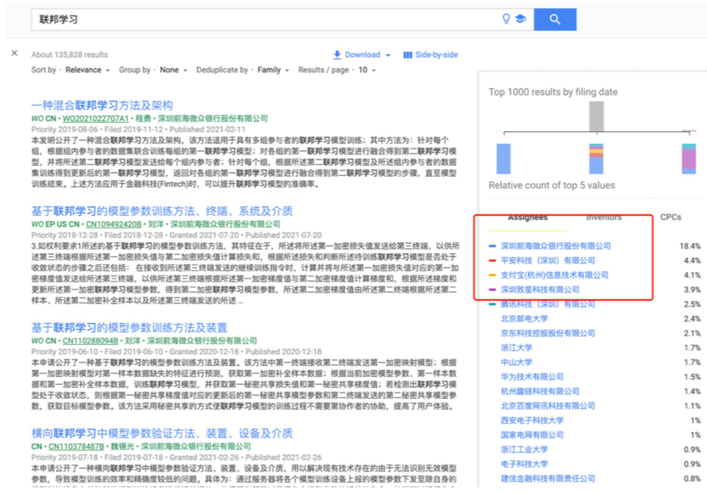
联邦学习应用调研

- 背景
  - 1 联邦学习这几年比较热络

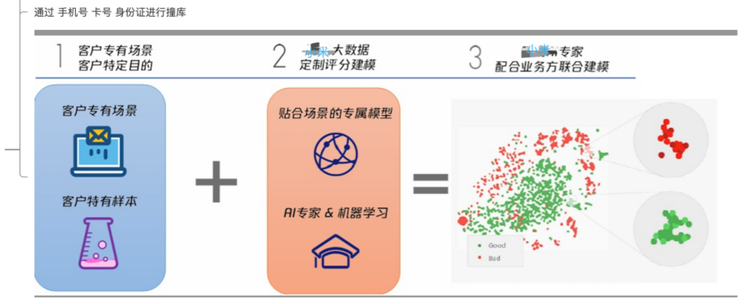


  - 2 我们团队对联邦学习了解不多 尤其在应用端

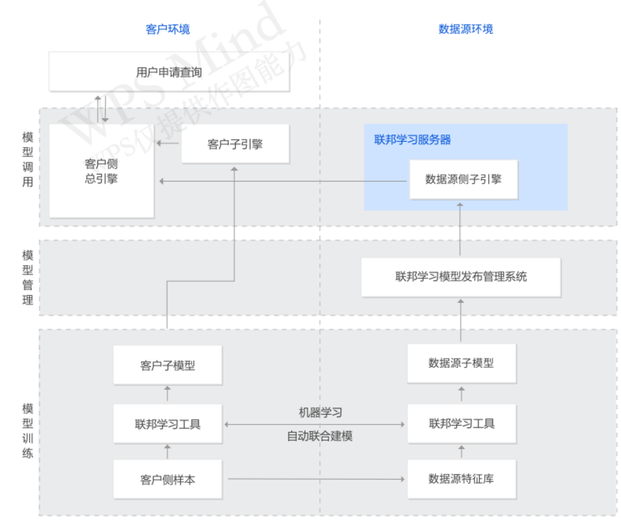
- 小米如何用联邦学习做落地的
  - 小米金融的整体业务介绍
    - 整体业务情况
      - 智能硬件+IOT
        - 各种摄像头 小爱音响等
        - 电视 路由器
      - 手机
      - 游戏
      - 小米金融
      - 电商有品网
      - 生态链 小米手环等
      - 小米云
    - 核心业务情况
      - 互联网广告
      - 手机
      - 小米金融
        - 供应链金融
        - 消费信贷
          - 联合贷场景
          - 杠杠率不得超过注册资金3倍
        - 互联网保险
        - 互联网理财
        - 第三方支付（小米钱包） —— 通过钱包入口 办理联名卡 —— 主要业绩 1.5年时间里 累计开卡200万张，单卡100元一张服务费，累计收入2个亿
      - 小米电商
    - 子主题

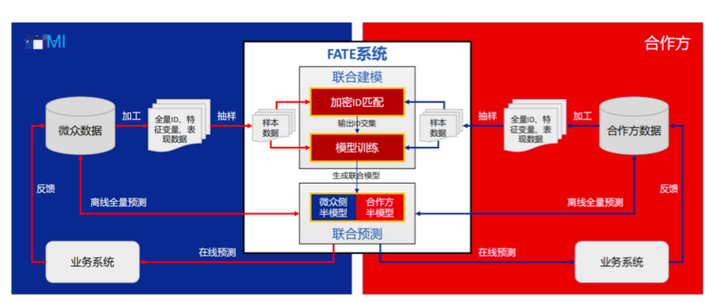
主要数据赋能的点 在这里

  - 核心业务的核心场景 —— 信用分 —— 本质上 还是卖用户流量的生意 —— 信用分构建
    - 构建信用分
      - 该系统的不足
        - 1 面向长期合作的 VIP客户 需要找到新的收入方式
        - 2 对通用模型效果不满意 要求更好的模型区分度
        - 3 客户独有的X输入定制化模型Y的输出
    - 联合建模 —— 联合建模环境的优化
      - 环境完全进行物理隔离，客户把数据带到我方的沙箱环境
      - 通过 手机号 卡号 身份证进行撞库



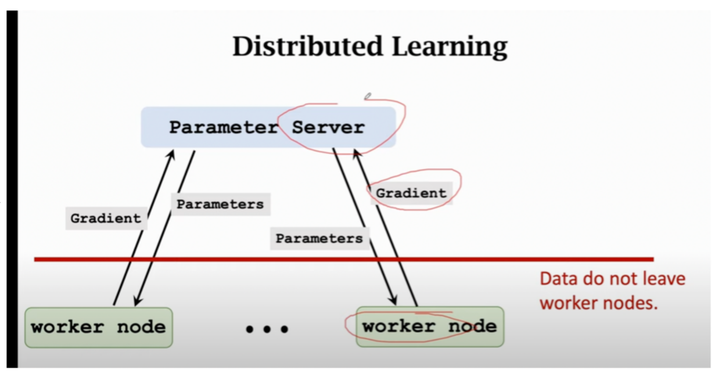
      - 该系统的不足
        - 1 由于需要双方撞库 撞库的成本正常的数据维度匹配率在5%~10%，模型要取5万样本 最终拿到小米的用户量在50万左右
        - 2 入场建模 需要人力出差 比如招商银行需要从深圳调人来出差 1~2周 反复沟通和确认 数据和模型
        - 3 客户不满足撞库的匹配率和模型最终的效果
    - 联合建模-联邦学习





结构示意图

- 联邦学习的几个待解决的问题
  - 1 为什么说联邦学习是分布式机器学习
    - 1. McMahan and others: Communication-Efficient Learning of Deep Networks from Decentralized Data. In *AISTATS*, 2017.
    - 2. Konevcny, McMahan, and Ramage: Federated optimization: distributed optimization beyond the datacenter. *arXiv:1511.03575*, 2015
    - 联邦学习最早是15年提出来的



    - 联邦学习本质上还是分布式学习
  - 2 联邦学习本质上还是分布式机器学习
    - 1 server对客户端是否有绝对的控制权（users control device and data）：用户对自己的设备和数据有绝对的控制权 随时可以停掉配合
    - 2 客户端不稳定(unstable)： 客户端可以是iphone ipad 等移动端设备， 传统节点 分布节点都是机房 非常稳定
    - 3 通信代价大(Communication)：不在一个国家 不在一个网域，局域网 连接稳定（传统的分布式节点都是局域网高速宽带）
    - 4 数据不一定是独立同分布(not IID)： 传统节点是独立同分布
    - 5 节点负载不平衡(imbalanced)： 每个设备的数据量也不同（传统的机器学习是支持 负载均衡的）
  - 3 联邦学习的通信问题（效率） —— 多本地计算 少通信
  - 4 联邦学习的安全问题（安全隐私）
    - 1 如果机器学习的模型是有用的 模型是一定带有数据信息的
    - 2 训练数据是可以通过模型反推断出来的
    - 3 联邦学习中 是可以通过梯度和模型参数 大致推断出数据（比如如果数据是图像数据 可以大致看出来这个是是男还是女 肤色等）
      - 1. Melis et al. Exploiting unintended feature leakage in collaborative learning. In IEEE Symposium on Security & Privacy, 2019.
      - 2. Fredrikson et al. Model inversion attacks that exploit confidence information and basic countermeasures. In CCS, 2015.
      - 3. Hitaj et al. Deep models under the GAN: information leakage from collaborative deep learning. In ACM SIGSAC Conference on Computer and Communications Security, 2017.
  - 5 联邦学习的鲁棒性问题 —— 由于server对客户端节点的控制性低，拜占庭将军问题 —— 某个节点 ，数据经过精心串改 恶因修改参数或者模型