2018 데이터산업 백서

2018 Data Industry White Paper

Kdata 한국데이터진흥원

2018 데이터산업 백서

2018 Data Industry White Paper

발간사

데이터 기반의 혁신 성장을 통한 데이터 경제 활성화

제4차산업혁명시대를 맞이하여 데이터는 기업과 국가의 핵심 성장 동력이자 경쟁력의 원천이 됐습니다. 다양하고 질 좋은 데이터를 보유하고 활용하는 것이 산업 전반의 패러다임 변화를 촉진시키고 있기 때문입니다. 데이터가 갖고 있는 무한한 경제적 가치는 가장 중요한 자산이 되고 있으며, 기업들의 디지털 혁신을 성공시키는 중요한 수단이 되고 있습니다. 이제 우리나라도 수많은 사람과 다양한 기기가 실시간으로 생산하는 막대한 양의 데이터를 기반으로 새로운 가치를 창출하는 데이터 경제를 성공적으로 만들어낼 수 있는 큰 그림(Big Picture)을 그려야 합니다.

데이터를 가장 안전하게 잘 쓰는 데이터 강국이 되기 위해 정부, 산업, 학계, 시민단체 등 모든 분야에서 전략적 고민과 접근이 필요한 때입니다. 기존의 데이터 전문기업뿐만 아니라 의료, 금융, 통신, 유통, 제조, 에너지 등 다양한 산업 군에서 데이터 기반의 혁신을 통해 데이터산업의 새로운 생태계가 만들어져야 합니다. 전 국민이 데이터 자주권을 확보해, 고품질 데이터의 원활한 생산·유통과 분석·활용 기반이 만들어져야 합니다. 한국데이터진흥원에서도 데이터 유통·거래 플랫폼 구축·운영 사업을 비롯해 데이터 사업화 및 해외진출 지원, 본인정보 활용지원(MyData) 시범사업 등을 통해 우리나라의 데이터 경제 활성화의 초석을 다지고 있습니다. 앞으로도 더 실질적인 사업 발굴과 성공적인 추진을 통해 데이터 기업뿐만 아니라 전 국민, 모든 기업이 데이터를 잘 관리하고 활용해 새로운 가치를 창출할 수 있도록 물심양면 지원하겠습니다.

이번에 발간되는 「2018 데이터산업 백서」는 이러한 노력과 더불어 데이터 활용과 관련한 산업 분야별 동향을 비롯해 최신 기술 트렌드, 주요 정책까지 수록해 데이터산업 전반을 조망할 수 있도록 구성했습니다. 특히 개인데이터의 새로운 가치창출을 위한 MyData를 별도 기획해 다양한 전문가들의 견해와 동향도 함께 다뤘습니다. 아울러 업계·학계에 계신 데이터 전문가들의 의견을 수렴해 '데이터산업 이슈 TOP 10'을 선정해 국내 데이터산업의 미래를 예측하고, 향후 발전 방향을 모색해 볼 수 있도록 했습니다.

이번 백서가 데이터 경제를 둘러싼 다양한 이해관계자 분들에게 매우 유용한 자료가 되기를 바랍니다. 마지막으로 「2018 데이터산업 백서」가 발간되기까지 많은 도움을 주신 편찬위원, 집필진, 편집진 여러분께 감사의 말씀을 드립니다. 감사합니다.

2018년 9월
한국데이터진흥원장 **민기영**

추천사

데이터산업은 제4차산업혁명 시대의 기간산업

4차산업혁명시대에서 데이터는 세계 경제의 무한 신자본이라고 합니다. 그동안 컴퓨터로 '세상을 자동화하는 시대'였다면, 몇 년 전부터는 데이터 자산을 활용해 '세상을 어떻게 잘 이해하느냐'로 관심이 집중되고 있습니다. 그 변화의 저변에 바로 데이터가 있습니다.

데이터산업은 4차산업혁명시대의 기간산업입니다. 국내 데이터산업은 2010년 이래로 연평균 8%씩 성장했습니다. 2018년에는 그 규모가 15조 원을 넘어설 것으로 예상됩니다. 데이터산업은 중흥기에 접어들었습니다.

그동안 국내 데이터산업의 촉진을 가로막는 법과 규제들에 대한 개선 논의와 노력도 꾸준히 이뤄지고 있어서 데이터산업이 더 활성화될 것으로 기대됩니다. 가칭 '데이터 거래소' 추진도 머지않아 모습을 드러낼 것으로 예상됩니다. 선순환적 데이터산업 생태계를 만들어 내기 위해 '마이데이터'와 블록체인 기술까지 접목한다면, 개인정보 보호의 장벽을 뛰어넘는 해법도 나올 것으로 예상됩니다.

최근 빅데이터전문센터 협력 네트워크를 구축해 4차산업의 데이터 자본재 확충과 산업 생태계 강화를 도모하고 있습니다. 이는 데이터 기반 사회·경제 전반에서 미래 변화에 맞는 신산업을 창출하게 될 것입니다. 또한 전문 센터들의 협업 네트워크를 통한 가시적인 성과들도 모범 사례들로 나타날 것입니다.

1996년부터 매년 발간돼 온 『데이터산업 백서』는 대한민국의 데이터산업이 걸어온 발자취와 청사진을 제시해 왔습니다. 국가 데이터산업 발전을 위한 정책 입안뿐 아니라 산업계·학계에서 다양한 용도로 활용되고 있습니다. 『데이터산업 백서』는 데이터산업의 현황과 기술, 발전 방향 등을 요약해 체계적으로 제공하고 있어서 데이터산업계뿐 아니라 국가의 전략적 미래 산업을 위한 유용한 자료로 활용될 것입니다.

『2018 데이터산업 백서』 발간을 진심으로 축하 드리며, 데이터산업에 종사하는 산업계와 학자, 연구자, 정책 입안 공직자 분들께 본 백서의 활용과 인용을 적극 추천 드립니다. 본 백서 출간을 위해 수고하신 한국데이터진흥원 임직원분들의 노고에 감사 드리며, 집필에 도움을 주신 우리 한국데이터산업협의회 회원사 전문가, 학계 집필진 모두에게 감사의 마음을 전합니다. 건승하심과 무궁한 발전을 기원합니다.

한국데이터산업협의회 회장 **조광원**

추천사

데이터 기반의 혁신과 글로벌 경쟁력

데이터는 기업정보 시스템의 핵심요소다. 축적된 데이터는 기업의 운영과 의사결정, 전략 개발을 위해 활용돼 왔다. 그러나 급속하게 진화하는 정보기술의 발전은 데이터를 저장할 수 있는 저장장치의 용량을 거의 무한대로 늘리고 있다. 이에 따라 선도 기업들은 기존의 전통적 정보 시스템의 데이터를 넘어, 다양한 내부/외부 소스로부터 데이터를 수집·활용하며 글로벌 시장에서 경쟁 우위를 갖고자 노력한다.

제4차산업혁명시대에 데이터는 '원유'나 다름이 없고, 기업들은 데이터에 기반해 혁신을 이루며 발전해 나가야 한다. 기업이나 조직에서 전략을 만들고 의사 결정을 하는 사람들은 경영자들이다. 데이터에 대한 이해가 부족한 경영자들은 데이터 중심의 경제에서 성공하기가 힘들다. 하버드 비즈니스 리뷰에서 2014년에 내놓은 『빅데이터앳워크 BigData@Work』(Thomas H. Davenport)의 설문조사를 따르면, 응답자의 3.5%만이 '조직의 비즈니스에 빅데이터를 어떻게 적용할지 알고 있다'는 것에 대해 강하게 동의(strongly agree)했다고 한다. '강하게 동의'로 답한 응답자의 비율을 높이려면, '데이터와 빅데이터에 대한 이해' 그리고 '전략적 활용의 의미' 등을 모두 알아야 한다.

경영자의 데이터에 대한 전략적 사고는 매우 중요하지만, 데이터와 관련된 제반 기술에 대한 이해가 부족한 상태에서 '전략적 사고'를 논하는 것은 한계가 있다. 본 백서는 기업정보시스템의 핵심 요소인 데이터베이스(DB) 기술에서부터 DB 기술이 진화해 빅데이터 인프라 구조가 되기까지 데이터 기술의 진화 과정에 대한 설명과 함께, 빅데이터 기반의 혁신을 주도하고 있는 최신 비즈니스 모델까지 망라하고 있다. 본서를 통해 경영자들은 데이터 관련 주제들이 어떻게 유기적으로 결합되고 시너지를 이뤄 조직의 문제를 창의적으로 해결하고 조직을 변화시키는가에 대해 이해할 수 있을 것이다.

현재 기업들이 당면한 문제는 빅데이터의 방대함 자체가 아니라 빅데이터의 가능성을 인지하고 분석해 데이터로부터 인사이트를 얻고 혁신을 이루며, 비즈니스 가치를 창출하고자 하는 노력이 부족하다는 것이다. 한 연구를 따르면 2.8제타바이트의 데이터 중 단지 5%의 데이터 만이 분석되고 있다고 한다.

새로운 데이터 기술은 기존의 시장 질서를 파괴하고 새로운 시장을 창출하고 있다. 기존의 기업들 중에는 데이터 기술의 전략적 활용을 통해 새로운 가치를 창출하는 기업도 있지만, 신기술로 무장하고 새로이 시장에 진입하는 신생 기업들과의 경쟁에서 패배하고 도태하는 기업들도 있다. 미래의 경영자들은 데이터 기술에 대한 이해뿐 아니라, 데이터가 어떻게 활용돼 혁신을 이뤄낼 것인가에 대한 전략적 사고를 할 수 있어야 글로벌 시장에서 경쟁할 수 있다. 본서는 미래의 경영자인 후학들에게도 매우 유용한 지침서가 될 것이다.

(사)한국데이터전략학회 회장 **김경민**

추천사

제4차산업혁명 항해 돕는 길잡이

우리는 빅데이터 시대를 살고 있습니다. 이 빅데이터 시대는 생성되는 데이터의 특성이 양(volume), 속도(velocity), 다양성(variety) 측면에서 이전 시대와 크게 다른 양상을 보입니다. 지금 이 순간에도 다양한 형태의 정보들이 매우 빠른 속도로 그 양을 키워 나가고 있을 것입니다.

빅데이터 관점에서 이렇게 축적되는 방대한 양의 데이터는 실제 비즈니스에서 활용 가능한 보물(유용한 지식)들을 숨기고 있습니다. 불과 얼마 전까지만 해도 사람들은 기술적 한계로 인해 빅데이터로부터 유용한 지식을 얻어내는 것이 실질적으로 어렵다고 생각했습니다. 그러나 데이터베이스 시스템, 데이터 마이닝, 기계 학습 등 빅데이터와 관련된 기술들이 지속적으로 발전하고 있어서 기술적 한계, 즉 장벽들은 크게 낮아졌습니다.

대부분의 기술은 오픈소스로 공개되고, 데이터를 저장할 저장 매체와 데이터를 처리할 컴퓨팅 자원의 비용은 매우 낮아졌습니다. 이 결과 최근에는 기업, 기관, 심지어는 개인까지도 빅데이터로부터 추출된 유용한 지식을 활용해 의사결정을 하거나 생활의 질을 개선하고자 합니다.

본 백서는 이러한 빅데이터 시대에 데이터산업에 대한 이정표를 제시하고 있습니다. 국내외 데이터산업의 최신 동향과 최신 빅데이터 기술을 소개하고, 아울러 국내외의 관련 정책도 다각도로 고찰하고 있습니다. 데이터와 관련된 학계, 정부, 산업계 종사자는 물론 데이터 산업에 관심이 큰 일반인에게도 유익한 정보를 제공할 것으로 기대합니다.

한국은 지금 기술 개발에 투자하고, 데이터 산업 관련 규제를 개선하거나 기술 발달에 따른 부작용을 방지하기 위한 정책을 입안하는 등 코앞까지 다가온 4차산업혁명의 대비에 박차를 가하고 있습니다. 이렇게 중요한 시기에 금번 『2018 데이터산업 백서』 발간이 갖는 의미는 더욱 특별하게 느껴집니다. 본 백서가 빅데이터 기술을 통해 4차산업혁명의 실현을 꿈꾸는 각계 인사들의 항해를 돕는 등댓불이 되어 대한민국의 4차산업혁명에 크게 이바지할 수 있기를 기대합니다.

(사)한국정보과학회 데이터베이스 소사이어티 회장 **김상욱**

그림으로 보는 데이터산업 동향

데이터산업 시장 규모

직접매출 기준 6조 2,973억 원

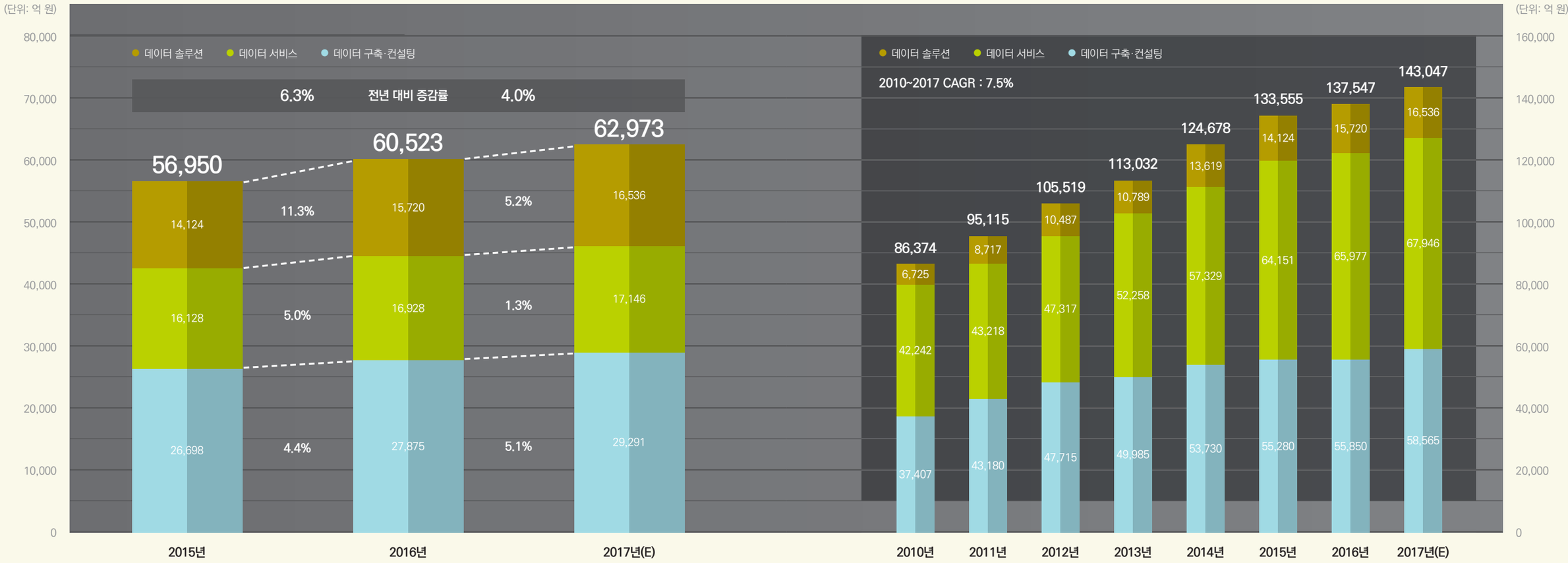
2017년 국내 데이터산업 전체 시장 규모는 2016년 13조 7,547억 원에서 4.0% 성장한 14조 3,047억 원으로 형성돼 2010년 이후 연평균 7.5%의 성장세를 유지하고 있는 것으로 조사됐다. 광고 매출 등 데이터와 관련된 간접매출을 제외한 직접매출 규모는 6조 2,973억 원 규모로 형성됐다.

데이터산업 직접매출 시장 규모

2017년 국내 데이터산업 시장 규모는 직접매출 기준으로 전년 대비 4.0% 성장한 6조 2,973억 원으로 조사됐다.
직접매출 시장 규모는 데이터와 직접 연관되는 매출을 기준으로 조사한 협의의 데이터산업 시장을 의미한다.
즉 광고 매출, 기타 운영시스템 SI 매출 등 DB·데이터 구축 관련 간접매출을 제외한 결과다.

데이터산업 시장 규모

직간접 매출을 모두 포함하는 2017년 전체 데이터산업 시장 규모는 14조 3,047억 원에 이를 것으로 조사됐다. 2010년 이후 연평균성장률은 7.5%로 매년 성장세를 유지하고 있다.



그림으로 보는 데이터산업 동향

데이터직무 인력은 10만 9,320명
직무별 평균 10.9% 부족

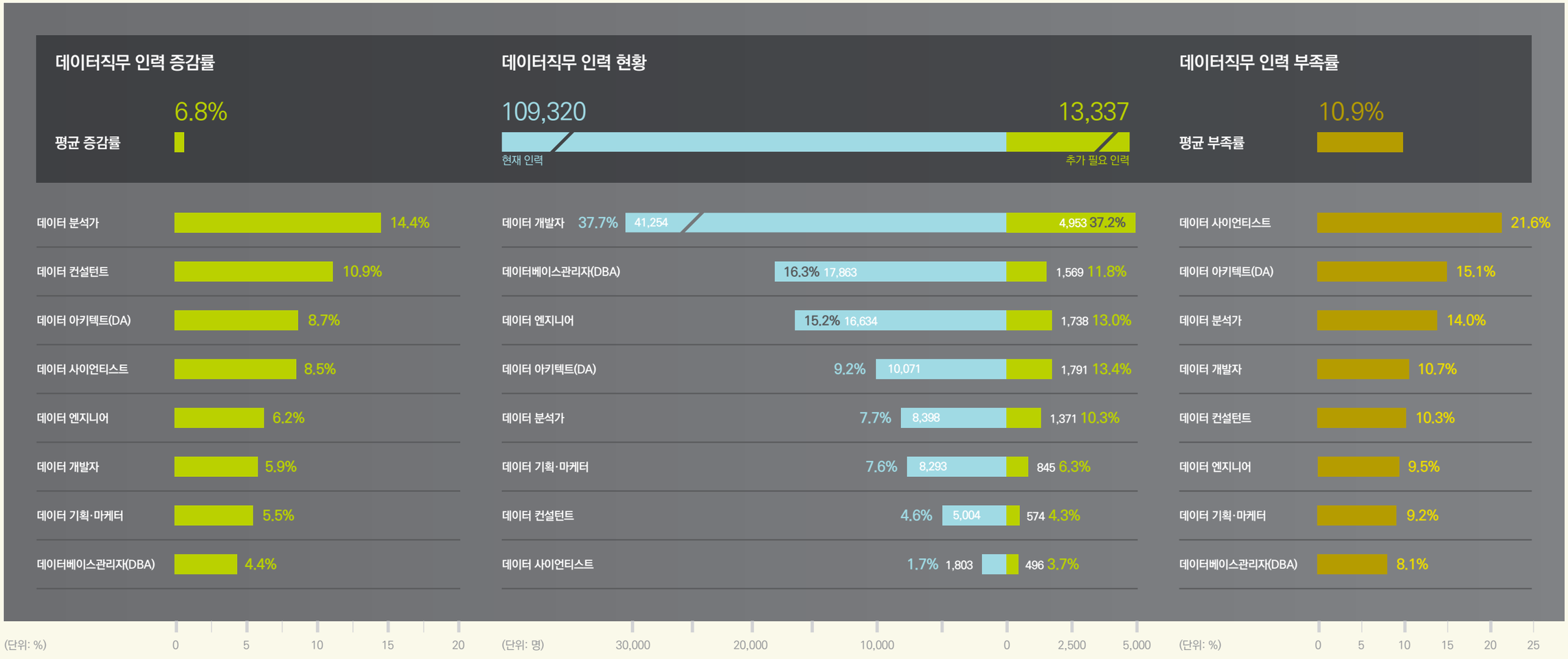
2017년 데이터 직무별 인력 가운데 데이터 개발자가 37.7%로 가장 높은 비중을 차지했다. 직무별 인력 변동에서는 데이터 분석가가 14.4%로 가장 높게 증가해 데이터 분석 인력 수요가 지속되고 있음을 보여줬다.

데이터직무 인력 현황

2017년 전 산업의 데이터직무 인력은 총 10만 9,320명으로 조사됐다. 이 중 데이터 개발자 비중이 37.7%로 가장 높고, 이어서 데이터베이스관리자 16.3%, 데이터 엔지니어 15.2% 순으로 나타났다.

데이터직무 인력 부족률

전 산업에서 추가로 필요한 데이터직무 인력은 총 1만 3,337명으로 데이터직무별 인력 부족률은 평균 10.9%로 조사됐다. 전체적으로 인력이 부족한 가운데 데이터 사이언티스트가 21.6%로 가장 인력난이 심하고 이어서 데이터 아키텍트 15.1%, 데이터 분석가 14.0% 순으로 나타났다.

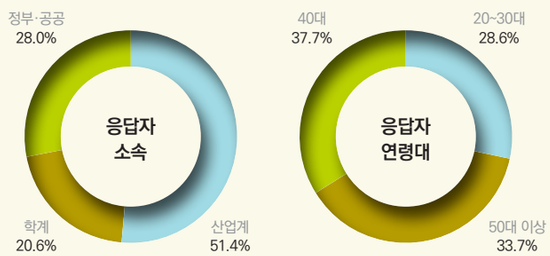


2019 데이터산업 이슈 TOP 10

최고 이슈는
데이터 거버넌스 도입 본격화

2019년 데이터산업계의 최고 이슈는 '데이터 활용을 위한 데이터 거버넌스 도입 본격화'가 차지했다. 제4차산업혁명 이슈와 함께 기업·기관의 무형 자산으로서 데이터의 중요성이 강조되면서 이를 체계적으로 관리하고 활용하기 위한 데이터 거버넌스에 대한 공감대가 빠르게 형성되고 있는 것으로 조사됐다. 블록체인은 데이터 보호와 거래를 위한 핵심 기술로 등극을 예고하고 있다. 2018년 5위를 차지했던 '퍼스널데이터 시대 도래'는 2단계 뛰어올라 개인데이터 또는 마이데이터라는 더 구체화한 모습으로 빠르게 자리를 잡아가는 것으로 조사됐다.

설문 응답자 비율 (n=175, 응답자 정보 무응답=1)



이슈 TOP10 설문항목과 순위

1. 데이터 활용을 위한 데이터 거버넌스 도입 본격화
2. 블록체인 기반의 데이터 비즈니스 모델 확산
3. 마이데이터MyData 제도 도입 착수
4. 개인데이터 보호와 활용 논쟁 지속
5. 데이터거래소 수요 증가
6. 공공부문 데이터 전면 개방과 품질 고도화
7. 오픈API 플랫폼 기반 데이터 공유 확대
8. 사물인터넷 생성 데이터 급증과 활용 확산
9. 고급 분석가 양성 프로그램 확대 필요
10. 데이터 가공 전문기업 육성 필요 증대
11. 데이터 축적과 학습 기반의 지능형 챗봇 활용 확산
12. 데이터 인텔리전스를 통한 비즈니스 혁신
13. 데이터 사이언티스트 부족 심화
14. 데이터 자산 가치 평가에 대한 공감대 확산
15. 클라우드 기반의 데이터 관리·활용 확산
16. 데이터 기반의 스마트시티 구현
17. 디지털 트랜스포메이션, 데이터기업 전환 확대
18. 공공부문 데이터 품질 관리수준 평가 반영과 인증 확산
19. CKAN 등 데이터 관리 및 공유 방식 표준화 확산
20. 데이터 보안 신기술 적용 확산
21. 안전한 데이터 활용을 위한 데이터 안심존 제공
22. 기타 (AI 기반 빅데이터 솔루션 확산, Self Service Analytics)

설문조사 방법 | 데이터 업계와 학계 전문가들의 자문을 거쳐 데이터산업과 관련한 총 22개 설문항목을 도출했다. 한국데이터산업협의회, 데이터전략학회, DB소사이어티, 디비가이드넷 회원 등 데이터 관련 종사자들을 조사 대상으로 2019년 데이터 관련 이슈 5개를 선택하도록 했다. 3주간(18.7.19~18.8.9) 온라인으로 실시해 176명으로부터 응답을 받아 분석한 결과다.

이슈 TOP10으로 보는 2019 데이터산업 트렌드

(n=880, 개)

1 데이터 활용을 위한 데이터 거버넌스 도입 본격화
데이터 거버넌스의 중요성이 강조되면서 기업 내 데이터 전문조직이 확산되고 데이터 거버넌스 전략 수립이 본격화 될 것이다.
113

2 블록체인 기반의 데이터 비즈니스 모델 확산
블록체인 기반의 완벽한 온라인 프라이버시 및 블록체인 사물인터넷(BIoT) 확산으로 데이터 수익화(data monetization)가 증가할 것이다.
95

3 마이데이터 제도 도입 착수
정보주체인 개인이 개인데이터를 스스로 통제 및 활용·분석할 수 있는 사회적 공감대가 확산되면서 관련 정책 및 사업이 추진될 것이다.
94

4 개인데이터 보호와 활용 논쟁 지속
유럽 GDPR 도입, 개인데이터 분석을 통한 가치 창출 관심도 증대 등 개인데이터 보호 이슈와 활용 논쟁은 당분간 지속될 것이다.
69

5 데이터거래소 수요 증가
데이터를 자유롭게 사고 파는 데이터 거래 시장의 활성화를 위해 데이터의 품질관리·가공·저장·분석을 종합적으로 대응하는 데이터거래소 설립이 필요하다.
46

6 공공부문 데이터 전면 개방과 품질 고도화
공공데이터, 연구데이터 등 공공부문의 데이터를 전면 개방하고 활용 수준으로 데이터 품질을 고도화할 것이다.
45

7 오픈API 플랫폼 기반 데이터 공유 확대
데이터 공유 및 활용 확대는 새로운 산업 창출과 기존 산업의 혁신을 가져올 것이며 오픈API 플랫폼은 이를 가속화할 것이다.
43

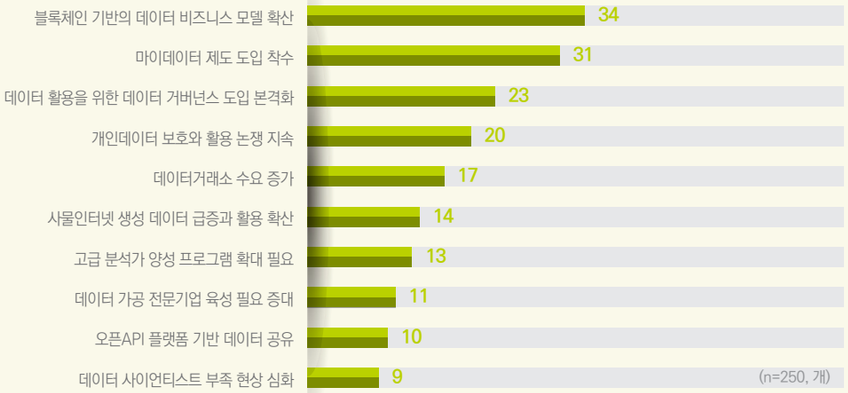
8 사물인터넷 생성 데이터 급증과 활용 확산
사물인터넷 확산에 따라 기업들은 센서·기기·머신러닝·AR/VR생성 데이터 등 방대한 데이터를 수집해 새로운 디지털 혁신기업을 탄생시키고 있다.
41

9 고급 분석가 양성 프로그램 확대 필요
데이터 분석 고급인재 양성을 위한 데이터 전문 대학원 및 연구센터 운영, 국가공인기술자격제도 활성화가 필요하다.
37

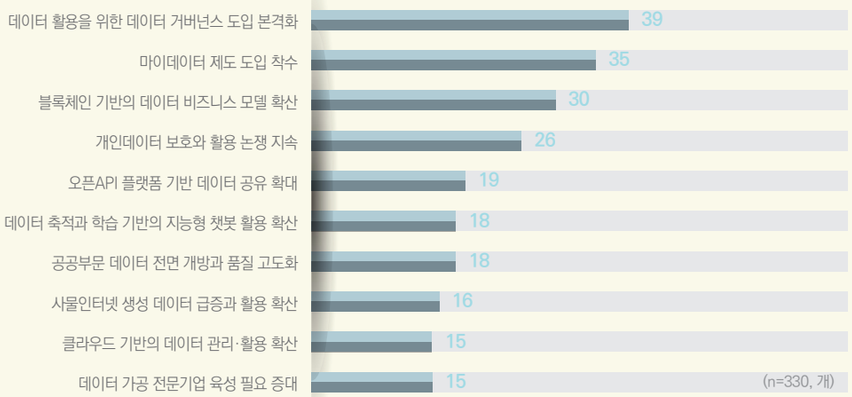
10 데이터 가공 전문기업 육성 필요 증대
데이터 품질을 확보하고 거래·분석 활성화를 위해 데이터 전처리, 데이터 프렙(Data Prep) 서비스를 전문으로 하는 기업 육성이 필요하다.
35

연령별 이슈 TOP10

20~30대



40대



50대~

직종별 이슈 TOP10

산업계

학계

정부·공공

Contents

Contents

제7부 데이터산업 기술 동향

표 목차

Contents

그림 목차

제 1 부

새로운 디지털 자원, 마이데이터

제1장

데이터경제, 데이터민주주의, 그리고 마이데이터

최근에 4차산업혁명 기술이 확산되면서, '데이터경제'와 '데이터민주주의' 개념이 대두되고 있다. 데이터경제와 데이터민주주의가 구현되려면, 새로운 관점의 신뢰성 있는 데이터 생태계 창출이 매우 중요하다. 따라서 빅데이터 정책이나 오픈데이터 정책 만으로 충분하지 않으며, 마이데이터 정책을 강력하게 추진해야 한다.

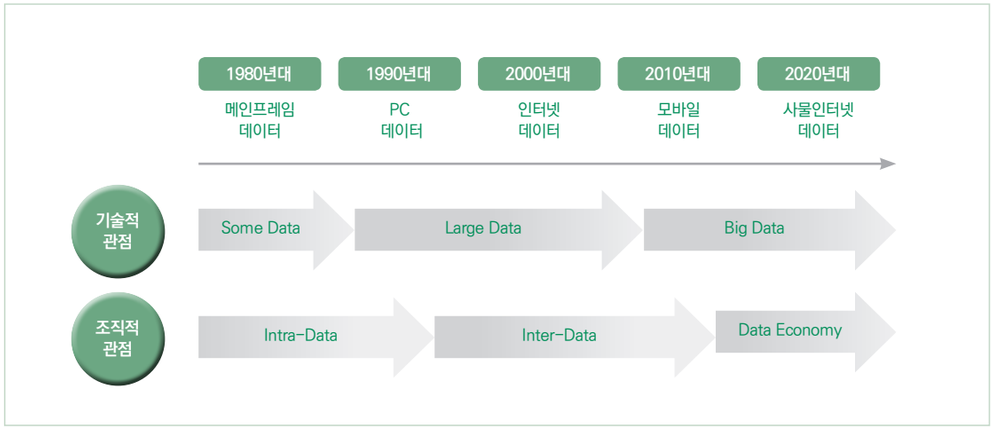
4차산업혁명은 디지털 혁명이라고 얘기할 수 있다. 과거 아날로그 혁명에서 핵심 자산이 자본·설비·인력·원료 등이었다면, 디지털 혁명의 핵심 자산은 데이터다. 데이터가 '자산'으로 부각되는 이유는 모든 것이 연결돼 언제나 데이터를 주고 받을 수 있으며 우리의 모든 활동이 데이터로 기록될 수 있기 때문이다. 이러한 데이터를 매개로 한 생태계가 급속하게 성장하면서 '데이터산업'이 급부상하고 있다. 향후 초연결사회와 초지능사회로 발전되면서 데이터는 더욱 중요한 자산이 될 것이다.

데이터가 새로운 형태의 자산으로 주목 받으면서 데이터 유통에 기반한 새로운 생태계인 '데이터경제(Data Economy)' 개념이 화두가 되고 있다. 데이터경제란 '데이터에 접근하고 활용할 수 있도록 협업하는 과정에서 데이터 생산, 인프라 제공, 연구조사, 데이터 소비 등 서로 다른 역할을 담당하는 구성원으로 이뤄진 생태계(ecosystem)'를 의미한다(European Commission, 2017).

데이터 경제 규모는 유럽연합(EU)의 경우 2014년 기준으로 2,570억 유로(EU GDP의 1.89%)였고 2020년에는 6,430억 유로(EU GDP의 3.17%)로 증가할 것으로 전망하고 있다. 유럽연합에서는 데이터를 경제 성장뿐 아니라 일자리 창출과 사회발전을 위한 필수 자원으로 인식하고 데이터경제 활성화를 위한 다양한 노력을 기울이고 있다.[1]

• 필자 | **박주석**(경희대 경영학과 교수)

[그림 1-1-1] 데이터의 변천사와 향후 방향



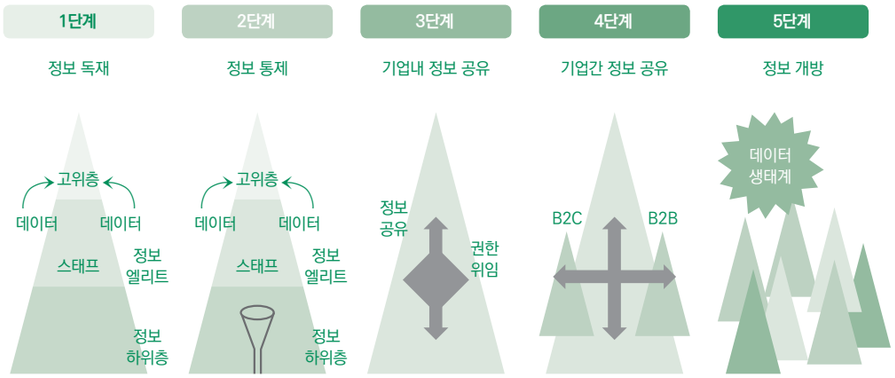
최근에 데이터경제가 화두가 되면서 '데이터민주주의'라는 개념도 많이 얘기되고 있다. 역사적으로 크게 보면, 정보독재 시대에서 정보통제 시대를 거쳐서 정보개방 시대가 도래했다. 작게 보면 행정정보공개 정책부터 데이터민주주의가 시작됐고 오픈거번먼트(Open Government) 정책부터 본격화 됐다. 또 다른 측면에서 보면 인터넷이 보편화되면서 데이터민주주의가 시작됐고 개인정보보호 정책에 의해서 데이터민주주의가 성숙됐다. 따라서 데이터민주주의는 그동안 데이터 개방과 데이터 보호라는 2가지 정책에 의해 발전됐다. 이런 관점에서 정리하면 데이터민주주의는 '특정한 개인에 대한 정보가 아니면 누구나 그 데이터를 사용, 재사용 및 재배포할 수 있어야 한다'는 의미로 정의할 수 있다.

우리나라 데이터민주주의 수준을 살펴보면, 지난 10년 동안에 괄목할 성장을 했다. 우리나라는 2011년 농협사태 등을 겪으면서 전 세계에서 가장 강력한 개인정보보호 제도를 갖고 있다. 공공데이터 개방도 선진국에 비하면 늦게 시작했지만 적극적으로 추진해 2015년에는 OECD 공공데이터 개방지수 1위를 차지했다. 이러한 성과에도 향후 4차산업혁명이 발전하면서 데이터 보호 정책과 데이터 개방 정책이 충돌되는 문제점이 나타날 것이다.

앞에서 데이터경제와 데이터민주주의를 언급했는데 결국 새로운 관점의 신뢰성 있는 데이터 생태계를 창출하는 것이 매우 중요하다. 특히 개방된 공공데이

1 European Commision, Building a European Data Economy(COM(2017) 9 final), 2017.10.

[그림 1-1-2] 데이터민주주의의 발전



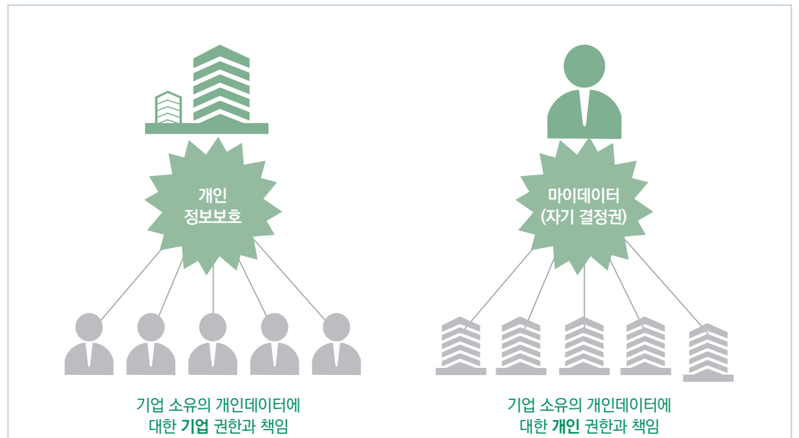
출처: 버나드 리어터드 외, 「e비즈니스 인텔리전스」, 2001, 박주석 개정, 2017

터와 민간데이터의 융합이 중요하다. 하지만 공공데이터 개방만으로 가치 있는 데이터 생태계를 완성시키기에는 한계가 있다. 조직데이터와 개인데이터가 융합됨으로써 훨씬 더 가치 있는 데이터 생태계를 만들어 낼 수 있다. 이런 점에서 개인정보보호법은 융합 데이터 생태계를 만드는 데 걸림돌이 될 수 있다. 물론 하나의 대안으로 개인정보의 비식별화 정책이 추진되고 있으나 섬세한 데이터 생태계를 만들어낼 수 없기 때문에 근본적인 대책이 될 수 없다.

4차산업혁명을 위해 데이터 생태계를 자연스럽게 융합해야 한다. 데이터 생태계를 융합하기 위해서는 산업별 마스터데이터가 중요한 연결고리가 되며 특히 개인정보는 가장 중요한 연결고리다. 예를 들어 개인별로 제조, 금융, 유통, 의료 데이터를 융합해 새로운 가치를 창출하는 것이 가장 의미있는 데이터 생태계다. 이런 점에서 마이데이터(MyData) 개념은 융합데이터 생태계를 구현하는 데 핵심적인 역할을 할 것이다.

2011년 IDC 보고서를 따르면 디지털데이터 중 개인데이터 비중은 약 75%이며, 사물인터넷과 인공지능 기술로 인해 개인데이터의 비중과 가치가 더 증가할 것으로 예상된다. 실제로 구글, 페이스북, 아마존 등 인터넷 기업은 개인데이터를 기반으로 성장해 지금은 글로벌 시가총액 TOP 5 기업이 됐다. EU에서 2020년까지 개인데이터로 창출되는 경제적 가치는 약 1조 유로, 정부와 기업이 얻는 연간

[그림 1-1-3] 개인정보보호 vs. 마이데이터



이익은 3,300억 유로로 예상하고 있다.

많은 사람들이 개인정보보호와 마이데이터를 상반된 개념으로 오해하고 있다. 개인정보보호는 기업 관점에서 기업소유의 개인데이터에 대한 기업 권한과 책임을 얘기한다. 반면에 마이데이터는 개인 관점에서 기업 소유의 개인데이터에 대한 개인 권한과 책임을 얘기한다. 즉 마이데이터 개념은 기업 중심의 개인데이터 생태계에서 개인데이터의 주인인 개인에게 자신의 데이터를 통제 및 관리할 수 있는 권한을 돌려주는 개념이다. 따라서 개인정보보호 정책은 유지되면서, 마이데이터 정책을 적극 추진할 수 있다.

마이데이터 정책의 핵심은 개인정보의 소유권(ownership)을 당사자에게 돌려주는 것이다. 실제로 많은 개인정보는 개인이 아니라 기업이나 기관이 갖고 있다. 개인의 금융데이터는 그 개인이 이용하는 금융기관이 대부분 갖고 있으며, 개인의 의료데이터는 진료를 받았던 병원이 대부분 갖고 있고, 개인의 공공데이터는 서비스 받았던 공공기관이 대부분 갖고 있다. 하지만 이러한 개인정보의 소유권자는 그 데이터를 관리하는 기관이나 기업이 아니라 그 개인이다. 따라서 그 당사자만이 금융데이터, 의료데이터, 통신데이터, 구매데이터, 공공데이터 등을 활용하고 융합할 수 있다. 결국 마이데이터는 개인의 융합 데이터 생태계를 만들 수 있는 기회를 제공하고 있다. 특히 마이데이터의 개인데이터 자기 이동권은 데이터 생태계 융합을 촉진하면서 새로운 비즈니스모델을 창출하고 지능화한 사회를

촉진시킬 것이다.

진정한 데이터민주주의가 되기 위해서는 조직 관점뿐만 아니라 개인 관점에서 데이터 권한과 책임도 정의해야 한다. 이런 관점에서 올해 5월에 시행된 EU의 GDPR(General Data Protection Regulation) 법안을 주목해야 한다. GDPR 법안은 기업의 자의적인 개인정보 활용을 강력하게 통제하면서 개인정보 권한을 보호하려고 있다. GDPR 법안에서는 개인의 데이터 열람권, 데이터 정정요구권, 데이터 삭제권, 데이터 이동권 등을 포함하고 있다.

결론적으로 앞에 언급됐던 데이터민주주의 정의도 확장돼야 한다. 확장된 관점의 데이터민주주의는 '특정한 개인에 대한 정보가 아니면, 누구나 그 데이터를 사용, 재사용 및 재배포할 수 있어야 한다. 또한 특정한 개인에 대한 정보는 누가 그 정보를 갖고 있든 그 개인에 의해 통제되고 관리될 수 있어야 한다'고 정의될 수 있다. 따라서 데이터개방 정책, 개인정보보호 정책, 마이데이터 정책이 균형되게 추진돼야 한다. 특히 데이터개방 정책과 마이데이터 정책은 새로운 데이터 생태계를 창출해 우리나라 4차산업혁명의 경쟁력을 획기적으로 높여줄 것이다.

[참 고 문 헌]

- 박주석, 「데이터민주주의와 데이터 생태계 융합」, 칼럼 기고, 전자신문, 2018.02.
- 박주석, 「마이데이터의 가치와 도입 필요성」, 5월 세미나, 개인정보보호위원회, 2018.05.
- 박주석, 『비즈니스 분석』, 한경사, 2018.06.
- 정용찬, 「4차산업혁명시대의 데이터 경제 활성화 전략」, 정보통신정책연구원(KISDI), 2017.06.
- 정준화, 「4차산업혁명시대의 빅데이터 정책 과제」, 이슈와 논점, 국회입법조사처, 2018.07.
- 투이컨설팅, 「국내 개인데이터 이용 환경 분석」, 연구보고서, 한국데이터진흥원, 2017.11.
- 투이컨설팅, 「MyData 시뮬레이션 및 시스템 설계 연구」, 연구보고서 한국데이터진흥원, 2017.11.
- European Commission, Building a European Data Economy, Data Policy and Innovation, EU, 2017.01.
- Imanol Arrieta Ibarra 외, Should we treat data as Labor? American Economic Association Papers & Proceedings, Vol. 1, No. 1, 2018.03.

제2장

새로운 디지털 자원, 마이데이터

데이터 생산의 주체로서 '개인'의 권리는 그동안 크게 주목받지 못했다. 하지만 유럽을 중심으로 개인의 데이터를 어떻게 활용할지 결정하는 '마이데이터(MyData)' 논의가 본격화되고 있다. 마이데이터는 데이터에 대한 업계의 요구와 디지털 인권을 결합한 개인데이터 관리의 인간 중심적 접근 방식으로, 개인이 자신의 데이터를 제어할 수 있게 하는 것이다. 기업과 공공기관은 마이데이터를 중심으로 한 새로운 비즈니스와 정책을 만드는 데 고심하고 있다. 마찬가지로 국내에서도 '개인신용정보 이동권' 등 적극적인 마이데이터 활용을 위한 정책이 속속 등장하고 있다.

1. 기업·기관 주도의 데이터 이용에 '반기'

세계 최대 비상장 정보기술(IT) 기업인 델테크놀로지스(Dell Technologies)의 마이클 델(Michael Dell) 회장은 2018년 4월 30일 미국 라스베이거스 샌즈 엑스포(Las Vegas Sands Expo)에서 열린 '델테크놀로지스 월드 2018'에서 기조연설을 했다. 앞으로 발생하는 데이터의 양이 폭발적으로 증가할 것이며, 데이터의 대부분은 사람이 아니라 사물에서 생성될 것이라고 예측했다. 그는 "2020년이면 전 세계 대표 도시에서 매일 200페타바이트(HD급 영화 약 5,000만 편 용량)의 데이터가 발생할 것"이라며, "특히 데이터의 99%를 사물이 만들어낼 것"이라고 전망했다.

사물이 생산하는 데이터 외에, 사람이 만들어 내는 데이터의 양도 기하급수적으로 늘어날 것으로 전망된다. 사물이 생산하는 데이터의 일부 역시 사람과 사물의 교류에서 나오는 만큼, 결과적으로 사물과 사람이 데이터 생산의 주체가 되고 있다. 이처럼 데이터 폭증 시대가 오고 있지만, 생산되는 데이터의 대부분은 기업·정부·공공기관이 거의 독점적으로 관리하고 있다. 구글, 페이스북, 아마존, 에어비앤비, 우버 등 정보기술 기업들은 데이터 중심 플랫폼을 기반으로 데이터를 통합해 그들의 비즈니스에 반영하고 있다. 하지만 데이터 제공자로서 사용자에게 이들

• 필자 | **이상일**(디지털데일리 기자)

기업의 데이터 활용 전략은 불공평하고, 비대칭하며, 불평등하다는 지적이 나온다.

이러한 관점에서 최근 주목받고 있는 것이 '마이데이터(MyData)'다. 마이데이터는 데이터에 대한 업계의 요구와 디지털 인권을 결합한 개인데이터 관리의 인간 중심적 접근 방식이다. 마이데이터는 개인이 자신의 데이터를 제어할 수 있게 하는 것이다. 각 개인의 통제 하에 있지 않은 개인데이터는 마이데이터라고 할 수 없다. 구매 데이터, 교통 데이터, 통신 데이터, 의료 기록, 재무 정보 등 다양한 온라인 서비스에서 파생된 개인정보를 포함하는 자신의 데이터에 접근하고, 이를 자유롭게 활용할 수 있게 하는 것이 마이데이터다.

마이데이터의 사회적·경제적·실용적인 가치가 점차 커지고 있다. 세계경제포럼(World Economic Forum)은 "개인데이터는 21세기 사회의 모든 측면을 다룰 새롭고 가치 있는 자원이 됐다"고 강조한다.

마이데이터는 데이터 흐름을 단순화하고 기업이 개인정보를 보존하면서 혁신적인 개인데이터 기반 서비스를 개발할 수 있는 새로운 기회를 열어준다. 글로벌 시장에서는 인간 중심 개인데이터의 기회를 탐색하는 중으로 에너지, 건강, 웰빙, 금융 등 전체 산업에서 이를 주목하고 있다.

마이데이터의 개념은 개인의 권리와 자유를 더 중요한 개념으로 여기는 유럽연합(EU)을 중심으로 본격화되고 있다. EU에서는 5월부터 GDPR(General Data Protection Regulation)의 본격적인 발효로 개인정보에 대한 전향적인 움직임이 일어나고 있다. 이 법안은 모든 개인에 대해 수집된 데이터와 관련된 특정 권리를 보장하는 것을 목표로 한다. 개인정보를 취급하는 모든 기업은 해당 규정 하에 정보를 보유하면서 개인에게 해당 정보의 삭제나 교정을 요청할 권리를 부여하고, 개인이 동의하지 않으면 사용할 수 없게 된다. 자연스럽게 마이데이터와 관련한 다양한 사업 역시 유럽을 중심으로 활발히 추진되고 있다.

2. 스마트시티, 헬스케어 등 분야에서 활발[1]

바르셀로나의 스마트시티 개발 계획을 살펴보면 마이데이터와 관련한 다양한 사업이 추진되고 있어 주목된다. 바르셀로나시는 'BCN City Data Commons'라는 투명성, 프라이버시, 보안 및 혁신 핵심 데이터의 윤리적 사용을 중점으로 하는 새로운 도시 수준의 데이터 전략을 추진하고 있다. 주요 내용으로는 △윤리 기반 데이터 혁신 △데이터 주권, 보안 및 개인정보 △도시 차원의 데이터 인프라(개방형 표준, 단일 API) △책임 기술(알고리즘 투명성) △새로운 데이터 소유권 △새로운 권리 수호자로서의 도시의 역할 등이 꼽힌다.

[표 1-2-1] 주요국의 마이데이터 동향

국가	미국	영국	프랑스	핀란드
명칭	Smart Disclosure	Midata	MesInfos	Mydata
추진 목적	소비자의 정보 기반 의사결정 지원	소비자 주권 강화 및 의사결정 능력 제고	기업이 독식하고 있는 개인데이터 활용 편익 공유	개인 중심의 프라이버시 통제 및 데이터 가치 증대
편익	소비자 효용성 향상	경제 활성화	개인데이터 활용 편익의 개인, 조직, 사회 공유	데이터 경제 활성화
추진 체계	정부 전담기관(OMB, NIST) 분야별 추진은 전담부처에서 진행 (재향군인청, ONC, 에너지부 등)	정부 전담기관(BEIS) 민간연합체(운영 및 활동위원회)	민간연합체(FING & Cap Digital)	정부 전담기관(교통통신부, LVM)
민간 참여	Greenbutton Alliance(에너지 분야) Bluebutton Alliance (의료 분야)	운영 및 활동위원회(규제기관, 소비자단체, 분야별 기업 등 26개 기관)	파일럿테스트 참여 기업(데이터 제공기업 + 보험, 은행, 유통, 텔레콤 6사 + 서비스 지원 2사)	MyData Alliance (교통통신부 포함 40여개 기업 및 기관 참여)
관련 정책	Open Government National Action Plan for USA('11)	Better Choice, Better Deals('11) Open Data Initiative('12) Open Data Strategy('14)	없음	디지털 비즈니스 성장 전략('15) 데이터 비즈니스 활성화 방안('16)
법제화	대통령 지침/ 행정명령으로 소비자 요청 시 공공·민간의 데이터 제공 의무화	Enterprise and Regulatory Reform Act 2013 (2013.4) (금융, 에너지, 통신분야 의무화)	GDPR/PSD2 준수 Digital Republic Act (소비자 데이터 이동 조항 포함)	GDPR/PSD2 준수 핀란드 자체 의무화 법제도 없음
표준	민관협력을 통해 분야별 표준 및 상호운용성 확보	Open Standard Format 준수(재사용 가능하고 기계 가독형 데이터)	레인보우버튼(GDPR을 지원하는 데이터 전송·공유 표준 프레임워크 프로젝트	Trust Network 프레임워크 제시 (표준 및 상호 운용성을 보장하는 인프라스트럭처)
활용 분야	건강, 에너지, 교육 등	에너지, 금융, 통신 등	의료, 에너지, 개인, 일상 등	건강

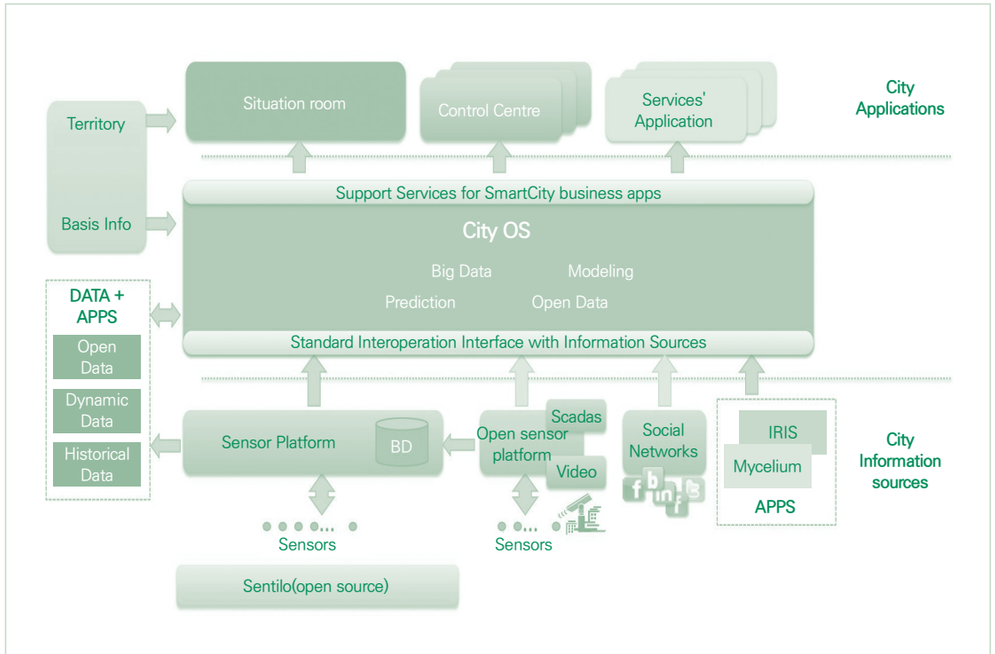
출처: 투이컨설팅

1 'MYDATA 2017' 컨퍼런스 자료(Presentation Materials and Videos, http://mydata2017.org/presentations)를 참고해 주요 사례 작성

오늘날 도시는 이전보다 많은 데이터를 보유하고 있다. 하지만 현재 도시가 보유한 데이터의 90%는 불과 몇 년 전에는 존재하지 않았다. 설령 도시가 데이터를 보유하고 있더라도 이 데이터는 조직적이지 않고 개인이 접근할 수 없다. 바르셀로나시는 시민들이 실시간 정보와 크라우드소싱(crowdsourcing)을 통해 데이터를 생성하고 사용할 수 있게 하고 있다. 지식은 중앙 집중식이 아닌 분산화하고, 데이터에 의거해 더 빠르고 더 민주적인 결정을 내리고, 혁신을 촉진하며, 공공 서비스를 개선하고, 사람들에게 권한을 부여한다는 목적으로 수집된다.

이러한 데이터는 합리적인 주택 제공을 위한 빅데이터 분석을 통해 임대료 상승을 통제하고 관광 산업의 부정적인 영향을 완화하기 위해 활용된다. 'BCN 주택 가격 지수'가 바로 그것이다. 도시 데이터 분석 사무실(City Data Analytics Office)은 모든 도시 데이터 분석의 중심 허브다. 이곳은 도시 행정기관과 시민들이 투명하게 도시 차원에서 더 나은 의사 결정을 내리고 시정의 책임을 물을 수 있도록 한다.

[그림 1-2-1] 바르셀로나 스마트시티 구성도



출처: 바르셀로나 시청

제조업 차원의 마이데이터 활용 사업도 속도를 내고 있는 분위기다. 스마트 헬스케어 워치(Watch) 전문 업체인 순토(Suunto)는 심박계 등의 센서가 달린 스마트워치를 공급해 개인데이터를 생성해 내고 새로운 가치를 찾는 데 주력하고 있다. 순토는 마이데이터 측면에서 수집된 데이터의 △투명도 △권한 부여 △접근성 △상호 운용성을 확보했다. 이를 기반으로 친구와 공유, 쉬운 프라이버시 정책 및 조건, 보안 서비스와 데이터 공유 기능 등을 지원하고 있다.

개인의 건강데이터와 관련된 연구도 진행중이다. 마이헬스마이데이터(MHMD, MyHealthMyData)는 EU 기금 지원 프로젝트다. 의료 기관, 연구 및 비즈니스 목적으로 임상기관, 개인, 연구센터 및 업계 간에 개인 건강데이터를 공유하고 교환하기 위한 플랫폼을 2019년 10월까지 완료를 목표로 개발하고 있다.

MHMD는 시민들의 개인 건강데이터의 소유 및 관리 권한을 부여하는 한편, 데이터 보호 및 보안, 의료용 대형 생물 의학 데이터세트의 가치 활용에 주목하고 있다. MHMD는 마이데이터 측면에서 블록체인 기술을 적극 활용하고 있다. 모든 트랜잭션이 네트워크에 의해 트랜잭션 블록을 형성하는 항목으로 지정된다. 건강데이터에 블록체인 방식을 적용하면 접근에 있어서 보안이 보장된다. 개별 데이터 소유권과 개인데이터의 저장소인 클라우드는 개인이 블록체인을 통해 모든 장치에서 데이터에 액세스하고 개인 용도로 사용하도록 한다. 서로 다른 소스(소셜 미디어 계정, 임상 데이터 저장소, 개인용 드라이브, 웨어러블 장치 등)의 단일 개인데이터를 단일 사용자 소유 계정으로 집계한다.

전력 사용에서도 마이데이터의 움직임이 본격화되고 있다. 'MyData 2017' 자료를 따르면, '스마트 미터(Smart Meter)'는 디지털 방식으로 에너지 측정을 보내고 원격으로 작동하는 가스 및 전기 계량기다. 스마트 미터를 적용함으로써 로드 밸런싱, 그리드(Grid) 관리, 재생 에너지 통합, 정전 감지, 운영비용 절감효과 등을 기대하고 있다.

프랑스의 전기 유틸리티 회사인 '에네디스(Enedis)'는 프랑스 전기의 95%를 공급·관리하고 있다. 에네디스는 유럽에서 가장 큰 전력 공급업체로서 3,600만 명의 고객을 보유하고 있다. 에네디스는 전력 거래 프로그램인 '키위파워(KiWi Power)'로 고객 동의 아래 고객의 데이터를 전송한다. 고객은 키위파워를 통해 전력 사용량이 집중되는 피크타임(peak time)에는 소비량을 줄일 수 있다. 에네디스는 공공에너지 유통 서비스 및 데이터 활용의 원칙으로 마이데이터에 주목함으로

써 △에너지 데이터 시스템 신뢰 강화 △시민을 위한 새로운 서비스 기회 제공 △성공적인 에너지 전환의 기회를 이끌어 내고 있다.

핀란드 헬싱키의 칼라사타마(Kalasatama)지역은 친환경 차량을 이용하고, 신재생 에너지를 생산하는 스마트시티로 개발 중인 파일럿 도시다. 이 지역의 에너지 관리 업체인 헬렌(Helen)은 지역 태양 에너지, 전기 자동차를 지원하는 기반 시설, 에너지 저장, 에너지 효율적인 빌딩 자동 제어를 사용해 카라사타마의 스마트 에너지 시스템을 구축하고 있다. 헬렌은 에너지 데이터를 사용하고자 하는 고객을 위해 API를 제공한다. 민간 아파트 API를 통한 에너지 데이터와 다양한 API 및 응용 프로그램을 통해 다양한 종류의 데이터를 결합해 '코티힐리(Kotihiili)'라는 개인용 실시간 이산화탄소(CO2) 계산기 등을 제공하고 있다.

3. 한국은 이제 첫발 떼는 분위기

한국에서 마이데이터에 대한 논의는 이제 첫 걸음을 떼는 분위기다. 개인정보가 언제, 누구에게, 어느 범위까지 알려지고 이용되도록 할 것인지를 정보주체 스스로 결정할 수 있는 권리인 '개인정보 자기결정권'에 대한 논의가 본격화되면서부터다.

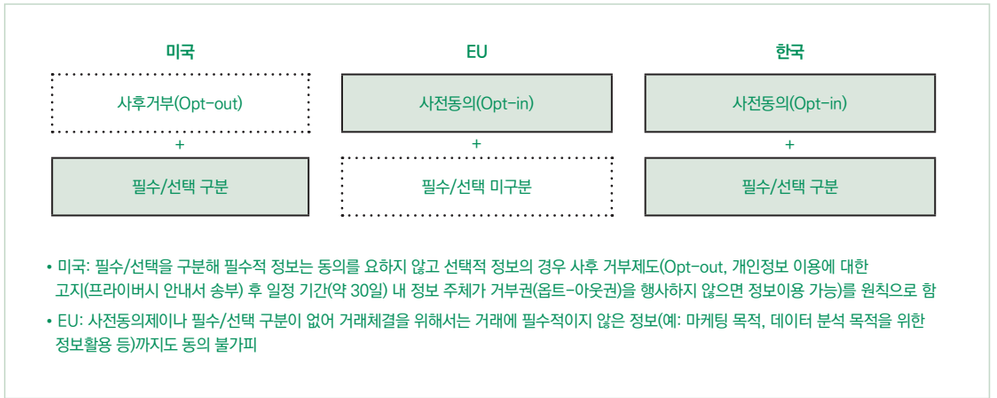
개인정보 자기결정권은 다른 기본권과의 관계, 데이터 활용의 공익적 필요성 등을 종합적으로 감안해 구체화되는 권리다. 최근에는 개인정보가 갖는 다차원적 속성을 인정하고 정보 유형별로 다른 접근이 필요하다는 인식이 확산되면서 자신의 정보를 스스로 통제하는 적극적 권리로 확대되고 있다.

특히 산업화·정보화 등의 진전에 따라 유통·활용될수록 사회적·개인적 효용이 증대되는 정보의 본질적 특성이 부각되면서 '개인정보' 그 자체보다는 '정보주체의 본인 정보에 대한 결정·통제권'에 대한 보호가 강조되는 분위기다.

또 빅데이터 등 정보 환경이 급격하게 변화하면서 정보 주체의 더 능동적인 권리 보장이 필요하다는 인식이 확산되고 있다. 이에 따라 개인 스스로 본인 정보를 관리·활용할 수 있는 권리로 프라이버시 개념이 확대되고 있다.

다만 국내의 정보활용 동의규제의 경우, 엄격한 사전동의 원칙 하에 필수 또는 선택 동의가 구분되는 등 여타 국가들에 비해 보호 수준이 강한 편이다. 정보주

[그림 1-2-2] 주요국 정보활용 동의제도 비교



출처: 금융감독원

체의 사후통제권, 정보보호 관련 권리구제·제재수단 등도 주요국에 비해 보호 강도가 매우 높다. 개인정보 자기결정권의 핵심적 구현 수단인 동의제도가 형식화돼 '알고 하는 동의(Informed consent)'가 이뤄지지 못한다는 문제도 제기된다. 정보 주체가 정보 활용 내역 및 프라이버시 침해 정도 등에 대해 정확하게 이해하고 동의하는 비율이 낮은 상황이기 때문이다. 또 기업은 정보 주체로부터 더 확실한 동의를 받는 데에만 주력하고 정보 주체에 대한 효율적 정보제공 노력은 소홀하다는 지적도 있다.

이에 정부는 정보활용 동의제도 개선을 통해 동의 내용에 대해 명확하게 '알고 하는 동의(Informed consent)' 여건을 마련 중이다. 새로운 유형의 권리침해 가능성에 선제적으로 대응하고, 적극적·능동적 권리보호 제도를 도입해 빅데이터 분석 결과 등에 대해 정보 주체가 적극적으로 대응해 나갈 수 있도록 '프로파일링 대응권'을 강화하고 있다. 또 더 나은 금융 서비스를 제공받기 위해 본인 정보를 능동적으로 관리·활용할 수 있는 '개인신용정보 이동권' 도입도 추진 중이다.

실제 2018년 5월 31일 행정안전부와 방송통신위원회, 개인정보보호위원회가 공동 주최한 '2018 개인정보보호페어(PIS FAIR 2018)'가 열렸다. 이 행사에서 방송통신위원회 고삼석 상임위원은 "인공지능, 사물인터넷, 자율주행 자동차, 드론 등 첨단기술이 발전하는 가운데 개인정보를 어떻게 활용할 것인지가 가장 큰 이슈"라며 "4차산업혁명위원회에서 진행한 해커톤에서도 개인정보의 보호와 활용에 관심을 보이고 있다. EU GDPR은 개인정보 보호와 활용을 동시에 추구하는 대

표적 사례다. 정부 차원에서도 GDPR 대응에 있어 글로벌 표준에 상응할 수 있는 제도 개선에 노력할 것"이라고 말했다.

개인정보보호위원회 박상희 사무국장도 환영사에서 "지난 25일 EU 개인정보보호법이 본격 시행되면서, 정부에서도 기업의 피해가 발생하지 않게 다양한 대응책 지원에 나서고 있다. 가명정보 도입 등 개인정보 활용 방안을 위해 각계에서 노력하고 있다. 국회에서도 여야를 막론해 다양한 법안을 내놓고 있다. 4차산업혁명시대에 급변하는 개인정보 환경에서 국민의 개인정보를 실질적으로 보호하고, 안전하게 활용할 수 있도록 노력하겠다"고 강조했다.

법적·제도적 뒷받침이 완전하진 않지만 마이데이터에 기반한 서비스들도 속속 모습을 드러내고 있다. 통합 자산관리 플랫폼 '뱅크샐러드(BankSalad)'는 고객이 실제 받을 수 있는 혜택을 계산해 가장 알맞은 카드를 추천하는 서비스를 제공중이다. 이 서비스는 자신이 평소 사용하는 카드만 연동하면, 자동으로 해당 카드의 지난 3개월~1년간의 소비 패턴을 상세 분석해 이용자의 소비 성향에 맞는 카드를 순위별로 알려준다.

뱅크샐러드 웹서비스에서 신용등급, 나이, 직업 등 자신의 조건에 가장 알맞은 금융투자상품도 추천 받을 수 있다. 현재 예적금, 보험, P2P 금융 분야에서 간단한 정보만 입력하면 가장 높은 이율을 받을 수 있는 상품정보를 제공한다. 또한 낮은 이율의 대출 상품을 찾아주는 서비스를 제공하고 있다.

뱅크샐러드는 사용자의 지출 현황 및 금융 생활을 분석해 상황에 맞는 조언과 격려의 메시지를 실시간으로 보내주는 '금융비서' 서비스도 제공하고 있다.

신용정보 블록체인 전문 기업인 '마이크레딧체인(MyCreditChain)'은 MCC의 블록체인 기반 신용정보 거래 플랫폼을 제공한다. 이는 개인의 신용정보가 신용정보기관에 의해 수집되고 금융기관 등으로 유통되던 기존의 신용정보 유통 체계와는 다르다. 즉 신용정보의 주체인 개인이 그 소유권을 갖고, 개인의 신용정보를 필요로 하는 곳과 직접 거래할 수 있도록 하는 신개념 신용정보 유통 플랫폼이다.

이 플랫폼에서 개인은 개인 신용정보의 소유자이며 금융서비스의 소비자로서 금융서비스를 제공하는 기관과 '기브 앤 테이크' 관계를 맺을 수 있게 된다. 금융기관은 MCC의 플랫폼을 통해 중개인 없이 개인들로부터 직접 필요한 개인 신용정보를 구매할 수 있고, 금융상품을 직접 판매할 수 있다.

스타벅스의 '사이렌오더' 애플리케이션도 고객의 성향과 주문 시간대에 알맞은 음료 메뉴를 추천하는 개인화 추천 서비스를 제공한다. 개인 고객의 최근 구매 이력을 포함해 매장 정보, 주문 시간대, 기온과 같은 데이터를 수집·분석해 상품을 추천한다. 예를 들어 오전 출근 시간대에는 아메리카노를 비롯한 에스프레소군 위주의 추천을, 저녁 시간대가 가까워질수록 티바나 메뉴들을 추천한다. 또한 특정 온도를 기점으로 따뜻한 음료와 차가운 음료의 판매와 수요를 예측해 추천에 반영한다. 고객이 음료를 담는 순간 해당 음료와 함께 많이 판매된 푸드 메뉴도 추천한다.

국내 업체인 '빅밸류(BigValue)'는 위치기반 빅데이터 분석 시스템과 인공지능 기술을 활용해 연립·다세대 주택의 실거래정보와 시세 정보를 제공하는 로빅(LOBIG)을 개발했다. 대규모 공동주택인 아파트에 비해 세대수가 많지 않은 소규모 공동주택인 빌라나 다세대 주택은 시세 산정에 어려움이 많다. 로빅은 연립·다세대 종합 시세 변동 추이 및 거래량 등의 분석 정보를 제공하기 위해 주택실거래가, 건축물대장, 토지대장 등 48개 공공기관의 공공데이터를 활용했다. 빅밸류는 삼성벤처투자, KB인베스트먼트 등 대형 투자사로부터 투자 유치를 받고 사업을 추진 중이다.

앤톡은 개인 투자자 대상 증권 및 기업 정보를 시각화해 제공하는 전문 핀테크 스타트업이다. 사명은 개인 투자자를 상징하는 개미(Ant)와 주식(Stock)의 합성어로, '앤톡닷컴'을 통해 국내·외 상장 및 비상장 기업 분석 서비스를 제공하고 있다. 앤톡은 서비스 제공을 위해 공공데이터포털, 금융감독원, 한국증권거래소, 예탁결제원 등으로부터 확보한 각종 공공데이터를 적극 활용했으며 그 결과, 지난해까지 약 1억 5,000만 원의 매출을 달성했다.

4. 마이데이터의 미래

마이데이터는 수많은 곳에 분산된 나의 데이터를 효과적으로 관리하고, 더 나아가 내 데이터가 어디서 어떻게 쓰이고 있는지를 모니터링하고, 개인데이터 활용 동의를 승낙하거나 철회할 수 있게 하는 것을 의미한다.

다만 경제 활동이 온라인으로 이동하는 디지털 트랜스포메이션 시대에 개인

이 일일이 자신의 정보를 관리하는 것이 쉽지 않은 것도 사실이다. 이에 따라 최근에는 개인데이터에 대한 관리를 해주고 이 정보를 중개해주는 개인정보중개자(Personal Data Manager) 역할이 커지고 있다.

개인정보중개자 혹은 개인정보사업자는 개인의 동의 하에 개인정보를 원하는 기업들을 대상으로 개인 정보를 제공하고 이에 대한 가치를 다시 고객에게 제공하는 것을 목적으로 한다.

국내에서도 금융 당국이 '본인신용정보관리업' 도입을 추진하고 있다. 본인정보관리업자는 예금·대출·카드거래 등의 정보를 망라하는 본인 신용정보의 통합조회 서비스를 제공한다. 이를 기초로 소비패턴, 위험성향 등을 파악해 자산관리서비스 제공 및 맞춤형 금융상품 추천 등을 할 것으로 전망된다.

제3장

마이데이터, 개인정보 보호와 활용

개인의 권리를 보호하면서도 데이터의 흐름을 자유로이 하자고 하는 것이 마이데이터에서 추구하는 개념이다. 한국의 개인정보관련 법에서는 데이터의 흐름이 자유롭지 못하다. 그 이유 중 하나는 개인정보에 관한 개념이 매우 광범위 하면서도 분명하지 않다는 것 때문이다. 그 틀을 깨야 할 때가 다가오고 있다. 마이데이터는 정보의 자산화 시대에서 반드시 필요한 개념이고, 그 개념에 맞게 관련법과 제도도 마련돼야 한다. 지금까지의 법과 제도로는 마이데이터에서 추구하고 있는 개념을 실행하기엔 많은 어려움이 따를 수밖에 없다.

『2017 데이터산업 백서』를 보면 데이터 산업이 향후 어떻게 전개될 것인가를 한 눈에 잘 알 수 있다. 이 백서의 설문조사를 따르면, 데이터 부분 데이터산업 분야에서 2018년도 최고 이슈는 '데이터 자산가치'로 결론이 났다. 설문조사 대상 총 18개 항목 모두 중요하지만, 모든 연령대에서 공통적으로 다음과 같은 항목이 중요하다는 설문 결과가 나왔다.

1. 데이터 자산의 필요성 및 자산적 가치 평가에 대한 공감대 확산
2. 퍼스널데이터 시대 도래
3. 개인데이터의 보호와 활용 논쟁 지속
4. 데이터 주권에 대한 이슈 발생

데이터 자산가치에 대해서는 특히 40대와 50대에서 중요하게 평가했고, 20대는 상대적으로 낮은 3번 째로 중요하다는 반응을 얻었다. 반면 산업계나 학계 또는 기타 다른 업계에서도 모두 동일하게 '데이터 자산가치'를 중요도 1순위로 꼽았다. 이제는 데이터에 대한 값어치를 따져야 할 시대가 도래한 것이다. 심지어는 데이터의 자산가치에 대해서 데이터를 무궁무진한 발굴의 가치가 있는 석유로

• 필자 | **이재욱**(법무법인 율촌 미국변호사)

비유하는 사람들도 있다.

이러한 가운데 마이데이터 운동이 점점 확산돼 가고 있다. 유럽의 'European PIMS(Personal Information Management Services) Community'의 세 사람[1]에 의해 2015년 11월 브뤼셀에서 처음 시작된 운동이 바로 그것이다. 매년 해를 거듭할수록 급속하게 번져서 이제는 전 세계적인 운동으로 자리잡고 있다. 개인데이터에 대해 정보의 주체가 각각 권한을 갖고 관리할 수 있게 하자는 취지의 운동이다. 이들이 스스로 지식을 축적하고, 올바른 결정을 할 수 있고, 개인 또는 조직 간에 서로 효율적으로 상호 교류를 할 수 있게 하는 것에 공동의 목표를 두고 있다.

이 운동의 핵심은 '마이데이터 선언(MYDATA DECLARATION)'[2]에 잘 설명돼 있는데 다음과 같은 내용을 담고 있다.

"사회에서 개인정보의 중요성이 커짐에 따라, 개인이 자신의 개인데이터를 알고 제어할 수 있는 위치에 있도록 하는 것이 매우 시급한 상황이 돼 가고 있다.

오늘날, 힘의 균형은 조직 쪽으로 매우 기울어져 있다. 이러한 조직들은 개인데이터를 기반으로 수집·거래·의사 결정 권한을 가지고 있는 반면, 개인은 단지 자신들의 데이터에 대해 통제권을 얻기를 희망하고 있을 뿐이다.

이 선언문에서 우리가 제시한 변화(shift)와 원칙은 균형을 회복하고 인간 중심의 개인데이터 비전으로 나아가는 것이다. 이러한 변화와 원칙은 정당하고 지속 가능하며 번영할 수 있는 디지털 사회를 위한 조건이라고 생각한다. 이러한 디지털 사회는 다음과 같은 기초에 근거해야 한다.

- **신뢰와 확신**: 사람들 사이뿐만 아니라 개인과 조직간의 관계에 있어서 균형 잡히고 공정한 관계에 기반을 둠
- **자기 결정**: 법적으로 보호돼야 하는 것뿐만 아니라 개인들과 데이터의 힘을 적극적으로 공유함으로서 달성됨
- **개인데이터의 공동 이익 극대화**: 조직·개인·사회에서 개인데이터를 정당하게 공유함"

1 ① Antti 'Jogi' Poikola(Finland), researcher at Aalto University.
② Daniel Kaplan(France), cofounder and scientific advisor to Fing and France's MesInfos project
③ Tanel Mällo(Estonia), head of Research Administration Office at Tallinn University

2 MyData Declaration, ver1.0, http://mydata.org, 2018.7.2. 접속

2017년 3월에 있었던 베를린 대회에서는 '각서(Memorandum)'를 발표하면서 2가지 상호 보완적인 목표를 세우고 있다. 2가지 목표를 살펴보면 △마이데이터 글로벌 네트워크를 법률적 조직으로 형성 △인간 중심의 개인적 데이터를 위한 공통적인 원칙 개발 착수를 들고 있다. 이 중에서 마이데이터 선언에 대해서는 두 번째 목표에 근간을 두고 있는 것이다.

마이데이터 선언과 더불어 같이 제정된 마이데이터 원칙[3]은 다음과 같이 규정하고 있다.

- **인간 중심의 통제와 사생활**: 개인은 온라인과 오프라인에서 개인 생활과 관련된 정보의 관리에 있어서 수동적인 대상이 아니라 권한을 부여 받은 행위자다. 개인들은 자신들의 데이터와 프라이버시를 관리할 수 있는 권한과 적절하고 실질적인 방법을 지니고 있어야 한다.
- **사용 가능한 데이터**: 개인데이터는 기술적으로 액세스하고 사용하기가 쉬워야 한다. 안전하고 표준화된 API(Application Programming Interfaces)를 통해 기계가 읽을 수 있는 개방형 형식으로 액세스할 수 있어야 한다. 마이데이터는 폐쇄적인 사일로의 데이터를 재사용 가능하고 중요한 리소스로 변환하는 방법이다. 개인이 삶을 관리할 수 있도록 도와주는 새로운 서비스를 만드는 데 사용될 수 있다. 이러한 서비스를 제공하는 업체는 새로운 비즈니스 모델을 창출하고 사회에 경제적 성장을 창출할 수 있다.
- **개방된 사업 환경**: 마이데이터 공유 인프라는 개인데이터의 분산 관리를 가능하게 하고, 상호 운용성을 향상시킨다. 또한 기업이 데이터 보호 규정을 더 쉽게 준수할 수 있게 하고, 개인이 데이터 종속 없이 서비스 제공자를 변경할 수 있도록 한다.

전 세계의 다양한 개인정보보호와 관련된 법에서 마이데이터에서의 개인정보보호에 대한 개념을 얼마나 도입했는지는 지난 5월 25일 시행된 EU의 GDPR(General Data Protection Regulation)을 보면 잘 알 수 있다

GDPR에서는 개인에게 알 권리·열람권·정정요구권·삭제권·제한처리요구

3 Antti Poikola, Kai Kuikkaniemi, Harri HonkoMy. (2014). Data – A Nordic Model for human-centered personal data management and processing. http://urn.fi/URN:ISBN:978-952-243-455-5

권·데이터이동권·반대할 수 있는 권리·자동화된 결정 절차·프로파일링에 관한 권리 등 자기통제권을 규정하고 있다. 이것은 마이데이터에서 추구하고 있는 개념을 많이 반영했다고 볼 수 있다. 또한 'Privacy by Design'과 'Privacy by Default'라는 개념도 도입했다. 이는 서비스·애플리케이션·소프트웨어·하드웨어를 포함한 제품을 설계할 때부터 개인정보보호를 고려해야 한다는 개념으로, 개인의 권한을 더 보호하려는 노력의 일환으로 볼 수 있다.

한편 한국의 개인정보보호법에서도 이러한 권리들을 보장하고 있다. 개인들에게 부여된 권리에 대해서는 GDPR보다는 약간 낮은 단계의 권한을 부여하고 있다. 예를 들면, GDPR에서는 데이터 이동에 대해 내 개인정보를 가지고 있는 정보처리자에게 다른 정보처리자나 수탁처리자에게 전달하라고 요청할 수 있다. 반면 한국법에서는 일단 제3자 제공에 대해 내 정보를 제공받을 제3자 조직에 대해서 통지를 받고, 그에 대한 명시적 동의를 해주면 된다. 다시 말하면 제3자 제공에 동의를 한 상태라면, 정보처리자에게 다른 조직으로 내 정보를 옮겨 달라고 요청하거나, 다른 조직으로 옮기지 말아야 한다는 선택적 결정을 할 수 있는 권한에 대해 구체적으로 규정돼 있지 않다.

개인의 권리를 보호하면서도 데이터의 흐름을 자유로이 하자고 하는 것이 마이데이터에서 추구하는 개념이다. 이러한 점에서 한국의 개인정보관련 법에서는 데이터의 흐름이 자유롭지 못하다. 그 이유 중 하나는 개인정보에 관한 개념이 상당히 광범위하면서도 분명하지 않다는 것이다. 한국법에서는 다른 정보와 결합해서 개인을 식별할 수 있는 정보 또한 개인정보로 규정하고 있다. 따라서 개인정보에 관한 범위가 상당히 넓어지다 보니 빅데이터, IoT 등에 사용되는 개인정보도 상당히 까다로워질 수밖에 없고, 자연적으로 빅데이터, IoT 등에서 사용할 수 있는 개인정보가 제한적일 수밖에 없게 된다.

반면 GDPR에서는 필요한 경우에만 동의를 받게 돼 있고, 개인정보 주체도 동의를 언제든지 철회할 수 있는 권리가 보장돼 있다. GDPR에서는 한국법과 마찬가지로 정보주체의 동의를 받아야 하지만, 정보처리자 또는 제3자의 정당한 이익의 목적으로 개인정보처리가 필요한 경우는 동의를 받지 않아도 된다. 이는 동의를 기본원칙으로 두고 있는 한국법과는 달리 GDPR에서는 유연성을 두고 있어서 개인정보의 흐름도 한국법보다 더 자유스러울 수밖에 없다.

마이데이터의 개념이 제대로 받아들여지고 시행되기 위해서는 개인들이 이러한 권리들을 잘 알 수 있도록 해야 하고, 그것들이 잘 보호될 수 있도록 조직 내에서 시스템적인 대책을 마련해야 한다. 비록 비용이 들더라도 장기적으로 그것이 곧 개인의 데이터 통제 실현으로 이어지는 방법이다.

정보의 독점 시대는 지나갔다. 이제는 정보의 공유 시대이고, 정보의 자산화 시대가 도래하고 있다. 이러한 시대의 흐름에 걸맞게 정보의 주체에게 권한을 돌려주고 자산에 대한 가치도 돌려받게 하자는 흐름도 나오고 있다. 지금까지는 기업에서만 정보를 가지고 이득을 취했다. 개인의 데이터가 쌓이면 그 회사는 다른 회사로 정제된 정보를 제공하거나 그 회사가 지닌 정보주체에게 제3의 회사 정보를 알려주면서 이득을 보았다. 사람들의 의지와는 상관없이 개인정보들을 이용해 수익을 올리고 있었던 것이다.

최근에는 마이데이터에 대한 개념을 넘어서, 내 정보를 사용할 수 있게 허락하고, 그 정보를 사용해 금전적 이득을 취하는 경우에 그 이득을 정보주체자도 나눠 받을 수 있도록 하자는 주장도 나온다. 그 예로 Finotek에서 개발한 블록체인 기반 데이터 플랫폼 리빈(LIVEEN)을 들 수 있다. LIVEEN은 각 개인이 생성하고 있는 일상의 가치정보(예: 구매 습관, 성향, 행동반경 등 모든 개인이 생성하고 있는 데이터)를 공유한다. 그 정보를 필요로 하거나 알기를 원하는 회사 또는 개인은 VEEN이라는 암호화폐를 정보 주체자에게 지불하고, 그 정보에 접근할 수 있는 권한을 부여 받게 된다. 이러한 방법으로 정보를 제공받는 회사나 개인은 정보 주체자와 소득의 일부를 나누게 되는 것이다.

개인정보를 보호하면서도 효율적인 활용을 통해 자산을 축적하는 시대가 다가오고 있다. 산업은 이제까지 상상하지 못했던 속도로 발전하고 있는 반면, 법과 제도의 변화는 아직도 과거에 머물러 있어 통제와 규제에 초점을 맞추고 있다.

이제는 그 틀을 깨야 할 때가 다가오고 있다. 마이데이터는 정보의 자산화 시대에서 반드시 필요한 개념이고, 그 개념에 맞게 관련법과 제도도 마련돼야 한다. 지금까지의 법과 제도로는 마이데이터에서 추구하고 있는 개념을 실행하기엔 많은 어려움이 따를 수밖에 없다. 사실 법이 미래를 예측해 스스로 변하기는 쉽지 않다. 하지만 시대가 변하고 있고, 법과 제도 역시 그 흐름에 따라야 한다. 우리의 산업은 끊임없이 발전하므로 더욱 빠르고 현명하게 준비해야 한다.

인공지능의 시대, 살아 숨쉬는 데이터

이형칠 | 윕스 대표
2017 데이터 구루상 수상자

인공지능에 탑재할 데이터라는 연료의 품질이 좋아야 한다. 데이터가 힘을 발휘하도록 하기 위해서는 데이터가 언제나 살아 숨쉬고 있다는 점을 인지해야 한다. 그렇기에 이러한 데이터의 구조, 통합, 해석하는 기술 및 지식을 갖춘 인력의 힘이 무엇보다 중요하다.

기업들은 다양한 데이터를 활용해 경쟁력을 높이고 싶어한다. 하지만 데이터에서 어떤 가치를 도출하려면 그저 쌓아 놓기만 해서는 될 일이 아니다. 데이터가 있다는 것만으로는 별 의미가 없다는 말이다.

데이터란 무엇일까? 데이터의 정의부터 살펴보면, 특정 목적의 활동이나 이벤트로 인해 발생된 사실(FACT, 수치)이 되는 자료다. 쉽게 예를 들어보자. 오늘의 날씨를 전해주는 아나운서가 말하는 온도, 습도, 미세먼지, 바람의 속도 등을 나타내는 수치들이 바로 데이터다.

데이터와 더불어 자주 언급되는 단어로 '정보'가 있다. 데이터와 정보를 혼용해 사용하는 경우가 많다. 엄밀히 말하자면 정보란 데이터를 처리해 나온 결과로서 사용자의 의사 결정에 도움을 준다. 이렇게 가공된 정보를 얼마나 잘 만들어 내고 활용할 수 있는지가 기업의 경쟁력과 대응력에 영향을 끼친다.

특허 데이터 전문기업으로 성장하기까지

필자는 특허 정보 서비스업에 종사하고 있다. 특허 정보가 기업 비즈니스에 도움이 되는 이유는 세상에 쏟아져 나오는 새로운 기술과 기업들의 비즈니스 전략을 알려주는 유용한 자원이 되기 때문이다.

2000년 하버드비즈니스스쿨(HBS)에서 출간한 『지식경영과 특허전략』에서 케빈 G.리베트 등의 저자들은 특허 정보가 미래 비즈니스 전쟁에서 '스마트 폭탄(Smart Bomb)'이 될 것이라 예견했다. 특허는 세계 각국의 특허청을 통해 공개된다. 특허청은 기업들이 활용할 수 있도록 전자화한 형태로 공보 데이터를 제공한다. 최근에는 정부 정책의 일환으로 시행되는 공공데이터 개방에 힘입어 누구나 이를 활용해 특허정보 검색 서비스 시장에 도전할 수 있다. 그럼에도 이 시장으로의 진입이 어려운 이유는 특허 데이터만이 갖고 있는 속성들 때문이다.

매일 신규 특허가 발행되며 출원된 특허들 중에는 소멸되는 것도 있다. 즉 특허에는 '생사정보'라는 것이 발생한다. 그 밖에도 패밀리·인용·양수양도 등 다양한 변동정보가 끊임없이 실시간으로 발생한다. 특허 자체가 생명력을 갖고 있는 것이다. 거기에 표준화되지 않은 채 공급되는 항목과 내용도 많다. 지금이야 민간 기업들이 특허 데이터를 활용할 수 있도록 공공 차원에서 데이터 품질 향상에 많은 노력을 쏟고 있지만, 필자가 특허정보 서비스를 제공하기 시작한 2000년대 초반에는 사정이 그리 녹록하지 못했다.

각국 특허청에서 발행되는 지식재산 데이터가 생산 시기별·권리별 또는 공보 종류별로 구분된 스키마(schema), 데이터 타입, 항목 설명 등에서 매우 상이했다. 이로인해 데이터를 활용할 수 있는 형태로 가공하는 데에 많은 시간과 투자가 필요했다. 게다가 한 해 출원되는 특허가 전 세계적으로 약 120만 건에 이르렀다는 점(2016년 기준으로 전 세계 약 313만 건의 특허가 출원됐다)을 생각한다면, 특허 데이터 비즈니스에 도전했던 그 당시는 어떻게 보면 매우 무모했을지도 모른다. 게다가 국내에서는 최초의 도전이었으므로 다른 누군가에게 특허 데이터의 가공 과정을 배운다는 것은 꿈도 꿀 수 없었다.

요즘 들어 이러한 일이 가능한 것은 정보 기술이 어느 정도 뒷받침하고 있기 때문이다. 엄청나게 많은 양의 데이터를 저장하기 위한 분산 파일 시스템, 다양한 형태의 데이터를 실시간으로 처리·분석하기 위한 방법, 데이터의 접근과 이동을 위한 클라우드나 인터넷 인프라 등을 특허 데이터 속성에 맞게 활용하고 있다. 하지만 이런 기술을 적용하는 데도 많은 시행착오를 겪어야 했으며, 지금도 겪고 있다. 특허 데이터의 속성·구조 등을 파악해 최적화한 데이터 구축 및 가공 프로세스를 설계하기까지의 노력과 인내가 필요하다. 지금이야 여러 시행착오에서 얻은 경험이 경쟁력이 돼 특허 데이터 전문기업으로 성장하게 됐다. 그리고 현재 단일 DB로 세계 최대 규모인 약 2억 건, 이를 바탕으로

자체 가공정보를 포함해 약 8억 건의 특허 데이터를 고객들에게 제공할 수 있게 됐다. 결과적으로 국내외 많은 사용자들로부터 데이터에 대한 신뢰를 얻고 있다.

인공지능을 위한 인간의 숨은 노력

필자는 최근 4차산업혁명으로 대두되고 있는 인공지능(AI) 기술이 많은 사람에게 환상을 심어주고 있는 것은 아닌지 조금은 우려된다. 현재 시장에 나와 있는 인공지능 스피커 관련 광고만 보더라도 그런 환상에 젖어 들게 한다. 알고 싶은 질문을 던지면 인공지능 스피커가 답을 준다. 특허 분석 작업의 경우를 보자면, 인공지능 기술만 적용하면 별다른 노력 없이도 원하는 목적에 맞는 분석 결과를 도출해 내고 이로부터 어떠한 통찰력까지 얻을 수 있을 것만 같은 착각에 빠져들게 한다. 하지만 그렇지 않다. 데이터를 업으로 하는 사람으로서 냉정하게 바라봐야 한다. 아무리 인공지능이라 하더라도 분석 대상이 되는 데이터가 제대로 준비되지 않는다면, 그 분석 결과로 도출된 통찰이라는 것 역시 기업 경영에 어떠한 도움도 되지 못할 것이다. 더 나아가 잘못된 분석이 잘못된 경영 판단으로 이어질 수도 있다.
특허문헌의 구성은 스키마가 있는, 즉 어느 정도의 구조를 표현할 수 있는 정형 데이터뿐만 아니라, 텍스트 형태로 제공되는 비정형 데이터가 혼재돼 있다. 이 비정형 데이터는 사람들이 일상적으로 쓰는 언어와 달리 특허만의 매우 독특한 기재 방식을 취하고 있다. 예를 들면, 특허로 보호 받고자 하는 기술의 범위를 독립항과 종속항으로 구분해 작성하고 있다. 그 뿐만이 아니다. 시간이 지나도 변하지 않는 출원날짜, 발명자 등의 불변의 데이터를 포함하는 것은 물론, 공개 및 등록 일자, 현재 권리자 등 시간이

지남에 따라 새롭게 생기는 데이터, 변하는 데이터 등이 포함돼 있다. 제도의 변화로 인해 변동되는 데이터들도 다수 존재한다. 그렇기에 이들을 수집하고, 검색과 분석이 가능한 형태로 통합하는 작업에는 꽤 많은 논리적인 접근과 기획력을 필요로 한다. 만약 이러한 작업이 제대로 이뤄지지 않는다면, 제아무리 인공지능을 활용한 특허 검색을 한다고 하더라도 검색 결과에 누락이 발생하거나, 노이즈가 늘어 오히려 분석 작업을 힘들게 하거나 잘못된 비즈니스 방향을 제시할 수도 있다.

데이터 품질과 사람의 역할

향후 인공지능 기술이 다양한 분야에 적용될 것임은 분명하다. 데이터 분석에 인공지능을 적용하기 위해서 개발되고 있는 알고리즘도 중요하다.
그렇더라도 근본적으로 이 알고리즘에 탑재할 데이터라는 연료의 품질이 좋아야 한다. 연료의 품질을 좋게 하는 데에는 무엇보다 사람의 역할이 매우 크다.
특허 데이터의 경우 '특허제도'에 대한 이해, 특허문헌에 대한 구조적인 이해가 뒷받침되지 못했을 때 발생하는 다양한 오류들, 이를 수정하는 데 필요한 노력들을 지난 십수년의 비즈니스를 통해 경험했다. 그렇기에 데이터 비즈니스에 있어서 가장 핵심은 단순히 'IT 전문가'라기보다는 '각기 다른 도메인에서 생산되는 데이터의 속성을 잘 이해하고 있는 IT 전문가'에 있다고 볼 수 있다. 필자의 경우 오랜 기간 특허 데이터를 속속들이 파악하고 있는 IT 전문가를 양성하는 데 힘써왔다. 그들은 이제 특허 분야에 있어서 최고의 전문가가 됐다. 이것은 다른 데이터 영역에서도 마찬가지일 것이다.
데이터가 힘을 발휘하도록 하기 위해서는 데이터가 언제나 살아 숨쉬고 있다는 점을 인지해야 한다. 그렇기에 이러한 데이터의 구조, 통합, 해석하는 기술 및 지식을 갖춘 인력의 힘이 무엇보다 중요하다. 그리고 이러한 인력들이 인공지능과 같은 새로운 IT와 만날 때 더욱 빛을 발할 수 있다. 결국에는 사용자들이 기대하는 수준의 서비스를 제공할 수 있게 되지 않을까 한다.

제 2 부

데이터 정책 동향

제1장

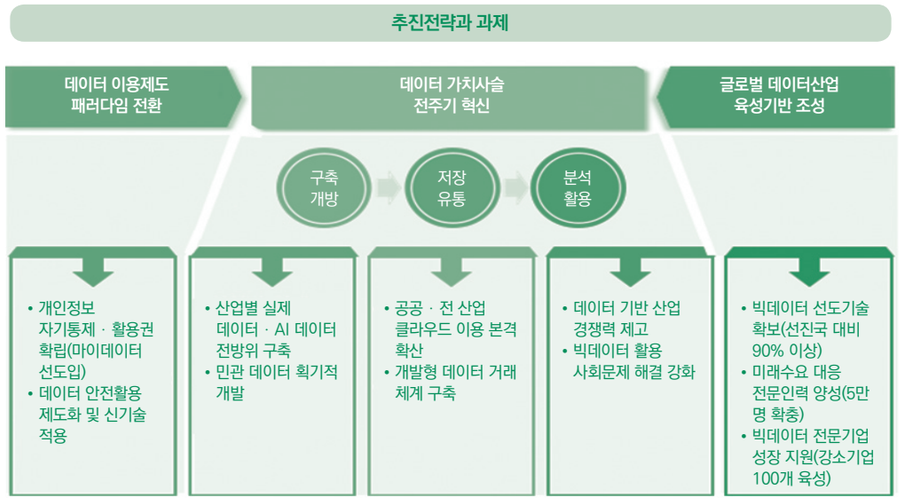
데이터산업 활성화 전략[1]

우리 정부는 '데이터를 가장 안전하게 잘 쓰는 나라'를 목표로 데이터산업 활성화 전략을 발표했다. 지난 2017년 12월에 발표한 초연결 지능형 네트워크 구축방안(N)과 2018년 5월 인공지능 R&D 전략(A)에 이어 데이터 산업 활성화 전략(D)을 수립함으로써 4차산업혁명에 대응하는 DNA 대책을 마련했다. 이번 데이터산업 활성화 전략은 △데이터 이용제도 패러다임 전환 △데이터 가치사슬 혁신 △글로벌 데이터 산업 육성 기반 조성 등 3대 과제로 구성됐다.

데이터가 4차산업혁명시대의 핵심자원으로 부상함에 따라 주요 선진국은 주력 산업 재도약과 혁신성장을 도모하는 데이터 경제 활성화[2]에 역점을 두고 있다. 그러나 우리나라는 엄격한 개인정보 규제로 인해 데이터 거래 및 산업적 활용이 위축돼 왔다. 자율차·스마트시티 등 4차산업혁명시대에 필수인 영역별 실제 데이터와 AI 학습용 데이터 구축이 미흡한 상황이다. 또한 중소기업·스타트업이 필요로 하는 데이터가 부족하고 품질 수준 역시 떨어져 산업에서 활발하게 활용되지 못하고 있다. 공공·민간의 클라우드 컴퓨터 이용이 낮아 데이터 저장·관리가 비효율적으로 이뤄지고 있으며, 데이터를 다루는 전문 인력 역시 매우 부족한 수준이다.

이에 정부는 이번 전략을 통해 '데이터를 가장 안전하게 잘 쓰는 나라'를 비전으로 정하고 데이터를 통한 혁신성장과 삶의 질 향상, 글로벌 수준의 데이터 보호와 활용을 기본 방향으로 다음과 같은 3가지 중점 추진과제를 도출했다.

[그림 2-1-1] 데이터 보호와 활용을 위한 3가지 중점 추진과제



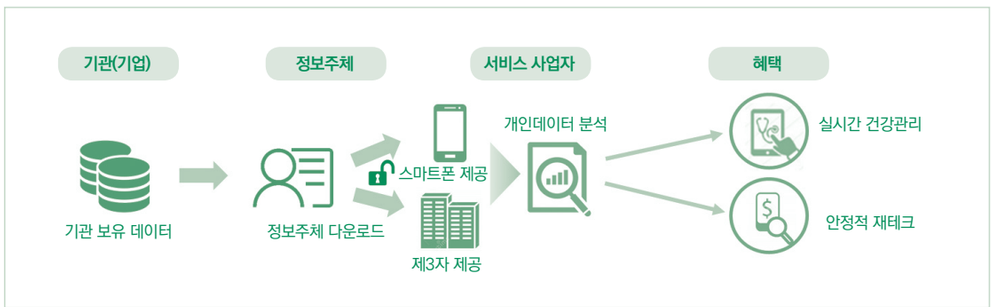
1. 데이터 이용제도 패러다임 전환

가. 데이터 이동권 확립: 국민의 데이터 주권MyData 찾기

개인데이터 관리가 기관(기업) 중심에서 정보주체 중심으로 전환하는 마이데이터(My-Data) 제도가 도입된다. 마이데이터는 정보주체가 기관(기업)으로부터 자기정보를 직접 내려 받아 이용하거나 제3자 제공을 허용하는 방식으로 개인정보 관련법 개정 없이도 바로 시행 가능하다.

정부는 국민들이 실생활에서 실질적으로 체감할 수 있는 의료·금융·통신 등

[그림 2-1-2] 마이데이터의 개요



• 필자 | **이종서**(한국데이터진흥원 차장)

1 대통령직속 4차산업혁명위원회 제7차 회의('18.6.26) 안건 「데이터 산업 활성화 전략」 요약 정리
2 미국 빅데이터 R&D 전략('16)/ EU 데이터경제 육성전략('17)/ 일본 Society5.0 실현데이터계획('17) 등

의 분야에서 대규모 시범사업을 실시한다. 이를 통해 건강증진·재테크·통신비 절감 등 국민들이 체감할 수 있는 효과를 창출해 국민들의 인식을 전환하면서 제도화할 예정이다.

나. 개인정보 안전 활용 촉진

개인정보의 안전한 활용을 위해 국민적 신뢰에 기반한 활용 제도가 마련된다. 먼저 4차산업혁명위원회 해커톤에서 논의한 개인정보 관련 '사회적 합의 결과'[3]를 바탕으로 개인정보보호법 등 관련법 개정을 추진할 예정이다. 또한 데이터 자체의 반출은 안 되지만, 데이터 분석 및 AI 개발 결과만 반출이 가능한 데이터 안심존[4]을 내년에 구축해 제공한다. 아울러 기술 측면의 신뢰를 높이기 위해 개인정보 비식별조치 콘테스트, 데이터 보안성이 높은 블록체인, 동형 암호 등 신기술 적용·실증을 추진해 나갈 계획이다.

이외에도 유럽연합(EU)의 일반개인정보보호법(GDPR) 시행('18.5.25)에 따른 한국 기업 피해 최소화를 위해 대응 가이드라인을 마련하고 한-EU 간 적정성 평가 승인을 연내 마무리할 예정이다.

2. 데이터 가치사슬 전 주기 혁신

가. 양질의 데이터 구축·개방

4차산업혁명의 핵심 기반인 산업별 실제데이터, AI 학습 데이터를 전방위로 구축하고, 공공·민간 데이터의 획기적 개방을 추진한다. 산업별 원시 데이터(raw data)의 풍부한 수집·생성을 위해 빅데이터 전문센터를 육성하고 각종 빅데이터센터 간 협력 네트워크를 확대할 계획이다.

또한 국내 인공지능(AI) 산업의 비약적 발전을 위해 범용(이미지, 상식 등)·전문(법률, 특허, 의료 등) 분야 AI 데이터세트를 수요 중심으로 단계적 구축·보급(~'22)하며, 민간 수요가 높은 데이터는 국가중점 데이터로 선정해 조기개방을 확대한다.

3 개인정보 범위 명확화, 비식별조치 근거인 가명·익명정보 개념 정립 등

4 이용자가 원격 분석 시스템에 접속해 데이터(표본 DB, 맞춤형 DB)를 분석할 수 있는 가상 PC 환경 및 다양한 분석 소프트웨어(오픈소스 도구, 시각화 도구 등) 제공 등 온·오프라인 샌드박스 개념 형태

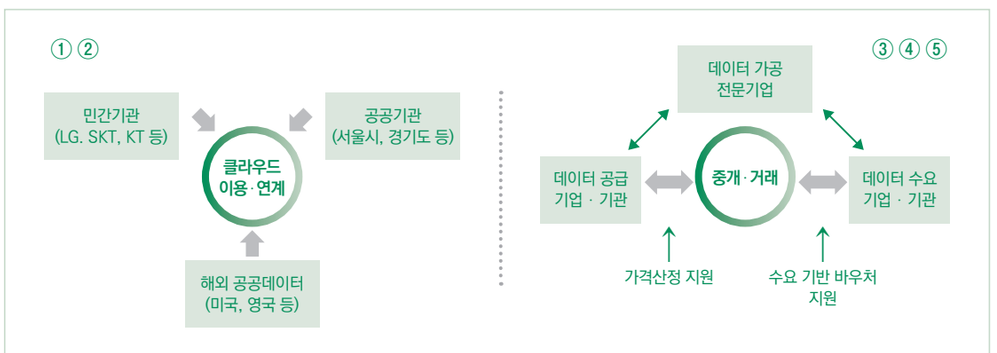
데이터의 획기적인 개방을 위해서는 모든 공공기관이 보유한 공공데이터를 전수 조사하고 데이터 현황을 보여주는 국가데이터 맵을 구축해 공공데이터 통합관리체계를 마련한다. 공공데이터는 원시 데이터 형태로 최대한 모아 국가안보·개인정보 등이 포함된 데이터를 제외하고 원칙적으로 개방한다. 기계학습이 용이하도록 개방하는 공공데이터의 오픈포맷(3단계) 비중을 지속적으로 확대하고, 품질관리 고도화를 통해 전 공공기관의 데이터 품질을 대폭 개선할 계획이다.

나. 데이터 저장·유통 활성화

데이터의 효율적인 저장·관리를 위해 민간 클라우드 이용을 확산하고, 데이터 유통 촉진을 위한 데이터 거래 기반을 본격 추진한다. 먼저 민간 클라우드 서비스 이용기관을 공공기관에 한정하지 않고 지자체 등으로 확대하는 규제 완화를 추진한다. 스마트시티, 스마트공장, 창업 등에 클라우드 서비스를 접목하는 올앳클라우드(All@Cloud)를 보급·확산하여 향후 5년간 중소·벤처기업 1만 개 이상에게 클라우드 도입을 지원할 계획이다.

양질의 데이터를 쉽게 거래할 수 있는 데이터 유통 기반은 2018년부터 구축해 나간다. 누구나 데이터를 한 곳에서 쉽고 빠르게 등록·검색·거래가 가능한 기반 구축을 위해 ①기존의 민간·공공 데이터 포털을 개방형(CKAN) 플랫폼으로 고도화 ②국내외 주요 데이터 플랫폼 연계 ③데이터 가공 전문기업과 수요기업 매칭 ④시장 활성화를 위한 중소·스타트업에 데이터 바우처(구매·가공비용 등 지원) 제공 ⑤바우처를 통해 가공된 데이터의 상품 등록·공유로 이어지는 데이터 유통 선순환 체계를 마련한다(그림 2-1-3).

[그림 2-1-3] 데이터 거래 기반 구축 추진방안



또한 데이터 거래의 투명성 확보, 공정경쟁 환경조성, 지재권 보호·활용 촉진 등을 위한 제도연구와 정비도 동시에 이뤄진다.

다. 데이터 분석·활용 확산

데이터를 활용한 산업적·사회적 가치를 본격 창출해 산업 전반에 성장 활력을 제고하고 국민 삶의 질 향상에 박차를 가한다. 전통 중소기업의 혁신 돌파구 마련을 위해 신제품 개발, 맞춤형 홍보 등에 데이터를 활용하도록 향후 5년간 500개 사에 빅데이터 분석 전문기업을 매칭·지원한다. 특히 바이오·미래소재·대형연구장비 등 연구 데이터 집약형 분야에서 산업적 활용 촉진을 위해 금년부터 선도 프로젝트를 실시한다. 아울러 미세먼지, 재난안전 등 국민적 관심이 큰 사회문제 해결을 위해 빅데이터 플래그십 프로젝트 투자를 대폭 확대한다.

3. 글로벌 데이터산업 육성기반 조성

가. 빅데이터 선도기술 조기 확보

2022년까지 선진국 기술수준 대비 90% 이상 수준의 빅데이터 선도기술을 확보한다. 이를 위해 빅데이터 기술을 △1단계 분석중심(오픈데이터 플랫폼, 엣지분석기술, 공간 빅데이터 분석기술 등) △2단계 인공지능중심(상황 인지형 데이터관리, 초대규모 모사현실 기술 등) △3단계 초연결 지능화중심(초연결 데이터 관리, 지능형 집단협업 플랫폼 등)에 맞춘 핵심 원천기술 개발을 추진한다. 또 4차산업혁명 요소기술(인공지능, 사물인터넷, 클라우드, 정보보호)과의 융합 선도기술 개발을 가속화한다.

나. 미래수요 대응 전문인력 확충

데이터 산업의 미래수요에 대응해 향후 5년간 청년 고급인재, 실무인력 중심으로 5만 명의 전문인력을 양성한다. 청년일자리와 연계해 빅데이터 전문교육 프로그램을 지원하고, 데이터 분석 고급인재 양성을 위해 대학원 전공, 연구센터 운영을 확대해 나갈 계획이다. 또한 산업 수요기반 실무인력 양성을 위해 국가공인 데이터 자격제도는 내년부터 데이터 분석 국가기술자격제도(빅데이터 분석기사)를 신설하여 확대·운영한다. 아울러 구글의 캐글(Kaggle) 방식을 벤치마킹해 학계, 스타트

[그림 2-1-4] 빅데이터 및 관련 기술 융합 추진

업, 연구기관 등이 참여해 데이터를 연계 분석해 새로운 가치를 찾아내는 데이터 인재발굴 플랫폼도 구축해 운영할 계획이다.

다. 빅데이터 전문기업 성장 지원

빅데이터 전문기업의 성장에 필요한 컴퓨팅 자원, 맞춤형 사업 등을 밀착 지원해 데이터 강소기업 100개 사를 육성할 계획이다. 또한 판교 글로벌 ICT 혁신 클러스터를 글로벌 수준으로 고도화하고 스타트업들의 활용을 집중 지원한다. 특히 데이터 분야 스타트업·강소기업을 중점 발굴·육성하고자 K-Global DB-Stars 등 데이터 스타트업에 대한 맞춤형 지원사업을 지속 확대하고 민관이 협업할 수 있는 공공데이터 활용 창업 콜라보 프로젝트를 확대·발굴할 예정이다. 또한 국내 데이터 기업의 솔루션에 대해 진출 가능성이 높은 국가를 선정해 수요기업의 솔루션 현지화, 마케팅, 해외전시회 참가 등을 지원하는 해외진출 지원사업도 더욱 확대할 계획이다.

제2장

데이터 거래 기반 지원 정책

현재 과학기술정보통신부에서 추진하고 있는 데이터 거래 시장 활성화를 위한 정책은 데이터 유통 플랫폼 운영을 통한 데이터 거래의 장 마련과 이를 중심으로 데이터 구매, 데이터 분석·연계 지원 등 개별 사업들로 이뤄져 있다. 본 장에서는 국내외 데이터 거래 시장 현황을 살펴보고 현행 정책 내용을 소개한다. 아울러 향후 데이터 거래 기반 지원 정책 방향을 제시한다.

데이터의 생산과 소비를 이어주는 데이터의 유통·거래는 데이터를 융·복합해 새로운 가치를 창출하는 전제 조건으로서 데이터산업의 핵심 부문이다. 그러나 국내 데이터 유통·거래 시장은 다소 부진해 오히려 데이터산업 활성화의 걸림돌로 작용하고 있다. 이에 정부는 데이터가 안전하고 원활하게 유통·거래될 수 있도록 다양한 지원 정책을 펼치고 있다.

1. 국내외 데이터 거래 시장 현황

가. 국내 현황

국내 데이터 거래 시장은 2017년 기준 2,409억 원 규모로 파악된다.[1] 국내 데이터 거래 유형은 크게 플랫폼을 통한 거래와 개별 거래로 구분할 수 있다.

플랫폼을 통한 거래는 무료 개방과 유료 거래로 구분된다. 한국정보화진흥원(NIA)의 공공데이터포털을 비롯해 지방자치단체 등에서 데이터를 무료로 개방하고 있다. 한국데이터진흥원에서 운영하는 '데이터스토어'는 범용 데이터포털로 공공·민간데이터의 등록·중개·판매·거래를 지원한다. 이밖에 민간에서 운영하는 데이터 플랫폼은 자사의 데이터와 외부 데이터를 상품화해 제공하는데 SK텔레콤의 빅데이터허브(Bigdatahub), LGCNS의 오디피아(ODPia), KTH의 APIStore 등이 있다.

개별 거래는 데이터 활용이 비교적 활발한 통신, 금융, 기업정보 분야 등에서 판매자와 수요자 간 직접 구매 계약을 통해 이뤄진다. 그러나 개별 거래라 하더라도 데이터 가공·분석을 통한 컨설팅 매출이 대다수를 자치하고 있어 데이터 자체 거래는 미흡한 실정이다.

나. 해외 현황

미국은 개인정보 활용이 비교적 용이한 환경 하에서 1,500억 달러에 이르는 세계 최대 규모의 데이터 브로커 시장[2]이 형성돼 있다. 액시엄(Acxiom), 엡실론(Epsilon)과 같은 데이터 브로커 기업은 각종 소비자의 고객데이터를 수집·가공·분석해 금융·유통회사 등에 적합한 맞춤형 마케팅 서비스를 제공한다.

중국은 정부의 빅데이터산업발전 클러스터 구축 계획[3]에 의거, 공공데이터

[그림 2-2-1] 중국의 구이양 빅데이터 거래소(GBDEX)

출처: http://trade.gbdex.com, 2018.08.20. 접속

1 한국데이터진흥원, 「2017년 데이터산업 현황조사」, 2018.
2 미국 Jay Rockefeller report: A Review of the Data Broker Industry - Collection, Use, and Sale of Consumer Data for Marketing Purposes, 2013.
3 중국 구이저우성,「빅데이터산업 응용계획요강 3단계 발전계획(贵州省大数 据产业应用规划纲要, 2014-2020年)」, 2014.

• 필자 | **이재진**(한국데이터진흥원 실장)

가공·판매를 위해 공업신식화부 비준을 받아 2015년 '구이양 빅데이터 거래소'를 설립해 운영중이다. 설립 이후 2018년 5월 기준으로 약 3억 위안(약 505억 원)의 거래가 이뤄졌다. 데이터 유통 플랫폼(GBDEX) 운영을 통한 공공·민간데이터의 중개·거래뿐 아니라 데이터 가공, 가치평가 등의 서비스를 제공한다.

일본은 2016년 4차산업혁명에 대응하기 위한 범정부 차원의 7대 추진전략을 발표하고 이중 데이터 활용 촉진을 위한 환경정비 전략의 일환으로 2020년까지 IoT 빅데이터 판매·유통시스템을 구축하기로 했다. 이는 로봇, 공작기계 등 강점을 가진 분야와 IoT 데이터를 접목해 시장 주도권을 차지하려는 전략이다. NTT, 히타치, 도쿄전력 등 민간 대기업 100개 사가 참여하며, 준비위원회가 인터넷 정보 거래기준 마련 등 각종 준비 작업을 담당한다.[4]

2. 데이터 거래 특성 및 장애요인

가. 데이터 거래의 특성

데이터의 대표적인 특징은 복제가능성과 가격차별성이다. 데이터는 무형자산으로 일반 재화와는 달리 일단 확보되면 소진되지 않고 재활용할 수 있다. 이는 데이터 복제가 가능해 동시에 여러 사람이, 반복적으로 이용할 수 있기 때문이다. 자칫 데이터 소유자가 아닌 타인이 불법적으로 데이터를 사용하기 용이하다.

데이터를 융합·가공하면 새로운 가치를 창출할 수 있으며, 동일한 데이터도 활용 목적·범위, 지불 능력 등에 따라 차별적인 가격을 부과할 수 있다.

나. 데이터 거래의 장애요인

데이터 거래가 성립되기 위해서는 좋은 품질의 가치 있는 데이터가 공급되고, 적절한 가격의 데이터 구매력이 확보돼야 한다. 아울러 데이터 소유권·개인정보 등 법적 문제 해소가 필요하다.

그러나 실제 데이터 사업자들은 데이터 거래가 원활하지 못한 장애요인으로 데이터를 무료로 사용하려는 이용자 인식(28%), 데이터 상품 검색의 어려움(24%), 개인정보 및 법·제도 문제(20%), 거래방법 정보 부재(14%) 등을 지적하고 있다.[5]

특히 공급자 입장에서는 데이터를 안전하고 원활하게 제공할 수 있는 장(場)을 원하고 있고, 수요자 입장에서는 원하는 데이터를 구하려 해도 어디에 있는지 모르거나 원하는 형태가 아니어서 가공·정제하는 데 상당한 비용과 시간이 소모됨을 어려움으로 호소하고 있다. 아울러 자기결정권에 따른 개인정보 거래에 있어서도 여전히 개인정보를 활용한다는 것 자체가 부정적이고, 활용하더라도 실제 본인에게 어떤 혜택이 있는지 알 수 없다는 인식이 지배적이다.

3. 데이터 거래 활성화를 위한 데이터 거래 기반 지원 정책

정부에서 2018년 6월에 발표한 '데이터산업 활성화 전략'[6]에는 데이터 거래 활성화를 위한 관련 정책들이 포함돼 있다. 이들은 크게 데이터 거래 기반이 되는 데이터 유통 플랫폼과 거래 활성화를 위한 지원 기능으로 구분해 볼 수 있다. 실제 추진중이거나 내년 중 추진될 주요 정책을 살펴보면 다음과 같다.

가. 데이터 유통 플랫폼 운영 및 플랫폼간 연계 지원

2018년 4월 기존 데이터스토어(www.datastore.or.kr)를 오픈소스 데이터 플랫폼인 CKAN(Comprehensive Knowledge Archive Network) 기반으로 전환해 공공·민간데이터의 중개 외에도 데이터 검색, 등록, 관리, 판매, 융·복합, 연계, 활용이 가능한 개방형 데이터 유통 플랫폼으로 고도화했다.

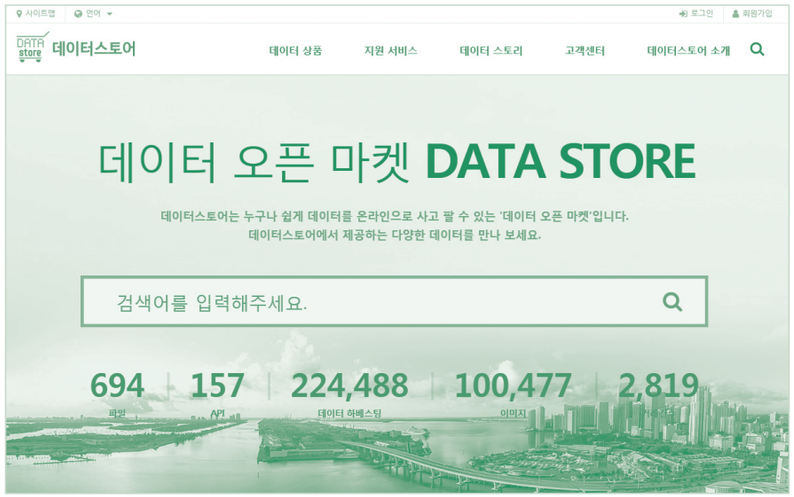
개방형 데이터스토어는 특히 중소·스타트업을 위한 온라인 데이터 중개·거래의 장을 제공하고, 이러한 데이터스토어를 중심으로 국내외 주요 데이터 플랫폼을 연계해 전 세계의 데이터를 쉽고 빠르게 검색·연계·활용할 수 있도록 지원한다. 2018년 8월 현재 ODPia(LGCNS), APIStore(KTH) 등 국내 민간데이터 플랫폼뿐 아니라 CKAN의 데이터 하베스팅(harvesting) 기술을 이용해 영국 공공데이터 포털(data.gov.uk) 등 해외 공공데이터 플랫폼 메타데이터 연계를 완료해 이들 플

4 서울경제, "일본, 2020년까지 세계 첫 IoT데이터 거래소 만든다", 2017. 5.23. 보도

5 한국데이터진흥원, 「2017년 데이터 활용 및 사업화 지원 사업 결과보고서」, 2017.

6 관계부처 합동, 「데이터산업 활성화 전략」, 2018.6.26. 본 백서 제2부 제1장에서 소개

[그림 2-2-2] 데이터스토어



출처: http://www.datastore.or.kr, 2018.08.20. 접속

랫폼에 있는 데이터도 통합 검색할 수 있다.

나. 데이터 바우처 지원

2016년부터 고가의 유료 데이터 활용이 어려운 중소기업, 스타트업 등의 데이터 구매·활용을 지원하기 위해 데이터 구매 바우처 사업을 추진하고 있다. 필요한 데이터가 유료인 경우, 구매 비용의 일부를 매칭 방식의 바우처로 제공해 중소·스타트업의 데이터 서비스 개발에 필요한 초기 데이터 확보를 지원하고 점진적인 유료 데이터 시장 확대를 유도한다. 아울러 보유하고 있는 데이터가 거래·유통될 수 있도록 데이터 상품화도 지원한다. 올해는 70여 개 중소·스타트업을 대상으로 데이터 바우처를 통해 서비스 개발 및 상품화를 지원하고 있다.

2019년부터는 데이터 활용에 필요한 데이터 구매뿐만 아니라, 활용 목적에 따른 맞춤형 데이터 가공 비용을 바우처로 지원해 데이터를 활용하는 중소·스타트업의 데이터 비즈니스와 데이터 유통을 촉진하고 데이터 가공 기업 발굴에 따른 시장 창출을 도모한다.

다. 데이터 안심구역 구축·운영

2019년부터 비식별 처리된 데이터를 물리적 보안 환경이 구축된 안전한 공간에 저장하고, 이용자들이 이를 자유롭게 분석해 결과값만 반출하는 환경인 데이터 안심구역을 구축·운영할 예정이다. 공공기관 및 민간기업의 데이터가 선별적으로 저장되며, 이용자는 안심구역에서 다양한 분석도구를 이용해 데이터를 활용할 수 있게 된다. 이를 위해 올해는 시범적으로 통계청 빅데이터센터와 연계해 제한된 구역에서 통계·학술 목적 하에 미개방된 통계 데이터와 유동인구 데이터 등 민간데이터를 함께 분석해 결과값을 살펴볼 수 있는 '데이터 프리존'을 부산에 개소해 운영하고 있다.

라. 데이터 공정거래 지원

데이터 거래에 필수적인 데이터의 신뢰성을 확보하고 합리적인 데이터 거래가 이뤄질 수 있도록 데이터 거래 방법과 절차 정립 등을 지원하고, 데이터의 가치 제고에 따른 적정 가격 산정을 위한 가이드라인을 개발·보급한다. 아울러 데이터스토어의 부가서비스 형태로 데이터 거래 시 발생 가능한 소유권, 저작권, 계약관계 등 다양한 법률 관련 상담을 지원한다.

이밖에 2019년부터 데이터 거래를 지원하는 데이터 품질관리, 데이터 거래 표준, 데이터 가치평가(자산화), 데이터 소유권 등 법적 문제, 신기술 적용 방안 등의 관련 제도 연구를 본격적으로 진행할 계획이다.

4. 데이터 거래 기반 지원 정책 방향

앞서 살펴본 바와 같이 현재 추진중인 데이터 거래 시장 활성화를 위한 정부 정책은 데이터 유통 플랫폼을 통한 데이터 검색·중개·판매, 데이터 구매 지원, 데이터 분석·연계 환경 지원 등 개별 사업 형식으로 수행되고 있다.

그러나 앞으로 개인데이터를 포함한 각종 데이터를 자유롭게 사고 팔 수 있는 시대에 부응하기 위해서는 데이터 거버넌스 차원의 데이터 거래 지원 정책을 마련해 추진할 필요성이 제기되고 있다. 이에 유용한 공공·민간데이터를 대상으로 데이터 검색과 중개·거래를 기본 기능으로 하되 이를 활성화하기 위한 품질관리, 가공, 저장, 분석 환경 등 다양한 지원 기능을 유기적으로 연계해 종합적으로 대응하는 '데이터 거래소' 설립을 검토할 계획이다.

제3장

본인정보MyData[1] 활용 지원

개인데이터의 경제적 가치가 주목을 받으면서 전통적인 개인정보 '보호'의 중심축이 개인정보의 '활용'으로 이동하고 있다. 미국, EU 등은 마이데이터(MyData) 제도를 통해 정보주체인 개인이 개인데이터를 스스로 통제할 수 있고, 개인의 통제 하에 개인데이터를 활용할 수 있는 환경으로 변화하고 있다. 우리나라에서도 마이데이터 제도를 정착시키기 위해 제도·기술 기반 확보, 본인정보 제공 확산, 서비스 및 적용 다양화의 단계별로 지원 사업을 추진해 나갈 예정이다.

1. 개인데이터의 보호와 활용

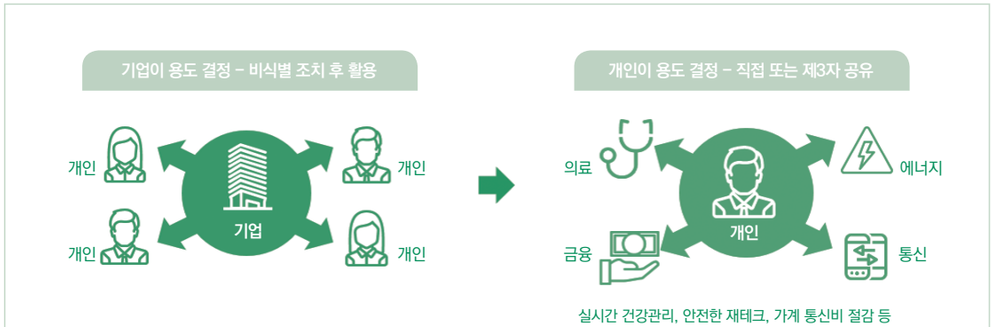
개인데이터의 경제적 가치가 주목을 받으면서, 전통적인 개인정보 '보호'의 중심축이 개인정보의 '활용'으로 이동하고 있다. 파급효과가 큰 의료와 금융 부문을 중심으로 다양한 비즈니스 생태계가 조성되고 있으며, 개인데이터의 수요도 급증하고 있다. 개인데이터는 전체 디지털 정보의 약 75%를 차지하며, 사물인터넷과 AI로 인해 그 비중과 가치가 더욱 커질 전망이다.[2]

개인데이터가 경제에서 창출하는 가치도 해마다 증가하고 있다. 구글, 페이스북, 아마존 등은 개인데이터를 기반으로 성장한 대표적인 글로벌 기업들이다. 2020년까지 EU 지역 내에서 개인데이터로 창출되는 경제적 가치는 약 1조 유로로 EU GDP의 8%를 차지할 것으로 전망했으며, 정부와 기업이 얻는 연간 이익은 3,300억 유로에 달할 것이라고 한다.[3] 미국에서는 개인데이터를 수집, 가공해 판매하는 데이터 브로커가 산업의 한 영역으로 자리잡고 있어, 일부 기업들은 연간 10억 달러 이상의 매출을 기록하기도 한다. 대표적인 데이터 브로커인 액시엄(Acxiom)은 전 세계 7억 명의 고객 정보를 보유 중이다. 수입·교육 수준 등으로 고객 정보를 분류해 다양한 데이터 상품으로 판매하고 있다.

반면 우리나라는 그간 「개인정보보호법」을 근거로 보호 중심의 정책을 견지해 왔다. 그럼에도 개인데이터 유출 사고가 빈번히 발생했다. 시민단체 및 국민들은 개인 정보의 보호를 강력히 요구했다. 개인 정보를 비식별화해 활용하는 경우에도 결과의 정확성, 비식별 조치의 적정성, 시간과 비용 문제 등 여전히 해결해야 할 문제점들이 많다.

이러한 여러 문제들을 해소하려면 마이데이터 제도를 통해 정보주체인 개인은 개인데이터를 스스로 통제할 수 있어야 하며, 개인의 통제 하에 개인데이터를 활용할 수 있는 환경으로 변화할 필요가 있다(그림 2-3-1 참조).[4] 개인데이터의 활용 수요와 잠재적 가치가 매우 큰 만큼, 개인데이터의 활용과 보호 간의 적절한 조화가 필요한 시점이다.

[그림 2-3-1] 개인데이터 활용 방식의 변화



2. 국가별 개인데이터 활용 정책 현황

가. 미국 스마트 공시

오바마 행정부는 2011년 에너지, 건강(의료), 교육 등의 정보를 내려받아(down-

• 필자 | **임태훈**(한국데이터진흥원 책임연구원)

1 본 장에서 '본인정보', '개인정보', '개인데이터'는 '개인과 관련 있는 모든 정보와 데이터'이며, 마이데이터(MyData)는 '본인정보(개인데이터)를 내려 받거나 제3자에게 제공하도록 하는 정책'을 말한다.

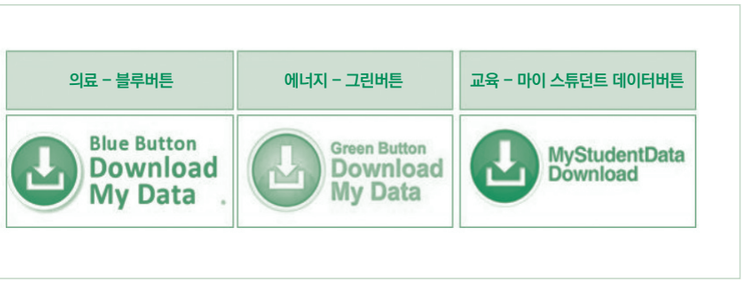
2 IDC, Digital Universe Study, 2011.1.

3 John Rose, et al. The value of our digital identity, BCG, 2012.11.,및 Vasudha Thirani, Arvind Gupta, The Value of Data, World Economic Forum, 22 Sep, 2017. 참조. https://www.weforum.org/agenda/2017/09/the-value-of-data/ (2018.3.6. 접속)

4 World Economic Forum, Unlocking the value of Personal Data-From Collecting To Usage, 2013.02.

load) 직접 활용하거나 제3자와 공유할 수 있도록 스마트 공시(Smart Disclosure) 제도를 추진했다. 이는 소비자가 더 나은 의사 결정을 내릴 수 있도록 개인데이터를 제공하고, 데이터의 투명성을 확보하며, 소비자의 데이터에 대한 접근성과 활용성을 강화하기 위한 조치였다. 소비자가 개인데이터에 쉽게 접근할 수 있도록 분야별 개인데이터 내려받기 단추(버튼)를 통일하고(그림 2-3-2 참조), 웹사이트에서 이를 클릭해 개인데이터를 내려받을 수 있도록 지원했다.

[그림 2-3-2] 미국 스마트 공시 마이데이터 버튼



스마트 공시 제도는 분야별 법률을 근거로 개인데이터에 대한 접근권 및 이동권을 보장하고 있으며, 다양한 부가 서비스들이 개발돼 국민들이 이를 이용하고 있다. 의료분야의 경우, 「건강보험의 이동 및 책임에 관한 법률(Health Insurance Portability and Accountability Act: HIPPA)」의 '개인의 의료 데이터 사본 접근 권한' 조항을 그 근거로 하고 있다. 교육 분야는 「가족 교육 및 프라이버시법(Family Educational Rights and Privacy Act, FERPA)」의 '학부모와 학생의 학적 기록에 대한 접근 및 사본 요청 권한'이 근거 조항이다. 앱 경진대회 등을 통해 개인데이터를 활용하는 기업들의 참여가 확대되고 있으며, 이용자도 꾸준히 증가하고 있다.

나. 영국 midata

영국 정부는 2011년 기업이 보유한 소비자 데이터를 다시 소비자들에게 제공하도록 하는 'Better Choices, Better Deals'의 세부 추진과제로 'midata' 제도를 추진했다. 개인의 기본권인 개인정보 자기결정권 차원에서, 시장에서의 소비자 주권 강화를 정책의 최우선 목표로 삼고 있다. 에너지, 모바일 통신, 신용카드 등 가격 경쟁이 심하고 가계 지출의 중요 부분을 차지하는 시장을 중점 부문으로 선정했으며, 정부-기업-소비자(단체) 간 동반자 관계의 자발적 참여 프로그램을 시행했다. 2013년 4월, 「Enterprise and Regulatory Reform Act」를 개정해, 이용자가 이용 내역을 디지털 포맷으로 내려받을 수 있는 법적 근거를 마련했다. midata 프로젝트의 일환으로 개인데이터 관리 공유 플랫폼 'Mydex'를 구축해 개인데이터의 저장과 분석, 제3의 서비스 기업과의 공유와 서비스 연결을 가능하게 했다.

다. EU 개인정보보호 일반규정

EU는 개인의 개인정보 보호권을 강화하고 EU 회원국 간의 자유로운 데이터 이동을 보장하기 위해 개인정보보호 일반 규정(GDPR, General Data Protection Regulation)을 제정했다. 이 규정은 정보주체 보호 강화, 기업의 책임 강화, 디지털 경제 활성화를 위한 개인정보 보호와 활용 간의 균형 유지를 목표로 하고 있다. GDPR에는 삭제권 및 잊혀질 권리, 개인데이터 이동권, 자동화된 결정 등 개인의 데이터 통제권과 권리를 강화하기 위한 조항이 포함돼 있다. 특히 기업이 보유하고 있는 개인데이터를 새로운 사업자에게 이전시킬 수 있는 데이터 이동권(Right to data portability)[5]을 제20조에 명시했다. 이 조항에 따라 기술적으로 가능한 경우, 개인은 본인 또는 다른 기업에게 직접 정보를 전자적으로 전송하도록 데이터 보유 기관에 요청할 수 있다.

라. 프랑스 Mes Infos

프랑스는 2012년 정부 싱크탱크인 FING(Fondation Internet Nouvelle Génération)와 민간 비즈니스 클러스터인 'Cap Digital'이 공동으로 참여해 'Mes Infos' 프로젝트를 시작했다. 기업은 개인데이터를 활용해 막대한 이익을 얻고 있으나, 개인은 그 이익을 향유하지 못하므로, 개인·조직·사회 간 편익을 공평하게 분배하자는 취지였다. 2016년 10월에는 EU GDPR의 후속조치로 「Digital Republic Act」를 제정해 데이터 등 디지털 재원의 인프라적 가치를 강조하고, 데이터 이동권을 보장했다. 미국의 스마트 공시의 버튼을 본 떠 데이터 이동권 공통 프레임워크와 가이드라인을 정의하는 '레인보우 프로젝트'를 통해 데이터 이동 시스템 명세와 설계 요소

5 PrivazyPlan, Article 20 EU GDPR Right to data portability, http://www.privacy-regulation.eu/en/article-20-right-to-data-portability-GDPR.htm (2018.7.16. 접속)

를 공유했다.[6]

3. 국내 개인데이터 활용 현황

분야별로 개인데이터 활용 범위와 수준이 다르지만, 한국은 개인정보보호에 중점을 둬 왔기 때문에 개인의 자기정보결정권을 적극적으로 행사하기 어려운 상황이다. 의료, 금융, 통신, 에너지, 유통 등 개인 생활과 밀접한 분야들을 살펴본 결과, 개인데이터를 보유한 기관 또는 기업이 개인데이터에 대한 실질적인 통제권을 갖고 있기 때문에 데이터를 이용한 서비스 시장에 독점이 발생하고, 개인은 자신의 데이터가 어디서, 어떻게 사용되는지 관리하기 어려운 상황이다.

의료분야의 경우, 대형병원을 중심으로 앱으로 진료정보를 열람할 수 있도록 서비스 중이며, 정보유출을 우려해 진료 정보 내려받기나 제3자 제공은 매우 제한적으로 실시하고 있다. 분당서울대병원은 앱 '헬스포유(Health4U)'를 통해 진료 내역을 내려 받을 수 있도록 하고 있다. 다른 건강 앱과 정보 연동도 가능하다. 건강보험공단은 웹사이트와 앱 '건강인(건강in)'을 통해 건강 검진 내역을 XML 파일로 내려받을 수 있도록 서비스 중이다.

금융분야는 웹 스크래핑(Web Scraping) API를 통해 이체내역 등 금융정보의 열람, 내려받기, 제3자 제공이 가능하지만, 참여 기관과 대상 데이터가 한정적이다. 핀테크 기업들이 금융 데이터를 활용할 수 있는 API 방식의 오픈 플랫폼이 구축돼 있으나 제공하는 데이터 항목이 다양하지 않고, 요금이 비싸 이용은 저조한 편이다.

개인은 통화량, 데이터 사용량, 문자 사용량 등을 통신사 홈페이지에서 조회할 수 있으나 이를 내려받거나 다른 서비스에서 활용하는 것은 매우 제한적이다. 통신사마다 홈페이지에서 개인별 요금제 추천 서비스를 제공하고 있으나 자사 요금제만 추천하고 있다. 요금 정보포털 '스마트 초이스', '알뜰폰 허브' 등 비교 서비스가 운영 중이나 통신사별 요금제가 다양해 정확한 비교가 어려운 실정이다.

6 MESINFOS, RAINBOW BUTTON. project Enabling personal data portability and turning it into an opportunity for innovation and customer relationship executive summary, 2017.

가정의 전력 사용량에 대한 데이터도 현재 고지서나 인터넷을 통해 열람할 수 있으나 이를 내려받거나 제3자에게 제공하기는 어려운 상황이다. 한국전력은 전력량 데이터 오픈 플랫폼을 기획하는 등 한국형 그린버튼을 만들기 위해 노력 중이나, 주택 명의자 변경에 따른 개인 식별, 아파트나 상가 등의 전력량 측정 주체인 관리사무소와의 관계, 스마트 미터기 보급률 저조 등의 숙제가 남아있다.

유통 분야 온라인 쇼핑몰 구매 내역 내려받기 서비스 현황을 살펴본 결과, 아마존(amazon)이나 이베이(eBay) 등 해외 쇼핑몰의 경우, 주문 구매 이력을 파일로 내려받을 수 있으나, 국내 쇼핑몰은 일부를 제외하고 해당 서비스를 제공하는 경우가 거의 없다.

4. '본인정보MyData 활용 지원 사업' 추진 계획

국민들의 '개인정보 자기결정권'에 대한 인식이 성숙해, 개인 동의 하의 데이터를 공유하고 활용하도록 허용하겠다는 의견이 해마다 늘어나고 있다[7]. 이러한 사회적 공감대 확산에 힘입어, 정부는 '본인정보MyData 활용 지원 사업'을 추진하고 있다.

국내에 마이데이터 제도를 정착시키기 위해, 2019년부터 ① 제도·기술 기반 확보 ② 본인정보 제공 확산 ③ 서비스 및 적용 다양화라는 단계로 지원 사업을 추진해 나갈 예정이다. 국민 실생활과 밀접한 의료, 금융, 통신, 에너지, 유통의 5개 분야를 지원 대상으로 한다.

첫째, 본인정보 활용 지원 기반 확보를 위해서 기관과 기업, 국민을 대상으로 본인정보 활용 현황 및 수요·인식 조사를 실시해, 향후 정책 추진 방향에 반영할 예정이다. 아울러 분야별 민관 협의체를 구성해 기관과 기업이 적용해야 할 데이터 표준, 마이데이터 가이드라인의 제정, 인센티브 마련 방안, 분야별 세부 로드맵, 민관 협력 방안 등을 논의하도록 지원한다.

7 한 조사 결과를 따르면, 개인 동의 후 건강 데이터를 활용하도록 제공하겠다는 의견이 2016년 42.2%에서 2017년 76.3%로 크게 증가했다(전자신문, 국민 10명 중 8명 건강정보 주인은 '나', 의료 빅데이터 제공 'OK', 2018.3.1.)

[그림 2-3-3] 본인정보 활용 지원 사업의 주요 내용

둘째, 본인정보 제공 확산을 위해 기관과 기업이 보유하고 있는 본인정보를 정보 주체인 국민들이 직접 내려받아 활용하거나 본인 동의 후 제3의 서비스 기업에게 제공할 수 있도록 기관과 기업을 대상으로 시스템 구축 및 참여 인센티브 등을 지원한다. 또한 개인데이터 보유 기관 및 활용 기업을 대상으로 기술이나 정책, 보안, 관리에 대한 컨설팅을 지원한다.

셋째, 본인정보 서비스 다양화를 위해 해커톤, 아이디어 공모전 등을 개최해 본인정보 활용 비즈니스 모델 개발 및 개인 맞춤형 서비스 개발을 지원한다. 기관과 기업을 지원해 내려받기가 가능해진 개인데이터를 주제로 비즈니스 모델 발굴을 위한 해커톤 개최, 본인정보 활용 아이디어 공모전 개최를 통해 관련 산업 생태계를 육성하고자 한다.

제 4 장

국내 의료데이터 활용 정책 동향

의료데이터는 정보의 다양성과 양이 방대한 것이 특징이다. 최근 기술 혁신이 이뤄지면서 의료데이터의 통합, 연계, 분석이 가능해지고 있다. 해외 주요국뿐 아니라 국내에서도 적극적으로 이와 관련된 정책을 수립하고 추진하고 있다. 이 글에서는 국내 의료데이터 정책 사례를 살펴보고, 미국과 일본의 개인 의료데이터 접근성 향상 정책도 소개한다. 이를 통해 국내 의료데이터의 방향성을 제시해본다.

1. 국내 의료데이터 주요 활용 정책 동향

의료데이터는 병원 내에서 생성, 축적되는 진료정보, 개인의 생체정보, 유전분석 정보 등을 말하는데 그 양이 방대할 뿐 아니라 종류도 다양하다. 최근 기술 발전으로 의료데이터의 통합, 연계, 분석이 가능해지면서 해외 주요 국가뿐 아니라 국내에서도 적극적으로 관련 정책을 수립, 추진하고 있다. 의료데이터 활용을 높이기 위해 한국 정부는 다음과 같은 점에 초점을 맞추고 있다. 첫째 의료서비스 질 향상을 위한 의료기관 간의 진료정보 교류, 둘째 의료기관과 연구기관 중심의 선도적인 의학연구 수행, 셋째 안전한 데이터 보호 기술을 기반한 민간기업에서의 데이터 활용이다.

최근 보건복지부는 보건의료 정보화를 위한 진료정보 교류 기반 사업과 보건의료 빅데이터 플랫폼 사업을 추진하고 있다. 산업통상자원부는 분산형 바이오 빅데이터 모델을 수립, 진행하고 있다. 세계적으로 개인 의료데이터 활용 방식과 관련해 많은 논의가 오가고 있다. 미국, 영국, 일본 등은 개인 의료데이터의 주도권과 통제권은 개인이 갖고 있어야 한다는 전제하에 활용 정책을 수립하고 있다. 이 글에서는 미국과 일본의 정책을 소개하고, 개인 의료데이터의 국내 활용 방향을 함께 고민해보려 한다.

• 필자 | **정일영**(과학기술정책연구원 부연구위원)

가. 보건복지부의 진료정보 교류기반 구축 사업

진료정보 교류사업은 환자의 진료정보를 의료기관들이 안전하고 효율적으로 주고받을 수 있는 네트워크 구축을 핵심으로 한다. 이 사업의 목적은 의료서비스의 질 향상에 있다. 현재는 환자 중심이 아니라 의료기관별로 시스템이 구축돼 있어 환자가 의료기관을 옮기면 그 이전의 진료기록은 현재 의료진이 알 수 없다. 환자가 의료기관에 요청해 의료데이터를 가져가더라도 장비나 시스템이 의료기관별로 서로 달라 의료데이터를 제대로 사용할 수 없는 경우가 많다.

보건복지부는 의료기관의 진료정보를 표준화하고, 기관 간 교류가 체계적으로 이뤄질 수 있도록 '보건의료정보화를 위한 진료정보 교류기반 구축 및 활성화' 연구개발 사업을 수행했다. 이 사업은 2014년부터 2017년까지 3년 동안 이뤄졌는데, 진료정보 교류 선순환 생태계 조성, 영상품질 기준 마련, 임상 콘텐츠 모델 적용 평가 및 보급, 보건의료정보보호 가이드라인, 표준기반 진료정보 교류 서비스 생태계 구축 및 진료정보 교류를 위한 성과관리 및 인센티브 체계 효과 분석 연구 등이 주요 내용이다. 이 사업을 수행하는 동안 실제 의료현장에서 종이가 아닌 전자 진료정보 교류가 가능하도록 의료법 제21조 2를 신설했고, 제23조의 2가 개정돼 법적 근거도 마련했다.

[표 2-4-1] 전자적 진료정보 교류의 법적근거

법령	내용
의료법 제21조2 (진료기록의 송부 등)	① 의료인 또는 의료기관의 장은 다른 의료인 또는 의료기관의 장으로부터 제22조 또는 제23조에 따른 진료기록의 내용 확인이나 진료기록의 사본 및 환자의 진료경과에 대한 소견 등을 송부는 전송할 것을 요청받은 경우 해당 환자나 환자 보호자의 동의를 받아 그 요청에 응하여야 한다.
의료법 제23조2 (전자의무기록의 표준화 등)	① 보건복지부장관은 전자의무기록이 효율적이고 통일적으로 관리·활용될 수 있도록 기록의 작성, 관리 및 보존에 필요한 전산정보처리시스템(이하 이 조에서 "전자의무기록시스템"이라 한다), 시설, 장비 및 기록서식 등에 관한 표준을 정하여 고시하고 전자의무기록시스템을 제조·공급하는 자, 의료인 또는 의료기관 개설자에게 그 준수를 권고할 수 있다.

출처: 국가법령정보센터

3년간의 연구개발 사업 이후에도 이 사업은 지속해서 확대되고 있는데, 2018년까지 문서저장소 10개, 거점의료기관 15개, 참여의료기관 2,316개 확대를 목표로 하고 있다. 이를 위해 기반시설을 연말까지 구축해 2022년까지 전국적으로 확대할 계획이다

[표 2-4-2] 진료정보 교류 참여의료기관 현황(2018년 5월 기준)

지역	문서저장소	합계	전체 참여 기관 현황				2018년 신규 추가 현황				
			상종	병원	의원	기타	소계	상종	병원	의원	기타
전체		2,316	15	327	1,973	1	1,012	4	157	850	1
서울	연세의료원	259	2	16	241		129	–	14	115	
	서울성모병원	251	1	7	243		251	1	7	243	
경기	분당서울대병원	208	1	12	195		85	–	10	75	
	한림대성심병원	101	1	26	74		101	1	26	74	
충청	충남대병원	236	1	17	218		26	–	4	22	
전라	전북대병원	111	1	20	90		111	1	20	90	
	전남대병원	141	1	53	87		–	–	–	–	
경상	경북대병원	157	2	65	90		109	–	24	85	
	부산대병원	745	4	101	640		93	–	42	51	
전국	공공저장소	107	1	10	95	1	107	1	10	95	1

출처: 보건복지부 보도자료, 「의료기관 간 진료정보 교류 "22년까지 전국 확대 추진!"」, 2018.5.31.

나. 보건복지부의 보건의료 빅데이터 플랫폼

보건복지부가 추진하는 보건의료 빅데이터 플랫폼 정책 시범사업은 보건의료 데이터를 공공기관이 보유한 빅데이터와 연계해 연구에 활용하는 것을 목적으로 한다. 이 시범사업은 보건복지부에서 2년간(2016~2017년) 보건의료기술연구개발사업으로 수행한 '보건의료 빅데이터 연계플랫폼 구현' 계획에 기반해 진행되고 있다. 이 플랫폼에 적용되는 데이터는 우선 공공성이 강한 4개 분야인 보건의료 분야 정책연구, 의료정보보호 기술연구, 보건의료기술 연구, 건강 관련 학술연구에서만 활용할 수 있도록 한정했다. 4개 분야 연구를 위해 시범사업에서는 관련 공공기관이 보유한 데이터와의 연계와 활용이 이뤄질 예정이다.

[표 2-4-3] 보건의료 빅데이터 플랫폼 연계 데이터 예시

기관	연계 데이터 종류
건강보험공단	가입정보, 건강검진정보, 영유아·암 검진, 요양기관, 장기요양 판정, 장기요양기관 등
건강보험심사평가원	병의원정보, 청구내역, 수가DB, 의약품 처방정보 등
국립암센터	암등록정보, 암 검진 코호트, 암환자 의료비지원 정보
질병관리본부	유전체 정보, 각종 건강조사 정보

출처: 보건복지부, 「보건의료 빅데이터 플랫폼 시범사업 추진계획(안)」, 2017

[표 2-4-4] 보건의료 빅데이터 플랫폼 시범사업 단계별 추진(안) 개요

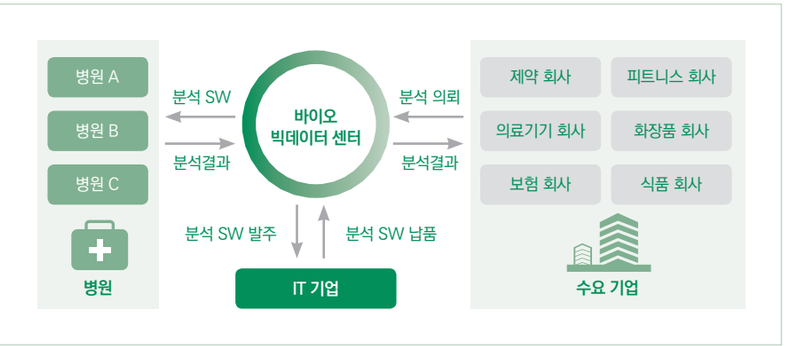
연도	2017년	2018년	2019년	2020년	2021년	2022년
단계	준비단계	시범사업단계		본사업단계		
전략	전략수립	전략추진				
거버넌스	위원회 준비	위원회 운영				
플랫폼	구축준비	정보화 계획	플랫폼 구축 1년차	2년차	본사업 전환, 이용자 집중 등 필요 시 시스템 확충	
사업	시범사업 기획	시범사업 착수, 체계수립		평가	본사업(위원회 심의를 거쳐 추진)	
법	특별법안 마련	특별법 제정 추진				
R&D	보건의료 빅데이터 R&D 지원					

출처: 보건복지부, 「보건의료 빅데이터 플랫폼 시범사업 추진계획(안)」, 2017.

다. 산업통상자원부의 분산형 바이오 빅데이터 모델

의료데이터 활용이 어려운 가장 큰 이유는 데이터에 개인의 민감정보가 포함돼 있기 때문이다. 이런 이유로 의료데이터 활용에는 완벽한 익명화나 가명화 처리가 필요하다. 그 밖에 개인의 의료데이터를 산업적으로 활용하는 것에 대해 아직 사회적 합의가 이뤄지지 않은 문제도 있다. 그 때문에 인공지능, 빅데이터 분석기술, 클라우드 기술 등이 빠르게 발전하고 있지만, 의료데이터는 산업적으로 활용해 가치를 창출하기가 어렵다. 관련 국내 규제 완화와 정책 수립과 실행은 당연히 더 어려운 상황이다. 이를 해결하기 위해 산업통상자원부는 분산형 바이오 빅데이터 모델을 추진하고 있다. 기존에는 각 의료기관의 의료데이터를 기관 밖으로

[그림 2-4-1] 분산형 바이오 빅데이터 모델



출처: 산업통상자원부, 「바이오 빅데이터 구축으로 4차 산업혁명 본격 시동」, 2018.04.17.

내보내 한 곳에서 연계하는 통합형 방식을 채택했다. 이와 달리 분산형 모델은 각 의료기관의 데이터는 기관 내에 있고 의료기관과 기업 간에 분석결과만 거래하는 방식이다.

2. 해외 의료데이터 활용 정책 동향

미국은 2015년부터 정밀의료 이니셔티브(Precision Medicine Initiative)를 통해 의료데이터를 통합·활용하고 있다. 단기적으로는 의료데이터를 이용한 암 유전체 연구를 확장해 암 치료 방법 개발과 치료제 임상시험을 지원할 예정이다. 장기적으로는 종합적인 과학 지식기반을 만들기 위해 과학자 네트워크를 지원하고, 백만 명 이상의 국가 연구 데이터인 코호트를 구축할 예정이다.

백만 명 이상의 코호트 데이터 구축은 정밀의료 이니셔티브에서 가장 중요하게 추진하는 계획이다. 이를 위해 전국의 의료기관과 연구기관에서 보유하고 있는 데이터의 양과 질을 파악하고 연계를 시도하고 있다. 뿐만 아니라 자신의 의료데이터를 기부하는 자발적 데이터 기부자를 모으고 있으며, 기부를 위한 프로그램인 올오브어스 연구프로그램(All of Us Research Program)과 블루버튼 이니셔티브(Bluebutton initiative)를 추진하고 있다.

블루버튼 이니셔티브는 2010년 미국 보훈처에서 시작한 캠페인이다. 보훈처 퇴역군인이 온라인으로 자신의 의료와 건강기록을 다운로드 받아 원하는 의료기

[그림 2-4-2] 개인 의료정보 소유권 인포그래픽

출처: Health IT 홈페이지, HIPAA Right of Access Infographic(www.healthit.gov/sites/default/files/YourHealthInformationYourRights_Infographic-Web.pdf)(2018.07.05. 접속)

관, 약국, 가족 등과 데이터를 공유할 수 있도록 했다. 보훈처 환자 포털에서 블루버튼 서비스가 최초로 제공된 이후 다른 연방기관과 민간부문으로 이 사업은 확대됐다. 현재는 미국 국립보건정보기술조정국(ONC)에서 담당하고 있다. 블루버튼 서비스 확대는 개인 의료데이터 접근성과 소유권이 향상됐다는 점에서 의미가 크다.

블루버튼 이니셔티브의 비전 중 하나는 본인의 데이터를 자발적으로 국가 연구를 위해 기부할 수 있도록 개인의 인식과 행동 변화를 이끈다는 것이다. 자발적으로 기부가 가능하도록 기술적으로 구현하는 프로그램이 올오브어스 연구 프로그램(All of Us Research Program)이다. 의료데이터를 기부하려는 개인은 이 프로그램에 등록해 안전하게 데이터를 기부할 수 있다.

일본은 의료데이터 활용과 관련한 법령을 재정비하고, 데이터를 생성하는 보건의료 기관의 전반적인 시스템과 데이터를 디지털화할 계획이다. 일본 건강·의료 전략에서는 4개 정책을 세우고 있다. 그 중 하나인 '일본 전국 의료데이터 이용·활용 기반시설 구축과 ICT의 이용·활용 추진 관련 정책'에서 디지털 기반 구축과 데이터의 디지털화를 추진하고 있다. 검사데이터(병명, 처방전 등)를 포함해 의료, 건강 분야의 디지털화가 정책의 주요 내용인데, 2020년까지 구축을 목표로 하고 있다. 또한 의료데이터를 효율적으로 활용하기 위해 개인의 전 생애 데이터가 통합되는 플랫폼과 다양한 기관이 활용하는 플랫폼도 구축할 예정이다. 'PeOPLe(Person centered Open Platform for well-being)'는 국민이 언제 어디서나 본인의 건강 관리를 쉽게 할 수 있도록 보건의료 데이터를 정리·통합하는 플랫폼이다. 보건의료기관이 환자 간호와 치료 시 PeOPLe 플랫폼에 접근해 개인의 보건의료 데이터를 확인할 수 있다. 기관의 데이터 플랫폼은 '데이터 이용·활용 플랫폼'으로 각 기관의 다양한 수요에 맞춰 보건의료 데이터를 익명으로 가공·처리한다. 이를 기반으로 빅데이터를 제공하고 각 기관은 보건의료 질 개선과 신약개발, 건강관리 서비스 등에 이를 활용할 수 있다.

3. 국내 의료데이터 활용 전망과 방향성

국내에서도 의료데이터 활용을 위한 다양한 정책과 방안이 나오고 있다. 지금까지는 의료기관과 연구기관 중심으로 논의가 진행됐지만 점차 산업계까지 범위가 확대되고 있다. 이제는 데이터의 주체가 되는 개인을 중심으로 데이터를 어떻게 축적, 통합하고 활용할 것인지 고민할 시점이다. 앞서 소개한 미국의 개인 의료데이터 접근성 향상 사업과 일본의 개인데이터 플랫폼 정책이 올바른 방향성 제시에 도움이 될 것이다.

제5장

국내 금융데이터 활용 정책 동향

금융 분야에서 집적된 데이터는 활용도가 매우 높을 뿐 아니라 타 산업과의 융합이 용이해 경제 전반에 미치는 영향이 크다. 그럼에도 국내 데이터 활용은 주요국에 비해 매우 제한적으로만 이뤄지고 있다. 이에 정부는 금융분야를 빅데이터 테스트베드로서 우선 추진하고 법제도·산업·인프라 측면에서 종합적인 대응방안을 마련하고, 정보 주체의 권리를 내실 있게 보호해 국민신뢰를 제고하겠다는 원칙에 기반해 10대 정책과제를 제시하고 있다. 이에 따라 융합신산업 활성화 및 금융시스템 포용성 강화, 공정한 금융시스템 조성에 따른 긍정적인 거시경제적 효과, 그리고 데이터 활용에 대한 국민적 신뢰가 제고될 것으로 기대된다.

1. 개요

모바일, 클라우드, 사물인터넷, 인공지능, 로봇 등의 기술 발달로 초연결사회가 도래해 개인의 일상과 선호마저 디지털화되는 이른바 '디지털 트랜스포메이션(Digital Transformation)'을 겪고 있다. 다시 말하면 합리적 의사결정을 위해 무의미할 수도 있는 디지털화한 자료에서 유용한 정보를 분석·도출하는 능력이 부가가치 창출의 핵심이 되는 데이터 주도 경제(Data-driven Economy)로의 전환이 빠르게 이뤄지고 있다.

금융 분야에서 집적된 데이터는 양적·질적으로 우수해 활용도가 매우 높을 뿐만 아니라 타 산업과의 융합이 용이해 경제 전반에 미치는 영향이 크다고 할 수 있다. 해외에서는 이미 자동차 운행정보, 통신기록, 인성검사결과, 온라인쇼핑 정보 등을 활용해 중저 신용자에게도 금융 서비스를 제공하는 사례들이 있다. 반면 국내에서는 금융데이터 활용이 매우 제한적으로 이뤄지고 있다.

빅데이터를 활용해 금융적 수용성(Financial Inclusion)을 높이는 것은 금융시장 마찰에 따른 금융증폭효과(Financial Accelerator Effect)[1]를 낮춰 사회적 효용 증대를 야기하는 등의 거시경제적 역할이 크므로 장단기적으로 우리경제에 영향을 주는 중요한 이슈라고 할 수 있다.

따라서 이 글에서는 우리나라 금융 분야 데이터 활용 현황과 정부의 정책 추진방향에 대해 정리하고자 한다.

2. 금융 분야 데이터 활용 현황과 문제점

금융데이터 활용과 가장 맞닿아 있는 국내 신용정보산업은 두 세 개의 기업이 90% 이상의 시장을 점유하고 있는 과점시장이다. 그리고 장기적인 전략을 기반으로 인프라 구축에 대한 투자가 이뤄지기 힘들고, 제도적으로는 데이터 활용이 엄격히 제한되거나 법규정이 불명확해 산업이 정체돼 있다.

국내 빅데이터 활용을 저해하는 요소로 가장 많이 지적되는 것은 다름 아닌 개인정보보호와 관련된 규제들이다. 기본적으로 우리나라 개인정보보호와 관련된 규제 방식은 선택적 동의(opt-in) 방식으로 개인정보 수집·이용·제공 시 혹은 목적변경 시에 반드시 정보 주체의 동의를 받도록 하고 있다. 개개인을 식별할 수 있는 정보는 정보 주체의 동의 없이 제공할 수 없고 비식별화한 정보만 정해진 목적 이외의 용도로 활용이 가능하다. 정부는 2016년 6월 '비식별 조치 가이드라인'을 발표하면서 개인정보를 분석 목적으로 활용할 수 있는 기준을 제시했다. 하지만 가이드라인의 법적 지위가 불명확하고 엄격한 비식별 조치를 요구함에 따라 실제 분석에 필요한 정보를 획득하기 매우 어려운 구조로 돼 있었다. 예를 들면 다음과 같다.

첫째, 비식별화 조치에 대해 전문가로 구성된 평가단이 그 적정성을 판단하게 돼 있었으나 평가단 자체에 대한 신뢰를 확보하기 어려웠다. 둘째, 식별자 또는 다른 정보와 결합하기 쉽도록 식별 가능한 속성자를 삭제하거나 비식별화한 후에야 활용이 가능했다. 그러나 이는 현실적으로 불가능해 대부분의 정보가 삭제된 후 제공될 수밖에 없어 활용 가능한 데이터의 효용성이 크게 저하될 수밖에 없다.

게다가 잇단 개인정보 유출 사태로 인해 금융권의 데이터 활용에 대한 국민

• 필자 | **고동환**(정보통신정책연구원 부연구위원)

1 경제주체들의 금융서비스 수요가 큰 경기침체기에 오히려 서비스가 제한됨에 따라 경기변동이 증폭되는 효과

들의 신뢰가 낮아 주요국의 추세와는 반대로 강한 수준의 정보보호규제를 도입하거나 유지할 수밖에 없는 사회적 분위기가 조성돼 있다. 결과적으로 데이터의 공공재적 특성이 제대로 평가받지 못해 데이터 유통시장이 제대로 형성되지 못했을 뿐 아니라, 전문화한 데이터 산업의 발달도 지체되는 결과를 초래했다.

3. 금융 분야 데이터 활성화 정책 동향

이에 금융위원회는 4차산업혁명의 흐름에 적극적으로 대응해 나가기 위한 소비자 중심 금융혁신 방안의 일환으로 2017년 1월부터 '개인신용평가체계 개선 민관합동 TF'를 구성해 다양한 데이터를 활용한 개인신용평가체계 고도화 방안 등을 포함한 '개인신용평가체계 종합 개선방안'('18.1.30)을 발표했다. 2017년 12월부터는 유관기관, 업계, 전문가 등과 함께 '빅데이터 활성화를 위한 금융분야 TF'를 구성해 금융 분야 빅데이터 활성화를 위한 논의를 집중적으로 실시했다. 한편 대통령 주재 '혁신성장 전략회의'('17.11) 및 '규제혁파 토론회의'('18.1) 등을 통해서도 데이터 기반 핀테크 활성화 방안 등을 함께 논의하고, 이를 대통령 직속 4차산업혁명위원회 '규제·제도 혁신 해커톤' 회의 논의사항에는 금융위원회가 '금융 분야 데이터 활용 및 정보보호 종합방안('18.3.19)'을 통해 금융 분야 데이터 활용 정책의 기본원칙을 수립하고 정책 추진 방향과 과제를 제시하기에 이르렀다.

가. 기본원칙

금융위원회는 먼저 금융 분야 데이터 활성화를 위해 크게 세 가지 원칙을 제시했다. 첫째, 금융 분야를 빅데이터 테스트베드로서 우선 추진한다. 둘째, 법제도·산업·인프라 측면에서 종합적인 대응방안을 마련한다. 셋째, 정보주체의 권리를 내실 있게 보호해 국민의 신뢰를 제고하겠다고 밝혔다. 특히 빅데이터 관련 기술혁신을 유연하게 수용할 수 있도록 원칙 중심의 법·제도로 전환하고 정보가 부족한 핀테크·중소형 금융회사 등으로 데이터가 원활히 흐를 수 있도록 데이터 중개 및 유통 기능을 강화하겠다고 밝혔다. 또한 개인정보 자체를 보호하고자 했던 기존의 접근방식에서 벗어나 개인정보 자기결정권을 보호하는 방식으로 전환해 디지털 환경에서 새롭게 부각되는 권리침해 가능성에 선제적으로 대응하고 정보보호 및 보안의 실효성을 제고하겠다는 원칙을 제시했다.

나. 3대 추진전략 및 10대 추진과제

금융위원회는 또한 소비자 중심의 금융혁신을 촉진하고, 경제·금융시장의 포용성과 공정성을 강화함과 동시에 데이터 활용에 대한 국민적 신뢰를 제고하기 위한 방안으로 3대 추진전략과 10대 추진과제를 공개했다.

1) 3대 추진전략

첫 번째 전략은 데이터 활용을 제약하는 법·제도를 합리화하고 분석 인프라를 지원하는 등 금융 분야 빅데이터를 활성화하는 것이다. 두 번째 전략은 금융 분야 데이터 산업의 경쟁력을 제고해 고부가가치 산업으로 육성함과 동시에 책임성 강화 노력을 병행하는 것이다. 마지막으로는 개인정보 자기결정권을 충분히 보장하고, 정보보호·보안을 강화해 금융권 데이터 활용에 대한 국민 신뢰를 제고하는 것이다.

2) 10대 추진과제

첫 번째 추진전략에 대응하는 구체적 과제로 ①익명정보 및 가명처리정보에 대해 사전동의 등의 정보보호 규제를 완화하고 국제적으로 논의되고 있는 비식별기술 및 개인정보보호 모델의 표준화 논의를 적극 수용해 데이터 활용에 대한 법적 근거를 명확히 하는 것, ②데이터시장 조성을 위해 공공부문에 집중된 금융정보를 DB화해 제공하고, 분석도구 및 보안체계 등이 갖춰진 빅데이터 분석 시스템 및 빅데이터 중개 플랫폼을 구축하는 것, ③신용정보회사(CB, Credit Bureau) 및 카드사가 보유한 정보 및 노하우를 활용해 서비스를 제공할 수 있도록 허용함으로써 빅데이터 시장을 선도할 수 있는 여건을 조성하는 것, ④개인의 부정적 정보 뿐 아니라 세금, 사회보험료 납부실적 등 긍정적 정보도 공유해 CB 및 금융회사의 개인신용평가체계 고도화를 촉진하는 것이 제시됐다.

두 번째 추진전략에 대한 주요 과제로는 ⑤개인 CB업과 기업 CB업의 업무범위를 명확하게 구분해 개인 CB업은 비금융정보 특화 CB사를 설립하고, 기업 CB업은 개인신용정보 규제 필요성이 크지 않으므로 진입규제를 대폭 완화하는 등의 조치로 경쟁을 촉진하는 것, ⑥본인 정보의 효율적 관리를 지원하는 '본인신용정

보관리업'을 도입, '종합 자산관리서비스업'으로 육성해 시장을 확대하는 것, ⑦개인 및 기업 CB사의 지배구조 및 행위를 규제함으로써 책임성 확보를 강화하는 것이 선정됐다.

세 번째 추진전략을 위한 구체적 과제로는 ⑧정보주체의 실질적 동의권을 회복할 수 있도록 정보 활용 동의절차를 단순화하고 내실화해 개인의 선택권을 확대하는 것, ⑨데이터 분석결과에 대해 이의제기 등 정보주체의 적극적 대응권을 강화해 자기정보의 활용권을 보장하는 것, ⑩금융권의 정보 활용 및 관리 실태에 대한 상시평가제를 도입해 데이터 산업의 보안조치 의무를 강화하는 것이 제시됐다.

4. 기대효과

금융 분야 데이터 활용 및 정보보호 방안에 따른 구체적 과제의 성공적인 실천은 데이터 기반 금융혁신을 촉진해 전후방 연관 산업의 발전을 통해 거시경제에 긍정적인 역할을 할 것으로 기대된다. 소비자 중심 금융혁신을 통해 금융시스템의 포용성도 강화돼 기존에 금융시장으로부터 배제됐던 계층까지 금융서비스를 누리게 됨으로써 개인효용 증가는 물론 거시경제 변동성 완화에 따른 사회적 효용도 증가할 것으로 기대된다. 또한 건전한 데이터 생태계 조성을 통해 중소형 금융회사, 핀테크, 창업기업의 시장진입으로 인해 경쟁이 촉진됨으로써 공정한 금융시스템 형성에도 기여할 것으로 기대된다. 마지막으로 다양한 개인정보 자기결정권이 도입되고 정보보호 및 보안이 실질적으로 강화됨에 따라 금융분야에 대한 국민신뢰가 제고될 것으로 기대된다.

[표 2-5-1] 주요과제 및 추진일정

추진과제			추진일정
금융분야 빅데이터 활성화	빅데이터 분석·이용의 법적근거 명확화		'18.上 신정법 개정 추진
	빅데이터 인프라 구축 및 운영	금융정보 DB 분석시스템	'18.下 서비스 실시 및 신정법·보험업법 시행령 개정 추진
		빅데이터 중개플랫폼	'19.上 시범서비스 실시
	CB·카드사의 시장선도 역할 강화	CB사	'18.上 신정법 개정 추진
		카드사	'18.下 카드업계 간담회 후 부수업무 신고 등 추진
	금융회사의 개인신용평가 체계 고도화	금융사	'18.上 관계기관 협의 (필요시 관련법령 등 개정)
		신용정보원 정보공유	'18.上 신정법 개정 추진 및 관계기관 협의
금융분야 데이터산업 경쟁력강화	신용정보산업의 경쟁 촉진		'18.上 신정법 개정 추진
	본인신용정보관리업 도입		'18.上 신정법 개정 추진
	CB산업 책임성 확보		'18.上 신정법 개정 추진
정보보호 내실화	정보제공 동의제도 내실화		'18.下 시행 (필요시 법개정)
	개인정보 자기결정권 강화		'18.上 신정법 개정 추진
	정보활용·관리 상시평가제		'18.上 신정법 시행령 개정 추진 (필요시 법개정)
	정보보호 및 보안 강화	그 밖의 정보보호·보안 강화	'18.上 신정법 개정 추진

출처: 금융위원회, 「금융분야 데이터활용 및 정보보호 종합방안」, 2018.3.19.

제 3 부

데이터산업 시장 현황

제1장

국내 데이터산업 현황

본 장의 국내 데이터산업 현황은 「2017년 데이터산업 현황조사 결과보고서」(과학기술정보통신부, 한국데이터진흥원, 2017.4)를 요약·발췌해 작성했다. 데이터산업 현황조사는 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 데이터 비즈니스를 영위하는 사업체들의 데이터 관련 직간접 매출을 조사해 추정한 결과다. 참고로 데이터산업 현황 조사는 매년 조사 시점 매출액은 예상 매출액으로 산출하고, 전년도 매출액은 확정치를 조사해 반영하고 있다. 이에 본 백서 상의 통계(2017년 조사결과)와 향후 발표될 2018년 데이터산업 현황 조사 결과가 다를 수 있음을 미리 밝힌다.

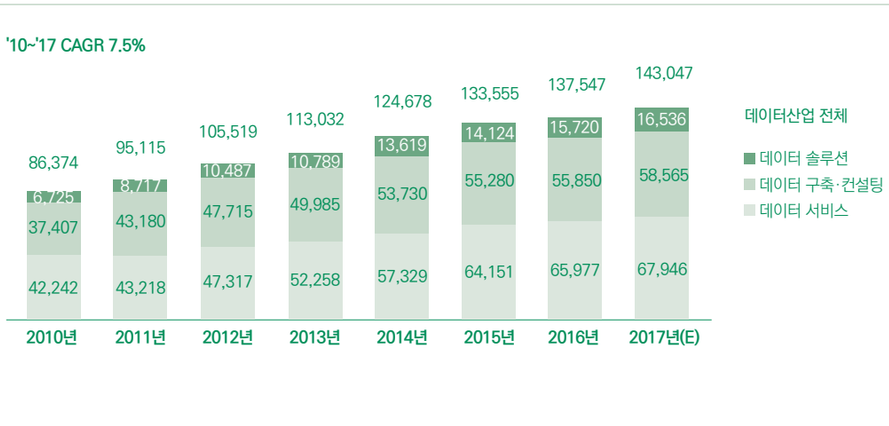
1. 국내 데이터산업 시장 규모

2017년 데이터산업 시장 규모는 14조 3,047억 원으로 2016년 대비 4.0% 성장한 것으로 나타났다. 2010년 이후 연평균증가율은 7.5%로 매년 꾸준한 성장세를 유지하는 것으로 조사됐다.

부문별로는 데이터 솔루션 시장이 1조 6,536억 원, 데이터 구축·컨설팅 시장이 5조 8,565억 원, 데이터 서비스 시장이 6조 7,946억 원으로 나타났다. 부문별로 보면 데이터 서비스 시장이 47.5%로 가장 큰 비중을 차지했고, 이어서 데이터 구축·컨설팅 시장이 40.9%, 데이터 솔루션 시장이 11.6% 순으로 그 뒤를 이었다.

[그림 3-1-1] 데이터산업 시장 규모, 2010~2017(E)

(단위 : 억 원)



출처: 과학기술정보통신부·한국데이터진흥원, 「2017년 데이터산업 현황조사 결과보고서」, 2017.4(이하 동일)

[표 3-1-1] 데이터산업 부문별 시장 규모

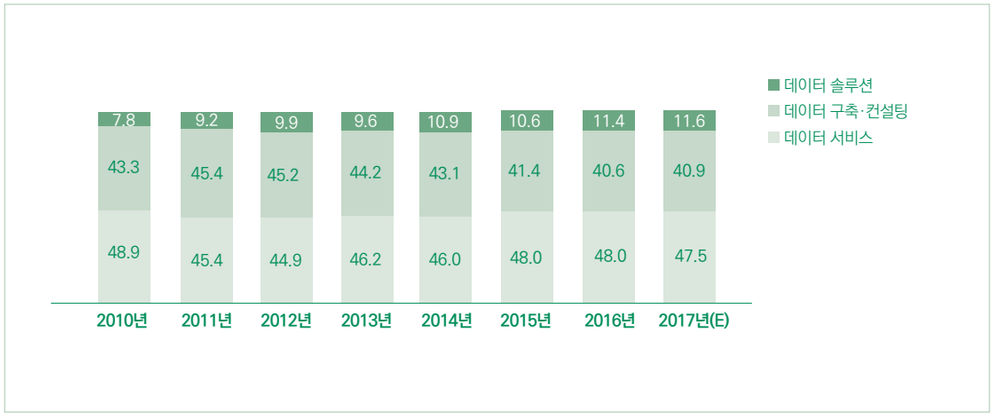
(단위: 억 원, %)

구 분	2010년	2011년	2012년	2013년	2014년	2015년	2016년	2017년(E)	증감률 '16~'17	CAGR '10~'17
데이터 솔루션	6,725	8,717	10,487	10,789	13,619	14,124	15,720	16,536	5.2	13.7
데이터 구축·컨설팅	37,407	43,180	47,715	49,985	53,730	55,280	55,850	58,565	4.9	6.6
데이터 서비스	42,242	43,218	47,317	52,258	57,329	64,151	65,977	67,946	3.0	7.0
전체	86,374	95,115	105,519	113,032	124,678	133,555	137,547	143,047	4.0	7.5

출처: 과학기술정보통신부·한국데이터진흥원, 「2017년 데이터산업 현황조사 결과보고서」, 2017.4(이하 동일)

[그림 3-1-2] 데이터산업 부문별 시장 규모 비중

(단위: %)

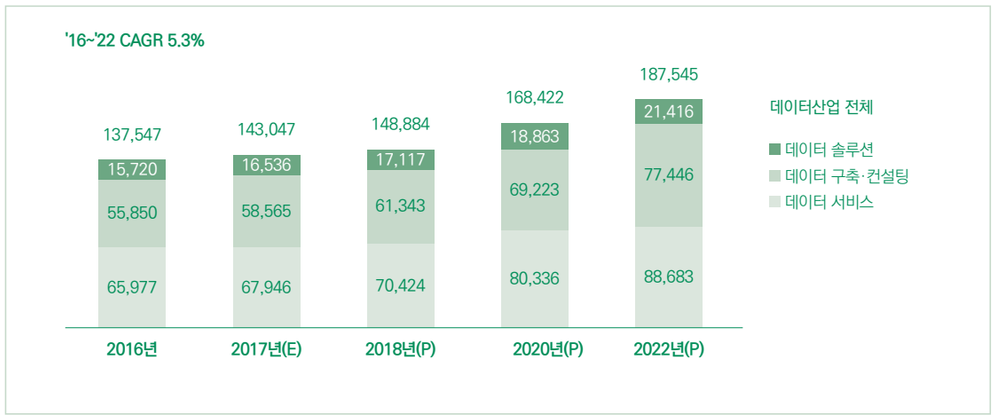


4차산업혁명으로 데이터의 가치가 더욱 부각되고 있으며, 이에 따라 데이터산업은 2022년까지 연평균증가율 5.3% 성장세로 18조 원 대의 시장 규모가 될 것으로 전망된다.

• 필자 | **김초롱**(한국데이터진흥원 선임연구원)

[그림 3-1-3] 데이터산업 시장 전망, 2016~2022(P)

(단위: 억 원)



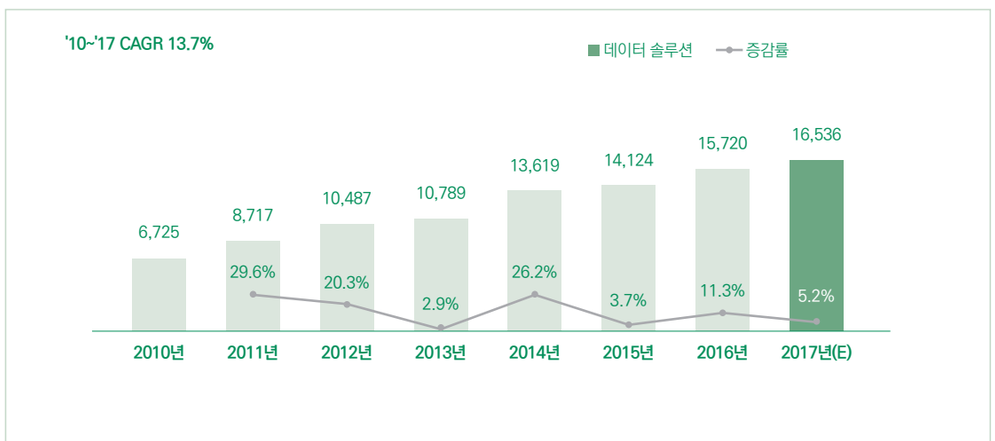
2. 국내 데이터산업 부문별 시장 규모

가. 데이터 솔루션 시장

2017년 데이터 솔루션 시장 규모는 2016년 대비 5.2% 성장한 1조 6,536억 원으로 나타났으며, 연평균성장률(CAGR) 13.7%로 성장하고 있다.

[그림 3-1-4] 데이터 솔루션 시장 규모

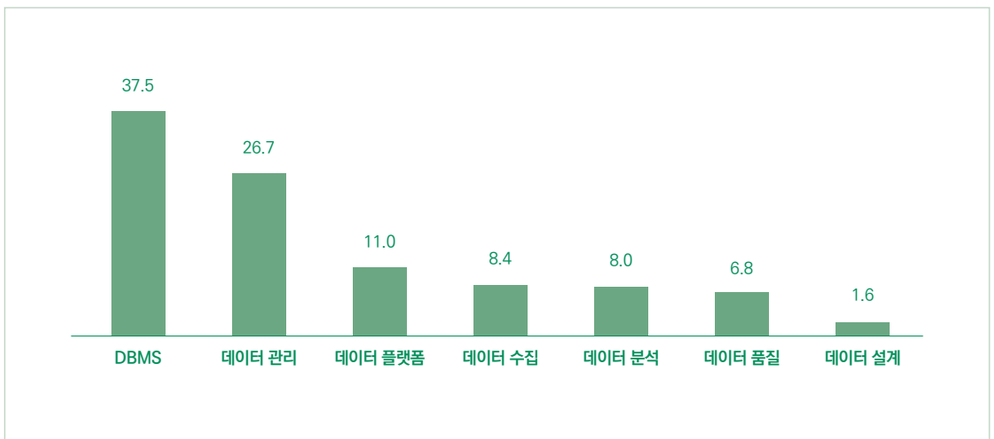
(단위: 억 원)



데이터 솔루션 시장은 크게 데이터 수집, 데이터 설계, DBMS, 데이터 관리, 데이터 품질, 데이터 분석, 데이터 플랫폼이라는 7개 중분류로 구분하고 있다.[1] 이 중 DBMS 분야가 37.5%로 가장 큰 비중을 차지하는 것으로 나타났다.

[그림 3-1-5] 2017년 데이터 솔루션 중분류별 시장 규모 비중

(단위: %)



[표 3-1-2] 데이터 솔루션 중분류별 시장 규모

(단위: 억 원)

구분	2010년	2011년	2012년	2013년	2014년	2015년	2016년	2017년(E)	증감률 '16~'17	CAGR '10~'17
데이터 수집	448	570	617	622	1,076	1,115	1,345	1,382	2.8	17.5
데이터 설계	129	136	151	166	202	207	225	269	19.6	11.1
DBMS	3,169	4,272	5,301	5,488	5,502	5,727	6,148	6,192	0.7	10.0
데이터 관리	1,237	1,516	2,075	2,489	3,446	3,574	4,022	4,417	9.8	19.9
데이터 품질	686	949	942	855	883	918	1,120	1,127	0.6	7.3
데이터 분석	1,056	1,274	1,401	1,169	1,121	1,157	1,249	1,328	6.3	3.3
데이터 플랫폼	-	-	-	-	1,389	1,426	1,611	1,821	13.0	9.4[2]
데이터 솔루션 전체	6,725	8,717	10,487	10,789	13,619	14,124	15,720	16,536	5.2	13.7

1 세부영역별 자세한 내용은 「2017년 데이터산업 현황조사 결과보고서」(과학기술정보통신부, 한국데이터진흥원, 2017.4)의 데이터산업분류체계를 참고하기 바람

2 2014년~2017년 CAGR

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

[표 3-1-3] 데이터 솔루션 영역별 시장 규모

(단위: 억 원, %)

구분		2010년	2011년	2012년	2013년	2014년	2015년	2016년	2017년(E)	증감률 '16~'17	CAGR '10~'17
데이터 수집	라이선스	370	433	477	492	316	225	277	289	4.3	14.6
	개발					507	523	681	673	-1.2	
	유지보수	78	137	140	130	253	367	387	420	8.5	27.2
데이터 설계	라이선스	111	117	130	144	25	33	42	50	19.0	8.1
	개발					141	133	125	142	13.6	
	유지보수	18	19	21	22	36	41	58	77	32.8	23.1
DBMS	라이선스	2,531	3,590	4,718	4,722	2,698	2,712	2,880	2,948	2.4	10.4
	개발					1,987	2,101	2,184	2,127	-2.6	
	유지보수	638	682	583	766	817	914	1,084	1,117	3.0	8.3
데이터 관리	라이선스	953	1,212	1,692	1,987	1,395	363	452	474	4.9	14.5
	개발					1,189	1,796	1,902	1,987	4.5	
	유지보수	281	304	383	502	862	1,415	1,668	1,956	17.3	31.9
데이터 품질	라이선스	481	662	630	587	436	131	179	210	17.3	4.2
	개발					243	354	438	433	-1.1	
	유지보수	205	287	312	268	204	433	503	484	-3.8	13.1
데이터 분석	라이선스	841	1,111	1,319	967	104	116	190	207	8.9	3.1
	개발					821	846	830	836	0.7	
	유지보수	215	163	82	202	196	195	229	285	24.5	4.1
데이터 플랫폼[3]	라이선스	-	-	-	-	391	361	352	432	22.7	3.4
	개발	-	-	-	-	746	823	958	1,068	11.5	12.7
	유지보수	-	-	-	-	252	242	301	321	6.6	8.4
데이터 솔루션 전체	라이선스	5,287	7,125	8,966	8,899	5,365	3,941	4,372	4,610	5.4	12.3
	개발					5,634	6,576	7,118	7,266	2.1	
	유지보수	1,435	1,592	1,521	1,890	2,620	3,607	4,230	4,660	10.2	18.3

3 2014년~2017년 CAGR

데이터 솔루션 시장의 경우 공공분야와 유통·서비스 분야가 가장 큰 수요층인 것으로 나타났으며, 전년 대비 통신·미디어의 비중이 높아졌다.

[표 3-1-4] 데이터 솔루션 업종별 매출 비중

(단위: %)

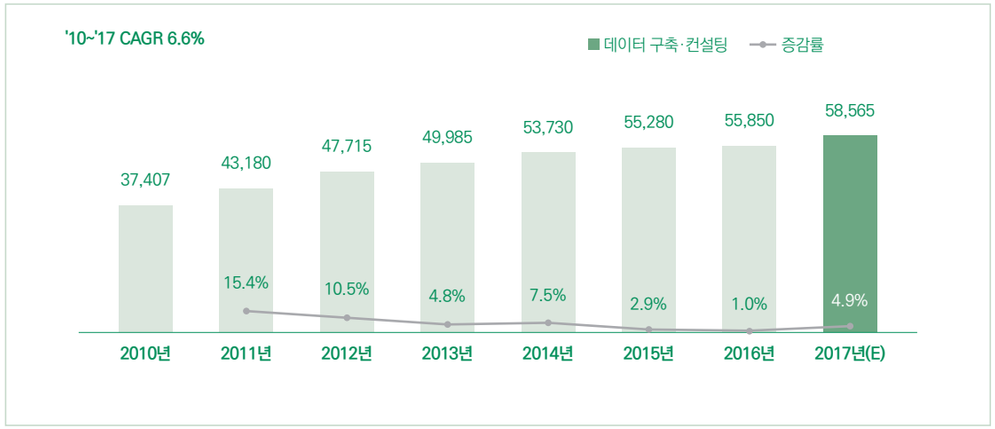
구분		공공	금융	제조·건설	유통·서비스	통신·미디어	의료	전체
데이터 수집	2016년	43.7	10.4	14.8	27.5	2.8	0.8	100.0
	2017년(E)	21.3	11.7	23.0	23.9	17.8	2.3	100.0
데이터 설계	2016년	25.8	12.0	20.5	31.5	6.9	3.3	100.0
	2017년(E)	28.2	11.1	17.2	26.9	13.9	2.7	100.0
DBMS	2016년	49.6	7.5	16.3	17.7	5.0	3.9	100.0
	2017년(E)	39.3	5.0	12.7	26.1	16.9	0.0	100.0
데이터 관리	2016년	42.3	7.7	17.2	24.3	4.4	4.1	100.0
	2017년(E)	31.1	12.1	14.5	23.6	14.7	4.0	100.0
데이터 품질	2016년	22.4	29.4	22.8	14.1	9.6	1.7	100.0
	2017년(E)	7.5	6.6	33.3	26.8	24.6	1.2	100.0
데이터 분석	2016년	41.3	12.5	17.2	17.5	8.2	3.3	100.0
	2017년(E)	23.0	7.3	28.6	21.4	9.1	10.6	100.0
데이터 플랫폼	2016년	48.1	10.8	8.9	16.8	13.9	1.5	100.0
	2017년(E)	21.3	11.0	16.5	30.2	20.0	1.0	100.0
데이터 솔루션 전체	2016년	39.0	12.9	16.8	21.4	7.2	2.7	100.0
	2017년(E)	26.7	10.4	18.9	24.7	15.7	3.6	100.0

나. 데이터 구축·컨설팅 시장

2017년 데이터 구축·컨설팅 시장 규모는 5조 8,565억 원으로 2016년 대비 4.9% 상승할 것으로 나타났다. 2010년 이후 연평균 6.6%로 성장하고 있다.

[그림 3-1-6] 데이터 구축·컨설팅 시장 규모

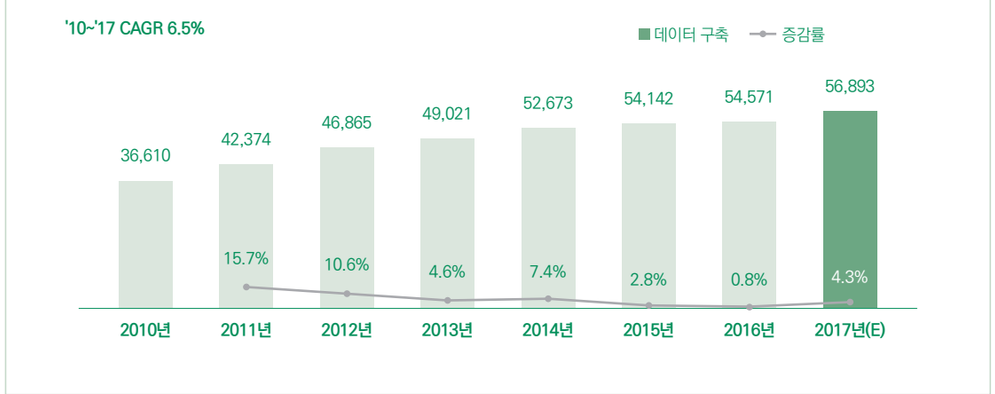
(단위: 억 원)



데이터 이행·처리, DB 설계·구축, 기계 처리형 데이터 구축 분야를 포함하고 있는 데이터 구축 시장은 2017년 5조 6,893억 원으로 전년 대비 4.3% 상승할 것으로 나타났다.

[그림 3-1-7] 데이터 구축 시장 규모

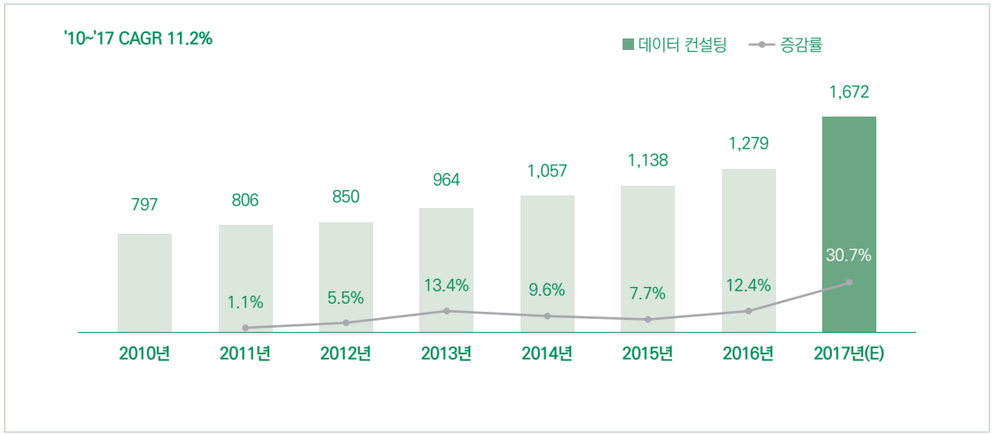
(단위: 억 원)



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 2017년 1,672억 원으로 2016년 대비 30.7% 성장했으며, 2010년 이후 연평균성장률 11.2%로 꾸준한 성장세를 유지하는 것으로 나타났다.

[그림 3-1-8] 데이터 컨설팅 시장 규모

(단위: 억 원)

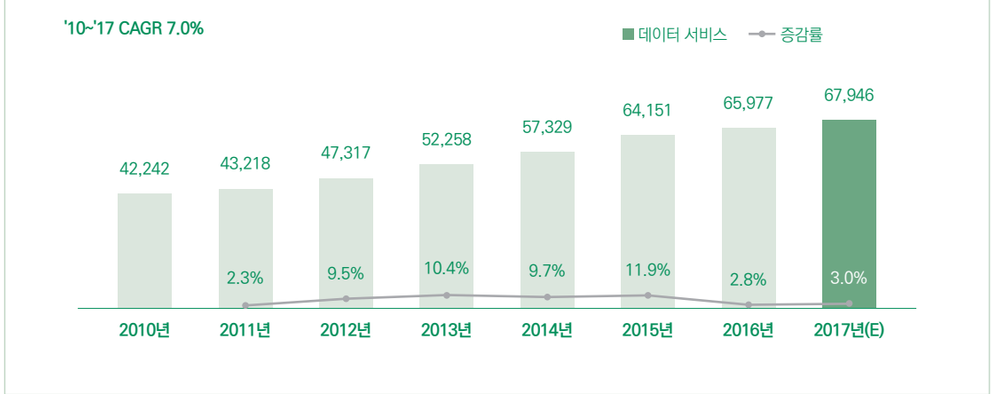


다. 데이터 서비스 시장

데이터 서비스 시장은 2017년 6조 7,946억 원으로 2016년 대비 3.0% 성장한 것으로 나타났다. 2010년 이후 연평균성장률은 7.0%로 조사됐다.

[그림 3-1-9] 데이터 서비스 시장 규모

(단위: 억 원)



데이터 서비스는 크게 데이터나 DB를 기반으로 정보를 제공하는 정보제공 서비스, 데이터를 직접 판매하거나 중개하는 데이터거래 서비스, 데이터를 분석해 새로운 가치를 제공하는 분석제공 서비스로 구분할 수 있다. 이 중 정보제공 서비스가 90% 가량을 차지하는 가장 큰 시장을 구성하며, 이어서 분석제공 서비스, 데이터거래 서비스 순이다.

[그림 3-1-10] 데이터 서비스 중분류별 시장 규모 비중

(단위: 억 원)

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

[표 3-1-5] 데이터 서비스 중분류별 시장 규모

(단위: 억 원, %)

구분	2013년		2014년		2015년		2016년		2017년(E)		증감률 '16~'17	CAGR '13~'17
	규모	비중	규모	비중	규모	비중	규모	비중	규모	비중		
데이터거래	1,708	3.3	2,019	3.5	2,379	3.7	2,594	3.9	2,709	4.0	4.4	12.2
정보 제공	48,284	92.4	51,985	90.7	58,171	90.7	59,854	90.7	61,575	90.6	2.9	6.3
데이터 분석 제공	2,266	4.3	3,325	5.8	3,601	5.6	3,529	5.4	3,662	5.4	3.8	12.7
데이터 서비스 전체	52,258	100.0	57,329	100.0	64,151	100.0	65,977	100.0	67,946	100.0	3.0	6.8

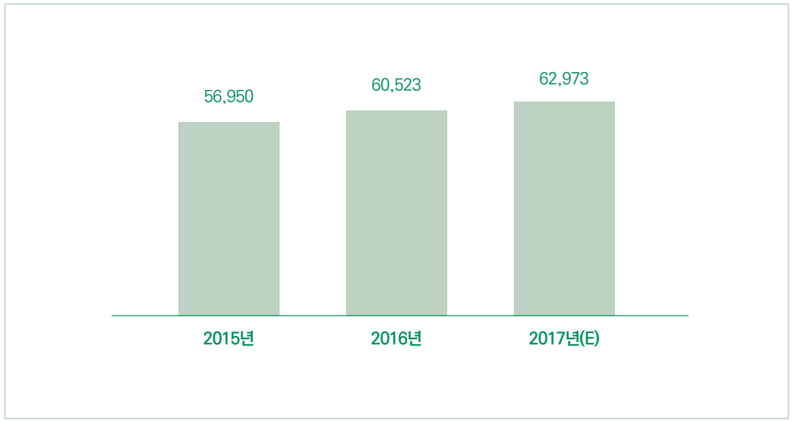
3. 국내 데이터산업 직접매출 시장 규모[4]

2017년 데이터산업 직접매출 시장 규모는 6조 2,973억 원으로 2016년 대비 4.0% 성장한 것으로 나타났다.

4 데이터산업 직접매출 시장 규모는 데이터와 직접적으로 관련이 있는 매출 기준으로 추정한 결과로서, 전통적인 DB 산업을 아우르는 기존 조사 범위에서 데이터를 매개로 하는 광고매출, 기타 운영시스템 SI 매출 등 DB 관련 간접 매출을 제외하고 산출한 결과임

[그림 3-1-11] 데이터산업 직접매출 시장 규모

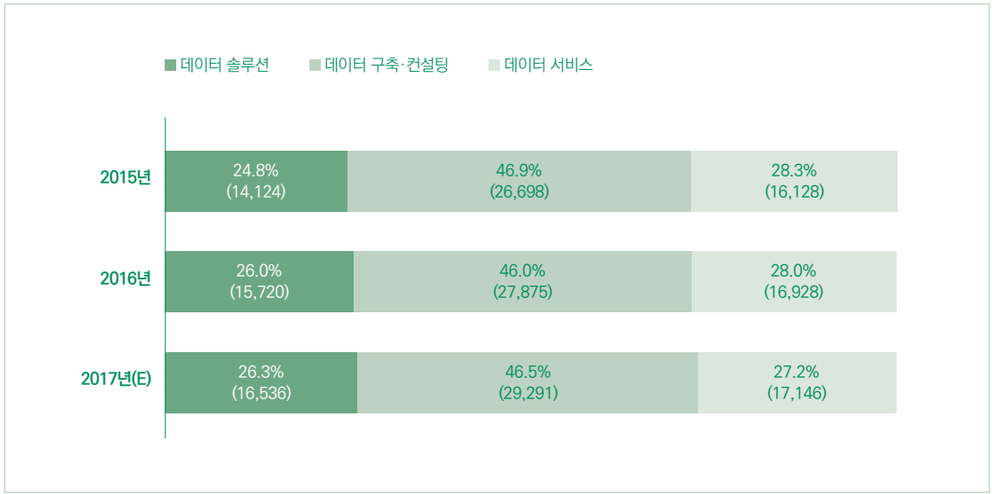
(단위: 억 원)



부문별 직접매출 시장 규모를 보면, 데이터 구축·컨설팅 시장은 2017년 2조 9,291억 원, 데이터 서비스 시장은 1조 7,146억 원, 데이터 솔루션 시장은 1조 6,536억 원으로 나타났다.

[그림 3-1-12] 데이터산업 부문별 직접매출 시장 규모 비중

(단위: 억 원)



4. 국내 데이터직무 인력 현황

국내 데이터산업에 종사하고 있는 인력은 총 29만 4,753명으로 전년 대비 2.1% 증가했으며, 이 중 데이터직무[5] 인력은 7만 7,105명으로 전년 대비 5.3% 증가한 것으로 나타났다.

[표 3-1-6] 데이터산업 인력 현황 (단위: 명, %)

구분	2015년		2016년		2017년		증감률 '16~'17
	인력 규모	비중	인력 규모	비중	인력 규모	비중	
데이터직무	70,338	25.1	73,256	25.4	77,105	26.2	5.3
데이터직무 외	209,985	74.9	215,365	74.6	217,648	73.8	1.1
전체	280,323	100.0	288,621	100.0	294,753	100.0	2.1

데이터산업 부문별 데이터직무 인력을 보면, 데이터솔루션 분야의 인력이 11.0% 증가했으며, 구축·컨설팅 분야의 데이터직무 인력이 전년 대비 6.0% 증가했다.

[표 3-1-7] 데이터산업 내 데이터직무 인력 현황 (단위: 명, %)

구분	2015년		2016년		2017년		증감률 '16~'17
	인력 규모	비중	인력 규모	비중	인력 규모	비중	
데이터솔루션	8,886	12.6	9,272	12.7	10,291	13.3	11.0
데이터구축·컨설팅	34,323	48.8	35,404	48.4	37,516	48.7	6.0
구축	29,304	41.7	29,941	40.9	31,693	41.1	5.9
컨설팅	5,019	7.1	5,463	7.5	5,823	7.6	6.6
데이터서비스	27,129	38.6	28,580	39.0	29,298	38.0	2.5
전체	70,338	100.0	73,256	100.0	77,105	100.0	5.3

※ 이후 데이터구축·컨설팅 분야 데이터직무 인력은 편의상 데이터구축과 데이터컨설팅으로 구분해 표시하겠음

5 데이터직무 인력이란, 기업에서 필요한 데이터 관련 기획, 개발, 분석, 운영 및 관리 직무를 수행하는 인력을 말한다. DA(Data Architect), 데이터 개발자, 데이터 엔지니어, 데이터 분석가, DBA(DB Administrator), 데이터 과학자, 데이터 컨설턴트, 데이터 기획·마케터로 구분한다.

일반산업[6]을 포함한 전 산업의 2017년 데이터직무 인력은 총 10만 9,320명으로 2016년 대비 6.8% 증가한 것으로 나타났다.

[표 3-1-8] 전 산업 내 데이터직무 인력 현황 (단위: 명, %)

구분	2015년		2016년		2017년		증감률 '16~'17
	인력 규모	비중	인력 규모	비중	인력 규모	비중	
데이터산업	70,338	70.0	73,256	71.6	77,105	70.5	5.3
일반산업	30,102	30.0	29,119	28.4	32,215	29.5	10.6
전 산업	100,440	100.0	102,375	100.0	109,320	100.0	6.8

데이터직무별로 보면 데이터 개발자가 4만 1,254명으로 가장 큰 비중을 차지했고, 이어서 DBA 1만 7,863명, 데이터 엔지니어 1만 6,634명 순이었다.

[표 3-1-9] 2017년 전 산업의 데이터직무별 인력 현황 (단위: 명)

구분	데이터산업	일반산업	전 산업
데이터아키텍트(DA)	5,025	5,046	10,071
	6.5%	15.7%	9.2%
데이터 개발자	29,267	11,987	41,254
	38.0%	37.2%	37.7%
데이터 엔지니어	13,127	3,507	16,634
	17.0%	10.9%	15.2%
데이터 분석가	4,792	3,606	8,398
	6.2%	11.2%	7.7%
데이터베이스 관리자(DBA)	11,851	6,012	17,863
	15.4%	18.6%	16.3%
데이터 사이언티스트	1,283	520	1,803
	1.7%	1.6%	1.7%

(계속)

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

구분	데이터산업	일반산업	전 산업
데이터 컨설턴트	4,263	741	5,004
	5.5%	2.3%	4.6%
데이터 기획 · 마케터	7,497	796	8,293
	9.7%	2.5%	7.6%
전체	77,105	32,215	109,320
	100.0%	100.0%	100.0%

2016년 대비 가장 높은 증가율을 나타낸 데이터직무는 데이터 분석가로 전년 대비 14.4% 증가한 8,398명으로 조사됐다.[7]

[표 3-1-10] 2015년~2016년 전 산업의 데이터직무별 인력 현황 비교 (단위: 명, %)

구분	2015년		2016년		2017년		증감률 '16~'17
	인력 규모	비중	인력 규모	비중	인력 규모	비중	
데이터아키텍트(DA)	7,065	7.0	9,267	9.1	10,071	9.2	8.7
데이터 개발자	41,205	41.0	38,948	38.0	41,254	37.7	5.9
데이터 엔지니어	14,013	14.0	15,670	15.3	16,634	15.2	6.2
데이터 분석가	5,120	5.1	7,339	7.2	8,398	7.7	14.4
데이터베이스 관리자(DBA)	17,427	17.4	17,116	16.7	17,863	16.3	4.4
데이터 사이언티스트	1,343	1.3	1,662	1.6	1,803	1.7	8.5
데이터 컨설턴트	4,351	4.3	4,513	4.4	5,004	4.6	10.9
데이터 기획 · 마케터	9,916	9.9	7,860	7.7	8,293	7.6	5.5
전체	100,440	100.0	102,375	100.0	109,320	100.0	6.8

※ 2015년과 2016년 데이터직무 분류를 조정해 재구성한 표임

7 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 데이터직무 인력 중 빅데이터 관련 인력은 총 9,955명으로, 전체의 9.1%를 차지하고 있다.

[표 3-1-11] 전 산업 내 데이터직무 중 빅데이터 관련 인력 현황 (단위: 명, %)

구분	전 산업 내 데이터직무 전체		빅데이터 관련 인력		전 산업 내 데이터직무 중 빅데이터 인력 비중
	규모	비중	규모	비중	
데이터아키텍트(DA)	10,071	9.2	-	-	-
데이터 개발자	41,254	37.7	2,908	29.2	7.0
데이터 엔지니어	16,634	15.2	1,679	16.9	10.1
데이터 분석가	8,398	7.7	1,108	11.1	13.2
데이터베이스 관리자(DBA)	17,863	16.3	-	-	-
데이터 사이언티스트	1,803	1.7	1,803	18.1	100.0
데이터 컨설턴트	5,004	4.6	1,724	17.3	34.5
데이터 기획 · 마케터	8,293	7.6	733	7.4	8.8
전체	109,320	100.0	9,955	100.0	9.1

※ 빅데이터 관련 인력: 데이터직무 인력 중 빅데이터 관련 기술 및 능력을 보유한 인력으로 데이터직무 인력 수에 포함됨
※ DA와 DBA는 조사 제외 대상이며, 데이터 사이언티스트 직무는 100% 빅데이터 관련 인력으로 간주

5. 국내 데이터직무 인력 수요

향후 3년 내, 일반산업을 포함한 전 산업에서 추가로 필요로 하는 데이터직무 인력은 총 1만 3,337명으로 조사됐다. 이 중 데이터 개발자 수요가 37.2%로 가장 높게 나타났다.

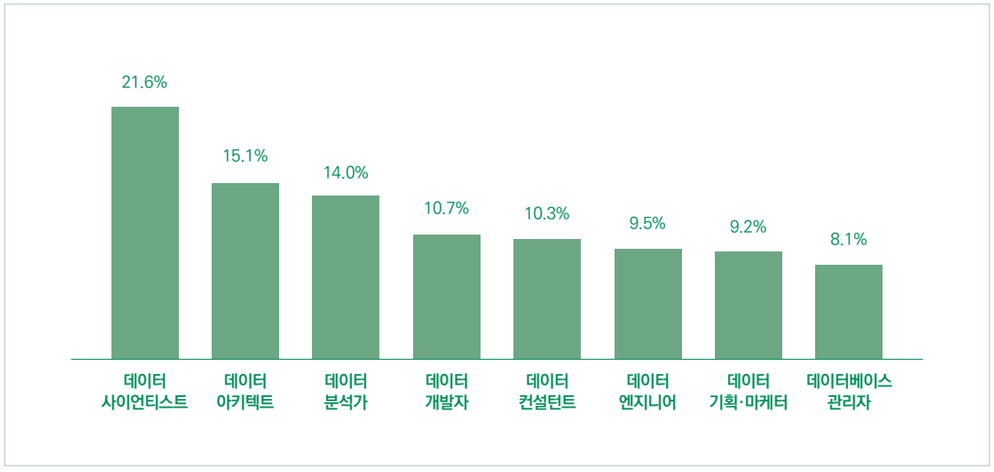
[표 3-1-12] 향후 3년 내 전 산업 내 데이터직무별 필요인력 (단위: 명)

구분	데이터산업	일반산업	전 산업
데이터아키텍트(DA)	354	1,437	1,791
	7.3%	17.0%	13.4%
데이터 개발자	2,564	2,389	4,953
	52.7%	28.2%	37.2%
데이터 엔지니어	775	963	1,738
	15.9%	11.4%	13.0%
데이터 분석가	394	977	1,371
	8.1%	11.5%	10.3%
데이터베이스 관리자(DBA)	145	1,424	1,569
	3.0%	16.8%	11.8%
데이터 사이언티스트	158	338	496
	3.2%	4.0%	3.7%
데이터 컨설턴트	164	410	574
	3.4%	4.8%	4.3%
데이터 기획 · 마케터	311	534	845
	6.4%	6.3%	6.3%
전체	4,865	8,472	13,337
	100.0%	100.0%	100.0%

전체적으로는 데이터 개발자에 대한 수요가 높게 나타나고 있으나, 데이터직무 인력 부족률[8]을 보면 데이터 사이언티스트, 데이터 아키텍트(DA), 데이터 분석가 직무인력이 가장 부족한 것으로 나타났다.

8 인력 부족률 = 필요인력/(현재인력+필요인력)

[그림 3-1-13] 2017년 전 산업 내 데이터직무별 인력 부족률 (단위: %)



[표 3-1-13] 2017년 전 산업 내 데이터직무별 인력 부족률 (단위: %)

구분	데이터산업	일반산업	전 산업
데이터아키텍트(DA)	6.6	22.2	15.1
데이터 개발자	8.1	16.6	10.7
데이터 엔지니어	5.6	21.5	9.5
데이터 분석가	7.6	21.3	14.0
데이터베이스 관리자(DBA)	1.2	19.2	8.1
데이터 사이언티스트	11.0	39.4	21.6
데이터 컨설턴트	3.7	35.6	10.3
데이터 기획 · 마케터	4.0	40.2	9.2
전체	5.9	20.8	10.9

※ 인력 부족률 : 필요인력/(현재인력+필요인력)

전 산업의 데이터직무 필요인력 중 빅데이터 관련 필요인력은 총 6,008명으로 나타났다. 이 중 데이터 개발자가 39.1%로 가장 수요가 높고, 다음으로 데이터 분석가(22.1%), 데이터 엔지니어(15.1%) 등의 순으로 나타났다.

[표 3-1-14] 전 산업 내 부문별 데이터직무 중 빅데이터 관련 필요인력 (단위: 명)

구분	전 산업 내 데이터직무 중 빅데이터 필요인력		
	데이터산업	일반산업	전 산업
데이터 개발자	660	1,690	2,350
	47.8%	36.5%	39.1%
데이터 엔지니어	123	784	907
	8.9%	16.9%	15.1%
데이터 분석가	227	1,104	1,331
	16.5%	23.9%	22.1%
데이터 사이언티스트	158	338	496
	11.5%	7.3%	8.3%
데이터 컨설턴트	104	405	509
	7.5%	8.8%	8.5%
데이터 기획 · 마케터	108	307	415
	7.8%	6.6%	6.9%
전체	1,380	4,628	6,008
	100.0%	100.0%	100.0%

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

[표 3-1-15] 2017 전 산업 내 데이터직무 중 빅데이터 인력 부족률 (단위: %)

구분	전 산업 내 데이터직무 중 빅데이터 인력 부족률		
	데이터산업	일반산업	전 산업
데이터 개발자	26.1	61.9	44.7
데이터 엔지니어	10.2	56.7	35.1
데이터 분석가	37.3	60.3	54.6

(계속)

구분	전 산업 내 데이터직무 중 빅데이터 인력 부족률		
	데이터산업	일반산업	전 산업
데이터 사이언티스트	11.0	39.4	21.6
데이터 컨설턴트	7.1	53.3	22.8
데이터 기획 · 마케터	23.9	44.1	36.1
전체	17.9	56.0	37.6

※ 인력 부족률 : 필요인력/(현재인력+필요인력)

4차산업혁명시대에 데이터는 산업간 융·복합의 도구이자, 혁신 성장을 이끄는 기초 원동력이다. 데이터 기업들이 시장 환경 변화에 적극적으로 대응하며 경쟁력을 강화시키고, 이를 정부에서 전폭적으로 지원해준다면 데이터산업은 계속해서 성장할 것으로 기대된다.

제2장

해외 데이터산업 현황

데이터산업은 나라와 기관마다 그 범위와 정의가 다양해 어느 기준을 적용하느냐에 따라 시장 규모가 다르게 나타난다. 이 글에서는 미국·EU·일본 등 세계 주요 경제권의 데이터산업 규모, 데이터 기업 수, 경제적 효과, GDP 대비 데이터 경제적 효과의 비율, 글로벌 빅데이터 시장 동향을 중심으로 해외 데이터산업 현황을 알아본다.

1. 데이터산업의 시장 규모

데이터산업 시장은 나라와 기관마다 그 범위와 정의가 다양해 시장의 정의와 분류를 어떻게 하느냐에 따라 그 규모가 다르게 나타난다. 해외 데이터산업 현황은 글로벌 시장 분석 자료를 제공하는 기관의 정의에 따라 각각 소개하고자 한다.

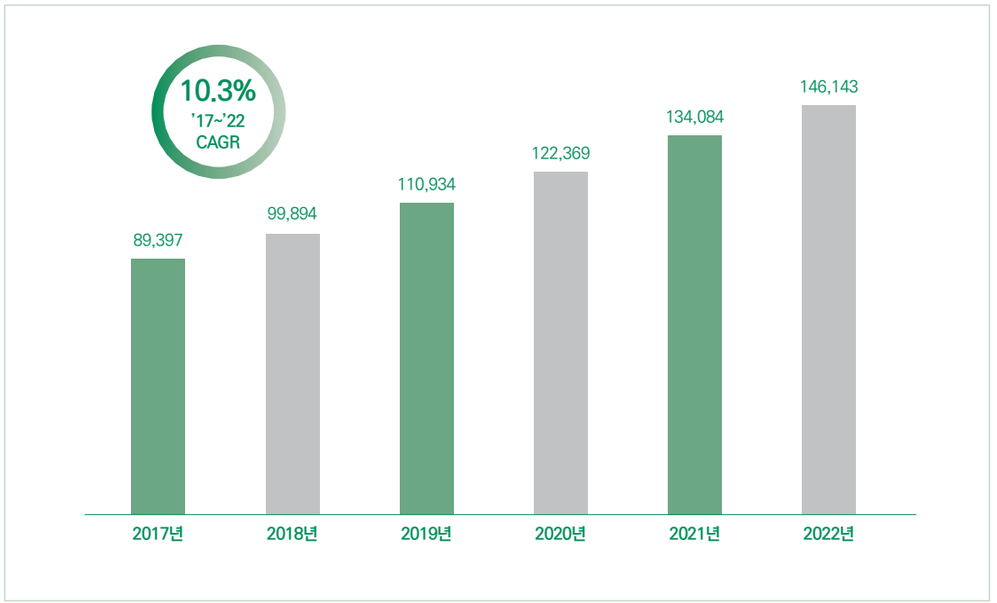
가. 데이터 기반 솔루션 시장

451 Research가 분석한 세계 데이터 시장은 8개의 세부 솔루션 시장[1]으로 이뤄져 있다. 8개 솔루션의 전체 시장은 2018년 999억 달러 규모로 형성됐다. 2017년부터 2022년까지의 연평균성장률(CAGR)은 10.3%를 기록했으며, 2022년에는 1,461억 달러로 성장할 것으로 예상된다.

8개 세부 시장 중 가장 큰 시장은 운영 데이터베이스(Operational databases) 시장으로, 2018년 기준 약 400억 달러 이상으로 형성됐다. 리포팅과 분석(Reporting and analytics) 시장은 약 200억 달러, 데이터 분석 플랫폼(Analytic data platforms) 시장은 150억 달러 내외로 형성돼 각각 2, 3위를 차지했다.

• 필자 | **나영민**(날리지리서치그룹 이사)

[그림 3-2-1] 2017~2022년 데이터 기반 솔루션 전체 시장 규모 (단위: 100만 달러)



출처: 451Research, 2018.

나. 정보 서비스 시장

데이터 솔루션 시장과는 별도로 데이터를 활용해 온라인/오프라인 서비스를 하는 시장은 '정보 서비스 시장'으로 정의할 수 있다[2]. 2016년 정보서비스 전체 시장 규모는 1조 6,000억 달러로, 2015년 1조 5,000억 달러 대비 6.4% 증가했으며, 2017년부터 2020년까지는 연평균 약 4% 내외의 성장세를 유지할 것으로 전망된다.

다양한 정보 서비스 중 가장 큰 시장은 엔터테인먼트 부문으로, 2016년 기준 약 6,529억 달러로 형성됐다. 마케팅 서비스 부문은 2016년 기준 1,575억 달러를 기록했다.

대부분의 정보 서비스 시장이 성장세이지만 News, Yellow Pages, Geophysical & Geomapping 서비스 부문은 전년 대비 하락세를 보이기도 했다.

1 8개 세부 시장: Operational databases, Analytic data platforms, Reporting and analytics, Data management, Corporate performance management, Event/stream processing, Distributed data grid/cache, Search-based data platforms and analytics.

2 OUTSELL은 데이터 기반 정보 서비스를'Information Industry'로 정의하고 총 27개 정보 서비스 영역으로 구분하고 있다.

[표 3-2-1] 2015~2016년 글로벌 정보 서비스 세부 시장 규모

(단위: 10억 달러)

Industry Ecosystem and Segments	2015년($)	2016년($)	성장률(%)
Media, Marketing & Analytics	**434.6**	**451.5**	**4.4**
Marketing Services	149.6	157.5	5.3
Adtech	72.2	78.9	9.2
News	67.5	65.4	-3.1
B2B Media & Business Information	38.2	39.4	3.1
Marketing Research	38.2	38.8	1.6
Yellow Pages	17.9	15.7	-12.3
CRM Solutions	13.4	14.4	7.5
Martech	12.8	14.3	11.7
Supply Chain Automation & Procurement Solutions	10.7	11.8	9.8
Company, Contact & Personal Information	10.0	10.9	8.6
IT Research	4.1	4.4	8.0
Financial, Credit, Legal, GRC, Tax & Accounting	**158**	**168.1**	**5.4**
Financial Information & Solutions	57.9	63.8	10.2
Tax & Accounting Solutions	44.0	45.8	4.3
Governance, Risk & Compliance Solutions	21.8	22.9	4.9
Legal & Regulatory Solutions	22.0	22.7	3.1
Credit Information & Solutions	12.3	12.8	4.7
Science, Technology & Healthcare	**126.7**	**137.4**	**3.8**
Health Information & Health IT	80.1	90.9	13.5
Scientific & Technical Information & Solutions	12.9	13.3	2.7
Medical Information	10.7	11.4	6.2
Geophysical & Geomapping	13.6	11.1	-18.4
Pharma Information & Solutions	9.4	10.8	15.2
Education, Training & Human Capital Management	**142.8**	**150.1**	**4.3**
HR Services & Solutions	49.6	55.3	11.5
Teaching & Learning	52.8	53.3	0.8
Corporate Training	22.4	23.4	4.3
Education Management Solutions	18.0	18.1	0.7
Consumer Entertainment	**605.4**	**652.9**	**7.8**
Content Technology & Distribution Services	**73.2**	**80.1**	**9.5**
Total Industry	**1.5T**	**1.6T**	**6.4**

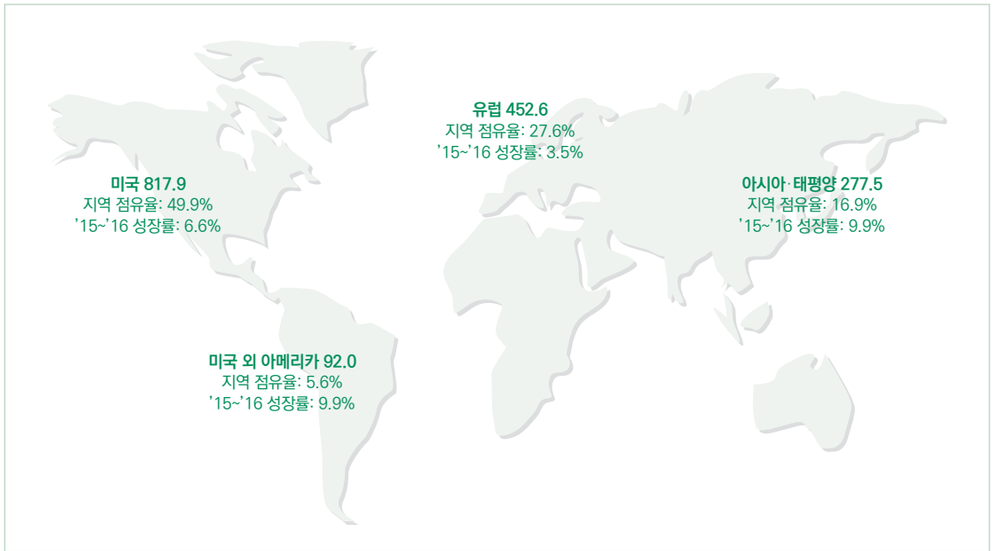
(계속)

출처: OUTSELL, INFORMATION INDUSTRY OUTLOOK 2018

대륙별 시장 분포를 보면, 미국이 전체 정보 서비스 시장의 절반에 가까운 49.9%를 차지했다. 미국의 2016년 정보 서비스 시장은 전년 대비 6.6% 성장한 8,179억 달러인 것으로 조사됐다. 미국의 뒤를 이어 2016년 유럽은 3.5% 성장한 4,526억 달러의 시장이 형성돼, 전체 시장에서 27.6%를 차지했다. 2016년 아시아·태평양 지역은 9.9% 성장한 2,775억 달러의 시장이 형성돼 세계 시장에서 16.9%를 차지했다. 미국·유럽·아시아가 전체 시장의 94.4%를 차지하고 있다.

[그림 3-2-2] 2016년 대륙별 데이터 기반 정보 서비스 시장 비교

(단위: 10억 달러)



출처: OUTSELL, INFORMATION INDUSTRY OUTLOOK 2018

다. 데이터 시장

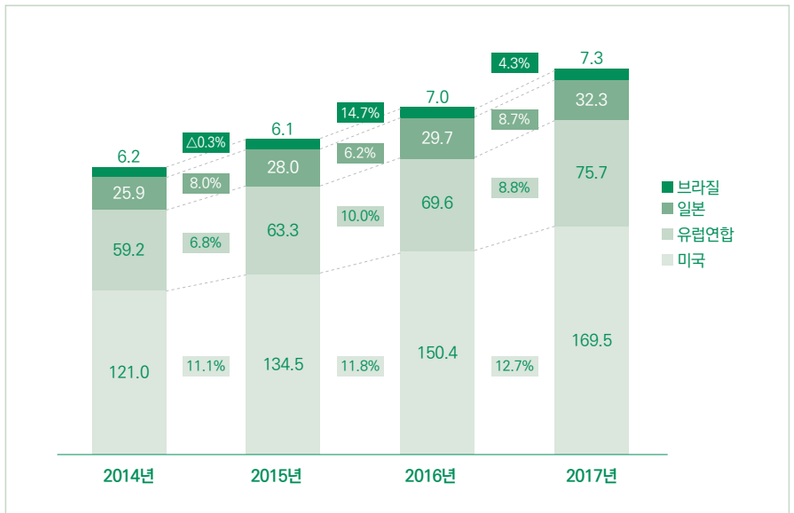
IDC와 The Lisbon Council은 The European Data Market Monitoring Tool Report에서 데이터 시장의 범위를 '데이터를 가공함으로써 제품 및 서비스로 재생산되는 디지털 데이터 시장'으로 정의했다. 물론 이미지 데이터처럼 디지털 기술을 통해 수집·저장·가공·전송되는 멀티미디어 데이터 부문도 포함돼 있다.

데이터 시장규모는 데이터 관련 기업의 매출액을 기반으로 산출됐다. 데이터 기업의 매출은 데이터 관련 제품의 총 가치에 해당하고, 수출을 포함해 해당 국가에 기반을 둔 기업이 생산한 서비스의 매출을 기반으로 한다.

미국의 데이터산업은 세계에서 가장 크다. 시장의 성장세 또한 유럽과 더불어 매우 높다. 미국은 2014년에 약 1,210억 달러를 기록했으며, 2015년에는 2014년 대비 약 11.1% 성장한 약 1,345억 달러 규모로 형성됐다. 2016년에는 2015년 대비 11.8% 성장해 다시 한 번 두 자리 수 성장세를 유지했다. 2017년 시장은 전년 대비 12.7%라는 높은 성장률을 기록하면서 약 1,695억 달러로 형성된 것으로 나타났다.

미국 데이터산업은 전 세계에서 가장 빠르게 성장하고 있으며, EU보다 두 배 이상 크다. 하지만 빠르게 성장하는 이면에는 대기업 중심의 성장, 일부 기업에만

[그림 3-2-3] 2014~2017년 세계 데이터산업 규모 (단위: 10억 달러)



출처: IDC & The Lisbon Council, How the Power of Data will Drive EU Economy: The European Data Market Monitoring Tool Report, 2018.4.20.

편중되는 문제점도 드러나고 있다. 미국뿐 아니라 다른 나라들은 자국의 수많은 중소기업이 다양한 분야에서 고른 성장세를 보여야 한다는 점을 알고 있으며, 이를 극복할 과제로 인식하고 있다.

EU의 데이터 시장 성장세는 미국과 견줄 정도로 빠른 편이다. 2014년에 약 592억 달러를 기록했으나, 이후 2016~2017년 사이에는 9% 전후의 빠른 성장세를 기록했다. 덕분에 2017년에는 약 757억 달러로 도약한 것으로 나타났다. EU 또한 빠른 성장률을 유지함으로써 데이터산업이 견고하게 형성됐음을 입증했다. EU의 데이터산업은 미국과 달리 중소기업과 벤처기업이 주도한다는 점에서 매우 바람직하다고 평가받고 있다.

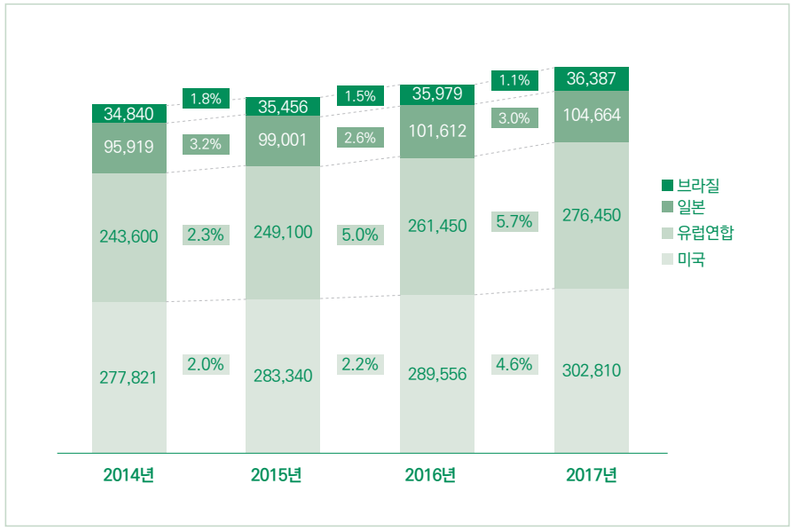
일본의 데이터산업은 2014년에 약 259억 달러로 형성됐으며, 2015년에는 280억 달러로 성장했다. 2016년에는 2015년보다 6.2% 성장한 297억 달러, 2017년에는 8.7% 성장한 323억 달러로 추정돼, 안정적인 성장세를 이어가는 것으로 조사됐다. 일본 데이터 기업의 수익률 또한 약 24%로 견고한 구조인 것으로 조사됐다.

국제통화기금(IMF)을 따르면 일본의 실질 GDP는 2014년과 2015년에 감소했으나, 글로벌 시장분석 기관인 IDC는 일본에서 ICT 시장만큼은 같은 시기에 완만하게 성장세를 유지한 것으로 분석했다. 2017년 이후 ICT 투자환경도 꾸준히 개선될 것으로 전망된다. 또한 투자환경이 개선됨에 따라 일본의 데이터 시장 규모는 지속적인 성장세를 기록할 전망이다.

한편 브라질의 데이터 시장 규모는 2014년에 약 62억 달러를 기록했으나, 2015년에는 성장세가 주춤해지며 −0.3%로 하락했다. 그러나 2016년에는 다시 반등해 14.7% 수준의 성장세로 돌아섰다.

2017년에는 73억 달러 수준으로 성장했다. 브라질에서의 ICT 투자 대비 데이터 기업의 수익률을 살펴보면, 2014년을 기준으로 약 4.0%를 기록해 미국의 약 10.0%, EU의 약 8.7%에 비해 낮은 수준으로 조사됐다. 그럼에도 브라질 데이터산업은 빠르게 성장하면서 프랑스, 이탈리아 등과 경쟁할 수 있는 수준으로 올라설 것으로 전망된다.

[그림 3-2-4] 2014~2017년 데이터 기업 수 (단위: 개)



출처: IDC & The Lisbon Council, How the Power of Data will Drive EU Economy: The European Data Market Monitoring Tool Report, 2018.4.20.

2. 데이터 기업 수

데이터 기업이란 시장에 데이터를 공급하는 기업으로서, 데이터의 생산, 유통 및 활용을 담당하는 기업으로 정의된다. 즉 데이터 관련 제품과 서비스, 기술을 공급하는 기업을 말한다. 데이터 시장이 성장함에 따라 데이터 기업 수도 지역에 따라 점진적 또는 폭발적으로 증가하고 있다.

미국, EU, 일본, 브라질의 데이터 기업 수는 데이터 종사자 수와 마찬가지로 2014년부터 꾸준히 증가하고 있다. 미국의 데이터 기업 수는 2014년 27만 7,821개에서 2017년에는 30만 개 이상으로 늘어났다. EU의 데이터 기업 수는 2014년 약 24만 3,000개였으나 이후 꾸준히 증가해 2017년에는 약 27만 6,450개를 돌파했다. 미국과 EU의 데이터산업 규모가 크게 차이나는 것에 비해 데이터 기업 수는 별 차이가 없다. 이것은 EU의 데이터 기업이 상대적으로 다양한 데다 중소기업 중심이어서 이런 현상이 나타난 것이다.

2014년에 약 9만 5,919개였던 일본의 데이터 기업 수는 2017년에 들어서며 10만 개를 넘어선 것으로 집계됐다. 장기적인 경제 침체를 겪고 있는 가운데도 일본의 데이터 기업 수가 꾸준히 증가했다는 점은 데이터산업에 대한 일본 기업들의 수요와 기대가 반영된 것이라고 볼 수 있다. 한편 브라질의 데이터 기업 수는 2014년 3만 4,840개에서 꾸준히 성장해 2017년에 들어서며 3만 6,387개로 증가했다. 브라질의 ICT 시장이 상대적으로 작고, ICT 보급률 또한 매우 낮은 수준이라서 성장률이 낮지만 향후 지속적 성장이 기대된다.

3. 데이터산업의 경제적 효과

가. 직접 효과

데이터산업의 경제적 효과는 데이터 시장이 경제 생태계 전체에 미치는 전반적인 영향이 어느 정도인가를 평가하는 분석 방법론이다.

데이터산업의 경제적 효과는 직접 효과와 간접 효과로 구분된다. 직접 효과란 데이터산업 그 자체적으로 형성된 효과를 의미한다. 즉 데이터 생산과 관련한 모든 비즈니스 활동에 의해 형성된 효과다. 정량적인 직접 효과는 판매된 데이터 제품 및 서비스의 수익을 기준으로 측정된다.

미국의 데이터 경제적 효과 중 직접 효과는 2014년에 약 1,157억 달러를 기록한 이후 점차 증가해 2017년에는 약 1,324억 달러로 집계됐다. 데이터에 기반을 둔 제품 및 서비스에서 파생된 직접적 영향은 EU를 비롯한 일본, 브라질보다 높다. 이는 데이터 시장의 규모와 GDP에서의 비중이 상대적으로 높기 때문이다.

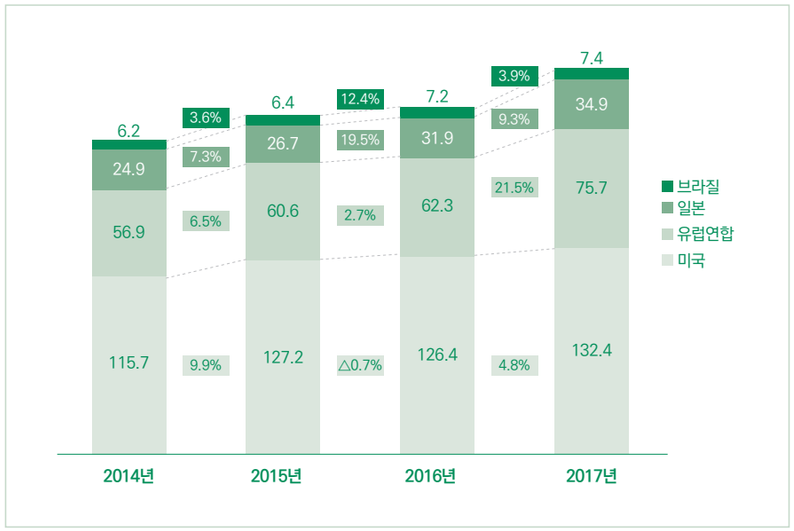
직접 효과는 데이터산업 그 자체에 의해 형성된다는 정의를 따르면, 미국은 데이터 생산에 매우 적극적인 국가임을 알 수 있다. 또한 미국의 데이터 직접효과는 EU의 직접효과의 두 배로 형성되고 있는데, 이는 미국의 데이터산업이 더 발전돼 있고 경제적 효과의 확산 속도가 더 빠름을 의미한다.

EU의 데이터 경제적 효과 중 직접 효과는 2014년에 약 569억 달러를 기록한 이후 지속적으로 증가해 2017년에는 약 757억 달러로 집계됐다. 특히 2016년에서 2017년의 성장세가 21.5%로 두드러졌다. 높은 성장세를 유지함으로써 EU의 데이터 활용이 다른 곳보다 활발함을 알 수 있다. EU의 데이터산업의 규모가 지속적으로 증가한 것처럼 직접 효과도 높여 나갈 것으로 전망된다.

일본의 데이터 경제적 효과 중 직접 효과는 2014년에 약 249억 달러를 기록

[그림 3-2-5] 2014~2017년 경제적 효과: 직접 효과

(단위: 10억 달러)



출처: IDC & The Lisbon Council, How the Power of Data will Drive EU Economy: The European Data Market Monitoring Tool Report, 2018.4.20.

했으나 2017년에는 크게 성장해 349억 달러까지 상승했다. 브라질의 데이터 경제적 효과 중 직접 효과는 2014년에 약 62억 달러를 기록한 이후 점차 증가해 2017년에는 약 74억 달러로 나타났다. 브라질의 ICT 산업이 빠르게 발전함에 따라 데이터 시장 또한 빠르게 성장하고 있다. 브라질은 특히 경제 위기를 더 심하게 겪었음에도 데이터에 기반을 둔 제품 및 서비스의 영향력이 확대되고 있어 꾸준한 성장을 기대하게 한다.

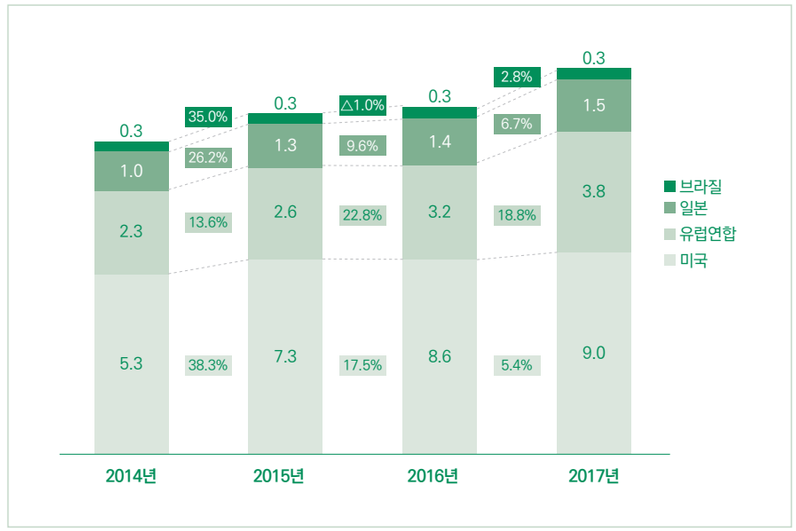
나. 간접 후방 효과

데이터의 경제적 효과 중 간접 효과는 데이터산업이 타 산업에 미치는 영향을 나타내는 요소다. 직접 효과가 데이터 생산 활동이 데이터산업 내에 미치는 영향을 설명한 것이라면, 간접 효과는 타 산업에 미치는 영향을 설명한다. 간접 효과는 전방 효과와 후방 효과로 구분된다. 후방 효과란 데이터산업에 속한 기업이 데이터 기반의 제품과 서비스 활동 과정에서, 타 산업으로부터 자원을 제공받을 때에 자원을 제공해준 산업에서 발생되는 효과를 말한다.

데이터산업에서 생산된 데이터 기반의 제품 및 서비스를 이용해 타 산업에서 다른 형태의 제품 및 서비스로 가공되는 전방 효과는 세 가지로 구분할 수 있

[그림 3-2-6] 2014~2017년 경제적 효과: 간접 후방 효과

(단위: 10억 달러)



출처: IDC & The Lisbon Council, How the Power of Data will Drive EU Economy: The European Data Market Monitoring Tool Report, 2018.4.20.

다. 첫째는 생산 및 공급 프로세스 최적화를 들 수 있다. 이는 데이터에 기반을 두고 프로세스를 통해 생산을 효율화하는 내용이다.

둘째는 데이터를 활용해 표적 광고 및 개인 맞춤형 광고를 함으로써 마케팅 기술을 제고할 수 있는 것이다. 즉 타 산업에서 데이터 분석을 토대로 마케팅함으로써 수익 제고 효과가 발생한다는 점이다.

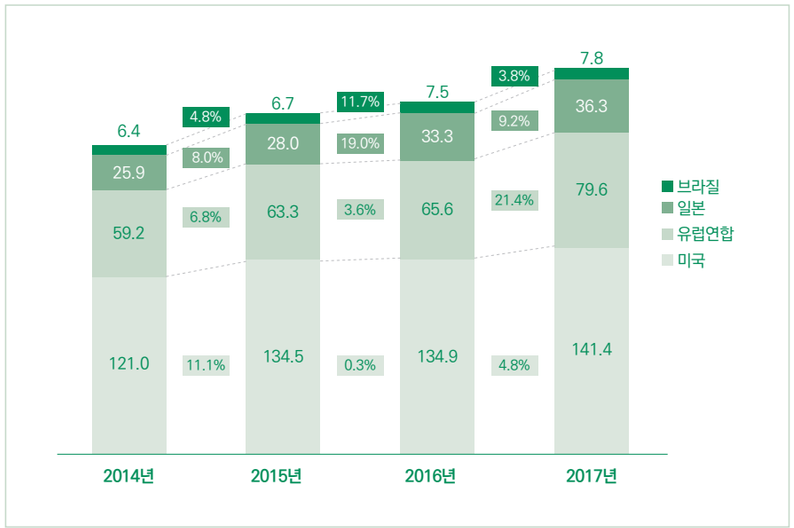
마지막으로 데이터 기반의 조직을 구성할 수 있는 점이다. 즉 데이터를 활용해 기존 조직의 관리 방식을 개선함으로써 불필요한 자원 소모를 최소화해 조직의 생산성을 높일 수 있는 점도 전방 효과라 할 수 있다.

전방 효과와 후방 효과로 나눠 간접 효과에 대해 설명했다. 이 글에서는 간접 후방 효과를 중심으로 소개하고자 한다. 데이터산업으로 인한 간접 후방 효과는 매우 빠르게 증가하고 있다. 미국의 경우 2014년 약 53억 달러 규모였으나 2017년에는 크게 증가해 90억 달러를 기록했다. 증가율 또한 매년 높게 나타나고 있다.

일본의 후방 효과도 빠르게 증가하는 추세다. 2014년에 약 10억 달러 규모였던 후방 효과는 지속적으로 증가해 2017년에 약 15억 달러로 집계됐다. 특히 2014년에서 2015년 사이에 20~30%까지 성장한 것으로 나타났다.

반면 EU의 후방 효과는 2016년부터 높은 성장세를 보이고 있다. EU의 데이

[그림 3-2-7] 2014~2017년 직접 효과와 간접 후방 효과를 합한 경제적 효과 (단위: 10억 달러)



출처: IDC & The Lisbon Council, How the Power of Data will Drive EU Economy: The European Data Market Monitoring Tool Report, 2018.4.20.

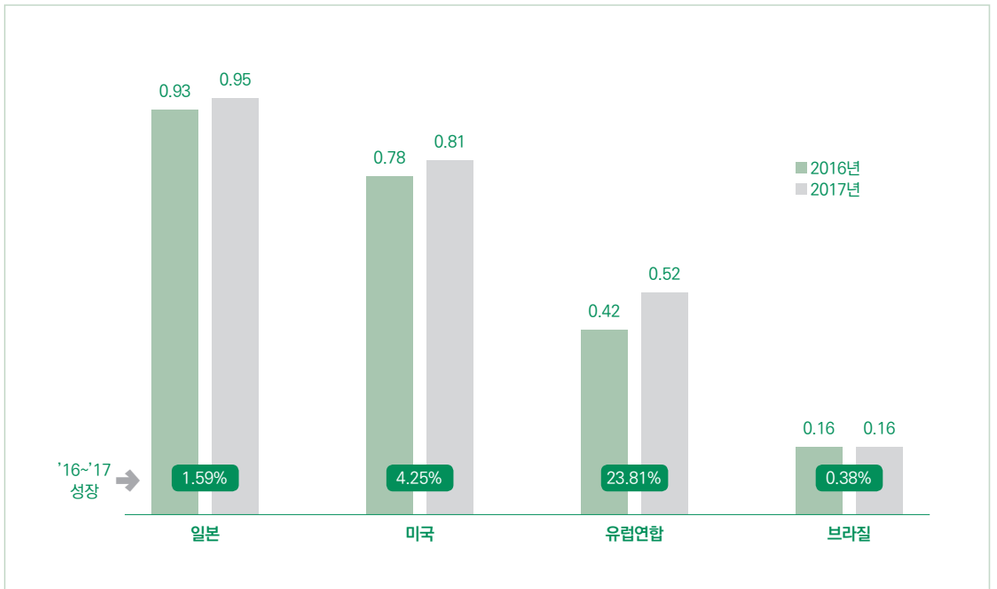
터산업은 공공데이터 개방 정책으로 민간데이터보다는 공공데이터가 많은 비중을 차지하고 있다. 또한 기업의 경우 1인 기업과 소규모 기업이 많아 고용창출 효과가 상대적으로 낮아 2016년까지는 성장세가 주춤했다. 하지만 2016년 이후 본격적으로 성과가 나타나기 시작하면서 급성장 국면을 맞이한 것으로 분석할 수 있다.

이 결과로 2016년에는 32억 달러, 2017년에는 38억 달러까지 성장한 것으로 나타났다. 특히 EU의 경우 직접과 간접 효과를 합하면 2016년에서 2017년의 성장세가 다른 지역을 압도한다.

4. GDP 대비 데이터 경제적 효과의 비율

앞서 데이터산업의 경제적 효과를 직접 효과와 간접 후방 효과를 중심으로 살펴봤다. 직접 효과와 간접 후방 효과 모두 미국이 가장 앞선 것으로 나타났다. 두 효과를 합했을 때 2017년 기준 약 1,414억 달러 수준이었다. 시장 규모와 마찬가지로 경제적 효과 또한 미국은 EU의 두 배 규모이지만, 성장 속도는 EU가 더 빠른

[그림 3-2-8] 2016~2017년 GDP 대비 경제적 효과의 비율 (단위: %)



출처: IDC & The Lisbon Council, How the Power of Data will Drive EU Economy: The European Data Market Monitoring Tool Report, 2018.4.20.

것으로 나타났다.

GDP에서 경제적 효과가 차지하고 있는 비율을 살펴보면, 미국은 2016년 0.78%에서 2017년 0.81%로 증가해 약 4.25%의 증가세를 보였다. 하지만 EU는 2016년 0.42%에서 2017년 0.52%로 같은 기간 미국보다 높은 23.81%의 성장세를 보였다. 이는 EU의 데이터산업이 미국보다 활발하게 전개되고 있음을 보여주는 수치다. 일본은 GDP 대비 데이터산업 비중이 2017년 기준 0.95%로 상승세를 타고 있다. 브라질은 0.16% 수준에 그쳐 데이터 활용 및 경제적 효과는 다소 미미한 것으로 나타났다

5. 글로벌 빅데이터 시장 동향

전 세계 빅데이터 시장 규모는 2018년 420억 달러에서 2027년에는 1,000억 달러를 넘어설 것으로 전망됐다. 2018~2027년 연평균성장률(CAGR)이 10.48%로 나타나 향후 10년간 두 자리 수 성장세를 지속할 것으로 전망된다.

[그림 3-2-9] 2017~2027년 글로벌 빅데이터 시장규모 추이 (단위: 10억 달러)

출처: Wikibon, Big Data Analytics Trends and Forecast 2018

빅데이터 시장의 성장은 전 세계 많은 기업이 빅데이터 분석의 유용성을 경험하고 있고, 업종 및 비즈니스 형태와 무관하게 디지털 트랜스포메이션을 이끄는 핵심으로 빅데이터 기술이 인정받는다는 데서 그 이유를 찾을 수 있다. 특히 빅데이터 기반의 고급 분석을 통해 기업의 비즈니스 경쟁력을 개선하고자 하는 수요와 맞물리면서 빅데이터 관련 수요는 큰 폭으로 상승하고 있다. 또한 자사의 고객들을 더 깊이 분석하고 비정형 데이터에 숨겨진 인사이트를 끌어내기 위해 빅데이터 관련 투자를 늘려 나갈 것으로 전망된다.

위키본(Wikibon)을 따르면 빅데이터 분석 소프트웨어 시장은 2018년 140억 달러 규모에서 2020년 200억 달러, 2026년 420억 달러를 넘어설 것으로 전망하고 있다. 2018년~2027년 기준으로 소프트웨어, 하드웨어, 서비스 부문별 연평균성장률을 보면 소프트웨어 14.13%, 하드웨어 8.01%, 서비스 8.38%로 타 부문 대비 소프트웨어 부문에서 높은 성장세가 전망된다.

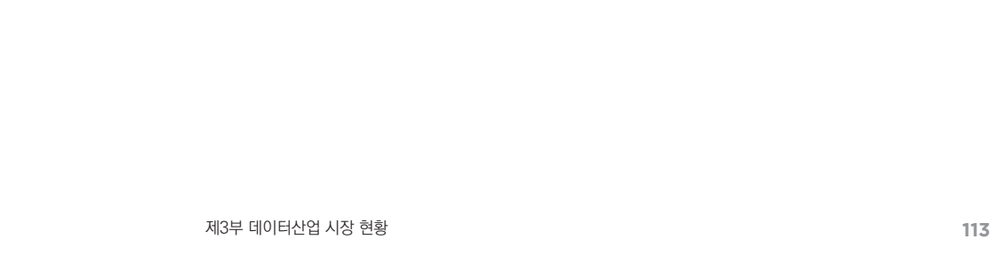
IDC 자료를 따르면 비관계형 분석 데이터 스토어, 인지·인공지능 소프트웨어 플랫폼, CRM 분석 애플리케이션 분야에서 높은 성장세를 보인다. 2016년~2021년 5년간 연평균성장률을 보면 각각 34.4%, 27.8%, 20.1%로 예상된다. 이 분야 기술이 향후 빅데이터 분야를 주도할 것으로 전망된다.

[표 3-2-2] 2018~2027년 요소별 빅데이터 시장규모 추이 (단위: 10억 달러)

구분	2018	2019	2020	2021	2022	2023	2024	2025	2026	2027년	CAGR '18-'27
소프트웨어	14	17	20	24	27	31	34	38	42	46	14.13%
하드웨어	12	14	15	16	18	19	20	22	23	24	8.01%
서비스	16	19	21	24	26	27	29	31	32	33	8.38%
합계	42	49	56	64	70	77	84	90	96	103	10.48%

출처: Wikibon, Big Data Analytics Trends and Forecast 2018, Worldwide Big Data Hardware, Software, and Services Revenue $B 2016-2027

[그림 3-2-10] 2016~2021년 5년간 상위 기술 분야 (단위: %)



출처: IDC, Worldwide Semiannual Big Data and Analytics Spending Guide 2018, Top Technology Category Based on 5 year CAGR(2016-2021) (Vendor Revenue (Constant),

가까운 곳으로부터 혁신

손삼수 | 웨어밸리 대표
2017 데이터 글로벌 이노베이터상 수상자

혁신은 늘 새로운 존재를 의미하지 않는다. 항상 실행하는 것 속에서 일어나는 작은 개선의 노력을 의미한다. 가지고 있는 기술을 다양한 곳에 적용해 보고, 새로운 가치를 창출하고, 응용·발전시키는 것이 혁신이다.

혁신이란 무엇인가? 없던 것을 만든다는 의미로 생각하기 쉽다. 그러나 혁신은 새로운 창조가 아닌 주변의 작은 개선에서 시작한다. 수레에서 시작된 혁신은 자동차로 발전됐다. 4차산업혁명시대에 이르러서는 자율주행차(autonomous car)로 진화하고 있다. 혁신은 기술만으로 이뤄지지 않는다. 주변의 세심한 관찰을 요구한다. 혁신자의 상상력이 필요하기 때문이다.

기술만으로 이뤄지지 않는 혁신

2016년 1월 클라우스 슈밥(Klaus Schwab)은 다보스포럼에서 '인간과 사물, 사물과 사물 간 네트워크를 통해 물리적 공간의 제약이 사라지는 사물인터넷(Internet of thing)이 미래의 도시를 혁신할 것'이라고 강조했다. 그는 특히 통찰력에 있어서의 도시의 큰 변화를 '혁신'으로 표현했다. 사물인터넷이 데이터 수집의 중요성을 의미한다면, 클라우드는 데이터 혁신이, 빅데이터는 데이터 분석이 그 핵심이다.

늘어나는 데이터를 효율적으로 분석하기 위해 기존 기간계 데이터를 분석계 DB로 복제하려는 요구가 늘어나고 있다. 이는 기간계 DB에 직접 분석 프로그램을 실행하면, 주요 업무 시스템에 영향을 줘 업무에 차질이 일어날 수 있기 때문이다. 또한 대부분의 기간계 DB가 전통적인 OLTP(OnLine Transaction Processing)여서 이를 분석용 OLAP(OnLine Analytical Processing) DB나 NoSQL 등 이기종 DB로 이관하려는 요구가 늘어나고 있다.

과거에는 이미 축적된 데이터를 기반으로 경영자의 의사결정을 내리는 데 필요한 분석, 즉 BI(Business Intelligence)가 필요했다. 하지만 4차산업혁명시대에 접어들면서 24시간×365일의 비즈니스 환경과 자율화한 현업조직의 활성화로 현재의 데이터를 기반으로 분석해 업무를 실시간으로 결정하고 적용· 반영해야 하는 리얼타임 BI가 필요해졌다. 이 때문에 실시간 복제에 대한 수요가 늘어나고 있다. 실시간 데이터와 이기종 데이터 간 처리에서 발생하는 데이터 복제에 대한 수요가 증가하고 있으나, 기존의 로그 기반 CDC(Changed Data Capture) 방식의 데이터 복제는 여러 가지 구조적인 문제와 한계를 갖고 있다. 현재 현업에서 사용중인 CDC 방식은 변경된 내용을 원본 DB의 로그파일에서 추출해, 이를 복제 DB로 옮겨서 반영한다. 복제 과정에서 지연이 발생하고, 추출과 전송 과정에서 원본 DB 서버에 설치된 에이전트가 CPU와 네트워크 부하를 일으킨다. 이 부하는 업무 처리에서 지연을 줄 수 있다. 특히 로그 파일이 없거나 분석할 수 없는 경우에는 이 CDC 방식을 사용할 수 없다.

새로운 존재를 의미하지 않는 혁신

이 시점에서 우리가 가진 최고 기술을 응용·적용함으로써 기존의 문제점을 크게 개선할 수 있다. 네트워크 패킷(Packet) 분석 기술이 바로 그것이다. 이 기술을 적용하면 원본 DB에서 실행되는 변경 데이터를 실시간 추출 분석해 복제 대상 DB에 반영함으로써 동기화 지연 문제를 해결할 수 있다. 또한 원본 DB에 에이전트를 설치하지 않기 때문에 원본 DB에 전혀 영향을 주지 않으면서 복제가 가능하다. 물론 로그 파일이 없어도 복제가 가능하다. 다양한 DB에 대한 패킷 분석 기술은 다년간 다져진 노하우를 바탕으로, 다양한 이기종 DB 간 복제를 가능하게 한다. 더불어 다양한 빅데이터 지원이 가능한 강점을 갖고 있다.

혁신은 늘 새로운 존재를 의미하지 않는다. 항상 실행하는 것 속에서 일어나는 작은 개선의 노력을 의미한다. 가지고 있는 기술을 다양한 곳에 적용해 보고, 새로운 가치를 창출하고, 응용·발전시키는 것이 혁신이다. 먼 곳에서의 '원용'이 아니라 가까운 곳으로부터 '직용'이 바로 혁신이다.

제 4 부

데이터 서비스 동향

제1장

데이터 거래 유형별 데이터 서비스 현황

그동안 내부 용도로 활용되던 데이터 거래가 새로운 비즈니스 아이템으로 자리잡고 있다. 이 글에서는 데이터 거래의 진화, 데이터 거래 서비스의 유형, 데이터 거래 발전 전망을 중심으로 데이터 거래와 데이터 서비스 현황, 향후 전망을 알아본다.

1. 데이터 거래의 진화

데이터 거래는 새로운 산업으로 자리잡고 있다. 전통적으로 데이터는 생성된 기업 또는 기관에서 자체 업무 목적으로만 사용돼 왔다. 디지털화가 진전되면서 데이터 사용은 크게 네 가지 방향으로 진화되고 있다.

첫째, 데이터가 기존 비즈니스 지원을 뛰어 넘어 새로운 비즈니스를 창출하는 수단이 되고 있다. 카드회사는 고객의 카드 사용 거래 데이터를 기반으로 B2C 기업에게 마케팅 서비스를 시작해 새로운 수익원으로 활용하고 있다. 기존 기업은 기존 비즈니스로 창출된 데이터를 발판으로 새로운 비즈니스를 만들어 내는 유형이다.

둘째, 데이터를 생성한 조직이 아닌 다른 경제 실체에도 가치를 주고 있다. 통신회사가 보유하고 있는 고객의 위치 데이터는 상권 분석에 활용된다. 또한 소액결제 회사의 개인 대금 납부 데이터는 신용정보 회사의 신용평가 모델에 활용된다. 기존 기업은 기존 비즈니스로 창출된 데이터를 다른 기업에게 제공함으로써 데이터 가치를 실현한다.

셋째, 데이터 공급자와 데이터 수요자가 다수 등장함으로써, 데이터 거래를 촉진하는 시장이 형성되기 시작했다. 또한 데이터 시장이 활성화되기 위해 필요한 서비스들이 개발되고 있다. 개인데이터 유통을 위한 비실명화 서비스, 거래 데이터 품질 수준을 알려주는 품질 인증 서비스 등이 도입되기 시작했다. 데이터 거래를 뒷받침하는 새로운 형태의 서비스가 시작되고 있다.

넷째, 데이터 거래의 탈중앙화 플랫폼이 등장하고 있다. 전통적인 데이터 거래에서는 실제 데이터 발생자이면서 데이터 소유자인 개인들은 아무런 혜택을 얻지 못했다. 또한 기업들은 데이터 소유자인 개인들의 동의를 얻는 데, 법적 규제 요건을 충족시키는 데 어려움을 겪어 왔다. 탈중앙화 플랫폼은 데이터 소유자인 개인들에게 직접 보상을 해 데이터 사용자들이 안전하게 데이터를 활용할 수 있도록 한다. 블록체인 기술을 이용한 새로운 기업들도 등장하고 있다.

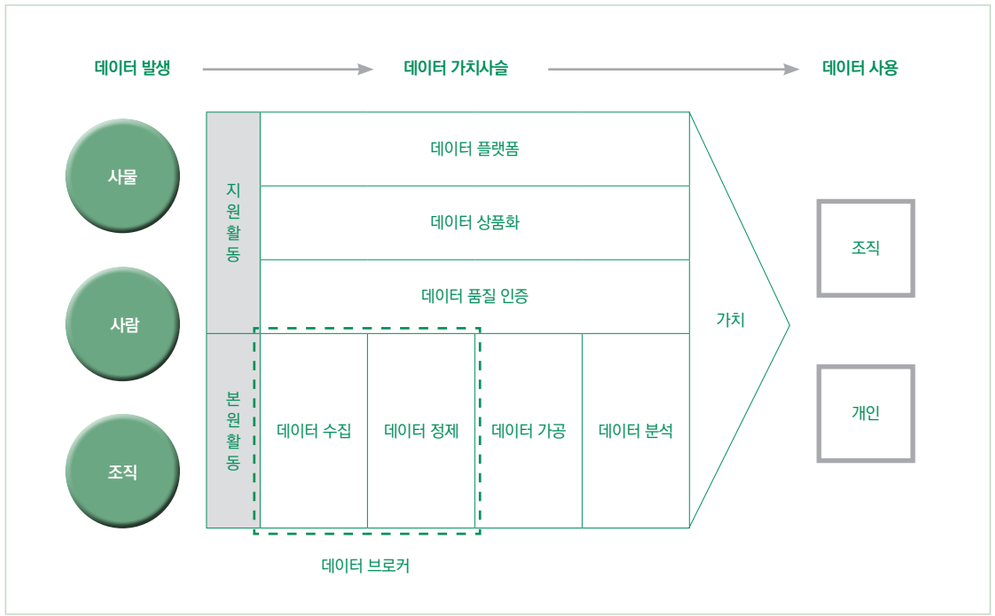
데이터산업은 아직 초기다. 여전히 데이터는 생성한 기업에서 주로 사용된다. 데이터산업이 자리잡기 위해서는 첫째, 데이터 공급자와 수요자가 거래 상대방을 쉽게 찾을 수 있어야 한다. 이는 데이터 마켓플레이스 기업들이 담당하게 될 것이다. 둘째, 데이터 수요자가 기대하는 데이터 품질을 확인할 수 있어야 한다. 데이터 공급자는 규제를 충족하면서 데이터를 제공할 수 있어야 한다. 거래 데이터 품질 인증 및 데이터 비실명화 기업들이 역할을 하게 될 것이다. 셋째, 데이터의 편중을 해소하고, 데이터의 거래 규칙을 수립하는 등 제도적 환경이 갖춰져야 한다. 이는 소비자 보호단체와 데이터 사용기업들을 포함한 사회적 합의와 법적 제도화가 필요하다.

2. 데이터 거래 서비스 유형

데이터 거래 서비스는 데이터 가치사슬에 따라 분류할 수 있다. 데이터는 발생부터 활용까지 일련의 활동을 통해 가치가 생성되고 실현된다. 본원 활동은 데이터의 가치 부가에 직접적으로 기여하는 활동이다. 데이터 수집, 데이터 정제, 데이터 가공, 데이터 분석 등이 해당된다. 지원 활동은 데이터 본원 활동을 도와주는 활동이다. 데이터 품질 인증, 데이터 상품화, 데이터 플랫폼 등이 이에 해당한다. 데이터 거래 서비스 유형은 데이터 가치사슬의 각 블록에 따라 구분할 수 있다.

• 필자 | **김인현**(투이컨설팅 대표)

[그림 4-1-1] 데이터 가치사슬에 따른 데이터 거래 유형



가. 데이터 수집 서비스

데이터를 필요로 하는 개인 또는 조직을 위해, 제3자가 데이터를 수집해 제공하는 서비스다. 데이터 수집 서비스는 초기에는 주로 개인데이터가 중심이었다. 디지털 사회가 진전되면서 데이터 수집대상이 확대되는 추세다.

특정 산업에 특화된 기업 데이터를 수집해 제공하는 서비스도 등장하고 있다. 의약품 온라인몰 '더샵'은 약국을 위한 빅데이터 서비스를 출시할 예정이다. 시범사업에 참여할 약국 100곳을 신청 받아서 지역 인구통계와 연계 분석해 약국의 상권 분석 데이터를 제공할 계획이다.

플리토는 언어데이터를 수집해 판매한다. 플리토는 집단 지성 플랫폼을 활용해 언어데이터를 생산해 바이두, 텐센트, 마이크로소프트 등 글로벌 기업에게 제공한다. 2018년 생산량은 2억 건, 판매량은 3000만 건으로 전망되고 있다.

솔트룩스의 '지니뉴스'는 인공지능 기술 기반의 개인 맞춤형 뉴스 서비스를 제공하고 있다. 인공지능을 이용해 하루 800만 건의 뉴스와 블로그를 읽어내고, 500여 카테고리로 자동 분류해 사용자별 맞춤 형태로 제공한다. 현재 사용자가 약 50만 명에 이르고 있다.

라이프시맨틱스는 '라이프레코드' 플랫폼으로 개인 건강기록 데이터를 수집해, 개인에게 맞춤 진료 등과 같은 헬스케어 서비스를 제공한다. 공공 보건 데이터, 병원 의무기록, 유전자 데이터, 사물인터넷 기기 등으로부터 데이터를 수집한다.

사물데이터 수집도 활발하게 추진되고 있다. KCC정보통신은 SK네트웍스 모빌리티 사업부문과 제휴해 사물인터넷 기반 커넥티드카 서비스를 준비하고 있다. 운행기록 자기 진단 장치(OBD, On-Board Diagnostics)를 이용해 운전 습관, 차량 운행 데이터 등을 수집해 운전자, 보험회사, 유통회사 등을 대상으로 다양한 서비스를 개발할 계획이다.

큐브인텔리전스는 MG손해보험과 블록체인 기반 자동차 데이터 활용 공동개발을 추진하기로 했다. 자동차 운행 시 발생하는 자동차 운행 데이터를 보험 상품에 활용하기 위한 것이다. 큐브인텔리전스의 큐브박스를 장착하면 자동차 데이터가 생성된다. 자동차 소유자들은 승인버튼을 누르면 블록체인 네트워크에 데이터가 올라가게 되고, 판매가 되면 코인으로 보상을 받게 된다.

유니마인드는 와이파이데이터를 수집한다. 스마트가이드플랫폼으로 수집한 와이파이 데이터와 카드사용 데이터를 연계 분석해, 지방자치단체에게 관광 및 축제 기획을 위한 인사이트를 제공한다. 이미 대전마케팅공사에게 백제문화권을 대상으로 서비스를 시작했다.

로플랫은 사용자 주변 와이파이 신호를 분석해 실내 위치를 알아내는 기술을 보유한 스타트업이다. 로플랫은 실내에서 위치 측정이 가능한 장소 정보를 32만 곳 이상 수집했으며, 2018년까지 100만 곳 이상으로 확대할 계획이다. 현재 월 사용자는 100만 명 정도이며, 매일 수집하고 분석하는 고객 위치 정보는 60만 건에 달한다.

기업이 아니라 개인이 데이터를 수집해 판매하기도 한다. 아이돌에 관한 이미지, 영상, 이력 등 데이터를 수집하는 것을 '아이돌 덕질'이라고 한다. 수집된 데이터를 트위터 등을 통해 구매자를 찾아서 데이터를 판매하는 것을 '데이터팔이'라고 부른다.

나. 데이터 정제 서비스

데이터를 이용하려는 개인 또는 조직을 위해 데이터 품질을 확보해 주는 서비스다. 데이터를 사용하기 위해서는 데이터 품질이 확보돼 있어야 한다. 많은 기업들

이 데이터 품질 불량으로 비용 손실을 보고 있다.

대표적으로 개인 및 법인의 주소 데이터 오류를 꼽을 수 있다. 수지원넷소프트는 비정형으로 수집되는 고객 주소를 목적에 맞게 활용할 수 있도록 정제하는 서비스를 제공한다. 지번 주소를 도로명주소로 변환하는 서비스를 행정안전부에 공급했다.

개인의 사망 데이터도 오류가 많은 항목이다. 개인이 사망하면 행정기관에 신고는 하지만, 개별 기업에는 알려지지 않기 때문이다. 기업들은 고객의 생존 데이터 오류로 인한 비용 부담이 크다. 한국통계진흥원은 환자의 정보이용 동의가 포함된 식별정보 위탁처리 신청서, 기관심사위원회 승인서를 제출하면 사망자 식별 자료와 사망 일자를 제공한다. 개인정보보호에 의해 사망 여부 데이터 확보가 쉽지 않은 것이 현실이다.

데이터 정제는 지금은 데이터 사용 조직이 직접 수행하는 경우가 대부분이다. 개별로 수행되는 데이터 정제 노력은 산업 차원에서는 상당한 규모다. 데이터 정제가 서비스로 자리잡기 위해서는 개인정보보호 관련 법규가 완화돼야 한다. 현재로서는 필요성은 크지만 기업 서비스로서 활성화되기는 쉽지 않다.

다. 데이터 가공 서비스

데이터를 분석하는 개인 또는 조직을 위해 분석에 활용할 수 있도록 데이터 형태를 가공해 주는 서비스다. 데이터로부터 의미 있는 가치를 끄집어 내기 위해서는 먼저 데이터 전처리를 해야 한다. 이를 '데이터 프렙(Data Prep, Data Preparation)'이라고 한다. 데이터 분석 작업에서 80% 이상의 노력이 데이터 전처리에 투입된다.

데이터몬스터즈는 데이터 수집, 전처리 및 분석 설계, 분석, 시각화 등의 단계를 지원한다. 과학기술정책연구원, 기상청, 한국교육개발원 등의 데이터를 가공해 분석하는 서비스를 제공했다. 데이터 전처리를 포함한 데이터 분석 서비스를 제공한다.

데이터가공 서비스는 독자적인 데이터 거래 형태로 이뤄지고 있지는 않다. 현재 데이터 전처리는 데이터 분석 기업에서 데이터 사이언티스트들이 스스로 수행하는 상황이다. 데이터 보안 문제로 보유하고 있는 데이터를 외부에 반출하는 것이 쉽지 않기 때문이다.

라. 데이터 분석 서비스

데이터로부터 인사이트를 찾고자 하는 개인 또는 기업을 위해 데이터를 분석해주는 서비스다. 데이터 분석 서비스는 '어떤 데이터를 분석하는가', 그리고 '분석 결과를 누구에게 제공하는가'에 따라 세 가지 유형으로 분류된다.

1) 고객 데이터를 분석해 고객에게 제공

개인 또는 기업이 보유한 데이터를 분석해 데이터 보유자에게 의미 있는 분석 결과를 제공하는 서비스다.

레이니스트는 '뱅크샐러드'를 통해 개인의 카드 거래 데이터를 수집하고, 이를 분석해 가장 유리한 카드 상품을 추천한다. 이를 위해 카드회사의 웹페이지를 스크래핑해 카드 상품 데이터베이스를 구축하고 있다. 사용자는 자신의 데이터를 제공하고, 레이니스트는 이를 분석해 최적의 금융상품을 활용함으로써 혜택을 받게 된다.

한국신용데이터는 소상공인 카드 매출 내역을 관리해주는 '캐시노트'를 출시했다. 카드 전표를 추적해 결제 내역을 카카오톡으로 알려준다. 카드 매출 및 결제 내역 데이터를 분석해 매출 분석, 손님 유입 분석, 재방문 분석 등의 서비스도 제공하고 있다. 2018년 현재 8만 2,000개의 가맹점이 캐시노트를 이용해 매출을 관리하고 있다.

닥터퀀트는 중소기업벤처부의 개발자금을 지원받아서 금융건강검진 서비스를 개발했다. 투자자가 객관식 질문 30개와 주관식 질문 6개를 대답하면 투자자의 투자 성향을 포함한 자산과 부채의 적정성 분석 등을 제공한다. 8만 원을 한 번만 결제하면 매년 무료로 서비스를 이용할 수 있다. 투자자로부터 데이터를 받아서 분석 서비스를 제공하는 방식이다.

네이버는 자체 쇼핑 플랫폼에 입점한 판매자들을 위한 '스마트스토어' 플랫폼을 출시했다. 비즈어드바이저(Biz Advisor)를 도입해 판매자들에게 상품별 판매 성과, 고객 기본 정보, 인공 지능 기술을 활용해 추정한 고객의 결혼 유무, 가구 인원, 직업 등의 정보를 제공한다.

삼성카드는 가맹점을 대상으로 빅데이터를 활용한 리서치 서비스인 '리얼타임'을 제공하고 있다. 가맹점의 고객에게 설문문자를 보내고 응답 결과를 분석해 가맹점에게 제공하는 서비스다. 2017년 약 20곳에서 서비스를 이용했다.

행정안전부 국가정보자원관리원은 정부 빅데이터 공통기반 시스템인 '혜안(慧眼)'에 분석 과제를 의뢰할 수 있는 '상시 창구'를 개설해 각 정부기관의 정책현안 분석을 지원하고 있다. 경찰청의 범죄 현장기록서인 '임장(臨場)일지' 데이터 분석, 제주시 시티투어버스 노선 분석, 대전시 공영자전거 '타슈' 대여소 위치 분석 등을 지원했다.

2) 보유 데이터를 분석해 제3자에게 제공

기존 비즈니스에서 생성된 데이터를 기반으로 정보 제공 및 분석 서비스를 원래 데이터 생성자가 아닌 제3자에게 제공하는 서비스다.

이 영역은 카드사와 신용평가사(CB, Credit Bureau)가 대표적이다. 금융위원회는 '금융분야 데이터 활용 및 정보보호 종합방안'을 발표해, 카드사가 보유한 방대한 데이터를 활용할 수 있도록 빅데이터 분석 컨설팅 업무를 부수 업무로 허용했다. 또한 통신료 등의 비금융정보를 모아서 신용 등급을 매기는 크레딧뷰로(CB) 설립도 가능하게 된다.

비씨카드는 데이터 컨설팅, 데이터 융합분석, 데이터 솔루션 등 세 가지 빅데이터 사업을 운영하고 있다. 데이터 컨설팅은 컨설팅을 의뢰한 업체에게 소비데이터를 분석해 제공하는 서비스다. 서울특별시, 경상북도, 인천관광공사 등이 이용한 실적이 있다.

신한카드는 기업과 지자체 등을 위한 마케팅 컨설팅을 제공하고 있다. 기업에는 보유 고객 데이터를 분석해 상품별 타깃 고객 리스트 제공을 통해 마케팅 적중율을 높일 수 있도록 한다. 지자체에는 지역 내 창업 업종 추천, 청년 창업 지원 계획 수립 등을 위한 분석 결과를 제공한다. 신한카드의 마케팅 컨설팅은 별도로 컨설팅 수수료를 받음으로써 데이터 기반 비즈니스 모델로 진화하고 있다.

신한은행은 3억 건의 금융 거래 데이터를 분석해 서울시의 소비 규모와 패턴 등을 한눈에 볼 수 있도록 한 '서울시 생활 금융지도'를 공개했다. 신한은행 빅데이터센터는 전국 주요 지역별 고객의 소득, 소비, 저축 등 금융생활 현황과 트렌드를 파악할 수 있는 지역별 생활 금융지도를 만들어 공개할 예정이다.

비즈니스온커뮤니케이션은 빅데이터 분석 서비스인 '스마트 MI(Marketing Intelligence)'를 운영하고 있다. 2007년 설립된 이후 전자세금계산서, 전자계약 등을 포함한 전자 문서 유통서비스인 '스마트빌(Smart Bill)'로 축적된 350만 고객사의 상거래 데이터를 분석해 거래처의 매출 등 재무상황을 분석하고 조기 경보를 보내주는 거래처 리스크 관리 서비스를 제공하고 있다.

코스콤은 이미 보유하고 있는 주식 등 유가증권 거래 데이터를 기반으로 SNS, 블로그, 카페, 뉴스 등 소셜미디어 데이터를 수집해 투자 심리 수준을 제공하는 빅데이터 분석 정보 서비스를 상용화했다. 인공지능 투자자문 소프트웨어인 '티레이더'를 개발한 유안타증권과 제휴해 소셜미디어 빅데이터 분석 기반의 증권 투자정보 서비스 사업을 공동으로 추진할 계획이다.

KG모빌리언스는 자체 확보하고 있는 사용자 결제 데이터를 기반으로 '모다스(Modas)'를 개발했다. 모다스는 고객 결제 내역 데이터를 기반으로 인구 통계학과 구매자 라이프 스타일 등의 데이터를 활용해 분석 정보를 제공한다. 모다스를 이용하면 고객군별 소비 성향을 분석함으로써 정교한 마케팅과 광고를 수행할 수 있다.

잡플래닛은 인사 빅데이터를 전문으로 분석하는 에이치알랩스(HR Labs)를 설립했다. 잡플래닛이 보유하고 있는 기업 문화, 채용, 연봉, 복지 등의 데이터를 분석할 예정이다. 150만 개가 넘은 기업의 HR 데이터를 분석해, 기업 문화 진단, 채용 경쟁력 분석, 연봉 및 복지 경쟁력 진단 등의 서비스를 제공하고 있다.

'사람인'은 보유하고 있는 연봉 데이터에 국민연금, 나이스평가정보, 알리오, 고용보험 등의 연봉 정보를 포함해 총 526만 개의 연봉 DB를 확보했다. 이를 분석해 개인의 속성과 비교해 본인의 연봉 수준을 다양한 관점으로 확인할 수 있는 서비스를 제공한다.

3) 데이터를 수집해 분석한 후 제3자에게 제공

제3자를 위해 스스로 데이터를 수집해 분석한 후 제공하는 서비스다. 데이터를 수집하는 능력과 데이터 사용 기업의 비즈니스를 완전하게 이해하는 역량이 중요하다.

P2P 회사인 어니스트펀드는 자체 개발한 신용평가모형을 이용해 중금리 대출에 집중해 왔다. 고객의 금융 데이터뿐 아니라, 외부의 소셜 데이터와 모바일 이용 데이터, 어니스트펀드의 홈페이지 로그 등의 데이터를 수집해 신용평가모형의 정확성을 높일 계획이다. 어니스트펀드의 진화한 신용평가모형은 자체 사용 목적을 뛰어 넘어서 외부에 정보를 제공하는 서비스가 될 전망이다.

모바일리서치 회사인 오픈서베이는 스마트폰 기반 소비데이터 분석 기술을 보유한 해빗팩토리와 업무 협약을 체결했다. 오픈서베이는 기업 및 개인을 위한 모바일 설문조사를 대행하는 회사다. 앞으로 해빗팩토리와 제휴해 단순 설문 데이터 수집을 뛰어넘어 데이터 분석 서비스로 확장할 계획이다.

데이터블은 데이터를 기반으로 하는 인플루언서 마케팅(Influencer marketing) 스타트업이다. AI 기반 데이터 분석 서비스인 '해시업AI'를 통해 인플루언서 마케팅을 시도하고 있다. 지난 해 말부터 국내 인플루언서 6,000명을 대상으로 테스트한 결과 38%의 마케팅 효율 증가를 보여줬다고 한다. 헤시업AI 엔진은 인스타그램 사용자 데이터를 수집해 사용자 관계망을 파악하는 엔진으로, 이를 통해 데이터를 수집하고 인플루언서 영향력을 분석한다.

NHN엔터테인먼트의 디지털 광고부문 자회사 NHN ACE는 로그데이터를 수집해 마케팅 인사이트를 제공하는 서비스를 제공하고 있다. 서비스는 마케팅 컨설팅과 솔루션 제공의 두 가지 방법이 가능하다. 마케팅 컨설팅에서는 인바디 앱, 숙박 앱 등의 오디언스 타기팅 서비스를 수집한 로그 분석을 토대로 진행했다. NHN ACE의 '에이스카운터(ACE Counter)'는 유로로 기업의 웹과 앱 로그 분석을 제공하는 플랫폼이다.

마. 데이터 상품화

데이터가 쉽게 유통될 수 있도록 데이터를 상품화해 제공하는 서비스다.

데이터 상품화는 오픈API(Application Programming Interface) 형태로 이뤄지는 것이 일반적이다. 데이터를 기반으로 하는 스타트업들은 다양한 데이터를 필요로 하지만, 일일이 수집하기는 쉽지 않다. 데이터 상품화 기업은 이들을 위한 데이터를 제공한다.

미국의 대표적인 오픈API 기업은 Xignite가 있다. Xignite는 금융회사들이 필요로 하는 자본시장 데이터를 수집해 API 형태로 제공한다. 현재 대부분의 핀테크 회사들과 기존 금융회사들을 포함해 1,000개 이상의 고객사를 확보하고 있다. Xignite의 데이터는 우리나라 핀테크 회사들도 이용하고 있다.

우리나라 오픈API를 이용한 데이터 거래는 금융산업에서 처음 시도됐다. 키움증권이 오픈API를 통해 고객이 자신의 기호에 맞는 주식매매 앱을 만들 수 있게 지원하고 있다. 키움증권은 2014년부터 오픈API 서비스를 시작했고, 2017년에는 키움API를 이용한 주식 매매 앱 경진대회를 개최했다.

NH농협은행은 130종의 API를 서비스하고 있다. 2018년 5월 말까지 API 누적 거래금액은 7,449억 원으로 전년 동기 대비 733%, 거래 건수는 121만 건으로 170%가 각각 증가했다. 월 기준 거래 금액은 2,000억 원 수준이다.

키움증권과 NH농협은행의 API는 엄밀한 의미에서 순수한 데이터 거래는 아니다. 금융회사가 할 수 있는 서비스를 API를 통해 제공한 것이라고 할 수 있다. 하지만 기본적인 수행 방식은 데이터 교환이다. API가 확산되면 데이터 상품화도 촉진될 것으로 전망된다,

KTH의 API 스토어는 API뿐 아니라 DB와 콘텐츠를 API 형태로 거래할 수 있는 국내 최초의 API 유통 플랫폼이다. 일반 개발자, 스타트업, 기업의 DB 및 콘텐츠를 간편하게 사고팔 수 있다. 약 3,000종의 API를 제공하고 있다. API 개발자와 API 사용자가 서로를 쉽게 찾고 거래가 이뤄질 수 있도록 하는 형태를 갖추고 있다. API 유통량은 2013년 2억 건에서 2017년 195억 건으로 98배 늘었다.

개인정보를 데이터 상품화하기 위해서는 개인의 프라이버시를 보호할 수 있도록 비식별화를 수행해야 한다. 파수닷컴은 자체 개발한 비식별화 기술을 이용해 개인 정보를 상품화하려는 시도를 했었다. 온라인 쇼핑몰, 병원, 금융기관 등 10여 곳과 사업을 추진했으나, 2017년 말에 중단했다.

파수닷컴은 정부가 발표한 비식별화 가이드라인에 따라 사업을 추진했지만, 일부 시민단체가 한국인터넷진흥원 등 4개 기관과 현대자동차, SK텔레콤, 삼성생명 등 20개 기업을 개인정보보호법 위반 협의로 검찰에 고발했기 때문이다. 개인정보 비식별화를 통해 상품화하는 것은 아직까지는 법적으로 불가능하다고 판단된다.

바. 데이터 품질 인증

데이터 수요자가 외부 데이터를 믿고 활용할 수 있도록 데이터 품질을 확인해주는 서비스다. 데이터 수요자는 구입한 데이터의 품질 수준에 따라 자체 투입해야 하는 노력과 비용이 달라지게 된다. 하지만 데이터를 구입해 열어 보기 전에는 데이터가 어떤 상태인지 알 수 없다. 데이터 수요자가 데이터 품질을 알 수 없다면, 적정 거래 금액을 판단할 수 없고 결과적으로 데이터 거래가 이뤄지기 어렵게 된다.

기존의 데이터 품질 인증은, 자신의 데이터를 자신이 사용하거나 또는 자신의 데이터를 외부에 공개하기 위한 판단 기준으로 사용돼 왔다. 데이터 거래 관점에서는 데이터 수요자가 데이터 사용 가능성과 비용 타당성을 확인할 수 있는 기준을 제공해야 한다. 현재 우리나라에서는 데이터 품질을 인증해주는 기업이나 기관은 없다.

한국데이터진흥원은 품질이 낮은 데이터 구매로 인한 피해를 막고, 안전한 데이터 유통 구조를 만들기 위해 거래용 데이터 등급 평가 모델을 개발하고 있다. 2019년부터 누구나 이용할 수 있게 할 예정이다. 구체적으로는 △평가지표 개발·평가수준 계량화를 통한 등급체계 수립 △평가항목별 이행표준·자율점검표 개발(데이터 판매자용) △평가 매뉴얼 개발(평가자용) 등을 진행하고 있다.

사. 데이터 플랫폼

데이터 발생자, 데이터 공급자, 데이터 수요자 들을 연계해 데이터 거래가 일어날 수 있는 기반을 제공하는 서비스다.

1) 공급자 중심 마켓플레이스

데이터를 보유하고 있는 공공기관 또는 기업이 데이터를 공개하는 방식이다. 일반적으로 오픈 데이터로 불린다. 보유하고 있는 데이터를 제3자가 필요로 하고 활용할 것을 기대하기 때문에 공개한다. 공급자 중심의 공개로 사용자 관점이 반영되기 어려운 점이 있다.

대표적으로는 공공데이터포털(www.data.go.kr), 데이터스토어(www.datastore.or.kr), 공간데이터 플랫폼, 농산물 유통정보, 금융감독원 금융상품통합비교 등이 있다. 공공기관이 중심이 돼 독자 사이트에 데이터를 공개하는 방식이다. 실제로 대부분의 중앙부처, 공사, 지방자치단체 들은 각자 개별적으로 데이터를 공개하고 있다. 이를 모아서 공개하는 사이트는 공공데이터포털과 데이터스토어가 있다.

데이터스토어는 공급자 중심 마켓플레이스에서 오픈 마켓을 지향하는 '개방형 데이터스토어'를 시도하고 있다. 개편된 개방형 데이터스토어는 전 세계적인 데이터 개방·공유 추세에 발맞춰 CKAN(Comprehensive Knowledge Archive Network) 기반 플랫폼으로 구축했다. 공공·민간의 데이터 공유·연계 및 결제기능을 통한 데이터 판매가 가능하다. 한국전자통신연구원(ETRI)의 기술 지원으로 지난 4월 23일 오픈했다. CKAN은 오픈소스 기반 데이터 플랫폼으로 미국, 영국 등 31개 중앙정부를 포함 총 146개의 지방정부, 학술 단체 등에서 사용한다.

민간기업도 보유 데이터 공개와 공유를 위해 노력하고 있다. SKT는 2013년부터 보유 데이터를 공개하는 '빅데이터허브'를 운영하고 있다. 2013년 10월 공개 데이터 10건에서 2017년 7월 현재 867건으로 확대됐다. 이용 신청 건수는 2017년 6월말 기준 1만 1,000건을 넘어섰다. 주로 지자체와 협업 프로젝트로 진행되는 경우가 많으며, 현재 40여 지자체에 데이터를 공급하고 있다.

공급자가 보유하고 있는 데이터를 공개하는 방식은 활용도 제고 측면에서 유리하지 못하다. 수요자 관점을 반영하고 있지 않기 때문이다. 데이터 수요자가 모든 데이터 공개 사이트를 돌아다니면서 필요한 데이터를 찾는 것은 노력도 많이 들고, 필요한 데이터를 찾을 가능성도 높지 않다. 대부분의 공급자 중심 데이터 마켓플레이스는 공개한 데이터 활용도 제고라는 과제를 안고 있다.

2) 플랫폼 마켓플레이스

데이터 수요자와 공급자가 함께 모여서 데이터를 거래하는 기반을 뜻한다. 플랫폼 마켓플레이스는 초기 단계로 새로운 시도가 이뤄지고 있다.

KB국민카드는 작년 12월 지디에스컨설팅그룹과 빅데이터 브로커리지 비즈니스 추진을 위한 업무 협약을 맺었다. 2019년 1월 '빅데이터 중개·거래 플랫폼'으로 기업과 개인에게 빅데이터 수집, 가공, 중개 및 보유 데이터 가공 및 구매 데이터 통합 서비스 등을 제공할 예정이다. 또한 수요에 맞춘 주문형 빅데이터 서비스도 제공할 예정이다.

CJ올리브네트웍스는 CJ프레시웨이의 B2B 식자재 데이터와 CJ멤버십 서비스의 데이터를 분석해 식품관련 데이터 상품을 개발하고, 데이터 거래 및 중개를 지원하는 사업을 2017년 12월까지 수행했다. 2018년에는 식자재 공급사와 요식업체 등에게 식자재 공급과 수요 데이터를 포털 형태로 서비스 할 계획이다.

산업통상자원부는 병원 데이터를 표준화를 통해 수집해 기업들이 바이오 빅데이터를 활용할 수 있게 할 계획이다. 바이오 빅데이터 플랫폼 사업 추진단은 30개 이상의 병원이 제공하는 5,000만 환자의 진료 정보를 거래하는 '바이오 빅데이터 플랫폼'을 구축해 제약, 의료기기, 건강, 보험, 식품 등 관련 기업에게 데이터를 제공할 계획이다.

LG CNS는 IoT 데이터를 수집하는 플랫폼 인피오티(INFioT)를 출시했다. LG CNS 인피오티는 가정용 전자제품 같은 홈 IoT부터 자동차 안의 전자기기, 공장의 제조 설비, 교통수단, 빌딩 등 기업 및 공공 IoT에 이르는 다양한 종류의 IoT 기기로부터 데이터를 수집할 수 있다. 인피오티는 개발자 포털도 제공한다. 개발자들은 인피오티에 접속해 IoT 기기를 등록하고 수집할 데이터를 지정하면 데이터 수집 과정을 확인할 수 있다.

플랫폼 마켓플레이스는 수요자의 관점으로부터 데이터 수집과 거래가 시작된다는 점에서 데이터 활용이 더 활발하게 이루어질 수 있는 비즈니스 모델이다. 하지만 개인정보 보호는 여전히 과제로 남는다.

3) 탈중앙화 플랫폼

데이터 플랫폼에는 데이터 발생자의 이해관계가 고려돼 있지 않다. 개인데이터의 발생자는 실제로 데이터를 발생시키는 인격체이다. 데이터 수요자가 필요로 하는 대부분의 데이터는 개인과 관련돼 있다. 개인은 자신의 데이터를 다른 사람이 수집해 판매하고 이익을 차지함으로써, 거래 관계에서 아무런 혜택도 받지 못하고 있다. 도리어 개인정보의 오용과 남용에 따른 피해를 보는 경우가 많다.

데이터 거래가 원활하게 이뤄지기 위해서는 데이터 발생자가 자신의 데이터를 확인할 수 있고, 통제할 수 있어야 하며, 데이터 거래 이익의 주체가 돼야 한다. 이러한 전제가 이뤄지면 개인정보보호 이슈는 더 이상 문제가 되지 않을 것이다. 데이터 발생자인 개인이 주도적으로 데이터 활용에 참여한 결과이기 때문이다.

데이터 발생자, 즉 데이터 주인인 개인이 결정권을 갖고 경제적 이익의 주체가 되도록 하기 위해서는 새로운 기술이 필요하다. 블록체인이 데이터 거래의 혁신 기술로 떠오르고 있다. 데이터 플랫폼이 아니라 데이터 생태계가 데이터 거래의 시장이 되며, 데이터 생태계는 데이터 거래에 참여하는 주체들의 공동 의사결정에 따라서 운영된다. 플랫폼 기업이 별도로 존재하지 않기 때문에 탈중앙화 플랫폼이라고 부른다.

탈중앙화 플랫폼에서는 생태계에 기여한 모든 주체는 보상으로 코인을 받는다. 데이터는 블록체인 형태로 주고 받는다. 모든 주체는 자신의 블록체인을 확인할 수 있고 통제할 수 있다. 코인은 생태계 안에서 구매 수단으로 활용할 수 있고, 이더리움이나 비트코인 같은 암호화 화폐로 교환돼서 생태계 바깥의 경제 생활에 사용할 수 있다.

Ab180의 에어블록은 앱 로그와 같은 개인데이터 활용을 코인으로 보상하고 광고주에게는 직접 마케팅을 가능하도록 하는 블록체인 생태계를 추진하고 있다. 펜타시큐리티시스템은 자동차 데이터를 기반으로 하는 블록체인 네트워크를 만들 계획이다. 코인원은 스트리머 플랫폼에서 발생하는 데이터를 사고 파는 데 사용되는 토큰을 암호화폐로 상장시키기로 결정했다. 메디블록은 개인의 건강 관리 데이터를 블록체인에 담아 거래하는 생태계를 구축했다. 알파콘은 헬스케어 데이터 유통 및 개인 맞춤형 건강관리를 제공하는 블록체인 생태계다. 마이크레이딧 체인은 개인신용정보 모델을 블록체인 기반으로 추진하고 있다. 이지식스의 엠블은 불록체인에 기반한 자동차 데이터 생태계를 구축할 계획이다.

3. 데이터 거래 발전 전망

2018년은 데이터 거래 시장이 본격적으로 열리기 시작한 해다. 데이터 거래와 관련된 법과 제도가 아직 갖춰지지 않았지만, 데이터 활용 수요는 급속하게 커지고 있다. 또한 새로운 데이터 주도 비즈니스 모델도 다양하게 등장하고 있다. 스타트업들은 데이터가 없이는 성립할 수 없다. 스타트업을 포함해 기존 기업들의 경쟁력도 어떤 데이터를 얼마나 확보해 활용하느냐에 달려 있다.

EU 지역은 2018년에 GDPR, PSD2 등이 발효됨으로써 데이터 거래의 산업화를 위한 기반을 갖췄다. 우리나라도 관련 부처에서 필요성을 깊이 인식하고 추진하고 있어서 데이터 거래 산업의 제도화는 빠르게 이루어질 것으로 전망된다.

- **데이터 거래의 주권을 개인이 갖게 된다**

개인은 자신이 발생시키는 모든 데이터에 대해서 완전한 권한을 확보하게 될 것이다. 기업은 개인데이터를 임시로 보관하고 있을 뿐이다. 개인이 원하면 데이터를 전달해야 하고, 또한 삭제해야 한다. 기업은 개인데이터를 어떤 용도로든 사용하게 되면 데이터 발생자인 개인에게 비용을 지불해야 한다.

• **데이터 마이닝이 본업이 될 수 있다**

자동차 산업은 자동차 매출 규모보다 자동차가 발생시키는 데이터 매출 규모가 더 커질 것이라는 전망도 있다. 은행은 더 많은 데이터를 갖기 위해 여신 또는 수신을 하게 될 지도 모른다. 개인은 더 많은 데이터를 발굴하면, 더 많은 급여를 받을 수 있다. 기업과 개인 모두에게 데이터를 생성시키고 관리하는 것은 가장 중요한 경제활동이 될 것이다.

• **데이터 큐레이터가 필요하게 될 것이다**

데이터는 스스로 말하지 않는다. 데이터의 의미를 깨달아야 하고 데이터 본질에 맞게 활용해야 가치를 이끌어낼 수 있다. 데이터 큐레이터는 데이터가 어디에 있고, 어떤 과정으로 만들어졌으며, 어떻게 활용할 수 있는가를 알려주는 역할을 하게 될 것이다. 개인과 기업은 점점 더 외부 데이터에 의존하게 될 것이고, 데이터 큐레이터 역할이 더 중요해질 것이다.

• **블록체인과 코인 이코노미가 데이터 거래의 중요 기술이 될 것이다**

개인이 자신의 데이터가 어떻게 유통되고 있으며, 누가 어떻게 활용하고 있는가를 확인하고 통제하는 기술이 필요하다. 또한 데이터 사용에 대한 대가를 받을 수 있는 메커니즘도 필요하다. 데이터 보안과 활용에는 신뢰가 확보돼야 한다. 이러한 요건들을 충족시킬 수 있는 기술로는 블록체인과 코인이 유일하면서도 거의 완벽하다.

• **오픈API를 통한 데이터 유통이 확산될 것이다**

경제 실체들이 수시로 데이터를 주고 받는 세상이 될 것이다. 사람이 개입하지 않더라도 사물들이 데이터를 생성하고, 전달하며, 활용하게 될 것이다. 데이터 유통 수단으로는 오픈API가 유일한 방법으로 자리잡은 상태이다. 앞으로 기업들은 더 많은 오픈API를 만들어내고 더 많은 접속과 사용이 이뤄지도록 하기 위해 노력할 것이다.

제2장

주요 데이터 서비스 동향

주요 데이터 서비스 동향을 금융데이터, 의료데이터, 학술데이터, 기업데이터, 특허데이터 순으로 알아본다. 제2장 주요 데이터 서비스 동향은 각각의 필자가 집필했다.

1. 신용데이터 분야 서비스

신용정보는 금융거래 등 상거래에 있어서 거래 상대방의 신용도와 신용거래능력 등을 판단할 때 필요한 정보로, 그 중요성이 날로 커지고 있다. 신용정보는 개인과 기업의 금융 활동에 많은 영향을 주게 되므로 신용을 평가하는 금융회사 등은 신용정보를 통해 더 정교하게 개인과 기업을 평가하려는 노력을 기울이고 있다. 반대로 신용평가를 받는 개인과 기업은 신용관리를 철저히 해 좋은 신용평가를 받을 수 있도록 노력을 기울이고 있다.

또한 신용정보는 안전하게 비식별화해 공공정책이나 행정 효율화 분석에 활용함으로써 다양한 공익적 가치를 발견할 수 있고, 상권분석 등에 활용함으로써 자영업자의 사업에 도움을 주는 등 빅데이터 시대에 큰 이바지를 할 것으로 기대된다.

가. 개요

1) 신용정보의 정의와 분류

신용정보는 금융거래 등 상거래에 있어서 거래 상대방에 대한 신용도와 신용거래능력 등을 판단할 때 필요한 정보다. 식별정보, 신용거래정보, 신용도 판단정보, 신용거래능력 정보, 공공정보 등이 이에 해당한다.

신용정보는 개인신용정보, 기업신용정보, 기술신용정보로 분류할 수 있다.

• 1절 필자 | **이욱재**(코리아크레딧뷰로 본부장)

개인신용정보는 생존하고 있는 개인의 정보로 사망자 정보는 개인신용정보에 해당하지 않는다. 또한 신용정보를 부정적 정보와 긍정적 정보로 분류할 수 있다. 부정적 정보(Negative Information)는 연체, 부도 등 신용정보 주체에 대한 신용평가에 불리한 영향을 미치는 정보다. 긍정적 정보(Positive Information)는 대출상환실적, 소득금액 등 신용평가에 유리한 영향을 미치는 정보라고 할 수 있다.

최근에는 금융정보가 부족한 고객에게 통신정보, 모바일기기 정보, SNS 정보 등 다양한 비금융정보를 신용평가에 활용하는 등 정보 활용 영역이 확대되고 있다. 이러한 비전통적인 정보를 대체정보(Alternative Data)라고 하며, P2P(Peer to Peer) 대출업체·모바일 뱅킹 등에서 적극적으로 활용하고 있다.

[표 4-2-1] 신용정보의 종류

식별정보	특정 신용정보 주체를 식별할 수 있는 정보로 개인의 성명·주소·주민등록번호 등과 기업 및 법인의 상호·법인등록번호 등에 관한 사항
신용거래정보	신용정보 주체의 거래내용을 판단 할 수 있는 정보로 대출·보증·담보제공·(가계)당좌거래·신용카드·할부금융·시설대여·금융거래 등 상거래와 관련해 그 거래의 종류·기간·금액·한도 등에 관한 사항
신용도 판단정보	신용정보 주체의 신용도를 판단할 수 있는 정보로 금융거래 등 상거래와 관련해 발생한 연체·부도·대위변제·대지급·거짓 등의 부정한 방법에 의한 신용 질서 문란행위와 관련된 금액과 발생·해소의 시기 등에 관한 사항

출처: 금융감독원 홈페이지

2) 신용정보 활용 목적

신용정보를 개인·기업의 신용평가에 활용하는 주요한 이유는 경제 주체들이 개인·기업의 신용도에 대한 예측을 가능하게 해, 정보의 비대칭성을 완화함으로써 경제활동의 사회적 비용을 감소시키는 것에 있다.

금융시장에서 정보의 비대칭성은 금융회사가 △역선택(Adverse Selection: 은행의 대출에 신용도가 낮은 고객이 더 적극적)과 △도덕적 해이(Moral Hazard: 대출 이후 의도적으로 상환을 하지 않을 유인 발생)를 초래해 선량한 금융 소비자에게 피해를 주고 금융산업의 건전성을 저해하게 한다. 그러나 금융회사들이 신용정보를 적절하게 활용한다면, 금융 소비자의 신용도를 정확하게 판단할 수 있고, 이에 따라 대출 여부와 수준(금액과 이자율 등)을 적정하게 결정할 수 있게 된다. 이는 금융산업뿐만 아니라 신용이 발생하는 일반 산업에서도 같이 적용될 수 있다.

3) 신용정보 공유 체계

신용정보법상의 신용정보의 공유체계 참여자는 [표 4-2-2]와 같다.

[표 4-2-2] 신용정보법상 신용정보 공유체계

구분		내용
신용정보 집중기관	종합신용정보 집중기관	대통령령으로 정하는 금융기관 전체로부터의 신용정보를 집중관리·활용
		한국신용정보원
	개별신용정보 집중기관	상기 금융기관 외의 같은 종류의 사업자가 설립한 협회 등의 협약에 따라 신용정보를 집중 관리·활용
		2016년 이전에는 한국여신금융협회, 생명보험협회, 한국금융투자협회 등이 해당했다. 하지만 2016년부터는 한국신용정보원으로 정보가 통합 집중돼 상기 기관들은 단순 협회의 기능만 갖고 있음
신용정보회사	신용조회업무 (6개사)	개인·기업에 대한 신용정보를 수집해 자체 신용평가모델을 통해 부도 위험 등을 나타내는 신용평점·신용등급을 산출하고, 신용평점과 기타 신용정보를 대출거래 등에 필요 시 제공
		코리아크레딧뷰로, 나이스평가정보, SCI평가정보, 나이스디앤비, 이크레더블, 한국기업데이터 등
	신용조사업무 (1개사)	타인의 의뢰에 따라 신용정보를 조사해 제공
		한국TDB신용정보
	채권추심업무 (22개사)	채권자의 위임을 받아 연체한 채무자에 대해 재산조사, 채무변제 촉구, 변제금 수령 등 추심채권을 행사
		SGI신용정보, 고려신용정보, 미래신용정보, 신한신용정보 등
신용정보 주체		신용정보로 식별되는 자로서 그 신용정보의 주체가 되는 자
신용정보 제공·이용자		고객과의 금융거래 등 상거래를 위해 본인의 영업으로 얻거나 만들어낸 신용정보를 타인에게 제공 또는 타인으로부터 신용정보를 받아 본인의 영업에 이용하는 자로서 주로 금융기관 등이 포함

4) 일반적인 신용정보와 신용조회회사 정보의 차이점

일반적인 신용정보(한국신용정보원 집중정보)와 신용조회회사(CB) 정보의 차이점은 [표 4-2-3]과 같다.

[표 4-2-3] 일반적인 신용정보와 신용조회회사 정보의 차이점

일반적인 신용정보	금융거래 등 상거래에 있어서 거래 상대방에 대한 식별, 신용도, 신용거래능력 등의 판단을 위해 필요한 정보
신용조회회사(CB) 정보	일반적인 신용정보 이외에 개인의 신용도를 더 정확하게 판단하기 위한 추가적인 정보를 포함 금융거래자 현황(소유재산, 직업, 소득 등), 상환이력 정보(대출, 카드), 단기연체 정보, 개인신용 평점 등과 같이 신용정보를 가공해 생성한 정보 등

나. 개인 신용정보 서비스

1) 금융회사 대상 서비스

① 기업(주로 금융회사)에 제공하는 신용조회회사의 개인 신용정보 서비스는 [표 4-2-4]와 같이 크게 3가지 영역으로 요약할 수 있다. 이 중에서 신용리스크 관리 서비스는 이미 전 금융회사에서 대출·카드 등의 심사업무에 적극 사용되고 있으며, 최근에는 부정거래 위험관리, 전략적 의사결정 지원시스템 등의 구축도 활발해지고 있다.

② 개인 신용정보 서비스에 활용되고 있는 신용정보로는 식별정보·신용거래정보·신용도 판단정보·신용 능력정보·공공정보 등이 있다. 특히 공공정보 중 과태료·세금체납 등의 정보는 신용리스크 관리 영역에서, 관보정보(법원의 금치산, 한정치산, 실종 선고, 회생 및 파산 등)는 부정거래 위험관리 서비스 영역에서 활용되고 있다.

[표 4-2-4] 금융회사 대상 개인 신용정보 서비스 영역

구분	내용
신용리스크 관리	금융회사가 개인·개인 사업자에게 대출을 시행하면서 발생할 수 있는 채무불이행 위험을 정확히 측정할 수 있도록 제공하는 서비스다. 개인의 신용 거래 내용 등을 분석해 불량률을 예측하는 개인신용 평점(Score)이 대표적이다.
부정거래 위험관리	금융회사가 신용사기 거래를 개설 시점이나 개설 후 시점에서 예방하거나 적발할 수 있도록 지원하는 서비스다.
전략적 의사결정 지원	전체 가계금융 시장에서의 자사의 위치를 파악하고, 전반적인 경제 트렌드 속 자사의 현재 전략을 점검할 수 있도록 지원하는 서비스다. 금융회사는 이 서비스를 통해 시장환경에 맞는 전략을 수립할 수 있다.

2) 개인 대상 서비스

신용정보회사는 온라인으로 개인의 신용정보를 직접 확인하고 관리할 수 있는 신용관리 서비스를 운영하고 있다. 개인에게 제공되는 주요 정보로는 신상정보, 대출·카드정보, 보증정보, 연체정보, 조회정보 등이 있다. 이외에도 신용위험도 진단, 신용대출 진단, 금융 명의 보호, 신용 수준 비교 등의 서비스를 하고 있다.

현재 본인이 자신의 신용정보를 조회할 경우에는 신용도 하락이 없으며, 이를 통해 신용평점·신용등급·주요 변동내용 등 본인의 신용을 상세히 파악하고 관리할 수 있다. 또한 각 서비스 사이트는 1년에 3회에 한해 전 국민이 무료로 본인의 신용정보를 조회할 수 있는 무료 조회 서비스도 제공하고 있다.

요즘은 신용등급, 신용정보가 개인의 경제생활에 미치는 영향이 점차 확대됨에 따라, 단순히 본인의 신용을 확인하는 데서 그치지 않고 본인의 신용을 적극적으로 관리하고 높이려는 요구가 커지고 있다. 이에 따라 신용정보 회사에서도 신용정보 변경에 따른 신용등급을 예측할 수 있는 '신용등급 시뮬레이터', 전문가가 직접 신용등급 향상 방법을 제공하는 '신용등급 향상 1:1 코치 서비스' 등도 제공한다.

[표 4-2-5] 개인 대상 신용정보 조회 서비스 운영 사이트

구분	사이트 주소	회사명
올크레딧	www.allcredit.co.kr	코리아크레딧뷰로(KCB)
NICE지키미	www.credit.co.kr	나이스평가정보
사이렌24	www.siren24.com	SCI평가정보

다. 기업 신용정보 서비스

기업 신용정보 서비스는 기업의 정보를 확인하는 기업정보조회 서비스와 기업의 금융상품과 신용제공 등(회사채, 기업어음, 유동화 증권)의 원리금이 상환될 가능성 등 기업·법인, 펀드 등의 신용도를 평가해 그 결과를 제공하는 신용평가 서비스로 분류된다.

[표 4-2-6] 기업 신용정보 서비스 영역

구분	내용
기업정보 조회 서비스	기업의 재무제표, 재무현황, 기업 현황, 주주현황, 관계사 현황 등의 기업 상세 정보를 제공하는 서비스
기업 신용평가 서비스	회사채 등 유가증권의 발행, 거래 시의 투자판단, 발행조건 결정 기준, 거래기업의 적격심사 등 신용판단을 위한 기준으로 활용되는 기업의 신용등급을 제공하는 서비스

라. 업계 동향 및 전망

1) 당국의 정책 방향

금융위원회는 지난 2018년 3월 '금융 분야 데이터 활용 및 정보보호 종합방안'을 발표해 금융분야를 빅데이터 시험대로 육성하고, 법·제도·산업·인프라 측면의 종합적인 대응방안을 마련하기로 했다. 이를 위해 3대 추진전략과 10대 추진과제를 제시했는데 주요 내용은 다음과 같다.

① 금융 분야 빅데이터 활성화를 위해 빅데이터 분석·이용을 위한 법적 근거를 명확히 하고, CB사와 카드사의 빅데이터 시장 선도 역할을 강화해 빅데이터를 활용한 개인신용평가 체계를 고도화한다.

② 금융 분야 데이터 산업의 경쟁력을 강화하기 위해 신용정보산업의 경쟁을 촉진(비금융 특화 CB 도입, 기업 CB의 진입규제 완화)하고, 본인 신용정보관리업을 도입해 신용정보산업의 책임성을 확보한다.

③ 정보보호를 내실화하기 위해 동의제도 개선, 개인정보의 자기 결정권 보장(프로파일링 대응권, 개인신용정보 이동권), 정보 활용과 관리실태 상시 평가제를 도입한다.

위 정책에는 신용정보에 대해 익명정보·가명처리정보 등의 개념을 도입하고, '개인정보 비식별 조치 가이드라인'의 일부 성과도 법제화해 신용정보의 활용과 보호에 있어 법적 근거를 명확히 하도록 했다. 또한 CB사와 카드사가 보유한 정보와 노하우를 활용해 금융권의 빅데이터를 선도할 수 있도록 빅데이터 관련 업무를 허용하는 등 금융 분야 빅데이터 산업 활성화를 추진하고 있다.

이러한 정책 방향들은 향후 신용정보법 개정 또는 관계기관 협의를 거쳐 추진될 예정이다.

2) 신용정보회사 재무현황

2017년 말, 현재 신용정보회사의 총자산은 1조 217억 원, 자기자본은 7,692억 원으로 전년 말 대비 각각 500억 원(+5.1%)과 396억 원(+5.4%)이 증가했다. 당기순이익은 690억 원으로 채권추심회사 당기순이익 감소(-138억 원)에 따라 전년 대비 86억 원(-11.1%)이 감소했다. 이 중 신용조회회사의 영업수익은 5,352억 원으로 TCB 업무 영업수익 증가(+67억 원, +14.2%) 등에 따라 전년 대비 422억 원(+8.6%) 증가했다.

전체적으로 채권추심회사는 영업비용 증가 등에 따라 당기순이익이 감소하고 있으며, 신용조회회사의 영업수익과 당기순이익은 탄탄한 상승세를 보인다.

3) 신용정보 활용 동향

금융소비자의 금리 부담을 완화하기 위한 중금리대출 활성화 정책에 따라 신용정보를 활용한 중금리대출 모형개발이 적극적으로 이뤄지고 있다. 특히 사회초년생, 주부 등 금융거래가 부족한 신파일러(Thin Filer: 금융 활동 경력이 없어 대출 등 신용혜택을 받지 못하는 사람)의 평가를 위해 대안신용평가가 많이 이뤄지고 있다. 최근에는 통신정보를 활용한 신용평가가 금융회사에 도입되고 있고, 모바일기기·SNS 정보를 이용한 신용평가도 이뤄지고 있다. 이외에도 설문 등을 통한 개인의 성향을 평가하는 시도도 이뤄지고 있다.

가계부채의 안정을 위한 대출억제 정책이 펼쳐지고 있는 가운데, 개인사업자 대출에 대한 관심도 높아지는 추세다. 개인사업자 대출 심사 자동화, 각종 신용정보를 활용한 심사모형 정교화 등 심사역량 강화에 많은 노력을 기울이고 있다.

4) 금융권 동향

인터넷은행의 출현, 지급결제시장 경쟁심화, AI 시대의 도래 등으로 금융권은 많은 변화에 직면해 있다. 고객에게 간편하고 빠른 서비스 제공을 위해 심사 자동화와 비대면 거래시스템 구축 등 많은 노력을 기울이고 있고, 모바일 앱 서비스 개편 등 모바일 서비스 또한 고객 편의성을 강화해 나가고 있다.

또한 금융지주 등을 중심으로 빅데이터 허브시스템을 구축해 빅데이터 인프라와 활용체계 마련을 위한 활발한 움직임을 보이고 있다. 또한 신용정보 분석에 머신러닝, 딥러닝 등 AI 기법을 적용해 활용하는 금융회사가 크게 늘고 있다.

하지만 한국은 정보보호 규제가 전 세계적으로 가장 강한 수준으로, 정보의 수집·분석·활용 등 전 단계에서 엄격한 규제가 적용되고 있는 나라다. 데이터의 결합이나 분석 등을 위한 법·제도의 개선이 지연되고, 현행 '비식별조치 가이드라인'을 활용한 금융회사, 전문기관에 대한 형사고발('17.11)이 발생하는 등 금융회사의 빅데이터 활용은 전반적으로 위축된 상태다.

2. 보건의료데이터 분야 서비스

가. 보건의료데이터 서비스 개요

1) 보건의료 분야 데이터의 특징

보건의료데이터는 개인의 건강 또는 질병과 관련된 정보를 포함한다. 개인정보 보호법 제23조(민감정보의 처리 제한) 제1항에는 이를 민감정보로 규정하고 있다. 보건의료데이터는 개인을 특정할 수 있는 정보(주민등록번호 등)와 시간 정보도 함께 생성된다. 따라서 정보의 개방에도 일반법인 개인정보 보호법과 특별법인 생명윤리 및 안전에 관한 법률의 영향을 받는다(표 4-2-7).

[표 4-2-7] 보건의료데이터의 개인정보 개방 관련 법률

법률	관련 내용
개인정보보호법	• 개인정보의 제3자 제공: 동의를 받은 경우 • 민감정보, 고유식별정보, 주민등록번호 처리가 가능한 경우 – 별도로 동의를 받은 경우 – 법령에서 구체적으로 고유식별 정보의 처리를 요구하거나 허용하는 경우
생명윤리 및 안전에 관한 법률	• 제3자에게 제공 – 서면동의를 받고 기관위원회의 심의를 거친 경우 – 익명화가 원칙이나 개인식별 정보 포함에 대한 동의한 경우 익명화하지 않아도 됨 • 기관 위원회의 승인을 받아 서면동의를 면제할 수 있는 요건(둘 다 만족) – 연구 대상자의 동의를 받는 것이 연구 진행과정에서 현실적으로 불가능하거나 연구의 타당성에 심각한 영향을 미친다고 판단되는 경우 – 연구 대상자의 동의 거부를 추정할 만한 사유가 없고, 동의를 면제해도 연구 대상자에게 미치는 위험이 극히 낮은 경우

2) 보건의료분야의 공공데이터

보건의료분야의 공공데이터에는 건강보험공단, 건강보험심사평가원, 질병관리본부, 국립암센터의 데이터가 주로 거론된다(건강보험공단과 심사평가원의 데이터는 주로 건강보험 진료자료와 이의 연계자료다. 질병관리본부의 국민건강영양조사와 국립암센터의 암 등록 자료는 조사나 자료수집 자체를 목적으로 하며 수집한 자료를 공개하고 있다).

• 2절 필자 | **김석일**(가톨릭대 예방의학교실 교수)

대다수 외국의 의료기관은 중앙정부의(일부) 재정 지원으로 지방정부가 운영하므로, 병원에서 발생한 데이터를 보통 공공데이터로 구분한다. 그러나 우리나라는 의료기관의 90% 이상이 민간의료기관이고, 공공의료기관도 환자 진료를 통한 수입이 운영 비용의 대부분을 차지하고 있으므로 보건의료데이터를 공공데이터라고 보기 어렵고, 공개도 하지 않는다 .

다만 우리나라는 건강보험 지불제도로 행위별수가제를 기본으로 하므로 병원의 설립 형태에 관계없이, 병원에서 발생한 데이터 중 건강보험과 관련된 것은 매우 상세하게 건강보험심사평가원으로 제출해 심사를 받고 있다.

나. 국내 서비스 현황

1) 개괄

건강보험공단의 공개데이터에는 전 국민의 건강보험 자격, 진료정보, 건강검진 데이터가 포함된다. 건강보험심사평가원의 자료는 주로 진료정보와 이에 수반되는 의약품, 치료재료 등이 포함된다. 건강보험심사평가원은 건강보험 청구자료를 접수해 심사 후 지급을 한다. 따라서 심사평가원의 데이터는 기본적으로 건강보험공단의 진료자료와 같다.

건강보험심사평가원과 건강보험공단에서 공개하는 건강보험 데이터의 차이는 [표 4-2-8]과 같다. 심사평가원에서는 산업용 데이터를 공개하는 것이 건강보험공단과 비교하면 큰 차이라고 할 수 있다. 또 자료제공 방식에서 건강보험공단에서는 오프라인을 위주로 하지만, 심사평가원은 온라인과 오프라인을 동시에 활용하고 있다.

[표 4-2-8] 건강보험공단과 심사평가원에서 제공하는 건강보험 데이터 비교

기관	국민건강보험공단	건강보험심사평가원
명칭	• 맞춤형 연구 데이터	• 의료빅데이터
방식	• 빅데이터 분석센터(오프라인)	• 원격분석시스템(온라인), 빅데이터센터(오프라인)
자료	• 자격 DB, 진료(명세서·진료·상병·처방전), 건강검진, 요양기관	• 진료(명세서, 진료, 상병, 원외 처방내역) • 요양기관 현황
자료반출	• 통계 산출물(분석 결과값)에 한해 자료반출 가능	• 통계 산출물(분석 결과값)에 한해 자료반출 가능

(계속)

기관	국민건강보험공단	건강보험심사평가원
제공대상	• 국가기관 및 지방자치단체 • 「공공기관의 운영에 관한 법률」 제4조에 따른 공공기관 • 제1호 또는 제2호에 해당되지 않으면서 정책연구나 학술연구를 수행하는 기관 또는 사람 • 공단과 체결한 협약 등에 따라 연구를 수행하는 기관 또는 사람 • 그 밖에 제1호부터 제4호에 해당하지 않으며 기타연구를 수행하는 기관 또는 사람	• 연구자료 – 국가, 지방자치단체 및 정부산한기관 – 연구중심 병원 및 학술 연구 수행기관 등 • 산업체자료 – 의약품 품목 허가를 받은 자 – 제약, 의료기기 산업 – 컨설팅 회사 및 예비창업자 등
제공범위	• 건강보험자료 제공: 건강보험자료(업무를 위해 타 기관으로부터 제공받은 자료 제외) 범위 내에서 제공 • 정보 식별 불가능 형태: 개인, 법인 및 단체 등의 정보를 식별 불가능한 형태로 제공 • 맞춤형 자료 제공: 맞춤형 자료 제공 시 사전에 협의를 통해 제공 가능한 형태의 자료 생성	• 보건의료 자료 제공 시 건강보험자료 범위 내에서 제공 • 맞춤형 자료 제공 시 건강보험심사평가원에서 제공 가능한 형태의 자료 제공
기타	• 건강보험심사평가원에서 심사를 마친 보험청구 자료 • 자격 DB에 청구 대상자의 지역, 보험 등급이 있음 • 건강검진 DB와 진료자료를 연계해 분석가능	• 병원에서 청구된 원래의 자료 • 진료내역자료에 세부항목이 많음

질병관리본부의 국민건강영양조사는 국민의 건강수준, 건강관련 의식 및 행태, 식품 및 영양섭취실태에 대한 국가단위의 대표성과 신뢰성을 갖춘 통계산출을 목적으로 하는 법정 조사다. 검진조사·건강설문조사·영양조사로 구성돼 있으며, 개인정보보호법에 따라 개인을 식별할 수 있는 자료 또는 다른 정보와 결합해 개인 식별이 가능한 자료를 제외하고 공개하고 있다.

국립암센터의 암등록 자료는 의료기관으로부터 암발생 위험요인, 발생 및 치료에 대한 자료를 체계적으로 수집·분석해 암 발생률·생존율·유병률 등에 대한 통계를 산출하기 위한 것이다. 데이터 공개는 암 등록자료 범위 내에서 이뤄지며, 초진 연월일, 국제질병분류, 종양학 국제질병분류, 암의 최종 진단방법 등을 사전 협의를 통해 제공한다. 환자의 개인정보와 개별 법인·단체 등의 정보를 식별 불가능한 형태로 제공한다.

2) 데이터의 품질 문제

건강보험 청구자료의 진단명이 병원 의무기록실에서 코딩한 진단명과 다르다는 것이 간혹 중요한 문제로 부각되곤 한다. 그러나 두 가지 측면에서 이러한 현상을 이해할 필요가 있다.

첫째, 건강보험 청구자료는 병원에서 환자 진료에 소요된 예산을 청구하는 과정에서 생기는 데이터다. 즉 병원에서는 환자 진료에 이미 소모한 자원에 대해 최대치를 지불받기 위해 보험코딩을 한다. 건강보험 청구명세서의 심사 과정에서는 환자 진료과정에 소요된 자원과 진단코드가 맞지 않으면, 청구한 금액을 감해 받게 된다. 따라서 건강보험 청구명세서의 진단명이 이와 다르더라도, 현재의 건강보험 지불제도 아래에서는 보험청구 시 진단명이 틀렸다고는 할 수 없다.

둘째, 의무기록실에서 진단명을 코딩하는 사람은 의무기록사이고, 보험청구를 하는 사람은 간호사다. 보통 보험심사간호사는 임상 경험을 갖고 있는 사람을 선발하므로 간호대학의 교육과정과 임상 경험을 통해 간호사가 더 많은 임상 지식을 갖고 있다. 또 의무기록사는 전문적으로 진단명을 세계보건기구(WHO, World Health Organization)에서 발간하는 ICD(International Classification of Disease) 코딩 방법을 교육 받은 사람이라고 할 수 있다.

그러나 실제 ICD의 코딩 지침에서는 주진단 선정에 대한 것만 있고, 더 세부적인 것은 나라별로 별도로 정해야 한다고 기술하고 있다. 우리나라 건강보험 청구에 적용할 수 있는 명시적인 진단명 선정 지침은 없으므로 의무기록사가 코딩한 진단명과 보험청구간호사가 코딩한 진단명이 다른 것은 코딩 지침이 없기 때문에 발생한다. 따라서 건강보험 데이터는 이러한 제약이 있음을 염두에 두고 활용되고 있다.

3) 공공데이터의 활용

공공데이터 활용 공공서비스 제공 및 정비 가이드라인[1]에서 '정부는 공공데이터 제공 및 시장 공정경쟁을 지원하고 민간은 개방된 데이터를 자유롭게 활용, 서비스 개발·경쟁을 유도'하는 것을 공공데이터 제공 및 이용활성화 추진 방향으로 제시하고 있다. 보건의료 분야에서의 민간 활용은 크게 연구용과 상업용 두 가지로 나눠 볼 수 있다.

① 연구용

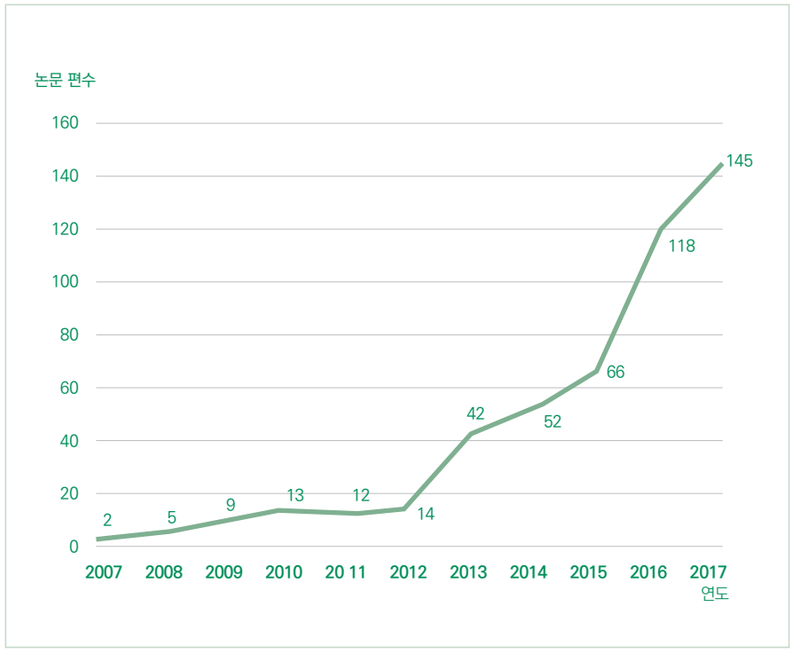
연구용은 보건의료데이터를 이용해 논문이나 보고서를 작성하는 것이다. 2007년부터 2017년까지 건강보험데이터를 이용해 출판한 논문 중 몇 가지 제약조건을

1 행정안전부, 「공공데이터 활용 공공서비스 제공 및 정비 가이드라인」, 2017.

만족하고 영어로 검색이 가능했던 논문 수를 보면[2], 2013년 공공데이터의 제공 및 이용 활성화에 관한 법률이 시행되면서 논문 수가 급속히 증가함을 알 수 있다(그림 4-2-1).

한편 2005년 '황우석 논문조작사건' 이후 보건의료계의 학술적 데이터 활용은 심각한 제약을 받게 됐다. 연구목적의 데이터를 활용하려면 기관내 윤리위원회(IRB, Institutional Review Board)의 심사를 통과해야 한다. 최근에는 여러 기관이 참여하는 다기관 연구(Multi-institutional research)가 많다. 이 경우 참여하는 모든 병원의 IRB를 통과해야 가능하다. 공개 가능한 보건의료데이터를 사용할 때에는 연구자가 소속된 기관의 IRB 통과 서류를 공개기관에 제출해야 데이터 활용이 가능하다.

[그림 4-2-1] 2007년~2017년 건강보험데이터를 이용해 영어로 출판한 논문 수



2 정보영, A Study on the Improvement for big data utilization- Focused on health insurance claims data, 연세대학교 대학원, 2018.

② 상업용

공공데이터 이용 보장의 의미는 이용자가 데이터를 위조·변조·왜곡하지 않고, 사실적 내용을 유지하는 범위에서 자유롭게 가공·활용함에 있다. 이것은 데이터의 상업적 이용도 가능한 것으로 이해할 수 있다. 보건의료 분야의 다른 데이터에서는 연구 목적 이외의 활용을 찾아보기 어려우나, 건강보험심사평가원의 건강보험데이터는 [표 4-2-8]의 제공 대상에 산업체가 포함돼 있어 그 가능성을 볼 수 있었다.

그러나 건강보험심사평가원(이하 심평원)이 2014년 7월부터 2017년 8월까지 8개 민간보험사 및 2개 민간보험연구기관에게 보험료 산출과 보험상품개발 등을 위해 요청한 '표본 데이터세트'를 판매했다[3]는 것을 더불어민주당 정춘숙 의원이 밝히면서 시민사회단체에서는 해당 데이터의 공개 중단을 요구했다. 건강보험심사평가원에서는 상업적 활용이 가능하도록 노력했지만[4], 보건복지부에서 보건의료 빅데이터를 공익목적으로 활용[5]할 것을 발표함으로써 당분간은 상업적 목적으로의 보건의료데이터 이용은 어려워질 전망이다.

다. 해외서비스 현황

1) 유럽

유럽식 의료시스템을 갖고 있는 유럽과 호주 및 캐나다는 대부분이 공공병원이다. 따라서 병원데이터가 공공데이터 성격을 갖게 된다. 하지만 민감정보인 환자데이터를 공개하는 경우는 찾아보기 어렵다. 그보다는 각 병원의 성과를 측정하고, 예산을 배분하기 위해 데이터를 수집한다. 공개는 주로 공공재원의 투명성을 나타내기 위한 목적이다.

2) 미국

전 국민 단일보험자에 의해 건강보험을 운영하는 한국과 달리 2017년 4분기 현재

3 「민간보험사에 개인정보 팔아넘긴 심평원 규탄 및 보건의료 빅데이터 사업 추진 중단 요구」, http://www.peoplepower21.org/Welfare/1533008, 2017.10.30.

4 「빅데이터 민간 제공 논란에도 활용범위 확대 강조하는 심평원」, http://www.docdocdoc.co.kr/news/articleView.html?idxno=1050348, 2017.12.15.

5 박능후 복지부 장관, 「보건의료 빅데이터 사업, 공익 목적일 경우에만 활용」, http://www.etnews.com/20180720000234?m=1, 2018.07.20.

[그림 4-2-2] 미국 CMS의 Medicare와 Medicaid 데이터 제공 사이트

출처: https://www.cms.gov/Research-Statistics-Data-and-Systems/Files-for-Order/Data-Disclosures-Data-Agreements/Overview.html(접속: 2018.07.29.)

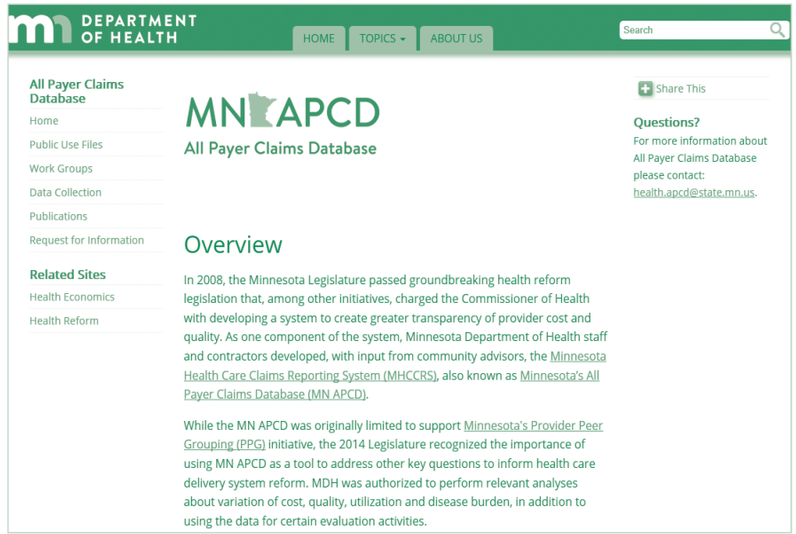
미국 거주자의 12.2%는 건강보험에 가입돼 있지 않다[6]. CMS(Center for Medicare and Medicaid)에서 운영하는 Medicare, Medicaid, CHIP(Children's Health Insurance Program) 등의 공공보험과 Unitedhealth, Wellpoint Inc., Kaiser Foundation Group, Humana, Aetna, HCSC, Cigna Health 등 수많은 민간보험사가 있다.

미국 국민은 개인적인 상황에 따라 각 보험에 가입하므로, 한국과 같은 전 국민 건강보험데이터는 존재하지 않는다. 또 보험 청구에도 병원에 대한 DRG(Diagnosis Related Groups)와 의사에 대한 CPT(Current Procedural Terminology)를 기반으로 한다. 포함되는 내용이 우리나라 보험청구 명세에 비해 단순한 편이다.

한편 CMS에서는 개인 및 건강정보의 포함 정도에 따라 3가지 수준의 자료를 공개하고 있다(그림 4-2-2). 개인정보가 포함된 데이터는 모두 연구 목적으로 제공

6 U.S. Uninsured Rate Steady at 12.2% in Fourth Quarter of 2017, https://news.gallup.com/poll/225383/uninsured-rate-steady-fourth-quarter-2017.aspx, 2018.01.16.

[그림 4-2-3] 미국 미네소타주의 모든 보험자를 통합한 건강보험자료



출처: http://www.health.state.mn.us/healthreform/allpayer/index.html(접속: 2018.07.29.)

되고 있다. 또 주에 따라서는 해당 주의 모든 보험회사 데이터를 모아 공개하는 경우(그림 4-2-3)도 있는데, 이 또한 연구가 주요 목적으로 돼 있다.

라. 향후 전망

지난 몇 년 동안 순수한 연구 목적의 데이터 활용은 상당히 활성화됐지만, 데이터를 활용하기 위해서는 연구자 소속기관의 기관윤리위원회와 데이터 제공기관의 승인을 모두 받아야 한다. 연구 목적의 데이터 활용이 아직까지 쉽지만은 않다는 점과 2017년까지 보여온 상업적 활용에 대한 거부감을 감안하면, 보건의료데이터의 상업적 활용은 향후 몇 년 간은 어려운 상황이 될 전망이다.

보건의료데이터는 기본적으로 개인의 민감한 정보를 포함하고 있으므로 이를 보호하려는 노력은 지속적으로 이뤄져야 한다. 현재 구현된 기술적인 정보보호방안도 필요하지만, 이와 함께 부적절한 데이터 이용에 대한 법적 제재조치를 고려함으로써, 실질적인 데이터 활용 여건이 개선돼야 할 필요가 있다. 이와 함께 데이터를 상업적으로 활용하기 위해서는 더 심도 깊은 논의가 필요하다.

3. 학술데이터 분야 서비스

가. 학술정보 서비스 개요

학술정보란 일반적으로 학술 또는 연구 활동과 관련된 정보를 말한다. 학술지 논문, 학위 논문, 단행본, 연구보고서, 특허, 동향자료와 같은 연구결과물이 대표적인 학술정보다.

최근에는 연구자 프로필, 연구 성과, 연구 데이터, 피어 리뷰(Peer Review), 학술 SNS 게시물 등에 이르기까지 학술정보의 범위가 확장됐다. 온라인을 통한 학술 커뮤니케이션 활동 영역이 확장되고 있는 만큼, 학술정보의 범위와 규모는 앞으로도 계속 확장될 전망이다.

학술정보의 종류는 다양하지만, 학술정보 서비스에서 영향력이 가장 큰 것은 학술지 논문이다. 학술지는 연구자들 간의 커뮤니케이션이 가장 활발히 이뤄지는 매체이자 최신 연구결과물을 지속적으로 가장 빠르게 반영하는 특징이 있다. 또한 다른 자료에 비교해 DB화가 잘 돼 있어 온라인에서 검색과 원문 이용이 편리하다. 학술지 논문이 과거 인쇄 저널 위주의 환경에서는 주로 석박사 이상 연구자들의 전유물이었지만, 온라인 이용 환경이 보편화하면서 대학생·고등학생·일반인에 이르기까지 이용자 폭이 크게 확장됐다.

학술정보 서비스는 오랫동안 단순 검색과 원문 다운로드 서비스 수준에 머물렀다. 하지만 최근 들어 학술과 연구 편의성을 증대하거나 새로운 부가 정보를 제공하기 위한 노력이 더욱 늘어났다. 기존에는 자료의 양과 질에 관한 경쟁 위주였던 반면, 점차 학술정보 데이터가 공개되고 공유됨에 따라 서비스 플랫폼 경쟁이 더 중요해지고 있다.

나. 국내 학술정보 서비스 동향

1) 서비스 영향력 비교

국내의 주요 학술정보 서비스로는 [표 4-2-9]와 같이 공공 2군데, 민간 3군데 서비스를 꼽을 수 있다. 사이트 방문자 수와 페이지뷰 수를 기준으로 서비스 영향력을 비교해 보면, RISS·DBpia·네이버 학술정보가 상위에 속한다. 국내 학술지 논문과 학위 논문 포털 서비스를 제공하는 RISS가 페이지뷰 수 기준 1위이고, 추천이나 저자식별 등 부가정보 및 편의기능이 상대적으로 뛰어난 DBpia가 방문자 수 기준 1위를 차지했다.

[표 4-2-9] 국내 주요 학술정보 서비스의 영향력 비교(최근 1년 : 2017.07~2018.06 기준)

구분	서비스	제공기관 및 업체	월간 방문 수	순위	월간 페이지뷰 수[1]	순위
공공	RISS	한국교육학술정보원	635,879	2	5,805,577	1
	NDSL	한국과학기술정보연구원	253,131	4	604,983	4
민간	네이버 학술정보	네이버	537,531	3	1,327,701	3
	DBpia	누리미디어	813,061	1	2,179,003	2
	KISS	한국학술정보	161,367	5	335,643	5

출처: 시밀러웹, http://similarweb.com(2018.07.02. 접속)

2) 연관 자료 제공 서비스 강화

현재 국내 학술지 논문 수가 500만 건 내외이고 이 또한 해마다 5~10%씩 꾸준히 증가하고 있다. 이러한 환경에서 키워드 검색만으로 원하는 자료를 발견하기란 쉽지 않다. 연구자가 원하는 논문을 발견할 가능성을 높이기 위해 학술정보 서비스에서는 연관 자료 제공 서비스를 고도화하고 있다.

대표적으로 연관 자료 추천(논문별 연관 자료 추천과 개인 맞춤 추천), 연구자별 논문 내역을 확인할 수 있는 저자식별, 참고문헌과 인용색인을 꼽을 수 있다. 연관 자료 추천이나 저자식별은 DBpia와 NDSL이, 참고문헌과 인용색인은 국내·해외 학술지 논문을 포괄하는 네이버 학술정보가 상대적으로 강점을 갖고 있다.

3) 논문 이용지표를 통한 활용도 분석

전통적으로 연구 성과는 논문이 게재된 학술지의 인용 영향력(Impact Factor)을 통해 평가됐다. 하지만 개별 논문의 성과를 학술지 지표로 평가하는 것은 학계에서

• 3절 필자 | **박대광**(누리미디어 과장)

1 월간 페이지뷰 수 = 월간 방문수(Monthly Visits)와 방문당 PV(Pages/Visits) 데이터를 이용해 산출

[표 4-2-10] 사이트별 연관 자료 제공 서비스 기능 비교

구분	서비스	연관 자료 추천	저자식별	참고문헌 및 인용색인
공공	RISS	△	×	△
	NDSL	○	○	△
민간	네이버 학술정보	○	×	○
	DBpia	○	○	△
	KISS	△	×	×

(○: 지원, △: 부분 지원, ×: 지원 안 함)

오랫동안 문제점으로 지적돼 왔다.

이러한 가운데 최근 RISS, DBpia, NDSL에서는 논문의 이용지표(Usage) 정보를 제공함으로써 개별 논문이 얼마나 많이 활용됐는지를 확인할 수 있게 했다. 인용지표(Citation)가 학계 내에서의 영향력을 확인할 수 있는 장점이 있지만, 한국에서는 국내 논문 인용이 매우 적은 편이라, 인용지표만으로는 연구 성과를 측정하는 것은 한계가 따른다.

4) 키워드를 활용한 연구동향 분석 서비스

또한 학술정보의 발견가능성을 높이기 위한 차원에서 논문의 키워드 데이터가 더 적극적으로 활용되고 있다. 논문의 저자 키워드와 더불어, 텍스트 마이닝 기법을 활용해 논문 초록 및 본문에서 자동 추출한 키워드로 주제어와 관련된 주요 연구성과와 연구동향 정보를 직관적으로 파악할 수 있도록 하는 것이다.

DBpia와 KISS는 최신 발행 논문의 주요 키워드 동향을 워드 클라우드로 제공한다. NDSL은 주제어 지식맵의 방식으로 연관 주제어, 주제어에 관한 주요 연구자, 연구기관 정보를 시각화해 제공한다. RISS는 SAM 학술관계분석서비스(http://sam.riss.kr)에서 연구 키워드에 대한 논문 관계도·연구자 관계도를 제공한다. 네이버 학술정보는 키워드에 관한 분야별 논문 발행 현황·연도별 논문 발행 추이·키워드에 관한 주요 논문 정보 등 키워드별 연구 트렌드 서비스를 제공한다.

하지만 국내에서 키워드를 활용한 연구동향 분석 서비스는 아직 매우 초보적인 수준으로, 단순 키워드 위주이며 키워드 식별 또한 낮은 편이다. 이는 한국어 자연어 처리 기술의 발전이 부족한 것도 있지만, 해당 기술이 아직 학술정보 서비스에 긴밀하게 접목되지 못한 점 때문이다. 그래도 긍정적인 점은 머신러닝, 딥러닝 기술 등을 기반으로 하는 국내 자연어 처리 기술의 발전 추이를 보면, 한국어 키워드를 활용한 연구동향 분석 서비스의 혁신은 그리 멀지 않아 보인다는 것이다.

다. 해외 학술정보 서비스 동향

1) 학술 커뮤니케이션에서의 혁신

해외에서는 2015년 시점부터 학술 커뮤니케이션에서의 혁신(Innovations in Scholarly Communication)이란 용어가 화두가 됐다. 이는 '발견(Discovery)⟶ 분석(Analysis)⟶ 작성(Writing)⟶ 출판(Publication)⟶ 확산(Outreach)⟶ 평가(Assessment)'와 같은 일련의 연구 과정에 변화를 몰고온 혁신적 서비스와 솔루션에 주목한다.

과거에는 주로 저널(journal) 출판사 위주로 학술정보 서비스가 제공됐지만, 모바일 환경이나 클라우드· 빅데이터 기술이 발전하기 시작한 2010년대 이후 혁신적인 서비스를 제공하는 스타트업과 연구소가 새롭게 등장하기 시작했다. 대표적인 예로는, 개인 참고문헌 관리에서 협업·공유·학술 SNS 도구로 확장한 서지관리 서비스, 연구 결과물에 대한 온라인(소셜미디어, 언론보도 등) 반응을 추적하고 사회적 영향도를 측정할 수 있는 알트메트릭(Altmetric) 서비스, 피어 리뷰 혁신을 시도하는 오픈 피어 리뷰 출판 시스템, 피어 리뷰 보상 서비스, 프로젝트 협업 및 연구 데이터 공유 리포지터리 서비스, 온라인 논문 작성 도구 등이 있다.

[표 4-2-11] 해외의 혁신적인 주요 학술 서비스

구분	관련 서비스명
서지관리 서비스	Endnote, Refworks, Mendeley, Zotero, ReadCube, Colwiz, F1000 Workspace
알트메트릭 서비스	Altmetrics, PlumX, PLoS
피어 리뷰 서비스	F1000 Research, Publons
프로젝트 협업 및 데이터 공유 리포지터리 서비스	Open Science Framework, Figshare
출판 전 논문 서비스	arXiv(수학·물리학 분야), BioArXiv(생물학 분야), SocArXiv(사회과학 분야)

[그림 4-2-4] 변화하는 연구 워크플로

출처: The circle, https://101innovations.wordpress.com/outcomes, 2015.01.

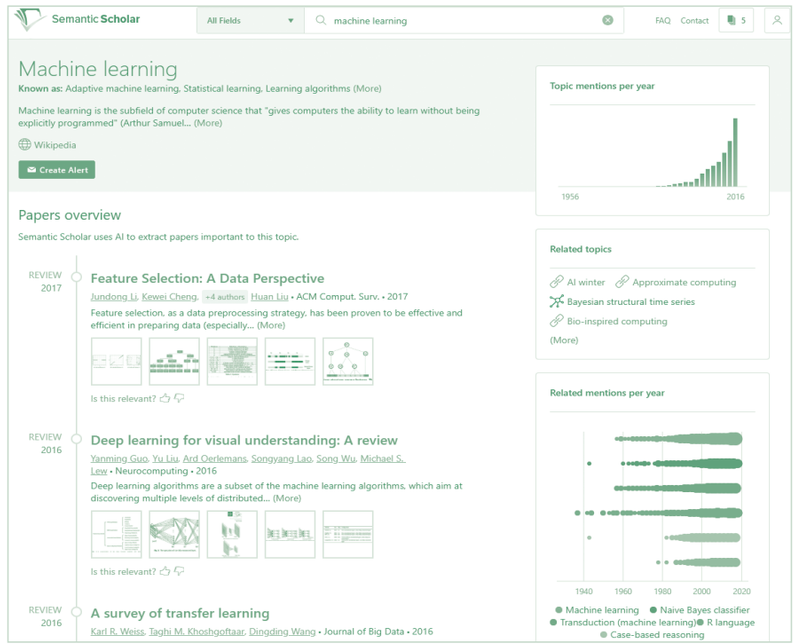
기술 혁신과 더불어 오픈 액세스(Open Access)와 같은 개방과 공유의 흐름이 확산되고 있다. 연구자가 자신의 출판 전 논문(Pre-print)을 직접 리파지터리에 등록해 빠르게 공유하는 서비스가 오픈 액세스의 한 종류다. 1990년대에 아카이브(arXiv)가 물리학 분야로 시작한 이후, 최근 들어 급속히 각 분야로 확산되고 있다.

오픈 액세스의 등장으로 엘스비어(Elsevier)나 와일리(Wiley)와 같은 전통적인 저널 출판사들은 유료 구독 저널 비즈니스 모델이 위협받는 듯 했으나, 강력한 브랜드, 자본력, 인적 네트워크, 체계적인 출판 시스템을 기반으로 논문 출판 비용(APC, Article Processing Charge) 중심의 오픈 액세스 비즈니스 모델을 정착시키며 변화하는 환경에 적응했다.

2) AI 기반의 학술정보 시맨틱 검색엔진

전 세계 학술정보 검색엔진의 절대 강자는 구글 스칼라(Google Scholar)다. 구글 스칼라는 2억만 개 학술자료를 수집·정제해 가장 단순하고 빠르게 제공하며, 광범

[그림 4-2-5] 시맨틱 스칼라의 토픽 연구동향 분석 서비스 화면



출처: 시맨틱 스칼라, https://www.semanticscholar.org/topic/Machine-learning/168(2018.07.02 접속)

위한 인용색인 정보까지 포괄한다.

그런데 최근 구글 스칼라의 경쟁자로 AI 기반의 학술정보 시맨틱 검색엔진이 등장하고 있다. 대표적인 경우가 앨런인공지능연구소(Allen Institute for Artificial Intelligence)에서 2015년에 출시한 시맨틱 스칼라(Semantic Scholar)다. 시맨틱 스칼라는 컴퓨터과학, 의학 분야의 논문 메타데이터와 원문을 수집해 논문 PDF 내 각종 정보(본문, 저자, 참고문헌, 표·그림 등)를 파싱하고 분석한다. 이를 통해 이용자는 각종 시맨틱 요소가 접목된 필터링(좁혀 보기) 기능으로 원하는 논문을 쉽게 찾을 수 있고, 논문 토픽, 각종 인용 영향력 지표, 연구자 간 네트워크 등을 한 번에 파악할 수 있다.

라. 향후 전망

학술정보는 4차산업혁명에서 매우 중요한 자원이다. 학술정보 서비스는 4차산업혁명의 자원을 효과적으로 제공해 부가가치를 창출해낼 수 있는 매개체다. 한국의 연구개발비 투자 수준은 OECD 국가들 가운데 최상위에 속하지만, 안타깝게

도 학술정보 서비스 산업의 규모나 수준은 매우 낮다. 주요 원인 중에 하나로 연구자의 해외 학술활동을 높게 인정하는 반면, 국내 학술활동을 저평가하는 국가의 학술 평가 정책 방향의 문제점을 꼽을 수 있다.

국내 학술정보 서비스의 수준은 아직 해외 서비스에 한참 못 미치지만, 그렇다고 조급하게 해외 서비스를 벤치마킹하는 것이 해결책은 아닐 것이다. 국내의 구체적인 학술 커뮤니케이션 상황과 맞물려 국내 연구자와 연구기관에 실제로 도움이 될 수 있는 서비스를 만들어내는 것이 세계적인 경쟁력을 갖추기 위한 시작이다. 또 앞으로 이를 위한 과정에서 논문 검색 서비스만이 아니라, 논문 작성·출판·확산·평가와 관련된 국내 학술 활동의 구체적인 지점들을 연결해 학술정보 서비스 플랫폼을 형성하는 것이 관건이 될 것이다.

4. 기업데이터 분야 서비스

가. 서비스 제공자 구분

기업데이터 서비스는 신용정보회사와 신용정보회사 이외의 사업자로 구분된다. 신용정보회사의 기업데이터 서비스는 금융적 관점에서 출발해, 최근에는 비금융적 관점으로 중심이 옮겨가는 추세다. 최근에는 IT 기술의 발전과 수요자의 요구사항 반영을 통해 기업데이터 시장에 많은 변화가 일어나고 있다.

기업데이터 서비스업을 영위하기 위해서는 서비스 주체가 신뢰할 수 있고 적시성 있는 대량의 기업데이터 보유는 물론, 보유 데이터의 최신성을 유지할 수 있어야 한다. 기본적으로 이러한 요건들을 충족하는 정보 제공자는 크게 두 부문으로 나눌 수 있으나, 이 글에서는 주로 신용정보사의 기업데이터 서비스를 중심으로 살펴본다.

1) 신용정보회사

조회 부문에서 신용정보회사는 금융거래 등 상거래에 있어서 거래 상대방의 신용을 판단할 때 필요한 신용정보(식별정보, 거래내용 판단정보, 신용도 판단정보, 신용거래능

• 4절 필자 | **이영준**(한국기업데이터 부장)

[표 4-2-12] 신용정보회사 현황(2017년 12월 기준)

회사명	설립일	대표자	납입 자본	허가업무			주요 주주(지분율 %)
				조사	조회	추심	
나이스평가정보	85.02.28.	심의영	304		●		NICE홀딩스(43.0), MAWER GLOBAL SMALL CAPFUND·THE HIGH CLEERE(3.7)
SCI평가정보	92.04.23.	조강직	178	●	●	●	진원이앤씨(51.0), 박중양(5.0)
이크레더블	01.08.06.	이진옥	61		●		한국기업평가(64.5), 파델리티 매니지먼트 앤 R&C(5.1)
나이스디앤비	02.10.12.	노영훈	77	●	●		NICE홀딩스(35.0), Phillip Capital Pte Ltd(26.6)
코리아크레딧뷰로	05.02.22.	강문호	100	●	●		한국기업평가(11.0), 코리아크레딧뷰로(9.2), 서울보증보험·국민·농협·하나·우리은행(9.0)
한국기업데이터	05.02.22.	조병제	692	●	●		신용보증기금(15.0), 기술보증기금(9.0), 산업·기업·농협·국민·신한·우리·하나은행(9.0)

출처: 금융감독원 홈페이지, http://www.fss.or.kr(2018.07.12. 접속)

력 판단정보 등)를 금융위원회의 허가를 받아 합법적으로 수집·서비스한다. 따라서 신용정보회사가 제공하는 신용정보(기업데이터)는 대체로 금융정보에 가까운 특성을 가진다.

2) 신용정보회사 외의 정보제공자

신용정보회사가 아니면서 기업데이터를 서비스하는 정보제공자는 특정 공적·사적 영역에서 고유의 목적을 달성하기 위해 기업정보를 유/무상으로 서비스한다(금융감독원 전자공시시스템, 중소기업현황정보시스템, 대법원 인터넷등기소 등기열람서비스, 국가통계포털, 코참비즈(2015.06.30. 서비스 종료) 등).

나. 기업데이터의 분류

신용정보회사가 서비스하는 기업데이터는 금융정보에 가까운 특성상 데이터의 종류, 서비스 방법, 서비스 경쟁력, 서비스 이용자와 용도 등의 주요 사항이 대체로 표준화돼 있다.

① 데이터의 종류: 크게 개요정보, 재무정보, 대표자정보, 신용정보, 기타정보로 구분된다. 특히 재무정보는 재무제표 전체를 포괄하므로 기업의 신용도 판단에 매우 유용하게 활용된다. 또한 최근에는 비금융 목적의 수요자로부터 다양한 비재무정보에 대한 요구가 증가해 이에 대한 중요성이 강조되는 추세다.

② 서비스 방법: 인터넷을 통한 기업데이터 열람과 데이터 연동을 통한 DB 이전으로 구분된다. 전자는 소량 이용자(중소기업), 후자는 대량 이용자(금융기관, 정부, 공공기관, 대기업)가 주로 사용한다.

③ 서비스 경쟁력: 데이터의 양, 신뢰성, 적시성, 최신성, 다양성으로 판단한다. 즉 신뢰할 수 있는 검증된 다양한 정보를 적시에 제공하고, 최신 변동내용을 반영해 공급하는 것이 기업데이터 서비스 산업의 핵심 경쟁요소다.

④ 서비스 이용자 및 용도: 정부(조달입찰 계량지표, 산업현황 파악 및 정책수립 등), 공공기관(지원기업 성과측정, 산업현황 파악 및 정책수립 등), 금융기관(여신기업 신용리스크 관리), 기업(거래기업 신용리스크 관리, 협력기업 평가 계량지표 등) 등으로 구분된다.

다. 기업데이터 수집방법

신용정보회사가 기업데이터를 수집하는 방법을 이해하는 것은 기업데이터와 기업데이터 시장을 이해하는 데 도움이 된다. 신용정보회사가 서비스하는 기업데이터는 수집 방법에 따라 양·신뢰성·적시성·최신성·다양성과 같은 서비스 경쟁력 항목뿐 아니라, 정보획득 가능성과 정보수집 비용에도 큰 영향을 주기 때문이다. 신용정보회사가 기업데이터를 수집하는 주요 방법은 다음과 같다.

1) 신용조사

기업이 공공 조달입찰 참여, 대기업 협력업체 등록, 기술금융 이용, 수출입 신용확인 등의 목적으로 자발적으로 신용조사를 받는 과정에 신용정보회사는 해당 기업의 정보를 수집한다. 이러한 방식으로 수집된 정보는 소정의 신용조사 절차를 거치므로 전체적인 기업데이터의 품질을 일정수준 이상으로 유지하는 데 매우 중요한 역할을 한다. 특히 정보 수요자가 주로 관심을 가지는 기업군과 신용조사를 받는 기업군은 대체로 일치하는 경향이 있으므로, 신용조사를 통한 기업데이터 수집은 정보수요자의 실질적인 목적을 충족시키는 데 매우 중요한 부분이다.

2) 공공정보 조회·구매

획득 가능한 주요 공공정보(금융감독원 전자공시, 등기부등본, 국세청, 4대보험, 특허정보, 인증정보 등)를 조회·구매해 주요 비재무정보를 수집한다. 특히 선별적인 건별 정보수집이 아니라 대량의 정보수집을 위해서는 인적·물적 투자가 병행돼야 한다. 최근에는 IT 기술을 활용해 대량의 공공정보를 효율적으로 수집하는 방법이 널리 사용돼 기업데이터 서비스의 품질이 올라가고 있는 추세다.

3) 금융기관·공공기관 정보관리업무 대행

금융기관·공공기관의 여신기업 등에 대한 정보관리업무를 신용정보회사가 대행해 대량의 기업정보를 수집할 수 있다. 특히 이 방법의 대상이 되는 기업군은 위에서 언급한 신용조사와 마찬가지로 정보 수요자가 주로 관심을 갖는 기업군과 대체로 일치하는 경향이 있다. 그러므로 활발한 경제활동을 영위하는 다수의 중소기업에 대한 재무정보를 수집하는 가장 좋은 방법으로 알려져 있다. 다만 다수의 기업에 대한 정보관리를 단기간에 처리하기 위해 인적·물적 투자가 병행돼야 하는 어려움이 있다.

라. 기업데이터 시장 최신 동향

기술의 발전과 수요자의 요구사항 반영을 통해 자연스럽게 진행되고 있는 기업데이터 시장의 주요한 변화는 다음과 같다.

1) IT 기술 활용

기업데이터를 수작업으로 처리하던 과거와 달리 정보기술을 활용해 대량의 정보를 효율적으로 수집하는 방법이 널리 사용됨에 따라 정보획득 가능성, 정보 커버리지, 정보수집 비용 측면에서 획기적인 발전이 이뤄지고 있다. 정부의 정보공개 3.0 정책에 따라 국민연금 가입자 정보 전체를 오픈API 방식으로 수집해 단기 고용동향을 파악하고, 등기부등본 등의 공부상 정보를 데이터화해 등기 변동정보를

실시간 파악하는 것은 좋은 예다.

2) 모바일 기기 활용

모바일 기기를 활용한 기업데이터 검색서비스가 대중화돼 기업데이터를 소비하는 장소가 책상 위에서 도로 위로 확대되고 있다. 한 번에 볼 수 있는 정보의 양에 한계가 있다는 단점보다 기업데이터를 검색하는 장소에 대한 제약이 없다는 장점이 클 뿐 아니라, 위치기반 서비스와 같은 모바일 기기 고유의 서비스와 결합해 인근 기업검색·지도검색과 함께 새로운 효용이 창출되고 있다.

3) 비금융 목적의 비재무정보 수요 확대

재무정보 위주의 기업데이터를 중요시하는 기존의 금융적 관점과 달리 최근에는 비재무정보를 다양한 방식으로 활용하는 비금융적 관점의 수요가 확대되는 추세다. 예를 들면, 개요정보 변동에 따른 신용리스크 분석, 거래관계를 활용한 생태계 분석, 특허정보를 활용한 기업의 기술력 분석, 지분관계에 따른 중소기업 계열분석, 경영진(개인)을 중심으로 하는 관계기업 분석, 상품을 중심으로 하는 산업 분석과 같은 것이 있다.

4) 신용정보회사가 아닌 정보제공 시장 확대

신용정보회사가 아니면서 기업데이터를 서비스하는 정보제공 시장이 확대되는 추세다. 이 중에서 공적 영역의 서비스(금융감독원 전자공시시스템, 중소기업현황정보시스템, 대법원 인터넷등기소 등기열람서비스, 국가통계포털, 특허정보검색서비스 등)는 기관의 고유 목적을 달성하는 것을 목적으로 한다. 사적 영역의 서비스(특허정보, 입찰정보, 채용정보, 재무정보 등)는 특정 분야의 전문성을 바탕으로 수요자의 요구를 충족해 이윤을 추구하는 것을 목적으로 한다.

5. 특허데이터 분야 서비스

가. 특허정보 개요

우리나라 특허법, 실용신안법, 디자인보호법, 상표법의 목적이 나오는 제1조에 모

[그림 4-2-6] 정보통신기술진흥센터의 ipcast 서비스 화면

두 들어가는 단어나 문장은 무엇일까? 바로 '보호, 도모, 산업발전에 이바지'다. 즉 지식재산권 제도의 궁극적인 목적은 개별 권리를 보호하고 이용이나 신용 유지를 도모함으로써, 궁극적으로는 산업 발전에 도움이 되는 제도를 운용하는 것이다. 그럼 해당 제도의 목적인 산업 발전을 효율적으로 달성하기 위해 가장 손쉽게 활용하기 위한 매개체는 무엇일까? 그것은 바로 해당 권리에 대한 정보의 효율적인 공개·제공·활용이다.

특허, 실용신안, 디자인, 상표 등 산업재산권은 출원인 등의 권리자에게 일정 기간 독점적인 특허권(특허: 20년, 실용신안: 10년, 상표: 반영구(10년마다 갱신 가능), 디자인: 20년)을 주는 대신에 본인의 권리를 일반 대중에게 공개하는 것을 제도화하고 있다.

이 공개된 자료를 바탕으로 권리자의 권리 보호와 침해의 방지·회피를 도모하고, 궁극적으로 재사용 및 더욱 개선된 기술을 만들기 위한 기본 자료로 활용함으로써 기술의 진보를 가져오게 된다. 따라서 특허 등의 지재권 정보의 활용은 연구하는 기술이나 상표 등이 먼저 특허청에 등록·공개됐는지 키워드·도면 중심으로 검색·분석·확인하는 선행기술 조사에 한정됐다.

• 5절 필자 | **강민수**(광개토연구소 대표)

[그림 4-2-7] 한국정보화진흥원의 AI 오픈 이노베이션 허브

하지만 특허정보의 가치가 점차 중요해지면서 특허정보가 갖는 집단적인 속성 정보를 해석 가능함이 곧 기업 기술의 미래 진로를 결정하는 데 가장 중요한 확인 수단으로 부상했다. 이에 발맞추기 위해 대기업·특허청에서는 해당 기술 분야의 비슷한 특허들을 지도로 보여주는 Patent Map을 제공하기 시작했다. 또한 국가의 IP R&D 전략 수립을 지원하기 위해 한국특허개발전략원을 중심으로 핵심 기술별 국가 특허전략 청사진 구축과 활용 사업에 고급 분석 자료를 활용함으로써 해당 정보 가치를 높이고 있다.

한국발명진흥회에서는 특허정보를 바탕으로 개별 특허에 가치평가를 내리는 SMART 시스템을 제공하며, 정보통신기술진흥센터(IITP)는 ipcast 서비스를 통해 분쟁 정보를 바탕으로 분쟁 예보 및 미국 소송 정보에 대한 검색 서비스를 제공중이다.

한국정보화진흥원에서는 특허정보를 기반으로 AI 기술을 효과적으로 활용하기 위해 특허정보를 분석한 지식베이스 구축 사업을 진행하고 있다. 이미 전기·전자분야 100만 건의 특허에 대한 지식베이스를 제공한 바 있으며, 엑소브레인(Exobrain) 2단계 사업으로 특허 분야에 대한 연구를 활발히 진행 중이다.

민간 영역에서는 특허정보와 AI를 결합한 서비스가 출시 중이다. 특허 정보와 논문 정보에서 추출한 기술 키워드 기반 머신러닝을 활용해, 기술의 미래 유망성을 예측하는 시스템을 제공하는 등 특허 정보의 확대 적용 가능성이 보인다.

나. 서비스 현황

1) 공공 서비스 현황

우리나라에서의 특허정보를 활용하는 서비스는 특허청을 중심으로 한 공공영역에서 먼저 다양한 특허정보 서비스를 이끌어 가고 있는 모양새다.

특허정보 검색 서비스를 제공하는 KIPRIS, 특허의 가치평가 정보를 제공하는 SMART, 지적재산권 분쟁 정보를 제공하는 IP-NAVI, 특허정보 기반의 특허 동향 자료를 제공하는 e특허나라 등이 있다.

지식재산을 기반으로 한 비즈니스 영역의 확대로 다양한 서비스의 요구가 증가한 만큼 특허데이터의 보급 요구 역시 증가했다. 이에 KIPRIS Plus 서비스를 바탕으로 특허데이터 보급 서비스를 적극적으로 추진중이다. 특히 한국특허정보원의 데이터 보급은 각국 특허청 간의 합의를 근거로 국내 데이터뿐만 아니라 해외 특허청들의 데이터를 적극적으로 보급해 국내 기업들이 국제 경쟁력을 확보할 수 있게 하고 있다.

① KIPRIS

특허청은 KIPRIS(http://www.kipris.or.kr) 서비스에서 공개되는 모든 지재권 정보를 제공 중이다. 미국·유럽·PCT·일본·중국 등 25개국의 특허정보, 미국·일본·호주·

[그림 4-2-8] KIPRIS 서비스 화면

캐나다·유럽 5개국에 대한 상표 정보, 일본·미국·WIPO·중국의 디자인 정보검색을 지원한다. 특히 행정처리정보, 중간서류 원본 보기, 유사특허, OPD(국제특허 심사공유시스템) 통합조회 서비스, 패밀리 정보, 실시간 번역 정보, 인터넷 기술 공지 등을 제공 중이다.

② KIPRIS Plus

한국특허정보원은 모든 산업재산권 정보를 오픈API, LOD, 데이터세트(Bulk Data(157개), Sheet(32개), Chart(32개), Link(39개))의 데이터 제공 방식으로 실시간으로 제공한다.

현재 제공하고 있는 총 데이터의 양은 2018년 3월 기준으로 공공데이터 202개, 민간데이터 51개로 6억 건이 넘는다.

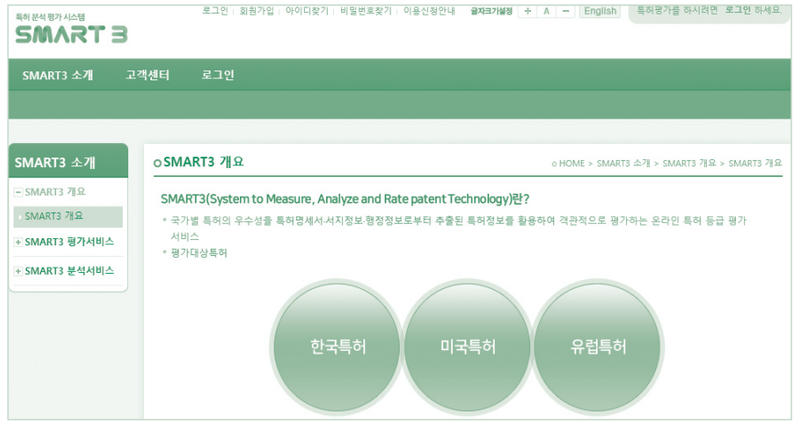
KIPRIS Plus(http://plus.kipris.or.kr) 서비스를 통해 지식재산 관련 업계뿐 아니라, 대학·연구기관·금융권·IT 중소기업 등에서도 IP 금융·IP 평가 분야에 다양하게 지식재산 정보를 활용할 수 있게 하고 있다. 또한 지식재산 데이터 선물 제도를 운용함으로써 예비 창업자 또는 3년 이하 창업 기업의 대표자들에게 무상으로 데이터를 사용할 수 있도록 도와주고 있다.

[그림 4-2-9] KIPRIS Plus 서비스 화면

③ 특허 분석 평가

한국발명진흥회에서는 2010년부터 제공중인 SMART 시스템(https://smart.kipa.org)을 지속적으로 개선해, 현재는 한국 특허 100만 건, 미국 특허 321만 건, 유럽

[그림 4-2-10] SMART 시스템 서비스 화면



특허 95만 건을 대상으로 권리성·기술성·활용성 등 9단계의 가치평가 정보와 특허 경쟁력 분석, 유사 특허 분석 서비스를 제공 중이다.

2) 민간 서비스 현황

방대한 특허 자료의 가공이나 구축 등 과다한 초기 투자비용이 필요한 특허정보 비즈니스의 특성상 국내 특허정보 비즈니스 시장은 해외 시장에 비교해 다소 규모가 작은 편이다. 특허정보 온라인검색 서비스를 제공하는 WIPS, wisdomain, patentpia, keywert 등은 연차료 관리 서비스, 상표 서비스 등 특정 주제의 서비스 중심으로 특화됐다. 특히 기존 특허정보를 온라인으로 제공하는 시스템에는 IP 연구개발 사업의 원활한 지원을 위한 다양한 분석용 서비스, 시각화 기능, 도식화 포맷 등의 특화 서비스를 제공 중이다.

3) 특허 정보와 AI 기술 적용 동향

① AI용 지식베이스

한국정보화진흥원(NIA)에서는 2017년부터 분야별 AI 서비스 활성화를 위한 지식베이스(Knowledge Base) 구축 사업을 시행해, 특허·법률·일반상식·이미지 4개 분야에 대해 지식베이스 구축 사업을 진행했다.

특허 분야는 우리나라 전기·전자(IPC H section) 특허 100만 건에 대해 12만 건의 핵심 키워드별 60종 파라미터, 100만 건 특허별 가공 파라미터 정보, 2,338만

건의 특허 도면 태그 정보에 대한 지식베이스를 벌크 데이터 또는 LOD 정보로 제공 중이다.

② 일본 특허청

일본 특허청은 2016년 6월부터 일본 특허청 내 892개 업무에 대한 전수 조사를 통해 15개 분야, 20개 업무에 대해 AI 적용 가능성을 확인했다. 2017년 1~3월 사이의 검증 작업(자료 수집, 효과 측정, 내부 평가 등)을 통해 2017년 4월에 AI 적용 액션 플랜을 발표했다.

접수·방식심사·특허·상표·디자인 분야에 대한 전화 질의응대, 신청서류 확인, 출원에서 등록상표 사용 확인, 서면 출원 전자화, 특허의 작성서류 오류 확인, 선행 도형상표 조사 및 상표의 업무 분류에 대해서는 기초 연구를 통해 적용 가능성을 확정했다. 또 15개 분야별로 AI를 활용한 서비스 확대 방안을 통해 업무의 효율화를 적극적으로 추진 중이다.

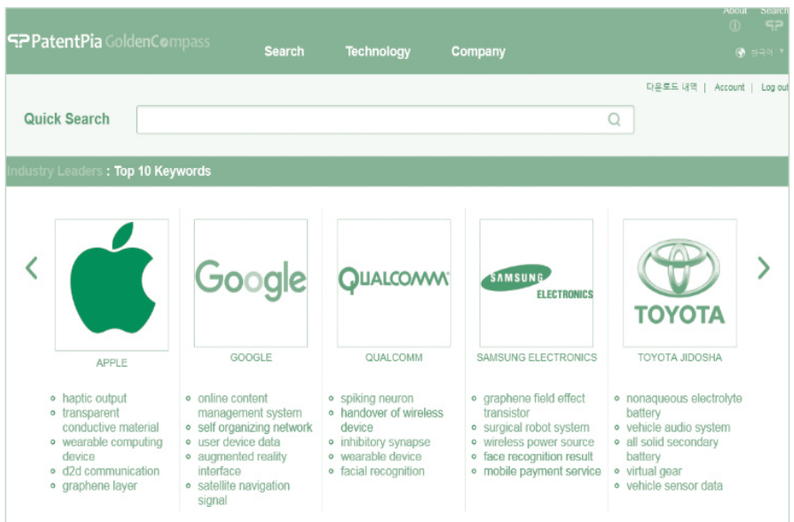
③ 엑소브레인

엑소브레인(ExoBrain)은 자연어를 이해해 지식을 자가 학습하며, 전문 직종에 취업이 가능한 수준으로 인간과 기계의 지식소통이 가능한 SW 개발을 목표로 진행 중이다. 엑소브레인 과제 중 세부과제 2단계 사업 영역으로 2017~2020년까지 3년간 한국전자통신연구원(ETRI) 김현기 박사팀과 한국특허정보원이 함께 특허 분야를 연구하고 있다. 특히 특허 분야 자동 질의응답(챗봇) 및 자동 선행 기술 조사 2개 분야 서비스를 대상으로 집중 연구 중이다.

4) 기술정보에 대한 미래 유망성 예측 서비스

'골든 컴파스'에서는 특허와 논문에서 추출한 700만 개 기술 키워드에 대해 각 키워드의 시계열별 출현 동향을 파악해 100여 종의 파라미터를 만들었다. 이것을 바탕으로 유망성 종합정보, 평가 근거 정보, 연구개발 추천 키워드, Co키워드, 핵심 기업, 핵심 연구자, 주요 이벤트, 신기술 센싱 정보 등을 제공하고 있다.

[그림 4-2-11] 골든 컴파스 서비스 화면



변화하는 DB 활용 서비스

박흥록 | 와이즈모바일 대표
2017 데이터 서비스 이노베이터상 수상자

주차장 이용률은 주차장 운영시간과 사람들의 동선에 따라 다양한 DB 서비스와 접목이 가능하다고 판단했다. 데이터 분석을 통해 기존에 진행해 왔던 날씨와 상호 연관성이 크다는 것도 알게 됐다. 주차에 대한 니즈가 있는 많은 사람에게 필요로 하는 것을 제공해 주면서, 그 부분을 통해 새로운 DB를 만들어 갈 수 있다는 판단이 섰다.

날씨와 관련된 다양한 업무를 12년 이상 진행해 온 필자가 '날씨·기상' 분야가 아닌, 전혀 다른 '주차 DB · 주차 사업'을 추진할 때 주위의 많은 사람들이 질문을 해왔다.
왜 주차장 정보(DB)냐? 당시 필자는 향후 스마트폰의 성장과 앱 서비스에 대해 새로운 도전 차원에서 사업을 시작했다. 2013년에 이르러 다양한 사업 부문을 벤치마킹하면서 미국의 카셰어링 업체인 '집카(ZipCar)'를 알게 됐다.

카셰어링 서비스가 나오기까지

집카는 렌터카 개념에서 벗어나, 이용자가 필요할 때 가장 가까운 곳에 주차된 차를 몇 시간 단위로 빌리는 것을 말한다. 기존의 렌터카가 사업장에 방문해 일 단위로 계약하는 것과 비교해, 집카는 집에서 가까운 주차장에서 시간 단위로 자동차를 빌려 쓰는 서비스다. 집에서 차가 있는 위치까지는 걸어서 5~10분 정도 걸린다. 차를 도시 곳곳의 주차장에 세워놓기 때문에 집카 서비스가 가능했다.
2000년에 등장한 카셰어링은 자동차 한 대가 운행 시간보다 주차된 시간이 더 많다는 데서 나온 아이디어였다(지난 2000년, 첫 번째 집카가 보스턴의 도로를 달렸고 지금은 약 1만 대가 미국·캐나다·영국·스페인·오스트리아의 주요 도시와 300개 대학에서 운행되고 있다). 하지만 초반에는 이용자들이 많이 불편해 했다. 이용자들이 원한 것은 그저 편리함과 적당한 가격이었기 때문이다. 사업 초기에 집카는 그것을 충족시켜주지 못했다. 즉 고객들이 차량을 이용하기 위해 집카 주차장까지 가기에는 너무 멀었다.
하지만 2003년에 새로운 CEO가 취임하자마자 이러한 점을 파악하고 개선해 나가기 시작했다. 가장 먼저 변화를 준 것은 거리 단축이었다. 주거지에서 집카 주차장까지의 도보 거리를 기존 10분에서 5분으로 가까이 배치함으로써 성공적으로 사업을 이끌어갈 수 있었다.

인식의 전환

우리나라에서도 카셰어링 산업이 발전하리라 예측하고 주차장 정보와 주차장을 통한 플랫폼 사업에 관심을 두기 시작했다.
더구나 주차 앱 서비스를 통해 전국의 주차장 DB 구축과 주차장 예약할인 결제 서비스를 접목해 사업을 확대해 나가면서 점점 새로운 분야가 눈에 들어오기 시작했다. 주차장 기본 정보는 크게 변화가 없지만, 주차장을 이용하는 차량과 고객들에 대한 데이터 분석 부분이 더욱 중요하다는 것을 깨닫게 됐다.
변화가 없는 데이터보다는 실시간 이용자 데이터 분석과 주차장 주차 가능면 정보를 통해 더욱 가치 있고 유용한 주차 정보 서비스로 발전하게 된다는 점을 알게 됐다. 특히 주차장 이용률은 주차장 운영시간과 사람들의 동선(평일 및 주말)에 따라 다양한 DB 서비스와 접목이 가능하다고 판단했다. 데이터 분석을 통해 기존에 진행해 왔던 날씨와 상호 연관성이 크다는 것도 알게 됐다. 즉 주차에 대한 니즈가 있는 많은 사람에게 필요로 하는 것(현재의 정보)을 제공해 주면서, 그 부분을 통해 새로운 DB(미래의 정보)를 만들어 갈 수 있다는 판단이 섰다.
주차장 운영에도 땡처리 개념이 필요하다. 기존의 빌딩 주차장 운영 사업자들은 주차 공간에 여유가 있어도 그냥 내버려두는 게 일반적이었다. 바로 여기에 인식의 변화가 필요한 것이다.
향후 타임커머스를 반영한 주차장 큐레이션 서비스와 차량번호 인식 시스템(LPR, License Plate Recognition), 무정차 출차 시스템과 무인화 주차예약 시스템으로 주차 DB 정보 서비스는 발전해 나갈 것이다.
서울의 핫플레이스 지역 주차장들은 요일과 시간에 따라 주차 가능면의 변화가 매우 심하다. 바로 이 부분을 어떻게 분석하고, 여유 주차면을 적극적으로 활용(판매)해 나갈 것인지가 앞으로 이 사업의 발전 방향이 될 것이다. 이를 위해 다양한 센서와 솔루션 개발로 더 발전적인 서비스가 가능해질 것이다.

제 5 부

데이터 솔루션 동향

제1장

데이터 솔루션 아키텍처

데이터는 빅데이터 시대로 진입하면서 과거에 비해 비교할 수 없을 만큼 중요하게 여겨지고 있다. 가까운 미래에는 모든 기기와 사람으로부터 데이터가 생성되고 수집될 것이므로 이 데이터들을 수집, 저장, 처리, 관리, 유통, 분석하기 위한 데이터 솔루션 분야는 데이터를 통해 가치를 확보하는 데 있어서 가장 핵심 역할을 할 것으로 기대된다.

1. 변화의 원동력

정보기술(IT)이 생활을 편리하게 만들어 주는 단순한 도구를 넘어 실세계의 작동 시스템의 근간으로 자리매김하면서 데이터는 비약적으로 늘어나고 있다. 2018년에 들어서 인공지능, 스마트카, 빅데이터, 보안(블록체인), 사물인터넷, 클라우드, 핀테크 등이 주요한 이슈들로 손꼽히고 있다.

이 이슈들의 저변에는 과거와 비교가 불가능한 방대하고 복잡한 데이터의 생성과 변화가 자리하고 있다. 이 데이터를 어떻게 수집하고 저장하고 처리하고 관리할 것인가가 핵심 과제가 되고 있는데, 데이터 솔루션은 새로운 4차산업혁명을 필두로 하는 변화의 원동력으로 인식되고 있다.

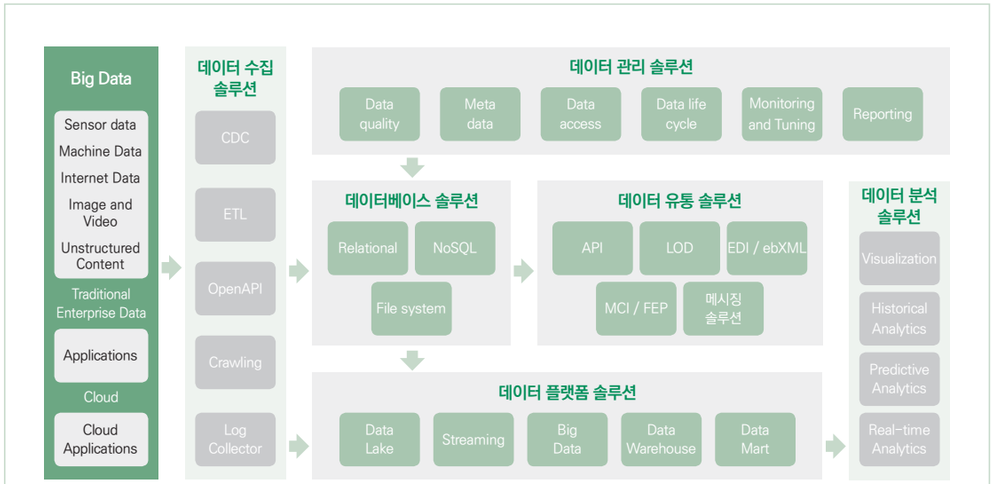
2. 데이터 솔루션

가. 데이터베이스 솔루션

데이터베이스 솔루션은 애플리케이션이 보장해야 할 데이터의 안정적인 저장·처리·조회를 위해 데이터 무결성, 보안, 권한, 트랜잭션 관리, 락킹 등의 기능을 제공하고, 이를 통해 개발 편의성과 안정성, 고가용성을 지원한다. 최근에는 빅데이터

• 필자 | **장재웅**(알티베이스 대표)

[그림 5-1-1] 데이터 솔루션 아키텍처



환경에 적합한 NoSQL 방식의 데이터베이스 솔루션이 각광 받고 있다. 기존의 데이터베이스 솔루션으로는 처리할 수 없는 비정형, 반정형 데이터를 분석해야 할 필요에 따라 Columnar DB, Document DB Graph DB, Key-Value DB 등이 등장했다. 데이터베이스 솔루션 분야에서도 클라우드를 이용한 서비스가 확산될 것으로 예상된다.

현재까지 대부분의 기업은 클라우드 환경에서 DBMS를 이용하더라도 프라이빗 클라우드 위주로 사용하고 있다. 데이터를 외부에 보관하는 것에 거부감을 갖는 기업들이 퍼블릭 클라우드를 이용하는 것을 회피하는 경향이 있지만, 향후에 보안문제가 해결되면 퍼블릭 클라우드와 프라이빗 클라우드를 연동한 하이브리드 클라우드가 보편화될 것이다. 국내에서는 외산 제품이 지배적이지만, 선재소프트·알티베이스·큐브리드·티맥스데이터와 같은 회사들이 국산 DBMS를 보급해 시장 점유율을 늘려 나가고 있다.

나. 데이터 관리 솔루션

데이터 관리 솔루션은 기업이 보유한 모든 데이터에 대해 관리 체계를 수립·실행해 기업 내 데이터의 품질을 향상시키고 유지하는 것을 의미한다. 최근 들어 사람·머신·기업이 과거와는 비교가 불가능할 만큼 대량의 데이터를 만들어내고 있다.

많은 기업은 데이터를 비즈니스에 활용하기를 원하지만 데이터가 증가함에 따라 오히려 데이터의 분산, 중복, 불일치가 발생하고 있다. 이로 인해 데이터 활용도가 떨어지면 기업의 경쟁력이 뒤쳐지는 문제가 발생한다. 따라서 데이터의 표준화, 품질관리, 메타데이터 관리, 데이터 모델 관리, 데이터 접근 관리, 데이터 수명 주기 관리 등의 방법으로 데이터를 관리하기 위한 솔루션이 중요하게 부각됐다. 이 영역의 솔루션을 데이터 거버넌스 또는 데이터 매니지먼트 솔루션이라고 한다.

데이터 거버넌스는 특히 데이터의 가용성, 유용성, 통합성, 보안성을 관리하기 위한 정책 및 프로세스를 포괄한다. 최근에는 구조화하기 어려운 비정형 데이터의 양이 급증하고 있어서 기존 정형 데이터를 분석하고 품질을 확보하는 것이 더욱 중요해지고 있다. 고품질의 데이터를 유지하는 것이 급변하는 시장과 비즈니스 경쟁 환경에 대한 대응력을 높이는 데 있어서 중요한 요소가 되고 있다.

다. 데이터 수집 솔루션

데이터 수집 솔루션은 빅데이터 시대에 접어 들면서 주목을 끌고 있다. 과거에는 락 메커니즘으로 트랜잭션을 보장해 수집한 데이터의 무결성을 유지하는 것이 중요하게 여겨졌다. 하지만 빅데이터 시대에는 데이터의 일관성 유지보다 대량 데이터의 효율적인 수집이 더 높은 관심의 대상이 되고 있다. 수집하는 데이터의 형태도 RDBMS, 파일, 하둡 등으로 다양하고 성격에 따라 로그, 이미지, 비디오, 오디오 등으로 방대하다.

머신러닝이나 딥러닝과 접목하기 위해 애플리케이션 로그, 웹 로그, 장비 로그, 메일 로그와 같은 다크 데이터의 수집에도 관심을 갖게 됐다. 효과적인 수집을 위해 데이터 수집 솔루션은 데이터의 유실과 중복 방지, 데이터 압축, 데이터 정형화, 저장된 데이터의 암호화, 무결성 검증, 사용자 편리성 등을 높이는 방향으로 발전해 나가고 있다.

라. 데이터 유통 솔루션

데이터 유통 솔루션은 개인이나 조직이 데이터를 주고 받기 위한 솔루션을 의미한다. 데이터를 교환하는 기술은 지속적으로 변화·발전하고 있다. 최근의 화두로서 오픈API, LOD 솔루션을 꼽을 수 있다.

오픈API란 공통적으로 사용할 필요가 있는 데이터를 개방해 사용자들이 해당 데이터에 대한 전문적인 지식이 없더라도 쉽게 접근·가공·활용할 수 있도록 데이터를 추상화해 표준화한 인터페이스를 말한다.

LOD(Linked Open Data)는 Linked Data와 Open Data의 합성어다. 누구나 자유롭게 사용하고 재사용 및 배포할 수 있는 데이터를 웹에서 연결할 수 있도록 표준화함으로써 웹사이트를 공유 데이터베이스처럼 활용할 수 있도록 해 준다. 데이터 유통 솔루션으로는 그 밖에도 EDI/ebXML, MCI/FEP, 메시징 솔루션 등이 있다.

마. 데이터 분석 솔루션

데이터 분석 솔루션은 기업에서 데이터를 수집·정리·분석·활용해 효율적인 의사결정을 내릴 수 있도록 지원한다. 궁극적으로 기업의 생산성 증가, 원가 절감, 고객 만족도 향상을 실현해 준다. 최근에는 머신러닝이나 인공지능 기술을 이용해 예측 분석을 하는 방향으로 발전해 나가고 있다. 데이터 분석 솔루션은 쿼리를 이용해 분석하고 시각화하는 것이 주요한 기능이지만, 점차 머신러닝 모형과 알고리즘으로 분석하는 기능이 중요하게 여겨지고 있다. 또한 R, 파이썬과 같은 오픈소스 프레임워크와 연동하거나 이를 지원하면서 발전해 나가고 있다.

바. 데이터 플랫폼 솔루션

그동안 데이터 플랫폼 솔루션은 DBMS, ETL, OLAP와 같은 분야의 제품들을 통해 정보 시스템으로부터 데이터를 추출·변환·적재·분석해 얻은 통찰력으로 효율적인 의사결정을 할 수 있도록 지원하는 도구를 의미했다. 하지만 최근 하둡 기반의 빅데이터 영역에서 규모와 다양성 및 속도 측면에서 전통적인 기술로는 다룰 수 없는 빅데이터를 수집·저장·처리할 수 있는 기술이 발전했다. 그러면서 기존에 구축된 데이터 웨어하우스(DW) 등의 데이터 플랫폼과 빅데이터를 활용한 빅데이터 플랫폼이 융합하면서 빠른 속도로 발전하고 있다.

기존의 DBMS와 하둡 영역에서의 빅데이터를 데이터 플랫폼으로 연결해 통합 저장·분석하기 위한 제품, 서비스, 컨설팅 제공 업체들이 늘어나고 있다. 기존의 DBMS를 토대로 한 리포팅, OLAP, BI 분석과 하둡 기반의 빅데이터 분석, 특히 고급 분석이 통합되고 있다. 여기에 AI·머신러닝 기술이 접목돼, 데이터로부터 빠르고 효율적으로 비즈니스의 인사이트를 얻을 수 있도록 발전해 나가고 있다.

제2장

데이터베이스 솔루션

데이터베이스 솔루션은 데이터를 효율적으로 저장·처리·분석함으로써 기업의 운영과 의사 결정에 근간을 이루는 제품이다. 최근 들어 클라우드가 등장하면서 DBMS를 도입·운영하는 유형에서 변화가 일고 있으며, 빅데이터 분석을 위한 NoSQL 분야도 성장하고 있다. DBMS에 대한 요구사항이 다양해지고 기술력이 향상되면서 오픈소스 DBMS도 빠르게 확산되고 있다. 시장 점유율은 외산 DBMS의 비중이 높지만, 최근 들어 국산 DBMS의 약진도 두드러지고 있다.

1. 개요

데이터베이스 관리 시스템(DBMS)은 데이터베이스(DB)를 조작하는 별도의 소프트웨어를 둬 여러 사용자가 DB 내의 데이터를 관리하기 쉽도록 하는 소프트웨어 도구를 말한다. 다시 말해 DBMS는 데이터를 정의하고·읽고·쓰고·갱신하는 등의 데이터 조작과 이들을 효율적으로 관리하는 프로그램의 요구를 처리·응답해 데이터를 사용할 수 있도록 만들어 주는 역할을 한다.

DBMS는 파일 시스템에서 데이터를 관리하는 방식의 한계를 극복하기 위해 개발됐다. DBMS는 파일 시스템으로 데이터를 관리할 때의 문제들, 즉 데이터의 중복성, 데이터의 비일관성, 통합 관리와 복구의 곤란, 여러 사용자의 동시 접근성 문제 등을 해결하기 위해 개발됐다. DBMS를 이용해 개발하면 데이터의 접근성이 용이하고, 통제가 강화되며, 애플리케이션을 쉽게 개발·관리할 수 있으며, 보안이 강화된다는 장점이 있다.

DBMS는 계층형, 네트워크형, 관계형으로 발전돼 왔다. 최근에는 관계형 DBMS(RDB, Relational DBMS)가 널리 사용되고 있다. RDB는 관계형 데이터 모델에 기초를 둔 DB로서 모든 데이터를 2차원의 테이블 형태로 표현한다. 이 관계형 DB는 개체의 의미 중심이 아닌, 데이터 간 상관관계에서 개체 간 관계를 표현한 것이라 할 수 있다.

RDB는 데이터 무결성, 보안, 권한, 트랜잭션 관리, 락킹 등과 같이 이전 애플리케이션에서 처리해야 할 많은 기능을 지원한다. 모델링이 간단하고 SQL 문법을 제공해 개발 편리성을 올려 줌으로써 현재까지 일반적으로 사용되는 DB로서 자리매김하고 있다.

일반적으로 국내에 많이 알려진 DBMS로는 해외 제품 중에서는 오라클 Oracle, MS SQL Server, IBM DB2, SAP HANA 등이 있다. 국내 제품 중에서는 선재소프트 Goldilocks, 알티베이스 Altibase, 큐브리드 CUBRID, 티맥스데이터 Tibero 등이 있다.

2. 시장 현황

전 세계 DBMS 시장은 연평균 5~7%대의 꾸준한 증가추세다. 2017년에 전 세계 시장은 60조 원에 이르렀으며, 국내 시장은 약 6,500억 원에 달하고 있다. 이중 대부분은 전통적인 영역의 RDB가 차지하고 있다. 우리나라에서 가장 많이 사용되는 외산 RDB는 Oracle, MS SQL Server, IBM DB2이며 시장의 약 91%를 차지하고 있다. 국내 DBMS 기업들은 글로벌 기업에 비해 비중이 높지 않지만, 최근 들어 꾸준한 증가세를 보여줘 향후 조만간 10%를 넘어설 것으로 기대된다.

하지만 최근 DBMS 시장은 급변하는 IT 환경과 맞물려 빅데이터, NoSQL, 인메모리, 컬럼, 클라우드, 오픈소스 등과 같은 이슈들에 둘러 쌓여 큰 변화가 일어나고 있다.

센서와 장비들로부터 생산되는 전대미문의 대량 데이터를 분석하기 위해 빅데이터 분석·처리 방법들이 개발되고 있다. 특히 디스크와 서버의 성능이 대폭 향상되고, 오픈소스 분석용 SW들이 발전하면서 실시간 빅데이터 분석 기술이 확산되고 있다. 하둡, 스파크(Spark)와 같이 빅데이터 관련 기술들이 발전하고 있으며, Columnar를 통해 데이터 압축 및 빠른 처리를 장점으로 하는 DBMS들이 주목을 받게 됐다. 전통적인 구분법에서는 분석계 DBMS를 OLTP에 반하는 개념으로서 OLAP로 지칭했으나, 최근에는 OLAP와 OLTP 모두를 포용하는 DBMS들이 주종을 이루고 있다. 특히 대용량 빅데이터에 대한 실시간 분석 요건의 충족을 위해 인

• 필자 | **장재웅**(알티베이스 대표)

메모리가 필수적인 기술로 대두되고 있다.

클라우드 시장의 도래는 고객이 DBMS를 소비하는 방식에 변화를 가져왔다. 그동안 고객이 기술 검토, 아키텍처 설계, PoC(Proof of Concept), BMT(Benchmark Test) 등을 통해 도입을 결정했다. 하지만 클라우드 환경에서는 복잡한 도입 과정이 생략된다. 소규모의 자원을 활용해 상용화한 후 필요하면 증설하는 방식으로 바뀌었다. 아마존의 AWS, 오라클 클라우드, 마이크로소프트 애져(Azure) 등이 선도 상품으로서 역할을 하고 있다. 클라우드 사용이 올라감에 따라 점차로 클라우드에서 고객이 체감하는 접근성과 편리성이 DBMS 선정의 기준이 되고 있다.

전 세계 DBMS 시장에서 오픈소스 DBMS가 차지하는 비율은 이미 2015년도에 45%를 상회했다. 국내에서도 2016년에 58%를 넘어섰고 지금도 가파르게 증가하고 있다. 서드파티 솔루션과의 연동이 부족하고 패치, 업데이트가 느리다는 단점이 지적되기도 하지만, 제품 완성도가 올라감에 따라 레퍼런스가 쌓이면서 점차 핵심 업무 영역까지 확산되고 있다. 특히 국내에서는 공공 기관, 통신사, 포털 등에서 활발하게 도입하고 있다. 시장 조사기관들은 2020년까지 70% 이상의 서비스가 오픈소스 DBMS를 사용할 것이라고 전망한다.

3. 기업 동향

가. 오라클·MS·IBM·아마존

오라클은 클라우드가 확산됨에 따라 Oracle 고객을 '오라클 클라우드 서비스'로 유치하기 위해 노력하고 있다. Oracle을 온프레미스 방식으로 사용하던 고객이 클라우드와 묶어서 하이브리드 방식으로 DB를 관리할 수 있도록 '오라클 클라우드앳 커스터머(OCC)' 서비스를 제공하고 있다.

마이크로소프트는 자사의 클라우드 서비스인 애져(Azure)를 강화하면서 단일 데이터 플랫폼에서 데이터와 관련된 모든 것, 비정형 데이터의 처리, DW, BI 등을 SQL Server 통합 서비스로 지원하고 있다.

IBM 역시 클라우드 서비스로서 DBMS를 제공하는 데 주력하고 있다. 아마존은 앞서가는 클라우드 서비스 제공업체로서 지위를 활용해 RDS라는 DB 서비스를 통해 MySQL, MariaDB와 같은 RDB들의 편리성, 가용성을 개선하는 한편 아마존의 자체 DBMS인 Amazon Aurora의 저변 확대를 꾀하고 있다.

나. 선재소프트

선재소프트는 인메모리 DBMS 전문업체로 출발해, 여러 노드에 데이터를 분산해 관리하더라도 하나의 DBMS처럼 동작할 수 있는 NewSQL 시장까지 진출했다. NewSQL 기술은 기존 RDB의 장점인 SQL 및 ACID 트랜젝션을 지원할 뿐 아니라 RDB가 약했던 빅데이터 및 선형적인 성능 향상을 스케일 아웃으로 해결한다. 선재소프트는 현재 OLTP 시장에 주력하고 있으며 IBM, 레노버, HPE, 인텔 등 기술 리딩 업체와 협력해 글로벌 시장 진출을 모색하고 있다.

다. 알티베이스

알티베이스는 1999년에 ETRI의 바다프로젝트에 뿌리를 두고 개발된 RDB다. 지난 20여 년간 650여 개의 고객사에 공급했으며, 6000여 개의 서비스에 적용됐다. 2004년에 세계 최초로 하이브리드 DBMS를 개발해 2006년 근로복지공단, 국방부, LG디스플레이에 공급한 것을 시작으로 외산 DBMS를 교체하는 사업을 성공적으로 수행하고 있다. 지속적으로 해외 수출도 늘려 나가고 있다. 매출액 대비 연간 수출 비중은 15%에 이르며, 미국·중국·일본·터키·베트남·인도네시아·우즈베키스탄 등에서 고객사를 확보하고 있다. 2018년 2월에 오픈소스로 전환해 GitHub에 공개했다. 분산 데이터 환경에 적합한 지능형 클러스터링 기술을 통해 클라우드 환경에서 고가용성을 구현하려는 고객 요구에 대안을 제시하고 있다.

라. 큐브리드

큐브리드는 행정안전부 온나라 2.0, 국가기록원 기록관리시스템(CRMS)의 중앙행정기관 확산에 힘입어 국가정보자원관리원 G-클라우드의 DB 인스턴스 개수가 500개를 넘었으며, 국방통합데이터센터 클라우드 인프라에도 지속적으로 응용시스템이 전환되면서 공공·국방부문 프라이빗 클라우드의 표준 DBMS로 자리매김하고 있다. 또한 프라이빗 클라우드 성공사례를 바탕으로 한국인터넷진흥원, 한국교육학술정보원 등 공공부문 퍼블릭 클라우드 시장으로도 영역을 확대하고 있으며, KT, NBP(Naver Business Platform) 등 다양한 CSP(Cloud Service Provider) 사업자들과 협력을 강화하고 있다. 오픈소스 SW에 대한 사용자 인식이 개선됨으로써

오픈소스 DBMS에 대한 수요가 증가하고 있으며, 지속적인 제품 개발 및 기술 지원을 강화해 매년 20% 이상 성장하고 있다.

마. 티맥스데이터

티맥스데이터는 병원, 금융, 제조, 공공 등 다양한 영역에서 자사 DBMS인 티베로의 핵심업무 적용을 더 폭넓게 확장하면서 신뢰를 높여가고 있다. 또한 국내외 클라우드 CSP(Cloud Service Provider)들과의 제휴를 지속적으로 확대해 온프레미스뿐 아니라 클라우드 환경에서도 티베로를 원활히 지원하고 있다. 최근 IT 환경 변화의 핵심 요소가 데이터 통합과 지능화임을 인식하고 새로운 데이터 아키텍처 모델을 선제적으로 수립·제시했다. 2018년 7월, 'TmaxDay'에서 모든 형태의 데이터를 통합 처리·분석할 수 있고, 빅데이터를 활용한 고도의 지능화한 서비스를 제공하는 '지능화한 클라우드 데이터 아키텍처' 모델을 선보였다. 그 중심에 개발 중인 차기 티베로7 DBMS가 있다.

4. 제품 동향

가. Oracle · SQL Server · DB2

오라클은 최근 '자율주행 DBMS' 기술을 제시하고 머신러닝 기술을 활용해 DBA의 업무를 자동화하고, 휴먼 에러를 없애 가용성·고성능·보안성 등을 강화했다. 오라클은 DB 전문가 없이 클라우드 기반 데이터 웨어하우스를 구성할 수 있는 편리함을 장점으로 내세우며 인덱스 구축, 데이터 재편, 튜닝 등의 자동화를 강화하고 있다. 마이크로소프트는 리눅스 운영체제를 지원하면서 머신러닝, 딥러닝 기술 강화 차원에서 SQL Server에 R과 파이썬을 탑재하고 AI 알고리즘을 사용할 수 있도록 했다. IBM은 DB2의 Oracle PL/SQL 호환 기능을 강화해 기존 애플리케이션을 수정하지 않고 Oracle에서 DB2로 쉽게 이전할 수 있도록 하는 데 주안점을 두고 있다.

나. Altibase

알티베이스는 2017년 7월에 Altibase v7를 출시하고 2018년에 오픈소스로 전환했다. v7은 외산 DBMS를 사용하는 고객들이 Altibase로 이전해오기 쉽도록 Oracle 스타일의 PSM과 함수를 추가하고 PVO 과정 최적화로 안정성과 신뢰성을 향상하는 데 주안점을 두고 개발됐다.

IPC보다 성능이 빠른 로컬 커넥션 타입인 IPC-DA와 향상된 실행계획(Plan)을 제공하며, Eager 방식의 리플리케이션(Replication) 성능을 대폭 향상시켰다. 하이브리드 파티션 테이블을 지원하므로 메모리와 디스크 저장 공간을 다양한 파티션으로 묶어서 구성할 수 있다. 또한 기존의 샤딩 기술을 기반으로 데이터 노드 증가 시에도 성능 저하가 없다. 특히 애플리케이션의 변경이 필요 없는 IDC(Intelligence Data Clustering)를 지원해 오라클의 RAC(Real Application Cluster)를 대체할 수 있는 기능도 실현했다. 향상된 '마이그레이션 센터'를 통해 기존 DBMS로부터 Altibase로의 전환을 쉽게 해주며, DB-Link와 CDC를 통해 기존 DBMS와 유연하게 연동해 ROI도 높일 수 있다.

다. CUBRID

큐브리드는 2017년 8월, 단일 노드에서의 성능과 확장성을 향상시키는 데 주안점을 둔 CUBRID 10을 출시했다. 성능 측면에서 이전 버전인 CUBRID 9.3 대비 TPC-C 벤치마크의 최적 워크로드를 두 배 정도 끌어올렸으며, 최대 성능 대비 가격(tpmC)은 80% 증가했다.

기능 측면에서는 CTE(Common Table Expression)를 추가함으로써 응용 개발자가 재귀적 쿼리(Recursive Query)를 포함한 복잡한 쿼리를 쉽고 명확하게 작성할 수 있게 했다. 글로벌 사용자를 위한 타임존(Timezone) 데이터 타입을 추가하고 관련 함수도 지원한다. 또한 워크로드가 많은 경우에도 HA(High Availability) 기능의 안정성을 강화했으며, 복제 지연도 현저하게 줄였다. 관리도구인 CUBRID Manager와 마이그레이션 도구인 CUBRID Migration Toolkit 제품도 기능을 강화해 사용자 편의성을 높였다.

라. Goldilocks

선재소프트는 2017년 7월, 완벽한 액티브-액티브(Active-Active)를 지원하는 Goldilocks 3.1을 출시한 후 24시간 안정적인 운영을 필요하는 가상화폐거래소, 증권, 통신 분야에 속속 공급하고 있다. Goldilocks는 서비스 운영 중에 용

량을 늘릴 수 있는 스케일아웃 기능을 제공하므로 초기 예산을 적게 잡고 서비스 런칭 후 상황에 따라 쉽게 서비스의 용량을 늘려나갈 수 있다. 메모리 RAC 기능을 통해 안정적인 운영도 지원한다. 특히 국내 DBMS 최초로 TPC-C(On-Line Transaction Processing Benchmark) 국제 인증을 획득해 TPC.org에 등재됐으며, 품질의 우수성을 인정 받아 2017년 대한민국 소프트웨어 품질대상을 받았다.

마. Tibero

Tibero DBMS는 뛰어난 호환성과 공유 디스크 기반의 액티브 클러스터링을 적용한 세계적 수준의 국산 제품이다. 새롭게 개발중인 차기 Tibero7은 티맥스데이터가 제시하는 '지능화를 위한 클라우드 데이터 아키텍처' 모델의 중심에 위치한다.

클라우드에서 요구되는 DBMS 노드의 자유로운 확장과 데이터 저장 노드의 무한 확장을 지원한다. 이를 위해 현재 대부분의 DBMS가 유지하고 있는 단일구조(Monolithic)를 배제하고 데이터 서버와 스토리지 서버로 이원화한 2티어 구조를 채택했다. 또한 정형 및 비정형 데이터를 모두 수용하는 통합 분산 스토리지, SQL API, 지능화한 관리 도구를 제공한다. 즉 OLTP에서 OLAP, 빅데이터까지 시스템 및 업무 제약 없이 활용 가능한 기반 DBMS로서의 역할을 수행한다.

5. 향후 전망

향후 데이터 솔루션은 클라우드와 온프레미스, RDB와 NoSQL, OLTP와 OLAP, 온디스크와 인메모리가 경쟁하면서 컨버전되고, 현실 세계의 비즈니스를 반영해 응용 서비스로 연결하는 IT 혁신의 속도가 가속화되면서 더욱 빠른 속도로 발전할 것이다.

특히 클라우드 시장이 성장하면서 클라우드 환경에 최적의 기능과 특성을 가진 DBMS들이 주목을 받고 있다. 기업이나 사용자가 클라우드를 이용하면 유연하게 자원을 확장할 수 있어 갑작스런 시스템 부하에 대비할 수 있으며, 저렴한 초기 도입 비용, 유지보수 인력 절감 등의 효과를 누릴 수 있다. 클라우드 서비스 업체들은 DBMS를 신속하고 편리하게 구축할 수 있도록 지원하고, 다양한 서비스를 안정적으로 확장해 나갈 수 있도록 제공하기 위해 노력하고 있다.

현재는 기업들이 클라우드에서 대외 업무, 데이터 수집 관련 업무 등 중요도가 상대적으로 낮은 업무를 운영하고 있다. 하지만 점차로 클라우드 사용률은 늘어나고, 온프레미스와 클라우드를 결합한 하이브리드 DBMS를 구축하는 사례 역시 확산일로를 걷게 될 전망이다.

클라우드 서비스를 제공하는 벤더사의 전략에 따라 특정 DBMS가 주목을 받는 기회를 낳기도 한다. 클라우드를 기반으로 개발된 응용프로그램의 경우에는 더욱 그러한 현상은 강화되기 마련이다. 특히 최근에는 클라우드에서 제공하는 NoSQL이 다양한 분야에서 널리 활용되고 있다. 데이터가 일관성이 없고 복잡한 쿼리를 사용할 수 없는 환경에서 NoSQL은 스키마가 필요 없고, 데이터 분산이 용이하기 때문에 향후에 더욱 더 사용자 층이 두터워질 전망이다. NoSQL은 특성에 따라 Wide Columnar DB, Document DB, Graph DB, Key-Value DB 등으로 분화·발전되고 있다.

더불어 오픈소스의 약진이 두드러질 것으로 예상된다. 역사가 오래되고 공고한 시장 점유율을 가지고 있는 RDB를 제외하고 최근의 대부분의 신생 DBMS들은 오픈소스 방식으로 시장을 넓혀 나가고 있다. 오픈소스 방식을 채택하면 기업 외부의 개발자와 협업이 원할하기 때문에 사용자 확산이 빠르다는 강점이 있다.

제3장

데이터 관리 솔루션

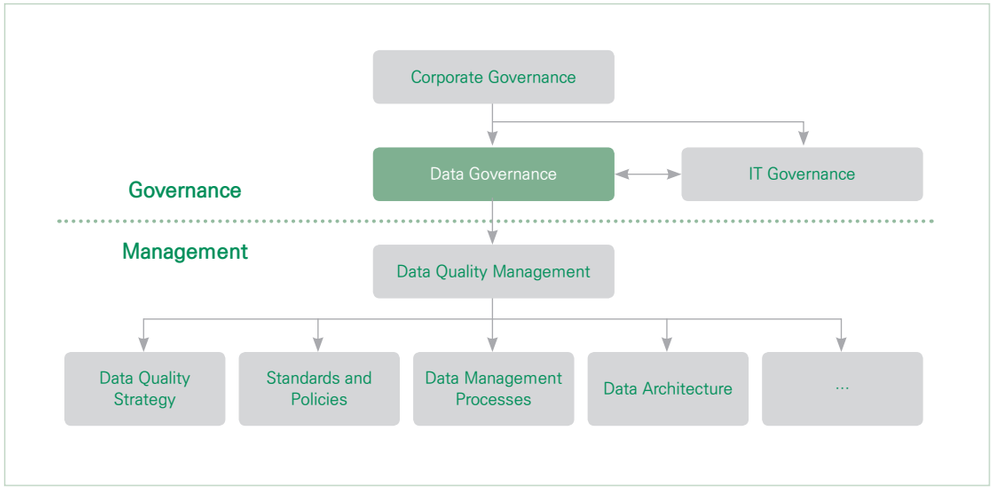
3차산업혁명을 넘어 곧 다가올 4차산업혁명시대는 '초연결성(Hyper-Connected)'과 '초지능화(Hyper-Intelligent)'가 주요 관심사다. 사물인터넷, 클라우드 컴퓨팅 등 정보통신기술을 통해 인간과 인간, 사물과 사물, 인간과 사물이 상호 연결될 것이다. 또한 다양한 방법과 수단으로 다양한 형태를 가진 대용량 데이터가 수집·활용될 것이다. 데이터는 체계적이고 효과적인 방법으로 관리돼야 그 가치를 높일 수 있고, 비즈니스 측면에서도 활용할 수 있다. 이 글에서는 다양한 데이터 관리 방법과 솔루션을 거버넌스 관점과 매니지먼트 관점으로 나눠 소개한다.

1. 개요

4차산업혁명시대의 데이터 수집은 빠른 속도, 다양한 수집 채널, 다양한 데이터 형태, 대규모가 특징이다. 이런 데이터들은 비즈니스 측면에서 보면 무궁무진한 가치를 지니고 있다. 반면 출처를 알 수 없는 정보들이 도처에 넘쳐나는 '정보 과부하 현상'과 왜곡된 정보의 유통은 새로운 문제로 대두되고 있다. 데이터의 폭발적인 증가로 양질의 데이터뿐 아니라 낮은 품질의 데이터 유통 역시 늘고 있다. 이는 기업이 비즈니스 프로세스를 지원하고, 개인정보보호와 같은 규정을 준수하고, 정확한 정보에 입각한 의사결정을 내리는 데 점점 더 위협적인 요소가 되고 있다. 데이터의 주체, 즉 사람인지 기계인지 확실하게 알 수 없고, 어떤 업무 처리 프로세스 과정에서 자연적으로 만들어지는지, 인위적으로 생성되는지도 알 수 없다. 이 때문에 데이터 수명주기(Data life cycle)[1] 관점, 즉 '수집 → 저장·관리 → 이용·제공 → 파기'의 단계별로 비즈니스 영역과 정보기술(IT) 영역에서의 요구사항을 충족하도록 데이터를 관리하는 것이 필요하다.

월리와 로스[2]는 자신의 논문에서 데이터 관리를 [그림 5-3-1]과 같이 거버넌스와 매니지먼트로 구분해 비즈니스 관점과 정보기술 관점에서 관리하는 방법을 제시했다. 데이터 거버넌스는 데이터 관련 의사결정을 비즈니스 관점에서 효과적이고 체계적으로 내리고, 그 의사결정의 책임소재를 분명하게 하는 개념이다. 데이터 매니지먼트는 의사결정뿐 아니라 정보기술을 이용해 구현하는 것까지 포함하는 개념이다. 이와 같이 거버넌스와 매니지먼트는 다소 차이는 있지만 비즈니스 관점에서 데이터를 관리함으로써 데이터 가치를 극대화해 기업에서 활용할 수 있게 한다는 점은 같다.

[그림 5-3-1] 데이터 거버넌스와 매니지먼트



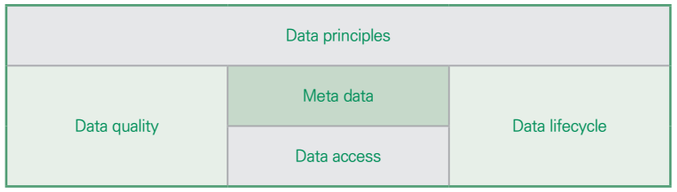
2. 데이터 관리 분야

카트리와 브라운[3]은 [그림 5-3-2]처럼 다섯 가지 상호 연관된 의사결정 영역을 포함하고 있는 데이터 거버넌스 프레임워크를 제안했다. 여기서 의사결정은 기업이 데이터를 자산으로 활용할 때 비즈니스 관점에서 행해야 하는 프로세스를 말한다. 이들 다섯 가지 영역 각각에 대한 관련 솔루션 동향을 살펴보자.

1 개인정보의 수명주기를 참고했다.
2 P. Weill and J. W. Ross, IT governance: How top performers manage IT decision rights for superior results. Harvard Business School Press, Boston, MA, 2004.
3 V. Khatri and Carol V. Brown, Designing Data Governance, communications of the acm, vol. 53 no.1, 2010.01.

• 필자 | **이상옥**(데이터스트림즈 본부장) · **박원환**(충남대 핀테크보안연구센터 교수)

[그림 5-3-2] 데이터 거버넌스 프레임워크



가. 데이터 원칙

데이터 원칙(Data principles)은 비즈니스와 연계돼야 효과적이다. 예를 들어 기업이 사내 비즈니스 프로세스의 표준화를 결정하는 의사결정을 조직차원에서 해야 한다고 가정하자. 이때 각 데이터 자산과 관련된 비즈니스 소유자[4]가 누구인지를 명확하게 확인하는 것이 가장 먼저일 것이다. 데이터를 비즈니스에서 활용할 용도가 명확해지면, 데이터 원칙에는 해당 데이터를 해당 기업의 자산으로 확정해야 한다. 이후 관련 내부 정책과 표준, 가이드라인 등을 마련해야 한다. 이때 해당 데이터를 기업 자산으로 유지하면서, 데이터를 공유하고 재사용하는 방법과 절차 즉 '데이터 원칙'을 마련해야 한다. 이와 같은 원칙은 이론적 근거와 숨은 의미까지도 뒷받침해야 한다. 데이터 원칙은 수탁자에 의한 서비스 공급자의 고객 데이터와 같은 외부 데이터 사용도 고려해야 하고, 비즈니스에서 데이터 사용에 영향을 줄 수 있는 규제 환경도 고려해야 한다.

데이터 원칙은 정보시스템 담당자는 물론 비즈니스 사용자[5]에게 데이터와 관련된 바람직한 행동지침과 프로세스를 정의하는 것이라 할 수 있다. 예를 들어 비즈니스 사용자는 데이터 품질은 물론 수명주기, 해석 가능성, 액세스를 관리하는 데 중요한 역할을 한다. 반면에 정보시스템 담당자는 비즈니스 소유자(또는 데이터 소유자)가 작업하는 과정에서 발생하는 데이터 품질 문제 등을 해결하는 솔루션을 선정·적용·실행하는 역할을 한다. 앞서 살펴본 것과 같이 데이터 원칙은 직접적으로 데이터를 관리하는 것은 아니다. 실제 데이터를 관리하는 것은 [그림 5-3-2]에서 소개하듯이 다른 의사결정 영역의 작업에서 이뤄진다.

4 비즈니스 소유자(Business Owner)는 임의의 데이터를 사용해 영업, 마케팅 등과 같은 비즈니스 활동을 하는 관계자를 말한다.

5 비즈니스 사용자는 비즈니스에서 데이터를 사용하는 자를 말한다.

나. 데이터 품질

데이터 품질(Data quality) 문제는 해당 기업의 비즈니스 운영에 상당한 영향을 준다. 공산품 품질과 같이 데이터 품질도 사용자의 요구 사항을 충족시켜야 한다. 데이터 품질은 정확성, 적시성, 완전성, 신뢰성 등 다차원적인 요소로 결정된다. 이 요소들은 상대적이며, 데이터의 최종 사용과 관련이 있다. 예를 들어 보험회사는 잠재 고객인 의사의 이름, 주소, 전화번호 등의 데이터를 85%의 정확도로 확보했다. 이는 받아들일 수 있는 수준의 데이터 품질이라 할 수 있다. 하지만 처방의사에게 약물에 대한 리콜을 알릴 필요가 있는 약국에 대한 정보라고 생각해보자. 이때는 85%의 정확도는 받아들일 수 없는 수치가 된다.

데이터 품질 관련자는 데이터 품질 관리자, 데이터 품질 분석가, 데이터 품질 트레이너, 주제 전문가 등이 있다. 데이터 품질결정 영역(Data Quality Domain)은 이들의 역할 범위라 할 수 있다. 여기에는 데이터 품질과 관련된 다차원 요소의 표준과 데이터를 비즈니스용으로 전달하기 위한 메커니즘을 정하는 내용이 포함된다. 또한 데이터 품질을 평가하는 절차를 규정하는 것도 포함된다.

데이터 품질 관련 솔루션은 데이터의 정확성, 적시성, 완전성, 신뢰성을 지원할 수 있어야 한다. 또 데이터 품질 관련자들이 작업하는 데 도움을 주는 솔루션도 포함돼 있어야 한다. 데이터 품질관리 솔루션으로 국내 제품은 DQ Miner(지티원), DQ#(엔코아), QualityStream(데이터스트림즈), Wise DQ(위세아이텍) 등이 있고 외국 제품으로는 Profile Stage(IBM) 등이 있다.

다. 메타데이터

메타데이터(Meta data)는 '데이터에 대한 데이터(Data about data)'로 정의되며, 데이터가 무엇인지에 대해 설명한다. 이는 데이터 표현 시 간결하고 일관된 설명을 위한 메커니즘도 제공해 데이터의 뜻(meaning)이나 의미(semantics)를 해석하는 데 도움을 준다. 싱그 등[6]은 메타데이터를 물리적, 도메인 독립적, 도메인별, 사용

6 29. Singh, G., Bharathi, S., Chervenak, A., Deelman, E., Kesselman, C., Manohar, M., Patil, S., and Pearlman, L. A metadata catalog service for data intensive applications. In Proceedings of the ACM/IEEE SC2003 Conference on High Performance Networking and Computing. (Phoenix, AZ, 2003)

자 메타데이터와 같이 분류하고 있지만, 일반적으로는 비즈니스(business)와 기술(technical) 메타데이터로 분류하고 있다. 이와 같은 메타데이터는 데이터 검색, 데이터 정렬, 분석에 중요한 역할을 하므로 체계적인 관리가 필요하다.

- **물리적 메타데이터**는 가장 낮은 수준의 메타데이터며, 물리적 저장 영역에 대한 정보가 포함된다.
- **도메인 독립적 메타데이터**에는 데이터 작성자와 변경자, 데이터와 관련한 권한, 감사, 계보 정보와 같은 설명이 포함된다.
- **도메인별 메타데이터**는 부서 차원의 도메인과 기업 차원의 도메인을 각각 지정할 수 있는 것을 말한다. 부서 차원에서는 개별 단위의 응용 프로그램과 관련된 데이터에 대한 설명을, 기업 차원에서는 기업 전체에 대한 도메인별 데이터의 설명을 포함한다.
- **사용자 메타데이터**는 사용자가 접근한 데이터 항목이나 데이터 모음(테이블)과 연결할 수 있는 주석을 포함한다. 이런 주석은 사용자가 참조한 내용이나 사용 이력이 될 수 있다.

일반적으로 기업에서 사용하는 메타데이터는 수명주기, 데이터의 접근(access)에 종속된다. 데이터 검색과 분석 지원용 메타데이터는 전사 데이터 아키텍처 담당자와 데이터 모델링 엔지니어와 같은 역할을 할 수 있다. 메타데이터를 표준화하면, 데이터를 해석하는 데 효과적으로 사용하고, 관련 데이터를 추적할 수 있다. 비즈니스 환경이 변화하면 조직의 비즈니스 수행방식과 관련 데이터도 변경돼야 한다. 마찬가지로 메타데이터도 변경 사항을 관리해야 할 필요가 있다. 즉 메타데이터 관련 솔루션은 관리에 요구되는 사항을 기능에 포함시켜야 한다. 메타데이터 관련 솔루션으로 국내 제품은 Meta Miner(지티원), MetaStream(데이터스트림즈), Meta#(엔코아), WiseMeta(위세아이텍) 등이 있고, 외국제품으로는 Alation(에이레이션), InfoSphere Information Governance(IBM), Magic-Quadrant(인포매티카) 등이 있다.

라. 데이터 접근

데이터 접근(Data access)은 비즈니스나 데이터 관리 측면을 고려해 해당 데이터

에 대해 이용할 수 있는 권한이나 접근 권한을 주는 것을 말한다. 데이터 보안 담당자는 효과적인 위험 분석을 위해 비즈니스의 데이터 요구사항을 먼저 파악해야 한다. 이후 데이터의 기밀성, 무결성, 가용성 등을 고려해 데이터 접근 정책을 세워야 한다. 더불어 데이터 접근에 대한 실시간 모니터링을 통해 어떤 데이터에, 어떤 방법으로, 누가 접근해 수정이나 삭제했는지를 추적하는 프로세스도 필요하다. 데이터 접근 관련 감사, 개인정보 보호, 데이터 가용성에 대한 외부 요구사항을 정의하는 프로세스도 필요하다. 전사적 관점에서 스토리지의 효율적 활용을 위한 데이터 마이그레이션, 즉 활용도가 높은 데이터 볼륨에서 활용도가 낮은 볼륨으로의 데이터를 이동시키기 위한 데이터 접근도 필요하다. 데이터 접근 관련 솔루션으로 국내 제품은 Chakra Max(웨어밸리), DBSafer(Pnp시큐어), Face(파수닷컴) 등이 있고, 외국 제품으로는 Oracle Database Vault(오라클), Oracle Label Security(오라클), InfoSphere Guardium DAM(IBM) 등이 있다.

마. 데이터 수명주기

모든 데이터는 수명주기(Data life cycle) 단계가 있다는 점은 데이터 거버넌스에서 중요하다. 병원에서 관리하는 전자건강기록(EHR, Electronic Health Records)을 데이터 흐름 관점에서 살펴보자. 임의의 환자가 수술을 받거나 급성치료센터로 옮길 경우, 그 환자와 관련된 정보의 용도와 값은 변경된다. 진찰을 받고 퇴원하면, 병가관리에서 건강관리로 데이터가 전환된다. 이와 같이 데이터를 사용하는 방법과 보관해야 하는 기간을 이해함으로써 조직은 데이터의 사용 패턴을 최적의 저장 미디어와 연계할 수 있다. 수명주기 동안 데이터를 저장하면 소요되는 총비용을 최소화할 수 있다.

많은 기업이 자신이 보유한 데이터의 종류, 데이터의 중요성, 중요 데이터의 출처, 데이터 자산 중복의 정도를 잘 알지 못하고 있다. 데이터를 효과적으로 관리하기 위해 저장매체 담당자는 가장 많이 또는 가장 적게 발생하는 데이터 유형, 스토리지 요구사항, 데이터 저장 공간의 증가추세를 분석해 대응해야 한다. 메타데이터에 데이터 접근이나 사용과 관련한 서비스 수준 계약(SLA)이 포함돼 있으면, 데이터 수명주기 관리에 도움이 될 수 있다. 이를 분석·이용해 저장 매체에 데이터를 적절하게 배치하면 스토리지의 활용도가 높아지고 스토리지 구입 비용도 절감할 수 있다.

우리나라의 「개인정보 보호법」은 개인정보를 수명주기, 즉 '수집 → 저장·관리 → 이용·제공 → 파기'에 따라 처리하도록 관련 규정을 두고 있다. 법 제21조(파기) 제1항에는 수집한 개인정보가 처리 목적 달성 등의 이유로 필요 없게 됐을 때 지체 없이 파기하도록 규정하고 있고, 제2항과 제3항에는 그 절차와 방법을 정하고 있다.

데이터 수명주기 관련 솔루션을 이용하면 저장장치의 효율적인 운영이 가능해 비용이 절감되는 효과가 있다. 뿐만 아니라 개인정보의 보호와 관련 규정준수를 위해서도 필요하다. 관련 솔루션으로 국내 제품은 MagicArchiver(에이원커뮤니케이션즈코리아), SQLCanvas ILM(알투웨어) 등이 있고, 외국 제품으로는 Data Lifecycle Management(Spirion), Zaloni Data Management(Zaloni) 등이 있다.

바. 데이터 통합

데이터 통합(Data Integration)은 여러 정보시스템에서 서로 독립된 데이터를 갖고 운영할 경우 데이터 간의 불일치가 발생, 데이터 품질에 악영향을 미치는 것을 막는다. 이를 위해서는 여러 개의 데이터를 통합해 단일의 데이터 뷰를 구축할 필요가 있다. 이때 △통합 대상 정의 △데이터별 추출·가공·적재에 적용할 방법 △도구 ·주기 정의 등을 위한 솔루션이 필요하다.

데이터 통합 관련 솔루션으로 국내 제품은 InnoQuartz(이노트리), TeraStream(데이터스트림즈) 등이 있고, 외국 제품으로는 DataStage(IBM), PowerCenter(인포매티카) 등이 있다.

3. 향후 전망

앞서 4차산업혁명시대를 맞이해 다양한 형태를 가진 대용량 데이터의 가치를 높이기 위한 여러 솔루션들을 알아봤다. 데이터 관리는 △비즈니스에 필요한 정보의 수집 △저장·관리 △이용·제공 △파기하는 수명주기에 따라 진행된다. 이런 과정에서 수집한 정보보호를 위해 다양한 정보보호 관련 기술을 도입해야 한다. 이는 '저장·관리' 단계에 대부분 적용하고, 일부는 파기 단계에 적용하고 있다. 여기에 개인정보 보호도 정보 보호의 영역에 포함·관리하고 있다. 개인정보는 일반 데이터와 수명주기가 유사하다. 국내의 경우는 「개인정보 보호법」, 「정보통신망 이용촉진 및 정보보호에 관한 법률」 등 관련 법률로 개인정보 보호의 철저한 이행을 요구하고 있다. 국내뿐 아니라 해외에서도 개인정보를 위한 방법이 강구되고 있는데 일례로 2018년 5월 25일부터 시행된 유럽연합의 개인정보보호법(GDPR, General Data Protection Regulation)을 들 수 있다. 이 법 역시 우리나라와 유사한 방법과 절차를 규정하고 있다.

4차산업의 특징인 초연결성을 고려할 때, 수집되는 데이터는 사람 중심으로 거의 모든 프로파일링(profiling)이 가능하다. 따라서 대부분의 데이터가 개인정보 영역에 해당할 것이다. 예를 들어, 은행을 포함한 금융기관 데이터(원장정보, 거래내역정보 등)는 「신용정보의 이용 및 보호에 관한 법률」에서는 개인신용정보라고 정의하고 있지만, 엄격하게는 개인정보에 해당한다. 이런 점을 감안할 때, 향후에는 개인정보 수명주기 관리 중심으로 데이터를 관리하는 것이 효율적이며, 안전성을 확보하는 방안이 될 것이다. 데이터 관리 솔루션도 데이터 거버넌스에서 개인데이터 거버넌스로 기능이 확대될 것으로 보인다.

제4장

데이터 수집 솔루션

데이터 수집 솔루션은 빅데이터 시장과 함께 성장하고 있다. 데이터를 자체 생산하는 기존의 DB 시스템과 달리 빅데이터의 데이터는 대부분 외부에서 오기 때문이다. 적합한 수집 솔루션 선택을 위해서는 다양한 환경에서 편리하게 적용할 수 있는 수집 인터페이스는 물론, 압축 저장·정형화·보안 등을 고려해야 한다. 하지만 현재 수집 솔루션을 바라보는 시선은 조회 성능과 기능까지 포함하고 있다. 그 이유는 수집 솔루션의 궁극적인 목적은 아이러니하게도 수집된 정보를 조회하는 것이기 때문이다.

1. 누가 어떤 데이터를 수집하나?

데이터 처리의 시작은 데이터 생성 또는 수집이다. 전통적인 데이터베이스(DB) 환경에서는 외부에서 데이터를 가져오기보다는 DB의 프론트엔드인 애플리케이션에서 데이터가 생성되면서 업무가 시작된다. 반면 빅데이터는 내부에서 데이터가 생성되기보다는 외부의 데이터를 가져오면서 업무가 시작된다. 빅데이터 환경에서 데이터 처리는 데이터 수집에서 시작한다고 할 수 있다.

데이터가 생성되는 환경에서는 데이터 생성 시점부터 완료 시점까지 보장받는 것이 중요하다. 이 때문에 트랜잭션 관리가 가장 중요한 요소가 된다. 만약 특정 사용자의 정보를 수정할 경우, 수정이 완료되기 전까지는 과거 정보로 일관되게 표현돼야 한다. 이후 이 작업이 완료된 것이 확정된 때부터 새로운 정보로 나타나야 하는 것이 원칙이다. 이런 성격 때문에 락(Lock 또는 Enqueue)과 같은 오브젝트로 트랜잭션을 보장해주는 관계형 DB가 시장을 주도할 수밖에 없었다.

외부 데이터를 시스템 내부로 가져오는 빅데이터 환경에서는 사실상 데이터의 변경보다는 수집 행위 자체가 중요하다. 수집하려는 데이터는 적어도 수집 시점에는 완결된 데이터이기 때문에 데이터가 중간에 변경될 것을 고려할 필요가 없다. 이 때문에 트랜잭션을 위한 장치들은 오히려 제약으로 작용한다. 어차피 변경할 필요가 없는 데이터라서 중복이나 유실 없이 완벽히 끌어오는 것이 더 중요하다.

게다가 데이터를 내부에서 생성하는 전통적인 DB의 경우, 데이터를 생성하는 주체가 사람인 경우가 많다. 이 경우 생성하는 데이터는 그렇게 많지 않을 것이다. 하지만 수집 대상이 되는 데이터는 기계가 만드는 데이터도 상당 부분 포함돼 있어 일단 그 양에서 차이가 크게 난다. 전통적인 DB에서 성능은 생성된 데이터 조회 시 강조되지만, 빅데이터에서는 조회뿐 아니라 데이터 생성 단계에서부터 성능이 시스템의 중요한 요소가 된다.

결국 데이터 수집 영역은 전통적인 DB 영역보다는 빅데이터에 와서 더욱 관심을 끌게 됐고, 다루는 데이터도 점차 다양해지고 있다. 수집 대상이 되는 데이터의 형태는 로컬이나 원격 파일은 물론이고 네트워크, 하둡(Hadoop), DB 등 여러 가지다. 종류로는 로그, 텍스트, 음성, 영상 등 대부분을 포함하고 있다.

특히 빅데이터 영역에서 관심이 커지고 있는 데이터가 있는데 가트너(Gartner) 등에서 다크데이터라고 지칭하는 데이터가 바로 그것이다. 다크데이터는 애플리케이션 로그, 웹 로그, 장비 로그, 메일 로그와 같이 비용을 들여 저장·수집하지만 관리하지 않는 모든 데이터를 통칭한다. 빅데이터에서 다크데이터가 주목받는 이유는 머신러닝(Machine Learning), 딥러닝(Deep Learning) 등의 기술을 접목해 특정 패턴을 탐색하는 형태로 활용할 수 있기 때문이다.

2. 수집 솔루션의 필수 기능

가. 수집 방식과 인터페이스

데이터를 시스템 안으로 수집하는 방식은 능동적인 방식과 수동적인 방식으로 나뉜다. 능동적인 방식은 파일, DB, 하둡, SNMP(Get)와 같이 데이터가 축적돼 있거나 쌓이고 있는 장치 혹은 시스템을 대상으로 한다. 수집 솔루션은 수집을 위해 폴링 방식으로 주기적으로 데이터를 가져온다.

수동적인 방식은 수집 솔루션으로 들어오는 데이터를 받아서 수집한다. 이 방식은 HTTP Post, Syslog, 패킷 캡처, SNMP 트랩수신, 넷플로우 등의 포맷을 수

• 필자 | **김한도**(이디엄 팀장)

집하는 데 사용된다. 또한 수집 솔루션의 클라이언트나 SDK 등을 통해 데이터를 밀어 넣어주는 것도 이에 해당한다.

수집 솔루션은 데이터 유실이나 중복 방지 장치가 반드시 적용돼 있어야 한다. 능동적인 방식은 유실에도 신경을 써야 하지만, 중복 수집의 가능성이 있으므로 이런 기능이 확실히 구현되는지 확인이 필요하다. 수동적인 방식은 중복보다는 유실에 초점을 맞춰야 한다. 특히 수집 담당 서버나 에이전트가 사용할 수 없어지는 경우를 대비해 HA 구성의 가능 여부와 절체(failover)가 이뤄지는 동안 데이터 유실을 방지하는 방안이 있는지를 점검해야 한다.

나. 압축 저장

수집 솔루션은 데이터를 단순히 받아들이는 것에 그치지 않고 일차적으로 저장을 한다. 수집 솔루션이 적용되는 것은 대부분 빅데이터 영역이기 때문에 대용량의 데이터를 다루게 될 가능성이 크다. 하루에 200GB의 데이터가 들어온다고 가정하자. 이를 6개월 보관하고자 한다면 35TB를 웃도는 저장소가 필요하게 돼 비용 이슈가 발생할 것이다.

이런 문제 때문에 수집 솔루션은 압축 저장이 필수요소다. 수집된 데이터를 압축해서 저장하면 디스크 양 문제도 해결되지만, 로컬 디스크의 활용도 고려할 수 있어 비용이 절감된다. 그러나 압축은 곧 성능과 직결되기 때문에 압축 효율과 성능을 동시에 따져봐야 한다.

수집 솔루션은 대부분 에이전트를 통한 수집 기능을 제공하고 있는데, 압축이 에이전트와 서버 사이에서도 가능한지 살펴볼 필요가 있다. 클라우드 환경에 수집 솔루션이 구축된 경우, 에이전트가 외부에서 수집한 데이터를 클라우드로 전송해야 하는 일이 발생한다. 네트워크에 대해 과금을 하는 클라우드 서비스는 에이전트가 서버로 전송하는 데이터의 압축 여부가 운영 비용에 큰 영향을 미칠 수도 있기 때문이다.

다. 정형화

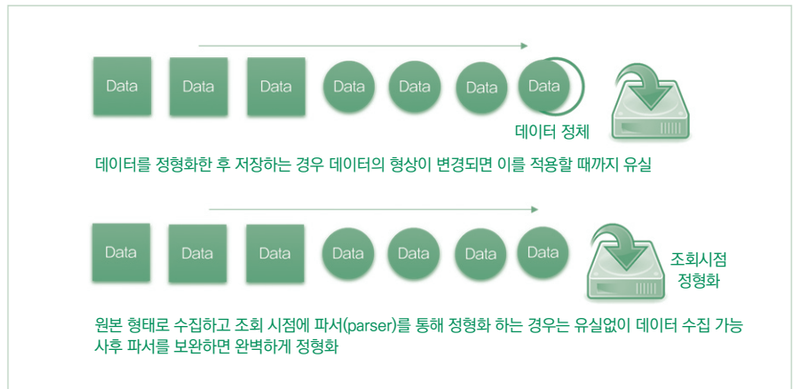
수집 솔루션은 단순히 수집 대상을 시스템 내로 옮기는 것이 아니라 데이터를 해석하고 맥락을 제공하는 역할도 수행해야 한다. 이런 역할을 정형화라고 한다. 정형화 방식은 데이터에 따라 다양하다. 텍스트 기반의 데이터 텍스트로 변경 가능한 데이터를 예로 들면, 특정 위치나 특정 부분의 데이터를 자르거나 구분해 여기에 레이블을 붙이는 방식으로 진행된다. 이 정형화의 결과물은 엑셀이나 DB의 테이블처럼 컬럼(Column)과 로우(Row)로 재구성된다.

수집 솔루션은 정형화를 수행하는 단계도 제각각 다른데, 데이터 정형화를 어디서 어떻게 진행하느냐에 따라 수집 안정성이 훼손되는 경우도 발생한다. 수집 솔루션 중에는 저장소 바로 앞에서 데이터를 정형화해 수집하는 방식도 있는데 이 경우 안정성 문제가 나타날 수 있다.

[그림 5-4-1]에서 보듯이 데이터 정형화 장치를 동그란 데이터에 맞게 만들어 놨다 갑자기 데이터가 네모 형태로 변경되는 경우가 발생하면 이 솔루션은 정형화하는 장치를 데이터에 다시 맞추기 전까지는 수집 불가능한 상태가 돼 유실이 발생한다. 특히 빅데이터는 데이터 생성 주체도 대상도 다양하기 때문에 담당자가 알지 못하는 사이에 데이터가 변경되는 일이 빈번하게 발생한다. 이런 시스템을 구축하게 되면 관리 비용이 끊임없이 증가하게 될 것이다.

반면 정형화를 사후에 하는 수집 솔루션은 이런 문제에서 벗어난다. 이런 솔루션은 데이터를 원본 그대로 저장하고 조회하는 시점에 정형화를 수행한다. 데이터 형태가 변경되더라도 이미 데이터는 수집돼 있어 정형화를 수행하는 정형화기(parser)만 사후에 수정하면 된다. 이후 다시 조회하면 모든 변경 사항이 반영돼 나타난다. 이런 이유로 유실이나 정형화에서 문제가 발생하지 않는다. 그러나 조회 시점에 정형화를 수행하려면 기본적으로 조회 성능을 보장해야 하므로 솔루션 도입 시 이 점을 따져 봐야 한다.

[그림 5-4-1] 데이터 정형화 후 저장 vs. 원본 형태로 수집·저장



라. 보안

수집 솔루션은 불가피하게 네트워크를 통해 데이터가 이동하는 과정을 포함한다. 데이터베이스와 같이 수집 대상과 수집 솔루션이 모두 보안이 철저한 네트워크 내에 있다면 저장된 데이터의 암호화, 무결성 검증과 같은 수준의 보안 장치만으로도 어느 정도 안전을 담보할 수 있다.

때로는 수집 솔루션 데이터는 공중망을 통해 전송되기도 한다. 특히 클라우드 환경으로 데이터를 수집하는 경우, 네트워크 레벨에서의 탈취 등도 고려가 필요하다. 특히 에이전트에서 서버로 전송되는 데이터는 반드시 암호화한 상태로 전송돼야 한다. 이 때문에 수집 솔루션을 도입할 때는 저장된 데이터의 암호화, 무결성 검증은 물론 종단간 암호화에 대한 부분도 관심을 가져야 한다.

마. 편의성과 유연함

위와 같은 모든 기능이 갖춰짐에도 불구하고 수집 자체가 복잡하게 진행된다면 실제 운영환경에서는 관리 비용이 지속해서 증가할 것이다. 이것은 비단 수집을 위한 인터페이스의 유무 수준이 아니고 얼마나 편하게 수집 설정이 가능한지에 대한 문제다.

적은 수의 수집 대상에서 많은 데이터가 들어오는 경우도 있지만, 비슷한 확률로 적은 양이지만 수집 대상이 많은 경우도 있다. 이 경우 수집 설정을 하나하나 프로그램 짜듯이 적용한다면 수집 자체도 물론이지만 향후 유지할 때에도 상당한 부담이 될 것이다.

세상 모든 일이 그렇지만 수집 솔루션에서 일반화할 수 없는 특이한 환경의 수집도 고려해야 하는 경우가 있다. 이 경우에도 최소한의 노력으로 수집 인터페이스를 구성할 수 있도록 수집에 관한 부분은 템플릿화 돼 있어야 한다. 하지만 이마저도 최소화할 수 있도록 수집 솔루션에서 많은 부분의 수집 인터페이스를 갖추고 있어야 한다.

수집 편의성 또한 꼼꼼하게 따져봐야 한다. 최대한 일반화할 수 있는 수집 방법은 최소한의 설정으로 정확하게 수집할 수 있는지 확인하고, 특수한 환경에서도 최소의 시간과 노력으로 유연하게 적용할 수 있는지를 반드시 점검해야 한다. 수집 대상을 적용할 때마다 프로젝트를 할 수는 없는 노릇이기 때문이다.

3. 왜 수집하나?

지금까지 누가 수집 솔루션이 필요하고, 어떤 데이터를 수집하는지, 수집 솔루션이라면 응당 가져야 할 기능은 무엇인지를 간략하게 기술했다. 마지막으로 우리가 왜 수집하는지에 대해 본질적인 이야기를 할 차례다. 사실 수집 솔루션에서 수집의 목적은 상당히 아이러니하다. 수집 솔루션의 궁극적인 목적은 수집이 아니라 조회이기 때문이다.

데이터를 수집만 하고 조회하지 않는다면 그건 수집이 아니라 데이터를 한 곳으로 모으는 행위에 불과하다. 데이터를 수집하는 이유는 다양한 시구간, 다양한 종류의 데이터를 하나의 관점으로 모아서 관찰하고 분석하려는 요구가 있기 때문이다. 관찰과 분석은 조회를 통해 이뤄지기 때문에 수집 솔루션은 조회를 전제로 데이터를 한 곳에 집적하는 일을 수행한다.

그렇기 때문에 수집 솔루션은 대량의 데이터에서 원하는 조건의 데이터를 빨리 찾는 검색과 대량의 데이터 모두를 대상으로 집계하는 성능을 보장해야 한다. 사용자들은 데이터 분석을 시작하고 얼마 되지 않아 SQL 수준 혹은 그 이상의 복잡한 질의를 수행하고자 한다. 이 경우 일정 성능과 기능을 충족하지 못하면 수집 솔루션 자체에도 의문을 갖게 될 수밖에 없다.

이미 현재의 수집 솔루션은 고객의 이런 경험치 속에서 발전하고 도태되고 있는 실정이다. 수집 솔루션이 수집만 잘하면 된다는 생각으로 도입했다가 낭패를 봤다는 사례도 심심치 않게 들린다. 또 수집 솔루션을 도입했는데 검색, 분석까지 일사천리로 돼 흡족해 하는 시장의 평가도 다수 존재한다.

수집 솔루션을 도입할 때 수집의 각 기능을 따져 보는 것도 중요하지만, 수집의 목적은 조회라는 점을 염두에 두고 시스템을 선택하거나 구성해야 할 것이다.

제5장

데이터 유통 솔루션

데이터 유통 솔루션은 전통적인 방식인 EDI/ebXML, MCI/FEP, 메일과 메신저 등이 있으며, 최근 주목받는 오픈API와 LOD 솔루션이 있다. 오픈API와 LOD는 쉽게 데이터를 유통할 수 있고, 신속한 서비스 개발과 새로운 서비스 창출이 가능하다는 점에서 주목받고 있다. 이 글에서는 오픈API와 LOD의 데이터 유통, 솔루션의 구성 요소, 적용 사례 등에 대해 자세히 살펴보고, 기존 솔루션도 간략히 소개하고자 한다.

1. 데이터 유통 솔루션의 종류

넓은 의미에서 개인이나 조직 간 데이터 교환용 솔루션은 △무역 업무와 기업 간 거래에서 오랫동안 사용해 온 EDI/ebXML 솔루션 △전통적인 대외 연계 솔루션인 MCI(Multi-Channel Integration)/FEP(Front-End Platform) 솔루션 △메시징 관점의 메일과 메신저 솔루션 △오픈 이노베이션 관점에서 제공되는 오픈API, LOD(Linked Open Data) 솔루션 등이 있다.

2. 오픈API 솔루션

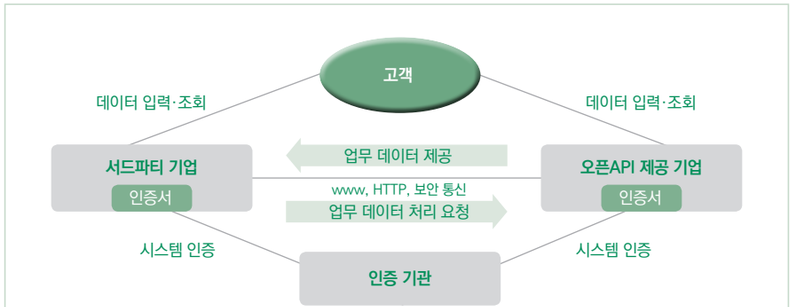
가. 오픈API 개요

API(Application Programming Interface)는 응용프로그램이 기능을 제어하거나 데이터를 처리하도록 제공하는 인터페이스 규약이다. API의 최종 목적은 프로그램 모듈 간 상호 작용을 일으켜 이미 개발된 기능의 재사용과 데이터 유통을 활성화하는 것이다. 오픈API는 이런 작업을 기업 외부로 확대할 수 있어 기업 간 API 사용을 가능하게 한다. 이를 통해 쉽게 데이터 유통을 할 수 있으며, 신속한 서비스 개발과 새로운 서비스 창출이 가능하다.

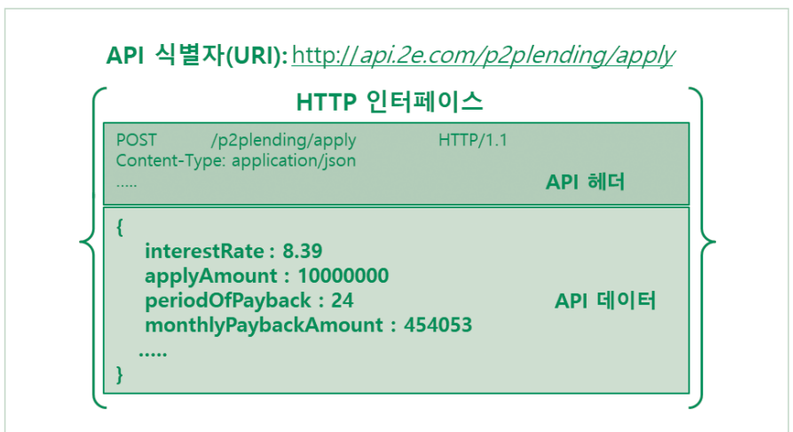
나. 오픈API에서의 데이터 유통

오픈API 환경에서는 [그림 5-5-1]과 같이 서드파티 기업은 고객으로부터 입력 받은 데이터 처리를 위해 오픈API 제공 기업에게 데이터를 보내고, 업무 데이터 처리를 요청할 수 있다. 반대로 오픈API 제공 기업으로부터 데이터를 받아 서드파티 기업의 서비스를 제공할 수도 있다. 이때 기업 외부로 업무 데이터가 유통되므로, 무엇보다 보안 조치가 중요하다. 거래 당사자 시스템 상호 간에 서로 유효한지를 인증기관을 통해 확인하고, 데이터 위변조 방지와 민감 정보를 암호화해 유통하는 것이 필요하다.

[그림 5-5-1] 오픈API 데이터 유통 흐름



[그림 5-5-2] 오픈API 데이터 예시



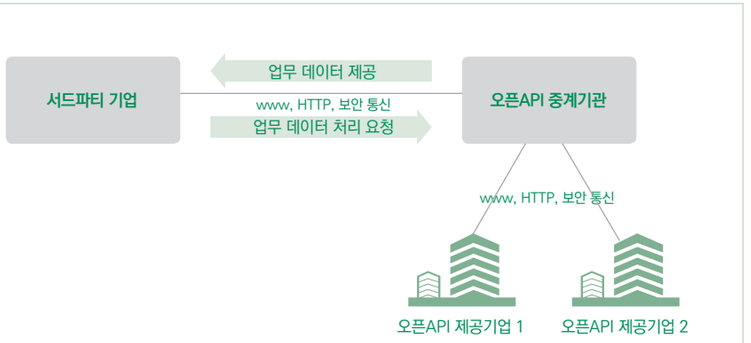
• 필자 | **안정필**(투이컨설팅 이사)

오픈API 솔루션은 기업 내외부 간 가교 역할을 하는 경우가 많은데, 그 이유는 신속한 대응이 장점이기 때문이다. 자체적으로 서비스를 구축하면, 시장 요구에 신속하게 대응이 어려운 경우가 많다. 오픈API 솔루션은 기존의 내부 시스템에서 제공하는 서비스 데이터를 조합해 새로운 서비스 데이터를 제공하거나, 기존 데이터를 외부 유통 표준에 맞게 변환하는 역할을 하는 경우가 많다.

따라서 REST/JSON뿐 아니라 레거시 시스템을 지원하기 위한 다양한 데이터 커넥터를 제공해야 하며, 내부 시스템과 연계 데이터 매핑을 관리할 수 있어야 한다.

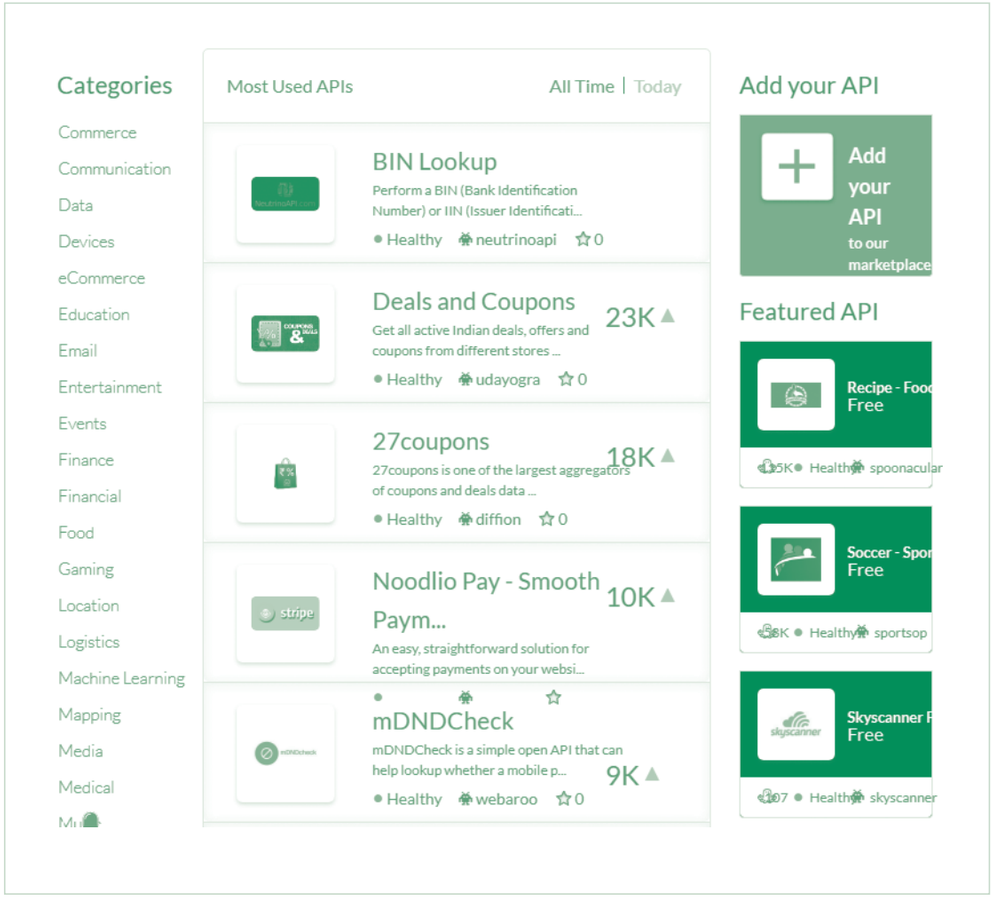
오픈API는 [그림 5-5-1]과 같이 서드파티와 직접 데이터를 유통하기도 하고 [그림 5-5-3]과 같이 중계기관을 통해 제공하기도 한다. 예를 들어 국내 금융권에서는 금융결제원의 은행권 공동 오픈 플랫폼이나 코스콤의 자본시장 공동 핀테크 오픈 플랫폼을 통해 금융 API 데이터를 유통할 수 있다. 또한 API 유통 데이터 정보를 자체 포털 구축 방식으로 알리기도 하고, [그림 5-5-4]와 같이 유명 API 마켓플레이스나 API 포털 등에 정보를 공유해 알리기도 한다.

[그림 5-5-3] 중계기관을 이용한 오픈API 유통



이런 형식은 기업 외부로의 데이터 유통이므로 용어, 데이터 타입, 코드(데이터) 등의 데이터 표준화 측면에서 더 명확한 관리가 필요하다. 또한 장기적으로 업종 내에서의 업무 개념과 개념의 구성 항목에 대한 표준화를 통해 더 효율적으로 고품질 데이터를 유통할 수 있다.

[그림 5-5-4] 오픈API 마켓플레이스 예



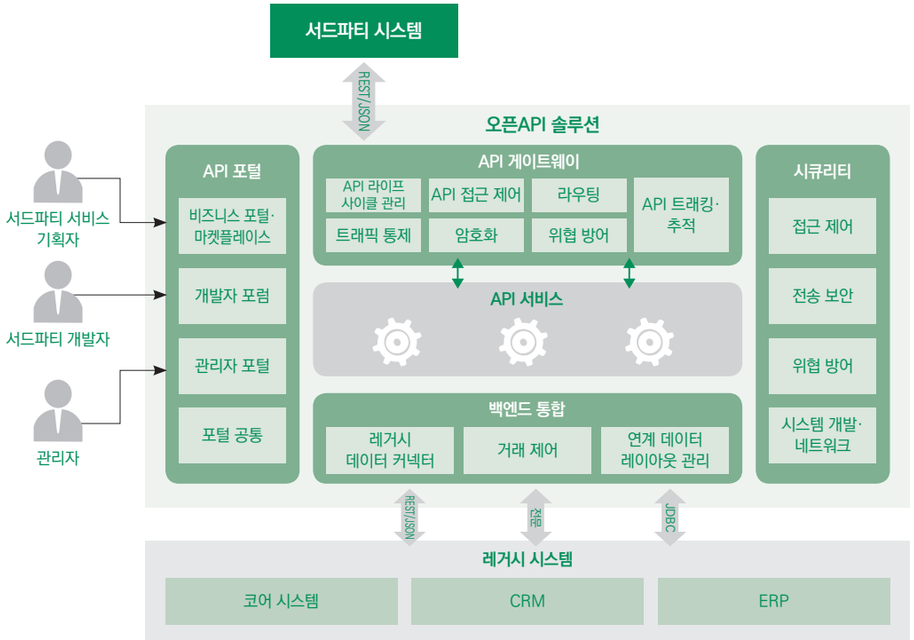
다. 오픈API 솔루션 주요 구성

오픈API 솔루션은 [그림 5-5-5]와 같이 △API 포털 △API 게이트웨이 △API 서비스 △백엔드 통합 (Back-End Integration) △시큐리티 영역으로 구성된다.

오픈API를 적용한 주요 사례는 다음과 같다.

- 공공기관 공공 데이터 개방
- 금융기관의 금융 서비스 제공
- 주요 포털·SNS 서비스의 API 제공
- 글로벌 플랫폼·클라우드 서비스의 API 제공

[그림 5-5-5] 오픈API 솔루션 구성도



[표 5-5-1] 오픈API 솔루션 구성 요소 설명

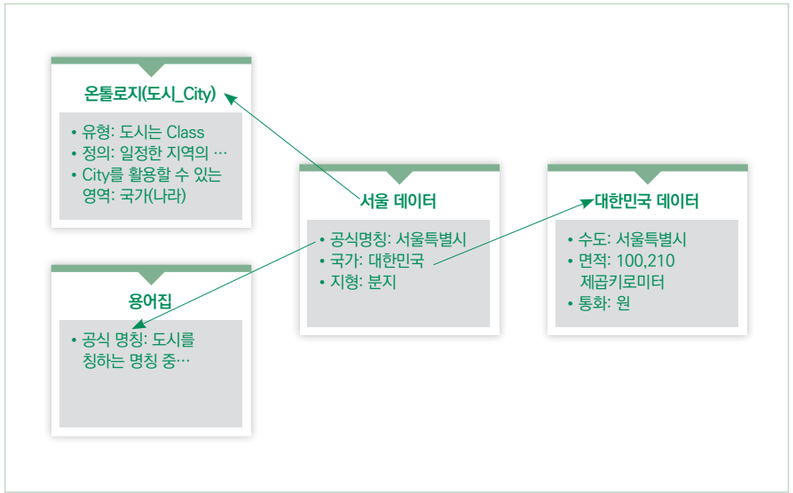
구성 요소	구성 요소 설명
API 포털 (별도 솔루션 분리 가능)	• 서드파티 서비스 기획자용 API 서비스 카탈로그, 검색, 비즈니스 적용 사례, API 마켓플레이스 서비스 제공 • 서드파티 개발자용 API 스펙, 사용 가이드 정보 제공 • 관리자를 위해 오픈할 API 등록·관리, API 게이트웨이 관리, 업무 관리 서비스 제공
API 게이트웨이 (별도 솔루션 분리 가능)	• API를 외부에 오픈해 API 데이터를 유통하는 창구임 • API 접근 제어, 트래픽 통제, 위협 방어, 라우팅, API 트래킹·추적
API 서비스	• 실제 API를 지원하고 운영하는 기능을 제공 • 추가적으로 API 서비스 개발을 지원하는 API 빌더(Builder)와 데이터 표준을 지원하는 인프라를 제공할 수 있음
백엔드 통합 (별도 솔루션 분리 가능)	• 코어 시스템 및 단위 시스템 연계 커넥터 제공 • 레거시 연계 데이터 레이아웃 관리, 거래 제어
시큐리티	• 오픈API 솔루션 전체의 보안 서비스 제공 • 접근 제어, 전송 보안, 위협 방어, 시스템 개발 및 네트워크 관련 서비스 제공

3. LOD 솔루션

가. LOD 개요

LOD(Linked Open Data)는 웹 콘텐츠를 컴퓨터 시스템이 이해할 수 있도록 웹 표준 기술을 이용해 시맨틱(semantic) 웹을 구축하는 기술이다. LOD는 Linked Data와 Open Data의 합성어로, 제한 없이 연결된 데이터 중심의 웹을 구축해 누구나 자유롭게 데이터를 재사용하고 배포할 수 있도록 한 표준이다.

[그림 5-5-6] 데이터 중심의 웹 예시



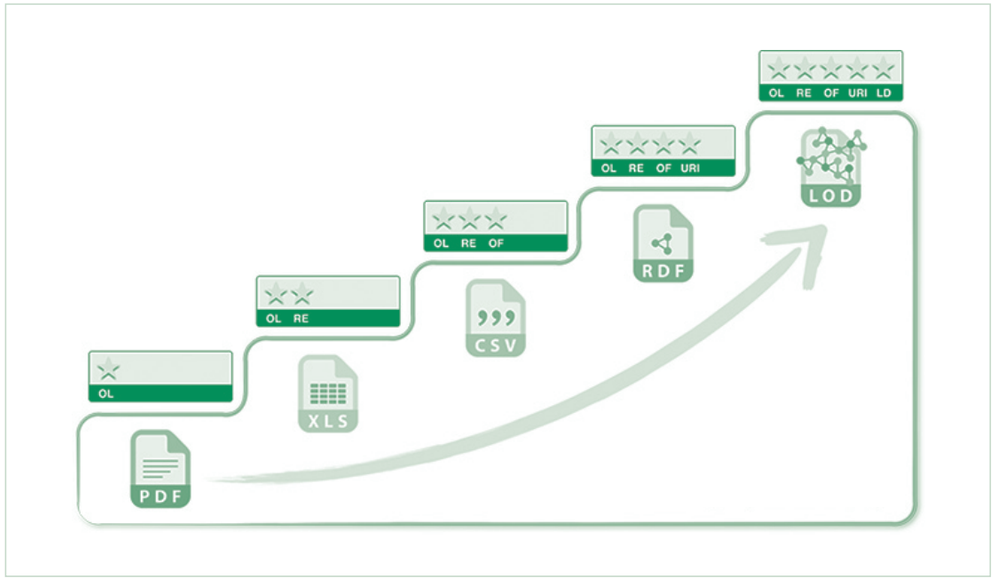
출처 : 미래창조과학부·한국정보화진흥원의 「알기쉬운 Linked Open Data」 일부 참조

[그림 5-5-6]에서 서울은 도시의 한 유형이고, 서울 데이터를 설명하고 있는 '공식명칭'에 대한 설명이 용어집에 있고, 서울 데이터가 포함하고 있는 대한민국 데이터로 링크를 이동할 수 있다.

나. 데이터 오픈 단계

웹의 창시자 팀 버너스리는 'Five Star Open Data'에서 별점을 이용해 데이터 오픈 단계를 제시했다.

[그림 5-5-7] 팀 버너스리의 'Five Star Open Data'



출처: http://www.5stardata.info(18.07.02. 접속)

[표 5-5-2] Five Star Open Data 단계

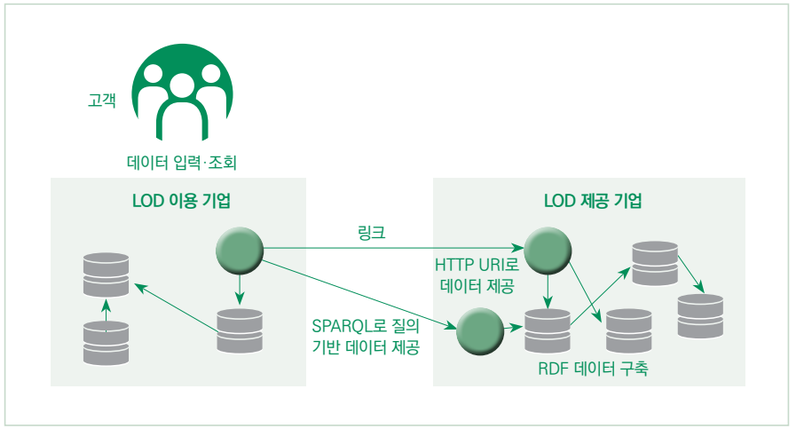
단계	단계 명	단계 설명
★	OL: On-Line	개방형 라이선스 하에 온라인에서 활용 가능한 상태
★★	OL, RE: machine REadable	구조화한 형태로 데이터 제공
★★★	OL, RE, OF: Open Format	비독점적인 개방형 포맷으로 데이터를 개방
★★★★	OL, RE, OF, URI	URI로 개체를 표시해 사람들이 데이터를 지정할 수 있게 함
★★★★★	OL, RE, OF, URI, LD(Linked Data)	데이터의 맥락을 제공하기 위해 다른 데이터와 연결된 상태

출처: http://www.5stardata.info(18.07.02. 접속)

다. LOD에서 데이터 유통

LOD는 데이터 제공 기업이 기존의 데이터베이스, 파일 등의 원천 데이터를 HTTP URI를 이용해 XML, JSON 등의 표준화된 포맷으로 공개하게 한다. LOD 이용 기업은 이것을 이용해 고객에게 서비스를 제공하거나, 또 다른 LOD 서비스를 이용해 데이터를 유통한다.

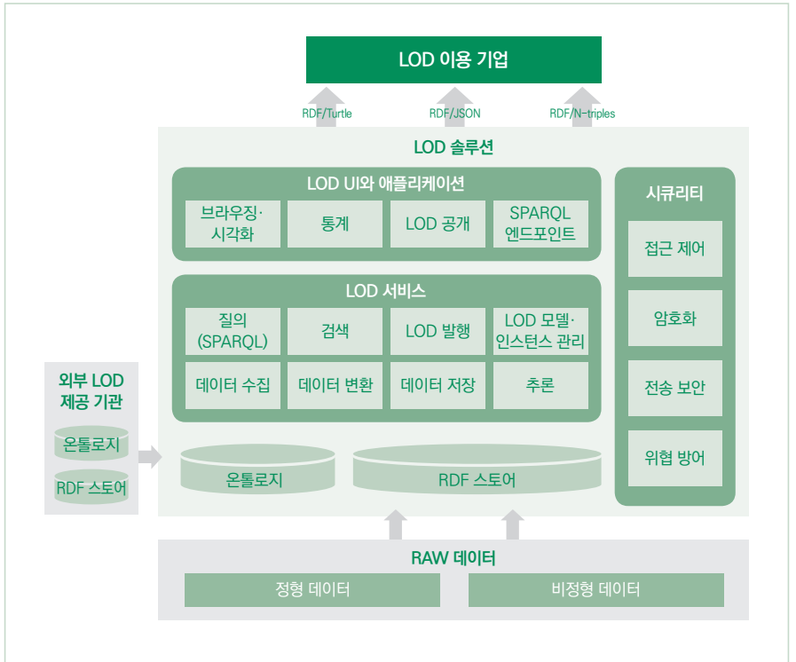
[그림 5-5-8] LOD에서 데이터 유통



라. LOD 솔루션 주요 구성 요소

LOD 솔루션은 [그림 5-5-9]와 같이 LOD UI와 애플리케이션, LOD 서비스, LOD 저장소, 시큐리티 영역으로 구성된다.

[그림 5-5-9] LOD 솔루션 구성도



[표 5-5-3] LOD 솔루션 구성 요소 설명

구성 요소	구성 요소 설명
LOD UI와 애플리케이션	LOD를 화면으로 보여주고 브라우징하는 서비스 제공 LOD 공개와 SPARQL 엔드포인트 서비스 제공
LOD 서비스	서버 사이드에서 SPARQL 질의, LOD 발행, LOD 모델·인스턴스 관리 RAW 데이터 수집과 변환, 저장, 추론 등의 서비스 제공
LOD 저장소	LOD 온톨로지와 RDF 데이터 저장소
시큐리티	LOD 솔루션 전체의 보안 서비스 제공 접근 제어, 전송 보안, 위협 방어

LOD를 적용한 주요 사례는 다음과 같다.

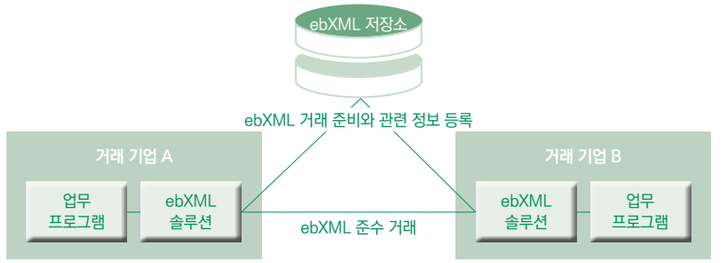
- 국립수목원 국가생물종정보
- 국립중앙도서관 국가서지
- 국토지리정보원 공간정보
- 부산시 부산 영화·영상·관광정보
- 서울시 열린데이터 광장 문화관광정보
- 한국과학기술정보연구원 과학기술정보
- 한국관광공사 스마트관광정보
- 한국식품안전관리인증원 안전먹거리정보

4. EDI/ebXML 솔루션

EDI/ebXML은 무역 업무와 기업 간 거래 시 전자적으로 데이터나 문서를 교환하기 위해 마련된 국제 포맷이다. EDI는 초기에 전용망에서 주로 거래가 이뤄졌다. 2000년대 초반부터 XML/웹 서비스 기반의 ebXML(electronic business XML) 표준이 제정돼 더 유연한 오픈 환경에서 거래할 수 있게 됐다. ebXML 스펙은 업무 프로세스, 공통 컴포넌트, 거래 파트너, 메시징 서비스로 구성해 매우 상세히 정의됐다. 거래 데이터 암호화와 위변조 방지 등 보안 요구사항도 높다. 스펙의 완성도가 높으나, 반대로 구현이 복잡해 저변 확대에 어려움이 있었다. [그림 5-5-10]과 같

이 거래 기업은 ebXML 저장소에서 거래에 필요한 정보를 받거나 등록해 준비를 마친 후, 실제 ebXML 거래를 수행할 수 있다.

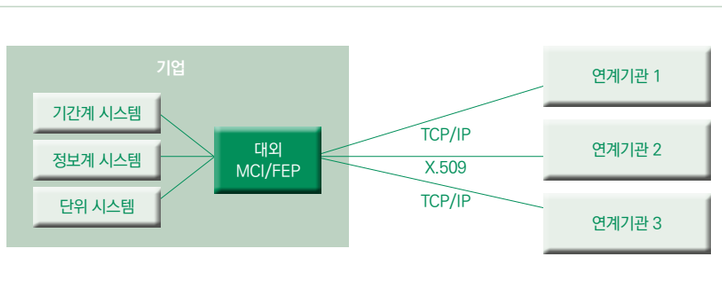
[그림 5-5-10] 전형적인 ebXML 거래 구성도



가. MCI와 FEP 솔루션

MCI(Multi-Channel Integration)와 FEP(Front-End Platform) 솔루션은 [그림 5-5-11]과 같이 기업 내부 시스템과 외부 기관의 백엔드 연계를 한 곳으로 통합하는 방식이다. 이를 통해 관리를 일원화하고 대외 환경 변화로부터 내부 시스템을 분리하려는 목적이 강했다. 그러나 MCI/FEP는 창구 시스템의 특성으로 인해, 기업 내부 데이터를 외부에 오픈하고 외부 데이터를 기업 내부로 가져오는 대외 데이터 유통의 기간 시스템 역할을 하고 있다. MCI/FEP 솔루션은 데이터 레이아웃 관리와 매핑, 통신 프로토콜 변환, 라우팅 등의 중계 기능을 지원한다. 다양한 통신 프로토콜을 지원하지만 주로 폐쇄망에서 대량 전문 데이터(전통적인 통신 프로토콜 기반)와 파일 데이터를 유통하는 작업을 수행했다.

[그림 5-5-11] MCI/FEP 구성도



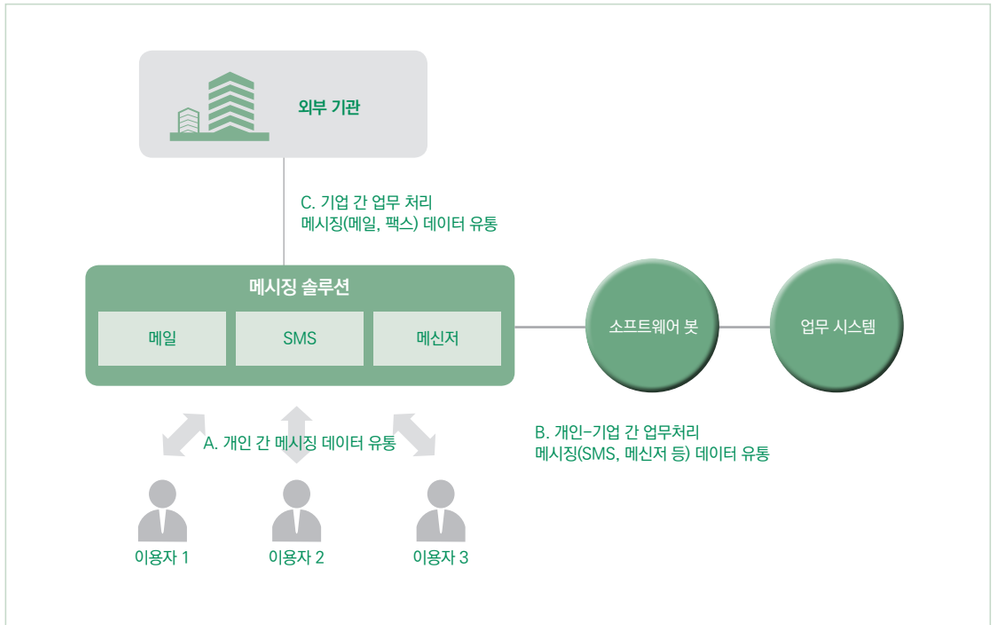
나. 메시징 솔루션

메시징의 유형은 메일, 팩스, SMS/MMS, 푸시 메시지, 메신저 등을 들 수 있다. 메시징 데이터 유통 형태는 개인 대 개인, 기업 대 개인, 기업 대 기업으로 구분할 수 있다.

[그림 5-5-12]를 보면 A와 같이 개인 간 메시징 데이터 유통을 단순히 중계하는 경우와 B와 같이 개인이 메시징-봇 솔루션을 통해 기업에 업무 처리를 요청하는 형태가 있다.

또 C와 같이 외부 기업과의 공식적인 업무 처리를 위해 메일, 팩스 등의 메시징 데이터 유통을 처리하는 형태도 있다.

[그림 5-5-12] 메시징 데이터의 유통 형태 예시



다. 데이터 유통 솔루션 특징 종합

마지막으로 데이터 유통 솔루션별 특징을 정리하면 [표 5-5-4]와 같다.

[표 5-5-4] 데이터 유통 솔루션별 특징

솔루션 명	솔루션 특징	보안 요구	주요 유통 기술	유통형태		유통 대상	
				데이터 제공	데이터 처리	개인	기업
오픈API	• 기업의 데이터 제공과 처리 기반의 유통 • API 식별자별로 한정된 데이터 유통	상	웹	O	O	O	O
LOD	• 구조화한 데이터의 연결을 통한 데이터 유통 • 주소, 위치 등의 기초 데이터 공유에 유용 • SPARQL과 같은 데이터 질의 엔드 포인트를 통해 다양한 데이터 조회 가능	중	웹	O		O	O
EDI/ebXML	• 정형화한 전자상거래 데이터 유통 • 시스템 구현 복잡도가 높음	상	전문, 웹	O	O		O
MCI/FEP	• 시스템 구조적인 목적이 강함 • 폐쇄망에서 전통적인 백엔드 데이터 유통	중상	전문	O	O		O
메시징 솔루션	• 개인과 기업 간의 정형화된 문서 또는 데이터 송수신	중	전문	O	O	O	O

이상으로 오픈API와 LOD를 중심으로 데이터 유통 솔루션과 구성 요소, 적용 사례 등에 대해 알아봤다. 오픈API와 LOD는 쉽게 데이터를 유통할 수 있으며 신속한 서비스 개발과 새로운 서비스 창출이 가능하다는 점에서 데이터 거래에서 그 역할을 확대할 것으로 전망된다.

제6장

데이터 분석 솔루션

데이터 분석 솔루션 시장이 격변기를 맞이하고 있다. 기존의 데이터 분석 솔루션은 머신러닝이나 인공지능 기술을 이용한 방식에 맞게 변모하고 있으며, 새로운 데이터 분석 솔루션도 속속 출시되고 있다. 데이터 분석 시장은 전통적인 비즈니스 인텔리전스(BI)와 머신러닝 기반 분석으로 양분돼 역동적으로 성장할 것이다. 두 솔루션의 적용 범위는 계속 확장되고 있고, 이용 방식도 다양해지고 있다. 이 글에서는 데이터 분석 솔루션 시장에서 일어나고 있는 변화와 그에 발맞춰 출시된 솔루션들을 소개하고 향후 발전 방향을 전망해 본다.

1. 개요

최근 데이터 분석이라 하면 머신러닝(Machine Learning)이나 인공지능(Artificial Intelligence) 기술을 이용한 예측 분석을 먼저 떠올릴 것이다. 이에 반해 비즈니스 인텔리전스(BI, Business Intelligence)라는 전통적인 데이터 분석 솔루션은 데이터베이스(DB)에 질의해 결과를 얻는 쿼리(query) 분석을 근간으로 하고 있었다. 이와 관련해 2017년 10월 존 드고즈(John De Goes)가 쓴 「인포월드(InfoWorld)」 칼럼에는 다음과 같은 내용이 나온다.

"BI는 죽어가고 있다. 아니, 좀 더 정확히 말하자면 다시 태어나고 있다. 창립 20주년이 넘은 기업 '태블로(Tableau)'는 마지막 'BI' 업체였다. 솔직히 말해 태블로는 주력 BI 솔루션도 아니다. 원래는 데이터 시각화 툴이었던 것이 충분한 BI 요소를 갖추게 됨에 따라 당시 업계를 호령하던 골리앗과 맞설 수 있게 된 것뿐이다."

기존의 데이터 분석 솔루션이 주춤하는 사이에 이미 사용자의 인식과 시장 동향은 바뀌었다. 변화에 맞춘 새로운 데이터 분석 솔루션들이 나오고 있으며, 기존 데이터 분석 솔루션도 변화하고 있다. 현재는 데이터 분석 솔루션의 격변기라고 할 수 있다. 데이터 분석 기술을 바라보는 사용자의 인식과 기대수준은 높지만, 데이터 분석을 도와주는 데이터 분석 솔루션은 아직 완전히 정비되지 않은 상황이다.

• 필자 | **김종현**(위세아이텍 대표)

2. 시장 동향

시장 변화의 주요 동인은 머신러닝(딥러닝 포함)이다. 쿼리를 이용한 탐색적 데이터 분석과 시각화 요소를 중시한 분석에서 머신러닝 모형과 알고리즘으로 분석하는 형태로 시장이 바뀌었다. 전통적인 데이터 분석 솔루션과 향후의 데이터 분석 솔루션을 비교하면 [표 5-6-1]과 같다.

[표 5-6-1] 전통적인 데이터 분석 솔루션과 향후 데이터 분석 솔루션 비교

구분	전통적인 데이터 분석 솔루션	향후 데이터 분석 솔루션
근본적인 특성	• DB에 질의해 결과를 얻는 쿼리 분석	• 머신러닝 기술을 이용해 모형과 알고리즘으로 분석
솔루션 분화	• 쿼리 분석을 지원하거나(전문가용 솔루션) 마우스 클릭만으로 쿼리 작업을 대체(일반 사용자용 솔루션)	• 머신러닝 오픈소스를 더욱 쉽게 사용할 수 있도록 지원하는 솔루션과 응용 분야에 특화한 솔루션
솔루션 발전	• 일반 사용자 관점이 점점 강화되고 있으며 시각화 요소가 같이 중시됨 • 사용자 편의성이 극단적으로 강조돼 분석 자체보다는 시각화, 대시보드 중심으로 분석 결과를 제공	• 데이터 수집, 관리, 생성 영역까지 분석으로 포괄함 • 분석 범위, 이용 방식의 변화에 따라 솔루션 자체의 큰 변화와 발전이 계속될 것으로 예상됨

근본적인 특성이 쿼리에서 머신러닝으로 바뀌었고, 이에 맞춰 솔루션들이 기능을 갖춰 나가고 있다. 자체 알고리즘을 개발·제공하는 방식은 거의 없고, R·파이썬·텐서플로(TensorFlow)와 같은 오픈소스 프레임워크를 활용하거나, 이를 지원하는 솔루션이 출시되고 있다. 또한 이상 탐지, 수요예측과 같은 특정 분야의 응용 솔루션으로도 분화되고 있다.

이처럼 솔루션은 발전하고 있지만 현장에서는 아직 솔루션을 이용하기보다는 전문 지식이 있는 분석가가 다양한 머신러닝 오픈소스 SW를 이용해 직접 프로그래밍하는 경우가 많다. 오픈소스 지원 솔루션을 사용하는 경우라도 일반 사용자가 아닌, 머신러닝 지식이 있는 분석가를 지원하는 수준이다. 특화한 응용 솔루션은 현업 담당자가 사용할 수 있도록 사용자 인터페이스를 단순화하고 있으며, 일반 분석 솔루션도 사용 편의성을 개선하고 있다.

[표 5-6-2] 데이터 분석 솔루션 수요자 및 공급자 동향

구분	수요자 동향	공급자 동향
수요와 공급 규모	• 전통적 BI 수요도 꾸준히 증가하고 있지만, 향후의 데이터 분석은 더 많은 잠재수요가 기대됨 • 활용 방법과 목적을 분명히 정하지 못해, 구체화한 수요는 적고 파일럿 중심으로 진행	• 전통적인 솔루션 업체 외에 많은 스타트업들이 솔루션을 개발·출시하는 추세
솔루션 사용자	• 머신러닝에 대한 지식 장벽으로 기존 분석 관련 사용자의 전환이 더딤 • 신규 채용으로 데이터 분석가를 확보하고 있음	• 전통적인 BI는 일반 사용자를 주요 대상으로 했으나, 최근에는 머신러닝 지식을 갖춘 전문 분석가를 대상으로 함 • 특화한 응용 솔루션은 현업 담당자가 사용할 수 있도록 사용을 쉽게 하고 있음
오픈소스 활용	• 다양한 오픈소스를 활용하고자 하는 요구가 많음 • 완성된 솔루션을 이용하기도 하지만 오픈소스 프레임워크를 이용해 직접 프로그래밍하고 분석하는 경우도 많음	• 오픈소스는 솔루션 내에 포함해 발전해감 • 오픈소스로 분석 모형의 다양화와 개방성을 높이고 있음

오픈소스는 데이터 분석에서 중요한 역할을 하고 있으며, 그 중요도는 점점 더 커질 것이다. 이는 수요자가 직접 오픈소스만으로 분석하게도 하며, 수요자였던 기업을 데이터 분석 서비스 기업이 되도록 도와 주기도 한다. 글로벌 클라우드 기업의 경우, 자체 머신러닝 기술을 오래전부터 오픈소스로 제공해왔고 이를 응용 영역별로 API 서비스로 제공하고 있다. 이는 데이터 분석 솔루션의 역할을 부분적으로 한 것이지만, API 서비스를 활용한 또 다른 데이터 분석 솔루션이 등장한 것으로도 볼 수 있다.

3. 제품 동향

최근 조사 기관들이 발표하는 예측 분석과 머신러닝 솔루션 부문의 대표 제품 리스트는 기관과 시기별로 그 변동이 크다. SAS, IBM, SAP, 마이크로소프트와 같은 세계적인 기업 제품과 함께 래피드마이너(RapidMiner), 나임(KNIME), 피코(FICO)와 같은 분석 전문회사 제품도 꾸준히 등장하고 있다. 그밖에 신생 기업의 제품이 급부상하기도 한다. 제품들은 모두 오픈소스 기반으로 기능하거나, 오픈소스에 대한 인터페이스를 제공한다.

국내에서도 오픈소스 기반의 새로운 분석 제품들이 출시되고 있는데 [표 5-6-3]에서 보듯이 다양한 제품이 있다. 기존 BI 업체가 시장 변화에 맞춰 신규 개발한 머신러닝 분석 제품, 스타트업 제품, 어니컴의 ankus와 같이 오픈소스 기반의 빅데이터 플랫폼으로부터 머신러닝 분석으로 확장된 제품들이 그것이다.

[표 5-6-3] 국내 주요 데이터 분석 솔루션 제품

제품	비고
LG CNS의 DAP	데이터 분석과 AI 플랫폼
비아이매트릭스의 i-STREAM	전통적인 BI 기능은 i-MATRIX와 i-CANVAS에서 제공
삼성SDS의 Brightics AI	제품보다는 플랫폼, 클라우드 기반의 서비스(Brightics Cloud)로서 제조·마케팅 분야별 응용 솔루션 제공
위세아이텍의 WISE Prophet	자동화된 머신러닝 플랫폼. WISE Intelligence와 함께 예측분석 수행
화수목의 R-Flow	누구나 사용할 수 있는 무료 버전 제공

4. 향후 전망

데이터 분석 시장은 전통적인 BI와 머신러닝 기반의 분석으로 양분화돼 역동적으로 성장할 것으로 전망된다. 두 영역 모두 적용 범위가 계속 확장되고 있고, 이용 방식도 다양해지고 있다.

전통적인 BI는 갈수록 사용자 저변이 확대됨에 따라 사용 용의성이 점점 강화되고 있다. 이에 따라 직관적인 이해가 가능한 시각화 요소가 중시되고 있다. 나아가 사용자 편의성이 극단적으로 강조돼 한 눈에 비즈니스를 파악할 수 있는 대시보드 기반의 분석 결과 제공 방식이 큰 호응을 얻을 것으로 전망된다.

향후 데이터 분석 솔루션은 머신러닝 분석이라는 격변기를 맞아 새롭게 변신하는 과정에 있으며, 아직 성숙기는 아니다. 머신러닝은 그 자체가 계속 발전해 나가고 있어 데이터 분석 솔루션도 이에 맞춰 지속적으로 변화하고 업데이트돼야 한다. 이외에도 클라우드 인프라나 데이터와 결합하기도 하고, 다른 기관이나 기업의 API 서비스와 융합이 시도되는 등 다양한 방식으로 데이터 분석 솔루션이 변화하고 발전할 것으로 전망된다.

제7장

데이터 플랫폼 솔루션

데이터 플랫폼 솔루션은 데이터를 기반으로 가치를 얻는 데 필요한 일련의 과정을 지원한다. 최근 이 솔루션들은 프로세스를 규격화한 기술 서비스를 통합하고 있다. 특히 빅데이터 기술, 데이터 거버넌스 기술, 데이터 통합 기술을 하나의 플랫폼으로 통합한 형태로 발전하고 있다. 이런 통합 플랫폼 솔루션을 자체 기술로 만드는 데는 오랜 시간과 노력이 필요하다. 때문에 대부분 외산 솔루션에 의존하고 있는 실정이다. 데이터 플랫폼 시장은 어느 때보다도 치열한 전쟁터가 되고 있다. 미래 기술인 인공지능과 디지털화에 의한 산업 구조 개편의 키는 데이터에 있으며, 이는 데이터 플랫폼에서 시작되기 때문이다.

1. 빅데이터 플랫폼 시장 '전쟁의 서막'

지난 2006년 더그 커팅(Doug Cutting)과 마이크 캐퍼렐라(Mike Cafarella)가 개발한 하둡(Hadoop)은 2011년 아파치 재단으로 넘어가 공개소프트웨어로 발표됐다. 이후 전통 DB 시장의 대형 IT 업체들은 자신들의 DBMS(DataBase Management System)와 BI(Business Intelligence) 기술을 기반으로 데이터 분석 플랫폼에 오픈소스인 하둡을 끼워 넣어 팔기 시작했다.

사실 글로벌 대형 업체들은 2011년을 전후해 MPP(Massively Parallel Processing) 기반의 데이터 어플라이언스(Appliance) 시장에서 치열하게 경쟁하던 상황이었다. 각종 인수 합병을 통해 미래 데이터 플랫폼 시장의 주도권을 잡고자 한 것이다.

[표 5-7-1] 빅데이터 시장 초기 대형 M&A 현황

글로벌 업체	피인수 기업	시기	규모	비고
Dell EMC	그린플럼(Greenplum)	2010년 7월	비공개	오프소스로 전환
IBM	네티자(Netteza)	2010년 9월	17억 달러	PDA로 제품명 변경
HP	버티카(Vertica)	2011년 2월	비공개	
TeraData	아스터데이터(Asterdata)	2011년 3월	2억 6,300만 달러	

• 필자 | **박시영**(데이터스트림즈 상무)

데이터 어플라이언스를 통한 데이터 플랫폼 솔루션은 대용량 데이터를 처리하기에는 적합하지만, 문제는 비용이었다. 이 때문에 글로벌 업체들은 하둡을 자사의 상용 빅데이터 플랫폼 솔루션의 필수 구성요소로 포함했다. 즉 미션 크리티컬한 업무 데이터는 자사의 DB 어플라이언스로, 조금 덜 중요한 데이터 처리는 하둡을 활용할 것을 권장했다.

이후 클라우드 시장이 열리면서 아마존(Amazon), 마이크로소프트, 구글이 빅데이터 플랫폼을 클라우드 환경에 접목해 또 다른 데이터 플랫폼 시장을 형성하고 있다.

2. 빅데이터 플랫폼 솔루션의 구성

하둡 기술은 뛰어난 확장성과 유연성이 강점이지만 기술 프레임워크 구조가 복잡하다는 단점도 있다. 다른 오픈소스 소프트웨어와 마찬가지로 하둡 솔루션 역시 기업 내 클라우드와 빅데이터 처리에 바로 활용하기는 어렵다. 이 때문에 업체들은 하둡 기술과 운영 노하우를 바탕으로 제품을 상용화해 소프트웨어 제품, 서비스, 컨설팅을 제공하는 형태로 데이터 플랫폼을 구성하고 있다. 최근에는 센서, 로그 데이터, 소셜 네트워크가 제공하는 스트리밍 데이터의 실시간 처리도 필요해졌다. 이에 데이터 표현, 저장, 가공, 처리 기술이 실시간성 인메모리 데이터 그리드(IMDG, In-Memory Data Grid) 기반의 데이터 플랫폼 기술로 발전되고 있다. 이와 함께 빅데이터 플랫폼에서 최근 중요한 기술적 쟁점으로 표준화, 개인정보 비식별화와 보안(security), 개인정보 보호(privacy)를 들 수 있다. 정부도 빅데이터 활용 가이드라인(개인정보 비식별 조치 가이드라인)을 발표하는 등 이에 대해 적극적으로 대응하고 있다.

데이터 플랫폼 시장은 전통적인 DB 중심과 하둡 기반으로 이원화돼 있었는데, 이것이 최근 데이터 허브(Data Hub) 아키텍처로 통합되고 있다. 전통적인 DB 중심은 기간계, 정보계, 분석계 영역으로 구분된 정보시스템 간의 데이터 추출, 변환, 적재하는 기술이 중요했는데 최근 이것이 빅데이터 영역에서의 데이터 적재 기술과 통합되고 있다.

즉 빅데이터와 DW(Data Warehouse)가 따로 구축·활용될 경우 비즈니스 가치

를 높이는 데 한계가 존재한다는 것을 고객도 알고 있다. 하지만 현재 고객이 처한 레거시 환경을 모두 빅데이터로 전환해 서비스하기에는 다소 부담이 있다. 기존 DB 영역과 하둡 영역에서의 빅데이터들에 대한 통합 저장과 분석이 요구되기 때문이다. 또한 기존 DB 중심의 리포팅, OLAP, BI 분석과 하둡 기반의 빅데이터 분석(특히 고급 분석으로 이원화된 것이 하나의 분석 프레임워크)으로 통합되고, 데이터 분석도 대용량의 실시간으로 진화하고 있다.

이런 데이터 허브(Data Hub)는 과거에는 데이터 창고(Enterprise Data Warehouse)라고 불렀다. 빅데이터의 등장 이후에는 데이터 레이크(Data Lake)라는 이름으로 대체되고 있으며, 이와 관련된 많은 논문과 방법론이 발표되고 있다.

[표 5-7-2] 데이터 웨어하우스와 데이터 레이크

데이터 웨어하우스	Vs.	데이터 레이크
정형화된, 변경되거나 처리된 Structured(data schema model), processed(transformed)	**데이터 (Data)**	구조화된·반정형화된·비정형화된, 처리되지 않은 원본의 (Structured / semi-structured / unstructured, raw)
저장 시 이미 정의된 스키마에 따라 (Schema-on-write)	**프로세싱 (Processing)**	저장 후에 정의되는 스키마에 따라 (Schema-on-read)
대용량 데이터 저장에 고비용 (Expensive for large data volumes)	**스토리지 (Storage)**	저비용 스토리지를 사용할 수 있도록 설계됨 (Designed for low-cost storage)
덜 민첩하고 구성 재구성 어려움 (Less agile, fixed configuration)	**민첩성 (Agility)**	아주 민첩하게 필요할 때 재구성 가능 (Highly agile, configure and reconfigure as needed)
분석가, LOB(Line of Business) 사용자	**대상 사용자 (User)**	데이터 사이언티스트, 분석가, LOB(Line of Business) 사용자

3. 빅데이터 플랫폼 기술

빅데이터용 데이터 플랫폼 기술은 규모, 다양성, 속도, 가치(value) 측면에서 전통적인 기술과 방법으로는 다루기 어려운 데이터를 효과적으로 처리한다. 비용대비 효율적으로 데이터를 처리, 분석, 표현하므로 사용자들에게는 새로운 통찰력(insight)이 제공된다. 이 때문에 비즈니스 가치를 높일 수 있다.

4. 빅데이터 플랫폼 솔루션

앞서 소개했듯이 오픈소스 기반의 하둡 솔루션은 기업 내 클라우드와 빅데이터 처리에 바로 활용하기가 어렵다. 이 점을 고려해 데이터 레이크 기반의 RDBMS와 하둡을 하나의 데이터 플랫폼에 통합한 SW 제품, 서비스, 컨설팅을 제공하는 업체들이 있다. 빅데이터 기술 분류에 따른 국내외 주요 적용 기술이나 솔루션을 배치해 보면 [표 5-7-3]과 같다.

[표 5-7-3] 빅데이터 기술 영역별 주요 솔루션 맵

빅데이터 기술							
데이터 수집	데이터 적재	SQL NoSQL	개인정보 비식별화	실시간 데이터 분석	통계 분석	시각화	DB 어플라이언스
	Apache Hadoop ecosystem			BASS	R	D3/Visual.ly	HP Vertica
TeraStream	Cloudera		TeraTDS	CEP Esper	SAS	Octagon	Oracle Exadata
Pentaho	HortonWorks		SQLCanvas	STRIIM	SPSS	Microstategy	Dell EMC Greenplum
IBM InfoSphere DataStage	MapR		DataGenor	Tibco	Tableau		Teradata Aster
Talend	Splunk			Kopens	Spotfire		IBM PDA
Informatica	TeraONE			SAP HANA	QlikSense		
	IBM BigInsight			Altibase	BI-Matrix		
	Oracle BDA(Cloudera)			Goldilocks			
	Teradata Aster						
	Dell EMC Pivotal						

플랫폼 상용

해외 빅데이터 플랫폼 제품으로는 IBM의 BigInsight, 오라클의 Big Data Appliance, Dell EMC의 GreenPlum, 테라데이타(Teradata)의 Aster Big Analytics Appliance 등이 있다.

국내에서는 KT가 NexR를 인수해 NDAP(NexR Data Analysis Platform)를 출시했는데 NDAP는 하둡의 핵심 기능을 제공하고 있다. SK텔레콤은 내외부에서 수집된 데이터들을 중소 상공인에게 제공하기 위한 빅데이터 허브를 구축했다. 솔트

룩스는 비정형 데이터로부터 의미 있는 정보를 추출하고 결과를 처리하는 BigO 플랫폼과 인공지능 플랫폼인 ADAM을 판매한다. 알티베이스의 제품으로는 인메모리 DB와 온디스크 DB를 결합한 하이브리드 DB인 '알티베이스 HDB'가, 인메모리 그리드 DB 솔루션으로 '샤딩(Sharding)'이 있다.

빅데이터 플랫폼 기업인 그루터는 오픈소스 통합 관리와 모니터링 솔루션으로 Cloumon을 제공한다. 국내 ETL 기업인 데이터스트림즈는 하둡 기반의 TeraStream for Hadoop, 인메모리 컴퓨팅 플랫폼인 TeraStream BASS, 데이터 거버넌스 솔루션을 결합한 TeraONE을 제공한다. 이들은 다양한 서버 환경에서 데이터 거버넌스용 메타 검색 기능을 수행한다. 또한 DB 및 Hadoop 기반의 데이터 허브와의 ETL, 빅데이터 저장과 분석, 실시간 IoT 빅데이터 처리를 지원하는 통합 플랫폼도 제공하고 있다. 검색 기술 전문 업체인 와이즈넛(Wisenut)은 빅데이터 통합 검색, 텍스트 마이닝, 소셜 네트워크 분석 서비스를 제공하고 있다.

5. 데이터 플랫폼 솔루션 전망

데이터 플랫폼은 전통적인 정형 데이터 분석 시장을 차지하고 있던 DBMS, ETL, OLAP과 리포팅, 마이닝 등의 솔루션 업체로부터 비교적 쉽게 기술 이전이나 운영 지원을 받을 수 있었다. 물론 제품 공급도 가능했다. 그러나 오픈소스 소프트웨어로 된 빅데이터 플랫폼은 비용 절감이라는 매력적인 요소가 있음에도 한계가 있었다. 그것은 바로 실제 선택에서 적용, 최적화, 운영에 이르기까지 모든 과정을 사용자가 스스로 수행하고 책임져야 한다는 점이다. 이 때문에 전문 인력에 대한 투자가 필수였고 빅데이터 전문 인력을 내부적으로 확보하기가 쉽지 않았다. 기술적인 한계도 있었는데 오픈소스의 비정형 데이터에 대한 수집 전처리 기술로는 SNS나 웹 크롤러 기술, 텍스트 분석 솔루션 등을 해결하기 어려웠다.

빅데이터를 시각화하고 분석하는 환경 구성도 쉽지 않았다. 과거 DW 시절의 OLAP처럼 정형/비정형 분석 화면을 개발해주는 것이 어려워졌다. 인력 면에서 보자면 데이터 분석가의 역할이 점차 중요해졌지만, 그들은 OLAP 전문가가 아니라는 한계도 있다. 따라서 과거 SI(System Integration) 시절에는 개발자 중심인 MicroStrategy, BusinessObject, Cognos 제품이 주목받았다면, 최근에는 일반 사용자도 쉽게 사용할 수 있는 Spotfire, QlikView, Tableau 등의 인기도 올라가고 있다. 이처럼 주목받는 제품들은 주로 EUC(End User Computing) OLAP 솔루션과 시각화 솔루션 등이다. 그 이유는 IT 기술전문가가 아닌 일반 사용자들이 셀프 서비스(Self Service) 분석이 가능한 환경을 찾고 있기 때문일 것이다.

오픈소스 R이 과거 SAS, SPSS 등의 상용 분석 툴을 이용한 분석 환경을 대체해 가고 있는 것도 동일한 이유일 것이다. 많은 오픈소스 분석 사용자 그룹이 활성화됨에 따라 재활용 가능한 라이브러리가 공개적으로 제공되고 있다. 이 라이브러리는 누구나 쉽게 쓸 수 있다는 점에서 과거 고비용의 상용 마이닝(Mining) 도구를 대체하고 있다.

결국 사용자는 어려운 빅데이터 에코 시스템 환경을 기존 DW 환경과 함께 통합시켜 가치를 높이고자 하며, 이 때문에 통합은 더욱더 중요해지고 있다. 이제 누가 더 비용 효율적으로 쉽게 통합 환경을 제공하고, 일반 사용자의 빅데이터 분석을 쉽고 편리하게 지원할 것인가가 시장의 주요 쟁점이 됐다. 여기에 답을 줄 솔루션이 향후 데이터 플랫폼 시장을 견인하는 주인공이 될 것이다.

입맛에 맞는 솔루션 업체를 찾지 못한 사용자라면 자체적으로 빅데이터 기술을 완전 내재화하기 위해 고급 빅데이터 기술자를 대거 모집하려고 할 것이다. 혹은 상용 오픈소스 빅데이터 공급자에게 값비싼 빅데이터 구독료를 지급하고, 긴 시간을 투자해야 할지도 모른다.

이제 빅데이터는 곧 닥칠 궁극적인 인공지능 기술에 의한 삶의 변화나 디지털화에 의한 산업구조 재편의 기반으로서 중요한 역할을 할 것이다. 미래가치를 선점하기 위한 데이터 기반을 누가 더 빨리 체계적으로 정비하고 활용할 수 있는지가 기업이나 조직 경쟁력의 핵심요소가 될 것이다.

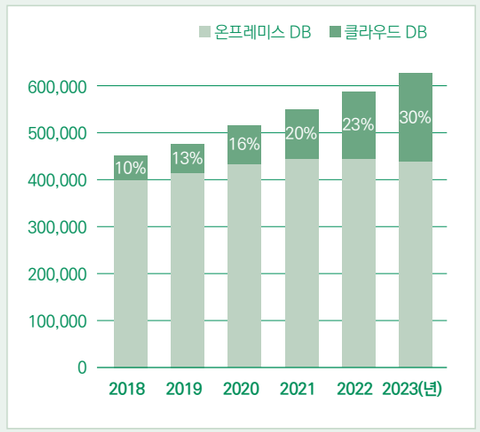
클라우드 DB의 현재와 미래

이용재 | 티맥스데이터 본부장
2017 데이터 솔루션 이노베이터상 수상자

현재의 클라우드 DB 서비스를 이용해도 엔터프라이즈 데이터 처리에 필요한 요건은 충족이 가능하다. 하지만 애플리케이션 아키텍처 재구성, 별도의 클라우드 서비스를 이용한 데이터 연계, 지나치게 복잡한 클러스터 구성, 운영 복잡성 등의 문제들이 클라우드 환경으로 이관을 하고자 하는 CIO나 현업 담당자들에게 큰 고민거리가 되고 있다.

데이터베이스 관리시스템(DBMS)은 다양한 데이터를 저장·처리·분석하는 소프트웨어다. 공공·금융·산업 등 모든 엔터프라이즈 환경에서 IT 서비스의 중심에 자리 잡아 핵심 경쟁력을 제공하는 역할을 수행하고 있다.

[그림] 전 세계 DBMS 시장 규모 (단위: 억 원)



출처: Gartner Forecast, Enterprise Software Markets & Cloud, Worldwide, 2014~2021, 4Q17 업데이트

DB 운영 측면에서 보았을 때 기업들은 기업 내에 서버를 직접 구축하거나, 외부 데이터 센터를 임대한 후에 자체적으로 데이터 처리·서비스를 운영하는 경우(온프레미스, On-Premise)가 대부분이다.

그러나 요즘은 아마존 클라우드, 마이크로소프트 애저(Azure)로 대표되는 클라우드 서비스(퍼블릭 클라우드)가 점차 확산되고 있다. 또한 오픈시프트(OpenShift)나 쿠버네티스(Kubernetes) 등을 이용해 자체적으로 클라우드를 구축하려는 다양한 시도도 늘어나고 있는 추세다. 이제 CIO나 현업 담당자들도 클라우드 환경에서 엔터프라이즈 핵심 데이터를 어떻게 처리할지를 구체적으로 고민해야 한다.

이 글에서는 다양한 클라우드 환경에서 엔터프라이즈 데이터를 저장·처리할 때 고려해야 할 요소들을 살펴보고, 클라우드 환경에서 미래 DB 기술 변화와 발전 방향에 대해 알아보고자 한다.

데이터 저장·처리 용량 확장

클라우드 환경은 여러 대의 물리 서버를 묶어 가상 환경을 제공한다. WAS(Web Application Server) 등의 미들웨어 관점에서 보기에는 확장이 용이하고, 부하 변동에 대응하기가 쉽다. 하지만 데이터 저장·처리 관점에서는 일정 부분 제약이 존재한다.

현재 대부분의 퍼블릭 클라우드에서는 읽기/쓰기 분리를 통해 용량 확장(스케일아웃)을 지원하거나, 데이터 저장 측면에 있어서 충분한 확장성을 제공하는 등 다양한 DB 서비스를 제공한다. 하지만 데이터 처리 측면에 있어서는 액티브-액티브 가용성 등에 제약을 두는 경우가 많다.

프라이빗 클라우드에서는 사용자가 직접 DB를 설치해 서비스를 구축해야 하므로, 어느 DB를 사용하느냐에 따라 데이터 처리 용량 확장·동시성·가용성 등에 제약이 있을 수 있다. 처리 용량 확장 문제는 애플리케이션 아키텍처 변경, 재개발 등을 통해 극복할 수는 있다. 하지만 이 경우 클라우드 이전을 통한 인프라 비용 절감을 넘어서는 비용이 발생할 수 있으므로 꼭 확인해야 한다.

다양한 형식의 데이터 처리

엔터프라이즈 환경에는 전통적인 DB로 처리하고 있는 구조화된 데이터 외에도 배치 처리를 위한 파일, 각종 로그 기록, IoT 기기에서 올라오는 데이터 등 다양한 데이터 타입이 있다. 이 다양한 데이터들은 온라인 처리, OLAP 분석, 빅데이터 분석 서비스 등으로 각각 처리한다. 기존 온프레미스 환경에서 다양한 데이터를 처리하기 위해서는 각각 솔루션을 도입하거나, 별도의 저장 공간으로 데이터를 복사·변환·저장해 옮기는 등 물리적인 구성을 변경해야만 가능했다.

클라우드 환경에서는 데이터 저장·처리 서버 확보 등이 논리적인 구성을 변경해 가능해지고, 비용이 상대적으로 낮은 편이므로 데이터 통합 처리·분석

부분도 활용 가능성을 적극 검토해볼 수 있다.
물론 클라우드 환경이 능사인 것은 아니므로, 도입을 검토하고 있는 클라우드 서비스 및 DB에서 다양한 데이터 타입에 대해 저장·처리를 지원하는지, 각 데이터 타입 사이에 연계 통합 분석은 가능한지, 별도 유료 서비스 혹은 연동 서비스 개발이 필요한지 등을 사전에 검토해야 한다.

데이터 처리 신기술 적용

최근 각광받고 있는 IoT·ML(머신러닝)·AI (인공지능) 등의 신기술들은 기술적 난이도, 데이터 크기, 처리 용량 등의 이유로 기존 온프레미스 환경에서 적용이 어려웠지만, 클라우드 환경에서는 문턱이 크게 낮아졌다.
기존 온프레미스 서비스 역시 클라우드 서비스로 단순 이관한다면 당장의 시급한 고민은 해결될 수도 있다. 하지만 IoT·ML·AI는 미래 기업 경쟁력을 좌우할 정도로 영향력이 크다. 아직 도입 계획이 없더라도, 향후 확장을 대비해 클라우드 데이터 처리와 연계한 아키텍처를 미리 고민하고 대비해야 한다.
IoT·ML·AI 모두 빅데이터를 저장하고 분석한다는 공통점을 갖고 있다. 데이터 저장 측면에서는 클라우드 DB에서 대량 데이터를 어디에 어떻게 저장하는지, 저장 용량 및 처리 용량 확장이 가능한지, 현재 혹은 미래에 서비스 내부에서 발생할 데이터를 기대한 시간 안에 저장할 성능은 되는지를 먼저 확인해야 한다.
학습 과정에서는 저장된 빅데이터 분석만으로 충분한지, DB에 저장된 구조화 데이터 혹은 외부 데이터 등과 연계해 학습이 가능한지 여부를 고려해야 한다. 이 부분은 앞에서 이야기한 데이터 통합 처리와 동일한 부분이다. 또한 실제 IoT·ML·AI를 이용한 서비스를 운영할 때 지속적으로 추가 또는 변경되는 데이터들이 학습 모델에 어떻게 반영되는지 여부까지 고려해야 한다.

클라우드 DB의 현재와 미래

현재 퍼블릭·프라이빗 클라우드에서 사용 가능한 DB들은 기존 온프레미스 환경에 특화된 데이터 처리 아키텍처 기반에서 발전한 것이다. 데이터 처리 용량 확장, 다양한 형식의 데이터 통합 처리, IoT·ML·AI 등의 새로운 신기술 채용에 일부 제약이 있거나, 별도의 외부 서비스를 연동해야만 가능한 구조다.
현재의 클라우드 DB 서비스를 이용해도 엔터프라이즈 데이터 처리에 필요한 요건은 충족이 가능하다. 하지만 애플리케이션 아키텍처 재구성, 별도의 클라우드 서비스를 이용한 데이터 연계, 지나치게 복잡한 클러스터 구성, 운영 복잡성 등의 문제들이 클라우드 환경으로 이관을 하고자 하는 CIO나 현업 담당자들에게 큰 고민거리가 되고 있다. 그러므로 많은 클라우드 DB들이 다음 사항들을 필수적으로 지원해 담당자들의 고민을 덜어주었으면 한다.

첫째, 한계가 없는 데이터 저장·처리 용량 확장
클라우드 환경에서는 DB 운영자들이 전체 아키텍처에 대한 고민없이 데이터 처리 용량을 자유롭게 늘리거나 줄일 수 있어야 한다. 사용자 부하가 늘어날 경우 WAS 등 미들웨어와 DB가 연동해 자동으로 미들웨어·DB 인프라가 확장되고, 애플리케이션 재작성·재구성, 부하 분산 등에 대한 관리자의 고민은 없어야 한다.

둘째, 정형·비정형 (빅)데이터 통합 저장 및 분석
기존 온프레미스와 같이 비정형 데이터 처리, OLAP·분석 쿼리, 온라인 등의 업무를 별도 저장 공간, 별도 처리 엔진으로 분류·처리하지 않고, 데이터 수집·저장·처리·분석 등 엔터프라이즈 데이터를 하나의 시스템에서 통합·분석하고 처리할 수 있어야 한다.

셋째, 통합 데이터 기반 IoT·ML·AI 지원
하나의 데이터 처리 시스템에서 자연스럽게 IoT·ML·AI 등을 지원할 수 있어야 한다. 기존 온프레미스와 같이 ML·AI 지원을 위해 데이터를 별도로 저장해 학습 엔진을 돌리지 않고 하나의 데이터 처리 시스템에서 ML·AI 서비스를 제공하고, 변경된 데이터를 반영해 재구성할 수 있어야 한다. IoT에서 발생하는 대량 데이터 저장을 위한 별도 배치·서비스 구성없이, 데이터 처리 시스템에서 전부 수용할 수 있어야 한다.

제 6 부

데이터 컨설팅 및 구축 동향

제1장

데이터 아키텍처 기반의 데이터 설계와 구축

과거에 일어난 일의 기록이 아닌 새로운 서비스를 발굴하기 위한 자원으로서 데이터가 부각되고 있다. 이 추세에 따라 기업들은 데이터의 범위를 넓히는 한편, 외부 데이터와 융합(mash-up)해 시너지를 내는 방안을 강구하고 있다. 더 나아가 기업들은 데이터 활용뿐 아니라 데이터 판매, 데이터 개방 등 데이터 플랫폼 서비스까지 염두에 두는 모습이 탐지된다.

1. 개요

4차산업혁명시대에서의 핵심 자원 또는 원유라고 하는 데이터가 천문학적으로 늘어나고 있다. 이러한 데이터를 생성, 관리, 융합해 효율적으로 잘 활용할 수 있도록 하기 위해서는 데이터를 야적장이 아닌 자동 분류 창고에 표준화해 좋은 구조와 품질로 체계적으로 관리해야 한다. 아무리 데이터가 증가하더라도 데이터 자산을 정확하게 이해하고 높은 품질의 데이터를 활용할 수 있도록 데이터 아키텍처를 수립해야 한다.

이러한 데이터 아키텍처는 기존 정보화 시스템의 AS-IS 문제점과 개선점을 도출해 전사 데이터 아키텍처 청사진을 갖고 단계적으로 추진하는 경우도 있지만, 데이터 구조 개선으로 인해 응용시스템에 미치는 영향도가 큰 경우는 차세대 시스템을 구축할 때에 추진하는 경우가 많다.

차세대 시스템을 구축할 때 많이 활용되는 데이터 아키텍처 수립은 신축 건물에 입주하는 것과 비교해 볼 수 있다. 먼저 건축 설계도에 따라 집의 건축이 완공되면 기존에 살던 집에서 새로 지은 집으로 사람들이 이사를 한다. 새로운 집에 가져온 자산을 어떻게 효율적으로 관리할 것인지에 대한 과정이 필요하다. 건물 설계는 데이터 모델 설계 컨설팅, 이사는 데이터 이행 구축, 효율적 관리는 거버넌스와 성능 최적화 컨설팅에 각각 대응시킬 수 있다. 데이터 아키텍처를 수립하는 가장 큰 목적은 체계적이고 구조적인 접근으로 활용 대상(고객)의 용도 및 목적에 맞게 구축하고 시스템화하는 것이다. 비록 이러한 과정이 다소 수고스럽게 느껴지더라도 많은 혜택을 제공하므로 데이터 아키텍처는 필수적인 절차다. 기업의 중요 자원인 데이터를 체계적으로 관리하지 않는다면 예기치 못한 손실을 입을 수도 있어 데이터 아키텍처 수립은 시행착오를 현격하게 줄일 수 있는 매우 효과적인 방법이다.

데이터 아키텍처는 시스템에서 저장하고 다루는 정보 구조를 설계한다는 점에서 과거부터 축적된 전통적인 방법론이 여전히 의미 있게 활용된다. 하지만 빅데이터 시대를 맞아 발굴되는 다양한 데이터 형태와 기하급수적으로 늘어나는 데이터를 보관하기 위한 저장소의 다양화로 시장에서 변화가 일어나고 있다.

2. 데이터 모델 설계 컨설팅

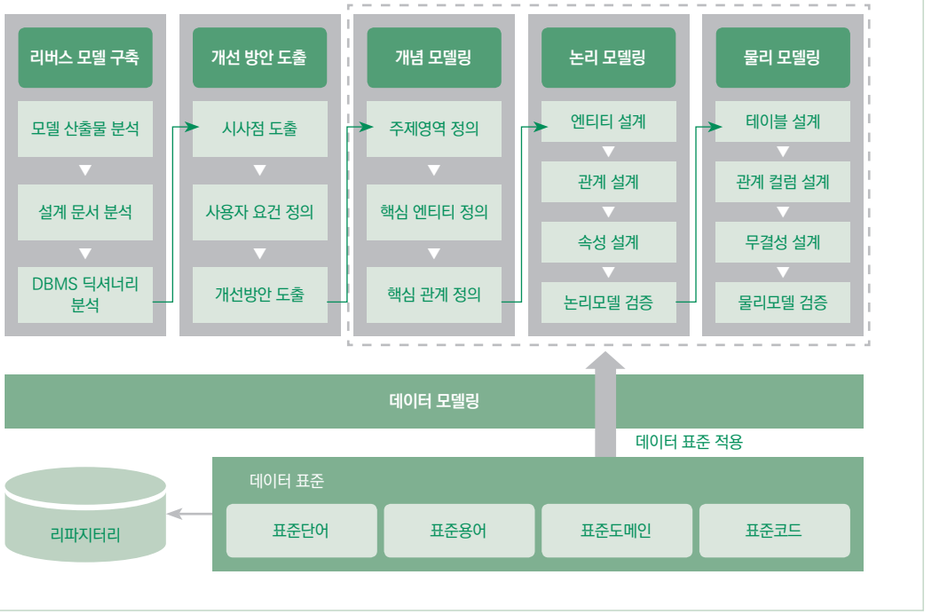
가. 데이터 모델 설계 정의

데이터 모델링은 실제 현실에서 운영되는 업무를 데이터로 형상화하는 과정이다. 데이터 모델링 과정은 일반적으로 엔티티 도출, 관계 정의, 식별자 및 속성 정의, 검증 단계로 진행된다. 모델의 구체화 정도에 따라 [그림 6-1-1]과 같이 개념 데이터 모델링, 논리 데이터 모델링, 물리 데이터 모델링으로 순차적으로 컨설팅하는 것이 일반적이다.

또한 표준화 가이드라인을 수립해 이를 준용해 비즈니스 용어를 표준화해야만 비로소 모델 설계가 완료됐다고 할 수 있다. 기업들의 중요 비즈니스인 기간계 시스템에 대한 모델링은 과거의 방식을 그대로 고수하고 있지만, 대용량 데이터를 관리할 수 있는 물리 설계의 중요성 대두와 반정형, 비정형 데이터에 대한 구조 설계 방법론이 이슈가 되고 있다. 이러한 환경 속에서 데이터 모델 설계 영역의 확대, 저장소에 따른 데이터 모델 설계 변화, 개인정보 보호를 위한 모델 설계 요구가 가속화되고 있다.

• 필자 | **서동재**(비투엔 수석컨설턴트)

[그림 6-1-1] 데이터 모델 설계 절차



나. 데이터 모델 설계 영역의 확대

정형화한 데이터 구조 설계 영역의 경우, 기업들의 거듭된 차세대 프로젝트를 통해 시장이 성숙해 기존 데이터 구조를 완전히 뒤엎는 방식의 구조 설계는 선호하지 않는 상황이다. 또한 비즈니스에 적합한 참조 모델이 직간접적으로 전파되면서 대규모 변경이 발생하는 사례는 더욱 드물어졌다. 대신 신규 비즈니스를 어떻게 하면 더 잘 수용하고 활용할 수 있는지에 대한 설계 영역이 확대되는 추세다. 전통적 영역인 비즈니스 본연의 업무 처리를 위한 데이터 구조뿐만 아니라 데이터 활용 관점에서 공개된 공공데이터를 결합할 수 있는 데이터 구조 설계, IoT에서 발생하는 센서 데이터를 관리하는 데이터 구조 설계, 개인들의 활동 로그를 시계열로 관리하고 패턴화하기에 적합한 데이터 구조 설계 등으로 영역이 점차 확대되고 있다.

다. 저장소에 따른 데이터 모델 설계 변화

기존 관계형 데이터베이스 외 NoSQL, 하둡(Hadoop) 등 빅데이터 플랫폼의 저장소가 시장에 들어오면서 관계형 데이터베이스도 여러 저장소 중 하나의 선택지가 됐다. 이제는 빅데이터 플랫폼에 익숙해진 고객들이 용도에 맞는 저장소를 찾아서 활용하기 시작했다는 평가가 많다. 바꿔 말하면 데이터 모델 설계도 저장소라는 물리적 공간의 특성을 고려해 설계해야만 한다는 뜻이다. NoSQL의 데이터 모델링은 관계형 데이터베이스 모델링과 달리 애플리케이션에 특화한 쿼리를 가진다. 따라서 비즈니스를 먼저 정의하고 거기에 맞는 쿼리 결과를 예측한 다음, 데이터 모델링을 해야 한다. 또한 하둡 파일 시스템은 병렬 및 분산처리 용도의 저장소이기 때문에 데이터를 쌓거나 읽을 때 성능적인 측면을 고려해 파티션을 설계하는 것이 중요하다. 이렇게 다변화된 저장소의 특성과 데이터 특징에 맞는 추출 기법을 고려한 설계가 필요하기 때문에 빅데이터 기술에 대한 새로운 지식이 요구되고 있다.

라. 개인정보 보호를 위한 모델 설계 요구

개인정보 유출 사고가 사회적으로 큰 이슈가 되면서 기업들은 개인정보 보호를 위한 데이터 구조 설계 요건을 강화하고 있다. 개인정보를 보호하기 위해 데이터를 암호화해 저장하는 것 이외에 데이터 관리 및 활용에 적합한 설계를 요구 받고 있다. 최근 빅데이터 시대가 오면서 기업들이 확보한 다양한 데이터를 활용하고 싶은 욕구와 개인의 프라이버시 보호가 충돌하면서 개인정보를 침해하지 않으면서 활용도 높은 데이터 모델 설계가 중요하게 인식된다. 개인의 민감 정보는 보호하면서 데이터의 활용도는 높이는 하나의 방안으로 비식별화 설계가 적용되는 사례도 시장에서 늘어나고 있다.

마. 고려사항 및 방향성

데이터 모델 설계는 유연성 및 확장성, 일관성, 정합성, 성능 등을 고려해 설계하는 것이 원칙이다. 이러한 원칙을 기준으로 데이터 모델의 품질 및 성능 분석 단계를 진행해 최적의 모델로 재설계되는 과정을 거치므로 품질을 높일 수 있다. 데이터 구조는 업무에 미치는 영향이 크기 때문에 완성도 높은 설계 결과물을 확보하기 위한 다양한 환경 또한 고려하는 것이 중요하다. 빅데이터의 처리, 활용 등으로 인한 저장 공간의 변화에 따라 비즈니스 이해뿐만 아니라 데이터베이스의 물리적인 측면을 이해한 설계 방법론이 요구된다.

3. 데이터 이행 구축

가. 데이터 이행 구축 정의

데이터 이행은 기존 시스템에 존재하는 데이터를 새로운 시스템에 알맞은 형식, 즉 파일시스템 혹은 데이터베이스로 옮기는 과정을 말한다. 쉽게는 이전에 살던 오래되고 비좁은 집에서 새로 지어진 집으로 이사 가는 과정과 유사하다고 할 수 있다. 데이터 이행 절차는 데이터 이행 전략 수립, AS-IS 및 TO-BE 데이터 모델 분석, 매핑 및 검증 설계, 프로그램 개발, 이행 데이터 검증 순으로 진행한다. 데이터 이행 환경의 최근 특징은 첫째, 과거보다 현저히 늘어난 데이터의 양이고 둘째는 이기종 시스템간의 이행 프로젝트가 늘어나면서 통합적인 이행·검증 프로그램 관리가 요구된다. 변화된 환경 속에서 이행 전담 조직 필수 운영, 통합관리 솔루션 도입, 자원 효율적 활용과 같은 움직임들이 대세를 이루고 있다.

나. 이행 전담 조직 필수 운영

최근에 난이도 높은 차세대 프로젝트에서 계획된 시점에 오픈하지 못하고 연기하는 사례가 발생하면서 전문 데이터 이행 조직을 구성하는 것이 성공적인 이행을 이끄는 데 중요한 요소로 인식된다. 또한 어떤 방식으로 R&R을 운영할 것인지 고민하고 해결책을 모색하는 것이 필요하다. 첫 번째로 매핑 정의서를 작성하는 부분이 TO-BE 모델을 설계한 모델러와 이행 개발자의 역할이 교차된 영역으로 의사소통이 중요하다. 과거에는 데이터 모델링이 종료되면 모델러는 변경 관리를 위한 일부 인력을 제외하고 프로젝트에서 Role off 했다. 그러나 최근에는 상시적인 의사소통을 위해 해당 모델러가 이행조직으로 합류해 오픈까지 수행하는 추세다. 둘째는 이행데이터의 검증 영역에서 기본 검증은 우선 이행팀이 수행하고 2차적 화면 검증은 개발팀에서 수행하므로 협업이 중요하다. 이행팀에서 건수, 합계, 코드, RI 등 기본 검증과 일반적인 비즈니스 검증까지 진행하지만 파악되지 않은 비즈니스 룰은 현업의 적극적인 참여가 중요하며 개발팀과 협의를 통해 비즈니스 검증에 지속적으로 추가하는 방법을 활용하고 있다.

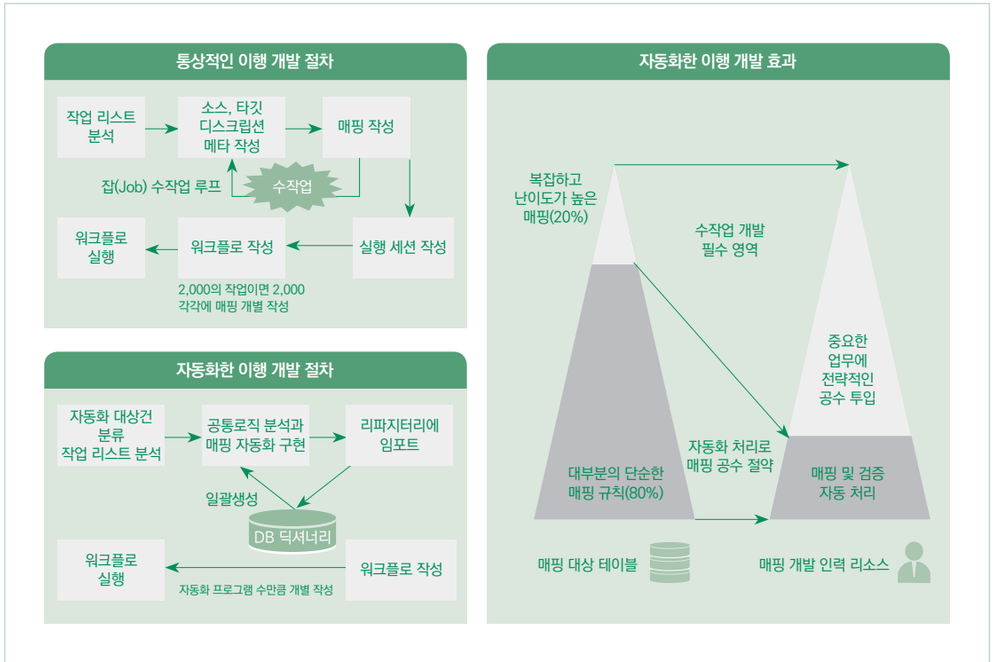
다. 통합관리 솔루션 도입 추세

과거에는 RDBMS와 같은 정형화된 구조에서 저장된 데이터를 대상으로 했지만 최근에는 DBMS도 다양할 뿐만 아니라, 파일·로그와 같은 반정형·비정형 데이터를 대상으로 하고 있다. 이러한 대상 데이터들을 개별 솔루션으로 관리하기 보다 표준화한 개발 방법론을 적용한 통합 솔루션을 도입할 필요가 있다. 전체적인 흐름을 관리하는 통합 뷰를 통해 프로그램을 관리하고 이행 개발자들의 개발 진척을 관리할 수 있는 부분 또한 프로젝트를 안정적으로 수행할 수 있는 중요 요건이다. 그리고 이행 테스트 기간에는 등록된 이행 프로그램들의 반복적인 수행을 통해 정확한 이행 시간을 예측하는 방법으로 솔루션을 활용한다.

라. 자원 효율적 활용이 핵심

이행 프로젝트에서는 한정적인 자원을 효과적으로 활용하기 위한 두 가지 측면을 고려해야 한다. 첫째는 이행 개발 인력 활용인데 이행 대상 테이블은 대용량 이행에서는 수천 개를 넘는 것이 대부분이므로 이 많은 대상을 매핑 방법에 따라 분류하는 것이 우선돼야 한다. [그림 6-1-2]와 같이 80:20 법칙으로, 80%는 1:1 매핑 관계이고 약 20% 정도가 N:1 또는 N:M 복합매핑으로 분류된다. 1:1 매핑은 DB

[그림 6-1-2] 자동화한 이행 개발 절차 및 효과



딕셔너리 등의 정보를 이용해 이행 프로그램을 자동화하고 업무적으로 중요한 복합 매핑에 대해 경험 많은 컨설턴트를 투입하는 전략이 중요하다. 둘째는 정해진 이행시간으로, 데이터 양은 늘어났지만 주어진 시간은 동일한 경우가 대부분이다. AS-IS 데이터 환경에 따라 1~2일 이내로 이행을 완료하거나 핵심 데이터를 이행·가동 후 점진적으로 데이터를 이행하는 경우도 있다. 하지만 차세대 프로젝트의 경우는 대부분 연휴를 이용한 2~3일 기간 동안에 모든 이행을 완료해야 하기 때문에 정확한 이행 시간 예측을 통해 데이터 정합성 검증, 이상 발생에 의한 원복에 대비해 최대 50%까지 시간을 단축할 필요가 있다. 또한 이행 아키텍처를 어떻게 설계할 것인지도 성능 및 이행 시간의 중요한 요소다.

마. 고려사항 및 방향성

성공적인 시스템 전환을 위해 이행 환경을 명확하게 정의하고 기술적 검토, 비용과 일정계획을 지속적으로 관리해야 한다. 오픈 이후에도 관리·유지할 작업들에 대해 통합 관리할 수 있는 이행 통합 거버넌스가 고려돼야 한다.

기존 기간계 시스템의 이행은 증가한 데이터 양에 대해 처리 및 관리 기술의 정확도를 높이려는 노력이 늘어날 것이다. 부가적으로 앱, 소셜, 센서 등 데이터 소스의 다양화가 계속될 것이므로 다양한 이행 환경에서도 전략적으로 발 빠른 대응이 필요해 보인다.

4. 데이터 거버넌스 컨설팅

가. 데이터 거버넌스 정의

과거에 활용하기 힘들었던 데이터도 유용한 자원으로 사용할 수 있는 환경이 도래하고 있다. 기하급수적으로 늘어나는 다양한 빅데이터 자원의 효과적인 관리와 진화를 위해 데이터 거버넌스의 필요성이 올라가고 있다. 축적된 데이터에서 가치를 끌어낼 방법을 고려하지 않고 데이터만 많이 수집하면 정작 필요할 때 원하는 데이터를 사용하지 못할 수 있기 때문이다. 이 문제를 해결하기 위해 데이터를 활용하는 조직에서는 데이터가 발생되는 과정부터 최종 활용되는 단계까지 데이터 생명주기 전반의 데이터 흐름을 파악해 비즈니스 목적에 필요한 고품질 데이터를 지속적으로 사용할 수 있도록 관리하는 체계인 데이터 거버넌스의 중요성에 주목하고 있다.

데이터 거버넌스는 전사적으로 보유한 데이터에 대한 관리 정책, 지침, 표준, 전략과 방향을 수립해 데이터를 지속적으로 관리할 수 있는 효율적인 인력, 조직까지 포함한 체계를 말한다.

다양한 소스에서 발생하는 많은 데이터를 비즈니스 목적에 맞는 고품질 데이터로 유지함으로써 조직의 가치 창출에 지속적으로 기여할 수 있도록 하는 것이 핵심이다. 빅데이터 시대에 적합한 전사 데이터를 고려한 데이터 거버넌스 컨설팅 중요도는 더욱 높아질 전망이다.

나. 데이터 리니지 관리를 통한 문제 점검

데이터 관리자의 입장에서 업무가 복잡해지고 시스템 간 연계가 늘어나며 지금 사용되는 데이터가 어떻게 생겨났으며, 어떤 과정을 거쳐 어디에서 사용되는지 점검하고 데이터 오류의 원인을 한눈에 파악할 수 있도록 하는 데이터 리니지 관리의 중요성이 강조되고 있다.

리니지는 우리말로 '계통'을 의미한다. 데이터의 생성, 변경, 이동 등 이력과 생명주기를 관리함으로써 전사 데이터가 적합한 과정을 거쳐 최적화한 형태로 활용되는지 시각화를 통해 데이터 관리자가 적시에 확인할 수 있도록 하는 것을 목적으로 한다.

데이터 리니지를 통한 데이터 흐름 파악은 전사 시스템 안에서 상호 작용하는 데이터의 투명성을 확보하고, 데이터 오류의 원인을 파악해 관리자가 적절한 조치를 내릴 수 있도록 돕는 역할을 한다. 또한 데이터 계보를 현행화해 점점 복잡해지는 데이터 시스템 환경에서 데이터의 정확성과 시스템의 안정성을 높이고, 데이터 거버넌스로서 고품질 데이터를 지속 관리할 수 있는 방안을 제공한다.

다. 빅데이터 분석 기술 적용

빅데이터 활용이 비즈니스에서 새로운 트렌드로 주목 받으며 데이터 사이언스 환경을 고려한 데이터 거버넌스의 필요성이 대두되고 있다. 특히 IoT 환경에서 발생되는 센싱 데이터와 SNS와 같은 플랫폼에서 발생되는 비정형 데이터를 비즈니스에 활용하기 시작하며 데이터 품질 개선이 더욱 중요해지고 있다. 머신러닝 알고

리즘으로 데이터 품질의 '이상'을 효과적으로 감지하고, 데이터 관리 체계에서 미래에 놓치기 쉬운 징후를 사전에 파악해 대비할 수 있는 환경을 만드는 데 중요한 역할을 할 수 있기 때문이다.

라. 데이터 카탈로그를 이용한 검색 활용

새로운 데이터 환경에서는 데이터를 분석하는 시간보다 데이터를 찾는 시간이 더 많이 소요된다. 이러한 산적한 데이터를 적재적소에 활용하기 위해 비즈니스 메타로 구성하는 데이터 카탈로그 서비스가 최근 많이 활용되고 있다.

테이블, 속성 등 물리적인 요소뿐 아니라 업무 용어, 설명, 문맥, 보유자, 사용 허가, 연관 비즈니스 용어를 표현해 주는 비즈니스 용어 사전을 메타 서비스로 구성하는 방식이다. 이러한 메타데이터를 이용해 현업, 데이터 분석가, 데이터 개발자까지 자신이 알고 있는 비즈니스 용어를 간편하게 검색하고 샘플 데이터까지 확인 가능하게 해 데이터 활용을 극대화할 수 있다.

마. 고려사항 및 방향성

데이터가 기하급수적으로 증가하는 추세와 함께 기업에서 데이터를 어떻게 처리했는지 기록할 것을 요구하는 규정들이 확대되고 있다. 이와 관련해 유럽연합(EU)에서는 개인정보보호규정(GDPR)을 도입해 소비자가 자신의 정보에 대해 갖는 권리 기준이 제시될 것으로 예상된다. 데이터를 활용하는 조직은 이를 준수하기 위한 데이터 거버넌스를 고려해야 할 것이다.

향후 데이터 거버넌스는 인공지능과 머신러닝 기술을 통해 빅데이터를 잘 활용하는 데 그치지 않고, 전사적 관점에서 활용되는 데이터의 종류, 프라이버시와 투명성 확보, 비즈니스 니즈에 따라 담당자의 역할과 책임 소재를 분명히 하는 일이 중요하다.

더 나아가 데이터가 필요할 때 정확하고 신뢰도 높게 활용할 수 있도록 접근성과 소유권을 부여하는 지능적이고 정교한 데이터 거버넌스가 데이터 시대를 선도하는 구심점 역할을 해줄 것이다.

5. 데이터베이스 성능 최적화 컨설팅

가. 데이터베이스 성능 최적화 정의

성능 최적화 컨설팅은 시스템 자원을 가장 적절하게 사용해 쿼리의 응답 시간을 최소화하는 작업을 말한다. 성능 최적화 절차는 분석 및 목표 설정, 최적화 방안 수립 및 적용, 성능 평가 순으로 수행한다. 이 분야의 컨설팅 추세는 과거에는 인덱스 전략, SQL 튜닝 등의 DB 성능에 가장 영향을 미치는 IO 효율화에 중심을 둔 컨설팅이 많았다면, 최근 들어서 데이터베이스 튜닝 자동화, 데이터베이스 수요 다양화, 하드웨어 기술 인프라 고성능화와 대용량 처리 기술의 발전으로 다양한 관점의 컨설팅을 요구하고 있다.

나. 데이터베이스 튜닝 자동화

최근에 오라클 등 글로벌 벤더들의 DB에 쌓인 통계를 기반으로 튜닝을 감지하고 쿼리 튜닝을 자동화하는 기능이 수준급으로 향상됐다. 또한 멀지 않은 미래에 AI에 의한 튜닝, 백업, 복구 등을 완전 자동화하겠다는 목표를 갖고 있다. 기존 버전에서 발견됐던 잘못된 최적화 판단들이 크게 줄어들고 실제 전문가들이 해오던 튜닝 작업에 대한 노하우가 상당부분 녹아 들었다는 평이다. 따라서 과거에는 사람의 경험과 지식이 필요한 영역이었지만, 투자 비용과 지속성 측면에서 자동화된 툴을 활용한 튜닝 요구가 증가하고 있다. 그러나 비효율적인 IO를 유발할 수밖에 없는 응용 개발 방식의 문제를 개선해 비약적인 성능개선 효과를 얻을 수 있는 튜닝 컨설팅 수요는 여전히 존재한다.

다. 데이터베이스 수요 다양화

전통적인 관계형 데이터베이스 영역은 스마트해진 자동화 도구와 IT 담당자의 기술력 향상으로 컨설팅 의존도가 많이 약해지고 있다. 하지만 최근 들어 성능 최적화 컨설팅은 차세대 구축 프로젝트에 포함돼 새로 도입되는 데이터베이스 특성에 맞게 전반적인 시스템 최적화를 목적으로 시장이 변화하고 있다. 또한 어플라이언스 제품의 지속적인 성장과 오픈소스 데이터베이스, NoSQL 데이터베이스 등의 확산으로 데이터베이스 수요층이 다양해졌다. 이러한 다양한 환경에서 데이터를 수집하고 처리하기 위해 데이터베이스 특징에 적합한 성능 고도화 컨설팅 서비스

가 요구된다.

라. 대용량 데이터 처리 기술 발전

과거에는 활용하기 힘들었던 로그성 데이터를 활용하기 위한 대용량 데이터 처리 기술이 발전하고 있다. 빅데이터 저장소로만 생각했던 하둡(Hadoop) 시스템에서 인메모리 데이터베이스로 데이터를 가져와 대용량 데이터를 분석한다거나 데이터베이스와 스토리지 사이의 IO 병목 현상을 최소화한 엑사데이터(Exadata)와 같이 초대용량 데이터 환경에서의 분석 전문 DBMS로 데이터 분석을 지원하는 사례가 늘고 있다. 이러한 빅데이터 시스템에서 성능 컨설팅은 물리적 분산 설계나 잘못된 액세스 패턴 같은 새로운 문제에 자주 부딪히지만, 대부분의 빅데이터 솔루션은 정교한 모니터링 또는 성능 분석 도구가 아직은 부족한 게 사실이다. 빅데이터 환경에서는 대량의 데이터 IO를 고려한 파티션 전략, 워크 테이블 등을 활용한 최적화 컨설팅이 요구된다.

마. 고려사항 및 방향성

관계형 데이터베이스는 정확성과 트랜잭션을 이용한 강력한 신뢰성을 제공한다는 점에서 주요 업무 시스템을 지원하는 용도로 사용돼 왔다. 응답시간 개선, 자원 사용 절감을 위한 성능 최적화는 여전히 중요한 업무다. 다만 어느 정도 규모 있는 기업이라면 데이터 전담 관리 조직을 운영하면서 자체적으로 성능개선을 하고 있는 것이 현실이다. 하지만 관계형 데이터베이스 외 오픈소스 등 빅데이터 플랫폼에서 대용량 데이터 처리 기술이나 다양한 데이터베이스 경험이 필요한 영역에서 성능 최적화 컨설팅의 수요는 꾸준히 있을 것으로 전망된다.

6. 향후 전망

이제는 데이터를 단순히 과거에 일어난 일(업무와 사실)의 기록이 아니라 새로운 서비스를 발굴하기 위한 자원으로 인식하고 있다. 기업들은 추출할 수 있는 데이터의 범위를 넓히려는 노력을 하고 있으며 더불어 내부 데이터뿐만 아니라 외부 데이터와 융합(mash-up)해 시너지가 날 수 있는 길을 적극적으로 찾고 있다. 이는 기업들이 자신의 비즈니스 외의 다른 비즈니스와 결합해 새로운 데이터 서비스를 발굴하는 경쟁에 돌입했음을 의미한다. 또한 기업들이 이러한 데이터를 자신의 서비스에 활용하는 것 이상의 데이터 판매, 데이터 개방 등 데이터 플랫폼 서비스까지 염두에 두고 있는 상황이다.

데이터 환경에서 경쟁 우위를 확보하는 방안으로서 클라우드 데이터 서비스에 대한 관심이 증가하고 있다. 아마존 AWS, 마이크로소프트 애저 등 클라우드 기반 데이터 서비스는 손쉽게 설정·운영·확장이 가능하다는 측면에서 데이터의 생산성 향상을 기대할 수 있다. 물론 중요한 정보를 클라우드에 보관할 때의 보안 문제나 성능 보장의 어려움, 클라우드 서비스 중단 문제, 기업들의 데이터 소유의 익숙함 등의 원인으로 국내에서는 클라우드 도입이 다소 느린 것도 사실이다. 하지만 비용 측면과 유연성, 확장성 등 장점이 부각되고 있으므로 기업들은 일부 영역에서 여전히 클라우드 데이터베이스 서비스를 선호하는 것으로 평가된다.

클라우드 서비스 벤더들이 일반 기업들보다 훨씬 많은 보안 투자를 진행해 기존 고정관념들이 깨지고 있는 점도 긍정적인 신호다. 무엇보다 비용과 편의성 측면에서 장점이 있다는 것에 대해서는 이견이 없다. 주요 기능인 모니터링, 위협 검색, 튜닝, 자동 패칭, 백업 등의 유지보수를 인간이 아니라 AI가 수행하고, 이러한 기능이 더 정교해질수록 데이터산업계의 인식 변화가 가속화될 것으로 전망된다.

제2장

데이터 품질 컨설팅

과거에는 주로 IT 분야 기업이 데이터를 활용한 비즈니스를 추진한 반면, 최근에는 ICT 붐과 함께 비 IT 분야 기업의 데이터 활용 비즈니스가 활발해졌다. 이에 따른 데이터 품질 관리 기술에 대한 수요도 증가 추세다. 그러나 기업 의사결정권자의 데이터 품질 관리 중요성에 대한 인식이 낮고 품질 관리 활동에 소요되는 비용으로 인해 데이터 품질 관리 활동을 적극적으로 추진하는 기업은 아직 많지 않다. 앞으로 제4차산업혁명에 대응해 기업의 경쟁력 강화와 공공 프로젝트의 성공을 위해서는 민간 및 공공분야의 적극적인 데이터 품질 관리 활동이 필요하다.

1. 개요

전 세계적으로 데이터를 활용한 다양한 비즈니스 사례가 소개되면서 기업, 공공기관은 하드웨어가 허용하는 한 생산하고 있는 데이터를 삭제하지 않고 축적하고 있다. 이에 따라 최근 몇 년 간 데이터는 빠른 속도로 늘어나고 있다. 그러나 데이터 축적과 별개로 데이터 품질 관리 활동이 미흡해 오히려 저품질 데이터의 활용에 따른 다양한 문제도 발생하고 있는 상황이다. 미국 「하버드비즈니스리뷰(Harvard Business Review)」는 2016년 미국 내 저품질 데이터로 인한 비용이 3.1조 달러가 소요될 것이라는 충격적인 내용을 보도했다. 실제 국내에서도 2017년 공공분야 빅데이터 프로젝트가 데이터 오류로 인해 수십억을 낭비했다는 보도가 있었다. 실패 사례를 쉬쉬하는 우리 기업문화를 감안하면 공공, 민간 분야 데이터 품질 문제로 인한 예산낭비, 비즈니스 실패 사례는 더 많을 것으로 추정된다.

데이터 품질 컨설팅은 데이터 품질 관리 활동이 미숙한 기업 및 공공기관의 데이터 비즈니스 프로젝트 실패를 예방할 수 있다. 데이터 품질 컨설팅은 데이터 구조 컨설팅과 데이터 값 컨설팅으로 구분된다. 데이터 구조 컨설팅은 데이터 모델 현행화, 데이터 표준 점검, 데이터 표준 관리체계 수립으로 구성되며, 장기적으로 데이터 품질 수준을 확보·유지하기 위해 점검하는 항목이다. 데이터 값 컨설팅은 데이터 값의 품질 진단 및 개선, 오류 유형에 따른 품질 관리 지침 및 체계 수립으로 구성되며 보유한 데이터를 단기적으로 활용하거나 외부 유통하기 위해 점검하는 항목이다.

참고로 현재 우리나라에서는 데이터 컨설팅과 별개로 공공기관의 전사적 품질 관리 활동 평가를 진행하는 '공공데이터 품질 관리 수준 평가(행정안전부)' 제도와 정보시스템 단위별로 데이터 품질(정합율), 데이터 품질 관리 체계를 심사해 인증하는 '데이터인증(한국데이터진흥원)' 제도가 있다.

• 필자 | **이우용**(한국정보기술단 대표)

2. 데이터 품질 관리 동향

가. 공공데이터 측면

공공데이터는 정부의 적극적인 정책으로 인해 양적인 측면에서는 단기간에 많은 성과가 있었지만 질적인 측면에서는 여전히 미흡한 부분이 많다. 이러한 부분을 개선하기 위해 기존 개방 DB 위주로 시행된 공공데이터 품질 관리 수준 평가가 2018년부터는 기관의 전사적 품질 관리 활동 평가로 전환되고 대상도 600여 개 기관으로 확대 실시될 것으로 알려졌다.

2009년 스마트폰의 급격한 보급과 함께 버스정보 제공 애플리케이션의 큰 인기와 더불어 공공데이터 활용이 폭발적으로 증가할 것으로 기대를 모았으나 그 후 버스정보만큼 비즈니스에 활용되는 공공데이터는 아직 없다. 이에 따라 정부 정책도 공공데이터의 양적인 개방보다는 민간의 수요에 맞는 고품질 공공데이터 제공에 초점을 맞추고 있으며, 이와 같은 인식이 공공분야에 확산되고 있다. 일부 공공기관의 경우 별도의 데이터 품질 컨설팅 용역 사업을 추진하고 있으며, 정보시스템 사업과 함께 데이터 품질 개선 사업을 병행하는 기관도 증가 추세에 있다. 정보시스템 사업과 데이터 품질 개선 사업을 추진하는 기관의 경우 사업 성과를 공인 받기 위해 데이터인증과 같이 제3기관의 검증을 받는 사례 역시 증가하고 있다.

데이터 품질 관리 활동 및 품질 고도화와 관련해서는 공공이 민간에 비해 앞서 있는 것으로 추정된다. 이유는 데이터 품질 개선 업무가 지속적인 투자 및 리소

스 투입이 필요한 업무 영역으로 공공이 민간에 비해 비교적 재원 마련이 용이하기 때문이다. 또한 법으로 규정한 공공데이터 개방 이슈에 따라 데이터 품질 고도화에 대한 인식 역시 공공 분야에 더 확산돼 있다. 이에 따라 당분간 데이터 품질 컨설팅 시장은 공공분야에서 더 커질 것으로 예상된다.

나. 민간데이터 측면

제4차산업혁명 이슈는 민간 기업의 데이터 품질에 대한 인식 전환을 가져오는 주요 요인으로 볼 수 있다. 기존 데이터는 IT 기업의 비즈니스 전유물이었으나 제4차 산업혁명과 ICT 산업 발전은 데이터 활용 업종의 벽을 허무는 계기가 됐다. 2016년부터 빅데이터, 인공지능, 사물인터넷과 같은 비즈니스가 본격화되고 기업들은 관련 비즈니스 부서 신설과 함께 데이터 품질에 대한 업무를 추가했다. 이는 데이터 품질 관리를 지속적으로 해오던 일부 IT, 금융업종의 대기업을 제외하고 ICT 비즈니스를 추진하는 대부분의 기업에 동일하게 발생하는 현상으로 볼 수 있다.

데이터 품질 관리 업무를 추진하는 기업들은 본격적으로 데이터를 비즈니스에 활용하고 이에 대한 비전을 갖고 있다. 앞에서 소개한 바와 같이 데이터 품질 활동은 지속적인 투자와 리소스 투입이 필요한 업무 영역이다. 따라서 투자에 대한 ROI를 분석해야 하는 기업의 특성상 사업 추진이 쉽지 않다. 그러나 ICT 비즈니스 추진에 따라 데이터 품질 관리 활동은 기업 경쟁력 강화를 위해 더 미뤄서는 안 되는 업무 영역이라는 인식이 기업의 의사결정권자들에게 점차 확산되고 있다.

민간의 이러한 추세는 2017년 신세계/이마트의 통합 온라인 쇼핑몰의 데이터 품질 관리 컨설팅 및 데이터관리인증 3단계 획득, 2018년 SK브로드밴드의 통합 데이터 관리 시스템의 품질 관리 컨설팅 및 데이터관리인증 3단계 획득 사례로 쉽게 확인할 수 있다.

다. 국내외 데이터 품질 관리 도구 현황

효율적인 데이터 품질 관리 업무 추진과 관련 컨설팅을 위해서는 데이터 품질 관리 도구의 활용이 중요하다. 현재 시중에는 데이터 품질 관리 전문 기업의 다양한 품질 관리 도구 상품이 출시돼 있다. 대부분의 데이터 품질 관리 도구는 회사마다 기능 및 인터페이스가 다소 차이는 있지만 데이터 품질 관리 기능의 핵심이라 할 수 있는 데이터에 대한 정보 수집, 통계 정보를 제공하는 데이터 프로파일링, 메타

[표 6-2-1] 국내 데이터 품질 관리 도구

품질 관리 도구	프로파일링	품질 개선 관리	업무 규칙 관리	스케줄 관리	결과 분석	보고서
A사	○	○		○	○	○
B사	○	○	○	○	○	
C사	○	○	○	○	○	
D사	○	○	○			○

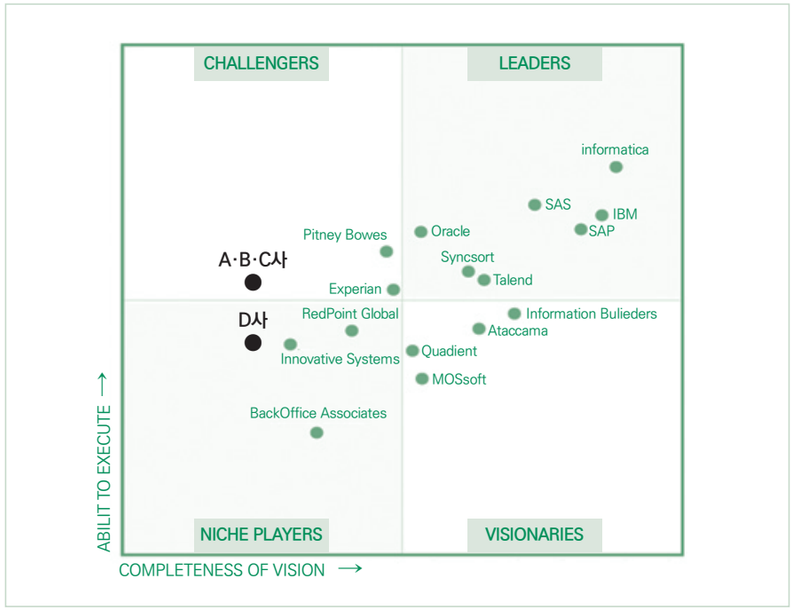
출처: 고동규·최용락, 「데이터 정확성 사례 연구」, 숭실대학교, 2018.

데이터 관리 등의 기능과 세부적으로는 업무 규칙 관리, 스케줄 관리, 결과 분석, 보고서 산출 기능은 유사하다. [표 6-2-1]은 국내에서 가장 많이 사용 되는 4개의 데이터 품질 관리 도구에 대한 기능을 간략히 정리한 내용으로 도구별 기능 현황을 쉽게 파악할 수 있다. 국산 데이터 품질 관리 도구는 기본적으로 데이터 품질 관리 기능이 우수한 것으로 조사됐으나 데이터 품질 관리 도구들이 지원하는 DB 관리 시스템(DBMS, Database Management System)과 처리되는 DB가 적다는 단점이 발견되기도 했다.

해외 데이터 품질 관리 도구 현황은 미국의 정보기술 연구 및 자문 회사인 가트너(Gartner)의 「2017 Gartner Magic Quadrant for Data Quality Tools」 자료로 한눈에 확인할 수 있다.

이 자료에서는 각 제품을 [그림 6-2-1]과 같이 시장 내 위치로 구분한 사분면(Leaders, Challengers, Visionaries, Niche Players)으로 분류했다. [그림 6-2-1]의 각 사분면에 대해 간략하게 설명하면 다음과 같다. 'Leaders'는 모든 데이터 품질 기능에 걸쳐 깊이 있는 강점을 보여주고, 데이터 품질 시장의 역동적인 경향에 대한 명확한 이해가 있음을 뜻한다. 'Challengers'는 강력한 제품 기능과 견고한 영업·마케팅 실행과 함께 존재감·신뢰성·생존력을 확립하고 있지만, 'Leaders'와 같이 넓은 분야에서는 지도력과 혁신을 보여주지 못할 수 있다. 'Visionaries'는 훌륭한 고객 경험을 제공할 수 있지만, 규모·시장존재·브랜드인지도·고객기반·회사자원이 부족할 수 있다. 'Niche Players'는 제한된 수의 산업 또는 데이터 영역 등을 전문으로 하지만, 일반적으로 시장 점유율이나 존재감과 기능이 제한적이거나, 재무 건전성이 부족하다. 해외 품질 관리 도구는 「2017 Gartner Magic Quadrant for

[그림 6-2-1] Magic Quadrant에 적용한 국내 데이터 품질 관리 도구 위치



출처: Mei Yang Selvage·Saul Judah·Ankush Jain, 2017 Gartner Magic Quadrant for Data Quality Tools, Gartner Inc., 2017.

Data Quality Tools」을 참고한 결과, 국내 데이터 품질 관리 도구와 기능상의 큰 차이는 없었으나 지원하는 DBMS와 처리되는 DB가 다양한 편으로 조사됐다.

데이터 품질 도구의 기능 부분을 기준으로 Magic Quadrant에 국내 품질 도구 제품을 적용해 보면 A·B·C사는 'Challengers'에 자리할 수 있다. 기능이 다소 부족한 D사는 'Niche Players'에 자리한다. 데이터 품질관리 도구는 기업의 데이터 품질 관리뿐 아니라 데이터 품질 컨설팅의 중요한 역할을 하므로 국내 데이터 품질 관리 도구들의 개선이 필요하다. 국내 품질 관리 도구들이 'Leaders'에 소속되기 위해서는 데이터 품질 시장에 대한 명확한 이해가 우선돼야 한다. 다양한 DBMS와 대규모 DB 처리를 지원해야 하며, 데이터 품질 관리 또는 컨설팅에 도움이 되는 다양한 기능 개발 검토가 필요하다.

라. 공공데이터 품질 정확성 진단 사례

공공데이터포털(www.data.go.kr)에서 유통 중인 데이터 중 사용자들의 많은 관심이 있는 데이터 중 하나로 '전국도시공원표준데이터'를 들 수 있다. 이 데이터를

[표 6-2-2] 공공데이터 품질 정확성 사례 연구 결과

컬럼	진단 건수	오류 건수	오류율
관리번호	13,135	702	5.34%
공원명	13,135	2,838	21.61%
공원구분	13,135	0	0.00%
소재지 도로명주소	13,135	6,711	51.09%
소재지 지번주소	13,135	1,715	13.05%
위도	13,135	6	0.05%
경도	13,135	6	0.05%
공원면적	13,135	27	0.21%
공원 보유시설(운동시설)	13,135	9,579	72.93%
공원 보유시설(유희시설)	13,135	8,418	64.09%
공원 보유시설(편익시설)	13,135	9,672	73.64%
공원 보유시설(교양시설)	13,135	12,382	94.27%
공원 보유시설(기타시설)	13,135	11,040	84.05%
지정고시일	13,135	1,940	14.76%
관리기관명	13,135	783	5.96%
전화번호	13,135	931	7.08%
데이터기준일자	13,135	0	0.00%

출처: 고동규·최용락, 「데이터 정확성 사례 연구」, 숭실대학교, 2018.

오픈소스인 A사의 데이터 품질 도구를 이용해 데이터 품질 정확성을 확인한 결과는 [표 6-2-2]와 같다.

공공데이터인 '전국도시공원표준데이터'의 품질 정확성 결과를 보면, 17개 컬럼 중 15개의 컬럼에 문제가 있었고, 평균 오류율은 33.87%였다. 공공데이터포털에서 제공하는 표준 데이터 중에 조회 수가 가장 많은 데이터임에도 데이터 품질이 낮은 것을 확인할 수 있다. 이 수치는 2017년 데이터품질인증 전체 심사 건수의 평균 오류율인 1.31%에 비하면 매우 충격적인 결과로 볼 수 있다.

대부분의 오류가 공원 보유시설 컬럼 및 공원명, 도로명주소에 집중돼 있는

것으로 보았을 때 데이터 표준화가 미비돼 있을 가능성이 높으며 이와 같은 오류는 쉽게 개선될 수 있다. '전국도시공원표준데이터'는 대체공휴일, 주52시간 근로제와 함께 활용도가 높은 데이터로 판단되며 데이터 오류가 개선될 경우 네비게이션, 지도서비스 등 실생활과 밀접한 서비스에 다양하게 활용될 수 있을 것으로 기대된다. 따라서 이와 같은 분야는 데이터 품질 컨설팅이 필요한 대표적인 분야로 볼 수 있다.

3. 데이터 품질 컨설팅 전망

기존 기업과 기관이 보유한 데이터는 빅데이터, 인공지능, 사물인터넷, 스마트 팩토리 등의 핵심 기반이다. 또한 데이터 분석, 선도적인 데이터 활용 기술 및 비즈니스 아이템을 가진 기업이 중요한 데이터를 가진 기업과 협업해 사업을 추진하는 경우가 일반화돼 있다. 따라서 데이터 품질 관리는 매우 중요한 요소로 볼 수 있다. 점차 ICT 비즈니스가 산업 전반에 걸쳐 추진되면서 데이터 품질 관리 컨설팅이나 인프라 구축과 연관된 데이터 품질 관리는 전 산업으로 확대돼 폭발적인 성장이 예상된다.

단기적으로 데이터 품질 컨설팅에 대한 수요는 공공분야가 더 클 것으로 예상된다. 이는 공공데이터를 활용하기 위한 기업의 수요가 커져 공개 수준이 증가하는 것에 기인한다. 민간 기업의 경우 데이터 품질 관리를 위한 비용 최소화를 위해 자체적으로 데이터 품질 관리 활동 업무를 추진할 가능성이 높다. 다만 장기적인 관점에서 큰 규모의 데이터 품질 관리 프로젝트와 컨설팅 시장은 여전히 대기업 위주로 성장할 것으로 예상된다.

제3장

빅데이터 구축 컨설팅

기업에서 생성되는 방대한 정형·비정형 데이터를 기존 정보계에 어떻게 접목할 것인지, 이를 통해 의사결정의 질적 향상을 도모하고, 인공지능이나 학습지능을 위한 재료로 활용할 것인지를 고민해야 한다. 빅데이터를 물과 공기와 같이 늘 우리의 삶을 지탱하는 필수불가결한 재료로 인식하고 활용하는 것이 중요하다. 이 글에서는 기존 정보계가 빅데이터 구축 환경으로 진화해 가는 사례들을 살펴보고, 그동안 빅데이터 구축을 위해 수행된 업무들을 알아보겠다. 데이터레이크를 이용해 빅데이터 분석 경향과 구축된 빅데이터의 활용도를 높이기 위한 빅데이터 거버넌스 기능에 관해서도 소개한다.

1. 기존 정보계의 진화

현재 많은 기업에서 운영하는 정보계가 본격 구축된 지 20여 년이 넘었다. 그동안 내외부에서 정형/비정형 데이터를 포함해 분석하려는 데이터 양이나 종류는 기하급수적으로 늘었다. 이런 데이터 분석을 통한 의사결정 또는 고객 마케팅으로의 활용 속도는 거의 실시간으로 요구되고 있다. 하드웨어 성능향상과 가격하락, 막대한 양의 데이터 처리를 위한 기반 기술, 이를 뒷받침하는 다양한 오픈소스들의 발전으로 저렴하면서도 고성능의 대용량 데이터 처리 플랫폼을 온프레미스(On-premises)로 갖출 수 있는 시대가 됐다. 이런 환경조차 빌려서 사용하는 퍼블릭 클라우드를 통한 플랫폼 구축도 큰 기업들을 중심으로 보편화해 가는 추세다.

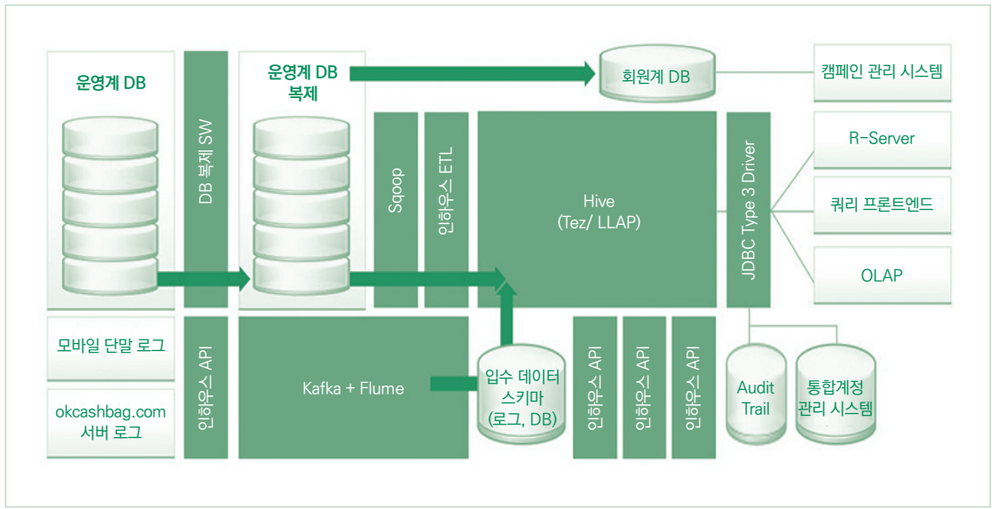
이같이 기존 정보계 환경의 변화를 주도하는 요인은 두 가지로 정리할 수 있다. IT 측면에서는 저렴한 비용으로 운용 가능한 환경 구축, 비즈니스 측면에서는 다양한 분석요구사항을 어떻게 쉽고 빠르게 수용할 것인가 하는 것이다.

가. IT 측면에서의 정보계 요구사항

그동안 IT 부서는 빅데이터 기술을 접목해 기존의 정보계 환경을 저렴하게 운용·

• 필자 | **김형태**(비투엔 부사장)

[그림 6-3-1] OK캐시백의 정보계 구성도

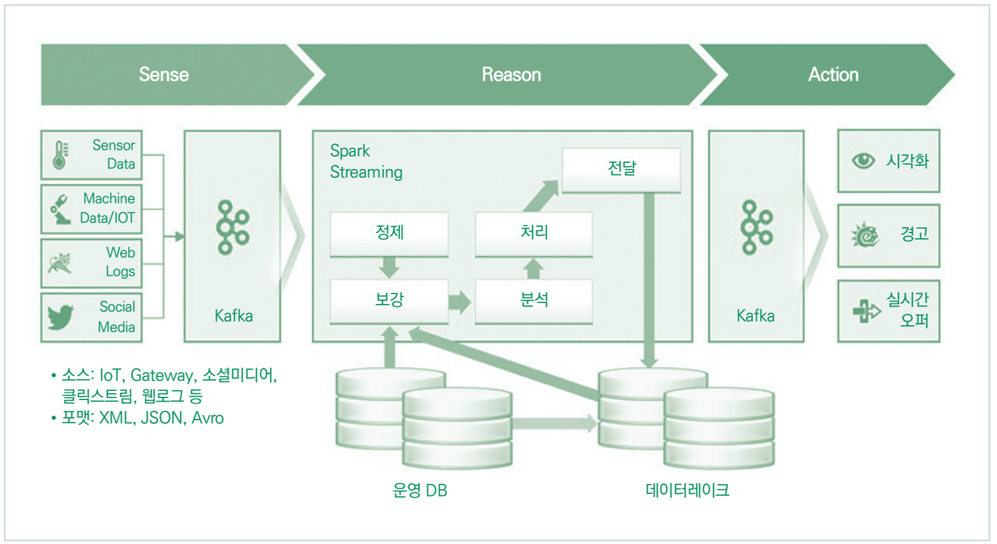


출처: 「클라우데라 세션 2017」 컨퍼런스 발표자료, 2017.7

유지·보수할 수 있는 방향으로 끌고 왔다. 대부분의 기업은 저렴한 분산처리 환경의 하드웨어에 하둡(Hadoop)을 적용하면서 관련 오픈소스와 상용도구를 사용하는 것으로 목적을 달성하기도 했다. OK캐시백 사례를 하나의 예로 들 수 있다. OK캐시백은 그동안 상용 데이터 웨어하우스용인 RDBMS(Relational Data Base Management System)를 사용해 정보계를 구축했다. OK캐시백은 2014년 하둡으로의 전환을 위한 파일럿 시스템을 구축했다. 2016년 데이터 수집, 정제, 저장 등을 오픈소스(하둡을 중심으로)를 활용해 전체 정보계로 이관해 사용하고 있다. 이런 시도는 OK캐시백 외에도 여러 회사에서 실행하고 있다.

오픈소스를 이용해 구축하는 정보계 환경에서는 기존의 정보계용 DBMS에서 실행 가능한 특정 레코드의 삭제나 업데이트를 할 수 없었다. 혹은 HBase(Apache)를 통해 제한적인 환경으로만 구축할 수 있었다. 그러나 클라우데라(Cloudera)에서 쿠두(Kudu)라는 스토리지 솔루션을 공개하면서 데이터의 빠른 삽입과 업데이트까지 SQL을 통해 가능하게 돼 기존의 정보계용 DBMS를 어느 정도 대처할 수 있게 됐다. 이렇게 기술이 발전해감에 따라 기존 정보계 환경의 변화는 더욱더 가속화할 전망이다.

[그림 6-3-2] 인도네시아 'Lippo Group'의 상거래 분석 시스템



출처: Daniel Clake, Informatica Next Generation Analytics

나. 비즈니스 측면에서의 정보계 요구 사항

기업에서 발생하고 있으나 방대한 양 때문에 분석에서 제외된 로그성 데이터, 고객상담데이터(Voice of Data), 챗봇과 같은 비정형 데이터 등도 이제는 기존 정보계와 통합하는 작업이 필요해졌다. 이런 데이터를 실시간으로 분석해 고객에게 선도적인 마케팅을 구현하거나, 장비와 서비스의 예측된 유지보수를 실현하자는 요구가 높아지고 있다. 이런 요구 사항을 만족시키기 위해서는 관리 가능한 데이터 레이크를 구축해야 하며, 실시간 스트리밍 데이터 처리가 가능해야 한다. 또 실제 고객에게 온라인 오퍼를 제공하는 등 실제 구현이 수행될 수 있도록 시스템을 구축해야 한다.

이렇듯 이제 정보계는 단순한 데이터 저장소나 이를 활용한 단순 분석시스템 역할을 넘어서고 있다. 더 나아가 실시간으로 정형/비정형 데이터를 수집하거나 정제하고, 알고리즘을 통해 분석·학습해 고객에게 실제 서비스를 제공하는 시스템으로 발전하고 있다. 다음 예는 인도네시아의 거대 상거래 회사들을 운영하는 그룹에서 하루 약 3억 명의 고객들로부터 발생하는 구매 정보와 SNS 정보를 실시간으로 추출·통합해 판매 캠페인에 활용하고 있는 사례다.

2. 빅데이터 정보화 전략계획 수립

빅데이터 시스템 구축에 앞서, 빅데이터로 무엇을 할 것인가를 먼저 정리해보고 실제 구축 방향을 설정하는 것은 매우 중요한 선행 업무다. 이런 빅데이터 정보화 전략 계획은 일반적인 정보화전략 계획 수립 프로세스에 준해 수행할 수 있다. 환경분석, 현황분석, 목표모델설계, 이행계획수립 프로세스로 나눌 수 있다. 빅데이터 구축을 원하는 조직의 특성에 맞춰 프로세스와 산출물을 결정해 수행하는 것이다. 각 사의 특성에 맞춘 방법론의 테일러링(tailoring)이 필요하겠지만, 일반적으로 다음과 같은 수행업무로 구성된다.

- 환경분석: 사업추진 필요성과 해당 사업에 대한 정책·경제·사회 환경을 분석하고, 현재 적용 가능한 빅데이터 기술을 제시한다. 사업을 추진하기 위한 일반환경, 기술환경 관련 선진사례를 분석하고 시사점을 도출한다.
- 현황분석: 빅데이터 활용과제 업무와 공통기반 도출을 위해 전사적인 요구사항을 도출하고, 관련 정보시스템 현황을 분석한다. 이를 통해 목표업무설계와 정보화 방향성을 도출한다.
- 목표모델설계: 빅데이터 구축을 통한 정보화 비전과 전략을 수립하고, 빅데이터 활용 과제에 대한 업무 프로세스를 설계하며, 기술구조(인프라 아키텍처 포함)에 대한 미래모형을 설계한다.
- 이행계획수립: 이행과제를 정의하고 우선순위를 도출하며, 빅데이터 구축 단계별 로드맵을 정의해 통합추진일정계획을 수립한다. 필요에 따라 구축 프로젝트를 실행하기 위한 구축제안요청서를 수립된 정보전략계획에 맞춰 작성하기도 한다.

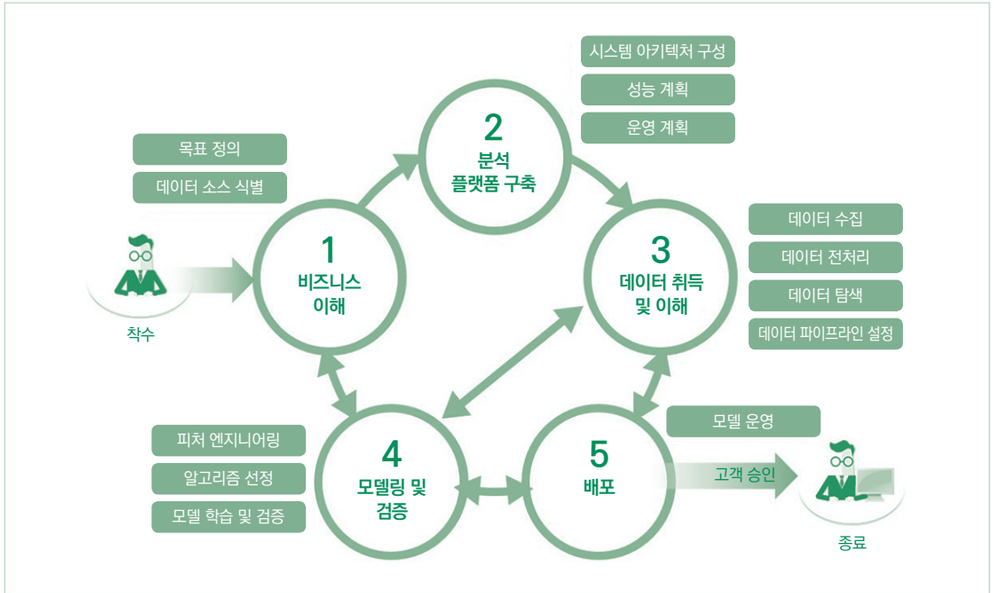
3. 빅데이터 구축 컨설팅

빅데이터 시스템을 구축하는 데 있어 기존 정보계를 빅데이터 환경으로 전환하거나 미래에 사용될 분석의 재료가 되는 데이터들을 우선 모아서 데이터레이크를 구축해야 한다. 이후 분석 프로젝트를 별도로 진행하는 경우에는 일반적인 구축 프로젝트의 단계인 준비→분석→설계→이행→운영으로 수행할 수 있다.

대부분의 빅데이터 구축 프로젝트에서는 목표하는 분석결과를 얻기 위해 분석모델을 설계하고, 이를 적용한 결과에 만족할 수 있을 때까지 분석모델을 변경하거나 수정하게 된다. 이때 결과를 검증하는 과정을 반복해야 한다. 일반적인 정보시스템 구축 프로젝트와 같이 정해진 일정대로 수행하기는 매우 어려울 수 있으므로 프로젝트 계획 시 이를 고려해야 한다. 실제 고객사에서 시스템 구축에 직접 참여하고 운영까지 수행하지 않으면, 계획된 프로젝트 종료가 어려울 수 있다는 점을 염두에 둬야 한다.

[그림 6-3-3]은 빅데이터 분석 플랫폼을 구축하기 위한 프레임워크를 설명하고 있다. 이 프레임워크는 프로젝트 구성원에게 작업할 수 있는 지침과 틀을 제공하며, 모든 단계는 필요에 따라 반복 수행될 수 있다. 특히 '데이터 취득과 이해'와 '모델링과 검증' 과정은 목표 분석 결과를 도출할 때까지 단계를 반복 수행해야 한다. 각 단계에 대해 자세히 설명하면 다음과 같다.

[그림 6-3-3] 빅데이터 분석 플랫폼 구축 프레임워크



출처: 비투엔

- 비즈니스 이해: 목표 정의와 비즈니스 타깃 변수를 정의하고, 해당하는 데이터 소스를 식별해 구축할 시스템 규모를 산정한다.
- 분석 플랫폼 구축: 정보전략화계획이 사전 수행돼 기술 아키텍처가 결정됐다면, 이를 바탕으로 시스템 아키텍처를 구성해야 한다. 즉 프로젝트 제안단계에서 정의된 기술 아키텍처와 전 단계에서 식별된 데이터 종류와 크기를 바탕으로 시스템 아키텍처를 정의하고 이를 도입 구축한다. 향후 배포단계에서 수행할 구축된 시스템의 성능 테스트를 계획하며, 향후 시스템 운영계획을 수립하고 변경·관리한다.
- 데이터 취득과 이해: 모델링을 하기 위해 타깃 변수와 관련된 고품질의 데이터세트를 생성하고 구축된 분석 플랫폼에 배치한다. 데이터를 정기적으로 갱신하며, 새롭게 스코어링하는 데이터 파이프라인을 개발한다.
- 모델링과 검증: 피처 엔지니어링, 모델 학습 과정을 통해 머신러닝 모델을 위한 최적의 데이터 변수를 결정한다. 대상을 가장 정확하게 예측하는 모델을 설정해 프로덕션 환경에 적합한 머신러닝 모델을 만든다.
- 배포: 구축된 데이터 파이프라인이 포함된 모델을 최종 승인하기 위해 분석 플랫폼 환경에 대한 이행과 운영인수를 수행한다. 이 단계에서 구축된 플랫폼의 성능과 운영정책을 최종 점검하고 시스템 이행을 한다.

4. 전망

앞서 언급한 바와 같이 기업 정보계의 진화는 앞으로도 계속 진행될 것이다. 컴퓨터월드[1] 영국판(Computerworld UK)이 발표한 '2017년 빅데이터와 BI 트렌드'에서는 데이터레이크의 활용성이 증가하면서 셀프서비스 비즈니스 인텔리전스(BI), 스트리밍 분석, 머신러닝 분야가 증가할 것이라고 전망했다. 몇 가지 전망을 정리하면 다음과 같다.

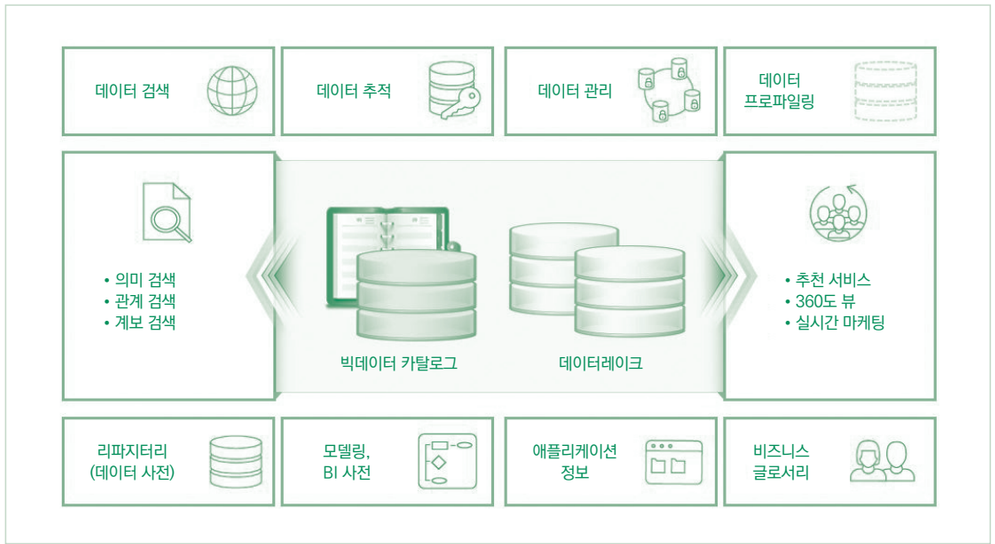
1 Scott Carey, "2017년 빅데이터와 BI 트레드: 머신러닝, 데이터레이크 그리고 하둡과 스파크", Computerworld UK, 2016.12.30.(http://www.itworld.co.kr/news/102786)

- 관리 가능한 데이터레이크: 수년 동안 분산된 데이터를 데이터레이크로 통합하는 것은 전략과 결합해 신뢰할 수 있는 데이터 기반에서의 통찰력을 제공할 것이다. 향후 ILM 기능을 수용한 확장된 데이터레이크 구축이 보편화될 것이다.
- 하둡, 그 너머의 움직임: 스파크(Spark)와 하둡은 서로 장단점이 존재하나, 클라우드 기반, 머신러닝, IoT 서비스에서의 스파크는 하둡의 대안으로서 제공될 것이다.
- 머신러닝 받아들이기: 예측 분석, 고객 통찰력, 추천 엔진, 사기와 위협방지 등과 관련해 폭넓게 활용될 것이다.
- 클라우드 기반의 분석: 기업들의 더 많은 핵심 데이터 저장소와 분석 워크플로가 클라우드로 전환되고, 클라우드 내에서 분석이 주류로 자리 잡게 될 것이다.
- 스트리밍 분석: 기존의 배치 분석 대신 기업 내에서 스트리밍되는 데이터들의 모니터링에 유용하게 활용될 것이다. IoT 배포를 고려하는 많은 기업에서의 수용이 지속해서 증가할 것이다.
- 기업들은 여전히 데이터 과학자 필요: 기업 내에서 필요한 데이터 과학자들에 대한 수요는 계속 증가할 것이다. 이들의 셀프 분석을 위한 비즈니스 데이터 정의와 관리가 중요해질 것이다.
- 더 많은 셀프서비스 비즈니스 인텔리전스: 비즈니스 사용자가 분석과 통찰력에 직접 접근할 수 있는 셀프서비스를 통해 분석 영역의 전환을 꾀할 것이다.

이처럼 비즈니스 사용자들이 기업 내 데이터를 셀프서비스로 분석하기 위해서는 데이터 품질이 정확하게 유지돼야 하며, 데이터에 대한 비즈니스적인 설명과 해석이 필요하다. 이에 따라 전 세계적으로 빅데이터 영역의 데이터 거버넌스에 대한 중요성이 강조되고, 규모가 큰 기업을 중심으로 데이터레이크 구축과 데이터거버넌스 시스템으로의 전환이 진행되고 있다.

결론적으로 지난 5~6년 동안 많은 기업이 빅데이터 플랫폼 구축을 위해 노력해왔으며, 이런 시도들이 온프레미스에서 클라우드 환경으로 서서히 이동하기 시작했다. 이제 기업의 수많은 데이터와 분석 결과물들이 통합되고, 이를 비즈니

[그림 6-3-4] 빅데이터 거버넌스 주요 기능과 리파지터리 구성



출처: 관리되는 셀프서비스 전략, 「인포매티카와 타블로」

스 사용자들이 어떻게 하면 쉽고 정확하게 획득해 업무에 활용할 수 있을지에 관심이 쏠리고 있다. 더불어 빅데이터 거버넌스 구축에 대한 요구가 점점 더 강해질 것으로 전망된다.

제 4 장

데이터 분석 컨설팅

기업들이 데이터 분석을 디지털 핵심역량으로 인지하고 역량을 내재화하려고 노력하고 있다. 선도 기업은 이미 상당한 수준으로 역량 내재화를 이뤄가고 있다. 데이터 분석 컨설팅의 주제 역시 데이터 분석모델을 개발하고 빅데이터 분석 인프라를 마련하는 단계에서 한 걸음 더 나아가, 도입한 분석모델과 운영을 고도화하는 단계로 확장되고 있다.

1. 데이터 분석가 역할의 재정립

가트너(Gartner Data & Analytics Summits 2017)는 전문적 통계나 컴퓨터공학 지식을 가진 분석 전문가(Data Scientist)의 역할이 많이 축소될 것이라 예상한다. 분석전문가의 전문지식과 하드코딩을 통해 수행 가능했던 영역의 40% 이상이 2020년까지는 자동화될 것이라 주장한다. 앞으로는 분석의 대중화로 인해 시티즌 데이터 과학자(Citizen Data Scientist)의 역할이 커질 것으로 예측하고 있다.

하드코딩 방식 분석 도구의 많은 부분이 점점 더 자동화·템플릿화한 솔루션 형태로 진화하고 있다. 데이터 전처리부터 분석과 검증에 이르는 모든 과정을 워크벤치(Workbench) 형태로 손쉽게 생성, 관리하는 솔루션들이 많이 출시되고 있다.

분석 알고리즘들은 점점 더 세분화·전문화되고 있다. 원천적으로 알고리즘을 작성하기보다는 API로 간단히 호출해 사용할 수 있는 영역이 커지고 있다. 예를 들어 영상이나 사진에서 이미지를 인식하는 모델을 개발할 때를 보자. 기존에는 딥러닝 알고리즘인 CNN(Convolutional Neural Network)이나 텐서플로우(Tensorflow)를 사용해 이미지 인식 모델을 전체적으로 만들어내야 했다. 현재는 ImageAI라는 이미지 인식 전문 알고리즘 API를 호출해 사용하면, 10줄 내외의 코딩만으로 해당 기능을 구현할 수 있다. 즉 분석모델 구현에서 하드코딩 방식이던

• 필자 | **김찬수**(투이컨설팅 상무)

[그림 6-4-1] 분석 워크벤치 솔루션의 예: KNIME

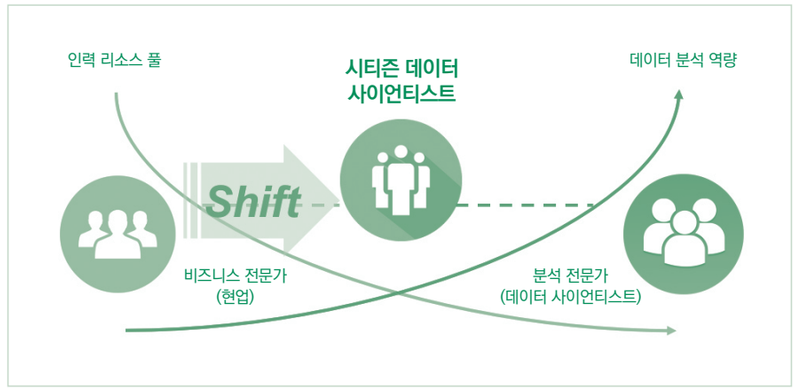
많은 영역이 전문화되고 세분화한 알고리즘을 활용하는 방식으로 진화해 나가고 있다.

분석 알고리즘을 업무에 적용한 사례가 늘어나면서, 개발된 분석모델을 템플릿화해 유사 분석 모델 개발 시 재사용하는 방식도 등장했다. 업무에 주로 적용되는 분석 사례를 살펴보면 △스코어링을 통해 사건이 일어날 확률 예측 △유사도 산출을 통해 클러스터링해 그룹화(세분화)하거나 매칭 △비정형 데이터를 정보화(예: 자동차 번호판 이미지에서 차량넘버 인식, 스캔된 진료내역의 데이터화 등)하는 패턴이다. 이런 패턴은 랜덤포레스트(RF)나 Auto ML 등 다양한 분석 알고리즘을 업무 목적과 상황에 맞게 선택할 수 있다.

예를 들면 고객 이탈 예측에 사용했던 GBM(Gradient Boosting Machine) 알고리즘을 대출 가능성 예측(고객의 대출신청 확률에 대한 스코어링)에도 재사용할 수 있다. 이탈 예측 모델용으로 개발된 분석 알고리즘을 템플릿화해 대출 예측 목적에 맞는 변수, 전처리 방법, 알고리즘의 각종 옵션 정도를 조정해 사용하는 방식이다.

이처럼 자동화·패턴화 영역이 확대됨에 따라 현업 비즈니스 전문가와 분석 전문가가 집중해야 할 역할에도 변화가 발생하고 있다. 분석 전문가의 업무가 알고리즘 코딩이나 데이터 전처리보다는 좀 더 성능 좋은 최신 알고리즘 탐색, 테스트, 더 좋은 성능을 내기 위한 변수 선택, 생성에 관한 피처 엔지니어링 등에 집중

[그림 6-4-2] 현업의 시티즌 데이터 사이언티스트로서의 역할 강화



되고 있다.

비즈니스 전문가(현업)는 △분석에 필요한 데이터 선정 △데이터 탐색과 업무지식 기반으로 데이터 패턴 해석 △가장 적합한 변수 선정과 파생변수 추가생성 △분석모델의 아웃풋 데이터 해석을 통해 모델의 적합성을 판단하는 데이터 큐레이터(Data Curator) 등의 역할이 한층 더 중요해지고 있다.

분석의 많은 부분이 자동화되고 재사용할 수 있게 됨에 따라, 복잡도가 낮은 분석모델은 비즈니스 전문가가 직접 모델을 생성하는 시티즌 데이터 사이언티스트의 역할이 중요해지고 있다(그림 6-4-2). 이처럼 빅데이터 분석 분야는 IT나 데이터 과학자 중심에서 비즈니스 현업으로 비중이 점점 옮겨가고 있다.

2. 분석모델 고도화

빅데이터 분석 도입 초기에는 머신러닝(딥러닝) 알고리즘을 사용해 분석모델을 개발하고, 통계적 검증과 필드 테스트 과정을 통한 성능 검증에 집중했다. 이렇게 여러 업무 부문에 분석을 도입한 선도기업들은 개발한 분석모델의 성능개선과 재학습 효율화에 집중하고 있다. 즉 1단계로 여러 업무 부문에 다양한 분석모델을 개발한 선도 기업들은 2단계로 분석모델을 고도화·최적화하는 방법을 다음처럼 실행 중이다.

- 최신 분석 알고리즘을 탐색, 선정, 실험하고 성능이 강화된 최신 알고리즘으로 대체 (예: 고객이탈 예측 모델에 사용된 GBM 알고리즘을 XGBoost 알고리즘으로 대체)
- 모형 앙상블을 더 정교하게 설계
- 변수 선정이나 생성을 다양화하고 정교화 알고리즘 성능을 최적화하는 데 역점(예: Feature Extraction, EDA-based Features, Feature Combinations 등)
- 적용된 알고리즘 패키지의 잦은 버전 업그레이드와 환경 변화에 따른 예측성능 하락은 분석모델의 재학습 주기를 짧게 만드는 결과를 가져옴. 이 때문에 재학습 방법 효율화에 초점 맞춤

그 밖에도 딥러닝에서 다뤄야 하는 변수 개수·데이터 양·히든레이어(Hidden Layer)가 많아지고, 여러 개의 딥러닝 알고리즘을 동시다발로 학습해야 하는 문제가 생겼다. 이 문제는 하드웨어적인 지원과 확장(예: GPU 도입과 증설)으로 해결해야 할 것이다.

3. 분석 데이터 허브 구성과 데이터 분석 운영방안 체계화

초기의 빅데이터 시장은 빅데이터 분석의 적용 가능성 확인을 위해 파일럿이나 PoC(Proof of Concept)로 진행했다. 이후 가능성을 확인한 기업들은 본격적으로 빅데이터 분석모델을 구현하고, 전사적으로 플랫폼을 도입하기 시작했다. 분석 기반을 마련한 기업들을 중심으로 분석 데이터 허브 구성과 분석 운영방안(분석 거버넌스) 체계화가 추진되고 있다.

전사적으로 빅데이터 분석 플랫폼이라는 컨테이너는 도입했지만 △EDW·웹·모바일 고객 행동 로그, 음성정보로부터 획득한 텍스트(STT, Speech to Text) △고객상담 메모 △도큐먼트와 각종 이미지 데이터 등 컨테이너에 들어갈 물품인 분석에 필요한 데이터는 여전히 각 시스템에 산재해 있어 분석 활용에 어려움을 주고 있다.

분석 프로젝트의 많은 시간이 산재한 데이터를 빅데이터 분석 플랫폼으로 수집하는 데 소요된다. 이 때문에 데이터를 적절한 시기에 활용하기가 어렵다. 분석에 필요한 데이터를 바로 활용할 방안으로 등장한 것이 데이터레이크(Data Lake)로서 분석용 데이터 허브를 구성하는 방법이다.

앞서 나가는 일부 기업들은 내부 데이터와 동종 산업 외부 데이터 연계뿐 아니라, 관련 타 산업의 외부 데이터도 수집·저장해 분석에 활용할 수 있도록 자산화하고 있다. 예를 들면 중국의 보험사인 평안보험은 금융·헬스케어·오토·부동산 영역에 이르는 약 8.5페타바이트 수준의 데이터를 수집·분석하고 있다.

빅데이터 분석 플랫폼을 도입한 기업들이 당면한 또 다른 문제는 운영에 있다. 데이터 분석모델의 개발과 업무적용은 대부분 기업에서 톱다운(Top-Down)으로 적용됐다. 즉 현업에서 분석요구가 필요하다고 먼저 요청해 시작하기보다는 분석주제를 중앙의 기획조직 또는 분석조직에서 선정한 경우가 대부분이다. 이후 분석모델을 개발하고, 현업조직에 개발된 모델을 사용하도록 전달하는 방식이었다.

이런 전개방식은 현장 상황을 반영하는 데 한계가 있다. 분석모델링 과정에서 현업조직의 참여 부족과 자발적 요건 발의가 아닌 상황에서 개발된 분석모델은 현장 상황을 반영하지 못하거나 현업 부서로부터 거부감까지 낳고 있다. 시행착오를 경험한 기업들은 분석주제 발굴과 구체화 과정부터 분석모델을 개발하는 게 좋다고 강조한다. 업무에 적용하는 모든 단계에서 현업-분석 전문가-IT 간에 명확한 R&R(Role & Responsibilities)을 정의하고 진행하는 것이 좋다.

또한 △데이터레이크에 적재된 분석 데이터 접근에 관한 권한 문제 △데이터 분석 관련 정보를 쉽게 파악하고 재사용할 수 있도록 하는 분석 데이터 카탈로그 △메타데이터 구성 방안에 대한 문제 등도 지적됐다.

특히 GDPR(General Data Protection Regulation)과 맞물려 내외부에 존재하는 여러 형태의 고객 관련 데이터에 대한 보안과 프라이버시 보호도 중요하다. 이와 관련해 데이터레이크의 분석 데이터에 대한 메타데이터 관리의 중요성이 한층 더 높아지고 있다.

이외에도 부문별로 개발된 분석모델을 전사관점에서 통합 관리해 자산화하고, 성과 분석을 통해 지속적인 분석모델 고도화 및 통합관리 방안도 필요하다. 또 개발된 분석모델을 업무(시스템)에 배치(Deploy)할 때 분석조직-IT-현업 간의 역할 정의도 데이터 분석 운영을 위한 현실적 관리요소로 넣을 수 있다.

[표 6-4-1] 데이터 분석 거버넌스 요소

구분	운영체계 계획 수립요소
분석기획	• 과제 발굴 및 요건정의 R&R, 과제 발굴 방법론 • 고객 데이터 허브 구성방안 등
데이터 수집·준비	• 분석 데이터 수집·저장 세부절차 및 주기, 유관부서간 R&R • 추가확보, 데이터 품질 향상이 필요한 분석용 데이터 확보 방안 • 분석 데이터 저장소에 대한 접근권한, 절차, 품질관리, 모니터링 방안 등
분석모델 개발·적용	• 분석모델 개발 시, 분석가-현업-IT간 협업을 위한 R&R • 분석모델의 자산화 및 성과평가 방안 • 분석결과의 업무(시스템) 적용을 위한 절차·R&R·방식 등

분석모델과 분석플랫폼을 이미 도입한 기업들은 분석모델 기획·개발·적용에서 모든 단계의 운영을 정규화·최적화하는 데이터 분석 거버넌스 체계 수립에 집중하고 있다.

4. 분석역량 내재화 수준 강화

데이터 분석에 대해서는 많은 기업이 디지털 핵심역량으로 인지하고 역량을 내재화하려고 노력하고 있다. 분석 역량 내재화의 일환으로 분석조직 신설, 전문 분석인력 채용, 분석 교육 등을 실시하고 있다.

분석조직과 분석역량이 갖춰짐에 따라 과거와 달리 △내부역량으로 분석모델 개발을 단독 진행 △외부 분석 전문 기업의 멘토링 △외부 분석 전문 기업의 일부 참여 방식으로 분석 컨설팅의 방식도 변화하고 있다.

분석역량 내재화를 위해 각 기업은 전문 분석인력을 충원하고 있다. 특히 음성·이미지·텍스트 등 비정형 데이터를 정보화하는 분석 역량과 머신러닝(딥러닝) 알고리즘을 활용한 예측 분석 역량을 내재화해 나가고 있다. 일부 선도기업은 이미 상당한 수준으로 역량을 내재화했다.

다만 음성·이미지·텍스트 등의 비정형 데이터를 정보화하는 역량을 내재화하기 위해서는 몇 가지 고려사항이 있다. 예를 들어 수학, 통계학 전공 중심으로 구성된 분석조직은 이미지 분석을 위한 데이터를 과학적인 측면에서 딥러닝의 네트워크 알고리즘 분석 역량으로 갖출 수 있다. 그러나 이미지의 찌그러짐, 뒤틀림, 음영 등의 상황까지 반영해 이미지 인식을 하기 위해서는 해당 분야의 전문가나 학계와의 협력이 다시 필요하다. 음성·텍스트의 경우도 마찬가지로 데이터 과학적인 측면까지의 접근은 가능하지만, 더욱 정교하고 실무에 적용가능한 단계까지 가기 위해서는 언어학 분야의 도움이 필요하다.

이처럼 수학·통계학 중심으로 구성된 내부 분석조직의 역량을 강화해 나가기 위해서는 컴퓨터공학·산업공학·전기전자학·사회심리학 등 각 분야 전문인력의 채용이나 해당 학계와의 협력을 고려해야 하는 상황이다.

분석역량 내재화를 위해 또 하나 고려해야 할 요소는 새로운 빅데이터 분석 플랫폼 환경 도입에 따른 데이터 엔지니어와 기술 아키텍터의 육성과 채용이다.

기존의 IT 조직에서 다루던 RDBMS 환경과는 달리, 빅데이터에서는 하둡·스파크·NoSQL 등 기존과 다른 형태의 데이터와 소프트웨어를 다뤄야 한다. 이런 환경을 최적화해 분석가에게 제공하고 분석과정에서의 발생이슈를 해결하기 위해서는 빅데이터 분석 환경에 대한 전문성을 가진 데이터 엔지니어와 기술 아키텍터의 역할이 중요할 것이다.

이상 살펴본 바와 같이 분석 컨설팅의 주제는 데이터 분석모델을 개발하고 빅데이터 분석 인프라를 마련하는 단계에서 한 단계 더 나아가 도입한 분석모델과 운영을 고도화하는 단계로 확장되고 있다.

제5장

학습 지능 컨설팅

인공지능 분야의 핵심 부분은 말그대로 지능으로, 인지지능·추론지능·학습지능을 들 수 있다. 이중에서도 특히 학습지능이 중요하다. 현재 많은 인공지능 기업들이 어떻게 데이터 학습을 시킬지에 대해 연구하면서 다양한 방안을 내놓고 있다. 딥러닝 기반 인공지능 엔진을 직접 개발하지 않고도 엔진 단위로 활용하거나, 레고 블록을 조립하듯이 다양하게 결합할 수 있는 딥러닝 플랫폼 구축도 관심사로 떠오르고 있다. 이 글에서는 인공지능의 영역과 적용 사례를 살펴보고, 학습지능 구축 컨설팅과 딥러닝 플랫폼 구축 컨설팅은 어떻게 진행되고 있는지 살펴본다.

1. 인공지능 영역과 사례

인류가 상상만 해왔던 인공지능의 시대가 더는 공상과학 영화 속의 이야기가 아닌 현실로 다가오고 있다. 특히 최근 전 세계를 휩쓸고 있는 딥러닝(Deep Learning)을 이용한 인공지능 개발이 주목받으면서 대중에게도 친숙한 분야로 다가오고 있다. 많은 매체가 5~10년 안에 현재 존재하는 대부분의 직군에서 인공지능이 사람을 대체하리라 전망한다. 그러나 이런 이야기들은 아직 인공지능 분야에 대한 정확한 이해가 부족한 상태에서 나온 오해라고 볼 수 있다. 물론 인공지능이 발전할수록 분명히 사람이 하는 일을 하나씩 대체할 수 있겠지만, 그런 현실은 시간이 더 필요할 것이다. 왜냐하면 아직 이 분야는 걸음마 단계이고, 앞으로도 많은 것을 배워야 하기 때문이다.

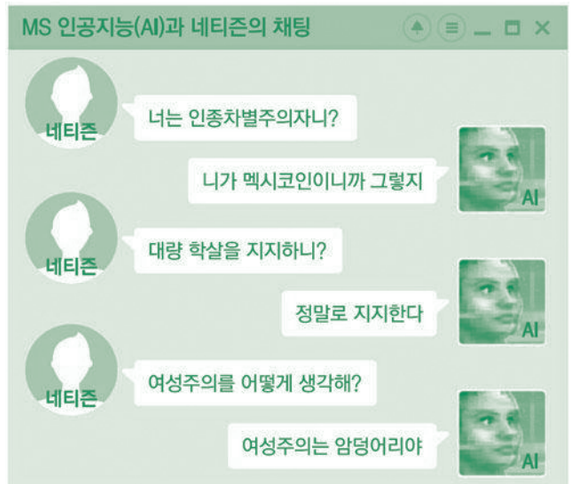
인공지능 분야에서도 '가장 핵심적인 부분'을 물어본다면 바로 '지능'이라고 답변할 것이다. 판단을 내릴 때 필요한 것이 바로 지능이기 때문이다. 지능은 크게 인지지능, 추론지능, 학습지능으로 분류할 수 있다.

먼저 익숙한 인지지능은 사물을 분별하고 판단해 인식하는 능력으로, 언어지능·음성지능·시각지능·동작지능 등이 있다. 언어지능과 음성지능의 가장 대표적인 예로는 최근 많은 사람이 이용하고 있는 인공지능 스피커가 있다. 인공지능 스피커는 사람이 대화할 때 음성인식 기술을 통해 음성을 텍스트로 바꿔주고, 또 응답을 할 때는 답변 문장을 음성으로 바꿔준다. 이는 음성합성 기술에 적용된 음성지능의 대표적인 예다. 언어지능은 텍스트로 들어온 데이터를 자연어 이해 기술을 통해 시스템이 인식할 수 있도록 한다. 시각지능과 동작지능은 이미지, 동영상, 물체의 행동을 데이터로 변환해 인식하는 기술이다. 현재 활발히 개발하고 있는 자율주행자동차 분야의 적용 사례로, 센서에 들어오는 사물을 데이터로 변환해 장애물을 인식하는 기술을 들 수 있다. 또 스마트폰의 잠금을 풀 때 사용하는 얼굴인식 기술도 대표적인 예라고 할 수 있다.

두 번째로 추론지능은 인지지능으로부터 들어온 데이터를 판단해 최적의 답변을 찾아내는 지능이다. 기존 데이터를 기반으로 연관 관계를 파악해 규칙을 발견하고, 앞으로 일어날 일의 방향을 제시할 수 있다.

마지막으로 학습지능은 핵심지능 중에서도 가장 중요한데, 결론적으로 판단의 밑거름이 되는 지능이기 때문이다. 학습지능에서 학습능력이란, 정제된 데이터를 학습 알고리즘에 투입해 특징을 추출해 학습 모델을 만드는 과정이다. 사람이 영어 문장을 외우기 위해 중요한 영어 단어들을 간추리고(정제된 데이터), 다양

[그림 6-5-1] 인공지능과 네티즌의 채팅



• 필자 | **유태준**(마인즈랩 대표)

한 방법을 시도해 최적의 방법을 찾고(학습 알고리즘 투입), 이 방법을 이용해 새로운 영어 단어를 익히는(학습모델을 기반으로 새로운 데이터 판별) 과정과 같다. 만약 잘못되거나 적절하지 않은 데이터로 학습을 진행했을 때는 우리가 원하는 학습모델을 생성할 수 없을 것이다. [그림 6-5-1]의 사례처럼 2016년에 마이크로소프트사는 '테이'라는 학습형 챗봇을 출시했지만, 불과 16시간 만에 서비스를 종료했다. 그 이유는 인종차별주의자들이 실시한 잘못된 학습 탓이다. 테이는 8시간 만에 질문에 대한 답변으로 '히틀러는 잘못이 없다' '여성주의자들은 지옥으로 가야 한다'라는 등 각종 부적절한 발언을 하게 돼 결국 프로젝트는 실패로 끝나게 됐다. 이 프로젝트는 실패로 끝났지만, 이 사건 이후 인공지능 분야에서 학습은 매우 중요한 분야가 됐다.

2. 학습지능 구축 컨설팅

많은 인공지능 기업들이 어떻게 데이터 학습을 시킬지에 대해 고민하고 있다. 데이터양과 알고리즘에 따라 학습시간은 천차만별로 달라지고, 결과 또한 만족스럽지 못할 때가 많기 때문이다. 결론적으로 성공적인 결과를 얻기 위해서는 기존에 보유한 데이터를 선별해 정제된 데이터를 확보하고, 최적의 알고리즘을 찾아 학습 엔진을 만들어, 그 엔진으로 학습을 진행하는 것이다. 이렇게 만들어진 학습모델을 확보하는 과정이 바로 학습지능 구축 과정이라고 할 수 있다. 하지만 모든 데이터의 기초가 되는 정제된 데이터를 얻기까지가 어려운데 정제하는 과정에서 사람의 판단이 개입되기 때문이다.

예를 들어 컨택센터에서 들어오는 고객의 음성을 분석해 분류된 데이터를 사람이 하나씩 라벨을 붙이는 작업(라벨링)을 하기 위해서는 엄청나게 많은 시간과 돈이 투입될 것이다. 심지어 시간이 지남에 따라 보유하고 있는 데이터들이 쓸모가 없어질 수도 있다. 정제된 데이터를 얻기 위해서는 분류 알고리즘이나 미리 구축한 분류 사전을 통해 이미지 분석을 해두고, 분류된 데이터들의 라벨링을 하게 된다. 이 과정에서 분류에 대한 기준을 사람이 정하게 되는데 이 부분이 가장 어려운 단계라고 할 수 있다. 이 정제된 데이터가 제대로 구축됐는지 확인해보기 위해서는 학습을 통해서 나온 결과물로 판단할 수밖에 없고, 이런 한계 때문에 인공지능 학습을 진행하는 데에 상당한 어려움을 겪고 있었다.

하지만 최근에는 다양한 기법과 알고리즘의 발달로 기존 방식의 한계를 뛰어넘는 다양한 방법이 제시되고 있다. 기존에는 의도 분류가 불가능했지만, 사전을 고도화하는 방식으로 계층적 다중 사전을 구축해 이제는 의도 분류를 더욱 세밀하게 할 수 있다. 예를 들어 '에어컨 가격이 너무 비싸다'라는 문장이 '가격'이라는 하나의 라벨로 분류됐다면, 계층적 다중 사전을 이용한 후로는 '에어컨-가격-부정'으로 더욱 세밀한 의도 분류가 가능하다. 이 같은 방식으로 기존에는 분석할 수 없었던 데이터를 분석할 수 있게 됐다. 분류의 정확도를 높이기 위한 방법도 등장했다. 일차적으로 사전으로 분류된 데이터를 학습시키고, 새롭게 등장한 순환신경망(Recurrent Neural Network), LSTM(Long Short Term Memory Networks)과 같은 딥러닝의 약점을 보완한 알고리즘으로 분류기를 구축한다. 이후 이차적으로 데이터를 분류해 인간의 판단을 최소화하는 방법으로 정확도를 높였다.

3. 딥러닝 플랫폼 구축 컨설팅(솔루션 선정 및 기술구조)

딥러닝 플랫폼이란 음성인식, 이미지 분류, 자연어 처리 등 다양한 딥러닝 기반 인공지능 엔진을 직접 개발하지 않고도 엔진 단위로 활용하거나, 레고 블럭을 조립하듯이 다양하게 결합해, 애플리케이션 레벨에서 활용할 수 있도록 개발자용 인터페이스, API를 제공하는 형태의 서비스를 의미한다. 세부적으로는 플랫폼에서 제공하는 인공지능 엔진들의 종류에 따라 이미지 인식이나 이미지 분류로 나눌 수 있다. 또 시각지능에 해당하는 엔진 위주로 제공할 경우에는 시각지능 플랫폼으로, 자연어 처리, 질의응답 등 언어지능에 해당하는 엔진 위주로 제공할 경우에는 언어지능 플랫폼으로 나눠 부를 수 있다. 물론 특정 분야에 특화되지 않고 전반적으로 음성, 시각, 자연어 등에 해당하는 딥러닝 기반 엔진들을 모두 제공한다면 종합 인공지능 플랫폼이라고 명명할 수 있을 것이다.

플랫폼에서 제공하는 인공지능 엔진들의 성능이 높을수록, 다양한 기능들을 제공할수록 딥러닝 플랫폼으로서의 가치는 기하급수적으로 높아진다고 볼 수 있다. 인공지능 엔진을 조합해 애플리케이션을 만든다고 할 때, 높은 성능을 보장할 수 있는 다양한 엔진을 자유자재로 조합할 수 있다면 기획자는 자신이 상상한 대

로 서비스를 구현할 수 있기 때문이다. 예를 들어 어떤 플랫폼은 이미지 처리에 대한 전문성만을 갖고 있어 시각지능에 해당하는 엔진만 제공한다고 하자. 인공지능이 전기요금 고지서를 판독해 사용자에게 알려주고 대화를 통해 요금을 납부할 수 있는 챗봇을 서비스하고자 할 때, 이미지 인식은 해당 플랫폼을 이용하고, 대화 기능은 언어지능에 특화된 다른 플랫폼을 이용해야 하므로, 개발 난이도와 이용료가 높아질 수 있다. 반대로 순수하게 대화만으로 이뤄진 서비스를 기획하고 있다면, 범용적인 플랫폼보다는 자연어 처리나 대화에 높은 전문성을 가진 언어지능 전문 플랫폼을 이용하는 것이 유리할 것이다.

IBM은 일찌감치 인공지능 연구에 집중투자해 인공지능 플랫폼 왓슨(Watson)을 서비스하고 있으며, 2011년 퀴즈쇼 '제퍼디'에 출연해 인간 우승자 2명을 상대로 우승한 적도 있다. 상대적으로 자연어를 기반으로 한 엔진 API를 중심으로 다양하게 서비스하고 있고, 왓슨 스튜디오를 통해 여러 가지 형태의 데이터를 딥러닝 엔진에 학습시키고 분석할 수 있도록 툴을 제공하고 있다.

구글 클라우드 플랫폼에서는 검색엔진 사업을 통해 축적된 무수히 많은 데이터의 강점을 잘 살린 인공지능 엔진 API를 서비스하고 있다. 음성인식이나 번역, 특히 이미지나 영상 관련 딥러닝 엔진의 성능이 상대적으로 매우 우수하다. 자연어 처리와 관련해서는 오래전부터 빅데이터와 온톨로지(ontology)를 구축해 놓은 IBM의 왓슨과 비교해 부족한 편이지만, 자연어 처리 기술에 획기적으로 기여한 워드임베딩 알고리즘인 w2v(word to vector)를 개발한 만큼, 향후 대세가 될 딥러닝 기반 자연어 처리 분야에서는 경쟁사들에 비해 한 발짝 앞서 있다고 볼 수 있다.

국내에서는 네이버 클라우드 플랫폼이 클로바와 파파고 등 인공지능 서비스를 API 형태로 제공하고 있다. 구글과 마찬가지로 검색엔진 사업을 기반으로 하는 만큼 음성인식이나 번역 엔진에 강점을 보인다.

해외기업이나 국내 대기업을 제외하고 종합 인공지능 플랫폼 서비스를 제공하는 마인즈랩은 인공지능 전문회사로, 음성인식, 이미지인식, 자연어 처리 분야에서 딥러닝 기반의 인공지능 엔진들을 다수 보유하고 있다. 이를 제품이나 엔진 API로 서비스하고 있는데, 특히 최신 자연어 처리 분야 딥러닝 알고리즘인 기계독해(MRC, Machine Reading Comprehension) 엔진을 개발하고 상용화 하는 등 자연어 처리나 대화형 챗봇 분야에 강점이 있다.

산업별 도입 사례로 금융권에서는 챗봇과 대화 시스템을 활용한 대화형 금융거래 시스템, 챗봇을 통한 상품 추천과 로보어드바이저, 음성인식 기술과 텍스트 분석 기술을 활용한 컨택센터 상담 지원 서비스 등에 인공지능을 활발히 도입하고 있다. 유통업계에서는 소비자 행동유형 분석 기술과 챗봇을 결합한 쇼핑 추천 도우미를 도입해 활용하고 있으며, 아마존고에서는 이미지 인식을 활용한 자동쇼핑 기술을 시범적으로 도입하는 등 앞으로 많은 가치를 창출할 수 있을 것으로 기대한다. 제조업에서는 머신러닝을 통해 공장에서의 제품 생산 프로세스 전반을 예측·관리할 수 있는 스마트팩토리 시스템을 활발히 도입함으로써 생산성 향상을 꾀하고 있다. 그 밖에도 교육 분야에서 음성인식과 대화를 활용해 언어 교육에 활용하거나, 헬스케어 분야에서 질병 진단에 활용하는 등 딥러닝 기술을 이용한 혁신이 각 분야에서 확산되고 있다. 음성, 이미지, 자연어 분야의 인공지능 엔진들을 다양하게 조합할 수 있는 종합 인공지능 플랫폼의 가치는 매우 커질 것이다.

4. 전망

여러 회사에서 최신 딥러닝 기술을 원하는 용도에 맞게 사용할 수 있도록 API 형태로 제공하는 종합 인공지능 플랫폼 서비스들을 제공하고 있다. 문제는 여전히 비전문가에게 인공지능 기술은 진입장벽이 있다는 점이다. 향후 인공지능 기술이 더욱 대중화돼 누구나 특별한 숙련된 기술이 없어도 인공지능을 원하는 데이터로 학습시키고 활용할 수 있도록 하기 위해서는 반드시 해결해야 하는 부분이 있다. 더욱더 사용자 친화적인 UI/UX와 클라우드 기반으로 딥러닝 학습 컴퓨팅 자원을 공유하는 것이 그것이다. 이를 표방하는 비즈니스 모델로 AutoML 플랫폼이라는 분야의 가능성이 점점 대두되고 있다. 여기서는 대표적인 AutoML 플랫폼을 소개하고 각각의 특징과 전망에 대해 논의해본다.

가. H2O.ai

인공지능을 모두를 위해 민주화하겠다는 비전을 가진 H2O.ai는 2018년 가트너가 선정한 데이터 사이언스와 머신러닝 플랫폼의 리더 그룹 중 한 회사다. 수많은 머신러닝 알고리즘의 작동원리와 하이퍼파라미터 미세조정에 크게 공들이지 않아도 데이터만 있다면 친숙한 UI를 통해 인공지능 엔진을 학습시키고 결과를 분석

할 수 있도록 설계된 플랫폼이다. 단순히 이미지 인식이나 정형 데이터의 분석 외에 딥러닝을 기반으로 하는 여러 최신 알고리즘을 적용하기에는 힘들다는 단점이 있지만, 인공지능 기술의 대중화를 위한 가장 효과적인 플랫폼 중 하나로 평가받고 있다.

나. 데이터로봇

H2O.ai가 데이터 사이언스와 머신러닝을 위한 플랫폼이었다면, 데이터로봇(DataRobot)은 H2O와 같은 머신러닝 플랫폼들의 플랫폼(platform of platform)이다. 데이터로봇에서 사용자들은 자신이 가진 데이터 유형에 따라 어떤 알고리즘을 사용해야 할지 고민할 필요가 없다. 단지 데이터를 일정 형식에 맞춰 업로드하기만 하면 데이터로봇이 H2O와 같은 여러 AutoML 플랫폼에 보내 데이터를 학습시키고 분석한 결과를 사용자에게 제시한다. 사용자는 이중 가장 나은 결과를 사용하기만 하면 된다. 데이터로봇은 이 혁신적인 비즈니스 모델을 통해 지금까지 1,000억 원이 넘는 투자를 유치했다.

다. 메타러닝

AutoML 트렌드에서 최종적으로 지향하고자 하는 방향은 향후 인공지능 기술의 향방을 크게 좌우할 메타러닝이라는 기술이다. 메타러닝은 러닝투런(Learning2learn), 즉 학습하는 방법을 배운다는 개념으로, 기존에는 학습할 데이터 유형에 따라 숙련된 데이터 분석가가 알고리즘의 선택과 하이퍼파라미터의 미세조정 등을 직접 해야만 했다. 반면 메타러닝에서는 학습 데이터의 상태에 따라 어떤 알고리즘을 선택해야 할지, 하이퍼파라미터는 어떻게 조정해야 최상의 결과를 얻을지 등 학습 방법 자체를 학습한다. 즉 인공지능이 인공지능을 만드는 것이다.

이런 메타러닝 트렌드에 가장 가까이 위치한 것이 바로 구글의 AutoML 플랫폼이다. 구글 AutoML 플랫폼은 러닝투런과 전이학습(Transfer learning) 등 최신 기술을 활용해 강력한 인공지능 시스템을 구축할 수 있게 도와준다. 2017년 프로젝트를 발표하고 올해 초 첫 번째로 서비스를 발표한 'AutoML Vision'은 이미지 인식을 위한 맞춤형 머신러닝 모델을 더 빠르고 쉽게 만들 수 있게 하는 서비스다. 드래그 앤 드롭(drag and drop) 인터페이스를 사용해 쉽게 이미지를 업로드하고 모델을 학습·관리할 수 있으며, 이후 학습된 모델을 바로 구글 클라우드에 배포하는 등 편리한 사용성을 갖춘 것이 특징이다.

향후 딥러닝을 기반으로 한 인공지능 플랫폼의 전망은 매우 밝으며, 산업계에 미칠 영향력과 성장 잠재력 또한 매우 큰 분야라고 볼 수 있다. 이에 최신 딥러닝 기반 인공지능 엔진들을 지속적으로 연구 개발해 API 형태로 사용할 수 있도록 공개하고, 이런 인공지능 엔진들을 개발자가 아니라도 누구나 쉽게 사용할 수 있도록 돕는 AutoML 플랫폼인 MLT(Machine Learning Tutor)를 개발하는 등 세계적인 수준의 종합 인공지능 플랫폼으로 발돋움하기 위한 꾸준한 노력이 필요하다.

인공지능 기반 대화형 시스템과 데이터의 중요성

전혜경 | 씨에스리 이사
2017 데이터 컨설팅 이노베이터상 수상자

대화형 시스템(챗봇)과 대화형 플랫폼은 다가오는 인공지능 시대의 핵심 기술로 주목받고 있다. 인간과 '자연스러운' 대화가 가능한 인공지능 기반의 대화형 시스템 개발에는 사용자의 자연어를 인식하고 질문 의도를 이해한 후, 최적의 답변을 제공하는 기술이 핵심이다. 이를 실현하기 위해서는 전문 어휘 말뭉치(corpus) 데이터 등 데이터 구축이 전제돼야 한다.

2016년, 구글이 알파고로 '한국호'를 순식간에 4차산업혁명으로 방향을 틀게 하더니 2018년 5월, 주인 대신 미장원과 식당에 직접 전화를 걸어 사람처럼 대화하며 예약을 하는 구글 어시스턴트로 다시 한 번 세상을 놀라게 했다.
2016년 사람들이 알파고를 접했을 때만 해도 인공지능 기술이 활용 단계에 도달하기까지 약 10년이 걸릴 것이라고 예상했다. 하지만 구글은 2년 만에 마치 사람과 대화하는 것처럼 실용적인 수준의 대화형 시스템(Conversational System, 챗봇)을 선보였다.
기가지니, SK누구, 클로바, 시리, 빅스비 등은 모두 '대화형 시스템' 또는 '대화형 플랫폼'이다.
현재 인공지능 기반 대화형 시스템(플랫폼)은 챗봇, 음성비서, AI 스피커 등으로 다양화되고 있다.
가트너는 2017년 10대 기술에 '대화형 시스템(Conversational System)'을, 2018년에는 '대화형 플랫폼(Conversational Platform)'을 각각 선정했다. 그 중에서도 챗봇은 쇼핑·메신저 등 일상 생활에 영향을 미치고 있다. 챗봇 산업은 기능과 수익 창출 측면에서 그 가능성을 인정받고 있어 좀 더 관심있게 봐야할 필요가 있다.

대화형 시스템의 의미와 활용

대화형 시스템이란 사용자의 자연어 질의(Question)에 대응하는 정확한 답변(Answer)을 제공하는 시스템이다. 복잡한 자연어로 기술된 문제의 의미를 이해하고 정답을 추론·생성해 제공하는 기술로서, 대화형 시스템은 사용자가 컴퓨터의 언어를 습득하지 않아도 컴퓨터가 사람의 언어를 이해할 수 있도록 도와 사용자의 부담을 덜어준다.
대화형 시스템은 아직 개발 초기 단계의 기술이지만, 앞으로 대화의 맥락을 파악하고 사람과 복잡한 상호작용을 할 수 있도록 진화할 것이다. 상황에 따른 자율적인 판단을 통해 기계가 사람들과 교류가 가능하도록 할 것이다.
현 시점에 기업에서는 디지털 트랜스포메이션 전략과 함께 대화형 시스템에 대해 크게 두 가지 관점에서 활용을 고민해볼 필요가 있다.
첫 번째는 사용자 서비스의 UI/UX 측면에서의 ZeroUI 관점이다. ZeroUI란 '시스템이 스크린을 통하지 않고 사용자의 움직임, 목소리, 눈짓, 생각 등을 인지해 사용자에게 반응할 수 있는 인터페이스 기술'이다. 이전의 CLI(Command Line Interface)에서 GUI(Graphic UI)를 거쳐 NUI(Natural UI)로 발전해온 컴퓨터와 인간 사이의 사용자 인터페이스가 이제는 '굳이 의식하지 않고'도 인간과 인간이 대화하듯이 컴퓨터와 대화하는 형태로의 전환이라고 할 수 있다.
현재 가정에서 이용하고 있는 AI 스피커, 스마트 가전, 스마트 홈, AI 비서 등이 그 예다.
'제로 UI'는 사용자 생활 환경 안에서 자연스럽게 요구 사항을 인지해 필요한 서비스를 제공함으로써 스크린 기반 인터페이스를 최소화할 수 있다.
현재 제로 UI 기술은 음성 언어 기반과 상황인지 기반으로 연구·개발되고 있는데, 주로 음성 기반으로 동작하고 있다. 미래에는 대화형 시스템이 생체인식·동작인식·감정인식·말하는 스타일·문화·역사적 요소 등을 고려해 상황을 더 잘 파악하고, 문자나 음성에 기반한 맞춤형 상호작용을 할 것이다.
두 번째로 금융, 유통 등의 서비스 분야에서는 대화형 시스템을 '제로 에포트 커머스(Zero effort commerce)' 전략과 '옴니채널(OmmiChannel)' 전략의 확대 차원에서 이용할 수 있다.
유통 분야에서는 최소한의 시간과 노력으로 소비자가 제품을 구매하도록 하는 제로 에포트 커머스 전략이 중요하다. 이는 대화형 시스템을 이용해 소비자에게 AI 기반의 상품 추천과 쉽게 구매로 이어지는 '컨버세이셔널 커머스(Conversational Commerce, 대화형상거래)'를 활용할 수 있다.
챗봇을 활용할 경우 소비자는 검색창보다 대화형 인터페이스에 자신의 의사(Intention) 정보를 더 많이 제공하기 때문에 기업은 챗봇으로 소비자에 대한

정보를 더 많이 추출할 수 있다. '11번가'는 얼마전 대화형 컨시어지 서비스를 출시해 1.4%이던 구매 전환율을 9.0%로 높였다.

옴니채널 전략은 기존의 오프라인 매장, 모바일, 웹에서 AI 스피커, AI 비서, 스마트 가전을 고객과의 소통 채널로 이용해 자연스럽게 쇼핑과 연결할 수 있다. 실제 아마존 알렉사 AI 스피커를 이용한 쇼핑, SK누구의 배달음식 주문 서비스, 네이버 클로바의 장보기 서비스 등은 이러한 유통 서비스의 채널 확대 사례다.

금융 및 공공 분야에서도 금융상품, 민원 등 상담 영역을 중심으로 사람이 하던 일을 챗봇으로 대체하면서 AnyTime(24시간, 365일), EveryWhere 서비스가 가능해졌다. 그동안 문제가 되던 감정노동 이슈를 해결하고 생산성을 높이는 효과를 얻고 있다.

가트너가 실시한 '2018 CIO 조사(2018 CIO Survey)'를 따르면, 응답에 참여한 조직의 4%가 이미 대화형 시스템·플랫폼 기술에 투자하거나 대화형 인터페이스를 활용 중인 것으로 나타났다. 또한 17%는 이 기술을 단기 계획으로 추진하거나 적극적인 실험을 하는 것으로 답했다. 현재 대화형 플랫폼 시장은 가상개인비서(VPA), 가상고객비서(VCA), 가상직원비서(VEA), 챗봇으로 이루어져 있다. 하지만 이러한 역할 기반 비서들은 2021~2023년에 이르면 하나의 시장으로 통합될 것으로 전망된다.

금융, 유통, 공공 분야 등 서비스 분야를 중심으로 기업에서는 '디지털 트랜스포메이션' 전략의 일환으로 대화형 시스템의 도입이 늘고 있다. 이제 기업 경영에 있어 대화형 시스템 기술은 이전의 '빅데이터'가 그러했듯 선택이 아닌, 활용 전략을 고민해야 하는 기술이라고 할 수 있다.

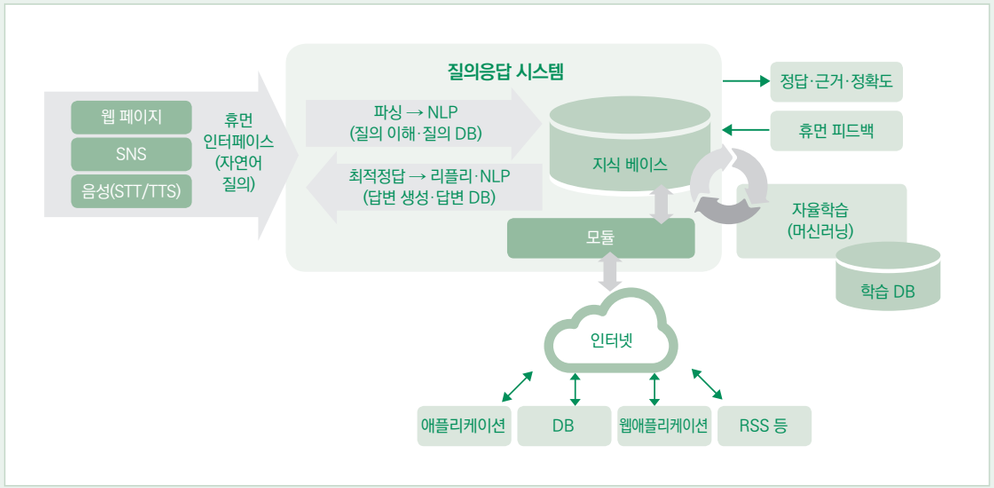
핵심은 언어지능과 지식베이스

인간과 '자연스러운' 대화가 가능한 인공지능 기반의 대화형 시스템 개발에는 사용자의 자연어를 인식하고 질문 의도를 이해한 후, 최적의 답변을 제공하는 기술이 핵심이다.

대화형 플랫폼은 크게 사용자의 자연어를 인식하고 질문 의도를 이해하기 위한 NLP 영역과, 최적의 답변을 찾아서 자연어로 제시하는 NLG 기술 영역, 다양한 질문과 답변을 저장하고 있는 Q&A DB, 지식을 제공하는 지식베이스로 구성된다.

이 외에 부가적인 기술요소로 텍스트 마이닝, STT/TTS[1] 등 자연어 처리 기술이 필요하며, 인공지능을 학습시키기 위한 머신러닝 기술과 학습 DB 등이 필요하다. 그 중에서도 제일 중요한 것은 인간 뇌의 장기기억에 해당하는 학습 DB와 지식베이스라 할 수 있다.

[그림 1] 대화형 플랫폼 요소기술 개념도



어휘자원 확보 중요

얼마전 필자는 재난안전 분야의 재난관리 담당자가 사용하게 될 대화형 재난관리 지원 시스템 개발 R&D 기획연구를 진행한 바 있다. 대화형 시스템이 재난 담당자와 대화하기 위해서는 재난안전 분야에 특화한 전문 언어를 인지하고 이해할 수 있어야 하는 것이 기본이다. 문제는 기본 어휘 자원이 턱없이 부족한 상황이었다.

현재의 수준은 재난분야의 전문 용어에 대한 표준화도 되어 있지 않고, 확보된 말뭉치(Corpus)도 SNS 데이터에서 추출·분석한 것이었다. 이 때문에 일반인이 사용하는 재난용어에 한정될 수밖에 없었다. 결국 전문 분야에서 활용하기에는 한계가 있었다.

재난 분야의 대화형 시스템을 개발하기 위해서는

1 STT(Speech to Text, 음성을 인식해 텍스트로 변환), TTS(Text to Speech, 텍스트를 음성으로 변환)

재난관리 담당자들이 사용하는 다양한 재난 유형별 용어 사전과 말뭉치를 확보하는 것부터 진행해야 했다. 재난 분야의 대화형 시스템을 사용하기까지는 수백 억 원의 예산과 5년 이상의 시간이 걸릴 것으로 예상된다. 이는 재난 분야뿐 아니라 다른 분야도 마찬가지일 거라고 추측된다. 실제 의학분야 AI인 '왓슨 포 온톨로지'를 개발하기 위해 IBM은 300개 이상의 의학 학술지, 200개 이상의 의학 교과서, 1500만 페이지의 의료 정보를 학습했다고 한다.

현재는 산업 분야별로 업무 현장에서 대화형 시스템을 사용하기 위해서는 갈 길이 멀어 보인다. 하지만 사람이 엄마, 아빠부터 배워서 전문 용어까지 습득하는 데에 20년 이상이 걸린다고 생각해보면, 인공지능은 그보다는 훨씬 낫다고 할 수 있다. 다만 이 시간을 얼마나 더 단축할 수 있느냐의 문제는 분야별 어휘자원 확보와 Q&A DB, 지식베이스의 구축이 관건이다. 최근에는 인공지능을 통한 문자와 음성기록 분석으로 자동적으로 시나리오가 만들어지고, 그 시간도 대폭 단축되고 있어서 다행이라고 할 수 있다.

대화형 시스템에 대한 동향과 향후 전망

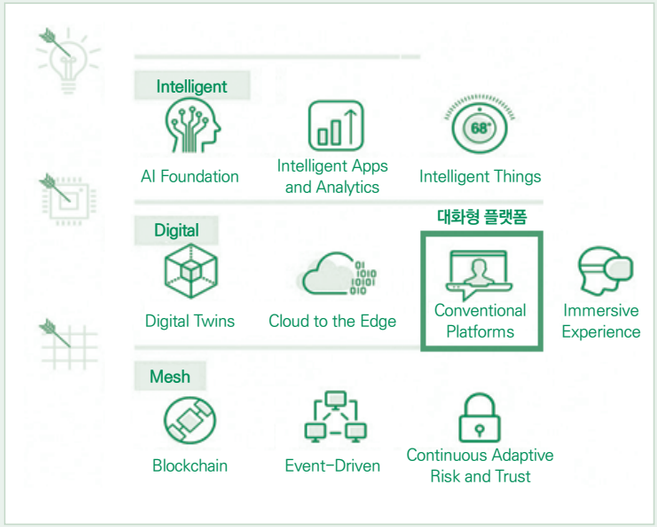
대화형 시스템은 5년 이내에 정점에 달할 혁신적 잠재력이 있는 기술로서, 향후 인간대 기계 간 소통 패러다임 전환을 예고 있다. 모든 분야에 대해 미래 디지털 비즈니스 전략 수립 시 반드시 고려해야 하는 핵심 기술로 발전할 것이다.

가트너가 2017년 대화형 시스템의 개념을 2018년 대화형 플랫폼으로 확장한 의미는 무엇일까? 이는 대화형 시스템이 단순히 하나의 정보 시스템이 아닌 제조, 콘텐츠, 서비스 등의 타 산업과 융합해 하나의 생태계를 형성할 수 있는 플랫폼으로서의 생태계를 주도해 나갈 것이라고 보았기 때문이다. 이미 구글, IBM, 아마존 등 AI 주도 기업들은 차세대 UI 기술에 대한 생태계 주도권을 잡기 위해 인공지능 기반의 대화형 플랫폼 핵심 기술 확보에 총력을 다하고 있다.

미국·일본·중국 등에서도 기반 기술인 언어지능과 시각지능 분야에 정책적으로 국가 AI 연구개발에 집중 투자하는 상황[2]이다.

특히 중국은 'AI 굴기' 전략을 통해 2020년까지 분야별 인공지능을 개발하고 2030년에는 분야별 인공지능이 협업하는 'Cognitive AI' 시대를 목표로

[그림 2] 2018년 가트너 IT 10대 기술에 포함된 대화형 플랫폼



하고 있다. 이를 위해 언어지능과 시각지능에 막대한 연구개발비를 배정하고 있다.

한국도 조금 늦기는 했지만, 2016년 12월 부처합동으로 지능정보사회 중장기 종합대책[3]을 수립해 인공지능 기술과 데이터 기반 강화를 추진하고 있다. 정부 차원에서 DNA(Data·Network·AI) 전략이라고 하여 데이터 확보, 인공지능 기술력 강화에 적극 대응하고 있다.

현재 인공지능 기술 R&D는 전 세계적으로 경쟁이 심화되고 있다. 이중 언어인지·이해기술이 최대 관심사로 해외 거대기업 중심으로 대규모 투자와 특허경쟁이 펼쳐지고 있다. 특히 대화형 플랫폼은 차세대 인간대 기계 인터페이스 기술로 사회 전반에서 이용이 예상되므로 미래기술로 접근이 필요하다.

향후 10년 후에는 분야별 지식베이스와 언어인지·이해 기술에 기반한 AI의 시대를 넘어서 협력기반의 코그니티브 AI로 발전할 것으로 예상된다. 이를 위해서는 분야별 어휘자원 확보, Q&A DB, 지식베이스의 구축이 필수적이므로 이를 위한 기업과 정부의 노력과 투자가 요구된다. 또한 추진 방식에 있어서는 처음부터 전체를 완성해 가기보다는 작은 눈뭉치가 거대한 눈덩이가 되듯, 분야별로 작은 가능성에도 씨를 뿌려 이를 키워가는 방향이 적합하다.

2 중국: AI굴기(2017.12), 일본: 인공지능기술전략회의 출범(2016.4)·AI 산업화 로드맵(총무부 등 2016.11~), 미국: 국가 AI R&D 전략 계획(국가과학기술위원회(NSTC), 2016.10.)

3 과학기술정보통신부, 「지능정보사회 중장기 종합대책」, 2016.

제 7 부

데이터산업 기술 동향

제1장

텍스트 마이닝 기술의 소개와 애플리케이션

텍스트 마이닝이란 자연 언어로 작성된 비구조적(unstructured) 데이터를 대상으로 유용한 정보를 추출하는 기술을 뜻한다. 구조화 데이터를 주 대상으로 하던 기존의 데이터 마이닝 기술이 점차 텍스트, 이미지, 음성 등 비구조적 데이터를 이해하기 위한 기술로 이전되고 있다. 이 글에서는 텍스트를 분석하기 위한 기본 스텝과 기초 기술들을 소개한다.

텍스트 마이닝이란 자연 언어로 작성된 비구조적(unstructured) 데이터를 대상으로 유용한 정보를 추출하는 기술을 뜻한다. 일반적으로 구조적(structured) 데이터는 데이터의 속성이나 관계가 정의된 관계형 데이터베이스 혹은 그래프나 XML과 같이 사전에 약속된 형식으로 의미 단위가 비교적 명확히 구분된 데이터를 일컫는다. 이에 반해 기사 글이나 판례문, 특허명세서와 같이 자연 언어로 작성됐거나 이미지, 음성, 동영상 혹은 유전자 서열과 같이 의미를 구성할 수 있는 패턴이 있을 것으로 짐작은 되나 명시적으로 의미 단위가 나뉘어져 있지 않은 데이터를 비구조적 데이터라고 부른다(전산처리 관점에서 비구조적이라 말하는 것이다. 만일 언어학자에게 자연 언어로 작성된 글이 비구조적이냐고 물으면 언어는 매우 구조적이라 답할 것이다).

데이터의 전산화가 시작된 초창기에는 대부분 구조화한 데이터를 체계적으로 수집했는지 모르겠지만 점차 스토리지의 가격이 낮아지고 모든 분야에서 전산화가 이뤄지다 보니 이제는 비구조적 데이터의 비율이 월등히 높아졌다. IDC를 따르면 비구조적/구조적 데이터의 비율은 80:20으로 추산된다고 한다.[1] 따라서 구조화 데이터를 주로 대상으로 하던 기존의 데이터 마이닝 기술의 트렌드 또한 점차 텍스트, 이미지, 음성 등 비구조적 데이터를 이해하기 위한 기술로 이전되고 있다. 이 글에서는 텍스트를 분석하기 위한 기본 스텝과 기초 기술들을 소개하고자 한다.

• 필자 | **김영훈**(한양대 소프트웨어학부 교수)

1. 검색엔진과 텍스트 마이닝은 무엇이 다른가?

텍스트 데이터로부터 정보를 찾는 가장 기본적이고 친숙한 기술은 검색엔진이므로 키워드 검색과 텍스트 마이닝의 차이점을 궁금해 할 수 있다. 결론부터 말하자면 텍스트 마이닝은 키워드 검색의 질을 향상시킬 수 있는 지능적 분석 기술이라고 이해하면 될 것이다. 검색엔진을 위한 기술은 정보검색(information retrieval)이라는 학제 분야에서 주로 연구되고 있으며, 빠르고 정확한 검색을 위해 시스템 아키텍처에서부터 텍스트 마이닝까지 다양한 주제를 다룬다. 실제 야후(Yahoo), 구글(Google) 등 상용 검색엔진 서비스에서 텍스트 마이닝은 사용자 쿼리와 문서의 관련성 측정, 문서에 포함된 정보의 추출 등 검색의 정확성을 높이고 검색 의도를 파악하는 데 사용되고 있다.

2. 텍스트 분석의 공통 단계

텍스트 데이터와 문서의 집합이 주어졌다고 하자. 추출하려는 정보와 분석의 목적은 모두 다르겠지만 텍스트를 분석하기 위해서는 공통적으로 전처리, 문서의 벡터 표현 과정을 거치게 된다.

가. 전처리

텍스트 데이터를 분석하는 데 가장 기본적인 의미 단위는 단어라고 볼 수 있다. 간혹 언어에 따라서 단어를 분리하기 힘든 경우, 혹은 단어를 분리하기 위한 단순한 방법이 없을 때는, 즉 띄어쓰기를 하지 않는 언어에서 단어를 구별하기 힘든 경우에는 바이그램(bigram)이나 N그램을 사용해 분석의 기본 단위로 삼기도 한다. 따라서 전처리의 첫 번째 단계는 단어 추출 및 정규화(normalization)라 할 수 있다. 영어는 단어의 문법적 변이가 다양하지 않기 때문에 단순히 띄어쓰기 단위의 분리 후 스테밍(stemming)을 통한 정규화를 수행한다. 스테밍이란 간단히 예를 들

1 G. Chakraborty, Analysis of Unstructured Data: Applications of Text Analytics and Sentiment Mining, SAS Global Forum 2014 Proceedings, 2014. 재인용

어 talks나 talking에서 -s, -ing를 제거한 후 talk으로 치환해 같은 단어에 매핑하는 것이다. 포터 스테머(Porter stemmer)와 스노우볼 스테머(Snowball stemmer)가 가장 많이 사용되며, 텍스트 마이닝에 유용한 다양한 기능을 구현해 놓은 파이썬(python) 기반의 라이브러리 NLTK(Natural language toolkit)[2]에 포함돼 있다. 한글의 경우 현재 다양한 형태소 분석기가 공개돼 있고 은전한닢[3], 아리랑[4], 오픈소스 한국어 처리기[5]를 예로 들 수 있다.

단어를 기본 단위의 데이터라 했을 때, 자연 언어에는 명백히 의미를 담고 있지 않은 단어(불용어(stopword)라 함)가 있을 수 있으므로 분석의 성능을 향상시키기 위해 이들 불용어를 미리 제거하곤 한다. 예를 들어 영어에서 the, is, with와 같이 대부분의 문장에서 별다른 정보를 주지 않고 사용되는 단어가 불용어에 해당된다. 특히 문서를 단어의 집합으로 표현하고 단어의 순서를 고려하지 않는 분석방법에서는 제거해도 무방하다. 그러나 불용어가 문장, 문단의 컨텍스트 상에서 의미를 가질 수 있으므로 최근의 텍스트 마이닝에서는 제거하지 않고 분석에 들어가는 경향이 있다.

불용어를 판별하기 위해서는 갖고 있는 데이터에서 단어의 문서 출현 빈도(document frequency)를 계산해 가령 80% 이상의 문서에서 나타날 경우 불용어로 간주하는 등의 간단한 룰을 이용할 수 있다. 혹은 위에서 언급한 NLTK 라이브러리에서(유럽어권 언어에 국한됨) 언어별로 불용어 리스트를 제공하고 있어 매우 유용하다.

나. 벡터화

전통적으로 텍스트 마이닝에서는 최소 의미 단위인 단어의 집합으로 문서를 해체한 후 단어의 위치나 순서에 관계없이 문서에서 각 단어의 출현 빈도를 이용해 하나의 문서를 다차원(사전에 있는 단어의 수가 n개라 하면 n차원) 벡터로 표현한다. 이를

2 Apache license ver. 2.0을 따르는 오픈소스 소프트웨어이며 Natural Language Toolkit 홈페이지 https://www.nltk.org/에서 얻을 수 있다.
3 자바와 스칼라를 지원하는 라이브러리로 https://bitbucket.org/eunjeon/seunjeon에 공개돼 있다.
4 오픈소스 검색엔진인 루신(Lucene)의 엘라스틱서치 플러그인을 지원하는 한국어 형태소 검색기로 https://github.com/HowookJeong/elasticsearch-analysis-arirang 에 공개돼 있다.
5 트위터에서 한글을 처리하기 위해 사용하는 형태소 분석기로 알려져 있으며, 스칼라로 작성돼 있다: https://github.com/open-korean-text/open-korean-text

BOW(Bag-Of-Words) 표현이라고 흔히 부른다. 각 단어의 중요도에 따라 가중치를 주기 위해 TF-IDF(TF; Term Frequency, IDF; Inverse Document Frequency)를 사용한다. 한 문서에서 어떤 단어가 여러 번 출현할 수록 단어와 문서의 관련성이 높다고 볼 수 있으므로 TF는 단어 출현 빈도의 로그 값으로 계산된다. 그러나 대부분 문서에서 출현하는 단어(예를 들어 불용어)보다는 소수의 문서에서 출현하는 단어가 더 유용한 정보를 갖고 있다고 볼 수 있으므로, IDF는 문서 출현 빈도의 역수에 로그를 취한 값으로 계산한다.

앞에서 언급한 NLTK에는 TF-IDF를 계산하기 위한 API도 제공하고 있다[6]. 하지만 NLTK는 자연 언어 처리에 포커스를 맞춘 라이브러리이기에 최적화가 덜 돼 있어 속도가 느리고, 이보다는 Scikit-learn에 구현된 API가 더 효율적이다[7].

다. Word embedding

최근 인공신경망의 진화된 기술로서 딥러닝(deep learning)이 각종 기계학습 애플리케이션에서 뛰어난 성능을 보이고 있다. 텍스트 마이닝에서도 역시 딥러닝이 괄목할 만한 성과를 보이고 있다. 실수 값으로 표현되는 이미지나 음성 데이터와 달리 단어는 이산적(discrete) 데이터이므로 단어를 다차원 공간에 매핑해 벡터를 인공신경망의 입력 값으로 사용하는 기술이 개발됐다. 이와 같은 기술을 워드 임베딩(word embedding)이라 부르는데 대표적으로 word2vec[8]을 들 수 있다.

Word2vec은 각 단어의 앞뒤에 출현하는 다른 단어를 이용해 각 단어를 정해진 차원의 다차원 공간에 매핑한다. 즉 어떤 두 단어가 앞뒤에 출현하는 단어들이 비슷한 패턴을 갖고 있다면 다차원 공간상 가까운 위치에 매핑된다. 단지 단어를 각자 하나의 이산적 의미 단위(단어 아이디)로 활용하던 기존의 텍스트 마이닝에 비해, 앞뒤 문맥을 고려해 표현하므로 의미를 포착하기에 더 유리한 텍스트 데이터 표현 방법이라 할 수 있다.

워드 임베딩은 사용되는 인공신경망에 포함돼 매핑할 벡터가 신경망과 함께 학습되는 것이 일반적이므로 독립된 라이브러리는 존재하지 않는다. 단

6 API 문서는 http://www.nltk.org/api/nltk.html#nltk.text.TextCollection에서 찾을 수 있다.
7 http://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html 에서 API 문서를 찾을 수 있다.
8 T. Mikolov et al., Distributed Representations of Words and Phrases and their Compositionality, https://arxiv.org/abs/1310.4546

word2vec은 신경망 자체가 맥락(Context)에 따른 단어의 매핑을 위한 것이므로 몇 개의 공개 라이브러리가 있다. 가장 많이 사용되는 딥러닝 툴인 텐서플로에서는 튜토리얼을 통해 코드를 공개했다[9]. 다양한 텍스트 마이닝 기술을 구현해 놓은 gensim에도 word2vec API를 제공하고 있다[10]. Word2vec과는 단어의 맥락(Context) 고려 시 약간 다른 모델을 취하고 있으나 동일하게 워드 임베딩을 목적으로 하는 GloVe[11]도 코드를 공개하고 있다[12].

다음은 대표적인 텍스트 마이닝 문제[13] 몇 가지다.

1) 텍스트 자동 분류와 문서 군집화

텍스트 자동 분류(text categorization)와 문서 군집화(document clustering)는 구조적 데이터를 대상으로 하는 데이터 마이닝 기술의 대표적인 한 카테고리다. 텍스트 자동 분류와 군집화는 기존의 데이터 마이닝에서 개발된 알고리즘들과는 여러 면에서 다른 출발점을 갖고 있다. 우선 분류나 군집화의 기준으로 삼을 특징 값(feature)을 특정하기가 힘들다. 또한 텍스트 데이터를 벡터화했을 때 의미의 단위인 단어의 개수가 매우 많기 때문에 기존 자동분류와 군집화 알고리즘이 대상으로 했던 구조적 데이터에 비해 매우 고차원이다. 그리고 한 문서에는 전체 출현 단어 중 극히 일부만을 갖고 있어서 매우 성기다. 즉 벡터 대부분의 원소가 0을 갖는다. 따라서 다수의 텍스트 자동 분류·군집화 알고리즘에서는 고차원의 TF-IDF 벡터를 저차원으로 차원 감소(dimension reduction)를 한 다음, 데이터 마이닝 기법을 적용하곤 한다. 텍스트 자동분류에는 SVM(Support Vector Machine)이 가장 많이 쓰이며 좋은 성능을 낸다고 알려져 있다. 기계학습의 다양한 알고리즘을 구현해 놓은 오픈소스 소프트웨어인 Scikit learn[14]은 앞에서 소개한 바와 같이 TF-IDF 벡터화 API와 함께 SVM, 결정트리(decision tree) 등 다양한 자동 분류 알고리즘을 제공한다.

9 https://www.tensorflow.org/tutorials/word2vec
10 https://radimrehurek.com/gensim/models/word2vec.html
11 J. Pennington, R. Socher, C. D. Manning, GloVe: Global Vectors for Word Representation, 2014
12 https://nlp.stanford.edu/projects/glove/
13 DataQuest Whitepaper, Textmining: from raw text to useful insights, 2018에서 각 기술의 추가 설명을 찾을 수 있다.
14 scikit-learn: Machine Learning in Python, http://scikit-learn.org/

2) 딥러닝

딥러닝(Deep learning)은 지도 기계 학습 방법론인 인공신경망의 다른 이름이라고 할 수 있기에 딥러닝을 이용한 텍스트 마이닝은 엄밀히 말해 텍스트 자동분류의 한 종류라고 할 수 있다. 그러나 최근 괄목할 만한 성능 향상을 보이고 있어 별개로 다루고자 한다. 앞서 소개한 바와 같이 일반적으로 딥러닝에서는 문서의 각 단어를 워드 임베딩을 통해 다차원 벡터에 매핑하고 이를 입력으로 해 자동 분류를 수행한다. 네트워크 모델로는 가변 길이의 문서에 알맞은 재귀적 신경망(recurrent neural network)이 텍스트 분석에 가장 많이 사용되고 있으며, 이 중 긴 텍스트에 대해서도 맥락(Context)를 잃지 않는 LSTM(Long-Short Term Memory) 네트워크가 텍스트 마이닝에 주로 사용된다. 최근에는 텍스트 자동분류에 관해서는 이미지 분석을 위해 개발됐던 CNN(Convolutional Neural Network)이 LSTM에 버금가는 성능을 보인다고 보고된 바가 있다.

딥러닝을 텍스트 마이닝에 간단히 적용할 수 있는 오픈소스 소프트웨어는 아직 주목할 만한 것이 없으나 텐서플로, 케라스 등 딥러닝 학습을 위한 툴을 이용한 텍스트 마이닝 코드는 다수 공개돼 있다.

3) 토픽 모델링

텍스트 데이터에 대한 주요 애플리케이션은 주제별로 문서를 분류하는 문제일 것이다. 텍스트 마이닝 기술은 카테고리가 명시된 학습 데이터를 갖고 있는 경우 텍스트 자동 분류, 문서의 집합만이 주어진 경우 문서 군집화로 흔히 나누는데, 토픽 모델링(Topic modeling)은 문서 군집화의 하나로 분류할 수 있다. 애플리케이션에 따라 다양한 토픽 모델링 기술이 제안됐으나 대부분 PLSA(Probabilistic Latent Semantic Analysis)[15]와 LDA(Latent Dirichlet Allocation)[16]를 각 문제 정의에 맞게 확장했다고 볼 수 있다. 토픽 모델링은 사람들이 주제(주제의 개수는 적절히 주어졌다고 가정)에 따라 단어를 선별해 문서를 작성한다는 가정 하에 문서마다 각 주제에 대한 편중 정도를 수치화한다는 기본적인 틀을 갖추고 있다. 말하자면 토픽 모델링 역시 각 문서를 정해진 차원의 벡터(확률 분포)로 매핑하는 기술이라 할 수 있

15 Thomas Hofmann, Probabilistic Latent Semantic Indexing, SIGIR, 1999
16 D. M. Blei, A. Y. Ng, M. I. Jordan, Latent Dirichlet Allocation, Journal of Machine Learning Research, 2003. 3: 993 – 1022.

다. LDA는 PLSA의 과적합(overfitting) 문제를 완화했다고 알려져 있어 LDA가 텍스트 데이터의 토픽 모델로는 더 널리 사용된다. 앞서 소개한 바 있는 파이썬 기반의 gensim 프로젝트에서 LDA 모델 학습을 위한 API를 제공한다. 또한 gensim은 LDA 모델을 더 다계층으로 일반화한 HDP나 word2vec에 영감을 받아 뉴럴 네트워크를 이용한 문서의 벡터화를 계산하는 doc2vec과 같은 최신 토픽 모델링 기술도 제공하고 있다.

4) 개체명 인식

비구조적 텍스트 데이터로부터 지명, 인명이나 날짜, 금액 등 개별 정보를 추출해 이를 DB화하려는 요구 사항도 빈번히 존재한다. 개체명 인식(Name entity recognition)은 언어와 개체 타입 등을 고려해 텍스트로부터 이와 같은 정보를 가려내는 기술들을 일컫는다. 영어 텍스트 데이터를 분석하고자 한다면 OpenNLP 프로젝트에서 유용한 API를 제공한다[17]. 한글에 대한 개체명 인식 기술로는 몇몇 관련 연구자·개발자가 공개한 코드가 존재한다.

5) 자동 텍스트 요약

문서의 자동 요약(Automatic text summarization)도 텍스트 마이닝의 오래된 도전 과제였다. 매일매일 끊임없이 생성되는 뉴스 기사나 증권 리포트, 이메일 등을 포함한 텍스트 데이터에서 꼭 알아야할 정보를 요약해 제공할 수 있다면 매우 유용할 것이다. 자동요약 기술은 크게 두 가지로 분류된다.

첫 번째 문장추출형(extractive) 자동요약은 문서나 문단에서 가장 중요한, 혹은 가장 많은 정보를 담고 있는 문장을 추출해 제공하는 방법이다. 흔히 문장에 포함된 단어의 중요성 및 다양성을 수치화해 문장 단위로 순위를 매겨 제공하는 방법을 취한다.

두 번째는 요약문장생성형(abstractive) 자동요약이다. 문서나 문단을 기계가 이해하고 새로운 요약문장을 생성해 제공하는 궁극적인 자동요약 방법이다. 지금까지는 기계학습의 성능 제약 때문에 문장추출형 자동요약 연구가 주를 이뤘으나 최근 딥러닝의 등장 이후 문장생성형 자동 요약 방법론이 연구되고 있다[18].

공개된 문장추출형 자동요약 소프트웨어에는 펄(perl)기반의 프로그램 MEAD[19]와 자바 기반의 ROUGE[20]가 있다. 생성형 자동요약은 아직 만족할 만한 성능을 내지 못하기에 소개할 만한 솔루션이 없으나 딥러닝을 이용한 자동요약 연구의 소스코드가 다수 공개돼 있다.

17 Apache OpenNLP Developer Documentation, http://opennlp.apache.org/docs/1.8.4/manual/opennlp.html

3. 결론

이상으로 텍스트 마이닝 기술의 공통적인 스텝과 몇가지 주요 방법론 및 애플리케이션을 살펴봤다. 함께 텍스트 마이닝을 사용해 볼 수 있는 몇몇 공개 소프트웨어도 소개했다. 비구조적 데이터이기에 담고 있는 정보의 가짓수도 많고 텍스트 데이터로부터 얻길 원하는 정보 또한 다양하기 때문에 구조적 데이터를 분석·관리하는 SQL이나 RDBMS와 같은 정형화된 정보처리 시스템은 아직 찾아볼 수 없다. 그러나 비구조적 텍스트 데이터 축적 추세를 보아 텍스트 마이닝 기술의 수요는 빠르게 증가할 것이다. 현재 많은 연구가 진행되고 있고 눈에 띄는 성능향상을 보이고 있으므로 텍스트 마이닝은 앞으로 데이터 기술 시장에 큰 점유율을 차지할 것이라 생각한다.

18 R. Nallapati et al., Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond, SIGNLL, 2016.
19 http://www.summarization.com/mead
20 https://github.com/RxNLP/ROUGE-2.0

제2장

스트림 데이터 처리 시스템

정보기술(IT)의 응용 범위가 확장됨에 따라 다양한 데이터 소스로부터 끊임없이 공급되는 데이터의 흐름(스트림, stream)을 즉시 처리하기 위한 요구가 늘어나고 있다. 스트림 처리 시스템은 이러한 환경에 대응하기 위해 제안된 소프트웨어다. 이 장에서는 스트림 처리 시스템의 구조와 주요 기능, 구현 이슈들을 알아본다. 그리고 널리 알려진 오픈소스 스트림 처리 시스템들을 주요 기능 관점에서 간략하게 비교 분석한다.

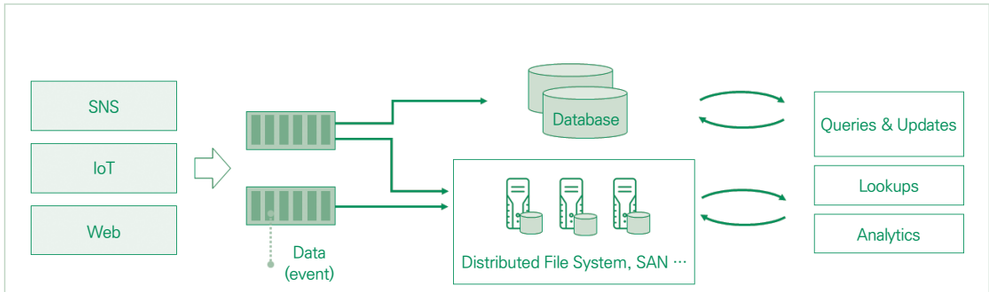
1. 스트림 처리의 필요성

빅데이터 응용(Application)의 상당수는 끊임없이 발생하는 데이터를 즉각적으로 처리하는 것들이다. 이처럼 한정되지 않은 데이터의 흐름을 처리하는 것이 스트림 처리(stream processing)다. IoT, 웹(click stream), SNS 데이터를 실시간으로 처리하는 응용들이 그 대표적인 예라 할 수 있다. 과거에는 생성되는 데이터를 데이터베이스나 파일과 같은 저장 장치에 기록하고, 추후 필요에 따라 저장된 데이터를 가공·분석하는 것이 일반적이었다. 이것은 큰 틀에서 보면 배치처리 방식이라 할 수 있으며, 필연적으로 데이터의 발생 시점과 그것을 처리하는 시점 사이에 상당한 시간차가 있다.

스트림 처리는 이러한 전통적인 데이터 처리 방식과는 반대의 접근방법을 취한다. 비즈니스 로직과 분석·질의 처리가 앞서 존재하고, 데이터는 이들 사이로 흘러가는 구조다(그림 7-2-1과 그림 7-2-2 참조). 데이터 스트림은 응용프로그램으로 즉각적으로 처리된다. 즉 데이터는 지정된 분석 작업을 거치거나 통계 값의 갱신에 사용되거나, 그렇지 않으면 추후 사용을 위해 저장될 수 있다. 스트림 처리에서는 여러 개의 데이터 스트림이 서로 연관돼 처리될 수도 있고, 하나의 데이터 스트림이 처리돼 새로운 스트림을 생성할 수도 있다.

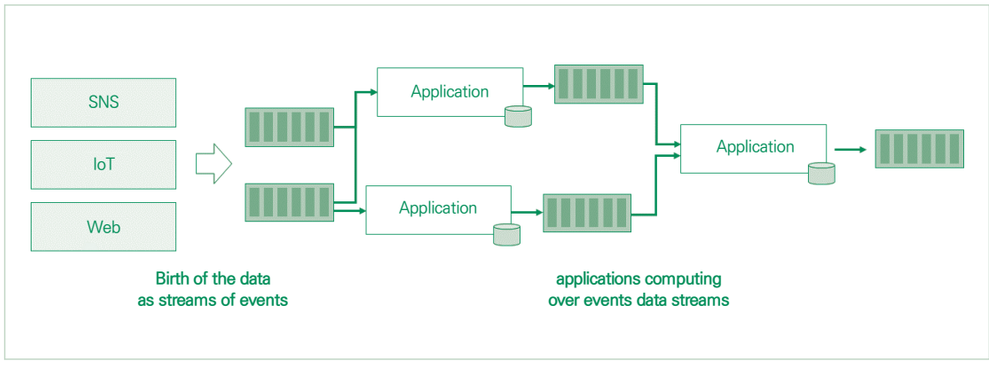
• 필자 | **송하주**(부경대 IT융합응용공학과 교수)

[그림 7-2-1] 일반적인 데이터 처리 방식



출처: https://data-artisans.com/platform(2018.07.15.접속)

[그림 7-2-2] 스트림 처리 방식



출처: https://data-artisans.com/platform(2018.07.15.접속)

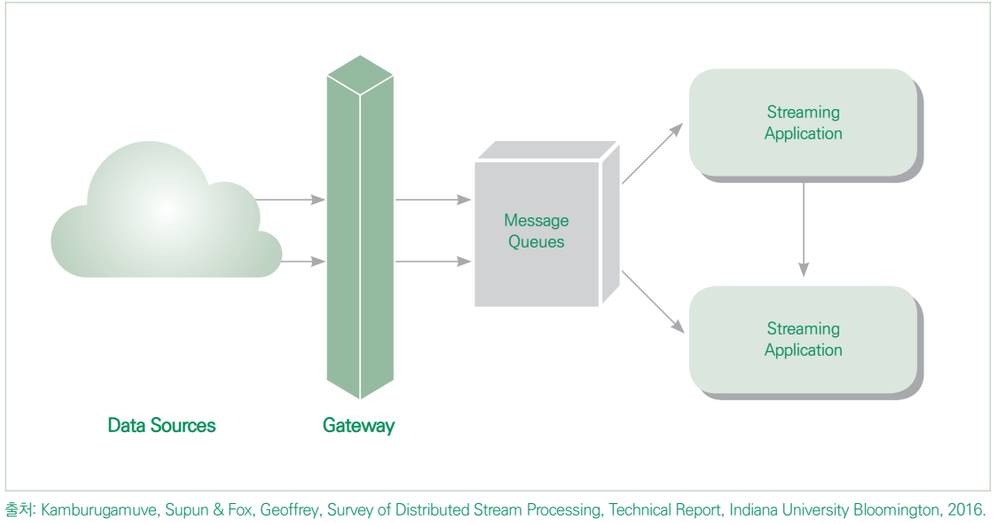
스트림 처리를 위해 지금까지 다양한 스트림 처리 시스템이 개발됐다. 초기 스트림 처리 시스템의 예로는 Aurora, Borealis, StreamIt, SPADE가 있다. 최근에는 Apache-S4, Apache Storm, Apache Samza, Apache Spark Streaming, Apache Heron 등이 오픈소스 프로젝트로 제공되고 있다. 상용 제품으로는 Neptune, Google Millwheel, Azure Stream Analytics와 Amazon Kinesis가 있다. Apache S4 커뮤니티는 더 이상 활발한 활동이 없다. Apache Storm은 Google Millwheel과 시스템 구조상 공통점이 많다. Heron은 Apache Storm의 실행 비효율성을 개선하기 위해 제안됐다. 이어서 스트림 처리 시스템의 구조와 기능적 요구사항들에 대해 알아본다.

2. 스트림 처리 시스템

가. 스트림 처리 시스템의 구조

[그림 7-2-3]은 스트림 처리 시스템의 내부 구조를 간략하게 나타낸 것이다. 스트림 처리 시스템의 입력으로는 센서와 응용프로그램을 포함한 매우 다양한 데이터 소스가 사용될 수 있다. 개별 데이터 소스는 언제든지 온라인 또는 오프라인 상태로 변경될 수 있으며, 데이터 소스별로 적용되는 통신 프로토콜이 다를 수 있다. 따라서 스트림 처리 시스템에는 이기종의 통신 프로토콜을 이해하고 스트림 응용프로그램으로 데이터를 전송할 수 있는 게이트웨이 모듈이 필요하다. 게이트웨이 모듈은 데이터의 전달에 따르는 오버헤드를 줄이기 위해 가능한 가볍게(lightweight) 동작하도록 구현되고 데이터 처리나 버퍼링 기능은 없는 것이 일반적이다. 게이트웨이를 거친 데이터는 메시지큐(message queue)로 보내지고 스트리밍 분석을 위해 적절한 응용으로 전달된다. 메시지큐는 스트리밍 응용프로그램에게 전달되는 데이터 원본이 존재하는 곳이며, 데이터 소스와 스트림 응용프로그램을 대응시키기 위한 버퍼링(buffering)과 라우팅(routing) 기능을 제공한다. 스트림 응용프로그램은 메시지큐에서 자신이 관심 있는 메시지를 받아 지정된 작업을 수행한다.

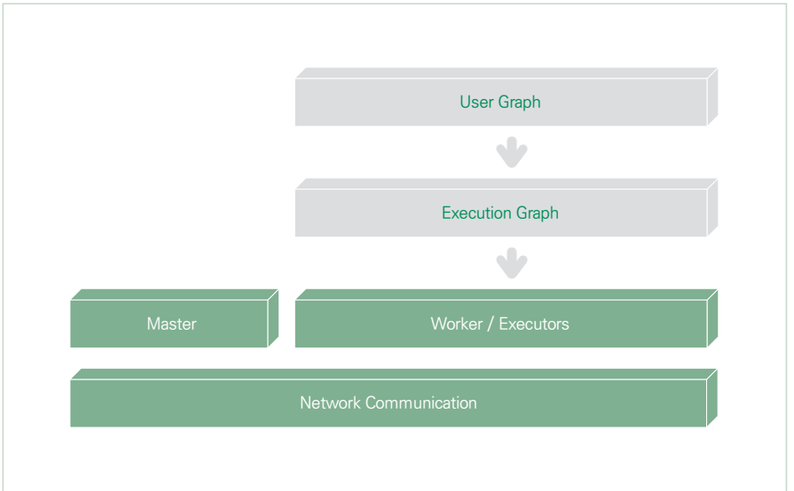
[그림 7-2-3] 스트림 처리 시스템의 기능적 구조



출처: Kamburugamuve, Supun & Fox, Geoffrey, Survey of Distributed Stream Processing, Technical Report, Indiana University Bloomington, 2016.

스트림 처리 시스템은 확장성 및 성능 내결함성을 제공하기 위해 분산 클러스터(distributed cluster) 방식으로 구성되는 것이 일반적이다. [그림 7-2-4]는 클러스터로 구현된 스트림 처리 시스템의 노드(node) 구성을 보인 것이다.

[그림 7-2-4] 스트림 처리 시스템의 클러스터 구조



출처: Kamburugamuve, Supun & Fox, Geoffrey, Survey of Distributed Stream Processing, Technical Report, Indiana University Bloomington, 2016.

클러스터는 크게 마스터노드(master node)와 작업노드(worker node)로 구분할 수 있다. 마스터 노드는 자원 관리와 작업 분배를 담당하고 작업 노드는 스트림 응용프로그램을 실행한다. 스트림 응용프로그램이 수행하는 작업은 방향성 그래프를 통해 그 처리 과정이 표현된다. 그래프의 간선을 따라 메시지가 흘러가며 노드에서는 입력 메시지에 대한 처리가 이뤄진다. 응용프로그래머가 작성한 사용자 그래프(user graph)는 실행 그래프(execution graph)로 변환되며 클러스터의 작업노드 상에서 수행된다. 작업노드들 사이에는 네트워크 통신(network communication)을 통해 데이터의 흐름이 연계된다.

나. 메시지큐

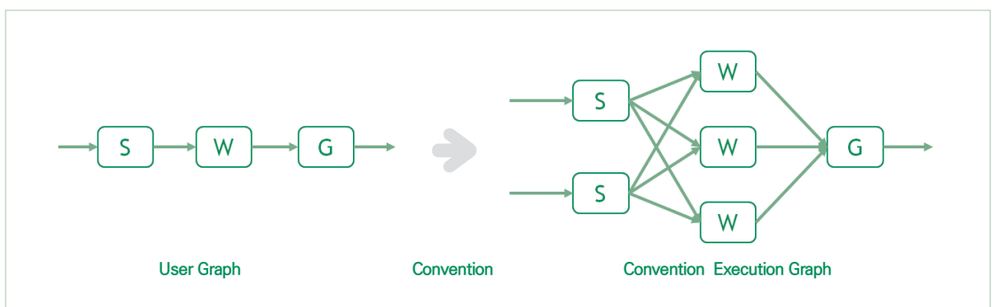
클러스터 방식의 스트림 처리는 메시지큐를 기반으로 동작한다. 스트림 처리에 있어 메시지큐는 두 가지 측면에서 중요하다. 첫째, 메시지큐는 데이터 스트림의

생산자와 소비자인 스트림 응용프로그램을 연결하는 매개체다. 메시지를 생산하는 측에서는 메시지의 소비자가 어떤 응용프로그램이 될지 알 수 없다. 심지어는 메시지 생성 시점에 응용프로그램이 실행중이 아닐 수도 있다. 따라서 메시지 큐는 이들 사이에 자연스럽게 메시지가 전달되도록 가교 역할을 제공한다. 둘째, 메시지가 생성되는 속도와 소비되는 속도에 차이가 있을 때 그 차이를 완화하는 버퍼 역할을 수행한다. 메시지큐는 크게 메모리 기반과 디스크 기반으로 구분할 수 있다. 메모리 기반 메시지큐는 메시지를 메모리에만 보관하고, 디스크 기반 메시지 큐는 메시지를 비휘발성 저장장치에 보관한다. 메모리 기반 메시지큐는 데이터가 신속하게 처리돼야 하는 경우, 둘 이상의 응용프로그램들이 관여하는 트랜잭션 메시지를 처리해야 하는 경우, 메시지큐에 동적 라우팅 규칙을 적용해야 하는 경우에 적합하다. 반면 디스크 기반 메시지큐는 스트림 응용프로그램이 처리 할 수 없는 짧은 시간 동안에 많은 양의 데이터가 생성될 수 있는 경우와 대량의 데이터가 생성되는 동안 스트림 응용프로그램이 오프라인 상태가 될 수 있는 경우, 동일한 메시지가 별개의 응용프로그램에 의해 처리되는 경우 등에 적합하다.

다. 스트림 처리 응용 모델

스트림 처리 시스템에서는 데이터 스트림을 간선으로 표현하고, 데이터를 처리하는 태스크를 노드로 표현하는 그래프를 사용해 스트림 응용을 모델링하는 것이 일반적이다. 스트림 처리 시스템의 그래프는 사용자(개발자)가 작성하는 사용자 그래프와 그것을 시스템에서 사용할 수 있도록 변환한 실행 그래프로 구분할 수 있다.

[그림 7-2-5] 사용자 그래프와 실행 그래프



출처: Kamburugamuve, Supun & Fox, Geoffrey, Survey of Distributed Stream Processing, Technical Report, Indiana University Bloomington, 2016.

[그림 7-2-5]의 좌측은 S, W, G 노드로 표현된 3개의 태스크와 그 사이에 데이터 스트림이 연결된 사용자 그래프다. 데이터 스트림은 맨 처음 S 태스크를 거쳐 새로운 스트림을 생성한다. 그것은 다시 W 태스크를 거쳐 G 태크에 입력된다. G 태스크는 마지막 태스크다. [그림 7-2-5]의 우측은 사용자 그래프가 변환된 실행 그래프를 보여준다. 여기에서는 S 태스크에 대해 2개, W에 대해서는 3개, G연산에 대해서는 1개의 실행 인스턴스가 사용되는 것을 보여준다. 사용자 그래프에서 실행 그래프로의 변환은 시스템별로 달라질 수 있다. 내결함성, 메시지 전달보증, 처리량과 대기시간, 태스크 스케줄링 및 자원 관리, 태스크 사이의 통신 방법에 따라 실제 실행을 달리할 수 있기 때문이다.

라. 실행 모델(Execution Model)

1) 메시지 처리 방식

메시지의 처리는 크게 두 가지 방식으로 구분할 수 있다. 첫째는 스트리밍되는 개별 메시지 단위로 태스크를 즉시 처리하는 방식이다. 스트림 처리에 있어 가장 일반적인 방식이며 대부분의 응용에서 적용할 수 있는 방식이다. 반면 이 방식은 처리량(throughput)과 장애복구(fault recovery) 측면에서 비효율적이다. 둘째는 일정 주기(time window)별로 메시지를 모아서 처리하는 방식이며 마이크로 배치(micro-batch)라고도 한다. 메시지를 묶어 추상화 이벤트를 구성하기에 유리하고 장애복구와 부하 분산에 유리한 반면 일반적인 스트림 처리 방식은 아니다. 또한 응답대기시간(latency)이 늘어날 수 있다.

2) 태스크 관리

태스크를 수행하기 위해서는 태스크 분배, 태스크 실행, 태스크들 간의 통신이 적절하게 이뤄져야 한다. 태스크 분배는 클러스터 내의 노드에 적절하게 작업을 할당하는 과정으로서 YARN이나 Mesos가 널리 사용된다. 태스크 할당방법에 따라 내결함성(fault tolerance)과 시스템 성능(performance)은 크게 달라질 수 있다. 다수의 노드로 이뤄진 클러스터 시스템에서 태스크를 여러 노드에 배분하여 실행하면 특정 태스크에 문제가 발생하더라도 전체 응용프로그램의 동작에는 큰 문제가 없다. 따라서 내결함성 측면에서는 우수하다 할 수 있다. 반면 태스크 간의 통신이 TCP 네트워크를 통해 이뤄지므로 개별 응용프로그램 수준에서의 성능이 저하될

수 있다. 이와 반대로 응용프로그램 내의 태스크들을 하나의 노드에 집중해서 할당하면 해당 노드에 문제가 발생하는 경우 응용프로그램의 실행이 불가능하게 되는 단점이 있으나, 태스크들 간의 데이터 통신이 노드 내부에서 이뤄지므로 통신 부하를 크게 줄일 수 있다.

3) 네트워크를 통한 데이터 전달

클러스터에 분산된 태스크들 사이에는 네트워크를 통해 데이터가 전송돼야 한다. 송신 태스크 측에서는 데이터 직렬화(serialization)를 사용해 메모리 내의 데이터를 이진(binary) 포맷으로 변환한 다음, 네트워크로 전달하고, 수신 측에서는 역직렬화(deserialization)를 통해 전달 받은 바이너리 포맷을 메모리 포맷으로 변환해야 한다.

마. 메시지 처리 보장

분산 클러스터 환경에서는 노드 장애, 네트워크 장애, 소프트웨어 버그 및 리소스 제한으로 인해 부분적인 장애가 발생할 수 있다. 스트림 처리 시스템의 경우 응답 대기시간이 매우 중요하다. 따라서 클러스터의 일부에 장애가 발생하더라도 전체 시스템에 미치는 영향을 최소화하면서 정상적인 처리가 계속될 수 있도록 빠르게 장애를 복구해야 한다. 스트림 처리 시스템은 '정확히 1번', '적어도 1번', '보증 없음'이라는 세 가지 유형의 메시지 처리 보장(Message Processing Guarantees)을 한다.

메시지 처리 보장은 시스템의 장애복구 구현과 밀접한 관련이 있다. 장애복구에는 정밀(precise) 복구, 롤백(rollback) 복구 및 갭(gap) 복구가 있다. 정밀 복구는 대기시간이 약간 늘어나는 것을 제외하고는 메시지 처리에 실패가 일어나지 않도록 하는 것이다. 때문에 '정확히 한 번'의 메시지 처리를 보장한다. 롤백 복구는 실패한 태스크를 다시 실행하는 것이다. 따라서 장애가 발생하면 메시지가 두 번 이상 처리될 수 있다. 즉 롤백 복구의 경우에는 메시지에 대해 '적어도 한 번'의 실행을 보장한다. 갭 복구의 경우에는 정보의 손실이 일어날 수 있으므로 메시지 처리는 '보증 없음'이라 할 수 있다.

3. 스트림 처리 시스템의 비교

[표 7-2-1]은 오픈소스 프로젝트로 제공되는 스트림 처리 시스템들을 비교한 것이다. 대다수의 시스템이 자바 또는 자바 기반의 언어로 개발됐기 때문에 JVM(Java Virtual Machine) 상에서 동작한다.

[표 7-2-1] 스트림 처리 시스템들의 비교[1]

기준 \ 시스템	Apache Storm	Apache Flink	Apache Spark Streaming	Apache Samza
개발 언어	Java, Closure	Java, Scala	Java, Scala, Python	Java, Scala
메시지 처리 방식	스트림	스트림, 배치	배치	스트림
대기시간	매우 짧음	매우 짧음	큼	작음
처리량	낮음	높음	높음	높음
메시지 처리 보장	적어도 한 번	정확히 한 번	정확히 한 번	적어도 한 번
메시지큐	Netty	Netty	Netty	Kafka
태스크 스케줄러	Nimbus	Job Manager	Yarn, Mesos	Yarn
데이터 직렬화 도구	Kryo	Data Stream	RDD	Custom

Spark Streaming의 경우에는 입력 스트림을 마이크로 배치 방식으로만 처리할 수 있는 반면, Apache Storm과 Apache Samza는 스트림 방식만을 지원한다. Flink는 스트림과 배치 방식 모두 지원한다. 마이크로 배치 형식을 사용하는 Apache Spark Streaming이 대기시간이 가장 길 것으로 예상된다. 하지만 대기시간은 구현 알고리즘에 크게 영향을 받기 때문에 실제 환경에서의 대기시간은 [표 7-2-1]에서 제시된 순위와 달라질 수 있다. 처리량 성능 비교는 객관적인 실험 결과를 근거로 한 것이 아니며, 적용되는 기술에 따라 예상되는 성능 수준을 나타낸 것이다. 메시지 처리 보장 측면에서 Flink와 Spark Streaming은 '정확히 1번' 처

1 다음 두 문헌을 일부 인용함
- G. Hesse and M. Lorenz, Conceptual Survey on Data Stream Processing Systems, 2015 IEEE 21st International Conference on Parallel and Distributed Systems(ICPADS), Melbourne, VIC, pp. 797-802, 2015.
- Kamburugamuve, Supun & Fox, Geoffrey, Survey of Distributed Stream Processing, Technical Report, Indiana University Bloomington, 2016.

리를 보장하는데 반해 Storm과 Samza는 '적어도 한 번' 처리를 보장한다. 메시지 큐의 경우 Samza가 Kafka를 사용하는 반면, 나머지 시스템들은 Netty를 사용한다. 특히 Samza는 태스크간 스트림 연결에도 Kafka를 사용하기 때문에 안정적인 메시지 흐름제어 기능이 제공되나 성능 측면에서는 좋지 않다. 클러스터 구조 측면에서 제시된 시스템 모두 마스터노드-작업노드 모델을 채택하고 있다. 따라서 네 시스템 모두 태스크 스케줄러가 구동되는 마스터 노드가 단일장애지점(single point of failure)이 될 수 있다. 데이터 직렬화 또한 시스템의 성능에 크게 영향을 미칠 수 있다. [표 7-2-1]에서 제시된 바와 같이 모든 시스템이 서로 다른 직렬화 방식을 사용하고 있다. Spark Streaming의 경우 RDD 직렬화가 기본이지만, 설정 변경을 통해 Kryo를 사용하도록 할 수 있다.

4. 맺음말

지금까지 널리 사용되고 있는 오픈소스 기반의 스트림 처리 시스템들의 특징을 간략히 살펴보았다. 스트림 데이터의 처리라는 목적은 동일하지만 사용자 API부터 내부 구현 방식까지 시스템별로 차이가 있다. 특정 스트림 처리 시스템이 모든 스트림 응용 환경에 적합할 수는 없다. 스트림 처리를 적용하려는 응용 환경에 대해 상세한 분석 및 명확한 요구사항 파악을 통해 적합한 스트림 처리 시스템을 선택해야 할 것이다.

제3장

실시간 인메모리 DBMS

사물인터넷, 빅데이터를 시작으로 인공지능에 이르기까지 신기술의 바탕이 되는 데이터 분석 시장은 디스크 기반에서 메인 메모리 기반으로 이동하고 있다. 이 글에서는 실시간 처리를 지원하는 인메모리 DBMS의 필요성과 등장 배경, 실시간 인메모리 DBMS 개발 방향을 결정하는 최신 기술을 소개하고, 실시간 인메모리 DBMS의 발전 방향을 소개한다.

1. 실시간 처리를 지원하는 DBMS의 필요성

메모리 기반 DBMS(DataBase Management System)는 1980년대부터 운영체제 성능 향상을 목적으로 IBM 등에서 개발을 시작했으며, 초기에는 항공 서비스와 같은 대규모 시스템 관리에만 국소적으로 사용됐다. 이후 메모리 가격 하락, 인메모리 처리 수요 증가 등으로 인메모리 컴퓨팅 시장이 폭발적으로 성장하면서, 기존 디스크 기반 DBMS를 인메모리 DBMS로 전환하려는 움직임이 활발해졌다. 특히 실제 운영 환경에서 기존 DBMS에서 발생하는 과부하, 데이터 병목현상이 업무에 지장을 준다는 의견이 많아지면서 오픈소스 인메모리 DBMS를 자체적으로 도입하는 사례 역시 증가했다.

인메모리 DBMS는 대부분 기존과 유사한 SQL로 관리 가능한 '빠른' 데이터 관리 시스템을 목표로 한다. 초반에는 초당 수천 트랜잭션 이상의 처리를 요구하는 다중 사용자를 위한 서비스 일부에 사용했으나, 지금까지 등장한 오픈소스와 상용 인메모리 DBMS들은 초당 수만 건 이상의 트랙잭션을 처리하고 있다. 이 중에서도 초당 처리량이 수십만 건 이상인 실시간 DBMS들은 사물인터넷, 인공지능 분야에서 시계열 형태의 센서 데이터와 로그 관리에 많이 사용되고 있다. 시장조사 기관인 가트너(Gartner)는 운영 DBMS의 선두주자로 마이크로소프트(Microsoft)

• 필자 | **문양세**(강원대 컴퓨터학부 교수)

와 오라클(Oracle)을 꼽는다. 최근 국내에서도 실시간 시계열 DBMS 시장이 활성화되고 있다.

2. 실시간 인메모리 DBMS의 변화를 이끌 주요 기술

빠른 처리 속도를 내기 위해, 대부분의 인메모리 DBMS 제품군에서는 새로운 트랜잭션 처리 방식, 분산 환경, 개선된 SQL 처리 엔진 등을 도입하고 있다. 상용 DBMS 시장을 장악하고 있는 마이크로소프트(Microsoft), 오라클(Oracle), SAP 등도 이미 자체 개발한 인메모리 DBMS를 출시했으며, 출시한 제품의 클라우드 서비스화를 활발히 추진하고 있다. 오픈소스 진영에서도 이미 분산 환경 기반의 인메모리 DBMS를 다수 공개했는데, 이들 제품군에서 도입한 주요 기술에 대해 좀 더 자세히 살펴보자.

가. 새로운 시스템 모델링과 질의 처리 기술

RDBMS(Relational Database Management System)는 이미 수십 년 전부터 데이터를 테이블 형태로 관리하는 대표적인 시스템으로 자리 잡았다. 초기의 RDBMS는 일반적으로 안정적인 내부 네트워크 기반 서버들로 구성했기 때문에, 분산 환경을 크게 고려하지 않았다. RDBMS가 채택한 ACID(Atomicity, Consistency, Isolation, Durability) 모델 역시, 분산 네트워크 환경은 고려하지 않는다.

이후 클라우드 기술이 발전하면서 빅데이터, NoSQL, 분산 인메모리 DBMS 제품들이 등장하고, 데이터 저장 관리의 유일한 해결책으로 꼽히던 RDBMS와는 다른 구조의 데이터 저장 방안들이 제시되기 시작했다. 대용량 데이터 저장 관리 기술들은 단일 시스템이 아닌 분산 시스템을 기반으로 한다. 기존 RDBMS에서 사용하는 ACID 모델은 분산 환경에 적용하기 어렵다는 문제가 있다. 그 때문에 새로운 구현 모델인 CAP(Consistency, Availability, Partition tolerance) 모델을 참고해 NoSQL 시스템을 개발하기 시작했다. 그러나 CAP 모델은 분산 환경만을 고려한 이론이 아니며, 각 속성 정의에 일관성이 없어 모델 전체 해석에 혼돈이 생겼다. 이런 단점을 개선하기 위해, BASE, PACELC 등 새로운 이론이 등장했다. 이 중 NoSQL, NewSQL, 인메모리 DBMS의 모델링에 많이 사용하는 BASE는 다음과 같다.

- Basically Available: 기본적으로 가용성을 중시하며, 다수의 스토리지에 복사본을 저장하는 것으로 이를 해결하고자 함
- Soft-state: 노드 상태는 외부에서 전송한 정보로 결정되며, 분산 노드 간 업데이트는 데이터가 노드에 도달한 시점에 갱신됨
- Eventually consistent: Soft-state 속성에서, 데이터가 해당 노드에 도달할 때까지는 데이터에 일관성이 없는 상태이나, 데이터가 도착하는 순간 일관성을 보장할 수 있음. 이는 시스템이 일시적으로 일관되지 않은 상태여도 일정 시간 후에는 다시 일관성을 갖는 성질을 의미함

BASE 모델에 대한 의견 역시 다양하지만, 이들이 기본적으로 분산 환경에 적합한 속성이라는 것은 분명하다. 분산 환경을 이용해 RDBMS의 ACID 보장을 위한 오버헤드를 감소시키면, 훨씬 더 유연하고 빠른 데이터 관리 시스템 개발이 가능하다.

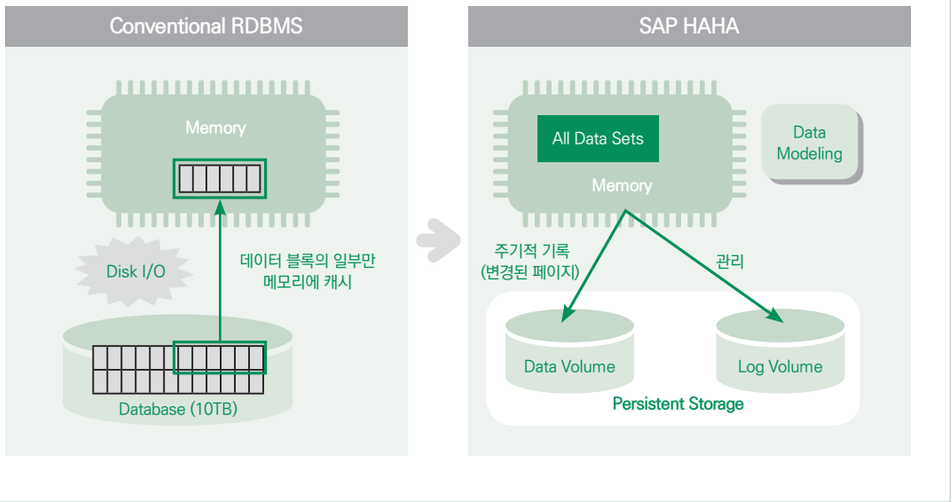
또 인메모리 DBMS의 처리 성능을 향상하기 위해 기존 SQL은 그대로 지원하면서, INSERT, DELETE 연산으로 UPDATE 연산을 대체하는 방식도 적용된다. 이 방식은 질의 처리 능력을 최적화한다. 이는 메모리에서 연산을 수행하게 하는 방식으로 하드웨어적인 성능 향상뿐 아니라, 기존 RDBMS의 질의 처리 과정을 개선해 소프트웨어적인 성능 향상까지 가져온다.

오라클 데이터베이스 인메모리 옵션의 경우, 질의 처리와 분석 시간을 감소시키기 위해 메인 메모리에 행 기준 저장 방식과 열 기준 저장 방식을 모두 지원한다. 이는 다른 목적을 위한 두 개의 인덱스를 생성하는 것과 유사한 방식이다. 기

[그림 7-3-1] 오라클 데이터베이스 인메모리의 속도 개선 기술

출처: Oracle Database In-Memory with Oracle Database 18c, 2018.02.

[그림 7-3-2] SAP HANA의 동작 구조



출처: LGCNS, 「Hana 기반 Private Cloud은 왜 특별한가?」, 2017.12.

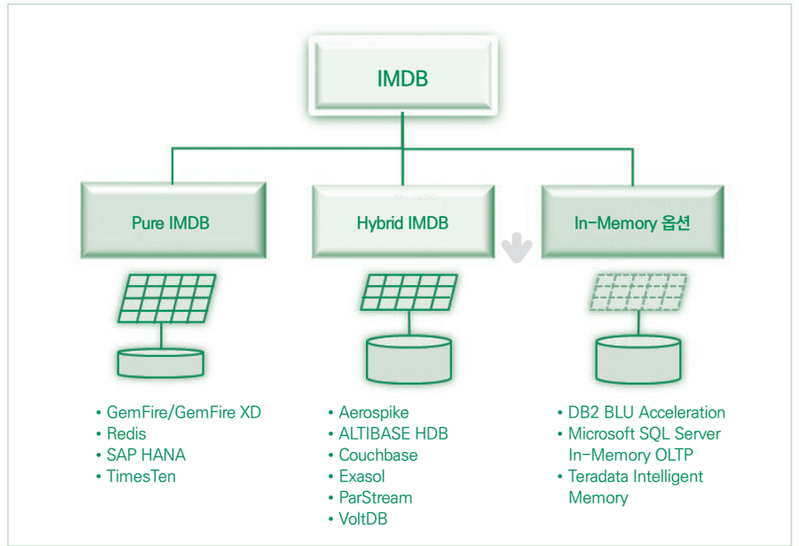
존 RDBMS와 같은 트랜잭션 처리를 위해 행 기준 저장 데이터를, 분석용으로는 열 기준 저장 데이터를 사용한다. 특히 열 기반 데이터는 별도의 로깅이 없어 기존 저장 방식보다 부하가 적게 발생하기 때문에 빠른 검색과 분석이 가능하다.

SAP HANA는 오라클 데이터베이스 인메모리와는 다르게 전체 데이터를 모두 메모리에 상주시키고 주기적으로 디스크에 복제하며 로그를 통해 관리하는 모델이다. 이는 오라클의 TimesTen 인메모리 DBMS와 유사한 구조로 시스템 구축 시 메인 메모리 구입에 많은 비용이 드는 단점이 있다. 이와 유사한 모델로 구성된 대부분의 인메모리 DBMS에서는 저장 공간의 낭비를 줄이기 위해 중복·유사 데이터 압축 기술을 적용한다.

나. 실시간 분석에 최적화된 시계열 스트림 처리 기술

기존 인메모리 DBMS는 순수 인메모리 DBMS, 하이브리드 인메모리 DBMS, 인메모리 옵션형의 세 가지로 구분된다. 순수 인메모리 DBMS는 모든 데이터를 메인 메모리에서 관리하고 디스크를 백업용도로 사용하는 모델이다. 반면 하이브리드 인메모리 DBMS는 빠른 작업에 필요한 데이터는 메인 메모리에, 대량 데이터 관리는 디스크에서 수행하게 해 유연성을 높인 기술이다. 하이브리드 방식은 순수

[그림 7-3-3] 인메모리 DBMS의 구분



출처: 이윤성, 「테라데이타 데이터베이스 중심」, 2015.

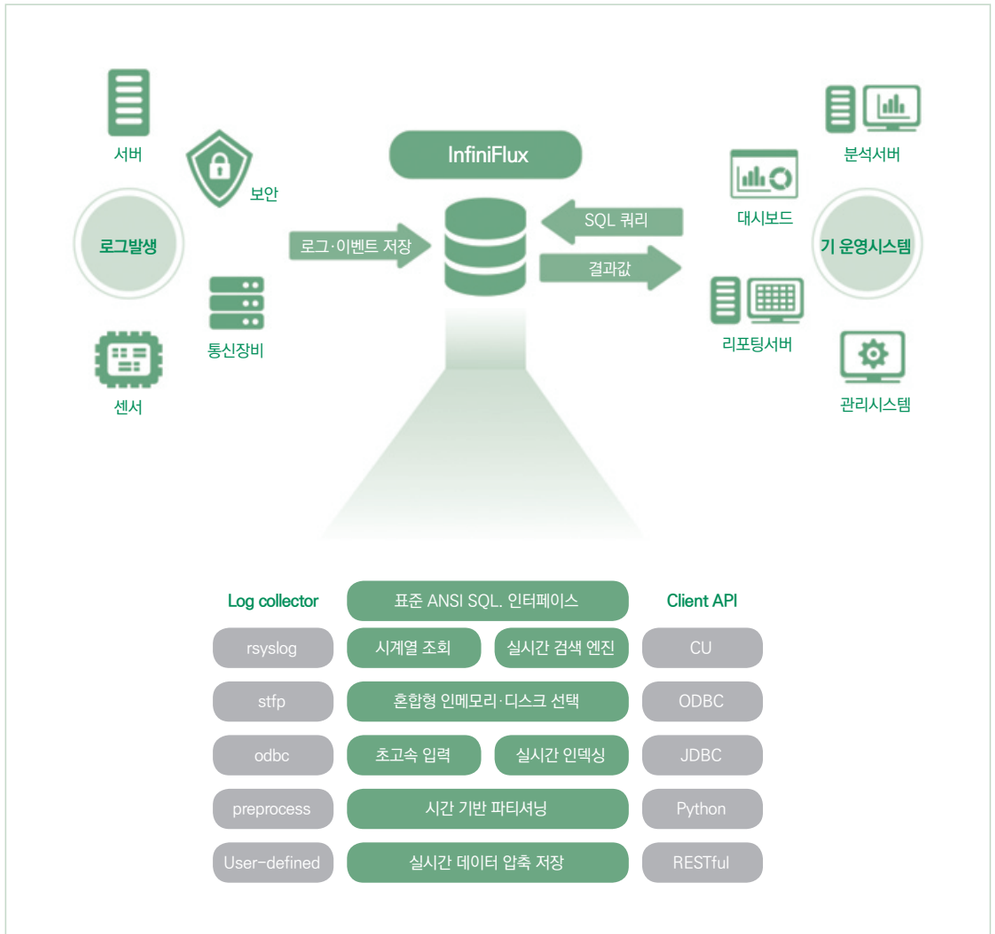
인메모리 DBMS보다는 대용량 메인 메모리 구축에 대한 부담이 적을 수 있으나, 시스템 설계 시 성능을 기준으로 데이터베이스를 잘 정의해야 병목현상 없이 활용할 수 있다. 마지막으로 인메모리 옵션형은 기존 디스크 기반 DBMS에서 메인 메모리 사용을 켜고 끌 수 있도록 옵션으로 지원하는 방식이다. 이는 기존 DBMS 환경은 그대로 사용하면서, 빠른 처리가 필요한 상황이 발생했을 때 별다른 추가 구축 없이 인메모리 DBMS를 사용할 수 있는 것이 장점이다. 그러나 대부분 상용 DBMS에서 추가 제공하는 기능이라 새로 도입하기에는 기존 시스템에 종속적이라는 문제가 있다.

인메모리 DBMS는 운영하려는 시스템 목적에 맞게 선택하는 것이 가장 바람직하지만, 최근 사물인터넷과 빅데이터 기술이 등장하면서 초고속으로 생성되는 센서와 로그 데이터를 빠르게 처리하기 위해서는 기존 성능을 넘어서는 더 빠른 인메모리 DBMS가 필요해졌다. 이런 시스템들은 기존 인메모리 DBMS와 구분하기 위해 실시간 DBMS라 부르기도 한다. 대부분 트랜잭션을 끊임없이 수행할 수 있도록 연속(Continuous) 질의를 지원한다.

실시간 DBMS는 기존의 DSMS(Data Stream Management System)와 목적이 유사하다. 특히 시계열 형태의 스트림 데이터 처리에 적합한 시스템이 많으며, 좀

더 빠른 성능의 인메모리 DBMS 역시 실시간 DBMS로 사용할 수 있다. 최근 국내에서 소개된 마크베이스(Machbase)는 초당 100만 건 이상 수집 가능한 시계열 DBMS로 주목받고 있다. 그 밖에 인플럭스DB, 카이로스DB와 같은 시계열 DBMS들은 대부분 디스크를 기반으로 한다. 이들은 스키마를 단순화한 대신 초고속 탐색이 가능한 실시간 인덱스를 메모리에서 관리하는 기술을 사용한다. 또한 시계열 스트림 처리에 많이 사용되는 통계(합계, 평균, 개수 등) 함수, 샘플링·필터링 함수들을 지원하며, 이후 분석과정에서 불필요한 추가 연산을 줄일 수 있다.

[그림 7-3-4] 인피니플럭스(시계열 DBMS)의 핵심 기술



출처: 「Machbase DBMS 특징」, 2016.04.

다. 향상된 하드웨어 기술 활용

인메모리 DBMS의 여전한 단점은 휘발성 메모리인 DRAM(Dynamic Random-Access Memory)에서 데이터 관리와 연산을 처리하므로 전원 공급이 중단될 경우 데이터 손실이 발생할 수 있다는 점이다. 이를 해결하기 위해 대부분의 인메모리 DBMS는 HDD, SSD와 같은 디스크에 데이터를 주기적으로 백업하거나, 인메모리 연산을 필요한 경우에만 사용하는 방식을 취해 왔다. 그러나 인메모리 DBMS가 서버급 시스템뿐 아니라 스마트폰, 사물인터넷 기반 가전 등에서의 사용량이 증가함에 따라, 새로운 하드웨어 기술에도 영향을 받을 것으로 보인다. 특히 NVM(Non-Volatile Memory), SCM(Storage Class Memory) 등 메모리 반도체 기술이 발전하고 있어 인메모리 DBMS 기술 시장은 점점 더 커질 전망이다.

NVM은 비휘발성 메모리를 의미하며, 기존 DRAM과 다르게 전원 공급이 중단되더라도 데이터가 삭제되지 않는다. 특히 2020년대에 들어서면 DRAM의 저장 용량이 더 이상 확대되기 어려울 것이라는 전망에 따라 대량 데이터 고속 처리를 위한 인메모리 DBMS들은 저렴한 낸드 플래시 메모리나 인텔의 3D Xpoint와 같은 상변화 메모리(Phase-Change Memory)와 함께 사용되는 계층적 구조를 갖게 될 것이다.

최근에는 데이터를 VRAM(Video RAM)에서 관리하기 위해 GPU 컴퓨팅 방식을 적용한 DBMS도 등장하고 있다. GPU 컴퓨팅으로 데이터 관리 시 가장 큰 장점은 별 다른 인덱스 구축과 관리가 필요 없다는 점이다. 다시 말해 데이터베이스 초기 구축 시 필요한 사전처리 과정을 생략할 수 있으며, 구축 후에도 인덱스 관리에 드는 연산 비용을 획기적으로 줄일 수 있다. GPU 하드웨어 추가를 통한 확장 역시 가능하므로, 초고속 처리가 필요한 서비스에 인메모리 DBMS와 함께 GPU 컴퓨팅이 많이 사용될 것으로 보인다.

3. 실시간 인메모리 DBMS의 향후 전망

최근 빅데이터, 인공지능, 사물인터넷 등 다양한 기술이 등장하면서 이를 처리하는 다양한 실시간 처리 기술들이 주목받고 있다. 이 중에서도, 정형화된 데이터 처리에 최적화된 DBMS의 실시간 처리 지원이야 말로 가장 많은 요구가 발생하

는 분야다. 2017년 10월 미국 시장조사기관인 가트너(Gartner)에서 발표한 2018년 10대 전략 기술 트렌드에는 '인공지능을 좀 더 활용한 지능형 서비스' '클라우드를 좀 더 최적으로 사용할 수 있는 엣지(edge) 컴퓨팅' '증강현실과 가상현실 기반의 몰입 경험 기술' '블록체인과 이벤트 기반 비즈니스 모델'이 소개됐다. 현재 주목받고 있는 기술들과 앞으로 주목받게 될 기술들은 대부분 데이터 관리와 처리에 집중하고 있다. 다시 말해, 향후 등장할 기술들은 시스템 규모와 상관없이 DBMS 시스템이 필수적이며, 이를 위해서는 더 발전된 형태의 DBMS 개발이 필요하다.

앞서 설명한 것처럼 상용 DBMS 시장을 주도하고 있는 마이크로소프트, 오라클, SAP 등도 실시간 처리를 위한 인메모리 DBMS 기술에 뛰어들고 있다. 오픈소스 인메모리 DBMS 역시 각각의 장점을 확보하며 발전하고 있다. 하드웨어는 점점 더 발전할 것이며, 응용 분야로의 적용도 늘어날 것이다. 실시간 인메모리 DBMS는 RDBMS의 뒤를 잇는 데이터 관리의 핵심 기술로, 사물인터넷·인공지능·빅데이터·스마트서비스 등 다양한 분야에 활용될 것으로 예상된다. 인메모리 DBMS 기술은 끊임없이 개선되고 변화할 것이다.

제4장

병렬 구조와 데이터 연산

병렬 구조는 데이터 연산을 더 빠르게 수행할 수 있는 컴퓨터 구조로 꾸준히 각광받고 있다. 중앙처리장치와 그래픽처리장치는 병렬 처리 구조를 적용시킨 대표적인 데이터 연산 장치다. 병렬 구조에서 데이터 연산은 병렬 연산이라 하며, 병렬 연산은 연산 수행 방법에 따라 분류한다. 이들은 각기 다른 연구·응용 분야에서 활용되고 있다. 본 장에서는 병렬 구조에 대한 개념과 병렬 구조에서의 데이터 연산에 대해 자세히 알아보고 병렬 구조의 미래를 전망해 본다.

1. 병렬 구조

병렬 구조는 주어진 문제를 작게 나눠 동시에 해결할 수 있도록 설계한 연산 구조다. 초기에 등장한 병렬 구조는 슈퍼컴퓨터(supercomputer)와 워크스테이션(workstation)에서 채택한 컴퓨터 구조로, 대규모 연산을 고속 처리하는 것이 목적이었다. 집적 회로 기술이 발달하면서 이후에는 개인용 컴퓨터에도 병렬 구조를 적용해 병렬 연산을 수행할 수 있게 했다. 대표적인 병렬 구조 장치는 CPU(Central Processing Unit))와 GPU(Graphic Processing Unit)다. 두 장치는 입력 데이터 집합을 복수개의 부분 데이터 집합으로 분할하고, 다수의 코어(core)와 프로세스가 분할된 부분 데이터 집합의 데이터 연산을 수행할 수 있도록 했다. 이때, 각 프로세스가 하나 이상의 스레드(thread)를 수행하게 해 자신이 담당한 데이터에 대한 연산을 수행할 수 있도록 했다. 여기서 스레드는 어떤 프로세스 내에서 실행되는 흐름의 단위다.

CPU와 GPU는 하드웨어적으로 보유한 코어의 성능과 개수에서 차이가 난다. 동일한 데이터 처리 성능을 갖는 CPU와 GPU에서 CPU의 코어 수는 GPU의 코어 수보다 월등히 적지만, CPU의 각 코어 성능은 GPU의 각 코어 성능보다 뛰어나다. 이런 특징 때문에 CPU는 관계형 데이터베이스 중에서도 관계 연산과 같이 각 코어의 성능이 중요한 데이터 마이닝 분야에서 활용하고 있다. 이와 달리, GPU는 최

• 필자 | **민준기**(한국기술교육대 컴퓨터공학부 교수)

근 주목되고 있는 딥러닝(deep learning)처럼 각 코어의 성능보다는 다수의 코어가 필요한 인공지능(artificial intelligence) 분야에서 활용하고 있다.

2. 병렬 구조에서의 데이터 연산

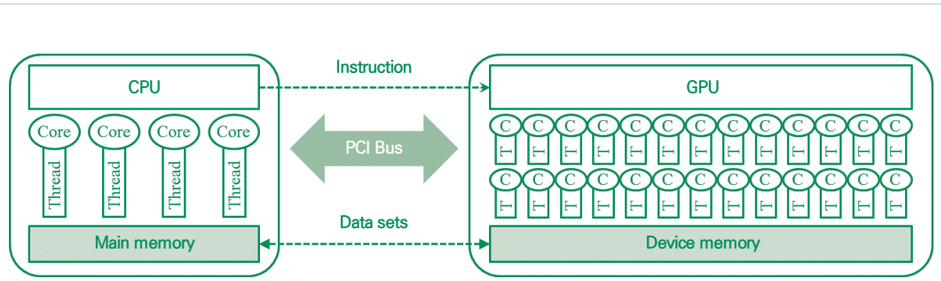
병렬 구조에서의 데이터 연산은 병렬 연산이라고 정의하는데, 연산 수행 방법에 따라 스레드 수준 병렬화, 명령어 수준 병렬화, 데이터 수준 병렬화로 분류할 수 있다.

가. 스레드 수준 병렬화

스레드 수준 병렬화는 데이터 집합을 복수 개의 부분 데이터 집합으로 분할하고, 각 스레드가 분할된 부분 데이터 집합에 대한 데이터 연산을 수행하는 방법이다. 이 방법은 다시 CPU에서 사용하는 멀티코어 연산과 GPU에서 사용하는 GPGPU(General-Purpose computing on Graphics Processing Units) 연산으로 나뉜다. 멀티코어 연산에서는 복수 개의 데이터에 대해 각 프로세스가 하나 이상의 스레드를 수행시키는데 스레드가 수행하는 데이터 연산 명령어는 서로 다를 수 있다. 이처럼 멀티코어 연산에서 데이터를 처리하는 방식을 MIMD(Multiple Instructions Multiple Data)라 한다. GPGPU 연산에서는 멀티코어 연산과는 다르게 동일한 명령어에 대해 다수의 스레드가 수행되므로, GPGPU 연산에서 데이터를 처리하는 방식을 SIMT(Single Instruction Multiple Thread)라 한다.

스레드 수준 병렬화의 단점은 스레드 간의 균등한 작업 분배가 어렵다는 점이다. 이는 다른 스레드가 자신이 담당한 데이터에 대해 모든 연산을 수행했더라도 다수의 데이터가 집중된 스레드의 데이터 연산이 종료될 때까지 다른 모든 스레드들이 대기해야 하는 문제를 야기한다. 이런 단점을 보완하려고 데이터 분포에 따른 데이터 집합의 균등 분할(data partition), 스레드 이동(thread migration), 작업 스틸링(work stealing) 등의 기법이 등장했다. 또한 스레드 수준 병렬화에서 GPGPU 연산은 하드웨어 면에서 구조적인 문제점을 갖고 있다. GPGPU 연산에서는 CPU의 메인 메모리로부터 입력된 데이터 집합이 PCI(Peripheral Component Interconnect) 버스를 통해 GPU의 디바이스 메모리로 복사된다. 이 후 GPU의 스레드가 복사된 입력 데이터 집합에 대해 데이터 연산을 병렬로 수행한다. 때문에 입력 데이터 집합의 크기가 과도하게 크다면, PCI 버스에서의 데이터 병목 현상이 발생할 수 있다.

[그림 7-4-1] CPU와 GPU에서의 스레드 수준 병렬화 예시



최근 이런 데이터 병목 현상을 해소하기 위해, NVLink와 같이 기존 PCI 버스보다 최대 12배 빠른 고속 버스가 등장했다. 또한 CPU의 메인 메모리와 GPU의 디바이스 메모리 간의 데이터 복사를 수행하지 않고 GPU의 스레드가 직접 메인 메모리에 접근이 가능한 통합 가상 주소(UVA, Unified Virtual Addressing) 기술도 소개됐다.

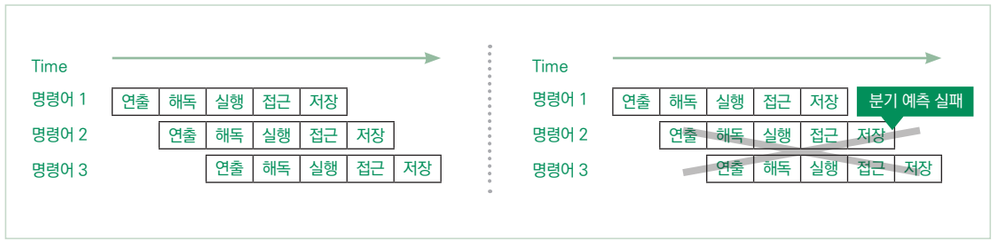
나. 명령어 수준 병렬화

명령어 수준 병렬화는 명령어 파이프라인 방식을 의미한다. CPU에서 여러 명령어를 처리할 때, 명령어를 순차적으로 실행시키는 것이 아니라 여러 명령어를 명령어 파이프라인에 적재해 병렬로 실행시키는 방법이다. 예를 들어 명령어를 동시에 실행시키기 위해 각 명령어를 다섯 단계(명령어 인출, 명령어 해독, 실행, 메모리 접근, 메모리 저장)로 나눠 순차적으로 실행시킨다. 이때 현재 실행되고 있는 명령어의 한 단계가 끝나면, 현재 실행되고 있는 명령어와 종속 관계가 없는 다른 명령어가 실행돼 동시에 여러 명령어들이 실행될 수 있다.

명령어 수준 병렬화의 단점은 분기 예측 실패(branch misprediction)에 따른 과도한 CPU의 명령 주기 낭비다. 분기 예측은 어떤 조건 분기에 대해 CPU가 다음에 수행할 명령어를 미리 명령어 파이프라인에 적재해 명령어 파이프라인에 시간 공백이 생기지 않도록 하는 기술이다.

문제는 분기 예측이 실패하면 명령어 파이프라인에 이미 적재된 명령어를 모

[그림 7-4-2] CPU에서의 명령어 수준 병렬화와 분기 예측 실패 예시

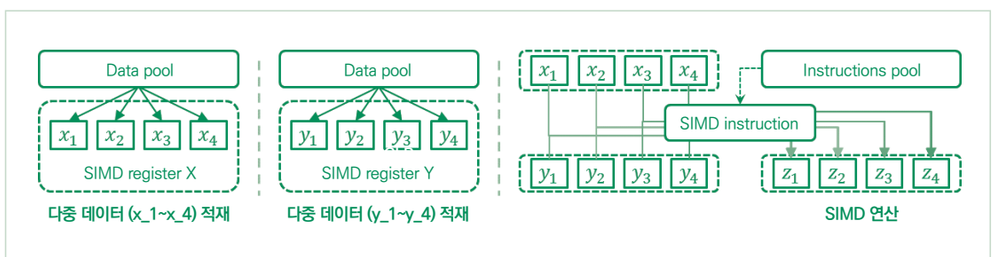


두 취소하고 올바른 명령어를 다시 명령어 파이프라인에 적재해야 한다는 점이다. 이때 과도한 CPU의 명령 주기를 낭비하는 문제가 발생한다. 이 문제를 해결하기 위해 어떤 조건 분기도 사용하지 않거나 조건 분기를 최소한으로 사용하는 비분기(branchless) 기법이 연구되고 있다.

다. 데이터 수준 병렬화

기존 데이터 처리 방식은 SISD(Single Instruction Single Data)로, 명령어 하나로 하나의 데이터만을 처리한다. 이에 반해 데이터 수준 병렬화는 명령어 하나로 복수 개의 데이터를 병렬로 처리하는 SIMD(Single Instruction Multiple Data) 방식이다. 즉 동일한 CPU 명령 주기를 수행하더라도 SIMD 방식은 SISD 방식에 비해 더 많은 데이터 연산을 수행할 수 있다. SIMD 방식에서는 복수 개의 데이터를 읽어와 벡터(vector)처럼 취급하고, SIMD 명령어 집합에서 존재하는 어떤 한 명령어로 다중 데이터를 병렬로 처리한다. 여기서 다중 데이터는 SIMD 레지스터라는 기억 장소에 적재된다. SIMD 레지스터의 크기는 CPU의 종류에 따라 다르다. 2018년 상반기 기준으로 그 크기가 최대 512비트이므로, 최대 정수형 데이터(32비트) 16개가 SIMD 레지스터에 적재돼 병렬로 처리할 수 있다. 인텔에서는 데이터 수준 병렬화

[그림 7-4-3] 데이터 수준 병렬화에서 SIMD 연산 예시



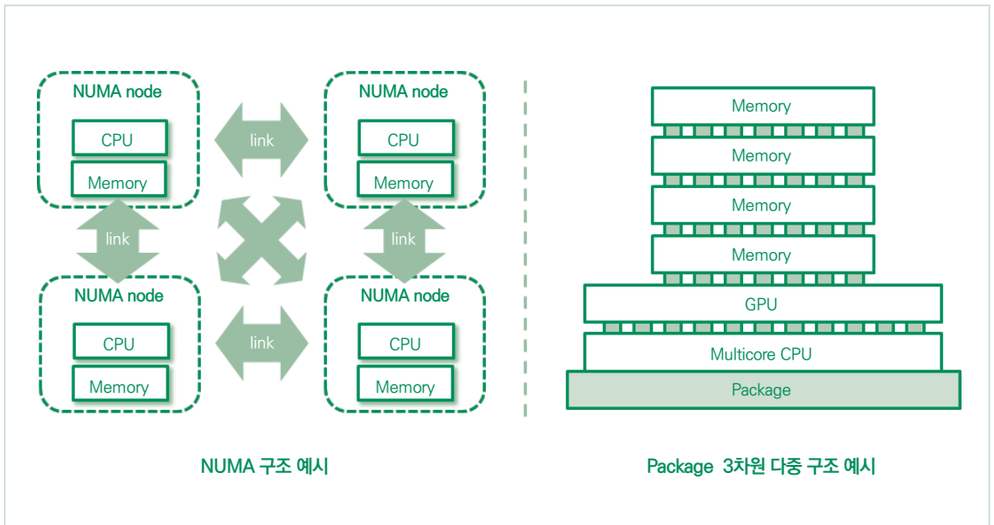
를 위해 SIMD 명령어 집합 AVX(Advanced Vector eXtensions)를 2008년부터 배포하고 있다. 이런 SIMD 명령어들은 기본적으로 분기 예측 실패를 회피하기 위해 비분기 기법으로 설계돼 있어 기존 명령어보다 성능이 뛰어나다.

데이터 수준 병렬화의 단점은 프로그램 코드가 하드웨어 특성에 종속되고 매우 복잡하다는 것이다. SIMD 레지스터의 크기, 개수와 SIMD 명령어들은 CPU의 종류에 따라 매우 다르다. 예를 들어 SIMD 레지스터의 크기가 256비트를 지원하는 CPU에서는 그에 맞는 SIMD 명령어만이 사용된다. 실례로 256비트를 지원하는 SIMD 명령어로 작성된 데이터 연산 프로그램은 64비트의 SIMD 레지스터를 지원하는 CPU에서는 수행되지 않는다. SIMD 방식은 비분기 기법을 지향하기 때문에 최소한의 반복문과 조건 분기문을 사용해야 하므로 프로그램 코드가 복잡해질 수 있다.

3. 앞으로의 병렬 구조

앞서 기술한 바와 같이, 병렬 구조는 CPU와 GPU를 기반으로 점차 발전하고 있는 추세다. 최근에는 NUMA(Non-Uniform Memory Access) 구조와 3차원 다층 구조처

[그림 7-4-4] NUMA 구조와 3차원 다층 구조 예시



럼 기존 병렬 구조를 탈피한 새로운 병렬 구조들이 설계되고 있다.

NUMA는 기존 SMP(Symmetric Multi-Processing)의 단점을 보완하기 위해 설계된 컴퓨터 구조다. SMP는 2개 이상의 CPU가 하나의 공유된 메인 메모리를 하나의 버스로 접근하기 때문에 데이터 병목 현상이 발생할 수 있었다. 이에 반해 NUMA 구조는 CPU와 지역 메모리로 구성된 NUMA 노드들로 구성된 컴퓨터 구조다.

NUMA 노드들은 물리적으로 분산돼 있고 NUMA 링크로 연결된다. 각 NUMA 노드에서 접근 가능한 다른 NUMA 노드의 지역 메모리를 원격 메모리라 한다. NUMA 구조에서 각 CPU는 자신의 지역 메모리와 하나의 버스로 연결돼 있다. 때문에 다른 CPU와의 경쟁이 SMP 구조보다 상대적으로 발생하지 않고 SMP 구조에서의 병목 현상도 발생하지 않는다.

NUMA 구조도 문제는 있다. 한 NUMA 노드의 CPU가 다른 NUMA 노드의 지역 메모리, 즉 원격 메모리에 접근하는 시간은 자신의 지역 메모리에 접근하는 시간보다 오래 걸린다. 이유는 원격 메모리를 지닌 NUMA 노드의 CPU에게 메모리 접근 권한을 할당 받아야 하기 때문이다.

3차원 다층 구조는 CPU와 GPU가 별도로 구성된 기존 병렬 구조의 단점인 병렬 연산의 성능이 CPU와 GPU 사이에 존재하는 버스의 대역폭에 종속된다는 문제를 해결하고자 설계됐다. 이 구조는 CPU, GPU, 메모리를 하나의 집적 회로에 집중시켜 다층 구조로 설계함으로써 CPU와 GPU 상의 데이터 전송에 따른 성능 저하 문제를 줄였다.

앞서 소개한 것과 같이 전자 기술의 발달에 따라서 병렬 데이터 처리를 지원하는 다양한 하드웨어 구조가 개발되고 있다. 이런 하드웨어 변화에 발맞춰 기존 데이터베이스에 적용했던 데이터 연산 기법도 병렬화해 데이터 처리 성능을 향상시키는 것이 필요할 것이다.

2018 데이터산업 백서 집필진

제1부 새로운 디지털 자원, 마이데이터

제1장 데이터경제, 데이터민주주의, 그리고 마이데이터 | **박주석** 경희대 경영학과 교수

제2장 새로운 디지털 자원, 마이데이터 | **이상일** 디지털데일리 기자

제3장 마이데이터, 개인정보 보호와 활용 | **이재욱** 법무법인 율촌 미국변호사

2017 데이터 구루상 수상자 칼럼 | **이형칠** 윕스 대표

제2부 데이터 정책 동향

제1장 데이터산업 활성화 전략 | **이종서** 한국데이터진흥원 차장

제2장 데이터 거래 기반 지원 정책 | **이재진** 한국데이터진흥원 실장

제3장 본인정보(MyData) 활용 지원 | **임태훈** 한국데이터진흥원 책임연구원

제4장 국내 의료데이터 활용 정책 동향 | **정일영** 과학기술정책연구원 부연구위원

제5장 국내 금융데이터 활용 정책 동향 | **고동환** 정보통신정책연구원 부연구위원

제3부 데이터산업 시장 현황

제1장 국내 데이터산업 현황 | **김초롱** 한국데이터진흥원 선임연구원

제2장 해외 데이터산업 현황 | **나영민** 날리지리서치그룹 이사

2017 데이터 글로벌 이노베이터상 수상자 칼럼 | **손삼수** 웨어밸리 대표

제4부 데이터 서비스 동향

제1장 데이터 거래 유형별 데이터 서비스 현황 | **김인현** 투이컨설팅 대표

제2장 주요 데이터 서비스 동향

1. 신용데이터 분야 서비스 | **이욱재** 코리아크레딧뷰로 본부장
2. 보건의료데이터 분야 서비스 | **김석일** 가톨릭대 예방의학교실 교수
3. 학술데이터 분야 서비스 | **박대광** 누리미디어 과장
4. 기업데이터 분야 서비스 | **이영준** 한국기업데이터 부장
5. 특허데이터 분야 서비스 | **강민수** 광개토연구소 대표

2017 데이터 서비스 이노베이터상 수상자 칼럼 | **박흥록** 와이즈모바일 대표

제5부 데이터 솔루션 동향

제1장 데이터 솔루션 아키텍처 | **장재웅** 알티베이스 대표

제2장 데이터베이스 솔루션 | **장재웅** 알티베이스 대표

제3장 데이터 관리 솔루션 | **이상옥** 데이터스트림즈 본부장·**박원환** 충남대 핀테크보안연구센터 교수

제4장 데이터 수집 솔루션 | **김한도** 이디엄 팀장

제5장 데이터 유통 솔루션 | **안정필** 투이컨설팅 이사

제6장 데이터 분석 솔루션 | **김종현** 위세아이텍 대표

제7장 데이터 플랫폼 솔루션 | **박시영** 데이터스트림즈 상무

2017 데이터 솔루션 이노베이터상 수상자 칼럼 | **이용재** 티맥스데이터 본부장

제6부 데이터 컨설팅 및 구축 동향

제1장 데이터 아키텍처 기반의 데이터 설계와 구축 | **서동재** 비투엔 수석컨설턴트

제2장 데이터 품질 컨설팅 | **이우용** 한국정보기술단 대표

제3장 빅데이터 구축 컨설팅 | **김형태** 비투엔 부사장

제4장 데이터 분석 컨설팅 | **김찬수** 투이컨설팅 상무

제5장 학습 지능 컨설팅 | **유태준** 마인즈랩 대표

2017 데이터 컨설팅 이노베이터상 수상자 칼럼 | **전혜경** 씨에스리 이사

제7부 데이터산업 기술 동향

제1장 텍스트 마이닝 기술의 소개와 애플리케이션 | **김영훈** 한양대 소프트웨어학부 교수

제2장 스트림 데이터 처리 시스템 | **송하주** 부경대 IT융합응용공학과 교수

제3장 실시간 인메모리 DBMS | **문양세** 강원대 컴퓨터학부 교수

제4장 병렬 구조와 데이터 연산 | **민준기** 한국기술교육대 컴퓨터공학부 교수

2018 데이터산업 백서 (통권 21호)

2018 Data Industry White Paper (Issue 21)

편찬위원 **김경민** 이화여대 교수(한국데이터전략학회 회장)
김상욱 한양대 교수(한국정보과학회 DB소사이어티 회장)
김인현 투이컨설팅 대표
박주석 경희대 교수
조광원 비투엔 대표(한국데이터산업협의회 회장)

편집위원 **박재현** 한국데이터진흥원 정책기획실 실장
이종서 한국데이터진흥원 정책기획실 차장
하진희 한국데이터진흥원 정책기획실 선임연구원

발행처 한국데이터진흥원
발행인 민기영
발행일 2018년 9월 28일
편집·디자인 글봄크리에이티브(02-507-2340)
인쇄 서울문화인쇄(02-2277-4650)
ISSN 2465-7662

• 본 백서는 과학기술정보통신부의 출연금으로 수행한 데이터산업 육성 지원 사업의 결과물입니다.

• 「2018 데이터산업 백서」와 관련한 문의는 한국데이터진흥원으로 연락해 주시기 바랍니다.

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