解决方案

# 英特尔® SGX 助力阿里云构建端到端隐私保护机器学习方案

## 概述

大数据、人工智能（AI）等数字化创新技术正在加速重塑世界面貌，其改变了社会经济发展模式，并给人们的生活带来更多可能。在此背景下，数据已经成为最重要的资产之一，越来越多的企业开展以数据为中心的变革。但同时，数据要素在存储、处理和流转等过程中，也面临着内外部安全风险的严重威胁。能否保护这些数据的安全，成为决定数字化转型策略成败的重要因素之一。

基于英特尔® Software Guard Extensions（英特尔® SGX）技术，英特尔打造了端到端的大数据和人工智能隐私计算解决方案 — BigDL PPML，并携手阿里云 DataTrust 平台联合验证了隐私保护机器学习的端到端工作流和相关业务场景，展示了基于 BigDL PPML 快速构建端到端隐私计算的最佳实践。

## 背景：面向大数据与 AI 的数据融通面临严峻安全风险

数字化转型凸显了数据的重要价值，加速了数据的流转，也带来了复杂的多方数据存储、流转与处理问题。特别是在 AI、大数据等应用中，由于单个组织难以掌握到 AI 与大数据应用所需要的全部数据，因此需要通过多方数据合作，实现数据资源的融通利用。例如，在训练金融 AI 算法时，单体金融机构的数据往往无法满足算法训练的需求，此时可以通过联合建模的方式，实现数据的共享，并共同创建、维护该 AI 模型。

由于跨机构、跨行业的数据融合、联合分析和建模的需求日趋增加，数据安全风险急剧增长。这一方面是由于数据本身可复制、易传播，在传统安全模式下，数据一经分享难以追踪。另一方面，数据持续流动会导致责任划分不明确、权限控制困难、以及问题难以追责等问题。保证数据的安全可靠成为重中之重。

但是，面向 AI 和大数据的传统安全防护方案常常会面临如下挑战：

- 联合分析、联合建模等应用涉及到频繁的数据共享、数据融合，而传统的数据安全解决方案可能有助于保护静态和传输中的数据，但难以保护正在使用的数据。这可能导致部分安全威胁突破安全防线，导致数据泄露等事件的发生。
- AI 和大数据应用涉及到数据输入、数据分析、机器学习、深度学习等多个流程，任何一个流程的漏洞都可能导致数据泄露等严重后果。因此，实现端到端的安全防护至关重要。
- 针对 AI 和大数据的攻击广泛涉及到已知与未知的安全威胁，覆盖多种攻击技术与工具。而传统方案往往从软件层面出发，难以保护硬件底层，这影响了其防护效率的进一步提升。
- 数据安全保护措施常会涉及到较为复杂的计算，可能会带来一定的性能损耗，对于数据银行的运行效率带来负面影响。

## 解决方案：基于英特尔® BigDL PPML 的阿里云端到端隐私保护机器学习

为了帮助企业在 AI 和大数据等应用中，更好地实现端到端的隐私保护，阿里云与英特尔合作，将英特尔® BigDL PPML 与阿里云 DataTrust 平台进行协同，联合验证了隐私保护机器学习的端到端工作流和相关业务场景。

### 英特尔® BigDL PPML

BigDL 是英特尔开源的统一的人工智能解决方案平台，数据科学家、数据工程师等开发者可以使用 BigDL 轻松创建端到端的分布式人工智能应用。BigDL 应用英特尔® SGX 可信硬件执行环境（Trusted Execution Environment, TEE），并集成了其他软硬件安全措施，构建了一个分布式的隐私保护机器学习（Privacy Preserving Machine Learning, PPML）平台，能够保护端到端（包括数据输入、数据分析、机器学习、深度学习等各个阶段）的分布式人工智能应用。

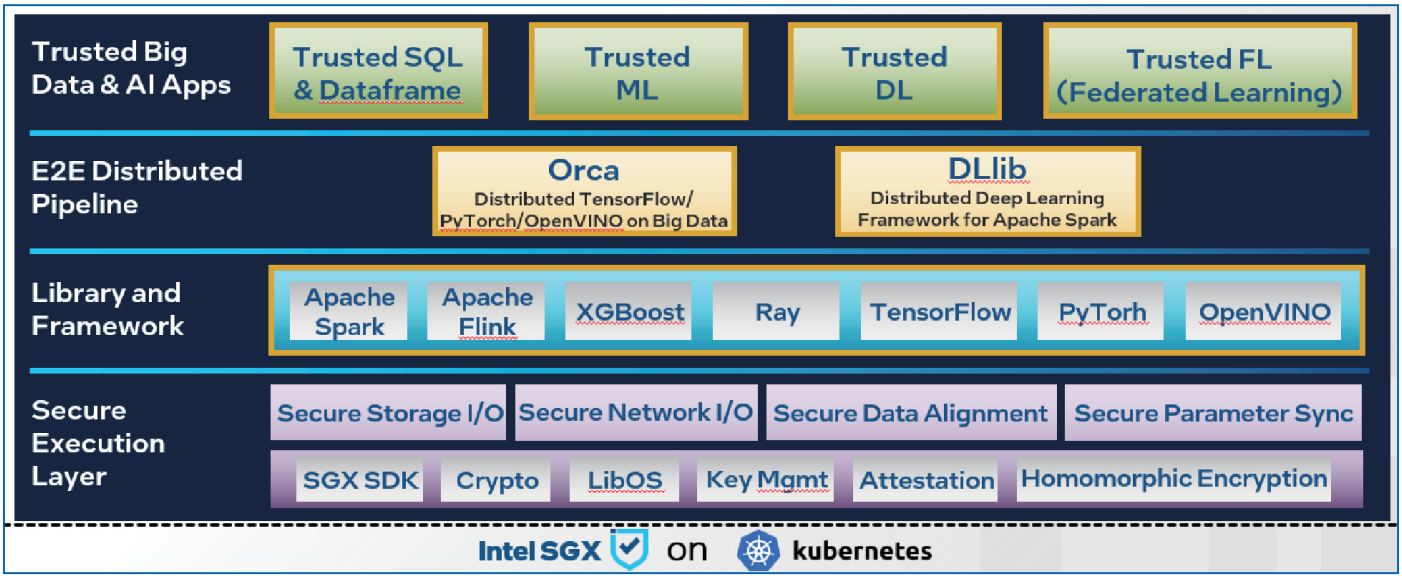


图 1. 英特尔® BigDL PPML 软件栈

作为英特尔® BigDL PPML 的重要基础技术，英特尔® SGX 通过绕过系统的操作系统和虚拟机软件层，能够更有效地抵御多种类型的攻击。它可显著加强数据安全，满足对于机密计算的广泛需求。英特尔® SGX 提供了一种基于硬件的内存加密机制，将内存中的特定应用代码和数据隔离开来。英特尔® SGX 允许为用户级代码分配专用内存区域（Encalve，安全飞地），以免受到拥有更高权限的进程的影响。

英特尔® SGX 经过了严格测试，是业界广泛部署的基于硬件的数据中心可信执行环境（TEE），大幅减少了系统中的攻击面。除了有助于防御基于软件的攻击外，英特尔® SGX 的验证机制还能够帮助用户确保应用程序及相关硬件没有受到攻击，且处理器安装了最新的安全更新。

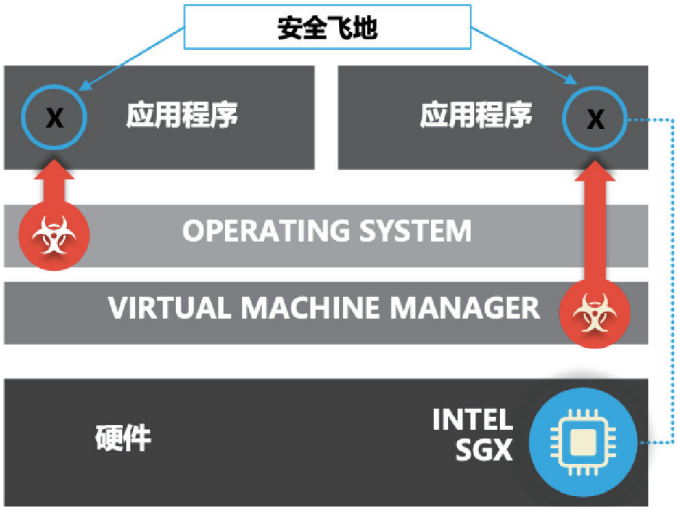


图 2. 英特尔® SGX 可在硬件底层提供保护

借助 BigDL PPML 平台，开发者可以：

- 在加密数据上开发并运行标准的分布式人工智能应用（如大数据分析、机器学习、深度学习等）。
- 利用基于硬件的安全技术技术（如英特尔® SGX）保护计算过程以及相应的内存数据。
- 为人工智能应用提供端到端的安全和隐私保护，例如：在具备英特尔® SGX 硬件能力的 K8s 环境中创建并认证可信的集群环境；通过密钥管理系统（Key Management System, KMS）为分布式数据提供加密和解密能力；通过英特尔® SGX、加解密技术、TLS 和安全认证等技术实现安全的分布式计算和数据通信。

## 阿里云 DataTrust 隐私保护计算平台

阿里云 DataTrust 是行业领先的基于可信执行环境、安全多方计算（Secure Multi-Party Computation, MPC）、联邦学习（Federated Learning, FL）、差分隐私（Differential Privacy, DP）等隐私增强计算（Privacy Enhancing Technique）技术打造的隐私增强计算平台，致力于实现数据价值的安全流动，为行业带来正确、易使用、高可用的安全数据流通产品。

阿里云 DataTrust 以英特尔® SGX 为底座，结合 MPC、FL 等技术，基于阿里云数据中台丰富应用场景实践，能够在保障数据安全的前提下完成多方数据联合分析、联合训练、联合预测，为企业提供立足数据业务原生的数据安全流通解决方案，助力企业业务增长。

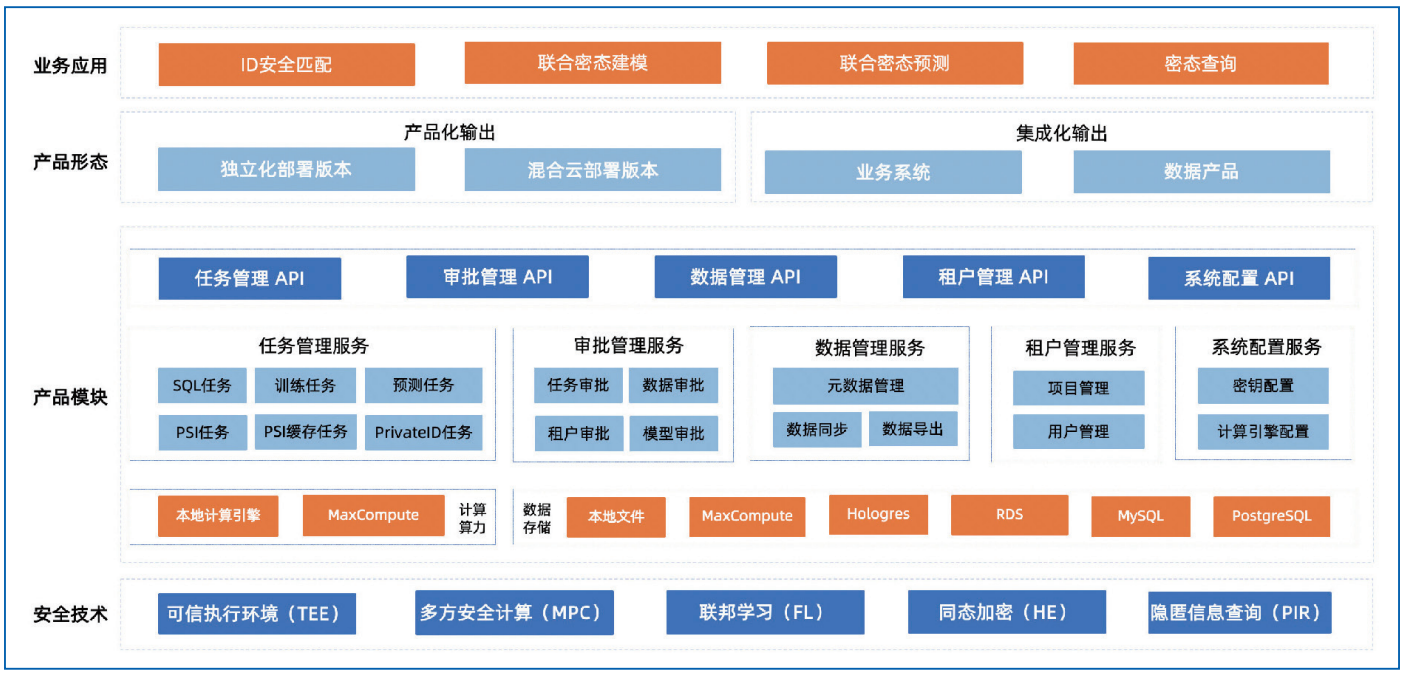


**图 3.** 阿里云 DataTrust 架构

## 端到端解决方案工作流程

基于隐私计算的核心功能，BigDL PPML 解决方案集成了端到端隐私保护计算工作流的更多组件，例如签鉴服务（Attestation Service），密钥管理（Key Management），以及基于 Kubernetes 的安全容器化部署方案。

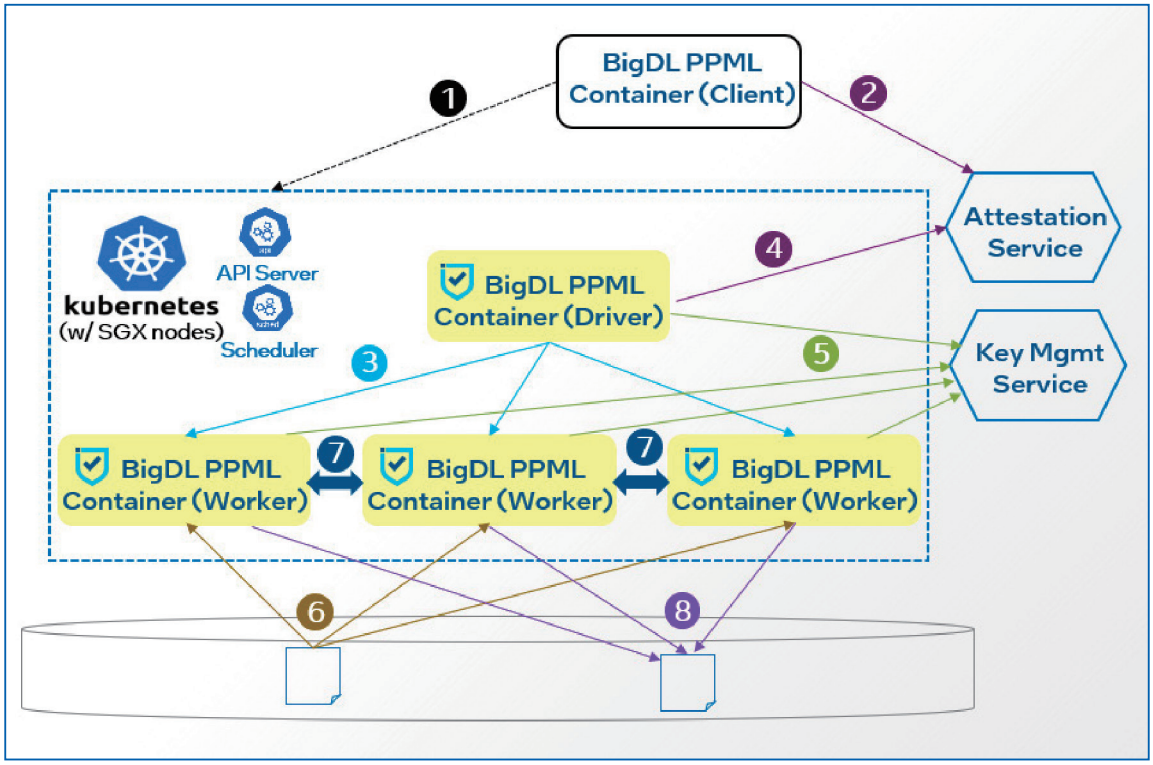


**图 4.** 基于 BigDL PPML 的端到端安全计算工作流程图

在基于 BigDL PPML 的端到端安全计算工作流中，各个流程的功能如下：


**图 5.** 基于 BigDL PPML 的端到端安全计算工作流程功能

BigDL PPML 提供了实现以上工作流的集成解决方案，包括：基于 SGX 的可信计算核心组件，支持 Apache Spark，Spark SQL 以及机器学习，深度学习等应用；经过抽象的认证服务客户端 API；经过抽象的密钥管理服务客户端 API；加密的数据传输和存储；以及定制的 K8s 容器镜像。

通过使用以上预制的工作流解决方案，开发者可以更加专注于业务逻辑的相关开发工作，利用 BigDL PPML 保障其应用的端到端安全性和隐私性。用户可以显著提高隐私计算应用的开发效率，大幅缩短实现隐私计算解决方案的开发时间（Time to solution）。

## 应用实践

阿里云 DataTrust 以运行 Spark SQL 为例，验证了 BigDL PPML 端到端解决方案。基本步骤如下：

### 1. 在阿里云上创建安全型 ECS 实例

创建带有加密内存的 g7t 实例，创建后的实例如图 6 所示。创建完成后，在实例列表中确认该实例规格：


**图 6.** ECS 实例规格

### 2. 准备 BigDL PPML 运行环境

首先部署 Kubernetes 集群、英特尔® SGX 插件、NFS 服务，获取 BigDL PPML docker 镜像，生成安全密钥和密码。随后进行 Kubernetes 安全配置，包括 RABC 配置及 Kuberentes secret 生成，最后启动 BigDL PPML 客户端容器。

### 3. 在 ECS 上利用 BigDL PPML 运行端到端安全保护的用户样例

进入 BigDL PPML 客户端容器，生成 appID、appKey 与 KMS 密钥，并使用 KMS 密钥加密输入数据，随后配置 spark-executor-template.yaml，将生成的加密数据与 KMS 密钥置于 NFS 路径下，最后向 Kubernetes 集群提交任务，运行样例程序。

通过上面的验证流程，在阿里云 ECS g7t.32xlarge 的实例上可以运行基于业界标准测试基准 TPC-DS 的查询语句。测试配置如下：

**表 1.** 测试配置

| 测试配置 | |
|---|---|
| **系统配置** | 3 节点集群（g7t.32xlarge Alibaba Cloud ECS 实例）<br><br>2x 英特尔® 至强® 铂金 8369B 处理器，64 核，启用超线程，256 GB 总内存，256 GB EPC，Ubuntu 20.04.2 LTS，5.17.0 kernel |
| **软件配置** | BigDL 2.1.2-SNAPSHOT，LibOS Graphene commit 1b8848b，Spark 3.1.2，Java 1.8.0_192 |
| **负载配置** | TPC-DS based queries implemented by databricks' spark-sql-perf |

图 7 为 100 GB 数据规模下，基于 TPC-DS 的查询语句在三个 ECS 组成的集群上的运行时间比较。以 99 个查询语句花费时间的几何平均作为度量标准，基于英特尔® SGX 硬件保护方案下的 BigDL PPML 方案的运行时间是完全未受英特尔® SGX 保护情况下的 1.89 倍[1]。

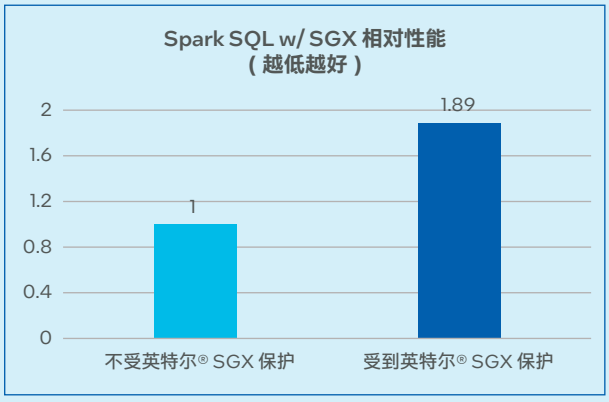


**图 7.** Spark SQL 基于英特尔® SGX 保护下的相对性能

从测试数据可以看出，虽然在开启英特尔® SGX 硬件保护之后，会带来一定的性能损耗，但是这一性能损耗在可接受范围内。而且英特尔® SGX 带来的性能损耗通常显著低于传统安全方案，能够在保护数据安全的同时，节约算力资源。

[1] 数据援引自阿里云于 2022 年 6 月开展的测试。测试配置：3 节点集群（g7t.32xlarge Alibaba Cloud ECS 实例），2x 英特尔® 至强® 铂金 8369B 处理器，64 核，启用超线程，256 GB 总内存，256 GB EPC，Ubuntu 20.04.2 LTS，5.17.0 kernel。分别测试在启用或不启用英特尔® SGX 下的运行时间。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。

## 收益：推动数据价值安全流动

基于英特尔® BigDL PPML 的阿里云端到端隐私保护机器学习解决方案继承了可信执行环境（TEE）的优点。和传统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，同时性能损耗较低。

通过应用该方案，企业能够构建端到端的安全保护流程，在数据输入、数据分析、机器学习、深度学习等 AI、大数据应用的多个阶段建立安全防护能力，避免安全威胁乘虚而入。同时，该方案实现了基于硬件底层的保护，具备更高的数据保护等级，能够防护传统安全方案难以抵抗的攻击形式，降低重要数据泄露的风险。

在该方案的支持下，企业能够提供安全的数据融通服务，联合分析、联合训练、联合预测等应用不透露原始数据以及基于数据应用逻辑层面的授权，保证场景化的数据融通安全需要；满足商业的自主性、可控性、安全性，为客户提供透明可控的安全流通环境，可随时管控和退出，永保数据控制权。同时，该方案运用前沿的安全技术，面向业务场景封装，可解决企业数据业务原生的海量级数据安全融通。

以下是该方案的典型应用场景：

- **全域精细运营：** 品牌方通过联动平台、第三方等全域数据，在保护个体隐私及数据安全的前提下，构建品牌数智化运营能力，优化人货场的配置，拉动业务增长。
- **联合智能风控：** 行业或企事业单位在原始数据不出自身环境的前提下，通过隐私增强计算技术，实现与多方数据的联合风控，提高风控识别有效性，助力业务良性增长。
- **广告搜索推荐：** 在保护消费者隐私与一方、二方数据安全的前提下，通过数据加持进行联合建模，提升算法准确率，提高广告投放有效性，推动业务持续高效增长。

## 总结与展望

随着数据安全和隐私保护的相关法律法规不断出台，对组织而言，保护客户数据的安全和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，组织能在继续探索强大的人工智能技术的同时，降低大规模敏感数据处理和分析的安全性风险。

BigDL PPML 隐私保护机器学习解决方案基于英特尔® SGX，BigDL 以及众多安全相关的组件共同打造，为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。阿里云和英特尔共同验证了 BigDL PPML 解决方案的工作流程。该合作展示了应用 BigDL PPML 开发端到端隐私保护应用的最佳实践，体现了 BigDL PPML 在加速开发隐私保护应用的显著作用。双方将在当前合作成果的基础上，进一步强化端到端隐私保护方面的创新与实践，帮助用户实现更加安全的数据融通，在安全的基础上加速数据价值挖掘。

intel 阿里云