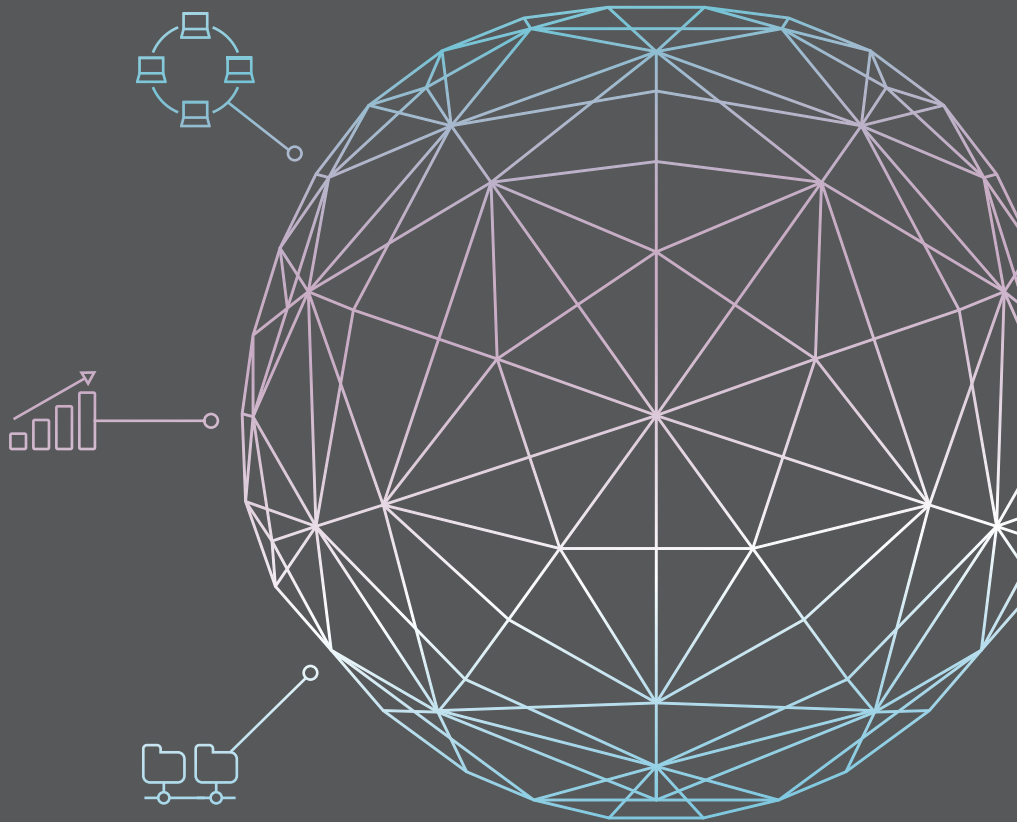
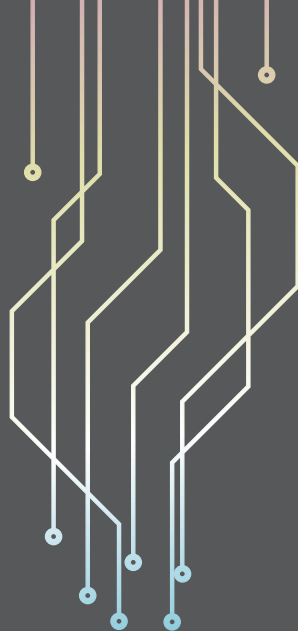


2023

데이터산업 백서

2023 DATA INDUSTRY WHITE PAPER



초거대 AI 시대의 경쟁력, 데이터 가치와 활용에서부터

안녕하십니까. 한국데이터산업진흥원 원장 윤혜정입니다.

2010년대 초 ‘데이터’가 새로운 화두로 주목받기 시작하던 시대가 엊그제 같은데, 불과 10여 년이 지난 2023년 현재 ‘데이터’는 우리 사회와 국가, 산업의 중심으로 자리 잡았습니다. 특히 지난해 광풍을 일으킨 생성형 AI 기술과 서비스의 급격한 발전은 데이터의 중요성뿐만 아니라 나아가 인간의 삶과 문화 그리고 사고에까지 큰 영향을 미칠 수 있는 데이터의 파급력을 다시금 확인시켜주는 계기가 되었습니다.

본격화되는 초거대 AI 시대를 맞이하며 데이터를 둘러싼 이슈도 그 어느 때보다 훨씬 더 복잡하고 다양화될 것으로 예상됩니다. 주지하는 바와 같이 데이터는 과거의 행동과 결정에서 산출되는 후행적 산물이면서 이를 토대로 의사결정과 전략수립 등의 불확실한 미래를 대비하는 합리적 선택에 이바지하는 선행적 도구로 기능하고 있습니다. 따라서 데이터를 ‘가진 자’와 ‘가지지 못한 자’의 시장 구도에서 이제는 ‘가진 국가’와 ‘가지지 못한 국가’의 국가 차원의 경쟁 구도로의 변화도 일어나고 있습니다. EU의 디지털시장법, GDPR로 촉발된 데이터 주권, 각국의 데이터 국지화 논의가 그 예라고 할 수 있습니다. 이를 대비하기 위해 우리도 범국가 차원에서 데이터의 권리, 개인 데이터 활용 등에 대한 다양한 논의와 명확한 정의를 통해 데이터가 공식적인 무형의 자원으로 사회에서 기능할 수 있는 제도적 방안을 강구해야 하며, 데이터 가치·품질 지원, 정부·대기업·의료 및 연구기관의 데이터 공유·접근 확산, 글로벌 협력 및 진출, 국내 중소·스타트업 등의 시장 참여 촉진과 같은 데이터 생태계 활성화를 위한 지속적 정책 지원 방안을 심도있게 고민해야 하는 시점입니다.

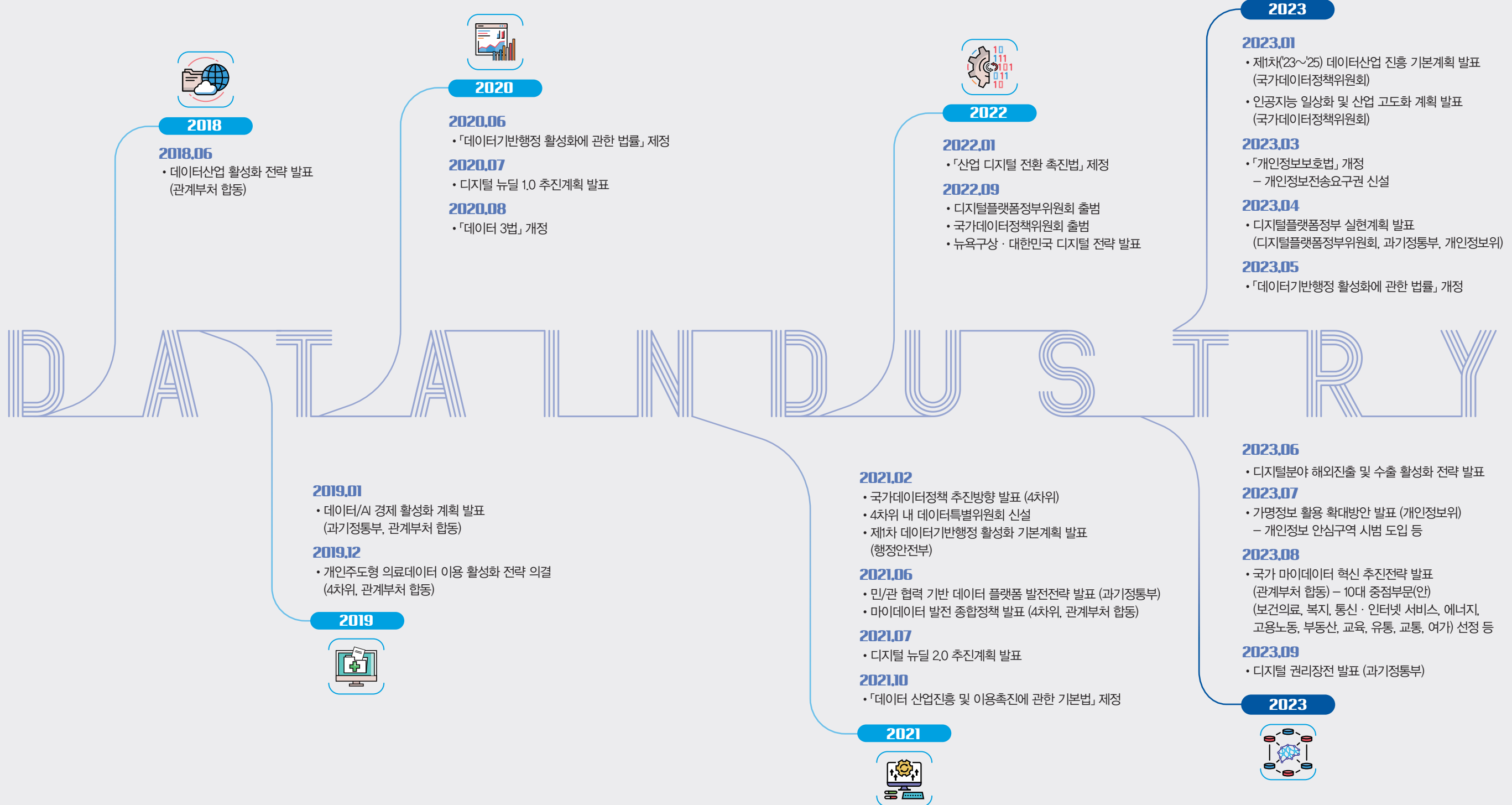


이러한 다양한 논의의 장을 마련하기 위해 올해 발간하는 「2023 데이터산업 백서」는 초거대 AI 시대의 데이터 가치와 활용을 큰 주제로 다루며, 1부에서 초거대 AI 시대의 데이터 활용의 중요성, 공유의 필요성, 활용 이슈에 대해 논의하였습니다. 2부는 국내외 데이터산업 관련 주요 정책 및 법·제도와 데이터 이슈에 대해 정리하였으며 3부에서는 국내외 데이터산업 시장 현황과 분석결과를 다루고 있습니다. 이어서 4부에서는 주요 산업별 데이터 활용 현황을, 5부에서는 ‘합성데이터’ 등 데이터 활용과 관련한 기술에 대해 기술하고 있습니다. 작년에 이어 근 6년 동안의 데이터산업 관련 정책 및 법제도 흐름과 데이터산업 시장 현황을 인포그래픽으로 정리하였으며, 데이터 분석을 바탕으로 국내 데이터산업 이슈를 선정하였습니다.

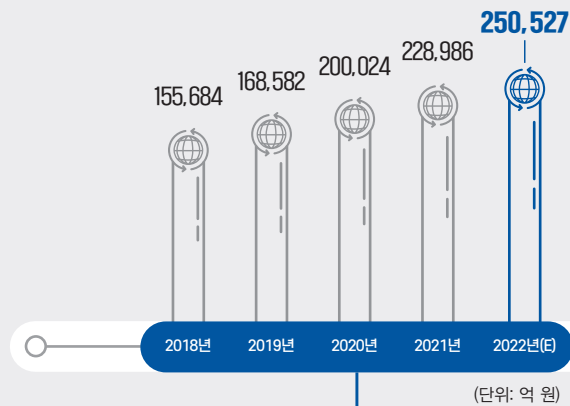
이번 백서가 한국 데이터산업과 정책의 미래를 논의하는 데 있어 유용한 마중물이 되기를 바랍니다. 마지막으로, 본 데이터산업 백서 발간을 위하여 힘써주신 집필진, 자문위원, 편집진 모두에게 깊은 감사의 인사를 드립니다.

한국데이터산업진흥원 원장
윤 혜 정

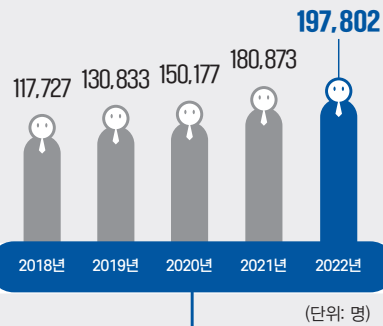
한 눈에 보는 데이터산업 정책 및 법제도



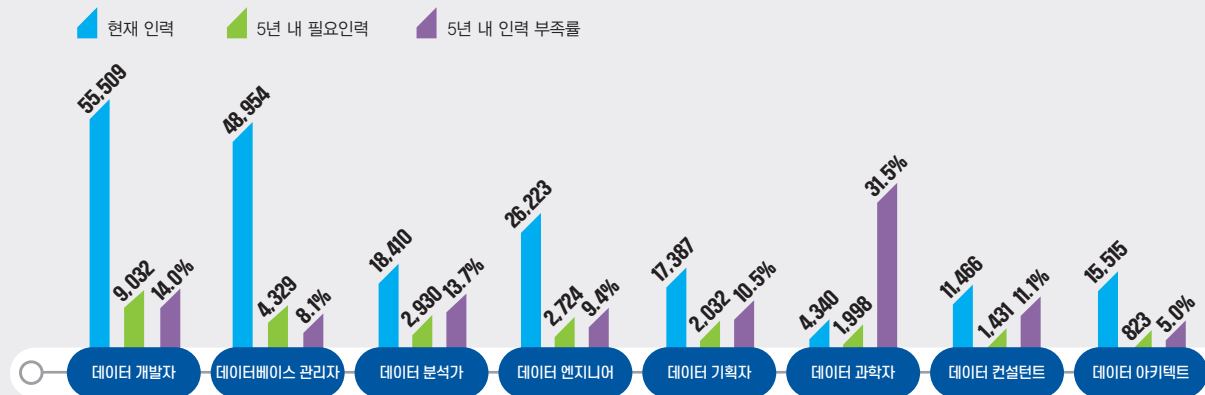
한 눈에 보는 데이터산업 시장 현황



데이터산업 시장규모



전 산업 데이터직무 인력 현황



데이터직무별 인력 현황 및 수요

2022년 기준 전 산업 데이터직무

- 인력은* **197,802** 규모
- 전년 대비 **9.4%** 증가
- 향후 5년 내 필요인력 **25,298명**
- 추가 인력** 필요 예측
- 인력부족률 **11.3%** 예측

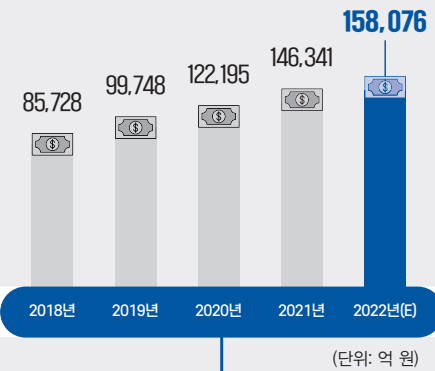
* 데이터산업 및 일반산업의 데이터직무 인력수
** 기업에서 현재 인력보다 추가로 더 필요로 하는 인력 수

직무별 부족률

데이터 과학자
31.5%

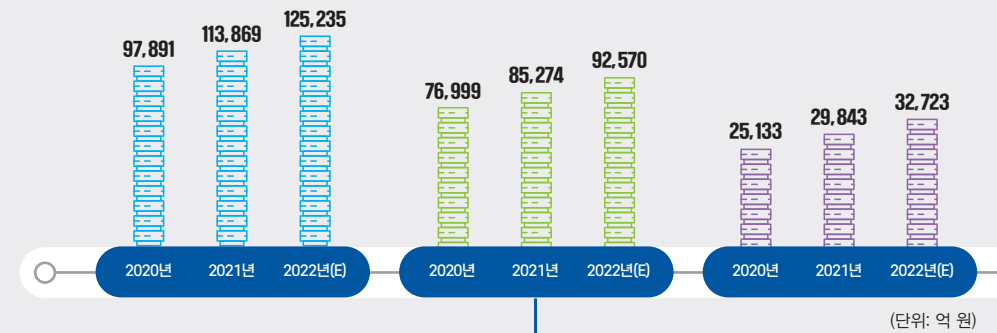
데이터 개발자
14.0%

데이터 분석가
13.7%



데이터산업 직접매출 규모

데이터 판매 및 제공 서비스업 데이터 구축 및 컨설팅 서비스업 데이터 처리 및 관리 솔루션 개발 · 공급업



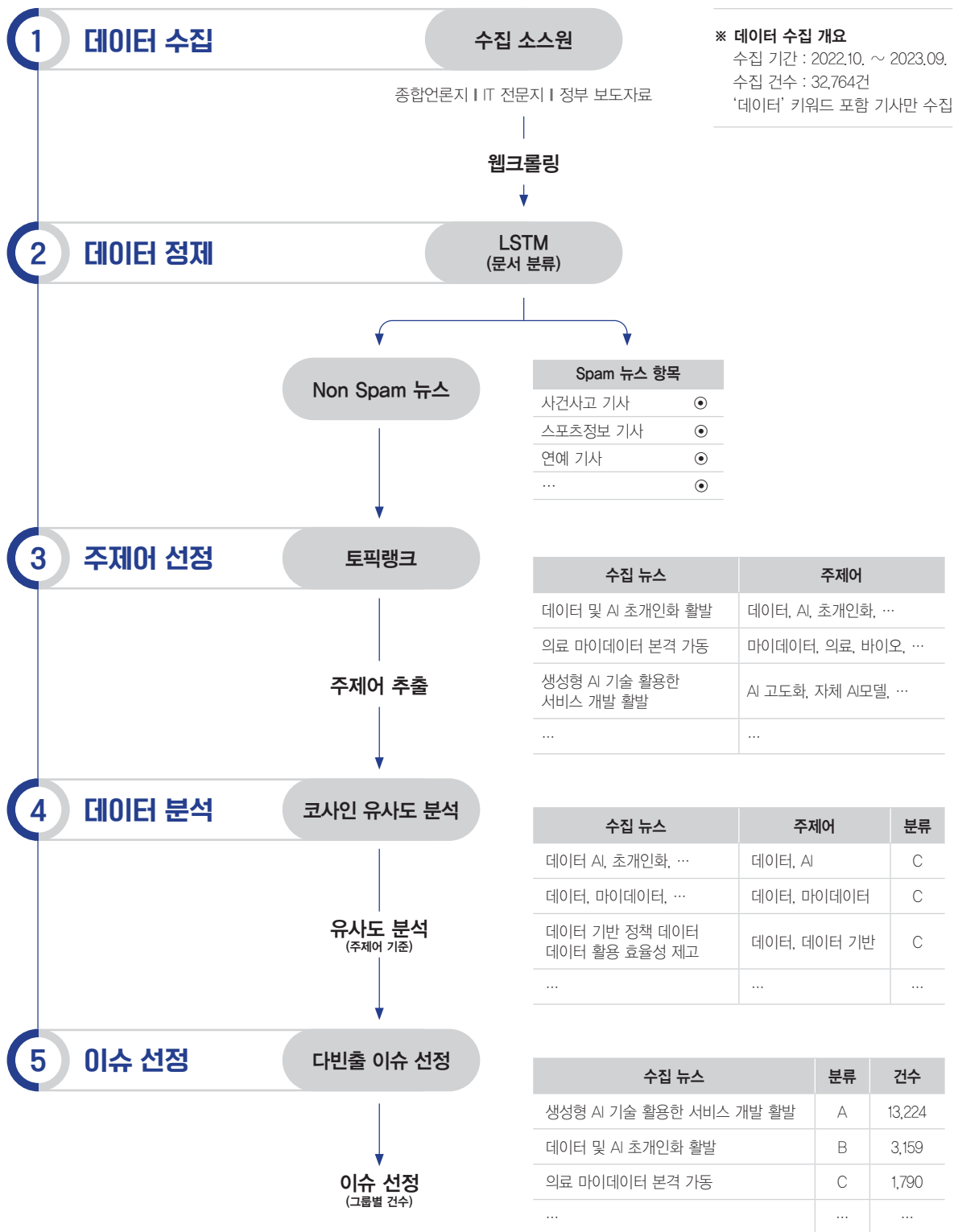
데이터산업 부문별 시장규모

출처: 과학기술정보통신부, 한국데이터산업진흥원, 「2022년 데이터산업현황조사」, 2023.4.

부록

2023 국내 데이터산업 이슈 TOP 6

데이터 분석 방법론



※ 데이터 수집 개요
수집 기간 : 2022.10. ~ 2023.09.
수집 건수 : 32,764건
'데이터' 키워드 포함 기사만 수집

01 데이터 학습 기반의 생성형 AI 개발 활발

주요 기업, 생성형 AI 기술 활용한 서비스 개발에 박차

데이터 저장·처리 기술의 큰 발전으로, 다양한 형태, 방대한 양의 데이터를 학습해 텍스트나 이미지, 음성 등의 새로운 데이터를 생성하는 '생성형(Generative) AI' 기술이 활발히 활용되고 있음. 특히나 약 7,000억 개의 데이터 토큰을 사전 학습했다고 알려진 OpenAI의 ChatGPT 등장이후로 생성형 AI시장의 경쟁이 본격화되고 있음. 국내에서도 많은 기업들이 생성형 AI를 활용한 새로운 서비스 개발에 박차를 가하고 있으며 이를 위한 고품질의 데이터를 확보하기 위한 경쟁도 치열해지고 있음.

[표 1] 주요기업별 생성형AI 기술

기업	출시 기술명	관련 내용
네이버	하이퍼클로바	<ul style="list-style-type: none"> 한국어 특화 언어모델(ChatGPT보다 한국어를 6,500배 더 많이 학습했다고 알려짐) 자체 생성형AI로 사용자(기업)가 자체 데이터 세트를 활용하는 맞춤형 서비스 제공
LG	엑사원2.0	<ul style="list-style-type: none"> 한국어·영어를 이해하고 답변할 수 있는 이중 언어 모델·텍스트·이미지 등으로 정보를 주고 받는 멀티모달이가능·특허·논문 등 약 4,500만 건의 전문 문헌과 3억 5,000만 장의 이미지를 사전 학습함
KT	민음	<ul style="list-style-type: none"> 대규모 데이터를 사전학습하여 다양한 Task별 AI 서비스를 생성할 수 있는 파운데이션 모델로, 기존 AI 모델의 한계점인 대량의 학습 데이터 수집과 학습 시간 단축
SKT	에이닷	<ul style="list-style-type: none"> 자체 개발 중인 거대언어모델(LLM)을 고도화해 에이닷이단답형 대화가 아닌 이용자의 맥락을 이해하고 복잡한 의도 파악
삼성SDS	브리티코파일럿/패브릭스	<ul style="list-style-type: none"> 생성형AI를 활용해 다양한 작업을 자동화하는 자체 솔루션·기존 시스템과 연계 및 프라이빗클라우드환경 지원 가능해 보안성높음 (*'24년 출시)

* 출처: 국내 데이터 발행 이슈 분석 (2022.10 ~ 2023.09)

정부, 데이터 및 AI의 안정적 활용을 위한 전략 마련

생성형AI가 많은 비즈니스 기회를 제공하고 있으나 데이터 및 정보 유출, 딥페이크, 저작권 문제, 편향 및 부정확하게 생성된 콘텐츠 등 해결 과제가 존재함. 이에 정부는 데이터 접근 보장, 디지털 리터러시향상, 디지털 격차 해소 등의 내용을 포함하는 '디지털 권리장전'을 발표함.

[그림 1] 생성형 AI 및 데이터 활성화 지원 정책

'AI 데이터 융합 네트워크' 발족 2023. 9. 8.	대한민국 초거대 AI 도약 방안 발표 2023. 9. 13.	'전 국민 인공지능 일상화' 추진 2023. 9. 13.	'디지털 권리장전' 마련 2023. 9. 25.
초거대 인공지능을 전산업으로 확산	분야별 특화 자율점검표·안내개발서 발행 확산	분야별 특화 자율점검표·안내개발서 발행 확산	새로운 디지털 질서의 기본방향
AI응용 서비스 개발에 필요한 전문 분야 데이터 확보	위험 요인과 성능을 제3기관 통해 평가	위험 요인과 성능을 제3기관 통해 평가하는 신뢰성 검증 인증 체계 마련	디지털 심화에 대한 범정부 대응 현황 분석
대규모 맞춤형 300억 토큰 구축 계획	초거대 AI의 한계 극복을 위한 기술 개발	초거대 AI의 한계 극복을 위한 기술 개발	누구나 자유롭게 토론할 수 있는 '디지털 공론장' 구축
	반도체, 플랫폼, ICT 인프라와 결합해 시너지 창출		학계, 업계, 소비자단체 등이 참여하는 민·관 협의체 구성

* 출처: 과학기술정보통신부 보도자료 기반 재구성 (2023.01 ~ 2023.09)

02 초개인화의 엔진, 'DATA'

데이터 기반의 초개인화 서비스 확대 및 고도화

개별 소비자들에게 데이터와 인공지능(AI)을 기반으로 한 맞춤형 서비스는 일상이 되었으며, 기업들은 고객들에 보다 가까이 다가가기 위해 노력하고 있음. 이를 위해 축적된 소비자 데이터를 기반으로 '타겟팅' 전략과 '초개인화' 기술을 활용하고 있음. 그 결과 여러 분야에서 생활 밀착형 초개인화 상품과 서비스가 개발되고 있음.

[그림 2] 데이터 기반의 개인 맞춤형 서비스 사례



* 출처: 국내 데이터 발행 이슈 분석 (2022.10 ~ 2023.09)

03 보건의료부문 데이터 활용 본격 활성화

데이터를 안전하고 주도적으로 활용할 수 있게 하는 움직임이 많아지고 있음. 특히나 민감한 개인정보가 많은 보건의료 부문에서 안전한 데이터 활용을 바탕으로 국민들이 혜택을 누릴 수 있도록 여러 정책 노력이 이뤄지고 있음.

개인정보보호법 개정과 마이데이터로드맵

'23년 3월 '개인정보보호법'이 개정됨에 따라 의료 등 전 분야 내 마이데이터 서비스 확산의 근거가 되는 개인정보 전송요구권이 신설됨. 정부는 연내 개인정보보호법 하위법령을 마련하고 2025년부터 제도를 본격 시행할 예정임.

[그림 3] 마이데이터 로드맵

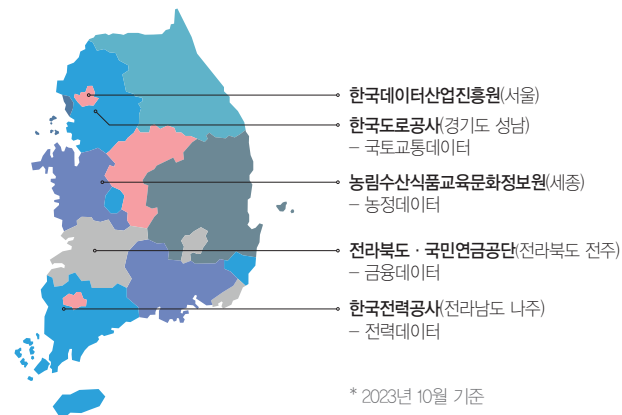


* 출처: 기획재정부, 과학기술정보통신부 외 관계부처 합동 보도자료 재구성

데이터안심구역 전국 확대

데이터를 유출 없이 안전하게 분석·활용할 수 있는 '데이터안심구역'이 확대되고 있음. 대표적으로 중부권 소재 연구기관 및 의료기관과 협력해 활용 사례를 발굴하고 있는 데이터안심구역 대전센터가 있으며, 그 외에도 보건의료데이터를 안전하게 활용할 수 있는 춘천 데이터안심구역을 포함해 데이터산업법에 따라 올해 전국 5개 데이터안심구역이 새롭게 지정되었음. 이렇듯 이용자들이 양질의 데이터에 접근할 수 있는 제공 거점과 협력체계가 확대될 것으로 예상됨.

[그림 4] 데이터안심구역 및 주요 보유 데이터



* 2023년 10월 기준

04 데이터 기반의 첨단 모빌리티 산업 성장

데이터 기반의 모빌리티 산업 확대

모빌리티산업 내에서 데이터를 기반으로 하는 디지털 전환이 빠르게 진행되고 있음. 특히 자동차 산업 전반이 소프트웨어 중심의 차량(SDV)으로 체제가 전환됨에 따라 데이터의 가치는 더욱 커지고 있음. 완성차 및 소프트웨어 기업들은 미래 모빌리티 부문을 주도하기 위해서 자율주행 시스템 설계, 서비스 제공 등을 위한 데이터 확보와 데이터를 수집 및 활용할 수 있는 역량을 갖추는데 총력을 기울이고 있음.

[표 2] 모빌리티 데이터 활용 사례

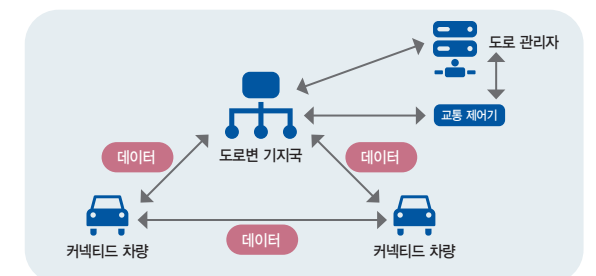
기업	관련 내용
현대자동차·기아	• 현대·기아의 커넥티드 카 이용자 확대로, 교통 신호 및 차량의 센서·운행 정보 등의 방대한 모빌리티 데이터가 확보됨. • 데이터와 AI 기술을 바탕으로 모빌리티 서비스 고도화 추진
한국타이어	• 내부에 장착된 복합 센서에서 수집한 데이터를 토대로 주행 중 타이어 마모 상태·압력·온도·주행 정보 및 노면 상태 진단 • 수집 데이터를 기반으로 안전한 주행과 효율적인 차량 관리를 구현하는 '아이 타이어' 출시
포티투닷	• 도심 교통OS 구현 위해 데이터를 기반으로 모빌리티플랫폼 시스템 구축·실제 도로주행 데이터를 확보 및 데이터 활용 기술을 고도화 본격 추진
카카오 모빌리티	• 수집 데이터를 기반으로 자체 운송관리시스템(TMS) 제공 • 데이터 분석 기반의 택시/주차/대리운전 등의 수요 예측, 사용자 편의성 제고
티맵 모빌리티	• 내비게이션·대중교통 데이터를 통합하여 개발한 슈퍼 앱인 '올 뉴 티맵'출시 • 연평균 44억 건에 이르는 22년간의 도로 안내 데이터를 수집, 기반으로 LLM 구축
쏘카	• 수집 데이터를 기반으로 수요 예측, 최적 가격을 책정하는 '다이나믹 프라이싱(Dynamic Pricing) 시스템' 보유 • 사내데이터 공유 시스템 '쏘카데이터 마켓'으로 데이터 산출 기준 및 집계 현황 공유
한국전자통신연구원 등	• 드론에 데이터·네트워크·인공지능(DNA)을 접목한 기술 개발 및 드론데이터 실시간 전송·AI 분석 가능한 DNA+드론 플랫폼 구축 • 플랫폼 기반 경진대회를 통한 서비스(삼육대 실시간 도로 결함 탐지, 경북대 안심귀가 서비스 등) 활성화 촉진

데이터 표준화로 2027년 완전자율주행 상용화

정부가 자율주행 차량 간(V2V), 그리고 차량과 인프라 간(V2I)을 오가는 V2X 데이터의 형식 표준화를 발표함. 데이터 형식이 표준화되면 타 제조사 차량 및 도로 인프라와 차량 위치·속도·브레이크·교통신호 상태 등 다양한 정보를 교환해 차량 단독의 자율주행보다 진일보된 협력형 자율주행을 실현할 수 있어 자율주행의 성능과 안전이 크게 향상될 것으로 기대됨.

* 출처: 한국정보통신기술협회 정보통신용어사전

[그림 5] 자율차 데이터의 차량·사물 통신(V2X)



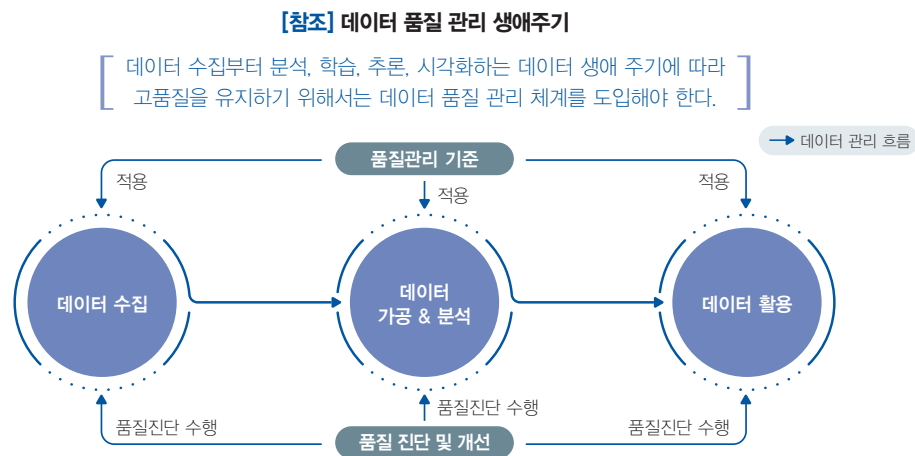
05 데이터 품질 관리 필요성 확대

AI 시대의 데이터 품질 관리 필요성

데이터의 품질이 기업에 매우 중요한 요소가 되고 있음. 정확하지 않은 데이터 등 저품질의 데이터를 이용하면 데이터 분석의 가치와 기업이 제공하는 서비스의 품질 역시 떨어지기 때문임. 특히 AI의 경우 학습 데이터에 따른 편향성이 존재하기에, 오류가 있는 데이터가 학습될 경우 윤리적·법적 문제가 발생할 수 있음. 그 때문에 신뢰할 수 있는 데이터 활용을 위해 민간 및 공공 분야에서 데이터 품질 및 활용관리 활동을 진행하고 있음.

데이터 품질 관리 생애주기

데이터 수집부터 분석, 학습, 추론, 시각화하는 데이터 생애주기에 따라 고품질을 유지하기 위해서는 데이터 품질 관리 체계를 도입해야 함.



* 출처: 2022년 데이터산업 백서 5부 3장 데이터 품질 관리 기술, 한국데이터산업진흥원(2022)

데이터 품질 향상을 위한 정부의 노력

양질의 데이터 유통을 활성화하기 위한 정부의 노력도 이어지고 있음. 대표적으로 과학기술정보통신부는 '데이터 산업진흥 및 이용촉진에 관한 기본법'에 따라 데이터 품질 인증기관을 지정하였으며 이를 계기로 데이터 품질에 대한 인식을 제고하고, 양질의 데이터 생산을 촉진하고 있음. 이를 통해 데이터의 유통·거래 시장이 더욱 활발해질 것으로 기대됨.

[참조] 데이터 품질인증 제도 시행

과학기술정보통신부는 데이터 품질인증기관으로 3개 기관 지정한다. '데이터산업법'에 따르면 품질인증기관은 데이터 내용의 완전성·유효성 및 정확성을 검토해야 하고, 데이터 구조의 일관성을 점검해야 한다. 또한, 데이터 관리체계의 유용성 및 접근성 등을 검토해야 한다. 이러한 품질기준에 따라 데이터 품질인증을 진행해야 한다.

데이터 품질 인증기관	(주)씨에이에스
	(주)와이즈스톤
	한국정보통신기술협회

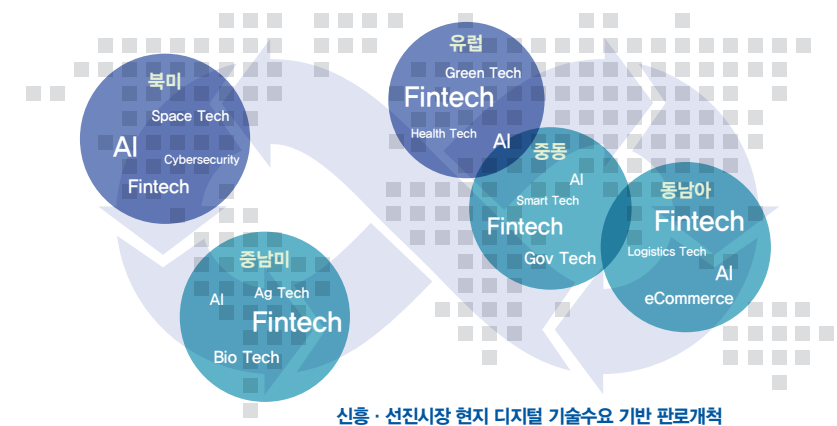
06 디지털 기업의 글로벌화

디지털 혁신기업 해외진출 지원

글로벌 시장을 주도하는 'K-디지털'을 실현하기 위해 대한민국 디지털 전략을 토대로 정부 지원이 확대되고 있음. 대표적으로 국내 디지털 기업이 글로벌 기업으로 도약할 수 있는 여건 조성하고자 민관협력 기반의 수출 지원체계가 구축되고 있음. 정부는 '디지털 수출개척단' 운영을 통해 아세안, 중동, 미주 등 각지에서 민간주도형 대·중소 해외진출 연대모델 발굴과 해외진출 확대를 지원하고자 함.

과기정통부는 디지털 혁신기업 해외 진출을 전문적으로 지원하기 위해 '글로벌 디지털 혁신 네트워크(GDIN)'를 출범함. GDIN은 '글로벌 디지털 로드' 구축을 목표로, 디지털전환 수요와 디지털 해외진출 기회가 높은 신흥시장 개척을 위해 기업을 지원하는 방안 등을 제시함. 또한 한-아세안 디지털 장관회의 및 정상회의와 같이 아세안 국가들과의 인공지능·데이터 관련 연대를 확산하기 위한 협력 체계도 마련되고 있음.

[그림 6] 글로벌 디지털 로드 모식도



* 출처: 디지털 혁신기업 해외진출 전문기관, 글로벌 디지털 로드를 개척한다, 과기정통부 보도자료, 2023.9.

데이터 기반 디지털 기업의 글로벌 시장 진출

고도화된 데이터·AI 기술을 기반으로 디지털 신기술·서비스 분야의 글로벌 시장에 진출하는 국내 기업이 확대되고 있음.

[표 3] 글로벌 ICT 미래 유니콘 선정 기업(2023)

기업	제공 서비스	진출/진출목표 국가
웨이센	AI 기반 소화기 내시경 영상 솔루션, 환자 데이터 분석 플랫폼	베트남
롤루랩	피부 데이터 기반 뷰티·헬스케어 솔루션	베트남
로보아르떼	실시간 모니터링 데이터 기반의 최적화 튀김 조리 로봇·시스템	미국, 싱가포르
뤼튼테크놀로지스	사용자 데이터 기반의 텍스트 생성 AI 솔루션	일본
프리월린	학습 데이터(수학문제) 기반의 에듀테크 솔루션	일본, 동남아시아
인터엑스	자율생산을 지원하는 AI 기반 제조 데이터 분석 서비스 플랫폼	독일

* 출처: 국내 데이터 발행 이슈 분석 (2022.10 ~ 2023.09)

2023 데이터산업 백서

2 0 2 3
D A T A
I N D U S T R Y
W H I T E
P A P E R



CONTENTS

발간사	- 02
한 눈에 보는 데이터산업 정책 및 법제도	- 04
한 눈에 보는 데이터산업 시장 현황	- 06
2023 국내 데이터산업 이슈 TOP 6	- 10

제1부 초거대 AI 시대의 데이터 가치와 활용

제1장 • 초거대 AI 시대 데이터의 가치와 활용의 중요성	- 20
제2장 • 초거대 AI 시대 데이터 공유	- 26
제3장 • 데이터 활용의 이슈 : 데이터 편향성/윤리성	- 34

제2부 데이터산업 주요 정책 및 법제도 현황

제1장 • 국내 데이터 관련 정책 및 법제도 현황	- 46
제2장 • 해외 데이터 관련 정책 및 법제도 현황	- 56
제3장 • 데이터 활용 이슈의 법제도적 측면 : 저작권/개인정보	- 66

제3부 데이터산업 시장 현황

제1장 • 국내 데이터산업 시장 현황	- 76
제2장 • 국내 데이터산업의 역동성 및 생산성 분석	- 89
제3장 • 해외 데이터산업 시장 현황	- 97

제4부 산업별 데이터 활용 현황

제1장 • 금융분야 데이터 활용 현황	- 110
제2장 • 헬스케어분야 데이터 활용 현황	- 118
제3장 • 모빌리티분야 데이터 활용 현황	- 126
제4장 • 제조분야 데이터 활용 현황	- 134
제5장 • 농업분야 데이터 활용 현황	- 144
제6장 • 에듀테크분야 데이터 활용 현황	- 153
제7장 • 新 데이터 비즈니스	- 163

제5부 데이터산업 기술 동향

제1장 • 합성 데이터 생성 기술 동향	- 174
제2장 • 클라우드 스토리지 기술 동향	- 184
제3장 • 데이터 분석 기술 동향	- 187
제4장 • 데이터 보안 기술 동향	- 196

2023 데이터산업 백서 집필진	- 204
-------------------	-------

1 PART

초거대 AI 시대의 데이터 가치와 활용

제1장 • 초거대 AI 시대 데이터의 가치와 활용의 중요성

제2장 • 초거대 AI 시대 데이터 공유

제3장 • 데이터 활용의 이슈 : 데이터 편향성/윤리성

제1장

초거대 AI 시대 데이터의 가치와 활용의 중요성

권호열 교수 강원대학교 컴퓨터공학과

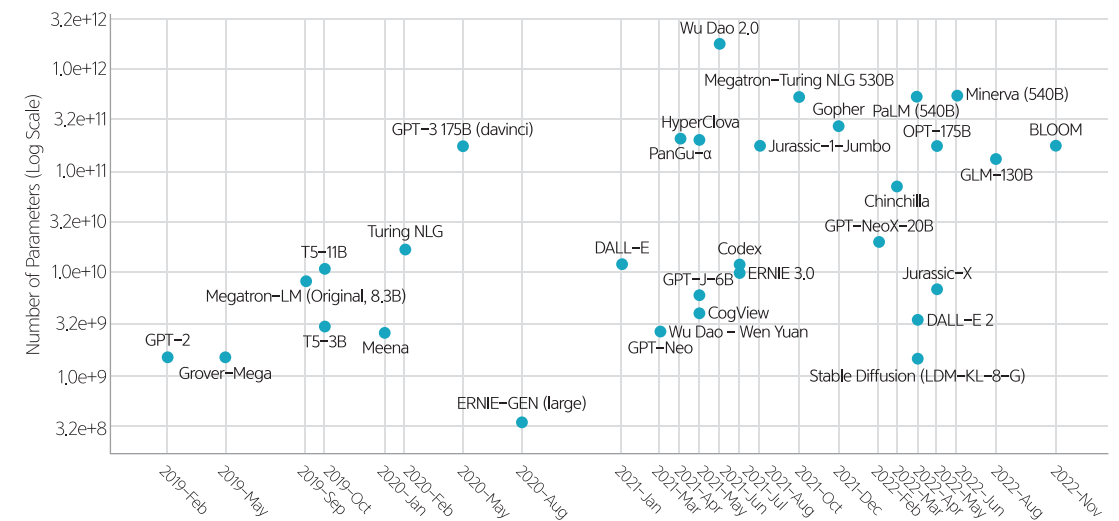
챗GPT의 출현과 함께 초거대 인공지능이 일상생활 속으로 들어왔다. 학습한 데이터를 활용한 인공지능은 문서 작성 및 요약으로부터 프로그램 코딩, 비즈니스 응용, 예술 작품 제작에 이르기까지 활용 범위가 매우 넓을 뿐 만 아니라 최적화된 자동화를 통하여 미래 국가 산업경쟁력의 핵심요소로 부상하고 있다. 이와 관련하여 인공지능과 디지털경제, 인공지능의 국가경쟁력을 살펴보고 데이터 품질과 가치, 데이터산업 인프라, 데이터와 인공지능의 활용을 제시하였다. 끝으로 데이터 기반 디지털 공동번영사회를 선도하기 위한 디지털권리장전을 소개하였다.

1. 초거대 AI와 디지털 경제

인공지능(AI)은 지각능력, 학습능력, 추론능력 등 인간의 지적 활동에 필요한 능력을 인공적으로 구현한 것으로서 인공지능 연구가 본격적으로 시작된 것은 지능을 시뮬레이션하는 기계를 만드는 방법을 논의한 1956년 다트머스 워크샵이다. 그 후 인공지능은 황금기와 침체기를 거듭 반복하다가 최근 인공지능경망의 효과적인 학습방법으로서 딥러닝이 발표되고 GPT, BERT 등 언어모델에 의한 자연어 처리 및 기계학습의 혁신적인 성과가 자율주행 자동차, 의료진단 등 다양한 산업으로 확산하면서 국가경쟁력의 핵심요소로 부상하고 있다.

특히 생성형 인공지능은 미리 학습한 데이터와 정보를 기반으로 텍스트, 이미지, 음성 등을 변형하거나 생성할 수 있으므로 문서 작성 및 요약으로부터 프로그램 코딩, 비즈니스 응용, 예술 작품 제작에 이르기까지 활용 범위가 매우 넓다. 2022년 11월 발표된 생성형 인공지능 챗GPT는 5일 만에 사용자 수 1백만 명을 확보하고 2개월 만에 월간 사용자 수 1억 명을 넘어섰다. 챗GPT의 등장은 그동안 일부 기업 및 전문가 영역에서 제한적으로 활용되던 인공지능이 일반인들까지 누구나 쉽게 사용할 수 있게 됨으로써 인공지능이 일상생활 속으로 본격적으로 도입되기 시작하였다는 데 큰 의미가 있다. 그림 1은 다양한 생성형 언어모델 인공지능 모델을 보여준다.

[그림 1-1-1] 초거대 AI와 파라미터 수



* 출처: Artificial Intelligence Index Report 2023, Stanford HAI, 2023

전 세계 62개 국가에 대하여 인공지능의 구현 역량, 혁신 역량, 투자 역량 등을 평가하는 〈Tortoise Global AI Index〉의 2023년 6월 발표에 따르면 미국(100)과 중국(61.5)이 각각 1위와 2위를 차지하며 치열한 패권 경쟁을 벌이고 있으며, 한국(40.3)은 싱가포르(49.7), 영국(41.8)에 이어 캐나다(40.3)와 함께 공동 5위를 기록하였다. 한국의 운영환경 및 정부전략 부문은 높이 평가되었으나 인재, 연구, 상용화 부문은 상대적으로 취약한 상태로서 향후 인재 양성, 연구 지원, 벤처 육성 및 투자에 대한 적극적인 지원 정책이 요구된다.

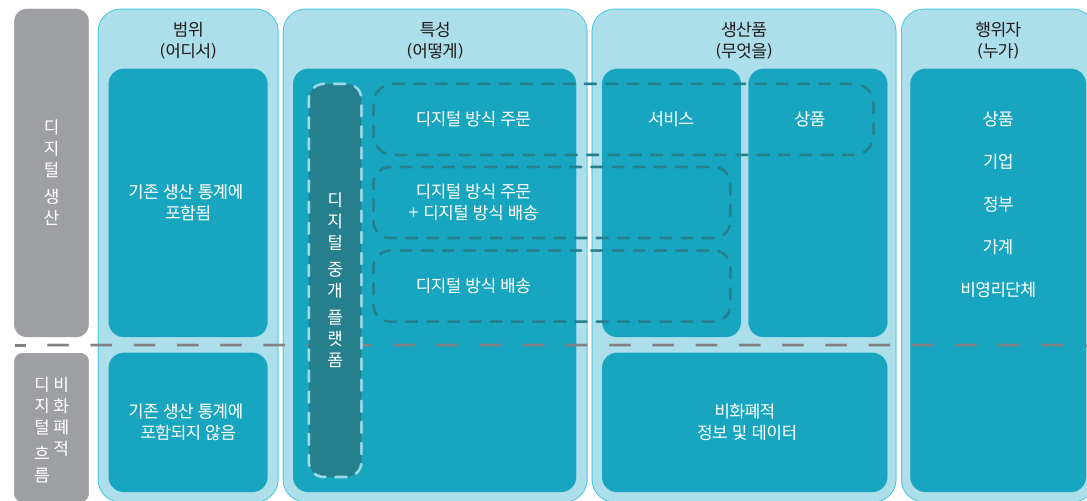
[표 1-1-1] 국가 AI 역량평가 (상위 10개국)

순위	국가	총점	구현 역량			혁신 역량		투자 역량	
			인재	인프라	운영환경	연구	개발	정부전략	상용화
1	미국	100.0	100.0	100.0	82.8	100.0	100.0	90.3	100.0
2	중국	61.5	30.0	92.1	99.7	54.7	80.6	93.5	43.1
3	싱가포르	49.7	56.9	82.8	85.7	48.8	24.4	81.8	26.2
4	영국	41.8	53.8	61.8	79.5	38.1	19.8	89.2	20.0
5	캐나다	40.3	46.0	62.1	93.1	34.0	18.9	93.4	18.9
5	한국	40.3	35.1	74.4	91.4	24.3	60.9	91.9	8.3
7	이스라엘	40.0	45.5	60.5	85.1	24.8	22.2	31.8	40.5
8	독일	39.2	57.0	68.2	90.7	29.3	19.5	93.9	10.3
9	스위스	37.7	44.5	68.0	81.9	41.3	24.9	9.0	13.3
10	핀란드	34.9	34.5	73.0	97.7	27.4	13.1	82.7	9.5
평균		48.5	50.3	74.3	88.8	42.3	38.4	75.8	29.0

* 출처: The Global AI Index, Tortoisemedia.com, 2023.6.28.

디지털 경제는 디지털 기술, 디지털 인프라, 디지털 서비스 및 데이터를 포함한 디지털 환경을 활용하는 모든 경제 활동을 포함한다. 그림 3은 OECD의 디지털 경제 프레임워크를 나타낸 것이다. 디지털 경제는 생산과 소비가 중간단계를 거치지 않고 직접 연결되어 비용을 절감할 수 있을 뿐만 아니라 디지털 중개플랫폼 참여자들의 거래활동에 대한 빅데이터를 인공지능으로 분석하여 최적화된 개인 맞춤형 서비스를 제공할 수 있다. 시장조사기관 Statista는 이러한 데이터의 저장, 처리, 유통을 담당하는 글로벌 데이터센터의 트래픽이 2013년 이후 매년 평균 26.7% 증가하여 2021년에 20.6 제타(10²¹)바이트를 넘어서는 것으로 추정하였으며 당분간 이러한 증가추세는 지속될 전망이다.

[그림 1-1-2] 디지털경제의 개념적 프레임워크



* 출처: A roadmap toward a common framework for measuring the Digital Economy, OECD, 2020

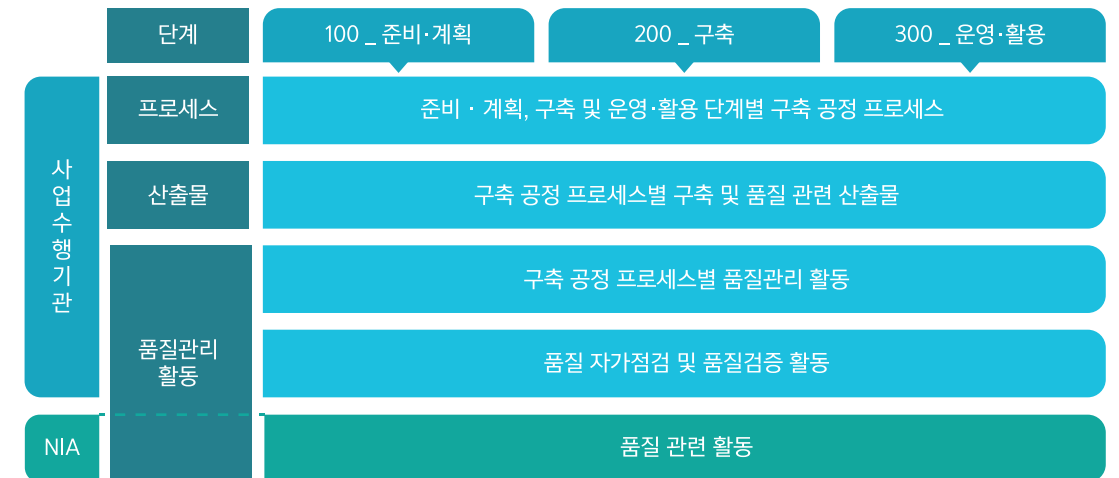
2. 데이터의 품질과 가치

지도학습과 강화학습을 사용하는 생성형 인공지능 GPT-3의 파라미터는 1750억개이며 WuDao 2.0의 경우 1조개가 넘는 파라미터를 사용한다. 이들 파라미터를 학습시키기 위한 데이터의 크기는 GPT-3의 경우 570테라바이트의 웹페이지 텍스트를 이용할 정도로 막대한 양의 학습용 데이터가 필요하며 이러한 인공지능 학습용 데이터의 품질은 학습 소요시간은 물론 최종적인 인공지능의 서비스 품질에 결정적인 영향을 미친다.

과학기술정보통신부는 지난 22년 시행된 '데이터 산업진흥 및 이용촉진에 관한 기본법'에 따라 데이터 품질과 신뢰성 확보를 위한 데이터 품질인증 제도를 본격 시행하고 있다. 데이터 품질관리 활동을 위한 가이드라인도 지속적으로 업데이트되고 있다. 그림 4는 인공지능 학습용 데이터 품질관리 프레임워크이다. 사업수행기관이 담당하는 품질관리 대상은 프로세스, 산출물, 품질관리 활동이며, 한국지능정보사회진흥원은 품질관리 활동 및 결과물의 적정성을 검토하고 검증하는 역할을 담당한다. 품질활동 단계는 준비·계획, 구축, 운영·활용 등 3단계로 구분되며 이 가운데 구축 단계는 세부적으로 데이터 획득/수집(원시데이터), 데이터 정제(원천데이터), 데이터 가공(라벨링데

이터, 데이터 학습(학습용데이터) 등으로 구성된다. 품질관리 지표는 데이터 생애주기 분석, 인공지능 학습용 데이터 구축 및 품질 관점의 일차성 분석, 데이터 품질관리 기준 분석을 통해 구축 및 활용 관점을 반영하여 준비성, 완전성, 유용성, 기준 적합성, 기술 적합성, 통계적 다양성, 구문 정확성, 의미 정확성, 알고리즘 적정성, 유효성 등 10가지 지표로 구성된다.

[그림 1-1-3] 품질관리 프레임워크 구성 요소



* 자세한 내용은 '품질관리 가이드라인 v3.0' 내 '인공지능 학습용 데이터 품질관리 프레임워크' 참고

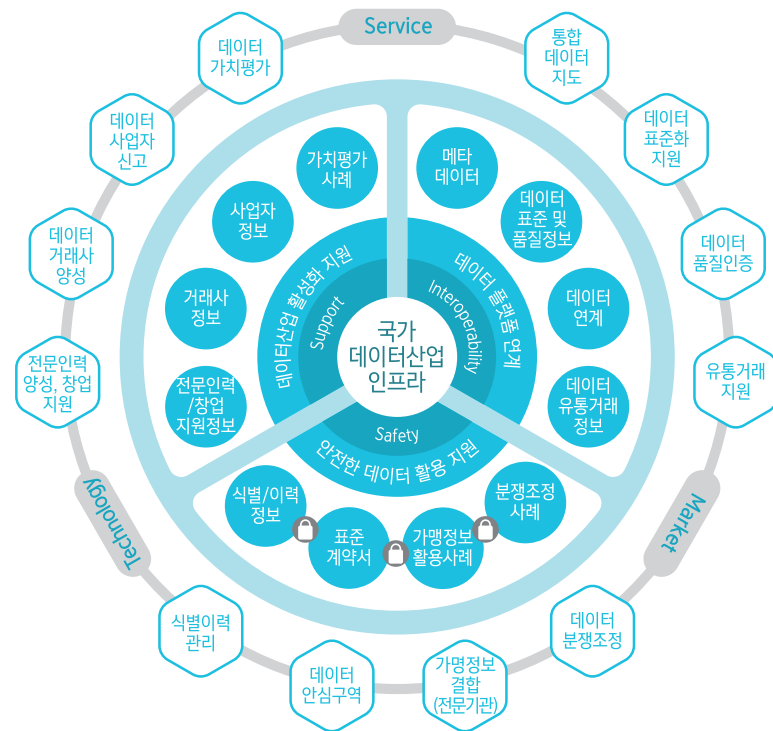
* 출처: 인공지능 학습용 데이터 품질관리 가이드라인 및 구축 안내서 v3.0, 과기정통부/NIA/TTA, 2023. 2.

3. 데이터산업 인프라

'데이터 산업진흥 및 이용촉진에 관한 기본법(데이터산업법)' 제4조 제1항에 따라 정부는 데이터 생산, 거래 및 활용을 촉진하고 데이터산업의 기반을 조성하기 위하여 3년마다 관계 중앙행정기관의 장과 협의를 거쳐 데이터산업 진흥 기본계획을 수립하여야 하며 이에 기초하여 2023년 1월 '제1차(23~25) 데이터산업 진흥 기본계획'이 발표되었다.

제1차 데이터산업 진흥 기본계획은 세계 최고 데이터 강국 도약을 비전으로 삼아 데이터 시장 규모 50조원, 데이터 활용역량 세계 10위권, 기업 데이터 도입률 30% 이상을 목표로 하고 있다. 이러한 비전과 목표를 달성하기 위한 중점 추진 과제는 생산공유개방, 유통거래, 보호활용, 산업기반 등 4개 전략 아래 17개 과제가 제시되었다. 주목할 점은 '2-1. 데이터 산업 생태계 지원 통합 국가인프라' 과제에서 산재된 민간·공공 데이터 정보를 종합 연계하여 데이터 접근성과 활용성을 제고할 수 있는 기반 구축 계획 수립(인프라 기획)과 함께 민간·공공의 다양한 데이터 플랫폼·포털을 연계하는 최상위 국가 인프라, 'Platform of Platforms' 구축(One-윈도우 구축)을 제시하였다는 점이다. 그림 5는 데이터 플랫폼 연계, 안전한 데이터 활용 지원, 데이터산업 활성화 지원 등으로 이루어진 통합 국가인프라 개념도이다.

[그림 1-1-4] 국가 데이터산업 인프라



* 출처: 제1차(2023-2025) 데이터산업 진흥 기본계획, 관계부처 합동, 2023. 1

4. 데이터의 활용

미국 스탠포드대학 인공지능100년(AI100) 보고서에 따르면 향후 데이터 및 인공지능 활용에 의해 가장 많이 변화될 분야로 교통, 의료, 교육, 복지, 안전, 고용, 로봇, 오락 분야 등이 예상된다.¹⁾

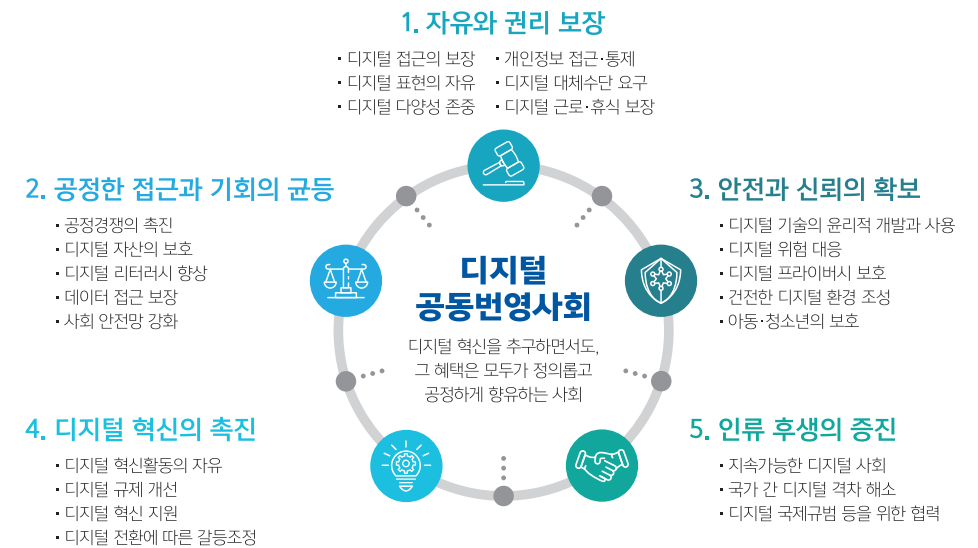
교통 분야에서는 교통 정보에 의한 자율주행이 일반화될 것이며 차량 공유가 확산되고 출퇴근 부담이 낮아져 도시 집중이 완화된다. 의료 분야는 임상 지원, 환자 모니터링, 수술이나 환자 치료를 지원하는 로봇, 의료 시스템 관리가 포함되며 의료 데이터의 수집 및 활용이 핵심사항이다. 교육 분야에서 인공지능은 대규모 개인화를 제공함으로써 모든 수준에서 교육을 향상시킬 것이며 인간 상호 작용과 인공지능 기술과 잘 통합하는 것이 과제이다. 복지 분야는 인공지능과 소셜 네트워크를 결합하여 대규모 인구를 적절한 때에 적절한 방법으로 지원할 수 있으며 인공지능에 의한 차별적 행동이 방지되도록 주의해야 한다. 안전 분야는 인공지능에 의해 사전에 범죄 가능성을 예측하는 것이 가능해지며 카메라, 드론 등을 통해 수집된 감시 데이터가 오남용되지 않도록 하여 대중의 신뢰를 얻는 것이 중요하다. 고용 분야에서 인공지능은 소량의 업무 대체 또는 강화부터 완전한 대체에 이르기까지 다양한 효과가

1) Artificial Intelligence and Life in 2030, One Hundred Year Study on Artificial Intelligence, Report of the 2015 Study Panel, Sept. 2016

나타날 것이며 효율적으로 아웃소싱된 노동 시장을 창출할 것이다. 오락 분야는 인터넷과 소셜 네트워크가 사회적 상호 작용 및 엔터테인먼트의 개인화된 플랫폼 역할을 수행하고 있으며 전통적인 엔터테인먼트인 프로 스포츠, 음악 작곡, 무대 제작, 영상 분석 등에서도 인공지능의 광범위한 도움을 받고 있다.

5. 디지털 권리장전과 데이터

[1-1-5] 디지털 권리장전의 5대 원칙



* 출처: 대한민국이 새로운 디지털 규범질서를 전 세계에 제시합니다. 과기정통부 보도자료, 2023.09.26.

미래의 국가 데이터 산업을 지원하는 정부 정책 가운데 최상위를 차지하는 정책 중 하나가 2023년 9월 발표된 전문 및 총 6장 28개조의 본문으로 이루어진 『디지털 권리장전』이다.

『디지털 권리장전』은 국제사회가 함께 추구해 나갈 모범적인 미래상으로 '디지털 혁신을 추구하면서도 그 혜택을 모두가 정의롭고 공정하게 향유하는 디지털 공동번영사회'를 제시하고, 그 실현을 위한 원칙들을 규정하였다.

세부적으로는 제1장에서 '디지털 공동번영사회' 구현을 위한 기본원칙을 규정하고, 제2장부터 제6장에서는 자유와 권리 보장, 공정한 접근과 기회의 균등, 안전하고 신뢰의 확보, 디지털 혁신의 촉진, 인류 후생의 증진 등 5대 기본원칙을 구현하기 위한 시민의 보편적 권리와 주체별 책무(국가·기업·시민)를 세부 원칙의 형태로 규정하였다. 특히 제15조(데이터 접근 보장)에서는 데이터 개방의 촉진과 함께 공공 데이터의 접근과 이용이 보편적으로 확대되도록 필요한 조치가 이루어져야 함을 규정하고 있다. 그림 6은 디지털 권리장전의 5대 원칙을 보여준다.

정부는 '새로운 디지털 질서'의 기본방향을 담은 『디지털 권리장전』을 기준으로 삼아, 디지털 심화시대의 쟁점들을 해소하고 구체적인 법·제도를 정비하기 위한 실질적인 노력들을 이어나갈 계획이다.

제2장 초거대 AI 시대 데이터 공유

전성민 교수 가천대학교 경영학부 / 서울대학교 AI연구원 객원연구원

생성형 AI 기술의 발전으로 데이터의 중요성이 더욱 커지고 있다. 생성형 AI는 기존의 AI 기술과 달리 데이터를 활용하여 새로운 정보를 생성할 수 있다. 이에 따라 창의성, 개인화 및 효율성 수준을 높일 엄청난 기회가 생겼다. 동시에 데이터 소유권과 프라이버시를 둘러싼 도전적이고 민감한 이슈도 떠오르고 있다. 개인에 대한 위험을 최소화하면서 생성형 AI의 장점을 극대화하는 데이터 관리 방안이 필요하다.

1. 생성형 AI와 데이터

챗GPT가 대중적인 서비스를 시작한 이후 생성형 AI(Generative AI)에 관한 관심이 급증했다. 생성형 AI 모델은 인간과 유사한 텍스트, 이미지, 심지어 음악까지 생성하는 능력을 크게 높였다. 이러한 발전 덕분에 창의적인 응용 프로그램에 대해 흥미로운 가능성이 생겼을 뿐 아니라 생산성도 향상했다는 평가를 받는다.

생성형 AI 모델은 대규모 데이터를 가지고 훈련받은 후 새로운 텍스트, 이미지, 음악을 생성할 수 있다. 이러한 모델은 다양한 형태의 창의적 응용 프로그램에 사용될 수 있다. 또한, 현장에서 실제로 업무 생산성을 높이는 데에도 적용할 수 있다. 예를 들어 텍스트를 생성하여 '문서'를 작성하거나, 이미지를 생성하여 '제품을 홍보하는 것'에 활용할 수 있다. 음악을 생성하여 '광고 음악'을 만들 수도 있다.

이렇듯 생성형 AI를 개발하고 사용하기 위해서는 방대한 데이터가 필요하다. 또한, 생성형 AI는 기존 데이터와 유사한 데이터를 생성해야 하므로, 데이터의 품질도 중요하다. 따라서 생성형 AI 시대에 데이터는 더욱 중요한 자원이 될 것이다.

그러나 생성형 AI 모델에 대해서 우려할 만한 점도 제기된다. 우선 생성형 AI 모델은 대규모 데이터로 훈련하다 보니 해당 데이터에 포함된 개인정보 노출 위험이 있다. 그뿐 아니라 가짜 뉴스나 가짜 콘텐츠를 생성하는 데 악용될 수도 있다. 이런 문제점을 주지하여 생성형 AI 모델을 책임감 있게 사용하고, 데이터 소유권 및 개인정보 보호에 대한 우려를 해결하는 것이 더욱 중요해졌다.

2. 생성형 AI 데이터 학습 기술

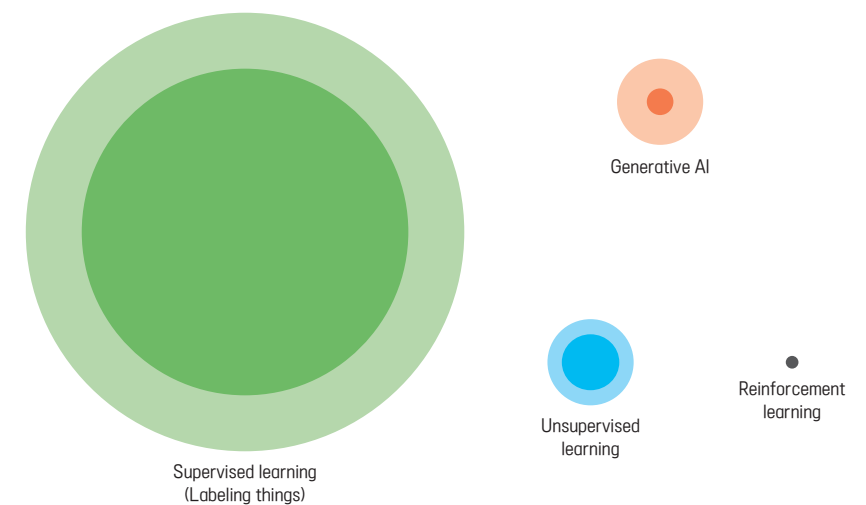
챗GPT가 일반에 공개된 지 채 몇 개월도 되지 않아 텍스트, 이미지, 비디오, 심지어 자동완성 컴퓨터 코드와 같은 새로운 콘텐츠를 생성하는 기술을 도입하면서 AI의 개념 자체가 크게 바뀌었다. 생성형 AI의 경우 일반에는 최근에 알려졌지만, 사실 새로운 기술은 아니다. AI 영역에는 전통적으로 지도학습, 비지도 학습, 강화학습이 이미 존재했다.

첫째, 지도 학습은 레이블이 지정된 데이터를 사용하여 모델을 학습하는 기술이다. 예를 들어 지도 학습 모델을 사용하여 '숫자 이미지'의 숫자를 인식하거나, 텍스트의 언어를 번역할 수 있다. 둘째, 비지도 학습은 레이블이 지정되지 않은 데이터를 사용하여 모델을 학습하는 기술이다. 예를 들어 비지도 학습 모델을 사용하여 이미지의 객체를 분류하거나, 텍스트의 주제를 추출할 수 있다. 셋째, 강화학습은 에이전트가 환경과 상호 작용하면서 보상을 통해 학습하는 기술이다. 예를 들어 강화학습 모델을 활용하여 로봇이 장애물을 피하면서 목적지까지 이동하거나, 게임 캐릭터가 적을 물리치면 보상을 얻을 수 있다.

생성형 AI는 지도학습, 비지도학습, 강화학습의 기술을 모두 적용하여 새로운 데이터를 생성한다. 지도학습은 생성형 AI를 학습시키는 데에, 비지도학습은 생성형 AI를 개선하는 데에, 강화학습은 생성형 AI를 제어하는 데 사용된다. 생성형 AI는 기존의 AI 기술을 활용하여 빠르게 발전하고 있다. 아래 <그림 1-2-1>에서 볼 수 있듯이 생성형 AI는 지도학습에 비해 규모는 적지만, 가장 빠르게 성장할 것으로 예측된다.

[그림 1-2-1] AI 기술의 가치

Value from AI technologies: Today ⇒ 3years



* 출처: 정보통신산업진흥원 NIPA, "AI 석학 앤드류 응 대담회(AI Talk with Andrew Ng)", 2023. 7. 21., 2023년 9월 3일 접속, <https://www.youtube.com/watch?v=lgzi0Tum4r4>

3. 생성형 AI 시대의 데이터

생성형 AI가 다양한 모델 형태로 작동하지만, 혁신적이고 현실적인 결과를 생성하기 위해 필수적으로 품질 좋은 데이터를 확보해야 한다. 생성형 AI에서 데이터의 역할, 모델 훈련, 정확성 향상, 창의적 발전 등의 측면에서 데이터의 역할을 검토해 보고자 한다.

가. 훈련 모델

생성형 AI 모델은 데이터로 학습하여 특정 작업을 수행하고, 예측하고, 새로운 결과를 생성하는 능력을 획득하는 일련의 과정을 거치는데, 이때 훈련 과정은 매우 중요한 단계다.

데이터의 경우엔 AI 모델 학습의 기초 역할을 한다. 이러한 모델은 방대한 양의 데이터에서 패턴과 기본 구조를 학습하여 새로운 콘텐츠를 생성하므로, 훈련 데이터의 품질, 다양성, 관련성 및 대표성은 모델의 출력에 큰 영향을 미친다. 구체적으로 보면 AI 모델의 정확성, 성능 및 공정성에 직접적인 영향을 미칠 뿐 아니라 학습, 패턴 인식 및 적응성을 촉진해 준다. 이런 이유로 AI 모델을 훈련하는 데 데이터는 매우 중요하다.

나. 정확성

데이터는 생성형 AI 모델의 정확성에 큰 역할을 한다. 대규모의 고품질 데이터를 사용하면 모델이 복잡한 세부 사항을 학습하여 정확하고 현실적인 결과를 얻을 수 있다. 또한 업데이트된 관련 데이터에 지속적으로 노출되면, 모델이 시간이 지남에 따라 적응하고 개선될 수 있다. 그렇게 인간의 기대에 부합하는 콘텐츠를 생성하는 능력이 향상된다.

다. 창의성

생성형 AI는 데이터를 기반으로 상상의 한계를 뛰어넘는다. 모델을 다양한 예술 스타일, 장르 또는 문화적 영향을 포함한 다양한 데이터 세트에 노출함으로써, 모델이 다양한 창의적 요소를 학습하고 결합할 수 있도록 한다. 이 프로세스를 통해 AI는 독특한 아이디어를 선보이고, 다양한 스타일을 융합한다. 또는 새로운 예술적 표현에 기여하고 뛰어난 조합의 콘텐츠를 생성한다.

데이터는 정확성을 높이고 상상력을 키우며 생성형 AI의 발전을 촉진하는 데 도움이 된다. 다양하고 관련성이 높은 고품질 데이터를 확보할수록 생성형 AI 모델의 개발을 촉진할 수 있다. 그럴 때 생성형 AI 모델을 통해 정확하고 혁신적인 콘텐츠 생산이 가능해진다.

또한 책임감 있고 윤리적으로 데이터를 활용할 때 생성 AI의 힘을 연결하여 복잡한 문제를 해결할 수 있고, 다양한 산업 전반에 걸쳐 전혀 없는 발전을 이루는 것이 가능해진다. 궁극적으로는 지금보다 혁신적이고 번영하는 미래를 만들 수 있다.

4. 생성형 AI의 데이터 활용 이슈

이렇듯 빠른 성장이 예상되는 생성형 AI는 사회와 경제에 큰 영향을 미칠 잠재력을 지니고 있다. 예를 들어 생성형 AI로 새로운 제품을 디자인하거나, 새로운 서비스와 경험을 제공할 수 있다.

이러한 생성형 AI의 잠재력을 최대한 활용하려면 데이터 소유권, 개인정보 등 윤리적 문제를 해결해야 한다.

일반적으로 논의되는 문제 중에는 저작권 침해, 생성 알고리즘에 내재한 편향의 존재, AI 기능을 과대평가하여 잘못된 출력(AI 환각)을 유포할 위험, 대중을 혼란에 빠지게 할 딥페이크 및 합성 콘텐츠 생성 등이 있다. 이러한 문제를 해결하려면, 윤리적이고 유익한 생성형 AI 구현을 위한 책임 있는 프레임워크를 개발하는 것이 필수적이다.

가. 생성형 AI와 데이터 소유권

생성형 AI가 계속 진화하면서 놀라운 결과를 만들어 내자 데이터 소유권, 데이터 프라이버시 및 윤리적 고려 사항 등이 중요한 토론 주제로 등장했다. 생성형 AI 모델은 원본 콘텐츠를 생성하도록 훈련받는다. 따라서 출력물과 함께 입력 데이터의 소유권이 복잡한 문제가 된다. 생성형 AI는 텍스트, 이미지, 비디오 등 다양한 종류의 콘텐츠를 생성할 수 있다. 이 콘텐츠는 종종 실제 콘텐츠와 구별할 수 없을 정도로 사실적이다. 따라서 생성된 콘텐츠의 소유권과 사용에 대한 법적 권리가 누구에게 있는지 명확히 하는 것이 중요하다.

이때 두 측면을 살펴야 한다. 우선 생성된 콘텐츠의 소유권은 일반적으로 생성형 AI 모델을 개발한 사람에게 있다. 그래서 생성된 콘텐츠를 상업적인 용도로 사용하는 경우 생성된 콘텐츠 소유자의 허락을 받아야 한다. 물론 생성된 콘텐츠를 개인적인 용도로 활용하는 경우엔 생성된 콘텐츠 소유자의 허락 없이 사용할 수 있다.

그런가 하면 둘째로 살펴야 할 사람은 데이터 제공자다. 생성된 콘텐츠에 사용된 데이터의 소유권은 데이터 제공자에게 있기 때문이다. 따라서 생성된 콘텐츠의 소유권을 결정하기 위해서는 생성된 콘텐츠에 사용된 데이터의 소유권을 고려해야 한다. 생성형 AI가 텍스트, 이미지, 오디오 및 기타 데이터를 활용하여 학습하기 때문이다. 보통 이 데이터는 공개적으로 사용 가능한 다양한 출처에서 또는 동의를 받아 수집한다. 생성형 AI에서 만들어진 출력물은 개인이나 개체에서 직접 파생되지 않으므로 기존 소유권 경계가 흐려진다. 이 때문에 생성형 AI에 활용되는 데이터의 소유권 문제가 발생할 수 있다. 기존 데이터의 지식재산권 소유자는 파생 생성된 데이터의 저작권을 주장할 수 있다. 이런 맥락에서 생성형 AI를 사용하여 생성된 데이터를 활용할 경우, 해당 데이터 세트의 소유자와 라이선스 계약을 체결해야 할 수도 있다.

생성형 AI에서 데이터 소유권 문제를 해결하려면 개발자, 연구원, 정책 입안자 및 사회 간의 협력이 필요하다. 투명성, 책임성 및 책임 있는 데이터 사용을 보장하기 위해 명확한 지침과 프레임워크를 수립해야 한다. 학습된 데이터에 포함된 데이터의 소유권을 가진 개인 또는 조직으로부터 명시적인 동의를 얻고 민감한 정보를 보호하기 위한 메커니즘을 구현하는 것이 중요하다.

구체적으로 데이터 소유권에 관한 명확한 지침을 수립하고, 데이터 사용의 책임을 명확히 해야 한다. 또한 데이터 사용의 투명성을 보장하면서도 민감한 정보를 보호해야 한다. 이를 위해 데이터 사용에 대한 윤리적 고려사항을 반영해야 한다. 이러한 조치를 선결 요건으로 적용할 때 생성형 AI에서 데이터 소유권 문제를 해결하고, 책임 있는 데이터 사용을 보장할 수 있다.

나. 생성형 AI와 개인정보 보호

생성형 AI는 보통 기존 데이터의 패턴을 활용하여 새로운 데이터를 생성하므로 데이터 내에 포함된 개인정보가 노출될 위험이 있다. 예를 들어 생성형 AI로 얼굴 이미지를 생성할 경우, 해당 이미지는 실제 사람의 얼굴과 매우 유사할 수 있다. 따라서 생성형 AI로 생성한 데이터는 개인을 식별하거나 추적할 때 활용할 수 있다.

그러나 많은 생성형 AI 플랫폼에선 데이터 프라이버시를 보장하지 않는다. 고객 데이터의 개인정보보호 및 기밀유지를 중시하는 기업에서는 매우 우려할 만한 일이다. 민감한 정보가 챗봇이나 생성형 AI 모델에 입력되면 기업은 해당 데이터가 어떻게 활용될지 예측할 수 없다. 개인정보에 대한 통제력이 부족하면 기업은 위험에 처할 수 있다.

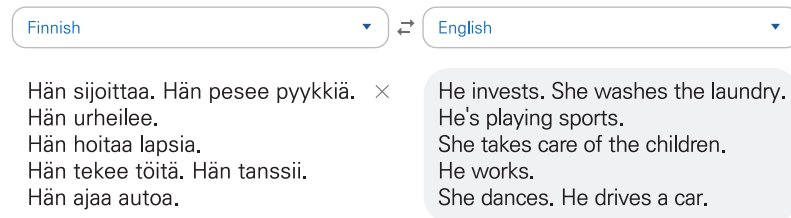
이러한 개인정보 보호 문제를 해결하기 위해서는 기업과 정부가 협력해야 한다. 기업은 생성형 AI 플랫폼을 개발할 때 데이터 프라이버시를 보장하는 기능을 포함해야 하며, 정부는 생성형 AI의 개발과 사용을 규율하는 법과 제도를 마련해야 한다.

다. 그 외 생성형 AI의 데이터 활용 이슈

- 데이터 편향

생성 AI 모델은 훈련된 데이터에서 패턴과 정보를 학습한다. 훈련 데이터에 편견이나 선입견이 포함되면 생성된 출력에 반영될 수 있다. 예를 들어 학습 데이터가 특정 인구 통계 또는 관점으로 편향된 경우 생성된 콘텐츠가 편향을 강화할 수 있다. 편향되지 않은 공정한 결과를 보장하기 위해 학습 데이터의 편향을 완화하고자 노력을 기울여야 한다.

[그림 1-2-2] 데이터 편향 사례 A



* 출처: Joanna J Bryson, Vuokko Aro, "2021년 3월 9일 21시 10분 트윗글", 2021.3.9., 2023년 9월 3일 접속, <https://twitter.com/j2bryson/status/1369259307540377606>

그림 1-2-2는 AI를 이용한 번역에서 성적 차별이 존재하는 사례를 보여준다. 핀란드어로는 그와 그녀를 구분하지 않는데, 영어로 번역할 경우 행동을 보여주는 동사에 따라 그와 그녀가 구분되는 예를 보여준다. 투자, 스포츠, 일, 운전 등은 남성으로 편향되고, 세탁, 육아, 춤은 여성이 특정되어 번역된 것이다.

[그림 1-2-3] 데이터 편향 사례 B

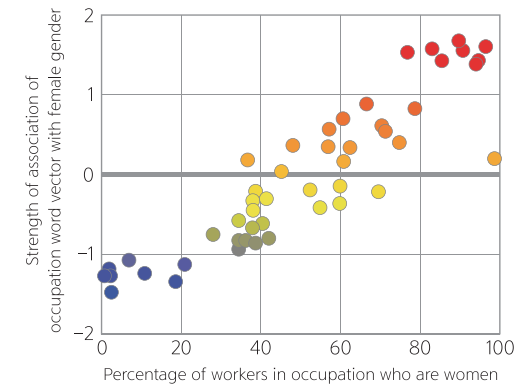


Fig. 1. Occupation-gender association. Pearson's correlation coefficient $p = 0.90$ with $P < 10^{-18}$.

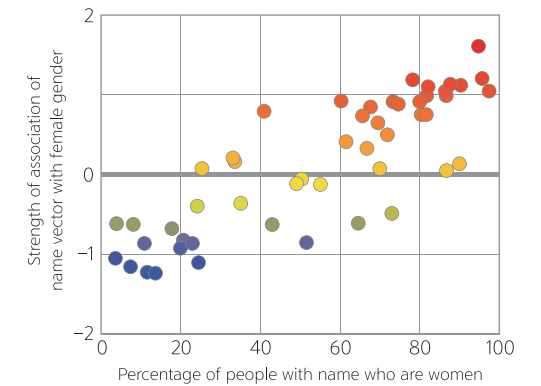


Fig. 2. Name-gender association. Pearson's correlation coefficient $p = 0.84$ with $P < 10^{-13}$.

* 출처: A. Caliskan, J. J. Bryson, A. Narayanan, "Semantics derived automatically from language corpora contain human-like biases", Science Vol.356 Issue.6334, 2017, p.183-186.

그림 1-2-3 역시 이러한 직업 및 이름이 성적 구별과 연관성이 있다는 점을 직관적으로 보여준다.

- 얼굴 인식 및 딥페이크

생성형 AI 기술의 발전으로 얼굴 인식 및 딥페이크에 대한 우려가 높아졌다. 이는 사생활 침해, 신원 도용, 명예 훼손에 대한 중대한 위험 요인이다. 또한 오해의 소지가 있거나 악의적인 콘텐츠를 생성하기 위해 기술을 오용할 경우, 조작된 이미지나 동영상이 만들어질 수 있다.

- 음성 합성 및 사칭

생성형 AI의 발전으로 사람과 같은 음성 합성이 가능해지면서 음성 사칭 및 조작에 대한 우려가 커졌다. 이것은 다른 사람의 목소리를 모방하고 다른 사람을 속이는 데 악용될 수 있으며 잠재적으로 사기 또는 기타 유해한 활동으로 이어질 수 있다.

- 합성 데이터를 통한 개인정보 침해

생성형 AI 모델은 실제 데이터와 매우 유사한 합성 데이터를 생성할 수 있다. 이는 민감한 정보를 노출하지 않고 AI를 학습시키는 데 유용할 수 있지만 합성 데이터의 무단 사용에 대한 우려가 있다. 자칫 합성 데이터에 섞인 개인정보의 침해 및 프로파일링, 감시 또는 기타 침입 목적으로 악용될 수 있다.

– 재식별 공격

생성형 AI 기술은 비공개 또는 익명 데이터를 역공학(Reverse Engineering) 하여 사용할 수 있다. 공격자는 합성 샘플을 생성함으로써 여러 데이터를 결합하여 개인을 재식별하고 민감한 정보에 다시 연결하여 개인정보를 손상할 수 있다.

– 동의 및 제어 부족

개인은 생성형 AI 모델에서 자신의 데이터가 사용되는 방식에 대한 제어 및 인식이 제한적일 수 있다. 동의 메커니즘이 없거나 데이터 사용에 대한 투명한 정보가 없으면, 개인정보 보호 권리와 개인의 자율성이 약해질 수 있다. 결과적으로 개인 데이터를 사용하거나 배포할 때 철저하게 통제하지는 못하는 바람에, 개인정보가 침해될 가능성이 있다.

– 알고리즘 편향 및 차별

생성형 AI 모델은 훈련된 데이터로 학습하는데, 이때 데이터는 편향적 요소가 포함될 수 있다. 이로 인해 사회적 편향이 지속되거나 증폭되어, 생성된 콘텐츠에서 차별적인 결과를 도출할 수 있다. AI 생성 결과의 공정성, 동등한 대표성 및 잠재적 차별에 대해 유의해야 한다.

– 타사 데이터 집계

생성형 AI 모델은 훈련 목적으로 공개적으로 사용 가능한 정보 및 타사 데이터 소스를 포함하여 방대한 양의 외부 데이터에 액세스해야 하는 경우가 많다. 명확한 경계가 없는 여러 데이터 소스를 집계할 때 다양한 소스의 개인 정보가 결합되고 연결된다. 이에 자연스럽게 개인정보 보호 문제가 떠오른다.

– 부적절한 규제 및 법적 프레임워크

생성형 AI의 급속한 발전으로 다양한 프라이버시 침해 문제가 발생할 것을 예상할 수 있다. 그런데 이 문제를 해결하기 위한 적절한 규정 및 법적 프레임워크의 개발 속도보다, 생성형 AI의 발전 속도가 빠르다. 이 때문에 규제의 허점을 오용하는 상황도 생기는데, 위반에 대해 책임져야 할 당사자에게 적절한 책임을 묻는 데 시간적 격차가 생긴다.

– 투명성 및 설명 가능성 부족

생성형 AI의 블랙박스 특성은 투명성과 설명 가능성을 제한한 탓에, 복잡하고 해석하기 어렵다는 특성이 있다. 또한 콘텐츠 생성 방법이나 잠재적 개인정보 위험 식별 방법도 이해하기 어렵다. 이처럼 투명성 부족 탓에 생성형 AI 기술의 프라이버시 영향을 평가하는 작업은 더욱 어려워진다.

5. 결론 - 생성형 AI와 데이터 활용의 미래

데이터 이슈에 접근하기 위해서는 우선, 데이터 사용에 대한 투명한 지침 및 동의 메커니즘을 설정해야 한다.

둘째로는, 개인으로서 처할 위험을 최소화하면서 동시에 생성형 AI의 이점을 실현하기 위해 개인정보 보호 기술 개발에 최선을 다해야 한다. 대표적인 최근 기술로는 차등 프라이버시, 연합 학습 및 안전한 다자 간 계산 등을 예로 들 수 있다. 생성형 AI 학습 중 사용된 데이터를 익명화하려는 기술을 연구하고 개발하려는 노력이 지속적으로 필요하다.

셋째, 개인 역시 생성형 AI를 이용하여 생성된 출력물을 적절히 제어할 수 있는 메커니즘이 필요하다. 예를 들어 특정 유형의 콘텐츠를 제외하는 상황도 생각해 볼 수 있고, 특정 주제나 민감한 정보가 생성되지 않도록 하는 옵션을 제공하는 것도 한 방법이다.

마지막으로 훈련 데이터의 소유권을 식별하고, 보존할 수 있는 메커니즘을 구축해야 한다. 이를 위해 데이터를 수집하고 처리하는 방법에 대해 접근이 가능해야 하고, 자세한 정보를 데이터 공급자에게 제공하는 방식으로 투명성을 확보해야 한다.

종합하자면 대규모 AI 시대에 걸맞은 기술, 법률, 윤리 및 사회적 고려사항을 결합하는 다차원적 접근 방식이 필요하다.

생성형 AI는 창의성, 개인화 및 효율성의 수준을 높일 엄청난 잠재력을 지녔다. 생성형 AI 기술이 빠르고 지속적으로 발전하면서 데이터의 중요성 역시 더욱 커지고 있으며, 수많은 기회와 함께 도전적인 문제도 꾸준히 생겨날 것이다. 새로운 시대로 건설적으로 나아가려면, 혁신과 보호 사이의 섬세한 균형을 유지하는 노력이 필요하다.

제3장 데이터 활용의 이슈 : 데이터 편향성/윤리성

변순용 교수 서울교육대학교 윤리교육과

인공지능 시대의 사회적·기술적 변화가 생기면서, 데이터 윤리가 변화의 핵심 쟁점이 되었다.

이 장에서는 데이터의 윤리적 문제를 제기하고, 데이터 윤리의 의미와 필요성을 살펴보고자 한다. 이를 위해 데이터 윤리를 데이터 권리와 데이터 책임의 관점에서 살펴보고, 데이터의 전체 과정에서 윤리적 고려의 필요성을 강조하고자 한다. 또한, 데이터 윤리 가이드라인에 대한 사회적 담론을 형성하고 협의하는 과정이 필요하다는 점을 설명하려고 한다.

1. 데이터 윤리와 데이터 윤리 가이드라인의 필요성

AI를 중심으로 하는 디지털 변혁(Digital Transformation) 시대에는 빅데이터를 강조하는 것을 넘어 인공지능 학습용 데이터의 중요성이 커지고 있다. 이에 인공지능이 어떤 학습 데이터로 학습했는지 검증하는 작업이 인공지능의 사회적 수용성 수준을 가능하는 중요한 기준으로 작용할 것으로 예측된다. 데이터 중심 사회가 형성되면서 데이터에 대한 윤리적 요구 사항은 여러 영역에서 다양하게 표출되고 있다. 공적 영역에서는 데이터의 성별이나 지역, 정치적 편향성의 여부에 대하여 윤리적 검증을 요구하는 목소리가 나오고 있으며, 사적 영역에서도 개인정보와 관련된 데이터의 올바른 수집과 활용을 위해 윤리적 기준을 시급히 마련해야 한다는 주장이 제기되고 있다.

기본적으로 데이터 윤리는 데이터의 책임 있고 지속가능한 사용을 목적으로 한다. 이러한 목적을 달성하기 위해서 데이터 윤리 가이드라인이 필요하다. 이때 데이터의 수집 및 처리 과정부터 결과까지 활용 전반에 걸쳐서 준수해야 할 윤리 원칙을 수립하고, 이에 근거하여 데이터 윤리 가이드라인에 대한 사회적 합의를 해야 한다. 개인의 프라이버시 보호뿐만 아니라 공적 데이터의 수집과 활용, 데이터 주권과 오픈 데이터의 갈등 등에 대한 윤리적 판단 역시 필요하며, 이를 위한 기준을 마련해야 한다.

최근 인공지능 학습용 데이터의 비윤리적 성격으로 문제가 발생하고 있다. 그래서 데이터의 윤리적 정제(the Ethical Cleaning of Data)의 필요성, 데이터 수집 시 윤리적 절차 수립의 필요성이 강조되는 실정이다. 데이터 정제(Data Cleaning)는 기본적으로 잘못된 데이터의 감지, 진단 및 수정하는 과정(Process of Detecting, Diagnosing, and Editing Faulty Data)을 의미한다. 윤리적 정제는 데이터의 내용적 차원에서 혐오나 증오 표현, 차별 등 비윤리적 요소의 제거 과정을 의미한다. 이를 위해서는 데이터 등급제¹⁾가 전제되어야 하겠지만, 데이터 윤리의 차원에서는 투명성, 책임성, 공정성이 매우 중요한 가치로 강조될 수 있다. 따라서 이와 같은 윤리적 가치를 포괄하는 데이터 윤리 가이드라인이 새롭게 마련되어야 한다.

1) 데이터 등급제는 데이터의 윤리적 정제 수준에 따라 데이터 라벨에 데이터의 윤리적 정제의 수준을 고지하도록 하는 체계로, 영화의 등급제와 유사하다. 이 제안대로 데이터의 윤리적 정제 수준을 등급으로 구분할 경우, 인공지능이 타깃으로 하는 연령층에 적절한 등급의 학습데이터로 학습할 수 있다.

2021년 유네스코는 국제기구 차원에서 처음으로 마련한 인공지능의 윤리적 사용에 대한 국제 표준 지침인 “인공지능 윤리 권고”를 채택하였다. 이 권고안에서 데이터 윤리와 관련된 부분을 살펴보면 다음과 같다.²⁾ 첫째, 오류 학습 및 모범사례 공유를 위한 피드백 메커니즘뿐만 아니라, 데이터 수집·선택 프로세스의 적정성, 적절한 보안 및 데이터 보호 조치 등 AI 시스템의 학습 데이터 품질을 연속적으로 평가할 수 있는 데이터 거버넌스 전략을 개발해야 한다. 둘째, 감시 등 프라이버시를 기본적 권리로 인식하고 보호하기 위한 장치(국제법 준수하는 법적 체계)를 마련해야 하며, 기업을 포함한 모든 AI 행위자가 윤리적 영향평가의 일환으로 프라이버시 영향평가를 수행할 것을 권장한다. 셋째, 데이터 주체가 (특히 AI 시스템에서 자신의 개인 데이터 접근권·삭제권에 대하여) 완전한 권리를 향유하도록 보장해야 하며, 넷째, 국경 간 데이터 이전 등 데이터를 상업적 목적으로 이용할 경우 데이터 보호법을 완전히 준수하는 데이터에 대해 적절한 수준의 보호가 필요하다. 그리고 데이터 주권(Data Sovereignty)과 ‘국제적으로 자유로운 데이터 이동’ 간의 균형을 고려해야 한다. 끝으로, 회원국은 오픈 데이터를 장려해야 하며 정책 프레임워크를 검토해야 한다. 특히 정책 프레임워크에서는 정보 접근성 확보, AI 관련 요구사항을 반영하는 열린 거버넌스 포함, 오픈 데이터 활성화를 위한 메커니즘(공공영역이 지원하는 개방형 저장소, 안전하고 합법적 공유를 위한 소스 코드 등)을 반영해야 한다.

이러한 필요성을 반영한 데이터 윤리 가이드라인을 구성하려면 가장 일반적인 윤리적 기준과 의미를 지닌 상위 원칙들로부터 연역적으로 분석하여 구성해야 한다. 이는 연구자나 개발자, 혹은 관리자나 사용자 모두가 데이터의 전 과정에 걸쳐서 숙고하고 검증해야 할 과제다. 여기에 덧붙여 사례분석으로부터 요청되는 윤리 원칙, 또는 기준 간의 중첩이나 갈등의 해결을 위한 메타 원칙까지 포함해야 할 것이다.

2. 데이터 윤리의 사례와 정의

2-1. 데이터 윤리 이슈의 사례

기존의 데이터 시장에서 일어나는 다양한 유형의 윤리적 이슈가 있었지만, 최근에는 인공지능과 결합한 데이터 윤리의 이슈 사례가 또 다른 양상으로 펼쳐지고 있다. 인공지능 시대에 알맞은 데이터 윤리의 새로운 정의를 발견하기 위해 구체적인 사례를 우선 검토해야 한다.

가. 인공지능 챗봇 이루다 사건

2021년 초에 인공지능 챗봇인 이루다가 출시 2주 만에 80만 명의 이용자를 모았지만 성소수자와 장애인 등 사회적 약자에 대한 혐오 발언으로 AI 윤리 논란의 중심에 섰다. 여기에 개인정보 보호 위반의 논란까지 겹치는 바람에 출시한 지 한 달도 안 되어 서비스가 종료되었다.³⁾ 결국 데이터 활용 과정에서 연구자의 비윤리적인 개입 문제뿐

2) UNESCO, Recommendation on the Ethics of Artificial Intelligence, UNESCO.org, (2021).

3) 김남영, “국립국어원 AI 데이터에도 ‘혐오·차별 발언’ 한국경제, 2021년 1월 16일, <https://www.hankyung.com/society/article/202101156377i>

아니라, 데이터 수집 과정에서 데이터 제공자의 동의 문제로 제동이 걸린 것이다. 인공지능 챗봇은 여타의 인공지능 기술보다 이미 인간 사회에 깊게 개입된 상태다. 이렇다 사태의 경우 텍스트형 인공지능 챗봇이 대상이었지만, 삼성의 빅스비나 애플의 시리 같은 음성형 인공지능 챗봇까지 더해진다면 데이터의 윤리적 쟁점은 심화된다. 연령과 무관하게 사람들은 인공지능 비서를 편하게 이용하지만, 인공지능 비서는 전원이 차단되기 전까지 쉼 없이 인간의 언어를 데이터로 받아들인다. 이는 사생활 침해 등의 데이터 권리와 관련된 심각한 윤리 문제를 일으킬 수도 있다.

나. 뉴스피드 알고리즘 조작실험

2012년 페이스북과 코넬대학교의 연구팀이 실시한 뉴스피드 알고리즘 조작실험의 경우에는 "남들의 행복한 소식이 뉴스피드에 나타날수록 우리는 불행해진다"라는 일반적인 믿음이 사실인지 실험한 것이 알려져 논란이 제기되었다. 페이스북 데이터 과학 연구팀은 페이스북 뉴스피드의 알고리즘을 조작했고, 일주일간 페이스북 사용자 68만 9003명의 뉴스피드에서 특정 감정과 관련된 단어를 삭제하여 사용자들의 감정에 미치는 영향을 연구한 것이다. 연구 결과, 부정적 감정과 관련된 단어를 삭제할수록 뉴스 피드에 '긍정적 기분'과 관련된 단어의 비중이 높아졌다. 즉, 뉴스피드가 행복할수록 유저는 불행함을 느낄 것이라는 일반적 통념과는 반대로 뉴스피드가 긍정적이면 우리의 감정도 긍정적으로, 뉴스피드가 부정적이면 우리의 감정도 부정적으로 변할 수 있다는 것이다. 2014년 미국 국립과학원회보(PNAS) 3월호에 이 논문을 발표한 후 연구팀은 사과문을 게재하였다. 가입약관에 있는 포괄적 동의에도 불구하고 심각한 사회문제를 일으킨 것이다. 페이스북은 이후에도 일명 '페이스북-케임브리지 애널리티카 정보 유출 사건'을 통해 데이터의 정치적이고 경제적인 결정력이 얼마나 영향력 있는가를 역설적으로 환기했다. 인공지능의 자원으로써 데이터의 지위는 인공지능의 기술 사회적 권위와 비례적 관계에 놓이므로, 인공지능 기반 사회에 돌입한 지금 데이터 윤리를 더 이상 무시할 수 없다.

다. 구글의 데이터 편향성 인지 주장

구글 매니저인 화이트(Becky White)는 데이터 안에는 편향성이 존재한다고 보며 이러한 편향성을 제거할 수 있는 해결책이 복합적이라고 주장했다. 또한 편향성 인지가 매우 중요한 출발점이 된다고 말한다. 그녀는 인공지능 편향성 중에서 선택편향, 확증편향, 자동편향을 강조한다.⁴⁾ 먼저 '선택 편향'의 대표적인 예로는 지리적 편향의 예를 들면서 "북미에서 데이터를 생성하고 라벨링 후 머신러닝한 AI는 북미 지역에 대한 편향이 발생한다고 지적한다. 두 번째는 '확증 편향'이다. 이는 사람들이 자신이 기존에 믿는 바에 부합하는 정보만 받아들이려고 하고, 자기 생각에 어긋나는 정보는 거부하는 편향을 뜻한다. 대개 데이터 수집 과정에서 조사자(리서처)는 무의식적으로 자신의 믿음과 일치하는 방향으로 수집하는데, 이는 데이터 처리 과정에도 영향을 미친다. 이 경우 조사자 입장에서 겉보기에 편향이 보이지 않기 때문에 더욱 문제다. 끝으로, '자동 편향'이다. 머신러닝은 자동 처리 데이터를 비자동 처리 데이터보다 선호한다. 이는 곧 데이터 배제로 이어지고, 결국 편향성이 강화되는 꼴이다.

4) 석대건, "AI 편향이란 무엇인가?...구글의 AI 원칙 '데이터 왜곡 없어야' 디지털투데이, 2019년 6월 26일, <http://www.digitaltoday.co.kr/news/articleView.html?idxno=211754>

데이터를 모으고 처리하는 과정에서 늘 편향성의 문제가 제기된다.⁵⁾ 물론 데이터 자체의 편향성과 데이터 처리 과정의 편향성은 구분해야겠지만, 이와 별도로 이 두 가지 모두 공통으로 데이터 공정성의 문제와 충돌하기 마련이다. 특히 데이터 객관성과 공정성 역시 상충할 소지가 충분히 있다. 데이터의 객관성은 데이터가 지향하는 대상과 데이터를 일치하는 것을 전제로 한다. 이는 데이터를 습득하는 과정에서 주관적 개입을 배제해야 확보할 수 있다. 이러한 근대적인 기계적 객관성이 보장되더라도 공정하지 않을 수 있는 문제가 발생한다. 데이터의 객관성의 경우 데이터 자체의 습득 과정에서 논의하겠지만, 데이터의 공정성의 경우 데이터의 활용 과정에서 문제를 제기하기 마련이다.

2-2. 데이터 윤리의 정의

데이터 개념 자체는 디지털 사회를 구축할 때부터 지금까지 지속적으로 발전했다. 이에 맞춰 데이터에 관한 윤리적 쟁점들도 드러났다. 데이터 윤리에 관한 개념 또한 수십 년을 지나며 데이터 개념과 함께 형성되었던 셈이다. 다만 인공지능 시대에 들어서면서 인공지능 기술의 특수성 때문에 인공지능의 학습을 위해 사용되는 데이터 관련 윤리 문제가 새로이 발생하였으니, 이를 반영하지 않을 수 없는 상황을 맞았다. 따라서 먼저 데이터 윤리 일반에 대한 개념을 정립해 보고, 이를 기반으로 인공지능용 학습 데이터 윤리에 대하여 논의하고자 한다.

데이터 윤리에 대한 프레임워크, 모델, 사고 리더십 및 사례 연구 등 3억 1,500만 개의 '데이터 윤리 정의' 검색 결과가 있음에도 아직 데이터 윤리에 대하여 전적으로 합의된 논의는 없는 상태로 보인다.⁶⁾ 우선 다양한 자료에서 데이터 윤리의 개념을 정의하는 데 주로 인용되는 것은 플로리다와 타데오의 논의다.⁷⁾ 그들에게 데이터 윤리란 데이터의 도덕적 문제를 연구하고 평가하는 윤리의 한 분야다. 이때 데이터 윤리는 도덕적으로 선한 해결책을 공식화하고 지원하기 위한 것이며, 여기서 데이터란 '데이터, 알고리즘 및 유사 관행'과 연관된 데이터를 의미한다. 즉 데이터는 사용 목적 자체에서 먼저 윤리성으로 출발하여야 하며, 데이터 활용 전 과정에서 도덕적 문제가 관리되어야만 한다는 것이다.

또한 노박과 파블리ček(Novak & Pavlíček)은 데이터 윤리의 특수성으로 나아가기 위해 먼저 일반적인 데이터 윤리에 대해 다음과 같이 다룬다.⁸⁾ 여기서 그는 관련된 응용 윤리로 디지털 윤리(digital ethics), 사이버 윤리(cyber ethics), 컴퓨터 윤리(computer ethics), 정보 윤리(information ethics) 등을 데이터 윤리와 함께 제시한다. 이때 디지털 윤리, 사이버 윤리 등과 데이터 윤리엔 차이가 있다면 데이터 윤리가 철학적인 접근 방식으로 논의된다고 설명한다. 유럽 최대의 방산 업체인 BAE Systems에서도 '데이터 윤리가 어떠한 가치를 추구하는가' 하는 철학적 질문과 연

5) 변순용, "데이터 윤리에서 인공지능 편향성 문제에 대한 연구", 윤리연구 128호 (2020), pp.143-158.

6) BAE Systems, An Introduction to Data Ethics, (2021).

7) L. Floridi, M. Taddeo, "What is data ethics?. Philosophical Transactions of the royal society", A mathematical, physical, and engineering science Vol.373 No.2083, 2016, pp. 334-337.

8) R. Novák, A. Pavlíček, "Big Data Ethics and Specific Differences from General Data Ethics". IDIMT 2020: Digitalized Economy, Society and Information Management—28th Interdisciplinary Information Management Talks 28, 2020, pp.223-230.

결해야 한다고 논했다.⁹⁾ 또한 그들은 데이터 윤리가 ‘데이터로 무엇을 할 수 있는가(What you could do with data)’라는 식의 선택 사항이 아니라면서, ‘데이터로 무엇을 해야 하는가(What you should do with data)’의 의무 사항으로 전환해야 한다고 제안했다. 즉 데이터 윤리는 단순히 행위 규범을 제시하려는 것을 넘어, 데이터 활용의 전부터 후까지를 모두 아우르며 그 철학적 기반을 다지는 것과 연관된다.

한편 데이터 윤리는 주체에 따라 정의가 달라질 수 있다. 기업의 관점에서는 언제나 데이터의 윤리적 문제가 발생해 이미지에 타격을 입을 수 있다. 그들에게 데이터 윤리는 기업 내 데이터 자체 또는 데이터 관리자에 대한 신뢰성을 높이는 수단이다. 그러나 주체를 사회 일반으로 확장한다면 데이터 윤리는 인권을 보장하고 프라이버시를 보호하는 등 사회의 공공선을 증진하는 방안이 된다. 덴마크 수출보증재단(EFK)은 자체 데이터 윤리 정책에서 데이터 윤리를 다음과 같이 정의한다.¹⁰⁾

‘데이터 윤리라면 시민과 기업의 권리를 넘어, 사회적 가치를 담을 수 있는 기술의 윤리적 차원을 다룰 수 있어야 한다.’

또한 트랜버그 외는 데이터 윤리에 관하여 ‘책임감을 지니고 지속가능하게 데이터를 사용하게 해주는 기준’으로 설명하며, 궁극적으로 데이터 윤리를 준수하는 행위는 사람과 사회에 대해 선한 일을 하는 것이라 보았다.¹¹⁾ 결론적으로 데이터 윤리의 경우 사회를 이롭게 할 수 있도록 정의해야 할 것이다.

데이터 윤리의 핵심 개념으로는 데이터 권리와 데이터 책임을 거론할 수 있다. 즉, 데이터 윤리는 데이터에 대한 각 주체의 권리를 보호하기 위하여 데이터의 전 과정에 걸쳐 책임 의무를 성실히 이행하는 것으로 요약된다. 덴마크 데이터 윤리에 관한 전문가 그룹(the Danish Expert Group on Data Ethics)은 데이터 윤리를 “개인이나 집단의 정당한 이익(the Legitimate Interests of an Individual or Group)”으로 보며, 권리에 반하여 사용되지 않도록 하는 능동적인 결정과 행동으로 정의했다.¹²⁾ 이는 데이터의 윤리적 문제가 데이터 권리를 침해하는 형태로 드러나는 것을 의미한다. 또한 그들은 데이터 권리와 충돌하는 윤리적 문제를 해결하기 위해 적극적으로 책임을 다해야 한다고도 설명했다. 여기서 책임이란 일반 데이터보호규정(GDPR, General Data Protection Regulation)과 같이 명시된 법률을 넘어서는 것이라 하였다. 바꿔 말하면 데이터 책임은 보통의 규범 준수 이상의 것이어야 하며, 과정에 대한 책임과 결과에 대한 책임을 모두 의미하는 것으로 보인다.

데이터 권리와 데이터 책임에 대한 논의는 ‘공공 및 민간 분야의 데이터를 어떻게 긴밀히 활용해야 하는가’라는 질문과 대단히 밀접하게 맞닿아 있다. 이는 최근 국제적으로 강조되는 방향성이다. 근래 들어 데이터를 몰샐트

없이 묶어두기보다, 안전하게 공유하여 기술을 발전시키거나 사회의 공공선을 증진할 때 데이터를 적극적으로 활용한다. 이러한 흐름을 반영하는 개념이 있는데, 데이터 분석 툴 제작사인 펜타호(Pentaho)의 CTO였던 제임스 딕슨(James Dixon)이 처음 제안하였던 ‘데이터 레이크(Data Lake)’나 한국 문재인 정부에서 고안한 ‘데이터 댐’이 그 예다. 데이터 레이크는 가공되지 않은 날것 그대로인 데이터를 정형 데이터와 비정형 데이터 상관없이 자연적으로 통합되도록 하는 저장소다. 이는 인공적으로 데이터셋을 구축하는 사업인 데이터 댐과는 일부 차이가 있다. 물론 데이터 레이크와 데이터 댐이라는 개념 모두 공통으로, 축적된 무수한 공공 데이터를 정부 주도하에 민간에 공급할 수 있다는 점을 반영한다.

또한 데이터의 개방성과 관련된 내용은 유네스코(UNESCO)의 인공지능 윤리 권고안에도 명시되어 있다(UNESCO, 2022). 권고안에 정책 활동 영역이 여럿 제시되어 있으며, 정책 3영역의 ‘데이터 정책(Data Policy)’에는 데이터에 관한 구체적인 안내가 되어 있다. 여기서 유네스코는 정확히 ‘개방형 데이터를 장려’하도록 제안했고, 이때 개방형 데이터의 경우 데이터 거버넌스 전략 및 메커니즘 속에서 통합적으로 관리하도록 제시했다. 그러면서 데이터 권리와 관련하여 다음과 같이 언급하고 있는데, 사생활 및 개인정보에 대한 권리, 데이터 주체의 개인 데이터 보유·접근·삭제·제어 권리, 데이터 활용에 관한 동의권 등이 바로 그것이다. 또한 데이터가 개방성을 지니려면 데이터 권리를 보호해야 할 책임도 중요해진다. 그에 걸맞은 데이터 책임도 강조해야 하는 것이다. 데이터 수집 및 선택 프로세스의 적절성, 적절한 데이터 보안 및 보호 조치, 문제 상황으로부터 피드백 메커니즘을 통한 데이터 품질의 지속적인 평가 보장, 개인정보 보호 영향평가와 같은 윤리적 영향평가, 데이터 책임을 위한 정책 및 프레임워크 수립 등이 강조된 데이터 책임의 구체적인 예이다.

선행된 연구와 여러 조직의 데이터 윤리에 관한 문헌¹³⁾을 분석한 결과 데이터 윤리는 최종적으로 다음과 같이 개념을 정립해볼 수 있다. 데이터 윤리란 ‘개발자, 관리자, 사용자가 데이터를 사회의 공공선을 위해 활용할 때 데이터의 각 주체의 권리를 보호하고자 데이터 수집부터 폐기까지 전 과정에서 책임의 기준을 제공하는 가치 체계이자, 윤리적 문제가 발생했을 때 그 결과에 대하여 책임을 다하려는 응용 윤리의 한 분야’이다.

3. 데이터 윤리 이슈의 사례 분석

가. 독일 울름시 – 데이터 처리 원칙

독일의 울름(Ulm)시는 도시의 디지털화를 대비하며 스마트 시티 전략을 세우고 4가지 주요 미래 과제를 제시하고 있다. 과제 중 한 가지가 ‘데이터 처리(Dealing with Data)’이다. 시의회는 2020년 10월 8일에 이와 관련하여 도시 행정에서 연결되는 다양한 도시 데이터를 처리하기 위한 원칙 수준의 ‘데이터 윤리 개념(Data Ethics Concept for the City of Ulm)’을 채택하였다.¹⁴⁾ 개념이라는 타이틀로 발표되었지만, 내용은 지침과 원칙의 차원에서 구성되었다. 특징

9) BAE Systems, An Introduction to Data Ethics, (2021).

10) EFK, Data Ethics Policy, (2021).

11) P. Tranberg, G. Hasselbalch, B. K. Olsen, C. S. Byrne, Dataethics: Principles and Guidelines for Companies, Authorities & Organisations, DataEthics.eu, (2018).

12) S. Rasmussen, Data for the Benefit of the People: Recommendations from the Danish Expert Group on Data Ethics, Økonomi- og Erhvervsministeriet, (2018).

13) 송선영, 김항인, “정보화시대의 빅데이터(Big Data) 활용에 대한 윤리적 논쟁과 전망”, 윤리연구 제1권 108호 (2016), pp.227-248.

14) Stadt Ulm, Data Ethics Concept for the City of Ulm, (2020).

은 다음과 같다. 첫째는 ‘프라이버시 보호(Securing Privacy)’로, EU의 일반 데이터보호규정(GDPR)의 의미 내에서 시민의 개인정보를 보호할 것을 요청하고 있다. 둘째는 ‘공공 데이터 또는 개방 데이터(Open data)’로, “기술 주권의 필수 요소(Necessary Element of Technological Sovereignty)”인 공공 데이터의 투명성 창출이 결과적으로 공공성을 보장할 것이라는 인식이 드러난다. 셋째는 ‘민주적 통제 확보(Securing Democratic Control)’로, 디지털 민주주의 측면에서 데이터 처리에 대한 시의회 및 위원회의 조언과 결정의 필요성을 강조하고 있다. 넷째는 ‘데이터, 알고리즘 및 자동화 시스템의 투명한 사용(Transparent Usage of Data, Algorithms and Automated Systems)’으로, 자동화된 행정적 의사결정 기준을 공개하는 등 데이터 투명성을 논하고 있다. 이외에도 안정성, 지속가능성, 책임성 등에 대한 개념이자 지침들이 제시되었다. 다만 데이터 전반에 대한 윤리보다는, 스마트 도시 구현을 위한 도시 행정 데이터 위주의 윤리 사항이라는 한계가 있었다.

나. 사피스 - 데이터 윤리 프레임워크

또한 사피스 및 기타 집필진은 건강과 연구 분야의 데이터로 발생하는 윤리적 문제를 다룰 프레임워크를 제안하고 있다.¹⁵⁾ 여기서 16가지 윤리적 가치를 크게 두 가지로 구분하였다. 의사결정의 결과로부터 고려될 수 있는 ‘실질적인 가치(Substantive Values)’와 함께, 의사결정 과정으로부터 고려될 수 있는 ‘절차적인 가치(Procedural Values)’가 그것이다. 실질적인 가치로는 ‘피해 최소화(Harm Minimisation)’, ‘진실성(Integrity)’, ‘정의(Justice)’, ‘자유/자율(Liberty/Autonomy)’, ‘프라이버시(Privacy)’, ‘비례(Proportionality)’, ‘공익(Public Benefit)’, ‘연대(Solidarity)’, ‘책무(Stewardship)’의 9가지 핵심 가치가 속한다. 또한 절차적인 가치로는 ‘책임성(Accountability)’, ‘일관성(Consistency)’, ‘참여(Engagement)’, ‘합리성(Reasonableness)’, ‘성찰(Reflexivity)’, ‘투명성(Transparency)’, ‘신뢰(Trustworthiness)’의 7가지 핵심 가치가 있다. 비록 건강과 연구 분야에 한정되지만 제시된 데이터 윤리의 핵심 가치들 덕분에 데이터 윤리 원칙이 더욱 정교해질 수 있다.

다. 독일 연방 정부 - 데이터 윤리 위원회(datenethikkommission)

독일 연방 정부에서 설립한 데이터 윤리 위원회(Datenethikkommission)는 2019년 데이터 윤리에 관한 의견을 발표했다(Data Ethics Commission, 2019).¹⁶⁾ 본 의견서에는 윤리적이고 법적으로 가장 기본적인 원칙이 ‘인간존엄성’, ‘자기 결정’, ‘프라이버시’, ‘안전’, ‘민주주의’, ‘정의와 연대’, ‘지속가능성’으로 거론되었다. 이러한 기저 위에 두 가지 주제의 지침을 마련하였는데, 논의 초기에는 ‘알고리즘 기반 의사결정, 인공지능, 데이터’의 3주제로 시작하였으나 인공지능을 알고리즘의 변형이라 판단하며 ‘데이터’와 ‘알고리즘’의 2가지 주제를 골조로 갖추게 되었다.

‘데이터’와 연관된 위원회의 권장 사항은 개인 데이터와 비개인 데이터 간의 차이를 중심으로 하여 다음의 일반 표준을 따라 만들어졌다. 즉 첫째, ‘예견되는 책임’ 기준에 따라 데이터 처리 과정의 잠재적 영향력을 고려해야 한다. 둘째, ‘당사자 권리 존중’ 사항에 따라 데이터 생성에 일부라도 관여한 이의 권리를 존중해야 한다. 셋째, ‘공익을

위한 데이터 사용 및 공유’ 사항에 따라 공공의 이익을 증진할 수 있어야 한다. 넷째, ‘목적에 맞는 데이터 품질’ 권장에 따라 단순한 적합성을 넘어서는 높은 수준의 품질로 보장해야 한다. 다섯째, ‘위험 적정 수준의 정보 보안’ 표준에 따라 내재한 위험 가능성을 해소해야 한다. 여섯째, ‘이해(Interest) 중심의 투명성’이라는 방향성에 따라 책임을 다할 수 있도록 데이터 관련 활동에 대한 설명이 가능해야 한다.

마찬가지로 ‘알고리즘’은 다음의 일반 표준을 따라 만들어졌다. 첫째, ‘인간 중심 설계’를 통해 인간성(Humanity)을 알고리즘의 설계 과정에서 확보하도록 한다. 둘째, ‘핵심 사회적 가치와의 호환성’을 통해 사회적 가치와 영향에 대한 고민이 필요하다. 셋째, ‘지속 가능성’과 넷째, ‘품질 및 성능’을 통해 알고리즘의 정확하고 안전한 작동을 실현한다. 다섯째, ‘견고성 및 보안’을 통해 외부 위협에 대한 보호와 시스템이 만들어 낼 부정적인 영향에 대한 보호를 포함한다. 여섯째, ‘편견과 차별의 최소화’를 통해 알고리즘 패턴에 따른 편향과 차별을 주의해야 한다. 일곱째, ‘투명하고, 설명가능하며, 이해가능한 시스템’을 통해 사용자와 더불어 영향 당사자가 권리를 행사할 수 있도록 작동 원리 등의 충분한 정보를 제공받을 수 있어야 한다. 여덟째, ‘명확한 책임 구조’를 통해 시스템 작동과 관련된 모든 책임이 정확하게 구분되어야 한다.

라. 네덜란드 - DEDA

마지막으로 네덜란드의 AI 및 데이터 프로젝트의 윤리적 검토를 위한 대화 프레임워크로 개발된 Data Ethics Decision Aid(DEDA)는 프로젝트의 데이터 윤리 사항 관련 점검 항목을 대화 형식으로 구성하였고, 이를 따라갈 수 있도록 나선형 구조로 제시하였다는 특징이 있다.¹⁷⁾ DEDA는 4가지 주요 단계로 구성되고, 1단계는 도입 단계로 본론 격인 질문에 들어가기 위한 개요에 해당한다. 2단계는 질문 단계로, 실제적인 DEDA의 질문이 제공되며 ‘데이터 고려사항(Data-related Consideration)’과 ‘일반 고려사항(General Consideration)’의 두 범주로 구분된다. 데이터 고려사항의 질문 세트에는 데이터, 데이터 품질, 데이터 소스 등의 6가지 클러스터에 각각 2~5가지 구체적인 질문이 나열되어 있다. 일반 고려사항의 질문 세트에는 책임, 사회적 영향, 편견 등의 6가지 클러스터에 각각 2~5가지 구체적인 질문이 나열되어 있다. 3단계는 가치 단계로, 클러스터마다 어떠한 윤리적 가치의 결과를 보이는지 확인하게 된다. 4단계는 결론 단계로, 나머지 단계를 거치며 확인한 내용을 전반적으로 검토하게 된다.

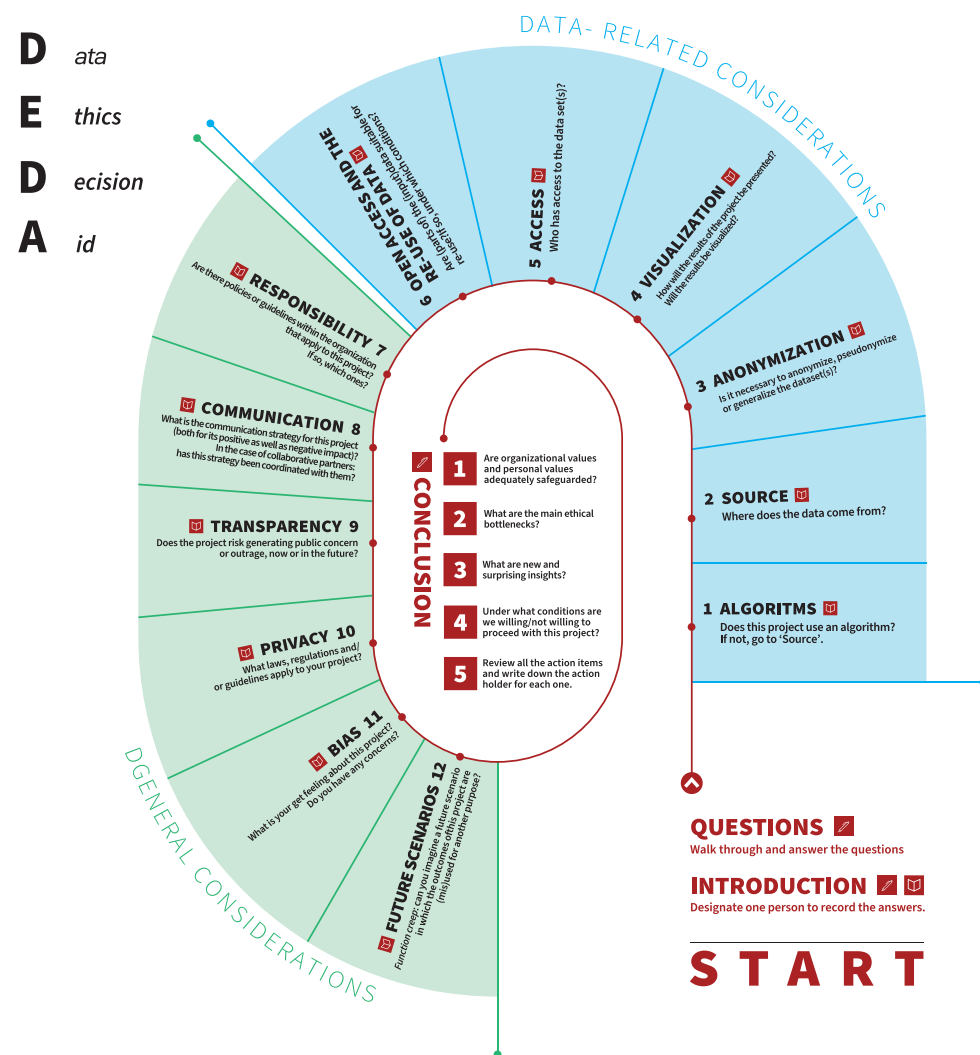
다음 <그림 1-3-1>은 DEDA 프레임워크의 최종 버전을 나타낸 그림이다.

¹⁵⁾ V. Xafis, G. O. Schaefer, M. K. Labude, I. Brassington, A. Ballantyne, H. Y. Lim, W. Lipworth, T. Lysaght, C. Stewart, S. Sun, G. T. Laurie, E. S. Tai, “An Ethics Framework for Big Data in Health and Research”, Asian Bioethics Review, 2019, pp.227-254.

¹⁶⁾ Data Ethics Commission, Opinion of the Data Ethics Commission, (2019).

¹⁷⁾ A. S. Franzke, I. Muis, M. T. Schafer, “Data Ethics Decision Aid(DEDA): a dialogical framework for ethical inquiry of AI and data projects in the Netherlands”, Ethics and Information Technology, 2021, pp.551-567.

[그림 1-3-1] 데이터 윤리 의사결정 자원(DEDA) 프레임워크

* 출처: <https://deda.dataschool.nl/en/deda-remote/>

4. 데이터 윤리: 데이터 권리와 데이터 책임

데이터 윤리를 정의하면서 핵심적으로 사용되는 개념으로 '데이터 권리'와 '데이터 책임'을 들었다. 첫째 '데이터 권리'는 '데이터가 어디로부터 생산되었는가' 하는 사안과 직접적으로 연관된다. 때로는 데이터의 출처가 지워질 수도 있겠지만, 그럼에도 데이터 권리는 유효하다. 인공지능과 빅데이터의 발전 덕분에 데이터의 공유와 개방만으로 사회의 공공성을 보장할 수 있기 때문이다. 그런 측면에서 보면 데이터의 권리는 개인 차원을 넘어, 공공의 권리로 환원되기도 한다. 그러므로 데이터 권리의 균형점을 찾아가는 과정은 데이터 윤리에서 대단히 주요한 과제다.

둘째 '데이터 책임'은 '데이터로 발생하는 윤리적 쟁점에 어떻게 대응할 것인가' 하는 사안과 연관된다. 결국 데이터 책임은 또 하나의 필수 사항으로, 윤리적 선택의 선명한 기준을 세워주는 개념이다. 그것을 통해 데이터 주체들은 여타의 데이터 윤리의 핵심 가치를 충분히 고려하고, 궁극적인 목표를 이룰 수 있다.

데이터 윤리로서 두 개념은 앞으로 더욱 중요해질 것이다. 기술의 발전 속도가 급격해지면서 기술의 지위는 계속 상승하기 때문이다. 인류의 기술 중심 사회가 고도화하면서 인간이 기술을 관리하는 것인지, 기술이 인간을 이끌어가는 것인지 갈수록 경계가 모호해진다. 이는 인권 문제로 연결된다.

인권은 세계대전의 역사로부터 마땅하게 존중받아야 할 권리로 평가받지만, 결국 인간이 지키는 만큼만 지켜진다. 스토아주의나 칸트의 인간론에서 비롯된 보편성으로 강조되더라도, 실제로는 상황에 따라 유동적인 셈이다. 그렇기에 매 순간 관심과 노력이 필요하다. 인간의 윤리로부터 발생한 인간의 존엄성을 지키겠다는 사회적 결의로 연대해야만 한다. 사실 그런 꾸준한 관심과 노력이 있었기에, 인권의 명맥이 여태까지 이어진 것이라고 볼 수 있다. 즉, 인권은 인간이 원하는 만큼만 지켜진다. 기술 분야에서도 마찬가지다. 기술로부터 윤리적 위기 상황이 연출되지 않도록 하는 노력도 결국 윤리로부터 시작해야 한다.

'정보혁명 시대'라는 타이틀은 결국 정보, 즉 데이터가 사회에서 중요한 역할을 담당한다는 사실을 단적으로 드러낸다. 데이터는 모든 소프트웨어를 운용하기 위한 이룰테면, 식수 자원이다. 이런 소중한 자원을 녹슨 파이프로 공급해서는 안 된다. 이를 방지하기 위해 녹 방지제와 같은 데이터 윤리가 필수적이다. 특히나 인공지능 시대에 데이터는 식수 그 이상일 수도 있다. 인공지능의 원리 자체가 데이터로부터 익히는 기계학습이고, 학습 속도가 압도적이므로 앞으로 어떤 변화를 이끌지 기대할 만하다. 동시에 데이터 윤리를 모색하는 작업은 데이터 시장이 활성화되었던 예전보다 더욱 중요해졌다.

2^{PART}

데이터산업 주요 정책 및 법제도 현황

제1장 • 국내 데이터 관련 정책 및 법제도 현황

제2장 • 해외 데이터 관련 정책 및 법제도 현황

제3장 • 데이터 활용 이슈의 법제도적 측면
: 저작권/개인정보

제1장 국내 데이터 관련 정책 및 법제도 현황

윤상필 연구교수 고려대학교 정보보호대학원

최근 디지털플랫폼정부를 중심으로 하는 데이터 중심 디지털 혁신 방안이 구체화되고 있다. 데이터 가치평가 기관이 지정되고 범분야 마이데이터 추진 근거도 마련되는 등 데이터 산업의 지형이 크게 변화하고 있다. 이에 따라 이 장에서는 데이터 정책 및 산업에 관한 국가정책과 ① 데이터 산업진흥, ② 데이터 활용 기반, ③ 데이터 안전을 기준으로 유형별 데이터 법제를 개괄하고 간략한 발전 방안을 소개한다.

1. 데이터 혁신과 국가정책

가. 대한민국 디지털 전략(2022.09.)

2022년 9월 정부는 “국민과 함께 세계 모범이 되는 디지털 대한민국”을 비전으로 내걸고 ‘대한민국 디지털 전략’을 발표하였다. 동 전략은 ① 세계 최고의 디지털 역량, ② 확장하는 디지털 경제, ③ 포용하는 디지털 사회, ④ 함께하는 디지털플랫폼정부, ⑤ 혁신하는 디지털 문화의 5개 과제를 수립하였는데 그 세부 과제는 다음과 같다.

[표 2-1-1] 대한민국 디지털 전략 세부과제

전략	추진과제
세계 최고의 디지털 역량	① 기술 패권에 대응한 6대 디지털 혁신 기술 확보 ② 충분한 디지털 자원 확보 ③ 더 빠르고 더 안전한 네트워크 구축 ④ 100만 인재 양성으로 디지털 인재 부국(富國) 달성 ⑤ 경계를 뛰어넘는 디지털 플랫폼 산업 육성 ⑥ 글로벌 시장을 주도하는 K-디지털 실현
확장하는 디지털 경제	① 서비스업 경쟁력 강화 ② 미래형 제조업으로 선진화 ③ 농·축·수산업의 신성장동력화
포용하는 디지털 사회	① 더 안전하고 쾌적한 삶의 터전 조성 ② 국민 누구나 디지털 혜택 보장 ③ 디지털로 재탄생하는 지역사회 구현
함께하는 디지털플랫폼 정부	① 혁신 인프라 구현 및 데이터 전면 개방·활용 촉진 ② AI·데이터 기술 기반 정부의 일하는 방식 혁신 ③ 안전하고 신뢰할 수 있는 이용환경 보장
혁신하는 디지털 문화	① 민간이 주도하는 디지털 혁신문화 정착 ② 혁신을 저해하는 규제 혁신 및 갈등 조정 ③ 디지털 경제·사회 기본법제 마련 ④ 디지털 혁신의 글로벌 확산

데이터 역량 강화를 위해 정부는 AI, 메타버스, 사이버보안 등 6대 혁신 기술에 집중 투자하면서 AI, Data, 클라우드 컴퓨팅, 소프트웨어와 같은 디지털 자원을 충분히 확보할 예정이다. 또한, 양질의 데이터를 공급하고 데이터의 객관적 가치평가를 위한 평가기법과 체계도 확립하며 공공 및 민간 데이터 플랫폼 연계, 마이데이터 확산 등도 추진한다.

[표 2-1-2] 대한민국 디지털 전략 내 데이터 관련 과제

<p>개방: 양질의 데이터 공급과 활용성 증진에 중점을 둔 공급체계 구축</p> <ul style="list-style-type: none"> 전 분야 AI 학습용 데이터 구축 확대: 교통, 물류 등 6개 분야에서 제조, 문화, 관광 등 다양한 분야로 확대 기업의 당면문제 해결을 지원하는 데이터 문제해결은행 구축 추진: 데이터, AI 기반 성공 및 실패 사례 등 축적, 제공하여 중소기업의 유사 당면문제 해결 지원 AI 개발자의 수요를 반영하여 인터넷 공개데이터 수집 및 제공: 빅데이터 활용 편의성 확대를 위한 저작권법 관련 규정 개정 검토
<p>거래: 데이터가 가치를 인정받고 유통되는 시대 본격화</p> <ul style="list-style-type: none"> (가치평가) 데이터의 객관적 가치평가를 위한 평가기법 및 체계 확립: 데이터 가치평가 제도 시행 (품질인증) 데이터의 품질향상을 위한 품질 대상 및 기준 등 확립: 데이터 품질인증 제도 시행 (과금·보상) 데이터 전송 과금 및 보상 체계 마련 (자산 보호) 데이터 자산 보호 및 부정사용 방지 (유통질서) 데이터의 합리적 유통 및 공정한 거래 질서 확립: 데이터 분쟁조정위원회, 표준계약서 등 제도 시행 (데이터 거래사) 수요자와 공급자 간 데이터 거래를 알선할 전문가 양성: 데이터 거래사 등록제 및 관련 교육 등 시행
<p>연계: 공공·민간 데이터 대통합으로 규모의 경제 실현</p> <ul style="list-style-type: none"> 데이터 정책 컨트롤타워로서 국가데이터정책위원회 출범 공공 및 민간 데이터플랫폼을 연계한 데이터 산업 통합 지원기반 구축 국가 차원의 데이터 표준화 체계
<p>서비스: 전 분야로 마이데이터 확산</p> <ul style="list-style-type: none"> 기업 보유 자기데이터에 누구든지 공정하게 접근, 활용하는 마이데이터 제도 전 분야 확대: 개인정보보호법 개정 및 표준화 추진 통신, 고용 등 새로운 분야, 이중 데이터 융합 기반의 마이데이터 실증 확대

아울러 동 전략은 개별 산업 분야의 데이터 혁신 계획도 포함하고 있다. 보건 의료 분야에서는 범부처 바이오 데이터 통합 수집 및 공유를 위한 국가바이오데이터스테이션 구축, 100만 명 규모의 임상, 유전체, 공공 보건 의료 정보 구축 및 개방, 식의약 데이터 전면 개방 및 데이터 플랫폼 구축 등이 추진된다. 제조업 분야에서도 제조기업 간 산업 데이터를 공동 활용할 수 있도록 협업시스템을 구축하고 제조 데이터 거래소를 통해 제조 데이터 거래와 활용을 촉진한다. 농축수산업 분야에서는 데이터 기반 스마트농업 전환과 가축 방역시스템 고도화를 추진한다. 이를 위해 온실·축산 등 데이터 수집·제공 및 AI 서비스 개발을 위한 클라우드 기반 플랫폼 구축, 데이터 기반 가축전염병 발생 위험도평가 모델 개발, 전염병 전파경로 분석 및 시뮬레이션을 통한 방역시스템 고도화 등의 사업이 진행된다.

나. 제1차 데이터 산업 진흥 기본계획(2023.01.)

2023년 1월 정부는 제2차 국가데이터정책위원회를 통해 ‘제1차 데이터 산업 진흥 기본계획’을 공개했다. ‘대한민국 디지털 전략’의 데이터 및 인공지능 분야 후속 계획으로 마련된 동 계획은 세계 최고 데이터 강국 도약을 비전으로 삼고 ① 데이터 시장 규모 50조 원, ② 데이터 활용 역량 10위권 내 돌입, ③ 기업 데이터 도입률 30% 이상 달성의 3가지 목표를 세웠다. 아울러 데이터 생산·개방·공유, 데이터 유통·거래, 데이터 보호·활용 및 데이터 산

업 기초체력 강화의 4개 전략별 추진 과제는 다음과 같다.

[표 2-1-3] 제1차 데이터산업 진흥 기본계획의 추진과제

전략	추진과제
데이터 생산 · 개방 · 공유	<ul style="list-style-type: none"> · AI 등 신산업 창출을 지원하는 전략적 데이터 생산 및 공유 · 민간 협력 기반 혁신적 데이터 공유 및 활용 · 축적된 공공데이터를 수요자 맞춤형으로 개방
데이터 유통 · 거래	<ul style="list-style-type: none"> · 데이터 산업 생태계 지원 통합 국가 인프라 구축(One-윈도우) · 새로운 데이터 유통 · 거래 제도 도입 및 안착 · 민간 · 공공 데이터 연계 활용을 위한 표준화 추진 · 데이터 유통 및 거래 민간 역량 강화
데이터 보호 · 활용	<ul style="list-style-type: none"> · 데이터 보호 및 활용 규제 혁신 · 정보주체의 데이터 주권 확대 · 가명정보 제도 활용 촉진 · 데이터 보호 및 활용 인프라 강화
데이터 산업 기초체력 강화	<ul style="list-style-type: none"> · 국민의 데이터 문해력(리터러시) 제고 · 산업 수요에 맞춘 데이터 인력 양성 · 기업의 데이터 활용 기반 디지털 전환 촉진 · 세계적 수준의 데이터 기업 육성 · 초일류 데이터 분석 · 활용 기술 경쟁력 확보

다. 디지털플랫폼정부 실행계획(2023.04.)

2023년 4월 정부는 ‘디지털플랫폼정부 실행계획’을 공개하였다. 동 계획은 인공지능과 데이터로 만드는 세계 최고의 디지털플랫폼정부를 비전으로 ① 하나의 정부, ② 똑똑한 나의 정부, ③ 민관이 함께 하는 성장플랫폼, ④ 신뢰하고 안심할 수 있는 디지털플랫폼정부 구현의 4개 전략과 전략별 추진 과제를 수립하였다.

[표 2-1-4] 디지털플랫폼정부 실행계획의 추진과제

전략	추진과제
하나의 정부	<ul style="list-style-type: none"> · 디지털을 기본으로 행정체계 전반 혁신(Digital by Design) · 데이터 칸막이의 근원적 해소 · 디지털플랫폼정부 혁신 인프라 구현
똑똑한 나의 정부	<ul style="list-style-type: none"> · 한 곳에서 한 번의 신청으로 끝나는 통합서비스 · 요구하지 않아도 알아서 챙겨주는 초개인화 서비스 · 국민 누구나 혜택을 누리는 환경 조성 · 인공지능 · 데이터 기반의 과학적 행정 일상화 · 투명하고 공정한 디지털 민주주의 실현
민관이 함께 하는 성장플랫폼	<ul style="list-style-type: none"> · 민관이 함께 사회문제를 발굴 · 해결하는 협업플랫폼 구축 · 민간의 공공 데이터 · 서비스 융합 · 활용 촉진 · 디지털 트윈을 통한 AI · 데이터 산업 쿼텀 점프 · GovTech 기업 성장 지원 강화 · 공공분야에 민간의 최신 AI기술 적극 활용

(계속 →)

신뢰하고 안심할 수 있는 디지털플랫폼정부 구현	<ul style="list-style-type: none"> · 개인정보에 대한 정보주체, 국민의 권리 강화 · 디지털플랫폼정부 안전을 보장하는 보안 체계 구축
------------------------------	--

데이터 산업에 관한 사항은 주로 민관 협업에 관한 전략에서 확인할 수 있다. 이에 따르면 정부는 온라인으로 도 데이터안심구역역을 확대하여 데이터 접근성을 제고하고 이중 데이터를 융합하여 활용할 수 있도록 추진한다. 민감도가 낮은 데이터를 잘 활용하면, 온라인에서도 원격 분석할 수 있는 환경을 마련할 수 있기 때문이다. 또한 국가 주요 인프라에 디지털 트윈 기술을 활용해 고품질 데이터를 실시간으로 확보하고 인공지능을 적용해 과학적 의사 결정을 구현하는 것도 용이해진다.

이런 이유로 마이데이터 전송인프라 구축을 추진해 개인정보를 손쉽게 유통, 활용할 수 있도록 하고 이중 분야 간 원활하고 안전한 데이터 이동을 위한 마이데이터 보안 및 표준화 가이드라인도 마련할 예정이다.

2. 데이터 법제도 현황

가. 데이터 산업진흥

1) 데이터 산업진흥 및 이용촉진에 관한 기본법

우리나라에서는 데이터 산업을 미래 성장의 핵심 동력으로 인식하고 2021년 10월 ‘데이터 산업진흥 및 이용촉진에 관한 기본법(약칭: 데이터산업법)’을 제정하여 2022년 4월부터 동법이 시행되었다.

데이터산업법에는 데이터 산업의 발전을 위한 여러 법적 근거가 담겨 있다. 우선 데이터 정책을 총괄하는 추진 체계로 국가데이터정책위원회를 설치하고 데이터의 생산, 활용 및 보호를 위해 데이터 결함 촉진 지원, 데이터 안심 구역 지정, 데이터 자산 보호 등에 관한 사항을 규정하였다.

둘째로 데이터 이용을 활성화하기 위해 데이터 가치를 평가할 수 있도록 하고, 데이터거래사업자와 데이터분석 제공사업자 등 데이터 사업 추진을 위한 법적 틀로서 데이터산업법이 그 역할을 할 수 있도록 했다.

셋째로, 데이터 유통과 거래 촉진을 위해서 유통 및 거래 체계를 구축하고 데이터 플랫폼 지원 사업을 추진할 수 있도록 했다.

넷째, 데이터 품질관리를 위해 데이터 품질인증 사업을 추진하고 공정한 거래를 위해 표준계약서를 마련하도록 했다.¹⁾

다섯째, 데이터 산업 기반을 조성하기 위해 데이터 기반 상품과 서비스 개발 지원, 데이터 기술 역량 교육 프로그램 실행, 데이터 산업 투자생태계 활성화, 전문 인력 교육, 고용연계, 데이터 호환성 확보, 데이터 사업 관련 세제 지원 및 보조금 지급 등에 관한 사항도 규정하고 있다.

1) 데이터의 내용, 구조 및 관리체계 등을 심사해 인증을 부여하는 데이터 품질인증제도가 도입되었으며 2023년 7월 데이터 품질인증기관 3곳이 지정되었다. (과학기술정보통신부, 데이터 품질인증기관 지정 공고, 과학기술정보통신부 공고 제2023-0693호 (2023))

여섯째로 데이터 생산, 거래 및 활용에 관한 분쟁을 조정하기 위해 데이터분쟁조정위원회²⁾도 두었다. 특히 그 간 데이터 산업 활성화의 핵심 조건 중 하나로 데이터 가치평가 기준이 요구되었는데, 2022년 7월 과학기술정보통신부는 ‘데이터 가치평가기관 지정 및 운영에 관한 지침(이하 가치평가 지침)’을 마련해 데이터 가치 평가기법과 가치평가기관 지정 및 운영 절차를 수립했다. 그 결과 2023년 3월 기술보증기금, ㈜나이스디앤비, 신용보증기금, 한국과학기술정보연구원이 데이터 가치평가기관으로 지정됐다.³⁾ 이에 따라 데이터 가치를 평가받고자 한다면 데이터 가치평가 신청서를 가치평가기관에 제출하면 된다. 이러한 신청이 있을 경우 가치평가기관은 가치평가 모델과 기법에 따라 데이터 가치를 평가할 수 있다.

2) 산업 디지털 전환 촉진법

산업 전반의 디지털 혁신을 지원하고 산업데이터의 활용을 확대하기 위해 2022년 1월 ‘산업 디지털 전환 촉진법’이 제정되었다. 이 법을 근거로 산업데이터를 생성하고 활용하는 활성화 과정을 거치고 지능정보기술을 산업에 적절히 적용한다면 산업의 디지털 전환을 촉진할 수 있다. 이를 통해 산업 경쟁력을 확보하고 국민 삶의 질 향상과 국가 경제발전에 이바지하는 것을 기대할 수 있다.

이 법의 핵심은 산업데이터의 사용·수익권에 있다. 이 용어는 ‘개인정보 보호법’ 등 기존의 권리보호 법령에서 규정하지 않은 산업데이터의 개념을 정의하고, 이에 관한 활용 및 보호 원칙을 제시하여 기업의 불확실성을 해소하고 산업데이터의 활용을 활성화하기 위해 도입되었다. 이와 관련해 법 제9조에서는 산업데이터를 생산한 자에게 해당 산업데이터를 활용해 사용하고 수익할 권리를 부여하고 약정에 따른 공동 권리, 제3자 제공 권리관계와 함께 공정한 상거래 관행과 경쟁 질서 유지, 이익의 합리적 배분, 산업데이터의 안전성과 무결성 보장 등에 관한 사항을 규정하고 있다. 이 외에도 동법은 산업데이터의 표준화(제12조), 산업데이터 품질관리 지원(제13조), 산업데이터 플랫폼(제14조), 산업 디지털 전환 선도사업의 선정 및 지원(제15조~제16조) 등에 관한 사항을 다룬다.

2023년 1월에는 제1차 산업 디지털 전환 위원회를 개최하고 ‘산업 인공지능 내재화 전략’을 심의 확정하였다. 동 전략은 산업데이터-X 플랫폼을 구축해 양질의 산업데이터를 다양한 주체들이 공유 거래할 수 있도록 하고, 노이즈가 최소화된 핵심 데이터셋을 선정해 보급하며, 산업데이터 품질 인증을 실시하는 등의 내용을 포함하고 있다. 아울러 산업 마이데이터 플랫폼 구축, 데이터 제공자에 대한 인센티브 제공 검토, 산업데이터 표준 개발, 산업데이터 사용 및 수익권자의 권리 보호를 위한 계약 가이드라인 마련 등도 언급하고 있다.

나. 데이터 기반 및 활용

1) 지능정보화 기본법

우리나라는 지능정보사회로의 전환에 대응하기 위해 2020년 ‘국가정보화 기본법’을 ‘지능정보화 기본법’으로

전부 개정하였다. 동법은 지능정보화 관련 정책을 수립하고 추진하기 위한 기반으로 기능한다. 구체적으로 지능정보기술의 개발, 표준화 및 전문 인력 양성, 지능정보서비스의 이용촉진, 관련 규제의 개선 및 선도사업 추진과 지원 등 서비스와 산업 활성화 정책을 담고 있다. 나아가 데이터의 기술적 활용 기반으로서 초연결지능정보통신망의 구축, 데이터센터 구축 및 운영, 데이터의 생산·수집·유통·활용·표준화·품질 제고 등을 촉진하기 위한 데이터 관련 시책의 마련, 민간 협력에 기반한 데이터 유통 및 활용 활성화, 데이터 통합지원센터의 설치 등에 관한 사항도 규정하고 있다. 이와 함께 정보 격차를 해소하고 양극화에 대응하며 사회적 영향평가를 할 수 있도록 하고 정보보호와 지능정보사회 윤리 등에 관한 내용도 규율하고 있다.

2) 정보통신망 이용촉진 및 정보보호 등에 관한 법률

‘정보통신망 이용촉진 및 정보보호 등에 관한 법률(약칭: 정보통신망법)’에서는 정보통신망과 인터넷 이용의 확산, 통신망 연계운영 및 표준화 등 정보의 공동 활용 체계를 구축하는 사항을 다루고 있다. 또 공공·지역·산업·생활 및 사회적 복지 등 각 분야의 정보통신망 이용촉진 관련 사업에 관해 규정하고 있으며, 인터넷 이용의 확산과 인터넷 서비스 품질 개선에 관한 사항을 규율하면서 데이터 활용의 기반에 관한 사항도 다루고 있다.

이와 함께 정보통신망의 안전성 확보를 위한 사전점검, 책임자 지정, 긴급 대응, 관리체계 인증 등 정보보호의무를 부여하고 정보통신망의 불법한 침해행위를 제재하고 있다.

3) 공공데이터의 제공 및 이용 활성화에 관한 법률

‘공공데이터의 제공 및 이용 활성화에 관한 법률(약칭: 공공데이터법)’에서는 공공기관이 보유한 양질의 데이터를 개방해 데이터 활용의 토대를 마련한다. 이 법을 근거로 하면 공공데이터 정책과 계획을 심의, 조정, 점검, 평가하기 위해 국무총리 소속으로 공공데이터전략위원회를 둘 수 있다.

무엇보다 동법에선 단순히 데이터를 개방하는 것에 그치지 않고 데이터 이용을 활성화하기 위해 필요한 데이터 품질 제고, 데이터 표준화, 데이터 재사용 지원, 데이터 제공 형태 정비, 데이터 제공 절차 및 분쟁조정에 관한 사항을 규율하고 있다. 아울러 민간 서비스와 중복되거나 유사한 서비스가 공공 부문에서 개발·제공되지 않도록 하고 공공데이터를 활용한 민간 협력과 창업 등을 지원할 수 있는 근거가 이 법에 담겨 있다.

2022년 11월 공공데이터전략위원회는 디지털플랫폼정부로의 패러다임 전환을 고려해 ‘공공데이터 혁신전략’과 ‘제4차(’23~’25) 공공데이터 제공 및 이용 활성화에 관한 기본계획’을 발표했다. 혁신전략을 통해 정부는 국민의 데이터 이용권 강화(일반 국민), 데이터를 통한 디지털 서비스 활성화(기업), 데이터 기반의 일하는 방식 혁신(정부)의 3개 중점 분야를 선정하고 10대 추진 과제를 수립했다.

²⁾ 데이터 산업법 제34조에 따른 데이터분쟁조정위원회는 2023년 10월 12일 출범하였다.

³⁾ 과학기술정보통신부, “데이터 가치평가기관 지정 공고”, 과학기술정보통신부 공고 제2023-0248호 (2023).

[표 2-1-5] 공공데이터 혁신전략의 10대 추진과제

중점분야	10대 추진과제
국민의 데이터 이용권 강화 (일반 국민)	• 비공개 데이터 개방 전환과 대체적 제공 방식 도입 • 일선 행정 · 공공기관 중심의 현장 데이터 제공 • 민관이 함께하는 공동 생성 데이터 구축 · 관리
데이터를 통한 디지털서비스 활성화 (기업)	• 서비스 창출에 도움이 되는 실시간 데이터 제공 • 다양한 서비스에 연계 · 활용될 수 있는 공공데이터 활용 체계 • 범정부 데이터 가공 · 결합을 위한 데이터 표준화, 품질관리
데이터 기반의 일하는 방식 혁신 (정부)	• 공공부문 데이터 공유 · 분석 환경 조성 및 활용 방식 고도화 • 온라인종합상황실 등을 통한 능동적 협치 구현 • 데이터 기반 행정을 위한 민간 기술 · 데이터 확보 • 데이터 활용의 기준이 되는 공공데이터 윤리 정립

제4차 공공데이터 기본계획은 디지털플랫폼정부 실현을 위한 모든 데이터의 개방과 연결을 비전으로 세우고 ① 국민을 위한 통합적 · 선제적 · 맞춤형 데이터 제공, ② 기업을 위한 데이터를 통한 새로운 기회 제공, ③ 데이터 기반의 과학적 행정 및 신뢰받는 사회 구현이라는 3개 목표를 세웠다. 해당 목표 달성을 위해 동 계획은 개방, 품질, 활용 기반의 4개 전략과 12대 추진 과제를 수립하였다.

[표 2-1-6] 제4차 공공데이터 기본계획의 전략별 과제

전략	추진과제
개방 네거티브 방식의 미개방 공공데이터 전면개방 체계 마련	• 네거티브 방식으로 미개방 데이터 전면 개방 추진 • 개방의 형태와 방식을 국민이 원하는 수요자 중심으로 전환 • 분야별 주요 공공데이터 개방 확대
품질 모든 데이터의 연결 및 융 · 복합 등 실질적 활용 제고를 위한 품질관리 및 표준적용 강화	• 고품질 데이터 기반 조성을 위한 품질인증제도 도입 및 공공데이터 표준 확대 • 범정부 공공데이터 표준 적용 지원 도구 제공 • 데이터 생애주기 전반의 표준 · 품질관리를 지원하는 「공공데이터 표준 · 품질 지원센터」 운영
활용 공공데이터 이용 활성화 및 민관협업을 통한 국정과제 실현 및 사회 현안 해결 지원	• 한 곳에서 공공데이터를 막힘없이 제공 · 활용하는 데이터 융합 · 분석 대국민 플랫폼 구현 • 공공데이터 활용 기업의 역량 및 성장 단계별 맞춤형 지원 확대 • 사회 현안 해결을 위한 지역 및 시민사회, 민간과의 파트너십 강화
기반	• 디지털플랫폼정부 구현을 위한 공공데이터법 개정 및 추진체계 강화
공공데이터 생산부터 활용까지 생태계 활성화를 위한 기반 강화	• 공공데이터 생성부터 개방과 활용을 촉진하기 위한 유관 제도 연계 및 개선 • 공공데이터 글로벌 선도 및 협력 강화 • 공공데이터를 이해하고 활용하는 공공 리터러시 수준 향상

4) 데이터기반행정 활성화에 관한 법률

‘공공데이터의 제공 및 이용 활성화에 관한 법률’이 공공부문에 축적된 데이터를 민간에서 활용해 데이터 경제를 활성화하는 데 중점을 두고 있다면 ‘데이터기반행정 활성화에 관한 법률’은 정부 내부의 데이터 공유와 활용을 통한 정부혁신에 초점을 두고 있다. 이를 위해 공공부문의 데이터 역량 강화를 핵심으로 하면서 민관협력 기반 데이

터 산업 발전의 내 · 외부 역량을 균형 있게 하는 데 기여한다.⁴⁾ 이와 관련해 동법에서는 공공기관이 보유한 데이터를 데이터통합관리 플랫폼에 등록할 수 있도록 하고 필요한 경우 데이터 제공을 요청하여 받을 수 있도록 했다. 또한 민간에서 생성하거나 취득, 관리하는 데이터가 필요한 경우 계약에 의한 구매나 업무협약 등을 통해 민간의 데이터를 제공받을 수 있는 규정도 마련했다. 데이터 역량을 강화하기 위해 데이터관리체계 구축, 데이터기반행정의 표준화, 데이터통합관리 플랫폼 구축, 데이터기반행정 책임관의 지정, 데이터분석센터의 설치, 전문기관의 지정 등에 관한 사항을 함께 규율하고 있다.

5) 국가공간정보기본법

‘국가공간정보기본법’의 목적은 국가공간정보체계를 효율적으로 구축하고 활용 · 관리하여 국토 및 자원을 합리적으로 이용해 국민경제의 발전에 이바지하는 데 있다. 2021년 3월 동법이 개정되어 공개가 제한되는 공간정보도 공간정보사업자나 위치정보사업자에게 제공할 수 있게 되었다. 본래 3차원 높이가 포함된 3D 공간지도나 고해상도 도로지도 등은 공개될 경우 공공의 안전을 해할 우려가 있어 공개가 제한되었고 학술연구나 공공복리의 목적으로만 제공되었다. 그러나 자율주행자동차나 디지털 트윈, 증강현실이나 가상현실과 같은 디지털 신산업 분야를 활성화하려면 공개가 제한된 고정밀 공간정보 또한 산업계에서 활용할 수 있어야 했다.

이에 따라 동법 제34조에 다음과 같은 사항을 규정하였는데, 우선 공간정보의 제공 및 제공제한에 관한 단서를 두면서도 공간정보사업자나 위치정보사업자가 사업 영위를 위해 제한된 공간정보 제공을 요청하는 경우 이를 제공할 수 있도록 하였다. 둘째로 필요한 절차와 보안관리 요구에 관한 사항 역시 명시하고 있다.

6) 국가지식정보 연계 및 활용 촉진에 관한 법률

2021년 5월 ‘국가지식정보 연계 및 활용 촉진에 관한 법률’이 제정되었다. 동법의 목적은 국민이 자유롭고 편리하게 국가지식정보를 이용할 수 있도록 국가지식정보의 연계 및 활용을 촉진하는 데 있다. 이를 위해 필요한 사항을 정함으로써 국민의 지식재산 창출 및 활용 역량을 제고하고 궁극적으로 국가경쟁력의 강화에 이바지하는 것 역시 주요한 목적이다. 이에 따라 지식 활용과 교육 목적으로 이용 가치가 있는 정보들은 이른바 ‘디지털 집현전’으로 불리는 통합플랫폼을 통해 활용할 수 있다. 이러한 정보에는 공공기관이 생산 · 보유 · 관리하고 있는 각종 과학 기술 · 교육학술 · 문화예술 · 사회경제 · 행정 등에 관한 내용이 있으며, 이 중에 지식 활용과 교육 목적에 부합하는 정보를 대상으로 한다.

이외에도 동법에선 지식정보의 활용을 촉진하기 위해 정보의 생산, 연계 등을 위한 표준화 추진, 민간과의 협력에 관한 사항을 규율하고 있다.

4) 권헌영, “데이터기반행정법과 데이터정책의 과제”, KISO Journal 제40호 (2020).

다. 데이터 보호 및 안전보장

1) 개인정보보호법

‘개인정보보호법’은 개인정보보호에 관한 일반법으로 기능하는 법이다. 이 법에도 2020년 큰 변화가 생기는데, 우선 데이터를 안전하게 활용하기 위해 추진된 데이터 3법 개정을 바탕으로 개인정보 감독 체계를 개인정보보호위원회로 일원화하고 가명정보 개념을 도입하였다. 또한 ‘정보통신망법’의 개인정보보호 내용을 가져와 특례로 규정하는 등의 변화가 있었다. 2023년 3월에는 ‘개인정보보호법’ 2차 개정을 통해 일반 분야의 마이데이터 사업추진 근거가 마련되었다. 전송요구권이 일반법인 ‘개인정보보호법’에 포함되면서 정보주체가 전송요구권을 전 분야에 행사할 수 있게 되었다. 이와 함께 자동화된 의사결정에 대한 거부, 이의제기, 설명요구권 등 정보주체의 권리도 강화되었다. 개인정보 국외 이전에 관한 사항이 문제가 있는 경우 개인정보보호위원회에게 국외 이전을 중지하도록 명령할 수 있는 권한도 부여되었다. 아울러 정보통신서비스제공자 등의 특례 규정이 삭제되어 개인정보처리자와 구분되어있던 체계가 일원화되었다. 나아가 개인정보처리자에 대한 형사처벌도 완화하고 과징금을 확대해 책임 규정을 합리적으로 정비하였다.

데이터와 관련된 가장 중요한 변화는 마이데이터다. 정부도 2023년 8월 ‘국가 마이데이터 혁신 추진 전략’을 발표했다. 이에 따르면 정부는 범부처 마이데이터 민간 협의회와 마이데이터 추진단을 구성하고, 마이데이터 서비스 중점 추진 분야를 선정하여 선도 서비스를 발굴하려고 한다. 또 마이데이터 플랫폼을 구축하고, 프라이버시 신고센터도 운영할 예정이다. 이를 위해 프라이버시 보호를 위한 데이터 엄정 관리 원칙을 마련해야 하고, 전송의무자의 단계적 확대 및 전송데이터의 단계적 혁신 과정이 필요하다. 마지막으로 생태계 활성화를 위해 참여자의 인센티브를 확산하는 방안을 마련하고, 연계산업으로 확산할 시스템을 구축하는 사업을 추진할 예정이다.

2) 신용정보보호법⁵⁾

‘신용정보보호법’은 개인신용정보에 우선 적용되는 법으로 신용정보 수집 및 처리의 원칙, 신용정보주체의 권리 보장을 위한 사업자의 의무 등을 규율한다. 특히 금융권 마이데이터 사업의 법적 근거로 기능하면서 제33조의 2를 통해 개인신용정보의 전송요구에 관한 사항을 다루고 있다.

비록 가명정보에 대해서는 정보주체의 전송요구권, 동의철회권, 열람 및 정정청구권 등을 적용하지 않지만, 가명 처리 및 익명 처리한 정보의 경우 부적절한 관리나 보안 위협으로 인해 식별성을 지니는 문제가 발생할 수 있으므로 이러한 상황을 방지하기 위해 필요한 조치를 취하도록 하고 있다.

3) 위치정보보호법

‘위치정보보호법’은 물건 또는 개인의 위치정보의 수집 · 저장 · 보호 및 이용 등에 관하여 특별히 우선 적용된다. 특히 특정 개인의 위치정보 또한 개인정보로서 동의를 받지 않고는 수집 · 처리하지 못하도록 하고 있으며, 관련 사업자에게 적절한 기술적 · 관리적 · 물리적 보호조치를 하도록 요구하고 있다. 사업자는 개인위치정보를 수집 ·

⁵⁾ 개인정보보호법 2차 개정과 함께 개인정보처리자는 공중위생 등 공공의 안전과 안녕을 위해 긴급히 필요한 경우에도 동의를 받지 않고 개인정보를 처리할 수 있게 됨. 신용정보보호법 제15조도 이를 반영하여 신용정보회사 등이 동일한 사유로 개인신용정보를 처리해야 하는 경우 신용정보주체의 동의를 받지 않고도 할 수 있도록 개정됨(‘23.3.)

이용 · 제공하려는 경우 개인위치정보주체의 동의를 받아야 하며, 개인위치정보주체는 언제든지 동의를 철회하거나 위치정보의 처리 중지를 요구할 수 있다. 이와 관련한 확인 자료를 열람할 권리가 있으며, 사업자에게 관련 자료의 고지를 요구할 수 있다.

3. 데이터 산업 법제의 발전 방향

가. 마이데이터 산업 기반 마련

개인정보보호법 개정을 통해 모든 분야에서 마이데이터 산업을 추진할 수 있는 근거가 마련되었다. 마이데이터는 자기정보통제권을 구체적으로 구현할 수 있는 개념이다.⁶⁾ 이는 프라이버시권의 관점에서 개인정보자기결정권을 적극적으로 실현하고자 하는 것으로, 정보통제권을 자기 영역 안에서 명확히 하려는 과정에서 성립되었다.

예전에는 정보주체가 최초 동의 이후 개인정보 처리에 개입하기 어려웠는데, 이제는 개인정보의 열람 · 정정 · 삭제 · 처리정지 · 전송요구와 같은 다양한 권리를 행사할 수 있는 법적 근거가 마련되었다. 또 이를 실질적으로 구현하기 위한 기술적 기반도 형성되고 있다. 이처럼 상세한 데이터 통제권을 정보주체에게 부여함으로써 데이터 처리자끼리 결정하고 책임지던 기존의 구조를 넘어, 이제는 정보주체의 허락을 받아 합법적으로 개인정보를 활용해 맞춤형 서비스를 제공해야 한다.

이러한 합리적 관행이 자리 잡을 때 데이터 산업 발전의 가능성이 확장된다고 할 수 있다. 물론 마이데이터 추진의 보편적인 근거는 마련되었지만, 아직 해결해야 사안은 많다. 예를 들어 실제 기술적 구현을 위한 작업이 진행될수록 예상치 못하게 이해관계자 간에 역할이 충돌하거나 책임 소재가 불분명해지는 문제가 불거질 수 있다. 이뿐 아니라 안심할 수 있는 보안 체계를 구축하여 신뢰를 확보하는 문제 등 쉽지 않은 과제들을 해소해 나갈 수 있어야 한다.

나. 합리적 규제체계의 모색

규제의 접근방식 역시 중요한 이슈다. 데이터 문제는 결국 해당 데이터를 처리하는 조직에서 가장 잘 알 수밖에 없다. 글로벌 경쟁 차원에서 산업 발전이 중대한 과제라는 점도 중요하지만 실제 규제를 논의하고 집행하는 과정에서는 데이터를 처리하는 조직이 데이터의 문제를 가장 잘 알 수 있다는 점도 고려해야 한다. 따라서 민간을 중심으로 하는 개방형 생태계를 활성화하고 구성원들의 자율규제가 이루어질 수 있도록 하는 접근이 요구된다.

다만, 자율규제라고 해서 공적 이해관계와 무관한 완전 자유방임을 의미하는 것은 아니다. 신뢰성이나 공공성의 확보, 자율규제 위반에 대한 실효적인 제재 수단이 모호해 공익을 해하는 결과로 이어질 수 있기 때문이다. 이에 따라 최근에는 정부가 자율규제를 위한 법적 근거와 원칙을 제시하고 민간이 자율규제를 집행하는 형태의 공동 규제 방안도 떠오르고 있다.⁷⁾ 공공성을 유지하면서도 민간의 혁신을 저해하지 않는 합리적 규제 방안을 모색하기 위한 범국가적 공론이 필요하다.

⁶⁾ 손형섭, “뉴노멀 시대에 데이터 이동권의 헌법적 수용에 관한 연구”, 유럽헌법연구 제37호 (2021), pp.253-260.

⁷⁾ 이강호, 이해원, “개인정보보호와 자율규제”, 법조 제69권 제6호 (2020), pp.34-35.

제2장 해외 데이터 관련 정책 및 법제도 현황

한국데이터산업진흥원 산업기획팀

데이터를 활용한 가치 창출이 전 세계적으로 더욱 활발해지면서 데이터는 이제 국제 경쟁력을 확보하는 데 없어서는 안 될 필수적인 경제적 자원이 되었다. 이에 주요국들은 데이터산업 관련 정책을 잇달아 발표하며 데이터 활용 및 공유 증대를 위해 적극적으로 노력하고 있다. 미국의 경우 지난 2020년 최고데이터책임자위원회(CDO Council, Chief Data Officers Council)를 설립하고 '연방데이터 전략(FDS, Federal Data Strategy)'을 기반으로 지속적인 데이터 산업 육성 정책을 시행하고 있다. 올해에 들어서는 이러한 노력을 국외로 확대하는 추세다. 그런가 하면 중국 정부는 2021년 발표한 '14차 5개년 빅데이터 산업발전 계획'을 기반으로, 정부 주도의 하향식 산업 육성책을 펼치고 있다. 이를 통해 2023년에는 특히 데이터산업 육성을 위해 시행 정책을 다각화하는 데 주력했다. 또한 유럽의회(EP)와 유럽이사회(Council of the European Union)는 올해, 유럽연합(EU) 통합의 '데이터법(안)(Data Act)' 잠정 합의에 이르렀다. 영국 역시 데이터법 개정안 재도입을 추진하였으며, 일본과 베트남은 사회 전반에서 디지털화 전환을 가속화하는 움직임이 두드러졌다. 이에 본 장에서는 미국, 중국, 유럽연합(EU), 영국, 일본, 베트남의 데이터 정책 수립 흐름 및 현황을 중심으로 해외 데이터 정책 현황에 대해 파악해 보도록 하겠다.¹⁾

1. 미국²⁾³⁾

미국: 데이터의 원활한 국외 이전 위한 정책 마련

미국 데이터산업 시장에서는 마이크로소프트(Microsoft), 구글(Google) 등 거대 IT 기업들뿐만 아니라, 벤처기업 등 많은 중소기업들까지도 글로벌 시장으로 사업을 확장하고 있다.

미국 정부는 올해 이러한 추세에 발맞추어 데이터의 국외 흐름과 이전이 손쉽게 이루어지게 하기 위한 정책 마련에 매우 적극적인 모습을 보였다. 일례로 미국은 지난 2020년, 최고데이터책임자위원회(CDO Council, Chief Data Officers Council)를 설립하고 '연방데이터 전략(FDS, Federal Data Strategy)'을 발표하였다. 이후 지속적으로 연방 정부 중심의 데이터산업 육성 정책을 펼치고 있다. 이러한 움직임이 올해에 들어서는 국외로 확장되고 있다.

그런가 하면 2023년 7월 10일, EU집행위원회는 'EU-미국 데이터 프라이버시 프레임워크(EU-US Data Privacy Framework)'에 대해 최종 적정성 결정을 채택했다. 이는 EU 회원국들과 미국 간의 개인정보를 자유롭게 이전하는

1) 이하 주요 국가의 데이터 정책 및 법제도에 대한 세부 내용은 한국데이터산업진흥원에서 발행된 월간 데이터이코노미 2023년 1~8호, 2022 데이터산업 백서 내용을 참조 및 인용하였다.

2) "데이터이코노미", 한국데이터산업진흥원 4월호(2023): 미국, EU 편.

3) "데이터이코노미", 한국데이터산업진흥원 7월호(2023): 미국, EU 편.

것을 허용하는 내용을 골자로 한다. 이로 인해 프레임워크에 참여하는 미국 기업들은 이제 EU에서 미국으로 개인정보 데이터를 이전할 수 있게 되었다. 기존의 EU-미국 간 개인정보 이전은 '프라이버시 실드'⁴⁾라는 미국 제도를 기반으로 이루어져 왔다. 그러나 올해 양국의 최종 적정성 결정 채택으로, 프레임워크에 참여하는 미국 기업들은 추가적인 보호조치 없이 데이터를 더 손쉽게 EU 지역으로 이전할 수 있다.

[표 2-2-1] EU-미국 데이터 프라이버시 프레임워크 최종 적정성 결정 채택 배경

기획 · 설계	기획 · 설계
EU-미국, 데이터 프라이버시 프레임워크 합의 ('22.03)	<ul style="list-style-type: none"> EU와 미국은 새로운 'EU-미국 데이터 프라이버시 프레임워크'에 합의 2020년 슈스 II 판결에서 유럽연합 사법재판소(Court of Justice of the European Union, CJEU)가 무효화한 미국 제도 '프라이버시 실드(Privacy Shield)'를 대체하고 '일반 데이터 보호 규정(GDPR)'에 따라 EU에서 제3국으로 개인 데이터 이동을 촉진하기 위해 제안
미국 바이든 대통령 행정명령 서명 ('22.10)	<ul style="list-style-type: none"> 미국 바이든 대통령이 행정명령에 서명함으로써 새로운 프레임워크에 대한 원칙적 합의를 미국 법률에 구현
EU, EU-미국 간 데이터 흐름 적정성 결정 초안 발표 ('22.12)	<ul style="list-style-type: none"> EU집행위가 행정 명령에 근거한 새로운 법적 프레임워크가 유럽 데이터 보호 표준과 비교할 수 있는 수준의 보안을 제공할 수 있으며, 유럽연합 거주자의 개인 데이터가 대서양 반대편으로 안전하고 합법적으로 전송될 수 있다고 판단
EU, EU-미국 데이터 개인정보 보호 프레임워크 결의안 채택 결정 ('23.05)	<ul style="list-style-type: none"> 유럽의회가 본회의에서 EU집행위가 제안하는 'EU-미국 데이터 프라이버시 프레임워크(EU-US Data Privacy Framework, DPF) 결의안'을 채택하기로 결정
EU, EU-미국 데이터 개인정보 보호 프레임워크 최종 적정성 결정 채택 ('23.07)	<ul style="list-style-type: none"> EU집행위원회가 EU-미국 데이터 프라이버시 프레임워크(EU-US Data Privacy Framework) 최종 적정성 결정 채택

* 출처: "데이터이코노미", 한국데이터산업진흥원 7월호(2023): 미국, EU 편.

2023년 6월 미국 정부는 영국 정부와도 유사한 합의를 했는데, 이는 'EU-미국 데이터 개인정보 보호 프레임워크'에 대한 내용을 영국으로 확장한다는 합의였다. 이에 미국 상무부(DOC, U.S. Department of Commerce)와 영국 과학혁신기술부(Department for Science, Innovation and Technology)가 미국-영국 간 새로운 '데이터 브릿지(Data bridge)'를 구축하고, 양국 간 개인 데이터의 자유로운 흐름을 촉진하겠다는 내용의 공동성명을 발표했다. 양국이 새로운 '데이터 브릿지'에 합의함에 따라 미국 기업은 이제 영국의 개인정보 데이터를 더 용이하게 이전할 수 있다.

기존에는 영국 기업이 '미국에 있는 서비스 제공업체 또는 기업'에 개인 데이터를 전송하려고 할 경우, 개인정보 보호 표준을 유지하기 위해 비용이 많이 드는 계약 조항을 마련해야 했었다. 데이터 브릿지를 통해 이러한 양국 간 데이터 이전에 대한 복잡한 절차의 부담을 없앴 뿐 아니라, 기업의 사업 진행프로세스를 간소화하며 비용을 절감할 것으로 보인다. 기업으로서는 더 손쉽게 국제적 거래 기회를 제공받을 것으로 기대된다.

4) EU-미국 프라이버시 실드(EU-US Privacy Shield): 유럽연합과 미국 간 상업적 목적의 대서양 횡단 개인 데이터 교환을 규제하기 위한 법적 프레임워크, 2016년 7월 12일에 발효되었다.

올해 2월에는 백악관이 미국-EU 간 공동의 AI 모델 구축 협정을 체결했다고 발표하며, 이를 기반으로 AI에 관한 연구를 더 심도 있게 발전시킬 것임을 공표했다. 양측은 체결된 협정을 토대로 주요 글로벌 과제를 해결하기 위한 AI 공동 개발 모델 및 통합 연구를 진행하게 된다. 구체적으로 극한의 기상 상황 및 이상 기후 예측, 비상 대응 관리, 건강 및 의료 개선, 전력망 최적화, 농업 최적화의 5가지 주요 영역에 초점을 맞추려고 한다.

사실 그동안 유사한 맥락의 노력은 꾸준히 있었다. 이는 2023년 3월 백악관 직속 산하의 과학기술정책실(OSTP, Office of Science and Technology)을 통해 발표한 ‘공평한 데이터 발전 진전 사항’에도 잘 드러나 있다. 이번 발표에서 미국 정부는 그동안 각 부문별로 산발되어 있는 데이터 표준을 확립하고, 개인정보 보호를 강화한 점을 부각했다. 또한 데이터 인벤토리, 카탈로그, 성과 대시보드 등을 통해 낮은 활용도의 데이터 사용 빈도를 크게 증대시킨 성과에 대해서도 공식적으로 발표했다.

이렇듯 미국 정부는 다양한 방법으로 데이터산업 활성화 전략을 펼치고 있으며, 데이터 안보에 대한 노력 또한 늦추지 않고 있다. 구체적으로 2023년 3월, 미국은 ‘국가 사이버 보안 전략(National Cybersecurity Strategy)’을 발표하였는데, 이를 기반으로 디지털 산업 전반에 걸쳐 기존의 사이버 보안 관행에 대한 규제를 강화하고 있다. 이를 위해 정부와 민간 부문 간 협력 역시 강화하는 추세다.

참고로 미국의 ‘국가 사이버 보안 전략’의 5가지 중점 목표로는 ① 주요 인프라 보호, ② 위협 행위자에 대한 교란 및 해체, ③ 보안 및 복원력을 촉진하기 위한 시장 환경 조성, ④ R&D 등 회복력 있는 미래에 대한 투자, ⑤ 공동의 목표를 추구하기 위한 국제 파트너십 구축이 있다.

2. 중국⁵⁾⁶⁾

중국: 데이터산업 육성을 위해 시행 정책 본격 다각화

중국은 지난 2021년 중국 공업정보화부(工业和信息化部)가 발표한 ‘14차 5개년 빅데이터 산업발전 계획(十四大数据产业发展规划)’을 기반으로 정부 주도의 하향식 데이터산업 육성 전략을 취하고 있다. 중국 정부는 올해 데이터산업 육성을 위해 시행 정책을 적극적으로 다각화하는 모습을 보였다.

2023년 1월, 공업정보화부(工业和信息化部), 국가사이버공간관리국(国家网信办) 등 16개의 중앙정부 부처가 공동으로 ‘데이터 보안 산업 발전 촉진에 관한 지침의견(关于促进数据安全产业发展的指导意见)’을 발표했다. 발표된 지침은 2025년까지 중국 내 데이터 보안 산업의 기본 역량을 강화하는 것을 목표로 한다.

또한 2월에는 중국 공산당중앙위원회(中共中央政治局常务委员会)와 국무원(国务院)이 ‘디지털중국건설규

5) “데이터이코노미”, 한국데이터산업진흥원 2월호(2023): 중국, 일본 편.

6) “데이터이코노미”, 한국데이터산업진흥원 6월호(2023): 중국, 일본 편.

획(数字中国建设整体布局规划)’을 발표했다. 발표된 계획에 따르면, 중국은 2023년 연내에 디지털 중국 건설과 관련된 주요 정책 시스템 구축을 완료하여 2025년까지 디지털 중국을 건설하기 위해 각 분야별 종합 발전을 도모하려 한다. 이후 2035년까지는 디지털 발전의 가시적인 성과를 내려고 하는데, 구체적으로 디지털 발전 수준에 있어서 세계 선두권에 진입하겠다는 목표를 수립했다.

[표 2-2-2] 디지털 중국 건설을 위한 의사 결정 배경 및 준비

항목	주요 내용
주요 의사 결정 배경	<ul style="list-style-type: none"> 2022년 국무원의 ‘디지털 정부건설 강화에 관한 지침 의견’ 발표 <ul style="list-style-type: none"> 새로운 과학기술 혁명과 산업 변혁 추세 적응, 디지털 경제와 디지털 사회의 발전 주도, 디지털 생태계 형성, 디지털 개발 가속화를 위한 디지털 정부 건설 제안 2022년 국무원의 ‘국가 통합정부 빅데이터시스템 구축 지침’ 발표 <ul style="list-style-type: none"> 통일된 표준,합리적 배치, 조정된 관리, 안전하고 신뢰할 수 있는 국가 통합 빅데이터 시스템 구축 촉진
디지털 중국 건설을 위한 디지털 정부의 수립	<ul style="list-style-type: none"> 디지털 정부 정책 수립 및 관련 규칙 및 규정 개정 정부의 디지털 역량 강화, 정보 시스템 네트워크 상호 연결, 데이터 공유, 효율적 협업 촉진 효율적인 정부 디지털 성과 역량 평가 시스템 구축, 포괄적인 보안 시스템 구축, 과학적이고 표준화된 규칙 시스템 구축, 데이터 자원 공유 시스템 구축

* 출처: “데이터이코노미”, 한국데이터산업진흥원 6월호(2023): 중국, 일본 편.

중국 국무원(中华人民共和国国务院)은 규칙 발표에 이어 2023년 3월, ‘국가데이터국(国家数据局) 설립 계획’을 발표했다. 이는 전국인민 대표대회에서 국가 제도개혁 방안 심의에 대한 동의에 따른 것으로, 앞서 발표한 ‘디지털중국건설계획(数字中国建设整体布局规划)’을 통해 공표한 바 있는 국가 데이터 관리 시스템 구축 계획을 실현한 것이다. 국가데이터국은 국가발전및개혁위원회(国家发展和改革委员会)가 관리하는 ‘국가국(国家局)⁷⁾으로, 데이터 인프라 시스템 구축을 촉진하고 데이터의 종합적인 개발 · 관리 · 활용이 원활하게 이루어지도록 뒷받침하는 곳이다. 이를 통해 중국은 디지털 경제 및 사회, 나아가 디지털 중국 건설을 본격화하고 있다.

이외에도 중국 정부는 데이터 거래 활성화를 위해 적극적인 움직임을 보였다. 지난 2021년 말 상하이 데이터 거래소 업무를 공식적으로 개시한 이후, ‘2022 금융 데이터 거래 번창 계획(2022 金融数据交易苗壮计划)’을 실시하여 금융 데이터 거래 프로젝트에 대한 특별 자금 지원을 하는 등 데이터 거래를 본격화했다.

이처럼 중국 정부는 다양한 정책으로 데이터 활용 활성화를 도모하면서, 데이터 보안 강화를 위해서도 노력하고 있다. 대표적으로 2023년 1월 3일 중국 공업정보화부(工业和信息化部), 국가사이버공간관리국(国家网信办) 개 발개혁위원회(发展改革委), 교육부(教育部) 등 16개의 정부부처가 공동으로 ‘데이터 보안 산업 발전 촉진에 관한 지침의견(关于促进数据安全产业发展的指导意见)’을 발표한 것을 들 수 있다. 발표된 지침을 보면 2025년까지 중

7) 현재 국무원 26개 부서(부처)에서 관리하는 16개 국가국(国家局)이 있으며, 국가데이터국이 설립되면 국가발전및개혁위원회가 관리하는 국가국이 3개가 된다. 기존의 2개 국가국으로는 국가에너지국, 곡물 및 재료 비축국이 있다.

국 내 데이터 보안 산업의 기본 역량을 증진시키는 것을 목표로 삼고 있으며, 국가 데이터 보안 산업 단지 조성, 혁신 애플리케이션 개발을 위한 첨단기술 시연 지역 공식 지정, 세계 수준의 데이터 보안 산업 기업 육성 등을 시행 정책으로 수립하였다.

3. 유럽연합⁸⁾⁹⁾

유럽연합: 데이터 공유 촉진 위한 법안 마련

올해 유럽 데이터산업에서 가장 큰 정책적 움직임으로는 지난 2023년 6월 유럽의회(EP)와 유럽이사회(Council of the European Union)가 유럽연합(EU) ‘데이터법(안)(Data Act)’에 합의한 것이 있다. ‘데이터법(안)’은 ‘데이터 거버넌스 법(Data Governance Act)’에 이어, ‘유럽 데이터 전략(A European Strategy for data)’에서 비롯된 두 번째 주요 입법 이니셔티브이기 때문이다. ‘유럽 데이터 전략(A European Strategy for data)’에서는 EU를 데이터 중심 사회의 리더로 만드는 것을 목표로 삼고 있다.

이에 맞춰 2021년 11월에 발효된 ‘데이터 거버넌스 법’이 기업, 개인, 공공 부문의 데이터 공유를 촉진하기 위한 프로세스와 법적 구조를 마련하는 것에 중점을 두었다면, 이번 ‘데이터법(안)’은 ‘데이터 거버넌스 법’의 법적 토대를 바탕으로 EU 역내 데이터 단일 시장을 형성하고 데이터 공유 활성화 방안을 마련하는 것에 초점을 맞췄다.

이러한 데이터법 합의에 앞서 유럽연합(EU)은 2022년 말(11월), ‘상호 운용 가능한 유럽법(Interoperable Europe Act)’을 채택했다. 이는 EU 국가 간 데이터 공유 및 재사용을 지원하는 것으로, ‘상호 운용 가능한 유럽법’¹⁰⁾ 채택을 통해 공공 행정 분야에서 EU 국가 간 상호운용성을 실질적으로 강화하는 것을 목표로 삼았다. 디지털 서비스의 상호운용성이 개선되면, 상호 운용 가능한 솔루션 사용을 통해 공공의 행정 목표를 더 효율적으로 달성할 수 있다. 이는 유럽 연합의 디지털 단일 시장 구축의 초석을 닦는 과정으로 평가할 만하다.

최근 유럽연합 회원국 사이에서 상호운용성에 대한 협력이 필요하다는 의견이 지속적으로 제기됐다는 점을 고려할 때, ‘상호 운용 가능한 유럽법’ 채택은 유럽 내 데이터 공유 활성화에 있어 큰 의미가 있다.

8) “데이터이코노미”, 한국데이터산업진흥원 4월호(2023), 미국 & EU.

9) “데이터이코노미”, 한국데이터산업진흥원 7월호(2023), 미국 & EU.

10) 상호 운용 가능한 유럽법은 유럽연합 내 공공 행정 간의 상호운용성에 대한 협력을 실질적으로 강화하려는 유럽연합 간 행정 연대를 위한 정책 시행 시 기반이 되는 법안이다.

[표 2-2-3] ‘EU 데이터법’ 입법 과정(좌측) 및 주요 목표(우측)

항목	내용	구분	내용
EU 데이터 전략 발표 ('20.02)	• EU의 주도권 확보를 위해 역내 데이터 단일시장 형성, 합법적,원활한 공유 환경 조성 목표	1	• 디지털 환경에서 데이터 가치의 공정한 분배 보장
EU 이사회 디지털 전환 회의 ('21.03)	• 성장, 번영, 안보, 경쟁력을 위한 디지털 전환의 중요성 강조	2	• 경쟁력 있는 데이터 시장 촉진
EU 데이터 거버넌스법 발표 ('22.05)	• 2020 ‘EU 데이터 전략’에 기초 • ‘EU 데이터 거버넌스 규정 보완	3	• 데이터 기반 혁신을 위한 열린 기회 조성
EU 데이터법 초안 발표 ('23.02)	• 데이터 재사용 증진, 데이터 공유에 대한 신뢰 제고, 데이터 수집 촉진 관련 내용 수록	4	• 개인들의 데이터 액세스 지원
EU 데이터법 잠정안 합의 ('23.06)	• 유럽 의회, 유럽 연합 이사회, 유럽 연합 집행위원회 최종 합의	5	• 소비자의 데이터 처리 서비스 제공업체 (클라우드 서비스 사업자)의 이용 전환 지원
		6	• 클라우드 서비스 사업자의 불법 데이터 전송 방지를 위한 보호 장치 마련
		7	• 부문 간 재사용 가능한 데이터에 대한 상호 운용성 표준 개발

* 출처: “데이터이코노미”, 한국데이터산업진흥원 7월호(2023); 미국, EU편.

한편, 유럽연합 역시 미국과 마찬가지로 데이터를 원활하게 해외로 이전하는 체계를 구축하기 위해 발 빠르게 대처하고 있다. 앞서 미국에서 언급한 유럽연합-미국 간 데이터 프라이버시 프레임워크(EU-US Data Privacy Framework) 최종 적정성 결정 채택이 그중 하나다. 이로 인해 유럽연합에서 활동하는 미국 기업들의 데이터 전송 절차가 크게 간소화되었으며, 양측 간의 데이터 교류가 더욱 활발해질 것으로 기대된다. 또한, 미국과 공동의 시 모델 구축(미국 참조) 결정을 통해서도 유럽연합이 단순히 데이터를 역내에서 활용하는 것에 그치지 않고, 글로벌 시장으로 뻗어나가려는 의도를 엿볼 수 있다.

이처럼 유럽연합은 다양한 방법을 통해 데이터 활용을 증대하고 활성화하고 있으며, 무분별한 데이터 활용으로 인해 발생할 수 있는 문제를 방지하기 위한 노력도 늦추지 않고 있다. 일례로 2023년 6월 유럽의회(European Parliament)는 인공지능을 규제하는 세계 최초의 법안인 인공지능(AI)법의 초안을 통과시켰다. 이를 통해 EU는 인공지능 규제 법안을 제정하기 위한 중요한 단계에 진입하였다. AI 규제법은 우선 유럽에서 개발되고 사용되는 AI가 인간 중심적이고 신뢰할 수 있는 AI의 활용을 촉진해야 한다는 내용을 담고 있다. 더 나아가 유해한 영향으로부터 건강, 안전, 기본권 및 민주주의를 보호하는 것을 목표로 한다. 이러한 AI 규제법안은 AI로부터 대량 감시를 피하기 위한 보호 장치로 작용할 것으로 예측된다.

4. 일본¹¹⁾

일본: 사회 전반에서 디지털화 전환 가속화

일본은 최근 몇 년 전까지 사회 대부분의 분야 행정이 아날로그 방식으로 처리될 정도로 디지털화 전환에 느린 편이었는데, 올해 들어 일본 정부는 사회 전반에서 빠른 디지털화 전환을 위해 매우 적극적으로 움직이고 있다.

대표적으로 일본 국토교통성(Ministry of Land, Infrastructure, Transport and Tourism, MLIT)은 2023년부터 도시개발을 고도화하기 위해 빅데이터를 사용하는 도시개발 사업에 보조금을 지급하기로 결정했다. 또 일본 내각부(CAO, Cabinet Office)는 의료분야의 데이터 활용 촉진을 위해 차세대 의료 인프라법을 개정하는 등, 올해 일본 정부는 각 산업군에서 데이터 활용 촉진을 위해 매우 적극적인 행보를 보였다.

특히 일본 경제산업성(METI)은 웹 3.0 유관 부처 간 협력을 기반으로 공통의 정책 수립하는 창구가 될 웹 3.0(Web 3.0) 정책실을 신설하였다. 이를 통해 세무, 회계, 법률제도, 지적재산권, 소비자 보호, 표준 등 새로운 비즈니스 환경에 적합하도록 시스템 구축을 진행함으로써, 그동안 시스템의 부재로 불편함을 겪던 일본 웹 3.0 사업 수행 기업들의 운영 편의를 대폭 제고하였다. 2022년 5월에는 기시다 후미오 일본 총리가 디지털 전환 계획을 발표하였는데, 이는 메타버스와 암호화폐(NFT)¹²⁾를 활용하는 웹 3.0 서비스 이용 확대를 지원하는 내용을 포함한다.

이런 추세에 따라 일본 정부는 정부 공식 포털사이트를 기존 인터넷 기반에서 클라우드로 전환하고 있으며, 최근 디지털 업무 관련 정원을 이례적으로 늘리는 방안 도입을 추진하는 등 디지털 인재 육성을 위한 정책도 시행하고 있다.

한편 일본 정부는 사회 전반에 걸친 사이버 보안 조치를 시행하고, 사이버 보안 분야에서 국제 협력을 다각화하고 있다. 대표적으로 2023년 1월 일본 경제산업성(METI)은 미국 국토안보부(DHS, Department of Homeland Security)와의 사이버 보안에 관한 협력각서(MOC, Memorandum of Collaboration)를 체결하였다. 또 태국, 인도네시아와는 각각 스마트 보안에 관한 사무총장급 협력각서를 체결하였으며, 그중에서도 특히 미국 및 아세안 국가들과 사이버 보안 연대를 공고히 하였다.

11) “데이터이코노미”, 한국데이터산업진흥원 3월호(2023): 일본 편.

12) NFT(Non-fungible token, 대체 불가능 토큰): 블록체인 기술을 이용해서 디지털 자산의 소유주를 증명하는 가상의 토큰

5. 영국¹³⁾¹⁴⁾

영국: 데이터법 개정안 재도입 추진

영국은 지난 2020년, ‘국가 데이터 전략(NDS, National Data Strategy)’을 발표하며 국경 간 데이터 이동을 적극 지원할 의사를 표명했다. 구체적으로 올해 6월 영국 과학혁신기술부(Department for Science, Innovation and Technology)는 미국 상무부(U.S. Department of Commerce, DOC)와 공동성명을 발표했는데, 영국-미국 간 새로운 ‘데이터 브릿지(Data bridge)’를 구축하여 양국 간 개인 데이터의 자유로운 흐름을 촉진하겠다는 내용을 골자로 한다.

같은 달인 2023년 6월, 영국은 싱가포르와 데이터 및 신기술 사용에 대한 연구 및 규제 협력을 강화하는 양해각서(MOU)를 체결하였다. MOU는 영국-싱가포르 디지털 경제 협정(UK-Singapore Digital Economy Agreement)과 영국-싱가포르 자유 무역 협정(UK-Singapore Free Trade Agreement)을 기반으로 체결되었다. 양국은 정부의 데이터 및 신기술 사용에 대한 새로운 발전을 주도하기로 약속했으며, 이를 통해 국가 간 연간 114억 파운드(약 19조 원)의 서비스 무역이 촉진될 것으로 기대된다.

한편 영국은 올해, 데이터법 개정안 재도입을 추진하였다. 2023년 3월, 영국 의회(UK Parliament)는 식별되거나 식별 가능한 개인 데이터 처리에 대한 규제를 규정하는 ‘데이터 보호 및 디지털 정보(제2호) 법안(DPDI, Data Protection and Digital Information(No.2) Bill)’을 발표했다. 이 법안은 2022년 여름 도입되었다가 데이터 전문가와의 정책개발 관련 협의 진행을 위해 일시 중지됐는데, 브렉시트 이후 영국의 국내 요구사항을 반영하여 개정되었다. 발표된 개정안은 데이터의 효과적인 사용이라는 목표를 달성하고, 데이터 보호 표준 유지를 위해 상호 강화된 협력을 하는 것을 기본으로 하고 있다. 이를 통해 조직에 ‘더 큰 유연성’을 제공함으로써, 향후 4년간 영국 경제에 10억 파운드 이상의 경제 효과를 기대할 수 있다.

6. 베트남¹⁵⁾¹⁶⁾

베트남: 사회 전반에서 디지털화 전환 추진

베트남 정부는 2023년 4월, 국가 최초의 포괄적 데이터 개인정보 보호법인 개인정보 보호에 관한 시행령을 발표했다. 인터넷 사용률이 높고 소셜 미디어 사용자가 많은 베트남에서는 개인 데이터 보호 규정 제정의 필요성이 오랫동안 제기되었다. 발표된 시행령은 ① 베트남 기관, 조직 또는 개인 ② 베트남에 있는 외국 기관, 조직 또는 개인, ③ 해외에서 활동하는 베트남 기관, 조직 또는 개인 ④ 베트남에서 개인 데이터를 처리하는 외국 기관, 조직 또는 개인

13) “데이터이코노미”, 한국데이터산업진흥원 5월호(2023): ASEAN, UK, 기타국가 편.

14) “데이터이코노미”, 한국데이터산업진흥원 8월호(2023): ASEAN, UK, 기타국가 편.

15) “데이터이코노미”, 한국데이터산업진흥원 5월호(2023): ASEAN, UK, 기타국가 편.

16) “데이터이코노미”, 한국데이터산업진흥원 8월호(2023): ASEAN, UK, 기타국가 편.

는 개인을 대상으로 하며, 베트남에 소재하거나 베트남에서 데이터 처리 활동을 수행하는 모든 베트남 및 외국 기업도 법령을 준수해야 한다.¹⁷⁾

베트남 정부는 올해, 사회 전반에서 디지털 전환을 이루기 위해 다양한 응용 프로그램을 개발했다. 일례로 베트남 교육부에 따르면 교육 데이터베이스와 사회보험 데이터베이스를 연동하여 졸업생들의 경력 동향을 파악하고 이를 기반으로 취업 통계를 구축할 것이며, 해당 데이터베이스를 연동하기 전에 보안 요구사항의 충족 및 규정 이행 여부를 확인할 것이다. 베트남 국세청은 납세 자료와 인구 통계의 동기화를 위해 MST(납세번호)를 개인 식별 번호로 사용 전환하는 정책을 추진하고 있다.

또한 베트남에서는 최근, 무현금 거래 애플리케이션 서비스가 본격화되는 등 '데이터 경제'가 실현되는 모습이 다. NAPAS(National Payment Corporation of Vietnam)는 약 5,000만 장의 비접촉식 카드를 발급하는 등 코로나19 팬데믹 이후 디지털 결제 및 인프라 공유가 점점 더 큰 역할을 하고 있다. 또 신용카드, 전자지갑, 모바일 머니 결제 등 다양한 결제 수단의 확산으로 사람들의 생활이 더욱 편리해지고 있다.

[표 2-2-4] 주요국 데이터 정책 및 법제도 현황

분류	미국	중국	EU	영국	일본	베트남
주요 정책	연방 데이터 전략	14차 5개년 빅데이터 산업발전 계획	유럽 데이터 전략	국가 데이터 전략	포괄적 데이터 전략	
데이터산업 육성 기구	최고데이터 책임자위원회	공업정보화부 국가데이터국	유럽연합 집행위원회	디지털문화 미디어스포츠포부	디지털청	정보통신부
세부 정책	연방데이터 전략 실행 계획, 유럽연합-미국 간 데이터 프레임 워크	디지털중국 건설계획, 금융 데이터 거래 번창 계획	유럽 오픈사이언스 클라우드, 유럽연합-미국 간 데이터 프레임 워크	데이터 퍼스트 프로그램	웹 3.0, 마이넘버카드, 데이터 대시 보드	국가데이터 포털, 디지털화 전환 정책, 데이터 현지화 정책
주요 규제 및 법령	국가 사이버 보안법, 각주(州)의 데이터/ 프라이버시 보호법	데이터보안법	일반데이터 보호 규정 (GDPR), 데이터 거버넌스법, 데이터법, 상호운용 가능한 유럽법, 인공지능(AI)법	일반 데이터 보호 규정, 데이터법	개정 개인정보 보호법, 디지털사회 형성기본법, 민간데이터 활용추진기본법	사이버 보안법 2022년 시행령, 개인정보 보호에 관한 시행령
특징	연방 기관의 데이터 역량 강화에 중점	정부 주도의 하향식 산업 육성	EU 통합 데이터 정책 마련에 중점	국경 간 데이터 이전 적극 지원	행정 데이터 기반 민관협력 추진	자국의 데이터 주권 법으로 명시

* 출처: 한국데이터산업진흥원

17) 2023년 7월 1일부터 시행됐으며, 위반하는 경우 규정에 따라 징계 조치, 행정 제재 또는 형사 처벌을 받게 된다.

제3장 데이터 활용 이슈의 법제도적 측면 : 저작권/개인정보

윤아리 변호사 김 · 장 법률사무소

데이터를 기반으로 한 AI 발전의 이면에는 다양한 이슈가 있다. 대개 그것은 어떠한 데이터가 어떻게 수집되고 활용 되는지와 관련된다. 그중 저작권과 개인정보 관련 이슈는 이해관계가 충돌할 수 있는 양면성을 가진 대표적 분야로 서, 최근 이와 관련된 다양한 논의가 활발히 이루어지고 있다. 이는 타인의 저작권 및 개인정보의 침해 유발하는 데이터의 무분별한 활용은 방지하면서도 AI 기술의 발전을 도모하기 위한 것이다.

이하에서는 AI 데이터 활용과 관련하여 논의되는 주요 저작권 및 개인정보 관련 이슈를 법제도 측면에서 살펴보고자 한다.

1. 저작권 이슈

가. 도입

오늘날 AI는 인간의 업무를 보조하는 역할을 넘어 인간의 고유 영역으로 여겨져 온 예술 및 창작 분야까지 영역 력을 넓혀가고 있다. 챗GPT(ChatGPT), 바드(Bard) 등으로 대표되는 초거대 AI는 전문가뿐 아니라 누구나 쉽사리 활 용할 수 있어 변화의 체감도가 특히 높다.

이러한 변화의 순기능을 온전히 향유하려면 그에 맞춘 법 · 제도의 정비가 필요하다. 이에 각국은 기존 법체계 로 해결하기 어려운 AI 관련 쟁점들에 대한 해결 방안을 활발히 논의하고 있다. 우리 정부도 인공지능 법제정비단¹⁾, AI-저작권법 제도개선 워킹그룹²⁾ 등을 통해 다양한 쟁점을 검토 중인데, 이 중 저작권과 관련해서는 첫째, AI가 데이 터 마이닝(data mining)³⁾ 과정에서 저작권에 의해 보호되는 다양한 저작물을 활용해도 되는지 여부와 둘째, AI가 만 들어낸 산출물의 저작권 인정 여부 및 저작권 인정 시 그 주체가 주로 문제 되고 있다.

나. 데이터 마이닝 과정의 저작권법 위반 가능성

1) 현행법상 관련 규정

‘저작권법’은 창작자의 배타적 권리인 저작권을 보호하기 위한 조항들을 열거하는 한편, 저작권 침해가 문제 될 수 있는 사안에 대한 면책 규정도 마련하고 있다. 저작물을 이용하기 위해서는 원칙적으로 사전 허락을 받아야 하

1) 관계부처 합동, 초거대AI 경쟁력 강화 방안, (2023), p.30.

2) 문화체육관광부 보도자료, AI 시대 새로운 저작권 해법 찾기 워킹그룹 첫 회의, (2023), p.1.

3) 아직 ‘데이터 마이닝(data mining)’에 대한 법상 정의는 마련되어 있지 않으나, 통상적으로 데이터에 접근, 추출 및 분석하여 결과물을 사용하기까지의 과정을 지칭하는 개념으로 사용된다.

는데, 데이터 마이닝 과정에서 사용되는 방대한 데이터에 대해 저작권자의 이용 허락을 일일이 받기는 사실상 불가 능하다. 따라서 저작권 여부와 관련해 ‘저작권법’상 면책 규정이 매우 중요한 의미가 있는데, 데이터 마이닝과 관련 해서는 ‘저작권법’ 제35조의5 제1항(저작물의 공정한 이용) 및 제35조의2(저작물 이용과정에서의 일시적 복제)가 주 로 거론된다.

우선 저작물의 공정한 이용에 관한 ‘저작권법’ 제35조의5 제1항에 따르면, 저작물의 통상적인 이용 방법과 충돌 하지 아니하고 저작자의 정당한 이익을 부당하게 해치지 아니하는 경우에는 저작자의 허락을 받지 않고도 저작물을 이용할 수 있다. 동조 제2항에 따르면 특정 저작물 이용 행위가 위와 같은 ‘공정한 이용’에 해당하는지 판단할 때에는 ① 이용의 목적 및 성격, ② 저작물의 종류 및 용도, ③ 이용된 부분이 저작물 전체에서 차지하는 비중과 그 중요성 및 ④ 저작물의 이용이 그 저작물의 현재 시장 또는 가치나 잠재적인 시장 또는 가치에 미치는 영향을 고려하여야 한다.

또 다른 면책 규정인 ‘저작권법’ 제35조의2에 따르면, 컴퓨터에서 저작물을 이용하는 경우 원활하고 효율적인 정보처리를 위하여 필요하다고 인정되는 범위 안에서 그 저작물을 그 컴퓨터에 일시적으로 복제할 수 있다. 다만, 그 저작물의 이용이 저작권을 침해하는 경우에는 이러한 면책 규정이 적용되지 않는다. 법원은 일시적 복제 자체가 저작물의 이용 등에 불가피하게 수반되는 경우는 해당 규정이 적용될 수 있지만, 독립한 경제적 가치를 가지는 경 우는 제외된다고 판시한 적이 있다(대법원 2018. 11. 15. 선고 2016다20916 판결).

이외에 어느 정도 기간이 ‘일시적’인지 등 실무상 문제가 될 수 있는 사항에 대해서는 명확한 기준이 확립되 지 않았다.

데이터 마이닝 과정에서 저작물을 이용하는 행위를 두고 위와 같은 두 면책 규정으로 포섭할 수 있을지는 아직 법원이 직접적인 판단을 한 사례가 없다. 따라서 구체적인 사안별로 평가가 달라질 여지도 있다.⁴⁾

2) 국내외 입법 동향

지난 2021년 1월 발의된 ‘저작권법 전부개정법률안(도종환 의원 대표발의, 의안번호 제2107440호)’은 현행법상 의 불명확성을 감안하여 제43조⁵⁾에 정보분석을 위한 복제 · 전송에 관한 조항을 신설하는 내용을 담고 있으며, 이후 발의된 2건의 ‘저작권법 일부개정법률안’⁶⁾들도 유사한 조항을 두고 있다. 2023년 7월 현재 3건 개정안 모두 소관 상 임위원회인 문화체육관광위원회에 계류 중이다.

4) 예컨대, 직접적인 경쟁상대의 데이터를 AI학습데이터로 활용하면서 원 데이터보유자와 유사한 목적의 유사한 결과물을 도출하는 경우에는 공정이용으로 인정받기 어려울 수 있다.

5) 제43조(정보분석을 위한 복제·전송) ① 컴퓨터를 이용한 자동화 분석기술을 통해 다수의 저작물을 포함한 대량의 정보를 분석(규칙, 구조, 경향, 상관관계 등의 정보를 추출하는 것)하여 추가적인 정보 또는 가치를 생성하기 위한 것으로 저작물에 표현된 사상이나 감정을 향유하지 않는 경우에는 필요한 한도 안에서 저작물을 복제 · 전송할 수 있다. 다만, 해당 저작물에 적법하게 접근할 수 있는 경우에 한정한다.

② 제1항에 따라 만들어진 복제물은 정보분석을 위하여 필요한 한도에서 보관할 수 있다.

6) 이용호 의원 대표발의안(2022. 10. 31., 의안번호 제2117990호) 및 황보승희 의원 대표발의안(2023. 6. 8., 의안번호 제2122537호).

해외에는 데이터 마이닝을 위한 구체적 면책 규정을 도입한 경우(일본⁷⁾, 유럽연합⁸⁾, 영국⁹⁾)도 있지만 아직 별도의 입법 없이 우리 ‘저작권법’ 제35조의5와 같은 ‘공정한 이용’ 법리를 통하여 해결하는 경우도 있다(미국).

다. 산출물에 대한 저작권 인정 관련

1) AI 산출물도 저작물인지 여부

‘저작권법’ 제2조 제1호 및 제2호에 따르면 ‘저작물’은 인간의 사상 또는 감정을 표현한 창작물을 말하며, 이때 ‘저작자’는 저작물을 창작한 자를 말한다. 즉, ‘저작권법’은 인간(자연인)이 창작한 것만을 저작물로 인정하고 있다. 법원 역시 “사상이나 감정에 대한 창작자 자신의 독자적인 표현”을 요구(대법원 2021. 6. 24. 선고 2017다261981 판결) 하거나 저작자의 “정신적 노력”에 의한 창작 행위를 요구(대법원 1993. 6. 8. 선고 93다3073 판결 및 대법원 1995. 11. 14. 선고 94도2238 판결 등)하고 있다.¹⁰⁾

참고로 미국의 경우는 저작권청이 일관되게 AI 산출물의 저작물성을 부정하면서 AI 산출물의 등록을 거절하고 있는데¹¹⁾, 2023년 3월 발표된 ‘AI 생성물의 저작권 등록지침에 관한 가이드라인’¹²⁾에서는 AI로 생성한 결과물에 대하여 기존 저작물과 동등하게 저작권 심사를 받을 수 있는 기회를 제공한다고 하면서도 핵심 요소인 사람의 창의성이 드러나야 한다는 취지를 밝힌 바 있다.

2) 저작권의 귀속 주체

위에서 살펴본 바와 같이 현행법상 AI 산출물을 저작물로 인정하기는 어려우나, AI 산출물을 저작물로 인정하더라도 그 저작권의 귀속 주체가 누구인지 판별하는 문제가 남는다.

위에서 살펴본 바와 같이 현행법상 자연인이 아닌 AI 자체를 저작자라고 인정하기는 어렵다. AI 자체를 제외할 경우에도, AI의 개발부터 제공, 활용 과정에 프로그래머 등 AI 제작자, AI 서비스 제공자, AI 서비스 이용자, 학습데이터의 저작권자 등 여러 이해관계자가 문제될 수 있는데, 법원은 복수의 주체가 저작물의 작성에 관여한 경우 창작적인 표현형식 자체에 기여한 자만이 저작자가 되고 아이디어나 소재 또는 자료를 제공하였더라도 직접적인 창작적 표현에 관여하지 않았다면 저작자로 보기 어렵다는 입장이므로(대법원 2020. 6. 25. 선고 2018도13696 판결) 위와 같은 이해관계자들의 기여분이 어느 정도인지, 즉 누가 창작적 표현 과정에 핵심적으로 기여했는가 하는 쟁점이

평가의 주요한 기준으로 작용할 것으로 보인다. 이와 관련해 논의¹³⁾를 진행하고 있는데, AI 산출물을 만인 공유로 보아야 한다는 입장도 있다.

2. 개인정보 보호 이슈

가. 도입

AI 데이터 활용과 관련해 개인정보 보호 이슈도 문제 된다. 개인정보는 AI 학습 단계부터 배포, 서비스 제공 단계까지 전 과정에 걸쳐 활용될 수 있는데, 크게는 AI 모델 학습 과정에서 개인정보가 포함된 데이터 세트(training data)를 활용하는 부분과, AI 서비스 제공 과정에서 개인정보를 처리하는 부분으로 구분할 수 있다.

나. AI 학습과 개인정보 활용

1) 공개된 개인정보의 활용

최근 주목받는 파운데이션 모델(foundation model)은 온라인상에 공개된 텍스트, 이미지 자료를 매우 광범위하게 크롤링하는 방식¹⁴⁾으로 학습용 데이터 세트를 구성한다고 알려져 있다. 이 과정에서 개인을 알아볼 수 있는 정보, 즉 개인정보가 포함되는 것은 불가피한 측면이 있다. 개인정보가 포함된 웹 페이지를 사전에 수집 대상에서 배제하거나 이미 수집된 정보에서 개인을 알아볼 수 있는 정보를 특정 후 임의의 내용으로 대체하거나 삭제하는 등의 조치를 통하여 개인정보 이슈를 해결하고자 하는 경우도 있는데, 이런 식의 문제 해결은 현실적으로 여의치 않다. AI를 개발하는 데 요구되는 정보의 양은 기하급수적으로 증가하고 있기 때문이다. 게다가 법상 개인정보의 범위는 매우 광범위하게 정의되어 있다 보니, 개인정보의 피해를 줄이려는 조치에도 불구하고 온라인상에 공개된 정보로부터 구축된 데이터 세트에 개인정보가 포함될 가능성이 여전히 잔존할 수 있다.

‘개인정보 보호법’은 개인정보를 수집·이용할 때에는 원칙적으로 정보주체의 명시적인 동의를 받도록 하고 있으나¹⁵⁾, 공개된 정보를 대량 수집하면서 일일이 정보주체의 동의를 받는 것은 현실적으로 불가능하다. 그 때문에 실무상으로는 정보주체의 동의 없이 공개된 개인정보를 처리할 수 있는 법적 근거가 필요해진다.

7) ‘저작권법’, 제30조의4.

8) 디지털 단일시장 저작권 지침(Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market), 제3조 및 제4조.

9) ‘저작권법’, 제29조.

10) 한국저작권위원회 홈페이지 게시물, 『인공지능(AI)을 이용한 창작물 안내』, 2022, https://www.cros.or.kr/psnsys/cmmn/infoPage.do?w2xPath=/ui/twc/sercen/notinf/notInf_dt.xml

11) 2018년 ‘A Recent Entrance to Paradise’ 사례 및 2023년 ‘Zarya of the Dawn’ 사례가 특히 화제가 된 바 있다.

12) Copyright Registration Guidance: Works Containing Material Generated by Artificial Intelligence, United States Copyright Office, (2023. 3. 16.).

13) 참고로 지난 2020년 12월 발의된 ‘저작권법 일부개정법률안(주호영 의원 대표발의, 의안번호 제2106785호)’는 인공지능이 제작한 저작물에 대한 저작권의 법적 근거를 마련하는 내용을 포함하고 있다. 동 개정안은 ‘인공지능 저작물’ 및 ‘인공지능 저작물의 저작자’의 정의를 신설하는 내용을 담고 있으며, 제2조 제2호의2는 ‘인공지능 저작물의 저작자’를 “인공지능 서비스를 이용하여 저작물을 창작한 자 또는 인공지능 저작물의 제작에 창작적 기여를 한 인공지능 제작자·서비스 제공자 등”으로 상당히 광범위하게 정의하고 있다. 2023년 7월 현재 개정안은 소관 상임위원회인 문화체육관광위원회에 계류 중인데, 해당 개정안에 대한 전문위원 검토보고서에서는 ‘인공지능의 창작물을 법적으로 보호할 필요성이 있는지,’ 및 ‘인공지능 저작물의 저작자를 판단하는 데 있어 창작 기여도가 충분한 기준이 될 수 있는지’ 등에 대한 사회적 합의가 부족하다는 입장을 취한 바 있어, 후속 논의를 지켜볼 필요가 있다.

14) 크롤링을 이용한 학습용 데이터의 수집과 관련해 개인정보 이슈 외에도 다양한 법적 쟁점이 논의되고 있다. 법상 스크레이핑·크롤링이 그 자체로 금지된 것은 아니나, 구체적 사실관계에 따라서는 ‘저작권법’, ‘정보통신망 이용촉진 및 정보보호 등에 관한 법률’, ‘형법’, ‘데이터 산업진흥 및 이용촉진에 관한 기본법’ 등 개별 법령의 위반 또는 민사상 분쟁이 발생할 가능성이 있다.

15) ‘개인정보 보호법’, 제15조 제1항 제1호.

그런데 ‘개인정보 보호법’에서는 공개된 개인정보의 처리에 관한 명확한 규정을 두고 있지 않다.¹⁶⁾ 다만, 대법원 2016. 8. 17. 선고2014다235080 판결(이른바 ‘로앤비 판결’)에서는 이 문제에 대하여 공개된 개인정보를 별도의 동의 없이 처리할 수 있는 법적 근거를 제시하였다.¹⁷⁾

위 사안에서 대법원은 공개된 개인정보의 경우 ‘정보주체의 동의가 있었다고 객관적으로 인정되는 범위’내에서 수집 · 이용 · 제공 등 처리를 할 때에는 별도의 동의가 필요하지 않다고 판단하면서, ‘정보주체의 동의가 있었다고 객관적으로 인정되는 범위’의 판단 기준으로서 ① 공개된 개인정보의 성격, ② 공개의 형태와 대상 범위, ③ 그로부터 추단되는 정보주체의 공개 의도 내지 목적뿐만 아니라, ④ 정보 처리자의 정보 제공 등 처리의 형태를 들었다. 또한 ⑤ 그 정보 제공으로 인해 공개의 대상 범위가 원래의 것과 달라졌는지, ⑥ 그 정보 제공이 정보주체의 원래의 공개 목적과 상당한 관련성이 있는지 살피는 것도 주요한 판단 기준이 된다고 하였다.

최근에는 온라인상에 공개된 개인정보를 AI 개발에 활용하는 것이 ‘정보주체의 동의가 있었다고 객관적으로 인정되는 범위’에 포함될 수 있는지 논의가 계속되고 있다. 이는 AI 학습에 활용된 원 데이터의 내용, 공개 매체, 방식 등을 고려한 정보주체의 공개 의도 등 구체적인 사안별로 달리 판단될 여지가 있다.

한편, ‘개인정보 보호법’에 따르면 개인정보처리자의 정당한 이익을 달성하기 위하여 필요한 경우로서 명백하게 정보주체의 권리보다 우선하는 경우에는 정보주체의 명시적인 동의 없이도 개인정보를 수집 · 이용할 수 있다.¹⁸⁾ 이에 해당 규정을 활용하여 공개된 정보를 이용하는 방안도 논의하고 있다.

개인정보보호위원회는 최근 ① 객관적 동의 의사가 추단되는 범위 또는 ② 정당한 이익이 정보주체 권리보다 명백히 우선하는 범위 내에서 공개된 정보를 수집 · 이용할 수 있다는 입장을 밝히면서, 공개된 정보 처리를 둘러싼 이익을 형량할 때 특히 다음과 같은 사항을 고려하도록 하였다.

〈 공개된 정보 처리 관련 이익 형량 시 고려사항¹⁹⁾ 〉

- ✓ 민감정보 처리제한(법 제23조), 노출된 개인정보의 삭제 · 차단(법 제34조의2) 등을 준수할 것
- ✓ 웹사이트 운영자가 ‘robots.txt’ 설정을 통해 로봇의 접근을 제한한 경우 해당 페이지에는 접근하지 않는 로봇배제표준을 준수할 것
- ✓ 시간 · 비용 · 기술을 합리적으로 고려하여 특정 개인을 식별하기 어렵도록 하는 조치*의 이행 여부 및 조치 수준
* 계정정보 등 분리, 특정 개인을 식별할 수 있는 주민등록번호 · 운전면허번호 · 카드번호 등 식별자 삭제 등
- ✓ 개인을 유추하거나 식별할 목적으로 데이터를 처리하는지 여부²⁰⁾
- ✓ 서비스 과정에서 개인정보 침해에 대한 모니터링을 실시하고 정보주체에게 대응 수단을 제공하며, 침해 발생 시 즉시 조치할 수 있는 체계를 갖추었는지 여부
- ✓ 학습데이터 출처, 개인정보 처리 방법 등을 투명하게 공개하는지 여부

2) 서비스 제공 과정에서 수집된 개인정보의 활용

AI 학습에는 특정 서비스 제공 과정에서 수집된 개인정보가 활용되는 경우도 있다.

이와 관련해 AI 챗봇 서비스 개발사가 자사의 다른 앱 서비스 과정에서 수집한 이용자의 개인정보(카카오톡 대화)를 AI 챗봇 서비스 개발에 활용한 행위가 개인정보를 수집 목적 범위를 초과하여 이용한 것으로 문제 되어 개인 정보보호위원회로부터 과징금을 부과받은 사례가 있다(이른바 ‘이루다 사건’).²¹⁾ 해당 사안에서 개발사는 기존 앱의 개인정보 처리방침에 ‘신규 서비스 개발’을 수집 · 이용 목적으로 기재하고 있었으나, 개인정보보호위원회는 이러한 기재만으로는 이용자가 개발사의 활용 방안을 예상할 수 있다고 보기 어렵다는 등의 이유로 법 위반을 인정하였다. 당시 개발사는 해당 앱에서 수집된 카카오톡 대화를 해당 앱과는 별개의 챗봇 서비스의 개발 · 운영에 이용했다.

이러한 결정례를 고려하면, 서비스 제공 과정에서 수집된 개인정보를 해당 서비스와 관계 없는 AI 학습 목적으로 활용하기 위해서는 정보주체가 이를 충분히 예상할 수 있도록 동의를 받을 필요가 있다고 볼 수 있다.

한편, 서비스 제공 과정에서 수집된 개인정보를 해당 서비스와 관련된 AI 학습에 이용하는 경우라도 여전히 동의가 필요하다는 견해가 있을 수 있으며, 이에 대해서는 개인정보의 추가적 이용 규정²²⁾을 활용하는 방안 등이 논의되고 있다.

3) 가명정보 특례를 활용한 AI 학습

‘개인정보 보호법’은 데이터 활용을 촉진하기 위한 취지에서 정보주체의 동의 없이도 과학적 연구 등 일정한 목적의 가명정보의 처리를 허용하고 있다(제28조의2). 과학적 연구에는 새로운 기술 · 제품 · 서비스 개발 및 실증을 위

16) 다만, 개인정보보호위원회가 고시한 ‘표준 개인정보 보호지침’에서는 개인정보처리자가 공개된 매체 또는 장소에서 개인정보를 수집하는 경우 사회 통념상 동의 의사가 있었다고 인정되는 범위 내에서만 이용할 수 있다고 규정하고 있다(제6조 제3항).

17) 해당 사안에서는 법률 정보를 제공하는 웹 사이트를 운영하는 회사(A)가 대학교수(B)의 개인정보(사진, 성명, 성별, 출생연도, 직업, 직장, 학력, 경력 등)를 해당 교수의 동의 없이 대학교 홈페이지, 교원명부, 교수 요람에서 수집하여 자사 웹 사이트에서 유료로 제3자에게 제공한 행위가 문제 되었고, 법원은 회사의 행위는 해당 교수의 개인정보자기결정권을 침해하는 위법한 행위로 평가할 수 없다고 판단하면서, 회사의 손해배상 책임을 부정하였다.

18) ‘개인정보 보호법’, 제15조 제1항 제6호.

19) 개인정보보호위원회, 인공지능 시대 안전한 개인정보 활용 정책방향, (2023), p.8.

20) 다만, 공적인 인물에 관한 정보로서 알 권리가 중요하다고 판단되는 범위에 대해서는 정확성 · 최신성 확보를 위해 합리적으로 필요한 범위 내에서 식별 목적으로 처리할 수 있다고 한다.

21) 개인정보보호위원회 2021. 4. 28.자 결정 제2021-007-072호. 이는 AI의 개인정보 처리가 문제시된 첫 번째 사례로 큰 주목을 받았다.

22) ‘개인정보 보호법’, 제15조 제3항.

한 산업적 연구가 포함되므로, AI 연구개발도 이에 해당할 수 있다. 다만 실무상 이미지 · 영상, 음성, 텍스트 정보 등 파운데이션 모델의 학습에 활용되는 비정형 데이터의 가명처리 기준이 명확하지 않은 점 등이 문제 되고 있다. 이에 개인정보보호위원회는 2023년 12월까지 비정형 데이터의 가명처리 기준을 마련하겠다는 계획을 발표하였다.²³⁾

다. AI 서비스 이용 · 제공 과정에서의 개인정보 활용

AI 서비스의 경우에도 다른 서비스와 마찬가지로 서비스 이용 · 제공 과정에서 개인정보가 처리되는 경우에는 동의, 개인정보 처리방침 공개, 정보주체 권리보장 등 개인정보 관련 규제가 적용된다. 특히, 정보주체의 권리보장과 관련해 향후 개정 '개인정보 보호법'상 도입될 자동화된 의사결정에 관한 정보주체의 대응권 역시 문제 될 수 있다.

지난 2023년 3월 14일에 공포된 개정 '개인정보 보호법'은 자동화된 결정에 대한 정보주체의 권리에 관한 조항을 신설하였다(제37조의2). 2024년 3월 15일부터 시행될 예정인 위 조항은 완전히 자동화된 시스템(AI 기술을 적용한 시스템도 포함)으로 개인정보를 처리하여 이루어지는 결정(자동화된 결정)에 대하여 설명 등을 요구하거나, 혹은 그 결정이 정보주체의 권리 또는 의무에 중대한 영향을 미치는 경우 해당 결정을 거부할 수 있는 권리를 창설하였다. 또한 자동화된 결정의 기준과 절차, 개인정보가 처리되는 방식 등을 모두 공개하도록 하고 있다. 따라서 개정 조항 시행 후에는 AI 기반 의사결정 서비스를 활용하는 경우, 정보주체가 해당 결정을 거부하거나 설명 등을 요구할 때 이를 적절히 반영하는 절차를 마련해야 할 것으로 보인다.

개정법에서는 자동화된 결정의 거부 · 설명 등을 요구하는 절차 및 방법, 거부 · 설명 등의 요구에 따른 필요한 조치, 자동화된 결정의 기준 · 절차 및 개인정보가 처리되는 방식의 공개 등에 필요한 사항을 대통령령으로 정하도록 명시했으므로, 상세한 내용은 향후 공개될 시행령을 통해 정해질 예정이다.

라. 규제 동향

개인정보 보호 측면에서 AI의 안전한 연구 및 개발을 활성화하기 위한 다양한 논의가 진행되고 있다.²⁴⁾ 정부는 2023년 4월 '초거대 AI 경쟁력 강화 방안'을 통하여 활용 가치가 높은 데이터를 개인정보 침해 우려 없이 안전하게 AI 학습용 데이터로 활용할 수 있는 방안을 마련할 예정이라고 밝혔다. 여기엔 비정형데이터 가명처리 기준 마련, 재현 데이터 활용 확대, 적극적으로 공개한 개인정보의 AI 학습용 데이터 활용 허용 검토 등을 포함한다.

이후 개인정보보호위원회는 2023년 8월 AI 개발 · 서비스 단계별 개인정보 처리기준과 보호조치, 고려 사항 등을 포함한 '인공지능 시대 안전한 개인정보 활용 정책 방향'을 발표하였다.

[표 2-3-1] AI 개발 · 서비스 단계별 개인정보 처리기준²⁵⁾

기획 · 설계	<ul style="list-style-type: none"> 개인정보 보호 중심 설계(Privacy by Design) 원칙 반영 AI 단계별 위험 분석 및 대응계획 수립 	
데이터 수집	일반 개인정보	<ul style="list-style-type: none"> 적법하게 수집한 정보는 '수집 목적 범위 내'에서 이용 가능 당초 수집 목적과 합리적으로 관련된 범위에서 정보주체 이익을 부당하게 침해하지 않는 경우 추가적 이용 가능
	공개된 정보	<ul style="list-style-type: none"> 이익형량 후 (1) 동의의사가 있다고 객관적으로 추단되거나 (2) 처리자의 정당한 이익이 명백히 우선하는 경우 수집 · 이용 가능 공개된 정보를 크롤링 등으로 수집, 가명처리 후 AI 학습에 이용 가능
	영상정보	<ul style="list-style-type: none"> (고정형) 당초 설치 · 운영 목적 관련 AI 개발 가능, 관련 없는 경우 익명 · 가명처리 필요 (이동형) 당초 촬영 목적 달성을 위해 필요한 범위 내에서 안전 조치 후 원격 관제, 최소한의 저장 가능
	생체인식정보	<ul style="list-style-type: none"> 별도 동의가 있거나 법령 근거가 있는 경우에만 처리 가능 대체 수단 마련, 원본 정보 분리 보관 등 보호조치 이행
AI 학습	<ul style="list-style-type: none"> 과학적 연구 등의 목적으로 가명처리하여 동의 없이 AI 개발 가능 활용목적 · 처리환경에 맞는 개인정보 보호 강화 기술(PET) 활용 	
서비스 제공	<ul style="list-style-type: none"> AI 학습데이터 수집 방법, 서비스 과정에서 생성되는 정보의 처리 방법 등 안내 삭제 · 처리정지 · 자동화된 결정 대응권 등 정보주체 권리행사 보장 	

또한, 변화 속도가 빠르고 데이터 활용 범위, 방식이 고도로 복잡한 AI의 특성을 고려하여 향후에는 원칙 중심의 규율을 정립해 나가겠다는 입장을 밝힌 바 있다. 이러한 원칙 중심의 규율은 비정형데이터 가명처리 기준(2023. 12. 예정), 공개된 정보 활용 가이드라인(2024. 3. 예정), AI 투명성 확보 가이드라인(2024. 6. 예정) 등 향후 마련될 데이터 처리 기준을 통해 더 구체화될 예정이므로 후속 논의에 주목해야 한다.

23) 개인정보보호위원회, 신뢰 기반 인공지능 데이터 규범, 첫 발 떼다, (2023), p.5.

24) 지난 2020년 2월 4일 개정된 바 있는 '개인정보 보호법'에서는 4차 산업혁명 시대의 데이터 활용을 촉진하기 위해 과학적 연구 등 목적의 가명정보 처리를 허용하였다. 또 개인정보보호위원회는 이루다 사건을 계기로 AI 개인정보보호 자율점검표를 발간하기도 하였다.

25) 개인정보보호위원회, 신뢰 기반 인공지능 데이터 규범, 첫 발 떼다, (2023), p.8.

3^{PART}

데이터산업 시장 현황

제1장 • 국내 데이터산업 시장 현황

제2장 • 국내 데이터산업의 역동성 및 생산성 분석

제3장 • 해외 데이터산업 시장 현황

제1장 국내 데이터산업 시장 현황

고태우 팀장 KDB산업은행

국내 데이터산업 시장 현황은 '2022 데이터산업 현황조사' 보고서(과학기술정보통신부 · 한국데이터산업진흥원, 2023.4)¹⁾를 요약하여 작성하였다. 2021년 데이터산업 분류에 따라 '데이터 처리 및 관리 솔루션 개발 · 공급업', '데이터 구축 및 컨설팅 서비스업', '데이터 판매 및 제공 서비스업'으로 구분하였다. 본 장에 수록된 시장 규모는 데이터 비즈니스를 영위하는 사업체의 데이터 관련 매출 기반의 추정치이다. 2021년 이전까지 시장 규모는 확정치인 반면, 2022년 시장 규모는 잠정치이기 때문에 향후 공표 예정인 확정치 통계 결과와 상이할 수 있음을 미리 밝힌다.

1. 국내 데이터산업 시장 규모

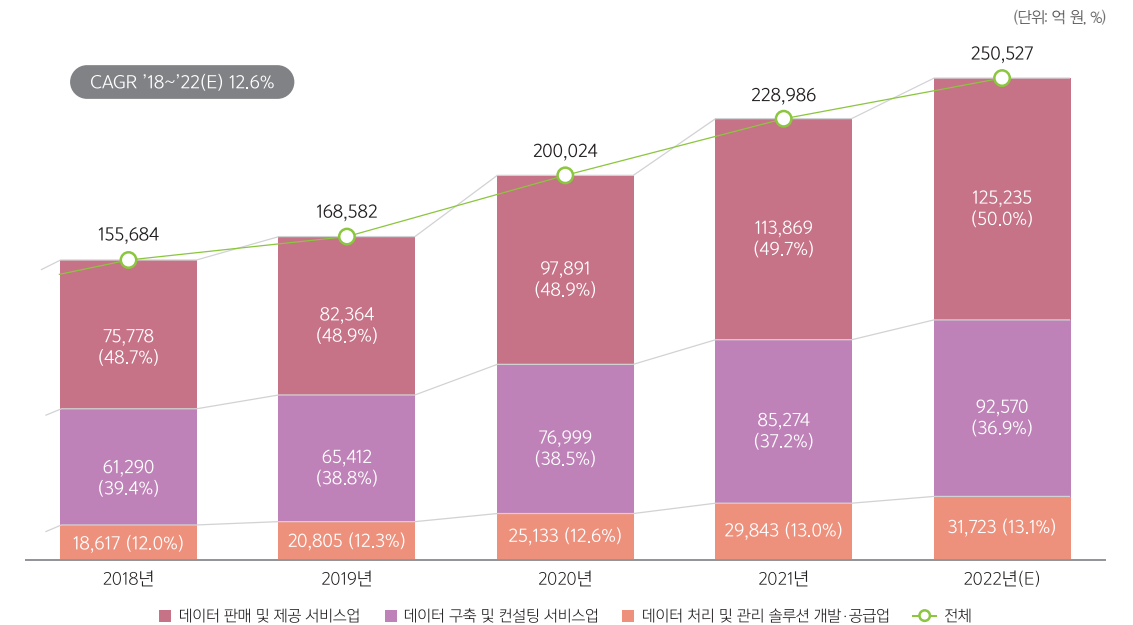
2021년 데이터산업²⁾ 시장 규모는 전년 대비 14.5% 성장한 22조 8,986억 원이며, 2022년에는 25조 527억 원 규모로 성장할 것으로 나타났다. 2018년부터 2022년(잠정치)까지의 5개년 연평균 성장률(CAGR, Compound Annual Growth Rate)은 12.6%로 나타나면서 지속적인 성장세를 이어갈 것으로 조사되었다.

2022년 데이터산업은 총 3개 대분류와 10개 중분류로 구성된다. 부문별 시장 규모 잠정치는 '데이터 판매 및 제공 서비스업' 시장이 12조 5,235억 원으로 가장 높은 비중을 차지하며, 다음으로 '데이터 구축 및 컨설팅 서비스업'이 9조 2,570억 원, '데이터 처리 및 관리 솔루션 개발 · 공급업'이 3조 2,723억 원으로 예상된다.

1) '2022년 데이터산업 현황조사'는 국가승인통계 127004호로 연간 발간되며, 해당 보고서는 2022년 6월말 기준으로 하반기 중 조사된 결과를 수록하여 2023년 4월에 발표하였다.

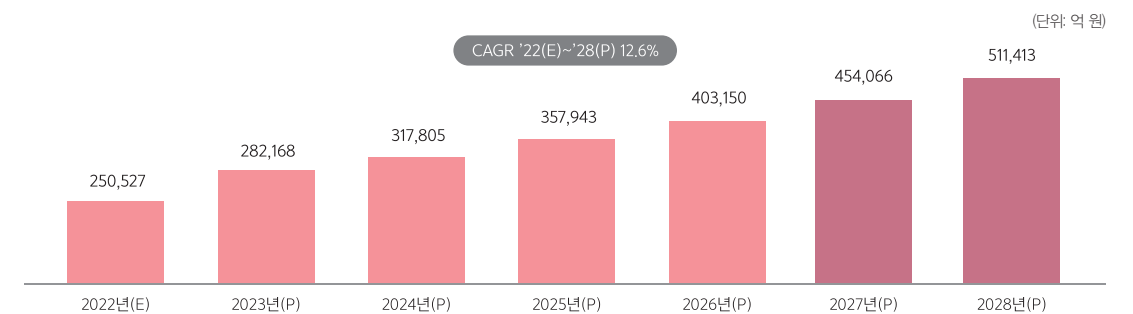
2) 본 장에서 '데이터산업'을 데이터의 생산·수집·처리·분석·유통·활용 등을 통해 가치를 창출하는 상품 및 서비스를 생산·제공하는 산업으로 정의한다.

[그림 3-1-1] 국내 데이터산업 시장 규모(2018~2022년(E))



국내 데이터산업 시장이 지난 5개년 연평균 성장률인 12.6%와 같이 지속적으로 성장한다면 2028년(추정치)까지 51조 원을 넘어설 것으로 전망된다.

[그림 3-1-2] 국내 데이터산업 시장 전망(2022(E)~2028년(P))



* E: 잠정치, P: 추정치

2. 국내 데이터산업 부문별 시장 규모

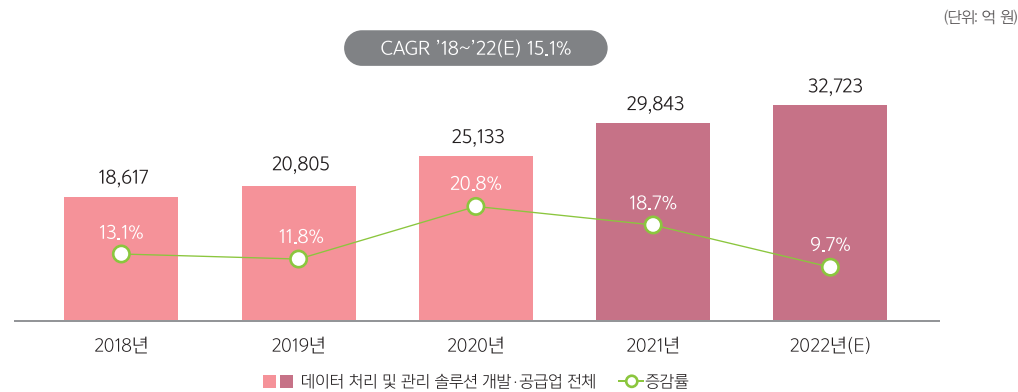
가. 데이터 처리 및 관리 솔루션 개발 · 공급업 시장

'데이터 처리 및 관리 솔루션 개발 · 공급업'은 '데이터 수집 · 연계 솔루션', '데이터베이스 관리 시스템 솔루션', '데이터 분석 솔루션', '데이터 관리 솔루션', '데이터 보안 솔루션', '빅데이터 통합 플랫폼 솔루션'을 포함하여 6개 중

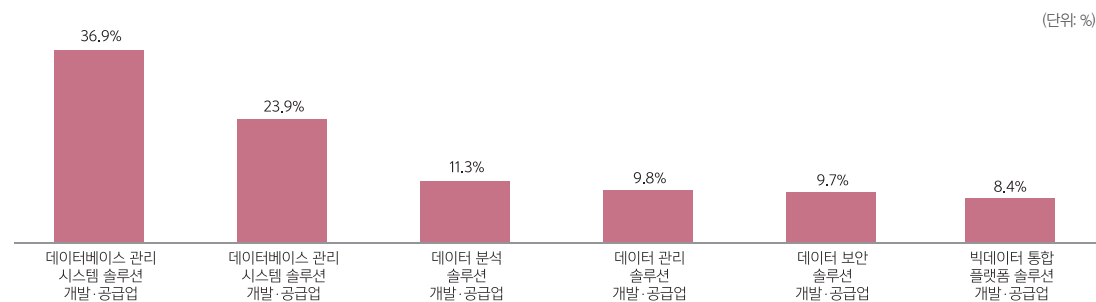
분류로 구분된다. 이들은 주로 솔루션 제품을 판매하는 비즈니스를 의미하고 라이선스, 개발 및 커스터마이징, 유지 보수 등을 통해 매출을 올리고 있다.

2022년 '데이터 처리 및 관리 솔루션 개발 · 공급업' 부문의 시장 규모 잠정치는 전년 대비 9.7% 성장한 3조 2,723억 원으로 전망되었으며, 2018년부터 5개년 연평균 성장률은 15.1%로 예상되었다.

[그림 3-1-3] 국내 데이터 처리 및 관리 솔루션 개발 · 공급업 시장 규모(2018~2022년(E))



[그림 3-1-4] 2021년 데이터 처리 및 관리 솔루션 개발 · 공급업 중분류별 시장 규모 비중



2022년도 부문별 예상 매출액은 '데이터베이스 관리 시스템 솔루션 개발 · 공급업'이 약 1조 1,787억 원, '데이터 관리 솔루션 개발 · 공급업'이 약 7,963억 원 순으로 높은 비중을 차지할 것으로 예상된다.

[표 3-1-1] 국내 데이터 처리 및 관리 솔루션 개발 · 공급업 중분류별 시장 규모(2018~2022년(E))

(단위: 억 원)

구 분	2018년	2019년	2020년	2021년	2022년 (E)	증감률 '20~'21	CAGR '20~'22 (E)
데이터 수집 · 연계 솔루션 개발 · 공급업	1,622	1,871	2,122	2,499	2,715	17.8%	13.1%
데이터베이스 관리 시스템 솔루션 개발 · 공급업	6,775	7,510	8,979	11,021	11,787	22.7%	14.6%

(계속 →)

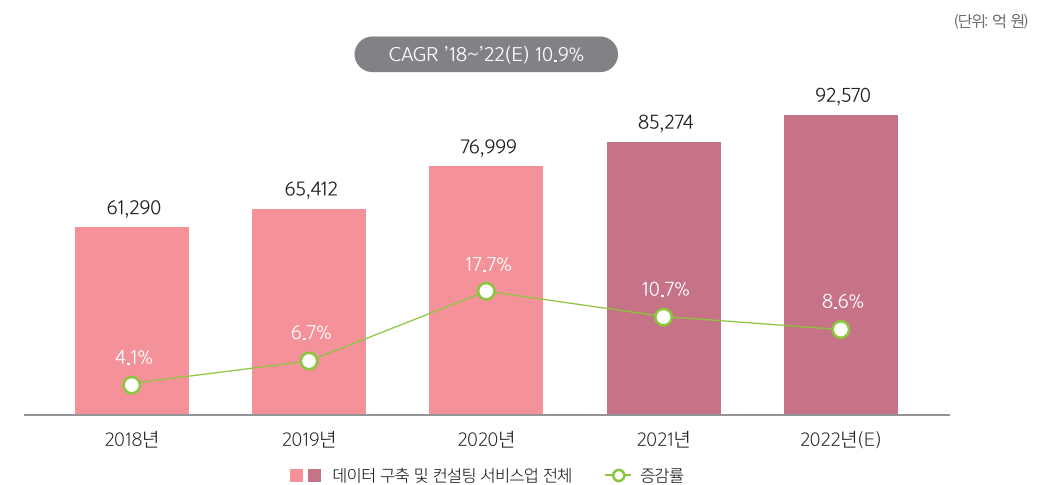
데이터 분석 솔루션 개발 · 공급업	1,782	2,014	2,586	2,932	3,247	13.4%	12.1%
데이터 관리 솔루션 개발 · 공급업	4,972	5,203	6,022	7,137	7,963	18.5%	15.0%
데이터 보안 솔루션 개발 · 공급업	1,517	1,975	2,558	2,894	3,015	13.1%	8.6%
빅데이터 통합 플랫폼 솔루션 개발 · 공급업	1,949	2,231	2,866	3,359	3,995	17.2%	18.1%
전체	18,617	20,805	25,133	29,843	32,723	18.7%	14.1%

나. 데이터 구축 및 컨설팅 서비스업 시장

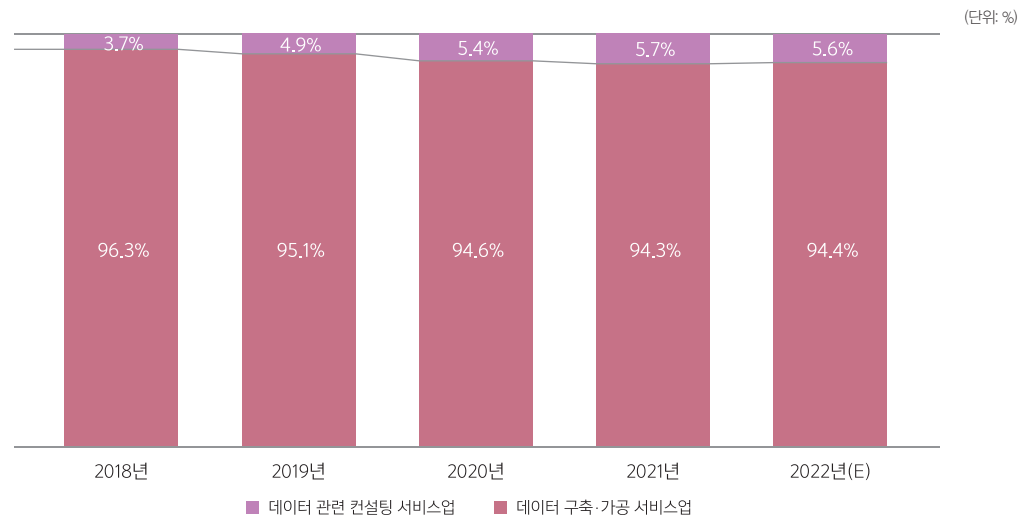
'데이터 구축 및 컨설팅 서비스업'은 '데이터 구축 · 가공 서비스업'과 '데이터 관련 컨설팅 서비스업'으로 구분된다. '데이터 구축 · 가공 서비스업'은 '데이터베이스 설계 · 구축 서비스', '데이터 이행 서비스', '데이터 가공 서비스'를 포함하고, '데이터 관련 컨설팅 서비스업'은 '데이터 설계 컨설팅', '데이터 품질 컨설팅', '데이터베이스 성능개선 컨설팅', '데이터 거버넌스 컨설팅', '데이터 분석 · 활용 컨설팅'을 의미한다. 비즈니스 매출은 주로 데이터 구축 · 개발, 유지보수 · 운영관리, 컨설팅을 통해 발생한다.

2022년 잠정적인 '데이터 구축 및 컨설팅 서비스업' 시장 규모는 9조 2,570억 원으로 전년 대비 8.6% 성장할 것으로 나타났다. 이 중 '데이터 구축 · 가공 서비스업' 시장 규모는 8조 7,366억 원이다. 이는 전체 '데이터 구축 및 컨설팅 서비스업' 시장의 94.4% 수준으로, '데이터 관련 컨설팅 서비스업'은 5.6% 수준인 5,204억 원으로 나타났다. '데이터 구축 및 컨설팅 서비스업' 시장의 2018년부터 5개년 연평균 성장률은 10.9%로 조사되었다.

[그림 3-1-5] 국내 데이터 구축 및 컨설팅 서비스업 시장 규모(2018~2022년(E))



[그림 3-1-6] 국내 데이터 구축 및 컨설팅 서비스업 중분류별 시장 규모 비중(2018~2022년(E))



[표 3-1-2] 국내 데이터 구축 및 컨설팅 서비스업 중분류별 시장 규모(2018~2022년(E))

(단위: 억 원)

구 분	2018년	2019년	2020년	2021년	2022년 (E)	증감률 '20~'21	CAGR '20~'22 (E)
데이터 구축·가공 서비스업	58,993	62,223	72,805	80,403	87,366	10.4%	9.5%
데이터 관련 컨설팅 서비스업	2,297	3,189	4,194	4,871	5,204	16.1%	11.4%
전체	61,290	65,412	76,999	85,274	92,570	10.7%	9.6%

다. 데이터 판매 및 제공 서비스업 시장

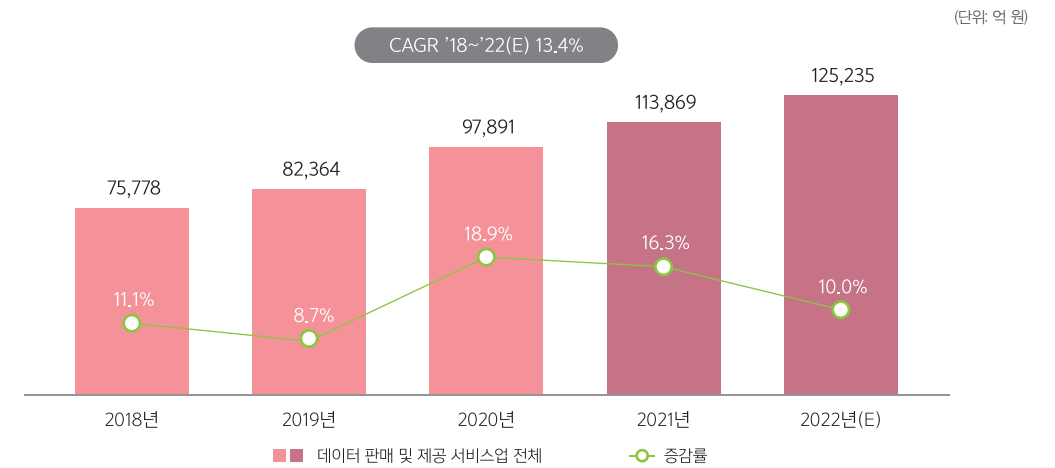
‘데이터 판매 및 제공 서비스업’은 원천 데이터와 데이터베이스를 수요자 맞춤형으로 가공·활용·분석하여 제공하는 비즈니스이다. 주로 데이터 이용료·수수료 등의 직접매출과 광고료 등의 간접매출로 수익을 창출한다.

세부 분류는 ‘데이터 판매·중개 서비스업’과 ‘정보제공 서비스업’으로 나뉜다. ‘데이터 판매·중개 서비스업’은 데이터 판매 서비스와 데이터 중개 서비스, 분석 데이터 제공 서비스를 포함하고, ‘정보제공 서비스업’은 포털·정보 매개 서비스와 정보제공 서비스를 포함한다.

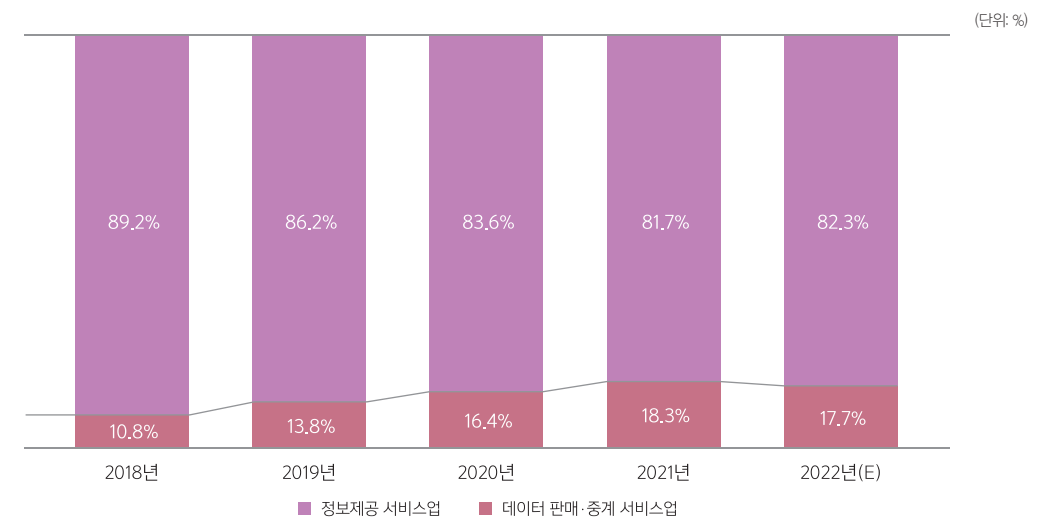
2022년 잠정적인 ‘데이터 판매 및 제공 서비스업’ 시장 규모는 전체 12조 5,235억 원으로, 전년 대비 10.0% 성장한 것으로 보인다. 2018년부터 2022년 예상 매출까지 5개년 연평균 성장률은 13.4%로 나타났다.

2022년 세부 부문별로는 ‘데이터 판매·중개 서비스업’ 시장은 2조 2,194억 원, ‘정보제공 서비스업’ 시장은 10조 3,040억 원으로 예상된다. ‘데이터 판매 및 제공 서비스업’ 시장에서 정보제공 서비스업의 비중이 82.3%를 차지한다.

[그림 3-1-7] 국내 데이터 판매 및 제공 서비스업 시장 규모(2018~2022년(E))



[그림 3-1-8] 국내 데이터 판매 및 제공 서비스업 중분류별 시장 규모 비중(2018~2022년(E))



[표 3-1-3] 국내 데이터 판매 및 제공 서비스업 중분류별 시장 규모(2018~2022년(E))

(단위: 억 원)

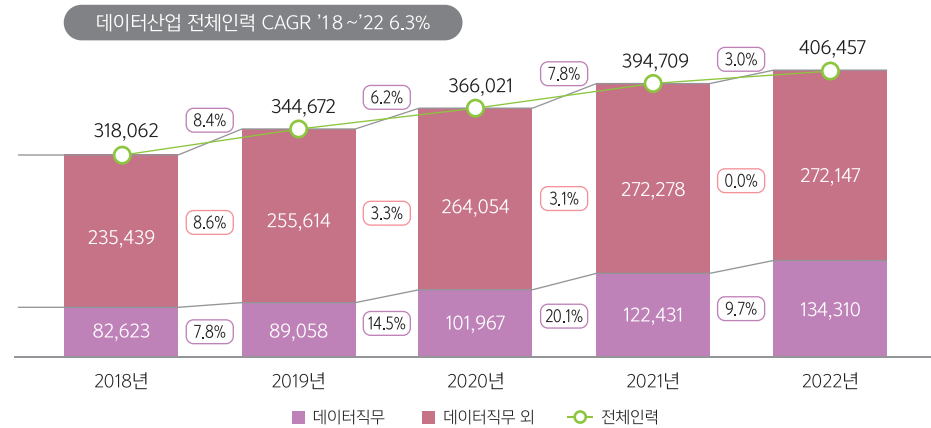
구 분	2018년	2019년	2020년	2021년	2022년 (E)	증감률 '20~'21	CAGR '20~'22 (E)
데이터 판매·중개 서비스업	8,198	11,332	16,054	20,861	22,194	29.9%	17.6%
정보제공 서비스업	67,580	71,033	81,838	93,008	103,040	13.6%	12.2%
전체	75,778	82,364	97,891	113,869	125,235	16.3%	13.1%

3. 국내 데이터 직무 인력 현황

가. 데이터산업 내 데이터 직무 인력 현황

2022년 데이터산업에 종사하는 전체 인력은 전년 대비 3.0% 증가한 40만 6,457명이다. 이 중 데이터 직무 인력은 13만 4,310명으로, 전체 데이터산업 종사자의 33%이며, 데이터산업 내 데이터 직무 인력은 전년 대비 9.7% 증가하였다.

[그림 3-1-9] 국내 데이터산업 전체 인력 현황(2018~2022년)



2022년 부문별로 데이터 직무 인력 현황과 추이를 살펴보면, '데이터 구축 및 컨설팅 서비스업' 분야의 데이터 직무 인력수가 6만 4,248명으로 가장 많았다. 반면, 전년 대비 데이터 직무 인력 비율이 가장 많이 증가한 분야는 '데이터 판매 및 제공 서비스업'으로, 2021년 대비 11.9% 증가하였다.

[표 3-1-4] 국내 데이터산업 부문별 데이터 직무 인력 현황(2018~2022년)

(단위: 명)

구 분		2018년	2019년	2020년	2021년	2022년	증감률 '21~'22	CAGR '18~'22
데이터 처리 및 관리 솔루션 개발·공급업	인력수	11,541	13,467	17,273	22,124	23,545	6.4%	19.5%
	비중	14.0%	15.1%	16.9%	18.1%	17.5%		
데이터 구축 및 컨설팅 서비스업	인력수	40,197	42,979	48,644	58,733	64,248	9.4%	12.4%
	비중	48.7%	48.3%	47.7%	48.0%	47.8%		

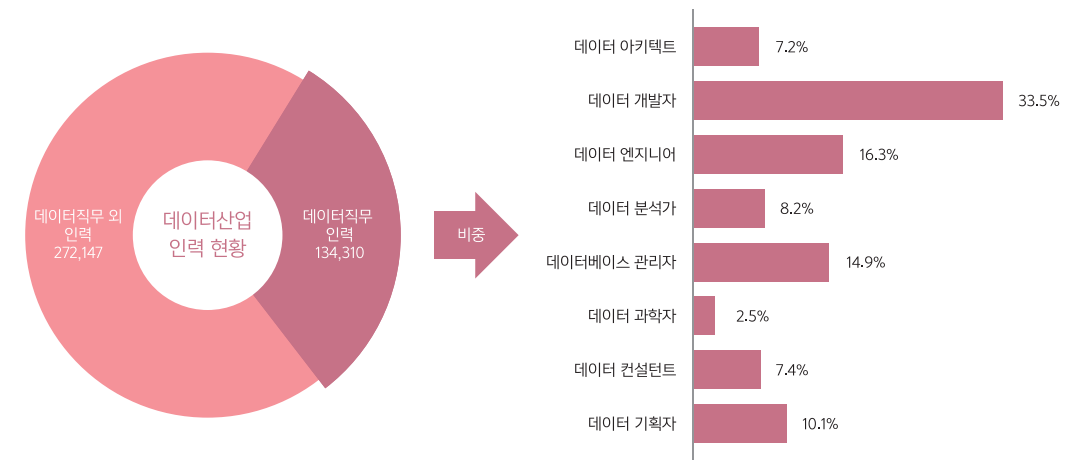
(계속 →)

3) 데이터 개발자, 데이터 엔지니어, 데이터 분석가, 데이터베이스 관리자, 데이터 과학자, 데이터 컨설턴트, 데이터 아키텍트, 데이터 기획자 8개 직무를 기준으로 한다.

데이터 판매 및 제공 서비스업	인력수	30,885	32,611	36,050	41,574	46,517	11.9%	10.8%
	비중	37.4%	36.6%	35.4%	34.0%	34.6%		
전체	인력수	82,623	89,058	101,967	122,431	134,310	9.7%	12.9%
	비중	100.0%	100.0%	100.0%	100.0%	100.0%		

데이터산업 인력의 직무별 비중을 살펴보면, 데이터 개발자가 4만 4,977명(33.5%)으로 가장 큰 비중을 차지하였고, 데이터 엔지니어 2만 1,923명(16.3%), 데이터베이스 관리자 1만 9,961명(14.9%) 순으로 나타났다.

[그림 3-1-10] 국내 데이터산업 인력 구성 및 데이터 직무별 인력 비중(2022년)



[표 3-1-5] 국내 데이터산업의 데이터 직무별 인력 현황(2022년)

(단위: 명)

구 분		데이터 처리 및 관리 솔루션 개발·공급업	데이터 구축 및 컨설팅 서비스업	데이터 판매 및 제공 서비스업	데이터산업 전체
데이터 아키텍트	인력수	2,067	6,227	1,416	9,711
	비중	8.8%	9.7%	3.0%	7.2%
데이터 개발자	인력수	8,286	22,458	14,233	44,977
	비중	35.2%	35.0%	30.6%	33.5%
데이터 엔지니어	인력수	3,375	11,549	6,999	21,923
	비중	14.3%	18.0%	15.0%	16.3%
데이터 분석가	인력수	1,836	4,815	4,323	10,974
	비중	7.8%	7.5%	9.3%	8.2%
데이터베이스 관리자	인력수	1,867	9,122	8,972	19,961
	비중	7.9%	14.2%	19.3%	14.9%

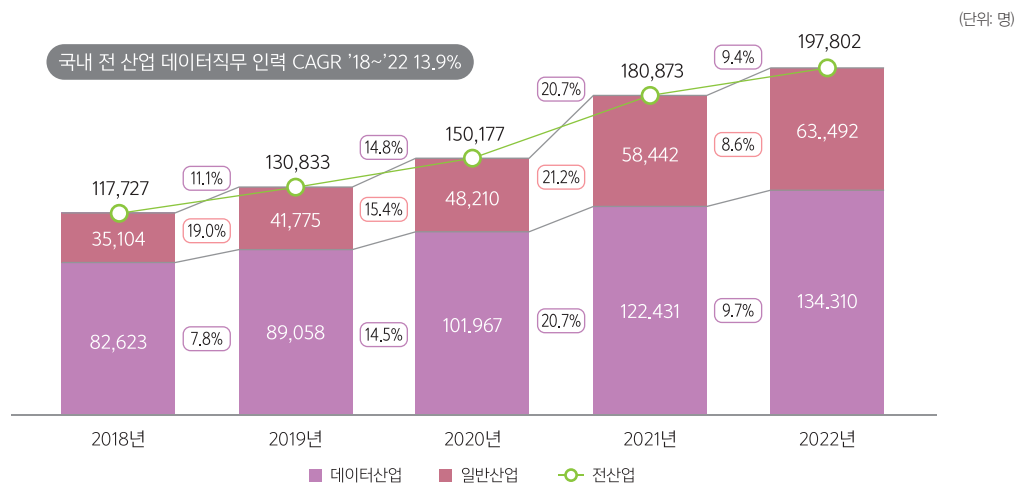
(계속 →)

구 분		데이터 처리 및 관리 솔루션 개발 · 공급업	데이터 구축 및 컨설팅 서비스업	데이터 판매 및 제공 서비스업	데이터산업 전체
데이터 과학자	인력수	706	1,647	945	3,297
	비중	3.0%	2.6%	2.0%	2.5%
데이터 컨설턴트	인력수	2,833	5,251	1,825	9,909
	비중	12.0%	8.2%	3.9%	7.4%
데이터 기획자	인력수	2,574	3,179	7,805	13,558
	비중	10.9%	4.9%	16.8%	10.1%
전체	인력수	23,545	64,248	46,517	134,310
	비중	100.0%	100.0%	100.0%	100.0%

나. 전 산업 내 데이터 직무 인력 현황

데이터산업과 일반 산업을 포함한 2022년 전 산업의 데이터 직무 인력은 총 19만 7,802명으로, 전년 대비 9.4% 증가하였다. 데이터산업을 제외한 일반 산업의 데이터 직무 인력은 6만 3,492명으로 전년 대비 8.6% 증가한 것으로 나타났다.

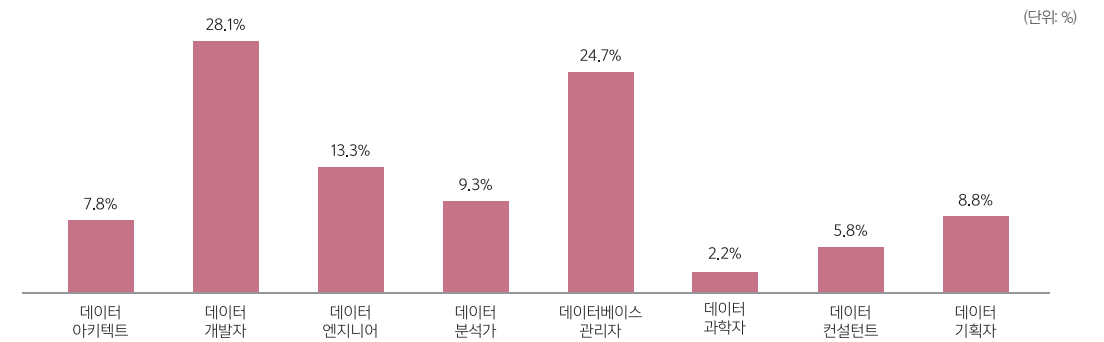
[그림 3-1-11] 국내 전 산업 데이터 직무 인력 현황(2018~2022년)



전 산업의 데이터 직무별 인력에서는 데이터 개발자가 5만 5,509명(28.1%)으로 가장 높은 비중을 차지했고, 데이터베이스 관리자 4만 8,954명(24.7%), 데이터 엔지니어 2만 6,223명(13.3%) 순으로 나타났다. 데이터산업과 일반 산업⁴⁾의 데이터 직무별 인력 비중을 살펴보면, 데이터산업은 데이터 개발자(33.5%)가 많지만, 일반 산업은 데이터베이스 관리자(45.7%)가 많은 것으로 나타났다.

4) 일반 산업은 데이터산업 외 13개 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 의미한다.

[그림 3-1-12] 국내 전 산업 데이터 직무별 인력 비중(2022년)



[표 3-1-6] 국내 전 산업 데이터 직무별 인력 현황(2022년)

구 분	데이터산업		일반산업		전 산업	
	인력수	비중	인력수	비중	인력수	비중
데이터 아키텍트	9,711	7.2%	5,805	9.1%	15,515	7.8%
데이터 개발자	44,977	33.5%	10,532	16.6%	55,509	28.1%
데이터 엔지니어	21,923	16.3%	4,300	6.8%	26,223	13.3%
데이터 분석가	10,974	8.2%	7,435	11.7%	18,410	9.3%
데이터베이스 관리자	19,961	14.9%	28,992	45.7%	48,954	24.7%
데이터 과학자	3,297	2.5%	1,043	1.6%	4,340	2.2%
데이터 컨설턴트	9,909	7.4%	1,556	2.5%	11,466	5.8%
데이터 기획자	13,558	10.1%	3,829	6.0%	17,387	8.8%
전체	134,310	100.0%	63,492	100.0%	197,802	100.0%

4. 국내 데이터 직무 인력 수요

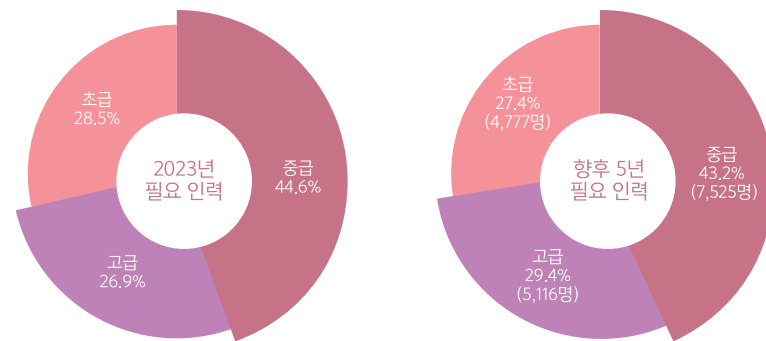
가. 데이터산업의 데이터 직무 인력 수요 및 부족률

2027년까지 향후 5년 내 데이터산업의 데이터 직무 필요 인력⁵⁾은 총 1만 7,418명으로 조사되었다. 직무별 수요는 데이터 개발자가 7,772명(44.6%)으로 가장 높았고, 데이터 엔지니어 1,955명(11.2%), 데이터 분석가 1,874명(10.8%) 순으로 나타났다.

향후 5년 내 데이터산업의 기술등급별 데이터 직무 필요 인력 비중을 살펴보면 현재와 마찬가지로 중급 인력에 대한 수요가 7,525명(43.2%)으로 가장 높게 나타났다.

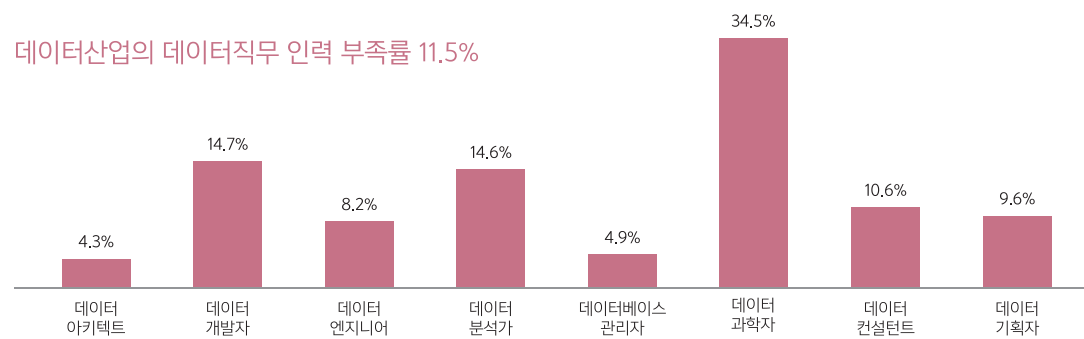
5) 필요 인력은 기업에서 현재 인력보다 추가로 필요한 인력수를 의미한다. 즉, 현재 부족한 인력수를 의미하며, 채용 계획 인력수는 아니다.

[그림 3-1-13] 국내 데이터산업의 향후 5년 내 기술등급별 데이터 직무 필요 인력 비중



데이터산업 내 데이터 직무 평균 부족률⁶⁾은 11.5%이다. 이 중 데이터 과학자의 부족률이 34.5%로 가장 높게 나타났다. 이어서 데이터 개발자(14.7%), 데이터 분석가(14.6%) 순으로 나타났다.

[그림 3-1-14] 국내 데이터산업의 향후 5년 내 데이터 직무 인력 부족률



* 통계 결과는 반올림되어 부분의 합계가 전체와 일치하지 않을 수 있음

나. 전 산업의 필요 인력 및 부족률⁷⁾

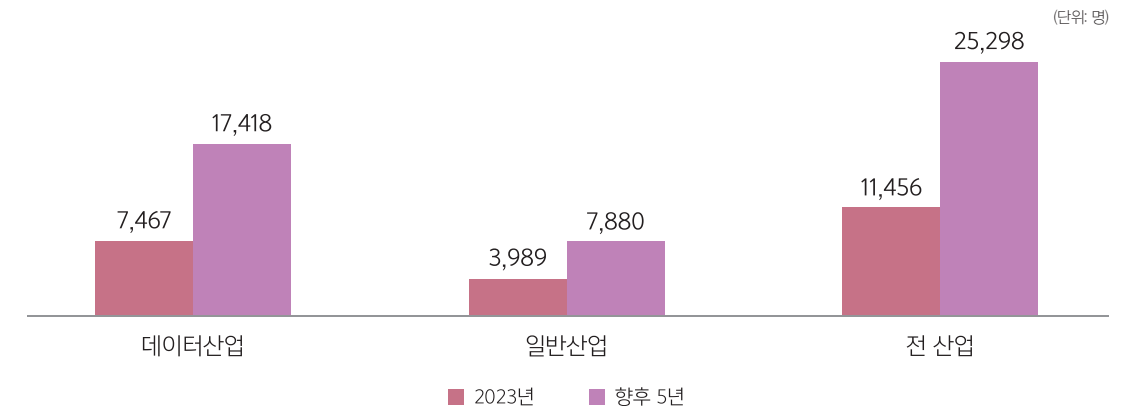
향후 5년 내 일반 산업을 포함한 전 산업에서 필요한 데이터 직무 인력은 총 2만 5,298명이다. 일반 산업은 데이터산업에 비하여 데이터 직무 필요 인력이 적었다. 향후 5년 내 전 산업 필요 인력 중 데이터 개발자가 9,032명으로 가장 크게 나타났고, 데이터베이스 관리자 4,329명, 데이터 분석가 2,930명, 데이터 엔지니어 2,724명 순으로 나타났다.

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

7) 부족률: {필요 인력 / (현재인력 + 필요 인력)} × 100

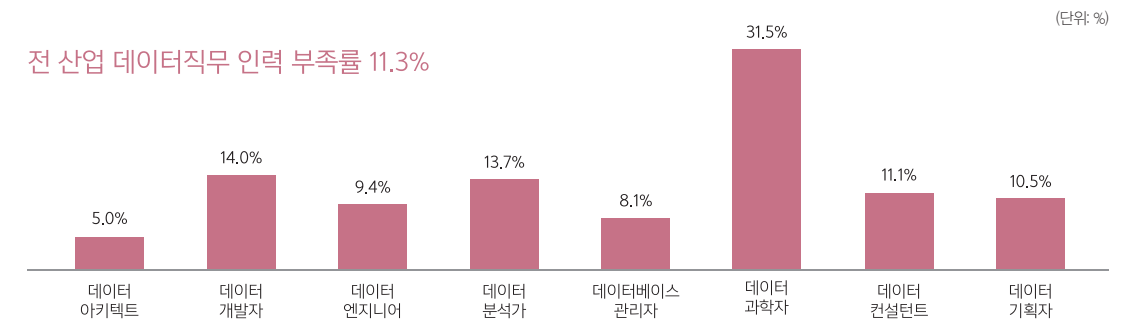
예) 현재인력 8명, 필요 인력 2명, 인력부족률은 2 / (8 + 2) × 100 = 20%

[그림 3-1-15] 국내 데이터산업/일반 산업/전 산업의 향후 5년 내 데이터 직무 필요 인력



향후 5년 내 일반 산업을 포함한 전 산업 내 데이터 직무별 인력 부족률은 평균 11.3% 수준이며, 데이터 과학자 부족률이 31.5%로 가장 높게 나타났다. 그리고 데이터 개발자(14.0%), 데이터 분석가(13.7%) 직무가 전 산업 평균보다 높은 부족률을 보였다.

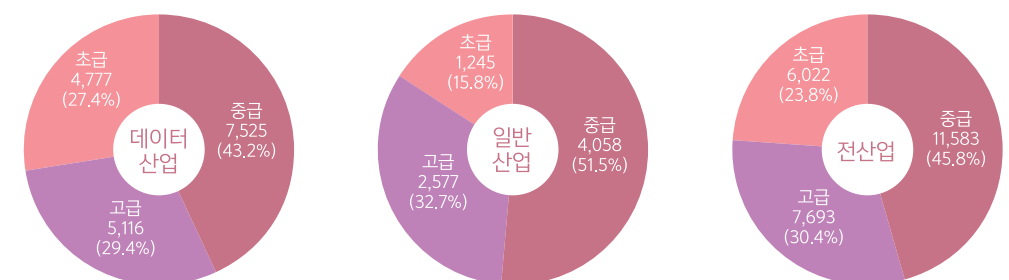
[그림 3-1-16] 국내 전 산업의 향후 5년 내 데이터 직무 인력 부족률



* 통계 결과는 반올림되어 부분의 합계가 전체와 일치하지 않을 수 있음

향후 5년 내 일반 산업을 포함한 전 산업에서 추가로 필요한 기술등급별 인력은 중급이 1만 1,583명(45.8%)으로 가장 높게 나타났고 고급 7,693명(30.4%), 초급 6,022명(23.8%) 순으로 나타났다.

[그림 3-1-17] 국내 전 산업의 향후 5년 내 기술등급별 데이터 직무 필요 인력 비중



* 통계 결과는 반올림되어 부분의 합계가 전체와 일치하지 않을 수 있음

5. 마무리

한국은행의 '2021년 기업경영분석'에 따르면 국내 전 산업의 연평균 성장률은 2017년부터 2021년까지 4.8%로 나타났다. 동일 기간, 데이터산업과 유사한 정보통신서비스업은 3.7%, 정보통신기술산업은 4.2%, 서비스업은 6.4% 수치를 보였다. 데이터산업은 12.4%로 비교적 높은 증가세를 보이고 있다.

[표 3-1-8] 국내 주요 산업별 시장 규모 추이(2017~2021년)

(단위: 억 원)

구분	2017년	2018년	2019년	2020년	2021년	CAGR '17~'21
전산업*	39,914,871	41,516,088	41,411,615	41,169,953	48,123,205	4.8%
제조업*	18,218,781	19,007,562	18,547,032	18,172,188	21,590,555	4.3%
서비스업*	15,880,736	16,526,281	16,995,849	17,018,094	20,319,920	6.4%
정보통신 기술산업*	6,060,007	6,471,342	6,243,279	6,067,018	7,146,479	4.2%
정보통신 서비스업*	2,183,098	2,393,394	2,517,847	2,203,726	2,519,994	3.7%
데이터산업**	143,530	155,684	168,582	200,024	228,986	12.4%

* 한국은행, 2021년 기업경영분석(2022년 12월)

** 한국데이터산업진흥원, 2022 데이터산업 현황조사(2023년4월)

데이터산업은 디지털 경제의 원천이며, AI, 클라우드컴퓨팅 등 디지털 기술 및 ICT 인프라를 바탕으로 데이터가 축적됨에 따라 성장세 지속 유지할 것으로 전망된다. 전 산업에서 데이터와 융합된 디지털 전환(Digital Transformation)과 이를 위한 데이터산업 생태계 강화 및 거버넌스 체계 구축이 강조되면서, 데이터 직무 인력에 대한 수요 역시 증대되고 있다.

위 조사 결과와 같이 중고급 이상의 기술 수준을 보유한 데이터 직무 인력에 대한 공급은 여전히 부족한 실정이나, 전년 대비 데이터산업의 데이터직무 인력 부족률(12.2%→11.5%) 및 전 산업 데이터직무 인력 부족률(11.9%→11.3%)은 각각 0.7%p, 0.6%p가 감소하였다. 이는 2022년 4월 '데이터 산업진흥 및 이용촉진에 관한 기본법' 시행을 통해 데이터의 생산, 거래 및 활용 촉진에 관한 필요 사항을 정함으로써 데이터산업 발전의 기반을 조성하려는 정부 정책 방향에 어느 정도 영향을 받은 것으로 판단된다.

제2장

국내 데이터산업의 역동성 및 생산성 분석

고동환 박사 정보통신정책연구원 디지털경제연구실

제2장은 2020년부터 2022년까지 3개년도의 「데이터산업 현황조사」의 모집단을 기준연도와 사업자등록번호를 기준으로 「기업활동조사」와 연계하여 분석한 결과를 바탕으로 작성하였다. 사업체 단위의 조사인 데이터 산업 현황 조사와 법인 단위의 조사인 기업활동 조사의 차이로 인해 분석 결과에 다소 차이가 있을 수 있다. 따라서 이 자료만을 근거로 삼아 연구개발의 효율성, 시장의 역동성과 시장 집중, 생산성 둔화 등과 관련해 단정적인 가치판단을 내리는 것은 부적절하며, 더 많은 시계열자료와 데이터의 특성을 고려하여 신중하게 해석할 필요가 있다.

1. '데이터 산업 현황 조사'의 모집단 분석

가. 「데이터 산업 현황 조사」의 모집단

한국데이터산업진흥원에서 제공한 데이터산업 현황조사의 모집단에 따르면, 조사의 모집단은 2020년 전체 8,615개의 사업체에서 2021년과 2022년 각각 9,342개, 9,566개로 증가하여 데이터 관련 서비스를 제공하는 사업체 수가 전체적으로 증가하고 있음을 알 수 있다. 사업체가 가장 많이 증가한 산업은 '데이터 처리 및 관리 솔루션 개발·공급업'이며, 2021년과 2022년에 각각 전년대비 12.4%, 7.7% 증가하였고, 가장 적게 증가한 산업은 '데이터 구축 및 컨설팅 서비스업'으로 같은 기간 7.5%와 1.5% 증가에 그쳤다.

[표 3-2-1] 데이터 산업 현황 조사 모집단 사업체 수(2020~2022년)

대분류	구분	사업체 수		
	중분류	2020	2021	2022
데이터 처리 및 관리 솔루션 개발·공급업	데이터 수집·연계 솔루션 개발·공급업	518	551	622
	데이터베이스 관리 시스템 솔루션 개발·공급업	149	186	241
	데이터 분석 솔루션 개발·공급업	365	446	471
	데이터 관리 솔루션 개발·공급업	1,062	1,105	1,137
	데이터 보안 솔루션 개발·공급업	235	302	317
	빅데이터 통합 플랫폼 솔루션 개발·공급업	158	206	222
소계		2,487	2,796	3,010
데이터 구축 및 컨설팅 서비스업	데이터 구축·가공 서비스업	2,158	2,229	2,229
	데이터 관련 컨설팅 서비스업	593	729	773
소계		2,751	2,958	3,002
데이터 판매 및 제공 서비스업	데이터 판매·중개 서비스업	494	567	605
	정보제공 서비스업	2,282	2,419	2,323
소계		2,776	2,986	2,928
데이터 산업 합계		8,014	8,740	8,940

데이터 인프라 서비스업	데이터 저장 장치·시설 서비스업	267	270	283
	데이터 네트워크 인프라 서비스업	334	332	343
소계		601	602	626
총계(인프라 포함)		8,615	9,342	9,566

나. 「기업활동조사」의 데이터 기업 모집단

본 장을 작성하기 위해 통계청이 제공하는 「기업활동조사」의 마이크로데이터를 획득한 후 「데이터 산업 현황 조사」의 사업자등록번호 및 법인등록번호를 활용하여 「기업활동조사」에서 데이터 기업을 식별하였다. 「기업활동조사」에서는 기업단위의 총요소생산성을 추정하는 데 필요한 매출액, 종사자 수, 사업 영위 기간, 유형자산, 원재료, 투자, 인건비를 포함한 영업비용의 세부 항목을 제공하고 있다. 법인 내에 복수의 사업체가 존재하기 때문에 법인의 표준산업분류와 사업체의 산업분류가 다를 수 있다. 연계된 법인은 연도별로 600개에서 800개 사이이며, 산업별 분포도 「데이터 산업 현황 조사」에 포함되는 KSIC-58, 62, 63뿐 아니라 KSIC-26과 같은 제조업을 비롯하여 다양하게 분포되어 있다. 본 장에서는 산업 분포를 구체적으로 논의하지는 않고 연계된 데이터의 기초통계량을 제시하였다.

[표 3-2-2] 데이터 기업과 ICT 산업¹⁾의 기초통계량 비교

(단위 : 백만 원 명)

		평균	중앙값	하위 10%	하위 25%	상위 25%	상위 10%
데이터 산업	매출액	287,438	23,756	4,604	11,123	68,868	247,528
	영업이익	29,855	860	-3,816	17	4,403	17,423
	상용근로자	439	130	36	70	244	664
	일용직근로자	25	0	0	0	2	29
	유형자산	85,970	2,464	92	482	10,326	41,263
	무형자산	28,996	707	2	60	2,983	11,173
ICT 제조업	매출액	380,892	34,460	9,199	16,986	80,784	209,358
	영업이익	36,192	777	-7,317	-785	3,847	12,485
	상용근로자	502	121	60	79	215	437
	일용직근로자	7	0	0	0	0	6
	유형자산	154,651	9,959	1,659	4,048	23,808	58,377
	무형자산	11,336	333	0	19	1818	5,455
ICT 서비스업	매출액	106,244	16,644	3,546	7,388	42,399	130,745
	영업이익	9306	450	-3,024	-174	2,641	102,72
	상용근로자	251	95	28	56	180	383
	일용직근로자	11	0	0	0	0	11
	유형자산	43,839	1,570	77	289	7,743	28,277
	무형자산	13,389	475	0	27	2,205	6,879

1) KSIC-26 산업은 ICT 제조업으로, KSIC-58~KSIC-63 산업, 즉 정보통신업(J)은 ICT 서비스업으로 분류하였다.

2. 국내 데이터 기업의 역동성²⁾

가. 데이터 기업의 진입률과 퇴출률³⁾

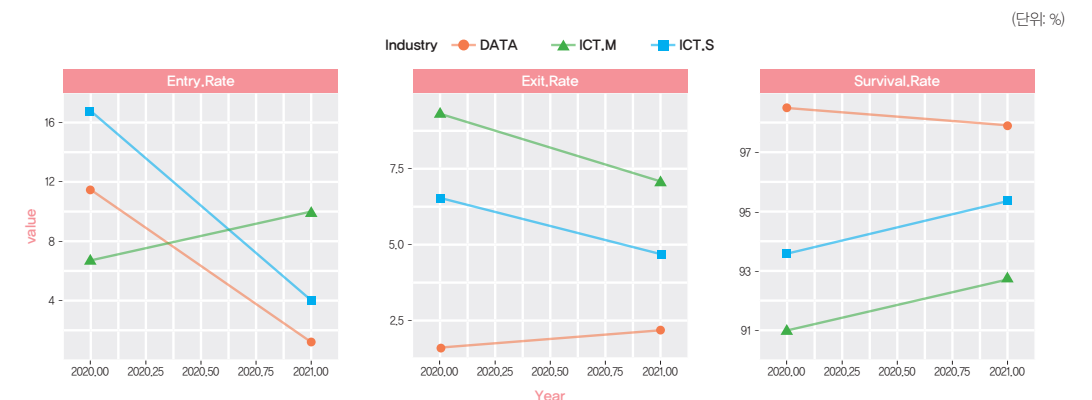
기업의 시장 진입과 퇴출 비율은 시장의 역동성을 나타내는 지표 중 하나로 시장의 규모가 커짐과 동시에 진입과 퇴출이 활발할수록 새로운 건전한 시장으로 볼 수 있다. 반대로 진입은 어렵고 퇴출이 지속된다면 생존 기업의 시장 장악력이 강화되고 있다는 점을 시사한다.

「기업활동조사」와 연계된 데이터 기업의 2020년 신규 기업의 진입률은 약 11.6%로, ICT 제조업의 6.7%보다는 크고 ICT 서비스 산업의 16.5%보다는 작다. 2020년 데이터 기업의 퇴출률은 1.5%로 ICT 제조업(9%)이나 ICT 서비스업(6.4%)보다 훨씬 낮은 수준이다. 그 결과, 데이터 기업의 생존율은 ICT 제조업이나 ICT 서비스업에 비해 약 5% 정도 높은 수준으로 나타났다. 특히 2020년에 발생한 코로나19의 부정적인 영향이 있었는데도, 생존율이 ICT 산업에 비해 높다는 점은 긍정적으로 해석할 수 있다.

그러나 증가율 관점에서는 2021년 데이터 기업의 진입률은 2020년에 비해 낮아졌지만, 퇴출률은 증가하여 생존 기업의 비중이 약 1%p 감소하였다. ICT 제조업은 반대로 진입기업 비중은 증가하고 퇴출 기업 비중은 감소하여 전년 대비 더 많은 기업이 시장에서 경쟁하고 있음을 알 수 있다.

일반적으로 경쟁이 활발한 시장에서는 기업 간 자원의 효율적 분배를 유도하고 생산성이 개선된다. 3개년도의 데이터만을 활용하였기 중장기적인 추세를 판단하기는 어려우나, 자산으로서의 데이터에 규모의 경제가 존재하고 네트워크 효과에 따른 플랫폼 기업의 시장 장악력 강화를 감안하여 데이터 산업의 경쟁 상황과 역동성을 계속 지켜볼 필요가 있다.

[그림 3-2-1] 데이터 기업의 시장 진입과 퇴출, 생존율 추이



* 붉은 색 실선은 데이터 산업, 파란색선은 ICT서비스업(KSIC-58 ~ KSIC 63), 초록색은 ICT 제조업(KSIC-26)을 나타냄

2) 「기업활동조사」와 「데이터 산업 현황 조사」는 연간 조사이므로 분석의 시간 단위(t)는 연(Year)이다.

3)
$$\text{진입률}_t = 100 \times \frac{\text{신규진입기업수}_t}{\text{생존기업}_t + \text{신규진입기업수}_t}$$

$$\text{퇴출률}_t = 100 \times \frac{\text{퇴출기업기업수}_{t+1}}{\text{생존기업}_t + \text{신규진입기업수}_t}$$

$$\text{생존율}_t = 100 \times \frac{\text{생존기업수}_{t+1}}{\text{생존기업}_t + \text{신규진입기업수}_t}$$

3. 국내 데이터 기업의 생산성 분석

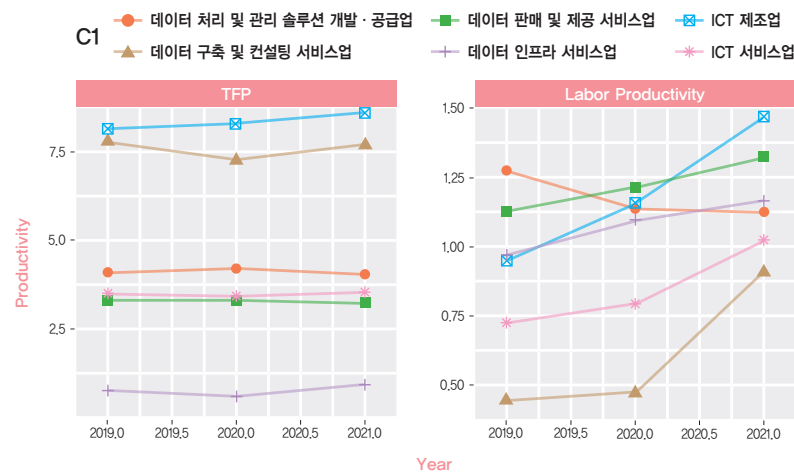
가. 생산성 추정 방법

본 장에서는 Akerberg et al.(2015)⁴⁾이 제안한 준모수 추정 방법을 활용하여 표준산업분류 중분류(KSIC-2digit) 단위의 생산함수를 추정하고 기업별 총요소생산성(TFP)을 도출하였다. 기업 단위의 노동생산성은 부가가치 가산법에 따라 산출된 기업의 부가가치를 종종사자 수로 나누어 산출하였다. 추정된 기업 단위의 생산성은 다시 데이터 기업과 ICT 제조 기업, ICT 서비스 기업으로 구분⁵⁾하고, 매출액을 기준으로 산업별로 가중 평균하여 산업별 생산성을 산출하였다. 데이터 기업은 또한 데이터 산업의 분류체계에 따라 대분류(C1)와 중분류(C2)로도 구분하여 분석하였다.

나. 데이터 산업 대분류 수준에서의 생산성 추이

데이터 산업 대분류 단위에서 총요소생산성을 비교한 결과, 데이터 산업 내에서는 '데이터 구축 및 컨설팅 서비스업'의 생산성이 가장 높았고, '데이터 인프라 서비스업'이 가장 낮게 나타났다. ICT 제조업은 데이터 산업보다도 더 높은 것으로 추정되었으며 최근 데이터 산업과의 격차가 확대되고 있음을 확인할 수 있다. 데이터 산업은 전반적으로 2019년에 비해 2021년에 총요소생산성이 소폭 하락하였으나, '데이터 인프라 서비스업'은 오히려 0.82에서 1.04로 증가하였다.

[그림 3-2-2] 데이터 산업 대분류별 총요소생산성과 노동생산성 추이



* TFP (Total Factor Productivity: 총요소생산성)과 Labor Productivity (노동생산성)는 각각 로그값

4) Akerberg, D. A., Caves, K., & Frazer, G. (2015). Identification properties of recent production function estimators. *Econometrica*, 83(6), 2411-2451.

5) 데이터 기업은 표준산업분류 체계에 따라 ICT 제조업이나 ICT 서비스업으로 분류될 수 있으며, 이에 따라 산업별 생산성 계산 시 중복되어 포함될 수 있다. 예를 들어, A 기업은 KSIC 기준으로는 ICT 제조업에 속하지만, 데이터 기업에도 포함될 수 있다.

데이터 산업의 노동생산성은 2019년과 2021년에 다소 큰 차이를 보였는데, 2019년에는 데이터 산업 내에서 '데이터 처리 관리 및 솔루션 개발·공급업'이 가장 높고 '데이터 판매 및 제공 서비스업'이 그 뒤를 이었으나, 2021년에는 '데이터 처리 관리 및 솔루션 개발·공급업'이 가장 높았고 '데이터 인프라 서비스업'이 그 뒤를 이었다. 총요소생산성이 가장 높았던 '데이터 구축 및 컨설팅 서비스업'의 노동생산성은 오히려 가장 낮게 나타났다. 노동생산성과 총요소생산성의 차이는 상대적인 자본집약도의 차이로 해석할 수 있는데⁶⁾, 총요소생산성에 비해 노동생산성이 낮다면 상대적으로 자본집약도가 떨어진 것이 중요한 원인이 될 수 있다. ICT 제조업과 ICT 서비스업의 경우 데이터 산업에 비해 지난 3년간 노동생산성이 더 빠르게 증가한 것으로 나타난다.

생산성 분석에 따르면, 대부분의 데이터 기업이 ICT 서비스 산업에도 속한다는 점을 감안할 때, 전반적으로 데이터 기업의 생산성이 ICT 서비스업에 비해서 생산성이 더 높다는 점도 확인할 수 있다. 또한, 2020년에 생산성이 다소 주춤했으나 2021년 다시 회복하거나 추가적인 하락이 나타나지 않는 것으로 보아 코로나 팬데믹의 영향이 다른 산업에 비해 크지 않았던 것으로 보인다. 다만, '데이터 처리 및 관리 솔루션 개발·공급업'은 유일하게 총요소생산성과 노동생산성이 모두 하락하는 모습을 보여 지속적인 모니터링이 필요한 것으로 판단된다.

[그림 3-2-3] 데이터 산업 중분류별 총요소생산성 추이⁷⁾



6) 간단하게 콥더글라스 생산함수를 가정하는 경우 노동생산성 = $TFP + \beta \times \text{자본집약도}$ 로 나타낼 수 있다. 여기서 β 는 자본에 대한 부가가치 탄력도를 의미하며, 자본집약도는 1인당 자본이다. 그러나 본 장에서 추정된 생산함수는 콥더글라스 생산함수에 비해 제약을 더 완화하여 추정하였다. 예를 들어 수익불변(Constant Returns to Scale) 가정을 하지 않아 실제로는 자본집약도 외 노동 투입에 대한 부가가치 탄력도 등도 영향을 줄 수 있다.

7) 이 자료는 데이터 산업을 중분류 단위에서 생산성 추이를 그린 그래프인데, 대분류별로 스케일을 달리하였기 때문에 해석할 때 주의가 필요하다. 예를 들어 ICT 서비스업의 TFP 증가 속도가 다른 산업에 비해 매우 빠른 것처럼 보이나, 실제로는 소수점 두 자리 수준의 변화를 보인다.

4. 국내 데이터 산업의 혁신 활동

가. 데이터 산업 대분류 단위에서의 연구개발과 지적재산권 취득 추이

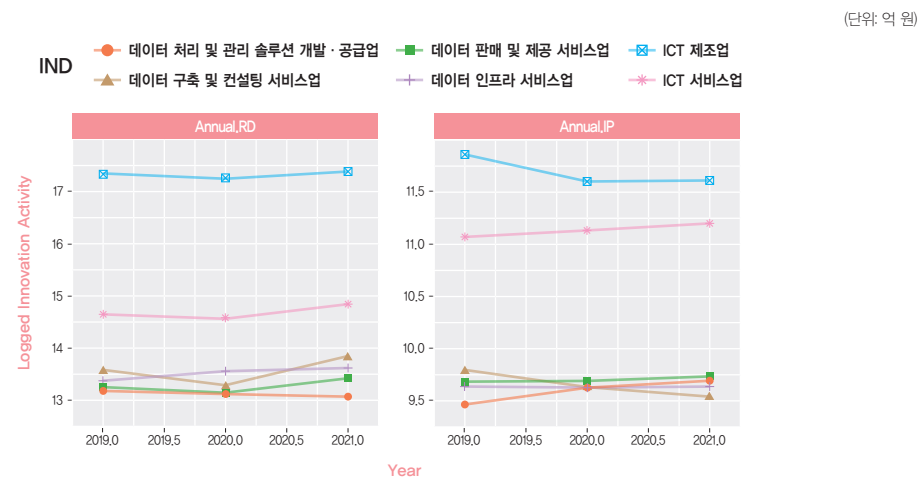
「기업활동조사」에는 연구개발활동과 관련하여 자체, 위탁, 위탁 연구개발 항목을 공개하고 있으며, 지적재산권에 관한 정보도 특허, 상표권, 디자인권, 실용신안권으로 구분하여 취득 현황을 제공하고 있다. 본 절에서는 데이터 기업의 혁신 활동을 연구개발비와 지적재산권 취득 현황을 통해 살펴보고, 이를 ICT 산업과 비교하였다. 연구개발비는 혁신 활동의 투입 요소로 지적재산권은 혁신 활동의 결과물로 해석할 수 있다.

[그림 3-2-4]는 각 기업의 연구개발비와 지적재산권을 연도별 대분류별 합계의 로그값을 그린 그래프이다. 혁신 활동은 ICT 제조업이 데이터 산업에 비해 월등히 많다는 점을 확인할 수 있다. ICT 제조업의 전체 연구개발비는 데이터 산업에 비해 14배가량 많은 것으로 나타났다. 반면, ICT 서비스업의 연구개발비에 비해서는 소폭(약 1.1배) 더 많은 것으로 나타났다.

데이터 산업 내에서는 '데이터 구축 및 컨설팅 서비스업'과 '데이터 판매 및 제공 서비스업'의 연구개발비가 증가하고 있으나 다른 세부 산업에서는 뚜렷한 증가세가 나타나지 않고 있다. 특히, '데이터 처리 및 관리 솔루션 개발 · 공급업' 분야에서는 연구개발비 하락이 지속되고 있다.

지적재산권 취득 측면에서 데이터 기업의 대부분은 증가하고 있으나 '데이터 구축 및 컨설팅 서비스업'만 유일하게 감소하는 패턴이 뚜렷이 나타나 연구개발비의 증가 패턴과 대조를 이루고 있다. 3개년도 추이만을 보고 연구개발 투자의 효율성을 판단하기는 이르지만, 지속적인 모니터링을 통해 효율성 문제가 식별되는 경우 적절한 대응 정책 마련이 필요할 것으로 보인다.

[그림 3-2-4] 데이터 산업 대분류별 혁신 활동 추이

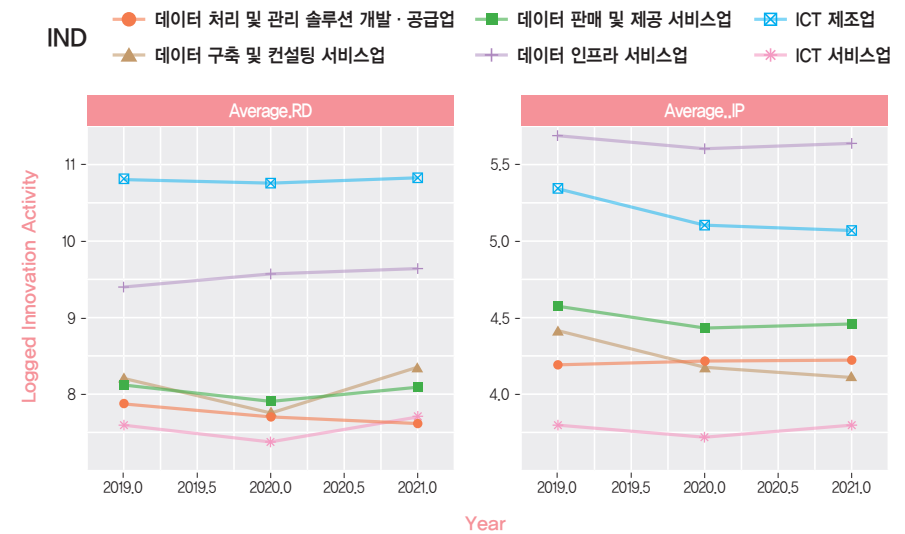


나. 데이터 산업 기업당 평균 연구개발과 지적재산권 취득 추이

산업 단위의 혁신 활동 총량은 지식의 유출효과(Spillover Effect) 등을 통해 산업 전체의 생산성에 영향을 줄 수 있다는 점에서 의미가 있다. 그러나 기업의 의사결정은 이러한 외부효과가 고려되기 어렵고, 기업의 이익 극대화를 위한 조건들을 평가한 후 혁신 활동의 수준이 정해지므로 기업 단위의 혁신 활동도 별도로 관찰할 필요가 있다.

[그림 3-2-5]는 데이터 산업 대분류 단위에서 기업당 평균 혁신 활동 수준을 그린 그래프이다. 산업별 총량의 비교와는 달리 연구개발비와 지적재산권 취득 수 측면에서 데이터 기업은 다른 ICT 서비스 기업에 비해 혁신 활동을 활발히 하는 것으로 나타났다. 특히 '데이터 인프라 서비스업'은 규모 면에서도 연구개발비 지출이 ICT 제조업에 근접한 수준이며, 지적재산권 취득 수는 오히려 더 많게 나타났다. '데이터 구축 및 컨설팅 서비스업'은 기업당 평균 수치에서도 연구개발비에 비해 지적재산권 취득이 감소하고 있다. 앞서 언급한 것처럼 연구개발의 효과를 지적재산권 취득 수로만 판단하는 것은 부적절하고, 지적재산권 취득도 시차를 두고 가능할 수도 있기 때문에 연구개발의 효율성을 현시점에서 선불리 판단하는 것은 부적절하다. 오히려 '데이터 처리 및 관리 솔루션 개발 · 공급업'과 같이 연구개발비의 하락이 지속되는 경우는 장기적 관점에서 문제를 일으킬 수 있다.

[그림 3-2-5] 데이터 산업 대분류별 기업당 평균 혁신 활동 추이



5. 마무리

본 장에서는 2020년부터 2022년까지 3개년도의 「데이터 산업 현황조사」의 모집단을 조사 기준연도와 사업자 등록번호를 기준으로 「기업활동조사」와 연계하여, 데이터 기업의 생산성과 데이터 산업의 역동성, 그리고 혁신 활동에 관해 분석을 시도하였다.

데이터 산업은 ICT 제조업이나 ICT 서비스업에 비해 기업의 생존율이 높으나 3개년도 추이를 관찰한 결과 신규 기업의 진입률은 감소하고 퇴출률은 증가하면서 생존기업의 시장점유율이 확대되는 것으로 보인다. 더 장기적인 추세를 관찰해 봐야 하겠지만, 자산으로서의 데이터에 대한 규모의 경제를 고려하고 플랫폼 경제의 네트워크효과를 감안할 때 소수기업이 시장을 장악하는 시장 집중이 발생하는지 유의할 필요가 있다. 시장 집중 현상 자체가 문제가 될 수는 없으나, 그로 인해 공정한 경쟁과 시장의 효율적 자원 배분을 저해할 우려가 있기 때문이다.

데이터 기업의 생산성은 대체로 ICT 제조업에 비해서는 낮고 ICT 서비스업에 비해서는 높은 것으로 나타났다. 데이터 산업의 대분류 단위에서 생산성을 분석한 결과 ICT 산업과는 달리 총요소생산성과 노동생산성이 높은 산업에 큰 차이를 보였다. 총요소생산성은 ‘데이터 구축 및 컨설팅 서비스업’이 가장 높았으나, 노동생산성은 ‘데이터 판매 및 제공 서비스업’이 가장 높았다. 데이터 산업의 세부 산업 중 ‘데이터 처리 및 관리 솔루션 개발 · 공급업’은 유일하게 총요소생산성과 노동생산성이 모두 하락하고 있는 모습을 보여 지속적인 모니터링이 필요한 것으로 판단된다.

데이터 산업의 연구개발비 지출은 ICT 제조업의 1/14 수준으로 매우 낮지만, ICT 서비스업에 비해서 약 1.1배 더 많은 것으로 나타났다. 데이터 기업이 표준산업분류에 따르면 대부분 ICT 서비스 기업에 해당하기 때문에 ICT 서비스 기업 중 데이터 기업의 혁신 활동은 더 활발하다고 판단할 수 있다. 데이터 기업의 지적재산권 취득 수는 대체로 증가하는 패턴을 보이나, ‘데이터 구축 및 컨설팅 서비스업’은 산업의 총량적 관점에서도 기업당 평균 관점에서도 연구개발비에 비해 지적재산권 취득 수준이 낮아 연구개발의 효율성에 대한 지속적인 모니터링이 필요하다.

한편 ‘데이터 처리 및 관리 솔루션 개발 · 공급업’의 기업당 평균 연구개발비의 하락이 지속되고 있다. 연구개발의 효과가 시차를 두고 나타나고, 지적재산권이 연구개발의 결과를 측정하는 유일한 지표가 아니라는 점에서, 혁신 활동 측면에서는 ‘데이터 구축 및 컨설팅 서비스업’보다 ‘데이터 처리 및 관리 솔루션 개발 · 공급업’에 대한 모니터링이 더 필요할 수 있다.

마지막으로, 본 장은 3개년도 데이터를 활용하여 기업 단위에서 분석한 결과를 바탕으로 작성된 것으로 더 많은 시계열 자료를 활용하여 분석하거나, 사업체 단위에서 분석한 결과와 다를 수 있다는 점을 밝힌다.

따라서 해석에 주의가 필요하다는 점을 다시 한번 강조한다. 즉, 이 자료만을 근거로 삼아 연구개발의 효율성, 시장의 역동성과 시장 집중, 생산성 둔화 등과 관련해 단정적인 가치판단을 내리는 것은 부적절하다.

제3장

해외 데이터산업 시장 현황

고태우 팀장 KDB산업은행

2021년 팬데믹, 2022년 러·우 전쟁 등 외부 요인으로 데이터는 언제나 실 새 없이 변하지만, 그럴수록 데이터 시장의 가치는 높아진다. 해를 거듭할수록 정부·기업·개인의 모든 영역에서 데이터를 생산·가공·유통·활용하여 부가가치를 창출하고 있다. 디지털 생태계와 직간접적으로 연결된 기업들로서 데이터의 수집·분석·활용이 생존을 위한 필수사항이 되었기에 새로운 흐름에 촉각을 곤두세워야 한다. 예를 들어 2023년에는 데이터를 기반으로 응용, 학습할 수 있는 대표적인 디지털 기술 중 하나인 LLM(Large Language Model)이 주목받는 시기였다는 점을 놓칠 수 없다. 발 빠른 대응을 위한 동향 파악은 늘 중요하다.

이는 해외에서도 마찬가지다. 선도국인 미국을 포함한 주요국은 국제연합(UN, United Nations), 유럽연합통계국(Eurostat), 경제협력개발기구(OECD) 등 국제기구를 중심으로 데이터의 공식통계 활용 방안과 관련한 협력 프로젝트를 진행하고 있는 한편, 빅테크 기업들은 주요 인프라 기술을 활용한 효율적인 데이터 시스템 구축을 위해 다양한 솔루션과 신기술을 도입하고 있다.

본 장에서는 데이터 산업의 해외 주요 시장을 소개한다.

1. 데이터 산업의 시장 현황

데이터 시장 규모의 차이는 조사기관별 기술적 가치 및 파급효과의 범위(기업·산업·경제) 등 산정 차이에 기인하므로, 공신력 있는 IDC & Lisbon Council이 매년 발표하는 ‘European Data Market Study’ 시장 통계를 참고했다. 또한 데이터 시장 및 데이터 경제의 성숙에 대응하여 주요 지표에 대한 정의 및 통계분석 방법이 매년 현실에 맞게 수정되곤 하는데, 이 글에서는 ‘European Market Study 2021–2023(2023년 2월)’를 기준으로 삼았다.

더 나아가 2021년은 팬데믹 영향을 고려하였으며, 2022년은 러·우 전쟁, 에너지 위기, 인플레이션 압력과 같은 거시경제의 불확실성과 함께 저조한 글로벌 경제 성장률이 주요한 외부 변수였다. 2023년 이후 전망치 또한 글로벌 공급망 단절, IT 기술력 부족, 포퓰리즘 위험, 에너지 위기 등이 몇 년간 연속적으로 영향을 주었다. 추가로 급격한 사회·경제적 변화로 표현되는 “뉴노멀(New Normal)”도 고려해야 했다.

주요 통계조사 방법론에는 크게 △‘정량 지표 측정 방법’과 △‘필드 리서치’가 있다. 또한 데이터 시장 가치 산정을 위한 하위 지표로는 △데이터 수익화(Monetization) 지표의 국가별·산업별·기업별 접근, △데이터 경제 지표의 데이터를 공유하고 재활용하는 것에 대한 간접적 효용 반영, △‘데이터 커뮤니티와의 상호작용 및 스토리를 통한’ 다른 측면의 데이터 가치 조사 등이 있다.

이 글에서는 'European Market Study(IDC, 2023년 2월)' 내용 중 유럽연합¹⁾에 집중된 통계 데이터의 특징을 정리하고, 영국 및 비유럽권의 주요 국가인 미국, 중국²⁾, 브라질, 일본과의 각 시장 규모를 비교·분석하였다.

[표 3-3-1] European Data Market Study 통계조사 방법론

항목 구분	변경 후
조사 기간	(현황) 2020~2022 (전망) 2025, 2030
추정 방법	시나리오법 - 고성장, 기본, 보수적 접근 (기준 연도) 2025 (추정 연도) 2030
주요 방법론	
정량 지표 측정 방법	데이터 가치, 데이터 공급기업과 이용 기업 수, 비즈니스 영향, 데이터 전문가 수
필드 리서치	- 일회성 서베이*

〈데이터 및 모델링 QC 관점의 방법론 개선 도식화〉



* 일회성 서베이(Ad-Hoc Survey): 데이터 공급자와 이용자 대상. 빅데이터 분석과 시 활용에 따른 효율성 증진 관련 조사(시간·비용 절감에 따른 비즈니스 영향도 조사 / 조직의 성과를 근간으로 한 비즈니스 영향도 조사)

* 출처: IDC, European Market Study, (2023년 2월)

1) 본 장에서 유럽연합(EU)은 27개국(오스트리아, 벨기에, 불가리아, 크로아티아, 사이프러스, 체코, 덴마크, 에스토니아, 핀란드, 프랑스, 독일, 그리스, 헝가리, 아일랜드, 이탈리아, 라트비아, 리투아니아, 룩셈부르크, 몰타, 네덜란드, 폴란드, 포르투갈, 루마니아, 슬로바키아, 슬로베니아, 스페인, 스웨덴)을 의미한다. 본 보고서에서는 스위스와 유럽경제지역 3개국(아이슬란드, 리히텐슈타인, 노르웨이)이 조사 대상국에 포함되었으나 본 장에서는 제외했다.

2) 중국의 사회경제적, 기술적 영향 증가로 국제 지표의 범위가 미국, 브라질, 일본에서 중국으로 확대되었다.

가. 데이터 시장

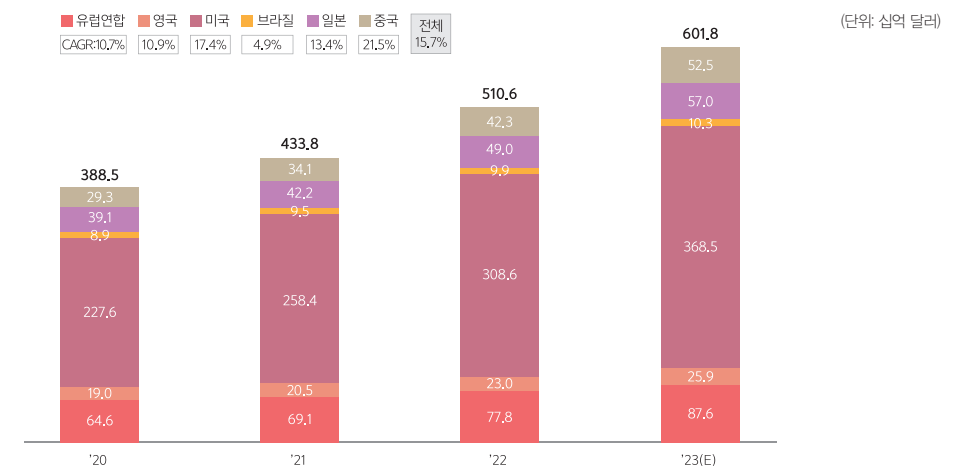
데이터 시장이란 데이터를 가공함으로써 제품 및 서비스로 재생산되는 '디지털 데이터 시장'으로 정의했다. 여기에는 이미지 데이터처럼 디지털 기술을 통해 수집·저장·가공·전송되는 멀티미디어 데이터 부문도 포함된다.

디지털 데이터 시장 규모는 유럽연합(EU), 영국, 미국, 중국, 일본, 브라질의 데이터 기반 기업의 매출액을 기반으로 산출되었다. 데이터 기업의 매출은 데이터 관련 제품의 총가치에 해당하고, 수출을 포함하여 해당 국가에 기반을 둔 기업이 생산한 서비스의 매출을 기반으로 한다.

미국의 디지털 데이터 시장은 규모 면에서 세계에서 가장 크다. 시장의 성장세 또한 높아서, 2020~2023년(E)까지 연평균 성장률 17.4%를 보였다. 미국은 시장 규모가 2020년 2,276억 달러였는데 2023년에는 3,685억 달러의 시장 형성이 예상될 만큼 지속적으로 두 자릿수 성장세를 보이고 있다.

유럽연합(EU)의 데이터 시장은 연평균 성장률 10.7% 수준으로 상대적으로 성장세가 약하다. 2020년 646억 달러에서 2023년 876억 달러에 이를 것으로 예상된다. 유럽연합 시장 규모 내에서는 2022년 기준 '독일, 프랑스, 이탈리아, 네덜란드, 스페인'이 약 2/3의 비중을 차지한다. 해당 국가들은 다른 국가보다 상대적으로 성장률이 높다. 국가별 성장률 차이는 ICT 기술에 대한 지출액 및 경제력과 상관관계가 높다. 또한 대다수 국가에서 우크라이나 전쟁으로 발생한 부정적 효과가 있긴 했지만, 그보다 코로나에 따른 경제적 충격에서 회복한 효과가 컸기 때문에 2022년엔 전년 대비 높은 성장률을 기록했다.

[그림 3-3-1] 글로벌 데이터 시장 규모(2020~2023년(E))³⁾



* 2020~2022: IDC, European Market Study, (2023년 2월).
2023(E): 위 보고서를 바탕으로 한 저자 추정치

3) 본 보고서에서는 스위스와 유럽경제지역(EEA) 3개국(아이슬란드, 리히텐슈타인, 노르웨이)이 조사 대상국에 포함되었으나 본 장에서는 제외했다. EU27+UK+스위스+EEA = 899억 달러(2020), 964억 달러(2021), 1,083억 달러(2022)
출처: IDC, European Market Study, (2023년 2월).

영국의 데이터 시장은 연평균 성장률 10.9%이며 2020년 190억 달러에서 2022년 259억 달러로 성장할 것으로 보인다. 일본의 데이터 시장 규모는 2020년 391억 달러에서 2023년 570억 달러로 연평균 성장률 13.4% 수준으로 추정된다.

중국의 데이터 시장은 2020년 293억 달러에서 2023년 525억 달러로 연평균 성장률 21.5% 수준의 가장 빠른 성장을 기록하였으며, 브라질의 데이터 시장 규모는 2020년 89억 달러에서 4.9% 성장해 2023년 103억 달러를 기록할 것으로 예상된다.

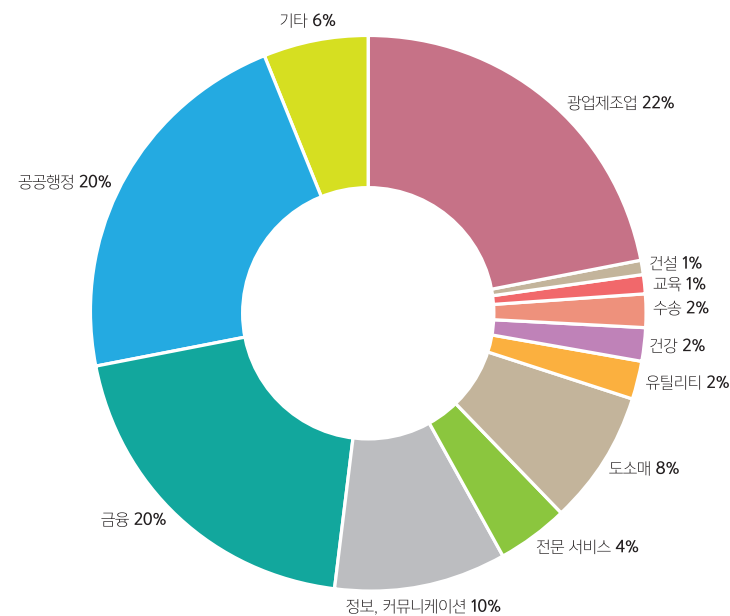
1) 산업별 기여도(유럽연합 기준)

데이터 기술의 산업별 적용 관점에서 기업 수로 측정되는 데이터 시장의 점유율에 대한 기여도를 산출하였다. 큰 비중을 차지하는 분야로는 광업 · 제조, 금융, 공공행정 산업이 각각 20% 이상으로 조사되었다.

해당 조사에서 산업에 적용되는 데이터 기술의 비중이 높을수록 데이터 기술의 산업별 적용이 잘되어 해당 산업이 성장하는 특징을 파악할 수 있었다. 시장이 성장하자 많은 기업이 데이터 툴과 서비스에 대한 선제적 투자를 하였고, 해당 시장 내 기관 혹은 기업이 데이터 기술의 가장 큰 소비자가 된 사실을 확인하였다.

해당 조사는 유럽연합(EU) 기준의 2020~2030년 추정치에 국한되어 있으나, 데이터 기술이 가지는 산업별 부가가치를 고려할 때 '디지털화가 진행되는 다른 주요 국가의 산업별 데이터 시장 기여도' 역시 비슷한 수준일 것으로 추정된다.

[그림 3-3-2] 유럽연합 내 산업별 데이터 시장 기여도(2020~2030년): 기본성장 시나리오



* 출처: IDC, European Market Study, (2023년 2월), Figure 48.

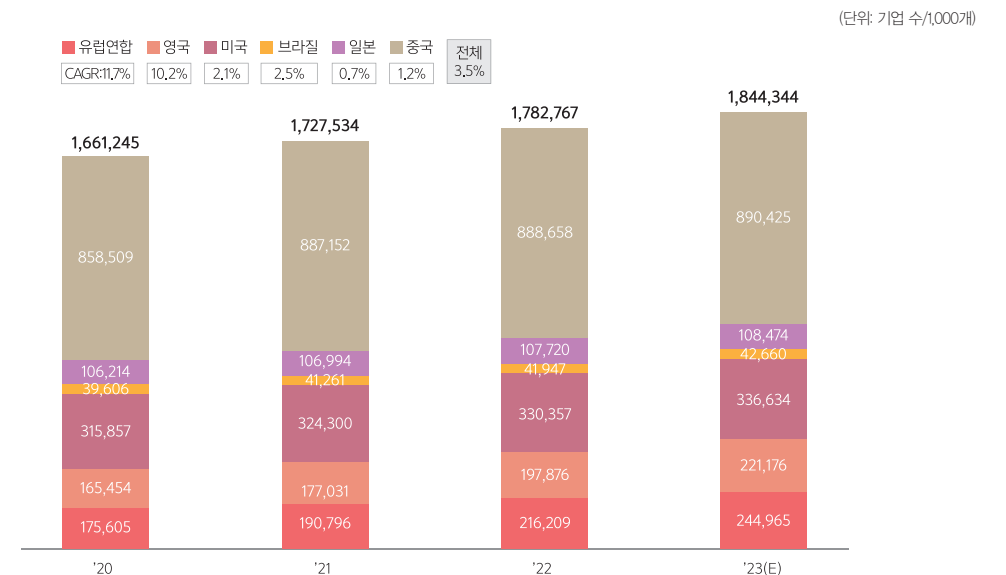
나. 데이터 공급기업 수

데이터 공급기업이란 시장에 데이터를 공급하는 기업으로서 데이터의 생산 · 전달 · 이용하는 기업으로 정의하며, 데이터 관련 제품, 서비스 및 기술에 대한 데이터의 공급자와 수요자로 구분할 수 있다.

데이터 시장 규모가 성장하면서 데이터 공급기업 수도 지속적으로 증가하고 있다. 특히 유럽연합의 데이터 공급기업 수는 2020년 1억 7,561만 개에서 2023년 2억 4,497만 개로 늘어 연평균 성장률 11.7%로 가장 높은 수준을 보였으며, 기업 수 측면에서는 중국이 가장 많았다. 상대적으로 미국은 앞서 언급한 시장 규모 측면에서는 가장 컸으나, 데이터 공급기업 수는 2020년 3억 1,586만 개에서 2023년 3억 3,366만 개로 늘어 연평균 성장률은 2.1% 수준이다.

시장 규모가 매출액을 기반으로 산출된 점을 고려한다면 미국의 주요 빅테크 기업인 구글, 애플, 메타(舊 페이스북), 아마존, 마이크로소프트 등의 기여도가 높고, 중국의 공급기업 중 소규모 기업 비중이 큰 것이 주요 요인으로 추정된다.

[그림 3-3-3] 글로벌 데이터 공급기업 수(2020~2023년(E))⁴⁾



* 2020~2022: IDC, European Market Study, (2023년 2월).
2023(E): 위 보고서를 바탕으로 한 저자 추정치

4) 본 보고서에서는 스위스와 유럽경제지역(EEA) 3개국(아이슬란드, 리히텐슈타인, 노르웨이)이 조사 대상국에 포함되었으나 본 장에서는 제외했다.

*EU27+UK+스위스+EEA = 3억 5,157만 개(2020), 3억 7,924만 개(2021), 4억 7,036만 개(2022)

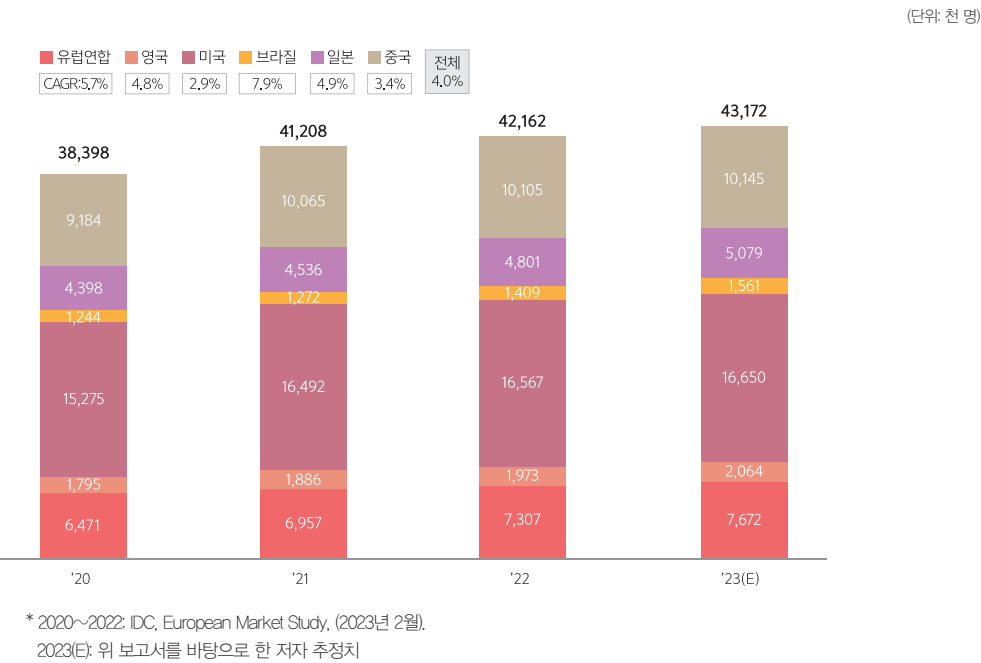
출처: 위의 보고서(IDC, 2023년 2월)

다. 데이터 전문인력

데이터 전문인력이란 새로운 데이터 분야 기술에 익숙한 능동자로 정의한다. 즉 대용량 데이터를 수집 · 저장 · 관리 및 분석 · 해석, 시각화할 수 있는 인력으로 정형 및 비정형 데이터의 사용에 능숙해야 하고, 막대한 양의 데이터를 원활히 활용할 수 있어야 한다.

미국은 2020년 1,527만 명에서 2023년 1,665만 명으로 예상되며, 연평균 성장률 2.9% 수준으로 꾸준히 증가하고 있다. 중국은 2020년 918만 명에서 2023년 1,015만 명으로 연평균 성장률 3.4% 수준의 성장률을 기록하고 있으며, 유럽연합(EU)의 경우 2020년에 650만 명에서 2023년 767만 명 수준으로 연평균 성장률 5.8%를 보이고 있다. 데이터 전문인력은 전 세계적으로 부족한 상황으로 각국은 데이터 전문인력을 양성하고 확보하는 데 높은 관심을 기울이고 있다.

[그림 3-3-4] 글로벌 데이터 전문인력 수(2020~2023년(E))⁵⁾



유럽연합(EU) 기준 분야별 데이터 전문인력은 전문 서비스, 도소매 영역에서 비중이 높으며, 뒤이어 정보 및 커뮤니케이션, 광업 · 제조 순이다.

⁵⁾ 본 보고서에서는 스위스와 유럽경제지역(EEA) 3개국(아이슬란드, 리히텐슈타인, 노르웨이)이 조사 대상국에 포함되었으나 본 장에서는 제외했다.
*EU27+UK+스위스+EEA = 862만 명(2020), 920만 명(2021), 966만 명(2022)
출처: 위의 보고서(IDC, 2023년 2월)

[표 3-3-2] 유럽연합의 데이터 전문인력 수(2020~2023년(E))

(단위: 천 명)

구 분	2020년	2021년	2022년	2023년(E)	CAGR('21~'22)
농업	35	37	39	41	4.8%
건설	131	142	148	153	3.7%
교육	480	509	538	568	5.6%
금융	609	652	680	710	4.4%
건강	518	555	586	619	5.6%
정보 및 커뮤니케이션	753	824	884	948	7.3%
광업 · 제조	745	808	844	881	5.0%
전문 서비스	1,385	1,497	1,583	1,673	5.7%
공공 행정	395	419	438	458	4.6%
도소매	1,131	1,203	1,246	1,290	3.5%
운송	197	211	219	228	3.9%
전력 · 가스 · 수도	92	99	102	105	2.7%
총 계	6,471	6,957	7,307	7,672	5.8%

라. 데이터의 경제 가치

데이터의 경제 가치는 데이터 시장이 경제 생태계 전체에 미치는 전반적인 영향이 어느 정도인지 평가하는 분석 방법론을 의미한다. 디지털 기술이 발달하면서 데이터의 수집 · 저장 · 처리 · 배포 · 분석 · 정교화 · 전달 및 활용이 쉽게 되었다. 이렇게 데이터의 활용이 쉬워지면서 데이터 산업과 시장이 형성되었고, 직접 또는 간접적으로 경제 전체에 영향을 미치고 있다.

이러한 데이터의 경제 가치는 직접 효과와 간접 효과로 구분된다.

1) 직접 효과

직접 효과란 데이터 산업 그 자체에서 형성된 효과를 의미한다. 즉 데이터 생산과 관련한 모든 비즈니스 활동으로 형성된 효과다. 정량적인 직접 효과는 판매된 데이터 제품 및 서비스의 수익을 기준으로 측정된다.

미국의 데이터 경제 가치 중 직접 효과는 2020년 2,276억 달러를 기록한 이후 점차 증가하여 2023년에는 3,685억 달러를 넘어설 것으로 보인다. 데이터에 기반을 둔 제품 및 서비스의 생산에서 파생된 직접적 영향은 중국, 유럽연합(EU), 영국, 일본, 브라질보다 높게 드러났다. 이는 다른 국가에 비해 데이터 시장의 규모가 크고, GDP에서 차지하는 비중도 크기 때문이다.

미국의 데이터 직접 효과는 데이터 시장 규모에서도 알 수 있듯이 유럽연합(EU)의 직접 효과와 비교할 때 거의 4배에 가까운 수치인데, 이는 미국의 데이터 산업이 상대적으로 발전되어 있고, 경제적 효과의 확산 속도가 더

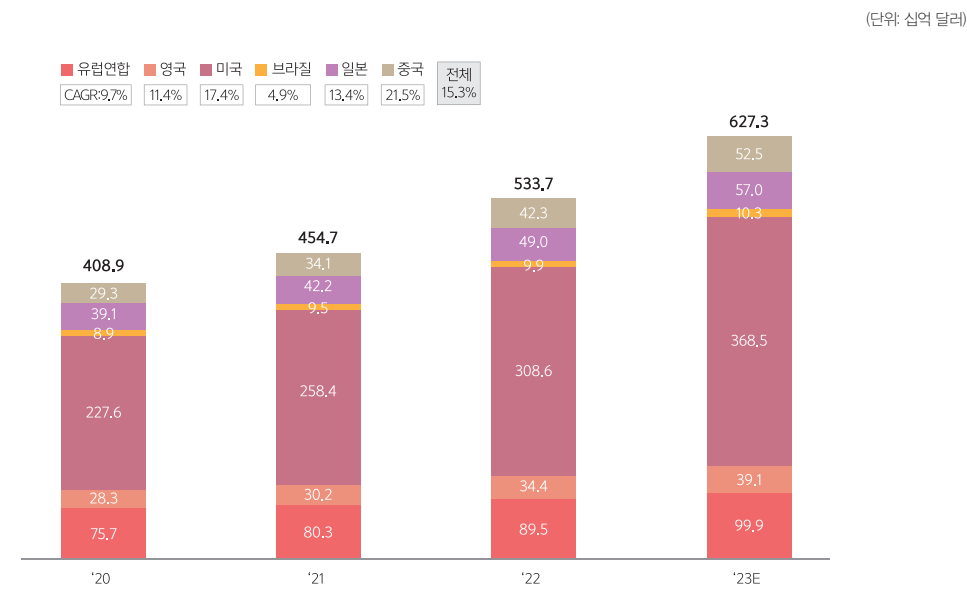
빠르다는 것을 시사한다.

중국의 데이터 경제 가치 중 직접 효과는 연평균 성장률 21.5% 수준으로 비교 대상국 중 가장 높으며, 2020년 293억 달러에서 2023년 525억 달러로 성장할 것으로 전망된다.

브라질의 데이터 경제 가치 중 직접 효과는 2020년에 89억 달러를 기록한 이후 점차 증가하여 2023년에는 103억 달러로 집계되었는데 최근 몇 년간의 성장세는 비교 대상국 대비 가장 낮은 것으로 평가된다.

영국의 데이터 경제 가치 중 직접 효과는 2020년 283억 달러에서 2023년 391억 달러로 성장할 것으로 예상되며 연평균 성장률은 11.4% 수준이다. 일본의 데이터 경제 가치 중 직접 효과는 2020년 391억 달러에서 2023년 570억 달러로 증가하여, 연평균 성장률 13.4% 수준으로 추정된다. 유럽연합(EU)의 데이터 경제적 효과 중 직접 효과는 2020년에 757억 달러를 기록한 이후 꾸준히 증가하여 2023년 999억 달러로 예상된다. 유럽연합(EU)은 미국 다음으로 데이터 시장 규모가 크긴 하나, 디지털화를 위한 투자의 구조적 한계, 파편화된 정책, 디지털 성장 잠재력에 대한 인지 부족, 디지털 전문 인력 부족을 이유로 미국 대비 1/4 수준에 머물러 있다.

[그림 3-3-5] 데이터 경제 가치: 직접 효과(2020~2023년(E))⁶⁾



* 2020~2022: IDC, European Market Study, (2023년 2월).
2023(E): 위 보고서를 바탕으로 한 저자 추정치

6) 본 보고서에서는 스위스와 유럽경제지역(EEA) 3개국(아이슬란드, 리히텐슈타인, 노르웨이)이 조사 대상국에 포함되었으나 본 장에서는 제외했다.

*EU27+UK+스위스+EEA = 1,273억 달러(2020), 1,384억 달러(2021), 1,571억 달러(2022)

출처: 위의 보고서(IDC, 2023년 2월)

2) 간접 효과: 후방 효과

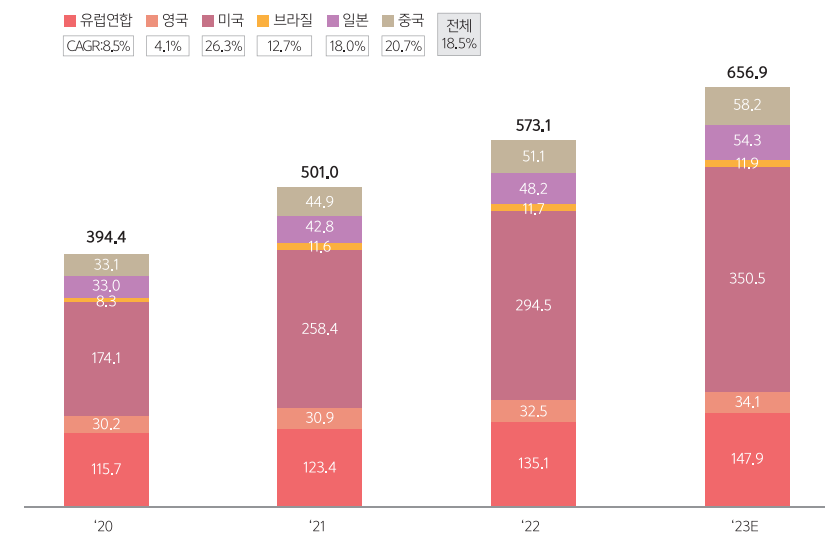
데이터 경제 가치 중 간접 효과는 데이터 산업이 다른 산업에 미치는 영향을 나타내는 요소이다. 직접 효과가 데이터 생산 활동이 데이터 산업 내에 미치는 영향을 설명하는 것이라면, 간접 효과는 타 산업에 미치는 영향을 설명하고 있다.

간접 효과는 전방 효과와 후방 효과로 구분된다. 전방 효과는 데이터 산업에서 생산된 데이터 기반의 제품 및 서비스를 이용하여 다른 산업에서 다른 형태의 제품 및 서비스로 가공되는 효과를 의미한다. 후방 효과는 데이터 산업에 속한 기업이 데이터 기반의 제품 및 서비스를 생산하는 활동을 하는 과정에서 다른 산업으로부터 자원을 받을 때 자원을 제공한 산업에서 발생하는 효과를 말한다.

데이터 산업으로 발생하는 후방 효과는 매우 빠르게 증가하고 있다. 특히 미국의 경우 2020년 1,741억 달러였는데, 26.3% 성장률로 2023년에는 3,505억 달러에 이를 것으로 전망된다.

유럽연합(EU)의 후방 효과는 2020년 1,157억 달러 규모였다가, 8.5%의 성장률로 증가하여 2023년 1,479억 달러 규모를 형성할 것으로 전망된다. 유럽연합의 경우 공공데이터 개방 정책을 펼치다 보니 민간 데이터보다 공공데이터 중심의 데이터 산업이 상당한 비중을 차지한다. 물론 민간기업의 데이터 또한 활용 폭을 넓히고자 다양한 정책을 마련하고 있다.

[그림 3-3-6] 데이터 경제 가치: 간접 효과 중 후방 효과(2020~2023년(E))⁷⁾



* 2020~2022: IDC, European Market Study, (2023년 2월).
2023(E): 위 보고서를 바탕으로 한 저자 추정치

7) 본 보고서에서는 스위스와 유럽경제지역(EEA) 3개국(아이슬란드, 리히텐슈타인, 노르웨이)이 조사 대상국에 포함되었으나 본 장에서는 제외했다.

*EU27+UK+스위스+EEA = 1,580억 달러(2020), 1,624억 달러(2021), 1,716억 달러(2022)

출처: 위의 보고서(IDC, 2023년 2월)

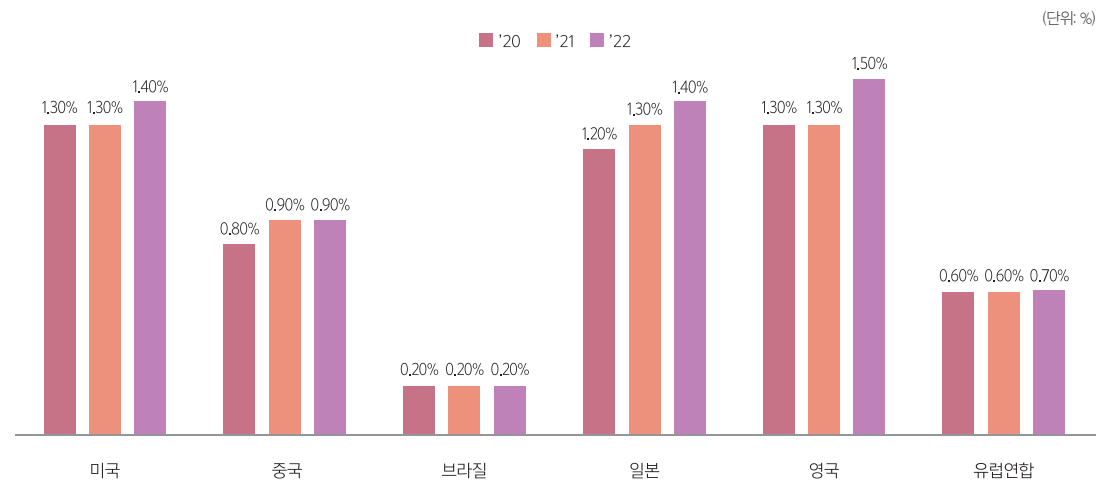
영국의 후방 효과는 비교 대상 국가 중 가장 낮은 연평균 성장률 4.1% 수준으로, 2020년 302억 달러에서 2023년 341억 달러로 성장할 것으로 보인다. 중국의 후방 효과는 연평균 성장률 20.7% 수준이며, 2020년 331억 달러에서 2023년 582억 달러로 성장할 것으로 전망된다. 일본의 후방 효과는 연평균 성장률 18.0% 수준이며, 2020년 330억 달러에서 2023년 543억 달러로 성장할 것으로 보인다. 브라질의 후방 효과는 연평균 성장률 12.7% 수준이며, 2020년에 83억 달러 규모였던 후방 효과는 2023년에 119억 달러 수준에 도달할 것으로 전망된다.

3) GDP 대비 데이터 경제 가치의 비율

앞서 데이터 산업의 경제적 효과를 직접 효과와 간접 후방 효과를 중심으로 살펴보았는데, 직접 효과와 간접 후방 효과 모두 규모 측면에서 미국이 가장 앞선 것으로 드러났으며, 두 효과를 합산할 때 2022년 기준 6,031억 달러 수준이었다.

GDP에서 경제적 가치(직접 효과)가 차지하는 비율을 살펴보면 미국은 2020년 1.30%에서 2022년 1.40%로 증가하여 0.10%p 증가하였다. 영국은 2020년 1.30%에서 2022년 1.50%로 0.20%p 상승하였으며, 일본의 경우도 2020년 기준 1.20%에서 2022년 1.40%로 0.20%p 증가폭을 보였다. 중국은 GDP 대비 경제적 가치(직접 효과) 비중이 2020년 0.80%에서 2022년 0.90%로 0.10%p 상승했다. 유럽연합(EU)은 3년 연속 0.60% 수준을 유지하였으며, 브라질의 경우 3년 연속 0.20% 수준에 그쳐 GDP에서 차지하는 데이터 경제 가치는 다소 뒤처지는 것으로 보인다.

[그림 3-3-7] GDP 대비 데이터 산업의 경제적 가치(직접 효과)의 비율(2020~2022년)



* 2020~2022: IDC, European Market Study, (2023년 2월).

마. 중장기 데이터 시장 전망(유럽연합)

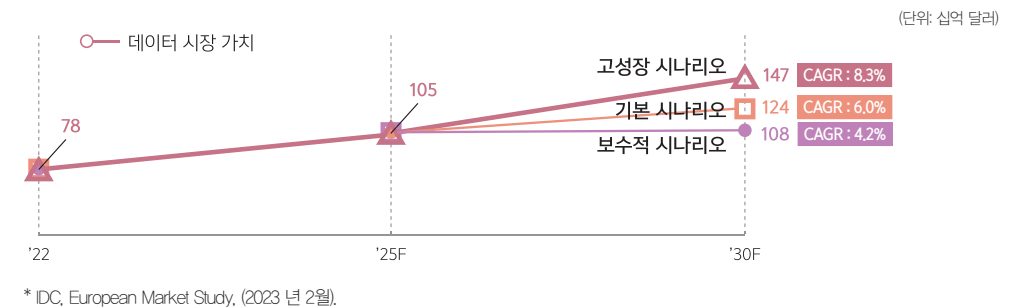
향후 데이터 시장은 지속적인 성장세를 기록할 것으로 전망된다. 'European Market Study(IDC, 2023년 2월)'에 따르면 기본 시나리오 기준으로 2022년에서 2030년까지 연평균 성장률은 6.0%로 예측되어 2030년 유럽연합(EU)의 데이터 시장 규모는 1,240억 달러까지 성장할 것이다. 만약 고성장 시나리오가 전개된다면 1,470억 달러까지 가

능할 것으로 전망된다.

데이터 기업의 매출 성장률은 같은 기간 전체 ICT 시장의 성장률을 훌쩍 뛰어넘을 것이라는 전망이 지배적이며, 경제 규모가 큰 국가일수록 데이터 기반 시장이 데이터 경제 확대에 크게 기여할 것으로 예측된다.

데이터 관련 인력 또한 데이터 관련 기업의 향후 성장세를 견인하는 중요한 자원으로 평가받고 있어 지속적인 증가세를 보일 것으로 보인다. 높은 고용률에 따라 고액 연봉 체제도 구축될 것으로 전망된다.

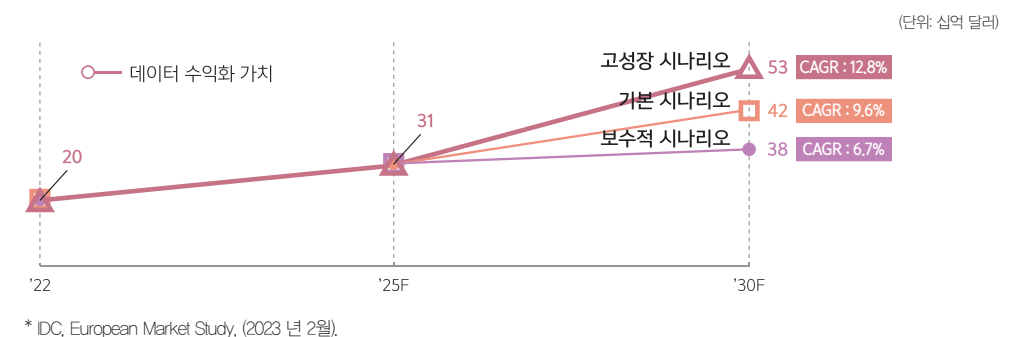
[그림 3-3-8] 유럽연합의 중장기 데이터 시장 전망(2022~2030년(E))



2022년 IDC는 데이터 시장 규모 산정의 하부 항목으로 '데이터 수익화 가치'를 산정하였고, 데이터의 유통 과정에서 창출되는 부가가치를 중요한 지표로 판단하고 있다.

이러한 데이터 수익화 시장의 성장을 이끌려면 몇 가지 주요 요인이 있는데, 우선 데이터 생성량을 늘려야 하고, 둘째로 데이터 저장 비용을 절감해야 한다. 셋째로 데이터를 고급 분석해야 하며, 마지막으로 데이터를 빠르게 시각화할 수 있어야 한다. 데이터 시장 성장 기여도 측면에서 '데이터 수익화 가치'의 기여도는 30% 수준인데, 데이터의 수익화 및 자본화 방법을 이해하고 사업화하는 과정에서 많은 비즈니스 모델이 생기고 있다. 향후 데이터 수익화의 중추적인 역할이 기대된다.

[그림 3-3-9] 유럽연합의 중장기 데이터 수익화 가치 전망(2022~2030년(E))



4 PART

산업별 데이터 활용 현황

제1장 • 금융분야 데이터 활용 현황

제2장 • 헬스케어분야 데이터 활용 현황

제3장 • 모빌리티분야 데이터 활용 현황

제4장 • 제조분야 데이터 활용 현황

제5장 • 농업분야 데이터 활용 현황

제6장 • 에듀테크분야 데이터 활용 현황

제7장 • 新 데이터 비즈니스

제 1 장

금융분야
데이터 활용 현황

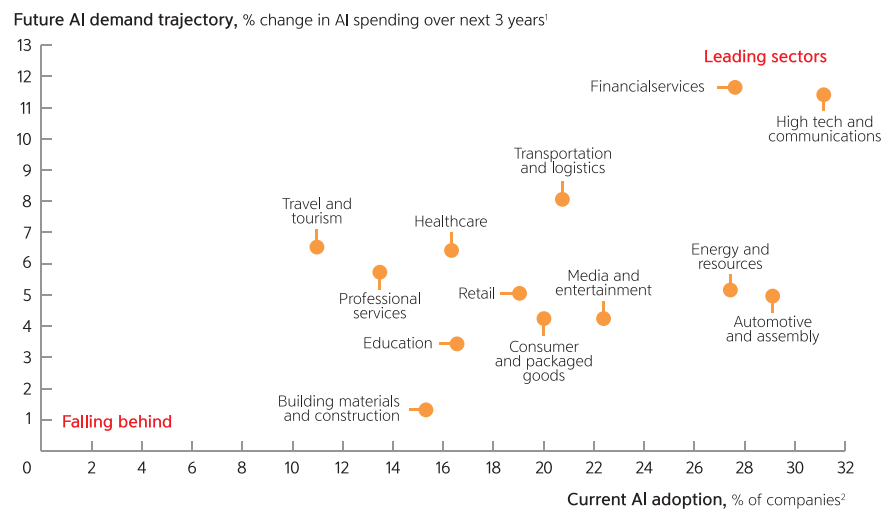
우지환 교수 고려대학교 기술경영전문대학원

2016년 이세돌 9단과 인공지능(AI) 알파고의 바둑 대결로 AI는 일반 대중에게 널리 알려졌다. 이는 ‘인공지능(AI)’ 역사에서 큰 사건이었다. ‘인공지능(AI)’이라는 단어가 연구실에서 과학자들만 사용하던 용어에서 벗어난 순간이었다. 대결이 벌어지기 전에는 사람의 낙승을 예상했다. 너무 많은 경우의 수를 고려해야 하기에 컴퓨터가 사람을 절대로 이기지 못할 것으로 보였다. 그런데 결과는 정반대로, 인공지능이 세계 최고의 바둑기사인 이세돌 9단을 너무나 쉽게 이겼다. 이세돌 9단과 알파고의 대국으로 발생한 파급효과는 단순히 바둑 게임으로 그치지 않았다. 대중들에게 인공지능이라는 단어가 널리 주목받은 것이다.

사실 인공지능 기술은 2016년도에 새롭게 등장한 기술이 아니다. 1950년대부터 꾸준히 연구되다가 2번의 겨울(연구에 대한 대중의 관심이 줄어서 연구의 성과가 미비하던 시기)을 겪고, 다시 등장했기 때문이다. 새로운 봄을 맞이한 인공지능 기술은 현재 제조, 의료, 법률, 금융, 서비스 등 다양한 산업에서 활용된다. 세계적인 경영 컨설팅 회사인 맥킨지(McKinsey)의 2018년 조사에 따르면 금융 산업은 인공지능 서비스에 대한 수요 및 활용 정도가 가장 높은 분야다.

이 글에서는 다양한 산업 분야 중에서 현재 금융 부문에서 인공지능 기술과 빅데이터가 어떻게 활용되는지 살펴보고, 이와 관련된 기술과 향후 발전 방향 등에 대해 살펴보고자 한다.

[그림 4-1-1] AI 활용 리더의 적극적인 미래 투자 계획(후발 기업과 비교)



¹Estimated average, weighted by company size; demand trajectory based on midpoint of range selected by survey respondent.

²Adopting 1 or more AI technologies at scale or in business core; weighted by company size.

Source: McKinsey Global Institute AI adoption and use survey; McKinsey Global Institute analysis

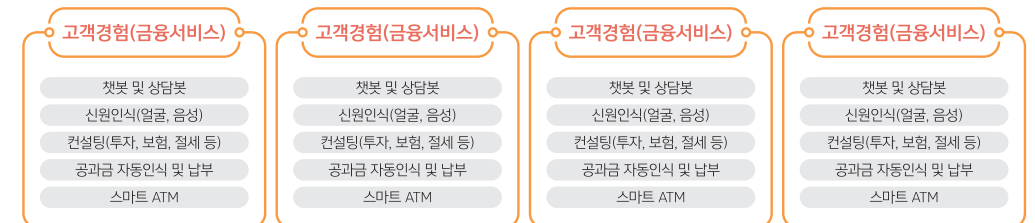
* 출처: Michael Chui, James Manyika, and Mehdi Miremadi, "What AI can and can't do (yet) for your business", 2018, 1. 11., 2023년 8월 25일 접속, <https://www.mckinsey.com/capabilities/quantumblack/our-insights/what-ai-can-and-cant-do-yet-for-your-business>

1. 금융 산업에 활용되는 인공지능 기술

가. 금융 회사 내부 업무에 활용되는 인공지능 기술

금융 산업의 핵심은 시시각각 발생하는 대용량의 금융 데이터를 잘 처리해서 고객들에게 가치 있는 서비스를 제공하는 것이다. 회사 내부적으로는 대용량의 금융 데이터를 수집하고 분석하는 업무가 있고, 외부적으로는 고객들에게 가치 있는 서비스를 제공하는 업무가 있다. 금융 산업에서 인공지능 기술은 두 분야 모두에 활용될 수 있다.

[그림 4-1-2] 금융권의 AI 적용 분야 현황

* 출처: 송민택, "생성형 AI 시대와 금융권의 AI 동향", 코스콤 뉴스룸, 2023. 3. 8., 2023년 8월 25일 접속, <https://newsroom.koscom.co.kr/33754>

주가나 환율을 보면 알겠지만, 금융 데이터는 시시각각 빠르게 변화하기 때문에 그 양이 방대하다. 지금 순간에도 세계 각국의 주식 시장에서 대용량의 데이터가 생산되고 있다. 이러한 데이터를 분석하는 업무는 그동안 금융 산업에 오랜 기간 종사했던 베테랑 직원들이 맡았지만, 이 업무를 인공지능이 맡으면서 더욱 빠르고 정확하게 데이터를 분석할 수 있었다. 금융 데이터 분석을 위해 활용되는 분야는 다음과 같다.

1) 주가/환율 데이터 예측

주가와 환율을 예측하는 일은 기존 데이터를 바탕으로 미래의 데이터를 예측하는 작업이다. 인공지능을 활용하기 전까지 미래 데이터를 예측하는 일은 전문가들의 업무였다. 전문가들이 한 분야에서 오랫동안 활동하면서 쌓은 능력을 바탕으로 전문가만의 고유 법칙을 만들고, 그 법칙을 바탕으로 다음 데이터를 예측하는 것이다. 즉 주가 데이터의 차트를 보고, 차트의 모양에 따라 주가를 예측하는 일을 떠올려볼 수 있다.

다만 사람이 데이터에서 법칙을 발견하는 업무는 분명히 한계가 있다. 우선 사람은 3차원까지만 데이터를 시각화할 수 있기 때문에 4개 이상의 특성을 가진 데이터는 시각화하기 어렵다. 당연히 이에 따른 데이터의 법칙을 찾기가 어렵다. 또한, 관찰해야 하는 데이터의 양이 많아질수록 사람들은 데이터 분석에 어려움을 겪는다.

그런데 AI 기술을 활용하여 데이터를 예측하면, 컴퓨터가 스스로 데이터를 분석해서 규칙을 찾는다는 점에서 장점이 있다. 인공지능은 4차원 이상도 스스로 시각화할 수 있고, 기억할 수 있는 데이터의 용량도 사람과 비교해서 크다.

이런 데이터 예측 기술은 로보어드바이저와 같은 애플리케이션에 적용할 수 있다. 그러면 인공지능을 통해 주

가 또는 환율을 분석해 수치가 상승할 수 있는 종목, 하락할 위험 등을 예측할 뿐 아니라, 주식들 사이의 상관관계 등을 자동으로 분석해 이것을 토대로 포트폴리오를 구성하거나 금융상품을 만들 수 있다. 즉 금융 기관들은 인공지능을 통해 데이터를 분석하고, 머신러닝을 활용하여 투자 포트폴리오를 최적화하고 고객의 투자 성향에 따라 맞춤형 자산 관리 서비스를 제공한다. 과거의 금융 데이터와 시장 동향을 분석하여 미래의 수익성을 예측하고 리스크를 최소화하는 포트폴리오를 구성할 뿐만 아니라, 개인 고객의 투자 성향과 목표에 맞춰 맞춤형 자산 관리 솔루션을 제공하여 고객의 투자 성공을 지원할 수 있다.

2) AI 기반 신용평가

금융에서 가장 중요한 일 중 하나는 신용을 평가하는 것이다. 돈을 빌려주거나 투자할 때, 신용을 정확하게 평가해야 채불 불능의 위험에서 벗어날 수 있다. 기존의 신용 평가 또한 금융 전문가들의 업무 영역이었다. 이처럼 신용을 평가하는 법칙을 사람이 만든다 보니, 개개인이 가진 다양한 특성을 고려한 신용 평가 법칙을 만들 수 없었다. 이에 필연적으로 신용 평가에서 소외받는 계층이 발생했다.

그러나 AI 기술을 활용하면서, 다양한 특성을 고려한 신용평가 모델을 만들게 된다. 우선 신용 평가에 빅데이터 분석 기술과 AI를 도입하면, AI가 대량의 데이터를 분석하고, 패턴 및 트렌드를 식별하여 개인 또는 기업의 신용 위험을 빠르고 정확하게 예측할 수 있다.

기존의 신용평가 모델은 주로 신용 점수와 같은 정형화된 데이터를 사용한다. 빅데이터와 AI를 활용한 신용평가 모델에서는 금융 정보와 같은 정형 데이터 이외에 SNS 데이터, 거래 이력, 상품 구매 기록 등 개인화 데이터와 같은 비정형, 비금융 데이터까지 활용할 수 있다. 따라서 고객별로 차별화된 신용 평가 모델을 더 정확하게 만들게 된다. 예를 들어 갓 대학을 졸업해서 사회에 첫발을 내딛는 청년이 높은 신용 점수를 받기는 어렵다. 그런데 그 청년이 휴대전화 요금을 미납 없이 꾸준히 지불했다는 정보, 또는 SNS상에서 신용이 높은 지인들과 관계를 맺고 있다는 정보, 또는 SNS상에서 과도한 소비 패턴을 보여주는 활동을 하지 않는다는 정보 등을 활용하여 AI가 정밀하게 분석한다면 신용 평가에 긍정적인 점수를 줄 수도 있다. AI를 활용하면 기존에 신용 점수가 낮은 개인 및 기업에 대해서도 다양한 데이터를 활용하여 다각도로 신용평가를 수행할 수 있다.

AI 기반의 신용평가를 하면 실시간 데이터 모니터링을 통해 신용 위험의 변화를 지속적으로 감지할 수 있고, 평가 과정이 자동화되어 결정을 신속하게 내릴 수 있다. 이런 이유로 AI를 활용한 데이터 분석 작업은 대출 승인과 대출 이자율 결정에 활용된다. 결과적으로 더욱 정확하고 신속하게 리스크 관리를 할 수 있고, 효율적인 대응이 가능해진다.

3) 이상데이터 감지

매초 빈번하게 발생하는 금융 데이터 중에는 보이스 피싱, 대포 통장 등 사기 거래에 이용되는 금융 데이터도 존재한다. 수백만 건 중 한 건 발생하는 사기 거래 데이터를 감지하기 위해서는 전문가들이 꼼꼼히 모니터들을 살펴

보면서 이상 여부를 판단해야 했다. 그런데 앞서 소개한 인공지능(AI) 기술을 활용하면 수백만 건의 거래 데이터 중에 섞여 있을 한두 건의 이상 데이터가 보이는 패턴을 학습하여 사기 거래를 탐지할 수 있다.

이처럼 금융 기관들은 데이터 분석과 머신러닝을 활용하여 금융 거래의 이상 행동을 감지하는 사기 탐지 시스템을 구축하고 있다. AI를 통해 이상 행동 패턴과 특이점을 식별하여 사기 거래를 예방하고, 금융 거래의 보안성을 강화할 수 있다. 또한, 데이터를 실시간으로 모니터링하여 금융 사기의 조기 탐지와 대응이 가능해진다.

이러한 업무 수행이 가능하기 위해서는 대표적으로 인공지능의 두 가지 기술이 필요했다. 우선 대용량 데이터 저장 기술이 그것이고, 그다음으로 빠른 연산을 통해 대용량의 금융 데이터에 숨겨진 패턴을 찾는 기술이 중요했다. 이러한 기술 덕분에 AI는 기존 소수의 금융 전문가에게 의존했던 업무를 맡아 더욱 빠르고 정확하게 수행할 뿐 아니라, 더 많은 영역에서 활용되는 추세를 보인다. 결과적으로 금융권에서 일하는 전문가들이 더욱 창의적인 일에 몰두할 수 있게 해줌으로써 업무 효율화에도 기여한다.

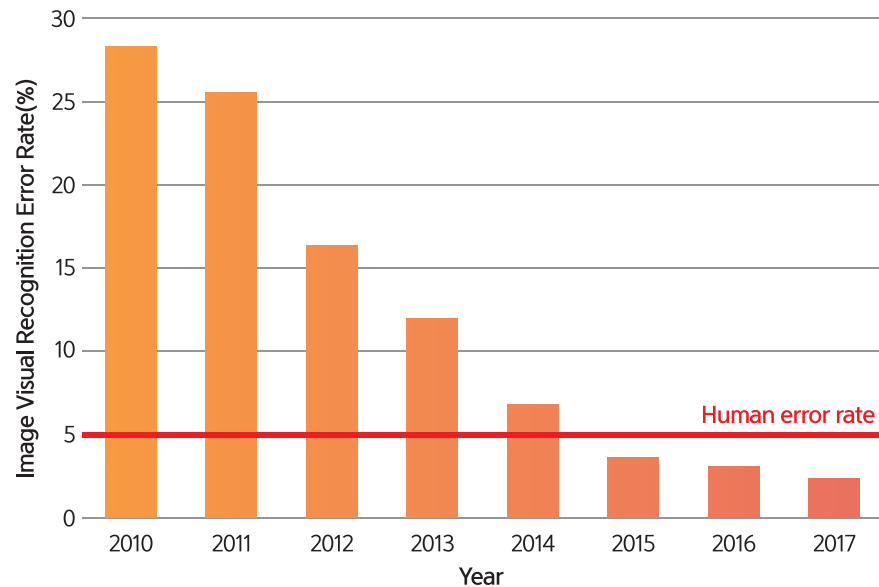
나. 고객에게 가치 있는 서비스를 제공할 때 활용되는 인공지능 기술

인공지능 기술은 회사 내부의 분석 업무에도 활용되지만, 외부적으로 고객들을 대응하는 영역에서도 활용된다. 금융 산업은 그 어떤 산업보다 고객과 밀접하게 소통하는 분야다. 특히 코로나19 팬데믹 이후 사회 분위기가 바뀌면서, ‘뉴노멀’이라 불리는 새로운 삶의 기준이 제시되고 금융 산업에서 고객과 소통하는 방식도 변화가 생겼다. 기존의 오프라인 점포에서 고객에게 서비스를 제공하던 것에서, 주로 모바일에서 고객과 소통하는 방향으로 변화했다. 앞서 소개한 데이터 분석에서는 주가와 환율 데이터 같은 정형화된 숫자 데이터를 분석하는 기술이 필요했다면, 고객과 소통을 위해서는 이미지, 음성, 또는 글자와 같은 형태가 일정하지 않은 비정형 데이터를 분석하는 기술이 필요해진다. 고객들 음성 또는 화상 통화, 또는 챗봇 등을 활용해서 금융 서비스를 받기 때문이다. 이러한 서비스를 위해서 필요한 기술은 이미지 인식, 음성 인식, 자연어 처리 기술 등이다.

1) 이미지 인식 기술

이미지 인식 기술은 사진 또는 동영상에 있는 사람이나 사물을 인공지능 기술을 활용해 인식하는 기술이다. 사람의 얼굴이나 지문을 인식해 보안 인증에 활용할 수 있다. 또는 사진 속 서류에서 글자를 인식하여 서류의 내용을 이해하는 데 활용하거나, 위성사진을 분석해 원자재 등의 투자에도 활용한다.

인공지능을 활용한 이미지 인식 기술은 앞서 소개했던 보안 인증과 서류 자동화 등에 편리하게 사용할 수 있지만, 성능이 좋지 않아서 실제 금융 산업에서 활용하는 일은 드물었다. 그런데 2013년 등장한 딥러닝 기술이 발전하면서, 인공지능을 활용한 기술이 사람의 인식 능력보다 우수하게 되면서 금융 산업에 적극적으로 도입되고 있다.

[그림 4-1-3] 2010년부터 2017년까지 ImageNet Large Scale Visual Recognition Challenge 우승팀의 오류율¹⁾

Error rates on the ImageNet Large-Scale Visual Recognition Challenge. Accuracy dramatically improved with the introduction of deep learning in 2012 and continued to improve thereafter. Humans perform with an error rate of approximately 5%.

* 출처: Olga Russakovsky et al., ImageNet Large Scale Visual Recognition Challenge, ILSVRC, (2015,1,30.)

2) AICC(AI기반 Contact Center)의 등장

앞서 소개한 바와 같이, 코로나19 팬데믹 이후로 많은 산업 분야에서 디지털 변환이 가속화되었다. 금융도 예외는 아니다. 대면으로만 가능했던 서비스를 디지털 환경에서 진행하도록 하는 것이다. 더 많은 고객에게 더 신속하게 금융 서비스를 제공하려면 AI 콜센터(AICC, AI Call Center)가 필수적이다. 사람들이 진행하던 콜센터 업무를 인공지능, 클라우드, 빅데이터 기술을 활용해서 대체하고 있다.²⁾ 즉 AICC는 인공지능이 상담사가 되어 고객의 말을 이해하고, 고객들에게 해결책을 말로 설명하는 기술을 의미한다. 이렇듯 고객들의 음성 요구사항을 인공지능이 이해하기 위해서는 음성 인식 기술이 필수적이다. 이미지 인식 기술처럼 음성 인식 기술 또한 딥러닝의 등장으로 정확도가 상승하면서 적극적으로 금융 산업에 도입되고 있다.

사실, 딥러닝 등장 전에도 컴퓨터를 활용한 챗봇이나 기계음을 활용하여 사람을 대체하려던 시도는 있었다. 하지만 딥러닝 이전의 기술로는 사람의 말을 이해하는 데 어려움이 있었다. 금융 산업의 특성상 정확도가 중요하다는 점에서 이는 분명한 한계였다. 그러다 딥러닝 기술의 등장으로 고객이 발화하는 것에 대한 성과를 보였고, 인공지능 최신 기술에 이르러서는 고객의 말뿐만 아니라 뉘앙스를 인식해서 고객의 감정까지 분석할 수 있다. 이 점을 활용

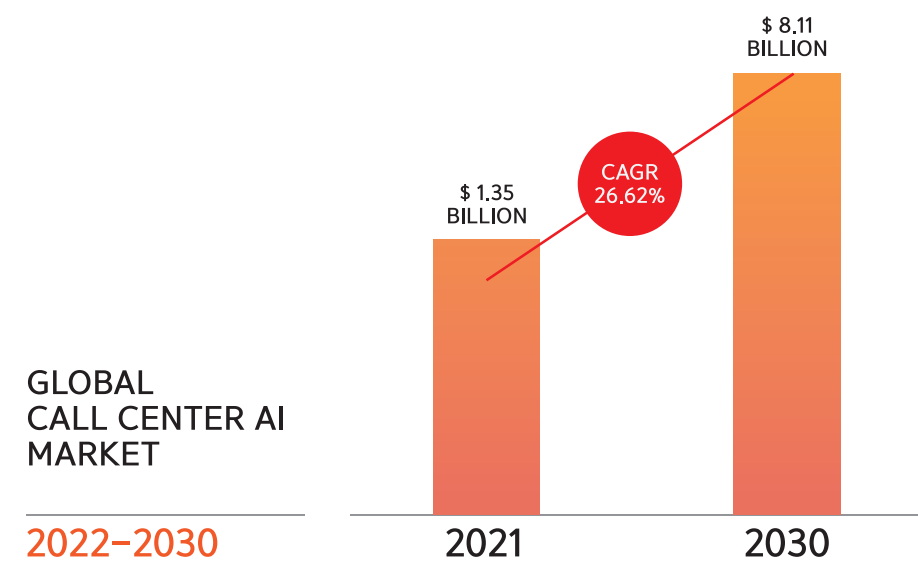
¹⁾ AI 성능을 평가하는 ImageNet의 역대 우승팀 정확도를 보여주는 자료로, 사람이 평가할 경우의 성능과 비교하고 있다. 2015년 기점으로 AI가 이미지를 분류하는 기술이 사람을 능가했다는 것을 보여준다.

²⁾ 김가은, “‘고객, 상담원 모두 만족’ 통신3사가 ‘AICC’를 주목하는 이유는” 테크M, 2023년 1월 16일, <https://www.techm.kr/news/articleView.html?idxno=105830>

하면, 고객의 기분과 감정까지 고려한 맞춤형 콜센터 서비스가 가능해진다.

앞서 소개했듯이 고객들이 말 또는 글로 문의한 질의 사항을 인공지능이 완벽하게 이해해서 해결책을 자동으로 제시하거나, 고객이 발화한 짧은 음성만을 이용해 고객의 신분을 인증하는 기술 등은 AICC의 핵심 기술이다. 이와 함께 이미지 인식 기술을 활용하여 신분증, 생체 인증, 서명 인증을 진행할 수 있는데, 이러한 기술도 AICC에 활발하게 적용된다. 결과적으로 AICC를 활용할 때 고객은 365일 24시간 쉬지 않는 금융 서비스를 경험할 것이다.

[그림 4-1-4] 글로벌 콜센터 AI 시장 규모



* 출처: “글로벌 콜센터 AI 시장 규모 분석 보고서”, verified market research, 2022. 7., 2023년 8월 25일 접속, <https://www.verifiedmarketresearch.com/product/call-center-ai-market/>

3) 자연어 처리 기술(NLP: Natural Language Processing)

금융 산업에서는 많은 정보가 문서를 통해 기록되고 전달된다. 이전에는 이런 문서들이 종이 형태였다면, 디지털 혁신이 진행되면서 전자 문서화되고 있으며, 고객도 모바일 기기에 직접 글을 입력하게 되었다. 문서 속의 글자 또는 채팅창에서 고객이 입력한 내용을 이해하는 것을 자연어 처리 기술 또는 NLP 기술이라고 부른다. AICC에서도 자연어 처리 기술이 동반되어야 고객에게 효율적인 서비스를 제공할 수 있다. 고객 서비스뿐만 아니라, 뉴스 기사를 이해해 금융 시장에 좋은 신호인지 판단하는 기술, 복잡한 경제 뉴스를 요약하는 기술, 문서 내용을 인공지능이 빠르게 이해해 질문에 대답하거나 검색을 쉽게 하는 기술 등이 모두 자연어 처리 기술로 현재 금융 산업에 적용되는 예시다. 사람들이 말로 소통하듯이, 자연어 처리 기술은 사람과 인공지능이 효율적으로 소통하는 데 활용될 뿐만 아니라, 글자 기반의 다양한 정보를 분석해 투자에 활용하는 데도 적용된다.

2. 금융 산업에서 인공지능 기술을 활용할 때 고려해야 할 사항

지금까지는 금융 산업에서 디지털 변화가 진행되면서 활용된 데이터, 인공지능 기술, 그리고 활용 분야를 살펴 보았다. 인공지능을 통해 빠르고 정확한 효율적 업무 처리가 가능해졌다는 장점이 있지만, 기술 도입에 앞서 고려해야 할 단점 역시 존재한다.

첫 번째는 인공지능 모델을 학습하는 과정에서 발생할 수 있는 사생활 침해 문제이다. 개인을 특정할 수 있는 데이터를 학습에 사용할 경우, 인공지능의 학습 결과가 개인의 사생활을 침해할 수 있다. 단순히 데이터를 익명화해 이용하더라도, 데이터의 조합 등으로 개인을 추론할 수 있기 때문에 익명화 이외에 최신 보안 기술을 활용해 개인의 프라이버시를 보호하는 것이 중요하다. 특히 금융 데이터는 소득이나, 연체 수준, 투자 정도 등 개인의 민감한 정보를 포함하기에 철저한 보안을 유지한 상태에서 데이터를 학습에 활용해야 한다.

두 번째는 인공지능 모형을 학습할 때 발생할 수 있는 데이터의 편향성이다. 특정 인종, 성별, 그리고 소득 수준에 속하는 데이터가 불균형적으로 학습에 사용될 경우에는 인공지능의 결과가 특정 계층에 편향될 위험이 존재한다. 즉 인공지능의 경우 신용 평가에서 특정 인종, 성별, 계층에 불리하게 평가를 진행할 수 있다. 금융 서비스는 의료와 더불어 개인의 삶을 좌우할 수 있는 중요한 서비스다. 그렇기에 다른 어떤 서비스보다도 정확하고 공정하게 처리되어야 한다.

마지막으로 설명 가능성이다. 인공지능은 블랙박스라는 별명도 가지고 있다. 데이터를 입력했을 때 정확한 결과가 도출되지만, 정확한 결과가 도출되는 과정을 사람들이 쉽게 이해할 수 없기 때문이다. 반면 사람이 만든 규칙은 그 논리가 보편화되어 익숙한 편이다. 그 때문에 그 결과 또한 쉽게 설명할 수 있지만, 인공지능이 고차원의 데이터를 분석해 패턴을 찾은 경우에는 설명이 어렵다. 인공지능을 금융 서비스에 도입하면 빠르고 정확하며 효율적으로 서비스를 진행할 수 있지만, 금융 산업에 인공지능 기술을 도입하기 위해서는 그 결과를 설명할 수 있어야 한다. 인공지능 기술을 이용해 신용 평가를 진행할 경우 '왜 이러한 등급이 나왔고, 등급을 올리기 위해서는 어떻게 해야 하는지' 고객에게 설명할 수 있어야 하는데, 그것이 쉽지 않다.

3. 금융 산업에서 인공지능 향후 발전 방향

인공지능 기술은 데이터를 정밀하게 이해하고 분석하는 기술에서 데이터를 생성하는 기술로 진화하고 있다. 이미지를 그리고, 글을 창작하고, 음악을 작곡하는 초거대 인공지능들이 생겨나고 있다. 초거대 인공지능의 중심에는 2022년 11월 오픈에이아이(OpenAI)사에서 출시한 Chat GPT가 있다.

초거대 인공지능 모델은 전 세계를 주름잡는 빅 테크 기업들이 대용량의 데이터를 기반으로 오랫동안 학습을 시켜 만든 인공지능 모델을 의미한다. 보통 인공지능 기술의 성능을 평가할 때 파라미터 수를 사용하는데, 파라미터 수가 많으면 인공지능이 기억하거나 할 수 있는 일들이 더 많기 때문이다. 초거대 모델은 10억 개가 넘는 파라미터

를 가진 인공지능 모델이다. 이런 모델을 만들기 위해서는 천문학적 비용과 시간 그리고 고급 엔지니어들이 필요하기 때문에, 아직은 전 세계에서 이를 만들 수 있는 기업은 손에 꼽히는 실정이다.

그럼에도 인공지능의 발전 방향과 흐름을 거스르긴 어려워 보인다. 금융 산업에서도 데이터를 생성하는 인공지능이 적극적으로 활용될 것이다. 재무제표를 자동으로 작성하고, 고객들에게 금융 리포트를 작성하는 인공지능이 본격적으로 개발될 것으로 예상된다. 인공지능 기술은 복잡한 금융 데이터를 분석하는 프로그램을 만드는 데도 활용하고, 고객들을 상대하는 가상 인간에도 적용할 수 있다. 이러한 생성형 인공지능 기술을 활용해 빠르게 금융 서비스를 만드는 능력이 금융 산업의 핵심 역량으로 발전할 것이다. 즉 새로운 인공지능 기술을 만드는 것보다, 초거대 인공지능 모델을 활용해서 빠르게 금융 서비스를 기획하고 구현하는 역량이 중요해질 것으로 예측된다.

이와 더불어 인공지능 기술의 표준화와 윤리 문제가 더욱 주목받을 것이다. 관련 애플리케이션에 접속해 다양한 금융사의 데이터를 활용한 금융 서비스를 받을 수 있듯이, 인공지능을 적용하기 위해 표준화된 학습데이터가 필요하다. 따라서 데이터에 대한 표준이 정립될 것으로 기대된다. 실제로 '산업 인공지능 표준화 포럼'이 국내에서 창립되었고, 세계적으로 인공지능 데이터에 대한 표준을 논의하고 있다.

인공지능 윤리도 중요하다. 앞서 언급한 바와 같이 편향 없는 인공지능 모델을 만들기 위해서는 인공지능 모델을 만들 때 관련 윤리를 고려해야 할 것이다. 인공지능 기술이 금융 산업에 적용될 때 소외되는 계층이 발생해서는 안 되기 때문이다.

마지막으로 신뢰할 수 있는 인공지능 기술도 주목해야 한다. 인공지능을 통해 의사 결정이 진행될 때, 그 결과를 신뢰할 수 있어야 한다. 이 신뢰의 기반은 결과를 제대로 설명할 수 있는 것에서 시작한다. 인공지능 기술이 만든 판단을 설명하는 것에 대한 연구 분야인 '설명가능 인공지능(XAI, explainable Artificial Intelligence)'은 금융 산업에서는 더욱 주목받을 것이다.

종합하면, 금융 분야에서 인공지능 기술은 데이터 활용을 폭넓게 하고 더욱 정교한 예측과 분석이 가능한 방향으로 발전하고 있다. 이를 통해 더 나은 개인 맞춤형 금융 서비스와 투자 컨설팅을 제공할 것이다. 인공지능이 발전할수록 금융 기관과 투자자에게 보다 효율적이고 정확한 서비스를 제공하며, 금융 시장의 안정성과 효율성을 높이는 데에 기여할 것으로 기대한다.

제2장 헬스케어분야 데이터 활용 현황

정규환 조교수 성균관대학교 삼성융합의과학원

의학 기술이 발전하면서 수명이 늘고 의료 접근성이 향상했다. 이에 따라 의료 데이터의 규모와 종류 모두 폭발적으로 증가했기에, 정확한 진단과 치료 의사결정을 위해 인공지능 기술의 활용이 더욱 중요해졌다. 이런 변화에 맞춰 지난 수년간 인공지능 솔루션의 개발 및 임상 검증, 임상 적용이 빠르게 진행되었다. 헬스케어와 의료 분야에서 대량의 의료데이터에 기반하여 이를 활용하려는 것으로, 최근에는 생성적 인공지능이 활발히 연구되고 있다. 대표적으로 챗지피티(ChatGPT)와 같은 초거대 언어모델이 다양한 분야에서 높은 성능을 보이며, 헬스케어와 의료 분야에서 활용 가능성을 타진하고 있다. 더 나아가 다양한 데이터를 복합적으로 분석할 수 있는 다중모달(Multi-modal) 모델의 발전도 가속화되고 있다. 이를 통해 헬스케어 분야에서 기반 모델(Foundation)의 개발 가능성이 커지고 있을 뿐 아니라, 높은 정확도로 다양한 작업을 수행하는 일반 인공지능(General Artificial Intelligence)의 구현 가능성도 커지고 있다.

1. 의료인공지능의 현황

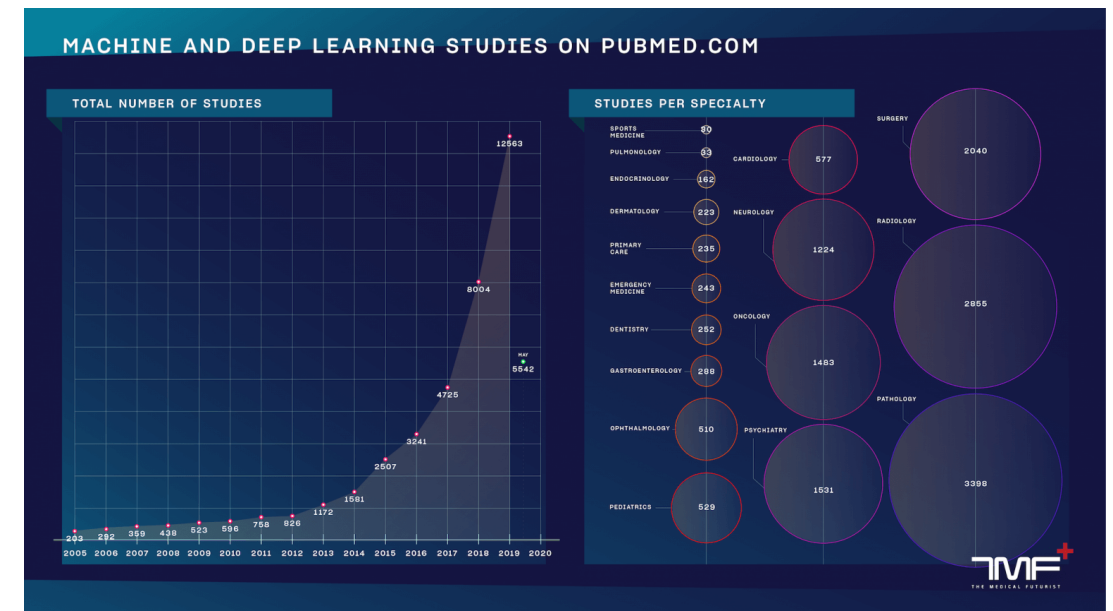
가. 의료인공지능의 필요성

정확한 진단과 치료 방법을 결정하기 위해 환자에게서 다양하고 높은 해상도의 데이터를 획득하는 방향으로 의료기술이 발전했다. 의료데이터는 텍스트, 영상, 생체신호 등의 시계열 데이터를 아우르며 다양해지고, 수명 연장 과 의료접근성 개선이라는 현상이 맞물려 의료데이터 규모는 급속도로 커지고 있다. 이러한 의료데이터는 본질적으로 빅데이터는 4V(Volume, Variety, Velocity, Veracity)의 속성을 보인다.

이처럼 의료데이터가 지나치게 다양해지고, 그 규모가 급속도로 커지면서 의료데이터 기반으로 최적의 의사결정을 내리는 것이 오히려 부담스러워졌다. 따라서 복잡하고 다양한 의료데이터의 효율적인 분석을 통해 임상적 의사결정의 일관성과 정확도를 개선하는 도구가 점점 더 필요해지고 있다.

최근 10년간 의료 전 분야에 걸쳐 의료 인공지능 관련 연구의 수가 폭발적으로 증가하였다. △대용량의 디지털화된 의료데이터, △효율적인 연산 하드웨어의 발전, 그리고 △의료데이터에 특화된 인공지능 알고리즘의 발전 등 긍정적 요인이 시너지를 낸 덕분이다. 특히 기존에 디지털화된 정보가 축적된 영상의학, 병리학 등의 분야에서 관련 연구를 선도했다. 이러한 학술 연구의 일부 성과로 소프트웨어로 개발되어, 현재 각국 규제기관의 승인을 받고 소프트웨어 의료기기(Software as a Medical Device)로 임상 현장에서 사용되고 있다.

[그림 4-2-1] 의료분야 기계학습 및 딥러닝 관련 논문 수의 증가



* 출처: "AI, Machine Learning & Deep Learning: The Number of Medical AI studies is on a rise," The Medical Futurist, 2023년 9월 10일 접속, <https://medicalfuturist.com/infographics/machine-learning-in-healthcare/>

나. 인공지능 기반 소프트웨어 의료기기 현황

각국의 규제기관 승인을 받은 의료기기의 수는 해마다 꾸준히 증가하여 미국의 경우 2022년 10월 말 기준 521개의 인공지능 및 기계학습 기반의 의료기기가 승인되었다.¹⁾ 한국의 경우엔 2018년 5월 최초 품목허가 사례 이후 빠르게 증가하여, 2022년 6월 기준 114개의 인공지능 기반 소프트웨어 의료기기가 품목 허가를 획득하였다. 현재 진행되는 임상시험 현황을 고려하면 향후 인공지능 기반의 소프트웨어는 품목 종류와 품목허가 건수 모두 가파르게 증가할 것으로 예상된다.

의료분야에서 이러한 솔루션이 임상 현장에 빠르게 도입되려면 지급체계의 구축이 매우 중요한데, 미국의 경우에는 2020년 Digital Diagnostics사의 당뇨병 망막병증 자동 스크리닝 솔루션인 IDx-DR이 인공지능 기반 의료기기로서는 최초로 정식 수가를 받는 사례가 되었다. 또한 Viz.ai사의 뇌혈관 질환 진단 보조 시스템인 Viz LVO는 혁신 의료기기에 대한 보조 수가인 신기술추가지불보상(NTAP, New Technology Add-on Payment)의 최초 대상이 되었다. NTAP 대상은 꾸준히 증가하고 있다.

1) "Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices," FDA, 2022. 10. 5., <https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-ai-ml-enabled-medical-devices>

[표 4-2-1] 국내 인공지능 기반 소프트웨어 의료기기 허가 품목 현황(2022년 6월 14일 기준)

구분		'18	'19	'20	'21	'22	계
계		4	10	44	37	19	114
진단 보조	소계	4	6	24	33	18	85
	의료영상	4	6	22	31	17	80
	생체신호	—	—	1	—	1	2
	병리조직	—	—	1	2	—	3
예방·치료		—	4	20	3	1	28
질환·예측		—	—	—	1	—	1

* 출처: 식약처, 디지털 치료기기 등 소프트웨어 의료기기 개발 증가, (2022.6.16).

[표 4-2-2] 미국 인공지능 기반 소프트웨어 의료기기의 수가 적용 사례

Table 1. Selected AI devices that are reimbursed by US Medicare.

Manufacturer	Technology	Description	Payment mechanism	Year reimbursement granted
Digital diagnostics	IDX-DR	Deep learning algorithm to diagnose diabetic retinopathy from fundoscopic images in the outpatient setting	CPT	2020
viz.ai	Viz LVO	Radiological computer-assisted triage and notification software that analyzes CT images of the brain and notifies hospital staff when a suspected large-vessel occlusion (LVO) is identified	NTAP	2020
Rapid AI Caption health	Rapid LVO Caption guidance	AI-guided medical imaging acquisition system intended to assist medical professionals in the acquisition of cardiac ultrasound images	NTAP NTAP	2020 2021
viz.ai	Viz SDH	Radiological computer-assisted triage and notification software that analyzes CT images of the brain and notifies hospital staff when a suspected subdural hematoma is identified	NTAP	2022 (candidate)
Rapid AI	Rapid aspects	Computer-aided diagnostic device characterizing brain tissue abnormalities on brain CT images	NTAP	2022 (candidate)
AIDoc	Briefcase for PE	Radiological computer-assisted triage and notification software that analyzes CT images of the chest and notifies hospital staff when a suspected pulmonary embolism is identified	NTAP	2022 (candidate)
PROCEPT BioRobotics Corporation	The AQUABEAM system	Autonomous tissue removal robot for the treatment of lower urinary tract symptoms due to benign prostatic hyperplasia (BPH)	NTAP	2020

* 출처: R. B. Parikh, L. A. Helmchen, "Paying for artificial intelligence in medicine", npj Digit. Medicine Vol.5 No.63, 2022.

국내에서는 2019년 영상의학 분야, 2020년 병리학 분야에서 AI 기반 의료기술에 대한 요양급여 여부 평가가 이드라인이 발표되었으나, 대부분을 차지하는 진단 보조 기술은 기존 기술로서 수가 고려를 위한 대상이 되지 않는 것으로 판단된 바 있다. 하지만 이후 신의료기술평가 유예제도나 혁신 의료기기 통합심사제도를 통해 기존 의료행

위가 아닌 새로운 의료행위를 가능케 해주는 여러 솔루션이 비급여로 시장에 먼저 진입하고, 수가 적용을 위한 근거로 다양한 사례가 꾸준히 쌓이면서 국내에서도 인공지능 기반 의료기기에 대한 수가 적용 가능성을 높이고 있다.

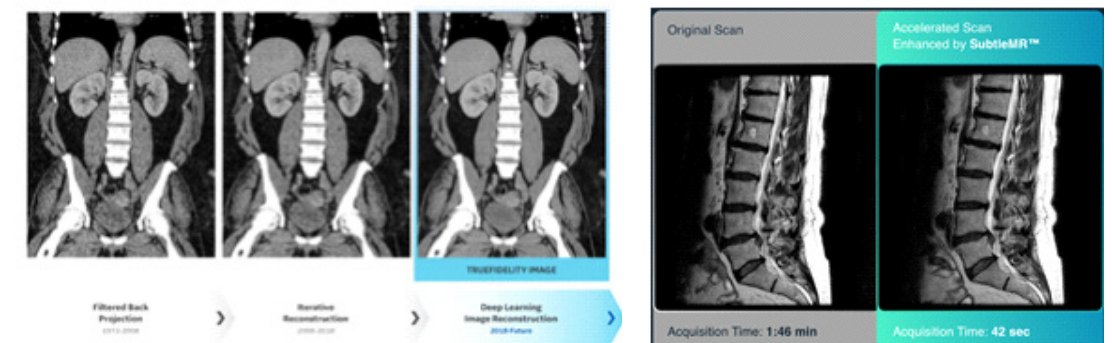
2. 의료 인공지능의 적용 사례

가. 의료데이터의 생성

대부분의 의료 인공지능 기술이 기존에 생성된 데이터의 분석에 활용되지만, 일부의 경우 의료데이터의 생성 단계에서 데이터의 품질을 높이거나 데이터 획득의 효율성을 높이는 데 활용된다. 대표적인 사례로 우선 인공지능 기술을 이용하여 저선량 전산화단층촬영(CT) 영상의 품질을 개선하는 솔루션을 들 수 있다. 또한 자기공명영상(MRI)의 획득 시간을 획기적으로 줄이는 솔루션도 있다. 이 솔루션들은 이미 다양한 영상진단 장비 생산자들이 상용화했고, 임상 현장에서 활용하고 있다. 그런가 하면 영상 간의 변환을 통해 MRI 기반으로 CT 영상을 생성하여 시간적·경제적 비용과 방사선 노출 빈도를 낮추는 등의 사례도 있다.

[그림 4-2-2] (좌) GE Healthcare사의 CT품질 개선 솔루션인 TrueFidelity

(우) Subtle Medical사의 MRI 가속 솔루션인 SubtleMR



* 출처: (좌) "Introducing a new era of image reconstruction," GE Healthcare, 2023년 9월 10일 접속, <https://www.gehealthcare.co.kr/products/computed-tomography/truefidelity>

(우) "AI Solutions for MRI and PET," Subtle Medical, 2023년 9월 10일 접속, <https://subtlemedical.com/>

나. 의료데이터의 처리 및 변환

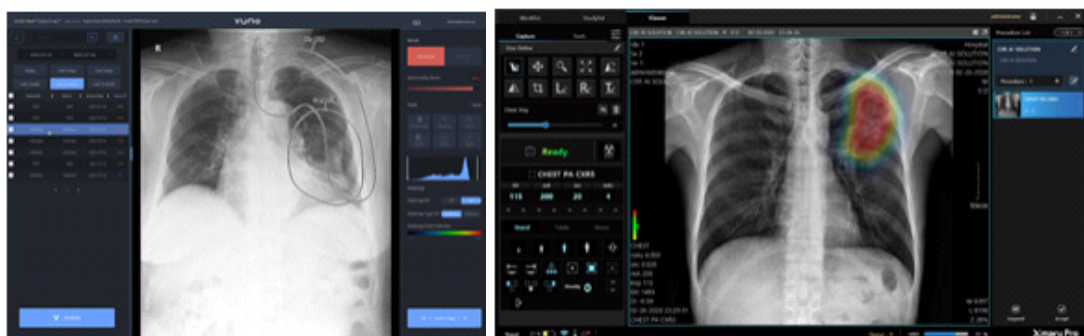
기존에 구축된 의료데이터의 활용성을 높이기 위해 데이터의 처리 및 변환에도 인공지능을 널리 활용한다. 대표적으로 의료영상의 해상도를 개선하거나 장비마다 다른 영상의 특성을 보정하는 등에 사용한다. 특히 학습 데이터에 의존적인 인공지능 모델의 특성상, 장비와 데이터 수집 환경의 변화에 따라 인공지능 모델의 성능이 민감하게 변화하는데, 별도의 인공지능 모델을 통해 입력 데이터의 일관성을 높임으로써 실제 임상 환경에서 의료 인공지능 솔루션의 성능을 안정화할 수 있다.

다. 진단 보조

진단 보조 솔루션은 규제기관의 인허가를 획득한 인공지능 기반 의료기기 중 가장 큰 비중을 차지하는 것으로, 의료진의 진단 정확도와 효율성을 높이는 것을 목적으로 개발되고 있다. 특히 의료영상을 분석하여 병변의 위치를 찾거나 정량화하고, 질환의 종류를 구분하는 솔루션들은 이미 다수 개발되어, 대중을 대상으로 하는 선별 검사나 암 환자를 위한 정밀 진단에도 폭넓게 활용된다. 더 나아가 심전도 등을 포함한 각종 생체 신호를 토대로 향후 발생할 질환이나 응급 사태 등을 예측하는 솔루션도 꾸준히 연구 개발되고 있다. 특정 치료에 대한 반응을 예측함으로써 최적의 치료 의사결정에 도움을 주려는 것이다.

[그림 4-2-3] (좌) VUNO사의 흉부 X선 영상 진단보조 솔루션의 독립 사용자환경

(우) 레이언스사의 장비에 탑재된 사용자환경



* 출처: (좌) "VUNO Med-Chest X-ray," vuno, 2023년 9월 10일 접속, <https://www.vuno.co/chest>
(우) "Xmaru pro with AI," rayence, 2023년 9월 10일 접속, https://www.rayence.com/sw_xmarupro

그런가 하면 동일한 의료 인공지능 모델도 임상 현장의 상황에 따라 다양한 방식으로 도입되고 있다. 간혹 의료 인공지능이 별도의 새로운 사용자 환경(UI)을 가지고 있는 사례도 있지만, 일반적으로는 병원의 기존 전자의료 기록(EMR, Electronic Medical Record)이나 의료영상저장 전송시스템(PACS, Picture Archiving and Communication System)과 연계되어 병원 시스템에 탑재된다. 때로는 의료데이터가 획득되는 장비 자체에 탑재되어, 검사 즉시 현장에서 결과에 대한 판정 결과를 제시해 주는 방식으로 솔루션들이 개발되기도 한다. 이는 긴급한 진단이 필요하거나, 의료 인프라가 부족한 곳에서 유용하게 활용된다.

3. 생성적 모델 및 기반 모델을 이용한 의료 분야 혁신

가. 생성적 인공지능(Generative AI)과 기반 모델의 정의 및 특징

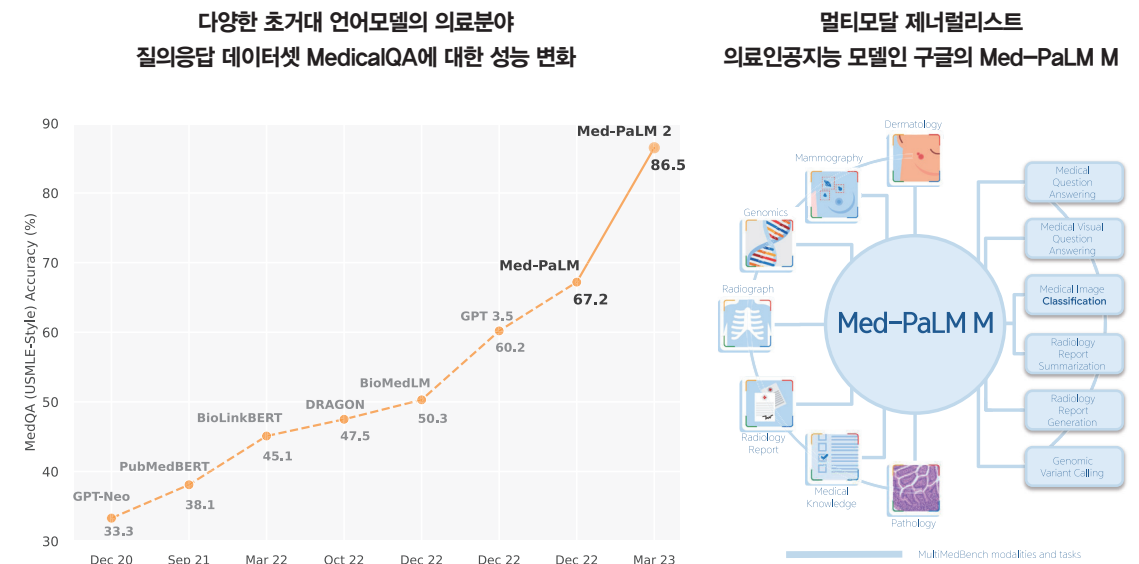
기존의 인공지능 모델, 특히 의료기기로서 의료분야에 활용되는 모델은 대부분 주어진 입력 데이터를 분류하거나 특정 값의 예측 등을 목표로 학습된 판별적 모델(Discriminative Model)이다. 하지만 최근 들어 원하는 조건에 맞는 텍스트나 영상 등의 데이터를 생성해 내는 생성적 인공지능이 빠르게 발전하고 있다. 대표적으로 텍스트 분야에

서는 오픈AI사의 초거대 언어모델(LLM, Large Language Model)로 GPT 기반의 챗봇인 챗지피티(ChatGPT)가 유명하다. 또 구글의 초거대언어모델인 팜(PaLM)에 기반한 바드(Bard)가 있다. 영상 분야에서는 오픈AI사의 달리(DALL·E)나 스테빌리티 AI(Stability AI)사의 스테이블 디퓨전(Stable Diffusion) 등이 잘 알려져 있다.

이런 모델들의 개발은 막대한 양의 학습 데이터를 이용한 자기지도 학습(Self-supervised Learning)을 통해 이뤄지는데, 모델의 규모가 커질수록 명시적으로 학습시키지 않은 다양한 과업(Task)에 대해 높은 성능을 보이는 특징이 있다. 예를 들어 챗지피티의 경우에도 기반이 되는 지피티3(GPT-3)는 주어진 단어의 나열에 대한 다음 단어를 예측하는 과업으로 학습되었지만, 모델의 규모가 커지고 학습 데이터가 충분히 많아지면서 번역이나 요약, 문서 작성, 질의응답 등 다양한 자연어 처리 과업을 높은 성능으로 처리할 수 있다.

또한 소량의 추가적인 데이터를 활용한 미세조정(Fine-tuning)을 통해 특정 분야의 성능을 추가로 개선할 수 있다. 이렇게 다양한 분야에서 활용 가능한 '하나의 기본이 되는' 모델의 성격을 지니므로 기반 모델(Foundation Model)이라고도 불린다. 최근에는 자연어와 영상 각각의 기반 모델뿐 아니라 시각-언어(Visual-Language) 기반 모델도 지속적으로 개발되어, 최소한의 예시 제공이나 미세조정을 통해 더 고차원적인 과업을 높은 정확도로 수행하는 사례가 발표되고 있다. 특히 CLIP(Contrastive Language-Image Pretraining) 방식으로 학습된 모델의 경우 시각적 정보와 언어적 정보의 연관성을 모델링함으로써, 추가적인 예시 제공이나 학습 없이도 새롭게 주어진 영상에 대해 광범위한 분류 작업을 수행할 수 있다는 사실이 다양한 연구를 통해 발표되고 있다.

[그림 4-2-4] 기반 모델의 의료분야 적용사례

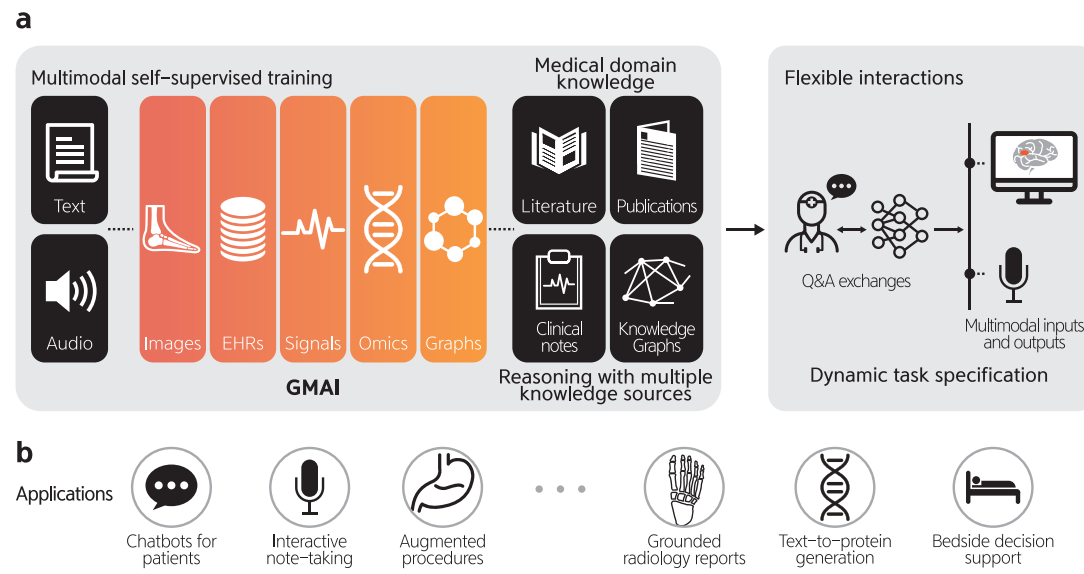


* 출처: (좌) K. Singhal et al., Towards Expert-Level Medical Question Answering with Large Language Models, Arxiv, p.2.
(우) T. Tu et al., Towards Generalist Biomedical AI, Arxiv, (2023).

나. 기반 모델의 의료분야 적용 사례

기반 모델의 특성상 추가 학습 없이도 다양한 분야에서 높은 성능을 보이므로, 초기에는 챗지피티(ChatGPT)와 같은 초거대 언어모델을 그대로 의학적 판단에 적용한 결과가 발표되었다. 특히 미국 의사면허 시험인 USMLE에서 최신 초거대 언어모델인 GPT-4가 커트라인 수준을 훨씬 초과하는 성적을 보이는가 하면, 다양한 임상분야의 질의 응답에서 전문의의 수준에 준하는 답변을 생성한다는 보고가 이어지고 있다. 구글도 Med-PaLM을 발표하였는데, 초거대 언어모델인 PaLM으로 추가적인 학습과 튜닝을 거친 뒤 의료 분야에 최적화한 경우였다. Med-PaLM의 경우 평균적으로 의료진의 답변 대비 만족도가 더 높은 의학적 응답을 생성할 수 있었다. 또 최근에는 Med-PaLM M을 발표하여, 범용 의료 인공지능의 가능성을 구체화했다. 범용 의료 인공지능의 수준에서는 단순 언어뿐 아니라 다양한 의료영상을 통합적으로 분석하여 광범위한 의료분야 과업을 수행하는 것을 기대할 수 있다.

[그림 4-2-5] 기반 모델을 이용한 제너럴리스트 의료인공지능의 개념도



Regulations: Application approval; validation; audits; community-based challenges; analyses of biases, fairness and diversity

* 출처: M. Moor, O. Banerjee, Z.S.H. Abad et al., "Foundation models for generalist medical artificial intelligence," Nature Vol.616, 2023, pp.259-265.

4. 의료 인공지능의 향후 전망

지금까지 살펴본 바와 같이, 초기 의료 인공지능은 의료기관에 축적된 대량 데이터를 이용하여 특정한 목적의 인공지능 모델을 개발하고, 임상 검증을 통해 의료기기로서 임상에 도입하는 과정을 거쳤다. 하지만 이러한 과정에서 학습을 위한 정답 데이터를 구축하기 위해 많은 비용과 시간이 필요해졌다. 또 미리 정한 목적 이외의 새로운 기능을 추가하려면 이러한 과정을 반복해야 하는 한계가 있었다.

하지만 생성적 인공지능과 기반 모델의 발전을 통해 대량의 의료데이터를 활용한 의료기반 모델의 연구 개발이 가속화되고 있으며, 궁극적으로 제너럴리스트 의료 인공지능(Generalist Medical AI)의 등장이 가시화되고 있다. 이를 통하면 의료진과 자연어를 이용해 유연하게 상호작용하면서도 다양한 데이터를 포괄적으로 이해하는 수준에 도달할 것으로 보인다.

물론 아직은 기반 모델을 개발하고 유지하는 데 필요한 막대한 연산 자원과 비용, 개인정보나 환자 안전에 대한 우려, 오류 발생 시 불명확한 책임 소재 등 다양한 이슈가 남아 있다.

이에 기반 모델을 활용한 의료기기의 안전성과 유효성을 평가하는 명확한 기준이 마련되어야 한다는 공감대가 형성되고 있다. 향후 수년간 다양한 이해당사자들 간의 소통과 협력이 활발해질 것으로 보인다. 이를 통해 다양한 기술적 규제적 발전을 기대해볼 만하며, 궁극적으로는 의료의 수준을 한 단계 더 높이는 결과를 끌어내는 데 핵심적인 역할을 할 것으로 전망한다.

제3장 모빌리티분야 데이터 활용 현황

아우토크립트

데이터는 수집 · 가공 · 분석을 통해 정보, 지식으로 발전하며 그 정보적 가치를 높인다. 자동차 및 모빌리티 사업 분야에서 데이터의 최대 경영적 가치는 '급격한 변동성 해소'의 역할을 한다는 점이다. 또한 일선에서는 데이터 분석을 통해 연구 · 개발, 제조 · 조립, 시장 및 공급망 관리 등 공정을 개선한다. 모빌리티 데이터는 그 내용에 따라 차량, 이동 데이터로 분류된다. 활용 목적에 따라서도 데이터를 적절하게 취급해야 한다. 데이터마다 성질이 판이하기 때문이다.

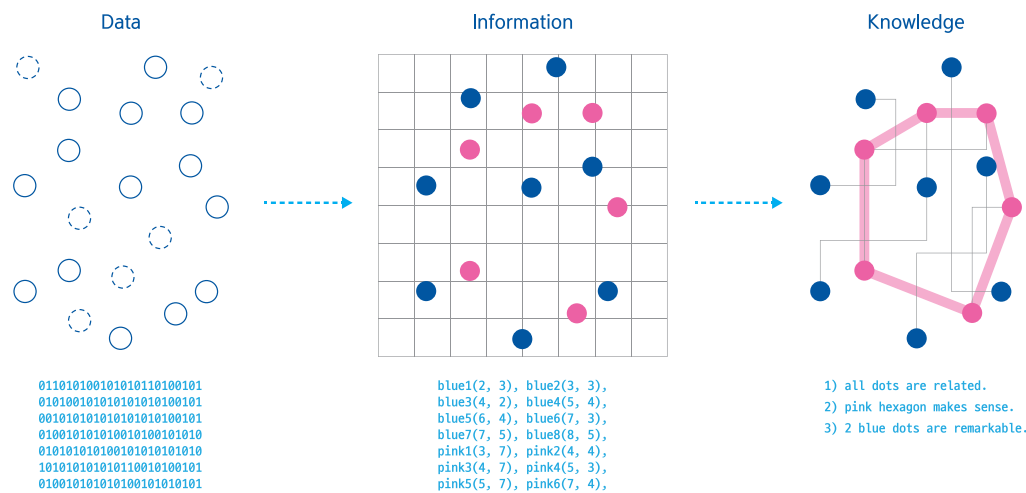
모빌리티 데이터 비즈니스의 대표적 예로는 'SDV(Software Defined Vehicle) 데이터 플랫폼'과 'MaaS(Mobility as a Service) 비즈니스'를 들 수 있으며, 공공 부문에서는 데이터를 통해 각종 사회 문제, 특히 도시 문제를 해결한다.

1. 데이터 가치

가. 데이터의 정의

'데이터(Data)'란 어떤 이론을 수립하는 데 기초가 되는 사실 또는 그 바탕이 되는 자료다. 그중에서도 컴퓨터가 처리할 수 있는, 즉 기계가 읽을 수 있는(Machine-readable) 문자, 숫자, 소리, 그림 등의 자료를 '데이터'라 부르는 것이 통상적이다.

[그림 4-3-1] 자료 → 정보 → 지식



* 출처: 아우토크립트 제작

'정보'가 아닌 '자료'임에 유의해야 한다. 자료(data)를 가공해 얻는 것이 정보(information)이고, 나아가 정보를 분석해 얻는 것이 지식(knowledge)이다. 일상에서는 '자료'와 '정보'란 말을 혼용해 쓰는 편이지만, 통계 분야에서는 둘을 엄격하게 구분하는 편이다.

통계 외 분야에서도 둘의 구분은 엄격하다. 일례로 '개인정보(personal information) 보호법'은 있지만 '개인 자료(personal data) 보호법'은 없다. 자료로부터 내용이 정리되고 구획된 정보에서는 해당 내용의 주체, 즉 어떤 개인 신원을 특정할 수 있기 때문이다. 따라서 일반적인 자료 형태를 띠고 있더라도 그로부터 특정인 신원을 알아낼 수 있다면 정보로 분류된다. 이는 아래 기술할 모빌리티 데이터, 특히 '이동' 데이터의 성질과 매우 유관하므로 주의할 부분이다.

이하 본문에서는 편의상 가공 · 분석된 정보 및 지식까지 포함하여 '데이터'로 통칭하려고 한다. 비록 데이터라고 칭했더라도 이는 결코 저절로 얻어지지 않고, 일정한 기술적 경지에 이른 가공 · 분석 작업의 가치까지 포함한 뜻을 강조해 둔다. 따라서 본문의 '데이터'란 말은 충분한 데이터 가공 · 분석 역량을 갖춘 조직의 경우를 전제한 용어다.

나. 데이터의 가치

1) 데이터 자체의 가치

데이터는 그 자체로도 가치를 지닌다. 데이터라는 무형 자산의 가치 평가에는 데이터 유형과 특성, 사업 절차, 활용 목적과 용도에 따라 여러 평가 방법이 동원된다. 대표적으로 수익 접근법, 원가 접근법, 그리고 시장 접근법을 떠올릴 수 있다.

수익 접근법은 데이터 활용으로 발생할 미래 효익을 현재 가치로 환산하는 방식이다. 원가 접근법으로는 대상 데이터나 동일한 경제성을 지녔다고 평가되는 데이터의 생산 또는 구입 금액을 산정하며, 시장 접근법의 경우 유사 데이터의 기존 거래 사례와 비교하여 가치를 산정하는 방법을 취한다.

하지만 이하 본문에서는 데이터 자체의 금전적 가치보다는 데이터 활용에 따른 무형적 가치, 특히 자동차 및 모빌리티 산업 분야에 한정해 그 경영적 가치를 해설하겠다.

2) 자동차 산업에서 데이터의 경영적 가치

기업 경영의 관점에서 데이터 분석 작업으로 급격한 환경 변화를 예측하여 급작스러운 변동성 위험을 막을 수 있다는 점이 매력적이다. 즉 급증하는 변동성에 대한 해소 기대감을 가장 큰 가치로 꼽을 수 있다.

예를 들어 자동차 산업의 경우 이전과 완전히 달라진 시장과 사업 환경 탓에 데이터의 중요성이 더욱 부각되었다. 전기차 보급에 따라 오랫동안 제품의 핵심이었던 엔진 대신, 모터와 배터리가 부품 체계의 중심에 섰다. 그러다 보니 기존에 경쟁자가 아니었던 IT나 가전 등 여타 분야 기업들이 경쟁자를 자처한다. 여기에 원자재 및 임금 상승 등으로 수익성은 나빠지고 있으며, 환경 영향이나 지속 가능성 등 기업에 요구되는 신종 규범의 압박 요소도 외면할 수 없다.

이처럼 총체적 변동성이 전에 없이 큰 부담으로 작용한다. 데이터의 경영적 가치가 절실하게 요구되는 시점이다.

변동성이란 예측 불가능성이지만, 그럼에도 이제는 확실해졌다고 보이는 요소도 있다. 이를테면 커넥티비티(Connectivity), 자율주행, 공유 모빌리티(Shared Mobility) 등 확실한 미래의 트렌드가 모두 하드웨어가 아닌 소프트웨어 관련 주제라는 사실도 데이터 가치를 높였다. 자동차 산업이 하드웨어보다 소프트웨어와 서비스 수익성 중심으로 재편되리라는, 즉 'SDV(Software Defined Vehicle) 시대'의 예견 또한 데이터의 가치를 한층 더 높인다.

이제 자동차 기업은 데이터 분석으로 변동성을 해소하려는 목적에서 더 나아가, 데이터를 활용한 새로운 가치 창출까지 기대하고 있다. 데이터를 확보하고 분석하는 과정이 '미래 시장 생존을 위한 필수 자원이자 역량'이라는 인식은 이제 업계의 상식으로 통한다.

2. 모빌리티 데이터

'모빌리티(Mobility)'란 모든 종류의 이동이다. 이는 무엇이 언제 어디에 있는지를 따지는 시공간 개념으로, 육·해·공 모든 교통수단을 아우른다. 또한 도보 이동뿐 아니라, 이동과 이동 사이 즉 정지 상태까지도 포함한다. '모빌리티 데이터(Mobility Data)'란 모빌리티 활동 과정에서 생산되는 모든 데이터를 말하며, 그 내용에 따라 개략하여 차량 데이터, 이동 데이터로 나뉘 볼 수 있다.

가. 차량 데이터

'차량 데이터'는 차량 환경 전반에 관련된 데이터로, 디지털 기기화된 차량에서 생산되는 데이터뿐 아니라 차량 제조-판매 전 과정의 디지털화에 따른 모든 데이터를 말한다.

기업은 데이터 분석을 통해 시장 수요를 예측하여 생산을 조절하고, 고객 취향 분석을 통해 시장 변화에 유연하게 대처한다. 미국 자동차 제조사 '포드(Ford)'는 '구글(Google)'과 손잡고 전 분야 공정의 디지털 혁신을 추진하고 있다. 구글의 AI(Artificial Intelligence) 역량을 활용하여 제품 기획, 공급망 관리, 연구·개발, 제조·조립, 유통, 애프터마켓(Aftermarket) 등 전 과정의 효율을 개선하고, 단계마다 최적의 포트폴리오 수립을 꾀한다. 이로써 미래형 자동차 공장 프로세스를 이룩하고 효율을 극대화하는 것이 최종 목표다.

일선 현장 관점에서 데이터의 효용성은 특히 가상 시험에서 두드러진다. 가상 환경에는 물리적 제한이 없고 혹여 실패하더라도 실제 시험에 비해 리스크가 적다. 그 때문에 더 다양한 시도를 간편하고 부담 없이 진행할 수 있다. 이를테면 'HILS(Hardware in the Loop System)' 검증이 그러하다. HILS 검증이란 실제 테스트 전에 가상으로 차량 환경을 구축하고 그 동작을 검증하는 과정으로, 차량 ECU(Electronic Control Unit)의 수가 급증하면서 과도하게 높아진 테스트 복잡도와 한계 상황을 극복하기 위해 마련된 것이다.

이때 주의할 점이 있다. 가상 시험 진행 시에는 기존 물리적 시험 과정을 그대로 재현하듯 특정 시나리오의 정상 이행 여부만 판단하고 그치는 관습이 있는데, 이는 시대 변화에 한참이나 뒤떨어진 일이다. 차량의 커넥티비티 즉 내외부 연결성이 높아지면서 소프트웨어적 돌발 상황이 빈번해졌다는 점에 주목해야 한다. 이에 따른 문제를 검증하려면 IT 분야의 테스트 방법론까지 추가해 검증해야 한다.

이를테면 '퍼징(Fuzzing)' 테스트가 필수다. 퍼징이란 비정상 데이터를 무작위로 생성·입력해 오작동을 유도함으로써 시스템의 취약성을 찾아내는 작업으로, 예기치 않은 외부 침입으로 시스템이 오염되는 상황을 파악하고 예전엔 발견된 적 없던 취약점까지 찾아내는 데 유용하다.

자동차 산업의 중심이 모빌리티로 이동하는 현상의 근간에도 데이터가 있다. 자동차 제조사는 소프트웨어 중심의 수익성을 노리는 사업 구조 변화에 따라 차량 판매 이후 일어나는 각종 서비스 사업, 즉 애프터마켓 데이터를 적극적으로 활용할 계획을 마련해 두고 있다. 이는 아래 기술할 'SDV 데이터 플랫폼'의 사업적 가치와 상통한다.

이렇듯 데이터는 자동차 산업의 경영·기술적 의사결정, 상품성 향상, 마케팅, 연구·개발 및 제조·조립 등 모든 분야에서 폭넓게 활용되고 있다. 그리고 데이터는 미래 차 개념의 가장 중요한 두 키워드, 즉 V2X(Vehicle-to-Everything) 커넥티드 카와 자율주행 구현의 핵심 요소다. 이는 수많은 전문가가 "데이터를 장악한 기업이 미래 모빌리티 시대를 주도할 것"이라고 예측하게 된 이유다.

나. 이동 데이터

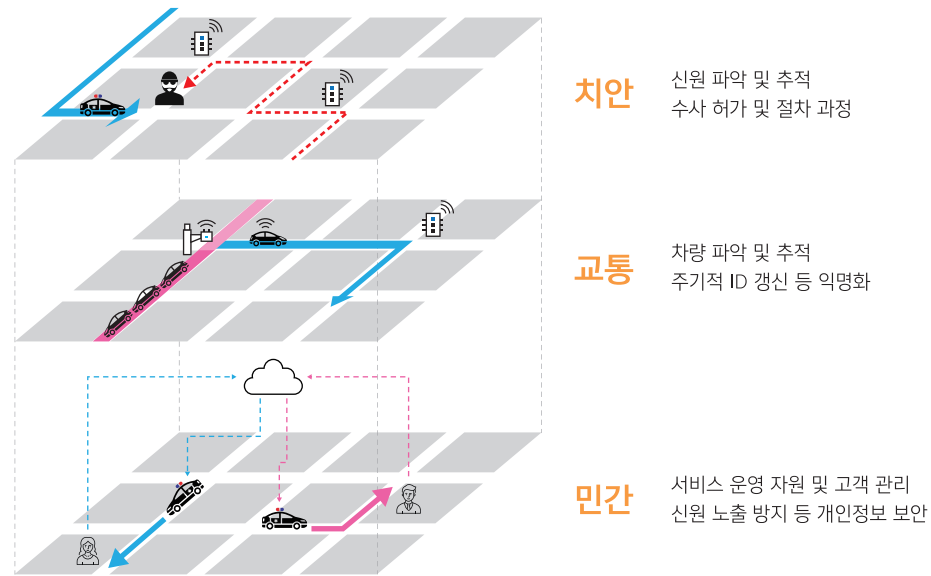
'이동' 데이터는 모빌리티 데이터 중 가장 중요하고, 모빌리티 데이터의 핵심 특징을 보인다. 바로 '다층 레이어(Multi-layer) 구조의 디지털 트윈(Digital Twin)'이다. 데이터는 그 활용 목적에 따라 서로 다른 레이어로 구분되고, 각 층의 데이터 취급 방법은 층마다 다르다.

'디지털 트윈'이란 우선 현실 세계를 디지털 가상 세계에 구현하고 실제 객체와 동기화하여 현실의 일을 모니터링하는 데에 활용된다. 더 나아가 이를 활용하여 시뮬레이션하면 미래 상황을 예측하여 잠재된 문제를 회피할 수 있다. 이를 바탕으로 할 때 의사결정의 품질은 높아진다.

모빌리티 객체가 이동하며 생산하는 시공간 데이터, 즉 모빌리티 데이터 또한 지도 형태의 가상 세계를 형성하고, 그 가상 세계를 관찰하여 실제 세계의 상황을 파악하고 예측하는 것이다.

'다층 레이어 구조'에선 모빌리티 데이터가 그 활용 목적에 따라 여러 층을 이룬다. 이때 층마다 수집·가공·분석되는 데이터의 종류가 다르다. 따라서 적용해야 하는 데이터 취급 방법 및 보안 조치도 각각 다르다. 아래 그림을 보자.

[그림 4-3-2] 모빌리티 데이터의 용례



* 출처: 아우토크립트 제작

데이터 활용 목적과 그 취급 방법, 특히 보안 조치가 서로 완전히 다른 3개 층의 예다. 물론 실제 모빌리티 데이터의 경우 이러한 범주로만 구성되지 않고, 단지 서너 개의 층만으로 구성되지도 않는다. 여기서는 설명을 쉽게 하기 위해 도식화된 예시를 제시했다.

이를 전제로 설명해보면 다음과 같다.

'민간', 즉 사기업의 영리사업 차원 레이어에서 모빌리티 데이터는 일반적인 웹 기반 비즈니스 데이터와 유사한 성질을 보인다. 기업은 자체 보유한 서비스 운영 자원과 고객 데이터를 관리한다. 고객의 신원은 개인정보 보호법 등의 법규로 엄중히 보호된다. 기업은 고객이 사전에 고지받고 동의한 범위에 한해서만 데이터를 활용할 수 있다. 이때 필요한 데이터 보호 조치는 데이터 암호화, 웹 애플리케이션 보안, 인증 보안 등 일반적인 IT 보안 조치다.

'교통' 목적의 모빌리티 데이터로는 사회 인프라인 교통체계를 관리하며, 데이터에 기반하여 교통 효율을 최대화한다. 이를 위해 각종 교통수단 그리고 보행자까지, 모든 교통 요소의 위치를 파악하고 그 이동을 감지한다.

그러나 시민의 사생활을 감시해서는 안 된다. 이동 내내 어떤 차인지를 알아야 하지만 누구의 차인지는 알아선 안 된다. 따라서 본 층 데이터의 가장 중요한 보호 조치는 가명화(Pseudonymization)다. 즉, 데이터가 특정인과 대응되지만 직접 연결되거나 추론되진 않아야 한다. 이를 위해 차량마다 임의의 유동 ID를 할당하여 추적하되 ID를 주기적으로 교체하는 등의 특수한 조치가 추가로 필요하다.

'치안' 및 안보 목적의 층은 맥락이 아주 다르다. 범죄자 등 특정인을 추적해야 하므로 신원 파악이 필요하다. 이는 헌법이 규정하는 선에서 제한적으로 가능하다. 당연한 기본권인 체포·구속·압수·수색으로부터 신체의 자

유, 사생활의 비밀과 자유, 개인정보 자기결정권 등과 심각하게 충돌하므로 엄격한 영장주의 법규 적용이 필요한 일이다. 최근 모바일 기기 추적 논란 때와 마찬가지로, 치안·안보 차원의 데이터 활용은 사회적 합의와 법제화가 무엇보다 중요하다.

3. 모빌리티 데이터 비즈니스

가. 민간 영역-모빌리티 데이터를 통한 문제 해결

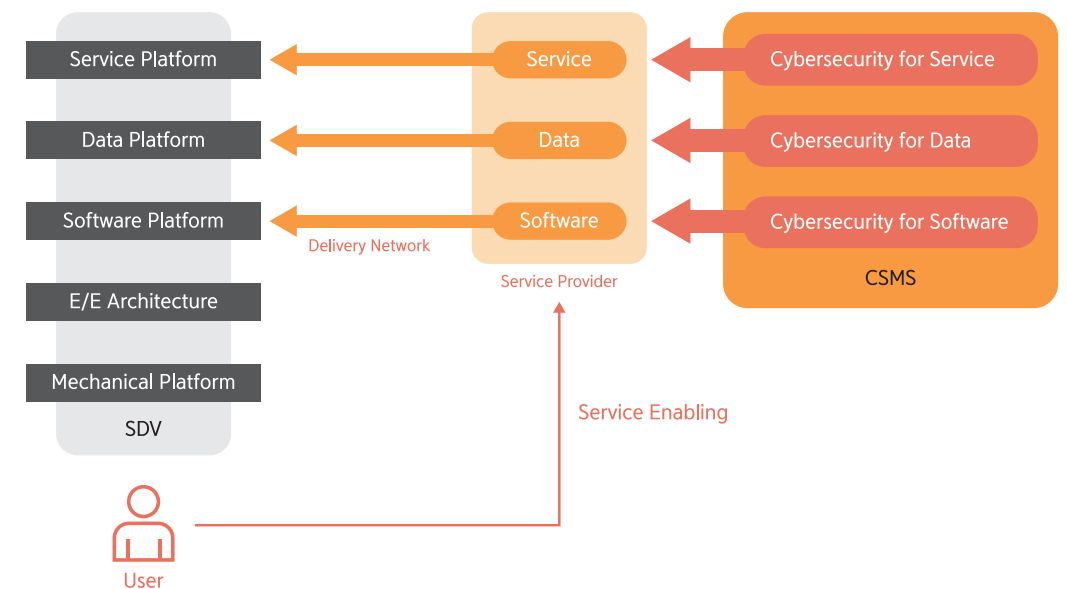
모빌리티 데이터를 활용한 민간 비즈니스 중, 앞서 분류한 차량 데이터, 이동 데이터 활용의 대표적 경우를 살펴보자. SDV 데이터 플랫폼, 그리고 MaaS(Mobility as a Service) 비즈니스다.

1) SDV 데이터 플랫폼

자동차 시장 판도가 SDV 중심으로 재편되고 있다. 일례로, '테슬라(Tesla)'는 자동차 하드웨어와 별개로 'FSD(Full Self Driving)' 등 소프트웨어를 별도 판매한다. 머지않아 소프트웨어의 영업이익이 하드웨어 영업이익을 추월할 것 이란 전망도 있다.

물론 가격이 높은 하드웨어 매출이 소프트웨어 매출보다 훨씬 더 크다. 하지만 둘의 원가 차이가 현격하고 양산 단계로 갈수록 차이가 더욱 벌어지니, 이익 즉 매출에서 원가를 제한 값을 따지면 소프트웨어 판매가 유리하다. 이것이 SDV 사업성의 핵심이다.

[그림 4-3-3] SDV 데이터 플랫폼 보안



* 출처: 아우토크립트 제작

기존 자동차에서 '데이터 플랫폼'이라 할 만한 것은 전장 아키텍처(E/E Architecture)와 이를 제어할 간단한 소프트웨어 정도였다. 전장 아키텍처(E/E Architecture)란 차량 내 CAN(Controller Area Network)을 통해 단순한 형식의 전기 신호가 오가는 것을 뜻하는데, 이것을 제어하는 소프트웨어까지가 데이터 플랫폼이었던 셈이다.

그런데 SDV에서는 E/E 중 상부의 '소프트웨어 플랫폼'이 본격화되고 '데이터 플랫폼'과 함께 '서비스 플랫폼'이 추가된다. 제조사의 계획에 따르면 자동차가 "상시적 기능 향상 제품"이 되기 때문이고, 그것이 SDV 개념의 핵심이다. 마치 공장 출시 시점에 기능이 고정되었던 '피쳐(Feature, 기능) 폰'에서 사용자가 기능을 수시로 추가하는 스마트폰으로 변화한 것처럼, 자동차 또한 소프트웨어 설치 및 업데이트를 통해 상시로 기능을 추가하는 방식으로 진화하고 있다.

SDV의 서비스 플랫폼은 차량, 이동 등 전반적 모빌리티 데이터에 기반한 비즈니스의 새로운 장이다. '자동차 앱 스토어'라 부를 만한 비즈니스 플랫폼을 통해 하드웨어 제조사, 소프트웨어 개발자, 그리고 소비자가 거래하는 일종의 생태계를 이룬다. 현재 스마트폰의 앱 마켓과 유사한 시장 구조를 그려도 크게 어긋나지 않을 것이다.

2) MaaS 비즈니스

MaaS는 모빌리티 데이터 비즈니스의 대표적 예다. MaaS란 "Mobility as a Service" 의미 그대로 '서비스로서의 이동'을 뜻하며, 구체적으로는 어떤 권역에서 운행되는 모든 대중 교통수단과 개인이 소유한 교통수단을 통합하여 이동 편의성과 효율성을 제고하려는 서비스다. 다시 말해, 파편화된 모든 교통수단을 통합하여 MaaS 플랫폼상에 얹음으로써 서비스 이용자에게 최대한 빠르고 저렴하며 효율적인 이동을 제공하는 것이 MaaS의 사업 모형이다.

MaaS는 일반 시민 입장에서 가장 자주 보게 될 모빌리티 비즈니스의 대표 격으로, 우리 일상에 매우 밀접하게 연계될 것이다. 하지만 크게 주의할 점이 두 가지 있다.

첫째, 보안이다. MaaS 사업에서는 하드웨어에서 소프트웨어까지 모두 노출된 광역 환경에서 일어나는 민감 데이터 공유 과정 전체에 걸쳐 공공 · 민간 사업자가 긴밀히 참여한다. 그 때문에 각종 보안 위험이 극대화된다. 서비스 연결만으로도 주변 여러 물리적 기기와 지속적으로 실시간 데이터 교환을 하기 때문에 수많은 단말과 인터페이스 장치에 모두 다 철저한 보안이 적용되어야 한다. 그중에 단 하나라도 감염되거나 인위적 조작이 가해진 기기가 있다면, 이는 단지 금전적 피해 정도로 그치지 않는다. 교통사고 피해는 물론, 나아가 테러에까지 악용될 수 있다.

둘째, 시장 공정성과 경제적 지속 가능성이다. 플랫폼 비즈니스가 대개 그러하듯 자칫 대기업의 독과점 현상이 심화될 우려가 있다. 대기업이 막강한 자본력과 장악력을 무기로 독과점 플랫폼 사업자가 되고 나면, 해당 부문에서 신성장 기회는 사라지고 모두가 한 기업에 종속되고 만다. 이는 단지 올바른 시장 공정성 문제일 뿐만 아니라, IT 서비스 독과점 기업의 시스템에 문제가 발생하면 가히 국가 재앙이라 할 만한 상황이 벌어지기도 하듯, 기술적 그리고 사회 · 국가적 인프라(Infrastructure) 관리 문제이기도 하다. 따라서 다수 기업 간의 상생 협력을 통해 건강한 산업 생태계를 이루려는 노력이 필요하다.

나. 공공 영역-모빌리티 데이터를 통한 문제 해결

모빌리티 데이터의 궁극적 목표는 문제 해결이다. 초기 도입 단계인 만큼 막연한 공포가 크지만 자율주행 기술의 최대 목표가 교통사고 절감이듯, 모빌리티 데이터 활용의 목표 또한 각종 사회 문제 해결이다.

그중에서도 도시 문제 해결이다. 인구 밀집, 교통 체증, 환경 오염 등 도시에서 발생하는 온갖 문제는 도시화(Urbanization) 추세에 따라 점점 더 심각해지고 있다. UN의 분석에 따르면 2050년에 전 세계 인구의 약 68%가 대도시에 거주할 것이다.¹⁾ 이토록 과도한 인구 밀집 탓에 지금도 심각한 도시 문제는 더욱 악화되고, 도시 문제가 곧 전체 사회 그리고 국가 문제가 될 것이다.

'스마트 시티(Smart City)'는 도시 문제를 기술로 해결하려는 포괄적 대책이다. 도시 문제 중에서 가장 큰 문제는 교통 문제이고, 자율주행, 친환경 자동차, 공유 모빌리티 등의 '스마트 모빌리티(Smart Mobility)' 기술이 문제 해결의 실마리를 제공한다. 즉 스마트 모빌리티, 그리고 거기서 발생하는 모빌리티 데이터는 스마트 시티의 핵심 요소이다.

도시화에 따른 교통 소외 문제 또한 모빌리티 서비스로 해결할 수 있다. 예를 들어 놓여온 교통 벽지 문제를 해소하기 위한 수요응답형 교통 체계(DRT, Demand Responsive Transit), 교통 약자 이동권 보장을 위한 배리어-프리(Barrier-Free) 서비스 등이 있다. 이를 위해 좀 더 안전하고 편리한 모빌리티를 광범위하게 보급 · 확산하려는 사회적 노력이 필요하다. 또한 이동성 취약점을 해소하는 전문 서비스 및 플랫폼을 개발하고 그 성능과 효율을 고도화하도록 해야 한다.

지속 가능한 경제를 위한 모빌리티, 그리고 우리 모두를 위한 모빌리티는 모빌리티 서비스가 나아가야 할 지향점이다. 그리고 이는 모빌리티 데이터에 기반한다.

1) "68% of the world population projected to live in urban areas by 2050, says UN", UN, 2023년 8월 25일 접속, <https://www.un.org/development/desa/en/news/population/2018-revision-of-world-urbanization-prospects.html#:~:text=News-,68%25%20of%20the%20world%20population%20projected%20to%20live%20in,areas%20by%202050%2C%20says%20UN&text=Today%2C%2055%25%20of%20the%20world's,increase%20to%2068%25%20by%202050>

제4장 제조분야 데이터 활용 현황

나혁준 연구소장 (주)비스텔리전스

수십 년 동안 기업들은 사물인터넷(IoT)을 통해 데이터를 수집하고, 응용 시스템을 구축하며, 경우에 따라 자동 제어 시스템 영역까지 공장을 '디지털화' 했다. 이로 인해 생산 현장의 많은 데이터를 컴퓨터나 모바일 장치에서 쉽게 볼 수 있지만, 한계도 있다. AI를 적용하는 분석 및 의사 결정 지원 솔루션의 최신 기술 측면에서 아직 대다수의 제조 회사는 다른 산업에 비해 상대적으로 디지털화 수준이 낮은 실정이다.

하지만 이제 제조업에서도 AI 적용 범위가 서서히 증가하고, 그 효과가 보인다. 바야흐로 "Manufacturing AI"의 시대가 오고 있다. 이번 장에서는 IoT를 통해 실시간 수집되는 방대한 제조 데이터에 AI를 활용하는 구체적인 방법을 알아보고, 그 한계점도 짚어볼 것이다. 특히 AI를 제조 설비의 이상 감지와 고장 예측에 활용하는 사례에 대해 소개하겠다.

1. 제조업 분야의 AI 활용

AI는 제조업의 여러 영역에서 사용된다. 특히 설비 예지 보전(Predictive Maintenance), 프로세스 및 품질 최적화(Process & Quality Optimization), 프로세스 자율 및 자동화 제어(Process Autonomous & Automation Control) 영역에서 그 활용 가치가 많이 증가하고 있다.

[그림 4-4-1] 커넥티비티 기술과 빅데이터 분석의 사용 증가율

	In use today	Change over the next five years	In use in five years
Predictive maintenance	28%	+38%	66%
Big data driven process and quality optimisation	30%	+35%	65%
Process visualisation/automation	28%	+34%	62%
Connected factory	29%	+31%	60%
Integrated planning	32%	+29%	61%
Data-enabled resource optimisation	52%	+25%	77%
Digital twin of the factory	19%	+25%	44%
Digital twin of the production asset	18%	+21%	39%
Digital twin of the product	23%	+20%	43%
Autonomous intra-plant logistics	17%	+18%	35%
Flexible production methods	18%	+16%	34%
Transfer of production parameters	16%	+16%	32%
Modular production assets	29%	+7%	36%
Fully autonomous digital factory	5%	+6%	11%

* 출처: R. Geissbauer, S. Schrauf, P. Bertram, F. Cheraghi, Digital Factories 2020: Shaping the future of manufacturing, PwC, (2017.4), p.11.

가. 설비 예지 보전(Predictive Maintenance)

제조 설비는 대부분 적절한 시점에 정비(Maintenance)를 해주어야 하는데, 생산 중에 예상치 못한 문제가 발생하거나, 생산 납기를 맞추기 위하여 무리하게 설비를 운영하는 경우, 계획되지 않은 고장(Unscheduled Downtime)이 발생할 수 있다. 이것은 계획 정비(Scheduled Downtime)보다 통상 2~10배의 비용이 더 소요될 수 있다. 단순히 정비 시간만 더 길어지는 것뿐만 아니라, 생산 중에 발생한 생산품의 폐기(Scrap) 및 재작업(Rework) 비용 때문에 그 손실이 커진다. 그렇다고 고장이 발생하기 전에 너무 미리, 그리고 너무 자주 설비에 대한 정비를 실행한다면 목표한 생산량을 맞추지 못할 수 있다. 그뿐 아니라 부품 등 정비 자체에 드는 비용이 불필요하게 증가할 것이다. 따라서 좋은 품질의 제품을 생산할 수 있도록 설비의 상태를 최적으로 유지하면서, 설비가 고장 나기 전에 적절한 정비를 수행할 수 있다면, 비용 절감뿐만 아니라 공장 운영 측면에서도 큰 도움이 될 것이다. 이에 다양한 설비 데이터와 고장 패턴에 AI를 접목하여 설비 예지 보전(Predictive Maintenance)을 구현하기 위한 연구가 다양하게 이루어진다.

나. 프로세스 및 품질 최적화(Process & Quality Optimization)

제조 현장에서 발생하는 데이터는 생산 중인 설비로부터 수집되는 데이터도 있지만, 생산된 제품의 품질 등을 검사해서 얻어지는 데이터도 있다. 이러한 생산 데이터와 품질 데이터는 매우 밀접한 관계가 있는데, 생산하는 동안 설비의 상태는 제품 품질에도 크게 영향을 미치기 때문이다.

이러한 생산 데이터는 다양한 센서 등을 통해 실시간 수집된다. 품질 검사는 보통 생산 공정의 제일 마지막 단계에 이루어지지만, 반도체 등 생산 스텝이 많고 복잡한 제조 공정에서는 중간중간의 중요 스텝마다 계측 테스트를 실행하기도 한다. 최종적으로 판매를 위한 완제품의 품질이 가장 중요하며, 이것은 수율(전체 제품 대비 정상 제품의 비율)로 관리한다. 이때 생산 데이터의 변화 내용을 분석하고 활용하여 품질과 수율을 예측할 수 있다면, 생산 물량 관리와 공정 운영 측면에서 큰 도움이 된다. 수율 및 품질 예측에 사용된 데이터의 경우 생산하는 설비 자체의 성능과 생산성 변화를 예측하는 데도 활용할 수 있다.

여러 AI 기법을 활용하면 수율 및 품질 개선을 위하여 불량률의 유형을 자동으로 분류하고, 품질 및 설비의 생산성 예측을 통해 프로세스 및 품질 최적화 목표를 달성할 수 있다.

다. 프로세스 자율 및 자동화 제어(Process Autonomous & Automation Control)

AI를 활용한 설비 예지 보전(Predictive Maintenance)과 프로세스 및 품질 최적화(Process & Quality Optimization)의 결과를 실제 제조 현장에 빨리 적용하여 효과를 보려면, 시스템이 자율적(Autonomous)으로 의사 결정을 하거나 엔지니어의 의사 결정을 지원해야 한다. 또한, 그 의사 결정 내용을 자동으로(Automation) 적용할 수 있어야 한다.

이를 위해 지식 저장소(Knowledge Base)를 구축한다. 과거 데이터를 이용하여 지식 저장소의 기초를 만들고, 새로운 데이터가 수집될 때마다 데이터 간의 상관성을 분석하여 지식 저장소를 업데이트하는 것이다. 여기에 더해, 의사 결정 시점에 이를 자동으로 활용할 수 있는 체계를 만들어야 한다. 예를 들어 설비 이상 등 새로운 이벤트가

발생했을 때 데이터 관점에서 저장된 지식 저장소에 구축된 과거의 유사 이상 패턴을 찾아내고, 엔지니어가 조치해야 할 내용까지 자동으로 생성해내는 것이다. 그런 식으로 엔지니어의 올바른 신속한 의사 결정을 지원하는 것이다. 이 과정에 다양한 AI 기법이 활용된다.

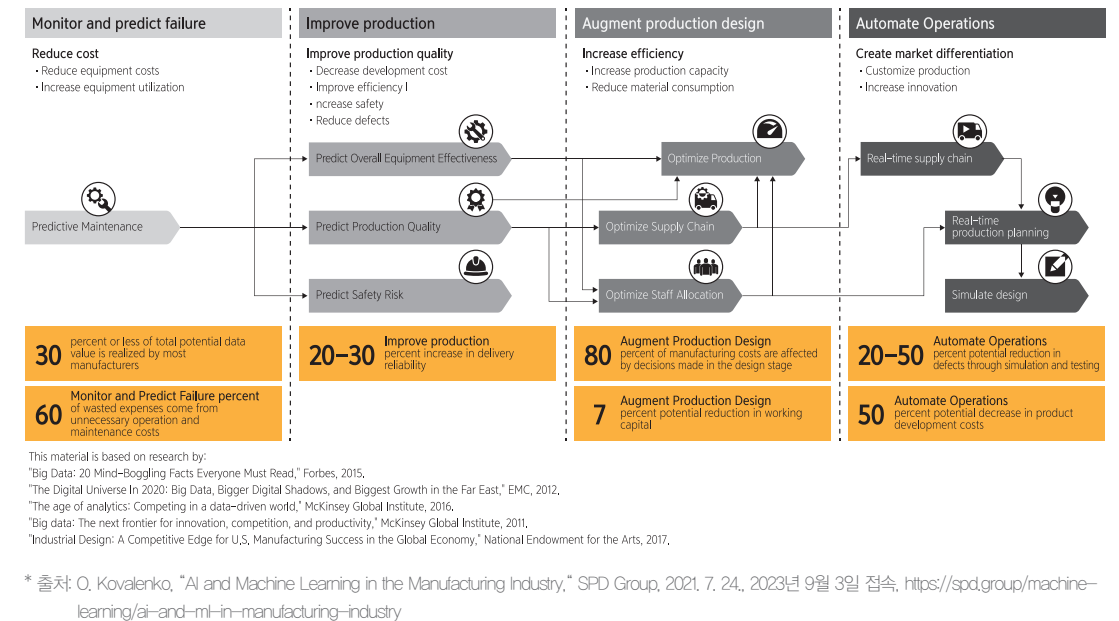
물론 축적된 과거 데이터에 근거하여, 때로는 이러한 이상 패턴에 별다른 조치가 필요 없다고 판별할 수도 있다. 이를테면 가상 알람(False Alarm)으로 취급한다. 반대로 조치가 필요한 경우, 정보기술(IT, Information Technology) 뿐만 아니라 운영 기술(OT, Operation Technology) 영역까지 아우르는 자동화 기술이 필요할 수 있다.

2. 제조업 분야의 AI 활용 가치

제조업에서 AI의 활용 가치는 이미 여러 영역에서 증명되었다. 특히 설비 고장 예측을 통한 가용성 확보, 수율 및 품질 향상, 설비 생산성 향상, 공정 운영성 개선, 안전사고 예방 등이 주요 영역이다.

- 제조 현장에서 발생하는 데이터는 AI를 통하여 그 활용도가 계속 높아지고 있으며, 대다수의 제조업체에서 수집되는 전체 데이터의 30%가 AI를 통하여 추가로 활용된다고 한다.
- 낭비되는 비용의 60%는 불필요한 운영 및 유지보수 비용에서 발생하는데, 설비 예지 보전(Predictive Maintenance)를 통하여 고장(Unscheduled Downtime)을 줄이고, 불필요한 운영 및 유지 보수 비용을 절감할 수 있다.
- 제품 품질 예측과 설비 생산성 예측을 통하여 생산 납기 준수율이 20~30% 증가하고, 예측된 값을 활용하여 공정 운영성이 개선된다.
- 제조 비용의 80%는 Design 단계에서 결정되는데, 여러 데이터와 AI 기법을 이용하여 다양한 Simulation을 실행하고, 이는 Design 개선 및 Production 최적화로 이어질 수 있다.
- Design 개선으로 운영 자본의 7%를 절약할 수 있다.
- 시뮬레이션 및 테스트를 통해 불량률의 20~50%가 감소할 수 있다.
- 자동화된 Operation으로 제품 개발 비용의 50%를 절감할 수 있다.

[그림 4-4-2] 제조업 분야에 산업화된 AI 적용



3. 제조업에서 AI 활용의 한계 및 극복 방안

앞에서 기술한 것과 같이 제조업에서 이미 AI가 많이 활용되고 있지만, 항상 기대하는 결과가 도출되는 것은 아니다. 제조업에서 AI 활용의 한계점이 무엇인지 알아보겠다.

가. 설비 예지 보전(Predictive Maintenance)

설비 고장을 예측하는 모델을 만들려면, 충분한 고장 사례를 학습해서 모델을 만들 수 있어야 하는데, 실제 현장에서는 고장 데이터가 매우 부족한 상황이다. 이제 막 데이터를 수집하기 시작한 공장들도 있고, 아직 활용하기 힘든 형태로 데이터가 저장된 곳도 있다. 주요 고장이나 정비 기록을 아직 메뉴얼로 관리하는 곳도 많다. 무엇보다 많은 설비들이 AI 모델을 만들 만큼 충분히 고장 나지 않고 있는데, 길게는 정비 주기가 3년 이상 되는 설비들도 산업 현장에 존재했다. 이러한 상황이라 적정 정비 시점을 예측하는 것이 거의 불가능하다. 결국 이상 감지(Anomaly Detection) 등 다른 목적으로 AI를 활용하는 것이 바람직하다. 만약 동일한 모델의 설비가 여러 대 존재한다면 데이터를 통합하여 단일 모델 생성에 사용할 수 있다. 다만, 이 역시 설비 간 편차가 크거나 특성이 많이 다른지 살펴야 한다. 무작정 동일한 모델이라고 데이터를 통합한다면, 좋은 모델 생성이 어렵다. 고장 데이터가 부족한 상태에서 비지도 학습(Unsupervised Learning) 기법을 활용하여, 이상 감지(Anomaly Detection) 및 예지 보전(Predictive Maintenance) 모델을 생성하는 방법은 뒤에서 자세히 다루겠다.

나. 프로세스 및 품질 최적화(Process & Quality Optimization)

제조 현장에서는 생산 설비의 고장 발생 데이터가 AI 모델을 만들기에 충분하지 않을 뿐만 아니라, 품질 이상이 나 저수율 데이터도 충분하지 않은 경우가 많다. 반도체 후공정의 경우 수율은 99%를 훨씬 웃돈다.

궁극적으로 불량률을 줄여 품질을 높이려면 불량 유형별 원인과 생산 데이터와의 상관성을 분석해야 한다. 이를 바탕으로 생산 조건을 조정하여야 하는데, 만약 불량 유형이 매우 다양하며 그 모수 자체가 많지 않다면, 좋은 분석 결과를 도출하기 어렵다. 특히 불량을 예측하는 작업은 더욱 어려워진다.

이때 비지도 학습(Unsupervised Learning) 기법을 활용하여, 모델링 없이 불량 유형을 그 특징별로 분류하여 원인을 분석할 수 있다. 이 방법론은 꽤 효과적인데, 이를 통해 일부 불량 데이터들을 도메인 지식(Domain Knowledge)에 근거하여 복제하거나 증식하여 모델 생성에 활용하기도 한다.

다. 프로세스 자율 및 자동화 제어(Process Autonomous & Automation Control)

엔지니어는 여전히 자신의 경험과 직관에 의지해 판단하려고 하는데, 이런 접근 방식을 취할 때 때로는 AI와는 다른 선택을 할 수 있다. 이때 과연 데이터 및 AI 모델 기반의 결과를 얼마나 잘 적용할 수 있을지 궁금해진다.

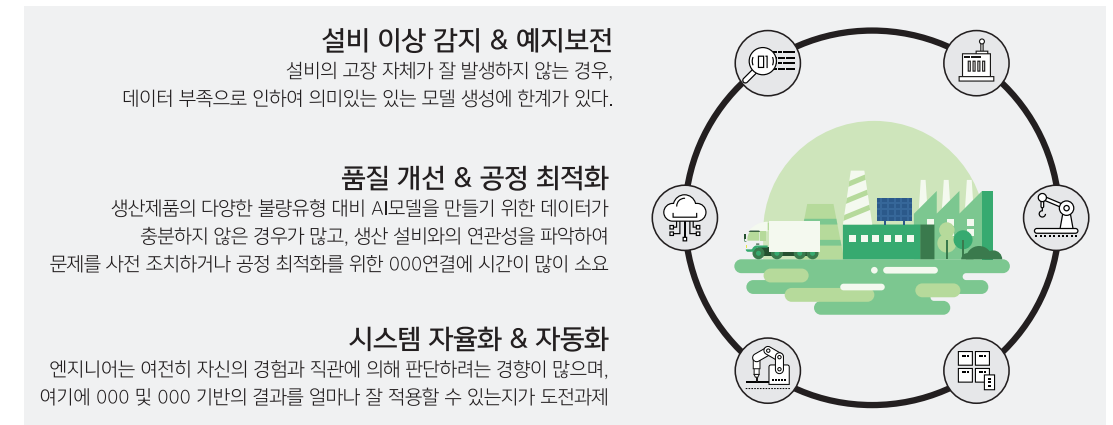
이것은 나름대로 도전 과제인데, 이에 적합한 시스템을 만들려면, 데이터 및 AI 기반의 모델로부터 얻은 결과를 활용해 엔지니어가 수행할 작업 지시로 전환하는 작업까지를 자동으로 할 수 있어야 한다. 이를 위해 지식 저장소 구축 초기에 엔지니어의 도메인 지식을 반영할 수 있는 프로세스 구축이 중요하다. 지속적으로 지식을 축적하고 의미 있는 연관성을 계속 업데이트해야 한다.

더 나아가 데이터 및 AI 모델 관점에서 이상 발생(Anomaly Detection) 시 이것이 진성 알람인지 가성 알람인지 구별해줘야 하며, 진성 알람의 경우에는 이에 대한 적절한 조치 내용까지 지식 저장소에 구축되어야 한다.

이처럼 궁극적으로는 시스템이 자율적(Autonomous)인 의사 결정을 하거나, 의사 결정 지원을 위해 과거 이력을 바탕으로 지식 저장소(Knowledge Base)를 구축하는 것이 중요해진다. 이때 비정형 데이터를 포함하여 다양한 형태로 산재된 엔지니어의 지식, 노하우, 사고 처리 이력 등을 정형화해야 한다. 또 이에 대한 의미 있는 연관성을 부여하는 작업이 중요해지는데, 이 과정들이 쉽지 않은 작업이다.

그런 과정을 잘 거친다면 지식 저장소가 구축된 후에 유사 문제 발생 시 이를 잘 활용할 수 있지만, 여기서 핵심은 '얼마나 풍성하고 정확한 지식 저장소를 구축할 수 있는가' 하는 것이다. 또 변화하는 공정 상황에 맞게 지속적으로 관리할 수 있는지도 중요하다.

[그림 4-4-3] 제조업 분야의 AI 활용 한계



* 출처 : 비스텔리전스 제작

4. 제조업에서 AI 적용 절차

여러 기업은 IoT 장치들로부터 데이터를 수집하기 시작했고, 이를 활용한 여러 시스템을 도입하고 있지만, 아직 많은 회사에선 숙련된 엔지니어들의 경험과 판단에 의존한다. 이처럼 전문적인 경험에 대한 의존도가 높기 때문에 퇴직 시 고도로 숙련된 엔지니어로 교체하기가 매우 어렵다. 이들의 이탈은 수익에도 큰 영향을 미칠 수 있으므로 AI를 통한 지식 보존, 개선 및 표준화 능력은 더욱 중요해진다.

따라서 숙련된 엔지니어들의 경험과 지식을 데이터 및 AI 모델 기반의 분석 결과에 빨리 접목하여 통합하고, 연결된 지식 체계를 구축하는 것이 매우 중요하다. 이러한 환경이 구축된다면 궁극적으로 AI 기반 시스템이 숙련된 엔지니어를 대체하고, 예측 가능하고 일관된 결과를 안정적으로 제공할 것이다.

아래 예제는 한 시멘트 회사가 AI를 적용하고, 이를 통해 얻은 이점을 설명하고 있다.

적용 절차

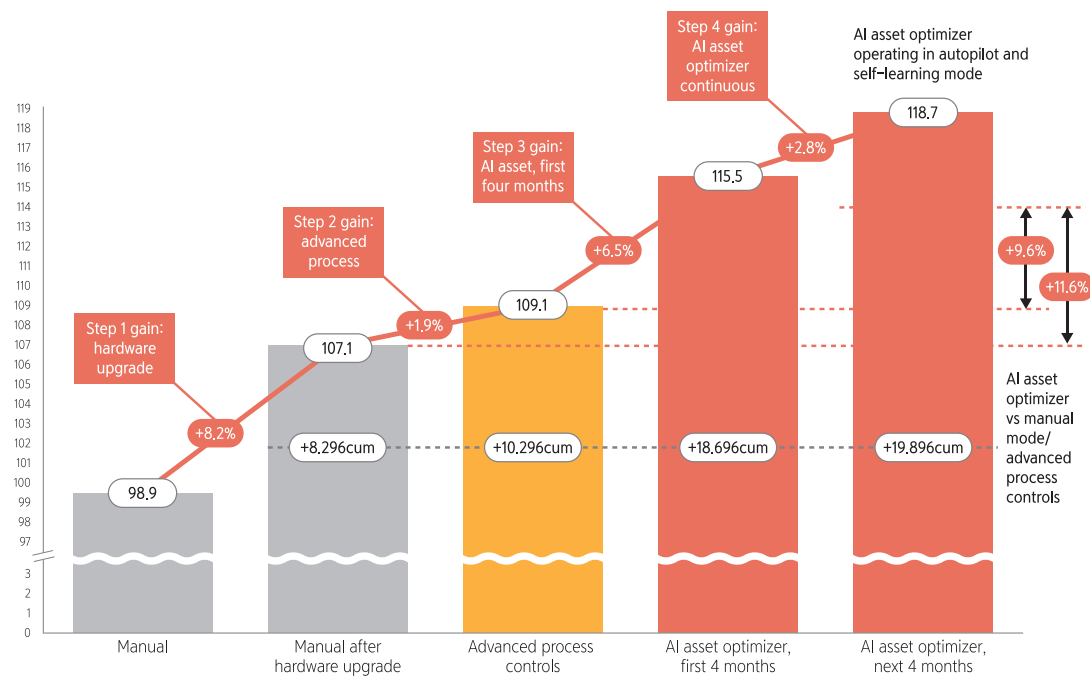
- 1) 수백 개의 프로세스 변수에서 수백만 건의 데이터 캡처
- 2) 고급 분석 도구 및 기술을 사용하여 데이터 준비 및 분석
- 3) 자동화 프로세스 흐름에 따라 데이터를 연결
- 4) 오프라인 분석을 통하여 자산 최적화(Asset Optimizing)를 위한 AI 모델을 생성하고 테스트
- 5) 생성된 모델의 온라인 적용. 데이터 인터페이스를 통해 자동화 및 제어 시스템에 연결

6) 운영자의 개입 없이 Asset Optimizer가 자율적으로 작동

적용 효과

- 1) Asset Optimizer 설치를 통해 몇 주 만에 수익이 크게 향상되었다. 설치 후 4개월 및 8개월의 성능 검토에 따르면 AI Asset Optimizer는 속도와 비용 측면에서 모두 상당한 비율로 기존 시스템보다 우수한 성능을 보였다. (약 11.6% 성능 향상을 달성하였음)
- 2) AI를 활성화하면 제조 설비에서 설비의 성능과 시간당 수익이 향상되는 동시에 정확하고 안전한 방식으로 프로세스 설정값을 자동으로 조절할 수 있다.

[그림 4-4-4] 제조업 분야의 AI 적용 효과에 대한 사례



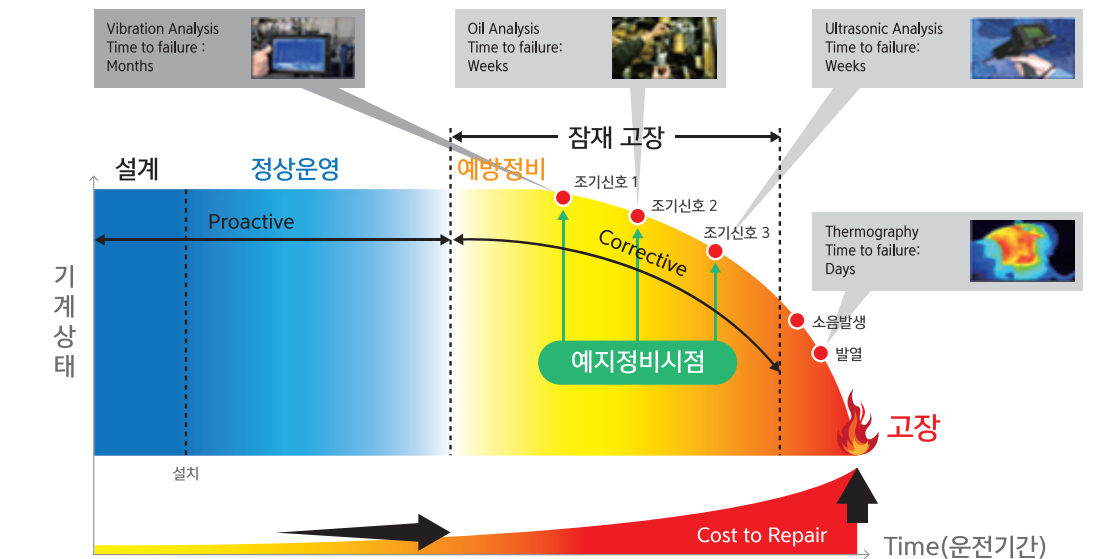
Source: McKinsey Report – “AI in production: A game changer for manufacturers with heavy assets”

* 출처 : E. Charalambous, R. Feldmann, G. Richter, C. Schmitz, “AI in production: A game changer for manufacturers with heavy assets”, QuantumBlack AI by McKinsey, 2019. 3. 7., 2023년 9월 3일 접속.
<https://www.mckinsey.com/capabilities/quantumblack/our-insights/ai-in-production-a-game-changer-for-manufacturers-with-heavy-assets>

5. AI를 활용한 제조 설비 이상 감지 및 고장 예측

설비 이상 감지와 고장 예측에 AI를 적용하기 위해서는 설비의 건강 상태를 잘 대변할 수 있는 파라미터 (Parameter) 선택이 중요하다. 기본적으로 설비로부터 수집되는 모든 데이터를 활용하여 AI Monitoring & Analysis 모델을 생성하는데, 다양한 데이터 중에서 진동 센서로 도출한 데이터는 특히 유용하다. 즉 진동 센서로 조기에 설비 이상 여부를 감지한다면 예지 정비에 매우 효과적이다. 진동 스펙트럼 데이터들은 회전체 설비의 고장 유형 분석에 탁월한 결과를 제공해 주기 때문이다. 즉 진동의 Summary 값과 다른 파라미터를 이용하여 만든 모델로 전반적인 설비의 건강도(Health Index)를 모니터링하고, 이상 발생 시 고속 푸리에 변환(FFT, Fast Fourier Transform) 작업을 통해 변환되었을 때 그 결과가 탁월하다.

[그림 4-4-5] 제조 데이터에 AI를 활용한 설비 예지 정비



* 출처: IoT Analytics Research & BISTelligence GV APM Material

설비 이상 감지와 고장 예측에 AI를 적용하는 것은 아래와 같이 크게 6가지 세부 항목으로 구분할 수 있다.

1) 파라미터 자동 선택(Auto Parameter Selection)

설비 고장에 영향을 줄 수 있는 파라미터들을 자동으로 선택한다. 값이 변하지 않는 등 모델 생성에 불필요한 파라미터들을 배제하고, 엔지니어의 경험을 반영하여 파라미터에 가중치를 부여할 수 있다. 파라미터의 원천 값(Raw Value)을 직접 사용하거나, 의미 있는 특징(Feature)을 추출하여 모델에 사용한다.

2) 자동 모델링(Auto Modeling)

선택된 파라미터와 도출된 특징을 활용하여 설비 건강도(Health Index) 모델을 생성한다. 모델 생성에는 여러

환경 설정이 필요할 수 있으며, 설비의 건강 상태를 가장 잘 대변할 최적의 설정을 찾아서 그 성능을 비교하고, 대표 모델을 생성한다.

3) 자동 모델 최적화(Auto Model Optimization)

초기 모델을 배포한 이후, 설비 또는 공정 상황의 중요 변화에 따라 적용 모델을 지속적으로 평가한다. 또 평가 결과를 반영하여 모델을 자동으로 업그레이드한다. 시간이 지나 여러 고장 케이스를 처리하는 동안 모델 스스로 최적화된다.

4) 설비 건강도 계산(Asset Health Assessment)

적용된 모델에 실시간으로 수집되는 파라미터 값들을 대입하여 설비 건강도(Health Index)를 계산한다. 건강도가 설정된 임계치(Threshold)를 넘을 경우 알람을 발생시키고, 그 문제에 대한 파라미터의 기여도 및 데이터 패턴의 변화 등을 근간으로 고장의 유형을 자동으로 분류한다.

5) 파라미터 프로파일 분석(Parameter Profile Analytics)

생성된 모델에서 의미 있는 결과가 도출되지 않은 경우, 마이크로 데이터 수준(Micro Data Level)의 파라미터 프로파일(Parameter Profile) 모델을 통해 추가적인 인사이트를 도출할 수 있다. 초 단위 또는 그 이하 데이터들의 미세한 패턴 변화가 설비 고장과 품질 문제에 영향을 주는 사례도 있기 때문이다.

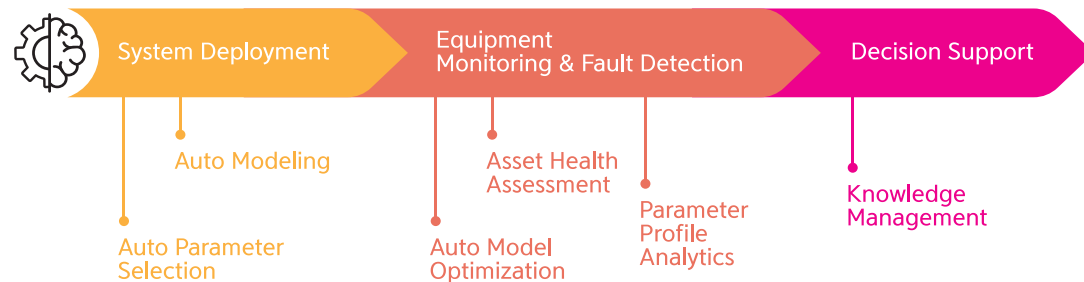
6) 지식 관리(Knowledge Management)

과거 데이터를 활용하여 지식을 신규로 구성하고, 새로운 이벤트 발생 시 유사한 사례를 탐색하여 엔지니어 조치 항목으로 활용한다. 또한 지속적으로 지식을 업데이트하고 관리한다.

[그림 4-4-6] 다양한 제조 설비 예지 보전 단계에 사용되는 AI

AI Everywhere

Artificial Intelligence improves productivity and efficiency throughout the monitoring system lifecycle



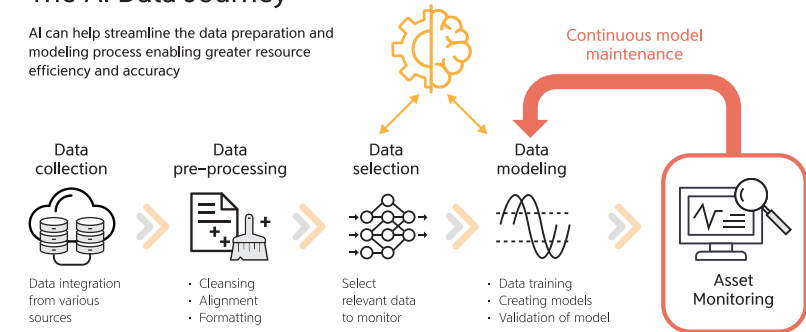
* 출처 : 비스텔리전스 제작

아래 그림은 설비 건강도(Health Index) Model 생성과 적용, 그리고 관리에 대한 전반적인 프로세스와 그 세부 내용을 도식하고 있으며, AI가 어떤 단계에서 사용되는지를 표현하고 있다. AI는 데이터 준비 및 모델링 프로세스를 간소화하여 리소스 효율성과 정확성을 높일 수 있게 하고, 사용자의 숙련도에 상관없이 언제나 일관된 값을 제시한다. 또한 스스로 학습하고 최적화함으로써 시간이 지날수록 그 정확도를 높일 수 있다.

[그림 4-4-7] AI 데이터 여정

The AI Data Journey

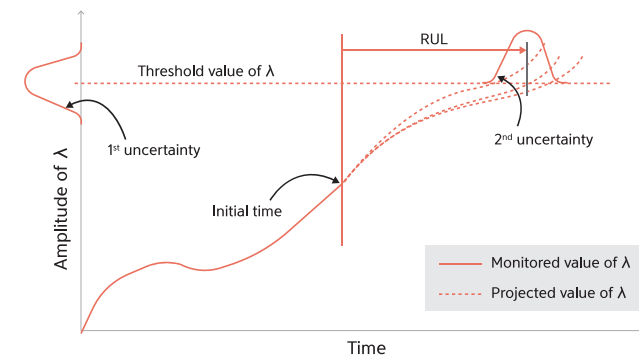
AI can help streamline the data preparation and modeling process enabling greater resource efficiency and accuracy



* 출처 : 비스텔리전스 제작

파라미터들의 정상 데이터 구간을 학습하여 얻어진 설비 건강도(Health Index) Y값은 설비의 건강도를 나타내며, 그 값이 0에 가까울수록 데이터 패턴이 모델 생성에 사용된 정상 데이터 구간과 유사하다는 의미다. 따라서 설비 상태가 건강하다고 말할 수 있다. 반대로 설비 건강도가 커지는 경우 설비 상태가 나빠진다고 말할 수 있다. 이 설비 건강도가 특정 임계치(Threshold)를 넘으면 시스템은 자동으로 알람 메시지를 생성하여, 엔지니어가 바로 문제를 식별하고 조치할 수 있도록 한다. 설비 건강도를 Y축을 기준으로 모니터링하면 설비 이상 감지(Anomaly Detection)에 활용할 수 있고, 같은 데이터를 X축을 기준으로 모니터링하면 설비의 잔존수명(RUL, Remaining Useful Life)을 예측하는 데 활용할 수 있다. 즉 현재 시점의 설비 건강도 값과 그 값이 변화한 이력을 파악하여, 그 추세와 패턴을 바탕으로 설비 고장에 관한 임계치(Threshold)에 언제 도달할지 예측할 수 있다.

[그림 4-4-8] AI를 활용한 설비 건강도 계산 및 잔존 수명 예측



* 출처 : 비스텔리전스 제작

제5장 농업분야 데이터 활용 현황

홍승길 농업연구관 농촌진흥청 디지털농업추진단

우리나라 농업은 기후변화와 이상기상으로 빈번한 자연재해 발생, 농업인구 감소와 고령화·여성화로 농업인력이 감소하면서 뚜렷해진 농촌 소멸화 위기, 쌀을 제외한 식량자급률 저조, 국제 시장 혼란으로 곡물 및 유가 인상 등 여러 문제에 직면해 있다.

이러한 문제를 해결하기 위해 국가 연구기관뿐만 아니라 기업에서도 농업 분야의 연구·서비스 개발 및 비즈니스 창출에 노력을 기울이고 있다. 즉 농업 생산 과정을 디지털화하여 빅데이터를 활용하는 생산 방식을 심도 있게 연구하고, 이를 바탕으로 농업 분야에서 지능화·자동화 수준에 이르도록 하려는 노력이다. 이는 신정부에서 세운 ‘농업 생산의 30%를 스마트농업으로 전환한다’라는 목표를 지원하기 위한 것이다.

2023년 미국가전박람회(CES)에서는 CES 사상 최초로 농기계 존디어의 회사 대표가 기조발표를 했다. 전자제품을 만드는 회사가 아니라, 농기계 회사였다. 이처럼 농기계 회사 존디어가 집중 조명을 받았던 것을 보아도 산업 분야의 벽이 허물어졌다는 것을 엿볼 수 있다. 또한, 가전의 개념이 농업 생산 과정으로도 확산한 상황을 짐작해 볼 수 있는 대목이다.

이제 인공지능 기술의 발달과 더불어, 농생명 분야는 급속도로 발전할 영역으로 평가된다. 농생명 분야에서도 자율주행, 로봇 등 최신 기술을 이용하기 위해 데이터 수집과 품질관리가 선행되고 있으며, 이를 활용한 가치 창출이 함께 이뤄질 것으로 전망한다.

기업 및 기관에서 제공하는 대표적인 데이터 기반 서비스를 바탕으로 국내외 농생명분야 내 데이터 비즈니스 현황을 살펴보고자 한다.

1. 국내 농업 분야의 데이터 비즈니스 및 인공지능(AI) 활용 사례

가. 인공지능 병해충 영상진단 서비스

농촌진흥청은 13개 대학과 4개 도농업기술원, 3개 민간 기업·연구소와 함께 ‘농작물 병해충 인공지능 영상진단 처방 앱 서비스’를 개발하였다. 이는 작물 생산량에 영향을 미치는 병해충을 스마트폰으로 실시간 현장 진단하고 처방할 수 있도록 해주는 서비스다. 병해충이나 진단하기 어려운 식물바이러스를 농업 현장에서 휴대전화로 촬영하면, 이미지 데이터를 바탕으로 인공지능을 활용해 즉시 진단하고 방제법 등을 제공하는 휴대전화 앱 서비스다.

스마트농업이 확산하면서 새로이 농촌에 유입된 귀농자와 청년 농업인들의 가장 큰 어려움이 병해충으로 인

한 농작물 피해인데, 이 인공지능 병해충 영상진단 서비스의 인식정확도는 작물 및 병해충에 따라 다르지만 평균 96.6%로 사람의 인지정확도 95.3%보다 더 정확하게 진단할 수 있다. 특히, 고추에 주로 발생하는 11개 병해충에 대한 인식정확도가 평균 99.45%로 상당히 높게 나타났다.

2024년까지 156개 전국 시·군농업기술센터에서 실증을 진행하고 있다.¹⁾ 이를 통해 8개 발작물, 15개 채소, 8개 과수 등 31개 작물에서 발생할 수 있는 병해충 및 바이러스 344종(병 136종, 해충 183종, 바이러스 25종)을 조기에 진단하는 시스템을 구축할 것으로 보인다.

[그림 4-5-1] (상) 병정부위 자동 인식 모델(Mask R-CNN 모델)

(하) 서비스 활용 병해충 진단과정



* 공동참여기관인 세종대학교 산학협력단 특허

* 출처: (상) DERBEL MohamedAziz, "Mask R-CNN for Instance Segmentation Using Pytorch", Analytics Vidhya, 2023. 2. 22., 2023년 9월 21일 접속, <https://www.analyticsvidhya.com/blog/2023/02/mask-r-cnn-for-instance-segmentation-using-pytorch/>
(상2) J. A. Jimenez-Berni, S. Foucher, J. Théau, P. St-Charles, "Convolutional Neural Networks for the Automatic Identification of Plant Diseases", frontiers, 2019. 7. 23., 2023년 9월 21일 접속, <https://www.frontiersin.org/articles/10.3389/fpls.2019.00941/full>
(하) 농촌진흥청, 내부 자료.

1) 이은용, "인공지능 병해충 영상진단 서비스' 시연회 성료", 농축유통신문, 2023년 8월 21일, <http://www.amnews.co.kr/news/articleView.html?idxno=54807>

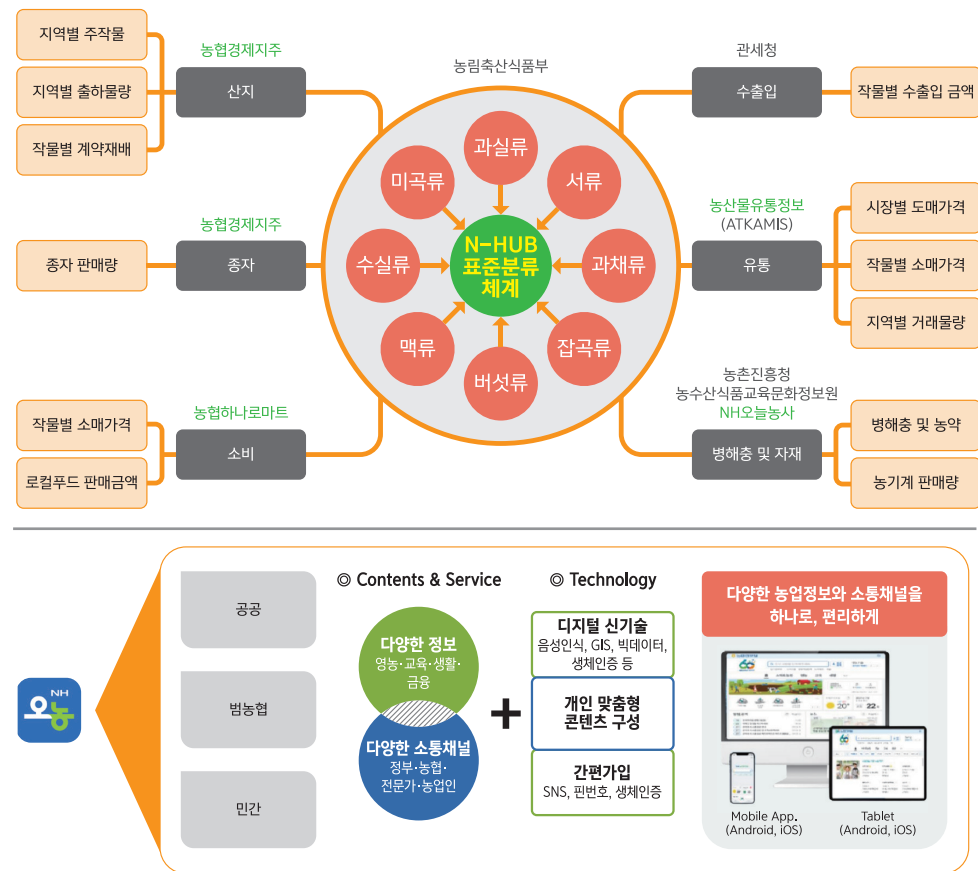
나. 빅데이터 플랫폼

농협에서 제공하는 N-Hub(NH Bigdata+Hub) 플랫폼에서는 435개 작물 데이터의 산지, 소비, 종자, 유통, 가공 등 농업 가치사슬에 대해 표준 코드 체계를 도입해 데이터 간의 연계를 강화했다. 이를 통해 농업인들과 관련 소비자들에게 농산물 가격 예측, 스마트팜 생산향상 가이드, 영농환경별 작물 추천 서비스를 제공하고 있다.

특히 농촌진흥청에서 개발한 스마트팜 생산량 향상 모델 기술을 이전받았는데, 이는 스마트팜 우수농가의 완숙토마토, 딸기, 파프리카 재배 환경과 생육 데이터를 수집하여 개발한 기술이다. 농협은 이를 토대로 스마트팜 생산량 비교 시뮬레이션 모델을 개발하였고, 다시 스마트팜 농가의 데이터기반 생산성 개선을 지원하고 있다.

또한, 종합영농플랫폼인 NH오늘농사에서는 농촌진흥청에서 제공하는 병충해·농약 정보, 재배기술 정보 등과 연계하여 작물가격정보 및 전망, 로컬푸드 판매 및 정산내역, 영농일지, 출하배차 정보 등의 서비스를 제공한다. 특히, 영농일지 서비스의 경우 공익직불금 증빙자료로 자동 연계되어 농업인의 이용 편의성을 높이고 있다.

[그림 4-5-2] (상) N-Hub 플랫폼 / (하) NH오늘농사



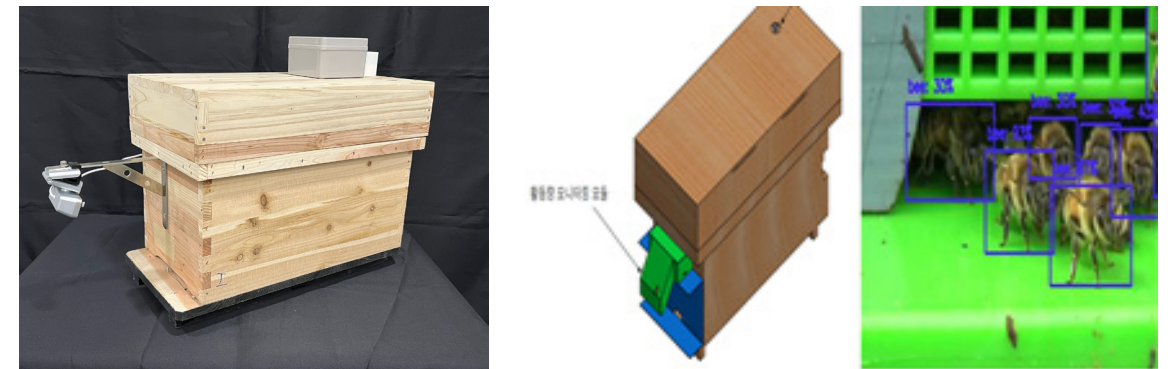
* 출처: 농협중앙회, 제공 자료, (2023. 9. 7.).

다. 스마트별통

스마트팜에서 주로 재배하는 딸기, 토마토 등 시설 과채류는 열매를 맺기 위해 화분매개용 벌을 활용해 수분을 한다. 그런데 최근 꿀벌 개체 수가 줄어드는 상황이라, 농촌진흥청 국립농업과학원에서는 화분매개용 스마트별통을 개발했다. 벌 관리와 화분 매개의 효율을 높이려는 시도라 할 수 있다.

스마트별통은 별통 내에 설치된 센서를 통해 환경정보를 수집한다. 이를 바탕으로 내부 온도센서와 연동된 환기팬(여름) 또는 열선판(겨울)이 자동으로 작동하여 별통 내부 환경을 최적으로 유지한다. 또한 내부에 설치된 카메라로 벌들이 별통에서 들어가고 나오는 수만 장의 사진을 수집한다. 이 데이터를 분석하여 0.1초 동안 벌의 움직임 변화에 따라 벌이 들어가고 나오는 행동을 계산하여, 사용자가 설정한 시간 동안에 별통을 출입하는 전체 벌의 수를 알려준다. 또한 벌의 색깔과 모양을 통해 벌을 판별할 수 있다. 이 정보들은 전용 애플리케이션으로 사용자에게 실시간 제공된다. 그 덕분에 벌에 익숙하지 않은 사용자도 별통 상태 점검 절차 및 벌 관리법, 별통 교체시기 등을 판단할 수 있다. 이러한 기술을 현장에 적용한 결과, 벌의 활동량은 1.6배, 벌무리의 수명은 68일 연장되었다. 농작물에서도 딸기의 상품과율은 6%, 토마토의 착과율은 15% 높아지는 결과를 보였다. 농촌진흥청은 신기술시범사업을 통하여 전국 8개 시군에 200여 개의 스마트별통을 시범보급 하고 있다.²⁾

[그림 4-5-3] (좌) 화분매개용 스마트별통 / (우) 실시간 화분매개벌 인식



* 출처: 농촌진흥청, 내부 자료.

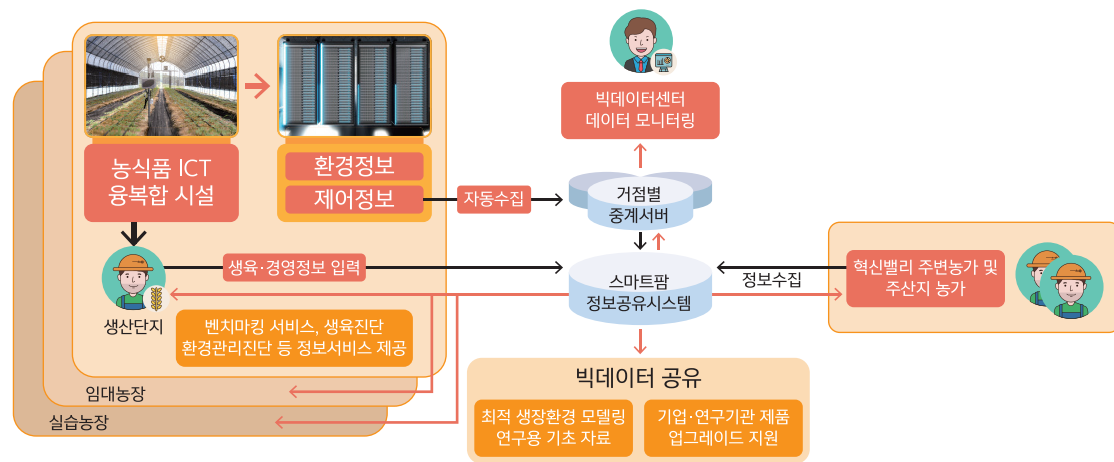
라. 스마트팜 혁신밸리 내 작물 데이터 연구

농림축산식품부는 2021~2022년에 걸쳐 스마트농업인력과 기술확산의 거점을 마련하기 위해 김제, 상주, 고흥, 밀양에 스마트팜 혁신밸리를 조성하였다. 각각의 혁신밸리는 스마트팜 관련 기업들이 다양한 스마트팜 제품 및 기술에 대한 품질 테스트, 성능 시험 등을 진행하는 실증단지, 그리고 혁신밸리 내에서 생산하는 모든 데이터를 수집·

2) 이호빈, "꿀벌이 사라지면 식량도 없다"... 농진청 '스마트별통' 개발 보급, 포인트데일리, 2023년 2월 16일, <https://www.thekpm.com/news/articleView.html?idxno=146334>

관리하는 빅데이터 센터 등으로 구성된다. 현재 혁신밸리에서는 각 지역별 특성에 맞게 딸기 · 토마토 · 파프리카 · 상추 · 가지 · 오이 · 멜론 · 만감류 등의 작물에 대한 생육 및 환경 데이터를 수집 · 가공 · 분석하고 있다. 또 이상 데이터 발생 시 즉시 대처할 수 있도록 실시간 모니터링을 시행한다. 이러한 데이터는 AI 기반의 맞춤형 모델 개발로도 이어지는 밑바탕이 되었으며, 이를 통해 생산성 최적화를 위한 재배 관리 등이 이루어질 수 있다. 농촌진흥청에서는 김제 스마트팜 혁신밸리의 토마토 농가에서 생산되는 환경정보(온도, 일사량, CO₂)와 생육정보(초장, 수확량)를 수집 · 분석하여, 토마토 재배 환경조절 의사결정을 지원하는 3D 시뮬레이터(메타팜)를 개발하였다. 3D 시뮬레이터인 메타팜은 실제 스마트팜과 동일한 메타버스 환경을 구축하여 효율성을 극대화하는 플랫폼이다.

[그림 4-5-4] 스마트팜혁신밸리 빅데이터 센터 운영체제



* 출처: "빅데이터센터 소개", 경북 상주 스마트팜 혁신밸리, 2023년 9월 21일 접속, <https://innovalley.smartfarmkorea.net/sangju/con/contents.do?cmsSeq=27>

[그림 4-5-5] (좌) 작물 생육 3D 시각화 모델 / (우) 시뮬레이터 내 재배 환경설정



* 출처: 농촌진흥청, 내부 자료.

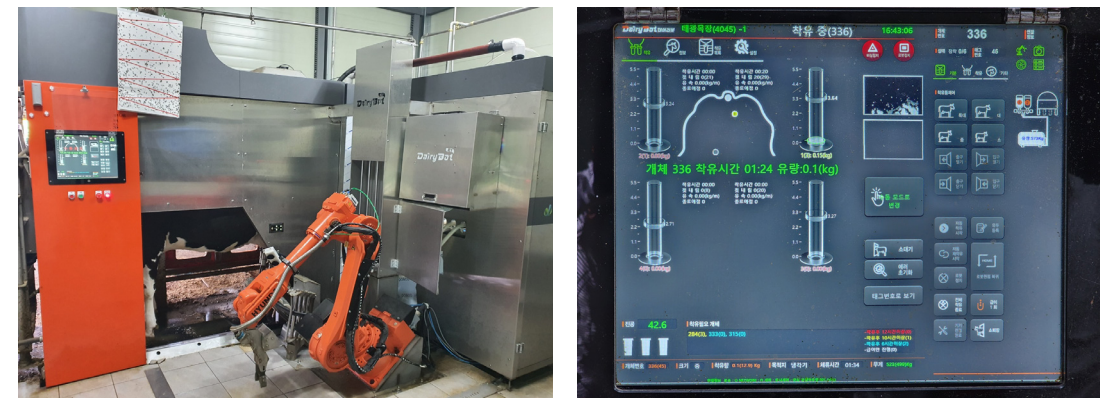
마. 로봇착유기

일반적으로 낙농가에서 투입하는 노동력은 착유 39.8%, 사료 급여 22.4%, 기타 작업 37.8%로 구분된다.³⁾ 로봇 착유기는 젖소가 착유실에 들어가 자동 급여된 사료를 먹는 사이 로봇이 착유컵을 부착해 우유를 짜는 방식으로 작동하는데, 세척, 착유컵 부착, 착유, 소독 등의 작업이 모두 자동으로 진행된다.

다만 국내에서 로봇착유기를 활용하는 농가는 5% 미만이었으며, 대부분 네덜란드, 스웨덴 등에서 고가의 로봇 착유기를 수입하여 사용하던 실정이었다. 이에 농촌진흥청 국립축산과학원과 ㈜다우이 공동으로 국산 로봇착유기를 개발하였다. 외국인 로봇착유기는 레이저와 3D카메라 조합으로 유두를 감지하는 반면, 국산 로봇착유기는 AI 기반으로 3D카메라를 사용하기 때문에 98%까지 정확하게 유두를 감지할 수 있다. 유두세척 · 착유 · 침지 등 전 과정을 한 번에 진행할 수 있는 일체형 착유컵을 사용하기 때문에 착유시간도 단축할 수 있다. 특히 착유 과정에서 우유의 유성분을 실시간으로 분석하여, 유방염이 발생한 젖소의 우유를 자동으로 분리하고 세척할 수 있도록 시스템화되어 있다.

또한 우유생산, 번식, 생체 정보 등 93개 항목에 대해서 빅데이터로 수집되고, 이는 농촌진흥청 농업빅데이터관리시스템(ABMS)에 실시간으로 연계 저장된다. 수집된 데이터의 경우 농장통합관리 프로그램으로 착유로봇을 통합 관리하는데, 원격진단으로 생산성 저하 요인, 잠재적인 질병 양상, 대사적인 문제 등을 사전에 예측하는 용도로 활용된다. 이를 통해 문제 발생 시 조기처방(예방)함으로써 개체 맞춤형 정밀 사양에 활용하는 것이 가능해진 덕분에, 고령화로 노동력 확보에 어려움을 겪는 낙농가에 도움이 될 것으로 전망된다.

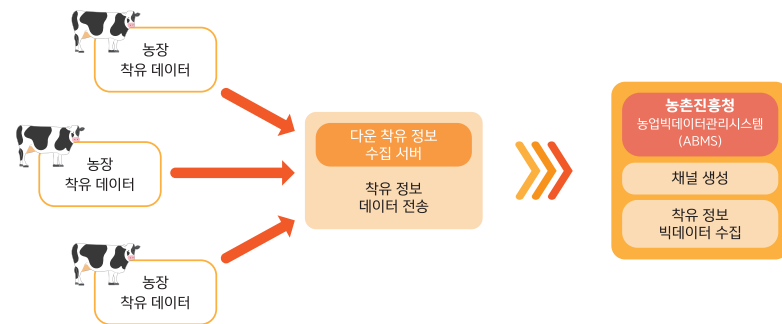
[그림 4-5-6] (좌) 국산 로봇착유기 / (우) 착유 주요 정보 대시보드 이미지 예시



* 출처: 농촌진흥청, 내부 자료.

³⁾ 통계청, 2022년 축산물생산비조사. (2022).

[그림 4-5-7] 로봇착유기 수집 빅데이터 수집 체계



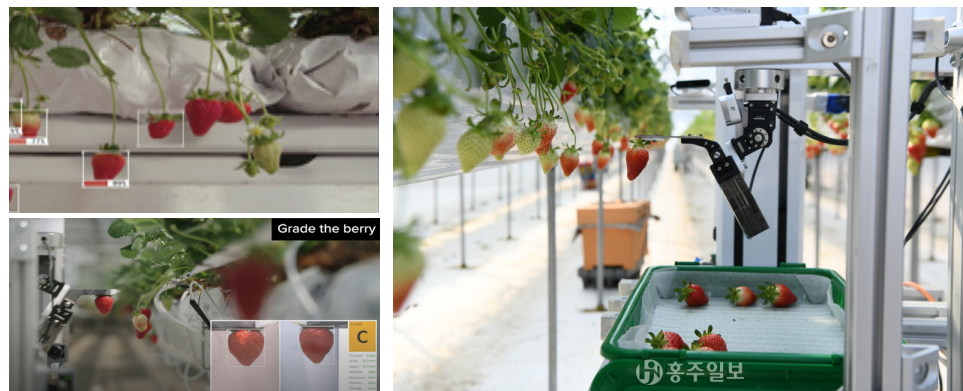
* 출처: 농촌진흥청, 내부 자료.

2. 해외 농업에서의 데이터 비즈니스 및 인공지능(AI) 활용 사례

가. 딸기 수확로봇

미국 로봇 스타트업 회사인 조르디(Zordi)는 센싱 기술을 활용해 스마트 온실 내의 환경정보 측정, 작물 인식, 숙성도 판별 등의 과정을 진행하여 딸기를 수확할 수 있는 로봇 기술을 개발하였다. 온실에 배치된 모바일 로봇과 센서를 통해 습득한 데이터를 바탕으로, AI 시스템은 딸기의 생육 단계를 파악하고 여러 해충 문제를 추적한다. 또한 수개월간의 작물 관리를 거치며 로봇에 장착된 카메라를 통해 딸기를 인식하고, AI 기계학습을 통해 딸기의 숙성도를 모니터링한다. 이런 기술 덕분에 로봇이 딸기에 직접 접촉하지 않으면서도, 딸기의 숙성도 등급을 매기며 그 등급에 따라 과일을 수확한다. 수확된 딸기 중 고품질 딸기는 자동으로 분류되며 그 외 일련의 과정을 거쳐 포장된다. 이렇듯 모니터링을 통해 관리 및 수확을 자동화함으로써 온실 내 수작업을 효율화할 수 있다.

[그림 4-5-8] (좌) 숙성도 인식 및 등급 판별 / (우) 로봇의 딸기 수확



* 출처: (좌) “딸기 수확 로봇에 대한 소개 동영상에서 발췌,” Zordi, 2023년 9월 21일 접속, <https://www.zordi.com/technology>
(우) 최효진, “로봇이 딸기 재배, 홍성의 희망 ‘홍희 딸기’ 미국 수출길 화창”, 홍주일보, 2023년 2월 15일, <https://hjn24.com/news/articleView.html?idxno=116800>

나. 정밀 파종 및 비료 살포 기술

농기계 제조회사인 존디어(John Deere)에서 개발한 이그젝트샷(Exact Shot)은 정밀 파종 및 비료 살포 기술로, ‘카메라 · 센서 등을 통해 파악한 데이터를 이용해 씨앗이 심어진 곳을 식별하는 용도로 쓰인다. 이를 통해 씨앗이 있는 곳에만 정확한 양의 비료를 살포할 수 있다. 그 결과 농경지 전체에 살포했을 때보다 비료량을 60% 이상 획기적으로 줄임으로써 이산화탄소 발생량도 줄이고 농업경영비도 절감할 수 있었다. 이는 기후변화 대응의 올바른 방향을 제시할 뿐 아니라, 다음 세대를 위한 지속 가능한 농업이 어떤 모습일지 엿볼 수 있게 해준다.

또 다른 기술인 ‘시 앤 스프레이(See & Spray)’는 자동 제초제 살포 기술로, 농기계에 부착된 카메라가 땅을 스캔한 영상 데이터를 학습하여 잡초만 감지하고 이를 통해 제초제를 살포하는 용도로 쓰인다. 투입하는 제초제를 최대 3분의 2까지 줄일 수 있어 비용 효율성을 크게 높인다. 농기계 분야에 데이터, AI, 카메라 센싱, 인공위성 정보 등 다양한 기술의 융 · 복합으로 구현된 산물이다.

[그림 4-5-9] (좌) ExactShot / (우) See&Spray



* 출처: (좌) H. Claver, “John Deere ExactShot applies starter fertilizer per seed”, FUTURE FARMING, 2023. 10. 01., 2023년 9월 21일 접속, <https://www.futurefarming.com/crop-solutions/john-deere-exactshot-planting-technology-to-reduce-starter-fertiliser-usage/>
(우) “John Deere See & Spray Ultimate: The Next Generation”, MACHINEFINDER BLOG, 2023년 9월 21일 접속, <https://blog.machinefinder.com/33587/john-deere-see-spray-ultimate-the-next-generation>

다. 자율주행 제초로봇

미국 카본 로보틱스(Carbon Robotics)의 레이저 제초기는 고해상도 카메라와 최첨단 컴퓨팅을 사용하여 잡초와 작물을 실시간으로 구별한다. 엔비디아(NVIDIA) GPU로 구동되는 레이저 제초기의 컴퓨터 비전 시스템은 고해상도 카메라 42대로 촬영한 이미지 데이터를 받아 딥러닝 기반 AI 기술로 분석해, 잡초와 작물을 구분한다. 이후 밀리미터 단위의 정확도를 보이는 30X 150W CO2 고정밀 레이저를 50밀리초(ms)마다 발사하여, 생장점을 태우는 방식으로 잡초를 제거한다. 시간당 20만 개의 잡초를 제거할 수 있으며, 정확도는 99%에 달한다.

레이저는 제초제와 달리 화학 농약을 사용하지 않기 때문에 토양 생물계를 교란하지 않는다. 따라서 작물 및 토양의 건전성을 유지해 주면서, 수확량 증가라는 긍정적인 결과를 얻을 수 있다. 또한 재배 기간 중 제초에 투입되는 노동력과 제초제 비용을 약 30% 정도 절감할 수 있을 것으로 기대된다.

[그림 4-5-10] (좌) 작물과 잡초 판별 / (우) 레이저를 활용한 잡초 제거



* 출처: (좌) "자율주행 제초로봇에 대한 소개 동영상에서 발췌", CARBON ROBOTICS, 2023년 9월 21일 접속, <https://carbonrobotics.com/laserweeder>
 (우) "Armed with lasers and blades and programmed to destroy all weeds, AI ag robots are coming to Colorado", COLORADO STATE UNIVERSITY, 2023. 6. 5., 2023년 9월 21일 접속, <https://engagement.source.colostate.edu/ai-agricultural-farm-robot-demonstration-rocky-ford-colorado/>

3. 농생명분야 데이터 비즈니스 개발의 향후 전망

농생명분야는 디지털 전환에 있어 향후 가장 잠재력이 큰 분야로 평가된다. 노동력이 절실히 필요한 농작업 분야 보니 로봇을 이용한 자동화가 중요했다. 이런 맥락에서 현재까지 로봇 자동화 기술이 개발되는 주된 영역이었다. 향후 5G 통신이 전국적으로 속도가 확보되어, 농기계와 로봇 간 IoT를 통한 연결이 가능해질 것으로 보인다. AI 지능화 기술이 농업에 본격적으로 적용되기 시작하면, 작물의 생산성 향상을 기대할 수 있다. 또한, 에너지 관리 효율화가 함께 농산업 가치 창출이 이루어질 것으로 전망된다.

올해 초부터 UAE, 사우디아라비아 등 중동 지역에 국내 기업의 스마트팜 수출이 활발히 진행되고 있다. 현재 진행되는 스마트농업 분야의 데이터 및 ICT 기자재 표준화가 구축된다면, 농생명분야 데이터 비즈니스 모델은 더 빠르게 확산할 것으로 기대된다.

제6장

에듀테크분야 데이터 활용 현황

차현승 대표이사 (주)카탐

코로나19가 발생하면서 비대면 환경에서 교육해야 함에 따라 교육의 디지털 전환을 사실상 강제적으로 진행해야 했다. 에듀테크 서비스에 의지해 빠르게 도입한 비대면 교육 시스템과 다양한 디지털 교육 서비스의 도입으로 초기에는 학업성취도 저하, 학력 양극화와 같은 부작용이 생겼다. 동시에 다양한 교수법과 에듀테크 서비스들의 발전으로 부작용을 완화하는 것을 넘어 기존의 대면 교육보다 더 나은 교육을 제공할 수 있다는 가능성을 보여주기도 했다. 대면 교육으로 복귀한 지금도 여전히 에듀테크 서비스가 활발히 도입되고 있고, ChatGPT를 중심으로 언어 모델이 교육에 활용됐을 때 그간 꿈꿔온 개별 맞춤형 학습이 한층 더 잘 이뤄질 것이라는 기대는 어느 때보다 높다. 공교육에서도 단순한 보조 도구가 아닌 주요 과목에 대한 디지털 교과서가 개발되고 있다. 이에 국내외 에듀테크 서비스들이 데이터와 인공지능을 활용해 교육의 어떤 문제를 어떻게 풀고 있는지 사례를 소개하고 이를 바탕으로 향후 어떤 기회가 있을지, 어떤 준비를 해야 할지 알아보고자 한다.

1. 교수자를 돕는 에듀테크 서비스

교육(Education)과 기술(Technology)의 합성어인 에듀테크(EduTech)는 이름 그대로 교육 서비스에 기술을 적용하는 것이다. Byju's, Chegg, Riid와 같이 에듀테크를 표방하는 국내외 스타트업은 물론이고, 국내외의 전통적인 교육 회사와 출판사도 기술을 통해 더 나은 교육을 제공하는 것을 목표로 하고 있다.

코로나 전후로 에듀테크에 대한 정부와 공교육의 태도가 크게 달라진 것으로 보인다. 그동안 보수적이고 조심스럽게 도입하던 것과 달리, 주요 과목의 교과서 자체를 AI 디지털 교과서 포맷으로 개발하고, 모든 학생에게 각자의 학습용 IT 기기를 보급하는 사업을 펼치고 있다. 또 사교육 영역의 다양한 에듀테크 서비스를 교실 안으로 적극적으로 들여오는 등 유례없이 적극적으로 바뀌었다. 이는 기술을 바탕으로 데이터를 활용했을 때, 기존 교육의 다양한 문제를 해결할 수 있을 것이라는 기대 때문이다.

교육에 데이터가 활용된 것은 결코 새로운 일이 아니다. IT 기술이 없던 시절에도 교수자들은 교육 과정에서 나타나는 학생들의 다양한 징후를 데이터 삼아 피드백하고, 더 나은 교육 내용을 구성하는 데 활용하곤 했다. 여러 다른 연구를 보더라도 교수자가 전문성을 가지고 학습자 개개인을 1대1 또는 소규모 그룹에서 오랜 시간을 들여 학생을 관찰하고 고민해 피드백하는 교육을 진행한다면, 에듀테크 기반의 디지털 교육 없이도 학업 성취도가 크게 향상한다는 것을 알 수 있다.

다만, 1대1 교육은 교수자 개개인의 역량에 영향을 크게 받을뿐더러 초, 중, 고등학교와 같은 공교육에서는 인

력 부족으로 실현하기 어렵다. 사교육에서도 높은 비용은 큰 걸림돌이 된다. 이에 시스템을 통해 교수자의 역할 일부를 자동화하거나, 보조함으로써 데이터를 균질적이고 합리적으로 활용하는 서비스가 개발되고 있다. 이는 교수자가 학습자의 이해도를 높이는 데 필요한 피드백을 줄 수 있는 시간을 늘리고, 피드백의 질을 높이는 데 유익할 것으로 예상된다.

이처럼 교육에 데이터와 기술을 활용하기 위해서는 교육 활동과 기록의 디지털화 작업을 선행해야 한다. 물론 정·오답 데이터를 바탕으로 학생별 수준에 맞는 문제를 추천해 주는 서비스는 오래전부터 다양한 곳에서 제공하고 있다. 하지만 여전히 활용률이 낮다. 추천 알고리즘의 성능이 낮아서가 아니라, 문제 풀이 등의 교육 활동과 기록이 전산화되어 있지 않기 때문이다.

이를 극복하기 위해서는 두 가지 방안이 있다. 우선 교육 활동은 아날로그 그대로 두고, 아날로그 기록을 디지털화한 이후에 기술을 적용하거나 기술을 활용해 대면 교육을 돕는 것이다. 둘째로 교육 활동과 기록 자체를 디지털화하는 방식을 통해 기존의 방식을 바꿀 수 있다.

가. 아날로그 방식으로 진행되는 대면 교육과 에듀테크 서비스

이에 해당하는 서비스 예시로는 Gradescope, Qanda, 매쓰플랫, 랑글이 있다.

Gradescope는 UC Berkeley 교수와 학생들이 창업한 회사로, 학생들의 과제, 시험, 프로젝트 등을 채점하고 관리하기 위한 온라인 플랫폼을 제공한다. Gradescope는 대학에서 대부분의 시험이 지필고사로 진행되는 것을 바꾸기보다는 현상을 수용하고 그에 대한 솔루션을 제공하는 방식을 택했다. 예를 들어 학생들의 종이 답안지를 사진 찍으면, OCR(Optical Character Recognition, 광학 문자 인식) 기술을 활용해 어떤 문제인지 인식하고, 해당 문제의 채점 기준표를 보여주는 등의 기능을 제공한다. 이를 통해 교수자는 채점할 때 시간을 절약할 수 있고, 일관된 기준에 따라 피드백할 수 있게 한다. 채점한 결과는 디지털 데이터 형태로, 교수자와 학생들 각각에 도움이 되는 통계와 함께 실시간으로 학생별 문제별로 피드백하는 데 활용된다. 채점 기준표를 수정해야 할 때는 일일이 재채점해야 하는 번거로움을 줄여주고, 학생별로 어떤 부분에서 어려움을 겪는지 제안해 주기도 한다. 또한, OCR과 NLP(Natural Language Processing, 자연어 처리) 기술을 활용해 문제별로 유사한 답변들을 그룹으로 묶어 한 번에 채점하는 기능을 제공해줄 뿐 아니라, 종이 답안지 이외의 디지털 시험에 대해서도 편리한 기능을 제공한다. 최신 AI 기술을 도입하기보다는 아날로그 답안지를 채점해야 하는 교수자의 시간을 절약하고, 더 나은 피드백을 제공하는 실질적인 가치에 집중하는 서비스로 보인다. 지금은 Turnitin이라는 회사에 인수되어 하버드, MIT, 스탠퍼드의 교수자들을 포함해 약 15만 명 이상의 교수자, 360만 명 이상의 학생들이 이를 활용하고 있다.¹⁾

Qanda는 매쓰프레소가 개발한 학생들의 수학 문제 풀이를 돕는 서비스로, 약 8천만 명의 학생들이 등록했고

약 60억 회의 질문에 대한 답변을 제공했다. Qanda 또한 대부분이 종이에 인쇄된 수학 문제를 풀고 있는 현상을 있는 그대로 받아들이고, 이에 대한 솔루션을 제공한다. 학생들이 문제를 카메라로 찍어 업로드하면, 그에 대한 풀이를 확인할 수 있는 서비스다. 여기서 더 나아가 인쇄된 문제뿐 아니라 손으로 쓴 문제들도 인식하고 이해하여 적절한 답변을 제공한다. 또한 초기에는 업로드된 문제를 일일이 풀어서 답변을 제공하였으나, 서비스가 성장하면서 수백만 데이터가 쌓이고 OCR, NLP, ML(Machine learning, 기계 학습) 기술을 활용하였다. 이에 따라 현재는 같은 문제에 대해 이전에 제공된 풀이를 제공할 뿐 아니라, 단계별 풀이를 제공하고 비슷한 문제를 추천하기도 한다. 학교별 기출 문제에 대해서도 문제 풀이는 물론, OMR 작성, 자동 채점, 해설과 같은 다양한 기능을 제공한다. 아날로그 기록을 디지털로 변환하였기에 가능한 것이다.

Gradescope와 Qanda가 문제와 답안을 인식하고 채점하는 서비스라면, 매쓰플랫은 교수자들의 문제 출제를 돕는 것에 특화된 서비스다. 다수의 EdTech 문제집이 디지털로 제작된 문제집을 일방향으로 제공했다면, 매쓰플랫은 교수자마다 활용하고자 하는 문제가 다르다는 점에 주목했다. 그리고 대부분의 문제가 종이로 출력되어 활용되는 현상을 수용하여, 교수자들의 문제 출제 과정을 돕는 것에 집중했다.

이 서비스를 활용하면 초·중·고 전 학년의 수학 과목에 대해 단원, 유형별, 시중 교재, 수능, 모의고사 등 원하는 형태의 문제 세트를 쉽게 제작할 수 있다. 출제 희망 범위를 선택하고 1부터 150까지의 문제 수를 고른 뒤, 5단계로 구분된 난이도와 주관식/객관식과 같은 문제 유형을 선택하고 클릭하면 문제 세트가 생성된다. 생성된 문제들을 검토하며 비슷한 유형의 유사 문제, 숫자만 바꾼 쌍둥이 문제 등으로 교체하는 것이 용이하다. 시중 교재나 수능, 모의고사 기출 문제의 경우 그대로 출제하거나, 숫자만 바꾸거나 난이도를 조정하여 출제하는 것 또한 가능하다. 매쓰플랫 출시 이후는 물론 그 이전에도 다양한 문제은행 서비스가 있었는데, 매쓰플랫 같지는 않았다. 매쓰플랫 출시 5년 만에 75만 명의 유료 사용자, 90% 후반대의 월별 재구매율을 기록한 것만 보아도 그 인기를 짐작할 수 있다. 현재 사교육 기관뿐만 아니라 300여 개의 중고등학교에서 매쓰플랫을 도입해 활용하고 있다. 단순히 문제를 모으거나 디지털화하는 것에 그치지 않고, 교수자들의 기존 업무 방식을 인정하면서, 그 과정에서 생기는 불편과 비효율을 해소하는 데 집중한 결과로 보인다. 교수자의 편의성을 높인 문제은행으로 출발한 서비스지만, 데이터가 축적됨에 따라 맞춤 학습자료(설명자료) 생성, 수학 코스웨어 제공, 학생들의 학습 진도 관리 등으로 기능을 확대하고 있다.

심지어 GPS, 심박수, 자세 등을 측정해 여전히 오프라인에서 이뤄지는 다양한 체육 활동에 대한 피드백을 제공하기도 한다. 대표적인 예로는, 현재 심박수가 학생의 최대 심박 대비 몇 퍼센트인지 측정해 운동 목표를 제시하는 서비스가 있다. 원래 달리기를 잘하는 학생이라면 3km를 10분에 뛰고도 더 뛸 수 있고, 반대의 경우에는 10분 동안 그 절반을 뛰는 것이 한계일 수 있다. 같은 학생이라도 당시의 컨디션에 따라 다를 수 있는데, 절대적인 양으로 운동 강도를 제시하기보다는 최대 심박수 대비 일정 퍼센티지 이상의 심박수를 기록한 구간의 길이를 기준으로 목표를 제시하는 것이다.

교육 데이터를 활용하려고 할 때 대부분의 교육 현장에 전산화된 데이터가 없다는 현실이 가장 먼저 마주하는 큰 걸림돌인데, Gradescope, Qanda, 매쓰플랫은 오프라인 교육 활동을 전산화된 데이터로 남길 충분한 유인 요소

1) "STUDENTS CAN NOW SUBMIT ASSIGNMENTS THROUGH THEIR PHONE ON THE GRADESCOPE MOBILE APP". EQUIT & ACCESS IN PRE K-12 EDUCATION, 2023년 10월 19일 접속, <https://www.ace-ed.org/students-can-now-submit-assignments-through-their-phone-on-the-gradescope-mobile-app/>

를 제공해 준다. 또 오프라인 교육을 더 쉽고 편리하게 준비할 수 있도록 돕는 데이터 활용 서비스를 제공한다. 상황별로 교수자와 학생들의 학습 패턴을 바꾸려는 접근보다는 차선택일 수 있지만, 현실적 한계를 수용하면서도 의미 있는 진전을 이룬다는 점에서 이 또한 어려운 작업이다.

나. 대면 교육 활동과 기록을 디지털로 전환하려는 에듀테크 서비스

물론 교육 활동과 기록 자체를 디지털화하려는 시도도 많다. 컴퓨터, 태블릿, 스마트폰과 같은 IT 기기를 활용해 콘텐츠를 읽고, 문제를 풀고, 답안을 제출하고, 이에 대한 피드백이 이뤄지는 서비스들로, 대다수의 EdTech 서비스가 이에 해당한다. 이들 대부분의 공통된 목표는 맞춤형 교육이다. 2022년 KDATA 백서에도 소개된 문항 응답 이론(Item Response Theory: 문항에 응답한 데이터를 분석해 학생과 능력과 문제의 난이도 등을 추정), 심층 지식 추적(Deep Knowledge Tracing: 학생이 수행한 학습과 문항에 대한 응답을 토대로 학습자의 지식수준을 평가), 학습 경로 개인화(Learning Path Construction: 학생별 문항 응답을 분석해 취약점을 찾고, 취약점을 강화하기 위해 효과적인 학습 경로를 제공)는 여전히 많은 디지털 교육 서비스에 활용되고 있다.

문항 응답 이론은 학원에서 수준별 반 편성을 위해 레벨 테스트를 보는 것, 각 단원이 끝난 뒤 이해도를 파악하기 위해 단원 평가 시험을 보는 것과 비슷하다. 에듀테크 서비스 중에는 미취학 아동부터 초등학교까지를 주 대상으로 삼는 서비스에서 주로 활용한다. 이를 통해 태블릿에서 학습하고 진단 평가를 볼 수 있기 때문이다.

그 과정을 살펴보면, 최초 도입 시에는 해당 과목에 대한 경험과 전문성을 갖춘 교수자 또는 문제 출제자가 문제별 난이도와 해당 단원 값을 입력한다. 이후 학생들의 응답 데이터가 쌓이면 문제의 난이도와 문제 간 연관성, 학생들의 수준을 더 정교하게 평가한다. 그리고 평가 후에는 맞춤형 문제 추천 성능이 개선된다. 웅진씽크빅의 스마트올(키즈), 단비교육의 윈크(키즈), 천재교육의 밀크티(아이/초등)와 같은 서비스가 이에 해당하고, Rind의 산타토익 또한 출시 초기에 이와 같은 기술이 활용되었다.

심층 지식 추적은 문항 응답 이론에 학습자의 문제 풀이 순서를 추가로 고려하는 기술이다. 한 달 전 학생의 이해도와 지금의 이해도가 다를 수 있기 때문이다. 문제에 대한 응답뿐 아니라 어떤 학습을 수행했는지 고려하기도 한다. 시계열(Time series) 정보를 다뤄야 하고, 시간별 문제에 대한 응답 가중치는 정해진 답이 없기 때문에 트랜스포머(이전에는 주로 RNN, LSTM)와 같은 딥러닝 모델이 활용된다. 코세라(Coursera)와 같이 학습 콘텐츠와 문제 풀이가 함께 제공되는 MOOC(Massive Open Online Course) 서비스에서 활발히 사용되고 있다.

학습 경로 개인화는 문항 반응 이론, 심층 지식 추적과 별개의 개념이 아니다. 이 두 기술뿐 아니라 교육학, 교육공학, 해당 학습 콘텐츠, 학생의 상태에 대한 이해를 모두 활용하여 더 나은 학습 경로를 제공하는 접근을 취해야 한다. 예를 들어 개별 학생의 문항별 적정 예상 정답률에 가까운 모델을 만들고, 그 모델의 예상 정답률이 정확하더라도 학생의 학습 동기를 떨어트리는 문제인지, 흥미를 유지하는 문제인지 하는 것에 따라 학습 효과는 크게 달라질 수 있다. 문제의 정답률을 예측하는 것과 달리, 그에 따라 학습해야 하는 내용과 그 순서를 만드는 작업을 하려면 학습 대상을 이해해야 하는 추가적인 노력이 필요하다.

학습 경로 개인화 서비스는 AI 코스웨어라고 불리기도 한다. 코스웨어는 교육과정을 뜻하는 코스(Course)와 소프트웨어의 합성어이다. AI 코스웨어는 AI를 활용해 교육 내용과 절차, 평가, 피드백 등을 제공하는 소프트웨어를 뜻한다. 한국 정부가 추진 중인 AI 디지털 교과서 또한 AI 코스웨어로 볼 수 있다. 다양한 기술과 경험이 필요한 영역인 만큼, AI 디지털 교과서의 경우 기존 교과서를 제작하는 출판사들과 AI 기술력을 갖춘 회사들이 파트너십을 맺어 제작에 참여하는 형태로 사업을 진행하고 있다. 예를 들어, 동아출판은 라이브데이터와 MOU를 체결하여 동아출판의 수학, 정보 과목에 라이브데이터의 AI 엔진을 적용하기로 했다. 비상교육과 미래엔은 엘리스그룹과 협약을 체결하여 비상교육의 정보 교과서 콘텐츠를 맞춤형 AI 디지털 교과서로 개발할 예정이다. 교육부 주도의 AI 디지털 교과서뿐 아니라 클래스팅, Rind, 엘리스, 클래스, 미래엔, 천재교육, 네이버 등의 다양한 기업이 AI 코스웨어를 직접 제공하거나 이를 제작할 수 있는 서비스를 제공하고 있다.

클래스의 경우 질문 기반의 AI 코스웨어를 제공한다. 교실에서 질문이 많지 않은 것은 개별 학생의 문제라기보다 질문하기 어려운 환경의 영향이 더 크다. 클래스는 단독방과 같은 온라인 채팅 공간을 만들어 학생들이 편하게 질문하도록 했다. 학생들이 더 편하게 질문할 뿐 아니라 교수자 입장에서는 같은 질문에 대해 한 번만 답해도 되고, 학생들 간에 서로 답을 해줄 수도 있다. 나아가 AI가 답변하기도 한다. 같은 수업 내용이라도 학생의 질문에 따라 다른 학습 경로가 제공되는 것이다.

클래스팅은 교사 출신의 창업자가 세운 회사로, 공교육에 깊숙이 침투한 에듀테크 서비스다. 출발은 교사와 학부모, 학생들이 온라인에서 소통할 수 있는 서비스로, 교사는 편하게 준비물과 과제 등을 공지하고 학생과 학부모는 쉽게 이를 확인할 수 있다. 필요에 따라 익명으로 교사와 상담할 수 있는 기능을 제공한다. 이후 기능을 추가해, 소통을 돕는 서비스에서 지금은 학습과 학습 관리를 돕는 서비스가 되었다.

사실 교사들이 생각하는 이상적인 학교 모습은 현실적으로 실현하기 어렵다. 예를 들어 다음과 같이 말하는 장면을 떠올려 보자.

“단원 수업을 시작하면서 진단평가로 학생이 해당 단원 학습에 필요한 선행 개념을 이해하고 있는지 확인하고 필요한 경우 보충을 진행합니다. 그 후 단원 안에 있는 차시별로 수업 후 형성평가를 하고, 마지막으로 총괄평가를 통해 수업을 마무리합니다.”²⁾

분명 바람직한 장면이지만 현실적으로는 수업 준비 시간도 부족할 뿐 아니라, 한정된 시간 안에 교사 혼자서 모든 학생의 학습 진도를 확인하고 피드백하는 것은 불가능에 가깝다. 그래서 클래스팅은 이들 요소를 나누고, 각각에 대한 자동화된 솔루션을 제공하는 접근을 취하고 있다. 같은 온라인 수업이라도 녹화된 강의보다 라이브 수업이 학생들의 집중도와 학업 성취도를 높인다는 연구 결과가 있다. 클래스팅의 접근 방법을 보면, 교사가 여전히 에듀테크 서비스보다 나은 영역인 대면 수업과 학습 독려에 집중하도록 한다는 점에서 흥미롭다. 수업과 학습 독려라

²⁾ 클래스팅, “다문화 학급 학력부진 개선을 위한 AI 코스웨어 활용법”, CLASSTING Blog, 2023. 8. 3., 2023년 10월 19일 접속, <https://blog.classting.com/ai-courseware-for-a-multicultural-class/>

는 두 가지 요소를 학급 전체에 공통으로 제공하면서도 개별 이해도에 따라 보완, 심화 학습 단계를 거치도록 자동화한 것이다. 현시점에서 효과적인 접근방법으로 보인다.

2. 교수자를 대체하는 에듀테크 서비스

앞서 대면 교육을 보완하거나 디지털로 진행 · 기록함으로써 기술을 활용해 더 나은 교육을 제공하고자 하는 서비스들을 알아보았다. 충분한 역량을 갖춘 교수자가 없거나 (비용 등의 문제로) 부족한 환경에 있는 학생들을 위한 서비스도 꾸준히 출시되고 있다. 그간 강의 자체는 훌륭하지만, 피드백이 부족해 집중도와 완강율이 떨어진다는 비판을 받아온 온라인 강의 서비스들은 생성형 AI를 활용한 피드백으로 단점을 빠르게 개선하고 있다. 이때 가장 활발히 활용되는 형태는 챗봇이다. ChatGPT 등장 이후 온라인 동영상 강의 서비스 회사뿐 아니라 수많은 회사가 교육에 챗봇을 활용하기 시작했다. 국내 기업들도 교육에 챗봇을 활용하고 있다. 하지만 현시점에서 영어로 된 챗봇들이 대체로 다른 언어 기반의 경우보다 더 나은 성능을 보인다. 한국어를 포함한 그 이외 언어의 언어모델 성능을 발전시키려면 시간이 더 필요하다.

가. 교수자를 대체하는 에듀테크 서비스 사례

칸 아카데미(Khan Academy)의 칸미고(Khanmigo)는 가장 널리 알려진 교육용 챗봇 중 하나로 수학, 과학, 영어 등 다양한 과목에 대해 튜터링을 돕거나 튜터링에 가까운 기능을 제공한다. 칸 아카데미는 ChatGPT를 기반으로 기존에 갖고 있던 칸 아카데미의 학습 콘텐츠를 학습시키고, 교육에 적합한 대화 흐름이 이어질 수 있도록 튜닝한 챗봇이다. 교육적으로 바람직하지 않은 질문에 대해서는 답변을 회피하기도 한다.

이는 ChatGPT의 뛰어난 성능에 기반한 것이기는 했지만, 칸 아카데미가 자체적으로 가이드라인을 설정하지 않았다면 ChatGPT를 교육에 활용했을 때 발생하는 부작용을 학생들에게 그대로 노출했을 것이다. 예를 들어, ChatGPT는 모르는 내용이거나 윤리적으로 문제가 있는 내용이 아닐 경우 대부분의 질문에 대해 한 번만 답변한다. 이와 달리 칸미고는 한 번에 답변할 수 있는 내용도 학생이 스스로 생각해 볼 수 있도록 최소한의 힌트를 주는 모습을 자주 보인다. 아직까지는 이와 같은 튜닝을 칸 아카데미의 교육 전문가가 설정한 기준에 따라 진행했겠지만, 이후 학생들의 학습 데이터가 쌓인다면 학생들의 질문에 반응하는 방법에 대해서도 더 다채롭게 학습할 것이다.

물론 ChatGPT의 성능이 완전하지 않다 보니 칸미고 또한 수학 문제에서 틀리거나, 사실이 아닌 답을 사실처럼 말할 때가 있다. 그러나 말 그대로 한 주가 멀다하고 문제점이 개선되고 성능이 올라가는 만큼 향후에는 더 많은 영역에서 활용될 수 있을 것이다. 앞서 언급한 엘리스와 클래스팅을 포함한 국내외 많은 기업도 자사 서비스에 챗봇을 접목해 교수자가 일일이 답하기 어려운 정도의 많은 질문에 대해 거의 실시간으로 답변을 제공하고 있다.

스픽(Speak)은 영어 회화 학습 서비스인데, OpenAI와 파트너십을 맺고 ChatGPT를 일찍이 자사 제품에 도입했다. ChatGPT는 기본적으로 언어 모델이고, 대부분의 학습 데이터가 영어로 된 까닭에, 영어에 대한 이해도와 지식

수준이 매우 높다. 다양한 상황과 역할을 설정하는 롤 플레이나 자유 토픽으로 AI와 대화를 나눌 수 있다. 이는 원어인 회화 서비스보다 훨씬 저렴할 뿐만 아니라, 학습자가 원하는 시간에 언제든지 회화를 할 수 있다는 장점도 있다. 이해하지 못한 문장이 있으면 터치 한 번에 번역과 설명을 제공해 주고, 영어로 표현하기 어려운 설명을 한국어로 말하면 이를 적절히 번역한 표현을 알려 준다. 학습자가 영어로 표현했을 때, 틀린 표현이나 어색한 문장을 고쳐 주고 설명하기도 한다. 듀오링고(Duolingo) 또한 일찍이 OpenAI와 파트너십을 맺고 ChatGPT를 기존 서비스에 접목했다. 듀오링고도 스픽과 비슷하게 학습자가 말하거나 작성한 문장에 대해 피드백 해주고 다양한 상황에서 롤 플레이를 할 수 있는 기능을 제공한다.

스픽과 듀오링고가 제공하는 기능은 기존 영어 학습 서비스들에서 오래전부터 제공되는 기능이다. 그럼에도 언급해야 할 만큼 ChatGPT를 도입하기 전과 후의 사용자 경험이 아주 크게 달라졌다. 다룰 수 있는 주제가 다양해지고, 내용은 전보다 훨씬 자연스러워졌다. 첨삭의 질 또한 크게 높아졌다. 직접 영미권에 가서 살거나 원어민 교사와 꾸준히 대화하고 피드백을 받는 것과 비교해 아직 확실히 더 낫다고 볼 수는 없지만, 비용과 편의성, 접근성을 고려한다면 대다수 사람들에게는 더 나은 선택지가 되는 수준까지는 온 것으로 보인다.

Riid는 스마트폰 등을 이용해 토익 객관식 10문제가량 풀면, 예상 점수와 수준별 문제를 추천해 주는 산타토익을 서비스하는 회사다. Riid는 토익 이외의 시험과 교과목에도 비슷한 기술과 접근방법을 적용하고 있고, 지금은 다수의 회사가 비슷한 접근을 취하고 있다. 그런데 최근에 공개한 Quizium은 이전의 제품들과 조금 다르다. 학습이 꼭 정해진 시험이나 교과목으로만 이뤄지는 것은 아니라는 점에 주목했다. 한 조사에 따르면, 86%의 미국 유튜브 사용자는 새로운 것을 배우기 위해 유튜브를 본다³⁾ 한다. 다만, 일방적인 비디오 시청의 학습 효율이나 시청 후 이해도는 높지 않다. 코세라와 같이 온라인으로 동영상 강의를 제공하는 MOOC 서비스들의 경우 이를 개선하기 위해 강의 중간에 Pop-up Quiz를 포함해 학습자가 잘 이해했는지 스스로 점검할 수 있도록 한다.

그런데 Riid의 Quizium은 사전에 준비된 동영상과 pop-up quiz 대신 사용자가 어떤 동영상을 보든 상관없이 그 동영상을 잘 이해했는지 확인할 수 있는 문제를 만들어 주는 서비스다. 기존 AI 코스웨어가 정해진 교육과정 콘텐츠 내의 경로 개인화를 추구했다면, 이것은 유튜브라는 사실상 무한정 열린 콘텐츠에 대해 점검할 수 있다. 유튜브 내 사용자의 흥미에 따라 콘텐츠를 선택해 학습하고, 학습이 잘 이뤄져있는지 확인할 수 있도록 하는 셈이다.

나. 생성형 AI를 교육에 활용할 때 데이터가 사용되는 방식

칸미고, 스픽, 듀오링고, Quizium과 같이 ChatGPT를 활용한 에듀테크 서비스가 데이터를 활용하는 방식과 순서는 다음과 같다. 높은 수준의 언어 모델을 만들기 위해, 해당 서비스와 직접적인 관련이 없는 다양한 언어 데이터를 사전 학습한다. 비용과 난이도의 문제로, 현재는 대부분의 에듀테크 회사에서 ChatGPT, Bard(또는 이들의 API)를 쓰거나 Llama와 같은 오픈소스 모델을 기본 모델로 활용하는 것으로 이 과정을 대체한다. 서비스 개발사가 직접 사

3) "86% of U.S. viewers say they often use YouTube to learn new things", Think with Google, 2023년 10월 19일 접속, <https://www.thinkwithgoogle.com/marketing-strategies/video/youtube-learning-statistics/>

전 학습을 하지 않을 뿐, ChatGPT, Bard, Llama를 개발할 때 데이터를 학습하는 것을 사전 학습이라고 볼 수 있다.

사전 학습 이후에는 각 서비스 개발사가 보유한 자체 데이터를 활용해 서비스의 목적과 구조에 맞도록 언어 모델을 튜닝한다. 데이터의 양이 많고, 학습해야 할 내용이 복잡한 경우에는 언어 모델의 파라미터 값을 직접적으로 바꾸는 형태의 파인 튜닝(Fine Tuning)이 이뤄지고, 간단한 톤 앤 매너는 프롬프트 엔지니어링을 통해 바꿀 수 있다. ChatGPT 출시 전에는 목적에 따라 데이터를 준비하고 파인 튜닝을 성공적으로 진행한 후에 실제로 쓸 수 있을 정도의 성능이 제공되었다면, ChatGPT 도입 후에는 파인 튜닝 없이도 상당한 수준의 성능을 얻게 되었다. (여전히 파인 튜닝은 태스크별로 성능을 추가로 개선하는 데 도움이 될 수 있다.)

앞의 두 단계를 통해 데이터를 학습한 서비스를 학습자가 사용하면, 실시간으로 사용자의 입력이 기록된다. 이때 언어 모델을 활용한 서비스들은 사용자의 입력을 데이터로 활용해 그에 맞는 답변을 제공한다. 맥락 학습(In-context learning)이라고도 한다. 서비스 개발사가 준비할 필요 없이, 사용자가 서비스를 사용하는 과정에서 데이터가 생성되고 반영되는 구조다.

ChatGPT가 출시되기 전까지 대다수의 에듀테크 서비스는 데이터를 의미 있게 활용하기까지 적지 않은 축적 시간이 걸리거나, 전문가가 일일이 데이터를 만들어야 했다. 그런데 ChatGPT가 도입되면서 이미 기존 에듀테크 서비스들이 축적한 데이터보다 훨씬 더 많은 데이터를 확보하게 되었고, 간단한 프롬프트 엔지니어링과 약간의 학습자 데이터만으로도 뛰어난 성능의 교육 서비스를 제공하게 되었다. 토론, 일기, 메일 작성, 보고서 작성, 번역, 리서치, 면접 준비, 새로운 개념 학습, 요약뿐 아니라 영어 기사를 어렵게, 쉽게, 짧게 재작성 하는 등 수많은 기능이 있다. ChatGPT를 있는 그대로 쓰는 것만으로도 매우 뛰어난 수준에 이른 것이다. 여전히 빠르게 성능이 개선되고 있고, 활용 사례도 늘어나고 있어 사례를 나열하는 것이 큰 의미가 없을 정도다. 일례로 국내에서 진행된 생성 AI 에듀 해커톤에서 학습자가 글을 읽을 때, 휴대폰 전면 카메라가 학습자 눈동자를 추적, 단어 인지 여부를 파악해 단어장에 저장하고 학습 과학에 근거한 간격 학습을 제공하는 서비스가 제안되기도 했다.⁴⁾

3. 추가로 고려해볼 사항

가. 교수자의 역할에 대한 재정의와 구분

에듀테크 서비스가 교수자들의 기존 업무를 일부 보조하거나 자동화할 수 있다는 사례는 꾸준히 쌓이고 있다. 더 나아가 교수자를 대체하는 에듀테크 서비스가 여럿 나왔음에도, 아직 해당 서비스가 시장에 보편적으로 침투하지는 못했다. 교수자가 앞으로도 한동안은 계속해야 하는 역할이 있을 것이고, AI 자동화나 보조를 통해 교수자의 짐을 덜어주는 보완 역할이 있을 것으로 보인다.

4) 이주영, “앨리스그룹 ‘생성 AI 에듀 해커톤에 1500명 참가’” AIT타임스, 2023년 8월 29일, <https://www.aitimes.com/news/articleView.html?idxno=153177>

지금은 개별 교수자가 이를 판단해 도입하고 활용해야 하는데, 이 경우 신뢰와 책임 소지의 문제가 생길 수 있다. 교수자가 개별적으로 에듀테크 서비스를 도입하고는 더 나은 교수자의 기존 역할과 그렇지 않은 부분을 임의로 구분한다고 해도, 그것이 통계적으로 충분히 신뢰할 만한 구분이라고 할 수 없다.

근래 증거 기반 교수라는 개념이 주목받는데, 이는 교육에 앞서 의학에서 활용된 개념이다. 즉, 의학 분야처럼 교육에서도 의사결정 또는 판단의 바탕이 되는 체계적이고 과학적인 근거를 마련해야 한다. 이를 위해 교육 현장(임상)의 데이터를 확보하면서 체계적인 연구가 함께 이뤄져야 한다. 그래야만 이미 시간이 부족해서 지친 교수자들에게 해야 할 업무를 하나 늘려주는 것 이상의 의미를 지닐 수 있다. 현실적인 여건을 잘 파악하여 역할과 책임을 기술에 부여해도 될 지점을 발견해야 한다. 꼭 필요한 지점에서 기술이 교수자를 도울 수 있도록 그 범위와 개입 정도를 정의하는 일이 함께 병행되어야 한다.

나. 학습 대상에 대한 재정의

교육을 통해 이루고자 하는 바는 하나의 단어나 문장으로 축약하기 어렵다. 2022년 개정 교육과정에서 추구하는 인간상의 주된 내용은 아니지만, 현실적인 면에 중점을 두고 생각해 보면, 단기적으로는 필기·실기 시험을 잘 보고, 그 후에는 좋은 일자리를 얻는 것, 나아가 사회에 기여하고 경제적 가치를 창출하는 것 또한 교육을 통해 얻고자 하는 목표라고 할 수 있다. 이미 AI가 주요 교과목 시험을 대개 사람보다 잘 본다. 의사, 변호사처럼 좋은 일자리를 얻는 보증수표로 여겨지는 전문직 자격증 시험 또한 예외가 아니다. AI가 대체하기 어려울 거라던 작가·사진가·미술가·작곡가 등의 창의적인 직업 또한 극소수 최상위 인력을 제외하고는 AI의 발전으로 위협받기는 마찬가지다.⁵⁾ AI는 상업적 가치가 있는 글과 그림, 동영상도 순식간에 만든다. 심지어 계속해서 빠르게 발전하고 있다.

이처럼 기존 교육에서 성공 사례에 해당하면 갖게 되는 주요 능력은 이제 AI를 통해 무료로 가까운 비용으로 활용할 수 있다. 코딩 또한 이미 상당 부분 AI가 주도하는 방식으로 대체되었고 앞으로 그 비중이 더 커질 것이 명확하다. 이런 상황에서도 뒤늦게 코딩 교육이 미래 지향적인 교육과정으로 도입되고 있다는 점을 고려한다면, AI의 변화를 따라잡지 못한다고 볼 수 있다.

학습하기로 한 것들을 더 잘 학습하도록 돕는 에듀테크 서비스들은 지금도 계속 출시되는데, 무엇을 학습할 것인지 제대로 정의하지 못한다면 목표를 달성하고도 실패한 것과 다름없는 상황을 맞을 수 있다.

5) M. Koivisto, S. Grassini, “Best humans still outperform artificial intelligence in a creative divergent thinking task”, Scientific Reports Vol. 13 article No. 13601, 2023.

4. 전망

가. AI 디지털 교과서를 축매로 한 공교육-사교육 간 협력

AI 디지털 교과서는 기존 교과서를 제작하는 출판사뿐 아니라 사교육 영역에서 기술력을 쌓은 에듀테크 회사들도 제작에 참여한다. AI 디지털 교과서가 취지대로 제작된다면, 7년 내외로 개정하던 기존 종이 교과서와는 전혀 다르게 운영될 것이다. 종이 교과서를 액자나 벽지와 같은 무생물 상품이라고 한다면, AI 디지털 교과서는 애완동물과 같은 유기체에 가깝다. 한 번 만들고 끝나는 것이 아니라 계속해서 관리하고 발전시켜 나가야 한다. 구조와 목표의 차이로 인해 공교육과 사교육은 서로 잘할 수 있는 영역이 다른데, 특히 AI 디지털 교과서를 구성하기 위한 기술 개발과 적용은 민간 기업이 더 잘할 것이므로 활발한 협력이 기대된다. 지금까지 시험은 공교육계에서 출제하고, 사교육계에서는 공교육 시험을 잘 보게 하는 역할을 주로 했다면, 이제는 함께 프로젝트를 진행하는 과정에서 서로 잘할 수 있는 일을 이해하고 배우며 각자의 역할을 분배하고 조정할 것이다.

나. AI 디지털 교과서와 교육마이데이터 간 시너지

2025년 마이데이터 전 분야 도입을 앞두고, 교육은 우선 적용 대상에 포함되었다. 어떤 데이터를 어떤 형태로 적용하게 될지 구체화해야 할 사안이 아직 많지만, 그럼에도 AI 디지털 교과서 도입과 교육마이데이터 간의 시너지는 기대할 만하다. 기존의 교육 데이터에도 다양한 정보가 담겨 있지만, AI 디지털 교과서 도입으로 평가 결과 및 의견 이외에 학습량, 학습 속도 등을 포함한 학습 과정 전반을 데이터로 남길 수 있는 토대를 마련할 수 있다.

학생 개인의 시험 성적과 구술에만 의존하던 때와는 다르다. 학교에서의 학습 과정 전반에 대한 실제 데이터를 활용할 수 있다면 학습을 도울 방법은 훨씬 다양해지고, 개별화될 여지가 많다.

다. AI코스웨어와 증거 기반 교수의 도입 증가

IT 기기 활용으로 교육 데이터의 수집과 축적이 용이해지고, 데이터 분석을 통해 얻을 수 있는 효용이 커지다 보면 점점 더 많은 데이터가 축적될 것이다. AI 코스웨어를 설계할 때 다양한 가설과 교수법이 활용되는데, 이전 종이 교과서를 보고 문제집을 풀 때와 달리 가설 별로 교육 효과를 측정하는 것이 용이해진다. 예를 들어, 학생별로 얼마나 어려운 문제를 풀어야 교육 효과가 가장 좋을지를 알기 위해서는 해당 문제를 풀었을 때 학생의 지식 변화 모델을 적용하는 것만으로는 부족하고, 문제 난이도에 따라 몇 퍼센트의 학습자가 실제로 해당 문제를 푸는 시도를 했는지, 어떻게 풀어냈는지, 얼마나 걸렸는지를 함께 봐야 교육 효과를 더 잘 측정할 수 있는데, 이들을 기록할 수 있다. 이에 따라 특정 개인의 주장만을 전적으로 수용한다거나 통계적으로 신뢰도가 낮은 소수 사례를 근거로 교수법을 결정하는 대신, 증거 기반 교수의 방법론이 주류가 되고 이를 통해 검증된 결과들이 차츰 보편화될 것으로 기대한다.

제7장

新 데이터 비즈니스

김인현 대표 투이컨설팅

디지털 경제로 진화함에 따라 데이터 산업이 등장했다. 데이터 산업은 전통적 산업 분류에서는 존재하지 않는다. 전통적 산업 분류는 재화와 서비스를 기준으로 하지만 데이터는 재화도 서비스도 아니기 때문이다.

데이터가 점점 더 중요해짐에 따라 우리는 데이터 산업을 정의하고, 데이터 산업으로 발생하는 비즈니스를 새롭게 시도해야 한다. 데이터 산업은 데이터 주도 비즈니스와 데이터 비즈니스로 구분할 수 있다. 데이터 주도 비즈니스의 경우엔 기업이 기존 비즈니스의 성과를 높이기 위해서 데이터를 사용하며, 데이터 비즈니스의 경우엔 데이터를 주산물로 가치를 만든다.

데이터 비즈니스에는 데이터 프로덕트, 데이터 마켓플레이스, 데이터프랩, 데이터 서비스, 데이터 리터러시 등이 있다. 이러한 데이터 비즈니스가 작동하기 위해서는 제도가 필요하다. 또한 데이터를 수익으로 창출하는 비즈니스 모델을 찾아야 한다. 현재 제도는 틀을 잡아가고 있고, 데이터 비즈니스 모델은 시작되는 수준이다.

장기적으로 기업의 데이터 주도 비즈니스와 데이터 비즈니스는 상호 배타적이라기보다는 상호 보완적인 관계다. 데이터 비즈니스를 잘하는 것이 데이터 주도 비즈니스에 도움이 된다. 조직은 데이터 기반 비즈니스에서 나아가 적합한 데이터 비즈니스를 찾아내고 도입하기 위한 시도를 계속해야 한다.

1. 데이터 산업의 정의

데이터 산업은 새로운 산업일까? 또는 기존 산업의 한 측면일까? 우리나라 통계청의 정의에 따르면, 산업은 유사한 성질을 보이는 산업 활동에 주로 종사하는 생산단위의 집합이다. 산업 활동은 각 생산단위가 노동, 자본, 원료 등 자원을 투입하여, 재화 또는 서비스를 생산 또는 제공하는 일련의 활동과정이다. 소비자는 재화와 서비스를 사용하거나 소비함으로써 소비 행위의 효용을 높인다. 이때 재화는 물리적 형태가 있고, 서비스는 물리적 형태가 없다.

데이터 산업이 성립하려면, 데이터는 재화거나 서비스여야 한다. 하지만 데이터는 재화나 서비스와는 다르다. 데이터는 직접적으로 소비자의 효용을 높이지 않는다. 데이터가 스스로 가치를 지니지 않기 때문이다. 데이터는 활용을 통해서 가치가 만들어진다. 데이터의 가치는 재화와 서비스가 제공하는 효용을 더 높이는 것이다.

그런데 데이터가 재화도 아니고 서비스도 아니라면 데이터 비즈니스는 산업이 될 수 있는 것일까?

데이터는 디지털경제의 핵심이며 모든 산업에서 중요한 요소다. 심지어 어떤 기업에선 비즈니스의 핵심이 재화와 서비스에서 데이터로 바뀌기도 한다. 이렇듯 기존에 통용되고 있는 산업의 정의로는 수용하기 어려운 점이 있

지만, 이미 데이터 산업은 크고 주요한 산업이 되었다. 때문에 국가 차원에서 데이터 산업을 기획하고 육성하지 않으면, 국가 경제의 지속적인 발전을 기대할 수 없을 것이다.

이러한 데이터 산업을 제대로 정의하려면, 데이터의 가치 창출 방식을 이해해야 한다. 데이터는 두 가지 방식으로 가치를 창출한다. 첫째는 데이터를 이용하여 기존 비즈니스의 성과를 높이는 것이고, 둘째는 데이터 자체로 수익을 올리는 것이다. 전자를 ‘데이터 주도 비즈니스(Data Driven Business)’라고 하고 후자를 ‘데이터 비즈니스(Data Business)’라고 한다. 데이터 주도 비즈니스와 데이터 비즈니스는 모두 데이터 산업에 포함하는 것이 맞다.

〈표 4-7-1〉 데이터 주도 비즈니스와 데이터 비즈니스 비교

	Data-Driven Company	Data Company
전략	비즈니스 최적화	비즈니스 전환
가치 창출	효율성 제고 수익 증대	데이터 프로덕트 새로운 비즈니스 모델
핵심 질문	우리가 하고 있는 일을 어떻게 더 잘할 수 있을까?	지금 하고 있는 일이 아닌 다른 일을 어떻게 할 수 있을까?

* 출처: D. Burbank, “Data Strategy Bootcamp”, Enterprise Data World (2021)

대부분 기업의 관심은 데이터 주도 비즈니스에 있다. 데이터 비즈니스는 아직 초기 단계다. 데이터 권리와 의무 등 관련 제도와 데이터의 가치 창출 방식 등 비즈니스 패턴에 대한 시도들이 진행되고 있다. 앞으로 데이터 비즈니스 쪽에서 더 많은 기회가 생길 것이다. 빅데이터의 3V 개념(Volume, Velocity, Variety)을 처음 이야기한 도그 레이니는 데이터 관점에서 기업을 세 가지로 분류하고, 기업가치를 비교했다.

〈표 4-7-2〉 데이터기업 유형과 기업가치

Average Company	Data Savvy Company	Data Product Company
데이터 활용이 보통 수준인 기업	데이터를 똑똑하게 잘 사용하는 기업	데이터를 상품화하여 수익을 창출하는 기업
시장가치 = 장부가치	시장가치 = 장부가치 2배	시장가치 = 장부가치 3배

* 출처: D. Laney, “발표 자료”, Data Monetization Workshop, (2021).

도그 레이니에 의하면 데이터를 잘 활용하는 기업의 시장가치는 장부가치의 2배인데 비하여, 데이터를 상품으로 수익을 창출하는 기업의 시장가치는 장부가치의 3배다. 데이터로 기존 비즈니스를 잘 지원하는 것보다, 데이터로 직접 가치를 창출하는 것이 효과가 더 크다는 주장이다. Data-Driven Company 또는 Data Savvy Company는 기존 산업에 속하는 기업들 중에서 데이터를 잘 사용하는 기업을 의미한다. 반면 Data Company와 Data Product Company는 데이터가 주산물인 기업이다.

2. 데이터 비즈니스

앞 문단에서 기존 비즈니스를 잘하기 위해 데이터를 활용하는 경우를 데이터 주도 비즈니스로, 데이터가 주된 가치 창출 수단이 되는 경우를 데이터 비즈니스로 정의했다. 이에 따르면 데이터 주도 비즈니스는 기존 산업 분류에 나타나는 비즈니스 형태로, 이미 존재하는 비즈니스다. 반면 데이터 비즈니스는 기존에는 존재하지 않는 새롭게 등장하는 비즈니스다. 재화나 용역이 아닌 데이터가 비즈니스 가치 창출의 주산물이다.

데이터 비즈니스는 데이터의 생명주기에 따라 구분할 수 있다. 데이터 생명주기는 데이터 생산, 데이터 유통, 데이터 준비, 데이터 활용 등이 있다. 또한 데이터 활용 역량 확보를 지원하는 데이터 리터러시 비즈니스의 중요성도 커지고 있다. 하트만(Haertmann) 등은 2016년에 데이터 비즈니스 스타트업 100개를 임의로 선택하고 이들의 특성을 분석하여 6개의 카테고리로 분류하였다¹⁾. 하트만의 분류는 데이터가 비즈니스의 원천이 되는 비즈니스 모델 유형을 밝혀주었다는 점에서 의미가 있다. 하지만 실제 운영되고 있는 비즈니스 모델 사례들을 조사하고 이를 분류하여 비즈니스모델 유형을 정의함으로써 유형 분류의 설득력은 떨어진다.

비즈니스모델연구소는 데이터비즈니스를 DaaS(Data as a Service), IaaS(Information as a Service), AaaS(Answers as a Service) 등의 세가지 유형으로 분류하였다.²⁾ 하지만, 이러한 분류는 데이터 수집과 생산, 유통 측면의 비즈니스 모델은 포함되어 있지 않다. 스타트업 컨설턴트인 아쉬시(Aashish)는 데이터로 수익을 창출하는 비즈니스 유형을 데이터 사용자, 데이터 공급자, 딜리버리 네트워크, 데이터 촉진자 등의 네 가지로 분류하였다³⁾. 아쉬시의 분류는 데이터주도 비즈니스와 데이터 비즈니스가 혼재되어 있고, 데이터 활용 비즈니스 유형에 대한 정의가 미흡하다.

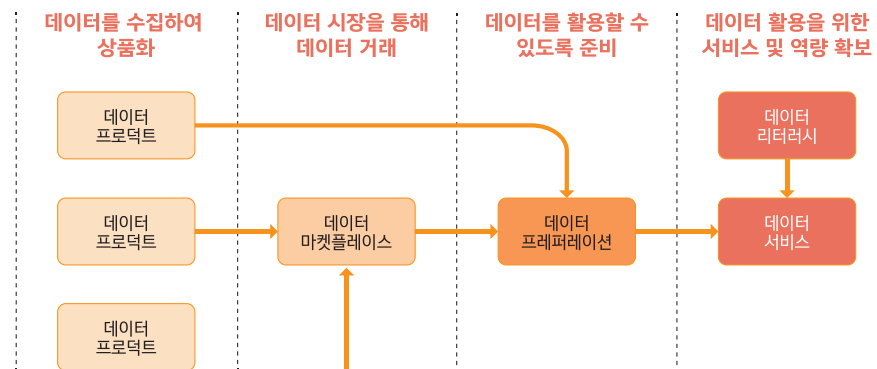
분석과 인공지능은 데이터의 소비자이다. 데이터는 발생하고 수집된 상태로는 사용될 수 없다. 데이터를 학습할 수 있도록 준비하는 활동이 필수적이다. 이를 담당하는 비즈니스가 데이터 프래퍼레이션이다. 또한 데이터 활용은 전문가에서 일반인으로 확대되고 있다. 조직의 데이터 활용 성패를 좌우하는 것은 조직과 개인의 데이터 활용 수준이다. 따라서 데이터 역량을 갖추도록 하기 위한 비즈니스가 떠오르고 있다. 기존 연구를 토대로 추가 요건을 반영한 데이터 비즈니스 유형 분류는 〈그림 4-7-1〉과 같다.

1) Capturing value from big data – a taxonomy of data-driven business models used by start-up firms, Hartmann, Philipp Max, Mohamed Zaki, Niels Feldmann and Andy Neely, International Journal of Operations & Production Management (2016).

2) The characteristics of data-driven business model development and how to succeed, BMI Lab(2022)

3) The Data Monetization : Big Data Business Models, Pahwa, Aashish, Feedough(2023)

〈그림 4-7-1〉 데이터 비즈니스 유형



* 출처: 투이컨설팅

가. 데이터 프로젝트

데이터를 상품화하여 제공하고 대가를 받는 비즈니스 모델이다. 본업에서 발생한 데이터를 상품화하는 경우가 있고, 수요를 예상하고 판매를 목적으로 데이터를 수집하여 제공하는 경우가 있다.

신용카드 회사들은 신용판매 비즈니스에서 확보한 데이터를 가공하여 프로젝트로 만들어서 제공하는 비즈니스에 적극적이다. 신한카드는 620여 개의 데이터 상품을 판매한다. 신한카드의 데이터 상품 매출은 2021년 이후 매년 100억 원대를 넘어섰다. 다른 카드사들도 데이터 비즈니스 전담 조직을 운영하거나 금융 빅데이터 플랫폼을 통해 카드 소비 데이터를 제공하고 있다.

프리랜서 거래 플랫폼인 크몽은 '데이터 구매 · 구축' 페이지를 통해 24개의 데이터세트 판매를 중개한다. 여기에는 N사 블로그 DB 추출, 건축인허가 데이터, 통신판매사업자 DB, 장기요양기관 DB, APT 실거래가 추출 등 24개의 데이터 상품이 등록되어 있다. 데이터 크롤링과 데이터 라벨링 등 주문형으로 데이터를 수집 및 준비해 주는 서비스도 이용할 수 있다.

기업 또는 기관이 보유한 데이터를 거래하는 경우 지금까지는 주로 프로젝트 방식으로 데이터를 제공했다. 수요처가 자신의 데이터 요건을 제시하면, 데이터 공급자가 이에 맞게 데이터를 추출하고 가공하여 제공하는 방식이다. 프로젝트 방식으로 데이터를 제공하는 것은 비용을 충당할 수 있는 수준의 수익을 확보하기 어렵다. 그러다 보니 데이터 수요자들로선 쓸 만한 양질의 데이터가 부족하다는 점이 가장 크고 절실하게 느끼곤 한다.

현재 데이터 수요는 빠르게 증가하는 데 비하여, 데이터 프로젝트의 공급이 부족한 실정이다. 수요자가 믿고 활용할 수 있는 양질의 데이터, 카탈로그 정보, 시각화 정보 등을 안정적으로 제공할 수 있다면 데이터 프로젝트 비즈니스는 확대될 것으로 예상된다.

나. 데이터 마켓플레이스

데이터 공급자와 데이터 수요자 간에 중개 역할을 하는 비즈니스 모델이다. 데이터 마켓플레이스는 소비 데이터, 공공 데이터, 금융 데이터, 시장 데이터, 센서 데이터 등 다양한 출처에서 확보할 수 있다. 데이터 수요자는 기업, 연구기관, 개발자, 데이터 과학자 등 다양하다. 데이터 마켓플레이스에선 데이터를 구매하고 판매하기 위한 플랫폼을 제공하여 데이터의 접근성을 높인다.

〈그림 4-7-2〉 국내외 데이터 획득 경로 비교

외국 (Q: 다음 각각의 데이터 획득 경로를 이용하는가?)	유형	국내 (Q: 다음 중 주로 이용하는 데이터 획득 경로는?)
72 Supplier data	내부데이터	온라인 회원 및 고객이 이용 동의한 데이터 24.9
69 Consumer usage data		자산의 데이터 가공(데이터화)에 의한 데이터 확보 13.3
67 Anonymous consumer data		마케팅 등 고객 커뮤니케이션을 통해 수집 12.4
		각종 센서를 통한 자동 데이터 수집 4.0
	공공데이터	내방객 등 오프라인을 통한 데이터 수집 3.5
97 Open data		공공기관에서 제공되는 데이터(공공데이터) 10.9
98 Publicly available competitor data		
77 Data from blogs/product reviews	웹사이트 데이터 수집	소셜(SNS), 인터넷 등을 통한 데이터 수집 18.6
77 Social media data		수집 솔루션(tools) 등에 의한 웹데이터 수집 6.5
92 Proprietary datasets from data aggregators	거래소 데이터 거래	데이터 거래를 통한 데이터 확보 5.9
90 Data from hyperscalers(ie., Google, Amazon)		
84 Data from distributors/partners		
67 Data from platform providers	기타	
91 Analyst/industry reports		

*자료: 1) 외국현황: 스테티스타 홈페이지, (최종방문일 : 2022.6.20.)(<https://www.statista.com/statistics/1235514/worldwide-popular-external-data-sources-companies/>)
2) 국내현황 : 과학기술정보통신부·한국데이터산업진흥원, 2021 데이터산업 현황조사, 2022

* 출처: 국회입법조사처, 데이터 거래 활성화를 위한 거래소, 거래사, 크롤링의 현황과 개선과제, (2022).

데이터 획득 경로를 비교해 보면 외국은 제3자를 통한 데이터 수집이 많아지고 있다. 반면 우리나라에선 데이터 수요자가 직접 데이터를 획득하고 있다. 국내 데이터 수요 기업들이 데이터 거래를 통해 데이터를 확보하는 비율은 5.9%에 불과하다. 또한 외국의 경우 민간 시장에서 데이터 거래가 이루어지는 반면, 우리나라에선 제도와 정책으로 탄생한 공공데이터 포털, 데이터 거래소, 빅데이터 플랫폼 등이 거래를 주도하고 있다. 예를 들어 데이터 거래소를 통하면, 증권이나 천연자원 등처럼 균일한 특성을 보인 제품을 다수의 판매자와 구매자의 호가로 가격이 결정되는 식이다. 그렇지만 데이터는 동일한 데이터 세트라고 하더라도, 수요자가 요구하는 데이터 품질 기준에 따라 가격

이 달라져야 한다. 데이터 상세화 수준, 적시성, 범위 등에 따라 데이터 수집과 가공의 난이도가 달라지기 때문이다.

데이터 거래소는 데이터 마켓플레이스와 성격이 다르다. 데이터 마켓플레이스는 공급자와 수요자가 자유롭게 만나고 편리하게 거래를 성립하도록 하는 기능을 하는 것이 중요하다. 큰 방향에서 보면 데이터 마켓플레이스로 나아가야 한다.

하지만 아직 갈 길은 멀다. 데이터 거래소, 데이터거래사, 데이터 전문기관, 데이터 결합기관 등의 다양한 제도가 준비되어 작동하지만, 데이터 마켓플레이스 비즈니스 확산을 위해 도움이 되지는 않는다. 정부가 선정하여 지원하는 빅데이터 플랫폼이 산업별로 선정되어 운영되지만 성과가 크게 나타나지 않았다.

성과가 나오려면 데이터 수집의 수단인 웹 크롤링을 금지하거나 제한하기보다는, 디지털 경제 진전의 추세에 맞춰 과감하게 이를 활성화하는 방향으로 제도가 준비되어야 한다. 또한 데이터 사용을 확산하기 위해서는 데이터 보유자(또는 데이터 생산자), 그리고 데이터 수요자(또는 데이터 소비자)가 쉽게 소통하고 협업할 수 있는 환경이 필요하다.

이때 데이터 마켓플레이스는 데이터 확산 생태계의 핵심 역할을 한다. 비즈니스 모델은 제도에 좌우된다는 점을 명심해야 한다. 공공 부문에서 데이터 거래소라는 개념을 중심으로 제도를 수립하고 빅데이터 플랫폼 비즈니스 확산을 주도하는 방식은 최적의 방식이 아니다. 그보다는 민간 생태계가 필요에 따라 형성될 수 있도록 지원하는 방식이 더 효과적이다.

다. 데이터 프레퍼레이션

데이터 프레퍼레이션(Data preparation)은 줄여서 데이터프렙(DataPrep)이라고 한다. 데이터 프렙은 분석 모델을 만들기 위하여 데이터를 준비하는 비즈니스다. 데이터프렙은 데이터 수집, 데이터 정제, 데이터 변환, 피쳐엔지니어링, 시각화, 데이터 분할, 데이터 저장, 자동화 등의 과정으로 진행된다. 데이터를 활용하여 인공지능 모델을 개발하는 경우에는 데이터 라벨링(Data Labeling) 작업도 포함된다. 데이터 라벨링은 데이터 포인트에 레이블, 태그 또는 주석을 부여하여 모델을 훈련하고 평가하기 위한 작업이다.

데이터 분석의 목적은 데이터로부터 의사결정에 활용할 수 있는 애널리틱스(또는 알고리즘)를 뽑아내기 위한 것이다. 애널리틱스를 만드는 과정에서 통계 기법 또는 머신러닝 등을 적용한다. 애널리틱스의 성능을 좌우하면서 노력이 가장 많이 드는 작업이 데이터프렙이다. 데이터프렙을 조직 내부의 프로세스로 구현하는 것이 데이터 엔지니어링이다. 데이터 엔지니어링은 데이터를 수집, 저장, 처리 및 관리하기 위한 기술과 프로세스를 개발하고 관리하는 용도로 활용된다. 데이터 엔지니어링을 자동화함으로써 데이터 활용의 속도를 높이고 효율화하기 위한 시도가 데이터옵스(Dataops)와 엠엘옵스(MLOps)이다.

다만, 데이터프렙을 위해 필요한 데이터를 충분히 수집해야 하는 문제가 있다. 또 수집된 데이터를 분류하고 표준화하고 정제하는 등의 상당한 노력을 해야 한다. 특정 분석 주제를 위하여 데이터를 준비한다는 것은 말처럼 쉽지 않다. 일단 데이터 자체를 모으기도 어렵고, 투자 규모를 감당하기도 쉽지 않다. 따라서 공동으로 데이터를 준비

하는 방안이 시도된다.

국가 차원에서는 정부 예산으로, 인공지능 서비스 개발에 필수적인 인공지능 학습용 데이터를 구축하고 이를 개방하는 노력이 진행되고 있다. 데이터 댐 사업의 일환으로 추진했고 성과도 있었지만, 구축한 데이터 댐을 유의미하고 폭넓게 활용해야 하는 과제가 남아 있다. 3년간 약 1조3천억 원의 예산을 투입했지만 실제 발생한 매출액은 투입 예산의 6% 정도였다. 데이터 수요자가 특정되지 않은 상태에서 데이터를 준비하기 때문에 활용 범위가 제한된다는 점이 문제로 지적된다.

데이터 공유 니즈를 해결하고 데이터 준비를 효율적으로 수행하기 위한 목적으로 민간 부문에서도 데이터 댐 사업이 시도되고 있다. 일례로 '디지털 라이프 데이터 댐'이 2022년 1월에 출범했다. 통신기업(LG유플러스), 은행(NH농협은행), 금융기관(KB국민카드·롯데카드·하나카드), 신용평가회사(NICE평가정보), 유통·제조 기업(LG전자), 메타버스 기업(바이브컴퍼니), 공공기관(한국데이터산업진흥원·경찰대학), 정보보호기술 스타트업 기업(크립토크), 빅데이터컨설팅 회사(NICE지니데이터) 등이 참여했다. 오픈라인과 메타버스 등 다양한 업종 간 데이터를 결합하여 고객분석, 마케팅 전략 모델, ESG지수 등 다양한 상품을 출시할 예정이다. 또 2021년에는 데이터 얼라이언스가 결성되었고 민간 데이터 댐을 구축하는 계획을 발표했는데 SK텔레콤, 신한카드 코리아크레딧뷰로, GS리테일, 부동산114 등이 참여했다.

데이터 프레퍼레이션을 추진하기 위해서는 인프라와 인력이 필요하다. 또한 데이터 기여와 활용의 규칙을 세우고, 비용 및 수익을 조정하기 위한 활동도 수행해야 한다. 대규모 데이터 보유 조직들이 데이터를 기반으로 동맹을 맺는 것에 대한 공정 거래 이슈도 해소해야 한다. 이는 제휴 협약 차원에서 해결될 수 있는 문제가 아니다. 당장의 필요로 데이터 동맹이 시도되지만 실질적인 효과를 거두려면, 독자 비즈니스 모델이 출발하고 지분 출자를 통해서 제휴하는 방식이 필요하다.

라. 데이터 서비스

데이터 서비스는 '데이터 또는 데이터 가공 결과'를 제공하는 비즈니스다. 그 대상은 데이터 소비자 또는 데이터를 사용하는 애플리케이션이다. 데이터 서비스 비즈니스를 위해서는 개인정보보호 등 필요한 법적 규제를 준수해야 한다. 데이터 서비스 방식은 필요에 따라거나 수요자 니즈에 따른다. 그에 맞춰 실시간 또는 배치 처리 형태도 지원한다. 실시간 제공을 위해 API(Application Programming Interface) 기술을 사용한다.

데이터 서비스는 데이터 활용 방식과 서비스 주제에 따라 다음과 같은 유형이 있다.

– **데이터중계서비스**: 어떤 조직이 보유한 데이터를 외부의 기관 또는 개인에게 전송해야 하는 의무가 있지만, 실제로 데이터 전송을 위한 인프라스트럭처를 구축하지 못했거나 이를 운영하기 위한 기술을 보유하지 않았을 때, 또 투자 여력이 부족한 경우 제3의 조직에게 데이터 전송을 위탁할 수 있다. 데이터중계서비스 비즈니스는 데이터 전송 의무를 안전하게 효율적으로 수행해주는 역할을 한다. 데이터중계서비스 비즈니스를 위해서는 관련 법규로부터 지정을 받아야 한다.

- **데이터결합서비스**: 두 개 이상의 개인정보를 융합해서 사용할 경우, 관련 규제는 제3의 기관에 적용된다. 개인정보 유출 및 오남용을 방지하기 위한 목적이다. 데이터결합서비스를 하기 위해서는 개인정보보호법의 데이터결합전문기관 인증을 받거나 신용정보법의 데이터전문기관 지정을 받아야 한다. 결합하고자 하는 데이터에 신용정보가 포함된 경우에는 신용정보법의 데이터전문기관 지정을 받는 것이 필수이다. 데이터결합전문기관은 필요한 데이터의 소싱 기능도 제공하는 추세다.
- **DaaS(Data as a Service)**: 가장 잘 알려진 데이터 기반 비즈니스 모델이다. 사용자가 필요로 하는 데이터를 특정하여 온라인으로 제공한다. 제공하는 데이터 건당 지급하는 방식을 주로 사용한다. 광고, 구독, 정액 요금 방식을 적용할 수도 있다. DaaS 비즈니스 경쟁력은 데이터 품질, 데이터 적시성, 반응 속도, 서비스 안정성 등이 좌우한다. DaaS 사용조직은 확보한 데이터를 활용하여 자신의 서비스를 제공한다.
- **IaaS(Information as a Service)**: 자체 수집한 데이터 또는 외부에서 확보한 데이터를 기반으로 분석한 정보 또는 보고서를 판매한다. 정보 조회 건당 비용을 지급하는 경우가 일반적이다. 프리미엄 요금제를 적용할 수도 있다. IaaS 비즈니스 경쟁력은 데이터를 분석하여 양질의 정보를 제공할 수 있는 능력에 달려 있다. 신용정보서비스 회사의 경쟁력은 신용등급 판정의 적합성이 좌우한다.
- **AaaS(Answer as a Service)**: 고객이 제시한 질문에 데이터를 활용하여 답변을 제공하는 비즈니스 모델이다. 고객은 답변을 토대로 전략을 수립하거나 조정하며 계획을 평가하고 피드백한다. 고객사의 질문에 답하기 위해서 AaaS 사업자는 산업에 대한 이해와 비즈니스 통찰력이 필요하다. 프로젝트 베이스로 과금하기도 하고, 일정 기간 질문에 답하는 구독 방식을 적용하기도 한다. 모바일 앱 사용자 분석, 디지털 광고 기획 등에서 효과를 내기 위해서는 데이터 확보 및 분석 능력이 중요하다. 거대언어모델(LLM)을 조직 구성원들이 활용할 수 있도록, 개발과 운영을 지원하는 비즈니스도 등장하고 있다.

마. 데이터 리터러시

데이터 리터러시 비즈니스는 조직의 데이터 활용 능력을 향상해 주는 비즈니스다. 데이터 리터러시 개념은 데이터를 분석하여 통찰력을 찾는 데 그치는 것이 아니라, 이를 실제 업무에 적용하여 성과를 내는 것까지 포함한다. 데이터 리터러시를 갖추기 위해서는 데이터 전문가를 확보하는 것으로 충분하지 않다. 조직 구성원들이 전반적으로 데이터를 이해하고 활용할 수 있어야 한다. 또한 데이터 기반 의사결정이 일반화될 수 있도록 프로세스가 변경되고 데이터 중심 문화가 정착되어야 한다.

단순히 데이터 교육 과정을 개설하고 운영하는 것만으로는 이러한 변화를 달성할 수 없다. 데이터 리터러시 비즈니스 모델은 다음 내용을 지원하여야 한다.

- **데이터 문화**: 데이터 자산의 가치를 인정하고, 데이터에 근거한 의사결정 체계가 확립된다. 전략과 계획은 데이터를 기반으로 수립한다. 조직 평가지표에는 데이터 관련 지표가 포함된다.

- **데이터 교육**: 조직 구성원에게 '데이터의 비즈니스 활용, 데이터 도구와 인프라 활용, 비즈니스에서 산출되는 데이터 카탈로그 정보' 등에 관해 교육함으로써 구성원 스스로 데이터를 조작하고 분석하고 활용하는 능력을 갖추도록 한다.
- **데이터 품질**: 내부 및 외부의 데이터 소스에 따른 품질 수준을 측정하고 활용 타당성을 확인한다. 데이터 품질의 바람직한 수준을 달성하기 위한 계획을 수립하고, 필요한 활동을 관리한다.
- **데이터 규제**: 데이터 보안, 개인정보보호, 소비자 및 고객의 신뢰 등을 유지 및 강화할 수 있는 데이터 거버넌스 체계를 수립하고 적용한다.
- **데이터 성숙 수준**: 업무 영역별로 또는 조직 단위별로 데이터 성숙 수준의 현재를 측정하고 목표를 설정한다. 목표 수준으로 이행하기 위한 과제를 도출하고 실행을 통제한다.

3. 데이터 비즈니스 과제

데이터 주도 기업은 하고 있는 비즈니스 성과를 높이기 위해 데이터를 활용하는 기업이다. 반면 데이터 기업은 데이터로 가치를 창출하는데, 데이터는 소비자의 효용을 직접 높이지 않는다. 소비자로서는 데이터에 대한 비용을 지불하고 싶은 동기가 약하다. 이런 점에서 데이터 기업의 비즈니스 모델을 설계하는 것이 쉽지 않다.

데이터 기업은 데이터를 수집하고 유통하고 가공하고 활용한다. 이때 데이터에는 민감한 개인정보가 포함된다. 데이터의 권리와 책임 문제가 발생할 수 있는데, 이는 기존 재화에서 적용하던 것과는 다른 면이 있다. 예를 들어 물건의 소유권은 판매를 통해 이전되지만, 데이터는 거래가 성사되더라도 데이터 주체의 권리가 소멸되지는 않는다.

데이터 권리는 2018년 유럽의 GDPR 제정으로 제도화가 시작되었다. 데이터 권리를 침해할 경우 페널티는 상당히 크다. 생성형 인공지능의 등장으로 데이터 규제 방법에 대한 논의는 더욱 중요해지고, 한편으로 복잡하게 전개되고 있다. 이러한 법적 지원이 체계를 잡는다면 분쟁 위험이 줄어든 상태로 데이터의 안정적 공급이 가능해질 것으로 보인다.

이런 흐름을 볼 때 데이터 비즈니스는 더욱 빠르게 성장할 것이며, 앞으로 두 가지 과제가 남아 있다. 우선 가치를 지속적으로 만들어낼 수 있는 비즈니스 모델을 찾아야 한다는 것이다. 물론 쉽지 않다. 제도는 만들어지고 있지만 아직 익숙하지 않고, 위반에 따른 리스크는 크다. 그럼에도 나아가야 할 명확한 방향은 있다.

데이터 기업 또는 데이터 주도 기업 중 어느 쪽으로 나아가야 할 것인가 하는 문제는 사실 상호 보완적이다. 언뜻 보기엔 두 가지 중 하나를 선택하는 것 같지만, 실제로는 어느 하나를 선택하는 것이 아니다. 어쩌면 답은 하나다. 어차피 데이터 주도 기업으로 성공하기 위해서도 데이터 기업으로 성공해야 한다. 그래야 제대로 된 시너지 효과를 낸다. 데이터 기업이 되기 위한 시도를 계속해야 할 이유다.

5^{PART}

데이터산업 기술 동향

- 제1장 · 합성 데이터 생성 기술 동향
- 제2장 · 클라우드 스토리지 기술 동향
- 제3장 · 데이터 분석 기술 동향
- 제4장 · 데이터 보안 기술 동향

제1장 합성 데이터 생성 기술 동향

황인준 교수 고려대학교 전기전자공학부

방대한 양의 데이터가 다양한 루트로 끊임없이 생산되는데, 이는 고도화된 인공지능 기술을 효과적으로 활용하는 사안과 밀접한 관련이 있다. 즉, 좋은 인공지능 모델을 만들기 위해선 필요한 양질의 데이터를 방대하게 확보해야 한다. 그래야 만들어진 인공지능 모델을 통해 실제로 활용할 수 있는 양질의 데이터를 생성할 가능성이 커진다. 이는 상호 보완적인 관계를 지니며, 바로 이것이 합성 데이터의 근간이 된다. 이 장에서는 합성 데이터에 대한 간단한 정의와 필요성을 언급하고, 많이 사용되는 주요 합성 데이터 기술 및 응용 분야에 대해 살펴보고자 한다.

1. 서론

최근 다양한 센서부터 다기능의 모바일 기기 및 이를 기반에 둔 거대한 사회관계망(Social Networks)을 빈번하게 활용하면서, 개인이나 조직에서 활용하고자 하는 데이터의 형태가 복잡하게 다양해지고 있다. 특히 끊임없이 생산되는 방대한 양의 데이터를, 고도화된 인공지능 기술을 이용하여 활용하려는 노력이 지속되고 있다. 이는 기계뿐만 아니라 개인과 조직이 주체가 되어 이뤄지고 있으며 교통이나 물류, 제조, 소매, 전자상거래, 헬스케어에 포함하는 다양한 산업 분야에서 시도되고 있다.

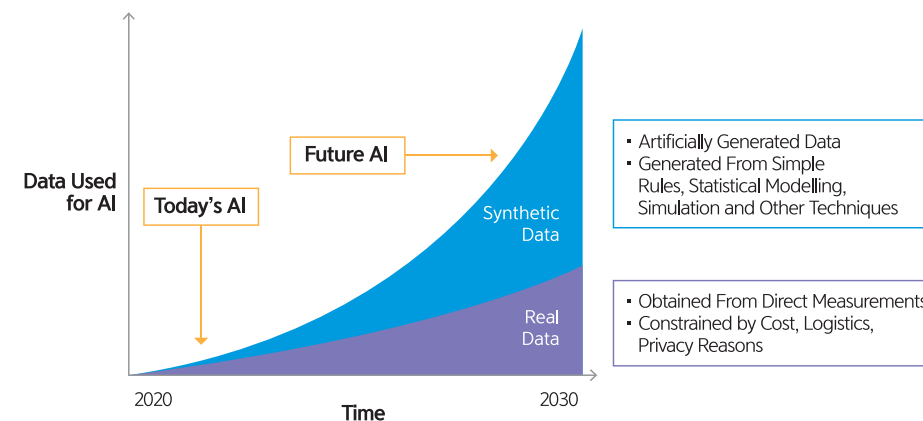
하지만 여러 미디어에서 생성된 이미지, 텍스트, 음성과 같은 데이터는 형식 없이 분산된 특성 탓에 적절한 데이터 라벨링 작업이 요구된다. 효과적인 인공지능 학습을 위해서는 그 목적에 맞춰 연관된 데이터를 수집해야 하기 때문이다.

데이터 라벨링 작업은 다음과 같다. 우선 데이터나 데이터의 메타 정보에 적절한 라벨이나 태그를 추가한다. 그러면 인공지능 모델이 해당 정보를 이해할 수 있도록 하는 데이터 전처리 작업을 진행한다. 이를 통해 비정형 또는 반정형 데이터를 수집하고 정답(Ground Truth) 주석을 추가한다. 이 작업을 거쳐 연관된 데이터를 지도 학습에 적용할 수 있게 해준다.

데이터 라벨링 과정은 개발하고자 하는 인공지능 모델의 목적에 따라 작업 방식이 달라지는데, 일반적으로 인적 자원을 활용하여 수작업으로 직접 라벨을 제작한다. 최근에는 라벨링 제작으로 소요되는 많은 인적 물적 비용을 최소화하기 위해 미리 학습된 다른 인공지능 모델 등에서 생성된 합성 데이터를 활용하는 방법들이 주목 받고 있다.

[그림 5-1-1] 합성 데이터와 실제 데이터의 활용 전망

By 2030, Synthetic Data Will Completely Overshadow Real Data in AI Models



* 출처: Leinar Ramos, Jitendra Subramanyam, "Synthetic Data Is the Future of AI," Gartner, 2021.6.24., 2023년 8월 15일 접속.
<https://www.gartner.com/en/documents/4002912>

2. 합성 데이터 개요 및 필요성

합성 데이터란 현실 세계에서 획득된 데이터가 아니라, 실제 데이터와 유사한 통계적 속성을 보이도록 인공적으로 생성된 데이터를 의미한다. 컴퓨터 시뮬레이션이나 인공지능 알고리즘을 통하면 '라벨이 있는' 합성 데이터를 생성할 수 있어, 데이터 라벨링을 위한 시간과 비용을 절감할 수 있다. 따라서 학습 데이터를 많이 필요로 하는 인공지능 모델을 개발할 때 필요한 학습 데이터를 확보하기 위해 합성 데이터를 보완적으로 활용하는 사례가 증가하고 있다. 이러한 경향은 2021년에 발행된 Gartner 'Synthetic Data Is the Future of AI' 보고서에¹⁾ 잘 나타나 있다. 이 보고서에서는 향후 2030년까지 합성 데이터를 활용하는 사례가 늘어 실제 데이터의 활용 빈도를 능가할 것으로 전망한다. 또한, MIT Technology Review는 인공지능을 위한 합성 데이터를 2022년도 10대 혁신 기술 중 하나로 선정²⁾하는 등 향후 큰 발전과 활용이 기대되는 분야이다.

대표적으로 개인정보나 민감한 정보 탓에 현실 세계에서 수집한 학습 데이터를 그대로 활용하는 것이 법적 제약으로 어려운 경우, 또는 현실 세계에서 드물게 발생하는 데이터를 학습 데이터로 사용해야 하는 경우에 합성 데이터 생성 기술을 활용할 수 있다. 합성 데이터에는 테이블 형식과 같이 구조화된 데이터와 텍스트나 영상이 있다. 또한 오디오와 같이 구조화되지 않은 데이터도 포함될 수 있다. 이러한 합성 데이터 내에 실제 데이터가 존재하는지

1) Leinar Ramos, Jitendra Subramanyam, "Synthetic Data Is the Future of AI," Gartner, 2021.6.24., 2023년 8월 15일 접속.
<https://www.gartner.com/en/documents/4002912>

2) Melissa Heikkilä, "MIT 테크놀로지 리뷰 선정 '2022년 10대 미래 기술'" MIT 테크놀로지 리뷰, 2022년 3월 3일.
<https://www.technologyreview.kr/mit-2022-10/>

여부에 따라 완전 합성 데이터와 하이브리드 합성 데이터로 구분될 수 있다.³⁾

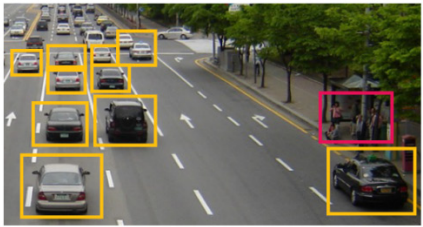
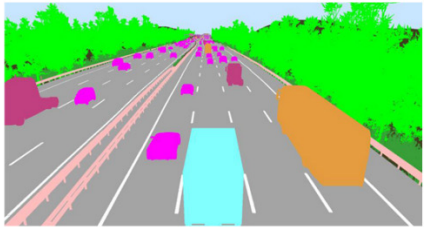
인공지능 모델을 학습시키기 위해서는 정확하게 레이블이 지정된 방대한 양의 학습 데이터가 필요하다. 그러나 대규모 데이터를 수집하고 각 데이터에 필요한 레이블을 지정하는 것은 현실적으로 시간과 비용이 많이 든다. 그 때문에 데이터 확보에 어려움을 겪고 있다. 또한 데이터 수집 후에는 데이터의 품질을 높이기 위해 데이터 정제 과정을 수행해야 하며, 그중에서도 개인정보 유출 위험이 있는 데이터의 경우 그 사용이 상당히 제한될 수 있다. 이러한 이유로 발생하는 소위 데이터 부족(Data Shortage) 문제는 인공지능 학습에서 과적합 및 과소적합을 초래하여 성능 면에서 치명적인 결과를 도출할 수 있다.

최근 합성 데이터의 경우 데이터가 불충분해서 발생하는 인공지능 모델의 성능 저하 문제를 극복하는 대안으로 부상했다. 이를 통해 비용 절감 외에도 실제 세계에서 수집된 데이터가 잠재적으로 민감한 경우, 예를 들어 개인정보 보호 문제를 해결할 때 합성 데이터를 활용한다.

또한 실제로 수집된 데이터의 경우 모든 상황을 적절하게 표현하는 자료를 확보할 수 없을 때가 있는데, 합성 데이터의 경우 적절한 자료를 수집하지 못했더라도 알맞은 데이터를 생성할 수 있으므로 이러한 제약을 극복할 수 있다. 그로 인해 수집된 데이터만을 사용할 때보다 더 풍부한 정보를 생성할 수 있어, 실세계에 대한 더 많은 정보를 데이터에 반영할 수 있다. 즉, 합성 데이터를 활용하면 실제 현상에서는 포착하기 어려운 드문 사례를 현실적인 가능성으로 제시할 수 있다. 결과적으로 자료의 폭과 깊이가 풍부하게 다양해진다.

[그림 5-1-2] 실제 데이터와 합성 데이터 비교

Real Versus Synthetic Data for AI Video Analysis

Real Data for Video Analysis	Synthetic Data for Video Analysis
<ul style="list-style-type: none"> Limited examples for infrequent scenarios (e.g., collisions, extreme weather) High cost of obtaining, processing, storing and labeling Inaccurate labeling (bounding boxes) Privacy issues 	<ul style="list-style-type: none"> Diverse scenarios can be generated (e.g., collisions, extreme weather) Zero marginal cost of obtaining additional data (fixed cost to create a data generator) High-granularity labeling (pixel level) No privacy issues
	

* 출처: Leinar Ramos, Jitendra Subramanyam, "Synthetic Data Is the Future of AI," Gartner, 2021.6.24., 2023년 8월 15일 접속.
<https://www.gartner.com/en/documents/4002912>

3) 김태원, '가짜' 데이터가 만드는 '진짜' 인공지능 시대, 한국지능정보사회진흥원12호(2022), p. 14.

3. 합성 데이터 생성 기술

1) 통계적 분포 접근법

통계적 분포 접근법은 합성 데이터를 생성하는 간단한 방법의 하나로, 실제 데이터의 통계적 분포를 관찰하고 이를 통해 실제 데이터와 유사한 분포를 보이는 데이터를 생성하는 것이다. 통계적 분포 접근법은 테이블 형식 데이터와 같이 단순한 구조의 소규모 데이터일수록 효과적인 기술이다. 일반적으로 정규 분포, 지수 분포, 카이-제곱 분포 및 로그 정규 분포 등과 같은 통계적 확률 분포를 사용한다.

2) 시뮬레이션

Unity, Unreal 등과 같은 게임 엔진을 활용하여 실제와 비슷한 형상의 데이터를 확보하거나, CAD(Computer-Aided Design)를 이용하여 실제와 유사한 데이터를 생성하는 방법이다. 시뮬레이션을 통해 획득한 합성 데이터가 얼마나 현실과 유사한 속성을 갖는지 살피게 되는데, 현실과 유사한 속성의 정도가 인공지능 모델의 성능에 영향을 미치는 주요 요인이다.

3) SMOTE(Synthetic Minority Oversampling Technique)⁴⁾

SMOTE는 데이터 불균형을 해결하기 위한 오버 샘플링 기법⁵⁾ 중 하나로 합성 데이터 생성을 위해 많이 사용되는 모델 중 하나이다. SMOTE는 K-최근접 이웃 알고리즘을 기반에 두고, 소수 클래스의 데이터 샘플과 가장 가까운 샘플을 연결하는 선분상에서 데이터를 생성하는 방식을 취하며 동작 방식은 다음과 같다.

첫째, 소수 클래스의 각 샘플에 대해 특징 공간에서 k개의 최근접 이웃을 선택한다.

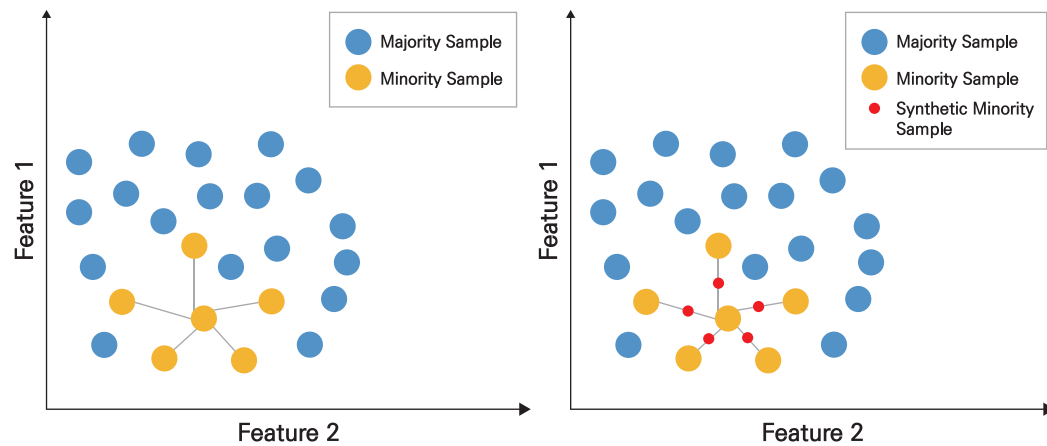
둘째, 그중에서 하나의 최근접 이웃을 무작위로 선택하고, 선택한 샘플과 원래 샘플의 특징을 결합하여 데이터를 생성한다.

셋째, 이 과정을 미리 지정한 횟수 또는 클래스 간 균형이 맞춰지는 수준이 될 때까지 반복한다. 이러한 과정을 통해 소수 클래스 샘플의 수를 증가시켜 클래스 분포의 균형을 맞춰줄 수 있으며, 모델의 소수 클래스 분류 및 전체적인 성능을 높이는 장점이 있다.

4) Nitesh V. Chawla et al. "SMOTE: synthetic minority over-sampling technique", Journal of artificial intelligence research Vol.16, 2002.

5) 낮은 비율 클래스의 데이터를 증식하여 학습을 위한 충분한 데이터를 확보하는 방법

[그림 5-1-3] SMOTE 적용 예시

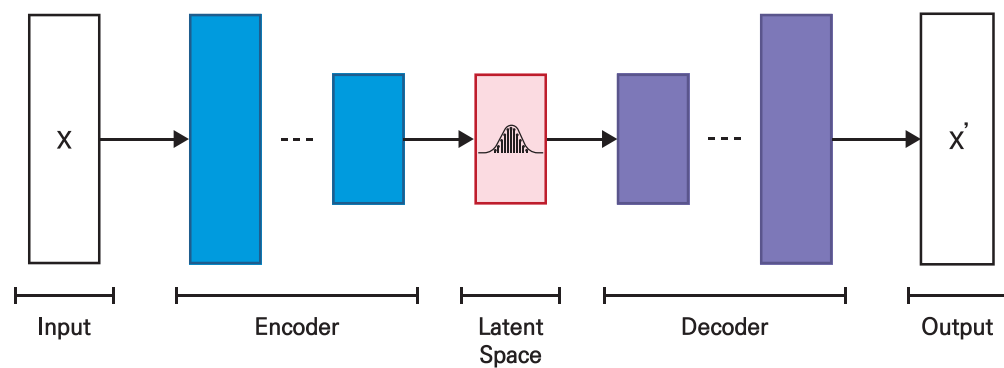


* 출처: A. Vijayvargiya, C. Prakash, R. Kumar, S. Bansal, & J. M. R. S. Tavares, "Human knee abnormality detection from imbalanced sEMG data", Biomedical Signal Processing and Control Vol.66, 2021.

4) VAE(Variational Auto Encoder)⁹⁾

VAE는 아래 그림처럼 원본 데이터를 낮은 차원의 잠재 공간으로 압축하여 새로운 데이터를 생성하는 모델이며, 인코더-디코더 구조로 구성된다. 인코더는 고차원 입력 데이터를 저차원 표현 벡터로 압축하며, 디코더는 주어진 표현 벡터를 원본 차원으로 복원하는 역할을 한다. VAE는 입력 이미지가 들어오면 그 이미지에서의 다양한 특징들을 잠재 공간의 확률 분포로 변환하며, 디코더에서는 이 확률 분포에서 샘플링한 값으로 원래의 데이터를 복원한다. 따라서 VAE 모델에서는 확률 분포를 이용한 표현 벡터 조종을 통해 새로운 합성 데이터를 생성할 수 있으며, 데이터의 노이즈에 강하다는 장점이 있다.

[그림 5-1-4] VAE 구조



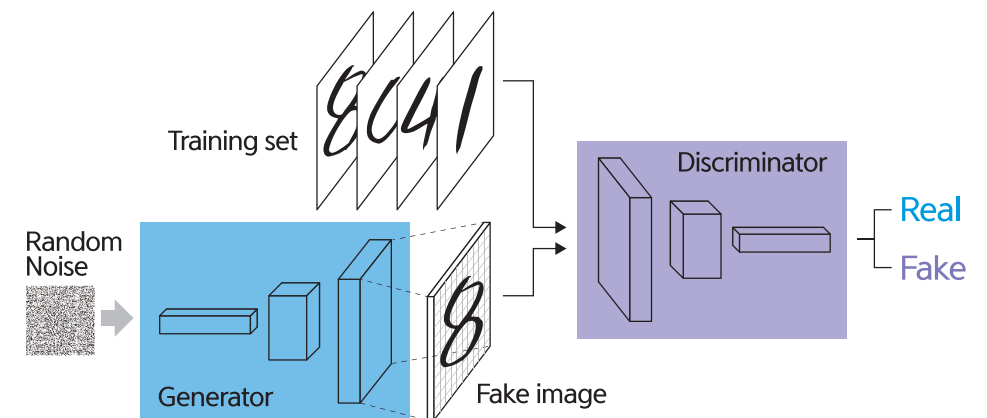
* 출처: "Variational autoencoder," 위키피디아, 2023.8.15., 2023년 8월 27일 접속, https://en.wikipedia.org/wiki/Variational_autoencoder

⁹⁾ D. P. Kingma, M. Welling, Auto-encoding variational bayes. arXiv. (2013).

5) GAN(Generative Adversarial Network)⁷⁾

GAN은 아래의 그림과 같이 가짜 데이터를 생성하는 생성기(Generator), 그리고 진짜와 가짜 데이터를 구분하는 판별기(Discriminator)로 구성된다. 생성기는 판별기를 최대한 잘 속이기 위해 노력하고, 판별기는 진짜 데이터와 가짜 데이터를 최대한 구별하기 위해 경쟁적으로 학습하여 진짜 데이터에 가까운 가짜 데이터를 생성한다. GAN은 진짜와 같아지는 학습을 통해 사용자가 입력한 조건에 가장 가까운 샘플을 만들어 좀 더 생생한 데이터를 생성할 수 있다. GAN의 장점으로서는 실제 데이터의 특성을 빠르게 학습하고 정확한 표현을 위해 더 빠르게 반복하는 능력을 들 수 있다. 비록 GAN을 훈련시키는 과정은 상당히 복잡하지만, 현재는 안정적인 훈련 방법들이 많이 연구되어 상당히 높은 퀄리티의 합성 데이터를 생성하는 것이 가능해졌다.⁸⁾ 또한, 2019년에는 Lei Xu가 테이블 형식의 데이터를 생성하기 위해 CTGAN(Conditional Tabular GAN)을 제안하여 주목 받기도 했다. 이는 GAN의 생성기와 판별기를 기존의 CNN(Convolutional Neural Network) 대신 LSTM(Long Short-term Memory network)과 MLP(Multilayer Perceptron)로 대체한 것이다.

[그림 5-1-5] GAN 구조 예시



* 출처: "A Short Introduction to Generative Adversarial Networks," Thalès' blog, 2017.6.7., 2023년 8월 15일 접속, <https://sthalles.github.io/intro-to-gans/>

6) Diffusion

Diffusion 모델은 역확산 과정을 통해 실제 데이터와 비슷한 합성 데이터를 생성한다. 예를 들어 이 모델에서는 이미지에 노이즈를 추가하여 이미지 손상 과정을 거친 다음, 추가된 노이즈를 제거하는 과정을 학습하여 점차 원본 이미지에 가까운 이미지를 생성한다. 역확산 과정에서 각 단계의 노이즈 변환은 확률 분포에 기반하여 이루어지므로 확률적인 생성을 통해 다양한 결과를 얻을 수 있다.

⁷⁾ Alvaro Figueira, Bruno Vaz. "Survey on synthetic data generation, evaluation methods and GANs". Mathematics Vol.10 No.15, 2022, p.2733.

⁸⁾ Ian Goodfellow et al. "Generative adversarial networks", Communications of the ACM Vol.63 No.11, 2020, pp.139-144.

GAN 모델과 비교하여 학습이 안정적이라는 장점이 있다. 다양한 연구를 통해 높은 품질의 데이터를 생성할 수 있다는 결과가 확인되면서 최근 새로운 추세로 떠오르고 있다.

4. 합성 데이터 응용 사례

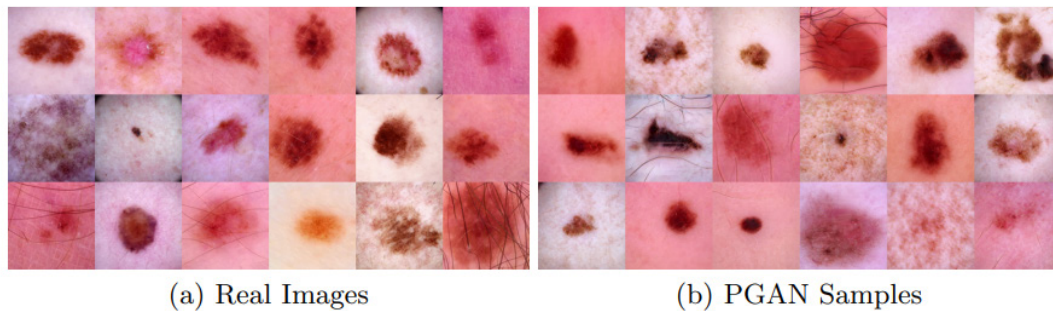
1) 의료 데이터

의료 데이터에는 개인정보 등 민감한 정보가 많이 포함되어 있어 실제 데이터를 생성하는 것이 제한될 때, 이를 보완하기 위해 합성 데이터를 사용하기도 한다. 민감한 정보들을 비식별화하여 사용할 수 있지만 이를 처리하는 데 많은 비용과 시간, 노력이 소모된다. 더구나 그런 과정을 거쳐도 민감한 정보들이 완벽하게 비식별화되었다고 보장하기도 어렵다.

그런 면에서 합성 데이터는 가상의 데이터라는 장점이 있다. 개인정보 유출과 같은 문제에서 실제 데이터 활용보다 자유롭기 때문이다. 그리고 희소 병변 사례와 같이 실제 데이터가 소량이고 이를 확보하는 데 오랜 시간이 드는 경우, 합성 데이터를 이용한다면 빠르게 학습 데이터를 생성할 수 있다. 미국의 의료 스타트업 큐라이(Curai)는 GAN을 이용하여 약 40만 건의 의료 데이터를 생성하여 인공지능 모델을 훈련시키는 데 활용한 사례가 있다.

실제로 Christoph Baur가 작성한 논문⁹⁾에 따르면 양성 및 악성 피부 병변의 이미지 10,000장의 데이터셋을 대상으로 Progressive Growing GAN(PGAN)¹⁰⁾을 사용하여 실제 데이터와 유사한 피부 병변의 이미지를 고해상도로 합성한 후, 3명의 피부과 전문의와 5명의 딥러닝 전문가로 하여금 Virtual Turing Test¹¹⁾를 해본 결과, 합성된 데이터와 실제 데이터의 구분이 어렵다는 결과가 도출되었다.

[그림 5-1-6] 기존 데이터 샘플과 생성된 합성 데이터 비교 예시



* 출처: Christoph Baur et al, Generating highly realistic images of skin lesions with GANs, Springer, (2018), p.2.

⁹⁾ Christoph Baur et al, Generating highly realistic images of skin lesions with GANs, Springer, (2018), p.2.

¹⁰⁾ 이미지를 생성하기 위해 저해상도부터 고해상도까지 점진적으로 학습하는 GAN 모델

¹¹⁾ 무작위로 섞어 놓은 합성된 이미지와 실제 이미지를 각 전문가가 분류한 후 분류의 정확도 평가

2) 자율 주행 데이터

NVIDIA는 합성 데이터 생성 관련 분야에서 약진하고 있으며 Lockheed Martin은 NVIDIA와 협력하여 합성 데이터 생성을 위한 플랫폼을 개발하고 있다. 또한, NVIDIA는 최근 공간 이미지 기반의 증강 기법을 적용해 인공지능 학습 데이터를 생성하는 프로세스를 구축하였다. 일례로 NVIDIA의 Omniverse Drive Sim은 사실에 가까운 현실 세계를 모사해 자율 주행과 관련된 학습 데이터를 획득할 수 있었다. 이는 시간 · 노력 · 비용 측면에서 현실 데이터를 수집하는 것보다 훨씬 경제적인 가능성을 보였다. 테슬라 역시 최근 사고 사례 중심의 합성 데이터를 만들어 인공지능을 학습시키는 것으로 알려졌다.

3) 금융 데이터

금융권의 경우 비정상 행위 탐지를 위해 인공지능 기술을 활용할 수 있다. 금융거래 특성상 비정상적인 거래는 소수이기 때문에 정상적인 거래와 비정상적인 거래 간의 데이터 불균형 문제가 발생할 수 있다. 이 경우 합성 데이터를 활용해 비정상적인 거래의 데이터를 충분히 확보하면, 데이터 불균형으로 발생하는 성능 문제를 해결할 수 있다. 실제로 미국 금융서비스 기업인 아메리칸 익스프레스는 GAN을 활용한 합성 데이터를 만들어 금융 사기를 탐지하는 인공지능 모델의 성능을 높였다.

4) 소매 데이터

소매 업계에서는 스타트업 케이퍼(Caper)와 같은 기업들이 제품 이미지 분류 모델을 학습시키기 위해 3D 시뮬레이션을 사용하고 있다. 3D 시뮬레이션을 통해 각 제품 당 5개의 이미지로부터 제품을 여러 각도에서 촬영하여 100~1,000장의 이미지로 생성한 후 분류 인공지능 모델을 학습시켰다. 그 결과 케이퍼는 최대 5만 개의 제품을 인식할 수 있는 AI 카트를 만들었다. 그 덕분에 고객들은 계산대 앞에서 기다릴 필요 없이 쇼핑 카트에 탑재된 기능을 이용하여 쇼핑한 물품들을 바로 스캔해 계산할 수 있다.

5) 보험 고객 관리 데이터

기존의 보험사에서는 서비스 개선을 위해 청구, 판매 및 이탈 데이터, 시장 및 설문조사 등의 데이터를 수집하여 사용했다. 하지만 이러한 방법은 최근 새롭게 바뀐 데이터 보호 규정을 근거로 규제받고 있다. 이에 스위스의 보험사 중 하나인 Die Mobiliar는 소비자의 프라이버시를 보호하고 데이터 보호 규정을 준수하기 위해 합성 데이터를 활용하였다. 그리고 실제로 합성 데이터를 기존 고객 이탈 예측 모델에 적용하였을 때 성능이 95%나 향상되는 것을 확인하였다.

6) 그 외 응용 사례

합성 데이터를 활용하면 대용량 데이터의 신속한 생산뿐만 아니라 데이터 불균형이나 개인정보 유출과 같은 심각한 문제를 해결할 수 있다는 장점이 있다. 위에서 열거한 사례들뿐만 아니라 결함 인식, 자동 공정 로봇 학습, 스마

트팜 작물 성장 및 발달 모니터링, 임상 시험 등 다양한 분야에 활발하게 적용되고 있다. 최근에는 Datagen, Stalioe, Synthesis AI와 같이 다양한 분야에서 합성 데이터 생성 솔루션을 제공하는 기업들이 글로벌 플랫폼 기업과 인수합병 단계를 거치고 있어, 앞으로 더 주목할 만하다.

[표 5-1-1] 그 밖의 합성 데이터 활용 사례

분야	활용 사례	업체명
식품	음식물 가공 로봇 학습	Soft Robotics
제조	제품 결함 인식	Replicator
의료	MRI 이미지 분석 및 진단	Mayo Clinic, MGH & BWH
제약	약물 개발을 위한 임상 시험 데이터 생성	Roche

5. 합성 데이터 시장

아래 표에 따르면 2018년 기준으로 세계 인공지능 학습 데이터 생성 시장 규모는 대략 7억 7천만 달러로 평가되며, 연평균 22.5%의 성장률로 2024년에는 26억 달러 규모로 성장할 것으로 전망된다. 또한, AIMultiple의 조사¹²⁾에 따르면 합성 데이터 분야 중 테스트 데이터 관리 시장은 연평균 12.7%로 성장하고 인공지능 학습 데이터 생성 시장은 연평균 성장률을 보일 것으로 예측된다. 이 두 서비스를 제공하는 합성 데이터 시장은 연평균 성장률이 10% 이상일 것으로 전망된다.¹³⁾

[표 5-1-2] 세계 인공지능 학습 데이터 생성 시장 규모

구분	'18	'19	'20	'21	'22	'23	'24	CAGR
세계시장	772.7	946.5	1,160	1,420	1,740	2,131	2,611	22.5

* 출처: Grand View Research, Inc, 2020.05.26.을 바탕으로 세계 AI시장과 국내 AI시장을 복합적으로 고려하여 네모아이씨지 재추정

중소벤처기업부의 조사에 따르면 국내 인공지능 학습 데이터 생성 시장 규모는 2018년 약 1,629억 원 규모에서 연평균 성장률 9.4%로 2024년에는 5,752억 원 규모로 성장할 것으로 전망된다.

12) 중소벤처기업부, 중소벤처기업부 전략기술로드맵 2021-2023, (2021), P.138.

13) 위의 보고서, P.138.

[표 5-1-3] 국내 인공지능 학습 데이터 생성 시장 규모

구분	'18	'19	'20	'21	'22	'23	'24	CAGR
국내시장	1,629	2,010	2,481	3,061	3,777	4,661	5,752	23.4

* 출처: Grand View Research, Inc, 2020.05.26., Revenues from the artificial intelligence software market worldwide from 2018 to 2025, 2020, Statista를 기반으로 네모아이씨지 재추정

6. 합성 데이터의 한계점

1) 데이터 왜곡

합성 데이터 생성 기술은 생성된 데이터가 단순하고 일련의 규칙이나 패턴으로 설명될 수 있을 때 가장 잘 작동한다. 반면, 자연어 텍스트나 이미지와 같은 복잡한 데이터를 생성하는 경우 모델의 한계로 인해 합성된 데이터가 실제 데이터와 다르게 왜곡될 수 있다. 또한, 이러한 왜곡은 다른 모델을 합성데이터로 학습시킬 때 실제 데이터에 대한 정확한 예측을 어렵게 만들 수 있다.

2) 합성 데이터 검증의 어려움

합성 데이터의 또 다른 한계는 검증이 어렵다는 점이다. 합성 데이터는 얼핏 보기에 실제 데이터와 상당히 유사해 보일 수 있지만 실제 데이터의 패턴을 정확하게 포착하여 생성되었는지는 확인하기 어렵다. 따라서 합성 데이터로 훈련된 모델이 실생활에 적용될 때 높은 성능을 보일 것이라고 보장할 수 없다.

3) 실제 데이터에 대한 의존성

합성 데이터도 결국 실제 데이터를 기반으로 생성되기 때문에, 기존 데이터셋이 가지는 통계에 편향된 데이터를 생성하게 된다. 이와 같이 편향된 데이터로 모델을 학습시키게 되면, 모델은 해당 데이터와 유사한 패턴을 강조하여 공정하지 못한 결정을 내릴 수 있으며, 결과적으로 모델 성능을 저하시킬 수 있다.

제2장 클라우드 스토리지 기술 동향

김재락 매니저 네이버클라우드 IaaS Product Planning

클라우드 컴퓨팅 기술이 발전하고 데이터가 폭발적으로 증가한 덕분에 클라우드 스토리지 기술은 현대 비즈니스 분야에서 핵심적인 역할을 수행하고 있다. 기업은 대용량 데이터의 저장과 공유, 높은 성능과 가용성 등 다양한 요구사항을 충족해 주는 클라우드 스토리지 옵션을 선택하고 구성하려고 한다. 경쟁 우위를 확보하려는 목적 때문이다. 이를 위해 클라우드 스토리지의 다양한 옵션과 각각의 특징, 응용 분야, 가격 모델, 기술 동향 등에 대한 종합적인 분석이 필수적이다.

1. 다양한 클라우드 스토리지 옵션과 특징

현재 클라우드에서는 다양한 스토리지 옵션은 제공한다. 각 옵션은 고유한 특징과 장점을 지닌다.

[표 5-2-1] 클라우드 스토리지 옵션과 특징

구분	주요 특징	애플리케이션 유형
Block Storage	주로 가상 머신(VM) 또는 컨테이너와 같은 데이터를 블록 단위로 읽고 쓰는 용도에 사용한다. 일반적으로 높은 성능과 낮은 지연 시간을 제공하며, 주로 데이터베이스, OLTP (On-Line Transaction Processing) 등의 응용 프로그램에 적합하다.	데이터베이스, OLTP(On-Line Transaction Processing) 시스템 사용예시: 온라인 상거래 플랫폼, 은행 시스템, ERP(Enterprise Resource Planning) 시스템 등
NAS(Network-Attached Storage)	파일 공유를 위한 스토리지로, 네트워크를 통해 파일을 공유하고 액세스하는 용도에 사용한다. 여러 클라이언트가 동시에 파일에 접근할 수 있으며 주로 파일 공유, 백업, 아카이브 등에 사용한다.	파일 공유, 백업, 데이터 공유 사용예시: 파일 서버, 협업 도구, 미디어 공유 플랫폼 등
Object Storage	대용량의 비정형 데이터를 저장하고 검색하는 용도에 사용한다. 파일마다 고유한 식별자를 가지고 있어 데이터의 무결성과 안정성을 보장한다. 주로 이미지, 비디오, 문서 등의 대용량 데이터를 저장하고, 웹 및 모바일 애플리케이션에서 사용한다.	대용량 비정형 데이터 저장, 웹 애플리케이션 사용예시: 이미지 및 비디오 호스팅, 대용량 파일 공유, 데이터 아카이브 등
Archive Storage	오랜 기간 동안 보존해야 하는 데이터를 위한 스토리지로, 주로 백업 및 장기 보관을 위해 사용한다. 저장 공간을 저렴하게 제공하며, 데이터에 대한 액세스 속도는 상대적으로 느리다. 주로 컴플라이언스 및 규정 준수 요구 사항을 충족시키기 위해 사용한다.	장기 데이터 보존 및 백업 사용예시: 장기 보관 데이터, 규정 준수 요구 사항에 따른 데이터 보존 등

(계속 →)

Server Local Storage	서버 자체에 내장된 스토리지로, 서버에서 직접 액세스할 수 있는 스토리지이다. 주로 데이터 처리가 빠르면서 저렴한 가격을 유지해야 하는 응용 프로그램에 사용한다. 서버의 로컬 디스크 또는 NVMe 드라이브 등이 속한다.	빠른 데이터 처리, 저렴한 가격 사용예시: 로컬 데이터베이스, 캐시 서버, 임시 작업 공간 등
대용량 데이터 마이그레이션(Migration)용 스토리지	대량의 데이터를 클라우드로 이전하는 용도로 사용한다. 네트워크 대역폭의 한계를 극복하기 위해 이전 전략을 사용하여 데이터를 효율적으로 옮긴다. 주로 대규모 데이터 마이그레이션 프로젝트나 초기 데이터 이전에 활용한다.	대규모 데이터 마이그레이션, 초기 데이터 이전 사용예시: 대규모 데이터 이전 프로젝트, 대용량 데이터 백업 및 복구 등

2. 클라우드 스토리지의 가격요소 및 가격모델

클라우드 스토리지의 가격 모델은 서비스 제공업체마다 약간 다를 수 있지만, 일반적으로 다음과 같은 요소들을 고려하여 책정한다.

[표 5-2-2] 클라우드 스토리지의 가격요소 및 가격모델

가격요소	가격모델
용량	사용량 또는 할당량 기반으로 요금을 책정한다. 보통 기가바이트(GB), 테라바이트(TB) 등의 단위로 용량을 측정한다.
데이터 전송량	데이터를 저장하거나 데이터에 액세스하는 데 필요한 네트워크 전송량에 따라 요금을 책정한다. 데이터의 입출력(IO) 작업량과 전송되는 데이터양에 따라 요금을 책정한다.
데이터 수명주기	일부 클라우드 서비스 제공업체는 데이터의 수명주기에 따라 가격을 다르게 책정할 수 있다. 예를 들어, 일반적인 데이터와 비교하여 아카이브용 데이터의 저장 비용이 저렴한 경우가 있다.
데이터 보관 기간	데이터 저장기간에 따라 가격이 달라질 수 있다. 일부 클라우드 제공업체는 장기적인 데이터 보관에 대해 저렴한 가격 모델을 제공한다.
서비스 수준 계약(SLA)	서비스 수준에 따라 가격이 결정될 수 있다. 더 높은 가용성, 데이터 내구성, 백업 및 복원 등의 기능을 제공하는 서비스는 일반적으로 더 높은 가격을 책정한다.
지역 및 리전	클라우드 서비스를 이용하는 지역과 리전에 따라 가격이 다를 수 있다. 서비스 제공업체는 지역별로 가격을 다르게 책정할 수 있으며, 일부 지역에서는 더 저렴한 가격을 제공할 수도 있다.

3. 클라우드 스토리지의 기술 동향

클라우드 환경을 효율적으로 이용하도록 데이터 관리, 보안, 성능, 확장성 관련 기술들이 발전하고 있다.

가. 객체 스토리지의 증가: 객체 스토리지는 파일을 조각내어 분산하고 저장하는 방식으로 데이터를 처리하는 기술이다. 이는 대용량 데이터의 저장과 처리에 효율적이며, 클라우드 스토리지에서 많이 사용한다. 비정형 데이터가 급격하게 증가하는 추세라, 객체 스토리지의 수요는 계속 증가할 것으로 보인다.

나. 데이터 보안 강화: 클라우드 스토리지에서 데이터 보안은 매우 중요한 요소다. 많은 기업이 데이터 보호를 강화하기 위해 암호화, 접근 제어, 보안 인증 등 다양한 보안 기술을 도입하고 있으며 클라우드 업체는 보안기술을 적극적으로 제공하고 있다.

다. 데이터 수명: 데이터의 생성, 저장, 삭제를 전반적으로 관리하는 Data Life cycle 관리에 대한 수요는 지속 증가할 것으로 보인다. 또한 Intelligent Tiering 기술 또한 클라우드 기술 동향 중 하나다. 이는 데이터의 액세스 패턴에 따라 자주 액세스 되는 데이터는 성능이 높은 스토리지에 저장하고, 자주 액세스 되지 않는 데이터는 성능이 낮은 스토리지에 저장하며, 액세스 패턴에 맞춰 주기적으로 데이터의 저장소를 변경해 주는 기술이다.

라. 성능: 스토리지의 성능을 측정하는 단위로는 우선 초당 입력/출력 작업 횟수를 의미하는 IOPS가 있고, 단위 시간당 처리할 수 있는 데이터량을 의미하는 Throughput도 있다. 메모리 기반 등 디스크 매체의 기술이 발전하면서 특히 IOPS의 성능이 지속적으로 개선되고 있다. 그 덕분에 클라우드 업체는 더욱 고성능의 스토리지 기술을 제공할 수 있게 되었다.

마. 백업/재해복구: 클라우드 스토리지를 활용한 백업 및 재해복구 기술은 데이터를 안전하게 보호하고, 비즈니스의 연속성을 유지해 준다는 점에서 중요하다. 클라우드상에서 파일 시스템 또는 DBMS를 백업하거나, 저장된 데이터를 지리적으로 떨어진 국가 또는 데이터센터에 저장하는 기술 덕분에 고객은 데이터의 안정성을 강화할 수 있다.

바. 인공지능(AI) 활용: 클라우드 스토리지에 저장된 데이터를 분석하고, 인공지능 알고리즘과 기계학습 기술 등에 접목하여 데이터의 패턴, 동향, 예측 등 데이터 관리의 효율성을 높이고 데이터의 가치를 극대화할 수 있다.

제3장

데이터 분석
기술 동향

박종건 매니저 네이버클라우드 Data Platform Planning

기업에서는 데이터의 활용 가치를 높이기 위하여 조직 내 데이터 분석 인력을 많이 두고, 도구에 유연하면서 데이터를 효율적으로 제공할 수 있는 환경을 만들고자 노력하고 있다. 현대적 데이터 아키텍처를 통해서 데이터를 찾고, 다양한 소스의 데이터를 조정 및 통합하며, 적절한 거버넌스를 적용하고, 조직 전체에서 사용할 수 있는 데이터 프로덕트를 생성하려는 것이다. 결과적으로 비용과 성능 면에서 최상의 조합을 이루고 비즈니스 민첩성과 유연성을 높일 수 있다.

데이터는 기업의 자산이자 전략적 리소스다. 기업은 데이터의 세부 정보로 고객 행동을 통찰하고 시장변동을 예측하여, 비즈니스 모델을 혁신하고 비즈니스 의사 결정을 내리기도 한다. 또한 데이터의 효과적 활용으로 상품 및 서비스의 개선, 효과적 마케팅 전략 수립 등으로 운영비용을 절감한다. 이를 통해 기업의 경쟁력을 강화하려고 한다.

그러나 기업 데이터의 잠재력을 제대로 활용하지 않아, 그 가치가 충분히 드러나지 못하는 경우가 많다. 현재 데이터 볼륨은 증가하고 다양한 소스에서 유입하기에, 다양한 데이터 스키마의 생산으로 복잡한 혼합 워크로드가 생성된다. 간단한 보고서에서 경영진의 주요 업무까지 지원하기 위해 정형, 반정형, 비정형 데이터를 처리해야 한다. 하지만 기존 온프레미스 환경에서 이루어지는 데이터 분석 접근 방식으로는, 증가하는 데이터 볼륨을 감당할 수준으로 확장하는 것이 불가능하며, 데이터 사일로 환경에 놓여 데이터의 이동이 자유롭지 않기 때문이다. 방대한 양의 데이터 분석 워크로드를 처리하려면 높은 수준의 성능과 비용이 필수적이다. 클라우드 환경에서는 단일 클라우드 공급자의 에코시스템에 종속되어 유연한 데이터 분석환경을 갖추기 어렵다¹⁾

기업에서는 데이터의 활용 가치를 높이기 위하여 조직 내 많은 데이터 분석 인력과 도구에, 유연하면서 데이터를 효율적으로 제공할 수 있는 환경을 만들고자 노력하고 있다. 누적된 데이터 처리 비용에 대한 현안을 해결하려면 하이브리드 또는 멀티클라우드 등의 퍼블릭 클라우드 기반 환경으로 전환하거나, 여러 사일로 환경에 있는 고품질 데이터를 쉽게 검색하고 접근할 수 있는 데이터 플랫폼 구축 등의 방안을 고려해야 한다. 이때 둘 중 하나를 선택해야 하는 경우가 자주 생긴다. 하나는 사일로 간에 데이터를 이동 및 중복시켜 다양한 분석 사용 사례를 지원하는 방안이고, 다른 하나는 데이터를 분산된 채로 두어 성능을 제약함으로써 민첩성 저하를 감내하는 방안이다. 기업에서는 이를 고려하면서 비용과 성능 면에서 최상의 조합을 제공하는 아키텍처를 선택해야 한다. 현대적 데이터 아키텍처를 통해 민첩성과 유연성을 높이고 비용과 보안 면에서 이상적인 균형을 이룬다면, 더욱 전략적인 IT 인시셔티브에 집중할 수 있다.

1) MIT Technology Review, "Modern data architectures fuel innovation", 2023.1.5., 2023년 8월 27일 접속,
<https://www.technologyreview.com/2023/01/05/1066239/modern-data-architectures-fuel-innovation/>

1. 현대적 데이터 아키텍처란²⁾

데이터 아키텍처는 수집에서 변환, 배포 및 소비에 이르기까지 데이터를 관리할 수 있는 구조로 되어 있다. 우수한 데이터 아키텍처는 데이터를 관리하고 유용하게 사용할 수 있도록 데이터 라이프사이클 관리를 지원한다. 특히 중복 데이터 스토리지를 방지하고, 정리 및 중복제거를 통해 데이터 품질을 개선하며, 새로운 애플리케이션을 지원할 수 있다.

- ① **중복성 감소**: 서로 다른 소스에 걸쳐 중복된 데이터 필드가 있을 수 있으므로 데이터 불일치, 데이터 부정확성의 요인으로 데이터 통합을 어렵게 할 수 있다. 우수한 데이터 아키텍처는 데이터 저장 방법을 표준화하고 중복을 줄여 더 나은 품질과 전체적인 분석을 가능하게 한다.
- ② **데이터 품질 향상**: 데이터 아키텍처가 잘 설계되어 있으면 흔히 "데이터 늪"이라고도 불리는, 관리가 제대로 되지 않는 데이터 호수의 문제를 해결할 수 있다. 데이터 아키텍처는 데이터 거버넌스 및 데이터 보안 표준을 적용하여, 의도한 대로 데이터 파이프라인을 작동할 수 있도록 지원한다. 데이터 아키텍처는 데이터 품질과 거버넌스를 개선하여 데이터가 현재와 미래에 유용하게 저장되도록 보장할 수 있다.
- ③ **통합 지원**: 기업 내에서 데이터 스토리지와 조직의 장벽에 대한 기술적 한계에 부딪혀 데이터가 종종 사일로화 된다. 오늘날 데이터 아키텍처는 도메인 간에 데이터통합을 촉진해야 한다. 이를 통해 다른 지역 혹은 다른 비즈니스 기능 간에 서로의 데이터에 액세스할 수 있도록 하는 것을 목표로 해야 한다. 이를 실현한다면 데이터 프로덕트들을 종합적으로 파악하여 의사 결정에 더 나은 정보를 제공할 수 있다.
- ④ **데이터 수명 주기 관리**: 현대의 데이터 아키텍처는 시간이 지남에 따라 데이터가 관리되는 방식을 해결할 수 있다. 일반적으로 데이터는 오래되고 액세스 빈도가 낮아질수록 유용성이 떨어진다. 시간이 지남에 따라 데이터를 더 저렴하고 느린 스토리지 유형으로 마이그레이션함으로써 고성능 스토리지 비용을 절감할 수 있다.

현대적 데이터 아키텍처는 모든 것을 한 곳에 저장할 때 발생하는 복잡한 사안 없이 비교적 간결하게 데이터 사일로를 해체한다. 이를 통해 부서 간, 지역 간, 또는 같은 도메인 간에 데이터를 통합할 수 있는 메커니즘을 제공한다. 최근의 현대적 데이터 아키텍처는 클라우드 플랫폼을 활용하여 데이터를 관리하고 처리한다. 비용이 더 많이 들 수 있지만 컴퓨팅 확장성을 통해 중요한 데이터 처리 작업을 신속하게 완료할 수 있다. 또한 스토리지 확장성을 통해 증가하는 데이터 볼륨에 대처하고, 모든 관련 데이터를 사용하여 데이터의 활용 품질을 향상하도록 지원한다. 기업은 현대적 데이터 아키텍처를 통해 데이터를 찾고, 다양한 소스의 데이터를 조정하고 통합해야 한다. 또한 적절한 거버넌스를 적용하고, 조직 전체에서 사용할 데이터 프로덕트를 생성할 수 있어야 한다. 데이터 엔지니어링 작업 역시 자동화하여 복잡성을 줄이고, 비용을 최소화하고, 비즈니스 가치를 최적화할 수 있어야 한다.

2) "What is a data architecture?," IBM, 2023년 8월 27일 접속, <https://www.ibm.com/topics/data-architecture>

2. 현대적 데이터 아키텍처의 주요 특징과 구축 준비

현대적 데이터 아키텍처의 특징은 다음과 같다.

- ① **클라우드 네이티브 및 클라우드 지원**: 데이터 아키텍처가 클라우드의 탄력적인 확장 및 고가용성을 활용할 수 있도록 지원한다.
- ② 분리 및 확장이 가능하여 서비스와 개방형 표준 간의 의존성이 없으므로 상호 운용성이 가능하다.
- ③ 하나의 프레임워크로 지능적 워크플로를 실현하고, 상황에 대한 분석 및 실시간 통합을 가능하게 하는 강력하고 확장성이 뛰어난 데이터 파이프라인이다.
- ④ 유효성 검사를 포함한 실시간 데이터 활용, 분류, 관리, 거버넌스 환경을 지원한다.
- ⑤ 표준 API 인터페이스를 사용하여 기존 애플리케이션과의 끊임없는 데이터 통합이 가능하다.
- ⑥ 비용과 단순성의 균형을 유지하도록 최적화했다.

그리고 현대적 데이터 아키텍처를 구축하기 위해 필요한 요건은 다음과 같다.³⁾

- ① 부서 혹은 지역, 물리적/가상적 요인으로 인해 분산된 도메인과 데이터 사일로의 통합
- ② 하이브리드/멀티클라우드 플랫폼을 사용하여 데이터 관리 및 처리
- ③ 데이터 관리가 용이하도록 유연성과 확장성을 고려하여 구축
- ④ 컴퓨팅 및 스토리지 확장성을 통해 증가하는 데이터 볼륨 처리
- ⑤ 데이터 공급자와 데이터 소비자 간의 가치 사슬에서 데이터 통합, 데이터 엔지니어링 및 거버넌스 자동화
- ⑥ 전 단계에 걸친 보안 환경

3. 현대적 데이터 아키텍처 구축

현대적 데이터 아키텍처를 효과적으로 구축하려면 데이터 관리시스템의 유형과 데이터 경계, 데이터 아키텍처 유형을 이해해야 한다.

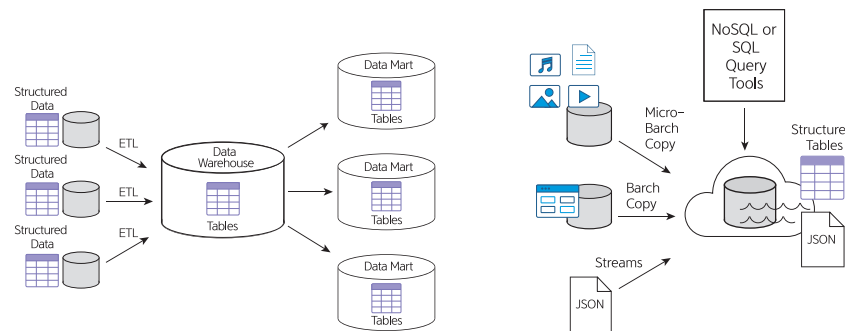
가. 데이터 관리시스템 유형

3) "Build a modern data architecture," IBM, 2023년 8월 27일 접속, <https://www.ibm.com/resources/the-data-differentiator/data-architecture>

데이터 관리 시스템 유형에는 데이터 웨어하우스, 데이터 마트, 데이터 레이크가 있으며, 이는 데이터 스토리지 저장소를 뜻하기도 한다. 기업의 방대한 데이터는 이와 같은 형태의 저장소에 담겨 엔터프라이즈 전체에 분산되어 있으며, 데이터는 여러 방법을 통해 다양한 형식으로 저장되고 검색된다.

- ① **데이터 웨어하우스**: 데이터 웨어하우스는 기업 전체에 걸쳐 서로 다른 관계형 데이터 소스의 데이터를 일관된 단일 중앙 저장소로 집계한다. 추출 후 데이터는 ETL 데이터 파이프라인을 통해 흐르며, 사전 정의된 데이터 모델을 충족하기 위해 다양한 데이터 변환을 거친다. 일단 데이터 웨어하우스에 로드되면 데이터가 활성화되어 서로 다른 비즈니스 인텔리전스(BI) 및 애플리케이션을 지원한다.
- ② **데이터 마트**: 데이터 마트는 데이터 웨어하우스의 중점 버전으로, 단일 팀 또는 HR 부서와 같은 조직 내의 특정 사용자 그룹에 중요하고 필요한 데이터의 하위 집합을 포함한다. 데이터 마트는 데이터의 하위 집합이 더 적기 때문에 부서나 비즈니스 라인에서 광범위한 데이터 웨어하우스 데이터 세트와 작업할 때, 좀 더 신속하고 집중적인 인사이트를 발견할 수 있다.
- ③ **데이터 레이크**: 데이터 웨어하우스는 처리된 데이터를 저장하는 반면, 데이터 레이크는 일반적으로 페타바이트 단위의 원시 데이터를 저장한다. 데이터 레이크는 정형 데이터와 비정형 데이터를 모두 저장할 수 있으므로 다른 데이터 저장소와 차별화된다. 스토리지 요구사항의 이러한 유연성은 데이터 과학자, 데이터 엔지니어 및 개발자에게 특히 유용하며, 데이터 검색 연습 및 머신 러닝 프로젝트를 위한 데이터에 액세스할 수 있다. 데이터 레이크는 데이터 웨어하우스가 증가하는 데이터의 양, 속도, 다양성을 지닌 데이터 처리의 한계로 생성되었다. 데이터 레이크는 데이터 웨어하우스보다 느리지만 수집 전 데이터 준비가 거의 또는 전혀 없기 때문에 비용도 저렴하다. 데이터 레이크는 데이터 수집 시 데이터에 대한 비즈니스 목표를 정의할 필요가 없기 때문에 광범위한 사용 사례를 지원한다. 정형 데이터와 비정형 데이터를 동일한 위치에 저장할 수 있는 기능의 이점을 제공한다. 애플리케이션이 개발되고 유용한 데이터가 식별되면 데이터를 데이터 웨어하우스로 내보내 운영에 사용할 수 있으며, 자동화를 사용하여 애플리케이션을 확장할 수 있다. 데이터 레이크는 저렴한 비용으로 확장할 수 있기 때문에 데이터 백업 및 복구에도 사용할 수 있다.

[그림 5-3-1] 데이터 관리시스템 유형: 데이터웨어하우스, 데이터마트, 데이터레이크



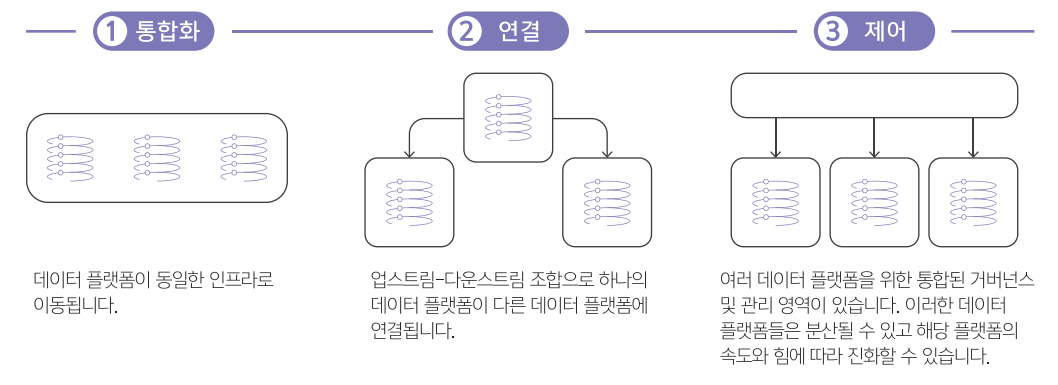
* 출처: Michael Nixon, "Data Warehouses, Data Lakes, and Data Lakehouses: Everything You Need to Know," snapLogic, 2023.4.20, 2023년 8월 27일 접속.
<https://www.snaplogic.com/blog/data-warehouses-data-lakes-data-lakehouses-everything-you-need-to-know>

나. 데이터 경계

두 번째는 데이터 경계에 대한 이해다. 데이터 실무자들은 빅데이터를 통해 데이터의 활용도를 높여 수익을 낼 기회는 많이 증가하였지만, 규모에 맞게 적합한 애플리케이션에 최적의 데이터를 제공하는 과제 또한 증가했다. 기업들은 다양한 데이터저장소에 위치한 데이터를 통합하기 위해 데이터 플랫폼을 구축하고 있다. 데이터 플랫폼은 일반적으로 각자의 보안 통제 범위 안에서 데이터를 활용하려는 사람들이 데이터에 액세스할 수 있게 하거나, 데이터를 사용자, 애플리케이션 또는 기타 기술에 제공하는 환경을 제공한다. 각 데이터 플랫폼의 경계는 일반적으로 저장되는 데이터의 유형 또는 데이터가 사용되는 방식에 따라 정해진다. 예를 들어 기업 내 재무 데이터 플랫폼, 인사 데이터 플랫폼, 사업부별 데이터 플랫폼이 필요할 수 있다. 이러한 데이터 플랫폼 간의 경계 내에서 각 부서는 데이터 마트와 같은 형태를 통해 집중적인 인사이트를 편리하게 얻고자 하였다. 그러나 특정 데이터 플랫폼에 국한된 인사이트만으로는 변화하는 비즈니스 상황에 적합한 새로운 비즈니스 인사이트 얻는 것에 한계가 있다. 특히 통합적 시각으로 필요한 데이터를 찾을 경우에는 각각의 단일 데이터 플랫폼으로부터 적절한 데이터를 찾아내기가 어려워졌다. 그런데 데이터 분석 기능이 향상되어, 데이터 플랫폼 설계자가 생각할 수 없었던 연결 관계와 인사이트를 여러 데이터 플랫폼 사이에서 찾아낼 가능성이 생겼다.

데이터 플랫폼 간의 경계를 관리하는 방법에는 통합, 연결, 제어의 방법이 있는데, 최근에는 '제어' 관리 방법을 통해 점점 경계를 허물고 있다. 즉, 분산된 여러 데이터 플랫폼을 통합한 거버넌스 환경과 데이터 액세스 환경을 제공하고 있다. 이는 분석 이니셔티브를 가속화하는 효과를 얻고 있다.

[그림 5-3-2] 데이터 경계 관리방법



* 출처: Varun Bijani, Sandipan Sarkar, Richard Warrick, "Weaving data fabric into hybrid multcloud," IBM, 2021.4.16, 2023년 8월 27일 접속.
https://www.ibm.com/thought-leadership/institute-business-value/report/data-fabric-multicloud?mhsrc=ibmsearch_a&mhq=Weaving%20data%20fabric%20into%20hybrid%20multicloud

다. 데이터 아키텍처 유형⁴⁾

마지막으로 데이터 아키텍처 유형에는 데이터 패브릭과 데이터 메시가 있다. 두 가지 유형은 기업의 데이터를 이해, 관리 및 사용하는 조직의 변화와 복잡성을 해결하기 위한 새로운 데이터 관리 개념이다. 이를 통해 사용 사례 중심의 설계를 따르고 데이터의 무질서한 증가, 데이터 거버넌스 및 데이터 가용성 문제를 해결하고자 한다. 데이터 패브릭 및 데이터 메시 접근 방식도 지속적인 데이터 검색 및 셀프서비스 데이터 지식 카탈로그에 의존하며, 두 개의 유형은 상호보완적이다.

1) 데이터 패브릭⁵⁾

현대적인 데이터 인프라를 구축하는 가장 효과적인 방법으로 데이터 패브릭이라는 새로운 유형의 아키텍처를 들 수 있는데, 이는 분산 환경 전반에 걸쳐 여러 소스의 데이터를 통합하고 조정한다. 데이터 패브릭은 데이터와 연결 프로세스의 통합 계층을 일컫는 설계 개념이다. 일반적으로 다양한 데이터와 데이터 프로세스를 구조 및 위치에 구애받지 않고 연결하는 역할을 한다.

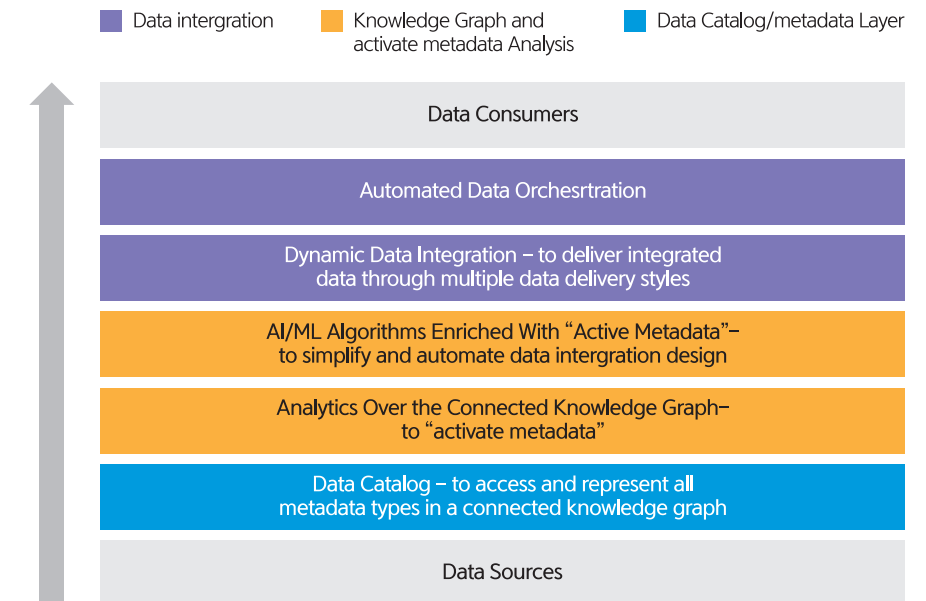
데이터 패브릭 아키텍처는 기존 인프라의 데이터를 통합하고, 의미 있는 인사이트를 사전에 얻을 수 있도록 데이터를 준비하는 운영 기술 계층을 갖추고 있다. 데이터 통합 기술을 기반으로 여러 출처의 실시간 데이터를 특정 위치로 식별하고 데이터 관리를 체계화하는 단일 통합 플랫폼을 제공한다. 메타데이터를 사용하여 중앙집중식 데이터 거버넌스와 데이터 보안 환경을 제공함으로써 다양한 데이터의 엔드포인트에 모두 같은 수준의 관리 및 보안 정책을 적용할 수 있다.

이를 통해 사용자는 셀프서비스 방식으로 조직 전체의 데이터에 액세스하고 사용할 수 있다. 그리고 조직은 데이터 거버넌스, 데이터 개인정보 보호 및 보안, 데이터 엔지니어링 작업 및 데이터 통합을 자동화할 수 있다. Gartner에 따르면 데이터 패브릭은 설계, 구축 및 운영을 포함한 데이터 관리 작업을 최대 70%까지 줄일 수 있다. 또한, 데이터 패브릭을 구축하면 데이터 활용 효율성이 증가하고 사람이 주도하는 데이터 관리 작업이 줄어들 것으로 예측한다.

4) Yash Mehta, IDG Connect, "메시 vs. 패브릭 : 대표 데이터 아키텍처 이해하기" IT World, 2022년 7월 15일, <https://www.itworld.co.kr/news/244933#csidx36c68b5b3416f38b70ca54c39d87bde>

5) "데이터 패브릭이란 무엇입니까?," TIBICO, 2023년 8월 25일 접속, <https://www.tibco.com/ko/reference-center/what-is-data-fabric>

[그림 5-3-3] 데이터 패브릭 계층



* 출처: "Data Fabric Architecture is Key to Modernizing Data Management and Integration," Gartner, 2021.5.11, 2023년 8월 27일 접속, <https://www.gartner.com/smarterwithgartner/data-fabric-architecture-is-key-to-modernizing-data-management-and-integration>

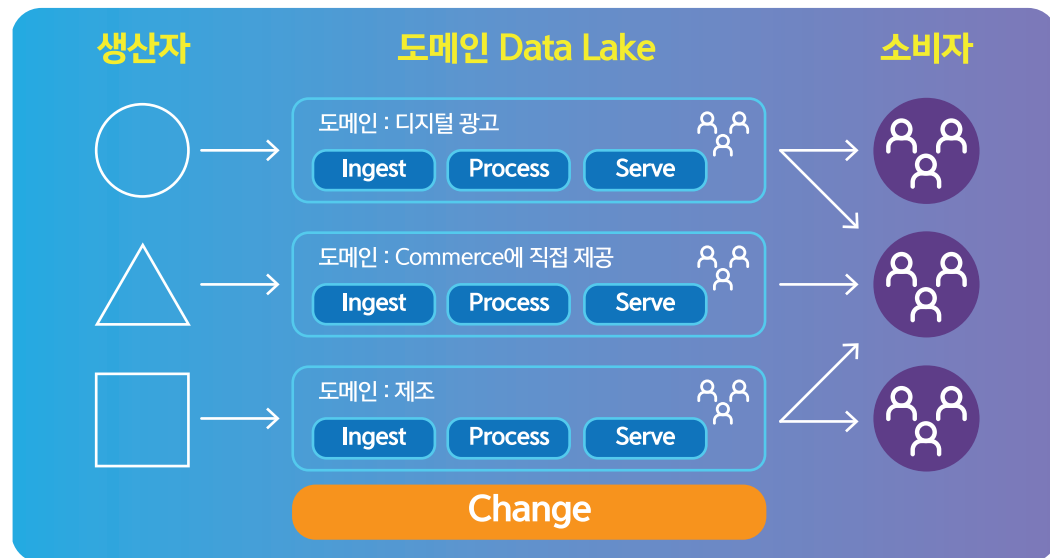
2) 데이터 메시

Forrester에 따르면, 데이터 메시는 복잡한 대규모 환경에서 분석 데이터를 공유, 액세스 및 관리하기 위해 도메인 단위로 분산시킨 접근 방식이다. 데이터를 분산시켜 비즈니스 도메인 또는 기능별로 데이터 소유권을 행사하도록 하여, 궁극적으로 데이터 소유권을 분산하는 것이다. 이를 통해 여러 도메인에서 발생하는 데이터 간의 일관성을 확립하고, 데이터 소유자는 각 도메인의 데이터 프로젝트를 만들 수 있다.

데이터 메시 아키텍처는 기존의 인프라를 확장하는 방식으로, 여러 부서에 새 아키텍처가 배치되며 직원들은 이 아키텍처를 기반으로 데이터를 더 효과적으로 관리할 수 있다. 이러한 탈중앙화된 접근 방식 덕분에 사용자는 소스 포인트에 데이터가 통합될 때까지 기다릴 필요 없이 바로 데이터에 액세스하고 쿼리할 수 있다.

각 도메인이 자체적으로 데이터 파이프라인을 처리하는 분산 데이터 토폴로지 방식을 기반으로 하며, 데이터 파이프라인의 생성/취급에 필요한 소유권은 데이터 이해도가 높은 도메인 전문가에게 일임한다. 도메인 전문가는 다양하게 산재한 데이터의 접근성을 높이고, 데이터 통합에 필요한 표준 포맷을 수립한다. 이를 통해 이질적 데이터를 상호운용성에 맞게 확보하고 호환할 수 있다. 특정 비즈니스 목적을 달성하기 위한 데이터 프로젝트를 생성하고 공유할 수도 있다.

[그림 5-3-4] 데이터 메시 개념도



* 출처: "AWS Summit 2022 강연 정리 - Data Mesh," Devnoris Tech Blog, 2022.5.15., 2023년 8월 27일 접속, <https://sungchul-p.github.io/aws-summit-data-mesh>

데이터 패브릭 아키텍처와 데이터 메시 아키텍처는 혼용되어 사용될 수 있다. 두 가지 아키텍처 모두 공통된 목적으로 구축되었기 때문이다. 즉 두 아키텍처 모두, 흩어져 있는 비즈니스 데이터에 사용자가 직접 접근하여 가치 있으면서 실행할 수 있는 인사이트를 창출하도록 돕는 역할을 한다.

데이터 패브릭을 사용하여 데이터 메시지를 구현하는 방법에는 다음과 같은 방법이 있다.

데이터 소유자에게 데이터에 대한 카탈로그, 변환, 통합 거버넌스 정책 준수와 같은 데이터 프로덕트를 생산하는 기능을 제공한다. 또한 데이터 소유자와 데이터 소비자가 데이터 프로덕트를 카탈로그화하고, 데이터 프로덕트를 검색할 수 있으며, 데이터 프로덕트에 쿼리 또는 데이터 프로덕트를 활용한 시각화 등의 다양한 방식으로 데이터 프로덕트를 사용할 수 있도록 해준다. 또한 데이터 패브릭 메타데이터를 통한 인사이트를 활용하여 특정 생산이나 작업에 대해 학습을 거쳐 자동화할 수 있다. 일례로 데이터 프로덕트의 일부 패턴을 통한 프로세스 생산이 있다. 또한 일부 프로세스를 통한 데이터 프로덕트의 모니터링 작업도 이런 예시에 해당된다. 즉, 데이터 패브릭은 데이터 관리와 관련하여 데이터 프로덕트를 생성하고 데이터 프로덕트의 라이프사이클을 관리하는 데 필요한 많은 작업을 자동화한다. 이를 통해 데이터 메시지를 구현하여, 필요한 기능을 최대한 제공하는 형태로 혼용할 수 있다.

현대적 데이터 아키텍처 환경은 여러 테크 서비스 기업에서 다양한 형태로 제공하고 있다. AWS는 확장 가능한 데이터레이크 환경을 기반으로 통합 거버넌스 및 보안 적용 환경을 제공하여 도메인별 비즈니스 문제의 해결을 돕는다. 구글클라우드의 '빅쿼리 옴니' 서비스로, 멀티 클라우드 환경에서 효율적으로 데이터에 액세스하고 안전하게 데이터를 분석할 수 있는 환경을 제공한다. 이를 통해 클라우드 간 분석으로 비즈니스 인사이트를 확보할 수 있게 한다. 네이버클라우드의 분산된 데이터의 메타데이터를 통해 데이터를 통합관리 할 수 있는 환경을 기반으로 다수의 서비스를 제공한다. 이처럼 기업의 현대적 데이터 아키텍처 구성을 실현하려는 다각적인 시도를 통해, 이용자가 데이터에 쉽게 액세스하고 편리하게 비즈니스 인사이트를 확보할 수 있는 다수의 서비스를 제공함으로써 기업의 현대적 데이터 아키텍처 구성을 실현시키고자 한다.

아마도 데이터 관리 아키텍처를 현대화해야 하는 가장 중요한 이유로는 효과적인 머신러닝과 AI를 가능하게 하려는 것을 꼽을 수 있다. 데이터 패브릭이 구축되면 조직은 데이터 관리의 대부분을 자동화할 수 있을 뿐만 아니라, 상용 머신러닝 운영 솔루션을 사용하여 AI를 구현하고 자동으로 관리할 수 있다. 예를 들어, MLOps 제품은 모델을 모니터링하고 새로운 데이터로 학습이 필요할 때 알림을 제공할 수 있다. 재무 운영 모델은 클라우드에 대한 지출을 관리하는 데 유용하다. 현대의 아키텍처로 데이터를 더 잘 관리하고 분석할 수 있는 기반을 마련된 덕분에, AI는 기업이 미래를 더 잘 준비하는 데 도움이 될 것으로 보인다. 그리고 데이터 관리 아키텍처를 현대화하는 조직은 AI 및 기타 새로운 기술을 지속적으로 채택하여 경쟁 우위를 유지할 수 있는 좋은 위치에 설 것이다.

제4장 데이터 보안 기술 동향

박병민 팀장 (주)신시웨이

4차 산업혁명 시대에 들어서며 인공지능(AI), 빅데이터, 사물인터넷 등이 핵심 기술로 자리 잡고 있다. 이러한 산업혁명의 주요 기술들은 공통적으로 데이터를 기반으로 한다. 과거 데이터(Data)가 단순히 자료의 개념을 지녔던 반면, 현대의 데이터는 단순 자료의 개념을 넘어, 데이터를 가공·처리·분석된 정보의 개념으로 접근해야 한다. 수집되는 데이터의 종류와 양이 증가하면서 데이터를 활용한 기술과 서비스 또한 다양해지고 있다. 이러한 기술과 서비스의 발전 덕분에 기업과 정부, 개인 상호 간에 다양한 정보를 교환할 수 있다. 이를 통해 생산성이 높아지고 업무가 편리해지지만, 이에 따른 보안 위협 또한 지속적으로 발생하고 있다. 따라서 데이터의 기밀성과 무결성, 가용성을 보장하기 위해 무단 접근 방지, 데이터 암호화, 데이터 비식별화 등 데이터를 보호하기 위한 기술적·관리적 조치가 필요하다. 이에 본 장에서는 데이터 보안 시장 현황과 기술 동향을 소개한다.

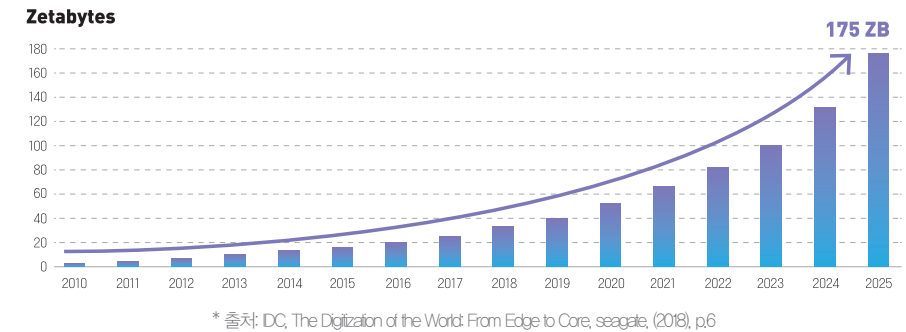
1. 데이터 보안 시장 동향

데이터 보안과 사이버 보안은 각종 디지털 정보와 자산을 보호한다는 관점에서는 공통점을 보인다. 하지만 디지털 영역에서 악의적인 공격이나 접근 등으로부터 시스템 및 디지털 자산을 보호하는 사이버 보안과 달리 데이터 보안은 데이터 자체의 무결성과 기밀성을 보장하는 데 중점을 둔다. 결국 두 보안 영역은 보호하고자 하는 대상이 서로 다르다. 데이터 보안은 사이버 보안 범주 안에 속하며, 사이버 보안에 필요한 기술적인 보안 조치들은 데이터를 보호하기 위한 기술적 조치를 포함하고 있다.

데이터 보안 시장은 데이터를 보호하기 위한 보안 솔루션들의 공급을 통해 시장이 형성된다. 그리고 지속적으로 발생하는 데이터의 생산과 각종 법적 규제, 컴플라이언스들은 데이터 보호를 위한 솔루션 공급과 기술 발전에 밀접히 연관되어 있다.

데이터 생산량이 증가하면 데이터의 활용 범위가 넓어지지만, 데이터를 활용한 새로운 서비스를 창출할수록 보안 위협도 늘기 마련이다. 이는 데이터 보안 시장의 규모 확대에 영향을 끼친다. 시장조사기관 IDC는 하루에 생산되는 데이터가 2016년 약 440억 기가바이트(GB)에서 2025년에는 10배가 넘는 4,630억GB로 늘어날 것으로 예측했다. 이 수치는 일일 단위로 생성되는 데이터량으로, 1년 단위로 계산하면 약 170제타바이트(ZB, 10의 21제곱)가 된다. 이는 2015년의 10제타바이트에 비해 17배 늘어난 수준이다.

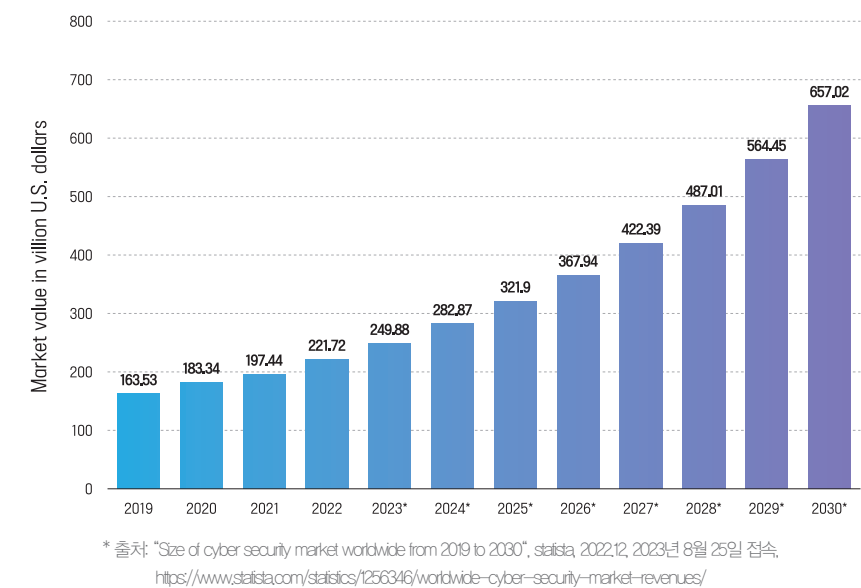
[그림 5-4-1] 전 세계 데이터 규모 추이



가. 글로벌 데이터 보안 시장 동향

글로벌 컨설팅그룹 맥킨지(McKinsey&Company) 조사에 따르면 디지털 경제가 성장함에 따라 사이버 공격도 함께 증가하고 있다. 맥킨지는 2025년 사이버 공격 피해액을 10조 5,000억 달러에 이를 것으로 본다. 이는 2015년 대비 300% 증가한 수치다. 이를 반영해 사이버 보안과 관련된 산업은 2021년 약 1,500억 달러의 가치가 있었으며 매년 12.4%씩 증가했다고 분석하고, 향후에는 보안 솔루션의 보급 상황에 따라 시장 규모가 1조 1,500억 달러에서 2조 달러에 이를 것이라 전망했다.¹⁾

[그림 5-4-2] 사이버 보안 시장 규모



¹⁾ Bharath Aiyer, Jeffrey Caso, Peter Russell, and Marc Sorel, "New survey reveals \$2 trillion market opportunity for cybersecurity technology and service providers," McKinsey & Company, 2022.10.27, 2023년 8월 27일 접속, <https://www.mckinsey.com/capabilities/risk-and-resilience/our-insights/cybersecurity/new-survey-reveals-2-trillion-dollar-market-opportunity-for-cybersecurity-technology-and-service-providers>

데이터 수집 분석 업체인 statista는 전 세계 사이버 보안 시장 규모를 2022년 약 2,222억 달러로 평가했다. 2030년에는 6,500억 달러를 넘어서며 2022년부터 2030년까지 연평균 12.83% 성장할 것으로도 보았다. 또한 전 세계 데이터 보안 시장은 2023년 59억 8,000만 달러였다가 2028년에는 107억 8,000만 달러에 달해 연간 12.50% 성장할 것으로 전망하였다.²⁾

나. 국내 데이터 보안 시장 동향

코로나 팬데믹에 따른 업무 환경 변화로 국내에서도 클라우드 시스템 도입 및 전환을 가속화하고 있다. 이에 데이터 보안은 클라우드 환경에서 더욱 중요한 보안 요소로 꼽힌다. 특히 데이터가 저장되는 운영체제와 데이터베이스의 데이터를 보호하기 위한 보안 솔루션의 매출이 증가하고 있다. 데이터를 보호하기 위한 규정도 국내에서 지속적으로 강화되면서 2021년 국내 정보 보안 제품 매출은 4조 5,497억 원이었다. 이는 전년 대비 16% 성장한 수치로, 2015년부터 2021년까지 연평균 13.7% 성장하였다. 그 중 데이터 보안 시장은 6,121억 원으로, 전체 매출의 19.6%를 차지하고 있다.³⁾

2. 데이터 보안을 위한 지침 및 인증

데이터의 안전한 보호를 위해 국제표준화기구 등 다양한 기관이 기술표준 가이드라인, 규정, 프레임워크 등을 마련하고 인증하는 제도를 운용하고 있다. 국가별로 별도의 법률을 제정하는데, 일반적으로 글로벌 표준보다 더 엄격하게 제정 및 개정하여 데이터를 실질적으로 보호할 수 있도록 하고 있다. 특히 의료 또는 금융 데이터처럼 특정 분야의 민감 정보에 대해 지속적인 법률 제정과 개정을 시행하고 있다.

이러한 표준화된 지침과 법률은 데이터 수명 주기 전반에 걸쳐 데이터가 안전하게 처리되도록 규정한다. 예를 들어 데이터 접근, 저장, 전송, 변경 및 삭제 등 데이터의 안전한 보호를 위해 생성부터 폐기에 이르기까지 전반적인 규칙이 담겨 있다.

가. ISO/IEC 27001

ISO/IEC 27001은 조직의 정보 자산을 안전하게 관리하기 위해 국제표준화기구(ISO)와 국제전기기술위원회(IEC)가 제정한 것으로, 정보보호 관리체계(SMS)에 대한 국제 표준이다. 데이터의 기밀성, 무결성, 가용성의 원칙을 기반으로 정보보호 정책, 규칙, 프로세스 등 정보 보안과 관련된 항목들을 다룬다. 지난 2022년 10월 26일 3차 개정 버전(ISO/IEC 27001:2022)을 발간하며 통제 항목을 조직 통제 37개, 인적 통제 8개, 물리적 통제 14개, 기술적 통제 34개로 재구성했다. 또한 데이터 마스킹과 데이터 유출 방지 등 7개의 기술적 통제 항목과 3개의 조직 통제 항목을 추가했다.

²⁾ "Data Security-Worldwide", statista, 2023.3., 2023년 8월 27일 접속.
<https://www.statista.com/outlook/tmo/cybersecurity/cyber-solutions/data-security/worldwide>

³⁾ 한국정보보호산업협회(KISIA), 2022 국내 정보보호산업 실태조사, (2022).

나. NIST SP 800-53(NIST Special Publication 800-53)

미국 상무부 산하 기관인 미국 국립표준기술연구소(NIST)에서 발간한 '정보 시스템 및 조직을 위한 보안과 개인 정보 보호'에 관한 문서이다. 기밀성, 무결성 및 가용성을 유지하기 위한 운영, 기술 표준 프레임워크로, 조직이 위험을 관리하고 미국 연방 정보보안 관리법(FISMA) 및 개인정보 보호 요건을 충족하는 데 필요한 기술적 조치를 포함한다. FISMA 규정에 따라 미국 연방 정부 기관 및 해당 기관과 협력하는 조직에는 의무적으로 적용되고 있으며 민간 기업에는 자발적 참여를 권고하고 있다.

다. 지불카드산업 데이터보안표준(PCI DSS, Payment Card Industry Data Security Standard)

VISA, MasterCard, American Express 등 5개의 신용카드사가 공동으로 설립한 PCI SSC(PCI Security Standards Council)에서 마련한 데이터 보안 규칙이다. 신용카드 거래를 처리하는 조직이 카드 사용자의 데이터를 보호하기 위해 마련한 인증 제도다. 데이터 암호화, 보안 네트워크 아키텍처, 접근 통제 등 사용자 데이터 보호에 필요한 12가지 요구사항을 포함하며, 연간 거래량에 따라 컴플라이언스 레벨을 달리한다.

라. 건강보험 이전 및 책임에 관한 법률(HIPAA, Health Insurance Portability and Accountability Act)

민감한 환자 건강 정보가 환자의 동의나 인지 없이 공개되지 않도록 보호하기 위해 국가 표준을 만들도록 요구하는 미국 연방법이다. 이 법에 따르면, 건강 정보를 전자적으로 전송하는 모든 의료 제공자가 환자를 식별할 수 있는 모든 정보에 대해 기밀성과 무결성, 가용성을 보장하고, 예측되는 보안 위협을 식별하고, 기술적 조치를 해야 한다. 주요 규칙으로는 개인정보 보호 규칙, 보안 규칙 및 위반 알림 규칙이 있다.

마. 일반 데이터보호규정(GDPR, General Data Protection Regulation)

개인 데이터의 처리 및 개인 데이터의 자유로운 이동과 보호에 관련된 규칙을 규정한 유럽연합의 포괄적인 데이터 보호 규정이다. 데이터 보호를 위한 기술 요구사항 및 권장 사항을 포함한다. GDPR은 전 세계에서 가장 강력한 데이터 보호 규정으로, 세계 표준으로 인정받고 있다. 국내 개인정보보호법이 GDPR의 요구 수준과 동등한 수준임을 인정하는 '개인정보 적정성 결정'이 2021년에 일본과 영국에 이어 3번째로 채택되었다.

3. 데이터 보안 기술 동향⁴⁾⁵⁾⁶⁾

효과적으로 데이터를 보호하기 위해 "아무도 신뢰할 수 없다"라고 가정하는 '제로 트러스트' 프레임워크가 기업과 조직을 중심으로 지속적으로 채택되고 있다. 제로 트러스트는 1994년 영국 스텔링대학교 스티븐 폴 마쉬(Stephen

⁴⁾ 김남일, 김창복, "상황인식 기반 적응적 접근제어 보안모델 설계에 관한 연구", 한국인터넷방송통신TV학회 제8권 5호(2008), pp.211-219

⁵⁾ 김원일, 하홍준, 이창훈, "RBAC을 위한 역할 정보 저장소의 설계(Design of Role Information Storage for Role-Based Access Control)", 제25회 한국정보처리학회 춘계학술발표대회 논문집 제13권 1호(2006.5).

⁶⁾ R. Sandu, D. Ferraiolo, R. Kuhn. The NIST Model for Role-Based Access Control: Towards A Unified Standard, NIST, (2000).

Paul Marsh)의 연구 논문을 통해 처음 알려졌다. 2020년 미국 국립표준기술연구소(NIST)에서는 제로 트러스트가 하나의 아키텍처가 아닌 워크플로 설계, 운영에 대한 가이드라인이라고 정의하며 구체적인 로드맵을 제시한 ‘제로 트러스트 아키텍처 기술서(SP 800-207)’를 발간했다.⁷⁾ 기술서에서는 제로 트러스트를 ‘데이터 유출 사고를 방지하고 내부 이동을 제한하도록 설계된’ 기업의 사이버 보안 아키텍처로 보면서, 구체적으로는 접근하고자 하는 주체에게만 접근을 허용하고 지속적인 인증에 따르는 보안 대책을 강구하는 방식을 규정했다. 즉, 제로 트러스트란 네트워크를 포함한 인프라 환경이 침해된 상황에서 정보 시스템 및 서비스에 대한 접근 요청을 정확하게 판단하려고 할 때 불확실성을 최소화하기 위한 개념과 발상의 모음이다. 제로 트러스트 아키텍처란 제로 트러스트 개념을 사용하는 사이버 보안 계획으로, 프로그램 모듈과의 관계, 워크플로 설계, 접근 정책을 포함한다. 이러한 정의에는 접근 통제를 세밀하게 하여 데이터 및 서비스에 대한 비인가된 접근을 방지하려는 목적이 담겨 있다.

데이터를 보호하기 위한 기술에는 데이터 암호화, 데이터 비식별화, 방화벽 및 침입 탐지, 접근 관리, 모니터링 등이 있다. 그 중에 가장 효과적이고 기본이 되는 보안 기술은 앞서 말했듯이 제로 트러스트 모델을 기반으로 하는 접근 통제 기술이다.

접근 통제는 ‘사용자, 프로그램 또는 프로세스에 부여된 액세스 권한 또는 이러한 권한을 부여하는 행위⁸⁾’이다. 국내에서는 접근 통제 시스템이 활성화된 시기인 2000년대 초중반부터 국내외 주요 컴플라이언스와 기업 및 조직의 요구 등에 따라 본격적인 데이터 접근 관리의 개념과 인식이 자리 잡기 시작했다.

기업 및 조직에서 사용하는 데이터 대부분은 데이터베이스와 운영체제 시스템 안에 저장된다. 그렇기 때문에 데이터 보안이라고 하면 가장 먼저 떠오르는 게 데이터베이스 보안과 서버 보안이다. 오늘날 우리가 흔히 말하는 데이터베이스 대부분은 관계형 데이터 모델을 기반으로 하는 관계형 데이터베이스(RDBMS)를 말하며, 접근 통제 기술로 인증과 권한 부여를 해 데이터를 보호한다.

접근 통제와 함께 적용되는 기술은 데이터 암호화다. 접근 통제 기술로 외부 침입을 막는 방화벽 역할을 실행한다면, 암호화 기술로는 금고의 역할을 실현하며 데이터를 확인할 수 없도록 높은 기밀성을 제공한다. 현재 데이터 암호화에 가장 많이 사용되는 암호 기법은 블록 암호화 기술이다. 일반적으로 데이터를 빠르게 처리할 수 있어 현재까지도 널리 사용하지만, 대량 또는 대용량의 데이터에는 적합하지 않다.

데이터 보안에서 가장 기본이자 핵심인 접근제어(Access Control)는 과거 물리적 제어부터 오늘날 정보기술로서 접근제어에 이르기까지 발전을 거듭했으며, 지금까지도 다양한 접근제어 모델과 기술들이 연구되고 있다. 접근제어 기술을 적용하면 식별(Identification)과 인증(Authentication), 인가(Authorization)를 거쳐 특정 주체(Subject)와 객체(Object)에 대해 접근 권한을 부여한다. 즉, 데이터에 접근하는 주체인 사용자와 시스템에 저장된 파일 또는 데이터베이스의 테이블 등을 통제한다.

⁷⁾ S. W. Rose, O. Borchert, S. Mitchell, S. Connelly, Zero Trust Architecture, NIST, (2020).

⁸⁾ NIST. Security and Privacy Controls for Information Systems and Organizations, (2020.12), NIST Special Publication 800-53 Revision 5

가. 접근제어 모델의 흐름

컴퓨팅에서 최초의 접근제어로 볼 수 있는 역사는 사용자가 시스템에 접근하기 위해 ID와 패스워드를 기입해야 하는 로그인 개념을 제공한 것이었다. 1960년대 초 MIT Computation Center가 최초로 시분할 운영체제(CTSS, Compatible Time-Sharing System)를 개발할 때 적용했다고 한다.⁹⁾

본격적으로 접근제어의 개념은 1971년 Butler Lampson이 제안한 접근제어 매트릭스(Access Matrix)로, 접근 권한과 사용자 식별 두 가지 요소로만 구성된 간단한 구조로 되어 있었다. 그 후 기밀성에 중점을 둔 Bell-LaPadula 모델(BLP)과 무결성에 중점을 둔 Biba 모델이 제안되었다. 이러한 보안 모델들을 기반으로 한 접근제어 방법에는 강제적 접근제어(MAC, Mandatory Access Control), 임의적 접근제어(DAC, Discretionary Access Control), 역할 기반 접근제어(RBAC, Role-Based Access Control), 규칙 기반 접근제어(RuBAC, Rule-Based Access Control), 속성 기반 접근제어(ABAC, Attribute-Based Access Control) 등이 있다.

강제적 접근제어(MAC)는 주체가 객체에 접근할 때, 관리자가 중앙에서 관리한다. 사전에 정의된 주체와 객체의 레벨을 비교하여 접근을 허가한다. 반면 임의적 접근제어(DAC)는 관리자의 개입 없이 주체 스스로 임의의 주체와 객체에 권한을 부여할 수 있다.¹⁰⁾ 두 가지 접근제어 방법 모두 개념은 동일하지만, 강제적 접근제어의 경우 조직 규모와 시스템 복잡성이 증가하면 고유의 규칙들을 많이 생성해야 하기 때문에 확장성의 문제가 발생한다. 임의적 접근제어의 경우엔 규칙의 속성을 세분화하면 복잡성의 문제가 발생한다. 이처럼 강제적 접근제어와 임의적 접근제어는 유연성과 보안성의 단점을 가지고 있어, 역할 기반 접근제어가 새로운 대안으로 떠올랐다. 유연한 조직 관리와 책임과 권한을 정교하게 제공해 주기 때문이다.

역할 기반 접근제어(RBAC)는 1992년 David Ferraiolo와 Richard Kuhn의 주도로 공식화되었다. 역할(Role)을 계층화한 개념의 가장 기본적인 모델인 “Flat RBAC(RBAC0)” 모델부터, 역할이 다른 역할로부터 권한을 상속하는 “Hierarchical RBAC”(RBAC1), 사용자에게 과도한 권한이 부여되지 못하도록 여러 역할 간의 권한을 분리하는 “Constrained RBAC”(RBAC2), 그리고 RBAC1과 RBAC2의 기능을 결합하여 오늘날의 역할 기반 접근제어의 특징을 갖게 된 “Symmetric RBAC”(RBAC3)로 구분된다. 역할 기반 접근제어는 규모가 크고 복잡한 구조 및 환경에서 보안 수준을 높일 수 있는 접근제어 방법으로, 데이터베이스 관리 시스템, 운영체제, 클라우드 컴퓨팅 플랫폼 등에서 많이 사용되고 있다. RBAC에 대한 NIST 모델은 2004년 2월 11일 국제 정보기술 표준화 위원회(ANSI/INCITS)에서 표준 기술로 채택되었으며, 2012년 INCITS 359-2012로 개정되었다.

RBAC는 MAC와 DAC의 단점을 보완한 것으로서 지금까지도 널리 사용되지만 조직의 역할과 권한을 정확하게 이해하고 있어야 효과적인 정책을 설정할 수 있다. 그 때문에 대규모 조직에서는 역할을 관리하는 데 비용이 많

⁹⁾ David Walden, Tom Van Vleck, Compatible Time-Sharing System(1961-1973), IEEE computer society, (2011), p.1-4.

¹⁰⁾ US Department of Defense, Department of Defense Trusted Computer System Evaluation Criteria, DOD 5200.28-STD, National Computer Security Center, (1985).

이 드는 등의 단점이 있다. 또한 유비쿼터스 시대가 도래하면서 통제 시스템에 접근하는 장소와 시간 등의 제약이 사라졌음에도 위치, 시간, 특정 작업 등 다양한 상황에 대처하지 못했다. 이러한 문제점을 보완하기 위해 2000년대 초 실시간 상황 정보를 기반으로 한 동적 접근제어 방식인 상황 인식 기반 접근제어(CAAC, Context-aware Access Contro)에 대한 연구가 활발하게 진행되었다.

RBAC 모델을 확장한 CAAC는 시스템의 리소스에 대한 액세스를 관리하는 동적 접근 방식이다. 이 방식으로는 역할뿐만 아니라 주체 및 객체에 대한 위치, 요청 시간, 사용 중인 장치, 네트워크 보안 수준, 시간 등의 상황 정보를 기반으로 접근을 제어한다. 그 때문에 복잡하고 실시간으로 변경될 수 있는 동적인 상황에 적합하며

위에서 소개한 모델 외에도 속성 기반 접근제어, 자격 기반 접근제어 등 많은 접근제어 기술이 있지만, 어느 한 가지 접근제어 모델을 통해 완벽한 제어와 관리를 할 수 없기 때문에 많은 데이터 보안 솔루션들은 접근제어 모델을 복합적으로 사용하는 하이브리드 형태의 모델을 채택하고 있다.

나. 향후 기술적 변화

IT 환경과 기술이 변화하고 개인정보 등에 대한 데이터 활용이 증가하고 있다. 이에 맞춰 접근제어 모델에 대해서도 새로운 방법들이 제안되고 있다. 특히 두 연구가 활발하게 진행되고 있는데, 하나는 스마트 콘트랙트(Smart Contract) 또는 데이터 관리 메시지를 기반으로 하는 블록체인 기반의 접근제어에 관한 것이고, 또 다른 하나는 머신러닝 기반의 접근제어에 대한 연구다.

블록체인은 활용 목적과 데이터 관리 방식, 참여자의 범위에 따라 퍼블릭(Public) 블록체인, 프라이빗(Private) 블록체인, 컨소시엄(Consortium) 블록체인으로 구분할 수 있다. 퍼블릭 블록체인은 누구나 접근할 수 있는 개방형 블록체인으로, 알고리즘을 통해 거래를 증명하여 거래 신뢰도를 높이고 익명성을 보장한다. 하지만 익명화된 상태에서 거래를 증명하는 알고리즘은 많은 계산을 필요로 하기 때문에 거래할 때 시간이 많이 든다. 프라이빗 블록체인은 사전에 승인된 사용자만 접근이 가능하도록 통제하는 형태로, 기업의 요구에 맞게 적용 가능한 기업형 블록체인이다. 중앙기관에서 트랜잭션을 증명하기 때문에 빠르고 효율적인 거래 처리가 가능하지만, 중앙기관에 의존하기 때문에 퍼블릭 블록체인보다는 안전성이 떨어진다. 컨소시엄 블록체인은 비즈니스에 가장 적합한 형태로, 프라이빗 블록체인처럼 승인된 사용자만 접근이 가능하며 조직 간의 합의 프로세스를 통해 접근을 허용한다. 사전에 합의된 규칙으로 빠르게 거래를 증명하며, 사용자 권한을 관리하여 민감한 정보를 효율적으로 제어할 수 있다.

블록체인 기반 접근제어의 경우엔 현재 사용되는 접근제어 기술들의 문제점을 해결할 방안이 있다. 블록체인 기반의 데이터 보호의 기본 개념은 데이터 저장소 계층 위에 블록체인 계층을 구축하는 것으로서, 데이터 접근을 정의하는 정책은 스마트 콘트랙트 또는 데이터 관리 메시지를 기반으로 한다. 블록체인 기반의 접근제어 기술에는 FairAccess의 ORBAC 접근제어 기술, Zyskind의 DHT 분산 저장소 기술, PrivacyGuard의 TEE 하드웨어 신뢰 환경,

Kanniche의 HIBE라는 ID 기반의 암호화를 기반으로 하는 기술 등이 있다.^{11) 12)}

블록체인 외에도 머신러닝 기반(Machine Learning)의 접근제어 기술이 전통적인 접근제어 기술들의 단점을 보완하는 새로운 대안으로 떠오르며 여러 연구 논문이 발표되고 있다. 하지만 불충분한 학습, 학습 중 실시간으로 변화하는 접근제어 정보 등으로 아직은 여러 상황에 대한 연구가 더 필요하다. 향후에는 블록체인과 머신러닝 기반의 접근제어 기술들이 데이터 보안 분야에서 가장 중요하게 널리 사용될 기술일 것으로 전망된다.

11) 김승현, 김수형, "블록체인 기반의 개인정보 관리를 위한 사용자 중심의 접근제어 서비스(User-Centric Access Control Service for Blockchain-Based Private Information Management)", 한국정보보호학회 31권 3호(2021), pp.341-351.

12) 김승현, 김수형, "블록체인 기반 접근제어 기술 동향(Analysis of Blockchain-based Access Control Technology)", 전자통신연구원 34권 4호(2019), pp.117-128.

2023 데이터산업 백서 집필진

제1부 • 초거대 AI 시대의 데이터 가치와 활용

제1장 • 초거대 AI 시대 데이터의 가치와 활용의 중요성

제2장 • 초거대 AI 시대 데이터 공유

제3장 • 데이터 활용의 이슈 : 데이터 편향성/윤리성

제2부 • 데이터산업 주요 정책 및 법제도 현황

제1장 • 국내 데이터 관련 정책 및 법제도 현황

제2장 • 해외 데이터 관련 정책 및 법제도 현황

제3장 • 데이터 활용 이슈의 법제도적 측면 : 저작권/개인정보

제3부 • 데이터산업 시장 현황

제1장 • 국내 데이터산업 시장 현황

제2장 • 국내 데이터산업의 역동성 및 생산성 분석

제3장 • 해외 데이터산업 시장 현황

제4부 • 산업별 데이터 활용 현황

제1장 • 금융분야 데이터 활용 현황

제2장 • 헬스케어분야 데이터 활용 현황

제3장 • 모빌리티분야 데이터 활용 현황

제4장 • 제조분야 데이터 활용 현황

제5장 • 농업분야 데이터 활용 현황

제6장 • 에듀테크분야 데이터 활용 현황

제7장 • 新 데이터 비즈니스

제5부 • 데이터산업 기술 동향

제1장 • 합성 데이터 생성 기술 동향

제2장 • 클라우드 스토리지 기술 동향

제3장 • 데이터 분석 기술 동향

제4장 • 데이터 보안 기술 동향

권호열 교수 강원대학교 컴퓨터공학과

전성민 교수 가천대학교 경영학부 / 서울대학교 AI연구원 객원연구원

변순용 교수 서울교육대학교 윤리교육과

윤상필 연구교수 고려대학교 정보보호대학원

한국데이터산업진흥원 산업기획팀

윤아리 변호사 김 · 장 법률사무소

고태우 팀장 KDB산업은행

고동환 박사 정보통신정책연구원 디지털경제연구실

고태우 팀장 KDB산업은행

우지환 교수 고려대학교 기술경영전문대학원

정규환 조교수 성균관대학교 삼성융합의과학원

아우토크립트

나혁준 연구소장 (주)비스텔리전스

홍승길 농업연구관 농촌진흥청 디지털농업추진단

차현승 대표이사 (주)카탐

김인현 대표 투이컨설팅

황인준 교수 고려대학교 전기전자공학부

김재락 매니저 네이버클라우드 IaaS Product Planning

박종건 매니저 네이버클라우드 Data Platform Planning

박병민 팀장 (주)신시웨이

2023 데이터산업 백서 [통권26호]

2023 Data Industry White Paper

편찬위원 김인현 투이컨설팅 대표
이형철 데이터산업협회장/(주)유플스 대표
윤석용 명지대학교 시빅데이터융합연계전공 교수
이상원 원광대학교 컴퓨터소프트웨어공학과 교수
황인준 고려대학교 전기전자공학부 교수

기획 및 편집 한국데이터산업진흥원 산업기반본부
하진희 본부장
안선빈 산업기획팀 주임
김혜빈 산업기획팀 인턴

발행처 한국데이터산업진흥원
서울특별시 중구 세종대로9길 42, 부영빌딩 7,8,11층 (04513)
Tel. 02-3708-5300
www.kdata.or.kr

발행인 윤혜정

발행일 2023년 10월 31일

디자인·편집·인쇄 사단법인 에스디워크 (서울지사) Tel. 02-2279-9938~9

ISSN 2465-7662

- 본 백서의 내용은 한국데이터산업진흥원의 공식 견해와 다를 수 있습니다.
- 본 백서 내용의 무단 전재를 금하며, 가공 인용 시에는 반드시 한국데이터산업진흥원, 「2023 데이터산업 백서」라고 밝혀 주시기 바랍니다.
- 본 백서는 과학기술정보통신부의 DB산업육성 사업의 결과물입니다.
- 「2023 데이터산업 백서」와 관련한 문의는 한국데이터산업진흥원으로 연락해 주시기 바랍니다.

2023 데이터산업 백서

2023 DATA INDUSTRY WHITE PAPER

(통권 26호)



 한국데이터산업진흥원
Korea Data Agency

서울특별시 중구 세종대로9길 부영빌딩 7,8,1층
T. 02.3708.5300 F. 02.318.5040
www.kdata.or.kr



ISSN 2465-7662

[비매품]