[빅데이터 개념]

0. 빅데이터란 단어의 정의(3V)를 기준으로 설명해주세요.

빅데이터란 high-volume, high-velocity, high-variety로 설명할 수 있다. High-volume이란 하나의 pc에서 담을 수 있는 용량을 초과하는 대용량을 의미한다. 즉 기존의 기술로는 해결할 수 없는 양의 데이터를 가지고 있어야 빅데이터이다. High-velocity는 데이터가 생성되는 속도가 빠르고 실시간 또는 적시에 데이터를 처리하고 분석하는 것을 의미한다. High-variety는 다양한 데이터가 존재하는 것을 의미한다. 구조화된 데이터, 반구조화된 데이터, 정형데이터, 비정형데이터 등등 다양한 데이터가 존재하고 이를 처리하고 다루는 것이 빅데이터 기술이다.

1. map reduce model의 개념에 대해서 조사해주세요. map-reduce의 간단한 사용 예를 하나들어주세요. 인터넷에서 찾으셔도 되고, 코딩까지가 아닌 개념 적인 예를 본인이 직접 들어서 설명해주셔도 됩니다.

Map reduce model은 분산 컴퓨팅 환경에서 대규모 데이터 세트를 처리하고 분석할 수 있는 프로그래밍 패러다임이다.
Map stage : 입력 데이터가 더 작은 청크로 분할되어 클러스터의 여러 노드에 분산된다. 각 노드는 지도 함수를 사용하여 할당된 데이터를 독립적으로 처리한다. 지도 함수는 입력 데이터를 가져와 중간단계 키-값 쌍을 내보낸다. 예를 들어 각 노드가 텍스트 문서의 일부를 수신하고 각 노드의 맵 함수는 할당된 문서를 읽고 개별 단어로 토큰화한 다음 단어가 키이고 값이 1로 설정된 키-값 쌍을 내보낸다. 따라서 문서의 각 단어에 대해 지도 함수 는 (단어,1)1)를 중간단계로 내보낸다. 문서에 “apple:, “banana”, “apple” 이라는 단어가 포함되어 있으면 지도 함수는 (apple, 1), (banana,1), (apple, 1)을 중간 키-값 쌍으로 내보낸다.
Shuffle : 이 중간 단계에서는 키를 기준으로 모든 중간 키-값 쌍을 수집하고 그룹화한다. (“apple”,[1, 1]) 및 (“banana”,[1])와 같은 그룹을 형성한다.
Reduce stage : 이 단계에서는 이전 단계의 그룹화 된 중간 키-값 쌍이 reduce 함수에 의해 처리된다. Reduce 함수는 키와 해당 키와 관련된 값의 목록을 가져와 이들에 대해 일부 집계 또는 계산을 수행한다. 위의 예에서는 reduce 함수는 (“apple”, [1,1])와 (“banana”, [1])를 수신할 것이다. 그런 다음 각 키 값을 합계하여 최종 단어 수를 얻는다. 따라서 reduce 함수는 (“apple”, 2) 와 (“banana”, 1) 을 최종결과로 출력한다.
이러한 방식은 map reduce의 병렬성과 분산 특성을 활용하여 대규모 데이터 세트를 효율적으로 처리할 수 있다. 데이터를 더 작은 정크로 분할하고 여러 노드에 걸쳐 맵을 병렬로 실행하고 기능을 줄여 확장성과 처리시간을 단축할 수 있다.

3.빅데이터는 보통 한 컴퓨터로 처리할수 없을 정도의 데이터를 말합니다. 하나의 컴퓨터에서 연산을 할때 장단점과, 여러대의 컴퓨터로 분산컴퓨팅을 통해 장단점을 조사하고 이둘을 비교해주세요. 결론적으로 어떤게 좋은지 의견도 내주세요.

하나의 컴퓨터에서 연산을 할 때의 장점 : (단순성) 단일 컴퓨터에서 작업하는 것은 여러 시스템을

관리하거나 클러스터에 작업을 분산시킬 필요가 없기 때문에 간단하다.
(비용효율적) 일반적으로 분산 컴퓨팅 환경을 설정하고 관리하는 것에 비해 단일 컴퓨터를 구입하고 유지 관리하는 것이 비용이 적게 든다.
(소규모 데이터셋의 성능) 소규모 또는 중간 크기의 데이터셋을 처리할 때 단일 컴퓨터가 워크로드를 효과적으로 처리할 수 있는 충분한 처리능력을 제공하는 경우가 많다.

하나의 컴퓨터에서 연산을 할 때의 단점 : (제한된 처리능력과 메모리) 단일 컴퓨터는 유한한 처리능력과 메모리를 가지고 있으며, 이는 대규모 또는 복잡한 데이터 세트를 처리할 때 병목 현상이 될 수 있다.
(확장성 제한) 단일 컴퓨터의 처리능력을 확장하는 것은 하드웨어 제약으로 인해 제한된다. 데이터 양과 복잡정이 증가함에 따라 단일 시스템이 합리적인 시간내에 워크로드를 처리하지 못할 수 있다.
(단일 고장 지점) 하드웨어 장애 또는 시스템 충돌 시 모든 처리 및 데이터에 영향을 미쳐 잠재적으로 상당한 다운타임과 데이터 손실이 발생할 수 있다.

분산컴퓨팅의 장점 : (처리능력향상) 여러 대의 컴퓨터에서 병렬 처리를 가능하게하여 더 빠르고 효율적인 데이터 처리를 가능하게 한다.
(확장성) 워크로드가 증가함에 따라 클러스터에 시스템을 추가할 수 있으므로 컴퓨터 리소스를 원활하게 확장할 수 있다.
(고장 허용범위) 한 시스템의 장애로 인해 데이터 손실이나 시스템 다운타임이 발생하지 않는다. 이는 데이터 처리의 신뢰성을 향상 시킨다.

분산컴퓨팅의 단점 :(복잡성) 시스템 클러스터 관리, 작업 조정, 노드 간 통신 및 데이터 동기화 처리가 포함된다. 이러한 복잡성을 설정하고 유지하기 위해서는 추가적인 전문 지식과 인프라가 필요하다.
(통신 오버헤드) 데이터를 기계 간에 교환해야 하므로 통신 오버헤드가 발생한다. 네트워크 대기시간 및 데이터 전송 시간은 분산 시스템의 전체 성능에 영향을 줄 수 있다.
(비용 증가) 여러 대의 컴퓨터와 관련 네트워킹을 포함한 분산 컴퓨팅 인프라를 설정하고 유지하는 것은 단일 컴퓨터 설정에 비해 비용이 많이 들 수 있다.

내 생각에는 빅 데이터 처리를 위해 분산컴퓨팅을 활용하는 것이 일반적으로 유리하고 생각한다. 분산컴퓨팅은 대규모의 복잡한 데이터 처리 작업을 효율적으로 처리하는데 필요한 확장성, 처리능력 및 내결함성을 제공한다. 따라서 기업은 합리적인 시간 내에 대규모 데이터셋을 처리하고 분석할 수 있다. 따라서 분산컴퓨팅이 더 좋다고 생각한다.

4. 기존 데이터를 다루는 기술이 아닌, 빅데이터를 다루는 기술을 사용해서 문제를 해결한 예를 하나 설명해주세요.

은행과 증권에서의 예시 : 상위 10개 투자 및 소매 은행의 16개 프로젝트를 대상으로 한 연구에 따르면 이 업계의 당면 과제는 다음과 같다. 증권 사기 조기 경고, 틱 분석, 카드 사기 탐지, 감사

추적 기록 보관, 엔터프라이즈 신용 리스크 보고, 거래 가시성, 고객 데이터 변환, 거래를 위한 소셜 분석, IT 운영 분석, IT 정책 컴플라이언스 분석 등을 수행할 수 있다.
은행과 증권업계에서의 빅데이터 활용 방안 : 증권거래위원회는 빅데이터를 활용해 금융시장 활동을 모니터링하고 있다. 이들은 현재 네트워크 분석과 자연어 프로세스를 사용하여 금융시장에서 불법거래 행위를 적발하고 있다. 소매 거래자, 빅뱅크, 헤지펀드 등 금융시장의 이른바 '빅보이'들은 고주파 거래, 거래 전 의사결정 지원 분석, 심리 측정, 예측 분석 등에 사용되는 거래 분석에 빅데이터를 활용한다. 또한 이 업계는 자금 세탁 방지, 요구 사항 엔터프라이즈 리스크 관리, "고객 파악" 및 부정 행위 방지를 포함한 리스크 분석을 위해 빅 데이터에 크게 의존하고 있습니다.

5. throughput : 중앙 처리장치가 단위 시간에 처리할 수 있는 데이터 처리능력이다.
Challenges in handling bigdata : (storage) 스토리지 비용, 처리량의 확장성
(Network) 국제 커뮤니티 간에 공유되는 대용량 데이터 세트, 로컬 영역 네트워크를 제공하여 연구자에게 즉시 데이터 제공
(Data integrity) 여러 개의 복사본 또는 일부 형태의 백업
메타 데이터, 출처 및 ontologies

7. 가트너 그룹은 "The Big Data Value Model"을 통해 빅데이터 분석의 주요 활용 목적을 아래와 같이 분류했습니다. 아래에 대해서 사례를 기반으로 정리 해주세요

Customer insight : (사우스웨스트 항공의 고객 맞춤형 광고) 저가 항공의 대명사인 사우스웨스트 항공은 비행기 좌석 스크린에 승객별로 다른 광고를 제공하고 있는데, 미국인의 96%를 비롯해 전세계적으로 5 억 명에 달하는 고객 정보를 갖고 있는 액시엄(Acxiom)社의 DB 에저장되어 있는 항공기 탑승객의 쇼핑 습관과 구매 패턴 등을 분석한 후 승객별 최적화된 광고를 제공하고 있다.
Process efficiency : (구글의 데이터센터 성능 및 에너지 사용 최적화) 구글은 데이터센터 서버와 기타 장비들의 사용시간과 에너지 사용량에 대한 방대한 분량의 운영 데이터를 분석하여 데이터센터의 성능과 에너지 사용량이라는 트레이드오프 관계에 있는 두 가지 지표를 최적의 상태로 운영하고 있다.
Digital products & service : (후지쯔(Fujitsu)의 농업용 빅데이터 분석 솔루션) 후지쯔는 농지작업 실적과 작물 이미지 등 데이터를 분석해 수확량 증가와 품질을 향상시키는 클라우드 기반의 농업용 빅데이터 분석 솔루션을 2012 년부터 제공하고 있다. 이는 기후와 토양환경 등에 대해 센서로부터 수집되는 데이터와 과거 수확실적 등을 비롯한 빅데이터를 분석하여 최적의 파종, 농약 살포, 수확 시점을 제공하는 솔루션이다
Operational excellence, digital marketing : (아마존의 예측 배송) 빅데이터 분석을 이용한 고객 이해와 구매 추천의 선구자인 아마존은 '예측 배송'이라는 또 다른 파격적 행보를 시도할 계획인데, 이를 위해 2013 년 12 월 고객이 구매하기 전에 배송을 준비하는 '예측배송(anticipatory shipping)' 서비스에 대한 특허를 취득한 바 있다.

8. GE 의 산업 인터넷(Industrial Internet)에 대해서 조사해 주세요.

GE 의 산업 인터넷은 산업 분야에서 사물인터넷(IoT)과 고급 분석의 힘을 활용하기 위한 GE 의 이니셔티브와 접근 방식을 의미한다. 기계 및 장비와 같은 물리적 산업 자산을 디지털 기술과 연결하여 데이터를 수집하고 통찰력을 얻어 운영 효율성, 생산성 및 비용 효율성을 개선하는 데 중점을 둔다.

Predix Platform: GE 는 산업 데이터 및 분석을 위해 특별히 설계된 클라우드 기반 플랫폼인 Predix 플랫폼을 개발했다. 연결된 산업 자산에서 데이터를 수집, 분석 및 저장할 수 있는 기반을 제공한다. Prefix 는 산업 운영의 실시간 모니터링, 예측 유지보수 및 최적화를 가능하게 한다.

자산 성과 관리: GE 의 산업 인터넷은 산업 자산에서 수집된 데이터에 고급 분석을 적용하여 자산 성과를 향상시키는 것을 목표로 한다. 예측 유지보수를 지원하므로 기업은 잠재적인 장비 장애가 발생하기 전에 이를 식별하고, 다운타임을 최소화하며, 유지보수 일정을 최적화할 수 있다.

Digital Twin: GE 는 "Digital Twin" 개념을 활용하여 물리적 자산의 가상 복제본을 만든다. 디지털 트윈은 센서의 실시간 데이터와 과거 데이터를 결합하여 물리적 자산의 동작과 성능을 시뮬레이션하는 동적 모델을 제공한다. 이를 통해 운영자는 자산 성능을 모니터링 및 분석하고, 다양한 운영 시나리오를 시뮬레이션하고, 정보에 입각한 의사 결정을 내릴 수 있다.

산업용 애플리케이션: GE 의 산업용 인터넷은 항공, 에너지, 의료, 제조 및 운송을 포함한 다양한 산업에 걸쳐 애플리케이션을 제공한다. 예를 들어, 항공 분야에서 GE Aviation 은 항공기 엔진의 데이터를 활용하여 연료 효율을 최적화하고, 유지보수 필요성을 예측하고, 비행 운영을 개선한다. 의료 분야에서 GE Healthcare 는 의료 기기의 데이터를 활용하여 환자 결과를 개선하고 병원 운영을 최적화한다.

9. 아래는 정부에서 발간한 2021 데이터산업 현황조사 결과보고서 입니다.

다 읽어보시면 좋지만 제 2 부의 "1 장 데이터산업 시장규모", "2 장 데이터직무 인력 현황 및 수요" 은 필수적으로 읽어보시고 정리해 주세요

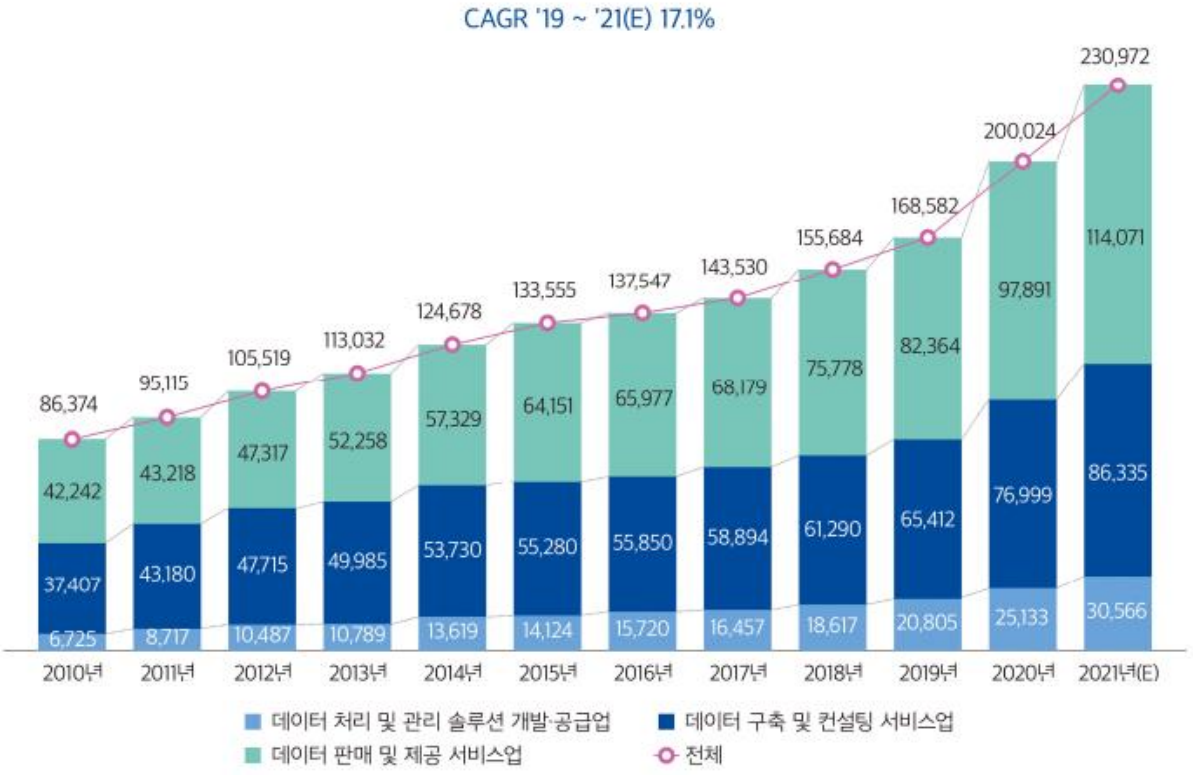
https://www.kdata.or.kr/kr/board/info_01/boardView.do?pageIndex=1&bbsIdx=33253&searchCondition=all&searchKeyword=

1 장 - 1. 전체 시장규모

데이터산업 시장규모는 2020 년 전년 대비 18.7% 성장한 20 조 24 억 원이며, 2021 년에는 23 조 972 억 원 규모로 성장할 것을 예상하였고, 지속적인 성장세를 이어 갈 것으로 조사되었다. 데이터 산업의 부문별 규모는 2020 년 기준 데이터 판매 및 제공 서비스업 시장이 9 조 7,891 억 원으로 가장 높고, 다음으로 데이터 구축 및 컨설팅 서비스업이 7 조 6,999 억 원, 데이터 처리 및 관리 솔루션 개발·공급업이 2 조 5,133 억 원으로 나타났다.

[그림 2-1] 2010~2021년(E) 데이터산업 시장규모

(단위 : 억 원)



▌2014년 이전 통계는 통계작성승인(2016년) 이전에 도출된 시범조사 결과임

향후 데이터산업 시장은 지난 5개년 연평균 성장률인 12.6%와 같이 지속적으로 성장한다면 2027년까지 47조 원을 넘어설 것으로 전망된다.

[표 2-2] 2021(E)~2027(P) 데이터산업 시장 전망

(단위 : 억 원)

구 분	2021년(E)	2022년(P)	2023년(P)	2024년(P)	2025년(P)	2026년(P)	2027년(P)
데이터산업 시장규모	230,972	260,144	293,000	330,005	371,685	418,629	471,501

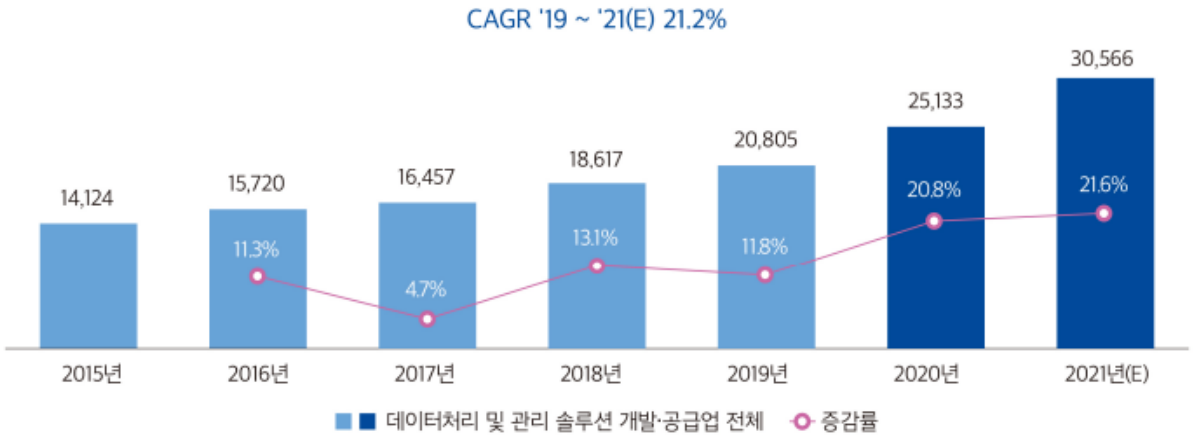
2. 부문별 시장 규모

(1) 데이터 처리 및 관리 솔루션 개발·공급업 시장

데이터 처리 및 관리 솔루션 개발·공급업 시장규모는 2020년 2조 5,133억 원으로 2019년 대비 20.8% 성장하였으며, 2019년부터 2021년 예상 매출까지 3개년 연평균 증감률(CAGR)은 21.2%로 데이터산업 전체 성장세(17.1%)보다 높게 나타났다.

[그림 2-4] 데이터 처리 및 관리 솔루션 개발 · 공급업 시장규모

(단위 : 억 원)

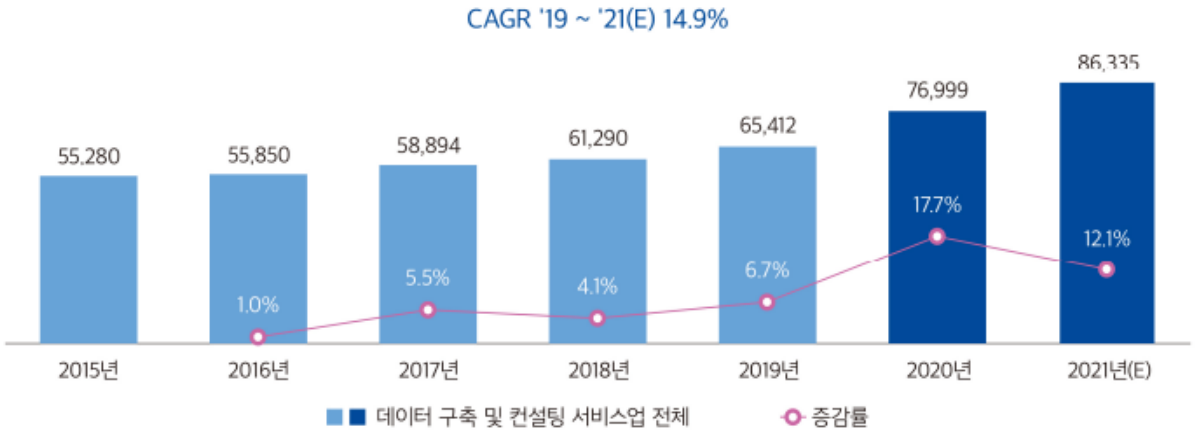


(2) 데이터 구축 및 컨설팅 서비스업 시장

데이터 구축 및 컨설팅 서비스업 시장규모는 2020년 7조 6,999억 원으로 2019년 대비 17.7% 성장하였으며, 2019년부터 2021년 예상 매출까지 3개년 연평균 증감률(CAGR)은 14.9%로 나타났다.

[그림 2-6] 데이터 구축 및 컨설팅 서비스업 시장규모

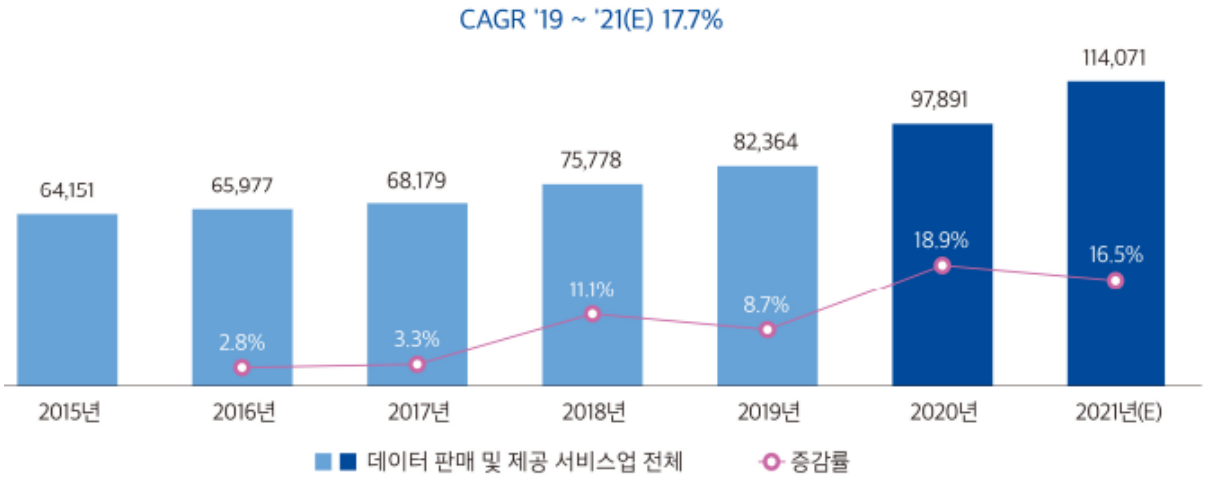
(단위 : 억 원)



(3) 데이터 판매 및 제공 서비스업 시장

데이터 판매 및 제공 서비스업 시장규모는 2020년 9조 7,891억 원으로 2019년 대비 18.9% 성장하였다. 데이터 판매 및 제공 서비스업의 시장규모는 2015년 이후 지속적으로 성장하고 있으며, 2019년부터 2021년 예상 매출까지 3개년 연평균 증감률(CAGR)은 17.7%로 나타났다.

(단위 : 억 원)



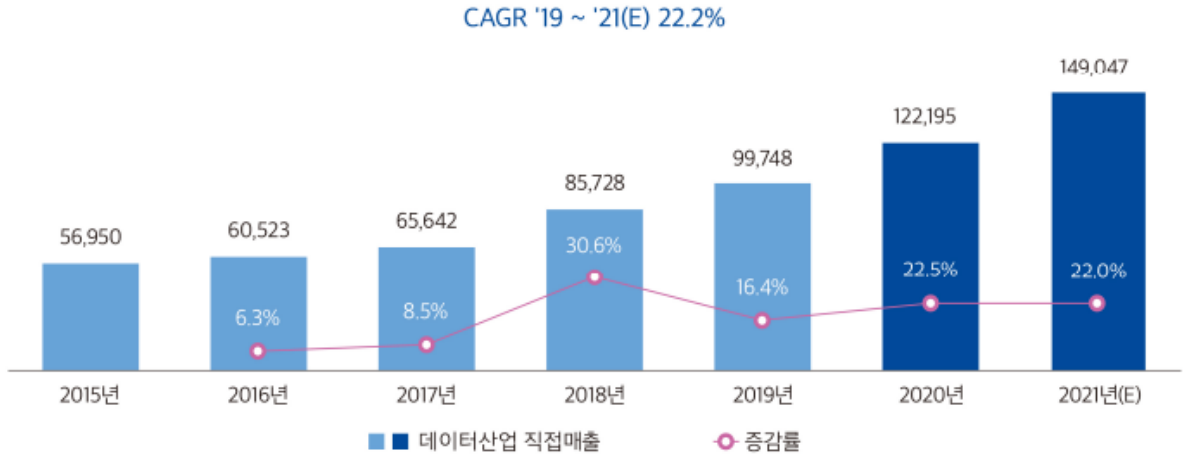
3. 직접매출 시장규모

(1) 전체 직접매출 시장

2020년 데이터산업 직접매출12) 시장규모는 12조 2,195억 원으로 전년 대비 22.5%가 성장하였으며, 데이터산업 전체 증감률(18.7%) 대비 높은 증감률을 보였다. 2019년에서 2021년 예상 매출의 연평균 증감률(CAGR) 또한 직접매출 시장이 22.2%로 전체 데이터산업의 동기간 연평균 증감률(17.1%)보다 높게 나타났다.

[그림 2-12] 데이터산업 직접매출 시장규모

(단위 : 억 원)



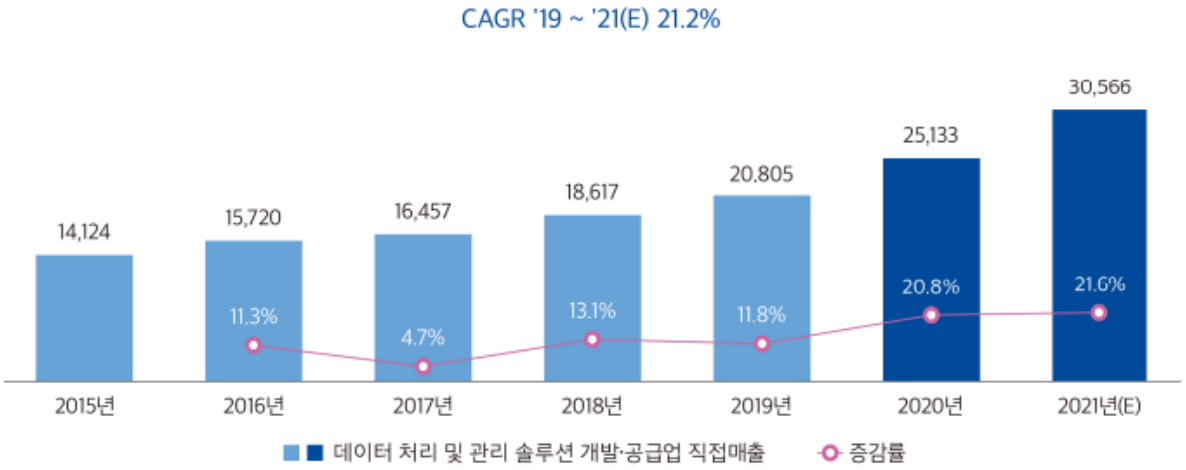
(2) 부문별 직접매출 시장

1) 데이터 처리 및 관리 솔루션 개발·공급업 시장

2020년 데이터 처리 및 관리 솔루션 개발·공급업의 직접매출은 전년 대비 20.8% 성장한 2조 5,133억 원으로 나타났다. 세부 영역별로는 데이터 보안 솔루션 개발·공급업이 전년 대비 29.5% 크게 성장하였고, 다음으로 데이터 분석 솔루션 개발·공급업, 빅데이터 통합 플랫폼 솔루션 개발·공급업(각 28.4%) 순으로 증감률이 높게 나타났다. 또한 2019년부터 2021년 예상 매출의 연평균 증감률(CAGR)은 21.2%로 꾸준한 성장세를 보였다.

[그림 2-14] 데이터 처리 및 관리 솔루션 개발 · 공급업 직접매출 시장규모

(단위 : 억 원)

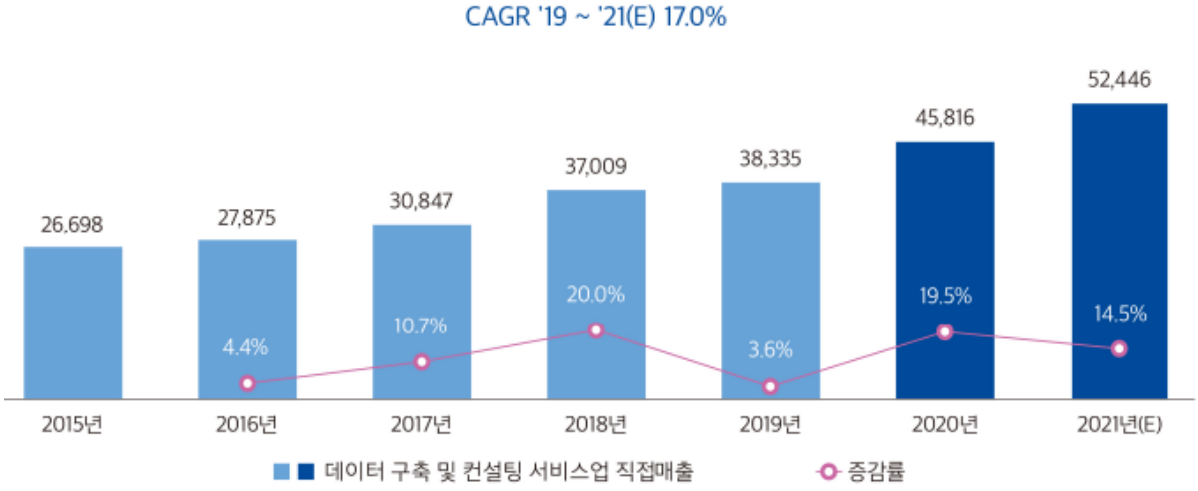


2) 데이터 구축 및 컨설팅 서비스업 시장

2020년 데이터 구축 및 컨설팅 서비스업의 직접매출은 전년 대비 19.5% 성장한 4조 5,816억 원이며, 2019년부터 2021년 예상 매출의 연평균 증감률(CAGR)은 17.0%로나타났다. 세부 영역별로는 데이터 관련 컨설팅 서비스업이 전년 대비 31.5% 크게 성장하였고, 데이터 구축·가공 서비스업은 18.4% 증가한 것으로 나타났다.

[그림 2-15] 데이터 구축 및 컨설팅 서비스업 직접매출 시장규모

(단위 : 억 원)

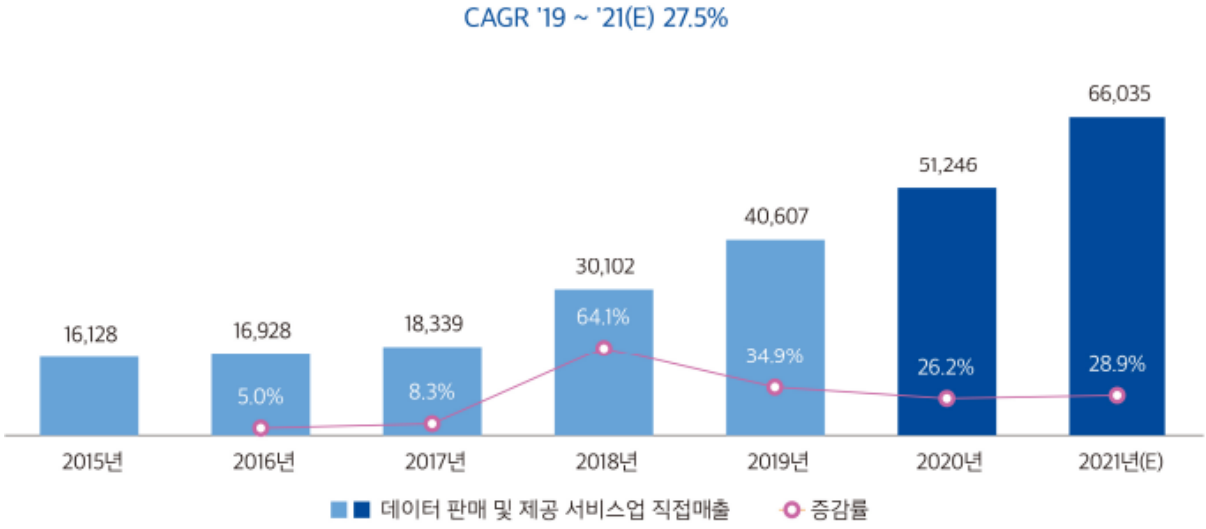


3) 데이터 판매 및 제공 서비스 시장

2020년 데이터 판매 및 제공 서비스업 시장의 직접매출은 5조 1,246억 원으로 전년 대비 26.2% 증가한 것으로 나타났다. 세부 부문별로는 데이터 판매·중개 서비스업에서 34.2%, 정보제공 서비스업에서 23.7%로 모두 크게 성장한 것으로 나타났다. 2019년부터 2021년 예상 매출의 연평균 증감률(CAGR)은 27.5%로 높은 성장율을 보였다.

[그림 2-18] 데이터 판매 및 제공 서비스업 직접매출 시장규모

(단위 : 억 원)



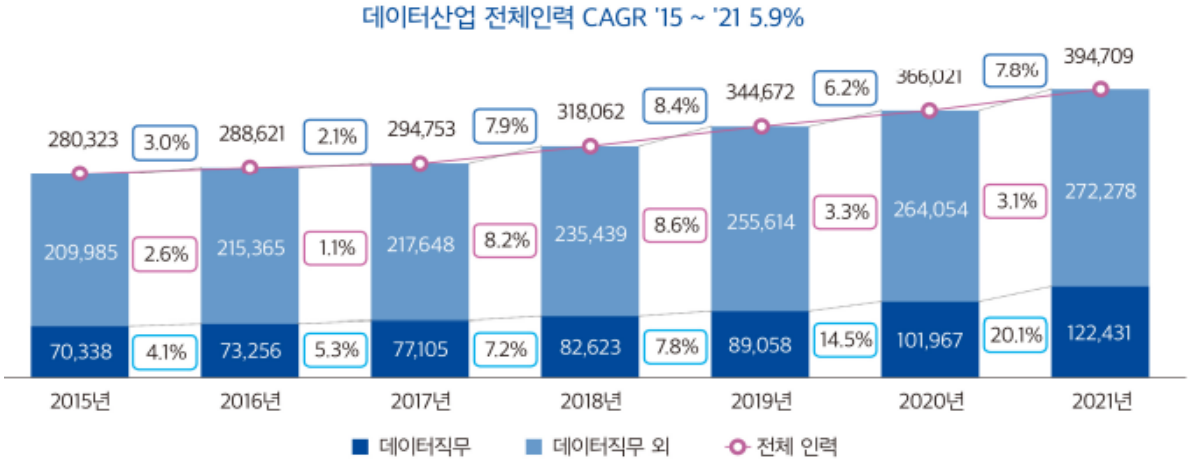
2장 – 1. 데이터직무 인력 현황

(1) 데이터산업의 종사자 현황

2021년 데이터산업에 종사하고 있는 인력은 총 394,709명으로 전년 대비 7.8% 증가했으며, 이 중 데이터직무14) 인력은 122,431명으로 전년 대비 20.1% 증가한 것으로 나타났다. 2015년부터 2021년까지 데이터직무 인력은 연평균 9.7% 증가하였고, 데이터직무 외 인력을 포함한 전체 인력의 연평균 증감률은 5.9%로 나타났다.

[그림 2-23] 2015~2021년 데이터산업 인력 현황

(단위 : 명)



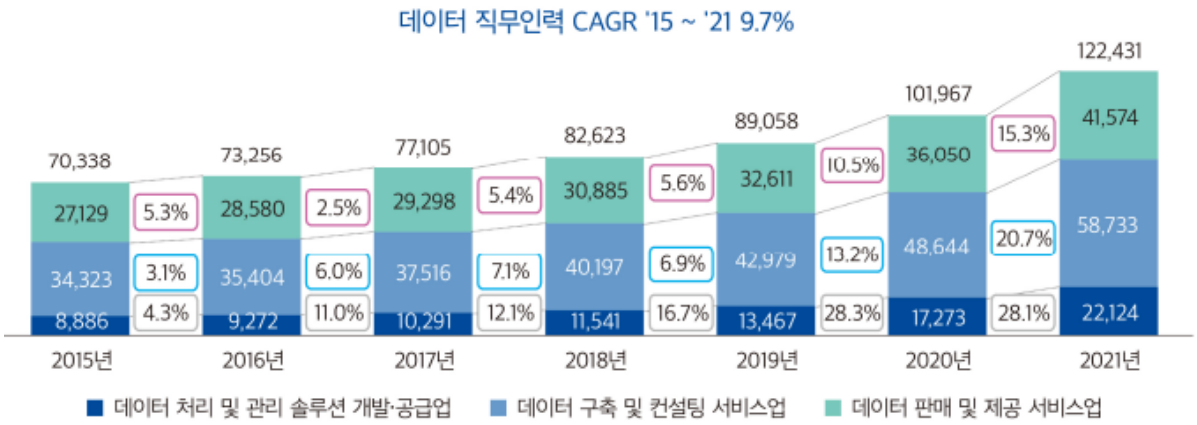
(2) 데이터산업의 데이터직무 인력 현황

1) 데이터산업 부문별

2021년 데이터산업 부문별 데이터직무 인력은 데이터 구축 및 컨설팅 서비스업 부문이58,733명으로 가장 많았고, 데이터 판매 및 제공 서비스업 41,574명, 데이터 처리 및 관리 솔루션 개발·공급업 22,124명 순으로 나타났다. 데이터산업 부문별로 증감률을 살펴보면, 데이터 처리 및 관리 솔루션 개발·공급업이 전년 대비 28.1%로 가장 높은 증가율을 보였고, 데이터 구축 및 컨설팅 서비스업은 20.7%, 데이터 판매 및 제공 서비스업은 15.3% 증가한 것으로 나타났다.

[그림 2-24] 2015~2021년 데이터산업 부문별 데이터직무 인력 현황

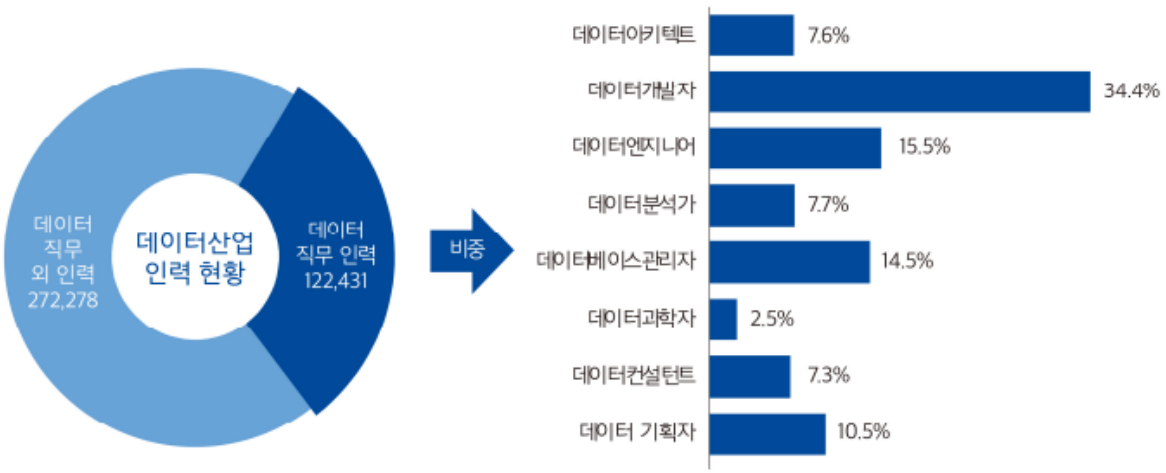
(단위 : 명)



데이터산업 인력의 직무별 비중을 보면, 데이터 개발자가 34.4%(42,128명)로 가장 많은 비중을 차지하였고, 데이터 엔지니어 15.5%(18,964명), 데이터베이스관리자 14.5%(17,706명) 순으로 나타났다.

[그림 2-25] 2021년 데이터산업 인력 구성 및 데이터직무별 인력 비중

(단위 : 명)



2. 데이터직무 인력 수요

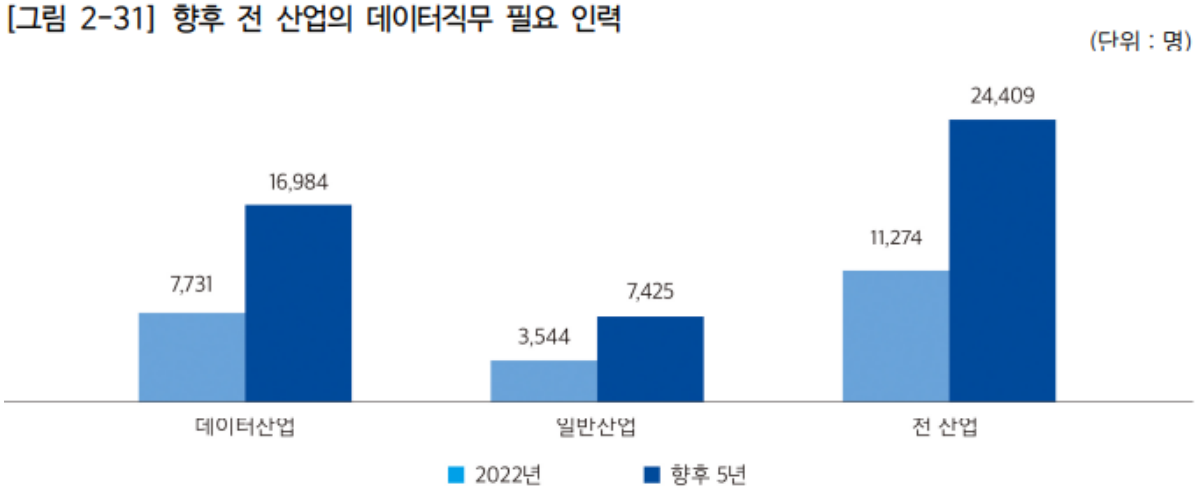
(1) 데이터산업의 필요 인력 및 부족률

향후 5년 내(2026년까지) 데이터산업의 데이터직무 필요 인력은 총 16,984명으로조사되었다. 이 중 데이터 개발자 수요가 8,035명(47.3%)으로 가장 높았고, 데이터 엔지니어 2,131명(12.5%), 데이터 분석가 1,744명(10.3%) 순으로 나타났다.

(2) 전 산업의 필요 인력 및 부족률

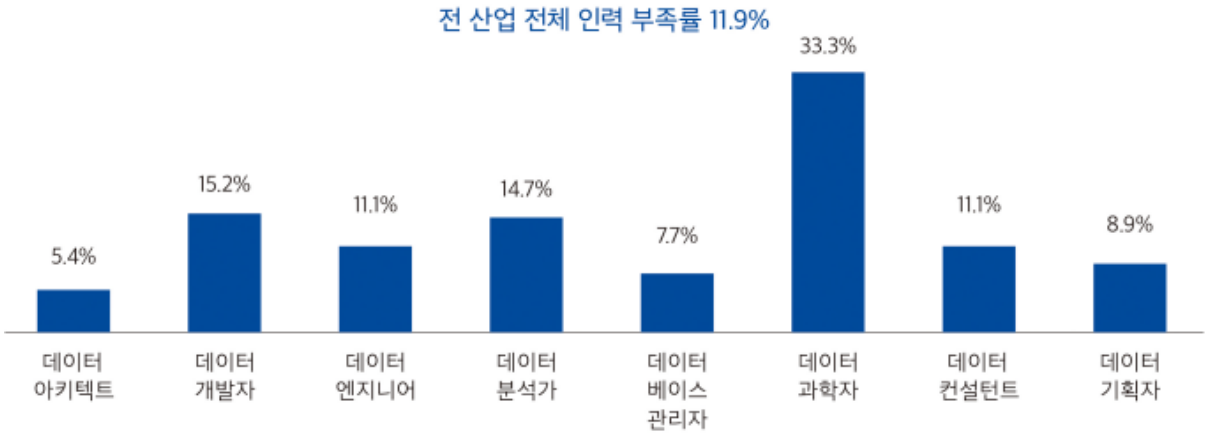
향후 5년 내 일반산업을 포함한 전 산업에서 필요한 데이터직무 인력은 총 24,409명으로, 이 중 데이터 개발자가 9,247명(37.9%)으로 가장 높게 나타났고, 데이터베이스관리자 3,765명(15.4%), 데이터 엔지니어 2,864명(11.7%), 데이터 분석가 2,821명(11.6%) 순으로 나타났다.

[그림 2-31] 향후 전 산업의 데이터직무 필요 인력



향후 5년 내 일반산업을 포함한 전 산업 내 데이터직무별 인력 부족률은 평균 11.9% 수준이며, 데이터 과학자 부족률이 33.3%로 가장 높게 나타났다. 데이터 개발자(15.2%), 데이터 분석가(14.7%) 직무가 전 산업 평균보다 높은 부족률을 보였다.

[그림 2-33] 향후 5년 내 전 산업의 데이터직무 인력 부족률



10. 데이터 엔지니어와 데이터 사이언티스트의 차이점은 무엇인지 조사해주세요 아래는 참조 링크 입니다.

데이터 엔지니어와 데이터 사이언티스트는 데이터 및 분석 분야에서 서로 다른 두가지 역할을 한다. 그들은 종종 협력하고 함께 일하지만, 이들의 책임과 기술은 다르다.

데이터 엔지니어
역할: 데이터 엔지니어는 주로 데이터 인프라 및 시스템의 개발 및 관리에 집중한다. 이들은 데이터 수집, 저장 및 처리를 가능하게 하는 시스템과 파이프라인을 설계, 구성 및 유지 관리하는 역할을 담당한다.

데이터 인프라: 데이터 엔지니어는 데이터 파이프라인, 데이터베이스 및 데이터 웨어하우스를 구축하고 관리한다. 데이터를 수집, 치료, 변환하고 분석을 위해 액세스할 수 있도록 보장한다.

프로그래밍 및 도구: 데이터 엔지니어는 일반적으로 Python, Java 또는 Scala와 같은 프로그래밍 언어에 대한 전문 지식을 가지고 있다. Apache Hadoop, Spark 및 SQL 데이터베이스와 같은 툴로 작동한다.

데이터 품질 및 통합: 데이터 엔지니어는 데이터 품질, 안정성 및 일관성을 보장한다. 다양한 소스의 데이터를 통합하고 데이터 무결성 및 보안을 유지한다.

스킬: 강력한 프로그래밍 기술, 데이터 모델링, 데이터베이스 설계, ETL(Extract, Transform, Load) 프로세스, 데이터 통합 및 데이터 파이프라인 개발 경험.

데이터 사이언티스트
역할: 데이터 사이언티스트는 복잡한 데이터를 분석하고 해석하여 패턴과 통찰력을 파악하고 데이터 중심의 의사 결정을 내리는 데 중점을 둔다. 그들은 귀중한 통찰력을 추출하고 비즈니스 문제를 해결하기 위해 통계 및 기계 학습 기술을 적용한다.

데이터 분석 및 모델링: 데이터 사이언티스트는 탐색적 데이터 분석을 수행하고, 통계 모델을 개발하고, 예측 및 규범적 모델을 만든다. 그들은 자신들의 도메인 전문 지식을 활용하여 실행 가능한 통찰력을 찾는다.

머신러닝: 데이터 사이언티스트들은 머신러닝 알고리즘과 기술에 대한 전문 지식을 가지고 있다. 그들은 분류, 회귀, 클러스터링 및 추천 시스템과 같은 작업을 위한 모델을 구축하고 배포한다.

프로그래밍 및 도구: 데이터 사이언티스트는 데이터 분석, 시각화 및 모델 구축을 위해 Python 또는 R과 같은 프로그래밍 언어를 사용한다. TensorFlow, scikit-learn, PyTorch와 같은 라이브러리와 프레임워크로 작업한다.

Business Context: 데이터 사이언티스트는 비즈니스 목표를 이해하고 데이터 통찰력을 사용하여 전략적 의사 결정을 안내한다. 그들은 비기술적 이해관계자들에게 결과를 효과적으로 전달한다.

기술: 통계 분석, 기계 학습, 데이터 시각화, 프로그래밍 및 도메인 지식에 능숙합니다. 강력한 문제 해결 능력과 의사소통 능력.