业务简介

值得数据中心负责人关注的人工智能数据科学要素

像数据科学家一样思考如何能够帮助您依据部署 IT 基础设施的轻重缓急调整人工智能 (AI) 目标。

基础设施和数据科学团队必须共同努力，才能满足人工智能应用的计算、延迟和吞吐量需求，同时保持数据中心的效率。

人工智能是过去十年中技术突飞猛进的一个领域，通常指机器学习和深度学习的训练与推理。过去，图像、文本、语音和音频等庞大或复杂的数据只能由人类解读。如今，随着人工智能领域的不断创新，各个公司已经可以构建有能力分析这类数据的应用，而且数据分析的规模和速度远远超过人类。但是，这类新应用具有极高的计算需求，为保持准确性往往对延迟和巨大吞吐量有着严苛的要求。

因此，基础设施团队必须与从事数据科学和人工智能应用开发工作的同事紧密合作，确保这些需求得到满足。同时，他们还必须持续关注更大规模数据中心优先考虑的那些因素，例如保持经济高效的运营和灵活性。为此，基础设施架构师必须在以应用为中心的同事中倡导和培养以数据为主导的全局观。

在基于 CPU 的基础设施上运行人工智能工作负载

您已经在 IT 基础设施上做了投资，因此，在增加新的工作负载时（无论是人工智能还是其他）必须先确保新功能可以优化现有资源，然后再考虑投资新计算资源。您当前的英特尔® 架构可以带来：

- **灵活性：** 英特尔® 至强® 平台原本就具有多种用途，因此能够支持广泛的工作负载，包括机器学习和深度学习（见图 1）。借助大量针对机器学习和深度学习所做的软件优化和集成英特尔® 深度学习加速技术的推理加速功能，基于 CPU 运行人工智能的速度大幅提高。
- **高效率：** 将您的人工智能工作负载需求与当前系统使用情况进行对比，确定在何处以及如何为这些工作负载优化资源配置。通常情况下，您可以利用现有 CPU 从备用容量中为人工智能工作负载提供更多使用空间。目前，大多数机器学习和深度学习推理都在 CPU 上运行，在许多情况下，CPU 是深度学习训练的理想选择。通过针对常见人工智能软件框架（例如 TensorFlow 和 PyTorch）、库和工具所做的优化，帮助保持较高的性能功耗比和性价比，使 PUE 比率尽可能接近 1。
- **可扩展性：** 您可以按需跨多个英特尔® 数据中心或边缘节点轻松扩展您的人工智能训练工作负载。在设计系统时，您可以使用诸如英特尔® 以太网 700 系列这样的网络技术和英特尔® 傲腾™ 这样的内存存储技术来优化网络和内存配置，从而做到扩展、效率两不误。这样您可以充分利用现有的硬件投资，扩展深度学习工作负载，从而获得更高的吞吐量，甚至有能力处理巨大的数据集。Facebook 已经充分证明了这种方法的可行性。

	推荐引擎	经典机器学习	循环神经网络	使用大数据样本的模型	其他实时推理	空闲时段训练
用途	推荐广告、搜索、应用等	从数据获取洞察	语音识别	医学影像、地震勘探、3D 环境	图像识别、语音识别、自然语言处理	任何用途
类别	多层感知器 (MLP)	回归、分类、集群等	循环神经网络 (RNN)	卷积神经网络 (CNN)	多种类别	任何类别
CPU 的优势	训练和推理。将更大的内存用于嵌入层	将速度更快的内核用于大型数据集和难以并行运行的算法	实时推理。将速度更快的内核用于顺序、难以并行处理的数据	训练和推理。需要更大的内存	将速度更快的内核用于难以并行处理的小批数据	数据中心容量

图 1. 基于英特尔® 技术的现有基础设施可以支持多种人工智能用例和工作负载

英特尔® 技术助力人工智能性能提升

第二代英特尔® 至强® 可扩展处理器能够为各类人工智能 (AI) 应用提供可扩展的性能，与上一代产品相比，**平均性价比提高 42%**[1]。

长期以来，英特尔和 Google 工程师紧密合作，面向英特尔® 至强® 平台不断优化 TensorFlow 这一灵活的人工智能开源框架，在使用英特尔® 深度学习加速技术时能将**推理速度提升多达 3.75 倍**[2]。

实施人工智能：数据科学家的视角

从事数据科学工作的同事可能会从不同的角度看待事物。他们很可能会认为，基于 GPU 的硬件平台能够为某些深度学习训练工作负载提供非常高的吞吐量，因此能够加快人工智能模型开发。

但实际上，开发速度很大程度上取决于涉及的人工智能工作负载、数据类型和要求。与数据科学家紧密合作，帮助他们保持开放的态度来选择适合的平台。首先向数据科学家提出以下问题，确定工作负载本身的特征：

- 您需要运行什么类型的模型？
- 这些模型的大小分别是多少（参数数量）？
- 用于建模的数据有多大？
- 每个模型的批大小通常是多少？
- 保持活跃的最大激活数量是多少？
- 每个模型的计算强度是多少？
- 有哪些延迟限制？

通过了解这些工作负载特征，您可以确定数据科学家对人工智能工作负载的基本计算要求。如果他们运行的是专门的深度学习训练，那么使用专用硬件加速顺理成章。在大多数其他情况下，要与数据科学家达成双方都满意的折中方案，在现有的计算基础设施上运行人工智能工作负载可能是理想的选择。这样既能满足数据科学家的加速和性能需求，又能帮助您达成效率、可扩展性和灵活性目标。

此外，还要考虑项目未来的扩展需求。数据科学家经常会在规模相对较小的平台上使用少量 GPU 来试验新的算法和工作负载。但是，当需要将这些项目以企业级规模部署到生产环境中时，几乎很少有 IT 团队拥有足够的预算来提供相应规模的平台。确保数据科学家在与大规模部署相当的平台上进行试验，可帮助他们减少一些不必要的挑战。

树立全局观，为人工智能奠定共同基础

接下来，集中精力确保您的数据科学团队完全赞同您提出的方法，并准备好解决他们可能提出的任何疑问或异议。之所以会出现这些疑问或异议，常常是仅考虑应用加速所致。作为全面掌握 IT 基础设施情况的专家，您可以提供宝贵的意见和见解，说明在 IT 环境中以更广阔的视角看待应用如何能够为数据科学家带来他们原本没有想到的优化机会。

向他们介绍从提取数据到使用数据的完整数据分析/人工智能管道（见图 2）。数据必须经过采集、提取和清洗后，才能用于人工智能算法。如今，这一过程大部分在英特尔® 至强® 平台上完成，这意味着如果最后一步人工智能操作也在同一平台上进行，就可以降低复杂性，节约整个管道的开发时间。

对于基于 Spark 和 Hadoop 构建集群的企业和机构，这种融合将获得进一步增强。因为英特尔针对这些集群进行的优化能够

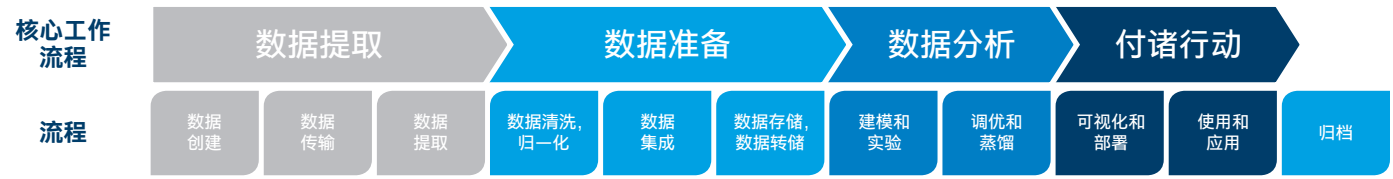

图 2. 人工智能应用的底层数据管道。使用基于 CPU 的现有基础设施运行您的人工智能工作负载

提高它们在运行机器学习和深度学习工作负载时的效率。此外，像端到端开源数据分析和人工智能平台 Analytics Zoo 这类工具同样也集成了英特尔的 Spark 和 Hadoop 优化方案。借助它们，您可以随着需求的增加，无缝扩展这些工作负载，同时仍保有完全控制和可见性。

建立合作伙伴关系，取得长期成功

与从事数据科学工作的同事建立合作关系可为您的人工智能策略奠定坚实的基础。英特尔拥有从边缘到云的灵活人工智能加速产品组合，能够满足您不断变化的功耗、性能和内存需求，助您在这一基础上大展拳脚。此外，英特尔依托英特尔® AI Builders 社区，建立了广泛的解决方案提供商生态系统，可帮助您快速开始。

与您所在的企业或机构中的数据科学家一样，您在关键的人工智能计划中也有着举足轻重的作用，因此你们必须保持沟通渠道的畅通。了解彼此的业务重点和关注点是迈向成功的第一步。

了解更多信息

- **业务简介**
  加速人工智能落地
- **解决方案简介**
  人工智能驱动型解决方案提升医疗保健服务与质量 (JLK)
- **案例研究**
  Kongsberg Maritime 海上解决方案
- **解决方案简介**
  运用人工智能分析时尚和奢侈品市场动态 (IFDAQ)
- **解决方案简介**
  人工智能赋能下一代联络中心

[1] **预估性能提高 36%，预估性价比提高 42%：对比十个全新双路第二代英特尔® 至强® 金牌处理器和上一代产品，在 SPECrate®2017_int_base(est)、SPECrate®2017_fp_base(est)、STREAM Triad 和英特尔® LINPACK* 分发版基准测试中的几何平均值。**第二代英特尔® 至强® 金牌 R 处理器：1 个节点，搭载 2 个第二代英特尔® 至强® 金牌处理器 (62xxR/$$) 的英特尔® 参考平台，总内存 384 GB（12 个插槽/32 GB/62xx@2933，52xx@2666），ucode 0x500002c，启用超线程技术（STREAM (GB/s) 除外），LINPACK (GFLOPS/s)，启用睿频，Ubuntu19.10（内核 5.3.0-24-generic），6258R/$3950：SPECrate®2017_int_base(est)=323，SPECrate®2017_fp_base(est)=262，STREAM=224，LINPACK=3305；6248R/$2700：SPECrate®2017_int_base(est)=299，SPECrate®2017_fp_base(est)=248，STREAM=224，LINPACK=3010；6246R/$3286：SPECrate®2017_int_base(est)=238，SPECrate®2017_fp_base(est)=217，STREAM=225，LINPACK=2394；6242R/$2529：SPECrate®2017_int_base(est)=265，SPECrate®2017_fp_base(est)=231，STREAM=227，LINPACK=2698；6240R/$2200：SPECrate®2017_int_base(est)=268，SPECrate®2017_fp_base(est)=228，STREAM=223，LINPACK=2438；6238R/$2612：SPECrate®2017_int_base(est)=287，SPECrate®2017_fp_base(est)=240，STREAM=222，LINPACK=2545；6230R/$1894：SPECrate®2017_int_base(est)=266，SPECrate®2017_fp_base(est)=227，STREAM=222，LINPACK=2219；6226R/$1300：SPECrate®2017_int_base(est)=208，SPECrate®2017_fp_base(est)=192，STREAM=200，LINPACK=2073；5220R/$1555：SPECrate®2017_int_base(est)=257，SPECrate®2017_fp_base(est)=220，STREAM=210，LINPACK=1610；5218R/$1273：SPECrate®2017_int_base(est)=210，SPECrate®2017_fp_base(est)=188，STREAM=199，LINPACK=1290，基于英特尔 2019 年 12 月 25 日所做的测试。第一代英特尔® 至强® 金牌处理器：1 个节点，搭载 2 个英特尔® 至强® 金牌处理器 (61xx/$$) 的英特尔® 参考平台，总内存 384 GB（12 个插槽/32 GB/61xx@2666，51xx@2400），ucode 0x500002c，启用超线程技术（STREAM (GB/s) 除外），LINPACK (GFLOPS/s)，启用睿频，Ubuntu19.10（内核 5.3.0-24-generic），6152/$3655：SPECrate®2017_int_base(est)=224，SPECrate®2017_fp_base(est)=198，STREAM=200，LINPACK=1988；6148/$3072：SPECrate®2017_int_base(est)=225，SPECrate®2017_fp_base(est)=198，STREAM=197，LINPACK=2162；6146/$3286：SPECrate®2017_int_base(est)=161，SPECrate®2017_fp_base(est)=175，STREAM=185，LINPACK=1896；6142/$2946：SPECrate®2017_int_base(est)=193，SPECrate®2017_fp_base(est)=176，STREAM=185，LINPACK=1895；6140/$2445：SPECrate®2017_int_base(est)=202，SPECrate®2017_fp_base(est)=183，STREAM=188，LINPACK=1877；6138/$2612：SPECrate®2017_int_base(est)=189，SPECrate®2017_fp_base(est)=195，STREAM=189，LINPACK=1976；6130/$1894：SPECrate®2017_int_base(est)=172，SPECrate®2017_fp_base(est)=165，STREAM=185，LINPACK=1645；6126(proj)/$1776：SPECrate®2017_int_base(est)=141，SPECrate®2017_fp_base(est)=157，STREAM=170，LINPACK=1605；5120(proj)/$1555：SPECrate®2017_int_base(est)=148，SPECrate®2017_fp_base(est)=148，STREAM=159，LINPACK=924，5118/$1273：SPECrate®2017_int_base(est)=134，SPECrate®2017_fp_base(est)=132，STREAM=149，LINPACK=818，基于英特尔 2020 年 2 月 18 日所做的测试。

[2] 英特尔® 人工智能推理精选解决方案将性能提升高达 3.75 倍。这一解决方案已经过 KPI 目标测试：2019 年 2 月 26 日的 OpenVINO/ResNet50 INT8 性能测试，测试所用软硬件配置如下：

基础配置：1 个节点，2 个英特尔® 至强® 金牌 6248 处理器；1 块英特尔® 服务器主板 S2600WFT；总内存 192 GB，12 个插槽/16 GB/2666 MT/s DDR4 RDIMM；超线程：启用；睿频加速：启用；存储（引导）：英特尔® 固态盘 DC P4101；存储（容量）：至少 2 TB 的英特尔® 固态盘 DC P4610 PCIe NVMe；操作系统/软件：CentOS Linux 版本 7.6.1810（内核），内核 3.10.0-957.el7.x86_64；框架版本：OpenVINO 2018 R5 445；数据集：来自基准测试工具的样本图像；模型拓扑：ResNet 50 v1；批大小：4；nireq：20。这一解决方案已经过 KPI 目标测试：2019 年 3 月 7 日的 TensorFlow/ ResNet50 INT8 性能测试，测试所用软硬件配置如下：

基础配置：1 个节点，2 个英特尔® 至强® 金牌 6248 处理器；1 块英特尔® 服务器主板 S2600WFT；总内存 192 GB，12 个插槽/16 GB/2666 MT/s DDR4 RDIMM；超线程：启用；睿频加速：启用；存储（引导）：英特尔® 固态盘 DC P4101；存储（容量）：至少 2 TB 的英特尔® 固态盘 DC P4610 PCIe NVMe；操作系统/软件：CentOS Linux 版本 7.6.1810（内核），内核 3.10.0-957.el7.x86_64；框架版本：intelaipg/intel-optimizedtensorflow:PR25765-devel-mkl；数据集：来自基准测试工具的综合数据；模型拓扑：ResNet 50 v1；批大小：80

性能测试中使用的软件和工作负荷可能仅在英特尔微处理器上进行了性能优化。

诸如 SYSmark 和 MobileMark 等测试均系基于特定计算机系统、硬件、软件、操作系统及功能。上述任何要素的变动都有可能导致测试结果的变化。请参考其他信息及性能测试（包括结合其他产品使用时的运行性能）以对目标产品进行全面评估。更多信息，详见 www.intel.cn/benchmarks。

英特尔编译器针对英特尔微处理器的优化程度可能与针对非英特尔微处理器的优化程度不同。这些优化包括 SSE2、SSE3 和 SSSE3 指令集和其他优化。对于非英特尔微处理器上的任何优化是否存在、其功能或效力，英特尔不做任何保证。本产品中取决于微处理器的优化是针对英特尔微处理器。不具体针对英特尔微架构的特定优化为英特尔微处理器保留。请参考适用的产品用户与参考指南，获取有关本声明中具体指令集的更多信息。

英特尔并不控制或审计第三方数据。请您审查该内容，咨询其他来源，并确认提及数据是否准确。

性能测试结果基于配置中所示日期进行的测试，且可能并未反映所有公开可用的安全更新。详情请参阅配置信息披露。没有任何产品或组件是绝对安全的。

您的成本和结果可能会有所不同。

英特尔技术可能需要支持的硬件、软件或服务得以激活。