




2016 杭州·云栖大会  
THE COMPUTING CONFERENCE

云栖社区  
yq.aliyun.com

# 阿里云机器学习PAI

## ——及其在广告营销中的应用

2016  
The Computing Conference

主办单位： 杭州

 Alibaba Group  
阿里巴巴集团

战略合作伙伴：

署名：褚巍（楚巍）  
职称：国家千人特聘专家



扫码观看大会视频

---

# 目录 content

---

- 一：产品简介
- 二：功能特点
- 三：案例分享



# 一、产品简介



# 机器学习平台产品



Amazon Machine Learning



Microsoft

Microsoft Azure Machine Learning



Alibaba Cloud



Platform of  
Artificial Intelligence



Google Cloud Platform



阿里云·数加  
aliyun.com

解决方案 产品 可视化 合作与活动 支持中心 控制台 basedem\*\*\*\*@aliyun... 三

## 大数据体验馆

免费体验 + 教程 = 大数据零距离！

### 推荐引擎

查看教程

敬请期待

### BI报表

查看教程

免费体验

### 机器学习

查看教程

免费体验

### 智能服务

机器翻译  
人脸识别 OCR

免费体验

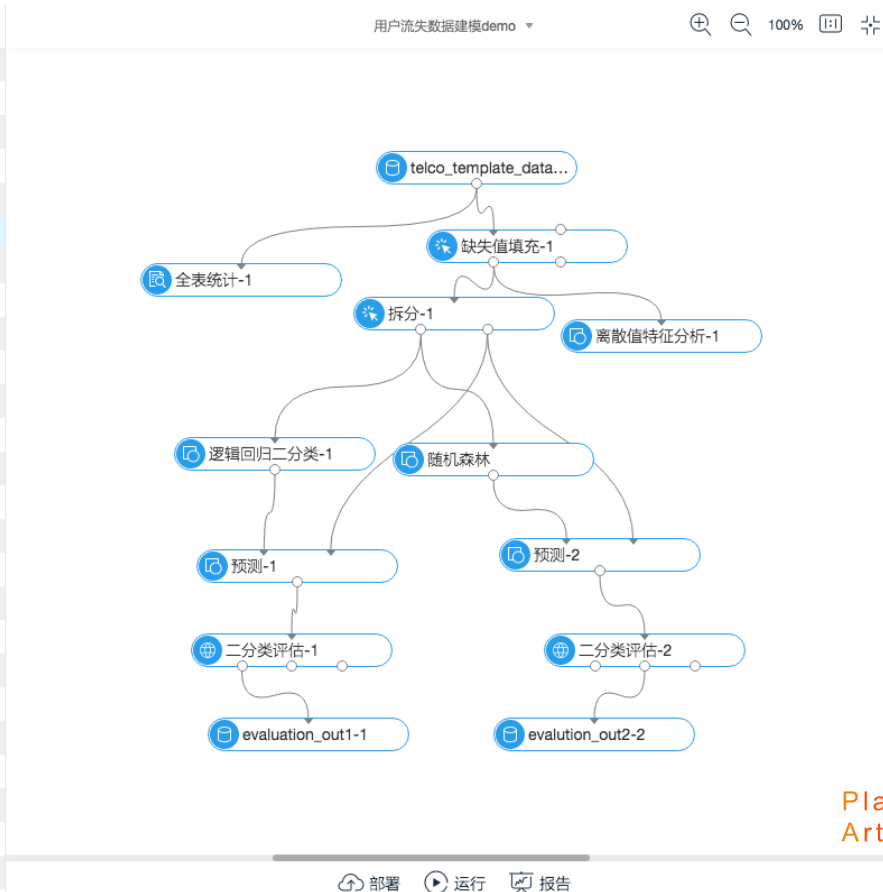
### 数据开发

查看教程



allpaytech\_dev

- 常用组件
- 源 / 目标
- 数据预处理
- 特征工程
- 统计分析
- 机器学习
  - 二分类
  - 多分类
  - 聚类
  - 回归
  - 关联推荐
  - 评估
  - 预测
  - 文本分析
  - 网络分析
  - 工具
    - SQL脚本
    - ODPS MR
    - ODPS GRAPH
    - ODPS Spark
  - Parameter Server
    - L1LR new
  - 金融板块(beta)
  - beta组件
  - 废弃栏(15天后会下线)



实验属性

创建日期 2016-06-01 22:07:56

名称  
用户流失数据建模demo

描述  
分别采用逻辑回归和随机森林两个算法建立用户流失模型并评估两个模型好坏

部署状态: 未部署



Platform of  
Artificial Intelligence

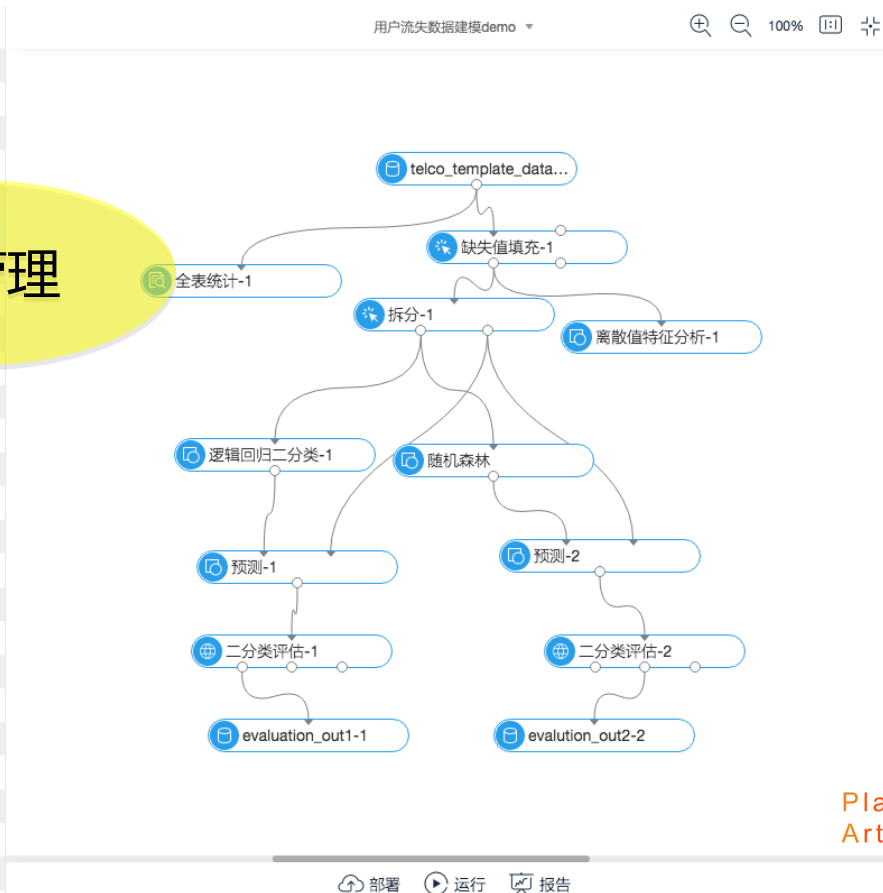


扫码观看大会视频

allpaytech\_dev

- 常用组件
- 源 / 目标
- 数据预处理
- 特征工程
- 统计分析
- 机器学习**
  - 二分类
  - 多分类
  - 聚类
  - 回归
  - 关联推荐
  - 评估
  - 预测
  - 文本分析
  - 网络分析
  - 工具
    - SQL脚本
    - ODPS MR
    - ODPS GRAPH
    - ODPS Spark
  - Parameter Server
    - L1LR new
  - 金融板块(beta)
  - beta组件
  - 废弃栏(15天后会下线)

实验管理



实验属性

创建日期 2016-06-01 22:07:56

名称  
用户流失数据建模demo

描述  
分别采用逻辑回归和随机森林两个算法建立用户流失模型并评估两个模型好坏

部署状态: 未部署



Platform of  
Artificial Intelligence



扫码观看大会视频



数据管理，  
支持多种数  
据源



实验属性

创建日期 2016-06-01 22:07:56

名称  
用户流失数据建模demo

描述  
分别采用逻辑回归和随机森林两个算法建立用户流失模型并评估两个模型好坏

部署状态: 未部署



Platform of  
Artificial Intelligence



扫码观看大会视频

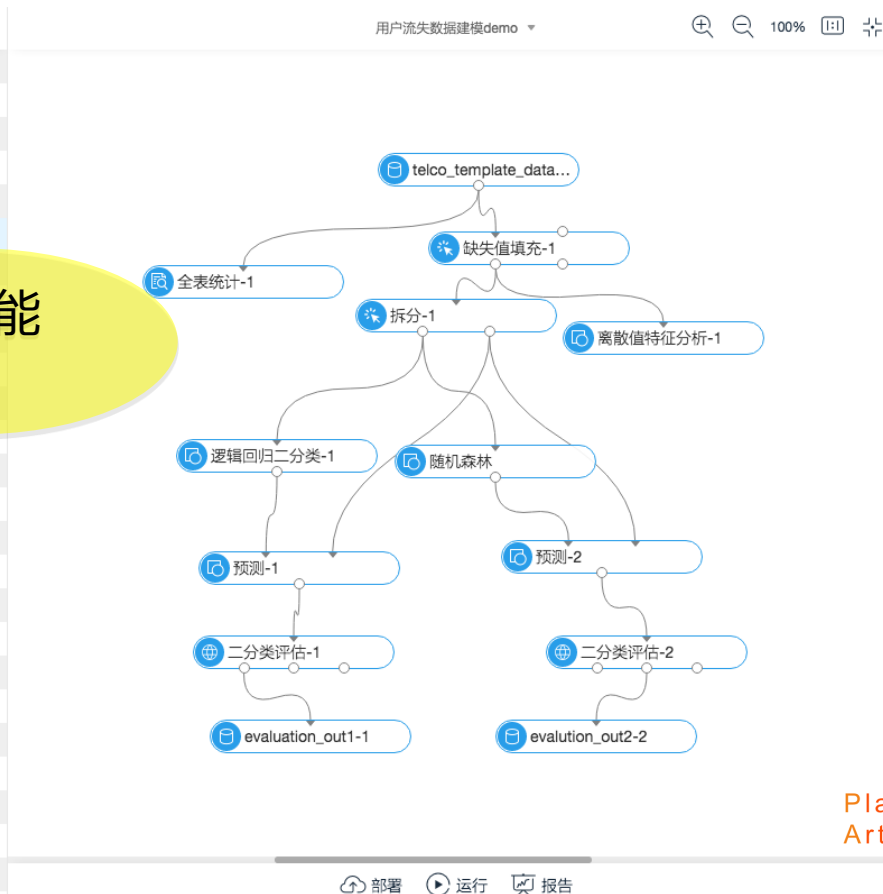


allpaytech\_dev

- 常用组件
- 源 / 目标
- 数据预处理
- 特征工程
- 统计分析
- 机器学习
  - 二分类
  - 多分类
  - 聚类
  - 回归
  - 关联推荐
  - 评估
  - 预测
  - 文本分析
  - 网络分析
  - 工具
    - SQL脚本
    - ODPS MR
    - ODPS GRAPH
    - ODPS Spark
  - Parameter Server
    - L1LR new
  - 金融板块(beta)
  - beta组件
  - 废弃栏(15天后会下线)

组件

算法功能  
组件



实验属性

创建日期 2016-06-01 22:07:56

名称  
用户流失数据建模demo

描述  
分别采用逻辑回归和随机森林两个算法建立用户流失模型并评估两个模型好坏

部署状态: 未部署



Platform of  
Artificial Intelligence



扫码观看大会视频

allpaytech\_dev

- 常用组件
- 源 / 目标
- 数据预处理
- 特征工程
- 统计分析
- 机器学习
  - 二分类
  - 多分类
  - 聚类
  - 回归
  - 关联推荐
  - 评估
  - 预测
  - 文本分析
  - 网络分析
  - 工具
    - SQL脚本
    - ODPS MR
    - ODPS GRAPH
    - ODPS Spark
    - Parameter Server
    - L1LR new
    - 金融板块(beta)
    - beta组件
    - 废弃栏(15天后会下线)

模型管理



实验属性

创建日期 2016-06-01 22:07:56

名称  
用户流失数据建模demo

描述  
分别采用逻辑回归和随机森林两个算法建立用户流失模型并评估两个模型好坏

部署状态: 未部署



Platform of  
Artificial Intelligence

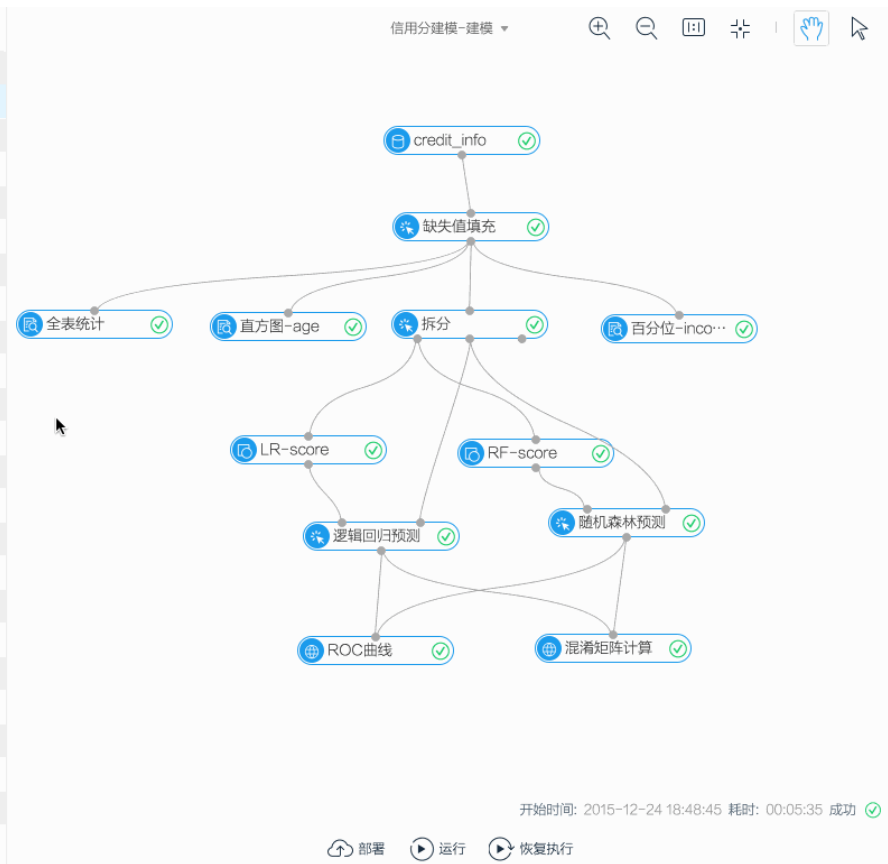


扫码观看大会视频

alipaydw\_dev

- 常用组件
- 源 / 目标
- 数据预处理
- 统计分析
- 机器学习
  - 分类
    - 随机森林
    - 逻辑回归
    - 线性支持向量机
    - GBDT二分类
    - 朴素贝叶斯
    - XGBOOST分类
  - 聚类
    - K均值聚类
  - 回归
  - 评估
  - 预测
  - 图像分析
  - 文本分析
  - 网络分析
  - 工具
  - beta组件
  - 废弃栏

搜索
 实验
 组件
 模型
 设置
 帮助



实验属性

创建日期 2015-12-24 18:24:09

名称  
信用分建模-建模

描述  
请输入描述文本



扫码观看大会视频

# PAI的特点



## 数据智能 触手可及

- 真正的云计算平台
- 支持海量数据计算
- 支持拖拽，无需编程
- 一站式的互联网服务
- 高效的机器学习算法
- 普惠更多用户



# PAI的架构



## 业务应用层

信用领域

安全领域

金融领域

图像识别

推荐系统

搜索引擎

金融云项目

公共云项目

## 平台化产品

项目管理

算法分享管理

可视化分析

部委定制 PAI

蚂蚁 MYPAI

## 模型与算法

数据预处理

特征工程

机器学习模型

数理统计

深度学习模型 - CNN / DNN / RNN / LSTM

## 计算框架

MR

SQL

MPI

PS

GRAPH

GPU 单机多卡 / 多机多卡

## 基础设施层

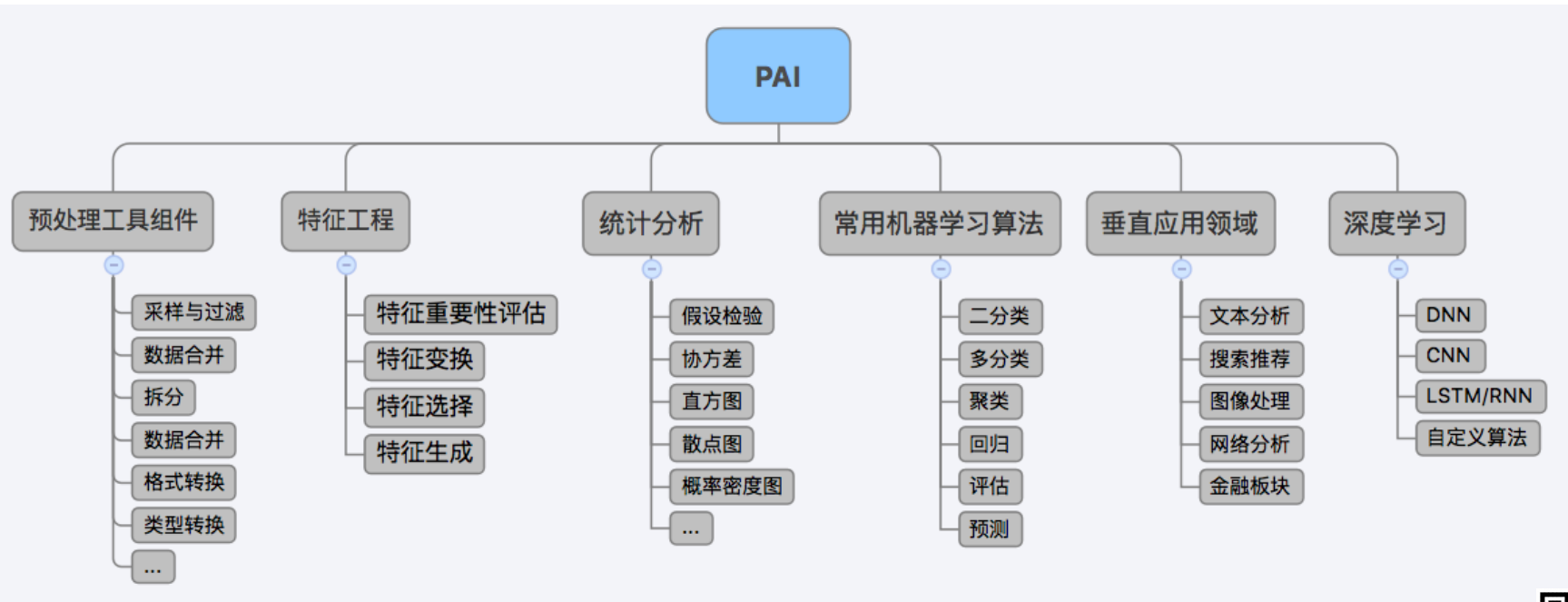
阿里云基础设施

CPU 集群

GPU 集群



# PAI的算法组件



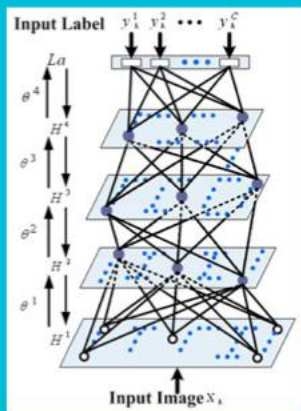
## 二、功能特点



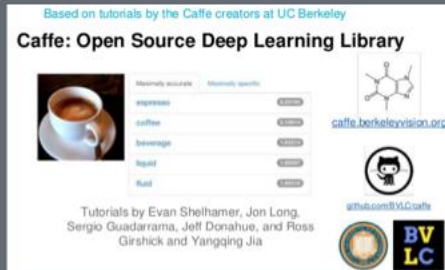
# 深度学习Pluto



## Deep Learning



## Caffe



## Pluto

- 基于开源Caffe
- GPU 多机多卡
- MPI 通信协议
- 分布式深度学习算法
- 支持常用模型 (DNN,CNN,RNN/LSTM)





# 深度学习Pluto



## Pluto 数据并行

Each computing node stores a model replica

## InfiniBand 高速通信

Speed in theory 56Gbps

## GPU 集群调度

Schedule mixed jobs on CPU/GPU cloud

## Caffe 开源社区

Support popular models, including CNN , LSTM

模块化

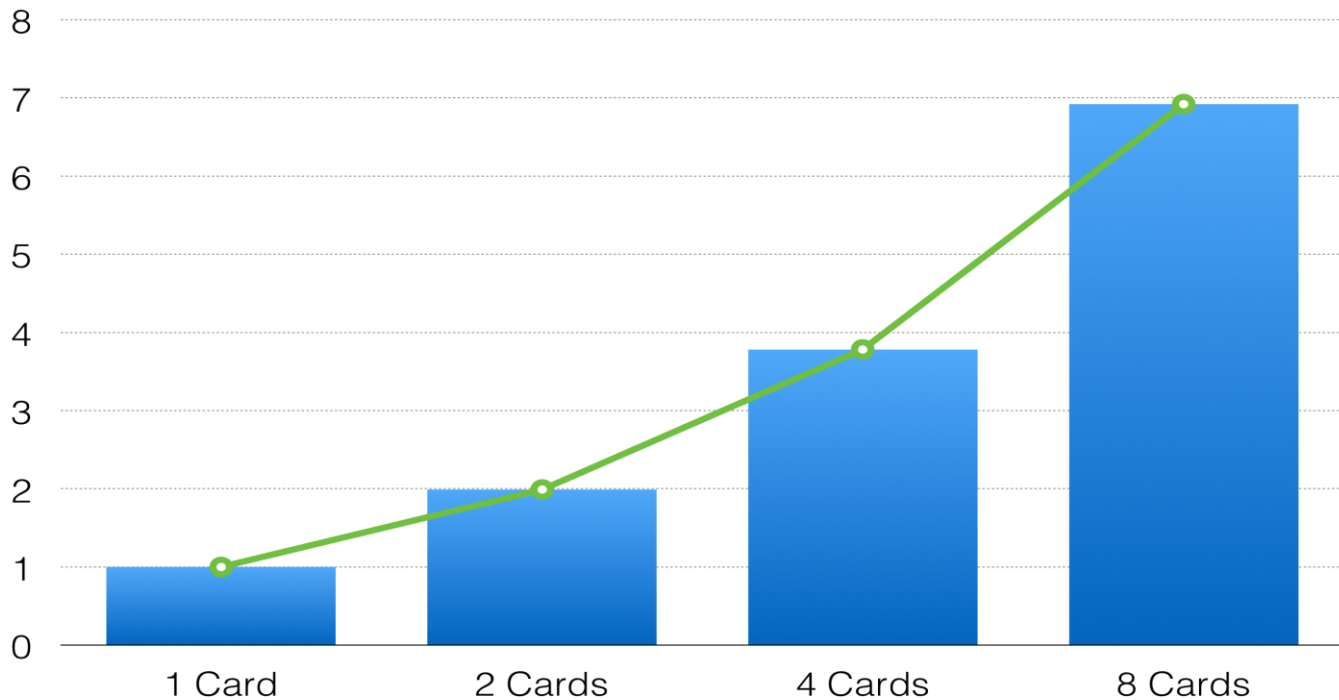
训练  
速度

可扩展性



# 深度学习Pluto

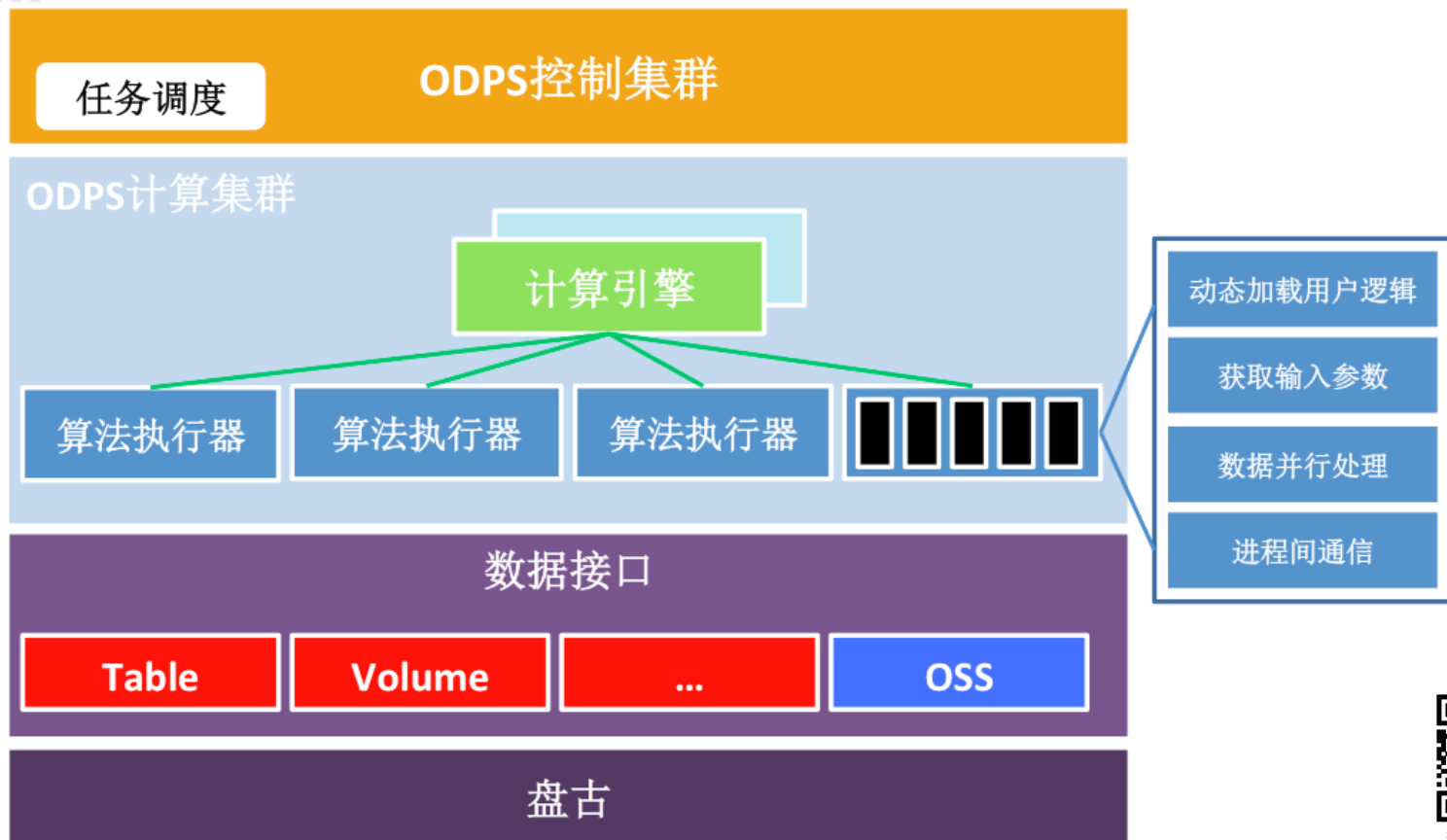
## Acceleration Ratio



AlexNet on ImageNet data with accuracy 57%



扫码观看大会视频



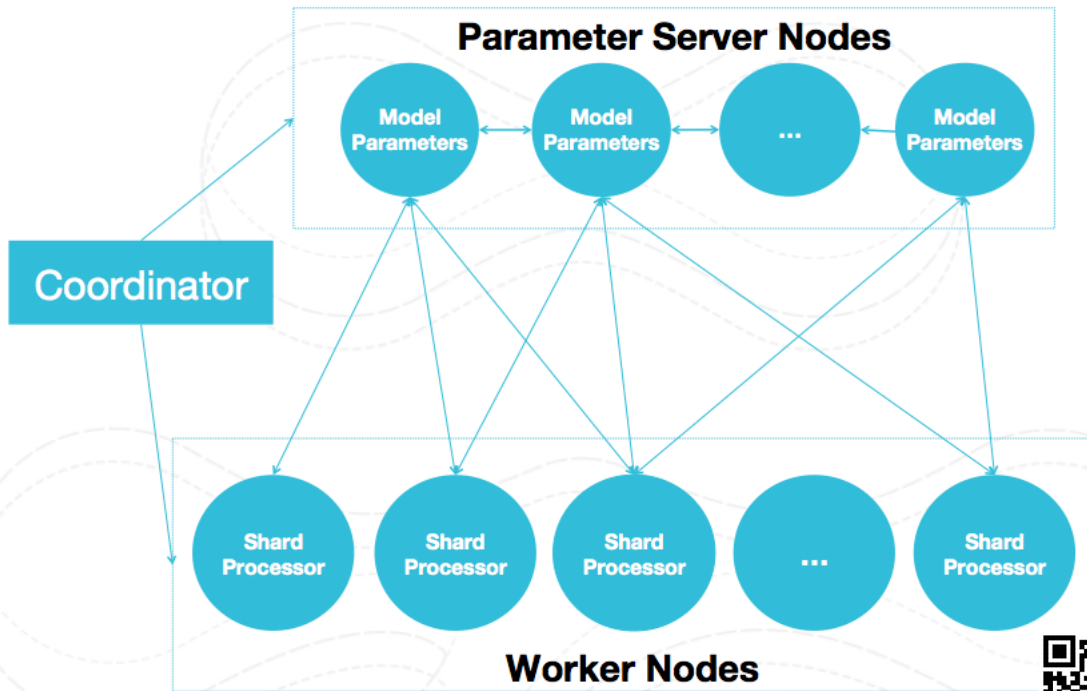
# 参数服务器



1 Complete Failover  
失败续跑

2 Asynchronous Update  
异步更新

3 Scalability  
超大模型



# 参数服务器

## 离线学习

- LR
- FTRL
- LDA
- GBDT
- FM
- WMD  
(开发中)
- Word2Vec
- DBScan
- 矩阵相关
- ...

## 在线学习

- Online Engine
- Online FTRL

## 深度学习

- DNN
- DSSM
- CNN (开发中)





## 三、案例分享



# 搜索推荐场景



- 1.Query: 连衣裙
- 2.User: 用户信息，个性化千人千面
- 3.Item: 可选购的商品
- 4.目标: 用户是否浏览或购买

Query x User x Item ---》(0 或 1)

线性逻辑回归模型



扫码观看大会视频

# 参数服务器



1

## 阿里妈妈广告数据

MPI实现只能支持最多千万特征，计算节点不能超过200

2

## 参数服务器

10亿特征，570亿样本，逻辑回归5小时完成

3

## 规模更大，速度更快，性能更稳定

支持100亿特征和1000亿样本，MPI实现，加速40%；失败续跑

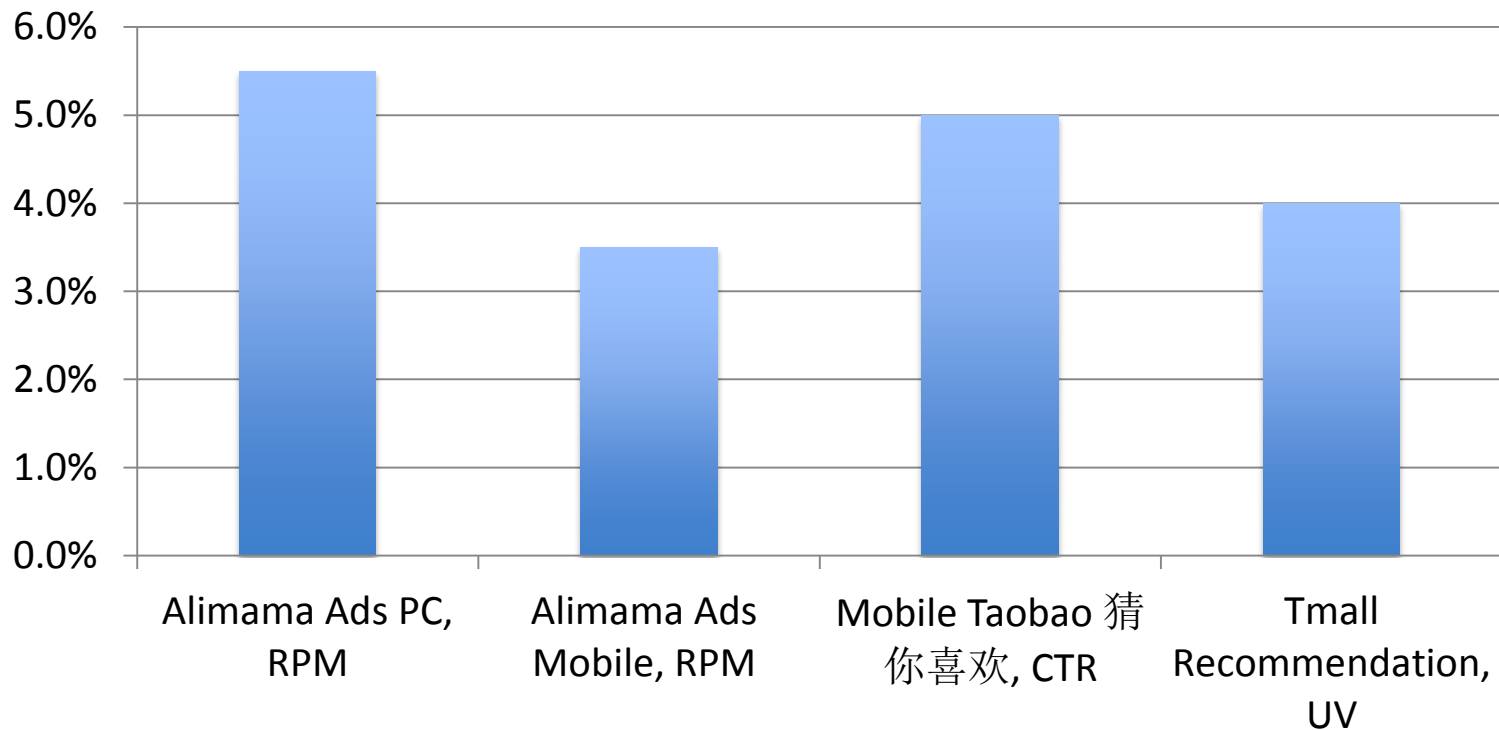




# 参数服务器



## Metric Improvement



# DSSM语义相似度



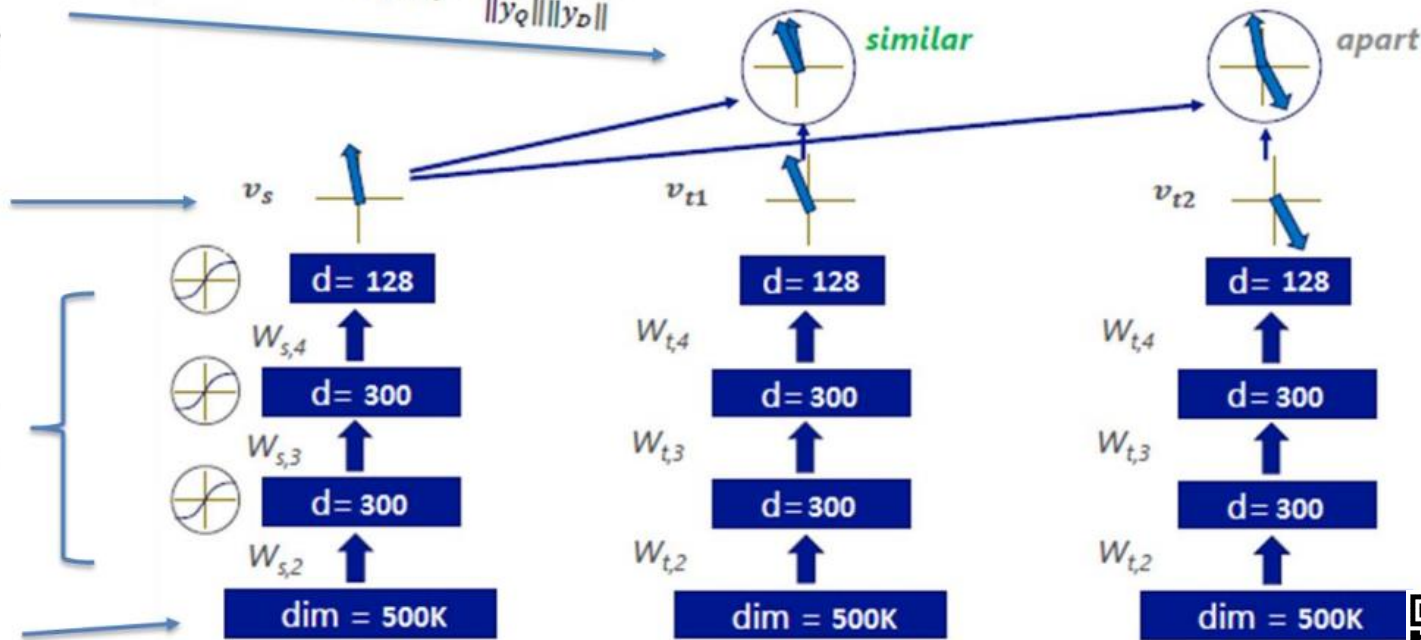
Relevance by  
Cosine Similarity

$$R(Q, D) = \text{cosine}(y_Q, y_D) = \frac{y_Q^T y_D}{\|y_Q\| \|y_D\|}$$

Semantic  
Features

Multi-layer Non-  
linear Projection

Term Vector  
(Bag-of-words)



Q: “杭州 哪里 好玩”

$D^+$ : “杭州 景点”

$D^-$ : “杭州 美食”



扫码观看大会视频

# DSSM语义相似度



1688.com

淘宝网  
Taobao.com



# 公共云CRM案例



## 明源科技

云采购平台

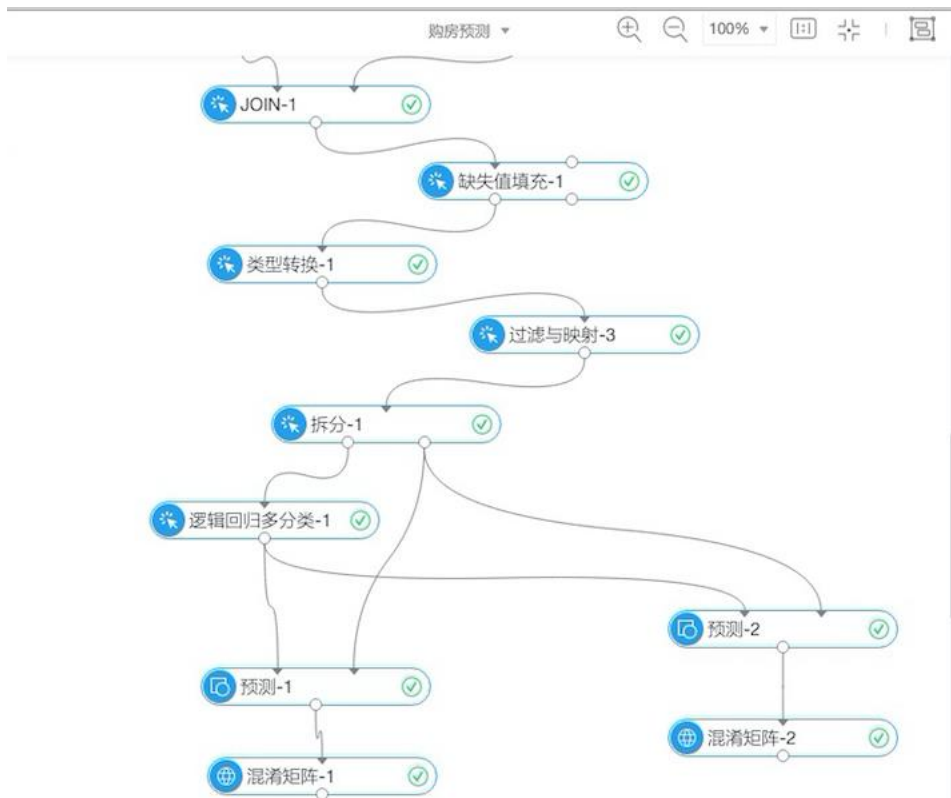
3000+的潜在建筑器材客户

4名销售

每名销售每天处理10个客户



# 公共云CRM案例



## 解决方案

1. 数据同步
2. 机器学习，包括特征工程，逻辑回归，模型预测，效果评估
3. 调度预测
4. 结果推送



- 图形界面，支持拖拽
- GPU集群调度
- 分布式深度学习算法
- 丰富的可视化分析
- 广泛的应用场景
- 赋能用户



SCAN BARCODE  
START YOUR TRIAL

数据智能 触手可及



Platform of  
Artificial Intelligence



扫码观看大会视频

2016 The  
Computing  
Conference  
**THANKS**

