蚂蚁集团 ANT GROUP | DataFun.

隐私计算新思路：可信密态计算

潘无穷 蚂蚁科技集团股份有限公司 高级专家

目录
CONTENT

蚂蚁集团
ANT GROUP

DataFun.

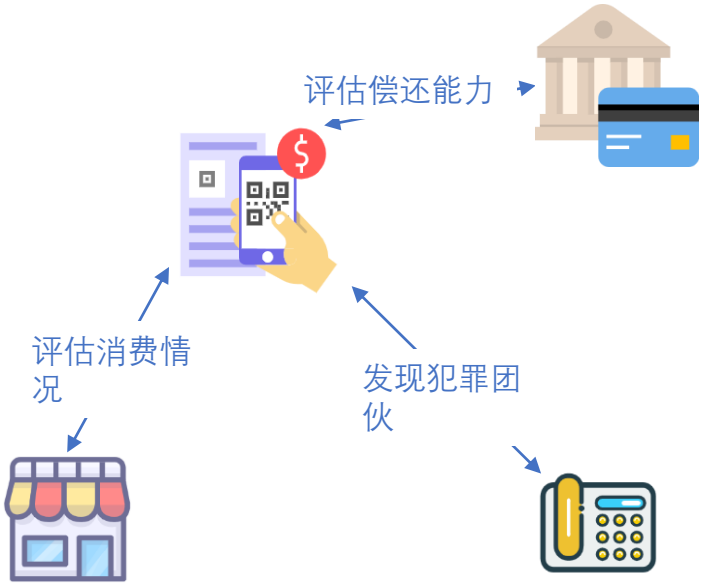
01 背景

哪些客观事实促使TECC的提出

迈入密态时代

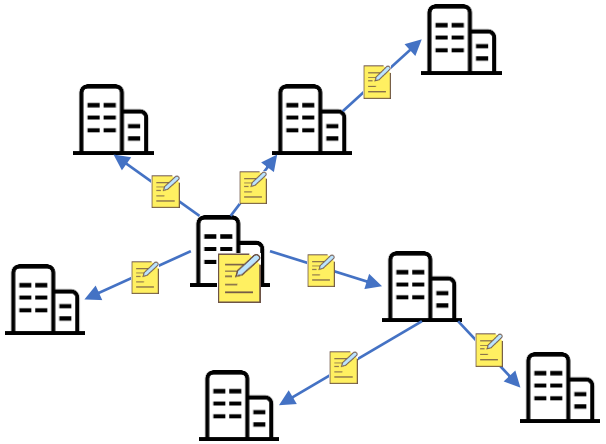
数据流通是必然的

• 同行业 • 跨行业 • 稀缺数据



密态流通是主要形式

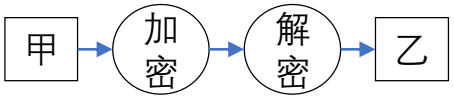
• 安全 • 合规 • 商业



明文传播导致泛滥

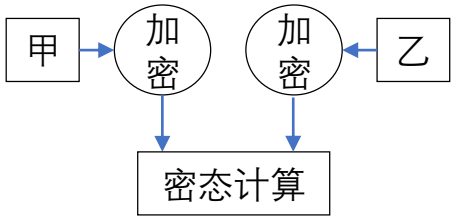
密态计算的难度远远超过传统密码算法

传统加密：在密文上做传输



解密

隐私计算：在密文上做计算



机器学习、数据分析

除、log、e^x、sinx、∫；地址访问、查询、排序、联合

加、乘；异或、与；比较、移位、选择

(类型转化)

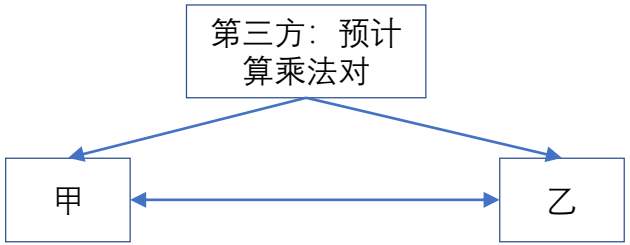
- 不是天然支持所有算子
- 支持的耗时不一致
- 底层差异导致上层需重新设计

部分算法存在恶意敌手攻击、合谋攻击

恶意敌手：攻击者不遵守协议。部分算法假设不会出现敌手。

会不会出现恶意敌手？
答：大概率会的，攻击者没有道理去遵守协议。

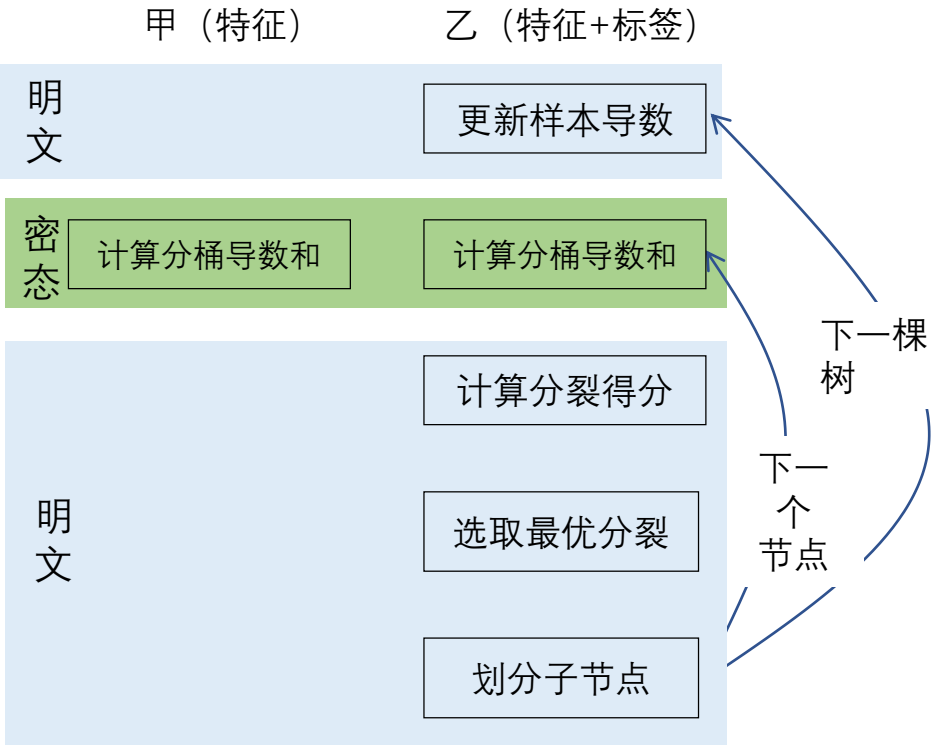
合谋攻击：多个攻击者合作。部分算法假设不会出现合谋攻击。



会不会出现合谋攻击？
答：在数据利益足够大的时候会。

部分算法存在信息熵泄露

某著名的隐私XGB算法：



信息熵泄露：原始信息的运算结果被泄露，导致原始信息的不确定性减少。

信息熵泄露会不会带来安全问题？
答：视规模而定

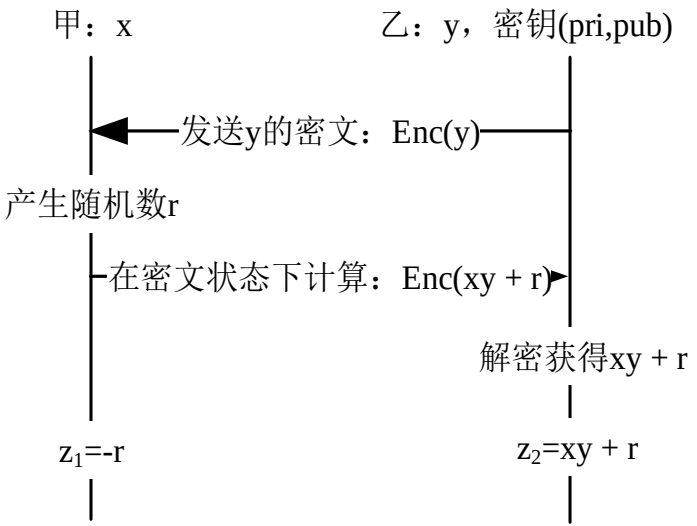
① 量小，风险可控。

② 量大，大概率存在有效攻击形式。

③ 泄露量会累加，长期可能很大

性能是算法类隐私计算技术最大障碍

一个抵御合谋攻击的密态乘法：$z_1+z_2=xy$



密态耗时分析：

- 加解密：1ms级
- 网络往返：10ms级
- 传输耗时：100us级（按1000比特、10mbps 网估算）

明文耗时：

- 一个CPU时钟周期，大概0.3ns

耗时膨胀5-7个数量级，少量场景适用，不适合大规模的密态数据中心。

系统实现漏洞

软件层

侧信道攻击

- SGAxe：CVE-2020-0549

内存安全、开放的动态反序列化（ODD）等

- CVE-2020-25459

硬件层

分支预测类攻击

- Specture：CVE-2017-5753、CVE-2017-5715、CVE-2017-5754、CVE-2018-3640、CVE-2018-3639、CVE-2018-3693

微体系结构攻击

- LVI：CVE-2020-0551
- CrossTalk：CVE-2020-0543
- ZombieLoad：CVE-2019-11135、CVE-2018-12130
- Fallout：CVE-2018-12126
- RIDL：CVE-2018-12127
- Lazy FP state restore：CVE-2018-3665

供应链

- Plundervolt：CVE-2019-11157

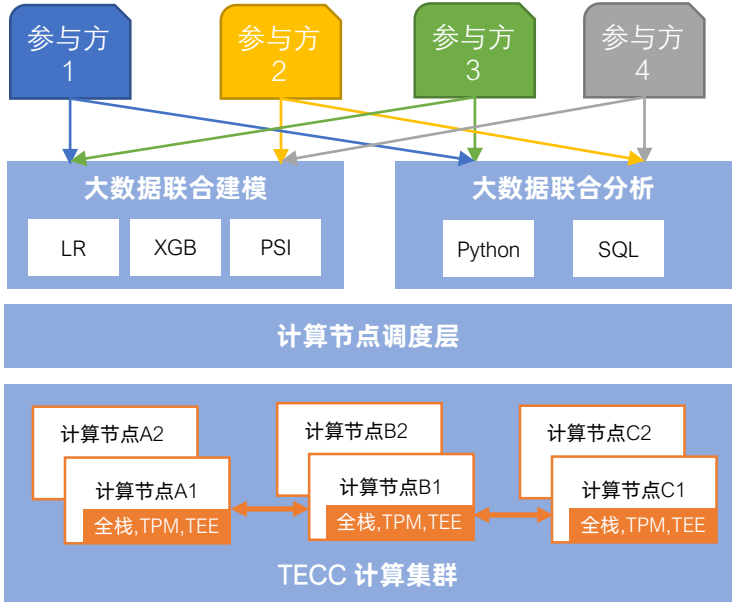
蚂蚁集团 ANT GROUP | DataFun.

02 TECC技术路线

TECC的主要性质是什么样的

可信密态计算TECC

TECC通过在多个高速互联的可信执行环境中运行密码协议，将两者有机地结合在一起，安全性可抵御现实攻击，成本低于一个量级，性能和稳定性接近明文。



主要原理：

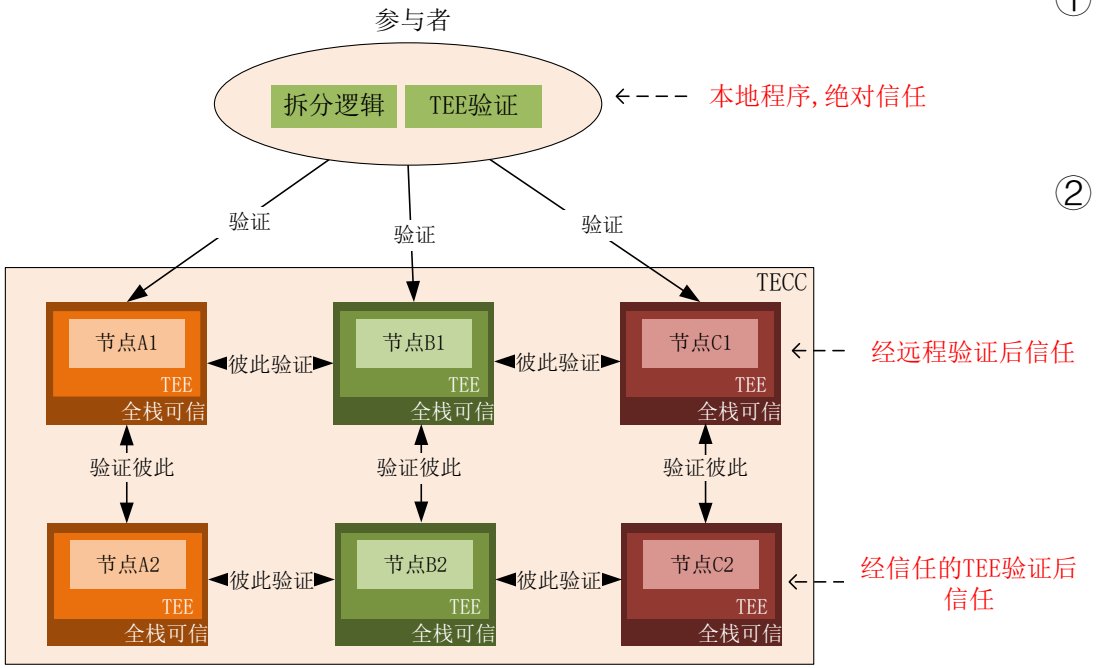
① 参与方将数据密态拆分，将每个分量传递给不同分区的可信计算节点。

② 每个可信计算节点只有分片数据，多个TEE分区通过密码协议完成目标计算。

③ 可信计算节点受TEE、TPM、全栈可信保护，运营者无法窥探。

④ 密码协议的同一个角色由一个TEE分区集群承担，可以进行并行化加速。

TECC整体安全模型

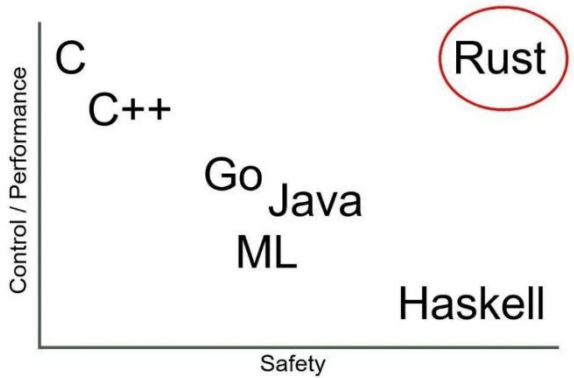


① 参与者通过TEE远程认证技术确认TECC节点行为如预期。

② 全栈可信、TEE、密态分片形成三层纵深防御，既防黑客也防运营者。

- 全栈可信阻断攻击TEE的前提条件。
- 密态（分片）形式消除TEE的大部分安全隐患。
- TEE可防御恶意敌手攻击、合谋攻击。

TECC实现安全性（软件层）



TECC的核心代码是Rust语言。采用Rust编程语言需要一些额外的开发成本，包括注明变量的所属关系及所属关系的转移等。

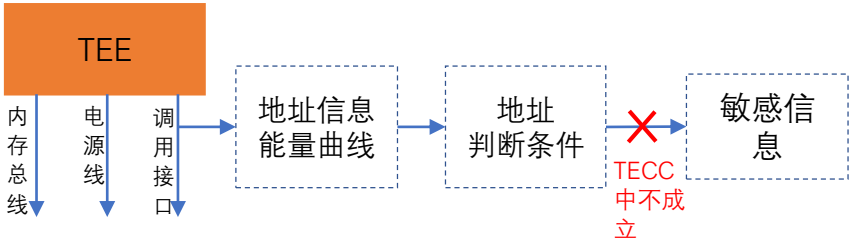
可以避免变量被非预期的代码访问、以及因为内存越界或多线程冲突导致的安全问题。

TRUST IN SOFT

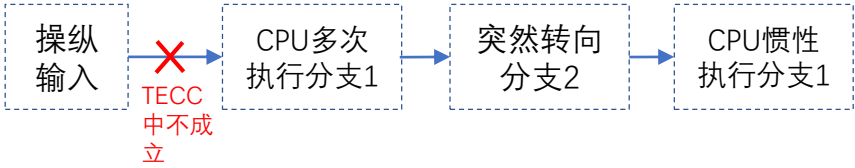
对于生态中的非Rust代码，使用形式化的方法进行验证，能够确保特定的安全问题不会出现。目前TECC的实现已经验证了内存安全问题。

TECC实现安全性（硬件层、供应链）

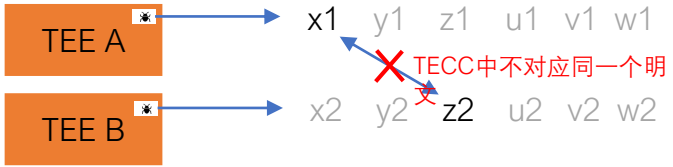
TECC抵消侧信道攻击：分片数据不作为地址和判断条件



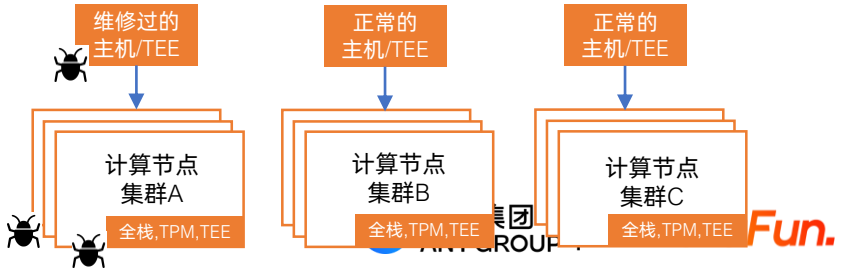
TECC抵消分支预测攻击：输入数据不会作为判断条件



TECC缓解微体系采样结构攻击：多个TEE泄露的分片数据不对应同一个明文



TECC缓解供应链攻击：有隐患的TEE放到同一个分区，全部沦陷无安全问题。

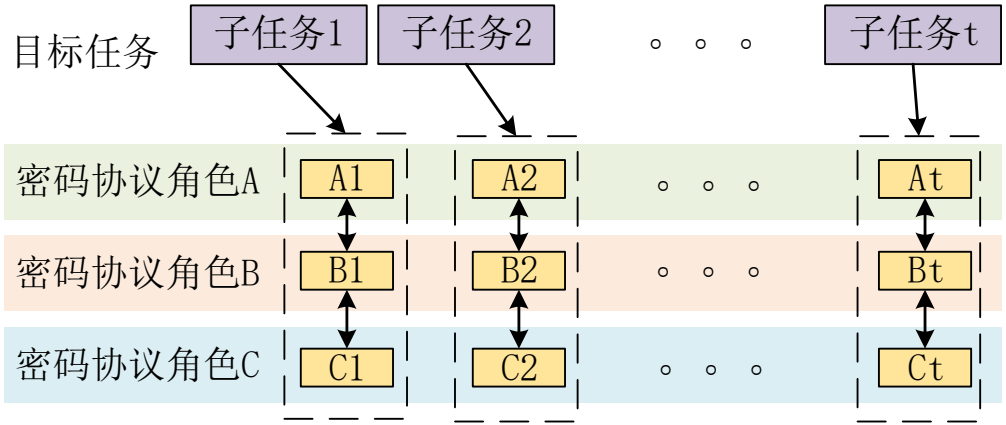


TECC性能可接近明文

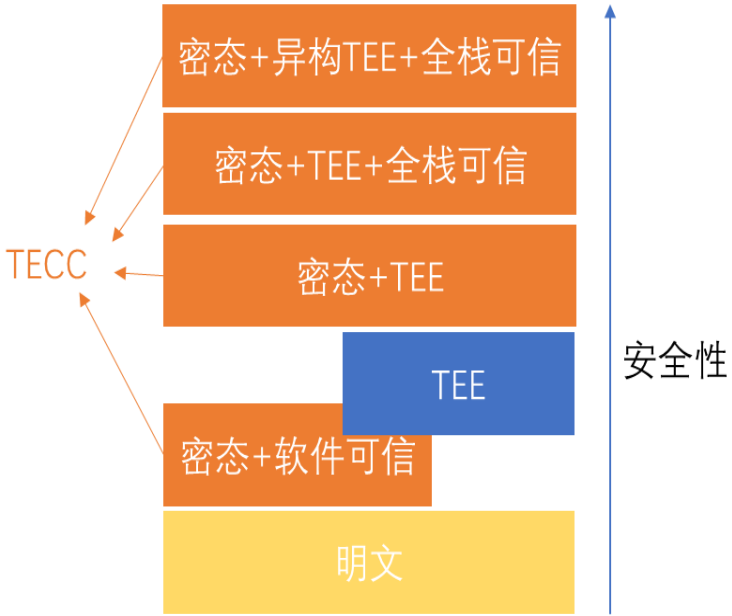
1、TECC对密码协议的需求仅为“密态数据”，可采用轻量级的密码协议。

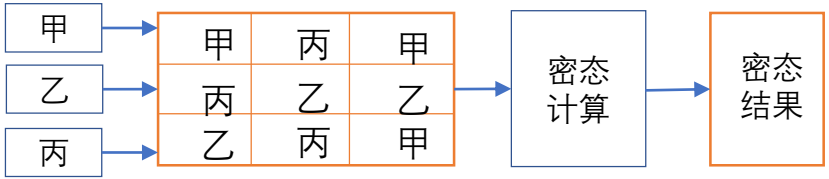
2、内网带宽高达25gbps以上。

3、通过并行化再提升1-2数量级。



优点一：多种安全和成本选择





优点二：

- 任意参与方数量、不同数据分割形式，代码相同
- 相同逻辑的多种情形，代码相同
  - 比如PIR、密态数据库

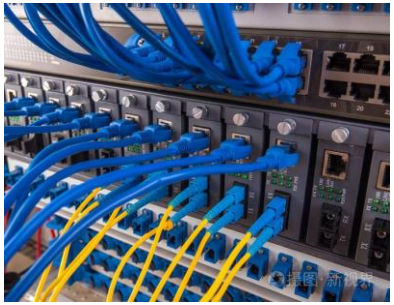
缺点一：需要将逻辑密态化。通过提供通用接口兼容现有生态，可缓解该缺点。

成本和稳定性

成本：

设备：

膨胀1个量级，性能可接近明文



网络：

无额外公网成本

人力：

接入、运维、部署成本低

稳定性：无公网交互，稳定性高。

蚂蚁集团 ANT GROUP | DataFun.

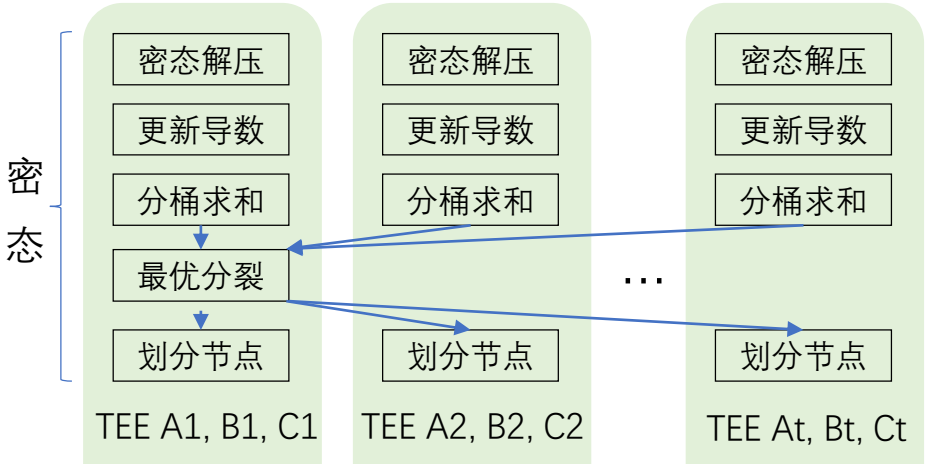
03 TECC实现现状

介绍我们团队现有的实现情况

TECC隐私机器学习

TECC XGBoost隐私机器学习训练：

- 全程密态运算、大运算全部分布式
- 密态压缩传输、传输均衡技术
- 有效位精控、多版本比较



耗时情况：

样本数	环境参数	机器学习参数	耗时
80w	16核128G容器12台	30棵/4层、分桶数13、特征数84和133	2.2分钟
1000w			16.2分钟

机器学习效果：

指标		明文训练 30/4/13	TECC训练 30/4/13	TECC训练 50/4/13
AUC		0.851	0.843	0.851
千一打扰	召回率	0.131	0.109	0.123
	准确率	0.508	0.462	0.493
百一打扰	召回率	0.334	0.316	0.330
	准确率	0.208	0.199	0.206

TECC密态数据分析

用户自定义Python代码

TECC Pandas接口层

表操作	统计运算
表提取、表拼接、选择、排序、Join、分组、分组后聚合	求和、平均、方差、最大最小值、最大最小值位置、分位数
基础运算	**数学运算**
+,-,*,/, =, +=, -=, *=, /=, >, <, >=, <=, ==, !=	Reciprocal, sqrt,log2,exp, sin, cos, tanh, sigmoid

密态数据结构算法

基础算法

TECC 密态数据分析:

- TECC Pandas原生接口、 Rust安全编程语言
- Encrypt in & Encrypt out、非交互式

  -> 保证密态分析可以连续进行
- O(n)的乱序操作

  -> 保证连续密态分析不会累加（或产生）

泄露量

排序耗时情况:

规模	环境	乱序	快排	重建	总时间(分钟)
1000万、200列	32核 128G容器6台	0.15	0.16	0.15	0.47
5000万、200列		0.72	0.83	0.71	2.33

蚂蚁集团
ANT GROUP | DataFun.

04 总结与展望

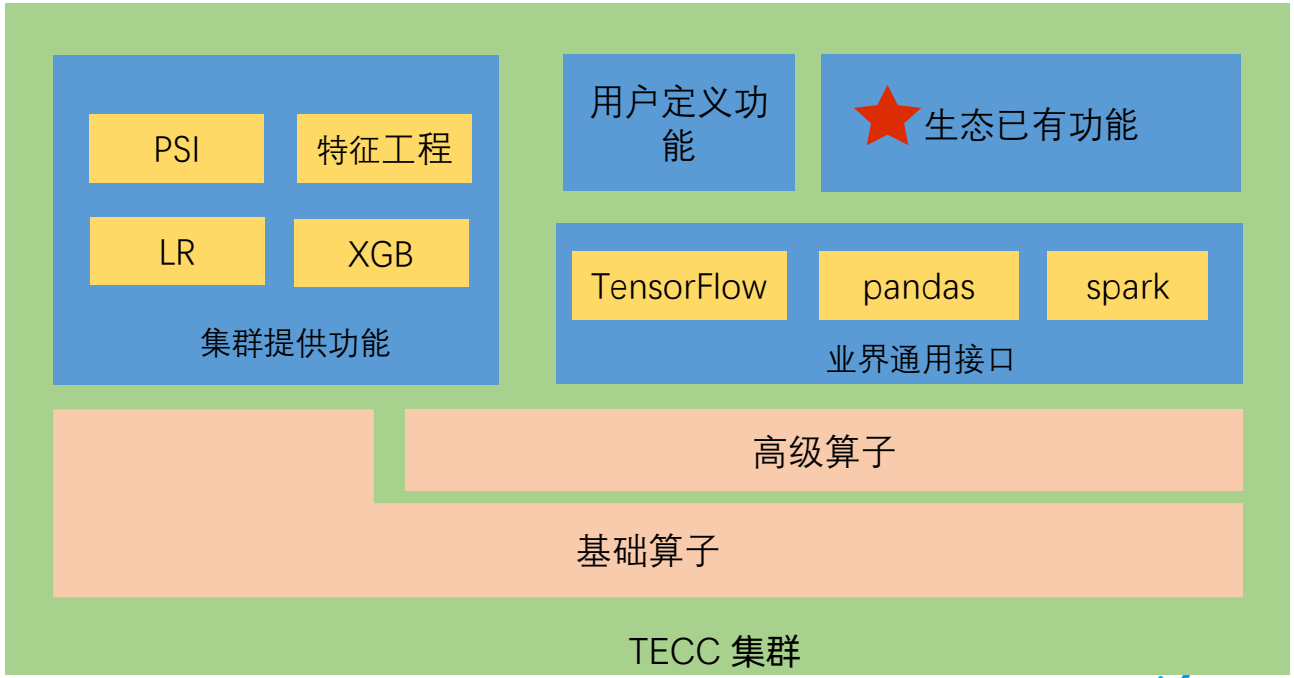
对之前讲过的主要内容进行总结

TECC将与计算生态进行深度融合

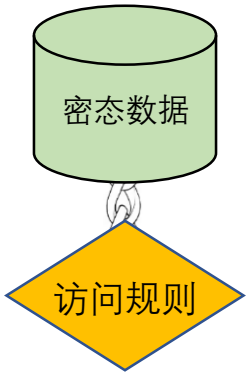
不限参与方数量

提供者1 提供者2 提供者3 … 使用者1 使用者2 … 不限数据形式

全密态功能



允许数据托管



总结

- TECC提供了非常优异的综合能力：
  - 安全：三层防御， Rust，抵消硬件漏洞
  - 性能：轻量级密码、内网、并行化
  - 适用性：多种成本选择、通用算法
  - 成本：一个量级以内
  - 稳定性：内网

1小时处理亿级样本隐私机器学习、10分钟处理亿级行数密态数据分析。

非常感谢您的观看

蚂蚁集团 ANT GROUP | DataFun.