삼정 KPMG

ISSUE MONITOR

제121호
March 2020
삼정KPMG 경제연구원

기업 운영 혁신을 위한 데이터 과학: 기업의 활용 방안

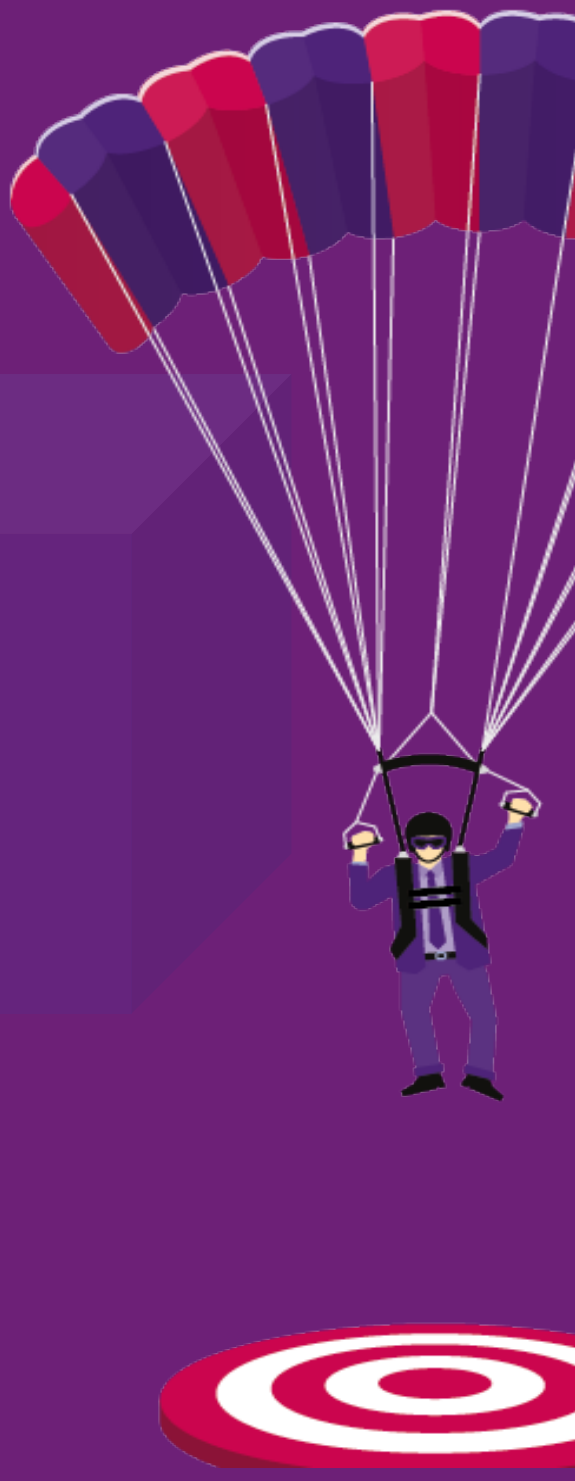
Contacts

삼정KPMG 경제연구원

김기범
선임연구원
Tel: +82 2 2112 7430
kkim28@kr.kpmg.com

이효정
이사
Tel: +82 2 2112 6744
hyojunglee@kr.kpmg.com

박도휘
책임연구원
Tel: +82 2 2112 0904
dohwipark@kr.kpmg.com



Contents

Executive Summary

데이터 자원과 데이터 과학은 석유 자원과 스트림라인(Streamline) 공정으로 빗대어 표현되곤 한다. 원유가 여러 단계의 공정을 거쳐 플라스틱, 섬유, 고무와 같은 제품으로 재탄생하듯 데이터 자원도 여러 단계를 거치면서 기존에 없던 새로운 가치를 창출해내기 때문이다. 본 Issue Monitor에서는 데이터 업무 수행 흐름을 5단계(데이처 수집→저장→처리→분석→활용)로 구분하고 각 단계에서 기업이 직면하고 있는 이슈사항과 이슈 해결에 도움을 줄 기술·기법 인에이블러(Enabler)를 제시했다. 각 단계에서 기업이 중점적으로 고려해야 할 사항과 함께 데이터 경제 시대에 기업의 운영 혁신 방안을 모색해보고자 한다.

Executive Summary

■ 데이터 자원과 데이터 과학

- 21세기에 접어들면서 4차 산업혁명 기반 기술의 도입과 발전으로, 전 산업에서 발생하는 데이터를 수집하고 분석할 수 있는 환경이 만들어고 있음
- 본 보고서에서는 데이터 업무 수행 단계를 5단계(데이처 수집→저장→처리→분석→활용)로 구분하고 각 단계에서 기업이 활용할 수 기술·기법 인에이블러(Enabler)와 운영 혁신 방안을 모색

■ 데이터 수집

- **기술·기법 인에이블러:** ① 자율형 사물인터넷(IoT)을 통한 데이터 센싱, ② 크롤링, 오픈API(Application Programming Interface) 등 데이터 수집 기술 활용, ③ 데이터 수집·연계·통합을 통한 Customer 360° 확보
- **비즈니스 고려사항:** ① 데이터로 해결할 비즈니스 이슈를 우선 정의, ② 비용과 효용을 고려해 데이터 수집 대상과 주기, 활용 기술을 선정, ③ 데이터의 양적·질적 측면을 고려해 데이터 품질 관리 강화

■ 데이터 저장

- **기술·기법 인에이블러:** ① 클라우드 플랫폼을 통한 혁신 환경 구축, ② 데이터 웨어하우스에서 데이터 레이크(Data Lake)로, ③ RDB(Relational Database)에서 그래프DB까지, 데이터베이스의 진화
- **비즈니스 고려사항:** ① 총소유비용(TCO)을 고려한 클라우드 마이그레이션 여정 설계, ② 데이터 계층화를 통한 데이터 생명 주기 관리, ③ 기업의 성격에 맞는 저장 플랫폼 선정

■ 데이터 처리

- **기술·기법 인에이블러:** ① 실시간 데이터 처리를 위한 엣지 컴퓨팅, ② 오토 레이블링으로 데이터 처리 간소화, ③ 암호기반 프라이버시 보호 기술로 데이터 비식별화
- **비즈니스 고려사항:** ①데이터의 상태를 확인하고 설계한 분석 요건에 맞도록 데이터 정제, ② IT 부서의 개입 없이도 현업 담당자가 직접 데이터 처리를 할 수 있도록 데이터 전처리 간소화 도구 활용

■ 데이터 분석

- **기술·기법 인에이블러:** ① 의사결정의 근간이 되는 기술 분석 및 진단 분석, ② 경영 활동의 불확실성을 낮추는 예측 분석, ③ 경험과 학습을 통해 진화하는 처방 분석
- **비즈니스 고려사항:** ① 사후적 분석에서 사전적 분석으로 진화하는 이행 로드맵 설계, ② 사업 목표와 일원화된 분석 프로세스 정립

■ 데이터 활용

- **기술·기법 인에이블러:** ① 자체적인 데이터 플랫폼 구축, ② 서드파티(Third-Party) 데이터 플랫폼 도입, ③ 생태계 기반의 디지털 트윈 아키텍처 구축
- **비즈니스 고려사항:** ① 조직의 비즈니스 목표에 대한 명확한 이해 필요, ② 고객의 니즈를 적극적으로 고려, ③ 소규모 팀과 업무 단위로 데이터 활용을 추진

기업 운영 혁신을 위한 데이터 과학: 기업의 활용 방안

데이터 자원과 데이터 과학

"데이터 산업은 데이터의 수집에서부터 활용에 이르는 모든 활동을 통해 상품을 제공하는 산업"

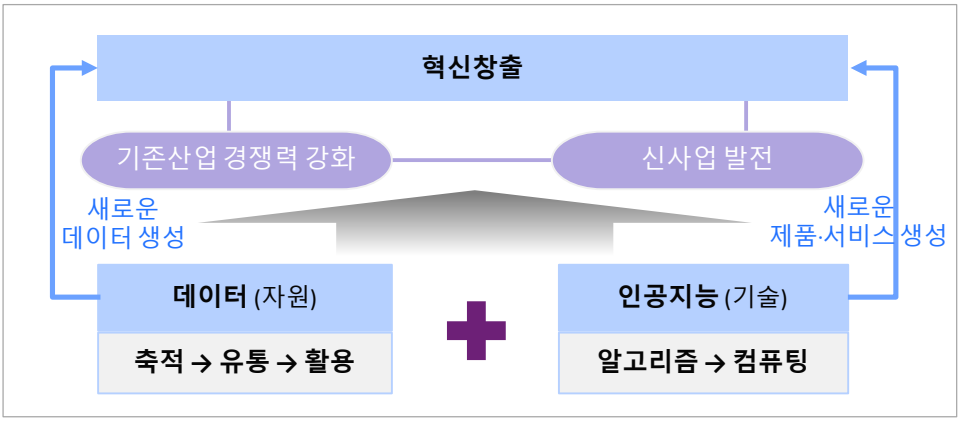
"본 보고서는 데이터 자원을 통한 가치 창출을 가능케 하는 기술·기법 인에이블러와 운영 혁신 방안을 제시"

데이터 산업의 부상과 성장 배경

데이터 산업은 일반적으로 데이터의 생산, 수집, 저장, 처리, 분석, 유통, 활용 등을 통해 상품을 생산·제공하는 산업으로 정의된다. 최근 국내 데이터 산업이 급격히 부상하고 있다. 한국데이터산업진흥원에 따르면 2018년 국내 데이터 산업의 전체 시장 규모는 15조 1,545억 원이며, 2017년의 14조 3,530억 원 대비 5.6% 성장한 것으로 나타났다.

데이터 산업이 부상하는 이유는 데이터 자원을 활용하여 새로운 비즈니스 가치를 창출할 수 있기 때문이다. 최근 부상하고 있는 인공지능 기술은 이러한 데이터를 분석하고 활용하여 새로운 제품 및 서비스를 만들 수 있도록 돕는다. 이렇게 창출된 새로운 비즈니스 가치는 산업 전반적인 혁신을 야기하고, 더 많은 데이터를 생성하여 데이터 과학의 선순환 고리를 완성한다.

>> 데이터 자원과 데이터 과학을 통한 산업 혁신 창출



Source: 과학기술정보통신부, 삼정KPMG 경제연구원 재구성

21세기에 접어들면서 4차 산업혁명 기반 기술의 도입과 발전으로, 전 산업에서 발생하는 데이터를 수집하고 분석할 수 있는 인프라가 빠르게 구축되었다. 산업 전체적으로 데이터의 양이 폭발적으로 증가하면서 소위 '빅데이터(Big Data)'를 분석하고 활용할 수 있는 역량이 기업의 핵심 역량으로 자리잡게 되었다. 데이터 산업의 성장을 촉진하는 요인으로 비약적인 발전을 거듭하고 있는 컴퓨팅 능력을 빼놓을 수 없다. 특히 병렬 연산처리에 특화된 반도체, GPU(Graphic Processing Unit) 등의 발전이 컴퓨팅 성능 발전에 기여하고 있다. GPU는 소수의 고성능 코어로 구성되어 있는 CPU(Central Processing Unit)보다 복잡한 연산에 대해서는 성능이 떨어지지만, 대량의 데이터를 빠르게 처리하는 데 장점을 가지고 있어 업계의 주목을 받고 있다.

역사상 그 어느 시기보다 데이터의 중요성이 부각되고 있는 현 시점에서, 기업들은 어떤 방식으로 데이터를 활용해 나갈 수 있을지에 대해 고민해야 한다. 본 보고서는 데이터 자원을 통한 가치 창출을 가능케 하는 기술·기법 인에이블러(Enabler)와 운영 혁신 방안을 모색해보고자 한다.

기업 운영 혁신을 위한 데이터 과학: 기업의 활용 방안

> 산업 혁신의 촉발과 기술의 발전에 따라, 시대가 요구하는 사고 능력의 형태 또한 급격히 변화해왔음

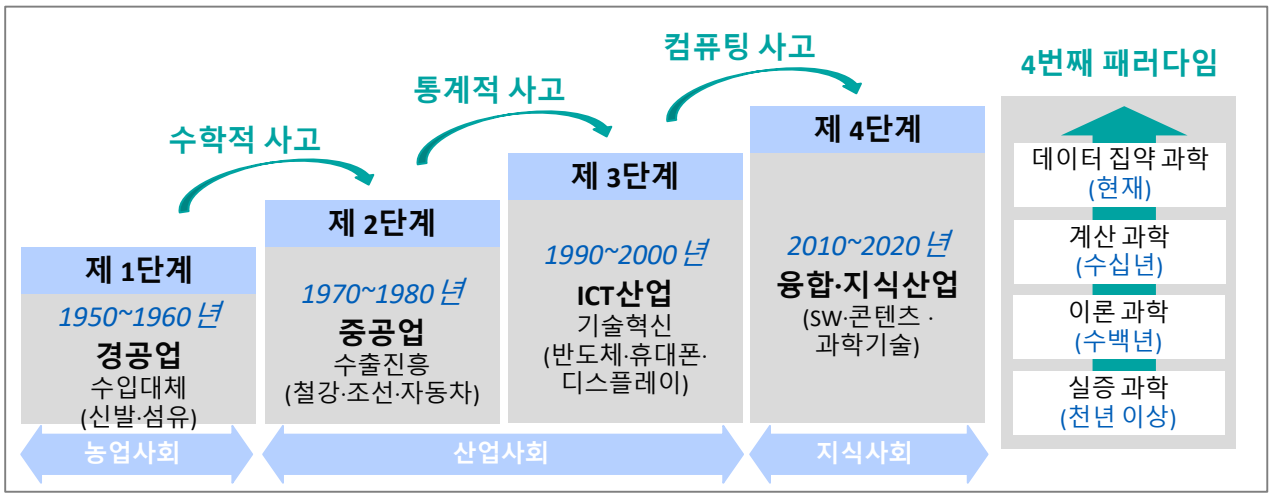
국가 주력 산업의 전환이 이끄는 사고 패러다임의 진화

한국은 선진국과 비교하여 늦은 시점에 산업 혁명의 시류에 동참하였지만, 시대가 요구하는 사고 패러다임 변화의 흐름을 그대로 답습해왔다.

과거 1950년대 신발·섬유 제조 등 경공업에 의지했던 한국의 경제는 1970년대에 접어들어 중공업의 중흥을 맞이하며 크게 성장했다. 동시에 조선·철강·건설 등 중후장대 상품의 품질을 담보하고 글로벌 경쟁력을 강화하기 위한 기계·장비 기술 및 공학이 발전하면서 수학적 사고의 중요성이 강조되었다. 1990년대에 접어들어, 한국은 ICT산업에 적극적으로 뛰어들며 다시 한 번 경제의 전환기를 맞이하게 된다. 반도체와 통신, 디스플레이 등 경박단소 제품 및 관련 서비스 개발을 위해 확률론에 기반한 통계적 사고의 중요성이 부상하였다. 최근에는 산업이 융합되는 컨버전스 추세로 인해, 수많은 데이터에서 기존에 없던 인사이트를 창출하는 컴퓨팅 사고가 필수인 시대로 접어들게 되었다.

최근 빅데이터가 10년도 되지 않는 짧은 시간에 성큼 우리곁에 다가왔고, 추상화(Abstraction)와 자동화(Automation)가 근간을 이루는 컴퓨팅 사고도 새로운 사고체계로 함께 받아들일 시점이 되었다. 컴퓨팅 사고 체계가 갖는 산업적인 영향력과 범위가 매우 커서, 미국, 영국, 이스라엘, 일본을 포함한 선진국뿐만 아니라 중국, 인도, 에스토니아 등 많은 나라들이 컴퓨터적 사고 체계를 갖춘 인력 확보가 미래 국가 경쟁력과 밀접한 연관이 있다고 내다보고 집중적으로 인력양성에 나서고 있다. 거시적인 측면 뿐만 아니라, 개인적인 측면에서도 컴퓨터적 사고와 통계적 사고를 갖춘 사람과 그렇지 못한 사람과의 차이는 산업경제 체제에서 빈부격차보다 더 큰 정보 불평등(Digital Divide)을 디지털 경제 시대에 야기할 것으로 예측되고 있다. 즉, 변화된 환경에 준비되고 잘 적응한 사람은 과거보다 훨씬 더 큰 혜택과 권한을 가질 것으로 예상된다.

>> 컴퓨팅 사고 패러다임의 진화



Source: 삼정KPMG

기업 운영 혁신을 위한 데이터 과학: 기업의 활용 방안

"데이터는 데이터 처리 과정을 통해 정돈된 데이터(Tidy Data)로 변환되야 분석에 활용할 수 있어"

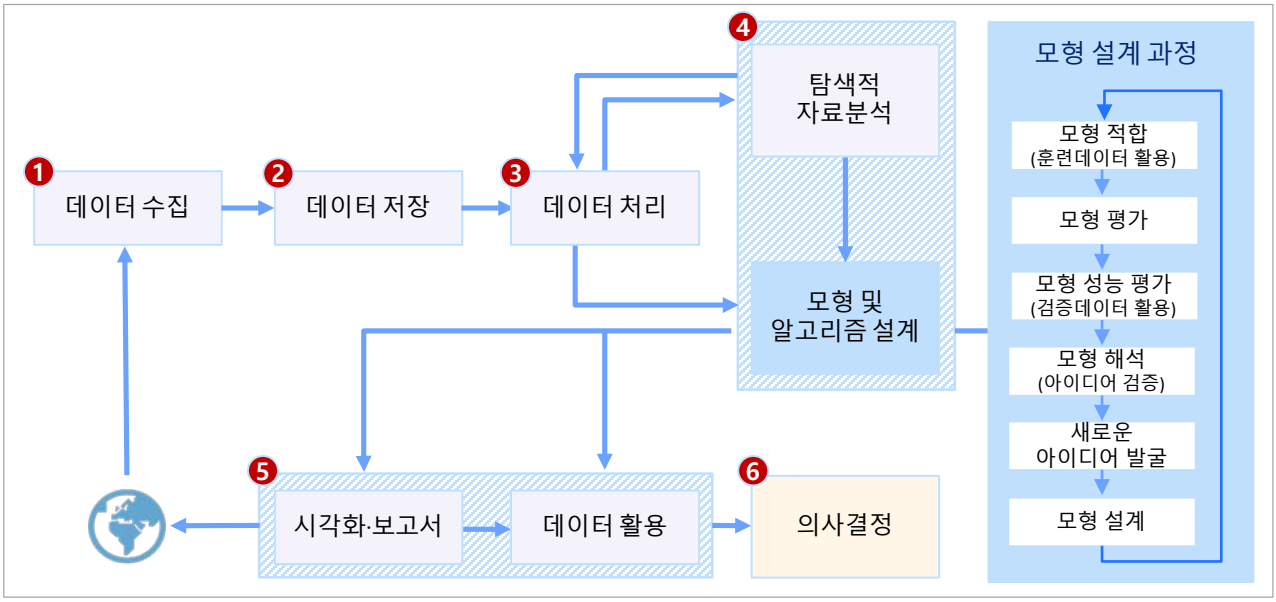
데이터 수집부터 분석·활용까지, 데이터 업무 수행 흐름도

기업은 데이터를 바탕으로 의사결정을 하기까지 여러 단계를 거친다. 그 첫 단계는 내·외부의 데이터를 수집하고 저장하는 단계다. 이후, 데이터 처리 과정을 통해 분석할 수 있는 정돈된 데이터(Tidy Data)로 변환된다. 변수와 관측점, 값이 일관된 체계를 갖춰야만이 분석 패키지와 소통할 수 있기 때문이다.

데이터 처리 이후 분석 단계에서 기업은 분석 목적에 맞게 모형(Model)을 설계한다. 모형 설계 과정에서 첫 번째로 해야 할 일은 탐색적 자료 분석을 통해 데이터의 특성을 파악하는 것이다. 가령, 데이터의 쏠림 현상(Skewness)이 없는지, 데이터를 왜곡시키는 아웃라이어가 없는지 등을 확인해야 한다. 통계 분석에서 결정계수(R^2)를 확인하는 것 또한 모형의 설명력을 가늠할 수 있는 정량적인 척도가 되기도 한다. 필요에 따라 데이터를 정규화하거나 연속형 변수를 범주형으로 바꾸는 등의 과정을 통해 분석 데이터의 신뢰도를 높일 수 있도록 해야 한다. 이처럼 데이터 과학은 결국 모형 개발자의 판단과 선택에 따라 이뤄지는 결과물로 봐도 과언이 아니다.

분석 목적에 따라 설계된 모형과 알고리즘은 지속적으로 평가된다. 시간이 흐르면서, 데이터의 성격이 바뀌는 경우도 있고, 분석 과정에서 더 간결하면서도 정확도가 높은 모형이 발견되기도 하기 때문이다. 기본적으로 오차가 최소화되는 모형이 좋은 모형으로 꼽히지만, 과적합에 대한 문제, 모형의 복잡성에 대한 문제, 컴퓨터 계산 자원 소모에 대한 문제, 연산시간에 대한 다양한 조건도 모형의 평가요소로 고려되곤 한다.

>> 데이터 업무 수행 흐름도



Source: 삼정KPMG 경제연구원

기업 운영 혁신을 위한 데이터 과학: 기업의 활용 방안

데이터의 분석 결과가 도출되면, 이를 바로 의사결정에 활용하기도 하지만, 일부는 의사결정을 돕는 시각화 도구를 활용하기도 한다. 또한 분석 결과는 데이터 수집을 개선시켜 기업의 데이터 자원을 더욱 풍부하게 만들기도 한다.

본 보고서에서는 데이터 업무 수행 단계를 수집, 저장, 처리, 분석, 활용으로 구분하고 각 단계에서 기업들이 직면하고 있는 이슈사항과 이슈 해결에 도움을 줄 기술·기법 인에이블러(Enabler)를 제시했다. 이후, 각 단계에서 기업이 중점적으로 고려해야 할 사항을 제시하고자 했다.

>> 데이터 업무 수행 단계별 기술·기법 인에이블러(Enabler)와 비즈니스 고려사항

단계	기술·기법 인에이블러	비즈니스 고려사항
데이터 수집	자율형 사물인터넷(IoT)을 통한 데이터 센싱 크롤링, 오픈API[1] 등 데이터 수집 기술의 활용 데이터 수집·연계·통합을 통한 Customer 360° 확보	▪ 데이터로 해결할 비즈니스 이슈 정의 ▪ 비용과 효용을 고려해 데이터 수집 대상과 주기, 활용 기술을 선정 ▪ 데이터의 양적·질적 측면을 고려해 데이터 품질 관리 강화
데이터 저장	클라우드 플랫폼을 통한 혁신 환경 구축 데이터 웨어하우스에서 데이터 레이크로 RDB[2]에서 그래프DB까지, 데이터베이스의 진화	▪ 총소유비용(TCO)을 고려한 클라우드 마이그레이션 여정 설계 ▪ 데이터 계층화를 통한 데이터 생명 주기 관리 ▪ 기업의 성격에 맞는 저장 플랫폼 선정
데이터 처리	실시간 데이터 처리를 위한 엣지 컴퓨팅 오토 레이블링으로 데이터 처리 간소화 암호기반 프라이버시 보호 기술로 데이터 비식별화	▪ 데이터의 상태를 확인하고 설계한 분석 요건에 맞도록 데이터 정제 ▪ IT부서의 개입 없이도 현업 담당자가 직접 데이터 처리를 할 수 있도록 데이터 전처리 간소화 도구 활용
데이터 분석	의사결정의 근간이 되는 기술 분석 및 진단 분석 경영 활동의 불확실성을 낮추는 예측 분석 경험과 학습을 통해 진화하는 처방 분석	▪ 사후적 분석에서 사전적 분석으로 진화하는 이행 로드맵 설계 ▪ 사업 목표와 일원화된 분석 프로세스 정립
데이터 활용	자체적인 데이터 플랫폼 구축 서드파티(Third-Party) 데이터 플랫폼 도입 생태계 기반의 디지털 트윈 아키텍처 구축	▪ 조직의 비즈니스 목표에 대한 명확한 이해 필요 ▪ 고객의 니즈를 적극적으로 고려 ▪ 소규모 팀과 업무 단위로 데이터 활용을 추진

Source: 삼정KPMG 경제연구원
Note: 1) API(Application Programming Interface)는 다양한 서비스와 데이터를 좀 더 쉽게 이용할 수 있도록 공개한 개발자를 위한 인터페이스. 2) RDB(Relational Database)는 관계형 데이터베이스

기업 운영 혁신을 위한 데이터 과학: 기업의 활용 방안

1. 데이터 수집

데이터 수집 과정에서의 이슈사항

기업들은 어떻게 기업 내·외부 데이터에 접근해 보다 많은 데이터를 수집할 것인가에 대해 고민을 하고 있다. 특히 오늘날의 클라우드 기업들이 비교적 저렴한 가격에 컴퓨팅 인프라와 기업들이 쉽게 활용할 수 있는 분석 알고리즘을 제공하고 있어, 기업이 얼마나 많은 양질의 데이터를 수집하고 축적하고 있는지는 중요한 차별화 요소가 되고 있다.

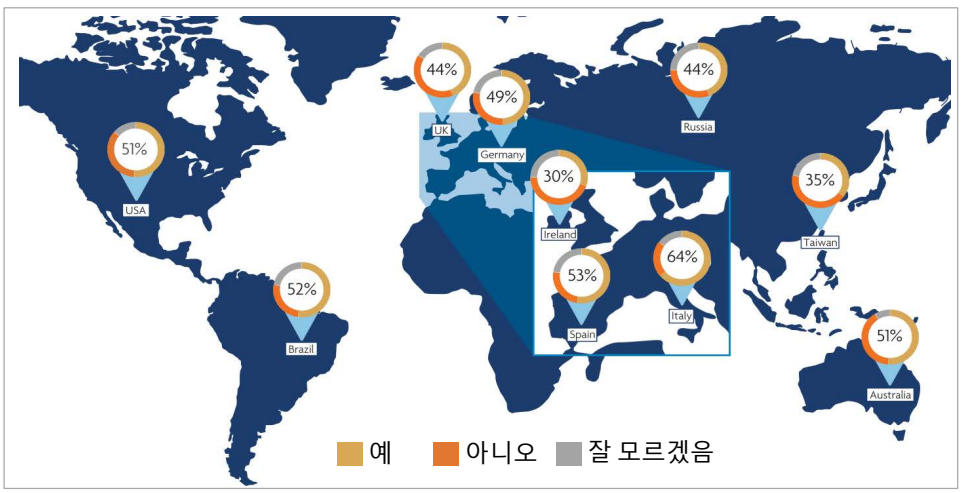
"기업은 필요 이상의 시간과 리소스를 데이터 수집 업무에 할애해, 기업 운영에 비효율성을 야기"

기업이 데이터 수집 단계에서 겪고 있는 이슈 중 하나는 필요 이상의 시간과 리소스를 데이터 수집 업무에 할애한다는 점이다. 기업 내·외부 데이터를 연결해 새로운 통찰력을 얻을 수 있을 거라는 생각에 기업들은 무작위로 데이터를 쌓는 경향이 있다. 하지만 아무리 많은 데이터를 보유한다 하더라도 기업에서 이를 활용하지 못하면 무용지물이며, 오히려 기업 운영에 비효율성을 야기시킬 수 있다. 데이터가 수집 단계에 명확한 목표와 체계가 확립되지 않은 채 수집되는 식별 불가능한 데이터는 결코 유의미한 가치를 지닌 정보와 인사이트로 전환될 수 없다.

"프라이버시 규제가 강화되는 가운데, 기업이 데이터 수집 과정에서 유의해야 할 사항이 많아져"

또 다른 이슈사항으로는 기업이 데이터 수집 단계에서 유의해야 할 사항이 많아지고 있다는 점이다. 전 세계적으로 프라이버시에 대한 규제가 강화되고 있지만, 기업들은 아직 새로운 규제에 완벽히 대응할 준비가 되어 있지 않은 상황이다. 2018년 KPMG와 법률 정보 매체인 'The Legal 500'이 공동으로 각 기업의 법무팀이 2018년 5월 유럽연합(EU)에서 발효된 GDPR(General Data Protection Regulation)에 어떻게 대응하고 있는지 살펴본 결과, 응답자 중 46%만이 GDPR에 충분히 준비되어 있다고 응답했다. GDPR 규정을 어길 경우, 전 세계의 매출의 4% 혹은 2,000만 유로 중 높은 금액이 과징금으로 부과될 수 있어 높은 주의가 요구됨에도 불구하고, 절반 이상의 조직에서는 GDPR 법규와 규제에 완전히 준비되어 있지 않은 것으로 나타났다.

>> 임직원이 GDPR에 대해서 충분히 준비되어 있다고 생각하십니까?



Source: The GC's Guide to GDPR, The Legal 500 & KPMG International
Note: KPMG와 법률 정보 매체인 'The Legal 500'이 공동으로 호주, 브라질, 독일, 아일랜드, 이탈리아, 러시아, 스페인, 대만, 영국, 미국의 448명의 Legal Counsel을 대상으로 설문조사를 진행

기업 운영 혁신을 위한 데이터 과학: 기업의 활용 방안

데이터 수집을 위한 기술·기법 인에이블러

① 자율형 사물인터넷(IoT)을 통한 데이터 센싱

모든 사물이 연결되는 초연결 시대에 사물인터넷(IoT)은 데이터 수집을 돕는 중추적인 기술 인에이블러로 자리매김하고 있다. IoT는 모든 사물을 인터넷으로 연결해 사람과 사물, 사물과 사물 간 정보를 교환하는 기술이다. 시장 조사 기관 IDC에 따르면 연간 생성된 데이터량은 2016년 16ZB(제타바이트; 약 10^9TB)에서 2020년 44ZB로 증가하고, 2025년에는 180ZB까지 늘어날 것으로 전망된다.

> 모든 사물이 연결되는 초연결 시대에 IoT는 데이터 수집을 돕는 중추적인 기술 인에이블러

제조 현장에 도입된 IoT는 생산설비에 대한 정보를 실시간 확보할 수 있도록 하며, IoT는 기존에는 확보할 수 없었던 다양한 종류의 데이터 수집을 가능케하고 있다. IoT의 요소 기술로는 유무선 통신 및 네트워크 인프라 기술, IoT 서비스 인터페이스 기술, 센싱 기술로 구분될 수 있다.

>> 사물인터넷(IoT)의 요소 기술

요소 기술	설명
통신·네트워크	와이파이, 3G/4G/5G, 블루투스 등 사람·사물·서비스를 연결할 수 있는 유무선 네트워크를 의미
인터페이스	IoT의 주요 구성 요소를 특정 기능을 수행하는 응용 서비스와 연동하는 역할
센서	기존의 전통적인 물리적 센서에서 다양한 서비스에 응용 가능한 스마트 센서로 발전. 개별적 역할을 수행하는 센서에서 다중 센서 기술을 통해 고차원적인 정보 수집

Source: 기계저널, 삼정KPMG 경제연구원 재구성

> IoT 기술은 데이터 센싱과 수집, 관리 등 상호 연결성이 강조되는 연결형 IoT에서 지능형 IoT, 자율형 IoT로 진화

IoT 기술은 데이터 센싱과 수집, 관리 등 상호 연결성이 강조되는 연결형 IoT에서 지능형 IoT, 자율형 IoT로 진화하고 있다. 지능형 IoT는 IoT와 클라우드, 인공지능이 접목되어 지능적 행위를 수행할 수 있는 단계를 의미한다. 향후 IoT 기술은 자율형 IoT로 넘어갈 것으로 예상되는데, 이는 디지털 트윈으로 실제 세계와 가상 세계가 지속적으로 상호작용하며 현장의 사물이 고도화된 지능적 대응을 하는 단계를 의미한다.

>> 사물인터넷(IoT)의 발전 단계

분류	연결형 IoT	지능형 IoT	자율형 IoT
기술적 특성	무선통신, 연결관리	IoT와 클라우드 기술의 연동	IoT와 클라우드 기술의 연동
동작 위치	디바이스, 플랫폼	클라우드	사물, 엣지
의사결정 주체	인간의 판단	+ 클라우드 지능	+ 사물 협업 지능

Source: 한국전자통신연구원, 삼정KPMG 경제연구원 재구성

기업 운영 혁신을 위한 데이터 과학: 기업의 활용 방안

② 크롤링, 오픈API 등 데이터 수집 기술의 활용

현재 데이터 수집 기술이 오픈소스를 중심으로 활발히 개발되고 있다. 특허청이 발간한 '2017 통계로 본 특허 동향'에서는 2002년부터 2016년까지 전 세계 특허 통계 데이터베이스(PATSTAT)에 공개된 빅데이터 분야의 특허를 분석했는데, 6,614건의 출원 특허 중 '빅데이터 수집'에 해당하는 특허가 가장 많은 것으로 나타났다. 76.5%(5,060건)에 해당하는 특허가 데이터 수집과 관련된 기술이었으며, 나머지 15.6%가 데이터 처리·저장·관리, 7.9%가 분석과 관련된 특허 출원인 것으로 나타났다. 데이터 수집 분야의 특허를 세부적으로 보면, 디바이스 데이터 수집 기술(2,134건)과 웹·소셜 미디어 데이터 수집 기술(1,627건)이 특히 많은 것으로 파악됐다.

"현재 데이터 수집 기술은 오픈소스를 중심으로 활발히 개발되고 있어"

>> 데이터 수집 기술의 세부 분류와 특허 건수

기술	내용	건수
디바이스 데이터 수집 기술	다양한 디바이스 데이터 소스에 접근하여 데이터를 수집하는 기술	2,134
웹·소셜 미디어 데이터 수집 기술	웹·소셜 미디어 서비스를 기반으로 데이터를 조합 및 질의하여 연계하는 기술	1,627
트랜잭션, 운영, 웨어하우스 데이터 수집 기술	운영 데이터베이스의 변경을 실시간으로 감지해 데이터 웨어하우스에 반영, 최신 데이터 기반의 의사결정을 지원하는 기술	1,133
빅데이터 유통 인프라 기술	빅데이터 유통 효율성을 위한 기술과 데이터 간 상호 연계를 위한 표준, 개인 정보 보안 기술	166

Source: 통계청, 삼정KPMG 경제연구원 재구성

"기업들은 수집하는 데이터의 유형과 형태에 따라 효율적인 방법으로 데이터를 수집하고 있어"

기업들은 수집하는 데이터의 유형과 형태에 따라 효율적인 방법으로 데이터를 수집하고 있다. 현재 기업에서 활용하는 데이터를 수집 기술은 웹상의 정보를 수집하는 크롤링(Crawling) 기술부터, 로그 수집기(Log Aggregator), RDB 수집기(RDP Aggregator), RSS 리더, 섀밍(Shaming), 오픈API 등 다양하다.

>> 주요 데이터 수집 기술

수집 기술	특징
Crawling	웹 로봇을 이용하여 조직 외부의 SNS, 뉴스, 웹정보 등 공개되어 있는 웹문서를 수집하는 기술
Log Aggregator	조직 내부에 존재하는 웹서버 로그, 웹로그, 트랜잭션 로그, 클릭 로그 등 각종 로그 데이터를 수집하는 오픈소스 기술
RDB Aggregator	관계형 데이터베이스에서 정형 데이터를 수집하여 HDFS(하둡 분산파일시스템)이나 Hbase와 같은 NoSQL에 저장하는 오픈소스 기술
RSS Reader	RSS(Really Simple Syndication)은 웹 기반 최신 정보를 공유하기 위한 XML 기반 콘텐츠 배급 프로토콜
Shaming	인터넷에서 음성, 오디오, 비디오 데이터를 실시간 수집 하는 기술
Open API	서비스, 정보, 데이터 등을 어디서나 쉽게 이용할 수 있도록 개방된 API로 데이터를 수집하는 방식

Source: 과학기술정보통신부, 한국정보화진흥원, 삼정KPMG 경제연구원 재구성

기업 운영 혁신을 위한 데이터 과학: 기업의 활용 방안

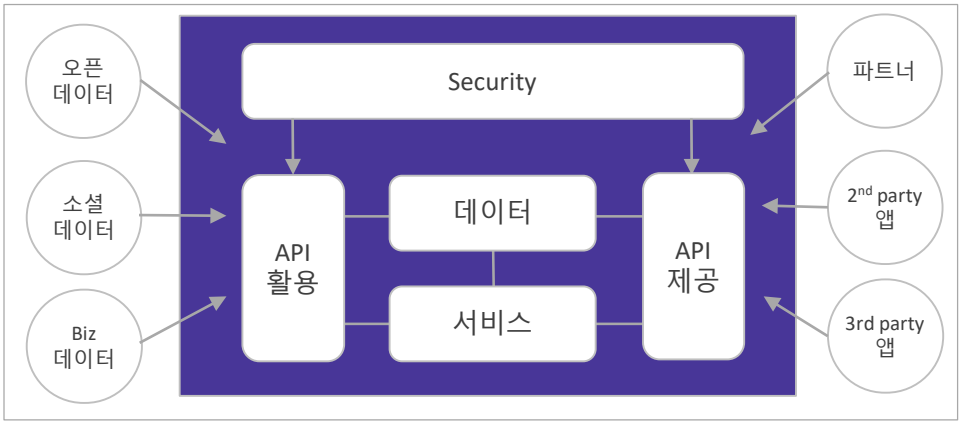
연결되지 않은 기업은 디지털 시대에 더 이상 생존이 어려워지게 되었다. 이에 따라 기업들은 데이터 수집 기술 중에서도 오픈API(Application Programming Interface)를 통한 데이터 거래·유통에 많은 관심을 기울이고 있다. 오픈API란 데이터 플랫폼을 외부에 공개해 외부 개발자나 사용자들이 다양한 서비스나 애플리케이션을 개발할 수 있게 하는 프로그램을 의미한다. 데이터 생산자들은 자사의 데이터를 제3자가 이용 가능하도록 오픈API 플랫폼을 공개할 수 있으며, 파트너사나 서드 파티(Third Party) 기업들은 이를 재가공하여 새로운 비즈니스 모델을 만들 수 있다.

"오픈API란 데이터를 외부에 공개해 제3자가 다양한 서비스를 개발할 수 있도록 공유하는 프로그램을 의미"

통신 사업자로 풍부한 데이터를 보유한 SK텔레콤의 경우도 데이터 산업 활성화를 위해 오픈API 전략을 적극 펼치고 있다. 2019년 9월 SK그룹은 'SK 오픈API 포털'을 구축하였고 SK하이닉스, SK브로드밴드, 11번가, SK실트론 등 SK그룹의 주요 ICT 회사들이 보유한 API와 활용 매뉴얼, 다양한 샘플을 공개했다. 데이터 취합자나 서비스 개발자들은 이를 활용해 응용 프로그램을 만들거나 새로운 서비스를 개발할 수 있다.

"오픈API로 제공받은 다양한 유형의 데이터를 재가공해 새로운 비즈니스 모델을 만들고 있어"

>> 오픈API로 연결된 기업 개념도



Source: 한국데이터산업진흥원, 삼정KPMG 경제연구원 재구성

>> 데이터 유통 생태계 형성을 위한 오픈API 전략

	데이터 생산자	데이터 취합자	서비스 개발자
역할	자사가 보유한 데이터를 개방하고 제3자가 이용 가능하도록 함	오픈 데이터 및 그 외 가치있는 데이터를 수집·취합하여 데이터의 가치 증진	오픈 데이터를 활용해 애플리케이션 또는 웹페이지 개발
예시	오픈 API 형태로 제공 (ex. 네이버 지도 API)	공공데이터를 재가공하여 API 형태로 제공	애플리케이션이나 웹 형태로 제공 (ex. 직방, 굿닥)

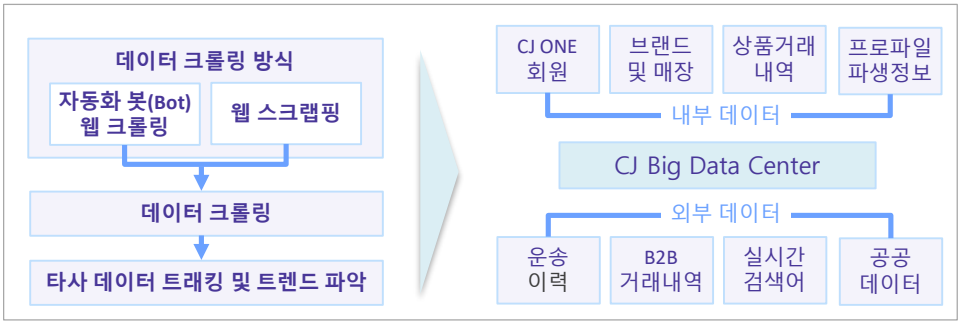
Source: SK텔레콤, 삼정KPMG 경제연구원

기업 운영 혁신을 위한 데이터 과학: 기업의 활용 방안

또 다른 예로, CJ올리브네트웍스는 온라인 유통 트렌드 파악을 위해 내·외부 데이터를 결합·분석하여 활용하고 있다. CJ올리브네트웍스는 웹크롤링과 웹 스크래핑 기술을 활용해 분석에 필요한 부가 데이터를 확보하고, 내부 데이터와 결합하여 온라인 시장의 흐름을 찾는 트렌드 분석 정보 제공 서비스를 하고 있다. 제조 및 유통사 기준의 카테고리, 브랜드 및 품목(SKU)별 구매 실적을 분석하여 경쟁사 동향 및 상품 연계 분석 등에 활용할 수 있다. CJ올리브네트웍스는 데이터 유통 생태계 조성의 일환으로 한국정보화진흥원(NIA)에서 주관하는 '2018년 빅데이터 전문센터'에 선정된 바 있다.

"CJ올리브네트웍스는 웹크롤링과 웹 스크래핑 기술을 활용해 분석에 필요한 부가 데이터를 확보하고 온라인 시장의 흐름을 찾아"

>> CJ올리브네트웍스의 데이터 크롤링 활용 방안



Source: CJ올리브네트웍스

③ 데이터 수집·연계·통합을 통한 Customer 360° 확보

기업들은 소비자와의 접점에서 확보한 외부 데이터만으로는 유용한 통찰력을 얻기 어렵다. 외부에서 확보한 데이터를 내부의 데이터와 연계·통합해야 고객의 추가적인 속성을 파악하고 고객에 대한 전방위적인 이해, 즉 'Customer 360° View'를 확보할 수 있다.

"내·외부 데이터를 연계해야 고객에 대한 추가적인 속성을 파악하고 고객에 대한 'Customer 360° View'를 확보할 수 있어"

삼성SDS는 전방위적으로 고객을 이해하기 위한 방안으로 내·외부 데이터 연계에 주목하고 있다. 기업은 개인식별정보(PII, Personally Identifiable Information)를 정의하고, 이를 기준으로 판매, VOC, 캠페인, e커머스 등으로부터 수집한 고객 데이터를 연계·통합해야 한다. 채널별, 사업부별, 시스템별로 산재되어 있는 고객 데이터를 PII를 기준으로 연결하고, 식별되지 않은 데이터의 경우, 지속적으로 데이터를 축적해 식별자를 확인할 수 있도록 해야 의미있는 고객 정보와 인사이트로 전환될 수 있다.

>> Customer 360° 확보를 위한 데이터 연계

내부 데이터 Enrichment	내·외부 데이터 연계
고객 기본 데이터 표준화 및 정제 고객 접점 채널별 상이한 포맷 표준화	**비식별 데이터 축적 및 모니터링** 축적된 비식별 데이터 모니터링
고객 데이터 매핑 및 그룹화 고객 데이터 통합을 위한 PII 정의	**Unique ID 기반 가상고객 정의** 비식별 고객 데이터 간 연결고리 확인
지속적인 데이터 품질 유지 신규 고객 데이터 모니터링 및 품질 관리	**내부 데이터와 연계** AI 기반으로 내·외부 데이터 연계

Source: 삼성SDS

기업 운영 혁신을 위한 데이터 과학: 기업의 활용 방안

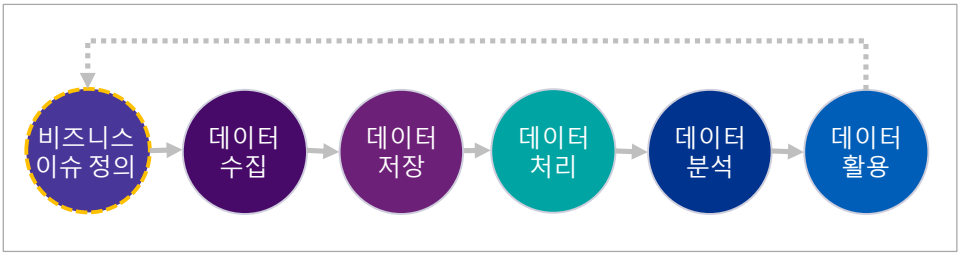
비즈니스 고려사항

데이터로 해결할 비즈니스 이슈 정의 필요

데이터를 수집하는 데 들이는 시간을 최소화하기 위해 기업은 우선 데이터를 통해 해결하고자 하는 비즈니스 이슈가 무엇인지를 정의할 필요가 있다. 데이터를 먼저 수집한 후, 이를 통해 무엇을 분석할 수 있을지 고민하는 것이 아니라, 데이터를 통해 어떤 통찰력을 얻고자 하는지, 누구에게 어떤 가치를 줄 것인지에 대한 명확한 설정이 데이터 수집 전에 선행되어야 한다. 데이터의 사용 목적에 맞춰 필요한 데이터의 종류와 수량을 미리 파악해야 데이터 수집에 필요한 인프라 구축 등에 과잉투자를 방지할 수 있다.

"데이터를 통해 어떤 통찰력을 얻고자 하는지에 대한 목표 설정이 데이터 수집 전에 선행돼야"

>> 데이터 업무 수행 단별에서 선행되는 비즈니스 이슈 정의



Source: 삼정KPMG 경제연구원

비용과 효용을 고려해 데이터 수집 대상과 주기, 기술을 선정

데이터는 사회학에서 통용되는 '양질전환의 법칙'이 적용된다. 양적 축적이 이뤄져야 질적인 변화를 가져올 수 있다는 말이다. 비록 당장은 숫자에 불과한 데이터일 수 있겠지만, 데이터가 축적되었을 때 유의미한 정보가 될 수 있다.

"기업은 데이터 수집에 필요한 비용과 효용을 고려해 수집할 데이터를 결정하고 수집 주기와 수집 기술을 선정해야"

하지만 현실에서는 기업의 제한된 자원으로 인해, 무한정으로 데이터를 늘릴 수 없다. 이에 따라 기업은 데이터 수집에 필요한 비용과 효용을 고려해 수집할 데이터를 결정하고 수집 주기를 설정할 필요가 있다. 실시간 스트리밍 데이터를 수집하기 위해서는 이를 저장하고 효율적으로 처리할 수 있는 IT시스템이 갖춰져야 하고, 배치 데이터의 경우 어느 주기로 데이터를 수집할 것인지에 대한 사전 설계가 필수적이다. 또한 데이터 유형에 따라 어떤 데이터 수집 기술을 활용하는 것이 효율적일지도 사전에 파악해야 한다.

>> 데이터 유형에 따른 수집 기술

분류	데이터 유형	수집 기술
정형 데이터	텍스트 데이터, 엑셀 스프레드시트	ETL, FTP, Open API
반정형 데이터	JSON/XML 포맷 데이터, HTML, 웹문서, 웹로그, 센서 데이터	Crawling, RSS, Open API, FTP
비정형 데이터	SNS, 이미지, 음성·영상 데이터	Crawling, RSS, Open API, FTP, Streaming

Source: 한국정보화진흥원, 삼정KPMG 경제연구원

기업 운영 혁신을 위한 데이터 과학: 기업의 활용 방안

데이터 수집 관련 규정 준수 및 데이터 보안에 유의

전 세계적으로 개인정보보호에 대한 규제와 처벌이 강화되고 있는 추세다. 2019년 9월 미국 연방거래위원회(FTC)는 아동의 개인정보를 부모 동의 없이 불법적으로 수집한 유튜브에 1억 7,000만 달러의 벌금을 부과했다. 같은 해 9월 페이스북은 개인정보 대량 유출 사건으로 50억 달러 규모의 벌금을 물게 됐는데, 이는 미국 FTC 명령 위반을 사유로 책정된 벌금 가운데 최대 규모다. 더불어 2018년 유럽에서 GDPR이 발효된 데 이어, 2020년 1월부터는 미국 캘리포니아 주에서도 새로운 소비자 프라이버시 법인 CCPA(California Consumer Privacy Act)가 도입된 바 있다.

여러 국가에서 사업을 하고 있는 기업들은 국가별 상이한 데이터 수집 및 관리 규정을 준수하고 있는지 주기적으로 모니터링 해야 한다. GDPR의 경우, 유럽연합(EU) 내 거주자의 개인정보를 처리하는 기업이라면 EU 외부에 소재한 기업이라도 GDPR 위반으로 처벌을 받을 수 있으므로, 규제의 대상이 되지 않기 위해 면밀히 대비해야 한다.

> “암묵지가 다양한 형태로 데이터화 되고 있는 가운데, 기업은 수집한 데이터가 유출되지 않도록 정보 보안에 신경써야”

아울러, 데이터의 수집 기술의 발전으로 기업이 축적한 암묵지가 다양한 형태의 문서로, 데이터로 전환되고 있다. 기업은 수집한 데이터가 유출되지 않도록 정보 보안에 신경을 써야 한다. 제품 개발 단계의 상세한 설계 정보가 담긴 CAD(Computer Aided Design)나 제품의 세부 부품별 원가나 구매처의 정보까지 담겨진 BOM(Bill of Material)의 유출은 기업 운영상 큰 문제를 불러일으킬 수 있어 기업들은 수집한 데이터 보안에 각별히 유의해야 한다.

데이터의 양적·질적 측면을 고려해 데이터 품질 관리

‘쓰레기를 넣으면 쓰레기가 나온다(Garbage in, Garbage out)’는 말이 있듯이, 기업이 아무리 많은 양의 데이터를 수집하고 좋은 시스템을 갖추고 있더라도 수집한 데이터가 엉망일 경우 무용지물이다. 데이터 수집 과정에서 기업은 양적인 측면에 집중하는 경향이 있는데, 데이터의 양과 질 모두를 고려해야 한다.

> “데이터 사이언티스트와 데이터 수집 단계부터 함께 하여 체계적인 데이터 수집 체계를 만들어야”

기업은 정확하지 않거나 정합성이 떨어지는 데이터를 수집하지 않기 위해 데이터의 품질 관리에도 신경을 써야 한다. 그러기 위해서는 데이터 사이언티스트와 같은 전문 인력을 분석 단계뿐만 아니라, 데이터 수집 단계에도 투입해 체계적인 데이터 수집 체계를 만들어야 한다.

기업 운영 혁신을 위한 데이터 과학: 기업의 활용 방안

2. 데이터 저장

데이터 저장 과정에서의 이슈사항

기업들은 한정된 저장 공간에서 기하급수적으로 늘어나는 데이터를 저장하고 처리하는 데 어려움을 겪곤 한다. 온프레미스(on-premise)로 IT 인프라를 구축한 일부 기업들은 예상치 못한 데이터 트래픽으로 서버 과부하에 걸리고 전체적인 시스템 기능 장애를 겪는다. 이런 상황에서, 기업들은 데이터의 중요도가 아닌 기업의 유휴 저장 공간에 따라 데이터의 보관 주기를 결정한다. 그러다 보면 자연스럽게 데이터 유실이 발생하고, 나중에 데이터를 찾고자 할 때 데이터가 없는 경우가 발생하게 된다.

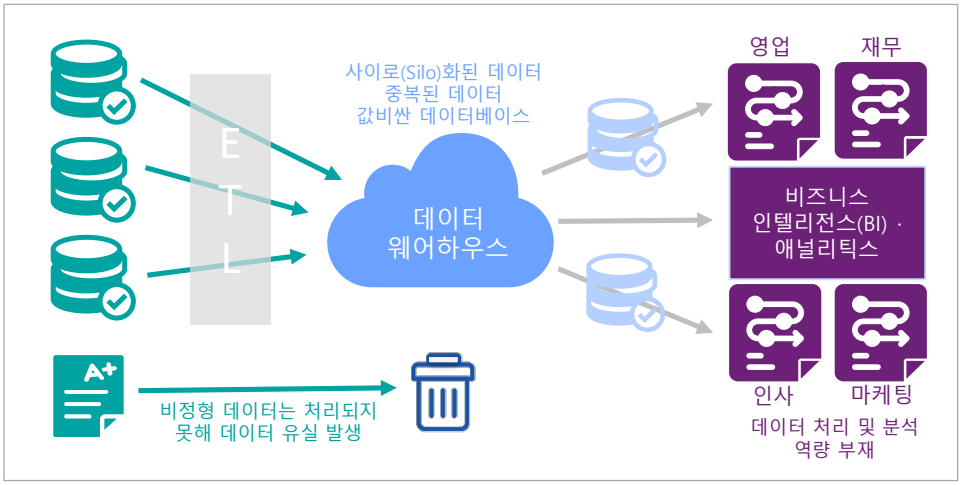
> “기업들은 한정된 저장 공간에서 늘어나는 데이터를 저장하는 데 어려움을 겪고 있어”

둘째는 데이터의 저장 위치가 분산되어 있다는 점이다. 각 부서에서 어떠한 데이터가 수집되고 있는지 파악하기가 어려우며, 일부 데이터는 시스템에 저장되지 않고 개인PC에 저장되거나 메일로 공유되곤 한다. 데이터 거버넌스 체계가 확립되지 않았거나 데이터 오너십이 분산되어 있을 경우, 여러 부서에서 수집되고 저장되는 데이터를 효율적으로 관리하는 데 한계가 있다.

마지막으로, 소셜미디어, 영상, 이미지 등 비정형 데이터를 저장할 수 있는 IT 환경이 마련돼 있지 않다는 점이다. 데이터 웨어하우스(Data Warehouse)는 여러 부서와 시스템에 산재되어 있는 데이터를 한 곳에 모은다는 취지에서 등장했다. ERP, CRM, SCM 등 기업에서 활용하는 다양한 IT시스템에서 생성되는 데이터를 데이터 웨어하우스에 담아두고, 분석이 필요할 때, 데이터에 접근해 활용하자는 접근법이었다. 하지만 최근 데이터의 종류가 다양해지면서, 데이터 웨어하우스의 한계점이 부각되고 있다. 가령 데이터 웨어하우스에 데이터를 저장할 때에는 사전에 정의된 스키마(Schema)대로 데이터를 변환하는 과정이 필요하다. 사전에 정의된 형태로 정제될 수 없는 비정형 데이터는 데이터 웨어하우스에 저장될 수가 없어 데이터 손실이 발생하게 된다.

> “사전에 정의되지 않은 비정형 데이터는 데이터 웨어하우스에 저장될 수가 없어 데이터 손실이 발생”

>> 데이터 웨어하우스의 아키텍처와 한계점



Source: 한국데이터진흥원, 삼정KPMG 경제연구원

기업 운영 혁신을 위한 데이터 과학: 기업의 활용 방안

데이터 저장을 위한 기술·기법 인에이블러

① 클라우드 플랫폼을 통한 혁신 환경 구축

기업이 클라우드로 이전함으로 인해 얻을 수 있는 이점 중 하나는 비용 절감이다. 필요한 만큼 클라우드 서비스를 이용하고 나중에 비용을 지불하는 페이고(Pay-as-you-go) 방식으로 기업은 사전에 비즈니스에 필요한 리소스를 면밀하게 계획하거나 예측할 필요가 없다. 둘째로, 기업은 클라우드 서비스를 통해 민첩성(Agility)을 확보할 수 있다. 클라우드를 활용할 경우, 가상 서버를 쉽게 확장하거나 축소할 수 있어 IT 인프라를 탄력적으로 운영할 수 있다. 한 예로, 알리바바는 2019년 광군제에서 초당 54만 건의 주문을 처리했는데, 평소의 100배가 넘는 트래픽을 처리할 수 있었던 배경에는 알리바바의 클라우드 서비스가 있었다고 한다.

> "클라우드를 활용할 경우, 가상 서버를 쉽게 확장하거나 축소할 수 있어 IT 인프라를 탄력적으로 운영할 수 있어"

마지막으로, 클라우드 기업들은 데이터의 저장 및 처리를 넘어 데이터의 수집부터 분석·활용 단계에 이르기까지 기업이 이용할 수 있는 다양한 애플리케이션을 제공한다는 점이다. 가령 기업은 AWS(Amazon Web Service)의 SageMaker나 구글의 AutoML과 같이 기능을 통해 손쉽게 머신러닝 분석 알고리즘을 자사에 적용하고 업무 혁신을 이룰 수 있다. 실제로, 인공지능, 사물인터넷, 블록체인 등 개별적으로 발전해오던 기술들은 클라우드란 플랫폼을 중심으로 재편되고 있는 모습을 보이고 있다.

> "클라우드 기업들은 데이터의 수집부터 분석·활용 단계에 이르기까지 다양한 애플리케이션을 제공"

해외의 대표적인 상용 클라우드 기업으로는 아마존과 마이크로소프트, 구글을 꼽을 수 있다. 아마존은 AWS란 이름으로 2006년 퍼블릭 클라우드 시장을 개화했고, 2010년에 마이크로소프트가 애저(Azure)를, 2011년에는 구글이 구글 클라우드(Google Cloud)를 출시하면서 클라우드 시장에 뛰어들었다. 현재 글로벌 시장점유율 1위인 아마존은 초기진입자의 우위를 바탕으로 고객에게 높은 비용 혜택과 편의성을 제공하고 있으며, 마이크로소프트는 기업들과 솔루션 개발에 직접 참여하는 등 파트너십 전략을 강화하고 있다. 구글은 자사가 축적한 인공지능 등 소프트웨어 기술력을 고객들이 손쉽게 이용할 수 있도록 하는 데 집중하고 있다.

>> 주요 클라우드 기업의 사업 전략 비교

기업명	사업 전략
aws	▪ 비용 절감과 고객 만족도 및 편의성에 초점을 맞춤 ▪ 우수한 기술력 지원 등에 역점을 두고 있음
Microsoft Azure	▪ 단순히 기술 공급에 그치지 않고 솔루션 차원에서 수요 기업의 어려움을 해결하는 데 초점을 맞춤 ▪ 맞춤형 서비스를 통한 파트너 확대에 초점을 맞춘 전략
Google Cloud	▪ 클라우드 TPU와 같은 하드웨어와 오토ML과 같은 소프트웨어 모두에서 산업 특화된 솔루션 제공 ▪ 축적된 AI 역량을 통해 수요 기업들의 AI 접근성을 높이고 있음

Source: KIET, 삼정KPMG 경제연구원

기업 운영 혁신을 위한 데이터 과학: 기업의 활용 방안

> 데이터 레이크란 조직의 모든 원시 데이터를 거대한 단일 저장소에 원본 그대로 저장하는 것을 의미

② 데이터 웨어하우스에서 데이터 레이크로

일반적으로 기업에서 사용하는 데이터베이스는 여러 개의 행과 열로 구성된 관계형 데이터베이스(RDB, Relational Database)다. 관계형 데이터베이스로 저장하기 위해서는 사전에 정해둔 데이터 스키마대로 데이터를 정제·처리하고 구조화해야 한다. 하지만 수집한 데이터를 저장할 때부터 목적에 따라 다르게 정리하면, 사용 목적이 바뀌었을 때 유연성이 떨어질 수 있다는 한계가 있다.

최근에 수집되는 많은 데이터가 구조화될 수 없는 비정형 데이터의 형태를 갖게 되면서 데이터 레이크(Data Lake)가 차세대 데이터 웨어하우스로 부상하고 있다. 데이터 레이크란 조직의 모든 원시 데이터를 거대한 단일 저장소에 모으는 것을 의미한다. 데이터 레이크에서는 저장될 때에는 사용자가 정해진 포맷에 맞춰서 저장하는 방식이 아니라, 데이터 본래의 형식과 스키마에 따라 저장하는 방식을 채택하고 있다. 저장되는 객체에 공통적인 스키마를 강제하는 대신에 데이터가 읽힘과 동시에 처리되는 스키마 읽기(Schema-on-read) 방식이 적용된다. 데이터 분석이 필요한 부서나 담당자들이 언제든지 데이터 레이크에서 데이터 원본을 끌어다 쓸 수 있어 유연성을 확보할 수 있다. 또한 데이터 레이크에는 변경 불가능한 다양한 형태의 데이터가 타임스탬프가 부여되어 저장되기 때문에, 특정 시점의 기업의 객관적인 상황을 파악할 수 있다.

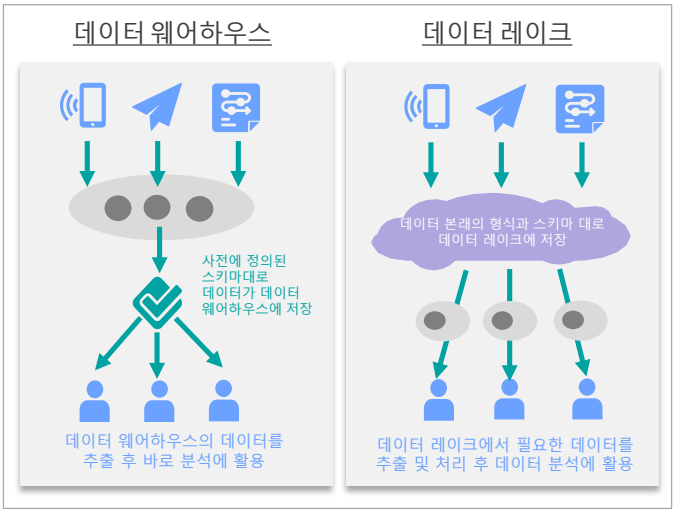
데이터 레이크를 어디에 구축할 것인가 또한 중요한 이슈다. 하둡(Hadoop)과 같은 분산 처리 기술의 발전으로 폭증하는 데이터를 빠르게 저장하고 처리하는 것이 과거에 비해 쉬워진 것은 사실이다. 하지만 그렇다고 해서 무한대로 하둡 노드를 늘릴 수가 없기 때문에, 온프레미스 방식보다는 퍼블릭 클라우드나 프라이빗 클라우드에 데이터 레이크를 구축하는 사례가 늘고 있다.

>> 데이터 웨어하우스와 데이터 레이크 비교

구분	데이터 웨어하우스	데이터 레이크
볼륨	테라바이트 규모	페타바이트 규모
스키마	Schema on Write	Schema on Read
데이터 유형	내부·정형 데이터	외부·비정형 데이터
적재 처리	배치, 준실시간	배치, 준실시간, 스트리밍
거버넌스	성숙단계	초기단계
조인	복잡한 조인, 다중테이블 조인	단순 조인, 풀테이블 스캔

Source: 데이터스트림즈

>> 데이터 웨어하우스와 데이터 레이크 개념



Source: Martin Fowler, 삼정KPMG 경제연구원 재구성

기업 운영 혁신을 위한 데이터 과학: 기업의 활용 방안

"시그널 리포지토리(Signal Repository)는 KPMG가 오픈소스로 수집해 보유하고 있는 데이터 저장소"

KPMG의 데이터 저장 플랫폼 '시그널 리포지토리'

기업의 내·외부 데이터를 하나의 데이터 저장소에 모아 데이터에 기반한 의사결정이 중요해지고 있는 가운데, KPMG의 Lighthouse*는 데이터 저장 플랫폼인 '시그널 리포지토리(Signals Repository)'를 상품화했다. 시그널 리포지토리는 KPMG가 오픈소스로 수집해 보유하고 있는 데이터 저장소로, 이를 기업의 내부 데이터와 연결하여 각종 분석에 활용할 수 있다. 시그널 리포지토리를 활용할 경우, 기업이 시장의 변화를 객관적이고 신속하게 파악해 민첩하게 대응할 수 있는 장점이 있다. KPMG의 시그널 리포지토리는 현재 수요 예측부터, 영업·마케팅, 리스크 관리, 인사 관리까지 다양한 영역에서 활용되고 있다.

한 예로는 기업의 리테일 매장의 입지 경쟁력을 분석하고 최적의 후보지를 선별하는 것이다. 적지 않은 리테일 기업들은 신규 점포 후보지와 기존 점포의 이전 후보지를 찾을 때 많은 시간을 정보 탐색에 들인다. KPMG의 시그널 리포지토리를 활용할 경우, 지역의 인구 통계부터 교통 편의성, 주변 매장의 매출, 임대료, 범죄 통계 데이터 등 후보지의 매출에 영향을 줄 수 있는 수천 개의 신호를 식별해내고 최적의 후보지를 선별할 수 있다. 더 나아가 객관적인 데이터를 바탕으로 점포 후보지의 매출이 입점 시부터 일별, 월별, 향후 3년간까지도 어떻게 될 것인지도 예측할 수 있다.

** KPMG Lighthouse는 회계감사, 세무, 딜, 컨설팅 등 KPMG의 모든 비즈니스 영역에서 이용되는 데이터 분석, 인공지능 및 기타 데이터 기반기술 능력을 배양할 목적으로 설립된 글로벌 조직*

>> KPMG의 사그널 리포지토리 활용 가능 영역

Source: 삼정KPMG

기업 운영 혁신을 위한 데이터 과학: 기업의 활용 방안

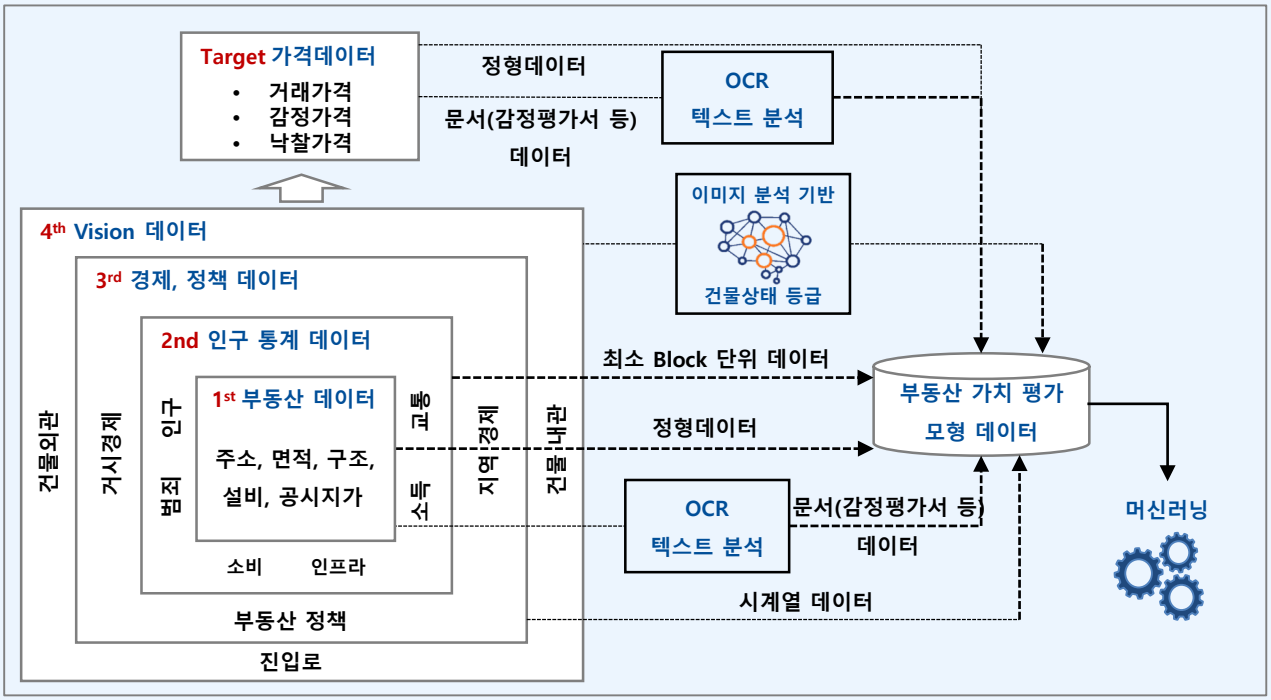
시그널 리포지토리는 미래의 수요를 예측하는 데에도 활용될 수도 있다. 기업이 얼마나 정확하게 수요를 예측하는지는 상품의 보충 계획, 투입 인력 계획, 투자 계획과도 직결되는 매우 중요한 사안이다. KPMG의 시그널 리포지토리를 활용할 경우, 지역 행사와 지역 환경 데이터를 종합해 단기부터 중장기적 수요를 보다 정확하게 예측할 수 있다. 더 나아가, 기업과 경쟁사, 업계 현황, 고객의 행동을 지속적으로 모니터링하고, 비정상적인 변화가 탐지될 시 이에 선제적으로 대응할 수 있도록 하는 데에도 활용될 수 있다.

"시그널 리포지토리를 통해 기업은 한층 더 신속하게 시장의 반응을 파악하고, 민첩하게 대응할 수 있어"

시그널 리포지토리의 또 다른 활용 예로는 데이터를 바탕으로 기업의 퇴직률을 감소시키는 것이다. 일반적으로 퇴사하는 사람들은 일정한 행동 패턴을 보이는데, 기업은 이를 데이터로 포착할 수 있다. 대표적으로 직원의 전화나 이메일을 통한 의사소통 패턴, 회의 참석과 같은 행동기반 신호와 업무 성과, 직원 만족도, 성과급 등의 데이터에서도 파악할 수 있다. KPMG는 시그널 리포지토리를 활용해 퇴직을 염두에 두고 있는 직원의 속성을 파악하고, 직원을 분류해 낼 수 있다. 고성과자 직원의 자발적 퇴직 징후가 감지되었을 때, 기업은 적절하게 개입하여 직원의 이탈을 예방할 수 있다.

마지막으로 활용 가능한 분야로는 데이터 리포지토리에 축적된 데이터를 활용해 부동산 가치 평가 모형을 정교화하는 것이다. 부동산 데이터부터, 인구 통계 정보, 규제와 같은 정책 등 13개 영역에서의 데이터를 부동산의 가치 평가 시 활용할 경우, 기존의 부동산 가치 평가 모형에서는 설명되지 않았던 부분도 설명이 가능해지고 모형의 정확도도 높일 수 있다.

>> KPMG의 시그널 리포지토리를 활용한 부동산 가치 평가 모형



Source: 삼정KPMG

기업 운영 혁신을 위한 데이터 과학: 기업의 활용 방안

③ RDB에서 그래프DB까지, 데이터베이스의 진화

현재 가장 범용적으로 사용되고 있는 데이터베이스는 2차원 테이블로 구성된 관계형 데이터베이스(RDB, Relational Database)다. RDB는 오랜 기간 검증되고 보장된 기술로 대표적으로 Oracle과 MySQL이 있다. RDB는 테이블마다 사전에 정의된 스키마대로 데이터가 저장되어 데이터 품질 측면에서 탁월하며 쿼리로 원하는 데이터를 추출하는 데 용이한 측면이 있다. 하지만 RDB는 테이블 간 관계를 매핑하는 작업이 복잡하다는 점과 대규모의 비정형 데이터를 RDB로 처리하기 어려워 확장성 측면에서 제약이 있다는 점이 한계로 꼽힌다.

"NoSQL은 데이터 구조를 사전에 확정하지 않아 다양한 형태의 데이터 저장이 가능"

이에 대한 대안으로 등장한 것이 대용량 데이터의 분산 처리가 가능한 비관계형DB NoSQL이다. NoSQL은 'Not only SQL'의 약자로 SQL(Structured Query Language)뿐만 아니라 부가적인 기능도 지원한다는 의미를 지닌다. NoSQL은 데이터 구조를 사전에 확정하지 않아 다양한 형태의 데이터 저장이 가능하며 구조 변경도 용이하다. 대표적으로 몽고DB, Cassandra, Hbase 등이 있다.

최근에는 데이터 간 복잡한 관계를 그래프 형태로 저장하고 표현하는 그래프 데이터베이스(Graph Database)또한 새롭게 주목을 받고 있다. 가트너는 2019년 10대 데이터 및 분석 기술 트렌드로 그래프 애널리틱스(Graph Analytics)를 꼽은 바 있다. 성능 개선을 위한 연구가 아직 더 필요한 상태지만 아마존의 넵튠(Neptune)과 같이 그래프 구조를 저장하고 표현하기 위한 도구들이 개발되고 있는 중이다.

"데이터 간 복잡한 관계를 그래프 형태로 저장하고 표현하는 그래프 데이터베이스가 주목을 받고 있어"

그래프의 표현 방법으로는 객체를 표현하는 노드(node)와 노드 간 관계를 나타내는 간선(edge)이 있다. 시각적인 그래프를 통해 데이터베이스를 조망할 경우, 각 개체 간의 관계와 객체의 활동을 직관적으로 파악할 수 있다. 한 예로, 그래프 데이터베이스를 이용할 경우 각 개체의 구매 내역, SNS상 활동, 친구의 구매 내역까지 무한하게 데이터를 확장할 수 있으며, 이들의 관계를 그래프로 표현할 수 있다. 한 예로, 자금 흐름을 파악해 조세 회피 내역을 찾거나 불법 상품·마약 등 밀수 범죄를 단속하기 위해 그래프 데이터베이스가 도입되고 있는 추세다.

>> 그래프 데이터베이스의 개요

정의	그래프 이론의 기본 개념을 바탕으로 객체를 표현하는 노드(node)와 노드 간 관계를 나타내는 간선(edge)으로 단순화해 표현한 데이터베이스
목적	▪ 각기 다른 시스템으로부터 연결 패턴을 찾아 분석을 수행하기 위함 ▪ 연결 관계가 복잡한 시스템으로부터 직관적인 분석을 수행할 수 있음
활용법	관계 정보 기반의 탐색 또는 분석을 통한 의사결정

Source: Bitnine, 삼정KPMG 경제연구원 재구성

기업 운영 혁신을 위한 데이터 과학: 기업의 활용 방안

비즈니스 고려사항

총소유비용을 고려한 클라우드 마이그레이션 여정 설계

데이터를 어느 저장 공간에 어떤 방식으로 저장할 것인지는 기업의 중요한 의사결정 중 하나다. 퍼블릭·프라이빗 클라우드부터, 멀티·하이브리드 클라우드까지 기업의 선택 옵션이 다양해진 오늘날, 기업의 의사결정은 한 층 더 어려워진 상황이다. 더불어, 수년째 사용하던 기존의 레거시 시스템을 클라우드로 이전(Migration)하는 것 또한 쉽지 않다.

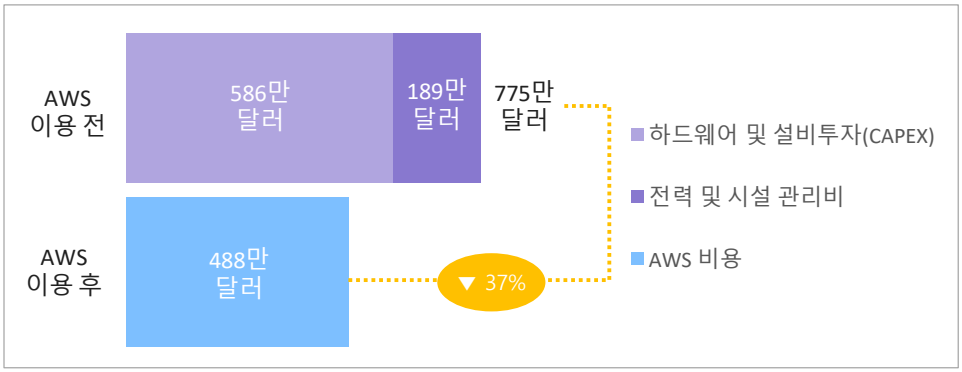
"기업은 총소유비용을 고려해 자사의 비즈니스에 적합한 클라우드 마이그레이션 전략을 마련해야"

이런 상황에서 기업은 총소유비용(TCO, Total Cost of Ownership)을 고려해 자사의 비즈니스에 적합한 클라우드 마이그레이션 전략을 마련할 필요가 있다. 클라우드로의 이전은 단순히 시스템 상의 문제가 아니다. 기업의 전략부터 업무 프로세스, 조직 구조까지 클라우드를 중심으로 새롭게 재편되어야 한다.

비용 측면만을 놓고 볼 때, 초기 장비 도입 비용은 크지만 일단 한번 구축해놓으면 이후 운영 비용이 크게 발생하지 않는 온프레미스(클라우드와 대비하여, 원격환경이 아닌 자체적으로 보유한 전산실 서버에 직접 설치) 방식이 매월 사용한 만큼 비용이 청구되는 클라우드 방식보다 저렴할 수 있다. 하지만 최근 클라우드 사업자 간 경쟁으로 인해 퍼블릭 클라우드 이용 금액이 낮아지고 있다. 더불어, 클라우드를 이용함으로써 얻을 수 있는 부가적인 관리 비용을 줄일 수 있어, 클라우드의 가격 경쟁력이 높아지고 있는 추세다. 한 예로, 시장조사기관인 IDC가 AWS 이용 고객을 대상으로 실시한 설문 조사에 따르면, 기업들은 AWS를 사용함으로써 평균적으로 37%의 IT인프라 비용을 줄일 수 있었다고 응답했다.

"클라우드를 통해 부가적인 관리 비용을 줄일 수 있어, 클라우드의 가격 경쟁력이 높아지고 있어"

>> AWS 이용 전후 IT인프라 비용 비교 (5년간)



Source: IDC(2019), 삼정KPMG 경제연구원

클라우드는 데이터 기반의 혁신 환경을 제공한다. 물리적 장비 구축 없이 빠르고 손쉽게 IT인프라를 구축할 수 있다는 점, 끊임없이 변화하는 비즈니스 환경 속에서 유연성과 확장성을 지닌 점, 유사 산업의 구축 레퍼런스를 활용할 수 있다는 점 등이 클라우드의 장점으로 꼽힌다.

기업 운영 혁신을 위한 데이터 과학: 기업의 활용 방안

반면, 클라우드 이용 시 시스템에 대한 완전한 통제권을 갖지 못한다는 점과 한번 특정 기업의 클라우드 서비스를 이용하기 시작하면 의존성이 심화되어 다른 업체로의 전환이나 온프레미스 방식으로 귀환가 어렵다는 점은 클라우드의 한계로 지적되고 있다. 기업은 클라우드의 장점과 단점을 비교분석하여 자사의 상황에 적합한 IT인프라를 구축할 필요가 있다.

데이터 계층화를 통한 데이터 생명 주기 관리 필요

"기업이 축적하는 데이터의 속성에 대해 파악한 후 핫데이터부터 콜드데이터까지 계층화해야"

대부분의 정보는 시간이 지남에 따라 무의미해지거나 가치가 희석되기 마련이다. 기업은 분석의 유효성에 따라 데이터의 생명 주기를 판단하고 데이터를 계층화해야 한다. 즉, 데이터가 얼마나 자주 활용되는지, 분석 요건의 시급성 및 유효성에 따라 저장위치를 달리해야 한다. 데이터를 언제 아카이브 영역에 저장할 것인지에 대한 정책을 결정해야 클라우드 환경에서 비용을 줄일 수 있다.

일반적으로 시스템에서 많은 이용자들이 찾는 데이터를 핫데이터(hot data) 또는 웜데이터(warm data)라 부르고, 반대로 호출이 뜸한 데이터는 콜드데이터(cold data)라 부른다. 기업은 자사가 보유한 데이터를 핫데이터부터 콜드데이터까지 계층화해야 한다. 그러기 위해서는 우선 자사가 축적하는 데이터의 속성에 대해서 면밀히 파악해야 한다. 값비싼 주 스토리지에 저장되는 데이터 중에서 아카이브에 담아놓아도 되는 데이터를 판별하는 등 비효율적으로 운영되는 부분을 찾을 필요가 있다.

기업의 비즈니스 성격에 맞는 저장 플랫폼 선정

"데이터 마트와 데이터 라이브러리를 어떻게 구성할 지에 대한 설계가 필요"

기업은 자사의 비즈니스 성격에 적합한 데이터 저장 플랫폼을 선정할 필요가 있다. 모든 기업이 데이터 레이크가 필요한 것은 아니다. 다양한 종류의 데이터를 한 곳에 모으는 것이 필요한 기업도 있는 반면, 정형화된 데이터를 사일로(Silo) 형식으로 모으는 방안이 더 효율적인 기업도 있기 때문이다.

실제로 적지 않은 기업들은 데이터 레이크를 적용하고는, 마치 데이터 늪에 빠진 것과 같다고 어려움을 토로한다. 데이터 레이크를 구축한다고 해서, 데이터 레이크가 기업의 모든 데이터를 자동으로 연결해주지 않는다. 현업 담당자가 데이터 레이크에 쌓인 데이터를 끄집어내고자 할 때, 데이터를 직접 선별하는 과정이 필요하며, 이 과정에서 정제되지 않은 원천 데이터로 인한 데이터 정합성 문제, 여러 소스로부터의 데이터 결합 문제, 동기화 문제 등이 발생한다.

데이터 레이크 환경에서 데이터를 효율적으로 관리하기 위해서 기업은 데이터의 출처와 속성을 명확히 파악하고, 데이터 간의 관계를 데이터맵을 통해 확인할 수 있어야 한다. 분석 요건을 중심으로 구성한 데이터 마트(Data Mart)와 데이터 라이브러리(Data Library)를 어떻게 구성할 것인지에 대한 설계 또한 필요하다.

기업 운영 혁신을 위한 데이터 과학: 기업의 활용 방안

3. 데이터 처리

데이터 처리 과정에서의 이슈사항

가공되지 않은 원시 데이터 자체만으로는 큰 효용 가치를 갖기 어렵다. 원시 데이터를 유용한 정보로 바꾸기 위해서는 데이터 처리라는 과정을 거쳐야 한다. 추출·변환·적재(ETL, Extract, Transform, Load) 등 데이터의 품질을 올리는 일련의 과정을 거쳐야 비로소 현업에서 분석에 활용할 수 있는 데이터가 마련된다. 기업에서 불완전한 데이터를 처리하는 기법으로는 데이터 정제, 통합, 축소, 변환 등이 있다.

>> 데이터 처리 기법

기법	설명
데이터 정제	불완전한 데이터는 채우고, 모순된 데이터는 수정하여 다듬는 작업
데이터 통합	다양하게 나뉘어져 있는 여러 데이터베이스, 파일을 합치는 작업
데이터 축소	일부 데이터만 샘플링하거나 분석 대상 데이터의 차원을 줄이는 작업
데이터 변환	평균값을 구하거나 로그를 씌우는 등 데이터를 정규화 또는 집단화하는 작업

Source: 고려대학교 디지털정보처, 삼정KPMG 경제연구원 재구성

"데이터 분석가들은 업무 시간의 70~80%를 데이터 수집 및 처리에 할애해"

데이터 처리 단계에서 기업이 당면하는 이슈사항 중 하나는 이 단계에서 병목 현상이 발생하고 기업의 의사결정을 지연시킨다는 점이다. 실제로 데이터 분석가들은 업무 시간의 70~80%를 데이터를 수집하고 처리하는 데 할애한다고 한다. 데이터 처리 과정은 시간 소모가 크며 필요에 따라 단순 노동 작업도 요구된다.

"기업은 서로 다른 소스로부터 수집한 데이터를 식별하고 데이터 간 관계를 파악해, 데이터를 정제해야"

데이터의 전처리를 담당하는 IT 부서와 현업 부서 간의 마찰도 간과할 수 없다. 현업 부서에서는 IT 부서에 데이터를 요청하고, 데이터를 받기까지 며칠에서 몇 주까지 걸리기도 한다. 현업 담당자는 일정한 기준에 맞춰 수집한 데이터를 추출하는 데 왜 그렇게 오래 걸리냐고 불만을 토로하곤 한다. 하지만 데이터 속을 자세히 들여다보면, 완벽한 데이터란 존재할 수 없고, 요즘처럼 데이터 수집 창구가 다양할 경우 데이터의 처리 과정은 더 까다롭다.

데이터 처리 단계의 마지막 이슈로는 클라우드 컴퓨팅 기술이 급증하는 데이터 트래픽을 실시간으로 처리하는 데 역부족이라는 점이다. 특히 5G 이동통신이 상용화되고 IoT를 통한 기기·서버 간 데이터 통신량이 폭발적으로 증가하면서 클라우드 컴퓨팅 환경에서 지연이 발생하거나 일시적으로 네트워크가 중단되는 등 기술적 한계가 표출되고 있다.

밀리 세컨드(ms)가 중요한 산업에서 트래픽 전송에서 지연이 발생하면, 그에 따른 비즈니스 영향도 클 수 있다. 제조 현장에서 센서 데이터를 통해 설비의 이상징후를 감지했을 경우, 즉각적으로 기계 작동을 중지하는 등의 제어가 필요하다.

기업 운영 혁신을 위한 데이터 과학: 기업의 활용 방안

VR(가상현실)·AR(증강현실)을 활용한 미디어·콘텐츠 분야에서도 즉각적인 데이터 처리가 이뤄지지 않는 경우 몰입감이 떨어질 수 있다. 더 나아가, 자율주행차에서 순간의 네트워크 지연이나 데이터 전송 및 처리 오류가 발생했을 때에는 치명적인 인명 사고로도 이어질 수도 있다.

클라우드는 복잡한 계산을 수행하는 데 탁월하지만 모든 데이터를 중앙 집중식 클라우드로 올려 처리할 때 지연이 발생할 수 있다는 한계가 있다. 모든 데이터를 클라우드로 내보내서 처리하거나 분석하는 것이 현재 기술로는 현실적이지 않다는 의견이 수렴되면서 클라우드 컴퓨팅의 한계를 보안해줄 엣지 컴퓨팅에 대한 기업의 관심이 높아지고 있다.

> 클라우드 컴퓨팅 환경에서 지연이 발생하거나 일시적으로 네트워크가 중단되는 등 기술적 한계가 존재

>> 클라우드 컴퓨팅 기술의 한계

한계	영역	설명
안전성	자율주행 자동차	순간의 네트워크 지연이나 데이터 전송 및 처리 오류가 치명적인 사고로 이어질 수 있음
	도심항공 모빌리티	
즉시성	연안 석유시추 시설	산업기계 자체가 중앙 서버에서 멀리 떨어진 곳에 위치해 있어 중앙서버와의 연결이 어려움
	사막에 위치한 물 분사 펌프	
	VR(가상현실)	사람의 시청각 반응 능력은 매우 예민하기 때문에 불과 몇 백 밀리 세컨드(ms) 차이만으로도 가상현실 몰입감이 떨어질 수 있음
	AR(증강현실)	
	생체 인식	
효율성	스마트 팩토리	스마트 팩토리에서는 대규모 센서 데이터가 발생하며, 이상징후 감지 시 즉각적인 제어 필요

Source: 삼성뉴스룸, 정보통신정책동향(29권 16호), 삼정KPMG 경제연구원

기업 운영 혁신을 위한 데이터 과학: 기업의 활용 방안

> 엣지 컴퓨팅은 중앙 서버가 아닌 데이터가 발생한 현장에서 실시간 분산 처리하는 기술을 의미

데이터 처리를 위한 기술·기법 인에이블러

① 실시간 데이터 처리를 위한 지능형 엣지 컴퓨팅

폭증하는 데이터를 데이터가 발생한 현장이나 그에 가까운 로컬 디바이스 단위에서 처리하는 엣지 컴퓨팅(Edge Computing)이 주목을 받고 있다. 가트너는 2025년에는 전 세계 기업에서 생성되는 데이터의 75%가 엣지에서 처리될 것으로 예상했으며, 2020년 기업이 주목해야 할 전략 기술 중 하나로 '자율권을 가진 엣지(The empowered edge)'를 선정한 바 있다.

엣지 컴퓨팅은 중앙 서버가 아닌 데이터가 발생한 현장이나 네트워크 종단에서 실시간 분산 처리하는 기술을 의미한다. 엣지 컴퓨팅을 활용할 경우 데이터 처리 부하를 단축시킬 수 있으며, 인터넷 대역폭 사용량도 줄일 수 있다. 비록 클라우드 컴퓨팅에 비해서 연산 능력은 떨어지지만, 데이터를 처리함에 있어 지연시간이 짧고, 해킹 가능성이 낮으며, 광범위한 이동성을 지원한다는 강점이 있다.

엣지 컴퓨팅은 실시간성이 중요한 영역에서 우선적으로 도입되고 있다. 엣지 컴퓨팅은 제조 현장에 도입되어 제품 결함을 예방하고 공정 효율 및 설비 자산의 생산성을 향상시키는 데 활용되고 있다. 또한 차량에 적용된 엣지 컴퓨팅으로 앞차 간 거리를 유지하거나 주변 도로 상황, 차량 흐름 등을 파악하고 돌발상황이 발생했을 때 신속하게 대처할 수 있도록 한다. 그 외에도 엣지 컴퓨팅은 게임 업계에도 도입되어 끊김 없는 게임 플레이가 가능토록 하고 있다.

엣지 컴퓨팅은 단말 장치와 가까운 위치에서 상황을 인지하고, 학습·판단해 대응이 가능한 지능형 엣지 컴퓨팅으로 진화하고 있다. 앞으로도 엣지 컴퓨팅과 클라우드 컴퓨팅은 서로의 부족한 부분을 보완해주는 관계로 발전할 것으로 예상된다.

>> 클라우드 컴퓨팅과 엣지 컴퓨팅 비교

속성	클라우드 컴퓨팅	엣지 컴퓨팅
지연시간	김	짧음
서비스 지역	인터넷	로컬 네트워크
지역식별	불가능	가능
해킹 가능성	높음	낮음
통신 방식	중앙집중식	분산형
서버 수	적음	많음
이동성 지원	제한적	광범위

Source: NxtGen, KB금융지주 경영연구소

기업 운영 혁신을 위한 데이터 과학: 기업의 활용 방안

② 오토 레이블링으로 데이터 처리 간소화

오늘날 대부분의 기계학습은 지도학습(supervised learning) 기법을 사용하고, 이를 위해서는 레이블링이 된 데이터 세트가 필요하다. 레이블링이란 모델에 사용되는 데이터에 정답을 달아놓는 과정을 의미한다. 가령 고양이와 개의 이미지를 분류하는 모델을 만들고자 할 때, 사람이 직접 고양이와 개의 이미지에 정답을 달아 놓아야 인공지능 알고리즘이 이를 학습을 할 수 있다. 좋은 학습 데이터가 모델의 정확도를 높여주므로, 인공지능 분석가들은 많은 시간을 학습 데이터의 품질을 높이는 데 소요한다.

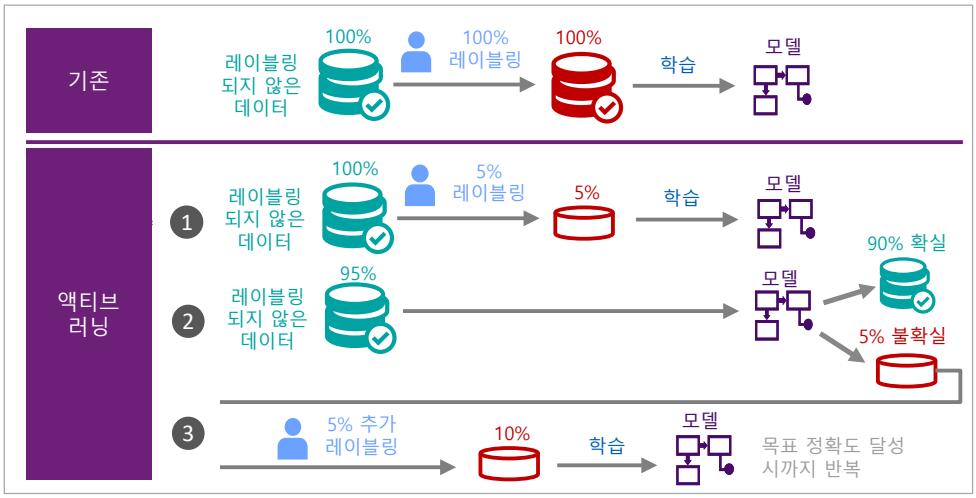
> “좋은 학습 데이터가 모델의 정확도를 높여주므로, 인공지능 분석가들은 많은 시간을 학습 데이터의 품질을 높이는 데 소요”

최근에는 IT 기업들이 제공하는 오토 레이블링 툴로 인해 데이터 레이블링 작업이 간소화되고 있다. 한 예로, 아마존의 세이지메이커 그라운드 트루스(SageMaker Ground Truth)는 레이블링 작업을 위한 워크플로우와 인터페이스를 제공한다. 이 툴을 이용할 경우, 여러 작업자들이 쉽고 편리하게 레이블링 작업을 할 수 있으며 데이터 레이블링에 투입되는 시간과 비용을 줄이고 레이블링 작업의 품질 또한 높일 수 있다. 그라운드 트루스가 제공하는 라벨링 작업은 크게 5가지로 텍스트 분류, 이미지 분류, 물체 감지, 의미론적 분할, 맞춤형 사용자 정의 작업이 포함되어 있다.

> “오토 레이블링 툴을 통해 데이터 레이블링에 투입되는 시간과 비용을 줄이고 레이블링 작업의 품질 또한 높일 수 있어”

삼성SDS의 경우도 액티브 러닝(Active Learning) 방법론을 오토 레이블링 작업에 적용하여 레이블링에 필요한 인력 투입을 줄이고, 분류 결과의 정확성도 높이고 있다. 기존의 방식은 사람이 100% 레이블링을 달아주면, 이를 가지고 학습하여 모델링을 하는 방식이었다. 액티브 러닝 방법론을 적용할 경우, 사람이 전체 데이터의 5%만 선별해 레이블링해주고 이를 바탕으로 1차적으로 학습 모델을 만든다. 그런 후, 학습된 모델을 가지고 나머지 95%의 데이터를 레이블링시키고 명확하지 않은 데이터에만 한해 사람이 추가적으로 레이블링 작업을 해주고 목표 정확도를 달성할 때까지 이 과정을 반복하는 방식이다.

>> 액티브 러닝(Active Learning) 방법론을 적용한 오토 레이블링



Source: 삼성SDS, 삼정KPMG 경제연구원 재구성

기업 운영 혁신을 위한 데이터 과학: 기업의 활용 방안

> 최근에 데이터 손실은 최소화하면서 분석가치를 극대화할 수 있는 암호기반 프라이버시 보호 기술이 부상하고 있어

③ 암호기반 프라이버시 보호 기술로 데이터 비식별화

유럽연합(EU)의 개인정보보호법(GDPR)과 미국 캘리포니아 주의 소비자 개인정보 보호법(CCPA)을 비롯해 전 세계적으로 개인 정보 보호에 대한 규제가 강화되고 있다. 2018년 11월에 ISO(세계표준화기구)는 IEC(국제전기표준회의)와 함께 ISO/IEC 20889(프라이버시를 강화한 비식별 처리 표준)를 발표하고 데이터 비식별 처리를 위한 표준화 기준 마련에 나선 바 있다. 미국 오레곤주과 뉴햄프셔주는 경찰이 바디 카메라를 이용한 안면인식 이용을 금지했고, 미국 샌프란시스코에서도 2019년 5월 경찰을 포함해 모든 행정기관들의 안면인식 기술의 사용금지 법안을 가결했다. 마이크로소프트는 2019년 6월 3년 전부터 연구 목적으로 축적한 얼굴인식용 데이터를 삭제하고 안면인식 기술에 대한 규제를 촉구하고 있다.

일반적으로 개인 정보 비식별 조치를 강화하면 데이터의 손실이 많아져 데이터의 분석 가치가 떨어지는 경향이 있다. 즉, 데이터 보호 장치를 많이 만들수록 데이터의 활용도가 떨어지는 트레이드 오프(Trade-off) 관계를 갖는다. 이에 따라, 기업들은 데이터를 안전하게 관리하면서 동시에, 데이터의 손실을 최소화할 수 있는 방안을 고민을 하고 있다.

개인 정보 비식별 기법으로는 가명처리, 총계처리, 데이터 삭제, 데이터 범주화, 데이터 마스킹 처리 기법 등 총 17개로 세부 기술로 구분될 수 있다. 기술별 효과가 다르고 장단점이 있어 업계에서는 데이터의 특성에 따라 여러 비식별화 기술을 복합적으로 사용하고 있다.

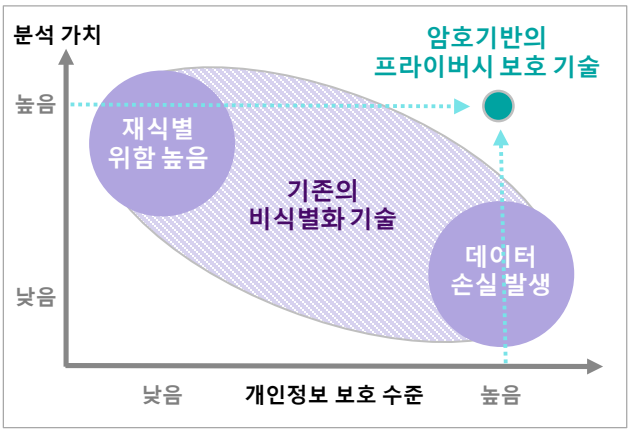
최근에 데이터 손실은 최소화하면서 분석가치를 극대화할 수 있는 암호기반 프라이버시 보호 기술이 부상하고 있다. 마이크로소프트, 구글, 인텔 등 글로벌 기업들은 현재 암호기반 프라이버시 보호 기술을 금융, 의료, 클라우드 등에 도입하고 있으며 국내에서는 삼성SDS가 서울대학교 암호랩과 협업하여 암호기반 프라이버시 보호 기술 중 하나인 동형 암호화 기술을 개발하고 있다.

>> 개인정보 비식별 기법

가명처리	휴리스틱 가명화	데이터 삭제	식별자 삭제
	암호화		식별자 부분삭제
	교환방법		레코드 삭제
총계처리	총계처리		식별요소 삭제
	부분총계	데이터 범주화	감추기
	라운딩		랜덤 라운딩
	재배열		범위 방법
데이터 마스킹	임의 잡음 추가		제어 라운딩
	공백과 대체		

Source: 정보통신기획평가원

>> 암호기반 프라이버시 보호 기술 개요



Source: 삼성SDS

기업 운영 혁신을 위한 데이터 과학: 기업의 활용 방안

동형 암호란 데이터의 손실·유출 없이 암호화된 상태에서 연산 및 분석이 가능하도록 지원하는 기술을 의미한다. 동형암호는 암호화된 상태에서 개인정보를 풀지 않고도 데이터 분석이 가능해 데이터의 유출 위험성을 원천적으로 막을 수 있다.

영상 데이터의 개인정보 보호가 중요해지고 있는 가운데, 기업들은 안전하게 영상을 처리하는 방법으로 영상 데이터의 익명화 기술에도 집중하고 있다. 영상 이미지 데이터를 익명화하는 기술은 영상 내에서 개인을 식별할 수 있는 기술과 탐지한 개인을 변형하는 기술로 구성될 수 있다. 개인 식별 영역을 변형하는 기술로는 일반적으로 이미지 필터링, 이미지 암호화, 얼굴합성, 인페이팅 등이 있다. 최근에는 영상 데이터를 수집한 이후 비식별화하는 것을 넘어, 카메라, CCTV등 디바이스 단에서 익명화 장치를 추가하여 영상을 익명화하고 있다. 수집 단계부터 익명화된 데이터는 데이터 손실 없이 원본과 동일한 분석 결과를 얻을 수 있고 동시에 데이터를 안전하게 보호할 수 있다.

> 영상 데이터의 개인정보 보호가 중요해지고 있는 가운데, 기업들은 영상 데이터의 익명화 기술에 집중하고 있어

>> 영상 이미지 식별 변형 기술

기술	설명	한계점
이미지 필터링	▪ 개인을 식별할 수 있는 영역에 여러 필터를 적용해 개인을 식별하지 못하게 함	▪ 딥러닝 기술의 발달로 필터링을 거친 이미지를 일정 수준 복원이 가능
이미지 암호화	▪ 영상을 암호화하여 허가된 대상에게만 공개	▪ 암호화 기법을 사용하면 연산량이 많아 실시간 영상 처리가 어려움
얼굴 합성	▪ 수집한 얼굴 이미지 집합 내에서만 유사한 얼굴을 합성	▪ 합성한 얼굴에서 원본 얼굴 복원이 불가능하여 유용성을 보장할 수 없음
인페이팅	▪ 영상에서 특정 부분을 제거하고 생긴 공백 또는 손상된 부분을 채우는 기법	▪ 부자연스럽게 복원한 경우, 영상 데이터의 유용성이 하락할 수 있으며 연산량이 많아 실시간 처리가 어려움

Source: K-ICT 빅데이터센터, 정보통신기획평가원, 삼정KPMG 경제연구원 재구성

기업 운영 혁신을 위한 데이터 과학: 기업의 활용 방안

비즈니스 고려사항

설계한 분석 요건에 맞도록 데이터 정제

기업은 데이터 분석 전, 기업이 보유한 데이터의 상태를 확인하고 설계한 분석 요건에 맞도록 데이터를 정제해야 한다. 첫 번째 단계로, 기업이 보유한 데이터의 유형(범주형, 연속형)과 데이터 타입(Character, Numeric, Date)을 확인해야 한다. 데이터의 결측치나 이상치(Outlier)가 있을 경우, 추세선에 영향을 줘 왜곡된 분석 결과를 야기할 수도 있으므로, 데이터 분석가의 판단에 따라 적절히 처리해줘야 한다. 결측치를 처리하는 방법으로는 평균값이나 중앙값, 혹은 회귀분석을 통해 예측한 값을 넣는 경우도 있다. 인적 오류로 인해 생긴 이상치의 경우 분석 대상에서 제외하거나, 자연발생적인 이상값의 경우, 튀는 값을 갖게 된 원인을 파악해보는 것이 필요하다. 마지막으로 데이터의 표현 방식을 정하는 것이 필요하다. 변수 간 관계가 잘 드러나지 않을 때 변수를 변환하는 방법으로 로그 함수를 취하거나, 범주형 변수로 만들거나 더미변수화하는 방법 등이 있다.

"데이터의 결측치나 이상치가 있을 경우, 추세선에 영향을 줘 왜곡된 분석 결과를 야기할 수도 있어"

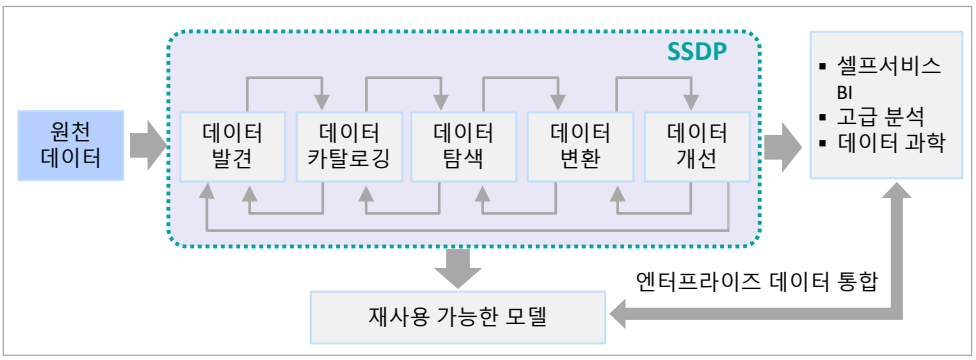
신뢰도가 낮은 데이터를 바탕으로 의사결정을 할 경우, 기업의 경쟁력 저하는 물론 사회적·경제적 비용까지도 발생시킬 수 있다. 인공지능으로 분석을 할 경우에도, 어떤 데이터로 모델을 학습시키는지에 따라 결과물이 180° 달라질 수 있다. 따라서 기업은 서로 다른 소스로부터 수집한 데이터를 식별하고 데이터 간 관계를 파악하고 데이터 정제 과정을 주의 깊게 진행할 필요가 있다.

데이터 처리를 간소화, 자동화, 지능화하는 SSDP 도구 활용

"SSDP 도구는 데이터 탐색, 통합, 카탈로깅, 정제, 변환 등의 업무를 자동화"

데이터를 수집하고 처리하는 업무는 더 이상 IT 엔지니어만의 영역이 아니다. 데이터를 다룰 수 있는 역량이 직원이나 부서별로 상이한 가운데, SSDP(Self Service Data Preparation) 도구는 현업 담당자가 데이터 업무를 하는 데 있어 상당 부분을 자동화해준다. SSDP 도구를 활용할 경우 IT부서의 개입 없이도 데이터 탐색, 통합, 카탈로깅, 정제, 변환 등의 업무를 현업 담당자가 직접 손쉽게 할 수 있다.

>> SSDP(Self Service Data Preparation) 개요



Source: LG CNS

기업 운영 혁신을 위한 데이터 과학: 기업의 활용 방안

4. 데이터 분석

"기업은 객관적인 분석을 통한 데이터 경쟁력의 진단과 지속적인 데이터 전략의 이행이 필요"

데이터 분석 과정에서의 이슈사항

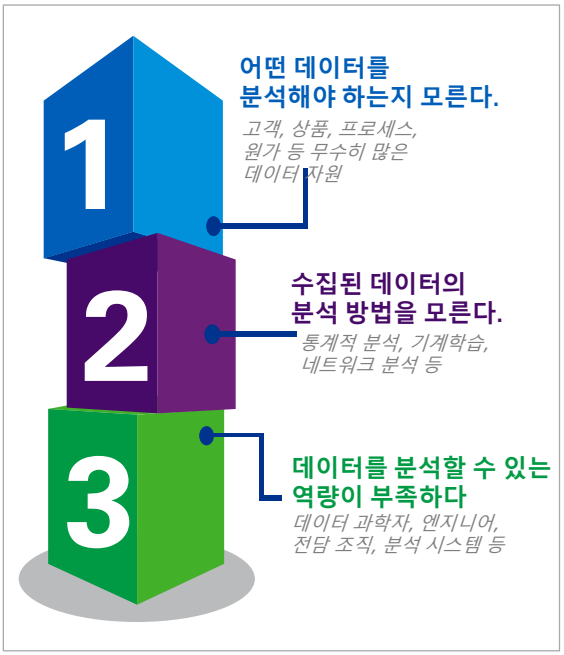
앞서 살펴본 가이드라인을 따라 데이터 자원의 수집, 저장 및 처리가 원활하게 진행되었다고 하더라도, 데이터의 분석이 유연하게 이루어지지 않으면 자원의 낭비가 불가피하다. 전 세계 144개 기업의 정보 책임자를 대상으로 조사한 2015년 KPMG의 데이터 및 애널리틱스 설문 보고서에 따르면, 기업들은 데이터 분석 경쟁력이 낮은 이유에 대해 '어떤 데이터를 분석해야 하는지 모른다', '데이터를 분석할 수 있는 역량이 부족하다', '수집된 데이터의 분석 방법을 모른다' 등으로 응답했다.

실제로 기업의 데이터 분석 경쟁력이 약화되는 요인은 크게 세 가지이다. 첫 번째로, 고객, 상품, 프로세스, 원가 등 기업의 경영 활동에서 무수히 발생하는 데이터 자원 중 어떤 데이터를 분석해야 하는지 모르는 것이다. 예를 들어, 고객관계 강화를 위해서 고객의 어떤 데이터를 분석해야 하는지, 장비의 이상 징후 판별을 위해서 어떤 데이터를 다루어야 하는지 모르는 경우가 해당된다.

두 번째로, 어떤 데이터를 분석해야 하는지 알고 있지만 분석 방법을 모르는 것이다. 예를 들어, 생산 품질 관리를 위해 수집한 라인의 이미지 데이터를 일반적인 규칙 기반 알고리즘으로 분석하고자 하는 경우가 해당된다.

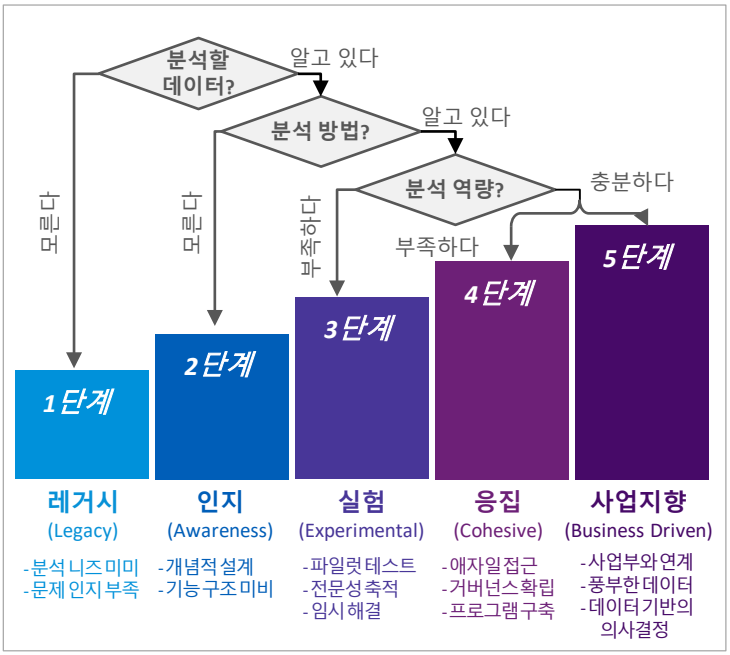
세 번째로, 보유하고 있는 데이터 자원도 풍부하고 분석 방법도 알고 있으나, 사내에 이를 수행할 수 있는 역량이 축적되어 있지 않은 경우이다. 일반적으로 데이터 사이언티스트(Data Scientist)의 부재는 데이터 가치 창출의 주요 저해 요인 중 하나이다.

>> 기업의 데이터 분석 경쟁력이 낮은 이유



Source: 삼정KPMG 경제연구원

>> 데이터 분석 경쟁력 강화 진단 트리



Source: 삼정KPMG 경제연구원

기업 운영 혁신을 위한 데이터 과학: 기업의 활용 방안

데이터 분석을 위한 기술·기법 인에이블러

데이터를 분석하는 목적은 크게 두 가지로 요약할 수 있다. 하나는 과거에 일어난 현상에 대한 인과성을 밝혀내어 기업이 활용할 수 있는 형태의 지적 자산으로 축적하는 것이고, 다른 하나는 불확실한 미래를 예측하여 변화에 선제적으로 대응하는 것이다.

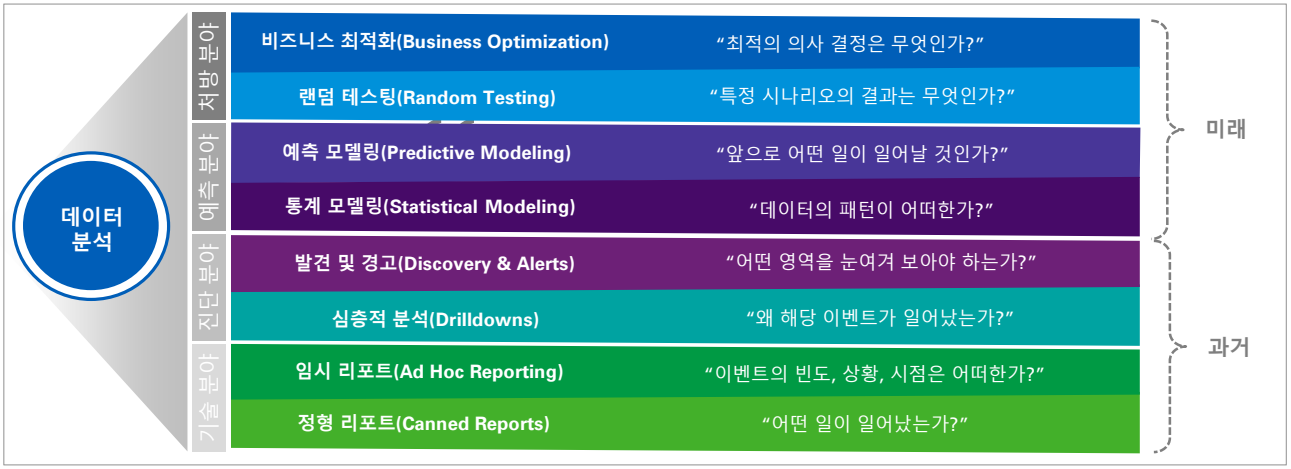
"데이터 분석은 그 목적에 따라 기술 분석, 진단 분석, 예측 분석, 처방 분석으로 분류"

과거에 일어난 현상을 대상으로 하는 분석의 활용 분야로는 기술 분석(Descriptive Analysis)의 영역인 '정형 리포트', '임시 리포트'와 진단 분석의 영역인 '심층적 분석', '발견 및 경고' 등이 있다. 기술 분석의 도구인 '정형 리포트'는 기업의 외부와 내부에서 발생한 이벤트를 일목요연하게 정리하는 보고서로서 고급 분석의 기초 자료가 된다. '임시 리포트'는 작성된 '정형 리포트'를 기반으로 특정 이벤트의 빈도, 영향, 가설적 요인 등을 분석하는 보고서이다.

진단 분석(Diagnostic Analysis)의 주요 활용 분야인 '심층적 분석'은 주로 사건의 전후 관계와 인과 관계를 명확히 하기 위해 수행된다. 예를 들어 기업의 매출이 일시적으로 증가했을 때, 그 원인이 무엇인지 시장 변수를 수집하고 분석하는 활동은 진단 분석의 영역에 포함된다. '발견 및 경고'는 전반적으로 기업이 집중해야 할 분야는 어디인지, 새롭게 개척해야 할 영역은 어디인지, 문제가 발생하는 영역이 어디인지에 대한 분석을 포함하는 진단 분석의 분야이다.

예측 분석(Predictive Analysis)은 주로 수집한 데이터의 패턴을 파악하고 이를 토대로 향후 어떤 일이 일어날 것인지에 대한 시나리오 분석을 의미하며, '통계 모델링'과 '예측 모델링' 등의 분야에 활용될 수 있다. 처방 분석(Prescriptive Analysis)은 예측한 시나리오를 바탕으로 영향도를 산출하여 대응 방안을 제시하거나 최적의 의사결정을 지원하는 등 인공지능을 활용한 융합 분석을 의미하며, '랜덤 테스팅'과 '비즈니스 최적화' 등으로 구분된다.

>> 데이터 분석의 정의 및 활용 분야



Source: 삼정KPMG 경제연구원

기업 운영 혁신을 위한 데이터 과학: 기업의 활용 방안

> 기술 분석은 사건과 현상에 대한 객관적인 해석이며, 진단 분석은 경영 변수의 관계와 패턴에 대한 심층적 이해

① 의사결정의 근간이 되는 기술 분석 및 진단 분석

기술 분석과 진단 분석은 공통적으로 기업의 내부와 외부 현황을 이해하고 향후 일어날 일에 대한 단초를 마련할 수 있도록 돕는다.

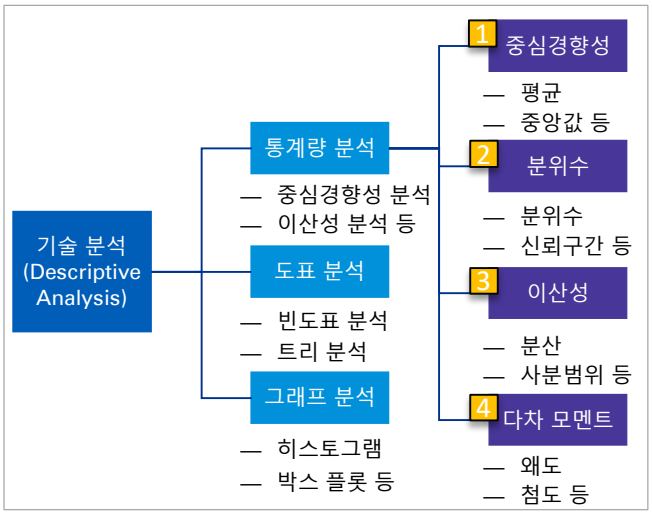
기술 분석은 활용 도구에 따라 크게 통계량 분석, 도표 분석, 그래프 분석으로 나눌 수 있다. 통계량 분석은 가장 일반적으로 쓰이는 데이터 분석 방법론이다. 이는 평균, 중앙값 등을 산출하는 '중심경향성' 분석, 통계량의 신뢰구간과 데이터 프로파일의 상대적 위치를 확인하는 '분위수' 분석, 분산, 사분범위 등 데이터 값의 변화 정도를 측정하는 '이산성' 분석, 수치 분포의 왜도와 첨도를 계산하는 '다차 모멘트' 분석으로 분류할 수 있으며, 경영 데이터 계량 분석의 기초적인 토대가 된다.

기술 분석의 도표 분석은 특정한 사건의 발생 빈도를 관찰하는 빈도표 분석과 경영 지표의 구조화를 통해 인사이트를 도출하는 트리 분석 등으로 구분할 수 있다. 그래프 분석은 빈도표를 시각화한 히스토그램 분석과 데이터의 대략적인 분포 및 개별적인 이상치를 보여주는 박스 플롯 분석 등으로 분류할 수 있다.

진단 분석은 기존 데이터의 연관 관계를 분석하는 심층적 분석과 빅데이터의 패턴을 이해하는 발견 및 경고 분석으로 구분할 수 있다. 심층적 분석 중 '상관 관계 분석'은 피어슨 상관계수, 스피어만 상관계수 등을 통해 변수 간의 연관성을 추적한다. '인과 관계 분석'은 z-테스트, t-테스트 등을 통해 경영 변수 간의 이론적인 인과관계를 실증적으로 증명해내는 방법론이다.

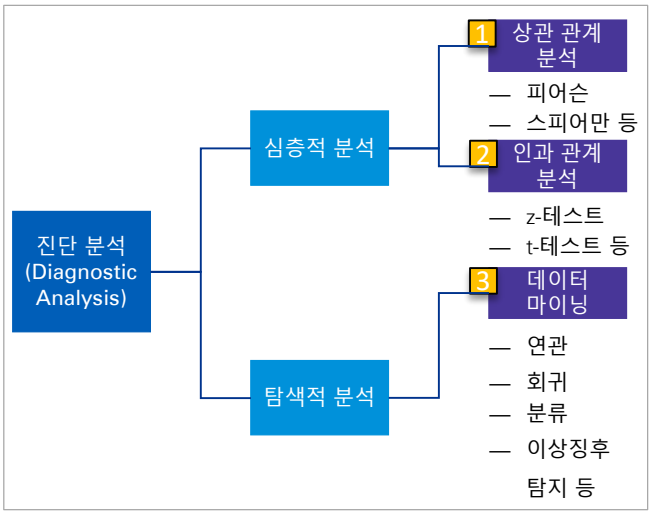
진단 분석의 탐색적 분석은 주로 데이터 마이닝의 기법을 활용하여 빅데이터의 패턴을 이해하고, 이를 통해 새로운 인사이트를 창출하는 방법론이다. 경영 현장에서 기술 분석과 진단 분석의 적절한 조합이 필요하다.

>> 기술 분석의 종류와 기법



Source: 삼정KPMG 경제연구원

>> 진단 분석의 종류와 기법



Source: 삼정KPMG 경제연구원

기업 운영 혁신을 위한 데이터 과학: 기업의 활용 방안

실무에서 기술 및 진단분석을 수행할 때 '일반화의 오류'와 '인과관계 추론의 오류'에 빠지지 않도록 주의해야 한다. 일반화의 오류는 데이터의 인위적 추출로 인한 샘플의 편향성을 간과한 채로 도출한 결론을 여과 없이 받아들이게 되는 현상이다. 인과관계 추론의 오류는 샘플 군집 간의 지표 비교 시 인위적 할당으로 인해 발생한 군집의 대표성 상실을 간과하여 상관관계를 인과관계로 이해하게 되는 현상을 의미한다.

"적절한 샘플링 전략의 채택을 통해 데이터 분석의 비용과 리드타임 감축 가능"

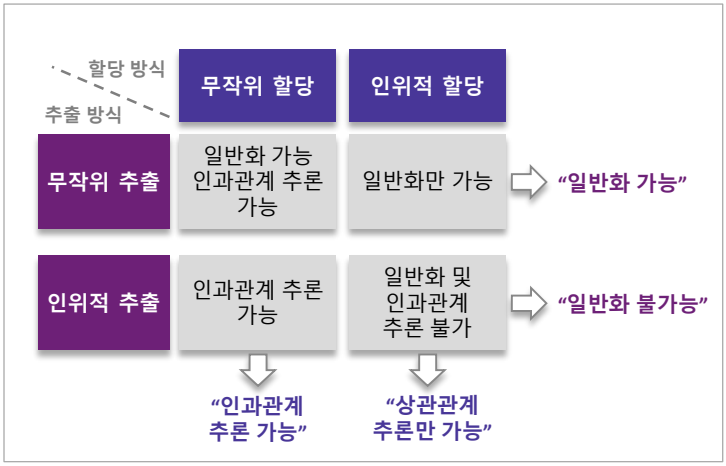
예를 들어, 온라인 액세서리 쇼핑몰의 특정 럭셔리 가방 매출이 급격히 상승한 상황에서 '해당 브랜드의 상품을 전면적으로 배치하고, 럭셔리 브랜딩을 위한 프로모션에 투자를 늘려야 한다'고 결론을 내리는 것은 일반화의 오류를 범하는 것이다. 최근 화제가 되고 있는 럭셔리 가방을 판매하는 온라인 쇼핑몰 중에서 가장 저렴한 가격을 내세우고 있을지도 모르기 때문이다.

또한 동일 쇼핑몰의 온라인 고객을 대상으로 발송한 이메일 쿠폰의 매출 기여 효과를 A/B테스트로 관찰하고자 할 때, 실험군과 대조군이 서로 다른 구매 시점이나 유입 경로를 가진 고객 표본에서 추출되는 경우에 인과관계 추론의 오류에 빠질 수 있다. 이 경우, '해당 온라인 쿠폰 프로모션의 효과로 매출이 3% 증가하는 효과를 얻었다'고 결론내리기 어렵기 때문이다.

일반적으로 데이터 분석 단계에서 무작위 추출과 무작위 할당 방식을 채택하는 경우, 도출된 결론에 대한 일반화가 가능하고 인과관계의 추론도 가능하다. 무작위 추출과 인위적 할당 방식을 채택하는 경우 인과관계의 추론은 불가능하고, 결론의 일반화만 가능하다. 반면 인위적 추출과 무작위 할당 방식을 채택하는 경우, 변수 간의 인과관계 추론은 가능하지만 일반화는 어렵다는 단점이 있다. 인위적 추출과 인위적 할당 방식을 채택하는 경우에는 일반화 및 인과관계 추론 모두 불가능하다.

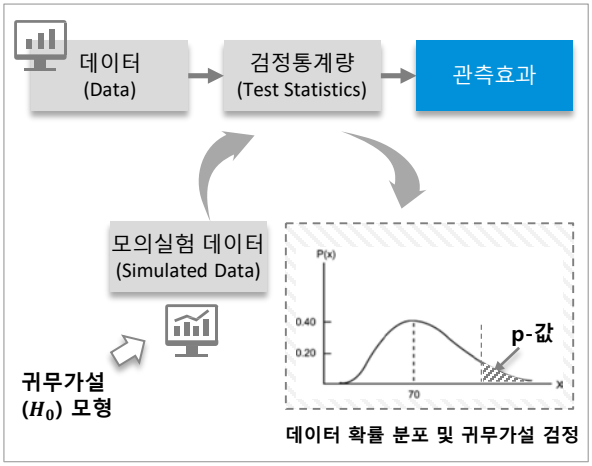
데이터 분석 단계에서, 검정하고자 하는 가설에 대한 적절한 샘플링 전략을 채택하는 것은 비용 효율화와 리드타임 단축의 측면에서 매우 중요하다.

>> 데이터 샘플 추출과 할당 방식에 따른 주요 고려사항



Source: Dr. Mine Çetinkaya-Rundel, Introduction to Probability and Data

>> 일반적인 가설 검정 프로세스



데이터 확률 분포 및 귀무가설 검정

Source: 삼정KPMG 경제연구원

기업 운영 혁신을 위한 데이터 과학: 기업의 활용 방안

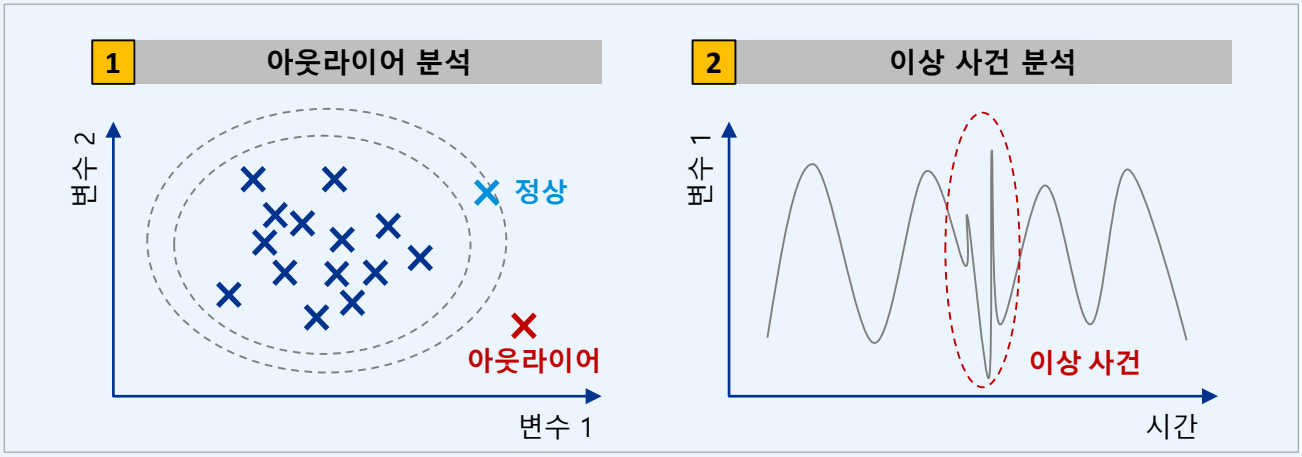
이상 탐지(Anomaly Detection) 분석의 부상

"이상 탐지 분석은 이상 징후를 판별하여 궁극적으로 기업 운영의 리스크를 감축하는 데 기여"

이상 탐지 분석은 데이터의 공간적, 시간적 패턴과 상이한 데이터를 판별하고 활용하여, 궁극적으로 기업 운영의 리스크를 감소시키는 데 그 목적이 있다.

이상 탐지 분석은 크게 경영 변수에 따른 아웃라이어 분석(Outlier Analysis)과 시간의 흐름에 따른 이상 사건 분석(Unusual Event Analysis)의 두 가지 영역으로 구분할 수 있다. 아웃라이어 분석은 고객 성향 분석 등에서 잘못된 결과를 초래하는 데이터를 제거하는 데 사용될 수 있으며, 이상 사건 분석은 공장 제조 라인의 일부 시스템 장애나 이상 동작을 판별하고 금융 범죄를 예방하기 위해 사용될 수 있다.

>> 이상 탐지 분석의 두 가지 영역



Source: 삼정KPMG 경제연구원

"이상 패턴 감지를 위해서 정상 데이터와 비정상 데이터의 특성을 먼저 정의하고, 목적에 맞는 모형을 디자인"

일반적으로 인간이 데이터에서 '이상 패턴'을 발견하기 위해서는 크게 세 가지 방법을 활용한다.

① 과거에는 잘 나타나지 않았던 패턴이 갑자기 빈번하게 발생하는지 확인
② 일반적인 상식으로 함께 나타날 수 있는 현상인지 타당성을 확인
③ 다른 그룹에서는 나타나지 않는 현상이 나타나고 있는지 확인

반면, 데이터 분석 기법을 통해 '이상 패턴'을 감지하기 위해서는 먼저 '정상 데이터'와 '정상적이지 않은 데이터'의 특성을 정의해야 하고, 비정상 데이터를 찾는 모형을 만들어야 한다. 비정상 데이터를 찾는 모형은 데이터의 종류, 형태, 원하는 아웃풋(Output)에 따라 상이한 모습을 갖추게 된다.

예를 들어 이상 탐지 모형에서 레이블링된 데이터가 매우 적거나 존재하지 않는 경우, 지도 학습 등 분류 기반의 알고리즘을 활용하기 어렵다. 또한 시계열 데이터인 경우 확률론적 모형을 사용하는 경우가 일반적이며, 원하는 아웃풋이 스코어링(Scoring) 형태인지, 혹은 판별 형태인지에 따라 모형의 구조가 달라지기도 한다.

기업 운영 혁신을 위한 데이터 과학: 기업의 활용 방안

> 예측 분석은 객관적 사실과 주관적 해석을 통해 불확실한 미래의 방향성을 탐색

② 경영 활동의 불확실성을 낮추는 예측 분석

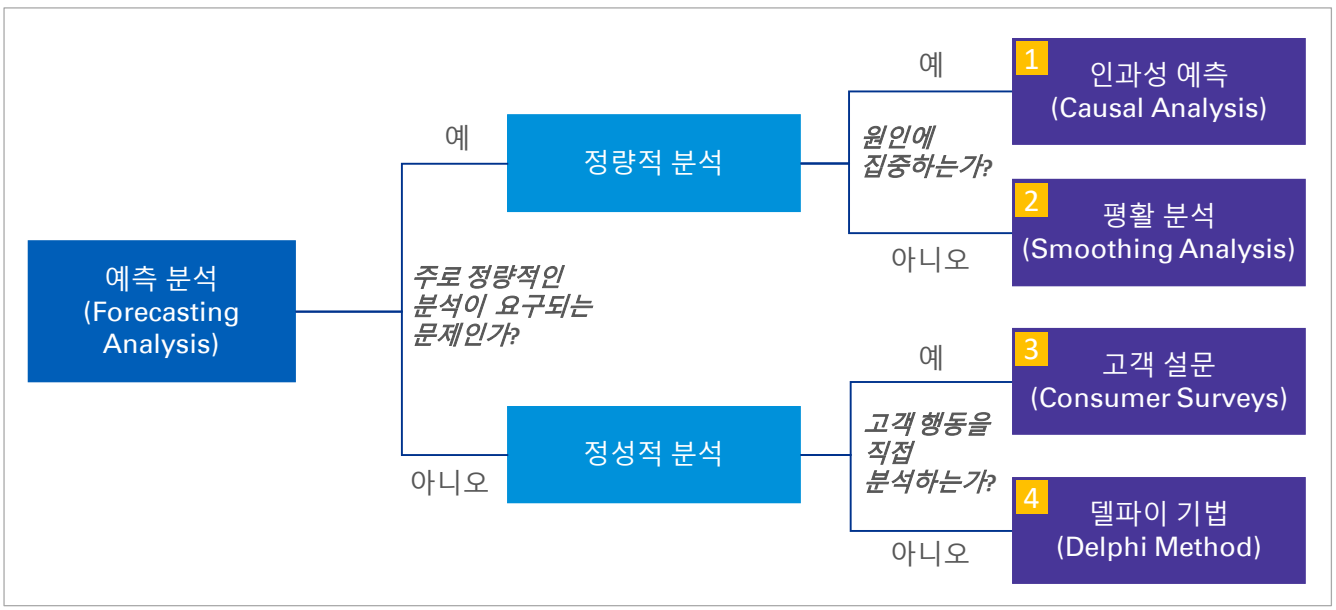
예측 분석은 기업이 보유하고 있는 데이터의 가용성과 범위에 따라 정량적 분석과 정성적 분석으로 분류할 수 있다. 정량적 예측 분석은 인과성 규명 여부에 따라 크게 '인과성 예측'과 '평활 분석'으로 나눌 수 있다. 정성적 분석은 고객의 행동을 대상으로 하는지 여부에 따라 '고객 설문'과 '델파이 기법' 등으로 구분할 수 있다.

정량적 분석은 미래의 변화를 예측 모형 등의 수리적 근거를 통해 추정할 수 있는 경우에 활용되는 방법론이다. 정량적 분석의 큰 축을 담당하고 있는 '인과성 예측'은 경영 변수 간의 인과성에 기반하여 미래 현상을 정량적으로 예측하는 방법론이며, 다른 한 축인 '평활 분석'은 시계열 데이터의 과거 추세나 패턴을 통해 향후 변수의 움직임을 관찰하는 방법론이다.

기업에서 정량적 분석을 활용해 미래를 예측하는 영역은 다양하다. 생산부서부터, 마케팅, 영업, 인사(HR)에 이르기까지 다양한 부서에서 경영 불확실성을 줄이기 위해 예측 분석을 활용하고 있으며, 기업의 데일리 오퍼레이션부터, 주단위, 월단위, 연단위 계획을 짤 때에도 데이터에 기반한 분석이 이뤄지고 있다. 과거에는 관리자의 경험이나 감에 의존해 미래를 예측을 했다면, 요즘은 데이터와 고도화된 분석 기법을 활용해 더 객관적이고 신뢰성있는 예측이 이뤄지고 있다.

치열한 경쟁 환경 속에서 기업이 얼마나 시장의 데이터를 정확하게 읽어내고, 선제적으로 대응하느냐는 기업의 생존과도 직결되어 있다고도 볼 수 있다.

>> 예측 분석의 종류와 기법



Source: 삼정KPMG 경제연구원

기업 운영 혁신을 위한 데이터 과학: 기업의 활용 방안

정략적 예측 분석의 활용 사례로 제조 기업의 판매량 예측을 들 수 있다. 기업이 판매량을 잘못 예측하여 과잉 생산을 할 경우 기업의 재고 비용은 증가한다. 만약 제품의 유통기간이 짧을 경우, 생산한 제품을 모두 폐기해야 하는 상황도 발생할 수 있다. 반면, 기업이 너무 보수적으로 예측해 주문을 다 소화하지 못하는 경우, 납기 일자를 맞추지 못하거나 거래처를 잃는 등 운영상 더 큰 문제를 일으킬 수 있다.

"최근에는 일부 예측 분석의 업무들이 간편해지고 자동화되고 있는 추세"

오늘날 판매량을 예측하는 시계열 모델로는 지수 평활법(Exponential Smoothing)부터 ARIMA, 요인 분해(Factor Decomposition) 등 다양하게 존재한다. 일반적으로 실무에서는 예측 모델에 활용할 훈련 데이터와 검증 데이터를 시계열로 나누고, 이를 다양한 모델에 적용시켜 예측 오차가 가장 적게 나오는 모형을 선정하곤 한다. 이 때, 기업이 얼마나 판매량에 영향을 미칠 수 있는 요인을 잘 선정하고 모델링을 하는지는 모델의 퍼포먼스와 직결되게 된다. 기업은 판매량에 영향을 미치는 가격뿐만 아니라, 프로모션, 계절적 요인, 경쟁 상황 등 다양한 요인을 고려하여 모델을 개발해야 한다.

기업의 중장기적 판매량에 대한 목표가 설정됐다면, 이후 기업은 생산판매 회의를 거쳐 월단위, 혹은 주단위로 세분화된 생산 계획을 짠다. 더 단기적으로는 일단위, 시간 단위로 업무를 스케줄링하고, 이 과정에서 병목현상이 일어나지 않고, 인력의 배치 또한 최적으로 운영될 수 있도록 계획해야 하는데, 이 때에도 정량적인 예측 분석이 활용된다.

"예측 분석은 객관적 사실과 주관적 해석을 통해 불확실한 미래의 방향성을 탐색"

최근에는 데이터 분석 전문가들이 모여서 한줄, 한줄 코딩하고 모델을 실험해보고 평가하던 방법에서, 일부 예측 분석의 업무들이 간편해지고 자동화되고 있는 추세다. 기업은 단계적 검색법을 사용하여 다양한 분석 모델 중에서도 가장 목적에 적합한 모델을 선택해 모형 탐색에 드는 시간과 비용을 줄여나가고 있다.

정성적 예측 분석은 기업 내 데이터 자원이 풍부하지 않은 상황에서 고객과 전문가의 의견을 통해 산업과 시장의 변화를 예측하기 위해 사용되는 방법론이다. 여기에는 '고객 설문'과 '델파이 기법' 등이 흔히 사용된다. 정성적 분석 방법론은 정량적 방법론에 비해 수치적인 근거를 마련하기 어렵다. 하지만, 예측 모형만으로는 발견하기 어려운 새로운 인사이트를 얻거나 위험 요소에 대해 인지할 수 있다는 장점이 있다.

본 분석에서 가장 중요한 점은 정량적 분석을 통한 예측과 정성적 분석을 통한 예측의 컨센서스(Consensus)를 마련하는 것이다. 단순히 시계열 모형을 활용하여 예측한 미래 지표의 변화가 모든 시장 상황을 반영한다고 보기는 어렵고, 반대로 전문가나 특정 고객의 의견이 항상 옳다고 보기 어렵기 때문이다. 당면한 비즈니스에 대한 깊은 이해를 바탕으로 수집된 데이터를 올바른 방법으로 분석했을 때, 파급력 있는 결과를 도출해낼 수 있을 것이다.

기업 운영 혁신을 위한 데이터 과학: 기업의 활용 방안

③ 경험과 학습을 통해 진화하는 처방 분석

처방 분석은 기업이 수익성, 운영 효율성, 고객 서비스 등의 다양한 비즈니스 목표를 달성할 수 있도록 데이터 자원으로부터 최적의 실행 방안을 도출한다. 처방 분석이 예측 분석과 다른 점은 실행 가능한 대안을 제시한다는 점이다. 기업은 기계학습, 시뮬레이션, 최적화 등의 수리적 기법을 통해 다양한 선택 가능 대안 중 최적의 실행안을 선정한다.

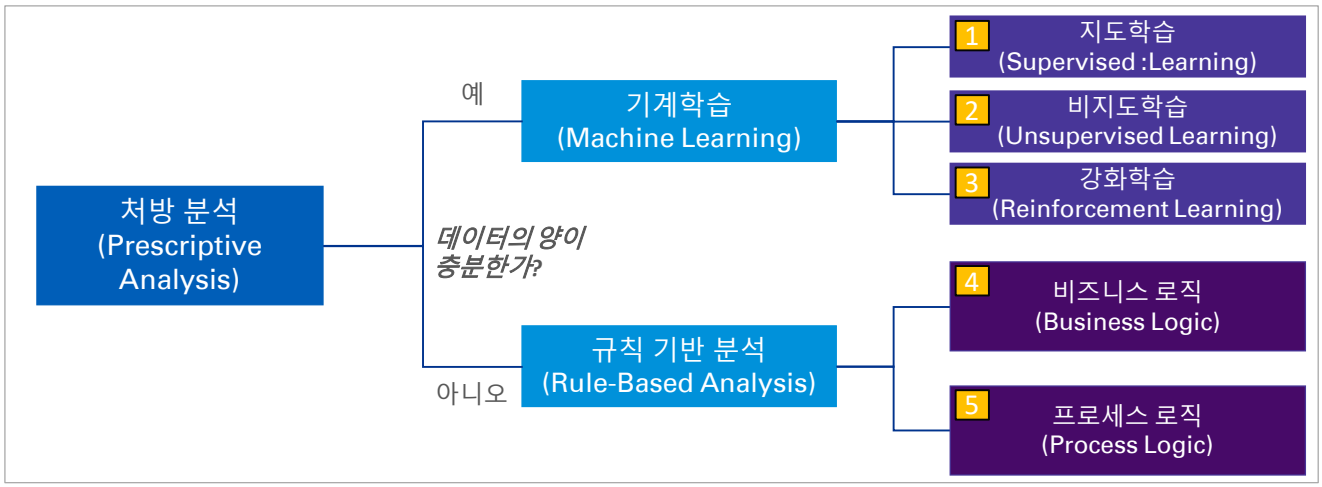
"처방 분석은 주어진 상황과 자원의 조건하에서 최적의 문제 해결 방안을 제시"

처방 분석은 크게 기계학습 기반 분석과 규칙 기반 분석으로 분류할 수 있다. 기계학습 기반의 분석은 주어진 데이터를 통해 스스로 학습할 수 있는 모형을 만들고, 이를 기업의 의사 결정에 활용하는 형태이다.

기계학습 기반 분석은 크게 '지도학습', '비지도학습', '강화학습'으로 나눌 수 있다. '지도학습 분석'은 훈련 데이터로부터 하나의 함수를 유추해내기 위한 기계학습의 한 방법론으로서, 훈련 데이터와 결과 값의 데이터가 상호 연결되어 있는 특징이 있다. 예측 분석에 기반한 시나리오 및 비즈니스 룰을 통해 산출한 영향도 데이터를 모형의 학습에 활용하고, 이 모형을 최적의 실행안을 도출하는 데 활용할 수 있다. '비지도학습 분석'은 데이터가 어떻게 구성되어 있는지를 알아내는 기계학습 방법론이다. 비지도학습 분석에는 일반적으로 군집화, 차원 축소, 이상치 탐지를 위한 알고리즘이 활용된다. '강화학습 분석'은 특정 환경 안에서 정의된 에이전트가 현재의 상태를 인식하여, 선택 가능한 행동들 중 보상을 최대화하는 행동 및 행동 순서를 선택하는 알고리즘을 통해 분석하는 방법론이다. 강화학습 과정은 최적의 정책 함수를 찾는 과정과 유사한데, 이 최적의 정책함수는 불확실한 미래에 얻을 수 있는 보상 함수의 기대값을 최대로 하는 행동을 고르기 때문에, 처방 분석의 기본 알고리즘으로 활용될 수 있다.

규칙 기반 분석은 보유하고 있는 데이터의 양이 부족한 경우, 기업의 경험을 토대로 비즈니스 및 프로세스 로직을 통해 최적의 안을 찾아내는 방법이다.

>> 처방 분석의 종류와 기법



Source: 삼정KPMG 경제연구원

기업 운영 혁신을 위한 데이터 과학: 기업의 활용 방안

> 처방 분석은 일반적으로 예측 분석, 대안 발굴, 영향 검토 등의 여러가지 실행 단계를 포괄

처방 분석은 앞서 기술한 기술·진단·예측 분석과 추천 알고리즘을 결합한 융합적인 분석 방법론으로서, 예측 분석, 대안 발굴 등의 여러 단계를 거친다.

처방 분석 이전 단계에서 수집된 기업 내부 및 외부의 데이터 자원은 처방 분석을 위한 여러 가지 프로세스를 거친다. 가장 먼저 처방 분석의 기반이 되는 '예측 분석'이 적용되는데, 이 과정에서 앞서 기술한 다양한 예측 모형, 통계 모형 등이 사용된다. 예측 분석 시, 수집되는 데이터의 특성을 파악하고 이에 적합한 분석 모형을 활용하는 것이 가장 중요하다.

'대안 발굴' 단계에서는 기업이 경험적으로 이해하고 있는 비즈니스 로직과 다양한 최적화 모델링 기법을 활용하여 실행 가능한 대안을 발굴한다. 예를 들어, 보험사는 예측 분석 단계에서 미래 금리 시나리오를 예측한 결과를 바탕으로 보험 부채의 규모와 변동성을 규명하고, 적절한 자본의 충당규모를 도출할 수 있다. '영향 검토' 단계에서는 민감도 및 시나리오 분석을 통해 다양한 대안 실행에 따른 경영 지표와 환경의 변화를 전망하고, 추천 알고리즘을 통해 최적의 안을 찾아낸다. 예를 들어, 보험사는 인공지능 처방 분석 시스템을 통해 금리 시나리오별 부채 변동 행태에 따른 최적의 자산운용 포트폴리오를 추천받을 수 있다.

다음 단계에서 조직의 책임자는 추천 받은 이행 안을 바탕으로 전략적·운영 의사결정을 내린다. 이러한 처방 분석의 프로세스는 좁게는 부서 차원에서 진행될 수도 있고, 넓게는 전사적인 의사결정 과정에 활용될 수도 있다. 분석 기법이 성숙되기 전에 가장 중요한 점은, 각 단계의 업무 담당자가 분석 결과와 기존의 지식이 일관성을 이루는지 확인하는 것이다.

>> 처방 분석의 프로세스

Source: 삼정KPMG 경제연구원

기업 운영 혁신을 위한 데이터 과학: 기업의 활용 방안

tidyverse는 데이터 과학의 새로운 패러다임으로서, 데이터 분석의 자동화를 촉진하고 통찰력을 배가

새로운 개념의 데이터 과학 운영체제 – tidyverse

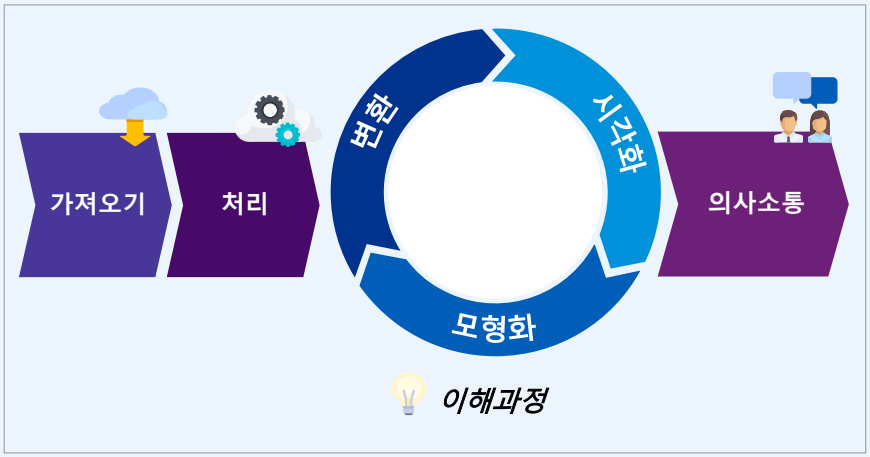
최근 SAS, SPSS, 미니탭과 같은 상용 패키지에서 R스튜디오, 파이썬과 같은 오픈소스 소프트웨어로 흐름이 바뀌면서, 데이터 과학자 및 실무자로부터 R스튜디오 기반의 타이디버스(tidyverse)에 대한 관심이 높아지고 있다.

일반적으로 데이터 시각화(Visualization)는 데이터에 대한 통찰력을 줄 수 있지만, 사람의 개입이 필요하기에 확장성(Scalability) 측면에서 한계가 존재한다. 반대로 모형(Model)은 자동화와 확장성 측면에서는 장점을 지니지만, 주어진 모형틀 안에서만 분석이 이뤄지기 때문에 새로운 통찰력을 주는 데 아쉬움이 있다. 타이디버스는 시각화와 모형을 통해 통찰력과 함께 자동화에 대한 부분도 충분히 반영한 체계적인 작업흐름을 제시하고 있다.

타이디버스는 데이터를 중심으로 다루기 때문에 정돈된 데이터(Tidy Data)에 대한 이해가 필요하다. 정돈된 데이터란 데이터를 통해 정보를 추출하고, 인사이트를 도출하기 위해서 시각화를 하고, 데이터를 모형으로 자동화를 하고, 수월한 자료구조를 갖는 데이터를 의미한다. 이와 반대되는 개념으로는 정돈되지 않은 메시 데이터(Messy Data)가 있다.

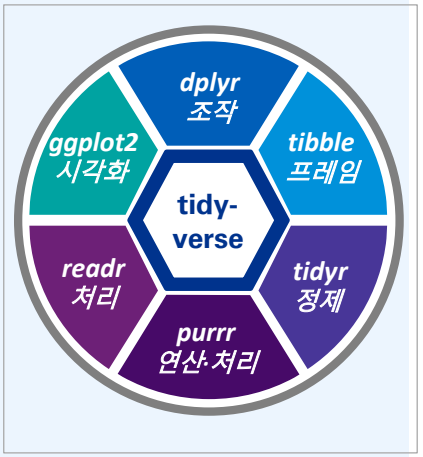
타이디버스는 데이터를 바라보는 관점과 분석 방식 측면에서 데이터 과학을 위한 새로운 운영체제로 바라볼 수 있다. 타이디버스의 핵심은 다양한 형태의 데이터를 가져와서 최종 산출물을 사람과 기계가 커뮤니케이션할 수 있는 형태로 제작하는 과정을 추상화했다는 것이다. 타이디버스의 구성요소로는 ggplot2(시각화), dplyr(조작), tibble(프레임), tidyr(정제), purr(연산·처리), readr(처리)가 있다. 타이디버스는 현재 많은 데이터 과학자와 실무자들로부터 호응을 얻으며 활용되고 있다.

>> tidyverse의 프로그램 내 작업 흐름



Source: 삼정KPMG

>> tidyverse 구성 패키지



Source: tidyverse

기업 운영 혁신을 위한 데이터 과학: 기업의 활용 방안

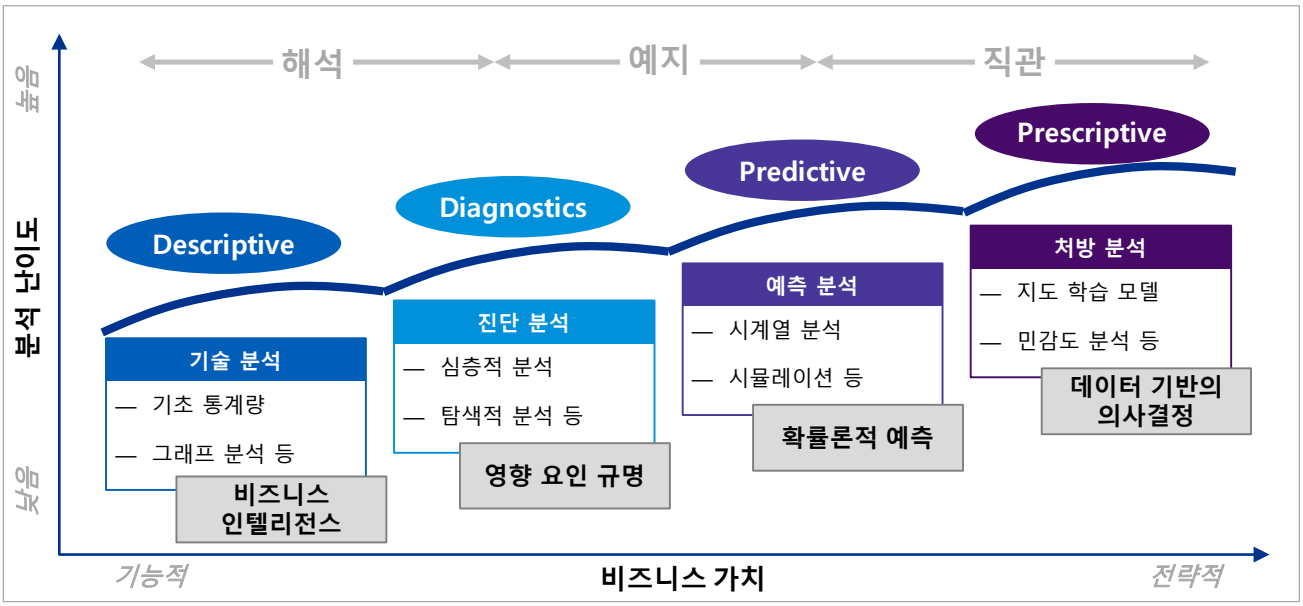
데이터 분석의 패러다임은 사후적 분석에서 사전적 분석으로 진화하는 중

비즈니스 고려 사항

사후적 분석에서 사전적 분석으로 진화하는 이행 로드맵 설계

기업의 데이터 기반 의사 결정 단계는 앞서 살펴 본 기술 분석, 진단 분석, 예측 분석, 처방 분석의 네 단계를 거치면서 완성된다. 기업의 내부와 외부에 가용한 데이터 자원이 많다고 가정하면, 일반적으로 데이터 분석의 가치는 기술 분석에서 처방 분석으로 패러다임이 진화한다. 왜냐하면, 도입되는 기술이 고도화될수록 기업 내·외부의 다양한 현상의 패턴에 대한 이해도가 높아지기 때문이다. 기존에는 의사결정을 위한 분석 시, 경영·경제학적인 가설을 기반으로 수리 모형을 만들고 외부적인 변화에 따라 이 모형을 수정하는 방식을 사용했지만, 향후에는 수집된 방대한 양의 데이터를 통해 기계학습 모형을 고도화하고 패턴에 대한 이해력을 높이는 방식이 널리 활용될 것으로 보인다.

>> 데이터 분석 패러다임의 진화



Source: 삼정KPMG 경제연구원

국내의 많은 기업들은 아직까지 기술 분석 및 진단 분석 패러다임에 머물러 있는 것으로 보인다. 이는 데이터 분석의 가치를 과소평가 하여 투자의 필요성을 느끼지 못하거나, 경쟁력 강화를 위한 인력, 조직, 문화 등의 기초 역량이 부족하기 때문으로 판단된다. 하지만, 급변하는 비즈니스의 세계에서 미래에 대한 예측 없는 기업 운영은 기업의 경쟁력을 점차 약화시키는 주요 요인이 된다.

지금부터라도 데이터 분석 역량을 확보할 수 있도록 투자를 아끼지 말아야 한다. 특히, 시장의 향방을 예측할 수 있는 예측 분석 및 인공지능을 활용한 처방 분석 시스템 구축을 위한 로드맵을 설정하고 실행해 나가야 한다.

기업 운영 혁신을 위한 데이터 과학: 기업의 활용 방안

> 데이터 분석은 사업 목표를 고려하여 문제 유형 정의, 알고리즘 선정, 시각화 및 모델링을 수행

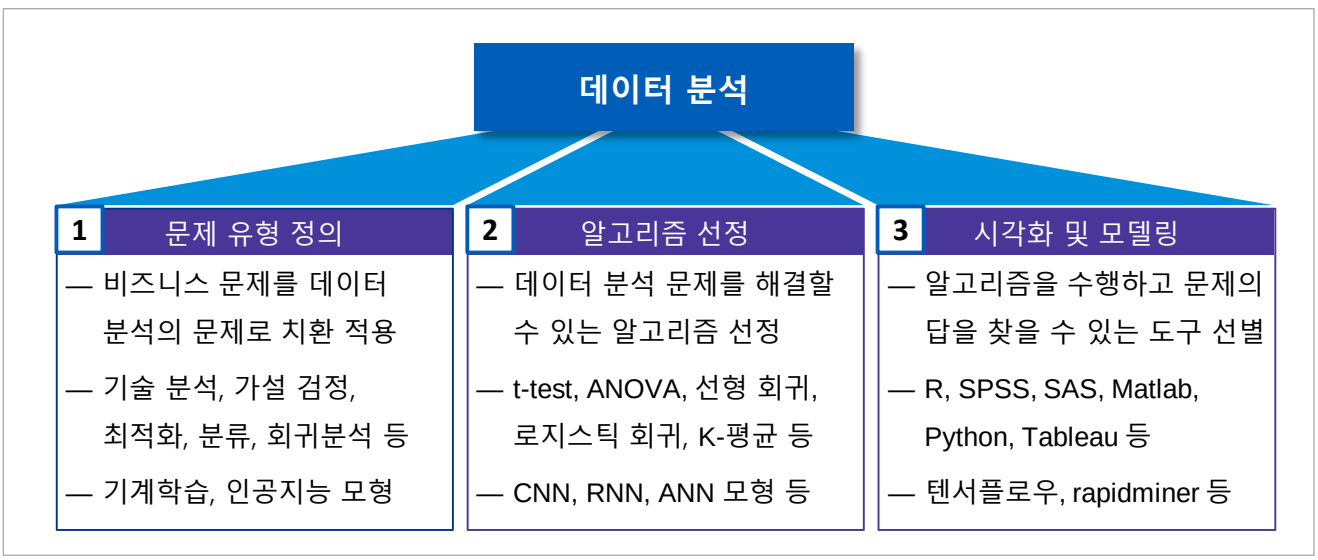
사업 목표와 일원화된 분석 프로세스 정립

전사적 데이터 분석 로드맵을 수립할 때, 조직이 고려해야 하는 사항을 크게 세 가지로 정리할 수 있다. 첫 번째로, 해결해야 할 문제를 명확히 인지하고 그 유형과 범위를 정의해야 한다. 조직에서 당면한 문제가 시장의 특성을 이해하는 것인지, 산업의 변화를 추적하는 것인지, 사건의 원인을 규명하는 것인지 등 분석의 성격을 파악해야 한다. 데이터 분석의 유형에는 기술 통계, 가설 검정, 최적화, 분류, 회귀분석, 시뮬레이션, 기계학습 등이 있으며, 비즈니스의 문제를 데이터 분석의 문제로 치환하는 역량을 갖추는 것이 중요하다.

두 번째로, 정의된 데이터 분석 문제에 적합한 알고리즘을 적용해야 한다. 데이터 분석의 문제는 적절한 알고리즘에 의해서만 해결될 수 있으며, 옳지 않은 방법론을 적용하는 경우 돌이킬 수 없는 결과를 야기할 수도 있다. 이 단계에서 중요한 점은 알고리즘 내 변수의 제한 조건과 범위를 명확히 하는 것이다. 예를 들어, 온라인 영상 스트리밍 플랫폼의 소비자를 적당한 기준에 의해 분류하고자 한다면, 분류 기준 변수의 수(차원의 수), 군집의 개수, 이상치 기준 등을 적절하게 설정해주어야 한다.

세 번째로, 설계된 알고리즘을 수행할 분석 도구를 선별해야 한다. 업계에 잘 알려진 데이터 분석 도구로는 R, SPSS, Matlab, Tableau 등이 있으며, 목적에 맞는 툴을 선택하여 데이터 분석을 진행한다. 시각화 분석의 경우 Tableau, R 등이 강점을 가지고 있는 것으로 알려져 있으며, 전통적인 통계 분석의 경우 R, SPSS 등이 흔히 사용되고 있다. 데이터 분석의 모든 요소들을 고려할 때 데이터를 통한 가치 혁신이 가능해지며, 기업의 데이터 경쟁력을 강화할 수 있다.

>> 데이터 분석의 비즈니스 고려사항



Source: 삼정KPMG 경제연구원

기업 운영 혁신을 위한 데이터 과학: 기업의 활용 방안

사내 보고를 위한 데이터 시각화 기법

사내 보고를 위한 데이터 시각화 기법으로 흔히 히스토그램, 파이 차트, 산점도, 선 그래프 등을 활용

컴퓨터를 기반으로 한 시각화 시스템은 시각적으로 데이터를 표현함으로 인해서 사람들이 작업을 더욱 효율적으로 수행할 수 있도록 돕는다.

① **히스토그램(Histogram):** 표로 되어 있는 도수 분포를 그래프로 나타내는 시각화 기법으로, 주로 고객이나 구매 행동의 빈도를 분석하기 위해 활용한다. 평균, 분산, 중앙값 등의 분석을 통해 표본의 일반적인 특성을 도출할 수 있다.

② **파이 차트(Pie Chart):** 전체 표본에 대한 각 계열의 비중을 원형 그래프로 나타내는 기법으로, 다수의 표본이 포함되는 계열을 표현하는 데 용이하다. 주로 시장 규모, 고객 및 상품 구성을 표현하는 데 사용된다.

③ **산점도(Scatter Plot):** 직교좌표계를 이용해 두 변수 간의 관계를 나타내는 기법으로, 주로 경영 변수 간의 영향도를 판단하고자 할 때 사용된다. 산점도를 기반으로 상관관계 분석, 회귀분석 등 통계적 분석이 수행된다.

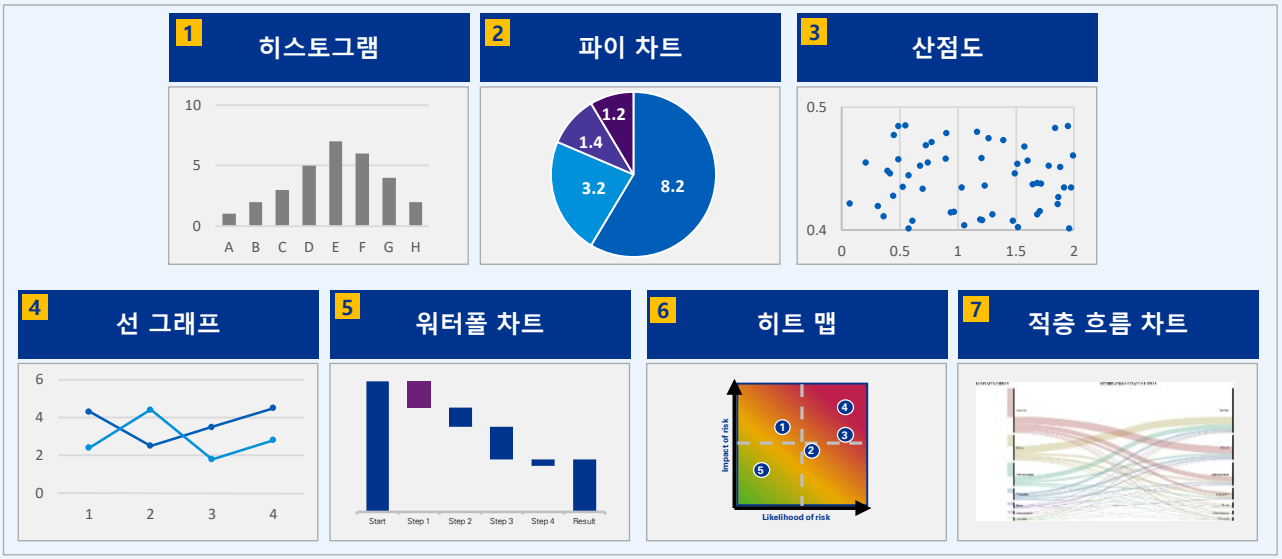
④ **선 그래프(Line Graph):** 지표를 점으로 표시하고 그 점들을 이어 그리는 기법이며, 추세나 변동성을 파악하기 위한 기초 자료로 활용된다. 주로 유가, 금리, 환율 등의 정량적인 지표의 추이를 분석하기 위해 사용된다.

⑤ **워터폴 차트(Waterfall Chart):** 변수에 영향을 미치는 순차적인 누적 효과를 막대로 나타내는 기법이며, 주 요인과 부 요인을 구분하기 용이하다.

⑥ **히트 맵(Heat Map):** 열을 뜻하는 히트와 지도를 뜻하는 맵을 결합한 단어로, 다양한 정보를 특정 이미지 위에 열 분포 형태로 표현하는 기법이다. 일례로, 특정 페이지의 요소가 얼마나 인기있는지를 분석하는 데 사용된다.

⑦ **적층 흐름 차트(Alluvial Chart):** 시간에 따른 네트워크 구조의 변화를 나타내기 위해 개발된 흐름 분석 기법이며, 주로 온라인 플랫폼에서 고객의 행동을 분석하기 위해 사용된다.

>> 사내 보고를 위한 데이터 시각화 기법 예제



Source: 삼정KPMG 경제연구원

기업 운영 혁신을 위한 데이터 과학: 기업의 활용 방안

5. 데이터 활용

데이터 활용 과정에서의 이슈사항

데이터의 활용은 데이터 자원에서 인사이트를 도출하는 과정을 의미하며, 기업의 성과 창출에 직접적인 영향을 미치는 중요한 활동이다. 넓은 범위에서의 데이터 활용은 앞서 기술한 데이터 수집, 저장, 처리, 분석의 전 과정을 포함하지만, 본 장에서는 데이터 분석 결과를 유의미한 성과로 전환하기 위한 활동만을 다루기로 한다.

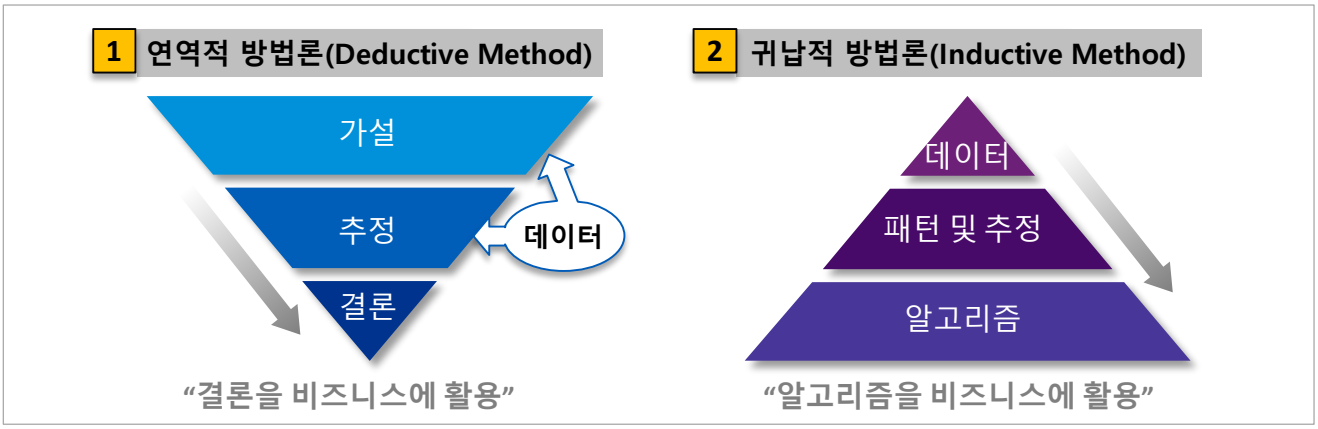
"데이터 활용 방법론은 크게 연역적 방법론과 귀납적 방법론으로 구분되며, 상황에 따라 유연한 적용이 필요"

기업이 데이터 자원을 실질적인 성과로 전환하는 과정에서 가장 큰 장애요인은 '사업 분석(Business Analysis)' 및 '데이터 엔지니어링(Data Engineering)' 역량이 부족하다는 것이다. 사업 분석 능력의 부재는 빅데이터에서 의미있는 결론을 도출하거나, 비즈니스 가설을 데이터를 통해 검증하는 데 부정적인 영향을 미친다. 데이터 엔지니어링 역량의 부재는 사업 분석가의 인사이트 창출 활동을 저해할 수 있다.

예를 들어, 사업 분석가가 초기 가설 설정의 오류로 옳지 않은 방향으로 분석을 진행하는 경우, 앞 단계의 데이터 수집·저장·처리·분석이 문제 없이 이루어졌다고 하더라도 의미있는 결론을 도출할 수 없다. 또한 데이터 엔지니어가 데이터 처리 단계에서 데이터를 유실했거나, 중복 데이터에 대한 처리를 해주지 않았을 경우에도 의미있는 결론을 도출하기 어려워진다.

또한, 사업 분석 체계가 미비한 경우 데이터 활용 능력이 저하될 수 있다. 일반적으로 사업 분석은 연역적 방법론과 귀납적 방법론을 통해 수행된다. 연역적 방법론은 초기 사업 가설을 설정하고 수집된 데이터를 통해 가설을 검정하기 위한 추정 단계를 거쳐, 최종적으로 도출된 결론을 비즈니스에 활용하는 방식이다. 또한 귀납적 방법론은 데이터 레이크에 저장된 빅데이터에서 패턴을 읽어내고, 도출된 알고리즘을 비즈니스에 활용한다. 이러한 사업 분석 방법론을 구체화할 체계가 미흡하거나, 시스템 내 기능 간 불균형이 발생하는 경우 데이터 활용이 제한될 수 있다.

>> 사업 분석에 데이터를 활용하는 두 가지 방식: 연역적 방법론과 귀납적 방법론



Source: 삼정KPMG 경제연구원

기업 운영 혁신을 위한 데이터 과학: 기업의 활용 방안

"데이터 활용 단계에서 자체적인 플랫폼을 구축하거나, 서드파티 서비스를 도입하여 새로운 인사이트를 창출"

데이터 활용을 위한 기술·기법 인에이블러

기업이 데이터를 활용하는 방법은 크게 두 가지로 나눌 수 있다. 자체적인 데이터 플랫폼을 구축하는 방법과 서드파티(Third-Party) 데이터 플랫폼을 도입하는 방법이다.

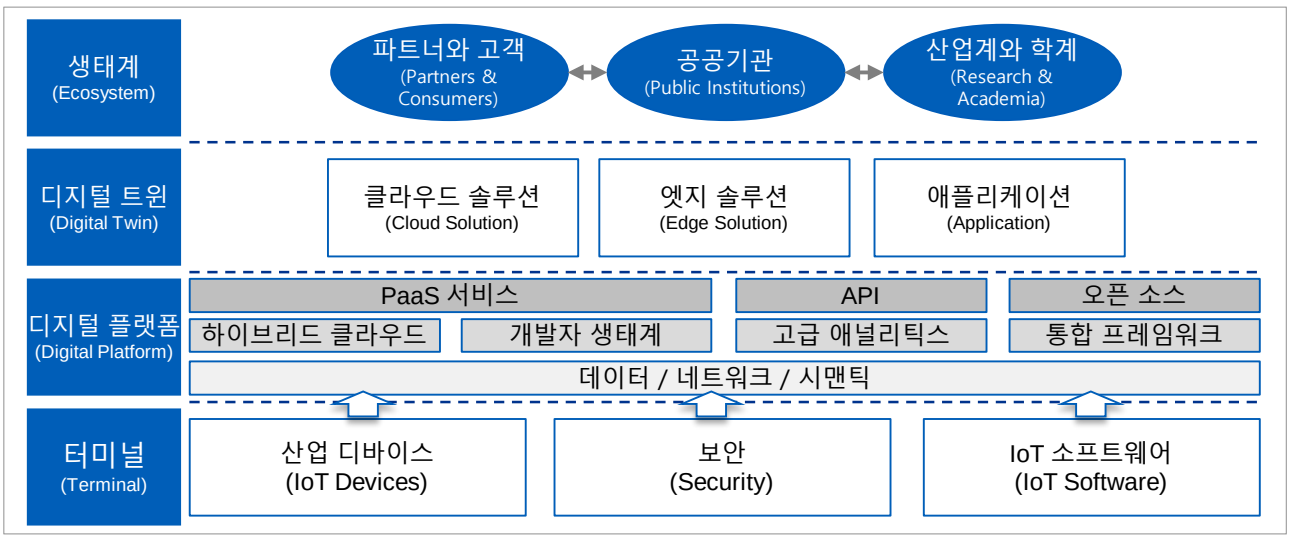
자체적인 데이터 플랫폼을 구축하는 방법은 조직의 구성원들이 익숙한 형태로 개발할 수 있고, 상황에 따라 신속한 커스터마이즈(Customize)가 용이하다는 장점이 있다. 반면 데이터 처리 및 연산 요구 능력 증가에 따른 스케일업(Scale-up)에 많은 비용이 들고, 특정 단계의 오류가 전후 단계로 전파되는 경향이 있어 운영 리스크가 높다는 단점이 있다.

서드파티 데이터 플랫폼을 솔루션 형태로 도입하는 방법은 구동 환경에 따라 스케일업이 용이할 수 있고, 데이터 활용 프로세스 내 오류 전파에 따른 운영 리스크가 획기적으로 낮아질 수 있다. 반면 기업의 경영 환경에 맞는 커스터마이징이 상대적으로 어렵고, 원격 구동 환경인 경우 독립 소프트웨어 벤더(ISV, Independent Software Vendor)의 신뢰성을 완벽히 보장받을 수 없다는 단점이 있다.

최근에는 클라우드 환경의 보편화로 인해 데이터 수집, 저장, 처리, 분석 등 데이터 과학의 모든 영역을 동시에 다룰 수 있는 '데이터 애널리틱스 플랫폼 서비스(Data Analytics Platform Service)'가 확산되고 있는 추세이다. 선도 기업들은 데이터 생태계의 일원으로서 외부의 비즈니스 솔루션을 적극적으로 도입하고 운영 혁신을 이어나가고 있다.

특히 데이터 생태계는 기업이 새로운 가치를 창출할 수 있는 토양의 역할을 하며, 개발자, 기획자, 파트너, 고객이 모두 함께 참여하는 비즈니스를 창출하여 해당 산업의 질적인 성장을 촉진한다.

>> 데이터 활용을 위한 '생태계 기반의 디지털 트윈' 아키텍처



Source: 삼정KPMG 경제연구원

기업 운영 혁신을 위한 데이터 과학: 기업의 활용 방안

데이터 활용 시 조직의 비즈니스 목표, 고객의 니즈, 전략 실행 환경을 종합적으로 고려하는 것이 필요

비즈니스 고려 사항

데이터 기반 의사결정 프로세스에 대한 이해

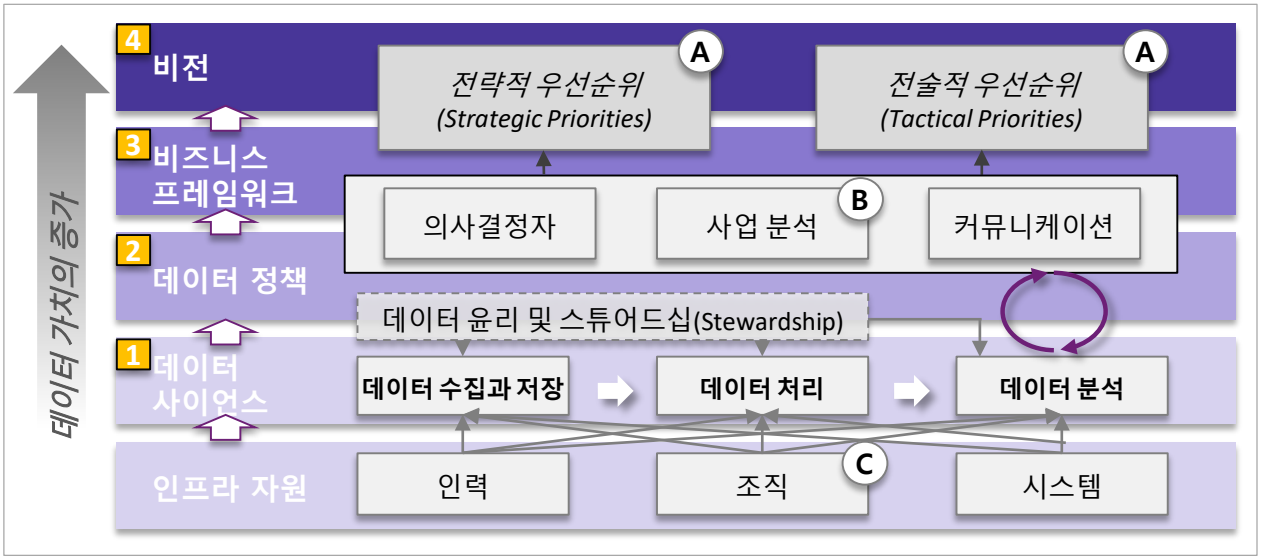
데이터 기반의 의사 결정 프로세스는 일반적으로 데이터 과학, 데이터 정책, 비즈니스 프레임워크, 비전의 네 가지 단계로 구성된다. 1단계(데이터 과학) 단계에서 수집된 데이터와 분석된 결과는 기업의 데이터 정책 하에서 관리·운영되며, 비즈니스 프레임워크 단계에서는 기획자, 마케터 등 여러 이해관계자들과 커뮤니케이션 과정을 거치게 된다. 이 단계에서 의사결정자는 전략적 우선순위와 전술적 우선순위를 고려하여 비전에 부합하는 의사 결정을 내린다.

데이터 기반의 의사결정 프로세스에서 고려할 사항은 크게 세 가지로 정리할 수 있다. 첫째, 조직의 비즈니스 목표를 명확히 이해해야 한다.(아래 그림의 'A' 부분) 기업에서 데이터 자원을 활용하는 이유는 궁극적으로 조직의 목표를 달성하기 위해서이다. 일부 기업의 경우 충분한 검토 없이 사내 데이터 사이언스 도입을 추진하여 미미한 성과를 거두는 데 그치기도 했다.

둘째, 고객의 입장에서 생각해야 한다.(아래 그림의 'B' 부분) 수익 창출의 원천인 고객에 대한 깊은 고민을 바탕으로 데이터 사이언스 전략을 추진해 나가야 한다. 단순히 업무에 데이터 사이언스를 도입하는 것만으로는 큰 의미가 없기 때문이다.

셋째, 처음에는 소규모 팀과 업무 단위로 데이터 활용을 추진해야 한다.(아래 그림의 'C' 부분) 기업이 처음 데이터 사이언스를 도입하고자 할 때 가장 어려운 점은 어디서부터 시작해야 할지 모른다는 점이다. 이런 경우, 가장 작은 업무로부터 적용 범위를 점차 넓혀 나가는 것이 중요하다.

>> 데이터 기반의 의사 결정 프로세스 및 고려사항



Source: 삼정KPMG 경제연구원

기업 운영 혁신을 위한 데이터 과학: 기업의 활용 방안

사례 분석

① 데이터 자산과 애널리틱스 역량으로 경쟁력을 만드는 우버(Uber)

데이터에 기반한 비즈니스를 활발히 전개하는 사례로 미국의 승차공유 플랫폼 기업인 우버(Uber)를 꼽을 수 있다. 우버는 2010년 처음 차량 공유 서비스를 미국 센프란시스코에서 런칭한 이후 지속적으로 우버 서비스를 이용하는 고객(수요자)과 드라이버(공급자) 양쪽 사이드의 데이터를 축적해왔다. 2019년 기준 1억명 이상의 이용자를 보유한 우버가 축적한 데이터 자산은 우버의 애널리틱스 역량과 결합되어 다른 플랫폼 기업이 모방할 수 있는 차별화된 경쟁력을 만들어내고 있다.

> "우버가 축적한 데이터 자산은 애널리틱스 역량과 접목되어 다른 플랫폼 기업이 모방할 수 있는 경쟁력을 만들어"

앞서 설명한 것처럼 데이터 분석의 궁극적인 목표는 문제를 진단하는 것에서 더 나아가 미래를 예측하고 선제적으로 대응 방안을 찾는 데 있다. 이를 위해 우버는 ARIMA 모델, Holt-Winters 지수평활법(Exponential Smoothing)과 같이 비교적 간단히 사용할 수 있는 전통적인 통계 모형을 사용기도 하고, RNN(Recurrent Neural Network), QRF(Quantile Regression Forest), GBM(Gradient Boosting Machine), SVR(Support Vector Machine), GPR(Gaussian Process Regression)과 같은 머신러닝 방법론을 사용하기도 한다.

기업이 활용할 수 있는 다양한 분석 모델이 존재하는 가운데, 우버는 문제 해결을 위해 가장 적합한 모델을 선정한다. 모델 선정 시, 우버가 보유한 데이터의 크기와 범위, 분석에 필요한 변수, 설명 가능성(Interpretability)의 필요 여부 등을 고려한다. 특히, 우버의 탑승 데이터는 시간대별, 주별로 패턴을 보이는데, 계절성이나, 경쟁 구도, 가격과 같은 외생 변수를 고려하여 다양한 모델을 시험해보고 최적의 모델을 선정한다. 모델을 만든 후 빠른 반복(rapid iteration)을 통해 모델을 검증하고 발전시키는 과정을 거친다. 경우에 따라, 우버는 다양한 예측 모델을 복합적으로 활용하기도 한다.

> "우버는 이용자들이 우버 플랫폼 내에서 가장 효율적인 운송 수단을 선택해 이동할 수 있도록 할 것"

기업에서 얼마나 정확하게 미래를 예측하는 것 말고도, 예측 구간을 얼마나 좁힐 수 있는지 또한 기업의 중요한 과제다. 아무리 예측 정확도가 높더라도, 예측 구간이 넓으면 경영 불확실성이 증폭될 수도 있기 때문이다. 이에 따라 우버는 모델을 평가할 때에는 모델이 얼마나 정확하게 미래의 상황을 예측하는지 뿐만 아니라, 정밀도(precision)와, 재현율(Recall)도 고려한다.

우버는 내부적으로 개발한 머신러닝 플랫폼 '미켈란젤로(Michelangelo)'에서 자사가 개발한 수천 개의 모델을 훈련시키고 있다. 이 시스템을 활용해 우버는 정확한 도착 시간 계산, 고객 응대 시간 개선, 탑승시 위험 요소 확인 등의 업무를 수행하고 있다.

우버는 승차공유 플랫폼에서 확보한 데이터와 분석 역량을 무기로 승차 공유 서비스를 넘어 전동킥보드, 대중교통, 비행 택시 등 모든 모빌리티 영역으로 사업을 확장해나가고 있는 중이다. 우버는 전 세계 이용자들이 우버란 플랫폼 내에서 가장 효율적인 운송 수단을 선택하여 이동할 수 있도록 할 계획이다.

기업 운영 혁신을 위한 데이터 과학: 기업의 활용 방안

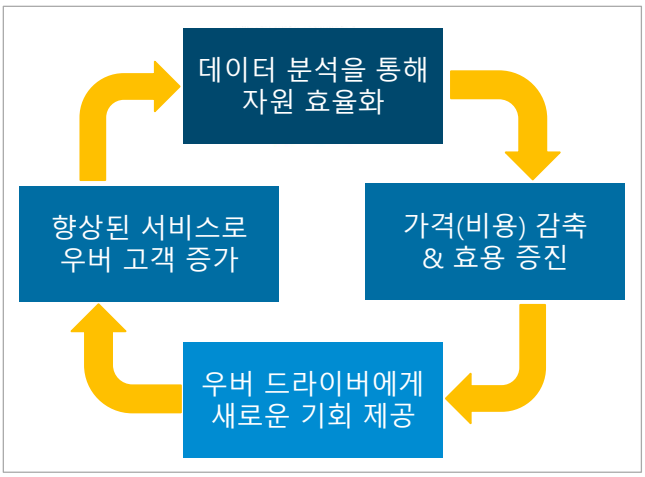
우버가 데이터 자산과 데이터 사이언스 역량을 발휘하는 대표적인 분야는 실시간 도로 현황을 분석해 최적의 이동 경로를 추천해주고 고객과 드라이버를 최적의 매칭 알고리즘으로 연결시켜주는 것이다. 우버가 축적한 데이터로 학습된 고도화된 알고리즘은 드라이버의 자원(운휴 시간, 이동거리, 비용, 회전율)을 효율화하고, 고객에게는 높은 편리성을 제공하는 등 드라이버와 우버 이용 고객을 지속적으로 유입시키는 우버의 핵심 경쟁력이다.

> "수요와 공급이 달라질 때마다 가격을 차별화하여 균형을 찾는 탄력 요금제는 드라이버와 이용자 모두에게 효용을 높여"

둘째는 가격이다. 우버는 설립 초기부터 서비스 이용자와 드라이버 양쪽에게 이득을 주는 다양한 가격 전략을 시험해왔다. 우버는 수요자와 공급자로부터 수집한 데이터를 바탕으로 최적의 가격을 결정하는 탄력 요금제(Surge Pricing)를 사용하고 있다. 이는 택시와 같이 이동 거리와 이동 시간에 따라 고정된 요금을 지불하는 것이 아니라 시장에서의 수요와 공급량에 따라 가격을 실시간 차별화 하는 가격 전략이다. 우버가 개발한 가격 결정 알고리즘은 수요가 많은 지역과 시간대에는 요금이 올라가게 하고, 그러지 않을 경우에는 가격을 낮춰 소비자들이 보다 저렴한 가격에 우버 서비스를 이용할 수 있도록 한다. 수요와 공급이 달라질 때마다 가격을 차별화하여 균형을 찾는 탄력 요금제는 드라이버와 이용자 모두에게 효용을 높이고 있다.

마지막으로, 우버는 기업 내부 서버를 운영할 때에도 데이터를 활용하고 있다. 우버가 글로벌 모빌리티 플랫폼으로 성장하면서, 우버의 트래픽은 폭발적으로 증가했으며 우버의 다양한 서비스를 호출하는 패턴 또한 복잡해졌다. 우버는 CPU, 메모리 부족으로 응답 속도가 늦어지거나 시스템이 과부하에 걸리는 것을 예방하기 위해 머신러닝을 활용하고 있다. 우버는 한주나 한달 동안의 초당 요청수, 응답 속도, CPU 사용률 등 핵심 지표들이 어떻게 변화할지를 예측하면서, 기업 내부적으로 데이터 흐름과 트래픽이 안정화될 수 있도록 하고 있다.

>> 우버의 데이터 활용을 통한 선순환 사이클



Source: Uber, 삼정KPMG 경제연구원

>> 우버의 데이터 분석 영역

1	**시장 예측 및 자원 최적화** 실시간 도로 현황을 분석해 최적의 이동 경로 탐색, 수요자와 공급자를 연결시켜주는 최적의 매칭 알고리즘 활용
2	**탄력 요금제(Surge Pricing) 도입** 수요와 공급에 따라 실시간으로 가격을 차별화하여 드라이버와 이용자에게 효용 증대
3	**내부의 데이터 흐름과 트래픽 관리** 머신러닝을 활용해 안정적 서버 운영

Source: 삼정KPMG 경제연구원

기업 운영 혁신을 위한 데이터 과학: 기업의 활용 방안

② 빅데이터 기반의 통합 리스크 관리 솔루션을 구현한 DHL

육상 물류 기업 또한 디지털 기술의 활용과 데이터 분석을 통해 기존의 가치 사슬과 그 확장 영역에서 새로운 기회를 창출할 수 있다. 최근 비상 상황에 대비한 클라우드 기반의 종합 컨트롤 타워 서비스와 적시 운송을 위한 도로 안전 시스템 등이 빅데이터 활용 신규 비즈니스로 부상하고 있다.

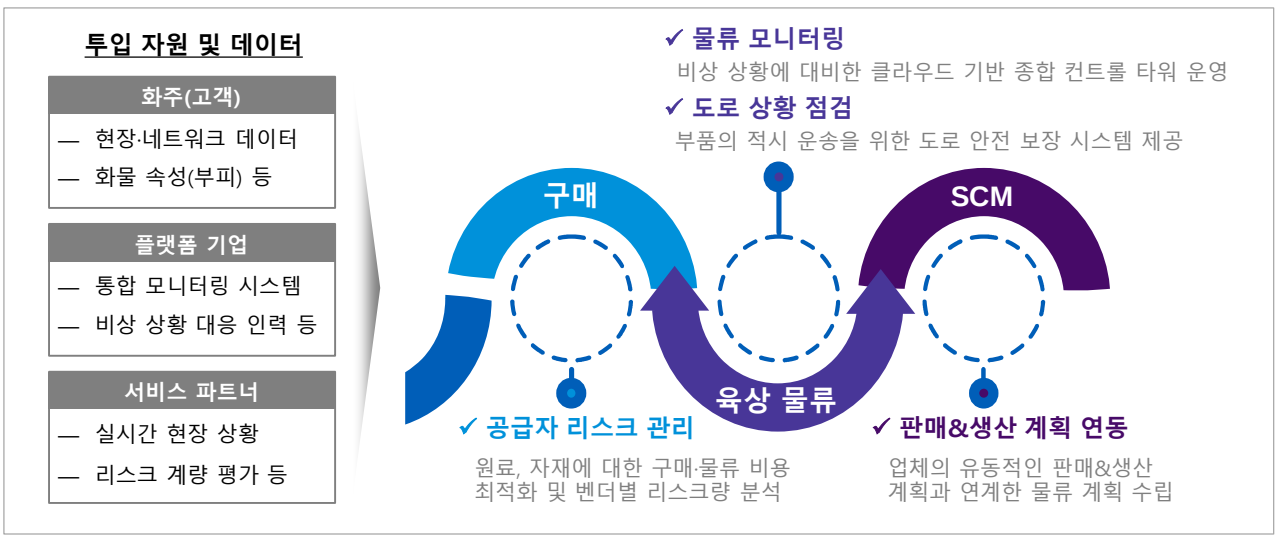
"DHL은 구매와 SCM 영역에 대한 리스크 관리 솔루션을 제공하고, 데이터 분석을 통해 기존의 육상 물류 가치 사슬을 확장"

글로벌 물류 업체인 DHL은 2016년, 공급망 네트워크를 가시화하여 물류의 리스크를 최소화하는 통합 리스크 관리 솔루션 'Resilience360'을 도입했다. 기업은 빅데이터 기반의 전용 포털을 통해 공급망 관리의 분석 영역을 확장할 수 있다. 예를 들어, 기업은 1차 벤더뿐만 아니라 2~3차 벤더까지의 공급망을 관리할 수 있게 되는 것이다.

DHL은 전통적인 리스크 관리 시스템에 다양한 ICT 기술과 데이터 분석을 도입하여 기존에 없던 새로운 고객 가치를 창출해냈다. 서비스는 배송 트럭의 이동 경로를 최적화하기 위해 디지털 지도, 위성 지도, 교통 패턴 인식, 소셜 네트워크 서비스의 데이터를 활용하였다. 또한 인공지능 시스템을 도입하여 '예측 물류'를 구현하고, 긴급한 화물의 수송에 대한 트럭의 고장이나 자연재해의 영향을 최소화하였다.

더불어, 빅데이터와 디지털 기술의 활용으로 구매와 SCM 분야에서 새로운 비즈니스 기회가 창출될 수 있다. 물류 기업은 고객의 구매·물류 비용 최적화 및 벤더별 리스크 관리 등을 통해 공급자 리스크를 총체적으로 관리하는 서비스를 제공할 수 있다. SCM 관련 가치 사슬에서는 업체의 유동적인 판매, 생산 계획과 연계하여 실시간으로 물류 계획을 수립하는 시스템을 구축하거나 서비스할 수 있다.

>> 디지털 기술을 통한 물류 리스크 관리 플랫폼의 영역 확대



Source: 삼정KPMG 경제연구원

기업 운영 혁신을 위한 데이터 과학: 기업의 활용 방안

③ AI 기반의 상품 혁신 전략을 추진하는 슈가크릭(Sugar Creek)

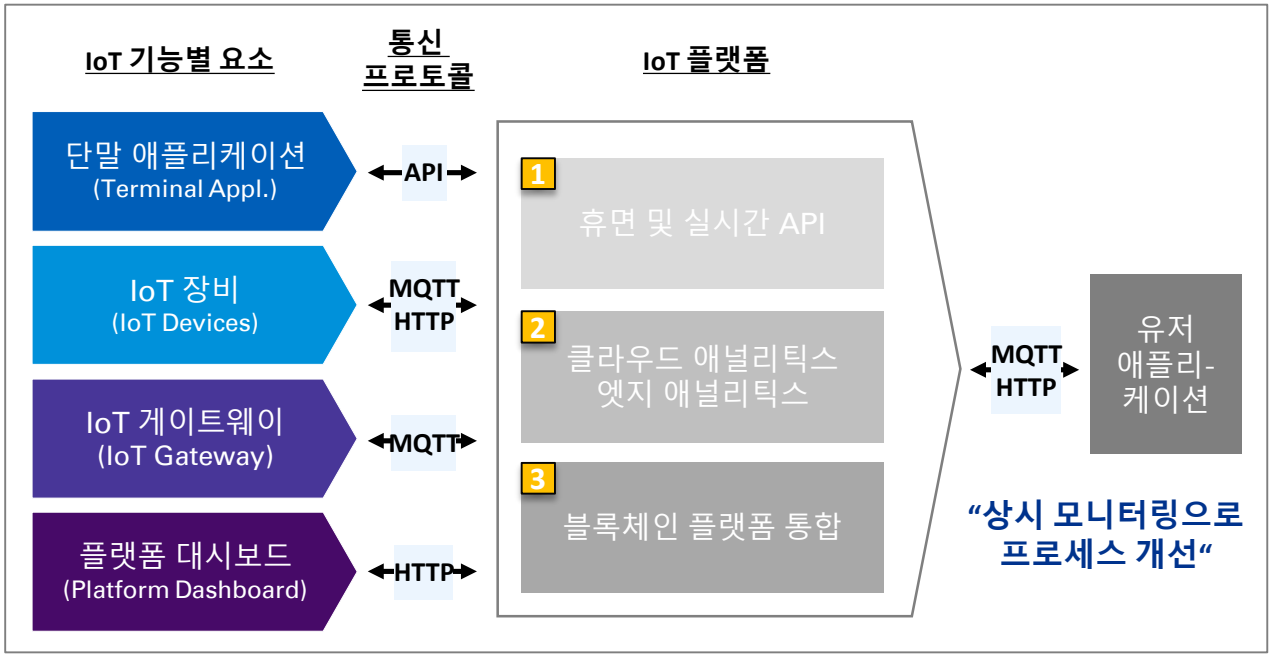
미국의 맥주 회사인 슈가크릭은 상품 및 생산 혁신에 데이터를 적극 활용하는 기업으로 알려져 있다. 슈가크릭은 맥주를 생산하는 과정에서 압력과 온도의 불균형 발생으로 인해 병에 맥주가 일정한 양으로 채워지지 않는다는 문제에 직면했다. 맥주를 병에 담는 과정에서 거품이 일정하지 않은 양으로 생성되는 것이었다. 균일한 양으로 생산되지 않는 맥주는 심각한 소비자의 반발을 야기할 수 있으며, 기업의 신뢰도에도 큰 영향을 미칠 수 있기 때문에 경영진에게는 반드시 해결되어야 할 문제로 인식되었다.

> 미국의 맥주 회사 슈가크릭은 IoT 솔루션 기반의 프로세스 개선으로 공정상의 기회 비용 감축

슈가크릭은 2019년 IBM의 '왓슨 IoT 플랫폼'을 도입한 이후, 정밀 유량계 및 사물 인터넷 센서로 수많은 데이터를 수집 및 분석하고 공정의 개선에 활용하였다. 특히, AI 분석을 통하여 압력과 온도 차에 의해 일정하게 채워지지 못한 맥주병의 생산율을 현저히 낮추어 한 달에 1만 달러 이상을 절약하는 성과를 거두었다.

일반적으로 IoT 플랫폼은 단말 애플리케이션, IoT 장비, IoT 게이트웨이 등과 연계하여 생산 라인의 데이터를 실시간으로 수집하고 분석한다. 특히, 클라우드 및 엣지 애널리틱스(Cloud and Edge Analytics)를 도입하여 분석의 신속성과 연계성을 강화한다. 데이터의 분석 결과는 '통합 관리 애플리케이션(Integrated Management Application)'을 통해 이용자에게 전달된다. 슈가크릭의 사례는 데이터 자원을 활용하여 고객 가치 혁신을 이끌어낸 대표적인 사례로 평가받고 있다.

>> IoT 기반 엣지 애널리틱스의 구조



Source: 삼정KPMG 경제연구원

기업 운영 혁신을 위한 데이터 과학: 기업의 활용 방안

④ 인공지능을 통해 제약 연구를 혁신하는 아톰넷(AtomNet)

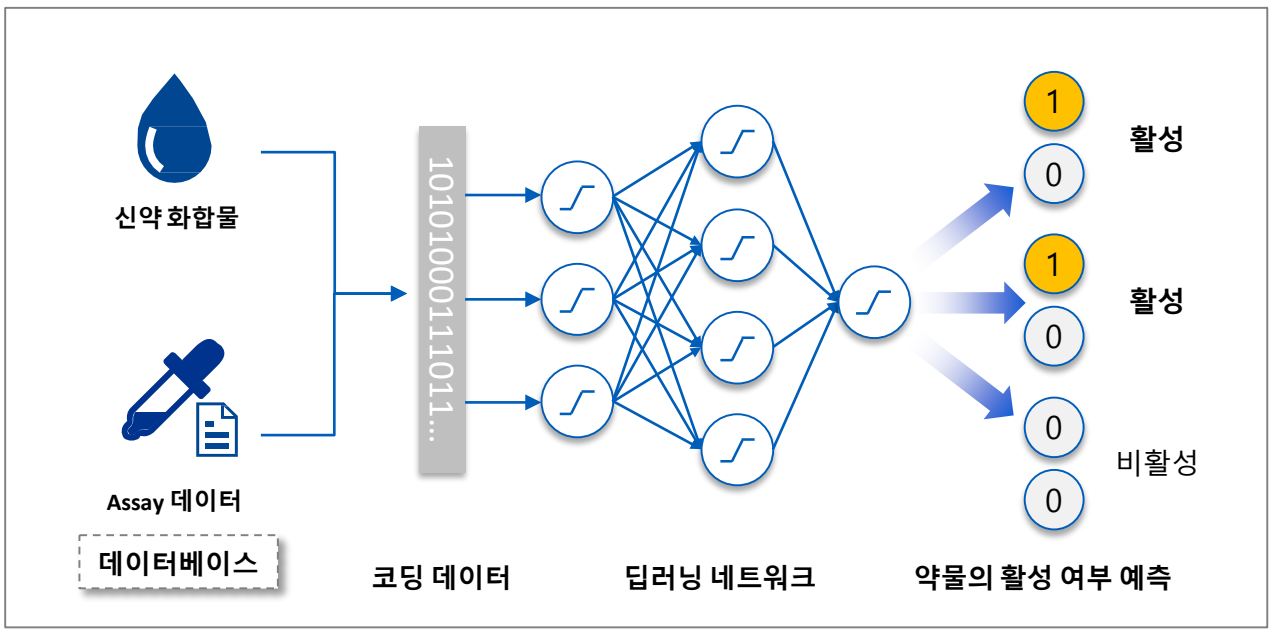
제약 산업에서 신약 개발은 그 파급력만큼이나 오랜 기간에 걸친 대규모의 투자가 이루어지는 활동이다. 신약 개발은 총 소요 기간 10년 이상, 최대 투자 금액 16억 달러 수준의 자본 집약적 활동으로 알려져 있다. 일반적으로, 신약 후보 화합물의 수가 초기 후보군의 2.5% 수준으로 줄어드는 데 걸리는 시간은 약 6~7년이며, 0.05% 수준으로 줄어드는 데 걸리는 시간은 약 13~14년이다. 잠정 후보 물질 중에서 약 1.5년이 소요되는 미국 식약청의 검토 프로세스를 거친 1개의 물질만이 승인되어 시장에 출시될 수 있다.

> "미국의 테크 기업인 아톰와이즈는 딥러닝 기반 솔루션인 아톰넷을 통해 신약 개발 소요 기간과 비용을 감축"

이러한 신약 개발이 데이터 혁신을 통해 변화하고 있다. 인공지능은 데이터를 활용하여 임상 실험 조건을 최적화하고, 투약 효과에 대한 사전 시뮬레이션을 용이하게 한다. 또한, 약물과 환자의 정보, 연구 결과 등에 대한 종합적인 분석으로 신약 개발 초기 단계의 소요 기간과 비용을 감축시킨다.

미국 샌프란시스코에 위치한 아톰와이즈(AtomWise)사가 2015년 공식 발표한 인공지능 신약 개발 솔루션인 '아톰넷'은 서로 다른 분자들의 상호작용을 분석하여 신약 개발 프로세스를 효율화한다. 수 만가지 화합물의 반응성을 일일이 시험해야 하는 연구·개발 활동을 딥러닝 모형을 통해 빠르고 정확하게 수행할 수 있다. 즉, 기존의 수많은 화합물과 Assay(실험) 데이터를 딥러닝 모형에 학습시켜 화합물의 형태를 통해 그 반응성을 예측할 수 있게 되었다. 아톰넷의 알고리즘은 공식 발표 이전부터 제약 연구에 활발히 사용되어 왔으며, 연구의 기술적 한계를 극복할 수 있는 차세대 인공지능 솔루션으로 평가되고 있다.

>> 인공지능 신약 개발 솔루션 '아톰넷'의 딥러닝 아키텍처



Source: 아톰와이즈

기업 운영 혁신을 위한 데이터 과학: 기업의 활용 방안

영국의 석유화학 기업 브리티시 페트롤륨은 사물 인터넷 기술을 통해 데이터 기반의 예측 정비를 실현

⑤ 데이터 기반의 예측 정비를 실현한 브리티시 페트롤륨(BP)

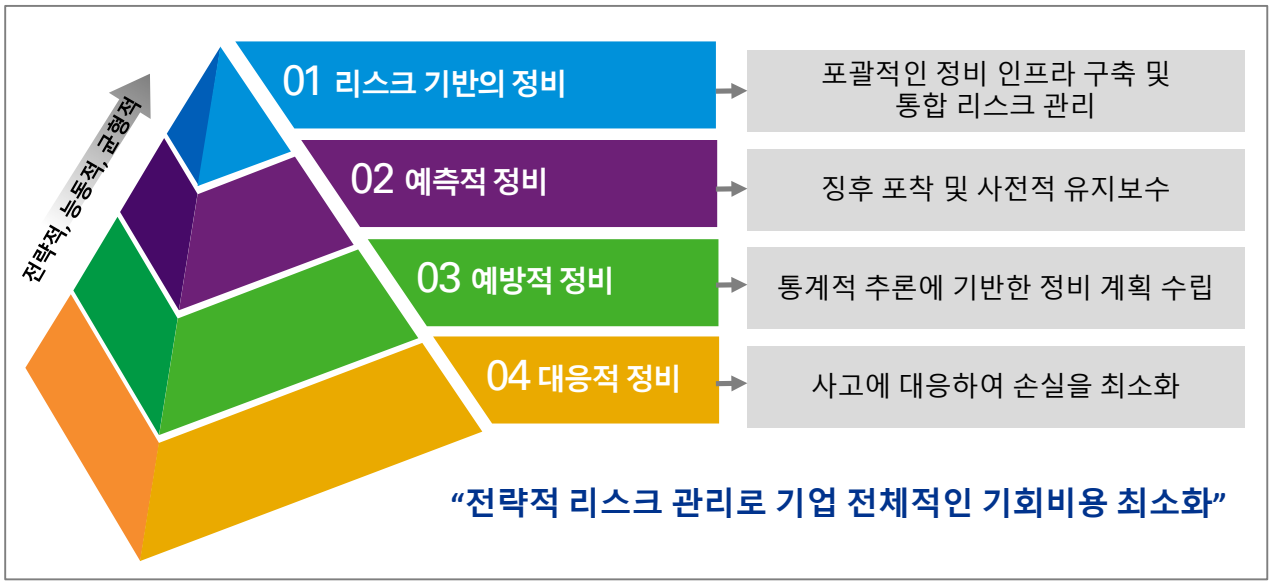
영국 런던에 본사를 두고 있는 석유화학 기업인 BP는 IoT 기반의 시추 운영 및 유지보수 시스템을 도입하여 비용 절감 등 효율성 개선을 이루어냈다. BP가 관리 중인 앙골라, 북해 등 약 650여 개의 시추 현장에 연결된 사물 인터넷 네트워크는 석유 및 가스 생산에 따른 설비 운영 데이터를 통합하고, 데이터 기반의 예측 정비를 현실화하였다. 사물 인터넷 시스템은 GE의 자산 성과 관리 시스템(APMS, Asset Performance Management System) 등 다양한 애널리틱스 플랫폼을 결합했다는 특징이 있다.

BP는 자체 고성능 슈퍼컴퓨터 허브인 CHPC(Center for High-Performance Computing)의 효율성 향상을 위해서 2,000km에 달하는 광케이블, 데이터 스트리밍, 저장, 처리 능력의 향상을 위한 빅데이터 기술에 투자하고 있다. 2020년까지 처리 가능한 데이터 용량을 약 1페타바이트(PB)에서 6페타바이트로 늘리는 작업을 진행할 것으로 알려져 있다.

BP가 투자하고 있는 데이터 처리 기술에는 아파치의 데이터 스트리밍 분야의 프로젝트인 카프카(Kafka), 소프트웨어 시스템 통신 자동화 프로젝트인 나이파이(NiFi), 아마존의 데이터 스트리밍 솔루션인 키네시스(Kinesis) 등이 있으며, 데이터 저장 기술로는 HDFS(Hadoop Disk Filing System) 등이 대표적이다.

리스크 기반의 정비 시스템 구축으로 BP는 1년에 약 2억 달러 규모의 비용 절감을 달성했으며, 전체적인 운영 효율성은 약 2~4% 수준 개선된 것으로 파악되고 있다.

>> 석유화학 기업의 전략적 운영 리스크 관리 이행 로드맵



Source: 삼정KPMG 경제연구원

기업 운영 혁신을 위한 데이터 과학: 기업의 활용 방안

결론 및 시사점

"데이터 과학은 데이터 기술의 집합이라기보다는 기업의 운영 철학이자 생존 전략"

데이터 과학의 핵심은 기술이 아닌 실행

데이터 자원과 데이터 과학은 석유 자원과 스트림라인(Streamline) 공정으로 빗대어 표현되곤 한다. 석유화학 산업에서 원유가 여러 단계의 공정을 거쳐 플라스틱, 섬유, 고무와 같은 제품으로 재탄생하듯 데이터 또한 여러 단계를 거치면서 새로운 가치를 창출해내기 때문이다. 원유 자원으로부터 가치가 창출되는 과정에서 증류, 냉각 등의 과정 중 단 한 가지 단계라도 간과하고 넘어가게 된다면 대규모 자본 투자의 성과를 기대하기 어렵다. 마찬가지로 기업이 데이터 자원을 활용할 때, 데이터 수집, 저장, 처리, 분석, 활용의 전 단계를 거치지 않으면 데이터의 숨겨진 가치를 발견할 수 없다.

오늘날 기업들은 데이터 과학과 기업의 비즈니스 모델을 별개의 영역으로 여기지 않고 있다. 데이터 과학을 일견 데이터 관련 기술의 집합으로 여길 수도 있지만, 오히려 기업의 운영 철학이자 생존 전략으로 해석하는 편이 타당하다고 본다. 왜냐하면, 기업이 데이터 경쟁력을 확보하기 위해서는 데이터 과학에 대한 로드맵과 실행 방안 수립이 필수적이기 때문이다. 액션 플랜(Action Plan)을 간과한 채로 신기술의 도입에만 집중한다면 데이터가 가지고 있는 숨겨진 가치를 발굴할 수 없다. 기업은 데이터 과학 역량을 확보하기 위해, 인력을 채용하고 시스템을 구축하는 등 다양한 데이터 전략을 추진할 필요가 있다.

아직 데이터 과학을 전사적으로 도입하지 못한 기업은 단순 데이터 활용 작업부터 실행해보는 편이 바람직하다. 기초적인 기술의 적용부터 시작해서 성공적인 활용 사례를 만들고, 점차 이를 시스템화해 나가는 편이 효율적이다. 이러한 방법론은 기업이 데이터 활용 사례를 학습하면서 목적에 맞는 데이터 활용 체계를 갖추기 용이하다는 장점이 있다.

"객관적인 사실과 그 해석을 기반으로 소통하는 기업만이 새로운 데이터의 시대에서 생존할 수 있을 것"

데이터의 언어로 소통할 수 있는 조직 문화

급변하는 시장의 형세와 빠른 기술의 발전은 기업이 데이터 기반의 사고방식을 체화하도록 요구하고 있다. 과거에는 경영자의 통찰과 감각이 기업 의사결정의 가장 중요한 요소였다면, 현재는 객관적인 사실과 현상으로부터 도출되는 패턴에 대한 신속한 대응이 중요해지고 있다.

이는 기업의 CEO나 임원에 국한되는 이야기가 아니다. 고객과 직접 대면하는 영업팀 사원, 재무팀의 매니저, R&D 부서의 연구원에 이르기까지 데이터 과학의 언어에 익숙해질 필요가 있다. 즉, 모든 직원이 업무를 하면서 축적되는 데이터를 처리, 분석하여 새로운 인사이트를 창출할 수 있는 기초 체력을 갖추어야 한다.

임직원의 데이터 분석 역량이 중요해지고 있는 가운데, 데이터 분석 툴을 교육하고 전파하는 커뮤니티 또한 늘어나고 있어, 이를 활용하는 방안도 고려해볼 수 있다.

기업 운영 혁신을 위한 데이터 과학: 기업의 활용 방안

한 예로, 1998년부터 그렉 윌슨(Greg Wilson) 박사 주도로 시작된 소프트웨어 카펜트리(Software Carpentry)를 들 수 있다. 1998년부터 시작된 이 커뮤니티는 전 세계에서 수백 차례 소프트웨어 교육 및 워크숍이 진행되었고, 한국에서는 2015년 한국전파진흥협회를 시작으로 연세대, 서강대, 한림대에서 매년 소프트웨어 카펜트리 프로그램이 진행되고 있다. 소프트웨어 카펜트리에서는 유닉스 쉘을 이용한 작업자동화, 소프트웨어 개발·협업 플랫폼 깃허브(GitHub), 파이썬과 R, SQL과 관련된 내용을 교육하고 있으며, 실습 위주의 강의로 과학 및 기술 종사자가 교육 내용을 업무에 활용하는 데 큰 도움을 주고 있다.

> 기업이 온톨로지적으로 쌓은 데이터를 가지고 작은 실험을 반복하며 빠르게 혁신하는 데브옵스(DevOps) 문화가 마련되야

더 나아가, 기업 내에서 데이터의 언어로 소통하고 온톨로지적으로 쌓은 데이터를 가지고 작은 실험을 반복하며 빠르게 혁신하는 데브옵스(DevOps) 문화 또한 마련되야 한다. 조직이나 개인의 주관적인 의견이 아니라, 객관적인 사실과 그 해석을 기반으로 소통하는 기업만이 새로운 데이터의 시대에서 생존할 수 있을 것으로 보인다.

외부의 데이터 사이언스 솔루션을 적극 도입

최근 클라우드 환경의 보편화로 PaaS, SaaS 환경에서 도입할 수 있는 데이터 사이언스 솔루션이 증가하고 있다. 아마존, 마이크로소프트 등의 클라우드 사업자는 대규모의 스토리지를 제공하는 동시에, 각사의 마켓플레이스(Marketplace)에서 다양한 기법으로 빅데이터를 분석할 수 있는 서드파티 애플리케이션을 제공하고 있다. 이러한 애플리케이션의 활용은 데이터 사이언스 시스템을 구축하는 시간과 비용을 줄여주고, 점차 증가하는 데이터량에 대응할 수 있는 확장성을 제공한다.

> 기업의 기존 역량 부족 시 외부의 데이터 사이언스 솔루션을 도입하는 전략이 필요

클라우드 환경으로 이행하지 않은 기업의 경우에도, 다수의 서드파티 데이터 사이언스 솔루션을 활용할 수 있다. 시장에서 각종 데이터베이스 소프트웨어, 애널리틱스 도구가 출시되고 있는 상황이다. 목적과 예산에 맞게 적절한 데이터 사이언스 솔루션을 구입하여 사용하는 것도 기업이 취할 수 있는 전략의 하나로 볼 수 있다. 외부 솔루션 도입 과정에서 기업은 데이터 분석을 위해 투자한 비용 대비 산출한 성과를 측정할 수 있는 지표를 함께 검토해야 한다. ROI(Return on Investment)에 대한 심사숙고 없이는 효과적인 데이터 전략이 지속적으로 추진되기 어렵기 때문이다.

데이터의 가치를 극대화하기 위해 데이터 연결성에 주목

기업은 고객 데이터를 효과적으로 센싱하고 이를 기업 내부 데이터와 연결할 경우 데이터의 가치를 증폭시킬 수 있다. 기업은 데이터 자원을 수집하고 저장, 처리, 분석하는 궁극적인 목적이 고객을 중심적 디지털 혁신하기 위함임을 인지해야 한다.

기업 운영 혁신을 위한 데이터 과학: 기업의 활용 방안

기업의 각기 다른 기능과 프로세스, 조직을 고객 데이터를 중심으로 연결되도록 재설계할 경우, 밸류체인 전반의 가시성이나 투명성을 제고할 수 있을뿐만 아니라, 더 나아가 제품과 서비스 혁신, 그리고 새로운 비즈니스 기회를 발굴할 수 있다.

KPMG의 '커넥티드 엔터프라이즈(Connected Enterprise)' 프레임워크는 기업의 내·외부 데이터를 연결하여 고객 중심의 디지털 혁신을 이룰 수 있는 방안을 모색하는 데 유용한 가이드라인을 제시해준다.

> "기업의 각기 다른 기능과 프로세스, 조직이 고객 데이터를 중심으로 연결"

데이터 전략부터 보안까지, 유연한 데이터 관리

데이터의 가치를 극대화하기 위해서는 데이터 관리에 또한 주목해야 한다. KPMG의 또 다른 프레임워크인 ADM(Advanced Data Management)은 데이터 관리 시 기업이 중점적으로 고려해야 할 사항을 종합해 놓은 툴이다. ADM의 세부적인 요소로는 데이터 전략 & 거버넌스, 데이터 아키텍처 & 모델링, 메타데이터, 데이터 품질, 마스터 & 레퍼런스 데이터, 데이터 오퍼레이션, 문서 및 콘텐츠 관리, 데이터 통합 & 호환성, 애널리틱스, 비즈니스 인텔리전스(BI), 데이터 보안이 있다. 기업은 ADM에서 제시한 데이터 관리의 11가지 핵심 요소를 고려해 전사적 차원에서의 변화를 이끌어야 한다.

>> KPMG의 ADM(Advanced Data Management) 구성 요소

데이터 전략 & 거버넌스	데이터 아키텍처 & 모델링	메타데이터	데이터 품질	마스터 & 레퍼런스 데이터	데이터 오퍼레이션
– 데이터 전략 & 로드맵 – 데이터 거버넌스 정책 & 원칙 – 커뮤니케이션 플랜 – R&R	– 데이터 아키텍쳐 디자인 – 데이터 모델링 – 데이터 디렉토리 – 데이터 흐름 & 리니지 툴	– 메타데이터 전략 – 메타데이터 표준 – 메타데이터 리포지토리 – 데이터 정의 – 메타데이터 관리	– 데이터 품질 전략 – 데이터 품질 프레임워크 – 데이터 품질 관리 – 데이터 품질 대시보드	– 마스터 데이터 전략과 로드맵 – 마스터 데이터 관리 가이드라인 – 마스터 데이터 평가/교육 툴 – 데이터 품질 툴 (KPI)	– 평가 및 영향도 분석 – 변화 관리 – 데이터 라이프 사이클 모델

문서 및 콘텐츠 관리	데이터 통합 & 호환성	애널리틱스	비즈니스 인텔리전스(BI)	데이터 보안
– 콘텐츠 관리 전략과 로드맵 – 템플릿 관리 – 콘텐츠 관리 – 콘텐츠 리포지토리 – 문서 검색 툴	– 마이그레이션 방식과 시나리오 – 데이터 통합 모니터링 프로세스 – 데이터 통합 원칙과 가이드라인	– 애널리틱스 전략 – 메타데이터 리포지토리 – 애널리틱스 레퍼런스 모델 – 애널리틱스 관리 로드랩	– BI 전략 – 메타데이터 리포지토리 – BI 로드맵 – 대시보드와 KPI	– 데이터 보안 전략과 로드맵 – 데이터 보안 리스크 평가 및 분류 – 데이터 보안 레포트

Source: KPMG International

기업 운영 혁신을 위한 데이터 과학: 기업의 활용 방안

> KPMG의 '커넥티드 엔터프라이즈'는 기업의 데이터와 프로세스, 기능을 고객중심적으로 재정의하고 전사적으로 디지털 혁신하는 프레임워크

KPMG의 '커넥티드 엔터프라이즈(Connected Enterprise)' 프레임워크

KPMG의 '커넥티드 엔터프라이즈(Connected Enterprise)'는 기업의 모든 데이터와 업무 프로세스, 기능 등을 고객중심적으로 재정의하고 전사적으로 디지털 혁신하는 프레임워크다. 일반적으로 고객중심 경영을 이야기할 때, 고객과의 접점에 있는 영업, 마케팅, 고객서비스(CS) 부분을 강조하는데, '커넥티드 엔터프라이즈'에서는 기업의 내·외부와 기업을 둘러싼 전체 생태계를 고객 중심적으로 변환시키고자 하는 데 차이점이 있다. 기업 내부적으로는 프런트 오피스부터 미들 오피스, 백오피스까지 전체 기능을 고객을 중심으로 변환시키는 데 중점을 두며 기업 내부뿐만 아니라 시장과 어려 채널, 기업 외부 파트너에게까지 고객 중심의 가치를 전달할 수 있도록 한다.

'커넥티드 엔터프라이즈'를 추진할 경우, 아래 8가지 효과를 기대할 수 있다.

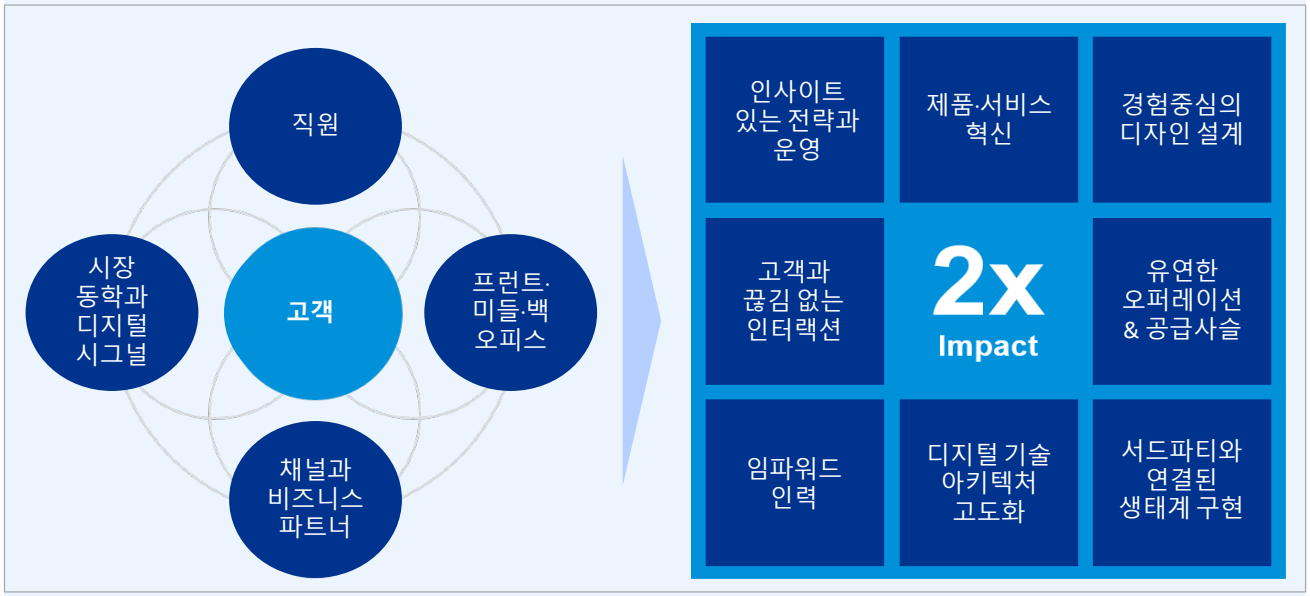
① **인사이트 있는 전략과 운영:** 데이터 자원과 애널리틱 기술을 활용해 경제적 이익(Data Monetization) 창출

② **제품·서비스 혁신:** 고객에게 새로운 제품·서비스, 가격 측면에서 가치 제안, 디지털 컴패니언 경험(Digital Companion Experience) 제공

③ **경험중심의 디자인 설계:** 비즈니스 목표와 연계하여 심리스(Seamless)한 디지털 고객 여정 지도를 디자인하고, 측정 및 가치 산정

④ **고객과 끊김 없는 인터랙션:** 마케팅, 영업, 고객 서비스 등 다양한 채널을 통해 고객과의 인터랙션을 강화하고 확보한 고객 데이터를 내부 데이터와 연결해 새로운 데이터 거래 구조 설계

>> KPMG의 '커넥티드 엔터프라이즈(Connected Enterprise)'의 개념과 기대효과



Source: KPMG International

기업 운영 혁신을 위한 데이터 과학: 기업의 활용 방안

① **유연한 오퍼레이션과 공급사슬: 시장의** 수요와 공급 데이터를 활용해 기업을 애자일하게 운영하고 밸류체인 전반에 가시성과 투명성을 제고

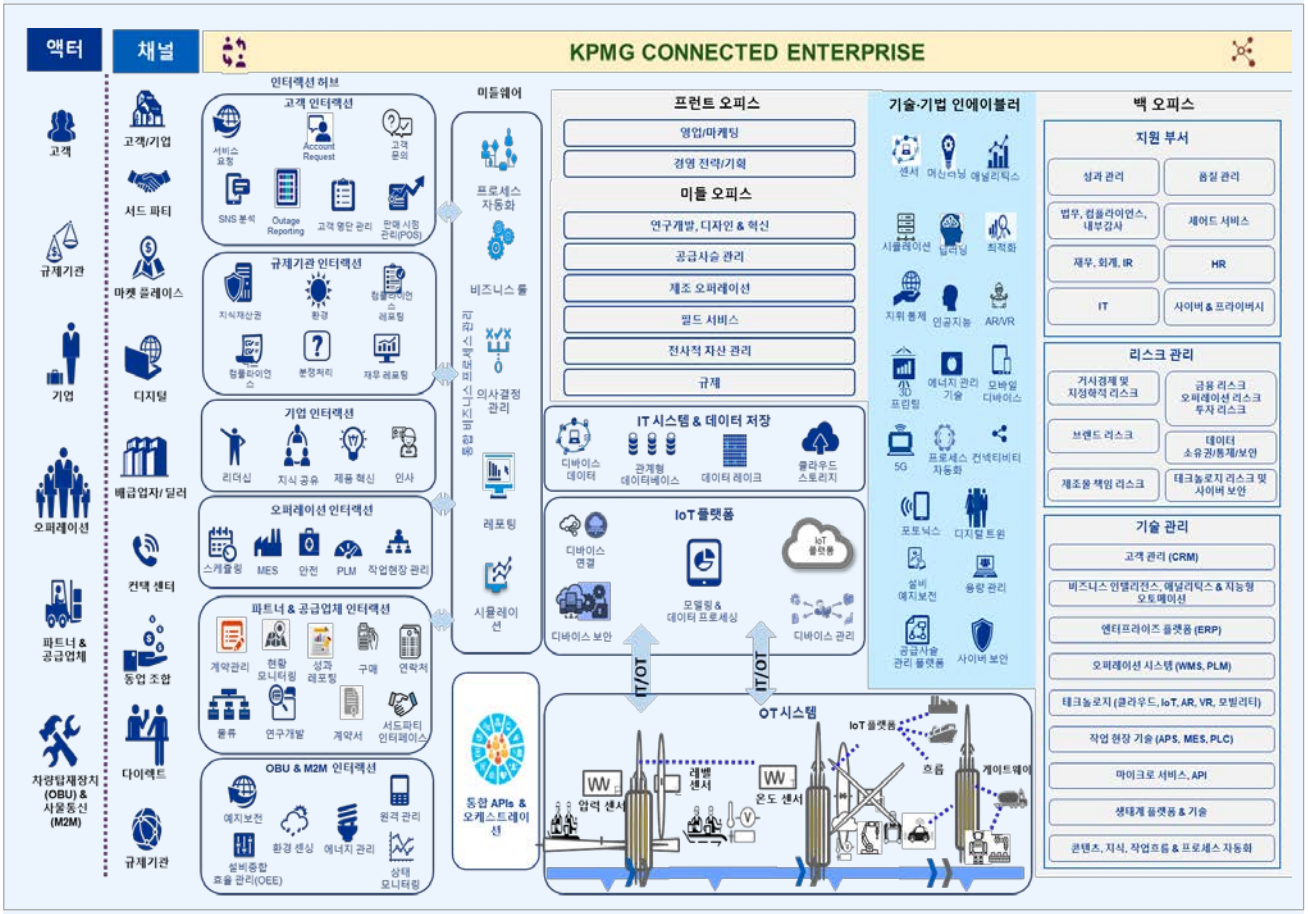
② **임파워드(Empowered) 인력:** 고객중심적인 조직 구조와 조직 문화를 형성하고 디지털 큐레이터가 기업의 '커넥티드 엔터프라이즈' 가치를 전사적으로 확산시켜 매시업(mash-up) 비즈니스 활성화

③ **디지털 기술 아키텍처 고도화:** 지능화되고 애자일한 솔루션을 구축하고 안전하면서도 확장 가능한, 비용 효율적인 신기술과 플랫폼 기반의 비즈니스 모델을 설계

④ **서드파티와 연결된 생태계 구현:** 디지털 신기술을 제공하는 기술 협력사, 거래처 등 서드파티(Third-party)를 활용해 제품의 출시 속도를 향상하고 비용을 절감, 리스크 완화

KPMG의 '커넥티드 엔터프라이즈' 프레임워크를 적용한 제조 산업의 아키텍처 모델은 아래 그림과 같게 구성될 수 있다.

>> KPMG의 '커넥티드 엔터프라이즈' 프레임워크를 적용한 제조 산업의 아키텍처 (예시)



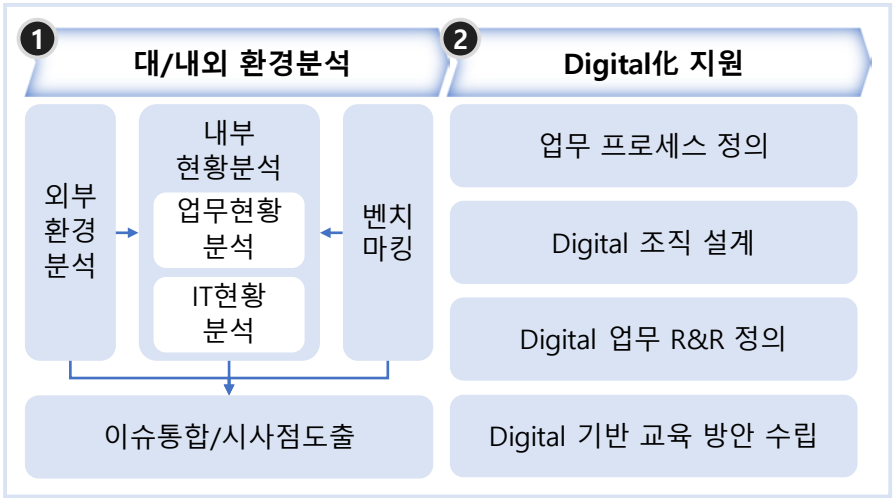
Source: KPMG International

HOW KPMG CAN HELP

삼정KPMG는 Digital 기반의 회사로 변화를 모색하는 기업에게 기업의 특색에 적합한 방향으로 보다 빠르게 Digital 기반의 회사로 변화하도록 Digital 체계, 기획 등 다양한 방면에 대한 자문 서비스를 제공하고 있습니다.

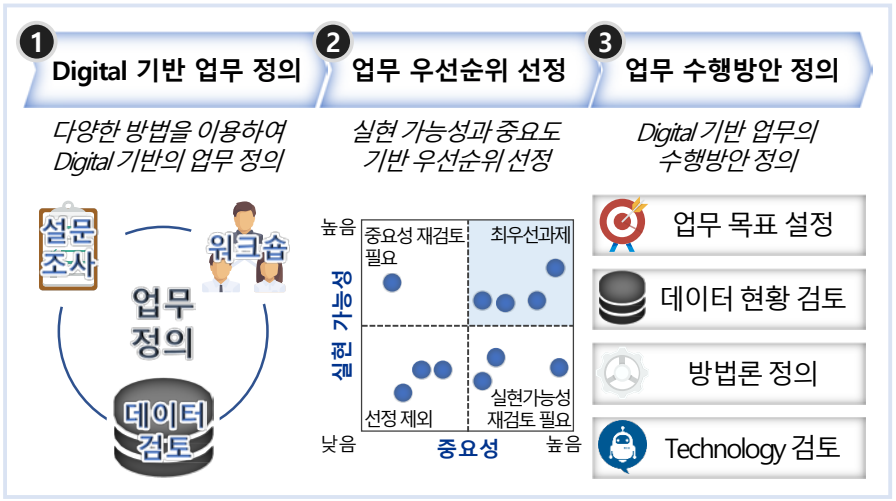
기업 Digital化 지원

1. Digital 기반 대/내외 환경 분석 및 벤치마킹
2. Digital 중심의 회사로 성장하기 위한 업무 프로세스, 조직 구성, 교육 등 다양한 체계 수립 방안 제안



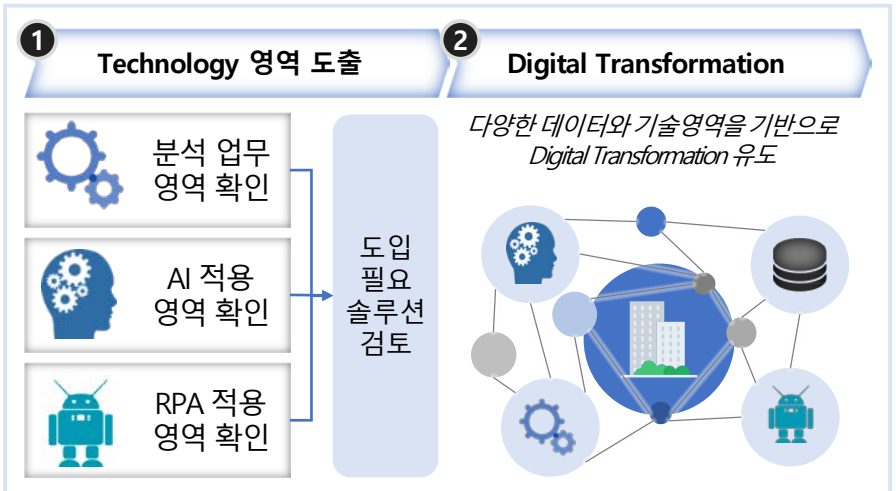
Digital 업무 기획

1. 기업 특성에 부합한 Digital 기반 업무 정의
2. 정의된 업무에 대한 우선순위 선정
3. Digital 기반 업무의 수행 방안 정의



Digital Transformation

1. Digital 기반 업무의 수행을 위해 필요한 기술 영역 도출
2. 기술 영역 적용 및 활용 방안 제안



삼정 KPMG

Business Contacts

데이터 과학 서비스 전문팀

Lighthouse (Advisory)

양현석
전무
T: 02-2112-3009
E: hyunseokyang@kr.kpmg.com

이광춘
상무
T: 02-2112-7748
E: kwangchunlee@kr.kpmg.com

디지털혁신센터 (Center of Excellence)

박문구
전무
T: 02-2112-0573
E: mungupark@kr.kpmg.com

kr.kpmg.com