博学谷
www.boxuegu.com
黑马程序员线上品牌

阿里PAI平台介绍



一样的教育，不一样的品质

博学谷
www.boxuegu.com
黑马程序员线上品牌

目录

Contents

1. 背景介绍
2. 阿里云人工智能平台PAI
3. PAI平台的意义
4. PAI的产品结构
5. PAI的架构

博学谷
www.boxuegu.com
黑马程序员线上品牌

01

背景介绍

AI generated content

AIGC的领域划分

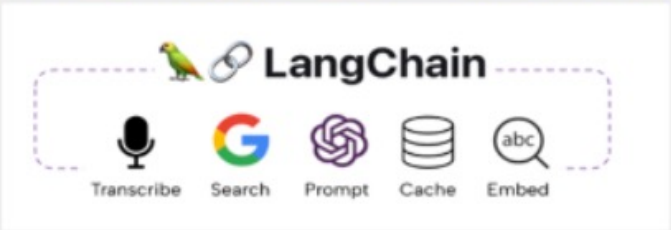
Wall-E 瓦力机器人
达利

代表产品

OpenAI	ChatGPT, GPT-4, DALL·E
ANTHROP\C	Claude 2
Meta	LlaMA 2
Google	Gimini Imagen Bard, PaLM 2
Midjourney	Midjourney
stability.ai	Stable Diffusion

开源生态

以LangChain为代表的开源框架，降低了开发人员基于大语言模型搭建应用程序的门槛。



在开源社区涌现了许多有趣的AIGC项目

- langchain-ChatGLM：基于ChatGLM的本地知识问答
- ChatMind： 使用ChatGPT生成思维导图
- ChatPaper：根据输入关键词，自动在arxiv上下载最新的论文，并对论文进行摘要总结。
- ……

下游应用

模型&服务

Jasper
帮助企业和用户生成智能营销文案，从诞生到成为独角兽仅用了18月。

LATITUDE
基于GPT构建了AI Dungeon 2 虚拟世界文字冒险游戏。

图像

Fashion Diffusion大模型颠覆传统服装设计行业，智能生成符合用户需求的风格款式图片。

视频

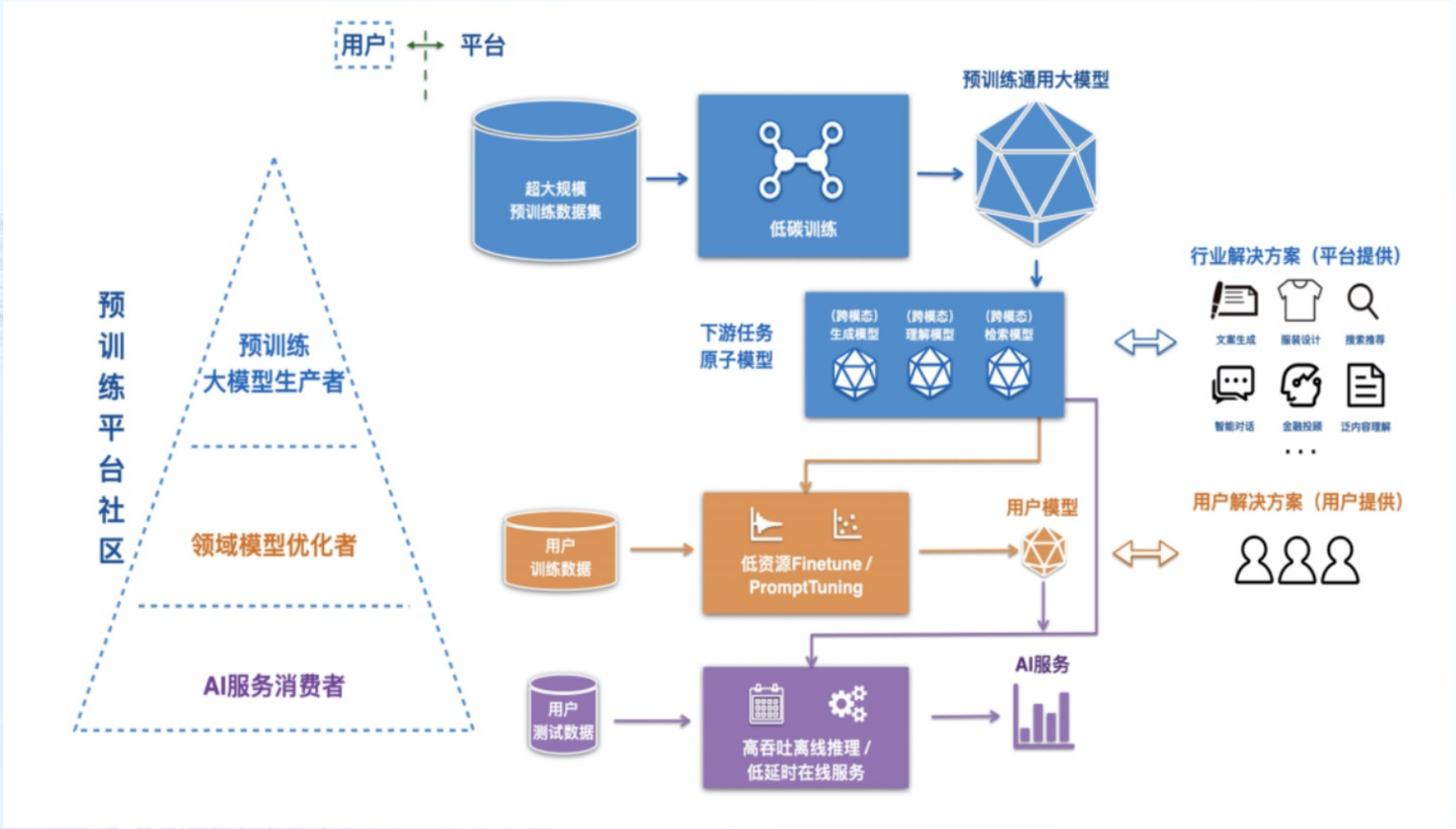
IBM
人工智能系统Watson为科幻电影《摩根》制作预告片。

runway
推出AI生成视频软件，赋能用户制作专业级视频特效。

音频

Mubert
基于AI算法，自动生成适合指定场景的音乐，如电子、古典、舞曲等等。

AI研发的新范式



三类用户：

上游：通用模型生产者与平台方（如阿里）

中游：利用垂直行业知识对模型进行优化（生态伙伴）

下游：最终AI应用的使用者（最广大的用户群体）。

博学谷
www.boxuegu.com
黑马程序员线上品牌

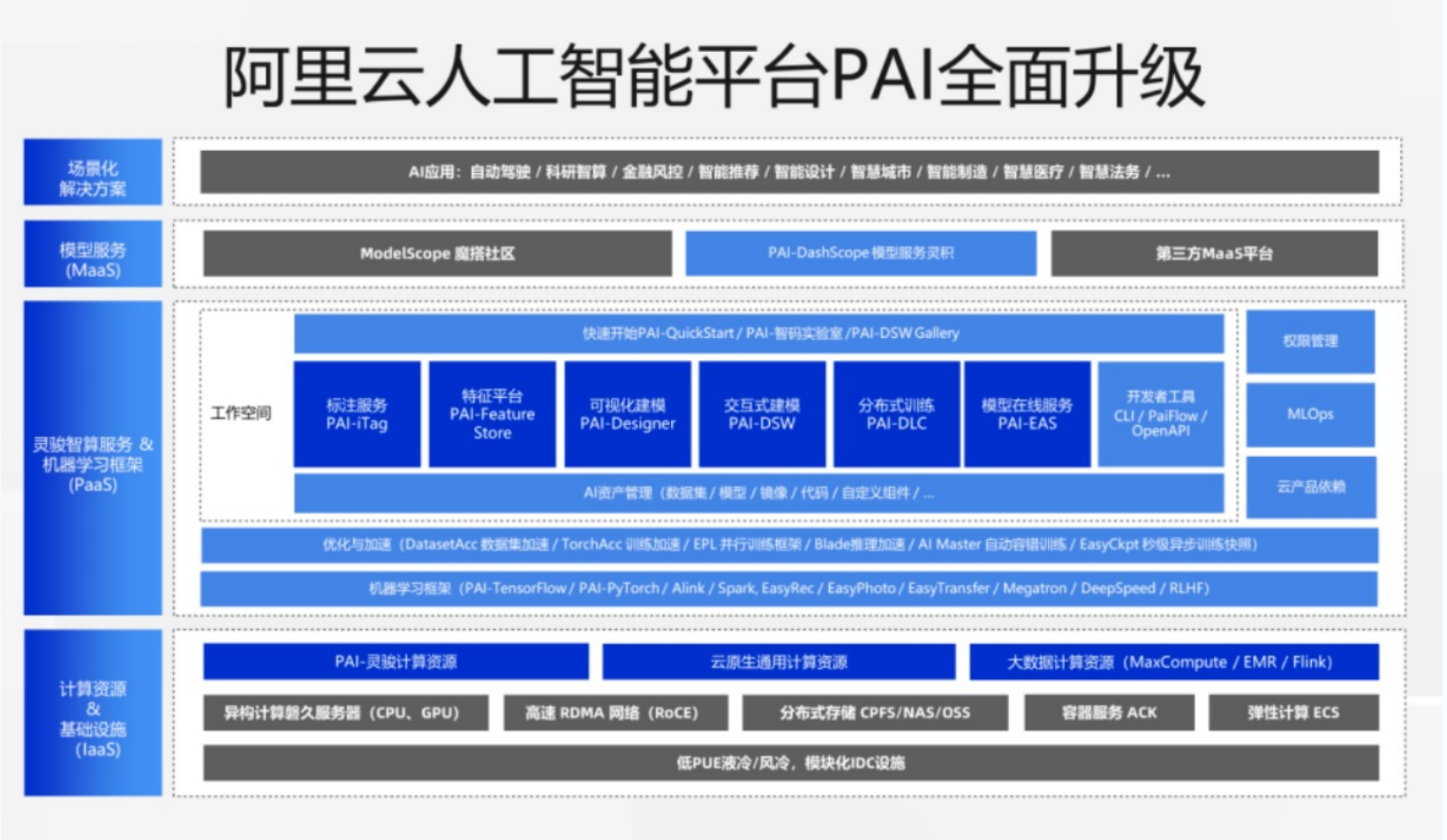
02

阿里云人工智能平台PAI

PAI4.0 架构

三层架构：

- 下层是强大的基础设施
- 中间层“PAI灵骏智算集群”，是专门面向超大规模分布式，重点解决预训练、Finetune、推理等工作。
- 最上层“Model as a Service”概念，能够让不懂算法，但需要应用AI的人进行全链路的AI创新。



博学谷
www.boxuegu.com
黑马程序员线上品牌

03

PAI平台的作用

提升开发效率：全生命周期化的AI平台

PAI - iTAG 智能标注

PAI 团队和达摩院深度合作共建，打造智能标注产品，提供全场景、高质量、智能化的标注平台服务和人力标注服务。



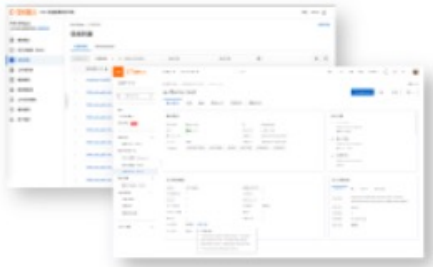
PAI - DSW 云原生交互式建模

交互式建模，支持直接编码 coding 来建模，服务算法专家。



PAI - DLC 云原生AI训练服务

DLC 给客户提供一个做训练的环境，支持分布式训练、单机的训练。



PAI - EAS 云原生AI在线预测服务

一键把模型变成一个在线服务，提供云上非常好的稳定性、弹性、规模以及性价比。



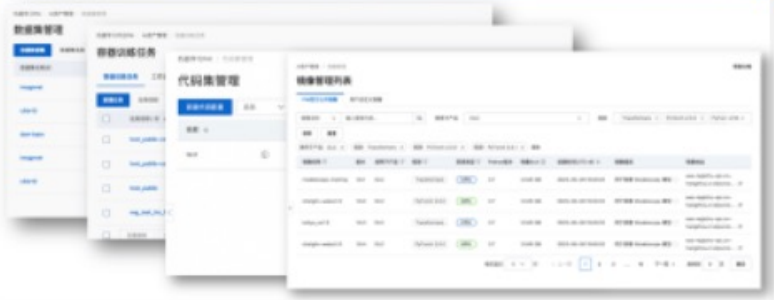
PAI – AI 工作空间

算力配额管理、集群监控、租户管理



PAI – AI资产

数据集管理、镜像管理、代码管理、模型管理、任务管理



PAI – OpenAPI

全链路OpenAPI

开发类							
iTAG	37个	DSW	26个	DLC	18个	EAS	52个

任务编排类	
Pipeline PAI-Flow	18个

资源管理类	
AI 工作空间	53个

PAI-DSW交互式建模

内置多种开发环境

- Notebook、WebIDE、Terminal，满足个性化开发需求

代码Copilot

- 内置基于LLM的智能代码助手

共享数据集挂载

- 支持OSS、NAS、CPFS等文件系统挂载和数据共享

实例精细化管理

- 支持定时关机、闲置关机等生命周期管理配置
- CPU/GPU/内存水位实时监控

自定义镜像开发及制作

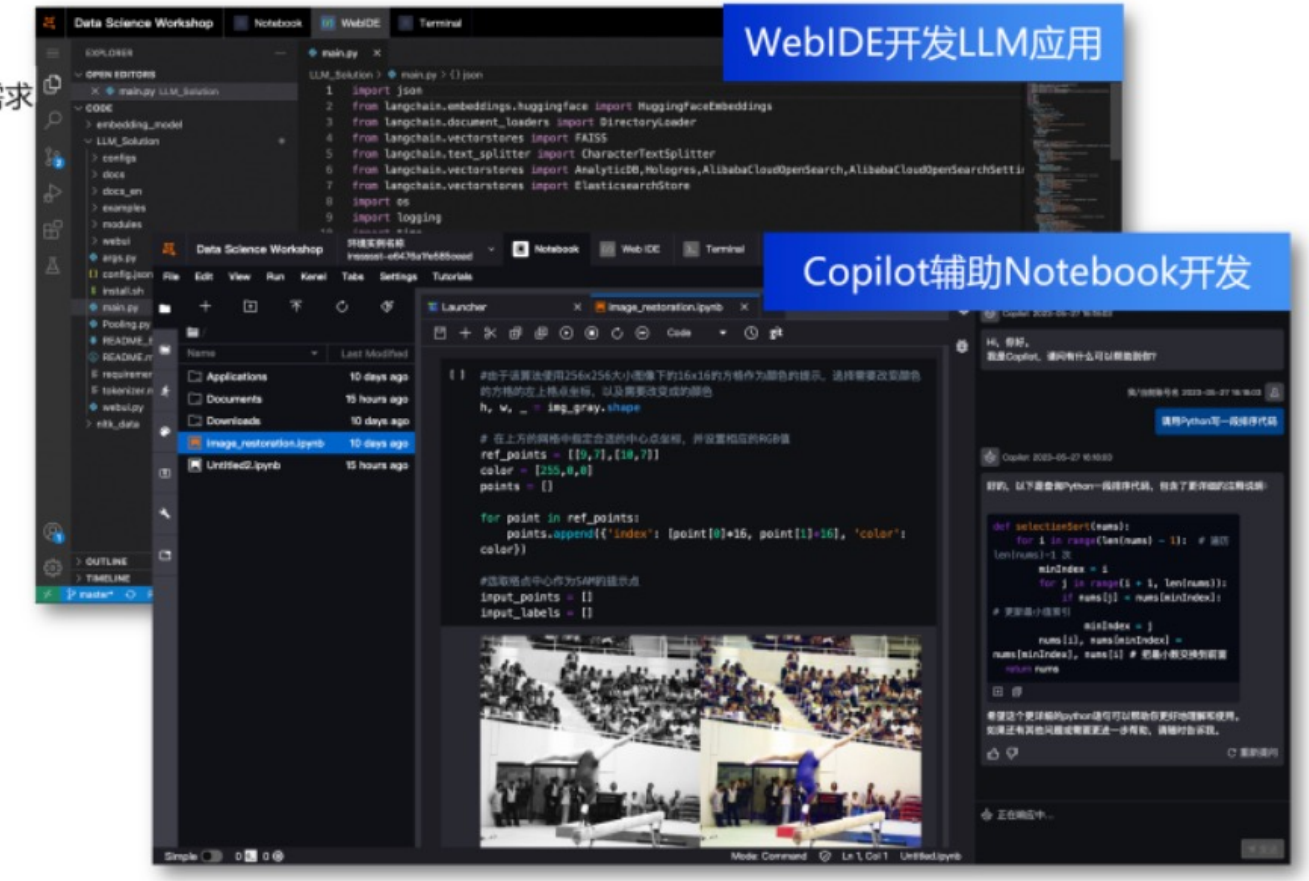
- 支持制作自定义镜像并推送至容器管理

DLC任务提交/EAS服务部署

- 提供DLC/EAS插件，支持用户在实例中快速训练及部署

DLC任务提交/EAS服务部署

- 支持远程连接DSW实例



PAI-DLC分布式训练

多资源组支持

- 支持弹性资源/用户专属资源
- 灵骏高性能资源/通用计算资源

灵活的环境/任务类型支持

- 丰富的官方镜像/自定义镜像
- 任务类型覆盖传统机器学习/深度学习等标准框架

监控及告警

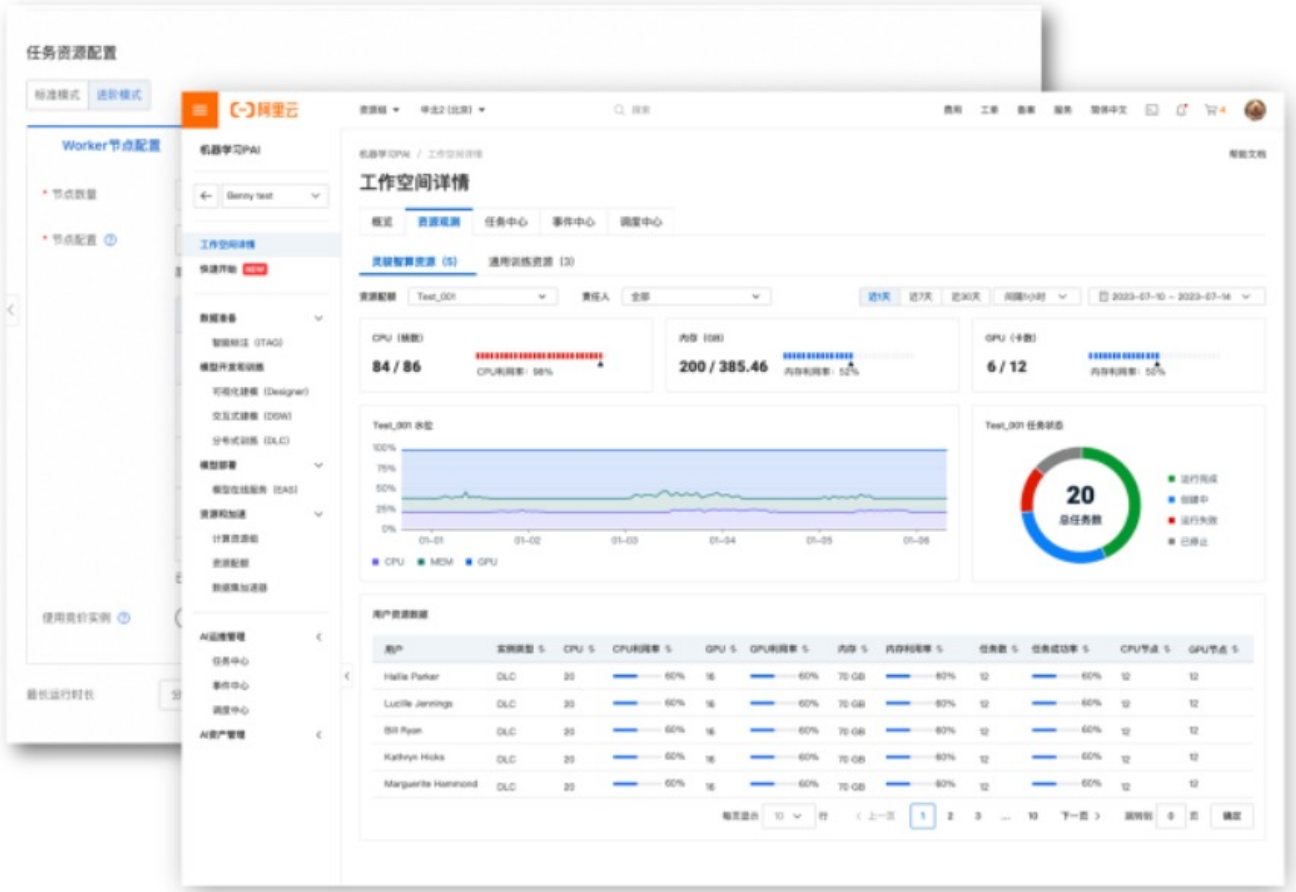
- 资源组/任务级别数十种监控告警指标
- 任务失败支持电话/短信/邮件等多种告警渠道

训练任务容错

- 基于AIMaster的容错监控
- AI大模型高性能状态保存恢复

任务优先级管理

- 支持多级Quota进行资源更精细化调度分配
- 支持按任务优先级/成员优先级/角色优先级多层次管理



PAI模型服务与AI推理



提升计算效率

EasyCkpt
秒级异步训练快照

- 秒级Checkpoint
- 训练精度无损
- 基于层次化架构的训练进度保存与恢复
- 支持任务销毁前自动checkpoint
- 简单易用、代码入侵性极小

AI Master
自动容错弹性训练

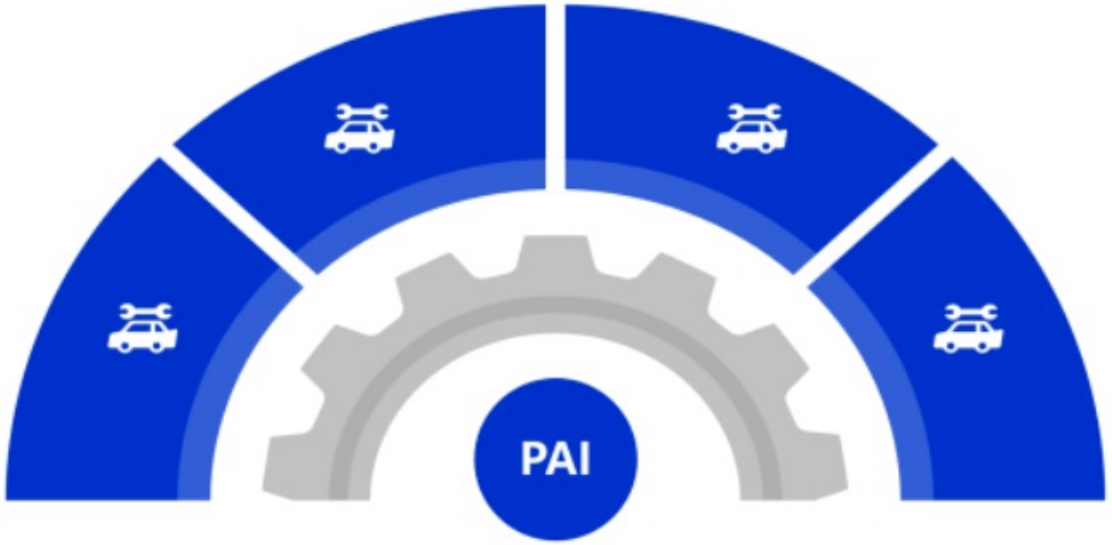
- per-job级别的组件，任务监控、决策、资源控制
- pod失败重启；Job退出异常检测；日志转储等
- 模型训练精度与DDP完全一致的分布式训练方式
- 训练任务作业弹性伸缩，支撑任务高效调度
- 训练任务作业容错处理增强平台稳定性
- 支持作业任务精细化管理，减少资源浪费

TorchAcc
模型训练加速框架

- 支持数据并行训练，优化计算和通信overlap，提升分布式训练效果
- 支持AMP训练加速
- 支持自动显存优化等

RLHF
基于人工反馈的强化学习

- PAI提供了完整的基于RLHF的大模型训练框架，一站式完成RLHF LLM训练。
- 提供了RLHF SDK快速开发RLHF训练中需要的SFT/RM/PPO模型。
- 通过PAI-DLC控制台，简单方便地提交RLHF训练任务，可视化地管理分布式任务。



PAI-Blade
通用推理优化

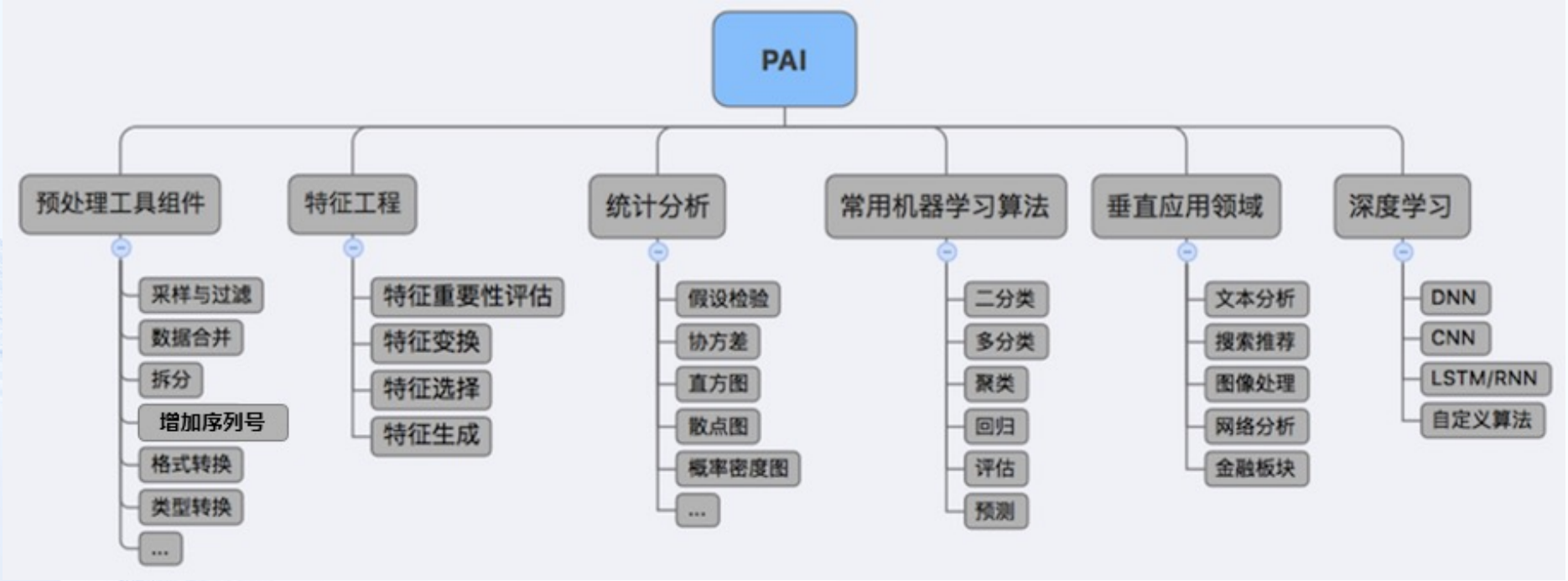
- 支持Tensorflow、PyTorch及主流加速设备：GPU、CPU、端侧设备。
- 支持计算图优化、厂商优化库、AI编译优化、高性能算子库、混合精度及自动压缩等多种优化技术。
- 易用，标准Python SDK，一个接口即可完成优化。

博学谷
www.boxuegu.com
黑马程序员线上品牌

04

PAI平台产品结构

产品结构

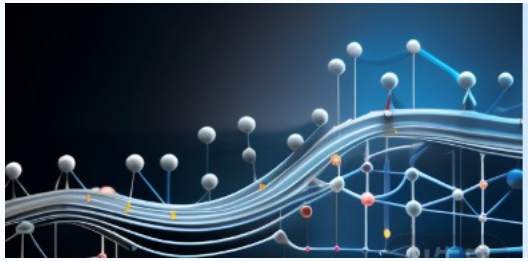

PAI
预处理工具组件
采样与过滤
数据合并
拆分
增加序列号
格式转换
类型转换
...
特征工程
特征重要性评估
特征变换
特征选择
特征生成
统计分析
假设检验
协方差
直方图
散点图
概率密度图
...
常用机器学习算法
二分类
多分类
聚类
回归
评估
预测
垂直应用领域
文本分析
搜索推荐
图像处理
网络分析
金融板块
深度学习
DNN
CNN
LSTM/RNN
自定义算法

支持机器学习算法



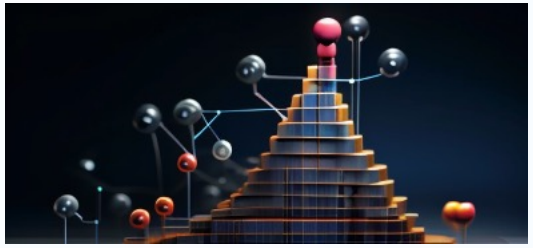
算法种类丰富

PAI平台支持基础的聚类和回归类算法，同时也支持文本分析和特征处理等复杂算法。



深度学习框架支持

PAI平台兼容TensorFlow、Caffe及MXNet等主流的机器学习框架，满足不同开发者的需求。



可视化建模方式

Designer工具封装了经典的机器学习算法，提供拖拽式搭建实验的可视化建模方式，无需编程即可实现模型构建。

一站式的机器学习体验

数据上传与预处理

PAI平台支持用户上传数据，并进行数据预处理操作。

特征工程与模型训练

PAI提供特征工程工具和分布式训练功能，助力用户高效完成模型训练。

模型评估与一键部署

PAI支持模型评估，同时提供一键式模型部署服务，实现模型到业务的无缝衔接。

主流深度学习框架支持

1 深度学习框架支持概述

PAI平台提供了对多种主流深度学习框架的支持，如Pytorch，TensorFlow、Caffe及MXNet等。

2 Pytorch框架应用

作为全球最受欢迎的深度学习框架之一，Pytor’ch在PAI平台上得到了全面支持和应用。

3 Caffe与MXNet框架

Caffe和MXNet作为另外两个主流深度学习框架，也在PAI平台上发挥了重要作用，满足了不同用户的需求。

可视化的建模方式

可视化建模工具-Designer

Designer封装了经典的机器学习算法，提供拖拽式搭建实验，无需编程。

交互式建模平台-DSW

DSW支持交互式建模，满足不同建模需求，让模型开发更灵活。

一键式模型部署服务

PAI支持将训练模型一键式发布为Restful API接口，实现模型到业务的无缝衔接。

博学谷
www.boxuegu.com
黑马程序员线上品牌

05

PAI平台架构

PAI平台架构


场景化
解决方案
AI应用：自动驾驶 / 科研智算 / 金融风控 / 智能推荐 / 智能设计 / 智慧城市 / 智能制造 / 智慧医疗 / 智慧法务 / ...
模型服务
(MaaS)
ModelScope 魔搭社区
PAI-DashScope 模型服务灵积
第三方MaaS平台
灵骏智算服务
&
人工智能框架
(PaaS)
工作空间
快速开始PAI-QuickStart / PAI-智码实验室 /PAI-DSW Gallery
标注服务
PAI-iTAG
特征平台
PAI-Feature
Store
可视化建模
PAI-Designer
交互式建模
PAI-DSW
分布式训练
PAI-DLC
模型在线服务
PAI-EAS
开发者工具
CLI / PaiFlow /
OpenAPI
AI资产管理（数据集 / 模型 / 镜像 / 代码 / 自定义组件 / ...
权限管理
MLOps
云产品依赖
优化与加速（DatasetAcc 数据集加速 / TorchAcc 训练加速 / EPL 并行训练框架 / Blade推理加速 / AI Master 自动容错训练 / EasyCkpt 秒级异步训练快照）
人工智能框架（PAI-TensorFlow / PAI-PyTorch / Alink / Spark, EasyRec / EasyPhoto / EasyTransfer / Megatron / DeepSpeed / RLHF）
计算资源
&
基础设施
(IaaS)
PAI-灵骏计算资源
云原生通用计算资源
大数据计算资源（MaxCompute / EMR / Flink）
异构计算磐久服务器（CPU、GPU）
高速 RDMA 网络（RoCE）
分布式存储 CPFS/NAS/OSS
容器服务 ACK
弹性计算 ECS
低PUE液冷/风冷，模块化IDC设施

- 基础设施层：包括CPU、GPU、FPGA及NPU。
- 计算引擎和容器服务层：包括MaxCompute、EMR、实时计算等计算引擎及容器服务ACK。
- 计算框架层：包括Alink、TensorFlow、PyTorch、Megatron、DeepSpeed、Caffe、MapReduce、SQL及MPI等计算框架，用于执行分布式计算任务。

- 按照机器学习全流程，PAI分别提供了数据准备、模型开发与训练及模型部署阶段的产品：
  1. 数据准备：PAI提供了智能标注，支持在多种场景下进行数据标注和数据集管理。
  2. 模型开发与训练：PAI提供了可视化建模（Designer）、交互式建模（DSW）、分布式训练（DLC）以及特征平台（FeatureStore），满足不同的建模需求。
  3. 模型部署：PAI提供了模型在线服务（EAS）和推理加速工具（Blade），帮助您快速地将模型部署为服务。同时，PAI提供了智能生态市场，您可以获取业务解决方案和模型算法，实现相关业务和技术的高效对接。
- 业务层：PAI应用于金融、医疗、教育、交通以及安全等各个领域。阿里巴巴集团内部的搜索系统、推荐系统及金融服务系统等，均依赖于PAI进行数据挖掘。

博学谷
www.boxuegu.com
黑马程序员线上品牌



扫码关注博学谷微信公众号