



# 云上大数据 一生于战火的云端计算平台：ODPS与阿里PAI

阿里巴巴ODPS, iDST



简介



ODPS计算引擎介绍



阿里PAI算法平台

## DT生态圈

完整的数据链路，  
丰富的交换场景。

数据平台

赋能：高效的生产  
数据价值

计算平台

强大的公有计算  
底层架构

阿里云

## 计算平台

### 平台化的服务

数据交换平台

算法平台

可视化分析

数据、预测服务等

...

### 模型与算法

数据挖掘

大规模机器学习

深度学习

统计、BI

...

### 计算引擎

SQL

MapReduc  
e

参数服务器

MPI

R

图计算

...

## 计算平台

平台化的服务

数据交换平台

阿里PAI算法平台

算法平台

可视化分析

数据、预测服务等

...

模型与算法

数据挖掘

大规模机器学习

深度学习

统计、BI

...

计算引擎

ODPS

SQL

MapReduce

参数服务器

MPI

R

图计算

...

 海量数据处理和分享需求

– EB级数据

 生于战火：没有数据，就没有计算能力

– 一路走来，在搜索、广告、电商、金融等高复杂场景下得到尽情锤炼

 相信生态、促进联创、成就大家

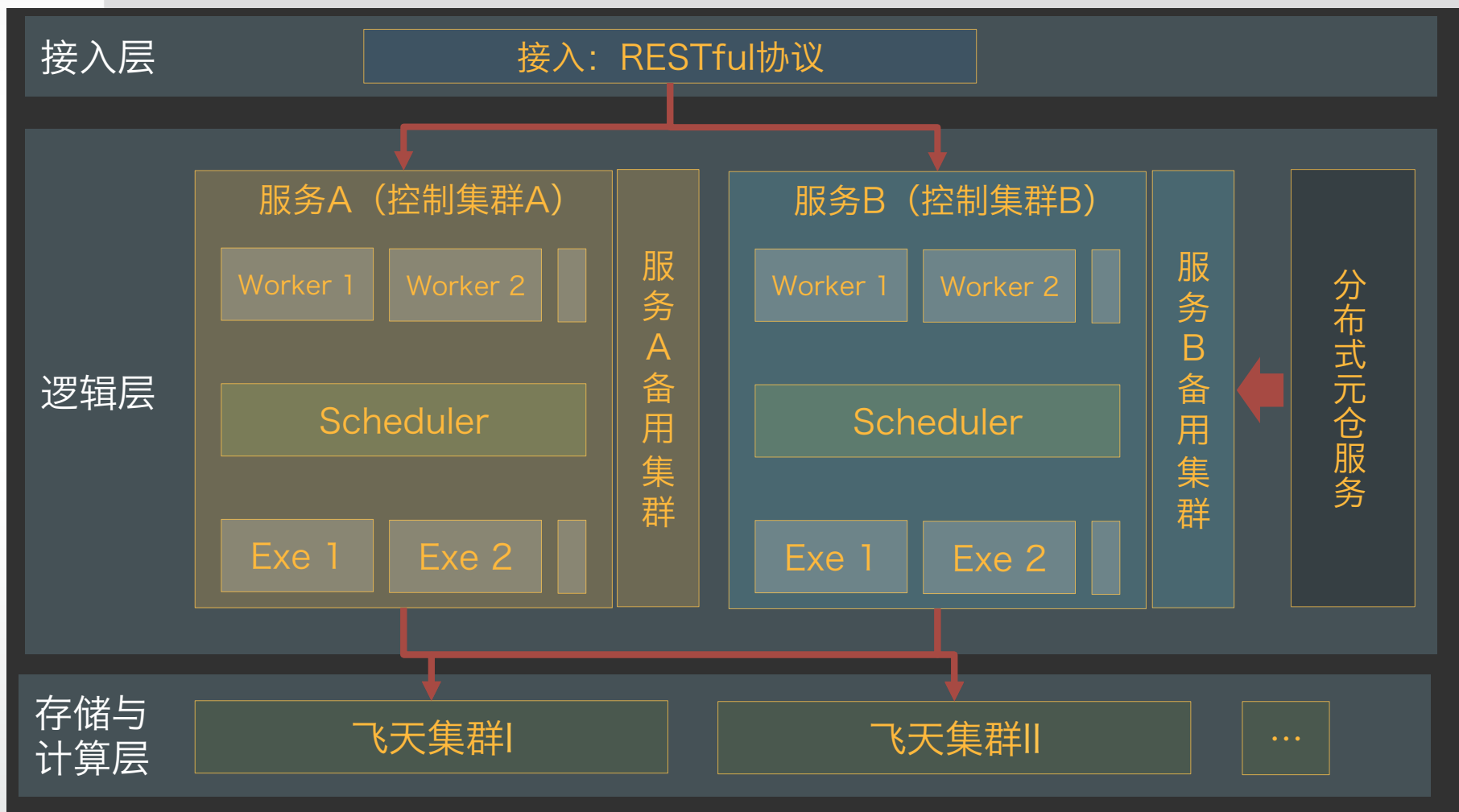


ODPS概述



ODPS计算框架

# ODPS系统架构











## 服务器

- 单一集群规模可以达到15000+
- 单个ODPS部署可以支持100万以上
- 支持同城、异地多数据中心模式



## 用户

- 10000+用户数
- 1000+项目应用
- 100+部门（多租户）
- 100万以上作业（目前单日平均提交任务）
- 20000以上并发作业

-  ODPS支持完善的多租户机制，通过存储的计算配额的方法可以让多个用户分享一个集群的资源
-  所有的计算任务都运行在安全沙箱中，通过进程和系统沙箱，配合运行时的签权方法，保障多用户共享集群资源时的数据安全
-  ODPS提供丰富的授权管理手段，包括ACL,角色， Policy以及 Label机制，可以提供精确到列级别的安全方案，满足一个组织或者跨组织间的授权需求
-  安全要求较高的项目，可以提供项目保护机制，防止数据流出。




支持以下调度方式:

- Fair Scheduler
- FIFO Scheduler
- 抢占
- 组内优先
- Mix/Max Quota







多租户资源控制

- CPU 通过抢占方式, 基于Lxc实现
- Memory
- Network QoS (在线、离线、混合业务场景控制)
- 信号、系统、用户
- Disk: 飞天 ChunkerServer 统一代理

 支持对存储和计算的压缩，只有当计算过程中要用到该数据时才解压缩，支持的压缩格式：

- 代管：Gzip、Snappy、LZO、LZ4
- RAID 通过纠错编解码来实现文件存储的可靠性

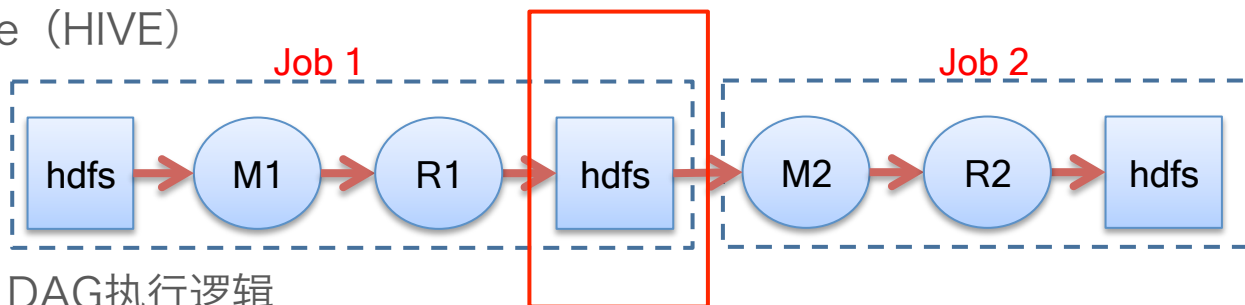
 采用（行）列存储模式

-  跨集群（机房）数据共享
-  数据生命周期功能可以极大的减轻对过期数据的管理成本，减少无效的存储空间占用。
-  数据质量检查（DQC）功能可以防止脏数据对生产任务造成的影响，降低运维的负担。
-  表的archive功能可以将冷数据的存储成本降低50%以上，使用上对应用完全透明。

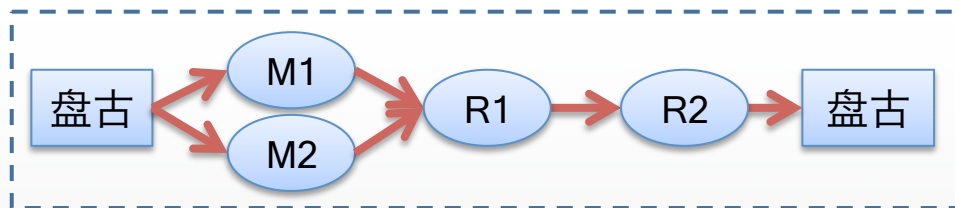
## 丰富的计算框架

例子: SELECT ... FROM a JOIN b ON a.id=b.id GROUP BY a.c;

MapReduce (HIVE)





ODPS SQL DAG执行逻辑



准实时SQL: 支持Service Mode的常驻进程, 利用内存和网络大幅提高效率。

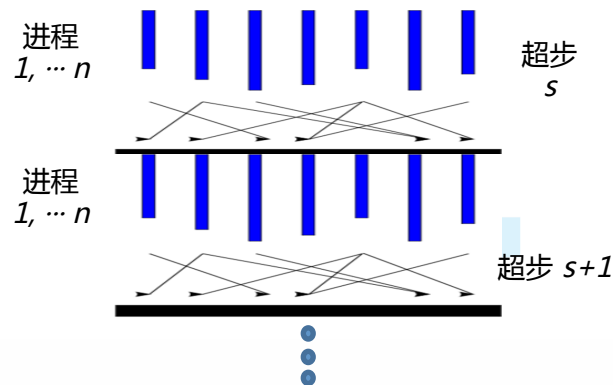
创新性的Stream SQL, 利用SQL高效的处理流数据

ADS: 通过index机制实现强大的实时SQL数据处理

-  场景：蚂蚁金服花呗、余额宝等产品
-  规模：上万计算节点，PB级数据处理
  - 离线SQL效果优于HIVE
  - 准实时SQL效果优于TEZ
  - 实时SQL系统ADS支持千亿级别数据毫秒级的响应

 面向迭代的分布式图计算框架支持JAVA编程接口，类似Pregel

- 磁盘IO -> 内存网络，换来更快的性能
- 面向图数据而设计，适合图算法开发.



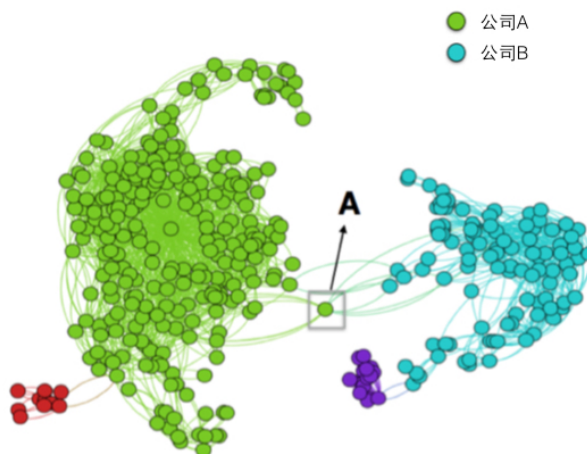
 计算规模:




- 最大顶点 (vertices) 数: 百亿, 最大边数 (edges) : 1500亿
- 最大迭代次数 (supersteps) : 120万, 最大发送消息量 (sent messages) : 6 千亿





场景：社交网络分析LPA

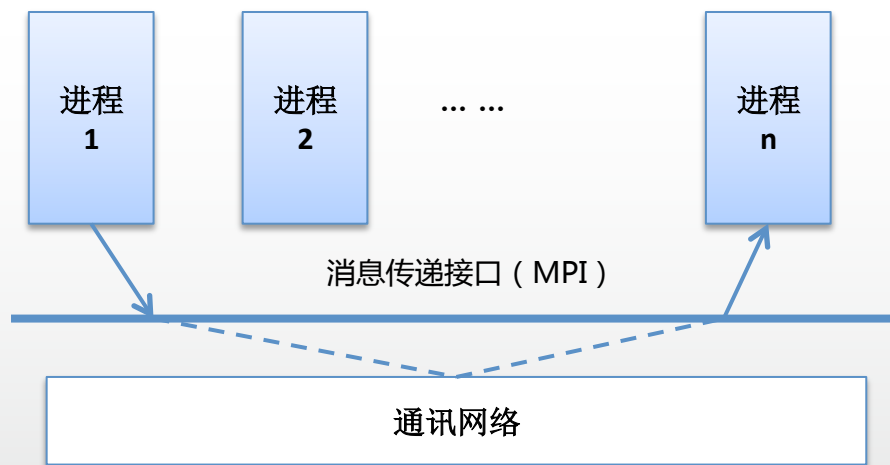




-  在物理集群上启动虚拟机运行R（单机）instance
-  打通ODPS数据源与R的联通，用户可以在安全可控的环境利用R的丰富资源
-  支持分布式自动调优：
  - 通过对同一问题并行运行多个R脚本（每个脚本代表对这个问题的不同解法）来达到自动选择模型和优化的效果。
  - 例如一个分类问题：同时起4个不同的分类器（LR、RF、SVM、GBRT），在每个算法选50套参数，同时训练200个模型，然后再同一份测试数据上很像比较最优解。

# Message Passing Interface

- 内存计算，适合大规模同步多迭代的算法实例
- 积累了成熟的算法库与开发流程
- 支持单机调试测试环境
- 支持Failover

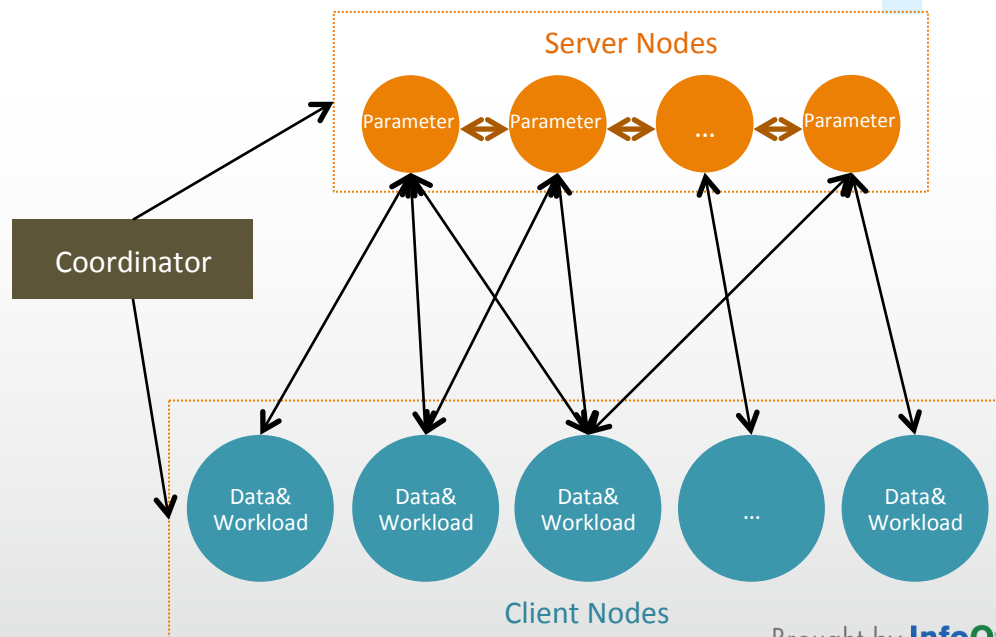
## MPI (*Message Passing Interface*)






-  场景：数据规模：80TB、集群规模：5k级别、逻辑回归训练
-  用时：~ 4 hours

# Parameter Server

- 🔗 模型、数据分片，支持超大模型
- 🔗 利用稀疏特性减小通信
- 🔗 支持异步迭代
- 🔗 各个角色有完善的Failover机制



-  场景：百亿级别的特征，千亿条的数据
-  计算时间：~8小时
-  在稀疏数据场景中的性能在相同硬件环境优于MPI

 提供核心、实际场景中锻炼过的算法库

- 特征工程
- 大规模机器学习与深度学习
- 在线学习过程

 开放的、易用的云产品：阿里PAI

## 计算平台

平台化的服务

阿里PAI算法平台

算法平台

可视化分析

模型与算法

大规模机器学习

深度学习

计算引擎

ODPS

SQL

MapReduce

参数服务器

MPI

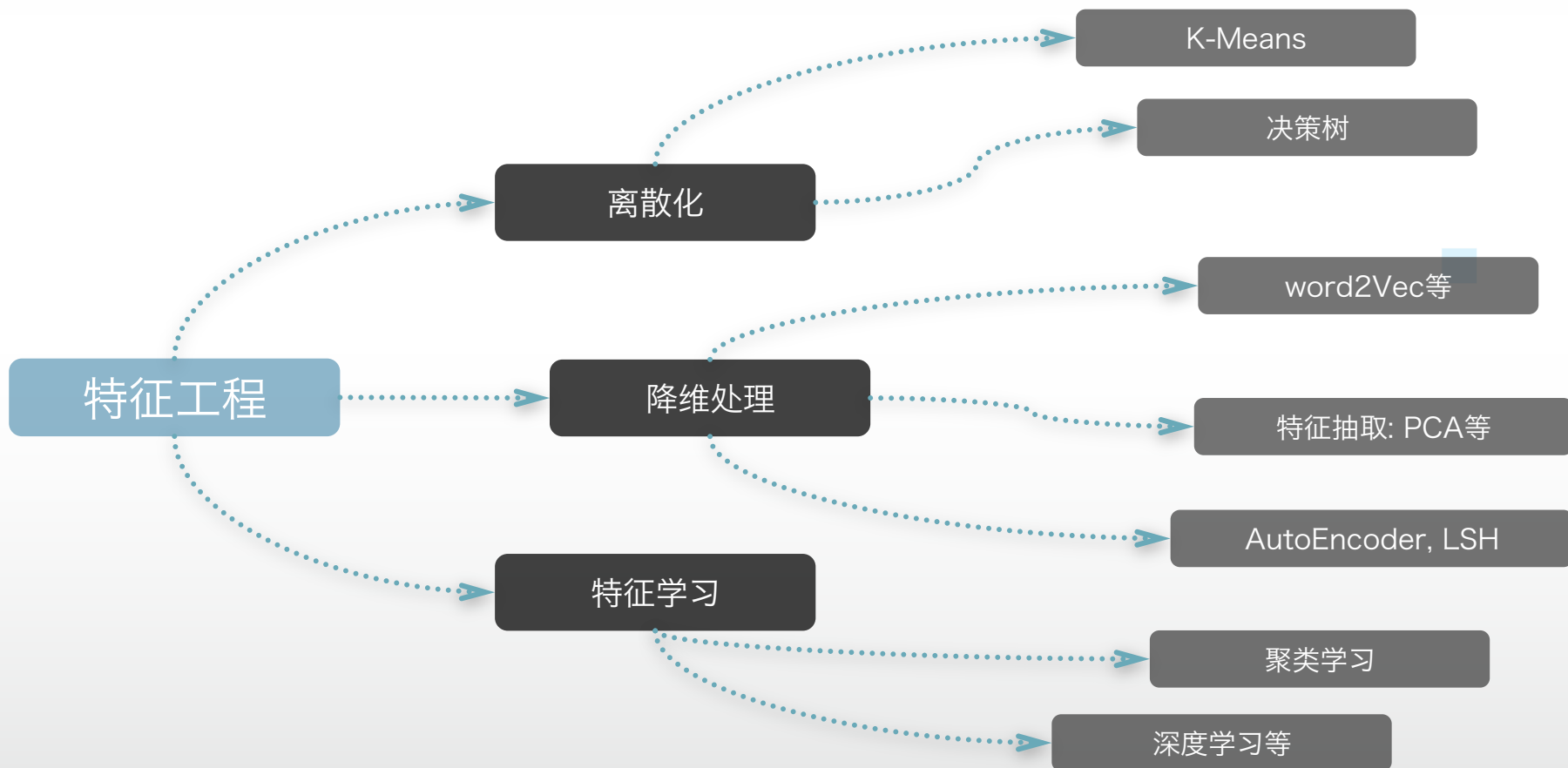
R

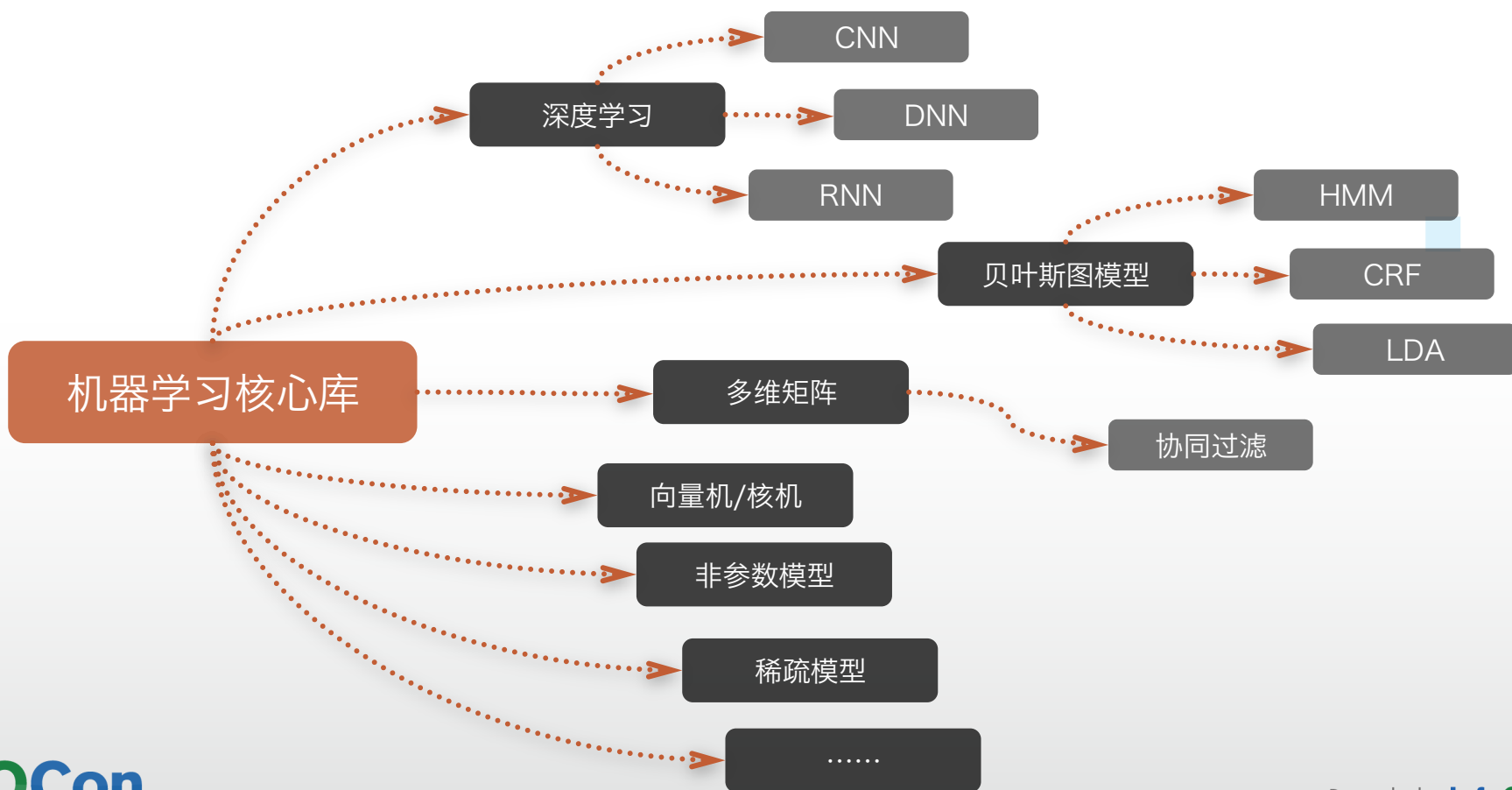
图计算

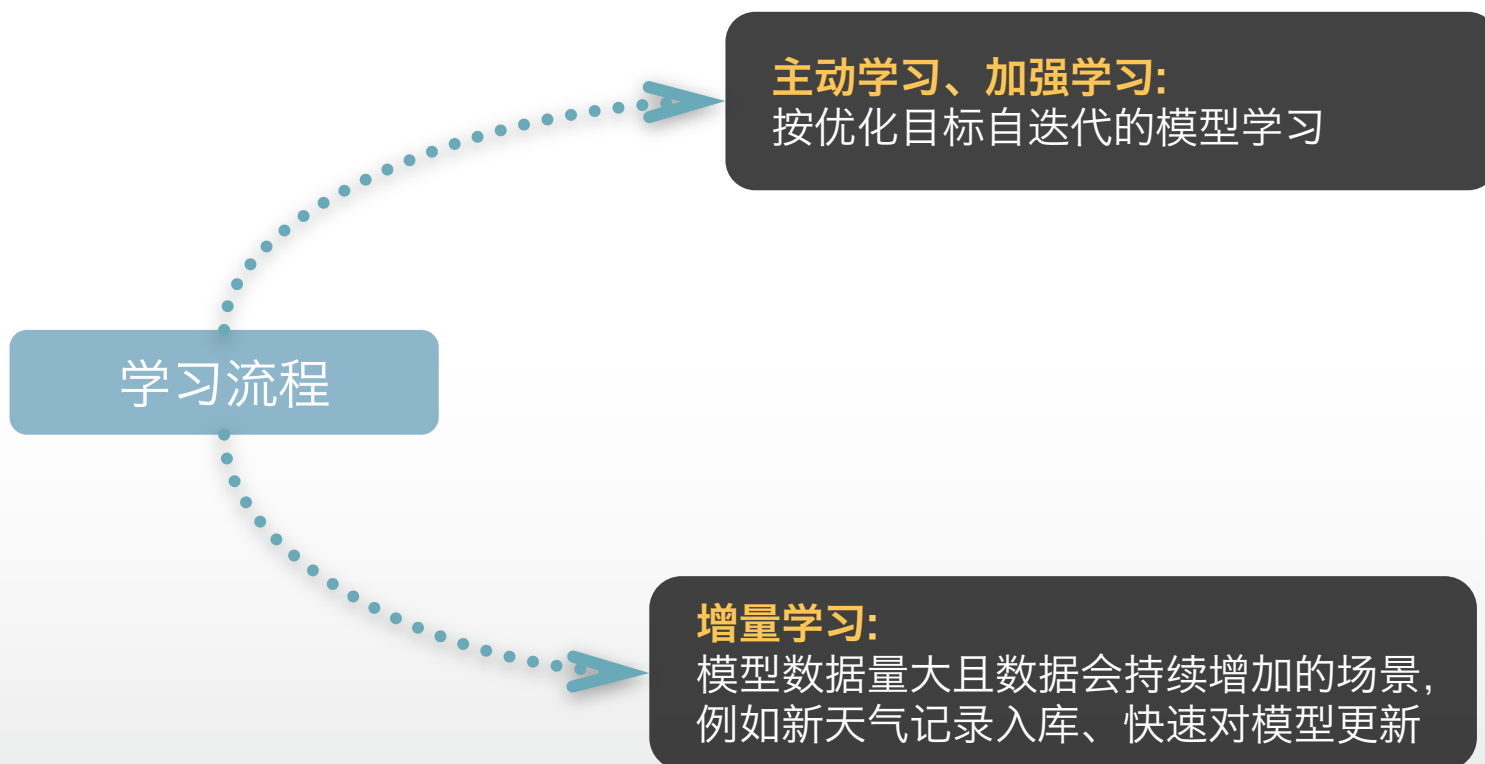


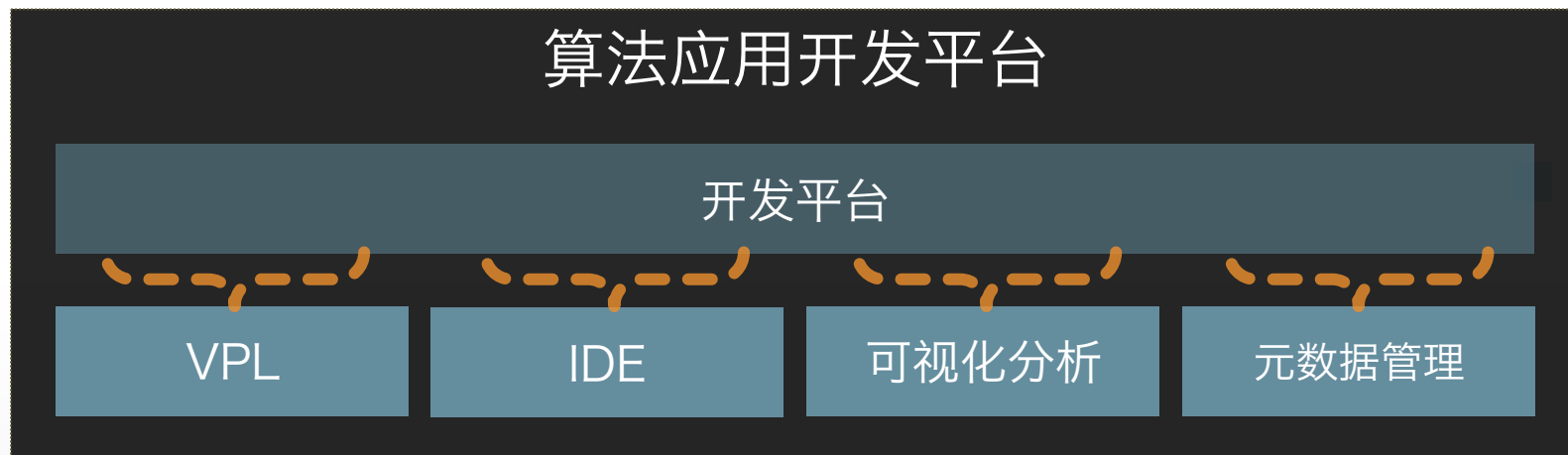
# 典型模型算法开发流程

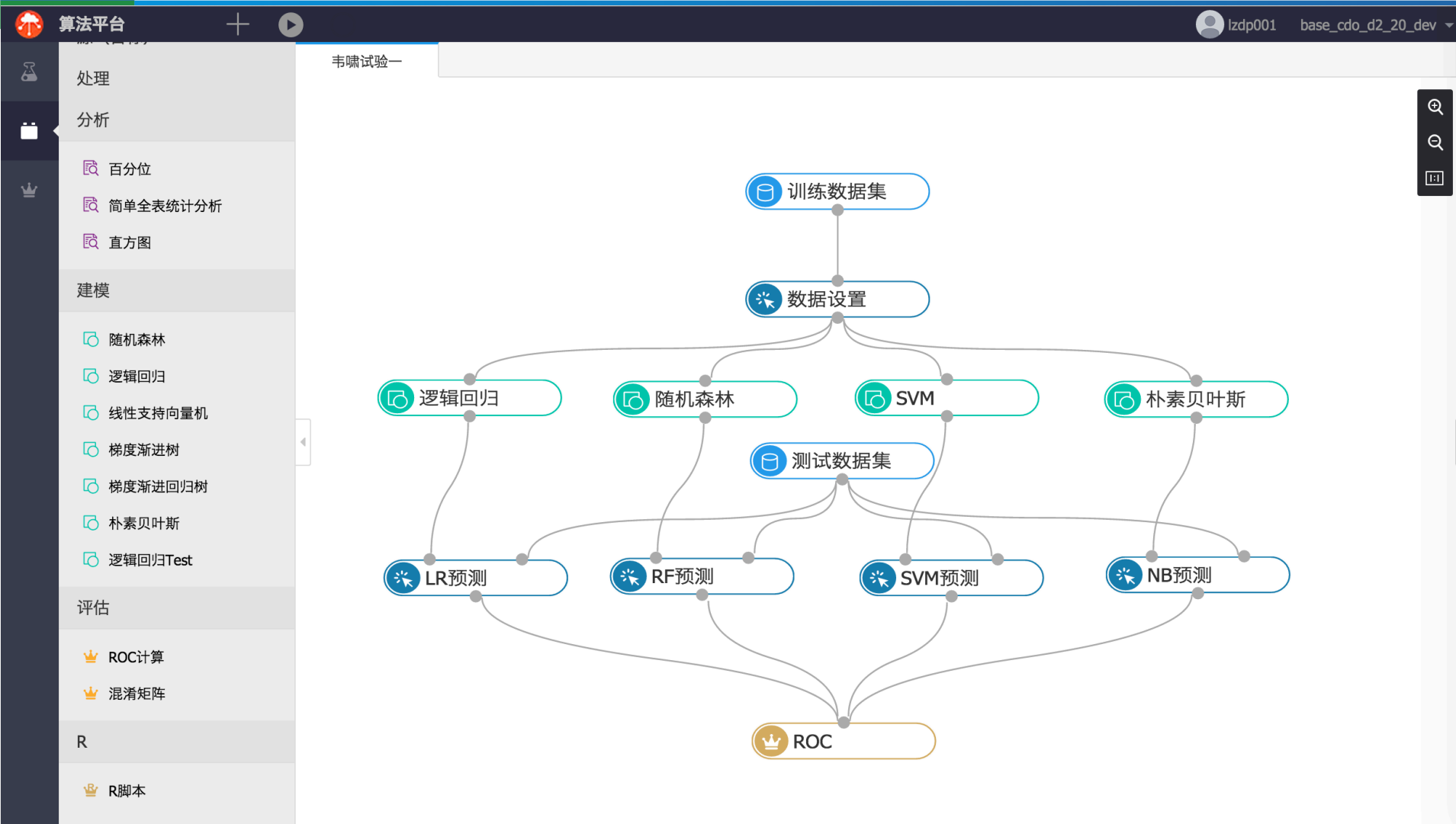




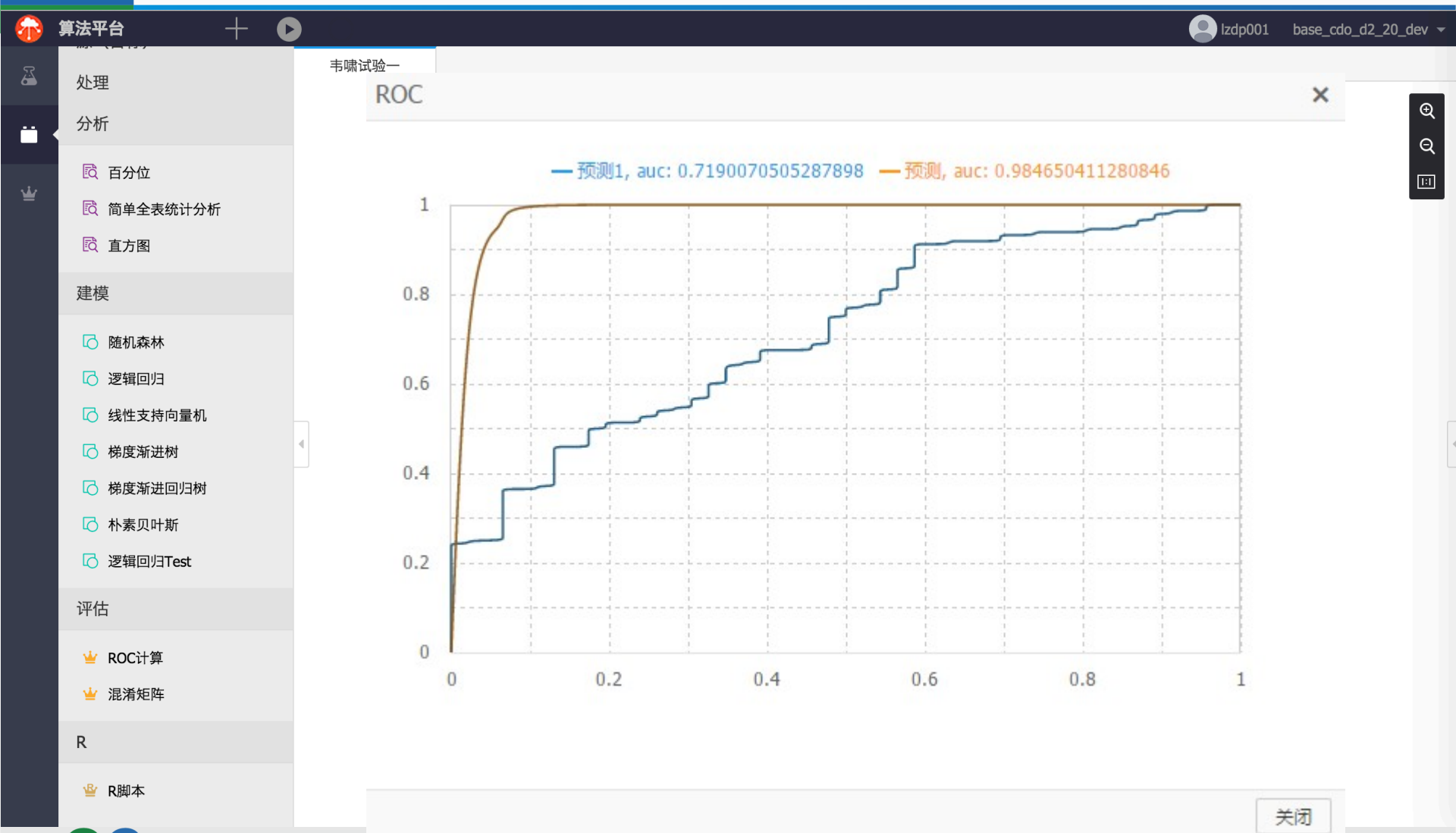








# 算法平台Demo





# 谢谢！