intel | DataFun.

隐私计算在大数据AI领域的应用实践

龚奇源 资深架构师

目录 CONTENT

intel | DataFun.

intel | DataFun.

01

隐私计算

隐私计算背景

个人的需求

- 隐私和安全的意识提高

隐私和安全合规要求

- 国外：欧盟GDPR，美国CCPA等
- 国内：网络安全法，数据安全法，个人信息保护法等

隐私计算背景

GDPR Fines Tracker & Statistics

Total Number of GDPR Fines	Total Amount of GDPR Fines
1087	**€1,631,665,322**

Largest Fine	Smallest Fine
€746,000,000	**€28**
Amazon Europe Core S.a.r.l. on July 22 , 2021 - Luxembourg	Unknown on November 18 , 2020 - Hungary

Most Recent GDPR Fines

*Only includes finalised cases

DATE	ORG	FINE
06/17/2022	Mayr Melnhof Packaging Romania S.R.L.	€1,500
06/08/2022	Wens Experience SRL	€1,500
06/06/2022	Esselmann Technika Pojazdowa Sp. z o.o. Sp. k.	€3,500
06/03/2022	Lodeju, S.L.	€3,000
06/03/2022	Private individual	€360

TOP 5 BIGGEST GDPR FINES

*Only includes final & binding fines

Amazon Europe Core S.a.r.l.	€746,000,000
WhatsApp	€225,000,000
Google LLC	€90,000,000
Facebook Ireland Ltd.	€60,000,000
Google Inc.	€50,000,000

All data is from official government sources, such as official reports of national Data Protection Authorities.

https://www.privacyaffairs.com/gdpr-fines/

总数：1087 (增长中)
总额：~110亿
单次最高：~50亿
前五罚款：Amazon, WhatsApp, Google, Facebook

intel | DataFun.

隐私计算现状

隐私计算成为热点

- 大量企业和投资涌入
- 大量研究成果
- 安全和隐私技术蓬勃发展
  - 差分隐私 (DP)
  - 可信执行环境 (TEE)
  - 同态加密 (HE)
  - 安全多方计算 (SMC)
  - 联邦学习 (FL/FML)

人民网 people.cn

首页 党政 要闻 观点 互动 可视化 地方 举报专区 多语言 合作网站

人民网 >> 经济·科技

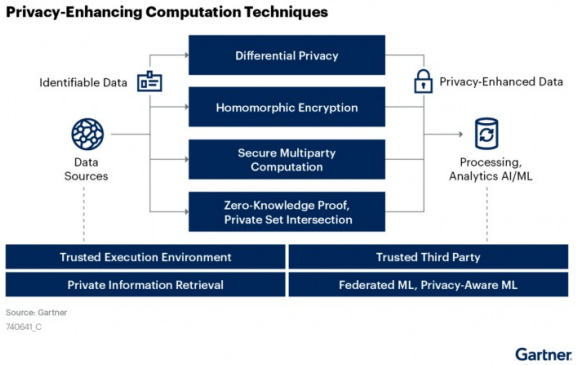
兼顾安全与应用 2021年中国隐私计算市场规模4.9亿元

陈果静

2022年04月24日09:15 | 来源：经济日报 T₁ 小字号

在数字经济时代，随着人工智能、大数据等产业的快速发展，数据要素的作用及其重要性愈发凸显。与此同时，大量挖掘和收集用户数据与信息，金融领域用户数据信息泄露事件多发。

随着《数据保护法》《个人信息保护法》等的落地，金融业如何做到既保护数据安全，又充分发挥数据资产价值、高效链接多方数据，成为一项亟需解决的课题。



http://finance.people.com.cn/n1/2022/0424/c1004-32407072.html

https://www.gartner.com/en/documents/4006926

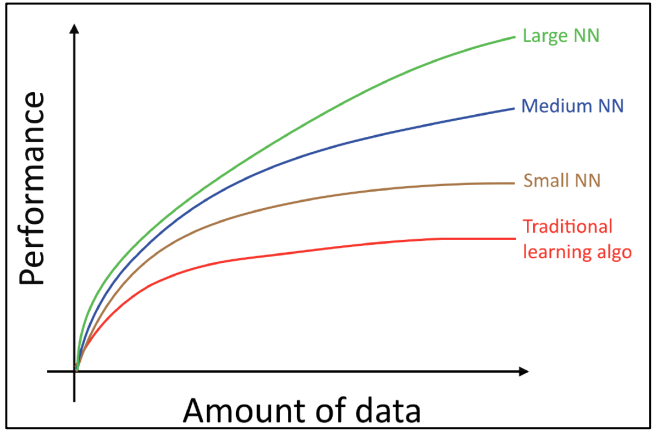
intel | DataFun.

02

大数据AI+隐私计算

大数据AI背景

- 大数据框架和技术已经大规模普及
  - 易用性提高
  - 方向逐步细化
    - 存储、处理更多数据
    - 分析（查询）更多数据
    - 实时分析
    - 建模和预测 (机器学习、深度学习)



“Machine Learning Yearning”, Andrew Ng, 2016

- AI 无处不在
  - 从实验室走向生产环境
  - 应用于大规模、分布式大数据

大数据AI背景

获取 / 存储 → 清洗 / 准备 → 分析 / 建模 → 部署 / 可视化

集成的数据流水线

Spark Streaming
hadoop HDFS
APACHE HBASE
APACHE kafka A distributed streaming platform
APACHE KUDU
APACHE Spark
Flink
RAY
Spark SQL
Spark MLlib
TensorFlow
ANALYTICS ZOO
Qlik
BigDL

数据管理

数据分析

数据科学及人工智能

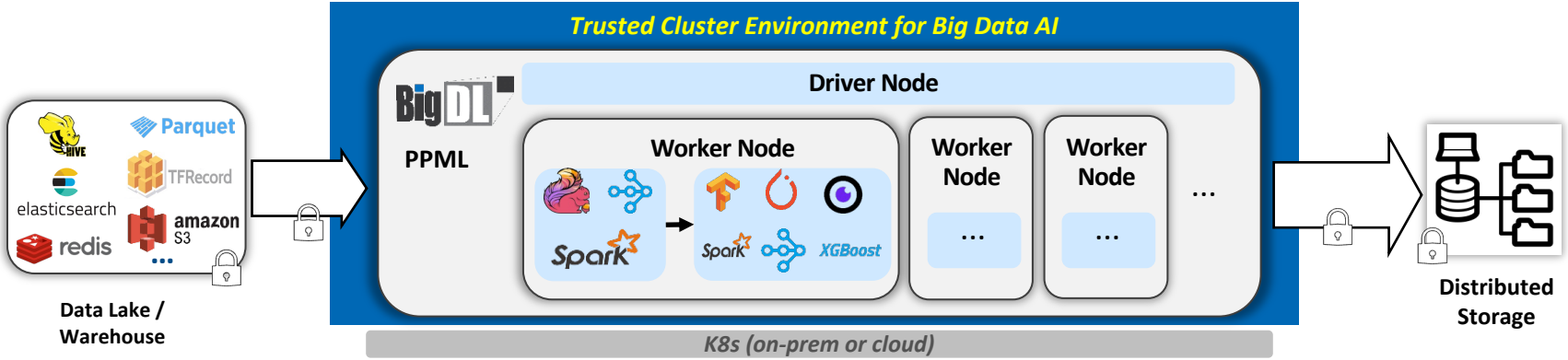
intel | DataFun.

常见痛点：

- 能否兼容现有的应用
  - 现有的应用(数据分析和AI)能否直接迁移
  - 对其他应用和设施是否有冲击
- 能否处理大规模数据
  - 能否支持大规模数据
  - 计算效率是否足够好
- 能否解决数据孤岛问题

BigDL PPML: 可信的大数据AI

HW (SGX/TDX) Protected Secure Big Data AI, even on Untrusted Cloud

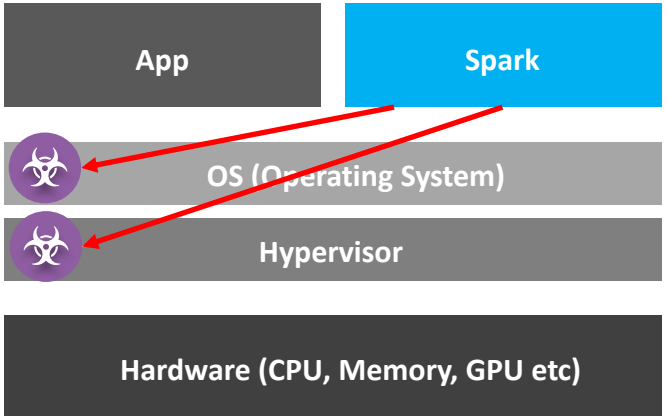


- **Standard, distributed AI applications on encrypted data**
- **Hardware (Intel SGX/TDX) protected computation (and memory)**
- **End-to-end security enabled for the entire workflow**
  - *Provision and attestation of "trusted cluster environment" on K8s (of SGX nodes)*
  - *Secrete key management through KMS for distributed data decryption/encryption*
  - *Secure distributed compute and communication (via SGX, encryption, TLS, etc.)*

intel | DataFun.

Apache Spark中的安全

- 网络加密（TLS/AES）
- 存储加密（AES）
- 计算（明文）

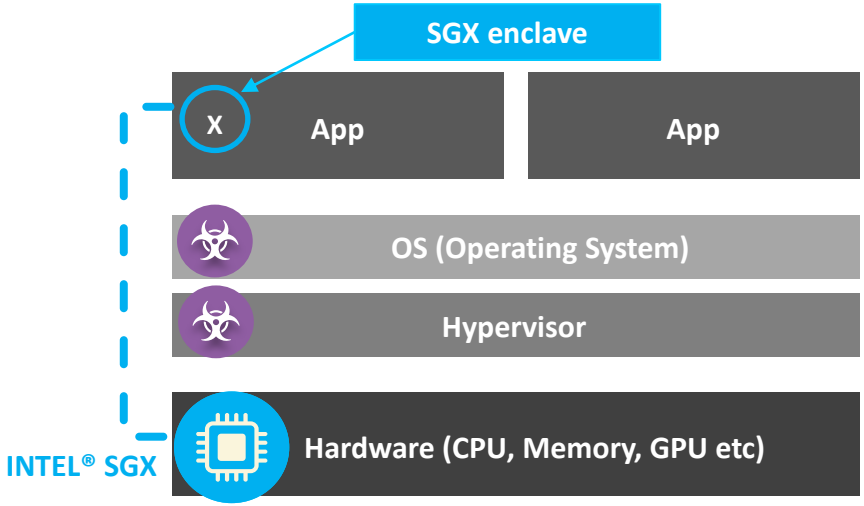


If OS/VM/Hypervisor/BIOS is hacked by adversaries, then they can dump sensitive data (input, temp, output etc) from Spark.

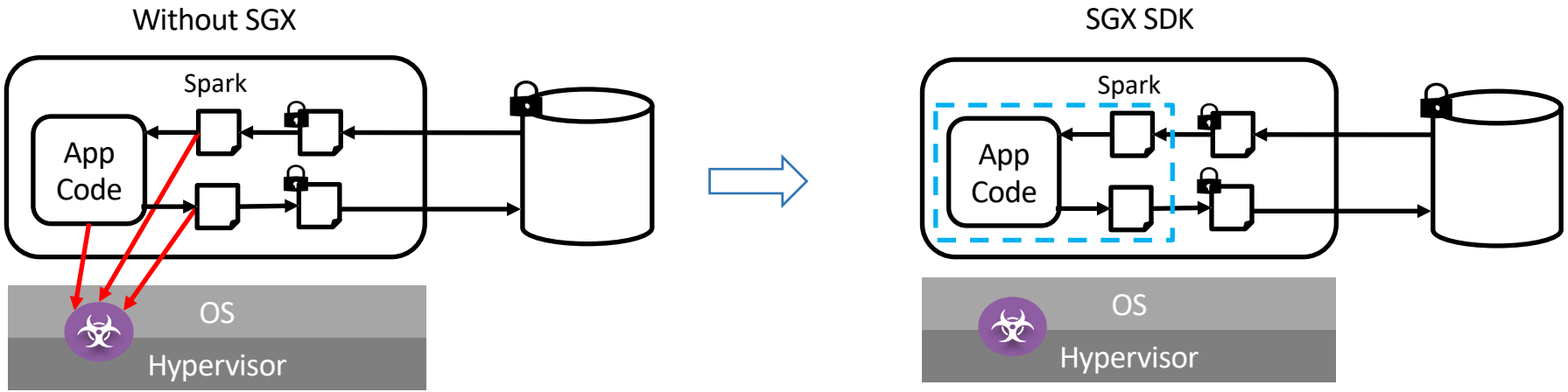
英特尔软件防护扩展SGX

- 硬件级的可信执行环境(TEE)
- 相对小的攻击面
- 性能影响小
- 足够大的飞地（最大1TB）

已经被广泛测试、研究和部署



Secure Spark with SGX



攻击者可以获取到应用和敏感数据

保护明文和敏感模块

攻击者无法获取明文数据

缺点：

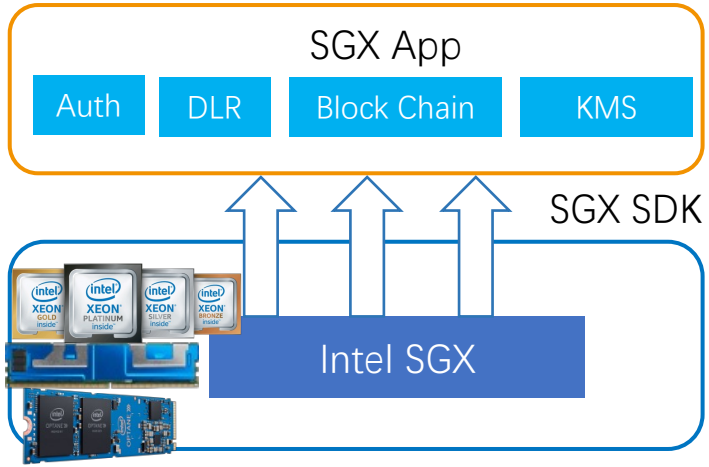
- 开发代价大
- 代码无法复用

Running in SGX

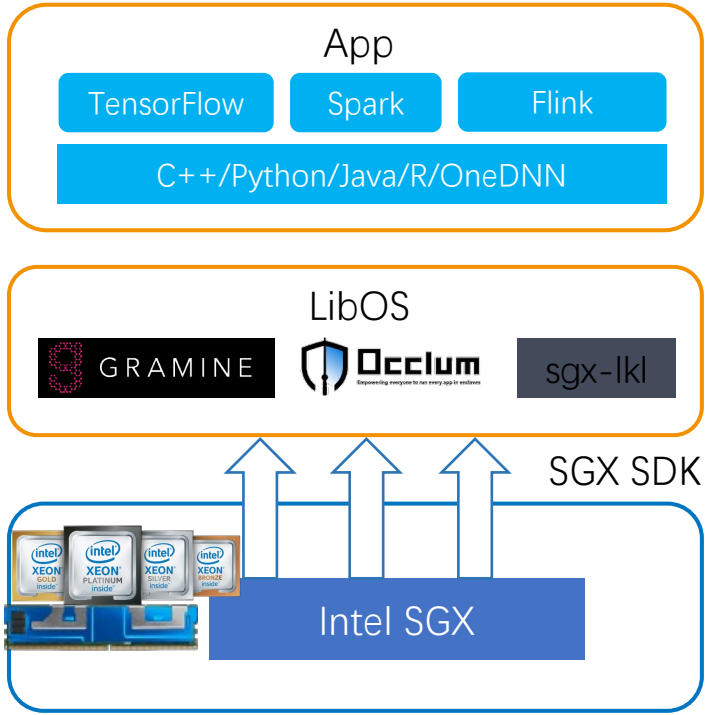
https://github.com/mc2-project/opaque-sql

大数据AI+隐私计算

安全

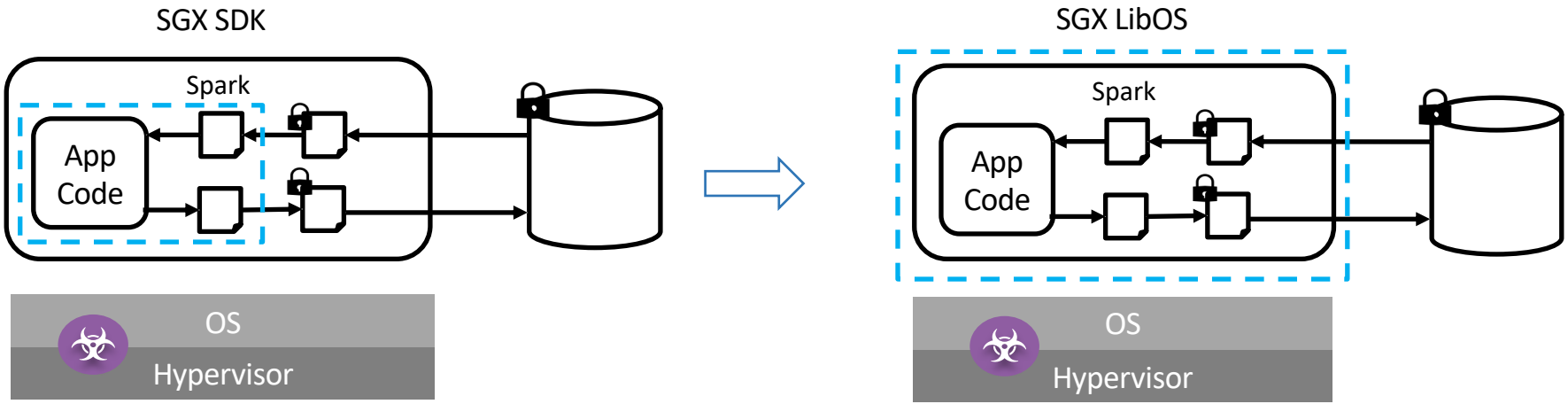


安全+易用性



intel | DataFun.

Running unchanged Spark Applications in SGX



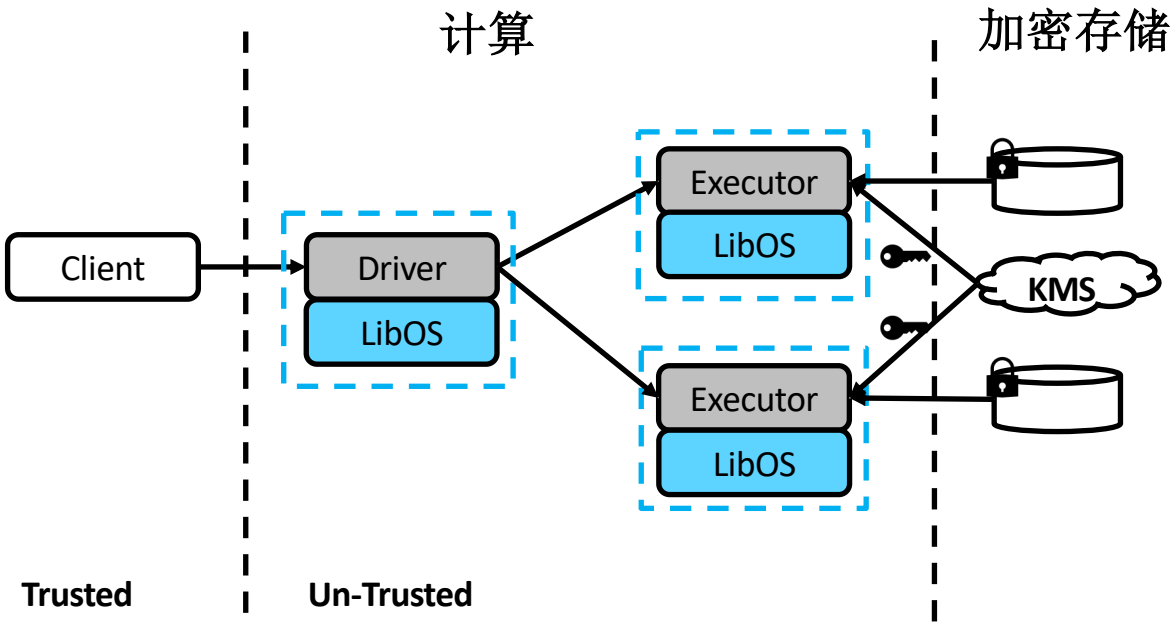
保护明文和敏感模块

保护整个Spark

优点：不需要修改Spark和Spark应用

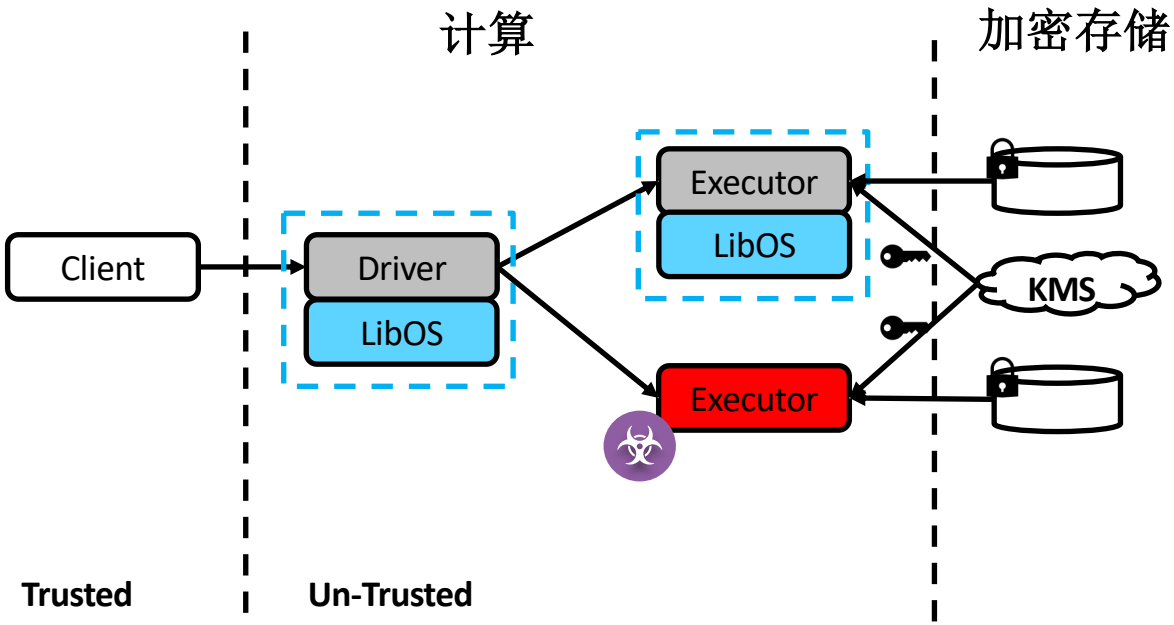
Running in SGX

Running unchanged Spark Applications in SGX



Running in SGX

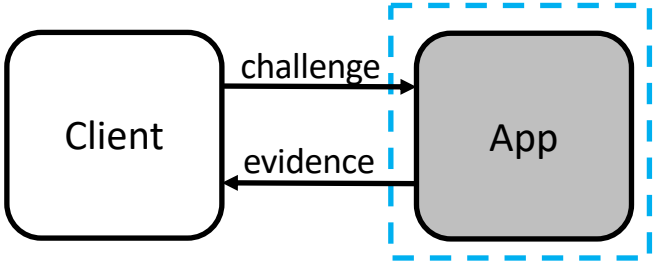
Attack on distributed Spark



Running in SGX

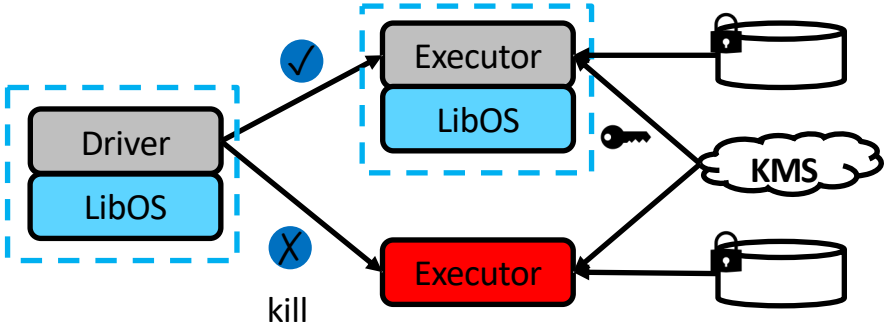
远程证明保证应用的完整性

- Attestation in short: Verify if an application is running in SGX
  - Application is expected
  - Within SGX
  - Running env is secured
  - …

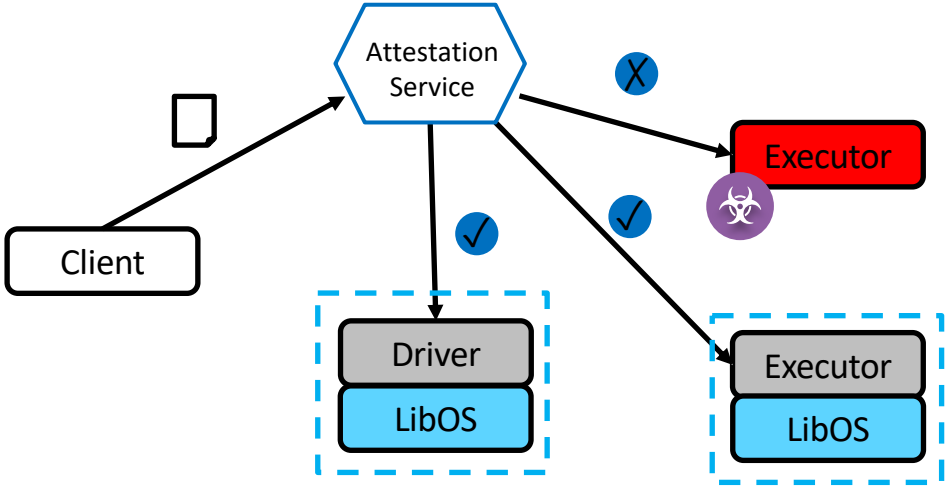


- Attestation result (verify evidence/quote)
  - Look good ✓
  - Not good ✗

远程证明保证应用的完整性



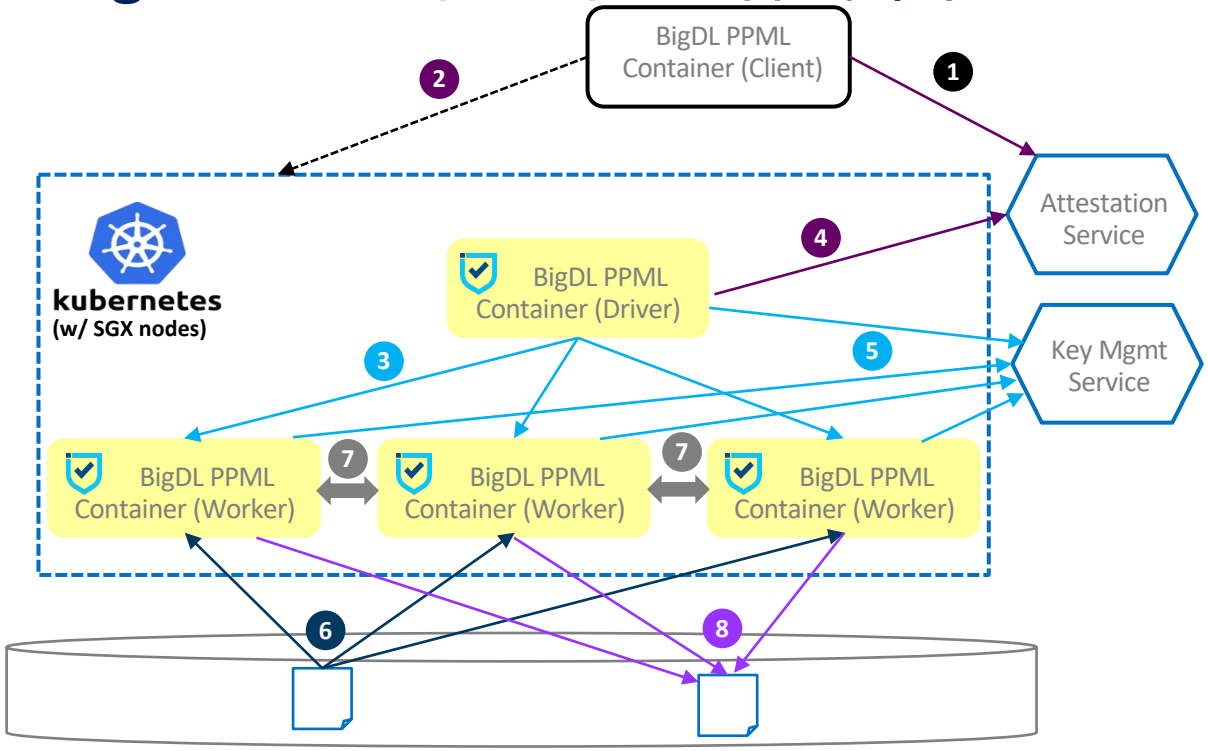
需要修改Spark的注册和Submit



无需修改Spark和Spark应用

大数据AI+隐私计算

BigDL PPML端到端一站式架构



1. User submits Policy
2. User submits job to K8s (using BigDL PPML CLI), which creates the driver node Driver creates more worker nodes
3. AS attests Driver/executor
4. Driver and workers request keys from KMS
5. Workers read and decrypt input data
6. Workers run distributed Big Data, ML and DL programs
7. Workers encrypt and write output data

大数据AI+隐私计算

BigDL PPML Workflow

集群管理员 | 开发者/数据科学家

Step 0 Deployment

- Set up K8s cluster
- Set up K8s-SGX plugin
- Set up Attestation service
- Set up KMS (key management service)

Step 1 Preparation

- Upload BigDL PPML docker image to K8s registery
- Encypt and upload data
- Submit Policy

Step 2 Build App

- Buid standard Big Data and ML applications (Spark XGBoost)
- Optionally use BigDL PPML APIs (ctypto, VFL, etc.)

Step 3 Submit Job

- Use BigDL PPML container and CLI to submit job to K8s

Step 4 Read Result

- Decrypt and read result of the job

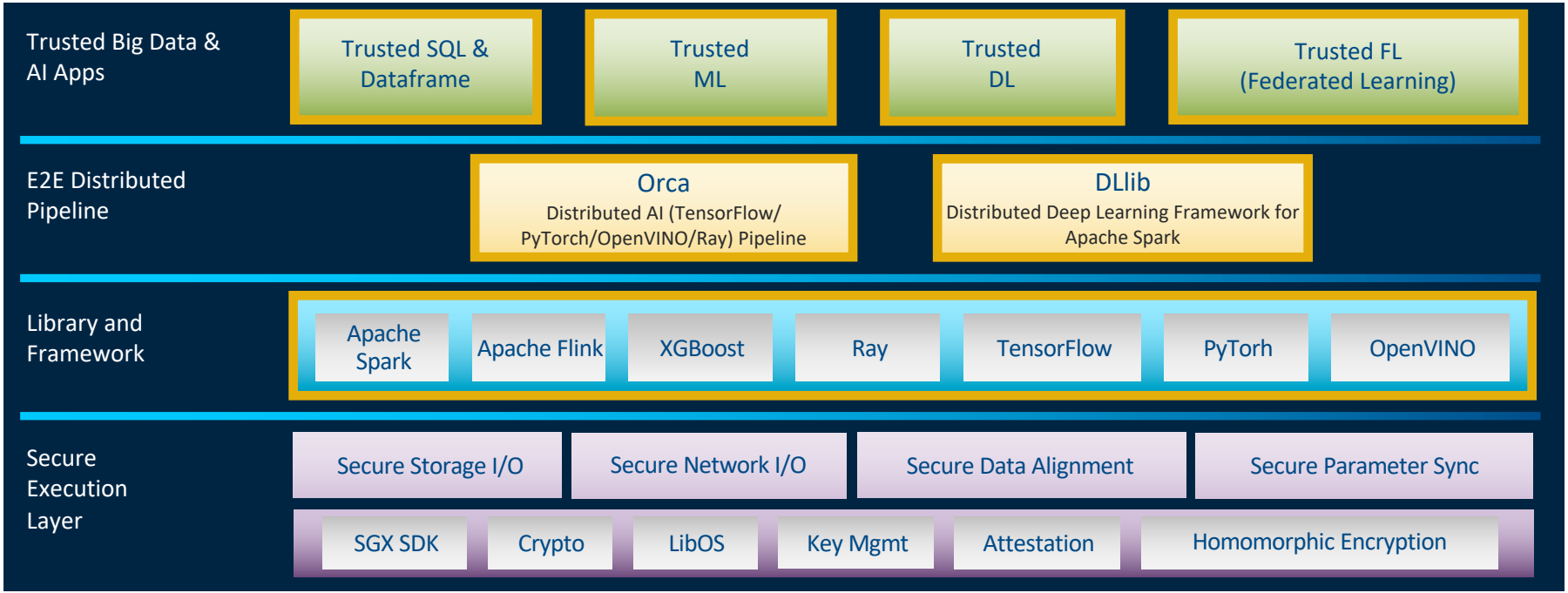
SGX相关的准备和开发

正常的建模和查询

https://hub.docker.com/r/intelanalytics/bigdl-ppml-trusted-big-data-ml-scala-occlum

BigDL 隐私保护的机器学习

Secure & Trusted Big Data and AI



Intel SGX on kubernetes

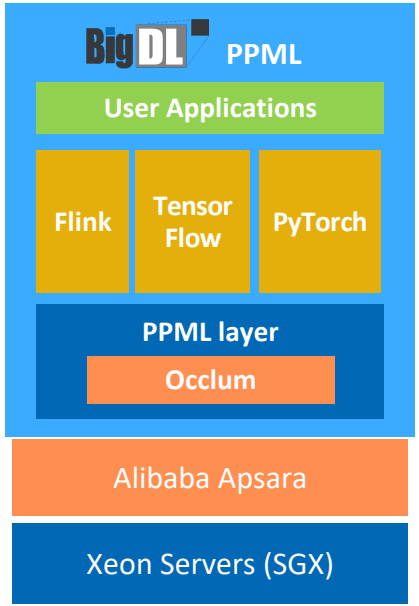
intel | DataFun.

intel | DataFun.

03

应用实践

实时的流计算-天池大赛



	Status	Season2	Teams
第三届 Apache Flink 极客挑战赛暨AAIG...	In Progress	2021-11-09	4537

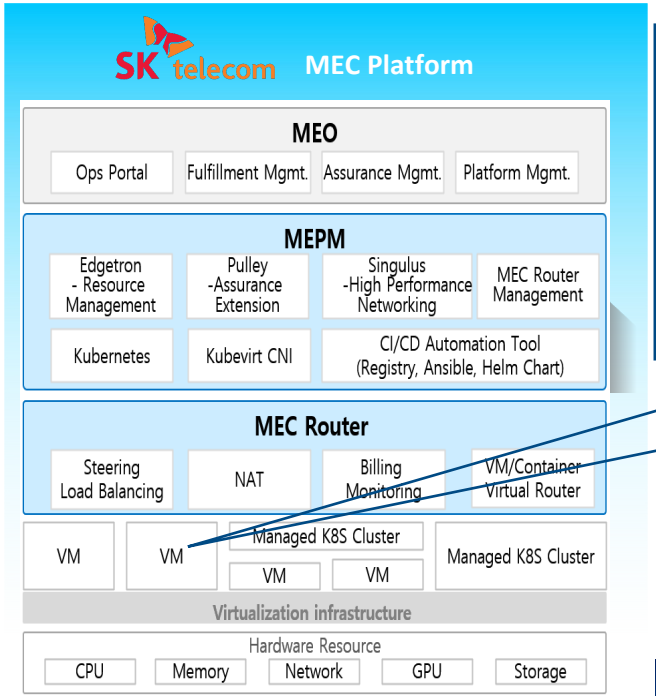
Alibaba, Intel and Occlum community co-host Kaggle-like PPML competition for spam detection in online e-commence recommendation.

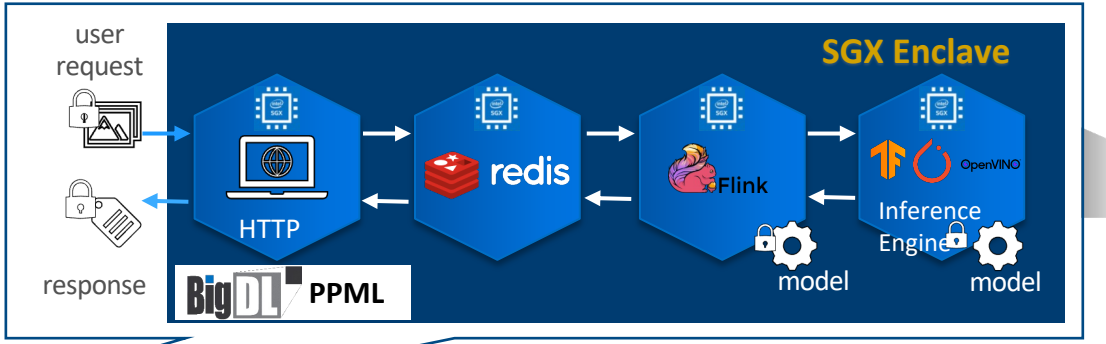
4500 Teams	Building PPML Applications	100+ IceLake Instances	Deployed on Alibaba Cloud

https://tianchi.aliyun.com/competition/entrance/531925/introduction

intel. | DataFun.

实时的流计算-SKT





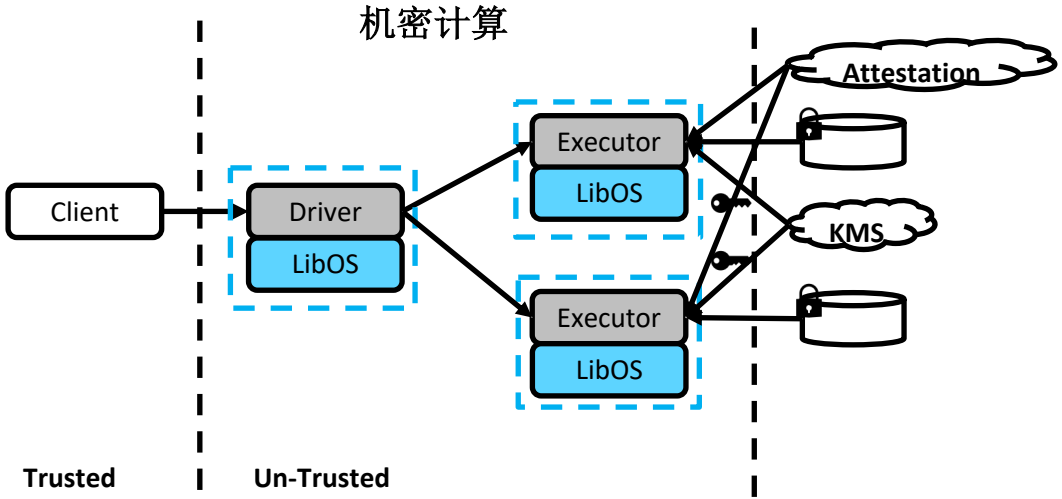
SKT Mobile Edge Computing provides common 5G services at the edge of the mobile telecommunication network. This POC runs *Trusted Model Serving* on BigDL PPML, providing secure, real-time, distributed DL model inference service across a cluster of Ice Lake servers

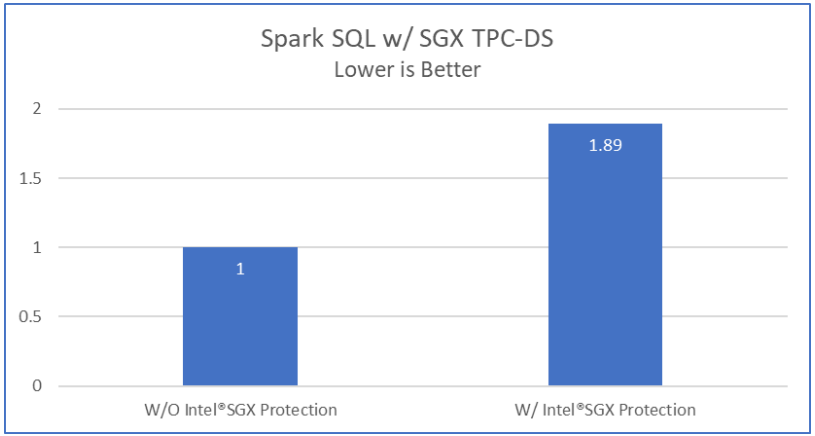
1300 image/sec	Secure Inference per MEC VM
<5% overheads	E2E Inference Pipeline Overhead

https://networkbuilders.intel.com/solutionslibrary/reference-architecture-for-confidential-computing-on-skt-5g-mec

intel | DataFun.

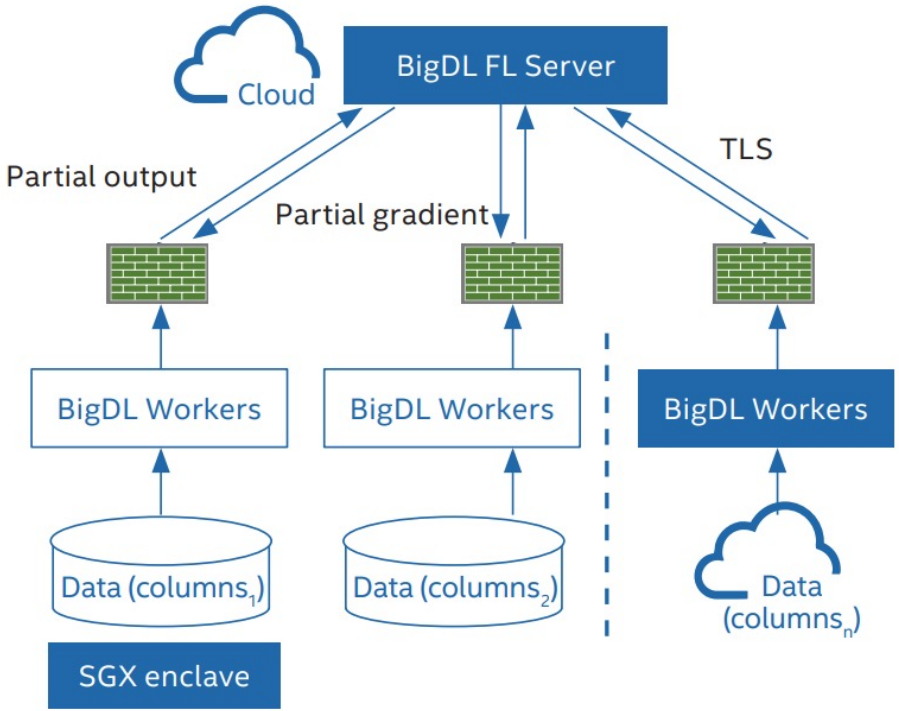
大规模数据分析SparkSQL TPC-DS





Spark SQL w/ SGX Performance Comparison®

联邦学习



Trusted Federated Learning

- Build united model across different parities
  - Training data remain local
  - Aggregation temp/partial results
- Secured computation environment with SGX

Win-Win for all parties

- End users
- Enterprises
- Cloud Service providers

https://www.intel.cn/content/www/cn/zh/now/data-centric/sgx-bigdl-financial-big-data.html

intel | DataFun.

04

总结和展望

总结和展望

隐私计算+大数据AI

- 若干痛点
- 用SGX构建安全的执行环境
  - LibOS帮助应用无缝迁移
  - 保证性能影响最小
  - 能够支持大规模数据
  - 联邦学习解决数据孤岛

BigDL PPML构建一站式的隐私计算方案

总结和展望

TEE发展趋势

- 易用性
  - TDX/Realm/SEV-SNP，机密容器
- 安全性：TEEOS, Micro kernel
- 拓展性
  - IO的支持
  - 加速器的支持：GPU/QAT/FPGA

非常感谢您的观看

intel | DataFun.