IDC TECHNOLOGY SPOTLIGHT

연구 의뢰: IBM

디지털 시대에 인공지능(AI)은 비즈니스 인텔리전스에서 가장 중대한 위치를 차지하게 되었습니다. AI가 보여줄 수 있는 역량과 약속이 많은 기대감을 갖게 하지만, 조직의 입장에서는 개념 증명을 거쳐 프로덕션에서 확장 단계까지 성공적으로 나아가야 하는 과제를 안게 되었습니다.

AI에 최적화된 인프라를 활용한 AI 배포 가속화 및 운영화

2018년 6월

작성자: RituJyoti,부사장

서론

개요

주요 통계

IDC는 2019년까지 DX 이니셔티브 중 40%가 AI 서비스를 활용할 것이고, 2021년까지는 상용 엔터프라이즈 앱의 75%가 AI를 활용할 것이며, 소비자 중 90% 이상이 고객 지원 로봇의 지원을 받고 새로운 산업용 로봇의 50% 이상이 AI를 활용할 것으로 예측합니다.

디지털 혁신(digital transformation, DX)은 거시경제의 규모로 확대되고 있습니다. 인공지능(artificial intelligence, AI), 머신러닝(machine learning, ML) 및 지속적 딥러닝(deep learning, DL) 기반의 지능형 애플리케이션은 소비자와 기업이 일하고 학습하고 즐기는 방식을 혁신하는 차세대 기술입니다. 데이터는 새로운 디지털 경제에서 핵심적 위치를 차지하지만 사용자가 환경을 인식하고 엣지부터 코어, 그리고 클라우드까지 데이터를 관리하고 거의 실시간으로 분석한 다음 그 결과를 통해 학습하고 이를 기반으로 조치를 실행하여 결과에 영향을 미치는 방식도 중요한 역할을 수행합니다. 사물인터넷(Internet of Things, IoT), 모바일 기기, 빅데이터, AI, ML, DL은 모두 결합되어 환경을 연속적으로 인식하고 이를 통해 총체적으로 학습합니다. 성공하는 조직을 차별화하는 요인은 조직이 이와 같은 기술을 활용하여 산업 공정, 의료, 경험적 참여 또는 기타 기업의 의사결정을 향상하기 위해 의미 있고 부가가치를 창출하는 예측과 활동을 제시하는 방식입니다. AI와 관련된 비즈니스 목표는 전술적 목표와 전략적 목표 사이에서 균형을 이루어야 합니다. 이러한 비즈니스 목표는 운영 효율성 향상부터 경쟁력 제고를 위한 차별화 강화, 그리고 기존 제품의 수익 극대화부터 새로운 디지털 수익원 창출에 이르기까지 다양할 수 있습니다.

AI는 수십 년 간 우리 주변에 존재해왔지만 점점 더 깊숙이 침투하고 있는 데이터, 클라우드 컴퓨팅의 확장성, 이용 가능한 AI 촉진 요인, ML 및 DL 알고리즘의 정교화로 인해 이제 비즈니스 인텔리전스에서 중심적 위치를 차지하게 되었습니다. IDC는 2019년까지 DX 이니셔티브 중 40%가 AI 서비스를 활용할 것이고, 2021년까지는 상용 엔터프라이즈 앱의 75%가 AI를 활용할 것이며, 소비자 중 90% 이상이 고객 지원 로봇의 지원을 받고 새로운 산업용 로봇의 50% 이상이 AI를 활용할 것으로 예측합니다.

그러나, DX 이니셔티브의 거의 절반에서 AI가 중대한 역할을 수행하게 되면 IT 부문에서 기술 인력을 확보해야 하는 새로운 부담이 발생할 것입니다. IDC는 2020년까지 데이터 중심 DX 프로젝트를 개발할 수 있도록 새로 고용하는 운영 부문 기술 인력의 85%가 분석 및 AI 관련 기술을 기준으로 선별될 것이라고 예측합니다. 이와 동시에, CIO는 새로운 운영 및 수익 창출 모델을 지원하는 통합된 엔터프라이즈 디지털 플랫폼을 만들고 지속적으로 개선해야 합니다. 회사 내에서 IT 조직은 개발, 데이터 관리, 사이버 보안 측면에서 AI를 활용하는 최적이자 최초의 사용 사례 환경 중 하나가 되어야 합니다.

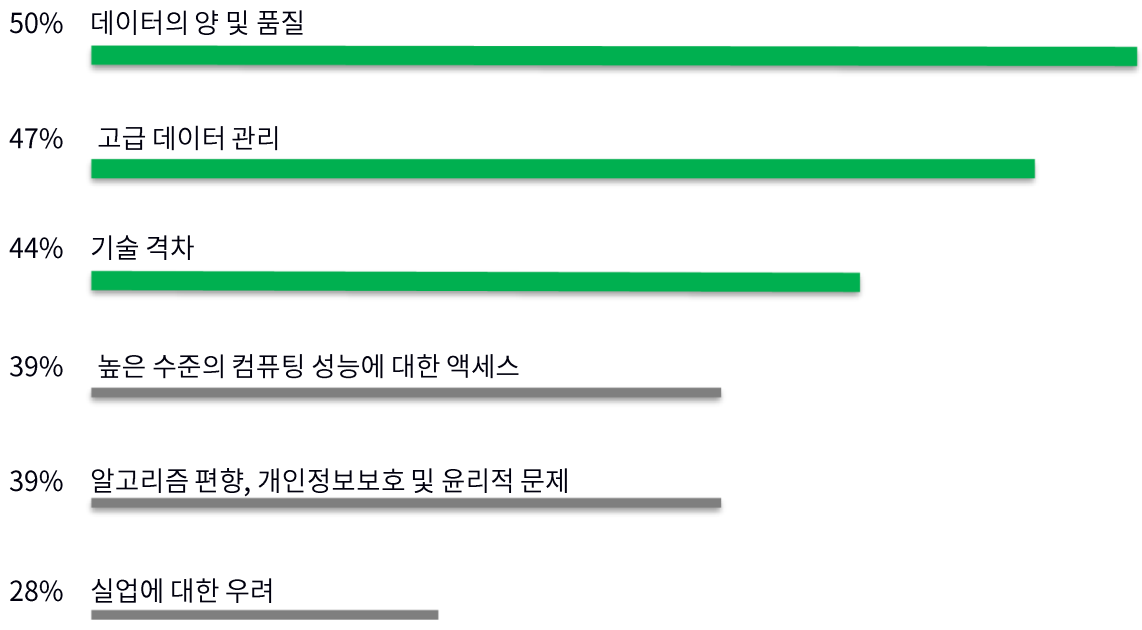
AI 모델 및 워크로드 배포: 과제 및 요구사항

AI는 디지털 시대에 비즈니스 프로세스를 수행하는 방식을 바꾸고 있습니다. AI가 보여줄 수 있는 역량과 약속이 많은 기대감을 갖게 하지만, AI 모델과 워크로드를 배포하는 일은 쉬운 일이 아닙니다. 부푼 기대에도 불구하고 대부분의 조직은 개념 증명(proof of concept, POC)을 완료하는 것조차 어려움을 겪고 있으며 완전한 프로덕션 단계로 진입한 경우는 극소수에 불과합니다.

문제는 ML 및 DL 알고리즘이 거대한 양의 훈련 데이터(일반적으로 기존 분석에 사용되는 데이터 양의 8~10배)를 필요로 한다는 점입니다. AI의 유효성은 고품질의 다양하고 역동적인 입력 데이터에 따라 크게 좌우됩니다. 과거에는 데이터 분석이 대규모 파일, 순차적 액세스, 배치 데이터를 중심으로 이루어졌습니다. 요즘에는 데이터가 소규모부터 대규모까지의 파일과 정형, 반정형, 비정형 콘텐츠로 구성되어 있습니다. 데이터 액세스는 무작위부터 순차적 방식까지 다양합니다. 2025년까지 전 세계 데이터 세트의 1/4 이상이 실시간적 성격을 띠게 될 것이고 실시간 IoT 데이터가 이 중 95% 이상을 차지할 것입니다. 그 외에도, 데이터가 온프레미스, 코로케이션 및 퍼블릭 클라우드 환경 전반에 걸쳐 점점 더 분산되는 양상을 보이고 있습니다.

IDC는 2018년 1월 미국과 캐나다의 IT 및 데이터 전문가 405명을 대상으로 설문조사를 실시했습니다. 이들은 AI 프로젝트를 성공적으로 완료한 경험이 있고 예산을 관리하거나 이에 대해 영향력이 있고 AI 워크로드를 실행할 플랫폼을 평가 또는 설계하는 일을 담당한 사람들이었습니다. 이 설문조사의 목적은 조직이 AI 지원 기술을 활용하고 관리하는 방법을 알아보고 코그너티브/ML/AI 워크로드 실행에 사용된 인프라, 기술 배포 위치, 그리고 이와 관련된 과제 및 요구사항을 파악하는 것이었습니다. 그림 1에서 볼 수 있듯이, 설문조사 응답자는 방대한 양의 데이터 처리 그리고 관련된 품질 및 관리 문제를 AI 배포 시 수반되는 주요 과제로 꼽았습니다.

그림 1: *AI 워크로드 배포와 관련된 과제*



출처: Cognitive, ML, and AI Workloads Infrastructure Market Survey(코그너티브, ML, AI 워크로드 인프라 시장 설문조사), IDC, 2018년 1월, n=405, 1,000명 이상의 직원(미국), 500명 이상의 직원(캐나다)

나쁜 데이터 품질은 편향되고 정확하지 않은 모델 구축과 직접적인 상관관계가 있습니다. 개발자가 모든 적절한 검사 및 검증 절차를 이해하고 예측하여 이를 코드에 반영하는 것이 사실상 힘들기 때문에 역동적이고 다양하며 분산된 다량의 데이터 세트에서 데이터 품질을 유지하는 일은 쉬운 일이 아닙니다. 이러한 과제를 해결하기 위해 기업은 자율적으로 데이터 품질을 유지하고 검증하는 솔루션을 원합니다. 이러한 데이터 솔루션은 데이터의 예상 행동을 자동으로 학습하고, 코딩 없이 수천 가지의 데이터 검사를 생성하며, 시간 경과에 따라 검사를 업데이트 및 관리하고, 예상되거나 예상되지 않은 데이터 품질 오류를 모두 제거하여 보다 신뢰할 수 있고 활용도 높은 데이터를 만들 수 있어야 합니다.

AI를 기반으로 하는 DX 이니셔티브에서 급부상 중인 분야를 지원하려면 인재(AI 엔지니어 및 데이터 사이언티스트)가 필요합니다.

AI를 기반으로 하는 DX 이니셔티브에서 급부상 중인 분야를 지원하려면 인재(AI 엔지니어 및 데이터 사이언티스트)가 필요합니다.

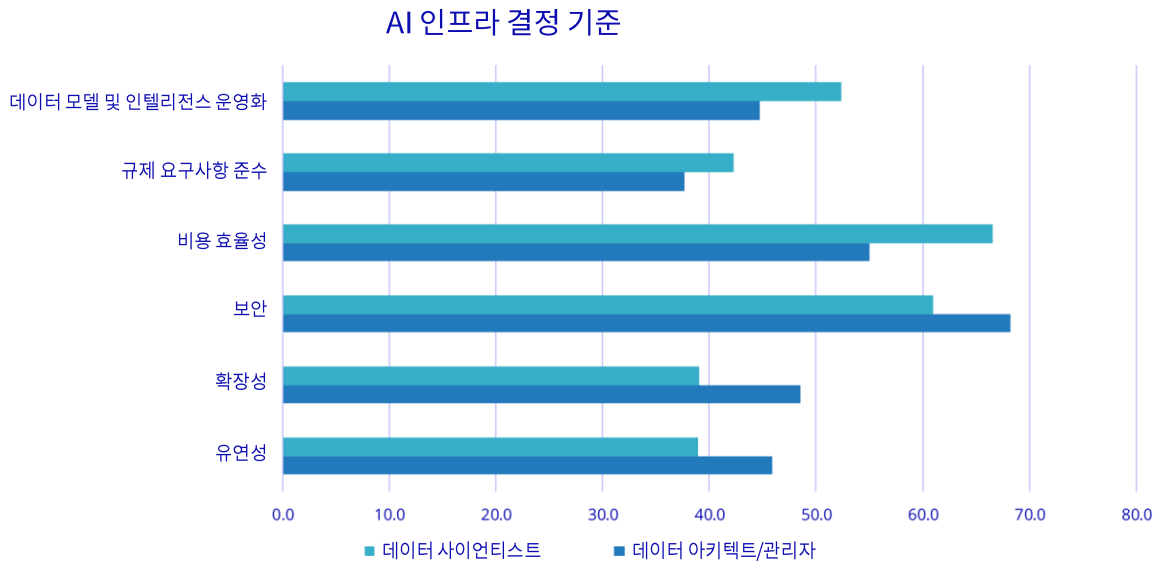
기술 격차 및 전문가 부족으로 어려움을 겪고 있습니다(그림 1 참조). 예를 들면, 모델 구축/최적화 및 훈련은 대부분의 데이터 사이언티스트가 보유하지 못한 기술인데 이러한 기술을 학습하려면 오랜 시간이 걸립니다.

기업이 AI를 도입하면 다음과 같은 인력 유형들이 복잡하게 얽혀 작업을 진행하게 됩니다.

- **데이터 사이언티스트**는 빅데이터를 취급하고 모델을 구축합니다. 이들은 한 덩어리의 데이터 포인트(비정형 및 정형)를 가지고 데이터를 정제하고 다듬고 조직화하기 위해 엄청난 수학적, 통계적, 프로그래밍적 기술을 사용합니다. 그리고 나서 산업 관련 지식, 컨텍스트에 대한 이해, 기존 가정에 대한 의심 등 모든 분석력을 적용하여 비즈니스 과제에 대한 숨겨진 해결책을 찾아냅니다. 데이터 사이언티스트가 해야 하는 일의 예는 다음과 같습니다.
  - 여러 내외부 소스에서 얻은 엄청난 양의 데이터를 추출, 정제, 가지치기
  - 정교한 분석 프로그램, ML, 통계 방법을 활용하여 예측적, 처방적 모델링에 사용할 수 있도록 데이터를 준비
  - 데이터를 살펴보고 검사하여 숨겨진 약점, 추세 또는 기회 파악
  - 문제를 해결하고 작업 자동화를 위한 새로운 툴을 만들기 위해 새로운 알고리즘 및 모델을 고안 또는 구축
  - 가장 긴급한 과제의 규모에 맞게 데이터 중심 AI 모델을 훈련, 최적화, 배포
  - AI 모델의 정확도 유지
  - 효과적인 시각화 자료와 보고서를 활용하여 경영진과 IT 부서에 예측 및 결과 전달
- **데이터 엔지니어/관리자**는 거대한 데이터 저장소를 구축합니다. 이들은 데이터베이스 및 대규모 데이터 처리 시스템과 같은 아키텍처를 개발, 구성, 테스트, 관리합니다. 이와 같이 필터링된 정보로 구성된 거대한 "풀"에 대한 지속적인 파이프라인이 설치되면 데이터 사이언티스트는 분석을 위해 관련 데이터 세트를 가져올 수 있습니다.
- **데이터/IT 아키텍트**는 AI 프레임워크를 인프라 구축 전략에 통합하며 IT 환경의 확장성, 민첩성 및 유연성을 지원하는 일을 담당합니다.

IDC가 이러한 인력 유형에 해당하는 이들에게 AI 솔루션을 선택할 때 가장 중요한 결정 기준이 무엇인지 묻자, 이들은 그림 2에서 볼 수 있듯이 데이터 모델/인텔리전스의 보안, 비용 효율성, 운영화(구축, 조정, 최적화, 훈련, 배포 및 추론)이라고 답변했습니다.

그림 2: *AI 인프라/솔루션 결정 기준*



출처: IDC, 2018

이러한 요인을 고려할 때 AI의 성공적 배포를 위한 필수 요소는 다음과 같습니다.

» **데이터 사이언티스트의 생산성**

모델의 구축, 테스트, 최적화, 훈련, 추론 및 정확도 유지는 AI 워크플로우에 필수적입니다. 이러한 신경망 모델은 구축하기 어렵습니다. 대규모 ML/DL 모델을 구축, 테스트, 배포하기 위해 데이터 사이언티스트는 일반적으로 R 또는 Python과 같은 프로그래밍 소프트웨어뿐만 아니라 Tensorflow 및 Caffe와 같은 오픈소스 프레임워크와 함께 RStudio, Spark와 같은 다양한 툴을 사용합니다. 그러나 오픈소스 프레임워크를 선택 및 설치하고 모델링 프로세스를 시작하는 작업은 복잡한 일이 될 수 있으며 프로세스를 작동시키는 데 수주 또는 수개월이 걸릴 수 있습니다. 모델 구축 및 최적화를 위해서는 수천 가지 조합의 하이퍼 파라미터를 수동으로 테스트해야 할 수 있습니다.

모델 훈련은 일부 사용 사례의 경우 완료하는 데 수주 또는 수개월이 걸릴 수 있습니다. 예를 들면, 한 의료 기관에서 조기 암 발견을 위한 의료 모델을 구축하고 훈련하는 데 1년이 걸렸습니다.

훈련은 반복적 속성을 지니므로 수 시간, 수일 또는 수주가 소요되는 수백만 가지 작업이 필요할 수 있습니다. 현재로서는 이 프로세스에서 훈련 작업이 성공적인지 확인하려면 작업을 완료해야 합니다. 즉, 조직은 일주일 동안 훈련 작업을 실행한 후에야 이 작업이 수렴하지 않는다는 것을 깨달을 수 있습니다. 서버/GPU에 장애가 발생할 경우에는 훈련 작업을 다시 시작해야 할 수도 있습니다. 이럴 경우 모든 것을 처음부터 다시 시작해야 합니다.

데이터 사이언티스트는 여러 소스에서 얻은 데이터를 정제하고 노이즈를 줄이는 작업을 효율적으로 수행하고 자동화하기를 원합니다. 또한, 하이퍼 파라미터 설정을 결정하는 작업의 간소화를 포함하여 모델에 적합한 기능을 구축, 조정 및 선택하는 데 도움이 필요합니다.

데이터 사이언티스트는 반복적이고 주기적인 프로세스를 신속하고 민첩하게 수행하기를 원합니다. 또한, 긴 훈련 시간을 줄이기 위해 훈련 단계에서 탁월한 성능을 활용하고 인프라 리소스를 유연하게 확장하기를 원하며, 분산 클러스터 환경에서 작업을 쉽게 실행할 수 있는 툴을 원합니다.

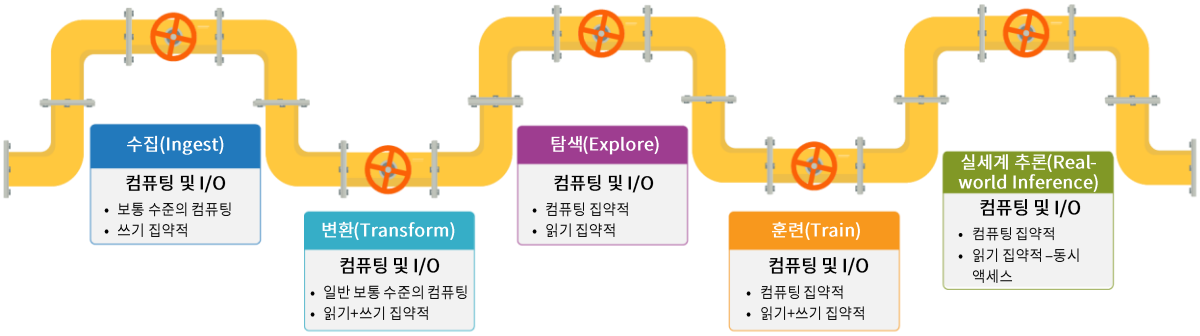
민첩한 워크로드 관리 능력, 특히 성능과 비용 효율을 위해 리소스 활용률을 극대화하여 작업을 더욱 효율적으로 실행할 수 있는 능력, 그리고 작업이 수렴하지 않을 경우 해당 작업을 시각화하고 중단할 수 있는 능력이 필요합니다. 또한, 이전 서버/GPU에 장애가 발생하면 다른 서버/GPU로 작업을 자동으로 할당하는 소프트웨어도 필요합니다.

» **인프라 최적화 및 데이터 관리 효율화**

그림 3에 제시된 AI 워크로드용 데이터 파이프라인을 보면, 애플리케이션 프로파일, 컴퓨팅 및 I/O 프로파일이 수집에서 실세계 추론으로 진행되는 것을 볼 수 있습니다. ML과 DL에는 대량의 훈련 데이터가 필요합니다. 훈련과 추론 모두 컴퓨팅 집약적이므로 빠른 실행을 위해 뛰어난 성능을 필요로 합니다. AI 애플리케이션은 수천 개의 GPU 코어 또는 수천 개의 CPU로 구성된 서버를 한계까지 밀어붙입니다. 따라서, AI 및 DL에는 기본적으로 GPU를 기반으로 하는 가속화된 새로운 유형의 인프라가 필요합니다. 신경망 모델을 훈련하는 데 필요한 선형적 수학 계산에는 가속화되지 않은 시스템 여러 대로 구성된 클러스터 하나보다 GPU로 구성된 하나의 단일 시스템이 훨씬 더 효과적입니다.

그러나, 모든 AI 구축 환경이 똑같지는 않습니다. 조직은 AI 구축 환경을 위한 성능, 운영 환경, 필요한 기술, 비용, 에너지 요구사항을 기준으로 다양한 이종의 프로세싱 아키텍처(예: GPU, FPGA, ASIC 또는 Manycore 프로세서)에 대해 알아봐야 합니다.

그림 3: *AI 워크로드용 데이터 파이프라인*



출처: IDC, 2018

우리는 또한 병렬 컴퓨팅에는 병렬 스토리지가 필요하다는 점도 알고 있습니다. 훈련 단계에서는 대규모 데이터 저장소가 필요하지만 추론 단계에서는 이에 대한 필요성이 줄어듭니다. 추론 모델은 흔히 DevOps 스타일의 저장소에 저장되며 이러한 저장소에서는 극도로 낮은 레이턴시로 액세스할 수 있는 이점이 있습니다. 실행 모델이 데이터를 기반으로 개발되고 워크로드가 추론 단계로 이동한 후에 훈련 단계가 설정되며, 새로운 데이터나 수정된 데이터가 발견되면 모델을 재훈련해야 하는 경우가 많습니다.

경우에 따라 애플리케이션의 실시간적 속성 때문에 거의 지속적인 모델의 재훈련 및 업데이트가 필요할 수 있습니다. 또한 추가적인 데이터 소스와 인사이트를 얻을 수 있으므로 시간 경과에 따른 모델 재훈련은 조직에 도움이 될 수 있습니다.

파이프라인을 따라 데이터가 원활하게 흐르지 않을 경우 생산성이 저하되므로 조직은 파이프라인 관리에 더 많은 노력과 리소스를 투입해야 합니다. 데이터 아키텍트와 엔지니어는 비용을 통제하면서 AI 워크로드의 민첩성, 유연성, 확장성, 성능, 보안 및 규정 준수 요구사항을 충족해야 하는 과제를 안고 있습니다.

분명한 점은 기업이 기존 인프라에서 AI와 같은 첨단 툴을 지원할 수 없다는 것입니다. 기존 인프라는 확장성, 탄력성, 컴퓨팅 능력, 성능 및 데이터 관리에 필요한 요구사항을 충족시키는 데 한계가 있습니다. 현재 조직들은 AI를 위한 데이터 파이프라인을 지원하기 위해 여러 가지 인프라 솔루션과 접근법을 활용하고 있습니다. 그러나 이러한 프로세스는 일반적으로 데이터 사일로를 초래합니다. 안정적인 애플리케이션을 방해하지 않으려고 파이프라인을 위해 중복된 데이터 복제본을 만드는 조직도 있습니다. 그러나, 이러한 조치를 취하기 보다는 동적으로 조정 및 확장 가능하며 지능적인(자가 구성, 자가 최적화, 자가 복구가 가능한) 인프라를 채택해야 합니다. 이러한 인프라는 다양한 데이터 형식과 액세스에 맞게 조정되어 있으며 대량의 데이터를 처리하고 분석할 수 있습니다. 또한, 더 빠른 컴퓨팅 계산 및 의사결정을 위한 속도를 구현하고 위험을 관리하며 AI 구축 비용을 전반적으로 낮출 수 있습니다.

» **엔터프라이즈 수준에서의 활용 가능성**

기업은 새로운 기술과 프레임워크를 기업 환경에 도입할 때 발생하는 영향에 대해 항상 우려합니다. 기업은 그림 2에서 언급한 대로 보안, 신뢰성, 지원 및 기타 기준과 관련하여 엔터프라이즈 수준에서 활용할 수 있는 준비된 기술과 프레임워크를 원합니다. 현재 사용 가능한 AI/ML/DL 프레임워크, 툴 키트, 애플리케이션은 대부분 보안 기능이 구현되어 있지 않아서 그 용도가 고립된 실험이나 연구소 환경에서의 실행으로 제한됩니다. 또한, 대부분의 기업이 AI 실행을 위해 별도의 클러스터에 투자하는 방법을 선택하는 데 이는 비용이 많이 들고 비효율적입니다. DIY 구축 시스템의 또 다른 문제는 다수의 벤더로부터 엔터프라이즈급 지원을 받기가 어렵다는 점입니다.

AI/ML/DL 워크로드를 위해 IBM의 솔루션을 고려할 경우

IBM의 전략은 더욱 액세스하기 쉽고 성능 기준에 적합한 AI/ML/DL을 구현하는 것입니다. IBM PowerAI, IBM Spectrum Conductor Deep Learning Impact 및 IBM Spectrum Scale 소프트웨어를 IBM Power Systems 및 IBM Elastic Storage Server(ESS)와 결합하면, 조직은 AI/ML/DL 워크로드를 위해 최적화되고 완벽하게 지원되는 고성능 플랫폼을 신속하게 구축할 수 있을 것입니다. 이러한 접근법을 취하면 오픈 소스에서 솔루션을 함께 통합해야 하는 어려움이 줄어듭니다. IBM 솔루션의 모든 구성요소는 IBM이 공급하고 획득하고 전체적으로 뒷받침하고 지원합니다(레벨 1 ~ 3 지원 포함). 지원이 필요할 경우 모든 IBM 구성요소에 대해 단일 연락 지점에 문의하면 되고 모든 케이스는 IBM에서 담당하고 관리합니다. 또한, IBM은 모든 소프트웨어 패치와 시스템 업그레이드를 관리합니다. 이러한 솔루션을 사용하면, 조직은 개발 환경을 단순화하고 AI 모델 훈련에 필요한 시간을 단축할 수 있습니다. 그리고, 이 솔루션을 통해 기존 IT 인프라에 플랫폼을 통합하는 동시에, 생산적인 PoC를 수행할 수 있으며, 다양한 환경과 프레임워크를 운영 중인 여러 데이터 사이언티스트가 조직에 맞게 확장되는 공통의 리소스 세트를 원활하게 공유할 수 있는 멀티테넌트 시스템으로 확장할 수 있습니다.

IBM은 데이터 사이언티스트 등 AI 구축 환경의 모든 중요 인력을 위한 툴과 소프트웨어로 구성된 가장

포괄적인 솔루션 스택 중 하나를 보유하고 있습니다. 경쟁사와 달리, IBM은 다양한 엔터프라이즈 고객을 보유하고 있으며 이러한 고객층은 점점 더 두터워지고 있습니다. 고객 중에는 시스템의 보안 및 데이터 보호 기능을 전적으로 신뢰하고 AI 솔루션을 배포하여 프로덕션 환경에서 실행하는 금융 기관도 포함됩니다.

그림 4에 제시된 IBM 레퍼런스 아키텍처는 다음과 같은 소프트웨어 구성요소 및 인프라 스택으로 구성되어 있습니다.

» **IBM PowerAI**는 TensorFlow, Caffe 및 관련 라이브러리와 같은 모델 훈련을 위한 주요 오픈소스 DL 프레임워크를 포함하는 소프트웨어 배포 패키지입니다. 이 프레임워크들은 IBM Power Systems 서버에 있는 CPU와 GPU 사이의 NVLink 인터커넥트를 활용하도록 최적화되어 뛰어난 처리량과 성능을 제공합니다.

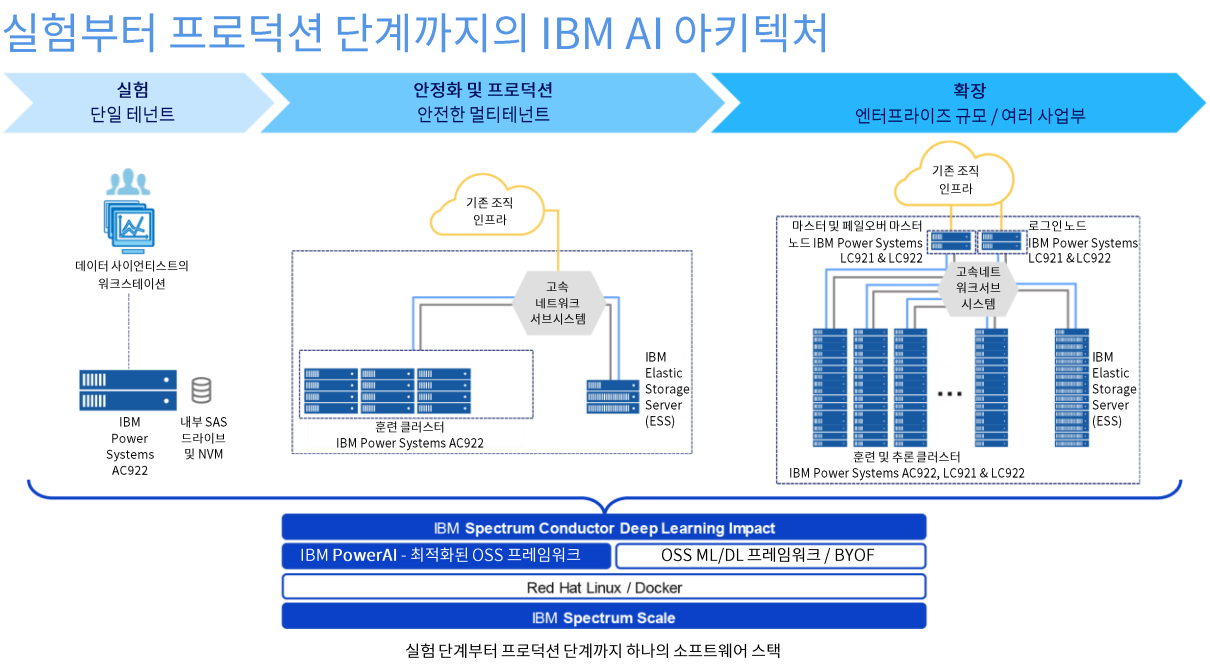
» **IBM Spectrum Conductor**는 Spark, Anaconda, TensorFlow, Caffe, MongoDB 및 Cassandra와 같은 최신 컴퓨팅 프레임워크와 서비스를 배포하고 관리하기 위해 공유된 엔터프라이즈급 환경을 구축하도록 설계된 고가용성 멀티테넌트 애플리케이션입니다. 데이터 변환과 준비는 수작업이 많이 필요한 단계들로 구성되어 있습니다. 즉, 데이터 소스를 찾은 후 연결하고, 스테이징 서버로 추출하고, 툴 및 스크립트를 사용하여 데이터를 조작하는 단계(예: 관련 없는 요소 제거, 큰 이미지를 GPU 메모리에 맞게 "타일" 크기로 나누기)가 필요합니다. 분산된 Apache Spark 및 애플리케이션 환경을 제공하는 IBM Spectrum Conductor는 프로세스를 자동화하고 속도를 향상하는 동시에 이러한 작업을 자동화하고 실행하는 데 도움을 줍니다. 스토리지 리소스에 연결하고 이러한 연결을 유지하며 데이터 형식 지정 정보를 포착하여 엔드 투 엔드 DL 프로세스 전반에서 빠른 속도로 반복이 가능하도록 합니다. Spectrum Conductor는 또한 중앙집중식 관리 및 모니터링 기능을 제공하고 프레임워크와 애플리케이션을 컨텍스트 내에서 실행하기 위해 엔드 투 엔드 보안을 구현하여 안전한 AI/ML/DL 환경을 조성합니다. 보안을 위해 사용되는 주요 기법은 다음과 같습니다.

  - **인증(authentication):** HDFS용 Kerberos 인증과 같이 Kerberos, SiteMinder, AD/LDAP 및 OS 인증이 제공됩니다.
  - **권한 부여(authorization):** 이 기능으로 미세한 액세스 제어, ACL/역할 기반 제어(role-based control, RBAC), Spark 바이너리 수명 주기 관리, 노트북 업데이트, 배포, 리소스 계획, 보고, 모니터링, 로그 검색, 실행이 가능합니다.
  - **위장(impersonation):** 다양한 테넌트가 프로덕션 실행 사용자를 정의할 수 있습니다.
  - **암호화(encryption):** SSL 및 모든 데몬 간의 인증을 지원합니다.

» **IBM Spectrum Conductor Deep Learning Impact**는 데이터 사이언티스트가 모델을 훈련, 조정 및 프로덕션 환경에 배포하는 데 집중할 수 있도록 해주는 엔드 투 엔드 워크플로우를 가진 DL 환경을 조성합니다. AI/ML/DL 프로세스의 구축/훈련 단계는 컴퓨팅 집약적이고 매우 반복적인 데다 모델 조정 및 최적화에 대한 전문성은 늘 부족합니다. 이 단계에서 조직의 DL 및 데이터 사이언스 부문 기술 격차가 가장 심합니다. PowerAI 및 Spectrum Conductor Deep Learning Impact는 하이퍼 파라미터를 제안하고 최적화하는 코그너티브 알고리즘을 사용하여 모델 선택과 제작을 지원합니다. 또한, 탄력적인 훈련을 지원하므로 실행 시간 도중 리소스를 유연하게 할당할 수 있습니다. 이 덕분에 리소스를 동적으로 공유하고 특정 작업에 우선 순위를 지정할 수 있으며 장애 발생 시 레질리언스를 발휘할 수 있습니다. 런타임 훈련 시각화를 통해 데이터 사이언티스트는 모델의 진행 상황을 확인하고 모델이 올바른 결과를 내지 못할 경우 훈련을 중단할 수 있으므로 보다 정확한 신경 모델을 더 빨리 제공할 수 있습니다.

» **IBM Spectrum Scale**은 레질리언스, 확장성, 제어, 스토리지 암호화를 통한 보안을 제공하는 엔터프라이즈급 병렬 파일 시스템입니다. Spectrum Scale은 요구사항이 많은 데이터 분석, 콘텐츠 저장소 및 기술적 컴퓨팅 워크로드를 처리하기 위해 확장 가능한 용량과 성능을 제공합니다. 스토리지 관리자는 플래시, 디스크, 클라우드, 테이프 스토리지를 성능이 더 뛰어난 통합 시스템으로 결합함으로써 기존 방법에 비해 비용을 더 낮출 수 있습니다.

그림 4: *IBM 레퍼런스 아키텍처*



출처: IBM

» 이 솔루션을 위한 **인프라**는 다음 요소로 구성됩니다.

- 컴퓨팅: IBM Power Server는 CPU:GPU NVLink 연결을 지원하므로 x86 서버보다 더 높은 I/O 대역폭을 제공합니다. 또한, 대용량 시스템 메모리도 지원할 수 있습니다.
  - » IBM Power System AC922는 NVLink를 사용하는 2~6개의 NVIDIA Tesla V100 GPU를 지원하므로 공랭식의 경우 초당 100GB, 수랭식의 경우 초당 150GB의 CPU:GPU 대역폭 속도를 제공합니다. 이 시스템은 최대 2TB의 총 메모리를 지원합니다.
  - » IBM Power System S822LC for HPC는 NVLink GPU와 2~4개의 NVIDIA Tesla P100 GPU를 지원하므로 초당 64GB의 CPU:GPU 대역폭 속도를 제공합니다. 이 시스템은 최대 1TB의 총 메모리를 지원합니다.
- 스토리지: IBM ESS는 IBM Spectrum Scale 소프트웨어와 IBM POWER8 프로세스 기반 I/O 집약적 서버 및 이중 포트 스토리지 엔클로저를 결합합니다. IBM Spectrum Scale은 이러한 IBM ESS의 핵심 요소인 병렬 파일 시스템입니다. IBM Spectrum Scale은 단일 네임스페이스를 제공하는 한편, 확장하면서 시스템 처리량을 늘립니다.

과제와 기회

IDC에 따르면 AI 모델 및 워크로드 배포에 퍼블릭 클라우드가 가장 많이 사용되고 그 다음으로 프라이빗 클라우드가 많이 사용되는 것으로 나타났습니다. AI 파이프라인을 퍼블릭 클라우드와 온프레미스 중 어디에서 실행할지 결정을 내릴 때는 일반적으로 데이터 중력(data gravity), 즉 데이터가 현재 있는 위치 또는 저장될 가능성이 높은 위치를 가장 많이 고려합니다. 컴퓨팅 리소스 및 애플리케이션에 대한 쉬운 액세스 그리고 이러한 기능을 살펴보고 사용할 때 구현할 수 있는 속도도 중요한 요인으로 작용합니다. 퍼블릭 클라우드 서비스에서는 데이터 사이언티스트와 IT 전문가에게, AI 모델을 훈련하거나 새로운 알고리즘으로 실험하거나 쉽고 민첩하게 새로운 기술과 방법을 학습할 수 있는 인프라와 툴을 제공합니다. 장기적으로는, 조직들이 IP와 인사이트의 외부 유출을 막고 보호하기 위해 모델 훈련을 온프레미스에서 실시할 가능성이 큽니다.

엣지 구축은 초기 단계에 있습니다. 엣지에는 리소스가 부족하지만 프로세싱이 엣지에서 이루어져야 하는 상황도 있습니다. 예를 들면, 공장 작업 현장에 있는 중요한 기계의 온도 센서가 엣지 인프라에 곧 장애가 발생할 수 있음을 알리는 측정 결과를 보내면 측정 결과가 신속하게 분석되고 기술자가 제때에 기계를 수리하도록 파견되어 비용이 많이 드는 운영 중단을 피할 수 있습니다.

IBM의 AI용 레퍼런스 아키텍처는 최적화된 소프트웨어 및 하드웨어 스택이라는 것이 IDC의 견해입니다. 이 레퍼런스 아키텍처는 테스트를 거치고 지원이 제공되며 즉시 사용 가능한 오픈소스 프레임워크, 높은 처리량의 GPU, 그리고 지능적이고 확장 가능하며 안전하고 메타데이터가 풍부하며 클라우드와 통합된 효율적인 멀티프로토콜 고성능 스토리지로 구성되어 있습니다. 이미 설명한 대로, 이 솔루션은 AI 배포 시 수반되는 복잡성을 줄여주며, 조직이 생산성과 효율성을 향상하고 획득 및 지원 비용을 낮추고 AI 채택을 가속하도록 지원합니다. 다음은 이 솔루션의 개선을 위해 활용 가능한 방법입니다.

- 계속 확장되는 하이브리드 멀티클라우드 구축 환경에서 IBM Cloud를 비롯한 Amazon Web Services, Google Cloud Platform 및 Microsoft Azure와 같은 다수의 퍼블릭 클라우드 서비스와의 원활한 통합.
- 엣지에서의 추론을 위해 임시 데이터를 보관할 수 있는 소형 폼 팩터의 엣지 인프라 지원.
- 이종의 프로세싱 아키텍처(예: GPU, FPGA, ASIC 또는 Manycore 프로세서)를 지원함으로써, AI 배포 환경을 위한 성능, 운영 환경, 필요한 기술, 비용, 에너지 요구사항을 기준으로 적합한 가속 기술을 선택할 수 있는 유연성 제공.

결론

오늘날 전 세계의 기업들이 우수한 비즈니스 성과를 내기 위해서는 AI/ML/DL 알고리즘이 적용된 애플리케이션을 실행하는 것이 매우 중요해졌으며 대다수의 DX 관련 활동과 활용 사례에도 이러한 작업은 꼭 필요한 것이 되었습니다. AI가 주도하는 비즈니스 성과를 더 빨리 얻고 배포 시 수반되는 장애물을 극복하려는 조직을 위해 IDC가 제시하는 지침은 다음과 같습니다.

- 비즈니스 성과에 집중하고 프로젝트 타임라인을 올바르게 정의하며 수익과 비용에 즉각적으로 영향을 주는 프로젝트에 우선 순위를 부여하십시오.
- 비즈니스 성과를 향상하기 위해 데이터 준비를 간소화 및 자동화하고, AI 모델 구축, 훈련 및 배포의 반복을 가속할 수 있는 소프트웨어 툴을 찾으십시오.

» 우수한 성능으로 훈련과 추론을 실시할 수 있도록 높은 용량, 높은 처리량 및 낮은 레이턴시를 지원할 수 있는, 동적으로 조정 가능하고 간편하며 유연하고 안전하고 비용 효율적이며 탄력적인 인프라를 찾으십시오.

» 지능형 인프라를 도입하고 활용하여 예측적 분석을 통해 가치 있는 인사이트를 얻은 다음, 데이터의 신뢰성과 품질이 확인되면 천천히 단계적으로 작업의 자동화를 진행하십시오.

애널리스트 소개:

Ritu Jyoti, *부사장*

Ritu Jyoti는 IDC의 엔터프라이즈 스토리지, 서버, 인프라 소프트웨어 팀의 시스템 인프라 프로그램 부문 부사장입니다. 시스템 인프라 프로그램은 연구 제안 및 분기별 트래커 보고서와 더불어 자문 서비스 및 컨설팅 프로그램 등을 제공합니다. Ritu Jyoti 부사장의 주요 연구 분야로는 코그너티브/인공지능, 빅데이터, 분석 워크로드 서비스 등이 있으며, 이러한 연구를 통해 새로운 머신러닝, Hadoop, NoSQL 데이터베이스, 분석 기술이 인프라 소프트웨어 및 하드웨어 시장 그리고 디지털 혁신, 즉 IT 혁신 데이터 인프라 전략에 미치는 영향을 살펴봅니다.

IDC Custom Solutions

IDC Corporate USA

5 Speen Street
Framingham, MA
01701, USA
T 508.872.8200
F 508.935.4015
Twitter @IDC
idc-insights-community.com
www.idc.com

본 간행물은 IDC Custom Solutions에서 제작했습니다. 특정 업체의 후원 여부가 명시되지 않은 한, 본 간행물에 제시된 의견, 분석, 연구 결과는 IDC가 독자적으로 수행하고 발표한 보다 세부적인 연구와 분석을 토대로 합니다. IDC Custom Solutions는 IDC 콘텐츠를 여러 기업에서 배포할 수 있도록 다양한 형태로 제공합니다. IDC 콘텐츠를 배포할 수 있는 라이센스를 부여했다고 해서 해당 라이센스를 받은 기업에 대한 옹호나 의견으로 간주할 수 없습니다..

IDC 정보 및 데이터의 외부 공개 - 광고, 보도 자료 또는 홍보 자료에 IDC 정보를 사용하기 위해서는 해당 IDC 부사장이나 지사장의 사전 서면 승인이 필요합니다. 사전 서면 승인 요청 시 제안 문서의 초안을 함께 제출해야 합니다. IDC는 임의의 사유로 외부 사용 승인을 거부할 수 있는 권한을 보유합니다..