IDC TECHNOLOGY SPOTLIGHT

赞助方：IBM

在数字时代，人工智能 (AI) 已成为商务智能的核心。尽管 AI 的优势与前景我们都十分看好，但组织若要实现从以往概念验证到 AI 生产化与规模化的成功过渡，还面临着诸多挑战。

借助 AI 优化的基础架构加速 AI 部署并实现其运营化

2018 年 6 月

作者：Ritu Jyoti（副总裁）

引言

概览

关键统计数据

据 IDC 预测，到 2019 年，40% 的数字转型计划将会使用 AI 服务；到 2021 年，75% 的商业性企业应用将使用 AI，超过 90% 的消费者将会与客户支持机器人进行互动，超过 50% 的新工业机器人将会利用 AI。

数字转型 (DX) 正在向宏观经济规模迈进。基于人工智能 (AI)、机器学习 (ML) 和持续深度学习 (DL) 的智能应用正在形成下一波的技术潮流，旨在推动消费者与企业工作、学习与娱乐方式的转型。尽管数据是新数字经济的核心所在，但它同样关乎您如何感知环境，从边缘到核心再到云端来管理数据，以近乎实时的方式分析数据，从数据中学习，然后基于数据采取行动，最终实现成效。物联网 (IoT)、移动设备、大数据、AI、ML 和 DL 将会结合使用，以持续感知环境并统一地从环境中学习。组织若想脱颖而出，其中的关键在于如何利用这些技术交付有意义的增值预测结果与行动，进而改善工业流程、医疗保健水平、实验性互动水平，或任何其他类型的企业决策。AI 业务的目标会兼顾战术型目标和战略性目标，从改善运营效率到提升竞争优势，从实现现有产品收入的最大化到开辟新的数字收入流，都属于 AI 业务目标的权衡范围。

尽管 AI 技术早在数十年前就已经出现，但由于数据无所不在、云计算的可扩展性、AI 加速器的可用性，以及 ML 和 DL 算法日益高级化，AI 已经成为商务智能的核心。据 IDC 预测，到 2019 年，40% 的数字转型计划将会使用 AI 服务；到 2021 年，75% 的商业性企业应用将使用 AI，超过 90% 的消费者将会与客户支持机器人进行互动，超过 50% 的新工业机器人将会利用 AI。

不过，在几乎一半的 DX 计划中，AI 的关键作用将会给 IT 领域的技能需求带来新的压力。据 IDC 预测，到 2020 年，85% 基于运营的新技术招聘岗位将会根据分析和 AI 技能来进行筛选，因为这些技能有助于数据驱动型 DX 项目的开发。与此同时，CIO 必须构建并持续提升集成式的企业数字平台，以便为新的运营与货币化模式提供支持。IT 部门将需要成为企业中最佳且首个 AI 用例环境之一，包括 AI 开发、数据管理和网络安全等等。

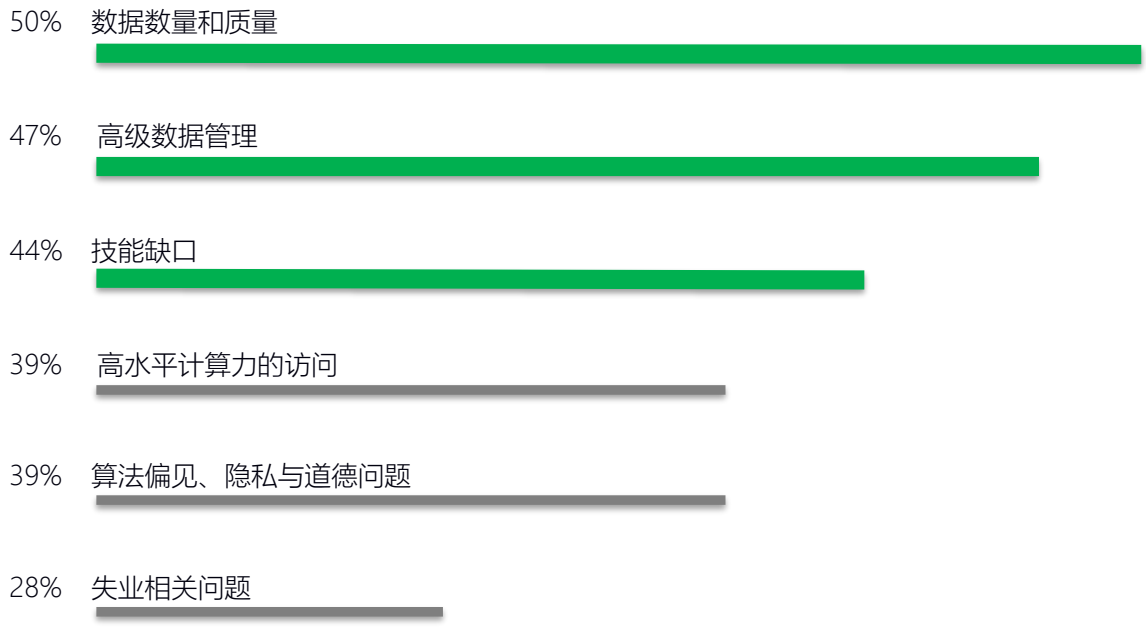
AI 模型与工作负载部署：挑战与需求

AI 正在改变着数字时代业务流程的执行方式。尽管 AI 的优势与前景我们都十分看好，但 AI 模型和工作负载的部署绝非易事。尽管 AI 炒的沸沸扬扬，但大多数组织仍旧在概念验证 (POC) 阶段挣扎，而仅有少数的组织实现了 AI 的生产化。

有待解决的问题在于，ML 和 DL 算法需要大量的训练数据（一般是传统分析所用数据量的 8 到 10 倍），而且 AI 的效率非常依赖于高质量、多样化且动态的数据输入。 从以往来看，数据分析以大文件、顺序访问及批数据为核心。但现代数据在来源和特性都与以往有所不同。就目前而言，数据的种类非常繁杂，有小文件和大文件，也有结构化数据、半结构化数据和非结构化数据。数据访问也包括随机访问和顺序访问。到 2025 年，超过四分之一的全球数据将会具有实时的特性，而实时物联网数据在其中的占比将会达到 95% 以上。此外，数据会日益在内部环境、主机代管及公有云环境中分布。

2018 年 1 月，IDC 对美国和加拿大的 405 位 IT 及数据专业人士进行了调研，受访对象都是曾经成功完成过 AI 项目、能够控制或影响预算，且负责 AI 工作负载运行平台的评估或架构设计的人士。此项调研的目的是确定组织在 AI 支持技术的使用和管理方面所采用的方式，并识别组织运行认知/ML/AI 工作负载所用的基础架构、技术的部署位置，以及相关的挑战和需求。如图 1 所示，受访者识别了他们在解决海量数据及相关质量与管理问题方面所遇到的关键 AI 部署挑战。

图 1: *AI 工作负载部署方面的挑战*



Source:IDC 认知、ML 与 AI 工作负载基础架构市场调研，2018 年 1 月；受访对象数量=405，1,000 多名员工（美国）；500 多名员工（加拿大）

数据质量不过关会直接导致模型构建出现偏差或不准确。对于海量的动态、多样化、分布式数据集，如何确保质量是一项非常困难的任务，因为开发人员很难了解和预测所有的适当检查项和验证项，而且相应的编码也非常困难。为了解决这些挑战，企业希望部署一款自主性的数据质量与验证解决方案。此类解决方案应能够自动学习数据的预期行为，在无需编码的情况下构建数千个数据验证/检查项，随时时间的推移不断更新和维护检查项，同时消除预期和非预期的数据质量错误，最终提升数据的可信性和可用性。

需要相关专业人员（包括 AI 工程师和数据科学家）为 AI 依赖型 DX 计划不断增多的细分领域提供支持。不过，IT 部门也面临着专业人员缺乏、技能缺口等问题（见图 1）。举例来说，在模型的构建/优化和训练方面，需要一个较高的学习曲线，而大多数数据科学家都不具备这一技能集。

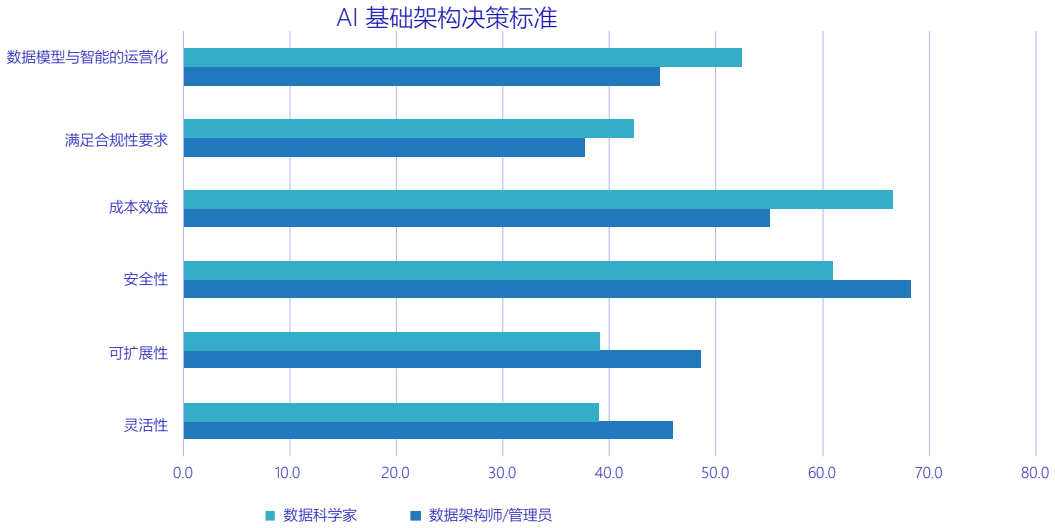
需要相关专业人员（包括 AI 工程师和数据科学家）为 AI 依赖型 DX 计划不断增多的细分领域提供支持。

在企业采用 AI 的过程中，下列用户会以错综复杂的方式牵涉其中：

- 数据科学家是大数据专家，也是模型的构建者。具体来说，他们首先会获取大量的数据点（非结构化数据和结构化数据），并使用他们在数学、统计及编程方面的娴熟技能清理、整理和组织数据。之后，他们会运用所有的分析能力（包括行业知识、情境理解、对现有假设事项的质疑）来发现适于解决业务挑战的潜在解决方案。对于数据科学家而言，可能需要完成的事项包括：
  - 从多个内部和外部来源中提取、清理和整理海量的数据
  - 使用高级的分析程序、ML 与统计方法准备数据，以供在预测性及规范性建模过程中使用
  - 探索并检查数据，以确定潜在的弱点、趋势或机会
  - 投资或构建解决问题所需的新算法和模型，并构建新的工作自动化工具
  - 针对最紧迫的挑战规模化地训练、优化并部署数据驱动型 AI 模型
  - 维持 AI 模型的准确性
  - 通过高效的数据可视化工具及报告向管理层和 IT 部门报告预测结果和发现结果
- 数据工程师/管理员负责构建海量的数据库。具体来说，他们负责架构的开发、构建、测试和维护，例如数据库及大规模数据处理系统。一旦持续进入这些大型过滤信息"池"的连续管道安装完毕，数据科学家就可以在其中放入相关的数据集，以便进行分析。
- 数据/IT 架构师负责将 AI 框架集成到基础架构部署战略之中，以及提供确保环境可扩展性、敏捷性与灵活性所需的支持。

当这些用户角色被 IDC 问及他们在选择 AI 解决方案时的最主要决策标准时，他们认为决策标准主要包括安全性、成本效益、数据模型/智能的运营化（构建、调优、优化、训练、部署及推理），详见图 2 所示。

图 2：*AI 基础架构/解决方案决策标准*



来源：IDC，2018 年

考虑到上述各个因素，成功部署 AI 的关键在于：

» 数据科学家的效率

构建、测试、优化、训练、推理并维持模型的准确性是 AI 工作流不可或缺的部分。这些神经网络模型很难构建。为了构建、测试和部署大规模的 ML/DL 模型，数据科学家通常会应用各种工具（如 RStudio 和 Spark）、开源框架（如 Tensorflow 和 Caffe），以及编程软件（如 R 或 Python）。不过，开源框架的选择与安装，以及建模流程的初始化都是非常繁重的事务，需要耗费数周或数月才能完成。构建并优化模型需要手动测试数千个超级参数组合。

在一些用例中，模型的训练可能要耗费数周或数月才能完成；举例来说，某家医疗保健组织耗费了 1 年的时间才完成了检测早期癌症所需医疗模型的构建和训练。

由于训练具有迭代性质，因此需要在数小时、数天或数周内完成数百万项任务。就目前而言，该流程需要完成相应的任务才能确定训练任务是否成功，这意味着组织或许在运行一周的训练任务之后，才发现该项任务并未奏效。如果某个服务器/GPU 出现故障，也必须重新开始训练任务，这意味着每个环节都得重新开始。

数据科学家希望提升清理多个来源的数据的效率并实现其自动化，同时降低干扰因素。他们还需要协助构建、调优和选择模型所需的正确功能，包括简化确定超级参数设置所需的流程。

数据科学家希望提升迭代与循环流程的速度和敏捷性。他们还希望在训练阶段基础架构资源能够实现一流性能和灵活扩展，以缩短训练时间，同时希望获得相应的工具，使其能够在分布式的集群环境中轻松运行训练任务。

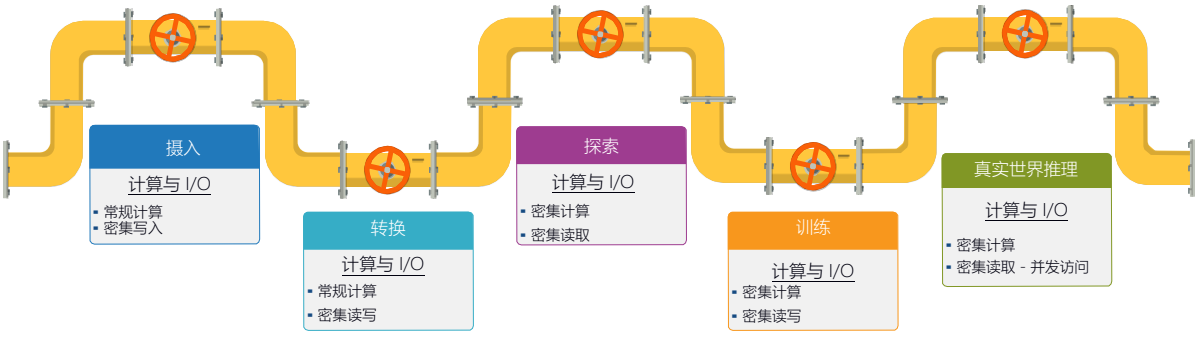
此外，还需要确保敏捷的工作负载管理，尤其是要能够更高效地运行任务，以实现资源利用率的最大化，确保性能和效率，同时实现任务执行情况的可视化，并在任务不奏效时，立即停止任务的执行。此外，还需要相应的软件，确保在之前的任务失败时，能够自动转至不同的服务器/GPU。

» 经优化的基础架构和高效的数据管理

通过图 3 所示的 AI 工作负载数据管道，我们可以看出，从摄入到真实世界推理，应用概要、计算及 I/O 概要会有所变化。ML 和 DL 需要大量的训练数据。训练和推理都属于计算密集型环节，需要高性能来确保快速执行。AI 应用的采用更是将硬件部署推到了极限，需要数千个 GPU 或数千个 CPU 服务器。AI 和 DL 需要一种主要基于 GPU 的新型加速基础架构。对于神经网络模型训练所需的线性数学计算而言，相比非加速系统的集群，通过 GPU 配置的单个系统则更加强大。

不过，并非所有的 AI 部署都是一样的。组织应根据 AI 部署的性能、运行环境、所需的技能集、成本及能耗需求，探索异构处理架构（如 GPU、FPGA、ASIC 或多核处理器）。

图 3：*AI 工作负载的数据管道*



来源：IDC，2018 年

我们还知道，并行计算需要并行存储。尽管训练阶段需要大容量的数据存储，但推理阶段对存储的需求较小。推理模型通常会存储在 DevOps 风格的资料库中，这种资料库的优势在于超低延迟的访问。尽管在执行模型基于数据开发完毕且工作负载转移到推理阶段之后就是训练阶段，但一旦出现新数据或数据有修改，往往就需要对模型进行重新训练。在一些情况下，由于应用的实时性质，可能需要近乎持续的重新训练和模型更新。由于数据来源的不断增加及相关洞察力发挥作用，组织还可以随着时间的推移通过模型的重新训练而受益。

如果数据在管道中流动不畅，就会影响生产效率，而组织将需要不断增加投入和资源来管理管道。数据架构师和工程师所面临的挑战在于如何在保持成本不超额的情况下确保 AI 工作负载的敏捷性、灵活性、可扩展性、性能、安全性与合规性要求。

显而易见，凭借传统的基础架构，企业根本无法支持 AI 等前沿工具，因为这种基础架构无法满足可扩展性、弹性、计算能力、性能及数据管理方面的需求。现在，组织已经开始使用不同的基础架构解决方案与方法来支持 AI 数据管道，但往往会导致产生数据孤岛。一些解决方案或方法会为管道创建数据的重复副本，其目的在于确保不会影响应用的稳定性。相反，组织需要采用具有动态适应性、可扩展性和智能化的基础架构（即能够自配置、自优化和自治愈的基础架构）。这种基础架构可根据数据格式和访问的变化进行调优，而且它可以处理和分析海量数据。它还能够确保速度，以支持更快的计算和决策，管理风险并降低 AI 部署的整体成本。

» 企业准备情况

企业往往都比较关注将新兴技术和框架纳入到企业环境所带来的影响。他们希望确保在安全、可靠性、支持及其他标准方面的企业就绪性，如图 2 所示。大多数可用的 AI/ML/DL 框架、工具包和应用不会实施安全策略，因此只能在未互联的实验及实验室实施项目中使用。此外，大多数公司都选择投资单独的集群来运行 AI，而这种集群不仅成本高而且效率低下。DIY 构建的系统的另一个挑战在于难以从多个供应商获得企业级的支持。

考虑 IBM 在 AI/ML/DL 工作负载方面的产品

IBM 的战略是让 AI/ML/DL 更具可访问性和可执行性。通过结合使用 IBM Power AI、IBM Spectrum Conductor、Deep Learning Impact、IBM Spectrum Scale 软件，以及 IBM POWER Systems 和 IBM Elastic Storage Server，组织便可快速部署并优化面向 AI 工作负载的支持平台，同时确保高性能。这种方法能够减少从开源拉取解决方案带来的挑战。IBM 解决方案的所有组件都来来自于公司，且由公司提供完全支持（包括 1-3 级支持）。就 IBM 的所有组件而言，如果需要支持，即可使用相同的单个联系点，而且所有的案例均由 IBM 所有并负责管理。此外，公司还会管理系统的所有软件补丁和更新。借助该解决方案，组织可以简化开发体验，减少 AI 模型训练所需的时间。该解决方案还支持生产性的概念验证，而且允许扩展为多租户系统，在这类系统中，运行不同环境和框架的多个数据科学家可以无缝地共享通用的资源集，确保组织可根据需求进行扩展，同时将该平台集成到现有的 IT 基础架构之中。

IBM 可提供最具综合性的内 AI 解决方案堆栈及相关工具和软件，这些解决方案面向 AI 部署中涉及的所有关键角色，包括数据科学家。与竞争对手不同的是，IBM 拥有多样化且不断发展的企业客户群，其中包括金融组织，此类组织在生产环境中运行他们的 AI 部署项目时可完全确保系统的安全与数据隐私。

图 4 所示的 IBM 参考架构由以下软件组件和基础机构堆栈构成：

» IBM PowerAI 是一个软件发布包，其中包含有模型训练所需的主要开源 DL 框架，如 TensorFlow、Caffe 及相关的资料库。这些框架可通过优化充分利用 IBM Power Systems 服务器的 CPU 和 GPU 之间的 NVLink 互联，进而实现一流的吞吐量和性能。

» IBM Spectrum Conductor 是一款高度可用的多租户应用，旨在构建一个共享的企业级环境，以便部署和管理现代化的计算框架与服务，例如 Spark、Anaconda、TensorFlow、Caffe、MongoDB 和 Cassandra 等。数据转化和准备涉及到许多手动密集的步骤：识别和连接数据来源、提取数据到加载服务器、使用工具和脚本来进行数据操作（如移除无关的要素、对大型镜像进行细分使其与 GPU 内存相匹配）。IBM Spectrum Conductor 可提供一个分布式 Apache Spark 和应用环境，因此有助于实现这些任务的自动化及并行运行，进而提升流程速度并实现其自动化。它能够建立并维持与存储资源的连接，捕获数据格式信息，进而通过端到端 DL 流程实现快速迭代。Spectrum Conductor 还可实现集中管理与监控，同时它能够框架和应用在其情境内的运行实施相应的端到端安全，进而构建一个安全的 AI/ML/DL 环境。安全实施所围绕的要点包括：

  o 认证：提供针对 Kerberos、SiteMinder、AD/LDAP 及 OS 认证的支持，就如同面向 HDFS 的 Kerberos 认证一样。

  o 授权：该功能可实现精细的访问控制、ACL/基于角色的控制 (RBAC)、Spark 二进制文件循环、Notebook 更新、部署、资源规划、报告、监控、日志检索和执行。

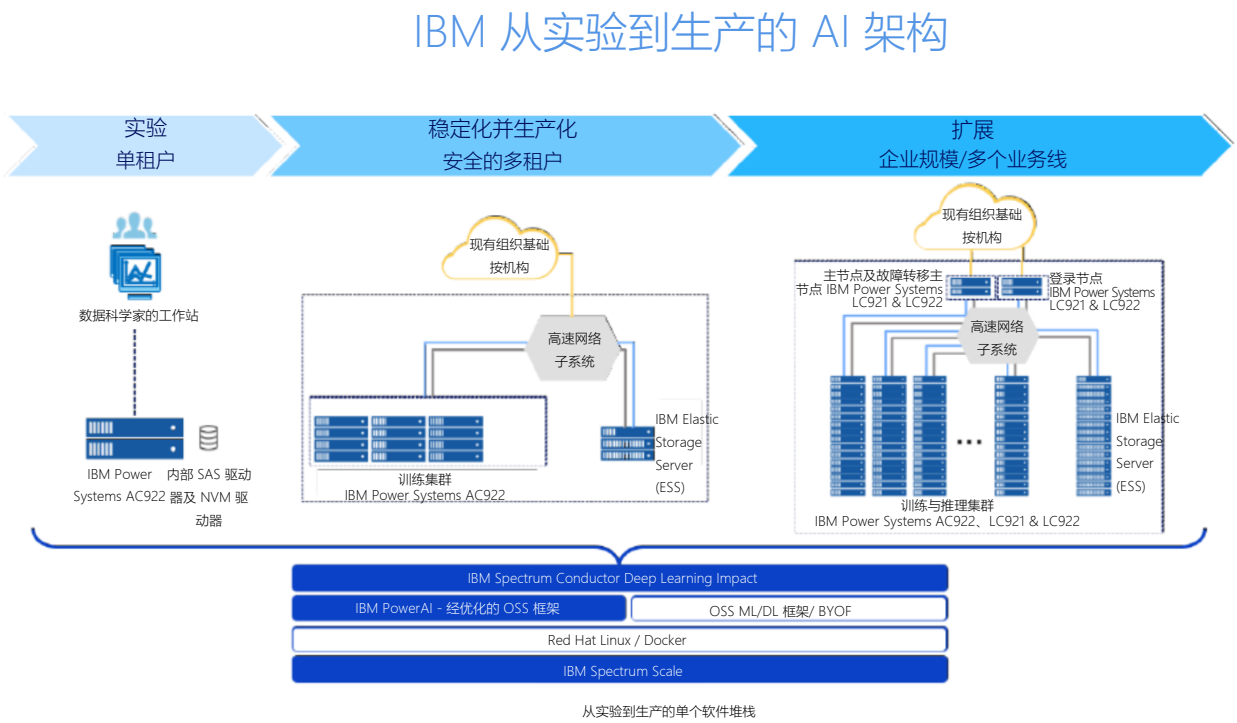
  o 身份模拟：不同的租户均可定义生产执行用户。

  o 加密：支持所有后台程序之间的 SSL 和认证。

» IBM Spectrum Conductor Deep Learning Impact 可构建带有端到端工作流的 DL 环境，使得数据科学家能够专注于模型的训练、调优及生产部署。AI/MI/DL 流程的构建/训练阶段需要大量的计算而且迭代性非常强，几乎不需要模型调优和优化方面的专业知识。在这个阶段中，组织在 DL 及数据科学技能方面的差距表现的最为明显。PowerAI 和 Spectrum Conductor Deep Learning Impact 采用认知算法来辅助和优化超级参数，因此有助于模型的选择与创建。它们还支持弹性训练，有助于运行时过程中资源的灵活分配，而这种分配可实现资源的动态共享，对任务进行优先排序，并确保故障情况下的弹性。通过运行时训练可视化，数据科学家可以查看模型训练的进展，一旦模型的训练结果出现错误，就可以理解停止训练，这有助于更快速地交付更准确的神经模型。

» IBM Spectrum Scale 是一个企业级的并行文件系统，可通过存储加密实现弹性、可扩展性、可控性及安全性。IBM Spectrum Scale 能够交付可扩展的功能与性能，轻松处理数据分析、内容库和技术计算工作负载。存储管理员可将闪存、磁盘、云和磁带存储合并到一个统一的系统之中，相比传统的存储方法，该系统不仅性能更好，而且成本更低。

图 4：*IBM 参考架构*



来源：IBM

» 面向解决方案的基础架构由以下几个要素构成：

o 计算：IBM Power Servers 采用 CPU:GPU NVLink 连接，相比 x86 服务器，可交付更高的 I/O 带宽。它还可支持更大的系统内存。

» IBM Power System AC922 可支持 2-6 个 NVIDIA Tesla V100 GPU，而且通过 NVLink，它可提供 100 GB/秒（风冷）的或 150 GB/秒（水冷）的 CPU:GPU 带宽。该系统最高可支持 2TB 的总内存。

» IBM Power System S822LC for HPC 可支持 2-4 个 NVIDIA Tesla P100 GPU，而 NVLink GPU 可提供 64 GBps 的 CPU:GPU 带宽。该系统最高可支持 1TB 的总内存。

o 存储：IBM ESS 结合采用了 IBM Spectrum Scale 软件与基于 IBM POWER8 处理器的I/O 密集型服务器及双端口存储机箱。IBM Spectrum Scale 是 IBM ESS 的核心并行文件系统。IBM Spectrum Scale 可随着时间的推移扩展系统吞吐量，同时提供单个域名。

挑战与机遇

IDC 的调查结果显示，公有云在 AI 模型和工作负载的部署方面处于主导地位，私有云部署紧随其后。组织就在公有云还是在内部系统上运行 AI 管道进行决策时，通常都会考虑数据重力，而数据管道是指数据目前所在的位置或可能在未来存储的位置。此外，能否轻松访问计算资源与应用，以及所需的功能探索及部署速度也是非常重要的考虑因素。公有云服务可为数据科学家及 IT 专业人员提供相应的基础架构及工具，以便他们训练 AI 模型、试验新算法，或以轻松且敏捷的方式学习新技能和技术。从长期来看，组织可能会在内部部署模型训练解决方案，因为他们希望确保自己的 IP 与洞察力。

边缘部署仍旧属于初期阶段，因为边缘缺乏资源，而且在某些情况下，必须在边缘进行处理。举例来说，车间内某个关键机器上的温度传感器会将与即将发生的故障相关的读数发送到边缘基础架构，之后会对该读数进行快速分析，再派遣相应的技术员及时修理机器，以避免成本高昂的停机。

IDC 认为，IBM 的 AI 参考架构是一个经优化的软件与硬件堆栈。它由经测试的、支持的且现成的开源框架、高吞吐量 GPU 及一流的存储组件构成，而且其中的存储组件兼具智能化、可扩展、安全、富于元数据、云集成、多协议、高性能且高效的特点。正如前文所说，该解决方案能够降低 AI 部署的复杂性，可帮助组织提升生产效率，降低获取和支持成本，并加速 AI 的采用进程。该解决方案的潜在提升机会包括：

» 在不断扩展的混合多云部署中实现与多个公有云服务的无缝集成，例如 Amazon Web Services、Google Cloud Platform、Microsoft Azure 及 IBM Cloud。

» 支持外形更小的边缘基础架构，以便暂时性地存储数据，以进行边缘推理。

» 支持异构处理架构（如 GPU、FPGA、ASIC 或 Manycore 处理器），有助于组织根据 AI 部署的性能、运行环境、所需的技能集、成本及能耗需求灵活选择相应的加速技术。

结论

对于全球各地的企业而言，通过运行将 AI/ML/DL 算法融入其中的应用来实现一流的业务成效至为关键，而且也是确保 DX 投入及用例所必需的。为了帮助组织加速实现 AI 驱动型业务成效并克服部署障碍，IDC 给出了以下指导意见：

» 专注于业务成效，严格执行项目时间表，并根据即时收入及成本影响对项目进行优先排序。

» 寻求相应的软件工具，以简化并自动化数据准备流程，加速 AI 模型的迭代构建、训练和部署，最终提升业务成效。

» 寻求具有动态适应性、简单、灵活、安全、成本高效且具有弹性的基础架构，以实现大容量、高吞吐量及低延迟，最终确保高性能的训练及推理体验。

» 采用智能基础架构并利用此类基础架构实现预测性分析、获取宝贵洞察力，然后在确保了可信性与数据质量之后逐步实现任务自动化。

分析师简介：

Ritu Jyoti（副总裁）

Ritu Jyoti 是 IDC 系统基础架构项目副总裁，领导着 IDC 的企业存储、服务器及基础架构软件团队，负责的领域包括研究产品及季度性跟踪服务，以及咨询服务与项目。Ms. Jyoti 的核心研究领域包括认知/人工智能、大数据、分析工作负载服务，旨在了解新机器学习、Hadoop、NoSQL 数据库及分析技术对基础架构软件和硬件市场及数字转型（IT 转型数据基础架构战略）的影响。

IDC Custom Solutions

本出版物由 IDC Custom Solutions 编制。本出版物中的意见、分析和研究结果来自 IDC 单独进行和公布的更为详细的调研和分析，但注明特定供应商赞助的情况除外。IDC Custom Solutions 会以多种格式提供 IDC 内容，以便由各个公司进行分布。IDC 内容的发布许可并不意味着对被许可方或其意见的认可。

IDC Corporate USA

5 Speen Street
Framingham, MA 01701,
USA

电话：508.872.8200

传真：508.935.4015

Twitter @IDC

idc-insights-community.com

www.idc.com