White Paper

엔터프라이즈 AI 를 위한 인프라 재고

스폰서: IBM

Peter Rutten
2018 년 6 월

IDC OPINION

IDC 는 단일 아키텍처가 데이터 센터의 모든 컴퓨팅을 주도하는 정형화된 컴퓨팅 시대는 끝났다고 강하게 믿고 있습니다. 점점 더 많은 기업이 인공 지능(AI) 이니셔티브를 출범하기 시작함에 따라 이러한 사실이 점점 더 분명해지고 있습니다. 많은 사람들은 AI 의 실험 단계에 있으며 일부는 생산 가능 상태에 도달했습니다. 하지만 이들은 모두 새로 개발된 AI 애플리케이션 및 서비스를 계속해서 실행하기 위해 인프라 옵션을 그동안 경험하지 못하였던 사이클로 빠르게 돌리고 있습니다.

인프라를 끊임없이 점검하게 되는 주요 원인은 데이터 센터에서 대량의 워크로드에 사용하는 표준 인프라가 대부분 AI 의 데이터 집약적인 특성에 적합하지 않기 때문입니다. 기존 서버의 성능과 I/O 는 딥 러닝(DL)에는 여전히 부족할 뿐만 아니라 AI 모델 개발에 기반이 되는 데이터 레이크가 이러한 중요한 임무를 수행하는 데 적합하지 않습니다. 데이터 레이크는 AI 모델링을 준비하는 데 몇 개월은 아니여도 몇 주가 걸리는 기존 스키마 기반의 느린 단일 방식으로 구성되어 있습니다. 이 데이터 레이크는 비즈니스에서 중요하지 않은 것으로 간주되지만 일단 AI 가 개발되기 시작하면 중요성이 높아질 것입니다.

AI 는 데이터 센터에서 새롭고 다양한 프로세서에 대해 이야기를 펼치는 연극에서 주연을 맡고 있습니다. 이러한 다양성은 특정 워크로드를 위한 GPU, FPGA, 많은 코어 프로세서 및 ASIC 가 증가하는 것 뿐만 아니라 다른 호스트 프로세서와 호스트와 가속기 사이의 향상된 링크로의 발전을 통해서 자체적으로 증명되고 있습니다. 가속기는 많은 성능 지연을 완화할 수 있을 뿐만 아니라 호스트 프로세서와 상호 작용하여 AI 와 같은 워크로드에 대해 진정으로 탁월한 성능을 제공할 수 있습니다.

이 백서에서는 이러한 당면 과제에 대해 논의하고 이를 위해 IBM 이 어떤 제안을 하고 있는지 알아봅니다.

SITUATION OVERVIEW

예전에는 "AI 가 오고 있다"고 말했지만 이제는 "AI 가 바로 여기에 있다."고 말합니다. AI 는 IT 조직 뿐만 아니라 CIO, CTO 및 CEO 도 긴급히 다음과 같은 질문을 던지면서 사업을 재고하게 만드는 대격변을 일으키고 있습니다. AI 를 활용할 수 있는 방법이 있을까? AI 를 어떻게 사용할 수 있을까? 목표는 어떻게 달성할 수 있을까? AI 의 5~10%를 차지하는 DL 과 학습은 많은 관심을 받고 있지만 AI 혁신을 준비하는 방법, 즉, AI 가 기업의 데이터 관리 방법과 기업의 인프라에 어떤 영향을 주게 될 것인지에 대한 인식은 상대적으로 낮은 편입니다.

우리에게는 수 많은 데이터가 있습니다. 연결된 자동차, 웨어러블 건강기기, 연결된 머신 및 센서 탑재 장치 등 끊임 없이 늘어나는 여러 기기가 데이터 홍수를 일으키고 있습니다. 많은 기업이 데이터 레이크를 구축했다는 점에서는 빅 데이터 혁신이 성공적이라고 볼 수 있습니다. 하지만 아직 데이터 레이크에서 가치를 이끌어내는 방법을 찾지 못했습니다. LOB(Line of business) 관리자가 데이터

레이크에서 조직의 데이터 과학자로부터 유용한 정보를 얻으려면 몇 주 또는 때로는 몇 개월이 걸릴 수 있습니다.

오늘날의 데이터 레이크는 일부 통찰력을 제공하지만 활용이 어려우며 AI 애플리케이션을 구축할 수 있는 최고의 기반이라고는 할 수 없습니다. 예를 들어 AI 혁명은 빅 데이터를 다시 논의하기 위한 초대장이라고 할 수 있으며, 이제는 새로운 AI 서비스를 더 잘 지원합니다. 데이터 홍수는 기계 학습, 특히 패턴을 식별하고 통찰력을 드러낼 수 있는 지능형 시스템을 만들고 교육하는 데 필요한 딥 러닝 기능을 강화합니다. 오늘날의 궁극적인 데이터 집약적인 워크로드는 대용량 딥 러닝이며, 다음과 같은 질문이 중요합니다. 기업이 AI 를 완전히 수용하는 데 필요한 서버 및 소프트웨어 플랫폼은 무엇인가?

데이터 통찰력과 전문 지식 및 데이터에서 가치를 창출할 수 있는 능력을 통해 회사는 비즈니스 운영 방식을 혁신하고 고객과 상호 작용할 수 있습니다. AI 는 회사가 하나의 사업체 또는 조직으로 운영하는 공간에서 깊이 숨겨진 복잡성을 진정으로 이해하기 위한 방식입니다. AI 및 특히 DL 은 기존 데이터 분석을 가속화하여 새로운 가치를 풍부하게 제공합니다. 그러나 DL 역시 수많은 도전 과제를 제시합니다. 특히 오픈 소스에서 매우 빠른 혁신과 함께 급속히 진화하는 기술의 집약체이므로 모든 부분을 잘 관리하고 원활하게 연동하는 것이 어렵습니다.

이제 해결해야 할 과제는 DL 에 적합한 인프라와 소프트웨어(사용하기 쉽고 신속하게 배포할 수 있을 뿐 아니라 하드웨어와 잘 통합되고 공급업체에 의해 완벽하게 지원되는)를 선택하고 데이터를 준비하는 것입니다.

이 중에서 특히 DL 을 위한 데이터를 준비하는 것은 AI 로 전환하는 과정에서 가장 큰 장애물입니다. 기업은 DL 을 학습하고, 학습을 위한 인프라를 최적화하게 되자 데이터를 준비하는 데 몇 달이 걸렸음을 알게 되었습니다. AI 프로세스 설정에서 가장 시간이 많이 소요된 작업은 데이터를 변환하고 그것을 A 지점으로부터 사용할 수 있는 B 지점으로 가져오는 것이었습니다. 또한, 진정으로 효과적인 AI 애플리케이션은 내부 데이터, 스트리밍 데이터 등 여러 소스로부터 데이터를 추출하여 이것을 의미있는 방식으로 결합하는 것이라는 사실이 일을 한층 더 복잡하게 만들었습니다.

IBM AI 인프라 접근 방식

AI 는 딥러닝 워크플로우에 연결되어야 하는 기존 데이터 저장소 및 애널리틱스를 활용해야 합니다. 이에 IBM 은 기업이 AI 환경을 구축하는 데 과도한 시간을 들이게 되는 요소들에 중점을 두었습니다. 이 중에서 가장 첫 번째 요소는 데이터를 준비하는 과정입니다.

IBM 은 AI 애플리케이션을 위한 데이터 준비 프로세스를 간소화하고 연결을 구축하고 이러한 AI 애플리케이션을 위한 다양한 데이터 소스를 합성하는 작업을 시작했습니다. AI 용 데이터 소스에는 기업의 통제 하에 있는 데이터 소스(예: 고객 환경설정), 기존의 고객 행동 패턴을 기반으로 하는 데이터, 그리고 스트리밍 데이터(예: 소셜 미디어 행동)와 같은 외부 데이터 등 다양한 유형이 있습니다.

IBM 은 AI 인프라와 AI 준비를 위해 기업이 데이터를 관리하는 방식을 재고할 필요가 있다는 견해를 갖고 있습니다. 기업의 이러한 전환을 지원하기 위해 IBM 은 데이터 파이프 라인, 서비스 및 AI 를 위한 단일 플랫폼으로 구성된 "AI 인프라"라는 개념을 도입했습니다. 이것은 본질적으로 기업이 자체적으로 AI 혁명을 지원할 수 있도록 설계된 엔드투엔드 서버, 스토리지 및 소프트웨어 플랫폼입니다.

이 플랫폼은 향상된 스토리지, 엔터프라이즈급 Hadoop 및 Spark, 향상된 데이터 관리 기능을 갖춘 최신 데이터 레이크를 기반으로 합니다. 이러한 기반은 다중 스키마(RDBMS, NoSQL, 그래프), 다중 아키텍처(CPU, GPU 가속화, 메모리 내) 및 탁월한 시스템 성능을 수용하는 데이터 플랫폼을 지원합니다. 보다 우수한 데이터 레이크와 데이터 플랫폼을 통해 CRM, IoT 및 부정 행위 탐지와 같은 실시간 정보를 위한 서비스를 제공하는 동적 스토리지, I/O 및 메모리가 탑재된 유연한 IT 환경을 구현할

수 있습니다. 이러한 빌딩 블록은 AI 로 기존 애플리케이션을 활용하고 새로운 AI 기반 애플리케이션을 개발하기 위한 딥러닝 학습의 발판이 됩니다.

AI 인프라의 핵심: IBM Power System AC922 및 IBM PowerAI

IBM Power System AC922

이 AI 인프라의 핵심에는 2017 년 12 월 IBM 에서 출시한 IBM 의 가속화된 POWER9 시스템이 있습니다. IBM Power System AC922 라고 불리는 이 시스템은 AI 워크로드 전용으로 설계되었습니다. Power AC922 는 NVIDIA GPU 를 통한 AI 가속화에 최적화되고 미세 조정된 하드웨어 및 소프트웨어와 고급 I/O 인터페이스가 탑재된 2 소켓 서버입니다. 이것은 또한 CORAL Summit 슈퍼컴퓨터의 기반이 되는 강력한 성능을 자랑합니다. 이 시스템에는 PCIe Gen 4, CAPI 2.0, OpenCAPI 및 NVLink 와 같은 매우 빠른 I/O 아키텍처가 포함되어 있어 데이터 집약적인 워크로드에 이상적입니다.

2 개의 POWER9 CPU 에는 각각 호스트 CPU 병렬 처리를 위한 4 방향 멀티 스레딩과 NVLink 와 함께 2~6 개의 NVIDIA Tesla V100 GPU 가 탑재되어 있어 CPU, HPC 용 GPU, 딥러닝 및 AI 워크로드 사이에서 탁월한 성능을 제공합니다. 또한 NVLink 가 프로세서에 통합되어 있어("2 세대 NVLink"라고 표시됨) 시스템의 처리량을 추가로 증가시킵니다. 이 시스템은 완전한 메모리 가간섭성(Coherency)을 제공하고 가속화된 애플리케이션이 시스템 메모리를 GPU 메모리로 사용할 수 있게 하여 GPU 의 16GB 또는 32GB 메모리 제한을 극복합니다. Power AC922 구성은 16 코어에서 44 코어까지 가능합니다.

IBM PowerAI

IBM 에서는 IBM PowerAI 를 "분산된 딥러닝을 위한 엔터프라이즈 제품"이라고 부릅니다. HPC 또는 IBM Power AC922 을 위한 IBM Power S822LC 상에 구축된 이 제품은 Caffe, Torch, TensorFlow, Theano 및 Chainer 와 같은 오픈 소스 딥러닝 프레임워크 및 도구와 함께 패키지로 제공됩니다. 또한 DIGITS, OpenBLAS, 분산 프레임워크, Bazel 및 NCCL 과 같은 보조 라이브러리가 포함되어 있습니다. 이 하드웨어/소프트웨어 패키지의 목적은 엔터프라이즈에서 바로 사용할 수 있고 엔터프라이즈에서 지원하는 필수 오픈 소스 소프트웨어, 고성능 하드웨어 및 데이터 과학자들의 생산성을 높여줄 통합 도구의 배포판을 제공함으로써 기업이 빠르게 딥러닝을 시작할 수 있도록 지원하는 것입니다.

IBM 이 PowerAI 를 통해 추구하는 목표는 원활하게 동작하는 AI 인프라 구축의 복잡성을 줄이고 최적화된 딥러닝 학습 프레임워크를 쉽게 갖추고 데이터 과학자가 TensorFlow 패키지를 디버그하는 데 시간을 낭비하지 않도록 지원하는 것입니다. 그 결과 약 45 분만에 베어메탈을 학습용 장치로 탈바꿈하여 완전한 딥러닝 환경을 완성할 수 있습니다.

IBM 은 베어 메탈에서부터 딥러닝 환경까지 전체 스택을 지원하는 PowerAI 프레임워크에 대한 업데이트를 발표했습니다. PowerAI 플랫폼에 대한 이 전체 스택 지원은 일반적으로는 AI 를, 특별하게는 DL 을 수용하고자 하지만 주로 오픈 소스 프로젝트 코드에 의해 주도되는 환경을 지원해야 하는 책임을 부담스러워 하는 기업 고객에게 좋은 소식입니다.

최근까지 PowerAI 플랫폼은 IBM 의 가속 시스템인 HPC 용 Power S822LC 를 기반으로 구축되었으며 현재 HPC 는 현재 구축 중인 대형 슈퍼 컴퓨터 두 대에서 사용되고 있습니다. 데이터가 가능한 한 자유롭게 흐를 수 있도록 설계되었습니다. HPC 용 Power S822LC 에서는 GPU 를 사용한 가속화가 서버 설계의 핵심입니다. 서버의 모든 컴퓨팅 엔진을 지점 간 고속 NVLink 로 연결합니다.

Power AC922 에서 NVLink 를 사용하면 물리적 연결이 프로세서 다이의 일부가 되어 GPU 를 데이터가 처리되는 방식 측면에서 CPU 에 종속되는 것이 아니라 CPU 에 대한 피어(peer)처럼 만듭니다. GPU 와 CPU 사이의 이중 NVLink 연결은 공유 PCIe 3.0 버스 또는 스위치에 의존하는 설계와 달리, 시스템

메모리에 대한 매우 직접적이고 거의 일관적인 액세스를 제공합니다. 이렇게 대기 시간이 짧은 연결은 더 크고 복잡한 신경 모델 및 훨씬 더 큰 딥러닝 세트를 지원합니다.

AI 인프라의 기초: 향상된 데이터 레이크

Power AC922 를 기업 AI 인프라 배치의 핵심으로 사용할지 여부는 선택할 수 있지만 이 환경의 기반은 AI 서비스에 적합한 향상된 데이터 레이크여야 한다는 사실을 명심하십시오. 향상된 데이터 레이크를 구축하기 위해서는 다음 다섯 가지 주요 측면을 고려해야 합니다.

- AI 시대의 데이터 아키텍처
- 고속 및 엔터프라이즈급 Hadoop 및 Spark 를 지원하는 고성능 서버
- 통합되고 호환성이 뛰어나며 엔터프라이즈에서 바로 사용할 수 있는 스케일아웃 스토리지
- 현재의 관습에 비해 더 쉽고 간편한 데이터 관리
- 여러 데이터 플랫폼 지원 - 스키마 및 아키텍처

다음 섹션에서는 이러한 다섯 가지 요소와 관련하여 AI 를 지원하는 데이터 레이크를 구축하기 위한 IBM 의 접근 방식에 대해 자세히 설명합니다.

AI 시대의 데이터 아키텍처

오늘날의 데이터 레이크는 "빅 데이터"와 이들에게 권한을 주는 소프트웨어 정의 인프라 혁명 이전의 데이터 아키텍처로 인해 병목 현상이 발생합니다. 데이터를 저장하는 소프트웨어 정의 스토리지나, 데이터를 분석하는 서비스나, 데이터를 이동시키는 인프라나, 거의 모든 것이 서버 상에서 실행되기 때문입니다. 또한 이들 서버의 데이터 아키텍처는 2010 년에 출시된 PCIe 3.0 으로, 이것은 Hadoop 최초 릴리스가 빅 데이터 혁명을 일으키기 시작하기 1 년 전이고 SNIA(Storage Networking Industry Association)가 소프트웨어 정의 스토리지에 대한 표준을 정의하려고 시도했던 시기보다 2 년 앞선 때입니다. 대부분의 서버를 구동하는 프로세서의 처리 속도는 빨라졌지만 I/O 라고 불리는 데이터 이동 및 관리 능력은 크게 변하지 않았습니다.

더 빨라진 데이터 서버

PCIe Gen 3 버스로 인해 엔터프라이즈 AI 를 실행하는 데 필요한 데이터 레이크 및 분석을 구동하는 대부분의 서버에 I/O 제약에 의해 병목 현상이 발생합니다. IBM Power 서버는 I/O 대역폭과 성능을 크게 향상시켰습니다. IBM Power 는 PCIe 3.0 보다 2 배 빠른 IBM POWER9 서버에서 PCIe 4.0 인터페이스를 도입하였습니다. 이는 PCIe 가 탑재된 스토리지 및 FPGA 의 성은을 크게 향상시킴으로써 서버 내에서의 데이터 이동을 향상시킵니다. 또한 GPU 가속 분석을 활용하는 데이터 레이크의 경우 2 세대 NVLink 인터페이스는 PCIe Gen3 보다 약 5~6 배 향상된 대역폭을 제공합니다. IBM Power 는 PCIe 3.0 보다 2 배 빠른 IBM POWER9 서버에서 PCIe 4.0 인터페이스를 도입했습니다. 이는 PCIe 부착 스토리지 및 FPGA 의 성능을 크게 향상시킴으로써 서버 내에서의 데이터 이동을 향상시킵니다. 또한 클러스터를 연결하는 네트워크 인터페이스의 속도를 두 배로 증가시켜 클러스터의 성능을 향상시킵니다. 또한 GPU 가속 분석을 활용하는 데이터 레이크의 경우 2 세대 NVLink 인터페이스는 PCIe Gen 3 보다 약 5-6 배 향상된 대역폭을 제공합니다.

더 쉽고 더 빠른 엔터프라이즈급 Hadoop 및 Spark

오늘날 대부분의 기업에서 Hadoop 과 Spark 는 미션 크리티컬 애플리케이션으로 간주되지 않으며 많은 조직에서 데이터 레이크는 여전히 비즈니스 운영에 필수적이지 않은 것으로 간주됩니다. 그러나 AI 플랫폼을 구축하게 되면 데이터 레이크가 지원 시스템에서 ERP 또는 CRM 과 같은 미션 크리티컬 애플리케이션을 실행하는 환경으로 빠르게 전환됩니다.

이는 기업이 데이터 레이크를 엔터프라이즈에서 바로 사용 가능한 미션 크리티컬 하드웨어 및 소프트웨어로 구축해야 한다는 것을 의미합니다. IBM 은 안정성을 보장하기 위해 전체 IBM Power 하드웨어 및 소프트웨어 스택을 검증하고 지원한다고 말합니다. 하드웨어 측면에서 IBM Power 는 다른 프로세서 아키텍처보다 객관적으로 더 높은 코어당 성능을 제공하기 때문에 몇몇 노드가 대체 아키텍처의 대형 클러스터보다 성능이 뛰어납니다.

소프트웨어의 경우에는 엔터프라이즈급 준비성이 Spark 및 MapReduce 의 다양한 HDFS 관련 문제를 해결해주는 IBM Elastic Storage Server(ESS) 플랫폼에서 시작됩니다.IBM 은 또한 다양한 소스 및 다양한 형식의 대용량 데이터를 저장, 처리 및 분석할 수 있는 대규모 확장형 오픈소스 플랫폼인 Hortonworks 와 함께 엔터프라이즈 지원을 제공합니다. Hortonworks 에는 MapReduce, HDFS, HCatalog, Pig, Hive, HBase, ZooKeeper 및 Ambari 가 포함됩니다.

향상된 스토리지

IBM 은 IBM ESS 가 고급 데이터 레이크를 위한 이상적인 스토리지 접근 방식이라고 믿습니다. IBM ESS 는 IBM Spectrum Scale 소프트웨어와 IBM POWER 서버 및 스토리지 인클로저를 결합한 소프트웨어 정의 스토리지 솔루션입니다. IBM ESS 의 핵심을 차지하고 있는 병렬 파일 시스템인 IBM Spectrum Scale 은 단일 네임 스페이스를 제공하면서 시스템 처리량을 확장합니다. 따라서 데이터 사일로의 생성을 피하면서 고성능을 구현하고 스토리지를 보다 쉽게 관리할 수 있습니다.

IBM 은 ESS 를 기반으로 Spark, MapReduce 및 특정 딥러닝 프레임워크와 같은 대규모 확장형 데이터 애플리케이션에 사용되는 HDFS 에 대한 엔터프라이즈급의 대안을 제시합니다. HDFS 는 3 대 1 복제 모델과 전용 데이터 사일로 때문에 비효율적일 수 있습니다. 단순한 작업에서는 빠르지만 크고 복잡한 프로세스에서는 느려질 수 있습니다. 또한 HDFS 는 표준 프로토콜 지원이 제한되어 있으며 다른 엔터프라이즈 인프라와 쉽게 통합할 수 없습니다.

IBM ESS 는 연산 집약적인 병렬 워크로드를 지원하기 위해 여러 산업 분야에 배포되는 소프트웨어 정의 스토리지인 IBM Spectrum Scale 에 의해 구동됩니다. IBM ESS 는 HDFS 와 외관이 비슷하고 HDFS 를 매우 잘 수행합니다. CIFS(Common Internet File System), NFS, 오브젝트, 블록 스토리지 등의 데이터에 대한 다중 프로토콜 액세스를 지원합니다. IBM Spectrum Scale 은 데이터 사용량 또는 기타 정의된 기준에 따라 플래시, 디스크, 테이프 및 클라우드를 계층화하는 데 사용할 수 있는 정책 기반 데이터 이동 기능을 갖추고 있습니다. IBM ESS 는 다른 애플리케이션과 함께 사용하거나 비용 효율적인 보관을 위해 데이터를 그대로 두면서 기업에게 다른 워크로드뿐만 아니라 빅 데이터 분석을 실행할 수 있는 스토리지 플랫폼을 제공합니다.

향상되고 쉬워진 데이터 관리

Spark 는 딥러닝을 위한 훌륭한 오픈 소스 빅 데이터 분석 프레임워크이지만 Spark 를 구현하는 것은 간단한 일이 아닙니다. 기업은 올바른 도구, 기술 세트 및 워크플로우뿐만 아니라 다른 프레임워크와의 통합을 통해 효율적이고 안전하게 작동하고 관리 효율성을 보장해야 합니다. IBM 은 기업이 이러한 장애물을 극복할 수 있도록 Spectrum Conductor 를 설계했습니다. 기업은 환경의 다양한 구성요소를 자체적으로 통합하는 대신 Spark 의 도입을 포함하여 Spark 및 기타 프레임워크의 다중 테넌트를 지원하며 HPC 인프라가 번창하는 동적 리소스 할당을 지원하는 완벽한 통합 솔루션을 얻을 수 있습니다.

IBM Spectrum Conductor 는 Spark 를 컴퓨팅 및 전송 계층으로 사용하여 데이터 레이크의 다양한 데이터 소스로부터 데이터를 가져옵니다. 이 솔루션은 사용자가 제공하는 정의에 따라 벡터 정보와 메타데이터를 데이터에 추가한 다음, 전체 데이터 변환을 실행합니다. 이것은 수백 개의 Spark 인스턴스를 생성하여 다양한 소스의 데이터를 처리하고 데이터를 Caffe 또는 TensorFlow 와 같은 데이터 세트로 변환합니다.

IBM Spectrum Conductor 는 분석 및 딥러닝 워크로드에 최적화된 워크로드 및 리소스 관리자입니다. IBM 은 YARN 대신 IBM Spectrum Conductor 를 사용하여 이러한 워크로드를 처리함으로써 작업을 예측 가능한 런타임으로 효율적으로 예약하고 적합한 GPU 및 CPU 사양, 올바른 양의 메모리 등과 같이 필요한 리소스를 확보 할 것을 제안합니다. 워크로드의 특성에 따라 IBM Spectrum Conductor 는 사용 가능한 리소스, 사용 가능한 리소스의 수, 리소스를 사용할 수 있는 위치 및 대기열 순서를 결정합니다. 이것은 활용도를 극대화하는데, IBM 에 따르면 평균 데이터 센터의 활용도인 약 20%보다 훨씬 높은 40% 이상을 달성할 수 있다고 합니다. 결과적으로 Spark 가 보다 효과적이고 효율적으로 작동할 수 있는 상당한 성능 향상을 가져옵니다.

IBM Spectrum Conductor 의 멀티 테넌시 기능을 통해 Spark 의 여러 인스턴스를 도입하여 최적의 리소스 활용 및 확장 성과 성능을 구현하는 동시에 개별 Spark 구현에 연결된 리소스 사일로를 제거할 수 있습니다. 이 솔루션은 또한 Hadoop, MongoDB 또는 Cassandra 와 같은 애플리케이션 프레임워크와 Spark 의 통합을 용이하게 합니다. IBM Spectrum Conductor 는 최종 사용자가 기존 클러스터 또는 새 클러스터에 구축할 수 있도록 허가되고, 지원되는 소프트웨어 패키지입니다.

IBM Spectrum Conductor Deep Learning Impact 는 데이터 과학자의 일상을 가속화 및 단순화하도록 설계된 딥러닝 환경입니다. 모든 인프라에서 실행할 수 있으며 IBM 은 Spark 내에서 GPU 최적화를 개발하여 Spark 인스턴스 내부에서 계산을 가속화합니다. 경쟁 인프라보다 높은 CPU-GPU 대역폭을 감안할 때 PowerAI 는 이 GPU 최적화를 진정 확실한 방식으로 활용할 수 있습니다. 따라서 딥러닝 학습에 사용되는 동일한 PowerAI 클러스터를 준비 및 교육 일환으로 IBM Spectrum Conductor 에서 데이터 준비에 사용할 수 있습니다. IBM PowerAI Enterprise 로드맵에서 IBM Spectrum Conductor Deep Learning Impact 기능이 2018 년 2 분기에 PowerAI Enterprise 패키지에 통합될 예정입니다.

향상된 사용자 인터페이스

오늘날 대부분의 데이터 레이크는 데이터 과학자의 영역입니다. LOB(line of business)의 보고서 요청에는 전문 기술이 필요하며 생성까지 몇 주 또는 몇 달이 걸릴 수 있습니다. 오늘날의 데이터 레이크가 예전의 데이터 웨어하우스와 동일한 방식으로 발전할 것이라고 가정하는 것도 무리는 아닙니다. 20 년 전에는 데이터 웨어하우스에서 유용한 정보를 얻는 데 수개월이 걸렸습니다. 그러나 개발자가 데이터 웨어하우스에 액세스하여 해당 데이터를 활용하는 애플리케이션을 개발할 수 있게 해주는 도구가 등장했습니다. 다음으로, LOB 직원들이 Excel 플러그인을 통해 데이터 웨어하우스에 액세스하여 비즈니스 보고서를 생성하기 시작했습니다.

이와 유사한 "개방"이 오늘날의 데이터 레이크에서 등장할 것입니다. 오늘날의 엔터프라이즈 개발자들은 AI 를 통해 애플리케이션을 혁신할 것입니다. 이를 위해서는 데이터 레이크의 데이터 및 조직의 AI 기능에 액세스해야 합니다. LOB 사용자들이 많이 뒤쳐지지는 않을 것입니다. 결론은 데이터 레이크 내 중요한 데이터에 대한 액세스와 해당 데이터를 AI 에 활용할 수 있는 기능이 대중화될 것이라는 사실입니다. 그러나 이를 달성하기 위한 도구가 좀더 사용하기 쉽고 직관적으로 개선되어야 할 필요가 있습니다. 이를 위해 IBM 은 데이터 과학 경험을 대표하는 DSX Local 을 개발했습니다. DSX Local 은 데이터 레이크 데이터를 사용하는 ETL 부분을 단순화하고 적절한 딥러닝 프레임워크에 연결합니다. DSX Local 은 IBM Power 뿐만 아니라 다른 프로세서 아키텍처에서도 온-프레미스 및 클라우드 환경에서 작동합니다. DSX Local 은 IBM 기술과 통합된 RStudio, Spark, Jupyter 및 Zeppelin 노트북과 같은 데이터 과학 도구 모음을 제공하는 데이터 과학자 및 데이터 엔지니어를 위한 기본적인 온-프레미스 엔터프라이즈 솔루션입니다. 사용자 인터페이스가 최대한 직관적으로 설계되었으며 이 도구는 데이터 과학자 및 개발자 팀을 위한 협업 프로젝트 공간을 제공합니다.

여러 최신 데이터 플랫폼 지원

데이터베이스에서는 프리사이즈란 없습니다. 여전히 기존의 관계형 데이터베이스(예 : CRM 및 ERP)로 실행될 수 있는 많은 작업이 있지만, 점점 더 많은 과제에서 최신 스키마를 필요로 합니다. 개선된 데이터 레이크에서 AI 학습 및 추론을 실행하는 AI 인프라를 구축하고자 하는 기업은 SQL, 특히 NoSQL(예: IoT 또는 컨텐트 워크로드) 및 그래프(부정 행위 감지용) 이외의 많은 스키마를 수용해야 합니다.

또한 기업들이 데이터를 위한 새로운 플랫폼에 투자하는 경우 메모리 내 데이터베이스와 그래프 및 NoSQL 용 GPU 가속 데이터베이스와 같은 최신 아키텍처를 조사해야 합니다. 이러한 데이터 집약적인 스키마에는 많은 스레드가 필요하며, 스키마 및 데이터베이스에 따라 일부는 CPU 에서만 처리되고 일부는 메모리에서, 다른 프로세스는 GPU 에서 처리됩니다. 예를 들어 그래프 데이터베이스를 사용할 때 데이터베이스 전체가 메모리에 있어야 합니다.

여러 데이터 플랫폼을 지원하게 되면 AI 를 위한 강력한 플랫폼이 탄생될 것입니다. 데이터 레이크 이외에도 기업들은 Redis, MongoDB 및 EDB Postgres, 그래프 데이터베이스 Neo4j, 그리고 GPU 가속화된 분산 메모리 내 데이터베이스인 Kinetica 와 같은 새로운 등급의 가속화된 오픈 소스 데이터베이스를 구축할 수 있습니다. IBM 은 이러한 데이터베이스에서 POWER9 가 가진 확실한 성능 이점을 주장하며 가격 대비 성능 보장을 제공합니다. 이들 데이터베이스는 Power 의 I/O, 빠른 상호 연결, 내장 RAS 및 대용량 메모리를 활용합니다. 이러한 새로운 서비스 중 많은 부분이 메모리에 저장되므로 데이터가 전례 없는 수준으로 이동하면서 많은 양의 메모리, 고성능 및 유연한 I/O 액세스를 필요로 하게 됩니다.

대형 모델 지원

IBM 은 4 방향 멀티스레딩, 코어당 성능, I/O 기능 및 완전한 일관성을 제공하는 내장 NVLink 덕분에 Power System 이 이러한 최신 스키마를 실행하기 위한 이상적인 플랫폼이라고 믿습니다. NVLink 는 GPU 의 프로세스가 성능 저하 없이 시스템 메모리를 활용하도록 허용합니다. 이것은 12GB, 16GB 또는 최근 32GB(GPU 에 일반적으로 탑재되어 있는)의 최대 메모리를 극복해주는 중요한 기능입니다. 이 메모리는 AI 비전, 4K 비디오 또는 다중 레이어의 행렬이 있는 복잡한 AI 모델은 충분하지 않습니다. 결과적으로 기업들은 저해상도 이미지 사용, 고화질 비디오가 아닌 웹 스케일 작업, 원하는 수준보다 낮은 네트워크 개발 등 절충안을 만들어야 합니다.

따라서 PowerAI 의 새로운 릴리스와 함께 IBM 은 "대형 모델 지원"을 도입했습니다. PowerAI 에서는 전체 모델을 CPU 메모리에 로드할 수 있습니다. 예를 들어, 4 개의 GPU 시스템에서는 16GB 또는 32GB 모델의 인스턴스 4 개 대신 230GB 모델의 인스턴스가 4 개 있을 수 있습니다. 따라서 하위 세트가 아니라 전체 데이터 세트를 모델에 로드할 수 있고 문제를 보다 정확하게 해결할 수 있으며 비디오 파일이 웹 스케일보다 고해상도가 됩니다. 다른 시스템에서는 PCIe 카드가 GPU 를 시스템 메모리에 연결하므로 PCIe 의 제한된 대역폭 때문에 성능이 저하될 수 있습니다.

IBM Power 는 PCIe 3.0 보다 훨씬 높은 대역폭의 CAPI 가 탑재되어 있어서 매우 빠른 SSD 에 연결한 다음 메모리급 성능의 메모리로 사용될 수 있기 때문에 플래시 메모리의 공유까지 가능합니다. 따라서 데이터 레이크의 많은 부분을 SSD 의 메모리에 넣을 수 있습니다. IBM 은 2017 년 8 월 Caffe 릴리스의 일환으로 대형 모델 지원을 기술 시연했으며, 이제 다양한 프레임워크 제공업체에게 소스를 공개하기 시작했습니다. 첫 번째는 일본의 Preferred Networks 에서 제공하는 오픈 소스 딥러닝 Chainer 프레임워크입니다.

FUTURE OUTLOOK

IDC 는 AI 가 매우 빠르게 발전할 것이며, 대부분의 기업이 향후 12-24 개월 내에 AI 문화를 도입해야 하며 그렇지 않으면 경쟁에서 뒤쳐질 위험이 있다고 내다봤습니다. AI 는 빠르게 진화하고 있을 뿐 아니라 모든 워크로드를 데이터 센터 또는 클라우드에 투입하고 있습니다. 따라서 AI 에 대한 장기적인 전략적 접근이 필수적입니다. AI 가 비즈니스를 어떻게 향상시킬 수 있을까? AI 가 비즈니스를 어떻게 보호해줄 수 있을까? AI 가 어떻게 비즈니스 경쟁력을 높여줄 수 있을까? AI 가 어떻게 비즈니스의 효율성을 높여줄 수 있을까? 이러한 수십 가지 질문에 대답해야 합니다. 동시에 데이터 과학자 및 앱 개발자에 의한 AI 개발은 AI 에 대한 장기적인 인프라 계획과 함께 보조를 맞춰야 합니다.

기업이 지능형 서비스를 위해 데이터 레이크를 설계하거나 재설계할 때는 AI 를 어느 영역에서 구현해야 할지 아직 모른다고 가정해야 합니다. 실시간 서비스들을 연계시키는 동안 데이터 레이크에 걸쳐 학습하는 것에서부터 소셜 미디어와 관련이 있는 유추 모델을 실행하는 것까지, 수 많은 잠재적인 시나리오가 있습니다. 기업은 앞으로 12 개월, 24 개월, 36 개월 내에 AI 를 얼마나 빠르게 도입할 것인지 불분명하다는 가정 하에 성능 저하 없이 AI 로 모든 것을 수용할 수 있는 방식으로 시스템을 구축해야 합니다. 따라서 샌드박스와 유사하면서 기업이 다른 모든 AI 접근 방식으로 운영할 수 있도록 지원하는 플랫폼을 보유해야 합니다.

이러한 미래 지향적인 접근 방식의 일환으로, 기업은 향후 2~3 년 동안 AI 가 인프라에 미치게 될 수요가 분명히 증가할 것이라고 가정해야 합니다.

AI 모델은 더 복잡해질 것이며, 학습에 사용되는 데이터, AI 를 활용하는 애플리케이션의 수, 대부분의 고객에 의해 발생하는 추론에 따른 부하 또한 천문학적으로 증가할 것입니다.

얼리 어댑터들이 2~3 년 동안 인프라를 3 번 바꾸는 등의 상황을 피하기 위해서는 오늘날과 같은 극도의 데이터 집약적인 AI 워크로드를 처리할 수 있는 인프라 솔루션을 바로 지금뿐만 아니라 앞으로 3 년 내에 사용해야 합니다.

CHALLENGES/OPPORTUNITIES

기업

오늘날의 기업들이 가장 먼저 해결해야 할 과제는 AI 를 계획적이고 능률적인 방식으로 조직에 도입하는 것입니다. 몇 년 전에는 AI 에 대한 무계획적인 시험이 유행이었고 상당한 비용에도 불구하고 필수적인 것으로 받아들여졌습니다. 이제는 그러한 시나리오를 벗어날 수 있습니다. 현재는 기업을 위해 일관되고 효율적이며 유용한 AI 로드맵을 구현하기 위한 다양한 솔루션이 출시되어 있습니다. 따라서 AI 는 해결해야 할 과제에서 비즈니스를 위한 기회로 바뀌고 있습니다. 물론 여전히 장애물은 있지만(전문 AI 데이터 과학자의 관심을 제외하고) 데이터를 준비하고 적절한 인프라를 구축한다는 측면에서 볼 때 더 이상 기업이 시간을 낭비하지 않아도 됩니다.

기업의 또 다른 과제는 우리가 이기종 데이터 센터로 전환한다는 사실입니다. 수년간 표준 아키텍처 서버에서 워크로드를 실행한 후, 기업은 GPU, FPGA, ASIC 및 많은 코어 프로세서와 같은 가속기를 통합, 프로그래밍 및 확장하는 방법을 이해해야 할 뿐만 아니라 특정 작업을 보다 효율적으로 수행하는 다른 아키텍처 호스트 프로세서에 익숙해지기 시작해야 합니다. 하지만 불안해하지 않아도 됩니다. Linux, 가상화 및 컨테이너화 덕분에 특정 워크로드에서 성능이 향상되는 것을 제외하면 다른 프로세서는 거의 눈에 띄지 않습니다.

IBM:

IBM 의 경우 가장 큰 과제는 앞서 언급한 기업의 두 번째 과제와 밀접한 관련이 있습니다. 일부 워크로드가 Power System 과 같이 다른 프로세서가 탑재된 시스템에서 객관적으로 더 잘 실행되는 사실과 관련하여 기업이 느끼는 무력감을 극복하는 것이 획기적인 시장 성공을 향한 IBM Power Systems 의 여정에서 여전히 가장 큰 장애물입니다. "표준" 아키텍처의 개념이 IT 문화에 너무 깊숙이 박혀있어서 대부분의 사람들은 "표준"이 공통적인 설계에서 나왔다는 사실을 잊어 버립니다. 다른 말로 하자면, 표준은 무난한 수준을 의미하지 탁월한 수준을 의미하지는 않습니다.

따라서 AI 는 IBM Power System 에게 일생일대의 기회입니다. 이것은 Power System 을 위한 이상적인 워크로드이며, 거대한 시장 기회로 빠르게 진화하고 있습니다. 이것은 분명히 IBM 에게 유리한 기회이며 IBM 이 성공하기 위해 가장 중요한 메시지는 기본 설계가 얼마나 복잡하든 상관 없이 최대한 간단하게 유지하고, 최대한 경제적인 가격을 유지하고, 데이터 센터에 대한 핵심 인플루엔서인 개발자들을 대상으로 마케팅하라는 것입니다.

마지막으로, IBM 이 간과해서는 안 되는 한 가지 기회는 Watson 과의 브랜드 인지도입니다. AI 인프라와 Watson 을 실제로 또는 개념적으로 연결하는 것은 별로 어려운 일이 아니며 IBM Power Systems 에 유리하게 작용할 수 있습니다.

CONCLUSION

AI 를 비즈니스의 운영, 제품 및 서비스에 도입하기 위해 여전히 AI 인프라를 시험하고 있는 기업들은 그 부담을 덜 수 있습니다. 불과 12 개월 전과는 달리 AI 추론뿐만 아니라 딥러닝이 어떤 종류의 인프라를 필요로 하는지에 대한 인식이 높아지고 있습니다. 공급업체들은 AI 애플리케이션에 대한 장애물을 줄여주는 AI 용 하드웨어, 소프트웨어 및 서비스 패키지를 함께 제공하고 있습니다. IBM 은 특히 Power System 프로세서 및 I/O 장점뿐만 아니라 2 세대 NVLink 를 통해 새로운 POWER9 프로세서와 NVIDIA V100 GPU 를 통합하기 위해 구현한 맞춤형 개발을 활용하여 PowerAI 시스템을 위한 토대가 되는 매우 적합한 서버를 제공했습니다. 이 회사는 또한 기업이 모든 종류의 데이터를 빠르고 포괄적으로 준비할 수 있도록 데이터 레이크를 업그레이드하는 작업에 어려움을 겪고 있음을 목격했습니다. 이것은 종종 간과되는 중요한 기능입니다. IDC 는 AI 가 수년 동안 데이터 센터의 표준이었던 인프라와는 다른 인프라를 필요로 한다고 믿고 있으며, 기업들도 이를 분명히 인식하고 있습니다. 그들은 미래를 내다보면서 다른 호스트 프로세서, 다양한 가속화 기술 및 대형 메모리 용량을 채택하고 있습니다.

IDC 정보

IDC(International Data Corporation)는 정보 기술, 통신 및 소비자 기술 시장을 위한 세계적인 시장 정보, 컨설팅 서비스 및 이벤트 공급업체입니다. IDC 는 IT 전문가, 기업 임원 및 투자 커뮤니티가 기술 구입 및 비즈니스 전략에 관하여 사실에 근거한 결정을 내릴 수 있도록 지원합니다. 1,100 명 이상의 IDC 연구원들이 전 세계 110 여 개국의 기술 및 산업 기회 및 동향에 대해 국제적, 지역적인 전문 지식을 제공하고 있습니다. IDC 는 50 년 동안 고객이 주요 비즈니스 목표를 달성할 수 있도록 전략적인 통찰력을 제공해오고 있습니다. IDC 는 세계 최고의 기술 미디어, 리서치 및 이벤트 회사인 IDG 의 자회사입니다.

국제 본부

5 Speen Street
Framingham, MA 01701
USA
508.872.8200
트위터: @IDC
idc-community.com
www.idc.com