백서

데이터 인프라로 AI 현대화 가속화

후원: IBM Corporation

Ashish Nadkarni Sriram Subramanian Matt Leib
2021 년 2 월

핵심 요약

인공 지능(Artificial Intelligence, AI), 머신 러닝(Machine Learning, ML), 그리고 종종 딥 러닝(Deep Learning, DL) 역량은 이제 디지털 혁신(Digital Transformation, DX) 이니셔티브의 필수적인 구성 요소가 되었습니다. AI 로 달성할 수 있는 비즈니스 기회는 무궁무진합니다. 경쟁사가 고객층을 확장하고 고객을 만족시키기 위해 이전에 활용하지 못한 인사이트와 역량을 풍부하게 획득할 수 있으므로 기업들은 AI 를 활용하지 않으면 비즈니스에 재앙을 초래할 수 있다는 점을 전보다 더 많이 인식하고 있습니다. 현재 "AI 는 우리 회사와 맞지 않다" 또는 "AI 에 대한 기대는 대부분 과장되었다"라고 생각하는 기업이 있다면 그 수는 몇 안 될 것입니다. 오히려 본격적인 AI 이니셔티브가 업종과 기업 규모에 관계 없이 전 세계에서 진행 되고 있습니다.

기업들은 AI 주도 혁신 및 현대화 이니셔티브를 추구하고 있습니다. 이러한 이니셔티브를 통해 AI 를 활용한 실험적인 도전에서 더 나아가 AI 투자로부터 비즈니스 가치를 창출할 수 있어야 합니다. AI 와 관련된 디지털 혁신에 대한 투자의 성공 여부는 규모에 맞게 AI 솔루션을 개발, 구축, 관리하는 데 필요한 광범위한 전문 지식을 갖추었는지에 따라 직접적으로 결정됩니다. 많은 조직의 사업부(Line of Business, LOB), IT 직원, 데이터 과학자, 개발자가 AI 에 대해 알아보고, 적용 사례를 이해하고, 비즈니스를 위한 AI 전략을 정의하고, 초기 AI 이니셔티브를 실행하고, 머신 러닝 알고리즘, 특히 딥 러닝을 사용하여 새로운 인사이트와 역량을 제공하는 AI 애플리케이션을 개발하고 테스트해왔습니다.

이러한 이니셔티브를 확장하는 과정에서 새로운 질문이 생깁니다. 이들은 표준 범용 컴퓨팅과 기존 또는 레거시 스토리지 인프라를 사용해서는 안 된다는 것을 알고 있습니다(실제로 이래서는 안 된다는 것을 직접 경험했을 수 있습니다). 또한, 이들은 AI 훈련(AI 모델의 훈련)과 AI 추론(훈련된 모델을 사용하여 이벤트를 이해하거나 예측)을 위해 다양한 유형의 확장 가능한 컴퓨팅, 그리고 이와 마찬가지로 확장 가능한 스토리지 인프라가 필요하다는 것도 알고 있습니다.

IDC 는 기존 인프라의 일부 또는 전부를 현대화하지 않고 사용하려는 조직은 실패할 가능성이 더 높다는 것을 알게 되었습니다. 현대화를 진행하는 경우에도 회사마다 중점 분야가 다릅니다. 기업들은 AI 기술에서 컴퓨팅의 역할을 잘 이해하고 있지만 스토리지의 가치는 과소평가하는 경우가 많습니다. 게다가, AI 애플리케이션 그리고 특히 기하급수적으로 많은 데이터를 구문 분석하는 딥 러닝 시스템은 굉장히 높은 사양을 요구합니다. 이러한 애플리케이션과 시스템은 매우 많은 컴퓨팅 코어를 기반으로 강력한 병렬 처리 기능을 요구합니다. 일반적인 스토리지 시스템은 이러한 AI 작업의 실행을 충분히 지원할 수 없습니다. 마지막으로, AI 관련 이니셔티브를 진행할 때 Kubernetes 또는 컨테이너를 포함하는 애플리케이션 현대화 노력과 하이브리드 클라우드 아키텍처를 활용하여 하나 이상의 퍼블릭 클라우드 서비스를 통합하는 것을 고려해야 합니다.

IDC 는 기존 인프라의 일부 또는 전부를 현대화하지 않고 사용하려는 조직은 실패할 가능성이 더 높다는 것을 알게 되었습니다. 현대화를 진행하는 경우에도 회사마다 중점 분야가 다릅니다. 기업들은 AI 기술에서 컴퓨팅의 역할을 잘 이해하고 있지만 스토리지의 가치는

IDC 연구에 따르면, 스토리지 인프라 측면에서 세부 사항에 충분한 주의를 기울이지 않는 경우, 그 즉시 AI 혁신의 진행을 방해할 수 있는 것으로 나타났습니다. 이 문제를 극복하려면 기존에 진행한 인프라의 실험적 시도를 통해 이를 프로덕션 단계로 확장할 준비가 된 기업들은 인프라를 정비하여 이 이니셔티브에 요구되는 병렬 처리 성능을 구현해야 합니다. 그리고 이 과정에서 대용량 확장을 위해 스케일아웃할 수 있고 글로벌 배포와 데이터 액세스를 위해 클라우드, 컨테이너, 성능 집약적 컴퓨팅에 통합된 현대화된 스토리지 솔루션에 투자해야 합니다. 이러한 과제에 적합한 IBM Spectrum Scale 및 IBM Elastic Storage System(ESS) 같은 솔루션은 AI 정보 아키텍처의 필수 구성 요소를 제공합니다. 이러한 솔루션은 AI 워크로드, 컨테이너화된 배포 그리고 AI 워크로드에 특히 중점을 두는 하이브리드 클라우드 배포에 적합합니다.

상황 개요

이미 도래한 AI/ML 의 세상

전 세계의 기업들은 디지털 혁신 이니셔티브를 촉진하기 위해 AI 에 투자함으로써 얻을 수 있는 새로운 기회에 열렬한 반응을 보이고 있습니다. 인공 지능은 자연어 처리(Natural Language Processing, NLP), 이미지/비디오 분석, 머신 러닝, 지식 그래프 그리고 질문에 응답하거나 인사이트를 발견하고 권장 사항을 제공하는 기타 기술을 활용하는 기술을 말합니다. 이러한 시스템은 활용 가능한 증거를 기반으로 가설을 세우고 가능한 답변을 만들며, 대량의 컨텐츠를 수집하여 훈련할 수 있고, 재훈련 또는 인간의 개입을 통해 실수와 실패를 기반으로 조정하고 학습할 수 있습니다. IDC 는 2022 년까지 이러한 AI 중심의 적용 사례 중 최소 60%가 Global 2000 에 속한 조직의 최소 65%에 배포될 것이라고 예측합니다. 이는 2019 년을 기준으로 34%의 증가에 해당됩니다.

IDC 는 2022 년까지 이러한 AI 중심의 적용 사례 중 최소 60%가 Global 2000 에 속한 조직의 최소 65%에 배포될 것이라고 예측합니다. 이는 2019 년을 기준으로

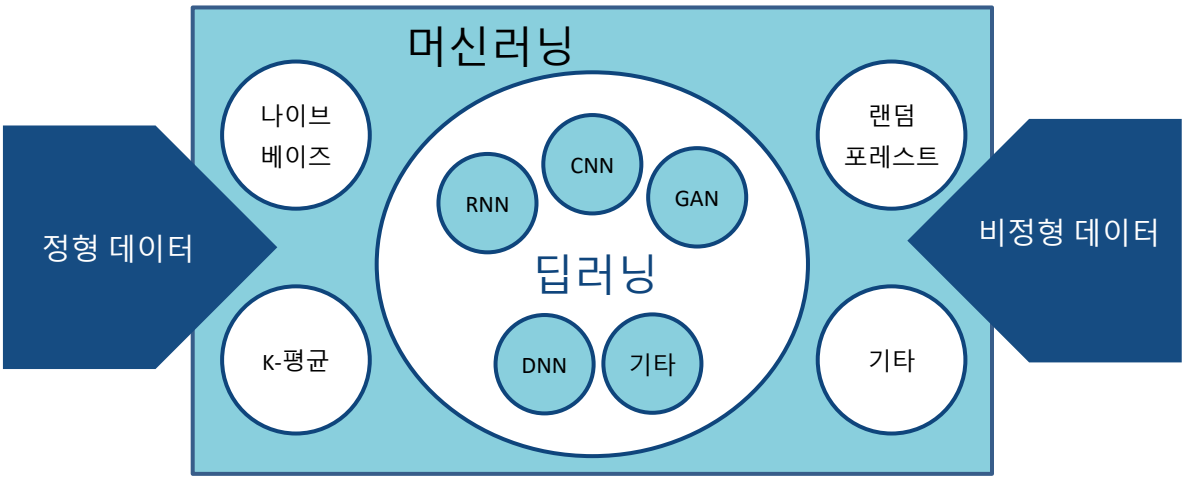
AI 는 매우 빠르게 비즈니스 전반에서 프로세스 및 워크플로우 자동화 분야에서 만연해지고 있습니다. 2019 년, IDC 는 고객 경험(Customer Experience, CX), 법무 및 기업 전략, 시설, 조달을 포함하는 8 개의 사업부 기능 영역에서 176 개의 디지털 혁신 적용 사례를 살펴보았으며, 이러한 적용 사례의 약 26%가 AI 에 의존하고 대다수의 조직에서 배포되어 있는 것으로 추정했습니다.

이는 곧 대다수의 선도 조직이 운영을 확장하고 비정형 데이터를 이해하고 지능적 비즈니스 인사이트를 제공하기 위해 조직 전반에서 자연어 처리, 머신 러닝, 딥 러닝, 스피치-텍스트 전환 등의 AI 기술을 활용할 것이라는 의미입니다. 한편, 개념 증명(Proof of Concept, POC)부터 프로덕션 단계까지 AI 기반 적용 사례를 발전시키는 방법을 아직 파악하지 못한 조직은 더욱 뒤쳐져 디지털 격차가 더 벌어질 것입니다.

머신 러닝은 컴퓨터 시스템을 통해 인간이 프로그래밍할 필요가 없이 주어진 작업을 위해 행동을 학습하고 개선하도록 지원하는 AI 기술의 한 종류입니다. 머신 러닝 모델은 작업(예: 사람의 얼굴을 인식하는 작업)을 "학습"한 것으로 여겨질 때까지 정형 및/또는 비정형 데이터를 대량으로 반복적으로 스스로 테스트하여 시간이 흐름에 따라 향상될 수 있는 알고리즘을 의미합니다. 그림 1 은 딥 러닝이 ML 의 한 종류임을 보여줍니다. 일반적인 DL 아키텍처로는 심층 신경망(Deep Neural Network, DNN), 콘볼루션 신경망(Convolutional Neural Network, CNN), 순환 신경망(Recurrent Neural Network, RNN), 생성적 적대 신경망(Generative Adversarial Network, GAN) 등이 있습니다.

그림 1

머신 러닝 및 딥 러닝 애플리케이션



출처: IDC, 2021 년

AI 소프트웨어 플랫폼은 디지털 어시스턴트와 같은 대화형 AI, 데이터에 숨겨진 관계를 파악하고 예측하기 위한 예측 분석, 테스트에서 가치를 인식하고 이해하고 추출하기 위한 텍스트 분석 및 자연어, 오디오, 음성 및 스피치에서 정보를 인식, 파악, 추출하기 위한 음성/스피치 분석,이미지 및 비디오에서 정보를 인식, 파악, 추출하기 위한 이미지 및 비디오 분석(예: 패턴 인식, 물체 및 색상 그리고 사람, 얼굴, 감정, 자동차, 풍경 등의 기타 속성 인식) 기능을 포함합니다.

많은 기업이 AI 이니셔티브를 상당히 진행했으며 AI 를 프로덕션 규모로 배포할 준비가 되었습니다. 다른 기업은 여전히 AI 로 실험적 시도를 진행 중입니다. 그리고 나머지 기업은 AI 애플리케이션이 조직에 가져다 줄 수 있는 이점을 여전히 평가하는 단계에 머물러 있습니다.

첫 번째 그룹(배포 준비 완료)의 경우, IDC 는 기업, 정부 그리고 기타 조직이 실행하기 시작한 다양한 AI 적용 사례를 확인했습니다. 기업이 적용 사례에 사용하는 하드웨어, 소프트웨어, 서비스의 양을 기준으로 순위를 매겼을 때 가장 일반적인 적용 사례 다섯 가지는 다음과 같습니다.

- **자동화된 고객 서비스 에이전트.** 예를 들면, 은행 산업에서 AI 애플리케이션이 고객의 요구 사항과 문제를 이해하고 은행이 고객의 문제를 해결하는 데 필요한 시간과 리소스를 줄이도록 지원하는 학습 프로그램을 통해 고객 서비스를 제공합니다. 이러한 에이전트는 여러 산업 분야에서 점점 더 널리 사용되고 있습니다.
- **영업 프로세스 권장 사항 및 자동화.** 다양한 산업에서 사용되는 이러한 AI 애플리케이션은 고객 관계 관리(Customer Relationship Management, CRM) 시스템과 작동하여 고객의 상황을 실시간으로 이해하고 영업 담당자에게 적절한 조치를 권장합니다.
- **자동화된 위협 인텔리전스 및 예방 시스템.** 여러 정부와 산업 전반에서 위협 예방 활동의 중요한 부분이 되고 있는 이러한 AI 애플리케이션은 인텔리전스 보고서를 처리하고 이러한 보고서에서 정보를 추출하고 다양한 정보 간의 관계를 파악한 후 데이터베이스, 시스템, 웹사이트 등에 대한 위협을 파악합니다.

- **사기 분석 및 조사.** 보험 산업 뿐만 아니라 다른 산업에서도 널리 사용되는 이러한 AI 애플리케이션은 규칙 기반 학습을 통해 사기 거래를 찾아내며, 자동 학습을 통해 다양한 보험 사기 책략을 찾아냅니다.
- **자동화된 예방적 유지보수.** 제조 산업에서 사용되는 이러한 AI 애플리케이션은 공장 및 기계 장애의 가능성을 정확하게 예측하는 모델을 구축하여 다운타임과 유지보수 비용을 줄여주는 머신 러닝 알고리즘을 기반으로 합니다.

엔터프라이즈 분야에서 인기를 얻은 다른 AI 적용 사례는 다음과 같습니다(하드웨어, 소프트웨어, 서비스에 대한 지출 규모 순으로 나열됨).

- 프로그램 어드바이저 및 권장 사항 제공 시스템
- 진단 및 치료 시스템
- 지능적 처리 자동화
- 품질 관리 조사 및 권장 사항 제공 시스템
- IT 자동화 및 엔터프라이즈 정보 취급자를 위한 디지털 어시스턴트
- 전문 쇼핑 어드바이저 및 제품 추천 시스템
- 공급 및 물류, 규제 인텔리전스
- 자산/운송 관리 및 자동화된 청구 프로세스
- 디지털 트윈/첨단 디지털 시뮬레이션
- 공공 안전 및 응급 대응
- 적응형 학습
- 스마트 네트워킹
- 화물, 자산 및 운송 관리
- 제약 연구 및 신약 발견

AI 이니셔티브를 위해 해결해야 할 과제

IDC 연구에 따르면 기업들은 데이터 관리의 관점에서 AI 이니셔티브와 관련하여 다음과 같은 과제를 해결해야 한다고 응답하는 경우가 많은 것으로 나타났습니다.

- 데이터 수집 및 준비 주기를 진행하는 데 시간이 너무 많이 소요됨
- 다양한 분석 적용 사례를 지원하여 발생하는 인프라 사일로
- 단일 데이터 소스가 없이 동일한 데이터의 복제본을 여러 개 사용
- 반복 가능성을 구현하기 위해 데이터 출처를 안전하게 관리하고 보호해야 함
- 글로벌 액세스 가능성(하이브리드 클라우드) 및 협업이 필요함
- 데이터가 캡처되고 큐레이션된 후 데이터 무결성 달성

AI 혁신을 위해서는 견고한 데이터의 토대가 필요함

ML 및 DL 이니셔티브는 입력되는 다양한 정형 및 비정형 데이터에 크게 의존합니다(그림 1 다시 참조). AI 프로젝트는 데이터 수집, 구성, 살균, 검증, 모델 구축, 훈련, 테스트, 추론, 데이터 폐기 등 여러 단계로 구성됩니다. 각 단계에서 빠른 액세스가 가능한 낮은 레이턴시 스토리지부터 저비용 보관 스토리지까지 다양한 요구 사항을 충족해야 하므로 기업은 이러한 단계를 위해 적절하고 비용 효율적인 스토리지 인프라를 선택해야 합니다. 또한, 라이프사이클에 걸쳐 AI 데이터 세트를 관리하기 위해 개별적 툴 세트를 사용해야 합니다.

5G 가 보급되고 IoT 기반 센서가 확산되면서 엣지 디바이스에서 더 많은 데이터가 생성 및 소비되고 있습니다. 카메라와 지능적 어시스턴트가 엣지 디바이스에 배포되어 더 빠르게 대응하고 더 나은 사용자 경험을 제공할 수 있습니다. 네트워크 엣지디바이스 또는 엔드포인트에서 처리되는 로컬 AI 에 대한 필요성이 증가하고 있습니다. 연결성이 제한된 엣지 디바이스에서 실행되는 레이턴시에 민감한 AI 애플리케이션에는 큰 확장성이 요구될 것입니다. AI 기반 엣지 컴퓨팅 적용 사례가 채택될 것입니다.

AI 혁신을 가속화하는 AI 데이터 인프라

많은 조직이 AI 에 투자하면서 여러 가지 인프라 설계 및 배포 요인이 매우 중요해지고 있습니다. 하이퍼스케일 클라우드 서비스 제공업체의 선례에 따라 조직들은 통합된 데이터 인프라를 구축하여 인프라 요구 사항을

충족하려고 하고 있습니다. 데이터 인프라는 효율성이 뛰어나고 확장 가능한 컴퓨팅 및 스토리지 계층을 포함하는, AI 이니셔티브를 위한 공통의 토대입니다. AI 워크로드에 균일하게 접근하는 대신, 데이터 인프라는 이러한 워크로드를 복합적인 방식으로 처리하며, 정형 및 비정형 데이터가 혼합된 데이터 세트에 따라 워크로드의 일부를 적절한 스토리지 계층 기반의 적합한 컴퓨팅 계층과 연결합니다. 이러한 복합적 접근법은 3 가지 차원에서 AI 혁신을 가속화합니다.

- **규모.** 규모라는 차원은 워크로드가 운영되는 규모를 의미합니다. 컴퓨팅, 네트워킹, 데이터 지속성(스토리지)과 같은 토대가 되는 하위 차원은 모두 하드웨어와 관련이 있습니다. 중요한 것은 오케스트레이션과 같은 소프트웨어 관련 하위 차원이 스택의 규모와 복잡성이 증가함에 따라 균형을 유지하기 위해 동등하게 중요해지고 있다는 점입니다.
- **이식성.** 이식성은 워크로드를 중앙 데이터센터, 엣지 디바이스, 엔드포인트 배포 환경 전반에서 이동할 수 있는 능력을 말합니다. 현재 이러한 워크로드 중 다수는 그 속성이 정적입니다(즉, 단일 배포 환경에서 실행되도록 설계됨). 점점 더 많은 기업이 하나의 배포 환경(예: 퍼블릭 클라우드)에서 워크로드를 개발하고 이를 (프로덕션 단계의) 다른 배포 환경(예: 엣지)에서 설치하려고 하고 있습니다. 이는 현재의 모바일 앱 개발 및 배포 모델과 유사합니다.
- **시간.** 이 차원은 워크로드 자체의 시간적 연속성과 관련이 있습니다. 많은 AI 워크로드가 고성능 컴퓨팅 또는 빅데이터 및 분석 배포 환경의 설계를 차용하며, 기본적으로 일괄처리되도록 설계됩니다. 고성능 액셀러레이터의 확산 덕분에 점점 더 많은 AI 워크로드가 실시간 또는 거의 실시간으로 스트리밍 데이터를 분석할 수 있습니다.

AI 데이터 인프라에 대한 오해와 필수 요소

간과되는 경우가 많지만 필수적인 토대는 스토리지 인프라입니다. 대규모로 AI 를 배포하는 경우 용량(확장) 및 성능(IOPS 및 대역폭) 측면에서 스토리지 인프라에 더 큰 부담이 발생합니다. 많은 경우 조직들은 다른 워크로드에 사용되는 내부 서버 기반의 스토리지 또는 엔터프라이즈 스토리지가 AI 애플리케이션을 실행하는 데 충분하다고 가정합니다. 그리고 인프라가 구축된 후 스토리지가 취약하다는 점을 깨닫습니다. 각 애플리케이션은 다른 요구 사항이 있으므로 IT 조직에 다른 과제를 부여합니다. 그러므로 IT 구매자와 공급업체는 "망치가 있으면 모든 것이 못으로 보인다"라는 속담과 같이 행동하는 것을 피해야 합니다.

더 신중하게 접근하려면 데이터 인프라를 포괄적으로 고려해야 합니다. 데이터 지속성 확장 및 액세스 메커니즘은 최소한의 요구 사항이며, 조직들은 시야를 넓혀 컴퓨팅, 스토리지 소프트웨어, 시스템 계층 간의 네트워킹 및 통합을 포함해야 합니다. 기업은 단지 다른 스토리지 시스템을 "연결"하는 인프라를 생각할 것이 아니라 일관적이고 전체적으로 데이터 인프라를 생각하는 사고 방식으로 이동해야 합니다. IDC 는 데이터 인프라 요구 사항, 특히 스토리지 관련 요구 사항을 다음 섹션에서 논의할 몇 개의 핵심 영역으로 나눌 수 있다고 생각합니다.

컴퓨팅 통합

모든 AI 워크로드는 컨테이너화된다고 가정하는 경우가 많습니다. 이는 잘못된 가정입니다. 오히려, 많은 AI 워크로드가 베어 메탈에서 실행되거나 심지어 가상화됩니다. 특히, AI 기반 애플리케이션은 베어 메탈 또는 가상화된 컴퓨팅을 기반으로 실행되는 경우가 많습니다. 많은 AI 워크로드가 액셀러레이터를 활용하도록 최적화되고 있습니다. 그러나 이것이 모든 AI 워크로드는 가속화된 컴퓨팅을 기반으로 가장 효과적으로 실행된다는 의미는 아닙니다. 가속화된 컴퓨팅은 일반적으로 워크로드에 다양한 문제를 야기합니다.

데이터 지속성 및 액세스

AI 워크로드를 위한 컴퓨팅 요구 사항이 다양하듯이, 데이터 지속성 요구 사항도 다양합니다. AI 워크로드 스택의 여러 측면 중 잘 드러나지 않고 사람들이 오해하는 측면이 데이터 지속성 계층입니다. 모든 AI 워크로드에는 대량의 고성능 스토리지가 필요하다고 가정하는 경우가 많습니다. 이러한 가정 역시 잘못된 가정입니다. 사실, 모든 AI 워크로드가 "대규모 데이터 세트"인 것은 아닙니다. 이러한 워크로드가 짧은 기간 동안 동시에 많은 소규모 데이터 세트를 샘플링하는 경우도 있습니다. 마찬가지로, 개방형 시스템 컴퓨팅

플랫폼에서 실행되는 베어 메탈 워크로드는 스케일아웃 블록 또는 파일 액세스를 사용하는 경우가 많습니다. 가상화된 워크로드가 하이퍼컨버지드 인프라(HyperConverged Infrastructure, HCI)에서 실행되는 경우도 드물지 않습니다.

스토리지 인프라에 정형 데이터와 비정형 데이터가 함께 수집되는 경우 당연히 멀티프로토콜 액세스가 사용됩니다. 많은 IoT 및 엣지 디바이스는 SMB 또는 NFS 를 통해 통신하며 몇몇은 S3 를 사용합니다. 스트리밍 데이터 액세스가 필요한 경우도 있습니다. 그리고 경우에 따라 네이티브 병렬 파일 시스템 클라이언트도 사용될 수 있습니다.

확장/축소 및 계층화

AI 와 ML 애플리케이션을 지원하려면 스토리지 시스템이 규모에 맞는 성능을 제공해야 합니다. 비정형 데이터 저장소의 경우 네트워크 액세스와 함께 병렬 파일 시스템을 활용하는 것은 스토리지 시스템입니다. 정형 데이터의 경우 플래시 기반 스토리지 시스템이 사용됩니다. 확장/축소는 기본적으로 AI 및 ML 애플리케이션의 요구 사항을 충족하기 위해 성능과 용량이 서로 영향을 받지 않는 상태에서 각각을 확장 또는 축소하는 것을 의미합니다.

또한, 미래 지향적 인프라를 지원하려면 시스템은 오래되었거나 자주 사용되지 않는 데이터를 S3 와 같은 알려진 오브젝트 스토리지 인터페이스가 있는 저비용 오브젝트 스토리지로 보내 간단하고 효율적으로 처리할 수 있어야 합니다.

소프트웨어 정의 스토리지

AI 와 ML 은 소프트웨어 정의 스토리지를 위한 촉매 역할을 수행합니다. AI 와 ML 은 하드웨어 위의 이질적 소프트웨어 제어 계층을 통해 IaC(Infrastructure as Code)와 자동화를 구현합니다. 이를 통해 AI/ML 과의 통합을 향상할 수 있으며, 그 결과 스토리지가 애플리케이션 요구 사항에 따라 원활하게 확장/축소될 수 있습니다.

배포 민첩성 및 유연성

적용 사례를 지원하는 애플리케이션은 조직이 개발한 맞춤형이거나 기성 AI 소프트웨어를 기반으로 하여 AI SaaS 로 제공될 수 있습니다. 맞춤형으로 개발된 애플리케이션과 기성 소프트웨어 기반 애플리케이션의 배포 고려 사항은 온프레미스, IaaS 기반 클라우드 또는 하이브리드 클라우드인지 여부입니다. 하이브리드 클라우드에서는 온프레미스 환경이 공통 자동화 및 오케스트레이션 계층을 통해 퍼블릭 클라우드 환경과 상호작용합니다.

AI 의 분산된 속성을 고려할 때, 컴퓨팅이 이루어지는 곳으로 데이터를 이동하는 것보다 데이터가 공급되거나 생성되는 곳 가까이에서 컴퓨팅을 수행하는 것이 가장 적절하다고 가정하는 것이 안전합니다. 최근, 중앙 데이터센터-엣지 디바이스-엔드포인트 모델은 사실상 AI 를 설명하는 방법이 되었습니다(중앙 데이터센터에는 클라우드가 포함되며 엔드포인트에는 임베디드 인텔리전스가 포함됩니다). 각 위치의 워크로드 프로파일이 다르므로 기반 인프라의 요구 사항도 다르다는 점에 유의해야 합니다.

다양한 배포 시나리오를 충족하기 위해 다음과 같은 역량을 갖춘 솔루션을 고려해야 합니다.

- **초고성능 AI 모델을 훈련하는 데 필요한 대량의 데이터를 안전하게 처리.** 딥 러닝 훈련에 필요한 성능을 구현하려면 고대역폭 데이터 수집과 더불어 GPU 를 사용한 대량 병렬 처리를 실행할 수 있어야 합니다.
- **초고성능 AI 모델이 추론을 수행할 대량의 데이터를 안전하게 처리.** 추론과 관련된 성능을 구현하려면 훈련된 AI 모델을 통해 유입 데이터를 처리하고 거의 실시간으로 AI 의 인사이트 또는 결정을 제공할 수 있어야 합니다.

데이터 과학자와 개발자에게는 클라우드에서 AI 이니셔티브를 시작하는 것이 더 쉬운 경우도 있습니다. 온프레미스 컴퓨팅을 준비할 필요가 없기 때문입니다. 딥 러닝의 경우 일반적으로 온프레미스 컴퓨팅을 가속화해야 합니다. 가속화된 AI 클라우드 인스턴스는 대부분의 퍼블릭 클라우드에서 보통 오픈소스 AI 스택과 함께 제공됩니다. 물론 AI 훈련을 위한 가속화된 클라우드 인스턴스의 경우 클라우드 서비스 제공업체(Service Provider, SP)가 프로세서, 코프로세서, 상호 연결, 메모리 크기, I/O 대역폭 등의 측면에서 최종 사용자에게 제공되는 것을 결정합니다. 이러한 구성 요소를 최적의 조합으로 제공하지 못하는 클라우드 SP 도 있습니다. 이러한 조합은 궁극적으로 데이터 과학자가 훈련 모델을 개발하는 속도와 품질을 결정합니다. 그래서 많은 조직이 온프레미스 배포를 선택합니다.

IBM Spectrum Scale 은 고성능 컴퓨팅 산업에서 쌓은 수년의 경험을 기반으로 구축된 AI 인프라에 매우 견고한 토대를 제공합니다

지난 몇 년 동안 AI 로 실험적 시도를 하면서 많은 조직은 표준 인프라 또는 기본적인 클라우드 인스턴스 때문에 "벽에 부딪혔습니다". 모델 훈련에 시간이 너무 많이 소요되었고 추론 속도는 너무 느렸습니다. IDC 연구에 따르면 응답자의 77.1%가 온프레미스 AI 인프라 때문에 한 가지 이상의 한계에 부딪혔다고 응답했으며, 응답자의 90.3%는 클라우드에서 컴퓨팅의 한계에 부딪혔다고 응답했습니다.

IBM STORAGE 를 사용하는 확장 가능한 글로벌 AI 정보 아키텍처

IBM Storage for data and AI 솔루션은 고객이 하이브리드 클라우드 환경에서 프로덕션 규모로 AI 이니셔티브를 원활하게 도입하도록 지원합니다. IBM 은 고성능 AI 및 빅데이터 솔루션을 위한 확장 가능한 고성능 워크로드와 효율적이고, 안전하며, 확장 가능한 고용량 스토리지 분야를 계속 이끌고 있습니다. IBM 의 스토리지 포트폴리오는 엣지 디바이스, 중앙 데이터센터, 퍼블릭 클라우드의 스토리지 및 데이터 관리를 통합하므로 AI 현대화를 가속화할 수 있습니다. 이 포트폴리오는 Kubernetes 컨테이너와 Red Hat OpenShift 플랫폼에 대한 광범위한 지원과 통합을 제공하며, 퍼블릭 클라우드 또는 데이터센터 워크로드를 위해 배포하고 액세스할 수 있습니다. IBM Storage for data and AI 는 조직 전체의 규모에 맞게 배포할 수 있는 AI 정보 아키텍처와의 향상된 심층적 통합을 통해 복잡성과 비용을 줄이고자 합니다.

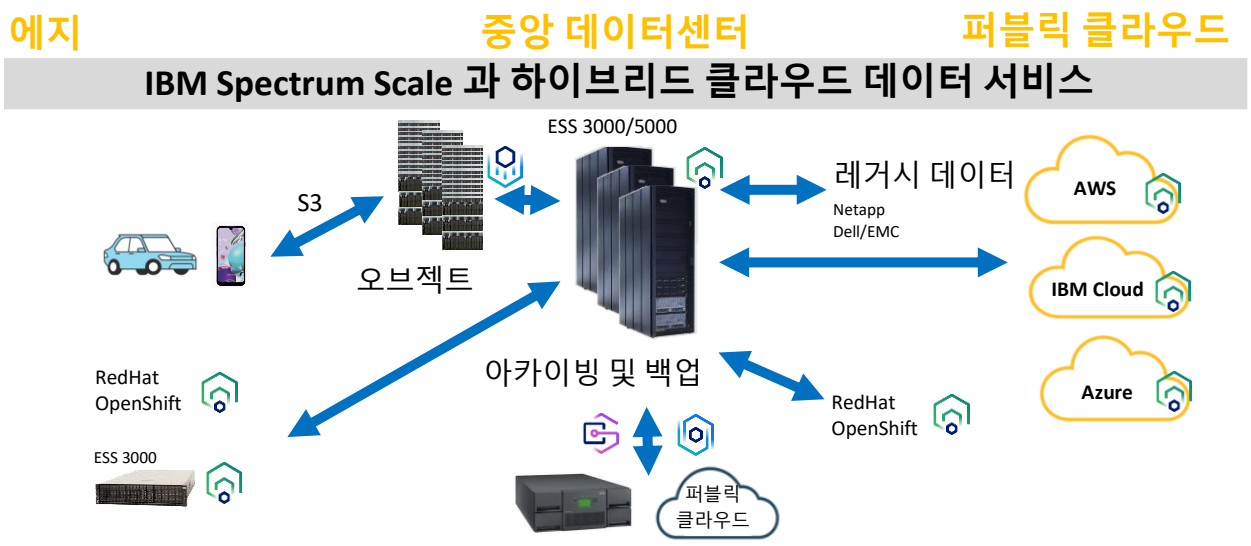
IBM Spectrum Scale

IBM Spectrum Scale 은 AI/ML, 모델링 및 시뮬레이션, 분석 등 강력한 성능을 요구하는 모든 워크로드를 위해 설계된 분산 컴퓨팅 아키텍처를 기반으로 구축되었습니다. 이 솔루션은 간편한 관리와 확장을 통해 운영 비용을 절감하고 정책 기반 데이터 최적화와 투명한 데이터 라이프사이클 관리를 통해 자본 비용을 절감하는 클러스터링된 병렬 파일 시스템입니다. IBM Spectrum Scale 은 단일 파일 시스템 또는 여러 노드의 파일 시스템 세트에 대한 동시 액세스를 제공합니다. 노드는 직접 연결 방식 또는 네트워크 연결 방식일 수 있으며, 이 두 가지 방식을 혼합할 수도 있습니다. 또는 무공유 클러스터 구성을 사용할 수도 있습니다. 이 스케일아웃 솔루션과 고가용성 플랫폼은 공통 데이터 세트에 고성능 공유 액세스를 지원합니다.

IBM Spectrum Scale 은 데이터 복제, 정책 기반 스토리지 관리, 멀티사이트 운영을 지원합니다. IT 운영 팀은 Kubernetes 컨테이너 노드, IBM AIX 노드, IBM Z 또는 LinuxONE 노드, Linux 노드, Microsoft Windows Server 노드 각각으로 구성된 클러스터를 생성하거나 이 다섯 가지 노드 모두로 구성된 클러스터를 생성할 수 있습니다. IBM Spectrum Scale 은 가상화 또는 컨테이너화 된 인스턴스에서 실행될 수 있으므로 여러 환경에서 공통 공유 데이터 액세스를 제공하며, 논리적 파티션 또는 다른 하이퍼바이저를 활용합니다. 여러 IBM Spectrum Scale 클러스터가 한 위치 내에서 또는 WAN(Wide Area Network) 연결을 통해 데이터를 공유하여 글로벌 데이터 협업과 데이터 액세스를 지원할 수 있습니다. IBM Spectrum Scale 은 고성능 컴퓨팅 산업에서 쌓은 수년의 경험을 기반으로 구축된 AI 인프라에 매우 견고한 토대를 제공합니다(그림 2 참조).

그림 2

IBM Spectrum Scale 과 하이브리드 클라우드 데이터 서비스



출처: IDC, 2021 년

IBM Spectrum Scale 은 글로벌 네임스페이스, IBM Spectrum Scale 클러스터 간에 공유된 파일 시스템 액세스, 다수의 노드에서의 동시 파일 액세스, 복제를 통한 높은 회복 가능성과 데이터 가용성, 파일 시스템이 마운트되는 동안 변경할 수 있는 기능, 대규모 환경에서도 간소화된 관리를 제공합니다. IBM Spectrum Scale 의 주요 차별화 요소는 다음과 같습니다.

- IBM Spectrum Scale 클러스터 간에 공유된 파일 시스템 액세스를 제공하므로 한 위치 안에서 또는 WAN 을 통해 별도의 클러스터들 사이의 데이터 공유가 가능함
- 고유의 병렬 파일 시스템을 사용하여 시스템 성능을 향상
- 토큰 관리를 활용하여 클러스터 전반의 클라이언트에 세부적으로 동시에 액세스할 수 있으므로 파일 일관성을 유지할 수 있음
- 파일 시스템 감사 로깅과 같은 기능 그리고 긴 거리를 포괄하는 지능적 마운트와 같은 구성 가능한 기능을 통해 데이터 가용성과 신뢰성 향상
- 시스템 유연성이 향상되어 파일 시스템이 마운트된 동안 디스크 또는 서버 리소스를 추가하거나 삭제할 수 있음
- 간소화된 스토리지 관리로 플래시부터 HDD, 클라우드, 테이프에 이르기까지 계층화된 강력한 정책 기반 자동 스토리지 관리 그리고 심지어 정책 기반 데이터 감축을 통해 정보 라이프사이클 관리(Information Life-cycle Management, ILM)를 달성하도록 지원함
- 대부분의 애플리케이션에서 직접 실행할 수 있는 여러 표준 파일 시스템 인터페이스를 통한 관리 간소화
- 데이터 가용성, 무결성, 보안, 최적화된 컨테이너 네이티브 스토리지 및 Red Hat OpenShift 와의 통합을 제공하는 하이브리드 클라우드 배포 환경

IBM Elastic Storage System

IBM Elastic Storage System 은 구성 요소의 구성과 관리를 용이하게 하도록 설계된, 현대적으로 구현된 소프트웨어 정의 스토리지로, IT 조직이 AI, 빅데이터, 분석 애플리케이션 등 강력한 성능을 요구하는 컴퓨팅 애플리케이션을 위해 확장성이 뛰어난 스토리지를 더 빠르고 쉽게 배포하도록 지원합니다. IBM ESS 의 특징은 다음과 같습니다.

- NVMe 플래시 스토리지로 구축되어 전체 인프라에서 엑사바이트 수준의 확장성과 일관적인 서비스 품질을 제공합니다.
- IBM Spectrum Scale 의 파일 관리 및 데이터 서비스 기능과 통합되어 페더레이션된 글로벌 스토리지 시스템을 제공할 수 있습니다.
- 기업이 엣지 디바이스부터 중앙 데이터센터의 데이터까지 스토리지 요구 사항을 병합하고 퍼블릭 클라우드와 통합되도록 지원하므로 비효율성을 낮추고, 획득 비용을 절감하고, 스토리지 관리를 간소화하고, 데이터 사일로를 제거할 수 있습니다.
- 베어 메탈 또는 가상화된 환경에서 배포된 까다로운 여러 애플리케이션에 일관적인 고성능 액세스를 제공합니다(Kubernetes 와 Red Hat OpenShift 통합을 지원하므로 더 쉽게 클라우드 네이티브 애플리케이션을 배포할 수 있음).

그림 3

IBM Elastic Storage System

출처: IDC, 2021 년

IBM ESS 는 3000 과 5000 이라는 두 개의 폼 팩터로 제공됩니다.

- IBM Elastic Storage System 3000(ESS 3000)은 분석을 위한 데이터 관리 문제를 해결할 수 있도록 설계되었습니다. 컴팩트한 2U 엔클로저에 패키지로 포함된 ESS 3000 은 스토리지가 모두 NVMe 를 사용하고 간편하고 빠르게 컨테이너화된 소프트웨어를 설치하고 업그레이드할 수 있어 인공지능/딥 러닝과 강력한 성능을 요구하는 컴퓨팅 애플리케이션의 가치 실현 시간을 단축합니다. ESS 의 하드웨어 및 소프트웨어 설계 덕분에 데이터에 목마른 컴퓨팅을 모두 활용하는 데 필요한 업계 최고 수준의 성능을 제공합니다.

- IBM ESS 5000 은 소프트웨어 정의 IBM Spectrum Scale 스토리지를 IBM POWER9 프로세서 기반 I/O 집약적 서버와 결합하여 처리량이 높은 페타바이트 단위의 스케일아웃 용량 노드를 제공합니다. IT 팀은 조직 전반의 스토리지 요구 사항을 IBM ESS 5000 과 NVMe 기반 ESS 3000 으로 통합하여 비효율성을 낮추고, 획득 비용을 절감하고, 까다로운 AI, HP, 분석 및/또는 고용량 스토리지 요구 사항(일반적으로 의료, 미디어, 정부, 금융 서비스 분야에서 요구됨)을 지원할 수 있습니다. ESS 5000 은 고비용을 초래하는 데이터 사일로를 제거하는 단일 통합 네임스페이스로 테라바이트 수준에서 시작하여 수백 페타바이트, 심지어 엑사바이트 규모로도 확장할 수 있습니다. IBM Spectrum Scale 은 확장에 따라 처리량을 증가시키는 IBM ESS 5000 의 핵심을 구성하는 병렬 파일 시스템입니다. 이 시스템은 투자 보호를 위해 이전 ESS 모델과의 통합을 지원하며, 클라우드 스토리지와 IBM Tape 와 같은 저비용 옵션을 제공합니다. IBM Spectrum Scale 을 사용하면 IT 팀은 데이터 사일로와 병목 현상을 제거하고, 스토리지 관리를 간소화하여, 데이터 액세스 속도를 높일 수 있습니다.

IBM Cloud Object Storage

IBM Cloud Object Storage(COS)는 업계를 선도하는 확장성이 뛰어난 비용 효율적 소프트웨어 정의 스토리지 솔루션으로 엣지 디바이스, 중앙 데이터센터, 프라이빗 또는 퍼블릭 클라우드에서 비정형 데이터를 저장하는 용도로 사용됩니다. IBM Cloud Object Storage 는 AI, 분석, IoT, 비디오 및 이미지 저장소를 위한 강력한 성능을 요구하는 인프라를 배포하거나 현대화하는 데 적합합니다. 이 스토리지는 새로운 18TB SMR(Shingled Magnetic Resonance) 드라이브를 사용하여 스토리지 비용을 최대 12% 절감하면서 12 노드 클러스터에서 처리량을 최대 55GBps 향상하도록 지원하므로 전례 없는 가치를 제공합니다. 기업은 까다로운 데이터 레이크와 대규모 용량 요구 사항에 맞게 맞춤화할 수 있는 IBM 의 로컬 또는 지리분산(geodispersed) 데이터 보호 기능으로 데이터를 보호할 수 있습니다.

그림 4

IBM Cloud Object Storage

IBM Cloud Object Storage

출처: IDC, 2021 년

IBM Cloud Object Storage 는 AI 인프 ss 라의 토대를 구성하는 시스템으로, 다음과 같은 주요 이점을 제공합니다.

- **확장성:** COS 는 기하급수적인 데이터 증가를 지원하여 성능과 용량을 테라바이트 단위에서 엑사바이트 단위로 확장할 수 있습니다.
- **보안:** COS 는 암호화 기능이 내장되어 있고 잠글 수 있는 정책 기반 WORM(Write Once, Ready Many) 스토리지를 갖고 있습니다.
- **단순성:** COS 는 어느 위치에서나 데이터에 동시에 액세스할 수 있으며, 자동 페일오버, 데이터 재구축, 자동 확장 및 리밸런싱 기능을 제공합니다.

- **비용 절감 효율성:** COS 는 IDA(Information Dispersal Algorithm) 효율성을 달성하여 지리보호(geoprotected) 데이터를 제공하며, 소프트웨어 전용 또는 완전히 지원되는 어플라이언스 솔루션으로 이용할 수 있습니다.
- **검색 기능:** COS 는 맞춤형 인사이트와 검색 기능을 제공하므로 시간을 절약할 수 있습니다. 조직은 맞춤형 메타데이터를 생성하여 가치를 향상할 수 있습니다.
- **향상된 파일 액세스:** SMB 또는 NFS 액세스를 사용하여 새로운 파일 액세스 소프트웨어 게이트웨이를 어느 Windows 또는 Linux 파일 시스템에나 원활하게 연결할 수 있으므로 파일 기반 애플리케이션을 오브젝트 스토리지에 쉽게 연결할 수 있습니다.
- **고속 전송:** IBM Aspera 의 고속 데이터 전송 옵션을 사용하면 데이터를 쉽게 전송할 수 있으며, 유연한 스토리지 클래스 계층 덕분에 데이터 액세스 요구 사항을 충족하면서 비용을 관리할 수 있습니다.

미래 전망

2022 에는 전 세계 GDP 의 65%가 디지털화될 것이며 이로 인해 2020 년부터 2023 년까지의 기간 동안 전 세계 IT 지출이 6 조 8,000 억 달러에 달할 것입니다(*IDC FutureScape: Worldwide Digital Transformation 2021 Predictions,* IDC #US46880818, 2020 년 10 월 참조). 디지털 인프라는 기존의 중앙 엔터프라이즈 서비스나 별개의 클라우드 데이터센터에 국한되지 않습니다. 디지털 인프라는 애플리케이션과 코드의 혁신적 변화를 지원하는 모든 자산과 리소스를 의미합니다. 이러한 인프라는 고객 경험 향상의 토대가 될 것입니다. 또한 비즈니스 운영에 인텔리전스/자동화를 구현하고 엔터프라이즈의 디지털 엣지와 업계에서 계속되는 혁신을 지원합니다. 성공적인 디지털 전략은 디지털 인프라를 혁신하여 사일로를 해소하고 기술 장벽을 부수고 기존의 툴과 애플리케이션에 대한 지원 활동 이상을 수행해야 합니다.

IDC 는 AI 가 디지털 인프라의 토대가 될 것이라고 생각합니다. IDC 의 Customer Insights and Analysis Group 은 최근 다양한 산업에 속한 모든 규모의 조직을 대상으로 현재와 미래의 IT 지출 및 채택 계획에 대해 알아보고자 설문조사를 실시했습니다. 이 연구에서 전 세계의 다양한 산업에 속한 IT 조직의 주요 IT 의사 결정권자 3,600 명 중 거의 76%가 인공 지능이 DX 전략의 주요 구성 요소이거나 향후 1~2 년 동안 주요 구성 요소가 될 것으로 예측한다고 응답했습니다. 설문조사 응답자의 22%만이 AI 가 향후 3~5 년 동안 DX 전략의 주요 구성 요소가 될 것이라고 응답했습니다. AI 가 향후 1~2 년 동안 DX 전략의 주요 구성 요소가 될 것으로 예측한 응답자 중에서 통신, 공공 사업, 교육, 전문 서비스 분야의 조직이 디지털 혁신 노력을 위해 AI 에 의존할 가능성이 가장 높았습니다.

AI 의 핵심은 "가치 실현 시간 단축"입니다. 여기서 가치란 최대한 짧은 시간 내에 기업이 데이터에서 얻을 수 있는 가치를 말합니다. IDC 의 인공 지능에 대한 FutureScape 는 "인공 지능이 우리의 일생에서 가장 파괴적인 혁신"이라고 예측했습니다. AI 는 더 이상 "갖고 있으면 좋은" 기술이 아닙니다. 세계적인 팬데믹으로 인해 AI 채택이 가속화되었고 AI 는 모든 비즈니스 프로세스에서, 어디에서나 점점 더 많이 활용되고 있습니다. 머신 러닝, 대화형 AI, 컴퓨터 비전이 지원하는 AI 솔루션은 비즈니스 회복탄력성을 향상하고 혁신을 가속화하고 혁신적인 고객 경험과 직원 경험을 제공하는 데 앞장서고 있습니다. 앞서 언급한 설문조사 응답자 중 약 51%는 현재 인공 지능을 평가 중이거나 이미 프로덕션 단계에 진입했다고 응답했습니다. 이 수치는 2019 년의 34%에서 상승한 수치입니다. AI 가 주는 가장 큰 영향은 직원이 업무를 더 잘 수행하도록 지원한다는 점입니다. 완전한 배포를 통해 실현되는 이점이 더욱 가시화되면서 엔터프라이즈의 AI 의 채택은 계속 늘어날 것입니다.

AI 스토리지는 2020 년에 11.4% 증가될 것으로 예측됩니다. 전년대비 43.1% 증가할 것으로 예측되는 AI 스토리지의 성장세에 힘입어 전체 하드웨어 시장이 2021 년에 전년대비 35.5% 성장하며 강한 회복세를 보일 것으로 예측됩니다.

IDC 는 앞으로 몇 년 동안 계속 AI 인프라에 대한 투자가 많이 이루어질 것으로 예측합니다. IDC 는 (서버와 스토리지를 합친) AI 하드웨어 매출이 전년대비

10.3% 성장하여 2020 년에 134 억 달러에 이를 것으로 예측합니다. 하드웨어 시장 내에서 AI 스토리지는 2020 년에 11.4% 성장할 것으로 예측됩니다. 전년대비 43.1% 증가할 것으로 예측되는 AI 스토리지의 성장세에 힘입어 전체 하드웨어 시장이 2021 년에 전년대비 35.5% 성장하며 강한 회복세를 보일 것으로 예측됩니다.

이러한 스토리지 중 많은 부분이 워크로드 계층이 중앙 데이터센터와 클라우드 그리고 엣지 사이를 이동하도록 지원하는 원활한 이동성 계층을 가진 하이브리드 클라우드 환경에 구축될 것입니다. 스토리지는 컴퓨팅을 데이터가 있는 곳으로 옮기도록 지원하면서 공통 액세스와 제어 영역을 제공하므로 하이브리드 클라우드의 중대한 토대가 될 것입니다.

스토리지, 특히 하이브리드 클라우드 환경의 스토리지는 AI 이니셔티브가 현재와 미래에 확장되도록 지원하는 기반 역할을 계속 수행할 것입니다. AI 그리고 AI 로 인한 데이터 현대화 이니셔티브에 대한 투자가 스케일아웃 파일 스토리지와 비정형 데이터에 대한 투자를 촉진할 것입니다. IDC 는 최근 IT 인프라 채택 트렌드를 알아보기 위해 전 세계의 IT 실무자와 운영 팀 구성원 624 명을 대상으로 설문조사를 실시했습니다. 이 연구를 통해 IDC 는 응답자의 65% 이상이 AI 와 같은 고성능 워크로드를 위해 로컬로 액세스되거나 NFS 를 통해 액세스되는 스케일아웃 파일 시스템을 선호한다는 것을 알게 되었습니다. 구체적으로 훈련 및 추론 워크로드를 포함하는 AI 워크로드의 경우 성능이 스토리지의 가장 중요한 요구 사항이었습니다. 그 다음으로 중요한 요구 사항은 하이브리드 클라우드에서의 배포 용이성과 서비스 품질이었습니다.

IT 구매자를 위한 필수 지침

기술 구매자는 자체 AI 인프라 스택을 구축하는 과정에서 혼란을 느낄 만합니다. 이들은 적용 사례를 정의하고 AI 이니셔티브를 실행하고 데이터 과학자와 애플리케이션 개발자를 고용하거나 훈련시키는 작업을 완료했지만, 갑자기 AI 모델을 개발할 기반 인프라가 주는 제약을 깨닫고 있습니다. 그리고 이들은 기존 인프라를 단기적으로 활용한 후 가속화된 인프라에 투자하는 경우가 많습니다. 데이터 과학자는 가속화된 인프라를 기반으로 스택을 통합하여 작동하게 하기 위해 노력합니다. 하지만, 이 작업은 엄밀히 말해 이들의 직무가 아닙니다. IT 인프라 팀은 데이터 과학자에게 필요한 스택을 잘 모르기 때문에 스택을 통합하여 최적화할 수 없습니다. 이로 인해 심각한 스킬 공백이 초래되고 서버 공급업체와 클라우드 SP 는 자사의 방식대로 자체 스택을 활용하여 이러한 공백을 메우려고 합니다. 현재 공급업체 수만큼 많은 스택이 존재하며 이러한 스택은 동일한 가치 사슬에 속한 여러 구성원이 개발한 경우 중복되는 경우가 많습니다.

비즈니스 성과를 먼저 고려하기

조직은 AI 인프라에 대한 투자가 도움이 되는 비즈니스 성과를 파악하기 위해 서비스 제약 사항과 적용 사례를 먼저 연결해야 합니다. 이들은 이러한 투자가 실현하는 이점을 정량화하고 측정해야 합니다. 예를 들어, 차별화된 경쟁력을 추구하는 경우, 얼마만큼을 언제까지 투자해야 하는지에 대한 질문을 해야 합니다. 그 다음에는 이러한 기준을 애플리케이션 아키텍처를 선택할 때 적용해야 합니다. 기업은 고객의 감정, 요구, 필요에 대한 이해를 높이고 응답성을 향상하여 브랜드 이미지를 향상하고자 하는 경우 AI 를 고려해야 합니다. 이해도와 응답성을 향상하면 매출과 수익을 증가시킬 수 있습니다.

포괄적인 접근

AI 이니셔티브를 실행할 때는 큰 그림을 보는 것이 중요합니다(즉, 포괄적인 시각을 갖는 것이 중요합니다). 하나의 문제를 그 자체로만 바라보면 여러 아키텍처와 솔루션 간의 통합 및 상호운용성의 부재로 인해 또 다른 사일로를 유발하거나 더 나쁜 경우 환경의 복잡성을 증가시킬 수 있습니다. 데이터 인프라는 엣지디바이스부터 중앙 데이터센터와 클라우드까지 아우르고, AI 및 기타 미션 크리티컬 애플리케이션 등의 적용 사례를 포괄하는 전체적인 솔루션으로 바라봐야 합니다. 데이터 인프라는 중앙 데이터센터, 엣지 디바이스, 클라우드에 배치된 AI 및 기타 미션 크리티컬 애플리케이션 같은 적용 사례를 지원하는 전체적인 솔루션 역할을 수행해야 합니다.

최적의 애플리케이션과 데이터 아키텍처 개발

AI 애플리케이션과 데이터 아키텍처를 개발하는 일은 복잡한 작업입니다. 이 작업을 수행하려면 비즈니스 요구 사항과 성과를 결정성을 갖는 AI 지원 워크플로우로 전환해야 합니다. 이러한 워크플로우는 AI 기능이 애플리케이션 행동을 어떻게 향상하는지, 데이터가 어떻게 수집되고 분석되는지, 애플리케이션이 어떻게 다른 비즈니스 애플리케이션 및 사용자와 상호작용하는지 설명해야 합니다. 여기서 초점은 애플리케이션이 데이터를 소비하고 생산하고 분석하는 방식, 그리고 이것이 하드웨어에 미치는 영향에 맞춰져야 합니다. 그린필드 계획을 수립할 때는 맞춤형(오픈소스 또는 독점) 소프트웨어 구성 요소와 기성품 소프트웨어 구성 요소의 조합에 초점을 맞춰야 합니다.

최적의 참조 스택 선택

여러 공급업체와 서비스 제공업체가 AI 인프라 실행을 위한 참조 스택을 내놓았습니다. 이러한 스택 중 다수는 "개방적" 속성을 가지고 있어 모듈식 "플러그 앤 플레이" 경험을 제공하며, 사용량 기반 지불 방식의 서비스로 소비할 수 있으므로 자본 비용 친화적입니다. AI 인프라 투자는 순식간에 비용이 크게 증가할 수 있으므로 이는 중요한 고려 사항입니다. IDC 는 앞으로 발표할 문서에서 널리 사용되는 공급업체 참조 스택에 대한 견해를 공개할 계획입니다.

참조 스택을 평가할 때 기억해야 할 IT 관련 이점으로는 비용 절감, 데이터 및 애플리케이션 가용성 달성, 효과적인 인프라 활용 및 통합, 그리고 가능한 경우 상호운용 가능한 단일 애플리케이션 제공 플랫폼 구현이 있습니다.

AI 를 위한 정보 아키텍처 구축

기업은 데이터의 위치에 상관없이 모든 유형의, 모든 데이터에 유연하고 체계적인 액세스를 제공하고 참조 아키텍처와 관련된 우려 사항을 해결하는 데이터 관리 전략을 수립해야 합니다. 현대화 노력을 통해 선택권 및 유연성과 함께, 다른 플랫폼과 통신할 수 있는 확장 가능한 개방적 토대를 제공하는 정보 아키텍처를 정의하고 구축할 수 있습니다. AI 로의 여정을 가속화하는 IBM 의 하이브리드 데이터 관리 전략은 수집, 구성, 분석, 주입이라는 4 단계로 구성된 AI Ladder 가 정의하는 처방적 접근법입니다.

- **수집:** 어느 데이터베이스에서나 또는 스토리지 시설에서나, 최적의 위치에서 데이터를 단순화하고 액세스할 수 있도록 합니다.
- **구성:** 데이터 프로파일링, 클렌징, 카탈로깅과 같은 정보 라이프사이클의 모든 단계에서 데이터의 신뢰성, 완전성, 일관성을 유지하고, 보호를 제공하고 규정 준수를 지원하며, 정책 기반 가시성, 탐지, 보고를 지원합니다.
- **분석:** 정형 및 비정형 데이터를 모두 탐색 및 분석하고 이를 안전하게 활용하는 통합된 툴을 사용하여 AI 모델을 구축, 배포, 관리합니다.
- **주입:** 제공된 솔루션과 서비스를 사용하여 모델이 추천한 결정의 신뢰성과 투명성을 달성하고, 결정을 설명하고, 성향을 탐지하는 등의 작업을 수행합니다.

AI 여정의 본질은 조직 전체에 손쉽게 주입되는 정보 아키텍처를 통해 수집한 데이터를 인사이트로 바꾸는 데 있습니다. AI Ladder 의 각 부분이 전체 여정에 통합되는 것이 중요합니다. 스토리지는 일반적으로 데이터 사일로를 유발하는 특정 스토리지 솔루션, 그리고 함께 통합되지 않았거나 포괄적인 인프라 솔루션 세트가 아닌 솔루션으로서 전술적으로 구축되어 왔습니다. 고객이 대용량 파일 또는 오브젝트 스토리지 시스템에 데이터를 저장한 다음, 데이터에 대한 세부 정보를 모르거나 데이터를 통해 추가적인 인사이트를 도출하지 못하는 경우가 있습니다. 고객은 여정의 한 부분에서 프로젝트를 시작하거나 프로젝트의 초점을 이 부분에 맞출 수 있습니다. 그러나 AI 워크로드 확장을 위해 리소스를 최적화하고 인프라를 현대화하려면 각 프로젝트는 전체 AI 정보 아키텍처를 고려해야 합니다.

스케일아웃 파일 시스템, 이기종 컴퓨팅, 분산 컴퓨팅 및 스토리지 액세스를 위한 고속 상호 연결과 같은 중요한 요소를 차용하여 AI 를 고성능 복합 애플리케이션(상호 연결된 여러 애플리케이션으로 구성됨)으로 간주하는 조직이 궁극적으로 AI 인프라를 확장할 수 있습니다.

최적의 파트너십 활용

IT 구매자가 장기적으로 성공하려면 엔드투엔드 솔루션 제공업체와 반드시 협력해야 합니다. 그러나 IDC 는 현재 시중의 어떤 공급업체도 이러한 엔드투엔드 환경을 아직 제공하고자 하나, 아직 까지는 어떤 공급업체도 제공하지 못하고 있다고 생각합니다. 그럼에도 불구하고 신뢰할 수 있는 파트너와 협력하면 조직은 비즈니스 성장을 위한 AI 접근법을 더 효과적으로 활용할 수 있습니다. 민첩성과 적응력을 향상하고 내부적 시너지 효과를 활용하여 수익성을 높일 수 있습니다. 마지막으로, 조직은 혁신을 통해 업계 내에 들이닥친 파괴적 혁신의 물결보다 앞서 나갈 수 있습니다. 이상적인 파트너는 다음과 같은 사항을 제공해야 합니다.

- 소규모 연구 배포 환경에서 대규모 글로벌 배포 환경으로 확장 가능한 것으로 입증된 솔루션
- 비즈니스 목표에 부합하는 부문에 대한 수직적 전문 지식
- 여러 ISV 와 인프라 제공업체로 구성된 다양한 생태계에 대한 통합 액세스
- 새로운 데이터 소스에 대한 장기적 투자 가치를 확보하고 극대화하도록 지원하는 데이터 중심적 시각
- 대규모 프로젝트의 하드웨어, 소프트웨어, 보안 관련 측면을 성공적으로 간소화하는 검증된 역량

문제/기회

조직의 경우

이 백서에서는 조직이 AI 애플리케이션을 프로덕션으로 확장할 준비가 되었을 때 직면하게 되는 다양한 문제에 대해 설명했습니다. 데이터 준비부터 모델 개발, 런타임 환경, 그리고 AI 모델의 훈련, 배포, 관리에 이르기까지, 기반 인프라의 요구 사항을 범용 하드웨어라는 구식 모델로 충족할 수 없습니다. 데이터 집약적 워크로드를 위해 설계되고, 탁월한 성능, 확장성, 데이터 액세스 및 통합 기능을 갖추었으며, 하이브리드 클라우드 환경으로 통합될 수 있는 인프라에 투자하면 장기적인 가치와 서비스 품질을 확보할 수 있습니다. 조직은 AI 를 위한 처리 작업에 맞게 구축된 스토리지 시스템으로 기존 범용 스토리지 플랫폼을 교체하거나 보충할지에 대한 결정을 내려야 할 것입니다. 이를 통해 첨단 AI 애플리케이션을 개발하고 실행하기 위한 토대를 마련할 수 있을 것입니다.

IBM 의 경우

IBM 의 어려움은 항상 시장의 인식과 관련이 있습니다. IBM 은 인프라 소프트웨어 스택(예: Red Hat OpenShift) 및 퍼블릭 클라우드와 통합된 포괄적이고 탁월한 AI 인프라 솔루션을 제공합니다. 그러나 잠재 고객은 이를 복잡하고 비용이 더 많이 드는 것으로 잘못 인식합니다. 매우 데이터 집약적인 워크로드를 위한 대규모 상용 스토리지 시스템 공급업체 중 하나에 뒤따르는 자동 반응은 진정한 이점을 제공할 수 있는 AI 인프라 솔루션을 이러한 조직으로부터 박탈합니다. 예를 들면, IBM Spectrum Scale 및 IBM Elastic Storage System 을 사용하는 IBM Storage for data and AI 솔루션은 가장 대규모인 데이터센터 중 다수를 위한 슈퍼컴퓨팅 구성 요소입니다. 이제 새로운 AI 워크로드가 HPC 요구 사항을 충족하는 슈퍼컴퓨팅 배포 환경을 진지하게 모방하기 시작하고 있고, 이러한 환경이 실행되는 온프레미스 및 클라우드 인프라에 도전을 안겨주고 있으므로, 지금이 IBM 이 무대를 장악하고 신규 고객을 확보해야 할 때입니다.

결론

지난 몇 년 동안 IDC 는 많은 조직에서 진행되는 AI 혁신, 그리고 이들이 어떻게 다양한 AI 역량을 개발하는지를 목격해왔습니다. 처음에는 상대적으로 경험이 부족한 직원의 실험적 시도로 시작되었고 당장 사용 가능한 인프라에서 실행된 이러한 이니셔티브는 이제 임계 질량에 도달하기 시작했습니다. 많은 조직이 광범위한 AI 전문 지식을 얻었고, AI 기능이 매우 빨리 비즈니스의 중대한 측면이 되고 있는 것을 직접 경험하고 있습니다.

또한 IT 조직 역시 AI 를 실행할 인프라와 관련하여 학습을 통해 혁신을 겪었습니다. 이제 딥 러닝 훈련 또는 추론을 위한 인프라 요구 사항과 이러한 환경을 프로덕션을 위해 확장하는 방법이 훨씬 더 분명해졌습니다. 딥 러닝 훈련에는 다른 애플리케이션과는 다른 인프라가 요구된다는 점은 당연한 사실입니다. 딥 러닝 훈련에는 강력한 프로세서와 코프로세서, 확장 가능한 고속 스토리지 지원 상호 연결, 대규모 I/O 대역폭, 그리고 충분한 메모리를 갖춘 클러스터링된 노드가 필요합니다.

현재 IT 조직이 내려야 하는 가장 중요한 결정은 동급 최고의 시스템으로 데이터 인프라 AI 애플리케이션을 가장 효과적으로 설계하고 배포하는 방법과 이들을 함께 최적화하는 방법에 관한 것입니다.

IDC 는 Spectrum Scale, Elastic Storage System, Cloud Object Storage 를 포함하는 IBM 의 데이터 및 AI 스토리지 솔루션이 전례 없는 가치와 성능을 제공한다고 생각합니다.

현재 IT 조직이 내려야 하는 가장 중요한 결정은 동급 최고의 시스템으로 데이터 인프라 AI 애플리케이션을 가장 효과적으로 설계하고 배포하는 방법과 이들을 함께 최적화하는 방법에 관한 것입니다.

IDC 정보

IDC(International Data Corporation)는 정보 기술, 원격 통신, 소비자 기술 시장에 대한 시장 인텔리전스, 자문 서비스 및 이벤트를 제공하는 최고의 글로벌 제공업체입니다. IDC 는 IT 전문가, 비즈니스 경영진 및 투자 커뮤니티가 기술 구매 및 비즈니스 전략에 대해 사실에 근거한 의사결정을 내리도록 지원합니다. 1,100 명 이상의 IDC 분석가가 전 세계 110 개 이상의 국가에서 기술 및 산업 기회와 트렌드에 대한 글로벌, 지역 및 현지 전문 지식을 제공합니다. IDC 는 50 년 동안 고객이 핵심 비즈니스 목표를 달성할 수 있도록 전략적 인사이트를 제공해 왔습니다. IDC 는 세계 최고의 기술 미디어, 연구 및 이벤트 기업인 IDG 의 자회사입니다.

글로벌 본사

140 Kendrick Street
Building B
Needham, MA 02494
USA
508.872.8200sss
Twitter: @IDC
idc-community.com
www.idc.com