요하다. 비식별화 솔루션을 제공하는 벤더들은 개인정보는 안전하게 보호하면서 빅데이터 활용 가치는 높이는 분석 방안에 대해 연구개발과 투자를 지속해야 한다. 빅데이터를 활용하는 기업 및 기관들은 전문가의 도움을 통해 분석할 데이터를 파악하고 재식별 위험을 진단한 후 그에 따른 위험 완화 방안 등을 체계적으로 적용해 나가야 한다.

우리가 받아들이든 받아들이지 않든 정보기술이 비약적으로 발전하면서 이런 변화의 와중에 문제는 늘 생기게 마련이다. 개인정보보호와 빅데이터의 활용 확산에 대한 논란의 여지는 있지만, 두 가지 사안을 모두 만족시킬 수 있는 최선의 방안을 찾아 내기 위한 솔루션 벤더와 빅데이터 활용 업계의 지속적인 노력과 협력이 필요한 시점이다.

제10장

머신러닝 솔루션

머신러닝은 기존의 데이터 마이닝과 통계, 예측 분석과 함께 신경망 계열의 머신러닝인 딥러닝을 포괄하는 개념이다. 최근 우수한 최신 오픈소스를 얼마나 잘 반영하고 있는지, 기존의 우수한 상용 컴포넌트를 잘 활용할 수 있는지 여부가 머신러닝 솔루션을 평가하는 중요한 잣대가 되고 있다.

1. 머신러닝 솔루션 유형

머신러닝 솔루션은 머신러닝 알고리즘을 소스코드와 라이브러리로 제공하는 프레임워크 유형과 프레임워크에 사용자 편의성과 성능 향상 등을 위한 추가적인 프로그래밍을 덧붙인 패키지 또는 서비스 유형, 이들 패키지가 특정 기능으로 특화되거나 특정 산업에 특화된 유형이 있다.

[표 4-10-1] 머신러닝 솔루션 유형별 특징

솔루션 유형	특징	솔루션 예
프레임워크	주로 오픈소스로 공개	TensorFlow, CNTK, Spark MLlib, R, Caffe, Veles
패키지·서비스 솔루션	인프라, 하드웨어와 결합해 다양한 형태로 확산	Azure ML Studio, AML, SAS, Matlab
기능 특화 및 산업 특화	데이터, 지식과 결합	Watson

* 필자 : 안동혁(화수목 대표)

프레임워크 유형은 대부분 오픈소스가 주도하고 있다. 오픈소스의 특성상 갑자기 출현해 큰 인기를 얻다가 곧 사라지는 경우도 많다. 패키지, 서비스 유형에는 기존의 통계, 마이닝 솔루션과 클라우드 인프라를 기반으로 한 머신러닝 서비스, 그리고 오픈소스 프레임워크를 활용한 다양한 솔루션들이 있다. 기능과 산업 특화 유형에는 특정 분야의 지식과 결합한 전문 솔루션들이 있으며, 데이터와 결합하는 형태로는 솔루션 내에 데이터를 포함하거나 데이터를 학습한 결과물을 제공하는 솔루션들이 있다.

2. 머신러닝 프레임워크

머신러닝은 용어 자체도 불분명한 경계를 가지고 있다. 머신러닝은 기존의 데이터 마이닝과 통계, 예측 분석과 함께 최근의 신경망 계열의 머신러닝인 딥러닝을 포괄한다.

머신러닝 프레임워크는 머신러닝 알고리즘을 소스코드, 라이브러리로 제공하기 때문에 지원하는 알고리즘 종류에 따라 특성이 다르다. 단순하게는 전통적인 통계와 마이닝 알고리즘을 중심으로 딥러닝을 일부 지원하는 프레임워크와 딥러닝만을 중심으로 하는 프레임워크로 구분할 수 있다.

[표 4-10-2] 머신러닝 프레임워크 분류

대표적인 프레임워크	R, Spark MLlib, CNTK, TensorFlow, Caffe, Veles				
프레임워크의 지원 알고리즘 특성	전반적인 통계, 마이닝 포함 딥러닝 중심				
머신러닝 알고리즘	정보 기반 머신러닝 (Decision Tree)	오차 기반 머신러닝 (Logistic Regression, SVM)	유사성 기반 머신러닝 (k-NN, CF)	베이지안 기반 머신러닝 (HMM, Bayesian Network)	신경망 기반 머신러닝과 딥러닝 (CNN, RNN)

이들 프레임워크는 그 자체로 머신러닝 분석 솔루션의 엔진으로 이용되기도 하고, 간단한 사용자 인터페이스를 붙여 바로 솔루션으로 사용된다. 대표적인 머신러닝 프레임워크의 특징은 [표 4-10-3]과 같다.

[표 4-10-3] 머신러닝 프레임워크 특징

프레임워크	특징
R	• 가장 다양하고 가장 많이 사용되는 프레임워크 • 교육용이나 파일럿 형태로 쉽게 이용하는 데 유리함 • 반대로 성능상의 한계점으로, 딥러닝도 사용할 수는 있으나 실제 대용량 데이터로 이용하는 것은 어려움
Spark MLlib	• 인메모리 처리 프레임워크인 아파치 스파크의 일부이므로 데이터베이스와 스트림, 다른 데이터 소스에 대한 액세스 기능이 뛰어남 • 빠른 속도의 인메모리 데이터에 적용할 수 있는 알고리즘 라이브러리가 늘면서 인기 상승 • 딥러닝 라이브러리는 부족
CNTK	• 마이크로소프트의 CNT(Computational Network Toolkit) • 모델과 알고리즘, 각종 지원이 매우 다양하고 마이크로소프트의 애져(Azure)와 연계해 확장 • 마이크로소프트는 Azure에서 CNTK와 GPU 클러스터를 함께 사용함으로써 코타나(Cortana)의 음성 인식 학습 속도가 높아졌다고 밝힘
TensorFlow	• 구글에서 공개해 전 세계적인 주목을 받은 프레임워크 • 배우기가 조금 어렵지만 성능과 확장성이 좋음 • 딥러닝에서 많이 사용하는 다양한 모델과 알고리즘이 들어 있으며 GPU나 구글 TPU를 장착한 하드웨어에서 탁월한 성능을 보임
Caffe	• 이미지 분류를 위한 CNN으로 유명 • 최근에는 개발이 정체되면서 다른 프레임워크에 밀려나는 분위기
Veles	• 삼성전자의 러시아 연구소에서 개발해 공개한 오픈소스 • 딥러닝 애플리케이션을 위한 분산형 플랫폼으로 학습된 모델을 바로 사용할 수 있는 장점이 있으며, 추가적인 개발 진행 여부는 지켜봐야 함

3. 머신러닝 패키지와 서비스

머신러닝 패키지와 서비스는 머신러닝 알고리즘 외에도 데이터 전처리, 머신러닝 모형의 성과 측정, 프로그래밍 없이 클릭만으로 이용하게 하는 사용자 인터페이스를 제공한다. 다양한 패키지와 서비스들이 있지만, 머신러닝이 수행되는 가장 이상적인 환경이 클라우드라는 것과 향후 확장성을 고려할 때 클라우드 기반의 머신러닝 서비스에 주목해야 한다. Azure ML Studio, AML(Amazon Machine Learning)이 대표적이다. 클라우드 인프라를 갖고 있지 못한 구글이 텐서플로(TensorFLow)를 오픈소스 프레임워크로 공개한 반면, 아마존과 마이크로소프트는 머신러닝용 API를 클라우드와 결합해 제공한다.

[표 4-10-4] 클라우드 기반 서비스의 특징

클라우드 기반 서비스	특징
Azure ML Studio	• Microsoft Azure ML Studio는 마이크로소프트 클라우드 플랫폼인 애져에서 월, 시간 단위로 머신러닝 서비스를 이용할 수 있게 함 • Azure ML Studio로 모델을 생성하고 학습한 후 다른 서비스가 사용할 수 있는 API로 바꿀 수 있으며, 제 3자가 제공하는 다양한 알고리즘도 사용 가능
AML	• AML은 아마존 클라우드에서의 머신러닝 서비스로 아마존 S3, Redshift, RDS 등에 저장된 데이터에 연결해 머신러닝 분석 모델을 구축하도록 함 • 아마존 데이터의 의존성이 심하고, 아마존 플랫폼 밖에서의 모델 활용이 제한적임 • 딥러닝 프레임워크로는 MX넷을 이용

머신러닝 솔루션을 제공하는 국내 기업으로는 솔리드웨어, 어니컴, 엑셈, 위세아이텍, 이씨마이너, 화수목 등이 있다. 자체 개발한 알고리즘도 일부 있으나, 대부분 오픈소스 프레임워크를 이용해 개발했다. 오픈소스를 이용하지 않고 자체 개발하는 것은 비효율적이며, 기능면에서나 성능면에서 뛰어나지 않다. 머신러닝 솔루션은 우수한 오픈소스를 얼마나 잘 반영하느냐, 기존의 우수한 상용 컴포넌트를 얼마나 잘 활용할 수 있는지가 중요한 평가 요소가 되고 있다.

머신러닝은 클라우드 인프라와 결합해 활용되기도 하지만, 각종 하드웨어와 결합해 새로운 서비스로 탈바꿈하기도 한다. 스마트폰이나 가전제품의 음성 비서에 적용되는 머신러닝 솔루션은 해당 영역의 프레임워크 혹은 완전히 특화된 서비스로 구현되고 있다. 많은 기업과 연구소, 대학 등에서 자율주행을 위한 머신러닝 솔루션, 드론이나 내비게이션을 위한 머신러닝 솔루션 등을 경쟁적으로 개발 중이다. 국내에서도 특화된 서비스로 응용이 가능한 음성인식 기술이 개발돼 API 형태의 솔루션으로 제공되고 있다.

4. 기능 특화 및 산업 특화 머신러닝 솔루션

IBM 왓슨(Watson)이 대표적인 특화 솔루션이다. IBM 왓슨 개발자 클라우드에서 20종이 넘는 API를 제공하고 있으며, 개발자는 이 API를 이용해 원하는 응용 기능을 구현할 수 있다. IBM 왓슨은 API를 제공할 뿐 데이터를 제공하는 것은 아니다. 데이터 학습이 중요한 머신러닝 서비스를 구현하기 위해서는 충분한 데이터

를 갖고 있어야 한다. 또한 이를 제대로 이용하려면 많은 자원, 즉 IBM 클라우드 인프라 비용과 API 사용 비용이 소요되기 때문에 가격이 올라갈 수 있다. IBM 왓슨은 법률·의료 부문의 적용 사례가 많은데, 최근에는 특히 한국과 중국의 병원에서 많이 도입하고 있다. 충분한 데이터 확보를 통해 효과적인 서비스를 구현하기보다는 홍보용으로 끝날 수도 있다는 지적이 있다.

마이크로소프트 애져에서도 자사 머신러닝 솔루션을 특화해 서비스로 제공한다. 머신러닝을 통한 장비의 장애 예측과 유지관리 기능을 사물인터넷 모니터링과 결합한 Azure IoT 서비스가 대표적이다.

산업 관점에서는 전 세계적으로 보안과 금융 분야 이상 탐지에 특화된 머신러닝 솔루션이 다수 출시되고 있고 경쟁도 치열하다. 해당 분야의 기존 솔루션 관점에서는 본래의 솔루션 기능에 머신러닝 기법을 활용한 예측 기능을 추가하기도 하고, 별도의 머신러닝 솔루션을 통합해 해당 분야에 독자적인 영역을 구축하기도 한다. 팍사타(Paxata)는 데이터 관리에 머신러닝 기법을 적용해 데이터 레이크 제품으로 시장에서 포지셔닝하고 있다. 머신 데이터를 수집하고 관리하는 스플렁크는 머신러닝 기반의 보안 업체인 캐스피다를 인수한 후 기존 솔루션을 머신러닝 기반으로 업데이트하고 있다.

국내에서도 추천과 이상 탐지에 특화된 위세아이텍, 금융 부문의 고객 세분화에 특화된 투이컨설팅의 솔루션 등이 있다.

5. 향후 전망

지금까지는 사용자가 보유한 데이터를 학습할 수 있게 하는 머신러닝 솔루션이나 초기 단계 학습만 돼 있어서 실제 활용을 위해서는 별도의 데이터 학습이 필요한 솔루션들이 대부분이었다. 최근에는 데이터를 갖춰 놓고 목적에 따라 데이터를 골라 학습하도록 하는 데이터 서비스를 제공하거나, 데이터를 학습한 최종 결과물을 제공하고 출력단만 조정해 바로 활용 가능한 방식이 제시되고 있다.

2016 데이터 솔루션 이노베이터상 수상자

“인공지능이 솔루션 시장의 향배 결정한다”

최용준 | 위세아이텍 상무

“인공지능과 빅데이터가 데이터 솔루션 분야 변화의 기폭제가 될 것이다. 정보시스템이 RDB에 집중이 되어 있다가 ‘빗장이 풀리듯이’ 새로운 변화가 일고 있다. 품질 관리 분야에서도 RDB로 대표되는 정형화된 데이터 품질 관리에서 빅데이터 품질 관리로 넘어가면서 인공지능을 활용하게 될 전망이다.”

솔루션 이노베이터로 선정되었는데 소감 한마디 해달라.

IT 업계에 입문한 지 17년 되었다. 그동안 개발자로 일하다 관리·기획 파트로 넘어와 현재는 데이터 품질 관리 파트 본부장을 맡고 있다. 데이터 솔루션 업계에 입문한 지 13년 만에 이런 큰 상을 받아 그간 해왔던 일에 보람을 느낀다. 솔루션에 인공지능 기능을 심는 프로젝트 진행 중에 상을 받게 되어 ‘앞으로 나아가는 이노베이터’가 되라는 격려로 받아들였다. 감사하다.

2017년 국내 데이터 솔루션 분야의 이슈라면.

단연 인공지능이다. 2017년을 인공지능 기술의 원년으로 보는 시각이 많다. 실제로 솔루션에 적용·출시된 것들도 많다. 예측이나 분석 분야에서는 인공지능이 대세가 됐고, 데이터 품질 분야도 마찬가지다. 보안 분야에도 인공지능을 이용해 이상 패턴을 탐지하는 솔루션이 나온 것으로 안다.

데이터 영역의 전문업체들에게 인공지능이 화두가 되고 있다는 얘기인데, 산업계에 어떤 변화가 올 것으로 예측하나.

데이터에 대한 시각이 예전과 달라지고 있다. 데이터가 중요하다는 것, 그리고 데이터로 부가가치를 창출할 수 있다는 것을 인정하고 있으며, 빅데이터 활용 사례도 늘어나고 있다. 물론 빅데이터는 수년 전에도 나온 개념이지만 요즘 들어 더 활발해졌다. 인공지능 알고리즘과 분석·예측 알고리즘이 포함된 것과 무관하지 않다. 이 시기가 아주 중요한 시기다. 2018~2019년이면 업계가 재편될 것이다.

데이터 품질 관리 분야에도 해당되는 얘기 아닌가.

이미지, 문서 등 분석 가능한 데이터가 늘어나고 있는 데에 비해 데이터 품질은 아직까지는 기존의 패러다임에 머물러 있는 것이 사실이다. 품질 측정이 되지 않은 다양한 유형의 데이터들이 여전히 많기 때문이다. 바람직한 것은 인공지능과 빅데이터가 변화의 기폭제가 될 것이라는 점이다. 정보시스템이 RDB에 집중되어 있다가 ‘빗장이 풀리듯이’ 새로운 변화가 일고 있다. 데이터가 정확하지 않으면 정확하지 못한 분석이 되어 결과에 대한 신뢰를 할 수 없게 된다. 데이터 품질을 점검하기 위해서는 품질 지표와 그에 맞는 기술적 방법이 필요하다. 하지만 이름 항목에 이름이 아닌 텍스트가 들어 있거나, 수치 데이터의 비교 검증 대상이 없거나, 비정형 문서와 DB화된 정보 또는 이미지와 그 내용 정보의 불일치 등을 점검할 수 있는 데이터 품질 지표가 없다는 것이 문제다. 기술적으로도 데이터 품질 관리 솔루션에는 접목되지 않았다. 기존 RDB의 데이터 품질 측정 방법으로는 이와 같은 사항을 점검할 수도 없다. 데이터 품질 관리에도 인공지능 기술을 활용해야 하는 이유다. RDB로 대표되는 정형화된 데이터 품질 관리에서 빅데이터 품질 관리로 넘어가면서 인공지능을 활용하게 될 전망이다.

품질과 분석 관련 이슈는 많은데 국산 솔루션 시장 규모가 작다. 무엇 때문이라고 보나.

분석이라고 하면 OLAP을 말하는 경우가 많은데 기존의 통계기법을 개선해 (향후에는) 예측이나 추천으로 시장이 형성될 것이다. 기술적인 과도기에 나타나는 일시적인 현상일 뿐이다. 2018년부터 품질과 분석 시장이 크게 성장할 것이다. 데이터 거버넌스 구축 또는 RDB 중심으로 형성되던 데이터 품질 관리 시장이 빅데이터 등 다양한 영역으로 확대될 것이다.

데이터산업계에 입문하고 싶어하는 후배에게 하고 싶은 말은.

데이터 분야라고 하면 분석을 먼저 떠올리는데, 생각을 바꿀 필요가 있다. 데이터 분야는 DB, 보안, 품질 분야 등 매우 광범위하다. 개발자, 아키텍트, 컨설턴트, 기획자 등 진출할 수 있는 직무도 다양하다. 스스로 자신에게 맞는 직무는 무엇인지 어떤 분야인지 깊이 탐구하길 바란다. 학교 졸업 후 현장에서 경험을 하며 쌓은 지식과 노하우가 많겠지만 적어도 남들보다는 뛰어난 기본 소양을 하나 이상 가지고 있다면 적응하는 데에 큰 도움이 될 것이다.

제 5 부

데이터 구축 및 컨설팅 동향

제1장

데이터 분석 컨설팅

빅데이터 분석 컨설팅은 기존의 시범사업 형태에서 분석 전략계획과 파일럿이 병행되는 형태로 확대되고 있으며, 향후에는 구현된 분석 모델을 재강화하는 관점에서 진행될 전망이다. 지속적인 실험과 실험환경에 대한 데이터 관리, 빅데이터 거버넌스 체계 확립, 분석과 데이터 지표(metric)의 자산화한 관리가 필요하다. 또한 분석 전략계획 수립은 빅데이터에서 한걸음 더 나아가 기업의 디지털 트랜스포메이션을 위한 관점으로 확장할 필요가 있다.

1. 분석 전략계획 수립 동향

가. 분석 컨설팅 접근 방법의 변화

2015년까지 국내에서 이루어진 대부분의(빅데이터) 분석 컨설팅은 특정 업무영역의 분석 주제를 선정해 파일럿 또는 시범사업 구축 형태로 진행돼 왔다. 대부분의 빅데이터 분석 컨설팅이 파일럿 구축 형태로 수행된 것은 몇 가지 이유에 기인한다.

첫째, 분석 컨설팅 성과에 대한 불확신 때문이다. 둘째는 차세대 시스템 구축 같은 대형 프로젝트를 지속적으로 진행함에 따라 비즈니스 현업의 분석 프로젝트 참여가 제한적이었던 점을 들 수 있다. 셋째는 그럼에도 불구하고 시장 전반에 불고 있는 빅데이터 트렌드와 정부의 적극적인 빅데이터 추진 정책이 진행되고 있다는 점이다.

그러나 2016년도를 기점으로 빅데이터 컨설팅 접근 방식에 변화가 생겼다. 기존에는 특정 업무를 대상으로 한 파일럿 구축 형태의 보텀업(Bottom Up) 접근 방식이 주류를 이뤘다. 하지만 2016년 이후 분석 전략계획 수립이라는 전사적인 분석 기획부터 출발하는 톱다운(Top Down) 접근 방식을 시도하는 기업이 늘기 시작했다. 파일럿이나 시범사업으로는 체계적이고 지속적인 분석 도입에 한계가 있다는 것을 체감했기 때문이다. 이에 따라 전사적으로 필요한 분석 기회 발굴과 구체화, 빅데이터 아키텍처에 대한 청사진 설계, 단계적 구축 방안 도입을 위한 로드맵 수립의 필요성이 제기됐다. 그러나 과거의 정보전략계획(ISP) 처럼 컨설팅 성격의 프로젝트로 진행하지 않고, 분석 전략계획 수립과 병행해 퀵윈(Quick Win) 분석 과제를 동시에 진행해 계획 수립과 분석 도입의 타당성 검증을 함께 진행하는 방식으로 컨설팅들이 진행되고 있다.

나. 분석 전략계획 수립의 목표

분석 전략계획 수립 컨설팅 진행을 통해 기업들이 얻고자 하는 것은 다음의 다섯 가지로 요약될 수 있다.

첫째, 전사 비즈니스 모델, 비즈니스 전략, 목표 달성 및 강화를 위한 기업의 핵심 의사결정 요소를 데이터 분석 기반으로 실행하고 핵심 분석기회를 발굴하기 위해서다. 분석 전략계획 수립은 과거의 정보계 전략에서 모든 통계성 분석을 도출하는 것과 달리 비즈니스 목표·모델 달성을 위한 핵심 분석 기회에 초점을 맞춘다. 둘째, 발굴된 핵심 분석 기회를 정의하고, 분석 질문을 도출해 구체화한다. 셋째, 분석 질문에 대한 대답을 얻기 위해 필요한 기업 내외부의 정형·비정형 데이터를 리스트업하고 해당 데이터의 가용 상태, 중요도, 수집 난이도 등을 평가한다. 넷째, 분석에 필요한 데이터를 수집·저장·처리·활용하기 위한 분석 아키텍처를 수립하는 일이다. 다섯, 도출된 분석의 비즈니스 관점의 중요도, 분석에 필요한 데이터의 활용가능 수준 등을 고려해 분석 컨설팅 수행을 위한 로드맵을 수립한다.

다. 분석 역량 내재화

최근에 진행되고 있는 빅데이터 분석 컨설팅의 주요 특징으로 분석 역량 내재화를 들 수 있다. 단순히 컨설팅·구축·운영 유지보수로 이어지는 과거의 방식에서 벗어나 분석 역량을 내재화 해 기업 스스로 분석을 지속적으로 실행하기 위한 프

* 필자: 김찬수(투이컨설팅 상무)

로그램을 컨설팅에서 요구한다. 즉 빅데이터 분석 컨설팅 기간 동안 분석에 대한 교육 프로그램 실행과 분석 과정에도 외부 수행사와 공동으로 수행하는 방식을 통해 내부 조직원의 분석역량을 강화하고 있다.

2. 분석 기법 적용 동향

가. 머신러닝 도입 일반화

알파고를 기점으로 머신러닝(딥러닝 포함)의 실효성에 대한 인식이 대폭 변화됨에 따라, 분석 구현 프로젝트에서 적용되는 알고리즘에 머신러닝 기법을 적용하는 것이 일반화 됐다. 현재 국내의 분석 컨설팅에서 머신러닝이 적용되는 유형은 세 가지 형태를 보인다.

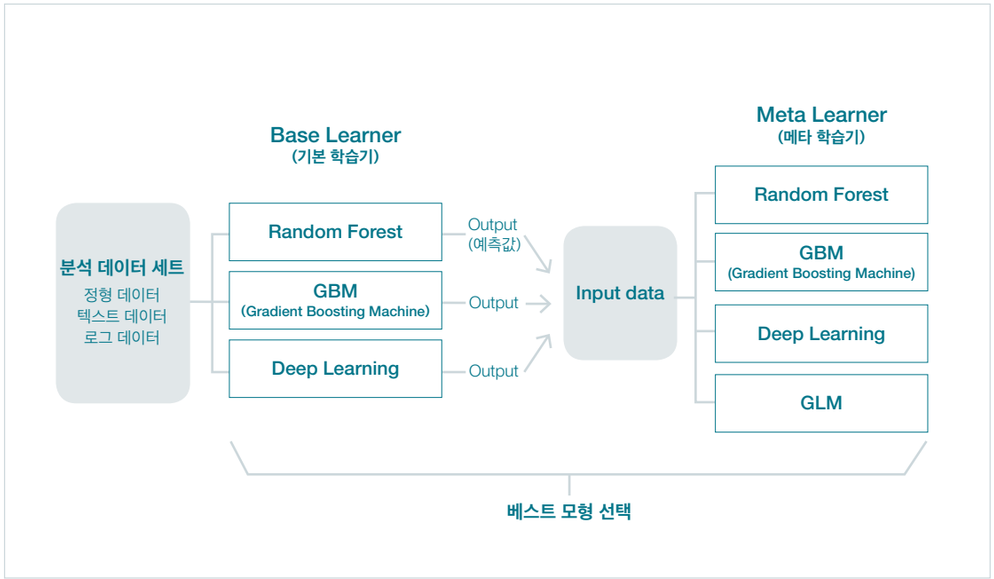
첫째, 기존에 분석이 제대로 적용돼 있지 않던 영역에 머신러닝을 신규로 도입하는 경우다. 이탈 예측 모델, 상품 추천 모델, 챗봇, 로보어드바이저 등을 예로 들 수 있다. 둘째, 기존의 전통적 분석 모델을 적용하던 영역에 머신러닝 모델과의 성능 비교를 통해 대체를 고려하는 경우다. 예를 들어 FDS(Fraud Detection System: 이상거래탐지시스템)에 적용되던 전통적인 회귀분석·룰 기반 스코어 모델과 신규로 머신러닝 모델을 적용해 스코어링을 통한 성능 비교를 통해 기존 모델의 대체 가능성을 검토하는 것이다. 특히 기존 전통적인 통계 모델을 머신러닝 모델로 대체를 하려는 주된 이유는 성능의 향상뿐 아니라, 비슷한 성능을 낸다면 주기적으로 모델을 갱신하는데 어려움을 겪었던 전통적 통계와 룰 방식에서 벗어나 운영의 편의성과 업데이트 용이성을 확보할 수 있기 때문이다. 셋째, 기존 전통적 분석 모델과 머신러닝 모델을 조합하는 경우다. 신계약 유지율 예측을 위해 전통적 스코어 모델과 머신러닝 스코어 모델 매트릭스를 조합하는 경우가 이에 해당된다.

나. 챔피언 모델, 앙상블(블렌딩) 모델 방식 적용

분석 알고리즘 구현에 있어 최근 진행되는 분석에서 특징적인 것은 챔피언 모델, 앙상블 모델 방식 적용이 보편화되고 있다는 점이다.

초기 분석 컨설팅에서는 한두 가지의 특정 분석기법을 적용해 분석을 수행하는 것이 일반적이었다. 그러나 고객의 수준과 눈높이가 높아지고, 분석가의 역량이 강화되면서 달라지기 시작했다. 여러 가지 분석기법을 적용해 경쟁시켜 가장 우수한 모델을 최종 모델로 선정하는 챔피언 모델 방식과 여러 분석기법을 혼합해 최적모델을 만들어 내는 앙상블(블렌딩) 모델 방식이 점점 보편화되고 있다.

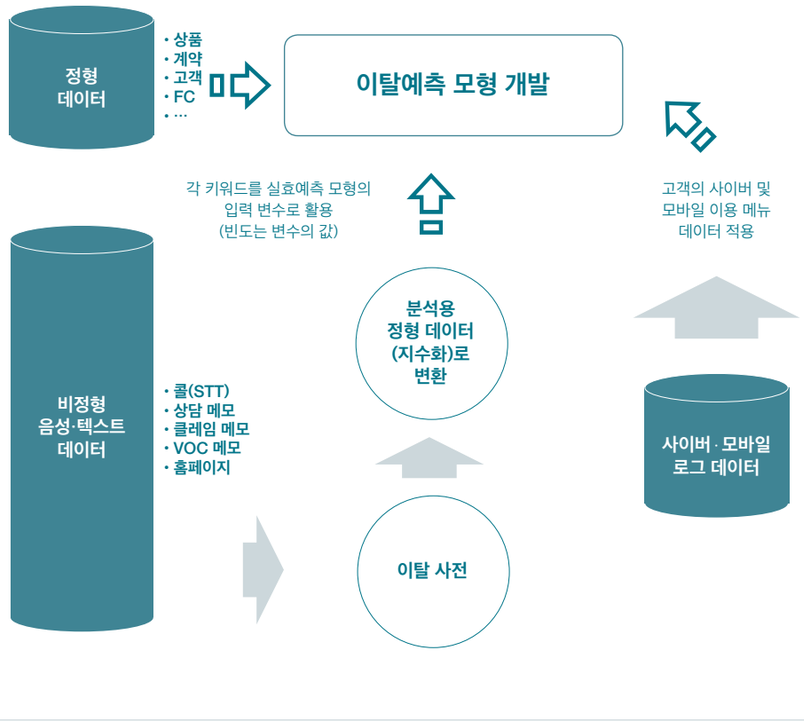
[그림 5-1-1] 앙상블 모델



다. 다양한 비정형 데이터의 분석 적용

분석에 사용되는 데이터 또한 점점 확장되고 있다. 콜센터 상담 녹취(음성) 데이터를 텍스트로 자동 전환하는 STT(Speech To Text) 엔진을 기반으로 고객의 소리를 지수화 해 고객의 상품 구매 니즈, 이탈 징후 등 각종 분석을 위한 주요 변수로 활용한다. 또한 텍스트 데이터뿐만 아니라 고객의 온라인 홈페이지 행동 로그의 패턴을 파악해 고객의 여정맵을 발굴하고 옴니채널에서의 이슈를 파악하는 CX(Customer Experience) 분석에도 활용한다. [그림 5-1-2]는 이탈예측 모형 분석에 사용되는 다양한 비정형 데이터에 대한 예시이다.

[그림 5-1-2] 이탈예측 모형에 사용되는 비정형 데이터



3. 향후 발전 방향

가. 재강화 학습 적용

지금까지 이루어지고 있는 거의 모든 분석 프로젝트는 해당 모델을 구현하는 시점에서의 최적 모델을 만드는데 초점을 맞추어 진행했다. 그러나 분석모델은 지속적으로 자동으로 강화될 수 있도록 만들어 주는 것이 향후의 중요한 과제다.

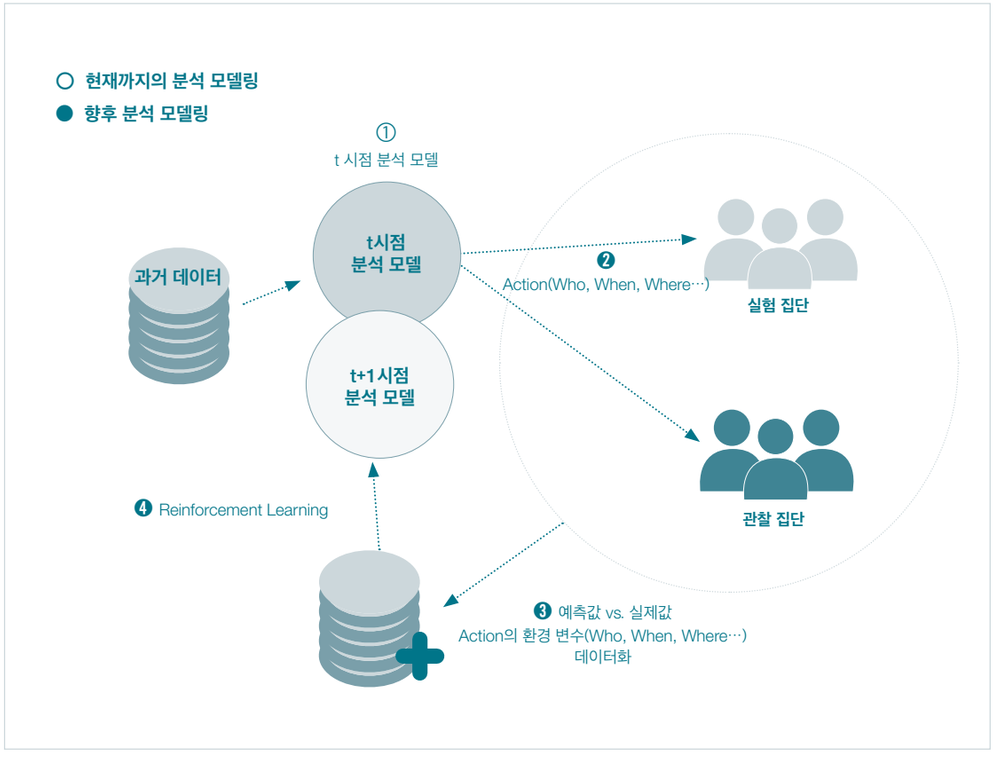
지난해 크게 이슈가 됐던 알파고는 재강화 학습(Reinforcement Learning)을 통해 분석 능력이 지속적으로 향상됐다. 알파고의 재강화 학습은 과거 기보를 기반으로 다양한 모델을 만들어 내고, 만들어 낸 모델끼리 시합을 벌인 결과를 반영해 모델에 보상과 패널티를 주는 방식이다. 즉 과거 기보를 학습해 만든 모델은 분석

모델을 만들어 낼 당시의 최적모델이라 할 수 있다. 그리고 해당 모델들끼리 시합을 하는 것은 모델에서 예측한 결과와 실제 발생하는 미래의 결과를 비교해 모델을 재강화하는 것이다.

예를 들면 이탈예측 모형을 과거에 발생했던 이탈 관련 여러 변수를 통해 현 시점에서 최적의 모델을 생성했다고 가정해보자. 해당 모델을 실제 업무에 적용시켜 이탈예측 모델에서 예측한 결과와 실제 결과와의 차이를 비교 분석해 모델에 다시 반영시키는 재강화 학습 과정이 필요하다. 또한 예측과 실제 값 사이에는 누가, 어떤 환경에서 개입했는가도 중요한 변수로 작용한다.

지금까지는 과거의 데이터를 기반으로 미래를 예측하는 현 시점의 최적 분석모델을 1회 생성하는데 초점을 맞추어 분석을 진행했다면, 향후에는 만들어진 분석모델을 재강화 학습을 통해 지속적으로 최적화 시켜나가기 위한 접근이 필요하다.

[그림 5-1-3] 재강화 학습 절차



나. 지속적 실험을 위한 준비

분석→설계→구현→테스트→적용으로 완결짓는 것이 기존의 업무 시스템이나 정보계 개발 방식이었다. 그러나 데이터 분석은 지속적인 실험의 과정이다. 분석 알고리즘의 구현으로 한 번에 최적의 결과를 낼 수 없다. 구현된 분석 알고리즘을 지속적으로 실험하고 업그레이드해야 한다. 즉 기업 내부의 분석 조직이 새로운 분석 알고리즘과 기존 분석 알고리즘을 지속적으로 만들고 실험할 수 있는 인프라와 조직이 필요하다.

여기서 실험할 수 있는 분석 인프라는 샌드박스를 의미한다. 기업 내외부의 다양한 데이터를 수집하고 조합해 새로운 알고리즘을 실험해 보고 기존에 만들어진 알고리즘을 보완해 볼 수 있는 안전한 실험 인프라 환경을 구축해야 한다. 또한 분석을 고도화 해나가기 위해서는 분석 전략을 통해 도출된 각 분석에 필요한 데이터를 단계적인 수집과 확장을 통해 분석 품질을 지속적으로 높이는 노력이 필요하다.

다. 빅데이터 거버넌스 체계 확립

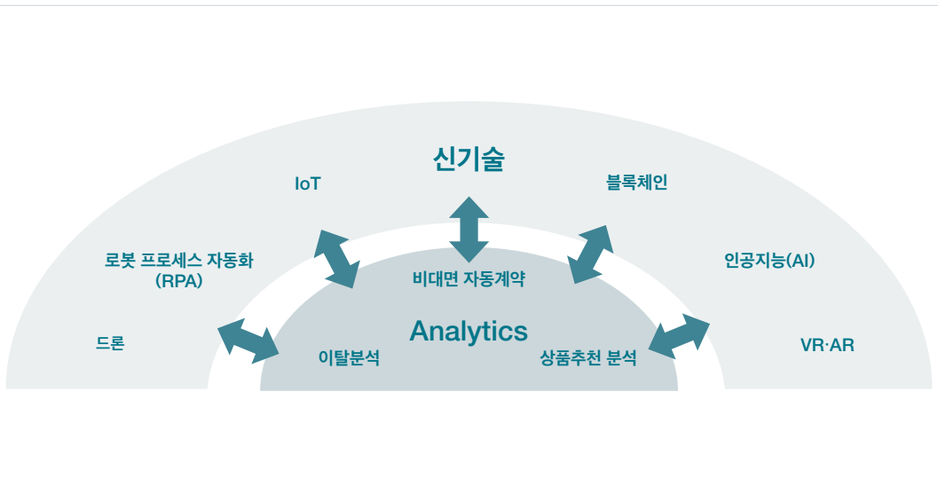
빅데이터 분석의 도입으로 기존의 거버넌스 영역에서 추가적인 고려사항이 발생하고 있다. 비정형 데이터, 외부 수집 데이터 등에 대한 거버넌스 정책과 분석을 지속적으로 수행할 수 있는 분석 조직에 대한 거버넌스 정책이 필요하다.

빅데이터를 분석에 적용하고 업무에 활용하기 위한 프로세스에 대한 거버넌스 체계 또한 필요하다. 이상의 관점에서 기존 거버넌스 정책을 확장하고 보완하는 활동이 요구되고 있다.

라. 분석의 자산화 관리

기존에 파일럿 형태 또는 다양한 부서에서 각각 분석 컨설팅을 수행하면서 분석 알고리즘, 분석 시스템의 파편화가 이미 발생하기 시작했다. 분석 알고리즘은 기업의 비즈니스 모델·전략 수행을 강화하고 가능하게 하는 디지털 자산이다. 전사 관점에서 분석 알고리즘을 체계적으로 생성하고 상호 관계를 관리하며 분석 요소·분석지표들을 재사용할 수 있도록 분석팩토리를 통한 분석 라이브러리 관리체계를 준비해야 한다.

[그림 5-1-4] 신기술과 빅데이터 분석과의 상호작용 관계



마. 디지털 트랜스포메이션과 맞물린 분석 전략계획 수립

최근에 제4차산업혁명과 온라인, 비대면 채널로의 전환에 따른 대응이 화두가 되고 있다. 이에 따라 핀테크, 신기술 도입, 스마트 팩토리 등을 기반으로 비즈니스 패러다임을 변화시키는 디지털 트랜스포메이션이 빅데이터에 이어 기업들의 새로운 주요 관심사로 떠오르고 있다.

그러나 빅데이터 분석과 디지털 트랜스포메이션은 다른 얘기가 아니다. 핀테크를 위한 금융기관의 오픈 플랫폼 콘텐츠 앱은 빅데이터 분석을 기반으로 생성될 수 있다. 또한 다양한 신기술 도입(드론, VR·AR 등)에 따라 기존보다 더 다양하고 방대한 데이터가 발생되고 이를 분석해 활용할 수 있는 기회가 존재한다.

신기술의 도입은 새롭고 다양한 데이터의 발생 측면뿐만 아니라, 신기술(챗봇, AI 등)이 작동하기 위해서는 빅데이터 분석이 기반이 돼야 한다. 그러므로 기업에서는 디지털 트랜스포메이션 전략 또는 신기술 도입을 통한 업무 혁신을 추진하기 위해서 빅데이터 분석 전략계획과 통합된 계획을 수립해야 한다.

제2장

데이터 품질 컨설팅

몇 년 전부터 금융권을 중심으로 개인데이터 보호 중심의 데이터 품질 컨설팅이 꾸준히 진행돼 왔다. 요즘 화두로 등장한 제4차산업혁명 준비 차원에서 공공 부문에서도 데이터 품질에 대한 관심이 커지고 있다. 공공데이터 품질 컨설팅은 공유·개방 데이터 개방 전략 수립, 개방 데이터 세트 발굴 등을 중심으로 전개되고 있다. 병무청의 병무행정 공공데이터의 품질 관리 사례를 소개한다.

1. 개요

데이터 품질 컨설팅은 데이터 구조 품질 컨설팅과 데이터 값 컨설팅으로 구분된다. 데이터 구조 품질 컨설팅은 데이터 모델 현행화, 데이터 표준화, 데이터 표준 관리체계 수립 등의 영역으로 나눌 수 있다. 데이터 값 품질 컨설팅 영역은 데이터 품질 진단 개선, 데이터 품질 관리 지침 수립, 품질 관리 체계 수립 등의 영역이 있다.

양쪽 영역을 모두 포함한 영역으로는 데이터 관리 프로세스 및 조직 수립, 데이터 개선 로드맵 수립 등을 들 수 있다. 몇 년 전부터 금융권을 중심으로 확대되기 시작한 개인데이터 보호에 의한 보안 컨설팅 영역도 최근까지 꾸준히 진행되고 있는 컨설팅 영역의 한 분야이다. 최근에는 공공데이터 공유·개방과 관련된 컨설팅도 꾸준히 성장 중이다. 이에 따라 공유·개방 데이터 개방 전략 수립, 개방 데이터 세트 발굴 등 이전에 존재하지 않았던 새로운 영역의 데이터 품질 컨설팅 영역이 탄생됐다.

* 필자: 오경조(지티원 상무)

2. 데이터 품질 관리 동향

가. 공공데이터 측면

정부의 적극적인 공공데이터 개방 및 확대, 데이터 품질 관리 지원을 통해 공공데이터 개방의 양적인 측면은 단기간에 달성이 되었으나 질적인 측면의 공공데이터 개방 및 확대는 아직 미흡한 측면이 많다. 특히 민간에서 활용할 수 있는 양질의 공공데이터는 많이 부족한 것이 사실이다. 향후 지속적으로 질적인 측면의 품질 확보 활동을 강화한다면 좀 더 활용 가치가 높은 양질의 공공데이터를 확보할 수 있을 것으로 예상된다. 최근 양적인 개방 데이터 세트를 발굴하기 보다는 공공·민간 모두 공유 데이터를 활용할 수 있는 의미 있는 데이터 세트를 발굴하기 위한 컨설팅 영역이 강화되고 있다. 데이터를 활용하는 민간기업을 적극적으로 조사해 새로운 융·복합 공유 발굴 및 공공기관이 앞으로 나아갈 개방 전략 수립 영역도 점차 시장이 확대되는 추세이다.

현재의 데이터 값 중심으로 이뤄지는 품질 확보 활동뿐만 아니라 이제는 전체 국가 공공데이터의 활용을 지원할 수 있는 구조적인 측면의 데이터 품질 관리 활동이 필요하다. 또한 단일의 공공기관에 한정되지 않고 관련 공공데이터의 공동 활용을 위해 추가로 통합 데이터 품질 관리 활동이 필요하다. 특히 제4차산업혁명과 관련해 앞으로 더 많은 데이터가 생산이 되며 데이터를 기반으로 모든 의사결정이 이루어지기 때문에 더욱 더 데이터에 대한 품질 활동이 중요한 활동으로 인식될 것으로 예상된다.

나. 민간데이터 측면

고품질 데이터를 확보하지 못하면 현재 진행형인 제4차산업혁명의 큰 흐름을 쫓아가지 못하기 때문에 기존 데이터 품질 관리 활동을 더욱 더 강화하며 특히 정형 데이터 품질 관리 활동뿐만 아니라 비정형 데이터까지 포함한 데이터 품질 관리 활동이 이루어질 것으로 예상된다. 특히 인공지능(AI)과 빅데이터 활용을 위한 음성 데이터, 텍스트 데이터 등의 영역이 1차적으로 확대된 데이터 품질 관리 영역이 될 것이다. 또한 여러 데이터의 융복합을 통해 새로운 비즈니스 기회를 창출하고, 기업이 보유한 양질의 데이터가 결국 4차산업의 핵심으로 떠오르면서 데이터 품질 확보를 위한 관련 인프라 확충, 관련 조직 및 프로세스 정비가 이루어 질 것

으로 예상된다.

과거 10년 동안 통신, 은행, 금융 등 특정 업무 영역에서만 데이터 품질 관리 시장이 존재했으나 앞으로는 전체 산업으로 확대될 전망이다. 기존에 추진됐던 여러 가지 사업들이 이제는 시행착오 단계를 지나 좀 더 현실적으로 실현 가능한 합리적인 방향으로 진행될 것이다. 특히 이전에 도입했던 초기의 데이터 품질 관련 컨설팅 및 솔루션(메타데이터 관리 솔루션, 데이터 품질 관리 솔루션) 등이 급변하는 현재 기업의 요구사항을 반영하지 못하고 있기 때문에 현 시점과 미래를 함께 포용할 수 있는 미래 지향적인 컨설팅 영역과 새로운 솔루션 도입 등이 추진될 것이다.

다. 공공데이터 품질 관리 사례

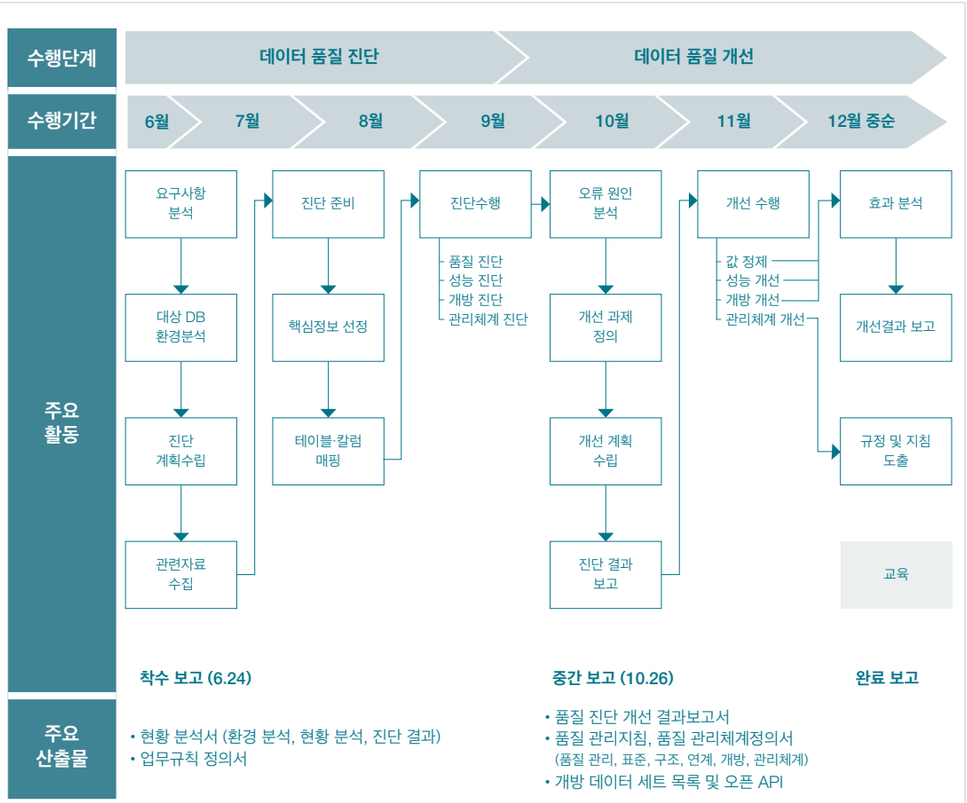
2016년도 공공데이터 품질 관리 지원사업의 여러 공공기관 중 품질 개선 효과가 가장 두드러졌던 병무청의 병무행정 DB의 데이터 품질 관리 사례를 소개한다. 병무청의 가장 중요한 DB인 병무행정 DB에 대해 데이터 품질 진단 및 정제, 성능 개선, 품질 관리 체계 수립, 개방 DB 구축 등을 약 6개월 간 진행했다.

사업 진행 전 일부 병역 데이터에 대해 오류 데이터가 발견됐고, 기 보유하고 있던 각종 관리규정과 지침이 문서적으로만 존재해 실질적인 적용이 힘든 상황이었다. 병역자료에 대한 인식이 부족해 수요자 관점의 활용가치가 높은 개방 데이터 세트 발굴이 어려운 상황이었다. 따라서 데이터 값 측면의 품질 진단 및 개선 활동, 기존에 병무청이 갖고 있던 각종 품질 관리 관련 지침 및 정의서를 최신화했다. 더불어 병무청의 조직과 프로세스에 맞게 현행화를 진행했으며 마지막으로 수요자 관점의 개방 데이터 세트 발굴 및 서비스 개발 등의 영역으로 진행했다.

본 사업을 통해 2016~2017년 사이의 병역 만료 자원에 대한 오류를 정비함으로써 데이터 개방을 위한 데이터 품질 개선 효과를 거뒀다. 더불어 업무규칙 55건을 반영해 공정하고 투명한 병역이행 업무 수행 체제를 구축해 지방 병무청을 통한 지속적인 데이터 품질 개선을 수행할 수 있게 됐다. 또 19개 프로세스에 대한 세부지침을 현행화함으로써 지속적인 품질 관리를 위한 체계 고도화를 이뤘다. 개발 수요자를 발굴하고 개방 종수 2종을 추가했으며, 데이터 세트 확대를 위한 총 18종(전환 16종, 신규 생성 2종)에 대해 개방 DB 구조를 개선했다.

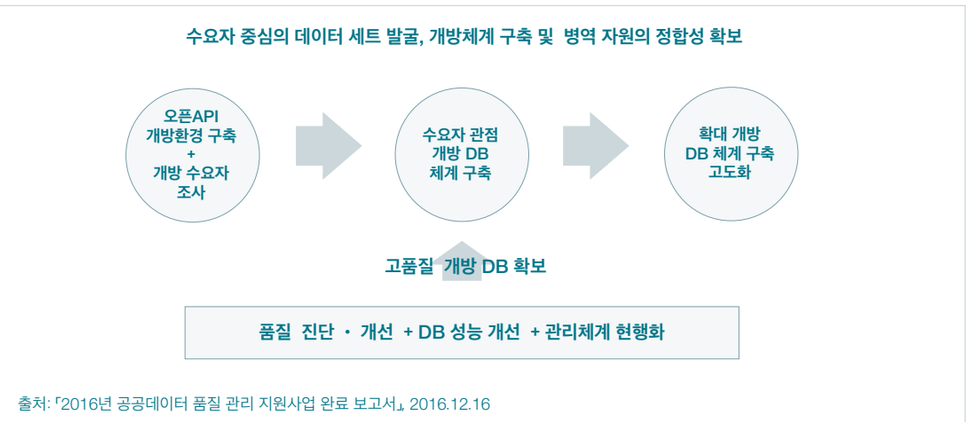
병무청은 공공데이터 품질 관리 지원사업을 통해 본청과 지방청 조직 및 역할을 재정의함으로써 대민 서비스 응용프로그램 35건의 응답속도 및 처리 성능

[그림 5-2-1] 병무청의 병무행정 데이터 품질 개선 프로젝트의 프로세스



출처: 「2016년 공공데이터 품질 관리 지원사업 완료 보고서」, 2016.12.16

[그림 5-2-2] 병무청의 병무행정 DB 중점 추진방향



출처: 「2016년 공공데이터 품질 관리 지원사업 완료 보고서」, 2016.12.16

을 7.4배 이상 끌어올리는 효과를 거뒀다.

더 정확한 고품질 데이터를 기반으로 공정한 병역 의무 부과, 다양한 병역 데이터의 공유와 개방을 통해 새로운 부가가치 창출을 유도했으며, 병무청 전청차원의 관리체계 현행화 및 고도화를 통해 지속적인 데이터 품질 관리 활동을 지원할 수 있는 기반을 마련했다. 사업 후에도 본청뿐만 아니라 지방 병무청까지 데이터 품질 관리 활동에 적극적으로 참여할 것으로 전망된다. 정량적인 기대 효과로는 데이터 정비 비용의 절감(연간 4,200만 원), 데이터 처리 성능 개선을 통해 병역자료 활용 시간의 절감(연간 3억 100만 원) 등이 예상된다. 병무청은 단순히 일회성으로 끝나는 데이터 품질 관리 프로젝트가 아니라 지속적인 데이터 품질 관리 활동을 내재화해 더 나은 병무행정 서비스를 위해 노력하고 있다.

3. 향후 전망

제4차산업혁명에 발 맞추어 나아가기 위해 빅데이터, 인공지능(AI), 딥러닝 등을 활용한 영역으로 관련 컨설팅 및 인프라가 확대될 전망이다. 또 다른 영역은 클라우드 환경이다. 아직 대다수 기업이 준비 단계에 머무르고 있으나 언제든지 폭발적인 시장으로 성장 가능성이 있다. 선제적으로 관련 기술들을 가지고 있는 기업들의 협업을 통해 상호 보완적인 컨설팅, 인프라 등 시장이 빠르게 형성될 것이다. 이런 데이터 품질 관련 활동을 통해 새로운 데이터 가치 발견, 새로운 데이터의 활용, 새로운 데이터의 융복합 등으로 인간이 생각하는 그 이상의 가치를 실현할 수 있게 될 전망이다.

제3장

비즈니스 인텔리전스 구축

고객이 원하는 시간에 전사의 모든 데이터를 한 곳에 모아 분석할 수 있는 정보 허브 역할을 하던 데이터 웨어하우스(DW)가 변화하고 있다. 막대한 양의 데이터에 대한 분석 요구가 증대되면서 거대한 양의 데이터를 저장·관리하기 위한 인프라로 빅데이터 기술이 발전하게 됐다. 비지니스에서 요구하는 데이터 분석 방향 역시 과거의 실적 분석 중심에서 분석 팩터(factor)를 포함한 예측 분석으로 발전하고 있으며, 이후에는 예측에서 한발 더 나아가 예측한 문제점에 대한 진단 및 처방이 가능한 수준으로 확대될 것이다.

1. 개요

초기 IT 환경에서는 업무 처리 중심으로 수행하는 기간계 시스템에서 비즈니스가 고도화되면서 데이터 및 트랜잭션이 증가함에 따라 기간계 시스템에 영향을 주지 않고 데이터를 분석해야 하는 제한 사항이 생겨날 수밖에 없었다. 이에 고객이 원하는 시간에 전사의 모든 데이터를 한 곳에 모아 분석할 수 있는 정보 허브 역할이 기업의 핵심 역량으로 자리잡으면서 데이터 웨어하우스(이하 DW)가 발전하게 됐다.

최근에는 SNS, 음성, 사물인터넷(IoT), 각종 장비의 로그 정보 등에서 발생하는 막대한 양의 데이터에 대한 분석 요구가 증대되는데, 이를 기존의 고사양 데이터베이스로 데이터를 저장하기에는 천문학적인 비용이 소요되기에 이르렀다. DW가 수용해야 하는 데이터 양이 수십 TB에서 수 PB까지 증가하면서 이를 저장·관리하기 위한 인프라로 새로운 빅데이터 기술이 발전하게 됐다.

기술적인 측면에서 보면 DW는 정형 통계, 데이터 마이닝(Data Mining) 또는

* 필자: 유수훈(비투엔 이사)

데이터 발굴 (Data Discovery), 통계를 통한 예측 등 분석 목적에 따라 고사양·고비용의 장비를 활용하는 DW 영역과 '데이터 발굴'을 목적으로 하는 빅데이터 영역이 공존할 것으로 예상된다. 또 DW 영역과 빅데이터 영역이 데이터를 서로 주고받는 하이브리드 DW 형태로 발전될 것이다. 이러한 경향은 이미 시장에서 수행되고 있는 여러 프로젝트를 통해 확인되고 있다.

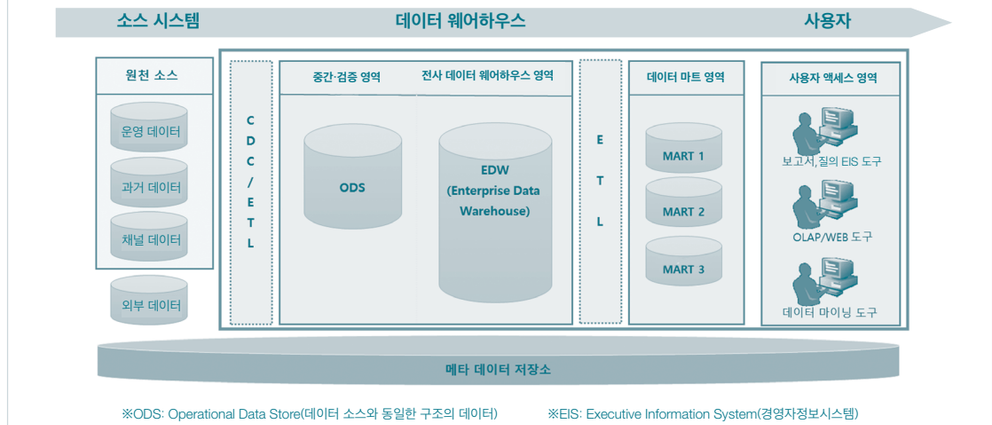
관리적인 측면에서 DW·BI(Business Intelligence)를 통해 데이터를 분석하는 과정을 요리하는 과정에 비유해 알아보자. 효율적인 식재료를 확보(수집)한 후 확보된 식재료를 잘 갖추어진 식료품 보관창고에 저장(보관·가공)해 두고 때에 따라 알맞은 재료를 레시피에 맞게 요리(분석)해 음식을 보기 좋게 전시한 후 고객이 맛있게 먹는 과정(활용)으로 볼 수 있다. 아무리 좋은 원천 데이터, 데이터 아키텍처, 기술 인프라를 갖추었다고 하더라도 원하는 데이터를 목적에 맞게 활용하지 못하면 소용이 없듯 고객이 원하는 정보를 적시에 제공하는 것이 더욱 중요하게 됐다. 목적지에 도달하기 위해 잘 닦인 도로와 같은 DW, 데이터 마트(Data Mart) 인프라를 잘 만드는 것도 중요하지만, 원하는 목적지를 잘 찾고 이를 활용하도록 결정하는 것은 최종적으로 사람이 하는 것이기에 데이터 거버넌스가 DW에서도 중요하다.

2. 기술 요소

DW의 아키텍처 측면에서는 전통적인 방식의 구조인 ODS(Operational Data Store), EDW(Enterprise Data Warehouse), MART라는 3 레이어에서 크게 벗어나지 않는 구조로 구축되고 있다. 최근에는 차세대 프로젝트를 통해 데이터 통합과 데이터 이력관리 등을 기간계에서 이미 적용하고 있으므로 ODS 영역의 역할이 축소되거나 최소한의 용도로만 사용되고 있는 경우가 있다. 심지어 기간계 DB를 아예 복제해 EDW로 사용하는 경우도 생겼다.

최근에는 기존 정보계 시스템에서는 수용하기 어려울 정도로 데이터 양이 증가하고 있기 때문에 기존 아키텍처 구조에서는 대용량 데이터를 저장할 수 없고, 저장하더라도 천문학적인 비용이 발생하게 됐다. 이에 하둡과 같은 빅데이터 플랫폼을 DW 구조와 연계해 관리하고 있는 추세다. 즉 전통적인 DW 구조와 빅

[그림 5-3-1] DW 아키텍처



※ODS: Operational Data Store(데이터 소스와 동일한 구조의 데이터) ※EIS: Executive Information System(경영자정보시스템)

데이터 플랫폼을 함께 관리하거나 EDW와 마트를 하둡으로 적용하고, 필요한 마트 또는 서머리 마트를 RDB 형태의 전통적인 방식의 마트에 적재한 후 기존 사용자에게 친숙한 OLAP 툴로 분석하는 구조를 가져가기도 한다.

데이터 모델링 기법으로는 마트 영역은 멀티 디멘셔널 모델링 기법에 의해 주로 스타 스키마 형태를 많이 사용해왔다. OLAP·BI 툴에서 성능적인 측면과 사용자 편의성을 고려해 반정규화를 고려한 팩트 테이블 생성을 통해 요구사항을 만족시키고 있다.

BI 툴은 기존 사용자가 사용하는 OLAP 분석 툴을 사용하는 영역과 통계 함수를 사용해 고급분석을 하는 영역으로 이원화해 사용됐다.

3. 구축 사례

가. 전통적인 방식의 정보계 구축

최근 통신, 금융, 공공 등 차세대 프로젝트의 정보계가 10TB에서 100TB로 DB 사이즈가 증가함에 따라 성능이 중요한 도입 요건이 되고 있다.

최근까지 많은 업계에서 상용화된 솔루션을 기반으로 Columnar DBMS 저

장 구조, 어플라이언스 장비, 인메모리 DB 등 세 가지 중요한 성능적인 기술 요소를 적용한 DBMS 기술구조로 차세대 정보계 프로젝트를 수행하는 추세다.

DBMS 저장 구조로 보면 마트와 같은 분석 영역의 경우 분석 항목별로 대규모 그룹핑이 필요하기에 컬럼별로 저장하고, 컬럼별로 인덱스를 설정해 빠른 조회를 보장하는 Columnar DBMS를 적용해 고객이 원하는 성능 요구사항을 만족시키고 있다. 수억 건에서 수십억 건의 데이터를 그룹핑 처리하기 위해 사이베이스(Sybase)나 버티카(Vertica) 등의 Columnar DBMS를 사용하는 추세인데, 한 예로 하나은행이나 삼성화재와 같은 대형 금융사에서 사이베이스를 기존 DW 구조로 사용하고 있다.

하드웨어 성능을 극대화할 수 있는 어플라이언스 장비를 사용하는 경우도 있다. 엑사데이터(Exadata), 네티자(Netezza) 등의 DBMS가 이에 해당하며 어플라이언스 장비 사용 시 초기에 quarter-rack*, half-rack**, full-rack*** 등 저장공간에 대한 적용 옵션을 선택해야 한다. 향후 저장공간이 부족해 확장할 경우 옵션별로 확장성의 제약이 있기 때문에 데이터 증가량을 고려해야 한다. OLTP 처리와 배치 처리가 모두 공존하는 Mixed Work Load 형태의 분석 환경에서는 오라클 엑사데이터(Exadata)를 통해 하드웨어에 메모리와 CPU를 갖추고 개별 셀(cell) 서버에서 연산을 통해 하드웨어 파워를 높여 성능을 유지하는 경우도 있다. 지금의 대용량 데이터 처리를 위한 시스템 구축 환경에서는 엑사데이터 도입이 일반화되고 있다.

최근에는 SSD와 같은 메모리 가격의 하락으로 SAP HANA와 같은 메모리 DB를 사용하는 경우도 있다. 아직은 많은 비용이 들지만 메모리 DB를 구축해 성능 요구사항을 만족시킬 수 있다. SAP HANA를 통해 정보계 시스템 구축을 진행 중인 기업은 삼성의료원과 삼성그룹의 금융사인 삼성생명, 삼성화재 등이 있다.

* 쿼터 랙(quarter-rack): 4대의 엑사데이터 스토리지 서버에서 1.5TB 엑사데이터 스마트 플래시 캐시, 96TB 디스크 스토리지, 48 CPU 코어 제공

** 하프 랙(half-rack): 9대의 엑사데이터 스토리지 서버에서 3.4TB 엑사데이터 스마트 플래시 캐시, 216TB 디스크 스토리지, 108 CPU 코어 제공

*** 풀 랙(full-rack): 18 대의 엑사데이터 스토리지 서버에서 6.75TB 엑사데이터 스마트 플래시 캐시, 432TB 디스크 스토리지, 216 CPU 코어 제공

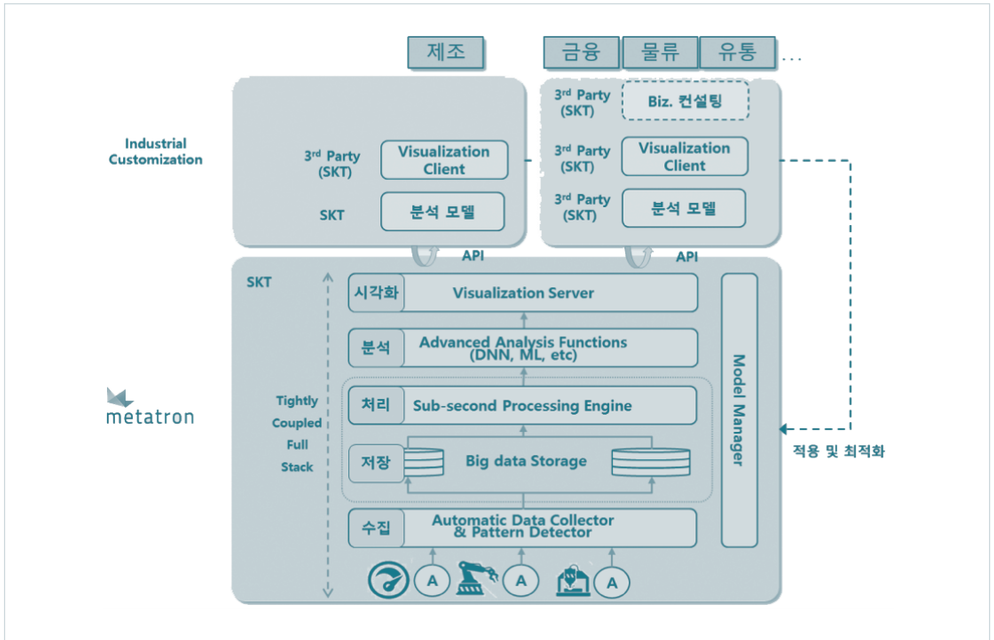
나. 기업의 전략적인 의사결정을 위한 분석

정보계의 범위가 기업내 데이터를 분석해 활용하려는 목적뿐만 아니라 경영에 도움을 줄 수 있는 다양한 정보를 활용해 인사이트를 도출하는 방향으로 확장되고 있다. 삼성화재의 경우 전통적인 분석 방법을 적용, 데이터 마트에서 실적 등 데이터 분석을 통해 비용 절감 및 효율화 분석 등을 수행했다. 이와 더불어 데이터를 경영에 활용하기 위해 MI(Market Intelligence) 마트를 구축해 내외부 데이터, 환경 데이터, 경쟁사 데이터, 공공 개방 데이터 등을 DW에 포함해 데이터를 통한 통찰력을 얻기 위해 다양한 마트를 구축·활용하고 있다. 변화하는 시장 환경에서 살아남기 위해 각 산업계에서는 다양한 분야의 데이터를 융합 분석해 경영에 활용하고 있고, 이러한 데이터를 바탕으로 고객의 경험을 지표화하는 고객경험정보(CEM) 시스템도 구축해 활용 중이다.

다. 빅데이터 아키텍처를 통한 분석계 구축

SK텔레콤은 고객들의 스마트폰을 통한 웹과 앱 접속 데이터가 하루에도 수 TB씩

[그림 5-3-2] SK텔레콤 메타트론 아키텍처



발생하고 있다. 이에 과금에 필요한 정보만 추출하고 향후 분석에 필요한 원천 데이터를 저장하지 못하고 버릴 수밖에 없었다.

현재는 기존에 버려야 했던 중요한 데이터를 필요한 기간 동안 하둡 환경에 저장·분석해 다양한 목적으로 분석하고 이를 활용하고 있다. 이러한 방대한 데이터를 활용하기 위해 먼저 개인데이터 비식별화 작업을 거쳐 데이터를 가공하고 상품 등 마스터 데이터 정보를 프로그램 작업을 통해 연결 데이터를 생성해 분석하고 있다. 과거에는 이용하지 못하던 대량의 데이터를 통계 분석 등에 대규모로 활용해 정확도를 향상하고, 다양한 분야에서 비즈니스 인사이트를 제안하고 있다.

또한 Data Discovery를 위한 도구인 SKT가 만든 '메타트론'은 하둡 환경에서 빅데이터의 모든 단계(수집-저장-처리-분석-시각화)에 최적화된 빅데이터 플랫폼으로, 대량 데이터를 대규모로 분석하는 데 최적화됐다. 메타트론 빅데이터 플랫폼을 활용해 네트워크 또는 반도체 생산 설비 등에서 발생하는 다양한 로그를 추출해 분석하고 이를 머신러닝, 딥러닝 기법을 활용해 통신장비의 예방 정비에도 적용하고 있다.

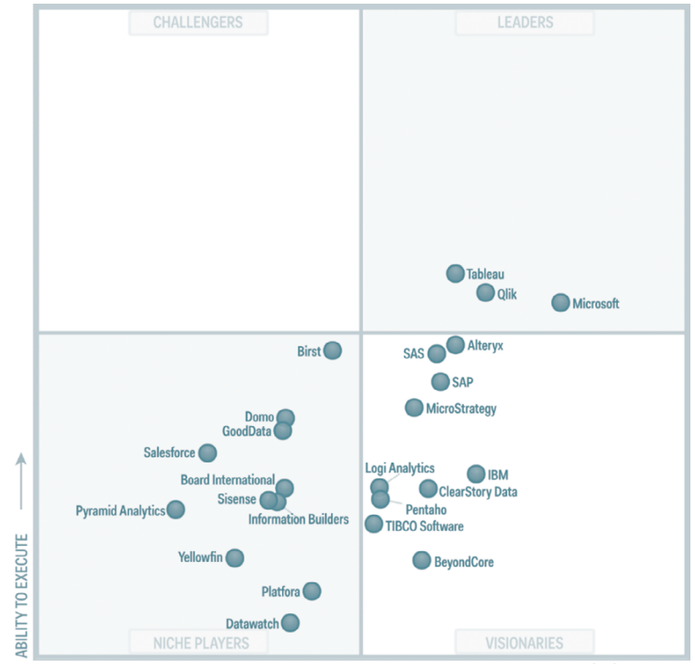
SK하이닉스에서는 다양한 공정에 사용하는 고가 장비에 대한 검사를 메타트론으로 수행 중이다. 반도체 장비의 수많은 공정 단계별 로그 데이터를 메타트론 빅데이터 플랫폼에 적재한 후 이를 분석해 고가 반도체 장비의 예방 정비에 활용하고 있다. 금융권에서도 빅데이터 플랫폼 구축에 나서고 있으며, 각종 분석을 위해 방대한 데이터 처리 장비와 도구 등을 구비하고 있다.

4. 향후 추이

홉스킨 박사는 "기존의 BI는 운전하면서 백미러를 보고 어디를 지나왔는지 뒤에 무엇이 있는지 확인하는 것과 같았다면, 지금의 BI는 앞을 향해 있으며 운전하면서 앞 유리를 통해 이 다음에 무슨 일이 일어날 것인지를 내다볼 수 있다"라고 비유한 바 있다.

BI 툴은 데이터 시각화, 빅데이터 지원, 비정형 모델 지원, 인메모리 분석, 예측분석 등을 중요하게 여기는 추세로 발전하고 있으며, 기존의 OLAP 분석과 통계함수를 사용해 고급 분석을 하는 영역으로 이원화해 발전하고 있다.

[그림 5-3-3]] Magic Quadrant Business Intelligence & Analytics



최근 국산 BI 툴들도 전통적인 DW뿐만 아니라 하둡 기반의 Hive, Tajo, Cloudera Impala, Spark 등을 접속해 데이터 분석이 가능하도록 적용하는 추세다.

비즈니스에서 요구하는 데이터 분석의 방향은 초기 과거의 실적 분석 위주에서 이제는 분석 팩터를 포함한 예측 분석으로 발전하고 있으며, 이후에는 예측에서 한발 더 나아가 예측한 문제점에 대한 진단 및 처방이 가능한 수준으로 확대될 것이다. 이를 위해서 머신러닝, 인공지능 알고리즘 등을 데이터 분석에 함께 사용할 것으로 보인다.

정보계 아키텍처 측면에서는 관계형 DBMS, 통합 DW, 마트, OLAP으로 대변되는 전통적인 DW 영역과 빅데이터, 데이터 레이크(Data Lake), 하둡, R 분석 툴 등으로 대변되는 데이터 디스커버리 영역으로 용도에 맞게 공존할 것이다.

제4장

빅데이터 구축

빅데이터 플랫폼 구축은 하둡에 데이터를 저장·관리하고, 맵리듀스로 추출한 통계 데이터를 다시 RDB로 저장하는 하이브리드 데이터 웨어하우스 구축 방법에서 벗어나고 있다. 하이브(Hive)와 같은 SQL on Hadoop 등의 분석 기술과 스파크(Spark) 기반의 실시간 처리, 클라우드 환경을 이용하는 쪽으로 바뀌고 있는 것이다. 빅데이터 분석 또한 비즈니스 인텔리전스(Business Intelligence, BI) 영역에서 벗어나 데이터 전처리와 인공지능 기술로 예측·분류해 분석 영역을 넓힌 비즈니스 애널리틱스(Business Analytics, BA)로 이전하고 있다. 데이터 분석 방법 측면에서는 오픈소스를 이용하거나 해외 분석 솔루션을 이용해 데이터에 대한 이해와 비즈니스 문제 해결에 더욱 집중하는 쪽으로 발전하고 있다. 한 온라인 미디어사의 빅데이터 구축 사례를 통해 최근의 빅데이터 분석 플랫폼 동향을 알아본다.

1. 빅데이터 시스템 구축

빅데이터 시스템의 축이 하드웨어 구축에서 분석 시각화와 문제 해결 쪽으로 옮겨가고 있다. 이 변화는 시스템 구축에도 영향을 끼치고 있다. 데이터를 수집·저장하는 것에서 벗어나 하둡 저장 시스템에서 스파크(Spark)를 통한 처리 속도 향상과 분석방법 개선으로 관심이 이동했다.

도입사들도 클라우드 서비스를 이용해 비즈니스에 더욱 집중하는 환경으로 관심이 이동하고 있다. OLTP 데이터 저장·분석과 빅데이터 분석 플랫폼도 클라우드 방식으로 바뀌고 있다. 즉 IaaS(Infrastructure as a Service) 방식으로 플랫폼을 구성하던 기존 방식에서 벗어나 클라우드 환경의 빅데이터 수집·처리 플랫폼으로 이전하고 있는 것이다.

빅데이터 분석 인프라 플랫폼은 전통적인 오픈소스를 사용하는 방식과 더불어 전문 운영인력 확보에 대한 부담 해소와 ROI 측면에서 맵알(MapR), 클라우데라와 같은 상용 패키지에 대한 관심도 올라가고 있다. OLTP 서비스가 클라우드

* 필자: 김상수(위세아이텍 연구소 센터장)

에서 이뤄지면서 데이터 이관 및 환경의 획일성, 비용, 리스크 관리 측면에서 클라우드 기반의 분석 인프라가 각광받고 있다.

중소 규모 비즈니스 환경에서는 하둡, 스파크 인프라를 활용하거나 클라우드에서 제공하는 스파크 가상화 서비스를 이용한다. 더불어 제플린 노트북(Zeppelin Notebook)과 같은 분석 환경을 구축해 데이터 분석부터 시각화에 이르기까지 통합 환경을 구성하고 있다. 하지만 이러한 인프라를 구축하려면 내부에 분석환경을 구성할 수 있는 전문인력이 필요하다. 사내에 인프라 구축 전담 인력이 없으면 AWS(Amazon Web Services) 같은 클라우드 서비스를 이용하면 된다. AWS 클라우드 분석 인프라는 데이터 저장소 영역과 구축 방법 및 범위에 따라 다음과 같이 크게 세 가지로 나뉜다.

[표 5-4-1] AWS 클라우드 분석 인프라 종류

서비스	제품 유형	설명
Amazon EMR	하둡	대량의 데이터를 신속하고 비용 효율적으로 처리할 수 있도록 관리형 하둡 프레임워크를 제공하며, Apache Spark, HBase, Presto, Flink와 같은 오픈소스 프레임워크를 실행할 수 있다.
Amazon Redshift	데이터 웨어하우스	속도가 빠른 페타바이트 규모의 완전 관리형 데이터 웨어하우스로, 간편하고 비용 효율적으로 모든 데이터를 기존 비즈니스 인텔리전스 도구를 사용해 분석할 수 있다.
Amazon Kinesis	스트리밍 데이터	AWS에서 스트리밍 데이터를 사용하는 가장 쉬운 방법으로, 분석과 실시간 데이터 수집을 지원한다.

AWS와 같은 클라우드 분석 플랫폼은 필요에 따라 단지 하둡 플랫폼만, 분산 데이터 웨어하우스 구축을 위한 ANSI SQL까지 동시에 사용할 수도 있다. 더불어 실시간 데이터 수집·저장하는 기능까지 제공하기 때문에 다양한 기능의 장점은 물론, 비용절감 효과까지 기대할 수 있다.

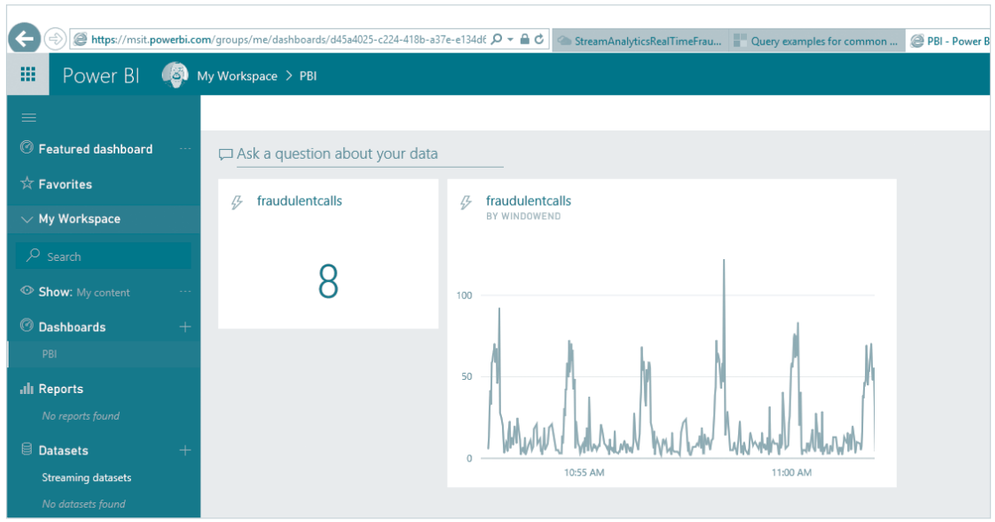
그렇더라도 데이터 이행이라는 어려움이 있다. 온프레미스(On-premise)로 구축했던 분석 인프라를 클라우드로 옮겨가려면 데이터 이행을 위한 시간과 비용이 드는 것이 사실이다. 또한 데이터 보안 측면에서 개방은 실제적으로 매우 어렵다. 이러한 문제점을 극복하기 위해 MS 애저(Azure)는 Azure Data Sync 기반의 데이터 동기화 기능을 하는 컴포넌트를 지원한다. OLTP에서 데이터를 생산·분석하는 하나의 프레임으로 구축하는 방법을 권장한다.

2. 빅데이터 분석 환경 구축

빅데이터 분석 방법 툴 시장은 현재 춘추전국시대를 맞고 있다. 앞서 소개한 오픈 소스 분석 툴인 제플린 노트북으로 분석하는 방법, 클라우드 데이터와 분석 툴을 결합한 MS Power BI, IBM Watson Analytics 등이 대표적인 제품이다.

[그림 5-4-1] 제플린 화면

[그림 5-4-2] 마이크로소프트 Power BI



[그림 5-4-3] IBM Watson Analytics

보편적으로 SQL, R을 활용하고 최근에는 파이썬(Python)이나 스칼라(Scala)를 이용해 분석 및 시각화를 자체적으로 구축하려고 한다. 이때는 시각화 및 분석 라이브러리가 필요하다. 이 환경에서 분석은 용이하지만 시각화는 시간이 많이 필요하기 때문이다. 그런 이유 때문에 마이크로소프트 Power BI와 같은 데이터와 시각화를 함께 지원하는 글로벌 벤더의 제품들이 속속 등장하고 있다.

상용 분석 솔루션 등장 이후 데이터 분석에 대한 가이드를 받으면서 편리한 분석 환경을 구현해 비즈니스에 집중할 수 있게 됐다. 분석 영역도 DBaaS(Database as a Service)로 DB와 분석 인프라를 클라우드에서 임대해 사용하는 것은 물론, 분석 시각화 BI 툴과 분석 방법까지 클라우드 서비스로 이동하려는 움직임이 포착된다.

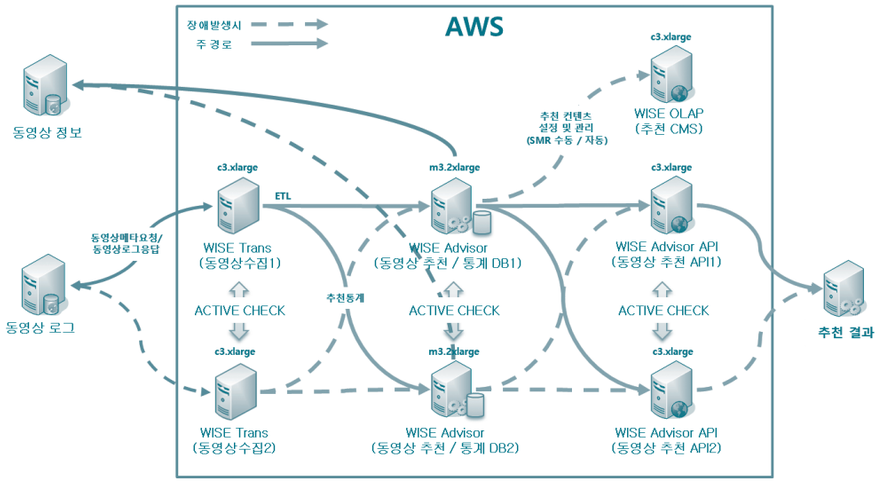
3. 온라인 미디어 콘텐츠 서비스 분석 사례

온라인 미디어 콘텐츠 서비스를 제공하는 'SMR(스마트미디어랩, Smart Media Representative)'은 동영상 및 광고와 관련해 일일 트랜잭션 2억 건에 이르는 빅데이터 기반의 추천 및 통계분석 시스템이 필요했다. 이를 위해 SMR은 클라우드 인프라 환경에서 'WISE Advisor BLU'를 적용해 빅데이터 플랫폼과 연계해 실시간 동영상 로그 처리와 맞춤형 추천 서비스를 제공하고 있다.

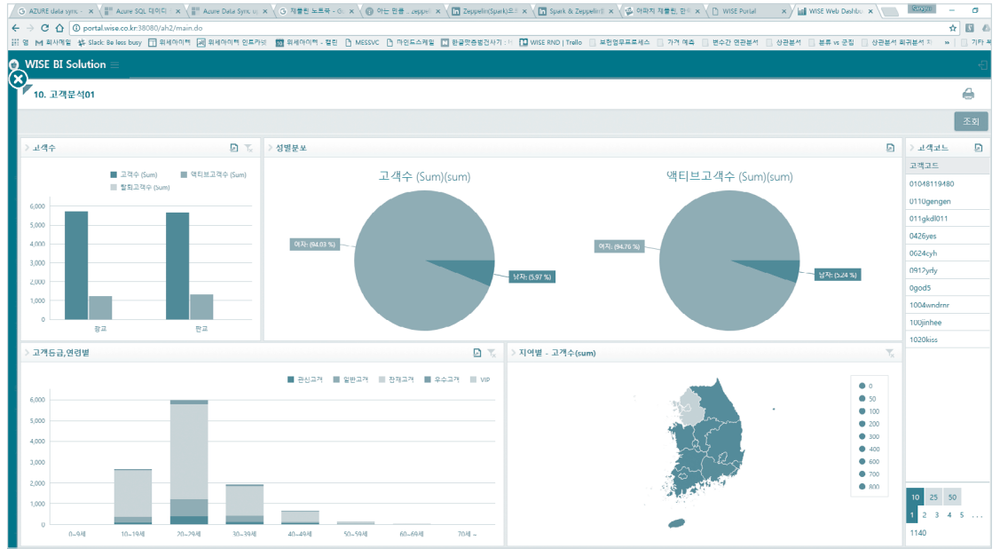
SMR은 클라우드에서 배치 처리를 해 데이터를 수집하고 통계 결과를 인메모리, 컬럼 저장 방식으로 RDB에 저장한다. 빅데이터 분석과 머신러닝으로 콘텐츠를 추천하고, R을 이용한 데이터 분석이 가능하게 구축했다.

구축 1년 후에는 10TB가 넘는 데이터가 축적돼 기존 저장·분석 방법으로 처리하기에는 한계점에 이르렀다. 이에 데이터를 클라우드에서 분산·저장하는 형태로 바꾸면서 여러 노드에 분산·복제·저장해 실시간 데이터를 수집·처리·저장하는 형태로 개선했다. 빅 테이블은 분산저장을 하고, 차원 테이블과 같은 분석 대상 영역은 복제해 빠르게 결과가 조인될 수 있도록 구성했다.

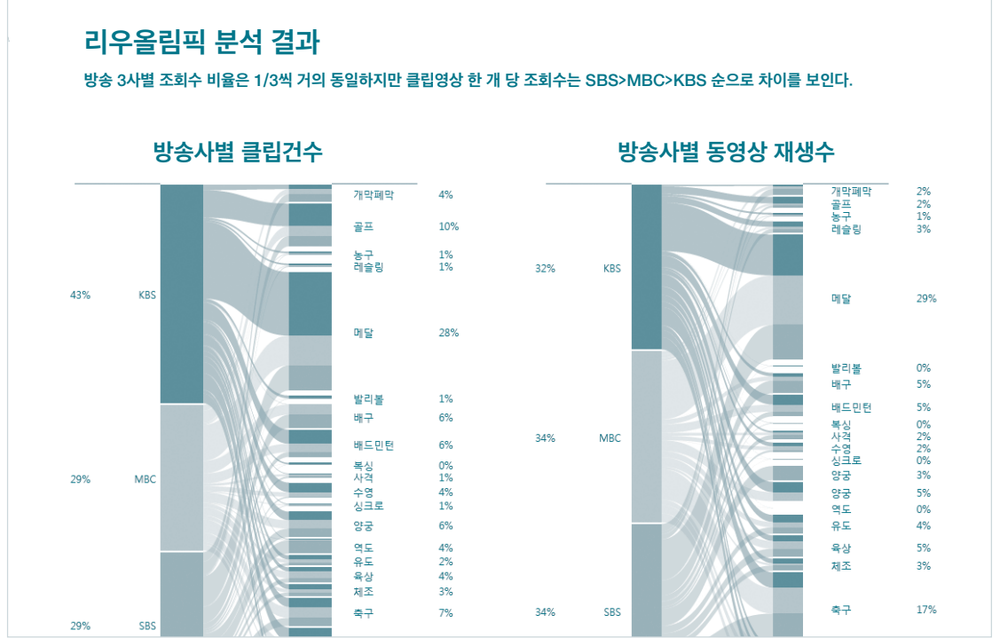
[그림 5-4-4] 클라우드 환경에서 WISE Advisor BLU 적용 사례



[그림 5-4-5] 분석 시각화



[그림 5-4-6] R 분석 결과의 시각화



[그림 5-4-7] 머신러닝을 이용한 콘텐츠 추천 결과 시각화

SMR은 향후 클라우드 자체에서 제공하는 AWS Redshift나 SoftLayer DashDB와 같은 클라우드 분석 DB로 전환해 데이터 증설과 분석 컴퓨팅을 효과적으로 사용할 수 있게 할 계획이다. 이를 통해 총소유비용(TCO) 절감을 위한 개선 작업을 계속해 진행하고 있다.

제5장

데이터 이행 구축

데이터 이행은 As-Is 시스템에서 To-Be 시스템으로 데이터를 이관하는 업무로 차세대 프로젝트에서 가장 중요한 부분 중 하나다. 데이터 양이 방대하고 매핑 조건이 복잡할 경우 오류 없이 새로운 시스템으로 이행하는 것은 결코 쉽지 않은 작업이다. 빅데이터 및 제4차산업혁명 시대를 맞아 변화의 기로에 선 국내 데이터 이행 현주소를 금융권을 중심으로 알아본다.

1. 개요

데이터 이행은 데이터 매핑 설계서를 바탕으로 데이터 추출, 매핑, 적재 프로그램을 작성하고, 단위 테스트, 통합 테스트, 이행 리허설 테스트를 거쳐 본 이행 작업을 수행하는 업무다.

업무팀이 작성하는 데이터 매핑 설계서는 이행 데이터의 품질을 좌우하며, 이행 데이터의 품질은 업무 프로그램의 품질에 영향을 주어 프로젝트의 성패를 좌우할 수 있다. 데이터 이행은 데이터 이행 파트와 업무팀, 인프라팀, 발주사, 수행사가 상호 유기적인 협조체계를 구축해 진행해야 성공할 수 있는 업무다.

데이터 이행 프로젝트 기간 중에 업무팀에서 테스트를 수행하기 위한 테스트 데이터를 제공하는데, 테스트 데이터의 품질은 업무 프로그램의 품질에 영향을 줄 수 있다. 따라서 데이터 이행은 테스트 단계부터 본 이행 오픈까지 프로젝트에 영향을 가장 많이 주는 업무다.

데이터 이행은 매핑 설계서 작성, 데이터 아키텍처 구성, 이행 지표 관리가 잘

* 필자: 이상일(LG히다찌 부장)

돼야 한다.

매핑 설계서는 As-Is 데이터와 To-Be 데이터에 대한 완벽한 이해를 기반으로 작성해야 하고, 데이터 아키텍처는 시스템, 데이터베이스, 데이터 용량을 파악해 구성해야 한다. 이행 지표는 매핑 설계 품질을 관리하는 매핑 지표, 이행 프로그램의 품질을 관리하는 이행 지표, 데이터 품질을 관리하는 오류지표를 작성해 관리한다.

데이터 이행 관리자 및 개발자는 데이터 모델, 매핑 설계서, 데이터베이스, 시스템, 네트워크에 대한 이해를 바탕으로 업무를 수행해야 한다.

2. 데이터 이행 현황

데이터 이행은 2000년대 초반부터 시작된 국내 은행권 차세대 프로젝트를 중심으로 발전해 카드, 보험 등 금융권 차세대 구축의 핵심 업무 중 하나다. 2000년대 초반부터 시작된 국내 은행의 1기 차세대는 완료됐으며, 포스트 차세대 구축 프로젝트를 완료한 은행도 있다.

대부분의 금융권 차세대 프로젝트는 연휴 기간(일부는 평일 또는 주말)에 데이터 이행 및 오픈을 하는데, 데이터 이행에 실패할 경우 짧게는 1주일에서 길게는 몇 달씩 프로젝트를 연장해야 한다.

메인프레임에서 오픈 시스템으로 전환되는 경우 메인프레임의 OIO(Open Infrastructure Offering) 계약 종료 시점에 오픈 일정을 수립한다. 데이터 이행에 실패할 경우 프로젝트 지연뿐만 아니라 OIO 연장 계약에 따른 비용을 추가로 지급해야 하는 상황이 발생할 수 있다. 아직까지 국내 은행 계정계 차세대 프로젝트는 오픈 일정을 연기한 경우는 없으나 데이터 및 업무 프로그램의 품질 저하로 오픈 이후 문제가 발생한 사례는 종종 있다. 데이터 이행은 이행 시간 내에 데이터 추출·매핑·적재·데이터 검증(이행 검증·논리 검증), 업무팀 화면 검증을 완료해야 하는데, 이행 시간은 데이터 이행 아키텍처와 매우 밀접한 관련이 있다.

데이터 이행 아키텍처는 데이터의 특성, 데이터 모델, 시스템, OS, 데이터베이스, 스토리지, 네트워크, ETL 툴, C 프로그램에 대한 이해를 바탕으로 구성해야 한다.

3. 데이터 추출 환경의 변화

데이터 이행은 As-Is 시스템에서 데이터를 추출해 매핑 작업을 거친 후 To-Be 시스템에 적재하는 업무다. As-Is의 시스템과 데이터베이스의 종류에 따라 데이터 추출 적용 기술을 다르게 해야 한다.

차세대 구축 초기에는 메인프레임의 데이터를 이행했으나, 최근에는 UNIX의 RDB 이행 사례가 증가하고 있다.

[표 5-5-1] 은행의 데이터 이행 프로젝트 사례

프로젝트명	시스템	데이터베이스	As-Is 특징	추출·전송	비고	오픈 연도
신한·조흥 통합 및 IT UPGRADE	유니시스	HDB	9BIT ASCII	SAM, FTP	9BIT를 8BIT로 변환	2006년
하나은행 차세대	IBM MF	IMS DB	EBCDIC	SAM, 스토리지 공유	EBCDIC을 ASCII로 변환	2009년
수협은행 차세대	유니시스	HDB	9BIT ASCII	SAM, FTP	0BIT를 8BIT로 변환	2011년
전북은행 차세대	유닉스	ORACLE	백업 DB 존재	Unload 툴	별도의 전환 원본 DB 구성	2013년
하나·외환 통합	유닉스	ORACLE	백업 DB 존재	DB복제	복제 DB 활용 전환	2016년
광주은행 차세대	유닉스	ORACLE	백업 DB 존재	DB복제	복제 DB 활용 전환	2016년

메인프레임에서 데이터 이행을 하는 경우 9BIT ASCII나 엡시딕(EBCDIC) 코드 변환 작업이나, HDB(Hierarchy DataBase)에서 마스터 테이블과 슬레이브 테이블을 분리하는 작업을 수행해야 한다.

UNIX RDB인 경우에는 ETL 툴이나 언로드(Unload) 툴을 사용해 추출할 수 있으며, DB 복제 방법을 활용할 수 있다. DB 복제 방법은 데이터 추출 시간만큼 이행시간을 단축할 수 있으며, 추출 오류를 방지할 수 있다.

데이터 추출 후 이행 서버로 데이터를 전송하는 방법은 FTP, SFTP, NAS, MFT, 스토리지 공유, 스토리지 전송 솔루션 등 다양한 전송방법 중에서 발주사 시스템, 데이터 양, 이행 시간을 고려해 발주사 시스템팀과 협의해 결정해야 한다.

4. 개인정보보호 규정 강화에 따른 데이터 이행 영향도

금융 분야는 개인정보보호 분야 일반법인 개인정보보호법('11)과 금융관련 법령인 신용정보법('95), 금융실명법('97) 및 전자금융거래법('07) 등을 통해 개인정보보호 관련 제도의 적용을 받고 있다. 데이터 이행은 해당 법규를 준수할 수 있도록 아키텍처를 설계하고 준비해야 한다.

고객 식별정보로 분류하는 데이터는 실명번호(주민등록번호, 여권번호, 거소증번호), 성명, 주소, 전화번호, 이메일 등이며, 실명번호의 경우 업무 특성상 사업자번호와 법인번호가 포함될 수 있다. 또한 금융회사의 보안 규정에 따라 계좌번호, 카드번호 등이 포함될 수 있다. 테스트를 수행하기 위해서는 고객 식별정보를 변환해 사용해야 하는데, 현재는 대부분의 금융회사에서 고객정보 변환 솔루션을 도입하고 있어 편리하게 사용할 수 있다.

고객정보 변환의 경우 실명번호 변환은 동일 실명번호에 대해 동일 변환 값으로 변환해야 하며 주민등록번호, 사업자번호, 법인번호는 검증 규칙에 위배되지 않도록 변환 값을 생성해야 한다.

데이터 이행은 변환되지 않은 실 데이터를 전환해야 하기 때문에 클린룸을 활용한다. 클린룸은 출입이 통제된 공간에서 데이터 유출 방지를 위한 기능을 완벽하게 갖춘 보안 PC를 이용해 제한된 인원이 작업을 수행하는 장소다. 클린룸을 활용해 데이터 이행을 수행하는 경우에도 금융감독원에 비조치의견서를 제출해 사전 승인을 받아야 한다.

5. 데이터 이행 조직의 변화

과거에는 별도의 데이터 이행 조직없이 프로젝트를 수행했으나, 최근에는 전문조직인 데이터 이행팀을 구성해 이행 작업을 수행한다.

데이터 이행팀은 데이터 이행 공통과 데이터 이행 개발 파트로 분리해 구성한다. 데이터 이행 공통 파트는 데이터 이행 아키텍처 설계와 개발 가이드 제공, 이행 지표 관리 등의 공통 업무를 수행한다. 데이터 이행 개발 파트는 이행 프로그램 개발과 테스트 및 이행 작업을 수행하는 조직이다.

[그림 5-5-1] 데이터 이행 조직 구성 예시

6. 데이터 이행 설계 방식의 발전

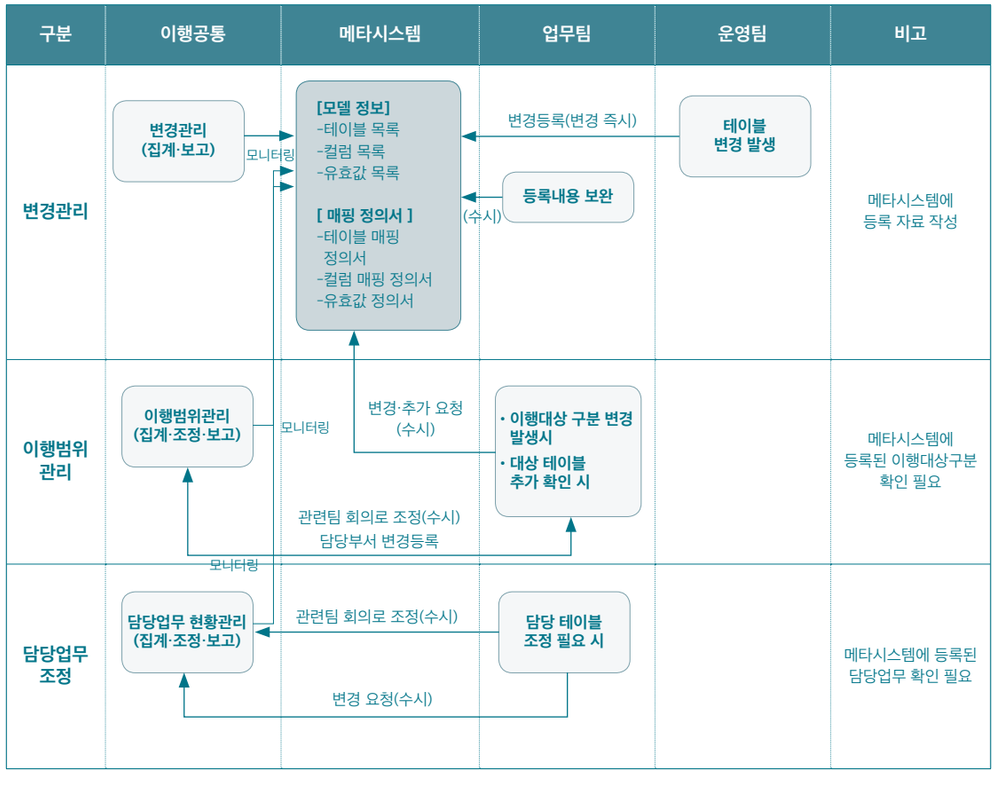
데이터 이행 작업의 핵심은 매핑 요건 정의서를 작성하고, 매핑 설계를 완성하는 것이다.

전통적인 매핑 설계 작업에는 엑셀을 이용했는데, 테이블 매핑, 컬럼 매핑, 코드 매핑을 하나의 엑셀로 관리하는 방식이다. 엑셀을 이용한 작업 방법은 모델 변경과 매핑 변경에 따라 변경되는 매핑 정의서의 변경 관리 및 집계에 어려움이 있으며, 모수의 누락이 발생할 가능성이 있다. 매핑 설계서를 데이터베이스에 등록할 수 있도록 정규화하고, 메타 시스템에서 등록 관리하면 편리하게 관리할 수 있다. 메타 시스템에서 모델 및 코드 관리를 하기 때문에, 소스 모델과 소스 코드를 등록하면 매핑 설계서 작성을 위한 기본 정보와 연계해 관리할 수 있다.

7. 데이터 이행 프로그램 개발 방법의 발전

데이터 이행 프로그램 개발은 셀, Pro-C, C에서 ETL 툴로 변화됐다. 최근에는 ETL 툴에 개발 대상 프로그램의 기본 정보를 자동생성 등록함으로써 데이터 이

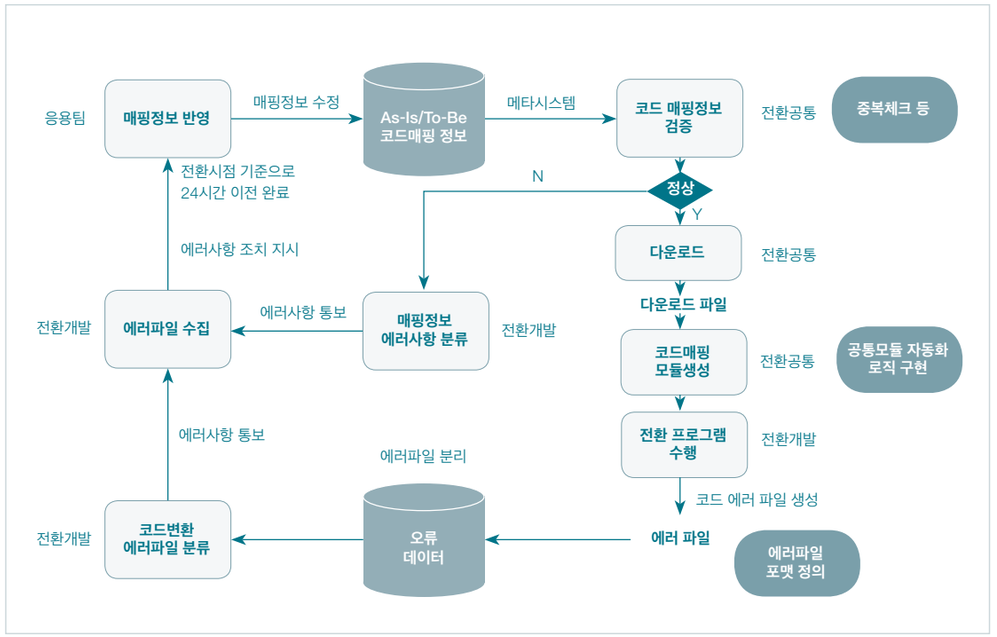
[그림 5-5-2] 메타시스템을 활용한 데이터 이행 설계 작업 흐름도



행 개발자가 프로그램 개발보다 데이터 정합성에 집중할 수 있도록 환경을 제공하고 있다.

금융권 데이터 이행의 경우 고객번호 매핑, 계좌번호 매핑, 코드 매핑, 날짜 정합성 검증, 채번 룰에 따른 일련번호 매핑, 영업일 매핑이 다수 존재한다. 매핑 데이터를 테이블로 만들어 조인할 경우 처리 시간이 오래 걸리고 오류 데이터에 대한 확인 작업이 어렵다. 공통 모듈을 이용할 경우 프로그램 개발이 단순해진다. 또한 오류 데이터를 자동으로 생성하고 DB에 적재하기 때문에 오류 데이터 정리가 편리하다. 공통 모듈을 메모리 매핑 방식으로 개발하면 데이터 이행 시간을 단축할 수 있다.

[그림 5-5-3] 공통 모듈을 활용한 데이터 이행 및 오류 처리 흐름도



8. 데이터 이행 검증

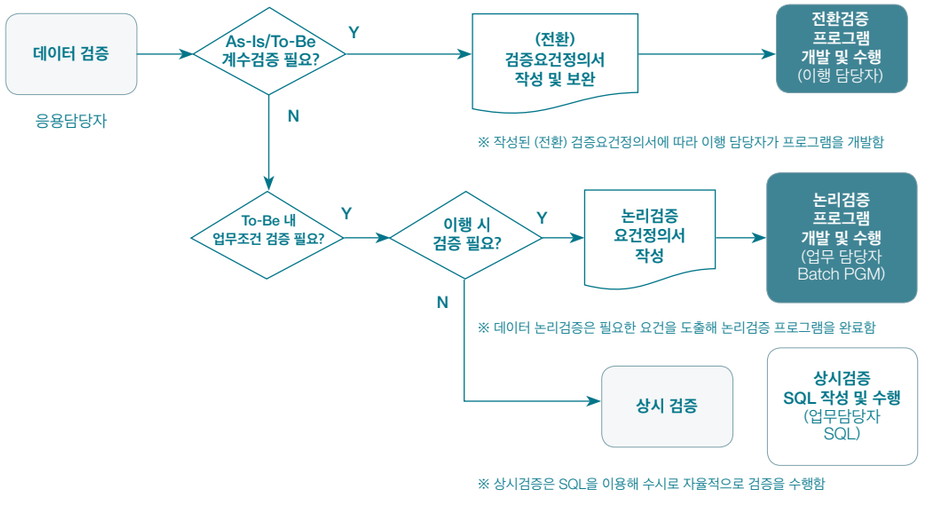
데이터 이행 검증의 종류는 이행 검증, 논리 검증, 상시 검증으로 분류해 진행한다.

이행 검증은 검증 요건을 데이터 이행 관점에 검증하는 작업이다. As-Is 시스템에서 To-Be 시스템으로 이행된 요건에 대한 계수를 비교 검증한다. 검증 요건은 매핑 설계서를 작성한 업무 담당자가 작성한다.

논리 검증은 데이터 이행 후 업무 관점에서 To-Be 데이터의 논리적인 업무 요건을 검증하는 것이다. 데이터 이행팀이 수행하는 이행 검증과 달리, 논리 검증은 이행 후 정해진 요건을 최적화해서 배치 SQL 프로그램을 통해 검증한다.

상시 검증은 데이터 완성도를 높이기 위해 업무 담당자가 필요에 따라 자율적으로 점검한다.

[그림 5-5-4] 데이터 검증 유형별 적용 흐름도



9. 향후 방향성

금융권은 현재 비대면 채널(CD, ATM, 인터넷뱅킹, 스마트뱅킹 등) 거래가 폭발적으로 증가하고 있다. 음성 데이터의 텍스트 변환(Speech To Text, STT), 외부 공개 정보 수집, 소셜미디어 정보 수집으로 데이터 용량도 기하급수적으로 증가하고 있다.

기존의 데이터 이행을 통해 데이터 처리 및 관리 기술이 급속도로 발전했으나, 빅데이터 및 제4차산업혁명의 시대에 대비하기 위해 대량 데이터의 처리 기술은 더 발전할 필요가 있다. 데이터 아키텍처를 더욱 발전시켜 빅데이터 구축에 대비해야 한다.

데이터 소스가 계정계(기간계), 정보계, 채널, 웹·앱 로그, 외부 데이터, 소셜 데이터로 다양화 되고, 테이블 및 데이터 양의 증가로 데이터 거버넌스 관리의 필요성 또한 커지고 있다. 기존의 매핑 정보, 이행 관리 시스템, 이행 프로그램이 결합된 데이터 이행 통합 거버넌스를 제공해야 한다.

제 6 장

인공지능 적용

챗봇은 제4차산업혁명의 핵심 기술인 인공지능을 구현하는 머신러닝, 자연어 처리, 의미 분석, 텍스트 마이닝, 검색, 온톨로지 등을 접목한 하이브리드 성장형 모델이다. 인공지능 관련 초기 산업이 지나친 기대감과 학술적이며 이론적인 고민 수준에 머물렀던 것에 반해, 현재는 현실적이고, 구체적이며 실용적으로 접근하고 있음을 입증하는 가장 대표적인 예다.

1. 데이터산업의 미래

2006년 출간된 엘빈 토플러의 '부의 미래'에서는 새로운 소통 수단(데이터)으로 특정 시간과 장소에 얽매이지 않아도 되는 시간과 공간의 혁명을 제4의 물결로 예견했다. 지금의 데이터 시대에서 빅데이터, 클라우드, 인공지능(AI), 사물인터넷(IoT) 등은 다양한 산업과의 융합을 통해 초생산사회, 초지능사회, 초연결사회로 이어지는 제4차산업혁명의 핵심 요소로 받아들여지고 있다.

「하버드 비즈니스 리뷰 2016」에서는 기업을 데이터와 인공지능을 가진 기업과 그렇지 못한 기업으로 분류했을 만큼, 데이터산업에서 데이터와 인공지능은 빼놓고 이야기할 수 없는 부분이 됐다. 인공지능 기술이 데이터와 함께 전 산업분야에서 핵심기술로 등장할 것으로 예상된다.

* 필자: 김영래(와이즈넛 팀장)

[표 5-6-1] 인공지능 기술 활용 분야

구분	활용분야
대화형 커머스 및 O2O	쇼핑, 비행기 예약, 숙소 예약, 레스토랑 예약 및 주문, 택시 호출
고객응대 서비스(챗봇)	금융(증권·보험·은행), 통신, 리테일(물류) 등
개인비서 서비스	헬스케어, 뉴스피드, 날씨 정보, 금융 상담, 일정 관리, 길찾기 등
로봇 컨시어지 서비스	국제회의장, 전시장, 백화점, 호텔 등
의료상담 서비스	병원, 약국 등
공공기관 고객응대 및 상담	관공서, 공단, 공사 등 법률 상담, 세금 납부, 부동산 정보, 구인구직
기업용 메신저	정보 검색, 파일 공유, 데이터 보관, 팀원 정보 공유, CRM

2. 인공지능 시장 통계

다수의 분석 기관이 인공지능 관련 시장 전망을 하며 인공지능 기술의 시장성 및 중요성을 강조하고 있다. 글로벌 시장 분석기관 트랙티카(Tractica)의 리서치 이사 아디탸 카울(Aditya Kaul)은 향후 10년간 인공지능 기술은 거의 모든 산업계에 영향을 미칠 것으로 전망했다. 글로벌 인공지능 시장은 현재 성장 중이며, 시장조사 기관 마다 규모의 차이는 있으나 공통적으로 높은 성장률을 예측하고 있다. 트랙티카의 「인공지능(AI) 시장 예측」 보고서에 따르면, 글로벌 인공지능 시장의 매출 규모는 2016년 6억 4,000만 달러에서 10년 후인 2025년에는 368억 달러로 성장한다고 한다. 조사기관 포레스터 리서치 역시 2017년부터 5년간 일반인들의 삶에 가장 큰 변화를 가져올 5대 혁신 기술 중 하나로 인공지능을 꼽았다.

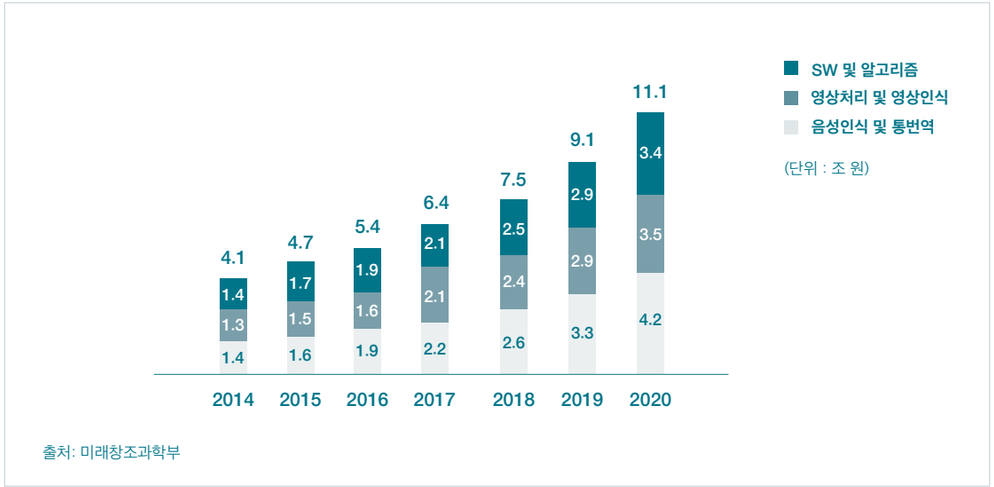
미래창조과학부가 최근 발표한 자료에 따르면, 국내 인공지능 시장 규모는 2016년 5조 4,000억 원에서 2020년 11조 1,000억 원으로 성장할 것으로 보인다.

[표 5-6-2] 세계 각국의 인공지능 관련 시장 전망

조사기관	대상	'15년 실적	전망	연평균 성장율
IDC	영상음성처리 분야	1,270억 달러	1,650억 달러	14%
	Cognitive SW 플랫폼	10억 달러	37억 달러 ('17년)	92%
BCC리서치	음성인식	840억 달러	1,130억 달러 ('17년)	16%
Market & market	서비스(광고, 미디어 등)	4억 2천만 달러	50억 달러('20년)	64%
트랙티카	AI 시스템	2억 달러	111억 달러('17년)	-
일본EY연구소	AI 관련 산업 전반(자국)	3조 7,450억 엔	23조 638억 엔('20년)	44%
IBM	2025년 2천조 원 시장 창출			
매킨지	2025년 6조 7천억 달러(7,000조 원) 파급 효과			

출처: KT경제연구소, 2016. 3

[그림 5-6-1] 한국 인공지능 시장 규모



출처: 미래창조과학부

3. 인공지능 기술 동향

인공지능은 컴퓨터에 인간과 같은 지능을 실현하기 위한 시도 및 일련의 기술을 의미한다. 세계적인 IT 자문기관 가트너는 2017년 주목할 만한 10대 전략 기술의 첫 번째로 인공지능과 머신러닝(Applied AI & Advanced Machine Learning)을 선정했다. 인공지능은 머신러닝, 자연어 처리 등의 다양한 기술로 이루어진다. 새로운 유형의 지능형 앱과 사물을 제공하는 동시에 다양한 디바이스와 소프트웨어, 서비스 솔루션을 위한 내장형 인텔리전스를 제공하게 될 것으로 전망했다. 머신러닝 역시 2016년에 이어 앞으로 지속적으로 발전할 것으로 보고 있다.

인공지능 SW의 핵심기술은 머신러닝 기술을 통해 자연어를 이해·표현해 지

[표 5-6-3] 인공지능 기술 분야

인공지능 기술분야	주요 내용
머신러닝 (Machine Learning)	• 새로운 정보를 학습하고, 습득한 정보를 효율적으로 사용할 수 있는 능력과 결부시키는 방법 • 명시적 프로그래밍 및 수학적인 모델 없이 데이터 패턴을 식별하는 알고리즘으로 미래 예측 및 이상 징후 감지 등에 활용
딥러닝 (Deep Learning)	• 머신러닝의 한 종류로 인간의 두뇌를 모방한 다중 처리 계층 및 파라미터를 사용해 데이터를 모델 분류 • 수많은 데이터를 컴퓨터에 주입하고 해당 데이터가 의미하는 바를 스스로 해독할 수 있도록 만듦
자연어 처리 (Natural Language Processing)	• 인간이 보통 쓰는 언어를 컴퓨터에 인식시켜 처리하는 것으로 정보 검색, 질의응답, 시스템 자동번역, 통역 등 포함
데이터 마이닝 (Data Mining)	• 데이터 가운데 숨겨져 있는 유용한 상관관계를 발견, 미래에 실행 가능한 정보를 추출해 내고 의사결정에 이용하는 과정
지능엔진 (Intelligent Agent)	• 인공지능적 기능을 가진 소프트웨어 엔진으로 사용자를 보조하고 반복된 업무를 인간을 대신해 실시하는 엔진
패턴인식 (Pattern Recognition)	• 기계로 도형, 문자, 음성 등을 식별
자동추론 (Automated Reasoning)	• 컴퓨터에 의한 자동추론을 가능하게 하는 소프트웨어 개발을 목표로 함

식 베이스를 구축하고 새로운 지식을 추론 및 생성해낼 수 있으며, 학습과 추론을 통해 진화가 가능한 인공지능 기술로 요약할 수 있다.

인공지능 관련 기술에 있어 해외의 경우에는 주로 정부나 글로벌 기업이 주도해 지식과 지능처리 소프트웨어 관련 연구가 진행되고 있다. 국내에서는 해외 기업들과의 기술격차를 줄이기 위해 정부 주도로 장기간 국책사업이 시행되고 있으며, 시장을 선도하는 기업을 중심으로 인공지능 적용 사업 사례가 증가하는 추세다.

[표 5-6-4] 인공지능 및 심층 질의응답 프로젝트 현황

프로젝트명	수행기관·기간	주요 내용
Watson DeepQA	IBM 2006 ~ 2011	• Factoid 형태의 복잡한 신뢰성 높은 빠른 응답 제시 • http://researchweb/watson/ibm.com/deepqa/
Knowledge Graph	Google	• 위키피디아 콘텐츠 위주 5억 개의 객체에 대한 Knowledge Graph를 구축, 질의응답 서비스 제공 • http://www.google.com/insidesearch/features/search/knowledge.html
프로젝트명	**수행기관·기간**	**주요 내용**
AQUAINT Program	NIST 2003 ~ 2008	• 다국어 기반의 Advanced QA 연구, 텍스트, 음성, 이미지, 비디오 지원 • http://www-nlpir.nist.gov/projects/aquaint
Project HALO	Vulcal Inc. 2003 ~ 2006	• Digital Aristotle: 어려운 과학문제 질문에 답변 • 파일럿 프로젝트(2004년, 6개월): domain expert 개발 • AURA(Automated User-Centered Reasoning and Acquisition System) 프로젝트(SRI, 2004~2006)
Wolfram Alpha	Wolfram Alpha 2009 ~ 현재	• Siri 서비스의 질의응답 부분 담당, 수식, 알고리즘, 모델을 적용해 사용자 의도에 맞는 정답 제시 • http://www.wolframalpha.com
Human Brain Project	EU 2013 ~ 2023	• 인간 뇌의 작동방식에 대한 정확한 이해, 활용을 통해 컴퓨팅 아키텍처, 신경과학, 의학 분야 등에 적용 예정 • EU 미래기술 주력 사업 프로그램의 6대 연구과제 선정 • http://www.humanbrainproject.eu
Google Brain Project	Google 2008 ~ 현재	• 9계층의 신경망, 10억 개의 연결 구조로 인간 인지방식의 자율학습 모델 연구 • 2만여 개의 객체 카테고리를 인식하는 데 15.8%의 정확도를 보임 • http://www.mpi-inf.mpg.de/yago-naga/yago

※ NIST(National Institute of Standards and Technology)·
※ SRI(Stanford Research Institute)
출처: 「빅데이터 지식처리 인공지능 기술 동향」, 전자통신동향분석, ETRI, 2014

[표 5-6-5] 인공지능 적용 사업 사례

구분	국내 연구 사례
미래창조과학부	• 2013년부터 10년간 엑소브레인 SW는 인간의 지식증강 서비스를 위해 빅데이터로부터 스스로 학습해 지식을 축적하고, 시스템 및 기기 간의 자율 협업방식으로 새로운 문제를 해결하는 엑소브레인 SW 기술 개발 과제 진행 중
삼성	• 인공지능 스타트업 바이캐리어스('15. 9월) 인수, 가정용 로봇 스타트업 지보(JIBO6) 투자 참여 등을 통해 인공지능 사업 기회 모색
엔씨소프트	• 인공지능 연구를 시작해 인공지능 기반 서비스를 개발(출시) 중
네이버	• 네이버랩스를 중심으로 대화형 엔진 '네이버 아이' 베타 버전 출시 • 빅데이터와 머신러닝으로 강화 '자연어 이해 기술'과 사람의 대화 처리 방식을 모사한 '대화 문맥 관리 기술'로 사용자와의 대화를 통해 정보를 제공하고 음성 명령으로 앱을 실행할 수 있다. • 인공지능 신경망 번역 '파파고', 여행 정보 제공 '코나', 뉴스 추천 '에어스' 등 출시 예상
카카오	• 음성인식과 인공지능 관련 연구개발을 전담하는 TF 별도 신설 • 검색, 추천, 데이터 커넥션 담당 조직과 'AI 부문'으로 통합
와이즈넛	• 인공지능 관련 국책연구과제 성공적 수행 • VOC 빅데이터를 활용한 지능정보기술 기반 채팅 상담(컨택트센터) • 자동화 시범서비스 개발 • 자율지능형 지식·기기 협업 프레임워크 기술 개발 • 지능정보 서비스를 위한 고성능 하이브리드 클라우드 컴퓨팅 기술 개발

기술집약적인 인공지능 시장에서 중요한 것은 지속적인 인공지능 기반 기술 확보에 있으며, 이를 위한 R&D 투자는 필수다. 미래의 성장은 현재의 성과가 뒷받침돼야 가능하다. 기업의 부진한 영업성과는 R&D 투자를 위축하게 하고, 그 결과 부족한 기술 역량과 인력을 해결하기 위해 상호 컨소시엄 형태의 변형적 사업이 진행되는 경우가 발생하게 된다. 이는 사업 과정뿐만 아니라 사업 종료 후 사업 관리 등 여러 측면에서 우려할 점이 많다.

4. 인공지능 산업 동향

인공지능 관련 초기 산업이 지나친 기대감과 학술적이며 이론적인 고민 수준에 머물렀던 것에 반해, 현재는 현실적이고, 구체적이며 실용적으로 접근하고 있는 점은 확실하다. 사람이 더 효율적이고 창의적인 업무를 수행할 수 있는 인공지능을 적용한 지능형 민원 상담 시스템(챗봇) 등으로 인공지능 산업이 성장하고 있다. 예를 들어 인공지능을 적용한 챗봇이 콜센터(컨택트센터) 본업의 업무 중 기능적 측면이나 서비스 측면에서 일부를 대체하고 있다. 이로써 콜센터 상담 직원의 역량, 자율성 및 전반적인 성과를 향상시키는 등 업무의 어려움을 줄여주고 있다.

[표 5-6-6] 해외 주요업체 인공지능 현황

회사명	프로젝트 현황
Microsoft	• 지능형 개인비서 코타나
Google	• 연구조직 신경네트워크 구축, 무인자동차, 음성인식 검색 등에 인공지능 적용
Zest Finance	• 미국의 은행들이 정형화된 변수(20개 내외)를 사용, 신용평가를 하는 것과는 달리 대출 실행전 인터넷 체류 시간, SNS 포스팅 주제 등 7만여 개의 변수에 대한 데이터를 수립하고 10개의 머신러닝 알고리즘을 이용해 개인의 신용도 분석
Kabbage	• 미국 소상공인 대출회사로 Data Context Engine이라는 독자적인 시스템을 이용해 대출자의 각종 데이터(배송, 회계, 소셜미디어, 전자상거래)를 분석 후 활용해 7분만에 간편 대출 서비스 제공
BillGuard	• 머신러닝 기술을 활용한 예측 알고리즘을 통해 고객 거래 데이터를 분석, 의심스러운 징후가 포착되면 즉시 고객 앱으로 경보 제공
Kaisto	• 인공지능과 음성인식 기능이 탑재된 모바일 개인 파이낸싱 어플을 제공해 이달의 지출액, 커피숍 사용 금액, 카드 잔고 등을 인공지능이 음성으로 분석해 알려주며, 결제도 진행

인공지능에 대한 여러 가지 우려에도 불구하고 인공지능의 발전 속도는 갈수록 빨라지고 있다. 이미 인공지능을 활용해 환자의 병을 진단하거나 인공지능이 언어를 자동 번역하기도 한다. 최근에는 고객이 문자로 질문한 것을 이해해 답을 문자로 보내는 챗봇(Chat bot)을 활용하는 기업도 늘고 있다.

[표 5-6-7] 국내 주요 인공지능 프로젝트 현황

구분		프로젝트 현황
금융 분야	대신증권	벤자민
	NH농협	금융봇
	동부화재	프로미 챗봇
	라이나생명	챗봇
와이즈넛 (WISEnut)		머신러닝, 자연어처리, 의미 분석, 텍스트 마이닝, 검색, 온톨로지 등을 접목한 하이브리드 성장형 모델인 인공지능 챗봇을 통해 국내 주요 금융(은행, 증권), 유통(쇼핑, 물류), 공공기관 고객만족센터, 의료, 제조, 쇼핑 등 다양한 분야에 공급 중
부노(VUNO)		딥러닝 알고리즘을 의료 분야에 특화 적용시켜 엑스레이·컴퓨터 단층촬영(CT)·자기공명영상(MRI) 및 생체 신호 데이터를 토대로 임상 진단
파수닷컴		파수닷컴의 주요 제품에 머신러닝 기술 적용
퓨이드(Riiid)		데이터와 머신러닝 기반 에듀테크 기업
코난테크놀로지		인공지능 기반 얼굴 인식기술을 개발, 방송국의 콘텐츠 관리 솔루션에 반영
스탠다임		10년 이상 1조 원의 비용이 투입되는 신약 개발 과정을 데이터에 기반한 머신러닝 기술을 활용해 개발 기간을 대폭 줄이는 기술개발 중
솔리드웨어		머신러닝 기술 이용 데이터 분석 솔루션 제공 보험, 신용카드, 은행·대출 등 금융 분야 필요 예측 모델 제공

5. 맺음말

현재 태동기인 우리나라의 제4차산업혁명은 인공지능 기술을 바탕으로 금융, 의료, 제조업 등 기술적, 산업적, 사회 문화적 측면에서 광범위한 파급효과를 가져올 것으로 전망된다. 특히 지능형 인공지능 챗봇에 대한 관심이 높아지고 있다. 2017년 하이브리드 성장형 모델을 적용한 인공지능 챗봇은 과거 비용을 많이 들여야 했던 일들을 보다 저렴하면서, 사람이 더 효율적이고 창의적인 업무를 수행할 수 있도록 산업혁신을 위한 미래 성장동력으로 자리잡을 것이다.

2016 데이터 컨설팅 이노베이터상 수상자

"제4차산업혁명의 기반인 데이터는 신뢰성 확보가 중요"

이해곤 | 비투엔 CIS본부 기술이사

2016 데이터 컨설팅 이노베이터상을 수상한 이해곤 비투엔 이사는 1994년 BYC생명보험에서 사회생활을 시작해 교보정보통신, 데이터스트림즈, 비투엔 등 IT 업계에서 데이터 품질 솔루션 및 데이터 품질 컨설팅 업무를 해오고 있다. 국내 데이터 품질 컨설팅 분야를 개척해온 대표 전문가들 가운데 한 명으로 교보생명·한화생명·특허청·금융감독원·국민연금·건강보험·수자원공사·행정자치부 등 굵직한 사이트의 데이터 품질 컨설팅을 진행했다.

데이터 컨설팅 이노베이터 상을 받은 소감은.

10여 년 동안 데이터 컨설팅 분야에서 일해왔는데, 더 분발하도록 배려해 주신 것으로 생각하고 개인적으로 해야 할 역할이 무엇인지 생각하고 있다. 컨설팅 분야 중에서도 데이터 품질 영역에서 오랫동안 일해와서 그 경험과 지식을 나누는 교육, 프로젝트 활동에 지속적으로 참여하겠다. 데이터 품질 분야에 이렇다 할 책이 국내에 없는 상황이므로 책 한 권을 쓰는 것도 생각해 보았다. 그 고민이 결과물로 나와야 하는데, 현장 프로젝트를 핑계로 계속 미루고 있다.

2017년 기준으로 데이터 컨설팅에 대해 간단히 소개한다면.

데이터 컨설팅을 데이터 솔루션과 비교하자면, 솔루션은 컨설팅에 비해 솔루션 구성 아키텍처에 포커스를 맞춰 접근하는 방식을 택하는 반면, 컨설팅은 새로운 트렌드의 변화와 기업의 운영 환경을 고려해 사용자 시각에서 더 다양한 전략을 제시한다. 최근 국내 데이터 품질 시장은 초기 단계를 벗어나면서 점차 데이터 솔루션과 컨설팅 사업 영역이 모호해지고 통합되고 있는 모습이다. 데이터 컨설팅사에서 솔루션을 내놓는 경우도 늘고 있다. 과거 데이터 컨설팅이 특정 업체와 특정 전문가 중심으로 펼쳐졌는데 시간이 지나면서 방법론이 체계화되고 보편화 되면서 솔루션과 컨설팅의 경계가 흐려지는 현상으로 본다. 데이터 컨설턴트 입장에서 보면 빅데이터, 사물인터넷(IoT), 인공지능(AI) 등 새로운 데이터 시대에 맞는 데이터 품질 컨설팅 방법론과 방향 연구에 정열을 투자해야 한다고 본다.

데이터 품질이 어느 때보다 강조되는 모습이다.

국가에서 건강에 대한 복지예산을 늘려 비용을 지출하면서 국민의 건강을 지키고자 하는 배경 중 하나는 투자대비효과(ROI)가 있다는 사실을 알기 때문이다. 한 사람의 생애 건강관리를 아기를 낳기 위한 부모의 몸부터 건강하게 관리하듯이 기업의 데이터를 건강하게 관리하기 위해 시스템 계획, 구축, 운영, 활용 단계에 체계적으로 접근할 뿐 아니라 단계별 요소마다 데이터 품질을 거버넌스할 수 있도록 데이터 품질 체계를 마련해야 한다. 제4차산업혁명 시대에 데이터 품질이 더욱 강조되는 이유는 실시간 빅데이터를 활용한 의사결정, 인공지능 서비스 시대에 실수(오류)를 회복하는 데 예전보다 더 심각하고 막대한 비용을 지출해야 하기 때문이다. 또한 국내외적으로 공공정보를 공유하고 개방함으로써 데이터 기반의 가치 창출, 융복합 활용 서비스가 활발히 일어나고 있다. 예를 들어 국가 재난 발생시 기상, 지진, 홍수, 가뭄 등 자연재난과 교통, 시설물, 원전 등 사회재난, 또 자연재난과 사회재난이 동시에 일어나는 복합재난 발생시에 공공정보, SNS 정보를 활용함으로써 재난 예방, 대비, 대응, 복구 등에 신뢰성 있는 데이터 제공 및 활용 인프라가 다양한 시각에서 추진되고 있다. 이를 위해 정부는 국가 중요 데이터 소관기관들의 데이터 품질(관리)수준 평가 등 제도 시행을 통해 품질을 제고할 수 있도록 앞장서고 있다.

데이터 품질을 높이기 위해서는 어떤 노력이 필요한가.

최근 규모가 있는 기업에서는 'CDO(데이터 책임자)' 등 데이터 관리 부서를 두고, 공공기관에서는 '공공데이터 제공 책임관'을 지정하도록 하고 있다. 또한 빅데이터 기술, AI 기술 등 발전적인 기술을 데이터 품질 분야에도 적용할 수 있도록 방법론(기술)을 체계화하고 이전보다 더 신속하고 정확한 품질관리를 위해 인간의 힘을 최소화할 수 있도록 자동화된 인프라를 구축할 필요가 있다. 2015년 실시간 수도정보 품질관리 프로젝트를 진행한 적이 있다. 센서로부터 수질 정보를 받아서 모니터링하는 것에서 끝나지 않고 기술적 오류에 의한 데이터를 보정해 품질을 높이려는 접근이었다. 현재는 관리자의 경험과 지식을 동원해 수작업으로 범위를 벗어나는 데이터를 추출·보정하는 수준이다. 향후 데이터 양과 활용도가 늘어나면, 이 같은 수작업은 한계에 이를 수밖에 없다. 그 차원에서 머신러닝 기법을 써서 사람의 관여를 줄임으로써 신속하고도 신뢰성 있는 정보관리를 위한 노력들을 추진하고 있다.

데이터 품질 분야에 진출하려는 젊은이들에게 하고 싶은 말.

IT하는 사람의 입장에서 데이터 분야는 데이터베이스라는 실체와 원하는 형태로 자유롭게 데이터를 추출(시각화)할 수 있는 기술(SQL 등)이 있어서 충분히 매력적이다. 데이터 분야에 진입하려는 이들도 반드시 소프트웨어 개발 분야도 경험해 보기를 바란다. 데이터 분야에서도 모델링이나 성능 튜닝 등에만 집중돼 있는데, 데이터 품질 분야에는 아직까지도 잘 모르고 관심을 덜 보이는 모습이다. 당장 필요한 모델링이나 낮은 성능을 끌어올리는 튜닝이 스포트라이트를 받기 때문이다. 데이터 품질은 당연한 것을 잘 해야 하고 애플리케이션처럼 눈에 들어나지 않으며, 감추고 싶은 부분(오류)을 드러내야 하는 조건이 따른다. 그런데 생각해보면, 제4차산업혁명 시대에는 데이터가 중요해지는 만큼 품질 이슈도 늘어날 것이다. 데이터 품질 컨설턴팅이 '블루오션'이라는 뜻이다. 데이터 품질 분야는 데이터 표준화 관리, 구조 관리, 성능 관리, 데이터베이스 구축 등을 포괄한다고 보면 된다. 어떻게 보면, ISP 사업의 성격까지 갖고 있다. 따라서 IT 분야의 다양한 영역들에 대한 이해도를 위해 폭넓은 지식과 경험을 쌓아가야 한다.

제 6 부

데이터 기술 동향

제1장

인공지능과 고급 머신러닝

빅데이터의 축적과 고성능 연산머신의 등장으로 막대한 규모의 데이터로부터 새로운 가치를 창출하는 데이터 분석이 현실화됐다. 본 고에서는 빅데이터의 생성 및 수집·통합·분석의 과정에서 활용되는 딥러닝(Deep Learning) 등 다양한 인공지능 및 고급 머신러닝 기술을 간략하게 설명하고, 전이학습(Transfer Learning)과 같이 주목할 필요가 있는 새로운 연구 동향을 소개해 인공지능 및 주요 고급 머신러닝 기술의 이해를 돕고자 한다.

1. 빅데이터와 인공지능

산업혁명 당시 영국의 경제학자 데이비드 리카르도(David Ricardo)가 제기한 기계에 의한 노동력 대체 현상의 함의에 대한 질문**은 산업혁명에 못지 않은 관심을 받았다. 현재 다시 주목 받기 시작한 인공지능(Artificial Intelligence, AI), 그리고 머신러닝 기술은 리카르도가 주목했던 기술 이상의 잠재력을 지니고 있다. 자연어 처리나 이미지 인식에서 달성된 딥러닝의 성과를 고려하면, 인공지능과 머신러닝의 발전에 대해 스티븐 호킹(Stephen Hawking)이나 일론 머스크(Elon Musk)와 같은 저명인사들이 표명하는 우려가 충분한 근거가 있는 것처럼 보이기도 한다.*** 딥러닝 혹은 고급 머신러닝 기법은 어떤 원리에 기반했기에 이처럼 놀라운 성과를 보이는 것일까?

인공지능 그리고 고급 머신러닝 기술이 달성한 놀라운 성과는 대용량 데이터(빅데이터)에 숨겨진 유용한 패턴 탐색 기능에 기반한 것이다. 주어진 태스크를 해결하기 위해 인간이 설정한 규칙에 의존하는 것이 아니라, 학습 규칙을 이용해 문제 공간의 적절한 표현 방법을 찾은 후 이에 기반해 임무를 달성한다. 이 과정에서 다양한 기법들이 결합하게 된다. 이 글에서는 현재 화제가 되고 있는 인공지능 및 고급 머신러닝 기술을 더 잘 이해할 수 있도록 고급 머신러닝 기술을 구성하는 단위 기술들과 주목해야 할 차세대 기술 동향을 소개하고자 한다.

* 필자: 석호식(강원대학교 IT대학 컴퓨터정보통신공학전공 조교수)

** Principles of Political Economy and Taxation, 1821.

*** 일론 머스크는 MIT AeroAstro Centennial Symposium (2014)에서 인공지능이 인류의 존재에 대한 가장 큰 위협이라는 견해를 밝힌 바 있다.

2. 인공지능과 고급 머신러닝의 구성 기술

가. 빅데이터

데이터 생성 및 수집, 데이터 통합, 데이터 분석의 과정을 고려하면 머신러닝 관점에서 빅데이터의 중요성을 더욱 명확하게 이해할 수 있다. 빅데이터 시대의 데이터 생성 및 수집의 가장 큰 특징은 다양성이다. 우선 행동 데이터(스마트폰의 센서와 결합된 움직임 데이터 혹은 위치 데이터), 거래 데이터, 지식 데이터 등 다양한 종류의 데이터가 여러 플랫폼에서 생성되고 있으며, 특히 전통적인 데이터와 달리 비정형 데이터도 큰 비중을 차지한다. 다양한 데이터를 '통합'하는 과정을 통해 빅데이터는 더 큰 가치를 창출하게 된다.

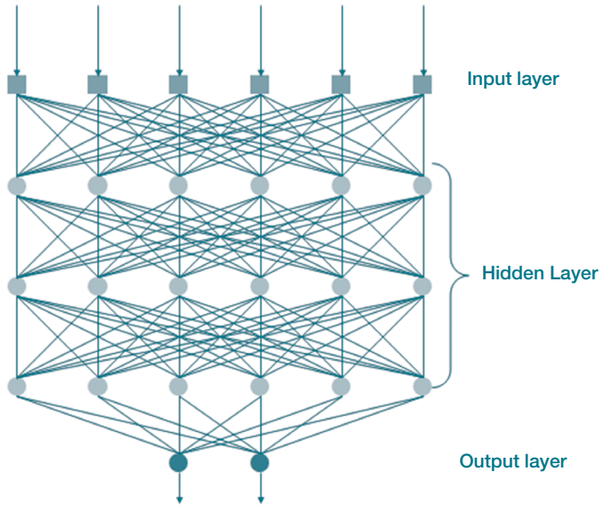
데이터 통합의 가장 큰 특징은 '여러 데이터 집합의 통합'을 통해 단일 데이터 집합에서는 찾을 수 없는 새로운 통찰(Insight)을 찾을 수 있다는 점이다. 데이터 통합은 특히 여러 데이터를 결합해 처리하기 어려운 데이터 집합에 적용할 때 가치가 증대되는데, 인공지능 및 고급 머신러닝 기술의 발전과 함께 빅데이터에서 추출되는 가치도 증대되고 있다.

데이터 통합 과정에서 머신러닝의 중요성은 1)데이터의 막대한 분량과 2)데이터의 큰 차원 때문에 부각된다. 다양한 공적·사적 데이터 소스에서 발생하는 데이터를 인간 분석자의 작업을 통해 처리해 유용한 정보를 추출한다는 것은 현실적으로 어려운 작업이다. 따라서 매우 많은 차원으로 설명되는 데이터를 분석해 데이터에 내재된 패턴을 추출할 수 있는 고급 머신러닝 기법을 활용해야 하는데 그 과정에서 현재 딥러닝이 부각되고 있다.

나. 딥러닝

딥러닝은 다양한 분야에서 높은 성능을 발휘하고 있지만 특히 표현의 학습(Representation Learning)에 강점이 있다.* 딥러닝에서 의미하는 표현의 학습은 전처리를 거치지 않은 원시데이터를 입력하면 자동으로 주어진 임무에 필요한 표현을 발견하는 학습을 의미한다. 딥러닝 기법에서는 특정 수준의 데이터 표현에 간단하지만 비선형 특성을 갖는 변환 과정을 적용해 약간 더 추상화된 데이터 표현을 획득하는 과정을 반복하면서 주어진 데이터에 대해 여러 수준으로 추상화된 데이터 표현을 확보하게 된다. 딥러닝에서는 인간에 의한 전처리를 거치지 않고 유효한 표현 방법을 학습을 통해 발견함으로써 놀라운 수준으로 성능이 향상된다. 딥러닝의 핵심 아이디어는 새롭게 등장한 것이 아니라 막대한 규모의 훈련 데이터와 강력한 연산 기능을 적용해 인공신경망(Artificial Neural Networks)의 근본 아이디어를 확대한 것이다. 딥러닝 혹은 인공신경망의 모델에서는 실제 두뇌와 유사하게 아주 많은 수의 뉴런(혹은 계산노드)이 모여 정보처리모델을 구성하게 되는데 보다 체계적으로 나누어보면 데이터가 네트워크로 공급되는 입력층과 연산 결과가 출력되는 출력층 그리고 정보가 처리되는 은닉층이 존재한다[그림 6-1-1]. 각 계산노드 혹은 뉴런은 다른 계산노드와 연결돼 있으며 계산노드와 계산노드 사이의 연결부에는 가중치가 부여되고, 개별 계산노드에서는 해당 계산노드의 출력을 제어하는 액티베이션 함수(activation function)가 동작한다.

[그림 6-1-1] 딥러닝의 처리 모델



* Yann LeCun, Yoshua Bengio & Geoffrey Hinton (2015), *Deep Learning,* Nature 521: 436-444

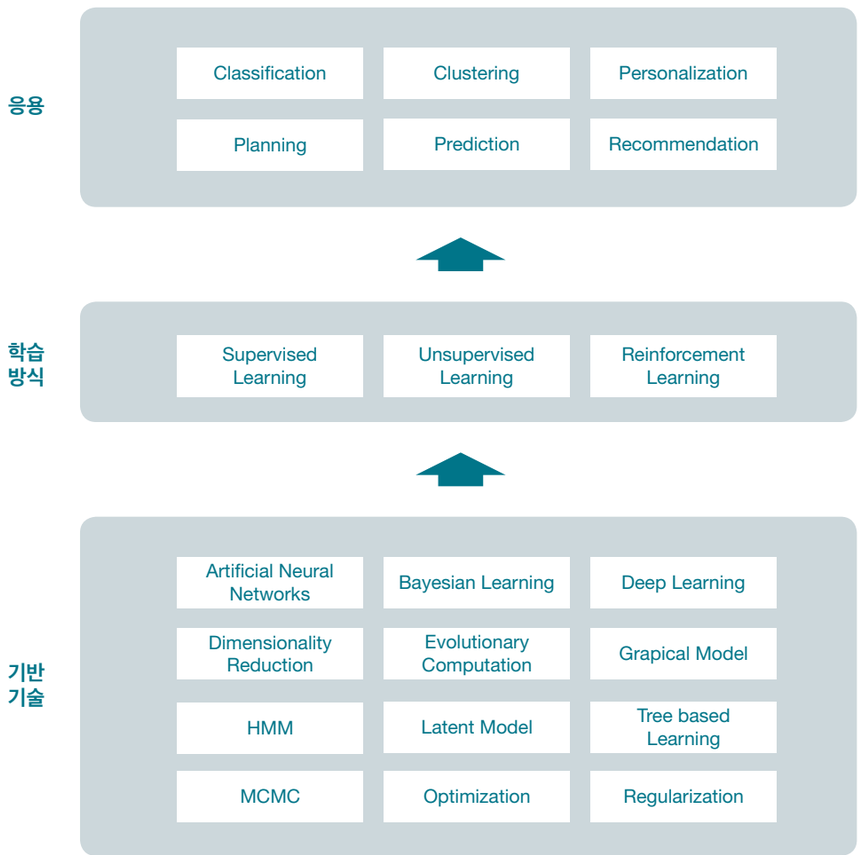
딥러닝은 간단한 모듈을 여러 층 축적함으로써 주어진 데이터에 내재된 매우 복잡한 특성을 입력 변화에 강건한 방식으로 획득할 수 있다. 그러나 딥러닝 혹은 딥뉴럴넷으로 사용자가 원하는 모든 기능을 수행할 수 있는 것은 아니며, 딥러닝이 제공하는 표현 학습 능력과 고급 머신러닝 기법의 복잡한 추론 능력이 결합할 때 인공지능의 잠재력이 구현될 수 있다. 따라서 딥러닝 외에도 다양한 고급 머신러닝 기술을 이해할 필요가 있다.

다. 인공지능과 고급 머신러닝의 구성 기술

인공지능 및 고급 머신러닝은 정체를 모르는 새로운 데이터 인스턴스의 카테고리 결정, 유사 데이터로의 그룹핑, 맞춤형 솔루션 제공을 위한 개인화, 주어진 태스크 해결을 위한 계획 수립, 미래에 대한 예측 그리고 사용자를 위한 맞춤형 추천 등을 위해 사용되고 있다. 이를 위해 감독학습(Supervised Learning), 무감독학습(Unsupervised Learning), 강화학습(Reinforcement Learning)의 학습 기법이 사용된다. 이러한 학습 방법들은 신경망, 베이지안 학습, 진화학습, 그래프 모델(Graphical Model), HMM(Hidden Markov Model), 은닉변수 모델, MCMC(Markov Chain Monte Carlo) 등의 기법들을 이용해 구현된다[그림 6-1-2].

현재 인공지능 및 고급 머신러닝 기술이 달성한 업적들은 상기 설명한 다양한 기능들이 결합된 것이다. 예를 들어 알파고의 경우 두 개의 딥뉴럴넷을 포함한 여러 개의 모듈이 상호 연결된 강화학습 시스템이다. 딥뉴럴넷 중 하나는 수백만 개의 기보를 분석해 얻은 모델을 이용해 복수 개의 가능한 움직임을 추천하며, 다른 모델은 랜덤 샘플링 기법에 기반해 추천된 움직임을 분석해 평가하는 역할을 한다. 하지만 빅데이터에 고급 머신러닝 기법을 적용해 데이터를 분석하는 일련의 활동은 따기 쉬운 과실에 해당하는 부가가치 창출 활동이다. 더 큰 가치를 창출하기 위해서는 이제까지 설명했던 머신러닝 기법 외에 새로운 고급 기법이 필요하다.

[그림 6-1-2] 인공지능과 고급 머신러닝의 구성 기술 및 학습 방식



(일부 대표적인 기술만 정리한 것임)

3. 인공지능 및 고급 머신러닝 기술의 주목할 연구 동향: 전이학습

2015년 백채널(Backchannel)에서 실시한 대담에서 딥마인드(DeepMind)의 창립자인 데미스 하사비스(Demis Hasabis)는 현재 딥마인드가 전이학습이라는 연구 주제에 집중하고 있다고 밝혔다.* 전이학습은 새로운 태스크를 해결하기 위해 매번 관련 데이터를 학습해 모델을 구축하는 것이 아니라 이미 획득한 지식(모델)을 활용해 새로운 모델을 만들려는 시도이다. 보다 구체적으로 설명하면 이미 학습이 완료된 관련 태스크의 지식을 새로운 태스크 해결을 위한 모델에 전이해 어느 정도 학습된 모델을 구축한 후 새로운 모델을 목표 태스크에 맞게 적응시키는 과정을 통해 새로운 모델을 확보한다.

전이학습은 학습된 딥뉴럴넷의 은닉계층 일부를 전이하는 방식(딥러닝을 통해 학습된 지식은 목표 태스크에 특화된 상태에서 일반화된 표현이기 때문에, 일반화됐으나 목표 태스크에 덜 특화된, 즉 입력 계층에 가까운 일부 은닉계층만 전이한다) 혹은 베이지안 학습 과정에서 다른 태스크에서 확보한 확률 분포를 사전 확률 분포로 활용하는 방식으로 구현되기도 한다. 더 나아가 타깃 태스크의 솔루션이 여러 개의 서브 솔루션이 결합된 계층 구조일 경우 다른 태스크에서 확보한 서브 솔루션을 결합하는 방식을 통해 솔루션의 계층 구조를 확보할 수도 있다. 전이학습은 머신러닝 모델의 가장 근본적인 전제, 즉 모델 훈련 과정에서 사용하는 훈련 데이터와 해당 모델을 적용할 테스트 데이터가 동일한 분포를 따른다는 가정을 극복할 수 있는 방법이라는 점에서 인공지능 및 고급머신러닝의 가능성을 넓힐 수 있는 방법이다. 그러나 전이학습의 잠재력을 구체화하기 위해서는 먼저 1) 표현 동기화 2) 유사성 기반 매핑 등의 방법을 확보해야 한다:

1) 표현 동기화: 전이될 도메인과 전이 받을 도메인의 표현 모델이 상이할 경우 서로 다른 표현 모델을 활용할 수 있도록 모델간의 표현을 동기화한다.
2) 유사성 기반 매핑: 인과관계 지식 등을 활용해 태스크간의 유사성을 판단하고 전이의 성능을 향상시킨다.

하지만 전이학습을 적용한다고 항상 타깃 태스크의 성능이 높아지는 것은 아니다. 표현 동기화, 유사성 기반 매핑 등 성공적인 전이학습을 위한 필수 관련 분야에서 만족스러운 성능을 보이는 머신러닝 기법이 확보되지 못한 관계로 전이학습을 잘못 적용할 가능성이 매우 높다. 전이학습을 잘못 적용할 경우 학습 성

* https://backchannel.com/the-deep-mind-of-demis-hassabis-156112890d8a

능이 저하되는 네거티브 전이(Negative Transfer) 현상*이 발생한다. 아직까지 전이학습을 실제 응용문제에 적극적으로 적용할 수는 없지만, 전이학습의 가능성을 고려할 때 적극적인 연구가 필요한 분야임에 틀림없다.

4. 요약

현재 데이터 분석을 위해 활용되는 인공지능 및 머신러닝 기법들은 각종 고급 통계기법과 실제 자연현상에서 착안한 아이디어를 활용해 놀라운 결과를 제시하고 있다. 딥러닝 모델은 전처리 과정 없이 원시데이터에서 타깃 태스크 해결에 필요한 표현을 확보할 수 있게 해주었으나, 실생활 문제를 해결하기 위해서는 딥러닝과 같은 생물학 기반 모델과 통계학 기반 학습 방법 혹은 강화학습과 같은 학습 방법의 통합이 필요하다. 그러나 생물학 기반 모델과 그 외 고급 머신러닝 기법을 통합한다고 해도 훈련 데이터가 확보돼야만 모델을 구축할 수 있으며, 훈련 데이터와 테스트 데이터의 분포가 동일할 경우에만 학습한 모델이 유용하다는 제약을 극복할 수 없다. 또한 상상할 수 있는 모든 태스크에 대한 데이터를 수집해 타깃 태스크를 위한 모델을 학습하는 방법 역시 계속해서 추구할 수 있는 방법이라고 볼 수 없다.

이미 확보된 데이터 및 향후 새롭게 축적될 데이터를 통합해 데이터에 숨겨진 가치를 발견하기 위해서는 학습 데이터의 제약에서 벗어나 데이터를 분석할 수 있는 분석 기법이 필요하다. 구글을 비롯한 해외 선도 집단에서는 이미 기 확보한 학습 모델을 활용해 새로운 모델을 학습하거나 시뮬레이션을 통해 해결하려는 문제 환경을 구성한 후 시뮬레이션 환경에서 모델을 확보하려는 접근법을 이용해 데이터의 제약에서 벗어나기 위해 다양한 가능성을 시도하고 있다. 이미 확보된 데이터를 분석해 새로운 가치를 찾아내는 것도 중요하지만 데이터의 활용도를 높여 보다 큰 부가가치를 창출하기 위해서는 새로운 관점에서 데이터를 통합하고 분석할 수 있는 인공지능 및 고급 머신러닝 기술의 개발 노력을 계속해야 할 것이다.

* 네거티브 전이: 관련성이 낮은 태스크간에 전이학습을 시도하면 도리어 성능이 악화되는 현상을 의미한다

제2장

빅데이터와 클라우드 기술의 컨버전스

클라우드 서비스의 고도화로 혜택을 입는 대표적인 서비스 중 하나는 빅데이터 기술이다. 대용량 데이터를 분산 시스템에서 처리하는 빅데이터 기술은 성격상 클라우드 기술과 궁합이 맞는다. 본 장에서는 컨테이너 기반 가상화 기술인 도커(Docker)의 기술적인 특징과 빅데이터 기술인 하둡의 최신 버전인 3.0의 중요 변화사항에 대해 자세히 알아본다.

1. 개요

클라우드 기술의 빠른 발전으로 인해 소프트웨어 기술의 판도가 바뀌고 있다. 클라우드 서비스의 활성화 정도에 따라 IT 업계의 순위도 변화하고 있다. 아마존의 제프 베조스(Jeff Bezos)가 세계 2위 부호로 등극하고 마이크로소프트 주식이 최고 수준을 돌파하는 직접적인 이유다. 이에 따라 오라클과 같은 전통적인 SW 강호들이 잇달아 클라우드 서비스에 출사표를 던지고 있다. 클라우드 서비스의 승자가 전체 IT의 지형을 바꿀 것으로 예상됨에 따라 클라우드의 대표적인 벤더인 아마존의 AWS(Advanced Web Services) 외에도 마이크로소프트도 자사의 클라우드 서비스인 애져(Azure)를 집중적으로 육성하면서 클라우드 서비스 가격 하락에도 불구하고 서비스 품질과 성능은 지속적으로 향상되고 있다.

클라우드 서비스의 고도화로 인해 혜택을 입는 대표적인 서비스로는 빅데이터 기술을 들 수 있다. 대용량 데이터를 분산 시스템에서 처리하는 빅데이터 기술은 성격상 클라우드 기술과 궁합이 잘 맞는다. 빅데이터 시스템은 분석에 필

* 필자: 하석재(한양대학교 컴퓨터소프트웨어학부 산업협력중점교수)

요한 빅데이터 클러스터의 성능과 사용할 시스템 수를 정하고 구축해 빅데이터 애플리케이션을 실행하고, 각각의 결과를 취합하는 과정을 거치게 된다. 클라우드 서비스를 사용하면 요구하는 성능의 시스템 자원(resource)을 프로비저닝(provisioning)을 통해 빠르게 구축해 빅데이터 애플리케이션을 실행할 수 있다. 또한 시스템 자원을 시간 단위로 이용 가능해 비용절감 효과도 뛰어나다. 아마존 AWS에서는 빅데이터 기술인 하둡을 PaaS(Platform as a Service) 형태로 제공하는 EMR(Elastic MapReduce) 서비스를 제공하고 있다. 마이크로소프트에서도 Azure HDInsight를 통해 제공 중이다.

한편 클라우드 구축을 하는 데 있어서 핵심 요소기술(enabling technology)이라고 할 수 있는 하이퍼바이저(예: 오픈스택의 KVM)의 약점인 오버헤드를 효과적으로 줄일 수 있는 컨테이너 기반의 가상화 기술이 등장하여 호응을 얻어가고 있다.

컨테이너기반 기술은 하이퍼바이저 기반 가상화와 비교해 전가상화나 반가상화(Full Virtualization/Para-Virtualization)가 아닌 실행환경의 격리(isolation)만을 제공한다. 이를 통해 시스템 효율성을 높이면서도 VM(virtual machine)과 유사한 환경을 제공한다.

이 글에서는 컨테이너 기반 가상화 기술인 도커(Docker)의 기술적인 특징과 빅데이터 기술인 하둡의 최신 버전인 3.0의 중요 변화사항에 대해 자세히 알아보도록 하겠다.

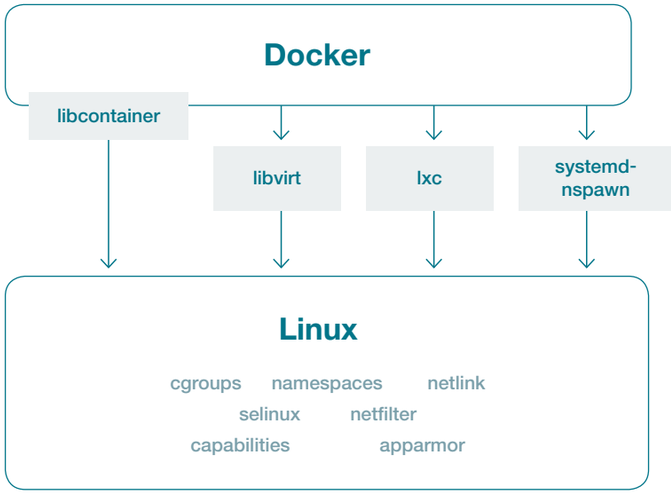
2. 컨테이너 기반 가상화 기술: 도커

클라우드를 구축할 때 기존에 사용하던 VMWare와 같은 하이퍼바이저보다 오버헤드가 낮으면서도 각 인스턴스 간의 실행환경에 대한 간섭을 없앨 수 있는 도커 기술의 인기가 매우 높아지고 있다. 현재 아마존에서는 Amazon EC2 Container Registry, Amazon EC2 Container Service와 같은 형태로, 마이크로소프트에서는 Azure Container Registry, Azure Container Service라는 이름으로 서비스가 제공되고 있다.

처음에는 실행환경의 격리를 위해 우분투(Ubuntu) 리눅스에서 제공하던 LXC(Linux Container) 기술을 사용했으나 정식 버전부터는 리눅스에서 모두 사용

이 가능한 libcontainer라는 드라이버로 대체해 RHEL(Red Hat Enterprise Linux), Fedora, CentOS와 같은 다양한 리눅스에서도 사용 가능하게 됐다. 현재는 runC로 통합됐다.

[그림 6-2-1] 도커 실행 환경



출처: http://www.zdnet.com/article/docker-libcontainer-unifies-linux-container-powers

도커는 클라우드 서비스인 PaaS(Platform as a Service)를 구축하기 위한 최적의 기술이다. 도커 자동화 기술인 Dockerfile을 사용하면 빅데이터 기술인 하둡이나 스파크, 자바 외에 필요한 소프트웨어가 설치된 이미지를 기반으로 한 컨테이너를 Large-scale(많은 수의 인스턴스 생성)로 실행하기가 매우 용이하다. 도커와 같은 컨테이너 기반 서비스를 사용하면 대량의 컨테이너를 빠른 시간 안에 배치(deploy)할 수 있다. 여기에 최근에 유행하는 개발과 서비스를 분리하지 않고 개발과 서비스를 빠르게 진행하는 DevOps 경향에도 맞는 기술이라고 할 수 있다.

도커의 구성 요소

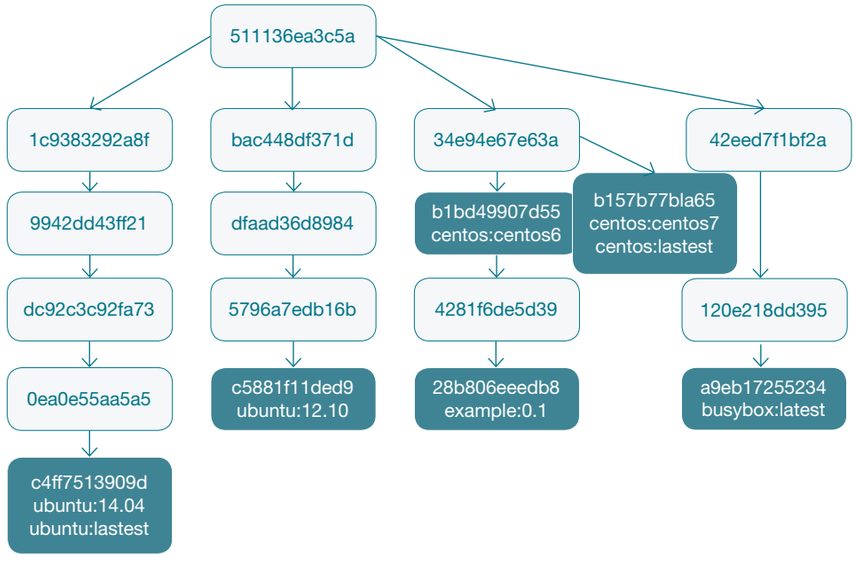
도커의 기술의 핵심은 LXC나 Libcontainer와 같은 실행환경 격리 기술과 자체

파일 시스템인 AUFS(Advance multi layered Unification File System)이다.

도커 버전 0.9부터는 libcontainer를 기본 실행환경으로 제공한다. LXC 외에도 libvert, lmctfy와 같은 다양한 격리 기술(isolation technology)을 지원해 특정 기술에 대한 의존도를 감소시켰다.

파일 시스템으로는 AUFS를 사용하는 데 브랜치(branch)를 관리해 서로 다른 이미지에서 공통되는 부분을 공유하고 차이만을 따로 관리하는 시스템이다. 이는 git과 같은 소스 버전 컨트롤과 유사한 방식으로 관리한다고 볼 수 있다.

[그림 6-2-2] AUFS 파일 시스템



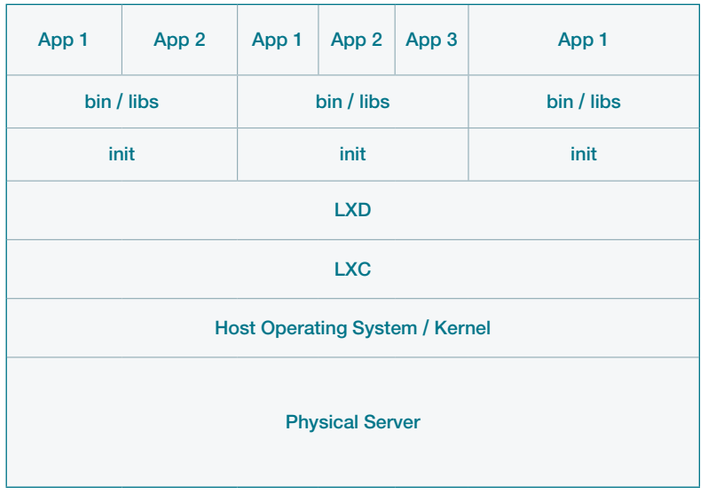
출처: http://pyrasis.com/book/DockerForTheReallyImpatient/Chapter01/02

또한 깃허브(http://github.com)와 유사한 도커허브(http://hub.docker.com)를 운영해 소셜 기능과 더불어 이미지 검색 및 공유가 가능하도록 했다.

3. 캐노니컬의 새로운 기술: LXD

우분투를 만든 캐노니컬에서는 LXC를 기반으로 더욱 발전된 형태의 컨테이너 하이퍼바이저(container hypervisior)인 LXD를 발표하고, 오픈스택의 기본 하이퍼바이저인 KVM과 적극적으로 경쟁에 나서고 있다.

[그림 6-2-3] LXD 구조



출처: https://hashprompt.blogspot.kr/2016/05/laymans-guide-for-lxd-canonicals.html

LXD는 LXC를 구성요소로 하는 컨테이너 기술로 보안기능 강화(Security by default)를 주요 특징으로 한다. 기존의 도커에서는 컨테이너를 생성하려면 관리자(루트) 권한으로만 가능했으나 LXD는 unprivileged 컨테이너를 지원해 관리자가 아니더라도 컨테이너를 생성할 수 있다. 아파치(Apache)나 엔진엑스(Nginx) 웹 서버와 같은 서버들이 기존의 루트 계정으로 실행하다가 해킹으로 인한 관리자 권한 획득 사례가 많아져 현재는 www나 www-data 계정으로 실행하는 것과 유사한 이유다.

캐노니컬에서는 도커를 애플리케이션 컨테이너(Application Container)로,

LXD는 머신 컨테이너(Machine Container) 기술로 부르면서 서로 다른 기술로 구분하고 있다. 캐노니컬에서는 LXD를 사용해 도커 컨테이너를 실행하는 것을 권장하고 있다. 이를 통해 도커가 갖는 장점과 LXD의 보안기술 측면에서의 장점을 함께 이용할 수 있다.

캐노니컬의 LXD가 노리는 시장은 오픈스택에서 사용하는 하이퍼바이저인 KVM을 대체하는 것이다. 이러한 접근은 기존의 VMWare와 같은 하이퍼바이저에 익숙한 사용자들에게 상대적으로 낯선 도커와 같은 컨테이너 기술 사이에서 대안으로 자리매김할 가능성도 있다.

LXD의 구성 요소

LXD는 세 개의 컴포넌트로 이루어져 있다. 각각 lxd 데몬(daemon), 명령어라인 클라이언트(command-line client)인 lxc, 오픈스택 Nova 플러그인인 nova-compute-lxd이다.

LXD는 RESTful API로 사용 가능하며 lxc위에 새로운 사용자 경험(User Experience)을 제공한다. 즉 내부적으로는 LXC를 기반으로 하는 기술이면서 LXD는 서버 형태로 제공되는 서비스로 볼 수 있다. 여기에 오픈스택의 하이퍼바이저를 교체할 수 있는 플러그인으로 구성돼 있다.

컨테이너 기반 가상화 기술의 최대 효용성이라면 게스트 OS 지원의 다양함을 포기한 대신 10%에 달하는 오버헤드를 3% 내외로 줄인 것이다. 리눅스 전용 기술인 도커는 각 컨테이너 간의 의존성(dependency) 문제를 해결하기 위해 우분투의 LXC(Linux Container) 기술을 근간으로 하는 libcontainer 드라이버를 적용했다.

정식 버전이 나온 지 3년 남짓 된 기술이지만 구글과 같은 대표적인 클라우드에서 적용돼 사실상의(de facto) 컨테이너 가상화 기술 표준으로 자리잡았다.

LXD는 도커와 서로 유사하지만 목표점이 다른 기술이 됐다. 도커는 실행환경의 분리를, LXD는 가벼운 하이퍼바이저로 발전하고 있다. 그리고 LXD는 현재로서는 우분투 서버에 통합·제공돼 서버로 반드시 우분투를 사용해야 하는 의존성이 있다. 도커는 전술한대로 libcontainer를 이용하여 우분투 의존성를 개선해 리눅스인 경우에는 모두 사용 가능하다.

4. 빅데이터의 대표 기술: 하둡3

분산파일 시스템인 HDFS(Hadoop Distributed File System)와 분산처리 프레임워크인 맵리듀스(MapReduce)로 되어 있는 하둡이 기존의 문제점들을 개선해 올해 정식버전(3.0)이 출시될 예정이다. 하둡3의 주요한 변경사항은 다음과 같다.

HDFS Erasure Coding은 데이터 블록을 여러 데이터 노드에 나눠 저장함에 따라 발생했던 스토리지 용량 문제를 해결하는 접근으로 RAID 5에서 사용한 것과 같은 패리티 기법을 HDFS에 적용해 스토리지 저장 효율을 50% 이상 끌어올렸다. 물론 네트워크 오버헤드와 CPU 계산 오버헤드는 감안해야 한다.(정확히는 3배의 복제수에 1.4배 오버헤드)

YARN Timeline Series는 애플리케이션의 정보를 Resource Manager(자원관리자)의 history store에 저장해 애플리케이션의 다양한 정보(컨테이너 이벤트, 메트릭, 맵리듀스 태스크의 수 등)를 확인할 수 있다. Yarn Timeline Series v.2에서는 자원관리자의 history store를 HBase를 사용해 고가용성과 확장성을 동시에 만족시켰다. 그리고 플로우(flow)와 애그리게이션(aggregation)까지 지원해 YARN 애플리케이션들의 단계별 상태 정보를 수집·관리하고 최적화가 가능하도록 했다.

이외에도 맵리듀스에서 시간이 많이 걸렸던 셔플링 작업과 관계된 map output collector를 JNI를 통한 C++ 코드로 대체해 성능을 30% 이상 개선했다. 이를 통해 셔플링 작업이 많은 애플리케이션의 속도 개선을 기대해 볼 수 있다.

그리고 액티브(Active)한 네임노드가 두 개 이상 가능하도록 해서 네임노드가 Active-Standby 교체 시 발생하는 지연에 따른 문제점을 개선했다. 이는 기존의 하둡2에서 네임노드 HA를 구축해 운영하더라도 특정 시점의 액티브한 네임노드는 하나이고 이를 변경해야 할 때 생기는 서비스 불능 시간을 개선하기 위해 추가되었다. 네임노드의 SPOF(Single Point of Failure) 문제를 3.0에서야 진정으로 해결했다고 볼 수 있다.

하둡3는 자바 8 기반으로 작성되었고, 셸 스크립트(shell script)를 전면 재작성했다. 또한 기존의 임시포트 대역(32768-61000)의 포트 번호를 새롭게 변경해 다른 애플리케이션과 충돌하는 현상을 개선했다. 여기에 데이터 노드의 내부 디스크 밸런서를 개선해 디스크 단위로 데이터의 불균등 저장 문제를 해결했다.

하둡3에서는 기존에 문제가 되던 사항에 대한 대대적인 개선이 이뤄졌다. 특

히 초기부터 계속 문제가 되어 왔던 네임노드의 장애를 최소화하도록 개선됐고, 데이터 손실에 대한 대비를 위해 필요한 디스크 요구사항을 최대한 줄였다. 여기에 맵리듀스 작업에서 문제가 되던 리듀서의 동작 성능을 개선하기 위한 결과가 반영됐다.

이를 통해 강력한 경쟁자인 아파치 스파크(Apache Spark)와의 치열한 주도권 싸움에서 승리할 수 있을지 주목된다.

5. 맺음말

본 장에서는 클라우드 기술의 요소 기술인 컨테이너 기반 가상화 기술에 대해 다뤘다. 아울러 빅데이터의 간판 기술인 하둡의 최신버전인 3.0의 주요 개선사항에 대해서 알아보았다.

빅데이터를 운영하기 위해 클라우드 서비스(정확하게는 PaaS)를 구축해 사용하면 비용이나 시간절약 등 여러가지 이점이 있다. 특히 클라우드 서비스를 컨테이너 기반 가상화 기술(주로 도커)을 사용해 구축하면 시스템 부하를 최소화하면서 자동화를 통한 빠른 인스턴스를 구축하는 것이 가능하다. 다시 말해 최적의 서비스 구축이 가능하게 된다. 이와 같이 빅데이터관련 서비스는 주류로 떠오른 클라우드 기술의 주요 서비스로 자리매김하게 될 것으로 예상된다.

제3장

NewSQL

지난 40여 년 동안 SQL 기반의 관계형 데이터베이스가 데이터를 저장하는 최적의 기술로 인정을 받았다. 그러나 사진, 동영상, 로그와 같은 비정형 데이터의 등장과 대용량의 빅데이터 처리의 필요성이 대두되면서 쉬운 비정형 데이터 처리와 확장성을 강조한 다양한 NoSQL 데이터베이스 시스템이 등장했고, 대부분 오픈소스 프로젝트에서 출발한 시스템이므로 저렴하게 데이터를 처리할 수 있다는 점에서 인기를 끌었다. 하지만 NoSQL이 한계를 드러내면서 기존 SQL(OldSQL) 기반의 관계형 데이터베이스 장점과 확장성, 분산 처리와 같은 NoSQL의 장점을 결합한 NewSQL이 등장하게 됐다.

1. NewSQL의 등장 배경

전통적인 SQL 데이터베이스는 수십 년 동안 사용돼 왔으며 오늘날 우리가 사용하는 거의 모든 응용 프로그램의 핵심 기반이 됐다. 이를 위해 구조화한 테이블 형태로 데이터를 저장하고 SQL 질의를 사용해 저장된 데이터에 대한 접근을 수행한다. 이 전통적인 관계형 데이터베이스시스템(RDBMS)을 뒤에 나올 다른 용어들과의 차이를 두기 위해 'OldSQL 시스템'이라고 부르기도 한다. OldSQL 시스템들은 디스크 기반의 데이터베이스 기술 및 표준 SQL을 사용하며, 임의 질의(ad-hoc query)를 지원한다.

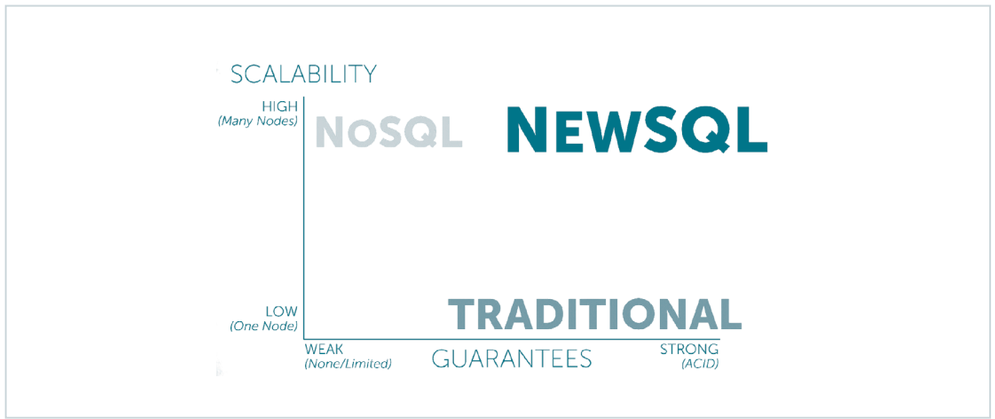
빅데이터와 사물인터넷(IoT) 기술 분야의 급속한 발전으로 인해 다양한 형태의 정형/비정형 데이터에 대한 효율적인 저장과 처리에 대한 요구가 증대됐다. 이에 따라 기존의 관계형 데이터베이스 시스템의 대안 성격으로 분할내성(Partition Tolerance)을 기본적인 특성으로 갖는, 즉 분산 구조를 채택한 다양한 종류의 NoSQL 데이터베이스가 등장했다. NoSQL 데이터베이스 시스템들은 스키마가

* 필자: 권준호(부산대학교 전기컴퓨터공학부 교수)

없으며(schema-free), 쉬운 복제를 지원하고, 간단한 API를 사용하며, RDBMS의 ACID*에 대비해 BASE** 특성을 가지고 있다. 많은 NoSQL 시스템들은 로그 메시지, XML과 JSON 문서, 구조화되지 않은(unstructured) 문서와 같이 비 관계형 데이터를 효율적으로 지원하도록 최적화돼 있다.

NewSQL이라는 용어는 2011년 451 Group의 애널리스트인 매트 애슬럿(Matt Asslet)이 처음 사용한 것으로, 그는 NoSQL과 같은 확장성과 유연성을 가지고 있으면서 SQL 질의와 ACID 특성에 대한 지원을 유지하고, 적절한 작업 부하에 대한 성능을 향상시키는 시스템을 NewSQL이라 정의***했다. 미국 MIT의 마이클 스톤브레이커(Michael Stonebraker) 교수는 NewSQL 데이터베이스는 주 인터페이스로서 SQL의 사용, 트랜잭션 처리를 위한 ACID 특성, 락을 사용하지 않는 병행 제어(non-locking concurrency control), 노드 단위의 고성능(high per-node performance), 확장성이 있는 shared-nothing 아키텍처(Scalable, shared nothing architecture)를 가지고 있는 시스템이라고 정의****했다. 전통적인 SQL(oldSQL), NoSQL과 NewSQL의 특성을 비교하면 [그림 6-3-1]과 같다.*****

[그림 6-3-1] OldSQL, NoSQL, NewSQL 비교



* ACID: Atomicity(원자성), Consistency(일관성), Isolation(독립성), Durability(지속성)

** BASE: Basically Available(기본적으로 가용적), Soft-state(소프트-상태), Eventually Consistency(점진적인 일관성)

*** https://www.451research.com/report-short?entityId=66963

**** http://cacm.acm.org/blogs/blog-cacm/109710

***** Júlio Alcântara Tavares, Angelo Brayner, José Maria Monteiro (2016), *What Comes After NoSQL? NewSQL: A New Era Of Challenges in DBMS Scalable Data Processing*, SSBD

[표 6-3-1] 세 종류 SQL 시스템의 특성 비교

특성	전통적인 SQL 시스템	NoSQL 시스템	NewSQL 시스템
ACID 특성	○	×	○
인메모리 DB	×	○	○
빅데이터	×	○	○
RDBMS	○	×	○

2. NewSQL 데이터베이스 분류

NewSQL 시스템을 크게 구분한다면 다음과 같은 세 범주로 분류할 수 있다.

- 새로운 아키텍처를 사용해 처음부터 새로 구축한 시스템
- 2000년대에 개발된 샤딩(sharding) 인프라를 다시 구현한 미들웨어
- 클라우드 컴퓨팅 공급자가 제공하는 새로운 아키텍처 기반 DaaS(Database-as-a-Service)

가. 새로운 아키텍처 시스템

이 범주는 처음부터 새로 구축된 데이터베이스관리시스템(DBMS)을 포함하므로 가장 흥미로운 NewSQL 시스템들을 포함하고 있다. 기존의 RDBMS를 확장하는 대신, 레거시 시스템의 아키텍처 없이 새로운 코스베이스로 설계됐다. 이 범주에 속하는 NewSQL 시스템은 shared-nothing 자원들 위에 동작하는 분산 아키텍처에 기반을 두고 있으며, 다중 노드 병행 제어(concurrency control), 복제를 통한 결함 허용(fault tolerance through replication), 흐름 제어 및 분산 질의 처리를 지원하는 구성 요소들을 포함한다.

분산 실행을 위해 만들어진 NewSQL 시스템을 사용하면 질의 최적화와 노드들 간의 통신 프로토콜을 포함한 시스템의 모든 부분을 다중 노드 환경에 맞게 최적화 할 수 있다는 장점이 있다. 예를 들면 미들웨어 시스템처럼 중앙 위치로 보내지 않고, NewSQL 시스템은 질의 내의 데이터(intra-query data)를 노드

들에게 직접 보낼 수 있는 기능을 제공한다. Clustrix, CockroachDB, Google Spanner, H-Store, HyPer, MemSQL, NuoDB, SAP HANA, VoltDB 등이 이 범주에 속한다.

나. 투명한 샤딩 미들웨어(Transparent Sharding Middleware)

현재 eBay, Google, Facebook 및 기타 회사가 2000년대에 개발한 것과 같은 종류의 샤딩 미들웨어를 제공하는 제품들이 있다. 이 제품들은 한 조직이 하나의 데이터베이스를 단일 노드 DBMS들로 구성된 클러스터에 저장되는 여러 조각(shards)으로 데이터베이스를 분할할 수 있게 해준다.

샤딩은 각 노드가 동일한 DBMS를 실행하고, 전체 데이터베이스의 일부분만을 가지며, 별도의 응용 프로그램에 의해 독립적으로 액세스되고 업데이트되지 않기 때문에 1990년대의 데이터베이스 연합 기술과는 다르다. 샤딩 미들웨어를 사용할 때의 주요 이점은 기존의 단일 노드 DBMS를 이미 사용하고 있는 응용 프로그램을 어떠한 코드도 변경하지 않고 대체할 수 있다는 점이다. AgilData Scalable Cluster, MariaDB MaxScale, ScaleArc, ScaleBase 등이 이 범주에 속한다.

다. DaaS

마지막 범주는 클라우드 컴퓨팅 공급자가 제공하는 NewSQL DaaS(Database-as-a-Service) 제품들이다. 이러한 서비스들을 사용하면, 특정 조직은 자체 하드웨어 또는 클라우드 호스트 가상 시스템에서 DBMS를 유지할 필요가 없다. 대신 DaaS 제공자는 시스템 튜닝(예: 버퍼 풀), 복제 및 백업을 포함한 데이터베이스의 물리적 구성에 대한 책임을 진다. 고객들은 대시보드 또는 시스템을 제어하는 API와 함께 DBMS에 대한 연결 URL을 제공받는다. Amazon Aurora, ClearDB 등이 이 범주에 속하는 시스템들이다.

3. NewSQL 시스템의 특징*

가. 메인 메모리 저장

메인 메모리 저장소 아키텍처를 기반으로 하는 NewSQL DBMS는 학계의 H-Store, HyPer와 상업용의 MemSQL, SAP HANA, VoltDB를 비롯해 여러 시스템이 존재한다. 이 시스템들은 메인 메모리 기반 특성으로 인해 OLTP 작업 부하에 대해 디스크 기반의 DBMS보다 훨씬 우수한 성능을 발휘한다.

데이터베이스를 전체적으로 메인 메모리에 저장한다는 아이디어는 새로운 것이 아니지만, 메인 메모리 NewSQL 시스템의 새로운 점 중 하나는 데이터베이스의 하위 집합을 영구 저장소로 내보내 메모리 풋 프린트를 줄이는 것이다. 이를 통해 DBMS는 디스크 지향 아키텍처로 다시 전환하지 않고 사용할 수 있는 메모리 양보다 큰 데이터베이스를 지원할 수 있다.

나. 파티셔닝/샤딩

거의 모든 분산 NewSQL DBMS가 스케일 아웃(scale out)하는 방식은 데이터베이스를 partition이나 shard라고 불리는 공통된 부분이 없는 부분 집합(disjoint subset)으로 분리하는 것이다.

다. 병행 제어

병행 제어(concurrency control)는 전통적인 트랜잭션 처리 DBMS 시스템의 거의 모든 측면에 영향을 미치기 때문에 가장 현저하고 중요한 사항이다. 1970~1980년대의 첫 번째 분산 DBMS들은 2단계 로킹(two-phase locking, 2PL) 방식을 사용했다. 새로운 아키텍처 기반의 거의 모든 NewSQL 시스템은 교착 상태 처리의 복잡성 때문에 2PL을 사용하지 않고, 직렬화 된 순서를 위반하는 인터리브된(interleaved) 연산들을 실행하지 않는다는 가정을 따르는 타임 스탬프 정렬(TO) 병행 제어나 그 변형 기법을 사용한다. NewSQL 시스템에서 가장 널리 사용되는 방법은 트랜잭션에 의해 업데이트 될 때 DBMS가 데이터베이스에 투플(tuple)의 새 버전을 생성하는

●○ NewSQL 시스템에서 가장 널리 사용되는 방법은 트랜잭션에 의해 업데이트 될 때 DBMS가 데이터베이스에 투플의 새 버전을 생성하는 분산 다중 버전 병행 제어 기법이다.

* Andrew Pavlo, Matthew Aslett (2016), *What's Really New with NewSQL?* SIGMOD Record vol.45, no.2, pp.45-55의 내용을 발췌해 요약 정리한 것임.

분산 다중 버전 병행 제어 (Multi Version Concurrency Control, MVCC) 기법이다. 어떤 NewSQL 시스템들은 2PL과 MVCC를 조합한 방법을 사용한다.

라. 보조 인덱스

보조 인덱스(secondary index)는 기본 키가 아닌 테이블의 일반 속성들의 서브셋으로 만들어진 인덱스를 의미한다. 이를 통해 DBMS는 기본 키와 분할 키 룩업 이외의 보다 빠른 질의들을 지원할 수 있다. 새로운 아키텍처를 기반으로하는 모든 NewSQL 시스템은 분산된 시스템이고, 분할된 보조 인덱스를 사용한다. 각 노드는 전체 사본을 가지고 있는 것이 아니라 인덱스의 일부를 저장한다.

마. 복제

조직에서 고가용성 및 데이터 내구성을 보장하는 가장 좋은 방법은 해당 데이터베이스를 복제하는 것이다. 강한 일관성을 지원하는 DBMS에서 트랜잭션의 쓰기 연산은 트랜잭션이 커밋으로 간주되기 전에, 모든 복제본에서 승인이 되고 실행돼야 한다. 이 접근법의 장점은 복제본이 읽기 전용 쿼리를 제공하면서도 일관성을 유지할 수 있다는 것이다. 또한 복제본이 실패하더라도 다른 모든 노드가 동기화돼 있으므로 업데이트가 손실되지 않는다.

이 동기화를 유지하려면 DBMS가 모든 복제본이 트랜잭션의 결과와 일치하는지 확인하기 위해 원자 커밋 프로토콜을 사용한다. 이것은 DBMS 시스템의 추가적인 오버헤드이며, 노드가 실패하거나 네트워크 파티션/지연이 있는 경우에 멈추는 현상을 일으키기도 한다. 따라서 NoSQL 시스템들은 약한 일관성 모델을 지원하는 경향이 있다. 하지만 우리가 알고 있는 모든 NewSQL 시스템은 강력한 일관된 복제를 지원한다.

[그림 6-3-2]는 여러 NewSQL 시스템들을 위에서 서술한 특징 위주로 분석한 결과를 요약해 보여준다.

[그림 6-3-2] NewSQL 시스템의 특성 비교

		Year Released	Main Memory Storage	Partitioning	Concurrency Control	Replication	Summary
NEW ARCHITECTURES	**Clustrix** [6]	2006	No	Yes	MVCC+2PL	Strong+Passive	MySQL-compatible DBMS that supports shared-nothing, distributed execution.
	CockroachDB [7]	2014	No	Yes	MVCC	Strong+Passive	Built on top of distributed key/value store. Uses software hybrid clocks for WAN replication.
	Google Spanner [24]	2012	No	Yes	MVCC+2PL	Strong+Passive	WAN-replicated, shared-nothing DBMS that uses special hardware for timestamp generation.
	H-Store [8]	2007	Yes	Yes	TO	Strong+Active	Single-threaded execution engines per partition. Optimized for stored procedures.
	HyPer [9]	2010	Yes	Yes	MVCC	Strong+Passive	HTAP DBMS that uses query compilation and memory efficient indexes.
	MemSQL [11]	2012	Yes	Yes	MVCC	Strong+Passive	Distributed, shared-nothing DBMS using compiled queries. Supports MySQL wire protocol.
	NuoDB [14]	2013	Yes	Yes	MVCC	Strong+Passive	Split architecture with multiple in-memory executor nodes and a single shared storage node.
	SAP HANA [55]	2010	Yes	Yes	MVCC	Strong+Passive	Hybrid storage (rows + cols). Amalgamation of previous TREX, P*TIME, and MaxDB systems.
	VoltDB [17]	2008	Yes	Yes	TO	Strong+Active	Single-threaded execution engines per partition. Supports streaming operators.
MIDDLEWARE	**AgilData** [1]	2007	No	Yes	MVCC+2PL	Strong+Passive	Shared-nothing database sharding over single-node MySQL instances.
	MariaDB MaxScale [10]	2015	No	Yes	MVCC+2PL	Strong+Passive	Query router that supports custom SQL rewriting. Relies on MySQL Cluster for coordination.
	ScaleArc [15]	2009	No	Yes	Mixed	Strong+Passive	Rule-based query router for MySQL, SQL Server, and Oracle.
DBAAS	**Amazon Aurora** [3]	2014	No	No	MVCC	Strong+Passive	Custom log-structured MySQL engine for RDS.
	ClearDB [5]	2010	No	No	MVCC+2PL	Strong+Active	Centralized router that mirrors a single-node MySQL instance in multiple data centers.

출처: Andrew Pavlo, Matthew Aslett (2016), *What's Really New with NewSQL?* SIGMOD Record vol.45, no.2, pp.45-55

4. 맺음말*

지금까지 NewSQL이 등장한 역사적인 배경과 최근의 NewSQL 시스템에 대해 살펴보았다. NewSQL 데이터베이스 시스템은 기존의 시스템 아키텍처와의 근본적인 차이가 아니라 데이터베이스 기술의 지속적인 개발에서 다음 단계를 의미한다. NewSQL 데이터베이스가 채택하고 있는 기술들은 학계와 산업계에 존재하는 이전의 DBMS에 존재했던 기술들이다. NewSQL 데이터베이스를 혁신적이라고 판단할 수 있는 근거는 여러 기술을 하나의 단일 플랫폼에 동작하도록 구현했다는 점이다. 머지않아 1980~1990년대의 RDBMS, 2000년대의 OLAP 데이터 웨어하우스, 2000년대의 NoSQL 데이터베이스, 2010년대의 NewSQL 시스템의 4개 시스템들의 특성들이 하나로 통합될 것이다.

* 상게서 결론 요약 정리함

제4장

아파치 스파크와 리얼타임 빅데이터 분석

스파크(Spark)는 지난 몇 년간 빠르게 발전하면서 빅데이터 분야에서 가장 '핫'하고 모든 분야에 활용되는 중요한 오프소스로 성장했다. 이 글에서는 스트리밍 시스템 구축을 위해 고려해야 할 점들과 스톰을 만든 네이선 마츠가 고안한 람다 아키텍처를 중심으로 스파크를 이용한 분산 머신러닝과 딥러닝 기술 현황을 알아본다.

1. 실시간 시스템

가. 속도를 강조하는 실시간 시스템

실시간 시스템의 정의는 '사용할 수 있는 자원이 한정된 상황에서 작업이 요청되었을 때, 주어진 시간 안에 처리해 결과를 보내주는 것을 말한다. 작업 요청에서 수행 결과를 얻기까지의 시간적인 제약이 존재하는 시스템으로, 그 제약의 엄격함에 따라서 경성 실시간 시스템과 연성 실시간 시스템으로 나뉜다.'이다. 여기서 경성 실시간 시스템(hard real-time system)은 무기·우주선·철도 등 아주 미세한 시간까지 지켜야 하는 시스템을 의미한다. 또 연성 실시간 시스템(soft real-time system)은 데이터 통신·게임·TV 등 실시간으로 전달돼야 하지만 어느 정도 프레임이 밀리더라도 큰 문제가 없는 시스템을 의미한다.

그러나 흔히 빅데이터 분석 환경에서 말하는 실시간 시스템은 이와는 다르다. 물론 최대한 시간 오차가 없어야겠지만 대략적으로 예상되는 시간 내에 응답이 나오면 실시간 시스템이라고 부르는 경우가 많다. 빅데이터가 그렇듯이 약간은 마케팅적인 요소가 들어간 용어라서 일부는 엄격한 기준의 실시간 의미와는 맞지 않을 수 있고, 사람마다 기준이 조금씩 다르기도 하다.

일반적으로 빠르게 처리해주는 실시간과 스트리밍이 혼용되기도 하지만, 좀 더 명확히 이야기하자면 둘은 서로 다른 개념이다. 실시간 처리는 제한된 시간 안에 처리해야 하는 제약이 따르는 것을 말하고, 스트림 처리(스트리밍)는 데이터가 한정되지 않고(unbounded) 끊임없이 흘러가는(stream) 형태에 대한 처리 방식을 말한다. 위에 언급한대로 두 가지는 엄격히 다른 내용이지만, 이 글에서 실시간은 빠르게 처리하는 것을 통칭하고 필요한 경우 스트리밍이라는 단어를 언급하도록 하겠다.

나. 빅데이터 환경에서의 실시간 처리·분석

지난 10여 년 간 데이터 프로세싱 분야는 많은 발전이 있었다. 특히 하둡(Hadoop)은 대용량 데이터를 분산 저장·분석을 하는 데 있어서 혁명을 가져왔다. 하둡은 글로벌 IT 기업들에서 대부분 사용하고 있을 정도로 일반화가 되어 국내에서도 많은 기업과 기관에서 사용하고 있다. 그러나 하둡은 배치성 처리에 특화된 시스템으로 실시간 처리에는 적합하지 않다.

보통 매번 요청이 있을 때마다 전체 데이터를 읽어서 맵리듀스(MapReduce)를 수행한다는 것은 매우 비효율적이다. 그래서 많은 기업이 중요한 데이터에 대해 맵리듀스로 매일 혹은 매시간 모든 페이지의 페이지 뷰를 구하는 것과 같은 배치 작업을 수행한다.

그래서 나온 것이 최근 이슈가 되고 있는 SQL On Hadoop 계열이나, 인메모리 처리, 컬럼 방식의 분산 DB 등이다. 분명 이 오픈소스들은 메모리 활용과 색인 등을 통해 빠른 응답성을 지원하지만, 필요할 때마다 대용량 데이터를 처리해야 하는 것은 마찬가지다. 또한 일반적인 방식에서는 데이터가 계속 들어오는 것이 아니라, 배치성으로 들어오므로 최신의 데이터는 분석하지 못한다. 배치는 결국 디스크에 저장하고 그 이후에 처리 작업을 지시해야 한다.

데이터가 Continuous, unbounded, rapid, time-varying stream과 같은 속성을 가진다면, 기존의 맵리듀스와 같은 패러다임이 아닌 새로운 접근이 필요하다. 이와 같은 데이터들은 여러 분야에서 많이 발생하지만, 위와 같은 한계 때문에 상당수가 버려졌다(물론 HDFS의 출현에 따라 저장 비용이 하락해 충분히 저장 가능).

* 필자: 이상훈(한국스파크 사용자모임 대표, SK C&C 데이터 기술팀 대리)

예를 들면 네트워크 모니터링, 센서 데이터, 통화 내역(Telecom Call Record), 웹로그, 생산관리 데이터, 금융 데이터 등이 이에 해당한다. 나열된 로그 예제에서는 최근 주목받는 사물인터넷(IoT) 관련 분야나 제조 분야 등이 포함된다.

2. 빅데이터를 위한 스트리밍 시스템

실시간 처리에서 스트리밍을 위한 시스템을 구축한다는 것은 단순히 데이터 처리 방식이 바뀌는 것만이 아니라 나머지 여러 부분까지 고려해야 한다. 빅데이터 스트리밍 시스템 구축을 위해 고려해야 할 점들을 가장 대표적인 오픈소스인 스파크와 스톰을 비교해 소개하면 다음과 같다.

●○
실시간 처리에서 스트리밍을 위한 시스템을 구축한다는 것은 단순히 데이터 처리 방식이 바뀌는 것만이 아니라 나머지 여러 부분까지 고려해야 한다.

가. 데이터 처리 방식

1) 레코드 단위 즉각 처리

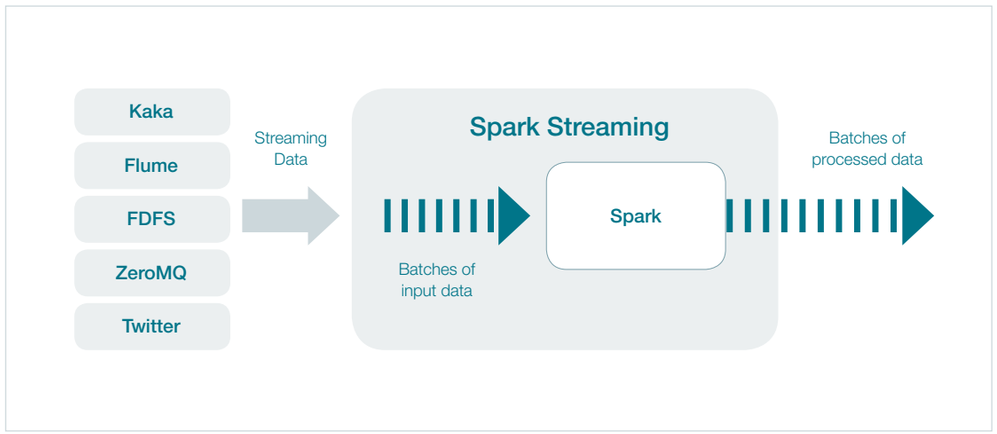
이 방식은 스톰이 주로 채택한 방식으로 데이터가 발생하면 레코드 단위로 내부적으로 받아서 즉시 처리하는 것(Native Stream)으로, 대부분의 로직을 구현할 수 있다. 그러나 모든 레코드를 각각 처리하기 때문에 처리량(throughput)이나 장애 처리에 대한 단점을 갖고 있고, API 추상화가 어렵기 때문에 구현 시 공수가 많이 든다.

2) 작은 단위 배치 처리

레코드 단위 즉각 처리(Native Stream)와는 달리 짧은 주기로(초 단위) 모아서 처리하기 때문에 마이크로 배치(Micro Batch)라는 이름이 붙었다. 장점으로는 추상화 API가 상대적으로 쉽고 장애복구, 로드밸런싱이 용이하다. 하지만 데이터를 묶다 보니 특정 로직은 구현하기 상대적으로 어렵다.

스파크에서는 [그림 6-4-1]과 같이 데이터가 스트림 형태로 들어오면 스파크 스트리밍 모듈에서 마이크로 배치 단위로 일정하게 나눠주고, 스파크 엔진에서 해당 마이크로 배치들에 대해 필요한 로직을 적용하는 방식이다.

[그림 6-4-1] 스파크의 스트림 처리



나. 장애 방지

스트리밍 시스템에서 중요한 점은 장애가 발생하지 않아야 한다는 것이다. 배치 시스템의 경우 장애가 발생하면 다음 배치 처리 전까지만 복구하면 되지만, 실시간 시스템은 장애가 발생하면 데이터 유실의 위험이나 실시간으로 처리해야 하는 데이터가 쌓이기 때문에 무중단 시스템 설계가 중요하다. 이에 대한 방식은 크게 다음 두 가지가 있다.

1) 레코드 추적(Ack)

스톰의 경우 ack라는 방식을 이용해 프로세스 끝까지 완료했는지 확인한다. 메시지 트리가 지정된 시간 안에 완전한 처리에 실패하면 해당 튜플은 실패 처리가 되며, 옵션에 따라서 큐(Queue)로부터 데이터를 다시 읽어오기도 한다. 그러나 이 모든 튜플을 모두 추적하게 되면 성능에 큰 영향을 미치기 때문에 64Bit Id와 XOR 연산을 이용해 전부 추적하지는 않으면서도 실패 시 재시도가 가능한 로직을 내부적으로 갖추고 있다.

2) 빠른 복구(Faster recovery)

스파크의 경우 RDDs(Resilient Distributed Datasets)라는 인메모리 데이터 형태를 기본으로 한다. 메모리는 휘발성이므로 데이터 유실에 취약하다고 생각할 수 있

지만, 데이터를 백업해두는 것이 아니라 데이터가 만들어지는 과정만 기록해 두는 형태로 데이터의 안정성과 속도 두 가지 장점을 모두 이용한다. 즉 내고장성(fault tolerance)을 위해 데이터 계보(Data Lineage)를 이용한 데이터 복구를 수행한다. 이를 토대로 각 로직이 만들어지는 계보를 내부적으로 기록해두었다가 데이터 복구 시 바로 이전 단계로부터 데이터를 생성해낸다. 최악의 상황에서는 맨 앞 단인 안정적인 저장공간(예: HDFS, Kafka 등)으로부터 데이터를 복원할 수 있다. 이와 같이 데이터 추적이나 Sync 등의 작업이 없기 때문에 데이터 유실 가능성은 낮추면서도 성능상 이점을 누릴 수 있다.

다. 성능 비교

성능 측정은 로직과 튜닝 방법에 따라 달라지므로 공평한 측정이 매우 어렵다. 그래서 공개된 각 오픈소스를 보면, 각각의 오픈소스마다 자신이 더 빠르다고 강조한다. 굳이 성능을 따진다면 처리량(throughput)과 지연·응답시간(latency) 두 가지를 비교해 볼 수 있다.

Micro-batch 시스템은 지연·응답시간에는 불리하고 처리량 측면에서는 유리하다. Native Streaming 시스템은 지연·응답시간 측면에서 유리하고 처리량 측면에서는 불리하다.

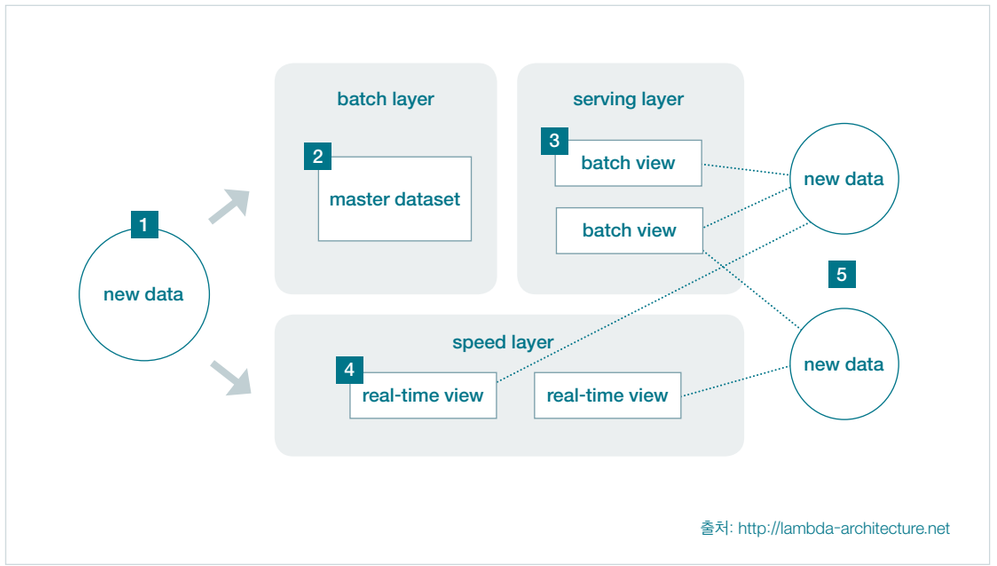
데이터의 전달 방식과 상태 관리, 내결함성(fault tolerance) 설정도 성능에 많은 영향을 미치는 요소다. 여기에 데이터 로컬리티(data locality), 그룹핑, 데이터 성향 등도 고려해야 한다. 따라서 어떠한 방법이 무조건 좋다기보다는 각 비즈니스 로직이나 제약 상황에 맞게 설계해야 한다.

3. 실시간 분석

가. 람다 아키텍처

람다 아키텍처(Lambda Architecture)는 스톰을 만든 네이선 마츠(Nathan Marz)가 빅데이터 시스템 경험을 바탕으로 범용, 확장 가능 및 내결함성 데이터 처리 아키텍처를 위해 만든 용어다. 이것은 크게 3개의 계층(layer)으로 이뤄진다. 제공 기능에 대해 정리해보면 다음과 같다.

[그림 6-4-2] 람다 아키텍처



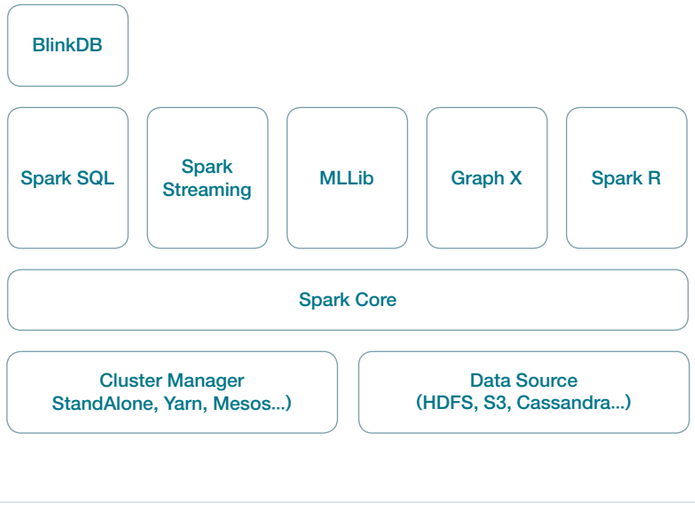
출처: http://lambda-architecture.net

- 모든 데이터는 전부 혹은 일부가 배치 계층과 스피드 계층으로 전달된다.
- 배치 계층은 마스터 데이터 세트(master dataset)를 관리하고, 전처리 혹은 미리 계산된(summary와 같은 로직) 배치 뷰를 지원한다.
- 서빙 계층는 low-latency, ad-hoc한 쿼리를 날릴 수 있도록 인덱싱된 뷰를 제공한다.
- 스피드 계층은 최근의 데이터(주로 스트리밍)에 대해 빠르게 처리한 리얼타임 뷰를 제공한다.
- 마지막으로 배치 뷰와 리얼타임 뷰를 조합한 결과를 얻을 수 있도록 한다.

[그림 6-4-2]의 람다(Lambda) 아키텍처는 스톰을 만든 네이선 마츠(Nathan Marz)가 고안한 그림이지만, 필자는 스톰보다는 스파크에 조금 더 어울리는 그림이라고 생각한다. 그 이유는 [그림 6-4-3]과 같이 스파크의 경우 배치성인 Programming(Core), SQL과 실시간성인 스트리밍에 대한 모듈을 제공한다. 또한 머신러닝을 위한 MLLib와 통계 쪽에서 유명한 R를 분산처리할 수 있는 스파크 R까지 거의 모든 영역을 커버하고 있다. 이것은 단순히 기능 제공 차원이 아니라, 하나의 일관된 방식으로 배치 분석(머신러닝 포함)과 실시간 분석을 가능하도록 해주

는 것을 의미한다. 실제로 스파크에서는 각 모듈에서 발생하는 결과를 쉽게 서로 연동 가능하도록 구조화돼 있다.

[그림 6-4-3] 스파크의 구조



나. 스파크를 이용한 분산 머신러닝

머하웃(Mahout)의 머신러닝은 하둡의 맵리듀스(MapReduce) 기반 하에 처리·계산된다. 그러나 맵리듀스는 반복적인 디스크 쓰기와 같이 속도 저하의 요인들이 있다. 머신러닝 특성상 반복(Iteration) 작업들이 많은데, 이러한 작업은 메모리를 사용하면 훨씬 빠르게 결과를 얻을 수 있다. 버클리 대학의 AMPLab(정확히는 그 전신)이 스파크를 만들게 된 계기도 맵리듀스를 이용한 분산 머신러닝 작업이 너무 느려 만든 것이다.

계속 이야기해 왔던 다른 채널 혹은 분석 방법과의 연계가 머신러닝에도 해당된다. 다른 프로그램과 마찬가지로 스파크 엔진 위에서 작동하며 스트리밍이나 SQL과 연동할 수 있다. 이것은 머신러닝이라는 것이 독립적인 분석이 아니라 데이터 전처리, 데이터 엔지니어링 작업 등이 동반해야 한다는 점에서 중요한 장점이다. 또한 스파크는 동일한 환경에서 여러 가지 분석 방법을 지원한다(same pipeline).

빅데이터에서 머신러닝의 강자는 원래 머하웃(Mahout)이었는데 맵리듀스라는 한계를 극복하지 못하고 내부 로직 개선, DSL(Domain-specific language), 스파크 전환 등을 하다가 2년 전부터 기여자가 많이 이탈했다. 이외에도 Sparkling Water, PredictionIO와 같이 스파크의 머신러닝 기능을 엔터프라이즈 환경에서 쉽게 쓸 수 있도록 분석 탬플릿, Rest API 제공 등을 해주는 오픈소스도 있다.

다. 스파크를 이용한 분산 딥러닝

딥러닝은 머신러닝 알고리즘 중 하나인 뉴럴 네트워크의 발전된 형태로 (정확히는 히든 레이어를 여러 층으로 쌓은 구조) 2~3년 전부터 뛰어난 성능 덕분에 굉장한 이슈가 되고 있다. 특히 우리나라에서는 작년에 이세돌 국수와 알파고의 승부 이후로 더 많은 관심을 받고 있으며, 산업계에도 이미 적용됐고 앞으로도 많은 영역에 적용 예정이다.

스파크는 이미 빅데이터 처리·분석 영역의 상당 부분을 커버하고 있고, 딥러닝에 대해서도 빠르게 분산처리하기 위한 많은 시도가 이루어지고 있다.

1) 스파크 기반의 기존 딥러닝 프레임워크

딥러닝 프레임워크에는 예전부터 많이 사용되던 Caffe, Torch 등이 있다. 또한 구글의 이름에 힘입어 최근 가장 이슈가 되고 있는 텐서플로(Tensorflow), MS의 CNTK 등이 있다. 딥러닝 연구자들이 새로운 API를 익히기 보다는 위와 같은 오픈소스가 갖고 있는 기존의 API를 최대한 활용하되 빅데이터를 처리하거나 모델을 학습시킬 때 가장 문제가 되는 분산처리를 도와주는 SparkNet(caffe), CaffeOnSpark, TensorFlowOnSpark와 같은 프레임워크도 있다. 특히 TensorFlowOnSpark의 경우에는 10라인 이하의 수정으로 텐서플로의 모든 기능을 사용할 수 있다. 이때 HDFS와의 연동, 파이프라인 제공, 쉬운 클라우드 deploy 기능 등을 제공한다.

오픈소스마다 조금씩의 차이는 있지만 대부분은 기존의 Caffe, 텐서플로와 같은 딥러닝 오픈소스 모듈을 스파크 위에 올려서 실행하도록 돼 있다. 좀 더 세부적으로 들어가면 파라미터 서버를 두어 데이터를 분산시켜서 학습시키는 방법과 모델 자체를 분산 학습할 수 있는 방법이 있다.

2) 스파크를 최대한 활용하는 프레임워크

Deeplearning4j는 자바(Java)와 스칼라(Scala)를 기반으로 만들어진 상용 수준의 딥러닝 오픈소스 프레임워크다. 다른 프레임워크와 달리, 연구 목적이 아닌 엔터프라이즈 환경을 위해 만들어졌으며(텐서플로는 중간에 걸쳐 있음), SkyMind라는 회사가 유료 기술지원도 한다.

모델의 자유도보다는 머신러닝, 딥러닝에 대한 지식이 부족한 사람도 쉽게 사용할 수 있도록 최대한 간단히 API를 만들었다. 물론 스파크 엔진을 기반으로 하므로 스파크의 데이터 구조와 연동이 매우 간단하며, 쉽게 분산 처리가 가능하다는 장점을 갖고 있다.

3) CPU를 최대한 활용하는 프레임워크

BigDL이라는 프레임워크는 앞서 설명한 것들과 같이 스파크 기반으로 작동하는 것은 동일하지만, 제온과 같은 서버 CPU 활용에 초점을 맞췄다는 점에서 차별점을 갖고 있다. 그 이유는 서버용 그래픽 카드는 일반 데스크톱에 비해 10배 가량 비싸고, 기존에 빅데이터 시스템이 구축된 경우 굳이 새로운 투자를 하기보다 성능을 조금 양보하고 기존 시스템을 최대한 활용하려는 기업들의 요구가 많기 때문이다.

이러한 환경 때문에 BigDL은 인텔이 주도해 인텔의 MKL과 멀티스레딩을 최대한 활용해 큰 모델의 경우 Caffe, Torch, 텐서플로보다 일부 빠른 결과를 얻기도 했다. 스파크를 이용해 주로 distributed synchronous, mini-batch SGD 작업을 수행하고, 하둡 에코 시스템과 연동하거나 다른 딥러닝 프레임워크에서 학습된 모델을 BigDL에서 불러오기 등 확장 기능에도 신경을 썼다.

4. 마치며

모든 방면에서 뛰어난 사람 혹은 소프트웨어를 찾기는 어렵다. 하지만 스파크는 지난 몇 년간 매우 빠른 발전을 하면서 빅데이터 분야에서 가장 핫하고 모든 분야에 활용되고 있는 중요한 오프소스로 성장했다. 여전히 발전 속도도 빠르다. 빅데이터 분석을 위해 꼭 하나의 오픈소스만 배워야 한다고 한다면 주저없이 스파크를 배워야 할 것이다.

제5장

데이터 시각화

가상현실과 인공지능과 같은 최신의 기술이 데이터 시각화와 결합하면서 새로운 차원의 시각화를 이끌고 있다. 가상현실에서 더욱 생생한 데이터 경험을 제공하는 몰입형 데이터 시각화가 더욱 중요해지고 있으며, 증강현실에서 더욱 정교한 시뮬레이션이 가능해짐으로써 과학 시각화가 의학, 공학, 교육 등의 분야에 적극적으로 활용되기 시작했다. 최근 몇 년 사이에 놀라운 성취를 이룬 인공지능은 딥러닝이 주도하고 있으며, 매우 복잡한 고차원 데이터를 다루는 딥러닝을 이해하고 구현하기 위해 고차원 데이터 시각화 기술이 다시 주목받고 있다. 나아가 인공지능이 스스로 의미있는 정보를 생성해내는 단계에 들어서면서 향후 데이터 시각화의 프로세스와 방법론에도 근본적 변화가 일어날 것으로 전망된다.

1. 데이터 시각화 기술 개요

데이터 시각화 기술은 과학기술의 발전과 사회적 변화에 많은 영향을 받으며, 오래된 기술과 최신의 기술이 공존하면서 다양한 유형과 수준에서 활용되고 있다. 전통적인 통계 분석 도구와 프로그래밍 환경의 발전과 더불어 전문가 도구 역시 더욱 세분화되고 정교해지고 있다. 데이터 분석 환경의 변화는 데이터 시각화 기술의 변화와 밀접하게 관련되며, 통계적 시각화에 적합한 기존의 방법에 비해 인터랙티브 시각화에 적합한 기술의 비중이 확대되는 경향을 보인다. 데이터 시각화가 전문가의 영역을 넘어서 대중적으로 저변이 확대되는 추세에 따라, 더 사용자 친화적인 대시보드를 제공하는 비즈니스 인텔리전스(BI)의 영향력이 더욱 증가하고 있다. 또한 세분화되고 전문적인 분야에서도 해당 분야에 특화된 데이터 시각화 기술을 통해 전문성을 더욱 강화하고 있다.

* 필자: 송한나(코그니텀랩 대표)

가. 데이터 시각화의 정의

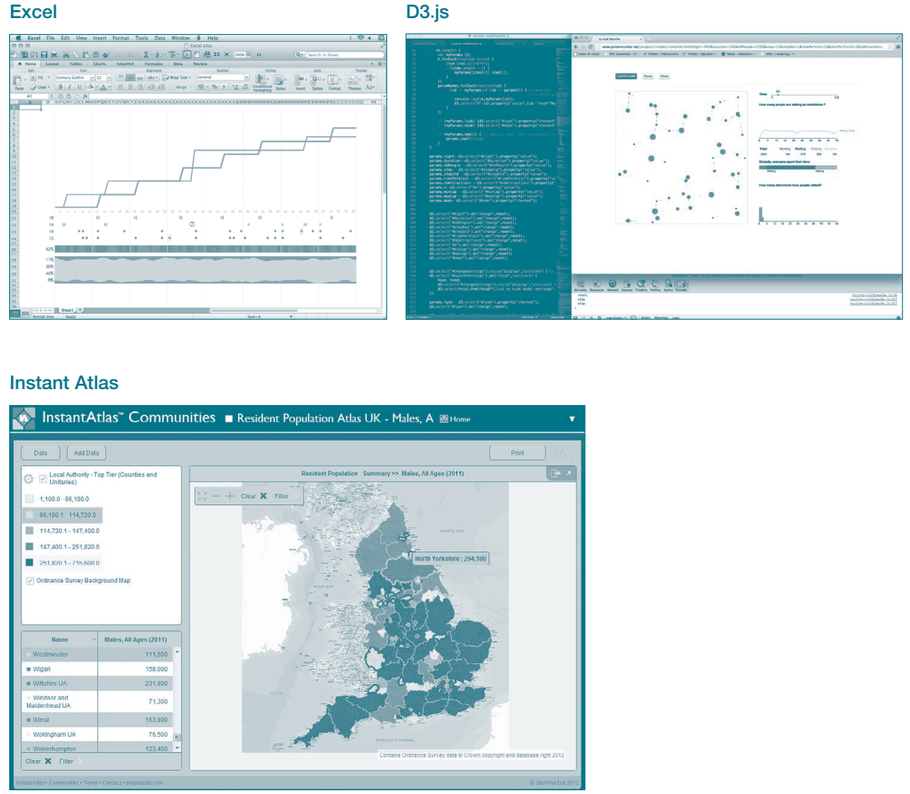
데이터 시각화(data visualization)는 '인지를 증폭시키기 위해 인간의 시지각 능력을 이용하는 데이터의 표현과 설명 방식(the representation and presentation of data that exploits our visual perception abilities in order to amplify cognition)'*으로 정의된다. 이 정의에 따르면 데이터 시각화는 다음과 같은 기능으로 이루어진다. 첫째, '데이터의 표현'은 선·막대·원과 같은 물리적 형태를 통해 데이터를 원료로 해 그 속성을 가장 잘 표현하는 방법을 뜻한다. 둘째, '데이터의 설명 방식'은 데이터의 표상을 넘어 색상, 주석 및 상호작용 기능의 선택을 포함하는 전반적인 의사소통 작업으로 통합하는 방법에 관련된다. 셋째, '시지각 능력의 이용'은 인간의 눈과 두뇌가 정보를 가장 효과적으로 처리하는 방법에 대한 과학적 이해와 관련된다. 넷째, '인지의 증폭'은 정보를 효과적이고 효율적으로 사고와 통찰력 및 지식으로 처리하는 능력을 최대화하는 방법과 관련된다.

나. 데이터 시각화 기술의 종류

데이터 시각화 기술은 매우 다양한 유형과 수준으로 활용되며, 모든 용도에 적합한 단일한 기술은 존재하지 않는다. 그러므로 각 기술이 가진 상대적인 강점과 약점, 기능과 목적을 확인하고 최신 기술 동향 등을 파악해 적절한 기술을 선택할 필요가 있다. 데이터 시각화 기술의 유형은 주요 목적에 따라 다음과 같이 분류될 수 있다. 첫째, 도표와 통계 분석 도구(charting and statistical analysis tools)는 주요 차트 생산성 도구와 강력한 시각화 기능을 제공하는 보다 효과적인 시각적 분석 또는 비즈니스 인텔리전스(BI) 응용 프로그램을 다룬다. 마이크로소프트 엑셀이나 태블로(Tableau)와 같은 프로그램을 예로 들 수 있다. 둘째, 프로그래밍 환경(programming environment)을 통해 시각화를 완벽하게 제어할 수 있으며, 이것은 차트·속성·이벤트 및 동작을 독창적으로 제어하기 위한 방법이다. D3.js와 프로세싱(processing), 그리고 R ggplot2를 예로 들 수 있다. 셋째, 전문가 도구(specialist tools)를 활용해 더 정교한 시각화를 제작하는 방법이 있다. 인포그래픽을 위해 어도비 일러스트레이터(Adobe Illustrator)를 활용하거나 모션그래픽을 위해 어도비 이펙트(Adobe Effect)를 활용하는 예를 들 수 있다. 또는 지리·공간 시각화를 위한 ArcGIS나 네트워크 시각화를 위한 지파이(Gephi)와 같이 특수한 종류의 시각화를 위해 제공되는 프로그램을 활용하는 예도 포함된다.*

* Andy Kirk (2012), *Data Visualization: A Successful Design Process*, Packt Publishing, pp.16-17

[그림 6-5-1] 차트와 통계 분석 도구(Excel), 프로그래밍 환경(D3.js), 전문가 도구(Instant Atlas)의 예



데이터 시각화 기술의 분류는 앞서 살펴본 것처럼 차트와 통계 분석 도구, 프로그래밍 환경, 전문가 도구로 크게 나눌 수 있는데, 서로 배타적인 것이 아니며 상호결합돼 기능을 보완하기도 하고 새로운 기술의 등장으로 인해 기존의 시각화 기술이 새로운 용도로 활용되기도 한다.

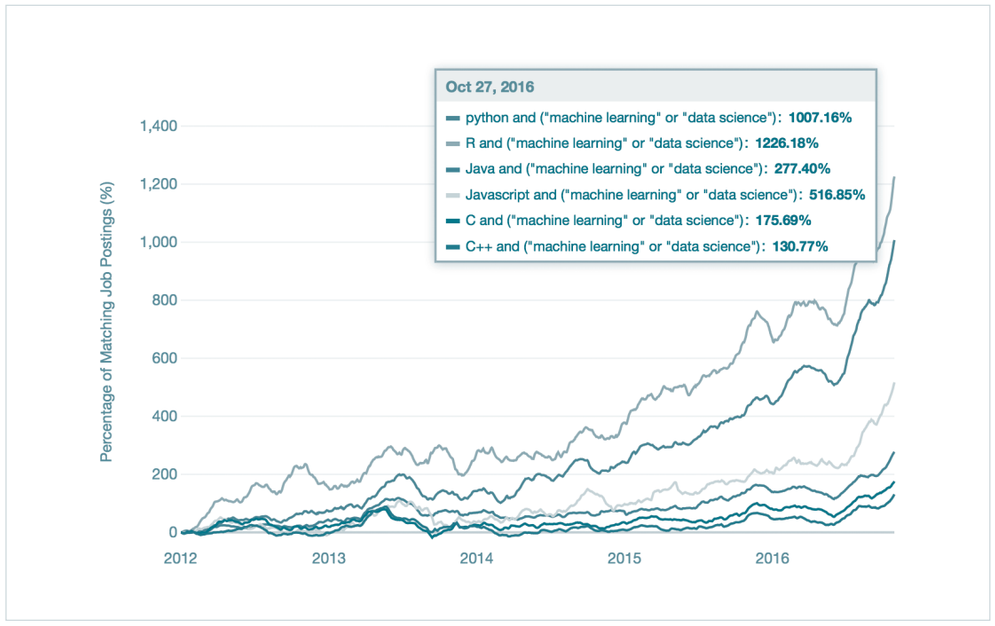
* Andy Kirk (2012), *Data Visualization: A Successful Design Process*, Packt Publishing, pp.161-169

2. 데이터 시각화 기술 동향

가. 데이터 분석 환경의 성숙에 따른 지배적 시각화 기술의 변화

데이터 시각화는 데이터 분석 과정에서 밀접하게 사용되기 때문에, 데이터 분석 환경의 변화는 데이터 시각화에 대한 사용자의 요구에도 영향을 미친다. 빅데이터, 데이터과학, 머신러닝과 같은 연관 분야에 대한 관심은 데이터 시각화에 대한 관심과 흐름을 같이 한다. [그림 6-5-2]에서 볼 수 있듯이 데이터과학 분야에서 가장 많이 쓰이는 언어인 R의 지위는 변함이 없으나, 최근 파이썬(Python)이 급격히 그 격차를 좁히고 있으며, 다음으로 javascript가 뒤를 잇고 있다. 이러한 프로그래밍 언어의 순위는 데이터 시각화 분야의 언어의 순위에도 직접적인 영향을 미친다.

[그림 6-5-2] 머신러닝과 데이터과학 분야에서 사용되는 언어 순위



출처: IBM Developer Works

1) 분석 전문가를 위한 R과 파이썬 기반의 시각화 기술

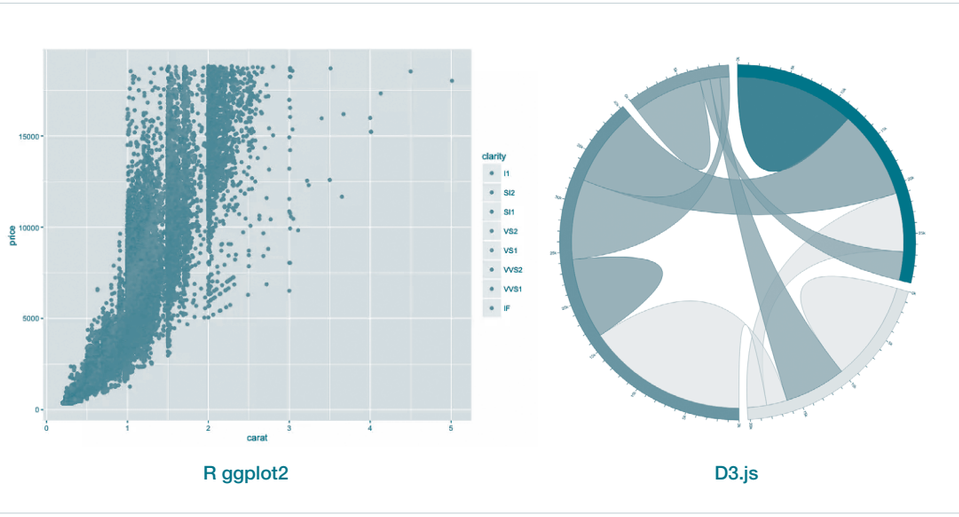
데이터 분석에서 R은 가장 많이 쓰이는 오픈소스 기술인 만큼, 매우 풍부한 시각화 라이브러리를 갖추고 있다. 체계적인 문법과 유려한 시각화가 특징인 ggplot2(http://ggplot2.org)는 R언어를 활용한 데이터 시각화를 대표하는 기술로서, 오랜 기간 사용되면서 지도 시각화를 위한 ggmap(https://github.com/dkahle/ggmap)과 같은 기술로 확장되고 있다. 그러나 동적인 시각화가 어렵다는 한계가 있는데, 인터랙티브한 시각화를 위해 shiny(http://shiny.rstudio.com/)와 같은 기술도 많이 활용되고 있다.

한편 파이썬은 최근 머신러닝 분야에서 중요한 역할을 담당하면서 더욱 주목을 받고 있으며, 통계 언어로 특화된 R에 비해 개발자 층이 두터워 R에 못지 않은 다양한 시각화 라이브러리를 갖추고 있다. 파이썬이 과학, 공학 및 교육 분야에서 널리 쓰이면서 MATLAB의 시각화를 계승하는 matplotlib(https://matplotlib.org)이 많이 활용된다. 그 외에도 통계적 데이터의 시각화에 적합한 seaborn(https://seaborn.pydata.org)이나 Altair(https://altair-viz.github.io), 그리고 인터랙티브한 시각화가 가능한 bokeh(http://bokeh.pydata.org)와 같은 시각화 라이브러리도 활용된다. 이와 같은 시각화 기술들은 다양한 분야에서 꾸준히 활용되고 있다.

2) 시각화 전문가를 위한 javascript 기반의 시각화 기술

데이터 분석의 결과를 공유하고 의사결정에 활용하기 위해 웹 환경에서 인터랙티브 시각화를 구현하려는 요구가 증가하면서, javascript 기반의 시각화 기술로 중심이 이동하고 있다. 가장 대표적인 웹 기반의 동적 시각화 라이브러리인 D3.js(https://d3js.org)는 다양한 데이터 타입에 대응해 기본 차트뿐만 아니라 복잡한 시각화의 구현에서 상당히 높은 자유도를 제공한다. 이와 같은 javascript 기반의 시각화 라이브러리는 꾸준히 증가하는 추세다. chart.js(http://www.chartjs.org)나 googlechart.js(https://developers.google.com/chart)와 같이 차트의 시각화 기술은 너무나 다양하며 목적과 특징에 따라 다양하게 활용된다. 오랜 기간 시각예술가들의 대표 언어로 자리매김한 processing(https://processing.org)은 본래 java로 만들어졌으나, 최근의 환경 변화를 반영해 p5.js(https://p5js.org)라는 자바스크립트 기반의 언어로 재탄생해 트렌드 변화에 대응하고 있다.

[그림 6-5-3] 프로그래밍 환경의 시각화 예



앞서 살펴본 R의 ggplot2, 파이썬(Python)의 matplotlib, 자바스크립트의 D3.js와 같은 각 분야의 대표적인 시각화 기술은 꾸준히 사용되고 있으며, 향후에도 두터운 사용자층으로 지속적으로 활용될 것이다.

나. 데이터 분석 서비스의 확산에 따른 사용자 친화적 대시보드 기술의 발전

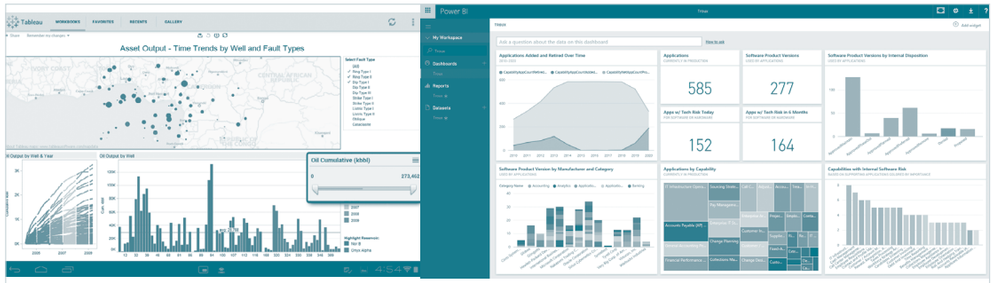
1) 비즈니스 인텔리전스의 사용자 친화적 대시보드의 발전

데이터 시각화가 핵심적인 기능을 담당하는 비즈니스 인텔리전스(Business Intelligence: BI) 솔루션 순위에 변동이 일어나고 있다. BI 분야에서 장기간 1위를 지켜온 SAP(https://www.sap.com)의 위치를 비주얼 애널리틱스에 강점이 있는 태블로(Tableau, https://www.tableau.com)가 차지했다. 이러한 변화에는 태블로가 다른 BI 솔루션에 비해 직관적 사용성이 우수한 점이 영향을 미쳤다. 또한 최근 데이터 분석과 인공지능 플랫폼 구축에 집중하고 있는 마이크로소프트의 Power BI(www.powerbi.com) 역시 기존의 분석 기능에 더해 더 강력한 시각화를 선보이고 있다. 이러한 경향은 데이터 분석의 대중화에 따라 대시보드의 사용빈도 역시 증가하면서 분석가뿐만 아니라 다양한 이해관계자들에게 사용성이 중요하게 여겨지는 경향을 반영한다. 구글의 데이터 검색 결과를 시각화해 보여주는 구글 트렌드(Google Trend, https://trends.google.com)가 기존의 시계열 방식뿐만 아니라 다양한 시각화를 구현한 Visualizing Google Data 페이지를 선보인 것도 이러한 경향을 반영하는 것으로 볼 수 있다.

[그림 6-5-4] Magic Quadrant for BI and Analytics Platforms 2016~2017

출처: 가트너

[그림 6-5-5] 비즈니스 인텔리전스의 예



2) 전문분야를 위한 데이터 시각화 기술의 활성화

데이터 시각화 기술은 범용적 용도뿐만 아니라 특정한 전문 분야나 사용자 그룹을 위해 특화돼 사용되기도 한다. 자연어 처리를 위한 ElasticSearch는 Kibana(https://www.elastic.co/kibana)를 통해 언어적 정보를 다양한 그래픽적 방법으로

시각화해 파악할 수 있도록 돕는다. 또한 데이터 저널리스트를 위한 시각화 스튜디오인 Flourish(https://flourish.studio)는 데이터에 기반한 스토리텔링을 만들 수 있도록 한다. 또한 데이터 기반의 서비스가 자체 시각화 기술을 오픈소스화하기도 하는데, 우버(Uber)는 deck.gl(https://github.com/uber/deck.gl)이라는 WebGL 기반의 3D 지리정보 시각화 프레임워크를 최근 공개했다. 이처럼 전문가 도구를 활용한 데이터 시각화가 특정 기업이나 분야를 넘어 일반화되는 경향이 지속될 것으로 전망된다.

[그림 6-5-6] 전문분야 시각화의 예

Elastic Search Kibana　　Uber deck.gl

다. 몰입형 데이터 시각화 기술의 부상

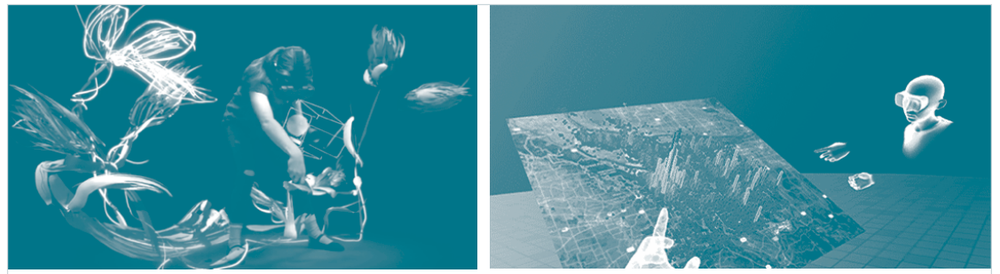
1) 3차원 시각화 기술의 중요성 증대

가상현실(Virtual Reality, VR) 기술이 보편화되면서 다양한 VR 헤드셋이 공급돼 선택의 폭이 넓어지고 최근에는 VR 전용 어트랙션이 등장하는 등 대중화 단계에 접어들고 있다. PC와 콘솔 기반 VR 헤드셋으로는 Facebook Oculus Rift, HTC Vive, Sony Playstation VR 등이 있고, 모바일 기반의 VR 헤드셋으로는 Samsung Galaxy Gear VR, Google Daydream 등 다양한 VR 디바이스가 보급되고 있다. VR 디바이스의 보급으로 스크린 중심의 2차원 데이터 시각화와 다르게 몰입감 있는 3차원 시각화의 필요성이 증가하고 있다. VR 보급에 장벽이 되고 있는 콘텐츠 부족 문제를 해소하기 위해 페이스북은 Oculus medium(https://www.oculus.com/medium)으로 쉽고 빠르게 콘텐츠를 제작할 수 있는 기술을 선보였다. HTC는 Vive Tracker(https://www.vive.com/eu/vive-tracker)로 3개의 트래커를 사용해 전신 추적을 할 수 있는 오픈소스를 공개해, 실시간으로 자연스러운 동작을 가상현실에서 구현할 수 있는 기술을 제공한다.

2) 가상현실 분야에 활용되는 몰입형 데이터 시각화

VR 환경에서 실감나는 경험을 제공하는 몰입형 데이터 시각화(immersive data visualization)의 중요성이 증대되고 있다. Virtualitics(https://www.virtualitics.com)는 데이터 시각화에 인공지능과 VR 기술을 결합해 가상현실 환경에서 마치 데이터 속에서 활동하는 듯한 경험을 제공함으로써 탐색적 데이터 시각화의 새로운 방법을 보여준다. 또한 구글 틸트 브러시(Google Tilt Brush, https://www.tiltbrush.com)는 VR 기술을 아티스트의 세계로 끌어들이고 있다. 디즈니(Disney)는 VR 영화(https://www.disneymoviesvr.com) 개발에 투자하고 있는데, Disney Research LA에서 개발한 iKinema(https://ikinema.com/Orion)는 VR 환경에서의 자연스러운 모션캡처와 바디트래킹 기술을 실시간으로 시각화하는 기술을 지원한다. 이와 같이 몰입형 데이터 시각화 기술은 VR 기술의 연구개발 및 상품화 과정에서 핵심적인 역할을 담당하고 있다.

[그림 6-5-7] 가상현실에서 데이터 시각화 기술의 예



Virtualitics　　구글 틸트 브러시

3) 증강현실과 혼합현실(Mixed Reality, MR) 기술의 보급

증강현실(Augmented Reality, AR)은 포켓몬고, 구글 워드렌즈와 같이 일상에서 활용되는 사례가 점차 증가하며 관심이 높아지고 있다. Microsoft HoloLens, Fove O,

Razer OSVR과 같은 AR 디바이스도 다양화되고 있으며, PTC Vuforia(https://www.vuforia.com)와 같은 AR 플랫폼이나 Augment(http://www.augment.com)와 같이 이커머스에 특화된 AR 플랫폼도 등장하고 있다. 증강현실 역시 콘텐츠 부족의 문제를 해결하기 위한 기술이 필요하다. 증강현실 환경에서 3D 모델링을 통해 편리하게 데이터를 시각화할 수 있는 DottyAR(http://dottyar.com)은 쉽고 빠르게 증강현실 콘텐츠를 제작할 수 있도록 한다.

4) 증강현실을 통해 새로운 차원으로 진입한 과학 시각화

데이터 시각화의 한 분야인 과학 시각화(scientific visualization)는 시뮬레이션 데이터 등 복잡한 데이터를 쉽게 탐색할 수 있도록 그래픽 기술을 활용해 시각화하는 분야로서 의학이나 과학, 공학 분야에서 활용돼 왔다. 증강현실 기술의 보급으로 과학 시각화 분야가 새로운 차원으로 접어들고 있다. Insight Heart(http://www.insight-heart.com)는 AR 기반의 시각화를 활용해 의학 교육에 새로운 경험을 제공한다. 또한 AR 기술은 건축 시뮬레이션 분야에서도 활용도가 높은데, Smart Reality(http://smartreality.co)는 AR 기반의 BIM(Building Information Modeling) 모델링 기술을 제공한다. 이와 같이 학문적 영역에 집중됐던 과학 시각화가 AR 기술을 통해 실생활에 응용되는 사례가 점차 증가할 것으로 보인다.

[그림 6-5-8] 증강현실에서 시각화 기술 활용의 예

DottyAR　　Insight Heart

라. 인공지능을 위한, 인공지능에 의한, 인공지능의 시각화

1) 딥러닝의 발전으로 인한 고차원 데이터 시각화 기술의 재부상

인공지능 분야가 오랜 침체기를 지나 딥러닝 기술을 통해 화려하게 부활하고 있다. 머신러닝의 개념을 직관적으로 이해하기 위해 시각화를 이용하는 방법은 오랜 기간 사용돼 왔으나, 최근 딥러닝의 복잡한 개념을 이해하기 위해 시각화가 적극적으로 활용되고 있다. 딥러닝 프레임워크의 표준이 돼 가고 있는 텐서플로(Tensorflow)는 딥러닝의 이해와 디버깅 및 최적화를 돕기 위한 시각화 도구인 Tensorboard(https://www.tensorflow.org/get_started/graph_viz)를 제공한다. 딥러닝 등장 이후 머신러닝의 데이터 차원이 수백 개 이상으로 폭증하게 되면서, 고차원 데이터를 차원축소해 시각화하는 기술인 t-SNE(https://lvdmaaten.github.io/tsne)가 다시 주목받고 있다. 또한 딥러닝은 인간의 신경망을 모방해 만들어진 알고리즘으로, 딥러닝의 고차원 데이터 처리 과정을 3차원 시뮬레이션으로 시각화함으로써 딥러닝과 인간 두뇌의 작동 원리를 이해하는 데 기여할 수 있다. Neural Network 3D Simulation(https://youtu.be/3JQ3hYko51Y)은 이러한 시도의 좋은 예이다.

[그림 6-5-9] 인공지능 시각화의 예

Tensorboard　　Neural Network 3D Simulation

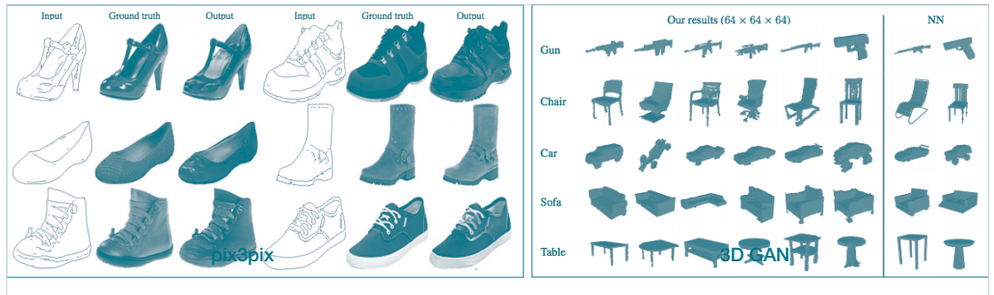
2) 인공지능이 생성하는 데이터 시각화의 가능성

딥러닝 기술 중 최근 주목받고 있는 GAN(Generative Adversarial Networks)*은 다

* Ian J.Goodfellow,et al. (2014), *Generative Adversarial Network*, Advances in Neural Information Processing Systems 27(NIPS 2014)

양한 분야에 적용된 연구가 쏟아지고 있다. Berkley AI Research(BAIR) Lab의 pix2pix(https://phillipi.github.io/pix2pix)*는 2D 이미지를 생성하는 방법을, MIT CSAIL의 3D GAN(http://3dgan.csail.mit.edu)**은 3D 모델링을 생성하는 방법을 각각 보여준다. 인공지능이 생성하는 시각화는 데이터의 입력 시점과 출력 시점, 데이터를 처리하는 방식과 시각화 방식을 결정하는 방식 등 모든 측면에서 기존의 데이터 시각화와는 전혀 다른 근본적인 패러다임의 변화를 가져올 가능성이 있다는 점에서 주목할 필요가 있다.

[그림 6-5-10] 3D GAN의 사례



3. 향후 전망

지금까지 데이터 시각화 기술의 동향에 대해 간략히 살펴보았다. 데이터 시각화는 컴퓨터공학, 통계학, 심리학, 인지공학부터 디자인까지 다양한 분야의 지식이 복합적으로 요구되는 분야로서, 기술의 발전과 사회적 변화에 영향을 받지 않을 수 없다. 데이터 시각화 분야는 수십 년 이상 쓰여온 오래된 기술과 가장 최신의 기술이 공존하면서 발전하고 있다.

데이터 시각화 기술은 전통적인 통계적 데이터의 시각화뿐 아니라, 웹 환경에 적합하고 인터랙티브한 구현이 가능한 기술, 그리고 사용자 친화적이고 직관적인 사용성이 높은 기술, 세분화된 전문 분야에 특화된 기술에 대한 선호가 올라가는 경향을 보이고 있다. 또한 VR 기술은 데이터 시각화가 구현되는 새로운 플랫폼으로서 가상의 환경 속에서 데이터를 직접 경험할 수 있는 몰입형 데이터 시각화의 중요성이 강조되고 있다. 또한 AR 기술은 보다 정교한 의학 교육이나 공학 시뮬레이션에 활용되고 있다.

앞으로 데이터 시각화뿐 아니라 거의 모든 분야가 인공지능의 영향으로부터 자유롭지 못할 것이다. 복잡한 고차원 정보를 시각화하는 기술은 딥러닝의 작동 원리를 이해하고 편리하게 구현하도록 하는 데 많은 기여를 하고 있다. 나아가 인공지능이 스스로 의미 있는 정보를 생성하는 기술이 개발돼 연구 결과가 쏟아지고 있으며, 인공지능은 지금까지 인간의 고유한 영역으로 여겨진 창작의 영역에 도전하고 있다. 인공지능은 전통적인 데이터 시각화 분야의 생산성을 개선할 뿐 아니라, 데이터 시각화의 프로세스와 방법론 그리고 결과물에까지 근본적인 변화를 가져올 것으로 전망된다.

* Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, Alexei A. Efros (2016), *Image-to-Image Translation with Conditional Adversarial Networks*, arXiv:1611.07004 [cs.CV]

** Jiajun Wu, et al. (2016), *Learning a Probabilistic Latent Space of Object Shapes via 3D GenerativeAdversarial Modeling*, arXiv:1610.07584 [cs.CV]

제6장

블록체인

블록체인(Blockchain)은 최근 관심이 집중되고 있는 디지털 암호 통화인 비트코인(Bitcoin)의 기반 시스템이다. 본 장에서는 블록체인의 기본원리에 대해 설명하고 최근의 동향과 관련 기술을 소개한다. 또한 블록체인이 갖고 있는 기술적인 의의에 대해 살펴보고 전망에 대해 알아본다.

1. 블록체인 개요

블록체인(Blockchain)은 2009년 나카모토 사토시로 알려진 미상의 인물에 의해 만들어진 비트코인(Bitcoin)의 기반이 되는 시스템으로 널리 알려지게 됐다. 블록체인은 분산 복제에 근거한 데이터베이스이며, 중앙집중식 관리자가 존재하지 않는 특징을 갖고 있다. 블록체인과 비트코인은 단일한 시스템에 근간을 두고 있는 물리적·논리적 계층이다. 즉 비트코인은 블록체인의 시스템을 유지하기 위한 논리적인 수단이며, 블록체인은 비트코인의 가치를 유지하기 위한 물리적인 수단으로서 상호 불가분의 관계를 갖고 있다.

비트코인은 2017년 4월 14일 현재 1BTC당 1,183.44USD의 가치로 거래되고 있으며, 이를 2009년 최초의 제시 가격인 1USD당 1,309.03BTC의 가치와 비교하면 백만 배에 가까운 가치 상승을 가져온 것이다**. 이러한 비트코인의 현실 가치의 상승은 자연스럽게 기반 시스템인 블록체인에 대한 관심을 불러일으키고 있다. 전 세계적으로 블록체인 기술에 기반한 다양한 스타트업 기업들이 등장하고 있고 이에 대한 투자가 활발하게 일어나고 있다. 우리나라도 정부와 민간기업에서 블록체인을 응용한 다양한 애플리케이션들에 대한 청사진을 제시하고 있는 상황이다.

2. 블록체인의 기본 원리

블록체인은 비트코인의 거래정보(트랜잭션)들을 담고 있는 거래원장(블록)의 연쇄적인 해시 체인* 구조라고 할 수 있다. 따라서 블록체인의 기본 원리를 이해하기 위해서는 트랜잭션과 블록의 구조를 이해하고 해시체인에 기반한 작업증명의 원리를 이해할 필요가 있다.

가. 트랜잭션

비트코인의 거래, 즉 트랜잭션은 실생활에서의 수표 거래의 배서와 매우 유사하다. A가 은행으로부터 발급받은 10만 원의 수표가 있다고 가정할 때, A가 B에게 10만 원을 지급하고 싶다면 본인이 갖고 있는 수표 뒷면에 본인의 정보와 서명을 배서해 B에게 건네 주는 방식으로 수표의 소유권은 B에게 넘어가게 된다. 만약 또 다시 B가 C에게 10만 원을 지급할 경우는 다시 수표 뒷면에 본인의 정보와 서명을 배서해 C에게 건네 줌으로써 수표의 소유권은 C가 갖게 된다. 실제 화폐를 갖고 있지 않더라도 이러한 수표의 소유는 해당 화폐 가치를 갖고 있다는 논리적인 증명수단이 된다.

이제 A가 B에게 10 BTC을 지급하는 비트코인의 트랜잭션을 [그림 6-6-1]을 통해 설명해 보도록 하자. A는 10 BTC 비트코인의 소유를 증명해주는 기존의 트랜잭션 T1으로부터 10 BTC의 소유권을 B에게 이전하는 새로운 트랜잭션 T2를 생성하는 방식으로 10 BTC을 지급하게 된다. 우선 비트코인의 거래를 위해서

* 필자: 김철연(숙명여자대학교 IT공학과 교수)

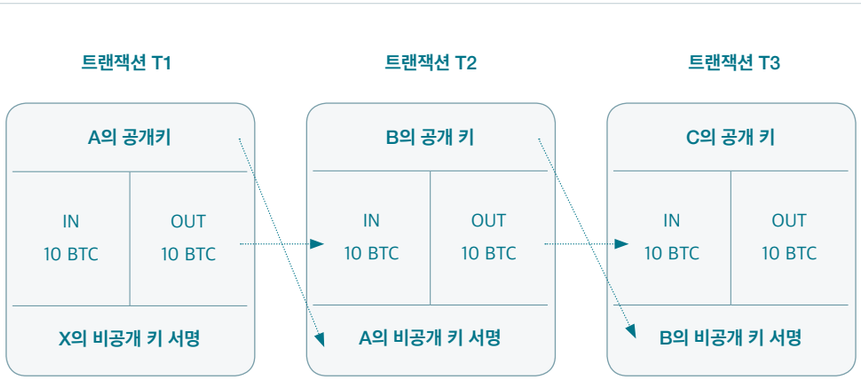
** 비트코인을 비롯한 다양한 디지털 가상화폐들이 전 세계적으로 민간 거래소를 통해 현실 화폐와 교환되고 있으며, 시장원리에 따라 교환비율이 동적으로 결정되고 있다.

* 해시함수는 해시키에서 해시값으로의 함수이며 해시체인은 하나의 해시함수의 해시값이 또다른 해시함수의 해시키에 포함되는 구조를 가진다. 따라서 선행하는 해시함수의 해시키가 후행하는 모든 해시함수의 해시값에 영향을 미치게 된다.

는 A와 B는 각각 비대칭키 암호화 기법*에 기반한 공개 및 비공개키를 갖고 있어야 한다. A는 이전 트랜잭션 T1의 정보와 이전할 금액인 10 BTC, 그리고 이를 이전받을 B의 공개키를 기록한 거래명세를 본인의 비공개 키로 암호화 서명을 한다. 이러한 정보와 서명이 바로 새로운 트랜잭션 T2이며 이러한 새로운 트랜잭션 T2는 생성과 동시에 블록체인 네트워크 상의 모든 사용자들에게 전파된다.

이때 B가 10 BTC을 소유하게 됐다는 것은 트랜잭션 T2로부터 새로운 트랜잭션 T3를 생성할 수 있다는 것을 의미한다. 새로운 트랜잭션 T3의 생성을 위해서는 이전의 설명과 마찬가지로 거래명세를 B의 비공개 키로 암호화 서명을 해야 한다. 이때 이러한 암호화 서명이 이전의 T2에 기록된 B의 공개키로 다시 풀릴 수 있어야 하며 이로써 T3의 서명자인 B가 T2로부터 T3를 생성하는 것이 정당하다는 것을 보여주게 된다.

[그림 6-6-1] 트랜잭션의 주요 구성과 연결 관계



나. 블록

트랜잭션을 통한 비트코인의 거래에서 한 가지 불완전한 문제점은 A가 트랜잭션 T1으로부터 T2를 생성한 후, T1으로부터 또 다른 트랜잭션 T4의 생성을 시도

* 암호화를 하는 열쇠와 암호화를 푸는 열쇠가 서로 다른 암호화 방식. 두 개의 열쇠를 통해 상호간의 암호화를 풀 수 있다.

하는 이중지불(double-spending) 문제가 있다는 것이다. 이때에 T2와 T4는 동시에 유효할 수 없으며, 하나의 트랜잭션만이 유효한 트랜잭션으로 인정돼야 한다. 따라서 트랜잭션의 생성과 전파만으로 해당 트랜잭션의 유효성을 담보할 수는 없으며, 유효하다고 인정된 트랜잭션만을 기록하는 원장의 개념이 필요하게 된다. 즉 전파되는 트랜잭션들 중 유효하다고 인정된 트랜잭션만이 원장에 기록돼야 하며, 이러한 원장이 바로 블록에 해당한다. 블록체인에서 하나의 블록은 최대 1MB의 용량을 가지며 해당 용량 내에서 유효한 트랜잭션들을 기록한다. 이후 계속해 생성되는 유효한 트랜잭션들을 저장하기 위해 새로운 블록들이 계속해 늘어나게 된다.

이때에 가장 중요한 문제는 어떠한 트랜잭션을 유효하다고 판단해 원장에 기록하는 권리를 누구에게 부여하고 이러한 원장의 위변조를 어떻게 방지할 수 있느냐는 문제다. 기존의 전통적인 거래 시스템에서는 참여자들로부터 이러한 권리를 양도받은 은행과 같은 중앙집중식 기관이 트랜잭션의 유효성을 판정하고 외부의 악의적인 공격으로 인한 위변조를 막기 위해 외부와 원장을 격리하는 폐쇄적인 시스템을 구성한다. 또한 이러한 중앙집중식 기관은 내부적인 위변조를 시도하지 않을 것과 해당 시스템의 안전성을 보장하는 책임을 지게 된다. 거래 참여자들은 이러한 중앙집중식 기관의 능력과 선의를 전적으로 신뢰하는 것 이외의 다른 선택이 없다.

블록체인은 이러한 전통적인 거래시스템과는 정반대의 방식으로 유효성의 선택과 위변조 방지를 구현한다. 즉 중앙집중식 기관의 존재를 배제하고 거래 정보를 기록하는 블록을 모든 참여자들에게 전파해 공유하는 방식을 택하고 있으며 위변조를 막기 위해 일종의 다수결의 원리를 채택하고 있다. 즉 거래정보의 무결성이 침해되는 블록들이 공존할 경우 다수결에 의해 더 많은 사용자가 지지하는 블록만을 유효한 블록으로 인정하는 방식이다.

따라서 블록체인에서는 다수결을 어떻게 정의할 것인가가 중요한 문제가 된다. 쉽게 생각해 참여자들의 아이디들 중 과반수의 아이디들의 지지를 다수로 간주한다면 악의적인 의도를 갖고 있는 사람이 허수의 참여자 아이디를 대량으로 만들어 내어 다수결의 결과를 뒤집으려 하는 시도를 할 수 있을 것이다. IP 주소*

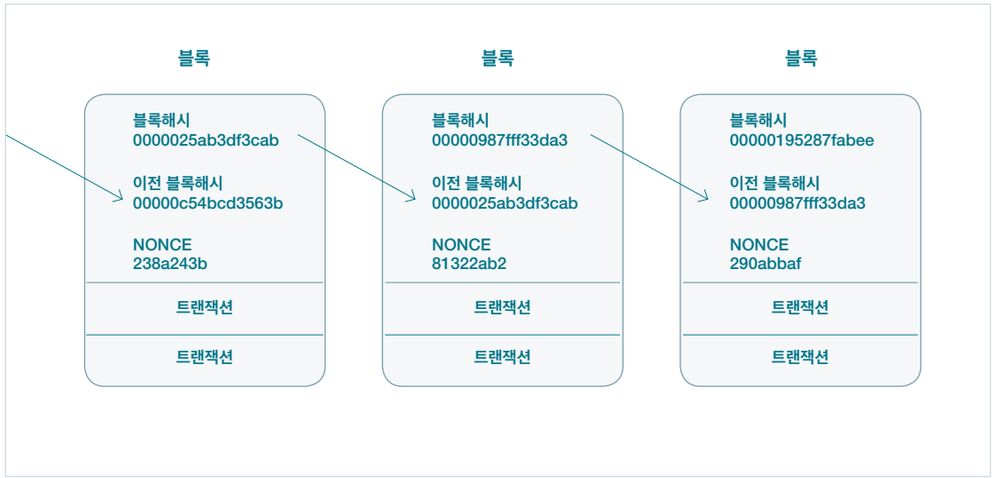
* 인터넷에 접속하는 모든 장치들이 서로를 식별하기 위해 부여하는 주소값

나 MAC 주소*와 같은 값도 어렵지 않게 허수의 값을 만들어 낼 수 있으며, 이로 인해 다수결의 과정이 교란될 수 있다. 이를 해결하기 위해 비트코인의 블록체인 시스템에서는 작업증명(proof of work)을 통해 허수의 참여자를 배제하는 방식을 택하고 있다. 즉 블록체인에서는 전체 참여자들의 총 작업량 중 과반의 작업량을 다수로 간주하는 원칙을 사용한다. 만약 악의적인 공격자가 블록체인을 위변조하기 위해는 전체 참여자들의 총 작업량의 과반 이상의 작업량을 수행해야 한다. 비트코인과 같이 수많은 참여자가 존재하는 블록체인 시스템 환경에서 공격자에게 이는 현실적으로 거의 불가능한 수준의 장벽으로 동작하게 된다.

다. 해시체인에 기반한 작업증명

블록체인의 작업증명은 '비트코인의 채굴'로 널리 알려져 있다. 해당 작업을 개념적으로 설명하면 해시 함수 Y = f(X)의 해시 값 Y를 특정한 값 k보다 작게 만드는 X값을 구하는 작업이다. 이때에 X는 다양한 정보의 결합으로 구성되는데 주요한 정보는 블록에 저장된 트랜잭션들의 정보, 이전 블록의 해시값, 그리고 임의의 값을 넣을 수 있는 NONCE** 항목이 있다. 이때 블록에 저장된 트랜잭션의 정보는 현재 블록이 담고 있는 거래기록이다. 이전 블록의 해시값은 블록들을 연쇄적으로 이어주는 연결정보가 되며, NONCE 값은 임의의 값으로 변경이 가능해 해시값 Y를 변경시키는 역할을 담당하게 된다. 즉 해시값 Y가 특정한 값 k보다 작아지도록 만드는 NONCE의 값을 찾는 것이 블록체인 작업증명의 작업이 된다. 이에 대한 대략의 예는 [그림 6-6-2]에 나타나 있다. 이때 해시함수* f는 256비트로 표현 가능한 모든 정수 영역에서 균등한 분포를 갖는 함수이기 때문에 k값에 따라 해당 작업의 작업량이 달라지게 된다. 만약 k가 모든 표현 가능한 정수 중 중간 크기를 갖는 수라면 평균적으로 1.5번의 시도로 NONCE 값을 찾을 수 있지만, 만약 k가 매우 작다면 더 많은 시도를 해야 원하는 NONCE 값을 찾을 수 있다. 비트코인의 블록체인은 이러한 NONCE 값이 평균적으로 10분 동안 작업을 한 후에 발견될 수 있도록 동적으로 k의 값을 조정할 수 있다.

[그림 6-6-2] 블록체인의 기본 구조

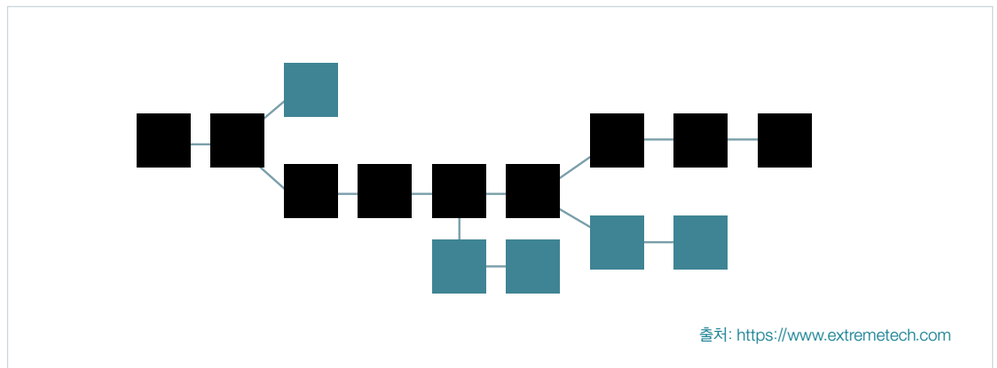


* 네트워크의 통신을 위한 네트워크 인터페이스에 할당된 고유 식별자

** 32비트의 정수값

비트코인의 블록체인에서는 모든 유효한 블록들은 1차원 연결구조를 갖게 된다. 이러한 1차원 연결구조의 최초 블록을 Genesis 블록이라 부르며, 시스템과 함께 최초 정의돼 소스코드에 하드코드돼 있다. Genesis 블록 이후 새롭게 생성되는 모든 블록은 현재 1차원 연결구조의 제일 말단 블록에 연결돼야 한다. 이때의 문제점은 새로운 블록의 생성을 시도하는 참여자가 복수명이 있을 경우 하나의 말단 블록에 복수 개의 신규 블록이 연결돼 1차원 구조가 깨지는 멀티스레드(thread)가 생길 수 있다는 것이다. 이에 대한 블록체인의 해결책은 이러한 복수 개의 멀티스레드 중 가장 길이가 빠르게 자라나는 스레드만을 유효한 1차원 연결로 인정하고 나머지 스레드들은 도태시키는 방식을 취한다. 이는 작업량에 근거

[그림 6-6-3] 블록체인의 성장



출처: https://www.extremetech.com

* SHA-256(Secure Hash Algorithm) 함수가 사용되며 함수값은 256비트(32바이트)로 구성된다.

한 다수결 원칙의 적용이라고 볼 수 있다. 즉 가장 많은 작업량이 지지하는 블록의 스레드가 가장 빠르게 자라날 수 있다는 원칙이다. 따라서 [그림 6-6-3]의 예에서는 검은색의 스레드가 유효하게 자라나는 블록체인이라고 할 수 있다.

비트코인의 신규 발행은 바로 작업증명에서 발생하는데, 유효한 블록을 새롭게 추가하는 사용자에게 일정량의 비트코인이 신규 발행돼 해당 작업에 대한 대가로 지급된다. 따라서 이러한 작업증명 과정이 비트코인 채굴로 더 널리 알려지게 된 것이다. 비트코인의 화폐가치 안정성을 위해 설계된 더 정교한 설정들이 있으나 이는 화폐의 희소성을 유지하기 위한 요소이기 때문에 이 글에서는 자세히 언급하지는 않겠다.

만약 악의적인 공격자가 블록체인의 어느 한 블록을 위변조하고 싶다면 크게 세 가지 가능성을 고려할 수 있다. 첫째 해당 블록의 내용을 수정해 이전의 해시값과 동일한 해시값을 갖도록 NONCE 값을 수정하는 방식이다. 이는 작업증명에서 사용하는 해시함수가 이론적으로 서로 다른 X에 대해 동일한 Y값이 생성되는 것을 회피하기 때문에 현실적으로 불가능한 방법이 될 수 있다. 둘째 해당 블록의 내용을 수정하고 이전 블록의 스레드보다 더 긴 스레드를 새롭게 위조한 블록의 뒤로 빠르게 만들어 내는 것이다. 이는 정상적인 참여자들의 총 작업량보다 더 큰 작업량을 산출해야 가능한 방식이기 때문에 이론적으로는 가능하나 이러한 작업량을 투입하는 것이 현실적으로 불가능하다. 마지막으로 해당 블록의 내용을 수정하고 이에 연결돼 있던 모든 블록들의 작업증명을 업데이트하는 것이다. 이 방식 또한 정상적인 참여자들의 총 작업량보다 더 큰 작업량을 산출해야 기존 스레드의 증가 속도를 따라잡을 수 있으므로 이론적으로는 가능하나 현실적으로는 불가능한 방식이다.

라. 블록체인에 기반한 거래의 완결시점

블록체인에 기반한 비트코인 거래는 크게 세 가지의 단계를 거치게 된다. 첫 번째로는 트랜잭션의 생성과 배포 단계, 두 번째로는 새로운 블록에 유효한 트랜잭션으로 기록되는 단계, 마지막으로는 속해 있는 블록이 유효한 스레드로 성숙(Maturation)되는 시점이다. 각각의 단계에 걸리는 시간은 트랜잭션의 생성과 배포는 수 초 정도의 시간이 소요되며, 블록에 기록되는 시점은 최소 10분, 길게는 그 이상의 시간이 소요될 수 있다. 마지막 단계는 보통 해당 블록부터의 스레드의 길이가 100개 이상이 되는 시점으로 최소 16시간 이상의 시간이 소요된다. 가장 안정적인 거래의 종료는 마지막 단계까지 완료된 후라고 할 수 있으나 거래에 소요되는 시간이 거의 하루 가까이 걸린다면 비트코인의 실용성은 매우 취약하게 될 것이다. 따라서 거래자들은 서로의 필요와 상호간의 신뢰도의 수준에 따라 실질적 거래의 완결시점을 조정하는 것이 일반적이다.

3. 블록체인의 최근 동향

가. 비트코인의 발전과 혼돈

화폐가치의 상승과 더불어 비트코인의 거래량이 급증하고 비트코인 채굴 경쟁이 과열되면서 대내외적으로 여러 문제점이 생기고 있다. 2017년 4월 14일 현재 비트코인의 화폐가치는 최초의 공시 가치에서 백만 배 가까이 상승했으며, 거래량은 1일 기준 30만 건에 가까운 트랜잭션들이 블록체인에 기록되고 있다.

최근 생성되는 트랜잭션이 급증함에 따라 이에 대한 처리 속도를 높이기 위해 블록의 크기를 키우기 위한 논쟁이 진행되고 있다. 현재의 블록의 크기는 1MB로 제한돼 있고, 새로운 블록의 생성은 10분의 시간 간격을 가지므로 블록체인이 단위시간 내에 처리할 수 있는 최대 트랜잭션의 수는 제한돼 있다. 따라서 블록의 크기를 키우거나 새로운 블록의 생성을 위해 필요한 작업증명의 요구시간을 줄이는 방식으로 블록체인의 처리 속도를 높일 수 있다. 이중에서 작업증명의 요구시간을 줄이는 방식은 통화량의 증가 속도와 관계가 있기 때문에 비트코인의 채굴자에게는 유리한 방식이지만 투자자에게는 불리한 방식이 될 수 있다. 이와 다르게 블록의 크기를 키운다면 통화량의 증가 속도가 그대로 유지되면서 처리 속도를 향상시키므로 투자자에게는 유리한 방식이지만 채굴자들이 반기지 않는 방식이라는 문제점이 있다. 따라서 블록의 크기를 키우는 논쟁은 아직도 진행 중인 문제이며 처리되지 못한 트랜잭션들의 적체 현상이 본격적으로 일어난다면 이는 비트코인의 실용성에 대한 심각한 위기가 될 수 있을 것이다.

또 다른 비트코인의 안정성에 대한 문제로는 거래소 해킹 사건들을 들 수 있다. 가장 대표적인 사건으로는 2016년 8월 홍콩의 한 비트코인 거래소가 해킹을 당해 현금가치로 700억 원이 넘는 비트코인이 도난당하는 사건이 있었다. 이 사

블록의 크기를 키우는 논쟁은 아직도 진행 중인 문제이며 처리되지 못한 트랜잭션들의 적체 현상이 본격적으로 일어난다면 이는 비트코인의 실용성에 대한 심각한 위기가 될 수 있을 것이다.

건은 매우 막대한 자산의 손실을 가져온 불행한 사건이었으나 블록체인의 위변조에 의한 사건이 아니라 사용자들의 공개/비공개 키의 도난으로 인한 불법 사용의 문제였다는 점에서 블록체인 자체의 보안의 문제라기보다는 또 다른 중앙집중식 시스템인 거래소의 보안 문제점을 보여주는 사례라고 할 수 있다.

●○
소수 집단의 과반 점유 현상은 이들의 단합에 의해 블록체인의 안전성이 근본적으로 위협받을 수 있다는 문제를 의미한다.

여러 문제점들 중 가장 근본적인 문제점은 채굴 작업량의 집중 현상이다. 비트코인의 화폐가치 상승으로 경제적 가치에 대한 확신이 확산되면서 전세계적으로 전문적인 채굴자들과 채굴집단들이 활발히 활동하고 있다. 이러한 채굴집단 중 특히 중국을 중심으로 활동하는 몇몇 채굴집단의 작업량은 전체 비트코인 블록체인의 총 작업량의 과반 이상을 점유하고 있다고 알려지고 있다. 최대의 채굴집단으로 알려진 Antpool*은 총 작업량의 15%를 차지하고 있다고 평가되고 있다. 이러한 소수 집단의 과반 점유 현상은 이들의 단합에 의해 블록체인의 안전성이 근본적으로 위협받을 수 있다는 문제를 의미한다. 이러한 사항은 비트코인에 비판적인 사용자들로부터 계속 지적되는 문제로, 비트코인을 대체하는 블록체인 기반의 새로운 디지털 화폐의 등장을 가속화시키는 중요한 요인이라고 할 수 있다.

나. 비트코인의 새로운 대체자들

비트코인의 가장 유력한 경쟁자인 이더리움(Ethereum)**은 2015년 7월 30일 비탈리 부틴(Vitalik Buterin)이 제안한 가상화폐로 비트코인과 동일하게 블록체인을 기반으로 하는 플랫폼으로 구현돼 있다. 비트코인과의 가장 주요한 차이점은 비트코인의 블록체인이 비트코인 거래정보의 저장만을 위해 설계돼 있다면 이더리움의 블록체인은 이보다 범용적인 정보를 저장하고 튜링 완전언어(Turing-Complete Language)로 프로그래밍할 수 있도록 확장돼 있다. 또한 새로운 블록의 생성시간을 15초 정도로 단축했고, 급증하는 작업량과 소수 그룹의 독점 현상을 초래하는 작업증명을 지분증명(Proof of Stake)으로 대체했다. 이러한 이더리움은 2017년 4월 14일 현재 47 USD 정도의 가치로 비트코인에 비해 아직은 매우 낮은 화폐가치를 갖고 있으나 화폐가치가 최근 정체중인 비트코인과 비교해 훨씬 빠르게 상승하고 있고 전체 가상통화의 거래 비중의 15% 가량을 차지해 비트코인에 이어 2번째로 거래가 활발하다.

●○
가상화폐들은 비트코인의 단점들을 극복하기 위한 다양한 방법론을 도입하며 비약적인 화폐가치의 상승을 꾀하고 있다.

이더리움 외에도 리플(Ripple)*, 라이트코인(Litecoin)** 등의 블록체인 기반의 다양한 가상화폐가 등장하고 있다. 이러한 가상화폐들은 비트코인의 단점들을 극복하기 위한 다양한 방법론을 도입하며 비약적인 화폐가치의 상승을 꾀하고 있다. 시장에서 경쟁하는 가상화폐 중 사용자들의 요구를 충족시키는 화폐의 선택 과정을 통해 각각의 가치가 안정화 될 것이며 새로운 시장이 열릴 것으로 기대된다.

4. 블록체인의 기술적 의의와 전망

데이터베이스의 전통적인 중요가치인 무결성을 보전하기 위해 기존의 데이터베이스 시스템은 인적 요소로는 관리자의 권한을 특화해 해당 권한을 갖는 관리자를 최소한의 수로 제한하고 시스템적으로는 해당 데이터베이스의 접근을 엄격하게 통제하는 방식을 택했다. 이러한 경우 데이터베이스를 관리하는 중앙 관리자의 시스템 관리에 대한 책임과 비용이 집중되는 것에 비례해 데이터에 대한 접근성과 권한 또한 집중되는 특징이 있다. 따라서 중앙 관리자의 의지에 따라 전체 데이터베이스 시스템의 무결성이 보존될 수도 있고, 이와 반대로 악의적인 의도가 있을 경우 전체 데이터베이스의 무결성이 신뢰를 잃을 수도 있다.

블록체인은 이러한 전통적인 데이터베이스와 정반대의 접근방식으로 데이터베이스의 무결성을 추구하고 있다. 즉 인적 요소는 데이터베이스에 대한 권한을 모든 사용자가 동등하게 가져 별도의 관리자 권한을 특화하지 않으며, 물리적으로는 데이터베이스를 복제해 공유함으로써 모든 사용자가 데이터베이스에 대한 동등한 접근성을 갖게 된다.

전통적인 중앙통제식 데이터베이스의 무결성이 중앙집중식 시스템에 대한 접근 통제의 기술과 관리자의 선의에 기반한다면 블록체인의 안정성은 공유와 작업증명의 결합에 기인하며 이외의 접근 통제 기술과 특정인에 대한 신뢰를 필

* https://www.antpool.com
** https://www.ethereum.org

* https://ripple.com
** https://litecoin.org

요로 하지 않게 된다. 따라서 블록체인의 가장 중요한 특징은 중앙집중식 시스템과 관리자의 존재 없이 데이터의 무결성을 확보하는 분산 복제 데이터베이스라고 할 수 있다.

블록체인의 장점은 기존의 전통적인 중앙통제식 데이터베이스 시스템보다 더 우월한 무결성을 제공한다는 점이며, 단점은 더 우월한 무결성 제공을 위해 필요한 전체 시스템의 총비용은 증가한다는 점이다. 물론 다수의 사용자에게 이러한 비용을 분산시키는 방식으로 증가되는 비용을 감당하며 가상화폐라는 유인책을 통해 자발적인 참여를 이끌어내는 상황이다.

블록체인의 응용영역과 발전성에 대해는 여러 의견이 정제되지 않은 상태로 대중에게 회자되고 있는 현실이다. 블록체인의 구체적인 미래상은 아직은 불명확하지만 분산 복제 기술에 기반한 데이터 관리 시스템은 다양한 기능과 구성을 가진 주요한 정보처리 시스템으로 자리잡을 것으로 예상된다.

제 7 부

데이터산업 정책 동향

제1장

주요 국가의 데이터산업 정책

본 장에서는 세계적으로 데이터 정책의 본보기를 제시하고 있는 미국과 영국의 데이터산업 정책을 오픈데이터, 빅데이터 중심으로 알아보고 정권 교체와 정책 변화에 따른 양국의 데이터산업 정책의 변화를 예측해 본다. 더불어 일본의 데이터산업 정책과 함께 제4차산업혁명을 향한 일본정부의 로드맵을 살펴본다.

1. 미국의 정책

가. 오픈데이터 정책

오바마 행정부는 임기가 시작된 2009년부터 DATA.GOV** 서비스로 대표되는 오픈데이터 정책을 적극적으로 추진해 여러 국가들의 선례가 되고 있다. 2017년 4월 현재 DATA.GOV에는 192,322개의 데이터 세트가 올라와 있으며, 이는 농업·기후·소비·교육·에너지·재정·의료·공공안전·과학기술·해양 등 크게 14개 분야로 분류·제공되고 있다. 또한 데이터 활용을 위한 소프트웨어 애플리케이션을 함께 제공함으로써 누구나 쉽게 데이터를 활용해 부가가치 창출에 기여할 수 있는 환경을 구축했다. 덕분에 연구자들은 방대한 양의 데이터를 통해 원활한 연구 수행이 가능해졌고, 시민들은 정부의 정책에 더 긴밀하게 관심을 가지게 됐으며, 각 정부 부처들은 서로 데이터를 공유함으로써 불필요한 통계조사 비용을 절감하고 업무 효율성을 증진시켰다.

* 필자: 김경훈(정보통신정책연구원 부연구위원)

** https://www.data.gov (2017.04.11.)

미국은 이렇게 적극적인 정부 주도형 오픈데이터 정책을 펼친 덕분에 2015년에 월드와이드웹(WWW) 재단에서 발표한 '제3차 오픈데이터 글로벌 보고서(ODB Global Report Third Edition)'의* 국가별 오픈데이터 지표(Open Data Barometer) 순위에서 총 92개 국가 중 영국에 이어 2위를 기록했다. 오픈데이터 지표는 준비성(Readiness), 실행력(Implementation), 영향력(Impact)의 세 가지 요소로 구성된다. 여기서 미국은 각 요소에서 97점, 76점, 76점을 기록하며 오픈데이터의 잠재적 가치를 구현하기 위한 인프라 구축 정도를 의미하는 준비성 면에서 높은 평가를 받은 것을 알 수 있다. 실행력은 데이터를 얼마나 다양하게 공개하고 있는지, 영향력은 오픈데이터가 정치·사회·환경·경제에 얼마나 긍정적인 변화를 야기했는지 나타내는 지표로, 두 지표 모두 준비성에 비해 상대적으로 낮은 평가를 받았다.

나. 빅데이터 정책

오바마 행정부의 빅데이터 주요 정책은 2010년 12월에 대통령 과학기술자문위원회(PCAST**)가 빅데이터 관련 기술 투자의 필요성을 역설한 것을 시작으로, 2012년 3월에 과학기술정책국(OSTP***)에서 빅데이터 기술 개발 및 활용, 빅데이터 전문인력 양성을 주요 목적으로 발표한 '빅데이터 R&D 이니셔티브(Big Data R&D Initiative)', 그리고 2016년 5월에 연방정부 네트워킹 IT R&D 프로그램(NITRD****) 산하 빅데이터 협의체(BDIWG*****)에서 발표한 '빅데이터 R&D 전략 계획(The Federal Big Data R&D Strategic Plan)'으로 이어진다.

* http://www.opendatabarometer.org (2017.04.11.)

** President's Council of Advisors on Science and Technology

*** Office of Science and Technology

**** Federal Networking and Information Technology R&D

***** Big Data Interagency Working Group. OSTP에서 BDSSG(Big Data Senior Steering Group)라는 이름으로 출범했으나, 현재는 NITRD 산하 그룹 중 하나로 분류됨

[표 7-1-1] 미국의 빅데이터 주요 정책

담당부처 (날짜)	정책 (또는 문서)	주요 내용
PCAST ('10.12)	Designing a Digital Future*	• 빅데이터 관련 기술 투자의 필요성 건의
OSTP ('12.3)	Big Data R&D Initiative	• 6개의 연방 정부기관이 참여해 빅데이터 기술 개발 및 활용, 빅데이터 전문인력 양성을 목적으로 2억 달러 규모의 예산을 투입해 추진계획 발표 • 참여기관별 추진계획** - 국립과학재단(NSF): 국립보건원과 공동으로 빅데이터과학 및 공학 향상을 위한 기술개발 추진, 대학연계 및 지원 프로그램을 통해 대용량 데이터 저장 및 활용방안 연구 - 국방부(DoD): 군사 관련 빅데이터 프로젝트 연간 25,000만 달러 투입, 전투원 및 군 분석가의 전투 수행 능력을 배가시키기 위한 빅데이터 기술 연구 주력 - 국립보건원(NIH): 신경과학 데이터 수집 접근 개선에 대한 연구개발, 1000 게놈 프로젝트를 통해 해독된 약 200TB 규모의 인체 유전자 데이터 공개 - 방위고등연구계획국(DARPA): 대용량 데이터에서 특정 정보 탐지하는 기술 개발에 초점을 둔 ADAMA 프로젝트 추진, 기계독해(Machine Reading) 프로그램 진행 - 미국지질조사원(USGS): 지구 시스템 과학 분야에 빅데이터 활용, '존 웰시 파월 분석 및 통합센터'를 통해 지구과학의 혁신 도모 - 에너지부(DoE): 고등과학컴퓨터연구소 및 기초에너지과학사무소에서 대용량 데이터의 관리 및 접근·보존·시각화·분석 관련 기술 개발
OSTP ('13.11)	Data to Knowledge to Action	• 빅데이터 R&D 이니셔티브 발표 이후 1년간의 작업 경과를 공유하고, 8개의 새로운 프로젝트 발표*** - Novartis, Pfizer, Eli Lilly and Company: 컨소시엄을 구성해 Lilly 임상실험 플랫폼의 API를 개선하고 임상실험연구에 대한 정보 접근성 향상 - N&C: IBM과의 전략적 제휴를 통해 모바일 애플리케이션에서 제공하는 주민참여시스템(civic engagement) 개선 - ISCC: 금융시장 규제에서 정보 공유 문제를 해결하기 위한 다차원 공동체 개발 - SAP: 스탠포드, 독일 국립암센터와의 제휴를 통해 건강 증진 및 맞춤형 제약을 위한 연구개발 추진 - Data&Society 연구소: 빅데이터의 사회적, 문화적, 윤리적 차원에 대한 워크숍 개최 - MIT 빅데이터 이니셔티브: 보스턴 시와 공동으로 보스턴 시내의 택시 수요를 예측하고 대중교통 시스템을 시각화해 이해하자는 취지로 빅데이터 챌린지 대회 개최 - PPFST: 백악관 주도로 OSTP, CDC, DoD와 함께 전염병 예측을 위한 계획 수립 - Splunk4Good: 시민들이 공공 의견에 대한 데이터를 수집하고 분석하는 eRegulation Insights 개발 및 제공

(계속)

* https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/pcast-nitrd-report-2010.pdf (2017.04.11)

** 「미국의 빅데이터 산업 육성정책」, 한국산업기술진흥원, 2016

*** https://www.nitrd.gov/nitrdgroups/index.php?title=Data_to_Knowledge_to_Action/Partnership_Updates (2017.04.11.)

담당부처 (날짜)	정책 (또는 문서)	주요 내용
NITRD ('16.5)	The Federal Big Data R&D Strategic Plan	• 빅데이터 R&D의 7가지 전략과 18가지 세부과제를 발표* (1) 미래 빅데이터 특성을 반영한 기술 개발로 차세대 능력 함양 - 데이터의 크기, 전달·처리 속도, 복잡성에 보조를 맞춘 기술 개발 - 미래에 요구되는 새로운 빅데이터 기술의 방법론 개발 (2) 데이터의 신뢰성 및 더 나은 빅데이터 기반 의사결정을 위한 R&D 지원 - 데이터의 신뢰성과 타당성을 제고시켜 더 나은 결과 도출 - 데이터 기반 의사결정을 지원하는 도구 개발 (3) 빅데이터 혁신을 가능하게 하는 사이버 인프라 구축 및 강화 - 국가 데이터 인프라 강화 - 빅데이터에 대한 응용과학 사이버 인프라 역량 강화 - 유연하고 다양한 인프라 자원 구축 (4) 데이터 공유 및 관리를 촉진하는 정책을 통해 데이터 활용 가치 향상 - 데이터 투명성과 효용성을 증가시키는 메타데이터의 모범사례 개발 - 데이터 자산에 효율적이고 지속적이며 안전한 접근 제공 (5) 개인정보보호, 보안 및 빅데이터의 수집·공유·활용의 윤리적 측면 이해 - 올바른 개인정보보호 - 안전한 빅데이터 사이버공간 구축 - 데이터 거버넌스를 위한 정보윤리 이해 (6) 국가의 빅데이터 교육 및 훈련 환경 개선, 폭 넓은 인력 확충 - 데이터 과학자 양성 - 데이터 영역 전문가 커뮤니티 확장 - 데이터 사용이 가능한 인력 확충 - 공공데이터 활용 역량 개선 (7) 정부기관, 대학, 기업, 비영리 단체와의 협력에 의한 빅데이터 혁신 생태계 지원 - 기관 간 빅데이터 협력 장려 - 빠른 대응과 영향력 측정이 가능한 정책과 정책추진 프레임워크 구축

한편 빅데이터 정책이 활성화됨에 따라 데이터 활용 범위가 민간 영역까지 광범위하게 확대되면서 미국 시민들의 개인정보 침해에 대한 우려가 확산됐다. 이에 2014년 5월 백악관에서는 자문단을 구성해 현재 추진 중인 빅데이터 정책에 대한 검토 및 개인정보보호 강화를 위한 내용이 담긴 연구보고서**를 발표했다. 백악관은 이 보고서를 통해 빅데이터 정책의 긍정적인 면을 부각시킴과 동시에 개인정보 침해 우려에 대한 설문결과와 이를 완화하기 위한 몇 가지 정책들을 제안

* 「미국의 빅데이터 R&D 전략계획」, 정보통신기술진흥센터, 2016

** Executive Office of the President (2014), *Big Data:Seizing Opportunities,Preserving Values*

했다. 우선 레이건 대통령 임기 중에 제정됐던 「전자 커뮤니케이션 정보보호법(Electronic Communications Privacy Act)」의 개정을 요구하고 있다. 180일 이상 저장된 메일 또는 웹 콘텐츠를 수색하기 위해서는 반드시 법원의 영장이 필요하다는 내용을 추가함으로써 경찰 기관의 개인정보 접근을 제한하자는 내용이다. 또한 오바마 행정부가 2011년에 제안한 '사이버보안 입법 제안(Cyber-security Legislative Proposal)' 이행과 2012년에 제안한 '소비자 개인정보 권리 장전(Consumer Privacy Bill of Rights)'의 통과를 촉구하고 있다. '사이버보안 입법 제안'이 이행될 경우 모든 연방기관들은 정보를 검열·수집·이용·보관·공유함에 있어 시민의 자유와 개인정보를 보호하기 위한 절차를 거치도록 법적 프레임워크를 도입해야 하며, '소비자 개인정보 권리 장전'이 통과될 경우 모든 시민은 자신의 개인정보가 무분별하게 수집되는 것에 대한 거부권 행사가 가능해진다. 그리고 교육목적으로 수집된 학생들의 자료가 교육 외에 다른 목적으로 사용되는 것을 주의해야 하며, 연방정부 차원에서 수집 자료를 이용해 시민들에게 차별적으로 서비스를 제공하는 행위에 대한 규제가 필요하다고 지적하고 있다.*

다. 트럼프 당선에 따른 미국의 데이터산업 정책 변화

현재 트럼프 행정부에서는 빅데이터 정책과 관련해 뚜렷한 입장을 보이고 있지 않는 가운데, 지난 2017년 3월 16일에 백악관에서 의회에 제출된 '2018년 회계연도 대통령예산안'**의 부처별 예산 배분내역을 통해 데이터산업 전망을 유추해 볼 수 있다.*** 이번 예산안에서는 국방부(Defense, +10.0%), 국토안보부(Homeland Security, +6.8%), 보훈부(Veterans Affairs, +5.9%)를 제외하고는 전부 예산이 삭감됐으며, 환경보호청(Environmental Protection Agency, -31.4%), 국무부(State, USAID and International Program, -28.7%), 노동부(Labor, -20.7%), 농무부(Agriculture, -20.7%), 보건복지부(Health and Human Services, -16.2%) 등의 예산 감축이 가장 두드러진다.

만약 이 예산안이 이대로 확정된다면 2012년에 발표된 빅데이터 R&D 이니셔티브 참여기관들의 희비가 엇갈릴 전망이다. 국방부에서는 증액된 예산만큼 관련 빅데이터 정책 또한 확장될 수 있는 반면, 58억 달러가 삭감(전년 대비 18% 감소)되는 국립보건원(NIH)의 빅데이터 연구개발 부문은 축소될 것으로 예상된다. 에너지부 또한 17억 달러(전년 대비 6% 감소)의 예산 삭감으로 인해 당초 기획했던 대용량 데이터 분석 관련 기술 개발이 위축될 가능성이 있다. 한편 트럼프는 정책의 최우선 과제 중 하나로 사이버보안 강화를 제시했는데 이에 따라 각 부처의 사이버보안 관련 예산액이 확대돼 데이터 관리·분석·보안 관련 시장이 함께 커질 전망이다.

* 「미국 백악관, 빅데이터 활용과 프라이버시 보호 강화를 위한 빅데이터 정책 검토」, 한국인터넷진흥원, 2014

** Office of Management and Budget (2017), America First: *A Budget Blueprint to Make America Great Again*

*** 의회 승인결과에 따라 예산안 내용이 바뀔 수 있음

2. 영국의 정책

가. 오픈데이터 정책

월드와이드웹 재단에서 발표하는 국가별 오픈데이터 지표 순위에서 3년 연속으로 1위를 기록한 영국은 2000년에 제정된 '정보공개법(Freedom of Information Act 2000)'을 시작으로 적극적인 오픈데이터 정책을 펼치기 시작했다. 2005년 1월부터 발효되기 시작한 이 정보공개법에 따라, 영국의 모든 정부기관은 '정보보호법(Data Protection Act 1998)'에 저촉되지 않는 범위 내에서 자체적으로 생산하거나 외부에서 이관된 모든 정보를 공개해야 할 의무를 지게 됐다. 정보공개의 대상은 중앙정부를 비롯한 지방의 공공기관, 건강·의료 분야의 공적 기관, 경찰, 학교 및 교육기관, 영국은행(Bank of England), 그밖에 정부가 지정하는 기관 등을 포함해 정부가 주주인 국영기업이나 국영기업으로부터 위탁받은 민간기관도 정보공개청구의 대상기관으로 규정하고 있다.* 같은 해 6월에는 유럽연합에서 제정된 '공공정보의 재활용에 관한 지침(Directive on Re-use of Public Sector Information)'을 바탕으로 '공공정보 재활용 규칙(Re-use of Public Sector Information Regulation 2005)'을 제정했으며, 2010년부터 인터넷 포털사이트(data.gov.uk)**를 구축해 본격적으로 공공데이터를 제공하기 시작했다. 2017년 4월 현재 data.gov.uk에는 4만

* http://world.moleg.go.kr/World/WesternEurope/UK/trend/1636?pageIndex=10 (2017.04.11.); 「영국의 공공정보 개방정책과 교육 분야 사례」 재인용, 교육정책네트워크, 2014

** https://data.gov.uk (2017.04.11.)

2,779개의 데이터 세트가 올라와 있으며, 이는 경영·경제, 환경, 지리, 범죄, 교육, 건강, 교통 등 12개 분야로 분류돼 제공되고 있다. 미국의 DATA.GOV와 마찬가지로 데이터 활용을 위한 소프트웨어 애플리케이션도 함께 제공된다.

[표 7-1-2] 오픈데이터 지표 상위 10개 국가

연도	순위									
	1	2	3	4	5	6	7	8	9	10
2013	영국	미국	스웨덴	뉴질랜드	노르웨이	덴마크	호주	캐나다	독일	프랑스
2014	영국	미국	스웨덴	프랑스	뉴질랜드	네덜란드	노르웨이	캐나다	덴마크	호주
2015	영국	미국	프랑스	캐나다	덴마크	뉴질랜드	네덜란드	한국	스웨덴	호주

출처: Open Data Barometer, http://www.opendatabarometer.org (2017.04.11)

이후 2012년 3월에 영국의 기업혁신기술부(BIS)*는 데이터전략위원회(DSB)**를 설립해 공공데이터의 가치창출을 도모했고, 같은 해 6월에는 '오픈데이터 백서(Open Data White Paper)', 2013년 5월과 6월에는 각각 '공공부문 정보에 대한 셰익스피어 검토(Shakespeare Review of Public Sector Information)'*** 및 '셰익스피어 검토에 대한 정부의견(Government Response to the Shakespeare Review)'****을 발표했다. 이 보고서에서는 데이터의 공개와 활용이 영국 정부는 물론 기업을 위한 새로운 기회를 창출하는 중요한 자원이 될 수 있음을 강조하고 있다.***** 그리고 2014년 7월에는 '오픈데이터 전략(Open Data Strategy 2014~2016)'을 발표함으로써 기존의 데이터 활용 전략을 개선했다. 한편 data.gov.uk를 개발한 팀 버너스리(Timothy Berners-Lee)와 나이젤 쉐드볼트는 2012년 정부의 지원으로 오픈데이터 연구소(Open Data Institute)를 설립했는데, 이 연구소에서 2014년에 발표한 '오픈데이터 로드맵 2015(Open Data Roadmap for the UK 2015)'******의 내용은 다음과 같다.

* Department for Business, Innovation and Skills
** Data Strategy Board
*** https://www.gov.uk/government/publications/shakespeare-review-of-public-sector-information (2017.04.11.)
**** https://www.gov.uk/government/publications/government-response-to-shakespeare-review (2017.04.11.)
***** 「2015 데이터산업 백서」, 한국데이터진흥원, 2015
****** https://theodi.org/roadmap-uk-2015 (2017.04.11.)

[표 7-1-3] 영국의 오픈데이터 로드맵 2015

단계	주제	내용
1	일관성 있는 오픈데이터 전략 수립	• 데이터 전략의 넓은 범위 안에 오픈데이터를 포함 • 최고데이터관리자(Chief Data Officer) 임명 • 정부 디지털 서비스 지원을 통해 최고 수준의 데이터 공개 플랫폼 구축
2	더 많은 양의 데이터 공개	• 아직 공개되지 않은 중요한 데이터의 공개를 위한 투자 필요 • 국가정보화기반(National Information Infrastructure)을 활용해 미래 계획 • 정부조달 관련 데이터 공개
3	오픈데이터 재활용 적극 지원	• 정부, 기업, 시민을 위한 데이터 교육 • 오픈데이터 활용 위한 인센티브 제공 • 오픈데이터 R&D 투자

출처: Open Data Institute, https://theodi.org/roadmap-uk-2015 (2017.04.11)

나. 빅데이터 정책

영국의 빅데이터 정책은 오픈데이터 정책의 연장선상에서 방대한 양의 공공데이터 활용에 초점이 맞춰져 있다. 2013년 10월에 발표한 '빅데이터 역량 강화 전략(UK Data Capability Strategy: Seizing the Data Opportunity)'*에 따르면, 영국이 빅데이터 역량을 강화하기 위해서는 빅데이터 전문인력 양성, 빅데이터 인프라와 SW 개발 및 R&D 협력, 그리고 공공데이터의 접근성과 안전성 확보 이렇게 세 가지 핵심 요소를 갖춰야 한다고 서술하고 있다.

2017년 2월, 영국 내각 산하에 위치한 정부디지털서비스(Government Digital Service)는 2020년까지 영국 정부가 이루어야 할 다섯 가지 목표와 이에 필요한 전략들이 담긴 '정부 혁신 전략(Government Transformation Strategy)'**을 발표했다. 이 보고서에서 데이터산업에 대한 영국 정부의 관심을 엿볼 수 있는 부분은 두 번째 목표인 '올바른 사람, 기술 및 문화 양성(Grow the right people, skills and culture)'과 네 번째 목표인 '데이터 활용 개선(Make better use of data)'이다. 두 번

* https://www.gov.uk/government/publications/uk-data-capability-strategy (2017.04.11.)
** https://www.gov.uk/government/publications/government-transformation-strategy-2017-to-2020/government-transformation-strategy (2017.04.11.)

[표 7-1-4] 영국의 빅데이터 역량 강화 전략

주제	내용
빅데이터 전문인력 양성	• 다양한 기관(민간, 대학, 비영리단체, 국립과학기술예술재단, 오픈데이터 연구소 등)과의 협력을 통해 초·중·고등학교 대상 교육 시스템 보강 - 데이터 분석 및 실무 수행에 필요한 컴퓨팅, 프로그래밍 SW 및 HW에 대한 이론적 내용을 포함한 커리큘럼을 구성하고 5세에서 16세까지 학생들의 교육 과정에 적용 - 컴퓨팅 프로그래밍 지도 교사 양성을 위한 펀딩자금(200만 파운드) 조성 • 대학 교육에서는 데이터 분석 역량 교육 및 평가 방식을 진단하고 개선 방안 도출 - 대학 졸업자들이 실제 현장에서 필요로 하는 전문성을 습득할 수 있도록 전문 기업과의 제휴를 통해 빅데이터 기술 교육 프로그램 개발 - 대학 졸업자들이 박사과정 트레이닝 센터(Centres for Doctoral Training)에서 4년 이상의 빅데이터 관련 전문 교육을 추가로 이수할 수 있도록 지원
빅데이터 인프라, SW 개발 및 R&D 협력	• 데이터 스토리지 시장의 글로벌 경쟁력 강화를 촉진하기 위한 정책 전개 - 데이터 스토리지 시장의 해외 투자 및 고객 유치를 지원하고, 동시에 영국 사업자들을 해외 시장으로 진출시키기 위한 방안 도출 • 고성능 컴퓨팅, 클라우드 컴퓨팅 등을 통합해 대량의 데이터 수집·저장·분석·시각화 등 빅데이터의 전 과정을 지원할 수 있는 'e-인프라 시스템' 구축 및 활용 • '빅데이터 분석 국가 네트워크(National Network of Centres in Big Data Analytics)'를 구축하고 민·관·학·연의 R&D 협력 체계를 견고화해 빅데이터 R&D 활동 촉진
공공데이터의 접근성과 안전성 확보	• 영국 정부 산하 연구위원회들로 구성된 TF 주도로 공공데이터 기반 대규모 분석 및 연구를 위한 '공공데이터 연구 네트워크(Administrative Data Research Network)' 발족 • 데이터 공유를 통한 역량 강화 차원에서 연구 데이터 활용·접근을 위한 지원 계획 마련 • 민간 기업에서 보유하고 있는 데이터 개방 범위 확대 • 소비자들의 개인정보를 보호할 수 있는 수단을 마련하되, 기업 활동이나 혁신 창출을 지나치게 제한하지 않도록 조치

출처: 「영국 정부의 빅데이터 역량 강화 전략」, 정보통신산업진흥원, 2014

째 목표를 이루기 위해 영국 정부는 부서별로 '디지털, 데이터 및 기술(DDaT)'*을 가장 잘 조직화할 수 있는 원칙을 수립하고, '디지털 아카데미(Digital Academy)'를 통해 DDaT 전문가를 위한 학습 및 개발 기회를 제공해야 하며, '데이터 사이언스 캠퍼스(Data Science Campus)'와 '데이터 사이언스 엑셀러레이터(Data Science Accelerator)'를 통해 정부의 데이터 사이언스 역량을 강화하는 등의 전략을 제시

* Digital, Data, and Technology

하고 있다. 그리고 네 번째 목표를 이루기 위해 API를 활용해 공공데이터 개방을 유지하고, '디지털 경제 법안(Digital Economy Bill)'의 데이터 공유 조항을 통해 부처 간 데이터 사용에 대한 장벽을 제거해야 하며, 최고데이터관리자(Chief Data Officer)를 임명해 데이터 사용에 대한 권한을 일임할 것을 제안하고 있다. 또한, 공공 부문의 근로자가 데이터 공유와 관련된 윤리를 이해할 수 있도록 하며, 정부 내외의 사용자를 위한 데이터 검색 도구를 개선하는 등의 전략도 함께 제시하고 있다.

다. 브렉시트에 따른 영국의 데이터산업 정책 변화

2016년 6월에 국민투표로 가결됐던 브렉시트(Brexit)가 2017년 3월 29일자로 영국 테리사 메이 총리의 서명과 함께 본격적으로 시작됐다. 이는 영국이 EU에 가입한 지 44년 만에 일어난 일이며, 앞으로 2년 동안 영국과 EU 사이에 탈퇴 협상이 진행될 예정이다. 탈퇴 보상금, 통상 협정 등 다양한 주제에서 난항을 겪을 것으로 예상되는 가운데, 개인정보보호 규정과 관련해 영국이 기존 EU의 방식을 그대로 따라갈지, 아니면 독자적으로 규정을 만들어낼지 관심이 집중되고 있다. 대표적인 예가 '세이프 하버 협정(Safe Harbor Agreement)'이다. EU와 미국이 맺은 개인정보 공유에 관한 이 협정에 따르면, 미국 기업들은 상무부의 세이프 하버에 미리 등록을 하면 EU 시민의 개인정보를 가져오는 것이 가능했다. 그러나 2015년 10월에 유럽사법재판소(European Court of Justice)가 이 협정이 무효라고 판결을 내렸고, 이에 따라 EU에서는 일반데이터보호규정(General Data Protection Regulation)을 제정해 EU 국가들의 「개인정보보호법」의 기준으로 삼고자 했다.

미국 다수의 기업 활동이 영국에서 이루어지고 있다는 점으로 미루어보아 만약 EU를 탈퇴한 영국이 독자적인 개인정보보호 규정을 만들게 되면, 미국 기업 입장에서는 영국에 위치한 데이터센터를 다른 EU 국가로 확장 또는 이전하는 등 새로운 데이터 전략을 수립해야 하는 고민을 안게 될 것이다. 한편, 브렉시트 이후 영국에 신규 데이터센터 4곳을 개관하는 IBM과 같이 브렉시트로 인한 변화를 우려하지 않는 기업도 있다.

한편 2017년 1월에 영국 정부는 브렉시트 이후의 산업 전략을 담은 '녹서

(Building Our Industrial Strategy: Green Paper)*를 발표했는데, 이 보고서를 통해 영국은 잠재적 미래 성장 동인으로 생산성 증대와 균형 성장을 핵심으로 열 가지 주요 전략방안을 제시하고 있다. 그중 첫 번째 전략으로 제시한 '과학·연구·혁신 분야에 대한 투자(Investing in Science, Research & Innovation)에서 영국이 지닌 순수과학 분야의 강점을 살려 외국 기업의 투자를 유치할 수 있다고 서술하고 있다. 대표적인 예로 IBM이 빅데이터 연구 촉진을 위해 Hartree Centre** 연구소에 200만 파운드를 투자한 내용이 제시돼 있다. 이처럼 브렉시트로 인해 많은 변화가 있을 것으로 예상되는 가운데, 특히 오픈데이터와 관련해 가장 선진화돼있는 영국이 앞으로 EU와 어떻게 다른 노선으로 정책을 펼칠지, 새로운 정책으로 인해 전 세계 데이터산업에서 영국의 입지가 어떻게 변화할 것인지 그 귀추가 주목된다.

3. 일본의 정책

가. 오픈데이터 정책***

2015년 국가별 오픈데이터 지표(Open Data Barometer) 순위에서 14위를 기록한 일본의 오픈데이터 정책은 2011년에 본격적으로 시작됐다. 2011년 3월 동일본 대지진 발생 이후 부처 간 수평적 정보 협력에 대한 필요성이 제기됐고, 이에 일본 정부는 2012년 7월에 공공데이터 활용 촉진을 위한 기본전략으로 '전자행정 오픈데이터 전략'을 제정했다. 정부가 직접 공공데이터를 공개, 기계판독이 가능한 형식 사용, 영리/비영리 목적을 불문한 데이터 활용 촉진, 그리고 공공데이터부터 신속하게 공개하는 것을 기본원칙으로 하는 이 전략은 내각관방(内閣官房, Cabinet Secretariat)에서 총괄하며, 총무부(総務省, Ministry of Internal Affairs and Communications), 경제산업부(経済産業省, Ministry of Economy, Trade and Industry) 등 각 부처에서 부처별 전략을 마련하는 식으로 추진됐다.

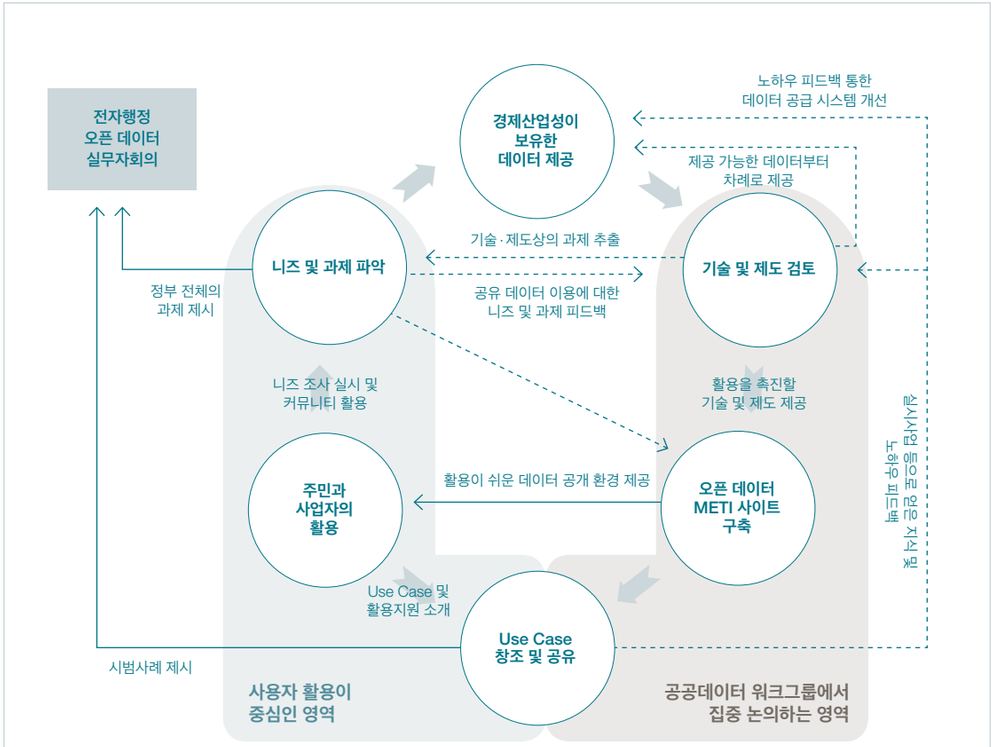
* https://www.gov.uk/government/consultations/building-our-industrial-strategy (2017.04.11.)

** 영국 과학기술위원회(Science and Technology Facilities Council)와 IBM의 합작으로 설립된 슈퍼컴퓨터 및 빅데이터 연구소

*** 「欧州オープンデータ政策に関する最新動向, オープン·データの課題と展望, オープンデータ戦略に係る総務省の取組」, 일본 총무부 정보유통행정국 & NTT DATA, 2012; 「일본, 주요국의 오픈데이터 정책」 글로벌 과학기술정책정보 서비스, 2013 재인용

전략 추진을 위한 사전 검토 작업으로 공공데이터 활용 수요를 파악하고, 데이터 제공 방식을 정비했으며, 공공데이터를 활용한 민간서비스 개발 사례를 축적했다. 2012년 10월, '공공데이터 워킹그룹 2차 회의'에서는 미국의 DATA.GOV, 영국의 data.gov.uk와 유사한 서비스를 제공하는 'DATA METI'* 구상 실현을 위한 로드맵을 제시했다. 이는 경제산업부가 보유한 데이터를 기업이 활용할 수 있도록 개방해 경제 활성화를 유도하고자 하는 정책으로 이를 그림으로 표현하면 [그림 7-1-1]과 같다. 'DATA METI'는 2013년 1월에 구축돼 2017년 4월 현재 3,691개의 데이터 세트가 올라와 있다.

[그림 7-1-1] 일본 경제산업부의 DATA METI 구상



출처: http://datameti.go.jp/datameti (2017.04.11); 「일본, 주요국의 오픈데이터 정책」 재인용, 글로벌 과학기술정책정보 서비스, 2013

* http://datameti.go.jp/datameti (2017.04.11.)

한편 2016년 7월에 일본 총무부에서 작성한 '2016년 정보통신백서'에 따르면, 총무부는 오픈데이터 활용 활성화를 위해 정보 유통 연계 공통 API 확립, 데이터 2차이용 지침 등 오픈데이터의 유통 환경 마련을 위한 다양한 정책 사업을 추진하고 있다. 사회와 시장에 존재하는 빅데이터와 오픈데이터를 상호 연결하는 '오픈데이터·빅데이터 활용 추진 사업'을 실시하고, '오픈&빅데이터 활용·지방 창생 추진기구'와 연계해 오픈데이터에 관한 기술 사양이나 사용규칙을 제시하는 등 산·학·관 협력연구체계를 구축하는 것이 그 예이다.*

나. 빅데이터 정책

일본은 2005년 문부과학부(文部科学省, Ministry of Education, Culture, Sports, Science and Technology)에서 '정보폭발(info-plosion) 시대에 대응하는 새로운 IT 기반 기술 연구 프로젝트'**를 수행한 것을 시작으로, 2007년 경제산업부의 '정보대항해(情報大航海) 프로젝트'***, 2011년 일본학술진흥회(日本学術振興会, Japan Society for the Promotion of Science)의 '초대형 데이터베이스 시대에 대응하는 최고속 데이터베이스 엔진 개발 및 해당 엔진을 핵심으로 한 전략적 사회 서비스 실증 평가'로 이어지며, 2012년 5월 총무부에서 발표한 '빅데이터 활용 방안' 이후로 구체적인 빅데이터 정책이 추진되기 시작했다. 여기서 발표한 내용은 2012년 7월에 총무부가 발표한 '일본 ICT 활성화 전략(Active Japan ICT)'에 반영이 돼 5대 중점 전략 중 하나인 '데이터 활성화(Active Data)' 전략으로 발표됐다. 이 데이터 활성화 전략은 [표 7-1-5]와 같이 7대 세부추진 과제로 구성된다. 이후 2014년 6월, 경제산업부는 빅데이터 서비스가 주로 공급자 관점에서 발전되고 있는 문제점 해결을 위해 '데이터 기반 혁신 창출 전략 협의회'를 발족해 이용자 관점에서 빅데이터 활용 방안을 모색하고자 했다.

* 「일본 2016 정보통신백서 조사분석」, 정보통신기술진흥센터, 2016

** 본 프로젝트는 2005년부터 2011년까지 추진되며, 폭증하는 정보로부터 필요한 정보를 추출하는 기술, 대량의 정보를 안전하고 안정적으로 관리·운용하는 기술, 인간과 유연한 상호작용으로 정보를 활용할 수 있는 기술, 정보를 활용해 선진적인 IT 서비스를 사회에 적용하기 위한 첨단기술 개발을 목표로 함 (「주요국의 빅데이터 추진전략 분석 및 시사점」, 한국정보화진흥원, 2013)

*** 본 프로젝트는 정보의 종류에 의존하지 않고, 많은 정보 중에서 사용자가 원하는 정보를 정확하게 검색·분석하는 일반적인 기술 개발을 목표로 추진하고 있음 (「주요국의 빅데이터 추진전략 분석 및 시사점」, 한국정보화진흥원, 2013)

[표 7-1-5] 데이터 활성화 전략을 위한 7대 추진 과제

추진 과제	추진 내용
데이터 개방 및 활용 가능한 환경 마련	• 오픈데이터 전략을 추진해 공공 및 민간 데이터 개방 • 오픈데이터 환경 기반 마련을 위한 국제 표준화 추진 • 데이터 재활용에 관한 법·제도 정비 • 개인정보보호를 기반으로 활용 가능한 가이드라인 제공
데이터 신뢰성·안정성 확보를 위한 연구개발	• 데이터 신뢰성·안전성을 확보하면서 데이터의 효율적인 수집, 해석 등이 가능하도록 통신 프로토콜, 보안 대책, 데이터 구조 등 연구개발 추진 • M2M, 매쉬, 센서 네트워크, 자동차용 무선통신 형태 연구개발 및 표준화 추진
데이터 사이언티스트 육성	• 데이터 분석을 위한 데이터 사이언티스트 육성
빅데이터 비즈니스 창출을 위한 M2M 보급 촉진	• M2M 통신제어 기술 확립 • 신뢰성·안정성 높은 통신규격 개발 및 표준화 추진
법·제도 정비	• 빅데이터의 효율적 활용 및 활성화를 위한 법·제도 정비
추진체계 정비	• 산·학·관 협력을 통해 활용 가능한 데이터 및 성공사례 공유체계 마련 • 데이터 자원의 수집, 활성화 방안 마련 등을 위한 인센티브 제도 도입
글로벌 협력 강화	• 유럽, 미국 등 빅데이터 활용 추진 국가들과 상호협력 체계 마련 • 빅데이터의 데이터 량과 활용에 따라 발생하는 경제적 가치 계측 및 평가방법 확립

출처: 「주요국의 빅데이터 추진전략 분석 및 시사점」, 한국정보화진흥원, 2013

한편 일본 정부는 2020년 이후의 미래에 대비해 범정부 차원에서 제4차 산업혁명 모델을 구상했으며 이는 2016년 6월에 '일본재흥전략(日本再興戦略) 2016'*이라는 이름으로 발표됐다. 본문만 226페이지에 달하는 방대한 분량의 이 보고서에서는 제4차산업혁명의 핵심 플레이어로 사물인터넷(IoT), 빅데이터, 인공지능(AI), 로봇 기술을 지정하고 이를 통해 2020년까지 30조 엔의 부가가치를 창출할 것으로 기대하고 있다. 이는 장차 데이터 주도 사회 실현을 위해서는 빅데이터 활용이 주요 역할을 수행하게 될 것이며, 빅데이터 수집을 위한 수단이

* 매년 수정되는 아베노믹스의 기본전략으로 제4차산업혁명을 밀려오는 변화와 도전으로 간주하고, 사물인터넷, 빅데이터, 인공지능이 가져올 충격에 대한 종합적인 로드맵과 민간이 공유하는 나침반이 될 비전을 제시(「제4차산업혁명을 리드하는 일본 정부의 추진 전략과 정책 시사점」, 한국표준협회, 2016)

사물인터넷이라는 판단 하에 만들어진 정책으로 평가된다. 목표를 실현하기 위해 아베 총리가 의장으로 있는 '산업경쟁력회의'를 중심으로 민관전문가 위원회를 구성해, 규제완화 도입, 데이터 공유·이용 촉진, 일본 내 혁신창조, 인적자원개발에 초점을 맞추어 추진할 계획이다.

[표 7-1-6] 제4차산업혁명을 향한 일본 정부의 로드맵

추진 전략	추진 내용
규제완화 도입	• 행정절차의 IT화 및 간소화
데이터 공유·이용 촉진	• 기업·조직을 넘어 빅데이터의 공유와 이용을 위한 공통 플랫폼 창출 • IT 이용 촉진을 위한 중소기업 지원 강화
일본 내 혁신 창조	• 최첨단 연구개발 관련해 최상위 전문가와 5개 R&D 센터 구축
인적자원개발	• 초등학교에 컴퓨터 프로그래밍 의무교육 도입 • IT 숙련을 통한 IT 이용에 의한 학습 도입 • 일본 내 영속적 거주 권리를 위한 그린카드로 비 일본인 전문가를 고용하기 위한 신속한 위임 절차 마련

출처: 「제4차산업혁명을 리드하는 일본 정부의 추진 전략과 정책 시사점」, 한국표준협회, 2016

제4차산업혁명과 관련해서 행정부처의 세부 추진계획은 다음과 같다. 경제산업부에서 발표한 '신산업구조비전, 제4차산업혁명을 리드하는 일본'*에서는 일본 정부의 제4차산업혁명에 대한 7가지 대응방침을 제시하고 있다. 이 문서에서도 데이터 이용·활용·촉진을 위한 환경 정비를 첫 번째 대응방침으로 내세우며 제4차산업혁명의 핵심이 데이터에 있음을 강조하고 있다. 구체적으로는 데이터 플랫폼의 구축, 데이터 유통시장의 창출, 개인 데이터 이용·활용·촉진, 보안기술 또는 인재를 키워내는 생태계 구조, 제4차산업혁명의 지적재산 정책의 방향을 언급하고 있다. 문부과학부에서는 2016년에서 2025년까지 1,000억 엔의 예산을 투입해 인공지능, 빅데이터, 사물인터넷, 사이버보안 통합 프로젝트(Advanced Integrated Intelligence Platform)를 수행하는 계획을 수립했으며, 사물인터넷을 통한 빅데이터의 실시간 수집·축적, 인공지능을 통한 신속한 판단으로 시스템을 제어하는 '보안 5.0(Security 5.0)' 구현을 비전으로 제시했다. 또한 문부과학부는 '제4차산업혁명을 향한 인재육성 종합 이니셔티브'를 발표해 인공지능, 사물인터넷, 사이버보안 및 그 기반이 되는 데이터 사이언스 등 인재육성 확보에 필요한 종합 프로그램을 체계적으로 실시 운영해야 할 것을 강조하고 있다. 특히 수리·정보 교육의 강화를 통해 고급 인재를 양성하는 방안으로 전 학년 교육연구조직을 정비하고, 수리·정보 관련 학부·대학원을 신설하고 교육정원을 확대해 전문교육을 중점 지원하며, 데이터 활용 분야를 선도할 수 있는 데이터 사이언티스트 및 사이버보안 인재 육성을 제시하고 있다. 구체적으로는 박사과정 또는 박사후과정의 젊은 인재를 대상으로 데이터 사이언스 기술을 획득하는 기회를 제공하고, 산업과 연계해 단기연수 시스템을 통해 데이터 사이언티스트로서의 능력 습득 및 경력개발을 지원하는 정책을 예로 들 수 있다.

* 「新産業構造ビジョン, 第４次産業革命をリードする日本の戦略」, 經濟産業省, 2016

주요 국가의 개인데이터 활용과 데이터 유통 활성화 정책 동향

데이터 유통 활성화를 위해서는 데이터 경제의 핵심이 되는 개인데이터의 개방 및 활용이 중요하다. 그러나 개인정보 보호라는 명목 하에 그동안 각종 법과 규제에 의해 개인데이터의 활용이 제한됐다.

이에 영국은 2011년 4월에 발표한 'Better Choice: Better Deals'에서 기업이 관리하고 있는 개인데이터를 해당 소비자에게 제공한다는 내용의 '마이데이터(midata)' 프로그램을 발표했다. 19개 주요 기업 및 소비자 단체 등과 협약을 맺고 개인데이터를 제공하기 위한 자발적인 프로그램 운영을 통해 소비자·기업·정부가 상생할 수 있는 개인데이터 생태계를 조성하고 데이터 유통과 관련한 신시장 창출을 도모했다.

미국은 2011년 9월 발표한 'Informing Consumers through Smart Disclosure'를 통해 '스마트공개(smart disclosure)' 촉진을 위한 지침을 제시, 시장에서 소비자들이 보다 현명한 의사결정을 내릴 수 있도록 개인데이터를 제공했다. 스마트공개의 대상이 되는 개인데이터는 접근성, 기계가독성, 표준화, 적시성, 시장 적응성 및 혁신, 상호운용성 그리고 개인식별정보 및 사생활 보호의 7가지 기준을 만족해야 한다. 미국 정부는 이를 통해 개인의 합리적 소비로 인한 시장 혁신뿐만 아니라 소비자가 현명한 의사결정을 내릴 수 있도록 지원하는 새로운 시장과 부가가치 창출을 기대했다.

한편, 일본에서는 지난 5월 23일 니혼게이자이신문을 통해 2020년에 세계 최초로 사물인터넷(IoT) 데이터를 매매하는 빅데이터 거래소를 개설한다고 발표했다. NTT, 히타치, 도쿄전력 등 일본의 민간기업 100개 사가 참여하는 이 프로젝트를 통해 일본은 현재 미국 기업 위주의 IoT 시장에서 데이터 유통 활성화를 통해 시장을 주도하겠다는 목표다. 이 과정에서 우려되는 점은 본인의 동의 없이 개인데이터의 거래를 금지하는 일본의 개인데이터 법과 규제다. 미국과 영국의 예시에서 보여주듯이 성공적인 거래소 정착을 위해서는 개인데이터 활용 규제에 대한 보완책을 마련해야 할 것이다.

제2장

국내 데이터 관련 법·제도 현황

정부는 데이터산업을 주요 전략산업으로 선정하고 적극적인 육성 정책을 추진하는 동시에 데이터의 인프라화를 통해 지능정보사회를 준비하고 있다. 법제 환경의 경우 데이터를 사용 권한에 따라 개방형·공유형·폐쇄형으로 구분하고, 각 영역을 관장하는 법률 현황을 분석해 데이터에 초점을 맞춘 최근의 제도 개선 노력을 소개한다.

1. 정부 시책 현황

가. 데이터산업 육성

미래창조과학부는 2015년 'K-ICT전략' 발표를 통해 빅데이터를 9대 전략산업 중 하나로 선정하고, 세계 3대 빅데이터 강국 도약을 목표로 빅데이터 선도 프로젝트와 전문인력 양성, 데이터 전문기업 육성 등을 위한 사업을 수행하고 있다. 특히 데이터 기반 심야버스 노선 수립 지원 등 공공과 민간이 함께 참여하는 빅데이터 선도 프로젝트를 통해 사회 현안을 해결하고 고부가가치를 창출하는 성공사례를 제시하고 있다. 또한 'K-ICT 빅데이터센터'(www.kbig.kr)를 운영해 빅데이터 분석 역량을 갖추지 못한 개인과 기업도 빅데이터를 활용할 수 있도록 빅데이터 인프라와 실습 환경을 지원하고 있다.

* 필자: 김형건(한국데이터진흥원 정책기획실 선임연구원)

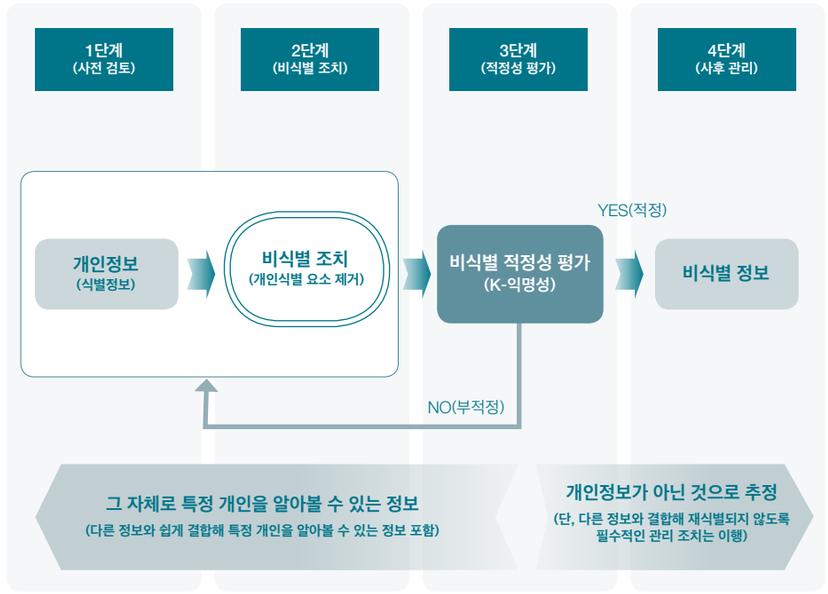
[표 7-2-1] 빅데이터 선도 프로젝트

연도	프로젝트 내용	참여기관
2015	제조 프로세스 분석을 위한 빅데이터 클라우드 서비스	현대중공업 UNIST
	빅데이터 기반 선박 신수요 예측 플랫폼 및 MRO 서비스 모델 개발	대우조선해양 더존비즈온
	빅데이터 분석을 통한 소비 트렌드 분석 및 예측 플랫폼 구축	비씨카드 LG CNS
	빅데이터를 활용한 스마트 에너지 관리 서비스	에스지에이 SKT, 엔코디
	빅데이터를 활용한 스마트 전시 컨벤션 서비스 구축	한화S&C 코엑스
2016	로밍 빅데이터를 활용한 해외 유입 감염병 차단 서비스	KT 질병관리본부
	빅데이터 딥러닝 기술 활용 스마트 T-커머스 서비스 개발	더블유쇼핑 한동대
	유가공 업종 제조 생산 최적화를 위한 빅데이터 플랫폼 개발	매일유업 한국그린비즈니스협회
	딥러닝 기술 기반의 대용량 제조 데이터 분석 서비스 플랫폼 개발	유라 충북대학교

출처: 「교통·의료 사회현안 빅데이터로 해결한다」, 미래창조과학부, 2017.03.15.

정부는 빅데이터 산업 활성화와 개인정보보호가 균형을 이룰 수 있도록 개인정보보호법 등 관련 법제의 개선을 함께 추진하고 있다. 그 일환으로 2016년 5월 규제개혁장관회의를 통해 빅데이터 분야 개인정보 규제혁신 방안을 마련하고, 같은 해 6월 범부처 합동으로 '개인정보보호법령 통합 해설서' 및 '개인정보 비식별 조치 가이드라인'을 발표했다. 통합 해설서와 가이드라인은 산재된 개인정보 관련 법령의 통일된 해석을 지원하고 비식별 조치의 기준과 비식별 정보의 활용 범위 등을 명확히 하기 위한 것이다. 이와 더불어 분야별 비식별 조치 전문기관을 선정해 비식별 조치, 기업 간 데이터 결합, 비식별 조치 적정성 점검, 교육·컨설팅 등을 지원하고 있다.

[그림 7-2-1] 비식별 조치 및 사후관리 절차

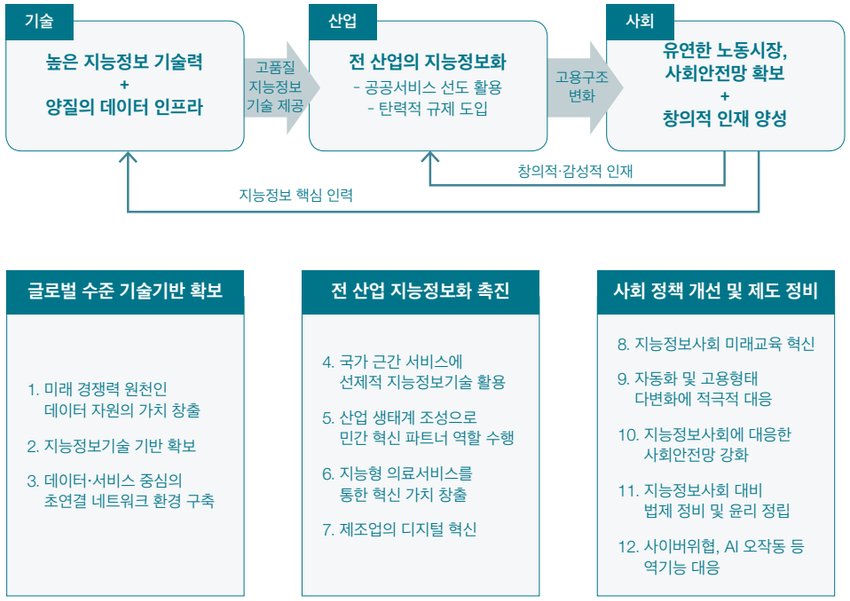


출처: 「개인정보 비식별 조치 가이드라인」, 관련 부처 합동, 2016

나. 데이터 인프라 구축

데이터의 사회적·경제적 가치 확대와 제4차산업혁명의 도래에 따라 데이터의 인프라 자원으로서의 가치가 강조되고 있다. 정부는 2016년 12월 제4차산업혁명의 주도권을 선점하고 지능정보사회 실현을 앞당기기 위한 범부처 플랜인 '지능정보사회 중장기 종합대책'을 발표했다. 종합대책은 지능정보사회 도래가 가져올 경제와 사회적인 변화를 조망하고 전문가 의견 수렴을 통해 '기술→산업→사회'로 연결되는 중장기 정책 방향을 제시했다. 특히 데이터를 산업 관점이 아닌 미래 경쟁력의 원천이자 기초 인프라로 상정하고 최우선 추진과제로 데이터 자원의 가치 창출을 선정했다.

[그림 7-2-2] 지능정보사회 중장기 종합대책 추진 전략



출처: 「지능정보사회 중장기 종합대책」, 미래창조과학부, 2016.12.27.

먼저 국가적인 데이터 관리체계를 확립해 기계가 학습할 수 있는 대규모 데이터를 구축할 계획이다. 정부 보유 공공데이터는 머신러닝이 가능한 오픈포맷으로 전환·개방하고, 스마트시티에서 생산된 사물인터넷(IoT) 기반 센서 데이터의 개방체계를 구축할 예정이다. 이 밖에 클라우드 이용 선도 프로젝트, 공공과 민간데이터의 현황을 알려주는 데이터맵 구축 등이 포함돼 있다.

둘째, 유형별 데이터가 프라이버시 침해 없이 안전하게 유통·활용되도록 차별화된 전략을 추진한다. 데이터의 자유로운 연결과 이동이 가능한 환경을 조성하기 위한 것이다. 우선 개인정보와 무관한 일반 데이터는 적정한 가치로 거래될 수 있도록 데이터 거래소를 도입할 계획이다. 현재 데이터 거래소 기반 마련을 위해 데이터 유통 플랫폼인 데이터스토어(www.datastore.or.kr)를 개방형 데이터 플랫폼으로 전환 중이다. 기업이 보유한 개인정보를 정보주체의 동의를 얻어 공유·활용할 수 있도록 하는 'K-MyData 제도' 도입 또한 핵심 추진과제에 포함돼 있

다. 이 밖에 기업들이 자유롭게 데이터 결합을 테스트해볼 수 있는 데이터 프리존 운영을 통해 비식별 처리된 개인정보의 안전한 유통과 활용을 지원한다.

마지막으로 데이터 구축 기반 조성을 위한 세부 과제로 데이터 전문 기업 및 인력 육성, 블록체인 기술 활용 지원 등을 추진할 계획이다. 데이터 수집·가공·정제 등을 수행하는 데이터 거래 전문 서비스 기업 육성과 현안 해결 중심의 데이터 과학자 양성을 목표로 선정했다. 또한 금융거래 외 공공기록물 관리, 저작권 관리 등으로 블록체인 활용 분야를 확대해 데이터 관리의 신뢰성을 제고할 계획이다.

다. 공공데이터 개방

정부는 2013년부터 공공데이터 개방 정책을 본격 추진해 오고 있다. 우선 「공공데이터의 제공 및 이용 활성화에 관한 법률」을 제정('13.7)해 제도적 기반을 마련했다. 또한, '국가오픈데이터포럼'을 출범('13.7)하고 '공공데이터활용지원센터'를 개소('13.11)하는 등 정책 추진체계를 함께 구축했다.

제1차 공공데이터 제공 및 이용활성화 기본계획('14~'16)은 공공데이터의 양적 확대와 개방·활용 인식 제고에 초점을 맞추었다. 국민이 직접 선정한 33개 국가 중점 데이터 등 2016년까지 총 695개 기관 20,679개의 데이터가 전면 개방됐다. 공공데이터 포털(data.go.kr)을 통해 총 1,663,390건의 데이터가 다운로드됐으며, 공공데이터를 활용한 서비스 또한 대폭 증가했다. 그 결과 공공데이터의 가용성, 접근성, 정부지원 정도를 평가하는 OECD 공공데이터 개방지수(OUR data Index)에서 조사국 30개국 중 1위를 기록하는 등 오픈데이터 리더 국가로 발돋움하게 됐다.

2016년 12월 공공데이터전략위원회와 국무회의를 거쳐 제2차 기본계획('17~'19)이 확정됐다. 1차 계획이 공공데이터 개방의 양적 확대와 인식 제고에 맞춰졌다면, 2차 계획은 데이터 기반 산업생태계 확산과 경제적 부가가치 창출이 주된 목표다. 정부 투명성과 국민의 알권리를 높이기 위한 수단으로 활용되던 공공데이터가 창업과 비즈니스 창출의 수단으로 진화한 환경 변화를 반영한 결과다.

[표 7-2-2] 제2차 공공데이터 기본계획 비전 체계

구분	내용
비전	데이터로 국민과 기업이 풍요로운 디지털 사회
목표	1) 데이터 기반의 산업 생태계 확산으로 새로운 부가가치 창출 2) 생활 속 데이터 활용 확산으로 보다 윤택해진 국민생활
주요 전략 방향	1) 질 높은 데이터 개방 및 산업 생태계 성장 지원 2) 국민참여 확대와 데이터 개방·활용 생활화 3) 데이터 기반의 플랫폼 정부 및 민관 거버넌스 조성
추진 영역 및 과제	**1. 기업·신산업에 꼭 필요한** 1) 융합형·지능형 고품질 데이터 개방 확대 2) 신산업 육성을 위한 데이터 활용 생태계 조성 3) 데이터 유통·거래 기반 조성 4) 공공데이터 활용기업 해외진출 지원 확대 **2. 국민의 눈높이에서** 1) 공공데이터 제공 및 활용에서의 국민 참여기반 조성 2) 사회적 현안 해결을 위한 데이터 활용 강화 3) 전 국민 공공데이터 활용 역량 제고 **3. 데이터 기반의 정부로** 1) One Gov 방식의 공공데이터 관리체계 구축 2) 민간 데이터·서비스 공동 활용을 통한 민관협업 체계 구축 3) 생애주기별 공공데이터 품질관리 강화 4) 범정부 공공데이터 성과관리 강화 5) 글로벌 공공데이터 파트너십 확대

출처: 「제2차 공공데이터 기본계획」, 공공데이터전략위원회, 2016.12.15.

2차 기본계획의 최우선 추진과제는 신산업에 활용성이 큰 융합형·지능형 데이터 개방의 확대이다. 차세대 국가성장을 견인할 자율주행, 가상현실(VR) 등 신산업 분야를 선정하고 관련 공공데이터 개방을 확대할 방침이다. 공공, 민간데이터를 자유롭게 유통·거래하는 오픈소스 기반의 플랫폼을 조성하고, 거래 활성화를 위한 거래시장 조성도 지원한다. 또한 국민 스스로 데이터 개방의 결정권을 행사할 수 있도록 하는 '개인데이터 개방 자기 결정 및 참여제도'의 도입도 추진할

예정이다. 이 밖에 데이터를 기반으로 부처 연계·통합을 강화하고, 국가적으로 통합적 관리가 필요한 주요 핵심 데이터를 대상으로 일원화된 '공공데이터 통합 관리체계(One Gov.)'를 마련할 계획이다.

라. 새 정부의 정책 방향

제4차산업혁명은 19대 대통령 선거의 화두였다. 새 정부 경제 정책의 핵심은 대규모 재정 투입을 통한 신성장 동력 확보와 양질의 일자리 창출이다. 새로운 성장 동력 확보는 결국 제4차산업혁명에 대한 대응 수준에 달려 있다. 이를 위해 새 정부는 대통령 직속 '4차산업혁명위원회' 설치를 공약으로 제시했다. 현재 각 부처별로 산재된 정책을 통합 추진하고 규제 개혁에 힘을 더해 제4차산업혁명 대응에 국가역량을 결집하기 위한 포석이다. 이에 더해 대선 공약집은 빅데이터 전문인력 양성과 세계 최고의 사물인터넷 기반 마련을 통해 제4차산업혁명을 가능하게 하는 ICT 생태계 조성 계획을 밝히고 있다.

혁신 창업국가는 미래 성장 동력 확충을 위한 새 정부의 또 다른 비전이다. 창업 촉진을 위해 공공빅데이터센터를 설립하고 국토공간정보의 무료 제공 등 공공데이터 개방을 확대해 데이터를 창업과 비즈니스의 소재로 제공할 계획이다. 또한 빅데이터를 포함한 신산업 분야에 대한 규제를 네거티브 방식으로 전환해 기업의 혁신과 창의를 최대한 보장할 것으로 기대된다.

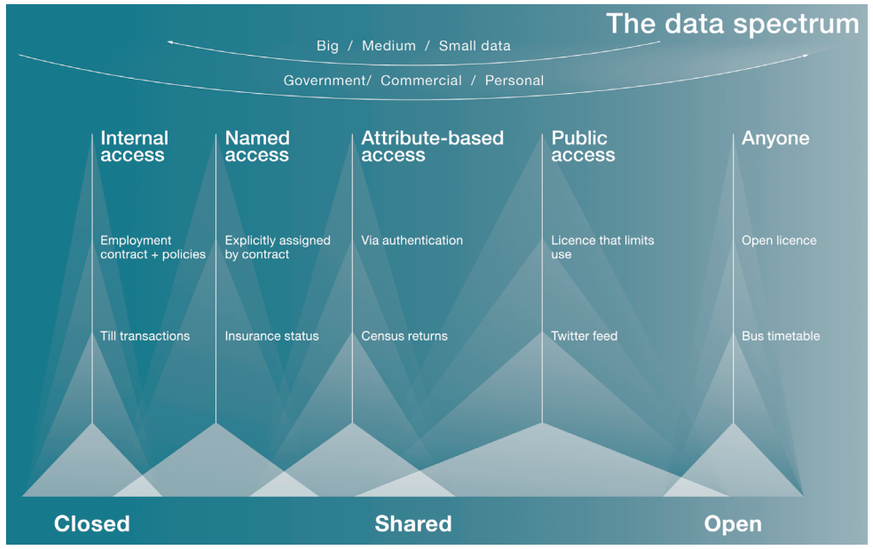
2. 법제 현황

가. 현황 분석

국내는 아직 데이터를 직접적으로 다루고 있는 법률이 존재하지 않는다. 2013년에 제정된 공공데이터법은 공공데이터의 제공과 이용만을 대상으로 한다. 최근 데이터베이스산업과 빅데이터산업 진흥을 위한 법률 제정 시도가 있었지만 국회를 통과하지 못했다. 클라우드, 사물인터넷(IoT) 등 데이터 기반 신기술의 등장에 맞춰 관련 입법 노력이 계속되고 있으나 산업 진흥 관점의 접근에 머무르는 한계를 보이고 있다. 제4차산업혁명에 적합한 ICT 환경을 마련하기 위해서는 데이터 고유의 특성과 데이터의 인프라 자원으로서의 가치를 반영한 새로운 법체계의 필요성이 제기되고 있다.

데이터 정책에 있어 선도적인 행보를 보이고 있는 영국은 모든 데이터를 사용권한에 따라 개방형 데이터, 공유형 데이터, 폐쇄형 데이터 세 가지로 분류(데이터 스펙트럼)하고, 각각의 기능을 강화하는 차별화된 정책을 추진하고 있다. 누구나 사용이 가능한 개방형 데이터 영역은 공공 부문 중심의 완전한 데이터 공개 확대를 정책 목표로 한다. 반면 데이터 사용권한이 소유자 등으로 최소화한 폐쇄형 데이터 영역은 신뢰할 수 있는 데이터 보호 환경 조성을 핵심으로 한다. 인증된 사용자에게 사용 권한이 부여된 공유형 데이터 영역은 데이터 자본화와 유통시장 활성화를 통한 자유로운 데이터 이동 환경 조성을 지향한다.

[그림 7-2-3] 사용권한에 따른 데이터 분류



구분	폐쇄형 데이터	공유형 데이터	개방형 데이터
사용권한(예시)	소유자 (영업비밀, 민감개인정보)	인증된 사용자 (판매 데이터)	누구나 (개방 공공데이터)
정책 목표	신뢰할 수 있는 데이터 및 프라이버시 보호	데이터 이동과 공유가 자유로운 환경 조성	공공 부문 중심의 완전한 데이터 공개 확대

출처: ODI, https://theodi.org/data-spectrum

국내 데이터 관련 법체계를 데이터 스펙트럼에 대입할 경우 개방형 데이터 영역은 공공데이터법이, 폐쇄형 데이터 영역은 「개인정보보호법」이 제도적 기반을 이루고 있다. 반면 데이터의 이동과 유통을 관장하는 공유형 데이터 영역에 대한 독립 법제는 현재까지 존재하지 않는다. 최근 영국, 프랑스 등 유럽국가를 중심으로 데이터 이동권과 공익 목적의 민간데이터 공유 의무를 법제화하는 등 공유형 데이터 영역의 기능 강화를 위한 제도 개선 노력이 확대되고 있다. 국내 역시 데이터의 원활한 이동과 이를 통한 산업 간 융합 활성화를 위해 데이터 중심의 법제 환경 마련이 필요한 때다.

나. 개방형 데이터 관련 법률

빅데이터 이용 활성화를 위해 공공데이터의 개방은 선택이 아닌 필수다. 「공공데이터의 제공 및 이용 활성화에 관한 법률」(이하 '공공데이터법')은 국민이 공공데이터를 최우선적으로 이용할 수 있도록 보장하고, 공공기관에 공공데이터 제공의무를 부여해 효과적인 민간 제공과 이용 활성화를 지원할 수 있는 법적 근거를 마련하기 위해 2013년 제정됐다. 본 법은 '공공기관이 보유·관리하는 데이터의 제공 및 그 이용 활성화에 관한 사항을 규정함으로써 국민의 공공데이터에 대한 이용권을 보장하고, 공공데이터의 민간 활용을 통한 삶의 질 향상과 국민경제 발전에 이바지함'을 목적으로 하고 있다. 여기서 공공기관이란 국가기관, 지방자치단체 및 「국가정보화 기본법」 제3조 제10호에 따른 공공기관을 말한다. 또, 공공데이터란 데이터베이스, 전자화한 파일 등 공공기관이 법령 등에서 정하는 목적을 위해 생성 또는 취득해 관리하고 있는 광(光) 또는 전자적 방식으로 처리된 자료 또는 정보를 의미한다.

공공데이터법은 공공데이터 제공과 이용의 기본원칙을 포함해 기본정책의 수립과 추진체계, 공공데이터 등록 등 제공기반 조성, 공공데이터의 제공 절차 등을 주요 내용으로 한다. 우선 공공데이터에 관한 정부 주요 정책의 심의·조정 등을 국무총리 산하의 공공데이터전략위원회가 담당하도록 했다. 공공기관은 원칙적으로 「공공기관의 정보공개에 관한 법률」에 따른 비공개 대상 정보 등을 제외한 모든 보유 공공데이터를 국민에게 제공해야 한다. 또한 공공데이터의 효율적인 제공과 편리한 이용을 위해 '공공데이터활용지원센터'와 '공공데이터 포털' 운영에 관한 근거를 규정하고 있다. 이 외에 공공데이터 품질관리, 표준화, 공공데이터 제공에 관한 분쟁조정 등 공공데이터 관리 및 이용·제공에 필요한 사항들이 포함돼 있다.

2016년에는 활용 중심의 공공데이터 생태계 조성을 위한 법 개정을 단행했다. 우선 공공데이터를 활용한 창업을 촉진하고 창업자의 성장·발전을 위해 필요한 지원을 할 수 있도록 했다. 또한 공공기관의 장이 공공데이터를 활용해 개인·기업 또는 단체 등이 제공하는 서비스와 중복되거나 유사한 서비스를 개발·제공하는 것을 금지해 민간의 창업 의욕 저하와 시장 왜곡을 방지했다. 행정자치부장관은 중복·유사 서비스의 현황을 주기적으로 조사해 정책의 실효성을 높이도록 했다.

공공데이터법은 공공데이터 개방 정책에 안정적인 추동력을 제공하고, 데이터 활용 중심의 정책 환경 변화를 반영하는 등 국내 데이터산업의 초석을 다지고 오픈데이터 정책의 선진화를 앞당겼다는 평가를 받는다. 이 밖에 「국가공간정보기본법」, 「기상법」 등 분야별 공공데이터의 이용·제공과 활용에 관해 규정하고 있는 개별법들이 존재한다.

다. 폐쇄형 데이터 관련 법률

2011년 공공부문과 민간부문을 망라하는 개인정보보호에 관한 일반법인 「개인정보보호법」이 제정됐다. 개인정보의 가치 증대로 사회 전 영역에서 개인정보의 수집과 이용이 일반화되고 있는 상황에서 국가 차원의 개인정보보호 정책 추진체계를 마련하고 이 전의 개별법 규율 체계에서 발생 가능한 개인정보보호의 사각지대를 방지하기 위한 방편이었다. 현재 개인정보보호 관련 법률은 일반법인 「개인정보보호법」 외에 여러 개별법에 산재해 있다.

[표 7-2-3] 국내 개인정보보호 관련 법률 현황

구 분		법률명	규율대상
일반법		개인정보보호법	개인정보
개별법	공공부문	전자정부법	행정정보, 개인정보
		주민등록법	주민등록전산정보
		공공기관의 정보공개에 관한 법률	행정정보
		공공기록물 관리에 관한 법률	기록물정보
	전기통신	정보통신망 이용촉진 및 정보보호 등에 관한 법률	개인정보
		위치정보의 보호 및 이용 등에 관한 법률	(개인)위치정보
		통신비밀보호법	감청자료, 통신사실확인자료
		전기통신사업법	통신비밀,통신자료
	전자거래	전자문서 및 전저거래 기본법	전자문서, 개인정보, 영업비밀
		전자상거래 등에서의 소비자 보호에 관한 법률	신원정보, 거래정보, 개인정보
		전자서명법	전자서명생성정보, 개인정보
	금융	전자금융거래법	금융정보, 개인정보
		신용정보의 이용 및 보호에 관한 법률	(개인)신용정보
		금융실명거래 및 비밀보장에 관한 법률	금융거래비밀
		보험업법	개인정보
		자본시장과 금융투자업에 관한 법률	투자자정보
	보건의료	의료법	의무기록, 개인정보
		약사법	개인정보
		감염병의 예방 및 관리에 관한 법률	질병정보, 개인정보
		국민건강보험법	보험정보, 개인정보

「개인정보보호법」은 개인정보의 처리 및 보호에 관한 사항을 정함으로써 개인의 자유와 권리를 보호하고, 개인의 존엄과 가치를 구현함을 목적으로 한다. 법률의 주요 내용은 본 법의 적용 대상을 공공·민간부문의 모든 개인정보 처리자로 하고, 개인정보의 수집, 이용, 제공 등 단계별 보호 기준을 마련했다. 개인정보 영향평가제도를 도입하고, 고유식별정보의 처리와 영상정보처리기기의 설치를 제한하는 근거를 규정했다. 또한, 정보 주체에게 개인정보의 열람청구권, 정정·삭제

청구권, 처리정지 요구권 등을 부여하고, 그 권리행사 방법 등을 규정했다. 이 외에 개인정보의 효율적 보호를 위해 집단분쟁조정제도와 단체소송을 도입했다.

그간 국내 데이터 이용 관련 법제의 최대 이슈는 단연 「개인정보보호법」이었다. 미국 정보통신 컨설팅 전문업체 애널리시스 메이슨(Analysys Mason)의 2014년 보고서에 따르면 국내 데이터 규제 수준은 세계적으로 매우 엄격한 편에 속한다. 실제로 현행 「개인정보보호법」 하에서 개인정보를 수집·결합·제공하기 위해서는 모든 단계마다 별도의 동의가 필요하다. 2011년 법 제정 이후 추진돼온 주요 개정 역시 주민번호 수집 법정주의 도입('13), 개인정보 수집 출처 고지 의무화('16) 등 보호 강화를 위한 조치들이 대부분이었다. 세계적으로 고객 맞춤형 혁신 서비스가 확산되면서 국내에서도 개인정보보호와 활용의 조화 필요성에 대한 요구가 증가했다.

[표 7-2-4] 주요국 데이터 규제 수준

국가	데이터 수집	데이터 보관	융합 및 타목적 사용	프로파일링 차별화	데이터 교환	평균
한국	4	4	3	3	4	3.6
유럽연합	3	2	3	3	3	2.8
미국	2	2	3	3	2	2.4
OECD	1	2	2	3	2	2.0
뉴질랜드	3	2	3	1	2	2.2
싱가포르	2	2	2	2	2	2.0
호주	1	2	2	3	2	2.0

(4: 장애 요인, 3: 상당한 수준의 장애 요인, 2: 약간의 장애 요인, 1: 규제가 약한 수준)
출처: AnalysysMason (2014), *Data-driven innovation in Singapore*

유럽연합(EU)의 경우 EU 개인정보보호 지침을 통해, 일본 역시 최근 개인정보보호법을 개정해 익명화한 정보에 대해서는 정보주체의 동의 없이 활용할 수 있도록 하고 있다. 이런 세계적인 추세에 맞춰 우리 정부 역시 지난해 6월 관계부처 합동으로 개인정보 비식별 조치 가이드라인을 공표했다. 그러나 법적 효력이 없는 가이드라인 형식을 취하고 있어 산업계는 비식별조치 활용에 소극적인 모습을 보이고 있으며, 해당 내용의 법제화를 요구하는 목소리가 커지고 있다. 이와 함께 개인정보 정의의 명확화, 개인정보 유형에 따른 부분적 옵트 아웃 도입 등

「개인정보보호법」의 개정 논의도 활발하게 진행 중이다.

2015년 「정보보호산업의 진흥에 관한 법률」(이하 정보보호산업진흥법)이 제정됐다. 국가안보와 안전한 ICT 성장을 위해 정보보호산업의 수요·공급 측면을 종합 반영한 법체계가 마련된 것이다. 정보보호산업진흥법은 정보보호 관련 시장에서의 불합리한 발주 관행을 개선하기 위해 정보보호제품 및 서비스의 적정 대가의 지급 노력 및 불공정 발주관행 개선을 위한 발주 모니터링체계의 운영 등을 규정했다. 또한 기업의 자발적인 정보보호 투자를 유도하기 위해 기업의 정보보호 투자·인력, 정보보호 수준 등을 평가해 등급을 부여하는 '정보보호 준비도 평가'와 정보보호 투자 및 인력 현황, 정보보호 관련 인증 등 정보보호 현황을 공개하도록 하는 '정보보호 공시' 제도의 근거도 마련했다. 정보보호 핵심 원천기술 개발을 촉진하기 위해 신규성, 독창성, 사업화 가능성이 있는 기술을 우수 정보보호기술로 지정해 시제품 제작비, 수출을 위한 비용 등을 지원한다. 기업 성장률, 기술개발 실적, 정보보호 인력, 고용창출 기여도 등을 평가해 우수 정보보호 기업으로 지정해, 국제협력·성능평가 등도 지원할 수 있도록 했다.

라. 최근의 입법 노력

데이터의 이용이 사회 전반으로 확대되고 데이터 자산의 경제적 가치가 강조되면서 데이터 자체에 초점을 맞춘 법제 마련의 필요성이 커지고 있다. 특히 해외 사례에서 보듯이 데이터가 인프라 자원으로 인식되면서 법률을 통한 규율 대상이 공공 부문을 벗어나 민간데이터의 이용과 공유 등으로 확대되고 있다.

전문가들은 올해부터 지능정보 확산이 본격화되고 2020년 이후 실생활에 영향을 미칠 것으로 전망하고 있다. 시장 선점을 위해 조직을 정비하고 인공지능(AI) 기술을 적용한 초기 제품을 출시하는 등 기업들의 경쟁도 치열해지고 있다. 때맞춰 정부는 올초 인공지능과 빅데이터 등의 지능정보사회의 핵심 기술을 육성하고 이용을 활성화하기 위한 「지능정보사회 기본법」(가칭) 제정을 2017년 말까지 추진한다고 밝혔다. 기존 「국가정보화 기본법」을 전면 개정하는 형태의 본 법을 통해 지능정보기술·사회를 정의하고 인공지능의 안전성, 사고 시 법적 책임의 주체, 기술개발 윤리 등에 대해서도 관련 법과 제도를 정비하기로 했다. 특히 본 법은 데이터의 이용과 데이터 재산권의 보호에 관한 내용 등을 포함시켜 데이터의 사회적 활용과 지능정보자원화를 촉진시킬 예정이다.

[표 7-2-5] 지능정보사회 규제혁신 방안

구분	현황		개선방향
안정성	AI에 대한 불안감, 사용 위축	증권가 자동매매시스템 오류 로봇 오작동	• 지능정보기술 안전성 심사 방안 마련 (기존 안전성 인증체계 보완, 기본법 관련 규정 마련, '17.12월)
법적 책임	인공지능 사고책임 불분명	자율주행차 사고시 책임은 누구에게? 제조사? 운전자? 알고리즘 개발사?	• 인공지능 결함에 대한 손해배상 법제 분석 (민법 등 책임 범위·입증 책임전환 연구, 개발 부문별 손해배상 법제 정비, 지능정보기술 특화 보험 연구, '17.12월)
윤리 신뢰	인공지능의 비윤리적 활용 우려	사진 판별 문제, 챗봇의 성차별 발언	• 기능정보기술 윤리헌장 제정('18년) (데이터 수집 및 알고리즘 개발단계 기준·절차연구, '17.12월)
데이터, 지적재산권	데이터 가치 및 AI 산출물 권리 보호 불분명	데이터를 상속? 인공지능이 만든 영화의 저작권 문제	• 빅데이터의 재산권적 가치를 인정하고 활용할 수 있는 방안 연구('17.12월) • AI 산출물에 대한 법적보호 방안 연구('17.12월)

출처: 「인공지능, 가상현실, 핀테크 규제혁신 방안 발표」, 미래창조과학부, 2017.2.17.

제도적 기반이 없는 공유형 데이터 영역에 대한 법제 마련도 시도 중이다. 정부는 데이터 유통과 공유에 관한 규범화를 주요 내용으로 하는 법률 제정 연구를 진행하고 있다. 본 법은 데이터를 일종의 자산으로 상정하고 적정 가격과 시장을 통해 유통·거래되기 위한 제도적 기반들을 담을 예정이다. 데이터 가치평가체계의 수립, 데이터 거래시장 활성화 및 민간데이터 공유 촉진을 위한 구체적 방안 등이 대표적인 예이다. 본 법이 제정될 경우 다양한 분야에서 생산되고 있는 고가치 데이터들이 활발하게 유통·공유될 수 있도록 해 견고한 데이터 인프라를 구축하는 데 큰 역할을 할 것으로 기대된다.

2017 데이터산업 백서 집필진

제1부 데이터산업의 미래

제1장 데이터 기반 혁신 : **박주석** 경희대학교 교수

제2장 데이터 기반 비즈니스 모델의 진화 : **김인현** 투이컨설팅 대표

제3장 지능정보시대의 데이터 전문가 전망 : **이종석** 신한카드 빅데이터센터장

2016 데이터 구루상 수상자 인터뷰 : **이경일** 솔트룩스 대표

제2부 데이터산업 시장 현황

제1장 국내 데이터산업 현황 : **하진희** 한국데이터진흥원 선임연구원

제2장 국내 빅데이터 시장 현황 : **최승우** 한국정보화진흥원 선임연구원

제3장 해외 데이터산업 현황 : **나영민** 날리지리서치그룹 실장

2016 데이터 글로벌 이노베이터상 수상자 인터뷰 : **강용성** 와이즈넛 대표

제3부 데이터 서비스 동향

제1장 지능정보시대를 향한 데이터 서비스 : **이정승** 호서대학교 교수

제2장 선진 데이터 서비스 사례와 우리의 과제 : **양정식** 투이컨설팅 이사

제3장 이용자 니즈를 이해하는 지능형 데이터 서비스 : **안세준** 카페인모터큐브 대표

2016 데이터 서비스 이노베이터상 수상자 인터뷰 : **김덕조** 시냅스엠 대표

제4부 데이터 솔루션 동향

제1장 데이터 솔루션 아키텍처 : **김종현** 위세아이텍 대표

제2장 데이터 수집 솔루션 : **신동선** 데이터스트림즈 상무

제3장 DBMS 솔루션 : **이동석** 티맥스소프트 상무

제4장 데이터 레이크 솔루션 : **공상휘** 티맥스소프트 상무

제5장 데이터 활용 솔루션 : **박민규** 비아이매트릭스 수석연구원

제6장 데이터 분석 솔루션 : **김선영** 화수목 대표

제7장 마스터데이터 관리 솔루션 : **오세창** 투비웨이 대표

제8장 데이터 품질 관리 솔루션 : **최용준** 위세아이텍 상무

제9장 데이터 보안 솔루션 : **안혜연** 파수닷컴 부사장

제10장 머신러닝 솔루션 : **안동혁** 화수목 대표

2016 데이터 솔루션 이노베이터상 수상자 인터뷰 : **최용준** 위세아이텍 상무

제5부 데이터 구축 및 컨설팅 동향

제1장 데이터 분석 컨설팅 : **김찬수** 투이컨설팅 상무

제2장 데이터 품질 컨설팅 : **오경조** 지티원 상무

제3장 비즈니스 인텔리전스 구축 : **유수훈** 비투엔 이사

제4장 빅데이터 구축 : **김상수** 위세아이텍 센터장

제5장 데이터 이행 구축 : **이상일** LG히다찌 부장

제6장 인공지능 적용 : **김영래** 와이즈넛 팀장

2016 데이터 컨설팅 이노베이터상 수상자 인터뷰 : **이해곤** 비투엔 기술이사

제6부 데이터 기술 동향

제1장 인공지능과 고급 머신러닝 : **석호식** 강원대학교 교수

제2장 빅데이터와 클라우드 기술의 컨버전스 : **하석재** 한양대학교 교수

제3장 NewSQL : **권준호** 부산대학교 교수

제4장 아파치 스파크와 리얼타임 빅데이터 분석 : **이상훈** SK C&C 대리

제5장 데이터 시각화 : **송한나** 코그니텀랩 대표

제6장 블록체인 : **김철연** 숙명여자대학교 교수

제7부 데이터산업 정책 동향

제1장 주요 국가의 데이터산업 정책 : **김경훈** 정보통신정책연구원 부연구위원

제2장 국내 데이터 관련 법·제도 현황 : **김형건** 한국데이터진흥원 선임연구원

2017 데이터산업 백서 (통권 20호)

2017 Data Industry White Paper (Issue 20)

편찬위원 **김인현** 투이컨설팅 대표
김종현 위세아이텍 대표(한국데이터산업협의회 회장)
이경일 솔트룩스 대표
이우기 인하대학교 교수(한국정보과학회 DB소사이어티 회장)
황재훈 연세대학교 교수(한국데이터베이스학회 회장)

편집위원 **박재현** 한국데이터진흥원 정책기획실 실장
이종서 한국데이터진흥원 정책기획실 차장
하진희 한국데이터진흥원 정책기획실 선임연구원

감수 **김인현** 투이컨설팅 대표

발행처 한국데이터진흥원
발행인 이영덕
발행일 2017년 7월 12일
편집·디자인 글봄크리에이티브(02-507-2340)
인터뷰 박세영, 김혜영
인쇄 서울문화인쇄
ISSN 2465-7662

• 본 백서는 미래창조과학부의 출연금으로 수행한 데이터산업 육성 지원 사업의 결과물입니다.

2017
데이터산업 백서

2017 Data Industry
White Paper

Kdata 한국데이터진흥원

서울특별시 중구 세종대로9길 42 부영빌딩 7층 한국데이터진흥원
Tel. 02-3708-5300 **Fax.** 02-318-5040 www.kdata.or.kr

9 772465 766005 75
ISSN 2465-7662