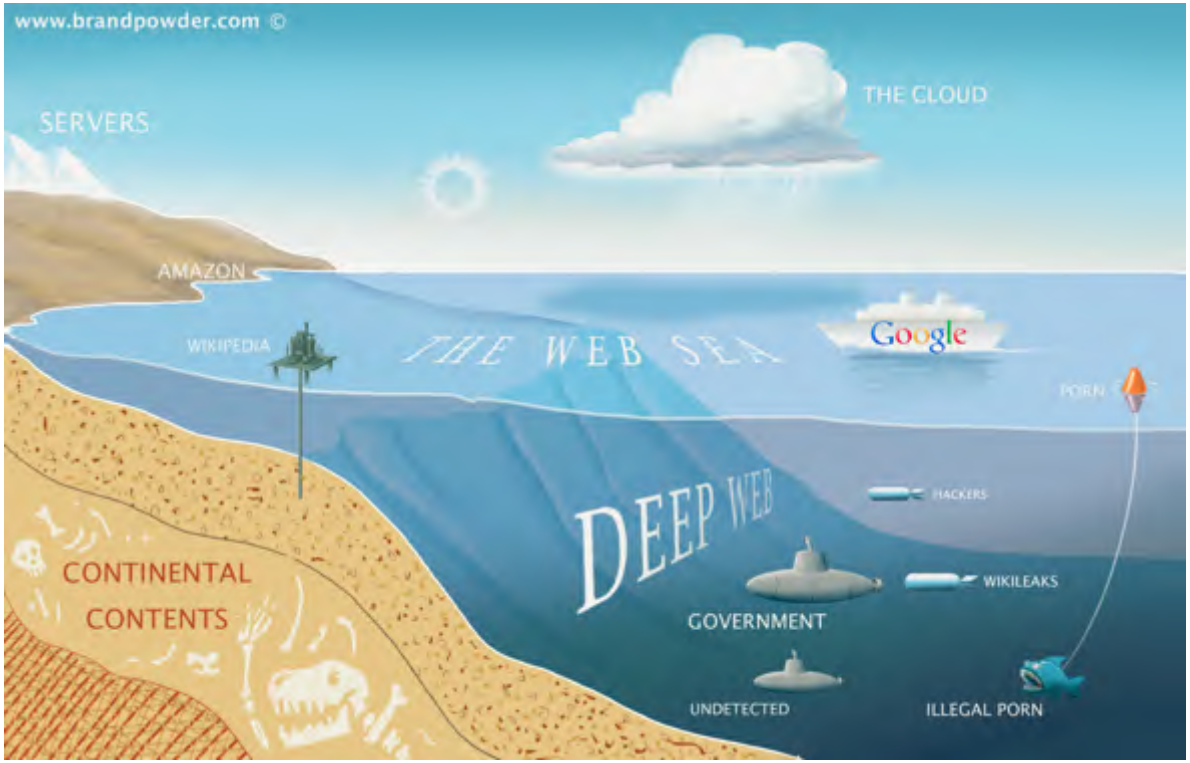
大数据应用
——数据安全和数据分析

周鑫
英特尔中国研究院

工作方向

- 关注实体行业
- 数据分析带动应用
- 数据将走上开放、共享和交易之路
- 关注沉淀的数据和实体来源的数据

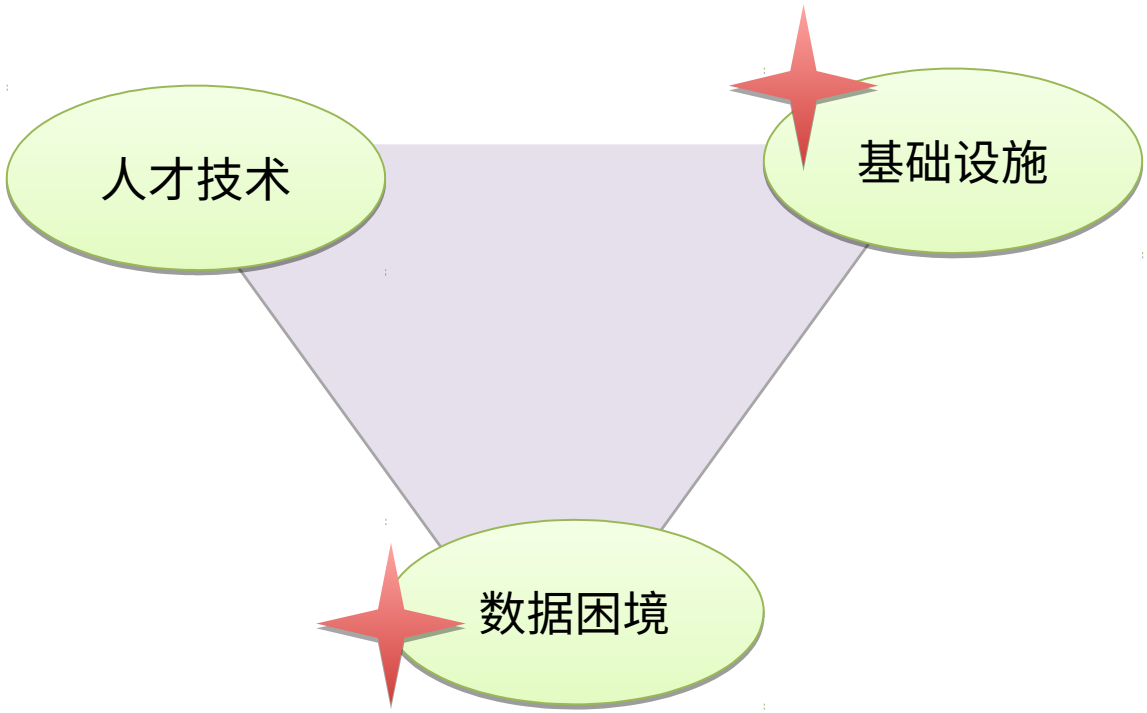


案例分享

乘法效应 + 外部效应

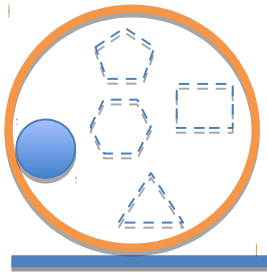
- 健康医疗
  - 肿瘤研究
- 云服务
  - ERP SaaS

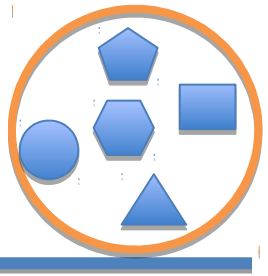
实体行业推广的问题小结



数据的困境

局部





全局

隐私

价值

统计特征

定向瞄准

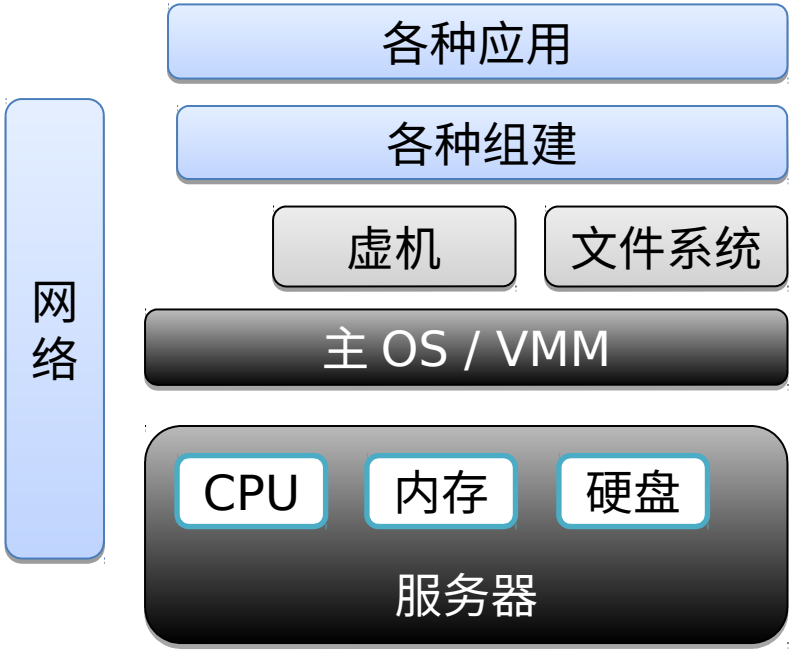
基础设施的困境

- 集中 vs. 分散
  - 拥有权和使用权
- 回报效益的决定性因素、渐进式研究
- 基于去隐私化的技术对数据分析算法和平台的重构
- 数据服务
  - 稳定持续的数据服务提供平台

工作分享 —— Data Coffeehouse

- 可信任云平台建设
- 数据分享和数据分析平台建设

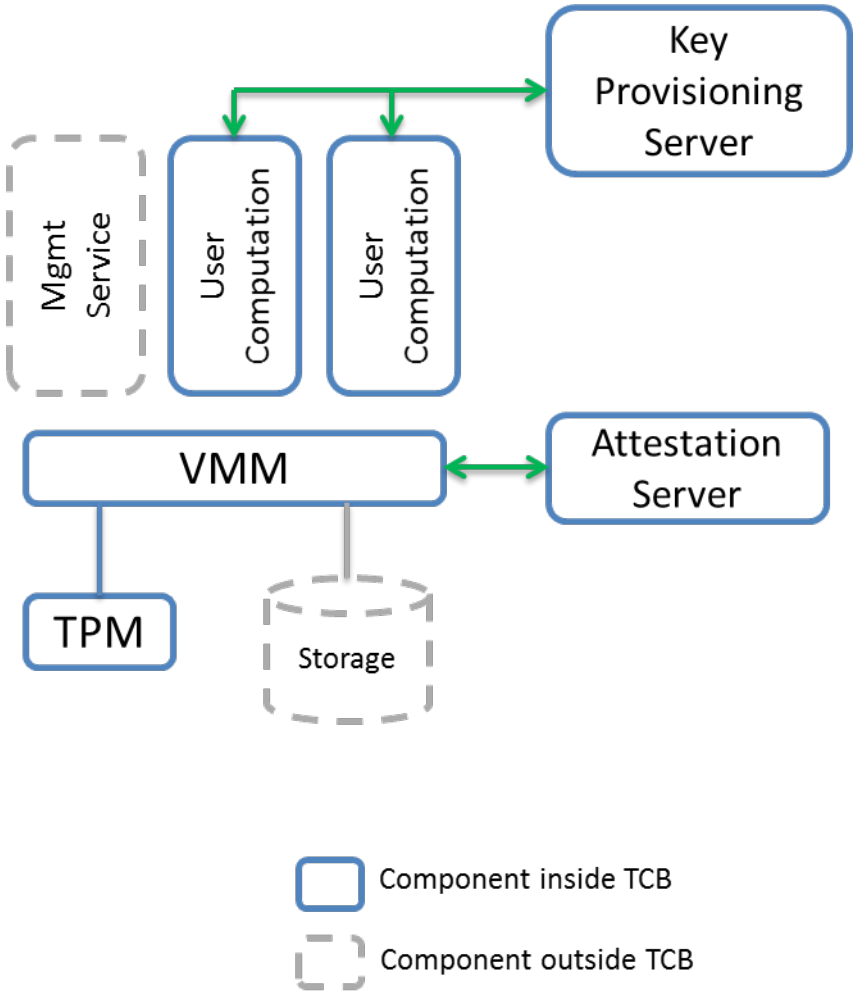
可信任云平台的目标问题



- 安全信任问题是一个深度的全栈的问题
  - 任意部位都可能成为突破点
- 通常的解决方案是基于软件，从 OS 开始
  - 用户认证、授权
  - 文件系统 ACL 、防火墙
- 云服务使得问题复杂，从 OS 开始的解决方案不再足够
  - 硬件隔离使得威胁多样化，立体化
- 必须从硬件开始
  - 硬件认证
  - 主 OS 认证
  - 远程信任链管理

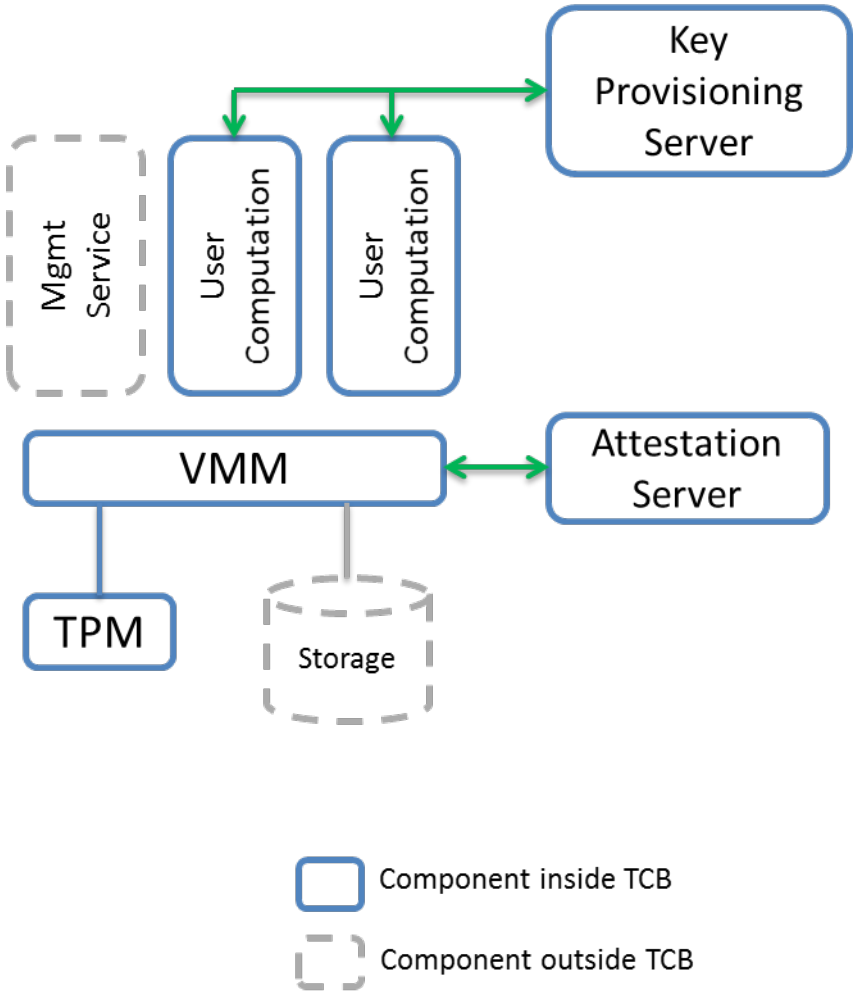
可信任云平台综述

- 可信的运行环境
  - 硬件上基于 Intel TXT/TPM 技术
  - 软件上基于 CloudProxy
  - 防止用户的任务遭到服务提供商的监控或破坏
- 硬件安全
  - 硬件的认证
  - 远程可控的硬件认证环境
  - 主机直连块设备安全
  - 主 OS/VMM 镜像安全
- 硬件配置的信任链
  - 可远程建立信任链，自由根服务器
- 不能防范全部的内存攻击
  - 等待 SGX

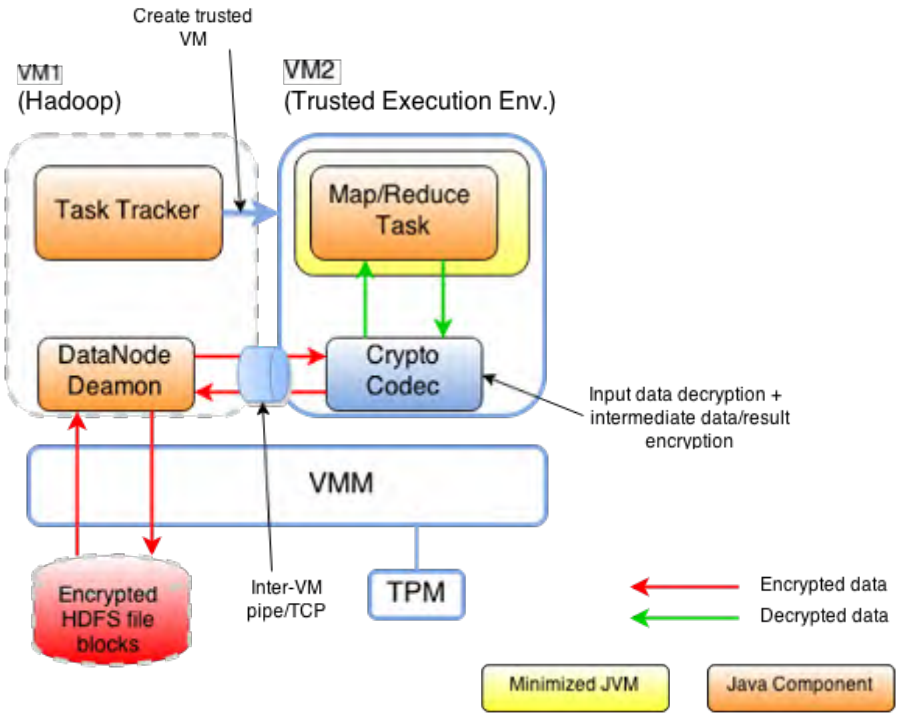


可信任云平台综述

- 计算隔离
  - 不同租户的任务运行在不同的虚机中
  - 防止不同租户间相互干扰或破坏
  - 不同虚机不同等级
    不保护、保护、保护且加密
- 可靠的网络通信信道
  - 同一租户不同虚机间，以及同密钥服务器、认证服务器间均使用加密信道
  - 防止通信数据泄漏或被篡改
- 数据加密
  - 所有离开 TCB 边界的数据均加密
    （输入、输出以及中间结果）
  - 保证的数据的机密性和一致性

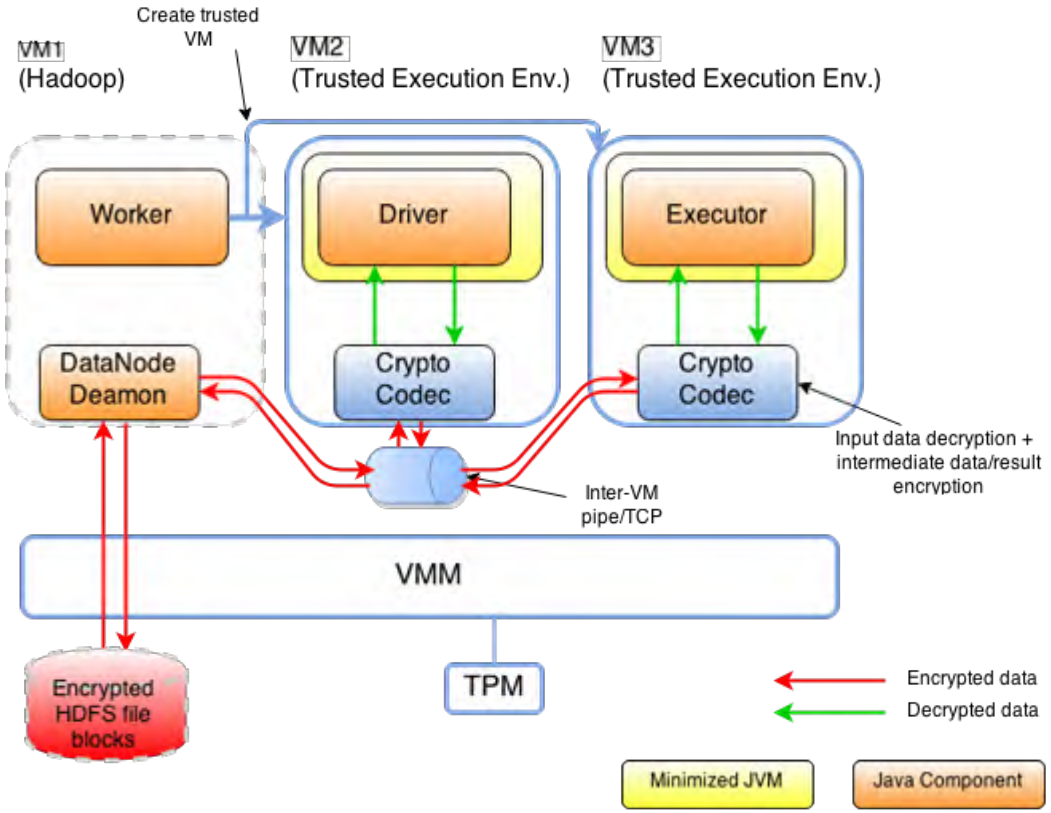


案例 1： Hadoop 改造



- 使用单独的可信虚机（VM2）将用户的计算与其他组件隔离
- 使用剪裁过的 JAR 包以减小 TCB
- 使用修改过的 Task Tracker 创建虚机 VM2
- 在 VM2 中添加一个加解密引擎负责对进出 VM2 的数据进行加解密。
- VM2 启动时，需要传入一个加密后的密钥用以初始化加解密引擎
- 可信虚机（VM2）通过普通管道或网络访问原有虚机（VM1）中的 HDFS API，进而读写 HDFS 文件

案例二： Spark 改造



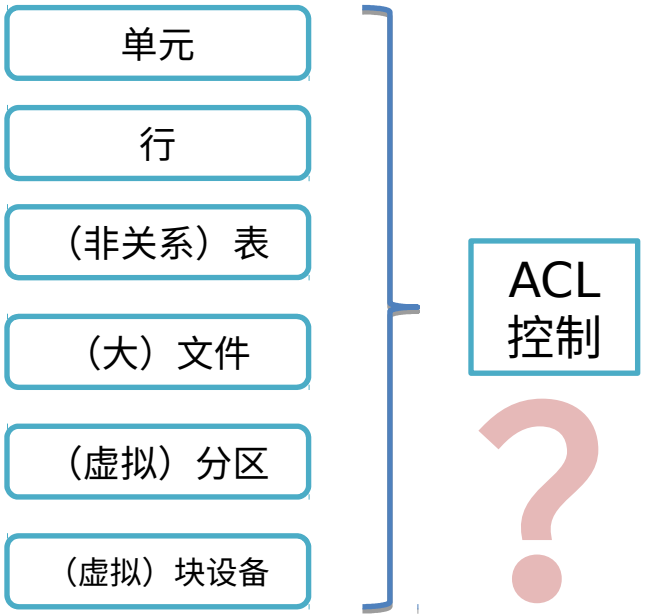
- 使用单独的可信虚机（VM2/3）将用户的计算与其他组件隔离
- 使用剪裁过的 JAR 包以减小 TCB
- 使用修改过的 Worker 创建可信虚机
- 在可信虚机中添加一个加解密引擎负责对进出 VM 的数据进行加解密。
- 可信虚机启动时，需要传入一个加密后的密钥用以初始化加解密引擎
- 可信虚机（VM2/3）通过普通管道或网络访问原有虚机（VM1）中的 HDFS API，进而读写 HDFS 文件

经验总结

- 性能！
  - 启动慢

- 解决
  - Docker
  - 但是问题很多
    - 隔离差
    - 镜像管理
    - Spark 结合不好

数据分享和数据分析平台的目标



- 传统的基于权限控制的方案
  - 数据安全不再是零和一的选择
  - 授权必须授权以后还得控制
- 加密可能是终极方案
  - 顶级大牛的说法
  - 但是，屌丝的思考可能跟不上大牛
- 学术界认识到基于语意，数据可以分成不同的语意敏感级别
  - ID
  - QI
  - SA
  - Public

  把基于语意分级的安全机制引入到大数据分析平台

数据安全的相关工作

- 静态数据安全
  - 访问控制： Apache Accumulo ， Hbase
  - 加密： HADOOP-10150

- 数据脱敏 / 匿名化
  - 去标识符，但基于准标识符（ quasi identifiers ）仍能重新标识化
  - k-anonymity 、 L-diversity 、 T-Closeness
  - 差分隐私（ differential privacy ）
  - 隐私安全性和数据可用性的平衡

- 动态数据安全
  - 动态审计能力：数据泄露防护（ Data Leakage Prevention ）

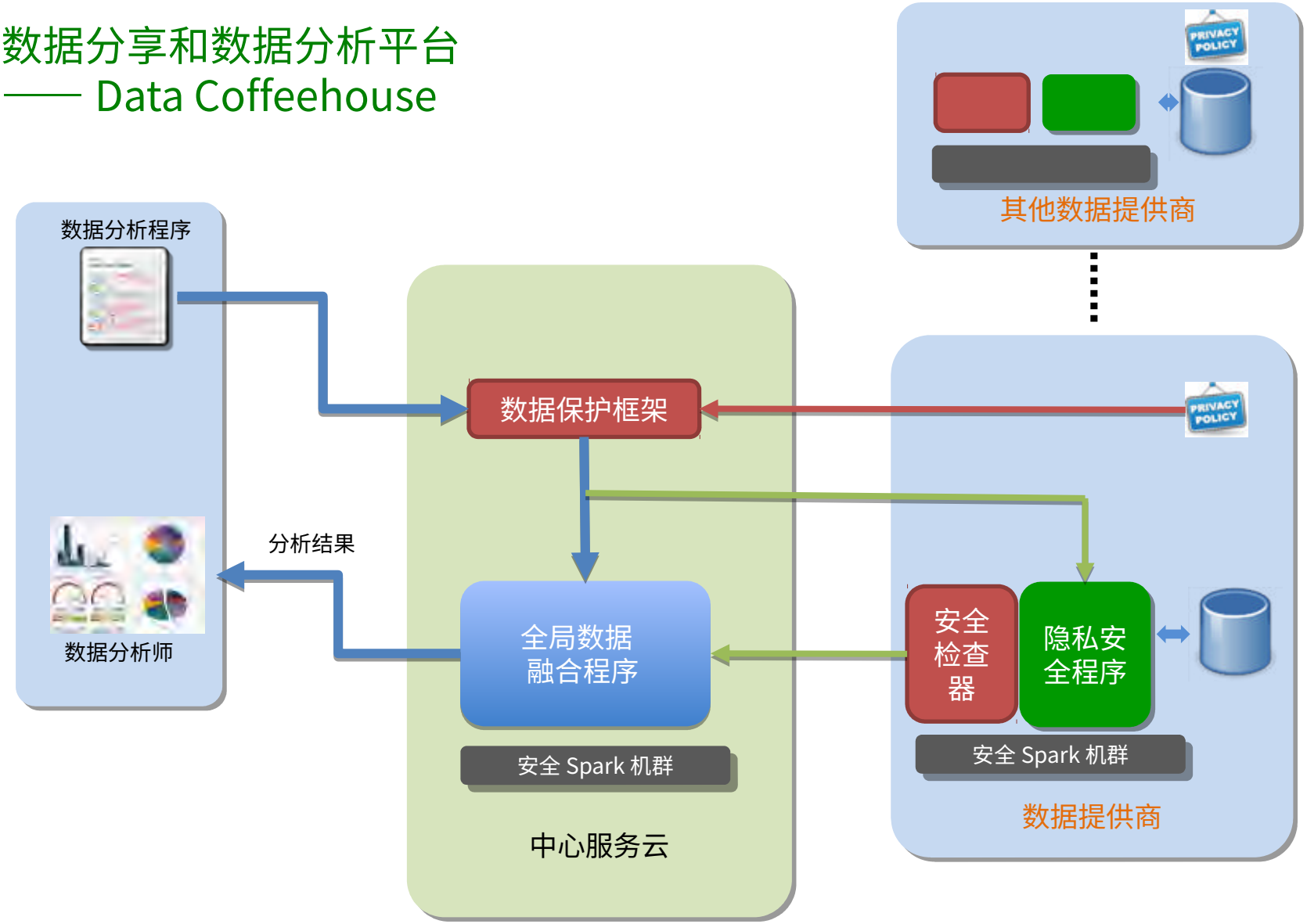
数据分享和数据分析平台综述

- 数据提供者
  - 数据收集
  - 法务部门
- 数据使用者
  - 数据分析师
- 运营方
  - 运维

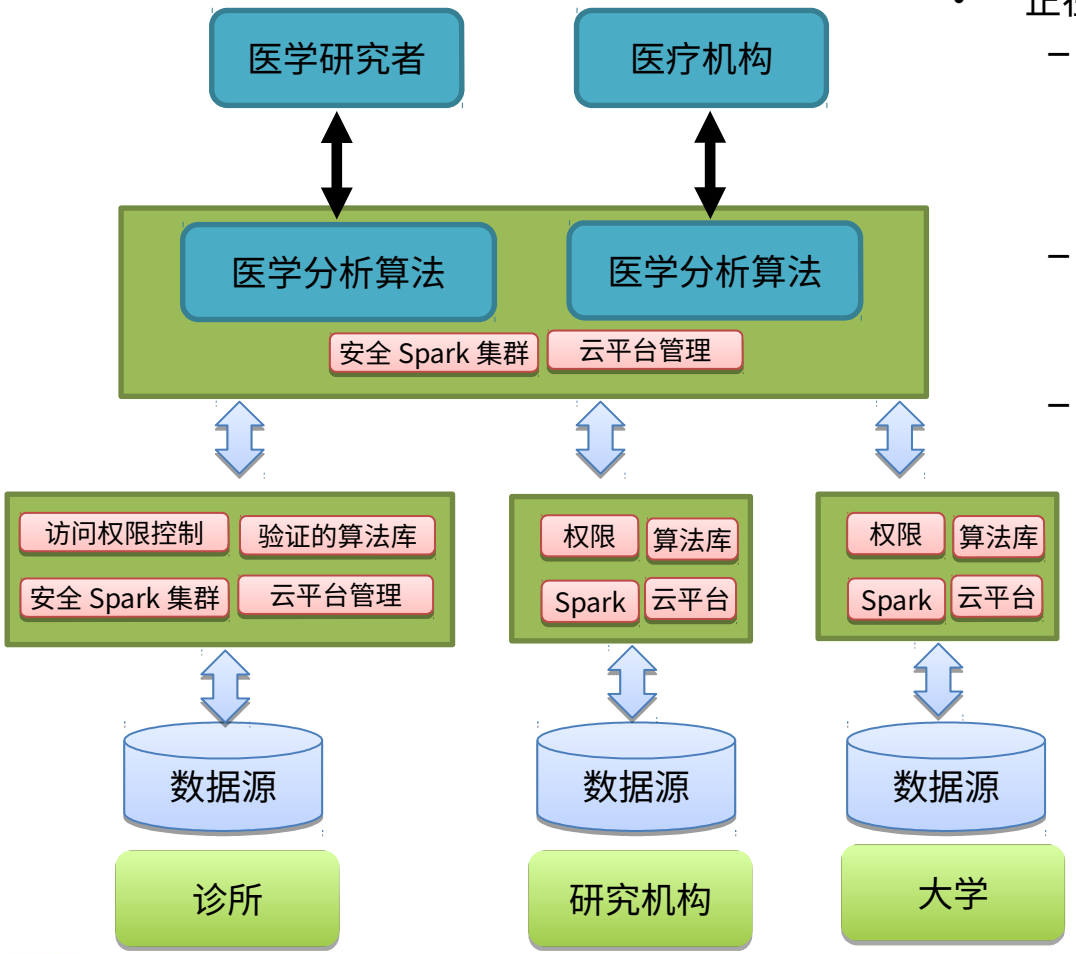
- 数据要求
  - 数据机密
    - 客户机密、业务机密
  - 数据大、流式数据
  - 保护规范
- 分析要求
  - 保护商业机密
  - 持续、反复运行分析程序
  - 降低学习曲线
- 运行要求
  - 成熟稳定平台

解决思路：
- 集中和分散结合 —— 中心云 + 私有云 / 本地云
- 简单明了的保护方案 —— 给予规则的数据访问保护

数据分享和数据分析平台
—— Data Coffeehouse



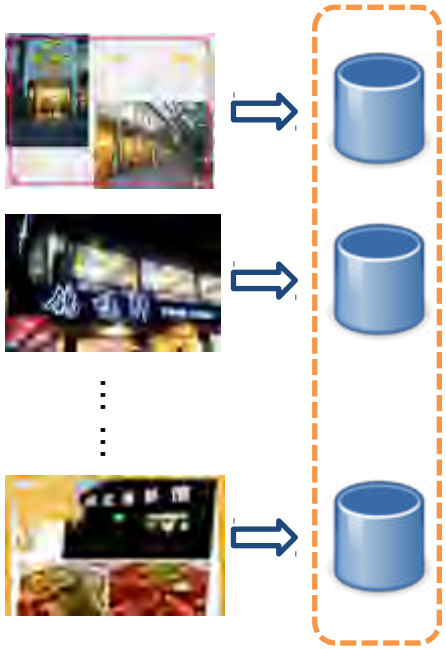
数据分享和数据分析平台实践之一



- 正在参与的共享医疗数据原型系统
  - 背景
    - 医疗数据具有长尾特性
    - 因为隐私保护等原因，诊所等医疗机构无法共享其数据
  - 问题
    - 如何将所有机构的医疗数据聚合起来，从而提高医学研究的成果
  - 解决方案
    - 创新性地使用基于 TPM 的软硬件协同方法建立一个云平台，实现安全的数据共享和数据处理

数据分享和数据分析平台实践之二

餐饮 ERP 系统的数据共享和数据分析平台

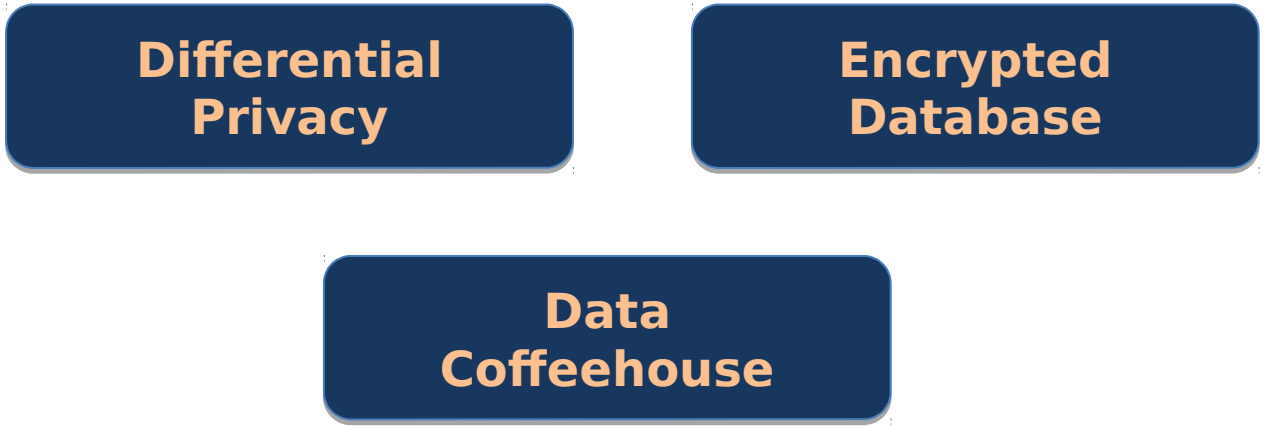


众多阻碍

- 小数据 ᚙ 大数据
  - 局部片面 ᚙ 全面综合
- 精准营销 ᚙ 数据安全
  - 顾客数据
  - 营业数据

数据分析服务商

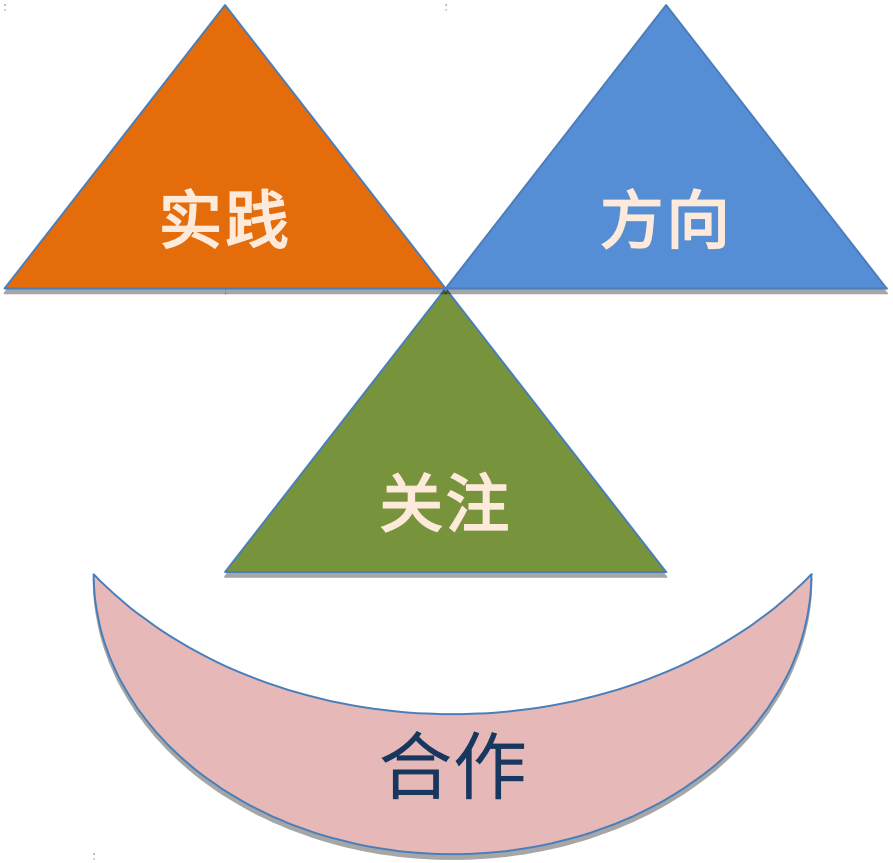
回顾和其他方案的比较



开发用户接受度：
- 数据分析程序难易
- 数据理解难易
- 数据更新频率

运维接受度：
- 角色清晰度
- 保护管理可行度
- 新业务加入难度

总结



xin.zhou@intel.com 微信：XinZhou_Bookworm

Thanks!