白皮书

重新思考基础架构，拥抱企业人工智能

作者：IBM

Peter Rutten

2018 年 6 月

IDC 观点

IDC 坚定地认为，一种架构控制数据中心所有计算的同构计算时代已然结束。随着越来越多的企业开始推行人工智能（AI）计划，这一事实变得愈发明显。许多企业都处于采用人工智能的试验阶段，少数已做好投产准备，但所有企业都在以异乎寻常的速度更替基础架构选项，以便运行他们新开发的人工智能应用和服务。

这种不断革新基础架构的做法的主要原因是，数据中心中用于处理大量工作负载的大多数标准基础架构都不太适合人工智能的极度数据密集特性。一方面，标准服务器在性能和 I/O 方面不具备支持深度学习 (DL) 的能力；另一方面，企业不具备人工智能模型开发温床 - 数据湖来执行这项关键任务。数据湖由基于传统模式的速度缓慢的单一文化组成，这些模式需要数周甚至数月时间才能为人工智能建模做好准备。企业认为这些数据湖无关紧要，但是一旦开始基于这些数据湖开发人工智能，它们将变得相当重要。

因此，人工智能已经成为讲述数据中心内新兴处理器多元性不断演变的故事这场戏剧中的主角 — 这种多元性不仅体现在用于处理特定工作负载的 GPU、FPGA、众核处理器和 ASIC 不断增加，还体现在迁移到其他主机处理器，以便更好地连接主机和加速器上。因为加速器不仅可以缓解很多性能延迟问题，还能与主机处理器相互作用，为人工智能之类的工作负载提供真正卓越的性能。

本白皮书探讨了这些挑战，并研究了 IBM 建议如何帮助企业战胜挑战。

市场概况

我们已经不再惊呼“AI 即将到来”，而是见证了“AI 就在身边”。人工智能是一种颠覆性的变革，不断迫使企业中的 IT 人员以及首席信息官、首席技术官和首席执行官重塑思考自身业务，还让他们开始自问：我们有办法利用人工智能吗？我们如何利用人工智能？我们怎样才能实现目标？虽然企业都关注人工智能最为重要的 5–10% 的作用，包括深度学习和训练，但却不太了解如何为人工智能革命做好准备，也就是说，企业不清楚人工智能将如何影响企业管理数据的方式，如何影响他们的基础架构。

存在大量的数据需要处理。这种数据洪流来自于多个日趋庞大的源头，包括互联汽车、可穿戴健康设备、

联网机器和传感器设备。这场大数据革命在某种程度上已经取得成功，因为当今许多企业都建立了数据湖。但是，企业还未弄明白的是，如何从数据湖中获取价值。企业的业务部门（LOB）可能要等待数周甚至数月时间，才能从数据研究员那里获取来自数据湖的有用信息。

当然，当今的数据湖提供了一些洞察，但是这些数据湖很难发挥作用，并且它们也的确不是构建人工智能应用的最佳基础。因此，人工智能革命要求人们重新审视大数据，这一次需要做的更为出色，以便支持全新的人工智能服务。数据洪流推动企业不断围绕机器学习（尤其深度学习）培养各种能力，以便创建和训练能够识别模式和发现洞察的智能系统。当今最重要的数据密集型工作负载就是大规模深度学习，那么问题是：企业需要选择什么样的服务器和软件平台才能全面采用人工智能?

敏锐的数据直觉、丰富的专业知识和强大的数据价值提取能力，是企业变革运营模式和与客户的互动方式的有力支持。人工智能是一种有效方式，能够真正了解深深隐藏在公司作为一家企业或一个组织所处的经营环境的复杂性。人工智能，尤其是深度学习，是传统数据分析加快的结果，能够提供大量的新价值。但是深度学习也带来了诸多挑战。它是一种快速演变的技术集合，这些技术创新速度很快，尤其开源技术，让人很难跟上它们的发展步伐，并协调各种部分高效工作。

挑战在于为深度学习选择合适的基础架构、选择合适的软件（该软件不但要易于使用，支持快速部署，还要能与硬件无缝集成，并获得供应商的全面支持）以及准备数据。

后一项挑战（为深度学习准备数据）显然是目前为止人工智能道路上的最大阻碍。企业发现，当他们开始进行深度学习训练，开始围绕训练优化基础架构时，已经在数据准备上花费了数月时间。构建人工智能流程时所用的大部分时间，都是在转换数据，以及将数据从 A 点迁移到 B 点，以便利用数据。更为复杂的是，真正有效的人工智能应用能够从多个来源提取数据（包括内部数据、流数据等），并对数据进行有意义的组合。

IBM 人工智能基础架构方法

人工智能需要利用现有数据存储和分析，这些数据存储和分析需要连接到深度学习工作流中。了解这一点后，IBM 开始致力于关注那些导致企业花费过多时间来构建人工智能环境的方面，其中最主要的就是数据准备环节。

IBM 着手简化为人工智能应用准备数据的过程，以及为这些人工智能应用构建连接和整合各种数据源的过程。用于人工智能的数据源包括各种类型：有些在企业的控制之下（例如，客户偏好），有些取自于现有的客户行为模式，有些则来自于外部，比如流数据（例如，社交媒体行为）。

IBM 认为，要想支持人工智能基础架构，为人工智能做好准备，企业就需要重新思考自己管理数据的方式。为帮助企业实现这种转变，IBM 推出了名为“人工智能基础架构”的概念，这个概念由一个用于数据管道、服务和人工智能的单一平台组成。本质上，它是一个端到端的服务器、存储器和软件平台，专为支持企业独特的人工智能革命旅程而设计。

这个平台的基础是一个现代的数据湖，具有拥有改进的存储功能、企业级 Hadoop 和 Spark 以及增强的数据管理能力。该基础支持包含多种模式（RDBMS、NoSQL 和图形）、多种架构（CPU、GPU 加速型和内存中）且拥有不折不扣的系统性能的数据平台。总之，更加完善的数据湖与更加卓越的数据平台二者结合可以实现灵活的 IT 环境，其中包含动态存储、动态 I/O 和动态内存，以及能够交付提供实时洞察的各种服务，比如 CRM、物联网、欺诈检测等。然后，这些构件块会成为深度学习训练的阵地，用于为现有应用注入人工智能，并开发人工智能驱动的全新应用。

人工智能基础架构的核心：IBM Power System AC922 和 IBM PowerAI

IBM Power System AC922

人工智能基础架构的核心是 IBM 于 2017 年 12 月推出的最新加速的 POWER9 系统。该系统被称为 IBM Power System AC922，专为人工智能工作负载而设计。Power AC922 是一个双插槽服务器，具有先进的 I/O 接口，以及为人工智能加速而优化和改进的硬件和软件，不仅性能强悍，还是 CORAL Summit 超级计算机的骨干。该系统包含各种极其快速的 I/O 架构，包括 PCIe Gen 4、CAPI 2.0、OpenCAPI 和 NVLink，所有这些架构都非常适合处理数据相当密集的工作负载。

该系统配备两个 POWER9 CPU，每个都具有四路多线程，支持主机 CPU 并行处理；搭载 NVLink 的 2-6 个 NVIDIA Tesla V100 GPU，可在 CPU 和 GPU 之间交付极致性能，支持高性能计算、深度学习和人工智能工作负载。NVLink 可以集成到处理器中（称为“NVLink 2”），这可以进一步提高整个系统的性能。该系统提供完整的内存一致性，使加速型应用能够利用系统内存作为 GPU 内存，攻克了 GPU 内存限制为 16 GB 或 32 GB 的难题。Power AC922 配置搭载 16 到 44 个核心，可以部署在任何系统上。

IBM PowerAI

IBM PowerAI 被 IBM 称为“用于分布式深度学习的企业产品”。PowerAI 基于 IBM Power S822LC for HPC 或 IBM Power AC922 而构建，带有开源深度学习框架和工具，比如 Caffe、Torch、TensorFlow、Theano 和 Chainer。还包含各种支持库，比如 DIGITS、OpenBLAS、分布式框架、Bazel 和 NCCL。该硬件/软件包旨在通过为数据研究员提供企业级、受支持版本的关键开源软件、高性能硬件和集成工具，提高他们的工作效率，从而为企业提供一种快速开始深度学习的方法。

IBM 希望通过 PowerAI 帮助降低构建顺利运行的人工智能基础架构的复杂性，确保轻松部署经过优化的深度学习训练框架，并且提供简单快捷的体验，这样数据科学家就不会将时间浪费在调试 TensorFlow 软件包上。最终，企业只需大约 45 分钟的时间就可以安装一个完整的深度学习环境，从裸机到开始作为训练平台。

IBM 还宣布对 PowerAI 框架的更新，可为从裸机到深度学习环境的整个堆栈提供支持。对于希望采用人工智能尤其是深度学习，但又不愿意承担主要由开源项目代码驱动的环境的支持责任的企业客户来说，这种针对 PowerAI 平台的全堆栈支持无疑是个振奋人心的好消息。

直到最近，PowerAI 平台都是基于支持高性能计算的 IBM Power S822LC 加速系统而构建。该系统还被用于最新构建的两台最大的超级计算机中。它旨在让数据能够尽可能自由地流动。在支持高性能计算的 Power S822LC 中，通过 GPU 加速是服务器设计的核心。它可以通过点到点的高速 NVLink 连接，将服务器上的所有计算引擎连接起来。

在使用 NVLink 的 Power AC922 中，物理连接是处理器芯片的组成部分，使得 GPU 在数据处理方面更像是 CPU 的对等设备，而不是依赖于 CPU。GPU 与 CPU 之间的双 NVLink 连接，支持极其快速且近乎一致地直接访问系统内存（不像其他设计，依靠共享的 PCIe 3.0 总线或交换设备）。这些更低延迟的连接可支持更大、更复杂的神经模型以及更大的训练数据集。

人工智能基础架构的基础：更完善的数据湖

虽然企业可以评估 Power AC922 是否应该成为他们人工智能基础架构部署的核心，但是他们需要牢记：这种环境的基础应该是更完善的数据湖，适合于运行人工智能服务的数据湖。为获得更完善的数据湖，他们需要考虑五个关键方面：

- 人工智能时代的数据架构
- 支持运行快速的企业级 Hadoop 和 Spark 的高性能服务器
- 整合式、可兼容、企业就绪的向外扩展存储器
- 与当前实践相比，更出色更轻松的数据管理
- 支持包含各种模式和架构的多个数据平台

接下来的部分将围绕这五个要素深入探讨 IBM 构建人工智能就绪的数据湖的方法。

人工智能时代的数据架构

当今的数据湖遇到了数据架构这个瓶颈，数据架构比赋予其能力的“大数据”和软件定义基础架构革命出现的更早。这是因为，无论是处理存储数据的软件定义存储，还是分析数据的服务，或者移动数据的基础架构，统统都需要在服务器上运行。这些服务器的数据架构是 PCIe 3.0，推出于 2010 年，比触发大数据革命的 Hadoop 初始版本早一年，比存储网络行业协会 (SNIA) 试图制定软件定义存储相关标准早至少两年。虽然支持全球大多数服务器的处理器提高了数据处理速度，但它们在数据移动和管理方面（常常被称作为 I/O）基本没有变化。

速度更快的数据服务器

由于 PCIe Gen 3 总线，推动企业人工智能所需的、支持数据湖和分析的大多数服务器都遇到了 I/O 限制的瓶颈。IBM Power 大大提高了 I/O 带宽和性能。IBM Power 在 IBM POWER9 服务器中引入了 PCIe 4.0 接口，速度是 PCIe 3.0 的两倍。这通过显著提高附加 PCIe 的存储器和 FPGA 的性能，改善了服务器的数据移动能力。它还通过将连接集群的网络接口速度翻倍，提高了集群的性能。此外，对于使用 GPU 加速型分析的数据湖，第二代 NVLink 接口可以提供是 PCIe Gen 3 大约 5-6 倍的带宽优势。

更简单、更快速的企业级 Hadoop 和 Spark

当今大多数企业都没有将 Hadoop 和 Spark 视为任务关键型应用，而且许多企业仍然认为数据湖对于业务运营并不重要。然而，数据湖一旦成为支持人工智能的平台，它就会华丽转身，从支持系统变成为运行企业的 ERP 或 CRM 等任务关键型应用的环境。

这意味着企业将需要使用企业就绪、任务关键型硬件和软件来构建数据湖。IBM 表示，它会验证和支持整个 IBM Power 硬件和软件堆栈，以便确保可靠性。在硬件方面，由于与其他处理器架构相比，IBM Power 客观上具有更高的单核性能，因此使一些节点具有与替代架构上更大集群相同的性能，甚至更高的性能。

在软件方面，企业级就绪始于 IBM Elastic Storage Server (ESS) 平台，该平台消除了 Spark 和 MapReduce 中与 HDFS 相关的各种问题。IBM 还通过 Hortonworks 提供企业支持，Hortonworks 是一种可大规模扩展的开源平台，可以存储、处理和分析来自许多来源和各种格式的海量数据。Hortonworks 包括 MapReduce、HDFS、HCatalog、Pig、Hive、HBase、ZooKeeper 和 Ambari。

更好的存储能力

IBM 相信，对于先进的数据湖，IBM ESS 是一种理想的存储方法。IBM ESS 是一种软件定义存储解决方案，结合了 IBM Spectrum Scale 软件与 IBM POWER 服务器和存储机柜。并行文件系统 IBM Spectrum Scale 是 IBM ESS 的核心，可以在传播时扩展系统吞吐量，同时仍然提供单个命名空间。这就能够获得高性能，同时避免产生数据孤岛，使存储管理更加容易。

IBM 决定使用 ESS 来提供 HDFS 的企业级替代方案，它可用于大型向外扩展的大数据应用（比如 Spark、MapReduce）以及某些深度学习框架。由于采用三合一复制模型和专用数据孤岛，HDFS 可能非常低效。虽然它处理简单事务速度很快，但是处理大型复杂流程就变得缓慢。此外，HDFS 的标准协议支持有限，不易与其他基础架构相集成。

IBM ESS 由 IBM Spectrum Scale 提供支持，是一种软件定义存储方案，受到了各行各业的热烈欢迎，用于支持计算密集的并行工作负载。IBM ESS 在外观和功能方面与 HDFS 很像；它支持对数据进行多协议访问，包括通用网络文件系统 (CIFS)、NFS、对象、块存储等。IBM Spectrum Scale 具有策略驱动的数据移动能力，可以根据数据使用情况或其他定义标准对闪存、磁盘、磁带和云存储进行分层。IBM ESS 为企业提供了一个存储平台，可以执行大数据分析以及其他工作负载，同时将数据留在原处以供其他应用使用或实现经济有效的归档。

更有效、更轻松的数据管理

Spark 是用于深度学习的卓越的开源大数据分析框架，但是实施 Spark 并非易事。企业需要合适的工具、技能集和工作流，以及与其他框架的集成，才能保证 Spark 高效安全地运转，并确保其可管理性。IBM 设计了 Spectrum Conductor 来帮助企业克服这些障碍。企业不是将环境的各种组件堆放在一起，而是获得一种全面受支持的集成解决方案，其中包含 Spark 发行版，支持 Spark 和其他框架的多租户架构，并使动态资源分配成为可能，推动 HPC 基础架构茁壮成长。

IBM Spectrum Conductor 使用 Spark 作为计算和传输层，以便从数据湖中的各种数据源提取数据。随后，该解决方案会根据用户提供的定义将矢量信息和元数据添加到这些数据，然后执行整个数据转换工作。它将促成数百个不同的 Spark 实例，遍历来自各个来源的数据，并将数据转换成为方便 Caffe 或 TensorFlow 使用的数据集。

IBM Spectrum Conductor 是一个工作负载和资源管理器，专为分析和深度学习工作负载进行了优化。IBM 建议企业使用 IBM Spectrum Conductor 而不是 YARN 来处理这些工作负载，这样就可以通过可预测的运行时有效调度作业，而且能够获得所需的资源，比如合适数量的 GPU 和 CPU，合适容量的内存等。根据工作负载的特征，IBM Spectrum Conductor 将确定哪些资源可用，有多少资源可用，资源在何处以及应该采用什么队列顺序。这样可以最大限度提高利用率，IBM 声称利用率可以达到达 40% 甚至更多，显著高于大约为 20% 的数据中心平均使用率。最终，性能实现显著提升，使 Spark 能够更加高效地运行。

IBM Spectrum Conductor 的多租户特性允许部署多个 Spark 实例，以便实现最优的资源利用率、更大的规模和更高的性能，同时消除与独立 Spark 实施相关的资源孤岛。该解决方案还促进了 Spark 与 Hadoop、MongoDB 或 Cassandra 等应用框架的集成。IBM Spectrum Conductor 是经过许可且受支持的软件包，最终用户可以将其构建到现有或新建的集群中。

IBM Spectrum Conductor Deep Learning Impact 是一种深度学习环境，旨在加快和简化数据研究员的工作。它可以在任何基础架构上运行，IBM 已经在 Spark 内开发了 GPU 优化，可以加快 Spark 实例内的计算速度。鉴于 PowerAI 拥有比其他基础架构竞争产品更高的 CPU-GPU 带宽，因此它可以显著利用这种 GPU 优化。因此，用于深度学习训练的 PowerAI 集群可以通过一种起伏变化的准备和训练模式，用于 IBM Spectrum Conductor 上的数据准备工作。在 IBM PowerAI Enterprise 路线图上，IBM Spectrum Conductor Deep Learning Impact 功能将在 2018 年第二季度整合到 PowerAI Enterprise 软件包中。

更出色的用户接口

如今，大多数数据湖都是数据研究员的天地 — 业务部门的报告请求需要专业的技能，可能需要数周或数月时间才能生成。认为当今的数据湖将以与之前的数据仓库相同的方式演变，并非没有道理。两年前，从数据仓库获取有用信息需要数月时间。然而，随后出现了一些工具，支持开发人员访问数据仓库，从而让他们能够开发出利用数据的应用。接着，业务部门人员开始使用 Excel 插件来访问数据仓库，以便生成业务报告。

当今的数据湖也可能发生类似的“开放”情况。目前，企业开发人员主要关注利用人工智能来彻底改变他们的应用。为此，他们需要访问数据湖中的数据，利用企业的人工智能能力。业务部门用户也不会远远滞后。最终，对数据湖中关键数据的访问和利用这些数据支持人工智能的能力将实现民主化。但是，这意味着，用于实现这个目标的工具需要改进，需要变得更易于使用和更加直观。

为此，IBM 开发了 DSX Local，用于提供卓越的数据科学体验。DSX Local 简化了数据湖数据使用过程中的提取转换加载 (ETL) 环节，并连接到适当的深度学习框架。DSX Local 可在本地和云端部署，不仅可以在 IBM Power 上运行，也可以在其他处理器架构上运行。DSX Local 是一种面向数据研究员和数据工程师的开箱即用的企业级本地解决方案，可提供一系列集成 IBM 技术的数据科学工具，比如 RStudio、Spark、Jupyter 和 Zeppelin Notebook。该用户接口采用极致的直观设计，可为数据研究员和开发团队提供一个协作项目空间。

支持多个现代数据平台

提到数据库，没有放之四海而皆准的产品。虽然许多任务仍然可以通过传统关系数据库（例如 CRM 和 ERP）来执行，但越来越多的挑战只有现代模式才能解决。打算构建人工智能基础架构（在改进的数据湖上执行人工智能训练和推理）的企业需要接受除 SQL 之外的更多模式，尤其是 NoSQL（例如，用于物联网或内容工作负载）和图形（例如，用于欺诈检测）。

此外，如果企业正在投资构建新的数据平台，那么他们应该研究内存数据库、GPU 加速型图形数据库和 NoSQL 等现代架构。这种数据密集型模式需要许多线程，有些可以仅在 CPU 中处理，有些在内存中处理，另一些则在 GPU 中处理，具体取决于模式和数据库。比如使用图形数据库时，整个数据库都需要在内存中处理。

支持多个数据平台将产生强大的人工智能平台。基于数据湖，企业可以构建一种全新的加速型开源数据库，比如 Redis、MongoDB 和 EDB Postgres；图形数据库 Neo4j；以及面向 GPU 的分布式内存数据库。IBM 声称 POWER9 在这些数据库上具有明显的性能优势，能够提供性价比保证。这些数据库利用 Power 的 I/O、快速互连、内置 RAS 以及庞大内存。许多这些新服务将在内存中处理，因为数据以前所未有的速度来回移动，因此需要大量内存、高性能、灵活的 I/O 访问。

大型模型支持

IBM 认为 Power Systems 是运行这些现代模式的理想平台，这主要归功于它的四路多线程、单核性能、I/O 能力和能够提供完全一致性的内置 NVLink。NVLink 允许 GPU 上的进程利用系统内存，几乎没有任何性能损失，这是克服 GPU 通常附带的 12GB、16GB 或 32GB 最大内存的限制的关键功能。这个内存不足以支持人工智能愿景、4K 视频或具有多个矩阵层的更为复杂的人工智能模型。因此，企业必须做出取舍，比如使用低分辨率图像，使用网络视频而不是高清视频，或者开发深度不及预期的网络。

为此，在 PowerAI 的最新发行版中，IBM 引入了所谓的“大型模型支持”。在 PowerAI 上，整个模型都可以加载到 CPU 内存中。例如，在一个包含四个 GPU 的系统中，可以有四个 230GB 模型的实例，而不是四个 16GB 或 32GB 模型的实例，允许将整个数据集而不只是子集加载到模型中，允许更准确地解决问题，或者使用高清而不是网络级别的视频文件。在其他系统中，PCIe 卡将 GPU 与系统内存连接起来，由于 PCIe 带宽有限，就会造成性能损失。

共享内容甚至也将成为可能，因为 IBM Power 具有 CAPI，拥有显著高于 PCIe 3.0 的带宽，可以连接到非常快速的 SSD，然后可以用作具备近场内存性能的内存。这使得将数据湖中的重要部分放在 SSD 上的内存中成为可能。IBM 在 2017 年 8 月通过技术预览宣布“大型模型支持”是其 Caffe 发行版的一部分，现在开始向各种框架提供商开放源码，第一个就是日本 Preferred Networks 的开源深度学习 Chainer 框架。

未来展望

IDC 预计，人工智能将迅速发展，大多数企业将需要在未来 12-24 个月内采用人工智能文化，否则将面临被竞争对手“智胜”的风险。人工智能不仅将迅速发展，而且将注入数据中心或云中的每一个工作负载。因此，制定长期的战略方法来采用人工智能至关重要：人工智能如何提高企业业绩？人工智能如何保护企业安全？人工智能如何提升企业竞争力？人工智能如何提高企业效率？企业有一系列诸如此类的问题需要解答。与此同时，数据研究员和应用开发人员进行的人工智能开发需要与长期的人工智能基础架构计划携手并进。

当企业为智能服务设计或重新设计数据湖时，应该考虑到他们目前并不知道自己将在哪里实施人工智能。存在许多潜在的场景，包括数据湖训练、关联实时服务以及运行与社交媒体关联的推理模型。企业在设计时应考虑到：他们并不清楚到底能以多快的速度在未来 12、24 或 36 个月采用人工智能，并且以这种方式构建系统，他们将能够让人工智能融入整个环境中，而不会造成任何性能损失。其中的理念就是拥有一个类似于沙箱的平台，允许企业运用所有不同的人工智能方法。

在这种前瞻性方法中，企业还应该考虑到，人工智能对基础架构的需求在未来两到三年将显著增加。人工智能模型将变得更加复杂，用于训练的数据将越来越多，利用人工智能的应用也将增多，大多数客户的推理负载可能规模相当庞大。早期采用者要想避免陷入困境（在两年或三年的时间里，某些情况下需要将基础架构切换三次），他们应该使用能够处理当前和未来三年数据极其密集的人工智能工作负载的基础架构解决方案。

挑战/机遇

对于企业

对于当今的企业来说，第一大挑战是通过一种有计划、快捷的方式将人工智能融入到企业当中。几年前，毫无计划可言的人工智能试验虽然让企业付出了很高的代价，但却是当年风靡一时的时尚，而且人们普遍认为非常必要。现在，这种情况可以避免。市场中有各种各样的解决方案，能够为企业提供一致、高效且有效的人工智能路线图。正因如此，对于企业来说，人工智能不再是挑战，而是机遇。当然，仍有很多的阻碍，这不仅在于吸引专业的人工智能数据研究员，还在于准备数据和构建合适的基础架构方面，而现在任何企业都无需进行无谓的重复劳动。

企业面临的另一个挑战是，他们正在转为采用异构数据中心。多年来，企业一直在标准架构服务器上运行工作负载；现在，企业不但需要了解如何集成、编程和扩展 GPU、FPGA、ASIC 等加速器和众核处理

器，还需要开始适应不同架构的主机处理器，这些处理器在处理特定工作时表现更出色、更高效。

不过，不用担心。因为有 Linux、虚拟化和容器化这些强大的主流技术，其他处理器很难引起人们的注意，除非它在处理特定工作负载方面更为出色。

对于 IBM

对于 IBM 来说，最大的挑战与前面提到的企业面临的第二个挑战密切相关。由于惯性，企业很难接受某些工作负载在使用不同处理器的系统（比如 Power Systems）上运行得更好这个客观事实，这种企业惯性仍然是 IBM Power Systems 取得市场成功的最大障碍。“标准”架构的概念已深深融入到 IT 文化当中，大多数人们都忘记了“标准”来自于共同特性设计。换句话说，如果它是标准的，那么它可以顺利完成所有工作，没有特别出色的地方。

因此，人工智能对于 IBM Power Systems 来说是个千载难逢的机遇。它是 Power System 的理想工作负载，并且正在迅速演变为一个巨大的市场商机。这个机遇是 IBM 不容错过的，IBM 最重要的成功秘诀就是：保持简单，无论基础设计有多复杂；保持经济实惠；面向开发人员（数据中心的主要影响者）大力营销推广。

最后，IBM 不应忽视的一个机遇就是它凭借 Watson 所获得的品牌知名度。无论是从实际上还是概念上来讲，将人工智能基础架构与 Watson 连接起来并非易事，但是这会对 IBM Power Systems 带来积极影响。

结论

那些仍在尝试通过人工智能基础架构来将人工智能融入到其运营、产品和服务中的企业，可以免去这种麻烦了。与 12 个月前不同的是，现在人们对深度学习和人工智能推理所需的基础架构类型有了越来越多的了解。供应商纷纷开始为人工智能整合各种硬件、软件和服务包，以便降低人工智能应用的门槛。IBM 表现尤为抢眼，它已经利用其 Power Systems 处理器和 I/O 优势以及它所完成的定制开发，通过 NVLink 2.0 将新的 POWER9 处理器与 NVIDIA V100 GPU 集成起来，提供性能极高的服务器，作为其 PowerAI 系统的基础。该公司还准确无误地意识到，企业纷纷努力升级数据湖，以便能够快速、全面地准备各种类型的数据。这是一个经常被忽视的关键能力。IDC 认为，人工智能需要的基础架构不同于多年来数据中心内的标准架构，而且企业显然已经意识到了这一点。在前进途中，他们将采用不同的主机处理器、各种加速技术和大量的内存空间。

关于 IDC

International Data Corporation (IDC) 是全球首屈一指的信息技术、电信和消费科技市场情报、咨询和活动服务供应商。IDC 致力于帮助 IT 专业人士、业务高管和投资机构以事实为基础，做出有关技术采购的决策，制定业务发展战略。IDC 在全球拥有超过 1,100 名分析师，他们从全球、区域和本地视角对 110 多个国家或地区的技术与行业机会和趋势提供专业化的指导意见。50 多年来，IDC 一直为客户提供战略洞察，帮助他们实现关键的业务目标。IDC 是 IDG 旗下子公司，IDG 是全球领先的技术、媒体、研究及活动服务公司。

全球总部

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