2021 Data Industry White Paper

2021 데이터산업 백서



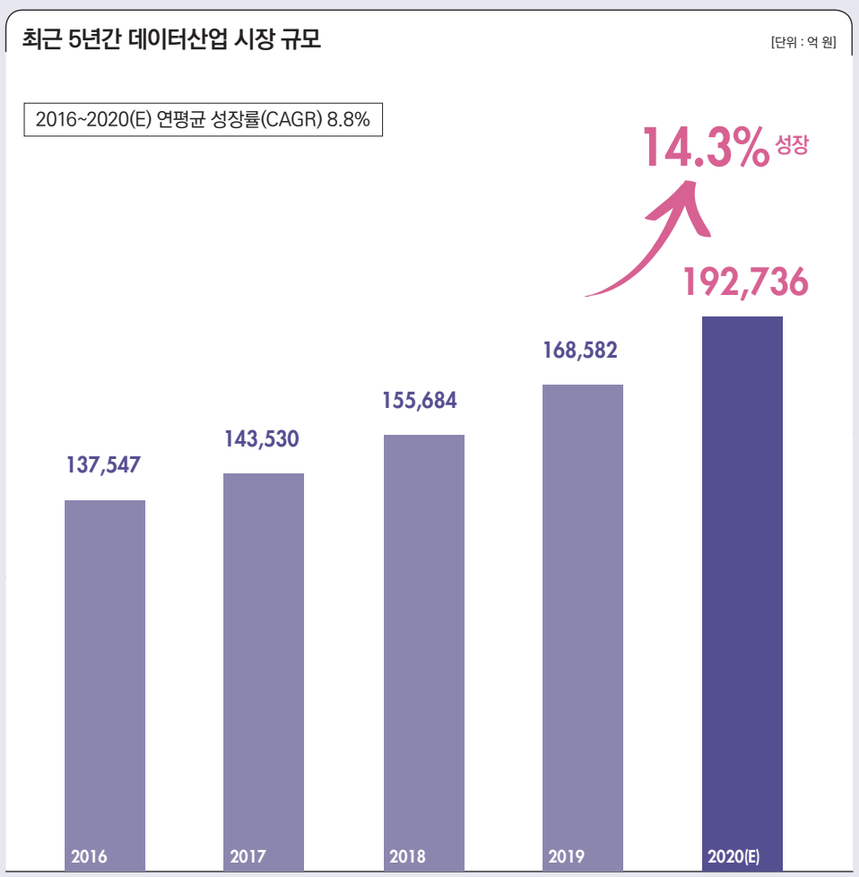
Kdata 한국데이터산업진흥원
Korea Data Agency

2020년 국내 데이터산업 시장 규모는 19조 2,736억 원, 전년 대비 14.3%의 가파른 성장세 예상

1. 전체 데이터산업 시장 규모

2020년 국내 데이터산업 전체 시장 규모는 2019년 16조 8,582억 원에서 14.3% 증가한 19조 2,736억 원으로, 2016~20년 연평균 성장률(CAGR) 8.8%의 1.6배를 넘는 성장세를 보였다.

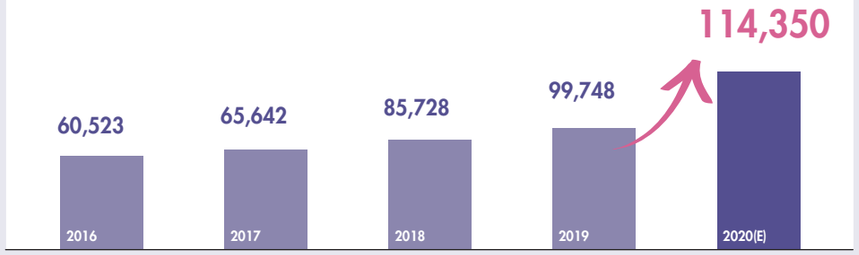
최근 5년간 데이터산업 시장 규모 [단위 : 억 원]



2. 데이터산업 직접매출 시장 규모

2020년(E) 기준 데이터산업 시장의 직접매출 규모는 2019년 대비 14.6% 성장한 11조 4,350억 원으로 추정된다.

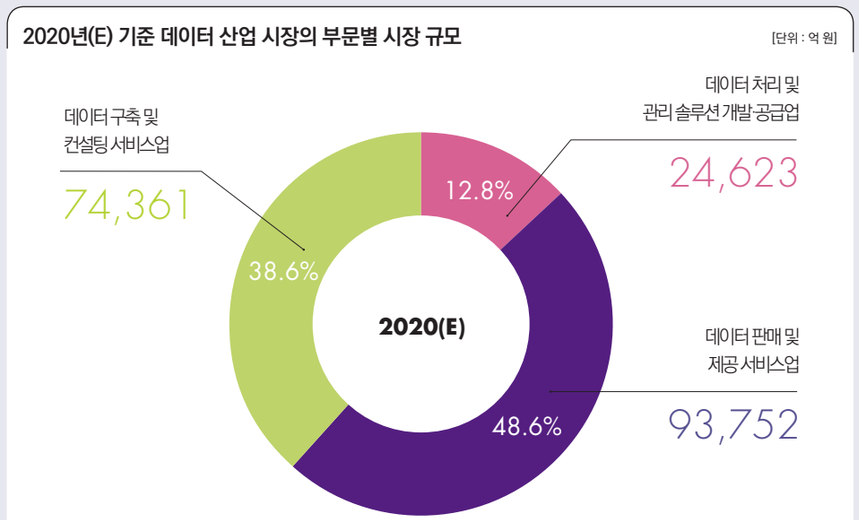
최근 5년간 데이터산업 시장의 직접매출 규모 [단위 : 억 원]



3. 부문별 데이터산업 시장 규모

2020년 부문별 시장 규모에서는 데이터 판매 및 제공 서비스업이 48.6%로 가장 많은 비중을 차지할 전망이다. 한편 전년 대비 가장 큰 성장률을 보인 부문은 데이터 처리 및 관리 솔루션 개발·공급업(18.3%)으로 나타났다.

2020년(E) 기준 데이터 산업 시장의 부문별 시장 규모 [단위 : 억 원]



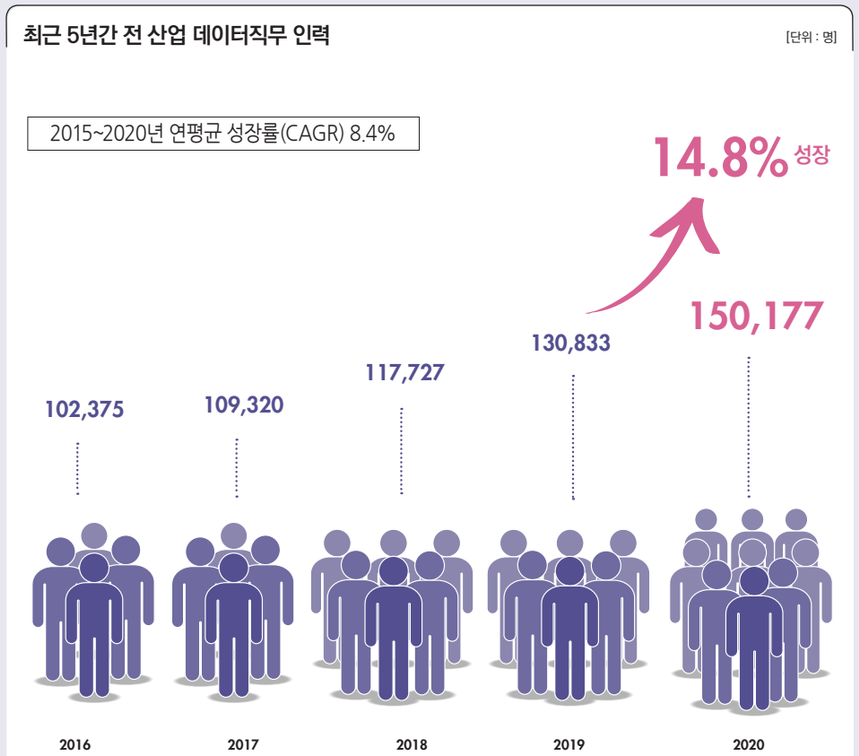
* 출처: 과학기술정보통신부, 한국데이터산업진흥원, 「2020년 데이터산업현황조사」, 2021.3.16.

2020년 전 산업의 데이터직무 인력은 15만 명 규모, 전년 대비 14.8% 증가 향후 5년 내 1.2만 명 추가 인력 필요 예측

1. 데이터직무 인력 현황

2020년 국내 전 산업의 데이터직무 인력 수는 총 15만 177명으로 전년 대비 14.8% 증가했다. 이는 2015~20년 연평균 성장률(CAGR)인 8.4%의 약 1.7배에 달한다. 이 중 데이터산업에 종사하는 데이터직무 인력은 총 10만 1,967명이며, 일반산업에 종사하는 데이터직무 인력은 총 4만 8,210명으로 나타났다.

최근 5년간 전 산업 데이터직무 인력 [단위 : 명]

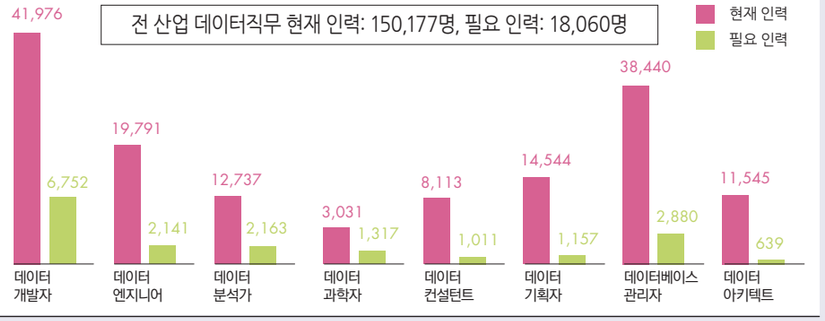


2. 데이터직무 인력 부족률

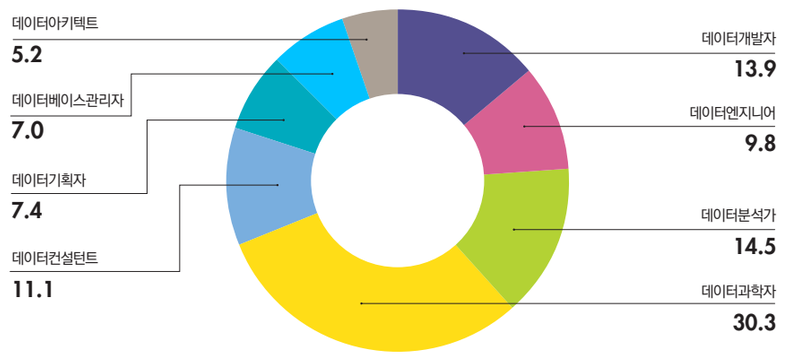
2021년 기준 향후 5년 내 국내 전 산업에서 필요로 하는 데이터직무 인력은 1만 8,060명으로 예상된다. 또한 향후 5년 내 전 산업의 데이터직무별 인력 부족률은 평균 10.7% 수준으로 전망된다.

이 중 데이터개발자의 필요 인력이 6,752명(37.4%)로 가장 높게 나타났고, 데이터베이스관리자 2,880명(15.9%), 데이터분석가 2,163명(12.0%) 순으로 나타났다.

2021년(E) 기준 전 산업의 데이터직무별 인력 현황 및 수요 [단위 : 명]



향후 5년 내 전 산업의 데이터직무 인력 부족률 [단위 : %]



* 출처: 과학기술정보통신부, 한국데이터산업진흥원, 「2020년 데이터산업현황조사」,2021.3.16.

국내 데이터산업 이슈 TOP 5

[국내 데이터산업 이슈]는 2021년 1월부터 2021년 10월까지 발생한 국내 데이터산업 이슈 1만 5,832건의 데이터를 바탕으로 분석되었다. 정부는 '디지털 뉴딜 2.0'을 발표하며 데이터와 밀접한 사업을 적극적으로 지원하겠다고 밝혔다. 정부의 대규모 투자 발표에 따라 관련 데이터산업 정책 마련 논의가 활발하게 진행되었으며 지역별 특성을 반영한 디지털 정책을 추진할 예정이다. 또한 정부의 마이데이터 본허가 및 빅데이터 플랫폼 구축 등 적극적으로 데이터 경제를 활성화하려는 움직임에 따라 관련 업권인 금융권·의료계의 사업 추진 또한 급속히 진행되었다.

국내 데이터산업 이슈 TOP 5

1. '데이터 산업진흥 및 이용촉진에 관한 기본법' 제정
2. 정부, 데이터산업 육성 정책을 포함한 '디지털 뉴딜2.0' 추진
3. 지자체, 데이터 활용한 '디지털 지역 균형 뉴딜' 정책 수립 및 실행
4. 마이데이터 사업 승인 확대, 금융산업 등 활용 본격화
5. 정부, 의료 빅데이터 플랫폼 구축 등을 통한 미래 의료 혁신 본격 추진

이슈1. '데이터 산업진흥 및 이용촉진에 관한 기본법' 제정

'데이터 산업진흥 및 이용촉진에 관한 기본법' 제정, 2022년 4월부터 시행

2021년 10월 12일 데이터 산업 발전 기반을 조성하고, 데이터 경제 활성화를 목적으로 한 '데이터 산업진흥 및 이용촉진에 관한 기본법(이하 데이터 기본법)'이 국무회의에서 의결되었다. 데이터 기본법은 지난해 12월 발의되어 올해 9월 과학기술정보방송통신위원회와 국회 본회의 의결을 통해 국무회의에 상정되었으며 내년 4월부터 시행될 예정이다. 데이터 기본법은 정부 차원의 데이터 컨트롤 타워 설치, 데이터 전문기업 육성 지원, 데이터 거래사 및 전문인력 양성, 데이터 자산가치·권리·이동권이 보장되는 시장 조성 및 관리 기구 설치 등을 주요 골자로 한다.

[표 1] 데이터 산업진흥 및 이용촉진에 관한 기본법 주요 내용(2021년 10월 기준)

구분	주요내용
데이터 산업진흥 기본계획 수립	3년마다 데이터 산업진흥 기본계획을 수립하여 데이터의 생산·거래·활용 촉진 및 데이터 산업 기반 조성
국가 데이터 정책 위원회 설치	데이터 산업진흥 기본계획을 총괄하는 국가 데이터 정책 위원회 설치
데이터 자산 보호	데이터의 무단 취득·사용·공개, 데이터 자산에 적용한 기술적 보호 조치 제거 등 금지
데이터 가치 평가 지원 및 품질관리	데이터 가치 평가 및 품질관리 기준 마련 및 관련 업무 전담할 기관 지정
데이터 사업자 신고	데이터 거래·분석 제공 사업자에 대한 신고제 도입과 함께 재정적·기술적 지원을 제공
데이터거래사 양성 지원	데이터 거래에 관한 상담, 거래, 알선 등을 수행하는 '데이터 거래사'의 등록 및 관련 교육 제공
창업 지원, 중소기업자 특별지원	데이터 산업 지원시책 시행 시 중소기업자 및 창업자 우선 고려 및 데이터 거래, 가공 등 필요 비용 일부 지원
전문인력 양성	데이터 전문인력 양성을 위한 시책 마련 및 양성기관 지정·지원
데이터분쟁 조정위원회 설치	데이터의 생산, 거래 및 활용에 관한 분쟁 조정을 위한 '데이터분쟁조정위원회' 설치

* 출처: 과학기술정보통신부(www.msit.go.kr, 접속일: 2021.11.3.)

데이터 기본법에 포함된 내용 중 데이터 이동권은 9월 28일 국무회의에서 통과된 '개인정보보호법 개정안'에서도 다루고 있으며, 이를 통해 정보 주체의 개인정보 통제권 및 데이터 자산 보호를 강화할 예정이다. 이외에도 개정된 개인정보보호법에서는 개인정보에 대한 국민의 권리 강화, 관련 법체계 정비, 글로벌 규제와의 정합성 확보 등을 포함하고 있다. 이러한 정책 및 법적 제도 정비를 통해 정부는 데이터 생태계를 조성하고 관련 산업을 육성하여 글로벌 경쟁력을 확보한다는 계획이다.

이슈2. 정부, 데이터산업 육성 정책을 포함한 '디지털 뉴딜 2.0' 추진

'디지털 뉴딜 2.0'으로 2025년까지 49조 원 이상 투자

2021년 7월, 과학기술정보통신부 및 관계부처는 데이터산업 육성 정책을 포함한 '디지털 뉴딜'의 성과와 향후 주요 정책 방향을 발표했다. 기존의 '디지털 뉴딜 1.0' 추진의 성과로는 ① 데이터 댐 확대 ② 데이터 시장 규모 16.8조 원 달성 ③ AI 학습용 데이터 191종 구축 ④ 빅데이터 플랫폼 및 센터 구축과 데이터산업의 국가 경쟁력 강화 및 국제 지표 순위 상승을 밝혔다. 위 성과를 바탕으로 정부는 2022년 '디지털 뉴딜 2.0'에서 데이터산업 육성을 재차 강조하였으며, 이를 위해 2022년 데이터 댐 구축사업에 디지털 뉴딜 예산의 51.7%인 1조 4,642억 원을 배정하였다.

[표 2] 디지털 뉴딜 2.0 추진 방향

항목	추진 내용
1	구축된 '디지털 인프라'를 바탕으로 '민간 주도의 활용 강화'
2	'중점 프로젝트'의 성과를 전 지역·전 산업으로 '디지털 확산'
3	'내수 진작'에 더해 '세계 진출·신산업 전략적 육성'

* 출처: 대한민국 정책브리핑(www.korea.kr, 접속일: 2021.11.3.)

금번에 새로 발표된 '디지털 뉴딜 2.0'은 2025년까지 49조 원 이상을 투자해 ① D.N.A(-DATA, Network, AI) 생태계 강화 ② 비대면 인프라 고도화 ③ 초연결 신산업 육성 ④ SOC 디지털화 4가지 주요 과제를 달성한다는 계획이다.

D.N.A 생태계 강화는 국민생활과 밀접한 분야의 데이터 구축·개방·활용을 바탕으로 진행될 예정이다. 해당 사업에는 공공데이터 개방, 공동 빅데이터 분석 시스템 구축, 바이오·축산·제조 데이터 플랫폼 구축, 마이데이터 활용 지원, 데이터 경제 활성화를 위한 법적 기반 마련 및 제도 개선 등이 포함되어 있다. 이외에도 인공지능 학습용 데이터 4.8억 건 구축, 데이터 댐 활용·가명정보 결합 강화, 5G 국가망 구축 및 특화망 활용 강화, AI+X 7대 융합 프로젝트 및 AI 윤리기준 마련, 지역 주력산업-AI 연계 대형 선도사업 등이 추진될 예정이다.

[표 3] '디지털 뉴딜 1.0' 주요 목표와 '디지털 뉴딜 2.0' 주요 목표

항목	디지털 뉴딜 1.0		디지털 뉴딜 2.0
D.N.A 생태계 강화	대규모 데이터 댐 구축	→	데이터 댐 전주기 활용 강화
	5G 등 네트워크 기반 구축		5G 특화망 등을 통해 산업 융화 강화
	AI 활용 기반 구축		AI 활용을 산업·지역 전반으로 확산
비대면 인프라 고도화	교육 인프라 디지털 전환		디지털 교육 인프라 강화
	스마트 의료 기반 조성		스마트 의료 성과 확산
	중소·소상공인 비대면 전환 지원		중소·소상공인 지원 내실화
초연결 신산업 육성	VR·AR 산업 육성		메타버스 산업 생태계 조성 지원
	클라우드 산업 육성		민간 주도 클라우드 산업 성장 지원
	블록체인 등 시범·실증사업		블록체인 등 대규모 확산 프로젝트
	주요 분야별 핵심 기술 개발		차세대 미래 핵심 기술 개발
SOC 디지털화	디지털 트윈 기반 국토 구축		사회·산업 전 영역의 디지털 트윈 확산
	취약계층 등의 디지털 접근성 강화		지역주민 일상의 디지털 활용성 강화

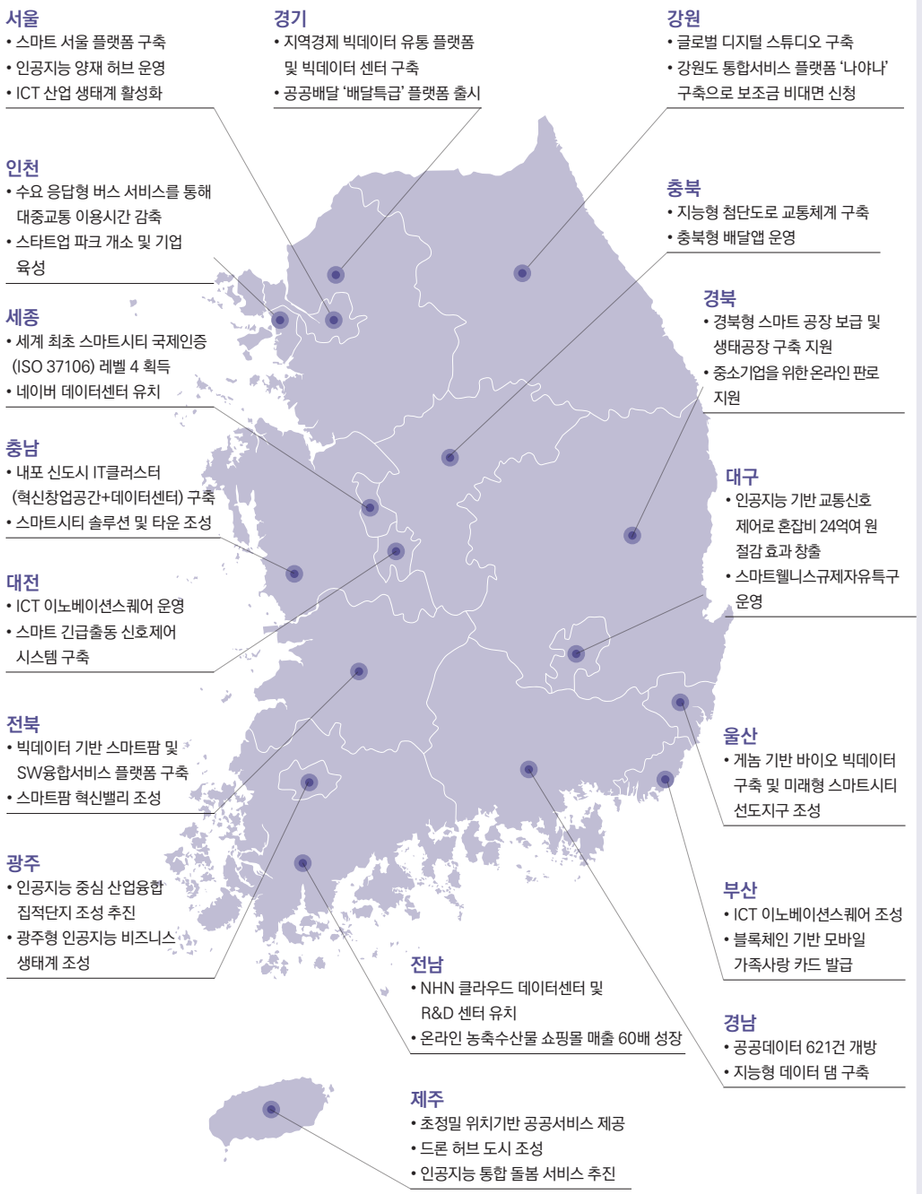
* 출처: 대한민국 정책브리핑(www.korea.kr, 접속일: 2021.11.3.)

이슈3. 지자체, 데이터 활용한 '디지털·지역 균형 뉴딜' 정책 수립 및 시행

주요 시·도, 데이터 기반 '디지털 산업 육성' 노력

각 지자체는 '디지털 뉴딜 1.0'을 바탕으로 데이터를 활용한 다양한 정책을 수립 및 실행하고 있다. 대표적으로 강원도의 경우 정밀의료 빅데이터 플랫폼을 구축하고 있으며, 경상북도는 데이터를 활용한 디지털 물류 서비스 사업이 추진 중이다. 세종시의 경우 주행

[그림 1] 주요 시·도 '디지털 산업 육성' 추진 성과



* 출처: 대한민국 정책브리핑(www.korea.kr, 접속일: 2021.11.3.)

데이터를 기반으로 한 자율주행 서비스 상용화 사업이 진행 중이며, 포항은 민관협력을 통한 포항형 빅데이터 프로젝트를 추진하고 있다.

과학기술정보통신부는 7월 21일 '제2차 디지털-지역균형 뉴딜 연석회의'를 개최하여 데이터 활용을 통한 디지털 산업 육성을 포함한 '디지털 뉴딜 1.0' 성과를 공유하고 지역 균형 발전에 근거한 '디지털 뉴딜 2.0' 실현 방안을 논의하였다.

이슈4. 마이데이터 사업 승인 확대, 금융산업 등 활용 본격화

금융위, 58개사에 마이데이터 사업 승인

금융위원회는 10월 13일 제18차 정례회의를 통해 본인신용정보관리업(마이데이터) 허가심사 결과, 본허가 신청 및 신용정보법령상 허가요건을 구비한 디셈버앤컴퍼니자산운용에 대한 본허가를 승인했다. 이로써 올해 7월 5개사였던 마이데이터 본허가 및 예비허가 사업 승인사는 10월 기준 58개사로 3개월 만에 53개사가 증가하였다. 이후에도 금융위원회는 신속한 심사절차를 통해 매월 심사 접수를 진행하겠다고 밝혔다.

이렇듯 마이데이터를 활용한 사업이 본격 승인됨에 따라, 금융기관·통신사·보험사 등 관련 기업들은 제공받은 개인정보를 통합·관리하는 서비스를 포함해 다양한 고객 맞춤형 서

[그림 2] 금융업계의 마이데이터 활용 방안



* 출처: 국내 데이터 뉴스매체 발행 이슈 15,832건 분석(2021.1.~2021.8.)

비스를 제공할 수 있게 되었다. 이러한 마이데이터 사업 확대는 향후 데이터 경제 전반 활성화에도 기여할 것으로 전망된다.

마이데이터 지정 업권 중에서도 은행권은 가장 다양하고 범용적인 서비스를 제공할 것으로 예상된다. KB국민은행은 금융데이터를 기존 자산관리서비스에 적용하여 신용관리 및 자동차 관리 서비스를 새롭게 출시했다. 이외에도 신한은행은 기존 자산 관리 서비스를 더욱 고도화할 예정이며, NH농협은행은 신용점수 조회, 맞춤상품 추천, 자산분석 등 개인 종합 자산관리 서비스를 제공할 계획이다.

[표 4] 주요 금융기업, 마이데이터 활용 계획 및 현황

기업명	활용 방안
KB국민은행	'KB마이머니'와 마이데이터 사업을 결합한 신용·자동차 관리 등 생활밀착 서비스
신한은행	전 금융기관의 상품 정보 데이터베이스화를 통한 고객 맞춤형 자산관리 서비스
NH농협은행	계좌 및 신용점수 조회·맞춤상품 추천, 자산 분석 등 개인종합 자산관리 서비스
교보생명	인공지능 음성합성 기술 및 마이데이터 사업을 결합한 생애 설계 자산관리 및 재무 컨설팅
신한라이프	보험·금융·카드·증권 등 금융 거래 데이터 융합을 통한 재무 설계, 상품 분석
메리츠화재	공공의료데이터 활용 승인 받아 헬스케어 및 마이데이터 사업 결합
토스	금융기관 데이터를 이용한 맞춤형 상품 추천 서비스 고도화

* 출처: 국내 데이터 뉴스매체 발행 이슈 15,832건 분석(2021.1 ~ 2021.8)

이슈5. 정부, 의료 빅데이터 플랫폼 구축 등을 통한 미래 의료 혁신 본격 추진

보건의료데이터·인공지능 혁신전략 추진…미래 의료 혁신 본격화

정부는 2021년 의료부문 데이터 활용을 위한 '보건의료데이터·인공지능 혁신전략' 수립을 비롯한 의료데이터 중심병원 지원사업, 의료데이터 분석 플랫폼 임시허가 승인, 정밀의료 SW 선도계획 등 다각화된 정책 사업을 추진 중이다. 이러한 정책은 데이터 기반 바이오헬스 경쟁력을 확보하고 미래 의료 혁신을 본격 추진하기 위해 마련되었다.

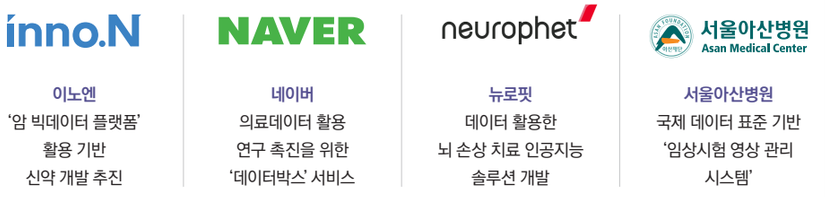
[표 5] 의료부문 데이터 활용 정책

구분	주요 현황
2021.1	**보건복지부, '의료데이터 중심병원 지원사업' 확대 추진** • 의료 빅데이터를 공개·연결하여 의료서비스 개선 및 공동연구 활성화 추진 • 25개 병원 및 38개 민간기업 참여, 대규모 임상데이터 기반 빅데이터 활용 추진 • 의료 빅데이터의 2차 활용을 위한 데이터 저장소 구축, 데이터 표준화 및 특화 데이터베이스 구축, 데이터 활용 파이프라인 체계화
	과학기술정보통신부, 정보통신기술 규제 샌드박스 심의·임시 허가 • 의료데이터 활용 활성화 및 신약 개발·의료연구 효율성 제고 전망 • 에비드넷은 각 의료기관 내 의료데이터 분석 플랫폼 구축, 통계값 추출하여 연구소·제약사·의료기관에 정보 서비스 제공
2021.6.	**보건복지부, '보건의료데이터·인공지능 혁신전략 2021~ 2025' 수립** • 데이터 기반 바이오헬스 경쟁력 확보 및 미래 의료 혁신 본격 추진 • 보건의료데이터 품질관리 인증제 도입·고품질 데이터 생산 체계 구축 • 건강보험 등 핵심 공공의료데이터 개방 및 신뢰할 수 있는 데이터 거버넌스 확립 • 빅데이터로 만드는 'K · Cancer' 사업 추진 • 2025년까지 의료데이터 중심의 3대 원천 데이터 플랫폼 구축 목표 • 개인생성건강데이터(PGHD) 표준화 및 희귀질환·암 등 100만 명 통합 바이오 빅데이터 구축
2021.8.	**비상경제 중앙대책본부, '한국판 뉴딜 2.0 실현을 위한 정밀의료SW 선도계획' 발표** • 디지털 의료 핵심사업은 클라우드 병원정보시스템, 닥터앤서, AI 앰뷸런스 등 • 4차산업혁명의 핵심인 D.N.A의 의료서비스 디지털화 추진 • AI 의료 SW 개발을 위한 데이터 댐 및 닥터앤서 2.0 데이터 공개 • 임상·검진·사망에 이르는 데이터 연계 암 임상 네트워크 지원

* 출처: 국내 데이터 뉴스매체 발행 이슈 15,832건 분석(2021.1.~2021.8.)

한편 정부의 적극적인 정책 추진에 힘입어 기업 또한 의료데이터를 활용한 신약 개발·질병 진단 및 치료·의료 플랫폼 구축 등 다양한 의료데이터 사업 투자 및 개발을 진행 중이다. 향후 의료데이터는 검사 비용 절감, 질병 조기 발견 등 의료 비용 절감 효과와 더불어 정확하고 효과적인 의료서비스 제공에 큰 역할을 담당할 것으로 전망된다.

[그림 3] 의료부문 데이터 활용 분야



*출처: 국내 데이터 뉴스매체 발행 이슈 15,832건 분석(2021.1 ~ 2021.8)

해외 데이터산업 이슈 TOP 5

[해외 데이터산업 이슈]는 2021년 1월부터 2021년 8월까지 발생한 해외 데이터산업 이슈 2만 3,850건의 데이터를 바탕으로 분석되었다. 전 세계적으로 시행 및 제정이 가속화되고 있는 데이터 보호법에 대한 이슈가 주를 이뤘으며 이외에도 디지털 거래 증가에 따른 '데이터 이동' 에 대한 논의도 활발하게 이어졌다. 또한 미국, 중국 EU는 거대 IT 기업의 데이터 독점에 대한 규제 필요성을 인식하고 이들 기업에 대한 반독점 규제 정책을 발의·시행하였다. 데이터산업의 전반적인 성장으로 데이터센터 시장 또한 성장이 재개될 전망이며, 자동차 산업에서의 데이터 활용이 다각화될 전망이다.

해외 데이터산업 이슈 TOP 5

1 주요국, 데이터 보호법 시행 및 제정 가속화

2 디지털 거래 증가에 따른 국가 간 '데이터 이동' 논의 시동

3 주요국, 거대 IT기업에 대한 데이터 반독점 규제 강화 추진

4 글로벌 데이터센터 시장, 대형 테크 기업 중심으로 성장 재개

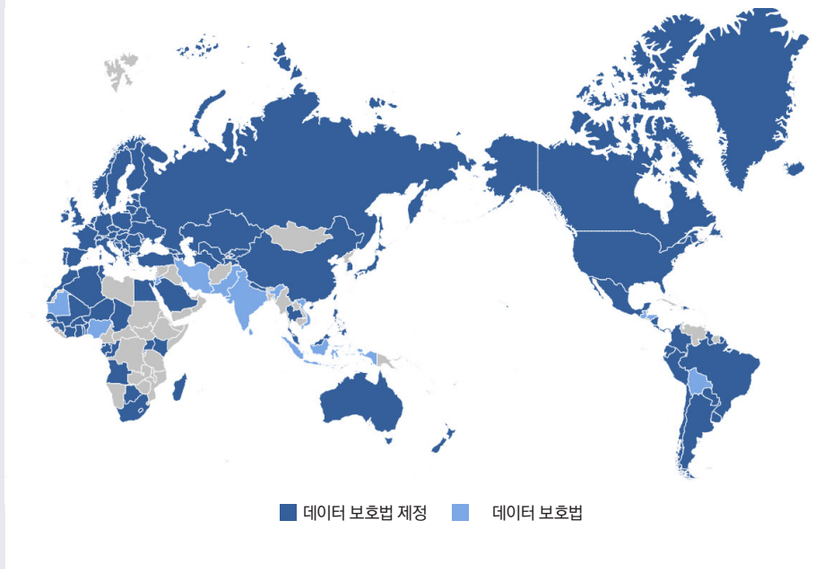
5 자율주행, 데이터 발전의 최대 수혜 기술로 부상

이슈1. 주요국, 데이터 보호법 시행 및 제정 가속화

유럽연합 일반개인정보보호법(GDPR), 글로벌 기준으로 자리매김

개인정보 및 데이터 활용이 증대됨에 따라 데이터 유출에 대한 우려 및 데이터 보호에 대한 필요성이 대두되고 있다. 2018년 5월 25일 발효된 유럽연합 일반개인정보보호법(General Data Protection Regulation, GDPR)은 글로벌 데이터 및 개인정보보호 법안에 기본 가이드라인이 되었으며, 이후 중동·아프리카·북아메리카 지역 국가를 비롯해 일본, 대한민국 등 16개 국가가 유사한 법안을 제정하였다.

[그림 4] 국가별 데이터 보호법 제정 현황(2021년 9월 기준)



* 출처: Data Guidance(www.dataguidance.com, 접속일: 2021.11.3.)

[표 6] 주요국 데이터 보호법 시행 및 제정 현황(2021년)

국가	현황	주요 내용
미국	캘리포니아 소비자 개인정보 보호법(California Consumer Privacy Act, CDPA) 시행	개인정보 수집은 소비자에게 공개된 바에 따라 합리적이고 필수적인 것에 한하며 데이터의 기밀성, 무결성, 접근성 보호를 위해 필요한 기술적, 물리적 보호 관행을 수립·유지하여야 함
중국	중화인민공화국 데이터안전법(中华人民共和国数据安全法) 제정	데이터 보안 및 개발, 데이터 보안 시스템, 데이터 보안·보호 의무, 정부 데이터의 보안·개방성·법적 책임 등 다양한 관점에서 데이터 보안·보호 의무 및 법적 책임을 규정함
EU	전자 프라이버시 규정(ePrivacy Regulation) 개정	전자 통신 데이터는 기밀이어야 하며 최종 사용자의 통신 데이터 감시 및 처리는 해당 법안에서 허용하는 경우를 제외하고 모두 금지함
일본	개인정보보호법(Act on the Protection of Personal Information, APPI) 개정	데이터 침해가 '개인의 권리와 이익을 해칠 가능성이 있는' 경우 지정된 시간 내에 관련 당국에 보고하고 영향을 받는 개인에게 적시에 알려야 함
베트남	개인데이터보호 관한 법령(Data Privacy Decree) 초안 발표	개인데이터 규정, 개인데이터 처리, 개인데이터 보호 조치, 개인데이터보호위원회, 개인데이터 침해 처리, 관련 기관, 조직 및 개인의 개인데이터 보호 책임을 규정함
인도네시아	개인데이터보호법안(Perlindungan Data Pribadi, PDP) 논의	데이터 주체가 자신의 데이터 사용 동의를 철회할 경우 이를 삭제해야 하며 컨트롤러는 저장하고 있는 개인데이터 보호를 위해 필요한 기술적 조치를 취해야 함

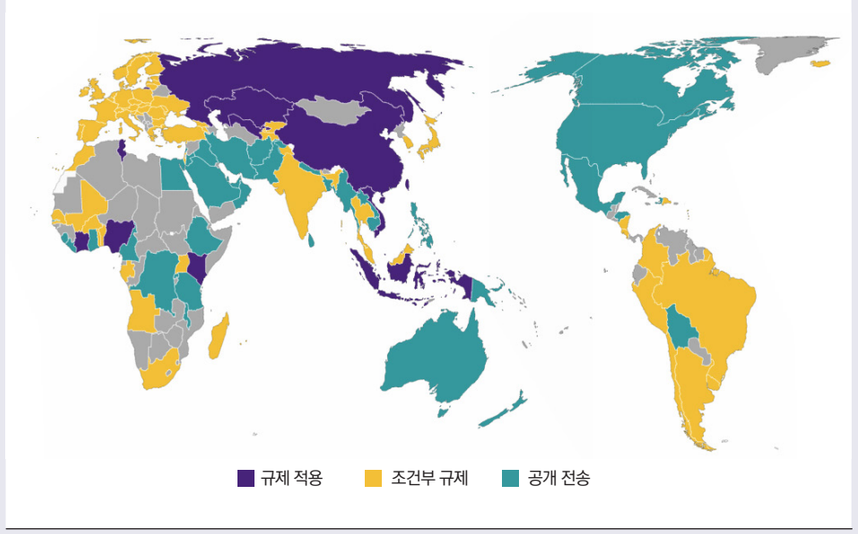
* 출처: 해외 데이터 뉴스매체 발행 이슈 23,850건 분석(2021.1 ~ 2021.8.)

이슈2. 디지털 거래 증가에 따른 국가 간 '데이터 이동' 논의 시동

EU, 제3국으로의 데이터 전송에 관한 가이드라인 발표

2021년 말까지 글로벌 GDP의 60%가 디지털화될 것으로 전망됨에 따라, 디지털 거래의 주된 요소 중 하나인 데이터 이동 및 전송에 관한 논의가 활발하게 진행되고 있다. 이러한 추세 속에서 2021년 6월 유럽 데이터보호위원회(European Data Protection Board, EDPB)는 디지털 거래 증가 및 데이터 이동 활성화를 위해 '제3국으로의 데이터 전송에 관한 최종 가이드라인'을 발표하여 데이터 이동에 관한 권장사항을 규정·강조했다. 해당

[그림 5] 국가별 데이터 이동 접근 방식



* 출처: World Bank(https://wdr2021.worldbank.org/stories/crossing-borders/, 접속일: 2021.11.3.)

가이드라인에서는 데이터 적정성 협정을 체결하지 않은 제3국으로의 데이터 전송은 일부 제한될 수 있음을 명시하고 있다.

주요국들은 디지털 거래 및 데이터 이동 활성화를 목표로 상호협력 방안을 마련하기 위한 논의를 진행하고 있다. 특히 EU의 경우 올해 들어 미국, 영국, 한국 등 데이터 산업 주요 국가들과 데이터 이동 관련 논의를 진행 중이다.

[표 7] 미국·EU, EU·영국, EU·한국, 아세안 국가 간 데이터 이동 논의 현황(2021년)

국가	논의 현황
EU·미국	• 6월 유럽 데이터보호위원회(EDPB), 미국을 포함한 데이터 적정성 협정을 체결하지 않은 제3국에 대한 데이터 전송 제한 지침 발표
EU·영국	• 6월 17일 유럽연합 집행위원회, 영국의 개인정보 보호 수준의 적정성 선언 • 개인데이터 전송 체제의 자유화 허용
EU·한국	• 3월 30일 유럽연합 집행위원회, 한국의 개인정보 보호 수준의 적정성 선언 • EU 일반개인정보보호법(GDPR)에 따라 추가 보호 조치 없이 EU 및 유럽경제지역(EEA)에서 한국으로의 개인데이터 전송 자유화(개인신용정보 제외) 허용
아세안	• 1월 22일 '제1차 아세안 디지털 장관회의', 국경 간 데이터 이동 모델 계약 조항 채택 • 자발적 성격의 규제로, 국가 간 데이터 이동 시 법률적 효력 발생 명시

이슈3. 주요국, 거대 IT 기업에 대한 데이터 반독점 규제 강화 추진

거대 IT 기업의 디지털 독점 및 권력 집중 우려 대두…규제 필요성 증가

거대 IT 기업의 시장 진출이 활성화되면서 각국에서는 데이터 보호, 콘텐츠 및 미디어 규제, 시장 독점의 문제 등 경제사회에 미치는 부정적 영향에 대한 우려가 높아지는 가운데 거대 IT 기업에 대한 규제 필요성을 인식하여 관련 정책을 논의·수립하고 있다.

[표 8] 주요국 빅테크 관련 규제 정책 논의 현황

국가	주요 내용
미국	**2021년 6월 11일 미국 의회, '빅테크 독점 권한 제재를 위한 5개 법안' 발의** • 아마존, 애플, 구글, 페이스북 등 주요 빅테크 기업 적용 • 데이터 독점, 시장 조작, 인수합병, 플랫폼 내 불공정 경쟁, 과다한 수수료 책정 등 빅테크의 불공정 독점 제한 및 플랫폼 독점 종식을 위한 법안
중국	**2021년 7월 30일 중국 공업정보화부, '빅테크 기업 및 인터넷 산업 전담 단속' 시행** • 반독점·부당 경쟁, 노동자 수익 보장 등 규제 강화 및 시정명령 조치 • 중국 내 개인정보 및 데이터 유출 방지 및 관련 규제 강화 • 데이터 안보 보장, 시장질서 교란 및 이용자 권익 침해 방지 추진 • 텐센트(Tecent) 온라인 음원 독점 판권 포기 명령
EU	**2020년 12월 EU집행위원회, '디지털 서비스법(DSA)', '디지털 시장법(DMA)' 발의** • 핵심 디지털 플랫폼의 서비스 행동 규제에 관한 디지털 산업 법안 • 데이터 공유·데이터 투명성·광고 및 콘텐츠 규정 포함 • 빅테크 기업의 불공정 경쟁 및 시장 지배 제재 • 데이터 투명성 및 소비자 보호 강화

* 출처: 해외 데이터 뉴스매체 발행 이슈 23,850건 분석(2021.1 ~ 2021.8.)

[표 9] 거대 IT 기업 주요 규제 현황(2021년)

기업	주요 내용
아마존 (Amazon)	**룩셈부르크 데이터보호 당국, 7월 16일 일반개인정보보호법(GDPR) 위반 혐의로 8억 8,660만 달러 벌금 부과** • 개인정보 처리 과정에서 일반개인정보보호법(GDPR) 위반 혐의
페이스북 (Facebook)	**유럽연합 집행위원회, 6월 4일 광고 시장 경쟁 왜곡 혐의로 조사 착수** • 데이터 이중독점 및 EU 시장경쟁 규칙 위반 혐의로 조사 착수
구글 (Google)	**프랑스 국가정보처리자유위원회, 6월 7일 데이터보호법 위반 혐의로 2억 2,000만 유로 벌금 부과** • 경쟁사 온라인 검색 광고주 차단으로 인한 반독점 규제법 위반 혐의
디디추싱 (Didi Chuxing)	**개인정보 수집 및 이용에 관한 위반 행위 혐의로 해당 애플리케이션 운영 금지 처분** • 중국 사이버보안관리국, 7월 4일 개인정보 수집 위반 혐의 적발
알리바바 (Alibaba)	**중국 사이버보안관리국, 4월 10일 반독점법 위반 혐의로 180억 위안 벌금 부과** • 불공정 경쟁 및 반독점 규정 위반 혐의
애플 (Apple)	**유럽연합 반독점규제 기관, 4월 27일 반독점 혐의 적발** • 경쟁사 진입 제한으로 인한 유럽연합 독점금지 위반 혐의 적용

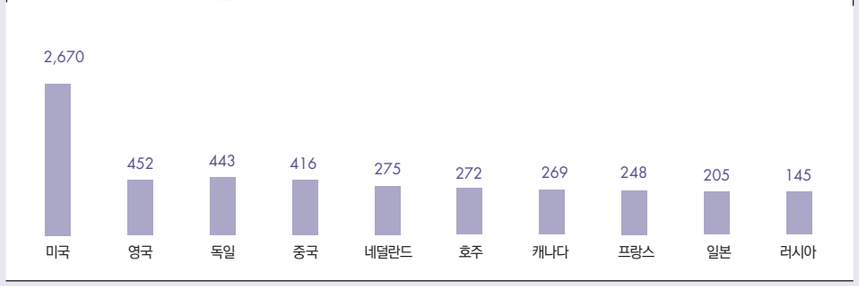
*출처: 해외 데이터 뉴스매체 발행 이슈 23,850건 분석(2021.1 ~ 2021.8.)

이슈4. 글로벌 데이터센터 시장, 대형 테크 기업 중심으로 성장 재개

데이터센터 시장 2025년 4,321억 4,000만 달러로 성장 전망

코로나19 확산으로 둔화되었던 데이터센터 시장이 글로벌 테크 기업을 중심으로 재성장하고 있다. 글로벌 데이터센터 시장은 2025년까지 연평균 성장률 9.9%를 기록하며 4,321억 4,000만 달러 규모를 달성할 것으로 전망된다.[1] 미국과 중국은 하이퍼스케일 및

[그림 6] 국가별 데이터센터 수



* 출처: Statista(www.statista.com, 접속일: 2021.10.31.)

1 Research And Markets, 「Increased Investment by Cloud and Colocation Providers Drives the Global Data Center Market」, 2021.1.

엣지 데이터센터에 적극 투자하고 있으며 스마트 데이터센터 기술에 대한 글로벌 선도 지위를 놓고 경쟁 중이다. 미국은 데이터센터 세금 감면 및 면세 인센티브 등 데이터센터 투자 및 유치를 위한 신규 법안이 마련되고 있으며 인도네시아는 동남아시아의 데이터센터 허브를 목표로 글로벌 기업과의 협력을 적극 추진하고 있다.

한편 최근 글로벌 IT 기업은 데이터센터의 원격 관리 전환, 친환경·고효율 데이터센터 운영 등을 기조로 신규 데이터센터 구축을 적극적으로 추진하고 있다.

[표 10] 글로벌 데이터센터 구축 계획

기업	설립 위치	주요 내용
페이스북 (Facebook)	미국 애리조나	• 300만㎡에 8억 달러 규모의 친환경 데이터센터 건립
	미국 오리건	• 총 20억 달러 투자 규모, 기존 데이터센터 9만m2 추가 확장
구글 (Google)	미국 19개 주	• 미국 19개 주 대상으로 70억 달러 투자(데이터센터, 사무실 포함)
	미국 사우스캐롤라이나	• 기존 데이터센터 확장 및 클라우드 컴퓨팅 사업에 활용
애플 (Apple)	미국 노스캐롤라이나	• 향후 5년간 미국 전역에 데이터센터 설립 4,300억 달러 투자 계획 • 노스캐롤라이나 카토바카운티 데이터센터 4억 4,800만 달러 투자 계획
	중국 남부 구이저우	• 10억 달러 규모 iCloud 서비스용 데이터센터 설립 및 운영
오포 (OPPO)	중국 둥관시	• 총 건축면적 102,200㎡, 12만여 개 서버 보유 • 지능형 제조 데이터센터
알리바바 (Alibaba)	중국 장베이현	• 3년간 클라우드 인프라 구축을 위해 2,000억 위안 규모 투자 진행
테슬라 (Tesla)	중국 상하이	• 중국 내 차량에서 수집된 데이터 처리·활용을 위한 데이터센터 설립
마이크로소프트 (Microsoft)	말레이시아	• 향후 5년 동안 10억 달러 투자, 복수의 데이터센터로 구성된 '데이터센터 지역' 구축 목표
	인도네시아	• 데이터센터 구축으로 경제적 효과 63억 달러, 일자리 6만 개 창출 전망
	미국 시카고	• 2억 달러 투자 및 20만m2 규모 데이터센터 2개 설립 • 자사 애저(Azure) 서비스 영역 확대에 활용
텐센트 (Tencent)	인도네시아	• 데이터센터 구축(2021년 말까지 2곳 목표)
화웨이 (Huawei)	태국	• 데이터센터 구축에 2,300만 달러 투자

* 출처: 해외 데이터 뉴스매체 발행 이슈 23,850건 분석(2021.1 ~ 2021.8)

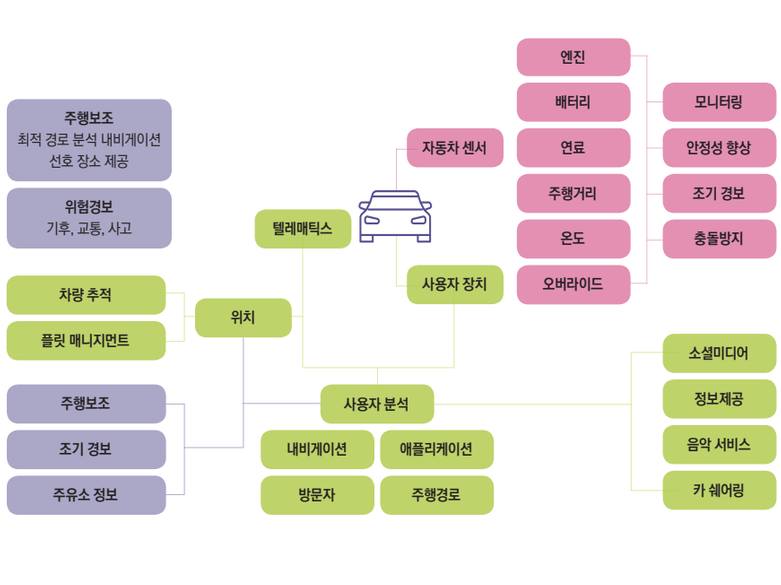
이슈5. 자율주행, 데이터 발전의 최대 수혜 기술로 부상

자동차 산업에서 데이터 시장 5년간 연평균 16.8% 성장

자동차 산업의 데이터 시장은 2025년까지 연평균 성장률 16.8%를 기록하며 89억 2,937만 달러 규모로 성장할 것으로 전망된다.[2] 데이터 기술은 커넥티드 카 개발, 지능형 제조 시스템 구축, 자동차 금융 서비스, 공급망 개선 및 마케팅 솔루션 등 광범위하게 도입되고 있다.

자동차 기업들은 자율주행 모빌리티 개발을 위해 데이터 기술 개발 및 활용을 적극 추진하고 있다. 그러나 일각에서는 자동차 주행과정에서 생성된 데이터가 개인 프라이버시 침해 문제를 일으킬 것이란 우려가 제기되고 있다. 특히 중국 사이버보안관리국(Cyberspace Administration of China, CAC)은 자동차 산업에 대한 데이터 보안 및 관리 규칙

[그림 7] 자동차 산업에서의 데이터 수집 및 활용 메커니즘



* 출처: UnfoldLabs(unfoldlabs.medium.com, 접속일: 2021.11.3.)

2 Research And Markets, 「Big Data Market in the Automotive Industry - Growth, Trends, COVID-19 Impact, and Forecasts (2021 - 2026)」, 2021.1.

을 발표하여 자동차 산업 내 데이터 활용에 관한 지침을 마련했다.

[표 11] 주요 자동차 기업의 데이터 활용 및 개발 현황

국가	주요 내용
마쓰다 (Mazda)	• 데이터 기술 및 클라우드 서비스를 도입한 커넥티드 모빌리티 개발 • 스즈키(Suzuki), 스바루(Subaru), 다이하쓰(Daihatsu) 등 자동차 제조기업과 차량 통신 기술 공동 개발
메르세데스 벤츠 (Mercedes· Benz)	• 광학 스캐닝 장치를 이용한 데이터 수집을 통해 생산 프로세스 최적화 진행 • 자동화 소프트웨어 도입으로 엔지니어링 과정에서 생산되는 데이터 수집 및 분석 수행
BMW	• 클라우드 데이터 허브를 통해 매일 수백만 대 차량에서 수집되는 테라바이트급 원격 측정 데이터 처리 수행 • 수집·처리된 데이터를 분석하여 문제 식별·신규 상품 개발 등 수요 예측 수행
포르쉐 (Porsche)	• 특수 소프트웨어 AMDA(Automated Measurement Data Analyics)를 통한 차선 변경 조기 감지 기술 개발 • 데이터 시뮬레이션을 통한 인공지능 및 자율주행 기능 향상
테슬라 (Tesla)	• 내장 카메라, 초음파 센서, 레이더를 활용한 도로환경 영상 데이터 기반의 딥러닝 기술 도입 • 고객 주행 패턴 및 주행 중 돌발상황 등 실제 도로주행 데이터 수집·분석을 통한 완전 자율주행 고도화
도요타 (Toyota)	• 데이터 수집·분석 플랫폼 도입 및 사물인터넷 플랫폼과의 결합 추진

* 출처: 해외 데이터 뉴스매체 발행 이슈 23,850건 분석(2021.1. ~ 2021.8.)

2021 데이터산업 백서

2021 Data Industry White Paper

Contents

발간사 02

그림으로 보는 데이터산업 동향 04

2021 데이터산업 이슈 TOP 5 08

국내 및 해외 데이터산업 이슈 TOP 5

34p **1부. 데이터 민주주의의 시대**

1장. 데이터 권리의 민주화(국문 번역본) 36

Data Sovereignty 46

2장. 데이터 활용의 민주화

: 시티즌 사이언티스트와 스마트시티 58

3장. 데이터 경쟁의 민주화 72

82p **2부. 데이터산업 관련 주요 정책·법제도**

1장. 국내 데이터산업 관련 정책·법제도 84

2장. 해외 데이터산업 관련 정책 96

해외 데이터산업 관련 법제도 102

3장. 디지털 무역과 국경 간 데이터 이동 108

114p **3부. 데이터산업 시장 현황**

1장. 국내 데이터산업 시장 현황 116

2장. 해외 데이터산업 시장 현황 130

142p **4부. 산업별 데이터 활용 현황**

1장. 금융분야 데이터 활용 144

2장. 헬스케어분야 데이터 활용 155

3장. 제조분야 데이터 활용 166

4장. 유통·물류분야 데이터 활용 175

5장. 광고분야 데이터 활용 186

6장. 모빌리티분야 데이터 활용 193

7장. 에듀테크분야 데이터 활용 205

8장. 농업분야 데이터 활용·비즈니스 213

9장. 데이터 동맹·제휴 현황 220

230p **5부. 데이터산업 기술 동향**

1장. AutoML 동향 그리고 대중화 232

2장. Cloud Computing 248

3장. 개인정보 비식별 처리 수준 및 프레임워크 259

4장. 안정적 AI/ML 서비스를 위한 MLOps 271

2021년 데이터산업 백서 집필진 282

[표 1-1-1] 마이데이터 원칙과 데이터 주권 38
[Table 1-1-1] MyData principles and data sovereign 49
[표 1-2-1] 지속적 지능화의 장애 요인 59
[표 1-2-2] 스마트시티데이터의 범위 68
[표 1-2-3] 시티데이터 거버넌스 범위 70
[표 1-3-1] 데이터 관련 경쟁 제한 우려 사항 79
[표 2-2-1] 주요국 데이터 정책 현황 101
[표 3-1-1] 국내 데이터 처리 및 관리 솔루션 개발·공급업 중분류별 시장 규모 (2016-2020년(E)) 119
[표 3-1-2] 국내 데이터 구축 및 컨설팅 서비스업 중분류별 시장 규모 (2016~2020년(E)) 120
[표 3-1-3] 국내 데이터 판매 및 제공 서비스업 중분류별 시장 규모 (2016~2020년(E)) 122
[표 3-1-4] 국내 데이터산업 부문별 데이터직무 인력 현황(2016~2020년) 123
[표 3-1-5] 국내 전 산업 데이터직무별 인력 현황(2020년) 125
[표 3-1-6] 국내 주요 산업별 시장 규모 추이(2015~2019년) 128
[표 3-2-1] 유럽연합의 데이터 전문 인력 수(2017~2021년(E)) 134
[표 4-1-1] 데이터 3법 주요 개정내용 146
[표 4-1-2] 개인정보, 가명정보, 익명정보 비교 147
[표 4-1-3] 금융사 대안신용평가모델 개발 현황 154
[표 4-2-1] 2021년 '보건의료데이터 중심병원 지원 사업' 참여현황 159
[표 4-2-2] 마이 헬스웨이 4개 분야 12개 과제 161
[표 4-2-3] 2021년 디지털 치료기기 R&D 프로젝트 공모 163
[표 4-2-4] 2021 AI융합신규감염병 대응 시스템 개발 164
[표 4-3-1] 수집 데이터의 종류 및 특징 170
[표 4-3-2] 주조 설비 주요 관리 항목 데이터 170
[표 4-3-3] 식품가공 업종 내 육가공 공장에서 수집한 데이터 예시 172
[표 4-5-1] 국내 온라인 광고시장 변화 추이 전망 189
[표 4-6-1] 국내 교통관련 공공데이터 플랫폼 203
[표 4-7-1] AIEd(AI in Education)의 세 가지 모델 209
[표 4-9-1] 결합전문기관과 데이터전문기관 지정 요건 223
[표 4-9-2] 2021년 공공과 민간의 데이터동맹 주요 사례 225
[표 4-9-3] 2021년 금융회사 중심 데이터동맹 주요 사례 226
[표 4-9-4] 2021년 기타 데이터동맹 227
[표 5-1-1] DataRobot의 표준 머신러닝 프로세스 참고 235
[표 5-1-2] AutoML 제공 유형 237
[표 5-1-3] Market Presence Metrics 244
[표 5-1-4] Gartner Peer Insights(2021)의 Customer Reviews 245
[표 5-1-5] AutoML 평가요소별 기능요소 246
[표 5-2-1] 컴퓨팅 환경의 변화 249
[표 5-2-2] 외국계 클라우드 국내 진출 현황 251
[표 5-2-3] 클라우드 서비스 세부 유형 분류 253
[표 5-2-4] Managed Service(MS) 내용 254
[표 5-2-5] 클라우드 서비스 유형별 성장률 전망 255
[표 5-2-6] 국내 주요 클라우드 257
[표 5-3-1] 재식별 공격 위협 260
[표 5-3-2] 비식별 처리의 유형 261
[표 5-3-3] ADF 구성요소 요약표 266
[표 5-3-4] 주요 비식별 프레임워크의 장·단점 267
[표 5-4-1] PSI 계산 예시 275
[표 5-4-2] 모델 학습방안 조건 278

그림 목차

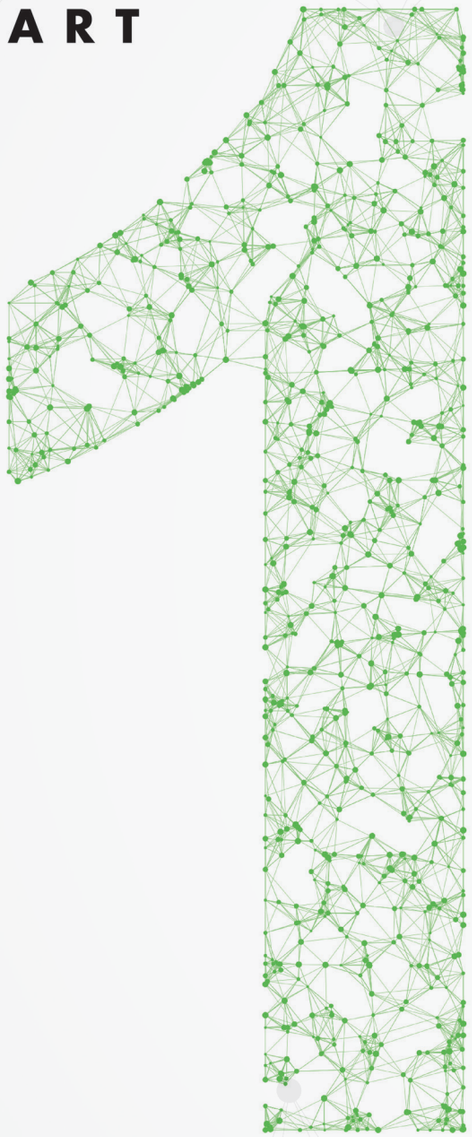
[그림 1-2-1] 데이터 기반 지속적 지능화 체계 59
[그림 1-2-2] '데이터 활용 민주화'를 통한 지속적 지능화 장애 요인 극복 60
[그림 1-2-3] 데이터팀의 역할 분산 61
[그림 1-2-4] 데이터 카탈로그의 확장된 메타데이터 제공 범위 63
[그림 1-2-5] Airbnb의 Dataportal, Lyft의 Amundsen 64
[그림 1-2-6] 데이터 분석 과정 자동화 영역 65
[그림 1-2-7] 비즈니스 현업의 데이터 리터러시를 위한 교육 영역 66
[그림 1-2-8] 스마트시티 생태계 구성 예시 69
[그림 2-1-1] 디지털 뉴딜 2.0에서 신설된 초연결 신산업 육성 과제 87
[그림 2-1-2] 디지털 뉴딜 1.0과 2.0의 과제 변화 88
[그림 2-1-3] 국가 데이터 정책 추진 체계도 90
[그림 3-1-1] 국내 데이터산업 시장 규모(2015~2020년(E)) 117
[그림 3-1-2] 국내 데이터산업 부문별 시장 규모 비중(2015~2020년(E)) 117
[그림 3-1-3] 국내 데이터산업 시장 전망(2020(E)~2026년(P)) 118
[그림 3-1-4] 국내 데이터 처리 및 관리 솔루션 개발·공급업 시장 규모 (2015~2020년(E)) 119
[그림 3-1-5] 국내 데이터 구축 및 컨설팅 서비스업 시장 규모 (2015~2020년(E)) 120
[그림 3-1-6] 국내 데이터 판매 및 제공 서비스업 시장 규모 (2015~2020년(E)) 121
[그림 3-1-7] 국내 데이터 판매 및 제공 서비스업 중분류별 시장 규모 비중 (2015~2020년(E)) 121
[그림 3-1-8] 국내 데이터산업 전체 인력 현황(2015~2020년) 122
[그림 3-1-9] 국내 데이터산업 인력 구성 및 데이터직무별 인력 비중(2020년) 123
[그림 3-1-10] 국내 전 산업 데이터직무 인력 현황(2015~2020년) 124
[그림 3-1-11] 국내 전 산업 데이터직무별 인력 비중(2020년) 124
[그림 3-1-12] 국내 데이터산업의 향후 5년 내 기술등급별 데이터직무 필요 인력 비중 126
[그림 3-1-13] 국내 데이터산업의 향후 5년 내 데이터직무 인력 부족률 126
[그림 3-1-14] 국내 데이터산업/일반산업/전 산업의 향후 5년 내 데이터직무 필요 인력 127
[그림 3-1-15] 국내 전 산업의 향후 5년 내 데이터직무 인력 부족률 127
[그림 3-1-16] 국내 전 산업의 향후 5년 내 기술등급별 데이터직무 필요 인력 비중 128
[그림 3-2-1] 글로벌 디지털 데이터 시장 규모(2017~2021년(E)) 131
[그림 3-2-2] 글로벌 디지털 데이터 기업 수(2017~2021년(E)) 132
[그림 3-2-3] 글로벌 데이터 전문 인력 수(2017~2021년(E)) 134
[그림 3-2-4] 데이터산업의 경제적 효과: 직접효과(2018~2021년(E)) 135
[그림 3-2-5] 데이터산업의 경제적 효과: 간접 후방 효과(2018~2021년(E)) 137
[그림 3-2-6] GDP 대비 데이터산업의 경제적 효과(직접 효과+간접 후방 효과)의 비율(2019~2021년(E)) 138
[그림 3-2-7] 유럽연합의 중장기 데이터 시장 전망(2020~2025년(E)) 139
[그림 3-2-8] 글로벌 빅데이터 시장 전망(2021(E)~2025년(E)) 140
[그림 4-1-1] 금융 빅데이터 5대 인프라 체계도 148
[그림 4-1-2] 금융표준종합정보DB 체계도 149
[그림 4-1-3] 금융 빅데이터 개방시스템(CreDB) 체계도 150
[그림 4-1-4] 금융데이터거래소 체계도 151
[그림 4-1-5] 마이데이터 지원센터 주요 역할 152
[그림 4-1-6] 금융분야 인공지능 가이드라인 주요 내용 153
[그림 4-2-1] 보건의료데이터·인공지능 혁신전략 추진 과제 156
[그림 4-2-2] 의료데이터 중심병원 빅데이터 시스템(예시) 158
[그림 4-2-3] 마이헬스웨이 개념도 160
[그림 4-2-4] 디지털치료기기 판단 기준 162
[그림 4-3-1] 스마트공장 제조데이터 활용 형태 167
[그림 4-3-2] 디지털 기술로 새로운 가치 창출 167
[그림 4-3-3] 지능화된 미래형 공장 개념 168
[그림 4-3-4] 스마트 공장 개념 168
[그림 4-3-5] 공정흐름도 및 데이터 수집 포인트 169
[그림 4-3-6] 대시보드 화면 구성 예시 171
[그림 4-3-7] 제어구간별 전류소모량 패턴의 가시화 및 분석 화면 172

그림 목차

[그림 4-3-8] 제조데이터 수집을 위한 국제 표준 가이드 174
[그림 4-4-1] CJ대한통운, 네이버쇼핑에 풀필먼트 서비스 개시 176
[그림 4-4-2] 통합GS리테일 사업 비전 178
[그림 4-4-3] AI 알고리즘 XGBoost와 GRU를 활용한 수요 예측 180
[그림 4-4-4] 대표적인 물류 로봇 182
[그림 4-4-5] 건물별 배송난이도 예시 184
[그림 4-4-6] 실시간 교통정보 서비스를 활용한 라스트마일 배송 경로 자원 산정 184
[그림 4-5-1] 전 세계 디지털 광고시장 지출 규모 전망치 187
[그림 4-5-2] 전 세계 유형별 디지털 광고의 분기별 투자 추이 188
[그림 4-5-3] 전 세계 광고 영역별 투자 전망: 2020 vs. 2021 188
[그림 4-5-4] 국내 광고시장 매출액 비중(2020년(E)) 189
[그림 4-5-5] 2020년 12월 기준 온라인 동영상 광고비(좌측) 및 추이(우측) 190
[그림 4-5-6] 네이버 게시글 수 추이(2003~2020년) 191
[그림 4-5-7] 메타버스 수익모델의 진화 단계 192
[그림 4-6-1] 슈퍼바이즈드 머신러닝 vs. 인간 운전자 수준 비교 196
[그림 4-6-2] 자율주행 데이터 소스 및 인공지능 적용 절차 197
[그림 4-6-3] 차량 데이터의 잠재적 수익과 시장진입 속도에 따른 서비스 유형 분류 200
[그림 4-6-4] 카페이먼트 에코시스템 202
[그림 4-7-1] 글로벌 교육 분야 에듀테크 지출 규모(2020~2025년) 206
[그림 4-7-2] 글로벌 교육분야에서의 첨단기술의 성장 207
[그림 4-7-3] 1965년에 출시된 교육용 애플리케이션 플라토(PLATO) 209
[그림 4-7-4] 상호보완적으로 운영 가능한 데이터 모델 211
[그림 4-7-5] Open Edx와의 연계를 위한 데이터 상호작용 관계 212
[그림 4-8-1] 목화 종자에 인디고 애그리컬처의 미생물 코팅 기술 적용 (왼쪽: 인디고종자) 214
[그림 4-8-2] 아쿠아바이트 수중 카메라(데이터 인식) 및 빅데이터 활용 214
[그림 4-8-3] 스마텍의 최적 사육 관리 시스템 215
[그림 4-8-4] 세계 농업 인공지능 대회 온실 공간과 IoT 센서 설치 중인 디지로그팀 216

[그림 5-1-1] 전 세계 2025년까지의 AI 소프트웨어 시장 성장 전망 233
[그림 5-1-2] 전 세계 자동 기계학습(Automated Machine Learning) 시장 234
[그림 5-1-3] DataRobot 모델 결과를 제공하는 리더보드 화면 236
[그림 5-1-4] Google Trends를 통한 AutoML 관련 검색어 추이 분석 (2015.7.1. ~ 2021.7.16.) 238
[그림 5-1-5] 오픈소스 - Auto sklearn, Auto WEKA, TPOT 등 관련 검색어 추이 분석 (2015.7.1. ~ 2021.7.16.) 239
[그림 5-1-6] 스타트업 - DataRobot, Dataiku, H2O.ai 등 관련 검색어 추이 분석 (2015.7.1. ~ 2021.7.16.) 239
[그림 5-1-7] 거대테크 - MS Azure, Google, AWS 등 관련 검색어 추이 분석 (2015.7.1. ~ 2021.7.16.) 240
[그림 5-1-8] AutoML의 영향력 241
[그림 5-1-9] Targetbase(2018)의 Automation 평가 243
[그림 5-1-10] AutoML 솔루션 새로운 흐름 243
[그림 5-2-1] 클라우드 사업구조 250
[그림 5-2-2] 클라우드 생태계 Value Chain 250
[그림 5-2-3] 외국계 클라우드 서비스 제공자별 매출 및 성장률, 매출 대비 비중 (2020) 252
[그림 5-2-4] 클라우드 시장 규모 전망 255
[그림 5-2-5] 국내 클라우드 시장 규모 및 전망 256
[그림 5-2-6] 국내 클라우드 시장 규모 및 서비스별 비중 256
[그림 5-2-7] 국내 업종별 클라우드 시스템 활용 현황 257
[그림 5-2-8] 국내 업무별 클라우드 사용 현황 257
[그림 5-3-1] Identifiability Continuum의 개념적 5단계 모델 263
[그림 5-3-2] ADF Workflow 265
[그림 5-3-3] 국내 데이터 상황 분류도 268
[그림 5-4-1] ML코드를 포함한 실제 ML 시스템 및 관련 인프라 272
[그림 5-4-2] 드리프트 모니터링 예시 275
[그림 5-4-3] 챌린저 모델의 정확도 모니터링 예시 277

PART



데이터 민주주의의 시대

제1장.
데이터 권리의 민주화(국문 번역본)
Data Sovereignty

제2장.
데이터 활용의 민주화
: 시티즌 사이언티스트와 스마트시티

제3장.
데이터 경쟁의 민주화

제1장.
데이터 권리의 민주화[1]

마이데이터 글로벌 리더십 팀 공저
대표저자: 비비 라티노야

MyData Global로서 우리는 개인 및 다수에 의해 생성된 개인데이터에 가장 관심을 갖고 있다. 이 장에서는 데이터 주권을 자신에게 특정한 디지털 영역에 관한 개인의 최고 권한으로 정의한다. 우리는 MyData Global의 맥락에서 이해되는 데이터 주권의 두 가지 측면을 강조한다. 1부에서는 마이데이터 원칙을 데이터 주권에 대한 정의와 연결짓는다. 2부에서는 개인데이터 주권의 다양한 측면을 실현하고 구체화하기 위해 진행 중인 이니셔티브에 대해 설명한다.

1. 전통적 주권 vs. 데이터 주권

데이터 주권은 완전히 새로운 개념은 아니고, EU집행위원회의 데이터 전략(2020.2)에서 '기술 주권(Technological Sovereignty)'이라는 용어를 사용하면서 특히 주목을 받았다. 이러한 유형의 주권의 의미와 동기에 대한 해석은 데이터 현지화 및 지역 보호주의로부터 경제적 및 지정학적 야심(ambition), 그리고 프라이버시와 디지털 자기 결정이라는 기본 디지털 인권 방어에 이르기까지 다양하다. 이러한 모든 해석의 공통점은 정확성은 차치하고 기술, 정보, 디지털 또는 데이터 주권으로 다양하게 불리는 이 새로운 종류의 주권을 지리적 영토라는 전통적 주권 개념의 측면에서 인지하고 있다는 점이다.

주권은 일반적으로 '영토에 대한 최고 권한'으로 이해되고, 전통적으로 영토 또는 일반적으로 국가와 관련하여 권력을 행사하는 가장 최고의 정당성에 관련된 정치 이론의 영역이었다(Philpott, 2020). EU가 주도하는 '유럽 기술 주권'에 관하여 EU에서 야기된 논의에서는 주권을 전통적인 국가에 기반한 것으로 간주하나 아날로그 세계에서 그려지는 국경에 한정된 주권이 아니라는 점을 강조한다. 디지털 세계에서도 국가의 경계가 그려진다는 것을 의미한다. 따라서, 정보와 데이터뿐만 아니라 기술과 인프라는 사용자가 속한 국가의 법률, 요구사항 및 이익에 기초하여 관리되어야 한다. 이러한 생각은 데이터 주권을 민족 국가(또는 EU의 경우, 민족 국가 집합)와 지리적 영역에서의 최고 권위라는 관점인 전통적 주권의 기본 개념을 확장한다.[2]

이 장에서는 주권의 다른 개념을 논하고자 한다. 최고 우위, 권위, 영토권이라는 전통적 개념을 그대로 가져오고, 영토권을 완전히 다른 것으로 대체한다. 본 장에서 정의하는 데이터 주권이란 '디지털 영역(domain)이라고 정의된 곳에서의 최고의 권한'을 의미하고, 국가나 지역을 더 이상 언급하지 않는다. 이 데이터 주권 개념은 행위자(개인, 그룹, 조직 및 기관)에 속한 것으로 그들의 물리적 위치나 현재 관할권과 관계없이 그들 자신에 의해, 그리고 그들 자신에 대해 생성된 데이터 주권을 나타낸다.

MyData Global[3]은 개인 혹은 다수에 의해 그리고 그들에 대해 생성된 데이터에 가장 관심이 많다. 이는 이 장에서 논의하는 데이터 주권이 조직과 기관 데이터에 효과적으로 적용될 수 없다는 의미는 아니다. 마찬가지로 국가 또는 영토의 데이터 주권에 대해 논의하지는 않지만 여기서 언급한 일부 사항이 해당 논의에 도움을 제공할 수 있다고 확신한다.

본 장의 목적은 우리가 이해하는 대로 데이터 주권의 두 가지 측면을 강조하는 것이다. 바로 데이터 주권의 '행사'와 '소유'다. 즉, 데이터 주권을 '실현'하는 방법과 이를 '올바르게' 하는 방법에 대한 비전을 간략하게 설명한다. 후자로 시작해서, 전자에 대해 논하고, 결론에서는 데이터 주권을 향한 조치를 취할 수 있는 역량을 가진 데이터 종사자들을 위한 실질적인 제언을 제공한다.

2. 데이터 주권의 정의, 마이데이터 원칙과의 연관성 및 제약성

가. 데이터 주권의 정의

> 데이터 주권은 특정한 디지털 영역(domain)과 관련된 개인의 최고 권위라고 정의된다.

1 위 원고는 해외 집필진의 영문 저작물을 국내 독자의 이해를 위하여 국문으로 번역한 것입니다. 정확한 이해를 위하여 46쪽 영문 원본을 참고 부탁드립니다.

2 여기서 주의해야 할 점은 탈식민 국가들의 토착 인구와 관련된, 그리고 보다 광범위하게는 국경에 의해 정의되지 않는 뚜렷한 정체성을 가진 사람들의 집단과 관련된 데이터를 구체적으로 다루는 데 전통적이고 지리적으로 묘사되는 데이터 주권 개념이 특별히 사용되기 시작했다는 점이다(참조 예: Rainie et al., 2017). 토착 인구의 데이터 주권에 대한 이러한 개념은 이하에서 설명된 데이터 주권에 대한 정의를 구축할 때 발견되는 유사한 전통적 주권 개념의 연장 내용과 관련이 있다.

3 MyData Global은 유럽이 중심으로 2018년 10월에 설립한 국제 협력 네트워크 조직으로, 6개 대륙에서 25개 로컬 허브가 활동하고 있다. 아시아에서는 일본에 이어 두 번째로 우리나라에 로컬 허브가 설립됐다. MyData Global 조직이 설립돼 마이데이터 사상과 표준을 정립해 나가는 것은 의미가 매우 크다.

권위란 의미는 단순히 강제력이 아니라 권리, 즉, 행사하고, 그에 따라 지켜야 할 권리로 이해되어야 한다(Wolff, 1970). 따라서 권위는 합법적인 것으로 이해되어야 한다.

한편, 최고 우위란 의미는 디지털 주권 영역과 관련하여 개인의 주권이 모든 것(국가 포함)에 앞선다는 것을 나타낸다.

개인에게 특정한 디지털 영역이란 의미는 해당 개인에 의해 또는 개인에 대해 생성된 데이터로 정의된다. 이러한 정의의 측면은 영역 기반 데이터 주권에 대한 일반적인 논의와 구별되는 것인데, 이 장에서는 그 논의에도 새로운 관점을 제공하고자 한다. 지리적 영역이 아닌 개인을 데이터 주권의 통합 지점이자 범위를 정하는 요소로 사용하면, 사람들과 그들에 관한 데이터가 지리적으로 엄격하게 제한되지 않는 현재 상황에 대해, 보다 유연하게 생각할 수 있을 것이다.

나. 마이데이터 원칙 및 데이터 주권과의 연관성

MyData Global은 2017년에 6가지 마이데이터 원칙을 선언하고, 이를 유지해왔다. 아래 표는 그 원칙이 앞에서 정의된 데이터 주권의 개념과 어떻게 관련되는지 설명한다.

[표 1-1-1] 마이데이터 원칙과 데이터 주권

마이데이터 원칙	주요 내용	데이터 주권과의 연관성
인간 중심적 통제	개인은 온·오프라인 모두에서 개인 생활을 관리할 수 있는 권한과 수단을 부여받아야 함. 누가 자신의 데이터에 접근하며 그 데이터가 어떻게 사용되고 공유되는지를 이해하고 효과적으로 통제할 수 있는 실질적인 도구와 방법 제공 필요	데이터 주권은 개인 및 그룹의 자기 결정권, 자율성 및 자신에 대한 개인데이터에 대한 통제권을 의미
구심점으로서의 개인	개인데이터가 다양해지면서 발생되는 가치가 큰 폭으로 상승하지만 사생활 침해에 대한 위협 또한 비례하여 증대. 이러한 상황은 개인이 개인데이터의 상호 참조가 일어나는 허브가 된다면 해결 가능	데이터 주권은 개인(또는 사람들의 집단)을 그들이 속한 국가 또는 지역 대신에 주권이 유지되고 행사되는 영역을 정의하는 요소로 여기는 것을 의미
개인의 주도권 강화	데이터 기반 사회에서는, 여느 사회와 마찬가지로, 개인을 단순히 서비스나 응용프로그램의 고객 혹은 사용자로 여기면 안 됨. 개인들은 자기 자신의 목표를 설정하고 추구할 수 있는 자유롭고 자율적인 행위자로 인식할 필요 있음. 또한 개인은 행동권과 주도권을 가져야 함	데이터 주권은 개인과 집단이 (반드시) 가져야 하는 것뿐만 아니라 그들이 데이터 생태계의 주체이자 적극적인 참여자로서 행사할 수 있어야 함
데이터 이동권: 접근 및 재사용	개인데이터의 이동권은 개인이 그들 자신의 목적이나 다른 서비스에 있어 자신의 개인데이터를 재활용할 수 있게 하며, 폐쇄된 사일로(silo)에서의 데이터를 재사용이 가능한 자원의 데이터로의 변화를 이끄는 열쇠임. 그런데, 데이터 이동권은 단순한 법적 권리로서만 존재하면 안 되며, 실질적인 수단과 결합해야 함	데이터 주권 행사의 중요한 측면은 자신에 대한 데이터에 접근하고, 연결하며, 재사용할 수 있는 권리이자 수단임
투명성과 책임	개인데이터를 사용하는 조직은 그 데이터로 무엇을 하고 왜 하는지 공지해야 하며, 공지한 것을 준수해야 함. 그들은 개인데이터를 보유하고 사용함으로써 발생한 의도된 결과뿐 아니라 보안사고 등 의도하지 않은 결과에도 책임을 져야 하며 개인은 그들에게 책임을 물을 수 있음	데이터 주권은 개인데이터 사용의 합법성이 항상 해당 데이터가 존재하는 개인 또는 집단에서 유래한다는 것을 의미. 조직이 개인데이터를 사용할 권리를 갖는 자격에 대한 투명성과 해당 조직에 책임을 물을 수 있는 능력은 데이터 주권 존중을 보장하는 데 필요
상호운용성	상호운용의 목적은 특정한 기관이나 기업에 종속되는 데이터 락인효과(lock-in)의 가능성을 제거함과 동시에 데이터 보유기관에서 데이터 활용기관으로 데이터가 이동할 때 생기는 마찰을 줄이는 것임. 상호운용성은 업무 관행의 일반화 및 기술 표준화를 통해 달성될 수 있으므로 이에 대한 지속적인 노력이 필요	데이터 주권을 행사하는 데에 외부의 간섭을 받지 않는 것을 의미. 상호운용성은 기술 또는 다른 장벽 없이 사용할 도구와 협력 주체를 선택할 수 있는 자유를 지원하는 원칙

다. 데이터 주권의 2가지 제약 요소

전통적으로 주권은 절대적인 것으로 인식되고, 절대적이라는 것은 영토 내의 모든 사항들에 제약 없이 적용되는 것을 의미한다[4]. 그러나 데이터 주권의 경우에는 그럴 수 없다. 상호연결된 디지털 영역에서 어떠한 사람의 권위가 절대적일 수 있다는 것은 상상할 수 없다. 당신과 다른 사람을 촬영한 사진을 공개하거나 혹은 공개하지 못하게 할 권리는 바로 한 명 이상의 개인과 관련된 개인데이터의 전형적인 예이며, 그 사항은 두 사람 모두의 권위에 종속된다. 사실, 우리에 대해 생성된 대부분의 개인데이터는 다른 사람이나 회사 및 정부와 같은 데이터 권리 보유자와의 상호작용을 통해 존재한다고 말할 수 있다.

둘째, 여기에서의 데이터 주권의 의미는 정의된 디지털 영역 내에서 즉, 개인에 의해 또는 개인에 대해 생성된 데이터와 개인과 관련된 보유되고, 행사되는 무언가를 말하는 동시에, 외부의 인정과 존중에 의존하며 또한 그러한 인정 및 존중을 요구하고 있는 무언가라고 말해야 할 것이다. 여기에서 우리가 생각하는 데이터 주권은 외부 영향으로부터 자유로운 기본 특권 또는 배타적 권리를 의미한다. 그러나 그 기본 특권이 정확히 무엇인가라는 질문은 보편적으로 대답하기 어려운 질문이다.

이러한 두 가지 제약 요소(데이터 주권의 비절대성과 데이터 주권자의 기본 특권의 구성요소에 대한 정의의 어려움)를 고려할 때, 데이터 주권은 우리가 살고 있는 주권 국가들로 구성된 세계의 주권 시스템처럼 상당히 많은 대화와 협상의 여지가 있다.

개인데이터는 공유되고, 다른 사람의 데이터와 결합될 때 그 가치와 유용성이 더욱 증대된

4 전통적인 영토 주권조차도 점차 절대적이지 않게 되고 있다는 점에 주목할 가치가 있다. UN, EU, 국제형사재판소 등의 구성원인 국가들은 초국가적 기관에 일부 권한을 양도한다.

다. 따라서, 개인데이터의 잠재적인 이익을 활용하려면 주권을 갖는 단체, 사람들, 집단이 협력하고, 그 자원을 같이 모을 수 있어야 한다. 이러한 협력은 연구, 정책 결정, 정부 당국의 의사 결정 등이 이루어지는 데 도움을 제공할 수 있다. 한편, 이러한 협력이 이루어지기 위해서는 공유 거버넌스 확보와 그 거버넌스의 이행 약속이 필요하다. 이때 좋은 데이터 거버넌스는 자신에 대한 데이터 주권을 기본 원칙으로 여긴다. 주권 국가가 EU나 UN과 같은 초국가적 프레임워크에 기초하여 협력하는 것과 마찬가지로, 주권을 갖는 개인과 그룹은 데이터 협업 또는 데이터 중개자 네트워크와 같은 공통 거버넌스 구조를 준수하고 충실히 지켜나갈 수 있다.

라. 데이터 주권의 적용 사례

개인데이터 주권은 MyData Global이 마이데이터 선언에서 설명한 것처럼, 미래 세계의 핵심적 구성 요소이다. 현행 데이터 사용 패러다임이나 그로 인한 이익이 사람과 조직들 간의 비대칭적인 힘의 관계에 의해 분배되는 방식은 지속가능하지도 공정하지도 않다. 사람과 집단이 원칙적으로 데이터 주권을 갖고 실제로 해당 데이터 주권을 행사할 수 있도록 보증하는 작업이 필요하다. 아래에서는 데이터 주권의 실제 적용에 대한 몇 가지 사례를 제시한다.

사례 1: 데이터 주권으로 가능해진 데이터 기부

데이터 주권은 사회(구성원들)에 이익을 가져다주는 개인데이터의 공유와 기여를 위해 중요하다. 현재 우리가 겪고 있는 팬데믹과 같은 경우 특히 공동체와 사회에 대한 개인의 의무감이 높아진다. 사람과 집단이 이러한 의무를 수행하는 한 가지 방식은 연구원과 의사 결정자를 돕기 위해 개인(건강)데이터를 기부하는 것이다. 이러한 형태의 기부는 사람들이 데이터 주권을 행사하고, 데이터를 원하는 목적에 사용할 수 있을 때 가능하다.

사례 2: 직원 데이터 주권

데이터 주권은 직원과 고객에 대한 데이터 사용의 책임이 있는 회사에도 중요하다. 직원 데이터를 윤리적으로 사용하면 인재를 유치하는 데 도움이 되고, 그 결과 고객 만족도를 높이고 혁신을 이룰 수 있다. 직원 데이터 주권이란 직원이 자신과 자신의 성과에 대해 모은 데이터에 접근할 수 있는 것을 의미한다. 나아가 직원 데이터 주권을 통해 직원이 사내에서 맞춤형 경력 서비스를 제공받거나 제3자 서비스를 사용함으로써 자신과 결과적으로 해당 고용주에게 혜택을 줄 수 있다.

사례 3: 데이터 주권에 대한 확신을 향한 센스메이킹[5]

개인이 데이터 주권을 행사하려면 데이터 생태계, 경제가 작동하는 방식, 그리고 그 안에서 탐색할 수 있는 능력에 대한 어느 정도의 이해가 필요하다. 이러한 종류의 이해와 능력은 집단적 센스메이킹에서 발생할 수 있고, 자신과 자신이 속한 공동체에 도움이 되는 방식으로 데이터 주권을 행사할 수 있다는 자신감 형성에 기여할 것이다. 데이터 주권은 데이터 주권을 가진 사람과 집단이 그것을 자신감 있게 행사하기 위해 주변의 생태계와 경제를 공동으로 그리고 지속적으로 이해할 것을 요구한다.

5 센스메이킹이란 환경의 여러 불확실한 요인들을 파악 및 이해하고 그러한 이해를 바탕으로 행동을 취하는 것을 의미한다.

3. 데이터 주권 관련 사례(사업)

두 번째 파트에서는 개인데이터 주권의 다양한 측면을 실현하고 구체화하려는 일부 진행 중인 주요 활동들에 대해 활동 주체들의 목소리를 통해 설명한다.

가. Findy

데이터와 프라이버시를 보호하고, 시민들이 데이터를 사용할 수 있도록 하며, 사기와 회색 경제와 싸우는 현재의 방법은 사용자와 서비스 제공자 모두에게 비용이 많이 들고, 어렵기도 하다. 익명화된 개인데이터의 자동화 및 실시간 사용을 통해 서비스 수준과 생산성을 개선하기 위해 데이터 주권을 행사할 때도 마찬가지이다. 주된 이유는 인터넷이 ID 계층 없이 구축되었기 때문이다.

자기주권 ID(Self-Sovereign Identity, SSI) 운동은 개방형 표준으로 이 문제를 해결하고, 개별 사용자에게 통제권을 이전하는 것을 목표로 한다. ToIP 재단(Trust over IP Foundation)이 표준화 작업의 대부분을 관리 · 운영하지만, 다른 조직들도 이에 관련되어 있다. 많은 프로그램들이 이미 이 접근 방식을 택하고 있다. 이와 관련된 활동을 하는 Findy(https://findy.fi/en/)도 그러한 예다.

Findy 협동조합은 e-서비스에 필요한 정보의 진본성(authenticity)을 보장하기 위한 목적으로 범용, 공유, 보안의 데이터 네트워크를 개발하는 민관 협력 비영리 단체이다. 혁신적이고, 실용적이며 사회적으로 영향력 있는 데이터 구동 서비스는 민간과 공공 부문 조직 간의 긴밀한 협력을 통해서만 실현될 수 있다는 의미에서 민간-공공 협력은 이를 위한 핵심적인 요소로서 여겨진다.

Findy의 몇몇 회원들은 MyData Global이 추진한 마이데이터 운동이 시작될 때부터 강력한 지지자였다. 개인데이터에 대한 시민들의 자기 결정과 개인데이터로부터 혜택을 받을 수 있는 능력을 향상시키기 위한 마이데이터 선언과 그 목표의 추진을 위해서는 실용적 솔루션이 필요하다.

Findy의 임무는 이러한 솔루션을 개발하고, 개인과 조직 모두에게 프라이버시를 보호하는 방식으로 광범위한 ID 데이터를 관리하고, 제어할 수 있는 힘을 부여하는 것이다. 기술 발전은 이제 특정 서비스공급자에 대한 락인효과(Lock-in)를 지양하면서 모든 시민과 조직을 위한 새로운 유형의 탈중앙화된 신뢰 네트워크와 ID 지갑을 구축할 수 있는 길을 열었다. Findy 협동조합 구성원의 목표는 윤리적, 경제적, 사회적으로 지속가능한 데이터 경제에 공

동으로 기여하는 것이다.

Findy는 2021년 6월 3일에 EU집행위원회에서 제안한 유럽 디지털 ID 프레임워크를 환영하고 있는데, 그 프레임워크의 주요 원칙은 다음과 같다.

1. 자기 주권 데이터 관리: 사용자는 자신의 ID 데이터에 접근할 수 있는 사람과 방법을 스스로 제어한다.
2. 투명성: 누가 데이터를 어떤 조건으로 처리하는지에 대해 항상 명확해야 한다.
3. 지역 전체에서의 구현을 위한 상호운용성: 각 회원국은 ID 지갑 구현을 위한 각자의 접근 방식을 선택하지만 서로 상호운용이 가능해야 한다.

Findy는 ToIP 재단의 디지털 신뢰 아키텍처를 위한 모델을 기반으로 작업하고 있고, 몇 년 동안 검증된 데이터를 공유하기 위한 자기 주권 원칙 기반 접근 방식을 연구해 왔다. Findy는 상호 합의된 책임과 규칙을 이행하려는 모든 조직들에 참여의 문이 열려 있다.

Findy의 실제 경험은 유럽을 넘어 사용되어야 한다. 왜냐하면 자기 주권 원칙에 기반한 공유를 통해 프라이버시, 서비스 수준, 생산성, 부문 간 및 국가 간 상호운용성, 범죄 억제 관련 데이터를 시민들에게 이용하도록 하는 것에서 발생하는 이익이 더욱 확대되기 때문이다 (Reed and Preukshat(2021) pp. 58-82).

나. MyData4Pandemics

코로나19는 2021년 가을까지 통제될 수 있을지 모르지만, 전문가들은[6] SARS-CoV-2보다 더 치명적이고 전염성이 강한 새로운 발병을 예측하고 있다. 감염병 발생 시 의료적 대응이 최우선이다. 두 번째 우선순위는 국내 및 국제 차원에서의 이동 제한, 사회적 거리두기, 격리와 같은 비약물적 개입(non-pharmaceutical interventions, NPI)을 개발하는 동시에 사회 생활, 교육, 경제의 혼란을 최소화하는 것이다.

NPI가 효과적이기 위해서는 끊임없이 변하는 상황과 가장 잘 문서화된 최신 개인데이터를 통해 조정해나가야 한다. 현재의 감염병 감시 시스템은 익명화된 인구데이터를 기반으로 한다. 이때 당국이 개인데이터에 접근하기 위해 공익이란 명목으로 개인의 권리를 침해할 가능성이 있다[7]. MyData Global이 조직한 MyData4Pandemics 내에서 이에 대한 대응 방안 마련을 추진하고 있다. 시민들은 상호운용 가능하고, 세분화된 개인데이터를 통해 기존 시스템을 보완할 수 있도록 자신들의 데이터를 기꺼이 공유하려 하는 책임 있는 행위자가 되어야 할 것이다. 이를 위해서는 개인이 자신의 데이터를 제어할 수 있는 신뢰할 만한 솔루션이 제공되어야 한다. 나아가 발병이 잦아질 경우 이러한 솔루션은 지속가능하도록 확장 및 축소할 수 있어야 한다.

MyData Global은 오픈소스의 확장 가능한 디지털 서비스인 ICAWAS(Individual Centric Advisory and Warning Services)를 개발할 것을 제안한다. 이 서비스는 다양한 신기술을 통합하여 주요 (기존) 데이터 중개자들에게 신속하게 배포할 수 있고, 발병 직후 바로 가치를 제공할 수 있는 것이다. 이 서비스는 다음과 같은 주요 요소로 구성되어 있다.

1. 데이터 중개자(Data Intermediary)는 ID 관리, 동의 관리, 데이터 전송 등을 담당한다. 그리고 이는 솔루션의 중추를 구성하는 개인데이터 저장소 및 거버넌스에 의해 보완된다. 개인데이터 공유 솔루션이 확장됨에 따라 데이터 중개자가 점점 더 증가할 것으로 예상된다.
2. 지능형 개인 비서(Intelligent Personal Assistant)는 개인의 데이터 및 건강 지식 수준에 따라 조정해주는 역할을 한다. 그것은 개인이 상황에 맞는 전염병 정보의 특정 값(value)에 접근하는 것을 도와주고, 신뢰할 수 있는 데이터를 관리하며, 또한 개인이 그들의 개인데이터 품질을 개선하도록 유도한다.
3. 인구데이터 커먼즈(공유자원)(Population Data Commons)는 인구 수준에서 개인들이 공유하는 정보를 통합한다. 데이터 커먼즈에서 운영 부문을 도출함으로써 정부 당국은 NPI를 정기적으로 정의하고, 모니터링하며 개선할 수 있다.
4. 디지털 지갑(Digital Wallet)에는 팬데믹의 발생 시점부터 생성되고, 이동의 안전성 증대를 지원하기 위한 필수노동자 및 모든 인구에 대한 감염 위험, 감염 및 면역 검사 또는 백신접종 증거 등 다양한 유형의 자격 증명이 포함되어 있다.

ICAWAS는 모든 이해관계자들(사람, 민간, 공공부문/당국)의 이익을 위해 기존 주요 데이터 중개자 또는 데이터 이타주의 조직에 배포되어야 한다.

- 사람들이 자신의 데이터를 제어하고, 데이터 공유 과정에서 그들의 프라이버시가 보호되는 것을 확인하며, 상황에 맞는 정보와 상황에 따른 제어 패스(QR 코드)를 통해 권

6 IPBES(2020), 세계보건기구(2018)
7 세계보건기구(2021)

한이 부여된다.

- 민간 부문은 감염 및 건강 위험을 제한하면서 비즈니스 연속성을 최적화할 수 있고, 그들 정보의 프라이버시가 보호된다.
- 공공부문/당국은 고품질의 최신 데이터에 대한 정당한 접근권을 갖게 되며, 의료 시스템, 경제, 사회에 대한 부담을 줄이면서 시민과 관련된 적절한 NPI를 정의하거나 조정할 수 있다.

실제로 모든 기술이 존재하지만, 팬데믹의 특정 필요에 맞게 통합되고 맞춤화되어야 한다. ID, 거주지, 이동성, 고용과 같은 핵심 데이터는 일반적인 것으로 취급하여 데이터 중개자 내에서 사용할 수 있어야 한다. 증상, 치료, 검사, 백신 접종에 대한 의료데이터는 팬데믹에 따라 특별하게 취급된다.

다. Skills Alliance

Skills Alliance는 개인데이터의 자유로운 흐름을 위한 인간 중심 표준의 조정, 거버넌스, 실험을 위한 국제 컨소시엄인 NewGovernance의 중심 활동이다. 이 조직은 사람들의 스킬을 강화(skill up)하고, 교육·기술·고용 영역에서 혁신을 가져오기 위해 시작되었다.

스킬과 직업이 빠르게 진화하는 글로벌 환경에서 적절한 훈련과 직업이 적절할 때에 제안될 수 있도록 사람들의 스킬업을 제공하는 것은 필수적이다. 코로나19 위기로 세계 12억 명의 사람들이 실업의 위기에 처했다. 위기에 처한 영역에서 수요가 많은 영역으로 스킬을 쉽게 이전할 수 있게 되면 시민들의 실업 기간이 단축된다. 교육 분야의 경우, 학습 결함을 감지하고, 각 학생에게 맞도록 학습 내용을 조정하기 위해 그들의 데이터를 사용할 수 있게 하는 것이 어린 학생들이 잘 성장할 수 있게 하는 데 필수적이다. 이것은 개인 서비스가 연결된 이해관계자들의 네트워크에 의해서만 수행될 수 있다. 개인들만이 자신의 동의를 통해 데이터 소스를 연결할 수 있고, 자신의 스킬 데이터가 교육이나 고용에 사용되도록 허용할 수 있다. 이러한 이유로 주요 스킬과 교육 관련 기관들(프랑스 교육부, 프랑스 노동부, 스웨덴 고용청, 핀란드 국가고용국, 헤이크 대학, 소르본 대학, 스탠퍼드 대학, 후지쓰, 비전스, HeadAI, Vastuu Group, MyData Global 등)이 세계 전역에서 모여 Skills Alliance를 구성하고 있다.

Skills Alliance의 회원들은 서비스를 개선하기 위한 스킬 데이터 이동성의 기회를 활용하고 싶어하고, 현재는 공통 표준을 통해 인간 중심 스킬 데이터 이동성을 가능하게 하기 위한 구체적인 솔루션을 만드는 작업 중이며, 국제적 인간 중심의 스킬 데이터 공간을 위한 길을 닦고 있다. Skills Alliance는 이러한 인간 중심의 데이터 흐름을 가능하게 하고, 해결해야 할 문제를 식별하기 위한 솔루션 및 거버넌스 모델에 대한 작업을 하고 있다. 이 모델은 동일한 인프라 및 거버넌스를 공유하는 것을 목표로 협력하는 세계 전역의 데이터 공유 활동을 통합한다. 회원들은 매달 가장 혁신적인 프로젝트를 공유하기 위해 의견을 모으고, 비전을 구현하기 위해 쌍방향 워크숍에서 공동작업을 한다. 그 조직의 하위 워킹그룹에서는 거버넌스, 스킬, 비즈니스, 사용 사례 등 다양한 트랙을 통해 지속적인 작업을 한다.

유럽에서 Skills Alliance는 EU집행위원회와 유럽 스킬 데이터 공간 구축에 대해 긴밀하게 협력하고 있으며, 스킬 데이터 공간 구축을 위한 기금이 마련되어야 한다고 주장했다. 현재는 디지털 유럽 프로그램이 유럽 스킬 데이터 공간 구축에 자금을 지원하고 있다.

프랑스에서는 스킬 및 교육데이터 공유를 위한 전국 공공·민간 인프라인 Prometheus가 관련 시스템의 국제 협력 및 상호운용성을 확보하기 위해 Skills Alliance와 협력하고 있다. 현재 보다 많은 구체적인 작업들이 진행되고 있고, Skills Alliance는 이러한 생태계를 통합하는 역할을 할 것이다. 이러한 국제적인 움직임을 통해 사람들이 각자의 스킬 데이터로 자기계발 및 전문성 개발 역량을 증진할 수 있도록 열심히 노력하고 있다.

제1장.
Data Sovereignty[8]

Co-Author: MyData Global Leadership Team
Lead Author: Viivi Lähteenoja

As MyData Global, we are most concerned with personal data generated by and about individuals and groups of people. In this paper, we define data sovereignty as the supreme authority of a person concerning the digital domain particular to them. We highlight two aspects of data sovereignty as understood in the context of MyData Global: what it means to exercise it and what it means to have it. In Part 1, we relate the MyData principles to our definition of data sovereignty. In Part 2, we describe ongoing initiatives seeking to realize and concretize different aspects of personal data sovereignty.

1. Introduction

Data sovereignty, while not an entirely new concept, has gained recent attention especially in the wake of the European Commission's data strategy and its use of the term "technological sovereignty" (European Commission 2020). Interpretations of the meaning of, and motivation behind, this kind of sovereignty have ranged from data localization and regional protectionism to economic and geopolitical ambition and again to the defense of basic digital human rights of privacy and digital self-determination. What's common to all of these interpretations, without commenting on their accuracy, is that they conceive of this new kind of sovereignty-variously called technological, information, digital, or data sovereignty-in terms of the basic element of traditional definitions of sovereignty that is geographic territoriality.

Sovereignty is generally understood as "supreme authority in a territory" and has traditionally been the domain of political theory regarding the most superior legitimacy to wield power with respect to territories or, usually, states (Philpott 2020). Discussions inspired by the EU's commitment to "European technological sovereignty" usually take this sovereignty to be of the traditional, nation-based kind, while emphasizing that the analogue world is no longer the only dimension in which the citizens of nations exist and function while subject to their nation state's sovereign authority: they are also active in the digital realm. Consequently, technology and infrastructure, as well as information and data, ought to be governed by and aligned with the laws, needs, and interests of that state in which their user is. This is a kind of extension of the basic concept of sovereignty which conceives of it in terms of nation states (or, in the case of the EU, a collection of nation states) and their supreme authority in a geographical territory.[9]

In this chapter, we discuss a different concept of sovereignty. Taking the traditional definition's elements of supremacy, authority, and territoriality, we substitute the last entirely. Data sovereignty, in our definition in this chapter, means "supreme authority in a defined digital space" and makes no reference to states or geographies. This concept of data sovereignty refers to actors-individuals, groups, organizations, and institutions-regardless of their physical location or present jurisdiction, and their sovereignty over the data generated by and about themselves-regardless of that data's physical location or present jurisdiction.

As MyData Global (mydata.org), we are most concerned with personal data generated by and about individuals and groups of people. This is not to say, however, that data sovereignty as we discuss it here cannot be fruitfully applied also to organizations and institutions who generate data and about which data is

8 1부 1장은 해외 집필진의 영문 저작물로 국내 독자의 이해를 위하여 국문 번역본을 수록하였습니다. 국문 번역본은 36쪽 참고 부탁드립니다.

9 It should be noted here that a special use of the traditional, geographically delineated data sovereignty has emerged that specifically addresses data related to indigenous populations in and between postcolonial states and, more broadly, groups of people with a salient identity not defined along the same lines as the borders between states; cf. e.g., (Rainie et al. 2017). These conceptions of indigenous data sovereignty rely on some of the same observations that we do in building our definition of data sovereignty, described below.

generated. But they are not our focus here. Similarly, while we do not here discuss the data sovereignty of states or territories, we are confident that some of the points made here can also inform that discussion.

Our objective in this chapter is to highlight two aspects of data sovereignty as we understand it: what it means to exercise it and what it means to have it. In other words, we outline our vision of how to make data sovereignty happen, and how to make it right. We begin with the latter and continue to the former before concluding with practical recommendations for data professionals in any capacity to take steps toward data sovereignty.

2. Make it right

A. Our definition of data sovereignty

> Data sovereignty is understood here as the supreme authority of a person with regard to the digital domain particular to them.

Authority must be conceived of not merely as coercive power but as a right: the right to command and correlatively the right to be obeyed (Wolff 1970). Authority must therefore also be understood as legitimate, i.e., as deriving from some mutually acknowledged source of legitimacy such as a body of law.

Supremacy, on the other hand, signals that other authorities with regard to this digital domain of sovereignty are subordinate to the sovereignty of the person themselves.

Finally, the digital domain particular to a person is defined as data generated by or about that individual person. This aspect of our definition is what sets it apart from the usual discourse on territorial data sovereignty and, we hope, also sheds new light on that discourse. Using the individual, as opposed to a geographical region, as the point of integration and the limiting factor for data sovereignty, allows us to think more flexibly about our current situation in which people-and-data about them-are not strictly geographically bound.

B. MyData principles and data sovereignty

MyData Global maintains the MyData declaration, first published in 2017, which describes the six MyData principles. The table below describes how

[Table 1-1-1] MyData principles and data sovereign

MyData Principle	Description	Relevance to data sovereignty
Human-centric control	Individuals should be empowered actors in the management of their personal lives both online and offline. They should be provided with the practical means to understand and effectively control who has access to data about them and how it is used and shared.	Data sovereignty means people's and groups' right to self-determination, autonomy, and control over personal data about themselves.
Individual as the point of integration	The value of personal data grows exponentially with their diversity; however, so does the threat to privacy. This contradiction can be solved if individuals become the "hubs" where, or through which, cross-referencing of personal data happens.	Data sovereignty means taking the individual person (or a group of people) as the defining element of the realm over which sovereignty is held and exercised instead of the state or other geography in which they are.
Individual empowerment	In a data-driven society, as in any society, individuals should not just be seen as customers or users of pre-defined services and applications. They should be considered free and autonomous agents, capable of setting and pursuing their own goals. They should have agency and initiative.	Data sovereignty is something that people and groups not only (should) have, but also something that can exercise as agents and active participants in data ecosystems.
Portability: access and re-use	The portability of personal data, that allows individuals to obtain and reuse their personal data for their own purposes and across different services, is the key to make the shift from data in closed silos to data which become reusable resources. Data portability should not be merely a legal right, but combined with practical means.	An important aspect of exercising data sovereignty is the right and, importantly, the means to access, port, and reuse data about oneself.
Transparency and accountability	Organizations that use a person's data should say what they do with them and why, and should do what they say. They should take responsibility for intended, as well as unintended, consequences of holding and using personal data, including, but not limited to, security incidents, and allow individuals to call them out on this responsibility.	Data sovereignty means that the legitimacy of personal data use is always derived from the person or group about whom that data exists. Transparency over where organizations derive the right to use personal data and the ability to hold those organizations accountable are necessary for ensuring data sovereignty is respected.

MyData Principle	Description	Relevance to data sovereignty
Interoperability	The purpose of interoperability is to decrease friction in the data flow from data sources to data using services, while eliminating the possibilities of data lock-in. It should be achieved by continuously driving towards common business practices and technical standards.	Data sovereignty means freedom from external interference in its exercise. Interoperability is a principle that support this freedom to choose the tools to use and the actors to collaborate with, without technological or other barriers.

C. The limits of data sovereignty

Sovereignty is traditionally conceived of as something absolute, and by absolute it is meant extending to all matters within a territory unconditionally.[10] In the case of data sovereignty, however, it is clear that this cannot be the case. In the interconnected digital realm, it is inconceivable that any person's authority could be absolute in this sense. My right to object to appearing in a photo you took of us and wish to publish is a classic example of personal data that is about more than just one individual and subject to authority by the both of us. In fact, it can be argued that most of the personal data generated about us is through interaction: either with other people or other data rights holders, such as companies and governments.

Secondly, it is also imperative to speak of data sovereignty here as something that is not only something held and exercised within a defined digital domain, i.e., with regard to data generated by and about a person, but also as something that both depends on and demands external recognition and respect. Data sovereignty, as we conceive of it here, means the freedom of external influence on its basic prerogatives or exclusive rights. But what exactly those basic prerogatives are is a difficult question to answer universally.

Taking into consideration these twin factors, the non-absoluteness of data sovereignty and the indefiniteness of what constitutes a data sovereign's basic prerogatives, it is clear that data sovereignty, like any system of sovereign entities like the world of sovereign states we live in, involves a great deal of dialogue and negotiation.

The value and utility of an individual's personal data is multiplied when it is shared and when it's combined with other people's data. In order to tap into this potential benefit from personal data, it is therefore critical that sovereign entities, people and groups, are able to cooperate and pool their resources. Benefits from such cooperation include support for research, policy making and authorities' decision making. Such cooperation, in turn, requires shared governance and shared commitment to that governance. Good data governance, then, takes people's sovereignty over data about themselves as a foundational principle. Just like sovereign states collaborate within and by committing to supranational frameworks like the EU and the UN, sovereign individuals and groups can abide by and commit to common governance structures such as data collaboratives or networks of data intermediaries.

D. Why is this important

Data sovereignty of people and groups is a critical component of the future world we, as MyData Global, want to see become real and which is described in the MyData declaration. The current paradigm of data use and how its benefits are distributed, with its asymmetries of power between people and organizations, is not sustainable or fair. Work is needed to ensure that people and groups have data sovereignty in principle and that they are empowered to exercise that data sovereignty in practice. Thus far we have focused on what data sovereignty is, next we give some examples of what having and exercising it might mean in different case examples.

10 It is worth noting that even traditional, territorial sovereignties are becoming increasingly non-absolute. The states that are a part of the UN, the EU, international criminal courts, and so on also yield some authority to those supranational institutions.

Case 1: **Data donation enabled by data sovereignty**

Data Sovereignty is critical for sharing and donation of personal data to benefit society - and therefore its members as individuals. In some cases, such as the pandemic we are now experiencing, there is an especially heightened sense of the individual's duty towards their community and society. One mechanism for people and groups to fulfill this duty is to donate personal (health) data to aid researchers and decision makers. This kind of donation is made possible when people are able to exercise their data sovereignty and put data to use for a purpose they choose.

Case 2: **Employee data sovereignty**

Data sovereignty is also important to companies that are responsible for the use of data about their employees as well as customers. Ethical use of employee data can be an advantage in attracting talent and therefore increasing innovation and customer satisfaction. Employee data sovereignty means giving employees access to the data gathered about them and their performance, and could enable offering tailored career services in-house or employees making use of third-party services to benefit themselves and so their employers.

Case 3: **Sensemaking toward confidence with data sovereignty**

Individual exercise of their data sovereignty requires a certain understanding of how data ecosystems and economies function and the ability to navigate within them. This kind of understanding and ability can result from collective sensemaking and will contribute towards the confidence to exercise one's data sovereignty in ways which benefit oneself and one's communities. Data sovereignty requires people and groups who have data sovereignty to make sense of the ecosystems and economies around them collectively and continuously in order to exercise that data sovereignty with confidence.

2. Make it happen

This second part of this chapter contains self-descriptions of some ongoing initiatives that seek to realize and concretize different aspects of personal data sovereignty.

A. Findy

The present ways to protect data and privacy and make data available to citizens and to fight fraud and grey economy are very costly and difficult for both users and service providers. The same goes when exercising data sovereignty to improve service levels and productivity with automation and real time use of both personalized and anonymised data. The main reason for this is that the internet was built without an identity layer.

The Self-sovereign Identity (SSI) initiative is aiming at solving this with open standards and to shift the control to the individual user. Much of the standardization work is organised by the Trust over IP Foundation (https://trustoverip.org) but also several others are involved.

Several programmes have already taken this approach. Findy (https://findy.fi/en/) is briefly described below.

The Findy Cooperative is a public-private not-for-profit organization that develops a general-purpose, shared and secure data network with the target to ensure the authenticity of information required in e-services.

The public-private aspect is seen as central as Innovative, practical and socially impactful data-driven services can only be realized through close cooperation between private and public sector organizations.

Several members of Findy have been strong proponents of the MyData movement since its inception. The MyData declaration and its aims to improve citizens' self-determination regarding personal data and their ability to benefit from it need practical solutions.

Findy's task is to develop these and empower both individuals and organizations to manage and control their wide identity data in a privacy-protecting way. The technology developments have now paved the way for establishing new types of decentralized trust networks and identity wallets for all citizens and organizations while avoiding vendor lock-in. The Findy cooperative members aim at jointly contributing to an ethically, economically and socially sustainable data economy.

Findy welcomes the framework for a European Digital Identity proposed by the European Commission on 3 June 2021. The key principles are:

1. Self-sovereign data control: users will themselves control who has access to their identity data and how.

2. Transparency: it should always be clear to the user who is processing their data and under which conditions.
3. Broad interoperability across local implementations: each member state will choose their own approach for implementing wallets, but they must be interoperable with each other.

Findy is building on the Trust over IP Foundation model for digital trust architecture and has been working on a self-sovereign principle-based approach for sharing verified data for several years. Findy is open to participation by all organizations that commit to jointly agreed responsibilities and rules.

Findy's practical experience should be used in Europe and beyond as the benefits of making data available for citizens with self-sovereign controlled sharing both for privacy, service levels, productivity (customers pay all the costs), cross-sector and cross-country interoperability and crime mitigation are so massive (as outlined in (Reed and Preukshat 2021) pp. 58-82).

B. MyData for Pandemics

While COVID-19 might be under control by autumn 2021, experts.[11] predict new outbreaks, more lethal, more contagious than SARS-CoV-2. The first priority in case of infectious disease outbreaks is medical countermeasures. The second priority is to develop non-pharmaceutical interventions (NPI)-such as mobility restriction, social distancing and confinement-at national and international levels while minimizing disruption on social life, education and economy.

To be effective, NPIs need to be adapted to a constantly evolving context, best documented through up-to-date personal data. Current infectious diseases surveillance systems are based on anonymized population data; to access personal data, authorities have a possibility to infringe on data privacy rights in the name of the common goods[12] Within MyData4Pandemics, we believe that alternative solutions are possible: we need to empower citizens to become responsible actors, willing to share their data to complement existing systems with interoperable, granular, personal data. This requires that individuals are provided with solutions enabling them to control their data in trust. In addition, as outbreaks come and go, these solutions must be able to scale up and down to be sustainable.

We propose to develop open-source, scalable, digital services-Individual Centric Advisory and Warning Services (ICAWAS)-integrating different emerging technologies that can rapidly be deployed on top of (existing) Data Intermediaries and deliver value soon after onset of an outbreak.

1. The Data Intermediary includes identity management, consent management, data transfer. It is complemented by Personal Data Stores and Governance, forming the backbone of the solution; it is expected that Data Intermediaries will increasingly be in place as personal data sharing solutions are expanding.
2. The Intelligent Personal Assistant adapts to the level of data and health literacy of individuals; it helps them to access pandemic information of specific value for their context, manages their data in trust, and incentivizes them to improve the quality of their own personal data profile.
3. The Population Data Commons consolidates information shared by the individuals at the population level. By deriving operational segments from the Data Commons, authorities are able to regularly define, monitor and refine NPIs.
4. The Digital Wallet includes different types of credentials - contagion risks for critical workers and for all the population, test of presence or immunity or vaccination proof ... - generated from the onset of the pandemic and supporting increased mobility in safety.

ICAWAS should be deployed on top of existing Data Intermediaries or Data Altruism organizations in a true PPPP for the benefit of all parties.

- People are in control of their data, ensured their privacy is preserved when

11 (IPBES 2020), (World Health Organisation 2018)
12 (World Health Organisation 2021)

sharing data, and empowered thanks to tailored information and context dependent control pass (QR code)

- Private sector is able to optimize business continuity while limiting contagion & health risks; privacy of their information is preserved.
- Public sector/Authorities have rightful access to high quality, up to date data and are able to define & adapt appropriate NPIs- in respect of citizens - while decreasing the burden on health care systems, the economy and the society.

While all technologies exist, they need to be integrated and customized for the specific needs of a pandemic: core data such as identity, residency, mobility, employment are generic and should be available within Data Intermediaries; medical data on symptoms, treatments, testing and vaccination are specific for each pandemic.

C. Skills Alliance

The Skills Alliance is a sectoral hub of aNewGovernance, an international initiative for the coordination, governance, and experimentation of human centric standards for the free flow of personal data. The Skills Alliance was launched to empower people and bring innovation in the education, skills and employment domain.

In a global context of quick evolution of skills and jobs, being able to empower people with their skills so that the right training and jobs are proposed to them at the right time is essential. The COVID-19 crisis threatens 1.2 billion people worldwide with unemployment; being able to easily transfer skills from sectors that are in crisis to in-demand sectors shortens the unemployment time for citizens. For education, being able to use pupil's data to detect learning deficiencies and adapt learning to each particular person is essential to allow each child to thrive. This can only be done by a network of connected stakeholders at the service of the individual. Only the individual can connect its data sources thanks to his consent and allow for his skills data to be used for its education and employment. It is for these reasons that major skills and education related organizations (French Ministry of Education, French Ministry of Labor, Swedish National Employment Agency, Finish National Employment Agency, University of the Hague, Sorbonne University, Stanford, Fujitsu, Visions, HeadAI, Vastuu Group, MyData Global, …) from all over the world are part of the Skills Alliance.

The Skills Alliance members want to seize the opportunity of skills data portability to better their services and are currently working on concrete solutions to allow a human-centric skills data portability, through common standards, and pave the way for an international human-centric Skills Data Space. The Skills Alliance works on solutions and governance models to enable this human centric data flow and identifies the problems to be addressed. It unites skills data sharing initiatives from all over the world that collaborate with the goal of sharing the same infrastructure and governance. Each month members unite to share the most innovative projects and collaborate in an interactive workshop to push the vision and its implementation forward. Subgroups work continuously on different tracks: governance, tech, business, use cases.

In Europe, the Skills Alliance works closely with the European Commission on the creation of a European skills data space. It advocated for funds to be released to build such a skills data space. Today, the Digital Europe Program will finance a skills data space in Europe.

In France, Prometheus, a national public-private infrastructure for skills and education data sharing, will collaborate with the Skills Alliance to ensure international cooperation and interoperability of its systems. Many more concrete works are happening and the skills alliance brings those ecosystems together.

This global movement is working hard to empower people with their skills data and make them the masters of their personal and professional development.

제2장.
데이터 활용의 민주화
: 시티즌 사이언티스트와 스마트시티

김찬수(투이컨설팅 상무)
이승봉(투이컨설팅 이사)

많은 조직이 데이터라는 금광에 앉아 있지만 데이터의 잠재적 가치를 실현하기 위해 데이터를 분석할 효과적인 리소스가 부재한 상황이다. 이를 해결하기 위해서는 데이터를 필요로 하는 모든 개인에게 쉽게 접근하고 활용할 수 있는 체계를 제공하는 '조직에서 개인으로의 데이터 활용 민주화'가 핵심적으로 필요하다. 또한 데이터 활용 민주화는 데이터 경제 시대와 맞물려 사회·경제 전반으로 확산되고 있다. 특히 스마트시티 구현을 위해서는 데이터 활용 민주화가 필수적이며, 스마트시티 생태계를 통해 사회 전방위적인 데이터 민주화를 촉진하고 있다.

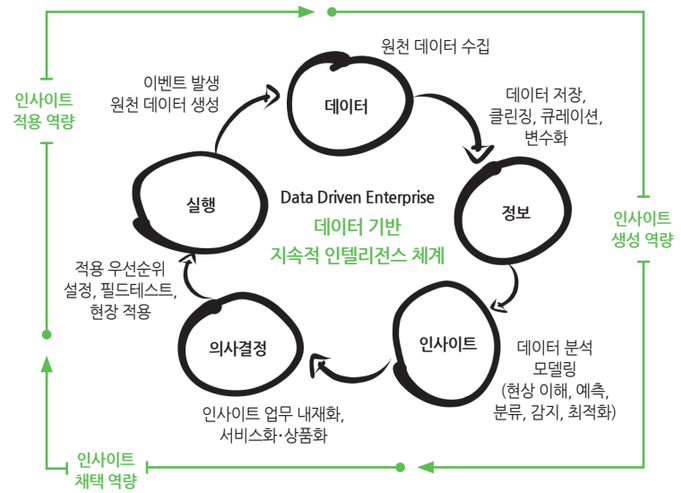
1. 데이터 활용 민주화의 등장 배경

가. 데이터 주도 기업(Data Driven Enterprise)으로의 변화

오늘날 선두 기업은 점점 더 데이터 중심적으로 변모하고 있다. 조직은 모든 의사 결정과 행동에 데이터 파워를 갖기를 열망하며, 비즈니스 가설과 아이디어 검증에 사용하고자 한다.

그러나 데이터 보유 자체만으로는 비즈니스의 성공을 보장하기에 충분하지 않다. 데이터를 사용하여 비즈니스의 가시적 성과를 만들어내기 위해, 그리고 데이터를 기반으로 지속적인 지능화 사이클을 운영하는 데이터 주도 기업 체계를 갖추기 위해 노력하고 있다. 데이터 기반 지속적 지능화 사이클이란 '데이터를 바탕으로 한 의사결정 및 실행까지의 주기'가 '일회성이 아니라 지속적으로 빠르게 순환'되는 것을 의미한다. 지속적 지능화 사이클을 갖추기 위해서는 인사이트 생성 역량, 인사이트 채택 역량, 인사이트 적용 역량을 갖추고 상호 유기적으로 연계되어 작동해야 하지만, 대부분의 기업들은 이러한 역량을 갖추고 연결하는 데 어려움을 겪고 있다.

[그림 1-2-1] 데이터 기반 지속적 지능화 체계



* 출처: 저자 작성

나. 지속적 지능화 장애 요인 해결을 위한 데이터 활용 민주화

대부분의 기업들은 데이터 수집-저장-처리-분석을 위한 리소스(애플리케이션, 인프라, 인력)를 소수의 데이터 전문가 팀에 의해 전담 운영하는 중앙집중적 방식을 취하고 있다. 즉, 조직을 위해 데이터를 적절하게 구성, 처리 및 해석하는 데 필요한 기술과 이해를 갖춘 소수의 데이터 분석 전문가에게 데이터 업무가 집중되어 왔다. 이는 여러 상황에서 데이터 기반 지능화의 장애 요인으로 작용하고 있다.

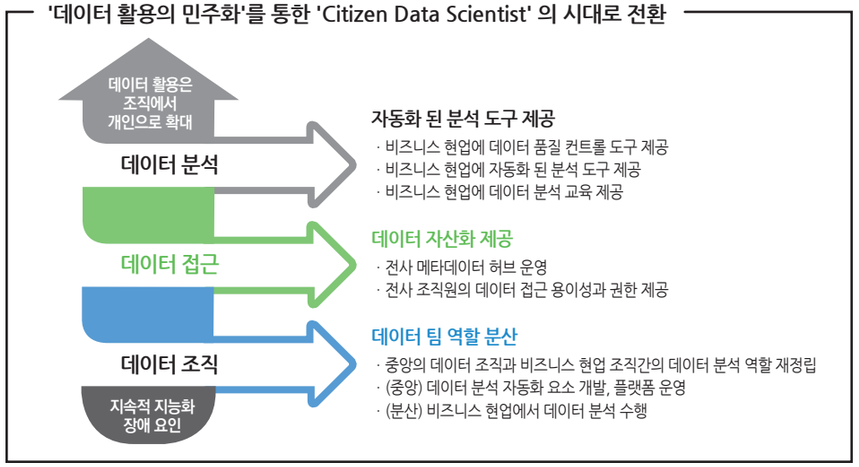
[표 1-2-1] 지속적 지능화의 장애 요인

구분	장애 요인
데이터 조직 데이터팀의 병목 현상	- 수많은 비즈니스 부서의 데이터 요청, 분석 모델 개발 요청 대응에 많은 시간 소요 - 데이터분석가는 분석 결과를 데이터·통계 이해도가 부족한 비즈니스 현업에 전달하는 데 어려움을 겪고 있음 - 데이터분석가는 비즈니스 현업 부서가 요청한 질문을 정확히 이해하지 못하여 적절한 분석 결과를 도출하지 못함(비즈니스 현업 부서가 데이터분석가에게 질문을 적절하게 설명하지 못하는 경우 역시 포함)
데이터 접근 비즈니스 현업 조직의 데이터 접근에 대한 제한	- 비즈니스 현업은 최종 정리된 데이터나 보고서에만 주로 접근 가능, 비즈니스 기회를 찾을 수 있는 다양한 데이터세트에 대한 접근 어려움 - 비즈니스 현업 조직에 데이터 프로세스 가시성 제공 미흡(데이터 계보 정보 제공 미흡) - 비즈니스 현업이 이해할 수 있는 비즈니스 관점의 데이터 설명 부족(IT 메타정보 중심, 비즈니스 메타정보, 운영 메타정보 제공 미흡)

구분	장애 요인
데이터 분석 비즈니스 현업 조직의 데이터 분석 모델링 과정 참여 제한	- 데이터 전처리, 변수화(Feature Engineering) 과정에 대한 비즈니스 현업 조직 미참여로 비즈니스 요구사항 반영의 한계 - 데이터 분석 전문성 부족에 따른 비즈니스 현업의 접근 한계

* 출처: 저자 작성

[그림 1-2-2] '데이터 활용 민주화'를 통한 지속적 지능화 장애 요인 극복



* 출처: 저자 작성

열거한 장애 요인을 해결하기 위해서는 데이터 조직, 데이터 접근, 데이터 분석의 3가지 관점에서 접근이 필요하다. 대부분의 데이터에 대한 접근 권한과 도구는 중앙의 데이터팀에 있으며, 이는 데이터 접근과 활용의 사일로(Silo)를 만들어 내고 있다. 데이터 분석 모델을 생성하고 비즈니스 가치를 실현하는 속도는 고용한 데이터엔지니어, 데이터분석가의 수에 정비례한다. 이러한 데이터팀의 확장성 문제는 데이터 분석 비용을 크게 증가시킨다. 또한 데이터분석가에 대한 수요를 시장에서 대응하지 못하고 있다.

그러므로, 데이터에 대한 접근을 전사 조직원으로 확대하고 비즈니스 현업이 스스로 분석할 수 있는 자동화된 분석도구 제공이 필요하다. 이를 통해 중앙에 집중된 데이터팀의 역할을 조직 전체로 확장하여 전 조직원 각자가 시티즌 데이터 사이언티스트(Citizen Data Scientist)로서 데이터를 활용하는 '데이터 활용의 민주화'가 요구되는 시점이다.

Gartner는 '2020년을 위한 10대 전략 기술 동향' 중 하나로 '데이터 활용 민주화 시대'의 도래를 꼽았다. 데이터 활용 민주화는 데이터팀이나 IT 조직의 개입 없이 비즈니스 현업이 데이터에 언제든지 쉽게 접근·이해·활용할 수 있도록 하는 것으로 정의할 수 있다.

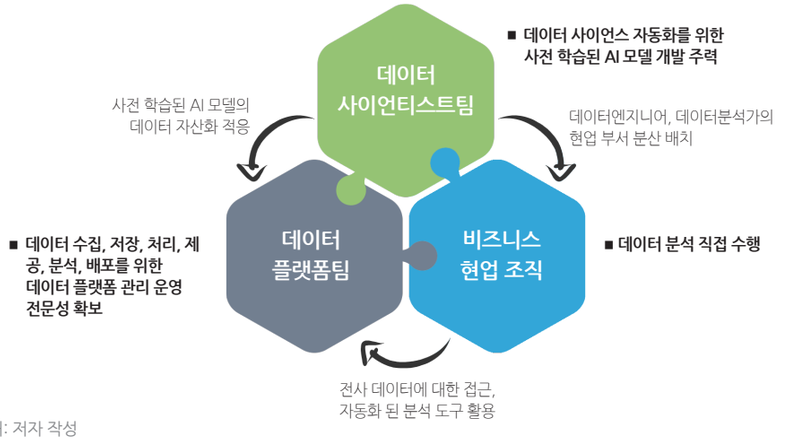
2. 데이터 활용 민주화 방안

데이터 분석·활용은 조직에서 조직원 개인으로 확대되는 데이터 활용의 민주화가 현재진행형으로 확산되고 있으며, 데이터 활용 민주화는 기업의 지속적 지능화를 작동시켜 데이터 드리븐 기업으로의 경쟁력을 강화하는 중요한 요소로 작용하고 있다.

가. 데이터팀의 역할 분산

가장 먼저, 기존의 중앙집중적 데이터팀의 역할 분산이 필요하다.

[그림 1-2-3] 데이터팀의 역할 분산



* 출처: 저자 작성

1) 데이터 사이언티스트팀 역할 변화

중앙의 데이터 전담 조직은 데이터 분석 자동화를 위한 AI 모델을 지속적으로 연구·개발

하는 팀(데이터 사이언티스트)과 데이터 플랫폼을 관리하는 팀으로 역할 분담이 필요하다.

즉, 중앙의 데이터 사이언티스트팀의 역할은 기존의 비즈니스 분석 모델 개발(추천모델, 이탈모델, 고객 세그먼트 등)과 비즈니스 현업 부서의 다양한 분석 요구 대응 역할에서 '데이터 사이언스 과정 자동화'가 가능한 AI 모델 개발에 주력하여, 비즈니스 현업 부서가 활용할 수 있도록 데이터를 제공하는 역할로 변화가 필요하다. 예를 들면 중앙의 데이터 사이언티스트팀에서는 이미지 인식에 대해 사전 훈련된 AI 모델, 텍스트 마이닝을 할 수 있는 NLP AI 모델, 데이터 전처리를 자동화할 수 있는 모델 등을 수집, 연구, 개발하여 제공하고, 비즈니스 현업 부서에서는 이를 활용하여 비즈니스 분석 모델을 직접 개발하는 형태로의 역할 변화가 필요하다.

2) 데이터 플랫폼팀의 전문성 강화

데이터에 대한 수집, 저장, 처리, 제공, 분석, 배포를 위한 데이터 플랫폼 관리·운영 전담 조직의 역할 강화가 필요하다. 기하급수적으로 발생하는 내외부 데이터와 다양한 유형의 데이터스택(Data Stack)으로부터 데이터를 수집, 저장, 처리, 제공, 분석, 배포하기 위해서는 과거의 정형 데이터, 배치중심 기술 인력에서 빅데이터·AI·실시간 기술을 이해하고, 이를 데이터 플랫폼에 지속적으로 반영하고 운영할 수 있는 전문성 있는 전담 인력 강화가 필요하다.

3) 비즈니스 현업 조직의 데이터 분석 직접 수행

비즈니스 현업 조직이 데이터 분석 역할을 이관받아 직접 수행하는 것이 필요하다. 앞서 언급했듯이, 중앙집중적 데이터 사이언티스트팀은 시간적, 비용적, 효과적인 분석을 수행하는 데 한계를 보이고 있다. 이를 해결하기 위해서는 우선적으로는 데이터엔지니어와 데이터 분석가를 각 팀에 분산시켜 현업의 분석업무를 현장에서 직접 지원하고, 이와 맞물려 데이터 자산화를 통한 현업의 데이터 접근 확대와 자동화된 데이터 도구의 제공, 교육을 통해 현업이 스스로 데이터를 분석할 수 있는 체계로 전환해 나가야 한다.

나. 데이터의 자산 활용 체계 제공

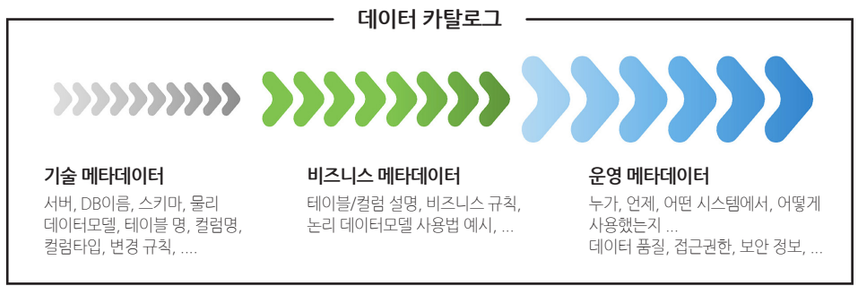
데이터 분석 업무를 비즈니스 현업으로 확산시키기 위해서는 모든 조직원이 전사 메타데이터 허브를 통해 분산되어 있는 데이터에 접근하고, 이해하고, 활용할 수 있는 데이터 자산 활용 체계가 필요하다. 전사 메타데이터 허브는 데이터 카탈로그와 데이터 검색 엔진이 포함된 데이터 포털의 형태로 구현되고 있다.

1) 확장된 데이터 카탈로그 접근 제공

데이터 카탈로그는 기존의 IT 관리 중심의 기술 메타데이터 정보에서, 비즈니스 현업 조직이 이해하고 활용할 수 있는 비즈니스 메타데이터와 운영 메타데이터로 확장되어 제공되어야 한다.

또한 데이터 카탈로그는 데이터 자체에 대한 메타데이터뿐만 아니라, 데이터가 어떻게 생성됐고, 어떤 과정을 거쳐 변경되었으며, 어디에 쓰이고 있는지에 대한 데이터 흐름 가시성 확보를 위한 데이터 계보(Data Lineage) 정보, 분석·활용을 위한 직관적인 도구와 사용자 인터페이스, 코드 조각 및 플러그인, 오픈소스 데이터 및 코드, 그리고 사전 훈련된 AI 모델 등이 포함되어 제공될 수 있도록 확장되고 있다. 즉, 메타데이터 자체가 또 하나의 중요한 데이터셋 개념으로 운영되는 형태로 변화하고 있다.

[그림 1-2-4] 데이터 카탈로그의 확장된 메타데이터 제공 범위

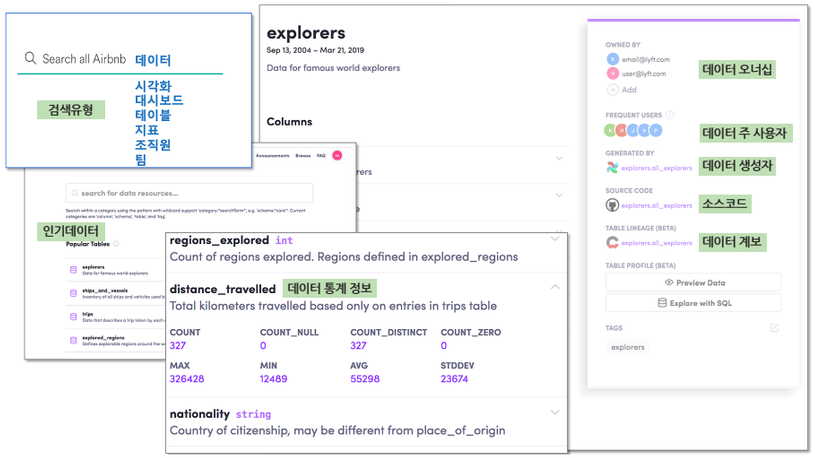


* 출처: 저자 작성

2) 데이터 검색 엔진과 데이터 포털 기능 강화

데이터 검색 엔진은 확장된 데이터 카탈로그를 기반으로 데이터 자체에 대한 검색뿐만 아니라, '데이터를 어떻게 사용하는지, 데이터는 어디에서 어떻게 변경되어 왔는지, 어떤 데이터가 가장 많이 활용되고 있는지, 해당 데이터의 오너십과 전문가는 누구인지, 해당 데이터는 누가 사용하고 있는지, 데이터 품질 수준은 어떤지' 등을 제공할 수 있어야 한다. 이러한 확장된 범위의 데이터 카탈로그와 강화된 검색 엔진은 Airbnb의 Dataportal, Uber의 Databook, Netflix의 Metacat, Lyft의 Amundsen, Google의 Data Catalog, LinkedIn의 DataHub 등의 글로벌 선진 기업에서 데이터 포털로 제공되고 있다.

[그림 1-2-5] Airbnb의 Dataportal, Lyft의 Amundsen



* 출처: 각사 해당 서비스 캡처

다. 자동화된 분석 도구 제공

비즈니스 현업 조직이 스스로 데이터를 분석하기 위해서는 데이터 품질 컨트롤 도구와 자동화된 분석 도구 제공이 필요하다.

1) 데이터 품질 컨트롤 도구 제공

각 조직에 걸쳐 분산되어 있는 데이터 테이블[1] 생성으로 중앙의 데이터팀에서 데이터 품질을 컨트롤하는 데 한계를 보이고 있으며, 이는 데이터 품질 저하로 이어지고 있다. 이를 해결하기 위해서는 각 팀이 생성한 테이블에 대해 직접 품질 컨트롤을 수행할 수 있는 도구 제공이 필요하다.

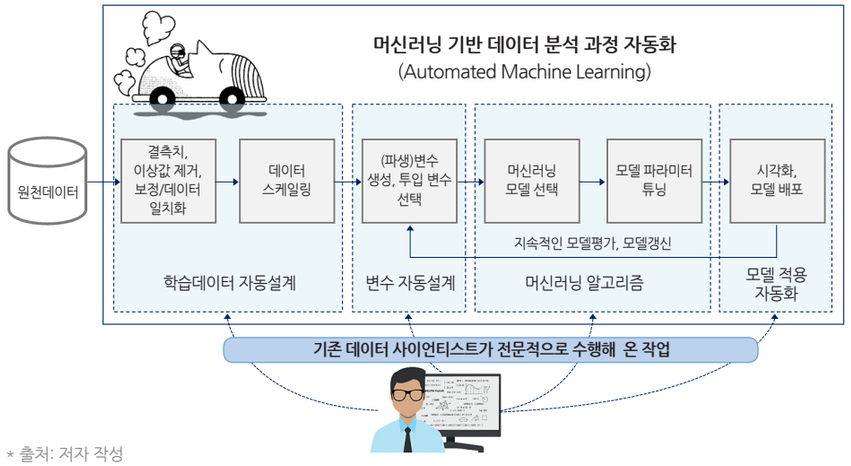
2) AutoML(Automated Machine Learning) 도구 제공

기존에는 데이터 분석을 위해서 데이터 분석 전문가와 IT 개발자가 데이터 수집-저장-처리-분석-적용에 이르는 각 단계별로 R과 Python 등의 데이터 처리·분석 언어를 이용하여 코딩을 수행해왔다. 이러한 전문성이 요구되는 데이터 분석업무를 프로그래밍 언어 지식과 통계지식이 취약한 비즈니스 현업 조직에 교육으로 이양하기에는 어려움이 존재했다. 그러나

1 행과 열로 구성된 데이터 모음

AutoML, Low Code·NoCode, Visual Analytics 등 여러 도구의 등장으로 데이터 분석업무의 자동화가 분석 과정의 많은 부분을 비즈니스 현업 조직에서 직접 수행할 수 있는 환경으로 변화시키고 있다. 코딩 최소화 또는 코딩 없이 분석할 수 있는 도구를 통해 비즈니스 현업 조직의 직접 분석이 가능해지고 있다.

[그림 1-2-6] 데이터 분석 과정 자동화 영역



* 출처: 저자 작성

대표적인 분석 자동화 솔루션 선두그룹으로는 DataRobot, H2O.ai Driverless AI 등이 있다. 이외에도 프로그래밍 없이 애플리케이션을 개발·배포할 수 있는 Microsoft Power Apps, 시각화 분석을 통해 데이터 패턴을 쉽게 파악할 수 있는 Tableau, QlikView, Power BI 등도 비즈니스 조직의 데이터 분석 접근을 용이하게 하는 데이터 분석 자동화 솔루션이다.

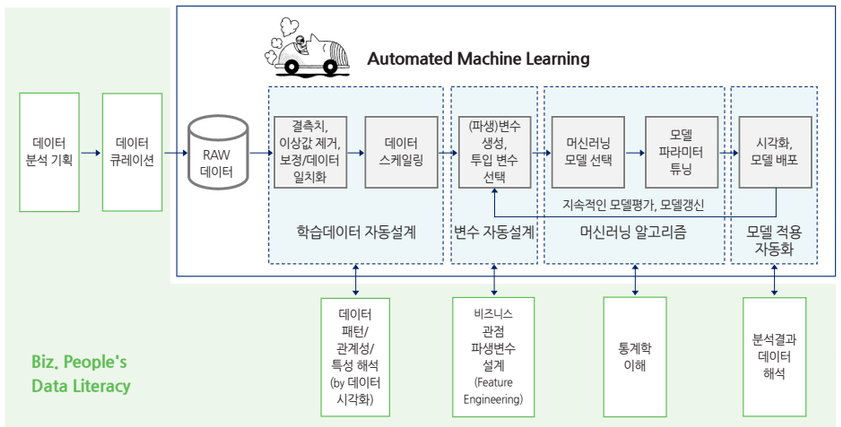
3) 데이터 분석 교육 제공

비즈니스 현업의 데이터에 대한 이해도와 데이터 분석의 직접 수행 역량 강화, 조직 전반에 걸친 데이터 기반 의사결정 문화를 정착시키기 위해서는 데이터 교육을 통한 전 조직원의 데이터 리터러시 확보 또한 병행되어야 한다.

비즈니스 현업의 데이터 리터러시 확보를 위한 교육은 분석 언어 교육에 치중하기 보다는 조직의 문제를 해결할 수 있는 적절한 질문을 찾아내고(Problem Framing & Right Question), 필요한 데이터를 찾아내고(Data Curation), 다양하고 적합한 변수를 설계하고

(Feature Engineering), 자동화된 분석모델의 옵션을 조정하고, 분석 결과의 의미를 제대로 해석할 수 있는 통계적 이해 역량 강화에 집중해야 한다.

[그림 1-2-7] 비즈니스 현업의 데이터 리터러시를 위한 교육 영역



* 출처: 저자 작성

라. 데이터 거버넌스와 인공지능 공학 체계 운영

데이터 활용 민주화를 위해 필요한 데이터 분석의 역할 분산, 데이터 자산화 관리와 운영, 자동화된 분석 도구의 도입과 교육을 효과적으로 기획·실행·관리하기 위해서는 비즈니스 전략과 데이터 전략을 일치(Alignment)시키고, 데이터에 대한 역할과 책임을 관리하고, 데이터 플랫폼 인프라와 기술을 관리하고, 전사 데이터 리터러시[2] 향상을 위한 교육을 기획하는 '데이터 거버넌스 체계'와 '인공지능 공학(AI Engineering: DataOps+ModelOps+DevOps) 체계'가 기반이 되어야 한다.

2 데이터를 목적에 맞게 활용할 수 있는 데이터 해석 역량

3. 스마트시티와 데이터 활용 민주화 확산

가. 사회 전 영역의 조직에서 개인으로의 데이터 주체 변화

데이터 경제에서 데이터 활용은 조직에서 개인으로 데이터 활용 주체가 변화하고 있다. 이는 기업 내부만의 움직임이 아니라, 우리 사회 경제 전방위적으로 나타나고 있다. 기업에서 고객으로, 도시나 지자체에서 시민으로 데이터 활용 주체가 변화하고 있다. 기업이 보유한 기존의 사일로(Silo)화된 기업 내부 데이터들은 협력사, 공공기관, 데이터마켓 등 기업 외부의 다양한 원천을 통해 수집된 데이터들과 결합되어 빅데이터로 만들어지고, 빅데이터 활용을 기업 일부 조직이 아닌 전 조직원이 다루도록 되었으며, 이제는 고객(시민)들이 주도적으로 데이터의 활용과 유통을 결정하는 시점에 있다.

나. 데이터 민주화를 통한 스마트시티의 촉진

마이데이터로 촉발된 개인의 자기정보(결정권)와 개인의 디지털 데이터 자산에 대한 의식 변화로 인해, 시민들의 마이데이터와 시티데이터들이 상호작용하면서 만들어지는 다양한 활용이, 스마트시티를 통한 새로운 데이터 민주주의의 성공 사례로 급부상하고 있다.

우리가 살아가는 장소이자 교통, 교육 등 다양한 기능들이 총체적으로 연계된 거대한 체계가 운영되는 공간인 스마트시티는 도시 운영체계의 중심축이 공공데이터에서 민간데이터 그리고 시민들의 개인데이터로 이동하고 있다.

스마트시티의 다양한 정의를 요약해보면, i) 시민이 참여하여 도시를 운영하고, ii) 다양한 ICT를 활용하여 도시를 지속적으로 혁신하여, iii) 시민의 삶의 질 향상과 도시문제를 해결하는 지속가능한 도시이다. 즉 시민 개개인이 도시의 운영주체로서, 시티즌 사이언티스트로서 스마트시티를 만들어가는 것이다.

진정한 스마트시티를 실현시키기 위해서는 무엇보다 다양한 시티데이터가 필수 요소이다. 스마트시티는 데이터시티로 불러도 될 만큼, 스마트시티 내의 모든 서비스와 도시 기능의 흐름은 데이터의 교환과 활용을 바탕으로 이루어진다. 우리나라의 스마트시티 시범도시인 부산 에코델타시티와 세종 5-1생활권 도시 계획에서도 도시가 지향하는 최종 목표로서 '데이터에 기반한 미래 도시'를 제시하고 있다.

다. 스마트시티데이터의 범위 확장

스마트시티에서 개인(시민)은 시티데이터의 생성·유통·활용 주체로서 도시 운영을 책임

짐과 동시에 도시 서비스를 이용하고 있다. 이렇듯 스마트시티는 데이터 민주화를 촉진할 뿐 아니라 필요로 하며, 이에 따라 스마트시티의 데이터는 다음과 같은 범위로 확장되고 있다.

[표 1-2-2] 스마트시티데이터의 범위

구분	내용
뉴타입 데이터	공공데이터 + 민간데이터 + 시민데이터가 결합되어 새로운 형태의 데이터를 생성하고 활용
시민데이터 멀티 도메인 모델	시민데이터 활용이 특정 도메인에 국한되지 않고 다양한 도시 도메인 모델로 활용
구독 도시 모델	모든 시티데이터는 결국 구독도시 모델로서 평가받고 활용 (과금을 하는 데이터부터 무과금 데이터까지 포괄적 구성)되며 구독의 주체는 시민이며, 서비스 제공자는 도시 운영 주체임

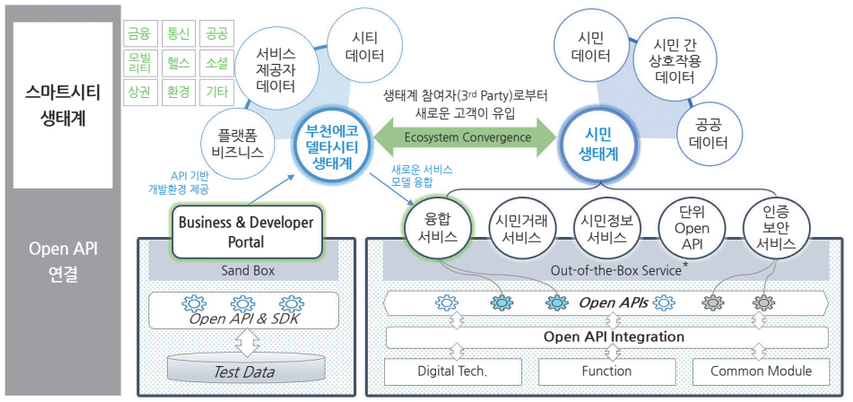
* 출처: 저자 작성

라. 데이터허브 기반 스마트시티 생태계 구성

스마트시티에서 시민 중심 도시 생활 혁신의 에너지원인 데이터는 도시 생명력과 지속가능성을 보장하는 핵심 역할을 수행하고 있다. 스마트시티 데이터허브를 통해, 시티데이터들의 경계와 구분을 제거하고 데이터 주체 혹은 사용 주체 관점에서 데이터 생태계를 새롭게 구성하는 것이 반드시 필요하다. 이 때문에 스마트시티에서는 셀 수 없이 많은 종류의 데이터와, 한정할 수 없는 데이터 사이즈, 실시간 데이터들을 조율, 수집, 저장 및 활용하는 데이터허브(시티허브플랫폼 혹은 스마트시티데이터플랫폼)가 필수적이다.

스마트시티의 데이터 인프라는 ▲다양한 도시 기능을 담당·처리하는 ICT 기반 정보시스템(H/W와 S/W가 결합된)들이 상호유기적으로 작용하고, ▲모든 시민의 주도적 참여를 보장하며, ▲다양한 이해관계자들이 디지털 경제를 생성·운영하는 생태계 환경을 제공하며, ▲다품종·대량의 시티데이터를 수집·처리·가공·분석·활용하여 디지털 도시 운영을 지원하는 기반이다. 또한 스마트시티 인프라는 다수의 시스템들과 플랫폼들이 상호운용성을 확보하여 구성된 거대한 디지털 체계이다.

[그림 1-2-8] 스마트시티 생태계 구성 예시



* 출처: 저자 작성

마. 시티데이터 민주화

우리가 살아가는 도시는 우리 개개인 인생(Life)의 시작부터 끝까지 전체 생애에 걸쳐 데이터를 생성·기록·활용하게 된다. 공동체에서 나의 데이터는 단계별(유치원, 초중고등학교, 대학교, 직장, 은퇴)로 타인들과 관계를 형성하게 된다. 즉 나의 도메인 영역(교육, 건강, 부동산/이사, 결혼, 육아, 창업, 교통, 여행/휴식 등)에 따라 데이터가 만들어지고 네트워킹된다.

우리가 만드는 데이터는 크게 두 가지 접근법으로 구체화될 수 있다. 첫 번째는, 데이터 생성 및 활용 주체인 우리 데이터를 마이데이터 개념에서 발전시켜 데이터 카탈로그 및 데이터 히스토리를 나를 중심으로 구성하는 것이고, 두 번째는, 도시 기능 운영 관점에서 각 기능별 데이터 구조와 관계를 구성하고 그중 시민데이터와 연계된 부분을 식별하여 시민과 상호작용하는 것이다.

스마트시티의 데이터는 민주주의 사상이 이상적으로 실현 가능한 형태로서 다양한 데이터들이 상호 공존하고, 함께 시너지를 만들어가기 때문에 그 무엇보다 개별 데이터에 대한 독립된 아이덴티티(Identity) 부여, 시티데이터 생애주기 거버넌스, 시티데이터 리터러시 확보를 위한 평생 교육, 시티데이터 라이브(Live)화가 실현되어야 한다.

1) 독립된 데이터 아이덴티티(Identity) 부여 및 시티데이터 거버넌스

스마트시티 뉴타입 데이터에 대한 아이덴티티 부여가 필수적이다. 아이덴티티는 시민데

이터, 민간데이터, 공공데이터를 포괄해야 한다. 시민데이터는 우리를 식별할 수 있는 마이데이터에 대한 아이덴티티를 구성하고, 민간데이터는 건물 등 민간 자산에 대한 데이터를 구분한다. 공공데이터는 각 데이터에 대한 목적과 활용도를 고려하여 아이덴티티를 부여해야 한다.

스마트시티 생애주기 거버넌스는 스마트시티의 핵심 자원인 데이터 구조 및 표준 정의를 중심으로, 스마트시티 서비스의 유연성과 안정성을 위해 데이터가 유통되고 활용되는 데이터허브를 효과적으로 지원할 수 있는 데이터 모델 관리를 수행한다.

[표 1-2-3] 시티데이터 거버넌스 범위

데이터 거버넌스 원칙 및 전략 수립	도시 전체 기능 영역을 포괄하는 원칙과 방향성(프레임워크)을 수립하고 향후 확장 가능한 전략 수립
데이터 표준 및 운영 관리체계 수립	데이터 프레임워크에 기반하여 표준, 구조, 흐름, 품질 영역을 대상으로 관리 체계(관리 대상 정의 및 대상별 관리절차) 세부 정의
데이터 모델 수립	도시 기능 영역에 대해 점진적으로 데이터를 식별하여 데이터 요건을 정의하고 개념 모델을 설계하며, 데이터 일관성 및 효율적 관리를 위한 메타데이터 기반 조성

* 출처: 저자 작성

2) 시티데이터 리터러시 확보를 위한 평생 교육과 라이브(Live)화

스마트시티 시민에게는 본인 데이터의 주권자로서 자연스럽게 데이터 리터러시를 확보할 수 있는 환경이 제공되어야 한다. 즉, 유치원에서 언어를 글로 배우듯이, 데이터에 대한 리터러시를 자연스럽게 습득할 수 있는 평생 교육 환경이 제공되어야 하며, 또한 데이터에 대한 쉬운 접근과 이해를 위한 데이터 카탈로그가 명확히 정의되어 제공되어야 한다.

시티데이터 라이브화는 데이터 주체로서 삶 자체와 결합된 데이터 분석·활용을 통해 자신(데이터 주체로서의 자신)과 도시가 동일시될 수 있도록 하는 것이다. 이러한 라이브 시티데이터를 지원하는 방법 중 대표적 방법으로는 리빙랩이 있다. 일반적으로 연구실은 운영조직과 독립되어 운영되고, 연구결과 접근과 활용이 제한적이었다. 스마트시티에서의 리빙랩은 연구결과의 사용 주체가 연구 주체가 되어 연구와 운영이 끊김 없이 상호작용하는 혁신 모델을 의미한다. 리빙랩은 4차산업혁명과 데이터 경제에서 스마트시티의 혁신 모델로 정의되고 있으며, 시민들이 다양한 도시 운영 정책과 기술을 열린 공간에서 기술지원기업 및 지자체와 함께 연구하여 우리 삶에 적용시키는 방법이다. 국내에도 서울특별시의 '천만상상오아시스'[3]

3 서울시민이 제안한 아이디어를 정책으로 실현하는 온라인 시민 제안 사이트

와 같은 성공적인 사례가 나타나고 있다.

스마트시티의 지향점은 폭발적으로 증가하는 시티데이터를 수집-융합-분석하여 시민생활 향상을 위한 도시의 의사결정에 활용하여 창의적 서비스를 발굴·적용하는 것이다. 시티데이터의 수집-융합-분석-의사결정-적용의 과정에 과거처럼 시정 주도의 톱다운(Top-Down) 방식만이 아닌, 시민의 참여와 협력을 통한 시민주도의 보텀업(Bottom-Up)의 데이터 민주화를 통한 개방형 혁신 체계 접목이 필수적이다.

4. 마무리

오늘날 모든 산업은 데이터로 가득 차 있고, 성공하는 조직은 다가오는 상황에 적응하기 위해 데이터를 가장 빨리 이해하고 활용하는 조직이다. 빠른 발견과 더 깊은 통찰력을 가능하게 하는 가장 좋은 방법은 조직 전체에 데이터 과학 전문 지식과 권한을 분산시켜 시티즌 데이터 사이언티스트의 역할을 강화하는 데이터 활용 민주화이다.

또한 이러한 데이터 활용의 민주화는 기업에서뿐만 아니라 시민사회 전반으로 확산되고 있다. 마이데이터로 시작된 시민데이터 권리와 활용에 대한 민주화를 기반으로, 거대한 데이터를 생산해내고 이러한 데이터를 기반으로 움직이는 스마트시티는 사회 전반으로 데이터 민주화의 확산을 더욱더 촉진하고 있다.

제3장. 데이터 경쟁의 민주화

유영상
(한국전자통신연구원 책임연구원)

데이터는 데이터 경제에서 가장 중요한 생산요소로서 기업들은 이를 획득하고 활용하기 위해 전 세계적으로 치열한 경쟁을 벌이고 있다. 경쟁자보다 우월한 지위를 유지하기 위한 데이터 경쟁은 데이터 독점으로 이어지고 나아가 시장 지배적 지위를 이용하여 공정한 거래 행위나 시장 경쟁을 해치는 결과를 가져올 수 있다. 데이터를 기반으로 하는 디지털 경제에서 플랫폼 사업자의 시장 지배력이 데이터 경쟁에 미칠 수 있는 요인들을 살펴보고 그와 관련된 쟁점 사항을 소개한다.

1. 데이터 경제와 데이터 독점

우리나라 혁신성장의 가장 중요한 키워드는 '데이터 경제'라고 할 수 있다. 데이터 경제는 데이터와 경제가 합성되어 만들어진 용어로 데이터가 주도하는 경제를 말한다. 데이터 경제가 활성화되면 데이터의 활용을 통해 다른 산업의 발전을 촉진할 수 있고, 지금까지 없었던 새로운 제품과 서비스를 창출할 수 있으며, 나아가 경제성장과 일자리 창출의 동력을 제공할 수 있다.

데이터 경제에서 가장 중요한 생산요소인 데이터는 금융 자본과 마찬가지로 장기적인 가치를 가질 수 있고, 계속해서 가치를 창출할 수 있다는 점에서 새로운 자본 형태라는 주장도 점차 힘을 얻고 있다.[1] 하지만 풍부한 자원일 것이라는 기존의 통념과 달리, 데이터는 매우 희귀하고 때로는 고유한 정보의 조각으로 구성되어 있다. 따라서 모든 산업에서 기업들은 경쟁자들보다 먼저 고유한 데이터 자본을 확보하고 이를 사용하는 방법을 만들어내기 위해 치열한 경쟁을 벌이고 있다. 그로 인해 데이터를 어떻게 확보할 수 있는가, 또 데이터를 어떻게 활용하여 경쟁자보다 우월한 지위를 유지할 수 있는가 하는 문제가 공정경쟁과 관련되어 새로운 이슈로 떠오르고 있다.

인터넷을 기반으로 이루어지는 경제 활동인 디지털 경제에서 경제 주체 간 연결의 매개체인 플랫폼 사업자는 소비자와 관련 기업에 새로운 서비스, 새로운 가치를 제공하면서 그 대가로 데이터를 수집하고 활용하게 된다. 이 과정에서 정보나 데이터를 독점하는 것은 그 자체로 불법적인 일은 아니다. 다만 그러한 시장 지배적 지위를 이용하여 공정한 거래행위나 시장 경쟁을 해치는 결과를 가져온다면 문제가 될 수 있다. 많은 이용자 정보를 가지고 있는 지배적 플랫폼 사업자가 데이터를 독점함으로써 반경쟁적 행위[2]를 할 수도 있고, 이용자의 데이터를 악의적으로 사용할 가능성도 존재하기 때문이다. 특히 글로벌 정보기술 기업들은 그들이 가진 플랫폼을 기반으로 데이터를 독점하고, 이를 통해 시장 지배력을 강화하고 있다는 논란이 끊임없이 이어지고 있다.

2. 데이터 경제의 시장 구조

빅데이터를 활용할 수 있다는 것은 기업에 중요하고 긍정적인 이득을 가져다준다는 것을 의미한다. 개별 소비자 욕구를 더 잘 이해하고 공략함으로써 제품의 품질을 개선하고 완전히 새로운 서비스를 개발할 수 있게 한다. 그러나 데이터 기반의 디지털 시장은 네트워크 효과[3]가 강해 시장이 집중화·독점화되기 쉬운 특징이 있다. 네트워크 효과란 어떤 제품이나 서비스를 사용하는 사용자가 많을수록 그 가치가 더욱 높아지는 것을 말하며, 지배적 플랫폼 사업자들은 이러한 네트워크 효과를 이용한 전략을 구사한다. 즉 킬러 서비스 제공을 통해 소비자를 끌어모은 후, 그 서비스에 대한 제공 대가로 소비자들로부터 데이터를 입수하고 분석함으로써 더 많은 소비자를 끌어모을 수 있는 새로운 서비스를 개발하고, 다시 이를 통해 시

1 데이터 경제에서 말하는 데이터는 일반적으로 빅데이터를 의미한다. 일반 데이터와 빅데이터는 빅데이터가 가지고 있는 4가지 요소 즉, 데이터의 크기(volume), 수집된 정보의 다양성(variety), 데이터가 수집·사용·전파되는 속도(velocity), 데이터의 가치(value)라는 "4V"에 의해 구분될 수 있다(OECD, Big Data: Bringing Competition Policy to the Digital Era, 2016, p.5).

2 반경쟁적 행위는 일정한 거래 분야의 경쟁이 줄어들어 특정 사업자 또는 사업자 단체의 의사에 따라 어느 정도 자유로이 가격, 수량, 품질 기타 거래 조건 등의 결정에 영향을 미치거나 미칠 우려가 있는 상태를 초래하는 행위를 의미한다(출처: 산업통상자원부, 지식경제용어사전).

3 네트워크 효과에는 직접 효과와 간접 효과의 두 종류가 있는데, 직접적 네트워크 효과는 전화나 사회적 관계망 서비스(SNS)와 같이 어떤 제품이나 서비스를 이용하는 사람이 많으면 많을수록 해당 제품이나 서비스를 이용하는 다른 사용자에게 더 큰 가치를 주는 것을 말하고, 간접적 네트워크 효과는 OS와 소프트웨어, 비디오 게임기와 게임 소프트웨어와 같이 서로 보완적 특성이 있는 제품이나 서비스에서 특정 제품이나 서비스를 더 많이 이용할수록 제3자가 호환 가능한 기술을 더욱 개발하게 됨으로써 이용자에게 원래 제품이나 서비스의 인기를 강화하는 것을 말한다. 이 두 가지 효과가 모두 데이터 경제 시장에 영향을 미쳐 시장을 더욱 집중화·독점화되게 한다 .

장을 지배하게 된다.[4] 그 일례로 구글(Google)과 페이스북(Facebook)은 메일이나 검색, 친구 찾기와 위치 검색 등의 킬러 서비스를 제공하고 그를 통해 광범위한 데이터를 축적한다. 그리고 이를 분석하여 이용자의 소비 성향 등을 파악함으로써 더 나은 제품이나 서비스 개발에 활용한다. 플랫폼 사업자들의 이러한 네트워크 효과로 인하여 후발 기업들은 좀처럼 이를 만회할 기회가 없게 되고, 이에 따라 승자독식의 시장 구조가 형성되는 경향을 보이게 되는 것이다. 검색 엔진 분야에서의 구글, 모바일 운영체제에서의 애플(Apple), 사회 관계망 서비스의 페이스북, 온라인 쇼핑의 아마존(Amazon)이 그 대표적인 사례다. 데이터 경제의 시장 구조를 살펴보면 다음과 같은 특징을 가지고 있다.

가. 승자독식의 시장

디지털 시장이 가지고 있는 네트워크 효과, 전환 비용의 존재, 그리고 규모 및 범위의 경제와 같은 특성은 승자독식의 시장 경제로 만드는 경향이 있다. 그 결과로 많은 기술 시장에서는 한두 개의 대기업에 유리한 쏠림현상이 발생하게 되고, 그로 인해 진입장벽이 생기게 됨으로써 새로운 기업은 기존 기업을 극복하기가 어려워진다. 그러나 승자의 독점적 이익은 오히려 새로운 진입과 투자를 유도하는 역할을 하게 되며, 이때 시장의 잠재적 경쟁자는 상당한 혁신을 통해서만 기존 기업의 이점을 극복할 수 있게 된다.

나. 집중화된 시장

디지털 경제는 고도로 집중화되어 있어서 소셜 미디어, 일반 온라인 검색 및 온라인 광고와 같은 다수의 주요 온라인 시장은 단지 한두 개의 회사가 지배하고 있다. 이러한 집중 현상이 나타나는 것은 지배적 디지털 플랫폼이 신생 기업들을 대량으로 인수했기 때문으로 여겨진다.[5] 지난 10년 동안 지배적 플랫폼들은 수백 개의 회사를 인수했으며 그 가운데 다수는 실제적 또는 잠재적 경쟁자들을 제거하기 위한 것으로 보인다. 또한 혁신적인 스타트업에 자금 제공의 역할을 하는 벤처캐피털 산업이 오히려 스타트업이 기존 기업에 매각되도록 조장함으로써 시장 집중화에 기여하고 있다는 증거도 제시되고 있다.[6]

4 한국언론정보학회, 「글로벌 기업의 데이터 독점 관련 이슈와 대응방안 연구」, 2018.12.

5 U. S. House of Representatives Subcommittee on Antitrust, Commercial and Administrative Law of the Committee on the Judiciary, Investigation of Competition in Digital Markets, 2020.10.6.

6 Lemley, Mark A. & McCreary, Andrew, Exit Strategy. Stanford Law and Economics Olin Working Paper #542, 2020.1.30.

다. 온라인 플랫폼의 시장 지배력

인기 있는 플랫폼이 디지털 시장에서 게이트키퍼[7]의 지위를 획득하면 이들은 플랫폼 입점 기업들을 통제할 수 있는 큰 힘을 갖게 된다. 게이트키퍼란 기업들이 최종 이용자에게 다가가는 중요한 관문 역할을 하는 핵심적 플랫폼 서비스의 제공자를 말하는데, 이들의 어떤 행위나 결정은 플랫폼에 의존하는 많은 기업의 존망을 결정할 수도 있고, 시장 지배력 남용의 가능성도 무시할 수 없다. 지배적 플랫폼들은 대부분 인접한 비즈니스 라인도 통합하였기 때문에, 이들은 타 기업의 주요 중개자이면서 동시에 타 기업의 직접적인 경쟁자이다. 이러한 지배적 플랫폼의 이중적 역할은 자사 제품이나 서비스를 경쟁사보다 더 우대하거나 경쟁사 서비스를 직접 배제할 가능성이 있어 경쟁이 제한될 우려가 있다.

현재 플랫폼 사업자들의 영향력이 세계 시장으로까지 확대되고 있어서 각국에서는 이에 대한 대응책을 모색하는 중이다. 세계 각국의 경쟁 당국은 이들 플랫폼 기업들의 시장 지배력과 반경쟁적 행위를 지속해서 조사하고 있으며, 이들의 시장 지배력이 미치는 영향에 대해서도 검토하고 있다.

3. 데이터 경쟁의 제한요인

경쟁법은 시장에서 경쟁을 확보하기 위한 법률로 우리나라의 경우 '독점규제 및 공정거래에 관한 법률(이하 공정거래법)'을 통해 규율하고 있으며, 미국은 연방 차원에서 '반독점법(Antitrust Law)'으로 총칭하는 '셔먼 반독점법(Sherman Antitrust Act of 1890)', '클레이튼법(Clayton Antitrust Act of 1914)', 'FTC법(Federal Trade Commission Act)' 3가지를 그 근간으로 하고 있다. 일본은 '사적독점의 금지 및 공정거래의 확보에 관한 법률(私的独占の禁止及び公正取引の確保に関する法律)'로, 유럽 각국은 '경쟁법(Competition Law)'을 통해 경쟁을 확보하고 있다.

2016년 5월에 프랑스의 경쟁 당국인 경쟁청(Autorité de la concurrence)과 독일의 경쟁 당국인 연방카르텔청(Bundeskartellamt)은 공동으로 '경쟁법과 데이터'라는 보고서를 출간

7 2020년 12월 유럽연합 집행위원회(European Commission)의 디지털 시장법 초안에서는 게이트키퍼로서의 자격 판단 기준으로 ①내부 시장에서 중대한 영향력이 있을 것, ②중요한 관문 역할을 하는 핵심 플랫폼 서비스를 운영하고 있을 것, ③운영 면에서 확고하고 지속성이 있는 지위를 누리고 있거나 가까운 장래에 그러한 지위를 누리게 될 것으로 예측 가능할 것을 기준으로 제시하고 있다.

하였다.[8] 그 보고서에서 두 기관은 빅데이터와 관련되어 경쟁이 제한될 가능성, 즉 시장 지배력과 경쟁법 간의 상호작용을 평가할 때 고려해야 할 몇 가지 주요 문제들을 검토하였다. 이들이 데이터 경쟁을 제한할 것이라 우려하는 요소들을 살펴보면 다음과 같다.

가. 진입장벽

경쟁 제한요인의 하나인 진입장벽은 데이터 경제에서 다음과 같이 네 가지 이유에서 발생할 수 있다. 첫 번째는 데이터 자체가 곧 시장 지배력과 연결된다는 점이다. 시장에서 대규모의 데이터 또는 다양한 데이터에 대한 접근은 차별화된 서비스 제공 등 시장 경쟁력을 확보하는 데 필수적이다. 즉 데이터가 시장 지배력의 원천이 된다. 따라서 데이터 독점은 곧 독점기업에게는 시장 지배력의 강화를, 신규 진입자에게는 시장 진입장벽을 초래할 수 있다.

두 번째는 네트워크 효과이다. 온라인 플랫폼은 서로 다른 시장 부문을 연결하기 때문에 강력한 네트워크 효과를 나타낸다. 예를 들어 아마존과 같은 온라인 상거래 플랫폼은 구매자와 판매자를 연결하는데 판매자와 구매자 모두 플랫폼에 참여하는 사용자 수가 증가함에 따라 아마존 시장의 가치가 증가하게 된다. 이러한 네트워크 효과는 진입장벽의 요소로서 경쟁을 제한한다고 여겨지지만, 새로운 시장 참여자들이 혁신적 특성과 같은 이유로 많은 수의 사용자를 끌어들일 수 있다면 오히려 네트워크 효과가 소비자 기반의 급속한 성장 잠재력을 줌으로써 경쟁을 자극하기 때문에 새로운 시장 참여자들에게 이로울 수 있다.

세 번째는 전환 비용의 존재이다. 전환 비용은 사용자가 현재 사용하고 있는 한 기술이나 제품, 서비스에서 다른 기술, 제품, 서비스로 전환할 때 발생하는 비용으로, 이 역시 잠재적 시장 진입자들에게는 또 다른 장벽이다. 사용자가 기존 기술에서 벗어나기가 쉽지 않기 때문에 대형 기술기업이 대부분 시장 지배력을 유지할 수 있게 된다. 사용자가 선호하는 제품이나 서비스를 제공하는 회사로 전환하는 대신 기존 회사에 머물러야 할 만큼 전환 비용이 매우 높을 때 시장에는 락인효과(lock-in)가 나타나게 된다. 시간이 지남에 따라, 락인효과는 경쟁을 낮추고, 시장 진입을 방해하며, 심지어 개인정보보호를 악화시킬 수도 있다.

네 번째는 규모의 경제나 범위의 경제로부터 진입장벽이 발생할 수 있다. 규모의 경제는 생산량 규모가 증가함에 따라 단위당 비용이 감소하는 것으로, 집중과 독점으로 기울기 쉬운 기술 시장의 또 다른 특징이다. 규모가 커지는 시장에서는 매출이 증가하면 평균 단가가 줄어든다. 이러한 시장에 진입하려면 상당한 선행 비용이 필요하므로 시장은 이미 규모가 큰 기업을 선호하게 되고 따라서 신규 기업이 시장에 진출해 대규모 기존 기업에 도전하기는 매우 어렵다. 한편, 범위의 경제는 한 기업이 두 개 이상의 재화나 서비스를 함께 생산할 때 발생하는 비용이 별도의 기업이 각각 생산할 때의 총비용 합계보다 작은 것을 의미한다. 범위의 경제를 누리는 지배적인 기업은 상대적으로 낮은 비용으로 경쟁할 수 있으며, 자사 제품의 광범위한 생태계를 통해 인접 시장 전반으로 그 영역을 쉽게 확장할 수 있다.

나. 인수와 합병

더 나은 데이터 액세스를 얻기 위한 첫 번째 기업 전략은 대규모 데이터셋를 소유한 다른 회사를 인수하거나 이들과 합병하는 것이다. 일반적으로 기존 기업과 혁신적인 신규 기업 간의 합병은 신규 기업의 낮은 시장점유율로 인해, 또는 수평적 중복이 없어서 기존 시장 구조에 미치는 영향이 적다. 그러나 데이터 관련 시장에서는 그 합병 대상인 신규 진입자가 다른 시장의 대규모 데이터베이스에 대한 접근을 보유하고 있다면 데이터에 대한 차별적 액세스를 초래할 수 있고, 이 시장과 관련된 데이터의 집중도를 증가시킬 수 있다. 특히 만약 서로 다른 데이터 조합을 통해 만들어진 정보를 경쟁업체가 복제하는 것이 불가능하다면 독점 우려를 초래할 수도 있다. 또한, 별도의 업스트림 또는 다운스트림 시장 데이터[9]에서 이미 강력한 시장 지위를 보유하고 있는 두 회사의 합병은 새로운 경쟁자를 이러한 시장에서 아예 배제할 가능성도 있다.

다. 배제적 행위

경쟁업체가 데이터에 접근하는 권한을 박탈하는 행위는 경쟁을 약화할 뿐만 아니라 어떤 상황에서는 경쟁사업자를 시장에서 배제할 수도 있다. 이렇게 경쟁사업자를 배제할 수 있는 행위에는 필수 설비에 해당하는 데이터에 대한 접근 거부, 데이터에 대한 차별적 접근 허용, 배타적 계약, 데이터와의 결합 판매 및 데이터의 교차 사용 등이 있다.

먼저 필수 설비에 해당하는 데이터에 대한 접근 거부는 해당 데이터가 접근을 요청하는 기업의 사업 활동에 '필수 설비'인 경우 반경쟁적일 수 있다. 그러나 원칙적으로 지배적인 회사라 하더라도 경쟁업체의 사업을 장려할 의무는 없으므로 유럽사법재판소에서는 필수 설비에 대한 의무적 접근 허용은 몇 가지 사례로 한정하여 인정하고 있다. 즉, 기존 사업자의 접근 거부가 해당 사업의 수행에 불가결한 제품과 관련이 있는 경우, 그 거부가 잠재적 소비자 수요가 있는 신제품의 출현을 방해하는 경우, 그리고 객관적으로 정당화되지 않고 2차 시

8 Autorité de la concurrence & Bundeskartellamt, Competition Law and Data, 2016.5.10.

9 공급망에서 업스트림(upstream)은 데이터의 생산, 수집이 이루어지는 활동 분야를 말하며, 다운스트림(downstream)은 데이터의 가공, 활용, 판매가 이루어지는 활동 분야를 말한다.

장의 모든 경쟁을 배제할 가능성이 있는 경우에는 설비나 네트워크에의 접근을 요청할 수 있다. 그러나 필수설비 이론을 적용한다는 것은 시장 지배적 사업자의 설비를 경쟁사업자에게 강제로 제공하도록 만드는 것인데, 데이터 제공과 관련하여 이 필수설비 이론을 적용하는 것은 사실상 어려울 것으로 보인다. 우선, 지배적 사업자가 보유한 데이터가 관련 시장에 참여하기 위해 필수적인가 하는 문제가 있다. 그를 대체할 수 있는 데이터가 존재할 수도 있으며, 데이터를 획득하는 새롭고 다양한 방법이 존재할 수도 있기 때문이다. 또한, 데이터는 비경합적인 특성이 있으므로 이를 복제할 가능성도 있다. 그리고 설령 경쟁사업자에게 데이터를 제공한다 하더라도 정보 주체의 동의 없이 제3자에게 개인정보를 제공하는 것은 어렵다는 문제가 있다.

데이터 접근을 차별적으로 거부하는 때도 반경쟁적인 것으로 간주할 수 있다. 일반적으로 수직적 통합은 경쟁을 왜곡하는 효과와 함께 전략적 정보에 대한 차별적 접근을 수반할 수 있다. 예를 들어, 플랫폼 사업자가 온라인 플랫폼에서 온라인 소매업자로도 활동하는 경우, 해당 시장에서 판매하는 경쟁업체와 소비자의 행동에 대한 정보에 쉽게 접근할 수 있다. 반면, 경쟁업체에는 정보를 제한함으로써 경쟁자들보다 더 우위에 있게 된다.

반경쟁적인 데이터 기반 전략에는 제3자 제공업체와의 배타적 조항을 통해 경쟁자가 데이터에 액세스하는 것을 방지하거나, 소비자가 기술이나 플랫폼을 채택하는 것을 어렵게 함으로써 경쟁자가 유사한 데이터를 조달할 기회를 차단하는 것도 포함될 수 있다. 특히 지배적인 회사들에 의해 체결된 배타적 계약은 경쟁상대를 배제하는 수단이 될 수 있다.

기업이 특정 시장에서 수집한 데이터를 다른 시장에서 반경쟁적인 방식으로 시장 지배력을 증가시키기 위해 사용할 수도 있다. 예를 들어, 가치 있는 데이터를 소유한 회사가 그 데이터를 자신이 제공하는 데이터 분석 서비스와 묶어서 판매하게 되면, 데이터 분석 시장에서는 경쟁사보다 데이터를 소유한 회사가 더 유리하게 됨으로써 경쟁을 감소시킬 수도 있다.

라. 가격 차별 수단으로서의 데이터

데이터는 가격 차별을 용이하게 한다. 기업은 고객에 대한 데이터를 수집함으로써 고객의 구매 습관에 대한 더 자세한 정보를 얻고, 제품이나 서비스에 대한 고객의 지불 의사를 평가할 수 있다. 만약 시장 지배력이 있다면, 기업은 수집한 데이터를 이용하여 식별한 여러 고객 그룹별로 시장가격을 다르게 정할 수도 있을 것이다. 이러한 데이터에 기반한 가격 차별의 결과로 어떤 소비자들은 주어진 제품이나 서비스에 대해 더 높은 가격을 지불해야 할 수 있지만, 또 어떤 소비자들은 차별이 없을 때보다 더 나은 가격을 제안받을 수 있으므로 가격 차별의 효과는 모호할 수 있다. 따라서 데이터로 인해 더 큰 가격 차별이 일어난다고 가정하더라도 그 영향을 평가하기는 쉬운 일이 아니다. 가격 차별이 경쟁법 적용 범위에 포함되는지 여부는 그 가격 차별이 시장 지배적 지위의 남용이나 수직적 거래제한[10]에 해당하는지에 따라 판별할 필요가 있다.

마. 개인정보보호 경쟁의 제한

개인정보보호에 대한 우려는 그 자체로 경쟁 당국의 개입 범위에 포함되지 않는다. 유럽사법재판소는 개인데이터의 민감성과 관련된 모든 문제가 경쟁법에 관한 문제가 아니라 데이터 보호에 관한 관련 규정에 근거하여 해결될 수 있다고 밝힌 바 있다.[11] 그러나 개인정보보호 문제를 경쟁법의 대상에서 제외할 수 없는 이유는 개인데이터 수집과 사용에 관한 기업의 결정은 경제적 및 경쟁적 차원에 모두 영향을 미칠 수 있기 때문이다. 특히 지배적 사업자의 제품이나 서비스에서 데이터가 주요한 생산요소로서 역할을 할 때 지배적 사업자의 정책이 경쟁에 영향을 미칠 수 있다. 그러한 경우, 기업의 지배력, 데이터 수집 과정 및 관련 시장에 대한 경쟁 사이에 밀접한 관계가 있을 수 있으며, 이는 경쟁 절차에서 개인정보보호 정책과 규제를 고려할 정당성을 가질 수 있게 된다. 이 때문에 경쟁 관점에서 개인정보보호 정책을 고려할 수 있다.

이상과 같은 데이터와 관련된 경쟁 제한 우려 사항을 요약하면 다음의 표와 같다.

[표 1-3-1] 데이터 관련 경쟁 제한 우려 사항

구분	내용
진입장벽	진입장벽은 다음과 같은 요인에 의해 발생 가능 · 시장 지배력 원천으로서의 데이터 · 네트워크 효과 · 전환 비용 · 규모 및 범위의 경제
인수와 합병	다음과 같은 경우의 인수와 합병은 경쟁 제한 가능성이 존재 · 합병 대상 신규사업자가 다른 시장의 대규모 데이터베이스 보유 · 데이터 조합을 통해 추출된 정보를 경쟁사가 복제 불가능 · 별도의 수직 시장에서 이미 강력한 시장 지위를 보유하고 있는 두 회사의 합병

10 수직적 거래제한은 생산이나 유통 등 각기 서로 다른 수준에서 영업하는 사업자 간에 제품이나 서비스의 판매·구입에 관한 제한 행위를 말한다.

11 ECJ, Asnef-Equifax and Administration Del Estado, C-238/05, Judgment of the Court(Third Chamber), 2006.11.23.

구분	내용
배제적 행위	경쟁자의 데이터 접근을 박탈함으로써 경쟁자를 배제할 수 있는 행위 · 필수 설비에 해당하는 데이터에 대한 접근 거부 · 차별적 접근 허용 · 배타적 계약 · 데이터와의 결합 판매 및 교차 사용
가격 차별 행위	고객데이터 수집을 통한 가격 차별이 다음에 해당하는 경우 · 시장 지배적 지위의 남용 · 수직적 거래제한
개인정보보호 경쟁의 제한	지배적 사업자가 데이터 기반의 시장 지배력을 이용하여 개인정보보호에 관한 경쟁을 제한하는 경우

* 출처: Autorité de la concurrence & Bundeskartellamt(2016), 저자 재정리

4. 마무리

데이터 경쟁을 제한할 수 있는 요인 가운데에는 네트워크 효과나 가격 차별과 같이 소비자에게 미치는 효과가 불분명한 것들도 있고, 필수설비 이론과 같이 데이터와 관련하여 경쟁법 적용이 어려운 예도 있다. 또한, 개인정보보호 문제는 그 자체로는 개인정보보호법으로 다루어야 한다는 것이 주류적인 흐름이지만, 비가격경쟁의 하나인 개인정보보호 경쟁을 제한하는 것은 경쟁법의 집행 대상이 될 수 있다는 시각도 있다.

우리나라의 공정거래법은 제1조에서 동 법의 목적을 '이 법은 사업자의 시장 지배적 지위의 남용과 과도한 경제력의 집중을 방지하고, 부당한 공동행위 및 불공정거래행위를 규제하여 공정하고 자유로운 경쟁을 촉진함으로써 창의적인 기업활동을 조장하고 소비자를 보호함과 아울러 국민경제의 균형 있는 발전을 도모함을 목적으로 한다.'라고 밝히고 있다. '공정하고 자유로운 경쟁'은 사업자들 간의 경쟁이 그들의 제품이나 서비스의 가격, 또는 품질과 같은 사업의 장점을 중심으로 이루어지는 능률 경쟁을 의미하며, 또한 시장의 진입과 퇴출이 자유롭고, 경쟁자 간에 경쟁이 제한되지 않는 것을 의미한다.[12] 경쟁은 경제 활동의 핵심 엔진이며 특히 디지털 시장에서의 경쟁은 기존 사업자에게는 혁신을 지속하도록 만들고, 신규 진입자들에게는 새로운 기술 개발과 자본 투자를 촉진하게 하는 강력한 동기가 된다. 그러므로, 공정하고 자유로운 경쟁이 이루어질 수 있도록 경쟁을 저해하는 요인을 개선해 나아가는 것이 무엇보다 중요하다.

12 손영화·차성민, 「경쟁법의 원리와 공정거래위원회의 역할」, 한양법학회, 한양법학 제19집, 2006.8.

PART



데이터산업 관련 주요 정책·법제도

제1장.
국내 데이터산업 관련 정책·법제도

제2장.
해외 데이터산업 관련 정책
해외 데이터산업 관련 법제도

제3장.
디지털 무역과 국경 간 데이터 이동

제1장. 국내 데이터산업 관련 정책·법제도

박소영
(국회입법조사처 입법조사관)

디지털 시대를 맞아 국내에서도 '데이터'를 공급하고 유통하고자 하는 정책 집행과 입법적 노력이 이루어지고 있다. 2020년 7월 시작된 디지털 뉴딜 사업은 데이터 댐과 데이터 바우처 사업 등을 통해 데이터 공급을 확대하였으며, 2021년에는 공급된 데이터를 통한 성과를 확산하기 위해 마이데이터 산업 활성화, 초연결 신산업 육성, 클라우드 강화, 디지털 복제물 활용 확대 사업 등을 추가적으로 진행한다. 또한 데이터 유통 환경 기반을 마련하기 위한 데이터 3법 및 데이터 관련 법률 개정이 진행되고 있다.

1. 국내 데이터 정책 현황

우리나라는 디지털 뉴딜 사업을 중심으로 데이터 정책을 추진하고 있다. 디지털 뉴딜 사업은 데이터, 인공지능, 메타버스 등 디지털 사업에 대규모 예산을 투자하여 기술 및 산업 기반을 마련하고 인력을 양성하며, 그에 알맞은 법제도, 규제를 마련하여, 우리나라의 디지털 대전환을 달성하고자 한다. 디지털 뉴딜 1.0이 데이터 댐 등 디지털 인프라를 구축하는 데 중점을 두었다면, 1년 뒤에 발표된 디지털 뉴딜 2.0은 디지털 인프라 구축을 지속하되 이를 활용하는 방안에 집중하고 있다. 한편 디지털 뉴딜이 본격적으로 추진되면서 데이터의 중요성이 무엇보다 높아짐에 따라, 정부는 4차산업혁명위원회에 데이터 특별위원회를 신설하고 대한민국 데이터 119 프로젝트를 국가 데이터 정책으로 추진하고 있다.

가. 디지털 뉴딜 1.0과 2.0

1) 디지털 뉴딜 1.0

정부는 2020년 7월 디지털 뉴딜 정책을 발표하고, 범정부 차원에서 2025년까지 총 58.2조 원(국비 44.8조 원)을 투자하여 디지털 경제로의 전환을 선도하기 위한 계획을 발표하였다. 데이터 관련 주요 성과는 다음과 같다.

① 데이터 댐 구축 및 활용

'데이터'가 디지털 뉴딜의 핵심인 만큼 디지털 뉴딜 1.0은 데이터를 구축하고 개방하는 부분에 집중하였다. 우선, 정부는 공공데이터를 개방하였다. 지금까지 10만 5,000개의 공공데이터가 개방되었고, 2021년에도 14만 2,000개의 공공데이터(공공 정밀 도로지도, 안전·취약 시설물 관리 데이터 등)가 개방된다. 정밀 도로지도, 스마트 전력 거래, 의료기기 안전정보, 산림 미세먼지 정보 등 46개 분야 국가 중점 데이터도 개방 범위를 확대하고 있다.

또한 정부는 시간·비용 등의 이유로 개인이나 기업이 개별적으로 구축하기 어려운 양질의 인공지능 학습용 데이터를 크라우드 소싱을 통해 대규모로 구축하였다. 2020년에는 170종을, 2021년에는 150종을 구축하여, 지금까지 총 인공지능 학습용 데이터 4.8억 건을 구축하였다. 현재는 산업정보 연계 주요국 특허 영-한 데이터, 한국인 전신 형상 및 치수 측정 데이터, 반려동물 안구 및 피부질환 데이터, 자동차 부품 품질 검사 영상 데이터, 영유아 행동 영상 데이터 등이 구축되고 있다.

데이터 댐 구축과 동시에 데이터 활용을 촉진하기 위해 정부는 중소·벤처기업을 대상으로 데이터 바우처 사업을 시행하고 있다. 데이터를 활용하여 비즈니스를 혁신하거나 신제품·신서비스를 개발하고 싶은 중소·벤처기업에 데이터 바우처를 발급하고, 해당 기업이 바우처를 활용하여 원하는 데이터 상품 또는 가공서비스 비용을 지불하는 방식이다. 2020년에는 2,040건, 2021년에는 2,580건의 데이터 바우처를 지원하였으며, 2025년까지 계속 지원할 예정이다. 데이터 바우처 사업으로 인해 기존 SW·SI 기업이 데이터 댐 사업에 본격적으로 참여하여 데이터 공급 기업 수가 2019년 393개에서 2021년 7월 기준 1,126개로 2.9배 증가하였다.

② 인공지능 산업 육성

정부는 다량의 데이터를 활용하는 인공지능 산업의 성장을 지원하였다. 데이터 바우처와 유사하게 인공지능 적용이 필요한 수요기업에 인공지능 바우처를 발급하고, 수요기업이 바우처를 활용하여 원하는 인공지능 솔루션을 이용하도록 하였다. 2020년 225건, 2021년 200건을 지원함으로써 수요기업의 디지털 전환뿐만 아니라, 인공지능 공급 기업 수가 2019년 220개에서 2021년 991개로(2021. 7. 기준), 4.5배 증가하는 등 인공지능 기반 성장이 확대되는 성과도 거두었다. 정부는 2025년까지 인공지능 바우처 사업을 지속할 예정이다.

③ 비대면 산업 육성

코로나19 상황이 장기화되면서 비대면·클라우드 바우처 사업을 통해 약 8.4만 중소기업의 비대면 역량을 강화하고, 비대면 비즈니스에 디지털 기술을 접목한 서비스 분야를 26개 발굴하여 관련 스타트업을 지원하였다. 또한 소상공인 점포 4,025개를 스마트 상점으로 전환하고, 모바일 헬스케어 및 인공지능·IoT 기반 어르신 건강관리 서비스 등을 통해 비대면으로 3.5만 명 국민의 건강관리를 도움으로써 일상생활에서의 비대면 기반을 마련하였다.

④ 사회기반시설(SOC) 디지털화

도로, 철도, 공항과 같은 사회기반시설에 인공지능 기술과 디지털 관리 체계를 도입하여 안전성과 효율성을 제고하고, 전국 3D 지도 구축 등을 통해 디지털 복제물(디지털 트윈) 혁신을 지원하고 있다. 2020년에는 철도통합무선망을 573km 구간에 구축하고, 광역철도 6,538개소, 경부고속 2,367개소에 전기설비 IoT를 설치하였다. 또한 경상·강원권역 3D 지형지도(2,020km^2, 국토 전체의 2%), 27개 시 지하공간 통합지도, 6,700km 구간의 정밀도로지도를 구축하였다. 2021년에도 △철도통합무선망 501km 구간 구축, △경부고속 3,190개소, 일반철도 6,954개소에 전기설비 IoT 설치, △도시지역 3D 지형지도 구축(누적 47,150km^2)을 진행하고 있다.

2) 디지털 뉴딜 2.0

정부는 2021년 7월 22일 디지털 뉴딜 1.0 추진 성과를 진단하고 이를 바탕으로 추진할 디지털 뉴딜 2.0을 발표하였다. 디지털 뉴딜 2.0 발표에서는 2025년까지 49조 원을 투자하겠다고 밝혔다. 이는 디지털 뉴딜 1.0의 44.8조 원보다 증가한 투자 금액이다.

디지털 뉴딜 2.0은 △디지털 뉴딜 1.0에서 구축한 디지털 인프라를 바탕으로 민간 중심의 활용을 강화하고 △중점 프로젝트의 성과를 전 지역·전 산업으로 확산하면서 △향후 세계 시장을 선도할 미래 신산업을 육성하는 세 가지 방향으로 추진한다. 이와 같은 방향에서 디지털 뉴딜 2.0 데이터 정책은 데이터 댐 구축, 바우처 사업 등 기존 정책을 유지하면서 마이데이터 사업 등을 통해 그동안 구축한 데이터 댐의 활용을 강화하고, 메타버스, 클라우드 등 디지털 신산업 지원을 강화하는 내용을 추가하였다.

① 데이터 댐 전주기 활용 강화

정부는 2025년까지 인공지능 학습용 데이터 1,300종, 분야별 빅데이터 플랫폼 31개 등 데이터 댐 구축을 계속 추진하면서, 가명정보 결합, 분야별 활용 경진대회 등을 통해 민간의 데이터 활용을 촉진하고자 한다.

정부는 국민이 일상생활에서 데이터 구축·개방·활용의 편익을 직접 체감할 수 있도록 산업 전 영역에서 마이데이터 생태계 활성화에 집중한다. 금융, 의료, 교통, 공공 영역 등에서 마이데이터 실증사업을 실시하여 지금까지 구축한 데이터를 활용하는 사례를 만들고 있다. 또한 '개인정보보호법'을 개정하여 마이데이터의 법적 기반을 일원화하여 마련하고, '가명정보활용지원센터'를 구축하여 기업 및 기관들이 안전하게 데이터를 처리할 수 있도록 지원할 예정이다.

② 메타버스 등 초연결 신산업 육성

메타버스(metaverse)라고 불리는 확장 가상세계에 대한 지원을 디지털 뉴딜 2.0의 새로운 과제로 추가하였다. 정부는 다양한 기업들이 데이터를 축적, 활용하고 새로운 콘텐츠와 서비스를 개발하는 개방형 확장 가상세계 플랫폼을 구축하고, 이를 기반으로 메타버스 콘텐츠 제작, 핵심 기술 개발 등을 종합 지원할 계획이다. 이를 위해 과학기술정보통신부는 2021년 5월 민간이 주도하고 정부가 지원하는 체계의 '확장 가상세계 얼라이언스'를 출범하였다. 공급·수요기업, 이동통신사, 미디어 업계 등 181개 기업·기관이 함께 참여한 확장 가상세계 얼라이언스를 중심으로 기업 간 공동 과제를 발굴·추진하고 콘텐츠 제작·실증, 개발자 교육 등을 지원할 예정이다.

[그림 2-1-1] 디지털 뉴딜 2.0에서 신설된 초연결 신산업 육성 과제



* 출처: 대한민국 정책브리핑, 「한국판 뉴딜 종합계획 발표자료(부총리 겸 기획재정부장관)」, 2020.7.14.

③ 클라우드 강화

정부는 인공지능과 데이터 경제를 가속화하는 내용에 초점을 두어 클라우드 정책을 추진할 계획이다. 클라우드는 업무 효율성 증진과 비용 절감을 위한 단순 인프라로서의 기능을 넘어 데이터, 인공지능 등 타 기술과 융합한 XaaS(Everything as a Service)로 진화하고 있

다. 그에 따라 국내 클라우드 산업 경쟁력을 강화하고 산업 생태계를 조성하는 방향으로 '제3차 클라우드 컴퓨팅 발전 기본계획'을 수립하여 추진할 예정이다. 우선 공공부문의 디지털 경쟁력을 높이고 국내 클라우드 산업 발전의 마중물이 될 수 있도록 2025년까지 공공부문을 클라우드로 우선 전환하고 경쟁력 있는 소프트웨어 서비스의 클라우드 전환을 지원한다.

④ 디지털 복제물 활용 확대

디지털 복제물(디지털 트윈)은, 현실의 실제 사물을 쌍둥이같이 가상세계에 동일하게 구현하고 이를 실시간 제어, 사고 예방 등에 활용하는 기술이다. 디지털 뉴딜 2.0은 디지털 복제물 활용을 확대한다. 제조현장 안전에 디지털 복제물을 적용하고, 분야별 복제물의 상호 연계를 위한 핵심 기술 개발도 추진한다. 이 과정에서 디지털 뉴딜 1.0에서 구축한 3D 정밀지도, 지하공간 통합 지도 등을 활용한다. 아울러 정부는 올해 하반기 중 디지털 복제물 활용 기반 구축과 산업생태계 조성 등을 포함한 '디지털 복제물 산업 활성화 전략(가칭)'을 발표할 계획이다. 현재 과학기술정보통신부는 지역 소재의 시장, 병원 등 다양한 공공시설물 및 중소·중견기업의 제조시설물을 대상으로 디지털 트윈 기술을 적용하는 '디지털 트윈 공공선도' 사업을 진행하고 있다.

[그림 2-1-2] 디지털 뉴딜 1.0과 2.0의 과제 변화

'20.7월, 「디지털 뉴딜」	⇒	'21.7월, 「디지털 뉴딜 2.0」
① D.N.A. 생태계 강화		① D.N.A. 생태계 강화
② 교육 인프라 디지털 전환		② **비대면 인프라 고도화 (통합)**
③ 비대면 산업 육성		③ **초연결 신산업 육성 (신설)**
④ SOC 디지털화		④ SOC 디지털화

* 출처: 관계부처 합동, 「한국판 뉴딜 2.0-미래를 만드는 나라 대한민국-」, 2021.07.14.

나. 대한민국 데이터 119 프로젝트

디지털 뉴딜이 본격적으로 추진되면서 범부처적인 데이터 정책 컨트롤타워의 필요성이 부각되었다. 이에 정부는 2021년 2월 대통령직속 4차산업혁명위원회에 데이터 특별위원회를 신설하고, '국가 데이터 정책 추진방향'으로 '대한민국 데이터 119 프로젝트'를 발표하였다. '대한민국 데이터 119 프로젝트'는 민간이 그동안 요구하였던 11대 실천과제와 국민이 데이터 활용 성과를 체감할 수 있는 9대 서비스로 구성되며, 주요 내용은 다음과 같다.

1) 미개방 핵심 데이터 제공

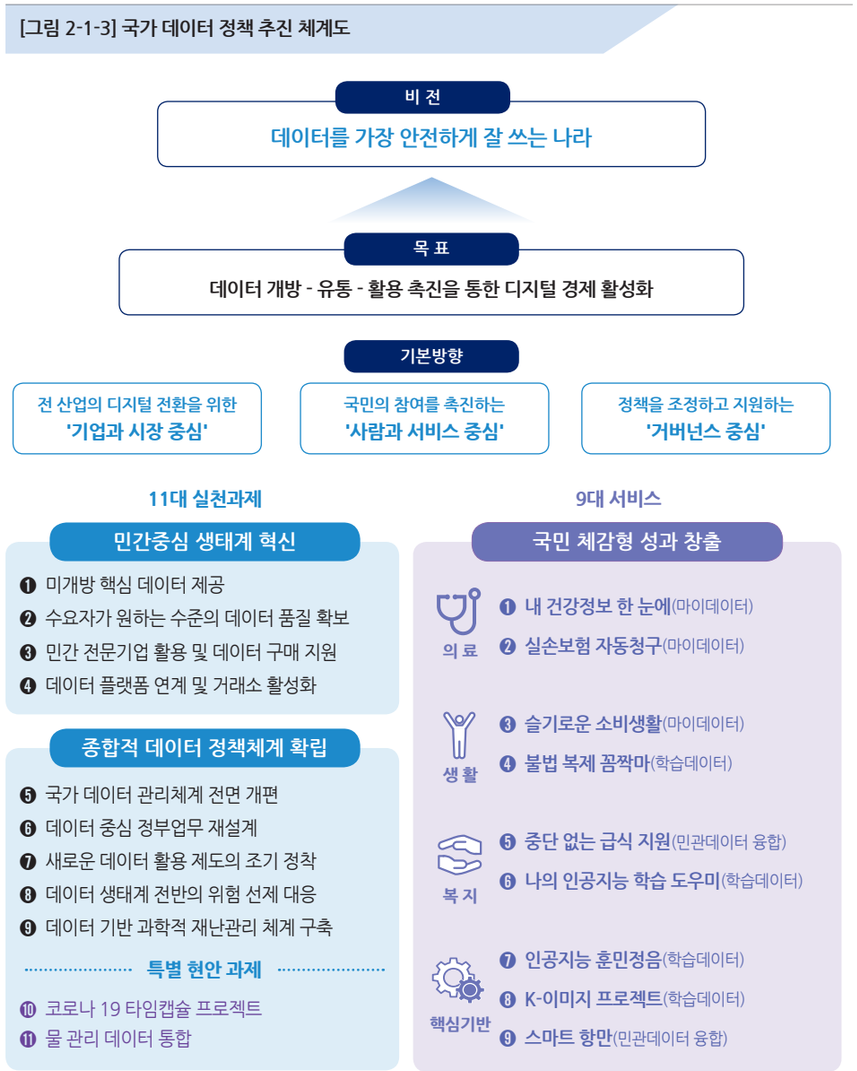
정부는 민간 수요가 높은 핵심 데이터를 개방한다. 우선 국세청의 사업자등록 및 휴·폐업 데이터, 건강보험공단의 보험데이터, 건강보험심사평가원의 진료데이터, 교육부·교육청의 학생 교육데이터를 올해 개방하며, 수요조사를 통해 2차 개방 대상을 선정할 계획이다. 상기 데이터는 민간 수요에 맞게 즉시 활용 가능한 다양한 방식으로 제공된다.

또한 공공기관이 보유하는 비정형 데이터(의료영상, 해저지형지도, 교통 CCTV, 문화 및 관광 사진 등)도 공개된다. 공공기관이 보유하고 있는 비정형 데이터 실태를 조사한 후, 메타데이터 중심의 표준화 작업과 국제표준 기반의 오픈 포맷 변환 작업을 거쳐 순차적으로 개방한다.

2) 데이터 품질 확보와 데이터 플랫폼 연계

정부는 공공데이터 품질을 제고한다. 공공데이터 제공 형태, 표준 등에 대해 지속적으로 제기되는 문제점을 개선하고, 앞으로는 공공의 정보시스템 구축 계획 단계에서부터 공공데이터 품질을 관리하도록 한다. 또한 고품질의 인공지능 학습용 데이터를 민간에게 제공하고, 공공·민간의 각종 데이터 플랫폼을 통합 데이터 지도로 연계하여 필요한 데이터를 손쉽게 찾아 활용할 수 있는 기반을 마련한다.

[그림 2-1-3] 국가 데이터 정책 추진 체계도



* 출처: 4차산업혁명위원회, 「국가 데이터 정책 추진방향」, 2021.2.17.

3) 민간 전문기업 육성과 공공의 데이터 구매

공공이 생산·개방한 데이터를 민간에서 활용하도록 데이터 가공을 전담하는 전문 기업을 육성하고, 민간이 판매하는 데이터를 공공에서 구매하여 민간데이터 시장을 활성화한다.

4) 마이데이터 사업 조기 정착

데이터 활용 사업으로 가장 주목받고 있는 마이데이터 사업을 확대하고 관련 제도 환경을 조성한다. 산업 분야별로 마이데이터 선정 기준과 데이터 수집 방법을 구체화하고 개인정보와 데이터 관련 법제도를 정비한다. 이와 더불어 개인 건강기록을 한 번에 확인할 수 있는 '내 건강정보 한눈에', 별도 서류 절차 없이 보험금을 청구하는 '실손보험 자동청구', 개인 소비활동을 분석하는 '슬기로운 소비생활' 등의 서비스를 구현하여 국민이 마이데이터 서비스로 인한 편의를 체감하도록 한다.

5) 국가 데이터 관리체계 전면 개편

정부는 국가 전체적인 데이터 현황을 파악하고, 데이터 중요도 등급에 따른 분류 기준을 산업 분야별로 마련하여 데이터 보호·활용 정책을 수립한다. 이를 위해 정부 부처, 지방자치단체 및 공공기관에 데이터 활성화 전략과 구현 전반을 담당하는 데이터기반행정 책임관(Chief Data Officer) 제도를 도입하였다. 데이터기반행정 책임관 협의체인 '데이터기반행정 책임관 협의회'는 2021년 3월 공식 출범하여 분기별로 데이터기반 행정 현안과 기관 간 협업 방안을 논의한다.

2. 국내 데이터 법제도 현황

국내 데이터 관련 법제도는 데이터 처리 과정에서 발생하는 타 권리(개인정보 등)와의 조화를 다루는 법률, 공공에서 생산·수집된 데이터를 공공·민간에서 활용하도록 돕는 법률, 민간데이터 시장 활성화를 위한 법률로 크게 분류해 볼 수 있다.

가. 개인정보·신용정보와의 조화

데이터 3법은 '개인정보보호법', '정보통신망 이용촉진 및 정보보호 등에 관한 법률(이하 정보통신망법)', '신용정보의 이용 및 보호에 관한 법률(이하 신용정보법)'을 말한다. 데이터를 활용하는 과정에서 개인정보 처리는 피할 수 없기 때문에 정부는 개인정보를 안전하게 보호하면서 데이터 이용을 활성화하겠다는 취지로 2020년 8월 데이터 3법을 재정비하는 입법을 추진하였다.

정부는 데이터 3법 개정을 통해 개인정보 감독기관을 '개인정보보호위원회'로 일원화하

여 중복 규제로 인한 혼란을 줄였다. 또한 '개인정보보호법'에 가명정보를 도입하고 합리적 관련성에 따른 동의 없는 정보 사용을 일정 부분 허용하였으며, 개인정보 결합 제도를 마련하여 데이터를 활용할 수 있는 길을 넓혔다. '신용정보법'에는 개인정보 자기결정권에 기반한 개인신용정보전송요구권, 본인신용정보관리업 등을 도입하여 2021년 8월부터 시작되는 금융 분야 마이데이터 사업의 법적 근거를 마련하였다.

데이터 3법 개정 이후에도 개인정보보호위원회는 데이터 활용과 개인정보 보호, 두 가치 간의 균형을 더욱 제고하기 위해 '개인정보보호법' 개정 작업을 추가적으로 진행하고 있다. 2021년 2월 입법예고를 마친 '개인정보보호법' 2차 개정안은 개인정보 전송 요구권 도입, 국제표준에 부합한 국외 이전 방식 다양화, 형벌 중심 제재를 과징금 등 금전적인 경제벌 위주로 전환, 정보통신서비스 제공자 등에 대한 특례 정비, 이동형 영상정보처리기기 운영 기준 마련 등을 담고 있다.

나. 공공 생산 데이터 활용

1) 공공데이터의 제공 및 이용 활성화에 관한 법률

국민 생활 전반에 걸쳐 생성되는 공공데이터는 데이터산업의 중요한 자원이 된다. 2013년 7월 30일에 제정된 '공공데이터의 제공 및 이용 활성화에 관한 법률'은 공공기관에 법령 등에서 정하는 목적을 위해 보유하는 공공데이터의 품질을 관리하고 이를 국민에게 제공할 의무를 부여함으로써, 국민이 공공데이터를 이용할 수 있는 법적 근거를 마련했다. 또한 공공기관이 민간과 중복되거나 유사한 공공데이터 서비스를 개발·제공하는 것을 금지하고 공공데이터를 활용한 민간 창업을 지원하고 있다.

2) 데이터기반행정 활성화에 관한 법률

데이터에 기반한 정책 수립과 의사 결정을 활성화하기 위해 2020년 6월 9일 제정되었다. 데이터 통합관리 플랫폼을 구축하여 공공기관이 데이터를 효율적으로 공동 활용하고, 공공기관이 필요한 데이터를 보유한 타 공공기관 또는 민간기업에 해당 데이터 제공을 요청할 수 있도록 규정하고 있다. 데이터기반행정 활성화 위원회 설립, 데이터기반행정 책임관 임명 등의 내용도 담고 있다.

3) 국가공간정보기본법

그동안 공개가 제한된 고정밀 공간 정보를 민간 신산업이 활용할 수 있도록 '국가공간정보기본법'을 2021년 3월 16일 개정하여, 메타버스, VR/AR, 자율주행, IoT 산업 등에서 3D 공간정보 데이터를 활용할 수 있게 되었다.

4) 국가지식정보 연계 및 활용 촉진에 관한 법률

일명 '디지털 집현전법'으로도 부르는 '국가지식정보 연계 및 활용 촉진에 관한 법률'이 2021년 6월 8일 제정되었다. 국가지식정보는 공공기관이 생산·보유·관리하는 정보 중 지식의 활용 및 교육을 목적으로 국가적 이용 가치가 있는 디지털화된 정보와 디지털화의 필요성이 인정되는 정보를 말한다. 분야별, 기관별로 제공되던 국가지식정보를 하나의 플랫폼에 모아 서비스하여 국민들이 국가지식정보에 쉽게 접근하고 자유롭게 활용할 수 있도록 한다.

다. 데이터 산업진흥 및 이용촉진

최근 법제도는 위험 통제뿐만 아니라 산업 진흥을 위한 기초를 제공하는 인프라로서 기능하고 있다. 현재 데이터산업에서는 '지능정보화 기본법'이 이와 같은 역할을 하고 있으며, 2021년 10월에는 민간데이터의 생산·거래·활용을 지원하는 '데이터 산업진흥 및 이용촉진에 관한 기본법'이 제정되었다.

1) 지능정보화 기본법

'지능정보화 기본법'은 약 25년간 우리나라 정보화의 법적 기반이 되었던 '국가 정보화 기본법'이 2020년 6월 9일 전면 개정된 법이다. 정보화에만 초점을 둔 '국가 정보화 기본법'이 인공지능, 빅데이터, 사물인터넷 등 지능기술정보기술의 급발전을 포용하기에 한계가 있다는 판단에 '지능정보화 기본법'으로 개정되었다.

'지능정보화 기본법'은 범국가적 데이터 정책을 수립하도록 규정하고 있다. '공공데이터의 제공 및 이용 활성화에 관한 법률'은 공공데이터 활용 측면만을 다루고 있어 민간데이터를 포함한 국가적 차원의 데이터 정책에 미흡한 측면이 있었는데, '지능정보화 기본법'이 데이터센터 구축 및 활성화, 데이터 시책 마련, 데이터 유통·활용 지원 등을 규정하며 이를 보완하였다. 한편 공공데이터에 관한 사항은 '공공데이터의 제공 및 이용 활성화에 관한 법률'을 그대로 따르도록 하고 있어, '지능정보화 기본법'이 사실상 민간데이터에 대한 정책 마련과 지원만을 담고 있다는 평가가 있다.

2) 데이터 산업진흥 및 이용촉진에 관한 기본법

'지능정보화 기본법'으로 민간에서 데이터를 생산, 거래, 활용하는 기반을 마련하기에는 부족하다는 인식하에, 제21대 국회는 '데이터 산업진흥 및 이용촉진에 관한 기본법(이하 데이터 기본법)'을 제정하였다.

2022년 4월부터 시행되는 데이터 기본법은, 지능정보화 기본법상 데이터 관련 규정을 구체화하는 동시에 민간데이터 활용에 대한 기본법적인 성격을 가진다. 데이터 기본법은 아래의 주요 내용들을 바탕으로 데이터 기반 산업과 서비스가 급증하는 데이터 경제에서 데이터로부터 다양한 가치를 창출하고 데이터 거래 시장을 조성하기 위한 기반을 마련하도록 하고 있다.

- 데이터 거버넌스 구축

정부는 3년마다 데이터산업 진흥을 위한 기본계획과 연차별 시행계획을 의무적으로 수립해야 하고, 국무총리 직속으로 데이터 생산·거래·활용 촉진에 관한 심의 기구인 '국가데이터정책위원회'를 신설하게 된다.

- 데이터 유통시장 활성화

또한 정부는 데이터 유통시장을 활성화하기 위한 기반을 마련하여야 한다. 정부는 ① 데이터를 원활하게 이동시킬 수 있는 제도와 ② 객관적인 데이터 가치평가 기준을 마련해야 한다. ③ 데이터 거래를 전문으로 하는 데이터거래사가 신설되며 ④ 데이터 생산, 거래 및 활용 관련 분쟁을 조정하는 데이터분쟁조정위원회 등을 통해 데이터 거래의 분쟁 비용을 낮춰야 한다. ⑤ 데이터 가치 평가 지원과 품질관리를 위해 관련 기준 및 전담기관을 지정해야 한다.

- 데이터 이용 활성화

데이터기본법은 정부가 ⑥ 데이터 플랫폼과 ⑦ 데이터를 안전하게 분석·활용할 수 있는 데이터 안심구역을 지원하여 데이터 이용을 활성화하도록 하고 있다. 이 외에도 정부는 ⑧ 데이터 자산 보호, ⑨ 세제 지원, ⑩ 표준화 등을 위해 노력해야한다.

이처럼 데이터기본법은 데이터 경제를 활성화하기 위해 정부가 앞장서 데이터 거래 시장과 기준을 조성하는 내용을 담고 있으며, 앞으로 이를 통한 성과를 지켜볼 필요가 있다.

3. 마무리

디지털 뉴딜 2.0은 데이터 댐 구축 및 활용, 인공지능 산업 육성 등 기존 디지털 뉴딜 1.0의 정책을 지속적으로 추진하면서도, 디지털 뉴딜 1.0의 성과를 확산하기 위해 마이데이터 산업 활성화, 초연결 신산업 육성, 클라우드 강화, 디지털 복제물 활용 확대 사업 등을 추가적으로 진행한다. 이러한 정책들을 통해 국민이 데이터 시대의 이점을 체감함으로써 우리나라가 디지털 대전환하는 데 디딤돌이 되길 기대한다. 다만, 여러 부처에서 다양한 데이터 정책을 추진하는 만큼 통합적인 체계를 통해 지속적으로 관리함으로써 정책 성과를 제고할 필요가 있어 보인다.

제2장. 해외 데이터산업 관련 정책·법제도

디지털 경제로의 전환이 가속화됨에 따라, 주요국들은 디지털 경제를 구성하는 기본 요소인 데이터 산업에 대한 지원을 강화해나가고 있다. 미국은 '산업 고도화를 통한 질적 성장'을 뒷받침하기 위한 데이터산업 육성정책을 수립하였으며, 중국은 '기업 성장 지원을 통한 산업 육성' 정책을 추진해나가고 있다. 이 가운데 유럽연합은 '유럽 통합형 데이터산업 발전 체계 구축'을 추구하며 글로벌 경쟁력 강화에 나섰고, 일본은 공공데이터 활용으로 '민관 협력을 통한 데이터산업 발전'을 이뤄 전반적인 사회의 디지털화를 계획하고 있다. 이 장에서는 미국, 유럽연합, 일본, 중국의 데이터 정책 수립 흐름 및 현황을 정리하고, 각 국가의 데이터산업 정책 특징과 관련된 법제도에 관해 서술해보고자 한다.

1. 해외 데이터 정책

이준호(이씨이십일 R&C 연구원)

가. 미국: 데이터 전략을 기반으로 '산업 고도화를 통한 질적 성장'

미국 정부의 데이터산업 정책은 전반적으로 '산업 고도화를 통한 질적 성장'을 추구하고 있다. 구글, 아마존, 애플 등 미국의 거대 IT 기업들이 이미 데이터 시장을 주도하고 있기에, 미국 정부는 데이터센터 효율화, 데이터 인프라 고도화 등 데이터산업의 질적 성장을 위한 제도 정비에 중점을 두고 있다.

미국 정부는 2012년 3월 백악관 과학기술정책실(Office of Science and Technology Policy, OSTP)이 발표한 '빅데이터 이니셔티브(Big Data Initiavies, BDI)'[1]를 기점으로 데이터산업 육성 지원을 본격화하였다. '빅데이터 이니셔티브'는 데이터 접근성 향상, 데이터 관리 기술 발전을 위한 지원, 데이터 전문 인력 양성 프로그램 확대 등의 내용을 포함하고 있다. 2016년에는 네트워크 및 정보기술 연구개발 프로그램(Networking and Information Technology Research and Development, NITRD)이 '연방 빅데이터 연구개발 전략 계획(Federal Big Data Research and Development Strategic Plan)'을 발표하고 빅데이터 전문 인력 양성, 데이터 혁신[2] 생태계 강화, 빅데이터 핵심 연구 기술 확보 등을 포함한 7대 전략 계획을 수립하였다.

2020년 1월 미국 최고데이터책임자위원회(Chief Data Officers Council, CDO Council)[3]가 '연방 데이터 전략(Federal Data Strategy, FDS)'을 발표하며 미국 데이터산업 육성 정책의 기본 방향을 구체화하였다. '연방 데이터 전략'은 연방 기관이 데이터 윤리를 준수하여 데이터 과제를 달성하고 공익을 실현하는 것을 목표로 한다. 해당 전략은 각 정부 부처가 부처에 맞는 세부 계획을 수립하고 실행하는 것으로, 10개의 원칙, 40개의 모범 사례, 20개의 과제로 구성되어 있다. 부처별로 세부 목표가 상이하나 공통적으로 데이터 거버넌스[4] 구축, 데이터 인프라 고도화, 데이터 인력 양성 등에 초점을 맞추고 있다. '연방 데이터 전략'을 반영하여 각 정부 부처별로 후속 정책이 나오고 있다.

대표적으로 최근에 발표된 미국 농무부(U.S.Department of Agriculture, USDA)의 '농무부 데이터 전략(USDA Data Staregy)'이 있다. 해당 정책의 첫 번째 목표는 농업 분야에서의 데이터 거버넌스와 데이터 리더십(data leadership) 강화로, 데이터 및 분석·개발, 인프라, 도구에 대한 접근을 높이는 것이며, 두 번째 목표는 데이터 분석 인력을 모집하고 재교육하여 유지함으로써 강력한 데이터 기반 환경을 조성하는 것이다. 세 번째 목표는 공공데이터에 대한 분석과 데이터 플랫폼 구축을 바탕으로, 활발한 데이터 공유 및 활용을 이루어 혁신을 추진하는 것이다. 네 번째 목표는 오픈 데이터(Open Data)[5]로 대중에게 깊은 인사이트(Insight), 가치, 투명성을 제공하기 위해 효과적인 데이터 공유를 지원하고 촉진하는 것이다. 농무부 이외에도 국방부(Department of Defense, DoD), 보건부(Health and Human Services, HHS) 산하 식품의약국(Food and Drug Administration, FDA) 등 다수의 정부 기관이 연방 데이터 전략을 기반으로 한 데이터 정책을 수립 및 발표하고 있다. 추가로 올해 4월 최고데이터책임자위원회가 발표한 '연방 데이터 전략 2021 실행계획(Federal Data Strategy 2021 Action Plan, FDS 2021)'에서는 기존 연방 데이터 전략에 제로트러스

1 이하 주요 국가의 데이터 정책에 대한 세부 내용은 한국데이터산업진흥원에서 발행한 「월간 데이터이코노미 2021년 1~8호」, 「데이터산업 동향 이슈 브리프 2021년 8호」, 「2020 데이터산업백서」 보고서 내용을 참조

2 새로운 데이터 소스와 다른 데이터 소스가 발전된 기술을 통해 더 많은 정보를 얻는 것을 의미함

3 미국은 데이터산업 육성을 지원하는 기관을 별도로 두고 있지 않음. 대신 주요 정부 기관들이 각각 최고데이터책임자(Chief Data Officers, CDO)를 두고 있으며, 이러한 최고데이터책임자들이 모이는 최고데이터책임자위원회에서 데이터산업 육성 정책을 수립하고 있음

4 데이터 활용, 보호, 정확성을 등을 위한 데이터 통합 관리 체계를 의미함

5 개방형 데이터라고도 불리며, 모든 이용자가 자유롭게 이용 및 재가공하여 배포할 수 있는 데이터를 의미함

트(Zero Trust)[6]에 기반한 데이터 보안 강화, 코로나19 데이터 처리를 위한 인프라 구축 확대 등의 내용이 포함되었다.

이렇듯 양적 성장보다 질적 성장을 추구하는 미국 데이터산업 육성의 정책 특성은 앞서 언급한 '연방 데이터 전략' 같은 주요 정책뿐만 아니라 세부 정책에서도 확인할 수 있다. 2016년부터 진행 중인 '데이터센터 최적화 이니셔티브(Data Center Optimization Initiative, DCOI)'를 통해 저전력 고효율의 데이터센터가 구축되면서, 핵심 시설을 제외한 신규 데이터센터 건설이 2017년 기준 5,426개에서 2020년 기준 1,613개로 크게 줄어들었으며, 매년 지속적으로 데이터센터 폐쇄를 진행하여 예산을 절감하고 있다. 이러한 미국의 정책 기조는 향후로도 지속될 것으로 전망된다.

나. 유럽연합: 유럽데이터 전략으로 '통합형 데이터산업 발전 체계 구축'

유럽연합은 유럽연합 집행위원회(European Commission, EC) 주도 하에 정책을 수립하고 있다. 특히 2010년 중후반부터 글로벌 데이터 시장에서 미국·중국의 영향력이 확대되면서, 페이스북, 아마존 등 타국 빅테크 기업을 규제하고 EU 내 데이터 생태계 확대 및 기업 육성에 집중하고 있다.

유럽연합 집행위원회는 2014년 '번영하는 데이터 기반 경제를 향하여(Towards a thriving data-driven economy)'라는 결의안을 채택하며 데이터 경제에 대한 초기 개념을 확립하였다. 2017년에는 데이터 경제에 대한 개념 정립을 바탕으로 데이터 경제 활성화를 위한 지원 내용을 담은 '유럽 데이터 경제(European Data Economy)'를 발표하였다.

이후, 2020년 2월 유럽연합 집행위원회는 유럽연합 내 데이터산업의 국제 경쟁력을 확보하고, 데이터 주권(Data Sovereignty)[7]을 보장하며, 단일 데이터 시장 구축을 추진하는 '유럽 데이터 전략(European Data Strategy)'을 발표하였다. 특히 유럽연합 27개 회원국 간의 분절된 데이터 규범·정책 간 조화를 추구하여 유럽 단일 데이터 공간(single European data space)을 구축하는 것은 해당 정책의 핵심 전략이다. 이외에도 데이터 접근 및 재사용에 관한 거버넌스 프레임워크 구축, 데이터산업에 대한 투자와 데이터 활용을 위한 역량 및 인프라 강화, 데이터 보호 권한 강화 등의 내용을 포함하고 있다. 2021년 7월 유럽데이터 전략의 첫 번째 입법 구성 요소인 '데이터 거버넌스법(Data Governance Act, DGA)'이 유럽의회위원회(European Parliament committee) 승인을 받았으며, 10월 1일에는 '데이터 거버넌스법'의 내용(text)이 유럽연합이사회(The Council of the European Union)에서 합의에 도달했다.

이렇듯 유럽연합은 '유럽 통합형 데이터산업 발전 체계 구축'을 목표로 다양한 전략 정책을 수립하고 있다. 데이터 인프라 개발 정책인 '가이아엑스(GAIA-X)', '유럽 오픈 사이언스 클라우드(European Open Science Cloud, EOSC)'도 그 중 하나이다. 이러한 통합과정에서 유럽연합은 데이터의 적극적인 활용을 장려하되, 데이터 보안 문제를 최소화하기 위해 '유럽연합 개인정보보호규정[8](General Data Protection Regulation, GDPR)'을 제정하였고, 이를 통해 궁극적으로 자유롭고 안전한 데이터 흐름(free and safe data flows)을 추구하고 있다.

다. 일본: 민관 협력 통한 '포괄적 데이터 전략' 수립

일본 정부는 데이터 정책 수립 시 '민관 협력을 통한 데이터산업 발전'을 우선순위로 고려하고 있으며, 최근에는 '디지털 전환(デジタルトランスフォーメーション)'을 통해 '전반적인 사회의 디지털화'를 중점 과제로 두고 있다.

2016년부터 '오픈 데이터 2.0(オープンデータ2.0)'을 통해 데이터를 활용하여 사회문제를 적극적으로 해결할 것이라고 발표하였다. 또 같은 해에 '민관 데이터 활용 추진기본법(官民データ活用推進基本法)' 제정을 통해 데이터 통신 환경 정비를 위한 정책적 프레임워크를 구축하였고, 이를 바탕으로 2017년 '세계 최첨단 IT 국가 선언·민관 데이터 활용 추진계획(世界最先端IT国家創造宣言·官民データ活用推進基本計画)'을 수립하였다. 2018년에는 데이터 표준화를 위한 기반 정비 및 행정서비스의 100% 디지털화와 데이터 활용에서의 민관 협력 강화 등의 내용을 담은 '세계 최첨단 디지털 국가 선언·민관 데이터 활용 추진 계획(世界最先端デジタル国家創 造宣言·官民データ活用推進基本計画)'을 발표하였다.

이후 2021년 6월 일본 내각관방(内閣官房)은 현재 일본의 대표 데이터 전략인 '포괄적 데이터 전략(包括的データ戦略)'을 수립하고 구체적인 구현 방안을 발표하였다. 해당 전략의 목표는 디지털 트윈(Digital Twin)[9] 구축으로 사회적 과제 해결과 경제적 발전을 동시에 달성함으로써 인간 중심의 디지털 사회를 구현하는 것이다. 이를 위해 일본 정부는 민관 협력을 바탕으로 한 트러스트 서비스 구축, 데이터 플랫폼·기초 데이터·디지털 인프라 정비, 데이터

6 엄격한 ID 확인 과정을 통해 사용자를 보호하는 보안 모델
7 특정 국가 또는 권역에 수집된 데이터는 해당 국가 또는 권역의 데이터 법률 및 거버넌스가 적용되어야 한다는 원칙
8 유럽연합 집행위원회에서 발표한 개인정보보호 법령으로 정보 주체의 권리와 기업의 책임성을 강화하고 개인정보의 유럽연합 역외 이전 요건을 명확히 한 법령
9 현실 공간과 가상 공간을 고도로 융합시킨 시스템

거래 시장 활성화 같은 세부 방안을 제시하였다. 이와 함께 '마이넘버카드(個人番号カード)' 보급 및 디지털 서비스의 UX/UI 개선 및 대국민 서비스 실현 등의 내용을 포함하고 있는 '디지털 사회 실현을 위한 중점 계획(デジタル社会の実現に向けた重点計画)'을 발표하였다.

일본 데이터산업 육성 정책의 특징으로 볼 수 있는 '민관 협력'의 맥락에서 일본 정부는 최근에 마이넘버카드[10]의 보급을 선결과제로 진행하고 있다. 행정, 의료, 사회 보험 등을 민간 금융 서비스와 연동시켜 일본 내 데이터산업을 육성한다는 계획이며, 관련 정책을 지원하는 '디지털청'이 2021년 9월 1일 발족하면서 '전반적인 사회의 디지털화'를 이룰 계획이다.

라. 중국: '14차 5개년 규획'에서 기업 성장 지원을 통한 데이터산업 육성 추구

중국 정부는 '기업 성장 지원을 통한 데이터산업 육성'을 우선적으로 고려하여 정책을 수립하고 있다. 이러한 정책 기조는 데이터산업 육성 주요 정책인 '14차 5개년 규획(十四五规划)'과 세부 정책인 '빅데이터 산업 발전 규획 2016-2020(大数据产业发展规划 2016-2020)', '빅데이터 산업 개발 시범 프로젝트(大数据产业发展试点示范项目)'를 통해 확인할 수 있다.

중국 정부는 2015년 '빅데이터 발전 촉진을 위한 행동 요강(促进大数据发展行动纲要)'에서 데이터 분류 표준 체계를 수립하고, 데이터산업에 대한 창업 지원 등의 계획을 발표했다. 2016년에는 '빅데이터 산업 제13차 5개년 규획(大数据产业十三五规划)'을 발표하여 데이터산업 발전 방향을 제시하였다. 해당 계획에서 중국 정부는 빅데이터를 활용한 기술 및 제품의 연구개발 강화, 산업 빅데이터 적용 강화 및 애플리케이션 개발, 빅데이터 산업체 육성, 빅데이터 표준 시스템 구축, 데이터 보안 등 데이터산업 육성에 대한 전반적인 가이드라인을 정립하였다. 이어 올해 4월 발표한 '14차 5개년 규획'에서는 데이터센터 설립 강화, 빅데이터 애플리케이션의 실물경제 전환, 산업 부문별 빅데이터 활용 강화(제조업, 건강관리, 교통, 농업, 에너지 등), 빅데이터 시범 산업단지 건설 가속화 등을 데이터산업에 대한 전반적인 발전 방향으로 제시하였다. 이를 바탕으로 한 세부적인 데이터산업 정책인 '빅데이터 산업 제14차 5개년 규획(大数据产业十四五规划)'은 현재 중국공업정보화부(工业和信息化部)[11]에서 수립 중이며, '광둥성 디지털 정부를 위한 14차 5개년 규획(广东省数字政府改革建设十四五规划)', '저장성 디지털 경제 발전을 위한 14차 5개년 규획(浙江省数字经济发展十四五规划)' 등 지방 정부인 성(省) 차원에서도 세부적인 데이터산업 지원 정책을 수립 중이다.

중국의 데이터산업 정책 특징은 '기업 성장 지원을 통한 산업 육성'이다. 특히 중국 내 산업 정책을 수립 및 총괄하는 기관인 공업정보화부를 통해 빅데이터 산업 발전 계획을 발표하여 데이터산업의 혁신 및 개발 역량을 강화하는 데 중점을 두고 있다. 또 데이터 기술 제품의 연구 개발을 정부 차원에서 지원하고 데이터산업 전체의 인프라 구축을 국가적 사업으로 진행하고 있다. 2020년에는 데이터산업 육성을 위해 200개 사업을 시범 사업으로 선발해 지원하는 빅데이터 산업 개발 시범 프로젝트를 추진하였다. 특히 산업용 데이터 부문이 90개로 가장 많이 선정되었는데, 이는 산업 부문에서의 데이터 활용 기업을 집중적으로 양성하려는 중국 정부의 의도가 반영된 것으로 볼 수 있으며, 기업 성장을 통해 데이터산업을 육성하려는 중국 정부의 움직임은 향후에도 계속될 것으로 전망된다.

[표 2-2-1] 주요국 데이터 정책 현황

분류	미국	EU	일본	중국
특징	산업 고도화를 통한 질적 성장	유럽 통합형 데이터 산업 발전 체계 구축	민관 협력을 통한 데이터산업 발전	기업 성장 지원을 통한 산업 육성
주요 정책	연방 데이터 전략 - 데이터 거버넌스 구축 - 데이터 인프라 개발 - 데이터 인력 양성	유럽데이터 전략 - 단일 데이터 시장 구축 - 데이터 접근 및 재사용에 관한 거버넌스 프레임워크 구축 - 데이터산업에 대한 투자와 데이터 활용을 위한 역량 및 인프라 강화 - 데이터 보호 권한 강화	포괄적 데이터 전략 - 트러스트 서비스 구축 - 데이터 플랫폼·기초 데이터·디지털 인프라 정비 - 데이터 거래 시장 활성화	14차 5개년 규획 - 데이터센터 설립 강화, 빅데이터 애플리케이션 실물경제 전환 - 빅데이터 시범단지 건설 - 각 산업 부문별 빅데이터 활용 강화
데이터산업 육성 기구	최고데이터책임자위원회	유럽연합 집행위원회	디지털청	공업정보화부
세부 정책	데이터센터 최적화 이니셔티브	유럽 오픈사이언스 클라우드	마이넘버카드	빅데이터 산업 개발 시범 프로젝트

10 마이넘버카드의 핵심은 원활한 데이터 이동(data flow)을 통한 '본인 인증'기능이며, 원활한 데이터 이동을 위해서는 정부기관, 민간기업, 개인 간 데이터 교환 인프라 구축이 선행되어야 함

11 공업정보화부는 중국 내 데이터산업 육성 전담 기구로 국가 데이터산업 육성 기본 전략 수립, 국가 데이터 시범 프로젝트 사업 운영, 통합형 데이터센터 구축 등을 추진하고 있음

2. 해외 데이터 법제도

이성엽(고려대 기술경영전문대학원 교수)

가. 미국의 법제도

개인정보를 포함하는 별도의 데이터 일반법은 존재하지 않으며, 연방정부를 규제하는 프라이버시법 외 분야별 개인정보보호법 등이 있으며, 다양한 데이터 개방 관련 법률이 존재한다. 먼저 '프라이버시법(Federal Privacy Act, 1974)'은 연방정부 차원에서 공공 부문의 데이터 및 정보를 보호하는 법규로서 일반법으로 제정되었다. 미국 정부 기관이 보유한 데이터 및 정보를 보호하는 규정이 명시되어 있으며, 일반 국민은 본인의 기록에 들어있는 정보의 내용과 사용 상황을 알 수 있어야 하고, 일정한 목적에 따라 수집된 정보의 목적 외 사용은 금지된다.[12] 공공 부문 프라이버시와 관련되어 제정된 법으로는 '컴퓨터정보 결합 및 프라이버시보호법(Computer Matching and Privacy Protection Act, 1988)', '운전자 프라이버시 보호법(Driver's Privacy Protection Act, 1994)' 등이 있으며, 민간 부문 프라이버시와 관련되어 제정된 법으로는 '가족의 교육권 및 프라이버시법(Family Education Rights and Privacy Act, 1974)', '금융프라이버시권리법(Right to Financial Privacy Act, 1978)', '전자통신프라이버시법(Electronic Communication Privacy Act, 1986)', '비디오 프라이버시 보호법(Video Privacy Protection Act, 1988)', '근로자 거짓말탐지기 보호법(Employee Polygraph Protection Act, 1988)', '아동 온라인 프라이버시 보호법(Child Online Privacy Protection Act, 1998)' 등이 있다.

'전자정보자유법(Electronic Freedom of Information Act, 1996)'은 웹 2.0 기반의 거버먼트 2.0 정책의 하나로, 172개 공공기관의 지리, 교육, 교통 등 약 38만 개의 공공정보를 Data.gov(2009.5 개설)로 개방한다. 연방 정부의 공공 문서에 기록된 공공 부문의 데이터 및 정보를 강제로 공개하게 하여 민간이 자유롭게 활용할 수 있는 권리를 부여하고 있다. 정보공개청구인의 범위를 적절하게 직접적으로 관계된 자(persons properly and directly concerned)로 규정하던 종전의 '연방행정절차법'과 달리, 모든 사람(any person)으로 확대하였다. 또한 전자정보 환경에 적합하도록 정부 기록의 온라인 접근을 허용하고, 정보공개청구인에게 가장 유용한 형태로 정보를 제공하도록 하며, 기록을 공개할 것을 요구받은 행정기관은 요구받은 기록을 찾아내기 위해 합리적으로 노력할 의무를 부과하고 있다.

'전자정부법(E-government Act, 2002)'은 전자정부를 국민을 중심으로 하여 통합적으로 추진하고, 행정기관 간의 협력과 민간-정부 간 협력을 촉진하여 국민의 권익을 보장하는 것을 목적으로 한다. '전자정부'란 웹 기반 인터넷 애플리케이션 및 기타 정보 기술을, 이를 구현하는 프로세스와 결합하여 공공 및 기타 정부 기관에 대한 정부 정보 및 서비스의 접근 및 전달을 강화하는 용도, 효과, 효율성, 서비스 품질 또는 혁신을 포함할 수 있는 정부 운영의 개선을 가져오는 용도로 사용하는 것을 의미한다고 정의하고 있다. 법안의 주요 내용으로는 전자정부 구축을 위한 전담기관인 전자정부국 신설, 정보화책임관협의회 법제화, 전자정부기금 설치 및 운용, 사이버보안 등이 있다.

'데이터 품질법(Data Quality Act, 2010)'은 공공 및 연방정부의 목적과 조항을 이해하기 위해 연방 정부가 배포한 정보 및 데이터의 품질, 객관성, 유용성 및 무결성 보장을 규정한다. '오픈 정부 데이터법(OPEN Government Data Act, 2019)'은 연방정부의 오픈 데이터 공약을 제도화하고, 모든 정부 데이터에 대해 '기본적으로 개방'하는 정책을 성문화하고 있다. 연방정부의 예산 사용 정보를 투명하고 정확하게 제공하기 위한 'Data Act of 2014(Digital Accountability and Transparency Act of 2014)'의 시행에 이어 공공데이터의 기계 판독성(Machine-Readable)을 강조하고 있다. 공공기관에서 이용할 수 있는 정부 데이터는 기계 판독이 가능한 데이터로 공개되어야 하고, 개방형 라이선스를 사용해야 하는 것이어야 한다. 오바마 행정부의 오픈 데이터 정책은 트럼프 행정부에 와서도 그 기조를 유지하고 있으며, 정부 데이터의 혁신적인 사용을 지원하고, 정부 전체에 걸쳐 일관된 데이터 관행을 채택하며, 개방형 데이터의 모범 사례를 개발하도록 각 기관에 지시한다.

한편 최근에는 주(州)법이기는 하지만 미국에서도 개인정보보호에 관한 일반법이 제정되고 있다. 2020년 1월부터 시행 중인 '캘리포니아 소비자 프라이버시 보호법(California Consumer Privacy Act 2019, CCPA)'이 대표적이다. CCPA는 캘리포니아주 이외에 거주하는 주민에게는 적용되지 않지만, 캘리포니아 내에 사업장이 없더라도 캘리포니아 주민을 대상으로 제공하는 서비스에는 원칙적으로 적용된다. CCPA에 의하면, 개인정보 수집 시 일정한 사항을 소비자에게 고지하여야 한다. 그리고 개인정보를 제3자에게 판매하는 경우 'Do not Sell my info' 문구와 링크를 웹사이트를 통해서 제공하여야 하고, 정보주체에게 사후 거부 의사를 밝힐 수 있도록 해야 한다. 나아가 CCPA는 데이터 주체에게 통지를 요구할 수 있는 권리, 데이터에 대한 접근을 허용할 권리를 부여하고 있다. 특히 CCPA는 개인정보처리방침에 캘리포니아의 개인정보 권리에 대한 설명을 명시하고 공개해야 하며, 개인정보의 판매 여부 및 목적의 고지 및 공개 의무 뿐만 아니라 이러한 사항에 대한 연 1회 이상의 업데이트 의무도 부과하고 있다.

12 이하 주요 국가의 데이터법제는 이성엽, 「데이터 유통 및 활용 확대를 위한 제도개선에 관한 연구」, 한국데이터산업진흥원, 2020.12월 보고서 내용을 참조

또한 미국 캘리포니아주에서 2020년 11월 3일 주민 투표로 '캘리포니아주 프라이버시 권리법(The California Privacy Rights Act of 2020, CPRA)'이 통과되어 2023년 1월 1일부터 시행될 예정이다. CPRA는 CCPA를 바탕으로 유럽연합의 GDPR에 보다 근접하도록 소비자의 프라이버시에 대한 권리를 확대함으로써 사업자에 대한 규제가 강화되었으며, 특히 미국 최초로 프라이버시 분야를 담당하는 규제기관인 캘리포니아 프라이버시보호국(California Privacy Protection Agency) 설립의 근거를 마련하였다는 점에서 의미가 있다.

나. EU의 법제도

EU는 데이터 보호와 활용을 위한 기본 법제로 '일반개인정보보호법(General Data Protection Regulation, GDPR)'을 시행하고 있다. 동법은 미국, 일본, 한국 등에도 영향을 미치는 모범 입법으로 역할을 하고 있다. 그 외 여러 데이터 활용과 보호를 위한 지침 등을 운영하고 있다.

'데이터베이스 지침(The Database Directive, Directive 96/9/EC)'은 데이터베이스 저작자나 제작자는 타인이 자신의 데이터베이스를 사용하는 것을 금지할 수 있는데, 데이터베이스를 전체적으로 또는 상당한 부분을 이용하는 행위와 데이터베이스의 통상적인 이용과 충돌하거나, 데이터베이스 제작자의 합법적인 이익을 부당하게 해치는 행위를 수반하는, 데이터베이스의 상당하지 않은 내용의 반복적이고 체계적인 추출 및 재이용 행위를 금지한다. 다만, 기계에 의해 생성된 데이터베이스에는 적용되지 않는다.

'유럽 일반 데이터 보호규칙 및 e프라이버시 지침(The General Data Protection Regulation and the lex specialise Privacy Directive)'은 개인데이터의 수집, 사용, 접근 및 이동성(portablility) 외에도 개인데이터의 양도나 전송 가능성 등의 개인데이터 처리 규정을 두고 있다. 데이터의 이동성에 대한 권리는 서비스 제공자에게 제공된 개인데이터 복사본을 입수하거나 또 다른 서비스 제공자(data controller)에게 전달하는 것을 개인들에게 허용하는 내용이다.

'비개인데이터의 자유로운 이동 체계에 관한 규칙(Regulation (EU) 2018/1807 on a framework for the free flow of non-personal data in the European Union)'은 데이터 현지화 요구, 관할 당국의 비개인데이터 활용 가능성, 전문적인 이용자를 위한 비개인데이터 포스팅에 관한 사용을 규정하여, 유럽연합 단일시장 내에서 비개인데이터의 자유로운 거래를 보장하는 것이 이 규칙의 목적이다(제1조). 여기서 비개인데이터란 '식별되거나 식별 가능한 개인과 연결되지 않은 모든 정보'를 말한다. GDPR에서 정의된 개인데이터가 아닌 모든 데이터가 해당되며, 이 규칙은 개인데이터와는 별도로 유럽연합 역내에서 디지털화된 데이터가 자유롭게 처리될 수 있도록 하고 데이터를 저장하거나 처리할 수 있는 장소에 대한 제한을 금지한다.

'개방형 데이터 및 공공 부문 정보의 재사용에 관한 지침(Directive (EU) 2019/1024 on open data and the re-use of public sector information)'은 공공정보에 초점을 맞추어 기존의 '공공정보 활용 지침(Directive 2013/37/EU)'을 대체하고 있다. 어떤 목적으로든 누구나 자유롭게 활용할 수 있는 개방형 데이터와 민간 또는 상업적 목적을 위한 공공부문 정보의 광범위한 활용을 장려하고 거래를 촉진한다.

유럽 데이터 거버넌스법(DGA)은 2020년 11월 25일 데이터 중개자에 대한 신뢰도를 높이고 유럽 전역에 걸친 데이터 공유 메커니즘을 강화하여 데이터 활용의 가용성을 높이는 것을 목적으로 제안되었다. 2021년 10월 1일 DGA의 내용이 유럽연합이사회에서 합의에 도달하며 유럽 전역에서 공공기관이 보유한 데이터의 재사용을 위한 조건을 제시하고 모든 형태의 보상을 대가로 한 기업 간 데이터 공유 활성화를 위한 통지 및 감독 체계 등 거버넌스 기반을 마련하고 있다. 정보 주체가 이타적인 목적으로 공익상의 데이터 활용을 위해 자신의 개인정보를 제공하여 데이터화할 수 있도록 하고, 이를 활용하는 거버넌스 기반을 마련한 것이다.

다. 일본의 법제도

공공과 민간을 포함하는 데이터 활용을 위한 기본법과 개인정보보호의 기본법으로서 '개인정보보호법(個人情報の保護に関 する法律)'을 운영하고 있다. 먼저 법률 제103호 '민관데이터 활용 추진 기본법(官民データ活用推進基本法)'이 있다. 이는 기본 이념, 민관데이터 활용 추진 기본 계획, 기본 시책, 민관데이터 활용 추진 전략 회의 등의 기본법적 내용으로 구성되어 있다. 민관데이터의 적절하고 효과적인 활용을 위하여 데이터의 원활한 유통 확보, 자립적이고 개성 넘치는 지역 사회의 형성, 새로운 사업의 창출, 국제 경쟁력 강화 등을 도모하는 것이 법의 취지이다(제5조). 한편 2020년 6월 '개인정보보호법(個人情報の保護に関 する法律)'개정으로 익명가공정보제도가 도입되며 개인데이터 활용 확대의 기반이 되었다.

법률 제25호 '생산성 향상 특별 조치법(生産性向上特別措置法)'은 '민관데이터 활용 추진 기본법'의 취지를 고려하여, 일본 총무성 장관 및 경제산업성 장관은 계획 실행 기간 안에 혁신적 데이터산업 활용에 관한 지침을 마련해야 한다(제21조 제1항)는 규정을 두고 있다.

산업의 생산성 향상을 단기간에 실현하기 위한 조치의 하나로 '혁신적 데이터산업 활용의 촉진'을 강구하며, 혁신적 데이터산업 활용에 관한 지침, 혁신적 데이터산업 활용 계획의 인정·변경, 국가기관 등에 대한 데이터 제공 요구를 규정하고 있다. 여기서 혁신적 데이터의 산업 활용이란 혁신적 사업 활동 중 전자적 기록에 기록된 데이터를 혁신적인 기술 또는 방식을 이용하여 수집하고 산업 활동에 이용하는 것을 말한다(제2조 제4항).

'AI·데이터의 이용에 관한 계약 가이드라인(AI·データの利用に関 する契約ガイドライン)'은 민간사업자 등이 데이터의 이용 등에 관한 계약을 체결할 때, 계약상 주요 이슈와 논점, 계약 조항의 예시나 작성 시 고려 사항 등을 정리한 것이다. 거래비용을 절감하고, 데이터 계약의 보급을 도모하며, 데이터 활용의 촉진을 이끌어 내는 것이 법의 목적이다.

2019년 7월 개정 시행된 '부정경쟁방지법(不正競争防止法)'은 한정제공 데이터의 개념을 규정하고, 한정제공 데이터를 부정하게 취득, 사용, 공개하는 행위를 영업비밀의 부정취득 등에 준하여 규율하는 방식을 채택하고 있다. 기술적 제한수단의 효과를 저해하는 행위의 보호 대상에 데이터의 처리를 포함한다. 법에서는 '한정제공 데이터'를 '업으로서 특정한 자에게 제공한 정보로 전자적 방법에 의하여 상당량이 축적, 관리되고 있는 기술상 또는 영업상의 정보(비밀로서 관리되고 있는 것은 제외)'로 규정하고, 이러한 한정제공 데이터를 부정취득, 사용, 공개하는 행위를 새로운 부정경쟁행위로 추가함으로써 저작권법 등 기존의 법률에서 보호되지 않는 데이터를 보호하도록 하였다.

라. 중국의 법제도

개인정보 및 데이터에 관해 다양한 법이 존재하고 있다. 중국 내 개인정보 보호에 관한 논의가 활발해지며 중국 내 데이터 활용, 역외 이동, 안전조치 등에 관한 사항을 체계적으로 다루는 관련 법제 제정이 이어지고 있다. 2017년부터 시행된 '네트워크안전법(网络安全法)'은 사이버 안전을 보장하고, 정보통신망의 주권 확보와 국가안전, 사회 공익을 수호할 목적으로 네트워크 운영 안전, 개인정보보호, 불법 정보 단속 등 규정과 네트워크 제한 조치, 법률 책임 등을 규정하고 있다. 이에 따라 사이버 테러리즘에 의한 공격이나 인터넷 사기 등을 방지하고, 긴급한 경우 정부가 네트워크를 차단하는 권한을 가진다. 2020년 10월엔 '개인정보안전규범(信息安全技术 个人信息安全规范)'이 시행되며 정보시스템 상에서 개인정보 처리시 원칙과 안전조치를 규정하였다.

개인정보 안전영향평가 관련 용어를 정의하고, 구체적 평가방법에 대해 규정한 '개인정보안전영향평가지침(信息安全技术 个人信息安全影响评估指南)'이 2021년 6월 시행되었으며, 2021년 9월 시행된 '중화인민공화국 데이터보안법(中华人民共和国数据安全法)'은 데이터 개발 이용 촉진, 개인 및 조직의 권익 보장, 국가 보안 유지, 발전이익 보호 등을 목적으로 데이터 보안 제도, 데이터 규제범위 확대 등을 규정한다.

한편 민법 관련 규정들을 통일한 '민법전(民法典)'이 2021년 1월 1일부터 시행되었다. 동법에서는 제1편 총칙에서 개인정보, 데이터를 민법상의 사적재산으로 인정하고 있으며, 제127조는 '데이터 및 온라인상 가상자산에 대한 민사적 보호가 가능하다.'고 규정하고 있다.

2021년 11월 시행을 앞둔 '중화인민공화국 개인정보보호법(中華人民共和國 個人信息保護法)'은 개인정보 및 민감정보의 정의, 개인정보주체의 권리 보장 확대, 역외적용 범위 및 개인정보의 역외 이동 등을 규정하고 있다. 동법은 '개인정보'를 전자 또는 기타 방식으로 기록되어 식별되었거나 식별이 가능한 자연인과 관련된 각종 정보로 정의한다. 익명화 처리된 후의 정보는 제외된다. 이는 기존의 '네트워크안전법'과 '민법전'상의 개인정보 정의와 유사하다.

공업·정보화부(工業和信息化部)의 '빅데이터산업발전규획 2016-2020(大数据产业发展规划 2016-2020)'은 데이터 개방 확대, 플랫폼 및 오픈소스 기술 지원, 빅데이터 전문 SW 수준 향상, 전문 인력의 공급, 데이터거래소 등 빅데이터 산업 생태계를 조성하고자 한다. 빅데이터 발전 장애 요소를 없애고, 기술 혁신 수준을 제고하는 정책 추진 방향에 위 법제도들이 역할을 할 것으로 예상된다.

제3장. 디지털 무역과 국경 간 데이터 이동

이성엽
(고려대 기술경영전문대학원 교수)

ICT기술의 급속한 발달과 모바일기기의 확산으로 기존의 전자상거래와는 다른 디지털 무역이 활성화되고 있다. 디지털 무역은 데이터 경제 시대를 맞이하여 그 자체로 경쟁력 요소가 된 데이터의 이동까지 포함하는 개념으로 확대되었으며, 이를 둘러싼 첨예한 문제들이 디지털 무역에 관한 다양한 협정의 주요 의제가 되고 있다. 이 장에서는 국경 간 데이터 이동을 비롯한 데이터 의제를 다루는 디지털 무역 관련 국제 협정의 주요 내용을 살펴보고, 미-EU 간 데이터 이동 현안, 미-중 패권 전쟁에서 확산된 데이터 전쟁 등 주요 글로벌 현안을 통해 한국이 취해야 할 방향에 대해 고찰하고자 한다.

1. 디지털 무역의 개념과 내용

이제 세계 최대 전자상거래 업체인 아마존을 통해 외국 원서를 클릭 몇 번으로 집에서 받아볼 수 있는 세상이다. 또한 외국의 영화, 드라마 등 디지털 콘텐츠도 넷플릭스 등 글로벌 플랫폼을 통해 안방에서 쉽게 즐기는 시대가 되었다. 이처럼 ICT 기술의 급속한 발달과 모바일 기기의 확산으로 인터넷을 기반으로 하는 글로벌 전자상거래가 확대되면서 기존의 전통적인 상품 중심의 무역 거래와는 다른 디지털 무역이 활성화되고 있다.

디지털 무역(digital trade)이란 결국 상품과 서비스의 주문, 생산 및 배송에 있어 인터넷과 인터넷 기반 기술이 중요한 역할을 하는 국제 거래라고 할 수 있다. 미국 무역대표부(USTR)에 따르면 디지털 무역이란 종래 주로 전자상거래를 지칭하는 개념이었지만 이제는 데이터의 이동까지 포함하는 개념으로 확대되고 있다.[1]

이는 최근 데이터 경제 시대를 맞이하여 데이터가 단순히 상품 등에 수반하는 것이 아니라 그 자체로 경쟁력 요소가 되었기 때문이다.

디지털 통상에 관한 다양한 협정들이 체결되었지만 그 협정들의 주요 의제들은 대동소이한데, 개인정보보호, 전자인증 및 전자서명, 국경 간 데이터 이동, 데이터 현지화(data localization) 금지, 공공데이터 개방, 소스코드 공개 금지 등이 그에 해당한다.

첫째, 개인정보보호는 개인정보보호를 보장하는 법률적 프레임워크를 채택, 유지해야 할 의무를 부여하는 것이고. 둘째, 전자인증 및 전자서명은 법령에서 규정하는 경우를 예외로 하고 당사국들이 어떠한 서명이 전자적 형식으로 이루어졌다는 이유만으로 그 서명의 법률효력을 부인해서는 안 된다는 것이다. 셋째, 국경 간 데이터 이동은 당사국들이 영업활동에 따른 전자적 수단에 의한 정보의 국경 간 이동을 금지해서는 안 된다는 원칙을 규정한다. 넷째, 데이터 현지화 금지는 어떠한 당사국도 그 국가에서 영업활동을 하는 조건으로 그 국가의 컴퓨팅 시설을 이용하거나 그 국가에 컴퓨팅 시설을 둘 것을 요구할 수 없다는 내용이다. 다섯째, 공공데이터 개방은 어떠한 당사자가 데이터를 포함하여 정부가 보유한 정보를 공공에 개방하는 경우, 기계가 읽을 수 있는 오픈된 포맷으로 검색, 회수, 사용, 재사용, 재분배를 보장해야 한다는 규정이다. 여섯째, 소스코드 공개 금지는 어떠한 당사국도 소프트웨어의 수입, 배급, 판매 또는 이용의 조건으로 다른 당사자가 보유한 소프트웨어의 소스코드에 대한 이동이나 접근을 요구할 수 없다는 규정이다.

이제 데이터 이동의 자유화와 디지털 무역 자유화의 흐름은 데이터 이동 제한과 대응될 수 있는 정도로 디지털 무역에서 데이터는 중요한 변수가 되고 있다. 그동안 미국을 비롯한 주요 선진국은 디지털 무역이 활성화됨에 따라 이와 관련된 국제규범의 제정 노력을 기울여 왔다. 아래에서는 디지털 무역협정 상 데이터 이동 규범의 내용과 시사점에 대해 살펴보고자 한다.

2. 디지털 무역 협정과 데이터 규범

세계 무역 규범 역할을 하고 있는 WTO 체제에서는 각국 간 이해관계의 차이로 디지털 무역을 규율할 수 있는 협정이 어렵다고 보고, 양자 간 FTA나 지역협정을 통해 디지털 무역에 대한 규범을 제정, 운용하고 있다. 데이터 이동 자유화를 담고 있는 대표적 디지털 무역 지역 협정이 일본, 캐나다, 호주, 싱가포르 등 태평양 연안의 11개국이 참여해 2018년 1월 23일 체결된 '아시아태평양동반자협정(Comprehensive and Progressive Agreement for

1 https://ustr.gov/about-us/policy-offices/press-office/fact-sheets/2017/march/key-barriers-digital-trade (접속일: 2021.11.2)

TransPacific Partnership, CPTPP)'과 미국, 멕시코, 캐나다 사이에 2018년 11월 30일 체결된 '미국멕시코캐나다협정(United States-Mexico-Canada Agreement, USMCA)'이다.

CPTPP는 국경 간 데이터 이동 자유화, 컴퓨터 설비의 현지화 요구 금지, 소스코드 공개요구 금지 등 3개 규범이 의무 조항으로 들어가 있다. 국경 간 개인정보의 이동과 관련해서 보면, 각 당사국들이 국경 간 데이터 이동에 관한 자체적인 국내 규제를 마련할 필요성을 인정하고 있다. 다만, 당사국들이 정당한 공공정책의 목적(legitimate public policy objective) 달성을 위하여 개인정보의 국경 간 이동을 금지하는 것은 허용되나, 자의적 또는 부당한 차별, 위장된 무역 제한 조치로써 이를 금지하여서는 안 된다고 규정하고 있다. 한편 CPTPP에서는 개인정보보호 조항을 단순한 선언적 규정이 아닌 의무규정으로 두어 개인정보에 대한 보호를 강화하고 있다.

USMCA는 디지털 무역의 장을 별도로 둔 현재까지 가장 자유화 수준이 높은 국제규범으로 미국의 디지털 통상에 관한 기본적 입장을 담고 있다. 당사국은 특정한 행위가 대상자의 사업수행을 위한 활동인 경우 개인정보를 포함하여 전자적 수단에 의한 데이터의 국경 간 이동을 금지하거나 제한하여서는 안 된다고 규정하고 있다. 다만, 당사국이 정당한 공공정책의 목적 달성을 위하여 데이터 이동을 제한하는 것은 가능하다. 그러나 해당 조치가 자의적이거나 부당한 차별의 수단 또는 무역에 대한 제한을 위장하는 수단 등과 같은 의미로 행해져서는 안 되며, 또한 해당 조치가 정당한 공공정책의 목적을 달성하기 위하여 필요한 정도를 초과하여 데이터 이동을 제한하여서는 안 된다고 규정하고 있다.

그 외 미국과 일본 사이에 2019년 12월 체결된 미-일 디지털 무역협정과 싱가포르, 뉴질랜드 및 칠레 사이에 2020년 6월 12일 체결된 '디지털경제동반자협정(The Digital Economy Partnership Agreement, DEPA)'도 데이터 이동 자유화를 선언하고 있다. 미-일 디지털 무역협정은 기본적으로 USMCA와 유사한 수준의 혁신적인 규범을 담고 있으며, CPTPP보다 강화된 디지털 무역 자유화 조항이 포함되어 'CPTPP 플러스' 수준인 것으로 평가된다. CPTPP에서 다루지 않는 알고리즘, 금융데이터, 조세, 암호화 기술, 온라인 플랫폼, 공공데이터에 관한 무역규범을 포함하여 현 무역협정 중 가장 높은 수준의 디지털 무역 자유화 협정이다. DEPA는 FTA 내 디지털 챕터 형식(CPTPP, USMCA 등)이 아닌 디지털 분야만 다룬 협정[2]으로서 디지털 제품의 비차별 대우, 암호를 사용하는 ICT 제품에 대한 기술규제 금지, 전자적 수단에 의한 국경 간 데이터 이동 자유화, 컴퓨팅 설비 현지화 금지 등의 디지털 규범을 담고 있다.

2 디지털 무역협정은 기존 FTA내에 디지털 챕터로 포함되는 방식과 FTA와 별개로 이루어지는 경우가 있다.

이러한 데이터 이동 자유화의 흐름과 달리 중국이 주도하고 한국, 일본, 호주, 싱가포르, 뉴질랜드, 베트남 등 15개국이 참여해 2020년 11월 15일 체결된 '역내포괄적경제동반자협정(Regional Comprehensive Economic Partnership, RCEP)'은 무역규모, 역내총생산(GDP), 인구 측면에서 전 세계의 약 30%를 차지하는 세계 최대 FTA로 USMCA·CPTPP보다 규모가 큰 협정이다. 다만, 전자상거래 챕터(Chapter 12, Electronic Commerce)는 CPTPP보다 자유화 수준이 낮은 것으로 평가된다. 특히, 주요 규정인 국경 간 데이터 이동과 현지화 금지와 관련하여 '정당한 공공정책의 목적 달성' 예외 규정 외에도 안보 예외 규정을 두고 있으며, '필요하다고 간주하는 모든 조치(any measure that it considers necessary)'를 포괄적으로 도입하고 있어 당사국의 규제 자율권을 더욱 확대하고 있다. 즉,공공정책의 정당한 목적을 달성하기 위해 필요하다고 인정하는 경우와 안보상 중대한 이익을 보호하는 데 필요하다고 인정하는 경우를 제외하고는 당사자의 영역에서 사업을 수행하기 위한 조건으로 ① 컴퓨터 설비를 자국의 영역 내에 설치할 것(data localization)을 요구하거나 ② 전자적 수단에 의한 데이터의 국경 간 이동(free flow of data)을 방해하면 안 된다는 의무를 포함하고 있다.

또한 중국의 경우 2018년 '네트워크안전법(网络安全法)'에 자국 내 개인정보와 중요 데이터가 해외로 이동되지 못하도록 서버의 중국 내 설치를 의무화하고, 당국의 승인 없이는 데이터의 해외 반출을 금지하였다. EU는 '디지털 시장통합(Digital Single Market) 전략'으로 역내에서 데이터의 자유로운 이동을 허용하지만 2018년 5월 시행된 '일반개인정보보호법(General Data Protection Regulation, GDPR)'은 역외로의 데이터 이동을 일부 제한하고 있다.

미국과 EU 간 데이터 이동 관련 갈등도 지속되고 있다. '세이프하버협정(Safe Harbor Agreement)' 무효화[3]에 따라 체결된 프라이버시 쉴드 협정(EU-US Privacy Shield)에 대해서도 유럽사법재판소는 2020년 7월 무효화 판결을 내렸다.[4] 유럽사법재판소는 국가안보를 목적으로 미 당국의 개인정보 접근을 가능하게 하는 법률의 제도적 보호조치 및 정보 접근과 관련하여 정보주체에게 명확한 정보 제공 등의 권리 보장이 필요하며, 미 당국의 정보접근 및 감시 프로그램을 감독하는 옴부즈맨 제도가 GDPR 제58조 제2항에 비례한 독립성을 보

3 2012년 당시 법학도였던 막스 슈렘스(Marx Schrems)가 페이스북이 유럽 정보보호법을 위반하고 개인정보를 무단 수집하여 미국으로 이동했다면서, 미-EU 간 데이터 이동 협정인 세이프하버협정에 대한 무효 소송을 제기(일명 슈렘스사건)한 데 대해 유럽사법재판소가 무효 판결을 내림

4 2015년 슈렘스가 세이프하버협정을 대체하여 체결된 프라이버시 쉴드 협정이 EU의 GDPR 수준에 미치지 못한다고 지적하면서 후속 소송을 제기했고, 유럽사법재판소는 이를 인정하여 프라이버시 쉴드를 무효화함

장하고 있지 못하며, 미 당국을 대상으로 EU 정보주체가 제기할 수 있는 권리 이행 방안이나 법적 구제 방안이 미흡하다고 판단했다. 특히, 미국의 '외국정보감시법(Foreign Intelligence Surveillance Act, FISA)' 제702조 및 대통령 정책 지침 제28조 등 미 당국의 개인정보 접근과 관련된 법률이 정보주체의 실질적 권리 이행 및 개인정보의 적정한 보호 수준을 갖추지 못한다고 판결했다.

한편 최근 미국과 중국 간 기술패권 전쟁이 데이터 전쟁으로 확산되고 있다. 미국은 국가안보 위협과 자국민 데이터의 중국 이동을 이유로 미국 내 틱톡 및 위챗 사용 금지명령과 틱톡의 모기업이자 AI 기술기업인 바이트댄스를 대상으로 90일 이내에 미국 내 사업 매각·퇴진과 미국 내 사용자 데이터 이동 명령을 내렸다. 나아가 2020년 4월 '클린 네트워크 프로그램(Clean Network Program, CNP)'을 발표해 미국 등이 사용하는 인터넷 인프라에서 중국의 통신회사, 앱, 클라우드, 해저케이블 퇴출을 추진하고 있다.

이에 대응해 중국은 '글로벌 데이터안전 이니셔티브'를 추진하고 있다. 이는 데이터 안보를 포괄적이고 객관적이며 증거에 기초한 방식으로 취급하고, 정보통신 기술과 서비스를 위해 개방적이고 안전하며 안정적인 공급망을 유지하며, 각국 정부에 데이터를 다루는 방식에 있어서 다른 나라들의 주권을 존중할 것을 내용으로 하고 있다.

3. 시사점

한국의 경우 데이터 이동에 관한 한 데이터 보호주의(data protectionism)에 가까운 것으로 평가된다. 데이터 보호주의란 개인정보보호와 자국 IT산업 육성을 위하여 자국민 데이터의 해외 이동을 제한하는 입장이다. 주요 데이터 및 서버를 현지 국경 안에 두도록 하는 데이터 현지화도 보호주의의 한 유형이라고 할 수 있다. 금융 및 의료데이터 이동 금지, 동의 기반의 데이터 이동, 5000분의 1 지도 반출 불허 등이 사례이다. 서버 현지화 등의 법안이 시도되었으나 아직 입법이 이루어지지는 않고 있다.

최근 데이터 경제는 플랫폼 경제로 수렴 중이며, 국내외 주요 플랫폼의 경쟁력의 원천은 데이터 특히, 개인정보의 확보 여부이다. 이에 EU, 중국은 미국의 플랫폼 견제를 위해 강력한 데이터 보호 정책과 규제를 시행 중이다. 미국과 같이 디지털 환경과 데이터에 있어 절대적 우위를 가진 국가들은 높은 협상력을 기초로 자국에게 유리하도록 다른 국가에 데이터 이동 자유화를 강력하게 요구하고 있다.

이런 상황에서 한국의 선택은 어떤 것이어야 할까. 한국의 빅테크 기업들이 이미 확보한 데이터 및 앞으로 데이터를 확보할 수 있는 기술적 능력이 어느 정도인지, 디지털 무역협정 체결을 위한 협상대상 국가의 기업들이 한국 기업들과 비교하여 어느 정도의 역량을 가지고 있는지 등을 고려하여 이중적인 전략을 취할 필요가 있다.

그러나 원칙적으로 개방체제와 무역을 통해 성장한 한국의 경제전략은 디지털 시대에도 여전히 유효하다고 봐야 한다. 이 점에서 우선 글로벌 스탠더드에 부합하는 데이터 이동 자유화 원칙이 중요하다. 글로벌 스탠더드 준수와 관련 사전동의 원칙 고수, 준수가 어려운 기술적 보호조치 기준, 과다한 형사처벌 등 비현실적인 개인정보 규제를 개선할 필요가 있다. 다만, 국내 개인정보보호 역시 중요하기 때문에 동의 이외의 데이터 국외 이동 수단으로 적정성 평가나 적절한 보호조치에 의한 데이터 이동 방안도 법제화할 필요가 있다.

데이터를 둘러싼 국가 간 갈등의 이면에는 기술, 산업, 안보를 포함하는 국가 전체의 생존과 번영을 위한 전략이 있다. 각국은 어떤 경우에도 국가안보, 자국민 및 자국 산업의 보호라는 목표를 포기할 수 없기 때문에 갈등은 필연적일 수밖에 없다. 더구나 미중 G2의 디지털 전쟁에서 한국과 같이 양국에 안보, 경제의 상당 부분을 의존하는 경우, 이런 목표의 성취에는 많은 고민이 따른다. 상황에 따른 유연한 대응이 필요하지만 그래도 제1순위는 자국민의 데이터와 자국 산업의 보호라고 할 것이다. 단, 통상마찰 등의 관점에서 국익을 고려한 데이터 정책과 규제 혁신 노력도 병행해야 할 것이다. 또한 산업통상자원부 등 통상 전담 부처는 물론 과학기술정보통신부, 개인정보보호위원회 등 ICT 및 데이터 소관 부처 간 협력 역시 중요하다고 할 것이다.

PART



데이터산업 시장 현황

제1장.
국내 데이터산업 시장 현황

제2장.
해외 데이터산업 시장 현황

제1장. 국내 데이터산업 시장 현황

한국데이터산업진흥원 산업기획팀

국내 데이터산업 현황은 2020 데이터산업 현황조사 보고서[1] (과학기술정보통신부·한국데이터산업진흥원, 2021.3)를 요약하여 작성하였다. 2019년 데이터산업 분류에 따라 데이터 처리 및 관리 솔루션 개발·공급업, 데이터 구축 및 컨설팅 서비스업, 데이터 판매 및 제공 서비스업으로 구분하였다. 본 지에 수록된 시장 규모는 데이터 비즈니스를 영위하는 사업체의 데이터 관련 매출기반의 추정치이다. 2019년 이전까지 시장 규모는 확정치인 반면, 2020년 시장 규모는 잠정치이기 때문에 향후 공표 예정인 확정치 통계 결과와 상이할 수 있음을 미리 밝힌다.

1. 국내 데이터산업 시장 규모

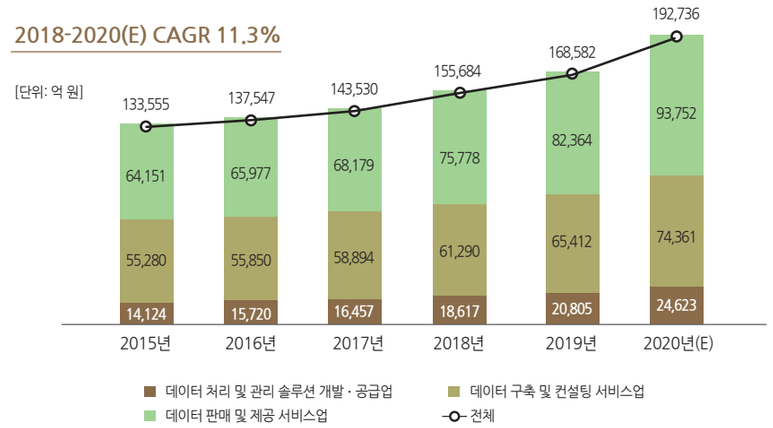
2019년 데이터산업[2] 시장 규모는 전년 대비 8.3% 성장한 16조 8,582억원이며, 2020년에는 19조 2,736억원 규모로 성장할 것으로 나타났다. 2018년부터 2020년(잠정치)까지의 3년간 연평균 성장률(Compound Annual Growth Rate, CAGR)은 11.3%로 나타나면서 지속적인 성장세를 이어갈 것으로 예상하였다.

2020년 데이터산업은 총 3개 대분류와 10개 중분류로 구성된다. 부문별 시장 규모 잠정치는 데이터 판매 및 제공 서비스업 시장이 9조 3,752억원으로 가장 높은 비중을 차지하며, 다음으로 데이터 구축 및 컨설팅 서비스업이 7조 4,361억원, 데이터 처리 및 관리 솔루션 개발·공급업이 2조 4,623억원으로 예상된다.

1 「2020년 데이터산업 현황조사」는 국가승인통계 127004호로 연간 발간되며, 해당 보고서는 2020년 11~12월에 조사된 결과를 수록하여 2021년 3월에 발표하였음

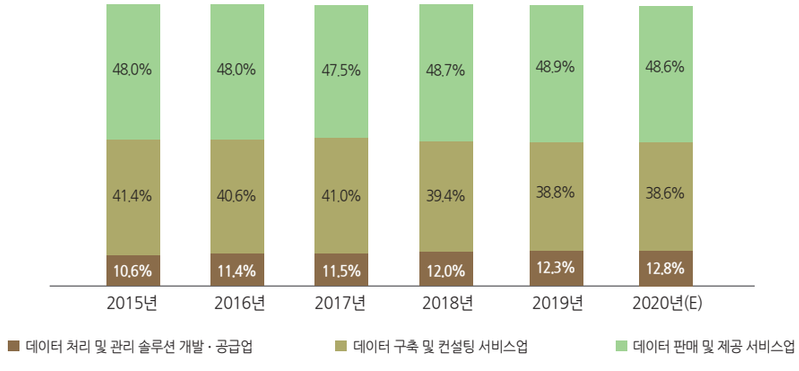
2 이 장에서 데이터산업이란, 데이터의 생산·수집·처리·분석·유통·활용 등을 통해 가치를 창출하는 상품 및 서비스를 생산·제공하는 산업으로 정의

[그림 3-1-1] 국내 데이터산업 시장 규모(2015~2020년(E))



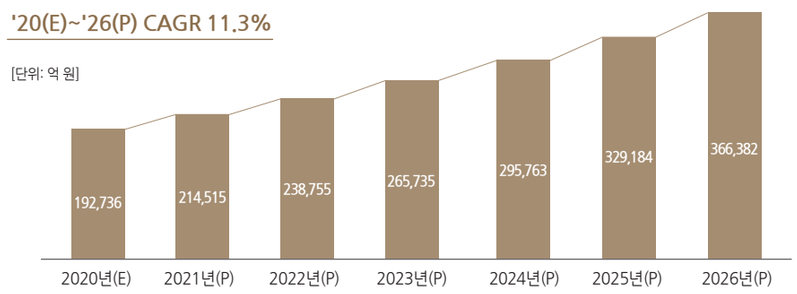
*출처: 과학기술정보통신부·한국데이터산업진흥원, 「2020 데이터산업 현황조사」 보고서 2021.3.(이하 표·그림 모두 동일)

[그림 3-1-2] 국내 데이터산업 부문별 시장 규모 비중(2015~2020년(E))



국내 데이터산업 시장이 지난 3개년 연평균 성장률인 11.3%로 지속적으로 성장한다고 가정 한다면, 2026년(추정치)에 36조원을 넘어설 것으로 전망된다.

[그림 3-1-3] 국내 데이터산업 시장 전망(2020(E)~2026년(P))



E: 잠정치, P: 추정치
* 잠정치(E): 조사 결과로 산출한 수치
추정치(P): 조사 결과를 바탕으로 전망을 추정한 수치

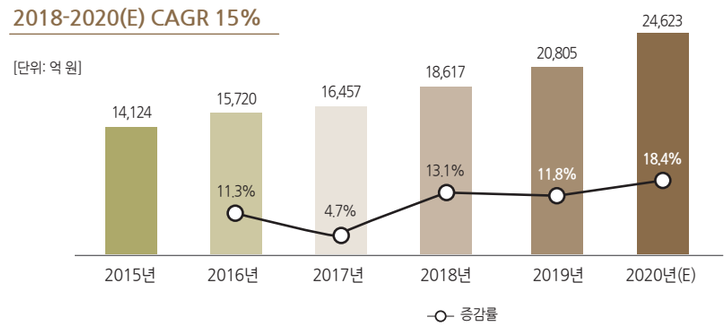
2. 국내 데이터산업 부문별 시장 규모

가. 데이터 처리 및 관리 솔루션 개발·공급업 시장

데이터 솔루션 분류는 데이터 처리 및 관리 솔루션 개발·공급업을 의미한다. 여기에는 데이터 수집·연계 솔루션, 데이터베이스 관리 시스템 솔루션, 데이터 분석 솔루션, 데이터 관리 솔루션, 데이터 보안 솔루션, 빅데이터 통합 플랫폼 솔루션을 포함하여 6개 중분류로 구분된다. 이들은 주로 솔루션 제품을 판매하는 비즈니스를 의미하고 라이선스, 개발 및 커스터마이징, 유지보수 등을 통해 매출을 올리고 있다.

2020년 데이터 처리 및 관리 솔루션 개발·공급업 부문의 시장 규모 잠정치는 전년 대비 18.4% 성장한 2조 4,623억원으로 전망되었으며, 2018년부터 3년간 연평균성장률(CAGR)은 15%로 예상되었다.

[그림 3-1-4] 국내 데이터 처리 및 관리 솔루션 개발·공급업 시장 규모(2015~2020년(E))



E: 잠정치, P: 추정치

2020년도 부문별 예상 매출액은 데이터베이스 관리 시스템 솔루션 개발·공급업이 약 8,835억원, 데이터 관리 솔루션 개발·공급업이 약 5,555억원 순으로 높은 비중을 차지할 것으로 예상된다.

[표 3-1-1] 국내 데이터 처리 및 관리 솔루션 개발·공급업 중분류별 시장 규모(2016-2020년(E))

[단위: 억 원]

구분	2016년	2017년	2018년	2019년	2020년(E)	증감률 '18~'19	CAGR `18~`20(E)
데이터 수집·연계 솔루션 개발·공급업	1,345	1,393	1,622	1,871	2,256	15.4%	17.9%
데이터베이스 관리 시스템 솔루션 개발·공급업	6,148	6,121	6,775	7,510	8,835	10.8%	14.2%
데이터 분석 솔루션 개발·공급업	1,249	1,325	1,782	2,014	2,779	13.1%	24.9%
데이터 관리 솔루션 개발·공급업	5,367	4,628	4,972	5,203	5,555	4.7%	5.7%
데이터 보안 솔루션 개발·공급업		1,213	1,517	1,975	2,495	30.2%	28.3%
빅데이터 통합 플랫폼 솔루션 개발·공급업	1,611	1,776	1,949	2,231	2,704	14.5%	17.8%
전체	**15,720**	**16,457**	**18,617**	**20,805**	**24,623**	**11.8%**	**15.0%**

나. 데이터 구축 및 컨설팅 서비스업 시장

데이터 구축 및 컨설팅 서비스업은 데이터 구축·가공 서비스업과 데이터 관련 컨설팅 서비스업으로 구분된다. 데이터 구축·가공 서비스업은 데이터베이스 설계·구축 서비스, 데이터 이행 서비스, 데이터 가공 서비스를 포함하고, 데이터 관련 컨설팅 서비스업은 데이터 설계 컨설팅, 데이터 품질 컨설팅, 데이터베이스 성능개선 컨설팅, 데이터 거버넌스 컨설팅, 데이터 분석·활용 컨설팅을 의미한다. 비즈니스 매출은 주로 데이터 구축·개발, 유지보수·운영관리, 컨설팅을 통해 발생한다.

2020년 잠정적인 데이터 구축 및 컨설팅 서비스업 시장 규모는 7조 4,361억원으로 전년대비 13.7% 성장할 것으로 예상된다. 이 중 데이터 구축·가공 서비스업 시장 규모는 7조 166억원으로 전체 데이터 구축 및 컨설팅 서비스업 시장의 94.4% 수준, 데이터 관련 컨설팅 서비스업은 4,195억원인 5.6% 수준으로 예상되었다. 데이터 구축 및 컨설팅 서비스업 시장의 2018년부터 3년간 연평균성장률(CAGR)은 10.1%로 조사되었다.

[그림 3-1-5] 국내 데이터 구축 및 컨설팅 서비스업 시장 규모(2015~2020년(E))

[표 3-1-2] 국내 데이터 구축 및 컨설팅 서비스업 중분류별 시장 규모(2016~2020년(E))

[단위: 억 원]

구분	2016년	2017년	2018년	2019년	2020년(E)	증감률 '18~'19	CAGR `18~`20(E)
데이터 구축·가공 서비스업	54,571	57,207	58,993	62,223	70,166	5.5%	9.1%
데이터 관련 컨설팅 서비스업	1,279	1,687	2,297	3,189	4,195	38.9%	35.1%
전체	**55,850**	**58,894**	**61,290**	**65,412**	**74,361**	**6.7%**	**10.1%**

다. 데이터 판매 및 제공 서비스업 시장

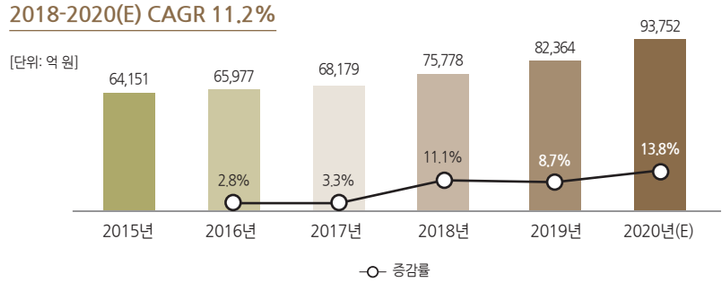
2020년 잠정적인 데이터 판매 및 제공 서비스업 시장 규모는 전체 9조 3,752억 원으로 전년대비 13.8% 성장할 것으로 예상된다. 2018년부터 2020년 예상 매출까지 3년간 연평균성장률(CAGR)은 11.2%로 나타난다.

데이터 판매 및 제공 서비스업은 원천 데이터와 데이터베이스를 수요자 맞춤형으로 가공·활용·분석하여 제공하는 비즈니스이다. 주로 데이터 이용료·수수료 등의 직접매출과 광고료 등의 간접매출로 수익을 창출한다.

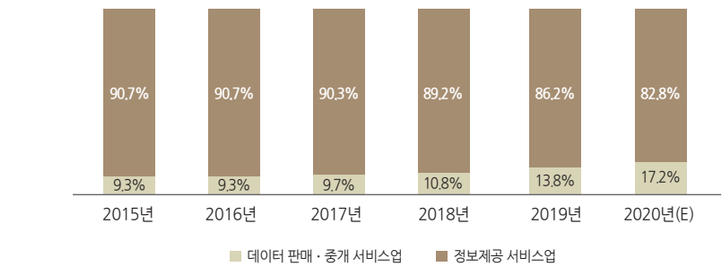
세부 분류로 데이터 판매 서비스와 데이터 중개 서비스, 분석 데이터 제공 서비스를 포함하는 데이터 판매·중개 서비스업과 포털·정보 매개 서비스와 정보제공 서비스를 포함하는 정보제공 서비스업으로 나누어진다.

2020년 세부 부문별로는 데이터 판매·중개 서비스업 시장은 1조 6,080억원, 정보제공 서비스업 시장은 7조 7,672억원으로 예상된다. 이 중 정보제공 서비스업의 비중이 82.8%를 차지하고 있다.

[그림 3-1-6] 국내 데이터 판매 및 제공 서비스업 시장 규모(2015~2020년(E))



[그림 3-1-7] 국내 데이터 판매 및 제공 서비스업 중분류별 시장 규모 비중(2015~2020년(E))



[표 3-1-3] 국내 데이터 판매 및 제공 서비스업 중분류별 시장 규모(2016~2020년(E))

[단위: 억 원]

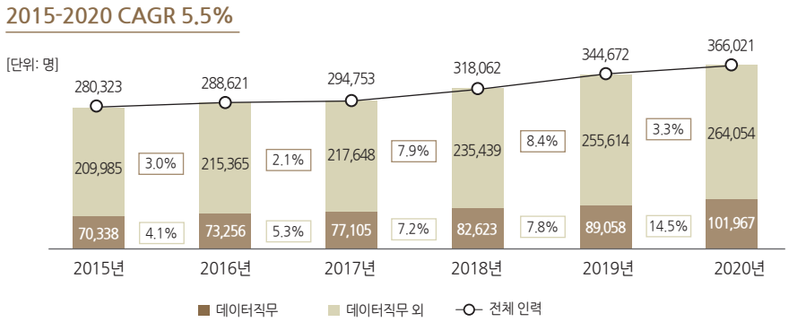
구분	2016년	2017년	2018년	2019년	2020년(E)	증감률 '18~'19	CAGR `18~`20(E)
데이터 판매·중개 서비스업	6,123	6,608	8,198	11,332	16,080	38.2%	40.1%
정보제공 서비스업	59,854	61,570	67,580	71,033	77,672	5.1%	7.2%
전체	**65,977**	**68,179**	**75,778**	**82,364**	**93,752**	**8.7%**	**11.2%**

3. 국내 데이터직무 인력 현황

가. 데이터산업 내 데이터직무 인력 현황

2020년 데이터산업에 종사하고 있는 전체 인력은 전년 대비 6.2% 증가한 36만 6,021명이다. 이 중 데이터직무 인력[3]은 10만 1,967명으로, 전체 데이터산업 종사자의 27.9%이며, 데이터산업 내 데이터직무 인력은 전년 대비 14.5% 증가하였다.

[그림 3-1-8] 국내 데이터산업 전체 인력 현황(2015~2020년)



2020년 부문별로 데이터직무 인력 현황과 추이를 살펴보면, 데이터 구축 및 컨설팅 서비스업 분야의 데이터직무 인력 수가 4만 8,644명으로 가장 많았다. 반면, 전년 대비 데이터직무 인력 비율이 가장 크게 증가한 분야는 데이터 처리 및 관리 솔루션 개발·공급업으로, 2019년 대비 28.3% 증가하였다.

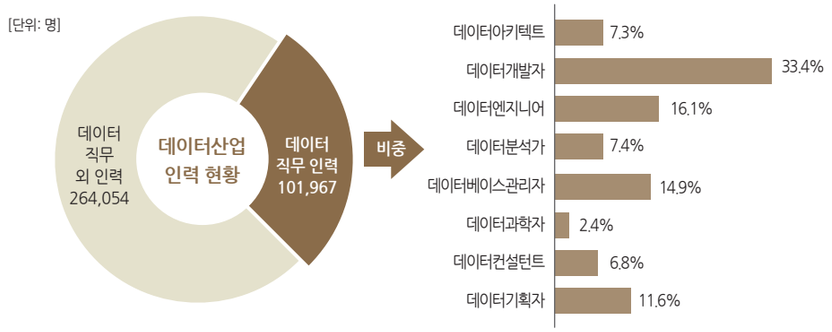
[표 3-1-4] 국내 데이터산업 부문별 데이터직무 인력 현황(2016~2020년)

[단위: 명]

구 분	2016년		2017년		2018년		2019년		2020년		증감률 '19~'20	CAGR '15~'20
	인력수	비중	인력수	비중	인력수	비중	인력수	비중	인력수	비중		
전체	**73,256**	**100.0%**	**77,105**	**100.0%**	**82,623**	**100.0%**	**89,058**	**100.0%**	**101,967**	**100.0%**	**14.5%**	**7.7%**
데이터 처리 및 관리 솔루션 개발·공급업	9,272	12.7%	10,291	13.3%	11,541	14.0%	13,467	15.1%	17,273	16.9%	28.3%	14.2%
데이터 구축 및 컨설팅 서비스업	35,404	48.3%	37,516	48.7%	40,197	48.7%	42,979	48.3%	48,644	47.7%	13.2%	7.2%
데이터 판매 및 제공 서비스업	28,580	39.0%	29,298	38.0%	30,885	37.4%	32,611	36.6%	36,050	35.4%	10.5%	5.9%

데이터산업 인력의 직무별 비중을 살펴보면, 데이터개발자가 3만 4,085명(33.4%)으로 가장 큰 비중을 차지하였고, 데이터엔지니어 1만 6,450명(16.1%), 데이터베이스관리자 1만 5,199명(14.9%) 순으로 나타났다.

[그림 3-1-9] 국내 데이터산업 인력 구성 및 데이터직무별 인력 비중(2020년)

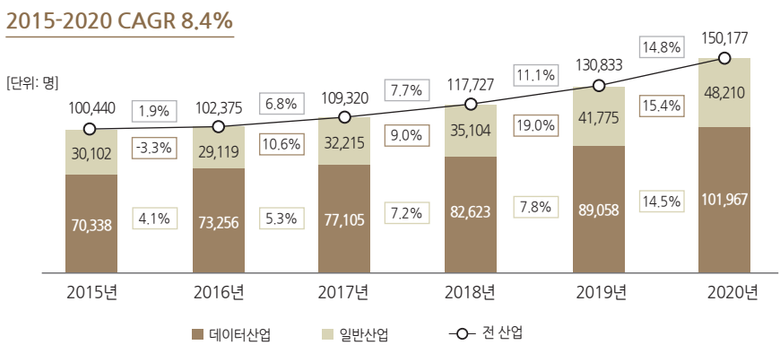


3 데이터개발자, 데이터엔지니어, 데이터분석가, 데이터베이스관리자, 데이터과학자, 데이터컨설턴트, 데이터아키텍트, 데이터기획자 8개 직무를 기준으로 함

나. 전 산업 내 데이터직무 인력 현황

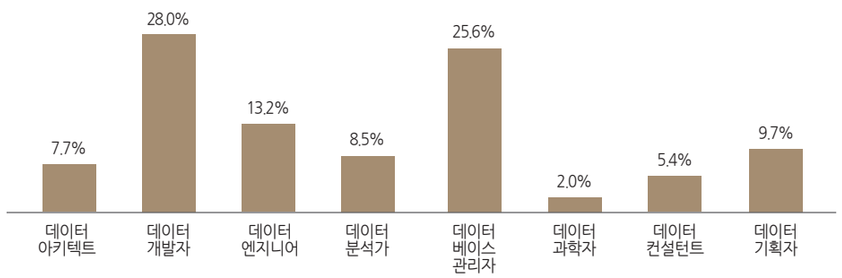
데이터산업과 일반산업[4]을 포함한 2020년 전 산업의 데이터직무 인력은 총 15만 177명으로, 전년 대비 14.8% 증가하였다. 데이터산업을 제외한 일반산업의 데이터직무 인력은 4만 8,210명으로 전년 대비 큰 폭(15.4%)으로 증가한 것으로 나타났다.

[그림 3-1-10] 국내 전 산업 데이터직무 인력 현황(2015~2020년)



전 산업의 데이터직무별 인력에서는 데이터개발자가 4만 1,976명(28.0%)으로 가장 높은 비중을 차지했고, 데이터베이스관리자 3만 8,440명(25.6%), 데이터엔지니어 1만 9,791명(13.2%) 순으로 나타났다. 데이터산업과 일반산업의 데이터인력 직무별 비중을 살펴보면, 데이터산업은 데이터개발자(33.4%)가 많은 반면, 일반산업은 데이터베이스관리자(48.2%)가 많은 것으로 나타났다.

[그림 3-1-11] 국내 전 산업 데이터직무별 인력 비중(2020년)



4 일반산업은 데이터산업 외 12개 분야(금융, 제조, 유통·서비스, 교육, 공공, 통신/미디어, 의료, 건설, 물류, 농림·축산광업, 숙박·음식점, 유틸리티(전기, 수도, 가스 등))의 100인 이상 사업체가 포함된 산업을 의미함

[표 3-1-5] 국내 전 산업 데이터직무별 인력 현황(2020년)

[단위: 명]

구 분	데이터산업		일반산업		전 산업	
	인력수	비중	인력수	비중	인력수	비중
데이터아키텍트	7,470	7.3%	4,075	8.5%	11,545	7.7%
데이터개발자	34,085	33.4%	7,891	16.4%	41,976	28.0%
데이터엔지니어	16,450	16.1%	3,341	6.9%	19,791	13.2%
데이터분석가	7,555	7.4%	5,182	10.7%	12,737	8.5%
데이터베이스관리자	15,199	14.9%	23,241	48.2%	38,440	25.6%
데이터과학자	2,421	2.4%	610	1.3%	3,031	2.0%
데이터컨설턴트	6,979	6.8%	1,133	2.4%	8,113	5.4%
데이터기획자	11,808	11.6%	2,736	5.7%	14,544	9.7%
합계	**101,967**	**100.0%**	**48,210**	**100.0%**	**150,177**	**100.0%**

4. 국내 데이터직무 인력 수요

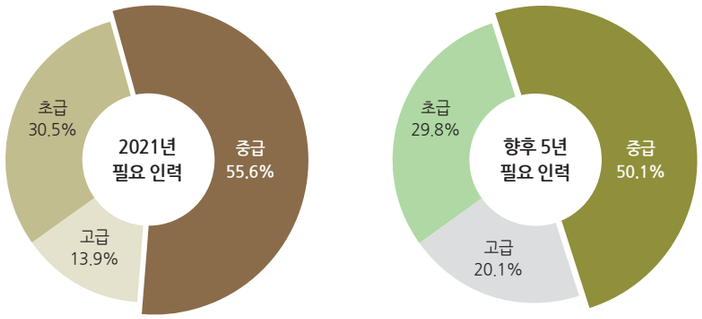
가. 데이터산업의 데이터직무 인력 수요 및 부족률

2025년까지 향후 5년 내 데이터산업의 데이터직무 필요 인력[5]은 총 1만 2,114명으로 조사되었다. 직무별 수요는 데이터개발자가 5,775명(47.7%)으로 가장 높았고, 데이터엔지니어 1,624명(13.4%), 데이터분석가 1,281명(10.6%) 순으로 나타났다.

향후 5년 내 데이터산업의 기술등급별 데이터직무 필요 인력 비중을 살펴보면 현재와 마찬가지로 중급 인력 6,065명(50.1%)에 대한 수요가 가장 높게 나타났다.

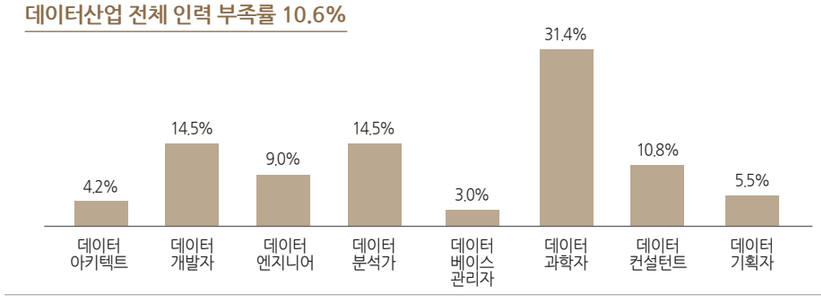
5 필요 인력은 기업에서 현재 인력보다 추가로 더 필요로 하는 인력 수를 의미함. 즉, 현재 부족한 인력 수를 의미하며, 채용계획 인력수는 아님

[그림 3-1-12] 국내 데이터산업의 향후 5년 내 기술등급별 데이터직무 필요 인력 비중



데이터산업 내 데이터직무 평균 부족률[6]은 10.6%이다. 이 중 데이터과학자의 부족률이 31.4%로 가장 높게 나타났다. 이어서 데이터개발자와 데이터분석가(각 14.5%), 데이터컨설턴트(10.8%) 순으로 나타났다.

[그림 3-1-13] 국내 데이터산업의 향후 5년 내 데이터직무 인력 부족률

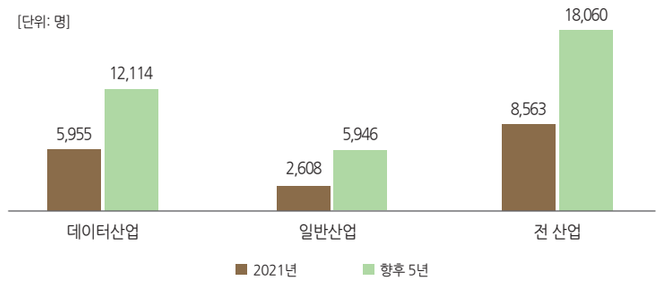


나. 전 산업의 필요 인력 및 부족률

향후 5년 내 일반산업을 포함한 전 산업에서 필요한 데이터직무 인력은 총 1만 8,060명이다. 일반산업은 데이터산업에 비하여 데이터직무에 대한 필요가 비교적으로 낮았다. 이 중 데이터개발자가 6,752명(37.4%)으로 가장 크게 나타났고, 데이터베이스관리자 2,880명(15.9%), 데이터분석가 2,163명(12.0%), 데이터엔지니어 2,141명(11.9%) 순으로 나타났다.

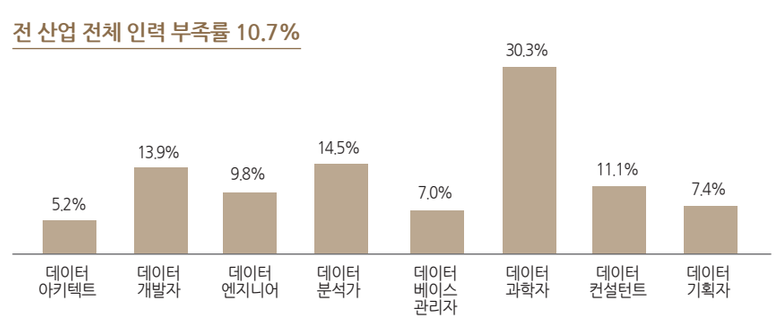
6 부족률(%)={필요 인력 수÷(현재 인력 수+필요 인력 수)}×100

[그림 3-1-14] 국내 데이터산업/일반산업/전 산업의 향후 5년 내 데이터직무 필요 인력



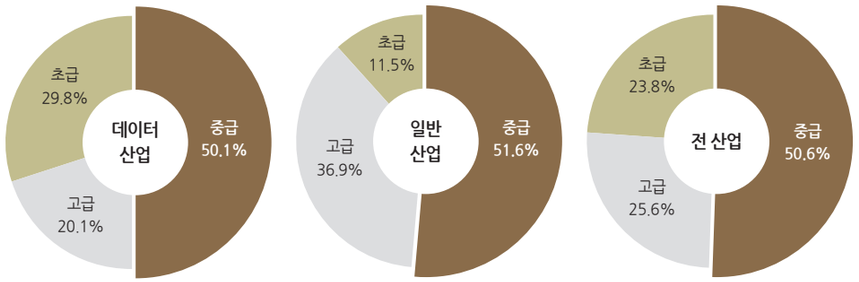
향후 5년 내 일반산업을 포함한 전 산업 내 데이터직무별 인력 부족률은 평균 10.7% 수준이며, 데이터과학자 부족률이 30.3%로 가장 높게 나타났다. 데이터분석가(14.5%), 데이터개발자(13.9%), 데이터컨설턴트(11.1%) 직무가 전 산업 평균보다 높은 부족률을 보였다.

[그림 3-1-15] 국내 전 산업의 향후 5년 내 데이터직무 인력 부족률



향후 5년 내 일반산업을 포함한 전 산업에서 추가로 필요한 기술등급별 인력은 중급이 9,133명(50.6%)으로 가장 높게 나타났고 고급 4,628명(25.6%), 초급 4,299명(23.8%) 순으로 나타났다.

[그림 3-1-16] 국내 전 산업의 향후 5년 내 기술등급별 데이터직무 필요 인력 비중



5. 마무리

한국은행 '기업경영분석'에 따르면 국내 전 산업의 연평균 매출 증감률(CAGR)은 2015년부터 2019년까지 3.6%로 나타났다. 동일 기간 동안 데이터산업과 유사한 정보통신서비스업은 6.5%, 정보통신기술산업은 2.6%, 서비스업은 연평균 -2.7% 수치를 보였다. 데이터산업은 6.0%로 비교적 높은 증가세를 보이고 있다.

[표 3-1-6] 국내 주요 산업별 시장 규모 추이(2015~2019년)

[단위: 억 원]

구분	2015년	2016년	2017년	2018년	2019년	CAGR '15~'19
전 산업*	35,889,830	36,676,988	39,914,871	41,516,088	41,411,615	3.6%
제조업*	16,962,408	16,645,144	18,218,781	19,007,562	18,547,032	2.3%
서비스업*	18,927,422	14,490,116	15,880,736	16,526,281	16,995,849	-2.7%
정보통신 서비스업*	1,955,693	2,082,409	2,183,098	2,393,394	2,517,847	6.5%
정보통신 기술산업*	5,627,440	5,638,768	6,060,007	6,471,342	6,243,279	2.6%
데이터산업 관련 업종*	672,866	742,494	833,333	895,587	956,257	9.2%
데이터산업**	133,555	137,547	143,530	155,684	168,582	6.0%

* 한국은행 경제통계시스템, 기업경영분석(2015-2019)
** 한국데이터산업진흥원, 「2020 데이터산업 현황조사」, 2021.3.16.

데이터산업은 디지털 경제를 이끄는 미래 핵심 동력으로 강조되고 있으며, AI 기술 및 새로운 ICT 인프라와 접목되어 향후에도 지속적인 성장세를 유지할 것으로 전망된다. 전 산업에서 데이터와 융합된 디지털 전환(Digital Transformation)과 이를 위한 데이터산업 생태계 강화 및 거버넌스 체계구축이 강조됨에 따라, 데이터직무 인력에 대한 수요 역시 증대되고 있다. 그러나 위 조사 결과와 같이 중고급 이상의 기술 수준을 보유한 데이터직무 인력에 대한 공급은 여전히 부족한 실정이다. 2021년 디지털 뉴딜 2.0 정책과 '데이터기본법' 제정을 기점으로, 수요에 대응할 수 있고 역량을 강화할 수 있는 데이터 전문인력 양성 기반 마련과 데이터산업에 대한 민관의 투자가 지속되기를 기대한다.

제2장.
해외 데이터산업 시장 현황[1]

나영민
(날리지리서치그룹 이사)

디지털 기업뿐만 아니라 제조 및 서비스 등 산업 대부분에서 스마트 기술을 이용한 자동화 시스템 구축이 확대되고 있으며 디지털 트랜스포메이션의 추진, IoT 및 데이터를 활용한 AI 시스템 확대가 빠르게 전개되고 있다. 디지털 기반 혁신이 전 세계적으로 확대되고 코로나19의 영향에 따라 데이터의 중요성은 그 어느 때보다 커지고 있다. 이에 따라 해외에서도 데이터 관리, 데이터 분석, 머신러닝 등 데이터 관련 산업도 놀라운 성장세를 이어가고 있다. 이러한 흐름 속에서 데이터 분석을 기반으로 하는 스타트업의 수가 증가하고 있으며 관련 시장 또한 빠르게 성장하고 있다. 본 데이터산업 백서에서는 글로벌 기반의 시장 자료를 제공하는 여러 기관의 데이터산업 관련 통계를 소개하고자 한다.

1. 디지털데이터 산업의 시장 현황

데이터 시장은 각 나라와 기관마다 그 범위와 정의가 다양하여 시장의 정의와 분류를 어떻게 하느냐에 따라 시장 규모가 다르게 나타나게 되는데 이 장에서는 글로벌 범위의 시장 자료를 제공하는 기관의 정의에 따른 시장 통계를 소개하고자 한다.

가. 디지털 데이터 시장

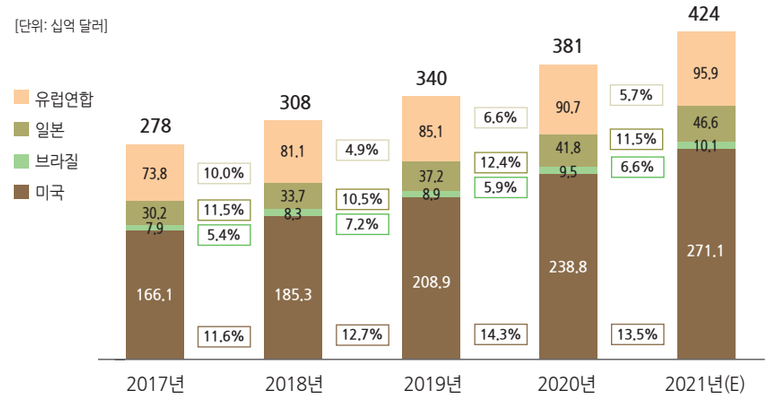
The European Data Market Monitoring Tool Report(IDC, 2020)에서는 데이터를 가공함으로써 제품 및 서비스로 재생산되는 '디지털 데이터 시장'으로 데이터 시장의 범위를 정의했다. 여기에는 이미지 데이터처럼 디지털 기술을 통해 수집, 저장, 가공, 전송되는 멀티미디어 데이터 부문도 포함되어 있다.

1 이 장은 「The European Data Market Monitoring Tool Report」(IDC, 2020.7.8.)의 데이터를 바탕으로 작성되었으며, 2021년도 추정치는 저자가 이 보고서의 2018~2020년 연평균성장률(CAGR)을 바탕으로 추정한 수치임을 밝힙니다.

디지털 데이터 시장 규모는 영국을 포함한 유럽연합, 미국, 일본, 브라질의 데이터 기반 기업의 매출액을 기반으로 산출되었으며 데이터 기업의 매출은 데이터 관련 제품의 총 가치에 해당하고, 수출을 포함하여 해당 국가에 기반을 둔 기업이 생산한 서비스의 매출을 기반으로 하고 있다.

미국의 디지털 데이터 시장은 세계에서 가장 큰 규모를 보이고 있으며, 시장의 성장세 또한 2021년 13.5%로 높게 나타났다. 미국은 시장 규모가 2017년 1,661억 달러, 2020년 2,388억 달러를 기록하였으며, 2021년에는 약 2,711억 달러의 시장 형성이 예상되는 등 지속적으로 두 자릿수 성장세를 보이고 있다.

[그림 3-2-1] 글로벌 디지털 데이터 시장 규모(2017~2021년(E))



* '17~'20: IDC, 「The European Data Market Monitoring Tool Report」, 2020.7.8.
'21(E): 위 보고서를 바탕으로 한 저자 추정치

미국은 전 세계 국가 중 가장 빠른 성장 속도를 보이며 유럽연합[2]의 3배 가까운 시장 규모를 형성하고 있다. 유럽연합의 데이터 시장 성장세 또한 미국과 견줄 정도로 빠른 편이었으나, 최근에는 5% 내외로 성장세가 완화되고 있는 모습을 보이고 있다. 2019년에 851억 달

2 이 장에서 유럽연합은 유럽연합 회원국 27개국과 영국을 포함: 오스트리아, 벨기에, 불가리아, 크로아티아, 키프로스, 체코, 덴마크, 에스토니아, 핀란드, 프랑스, 독일, 그리스, 헝가리, 아일랜드, 이탈리아, 라트비아, 리투아니아, 룩셈부르크, 몰타, 네덜란드, 폴란드, 포르투갈, 루마니아, 슬로바키아, 슬로베니아, 스페인, 스웨덴, 영국

러 수준을 기록하였으며 2020년에는 6.6% 성장한 907억 달러, 2021년에는 959억 달러에 이를 것으로 예상되고 있다.

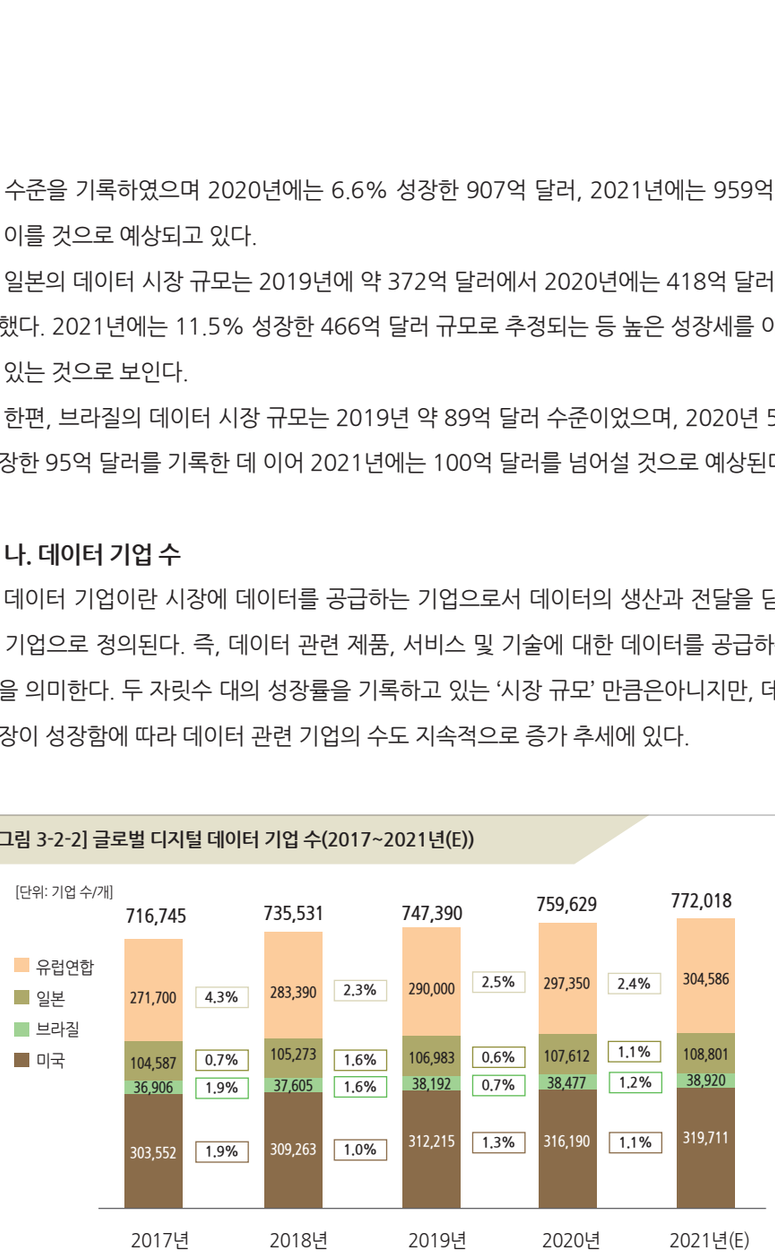
일본의 데이터 시장 규모는 2019년에 약 372억 달러에서 2020년에는 418억 달러로 성장했다. 2021년에는 11.5% 성장한 466억 달러 규모로 추정되는 등 높은 성장세를 이어가고 있는 것으로 보인다.

한편, 브라질의 데이터 시장 규모는 2019년 약 89억 달러 수준이었으며, 2020년 5.9% 성장한 95억 달러를 기록한 데 이어 2021년에는 100억 달러를 넘어설 것으로 예상된다.

나. 데이터 기업 수

데이터 기업이란 시장에 데이터를 공급하는 기업으로서 데이터의 생산과 전달을 담당하는 기업으로 정의된다. 즉, 데이터 관련 제품, 서비스 및 기술에 대한 데이터를 공급하는 기업을 의미한다. 두 자릿수 대의 성장률을 기록하고 있는 '시장 규모' 만큼은아니지만, 데이터 시장이 성장함에 따라 데이터 관련 기업의 수도 지속적으로 증가 추세에 있다.

[그림 3-2-2] 글로벌 디지털 데이터 기업 수(2017~2021년(E))



* '17~'20: IDC, 「The European Data Market Monitoring Tool Report」, 2020.7.8.
'21(E): 위 보고서를 바탕으로 한 저자 추정치

미국, 유럽연합, 일본, 브라질의 디지털 데이터 기업의 수는 데이터 종사자 수와 마찬가지로 매년 꾸준히 증가하고 있는데, 미국의 디지털 데이터 기업 수는 2020년 31만 6,190개에서 2021년에는 31만 9,711개로 1.1% 증가할 것으로 예측되었다. 유럽연합의 디지털 데이터 기업 수는 2020년 약 29만 7,000여 개였으나 2021년에는 30만 개 이상으로 증가할 것으로 전망된다. 미국과 유럽연합의 데이터 시장 규모가 크게 차이 나는 것과 비교하여 데이터 기업의 수는 큰 차이를 보이지 않는다. 이것은 유럽연합의 디지털 데이터 기업이 상대적으로 다양한 데다 중소기업 중심으로 나타나고 있어 상대적으로 바람직한 상태라고 할 수 있다. 미국의 경우 구글, 아마존, 마이크로소프트 등 소수의 글로벌 공룡 기업이 글로벌 데이터 시장을 주도하는 양상이 나타나고 있으며 그 외에도 대체로 대기업 위주의 데이터 기업들을 중심으로 시장이 움직이고 있다.

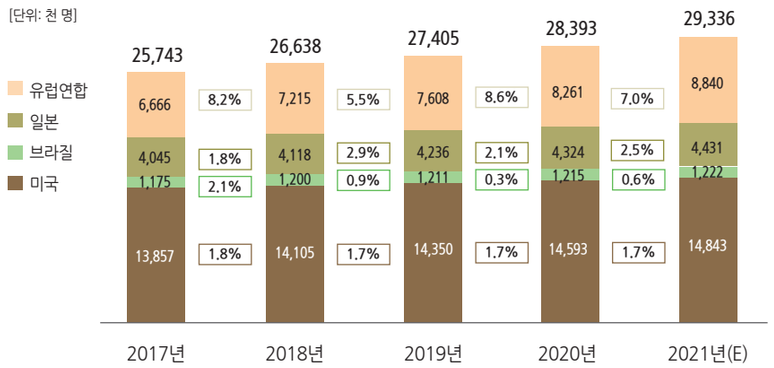
일본의 데이터 기업 수는 2020년 약 10만 7,612개였으나 2021년에는 1.1% 증가한 10만 8,801개로 집계됐다. 장기적인 경제 침체를 겪은 일본이지만 그 가운데에서도 데이터 기업 수가 꾸준히 증가하였다는 것은 데이터산업에 대한 일본 기업들의 수요와 기대가 반영되어 있기 때문이다.

한편, 브라질의 데이터 기업 수는 2020년 약 3만 8,477개에서 2021년 3만 8,920개로 소폭 증가하는 모습을 보였다. 브라질은 ICT 시장이 상대적으로 작고, ICT의 보급률 또한 매우 낮은 수준이라서 상대적으로 기업 수는 적지만 향후 지속적 성장이 기대되고 있다.

다. 데이터 전문 인력

데이터 전문 인력은 대용량 데이터를 수집, 저장, 관리 및 분석, 해석, 시각화할 수 있는 인력으로 정형 및 비정형 데이터의 사용에 능숙해야 하며, 막대한 양의 데이터를 핸들링할 수 있고, 새로운 데이터 분야 기술에 익숙한 능통자로 정의된다. 유럽연합의 경우 2017년에 666만 명 수준이었으나 2021년에는 884만 명 수준으로 예상되며 가장 높은 증가율을 보였다. 미국은 1,385만 명이었던 2017년에 비해 2021년에는 1,484만 명 수준으로 예상되며 꾸준히 증가하고 있다. 데이터 전문 인력은 전 세계적으로 부족한 상황으로 각국은 데이터 전문 인력을 양성하고 확보하는 데 높은 관심을 기울이고 있다.

[그림 3-2-3] 글로벌 데이터 전문 인력 수(2017~2021년(E))



* '17~'20: IDC, 「The European Data Market Monitoring Tool Report」, 2020.7.8.
'21(E): 위 보고서를 바탕으로 한 저자 추정치

유럽연합 기준 분야별 데이터 전문 인력은 전문 서비스와 도·소매 영역에서 비중이 높게 나타나고 있으며 인력 확대의 속도를 의미하는 성장률 측면에서는 정보 및 커뮤니케이션 분야와 전문 서비스, 건강 분야가 상대적으로 높게 나타났다.

[표 3-2-1] 유럽연합의 데이터 전문 인력 수(2017~2021년(E))

[단위: 천 명]

구분	2017년	2018년	2019년	2020년	2021년(E)	'19~'20 증감률	'20~'21 증감률(E)
건설	138	152	156	168	177	8.0%	5.0%
교육	489	530	566	616	664	8.9%	7.8%
금융	651	704	728	781	823	7.3%	5.3%
건강	529	573	610	666	718	9.1%	7.8%
정보 및 커뮤니케이션	750	835	880	967	1,040	9.9%	7.6%
채광, 제조	779	836	883	956	1,023	8.3%	7.0%
전문 서비스	1,392	1,511	1,621	1,777	1,927	9.6%	8.4%
공공 행정	417	447	469	507	540	8.1%	6.5%
도·소매	1,215	1,299	1,352	1,452	1,535	7.4%	5.7%
운송	207	221	233	252	270	8.2%	6.9%
전력·가스·수도	99	106	110	118	125	7.2%	5.8%
총계	6,666	7,215	7,608	8,261	8,841	8.6%	7.0%

* '17~'20: IDC, 「The European Data Market Monitoring Tool Report」, 2020.7.8.
'21(E): 위 보고서를 바탕으로 한 저자 추정치

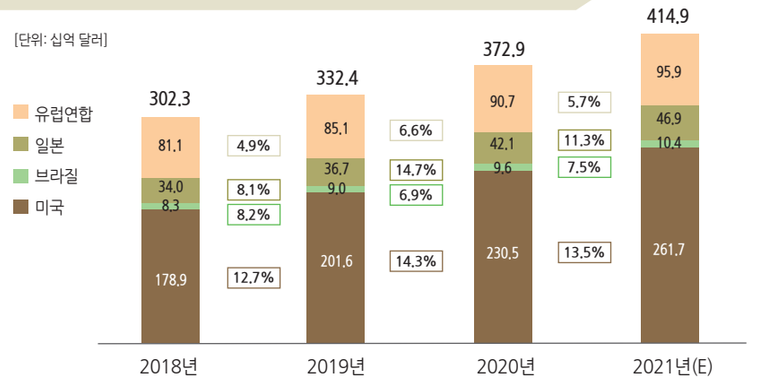
라. 데이터산업의 경제적 효과

(1) 직접 효과

데이터산업의 경제적 효과는 데이터 시장이 경제 생태계 전체에 미치는 전반적인 영향이 어느 정도인가를 평가하는 분석 방법론을 의미한다. 디지털 기술이 발달함에 따라 데이터의 수집, 저장, 처리, 배포, 분석, 정교화, 전달 및 활용이 쉽게 되었다. 이렇게 데이터의 활용이 쉬워짐에 따라 데이터산업과 시장이 형성되었고 이러한 시장은 직접 또는 간접적으로 경제 전체에 영향을 미치게 되었다.

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

[그림 3-2-4] 데이터산업의 경제적 효과: 직접효과(2018~2021년(E))



* '18~'20: IDC, 「The European Data Market Monitoring Tool Report」, 2020.7.8.
'21(E): 위 보고서를 바탕으로 한 저자 추정치

미국의 데이터 경제적 효과 중 직접 효과는 2018년 약 1,789억 달러를 기록한 이후 점차

증가하여 2020년에는 2,300억 달러를 넘어섰다. 즉, 데이터에 기반을 둔 제품 및 서비스의 생산에서 파생된 직접적 영향은 유럽연합을 비롯한 일본, 브라질보다 높게 나타났다. 이는 데이터 시장의 규모와 GDP에서의 비중이 상대적으로 크기 때문이다.

직접 효과는 데이터 사업 그 자체에 의해 형성되는데 이러한 정의에 따라 미국의 경우 데이터 생산에 매우 적극적이라는 것을 알 수 있다. 또한, 미국의 데이터 직접 효과는 데이터 시장 규모에서도 나타났듯이 유럽연합의 직접 효과의 거의 세 배에 가까운 수치인데, 이는 미국의 데이터산업이 상대적으로 발전되어 있고, 경제적인 효과의 확산 속도가 더 빠르다는 것을 시사한다.

유럽연합의 데이터 경제적 효과 중 직접 효과는 2018년에 약 811억 달러를 기록한 이후 점차 증가하여 2020년에는 약 907억 달러로 집계되었는데 최근 몇 년간의 성장세는 미국, 일본에 비해서는 다소 낮은 것으로 평가되고 있다.

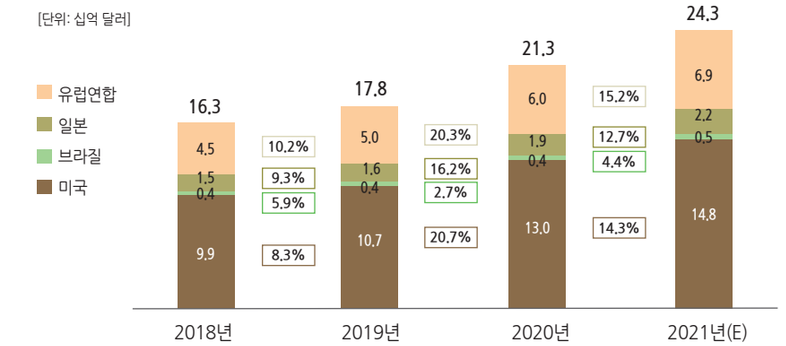
일본의 데이터 경제적 효과 중 직접 효과는 2016년에서 2017년 사이 소폭 하락세를 보였으나, 2018년에는 반등에 성공하였으며, 2019년에는 8.1% 증가한 367억 달러를 기록하였다. 2020년에는 14.7%, 2021년에는 11.3%의 높은 성장세를 보일 것으로 예측되었다.

브라질의 데이터 경제적 효과 중 직접 효과는 2018년에 약 83억 달러를 기록한 이후 꾸준히 증가하여 2020년 96억 달러를 기록하였다. 브라질의 ICT 산업이 빠르게 발전하고 있는데다 기반이 되는 데이터 시장 또한 빠르게 성장하고 있기 때문이다. 특히 브라질은 경제 위기를 심하게 겪고 있으나 데이터에 기반을 둔 제품 및 서비스의 영향력은 유지하고 있어 브라질 데이터 시장의 성장을 기대하게 한다.

(2) 간접 후방 효과

데이터의 경제적 효과 중 간접 효과는 데이터산업이 타 산업에 미치는 영향을 나타내는 요소이다. 직접 효과가 데이터 생산 활동이 데이터산업 내에 미치는 영향을 설명하는 것이라면, 간접 효과는 타 산업에 미치는 영향을 설명하고 있다. 간접 효과는 전방 효과와 후방 효과로 구분된다. 후방 효과란 데이터산업에 속한 기업이 데이터 기반의 제품 및 서비스를 생산하는 활동을 하는 과정에서, 타 산업으로부터 자원을 받을 때 자원을 제공해준 산업에서 발생하는 효과를 말한다.

[그림 3-2-5] 데이터산업의 경제적 효과: 간접 후방 효과(2018~2021년(E))



* '18~'20: IDC, 「The European Data Market Monitoring Tool Report」, 2020.7.8.
'21(E): 위 보고서를 바탕으로 한 저자 추정치

반면에 전방 효과는 데이터산업에서 생산된 데이터 기반의 제품 및 서비스를 이용하여 타 산업에서 다른 형태의 제품 및 서비스로 가공되는 효과를 말한다. 즉, 타 산업에서 데이터를 활용해서 제품 및 서비스를 제공함으로써 얻는 수익을 일컫는다. 전방 효과는 세 가지로 구분할 수 있는데, 첫째는 생산 및 공급 프로세스 최적화를 들 수 있다. 이는 데이터에 기반을 두고 프로세스를 통해 생산을 효율화하는 내용이다.

둘째는 데이터를 활용하여 표적 광고 및 개인 맞춤형 광고를 제공함으로써 마케팅 기술을 향상할 수 있는 것이다. 즉, 타 산업에서 데이터 분석을 통한 마케팅을 함으로써 수익이 증가하는 효과가 발생한다는 점이다.

마지막으로 데이터 기반의 조직을 구성할 수 있는 점이다. 즉, 데이터를 활용하여 기존 조직의 관리 방식을 개선함으로써 불필요한 자원 소모를 최소화하여 조직의 생산성을 증대할 수 있는 점도 전방 효과라 할 수 있다.

데이터산업으로 인한 간접 후방 효과는 매우 빠르게 증가하고 있다. 특히 미국의 경우 2016년 약 82억 달러 규모였으며, 이후 8%가 넘는 높은 증가세를 보이면서 2019년에는 107억 달러를 기록했다. 또한 2019년 이후로는 2021년까지 두 자릿수 성장을 이뤄내면서 2021년에는 148억 달러에 이를 것으로 전망되고 있다.

유럽연합의 간접 후방 효과는 2019년에 50억 달러 규모였으나 2020년 20%가 넘는 성장률을 보이면서 60억 달러 규모를 형성했다. 유럽연합의 경우 공공데이터 개방 정책을 통해서 민간데이터보다 공공데이터 중심의 데이터산업이 상당한 비중을 차지하고 있

고, 민간기업의 데이터 또한 활용 폭을 넓히고자 다양한 정책을 마련하고 있다. 산업 내에 1인 기업 및 소규모의 기업들이 많아 건전한 시장 상태의 성과가 본격적으로 나타나기 시작한 것으로 판단된다.

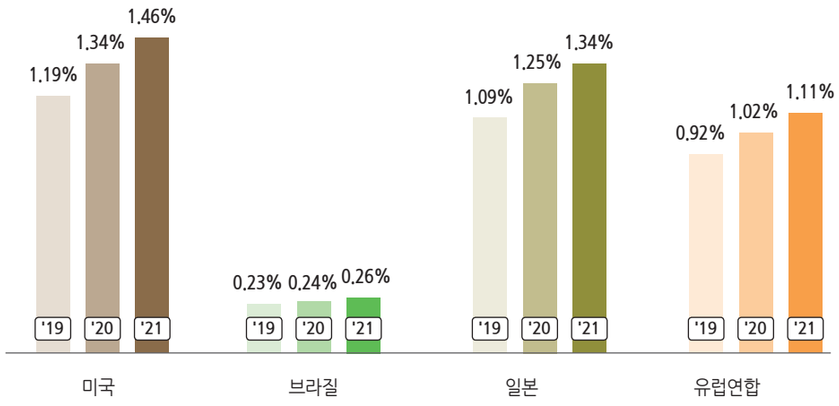
일본의 간접 후방 효과도 빠르게 증가하는 추세이다. 2018년에 약 15억 달러 규모였던 간접 후방 효과는 증가세를 지속하면서 2021년에 약 22억 달러 수준에 도달할 것으로 전망된다.

(3) GDP 대비 데이터 경제적 효과의 비율

앞서 데이터산업의 경제적 효과를 직접 효과와 간접 후방 효과를 중심으로 살펴보았는데, 직접 효과와 간접 후방 효과 모두 미국이 가장 앞선 것으로 나타났으며, 두 효과를 합하였을 때 2020년 기준 약 2,435억 달러 수준이었다. 시장 규모와 마찬가지로 경제적 효과 또한 미국은 유럽연합의 두 배 이상의 규모를 기록하고 있다.

GDP에서 경제적 효과가 차지하고 있는 비율을 살펴보면 미국은 2019년 1.19%에서 2021년 1.46%로 증가하여 약 0.27%포인트 증가하였으나, 유럽연합은 2019년 0.92%에서 2021년 1.11%로 상승하여 0.19%포인트의 성장세를 기록했다. 일본의 경우 GDP 대비 비중이 2019년 기준 1.09%였으나 2021년 1.34%로 약 0.25%포인트 증가하여

[그림 3-2-6] GDP 대비 데이터산업의 경제적 효과(직접 효과+간접 후방 효과)의 비율(2019~2021년(E))



* '19~'20: IDC, 「The European Data Market Monitoring Tool Report」, 2020.7.8.
'21(E): 위 보고서를 바탕으로 한 저자 추정치

미국에 견줄 정도의 성장세를 보였다. 브라질의 경우 2021년에도 0.26% 수준에 그쳐 GDP에서 차지하는 데이터 활용 및 경제적 효과는 다소 뒤처지는 것으로 나타났다.

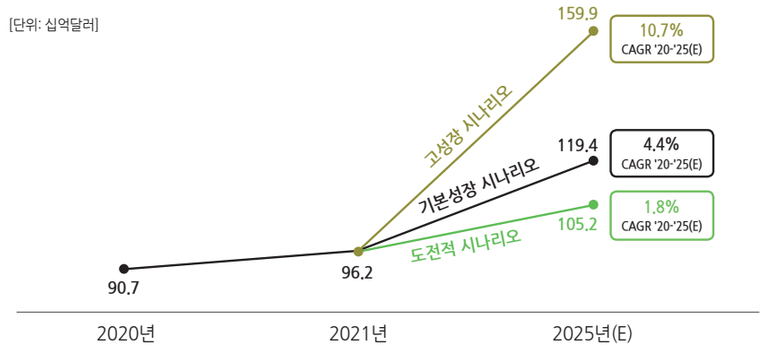
마. 중장기 데이터 시장 전망

향후 데이터 시장은 지속적인 성장세를 기록할 것으로 전망된다. 'European Data Market Monitoring Tool, IDC 2020 리포트'에 따르면 기본 시나리오 기준 2020년에서 2025년까지 연평균 성장률은 4.4%로 예측되어 2025년 유럽연합의 데이터 시장 규모는 1,194억 달러까지 성장할 것으로 보고 있으며 만약 고성장 시나리오가 전개된다면 1,599억 달러까지 내다볼 수도 있다는 평가가 제시되고 있다.

데이터 기업의 매출 성장률은 같은 기간 전체 ICT 시장의 성장률을 훌쩍 뛰어넘을 것이라는 전망이 지배적이며, 경제 규모가 큰 국가일수록 이러한 데이터 기반 시장은 데이터 경제 확대에 크게 이바지할 것으로 예측되었다.

데이터 관련 인력 또한 데이터 관련 기업의 향후 성장세를 견인하는 중요한 자원으로 평가받고 있어 지속적인 증가세를 보일 것으로 예측될 뿐만 아니라 높은 고용률과 고액 연봉 체제를 구축해 나갈 것으로 전망된다.

[그림 3-2-7] 유럽연합의 중장기 데이터 시장 전망(2020~2025년(E))[3]



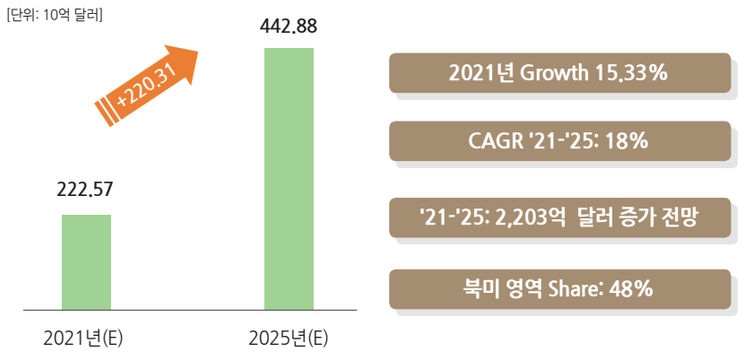
* '20, '25(E): IDC, 「The European Data Market Monitoring Tool Report」, 2020.7.8.
'21(E): 위 보고서를 바탕으로 한 저자 추정치

3 3가지 시나리오를 산출하기 위해 활용된 데이터는 다음과 같다. (1) IDC의 최근 전 세계 시장 예측 추정 및 black book(표준판), (2)IDC의 빅데이터 분석, Digital Transformation, Innovation accelerators 및 2019년 시장 동향, (3) EU GDP 및 ICT 지출의 2025 예측 업데이트, (4) 2013~2018년 EDM 모니터링 Tool에서 나타난 추세, EDM 지표를 모델링하고 예측에 사용된 시나리오를 참조

2. 글로벌 빅데이터 시장 전망

전 세계 기업 대부분은 다양한 디지털 플랫폼 서비스를 제공하면서 B2C(소비자 기반 서비스) 영역에서 비즈니스 기회를 포착하고 있다. 이러한 경향은 엄청난 양의 정형 데이터와 비정형 데이터가 생성되는 결과를 초래하며 기업 영역에서는 Industry 4.0과 디지털 트랜스포메이션 구축이 증가함에 따라 조직 간 데이터 생성량이 더욱 증가하고 있다. 이는 기업들이 빅데이터 분석과 같은 기술에 투자하여 인사이트를 끌어내도록 유도하고 있으며, 기업의 중요한 운영 및 경영 결정을 내리는 데 큰 도움을 주고 있다.

[그림 3-2-8] 글로벌 빅데이터 시장 전망(2021(E)~2025년(E))



* 출처: Cision PR Newswire, 「Big Data Market by Type, Deployment, and Geography | Global Forecast to 2025」, 2021.9.22.

글로벌 빅데이터 시장은 2021년 2,226억 달러에서 2025년 4,429억 달러로 크게 증가할 것으로 전망된다. 이미 2021년에도 15.3%의 높은 성장세를 보였으며 2021년부터 2025년까지의 연평균 성장률은 18%에 달할 것으로 예측되고 있다. 또한 같은 기간 글로벌 빅데이터 시장에서 북미 지역이 전 세계 빅데이터 시장의 약 48%를 차지할 것으로 예상된다. 수년간 미국 기업들은 빅데이터 및 고급 분석, IoT, 다양한 모빌리티 프로젝트 등 새로운 기술에 투자하여 다양한 디지털 전환 이니셔티브를 수행해 왔다. 또한, 유통 기업들의 서비스 자동화 및 플랫폼화는 기업들에 상당한 성장 기회를 창출하고 있다.

2020년 기업들은 빅데이터 서비스 부문에서 높은 수익을 창출했으며 이러한 성장은 빅데이터 분석 서비스에 대한 수요 증가를 의미하고 있다. 스마트화된 IT 인프라의 필요성이 지속적으로 증가함에 따라 빅데이터 서비스 부문의 성장으로 이어질 것으로 전망된다.

PART



산업별 데이터 활용 현황

제1장.
금융분야 데이터 활용

제2장.
헬스케어분야 데이터 활용

제3장.
제조분야 데이터 활용

제4장.
유통 · 물류분야 데이터 활용

제5장.
광고분야 데이터 활용

제6장.
모빌리티분야 데이터 활용

제7장.
에듀테크분야 데이터 활용

제8장.
농업분야 데이터 활용 · 비즈니스

제9장.
데이터 동맹 · 제휴 현황

제1장. 금융분야 데이터 활용

정구태
(카르도 전략부문 이사)

핀테크로 시작된 금융업 변화의 물결이 디지털 트랜스포메이션 시대를 맞이하여 더욱 가속화되고 있다. 금융산업도 더 이상 기존 금융사의 전유물이 아니며 새로운 기업, 서비스와 경쟁하는 시대에 접어들고 있다. 신구 경쟁자 간우위를 점하기 위해 가장 중요한 것은 무엇보다도 고객을 가장 잘 이해하는 것이다. 고객을 보다 잘 이해하기 위해서는 다양한 양질의 '데이터'를 가장 많이, 빠르게 확보하는 것이 중요하다. 데이터 확보는 각 산업의 전쟁에 비견될 정도로 치열하게 이루어지고 있다. 금융당국도 데이터 개방, 활용에 적극 동참하여 데이터 활용 경쟁을 통한 금융업의 발전과 금융소비자 혜택 확대를 지원하고 있다. 데이터 3법 개정을 시작으로 마이데이터 도입, 금융데이터거래소 출범, 금융공공데이터 개방, 금융 빅데이터 시스템 구축 등 새로운 제도와 금융 인프라가 속속 도입되고 있다. 최근에는 최적의 데이터 분석, 활용을 위한 인공지능 기술도 크게 각광받고 있다. 데이터와 인공지능이 바꿀 금융업의변화 모습이 더욱 기대되는 시점이다.

1. 데이터 기반 디지털금융

가. 핀테크와 디지털 트랜스포메이션[1]

2005년 핀테크가 등장한 이후 최근 디지털금융에 이르기까지 국내외 금융산업은 변화를 거듭하고 있다. 국내 금융업에서도 토스, 뱅크샐러드 등 신생 핀테크 기업을 비롯하여 카카오페이, 네이버 파이낸셜 등 빅테크 기업들도 금융혁신에 동참하고 있다. 2016년 케이뱅크를 필두로 국내 첫 인터넷전문은행이 등장하면서 금융업의 변화는 더욱 가속화되고 있다.

새로운 기술과 서비스의 등장은 기존 금융사에게도 커다란 위협이 되고 있다. 은행, 증권, 보험, 카드 등 전통 기업들도 금융업에 특화된 디지털 트랜스포메이션 시대를 준비하고 있다. 2021년 3월 과학기술정보통신부에서 발표한 '2020 인터넷 이용 실태조사'에 따르면 국내 가구인터넷 접속률은 99.7%, 국민 이용률은 91.9%에 달할 정도로 인터넷 보급률은 세계 최고 수준이다. 특히 인터넷뱅킹 이용률은 76.5%로 전년대비 11.6%가 증가할 정도로 가파르게 상승하고 있는데, 2021년 4월 한국은행에서 발표한 '2020년 국내은행 인터넷뱅킹서비스 이용현황'에 의하면 지난해 인터넷뱅킹을 통한 대출신청서비스 이용금액은 전년대비 151%로 크게 증가하였다. 이러한 금융소비자의 눈높이에 맞춰 금융기업의 업무들도 대부분 디지털로 전환 중이다. 이제 금융의 디지털 트랜스포메이션은 선택이 아닌 필수가 되고 있다. 금융기업들이 금융업의 본질을 재정립하고, 생존을 위해 독자적이고 창의적인 디지털 트랜스포메이션 전략을 추진하기에 이르렀다.

나. 데이터금융[2]의 중요성

국내 금융시장도 오픈뱅킹, 마이데이터(본인신용정보관리업), 마이페이먼트(지급지시서비스업), 종합지급결제업 등 새로운 제도가 속속 도입되고 있다. 은행, 증권, 카드, 보험 등 기존 금융사의 전유물로만 여겨졌던 금융산업이 핀테크, 빅테크 등 새로운 플레이어의 등장으로 디지털금융 선점을 위한 각축장으로 변모하고 있다. 경쟁을 위한 필수 요건은 '데이터'와 이를 잘 활용할 수 있는 방법을 찾는 것이다. 데이터 활용에 가장 효과적 방법은 인공지능과 같은 디지털 신기술을 활용하는 것이다. 데이터와 디지털 신기술은 그 자체로 의미있는 것이 아니라 이를 활용하여 비즈니스에 접목할 때 비로소 새로운 가치가 생겨난다. 앞으로 양질의 데이터를 누가 어떻게 활용하고 선점하느냐에 따라 디지털금융 시장의 패권 선점은 달라질 수 있다.

2. 데이터 3법과 금융데이터[3] 활용

가. 데이터 3법의 시행

2020년 2월, '데이터 3법' 개정안이 1년 2개월 만에 국회 본회의를 통과했다. 데이터 3법이란 개인정보보호법, 정보통신망 이용 촉진 및 정보보호 등에 관한 법률(이하 '정보통신망법'), 신용정보의 이용 및 보호에 관한 법률(이하 '신용정보법') 등 3가지 법을 통칭한다. 기존

1 인공지능, 클라우드, 빅데이터, 블록체인 등 디지털 기술을 기반으로 상품, 서비스, 프로세스, 업무방식 등을 혁신하고 기존 다른 새로운 디지털 기반의 비즈니스 모델로 전환하는 과정

2 데이터를 기반으로 하는 금융시스템 및 서비스
3 금융업에서 활용되는 데이터

의 법체계는 정보보호 측면만을 강조하여 데이터 활용을 어렵게 하고 이를 기반으로 한 산업 발전에 저해를 가져온다는 지적이 많았다. 정부당국은 데이터 3법 개정을 통해 데이터 경제의 토대를 만들고 데이터 이용 활성화와 관련 산업 진흥을 꾀하고 있다.

[표 4-1-1] 데이터 3법 주요 개정내용

구분	개정내용
개인정보보호법	- 개인정보 관리 감독 기능을 개인정보보호위원회로 일원화 - 가명정보 개념을 도입하여 제품 서비스 개발에 활용 - 가명정보 이용 시 안전장치 및 통제수단 마련
신용정보법	- 가명정보를 금융 분야 빅데이터 분석에 활용 - 일부 목적에 한해 신용정보 주체자 동의 없이 가명정보 이용·제공 허용 - 신용정보 주체자의 본인정보 통제 기능 강화
정보통신망법	- 온라인상 개인정보 관리·감독 권한을 개인정보보호위원회로 이관

* 출처: 대한민국 정책브리핑, 「데이터 3법」, 2020.3.30.업데이트
(https://www.korea.kr/special/policyCurationView.do?newsId=148867915, 접속일: 2021.11.5.)

데이터 3법 개정으로 말미암아 정보제공 기관에서 수집한 고객의 신용데이터를 외부 데이터와 결합할 수 있는 법적 근거가 마련되었다. 데이터 3법 개정안은 '가명정보[4]' 개념을 도입하여 명확한 개인정보 판단 기준 마련 및 이종 데이터 결합을 촉진하고자 하였다. 가명정보는 완전한 비식별 조치가 된 익명정보와 식별 가능한 개인정보 사이의 중간 단계라고 할 수 있다. 가명정보는 통계 작성, 과학적 연구, 공익적 기록 보존을 목적으로 정보주체의 동의 없이 활용이 가능하다. 은행, 증권, 보험 등 정보제공 기관이 내부 데이터를 외부로 반출할 시엔 반드시 가명조치 후 중앙기관의 승인 하에 제공해야 한다. 서로 다른 기업이 보유하고 있는 가명정보는 정부에서 허가받은 전문기관을 통해서만 결합할 수 있다.

4 개인정보 일부를 삭제·대체하는 등 가명으로 처리하여 추가 정보없이 개인을 알아볼 수 없도록 한 개인정보

[표 4-1-2] 개인정보, 가명정보, 익명정보 비교

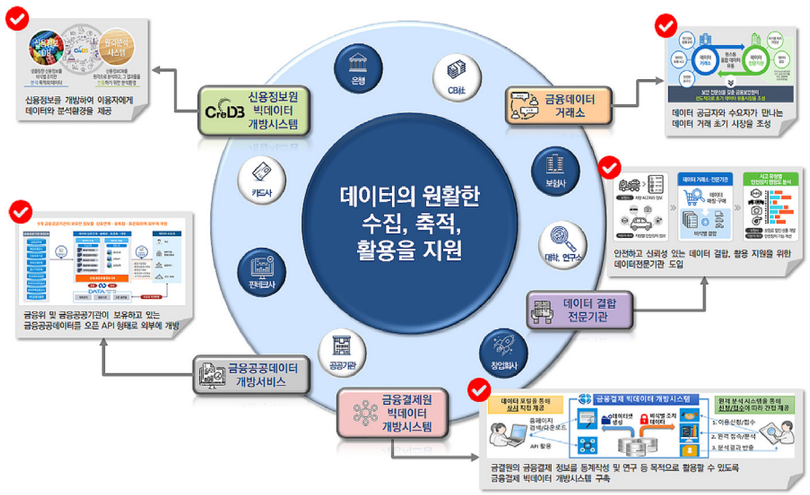
	개념	활용가능 범위
개인정보	특정 개인에 관한 정보, 개인을 알아볼 수 있게 하는 정보	사전적이고 구체적인 동의를 받은 범위 內 활용 가능
가명정보	추가정보의 사용 없이는 특정 개인을 알아볼 수 없게 조치한 정보	다음 목적에 동의 없이 활용 가능(EU GDPR 반영) ① 통계작성(상업적 목적 포함) ② 연구(산업적 연구 포함) ③ 공익적 기록보존 목적 등
익명정보	더 이상 개인을 알아볼 수 없게 (복원 불가능할 정도로) 조치한 정보	개인정보가 아니기 때문에 제한없이 자유롭게 활용

* 출처: 대한민국 정책브리핑, 「데이터 3법」, 2020.3.30.업데이트
(https://www.korea.kr/special/policyCurationView.do?newsId=148867915, 접속일: 2021.11.2.)

나. 금융데이터의 활용

데이터의 활용 가치가 높아짐에 따라 이를 통한 새로운 기술, 서비스의 결합 시도가 지속되고 있다. 금융업 역시 핀테크를 비롯한 빅테크 기업의 데이터 융복합을 시도하고 있다. 이에 따라 각 산업 간 데이터 활용 경계도 점차 희미해질 것으로 예상된다. 특히 금융기업이 보유한 방대한 금융 정보와 통신, 유통, 콘텐츠 등 다른 분야의 고객데이터가 결합되면서 맞춤형 금융상품과 서비스 등장을 예고하고 있다. 이전에는 금융기업 자체적으로 축적한 데이터만으로 금융상품과 서비스를 설계·개발하였다면, 앞으로는 경쟁 금융기업의 데이터(입출금 내역, 카드거래명세, 보험계약 정보 등)뿐만 아니라 통신료 납부내역, 유통사의 장바구니 정보 등 이종 산업의 데이터와도 결합하여 상품 및 서비스 개발에 활용할 수 있게 되었다. 이를 통해 고객의 소비패턴을 분석하고 고객 성향에 대한 종합적인 컨설팅과 초개인화된 서비스 제공이 가능하게 된다. 금융위원회도 '금융 빅데이터 5대 인프라 체계'를 마련하여 각종 금융 데이터의 원활한 수집, 축적, 활용을 지원하고 있다.

[그림 4-1-1] 금융 빅데이터 5대 인프라 체계도



* 출처: 금융위원회, 「금융 빅데이터 개방으로 디지털 금융혁신을 촉진합니다」, 2020.7.1.

3. 금융데이터와 비즈니스 활용 사례

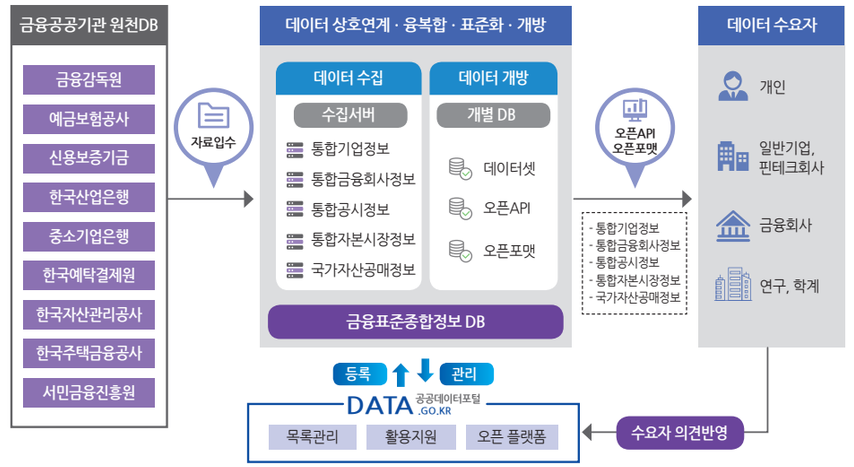
가. 금융공공데이터[5] 개방

2020년 4월부터 금융위원회를 비롯하여 금융감독원, 예금보험공사, 한국산업은행, 중소기업은행, 신용보증기금, 한국예탁결제원, 한국자산관리공사, 주택금융공사, 서민금융진흥원 등 9개 금융공공기관이 보유한 금융공공데이터를 공공데이터포털[6](www.data.go.kr)을 통해 오픈API[7]로 제공하기 시작하였다. 개방되는 데이터는 통합기업, 통합금융회사, 통합공시, 통합자본시장, 통합국가자산공매 등 5개 핵심분야 테마정보와 50개 서비스로 5,500개 항목, 4,400만 건으로 구성되어 있다. 기존에는 금융서비스 개발 시 신용정보회사를 통해 유료로 해당 데이터를 구매하였으나 이제 금융공공데이터를 통해 무료로 활용할 수 있다. 또한 핀테크 기업들은 비외감법인 정보, 통합 공시정보 등을 활용하여 거래 기업의 재무현황을 파악하고 새로운 영업 기회를 발굴할 수 있다. 연구기관들도 금융기관 통계정보, 공시정

5 금융업에서 활용되는 데이터
6 공공데이터의 제공 및 이용 활성화에 관한 법률(제11956호)에 근거하여 국가에서 보유한 다양한 데이터를 제공하는 통합창구 사이트
7 다른 프로그램이 특정 기능·데이터에 접근할 수 있도록 사전에 규정한 통신규칙(Application Programing Interface)

보 등 연구에 필요한 데이터 활용이 가능하다.

[그림 4-1-2] 금융표준종합정보DB 체계도

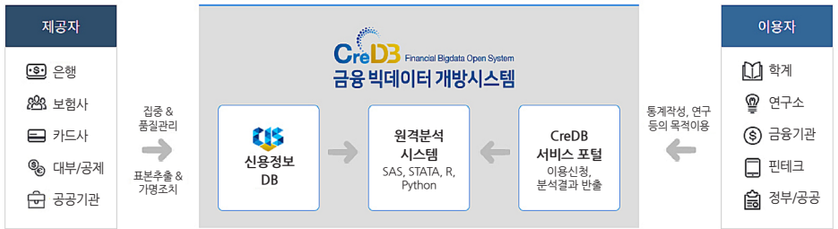


* 출처: 금융위원회, 「금융공공데이터 전면개방」, 2020.6.5.

나. 금융 빅데이터 개방시스템

한국신용정보원은 '신용정보의 이용 및 보호에 관한 법률'에 근거하여 2019년 6월부터 '금융 빅데이터 개방시스템(CreDB)'을 개발·운영하고 있다. 금융 빅데이터 개방시스템은 다양한 금융, 신용데이터를 가공하여 분석용 데이터와 분석 시스템을 제공하며, 금융업에 필요한 통계 작성 및 빅데이터를 분석하여 다양한 분야의 정책, 학술 연구를 지원한다. 금융 빅데이터 개방시스템은 그동안 한국신용정보원에만 집중되었던 신용정보를 표본추출하고 비식별 조치하여 제공한다. 제공되는 데이터는 크게 표본DB와 맞춤형DB 2가지 종류로 나뉜다. 표본DB는 사전에 정의한 제공 항목에 대해 전체 모수를 기준으로 표본 추출하여 구축한 분석용 데이터로써 개인신용DB, 기업신용DB, 보험신용DB로 구성된다. 맞춤형DB는 이용자의 요구사항과 연구목적에 따라 커스터마이징된 데이터를 제공한다.

[그림 4-1-3] 금융 빅데이터 개방시스템(CreDB) 체계도



* 출처: 금융빅데이터 개방시스템 홈페이지(http://credb.kcredit.or.kr, 접속일: 2021.8.9.)

다. 금융데이터거래소

2020년 5월에 출범한 금융데이터거래소는 데이터 공급자와 수요자를 상호 매칭하는 중개시스템이다. 정보 유출 방지 등 보안성을 유지하면서 편리한 거래 기능을 제공하고 데이터 거래가 원활히 이루어질 수 있도록 지원하고 있다. 금융데이터거래소는 금융업과 이종 산업을 연결하는 개방형 데이터 거래소를 운영하는 것이 목표이다. 금융정보 외에도 다양한 분야의 데이터가 함께 거래되도록 통신, 유통 등 일반 상거래 기업들도 참여가 가능하다. 금융데이터거래소는 수요자가 원하는 데이터, 제공 형태 등을 공급자에게 직접 요청하는 수요자 중심의 거래를 지향한다. 이용자와 개별 연락 없이 거래소 시스템만으로 모든 거래 절차가 진행되는 편의성이 있다. 이를 통해 금융데이터 검색, 계약, 결제, 분석 등 데이터 유통 전 과정을 원스톱으로 지원하게 된다. 금융데이터거래소는 금융업에 새로 진입하는 핀테크 기업의 창업과 비즈니스 기회를 확대하는 것뿐만 아니라, 이종 분야 데이터 활용을 통해 새로운 서비스 개발의 활성화를 지원한다. 예를 들어 특정 지역에 필요한 맞춤형 마케팅 정보 창출이나 상권분석 앱을 개발한다면 해당 지역의 소득수준, 소비패턴 등 기존 금융정보와 함께 지리정보, 학군정보, 상권정보 등 비금융정보를 결합하여 만들 수 있다.

2020년 8월 신한카드는 SK텔레콤과 함께 금융, 통신 가명정보 결합상품인 '카드소비×이동동선 데이터'를 만들었다. 이 상품은 2020년 여름 부산지역 주요 관광지별 고객 체류시간, 소비금액, 소비건수 등을 결합·분석한 데이터다. 한국문화관광연구원은 이를 활용하여 국민 관광형태 분석에 활용하고 있다.

[그림 4-1-4] 금융데이터거래소 체계도



* 출처: 금융데이터거래소 홈페이지(https://www.findatamall.or.kr, 접속일: 2021.11.2.)

라. 마이데이터

2021년 2월 '마이데이터 가이드라인'이 발표되었다. 향후 마이데이터 사업자들은 표준화된 오픈API를 통해 개인신용정보를 수집·활용하여 소비자에게 서비스를 제공하게 된다. 가이드라인에 따르면 신용정보의 제공범위는 크게 5가지 부문으로 구분된다. 여·수신 금융투자 부문은 예·적금(납입액, 금리, 만기 등), 대출(잔액, 금리, 만기 등), 투자상품(예수금, 매입종목, 거래단가·수량, 평가금액 등)이며, 보험 부문은 가입상품(계약, 특약, 납입내역, 자기부담금 등), 대출내역(잔액, 상환내역 등)을 제공한다. 카드 부문은 월 이용정보(금액, 일시, 결제예정총액), 카드대출, 포인트 등이며, 전자금융 부문은 선불발행정보(잔액, 충전계좌), 거래내역(일시, 금액) 및 12개 범주화[8] 된 주문내역정보 등을 제공한다. 가이드라인에 따르면 개인신용정보 전송요구를 보다 명확하게 규정하여, 정보주체가 정보제공기관에 개인정보 이동을 요구하면 마이데이터 사업자가 해당 제공정보를 안전하게 전송할 수 있도록 하였다. 개인이 정보제공기관의 종합포털을 통해 개인정보 전송을 요구하면 개인별 PDS[9](Personal Data Storage)로 전송하게 된다.

2021년 2월에는 마이데이터 지원센터도 개소하였다. 마이데이터 지원센터는 마이데이터 프로세스 운영방안, 사업자 허가 절차, 제공 정보 종류, API 규격 등 마이데이터 제도 안내를 비롯하여 참여기관 관리, 등록 등 마이데이터 생태계 전반을 지원할 계획이다.

8 가전/전자, 도서/문구, 패션/의류, 스포츠, 화장품, 아동/유아, 식품, 생활/가구, 여행/교통, 문화/레저, 음식, e쿠폰/기타
9 마이데이터 종합포털 내 구축되는 본인의 데이터를 안전하게 저장·관리하는 플랫폼

[그림 4-1-5] 마이데이터 지원센터 주요 역할

* 출처: 마이데이터 지원센터 홈페이지(http://www.mydatacenter.or.kr/, 접속일: 2021.7.22.)

마. 인공지능(AI) 활용

2021년 7월 금융위원회는 인공지능 기반 금융서비스 개발을 위한 '금융분야 인공지능 가이드라인'을 발표하였다. 가이드라인은 금융기업들이 보다 적극적으로 인공지능 기술을 활용하고 금융서비스에 대한 고객 신뢰를 확보하는데 목적이 있다. 가이드라인에 따르면 금융기업은 인공지능 서비스 운영의 책임 강화를 위해 인공지능 윤리원칙 마련, 인공지능 조직 구성, 위험관리 정책 수립 등 3중 내부 통제장치를 마련해야 한다. 이를 통해 인공지능 학습용 데이터의 정확성, 안정성을 확보하고 인공지능 금융서비스의 투명성, 공정성을 강화하게 된다. 2021년 내에 가이드라인이 본격적으로 시행되면, 인공지능 기술을 활용한 신용평가, 대출심사, 자산관리 등 핀테크 기업들의 신규 서비스 등장과 핀테크 사업의 활성화가 기대된다. 또한 은행, 증권, 카드, 보험 등 기존 금융사들도 금융상품 설명, 고객상담 등을 인공지능 기반으로 설계하여 업무 효율성 측면에서도 많은 도움이 될 것으로 예상된다.

[그림 4-1-6] 금융분야 인공지능 가이드라인 주요 내용



* 출처: 금융위원회, 「금융분야 인공지능 가이드라인이 시행됩니다」, 2021.7.8.

바. 대안신용평가

대안신용평가란 금융거래 정보가 아닌 통신·전기·가스요금 납부이력, 온라인 구매정보, SNS 정보, 이메일 정보 등 비금융정보를 활용하여 신용을 평가하는 방식을 의미한다. 대안신용평가 도입을 통해 금융정보가 부족해 신용평가가 어려운 사회초년생, 주부 등 씬파일러(Thin Filer)[10]에게도 금융 혜택을 제공할 수 있다. 신용회복위원회 조사에 따르면 국내 씬파일러는 약 1,300만 명에 달하는데, 대안신용평가를 활용하면 중금리 시장을 확대하고 리스크가 낮은 우량 고객을 발굴하여 대출 승인율은 높이면서 연체율은 감소시킬 수 있다.

네이버 파이낸셜은 기존 신용평가 회사가 보유한 금융데이터에 자체적으로 분석한 다양한 데이터를 활용하여 대안신용평가를 도입하였다. 미래에셋캐피탈, 우리은행과의 제휴를 통해 네이버 쇼핑몰인 스마트스토어에 입점한 온라인 소상공인에게 대출중개 업무를 하고 있다. 소상공인은 기존 평가측정 모델로는 신용도가 낮아 대출을 받을 수 없었지만, 자재구매, 상품판매, 경제활동 등 수집가능한 데이터 등을 인공지능으로 분석하여 위험도가 낮은 적정한 대출규모를 산출·제공 받게 된다. 신한카드는 자사가 보유한 결제 데이터와 신용평가업체, 음식배달 플랫폼, 프리랜서 마케팅 플랫폼 등과 컨소시엄을 구성하여 플랫폼 종사자를 위한 대안신용평가 모델을 구축할 예정이다. 카카오뱅크도 한국신용데이터와 함께 중금리 혁신법인을 공동 설립하여 개인사업자 대출시장에 뛰어들었다.

10 신용카드 사용내역, 대출 실적 등 금융거래가 거의 없어 금융정보가 없는 고객

[표 4-1-3] 금융사 대안신용평가모델 개발 현황

구분	내용
네이버 파이낸셜	스마트스토어 입점 소상공인 전용 대안신용평가 및 대출상품 개발
카카오뱅크	한국신용데이터와 조인트벤처 설립 및 중금리대출시장 진출
금융결제원	한국기업데이터와 소상공인 전용 신용평가모델 구축
롯데카드	나이스 평가정보와 자영업자 전용 대안신용평가모델 공동 개발

* 출처: 각사 발표자료 정리

4. 기대효과 및 향후 전망

금융정보 이용 활성화를 위한 제도 도입 및 정비를 통해 데이터금융 시대가 더욱 가속화될 것으로 예상된다. 2020년 8월 금융당국도 디지털금융 종합혁신방안을 내놓고 마이데이터, 마이페이먼트, 종합지급결제업을 신설하였다. 2020년 7월 금융위원회는 금융 빅데이터 5대 인프라 체계와 함께 금융데이터 생성, 결합, 거래 확대 계획을 발표하였다. 데이터 개방 확대와 신규 사업 육성을 통해 디지털금융 경쟁과 혁신의 촉진 환경이 마련되고 있다. 이러한 환경 변화에 맞춰 금융업에 새롭게 진출하는 기업들도 혁신적인 아이디어와 기술을 무기로 서비스 개발에 박차를 가하고 있다.

기업들이 효과적인 데이터 활용을 위해 인공지능 기술도 활발하게 사용하고 있어, 앞으로는 데이터 자체의 중요성뿐만 아니라 데이터를 다루기 위한 인공지능 기술도 각광받게 될 것이다. 인공지능을 통해 비정형, 비금융데이터 활용이 활성화되면서 금융데이터가 기존에 비해 더욱 효율적으로 사용될 수 있다. 이에 따라 금융데이터 부족으로 인해 소외될 수밖에 없었던 취약계층을 제도권 금융으로 포용할 수 있게 되고 데이터 처리 속도와 정확성이 증가되어 금융거래 비용이 낮아지게 된다. 궁극적으로 금융업에서 추구하는 금융 중개기능의 효율성은 더욱 증대될 것으로 예상된다.

제2장. 헬스케어분야 데이터 활용[1]

한현욱
(차의과학대학교 의학전문대학원 정보의학교실 교수)

헬스케어데이터는 인간이 평생 살아가면서 생산하는 데이터 중에 가장 다양하고 많은 양으로 생산하는 데이터 중에 하나다. 산업계에서는 이러한 데이터를 이용해 새로운 산업 생태계를 만들기 위해 노력중이다. 헬스케어분야 데이터 활용은 정부-병원-기업이 유기적으로 협력해야 성공할 수 있다. 왜냐하면 의료는 민간영역 뿐만 아니라 공공영역이 복잡하게 얽혀 있으며 여러 이해당사자가 있기 때문이다. 이 장에서는 헬스케어분야 데이터 활용 현황을 여러예를 들어 설명하고 산업 활성화를 위한 주요 국가 R&D 과제 및 동향에 대해 살펴본다. 이를 통해, 헬스케어분야에서 미래 데이터산업의 발전 가능성과 향후 과제를 제시하고자 한다.

1. 헬스케어데이터 산업 정책 동향

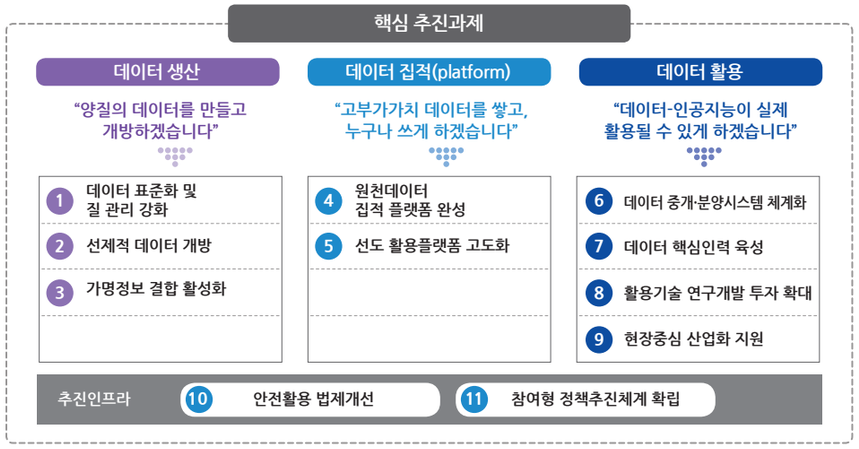
인간이 만들어낸 가장 다양하며 방대한 종류의 데이터 중 하나가 바로 헬스케어데이터이다. 헬스케어데이터 산업은 바이오기술과 정보기술이 눈부시게 발전함에 따라 비약적인 성장을 하고 있다. 하지만 헬스케어데이터 산업은 개인의 민감한 의료정보를 다루는 의료법과 생명윤리 및 안전에 관한 법률의 지배를 받는 강력한 규제의 산업이기도 하다. 헬스케어분야는 민간영역과 공공영역이 복잡하게 얽혀있는 분야로 산업 비즈니스를 영리기관인 기업에 한정하기는 어렵다. 왜냐하면, 헬스케어분야는 비영리 기관인 병원의 역할이 매우 크기 때문이다. 헬스케어데이터 산업 분야는 정부-병원-기업이 유기적인 협력을 통해 산업 비즈니스가 이루어지고 있다. 헬스케어데이터 산업 활성화를 위해 국가에서는 데이터의 생산·집적·활용에 관한 3대 분야와 이를 실제 구체화하기 위한 9개의 핵심과제를 선정하여 추진한다는 계획이다. 즉, 양질의 데이터를 생산하여 개방하는 방법, 고부가가치 데이터를 쉽고 편리하

1 이 장에서 저자가 사용한 '헬스케어데이터'와 '의료데이터', '헬스케어분야'와 '의료분야'는 의미에 있어 차이가 없음을 밝힙니다.

게 누구나 사용할 수 있도록 플랫폼을 구축하는 방법, 마지막으로 구축된 빅데이터를 이용해 인공지능을 개발하기 위한 방법이 그것이다.

정부에서는 이미 한국판 뉴딜을 통해 10대 중점과제 중 하나로 '스마트의료 인프라'를 선정하여 의료 빅데이터와 인공지능 기술을 활용해 스마트병원을 집중 육성·발전시키겠다고 계획한 바 있다. 최근 이를 뒷받침 해주기 위해 헬스케어분야에서 다양한 국가 R&D투자가 이루어지고 있다. 이 글에서는 헬스케어데이터 산업의 주요 예를 소개하고, 관련 국가 R&D 투자 및 기술 동향 파악을 통해 향후 헬스케어 산업 비즈니스 전망에 대해 논하고자 한다.

[그림 4-2-1] 보건의료데이터·인공지능 혁신전략 추진 과제



* 출처: 보건복지부, 「보건의료 데이터 인공지능 혁신전략」, 제126차 국정현안점검조정회의, 2021.6.3.

2. 헬스케어 산업에서 데이터 활용 현황 및 동향

헬스케어분야는 전자의무기록의 보편화와 유전체 분석 기술의 고도화에 힘입어 광범위한 디지털화가 진행 중이다. 전통적인 생명과학 분야의 연구 방법인 세포실험(In-vitro)과 동물실험(In-vivo)이 초고속 컴퓨팅 기술을 이용한 데이터 분석에 근거한 생명정보분석실험(In-silico)의 도움을 받고 있다. 또한 웨어러블기기의 대중화를 통해, 개인중심의 라이프로그 헬스케어데이터의 양이 기하급수적으로 증가하고 있다. 기존의 전통적인 정보통신 기업은 그동안 전혀 경험해 보지 못한 헬스케어데이터를 마주하며, 새로운 비즈니스 전략을 수립하기 위해 부심 중이다. 해외에서는 이미 아이비엠 왓슨(IBM Watson)을 시작으로 애플 및 구글과 같은 IT 공룡 기업들이 헬스케어데이터 산업 분야에 직간접적으로 진출하였다. 환자의 헬스케어데이터를 플랫폼화하여 체계적으로 수집하기 위해 이미 포석을 깔아놓은 상태이다.

최근 국내에서도 헬스케어데이터 산업 분야에 도전장을 내는 기존 기업들이 속속 등장하고 있다. 예를 들어, 네이버 클라우드는 헬스케어분야에 진출하기 위해 의료분야 클라우드 인증을 받아 놓은 상태이다. KT는 데이터기반 감염병 대응 기술 개발을 위해 '샤인'이란 앱을 출시하였다. 이 프로젝트는 자선단체인 빌&멜린다 게이츠 재단을 통해 지원받아 진행하는 연구 프로젝트로 인공지능 및 빅데이터 기반 감염병 조기진단 알고리즘 개발과 통신 데이터 기반 감염병 확산 경로 분석 연구를 진행 중이다.

한편, 최근 국내에서 헬스케어데이터산업 분야 스타트업이 선전하고 있다. 특히, 헬스케어데이터 산업 분야에서 스타트업으로서 가장 발전한 분야가 의료영상 인공지능 분야이다. 국내의 의료영상 인공지능을 선도하는 기업으로는 뷰노, 루닛과 같은 기업이 있다. 이미 이러한 회사는 창업 10년도 되지 않은 짧은 시간 내에 코스닥 상장을 하며 기술적 우위를 점하였다. 최근 빅데이터와 인공지능 기술을 바탕으로 신약개발에 문을 두드려 IPO에 성공한 신테카바이오와 같은 기업도 있다. 신테카바이오는 다양한 종류의 유전체 및 임상데이터를 수집하여, 인공지능을 통한 신약 후보물질을 발굴하는 기업이다. 최근에는 개인화된 의료정보를 유통 및 관리하기 위한 헬스케어 기업도 등장하였다. 메디블록은 헬스케어데이터를 안정적으로 유통하기 위해 블록체인기술을 활용한다. 레몬헬스케어는 개인의 진료정보를 보험사와 연계하여 실손보험을 청구할 수 있는 시스템을 구축하고 있다.

최근의 산업 경향은 헬스케어데이터 산업분야의 비즈니스 모델이 과거보다 훨씬 다양화·다각화 되고 있다는 것이다. 또한 병원과 기업이 협업하는 비즈니스 모델이 점점 증가하고 있다. 의사가 진료실을 떠나 기업의 환경을 경험하고 창업하는 수가 지속적으로 증가하고 있다. 이들은 자신의 진료경험을 바탕으로 전통적인 바이오 기업뿐만 아니라 헬스케어데이터에 근거한 임상결정지원시스템(CDSS) 및 디지털치료기기 분야에 도전장을 내고 있다.

정부에서도 헬스케어분야에서 빅데이터 산업 분야를 육성·발전시키기 위해 다음 장에서 소개할 다양한 국가 R&D계획을 마련하고 지속적으로 지원할 계획이다.

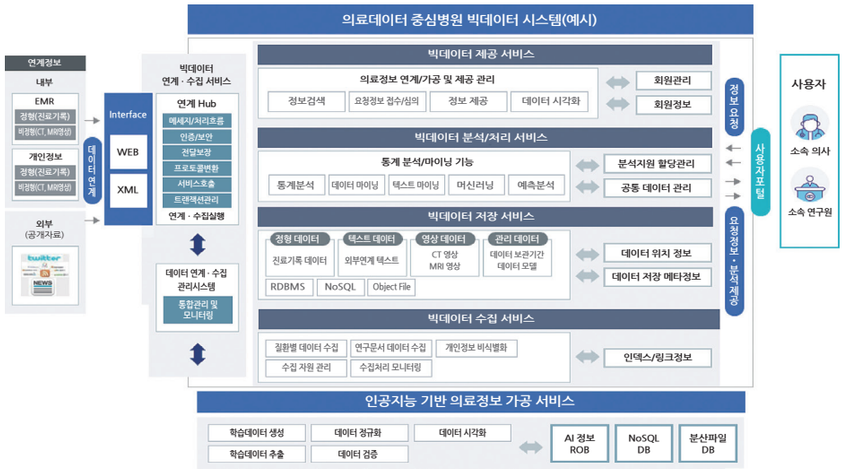
3. 헬스케어데이터 분야 주요 정부 R&D

가. 보건의료데이터 중심병원 지원 사업

'보건의료데이터 중심병원 지원 사업'은 민간병원을 중심으로 빅데이터 기반의 병원 의료 혁신 생태계 확보를 위해 보건복지부가 지원을 하는 대규모 국책 프로젝트이다. '보건의료데이터 중심병원 지원사업'은 민간병원에 축적되어 있는 대규모 임상데이터를 활용해 새로운 치료기술, 신약, 의료기기 및 인공지능 등의 개발을 지원하기 위한 사업이다. 이를 통해 병원의 효과적인 데이터 활용 환경을 구축하고, 표준화·보안·전문인력을 확보해 전문분야별 특화 데이터를 구축하고 연구자가 의료데이터를 이용한 다양한 연구 사업을 추진하는 원동력을 마련하고자 한다. 현재 서울대학교병원 및 세브란스병원 등 총 7개 컨소시엄, 40개 병원 및 41개 기업이 참여하고 있으며 총 200억원이 넘는 국비가 투입되었다.

이 사업의 주요 내용은 첫째, 병원이 충분한 전산장비를 도입·보강하여 의료데이터 활용 환경을 구축하고, 둘째, 데이터 활용기반을 마련해 국제적인 수준의 데이터 표준화·거버넌스를 구축·지원 할 수 있는 환경을 조성하며, 셋째, 신생물·순환계통질환·호흡계통질환 또는 질환·연령·환자군에 따른 분야별 특화된 의료데이터 개발·활용을 할 수 있도록 임상데이터를 제공하고, 넷째, 시나리오에 근거한 의료데이터 셋을 정의 및 구축하여 신약·의료기기, AI 등 개발을 위한 연구 체계를 마련하는 것이다.

[그림 4-2-2] 의료데이터 중심병원 빅데이터 시스템(예시)



* 출처: 보건의료정보원 홈페이지「보건의료데이터 중심병원 지원 사업」, (접속일: 2021.11.2.)

[표 4-2-1] 2021년 '보건의료데이터 중심병원 지원 사업' 참여현황

주관병원	참여병원	참여기업
부산대학교병원	양산부산대병원, 전남대병원, 화순전남대병원, 경북대병원, 칠곡경북대병원	㈜아이티아이즈, ㈜비주얼터미놀러지
삼성서울병원	한양대병원, 강북삼성병원, 삼성창원병원	삼성SDS㈜, GE헬스케어, 한미약품, JW중외제약, 제약바이오협회, 대웅제약, 한국존슨앤드존스메디컬㈜, 성균관대산학협력단
서울대학교병원	계명대동산병원, 동국대일산병원, 분당서울대병원, 서울보라매병원, 이화여대의대병원, 충남대병원, 충북대병원	이지케어텍㈜, ㈜인피니트헬스케어, 사이앱스코리아 유한회사
서울아산병원	서울성모병원, 국립암센터	㈜라인웍스, ㈜인브레인, ㈜유젠스, 가톨릭중앙의료원, 평화이즈, 데이터스트림즈, 메타빌드
세브란스병원	강남세브란스병원, 용인세브란스병원, 원주세브란스병원, 아주대병원	연세대학교 인공지능대학원, 한국전자통신연구원, ㈜파이디지털헬스케어, 대아정보시스템, 솔비트, ㈜태영소프트, ㈜미소정보기술, 레몬헬스케어, 지니너스 주식회사, ㈜위세아이텍, ㈜에이아이트릭스, ㈜메디웨일, ㈜헬스올, 버드온
고려대학교의료원	고대구로병원, 고대안산병원, 건양대병원, 길의료재단, 김안과병원	네이버클라우드㈜, 티맥스티베로
한림대학교 성심병원	강동경희대병원, 강릉아산병원, 강원대병원, 울산대병원, 춘천성심병원, 세종병원, 한길안과	㈜헬스리안, ㈜메디컬에이아이, ㈜미소정보기술, ㈜웰리시스, 에비드넷, Lux Mind

* 출처: 보건복지부,「의료데이터 중심병원 지원 사업 착수보고회」, 2021.4.1.

나. 헬스케어데이터 댐 사업

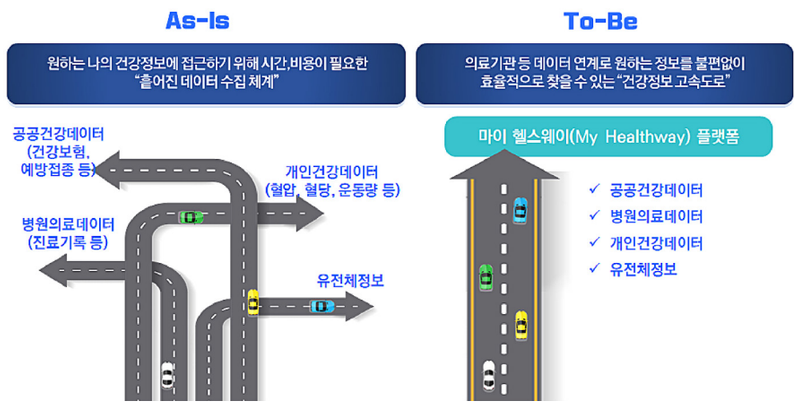
한국판 뉴딜의 사업으로 추진 중인 10대 과제에서 헬스케어분야와 연계된 과제로는 데이터 댐, 스마트 의료 인프라, 디지털 트윈이 있다. 이 중 데이터 댐 사업은 광범위한 데이터를 플랫폼화된 '댐'에 가둬두고 필요한 곳에 적재적소에 사용할 수 있게 하여 데이터의 사용 효율을 높여 부가가치를 만들고자 하는 사업이다. 그동안 많은 종류의 막대한 데이터가 여러 민간기업 및 공공기관에서 다양하게 생산되었으나, 데이터를 보유한 기관이 시간과 비용의 문제로 이를 체계적으로 개별 구축하는 것이 사실상 어려웠다. 이를 위해 헬스케어분야에서는 질병 중심으로 여러 병원에 분산 저장되어 있는 의료데이터를 통합 및 정제하여 고품질 학습용 데이터를 구축하

기 위한 작업에 착수했다. 예를 들어, 척추질환 환자로부터 영상데이터 및 임상데이터를 표준화 및 레이블링하여 민간에 개방하는 것이 그것이다. 이를 통해 질병 중심으로 여러 의료기관에 저장되어 있는 데이터를 통합하여 개방할 수 있게 되었다. 특히 데이터 댐 사업의 전제 조건은 민간기업이 주도한다는 데 있다. 그동안 의료분야에 진출하고자 하는 기업이 많았지만, 병원으로부터 의료데이터를 수집하는 것이 사실상 불가능에 가까웠기 때문에 이 사업을 통해 양질의 의료데이터를 확보할 수 있는 루트가 확보되었다는 면에서 의미가 크다고 할 수 있다.

다. 마이 헬스웨이(건강정보 고속도로) 사업

여러 의료기관에 흩어져 있는 개인의 의료정보를 자신의 스마트 기기에서 손쉽게 확인하고, 다른 의료기관으로 해당 진료기록을 간편하게 보낼 수 있고, 자신이 겪는 질병에 대해 건강관리 서비스를 받을 수 있는 체계를 만드는 사업이다. 이 사업은 대통령직속 4차산업혁명위원회의 디지털헬스케어 특별위원회가 구성된 이래, 개인 주도 의료데이터(Personal Health Record) 활용 생태계 조성을 통해 의료분야 마이데이터 활성화 방안의 논의로 구체화되었다. 의료분야 마이데이터는 환자가 자신의 건강정보에 대한 자기결정권을 갖고 건강증진 혜택을 누릴 수 있고, 해당 데이터는 적재적소에 원하는 방식으로 활용될 수 있다는 데에 큰 의의를 갖는다. 이는 어느 한 기관이 의료데이터의 주권을 갖지 않고, 개인이 의료데이터 주권을 가질 수 있도록 API 형태로 제공되며, 어떤 사업체든 자유롭게 사업을 할 수 있는 생태계를 마련한다는 면에서 큰 의의가 있다. 데이터 보유기관은 본인 또는 데이터 활용기관으로 건강정보가 흘러가는 고속도로 역할을 수행한다.

[그림 4-2-3] 마이헬스웨이 개념도



* 출처: 관계부처 합동 보도자료 「마이 헬스웨이 구축 시작」, 2021.2.24.

이 사업의 3대 추진전략은 첫째, 건강정보 제공 항목의 순차적·단계적 확대, 둘째, 국민들이 안심하고 이용할 수 있는 안전한 플랫폼 구축, 셋째, 정부는 법·제도적 기반을 마련하고, 민간은 혁신적 서비스를 제공하는 역할을 분담하도록 하고 있다. 이를 위해 다음과 같은 4개 분야 12개 과제를 마련하였다.

[표 4-2-2] 마이 헬스웨이 4개 분야 12개 과제

1. 건강정보 수집체계 마련	2. 마이 헬스웨이 플랫폼 구축	3. 개인 주도 건강정보 활용지원	4. 의료분야 마이데이터 도입 기반 마련
- 데이터 유형별 수집항목 정의 - 플랫폼 제공 데이터 표준화 - 데이터 제공기관 참여 유인 방안 마련	- 플랫폼 공통 인프라 구축 - 사용자 인증·동의 체계 구현 - 데이터 연계 네트워크 구축	- '나의건강기록'앱 개발 - 활용서비스 연계·관리 방안 마련 - R&D를 통한 서비스 개발 지원	- 생태계 활성화를 위한 법·제도 개선 - 민·관 협업을 위한 거버넌스 구축 - 대국민 소통 전략 마련

* 출처: 대통령직속 4차위 디지털헬스케어특위, 「마이 헬스웨이 도입방안」, 2021.2.24.

마이헬스웨이 사업과 타 데이터사업의 근본적인 차이점은 식별 가능한 개인정보를 활용한다는 데 있다. 개인이 이 서비스를 이용하기 위해서는 가명정보가 아닌 개인정보를 직접 이용해야 하기 때문이다. 마이헬스웨이 사업이 성공하기 위해서는 충분한 제도적·재정적 뒷받침이 필요하다. 마이헬스웨이사업을 통해 진료기록을 발급할 경우 과금에 대한 비용 문제를 해결하기 위한 제도적 방법을 고려해야 하고, 의료정보분야 인프라에 투자하기 어려운 영세한 병원이 참여할 수 있도록 재정적 지원이 충분히 뒷받침되어야 한다.

라. 디지털 치료기기 사업

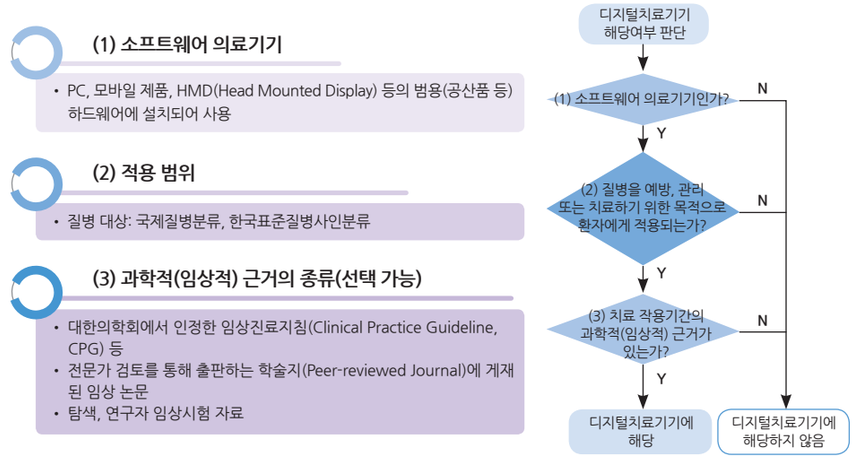
지금까지의 전통적인 환자 관리 및 치료는 진료 후에 의사가 환자에게 생활습관 개선을 위한 교육 및 약물을 처치하는 것이었다. 이러한 전통적 개념을 넘어서 최근에는 약물의 처방뿐만 아니라 심리적 및 행동학적 변화를 일으키게 하기 위한 기술들이 고안되고 있다. 바로 디지털 치료기기(Digital Therapeutics)이다. 디지털 치료기기 혹은 디지털 치료제는 의학적 장애나 질병을 예방, 관리, 치료하기 위해 환자에게 근거 기반의 치료적 개입을 제공하는 소프트웨어 의료기기로 정의한다. 즉, 스마트폰에 설치된 앱을 실행시켜 질환을 개선하거나 치료한다는 개념이다. 하드웨어에 의존적이지 않고 특정한 의료기기의 목적을 띠고 독립적인 형태의 소프트웨어만으로 이루어진 의료기기를 말한다. 여기에는 의료용 모바일 앱

뿐만 아니라, 가상·증강현실(VR·AR), 게임 등의 형태도 포함된다. 예를 들어, 우울증이나 경도인지장애 등 정신·신경계 질환의 예방 및 치료뿐만 아니라, 고혈압 및 당뇨 등 만성 질환의 치료 및 관리에도 적용될 수 있다.

디지털 치료기기를 개발하기 위해서는 기존의 모바일 어플리케이션을 개발하기 위한 다양한 정보통신 기술뿐만 아니라, 전통적인 약물 개발과 거의 유사하게 임상적인 유효성을 확보하기 위한 과정이 필요하다. 이를 위해 다양한 수준의 헬스케어데이터가 확보돼야 하고, 제작된 제품의 유효성 확인을 위해 데이터가 생산되기도 한다.

디지털치료기기 개발을 촉진하기 위해 2020년 5월 '의료기기산업 육성 및 혁신의료기기 지원법'이 시행됨에 따라 식품의약품안전처는 디지털치료기기를 포함하는 혁신의료기기의 신속한 허가심사를 지원하기 위해 디지털치료기기 허가·심사 가이드라인을 마련하였다. 이를 통해 디지털치료기기의 판단 기준, 제품 사례, 허가신청 자료 등에 대한 기초적인 가이드라인을 제공하고 있다.

[그림 4-2-4] 디지털치료기기 판단 기준



* 출처: 식품의약품안전처, 「디지털치료기기 허가·심사 가이드라인」, 2020.10.6.

또한 2021년 산업통상자원부에서는 기업과 병원의 디지털치료기기 개발지원을 위해 '디지털치료기기 산업 기술 개발·실증 및 상용화 지원'사업에 관한 제안요청서(RFP)를 내고 다음 표와 같은 디지털 치료기기 R&D프로젝트를 공모했다.

[표 4-2-3] 2021년 디지털 치료기기 R&D 프로젝트 공모

과제명
디지털치료기기 산업원천 기술 개발·실증 및 상용화 지원 연구
코로나 블루 시대 동반 우울증 디지털치료기기 기술 개발
공황 장애 치료와 연관 스트레스 관리 디지털치료기기 기술 개발
소아청소년 또는 성인 대상 섭식장애 디지털치료기기 기술 개발
금연 치료와 모니터링 디지털치료기기 기술 개발
기립성 또는 식후 유발성 저혈압 디지털치료기기 기술 개발

* 출처: 한국산업기술평가관리원, 「2021산업기술개발사업, 바이오산업기술개발 사업 RFP」, 2021.2.4.

또한 식품의약품안전처에서는 '디지털치료기기의 신속제품화 지원을 위한 안전성 및 성능 평가기술 개발 사업'을 통해 기초적인 가이드라인 제시 이후에 구체적인 안전성 및 성능 평가를 위한 가이드라인을 제시할 예정이다.

마. 감염병 연구개발 사업

코로나19가 유행하기 이전 2019년부터 정부는 메르스 등 신종·해외 유입 감염병의 유입 및 확산을 막기 위한 기술을 개발하기 위해 '방역연계 범부처 감염병연구개발 사업단'을 구성하여 7개 중점과제, 33개의 세부과제를 추진 중이었다. 이 연구개발 사업의 주된 목적은 감염병의 유입차단, 현장대응, 확산방지를 위한 기술을 개발하는 것이다. 유입차단을 위해 감염병 생물감시 마스터플랜 수립, 임상감시 네트워크 구축, 국내외 정보 수집 및 통합·분석을 통해 실시간으로 의사결정 등을 지원할 수 있는 한국형 통합생물감시체계를 구축한다. 현장대응에서는 감염병 환자 치료 및 확산 저지 작업을 위해 살처분 현장 및 의료현장 등 다양한 환경과 사용자 편의성 최적화로 국제 기준에 적합한 개인보호구 개발 등을 수행한다. 확산방지에서는 정보통신기술을 활용한 자가격리자 및 접촉자 관리 시스템의 평가방법 개발 및 모니터링 시스템을 개발한다. 특히, 이 사업단에서는 질병관리청의 긴급상황실(EOC)에 실시간으로 수집되는 데이터 관리를 위한 소프트웨어를 개발하여 신종 및 원인 불명 감염병 등의 임상 감시네트워크를 구축·운영할 계획이다. 또한 최근 '범부처 의료기기 연구개발 사업단'과 정보통신산업진흥원을 통해 코로나19를 극복하기 위한 데이터기반의 감염병 차단 기술에 관한 RFP를 내고 여러 민간 사업자가 선정되어 기술 개발 중에 있다.

[표 4-2-4] 2021 AI융합신규감염병 대응 시스템 개발

과제명	연구내용
AI기술을 활용한 감염병의 전파 관련 매개변수 분석	감염병 전파 관련 매개변수를 선정하고, 다양한 데이터로부터 AI기술을 활용한 상관 분석 알고리즘을 개발하여 감염병 매개변수를 도출
AI기술을 활용한 공간 내 감염원 전파 양상 분석	다양한 공간의 공기흐름 관련 실험 데이터와 유체역학 시뮬레이션에 AI기술을 적용하여 공간 내 감염 양상 분석기술 개발
AI활용 감염병 유행 및 완화 전략 예측 모델 개발 및 고도화	유행 예측 모델 개선·개발과 AI알고리즘 등의 활용
온라인 빅데이터 분석 모델 및 지리정보시스템(GIS) 시각화 기술 개발	온라인 공간에서 생산되는 비정형 빅데이터를 AI기술로 분석하여 유행 양상을 파악하고 시각화

* 출처: 정보통신산업진흥원, 「2021 AI융합신규감염병 대응 시스템 개발」, 2021.3.

4. 데이터 3법과 의료정보이용 가이드라인

2020년 국회를 통과한 데이터 3법 개정은 의료분야에 적용하기에 어려움이 있는데, 일반법을 의료분야에 적용하다 보니 특수법인 '의료법'과 '생명윤리 및 안전에 관한 법률'[2]을 우선 적용하기 때문이다. 이를 위해서 보건복지부 개인정보보호위원회에서는 보건의료데이터 분야의 유형별·목적별 세부기준과 절차를 제시하고, 거버넌스, 안전조치, 윤리사항 등에 관한 법적 안정성 도모를 위해 지난 2020년 9월에 '보건의료데이터 활용 가이드라인'을 배포하였다. 이 가이드라인의 핵심은 가명정보에 대한 명확한 기준을 두고 가명정보의 처리 및 결합에 대한 지정기준을 마련하였다는 데 있다. 또한, 여러 의료기관 및 공공기관에 분산 저장되어 있는 의료데이터를 결합하기 위한 구체적인 방법들을 명시하고 있다. 데이터의 내부·외부 반출 방법들에 관한 기술들을 제시하고 있고 데이터 결합을 위한 데이터심의위원회의 역할에 대해서도 기술하였다. 그동안 의료데이터 결합 주체에 관한 명확한 근거가 없었고, 구체적인 결합 방법에 관한 가이드라인이 없어, 보건복지 분야 빅데이터의 정보집합물 결합은 사회보장정보원이 맡고 있었다. 최근 의료분야에서는 국민건강보험공단, 건강보험심사평가원, 보건산업진흥원을 결합전문기관으로 지정하여 운영 중이다. 그리고 데이터결합에 대한 법적 근거도 명확하게 제시하여 가명정보에 관해서만 결합을 할 수 있도록 도모하였다.

하지만 아직 해결되지 않은 문제들이 있다. 병원에는 진료데이터 이외에 연구데이터도 많은데, 진료데이터는 '의료법'의 영역인 반면, 연구데이터는 '생명윤리 및 안전에 관한 법률'의 영역이라는 것이다. 연구에서는 두 가지 데이터를 혼용하는 경우가 많기 때문에 복지부 조직에서 이런 데이터 영역을 일원화하여 총괄 관리할 필요가 있다.

또 다른 문제는 유전체 데이터 및 영상데이터 등에 관한 가명화 처리 방안이 명확히 제시되어 있지 않다는 것이다. 즉, 유전체 데이터를 가명화하기 위한 기술이 명확하지 않은 상황에서 가명정보와 개인정보를 혼재해서 사용해야 하는 경우가 발생할 수 있다.

5. 향후 과제

COVID-19는 여러 산업의 디지털 혁신의 촉매제로서 작용하였다. 특히 국내 헬스케어분야에서는 원격의료의 현실적 반영, 감염병의 위험에 대한 기술적 차단과 방역, 공공의료의 질적 수준의 제고, 가정 방문 건강돌봄 등에 대한 요구가 증가하였다. 그리고 오늘날의 의료는 디지털 대전환의 기점에서 인공지능, 클라우드컴퓨팅, 사물인터넷, 모바일기술 등과의 기술적 융합을 도모하고 있다. 이러한 시대에 양질의 대규모 의료데이터를 체계적으로 수집 및 관리하는 것은 헬스케어데이터 산업 생태계를 만들기 위한 최우선 과제이다. 또한 수집된 데이터를 이용해 의료의 질적 향상을 위해 다양한 종류의 의료 인공지능이 개발되어야 한다. 결국 의료의 질적 향상을 도모하며 동시에 천문학적인 의료비를 줄이기 위한 방법을 데이터로부터 찾고 있다. 미래에는 지금과는 상상할 수 없는 수준의 의료데이터 자원을 이용한 의료행위가 이루어질 것이다. 따라서, 현재 정부가 추진 중인 다양한 헬스케어분야 데이터 R&D 과제 및 정책이 보다 체계화되길 기대한다. 아울러, 의료데이터 이용에서 산업계의 혼선을 방지하기 위해 보다 구체화된 데이터 사용 가이드라인이 마련되어야 할 필요가 있다.

2 생명윤리 및 안전에 관한 법률: 인간과 인체유래물 등을 연구하거나, 배아나 유전자 등을 취급할 때 인간의 존엄과 가치를 침해하거나 인체에 위해(危害)를 끼치는 것을 방지함으로써 생명윤리 및 안전을 확보하고 국민의 건강과 삶의 질 향상에 이바지함을 목적으로 함

제3장. 제조분야 데이터 활용

김창원
(건국대학교 신산업융합학과 겸임교수)

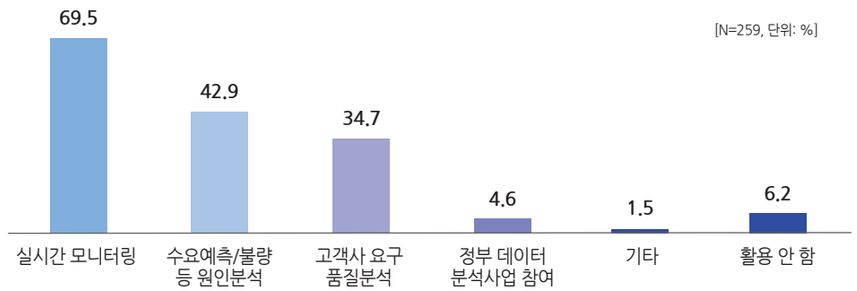
제조 현장에서 생산되는 제조데이터는 기업의 혁신을 가능하게 하는 핵심 자원이다. 업종별 스마트 공장 구현을 통해 다양한 제조데이터 활용에 대한 실증 사례를 찾아볼 수 있다. 스마트 공장에서 양질의 제조데이터를 확보하기 위해서는 데이터의 수준을 일치시키고 표준화할 수 있는 과정이 필요하다.

1. 제조데이터의 중요성

최근 제조 현장에서 생성되는 실시간 생산 데이터, 계측 데이터, 주문 데이터 등 다양한 제조데이터의 활용은 스마트 공장 및 스마트 제조 구현을 위한 핵심 자원으로 인식되고 있다. 제조데이터는 생산 관리 시스템(Manufacturing Execution System, MES), 전사적 자원 관리(Enterprise Resource Planning, ERP), 제품 수명 주기 관리(Product Lifecycle Management, PLM) 등 제조 IT 응용시스템에서 수집 활용되고 생산되는 데이터를 말한다. 이러한 제조데이터는 기업 간의 수평적 수직적 통합에 있어서 중요한 역할을 담당하게 된다. 제조데이터를 잘 활용하게 되면 이는 기업이 공정 혁신, 불량 예측, 신규 서비스 창출 등 새로운 부가가치를 창출할 수 있는 원동력이 될 수 있기 때문이다.

중소기업중앙회에 따르면 스마트 공장을 운영하는 기업이 제조데이터를 수집해 다양한 제조공정에 활용하고 있는 것으로 나타났다. 조사에 응답한 기업들은, 제조데이터가 실시간 모니터링(69.5%), 수요예측/불량 등 원인 분석(42.9%), 고객사가 요구하는 공정 품질 분석 자료(34.7%) 등에 활용되고 있다고 응답했다. 즉, 스마트 공장의 데이터가 단순히 쌓아두기만 하는 수준을 넘어 다양한 제조공정에 활용되고 있는 것으로 조사되었다.

[그림 4-3-1] 스마트공장 제조데이터 활용 형태

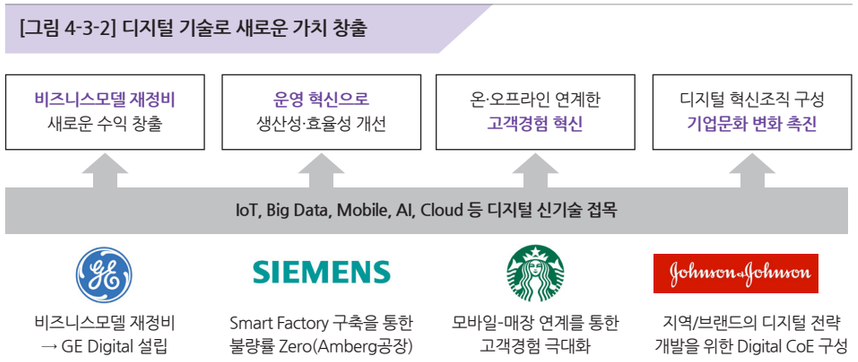


* 출처: 중소기업중앙회, 「스마트공장 제조데이터 활용실태 및 분석 수요조사」, 2020.6.22.

2. 제조데이터 실증사례

디지털 기술로 새로운 가치를 창출하는 혁신기업의 등장으로 기존 사업 경계가 붕괴되고 있다. 특히 글로벌 선진 기업에서 진행 중인 스마트화 및 디지털 전환은 신기술을 활용하여 비즈니스모델, 운영프로세스, 고객관리 등 조직 전반에 변화를 촉진하고 있다.

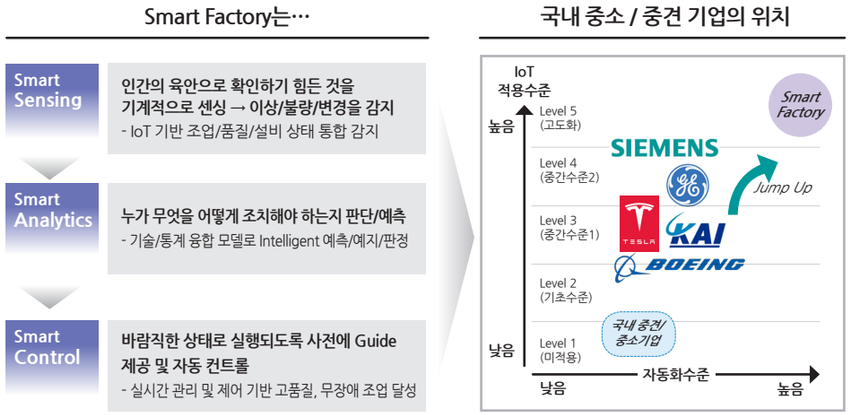
[그림 4-3-2] 디지털 기술로 새로운 가치 창출



* 출처: 저자 작성

디지털화의 핵심축인 스마트 공장은 '인간을 닮아가는 지능화된 미래형 공장'이다. 이에 이미 글로벌 선도 기업들은 사물인터넷 기술 기반으로 공장설비에서 산출되는 제조데이터 분석을 통해 지능형 공장으로 변화를 시도하고 있다.

[그림 4-3-3] 지능화된 미래형 공장 개념

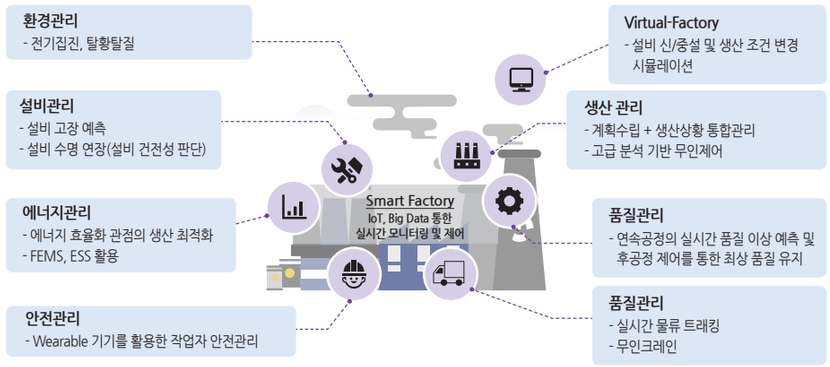


* 출처: 저자 작성

제조데이터를 핵심으로 하는 스마트 공장은 인간이 보고 듣고 느낀 것으로 판단·예측하고 알리고 행동하는 것과 같이 공장 스스로 센싱(Sensing)하고, 분석·예지하고(Analyze), 통제하는(Control) 지능형 공장을 일컫는다.

스마트 공장에서의 업종별 제조데이터 활용을 통한 실증사례를 살펴보면 다음과 같다. 제조데이터 실증사례는 중소기업기술정보진흥원 「클라우드 기반의 중소 제조데이터 수집 및 활용에 관한 연구(2020)」를 기반으로 하였다.

[그림 4-3-4] 스마트 공장 개념

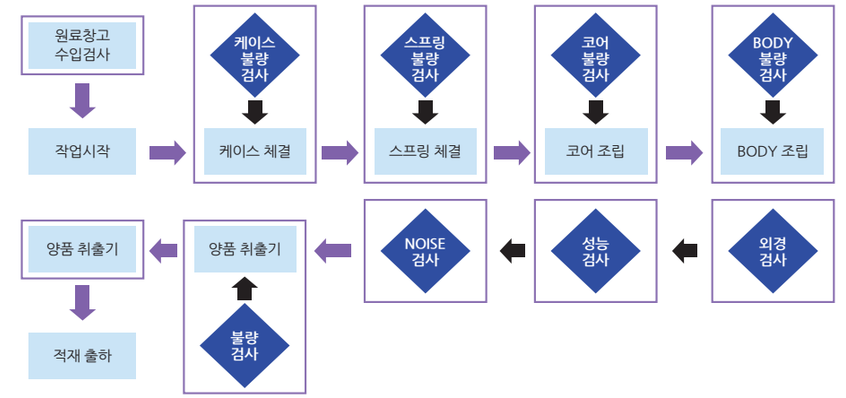


* 출처: 포스코ICT, 키움증권, 「포스코 ICT(022100)」, 2017.4.3.

• **기계부품 조립업종**

기계부품 조립업종 중 조립기, 소음기 등의 자동차 부품 생산공장에서 불량이 발생한 위치 파악 불가, 수기입력(Key-in) 방식의 생산·공정 관리, MES 연동 불가능 등의 문제를 해결하기 위해 데이터 수집 장치를 통해 작업환경·생산·작업 데이터 등 22가지 데이터를 분당 1회 수집한다. 이렇게 수집한 데이터를 로컬 서버 및 클라우드 내에 저장하였으며, 이러한 데이터를 기반으로 실시간 모니터링 시스템을 구축하여 불량률을 33.9% 낮추고 원가절감액을 130.4% 증대시켰다.

[그림 4-3-5] 공정흐름도 및 데이터 수집 포인트



* 출처: 중소기업기술정보진흥원, 「클라우드 기반의 중소 제조데이터 수집 및 활용에 관한 연구」, 2020

• **전자부품 조립업종**

전자부품 조립업종 중 정밀모터 조립라인에 대한 스마트화 참조 모델에 따라 12개 셀로 구성되어 있는 테스트베드 내 관할제어장치(Programmable Logic Controllers, PLC), 전력, MES 데이터 등 3가지 종류의 데이터를 수집하여 템플릿 기반 품질 분석을 하여 품질 모니터링 관리 방식을 구축하였다.

[표 4-3-1] 수집 데이터의 종류 및 특징

데이터 종류	데이터의 특징
모터 조립 설비 PLC 데이터	모터라인 조립 설비의 PLC에서 획득 가능한 데이터로서 공정별로 정의된 단계에 따라 데이터가 시계열 형태로 저장(저장데이터: 온도, 압력, 공정번호, 취득시간 등)
전력데이터	모터라인의 조립 설비 각각에 전력 미터기를 설치하여 각 설비에서 소모되는 전력 관련 데이터를 수집
MES 데이터	공정 데이터로부터 전달된 데이터를 MES의 공정시스템으로 연동된 Log 데이터로서, MES에서 정의된 4M(Man, Machine, Material, Method) 데이터와의 연계 상태 확인이 가능

* 출처: 중소기업기술정보진흥원, 「클라우드 기반의 중소 제조데이터 수집 및 활용에 관한 연구」, 2020

• 주조 설비 업종

주조 품질관리 업종 중 내부 품질 문제 인라인 검사 및 품질관리의 어려움을 해결하기 위해 금형 내에 센서를 설치하고 온도 등의 공정 데이터를 수집하여 4가지 기계학습 모델을 적용하여 예측치와 측정치를 비교 검증하였다. 적용 결과, 양품/탕경불량/기포불량 예측률이 93.8%로 정확하였고 생산성은 3.8% 증대하였다.[1]

[표 4-3-2] 주조 설비 주요 관리 항목 데이터

설비명	관리항목	관련공정	관리방식
중량저울	무게(kg)	용해, 주조	상, 하한 오차 한계 / 투입량/ 실시간 그래프
용해로	온도, 투입량	용해	실시간 온도 모니터링 / 투입량 / 실시간 그래프
다이케스팅	온도, 무게, 실적, 이형제, Gas	주조	실시간 온도 모니터링 / 상, 하한 오차 한계 / 불량 수 / 불량유형 / 실시간 그래프
보온로	온도	주조	실시간 온도 모니터링 / 실시간 그래프
펀칭기	Burr	트리밍	불량관리 / 불량유형관리
디버링기 샌드브러싱기	속도, 압력, 입방 수	사상	Rework / 실시간 그래프
열처리로	온도, 시간	열처리	실시간 온도 모니터링 / 실시간 그래프
성분분석기	성분검사	검사	Xbar-R 관리도, C_{pk}
3차원측정기	외관검사	검사	Xbar-R 관리도, C_{pk}

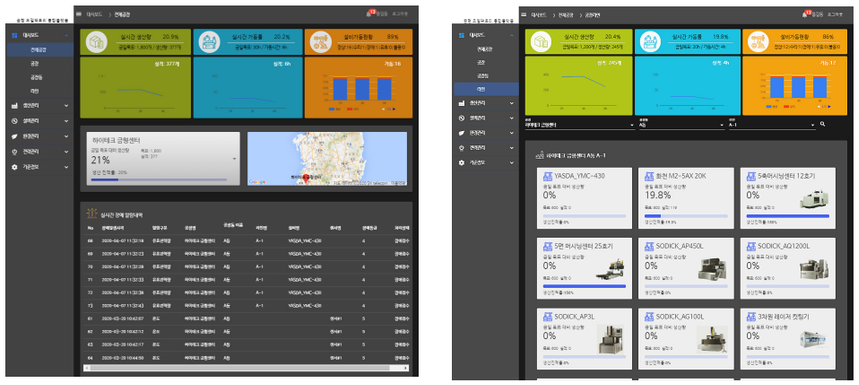
* 출처: 스마트공장추진단, 「스마트공장 참조모델 3차 개정증보판」, 2017.7.

1 중소기업기술정보진흥원(2020)

• 금형업종

금형업종 중 금형을 기반으로 한 제조업 전반의 경쟁력 저하를 해결하기 위하여 설비, 전력, 환경센서 데이터 관리가 필요하다. 이를 위해 IoT 게이트웨이와 IoT 플랫폼을 적용하여 대시보드를 구축하였다. 적용 결과, IoT 게이트웨이와 IoT 플랫폼을 통하여 설비데이터, 전력데이터, 환경센서데이터 등을 관리할 수 있게 되었다.

[그림 4-3-6] 대시보드 화면 구성 예시



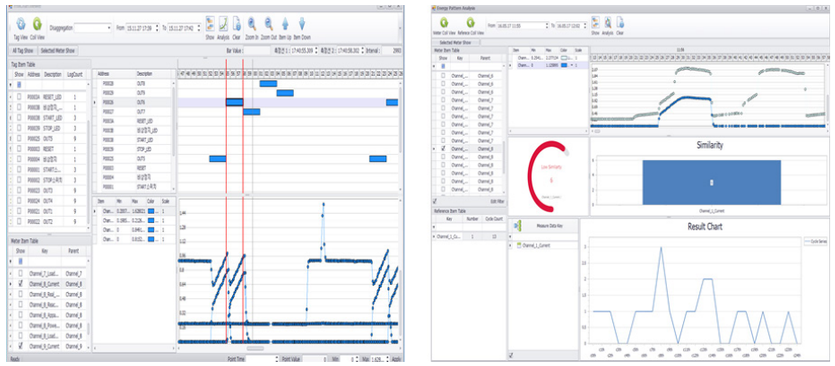
* 출처: 중소기업기술정보진흥원(2020)

• 열처리 업종

열처리 업종 중 열처리 공정 설비의 이상 탐지 및 고장 분석을 위하여 PLC와 OPC-UA[2] 기반 데이터 수집체계를 구축하고, 제어 구간별 전류 추이 분석을 딥러닝 기반으로 수행하였다. 그 결과로 나온 제어 구간별 전류 소모량 패턴을 가시화하여 이상 상태 탐지 기준으로 활용하였다.

2 설비 간의 통신을 위한 규약의 일종

[그림 4-3-7] 제어구간별 전류소모량 패턴의 가시화 및 분석 화면



* 출처: 중소기업기술정보진흥원, 「클라우드 기반의 중소 제조데이터 수집 및 활용에 관한 연구」, 2020

• **식품가공 업종**

식품가공 업종 중 수기로 기록한 문서들로 인해 전체 흐름에서 데이터 손실(Data Loss)이 발생하는 문제를 해결하기 위해 온도센서 및 스마트 저울을 활용하여 데이터를 자동 수집하여 클라우드 기반 서버에 저장하였다. 적용 결과 인위적 정보 조작을 방지하고, 제조 작업에 대한 전산화와 장비 이상 및 제품 생산 중 불량 발생 등에 관한 상황에 실시간으로 대응할 수 있는 체계를 구축하였다.[3]

[표 4-3-3] 식품가공 업종 내 육가공 공장에서 수집한 데이터 예시

순위	작업·장비 데이터	순위	생산 데이터	순위	작업환경 데이터
1	공정번호	1	투입 수량	1	미세먼지
2	장비번호	2	양념 투입 비율	2	가연성 가스
3	작업자	3	LOT No.	3	작업장 온도
4	가동시간	4	가열 온도		
5	가동속도	5	불량 수량		
6	가동상태	6	양품 수량		
7	생산모델명	7	하중불량		
8	에러	8	높이 치수 불량		
9	전력, 진동	9	외경 치수 불량		
10	온도				

* 출처: 중소기업기술정보진흥원, 「클라우드 기반의 중소 제조데이터 수집 및 활용에 관한 연구」, 2020

3 중소기업기술정보진흥원, 「클라우드 기반의 중소 제조데이터 수집 및 활용에 관한 연구」, 2020

• **패션 업종**

패션 업종 중 수기 기록을 통한 생산현황 일일 집계 문제를 해결하고자, 제조데이터와 IoT 데이터를 IIoS(Industrial Internet of Services) 플랫폼 내에서 관리하고 클라우드 기반 품질분석시스템을 활용하였다. 적용 결과, 모바일 디바이스를 활용하여 실시간 공정 및 품질 데이터를 수집하고 공정을 모니터링할 수 있었으며, 라인별 실시간 생산수량 확인을 위한 기능, 품질 불량 이력 확인 기능, 완제품 창고의 실시간 온 · 습도 모니터링 기능 등이 가능하게 되었다.

• **염색 업종**

염색 업종 중 염색 불량과 재염으로 인한 에너지 효율 및 생산성 저하 문제를 해결하기 위해 염색 공정에서 데이터를 수집 및 분석하여 작업을 표준화하고, 에너지에 대한 유의 변수들을 추출하여 효율적 운영을 달성하였다. 적용 결과 생산성 17% 향상과 에너지 사용량 13% 절감 효과를 실현하였다.[4]

3. 제조데이터 정책 방향

산업 현장에서 제조데이터는 업무 부서별 또는 비즈니스 영역별로 파편화되어 구축되어 있어 통합적 분석을 통한 인사이트 도출이 어려운 상황이다. 특히, 중소 제조기업의 발주/구매-생산, 관리-현장, 사람-기계 등 생산활동과 관련된 모든 데이터 수집 및 연결성 개선을 통한 운영 효율 고도화 지원이 필요한 실정이다. 제조 분야 데이터 활성화를 위한 업계 간담회(2020년 한국산업지능화협회)에서는 제조 분야 데이터는 보유하고 있더라도 쓸만한 데이터가 없는 실정이다. 처음부터 용도가 분명한 데이터를 수집해야 하지만, 제조기업들은 영업비밀 이슈로 데이터의 외부 공개를 기피하며 기업 간 업종 간 데이터 공유는 쉽지 않을 것이라는 전망 등의 의견이 나오고 있다.

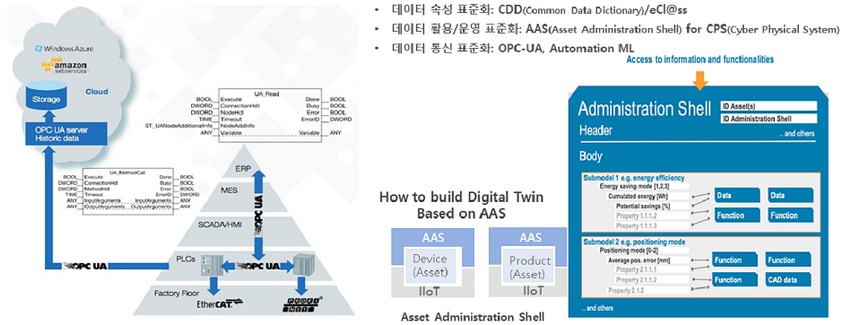
이를 해결하기 위해 제조현장의 4M1E(Man, Machine, Material, Method & Energy) 관점에서 비즈니스 프로세스 통합을 위한 제조데이터 수집 표준이 필요하다. 특히, 중소 제조기업은 수율, 불량률, 작업자 효율성(조립공정), 설비 가동률(장치산업), 에너지 소비 효율 등 제조 수준 및 특성에 맞는 핵심성과지표(Key performance indicator, KPI) 관리가 필요하다.

독일은 2012년부터 시스템 통합 능력이 곧 기업 경쟁력이라는 점을 강조하며 공장 최하위 레벨의 센터부터 최상위 레벨의 기업운영 시스템까지 통합 능력을 정립하기에 이르렀는

4 중소기업기술정보진흥원, 「클라우드 기반의 중소 제조데이터 수집 및 활용에 관한 연구」, 2020

데, 이른바 Industrie 4.0(I4.0)이다. 이는 4차산업혁명의 시초가 되었고, 현재 세계 전 산업에 큰 영향을 미치고 있다. 이에 우리나라에서는 스마트 공장의 개념, 용어, 진단모델에 대해 I4.0의 참조 모델인 RAMI4.0(Reference Architecture Model I4.0)이 반영된 국가표준 KSX-9001을 2016년 제정하였고, 이는 스마트 공장 수준 확인제의 근간이 되고 있다. 특히 I4.0에서 권고하는 국제표준이 적용된 생산 현장 데이터 수집 도구는 자산관리쉘(Asset Administration Shell, AAS)인데, 이는 제조현장의 수직통합을 위한 가장 적절한 대안으로 논의되고 있으며 이를 효과적으로 적용·확산하고 국내 제조 환경 적합성을 확보하기 위한 선제적 실증 작업이 요구되고 있다.

[그림 4-3-8] 제조데이터 수집을 위한 국제 표준 가이드



* 출처: 스마트제조혁신추진단(2020)

중소 제조 현장에서의 4차산업혁명은 빅데이터, 클라우드, AI, IoT와 같은 최신 기술을 적극적으로 활용하여 제조기업의 고도화를 이끄는 스마트 제조가 핵심이라고 할 수 있다. 스마트 제조에서 양질의 데이터를 확보하기 위해서는 데이터의 수준을 일치시키고 표준화할 수 있는 과정이 필요하며, 제조데이터 관리 가이드라인을 작성 및 배포하는 실행 방안 또한 필수적이라고 할 수 있다. 수집된 제조데이터를 체계적으로 수집·가공·분석하고 개방·공유의 선순환 구조를 구축하기 위해서는 체계적인 제조데이터 인프라 구축과 거버넌스 정책이 필요하다. 이를 통해 우리나라 중소 제조기업이 데이터 팩토리(Data Factory)의 경쟁력을 확보하고, 향후 AI Factory, Digital Twin Factory로 발전할 수 있기를 기대한다.

제4장. 유통·물류분야 데이터 활용

송성렬
(CJ대한통운 TES물류기술연구소 부장)

유통 시장이 오프라인 중심에서 온라인 중심으로 급속히 전환됨에 따라 이커머스 플랫폼 기업들이 물류영역으로 들어오고 온·오프라인 연계를 극대화하면서, 기업들은 시장에 대응하기 위해서 변화를 꾀하고 있다. 소비자들이 다 품종 소량 합포장 형태의 주문을 하며 더 빠른 배송을 요구함에 따라, 유통·물류업체들은 소비자 요구에 대응하기 위해서 맞춤 인프라를 갖추어야 한다. 현재 이커머스 시장에서는 이러한 고객 주문에 따라 물류센터에서 상품보관- 제품피킹-포장-배송에 이르기까지 전체 서비스를 담당하는 풀필먼트 서비스가 중요해지면서 AI, 자동화, 물류로봇, 빅데이터 기술을 갖춘 스마트 물류센터가 등장하고 있다. 따라서 풀필먼트 서비스를 위해서 데이터 기반의 효율적 인 풀필먼트 센터 운영과 출고 이후 최적의 배송을 위한 라스트마일에 대해서 알아보겠다.

1. 유통·물류 시장 변화

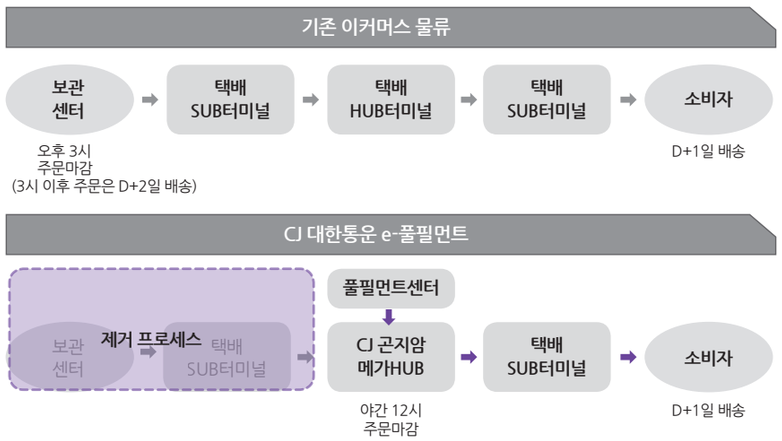
코로나 시대를 거치면서 이커머스가 급속히 성장했다. 이에 따른 소비 형태의 변화, 빠른 배송, 비대면 가속화와 같은 환경에 대응하고 시장을 선점하고자, 유통물류의 혁신이 가속화되고 있다. 이업종 간의 연합, 온·오프 중심 회사 간의 인수 또는 합병으로 유통업계의 기업들은 빠르게 높은 수준의 인프라를 갖추며 시장을 이끌고 있다.

이런 시대적 변화를 선도하기 위해서 업계는 주문한 상품이 물류창고를 거쳐 고객에게 배달 완료까지 자동화되는 풀필먼트 서비스 도입에 적극적일 것이며, 이러한 풀필먼트 시스템이 일반화될 것이다. 또한, 풀필먼트 센터 내 인력 부족 문제에 대한 대응과 해결 방안 중 하나인 자원 운영의 효율 향상을 극대화하기 위해서 물류로봇 도입도 가속화될 것이다.

가. 네이버 CJ대한통운 연합

IT 중심의 플랫폼 기업인 네이버는 CJ대한통운과 자사주 맞교환을 함으로써 네이버의 쇼핑·결제 시장 지배력과 CJ대한통운의 풀필먼트 운영 역량을 결합하여 한국판 아마존을 그리며 물류 영역으로 진입했다. CJ대한통운은 네이버 스마트스토어 주요 입점 상품을 처리하는 곤지암 풀필먼트 센터에 이어 군포와 용인에 풀필먼트 센터를 연이어 오픈할 계획이다. 군포·용인 센터는 상품 보관 온도에 따라 맞춤형 물류 서비스가 가능한 센터로 스마트스토어 입점 상품 중 의류, 생활용품, 가정간편식, 건강기능식품 등 소비재 상품을 보관하며 배송시간을 앞당기고 새벽배송, 당일배송 등 새로운 서비스를 소비자들에게 제공할 것이다.

[그림 4-4-1] CJ대한통운, 네이버쇼핑에 풀필먼트 서비스 개시



* 출처: CJ대한통운

일반 택배는 포장과 집화 등 작업 시간을 고려해 오후 3시에 주문이 마감되지만, CJ대한통운 풀필먼트 센터는 출고된 상품이 바로 1시간 거리에 있는 곤지암 메가 허브로 발송해 처리된다. 곤지암 메가 허브는 아시아 최대 규모의 물류 터미널로 CJ대한통운의 전국 범위의 택배 배송을 담당한다.

이에 네이버는 쿠팡의 '로켓배송' 같은 더 빠른 배송을 위해서 자사 커머스 상품 대상으로 '지정일 배송'과 '오늘 도착' 등이 포함된 배송 서비스 추진을 계획하고 있다. '오늘 도착' 서비스는 소비자가 네이버 스마트스토어, 브랜드스토어 등에서 오전 10시까지 주문한 상품이 당일 오후까지, 오후 2시까지 주문한 상품이 당일 저녁에 배송되는 것으로 계획하고 있다.

계획 중인 풀필먼트 센터는 AI 수요예측, 물류로봇, 자동화 분류 설비 등 디지털 기술이 대거 적용된 스마트 물류 체계를 실험할 수 있는 센터가 될 것이고, 촘촘한 물류망을 통한 라스트마일 경쟁력을 추가로 갖추어 대형 고객사부터 중소상공인까지 서비스 제공을 확대할 계획이다.

나. 신세계, 이베이 코리아 인수

신세계는 '미래 유통은 온라인 강자만이 살아남을 것'이라며, 이베이 코리아 인수에 있어서 단순히 기업을 사는 것이 아니라 시간과 기회를 사는 딜(Deal)이라고 표현했다. 이번 인수는 신세계 그룹의 사업 구조가 오프라인에서 온라인으로 전환된 것을 의미하며, 이베이 코리아 인수로 이마트는 온라인 비중이 약 50%를 차지할 정도로 사업 포트폴리오와 디지털 전환을 더욱 더 가속화할 방침이며 온-오프 통합 국내 1등 유통 사업자를 목표로 하고 있다.

이미 첨단 풀필먼트 센터를 보유한 신세계 그룹은 오프라인 거점을 온라인 물류 전진기지로 활용해 물류 경쟁력을 극대화하려고 하며, 온라인 풀필먼트 센터에 향후 4년간 1조원 이상을 집중 투자할 계획이다.

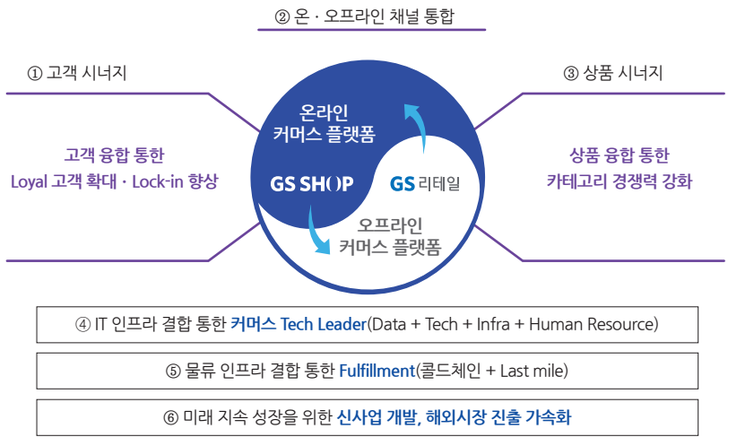
빠르게 성장하고 갈수록 경쟁이 치열해지는 이커머스 시장에서 이베이 코리아를 인수한 신세계는 투자 시간과 비용을 줄여서 단시간 내 고객과 판매자를 확보하게 되고 플랫폼 및 풀필먼트 운영에 있어서 IT 인력까지 확보하게 되는 것이다. 단기간 내 자원확보를 통해 규모의 경제를 이루어 사업을 신속히 성장시키기 위한 전략으로 볼 수 있다.

다. GS리테일 · GS홈쇼핑 합병

강자인 네이버, 쿠팡에 이어 롯데, 이마트와 같은 대형 유통업체들과의 생존 경쟁에 대응하기 위해서 GS리테일과 GS홈쇼핑이 합병해 올해 7월 1일 통합GS리테일이 출범했다.

이날 임시 주총에서 허연수 GS리테일 부회장은 "디지털 커머스를 중심으로 온라인과 오프라인을 연계하고, 온·오프라인 구분 없이 소비자들이 모든 쇼핑 니즈를 해결할 수 있는 토털 솔루션을 제공해 온·오프라인 통합 커머스 플랫폼 리딩기업으로 거듭나겠다"고 선언했다.

[그림 4-4-2] 통합GS리테일 사업 비전



* 출처: 유안타증권, 「Sector Report」, 2020.11.11.

통합GS리테일은 원가를 낮추고자 상품 통합 구매를 고려하고 있으며, 고객 통합으로 공통 마케팅에 나설 계획이다. 양사의 중복 회원 600만명을 제외한 통합회원은 2,600만명으로, 네이버, 쿠팡, 이마트에 맞설 수 있다. 통합GS리테일은 통합 온라인몰 '마켓포'를 론칭하고 GS숍은 빠른 배송 경쟁력 강화로 ㈜매쉬코리아 지분도 확보하며 다회전 당일배송과 즉시배송으로 고객에게 접근할 계획을 가지고 있다.

신축 물류센터를 포함해서 전국 60개 물류센터망과 3,300여 대 차량, 2,200여 명의 인력으로 초대형 물류 기업을 목표로 하고 있으며, 편의점 GS25, 슈퍼마켓 GS더프레시의 배송 물량을 소화할 퀵커머스 플랫폼과 도심형 마이크로 풀필먼트를 준비하고 있다. 통합GS리테일은 사업 간 시너지를 극대화하고 신성장 동력을 확보하기 위해서 투자를 결정했다.

2. 시장 대응을 위한 데이터 기반 필요 기술

회사 간 인수·합병이 일어나는 유통·물류 시장에서 변화에 대응하고 최종 소비자의 마음을 사로잡기 위해서는, 데이터 중심으로 즉시 판단하며 일사불란하게 운영될 수 있는 높은 수준의 인프라 기반의 풀필먼트가 필요하다. 대표적으로 미국 아마존 풀필먼트는 인공지능, 로봇, 빅데이터 등 첨단 기술로 무료·빠른 배송, 24시간 즉시 환불·재발송 등과 같은 서비스로 소비자를 사로잡았다고 볼 수 있다.

풀필먼트 서비스를 운영하는 데 있어 필요한 기술을 보면, 고객 주문을 미출고 없이 원활하게 처리하기 위한 수요예측 기술과 센터 내 운영 생산성 확보를 위한 설비 및 로봇의 자동화가 중요하고, 그 다음으로는 출고된 고객 상품을 당일배송, 익일배송과 같이 원하는 시간에 맞추어 배달하기 위해 촘촘한 네트워크망에서 라스트마일 관점의 경로를 최적화하는 기술이 필요하다. 따라서 데이터 중심의 주요한 기술 키워드 3가지인 수요예측, 센터 내 설비 자동화, 라스트마일에 대해서 살펴보겠다.

가. 물량과 이벤트 정보 분석을 통한 수요 예상 및 대처

이커머스 시장이 급성장함에 따라 업체 간의 치열한 경쟁이 지속되고 물류센터 내 원가는 상승함에 따라 업계에서 고민하는 부분이 있다.

제조·유통업체의 경우는 재고가 없어서 고객들이 주문한 상품이 출고되지 못하는 상황이 발생하지 않도록, 특정 이벤트 행사에도 늘어날 출고량을 감안해서 풀필먼트 센터 내 상품을 보관해야 한다. 미출고로 최종 고객이 원하는 일자에 상품을 받지 못하면 서비스 만족도가 떨어져서 궁극에는 고객 이탈을 초래할 수 있다. 그렇다면 물류의 경우, 많은 주문량으로 인하여 작업 인력이 부족하면 미출고가 발생되고, 반면 주문량이 적은 경우 과다한 작업 인력이 투입된다면 비용이 증가할 수 있다. 작업 인력의 원가 상승에 따라 미출고 없이 비용 효율화를 높이기 위한 대처가 필요하다.

이와 같이 재고관리와 비용 효율화를 위해서 상품 수요를 정확히 예측하고 관리하는 것이 기업들에게는 가장 중요한 과제이다. 제조·유통기업은 이벤트 상품, 할인율, 기간 등을 고려하여 당일 또는 익일 판매량을 예측하여 사전에 재고를 비축하여 미출고 없이 최종 고객 주문에 대응해야 하고, 물류업체는 상품 이벤트에 따른 주문량을 예측하여 사전 출고 작업 등으로 작업 인력의 생산성을 높여서 원가를 줄여야만 한다. 이 모든 것의 기반은 이벤트를 고려한 제조·유통기업의 주문량 예측에서 출발하기 때문에 주문량 예측이 중요한 것이다.

제조·유통기업이 직접 운영하는 물류의 경우는 예측된 당일·익일 물동량으로 재고 관리와 인력 효율화를 할 수 있지만, 물류를 위탁한 경우에 물류업체는 고객사의 이벤트에 따른 당일·익일 물량을 예측하기 어려우므로 제조·유통기업과 협업하여 예측된 주문량으로 계획적으로 재고를 운영하고 작업 인력을 관리해야 한다.

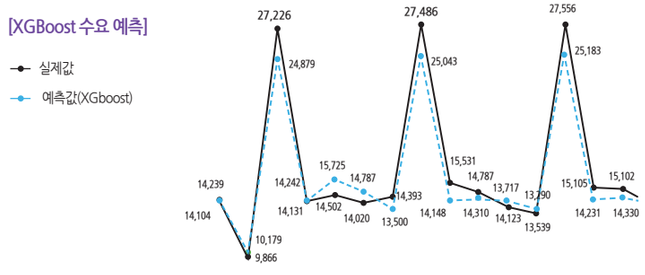
유통과 물류를 같이하는 아마존도 클라우드 서비스인 AWS와 인공지능 기술을 활용하여

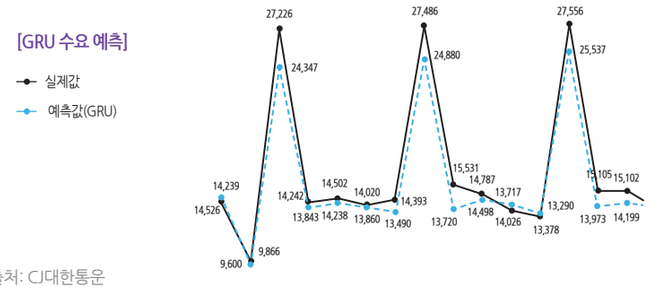
방대한 고객 정보 데이터를 분석했으며, 이를 통해 고객의 수요를 예측하고 고객이 주문하기 전에 배송을 미리 준비하는 시스템을 구축했다. 또한 미국 이베이는 인공지능 기술을 토대로 판매량을 예측하고 잠재수요를 파악하는 기술력을 획득하고자 이스라엘 스타트업 세일즈프레딕트(SalesPredict)를 인수하기도 했다. 이와 같이 해외 유통사들도 인공지능을 활용한 수요 예측에 집중하고 있다.

국내의 경우를 보면, 네이버는 자체 개발한 물류 수요 예측 AI인 '클로바 포캐스트'를 CJ대한통운의 군포와 용인센터에 적용할 예정이며, 네이버 쇼핑 주문량을 하루 전에 예측하여 물류센터의 인력 배치와 운영 효율화에 활용하고 있다.

또한, CJ대한통운은 인력의 효율적인 비용 관리를 위해서 물류창고를 운영하는 D2C(Direct to Consumer) 고객사가 상품을 판매하는 주요 쇼핑몰의 당일 이벤트 정보를 수집하고 과거 주문 정보와 결합하여 수요 예측 모형을 개발하고 있으며, 이를 현장에서 테스트하며 고도화하고 있다. 주문량 예측뿐만 아니라 추가적으로 상온과 저온으로 구분한 주문량, 물류센터 내 운영을 위해 단일 품목 또는 이종 품목의 합포에 대한 주문량을 예측하여 좀 더 정교한 인력 수급에 활용하고 있다.

[그림 4-4-3] AI 알고리즘 XGBoost와 GRU를 활용한 수요 예측





* 출처: CJ대한통운

나. 최적의 센터 운영을 위한 설비 자동화

최근 이커머스 산업의 발달로 인하여 창고의 역할은 단순히 물건을 보관하는 역할에서 구매자가 원하는 조합으로 분류 및 재포장, 출고까지 점차 확장되고 있다. 또한 오프라인 시장이 활발하던 과거의 단일 품목, 대용량 단위의 출고에서 현재는 창고와 소비자를 직접 연결해주는 온라인 플랫폼의 보급으로 인해 복수 품목, 소용량 단위 위주의 복잡한 과정을 통한 출고 형태로 변했다. 이와 같은 다품종 소량 합포장에 대응할 수 있는 이커머스 전용 창고에 대한 수요가 증가하면서 복잡한 분류 작업을 자동화할 수 있는 설비 및 로봇에 대한 요구가 늘고 있다. 독일의 시장조사 전문기업 Statista에 따르면, 창고 자동화 시장은 2021년 156억 달러에서 2026년에는 두 배가 증가한 302억 달러로 증가할 전망이다[1].

단일 품목의 대용량 단위로 창고에 입고된 상품을 고객이 원하는 소품목으로 분류 및 재포장하기 위해서, 현재 QPS(Quick Picking System), DPS(Digital Picking System)의 성능을 강화한 CJ대한통운 피킹 시스템(Picking System), QAS (Quick Assorting System, DAS(Digital Assorting System)의 성능을 강화한 CJ대한통운 분배 시스템(Assorting System) 등과 같은 분류 자동화 설비들이 운영되고 있을 뿐만 아니라 종이 박스를 자동으로 접어주는 제함기, 포장을 자동화하는 자동 포장기, 포장 검수를 자동화할 수 있는 중량 검수기 등에 대한 보급이 활발하게 이루어지고 있다. 이런 자동화 설비의 성능을 최대한으로 사용하기 위해서는 설비 로그 데이터 분석 및 주문량 예측 등을 활용하여 상시적으로 프로세스 최적화를 진행해야 한다. 또한 자동화 설비의 예지 정비(Predictive Maintenance) 및 오류율 개선에는 설비 로그뿐만 아니라 창고 환경 데이터 등 다양한 빅데이터가 활용되고 있다.

로봇을 활용한 창고 시스템은 제품을 구매하는 소비자에게 다양한 혜택을 보장할 수 있는데, 기업에서는 창고의 24시간 무중단 운영이 가능해지면서 인력에 의존하던 과거와는 달리 생산성 향상과 비용 절감을 보장할 수 있다. 이렇듯 유통사가 고객의 주문을 실시간으로 처리함으로써 양질의 서비스를 제공할 수 있을 뿐만 아니라 야간근무에 대한 비용 또한 절감할 수 있어서 기업의 부담을 덜 수 있다는 장점이 있다. 창고는 주로 보관하는 물품 종류에 따라 레이아웃의 복잡도가 결정되는데, 단순한 구조 레이아웃에는 바코드 기반의 무인운반로봇(Automated Guided Vehicle, AGV)이 주로 사용되었지만, 다양한 상품을 취급하는 창고에 대한 수요가 증가됨에 따라 자율 주행 기반 무인 지게차와 자율이동로봇(Autonomous Mobile Robot, AMR) 또한 보급이 증가하고 있다.

또한 물류창고는 특성상 중량이 무거운 상품을 취급하기 때문에 상품을 운반하기 위한 지

1 Statista, 「Size of the warehouse automation market worldwide from 2012 to 2026」, 2020.10.

게차와 입·출고를 위한 대형 화물차량들이 상시 이동하고 있어서 안전에 대한 관심도가 매우 높다. 이 때문에 이미 일부 창고에서는 로봇들만 이동할 수 있는 무인 구역 운영을 목표로 하고 있다.

대표적인 물류로봇으로는 아마존의 오더 피킹용 로봇 Kiva가 있으며 주문에 맞는 상품을 창고 안에서 이동시키는 기능을 갖고 있다. 아마존에 따르면 현재 20여만 대의 로봇이 미국 내 다수의 창고에서 운영되고 있으며 그 수가 지속적으로 증가하고 있다고 한다. 국내에서도 물류기업들이 로봇 도입을 검토하고 있다. 대표적으로 CJ대한통운이 여러 대의 무인 로봇을 일부 센터에 도입하여 활용하고 있고 지속적으로 확대할 계획이다.

[그림 4-4-4] 대표적인 물류 로봇

[아마존 AGV 키바(Kiva)]



[Boston AMR]



* 출처: 아마존, Boston Dynamics

물류 로봇은 제조 로봇과는 다르게 특정 공간을 이동하고 다양한 작업을 수행하기 때문에, 각 로봇은 자동화된 프로세스를 통해서 작업을 지시받으며 창고 내에서 발생하는 모든 상황에 대한 데이터가 실시간으로 수집되어 서로 공유되어야 한다. 이를 위해 대용량 데이터 기반 최적화가 요구되며, 많은 물류기업 도입을 검토할 때에는 단순히 로봇의 기능뿐만 아니라 로봇 운용을 위한 데이터 분석 및 활용 기술에 대한 검토가 동시에 이루어져야 할 것이다.

다. 라스트마일을 위한 경로 최적화

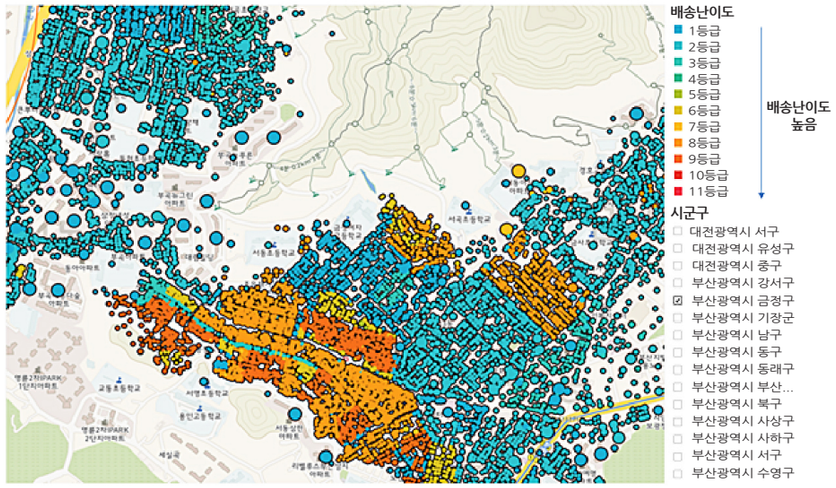
유통·물류에서 최종 고객의 만족도를 높이는 최우선의 척도가 배송서비스라고 볼 수 있다. 업계는 배송에 대한 다양한 시도들을 하고 있으며 대표적으로는 국내 택배 산업의 익일 배송뿐만 아니라 최근에는 일부 당일 주문에 대한 새벽배송과 같이 실시간에 가까운 라스트마일 배송서비스를 제공하여 고객 유치에 힘쓰고 있는 상황이다.

과거 라스트마일 배송서비스와 다르게 최근에는 유통전문회사와 물류전문회사가 서로 합작하여 시너지를 극대화하는 방식으로 발전하고 있다. 대표적인 사례로는 네이버, 마켓컬리 등과 같은 대형 온라인 쇼핑몰과 물류 전문기업인 CJ대한통운의 합작 서비스가 있다. 차세대 이커머스 배송서비스는 집화 처리 및 재분류 과정 없이 소비자에게 직접 배송하는 구조로 이루어지기 때문에 속도 측면에서 차별화된 배송서비스를 제공할 수 있다는 장점이 있다.

차별화된 라스트마일 서비스를 제공하기 위해서는 소비자의 주문에 따라 실시간으로 배송 경로를 변화시켜 반영할 수 있는 데이터 및 최적화 기술력이 보장되어야 한다. 배송지역에 대한 객관화된 배송난이도에 기반하여 각 배송 기사가 배송을 처리할 수 있도록 자원관리를 해야 한다. 기존 택배 배송은 고정된 지역 내 고정 인력을 통해 배송하는 구조였지만, 모든 상황이 가변적인 라스트마일 서비스는 배송난이도에 따른 자원관리가 이루어지지 않을 경우에 특정 담당자에게 과도한 업무량이 부과되는 상황이 발생할 수 있다.

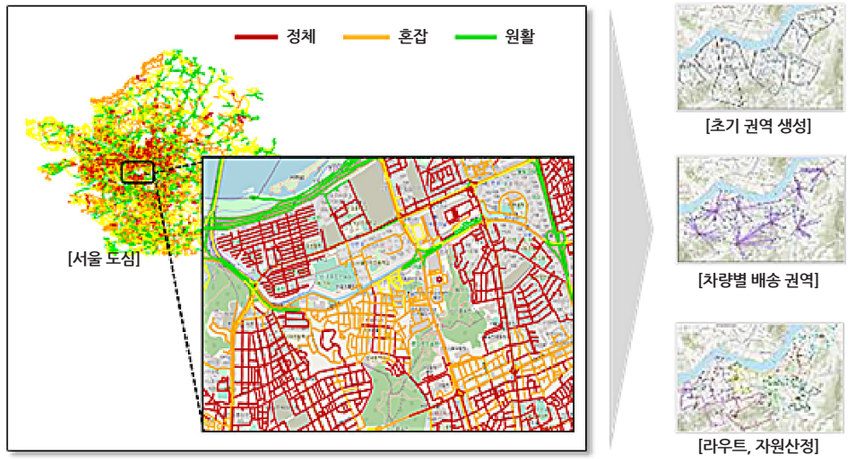
지역별 배송난이도 정의를 위해서는 건물정보, GIS 데이터, 도로망 데이터 등의 정보를 활용하는데, 예를 들면 각 건물 층수, 엘리베이터 유·무, 권역별 건물 밀집도, 거주 세대 수, 건물용도, 도로 넓이와 같은 건물 주변 정보 등이 있다. 이러한 데이터를 다양한 형태로 조합하여 산출한 각 건물 혹은 지역별 배송난이도를 배송 시 필요한 총 배송시간 예측에 활용함으로써 특정 배송기사에게 과도한 배송을 요구하는 상황을 피할 수 있으며, 회사와 배송기사가 서로 만족할 수 있는 적정한 수준의 배송 스케쥴을 정의할 수 있다.

[그림 4-4-5] 건물별 배송난이도 예시



* 출처: CJ대한통운

[그림 4-4-6] 실시간 교통정보 서비스를 활용한 라스트마일 배송 경로 자원 산정



* 출처: CJ대한통운

기존의 택배 배송은 배송 지역 전담을 통해서 일정한 배송 경로를 유지하는 형태이기 때문에 배송 경로는 배송 기사의 장시간 경험에 기반하고 있다. 하지만 라스트마일 서비스는 매일 새로운 패턴의 배송 목적지가 지정되고 배송 지역도 변하기 때문에 경험에 기반하여 배송이 이루어지기 힘들다는 특징이 있다. 이러한 환경에서 양질의 서비스를 고객에게 제공하기 위해 실시간 교통정보 데이터, 각 배송지 및 배송 환경에 대한 배송난이도 등의 정보로 최적의 배송 경로를 배송기사에게 제공하고 있다.

3. 시사점

이커머스 성장과 함께 고객 니즈도 고도화됨에 따라, 유통업계의 물류 운영이 질적 성장을 해야만 하는 도전에 직면해 있다.

빠른 배송과 신속 출고를 위해서 재고 보관량이 증가함에 따라 보관 효율의 극대화도 필요하고 다품종 소량 합포장이 증가함에 따라 물류 작업의 난이도가 높아져 오분류로 인한 오출고가 증가하게 된다. 이를 해결하기 위해서 물류 작업 노동의 부담을 줄이고 인력 안전과 자동 검수를 위해서 물류로봇 및 자동화 설비가 급속히 도입될 것이다.

코로나 시대 이후, 풀필먼트 도입, 데이터에 근거한 센터 운영, 설비와 로봇에 의한 센터 운영은 기본이 될 것이며, 풀필먼트 인프라 구축 이외에도 센터에서 출발하여 최종 고객까지 도착하는 라스트마일 배송은 원하는 상품을 빠르게 받아보는 구매 경험의 가치로 대두되기 때문에 배송 네트워크 설계 또한 매우 중요한 물류 역량 중 하나가 될 것이다. 빠른 배송을 위해서는 풀필먼트 센터의 거점이 도심 인근에 있어야 하므로 원가 상승에 따른 생산성을 높이기 위해서 운영 최적화가 상시 이루어져야 한다.

유통·물류기업들은 현재 변하지 않으면 안 되는 절체절명의 상황에 놓여있다. 고객의 마음을 사로잡기 위해서 상품의 품질도 중요하지만, 고객이 주문한 상품을 기다림 없이 빨리 배송받아 만족스러운 경험을 갖게 해야 한다. 이를 위해서 사전 물량 예측, 재고 준비, 빠른 출고, 빠른 배송을 위한 인프라와 서비스를 만들어야만 하고 과감한 투자를 지속해야 한다.

제5장. 광고분야 데이터 활용

최홍규
(한국교육방송공사 연구위원)

2021년 기준 광고 비즈니스 분야는 전 세계적으로 '온라인'과 '모바일'의 키워드로 수렴한다. 코로나19 시기를 거치면서 비대면의 일상이 확대되고 미디어 이용량이 늘어나면서 스마트 기기를 통한 광고 소비가 더욱 늘어났기 때문이다. 한편 동영상 서비스 시장 정착, 블로그를 통한 개인 기록물 증가, 버추얼인플루언서를 통한 소통 증대, 메타버스 서비스를 통한 마케팅 확대 등의 현상은 광고 데이터 비즈니스에 더욱 활력을 주고 있다.

1. 데이터 기반 광고시장의 성장 추이

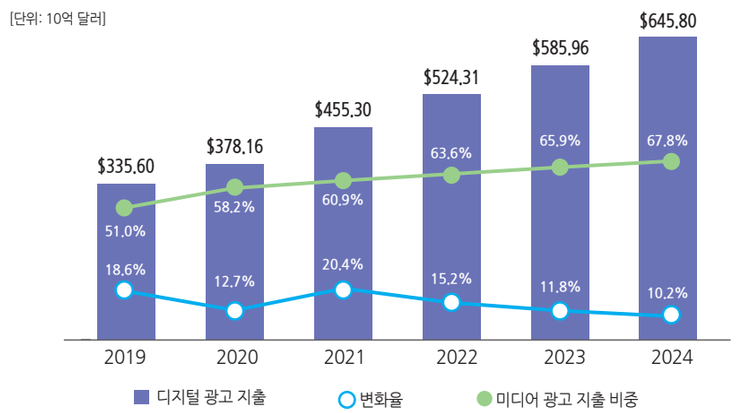
전 세계 데이터 기반 광고시장은 디지털 광고를 중심으로 성장세가 뚜렷하다. 국내의 경우 온라인 광고시장 매출액 비율이 압도적으로 높고 모바일 중심의 성장세가 지속하고 있다.

가. 전 세계 광고시장 규모

인터넷 등장 이후 그 규모가 지속적으로 증가하고 있는 디지털 광고는, 온라인에서의 이용자 행동 데이터 분석과 대용량의 빅데이터 활용이 가능하다. 이 때문에 기존 아날로그 방식의 광고와 차별적인 방식으로 다양한 포맷의 광고를 집행할 수 있어 매년 디지털 광고에 대한 수요는 늘어나는 추세다.

전 세계 디지털 광고시장의 지출 규모로 살펴볼 때, 스마트폰의 등장과 함께 2010년대 이후에는 미디어 광고가 기존 디지털 광고시장을 대체하고 있다고 볼 수 있다. 수치상으로도 전 세계 디지털 광고시장 지출 규모는 2019년 이후 매년 10% 이상 증가하여 2024년에 이르러서는 6,450억 달러 규모로 성장할 것이 예상된다. 전체 디지털 광고 지출 규모에서 미디어 광고의 비중은 매년 50% 이상을 유지할 것으로 예측되며, 2024년에는 67.8%에 달할 것으로 예측된다.

[그림 4-5-1] 전 세계 디지털 광고시장 지출 규모 전망치



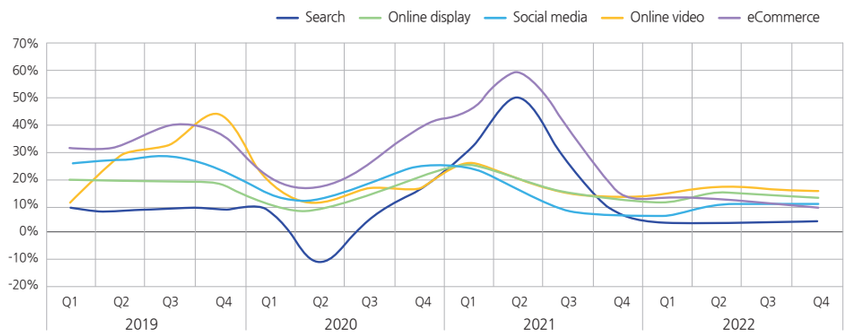
* 출처: eMarketer, 「Worldwide Digital Ad Spending 2021」, 2021.4.29.

전 세계 디지털 광고의 유형별 투자 추이를 살펴보면, 이커머스 광고는 2020년 1분기 이후에 온라인 동영상 광고를 추월하나, 2022년에 접어들어 다시 온라인 동영상 광고가 가장 많은 투자를 유치할 것으로 보인다. 2021년 1분기 이후 코로나19의 영향세가 점차 감소하는 시기에는 검색 광고가 온라인 동영상 광고를 추월하는 양상을 보이겠지만 2021년 말 이후에는 디지털 광고시장에서 검색광고에 대한 투자액이 가장 작을 것으로 예측된다.

점차 코로나19 이후를 대비하는 시기로 접어든 2021년 2분기 이후 전 세계 광고의 투자액은 전반적인 하향세로 돌아설 가능성이 있다. 이는 코로나19 이후 스마트 기기를 통한 미디어 콘텐츠 소비량이 다소 줄어들 수 있어, 이에 따른 여파를 원인으로 꼽을 수 있다.

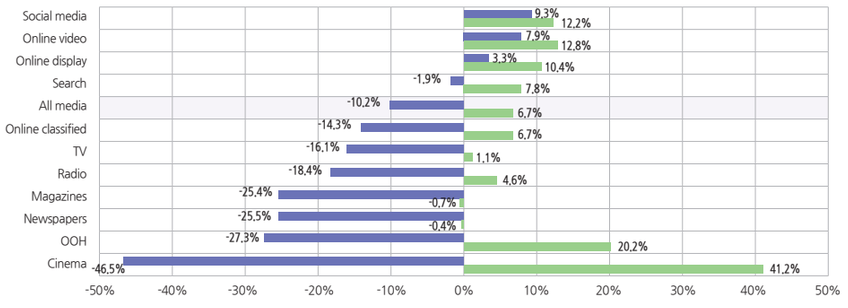
광고 데이터가 비즈니스 영역에서 활용되어 더 많은 부가가치를 창출하려면 온라인 공간에서 더 많은 광고 데이터가 축적되고 이것이 상품 판매에 영향을 미치는 정보로 가치를 지녀야 한다. 전 세계적으로 코로나19가 확산된 2020년과 코로나19의 확산이 점차 안정세로 접어든 2021년을 비교해보면, 2021년에 온라인 공간에서 더욱 광고 데이터 확산 여건이 마련되는 현상을 보인다고 할 수 있다. 2021년 소셜 미디어, 온라인 비디오, 온라인 디스플레이 영역 등에 지출되는 광고비는 더욱 증가하고 있다. 그러나 TV, 라디오, 잡지, 신문 등 전통적인 미디어 영역의 광고 투자액은 전년 대비 소폭의 증가세만을 기록하거나 여전히 마이너스 성장률을 보이고 있다.

[그림 4-5-2] 전 세계 유형별 디지털 광고의 분기별 투자 추이



* 출처: WARC, 「Global Ad Trends: Ad Investment 2021/22」, 2021.7.

[그림 4-5-3] 전 세계 광고 영역별 투자 전망: 2020 vs. 2021

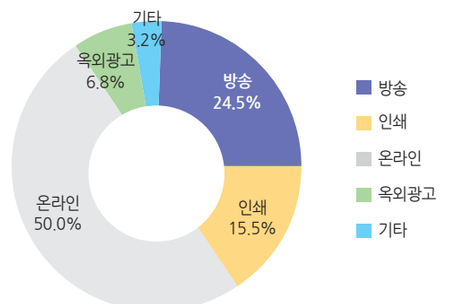


* 출처: WARC, 「Global Ad Trends: State of The Industy 2020/21」, 2020.

나. 국내 광고시장 규모

2020년 국내 광고시장 매출액 비중을 살펴보면, 온라인 광고가 50%를 차지하며 이는 방송(24.5%), 인쇄(15.5%), 옥외광고(6.8%) 등을 압도하는 수치다. 국내에서도 온라인 공간에서 광고 데이터가 생성되거나 이를 활용할 여건이 갖춰지고 있다고 볼 수 있다. 특히 광고 데이터를 전산화하는 데 상대적으로 취약할 수 있는 인쇄와 옥외광고 비중은 여타 영역에 비해 낮은 비율을 보이는 것이 특징이다.

[그림 4-5-4] 국내 광고시장 매출액 비중(2020년(E))



* 출처: 과학기술정보통신부, 한국방송광고진흥공사, 「2020 방송통신광고비 조사 보고서」, 2020.12.

국내 온라인 광고시장을 좀 더 세부적으로 살펴보면 고정형 PC에서의 광고시장 규모는 점차 감소하거나 성장 정체기에 접어든 반면, 모바일 광고시장은 성장을 이어가고 있다. 2020년 모바일 영역에서 전년에 비해 디스플레이광고는 19%, 검색광고는 18.6%의 성장이 전망되며, 2021년에는 디스플레이광고 11.3%, 검색광고 11.6%의 성장을 이어갈 것으로 관측된다. 모바일 광고시장이 지속적으로 성장하고 있으나, 전년 대비 증가율은 낮아지고 있어 2021년 이후에도 두 자릿수 성장이 이어질 수 있을지는 지켜봐야 한다.

[표 4-5-1] 국내 온라인 광고시장 변화 추이 전망

[단위: 백만 원, %]

매체 유형	세부 광고유형	'14년	'15년 (전년대비 증감률)	'16년 (전년대비 증감률)	'17년 (전년대비 증감률)	'18년 (전년대비 증감률)	'19년 (전년대비 증감률)	'20년(E) (전년대비 증감률)	'21년(E) (전년대비 증감률)
온라인		3,050,949	3,427,814 (12.4%)	4,154,724 (21.2%)	4,775,137 (14.9%)	5,717,205 (19.7%)	6,521,929 (14.1%)	7,273,275 (11.5%)	7,956,911 (9.4%)
PC		2,141,046	2,053,373 (-4.1%)	2,173,087 (5.8%)	1,909,192 (-12.1%)	2,055,449 (7.7%)	1,871,643 (-8.9%)	1,748,841 (-6.6%)	1,799,048 (2.9%)
	디스플레이 광고	720,795	775,198 (7.5%)	1,022,403 (31.9%)	904,407 (-11.5%)	965,785 (6.8%)	919,198 (-4.8%)	860,709 (-6.4%)	896,955 (4.2%)
	검색광고	1,420,252	1,278,174 (-10.0%)	1,150,684 (-10.0%)	1,004,785 (-12.7%)	1,089,664 (8.4%)	952,445 (-12.6%)	888,132 (-6.8%)	902,093 (1.6%)
모바일		909,903	1,374,442 (51.1%)	1,981,637 (44.2%)	2,865,945 (44.6%)	3,661,755 (27.8%)	4,650,286 (27.0%)	5,524,434 (11.5%)	6,157,863 (9.4%)
	디스플레이 광고	415,214	752,135 (81.1%)	1,002,117 (33.2%)	1,519,253 (51.6%)	1,839,956 (21.1%)	2,353,106 (27.9%)	2,800,817 (19.0%)	3,117,954 (11.3%)
	검색광고	494,689	622,307 (25.8%)	979,521 (57.4%)	1,346,693 (37.5%)	1,821,799 (35.3%)	2,297,180 (26.1%)	2,723,617 (18.6%)	3,039,909 (11.6%)

* 출처: 과학기술정보통신부, 한국방송광고진흥공사, 「2020 방송통신광고비 조사 보고서」, 2020.12.

2. 코로나19 이후 광고 데이터 비즈니스 이슈

2020년 광고 데이터 시장에 영향을 미친 이슈는 여러 가지가 있지만, 그중에서 동영상 서비스 시장의 정착, 라이프 블로그를 통한 개인 기록물 증가, 버추얼인플루언서를 통한 소통 증대, 메타버스 서비스를 통한 마케팅 확대 등을 핵심적인 이슈로 꼽을 수 있다. 이들 이슈는 코로나19를 거치며 미디어 이용행태 변화를 수반하거나, 광고시장에 새로운 활력을 불어넣을 수 있는 요소로 여겨진다.

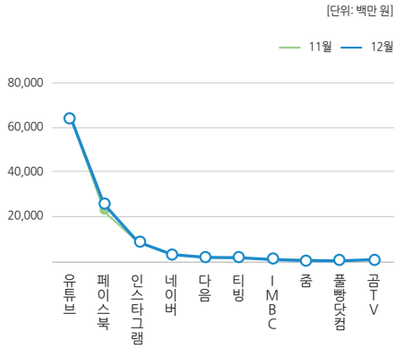
가. 동영상 서비스 시장 정착

2020년 12월 기준 국내 온라인 동영상 광고비는 유튜브, 페이스북, 인스타그램, 네이버, 다음, 티빙 순이다. 유튜브에서의 광고비는 전월 대비 2.7%의 감소율을 보였지만 여전히 시장에서 가장 많은 광고비가 집행되었다. 소통형 소셜미디어 서비스로 시작한 페이스북과 인스타그램의 광고비가 각각 2, 3위를 차지하고 있다는 점도 눈에 띄는 지점이다. 코로나19 시기를 거치면서 유튜브를 중심으로 한 동영상 광고시장의 점유율이 높아졌고 비대면 상황으로 인해 소통형 미디어를 매개로 한 광고의 비중이 높아지고 있다는 점을 보여준다. 이러한 소통형의 동영상은 이용자의 참여를 토대로 서비스 이용이 활성화되므로 광고에 활용될 수 있는 이용자 데이터 생성을 더욱 촉진할 수 있다.

[그림 4-5-5] 2020년 12월 기준 온라인 동영상 광고비(좌측) 및 추이(우측)

[단위: 천 원]

순위	매체	11월	12월	증감률
1	유튜브	65,496,472	63,755,931	-2.7%
2	페이스북	23,156,595	26,344,395	13.8%
3	인스타그램	10,049,850	8,954,334	-10.9%
4	네이버	2,943,936	2,853,576	-3.1%
5	다음	1,904,958	2,040,390	7.1%
6	티빙	1,720,606	2,031,542	18.1%
7	IMBC	892,328	832,496	-6.7%
8	쿰	856,776	824,022	-3.8%
9	풀빵닷컴	615,735	687,609	11.7%
10	곰TV	679,709	639,307	-5.9%

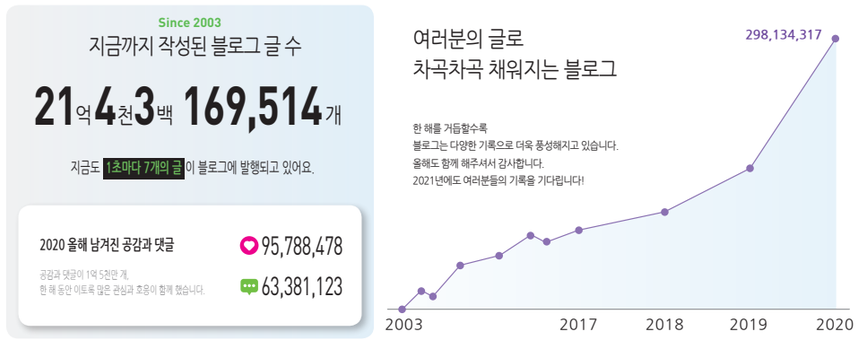


* 출처: 리서치애드, 「2020년 12월 온라인 동영상광고 현황 보고서」, 2021.1.

나. 라이프 블로그를 통한 개인 기록물 증가

닐슨 코리안클릭에 따르면, 기록용 SNS인 네이버 블로그는 2019년 4분기에서 2021년 2분기까지 평균적으로 250만 명에서 280만 명 수준의 이용자 추이를 보이는 것으로 나타난다. 대면 활동보다 집에 머무는 시간이 많아지고 개인이 온라인에 정보를 공유하는 것이 일상화되면서 나타나는 현상이다. 개개인의 라이프가 하나의 정보로 블로그 소재로 활용되며 이를 통해 블로그가 온라인에서 주요 광고 집행 공간으로 부각되고 있다. 개인의 기록물에 포함된 정보들은 광고 상품의 특성과 매개하여 광고효과를 극대화하는 데이터로 기능하기도 한다.

[그림 4-5-6] 네이버 게시글 수 추이(2003~2020년)



* 출처: 네이버, 「2020 네이버 블로그 리포트」, 2020.

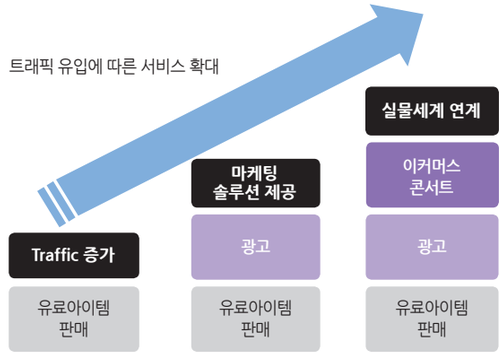
다. 버추얼인플루언서를 통한 소통 증대

2020년부터 광고시장에서 부각한 현상 중 하나는 디지털 휴먼이다. 광고시장에서 가상의 인물이 광고모델로 등장하는가 하면, 이들이 소셜미디어에서 버푸얼인플루언서로 활약하며 광고 소비자들의 관심을 모았다. 이러한 버추얼인플루언서는 전 세계적으로 활발히 활동 중인데 대표적으로 국내에서는 신한라이프 광고에 등장한 '로지(Rozy)', 미국에서 샤넬과 캘빈 클라인 등 명품 브랜드 모델로 활동하고 있는 '릴 미켈라(Lil Miquela)', 일본에서 이케아 광고모델로 유명세를 얻은 '이마(Imma)'가 있다. 이들 모두는 가상의 인물이며 광고모델로 시작해 소셜미디어 인플루언서로 활약 중이라는 공통점이 있다. 많은 팔로어 수를 자랑하는 만큼 이들 광고모델에 대한 팬덤 데이터가 향후 광고시장에서 큰 영향을 미칠 것으로 판단된다.

라. 메타버스 서비스를 통한 마케팅 확대

광고 데이터 비즈니스 영역에서 빼놓을 수 없는 시장이 바로 메타버스 시장이다. 유니버스(UNIVERSE), 제페토(ZEPETO), 점프(Jump), 로블록스(Roblox), 마인크래프트(Minecraft), 포트나이트(Fortnite)까지 다양한 서비스가 부각하며 광고시장에서 새로운 기대감을 불러일으키고 있다. 메타버스에서는 가상의 옥외광고, 콘서트 광고, PPL, 3D 네이티브 광고 등 현실 세계의 광고 유형을 모두 시도할 수 있다. 따라서 현재 메타버스 서비스에서는 다양한 마케팅 솔루션을 통한 광고가 주요 수익원으로 자리 잡는 추세다. 메타버스에서는 자기 정체성을 가진 캐릭터로 이용자 남녀노소가 활동할 수 있어, 이용자의 활동 데이터를 보다 다면적으로 분석하여 효율적인 광고 집행이 가능하다는 특징도 있다.

[그림 4-5-7] 메타버스 수익모델의 진화 단계



* 출처: KB증권, 「메타버스, 디지털 평행세계」, 2021.3.8.

3. 향후 전망 및 기대

2020년 이후 광고 데이터 비즈니스는 온라인 공간에서 디지털화된 정보를 토대로 하는 광고시장이 더욱 조명을 받는 추세로 가고 있다. 따라서 광고 콘텐츠를 소비하는 이용자 정보를 어떠한 방식으로 수집하여 더욱 효과적인 광고를 집행할 수 있는가가 광고시장의 주요 과제로 부각할 것으로 본다. 전 세계 광고시장이 모바일 네트워크를 통한 온라인 광고시장으로 수렴되고 있고, 이용자 개인의 데이터를 토대로 하는 소통형 서비스가 확산되고 있으며, 가상의 인물과 공간에 대한 수요가 늘어나는 만큼 광고 데이터를 활용한 비즈니스는 더욱 다양한 전략들을 모색할 것으로 전망한다.

제6장. 모빌리티분야 데이터 활용

차두원(차두원모빌리티연구소 소장)
금효영(슈퍼브 에이아이 이사)

첨단 산업과 서비스의 집합체로 많은 관심을 받고 있는 모빌리티 산업의 핵심 요소는 바로 데이터다. 자율주행 자동차 인공지능 학습데이터뿐만 아니라, 커넥티드카 증가에 따른 모네타이제이션, 자율주행과 커넥티드 기술 발전에 따른 카커머스와 카페이먼트 시스템 등 모빌리티 산업에 있어 데이터는 기존 자동차 관련 하드웨어와 소프트웨어 기업들의 모빌리티 트랜스포메이션(Mobility Transformation)을 가능하게 하고 새로운 스타트업들의 등장을 촉진하고 있다. 이에 이 장에서는 모빌리티 산업에서의 주요 데이터 활용 사례와 향후 전망을 기술한다.

1. 모빌리티 산업

가. 모빌리티의 정의

최근 모빌리티 산업은 국내뿐만 아니라, 세계 시장에서도 신산업이자 기술 융합과 투자를 이끄는 견인차 역할로 많은 관심을 받고 있다. 사람과 물건의 이동을 위한 하드웨어, 소프트웨어, 서비스뿐만 아니라 더욱 높은 품질과 정교한 서비스, 새로운 사용자 경험과 높은 만족도를 얻기 위해 사물인터넷, 인공지능, 클라우드 컴퓨팅, 빅데이터, 이동통신, 블록체인 등 다양한 첨단 기술들이 통합된 새로운 움직이는 플랫폼 산업으로 자리매김하고 있기 때문이다. 그만큼 국가, 기업 간의 경쟁도 심화되고 있으며, 관련된 규제, 일자리 문제 등 사회에 미치는 영향도 중요한 이슈로 다루어지고 있다.

하지만 아직 명확하고 이해 주체 간에 합의된 산업분류나 정의가 없는 것이 현실이기도 하다. 현재 국회에 계류 중인 '모빌리티 활성화 및 지원에 관한 법률안'에서 포함된 모빌리티의 정의는 사람과 물건의 이동에 관련된 수단, 인프라, 서비스 등 관련된 모든 과정에 필수적인 소프트웨어, 하드웨어, 인프라, 서비스 등 거의 모든 것들을 포함하는 광범위한 개념으로 해석할 수 있다.

모빌리티란 지역 내에서 사람 또는 물건을 한 장소에서 다른 장소로 이동하거나 운송하는 행위, 기능 또는 과정으로서 이와 관련한 수단, 기반시설 및 일련의 서비스를 통하여 확보할 수 있는 다음 각 항목의 이동성을 말한다.[1]

가. 한 장소에서 주요 교통망 또는 교통거점까지의 접근 이동성

나. 주요 교통망 또는 교통거점에서 다른 장소까지의 연계 이동성

다. 지역 내 서로 다른 장소 사이의 상호 이동성

또한 국내에서는 출발지에서 목적지까지 서비스 측면을 강조하는 MaaS(Mobility as a Service)라는 개념이 광범위하게 모빌리티 업계 및 정책 분야에서 사용되고 있다.

다양한 이동수단 활용을 위한 효과적인 '접근성(accessibility)'과 이들을 효과적으로 연계하는 '연결성(connectivity)'을 강조하며 공급자 중심 개념 측면이 강한 것도 공공주도 성격이 강한 국내 모빌리티 정책의 특징 가운데 하나로 볼 수 있다.

2020년 6월 4일 혁신성장전략회의에서 발표한 '한국형 도심항공교통(K-UAM) 로드맵'에서도 UAM을 기존 버스, 택시, 철도, 퍼스널 모빌리티가 혼합된, 연결성이 강조된 'Seamless MaaS'에 포함된 하나의 연계 이동수단 관점으로 해석하기도 했다.

반면 우리나라보다 비교적 이른 시점에 MaaS 개념과 기업들이 등장한 해외에서는 주로 수요자 중심의 온디맨드 서비스를 강조하고 있다. MaaS를 '온디맨드로 접근할 수 있는 단일 모빌리티 서비스들의 통합'으로 정의하며[2], 이동수단보다는 서비스 관점, 출발지에서 목적지까지 다양한 이동수단 정보와 하나의 결제 수단으로 한 번에 요금을 지불하는 '통합(integration)'을 강조하는 것이 특징으로 볼 수 있다.

나. 모빌리티 서비스와 데이터

모빌리티를 표방하는 업체들이 제공하는 서비스도 매우 다양하다. 모빌리티 업체를 표방하는 기업들을 이동수단 관점에서 구분해 보면 자전거, 전기자전거, 전동킥보드, 이륜차, 택시, 지하철과 버스 등 대중교통수단, 완성차, 자율주행차뿐만 아니라, 최근 많은 관심을 받는 도심항공모빌리티도 포함된다.

서비스 관점에서 보면 위의 이동수단의 호출, 공유, 구독, 렌털, 온디맨드 등의 유형과 이들을 연결해 주는 MaaS 서비스 제공 기업들도 있다. 전기차 원년으로 인정받는 2021년은 전기차 에코시스템 구축을 위한 충전 및 서비스 인프라 확장 등과 관련된 정보 및 서비스를 제공하며 전기차 생태계를 구성하는 모든 업체들 역시 모빌리티 기업으로 스스로를 정의하고 있다.

특히 코로나 19가 확산하면서 재택근무 확대 등에 따른 라이프스타일의 변화는 물류와 배송 산업 규모를 확대하였고, 딜리버리 로봇, 무인 배송, 이륜차 등을 이용한 라스트마일 배송 등의 업체들도 모빌리티 산업군으로 포함하는 등 'MaaS=TaaS(Transportation as a Service)+LaaS(Logistics as a Service)' 형태로 모빌리티 산업 범위는 점차 확대 추세에 있다.

이러한 트렌드와 모빌리티 기업을 표방하는 업체들의 비즈니스 모델들을 취합하면 서비스 관점에서 모빌리티는 '디바이스 다양성, 손쉬운 접근성, 새로운 사용자 경험을 제공하며 안전한 이동을 위한 모든 서비스'로 정의할 수 있다. 즉 지상에서 하늘까지, 실내에서 실외까지, 위에서 언급한 다양한 이동수단을 활용해 스마트폰을 통한 온디맨드 혹은 구독모델 등 다양한 접근 방법으로 끊임없는 연결과 통합 등 새로운 경험을 제공하는 서비스를 통칭한다고 볼 수 있다. 이러한 서비스 중에서도, 이동수단의 사고 없는 이동 과정은 매우 중요하다. 하지만 코로나 19 확산 이후 전통적 사고에 대한 안전과 함께 코로나 19 감염 및 확산 방지 등 바이러스 안전도 중요한 모빌리티 서비스 기본 요건으로 자리 잡고 있다.

하지만 이러한 서비스 제공을 위해 모빌리티 산업의 가장 중요한 자원 가운데 하나는 바로 데이터다. 자율주행 자동차 운행을 가능하게 하는 인공지능 학습을 위한 영상데이터, 각종 호출서비스와 퍼스널 모빌리티에서 출발지와 목적지까지의 경로데이터, 커넥티드카의 차량과 사용자데이터 등은 다양한 새로운 서비스의 창출과 신규 일자리, 무엇보다 도시의 이동 패턴과 특성을 파악할 수 있어 그 가치와 활용에 대한 잠재력은 점차 높아지고 있다.

2. 모빌리티 분야의 데이터 활용 사례

가. 자율주행 자동차

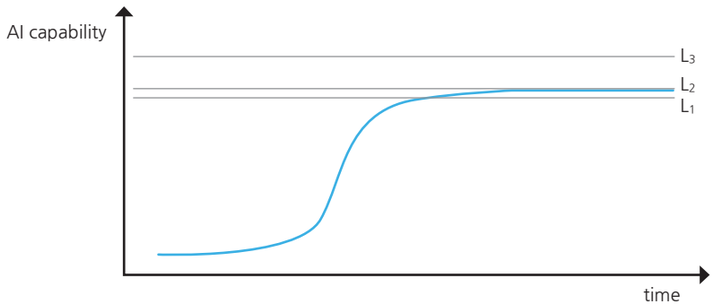
2020년 5월 미국의 대표적 자율주행 트럭 스타트업인 스타스키 로보틱스가 폐업을 선언했다. 트럭 외부에 장착된 6개 카메라에서 전송되는 실시간 영상을 보며 비디오 게임과 같이 차량을 원격조종하는 기술은 자율주행기술과 함께 스타스키 로보틱스가 자랑하는 기술로, 고속도로에서는 자율주행기술을 활용하고, 물류센터나 배송거점으로 이동하는 복잡한 시내에서는 원격조종 기술을 활용해 안전한 이동을 위한 기술과 비즈니스 모델을 개발했던 기업이다. 이러한 성과로 2019년 미국 CNBC에서 전 세계 스타트업을 대상으로 선정하는 2019

1 「모빌리티 활성화 및 지원에 관한 법률안」(박상혁 의원 대표 발의), 2020.9.18.
2 Transit Protocol MEDIUM, 「What is Mobility as a Service?」, 2019.7.19.

Upstart 100, 프라이트웨이브가 선정하는 FreightTech 25 가운데 12위로 선정되기도 했다. 이렇듯 자율주행 트럭 업계에서 가장 주목받았던 기업이었기 때문에 스타스키 로보틱스의 폐업이 자율주행업계에 던지는 충격이 작지 않았다.

당시 공동창업자 스테판 셀츠 아크마허(Stefan Seltz-Axmacher)는 인공지능 관련 블로그 포스팅에서 "인공지능 기술이 급격히 발전했지만, 자율주행을 위한 인공지능이 인간 운전자 수준을 능가했다면 자율주행기술을 개발하는 기업들은 상용화를 위한 안전성만 증명하면 되고, 수준이 유사하다면 안전성을 확보하기 위한 기술 개발을 위한 투자가 필요하다. 하지만 실제로는 아직 인공지능이 인간의 수준에 미치지 못하기 때문에([그림 4-6-1]의 L3) 완벽한 자율주행기술 개발이 쉽지 않다."고 주장했다. [그림 4-6-1]과 같이 인공지능 기술 수준(파란색 라인)은 최근 급격한 속도로 기술이 발전하며 자율주행 자동차에 적용되고 있다. 하지만 자동차가 스스로 안전한 주행을 담당하려면 최소 인간 운전 수준과 동일하거나(L2: 인공지능과 동일한 인간 운전 수준) 인간의 운전 수준을 능가해야 하지만(L1: 인공지능보다 낮은 인간의 운전 수준), 현재 인공지능 기술 수준은 인간 운전 수준보다 낮기 때문이라는 것이다(L3: 인공지능 기술보다 높은 인간 운전 수준).

[그림 4-6-1] 슈퍼바이즈드 머신러닝 vs. 인간 운전자 수준 비교

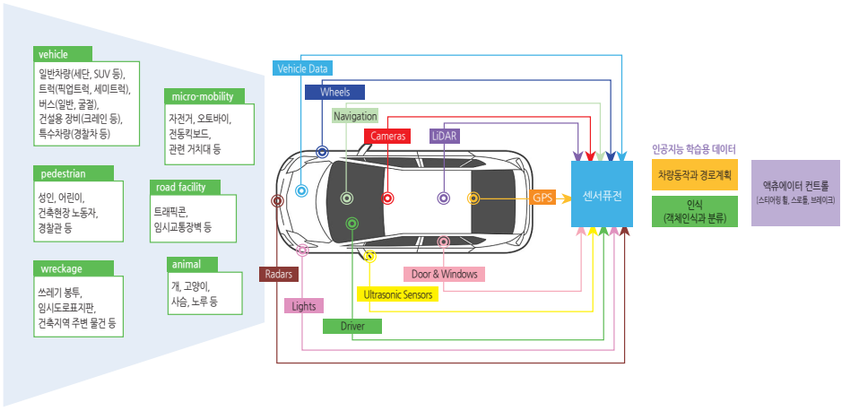


* 출처: 스타스키 로보틱스 블로그[3]

인공지능은 빅데이터와 머신러닝 기술이 없으면 구현되기 어렵기 때문에 인공지능 학습 재료로서 빅데이터 확보는 매우 중요하다. 자율주행 자동차가 주행하는 환경에 존재하는 모든 객체의 이미지를 다양한 상황별로 확보해 이미지 데이터셋을 생성하고 인공지능이 이를 학습해 자율주행 자동차의 안전하고 정확한 주행을 위한 기본 데이터로 활용한다.

절차를 살펴보면 수집된 사진데이터에서 개인정보보호를 위한 익명화 과정, 날씨, 상황, 객체 유형에 따른 데이터를 추출하고, 추출된 데이터를 라벨링과 주석 작업을 통해 구분 가능한 명확한 데이터로 분류하며 속성을 입력하는 어노테이션(annotation) 작업을 거쳐 인공지능의 기반 정보로 사용한다. 자율주행 자동차 조작을 위한 인지 → 판단 → 제어 과정에서 특히 정확한 인지를 통한 안전 확보를 위해서는 풍부한 고품질의 학습데이터와 어노테이션은 매우 중요한 기본 작업이다.

[그림 4-6-2] 자율주행 데이터 소스 및 인공지능 적용 절차



* 출처: 「자율주행차 상용화를 위한 데이터 이슈」, 한국지능정보사회진흥원, 「AI DATA INSIGHT 12호」, 2021.8.

자율주행기술을 개발하는 거의 모든 업체들에 인공지능 학습데이터 확보 및 어노테이션은 필수 과정이다. 일반적인 자율주행기술 개발 기업들이 라이다(LiDar)와 HD맵을 사용하는 반면, 2021년 6월부터 북미 지역에서 판매하는 모델3와 모델Y에서 레이더(Radar)를 제거한 테슬라는 많은 우려에도 불구하고 카메라 8개만 사용해 자율주행 기능을 제공하기 시작했다. 현재 전 세계에서 운행되고 있는 테슬라 차량은 100만 대 이상이며, 테슬라는 실시간으로 차량과 주행 관련 모든 데이터를 전송받아 수집하고 있다. 현재 완성차 업체들 가운데 가장 많은 데이터, 특히 오토파일럿 관련 데이터를 수집하고 있어 테슬라를 데이터 플랫폼이라고도 부르는 이유기도 하다. 하지만 이와 같이 인공지능을 강조하는 테슬라에도 더욱 정교한 인공지능의 판단을 위해 약 1,000명의 어노테이션 담당자가 소속된 것으로 알려져

3 https://medium.com/starsky-robotics-blog/the-end-of-starsky-robotics-acb8a6a8a5f5 (접속일: 2021.11.2.)

서 자율주행차 안전을 위한 빅데이터 수집과 처리의 중요성은 매우 높다고 할 수 있다[4].

나. 커넥티드카와 데이터 모네타이제이션(Data Monetization)

1) 커넥티드카의 차량 데이터

2014년부터 자동차에 내장형 셀룰러 모뎀(Embedded Cellular Modem)이 본격적으로 장착되기 시작하면서 자동차와 인터넷의 연결이 확산되기 시작했다. ABI Research(2020.7.28.)에 따르면 코로나 19로 차량 판매가 19% 감소한 상황에서 2020년 글로벌 차량 판매의 약 41%인 약 3000만 대에 셀룰러 모뎀이 내장되었다. 지역별로 살펴보면 각 지역의 신차 판매량 가운데 미국은 91%, 아시아 태평양은 51%, 라틴 아메리카는 37%가 커넥티드였던 것으로 분석했다. 하지만 2020년 15% 축소된 커넥티드카 시장은 2021년부터는 회복세로 접어들어 2023년 75억 달러, 2025년에는 1억 1,500만대의 커넥티드카가 판매되고 시장 규모는 830억 달러로 성장할 것으로 예측했다[5]. 보다 긍정적으로는, 마켓츠앤마켓츠(MarketsandMarkets)에서 2020년 539억 달러에서 2020년부터 2025년까지 연평균성장률(CAGR) 25.2%를 기록하며 1,660억 달러 규모로 성장할 것으로 분석하기도 했다[6].

10인치에서 30인치에 이르는 대형 인포테인먼트 디스플레이를 장착한 완성차들이 많아지고, 프리미엄 모델이 아닌 일반 차량 모델에도 커넥티드 인포테인먼트와 내비게이션을 기본 옵션으로 장착하고 있다. ABI Research에 따르면 멀티미디어 스트리밍, 와이파이 핫스팟, 음성비서 등을 내장한 차량은 2025년 약 5,700만대가 판매될 것으로 예상되고, 2023년까지 북미와 유럽은 거의 모든 차량이 음성제어 기능을 보유할 것으로 예상하고 있다[7].

이러한 커넥티드 차량의 확산은 차량 데이터와 운전자 데이터를 활용해 수익을 올릴 수 있는 차량 데이터 모네타이제이션(Data Monetization) 산업 성장을 촉진하고 있다. 현대차는 현대디벨로퍼스(Hyundai Developers), BMW는 카데이터(Car Data), GM은 디벨로퍼스, 폭스바겐은 커넥트 고(Connect Go) 등의 서비스를 통해 차량 데이터를 확보하여 다양한 부가 서비스를 제공하고 있다.

일반적으로 차량 데이터는 ECU(전자 제어 장치), CAN(컨트롤 액세스 네트워크) 및 인포테인먼트 시스템에서 생성되며, 위치, 엔진 상태, 도어 잠금 상태, 속도, 타이어 공기압, 안전벨트 착용 여부, 외부온도 등 차량과 관련된 거의 모든 데이터 수집이 가능하다. 완성차 업체들뿐만 아니라, 해당 데이터를 재가공해 보험사, 금융사, 운송회사, 응급 서비스 업체 등에 판매하는 오토노모(Autonomo), 위조(Wejo)와 같은 전문업체들도 등장하기 시작했다.

해당 업체들은 차량 데이터를 수집해 긴급 서비스, 전기차 관리, 구독 기반 서비스, 주차, 예측 유지보수, 보험, 미디어, 차량 내 서비스, 교통 관리 및 스마트 시티 솔루션 등을 완성차 업체나 관련 서비스 업체에 제공한다. 차량 빅데이터 분석 업체 오토노모는 벤츠의 차량 데이터를 바탕으로 GPS를 활용한 차량 위치 추적, 원격 진단, 운전자 안전 모니터링, 서비스 타임 최적화, 자동주차 페이먼트, 주유와 충전 관리, 전기차 경로 계획 수립, 차량 사용 모니터링, 효과적 차량관리 등의 서비스를 제공하고 있다. 그리고 긴급 구조 업체에는 실시간 도로 상황 정보를, 렌터카 기업들에는 원격으로 연료 상황이나 차량 잠금 실행 서비스를, 전기차에는 경로상에 효과적인 충전소 정보 등을 제공하기도 한다[8].

2) 차량 데이터 모네타이제이션: 보험 분야

컨설팅업체 캡제미니(Capgemini)에 따르면 차량 데이터를 활용한 잠재적 수익과 시장진입속도(Go-to-Market Speed)가 가장 높은 모네타이제이션 분야는 보험이며, 그 중에서도 최초 보상신고 개념인 FNOL(First Notice of Loss)이 잠재적 수익과 시장진입속도가 가장 빠른 분야다. FNOL은 보험에 가입한 운전자에게 사고, 도난, 부상 등이 발생했을 때 계약자가 보험사에 최초로 통지하는 절차를 의미한다. 일반적으로 계약자의 성명, 생년월일, 주소 등 개인정보, 사고 날짜와 시간, 위치, 사고 발생 경위, 차량 번호 및 관련 정보, 과실 당사자 보험정보와 사건 세부 정보 등을 전달한다.

최근 보험업계에서 많은 관심을 받는 PAUD(Pay-As-You-Drive) 및 PHUD(Pay-How-You-Drive)와 같은 UBI(Usage-Based Insurance) 보험에도 차량 데이터는 유용하게 활용되고 있다. 커넥티드카는 관련 프로세스를 인공지능과 인터넷을 통해 자동화된 프로세스를 제공하는 장점이 있다. 이러한 데이터를 활용해 테슬라는 2019년 9월부터 보험 사업도 시작했다. 테슬라가 출시한 차량에 탑재된 8대의 카메라, 레이더, 센서 등을 통해 다양한 데이터를 수집하고 있다는 것은 널리 알려진 사실이다. 테슬라가 수집하는 데이터는 매우 방대

4 Introduction of Giant Self-driving Companies Tesla and BASIC-How Data Labeling Services Empower Self-Driving Industry 2021?, ByteBridge, 2021.5.28.

5 The Connected Car Market Will Endure a 15% Shipment Decline, Flat Revenues in 2020; Sales Return on Trend Early 2022, ABI Research, 2020.7.28.

6 Connected Car Market by Service(OTA Update Navigation, Cybersecurity, Multimedia Streaming, Social Media, e-Call, Autopilot, Home Integration & Other), Form, End Market(OE, Aftermarket), Network, Transponder, Hardware and Region - Global Forecast to 2025, MarketsandMarkets, 2020.7.

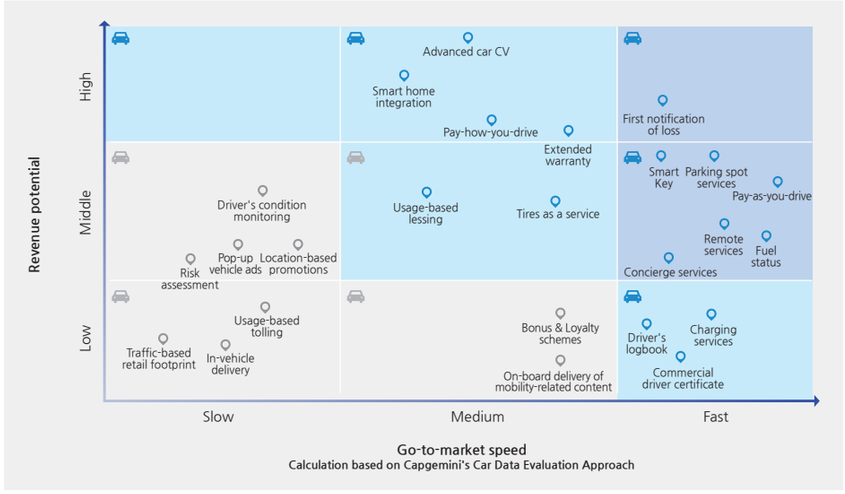
7 The Connected Car Market Will Endure a 15% Shipment Decline, Flat Revenues in 2020; Sales Return on Trend Early 2022, ABI Research, 2020.7.28.

8 autonomo website(https://autonomo.ai, 접속일: 2021.11.7.)

하다. 데이터 출처별로 구분하면 ①고객과 사용 디바이스(성명, 주소, 전화번호, 이메일, 지불 내역 정보), ②차량(차대번호, 속도, 주행거리계, 배터리 및 충전 이력, 각종 전자시스템 사용 이력, 소프트웨어 버전, 인포테인먼트 시스템 사용 데이터, 안전벨트 착용 여부 등 안전 데이터, 카메라 이미지, 에어백 정보, 가감속, 보안, 사고 발생 시 비디오 클립, 차량 호출기능(Summon), 차량 퍼포먼스 데이터, 리모트 서비스 이력, 차량 외부 비디오 클립), ③개인정보(연락처, 인터넷 검색 이력, 내비게이션 사용 이력, 라디오 채널 사용 이력, 현재 위치, 사용 충전소 위치), ④신용정보(은행 라우팅 정보, 사회보장번호, 신용카드 사용 이력 및 한도, 운전면허증, 생년월일, 기타 카드 관련 정보) 등으로 거의 모든 차량과 운전자 정보를 활용한다.

데이터 기반의 개인 맞춤형 설계로 경쟁사보다 20~30% 보험료가 저렴한 것이 특징으로 일론 머스크는 향후 보험 비즈니스가 테슬라 비즈니스의 30~40%의 가치를 차지할 것이라고 주장했다[9]. 이렇듯 차량 데이터 활용 범위와 가치, 새로운 서비스 제공 범위는 점차 확장되고 있는 유망한 분야로 성장이 기대된다.

[그림 4-6-3] 차량 데이터의 잠재적 수익과 시장진입 속도에 따른 서비스 유형 분류[10]



* 출처: 「Monetizing Vehicle Data- How to Fulfill the Promise」, Capgemini, 2020.9.

9 Johnna Crider, Tesla Insurance Could Be 30-40% Of The Value Of Tesla's Car Business, 2020.10.24.

10 X축 시장진입속도(Go-to-Market Speed)는 해당 서비스들이 실제 서비스화되어 제공되기까지 걸리는 시간을 의미하며, Y축 잠재적 수익(Revenue Potential)은 해당 서비스가 출시되었을 때의 수익 규모 예측 수준을 의미

다. 카커머스와 카페이먼트 시스템(In-Car Payment System)

최근 자동차와 페이먼트 업계는 과거 완성차 기업들이 강조했던 파워, 스타일링, 안전 중심의 시스템 개발 철학과 마케팅 키워드를 넘어 이커머스(e-Commerce)와 결합된 카커머스(Car-Commerce) 서비스 제공을 위한 노력을 강화하고 있다.

미국의 예를 들면 2018년 기준 일주일 중 평일 5일 동안 매일 평균 51분을 운전하는 미국의 1억 3,500만 운전자 가운데 73%는 스마트폰 혹은 차량 인포테인먼트 시스템을 이용해 인터넷과 연결하며 2,300억 달러를 지출했다[11].

주니퍼 리서치(Juniper Research)는 전 세계에서 2018년 3억 3,000만 대에서 2023년까지 7억 7,500만 대의 소비자용 차량이 텔레매틱스 또는 차량 내 앱을 통해 연결될 것으로 분석했다. 자율주행과 커넥티드 기능이 발전함에 따라 운전자가 이동 중 서비스와 상품을 검색하고 지불하는 빈도가 높아져 시장 규모도 2023년 2,650억 달러 규모로 성장하여 전체 이커머스의 1%를 차지할 것으로 예상되는 등 앞으로의 성장 가능성도 무한하다[12].

이미 피아트크라이슬러(FCA)는 커넥티드카를 위한 OTA(Over-the-Air)로 가능한 차량 내 상거래 플랫폼 Uconnect Market, BMW는 ConnectedDrive, 혼다는 DreamDrive, 벤츠는 주차결제 서비스인 on-street, 실시간 교통정보 구독 서비스인 in-car store, 재규어 랜드로버는 스마트 월렛(Smart Wallet) 서비스를 운영하고 있다. 폭스바겐은 2025년을 목표로 10억 유로 이상의 수익을 창출하기 위해, 자율주행 차량 승객이 이동 중 서비스를 사용할 수 있는 클라우드 기반 시장을 개발하고 있다[13].

현재 글로벌 시장에서 제공되는 주요 서비스는 주유소와 충전소, 음식점, 주차장 등의 검색, 예약, 주문, 결제, 커피와 편의점 등 드라이브 스루, 다양한 사전 물품 주문과 픽업, 수리와 같은 자동차 관련 서비스 예약, 호텔 예약, 콘서트 등 티켓 예약, 음악 및 영상 스트리밍 결제 등이 있다. 향후 관련 서비스 대상은 점차 확대될 것으로 예상되며, 자동차는 새로운 미래 금융서비스 성장의 원천으로 자리 잡고 있다.

특히 카페이먼트 생태계는 새로운 서비스보다는 음성인식 등 서비스를 활용하기 위한 사용자 경험이 핵심으로, 현재 제스처·홍체 인식 등 다양한 시도가 진행되고 있다.

결국 완성차 업체 등 카커머스를 추진하고 있는 기업들의 전략의 핵심에는 데이터가 있

11 「Digital Driver Report 2019」, P97, 2020.1.

12 Junper Research, 「In-Vehicle Commerce Opportunities Drive Total Connected Cars to Exceed 775 Million by 2023」, 2018.10.30.

13 John Maslen, Volkswagen targets €1bn revenue from connected car services, Asset Finance International, 2019.1.9.

다. 운전자 혹은 탑승자가 구매하는 서비스나 물품의 빅데이터를 기반으로 상황에 따른 구매를 분석하고 상황에 적합한 서비스나 물품을 사전에 예측해 적절한 시기에 선호도를 반영해 정보를 제공해 개인화(Personalization) 서비스를 확대하고 마케팅 자료로도 활용하겠다는 전략이다.

카커머스 핵심인 카페이먼트는 서비스 제공기업, 결제 솔루션 서드파티 기업, 완성차 기업, 결제 기업과 함께 가상비서, 구매 대상 물품과 서비스 제공자로 구성이 되어 있다. 성공적인 에코시스템 구축을 위해서는 해당 기업들 간의 긴밀한 협력과 함께 API 및 개발 플랫폼을 제공하여 타사 제공기업에서 결제 가능한 차량 내 앱을 제공할 수 있어야 한다.

[그림 4-6-4] 카페이먼트 에코시스템



* 출처: Krishna Jayaraman, 「Car-as-a-Marketplace: A Transaction Center on Wheels」, BrightTALK, Webinar 발표자료, 2020.1.22.

3. 향후 전망과 과제

2021년 6월 개최된 4차산업혁명위원회 해커톤에서는 '자율주행 자동차 영상정보 활용 방안'에 대한 의제가 논의되었다. 업계에서는 비식별 처리 데이터 사용 시 인공지능 학습 과정에서 오류 발생이 높아져 영상 데이터 원본이 필요하다는 입장이며, 시민단체 등에서는 개인정보 보호를 위해 영상 데이터 원본 사용은 곤란하다는 입장이다. 결론적으로 개인정보 보호를 위한 적절한 안전장치 확보 및 유출사고 발생 시 법적 책임·피해 배상을 전제로 영상 데이터 원본을 R&D에 활용할 수 있어야 한다는 점에 상호 동의했다. 이에 따라 관련 부처인 개인정보보호위원회, 국토교통부, 산업통상부는 관련 법령 개선을 위해 노력하고, 4차산업혁명위원회는 반영 사항을 모니터링한다는 결론을 내렸다. 자율주행기술이 발전하고, 국가·기업별 경쟁이 심화됨에 따라 관련 데이터에 대한 중요성 및 기업들의 요구 사항이 증가할 것으로 예상된다.

단, 민간기업 데이터 개방에 대한 이슈는 한동안 지속될 것으로 예상된다. 자율주행 자동차 인공지능 학습을 위한 기업들의 영상 데이터와 민간 모빌리티 기업들이 서비스 제공을 통해 획득한 데이터에 대한 공개 및 공유 이슈는 계속 제기되고 있다. 공공도로를 활용하는 모빌리티 서비스 데이터를 공공재로 보는 관점으로 기존에 개방된 공공데이터와 같이 새로운 비즈니스 창출의 한계를 극복하자는 관점이다.

[표 4-6-1] 국내 교통관련 공공데이터 플랫폼

구분	시스템명	관리기관	주요 정보
차량운행	디지털운행분석기록 시스템(eTAS)	한국교통안전공단	- 버스, 택시, 화물차의 실시간 운행행태 정보 - 운행기록, 위험운전행동 등 운행 전반 정보
	교통안전정보관리 시스템(TMACS)		- 전국에서 발생한 교통사고에 대한 통계적, 시각적 정보
	택시운행정보관리 시스템(TIMS)		- 시간대별 도로구간별 택시정보 시각화 정보 - 이용자별 맞춤형 승하차 지점 정보
대중교통	전국대중교통정보 (TAGO)		- 전국 대중교통 노선 및 운행 관련 정보
	대중교통DB 시스템(PTDB)		- 교통수단별 대중교통 이용 관련 통계자료
운수종사자	운수종사자종합관리 시스템		- 사업용자동차 운수종사자 사고이력자료 - 사고일시, 장소, 위치 등 사고 상세정보
	자동차검사통합 시스템(VIMS)		- 차량 운행기록 검사자료 - 차량용도, 검사일자, 주행거리 정보
기타	온실가스관리시스템 (KOTEMS)		- 교통부문 온실가스 배출량 - 지역별, 교통수단별 온실가스 배출현황 정보
	화물운송실적관리 시스템(FPIS)		- 화물 운송 전반에 관한 정보 - 화물 하차일, 운송료, 운송실적 등
공간정보	통계지리정보서비스 (SGIS)	통계청	- 인구, 교통, 복지, 경제와 융합된 공간정보 - 읍면동별 시계열 통계자료
	국가공간정보포털 (브이월드)	한국국토정보공사	- 토지에 대한 권역별 지리정보 - 부동산, 건물에 대한 이용정보

구분	시스템명	관리기관	주요 정보
교통사고	교통사고분석시스템 (TAAS)	도로 교통공단	- 교통사고 사례 유형, 차종, 기상상태, 기하구조 - 대중교통을 비롯한 전체 통행의 미시적 특성
통행특성	국가교통 DB(KTDB)	한국교통연구원	- 가구통행실태조사, 주말 통행 실태조사, 승차 인원 - 대중교통을 비롯한 정체통행의 미시적 특성
차량운행	버스운행관리시스템 (BMS)	각 지방자치단체	- 정류장별 · 노선별 버스 출 · 도착정보

* 출처: 이지민, 「모빌리티 활성화 및 지원에 관한 법률안 검토」 보고서, 2020.11.

모빌리티 분야 최상위 법으로 2020년 9월 발의되어 국회에 계류 중인 '모빌리티 활성화 및 지원에 관한 법률안'에는 민간기업이 보유한 데이터의 제출 혹은 교통빅데이터 플랫폼 연계를 중앙행정기관과 지방자치단체장이 요청할 수 있도록 하는 조항이 포함되어 있다.

하지만 국회교통위 검토 보고서는 '교통 관련 보유 데이터를 중심으로 새로운 비즈니스를 창출하려는 민간기업 등에 과도한 부담이 될 소지가 있다'는 결론을 제시했다. 실제로 중앙행정기관과 지방자치단체뿐만 아니라, 데이터 관련 스타트업들도 새로운 비즈니스 창출을 위해 해당 데이터들의 개방을 요구하고 있지만, 서비스 운영 업체들은 데이터 개방 요구를 거부하고 있는 상황이다.

향후 다양한 모빌리티 관련 데이터 확보를 통한 모빌리티 산업 활성화 기반 확대도 중요하지만 공공과 민간데이터의 공유, 개방, 협력을 유도할 수 있는 거버넌스 체제 구축도 중요한 이슈다.

제7장. 에듀테크분야 데이터 활용

유인식
(유비온 전무이사)

코로나 상황을 맞아 전 세계 교육시장의 큰 전환이 일어나고 있다. 하지만 이것은 4차산업혁명시대를 맞아 서서히 변화되고 있었던 교육시장의 변화를 촉진하는 것이었을 뿐 근본적인 원인은 아니라고 할 수 있다. 좀 더 근본적인 전환의 요인은 시대가 바라고 있는 인재상이 바뀌고 있고 여기에 더해 새로운 기술들이 그 교육방식의 변화를 촉진하고 있다는 사실이다. 이러한 변화의 흐름 속에 데이터 기술은 자동화, 개인 맞춤화와 같은 역할을 통해 교육 전환의 시대에서 핵심적인 기능을 맡고 있다.

1. 글로벌 에듀테크 산업 동향

가. COVID-19 상황으로 에듀테크에 대한 관심 급증

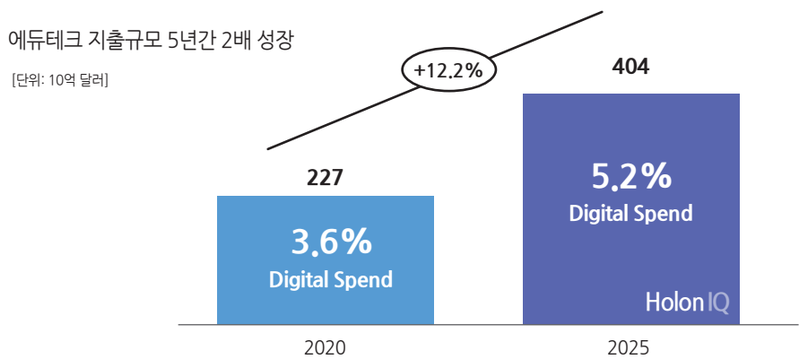
코로나 상황을 맞아 교육 전반에 많은 변화들이 생겨나고 있다. 코로나19 확산이 장기화되면서 많은 학교가 폐쇄되었고 노트북이나 태블릿, 화상채팅 등 IT를 활용한 원격교육이 활성화되었다. 이에 따라 인공지능(Artificial Intelligent, AI)을 비롯한 메타버스(Metaverse), 클라우드 컴퓨팅(Cloud Computing)과 같은 다양한 ICT 기술이 에듀테크분야에 등장하였고, 작년 한해 비대면 교육현장에 바로 적용되기 시작하며 본격적인 에듀테크 시대의 서막이 열리고 있다. 이들 신기술과 교육이 단순한 결합 이상의 시너지를 내고 있는 이유는 교육의 편리성과 효율성 강화를 추구하고 있는 교육시장의 변화와 관계가 있다.

2020년도는 하버드대학이 1930년 대공황이후 처음으로 적자를 기록한 해이다. 코로나로 인해 원격교육시스템 구축 등을 위한 투자가 늘어난 반면 다른 수익들은 감소한 탓이다. 반면 온라인 공개수업(Massive Open Online Course, MOOC)의 대표주자인 코세라(Coursera)는 최근 상장을 통해 6조 원 가치의 기업으로 상장되었다. 에듀테크산업이 전통적인 교육방식을 대체할 것이라는 기대가 시장에 모이고 있는 것이다.

나. 에듀테크 분야 전망

글로벌 교육시장 내 에듀테크 분야 지출규모 비중은 2020년 2,270억 달러(3.6%)에서 2025년 4,040억 달러(5.2%)로 12.2% 성장할 것으로 기대가 되고 있다.

[그림 4-7-1] 글로벌 교육 분야 에듀테크 지출 규모(2020~2025년)

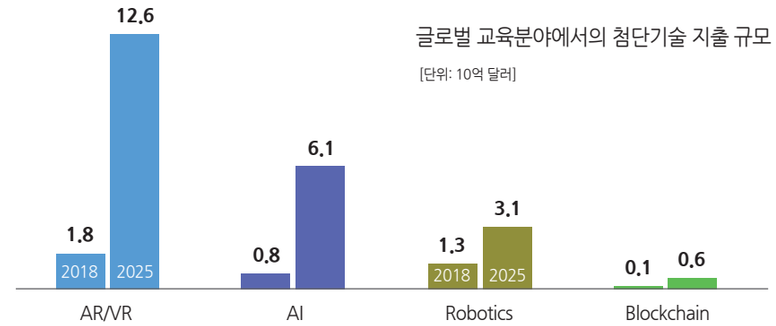


* 출처: HolonIQ, 「Global Education Technology in 10 Charts」, 2021.1.

이러한 성장이 예상되는 에듀테크 분야에서 글로벌 트렌드 중 하나로 꼽히는 '실감화'는 메타버스로 대변되는 AR/VR 기술을 통해 학습 콘텐츠의 물리적 제약의 해소뿐만 아니라, 교육의 공간을 교실에서 가상현실로 확장할 수 있을 것으로 기대되고 있다. 이러한 트렌드에 따라, 교육 분야에서의 메타버스 시장 규모는 2018년 18억 달러에서 2025년 126억 달러로 약 7배의 높은 성장을 할 것으로 전망된다.

에듀테크에서 인공지능 기반의 개인 맞춤형 교육에 집중할 뿐만 아니라, 미국, 일본, 영국, 프랑스 등 각국에서 인공지능 인재 양성을 위한 교육 혁신 계획을 추진하는 등 AI에 대한 관심이 높아지고 있다. 에듀테크 내에서의 AI 기술 활용의 규모 역시 2025년은 61억 달러로 2018년에 비해 7배 성장할 것으로 예상되고 있다.

[그림 4-7-2] 글로벌 교육분야에서의 첨단기술의 성장



* 출처: HolonIQ, 「Global Education Technology in 10 Charts」, 2021.1.

다. 에듀테크 생태계 트렌드

최근 들어 국가 단위에서 에듀테크에 대한 관심이 높아지고 있는 이유는 코로나19 상황과 함께 4차산업혁명 시대에서 요구되고 있는 인재상이 예전과 달라지고 있기 때문이다. 새로운 역량 개념이 등장하고 있고 그에 따른 교육 방식의 변화가 교육 분야에서 에듀테크 도입의 필요성을 높이고 있다.

앞서 설명했듯이 미국, 영국, 일본 등은 급변하는 시대에 대응하여 에듀테크를 활용한 교육 혁신 정책을 수립하고 추진하고 있다. 우리 정부도 2021년 7월 교육빅데이터위원회를 구성하고 교육데이터와 관련된 발전방향을 논의하기 시작했다. 이렇듯 각 정부의 지원 정책과 메타버스, AI, 로봇 등의 대표적인 기술을 기반으로 에듀테크 산업은 급속도록 발전하고 있다.

글로벌 에듀테크 산업에서는 벤처캐피털들이 에듀테크 산업의 성장 가능성에 주목하며 관련 스타트업에 투자 규모를 급속도로 확대하고 있는 추세이다. 지금까지는 미국, 중국, 인도 3개 국가가 글로벌 에듀테크 시장 및 투자 규모의 90%를 차지하고 있지만 최근 들어 인도네시아, 베트남 등 신흥국가에서 에듀테크 산업의 성장세도 두드러지고 있다. 조만간 교육 분야가 전 세계에서 가장 주목받은 신흥 산업으로 성장할 가능성도 배제할 수 없다.

2. 에듀테크에서 데이터의 의미

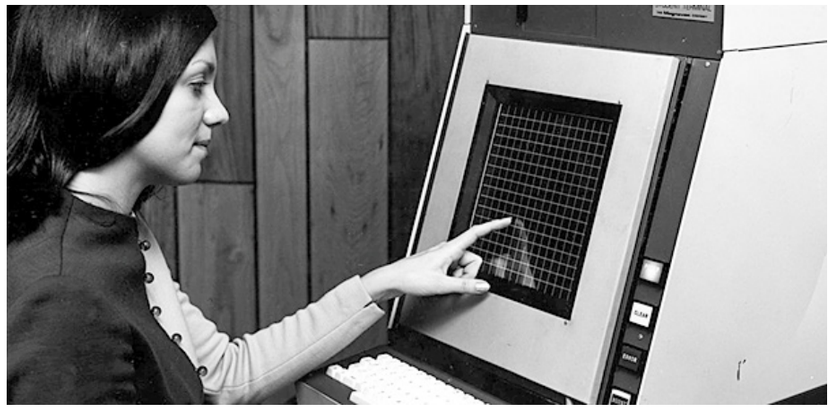
최근 우리 사회를 일컫는 대표적 수식 어구는 4차산업혁명과 지능정보화일 것이다. 4차산업혁명은 산업 전반이 클라우드, 인공지능, 메타버스 등의 정보통신기술과 융합하여 생산성이 고도로 향상되고, 제품과 서비스가 지능화되면서 산업 구조가 근본적으로 변화하는 것을 의미한다. 4차산업혁명의 특징으로 가장 많이 회자되는 두 가지는 '초연결성(Hyper connectivity)'과 '초지능성(Hyper intelligence)'이라고 할 수 있다. 이러한 초연결성과 초지능성이 실현된 '지능정보화'란 "방대한 정보를 수집, 분석, 처리, 저장할 수 있는 컴퓨터 및 지능형 시스템의 발달, 컴퓨터 및 시스템 간의 상호작용 기술에 기반하여 수요자에게 맞춤형 정보가 최적의 형태로 제공되는 것"을 의미한다.[1]

이러한 지능형 시스템이 '교육' 분야에 접목되는 것은 시간문제라고 할 수 있다. 학생에 관한 정보, 학생의 학습 활동정보, 학습과정 정보, 학업성취 및 학습 성과에 관한 정보들, 독서활동 정보, 교사와 학생·학교에 관한 각종 행정지원 정보 등 가르치고 배우는 교수 학습, 내·외적으로 수많은 교육 정보들이 통합 시스템을 통해서 취합, 저장되고 있다. 데이터를 기반으로 한 이러한 '지능형 학습 서비스'가 학생들의 학습활동 데이터를 수집하고 학생들의 학업성취도를 진단한 후 그에 맞는 맞춤화된 개인화 학습을 제공할 가능성이 높아지고 있기 때문이다.

지능형 학습 서비스의 개념은 최근의 것이 아니다. 이미 40년 전 컴퓨터가 교육에 활용되면서 컴퓨터 보조학습(Computer Aided Instruction, CAI) 및 컴퓨터 기반 수업(Computer-based Instruction, CBI)이 학습자의 '개별화 학습(Personalized learning)'을 지원하면서부터 일찌감치 시작되었다. 1965년에 플라톤의 이름을 따서 만든 PLATO (Programmed Logic for Automatic Teaching Operation)는 개별화(Personalized) 혹은 적응형(Adaptive) 교육을 지향하면서 탄생된 대표적인 교육용 애플리케이션이다.

1 홍선주 외, 「지능정보사회 대비 학교 교육의 방향 탐색」, 2016

[그림 4-7-3] 1965년에 출시된 교육용 애플리케이션 플라토(PLATO)



* 출처: https://grainger.illinois.edu/news/limitless-magazine/plato (접속일: 2021.11.2.)

현재 개별화 혹은 맞춤형 교육과 관련된 연구에서는 ITS(Intelligent Tutoring System)와 AIS (Adaptive Instuctional System)를 혼용해서 쓰고 있다. 최근 들어서는 ITS는 AIS를 구현하기 위한 하나의 모델로 정의하고 있고 AIS를 좀 더 포괄적인 의미로 활용하고 있다. AIS를 적용하는 데 가장 큰 난관은 AIS의 적용에 대한 효율성 부분이다. 지금까지 나온 대부분의 AIS는 특정 과목에만 적용될 수 있는 방식을 쓰고 있기 때문이다. 수학에서 사용되고 있는 AIS 엔진은 영어 분야에 적용하기 어렵기 때문이다. 이러한 문제들을 해결하기 위해 인공지능을 교육에 범용적인 방식으로 적용할 수 있는 접근 방법을 다음 세 가지로 요약하고 있다.

[표 4-7-1] AIEd(AI in Education)의 세 가지 모델

AIEd	의미	사례
Pedagogical model	교육방법에 대한 지식과 교육 전문성	생산적 실패(Productive failure): 학생들이 정답을 맞히기 전에 개념을 탐색하고 실수를 허용함
		학생의 행동에 따른 피드백(feedback): 질문, 힌트, 햅틱스 등을 제공하여 학습을 증진시킴
		학습의 진단과 측정
Domain model	특정 주제 관련 지식과 내용 전문성	분수의 덧셈, 뺄셈, 곱셈
		뉴튼의 물리법칙
		1차 세계대전의 원인
		텍스트 읽기의 서로 다른 접근들(예: 문장 읽기 등)

AIEd	의미	사례
Learner model	학습자에 관한 지식	학생의 선수 학습, 사전 지식, 경험
		학생의 감정 상태
		학생의 학습 몰입(예: 과제수행 시간: time-on-task)

* 출처: KERIS, 「지능형 학습 분석을 위한 데이터 수집·분석 API 고도화 연구」, 2018

우선 효과적 지식전달을 위한 ID(Instructional Design) 기반의 티칭 모델(Pedagogical Model), 학습 내용과 주제 중심의 주제 모델(Domain Model), 학습자의 태도·성향·성과 분석을 위한 학습자 모델(Learner Model)은 각각 구현되기도 하지만 대개의 경우 통합적인 방식으로 개발된다. 티칭 모델은 주로 LMS(Learning Management System)와 같은 학습 플랫폼에서 관심을 가지고 있는 반면 주제 모델은 수학, 영어 등 특정 과목을 서비스하는 학습 분야에서 주로 활용되었던 방식이다. 학습자 모델은 최근 들어 실시간 강의 혹은 시험감독 서비스에서 학습자의 학습상황을 분석하기 위해 주로 활용되고 있다.

이와 같은 AIS 모델은 각각 활용되기도 하고 두 가지 이상 병행해서 활용되기도 한다. 관건은 이들 모델이 특정 영역에서만 적용되는 방식이 아닌 범용적인 방식으로 여러 분야에서 적용될 수 있는 방법을 찾아야 한다는 것이다. 이를 위한 다양한 연구 개발이 진행되고 있다. 이러한 연구개발 노력들 중 하나가 IMS Global, ADL Initiative, IEEE와 같은 표준화 기관들이 공동으로 진행하고 있는 표준화 연구다. 이를 범용 도메인에서 적용하기 위해서 도메인 간의 호환성이 보장될 수 있는 데이터모델의 표준화가 우선적으로 필요하다.[2] 이 표준들을 각각 살펴보면 개별화 학습에 접근하고 있는 데이터표준 연구의 방향성을 짐작할 수 있다. 학습활동과 콘텐츠 단위에 대한 메타데이터 표준(LOM 2.0), 도메인별 호환 가능한 형식의 역량표준(RCD), 개인별 학습 활동 데이터 수집의 표준(xAPI 2.0)과 학습자 프로파일 표준(ILR, IGCL)은 개인화된 맞춤형 교육을 실현하기 위한 기본 데이터세트로 이를 조합해서 범용적인 방식으로 개별화된 학습분석을 할 수 있도록 시도하고 있는 것이다. 여기서 역량은 교육의 방향성을 제시하는 역할을, 학습자의 학습 상태 관리는 학습자 프로파일에, 학습에 활용되는 학습활동과 콘텐츠는 메타데이터 표준에, 그리고 마지막으로 학습자의 학습활동 로그는 LRS(Learning Record Store) 형식의 데이터 수집표준 스키마에 저장하게 된다.[3] 아래 그림은 이들 각각의 데이터모델이 서로 상호보완적으로 운영될 수 있다는 것을 보여주고 있다.

[그림 4-7-4] 상호보완적으로 운영 가능한 데이터 모델



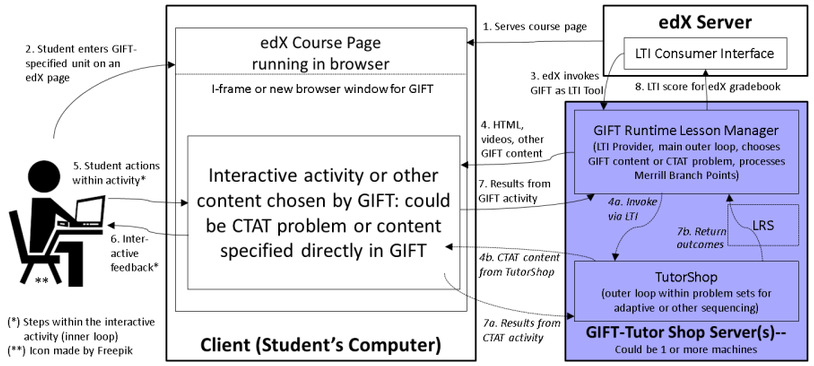
* 출처: ADL Initiative, 「Total Learning Architecture 2019 Report」, 2020.3.10.

범용 AIS 개발을 위한 또 하나의 과제는 다양한 환경에서 작동될 수 있는 프레임워크를 개발하는 것이다. 대표적인 사례가 미 육군 연구소(US Army Research Laboratory, ARL)의 LITE(Learning in Intelligent Tutoring Environments) Lab에서 개발한 GIFT(Generalized Intelligent Framework for Tutoring)이다. GIFT는 오픈소스 아키텍처를 채택하여 다양한 시스템과의 연계성을 목표로 개발되고 있다. 일례로 대표적인 오픈소스 MOOC 플랫폼인 Open edX와 고등교육 분야에서 가장 활발히 사용되고 있는 학습 플랫폼인 Moodle과의 연계를 시도하고 있다. [그림 4-7-5]는 Open Edx와의 연계를 위한 데이터 상호작용 관계를 보여주고 있다.

앞서 이야기한 AIS 표준화와 범용 프레임워크의 사례는 AI를 다양한 교육 분야에 적용하기 위한 노력의 일환으로 진행되고 있다. 하지만 여전히 인간화된 AI가 교육 분야에 적용되기까지는 많은 시간이 걸릴 것으로 예상된다. 그럼에도 불구하고 인공지능 기술을 교육 부문에 활용하고자 하는 노력은 지속될 것이고 그 근간은 데이터에 대한 관심으로 시작될 것이 틀림없다.

2 러닝스파크, 「교육분야에서 인공지능 활용의 가능성」, 2020.10.26.
3 러닝스파크, 「교육분야에서 인공지능 활용의 가능성」, 2020.10.26.

[그림 4-7-5] Open Edx와의 연계를 위한 데이터 상호작용 관계



* 출처: 「Integrating MOOCs and Intelligent Tutoring Systems: edX, GIFT, and CTAT」, 2017

3. 마무리

교육 관점에서 데이터를 바라보는 시각이 환경적 변화로 인해 최근 들어 많은 변화가 생기고 있다. 학습을 단순히 지식전달적인 방식으로 바라보던 시대에서 학습을 경험 중심적으로 역량과 스킬을 내재화하는 것으로 바라보는 시대로 전환되고 있기 때문이다. 지식전달적인 학습환경에서 데이터는 수월성을 판단하기 위한 몇 가지 정형화된 데이터를 수집하고 분석하는 것으로 충분했지만 경험 중심적인 학습환경에서는 기존 정형 데이터뿐만 아니라 다양한 비정형 데이터를 수집하는 것으로 수집의 대상과 방식이 변화되고 있으며 분석 방식 또한 이전과는 매우 다른 형식으로 전환되고 있다.

기술의 발전으로 인해 학습에서 데이터를 다루는 방식도 급격한 전환을 하고 있다. 특히 인공지능과 빅데이터 기술은 교육에서 데이터를 바라보는 시각을 전환하는 속도를 앞당기는 기폭제 역할을 하게 될 것이다. 하지만 정부 차원에서 산업을 육성하는 데 있어 각각의 기술을 중심으로 두고 개별 도메인 전략을 취하게 될 경우 근간이 되는 Data Science적인 접근의 중요성이 간과되거나 누락될 수 있다. 에듀테크 분야에서도 동일한 현상들이 벌어질 가능성이 있으므로 기존 기술 중심의 수직적(Vertical) 도메인 전략뿐만 아니라 Data Science를 중심으로 수평적(Horizontal) 에듀테크 산업 육성도 함께 고려해야 할 필요가 있다.

제8장. 농업분야 데이터 활용·비즈니스

최대근
(파미너스 대표)

현재 농업은 생산, 가공, 유통산업과 다양한 ICT 기술 등과 융합하며 계속해서 진화하고 있다. 특히 사물인터넷(Internet of Things, IoT)을 통해 수집된 농업데이터는 향후 인공지능 (Artificial Intelligence) 기술에 의해 세밀하게 분석 및 가공이 되어 새로운 농업 비즈니스 기회로도 발전할 수 있는 가능성이 높다. 과거에는 농업인의 환경 데이터가 공유되지 못하고 단기적인 개인 의사 결정 활용에만 그쳤다면, 앞으로는 농업에서 활용되는 다양한 센서와 클라우드 플랫폼 기술들이 서로 연계 및 호환될 수 있는 데이터 표준화가 진행될 것이고, 나아가 인공지능 시스템을 활용하는 것이 농업 경쟁력의 관건이 될 것이다.

1. 농업에서의 데이터 비즈니스 및 AI 활용 사례

가. 인디고 애그리컬처(미국 농업 유니콘 기업)

인디고 애그리컬처(Indigo Agriculture)는 농업 분야 최초로 유니콘 기업[1]으로 선정된 미국 스타트업이다. 주요 사업 모델은 '종자코팅(seed coating)'으로, 생산성 향상을 위해 종자에 미생물로 코팅을 입혀, 병충해나 열악한 자연환경에서의 생산성을 높이는 기술이다. 이 기술을 실현시키기 위해 인디고 애그리컬쳐는 미국 전역에서 오랜 기간 축적된 방대한 농업데이터를 수집하고 분석하였다. 현재도 빅데이터와 머신 러닝(Machine Learning)[2] 기술을 사용하여 종자 개발 연구에 1조원 이상의 가치를 가지는 데이터를 축적 및 관리하고 있으며 빅데이터가 활용될 수 있는 농산물 가격 책정과 생산 단계 이후의 사업 모델로도 확장시키고 있다.

1 기업가치가 10억 달러 이상인 스타트업 기업을 유니콘에 비유하여 지칭하는 말
2 컴퓨터가 스스로 방대한 데이터를 분석해서 미래를 예측하는 기술

[그림 4-8-1] 목화 종자에 인디고 애그리컬처의 미생물 코팅 기술 적용(왼쪽: 인디고종자)

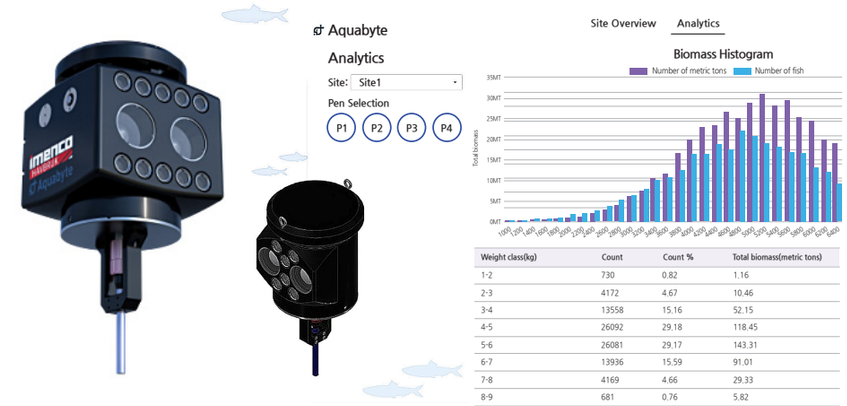


* 출처: Indigo Agriculture AG

나. 아쿠아 바이트(노르웨이 수산 AI 스타트업)

노르웨이의 수산 인공지능 스타트업인 아쿠아 바이트(AquaByte)는 빅데이터 및 인공지능 활용의 대표적인 기업으로 볼 수 있다. 빅데이터 수집을 위해 양식장 안에 설치된 수중카메라는 1,000 테라바이트가 넘는 250만 마리 이상의 생선 이미지를 학습하여 0.8% 오차 범위 내로 물고기의 무게를 예측한다. 연어의 안면과 길이로 개별 인식(Fish Facial Recognnition)을 하는데, 딥러닝(Deep Learning)[3] 기반의 데이터 분석을 통해 물고기의 건강, 크기, 최적 사료량을 결정하고, 양식장 운영 비용과 폐사율, 그리고 주변 생태계에 끼치는 영향을 최소화하는 데 성공하였다. 이 기업은 2019년 12월 아마존 AWS 혁신 우수 사례로 선정되었고, 향후 빅데이터와 AI의 학습을 통한 100% 완전 자동화된 연어 양식장을 구현하는 것을 목표로 하고 있다.

[그림 4-8-2] 아쿠아바이트 수중 카메라(데이터 인식) 및 빅데이터 활용



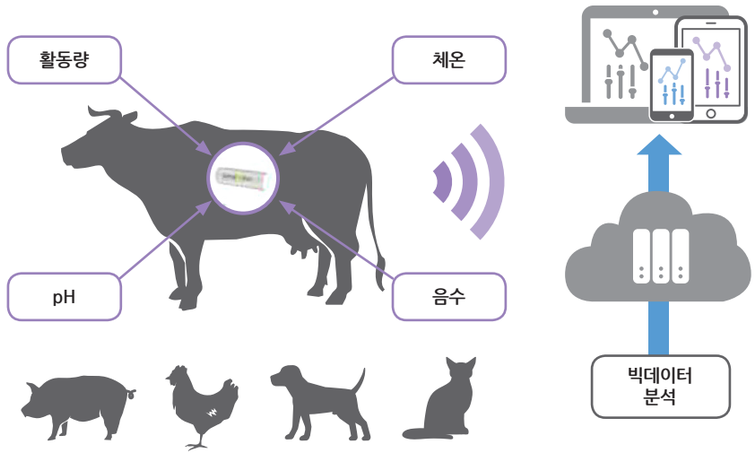
Weight class(kg)	Count	Count %	Total biomass(metric tons)
1-2	730	0.82	1.16
2-3	4172	4.67	10.46
3-4	13558	15.16	52.15
4-5	26092	29.18	118.45
5-6	26081	29.17	143.31
6-7	13936	15.59	91.01
7-8	4169	4.66	29.33
8-9	681	0.76	5.82

* 출처: Aquabyte(http://www.aquabyte.ai/index.html, 접속일: 2021.11.8.)

다. 스막텍(오스트리아 축산 AI 스타트업)

오스트리아의 축산 스타트업 스막텍(smaXtec)은 센서가 내장된 소형기기를 젖소의 체내에 삽입하여 젖소의 질병상태와 건강상태 등을 모니터링하면서 데이터를 축적 및 분석한다. 이 기기의 특징은 체외에 부착하는 목걸이 형태의 제품보다 훨씬 더 신뢰성 있는 데이터 확보가 가능하고, 특수 플라스틱을 이용한 센서이기 때문에 반영구적으로 사용이 가능하다는 점이다. 젖소가 하루에 몇 발자국을 움직였는지, 젖소의 가임 기간이 언제인지와 같은 다양한 가축 정보와 지역별 기후정보를 실시간으로 클라우드[4]에 전송하여 전 세계 목장의 빅데이터화를 이룸으로써, 데이터 축산 농업을 실현해 나가고 있다.

[그림 4-8-3] 스막텍의 최적 사육 관리 시스템



* 출처: RDA Interrobang(제199호), 「빅데이터가 바꾸는 농업의 미래」, 민승규, 서현권, 2017.7.14.

3 다층구조의 신경망을 기반으로 하는 머신 러닝의 한 분야

4 소프트웨어와 데이터를 인터넷과 연결된 중앙 데이터에 저장하여 인터넷에 접속하기만 하면 언제 어디서든 데이터를 이용하게 하는 것

라. 에이드믹스 바이오테크놀로지(대만 종돈 스타트업)

종돈[5] 스타트업인 에이드믹스 바이오테크놀로지(Aidmics Biotechnology)는 대만의 30대 연구자 2명이 창업한 농업 스타트업으로, 사물인터넷 기술과 빅데이터를 활용하여 종돈의 선발, 도태 검사[6] 등의 서비스를 제공하여 효율성을 극대화하고 있다. 특히, 기존 현미경 검사 대신 태블릿과 웹을 이용하여 돼지 정액성상을 분석함으로 검사 시간과 비용을 절약하고 개체 관리의 용이성을 증가시키는 등의 특징이 있다. 현재는 검사한 돼지 정액과 이를 이용해 생산된 돼지의 빅데이터를 활용하여 서비스를 제공하고 있으며, 다양하게 축적된 데이터를 이용하여 새로운 종돈 시장에 진출할 수 있는 가능성이 높다.

마. 한국 '디지로그팀'의 세계 농업 인공지능 대회 참여

빅데이터와 AI가 세계 농업의 트렌드가 되어 가고 있는 지금, 한국도 이러한 흐름에 도태되지 않기 위해 '디지로그(DigiLog)'라는 이름으로 한국 대표팀을 구성해 2019년 네덜란드 와게닝겐 대학(Wageningenn University and Research)과 중국 IT 기업 텐센트(Tencent)가 후원하는 제2회 세계농업 인공지능 대회(Autonomous Greenhouse International Challenge 2019)에 참여하였다. 1회 대회에서는 4개월간 인공지능 기술을 활용하여 원격으로 오이를 재배하였고, 그 결과 마이크로소프트 연구원과 와게닝겐 대학교 연구원이 소속된 소노마(sonoma) 팀이 우승을 거머쥐었다. 한국 대표로 디지로그팀이 참여한 2회 대회도 1회 대회와 마찬가지로 예선과 본선으로 나뉘어 진행되었는데, 예선은 온실 시뮬레이션 프로그램에서 가상으로 방울토마토를 재배하고, 가장 높은 이익(Net Profit)을 내는 상위 5개의 팀이 본선에 진출하는 방식이었다. 5개월의 시뮬레이션 재배 기간동안 온도와 습도, 환기, 관수, 보광, CO_2 시비 등 다양한 변수의 데이터를 제어해야 했다. 디지로그팀은 시뮬레이션에서 받은 데이터를 학습하고 원예 재배 전문가 팀원들이 제시한 변수의 범위를 조합하여 좋은 점수를 내는 방향으로 접근하였고, 24시간의 예선전에서 2위로 본선에 진출하였다. 본선에서는 네덜란드 실제 온실 공간에 디지로그팀의 IoT 센서 및 장비들을 설치한 후, 네덜란드가 아닌 한국에서 5개월 동안 원격으로 방울토마토 온실을 제어했다. 결과적으로 디지로그팀은 3등이라는 쾌거를 이루었다. 비록 짧은 기간 동안 방대한 데이터를 얻기는 어려웠지만, 다양한 분석 및 시도를 해볼 수 있었던 좋은 기회였으며, 이는 한국이 데이터 및 AI 농업의 테스트베드가 될 수 있다는 가능성을 보여준 지표였다.

[그림 4-8-4] 세계 농업 인공지능 대회 온실 공간과 IoT 센서 설치 중인 디지로그팀



* 출처: DigiLog

5 인위적 혹은 의학적 기술을 이용하여 돼지의 임신을 할 수 있게 수정되도록 하는 과정
6 동일한 종(種)에 속한 다수의 개체 중 특정 개체는 보존되고 다른 개체는 사라지는 현상을 검사하는 일

2. 국내 농업에서의 데이터 활용 기술 중요성 및 시사점

현재 농업 선진국은 계속해서 빅데이터를 통해 모든 농업데이터를 수집 및 가공하여, 새로운 농업 시장의 리더가 되고자 한다. 한국 역시 농업 생산, 유통 및 가공 부문에서 빅데이터 활용을 통한 시스템 정비와 사회적 합의가 더욱 필요할 것이다. 앞으로는 빅데이터가 산업 환경변화의 촉매제 역할을 담당하며 국가 간의 농업 경쟁력 부문으로 자리매김할 것이고, 이에 따라 수집된 데이터를 분석 및 가공하여 농업인에게 유용한 정보로 판매하는 서비스가 농업 비즈니스 모델이 될 수 있다. 온실 자재나 장비 등의 하드웨어는 제품 수명이 짧고 교체가 필요한 반면에, 정보는 축적될수록 가치가 높아질 뿐 아니라 다른 경쟁자들이 복제하기 어렵다는 장점이 있어서 앞으로 데이터 시장이 더욱 활발해질 것으로 예상된다. 만약 한국이 빅데이터 활용 측면에서 선진 농업 국가와 같은 경쟁력을 확보하지 못한다면, 다른 플랫폼 회사들이 가진 농업데이터에 의존할 수밖에 없는 상황을 초래할 수 있다. 가치 있는 농업데이터의 수집과 분석, 가공이 미래 농업 경쟁력의 원천이 될 것이다.

현재 외국 기업의 환경 제어 시스템을 설치한 국내 스마트팜 농가의 경우에는 계속해서 농가의 재배 및 환경 데이터의 일부가 외국으로 유출되고 있다. 한국에 자동화 스마트팜이 도입된 지 20년이 지났지만 그동안의 데이터는 개인 의사 결정 사용 등에만 국한되고, 다양

한 현장 데이터들을 활용하지 못했다. 이제는 데이터의 표준화를 통해 우리의 다양한 재배 및 환경 데이터를 활용하여 데이터 농업 발전으로 연결해야 한다.

앞으로 한국 농업의 성공적인 빅데이터 활용을 위해서는 크게 5가지 측면을 고려할 필요가 있다. 첫째, 센서와 장비, 플랫폼 기술들이 서로 연계되고 호환될 수 있도록 표준화를 진행하여 시스템 기반을 확보하는 것이 중요하다. 현재는 대부분의 장비와 시스템 간의 호환성이 부족하기 때문에, 표준화 작업을 통해 이런 문제를 해결해야 한다. 둘째, 국가 농업 기관이 보유하고 있는 방대한 데이터 진단과, 실제 농업에 적용할 수 있는 방법 모색이 필요하다. 셋째, 민간 분야의 다양한 비즈니스 주체와 개발 및 협업을 장려하여 그들이 농업에 참여할 수 있도록 참여를 독려하고, 개발 및 협업을 장려해야 한다. 넷째, 농업 빅데이터 축적 기술에 대한 재산권과 같이 향후 예상되는 법적 제도 정비가 필요하다. 농업인의 오랜 경험을 바탕으로 축적해온 노하우나 재배기술들에 대한 재산권 개념이 아직 불분명하기 때문에, 데이터를 제공하는 농업인과 이를 수집하는 기업 간의 소유권 문제가 발생할 수 있다. 마지막으로, 규모가 작은 소농업인의 경쟁력 강화를 위한 별도의 지원과 맞춤형 시스템 개발이 필요하다. 소농의 경우 ICT 기술을 활용하여 데이터의 수집 및 활용하는 데 적지 않은 비용이 발생하기 때문에 그들을 위한 행정적 지원 방법을 강구해야 한다.

3. 향후 전망: 농업, 이제 인공지능 기술과 데이터의 영역으로!

지금까지 농업데이터 비즈니스에 대한 기업 사례를 설명하고 시사점을 제시하였다. 그러나 농업 선진국의 대규모 농가를 대상으로 하는 '선택과 집중'형 모델은, 현재 소규모 농가가 많은 한국 농업에 적용하기에는 우려스러운 부분이 많다. 따라서 우리의 여건에 맞는 한국형 농업데이터 활용 모델을 개발하는 것이 우선이라고 생각한다.

현재 세계적으로 농업에 첨단 기술이 결합되어 많은 새로운 가치들이 창출되고 있으며, 한국 농업에도 다양한 센서와 소프트웨어 기술을 통해 빅데이터 수집과 같은 AI 기술이 활발하게 적용되고 있다. 특히 앞서 말한 제2회 농업 인공지능 대회를 통해 농가의 기존 재배 방식과 노하우가 작물 재배에 있어 최선의 선택이 아닐 수도 있다는 교훈을 얻었다. 당시 결승에 참여한 5개 팀은 모두 AI 전문가들로 이루어졌고, 그들은 모두 30년 경력의 사람 농부 팀보다 좋은 성적을 거두었기 때문이다. 대회 기간 동안 AI는 사람이 하지 않았을 선택을 하였고, 이는 단기적이지만 결과적으로 생산성을 높일 수 있는 판단이었다. 이를 통해 미래에는 빅데이터 및 AI 기술의 발전으로 보다 스마트한 과학 농사를 짓게 될 것으로 예상된다. 물론 식물이 살아있는 유기체이기 때문에 자동차의 자율 주행과는 달리 다양한 변수들이 있고 더 많은 양질의 데이터 학습이 되어야 한다는 선행조건이 따른다. 미래의 과학 농사는 스마트팜 내부에서 사람의 눈을 대신해 AI가 영상을 보고, 식물의 생육 및 병해충 상태를 판단하는 등의 분석이 가능할 것이다. 또한 스마트팜 온실 내에 부착된 센서와 데이터를 기반으로 과거 패턴을 분석해 질병에 걸릴 확률을 예측하고, 미리 선택적 방제를 할 수 있을 것이다. 이러한 농업의 미래를 위해 국내에서도 많은 노력을 기울이고 있다. 그 중 디지로그팀이 미래에 크게 활용될 수 있는 농업용 로봇을 개발 중이다. 현재 농가에서 가장 큰 문제 중 하나인 인건비 부담, 농약의 사용을 줄이고, 생산성을 늘리는 것이 이 로봇의 목표이다. 디지로그팀은 이러한 데이터 및 AI 기술 작물 재배 솔루션을 통하여 한국 농업데이터와 AI 농업 발전에 이바지하고자 한다. 마지막으로, 앞으로의 기후 변화와 인구 증가로 인한 식량문제를 해결하기 위해서는 농업데이터와 AI 기술의 개발이 필수적이다. 이러한 기술의 활용은 이제 국가의 농업 경쟁력이 될 것이다. 이러한 기술의 발전을 통해 한국 농업에 사람의 개입이 없는 "완전 자율형 온실" 구축과 같은 새로운 혁신의 바람이 불길 기대한다.

제9장. 데이터 동맹·제휴 현황

김인현
(투이컨설팅 대표)

데이터3법이 발효되면서 데이터 공유를 위한 법적 환경이 갖추어졌다. 데이터 결합기관과 데이터 전문기관이 지정됨으로써, 데이터 가명화를 통한 공유도 가능하게 되었다. 일회성이 아니라 지속적으로 데이터를 공유하기 위한 방법으로 참여 조직들이 협약을 맺는 데이터 동맹이 확산되고 있다. 데이터동맹은 카드사를 중심으로 금융회사들이 보다 활발하게 진행하고 있다. 통신사도 적극적이다. 데이터 동맹은 앞으로 더욱 확대될 것으로 전망된다. 데이터 동맹이 지속적으로 작동하기 위해서는 비즈니스 관점, 데이터 관점, 기술 관점의 과제를 해결해야 한다.

1. 데이터 동맹의 배경

가. 데이터 공유

개인데이터에 속하지 않는 데이터의 공유는 법적 제약이 없다. 예를 들어서 기상데이터, 주가데이터, 기업재무데이터 등은 소유자의 판단에 따라 판매하거나 무상으로 제공되기도 한다. 또한 이러한 데이터는 오픈데이터 또는 데이터상품 형태로 유통되고 있다.

반면 개인데이터의 공유는 법의 제약을 받는다. 이는 데이터 주체인 개인의 권리를 보장하고 예상되는 피해를 방지하기 위한 목적이다. 우리나라는 2020년 1월 '개인정보보호법' 등 데이터3법의 개정으로 개인데이터를 공유할 수 있는 길이 열렸다. 개인데이터를 가명화하고, 이를 결합전문기관을 통해 결합한 후 사용할 수 있게 되었다. 다만, 통계 작성, 과학적 연구, 공익적 기록 보존 등의 목적일 경우에만 허용된다.

나. 데이터 동맹의 등장

데이터 동맹은 둘 이상의 조직이 공동의 목적을 달성하기 위해 또는 각자의 이익을 위해 각자가 보유하고 있는 데이터를 공유하는 것을 뜻한다. 데이터 동맹이 필요한 경우는 대부분 개인데이터를 대상으로 한다. 한 조직이 확보할 수 있는 데이터는 한계가 있다. 그 이유로는 첫째, 충분한 규모를 확보하기 어려우며, 둘째, 다양한 관점을 포함하기 어렵다는 것이다.

첫 번째 한계점으로 인해 동일한 데이터를 보유하고 있는 조직끼리 동맹을 맺는 경우가 발생한다. 건수가 많을수록 데이터 분석 결과의 예측력이 높아지기 때문이다. 예를 들어 은행들이 개인의 여신데이터를 공유함으로써, 신용평가모델의 적중도를 높일 수 있다. 두 번째 한계점으로 인해서는, 서로 다른 데이터를 보유하고 있는 조직끼리 동맹을 맺게 된다. 폭이 커질수록 데이터 활용 분야가 넓어지기 때문이다. 예를 들어 카드회사의 매출데이터와 통신회사의 위치데이터를 결합하여 상권분석을 하는 경우가 있다.

디지털 경제가 본격화되고 있는 현재는 데이터 활용 능력이 조직의 성과를 좌우한다. 이는 공공기관과 민간기업 그리고 개인까지 모두 마찬가지이다. 데이터 분석을 통해서 빠르고 적확한 의사결정, 맞춤형 마케팅, 서비스 혁신 등이 가능하기 때문이다. 자신의 데이터만으로는 충분하지 않다. 외부 데이터를 활용할 수 있어야 한다.

이 때문에 외부 데이터 활용에 대한 수요가 급격히 증가하고 있다. 이러한 데이터 수요를 해결하기 위해서 데이터 거래 활성화를 위한 노력이 진행되어 왔지만 큰 성과를 내지는 못했다. 2020년 금융데이터거래소가 개설되었지만, 2021년 5월 현재 데이터 거래에 참여하고 있는 기업은 104개이고 이 중 53.8%인 56개 기업은 데이터 등록 건수가 한 건도 없다. 출범 후 1년간 유료 거래 누적 금액도 약 11억 원에 불과하다[1].

데이터거래소가 성립하기 위해서는 두 가지 조건이 필요하다. 첫째, 데이터가 상품화되어야 한다. 둘째, 데이터 생산자와 데이터 판매자가 다수 존재해야 한다. 데이터가 상품화된다는 뜻은 다수의 사용자가 같은 용도로 사용할 수 있는 상태로 제공되어야 한다. 현실에서는 데이터 수요자들이 필요로 하는 데이터 요건이 각각 다르다. 데이터거래소를 통해 데이터 거래 니즈가 충족되기까지는 조금 더 시간이 필요할 전망이다.

데이터를 필요로 하는 조직은 데이터를 보유하고 있는 조직과 직접 관계를 설정하는 것이 효과적이다. 데이터 동맹은 이러한 관점에서 주목을 받고 있다. 데이터 동맹을 맺는다고 하더라도 실제로 데이터를 주고받는 것은 데이터 보유조직이 임의로 할 수 있는 일은 아니다.

데이터 동맹이 목적을 달성하기 위해서는 데이터라는 자산을 공유하는 방식에 대한 법적 제도적 규칙이 마련되어야 하고, 실제로 데이터를 공유할 수 있는 기술 환경이 지원되어야 한다. 우리나라의 데이터 동맹은 2020년 이후 활발하게 탄생하고 있다. 데이터3법에 의해서 데이터 공유의 기준이 마련되었고, 결합전문기관이 지정됨에 따라서 데이터 공유가 실질

1 변준석, 박민정, 천정희, 「데이터거래 활성화를 위한 데이터구독플랫폼 구축 사전 연구」, 데이터리서치 브리프 2021년 9월

적으로 가능하게 되었기 때문이다.

2. 결합전문기관과 데이터전문기관 현황

가. 결합전문기관의 정의

결합전문기관은 서로 다른 조직이 보유하고 있는 개인의 정보를 결합하는 역할을 수행한다. 개인정보보호법 제28조의 3에서는 결합전문기관의 역할을 '서로 다른 개인정보처리자 간의 가명정보의 결합을 수행'하고, '결합된 정보의 외부 반출을 승인하는 것' 으로 정의하고 있다[2]. '신용정보의 이용 및 보호에 관한 법률'(이하 신용정보법) 26조의 4에서는 결합전문기관을 '정보의 결합과 익명처리의 적정성 평가를 전문적으로 수행하는 법인 또는 기관'으로 정의한다[3].

결합전문기관의 명칭은 각 법령에 따라 다르다. 개인정보보호법에서는 '전문기관'으로 표현하고 있고, 관련 고시에서 '결합전문기관'으로 표현하고 있다[4]. 신용정보법에서는 데이터전문기관으로 표현하고 있다. 개인정보보호법은 전체 산업을 아우르는 법률인 반면에 신용정보법은 신용정보와 관련된 조직에 관한 법률이다. 결합전문기관과 데이터전문기관의 역할은 사실상 동일하지만, 대상 데이터 영역이 신용정보인 경우는 데이터전문기관으로 호칭하고 있다. 언론에서는 '가명정보 결합전문기관', '데이터결합전문기관', '가명정보결합기관' 등으로 표현하는 경우도 있다.

다양한 호칭이 사용되고 있지만, 결합전문기관은 서로 다른 조직들이 보유하고 있는 가명정보를 각 조직이 가명화하여 제출하면 이를 결합해서 전달함으로써, 서로 다른 조직의 데이터 공유를 지원하고 법적 제약 요인을 충족시키는 것을 확인하는 역할을 수행한다.

나. 결합전문기관과 데이터전문기관 지정 현황

결합전문기관은 개인정보보호위원회 또는 관계 중앙 행정기관의 장이 정한다. 데이터전문기관은 금융위원회가 지정한다, 결합전문기관과 데이터전문기관으로 지정받으려면 [표 4-9-1]의 조건을 충족해야 한다.[5] [6] 결합대상 데이터에 금융데이터가 포함된 경우에는 데이터전문기관만이 데이터 결합을 수행할 수 있다.

[표 4-9-1] 결합전문기관과 데이터전문기관 지정 요건

요건	결합전문기관	데이터전문기관
기본요건	–	- 비영리법인 또는 공공기관 (이후 자본금, 매출액 등 요건을 갖춘 법인으로 확대)
인력·조직	- 8인 이상 담당 조직 - 전문가 3명 이상 상시 고용	- 8인 이상 전문인력 보유 - 임원의 개인정보보호법, 신용정보법 위반 사실 없을 것
시설·설비	- 결합, 가명처리, 반출 심사 등을 위한 업무 공간 및 시설 - 물리적 안전 조치	- 데이터 결합 등을 위한 정보처리, 정보통신 설비, 보안체계 등 구비
시스템 구축	- 결합, 가명처리, 반출 심사 - 데이터 송수신 - 결합전문기관 업무 기능	–
보안 조치	- 데이터 유출 방지 등 보안 조치	–
정책 절차 관리 체계	- 결합, 가명처리, 반출 심사 등을 위한 정책 및 절차 등	- 위험관리체계, 내부통제장치 - 설립 취지, 사업영역 적합
재정 능력	- 손해배상책임 보험 또는 공제 가입 최소 10억원 이상 - 자본금 50억원 이상	- 순자산 대비 부채 총액 비율 200% 이내 등

개인정보보호위원회는 2020년 11월 통계청과 삼성SDS를 1호 결합전문기관으로 지정했다. 금융위원회는 2020년 8월에 신용정보원과 금융보안원을 1호 데이터 전문기관으로 지정했으며, 보건복지부, 국토교통부, 과기정통부, 교육부, 산업통상자원부도 각각 결합전문기관을 지정했다.

금융위원회는 2021년 5월에 데이터전문기관을 통해 총 111개의 데이터가 결합되었다

2 개인정보보호법(2020.8.5. 시행), 제28조의 3(가명정보의 결합 제한)
3 신용정보법(2021.8.4. 시행), 제26조의 4(데이터전문기관)
4 가명정보의 결합 및 반출 등에 대한 고시(2020.9.1. 시행), 제2조(정의)
5 결합전문기관 지정 기준, 가명정보의 결합 및 반출 등에 관한 고시[별표 1]
6 데이터전문기관 지정 사전신청 공고, 금융위원회 사이트 (https://www.fsc.go.kr/po010101/74369?srchCtgry=1&curPage=22&srchKey=&srchText=&srchBeginDt=&srchEndDt=, 접속일: 2021.11.2.)

고 밝혔다[7]. 온라인 소상공인 대출심사 모형 상품 개발, 청년 등 금융이력 부족자를 위한 신용평가모형 개발, 취약계층에 대한 금융지원정책 효과 분석 등을 위한 데이터 결합이 진행되었다. 금융위원회는 다양한 기관이 보유한 데이터의 결합, 개방 등을 통해 특화된 결합서비스를 제공하기 위해 전문기관 추가 지정 방안을 마련할 것이라고 발표했다. 결합 과정에서, 데이터 성격에 따라 해당 부처가 발표한 가이드라인을 준수해야 한다. 예를 들어서 헬스 데이터를 결합하는 경우 보건복지부의 '보건의료데이터 활용 가이드라인'을 준수해야 한다.

3. 데이터 동맹 현황

가. 데이터 동맹 특징

데이터 동맹은 2020년 이후 급속하게 확산되고 있다. 이는 데이터3법 제정으로 개인정보의 합법적 거래가 가능하게 되었고, 마이데이터 사업의 도입으로 새로운 비즈니스를 시도할 수 있게 되었기 때문이다. 데이터 동맹의 특징은 다음과 같다.

첫째, 금융회사와 통신사가 주도하고 있다. 개인데이터를 대규모로 보유하고 있는 조직이 데이터 활용을 위해서 제휴를 추진하는 경우이다. 또한 금융은 산업계 최초로 마이데이터 사업 허가가 진행되어, 2022년부터 금융 마이데이터 서비스를 시작할 수 있게 된 점도 작용하고 있다. 금융 마이데이터 서비스 차별화를 위해 금융사들은 개인의 일상 및 상거래 데이터 확보를 위해 노력하고 있다. 통신사들은 새로운 비즈니스 창출과 함께 공익 목적의 데이터 활용에도 적극적이다.

둘째, 이종업종 간 동맹이 주류를 이루고 있다. 서로 다른 업종의 동맹이 보다 활발하다. 그중에서도 금융과 통신, 금융과 커머스의 제휴가 활발하다. 고객에 대해서 서로 다른 서비스를 하고 있는 조직들이 데이터를 공유함으로써 데이터 시너지를 높이는 전략이 선호되고 있다. 고객을 보다 잘 이해하기 위해서는 다른 관점의 데이터를 활용하는 것이 보다 효과적이다.

셋째, 비즈니스 및 서비스 협력으로 이어지고 있다. 데이터 공유는 비즈니스 협업 및 서비스 제휴로 확산되는 경향이 있다. 데이터를 서로 주고받는 것을 뛰어넘어서 타깃 마케팅 기획 지원, 신용 평가 모형 개발 및 공유, 연구 과제 공동 추진 등의 활동도 추진될 계획이 있다.

7 「빅데이터 인프라 확충과 데이터산업 선도를 위하여 금융 분야 가명정보 결합·활용을 가속화하겠습니다」, 금융위원회 보도자료, 2021.5.27.

나아가서, 데이터 결합을 통해 새롭게 만들어지는 데이터를 거래소를 통해 공동으로 판매하는 서비스 개발을 추진하는 사례도 있다.

나. 공공 중심 데이터 동맹

공공 부문이 보유하고 있는 데이터는 민간기업의 비즈니스에 큰 도움을 줄 수 있다. 또한 민간기업의 데이터를 활용한 공공서비스 혁신도 성과를 내고 있다. 공공과 민간의 데이터 동맹은 공공 부문이 선도하고 있다.

과기정통부는 데이터얼라이언스전략위원회를 2019년에 발족했다. 빅데이터플랫폼 사업을 통해 형성되는 빅데이터플랫폼과 센터를 통한 민관데이터 공유 활성화가 목표이다. 행정안전부는 SK텔레콤과 사회적 가치 실현을 위한 데이터 공유 및 분석 공동 수행 협약을 체결하였다. 국토교통부는 국토부, 인천국제공항 등 산하기관과 연구기관, 지자체 등이 참여하는 협약 체결을 주도했다. 서울신용보증재단은 SK텔레콤, KB국민카드와 소상공인 지원을 위한 데이터 동맹을 추진하고 있다.

[표 4-9-2] 2021년 공공과 민간의 데이터동맹 주요 사례

동맹	시기	참여 조직	내용
소상공인 지원을 위한 민관 빅데이터 동맹	2021년 7월	SK텔레콤, 서울신용보증재단, KB국민카드	SKT의 이동통신 정보 기반 유동인구데이터, KB국민카드의 소상공인 가맹점 데이터, 서울신용보증재단의 상권활성화지수 등 상권분석에 핵심적인 데이터 융합
국토교통데이터 산업 생태계	2021년 7월	국토교통부, 국토교통공공기관 협의체의 25개 참여기관	참여기관 간 협력을 강화하기 위한 업무협약 체결. 데이터 공유 및 공동 연구과제 추진 등 국토교통 데이터산업 생태계 조성을 위한 환경 구축

다. 금융 중심 데이터 동맹

카드회사들이 데이터 동맹에 가장 적극적이다. 카드회사는 일회성 협약보다는 데이터 동맹의 생태계를 만들고 이를 확장하는 관점으로 추진한다. KB국민카드는 이업종 데이터 융합 플랫폼을, 신한카드는 오픈 얼라이언스를, 삼성카드는 LINK파트너 등을 운영하고 있으며, 참여사를 지속적으로 확대해가고 있다. KB국민은행, 신한은행, 우리은행, 하나은행, 카카오뱅크 등은 상거래 기업과 데이터 제휴를 통해 금융데이터의 한계를 확장하는 방향으로 노력

하고 있다.

민간 부문의 데이터 댐 구축을 목표로 하는 동맹도 있다. 우리은행과 교보생명 등은 초대형 민간 금융데이터 댐을 구축하기 위해 협력한다. 신한카드는 SK텔레콤 등과 제휴하여 민간 영역의 모든 데이터를 수집하여 활용할 수 있게 하는 데이터 댐을 구축하고자 한다.

[표 4-9-3] 2021년 금융회사 중심 데이터동맹 주요 사례

동맹	시기	참여 조직	내용
민간데이터 댐 구축	2021년 2월	신한카드, SK텔레콤, 코리아크레딧뷰로, GS리테일홈쇼핑, 다날, 이니시스, 부동산114	- 민간 영역의 모든 데이터를 수집, 결합, 분석, 유통하는 민간 데이터 댐 구축 - 오픈형 얼라이언스를 지속 확대해 자동차, 제조, 패션, 의료, 교통, 숙박 등의 데이터 기업과 데이터분석 전문기업, 컨설팅업체, 시스템·솔루션 기업 등 추가 합류 예상(신한카드)
도메인 갤럭시	2021년 2월	현대카드, 이마트, 코스트코, 스타벅스, 배달의민족, 대한항공 등	- 상업자 표시 신용카드(PLCC)를 매개체로 하는 협력 관계 - 데이터는 공유하지 않지만 데이터를 기반으로 현대카드가 각 파트너사에 마케팅 지원
신용평가모형 개발과 고도화를 위한 데이터 협력	2021년 3월	카카오뱅크, 카카오페이	- 카카오페이의 결제 및 고객 행동데이터, 카카오뱅크의 대출 수신 데이터 협력으로 신용평가모형 개발에 활용. - 카카오 주요 계열사 데이터 협력 추진 계획
빅데이터 기반의 보험상품 제휴 마케팅	2021년 3월	오렌지라이프, SK플래닛	- 각 사가 보유한 고객데이터를 기반으로 고객 맞춤형 서비스 개발 및 종합 금융솔루션 서비스 제공 추진
데이터사업 제휴 협약	2021년 4월	삼성카드, 이마트24	- 양사 데이터 교류 및 분석, 빅데이터 기반 공동 리서치 및 데이터 판매, 삼성카드 LINK파트너를 통한 빅데이터 기반 마케팅 등 - LINK 파트너 참여사 확장 추진(삼성카드)
데이터 유통 활용 촉진을 위한 업무협약	2021년 5월	비씨카드, 한국데이터산업진흥원, 금융보안원	- 각 기관은 표준계약서 등 금융데이터 거래 활성화를 위해 협력
금융데이터 융합 기반 금융 트렌드 공동연구를 위한 업무협약	2021년 5월	교보생명, 우리은행, 우리카드, 미래에셋증권, 한화손해보험, NICE평가정보	- 초대형 민간 금융데이터 댐 구축을 위한 협력 - 데이터 수집부터 결합·분석·유통 등의 프로세스 간소화, 금융·소비·건강 등에 이르는 통합 고객 상황판, 가명처리 정보 수집·결합·분석 등
디지털 트랜스포메이션 업무협약	2021년 7월	GS리테일, KB국민은행	- GS페이 개발을 통한 핀테크 비즈니스 확대, 공동 프로모션 및 홍보 마케팅 협력, 양사의 데이터 협업을 통한 빅데이터 시너지 창출
데이터 상호협력을 위한 업무협약	2021년 8월	BC카드, 이마트24, 닐슨컴퍼니 코리아	- 데이터 교류 및 분석, 빅데이터 기반 공동 리서치 및 데이터 비즈니스 모델 개발, 빅데이터 기반의 마케팅 등에 협력

라. 비금융 데이터 동맹

통신사 중심의 데이터 동맹, 모빌리티 서비스 개발을 위한 데이터 동맹, 엔터테인먼트 서비스 혁신을 위한 데이터 동맹 등이 가동되고 있다. SK텔레콤, KT, LG U플러스 등은 공공, 금융, 미디어 등과 제휴를 확대하고 있다. 현대자동차는 모빌리티 서비스의 선제적 개발을 위해 GS칼텍스, 렌터카회사 등과 협약을 체결했다. 엔터테인먼트 분야에서 미디어데이터얼라이언스는 콘텐츠 소비행태 분석을 위해, 한국음악콘텐츠협회는 K-Pop 아티스트 인기 분석을 위해 데이터 동맹을 맺고 있다.

[표 4-9-4] 2021년 기타 데이터동맹

동맹	시기	참여 조직	내용
미디어 데이터 얼라이언스	2021년 3월	LG유플러스, LG헬로비전, CJ CGV, KDX	- 미디어/콘텐츠 데이터 레이크 구축 계획 - 극장 상영 영화 실시간 방영, VOD 시청 이력 데이터 공유, 전체 시장의 콘텐츠 소비 행태 분석
소셜차트 2.0	2021년 6월	가온차트, 유튜브, Mubeat, V LIVE, 마이셀럽스, 틱톡	- 한국음악콘텐츠협회는 K-Pop 아티스트의 글로벌 인기를 확인할 수 있는 소셜차트 2.0을 운영하기 위해 데이터 제휴 - 2021년 6월에 틱톡 데이터 추가하기로 함
데이터 협력 및 사업 제휴를 위한 업무협약	2021년 7월	KT 아이지에이웍스	- IPTV와 모바일 이종 영역 데이터를 결합하여 TV 광고 시청과 타깃 고객의 행동 변화 간 연관성 도출 등

4. 데이터 동맹 향후 과제

데이터 공유의 필요성은 점점 더 커질 것이다. 데이터 공유의 방법으로는 데이터 거래와 데이터 제휴, 데이터 플랫폼이 있다. 데이터 동맹은 참여자의 필요에 따라서 다양한 방식이 적용될 것으로 보인다. 데이터3법이 제정되고 데이터결합기관과 데이터거래소가 가동되고 있지만, 데이터 동맹이 활발하게 작동하기 위해서는 넘어야 할 과제가 있다.

가. 비즈니스 관점

동맹을 맺은 조직들 사이에 보유하고 있는, 그리고 제공할 수 있는 데이터의 내용과 품질이 동일하지 않다. 동맹으로 인해 발생한 가치에 각 참여자가 기여한 정도에 따라 수익과 비용의 배분이 이루어져야 할 것이다. 데이터 제공에 따라 기회 손실이 발생한다고 판단하는

참여자는 동맹에서 이탈할 가능성이 크다. 하지만 데이터로 인한 기여를 명확하게 산출하기는 쉽지 않다.

나. 데이터 관점

데이터는 보유 조직이 소유권을 완전하게 갖고 있다고 볼 수 없다. 동의를 받았다고 하더라도, 제한된 범위에서 사용 또는 제공 권한을 갖고 있을 뿐이다. 따라서 동맹에 의하여 임의로 데이터를 주고받는 것은 관련 법규 및 동의 관리에 의해서 통제되어야 하며, 동맹에 참여한 조직들 사이에는 데이터 표준이 공유되어야 한다. 또한 데이터 동맹의 데이터 공유가 지속되기 위해서는 데이터 거버넌스가 정립되어야 한다.

다. 기술 관점

데이터 동맹에 참여한 조직들 사이에 데이터 교류 방법으로는 몇 가지가 있다. 바람직한 것은 참여 조직들이 공통으로 사용하는 플랫폼을 운영하는 것이다. 플랫폼을 전담하는 기능을 동맹에 참여하는 어느 조직에서 전담하는 방안도 고려할 수 있다. 이러한 데이터 교류 기술 인프라를 잘 갖추는 것이 데이터 동맹이 실질적으로 작동하게 하는 전제 조건이다.

PART



데이터산업 기술 동향

제1장.

AutoML 동향 그리고 대중화

제2장.

Cloud Computing

제3장.

개인정보 비식별 처리 수준 및 프레임워크

제4장.

안정적 AI/ML 서비스를 위한 MLOps

제1장. AutoML 동향 그리고 대중화

조치선
(바이브컴퍼니 그룹장)

AI 생산자 플랫폼은 2025년까지 130억 달러 규모로 빠르게 성장할 것으로 전망되며, AutoML(Automated Machine Learning) 관련 솔루션 또한 2030년까지 145억 달러 규모의 성장세가 예상된다. 이러한 성장 배경에는 데이터 사이언티스트가 해결해야 하는 다양한 비즈니스 이슈들을 풀어나갈 인력이 한정적이라는 현실이 있다. 때문에 피처 엔지니어링, 튜닝, 분석모델 구현, 검증, 배포, 운영을 자동으로 수행하는 AutoML 솔루션의 필요성이 대두된다. 이는 데이터 분석에 발생되는 여러 이슈들을 해결하고, 데이터 사이언티스트의 부족을 보완해 준다. 향후에 는 기초적인 분석 역량만을 보유한 시티즌 데이터 사이언티스트 만으로도 이슈 해결과 기업의 생산성 증대와 투자 수익률(ROI)을 개선하는 데 기여할 수 있게 될 것이다.

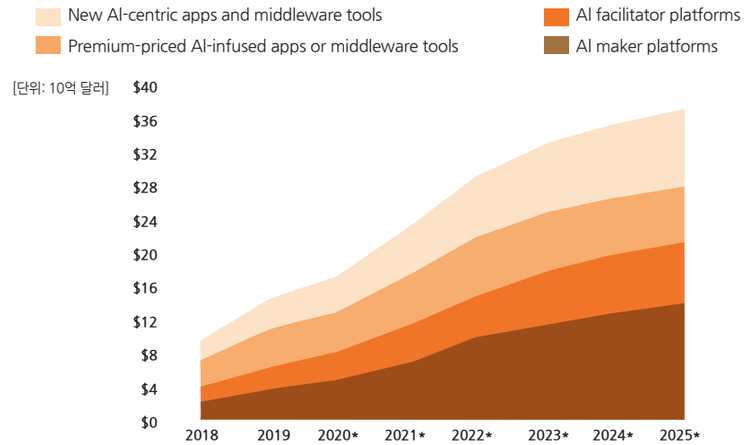
1. AI/ML과 AutoML 시장의 전망

가. AI/ML 시장의 전망 및 AutoML 정의

전 세계적으로 인공지능(Artificial intelligence, 이하 AI) 및 기계학습(Machine Learning, 이하 ML) 시장에 대한 관심과 수요가 증가하는 가운데 Forrester의 보고서는 2025년까지 AI 소프트웨어 시장이 370억 달러로 성장할 것으로 전망했다. AI 시장 중 AI 생산자 플랫폼이 가장 빠르게 성장하여 2025년까지 130억 달러의 규모에 도달할 것으로 예상된다.[1]

[그림 5-1-1]는 4가지의 AI 소프트웨어 부문을 ① 범용 AI 생산자(maker) 플랫폼, ② AI 촉진자(facilitator) 플랫폼,③ AI 중심(centric) 애플리케이션 및 미들웨어, ④ AI 인퓨즈드 애플리케이션(AI infused application)으로 구분하였다. 이러한 4가지 중 AutoML 솔루션이 범용 AI 생산자 플랫폼에 해당되는데, AutoML[2]란 Automated Machine Learning의 약칭으로 머신러닝 비전문가에게 머신러닝을 제공하고, 머신러닝의 효율성을 향상시키며 머신러닝에 대한 연구를 가속화시키는 방법과 프로세스를 의미한다.

[그림 5-1-1] 전 세계 2025년까지의 AI 소프트웨어 시장 성장 전망



*출처: Forrester, 「The AI Software Market Will Grow To $37 Billion Globally By 2025」, 2020.12.10.

나. AutoML 시장의 전망

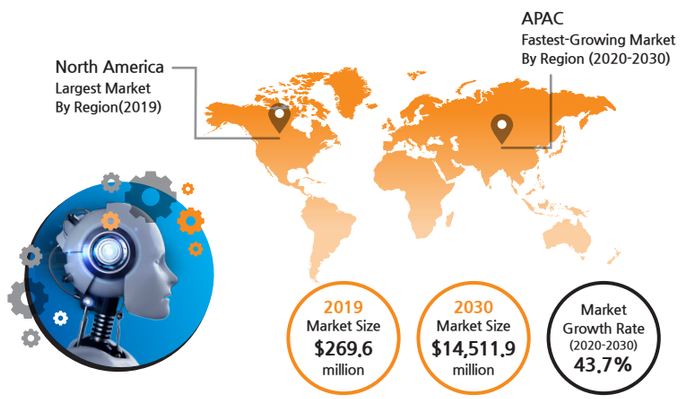
P&S Intelligence의 보고서 [그림5-1-2]는 전 세계 AutoML 시장이 2019년 2억 6,960만 달러, 2030년까지 145억 1,190만 달러의 수익을 창출할 것으로 추정했으며, 2020~2030년까지 CAGR 기준으로 43.7%의 높은 성장세를 예상했다[3].

1 Forrester, 「Sizing The AI Software Market: Not As Big As Investors Expect But Still $37 Billion By 2025」, 2020.12.10.

2 https://www.automl.org (접속일: 2021.11.7.)

3 https://www.psmarketresearch.com/market-analysis/automated-machine-learning-market/ (접속일: 2021.10.30.)

[그림 5-1-2] 전 세계 자동 기계학습(Automated Machine Learning) 시장



*출처: https://www.psmarketresearch.com/market-analysis/automated-machine-learning-market/ (접속일: 2021.10.30.)

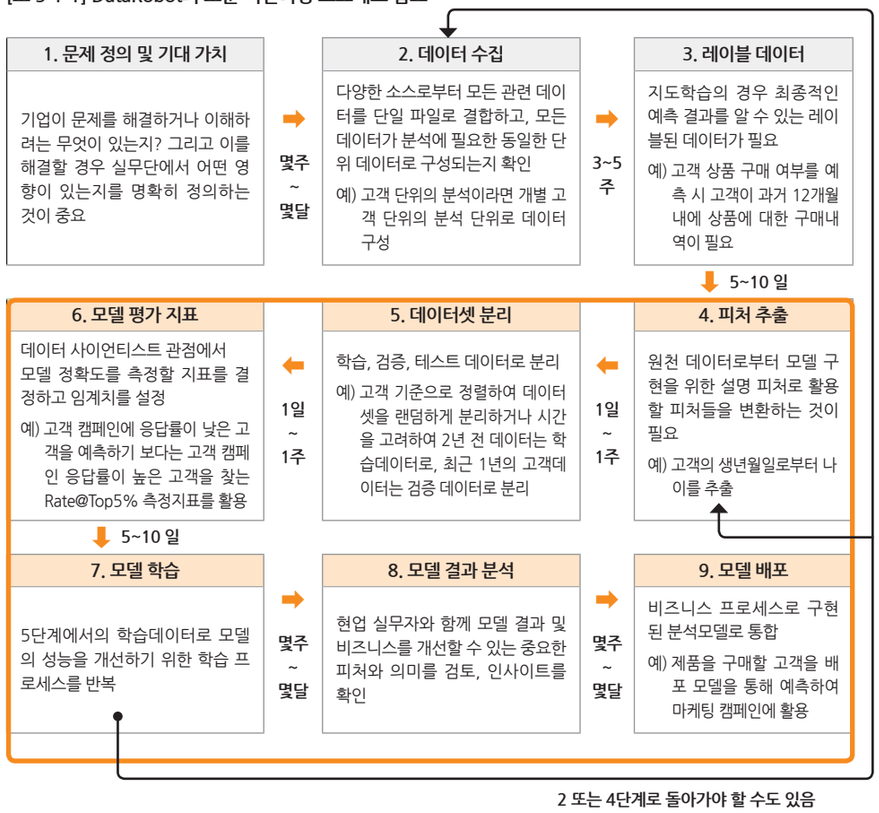
2. AutoML의 절차 및 제공 유형

가. AutoML의 절차

AutoML은 머신러닝 전문가에 의존한 적절한 피처 선택, 워크플로우, 전형적인 머신러닝, 알고리즘, 하이퍼 파라미터 튜닝을 벗어나 전문 지식 없이도 쉽게 사용할 수 있는 머신러닝 방법을 점진적 자동화하는 것을 목표로 두고 있다.[4] DataRobot에서는 [표5-1-1]과 같이 표준 머신러닝 프로세스를 9단계로 제시하였다. 각 단계는 ① 비즈니스 문제 정의 및 기대 가치 ② 데이터 수집 ③ 레이블 데이터 ④ 피처 추출 ⑤ 데이터셋 분리 ⑥ 모델 평가 지표 ⑦ 모델 학습 ⑧ 모델 결과 분석 ⑨ 모델 배포 순으로 나누어진다. 그중 AutoML이 가능한 영역을 ④ ~ ⑨ 단계로 보고 있으며, 이러한 단계에서 데이터셋이 업로드된 이후 분석이 자동으로 이루어져서, 프로그래밍 지식이 없는 사람도 비즈니스 문제를 해결할 수 있다.

4 Hutter F, Caruana R, Bardenet R, Bilenko M, Guyon I, Kegl B, and Larochelle H. "AutoML 2014@ICML". AutoML 2014 Workshop@ICML. Retrieved 2018.3.28.

[표 5-1-1] DataRobot의 표준 머신러닝 프로세스 참고



* 출처: https://www.datarobot.com/wiki/automated-machine-learning/ (접속일: 2021.10.31)

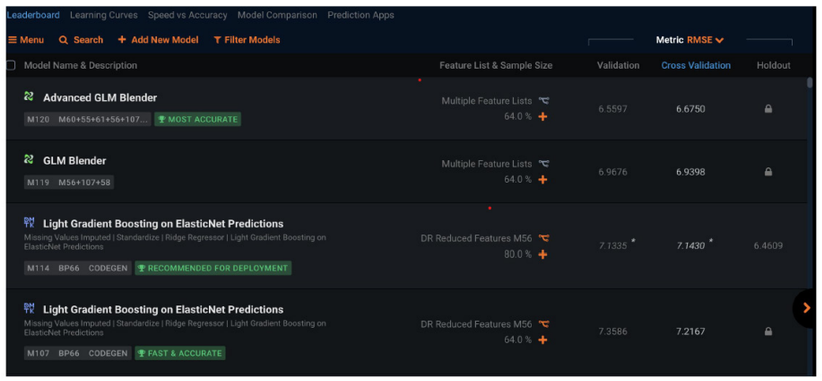
나. AutoML의 등장 배경 및 필요성

과거부터 지금까지 많은 기업들이 통계적, 수학적 알고리즘을 기반으로 데이터마이닝, 빅데이터 분석, 머신러닝, 인공지능을 활용한 분석을 수행해왔으나, 각 기업들의 데이터 사이언스 성숙도는 천차만별이다. 기업들의 데이터 사이언스 성숙 과정을 살펴보면, 1단계에서는 데이터 사이언스를 통한 비즈니스를 전환하는데, 이에 필요한 분석과 데이터에 대한 지식 그리고 관련 인력이 부족하다. 가장 큰 문제는 분석에 필요한 데이터가 구비되어 있지 않고, 있더라도 분석가에게 데이터를 설명하고 이해시키는 데 오랜 시간이 걸린다. 그리고 현업 실무자 입장에서는 분석 이해도가 떨어져서 비즈니스에 접목시키는 것이 무척 어렵다. 2단계에서는 데이터 처리에 대한 어려움과 인사이트에 대한 이해를 돕기 위한 솔루션을 고려해서

빅데이터 플랫폼, 데이터 전처리 솔루션, 시각화 솔루션을 활용하여 가시적인 성과를 내는 듯 하지만, 이 또한 현실의 비즈니스 프로세스에 접목시키는 경우는 드물다. 3단계에서는 전문 데이터 사이언티스트와 데이터엔지니어, 현업 실무자, IT실무자 등이 조직을 구성하여 예측이 필요한 문제에서 실현 가능한 비즈니스 문제(실효성 평가)를 중심으로 과업을 정의한다. 실무 관점에서 정의된 문제는 전문 데이터 사이언티스트가 직접 R/Python 코딩해서 머신러닝과 딥러닝 알고리즘을 활용하여 예측 모델을 만들고 모델을 배포하여 비즈니스 프로세스에 접목시켜본다.

앞선 1~3단계는 기업들의 데이터 사이언스 성숙 과정에서 일반적으로 겪게 되는 시나리오이다. 그러나 기업들이 오랜 시간을 투자하여 이러한 문제를 해결하기에는 전문 데이터 사이언티스트와 데이터엔지니어 인력이 턱없이 부족하고 지속적으로 문제를 해결하는 데도 한계에 부딪힌다. 데이터 분석에 있어 겪게 되는 문제로는 ① 피처 엔지니어링, 튜닝, 분석 모델 구현, 검증에 많은 시간이 소요됨 ② 다양한 알고리즘을 검증하기 위한 코딩 능력과 지식이 사람의 경험과 숙련도에 따라 유동적임 ③ 피처 중요도와 해석 설명이 표준화되지 않아 이해도가 떨어짐 ④ 사람에 의존하여 머신러닝 코딩을 수행하기 때문에 병렬로 머신러닝 문제해결을 수행할 수가 없음 ⑤ 데이터 변화에 따른 모델 재학습 및 변경 대응이 느림 ⑥ 모델 배포와 운영 서비스 관리가 어려움 ⑥ 현업 실무자가 데이터 사이언스를 수행하기 위한 진입 장벽이 높음 ⑦ 도메인 지식을 가진 데이터 사이언티스트의 확보 및 육성의 어려움 등이 있다. 이러한 이유로 인하여 AutoML의 필요성이 부각된다고 할 수 있다. AutoML 솔루션을 활용한다면 [그림 5-1-3]과 같이 자동으로 수많은 머신러닝, 딥러닝 알고리즘들을 병렬로

[그림 5-1-3] DataRobot 모델 결과를 제공하는 리더보드 화면



* 출처: DataRobot

피처 엔지니어링, 튜닝, 분석 모델 구현, 검증을 수행하여 빠른 시간에 최적의 모델을 추천받아 비즈니스에 접목하는 것이 가능하다.

다. AutoML의 제공 유형

AutoML의 제공 유형은 오픈소스(Open Source), 스타트업(Startups), 거대 테크(Tech giants)로 나누어 볼 수 있다. 오픈 소스 AutoML 기반 도구로는 Python 또는 R의 개발환경이 필요하며, 사용자가 자동화된 머신러닝 프로세스를 수행하기 위해서는 코드작성 및 실행 능력이 필요하다. 스타트업이 제공하는 AutoML은 데이터 사이언티스트가 부족한 현실에서 분석 역량이 떨어지는 기업이 보다 효과적으로 투자수익률(ROI)을 극대화할 수 있다는 점에서 최근 많은 관심을 받고 있다. 분석 역량이 부족한 사용자를 위해 깔끔한 인터페이스를 제공하고 있고, 분석 결과와 모델 설명에 있어 필요한 시각화뿐만 아니라 모델 배포과 운영 관리 기능도 제공하고 있다. 대표적인 거대 테크 기업으로는 구글(Google), 아마존(Amazon), 마이크로소프트(Microsoft) 등이 있으며, 대표적인 AutoML로는 Google Cloud AutoML, AWS Sagemaker, Microsoft Azure 등이 있다.[5]

[표 5-1-2] AutoML 제공 유형

소프트웨어	제공유형	출시연도	고용자 수
Xpanse Analytics	Startup	2014	1-10
TPOT	Open source(Python)	–	–
Tazi.ai	Startup	2015	1-10
PurePredictive	Startup	2011	11-50
MLJAR	Startup	2016	1-10
Kortical	Startup	2016	11-25
H2O.ai	Startup	2012	51-100
Google Cloud AutoML	Tech giant	2018	–
Enhencer	Startup	2017	1-10
DMWay Analytics	Startup	2014	1-10
DataRobot	Startup	2012	251-500

5 「AutoML Software/Tools in 2021: In-depth Guide」, 2018.6.3. (https://research.aimultiple.com/auto-ml-software/, 접속일: 2021.11.2.)

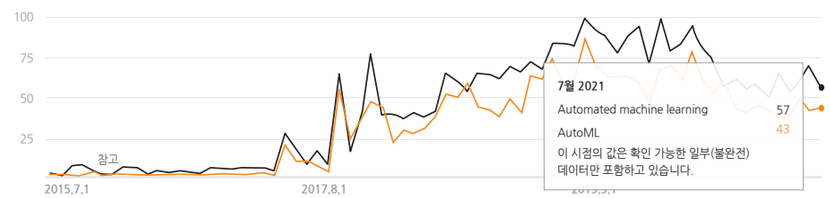
소프트웨어	제공유형	출시연도	고용자 수
Xpanse Analytics	Startup	2014	1-10
Caret	Open source(R)	–	–
Business Insight's TIMi Suite	Startup	–	11-25
BigML OptiML	Open source	–	–
Azure Machine Learning	Tech giant	–	–
Auto-WEKA	Open source	2013	–
Auto-Sklearn	Open source(Python)	–	–
Auger.ai	Startup	2017	–

* 출처: AI Multiple, 「AutoML Software/Tools in 2021: In-depth Guide」, 2021.1.18.

3. AutoML 관심도 추이 및 주요 제공 유별별 검색 추이

구글 트렌드를 통해 AutoML에 대한 관심도 추이를 살펴보면, 2016년 하반기부터 본격적으로 관련 키워드의 검색량이 증가했고, 2018년 ~2019년 꾸준히 증가했다가 2020년부터 감소세를 보였다.

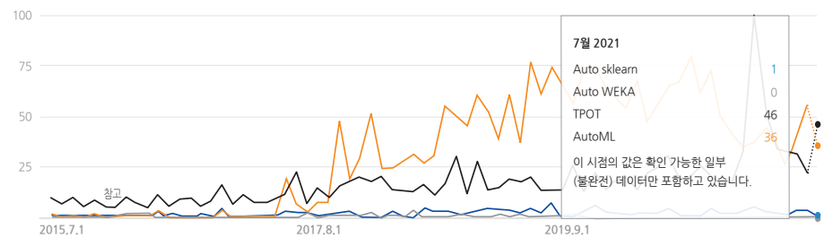
[그림 5-1-4] Google Trends를 통한 AutoML 관련 검색어 추이 분석(2015.7.1. ~ 2021.7.16.)



* 출처: Google Trends

비록 최근에 감소세를 보이고 있으나, AutoML에 대한 관심은 지속적으로 계속되고 있는 것으로 판단된다. [그림 5-1-5]와 같이 주요 AutoML 오픈소스 중 Auto-Sklearn, Auto WEKA, TPOT 등의 경우 전반적인 AutoML 키워드 검색량에 비해 낮은 추이를 보이나. 최근에는 TPOT에 대한 관심도가 증가하는 경향을 보인다.

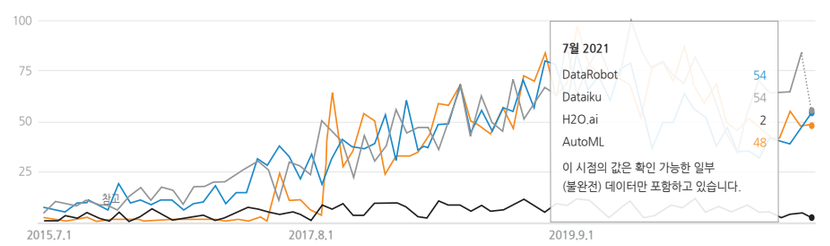
[그림 5-1-5] 오픈소스 - Auto sklearn, Auto WEKA, TPOT 등 관련 검색어 추이 분석(2015.7.1. ~ 2021.7.16.)



* 출처: Google Trends

스타트업에서 제공하는 AutoML 제품군이 일반적인 기업 비즈니스에 활용하는 상용 솔루션이며, [그림 5-1-6]에서의 분석 결과와 같이, AutoML과 관련된 상업용 솔루션에서는 DataRobot과 Dataiku의 검색량이 비슷하며, 함께 높은 관심도를 나타내고 있다.

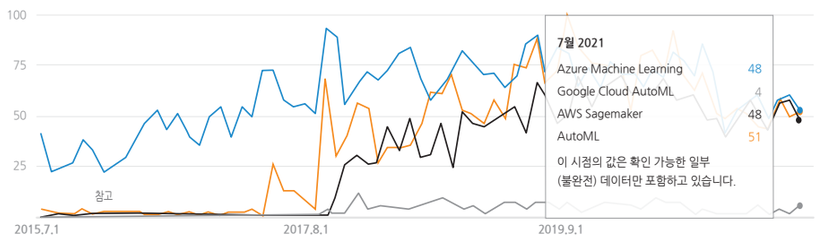
[그림 5-1-6] 스타트업 - DataRobot, Dataiku, H2O.ai 등 관련 검색어 추이 분석(2015.7.1. ~ 2021.7.16.)



* 출처: Google Trends

대표적인 거대 테크 기업인 구글, 아마존, 마이크로소프트에서도 AutoML 제품군을 출시했고, 자체 클라우드 환경에서 AutoML 제품을 활용할 수 있도록 지원하고 있다. [그림 5-1-7]에서 2017년 8월 이후 아마존의 AWS Sagemaker 검색량이 증가했는데, 이는 기업들이 클라우드 환경으로의 전환에 따른 것으로 보이며, 반면에 Microsoft Azure의 경우 AutoML과 상관없이 검색 추이가 꾸준히 증가하는 경향을 보인다.

[그림 5-1-7] 거대테크 - MS Azure, Google, AWS 등 관련 검색어 추이 분석(2015.7.1. ~ 2021.7.16.)



* 출처: Google Trends

4. AutoML 역할 및 이점

가. AutoML 역할

AutoML에 대한 관심 증가의 원인과 배경을 살펴보면, 첫째, 가장 큰 원인으로 데이터 분석의 수요는 계속 늘어나고 있으나 데이터 사이언티스트가 턱없이 부족하며, 둘째, 기업의 비즈니스 투자수익률(ROI) 측면에서 신속하고 빠르게 분석 결과를 내기 위한 수요가 증가하고 있다는 것이다. 최근에는 MLOps를 통한 모델 관리에 대한 중요성도 커지고 있는 상황이기도 하다. Bip Consulting(2020)[6]은 전문 데이터 사이언티스트와 시티즌 데이터 사이언티스트로서의 역할에 대한 AutoML의 영향력에 차이가 있음을 제시하였다. 비즈니스 문제를 해결하기 위해 잘 훈련된 데이터 사이언티스트를 발굴하고 육성하는 것은 어렵다. 그래서 앞으로 기업들은 어렵고 특화된 문제는 고비용의 전문 데이터 사이언티스트가 해결하게 하고, 일반적인 문제는 데이터 사이언스, 통계학 또는 수학이 부족하더라도 비즈니스 전문 지식과 데이터를 이해하고 있는 시티즌 데이터 사이언티스트가 AutoML 솔루션을 활용하여 비즈니스 문제를 해결하게 할 것이다.

6 Bip Consulting, 「Overview on AutoML Technologies v02」, 2021.10.
https://www.bipconsulting.com/wp-content/uploads/2020/10/a.storyblok.com-f-68014-x-8f2b26c4f6-overview-on-automl-technologies_v02.pdf (접속일: 2021.11.2.)

[그림 5-1-8] AutoML의 영향력



* 출처: Bip Consulting, 「Overview on AutoML Technologies v02」, 2020.10.

나. AutoML 이점

① 비용 절감 및 생산성 증대: 데이터 사이언티스트가 코딩에 사용하던 또는 소모하던 시간을 절약할수 있고, 다양한 모델의 성능을 비교하여 빠른 시간 내에 모델 성능을 높일 수 있기 때문에 생산성도 높일 수 있다. 또한 데이터 사이언티스트 교육 및 성장에 필요한 투자비용도 줄일 수 있다.

② 성과 개선: 분석의 숙련도가 낮더라도 신속하게 다양한 모델을 학습, 검증하고 최적 모델을 통해 비즈니스 관련 중요한 요인들과 예측 결과를 검토할 수 있어 새로운 인사이트를 찾고 확인하는 데 도움이 된다. 신속하게 현실 비즈니스에 접목하는 의사결정을 할 수 있고, 기업의 투자수익률(ROI) 개선 정도도 빠르게 확인해 볼 수 있다.

③ 모델 운영 관리: 서비스 중단 없이 모델이 운영될 수 있도록 서비스 상태를 모니터링하고 모델 정확도, 데이터 변화 관리를 모니터링함으로써, 안정적이고 지속적인 모델 운영 관리가 가능하다. 모델의 정확도가 떨어지는 시점에서 모델 재학습 및 교체가 편리하다.

5. AutoML 주요 활용 사례

AutoML은 금융, 보험, 핀테크, 건설, 부동산, 의료, 제조, 유통, 공공 등 다양한 산업에서 활용될 수 있다. 예를 들면 거래이상 탐지, 보험율 산정, 매출 예측, 영업관리, 마케팅, 수요예측, 고객 관리, 생산 수율 등 더 나은 의사결정과 예측을 위해 활용되고 있다. 미국 Avant사의 경우 DataRobot를 통해 대출 의사결정의 시간을 절약하고 보다 정확하게 위험을 식별한 성과를 냈고, 미국 Trupanion사는 DataRobot를 통해 판매 프로세스 속도를 높여 생산성을 10배 향상시켰으며, 고객의 2/3가 이탈하기 전에 이탈 가능성을 확인하고 대응하였다. 브라질 레노버사는 DataRobot를 통해 판매 및 제조 작업 모델의 정확도를 80%에서 87.5%로 증가시켰으며, 모델 생성 시간을 4주에서 3일로 단축시킨 성과를 내기도 하였다. 또한 미국 부동산업체 G5사는 H2O.ai를 통해 마케팅과 콜센터 관리에 있어 모델을 5배 빠르게 구현했고, 정확도는 95%까지 개선하는 효과를 보았다.[7] 국내에서는 SK하이닉스, 포스코, 미래에셋, KT, 쏘카, 카카오뱅크, 라이나생명, 신한은행, 현대캐피탈, 현대제철, SK가스 등 대기업에서 주로 AutoML를 활용하고 있고, 이러한 주요 대기업의 AutoML에 대한 활용 사례가 아직 대외적으로 공개되지는 않았다. 앞으로는 민간 대기업뿐만 아니라 자체 데이터를 확보하고 있는 공공기관, 공기업에서도 AutoML 솔루션의 수요(Needs)가 있을 것이라고 전망한다.

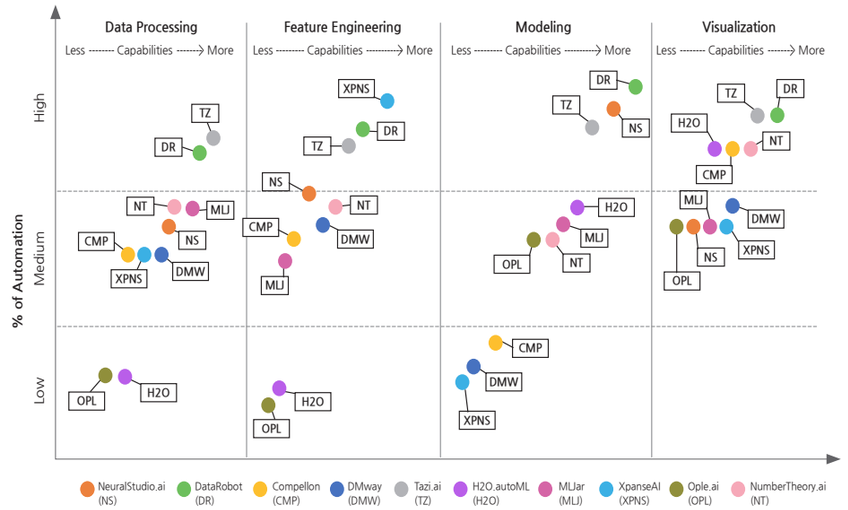
6. AutoML 공급업체에 대한 평가 및 요구사항

가. 상업용 AutoML에 대한 평가

Targetbase(2018)의 “Automated Machine Learning: An Agile Approach to AI Deployment"에 따르면 DataRobot과 H2O.ai가 다른 AutoML 공급업체보다 더 나은 성능을 보인다고 평가를 하였으며, [그림 5-1-9]와 같이 DataRobot과 Tazi.ai가 평가된 데이터 프로세싱, 피처 엔지니어링, 모델링, 시각화에서 다른 공급업체에 비해 더 높은 자동화 비율을 나타내고 있다.

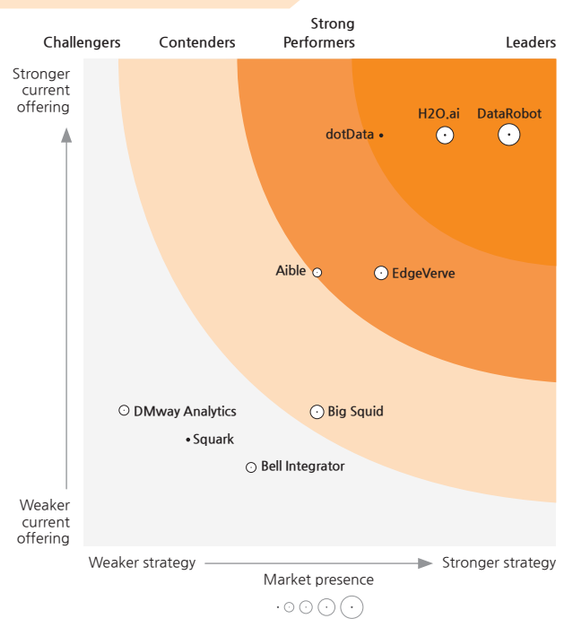
7 「22 AutoML Case Studies/Examples: In-Depth Guide [2021]」, 2019.8.9. https://research.aimultiple.com/automl-case-studies/ (접속일: 2021.11.2.)

[그림 5-1-9] Targetbase(2018)의 Automation 평가



* 출처: Targetbase, 「Automated Machine Learning: An Agile Approach to AI Deployment」, 2018.7.24.

[그림 5-1-10] AutoML 솔루션 새로운 흐름



* 출처: H2O.ai, 「H2O.ai is a Leader in The Forrester New Wave™: Automation-Focused Machine Learning Solutions, Q2 2019」, 2019.5.

Forrester(2019)의 보고서[8]는 [그림 5-1-10]과 같이 AutoML 시장에 대해서 현재 오퍼링(Current offering)과 시장 상황 (Market presence)의 정도에 따라 4가지 영역(Leaders: 리더, Strong Performers: 강력한 실행자, Contenders: 경쟁자, Challengers: 도전자)으로 구분하였다. 이 보고서는 해당 Aible, Bell Integrator, Big Squid, DataRobot, DMway Analytics, dotData, EdgeVerve, H2O.ai, Squark 등의 솔루션 중에서 DataRobot과 H2O.ai를 리더그룹으로 평가하였다. 특히 DataRobot은 사용자 경험, 모델 운영, 비전, 로드맵을 포함한 10가지 평가 기준 중 9가지에서 높은 평가를 받았다. 최근에 2021년 AIMultiple의 AutoML Software[9]에 따르면 [표 5-1-3]과 같이 브랜드 인기도과 웹 트래픽에 있어 DataRobot, Dataiku DSS, H2O.ai가 Top3로 조사되었다.

[표 5-1-3] Market Presence Metrics

	DataRobot	Dataiku DSS	H2O.ai	Others
인기도	43%	38%	6%	13%
웹 트래핑	45%	45%	4%	6%

* 출처: AI Multiple, 「Top 24 AutoML software of 2021: In-Depth Guide」, 2021.

국내 상용 AutoML솔루션 시장에서도 현재 DataRobot, Dataiku DSS, H2O.ai가 제품 지원과 마케팅, 영업활동을 하고 있다. Gartner peerinsights(2021)[10]의 Data Science and Machine Learning Platforms 평가 기반으로 DataRobot, Dataiku DSS, H2O.ai 등에 대해서 [표 5-1-4]과 같이 정리해보았다. 각 솔루션에 대한 사용자의 선호적 평가와 비판적 평가를 살펴보면, DataRobot은 데이터 사이언티스트가 아닌 경우도 현실 비즈니스 문제를 빠르게 처리하여 생산성이 높다는 측면에서 AutoML 철학에 부합되는 솔루션이라는 평과 함께 최소한의 유연성만 허용되는 점이 아쉽다는 지적이 있다. DataRobot의 특징은 아무리 복잡한 모델이라 하더라도 PDP, Permutation Feature Importance, XEMP, Shapley Values 등을 통해 피처 및 모델에 대한 설명이 가능한 일관성 높은 인터페이스 화면을 제공하고, DataRobot 노하우가 반영된 블루프린트 기반의 다양하고 수많은 알고리즘을 병렬로 학습, 검증, 블랜딩하여 높은 정확도의 모델을 추천하다는 점이 특징적이다.

8 H2O.ai, 「H2O.ai is a Leader in The Forrester New Wave™: Automation-Focused Machine Learning Solutions, Q2 2019」, 2019.5.
9 https://aimultiple.com/automl-software (접속일: 2021.11.2.)
10 https://www.gartner.com/reviews/market/data-science-machine-learning-platforms (접속일: 2021.11.2.)

손쉽고 편리하게 모델 배포와 모니터링을 관리하고, 자동으로 모델 생애주기(Lifecycle)가 관리되어 서비스 중단 없이 모델 재학습, 교체 등의 운영이 가능하다. 또한 모델 운영 관리에 필요한 서비스, 정확도, 데이터 드리프트 등과 같은 MLOps 서비스 화면이 잘 정비되어 있다. Dataiku DSS의 경우 데이터 사이언티스트와 함께 워크플로우를 공유하면서 쉽게 분석할 수 있고, R, Python, SQL등의 코딩도 활용할 수 있다는 점이 특징이다. 다만 폭넓은 범위의 산업에 활용되기 위해서는 대규모 데이터셋 처리에 대한 개선이 필요하다는 평도 있다. Dataiku DSS의 강점은 워크플로에 기반한 유연한 데이터 처리와 피처 생성 그리고 수동 (R/Python)코딩을 워크플로우에 삽입할 수 있다는 점이 특징이다. H2O.ai의 경우 머신러닝 수행을 위한 적절한 옵션 및 오픈 API를 활용하여 맞춤 설계와 배포가 가능하다. 다만 여러 번 실행 시 클러스터 안정성이 부족하고 웹 인터페이스에서 소프트웨어 활용이 제한적이라는 평도 있다. H2O.ai는 피처 엔지니어링 관점에 자동으로 상호작용이 높은 피처를 생성하고 모델 설명을 위한 PDP, ICE, Permutation Feature Importance, Shapley Values, LIME 등 다양한 시각화 화면을 제공하고 있다.

[표 5-1-4] Gartner Peer Insights(2021)의 Customer Reviews

평점 요소	DataRobot	Dataiku DSS	H2O.ai
전반적 평점	4.7	4.8	4.7
평가 및 계약	4.5	4.5	4.7
통합 및 배포	4.5	4.5	4.6
서비스 및 지원	4.7	4.8	4.9
제품 기능	4.6	4.8	4.6

* 출처: Gartner Peer Insights, 「Data Science and Machine Learning (ML) Platforms Reviews and Ratings」, 2021

나. AutoML 요구사양

AutoML 요구사양에 있어서 사용자의 접근성이 높아야 한다. 데이터 사이언티스트, 데이터엔지니어, IT 전문가, 현업 비즈니스 실무자 등 폭넓은 사용자들이 편리하게 사용할 수 있어야 한다. 더불어 데이터 거버넌스, 데이터엔지니어링, 피처 엔지니어링, 모델 최적화 구현, 모델 설명 해석 가능성 그리고 MLOps 기능 까지 모든 영역에서 사용자의 편리성, 모델의 다양성, 효율성, 정확성, 모델 운영/관리성 등이 확보되어야 한다.

[표 5-1-5] AutoML 평가요소별 기능요소

편리성	다양성/확장성	효율성/정확성	운영/관리성
• 자동화 모델 생성 • 자동 EDA과 데이터 품질 추적과 관리 • 하이퍼 파라미터 튜닝 • 모델 성능 평가 지표 및 모델 비교 제공 • 피처 중요도 • 피처 관계 해석 • 피처 엔지니어링 기능 • API 배포코드 및 보고서 기능	• 다양한 소스(CSV, DB, Hive, Spark, HDFS 등) 연계 • 다양한 데이터 유형(정형, 이미지, 텍스트, 공간정보 등)에 따른 분석 활용 • 최신의 다양한 머신러닝/딥러닝(지도학습, 비지도학습, 시계열, 이상 탐지 등) 알고리즘 제공	• 다수 모델링 자동 병렬 실행을 통한 신속한 모델 구현 • 임계치 변동시뮬레이션 기능 • 모델 투자수익률(ROI) 시뮬레이션 기능 • 다수의 모델 설계 노하우 기반 분석 시나리오 • 모델 간 오토 앙상블 및 Blend 기능	• 배포 모델 서비스 모니터링 기능 • 배포 모델 정확도 및 데이터 변화 추적 모니터링 기능 • 배포 모델 재학습과 모델 교체 기능 • 프로젝트 공유 및 이력 관리, 계정 권한 관리 • 프로젝트 자원 및 리소스 관리

7. 시티즌 데이터 사이언티스트 관점에서의 AutoML

최근에는 AutoML 제품을 활용하여, 비전문 분석인력(현업 비즈니스 실무자, IT 전문가, 데이터엔지니어)들도 직접 코딩하지 않더라도 머신러닝 프로세스에 참여 할 수 있게 되었다. 일반적으로 시티즌 데이터 사이언티스트는 영업, 마케팅, 재무 또는 인적 자원과 같은 비즈니스 라인에서 직면한 비즈니스 과제에 대한 도메인 지식을 가지고, 데이터 과학자, 통계학자 또는 수학자들이 했던 작업의 지원을 받아 분석한다. 시티즌 데이터 사이언티스트는 해당 분야에서 일반적인 분류, 회귀 분석과 같은 비교적 용이한 문제를 해결하는 역할을 담당하고, 난이도가 높은 지식이 필요한 경우, 데이터 사이언티스트의 지원을 받아 문제를 해결하는 것이 바람직하다. 데이터 사이언티스트의 경우, 전문적이고 복잡하고 난이도가 높은, 잘 해결되기 어려운 분석 프로젝트에 집중할 수 있게 하거나 기업에서 최대한 많은 비전문 분석 인력이 시티즌 데이터 사이언티스트의 역량을 높일 수 있도록 기술 지원과 자문하는 것이 바람직하다. 시티즌 데이터 사이언티스트은 머신러닝과 인공지능에 대한 전문 지식과 코딩 능력이 없는 경우가 대부분이기 때문에 오픈소스 기반보다는 모델 유지 관리가 편리하고 인터페이스가 잘 갖추어진 상업용 AutoML를 활용하는 것이 효과적이다.

지금까지는 데이터 사이언티스트가 직접 R이나 Python 코딩을 하지 않았으며, 머신러닝 모델을 구현할 수가 없었고, 최신 알고리즘보다는 기존에 자주 사용했던 알고리즘만 사용하는 경우가 많아서 배포된 모델이 최선이라고 보장할 수도 없었다. 그리고 분석 중 최신 알고리즘을 검토하고 적용할 시간도 충분하지 않았다. AutoML 제품은 데이터 준비가 완료되면 피처 엔지니어링와 모델 학습 및 검증, 배포, 서비스 구현까지 실무에 바로 적용할 수 있도록 종단간(End to End) 플랫폼으로 발전하고 있다. 앞으로 AutoML를 활용하면 데이터 기반 의사결정과 문제해결에 대한 진입장벽이 낮아질 것이다. 그래서 시티즌 데이터 사이언티스트가 기존보다 많은 비즈니스 문제에 있어 쉽고 빠르게 최적의 모델을 구현하게 될 것이며, 기업의 생산성 증대와 투자수익률(ROI)을 개선하는 데 더 가치 있게 기여하게 될 것이다.

제2장.
Cloud Computing

고태우
(KDB산업은행 산업기술리서치센터 차장)

클라우드 기업은 소프트웨어 아웃소싱 증가, 소비자 트렌드에 맞는 IT 솔루션 수요 확대에 따라 인프라서비스(Infra as a Service, 이하 IaaS), 소프트웨어서비스(Software as a Service, 이하 SaaS), 플랫폼서비스(Platform as a Service, 이하 PaaS)를 On-premise, Private, Public, 하이브리드 등 고객 환경에 적합한 형태로 제공 중이며, 본문에서는 클라우드 컴퓨팅의 개요, 클라우드 산업 생태계, 국내외 시장현황을 살펴보고자 한다.

1. 클라우드 컴퓨팅 개요

가. 개념과 특징, 그리고 변화

클라우드 컴퓨팅(Cloud Computing, 이하 클라우드)는 사용자가 필요한 IT인프라(스토리지, 네트워크, 서버 등)를 직접 물리적으로 구축하여 소유 관리하는 기존 방식(On-premise)과는 달리, 클라우드 서비스 업체가 구축해 놓은 가상화된 IT 자원을 네트워크를 통해 빌려서 필요에 따라 사용하고, 사용한 만큼 비용을 지불하는 방식의 서비스로, 소유와 관리가 분리된 방식이다.

클라우드를 활용하면 개별 사용자의 즉각적인 요구에 따라(On-demand) 최적화된 서비스 이용이 가능하므로, 시간과 장소, 기기의 제약이 없고 급격한 이용량 증가에도 유연한 확장이 가능하다는 특징이 있다.

1980년대 대형 범용 컴퓨터(메인 프레임) → 1990년대 단말컴퓨터의 성능이 향상되며 분산처리가 가능 → 2000년대 네트워크의 속도가 향상되어 기업 내 인트라넷 구축 → 2010년대 클라우드 컴퓨팅이 등장하며 기업들의 IT인프라(계산, 저장, 네트워크)를 특정 클라우드 서비스 업체가 제공한다.

[표 5-2-1] 컴퓨팅 환경의 변화

컴퓨팅 환경	개인용 컴퓨팅 환경	서버-클라이언트 환경	클라우드 컴퓨팅 환경
데이터 위치 및 컴퓨팅 주체	개인용 PC, 노트북	서버/클라이언트	클라우드 서버(온라인)
자원 구매/폐기	이용자	이용자	서비스 제공자
사용자 컴퓨터 설치 S/W	OS, 응용 S/W	OS, 응용 S/W, 클라이언트	클라이언트(웹브라우저)
데이터의 소유 및 관리	소유와 관리가 동일	소유와 관리가 일부 분리	• 소유와 관리 분리 - 소유: 이용자 - 관리: 서비스 제공자
제공 서비스	오프라인 컴퓨팅 서비스 (문서작성, 통계 계산, 그래픽 작성 등)	• 기본 인터넷 서비스 - 웹, FTP, 이메일 등 • 응용 인터넷 서비스 - 웹하드, SBC, ASP 등 • IT 융합서비스 - VoIP, IPTV 등	• 가상 서버/데스크탑 서비스 • 스토리지 제공 서비스 • S/W 임대서비스 등

* 출처: 한국인터넷진흥원, 「클라우드 서비스 정보보호 안내서」, 2011.10.11.

나. 클라우드의 일반화

iCloud, Gmail, Google Photo, Google Drive 등의 서비스를 통해 클라우드 기술은 일상생활 속에서 쉽게 접할 수 있다.

클라우드의 기본개념은 1960년대 인터넷의 전신인 ARPANET(Advanced Research Project Agency Network)의 개발자인 릭라이더(Licklider)가 'Intergalactic Computer Network'에서 최초로 언급했으며, 초기 아이디어를 바탕으로 그리드컴퓨팅, 유틸리티컴퓨팅, 애플리케이션 서비스 제공자(Application Service Provider, ASP), 네트워크컴퓨팅 등 다양한 개념으로 진화하였다.

세일즈포스닷컴(Salesforce.com)이 웹사이트 상에 기업 애플리케이션을 제공하는 서비스를 출시함으로써 클라우드의 개념이 현실에 구현되기 시작했다. 이어서 아마존이 2002년도 AWS(Amazon Web Services)를 시작으로, 2006년 EC2(Elastic Compute Cloud)라는 상업용 클라우드 서비스를 제공해 컴퓨터 구입 없이 빌려 쓰는 형태로 응용프로그램을 실행할 수 있게 되면서 클라우드가 일반화되기 시작하였다.

다. 클라우드의 장점

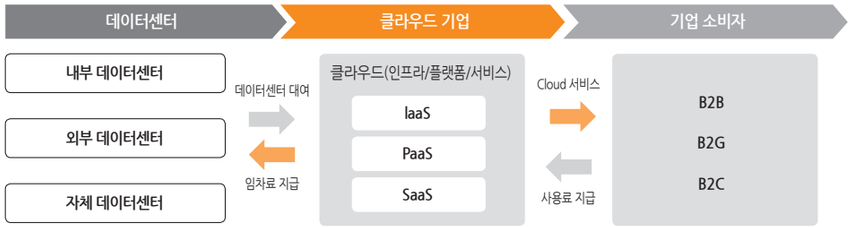
많은 기업체나 기관들이 필요한 IT 인프라 관리 및 데이터 백업, 재해 복구 지원을 위해 지출하는 비용을 절감하고, 하드웨어 구입 시 필요한 배송 및 설치시간 감소, 필요한 전산 자원(컴퓨터 성능, 저장소, 대역폭)의 쉬운 확장, 전산 인프라 관리, 하드웨어 설치, 소프트웨어 패

치와 같은 반복적인 IT 관리 작업 감소로 필요한 핵심 목표에 집중 가능, 최신 컴퓨팅 자원 제공 및 지역에 관계없이 서비스 제공이 가능하다.

2. 클라우드 산업 생태계

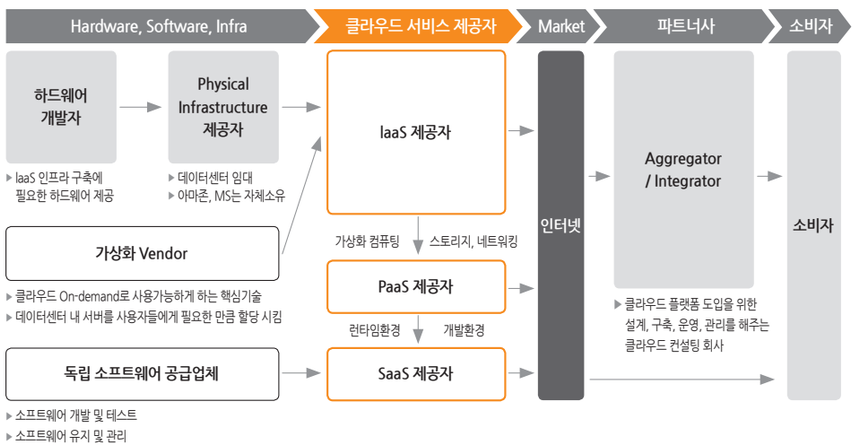
온라인 서비스의 확대로 성장 중인 클라우드 산업 생태계는 후방산업(데이터센터)을 기반으로 클라우드 서비스 제공자와 클라우드 파트너사의 협업으로 다양한 온라인 서비스 및 애플리케이션 서비스를 제공한다.

[그림 5-2-1] 클라우드 사업구조



* 출처: 저자 작성

[그림 5-2-2] 클라우드 생태계 Value Chain



* 출처: 저자 작성

가. 데이터센터

구글, 아마존, MS 등의 해외 기업이 자사의 인프라(대여한 데이터센터와 클라우드 자원)를 기반으로 다양한 플랫폼 기술을 제공하며, 국내 클라우드 서비스 제공자(KT, NHN, NBP, 더존비즈온 등)는 자체 데이터센터를 보유 중이다.

[표 5-2-2] 외국계 클라우드 국내 진출 현황

기업명	시기	내용	위치(임대기업)
아마존	2016년	서울 데이터센터 오픈(1개 리전, 2개 가용영역)	목동(KT), 일산(SKB)
	2019년	서울 데이터센터 확장(1개 가용영역 추가)	용인(롯데정보통신)
	2020년	서울 데이터센터 확장	평촌(LGU+)
MS	2017년	서울과 부산에 데이터센터 구축(2개 리전)	평촌(LGU+), 김해(LG-CNS)
	2018년	부산 데이터센터 2개 추가 구축	부산(MS 자체 구축)
	2019년	서울 데이터센터 확장	목동(KT 2센터)
	2021년	서울 데이터센터 확장	용산(KT)
구글	2020년	서울 리전 설립(3개 가용영역)	평촌(LGU+), 가산(LGU+)
	2021년	서울 리전 확장(3개 가용영역)	용산(KT)
IBM	2016년	서울 데이터센터 오픈(1개 리전, 1개 가용영역)	판교(SK C&C)
오라클	2019년	서울 데이터센터 오픈(1개 리전, 1개 가용영역)	목동(KT)
	2020년	춘천 리전 오픈	춘천(삼성 SDS)

* 출처: 저자 작성

구글 클라우드의 서울 데이터센터 리전(Region) 설립 등 외국계 클라우드 기업들(아마존, MS, 오라클 등)이 시장 공세를 강화 중에 있으며, 국내 기업인 KT, 네이버의 NBP(Naver Business Platform)는 시장의 약 20% 점유율로 열세를 보이고 있다.

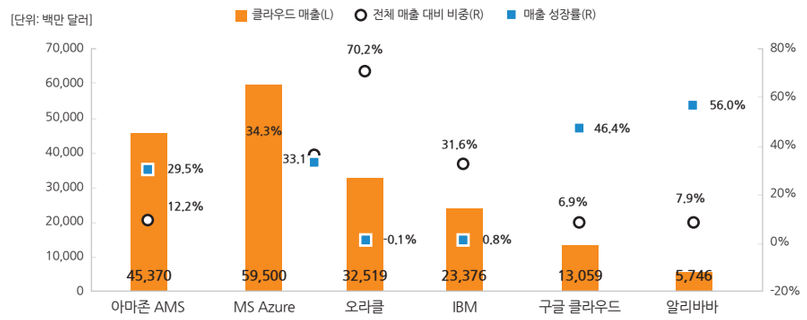
나. 클라우드 서비스 제공자와 파트너사

클라우드 사업구조의 Value Chain의 중심에 있는 주요 기업들은 설비투자를 통해 자체 데이터센터를 확보하거나, 내외부에 있는 데이터센터를 대여하여 대 고객 클라우드 서비스를 제공한다.

외국계 주요 벤더사는 클라우드 인프라, 플랫폼을 제공하고 있으며, 2020년 매출은 IaaS,

PaaS, SaaS를 합산한 기준으로 마이크로소프트(MS), 아마존, 오라클, IBM, 구글, 알리바바 순이다.

[그림 5-2-3] 외국계 클라우드 서비스 제공자별 매출 및 성장률, 매출 대비 비중(2020)



* 출처: 미래에셋증권 리서치센터, 「글로벌소프트웨어: 글로벌 클라우드 산업 현황」, 2021.4.14.

클라우드 파트너사는 엔드유저에게 클라우드 구축 및 운영 등의 공급과 특화된 솔루션을 제공하며, 클라우드 서비스 제공자에게 자격을 취득하여 클라우드 사업을 수행한다.

최근 IaaS에서는 선두업체의 과점화가 진행되고 있으며 IaaS·PaaS 통합 플랫폼화와 SaaS 분야의 참여자 확대도 동시에 이루어지면서 클라우드를 효율적으로 활용하기 위한 클라우드 파트너의 필요성이 점차 중요시되고 있다.

다. 클라우드 서비스

클라우드 서비스는 온라인 서비스와 비즈니스 애플리케이션 형태로 제공되며, 클라우드 파트너사를 통한 구축형 프로젝트로 서비스를 개발한다.

IaaS분야는 시장 선점을 위해 서버, 네트워크, 스토리지 등 IT 인프라 대상으로 대규모 투자가 필요한 분야로 자본력을 갖춘 글로벌 기업들이 시장을 과점해 가는 상황이며, 최근에는 대규모 IT 인프라 구축을 기본으로 데이터베이스, 분산처리, AI 등의 PaaS 플랫폼이 결합되는 형태로 경쟁방식이 다양해지는 상황이다.

SaaS분야는 SW솔루션 기업들이 기존 구축형 제품을 클라우드 기반으로 변형하면서 자연스럽게 참여가 이루어지는 분야로 다양한 기업들이 경쟁하고 있으며, 최근에는 IaaS의 경쟁력 강화에 SaaS솔루션 도입의 효용성이 입증됨에 따라 IaaS기업과 다양한 Third Party 솔루션 개발사의 연계가 진행 중에 있다.

MS가 MS office를 SaaS(Office365) 형태로 제공하여 동사의 IaaS 경쟁력 강화에 활용하는 것처럼 구글도 Google Workspace라는 유사 애플리케이션을 제공하면서 시장에 참여했다.

[표 5-2-3] 클라우드 서비스 세부 유형 분류

서비스 유형		세부내용
IT 자원 유형	IaaS	• 스토리지, 서버 등 컴퓨팅 인프라 구축에 필요한 가상 하드웨어 자원 제공 • 자체 인프라 구축에 부담을 느끼는 중소기업이 주요 대상 • MS(Managed Service)를 통한 효율적 관리가 중요
	PaaS	• 프로그램이나 애플리케이션을 개발하는데 필요한 툴이나 프레임워크 등 제공 • 고가의 장비나 개발 툴을 구매하지 않고 소프트웨어 개발이 가능 • MS를 통한 효율적 관리가 중요
	SaaS	• 소프트웨어나 애플리케이션을 엔드유저에게 제공 • 사용자는 소프트웨어를 직접 구매해 자신의 단말기에 설치하는 대신 웹 접속을 통한 임대 형태로 사용
서비스 개방 여부	퍼블릭	• 사용대상에 제한을 두지 않으며, 데이터센터 내의 컴퓨팅자원을 불특정 다수의 사용자들이 공용으로 이용(Multi-tenant) • 초기 투자비용이 없기 때문에 경제적이나, 서비스 제공자에 대한 의존도가 큼 • 가장 기본적인 클라우드 서비스로 서비스 제공자가 운영하는 IaaS와 PaaS 시장의 대부분을 차지 • (싱글/멀티 방식) 싱글 방식은 한 업체에 종속되며, 멀티 방식은 퍼블릭-퍼블릭 결합 형태로 단일 클라우드의 취약점(서버 부하 등)을 보완하기 위해 활용함
	프라이빗	• 단일 고객에게 클라우드 서비스 제공 • 대규모의 투자비용이 소요되지만 데이터 및 인프라에 대한 통제권이 크기 때문에 보안에 강점이 있음 • 컴퓨팅 자원을 특정 사용자(Single-tenant)에게 할당하여 배타적으로 사용 가능 • 사용자의 기존 데이터센터 內 데이터센터를 구축하는 방식도 가능(On-premise private cloud)
	하이브리드	• 퍼블릭-프라이빗 클라우드 서비스를 혼용하는 방식 • 시스템 자원을 효율적으로 사용하면서도 비용절감이 가능 • 공유를 원하지 않는 일부 데이터 및 서비스에 대해서는 프라이빗 클라우드 형태를 활용

* 출처: 미래창조과학부, 한국과학기술정보연구원, 「퍼스널 클라우드의 기술동향」, 2013.9., 저자 재구성

※ MSP(Managed Service Provider) 서비스의 발전

1) IaaS의 과점화와 SaaS솔루션의 다양화는 클라우드 도입기업 입장에서 최적화 문제를 야기함에 따라 적합한 클라우드 사용환경을 제시해 줄 수 있는 MSP가 등장하게 되었다.

2) IaaS 기업들이 경쟁력 확보를 위해 PaaS, SaaS서비스를 확대한 결과 AWS가 100개가

넘는 서비스를 제공하는 등 기업들의 선택권은 확대되었으나, IT예산 제약을 가진 기업은 도입 장애 요인으로 인식하기 시작하였다.

3) 2008년 글로벌 금융위기를 기점으로 등장한 MSP는 도입 기업의 안정적인 클라우드 운영을 보장함에 따라 클라우드 최적화 및 운영 지원 서비스와 관련된 비즈니스를 발전시키게 된다.

4) MS 분야의 성장에 따라, 기존의 웹/서버 호스팅 업체, IT컨설팅 업체(Accenture) 등 다양한 기반의 MPS들이 등장하기 시작하였으며 최근에는 경쟁에서 뒤쳐진 IaaS 후발주자(Rackspace 등)까지 참여가 확대되고 있다.

한편, IaaS 사업자 입장에서도 대규모 IT 인프라를 관리하면서 지역, 언어, 통화 등을 감안한 전문화된 서비스 제공에는 한계가 있어 MSP와의 파트너십이 강화되는 추세다.

[표 5-2-4] Managed Service(MS) 내용

자원관리	서버(컴퓨팅)자원, 패치관리(최신 업데이트, 패치 적용), 액세스관리(여러 인증 체계의 복잡성을 관리)
보안관리	침입탐지, 침입방지시스템 등을 통한 정보자산 보호
연속성관리	비즈니스 장애 및 중단 시 복원, 예약 백업 등 실시
보고	모든 유형의 인프라를 기업이 통합된 형태로 파악할 수 있도록 IT 서비스 관리 도구와 통합 서비스 제공

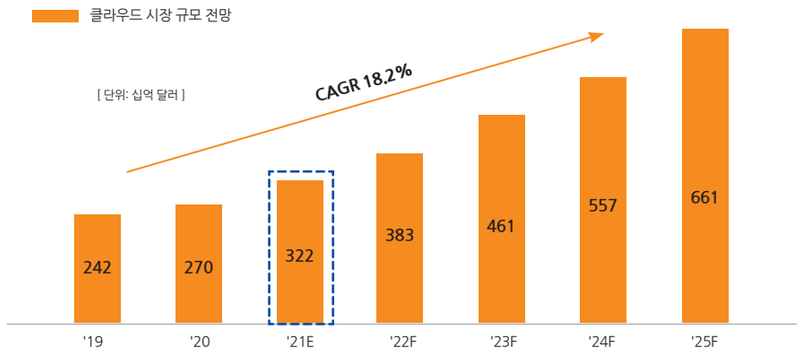
* 출처: 저자 작성

3. 국내외 시장현황

가. 글로벌 클라우드 시장

2021년 기준 시장 규모 3,220억 달러로 2019년부터 2025년까지 연평균 18.2% 수준의 성장을 지속할 것으로 예상된다.

[그림 5-2-4] 클라우드 시장 규모 전망



* 출처: 가트너, 「Forecast : Public Cloud Services, Worldwide, 2019-2025」, 2021.4., 미래에셋증권, 「글로벌 클라우드 산업 현황」, 2021.4., 저자 가공

클라우드 서비스 유형별로는 2021년 기준 IaaS의 성장률이 29.5%로 가장 높고, PaaS 19.7%, SaaS 17.8%, BPaaS(Business Process as a Service) 아웃소싱 서비스 5.2% 순이다.

[표 5-2-5] 클라우드 서비스 유형별 성장률 전망

구 분	'19	'20	'21	'22	'23	'24
PaaS	23.6%	24.8%	19.7%	23.6%	24.6%	19.9%
SaaS	0.9%	15.7%	17.8%	18.7%	18.8%	17.2%
BPaaS	2.6%	6.1%	5.2%	6.7%	8.2%	7.6%
보안/관리	12.2%	8.3%	11.6%	10.4%	10.4%	9.7%
IaaS	33.3%	33.4%	29.5%	29.4%	27.5%	25.4%
전체	11.5%	19.3%	18.9%	20.4%	20.7%	18.8%

* 출처: KRG, 「4차 산업혁명 관련 주요 이슈 브리핑」, 2019., 저자가공

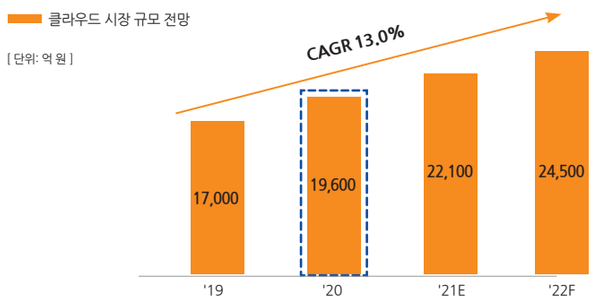
나. 국내 클라우드 시장

2017년 이전에는 국내 대기업들은 클라우드 서비스로의 전환이 신중한 편이었으나, 2017년 쿠팡을 시작으로 2018년에는 대기업 최초로 대한항공이 클라우드를 전사적인 IT

인프라로 전환하면서 클라우드 서비스 시장에 대한 적극적인 진출이 시작되었다.

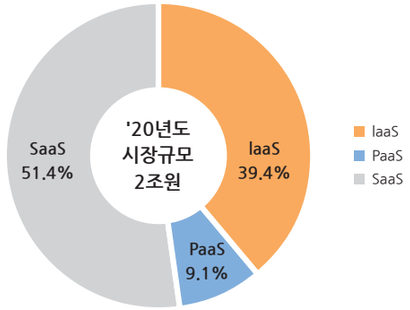
국내 퍼블릭 클라우드 시장 규모는 2020년 기준 2조원 수준이며, 각 서비스별 비중은 Saas(51.4%), IaaS(39.4%), PaaS((9.1%) 순이다.

[그림 5-2-5] 국내 클라우드 시장 규모 및 전망



* 출처: KRG, 「IT 시장백서(2020)」, 저자 가공

[그림 5-2-6] 국내 클라우드 시장 규모 및 서비스별 비중

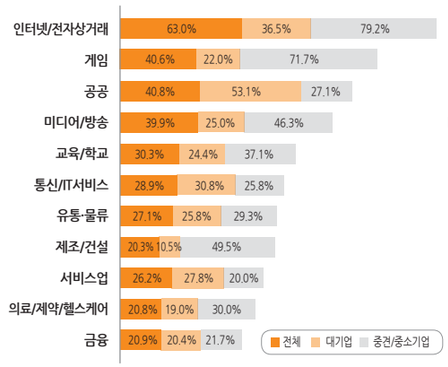


* 출처: IDC, 「Semiannual Public Cloud Services Tracker」, 2021

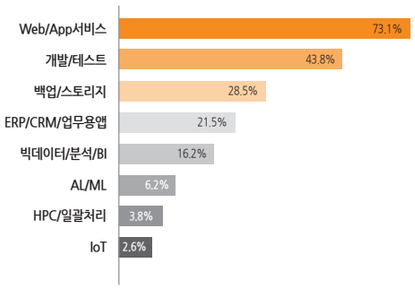
국내 클라우드 주요 사업자는 자체 데이터센터 기반의 특화된 서비스로 외국계 클라우드 기업과 경쟁 중이며, 온라인 관련 서비스를 중심으로 클라우드 활용 비중이 높다.

특히 인터넷/전자상거래(63%), 게임(40.6%), 미디어/방송(39.9%) 등에서 높은 클라우드 활용률을 보이고 있으며, 클라우드 활용 업무 영역에서는 Web/App서비스가 73.1%의 활용률을 보여 압도적인 비중을 차지하고 있다.

[그림 5-2-7] 국내 업종별 클라우드 시스템 활용 현황



[그림 5-2-8] 국내 업무별 클라우드 사용 현황



* 출처: 베스핀글로벌, 「2019 국내 클라우드 도입의 현주소」, 2019.5.29.

국내 기업들은 규모와 무관하게 빅데이터, 미디어, AI 등의 분야에 클라우드 기술을 활용하는 것을 선호한다.

[표 5-2-6] 국내 주요 클라우드

amazon	개발자 생태계를 초기에 장악하여 개발자들이 가장 선호 가격을 선도, 지속적 가격인하에도 매출, 영업이익 상승 머신러닝, AI기반의 PasS, SaaS로의 진화
kt	국내 기업 중 가장 큰 클라우드 시장 점유율 보유 공공기관 대상의 G-클라우드에 강점 글로벌 회사 대비 가성비가 좋다는 평가를 받음
servicenow	포트폴리오 다각화를 통해 매출을 크게 견인 토탈 기업솔루션을 목표로 모든 기업 내 비즈 앱을 하나의 시스템으로 통합시키고자 함
Microsoft Azure	• 하이브리드 클라우드 시장을 선점하여 고성장 지속 • AI기술을 활용한 개발 툴과 플랫폼 제공, 오피스365 점유율을 바탕으로 IaaS, PaaS 고객 유치
NAVER	• 한국은행, 코레일 등 고객사 확보(기업 고객 6,000여개사) • 최근 금융/공공 영역에 집중해 클라우드 포털 별도 운영 • 현재 15개 카테고리 119개 상품 보유 중
salesforce	• SaaS 시장에서 포트폴리오 다각화, 손쉬운 개발환경을 제공해 자체 생태계 구축 • 적극적인 M&A로 데이터분석 및 시각화에 강점 보유

4. 마무리

가트너는 2019년부터 2025년까지 전 세계 클라우드 시장이 연평균 18.2% 성장할 것으로 예상하였다. 가장 큰 시장은 SaaS이지만 최근 IaaS 영역이 빠르게 성장 중이며, 글로벌 빅테크 기업인 MS와 아마존의 압도적인 시장 점유율로 해당 시장을 선도하고 있다. 국내 시장은 2020년 기준 2조원 규모이며, 국내 클라우드 주요 사업자는 자체 데이터센터 기반의 특화된 서비스로 외국계 클라우드 기업과 경쟁 중이다.

최근 글로벌 빅테크 기업들의 한국 내 데이터센터 확보가 급속히 늘어남에 따라 국내 시장의 성장세가 가속화되고 있으므로, 클라우드 생태계 내 데이터센터, 클라우드 기업(클라우드 서비스 제공자, 파트너사)에 대한 관심을 가지고 공공과 민간이 잘 활용할 수 있는 기반을 갖추는 것이 필요하다.

제3장. 개인정보 비식별 처리 수준 및 프레임워크

윤덕상
((주)파수 전략사업본부장)

2020년 8월 5일 데이터 3법이 발효되면서 드디어 우리나라에서도 개인정보를 가명·익명처리를 하면 수집 목적 외의 활용이 가능해져 그동안 답보 상태에 있던 데이터산업이 인공지능(AI), 빅데이터 등 4차산업혁명을 기반으로 크게 성장할 것이라고 기대하고 있다. 하지만 법에서 요구하는 안전한 비식별 처리 수준의 정의와 기준의 모호성은 아직까지도 많은 기업들이 데이터 비식별 처리에 선뜻 나서지 못하게 하는 장애 요인이 되고 있다. 이러한 고민은 비단 우리나라뿐만 아니라 우리보다 앞서 개인정보 활용을 허용해온 다른 많은 국가들의 공통된 과제이기에 하다. 이를 위해 북미나 유럽연합(EU)에서는 정부기관이나 연구기관들이 앞장서서 개인정보 비식별 처리담당 실무자들이참고하여 비식별 수준을 설정하고 위험도를 평가할 수 있도록 비식별 처리 방법론이나 프레임워크를 제공하고 있다. 우리나라도 각 분야별 「가명정보 처리 가이드라인」을 통해 가명처리 전 가명화 수준을 설정하고, 가명처리 후 가명화 수준의 적절성을 평가하는 방법을 제공하고 있다.

1. 비식별 위험 요소와 데이터 상황

가. 비식별 데이터의 안전성

세계적으로 개인정보 익명화와 관련하여 다양한 연구와 지침서를 제공하고 있는 영국의 민간조직인 영국 익명화 네트워크(UKAN)[1]는 "비식별화는 목적에 맞는 유용한 데이터가 생성되어야만 의미가 있으며 유용하면서도 리스크가 없는 데이터를 생산하는 것은 현실적으로 불가능하다"를 비식별 원칙으로 선언하고 있다. 따라서 개인정보 데이터는 완벽히 안전하게 비식별 처리를 하기는 불가능하며 이렇게 생성된 데이터는 다양한 위협 주체들로부터 여

1 영국 익명화 네트워크(The UK Anonymisation Network, UKAN)는 개인정보의 활용에 대한 실용적인 조언과 정보를 제공하기 위한 목적으로 2012년 설립되어 개인정보 익명화와 관련한 다양한 연구 및 지침서를 제작하여 제공하고 있다.

러 유형의 재식별(Re-identification) 공격 위협에 노출되어 있다는 것을 의미한다.

[표 5-3-1] 재식별 공격 위협

위협요인	내 용
내부 공격자	특정 지식·기능·권한을 보유한 내부자가 비식별화에 이용된 기밀 또는 여타 중요한 정보를 획득할 수 있으며, 데이터 컨트롤러[2], 데이터 프로세서[3], 제3자가 모두 내부 공격자가 될 수 있음. 단, 감독관이나 법 집행 기관 등 개인정보에 합법적으로 접근할 수 있는 제3자는 잠재적인 내부 공격자에서 제외
외부 공격자	비식별 처리의 기밀이나 기타 관련 정보에 직접 접근할 수 없으나 비식별화 데이터셋에 접근하여 추가적인 정보를 파악 또는 적용함으로써 공격을 시도하며, 일반적인 외부자 이외에 악의적인 데이터 컨트롤러와 악의적인 데이터 프로세서도 외부 공격자 역할을 할 수 있음
공격의 목적	공격이 발생하는 상황과 가명처리 방법 등에 따라, 공격자는 ①가명처리 기밀 인출(retrieval of the pseudonymisation secret)[4], ②완전한 재식별화(complete re-identification), ③판별(discrimination) 등을 목적으로 공격을 시도
주요 공격 기술	가명화된 데이터에 대한 공격은 ①무차별 대입 공격(brute force attacks 또는 exhaustive search), ②사전검색(dictionary search), ③추측(guesswork) 등의 방식으로 이루어지며, 이 같은 공격 기술의 효과는 정보의 양, 배경지식, 가명처리 기법의 특징 등에 따라 달라질 수 있음

* 출처: 한국인터넷진흥원, 해외 개인정보보호 동향 보고서 「ENISA의 가명처리 기법 및 활용사례 보고서 분석」, 2020.1.

나. 비식별 처리 유형

비식별화의 주요 목적는 비식별 처리된 데이터셋(dataset)과 정보주체 간의 연결성을 제한하여 정보주체의 신원을 보호하는 것이지만, 앞서 언급한대로 유용성을 감안해야하기에 현실적으로 완벽한 비식별 처리는 불가능하다. 이러한 상황을 고려했을 때 Elliot[5]등은 비식별 처리영역과 처리방법 등을 고려하여 선택 가능한 비식별 처리 유형을 다음과 같이 4가지 유형으로 분류하였다.

2 컨트롤러는 개인정보의 처리 목적 및 수단을 단독 또는 공동(jointly)으로 결정하는 자연인, 법인, 공공기관, 에이전시 기타 단체를 의미

3 프로세서는 컨트롤러를 대신하여 개인정보를 처리하는 자연인, 법인, 공공기관, 에이전시, 기타 단체를 의미

4 가장 심각한 공격으로, 해당 데이터 세트에서 재식별이 이루어지는 것은 물론 해당 데이터 세트에 대해 임의로 가명처리 프로세스를 수행하는 것도 가능

5 Elliot, M. J., Dibben, C., Gowans, H.,Mackey, E., Lightfoot, D., O'Hara, K., and Purdam, K, 「Functional Anonymisation: The crucial role of the data environment in determining the classification of data as (non-) personal,」 CMIST work paper, 2015

[표 5-3-2] 비식별 처리의 유형

처리 유형	방법
정형 비식별 (Formal Anonymisation)	특정 개인을 구분할 수 있는 식별자를 제거하거나 마스킹(masking)하여 방법(재식별화 가능성을 1이하로 낮추는 메커니즘)
보장형 비식별 (Guaranteed Anonymisation)	특정 개인이 식별될 수 있는 위험을 제로(zero)에 가깝도록 확실히 보장하고 복구가 불가능하게 데이터를 처리하는 방법. 하지만 유용한 데이터를 공유하고자 할 때 일반적으로 보장형 익명처리는 현실적으로 가능하지 않을뿐더러 실용적이지 않음(재식별 가능성을 0으로 만드는 메커니즘)
통계적 비식별 (Statistical Anonymisation)	재식별 가능성을 0의 확률로 축소하는 것이 어렵기에 특정 속성값에 대한 노출 빈도를 통계적으로 균일하게 처리[6]하는 비식별 처리로써 데이터의 속성값에 적용한다는 부분에서 정형 익명처리 대상과 차이가 있음(정형, 보장형 익명처리 모두 통계적 익명처리의 특별한 경우임)
기능적 비식별 (Functional Anonymisation)	데이터 속성 자체뿐만 아니라 데이터 상황이라 불리는 데이터 활용 목적과 환경에 따라 비식별 처리 기법을 결정하는 방식으로 재식별이 미치는 영향, 노출이 가능한 데이터, 비식별 데이터의 관리 방법, 재식별 가능성을 분석하여 적합한 방식을 선정하는 방법

* 출처: UKAN, 「The anonymisation decision-making framework」, 2016.7., 저자 재구성

다. 비식별 위험결정 요인

UKAN은 비식별 데이터의 위험은 특정 위험 결정요인에 의해 발생한다고 정의하고 다음과 같이 6가지 비식별 위험결정 요인을 제시하였다.

- 비식별 처리된 데이터에 포함된 개인을 재식별하려는 공격동기
- 개인정보 노출(Disclosure)에 따른 결과(공격 동기와 연관이 있음)
- 악의 없는 우발적 재식별(Spontaneous Identification)이슈
- 위험에 영향을 주는 정부의 데이터 접근관리 정책, 데이터 보안 및 다른 인프라들
- 해당 데이터와 연관 가능성이 있는 다른 데이터나 지식
- 해당 데이터와 다른 데이터나 지식의 차이(Data Divergence)

라. 데이터 상황

「통계적 비식별 처리유형」에 「위험 결정요인」들을 추가적으로 감안하여 비식별 처리를 한 것이 바로 「기능적 비식별 처리」이다. 이렇듯 기능적 비식별 처리는 통계적 측면뿐만 아

6 통계적 노출 제어(Statistical Disclosure Control, SDC)라는 기술 분야와 관계가 있음

니라 기본적인 위험을 판단할 수 있는 모든 정보는 데이터 환경이라고 하는 일종의 맥락(Context) 안에 포함되어 있다고 생각하며, 데이터와 관련된 모든 환경을 종합적으로 고려하여 재식별 위험을 평가한다. UKAN의 ADF[7]가 바로 해당 개념을 반영한 프레임워크이다.

UKAN에서는 「데이터 상황」을 데이터와 데이터를 둘러싼 환경과의 관계로 정의하고 데이터 환경을 고려하지 않고는 데이터가 비식별화되었는지의 여부를 판단할 수 없다고 한다. 데이터 환경은 세부적으로 다른 데이터, 행위자(에이전트), 거버넌스(법, 제도적 규칙), 인프라의 4가지 구성요소[8]로 이루어지며, 다음과 같이 누가(Who), 무엇을(What), 어디서(Where), 어떻게(How)라는 질문을 이용하면 좀 더 구체적인 데이터 환경 제어 방법을 특정할 수 있을 것이다.[9]

- 누가(Who) 데이터에 접근할 수 있는가?
- 무슨(What) 분석이 이루어지거나 이루어지지 않는가?
- 어디(Where)에서 데이터 접근/분석이 이루어지는가?
- 어떻게(How) 데이터에 접근하는가?

2. 주요 비식별 처리 프레임워크 및 위험평가 방법

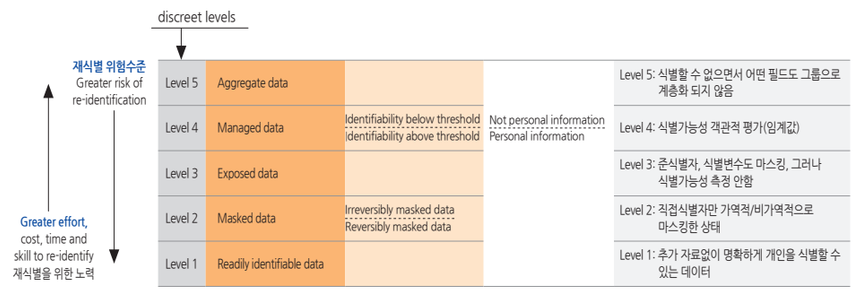
가. 미국 건강정보신뢰협회[10]의 리스크 기반 비식별화(Risk-Based De-Identification)

HITRUST에서는 개인정보의 식별성 판단을 비식별 데이터 처리수준이 특정 임계값(Threshold)을 넘으면 '개인 식별이 가능한(Identifiable) 데이터'라 하고, 넘지 않으면 '개인을 식별할 수 없는 데이터'라고 이분법적으로 판단을 하지만, 데이터의 식별성(Identifiability) 수준을 특정 단계로 명확히 나눌 수 없는 일종의 연속체[11]로 정의하고 데이터에 가해지는 비식별 조치 결과에 따라 재식별 위험 수준을 5단계로 구분하였다. 여기서

7 익명처리 의사결정 프레임워크(The Anonymisation Decision-making Framework, ADF)
8 내용참고: 김동현, 김순석, 「개인정보 비식별 조치를 위한 데이터 상황기반의 위험도 측정에 관한 새로운 방법」, Journal of The Korea Institute of Information Security & Cryptology, VOL 30, NO.4, 2020.8.
9 Duncan, G. T, Elliot, M. J and Salazar-Gonzalez, J. J, 「Statistical Confidentiality,」 New York: Springer, 2011
10 미국 건강정보신뢰협회(HITRUST): 2007년에 설립된 민간기구로서 공공 및 민간 부문의 개인정보보호, 정보 보안 및 위험 관리 리더와 협력하여 의료산업 및 공급망 전반에 걸쳐서 필요한 민감정보 보호에 대한 공통 위험 및 규정 준수 관리 프레임워크, 관련 평가 및 보증 방법론을 개발, 유지 관리하고 있다.
11 (Continuum)Emam, K.E., Risk-Based De-Identification of Health Data. IEEE Security and Privacy, 8, 64-67. 2010

의 척도는 단순히 서수이기 때문에 레벨 간의 개념적 거리를 의미하는 것은 아니다.(즉, 레벨1, 2 사이의 거리 ≠ 레벨4, 5 사이의 거리) 또한 레벨 간 위험 수준이 비율 척도를 나타내는 것도 아니기에 레벨1에서 레벨3로 이동 시 위험 수준이 레벨1에서 레벨2로 이동 시 위험 수준의 2배라는 의미로 해석해서는 안 된다.

[그림 5-3-1] Identifiability Continuum의 개념적 5단계 모델



* 출처: PharmaSUG, 「Practical Implications of Sharing Data: A Primer on Data Privacy, Anonymization, and De-Identification」, 2015.

여기서 중요한 것은 식별가능성 판단 기준이 되는 임계값(Threshold)을 설정하는 것인데, HITRUST는 임계값 설정 기초자료로 검사(Prosecutor)[12], 기자(Journalist)[13], 마케터(Marketer)[14] 공격자 모델 기반의 데이터 자체 위험도를 측정하고 다음과 같이 3가지 요소에 대한 Check List 기반 평가 결과를 바탕으로 비식별 수준을 결정하는 임계값을 구한다.

- 데이터 사용자 환경에 대한 데이터 보호를 위한 통제 상황(Mitigating Controls)
- 데이터 재식별 시 정보주체에 대한 프라이버시 침해 정도(Invasion of Privacy)
- 데이터 요청자의 재식별 동기와 능력(Motives and Capacity)

12 검사(Prosecutor) 리스크: 특정인이 정보 집합 안에 있음을 공격자가 알고 있는 경우
13 기자(Journalist) 리스크: 정보 집합 안에 재식별 할 수 있는 사람이 최소한 1명이 있는 경우로 누군가를 재식별 할 수 있음을 증명하여 비식별 처리를 수행한 기관을 곤경에 처하도록 하거나 신뢰를 떨어뜨리는 것이 목표인 경우가 대부분임
14 마케터(Marketer) 리스크: 정보 집합 안의 신원의 비율을 정확하게 재식별 할 수 있는 경우

나. UKAN의 익명화 의사결정 프레임워크(ADF 개정판[15])

UKAN은 데이터가 익명화되었는지 여부를 데이터 속성뿐만 아니라 데이터와 그 맥락 간의 관계를 통해 결정하는 기능적 익명화(Functional Anonymisation) 방식의 프레임워크인 익명화 의사결정 프레임워크를 다음의 4가지 원칙을 기반으로 제정하였다.

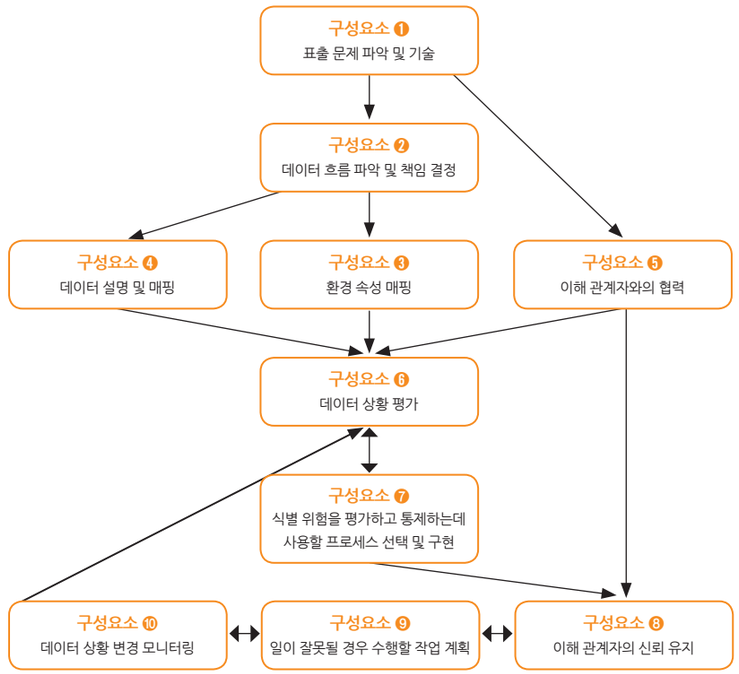
- 포괄성 원칙(Comprehensiveness Principle): 데이터 자체만으로는 공유나 공개의 안정성 여부를 판단할 수 없지만 그래도 데이터를 살펴보아야 한다.
- 유용성 원칙(Utility Principle): 익명화는 안전한 데이터를 생성하기 위한 프로세스이지만 생성된 데이터가 목적에 맞는 유용한 데이터이어야만 의미가 있다.
- 현실적 위험 원칙(Realistic Risk Principle): 유용하면서도 리스크가 없는 데이터를 생산하는 것은 현실적으로 불가능하다.
- 비례원칙(Proportionality Principle): 위험관리를 위해 적용하는 통제는 위험의 정도와 해당 위험으로 야기될 수 있는 영향에 비례하여 적용해야 한다.

이 프레임워크는 기술적(Technical)요소와 맥락적(Contextual)요소의, 두 가지 요소가 하나로 통합된 프레임워크로, 데이터뿐만 아니라 데이터의 환경까지를 모두 검토하여 현실적인 위험을 판단하는 재식별 문제에 대한 새로운 사고의 접근방식인 데이터 상황 접근법을 가장 처음 도입한 프레임워크이다.

- 기술적(Technical) 요소: 재식별 리스크의 정량화와 관리 방법 파악
- 맥락적(Contextual) 요소: 위험에 영향을 미치는 요소를 파악하고 처리(데이터 흐름, 법적 및 윤리적 책임, 거버넌스 관행, 공유 또는 릴리스 후의 책임, 문제 발생 상황에 대한 대응계획)

15 ADF는 2016년 9월에 실무자들이 개인정보의 익명화를 체계적으로 수행할 수 있도록 지원하기 위해 최초로 발간되었으며, 2018년 EU 전체에 '일반개인정보보호법(General Data Protection Regulation, GDPR)'이 시행되면서 새로운 법적 요구사항을 반영한 ADF 개정작업이 진행되었고, 기존 버전 사용자 커뮤니티의 의견까지 반영하여 2020년 개정판(2nd Edition)을 발간하였다.

[그림 5-3-2] ADF Workflow



* 출처: Mark Elliot, Elaine Mackey and Kiero O'Hara, 「The Anonymysation Decision-Making Framwork: European Practitioners' Guide」, UKAN Publications 2020.

ADF는 총 3개 분야 10개의 구성요소로 구조화되어 있으며, 구성요소가 순서별로 목록화되어 있기에 일반적으로 구성요소1 에서 시작하여 차례대로 적용하지만 그렇다고 엄격하게 순차적으로 적용할 것을 요구하거나 확인하지 않기 때문에 필요할 경우 순서 없이 자유로운 적용이 가능하다. 예를 들어, 이해 관계자 참여(구성 요소 5)는 프로세스의 모든 지점에서 발생할 수 있으며, 데이터 상황 평가(구성 요소6)의 경우도 데이터 상황을 새롭게 제안하거나 수정이 필요할 때는 여러 번 반복할 수 있다.

[표 5-3-3] ADF 구성요소 요약표

분류		구성요소	내용
데이터 상황 감시	1	표출 문제 파악 및 기술	데이터 상황을 이해하기 위한 단계로 익명처리 목표를 명확히 하고 위험을 식별하며 기술적 처리를 하기 위한 준비 단계
	2	데이터 흐름 파악 및 책임 결정	
	3	환경 속성 매핑	
	4	데이터 설명 및 매핑	
	5	이해 관계자와의 협력	
	6	데이터 상황 평가	
식별 위험 평가 및 통제	7	식별 위험을 평가하고 통제하는 데 사용할 프로세스 선택 및 구현	데이터의 식별 위험을 평가하고 적절한 통제를 구현하는 단계
영향도 관리	8	이해 관계자의 신뢰 유지	의도하지 않은 식별이나 법위반 등에 대비하여 대응계획을 수립하고 위험수준을 적절히 유지하기 위한 상황 모니터링 단계
	9	일이 잘못될 경우 수립할 작업 계획	
	10	데이터 상황 변경 모니터링	

* 출처: Mark Elliot, Elaine Mackey and Kiero O'Hara, 「The Anonymysation Decision-Making Framwork: European Practitioners' Guide」, UKAN Publications 2020., https://msrbcel.files.wordpress.com/2020/11/adf-2nd-edition-1.pdf (접속일: 2021.11.2.)

다. 우리나라 데이터 상황 기반의 가명·익명 프레임워크[16]

1) 프레임워크 제정 배경

우리나라의 데이터 상황 기반 비식별 프레임워크는 2019년 한국인터넷진흥원에서 위탁하고 한라대학교 김순석 교수팀에서 수행한 「개인정보 비식별 조치 프레임워크 및 정보 활용절차 개선 연구」를 통해 처음으로 제안되었다. 이 연구 보고서는 기존에 다른 나라에서 제안한 데이터 상황 기반 접근법이 우리나라의 법과 제도적인 요소에 맞추어 적용하기에는 적절치 않고, 의료 등 특정 분야에 치우쳐 있어 범용적으로 적용하기 어렵기에 새로운 연구를 통해 최근 개정된 우리나라의 데이터3법을 따르면서 우리나라 법과 환경에 맞게 체계적인 분류가 가능하고 각 구성요소들의 위험도를 정량화하여 측정할 수 있는 새로운 방안을 제시하였다.

16 김동현, 김순석, 「개인정보 비식별 조치를 위한 데이터 상황기반의 위험도 측정에 관한 새로운 방법」, Journal of The Korea Institute of Information Security & Cryptology, VOL 30, NO.4, 2020.8.

[표 5-3-4] 주요 비식별 프레임워크의 장·단점

프레임워크 및 가이드	장점	단점
영국 UKAN 익명처리 의사결정 프레임워크	단순하고 명료함	실제 데이터를 둘러싼 모든 상황을 고려하기에 부족
미국 NIST SP800-188(2nd Draft) 정부 데이터셋 비식별화 가이드	데이터뿐만 아니라 데이터 주변 여러 다른 측면을 고려토록 함	고려사항일뿐 구체적이지 못함
HIPAA 프라이버시 규칙 [17]	비식별 조치 전 데이터 중요도, 잠재적 피해, 공개 타당성 검토	측정 자체가 정성적이고 의료분야에 한정, 모든 환경요소 미고려
HITRUST & Privacy Analytics사 상황기반 위험도 측정 방법론	데이터 상황에 대한 구체적인 접근법 제시	우리나라의 법제도, 상황에 부적합하며, 의료분야에 한정됨
대한민국 비식별조치 가이드라인(2016)	재식별 시도 가능성과 재식별 시 영향분석 등 일부 내용 포함	영향분석이 비식별 처리 이후 실시 데이터 상황 전체 구성요소 중 일부만 고려됨

* 출처: 김동현, 김순석, 「개인정보 비식별 조치를 위한 데이터 상황기반의 위험도 측정에 관한 새로운 방법」, Journal of The Korea Institute of Information Security & Cryptology, VOL 30, NO.4, 2020.8., 저자 재구성

2) 국내 가명·익명 프레임워크

실제 국내에 적용되고 있는 프레임워크는 개인정보보호위원회와 한국인터넷진흥원이 진행하는 2021년 가명정보 전문가 양성교육 과정에서 확인할 수 있다. 연구보고서의 내용을 좀 더 구체적이고 현실적으로 변형하여 아래와 같이 총 3개 분야 6가지의 관점으로 데이터 상황을 분류하였다.

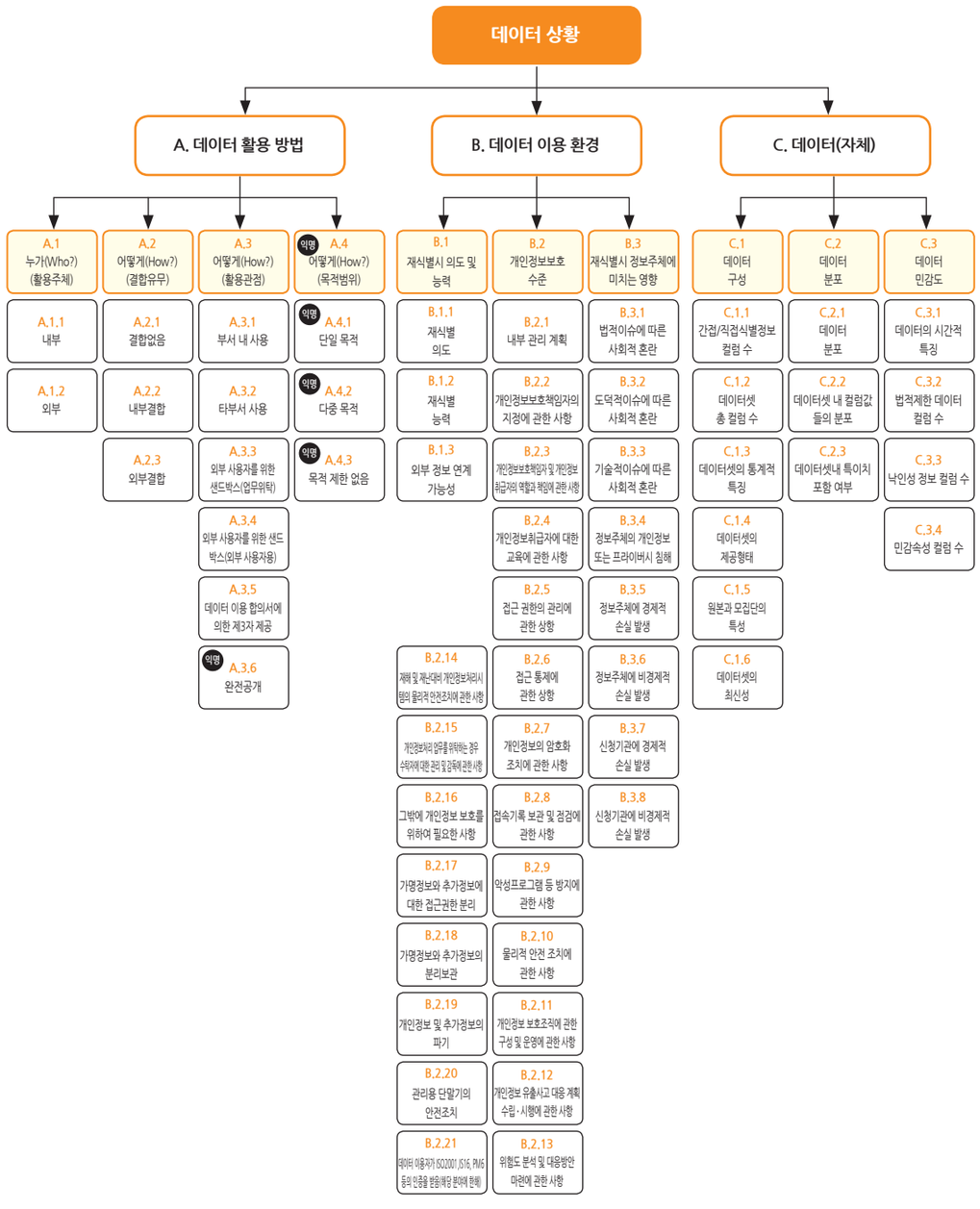
(가) 데이터 활용 방법(A)

데이터 활용 방법에서는 가명 처리 시와 익명 처리 시로 나누어 앞서 언급한 데이터 상황 제어 요소들을 적용하였다. 먼저 가명 처리 시에는 누가(Who, 활용 주체), 어떻게(How, 결합 유무), 어떻게(How, 활용 관점) 등 3개 분야로 나누고 세부적으로 10가지의 상황을 점검한다.

- A.1 누가(Who)(활용 주체) 분야에서는 내부 사용인지 외부 사용인지를 판단한다.
- A.2 어떻게(How)(결합 유무) 분야에서는 다른 데이터와 결합을 하는지, 만약에 한다면 자사가 보유한 다른 데이터와의 결합을 위해 내부 결합을 하는지, 아니면 타사 보유

17 HIPAA(Health Insurance Portability and Accountability Act) Privacy Rule, 2012.12.

[그림 5-3-3] 국내 데이터 상황 분류도[18]



18 개인정보보호위원회, 한국인터넷진흥원, 2021년 가명정보 전문가 양성교육(심화과정) 교재

데이터와의 결합을 위해 외부 결합을 하는지 판단한다.

- A.3 어떻게(How)(활용 관점) 분야에서는 조직 내부 사용 시에 데이터 관리 부서에서 사용하는지 아니면 타 부서에서 사용하는지, 사용자가 외부일 경우 업무위탁 형태의 사용인지, 형태 제공인지, 제3자 제공인지 등을 꼼꼼히 살핀다.

익명처리 시에는 추가적으로 어떻게(How, 목적 범위)를 추가하여 총 4개 분야 14가지의 상황에 대해 점검을 통해 상황을 평가한다.

- A.3 어떻게(How)(활용 관점)에서 익명 처리된 데이터가 완전 공개 데이터인지 여부를 추가로 판단한다.
- A.4 어떻게(How)(목적 범위)에서는 활용 목적이 단일한 목적인지 다중의 목적인지, 아니면 목적에 제한이 없는지 등을 확인한다.

(나) 데이터 이용환경 위험도 분석(B)

데이터 이용환경에 대한 위험도 분석은 총 3개 분야 32가지 상황에 대해 세부적으로 평가를 해서 위험도를 산출해야 한다.

- B.1 비식별 의도 및 능력 분석 단계에서는 의도나 능력뿐만 아니라 가지고 있거나 확보 가능한 외부정보와의 연계 가능성까지 점검해야 한다.
- B.2 개인정보보호 수준 분석 단계에서는 내부관리계획부터 기술적, 관리적, 물리적 보호조치가 적절히 이루어지고 있는지에 대해 21가지 상황에 걸쳐 점검한다.
- B.3 재식별시 정보주체에 미치는 영향 분석 단계에서는 재식별시 발생할 수 있는 법적, 도덕적, 기술적 이슈에 따른 사회적 혼란 가능성이나 정보주체나 신청기관에 발생할 수 있는 침해 정도, 경제적·비경제적 손실까지도 꼼꼼히 점검해야 한다.

(다) 데이터 자체 위험도 분석(C)

마지막으로 데이터 자체 위험도 분석 단계에서는 총 3개 분야 13가지 데이터 자체에 대한 정보를 바탕으로 데이터가 가지고 있는 위험을 평가해서 위험도를 산출한다.

- C.1 데이터 구성에서는 간접/직접 식별정보의 컬럼 수나 데이터 셋의 총 컬럼 수, 통계적 특징, 제공형태, 모집단의 특성, 데이터의 최신성 등 주로 데이터 구성과 관련한 정보를 검토한다.
- C.2 데이터 분포 분야에서는 데이터의 분포나 컬럼값들의 분포, 특이치 포함 여부 등을 평가한다.
- C.3 데이터 민감도 분야에서는 시계열 정보 여부를 확인하기 위한 시간적 특징, 법적 제한 정보나 낙인성 정보, 민감속성 정보의 컬럼수 등을 파악한다.

3) 위험도 평가 결과의 활용

이렇게 측정된 데이터 상황 및 데이터 자체에 대한 위험도는 '가명처리 단계별 세부절차[19]'의 가명처리 수준 정의의 기초자료로 활용되고, 각 속성 및 컬럼별 처리 수준을 정의하고 관련된 처리 기술을 적용하여 최종적인 가명처리를 하게 된다. 이렇게 가명처리된 데이터는 바로 사용되는 것이 아니라 내외부 전문가들로 구성된 적정성 평가위원회를 통해 평가를 받고 적정하다고 판단되면 목적에 따라 활용이 되며, 적정하지 못할 경우는 다시 가명처리 단계를 반복적으로 수행하게 된다.

19 개인정보보호위원회, 「가명정보 처리 가이드라인」, 2020

제4장. 안정적 AI/ML 서비스를 위한 MLOps

홍운표
(데이터로봇 코리아 상무)

한 조직 내 기계학습(Machine Learning, 이하 ML)의 적용 및 확산 시 반드시 운영 단계에 대한 준비 및 고려가 필요하다. 데이터에 기반한 의사결정 또는 Action item 발굴을 염두에 둔 케이스라면 더욱 안정적이고 연속적인 운영을 위해 컴퓨팅 리소스 관리 외에도 데이터 수급, 변이 대응, 이상치 대응 전략이 필요하다. 이와 관련하여 최근 AI/ML 결과물의 운영 적용을 위해 MLOps(Machine Learning Operations)가 크게 주목받고 있다. 이 장에서는 이러한 MLOps와 주요 기능 및 가용 소프트웨어를 소개한다.

1. MLOps 등장 배경

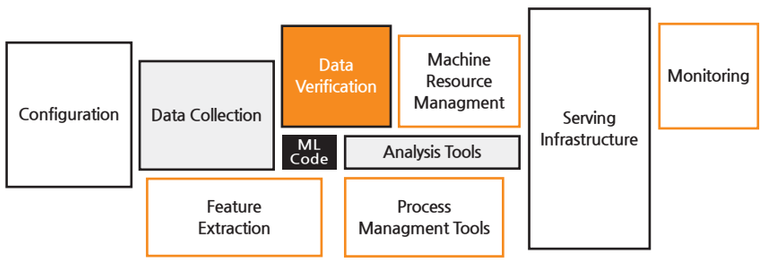
많은 기업들이 데이터 사이언스 팀을 만들고 툴에 투자한 이후에도 겪게 되는 어려움이 있는데, 이는 ML 프로젝트 마지막에 운영을 위한 작업을 진행할 때이다. 왜냐하면 ML은 배포(Deploy)하기가 쉽지 않기 때문이다. 모든 데이터 사이언티스트는 각자 선호하는 툴, 언어와 기법이 있다. 또 각 팀은 각자 다른 문제를 다른 플랫폼으로 해결하기 마련이다. 더욱 어렵게 만드는 것은 배포(Deploy) 이후 IT팀이 데이터 사이언스에서 사용하는 R, Python과는 다른 Java, C언어에 익숙하기 때문에 자신들이 이해하지 못하는 언어로 만들어진 모델을 배포(Deploy)하기 꺼려 하기 마련이다.

IT팀은 ML모델을 하나의 소프트웨어 구성요소로 간주하기 쉬우나, 데이터 사이언티스트 관점에서 보면 일반적인 소프트웨어와는 다르다. ML모델들은 학습데이터에 특정 알고리즘과 파라미터가 튜닝된 결과물이어서 학습데이터가 바뀌면 (실제 운영환경에서 자주 발생하는) 모델의 성능도 저하되고 재작업이 필수적이기 때문이다. 또한, 재작업된 모델은 앞서 운영을 위해서 추가된 모니터링, 보안, 거버넌스 작업도 다시 수행되어야 하는 것이다. 이러한

이유로 기존 DevOps 개념처럼, ML 모델의 지속적인 학습과 운영을 지원하기 위한 MLOps 개념이 나오게 되었고, MLOps는 초기 개발된 모델의 배포 및 운영과 개선방법을 제공하는 기능을 통칭한다. MLOps 가 지원해야 하는 기능들은 다음과 같다. 첫째, 만들어진 모델이 예상대로 동작하는지 어떻게 모니터링할 것인가? 둘째, 모델 성능 저하 시 업무 프로세스 중단 없이 어떻게 모델을 교체할 것인가? 셋째, 어떻게 운영환경 접근에 관한 통제를 구현할 것인가? 넷째, 운영 지속 시 발생할 수 있는 모델 교체 등의 이벤트 발생 시에도 규제 사항을 충족하면서 운영환경을 유지할 것인가? 이러한 이유로 데이터 사이언스 팀은 IT 팀이 요구하는 보안이나 거버넌스 정책을 회피하고 스스로 만든 플랫폼을 통해 배포(Deploy)하는 경우가 적지 않게 된다.

ML 프로젝트를 진행하는 경우, 정확도 높은 모델을 얻기 위한 코딩(R/Python)이 대부분의 작업이라고 생각하기 쉽다. 이는 전통적인 데이터 분석 방법론의 대부분이 데이터 탐색 및 모델링에 맞추어져 있기 때문이기도 하고, 대부분의 ML 프로젝트가 실제 운영까지 가기보다는 데이터로부터 인사이트를 얻는 데 그치는 경우가 많기 때문일 것이다. 그러나, 다루고 있는 주제가 비즈니스에 실질적인 영향이 있는 것이라면 인사이트를 얻고 끝내는 대신, 실제 운영환경으로 배포(Deploy)를 해야 하고, [그림 5-4-1]과 같이 실제 운영으로 넘기기 위해서 해야 하는 전체 작업 중에 코딩 작업은 극히 일부분에 불과하며 코딩 작업과 전혀 다른 성격의 일들이 존재하게 된다.[1]

[그림5-4-1] ML코드를 포함한 실제 ML시스템 및 관련 인프라



* 출처: Sculley, D. a.-F., 「Hidden Technical Debt in Machine Learning Systems」, Curran Associates, Inc., 2015.

1 Sculley, D. a.-F. (2015). Hidden Technical Debt in Machine Learning Systems. Advances in Neural Information Processing Systems. Curran Associates, Inc.

2. MLOps 구현 전략

전통적인 ML 개발 방법론을 준수하여 개발된 모델은 문제정의-데이터준비-탐색분석(EDA)-학습 및 튜닝-배포/운영의 생애주기를 갖게 된다. 각 단계별로 모델의 신뢰성 확보/유지의 위험 요소가 존재하나 MLOps 관점에서 주요하게 검토하고 대비해야 하는 위험은 배포/운영 단계에 맞추어져 있다.

MLOps를 구현할 때 두 가지 접근을 할 수 있는데, 성능 관점의 모델 확보를 위해 배포된 모델의 관리에 중점을 두고 모델의 지속적 개발(재학습, 재개발)은 기존의 ML 개발 방법론을 따르도록 하거나 (이하 모델관리 중점 MLOps), 주기적인 재학습에 초점을 두고 자동 워크플로우 구현에 중점을 두는 방식을 취할 수도 있다.(이하 재학습 중점 MLOps)

특히 모델관리 중점 MLOps에서는 배포/운영 단계에서 발생할 수 있는 위험을 서비스 관점과 성능 관점으로 나누어 볼 수 있다. 서비스 관점에서는 안정적 예측 값 서비스를 위한 응답시간, 캐쉬율 등 일반적인 SW/HW 관리 측면이 중요하다. 성능 관점 또는 데이터 사이언스 관점에서 보다 중요한 위험은 운영 중인 모델의 성능이 학습단계에서 보였던 수준으로 얼마나 유지될 수 있을 것인지에 있으며, 이를 위협하는 다양한 요소를 사전에 방지하고 관리하기 위해 요구되는 기능들과 추가로 재학습 중점 MLOps를 위해 고려해야 할 기능들을 아래에서 기술한다.

3. 모델 관리 중점 MLOps

가. 드리프트(Drift)

전형적인 ML개발방법론에 따르면, 학습된 모델의 운영/배포 이후 학습데이터의 패턴에 맞추어진 ML모델은 서비스 시간이 지날수록 성능이 저하될 위험이 있다. 이러한 위험의 원인은 크게 세 가지로 간추려지는데, 1) 타깃 정의(Concept) 드리프트, 2) 데이터 정의 드리프트, 3) 데이터 드리프트가 존재한다[2]. 타깃 정의 드리프트는 비즈니스 관점에서 예측하고자 하는 타깃의 변화 때문에 발생하는 것이다. 예를 들어, 3개월 내 경쟁사로 이동할 확률이 높은 고객 예측을 타깃으로 배포하였으나, 시장 환경 변화로 3개월이 아닌 1개월 내 이동이

2 Thomas, J., Databricks Blog. Retrieved from https://databricks.com/blog/2019/09/18/productionizing-machine-learning-from-deployment-to-drift-detection.html (접속일: 2021.11.2.)

예상되는 고객을 예측하는 것이 필요해진 경우이다. 데이터 정의 드리프트는 데이터 내 컬럼의 포맷이 변화되는 경우를 말하는데, 예를 들어 무게를 나타내는 컬럼이 킬로그램(kg) 단위에서 파운드(lb)로 바뀐 경우, 해당 모델을 변경하지 않으면 엉뚱한 결과를 초래할 수 있다. 이 두 가지 요인도 무시할 수는 없으나, 데이터 드리프트에 비하면 해결책이 단순하기 때문에 이 장에서는 첫 번째 형태의 드리프트에 대해서만 고찰한다.

대부분의 ML 알고리즘들은 데이터의 통계적인 특성 또는 패턴에 집중하여 정확하고 강건한 예측 모델을 얻어내기 때문에 학습데이터와 운영 단계에서 예측 요청에 들어오는 데이터의 통계적인 특성들이 달라진다면 성능 저하를 피할 수 없게 된다. 여기서 통계적인 특성이라 함은 확률분포(Probability density/mass function)로 완벽하게 표현 가능하고, 이를 운영단계에서 추적하고 모니터링한다면 데이터의 드리프트를 완벽하게 검출하고 대비할 수 있을 것이다. 그러나, 확률분포의 정밀한 함수 표현식을 기준으로 놓고 학습데이터와 예측데이터의 차이(Skew)를 검출한다는 것은 모수의 부족 (예를 들어 실시간 예측 서비스가 필요한 경우) 및 통계 잡음으로 인해 판단 착오로 이어지기 쉽다.

보다 근사적인 방법으로 확률분포의 수식적인 특성 차이를 보기보다는 평균과 분산처럼 모멘트값을 모니터링하는 방법이 있으나, 정규분포를 따르는 데이터가 아니라면 (예를 들어 감마 분포) 이차항으로 전체 확률분포의 차이(skew)를 모니터링하는 것에는 한계가 있다. 이에 대한 대안으로 모집단의 구간을 일정하게 나누어 각 구간별 빈도수를 모니터링하는 방법이 있다. 아래의 수식 PSI(Population Stability Index) 개념은 통계적인 안정도를 판별하기 위해 도입된 개념이며 구현이 쉽고 직관적이다.

$$PSI = \sum_{bin} (Scoring\% - Training\%) \ln(Scoring\% / Training\%)$$

즉, 하나의 피처를 구간(Bin)으로 나누어, 각 구간별로 전체 데이터 중의 비중을 구한 후, 운영 시 예측 데이터로 들어온 피처의 구간별 비중의 차이를 구해 전체로 합산한 값이다. 예를 들어, [표 5-4-1] PSI 계산 예시와 같이[3] PSI 금융권 사례 중 신용도 구간이라는(Score Range) 피처에 대해 학습데이터의 구간별 분포(Training%)를 구하고, 예측 데이터의 분포(Scoring%)를 구해서 각 구간별 PSI를 구하고 이를 합산하면 0.013이라는 값이 나온다.

3 https://www.listendata.com/2015/05/population-stability-index.html (접속일: 2021.11.2.)

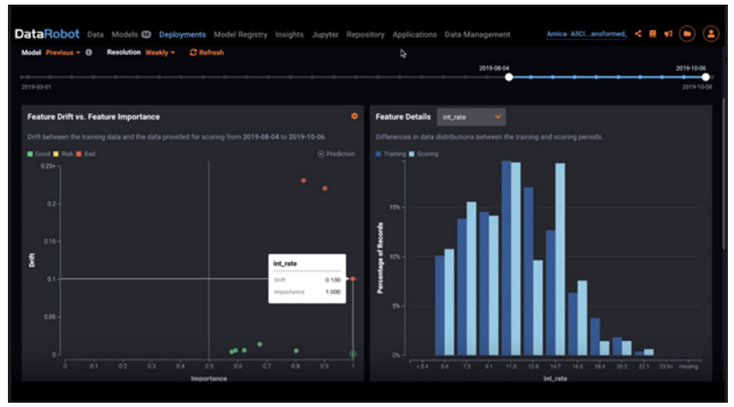
[표 5-4-1] PSI 계산 예시

Score Range	분위수(Decile)	Scoring%(A)	Training%(B)	A - B	ln(A/B)	PSI
> 720	1	12%	11%	1%	0.09	0.001
671-720	2	11%	11%	0%	0.00	0.000
641-670	3	14%	12%	2%	0.15	0.003
611-640	4	12%	13%	-1%	-0.08	0.001
581-610	5	12%	11%	1%	0.09	0.001
551-580	6	10%	11%	-1%	-0.10	0.001
521-550	7	12%	13%	-1%	-0.08	0.001
491-520	8	6%	5%	1%	0.18	0.002
451-490	9	6%	7%	-1%	-0.15	0.002
< 451	10	5%	6%	-1%	-0.18	0.002
						0.013

* 출처: LISTEN DATA, 「POPULATION STABILITY INDEX AND CHARACTERISTIC ANALYSIS」, 2015.5.

일반적으로는 PSI 값이 0.1 이하이면 별도의 작업이 필요없고, 0.1에서 0.25 사이 값이라면 작은 변동으로 인식하여 지속적 모니터링이 필요하고, 0.25를 넘는 경우 심각한 차이(Skew)가 발생했다고 보고 모델의 신뢰성을 확보하기 위한 후속 작업에 대한 검토를 거쳐 재학습 또는 모델 교체 작업을 수행하게 된다[4]. 그러나, PSI 값 자체는 사용된 구간의 수 [표 5-4-1]의 경우 10개 구간, 샘플 수에 종속적이기 때문에 0.25라는 기준은 절대적일 수는 없다[5].

[그림 5-4-2] 드리프트 모니터링 예시



* 출처: https://www.datarobot.com/blog/all-data-drift-is-not-created-equal-dot-dot-dot/ (접속일: 2021.11.2.)

4 Upadhyay, R. (n.d.). YOU CANalytics. Retrieved from http://ucanalytics.com/blogs/population-stability-index-psi-banking-case-study/ (접속일: 2021.11.2.)

5 Naranjo, B. Y., 「Statistical properties of the population stability index」, Journal of Risk Model Validation, Vol.14, No.4, p89-100, 2020

오히려 당면 ML 사례에 대한 데이터 사이언티스트와 현업 지식을 가진 담당자가 협업하여 각 피처별로 기준을 정하고 이에 대한 모니터링을 통해 드리프트 발생 시 대응 전략을 수립하는 것이 적절하다. DataRobot사의 사례 [그림 5-4-2]를 보면, 모델 개발 시 확인된 주요 변수 중 이자율(int_rate)에 대해 PSI 기준값을 0.1로 놓고, 특정 기간으로 필터링하여 모니터링이 가능하도록 구현되어 있는 것을 볼 수 있다.

나. 정확도 모니터링과 및 챌린저 모델

배포/운영 전환 이전에 모델 개발 단계에서 모델의 평가는 여러 가지 적용 가능한 평가 기준(metric)이 사용될 수 있다. 이때 사용된 평가 기준은 배포/운영 단계에서도 동일하게 적용되어 개발단계에서의 평가치와 실제 운영 단계에서의 평가치가 비교 관리되어야 한다. 관리되어야 하는 평가 기준은 운영 중인 모델의 타입에 따라 달라질 수 있다. 즉, 분류 모델의 경우는 Logloss, AUC, 민감도/특이도 등이 적용될 수 있고, 회귀 모델의 경우는 RMSE, MAPE, R^2 등이 적용될 수 있다.

운영 시 배포된 모델은 예측 데이터에 대해 예측치를 계산하는데, 실제 값이 확보되기 전까지 실제 정확도는 측정할 수 없으므로, 예측 건에 대한 실제 데이터가 확보되는 시점에 계산할 수밖에 없다. 따라서, 모델의 실제 성능 저하는 시간차가 발생하게 되므로, 앞서 언급된 드리프트를 참고하여 먼저 성능 저하 여부에 대해 판단하고, 정확도를 기반으로 한 정보는 참고자료로 하여 최종적인 모델 교체에 대한 의사결정을 내릴 수밖에 없다.

예측치에 대한 실제치가 확보되는 시점에는 기존에 발행된 예측치와 연관 지을 수 있는 체계가 미리 구비되어 있어야 하며 (예를 들어 조인 키(Join Key)), 해당 모델에 적용 가능한 평가 기준에 대해 계산 가능한 기능도 구현되어 있어야 한다.

정확도 관점에서 보다 기민한 대응이 필요한 경우라면 모델 개발 시 여러 가지 모델을 동시에 배포하여 성능 저하에 대비할 수 있다. 즉, 가장 성능이 좋았던 모델을 실제 예측치로 서비스하되, 실제 예측 서비스는 하지 않더라도 동일 문제에서 확보된 다른 모델의 예측치를 시스템 내부에 저장하였다가 실제치가 확보되는 시점에 계산을 통해 후보 모델들의 정확도를 산출하여 비교해 보는 것이다. 많은 벤더사들이 이러한 개념을 챌린저 모델로 부르며 다수의 챌린저 모델을 동시에 배포하여 정확도 추이를 모니터링하는 데 활용하고 있다.

[그림 5-4-3] 챌린저 모델의 정확도 모니터링 예시



* 출처: https://www.datarobot.com/blog/how-mlops-helps-you-dodge-the-wrecking-ball-of-underspecification/ (접속일: 2021.11.2.)

다. 이상치 대응

운영 시 발생할 수 있는 또 다른 문제 중 하나는 배포된 모델로부터 산출된 예측치가 해당 모델의 비즈니스 측면에서 받아들일 수 없는 값을 내는 경우이다. 예를 들어 방문자 수를 예측하는 모델이 음의 값을 낸다거나 정상 동작범위를 벗어나는 센서 값을 예측해야 되는 배포의 경우, 또는 센서의 정상 동작 범위를 벗어나는 예측 값이 예측된 경우 예기치 못한 문제를 야기할 수 있다.

이러한 원인은 예측을 위해 입력된 종속변수들이 잠재적인 가정을 벗어나는 값이거나, 운영 중인 모델 자체가 개발 단계에서 충분한 고려를 못했을 경우 발생할 수 있다. 이에 대한 대비책으로 예측치 서비스를 전달하면서 기록을 남기거나, 또는 예측 값을 제공하지 않고 에러값만을 반환하여 비즈니스에 파급효과를 차단하는 방법으로 대응할 수도 있다.

4. 재학습 중점 MLOps

가. 모델 리프레시

위에서 살펴본 데이터 드리프트, 정확도 모니터링 및 이상치에 대한 모니터링을 통해 운영 중인 모델의 신뢰성이 저하된다면 운영팀뿐만 아니라 데이터 사이언티스트 및 현업 담당자들 간의 협의를 통해 대처 방안을 수립해야 한다. 내릴 수 있는 결정은 모델을 교체(Replace)하거나, 모델 및 해당 비즈니스 케이스를 폐기(Retire)하거나 재학습(Retrain)을

통해 새로운 모델을 개발하는 세 가지이다. 각각의 결정에 도달하는 조건은 조직마다 다를 수 있으나 일반적으로는 아래 [표 5-4-2]와 같은 조건을 포함하여 정의할 수 있다.

[표 5-4-2] 모델 학습방안 조건

	교체(Replace)	폐기(Retire)	재학습(Retrain)
드리프트	설정된 기준 내에서 미미한 수준의 변화를 보인다.	콘셉트 드리프트 발생으로 기존 모델의 문제 정의와 달라진 문제를 해결해야 한다.	기준선 이상의 데이터 드리프트가 발생한다.
정확도	학습 시 보다 저하되었으나 챌린저 모델의 성능이 현재 배포된 모델의 성능을 능가하여 상당기간 유지되고 있다.	고려 불필요	정확도는 학습 시 보다 하락하였고, 챌린저 모델의 성능도 하락 추세를 보인다.
이상치	기존 모델로 처리 가능한 이상치가 발생하며, 주요하게 고려해야 할 이상치는 발생하지 않는다.	고려 불필요	기존 모델이 처리할 수 없는 이상치의 잦은 발생으로 이상치 처리가 포함된 신규 모델이 필요하다.

나. 피처 스토어(feature store)

운영 중인 모델에 특정 시점까지 드리프트나 정확도의 추세 변화가 없다 하더라도 주기적으로 재학습을 진행해야 되는 ML 모델들이 있다. 예를 들어 24시간 가동이 필요한 제조 설비의 경우 주기적인 정비(Periodic Maintenance, PM)가 필수적인데, 제조라인에서 수집된 데이터로 학습된 모델의 경우 PM 전후 드리프트 발생이 필연적이기 때문에 주기적인 재학습이 요구되는 것이다. 이런 경우에 학습에 필요한 데이터를 처음부터 다시 수집하고, 정제하고 추가적인 피처 엔지니어링 작업을 거치는 것은 모델 개발의 속도를 저하하는 주요 요인이 된다. 이를 방지하기 위해 도입된 개념이 피처 스토어이며, 재학습이 필요한 시점에 데이터엔지니어링 작업을 최소화할 수 있도록 이전 모델에 사용된 피처들을 저장해 두는 것이다.

그러나 이전에 사용된 모델이 향후 꾸준히 사용되어야 하는 경우라면 피처 엔지니어링이 포함된 피처 스토어가 의미가 있을 수 있지만 알고리즘을 바꾸는 게 성능상 의미가 있는 경우라면 새로운 피처 엔지니어링이 필요할 수 있다. 그래서, 피처 스토어를 좀 더 넓은 의미로 정의하여 데이터의 저장 및 관리를 주 목적으로 한정하는 경우도 있고, 자주 사용되는 파생변수를 포함하는 경우도 있다[6][7][8][9].

5. 공통 기능

가. 모델 거버넌스 (승인, 버전관리)

위에서 살펴본 모델의 Replace/Retire/Retrain은 관련된 이해 당사자들의 협의와 노력을 거쳐, 하나의 배포라고 하더라도 다양한 인스턴스가 발생하게 된다. 모델의 버전 관리, 승인 프로세스, 배포 권한관리 기능이 필요한 이유이다.

나. 모델 메타데이터 관리

운영되는 모델의 안정적 관리를 위해서는 메타데이터를 정의하고 이의 관리를 통해 해당 모델의 라이프사이클 동안 유지·관리되어야 한다. ML 모델 자체에 대한 정보들(알고리즘, 하이퍼파라미터, 모델 사이즈 등)과, 피처 스토어와 연관된 정보들(데이터셋 아이디, 피처리스트 등)이 운영되는 동안 수집되고 관리되어야 한다.

다. 서비스 모니터링 및 관리

운영되는 모델의 성능 외에도 예측 요청에 대한 응답시간, 시간당 요청 건수, 배치 및 실시간 요청 통계 정보 등이 수집되고 관리되어야 한다. 이는 예측 서비스의 안정적 제공 외에 모델 Replace/Retire/Retrain 의사결정에도 중요한 판단 자료로 쓰일 수 있기 때문이다.

라. 컨테이너 기반 컴퓨팅 자원

서버 인프라에 대한 고려는 각 기업 또는 조직이 처한 환경에 맞게 설계되어야 하는데 최근의 추세만을 언급하자면, 가상 머신 기반으로 배포하는 것보다는 도커(Docker) 기반 컨테

6 Khalid Salama, J. K. (n.d.). Google. Retrieved from mlops-whitepaper: https://cloud.google.com/resources/mlops-whitepaper (접속일: 2021.11.2.)

7 Thomas, J. (2019, September 18). Databricks Blog. Retrieved from https://databricks.com/blog/2019/09/18/productionizing-machine-learning-from-deployment-to-drift-detection.html (접속일: 2021.11.2.)

8 AWS. (2021). Amazon SageMaker Feature Store. Retrieved from https://aws.amazon.com/sagemaker/feature-store/ (접속일: 2021.11.2.)

9 Google. (2021). Google Cloud. Retrieved from MLOps https://cloud.google.com/architecture/mlops-continuous-delivery-and-automation-pipelines-in-machine-learning (접속일: 2021.11.2.)

이너에 Flask 같은 작은 크기의 웹서버를 통해, 각 배포별로 독립된 서비스를 제공하는 것이 일반적인 추세이다. 최근에는 도커 기반 컨테이너 환경의 문제점을 해소하고자 관리 도구인 쿠버네티스(Kubernetes)를 채택하여 구축하는 경향도 최근 들어 나타나고 있다[10]. 아울러 쿠버네티스 환경에서 ML 파이프라인 수행을 돕는 오픈소스인 Kuberflow를 이용해 ML의 전반적인 라이프사이클을 지원하고자 하는 경향도 존재한다.

6. MLOps 상용 소프트웨어

가. SAS

모델 관리에 중점을 둔 제품으로 MLOps 기능을 제공하며, 재학습 관점에서는 전통적인 ML 개발 방법론으로 해결책을 제시한다. 별도의 패키지로 Model Manager를 판매하며 모델 관리 관점에서 위에 언급된 기능 외에도 해당 모델의 내재적인 특성을 모니터링하고 관리할 수 있는 기능을 풍부하게 제공한다[11].

나. Google

다양한 오픈소스를 활용한 MLOps 개념을 활발히 적용하고 발표해 오고 있으며, 모델 관리 관점에 대한 기능들보다 재학습 중점에 보다 많은 기능들을 제시하고 있다. 주로 ML Pipeline 자동화와 지속적 개발(Continuous Development) 지속적인 통합에(Continuous Integration) 주안점을 두고 코드 기반의 MLOps 구현을 제시하고 있다[12].

다. AWS

AWS Sagemaker 환경에서 만들어진 모델을 SDK를 활용하여 코딩을 통해 API 엔드포인트(Endpoint)로 구현하거나 배치 서비스로 구현 가능하다. 최근에 MLOps 프레임워크를 발표했으나 모델 관리 관점에서 위에 제시한 기능들에 대한 구현은 비교적 초기 상태이다[13].

라. Microsoft Azure

MS Azure Machine Learning 환경에서 개발된 모델을 웹서비스 형태로 배포할 수도 있고, SDK기반으로 배포하고 관리하는 다양한 기능을 구현할 수 있다. 다양한 타깃(컨테이너, Kubernetes, 웹서비스 등)에 배포하고 MLOps 개념을 구현할 수 있으나 드리프트 검출 기능 등은 아직 진행 중이다[14].

마. AutoML 개발사

최근 들어 데이터 사이언티스트 부족에 따른 분석 자동화를 위해 AutoML 개발사들이 대두되었다. DataRobot, H2O.ai 및 Dataiku 등이 해당하는데 분석 자동화를 모델 개발에 한정하지 않고, 모델 관리 및 재학습까지 범위를 넓히기 위해서 MLOps 기능은 자연스럽게 지원되는 추세이다. 그러나 대부분은 모델 관리에 중점을 둔 기능을 제공한다. 다만, DataRobot은 최근 Continuous AI를 선언하며 모델 관리 외에도 피처 스토어 및 재학습 워크플로우 기능을 포함하여 제공하고 있다.

7. 마무리

이 장에서는 최근 AI/ML 도메인에서 관심도가 증가하는 MLOps에 대해 기술하였다. 전통적인 모델 관리 관점에 초점을 맞춘 기능 요소와 더불어 재학습에 중점을 둔 MLOps에 요구되는 기능들도 포함하여 고찰하였다. 기존에는 기업들이 AI/ML이 모델 개발에 집중하여 신뢰성이 있는 모델을 확보하는데 노력을 기울였다면 최근에는 개발된 모델로부터 지속적인 비즈니스 효과를 거두기 위해 모델을 관리하고 재학습을 통해 지속적으로 신뢰도 높은 모델을 얻기 위해 MLOps 기능을 고려할 필요가 있다.

10 Singh, P. (2021). Deploy machine learning models to production: with flask, streamlit, docker, and kubernetes on google cloud platform. Apress.
11 SAS. (n.d.). Model Manager. Retrieved from https://www.sas.com/en_us/software/model-manager.html (접속일: 2021.11.2.)
12 Google. (2021). Google Cloud. Retrieved from MLOps: https://cloud.google.com/architecture/mlops-continuous-delivery-and-automation-pipelines-in-machine-learning (접속일: 2021.11.2.)
13 AWS. (2021). AWS MLOps Framework. Retrieved from https://aws.amazon.com/solutions/implementations/aws-mlops-framework/ (접속일: 2021.11.2.)
14 Microsoft. (2021). Azure. Retrieved from MLOps: https://docs.microsoft.com/en-us/azure/machine-learning/concept-model-management-and-deployment (접속일: 2021.11.2.)

2021 데이터산업 백서 집필진

제1부. 데이터 민주주의의 시대

1장. 데이터 권리의 민주화(국문 번역본) Data Sovereignty	**마이데이터 글로벌 리더십 팀 공저** **대표저자: 비비 라티노야**
2장. 데이터 활용의 민주화	**김찬수** 투이컨설팅 상무 **이승봉** 투이컨설팅 이사
3장. 데이터 경쟁의 민주화	**유영상** 한국전자통신연구원 책임연구원

제2부. 데이터산업 관련 주요 정책·법제도

1장. 국내 데이터산업 관련 정책·법제도	**박소영** 국회입법조사처 입법조사관
2장. 해외 데이터산업 관련 정책	**이준호** 이씨이십일 R&C 연구원
해외 데이터산업 관련 법제도	**이성엽** 고려대 기술경영전문대학원 교수
3장. 디지털 무역과 국경 간 데이터 이동	**이성엽** 고려대 기술경영전문대학원 교수

제3부. 데이터산업 시장 현황

1장. 국내 데이터산업 시장 현황	**한국데이터산업진흥원 산업기획팀**
2장. 해외 데이터산업 시장 현황	**나영민** 날리지리서치그룹 이사

제4부. 산업별 데이터 활용 현황

1장. 금융분야 데이터 활용	**정구태** 카르도 전략부문 이사
2장. 헬스케어분야 데이터 활용	**한현욱** 차의과학대학교 의학전문대학원 정보의학교실 교수
3장. 제조분야 데이터 활용	**김창원** 건국대학교 신산업융합학과 겸임교수
4장. 유통·물류분야 데이터 활용	**송성렬** CJ대한통운 TES물류기술연구소 부장
5장. 광고분야 데이터 활용	**최홍규** 한국교육방송공사 연구위원
6장. 모빌리티분야 데이터 활용	**차두원** 차두원모빌리티연구소 소장 **금효영** 슈퍼브 에이아이 이사
7장. 에듀테크분야 데이터 활용	**유인식** 유비온 전무이사
8장. 농업분야 데이터 활용 · 비즈니스	**최대근** 파미너스 대표
9장. 데이터 동맹·제휴 현황	**김인현** 투이컨설팅 대표

제5부. 데이터산업 기술 동향

1장. AutoML 동향 그리고 대중화	**조치선** 바이브컴퍼니 그룹장
2장. Cloud Computing	**고태우** KDB산업은행 산업기술리서치센터 차장
3장. 개인정보 비식별 처리 수준 및 프레임워크	**윤덕상** (주)파수 전략사업본부장
4장. 안정적 AI/ML 서비스를 위한 MLOps	**홍운표** 데이터로봇 코리아 상무

2021 Data Industry White Paper

2021 데이터산업 백서 [통권24호]

편찬위원 김인현 투이컨설팅 대표
김종현 쿠콘 대표
이성엽 고려대 기술경영전문대학원 교수
함유근 건국대 경영대학 교수

편집위원 한국데이터산업진흥원 데이터산업본부
하진희 본부장
안선빈 산업기획팀 주임
허은희 산업기획팀 인턴

발행처 한국데이터산업진흥원
서울특별시 중구 세종대로9길 42 부영빌딩 7,8,11층(04513)
Tel. 02-3708-5300
www.kdata.or.kr

발행인 윤혜정

발행일 2021년 11월 12일

제작 넥스텔리전스(주) Tel. 02-783-3801

디자인·편집 (주)엔더블유기획정보 Tel. 02-2235-1411

ISSN 2465-7662

• 본 백서의 내용은 한국데이터산업진흥원의 공식 견해와 다를 수 있습니다.
• 본 백서 내용의 무단 전재를 금하며, 가공 인용 시에는 반드시 한국데이터산업진흥원, 「2021 데이터산업 백서」라고 밝혀 주시기 바랍니다.

• 본 백서는 과학기술정보통신부의 DB산업 육성 사업의 결과물입니다.

• 「2021 데이터산업 백서」와 관련한 문의는 한국데이터산업진흥원으로 연락해 주시기 바랍니다.

2021 Data Industry White Paper

2021 데이터산업 백서

9 772465 766005 75
ISSN 2465-7662 [비매품]

한국데이터산업진흥원
Korea Data Agency
서울특별시 중구 세종대로9길 42 부영빌딩 7, 8, 11층
Tel. 02.3708.5300 Fax. 02.318.5040
www.kdata.or.kr