阿里广告中的机器学习平台

昙宗

个人简介

- 蒋龙
- 花名：昙宗
- 阿里巴巴高级技术专家

- 2006-2011，微软亚洲研究院自然语言计算组
  - 微软对联，机器翻译，实时搜索
- 2011-，阿里巴巴广告部（阿里妈妈）
  - 研发广告中的机器学习和NLP技术

- Email: tanzong.jl@taobao.com
- Weibo：蒋龙_淘昙宗

提纲

- 广告中的机器学习
- 阿里广告机器学习平台
  - 特征平台
  - 算法平台
    - MPI并行计算集群
    - 算法平台
    - 评测平台
  - 应用现状

搜索广告

定向广告

热卖单品 精品凉鞋 热卖女鞋 新款凉鞋 坡跟女凉鞋 休闲凉拖 女鞋凉鞋 美鞋 拖鞋 性感鱼嘴 舒适单鞋 更多热卖

疯抢!漫步云端 女士包臀收腰蕾丝花边连衣裙

￥598.00

Amovo纯可可脂巧克力8口味礼盒 零食 包邮

￥109.00

思加图 专柜正品 2012夏季新款拖鞋凉拖凉鞋

￥238.00

0利润大牌高档退换包邮真丝连衣裙送彩票

￥800.00

您可能对这些宝贝感兴趣

排序分数：

f(点击率，竞价)

=>点击率预估p(click|ad,user)

商品推荐

主宝贝

排序分数：
f(转化率，商品价格)
=>转化率预估
p(buy|item,user,curItem)

相关推荐

包邮Apple/苹果 iPhone 5 联通电信苹果5 iphone5 限时抢购

¥ 4699.00

Apple/苹果 iphone 4 (8G) 苹果4代 8G 行货正品全国联保

¥ 2899.00

MIUI/小米 M2小米二代小米2四核顶配 正品行货手机16G 现货

¥ 2099.00

顺丰包邮Samsung/三星 GALAXY Note II N7100 Note2 牛2 正品现

¥ 3894.05

搭配推荐

洛克rock 苹果5 iphone 5手机壳 TPU软壳 保护壳保护套 畅动系列

¥ 48.00

手机剪卡器 搭配手机一起买价格15元

¥ 30.00

Zupoo/触宝 苹果4 iPhone4/4S iphone4s保护膜屏幕贴膜 晶透系列

¥ 25.00

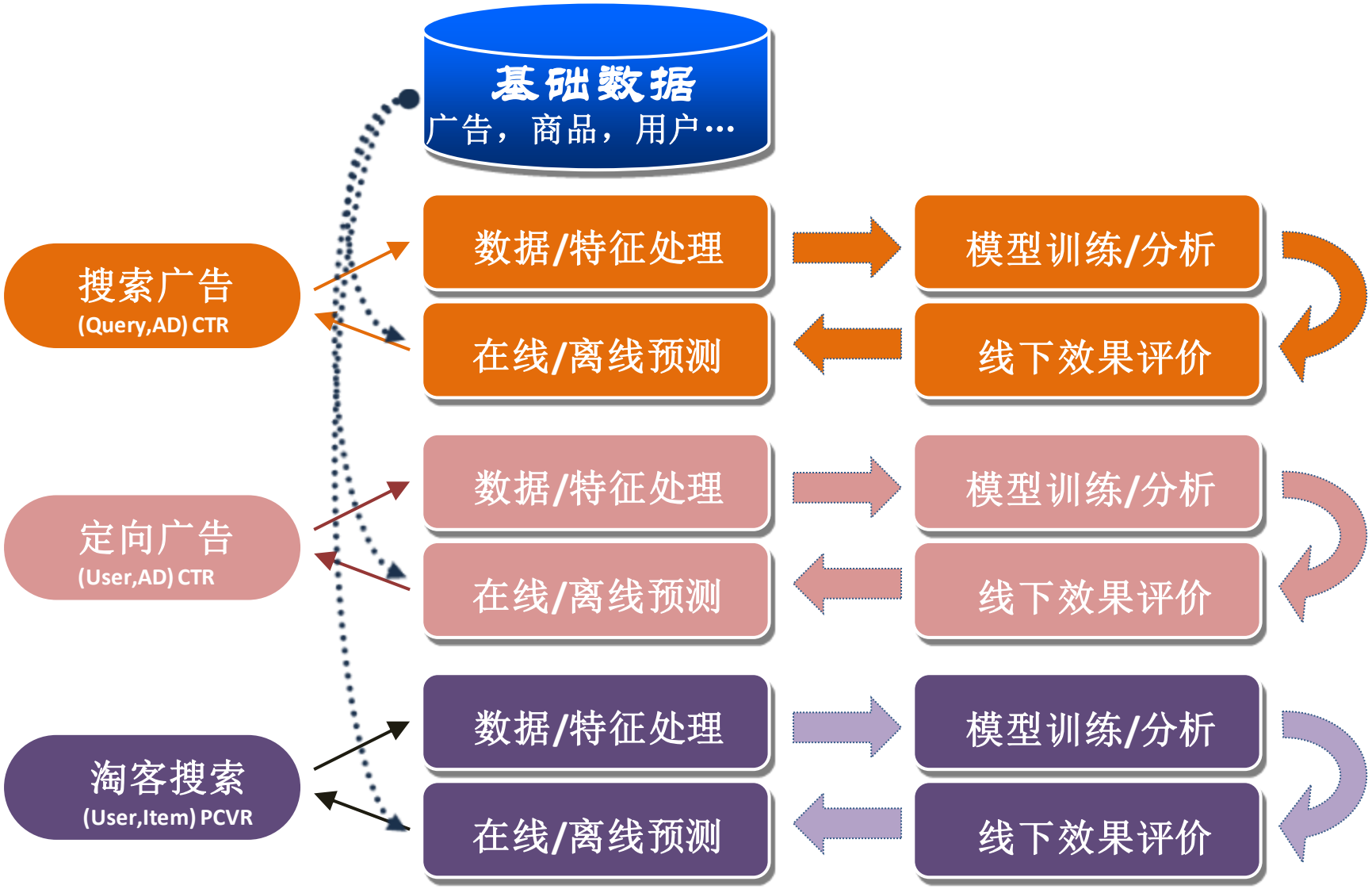
SanDisk/闪迪TF 16G TF卡 MicroSD手机内存卡存储卡70元

¥ 70.00

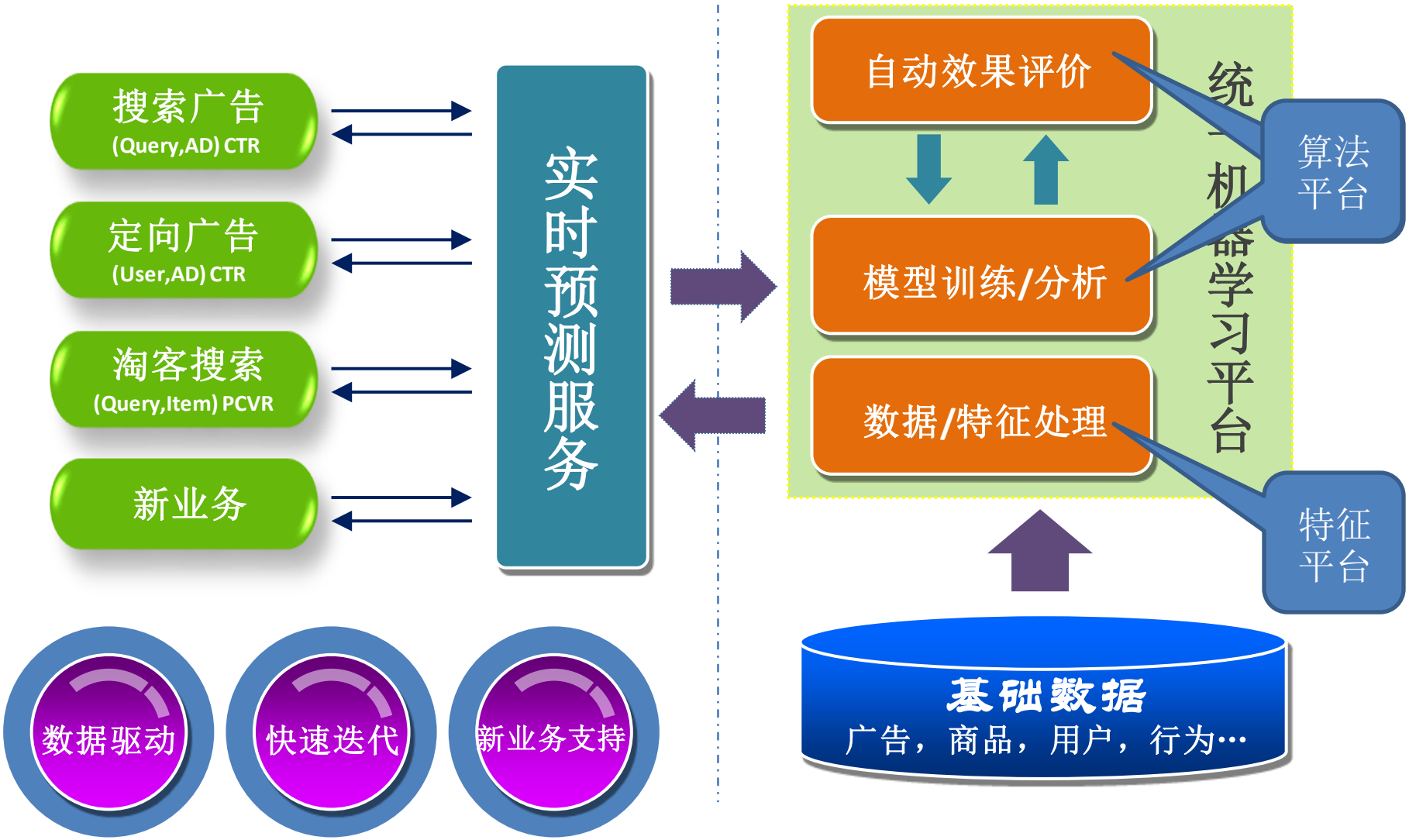
广告中的机器学习

- 点击率预估
  - 搜索广告：p(click|ad,query)
  - 定向广告：p(click|ad,user)
- 转化率预估
  - 淘客：p(buy|item,query)，p(buy|item,user)
  - 推荐：p(buy|item,user,curitem)
- 相同的流程
  - 基础数据+日志数据，特征抽取和处理，训练数据收集，模型训练，模型线下评估

重复的数据流


基础数据
广告，商品，用户…
搜索广告
(Query,AD) CTR
数据/特征处理
模型训练/分析
在线/离线预测
线下效果评价
定向广告
(User,AD) CTR
数据/特征处理
模型训练/分析
在线/离线预测
线下效果评价
淘客搜索
(User,Item) PCVR
数据/特征处理
模型训练/分析
在线/离线预测
线下效果评价

统一机器学习平台

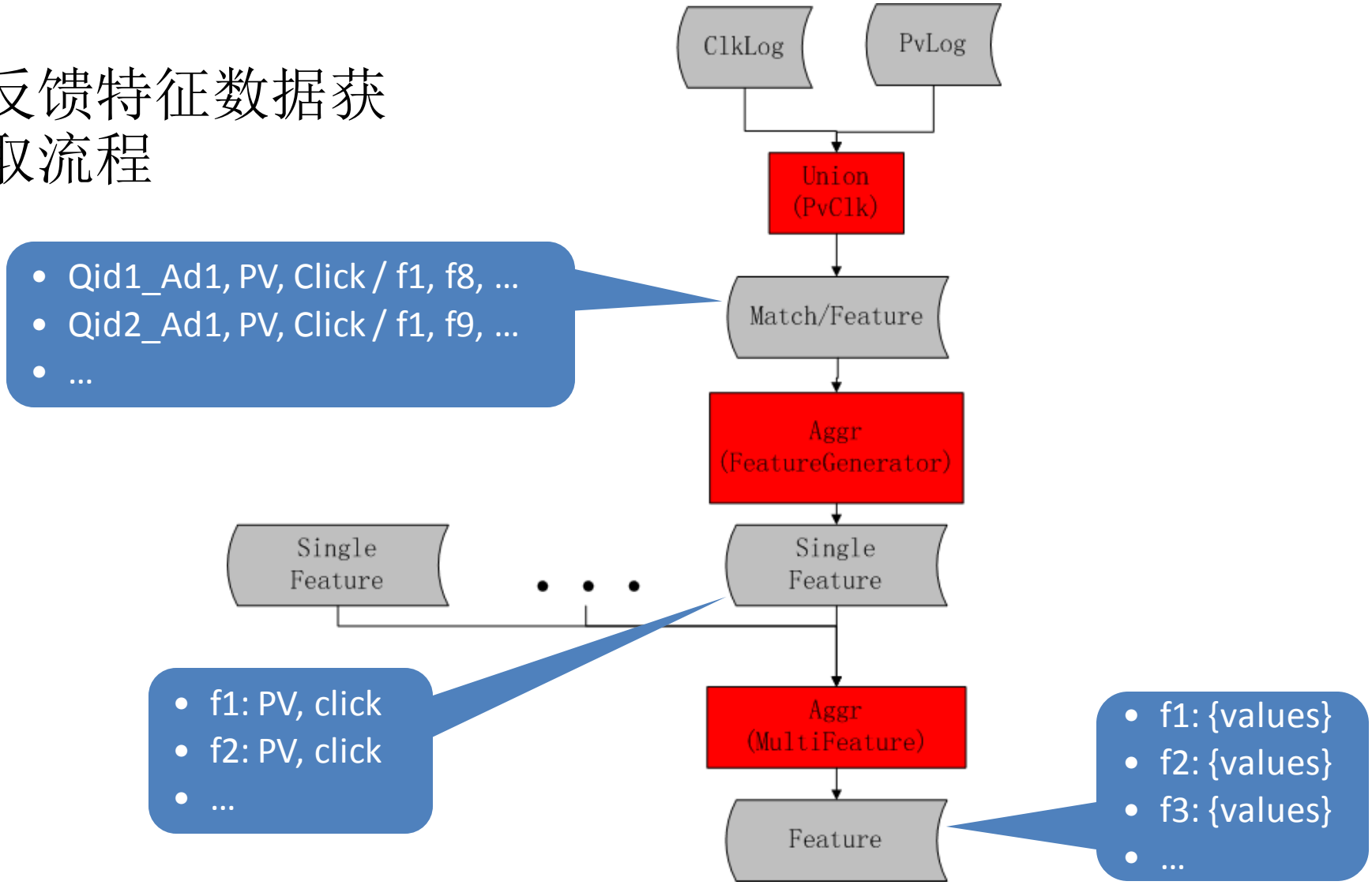

搜索广告
(Query,AD) CTR
定向广告
(User,AD) CTR
淘客搜索
(Query,Item) PCVR
新业务
实时预测服务
自动效果评价
模型训练/分析
数据/特征处理
统一机器学习平台
算法平台
特征平台
基础数据
广告，商品，用户，行为…
数据驱动
快速迭代
新业务支持

特征平台

- 特征处理
  - 特征抽取
    - 公用数据(Query, Ad, Item, Shop, User)
  - 特征处理
    - 变换，规范化，交叉组合，离散化
  - 特征分析
    - 覆盖率，均值，方差
    - 相关系数，Chi-square，互信息

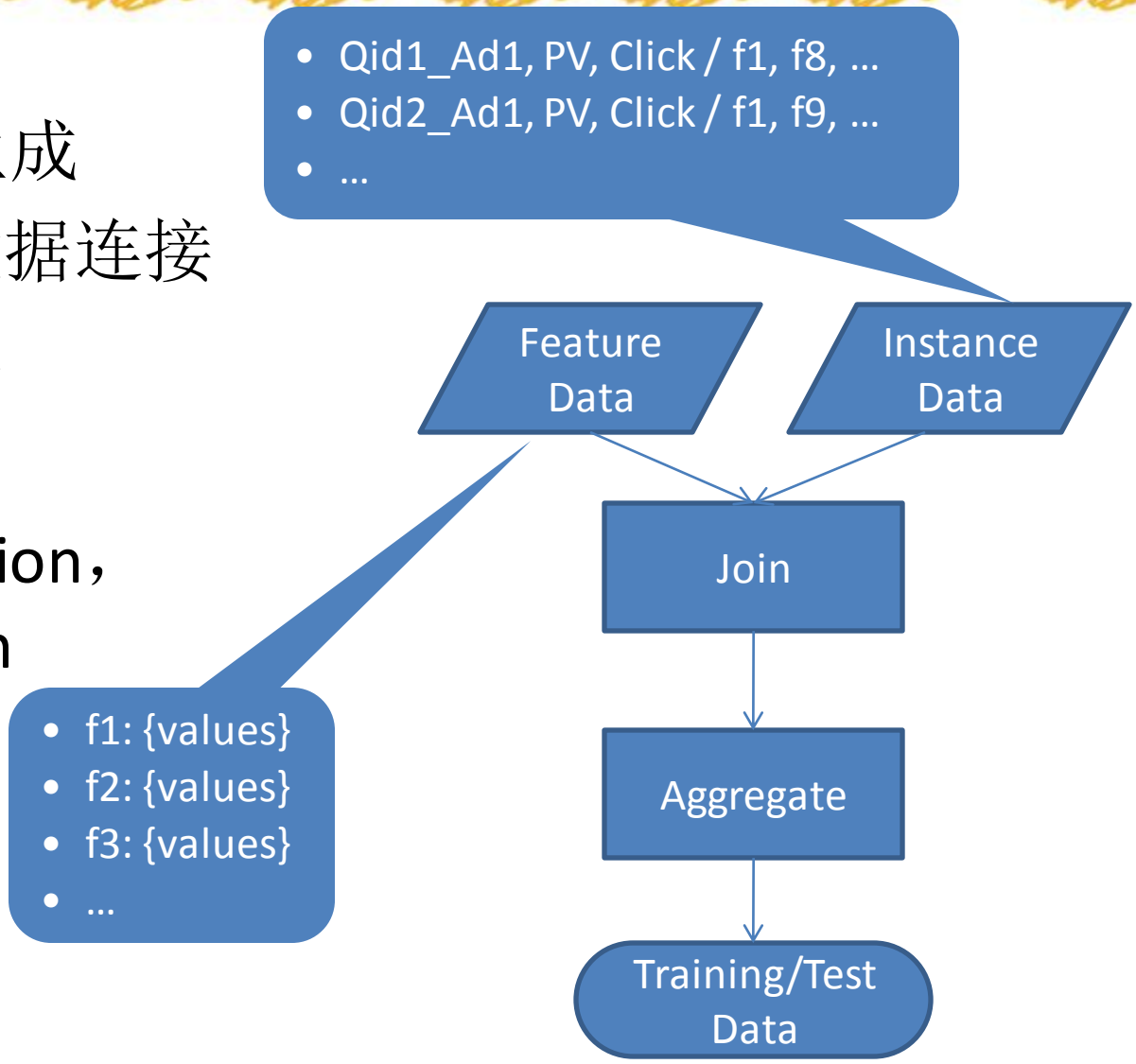
特征平台

- 反馈特征数据获取流程



特征平台

- 训练/测试数据生成
  - 目标和特征数据连接
  - 训练数据选择
  - 多天数据集成
- 大数据操作：Union，Aggregation，Join



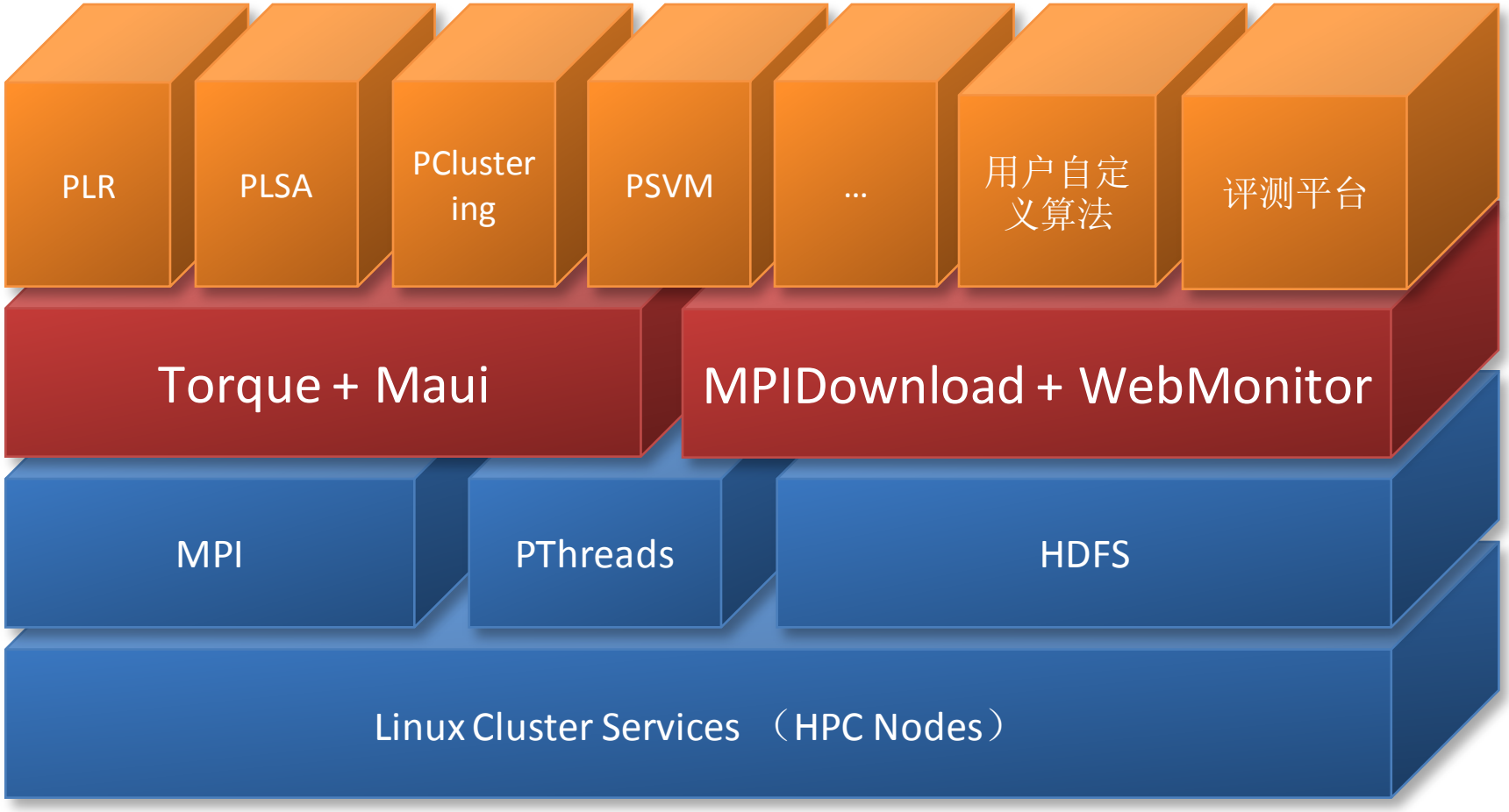
基于Hadoop的特征平台

- 特征平台特点
  - 海量数据
    - 原始日志，训练数据：几十T
  - 计算逻辑简单，迭代少
  - 稳定，容错
- 基于Hadoop开发
  - 大规模MapReduce+HDFS集群
    - 阿里云梯：4000+节点，~80PB存储

算法平台

- 算法平台：模型训练和评测
  - 海量数据，大规模特征
  - 计算逻辑复杂，迭代过程很多
  - 速度要求很高
- 基于MPI框架
  - 数据常驻内存
  - 单机高性能，大内存
  - 允许自定义复杂的数据通信

基于MPI的算法平台


PLR
PLSA
PClustering
PSVM
...
用户自定义算法
评测平台
Torque + Maui
MPIDownload + WebMonitor
MPI
PThreads
HDFS
Linux Cluster Services （HPC Nodes）

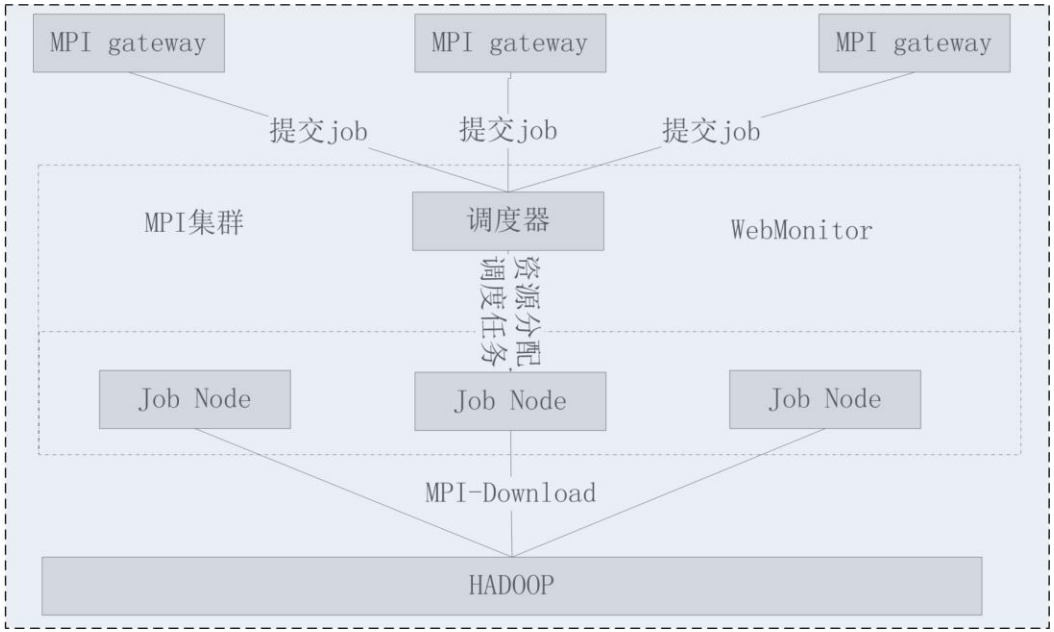
阿里MPI集群

- 集群概况
  - ~500台服务器
    - 单机12核（超线程24核），100G内存，千/万兆网卡
  - 生产、开发、测试集群
  - 部署MPICH2框架
  - 共享云梯存储
    - MPI-Download：统一管理与Hadoop的数据交换
      - 分布式下载，自动解压合并，自动流量控制

MPI集群任务管理

- 使用Torque+Maui为基础的调度器
  - Job模块自动分发，计算资源均衡
  - 调度队列区分优先级
- Cgroups确保资源隔离
  - 动态资源调整

整体架构



并行机器学习算法

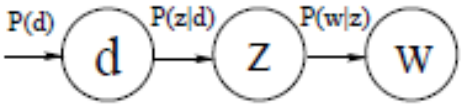
- 并行算法库
  - LR，MLR，LDA，PLSA，GBRT，SVM，MAXENT，CF，Spectral Clustering，…

基于MPI的并行PLSA

- PLSA (Probabilistic Latent Semantic Analysis)模型

$$p(w|d) = \sum_{z \in Z} p(w|z)p(z|d)$$

  - Z代表隐藏topics
  - <d,w>矩阵： d先产生z，再由z产生w



- 应用
  - 文本语义分析，用户兴趣分析

基于EM算法的参数估计

P(z1|d1,w1)=xx
P(z2|d1,w1)=xx
…
P(zk|d1,w1)=xx

– E-step

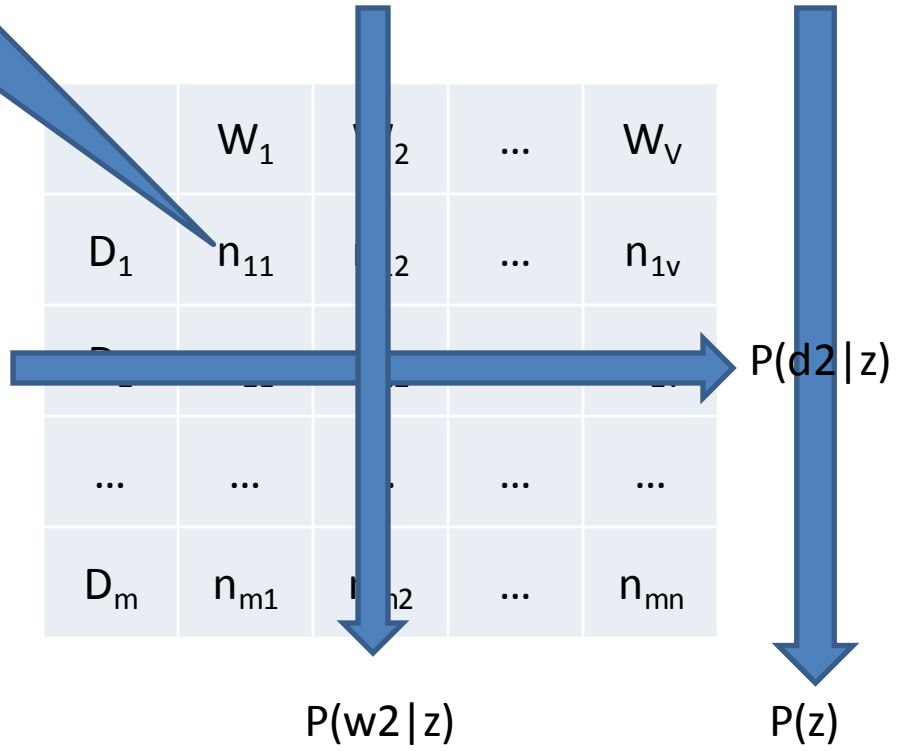
$$p(z \mid d, w) = \frac{p(z)p(d \mid z)p(w \mid z)}{\sum_{z' \in Z} p(z')p(d \mid z')p(w \mid z')}$$

– M-step

$$p(d \mid z) \propto \sum_{w \in W} n(d,w)p(z \mid d,w)$$

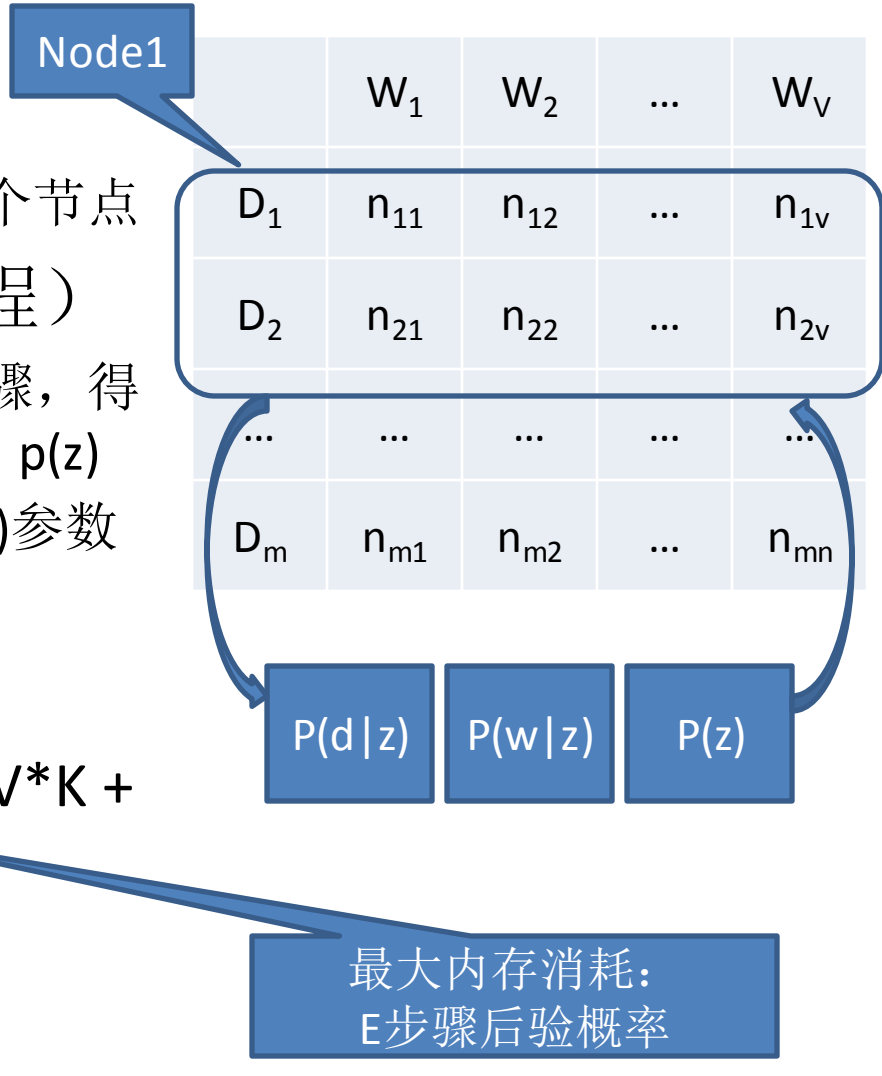
$$p(w \mid z) \propto \sum_{d \in D} n(d,w)p(z \mid d,w)$$

$$p(z) \propto \sum_{d \in D} \sum_{w \in W} n(d,w)p(z \mid d,w)$$



基于MPI的并行PLSA

- 数据并行化
  - <d,w>矩阵按行(d)均匀划分到每个节点
- 计算并行化（N个节点/进程）
  - 节点同步完成单机数据的E-M步骤，得到新一轮的p(d|z)和局部p(w|z)，p(z)
  - 节点间通信得到全局p(w|z)，p(z)参数更新
- 单节点内存占用
  - |<d,w>|/N + |<d,w>|*K/N + V*K + K*D/N + K



优化的并行PLSA

- 内存优化
  - E步M步交叉进行，不存储后验概率p(z|d,w)
    - 计算完每个词的p(z|d,w)，直接更新对应的p(d|z)，p(w|z)参数
    - 需要存储新旧两个模型p(d|z),p(w|z)
  - 按行(d)扫描数据计算，省下内存p(d|z)
    - 每行计算完直接在旧p(z|d)上更新新值
- E-step计算优化

$$p(z|d,w)=\frac{p(z)p(d|z)p(w|z)}{\sum_{z'\in Z}p(z')p(d|z')p(w|z')}$$

$$p(z|d,w)=\frac{p(w|z)p(z|d)}{\sum_{z'\in Z}p(z'|d)p(w|z')}$$

基于MPI的自动评测

- 基于MPI的模型评测
  - 无需回传model到hadoop
  - 训练完直接评测，也可单独评测
  - 自动化的参数调整
- 指标种类
  - MAE, MSE, Accuracy, Precision, Recall, AUC, GAUC

集群运行状况

- 使用方
  - 阿里妈妈事业部，搜索事业部
  - 天猫算法，交易算法团队
- 集群稳定性>**99.9%**
- 集群资源CPU周级峰值平均利用率60%+，平均利用率30%+
- 周Job数**1000+**

应用案例

- 基于MPI的LR/MLR模型
  - 搜索广告ctr预估
  - 淘客搜索CVR预估
  - Tmall推荐融合排序
  - 定向广告CTR预估

- 基于MPI版本PLSA模型
  - 外投广告个性化
  - 搜索个性化

Thanks!