共创美好未来：英特尔® SGX 和英特尔® DL Boost 赋能隐私保护机器学习

Zongmin Gu
Hongliang Tian
Qing Li
Chunyang Hui
蚂蚁集团

Qiyuan Gong
Dongjie Shi
Wesley Du
Yabai Hu
Jack Chen
Yuan Wu
Ban Hsu
英特尔

引言

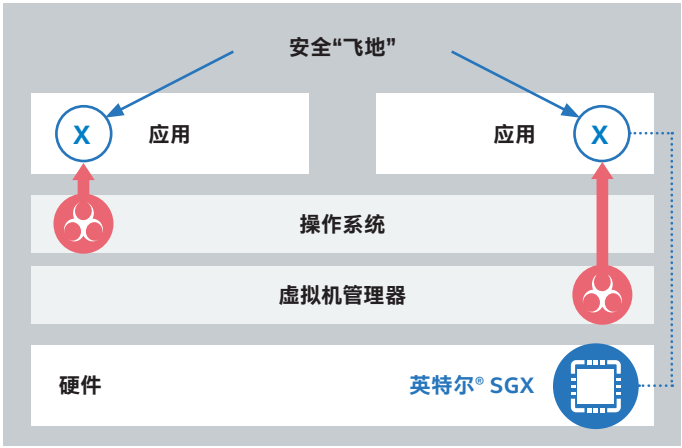
机器学习 (ML) 和深度学习 (DL) 在众多真实的应用场景中愈发重要。这些模型使用已知数据进行训练，并部署在图像分类、内容推荐等场景中进行新数据的处理。总体而言，数据越多，ML/DL 模型就越完善。但囤积和处理海量数据也带来了隐私、安全和监管等风险。

隐私保护机器学习 (PPML) 有助于化解这些风险。其采用加密技术差分隐私、硬件技术等，旨在处理机器学习任务的同时保护敏感用户数据和训练模型的隐私。

在英特尔® 软件防护扩展 (英特尔® SGX) 和蚂蚁集团用于英特尔® SGX 的内存安全多进程用户态操作系统 Occlum 的基础上，蚂蚁集团与英特尔合作搭建了 PPML 平台。在本篇博客中，我们将介绍这项运行在 Analytics Zoo 上的解决方案，并展示该解决方案在第三代英特尔® 至强® 可扩展处理器上得到英特尔® 深度学习加速 (英特尔® DL Boost) 技术助力时的性能优势。

英特尔® SGX 和 Occlum

英特尔® SGX 是英特尔的受信任执行环境 (TEE)，它提供基于硬件的内存加密，隔离内存中的特定应用代码和数据。英特尔® SGX 使得用户层代码可以分配内存中的受保护区域，即“飞地”，这些区域不受更高权限等级程序运行的任何影响 (如图一所示)。



图一 通过英特尔® SGX 加强防护

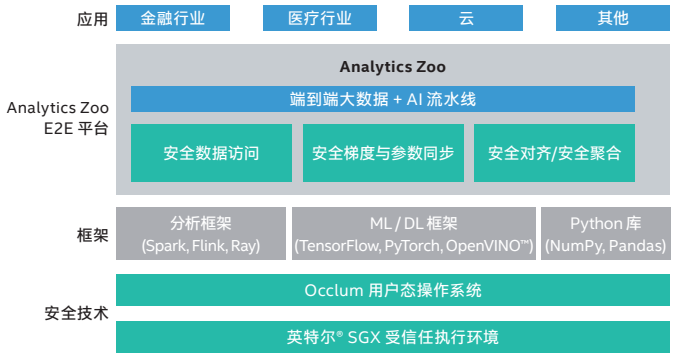
与同态加密和差分隐私相比，英特尔® SGX 在操作系统、驱动、BIOS、虚拟机管理器或系统管理模型已瘫痪的情况下仍可帮助防御软件攻击。因此，英特尔® SGX 在攻击者完全控制平台的情况下仍可增强对隐私数据和密钥的保护。第三代英特尔® 至强® 可扩展处理器可使 CPU 受信任内存区域增加到 512GB，使得英特尔® SGX 技术能够为隐私保护机器学习解决方案打下坚实的基础。

2014 年正式成立的蚂蚁集团服务于超 10 亿用户，是全球领先的金融科技企业之一。蚂蚁集团一直积极探索隐私保护机器学习领域，并发起了开源项目 Occlum。Occlum 是用于英特尔® SGX 的内存安全多进程用户态操作系统 (LibOS)。使用 Occlum 后，机器学习工作负载等只需修改极少量 (甚至无需修改) 源代码即可在英特尔® SGX 上运行，以高度透明的方式保护了用户数据的机密性和完整性。用于英特尔® SGX的 Occlum 架构如图二所示。

Analytics Zoo 赋能端到端 PPML 解决方案

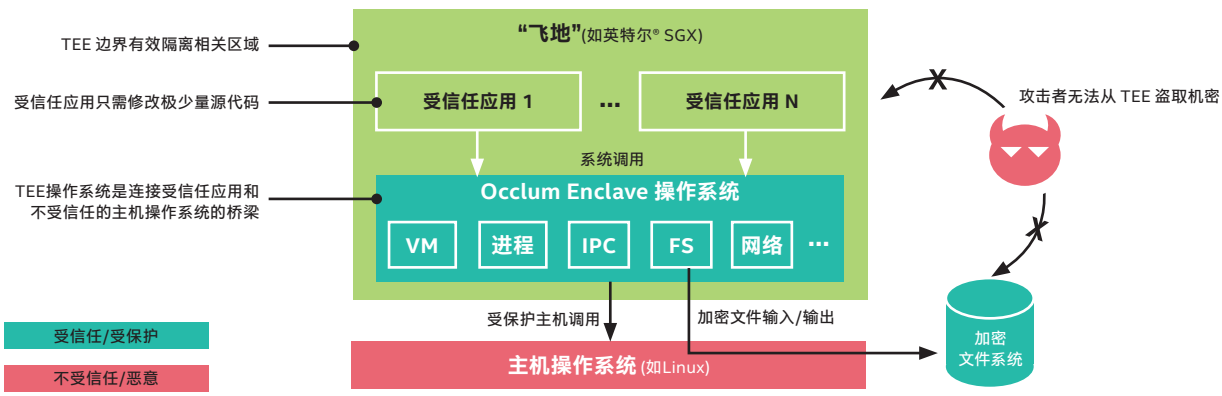
Analytics Zoo

Analytics Zoo 是面向基于 Apache Spark、Flink 和 Ray 的分布式 TensorFlow、Keras 和 PyTorch 的统一的大数据分析和人工智能平台。使用 Analytics Zoo 后，分析框架、ML/DL 框架和 Python 库可以在 Occlum LibOS 以受保护的方式作为一个整体运行。此外，Analytics Zoo 还提供安全数据访问、安全梯度与参数管理等安全性功能，赋能联邦学习等隐私保护机器学习用例。端到端 Analytics Zoo PPML 解决方案如图三所示。



图三 端到端 PPML 解决方案为金融服务、医疗卫生、云服务等应用领域提供安全分布式计算

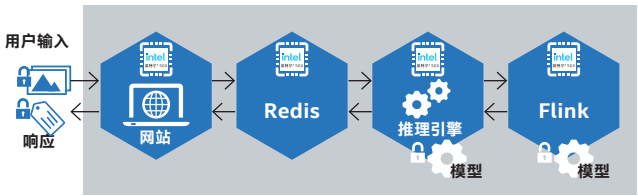
在 Analytics Zoo PPML 平台上，蚂蚁集团与英特尔共同打造了一个更加安全的分布式端到端推理服务流水线 (如图四所示)。该流水线采用 Analytics Zoo Cluster Serving 打造，后者是轻量级分布式实时服务解决方案，支持多种深度学习模型，



图二 用于英特尔® SGX 的 Occlum 架构 (图片来源: Occlum · GitHub)

包括 TensorFlow、PyTorch、Caffe、BigDL 和 OpenVINO™。Analytics Zoo Cluster Serving 包括 web 前端、内存数据结构存储 Redis、推理引擎 (如面向英特尔® 架构优化的 TensorFlow 或 OpenVINO™ 工具套件), 以及分布式流处理框架 (如 Apache Flink)。

推理引擎和流处理框架在 Occlum 和英特尔® SGX "飞地"上运行。web 前端和 Redis 受到传输层安全 (TLS) 协议加密, 因此推理流水线中的数据 (包括用户数据和模型) 在存储、传输、使用的过程中都受到更多地保护。



图四 推理服务流水线

共创美好未来: 英特尔® DL Boost 加速端到端 PPML 解决方案

该解决方案执行如下端到端推理流水线:

1. RESTful http API 接收用户输入, Analytics Zoo pub/sub API 将用户输入转化成输入队列, 并由 Redis 管理。用户数据受加密保护。

2. Analytics Zoo 从输入队列中抓取数据。它在分布式流处理框架 (如 Apache Flink) 上采用推理引擎进行推理。英特尔® SGX 使用 Occlum 来保护推理引擎和分布式流处理框架。英特尔® oneAPI 深度神经网络库 (oneDNN) 利用支持 Int8 指令集的英特尔® DL Boost 提高分布式推理流水线的性能。

3. Analytics Zoo 从分布式环境中收集推理输出, 并送回到由 Redis 管理的输出队列。随后, 解决方案使用 RESTful http API 将推理结果作为预测返回给用户。输出队列中的数据和 http 通信内容都被加密。

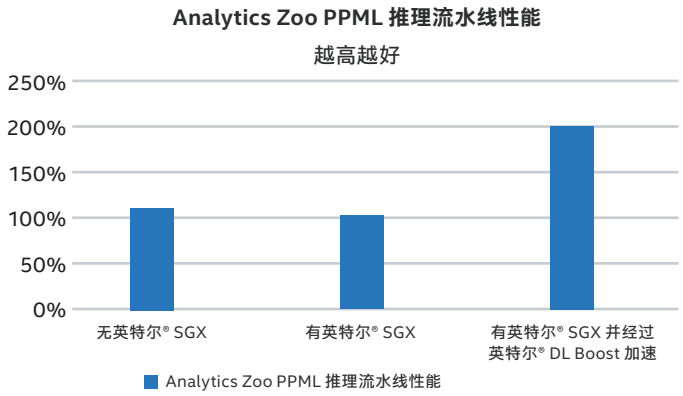
性能分析

Analytics Zoo PPML 解决方案的性能进行了验证。

服务器	• 服务器: 基于第三代英特尔®至强®可扩展处理器的服务器(2), 1024GB内存, 英特尔® SSD, 英特尔® 10GbE网络
系统软件	• Occlum用户态操作系统 • Ubuntu操作系统
应用软件	• Analytics Zoo • OpenVINO™ 工具套件 • 英特尔® oneAPI oneDNN 0.19 • Redis • Apache Flink
工作负载	• ResNet-50 深度学习模型

表一 测试配置

图五为测试结果。与不受英特尔® SGX 保护的推理流水线相比, 当推理解决方案受到英特尔® SGX 保护, ResNet50 推理流水线的吞吐量会有少许损失。而采用支持 INT8 指令集的英特尔® DL Boost 后, 受英特尔® SGX 保护的推理流水线吞吐量翻了一番。



图五 英特尔® SGX、英特尔® DL Boost 和第三代英特尔® 至强® 可扩展处理器提供高性能安全能力

基于英特尔® SGX 打造的 Analytics Zoo PPML 解决方案继承了受信任执行环境 (TEE) 的优点。和其它数据安全解决方案相比, 它的安全性和数据效用性十分突出, 性能方面仅略逊于纯文本。英特尔® DL Boost 和英特尔® oneDNN 则进一步提升了 Analytics Zoo PPML 推理解决方案的性能。表二总结了该解决方案 (TEE) 相对于同态加密 (HE)、差分隐私 (DP)、安全多方计算 (MPC) 和纯文本的优势。

	TEE	HE	DP	MPC	Plain 纯文本
安全性	★★★★★	★★★★★	★★★★	★★★★★	NA
性能	★★★★	★	★★★★	★★★	★★★★★
数据效用性	★★★★★	★★★★	★	★★★	★★★★★

表二 Analytics Zoo PPML 解决方案 (TEE) 与其他方案的比较

总结

在日益复杂的法律和监管环境中，对于企业和组织来说，保护客户数据隐私比以往任何时候都更加重要。在隐私保护机器学习的助力下，企业和组织就能在继续探索强大的人工智能技术的同时，面对大量敏感数据处理降低安全性风险。

Analytics Zoo 隐私保护机器学习解决方案基于 Occlum、英特尔® SGX、英特尔® DL Boost 和 Analytics Zoo 打造，为助力确保数据的安全性和大数据人工智能工作负载性能提供了平台解决方案。蚂蚁集团和英特尔共同打造并验证了这一 PPML 解决方案，并将继续合作探索人工智能和数据安全性领域的最佳实践。

测试配置

系统配置： 2 节点，双路英特尔® 至强® 铂金 8369B 处理器，每路 32 核心，超线程开启，睿频开启，总内存 1024 GB (16 个插槽 / 64GB / 3200 MHz)，EPC 512GB，SGX DCAP 驱动程序 1.36.2，微代码：0x8d05a260，Ubuntu 18.04.4 LTS，4.15.0-112-generic 内核，英特尔截至 2021 年 3 月 20 日的测试。

软件配置： LibOS Occlum 0.19.1，Flink 1.10.1，Redis 0.6.9，OpenJDK 11.0.10，Python 3.6.9

工作负载配置： 模型：Resnet50，深度学习框架：Analytics Zoo 0.9.0，OpenVINO™ 2020R2，数据集：Imagenet，BS=16 / 实例，16 个实例 / 双路，数据类型：FP32 / INT8

所有性能数据均为实验室环境下测试所得。

了解更多

- 英特尔® SGX
- 英特尔® DL Boost
- 第三代英特尔® 至强® 可扩展处理器
- Occlum
- Analytics Zoo