cube-studio

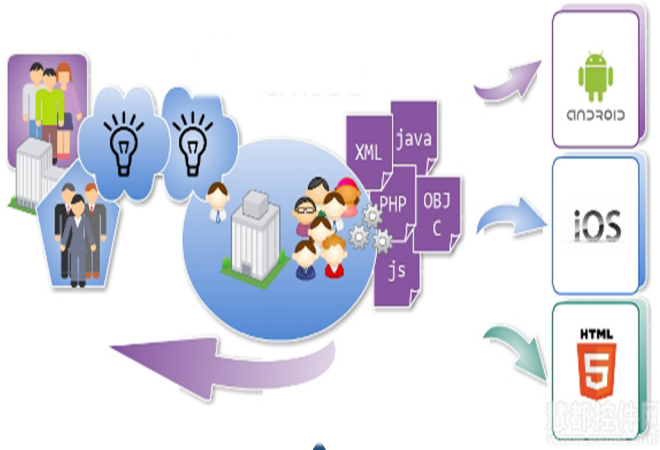
一站式云原生机器学习平台

AI | 用人工智能建设美好世界
CREATE A BETTER WORLD WITH ARTIFICIAL INTELLIGENCE

行业现状：AI平台发展现状和痛点

现状

- ✓ 数据智能化+大模型带来AI平台刚性需求
- ✓ 智能化转型在早期，一般不具备AI平台能力
- ✓ 数据和算力都不想上云

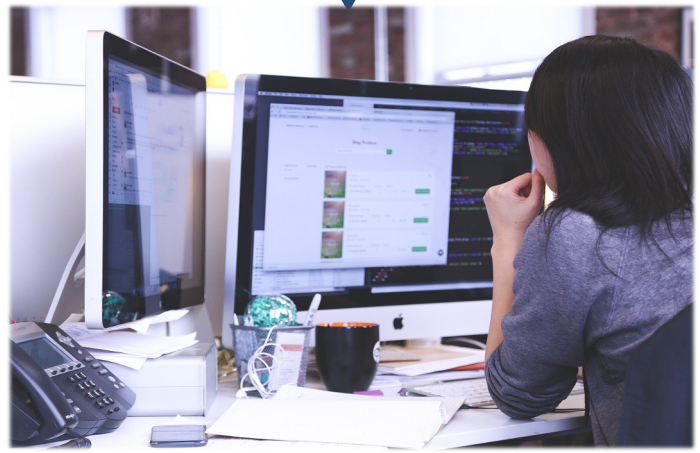


难点

- ✓ 专业门槛高，培养周期长
- ✓ 投入产出比不明朗
- ✓ 被破内卷，无法找到适合自己的落脚点

自主可控

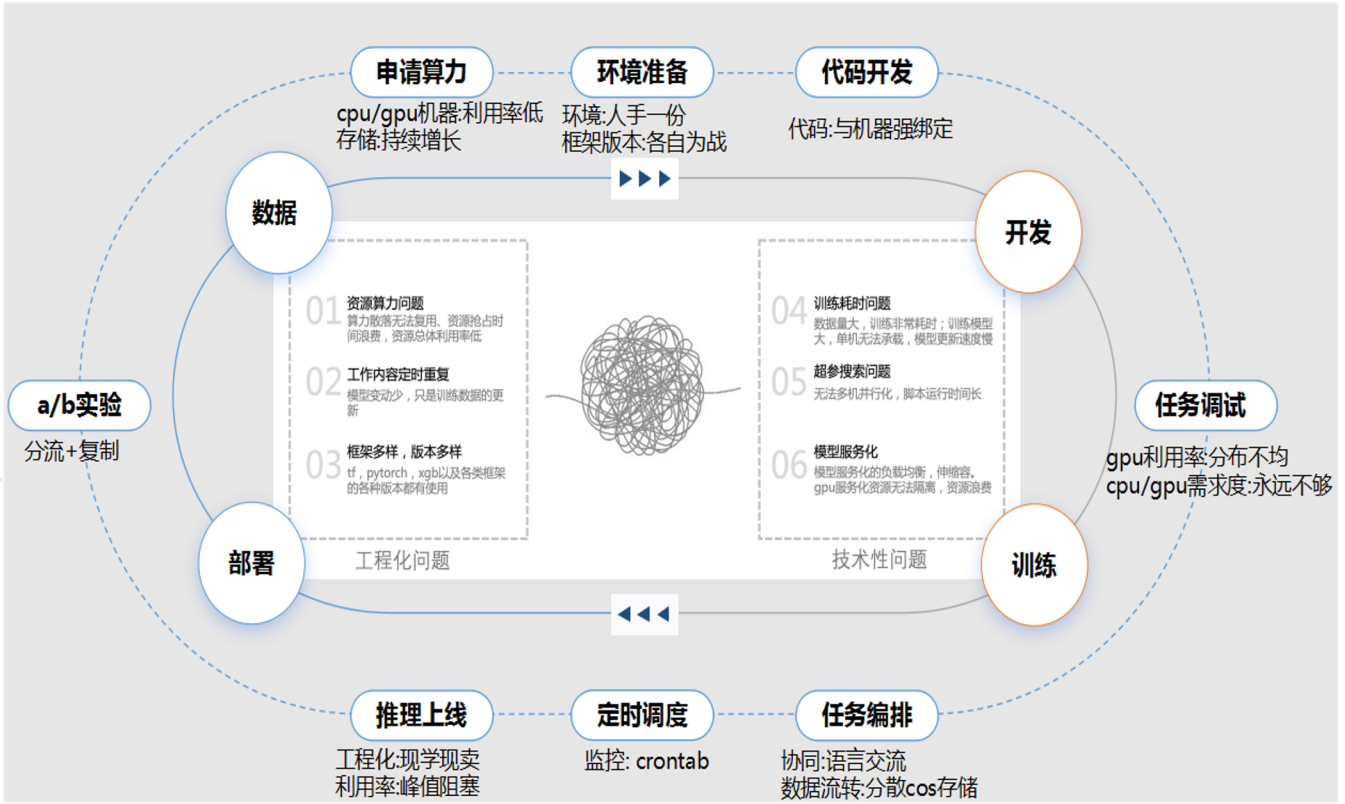
- ✓ 代码自主可控，可自主改造定制
- ✓ 私有化部署
- ✓ 算力可控，国产gpu/npu等



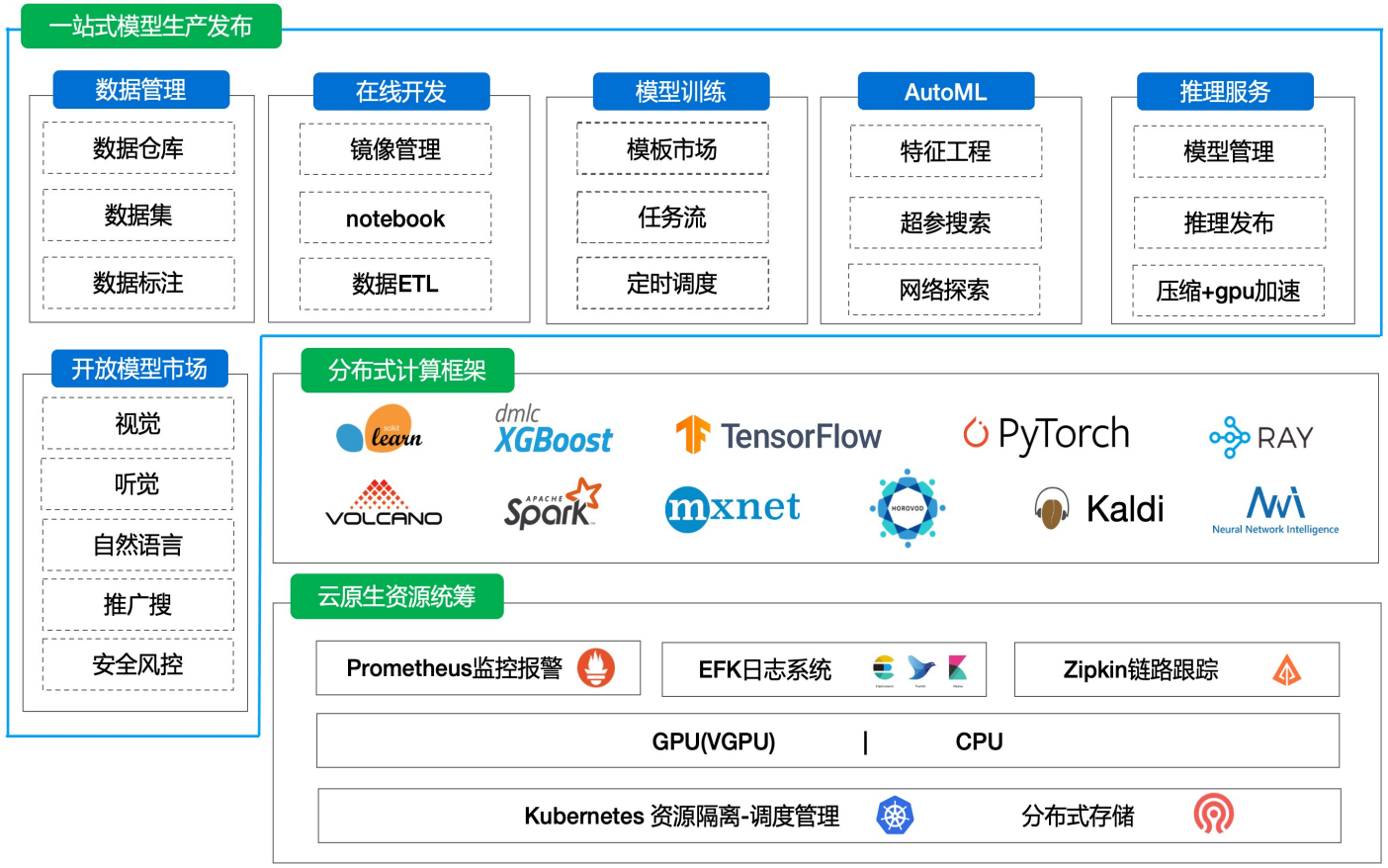
产品介绍：cube-studio是什么？

cube studio开源云原生一站式机器学习/深度学习AI平台，支持sso登录，多租户/多项目组，数据资产对接，notebook在线开发，拖拉拽任务流pipeline编排，多机多卡分布式算法训练，超参搜索，推理服务VGPU，多集群调度，边缘计算，serverless，标注平台自动化标注，数据集管理，大模型一键微调，私有知识库，AI应用商店，支持模型一键开发/推理/微调，私有化部署，支持国产cpu/gpu/npu芯片，支持RDMA。

对外共享和交换



产品功能架构



AI中台一站式　数据中台对接　AIHub模型市场　gpt大模型

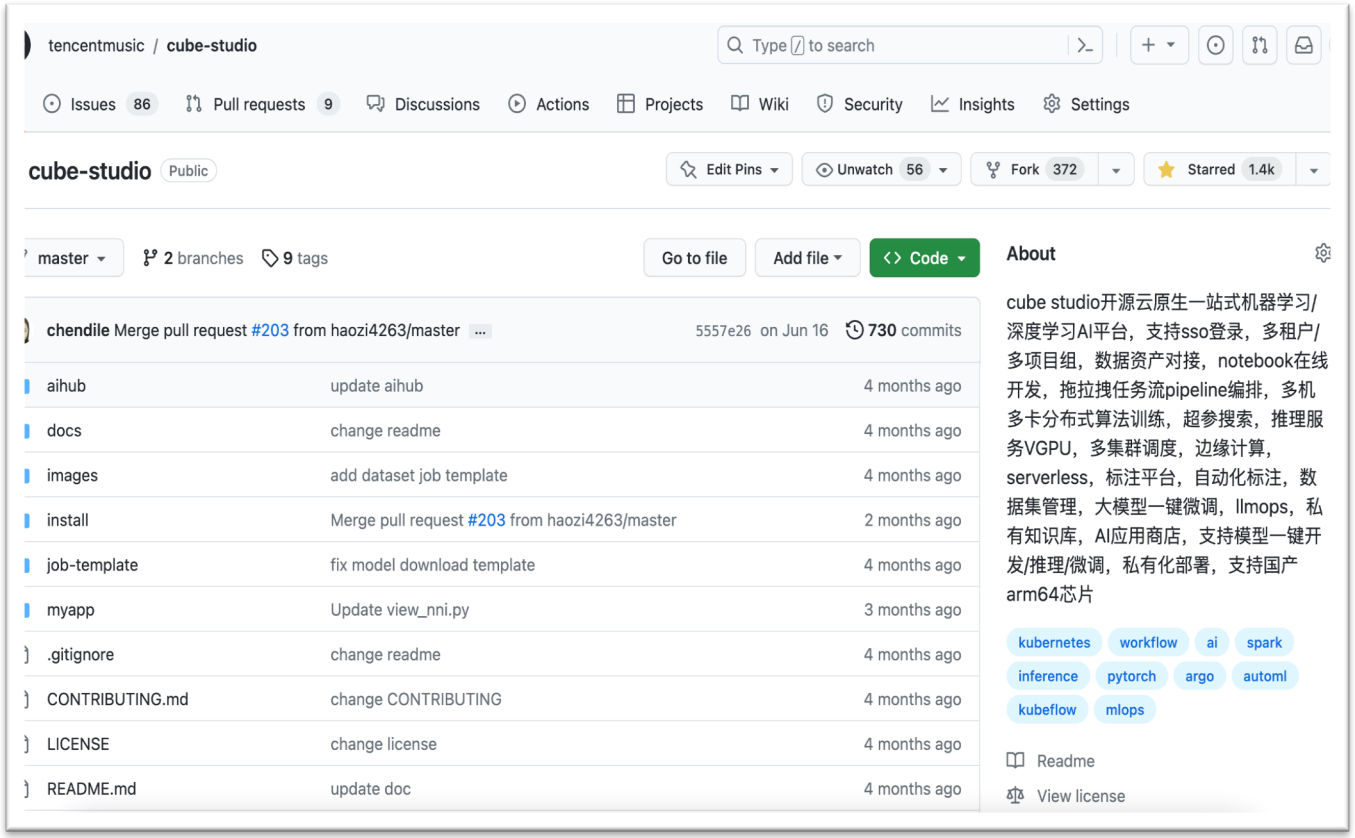
产品价值

90%	100%	60%	80%
完备性+稳定性	**源码交付**	**算力需求**	**合作共赢**
快速具备完备的算法平台：cube-studio包含完备的一站式学习平台能力，开源社区的补充，使其具备了各种场景的兼容性，腾讯内部实践，使其具备较好的稳定性。	细分市场的适配性：支持源码交付，可自行为客户进行定制，或加工改造为自己的专有产品。平台开发框架适合人力少情况的定制开发。	算力需求多样性：可私有化部署，支持国产算力，支持arm64架构，支持serverless弹性伸缩，支持边缘集群模式，支持多k8s集群。	与解决方案商合作共赢：不直接对接细分领域客户，与解决方案商无竞争关系，协助解决方案商/算力厂商与客户项目落地，参与分成。

友商功能对标

指 标	Cube studio	火山 引擎	百度Paddle	kube flow	微众 Prophecis	Meta Spore	阿里 PAI
是否自主研发	✓	✓	✓	/	✓	✓	✓
功能全面性	★★★	★★★	★★★★	★★★	★★	★	★★★★
灵活性	高	中	中	低	中	中	中
兼容性	★★★★	★★★	★★★	★	★★	★★	★★★
源码交付	是	否	否	是	否	否	否
用户体验	★★★★	★★★★	★★★★	★	★★	★★	★★★★
使用成本	★	★★★	★★★★	★	★★	★★	★★★★
是否开源	开源	不开源	部分	✓	✓	✓	部分
本地化服务	★★★★	★★★	★★★	/	/	/	★★★
销售渠道	★	★★★	★★★★	/	/	/	★★★★

国内mlops github开源第一



国内开源mlops平台第一名

日访问量3000+，日clone量100+， 社区参与900+人

上百家企业私有化部署

开源星「001号」落地（cube-studio），开源贡献赢神秘大礼包！

腾源会 TME数据智能 2022-08-17 19:45 发表于广东

第一波入选腾讯摘星计划

今年 5 月 6 日，腾讯·腾源会社区联合（项目名）等在内的 80 余家开源社区、国内外开源基金会等，共同发起「开源摘星计划」。开展 3 月以来，我们累计为近百位优秀摘星贡献者，送出激励大礼包 300 余份；同时为 700 位的开源爱好者搭建了共同的交流乐园，帮助很多人完成了从开源萌新到「过来人」的成长、蜕变。

主要功能介绍-算力/存储/用户管理

算力

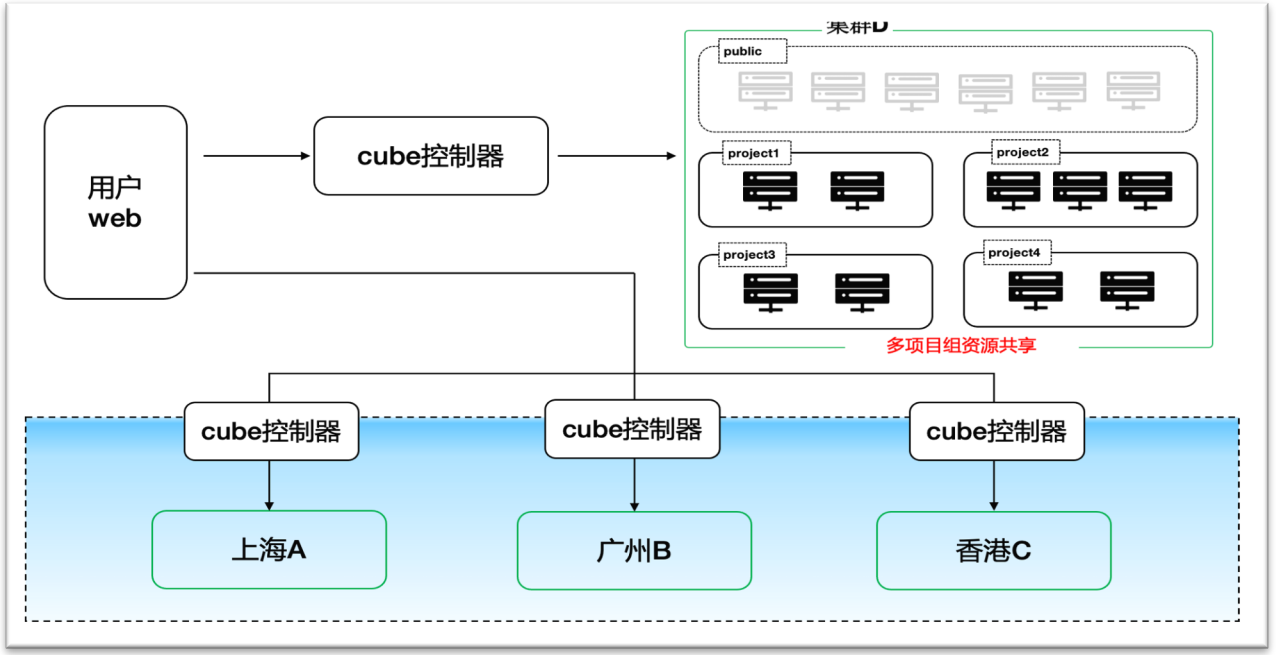
- ✓云原生统筹平台cpu/gpu等算力
- ✓支持划分多资源组，支持多k8s集群，多地部署
- ✓支持T4/V100/A100/昇腾/VGPU等异构GPU/NPU环境
- ✓支持边缘集群模式，支持边缘节点上开发/训练/推理
- ✓支持鲲鹏芯片arm64架构，RDMA

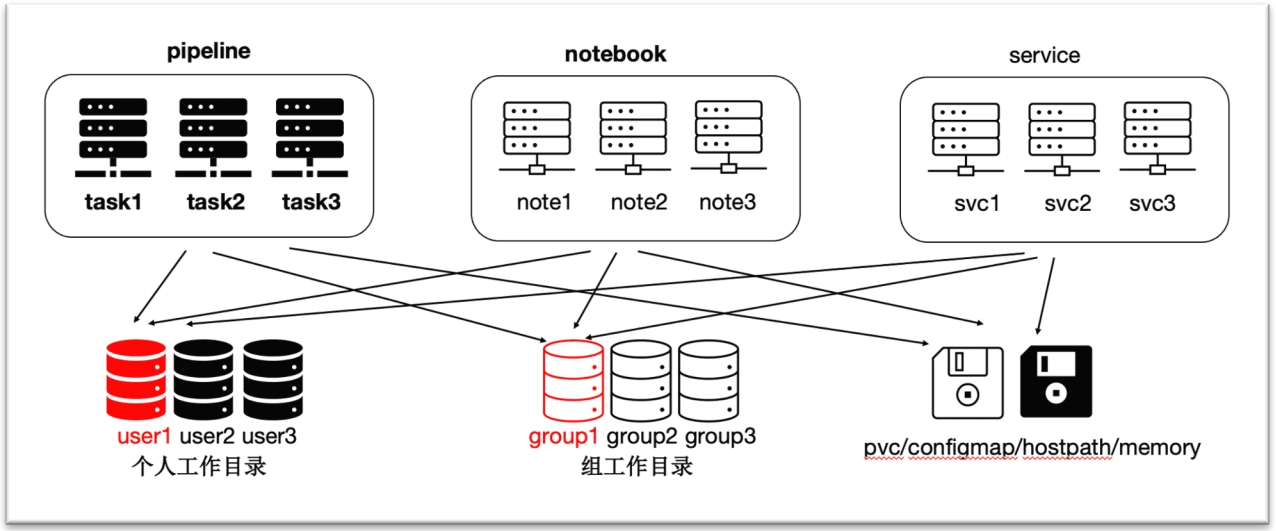
存储

- ✓自带分布式存储，支持多机分布式下文件处理
- ✓支持外部存储挂载，支持项目组挂载绑定
- ✓支持个人存储空间/组空间等多种形式
- ✓平台内存储空间不需要迁移

用户权限

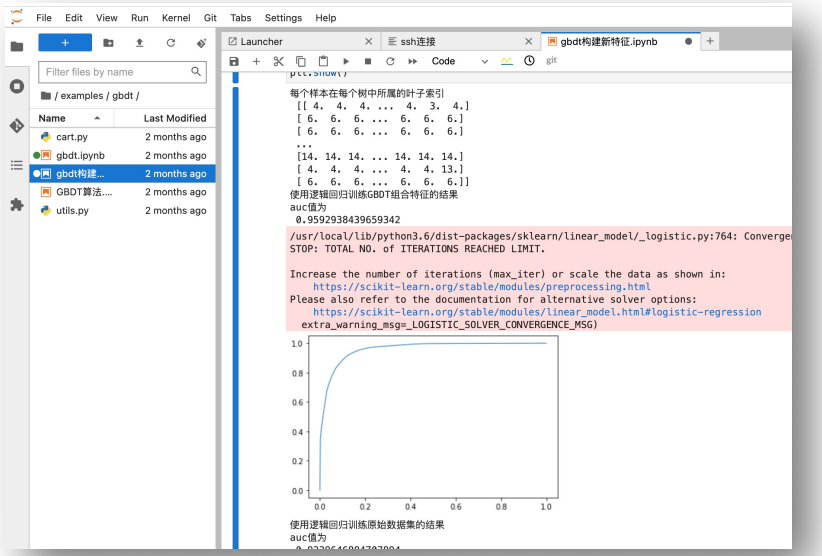
- ✓支持sso登录，对接公司账号体系
- ✓支持项目组划分，支持配置相应项目组用户的权限
- ✓管理平台用户的基本信息，组织架构，rbac权限体系



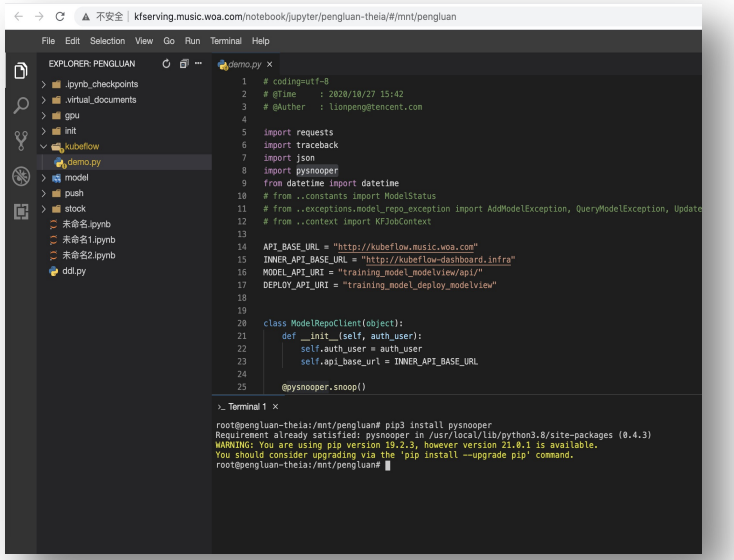


MLOPS-一站式开发工具

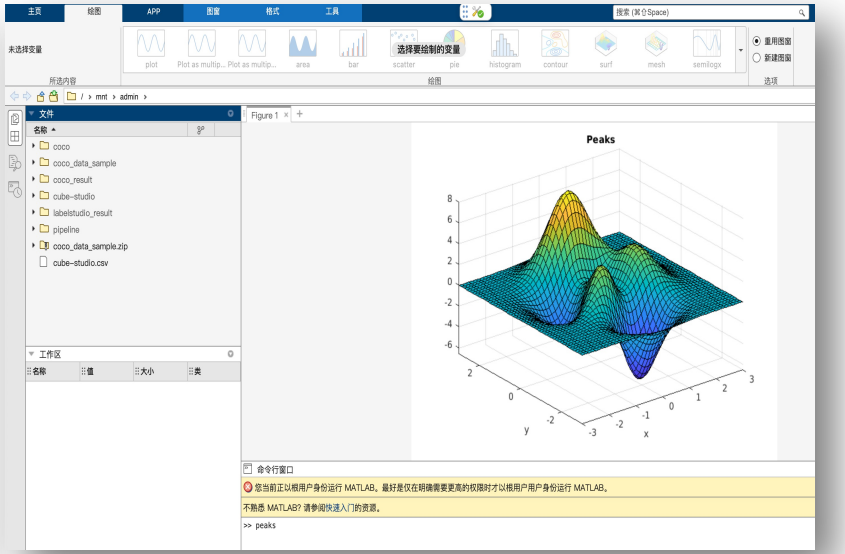
- ✓ 系统多租户/多实例管理，在线交互开发调试，无需安装三方控件，只需浏览器就能完成开发。
- ✓ 支持vscode，jupyter，Matlab，Rstudio等多种在线IDE类型
- ✓ Jupyter支持cube-studio sdk，Julia，R，python，pyspark多内核版本，



jupyter



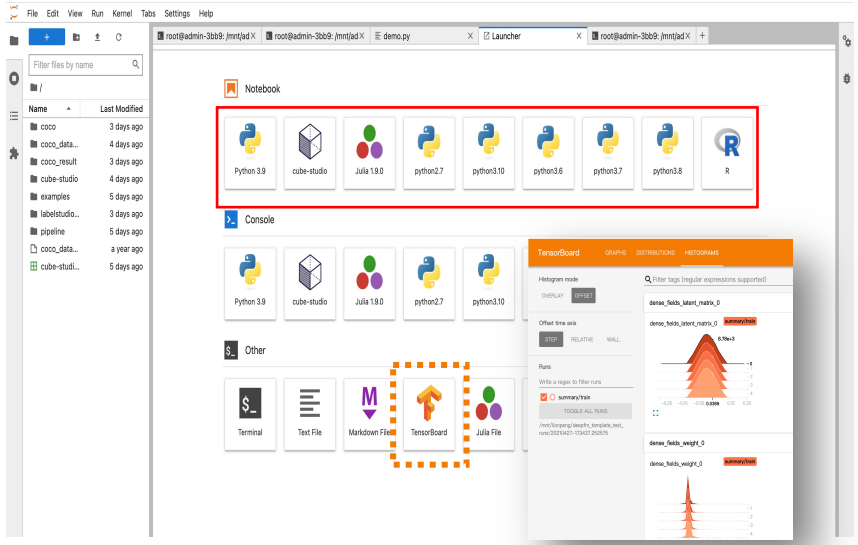
vscode



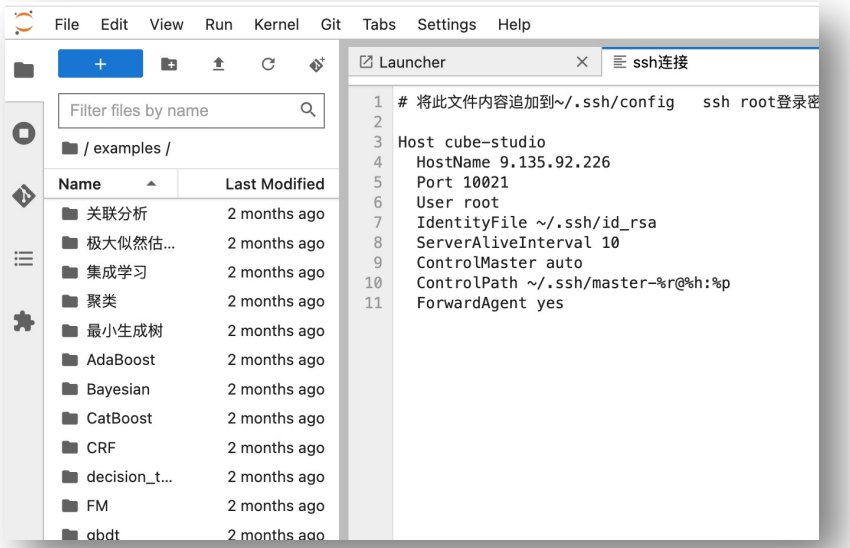
matlab

MLOPS-一站式开发工具

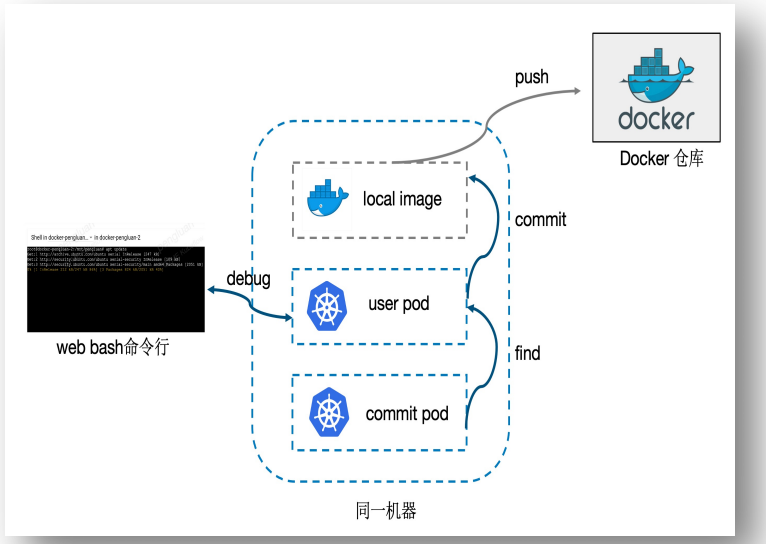
- ✓ 支持c++，java，conda等多种开发语言，以及tensorboard/git/gpu监控等多种插件
- ✓ 支持ssh remote与notebook互通，本地进行代码开发
- ✓ 在线镜像构建，通过Web Shell方式在浏览器中完成构建；并提供各种版本notebook，inference，gpu，python等基础镜像



多种内核/插件



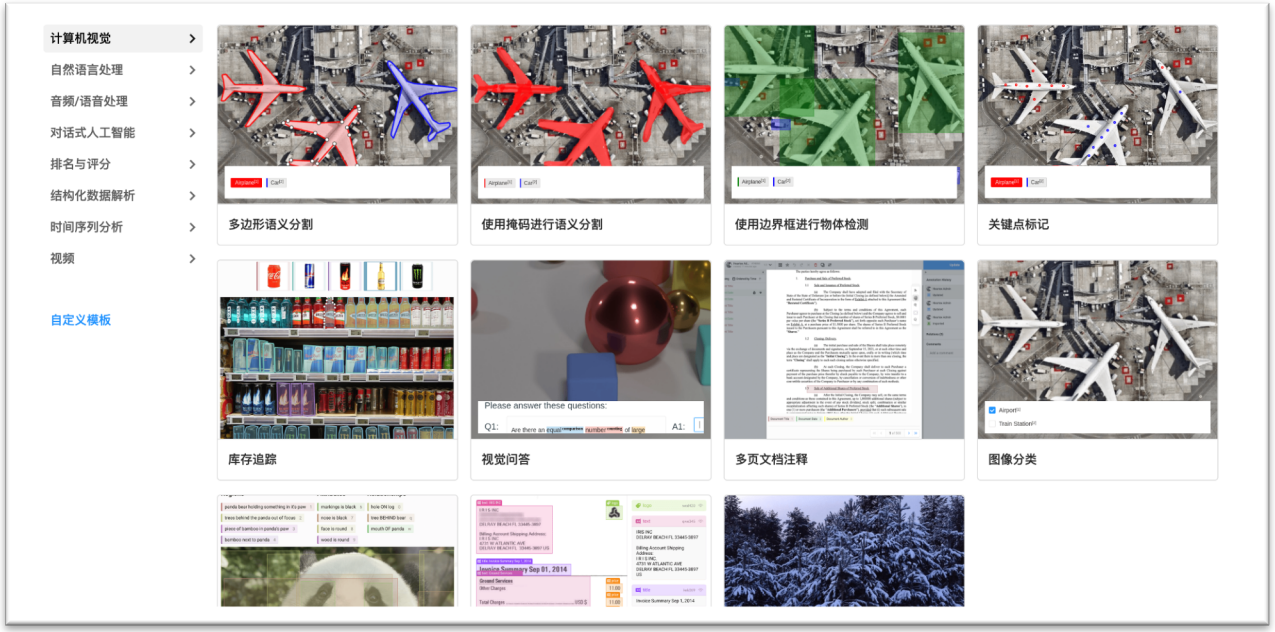
ssh远程开发



在线构建镜像

MLOPS-可视化数据标注

- ✓ 支持图/文/音/多模态/大模型多种类型标注功能，用户管理，工作任务分发
- ✓ 对接aihub模型市场，支持自动化标注；对接数据集，支持标注数据导入；对接pipeline，支持标注结果自动化训练



图文音多模态大模型类型标注

对接aihub/大模型支持自动化标注

MLOPS-拖拉拽建模流程

模板列表　　拖拉拽编排　　任务配置

MI全流程

1、数据导入，数据预处理，超参搜索，模型训练，模型评估，模型压缩，模型注册，服务上线，机器学习/深度学习/大模型全流程

灵活开放

2、便捷的拖拉拽方式，编排算法dag，支持单任务和pipeline整体等多种调试运行方式

3、丰富多样的计算任务模板。支持自定义模板

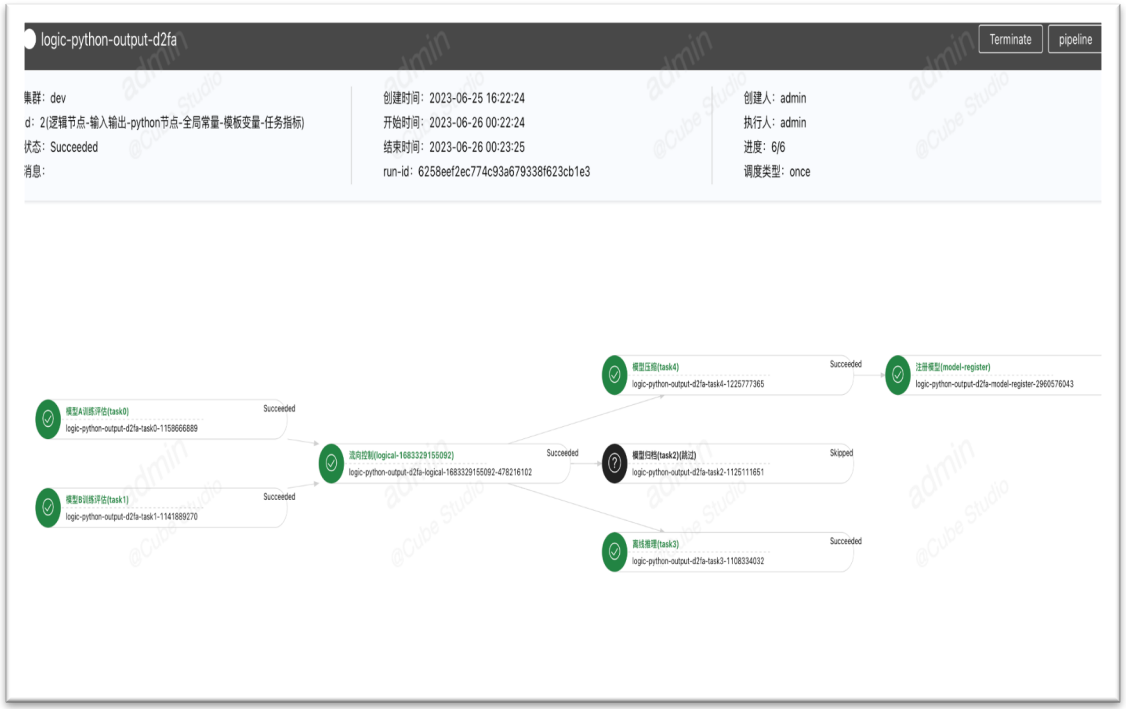
分布式计算

4、便捷的多机多卡分布式训练，通过标准化方式提供分布式多机多卡计算和训练的能力, 提升平台落地大规模计算所需要的效率.

MLOPS-多层次多类型算子

MLOPS-流水线调试

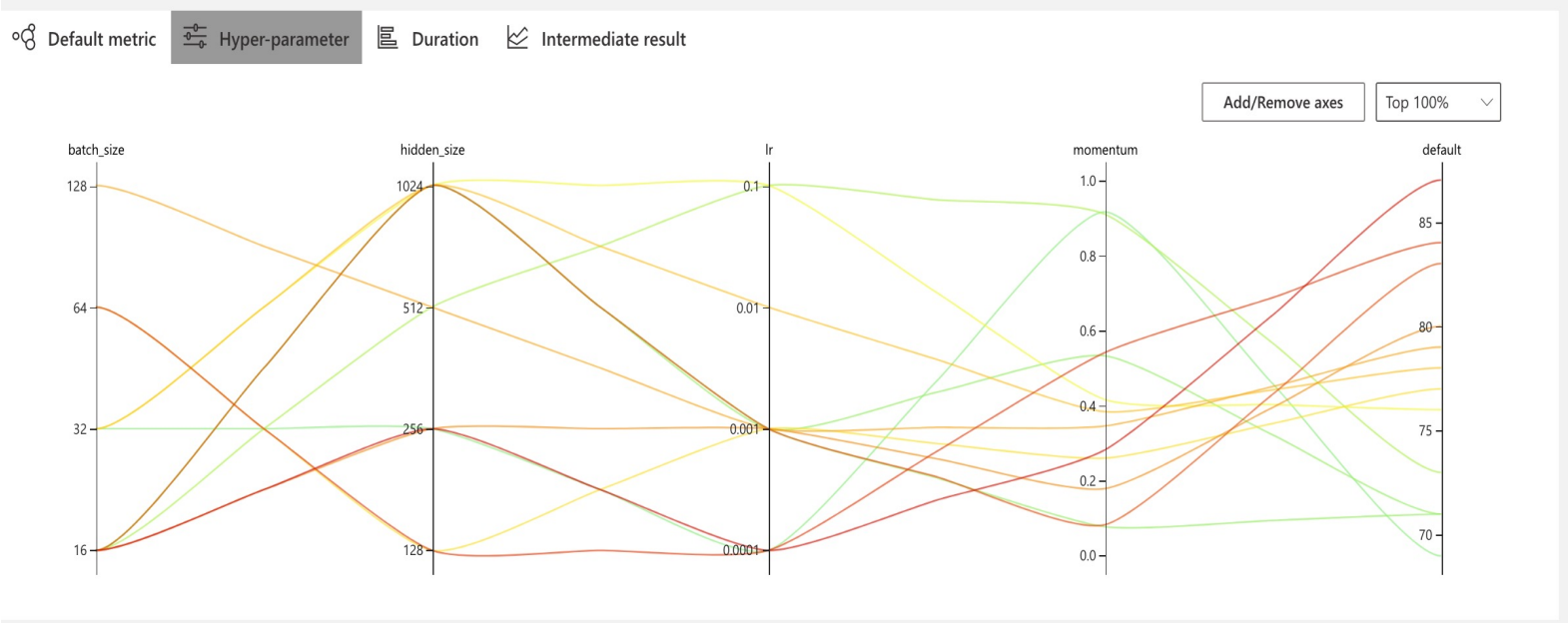
- ✓ Pipeline调试支持定时执行，支持，补录，并发限制，超时，实例依赖等。
- ✓ Pipeling运行，支持变量在任务间输入输出，全局变量，流向控制，模板变量，数据时间等
- ✓ Pipeling运行，支持任务结果可视化，图片、csv/json，echart源码可视化

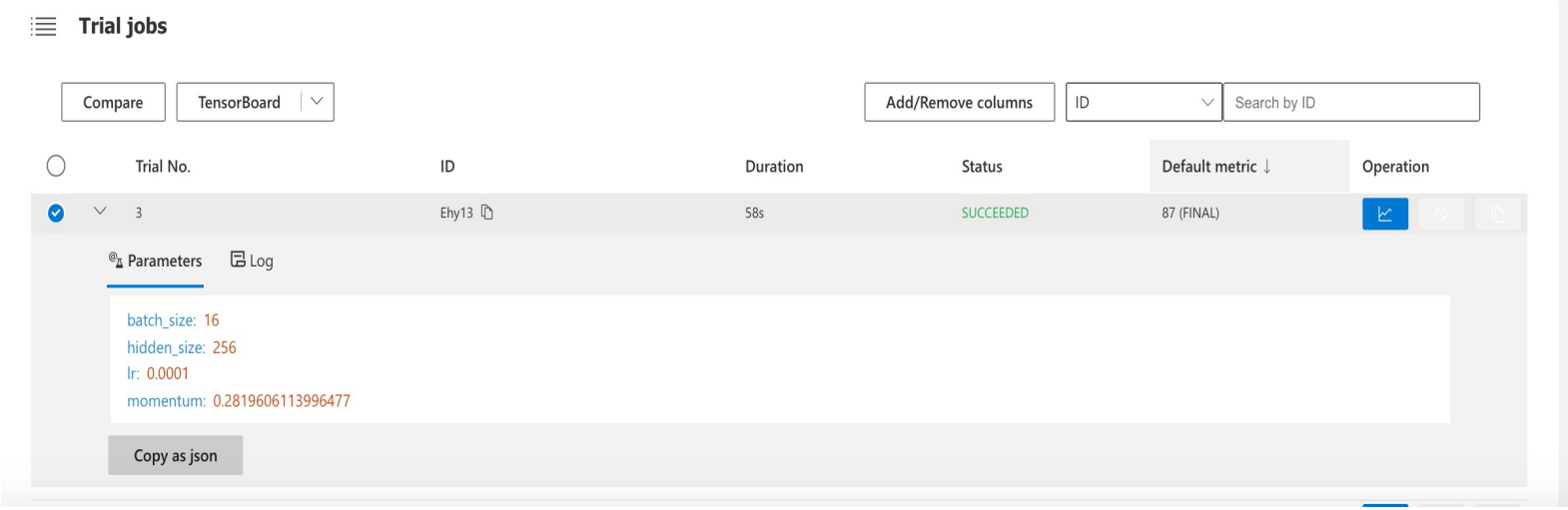


MLOPS-超参搜索

✓ 界面化呈现训练各组数据，通过图形界面进行直观呈现。

✓ 减少以往开发调参过程的枯燥感，让整个调参过程更加生动具有趣味性，完全无需丰富经验就能实现更精准的参数控制调节。

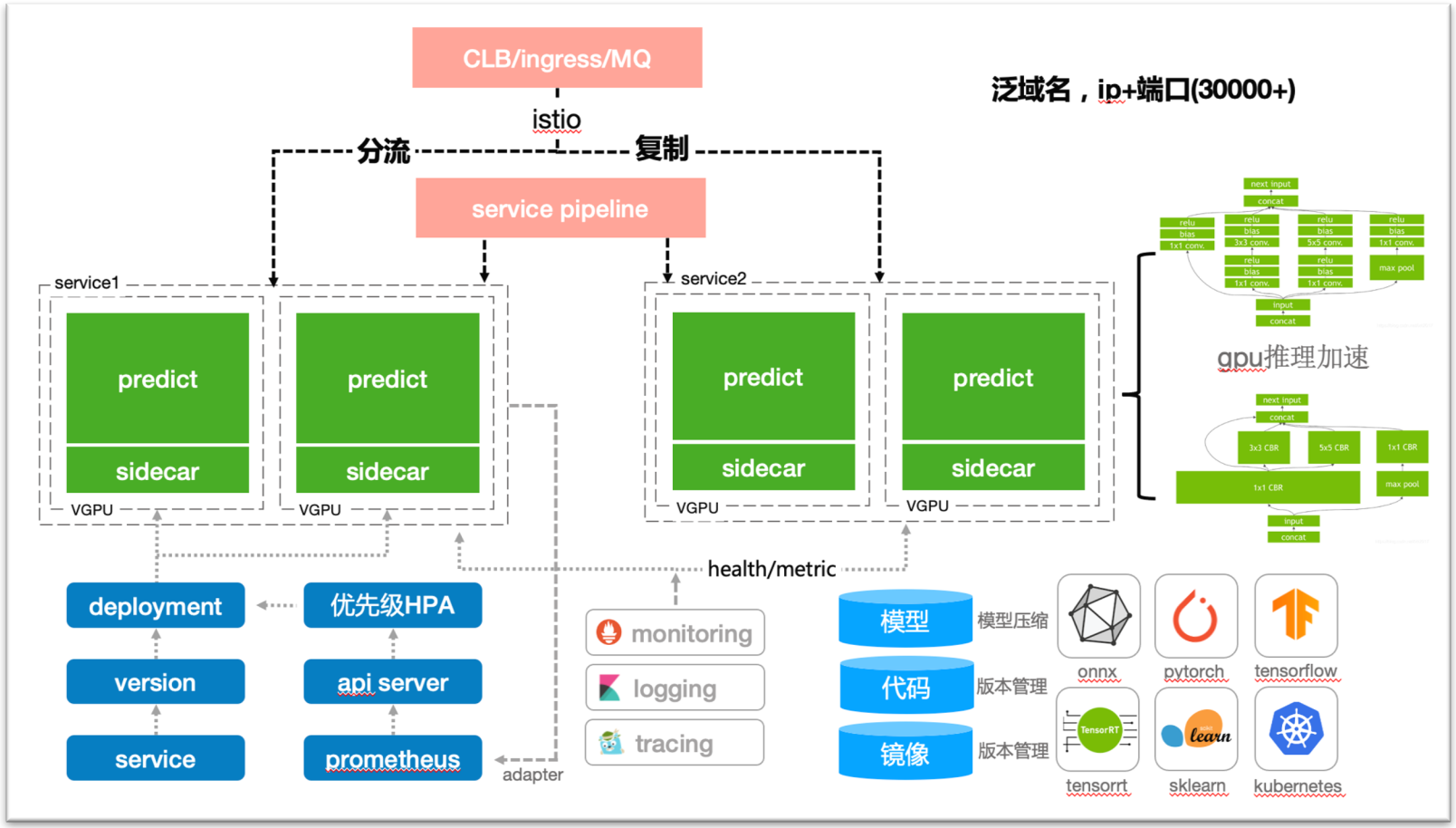




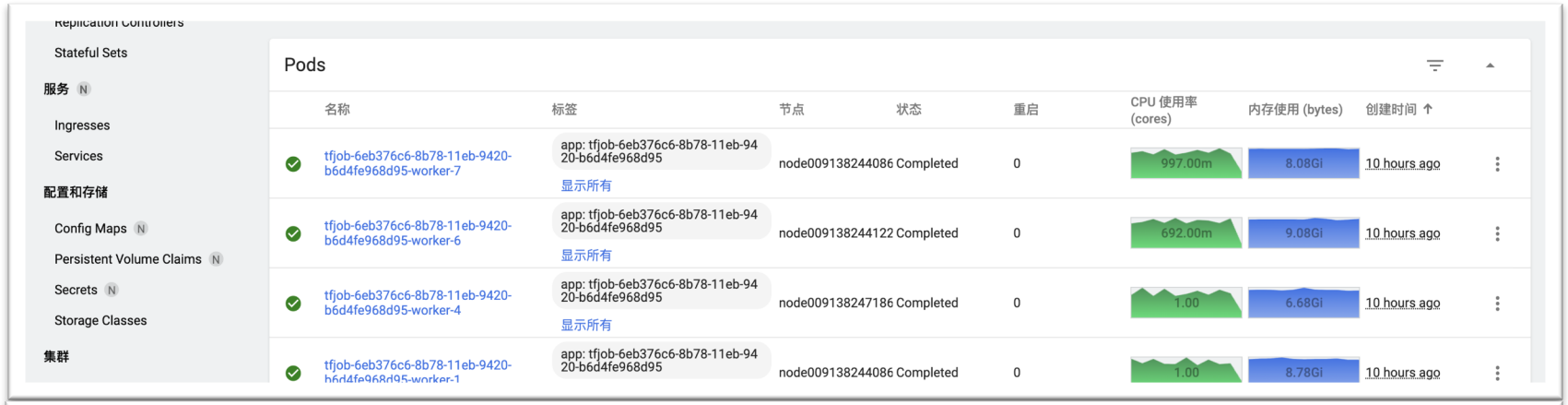

```
# 上报当前迭代目标值
nni.report_intermediate_result(test_acc)
# 上报最终目标值
nni.report_final_result(test_acc)

# 接收超参数为输入参数
parser.add_argument('--batch_size', type=int)
```

MLOPS-自动化零代码推理部署

- ✓ 支持模型管理注册，灰度发布，版本回退，模型指标可视化，以及在piepline中进行模型注册
- ✓ 推理服务支持多集群，多资源组，异构gpu环境，平台资源统筹监控，VGPU，服务流量分流，复制，sidecar
- ✓ 支持0代码的模型发布，gpu推理加速，支持训练推理混部，服务优先级，自定义指标弹性伸缩。



MLOPS-监控报警体系



运行通知

pipeline: crontab-standalone-train(调度测试)
namespace: pipeline
status: Running
start_time: 2021-05-30 03:01:02
finish_time:
pod 详情

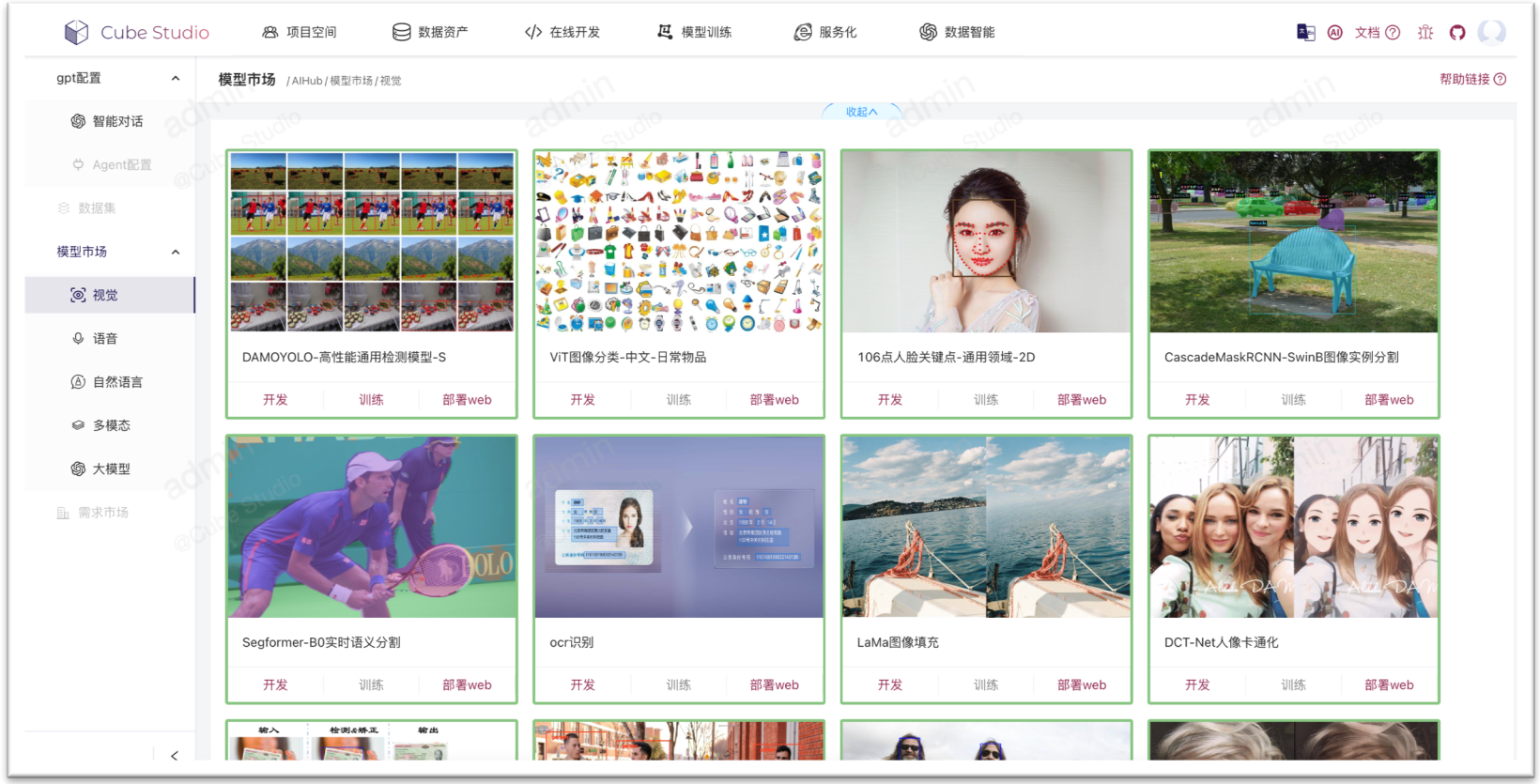
星期日 上午 4:01

workflow: crontab-standalone-train-bgm2x
pipeline: crontab-standalone-train(调度测试)
namespace: pipeline
status: Succeeded
start_time: 2021-05-30 03:01:02
finish_time: 2021-05-30 04:01:11
pod 详情

调度测试，各 task 耗时，酌情优化:
tlinux 命令测试:1.0(h)
tlinux 命令测试 2:1.0(h)

AIHub-模型市场

- ✓ 系统自带通用模型数量**400+**，覆盖绝大数行业场景，根据需求可以不断扩充。
- ✓ 模型开源、按需定制，方便快速集成，满足用户业务增长及二次开发升级。
- ✓ 模型标准化开发管理，大幅降低使用门槛，开发周期时长平均下降30%以上。



AIHub-模型市场

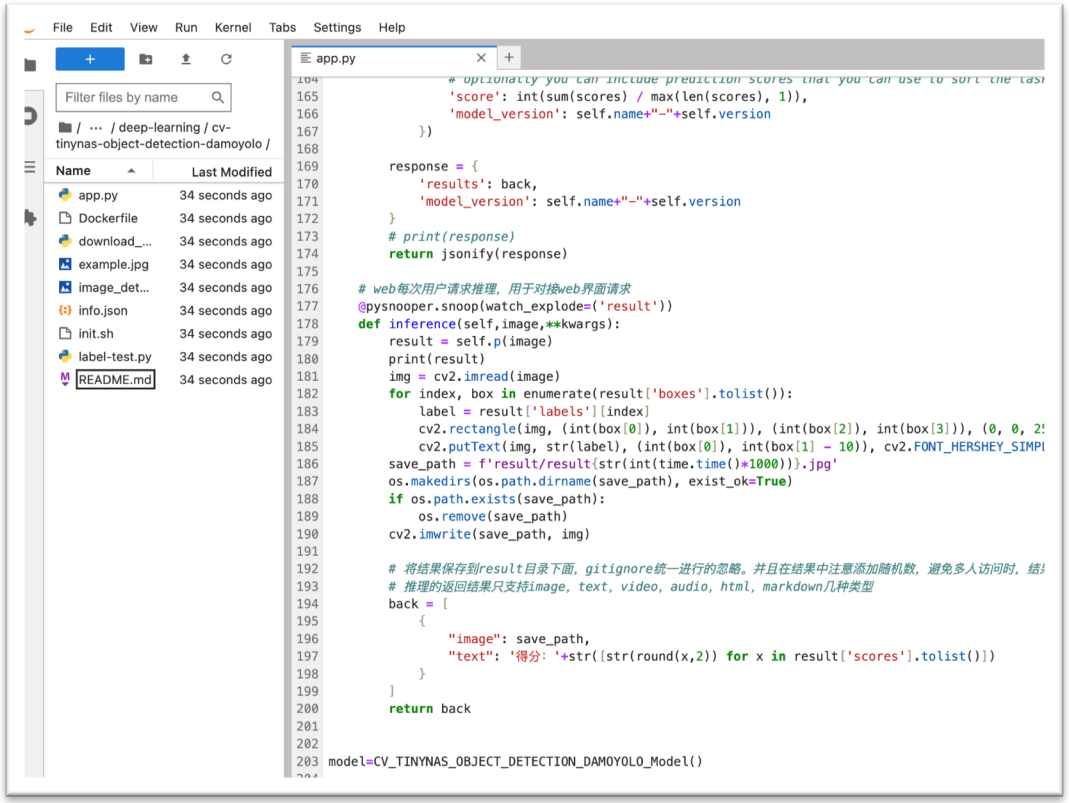
- ✓ AIHub模型可一键部署为WEB端应用，手机端/PC端皆可，实时查看模型应用效果
- ✓ 点击模型开发即可进入notebook进行模型代码的二次开发，实现一键开发
- ✓ 点击训练即可加入自己的数据进行一键微调，使模型更贴合自身场景

PC版

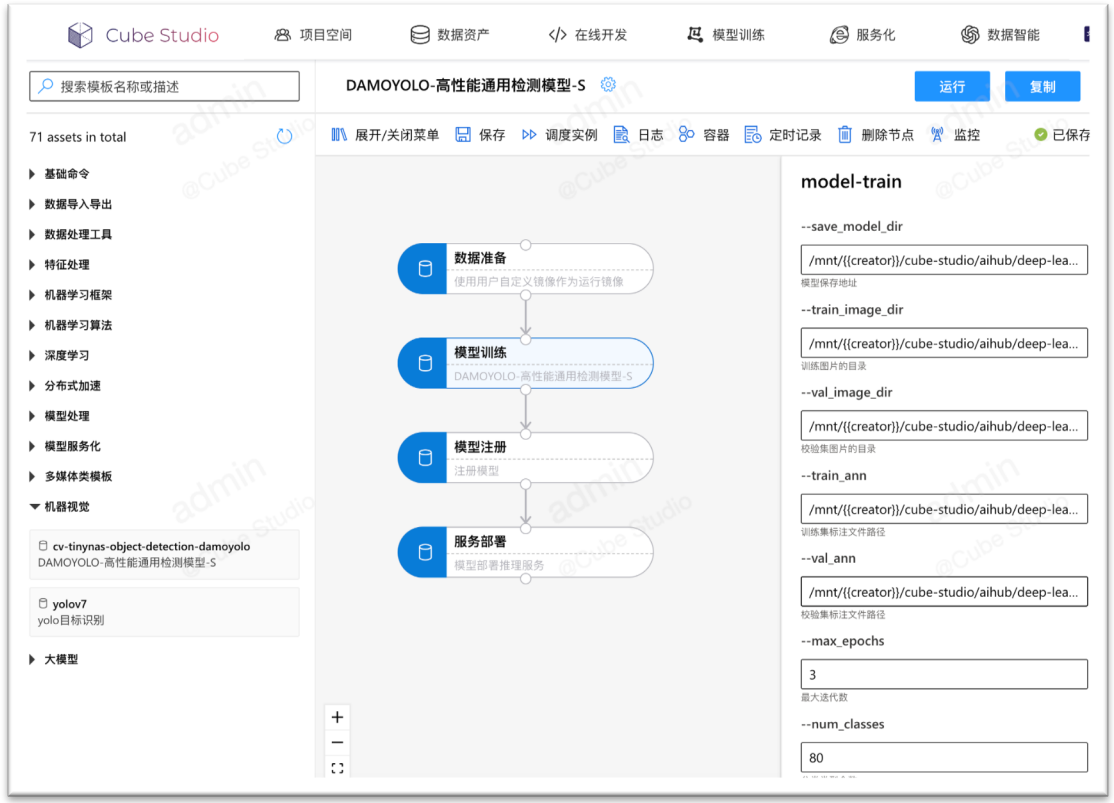
手机/微信版

AIHub-模型市场

- ✓ AIHub模型可一键部署为WEB端应用，手机端/PC端皆可，实时查看模型应用效果
- ✓ 点击模型开发即可进入notebook进行模型代码的二次开发，实现一键开发
- ✓ 点击训练即可加入自己的数据进行一键微调，使模型更贴合自身场景



Notebook在线开发(包含所有环境)



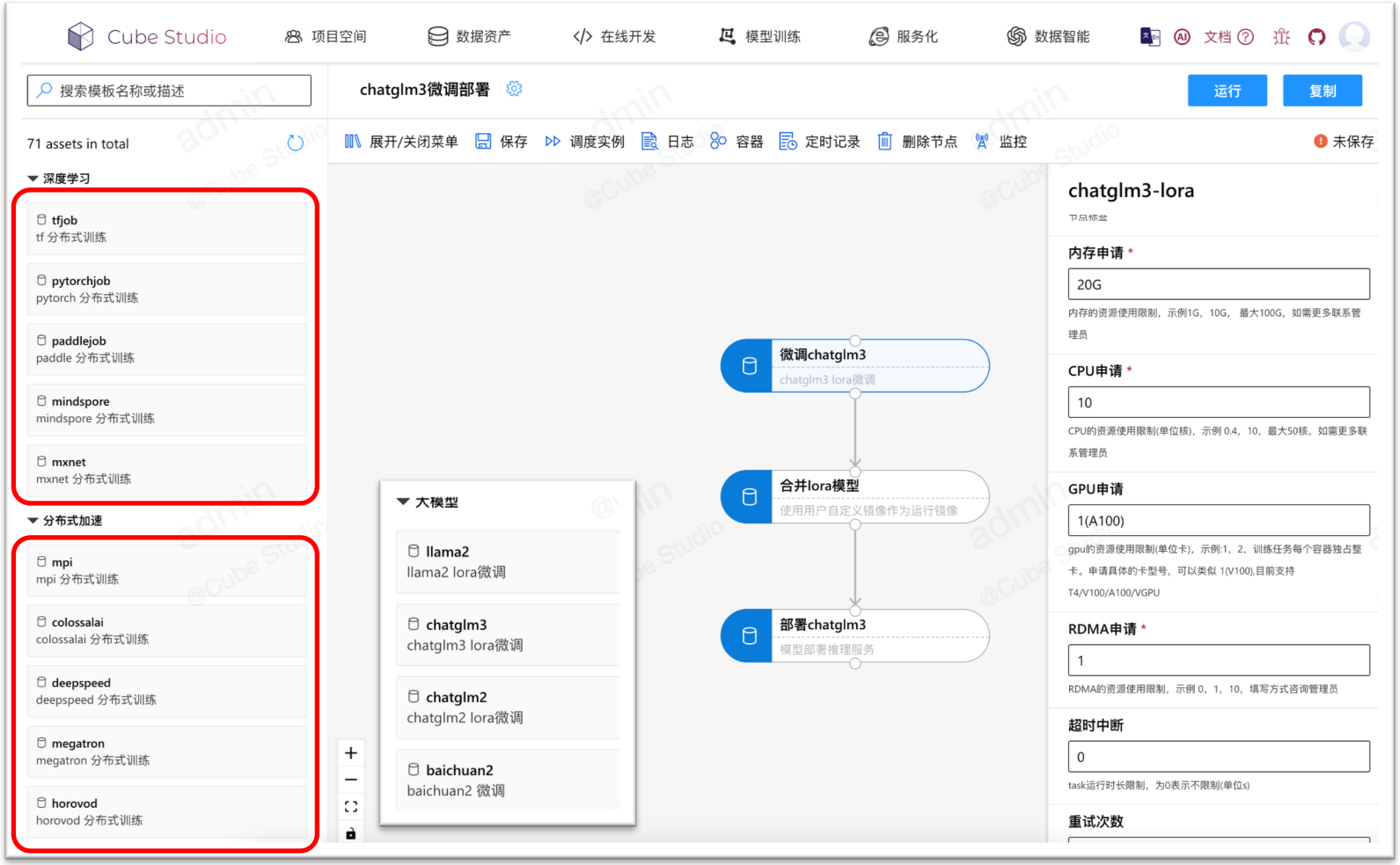
pipeline微调

GPT-自动化标注

✓ 借助大模型进行常规nlp任务的自动化标注

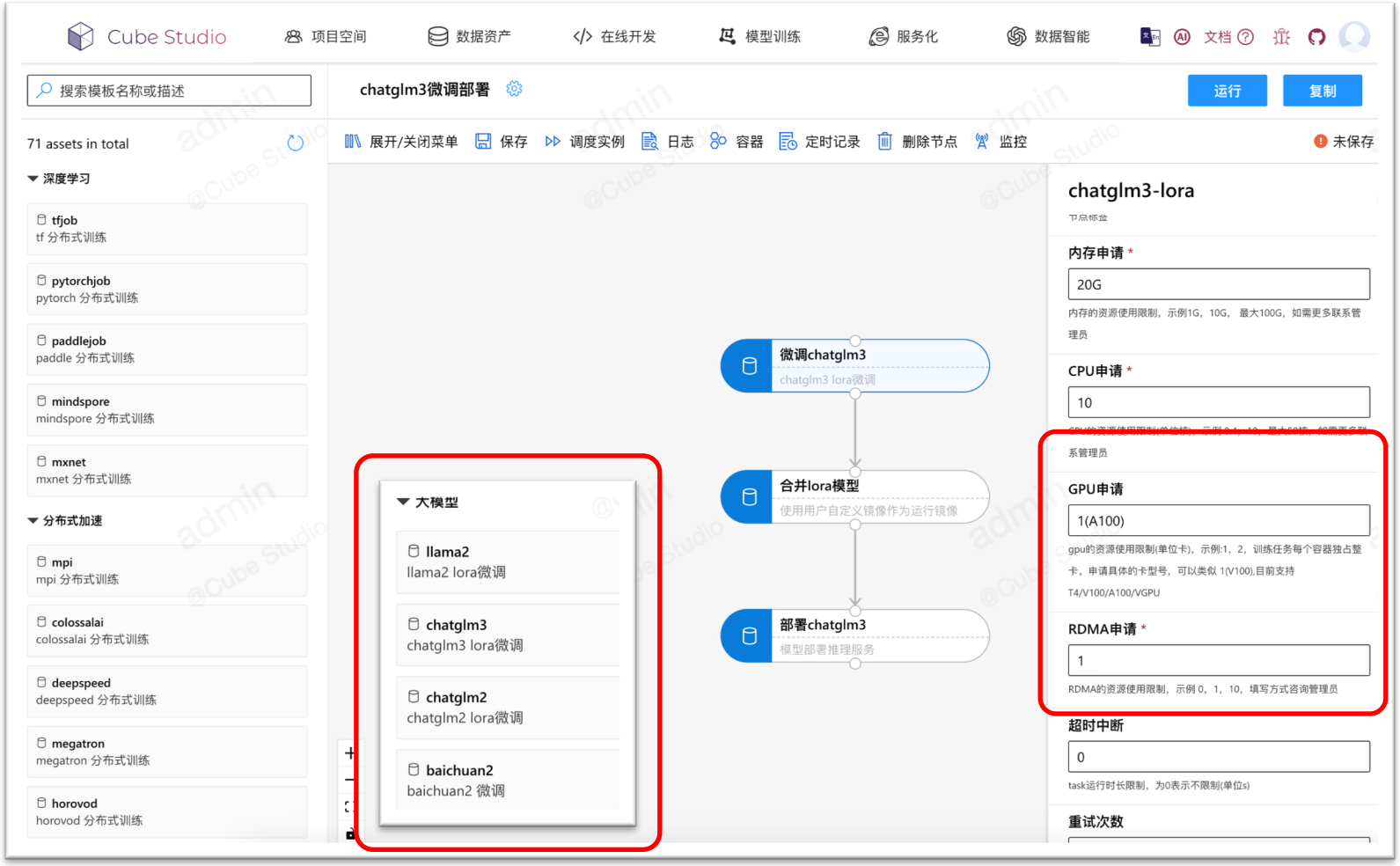
GPT-大模型微调

- ✓ 支持tfjob/pytorch/paddlejob/mindspore/mxnet等分布式训练框架
- ✓ 支持deepspeed/megatron/colossalai/horovod/mpi等分布式加速框架



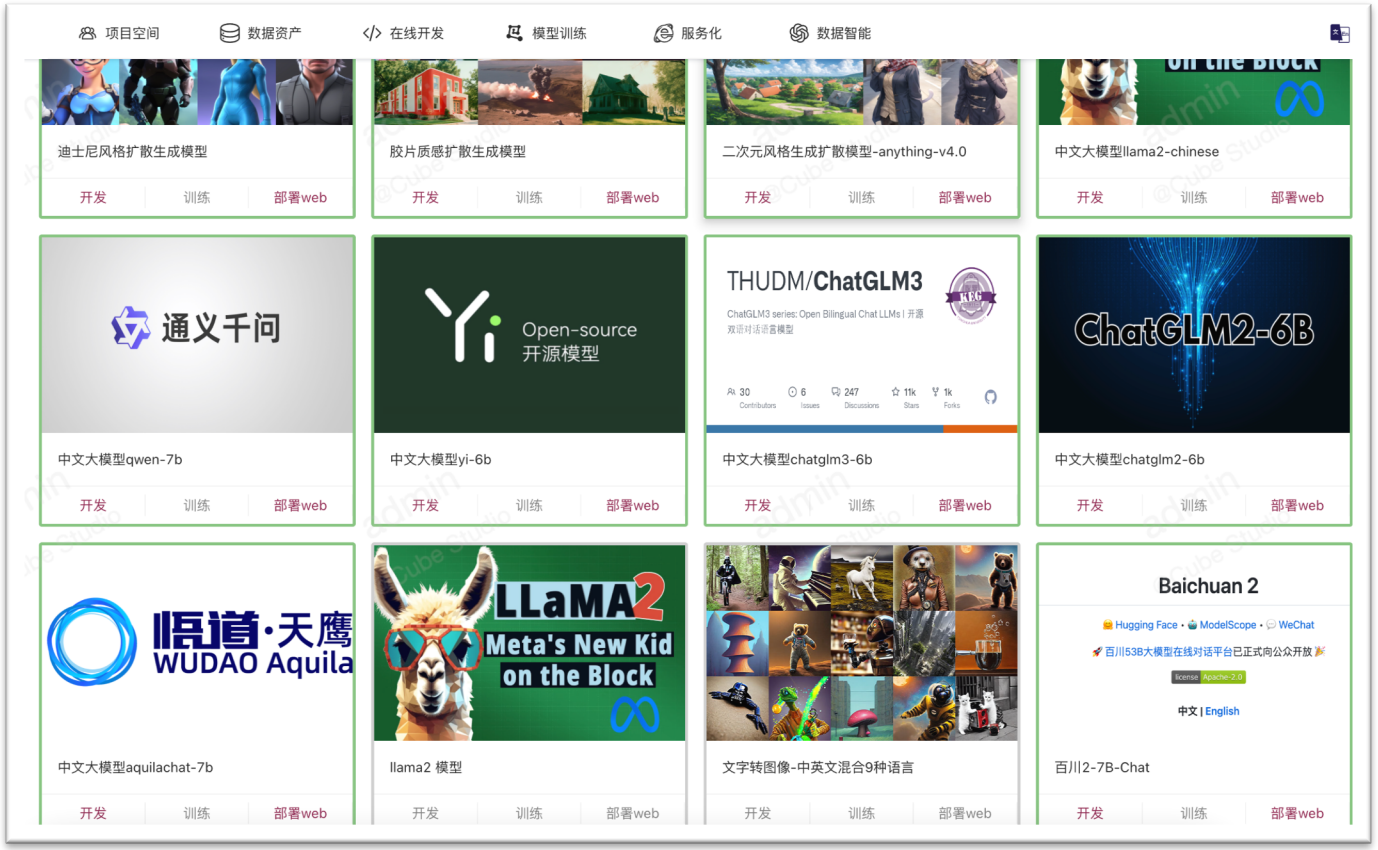
GPT-大模型微调

- ✓ 支持chatglm2/chatglm3/baichuan2/llama2等lora微调，提供chatglm微调/模型合并/部署全链路
- ✓ 支持RDMA，支持ib交换机，国产gpu卡，不同gpu卡型选择



GPT-推理部署

- ✓ AIHub形式部署开源大模型的部署
- ✓ 支持fastchat+vllm推理加速，形成openai接口

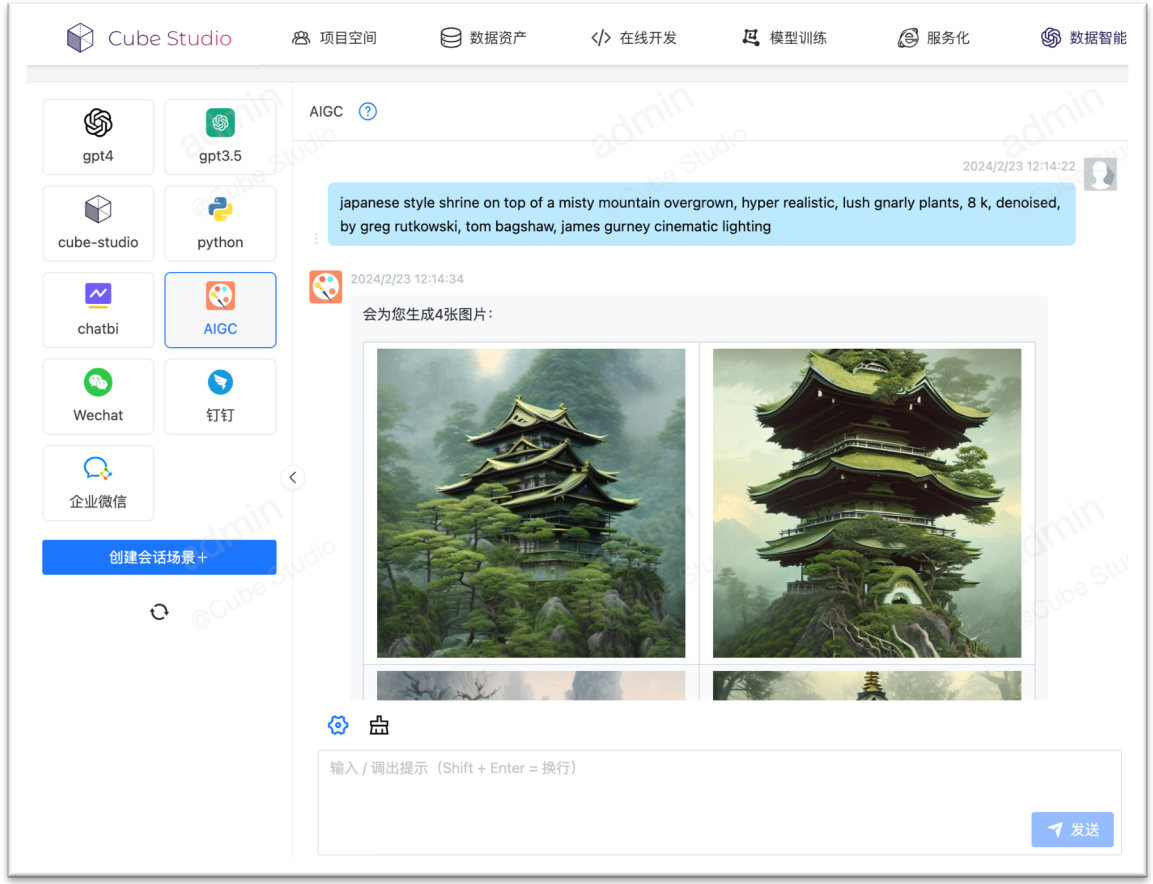


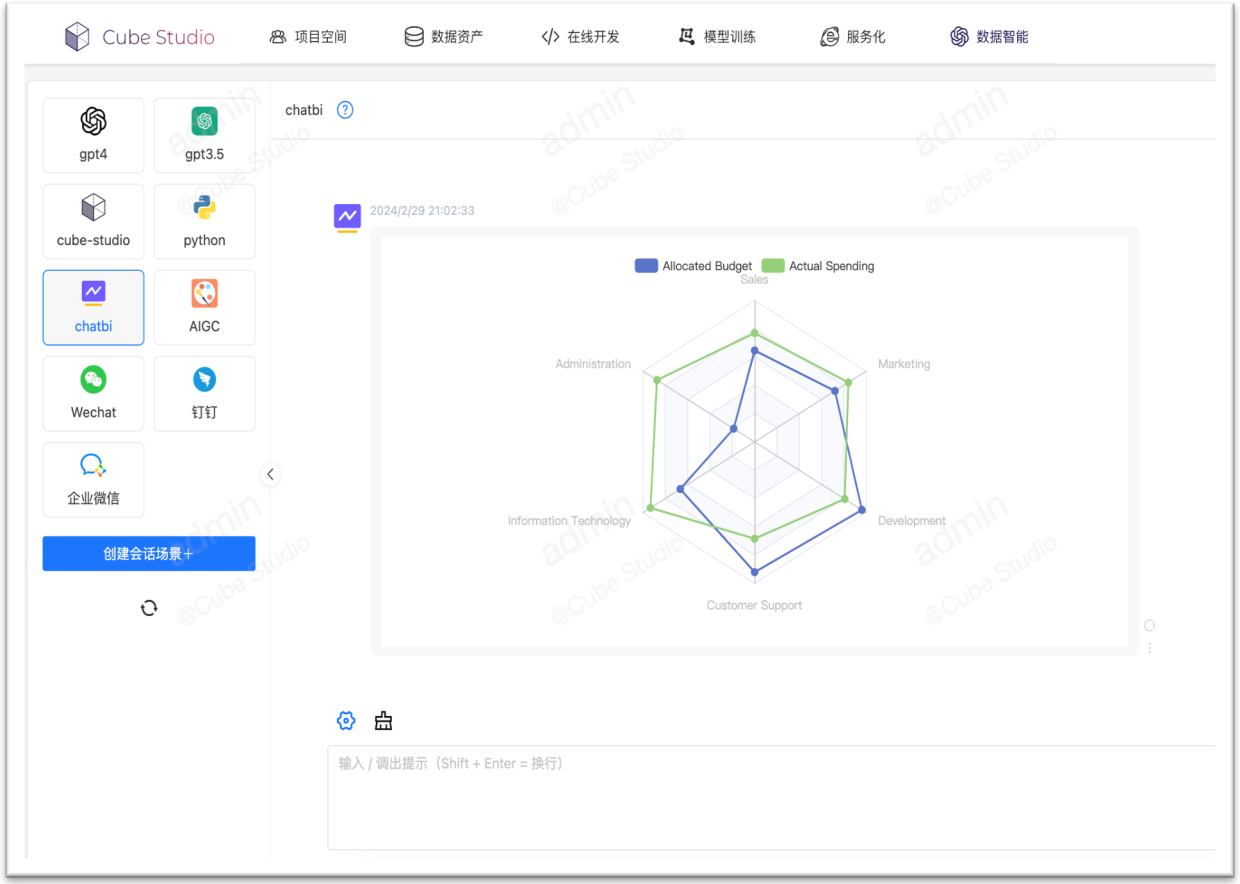
GPT-私有知识库

- ✓ 数据智能模块可配置专业领域智能对话，快速敏捷使用llm
- ✓ 可为某个聊天场景配置私有知识库文件，支持主题分割，语义embedding，意图识别，概要提取，多路召回，排序，多种功能融合

GPT-AIGC与chatbi

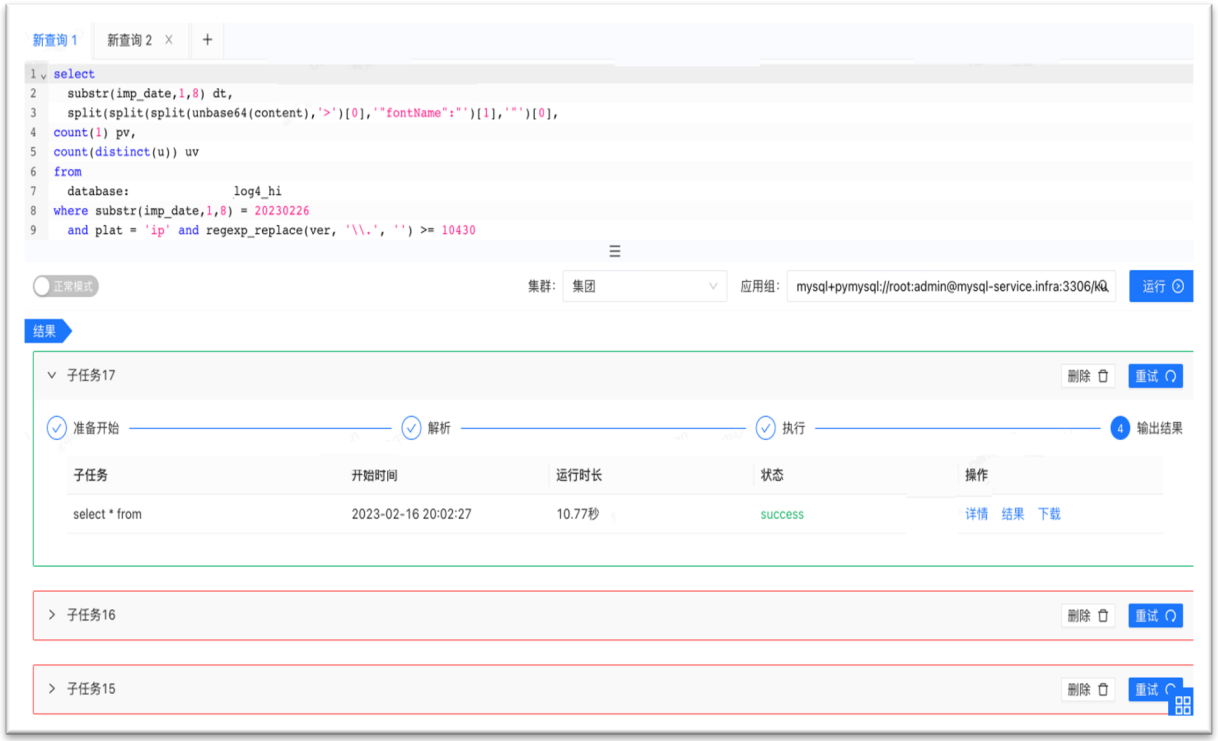
- ✓ AIGC：可以将智能会话与AIHub相结合，例如下面AIGC模型与聊天会话
- ✓ chatBI：智能问询，看板形式展示数据结果
- ✓ 机器人：智能会话与微信公众号/企业微信/钉钉机器人打通，可在机器人中使用知识库



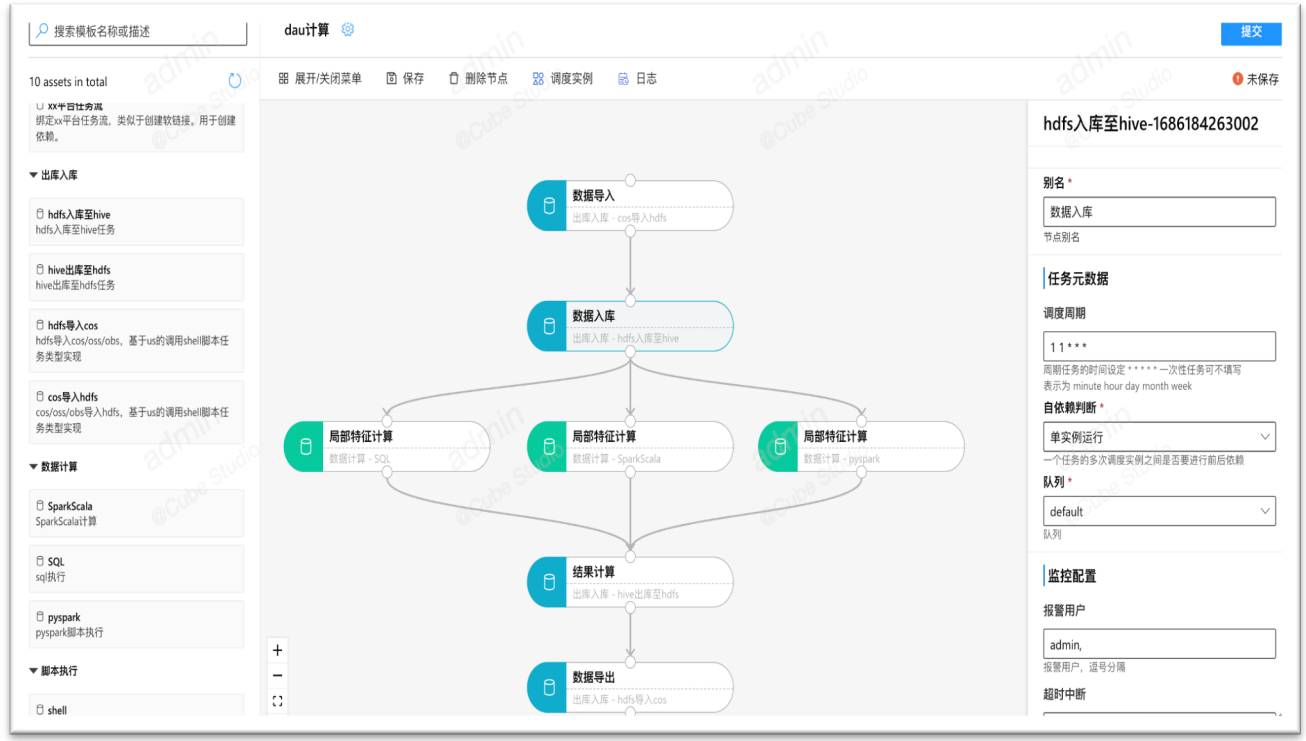


数据中台对接

为了加速AI算法平台的使用，cube-studio支持对接公司原有数据中台，包括数据计算引擎sqllab，元数据管理，指标管理，维表管理，数据ETL，数据集管理



数据资产开发工具



数据ETL可视化链路

三种部署模式灵活选择

针对企业和家庭的用户需求，根据不同场景对**计算实时性**的不同需求，可以提供三种建设模式

- **模式一：私有化部署——对数据安全要求高、预算充足、自己有开发能力**
- **模式二：边缘集群部署——算力分散，多个子网环境的场景，或边缘设备场景**
- **模式三：serverless集群——成本有限，按需申请算力的场景**

实时性要求低或管理性需求场景（公有云服务）

公有云算力服务

明厨亮灶 | 摄像头巡检 | 文旅景区 | 应急灾害

实时性要求较高场景（私有化部署）

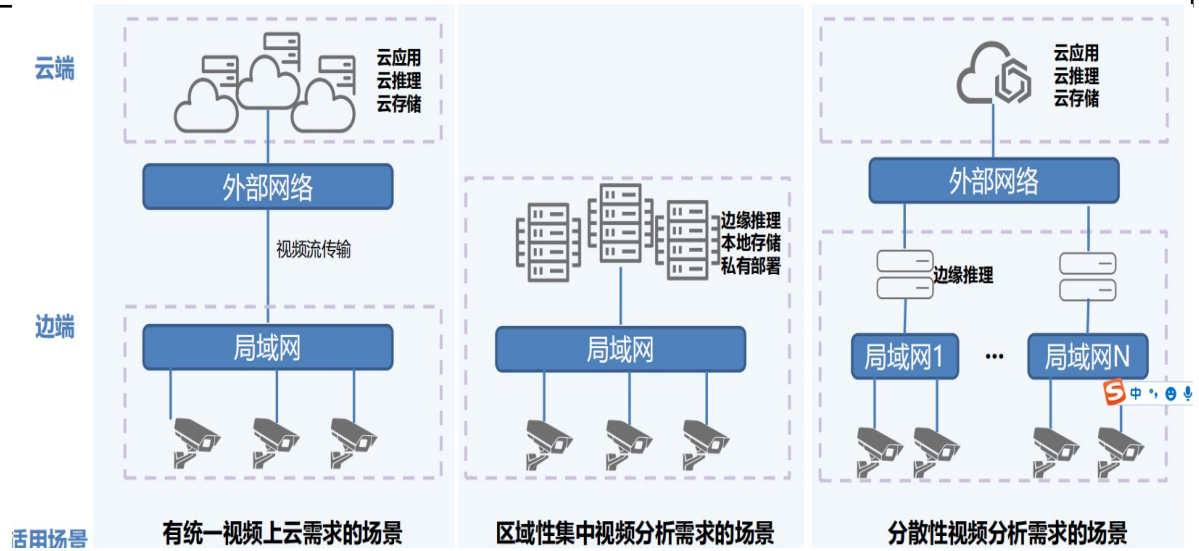
私有化部署服务

考勤管理 | 客流统计 | 人员布控 | 车辆布控

准实时流场景（混合部署）

混合边缘云部署服务

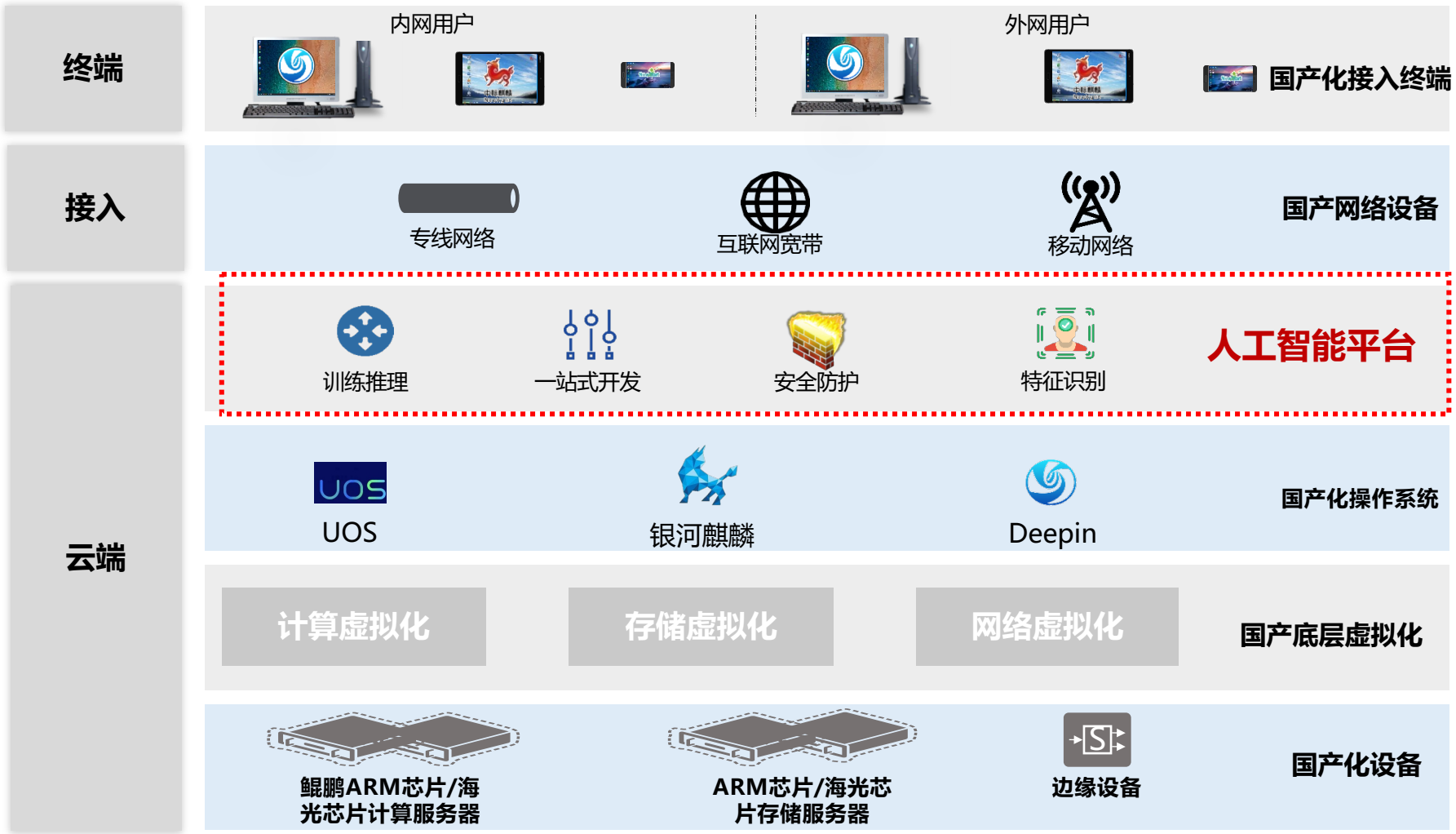
区域雪亮 | 智慧社区 | 火焰识别 | 行为动作



■ 三种模式对比

类型	模式一	模式二	模式三
计算实时性	低	中	高
部署难度	低	中	高
成本	低	中	低
抽帧解码能力平台	无需要	需要	需要
AI计算资源	无需	需要	需要
带 宽	需要	需要	需要

国产化部署支持



产品在国产化链路中的定位

平台原生支持国产化平台部署，以满足客户对国产化产品安全可信要求！