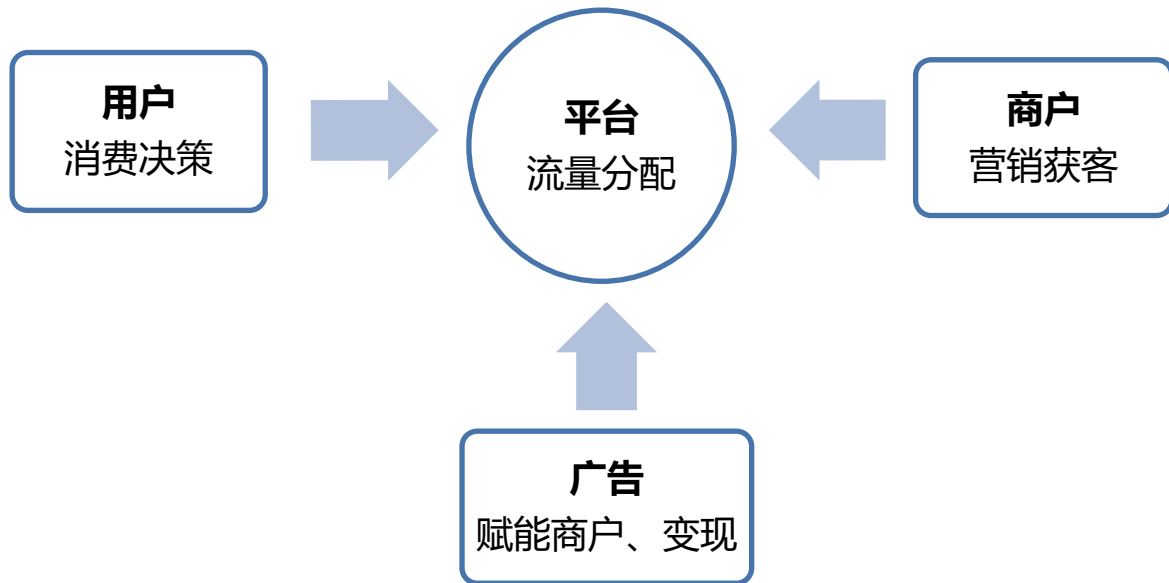


推荐广告机器学习实践

程佳

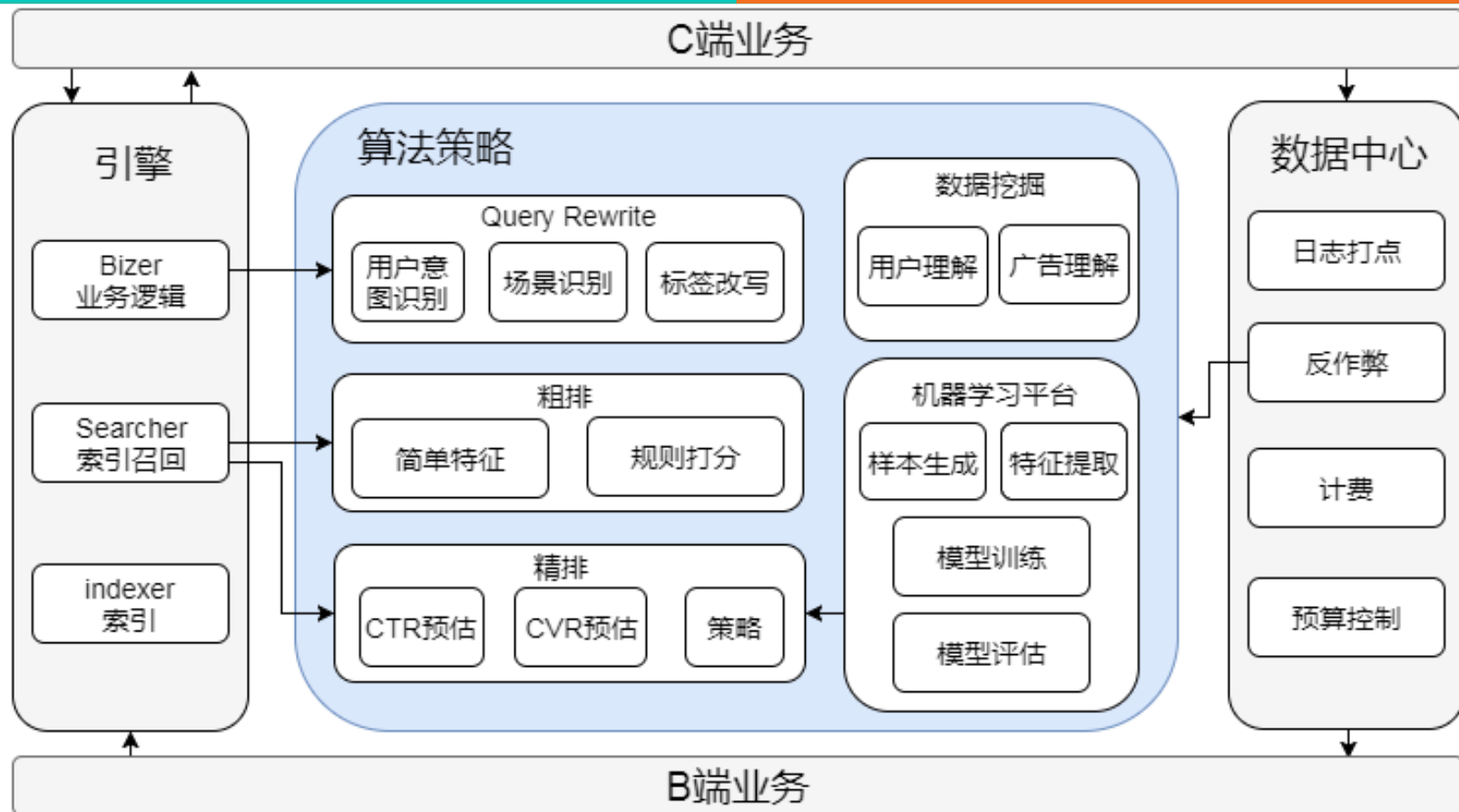
- 业务背景
- 机器学习平台
 - CTR预估平台
 - 模型平台
- 排序机制
- 总结思考



主要位置



业务背景/推荐广告业务架构



- **双平台合并**

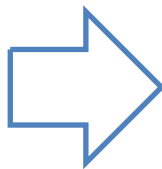
- 如何提高效率，优化人力成本？

- **业务初期**

- 发展快，团队初创，基础技术储备弱

- **O2O广告主**

- 互联网新人，如何优化广告主体验？



- **搭建机器学习平台**

- CTR预估统一框架、平台化，解耦业务
- 构建模型平台，支持大规模模型的优化迭代

- **优化排序机制**

- 广告投放更简单、高效

- 业务背景
- 机器学习平台
 - CTR预估平台
 - 模型平台
- 排序机制
- 总结思考



- 双平台多套系统
- 按广告类型分工优化

- 优化效率低：离线在线不一致；重复开发
- 人力成本高：面面俱到，无法做深

多套流程

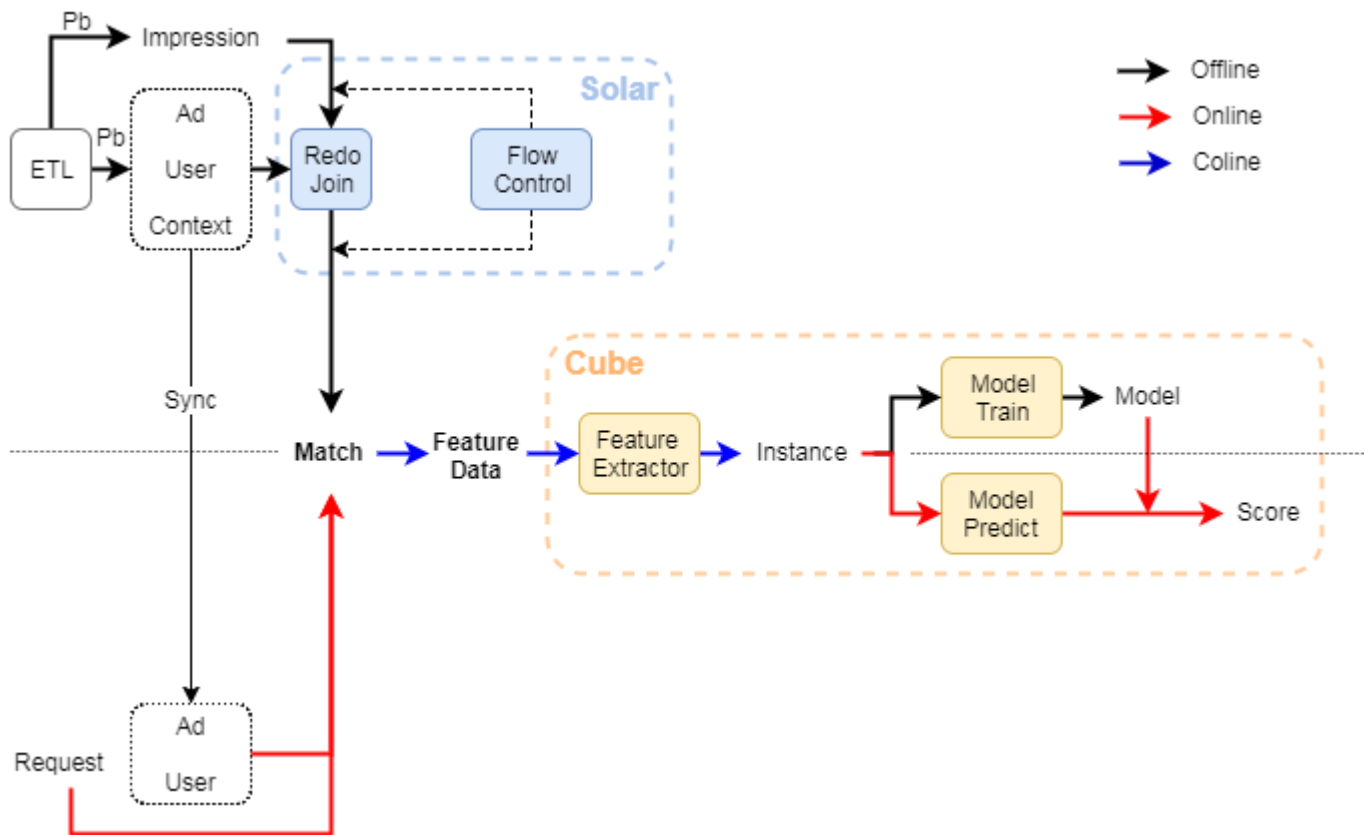
统一框架

框架平台化

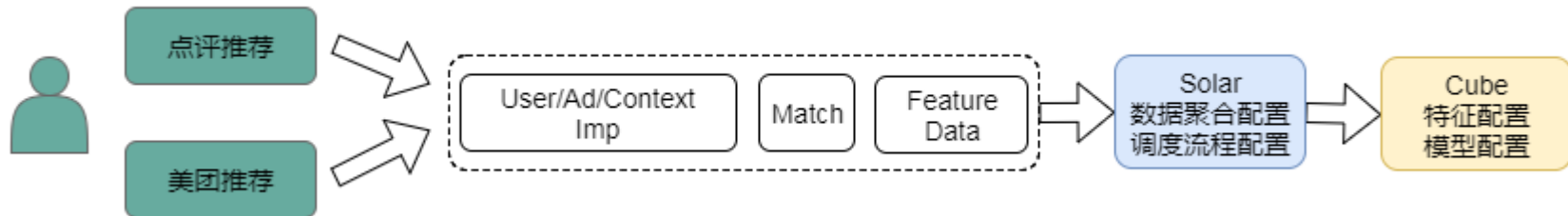
- 挑战

- 业务框架如何抽象？
- 线上线下一致性如何保证？
- 如何高效的数据聚合及回溯调度？

CTR预估平台/统一框架



- 业务：数据
- 框架：流程
- 线上线下数据一致，代码一致
- Solar：数据聚合及回溯控制



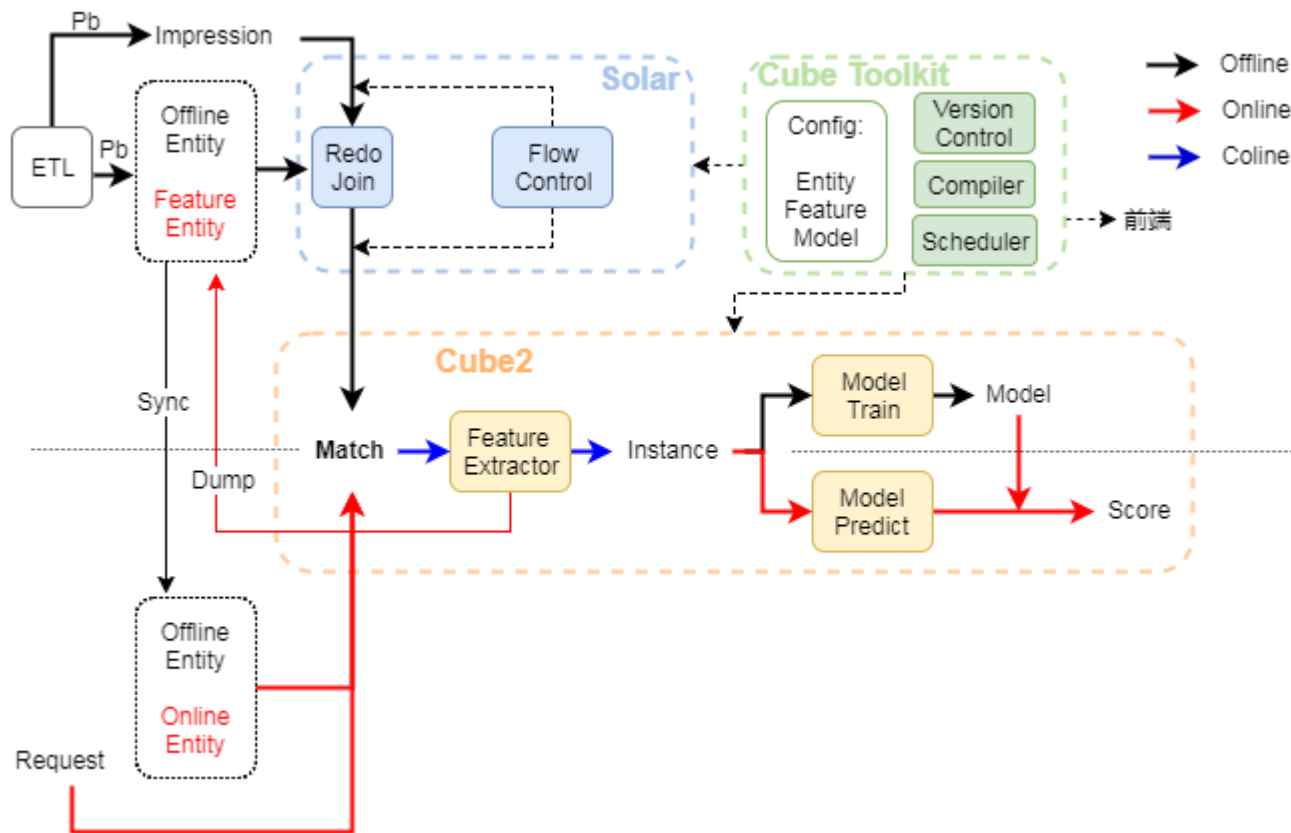
- 总结

- 一套代码，线上线下一致
- 人力成本减少一半

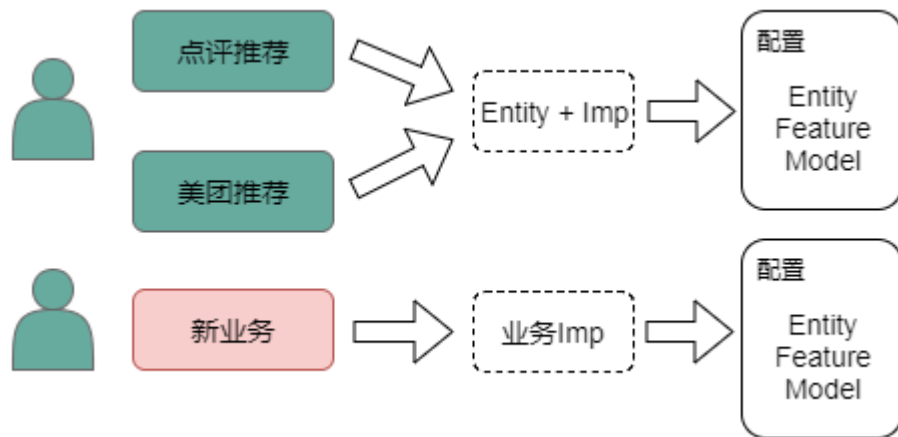
- 够好用吗？No！

- 数据抽象不够：需要实现少量代码(Match+FeatureData)
- 配置复杂冗余、流程多：新人熟悉成本高
- 实时特征需要额外开发

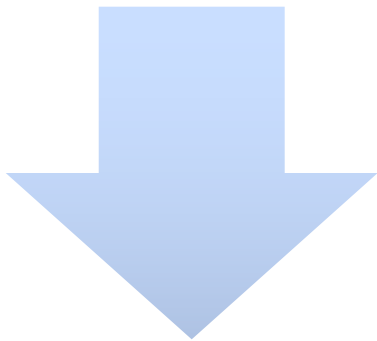
CTR预估平台/平台化



- 数据抽象：Entity + Imp
- Toolkit：配置简化、版本自动依赖
- 回溯与实时统一



- 生成Entity、Imp数据
- 简单配置



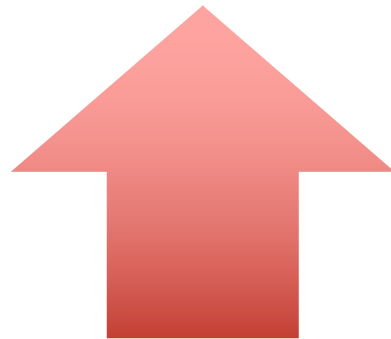
• 简单易用

- 模块化、简单化
- 全配置
- 拖拽可视化



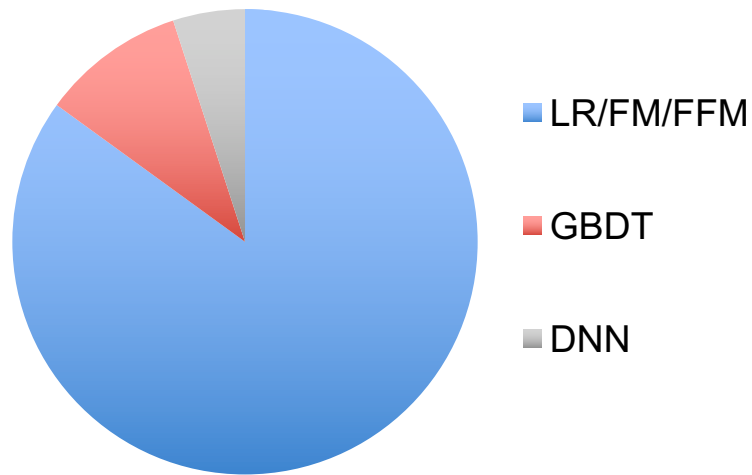
• 能力强健

- 支持海量数据和特征处理
- 离线在线统一
- 回溯和实时统一

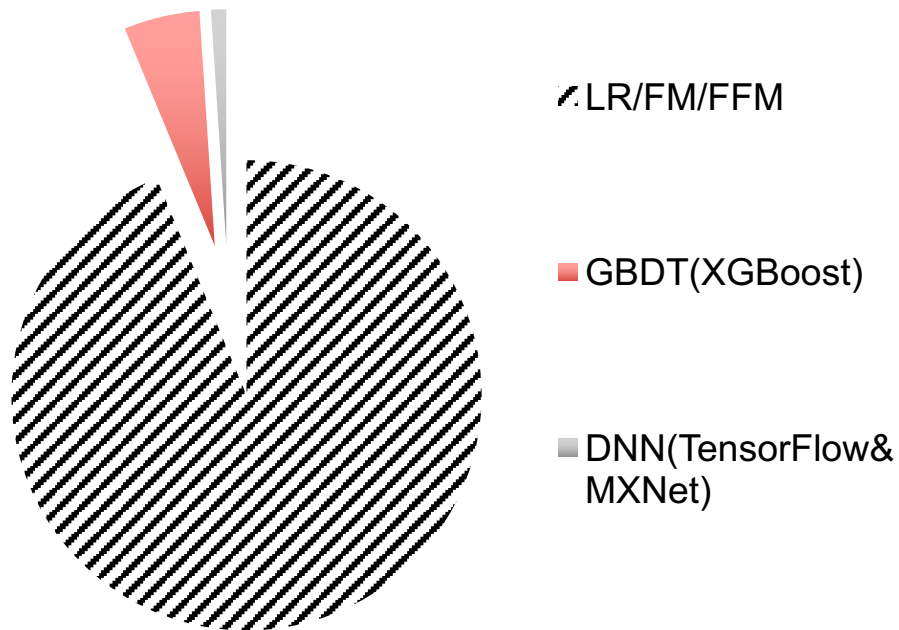


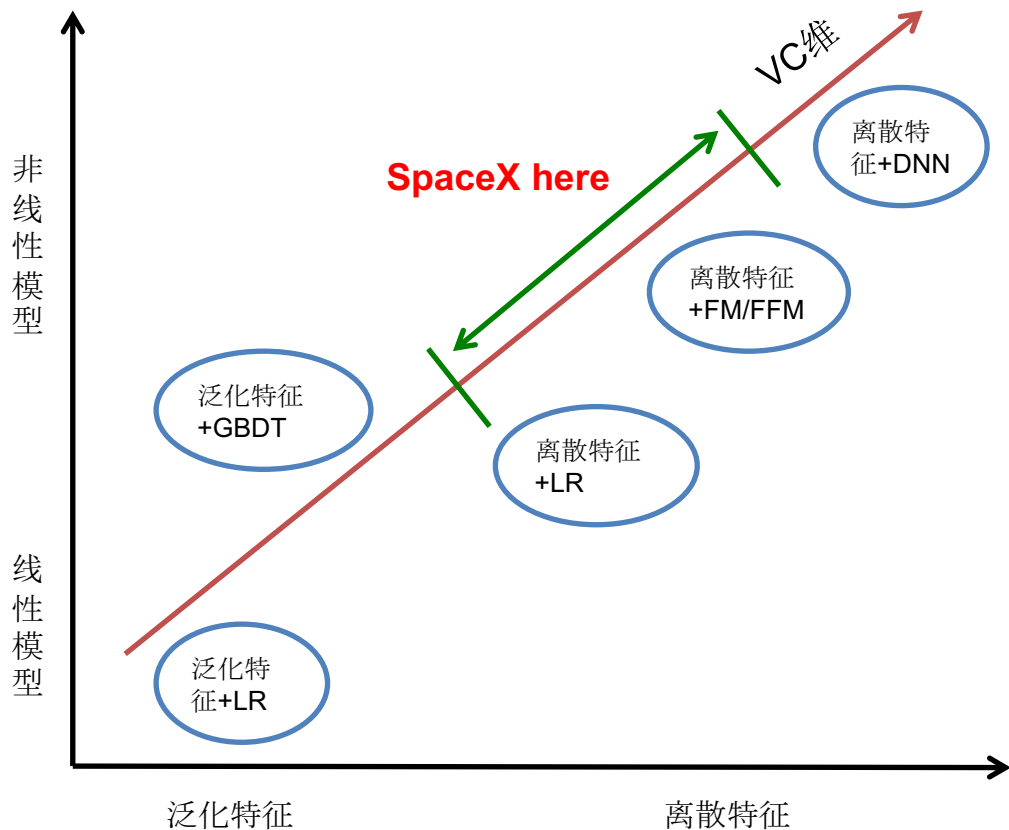
- 业务背景
- 机器学习平台
 - CTR预估平台
 - 模型平台
- 排序机制
- 总结思考

业界CTR模型



美团点评CTR模型



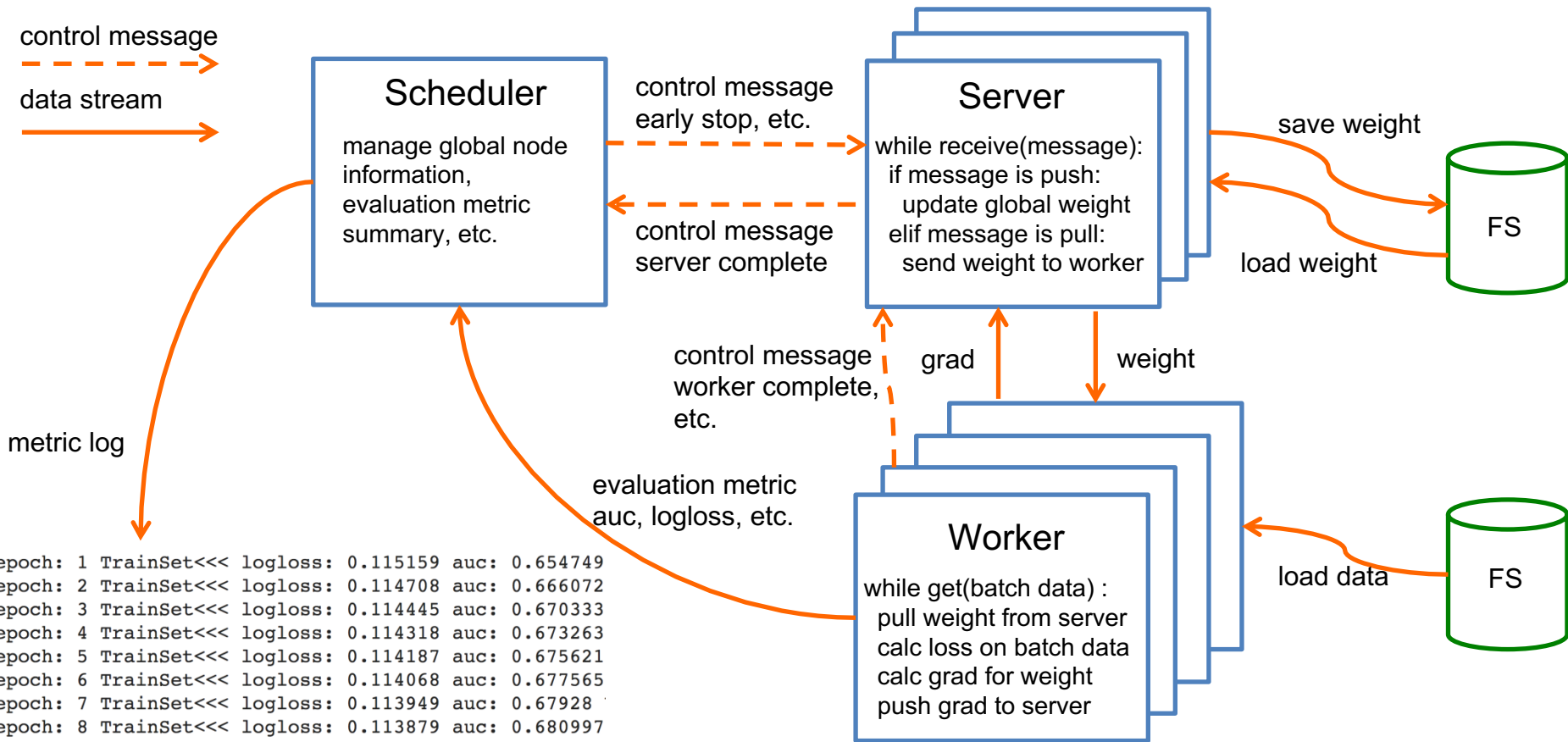


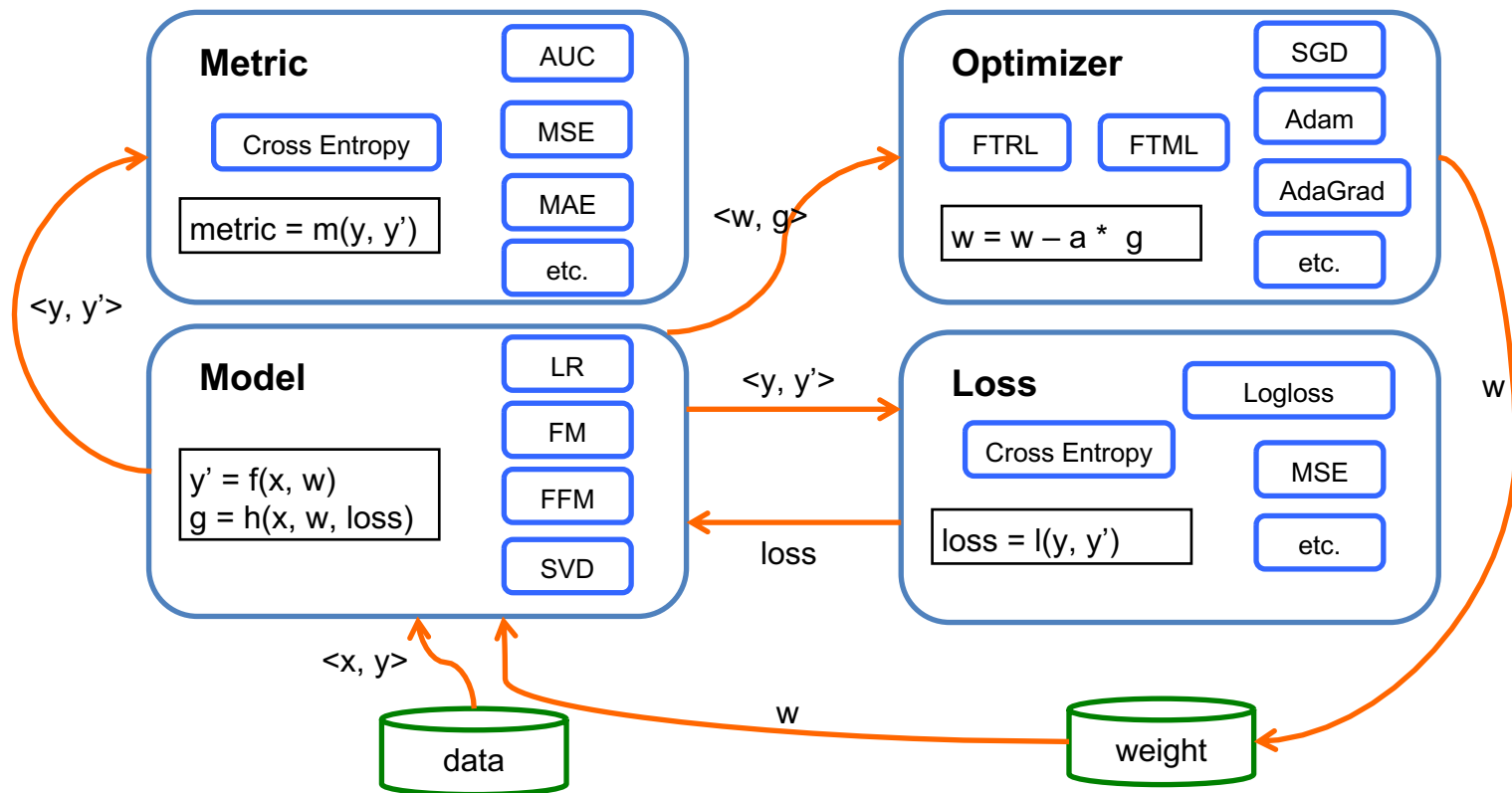
- 更高VC，更高的拟合能力
- DNN > GBDT > 线性
- 离散特征 > 泛化特征
- SpaceX：支持大规模离散特征

模型平台/分布式架构

control message
----->

data stream
----->





- FTML算法：参考2017 ICML 《Follow the Moving Leader in Deep Learning》



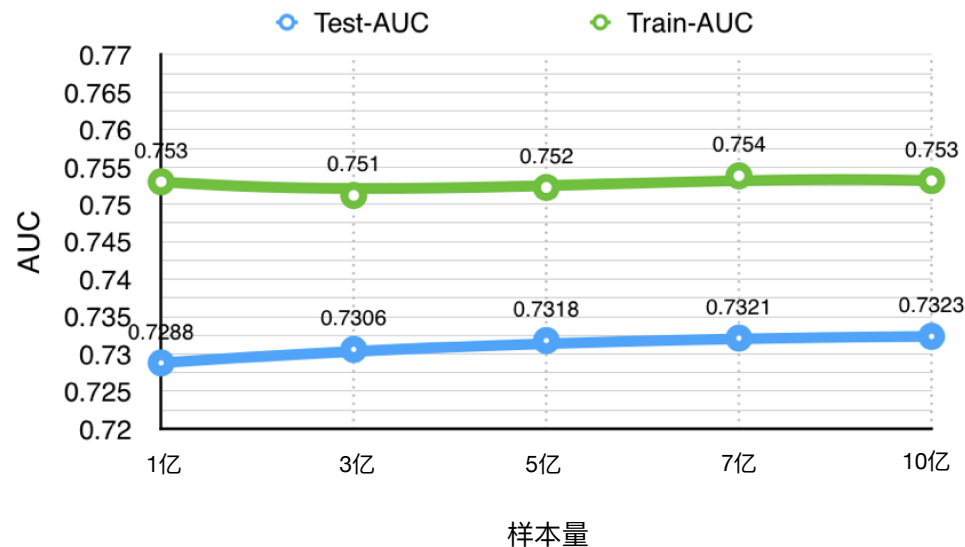
模型能力越来越强



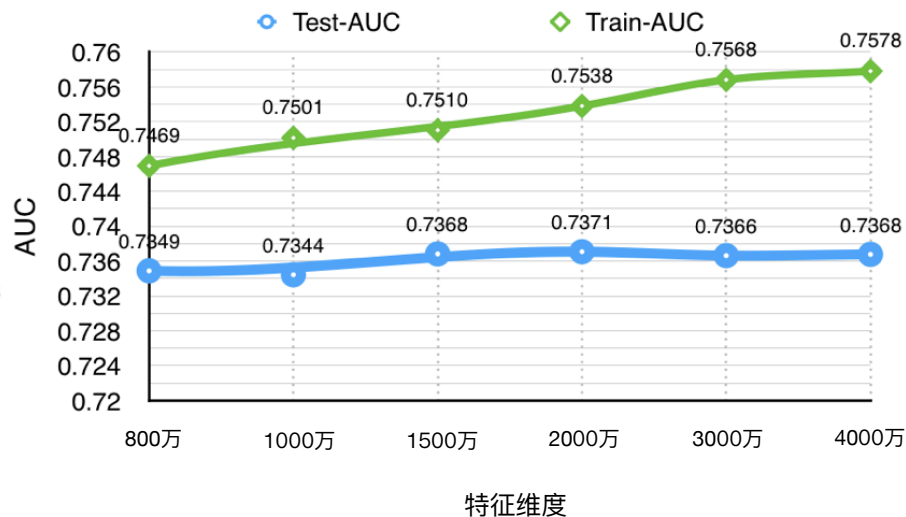
样本量级越来越大



特征维度越来越高



Test-AUC: 随着训练样本增加, 效果整体提升并趋于收敛



Test-AUC: 随着特征规模增加, 效果先变好后变差。

原因分析:

前期: 特征变多, 模型表达能力变强

后期: 稀疏特征变多, 过拟合严重, 泛化能力变弱

- **高效支持海量特征的模型训练**

- 现有集群上支持百亿样本、十亿级别特征
- 10亿样本，2500万特征，LR1小时20分钟

- **灵活易用**

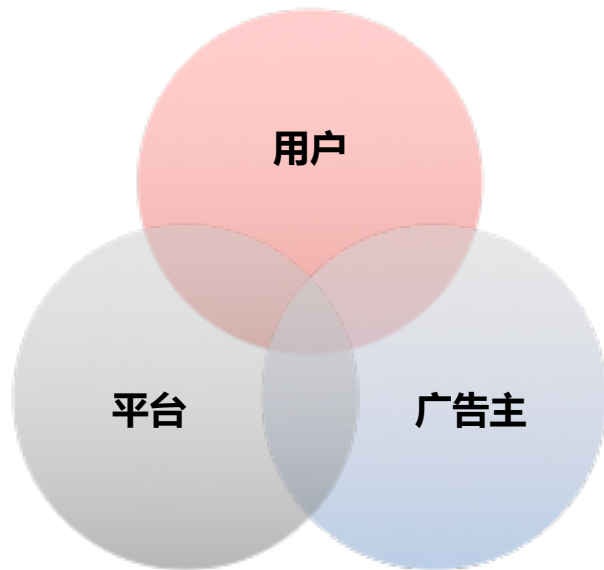
- 与公司数据平台对接，开放其他业务使用
- 支持在线预测

- **更多特性**

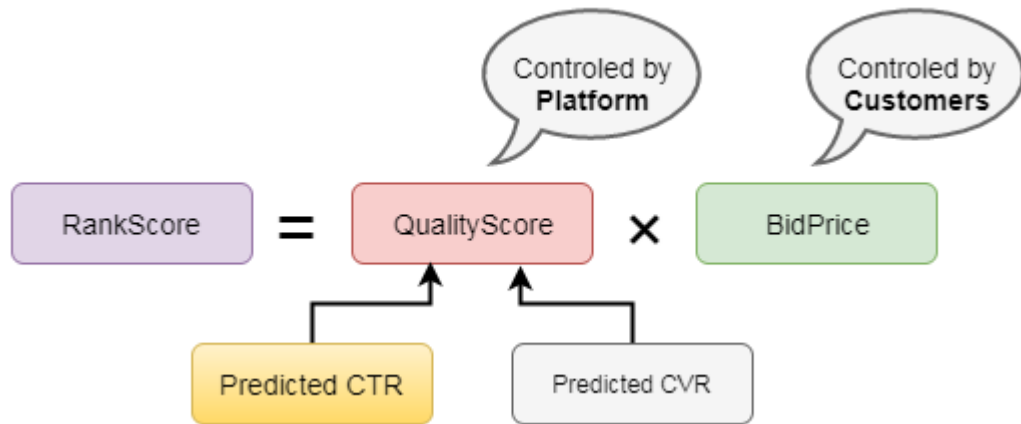
- 支持优化算法和模型多种组合方式
- 支持FFM定制化Field组合

- 业务背景
- 机器学习平台
 - CTR预估平台
 - 模型平台
- 排序机制
- 总结思考

- 如何综合考虑各个因素对广告排序的影响



- 综合考虑CTR与CVR



- $$\text{RankScore} = \text{CTR} * (a + b * \text{CVR}) * \text{BidPrice}$$

- 优点：

- 转化率提高较大

- 缺点：

- 参数不够稳定
 - 收入下降
 - 出价固定，无法差异化流量价值

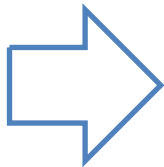
- **问题**：如何为不同流量出不同的钱

- **业界解决方案**

- oCPA：腾讯、头条
- oCPM：Facebook
- oCPC：淘宝

- **O2O广告特点**

- 广告主对互联网广告不了解
- 流量类型多样、差异大，多个出价

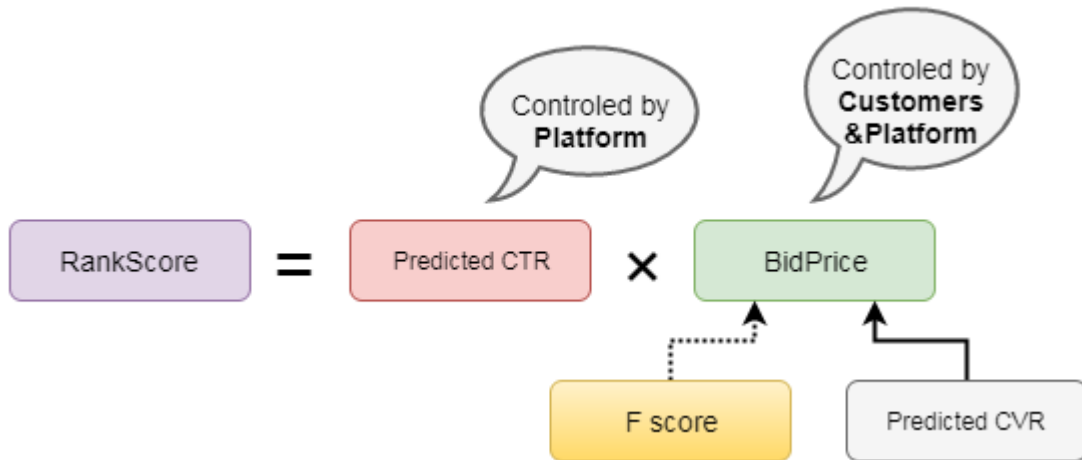


- 统一出价：简单
 - 自动打折，对广告主透明
- oCPC排序：双赢
 - 综合优化平台及广告主收益

- oCPC算法：参考2017 KDD 《Optimized Cost per Click in Taobao Display Advertising》

- 原理：

- eCPM排序
- 保证roi的前提下，调整出价b*，优化业务目标F



- 如何构造F函数：

$$f(k, b_k^*) = pctr_k * b_k * (1 + \sigma(\frac{pcvr_k * \|A\|}{\sum_{i \in A} pcvr_i}, w) * r_a)$$

- 物理意义：

- 对于CVR高于自身历史期望水平的广告提高出价
 - 对CVR高于候选平均水平的广告提权

- 如何处理冷启动？

- 品类+位置近似CVR + 阈值限制

- 优点

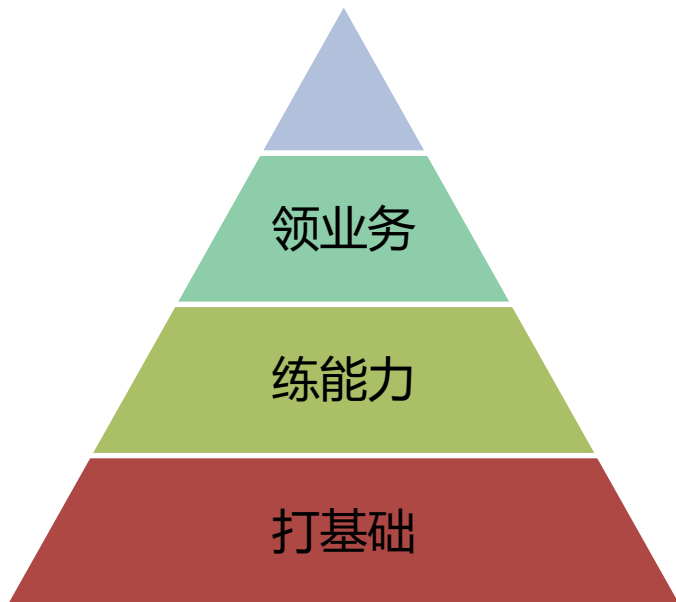
- 差异化流量价值，双赢：收入提升、ROI提升
 - 参数鲁棒，对Scale不敏感



- 单目标优化到多目标优化
- 不仅是广告
 - 智能营销

- 业务背景
- 机器学习平台
 - CTR预估平台
 - 模型平台
- 排序机制
- 总结思考

- 推荐广告机器学习优化之路：



- 心得体会
 - 立足业务
 - 重视基础
- 团队口头禅：一切都是原因的

Thanks !

Q&A

