业务简介

第四代英特尔® 至强® 可扩展处理器
AI 和机器学习

采用数据分析和 AI 来驱动关键成果的产出

借助英特尔® 技术，您可以提升洞察质量，驱动关键业务成果的产出。

将人工智能 (AI) 从概念落实到规模化实践始终是一项挑战。先进的 AI 模型曾需要使用专用硬件、高级技能和定制化工具才能将数据转化为业务成果。在整个端到端流水线上运行 AI，无论是在本地、在云端还是使用两者混合的部署方式，往往都意味着要额外增加支出并且难以招聘到合适的人才。对于力求在整个企业业务层面扩展 AI 应用的企业领导层来说，降低复杂性是关键所在。

在企业和机构寻求扩大规模、降低成本和提供新服务的过程中，通过技术来实现商业价值的重要性日益凸显。面对新的应用场景，他们无需定制系统（这可能会增加复杂性），而是可以通过易于扩展的平台来满足当下和未来各种部署的性能需求。

84% 的高管 **认为他们需要借助 AI 来获得成功[1]**

70% 的数据中心 AI 推理任务 **在英特尔® 至强® 可扩展处理器上运行[2]**

90% （到 2025 年） **的企业应用将使用嵌入式 AI[3]**

运用英特尔® 技术加速 AI

第四代英特尔® 至强® 可扩展处理器内置众多加速器，可为 AI、数据分析、网络、存储和科学计算等快速增长的工作负载提供性能和能效优势。在全新英特尔® 高级矩阵扩展（Intel® Advanced Matrix Extensions，英特尔® AMX）这一加速器的支持下，第四代英特尔® 至强® 可扩展处理器具有更为出色的 AI 训练和推理性能。为实现新的内置加速器功能，英特尔为生态系统提供了操作系统级软件、库和 API 支持。

通过内置加速器和软件优化，上一代英特尔® 至强® 可扩展处理器已被证明可以在真实场景下的目标工作负载上实现出色的性能功耗比[4]。这不但可以提高 CPU 利用率，降低功耗，并提高投资回报率 (ROI)，同时还能帮助企业实现可持续发展目标。

性能证明

高达 5.7 倍至 10 倍 的 PyTorch 实时推理性能提升[5]

高达 3.5 倍至 10 倍 的 PyTorch 训练性能提升[6]

数据基于启用内置英特尔® AMX (BF16) 的第四代英特尔® 至强® 可扩展处理器与上一代产品 (FP32) 的比较

AI 关键用例

深度学习：推荐系统和自然语言处理（NLP）

为了根据实时行为为信号和上下文队列提供个性化用户体验，企业可以部署基于深度学习的推荐系统以及使用自然语言处理，同时平衡总体拥有成本（TCO）。推荐系统可帮助企业通过个性化推荐为每个客户提供更好的服务，而自然语言处理则使设备能够更好地理解文本的含义，从而让企业能够更好地了解并满足客户的需求。

需求：

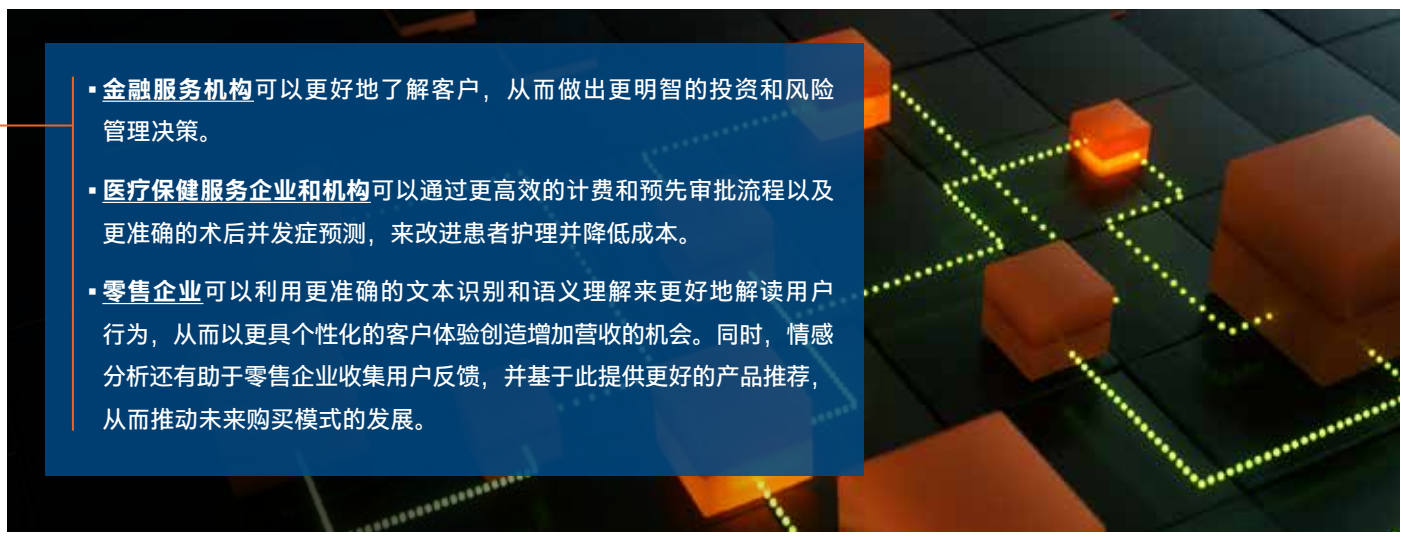
提供个性化的用户体验能够驱动客户的需求并持续吸引他们的关注，在各个行业中都有巨大的营收潜力。为了经济高效地提供更优的用户体验以及为他们提供高质量支持，必须持续改进计算机和推荐系统在解读文本方面的完善程度。

答案：

应用于非结构化数据时，情感分析和内容扩充有助于企业对大量看似无意义的数据进行深度分析，包括实时分析和事后分析。在不同行业，自然语言处理都可帮助提高用户参与度和运营效率，并有助于充分利用新出现的营收机会。

英特尔® AMX 强化了第四代英特尔® 至强® 可扩展处理器的 AI 加速能力，无需额外硬件即可加速深度学习训练和推理。该内置加速器可为自然语言处理、推荐系统和图像识别等 AI 应用提供更强的支持。

实现：



- **金融服务机构**可以更好地了解客户，从而做出更明智的投资和风险管理决策。
- **医疗保健服务企业和机构**可以通过更高效的计费和预先审批流程以及更准确的术后并发症预测，来改进患者护理并降低成本。
- **零售企业**可以利用更准确的文本识别和语义理解来更好地解读用户行为，从而以更具个性化的客户体验创造增加营收的机会。同时，情感分析还有助于零售企业收集用户反馈，并基于此提供更好的产品推荐，从而推动未来购买模式的发展。

机密计算：多方机器学习

需求：

在不影响客户敏感数据的机密性和隐私性的前提下充分利用机器学习能力。

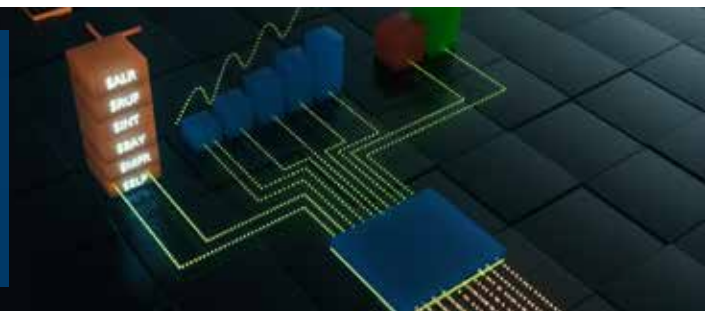
答案：

基于机密计算的多方机器学习特别适用于金融服务、欺诈检测和研究。

- **英特尔® 软件防护扩展（Intel® Software Guard Extensions，英特尔® SGX）**是目前市场上经过深入研究、多次更新和广泛部署的数据中心级机密计算技术，拥有极小的信任边界。
- **英特尔® 高级矢量扩展 512（Intel® Advanced Vector Extensions 512，英特尔® AVX-512）**是一种通用的性能强化型加速器，适用于广泛的数据类型和机器学习应用，第四代英特尔® 至强® 可扩展处理器也会继续内置这一加速器。英特尔® AVX-512 可以加速来自多种来源、用于模型训练的非结构化数据的预处理，并可以加速数据传输，从而缩短待处理数据集的准备时间。与英特尔® AVX-512 配合使用的英特尔® Extension for Scikit-learn 还加速了用于训练和推理的机器学习算法。

实现：

- **医疗保健服务企业和机构可以在更好地保护患者机密信息的情况下利用数据进行更先进的研究。**
- **金融服务机构可以更好地预测潜在的欺诈活动，同时打击洗钱和非法融资。**



提高工作效率

使用集成到 TensorFlow 和 PyTorch 中的库，开发人员无需任何额外工作即可利用第四代英特尔® 至强® 可扩展处理器内置 AI 加速技术的优势。此外，只需更改几行代码，开发人员即可无缝加速单节点和多节点配置中的 scikit-learn 应用。多年来，英特尔一直与 AI 社区携手，共同优化常见的 AI 框架、软件和工具，使其主流发行版能够在英特尔® 产品上更顺畅地运行。建议您充分利用额外工具，例如利用英特尔® 分发版 OpenVINO™ 工具包优化推理模型，利用英特尔® BigDL 在 Apache Spark 上进行分布式深度学习，或利用 cnvrg.io MLOps 平台在数据中心或云端的任何基础设施上编排机器学习流水线。

利用既有基础设施轻松集成

与英特尔合作，企业可以利用他们已经熟悉和正在使用的大规模合作伙伴生态系统缩短部署时间。全球各地的硬件和软件供应商以及解决方案集成商都在使用英特尔® 至强® 可扩展处理器构建其产品，并通过数以千计来自真实场景的实现案例提供更多选择和更好的互操作性。

数十年积累下来的生态系统支持有助于英特尔将其可信计算基础扩展到超大规模数据中心和新的边缘环境之中，并且使英特尔有能力去推进更多构建、扩展和转型工作，帮助企业实现更优的运营敏捷性。这种生态系统上的开放性使得用户可在硬件、软件、云和服务提供商中自由选择。

英特尔® 架构具有很强的灵活性，可在其上选择不同的云服务，并对其进行进一步的量身定制，以支持特定工作负载的需求，因此其应用规模可扩展到全球范围，并覆盖各个主流云服务提供商。

通过英特尔® 合作伙伴联盟，您可访问 AI、云、科学计算和其他解决方案领域的更多专属资源，帮助您规划、构建并为客户提供更多价值。

与强大的生态系统协作，打造更具创新性的解决方案，加快业务发展，创造非凡机遇，推动全球进步，让生活变得更加丰富多彩。

支持性统计数字

利用英特尔**超 50,000 个**独特实例类型、规格和区域，获得更多选择。

领导层在数字化转型之旅中的首要业务目标

2022 至 2024 年间，企业和机构领导层（科技和商业公司）在数字化转型方面的投资预计将达到 6.3 万亿美元，到 2024 年，这一数字将占所有 IT 支出的 55%[7]。本业务简介是一系列业务简介的一部分，旨在阐明领导者在未来的变革中实现业务成功所关注的首要业务目标，以及英特尔® 硬件、软件和服务（包括第四代英特尔® 至强® 可扩展处理器）将如何帮助他们实现这些目标：

- **AI（本简介）：** 采用数据分析和 AI 来驱动关键成果的产出
- **安全性：** 实现更严格的安全性，推动零信任安全策略
- **云：** 启动跨混合云、多云与智能边缘的策略
- **重新定义员工体验：** 支持无边界互动式员工体验
- **ESG：** 促进环境中的公平结果与责任 | 社会 | 治理（ESG）

了解更多信息

www.intel.cn/xeon/scalable

www.intel.cn/content/www/cn/zh/artificial-intelligence/overview.html

intel®

[1] Accenture，2019 年 11 月 19 日，“AI: Built to Scale”（AI：为扩展而构建），https://www.accenture.com/us-en/insights/artificial-intelligence/ai-investments。

[2] 基于英特尔对截至 2021 年 12 月运行 AI 推理工作负载的全球数据中心服务器装机容量的市场建模。

[3] Grand View Research，“Artificial Intelligence Market Size, Share & Trends Analysis Report By Solution, By Technology (Deep Learning, Machine Learning, Natural Language Processing, Machine Vision), By End Use, By Region, And Segment Forecasts, 2022 - 2030" [2022-2030 年按解决方案、技术（深度学习、机器学习、自然语言处理、机器视觉）、最终用途和区域划分的人工智能市场规模、份额和趋势分析报告以及细分市场预测]，https://www.grandviewresearch.com/industry-analysis/artificial-intelligence-ai-market。

[4] 配置详情请见以下网址的 [126-130]：https://edc.intel.com/content/www/cn/zh/products/performance/benchmarks/3rd-generation-intel-xeon-scalable-processors/。

[5] 详情请见以下网址的 [A17]：https://edc.intel.com/content/www/cn/zh/products/performance/benchmarks/processors/（第四代英特尔® 至强® 可扩展处理器）。结果可能不同。

[6] 详情请见以下网址的 [A16]：https://edc.intel.com/content/www/cn/zh/products/performance/benchmarks/processors/（第四代英特尔® 至强® 可扩展处理器）。结果可能不同。

[7] IDC，2021 年 10 月，“IDC FutureScape: Worldwide Digital Transformation 2022 Predictions”（IDC FutureScape：2022 年全球数字化转型预测），https://www.idc.com/getdoc.jsp?containerId=US47115521。

加速器是否可用视 SKU 而定。更多产品详情，请见英特尔® 产品规格页面。

实际性能受使用情况、配置和其他因素的差异影响。更多信息请见 https://intel.cn/PerformanceIndex。

性能测试结果基于配置信息中显示的日期进行的测试，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。

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。

具体成本和结果可能不同。

英特尔技术可能需要启用硬件、软件或激活服务。

1122/MH/MESH/350497-002US