个性化推荐系统的演变

吴官林

1 推荐业务背景介绍

2 精准推荐架构的演进

3 推荐引擎设计与实现

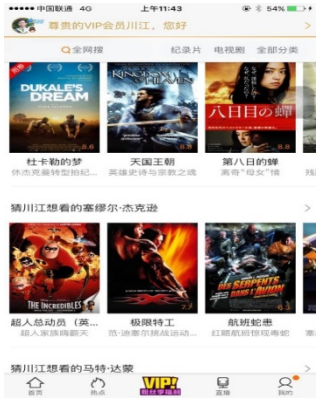
4 未来展望

个性化@Everywhere

什么是个性化推荐？

效果广告

视频推荐



新闻推荐

电商推荐

海量@唯品会

海量货品，无搜索限量抢购模式，用户需要花费大量时间

数千档期

数百万商品

数次浏览

男士
百万级

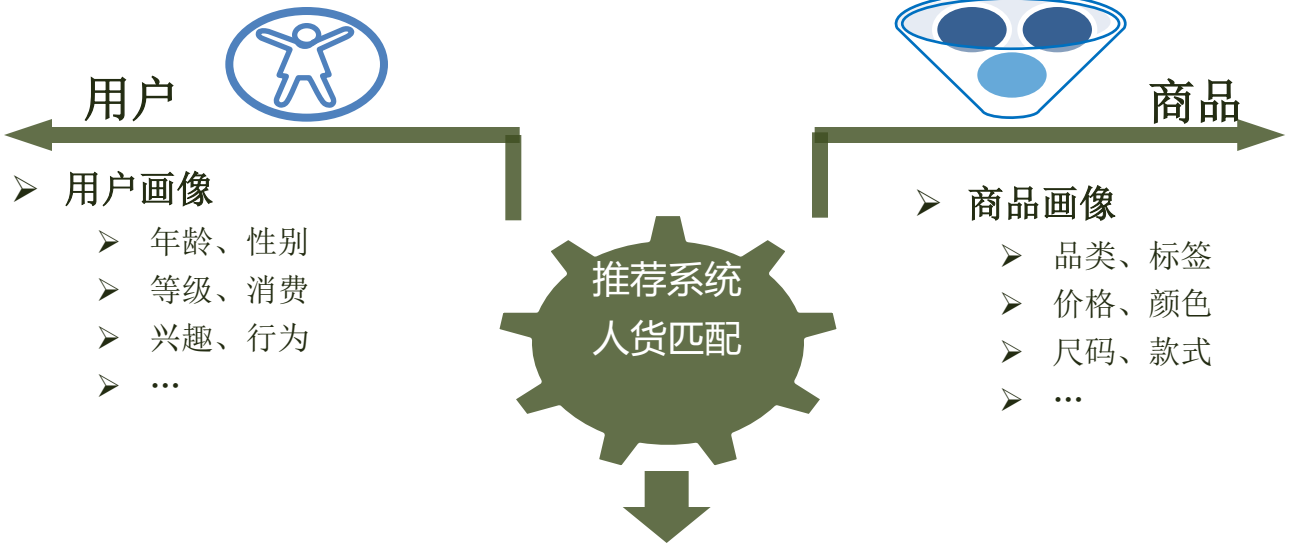
90后
百万级

70后
百万级

主流
千万级

海量用户，四大人群精细化运营，不能很好的满足用户偏好

个性化目的



- 用户画像
  - 年龄、性别
  - 等级、消费
  - 兴趣、行为
  - …
- 商品画像
  - 品类、标签
  - 价格、颜色
  - 尺码、款式
  - …

90后 - **夏娜Shinena**：甜美、俏皮、靓丽；单价100+ ~ 300

80后 - **丽丽Lily**：简约、职业、个性；单价300 ~ 500

70后 - **雅莹EP**：成熟、稳重、调性；单价1000

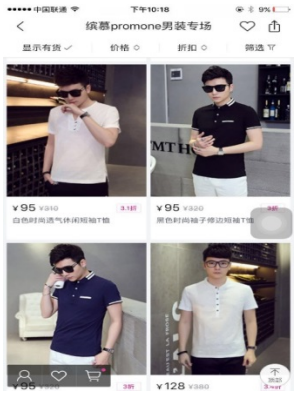
个性化目标

单位访客金额 = 金额/访客 = 转化率（购买人数/访客人数） × 客单价(购买单价)

单位曝光价值 = 金额/曝光 = **CTR** × **CVR** × **Price**

CTR = P(click/impression)
- 品牌熟悉吗？
- 图片漂亮吗？

CVR = P(order/click)
- 款式喜欢吗？
- 价格实惠吗？



我们所做的

实时 预测 每个人的 未来

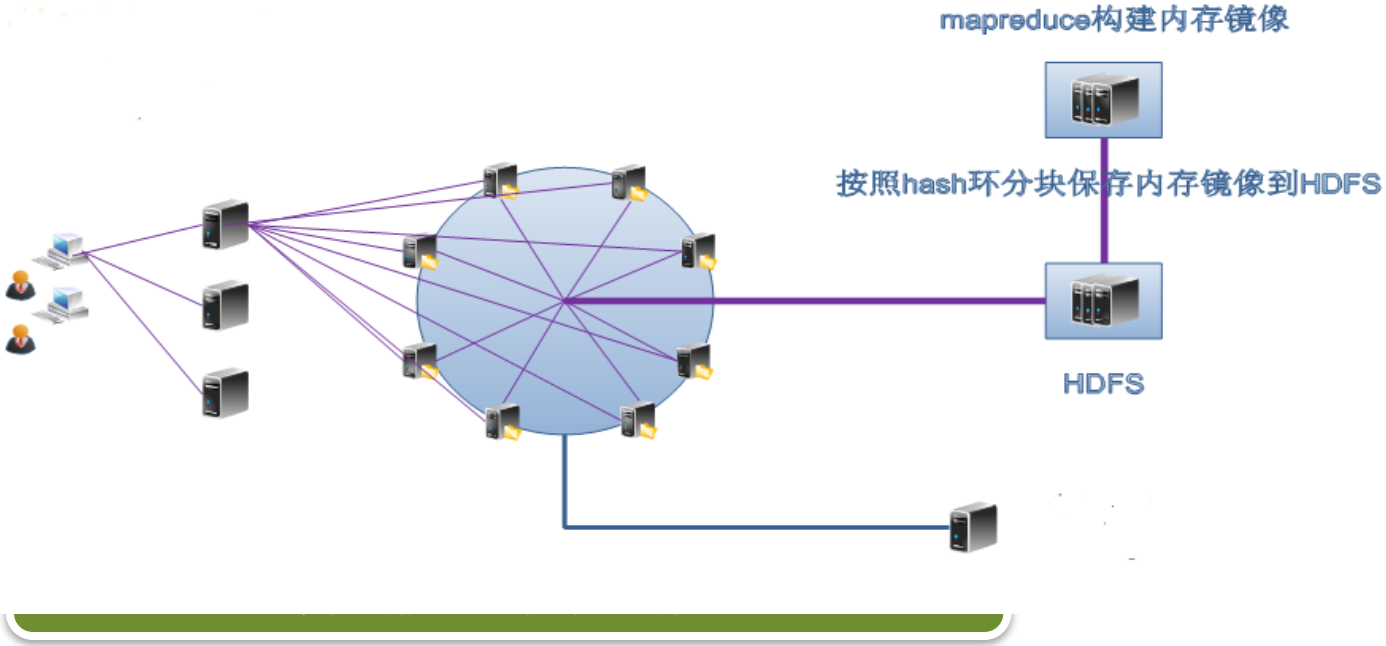
核心挑战

100MS,KW活跃用户，如何精准预测

技术难点

大量档期、大量用户、100MS时间、10亿次预测请求

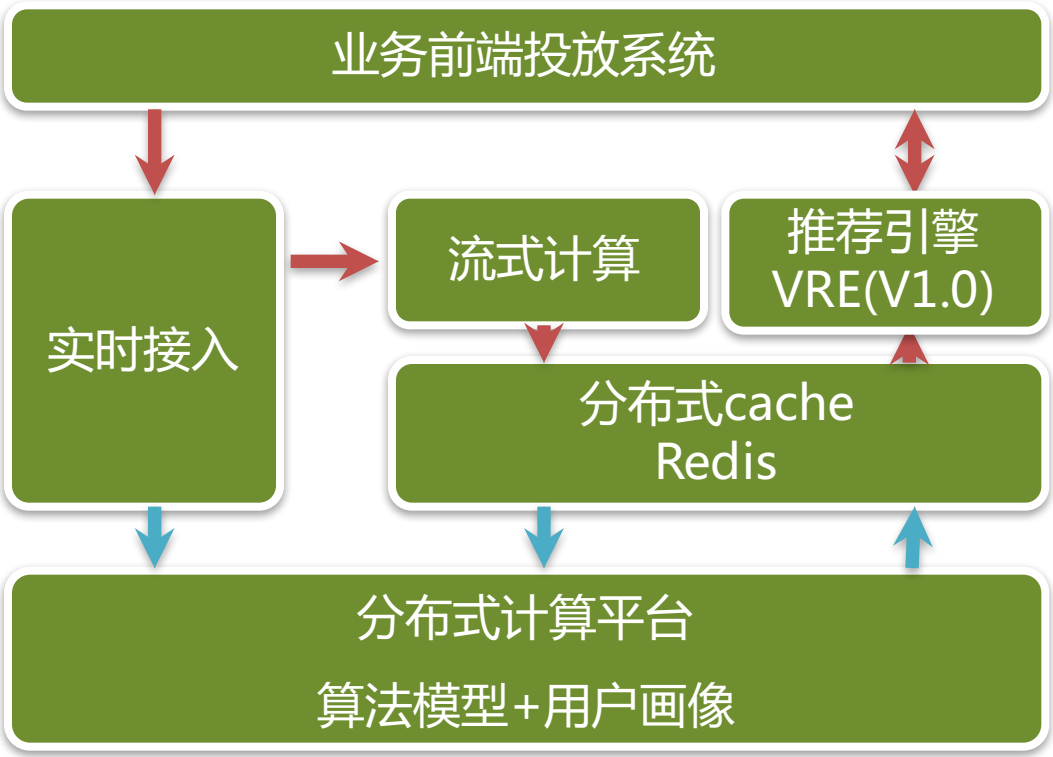
精准推荐第一代架构



- **核心特征**：
  - 离线文件传输
  - 批量暴力计算
  - 实时匹配查询

- **存在的问题**：
  - 数据时延高
  - 人群聚类粗
  - 扩展性差

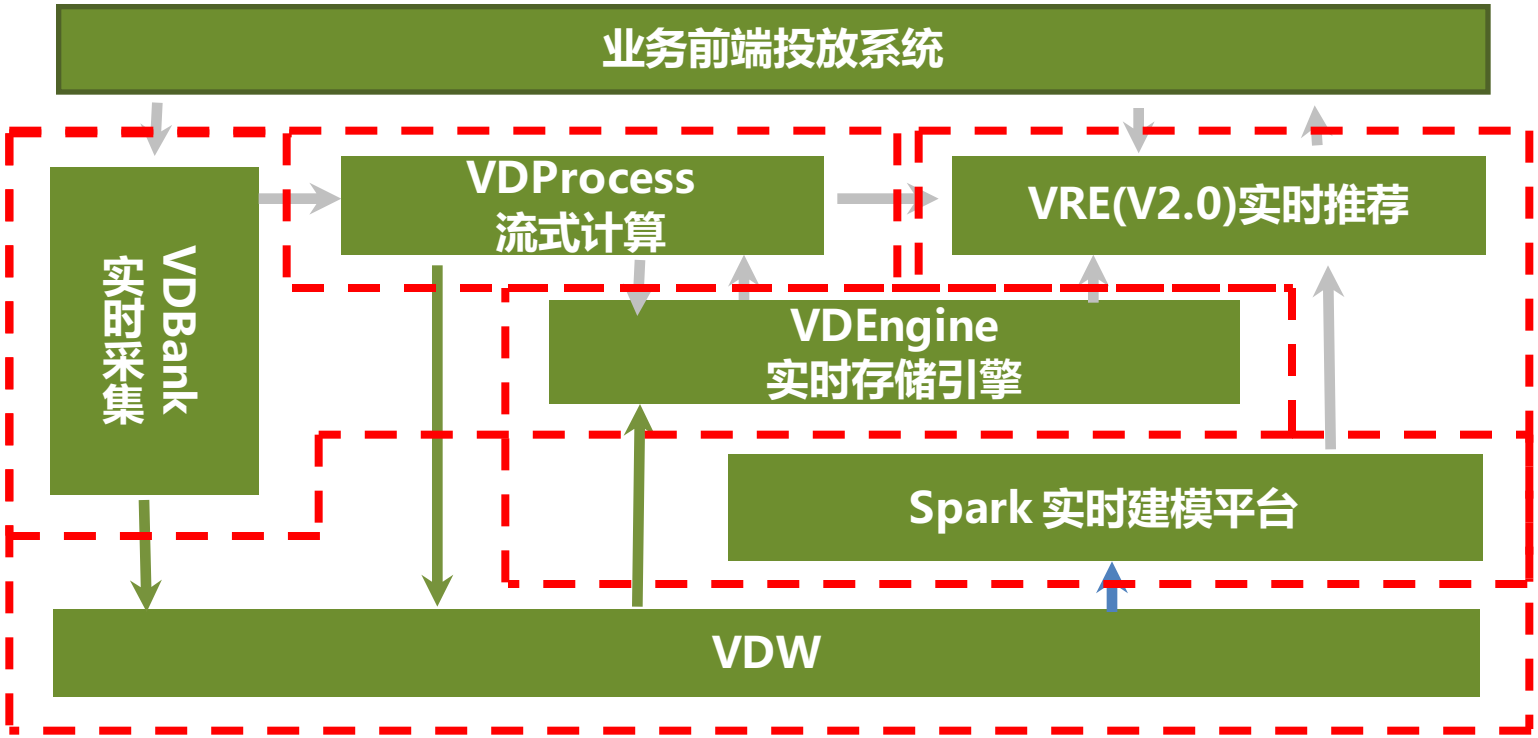
精准推荐第二代架构



- 特征：
  - 实时数据传输
  - 在线实时计算

- 优点：
  - 秒级延迟
  - 低耦合
  - 易扩展

精准推荐第三代架构


业务前端投放系统
VDBank
实时采集
VDProcess
流式计算
VRE(V2.0)实时推荐
VDEngine
实时存储引擎
Spark 实时建模平台
VDW

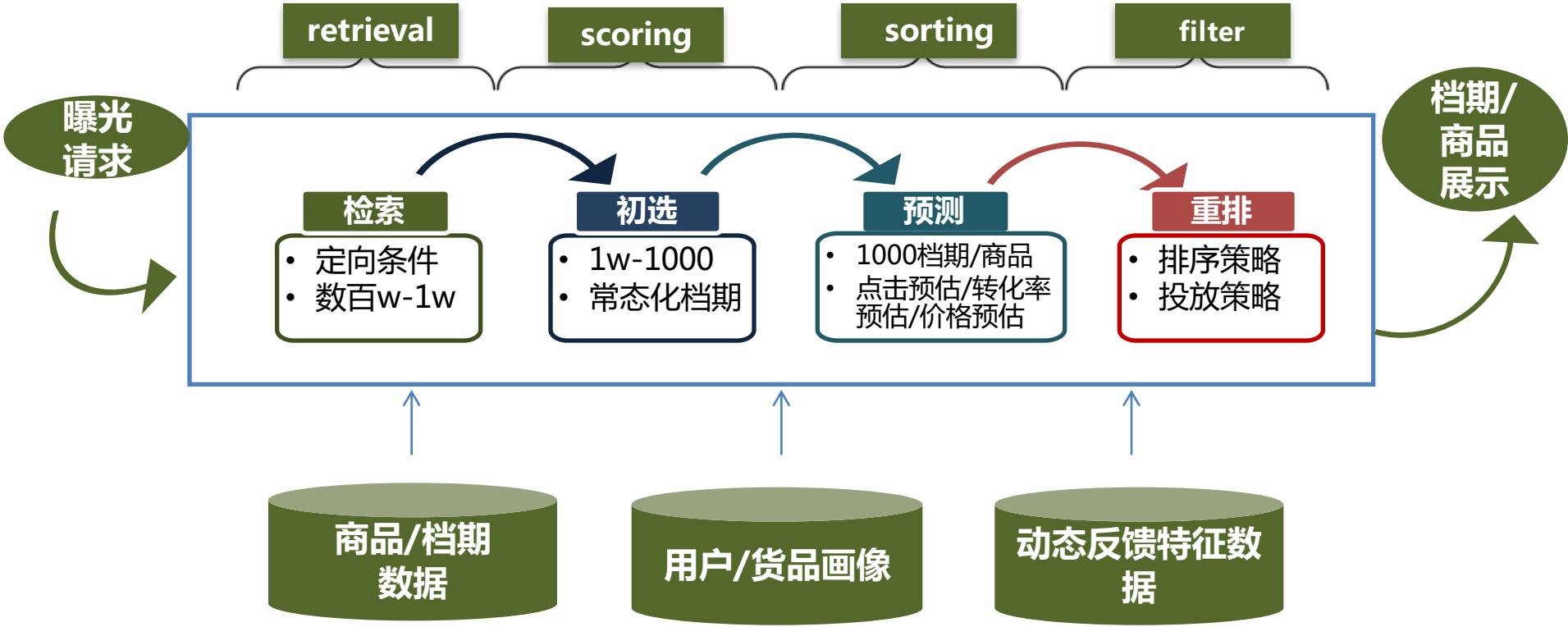
VRE推荐引擎挑战

特卖会个性化推荐的**核心技术挑战**：

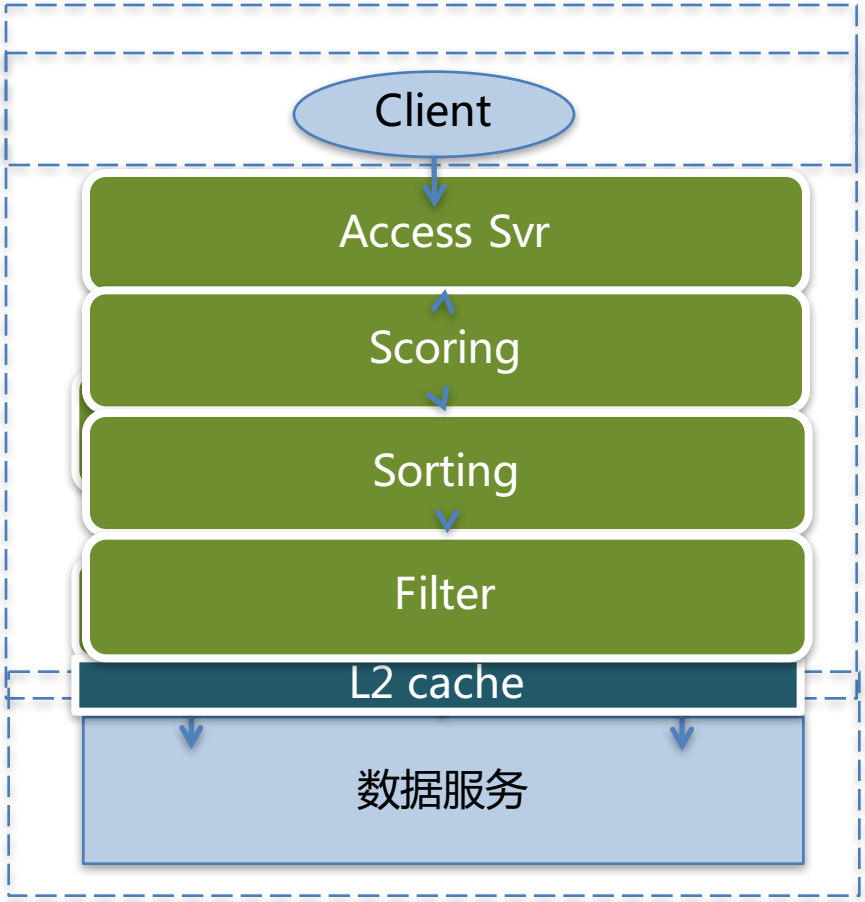
如何在100ms的时间内， 面对数kw活跃用户中的任意一人，预测其在未来下一秒内，最可能购买上w个档期中所包括的数百w个商品中的哪一批？

- “数亿用户↔ 数百万实时变化商品”的相关度计算非常复杂
- 100ms内完成
- 特卖会每天执行数十亿次这样的预测

推荐实时计算在线业务流程



VRE一代架构



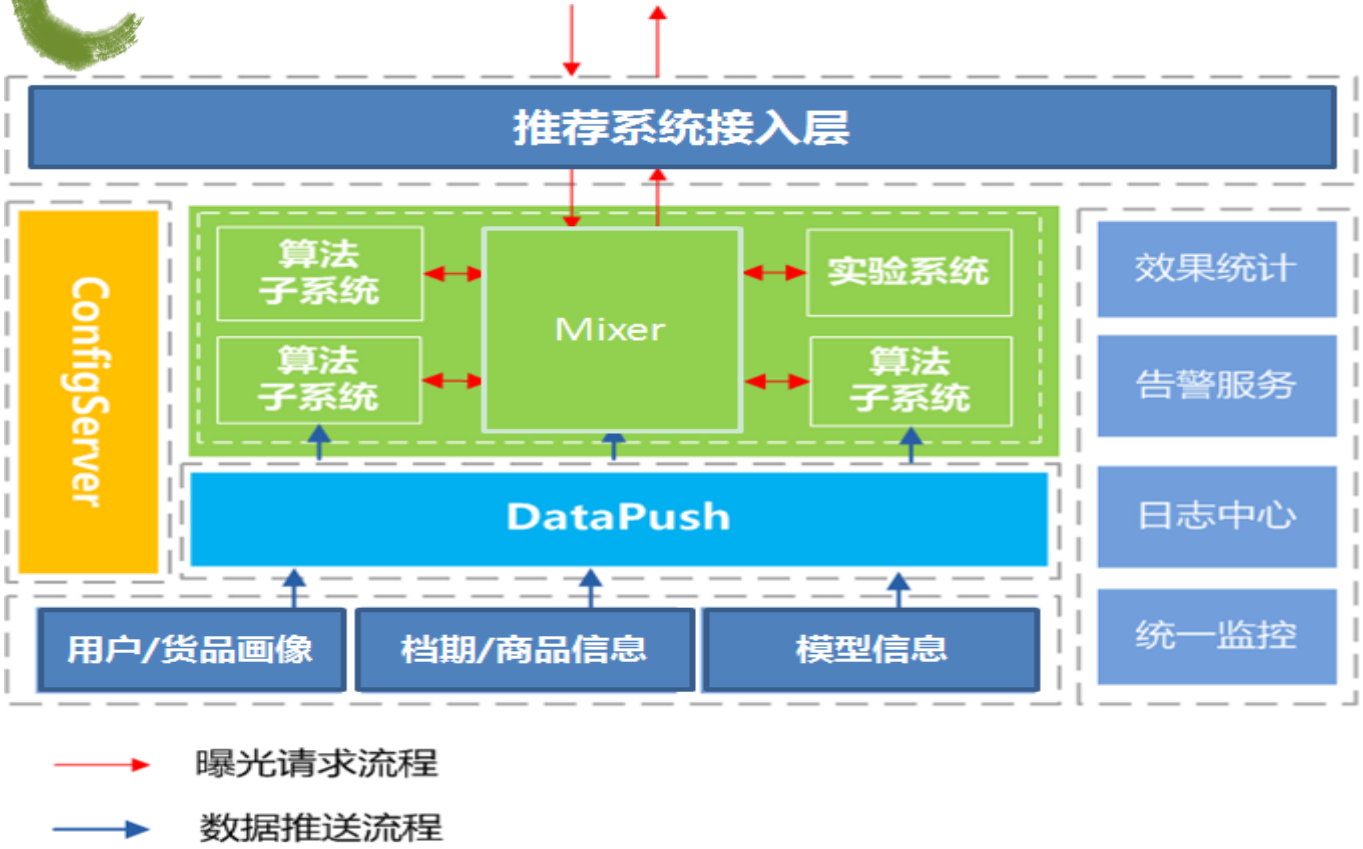
核心需求

- 海量+实时：计算复杂响应时耗要求高；
- 大规模分布式系统流量、算法和数据管理。

主要痛点

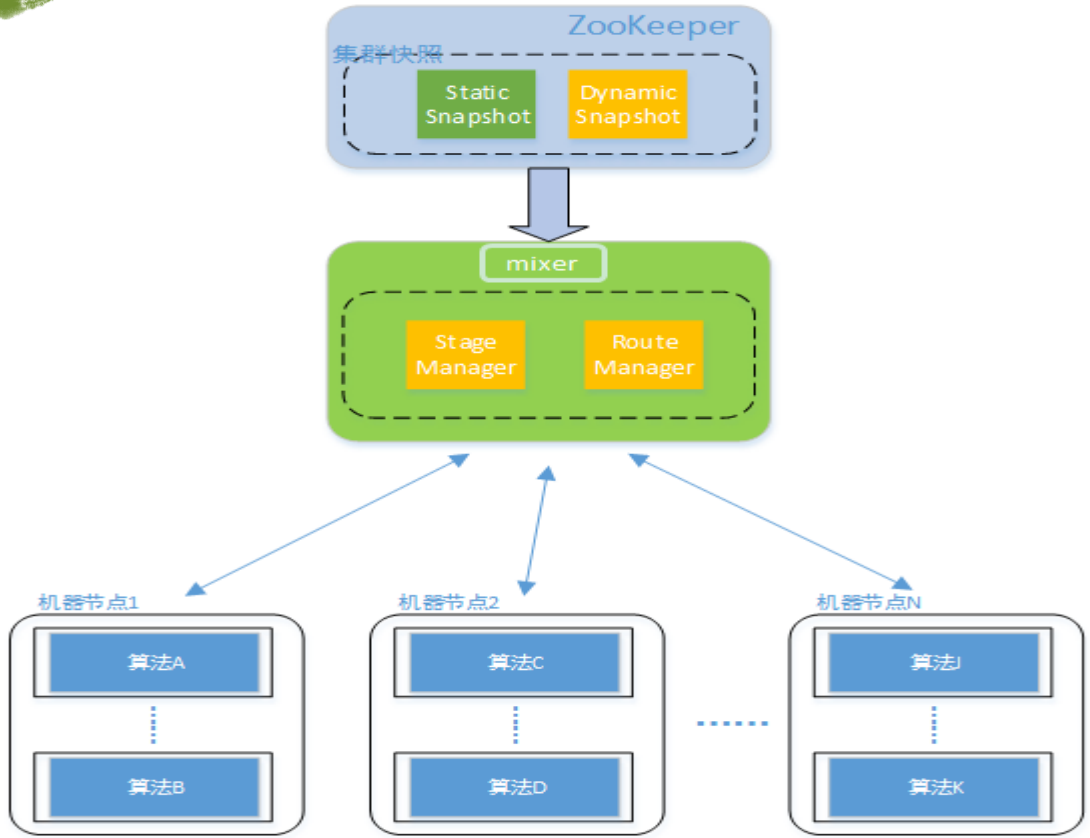
- 扩展难;
- 静态路由，关联系统各自为政;
- 可运营性差。

VRE二代架构



- **主要优点**
- 易扩展；
- 动态智能路由；
- 集群透明。

VRE算法管理



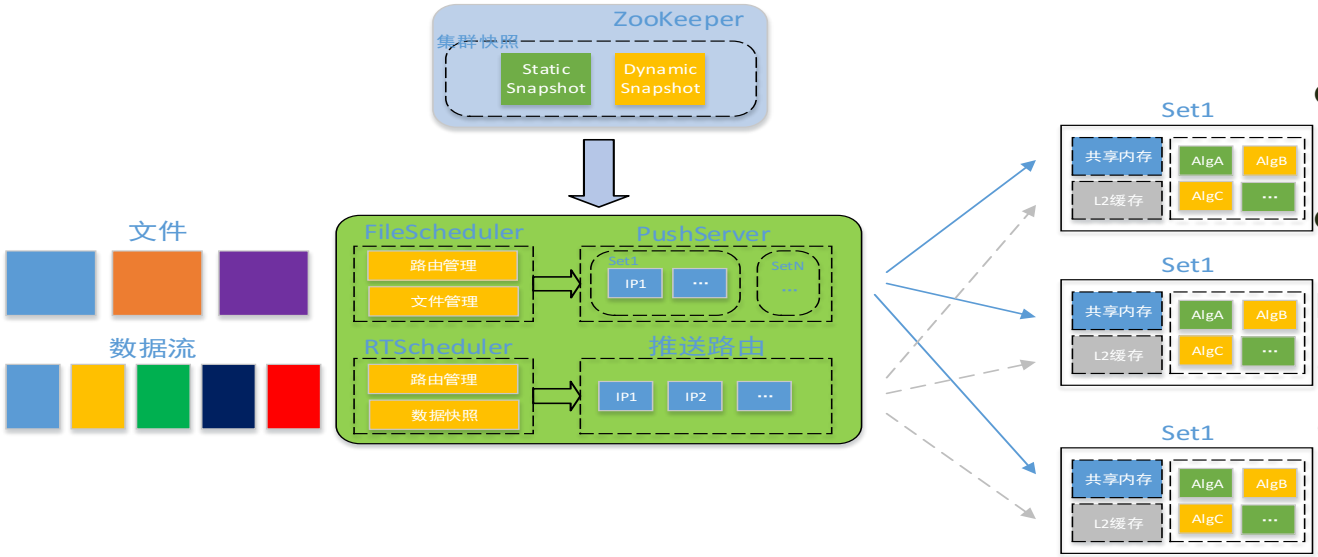
核心需求

- 针对不同流量建模
- 支持100+在线，实验算法效果调优，频繁更新，上下架
- 支持多种业务流程

特色功能

- 插件式管理
- 支持动态上下架
- 算法作为系统调度的路由依据
- 接口解耦，状态机分离
- 配置驱动流程调度

VRE数据更新



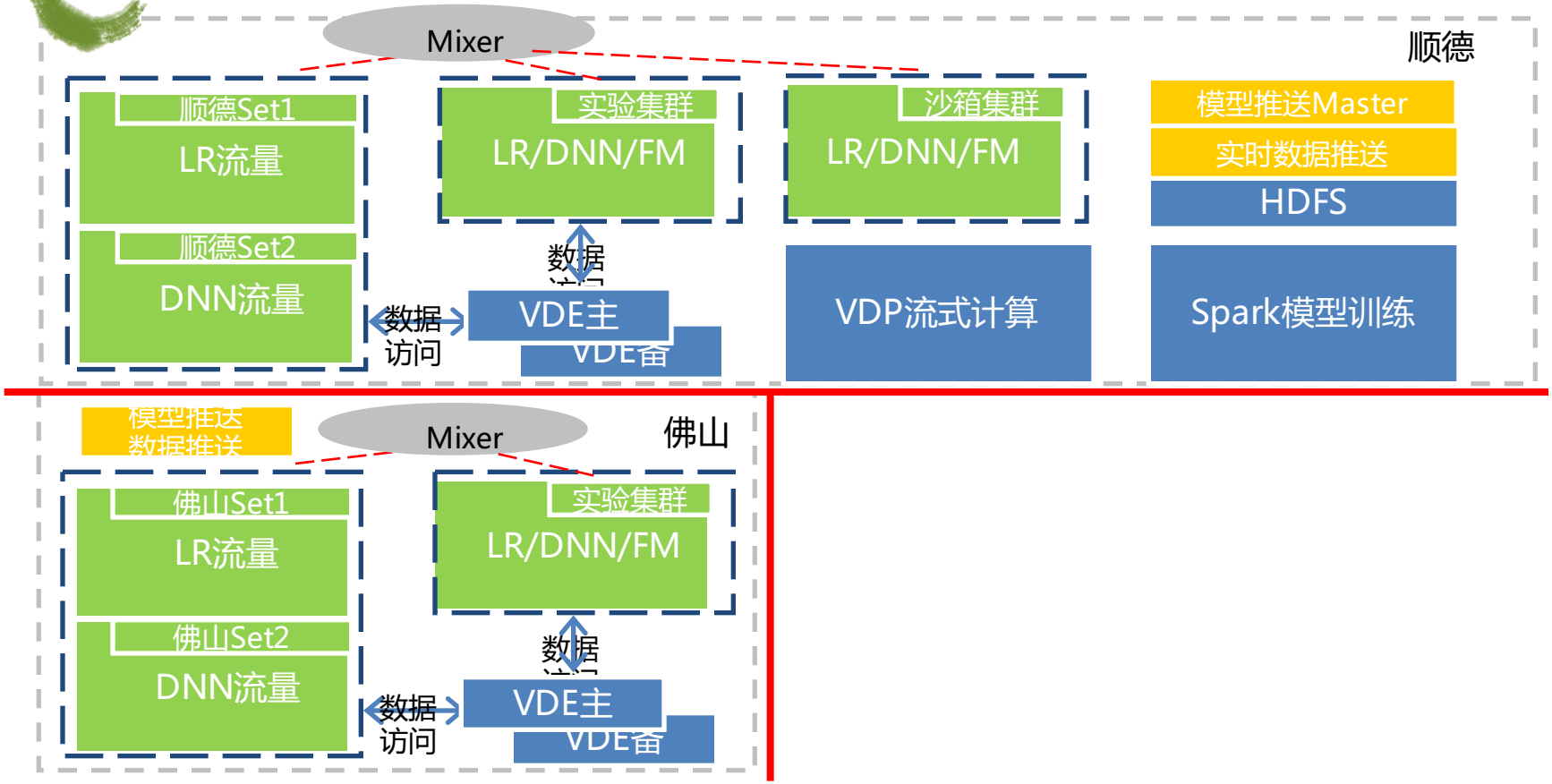
核心需求

- 保证算法模型一致性
- 支持100+在线实验算法模型数据推送
- 每15分钟粒度定期推送到线上几百台服务器
- 保证关键性数据查询命中

特色功能

- 文件Pipeline任务调度
- 100MB文件，10S内发送到集群所有集群上
- 实时流数据毫秒级更新到server缓存
- 数据快照，支持计算节点快速恢复

VRE多地容灾


Mixer
顺德
顺德Set1
LR流量
顺德Set2
DNN流量
实验集群
LR/DNN/FM
沙箱集群
LR/DNN/FM
模型推送Master
实时数据推送
HDFS
数据
数据
访问
VDE主
VDE备
VDP流式计算
Spark模型训练
模型推送
数据推送
Mixer
佛山
佛山Set1
LR流量
佛山Set2
DNN流量
实验集群
LR/DNN/FM
数据
数据
访问
VDE主
VDE备

未来展望

- **更加实时**
  - Online Model Training
  - 实时用户画像
- **更加精准**
  - 更多特征
  - 深度学习
- **更加通用**
  - 多业务支撑
  - 第三方开放