



应用托管平台 用户操作手册

天翼云科技有限公司

一、产品介绍	5
1.产品定义	5
2.产品优势	5
3.功能特性	6
4. 应用场景	7
5.使用限制	10
二、 计费说明	11
1. 计费概述	11
2. 计费项	14
3. 欠费说明	15
三、快速入门	16
1.准备工作	16
1.1 注册天翼云账号	17
1.2 实名认证	17
1.3 账户充值	17
2.购买应用	17
订购应用市场应用	17
订购 MCP 应用	19
自定义部署应用（公共应用资源）	20
四、用户指南	23
1.应用市场	23
2.MCP 市场	25

3.应用实例	27
3.1 创建公共资源应用实例	27
3.2 创建独享资源应用实例	30
3.3 查看应用实例	32
3.4 访问服务	40
3.5 启动服务	41
3.6 停止服务	43
3.7 更新服务	45
3.8 扩缩容	51
3.9 弹性伸缩	53
3.10 增减组件	57
3.11 删除服务	62
4.应用访问策略	62
4.1 创建访问策略	63
4.2 访问策略管理	67
5.项目空间	70
5.1 创建项目空间	70
5.2 项目空间管理	73
6.存储管理	83
6.1 创建存储	83
6.2 存储生命周期管理	85
6.3 文件上传工具	93

7.应用门户	98
8.资源组管理	101
8.1 查看资源组	101
8.2 管理资源组	101
五、 常见问题	101
1. 计费类	101
2. 操作类	102
3. 技术运维类	104
3.1 应用部署常见问题	104
3.2 调试和排查常见问题的方法	105
六、 最佳实践	109
1.基于 Dify 和 DeepSeek 构建本地知识库的最佳实践	110
2.通过应用市场部署 Qwen3 推理 api 服务	125
3.在 Dify 中集成 MCP Server	137
4. 云主机通过内网访问应用实例	146
5. 应用实例通过内网访问数据库等云资源	151
6.使用容器镜像服务部署应用	155

一、产品介绍

1.产品定义

一站式应用托管平台，免运维底层资源，提供应用构建、部署、交付和运维全生命周期管理，构建息壤应用生态体系。



2.产品优势

开箱即用

提供应用一键部署的功能，极大简化应用的安装和配置过程，分钟级完成应用部署；为部署的应用提供公网访问能力，可通过公网便捷访问已部署的应用。用户无需学习复杂的 K8s 等容器知识，即可轻松上手。

丰富的生态与扩展



汇聚不同类型的应用，覆盖多种用途和行业需求。可方便地浏览、搜索选择适合的应用；集成 MCP 相关的服务资源，涵盖内容生成、网页搜索、效率工具等多种类型，全面支撑企业各类智能应用场景。

降低 IT 资源投入

支持根据业务需求灵活增减配资源，避免重资产投入，高效管控成本；按小时计费，暂停不计费。

少人化运维与成本优化

基于云原生和部署工具链的集成，大幅减少人工操作，少量运维人员即可管理大量应用系统，提升维护效率，降低成本；平台管理并分配算力，屏蔽算力底座差异，用户无需关心底层硬件和 K8s 集群，聚焦业务应用开发和管理。

应用交付效率提升

为供应商提供独立的测试空间，灵活分配计算资源避免干扰，助力供应商快速完成部署与测试，提升协作效率；完成测试并审核通过的应用可生成“应用模板”，并可发布应用市场，实现用户环境秒级部署上线，大幅降低运维成本。

3.功能特性

应用全生命周期管理

提供从构建、交付、管理到运维的全生命周期应用管理能力，支持多种方式灵活构建应用、通过模板标准化交付、一站式管理应用访问与资源，提供持续运维能力。

高效便捷的应用市场和 MCP 市场

提供丰富多样的应用资源库，涵盖开源工具、模型算法、AI 应用、MCP 服务等多种类型，满足不同业务场景下的功能扩展与集成需求。支持一键部署，简化传统应用安装与配置的复杂流程，显著提升部署效率。

持久化应用存储

为应用提供持久化存储，支持多种存储类型和性能规格，提供多途径文件接入方式。

灵活资源分配与团队协作

支持完善的项目空间管理能力，实现多项目、多团队的高效协作与资源隔离，灵活配置人员、角色与权限体系，实现对成员访问范围与操作权限的精细化控制。

弹性伸缩

具备弹性伸缩能力，一种根据系统负载情况自动调整资源配置的机制，旨在确保系统在不同的业务场景下都能保持良好的性能和稳定性，同时避免资源的浪费。

4. 应用场景

企业应用构建、部署、发布和运维

适用场景



在企业 IT 管理下，应用托管替代虚拟机/裸金属/k8s 等基础设施，满足企业数字化云原生转型需求，构建企业智能应用

产品优势

低门槛应用全生命周期管理：无需掌握 K8s 底层技术，通过可视化界面可完成应用部署、升级、扩容等全流程操作，提供丰富工具，非专业人员快速上手管理应用

高效部署与智能运维能力：支持高可用、负载均衡与资源扩缩，同时提供公网访问、可观测等能力，可同时管理百套应用

优化运维成本：少量运维人员即可管理大量应用系统，降低人力成本，同时避免操作底层资源，提升运维稳定性与效率

一站式订阅和自动化部署应用

适用场景

行业应用、热门应用等一站式部署和使用

产品优势

一键部署：提供应用一键部署的功能，极大简化应用的安装和配置过程，轻松上手

多样的应用资源：汇聚不同类型的应用，覆盖多种用途和行业需求。快速满足业务需求

便捷访问：为部署的应用提供公网访问能力，可通过自定义域名在公网便捷访问已部署的应用



应用托管+云资源全栈业务场景解决方案

适用场景

客户在应用托管中部署自研应用/第三方模型，与现有弹性云主机、数据库等云资源联动协同，实现完整复杂业务流程。

产品优势

资源协同，性能高效：共池内网打通，应用托管与弹性云主机、数据库、消息队列等云服务实现低延迟、高带宽内网通信，同时内网访问更具安全保障

业务整合成本低：已有公有云用户可直接复用弹性云主机、数据库等云上资源，无需额外迁移 / 重构，快速搭建完整业务架构，降低整合成本

部署灵活，场景覆盖广：支持自定义模型服务、MCP Server、智能体应用、自定义业务应用等多类型部署，结合丰富的云产品，满足业务的全场景需求

Aone 赋能应用与企业内网网络打通

适用场景

当企业同时使用公有云的应用托管服务和企业内网的各类系统时，需要实现两者之间的安全网络连通

产品优势

网络连通性优势：Aone 打通公有云应用与企业内网网络，保障安全稳定，实现公有云应用本地式访问企业内网资源，破解跨网资源访问壁垒

资源整合优势：整合公有云的应用托管能力和企业内网系统资源，实现资源的高效利用和协同工作，提升企业整体

IT 架构的运行效率：安全性优势 在打通网络的同时，保障了企业内网资源的安全访问，避免了因网络开放带来的安全风险，满足企业对数据和系统安全性的高要求

5.使用限制

可用区限制

当前开放西南 1、昆明 2、安顺 4 可用区，其他可用区请持续关注。

配额限制

使用应用托管平台时，需注意配额限制，每个可用区的配额具体如下表所示。如果可用区配额不能满足业务使用，可通过联系客户经理或工单申请调整资源池和配额。

公共资源算力配额

分类	限制数量	单位
CPU	500	核
GPU	50	卡
内存	1000	GB
存储	5000	GB

项目空间算力配额

使用项目空间时，支持为项目空间成员分配算力配额（全部可用区总额），具体如下表所示。

如果默认配额不能满足成员业务使用，可通过项目空间管理员调整配额。

分类	默认限制数量	单位
CPU	500	核
GPU	50	卡

资源售罄

在购买算力资源时，所选的产品规格可能已售罄，如果没有需要的产品规格，请联系客户经理或通过工单反馈，我们会根据资源库存情况，补充对应的产品规格。

二、计费说明

1. 计费概述

包年包月

适用场景

包年包月是一种计费模式，适用于多种场景，尤其是需要稳定资源并长期使用的情况：

- 长期稳定使用：适用于需要长期稳定运行的服务。
- 成本可控和节约：对于需要长期使用的资源，包年包月可以降低长期成本，相比按量付费更为经济。

- **稳定资源需求：**对于有稳定资源需求的业务，包年包月提供稳定的资源保障。

收费方式

一种预付费模式，即先付费再使用。

按月计费，以自然月为计费单位。

续订规则

用户可以续订包月包年且未退订、未释放的资源，延长相关资源的使用时间。

手动续订

控制台费用中心续订管理页面找到所需续订的资源，点击“手动续订”，根据使用需要调整续订周期后，提交续订订单。

自动续订

控制台费用中心续订管理页面找到所需续订的资源，点击“开通自动续订”，根据使用需要调整续订周期后，提交续订订单。

退订规则

客户（天翼云用户）可根据需要，在符合天翼云退订规则的前提下，灵活退订资源。目前退订包含七天无理由全额退订和非七天无理由退订以及其他退订。

资源实例一旦被退订，将无法恢复，请提前导出、备份您的资源，避免造成损失。

实例到期

包年包月资源到期后，将被冻结无法使用，并在 15 天后自动释放，具体以您收到的邮件或短信提示为准。为避免影响您的正常使用，请及时在控制台费用中心续订管理，手动续订或开通自动续订。

按量付费

适用场景

按量付费是一种灵活的计费模式，适用于以下场景：

- 不确定性需求：对于短期项目或存在季节性高峰期的业务按量付费更适合，无需长期绑定资源。
- 需求波动较大：应对突发事件或需求波动，按量付费能够灵活调整资源规模。
- 不稳定业务模式：用于试验性质的项目或业务，需要灵活、短期的资源支持。
- 资源使用不稳定：对资源使用量不稳定，难以预测的业务需求，按量付费可以根据实际使用量灵活付费。
- 资金有限或预算不确定：对于预算有限的小型企业或者业务预算不稳定的场景，按量付费不需要一次性大额支出。

收费方式

一种后付费模式，即先使用再付费。

以小时为出账周期，在每个结算周期生成账单并从账户中扣除相应费用。您可以通过控制台费用中心账单管理页面，了解扣费的详情。

2. 计费项

计费说明

在应用托管全流程的使用中，涉及到计费项主要包括资源费用和软件费用。

- 资源费用：使用应用托管平台的计算、存储和网络资源产生的费用。
- 软件费用：使用应用托管平台的付费应用产生的费用。

资源费用

所有资源价格，以订单为准。价格具体如下：

计费项	按需标准价格（元/核/小时）	包月标准价格（元/核/月）
CPU	0.05	27

计费项	按需标准价格（元/GB/小时）	包月标准价格（元/GB/月）
内存	0.02	9.6

计费项	显卡	显存（GB）	按需标准价格（元/卡/小时）	包月标准价格（元/卡/月）
GPU	A10	24	3.33	1620

计费项	类型	按需标准价格 (元/GB/小时)	包月标准价格 (元/GB/月)
存储	基础性能版	0.000573	0.3
	专业入门版	0.0009	0.4
	企业版	0.0017	1.2

说明：按量付费（按需计费）的存储在开通之后即开始按小时计费，请保证余额充足，防止欠费导致资源冻结。

计费项	类型	按需标准价格 (元/GB)
网络	流量	0.2

软件费用

计费项	包年标准价格 (元/年)
免费应用	0
付费应用	按年收取费用，详见订购页

非应用托管费用

在应用托管平台授权使用天翼云其他产品，如弹性云主机、息壤模型推理服务、息壤智能体、智能边缘云、云日志、VPC 终端节点等，产生的费用由对应产品收取。

3. 欠费说明

欠费原因

购买了按需计费的资源实例，账户余额不足以抵扣实例结算费用时，会造成余额不足，导致欠费。

欠费停服说明

- 用户欠费后，占用的计算资源将进入冻结状态。
- 进入冻结状态时，服务会自动停止，无法在平台新建按量付费资源，也无法启动已有按需资源。
- 如果在 15 天内充值补足欠款，资源将自动解冻。
- 欠费超过 15 天后，将视为您主动放弃该服务，计费资源将会被释放，具体时间以短信及邮件通知为准，资源释放后将不可恢复。如因欠费资源释放导致的业务影响请您自行负责。
- 如用户账户出现欠费，账户一旦充值，系统将会自动优先扣除欠费金额。

续费说明

- 请在欠费 15 天内，完成账号充值补足欠款，保证账号中的余额充足，资源实例将自动解冻。
- 欠费仅影响按量付费资源的使用，对已购买的包年包月资源无影响。

三、快速入门

1.准备工作

1.1 注册天翼云账号

如果您已经拥有一个天翼云实名账号可直接跳过本步骤，如果您还没有天翼云账号，具体可参见[注册天翼云账号](#)。

1.2 实名认证

具体可参见[实名认证](#)模块对天翼云账号进行实名认证。

1.3 账户充值

您需要确保天翼云实名账户中的余额充足，具体天翼云账户充值步骤请参见[账户充值](#)。

2.购买应用

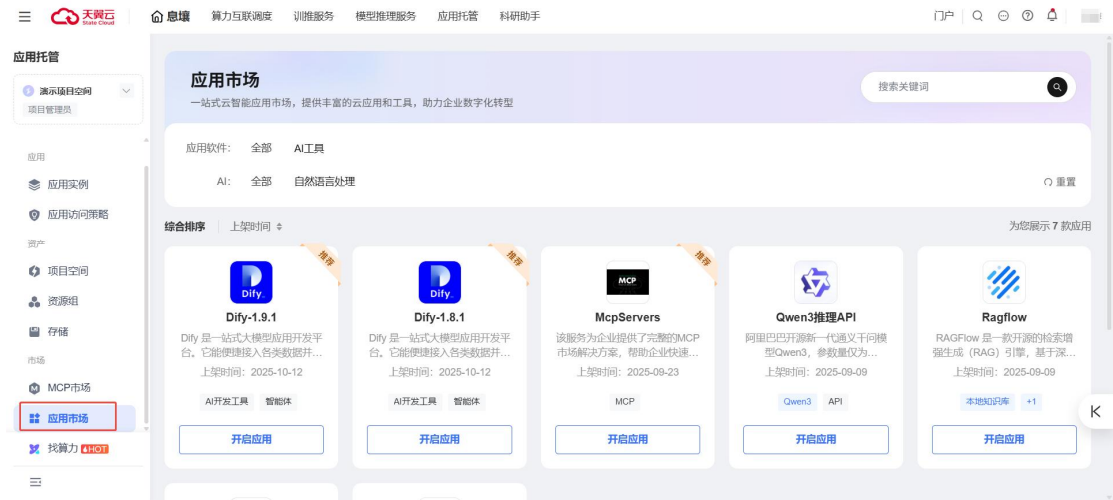
应用托管提供两种部署应用的方式。

订购应用市场应用

应用市场提供优质的应用资源，支持通过关键词搜索、分类筛选等方式快速定位所需应用，一键部署。

操作步骤

1.进入天翼云应用托管平台，点击左侧【应用市场】。



2.选择应用点击对应的【开启应用】按钮，进入应用详情页。



3.填写应用所需要的参数，选择访问策略和公网访问配置，并勾选协议后，点击【开启应用】按钮，提交应用订购。

4.应用购买成功后，可在控制台【应用实例】，查看当前已部署的应用实例。

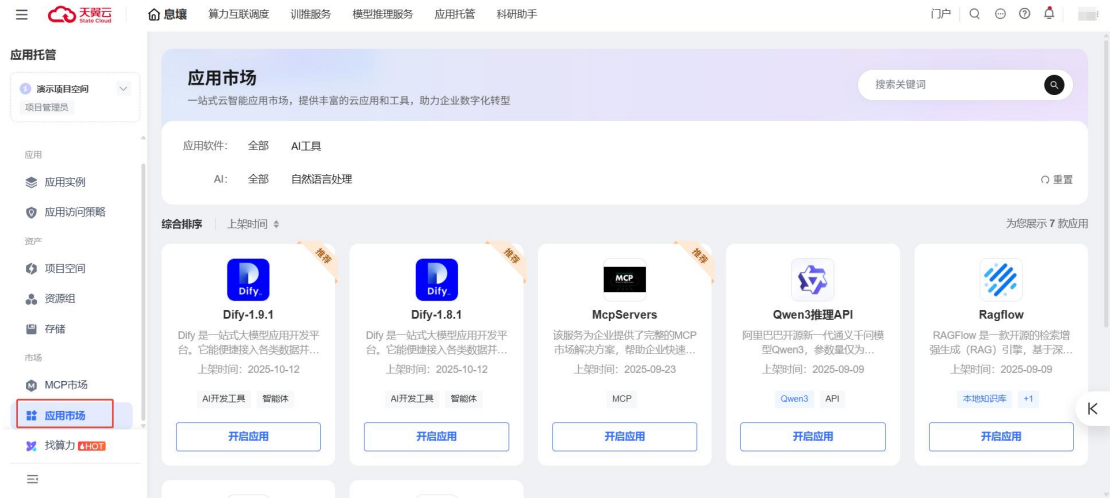
公共应用资源		专属应用资源				
+ 创建应用实例		一键快速部署？去应用市场				
实例名称/ID	状态	应用信息	付费类型	创建时间	更新时间	操作
MCP-Server-941163 ID:d4a3hihgkt1qmi77m05g	运行中	MCP-Server app-d2h9j71gk1i5348prb0	按量付费	2025-11-12 15:35:52	2025-11-12 16:59:23	访问 启动 停止 更多
MCP-Server-256285 ID:d4a43l1gk1i77m4l0	运行中	MCP-Server app-d2h9j71gk1i5348prb0	按量付费	2025-11-12 16:14:19	2025-11-12 16:16:41	访问 启动 停止 更多
121324-915473 ID:d49vs6ld3icvf6jp3lko	删除中	121324 app-d104b5td3ickrq2kkl2g	按量付费	2025-11-12 11:25:17	2025-11-12 12:19:29	访问 启动 停止 更多
lss-anxuanxu1101 ID:d49f0ehb9tnradots0	运行中	自定义应用	按量付费	2025-11-11 16:51:20	2025-11-12 12:03:55	访问 启动 停止 更多

订购 MCP 应用

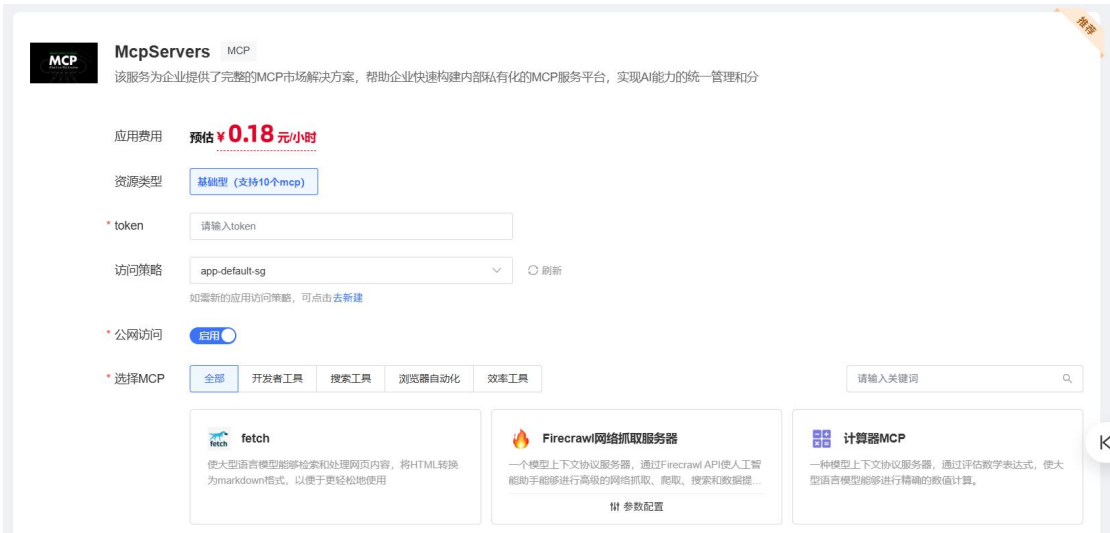
MCP 市场聚合优质 MCP 应用资源，赋能 AI 高效调用多元服务。

操作步骤

1.进入天翼云应用托管平台，点击左侧【应用市场】。



2.选择 “McpServers” 应用，点击对应的【开启应用】按钮，进入应用详情页。



3.填写 “McpServers” 应用所需要的参数。

字段	说明
资源类型	可选项“基础型”，一个基础型的 McpServs 支持运行 10 个 MCP 服务。
token	填写调用 MCP 服务所需要的 token。
访问策略	选择应用访问策略，如没有合适策略可新建，默认使用系统策略。
公网访问	选择是否启用公网访问，默认启用。
选择 MCP	选择需要部署的 MCP 服务，最多选择 10 个。

- 4.勾选协议后，点击【部署应用】按钮，提交应用订购。
- 5.应用购买成功后，可在控制台【应用实例】，查看当前已部署的应用实例。

公共应用资源 专属应用资源

+ 创建应用实例

一键快速部署? 去应用市场

实例名称

搜索关键词

Q

刷新

实例名称/ID	状态	应用信息	付费类型	创建时间	更新时间	操作
MCP-Server-941163 ID:d4a3hihgkt1qmi77m05g	运行中	MCP-Server app-d2h9j71gk1i5348prb0	按量付费	2025-11-12 15:35:52	2025-11-12 16:59:23	访问 启动 停止 更多
MCP-Server-256285 ID:d4a43l1gkt1qmi77m4l0	运行中	MCP-Server app-d2h9j71gk1i5348prb0	按量付费	2025-11-12 16:14:19	2025-11-12 16:16:41	访问 启动 停止 更多
121324-916473 ID:d49vse6ld3lcvf6jp3lko	删除中	121324 app-d104b5d3lckrqa2kk2g	按量付费	2025-11-12 11:25:17	2025-11-12 12:19:29	访问 启动 停止 更多
lss-anxuanxu1101 ID:d49f0ehb9tnradotss0	运行中	自定义应用	按量付费	2025-11-11 16:51:20	2025-11-12 12:03:55	访问 启动 停止 更多

自定义部署应用（公共应用资源）

通过自定义部署参数进行应用的部署，适用于管理与运维自己的应用服务。

操作步骤

- 1.登录应用托管控制台，在左侧导航栏单击【应用实例】，在公共应用资源 Tab 点击【创建应用实例】按钮。

公共应用资源 专属应用资源

+ 创建应用实例

一键快速部署? 去应用市场

实例名称

搜索关键词

实例名称ID	状态	应用信息	付费类型	创建时间	更新时间	操作
MCP-Server-941163 ID:d4a3hihgkt1qmi77m05g	运行中	MCP-Server app-d2h9/71gk1i5348prb0	按量付费	2025-11-12 15:35:52	2025-11-12 16:59:23	访问 启动 停止 更多
MCP-Server-256285 ID:d4a43l1gk1qmi77m4l0	运行中	MCP-Server app-d2h9/71gk1i5348prb0	按量付费	2025-11-12 16:14:19	2025-11-12 16:16:41	访问 启动 停止 更多
121324-915473 ID:d49vvs6ld3lcvf6jp3lk0	删除中	121324 app-d104b5td3lckrqa2kk2g	按量付费	2025-11-12 11:25:17	2025-11-12 12:19:29	访问 启动 停止 更多
lss-anxuanxu1101 ID:d49f0iehb9tnradotss0	运行中	自定义应用	按量付费	2025-11-11 16:51:20	2025-11-12 12:03:55	访问 启动 停止 更多

2.应用配置：选择自定义应用，输入实例名称。

* 应用名称

请输入内容

名称为4-32字符，支持大小写字母、数字、_、-

3.应用配置：填写镜像地址（私有镜像仓库需配置访问凭证），选择付费类型，选择可用区、算力规格和实例数量。

应用配置

* 镜像地址

请输入镜像仓库地址/镜像名/版本

☐ 私有化镜像仓库授权

* 计费类型

按量付费 包年包月

* 可用区

可用区1 可用区2 可用区3 可用区4

* 算力规格

CPU: 2核 | 内存: 4GB

资源充足

* 数量

- 1 +

4.网络配置

字段	说明
网络	中文名称：网络的中文名称
	英文名称：网络的英文名称
	容器端口：容器的监听端口
变量	url 地址：应用访问的路径，共享网络必填
	对外服务：对否开启对外服务，开启后将为服务分配公网 IP

字段	说明
访问域名	应用的访问域名，独享网络选填，共享网络必填。
访问策略	访问策略为应用的公网访问提供安全防护，可根据需要调整访问策略内容。默认引用系统策略，一个应用只能绑定一个访问策略。

网络配置

* 网络变量

共享网络

中文名称	* 英文名称	* 容器端口	* url地址	对外服务	操作
<input type="text" value="请输入"/>	<input type="text" value="请输入"/>	<input type="text" value="请输入"/>	<input type="text" value="请输入"/>	<input checked="" type="checkbox"/> 关闭	删除
添加网络变量					

* 访问域名

* 访问策略

app-default-sg-1(白名单)

刷新

如需新的应用访问策略, 可点击 [去创建](#)

说明：使用 80、8080、443、8443 敏感调用端口时，需先备案，再联系运营人员启用端口。仅独享网络支持开放敏感端口。

5.高级配置：

参数名称	参数说明
运行命令	配置服务的启动命令
环境变量	配置服务的环境变量
配置文件	配置服务的配置文件
存储	挂载持久化存储

高级配置

运行命令

请输入Shell启动命令
例如: /webui.sh --listen --port 8000 --skip-version-check --no-hashing --no-download-sd-model --skip-prepare-environment --api --filebrowser

环境变量

支持批量输入, 格式为key=value, 多变量用英文分号分隔。点击「内容识别」或输入框外任意位置, 系统将自动解析并填充环境变量
例如: mongoUrl=127.0.0.1:8000,redisUrl=127.0.0.1:8001,env1=test

内容识别

添加环境变量

配置文件

文件路径	操作
暂无数据	
添加配置文件	

存储

存储	选择文件路径	挂载路径	操作
创建新存储			
请选择存储	+ 请选择	例如 /tmp/username	删除
添加存储			

6.完成以上配置，确认部署应用的资源配置费用，并勾选协议点击【部署应用】后，完成实例的创建。

四、用户指南

1.应用市场

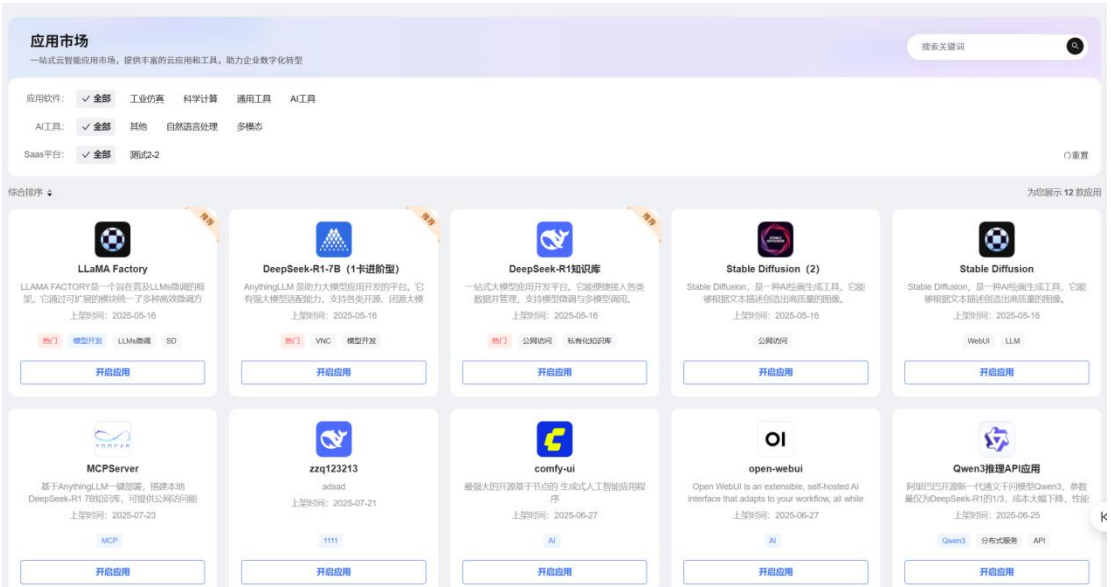
应用是指软件应用程序，例如开源软件工具、工业仿真软件、游戏软件等。

应用托管平台平台提供全栈式应用托管方案，涵盖应用构建与部署、模型算法、SaaS 应用、业务调度、立体化监控等应用生命周期管理，基于异构算力统一调度引擎，实现对算力芯片，算力架构无差异化的应用部署，开箱即用。

操作步骤：



1.登录应用托管控制台，在左侧导航栏单击【应用市场】。

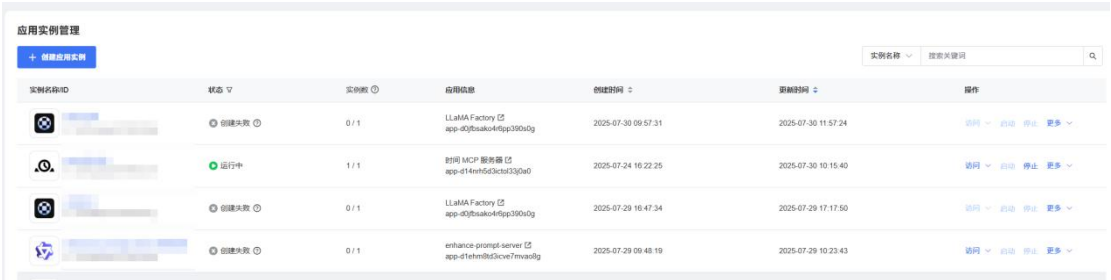


2.选择应用点击对应的【开启应用】按钮，进入应用详情页。







3. 选择访问策略，填入需要的参数，选择是否公网访问，勾选协议，点击【开启应用】按钮。

4.应用购买成功后，可在【应用】->【应用实例】， 查看已购买的应用实例。



5.待应用处于运行中状态时，可点击【访问】操作，点击访问地址进行应用使用。

应用实例管理						
+ 新建应用实例						
实例名称ID	状态	实例数	应用信息	创建时间	更新时间	操作
 [实例名称]	创建失败	0 / 1	LLaMA Factory 应用 app-d59ba0e4f99c390a0g	2025-07-30 09:57:31	2025-07-30 11:57:24	访问 启动 停止 更多
 [实例名称]	运行中	1 / 1	增强 MCP 服务应用 app-d14e1b5d3c0a7f9a0g	2025-07-24 10:22:25	2025-07-30 10:15:40	访问 启动 停止 更多
 [实例名称]	创建失败	0 / 1	LLaMA Factory 应用 app-d59ba0e4f99c390a0g	2025-07-29 16:47:34	2025-07-29 17:17:50	访问 启动 停止 更多
 [实例名称]	创建失败	0 / 1	enhance prompt server 应用 app-d14e1b5d3c0a7f9a0g	2025-07-29 09:48:19	2025-07-29 10:23:43	访问 启动 停止 更多
mcp-cal ID: d15a835d3c0a7f9a0g	已停止	0 / 1	mcp-计算应用 app-d15a835d3c0a7f9a0g	2025-06-25 17:46:52	2025-07-23 17:58:01	mcp-enhance-prompt 访问 启动 停止 更多
oeshi ID: d15a835d3c0a7f9a0g	已停止	0 / 1	数据应用 app-d0a7b0d0b0f9a0g	2025-07-09 15:17:35	2025-07-10 08:09:07	访问 启动 停止 更多

2.MCP 市场

MCP 是指 Model Calling Platform 相关的服务资源，涵盖内容生成、网页搜索、效率工具等多种类型，全面支撑企业各类智能应用场景。

操作步骤：

1.登录应用托管控制台，点击【MCP 市场】。


MCP市场

一站式MCP应用市场，聚合优质MCP应用资源，赋能AI应用开发多元服务

MCP类型: 全部 内容生成

绿色排序


为应用 13 款应用

 中文趋势聚合

基于 Model Calling Protocol (MCP) 协议的全网趋势聚合一站式聚合服务

上架时间: 2025-06-16


访问 部署

 Wikipedia MCP 服务端

这是一个 Model Calling Protocol (MCP) 服务端，可以连接维基百科知识库，为大型语言模型 (LLM) 提供上下文。该工具帮助 AI 助手从维基百科知识库中获取信息，可以

上架时间: 2025-06-11


访问 部署

 mcp-server-omdb

该MCP服务端与The Movie Database (TMDB) API集成，提供电影信息、搜索功能

上架时间: 2025-06-26


访问 MCP 部署

 bing-cn-mcp-server

一个基于 MCP (Model Calling Protocol) 的中文必应搜索工具，可以通过 Bing 搜索引擎 (Bing) 的 AI 搜索接口进行实时信息查询。

上架时间: 2025-06-11


访问 部署

 fetch-mcp

fetch-mcp 是基于 Anytrig LLM 一键部署，通过调用 DeepSeek-R1 72B 模型，可

上架时间: 2025-05-29


访问 部署

 enhance-prompt-server

EnhancePrompt 是一个基于人工智能模型的应用程序，旨在帮助用户为各种生成式 AI 模型生成高质量的提示词。它提供了一个简单易用的接口，让用户可以快速生成

上架时间: 2025-06-26


访问 部署

 headline-vibes

一个基于上下文生成新闻标题，用于分析新闻主要内容和生成标题的

上架时间: 2025-06-26


访问 MCP 部署

 mcp-github-trending

一个通过调用 GitHub API 提供 GitHub 仓库和开发者的 MCP 服务端。

上架时间: 2025-06-25


访问 MCP 部署

 数据查询MCP服务

一个基于上下文生成 (MCP) 服务端，通过调用 Alpha Vantage API 提供实时财经

上架时间: 2025-06-13


访问 部署

 实时 MCP 服务端

一个提供实时数据和实时生成内容的上下文生成服务端。该服务端通过 LLM 模型

上架时间: 2025-06-11


访问 部署

 mcp-levily

The Model Calling Protocol (MCP) is an open standard that enables AI systems to

上架时间: 2025-06-11

访问 MCP 部署

 mcp-airbnb

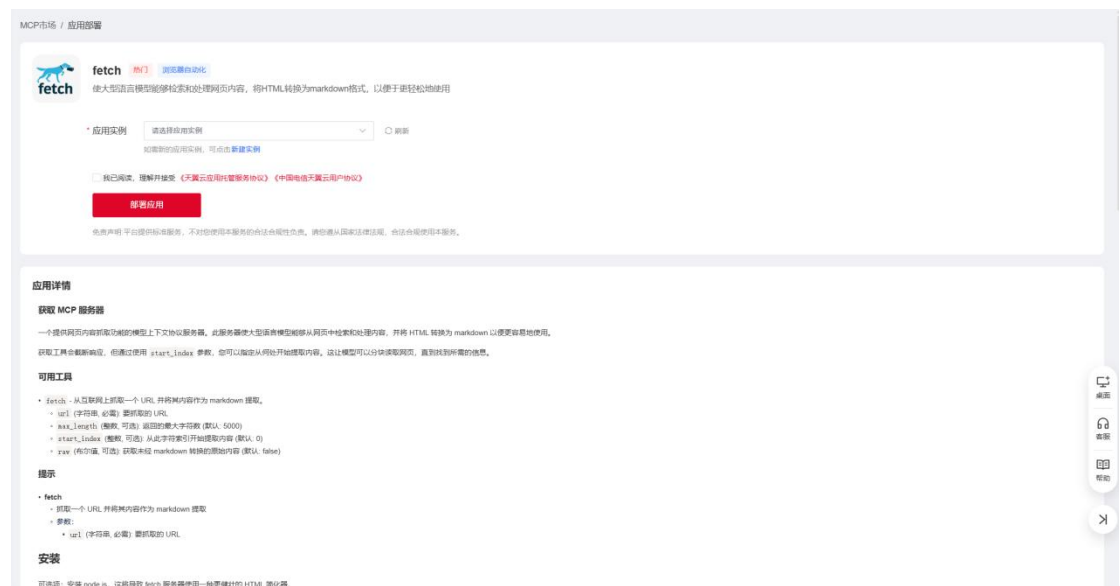
用于搜索 Airbnb 并获取详细信息的 MCP 服务端。

上架时间: 2025-06-11

访问 MCP 部署

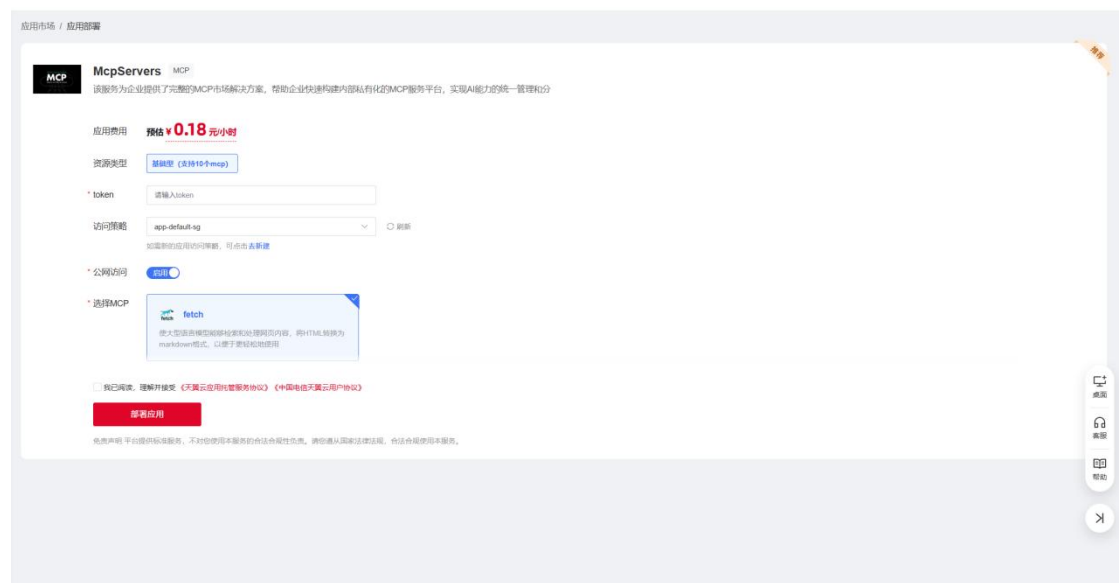


2.选择应用点击对应的【立即部署】按钮，进入应用详情页。



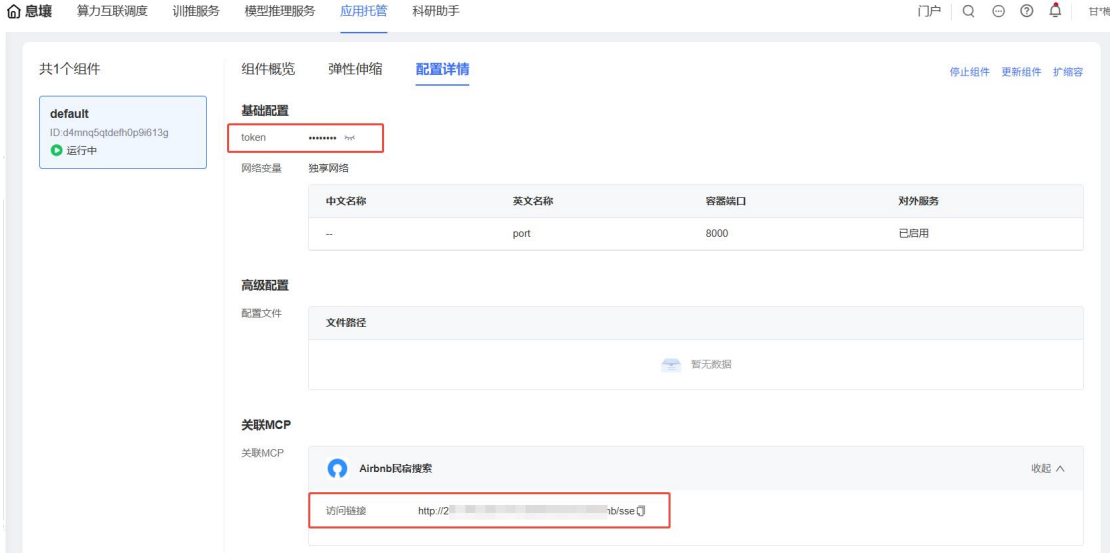
3.选择已经部署的 MCP Servers 应用实例，填写必要的参数，勾选协议，点击【立即部署】按钮。

4.若无可用 MCP Servers 应用实例，可以点击“新建实例”，创建 MCP Servers 实例并部署选择的 MCP 服务。



5.在创建 MCP Servers 应用实例页面填写必要的参数，勾选协议，点击【部署应用】按钮。

6.应用实例创建完成后，进入应用实例详情页，点击配置详情 TAB，可以看到对应 MCP 的访问 token 和 URL 地址。



3.应用实例

3.1 创建公共资源应用实例

3.1.1 创建应用/MCP 市场应用实例

操作步骤

1. 登录应用托管控制台，在左侧导航栏单击【应用实例】， 点击【去应用市场】按钮。



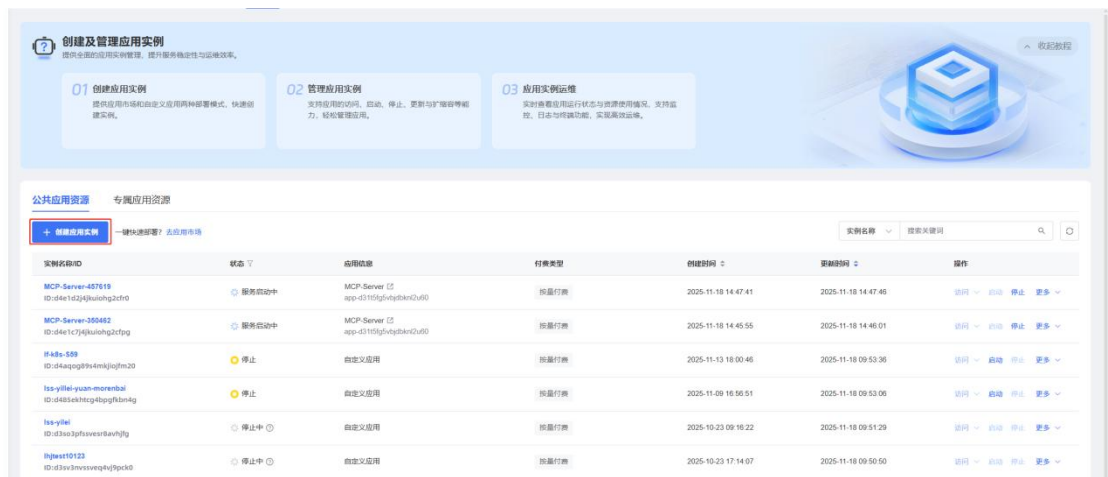
2.进入应用市场后，选择应用市场应用进行部署。

3.1.2 创建自定义应用实例自定义部署应用

通过自定义部署参数进行应用的部署，适用于管理与运维自己的应用服务。

操作步骤

1. 登录应用托管控制台，在左侧导航栏单击【应用实例】， 单击【创建应用实例】按钮。



2. 输入实例名称。

* 应用名称

名称为4-32字符，支持大小写字母、数字、_、-、.

3. 应用配置配置：填入镜像地址，付费类型当前支持按量付费，选择可用区，算力规格，数量。

应用配置

* 镜像地址

☐ 私有化镜像仓库授权

* 计费类型

按量付费

包年包月

* 可用区

一类可用区-chungang-1-A10

中卫- NVIDIA T4

内蒙呼和浩特5-ek-准生产-分布式资源- NVIDIA T4

* 算力规格

CPU 1核 | 内存 2GB | GPU 1* A10

选择实例

* 数量

-

1

+

4. 网络配置：配置容器端口及是否开启对外服务。访问策略为服务的公网访问提供安全防护，可根据需要调整访问策略内容。一个应用只能绑定一个访问策略。在共享网络模式下，还需要输入访问域名，应用实例通过域名进行访问，同时可配置是否需要开启 websocket。
- a. 中文名称：用于直观展示端口用途，在应用实例列表中可下拉展示。
 - b. 英文名称：标识端口。
 - c. 容器端口：应用组件需要暴露的服务端口，系统会根据网络类型不同自动分配端口。
 - d. url：指应用组件对应端口访问的 url 表达式，支持正则表达式。

网络配置

网络配置

共享网络

中文名称

英文名称

容器端口

url地址

对外服务

操作

请输入

请输入

请输入

请输入

开启

新增

添加网络变量

访问域名

请输入访问域名

websocket

关闭

访问策略

app-default-sg(默认)

删除

如需新的应用访问策略，可点击 去创建

5. 高级配置：

参数名称	参数说明
运行命令	配置服务的运行命令
环境变量	配置服务的环境变量
存储	挂载持久化存储
配置文件	将文件内容挂载到容器中指定文件

高级配置

运行命令

请输入Shell启动命令
例如: heroku sh --listen -port 8000 --skip-version-check --no-hashing --no-download-sd-model --skip-prepare-environment --api --filebrowser

环境变量

支持批量输入，格式为key=value，多变量用英文分号分隔。点击「内容识别」或输入框外任意位置，系统将自动解析并填充环境变量
例如: mongoUrl=mongodb://127.0.0.1:27010;redisUrl=mongodb://127.0.0.1:6379;env1=test

内容识别

添加环境变量

配置文件

文件路径	操作
	删除
	删除

暂无数据

添加配置文件

存储

存储	创建新存储	选择文件路径	挂载路径	操作
请选择存储	▼	+ 请选择	例如: /tmp/username	删除

添加存储

6.完成以上配置，确认部署应用的资源配置费用，并勾选协议点击【部署应用】后，完成实例的创建。

3.2 创建独享资源应用实例

3.2.1 创建自定义应用实例

操作步骤

1. 登录应用托管控制台，在左侧导航栏单击【应用实例】， 点击【创建应用实例】按钮。

应用实例管理

创建应用实例

实例名称 ▼ | 搜索关键字

实例名称	状态	实例ID	应用配置	创建时间	更新时间	操作
mcp-123 id: 2f1a23d3c0a1b0e0d	已停止	9 / 1	mcp-计算节点 app-4f19c3d5d3d3e9d54g	2025-06-26 17:46:52	2025-07-23 15:35:55	详情 删除 停止 重启

2. 应用配置：选择自定义应用，输入实例名称。

应用配置

* 应用类型

应用软件

AI工具

Saas平台

MCP

自定义应用

* 实例名称

请输入内容

长度为4-53字符，只能包含小写字母、数字及分隔符，分隔符不能连续出现，且需以小写字母开头，数字或小写字母结尾

3.资源配置：资源类型选择【独享资源】，选择资源组进行应用部署。

资源配置

* 资源类型 ☐ 公共资源 ☒ 独享资源

* 选择资源组

4.参数配置：

参数名称	参数说明
Helm chart 地址	指定 Helm Chart 的存储位置。如果您的 Helm Chart 存放在私有仓库中，需要进行授权设置
命名空间	制定应用部署的命名空间
chart values	自定义 Helm Chart 的配置参数, 根据实际需求调整各个参数, 如镜像版本、副本数量、服务类型等
超时等待时间	定义 Helm 部署过程中拉取 Chart 的最大等待时间

参数配置

* Helm chart 地址

☐ 私有仓库授权

* 命名空间

chart values

说明: 如需要修改配置项, 本界面的值将自动同步

* 超时等待时间 秒

Helm chart 说明

[简介](#)

1. templates 模板文件库每个一个或模板必须包含 metadata, metadata.labels, metadata.name 等

2. name 模板文件库的 metadata.labels 必须包含: include "子资源 label" / include 4 5. 模板必须包含 3 个资源

3. metadata.labels 必须包含: app.kubernetes.io/instance 和, 值为 release 名称 (即: Release Name)

4. metadata.name 必须包含: release 名称 (即: Release Name)

5. 模板必须包含: deployment, statefulset 等, 需要在 spec.template.metadata 中包含: app.kubernetes.io/instance 和

5.完成以上配置，并勾选协议点击【部署应用】后，完成实例的创建。

3.3 查看应用实例

3.3.1 查看实例概览

1. 登录应用托管控制台，在左侧导航栏单击【应用实例】，点击应用实例名称进入详情页。



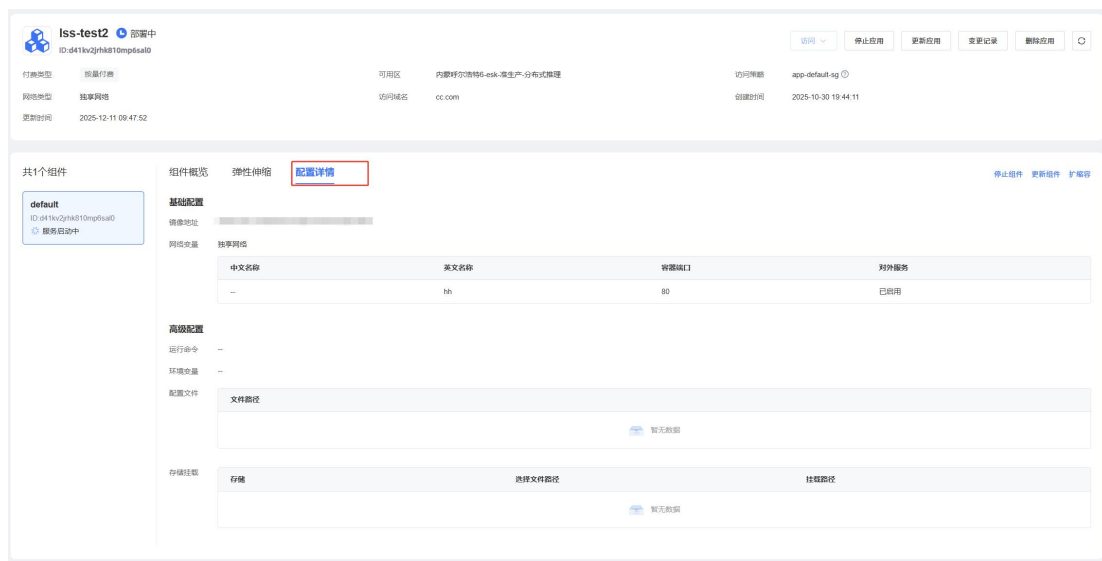
2. 支持查看部署的应用信息、组件信息、组件配置。



3.3.2 查看配置

1. 登录应用托管控制台，在左侧导航栏单击【应用实例】，点击应用实例名称进入详情页。

2. 选择组件，点击【配置详情】，可查看该组件实例的部署配置信息。



1个组件

mcp-test

ID: d485g13go050yfbocag

服务启动中

组件概览

弹性伸缩

配置详情

停止组件

更新组件

扩容组件

基础配置

网络配置

实例列表

中文名称	英文名称	资源端口	对外服务
11eo	aa	8000	已启用

高级配置

配置文件

文件路径

文件路径

暂无数据

关联MCP

关联MCP

Fetch

收起

访问链接

API_ACCESS_TOKEN

END

4.通过应用实例自定义部署的应用，展示部署时完整的部署参数。通过公共资源部署的应用展示镜像地址、运行命令、环境变量、网络配置等参数。

lss-test2

部署中

ID: d418v2jrh810mp6salo

访问

停止应用

更新应用

变更记录

删除应用

应用类型: 应用模板

可用区: 内网华东3-eks-资源生产-分布式推理

访问策略: app-default-eg

网络类型: 独享网络

访问域名: cc.com

创建时间: 2025-10-30 19:44:11

更新时间: 2025-12-11 09:47:52

共1个组件

default

ID: d418v2jrh810mp6salo

服务启动中

组件概览

弹性伸缩

配置详情

基础配置

网络配置

高级配置

存储配置

网络配置

中文名称	英文名称	容器端口	对外服务
--	hth	80	已启用

高级配置

运行命令

环境变量

配置文件

文件路径

存储配置

存储	选择文件路径	挂载路径

5.通过独享资源部署的应用展示 helm 地址、命名空间、chart values 等参数。

查看配置

Helm chart地址

oci://harbor.ctyuncdn.cn/esx-dev/chart-test/http-echo:0.1.3

命名空间

dadada

chart values

```
1 # Default values for lzy-test.
2 # This is a YAML-formatted file.
3 # Declare variables to be passed into your templates.
4
5 # This will set the replicaset count more information can be found
6 replicacount: 1
7
8 # This sets the container image more information can be found
9 image:
10   repository: harbor.ctyuncdn.cn/esx/hashicorp/http-echo
11   tag: latest
12
13 args:
14   - "-text=Hello from Helm chart!"
15   - "-listen=:8080"
16
```

超时时间

300 秒

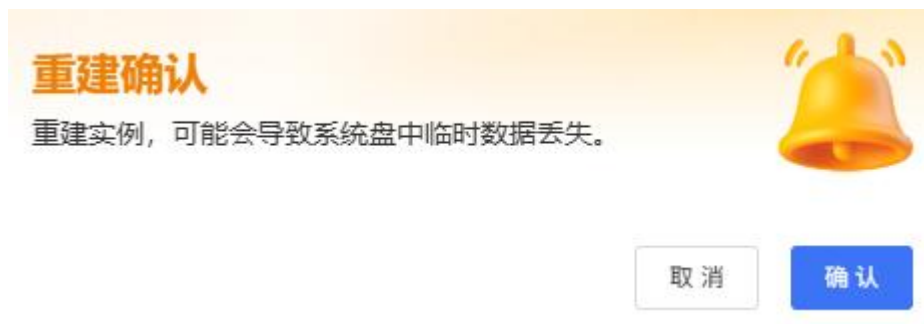
关闭

3.3.3 查看资源信息

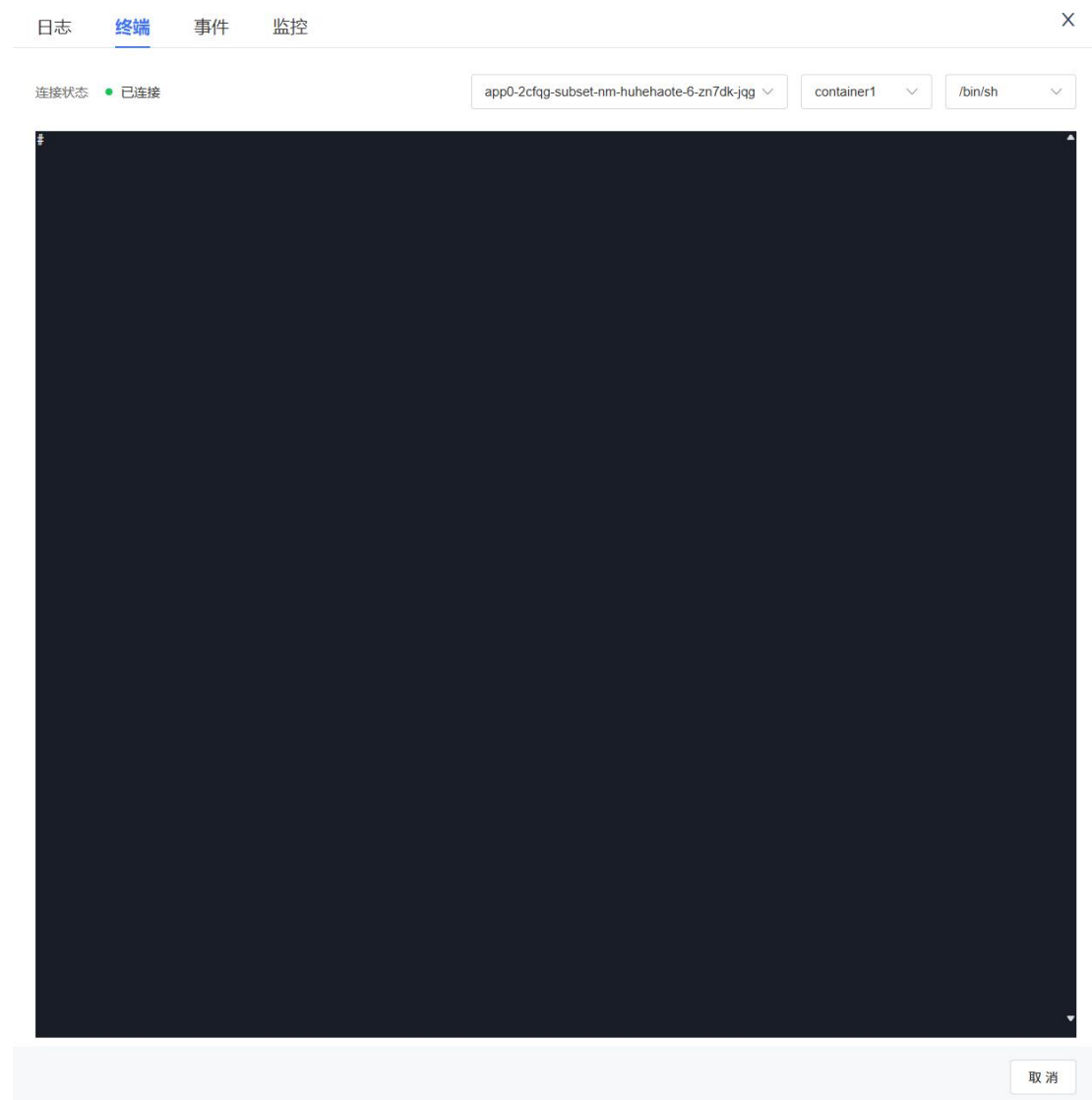
1. 登录应用托管控制台，在左侧导航栏单击【应用实例】，点击应用实例名称进入详情页。
- 2.选择组件，可查看该组件实例的部署实例资源信息。



- 3.点击【重建】或【批量重建】，可手动对实例进行单独或批量的重启操作。

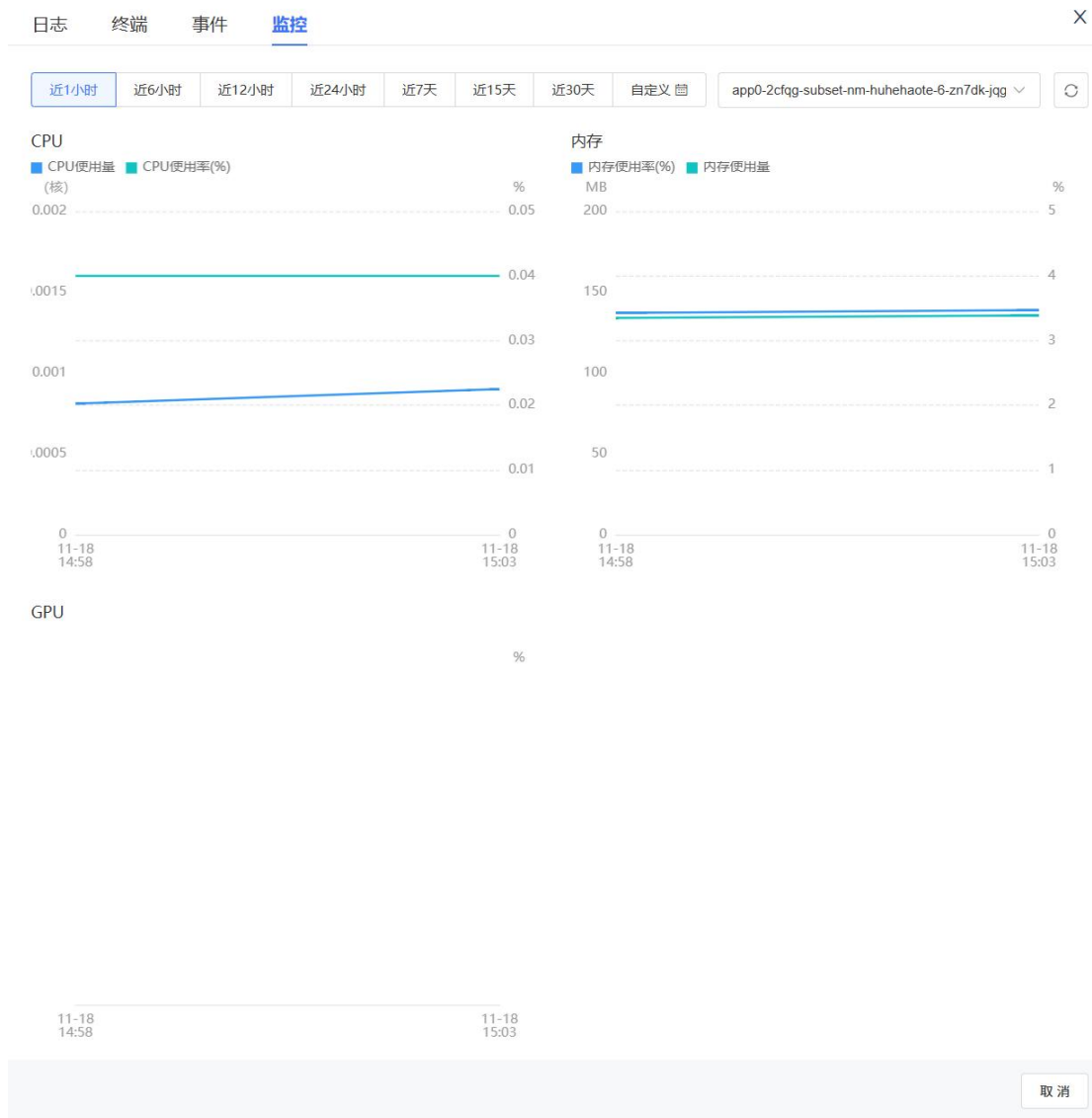


- 4.点击【终端】进入到资源的控制台页面，支持输入命令进行故障的排查。



3.3.4 查看实例监控

1. 登录应用托管控制台，在左侧导航栏单击【应用实例】，点击应用实例名称进入详情页。
2. 选择组件，点击【监控】，可查看该应用实例的监控信息，支持切换实例，可查询实例的CPU、内存和显卡数据。



3.3.5 查看实例日志

1. 登录应用托管控制台，在左侧导航栏单击【应用实例】，点击应用实例名称进入详情页。
2. 选择组件，点击【日志】，可查看该应用实例的日志信息，支持切换实例。
3. 可查询实时日志。

日志 终端 事件 监控 ×

连接状态 ● 正在连接中

app0-2cfs0-subset-nm-huhehaote-6-6cfkh-xdh ▼ container1 ▼ 最近500条 ▼ ☐ 容器退出日志 ☒ 自动滚动 ☒ 自动刷新(每5秒) ↻

1 加载中...

↑

↓

⌂

取消

3.3.6 查看实例事件

1. 登录应用托管控制台，在左侧导航栏单击【应用实例】，点击应用实例名称进入详情页。
2. 选择组件，点击【事件】，可查看该应用实例的最近一个小时事件信息，支持切换实例。

展示最近 1 小时产生的事件数据

app0-2cfqg-subset-nm-huhehaote-6-zn7dk-jqg

搜索关键词



事件名称	类型	事件描述	发生次数	最近发生时间	首次发生时间
容器调度完成	正常	Successfully assigned cps-23/app0-2cfqg-subset-nm-huhehaote-6-zn7dk-jqgqn to vk-nm-huhehaote-6-f26e787a911c4cceb168b7307a66776	--	2025-11-18 14:46:01	2025-11-18 14:46:01
资源准备完成	正常	Create pod in provider successfully	1	2025-11-18 14:46:08	2025-11-18 14:46:08
拉取镜像中	正常	Pulling image ***	1	2025-11-18 14:54:03	2025-11-18 14:54:03
镜像拉取完成	正常	Successfully pulled image *** in 1m28.991613488s	1	2025-11-18 14:55:32	2025-11-18 14:55:32
容器创建中	正常	Created container container1	1	2025-11-18 14:55:37	2025-11-18 14:55:37
容器启动	正常	Started container container1	1	2025-11-18 14:55:37	2025-11-18 14:55:37

取消

3.3.7 查看变更记录

- 1.登录应用托管控制台，在左侧导航栏单击【应用实例】， 点击应用实例名称进入详情页。
- 2.点击【变更记录】按钮，可查看该应用实例的操作变更记录。



应用实例 / 公共应用详情

mcp-cjc-394891

ID: d465g0nqo550j6ksoa0

部署中

访问

停止应用

更新应用

变更记录

删除应用

付费类型: 按量计费

网络类型: 独享网络

可用区: 内蒙呼和浩特6-esk-准生产-分布式部署

访问策略: app-default.sg

应用版本: 1.0.0

创建时间: 2025-12-11 14:43:19

应用信息: mcp-cjc-394891

更新时间: 2025-12-11 14:43:19

共1个组件

mcp-test

ID: 9hmg7l9p000gfb0oaq

服务应用中

基础信息

弹性伸缩

配置详情

算力规格: CPU 2 核 | 内存 4 GB

内网地址: ip aa | 内蒙呼和浩特6-esk-准生产-分布式部署 | 198.19.0.261...

外网地址: ip aa | 内蒙呼和浩特6-esk-准生产-分布式部署 | 182.42.228.171.61...

创建时间: 2025-12-11 14:43:19

更新时间: 2025-12-11 14:43:26

实例管理

批量重建

搜索关键词

实例名称	状态	CPU使用率	内存使用率	GPU使用率	创建时间	操作
app01-0qjfboc0-subset-nm-huhehaote-6-b090y-r7j6d	等待中	--	--	--	2025-12-11 14:43:26	日志 终端 事件 更多

监控 重建

3.4 访问服务

前提条件

确认当前实例状态处于运行中状态，且服务端口配置正确。

操作步骤

- 1.登录应用托管控制台，在左侧导航栏单击【应用实例】，在列表页，找到需要访问的实例，在操作栏点击【访问】按钮。

创建及管理应用实例

提供全面的应用实例管理，提升业务稳定性与运维效率。

01 创建应用实例

提供应用市场和自定义应用两种部署模式，快速创建实例。

02 管理应用实例

支持应用的访问、启动、停止、更新与扩容等能力，轻松管理应用。

03 应用实例运维

实时监控应用运行状态与资源使用情况，支持监控、日志与故障功能，实现高效运维。

公共应用资源

专属应用资源

一键快速部署

去应用市场

实例名称

搜索关键词

实例名称ID	状态	应用信息	付费类型	创建时间	更新时间	操作
MCP-Server-678939 ID: d3c3c0w0h3n176wup000	资源准备中	MCP-Server (5 app-d310d9fvqy8kwo2u00	按量计费	2025-09-28 09:58:08	2025-11-18 15:00:59	访问 启动 停止 更多
MCP-Server-355462 ID: d4e1c7j9kuishg2cfpg	运行中	MCP-Server (5 app-d310d9fvqy8kwo2u00	按量计费	2025-11-18 14:45:55	2025-11-18 14:56:02	访问 启动 停止 更多
if-48s-569 ID: d4aqg08t94mkajqjfm20	停止	自定义应用	按量计费	2025-11-13 18:00:46	2025-11-18 09:53:36	访问 启动 停止 更多
lss-yflei-guan-morenbal ID: d445ekmzg4bpgfk0n4g	停止	自定义应用	按量计费	2025-11-09 16:56:51	2025-11-18 09:53:06	访问 启动 停止 更多
lss-yflei	停止中 (0)	自定义应用	按量计费	2025-10-23 08:16:22	2025-11-18 09:51:29	访问 启动 停止 更多

2.可进入实例详情页，可查看访问的域名或地址，复制链接到浏览器中可访问对应服务。



应用实例 / 公共应用详情

mcp-cjc-394891

ID: d486g0rtgs550qkbaa0

部署中

访问

停止应用

更新应用

变更记录

删除应用

行集类型

独享行集

可用区

内网呼和浩特6-eks-准生产-分布式部署

访问策略

app-default-sg

网络类型

独享网络

访问域名

—

应用信息

mcp-cjc-394891

应用版本

1.0.0

创建时间

2025-12-11 14:43:19

更新时间

2025-12-11 14:43:19

共1个组件

组件概览

弹性伸缩

配置详情

停止组件

更新组件

扩容

mcp-test

ID: d486g15gr550qkbaa0

服务启动中

基础信息

算力规格

CPU 2核 | 内存 4GB

内网地址

aa | 内网呼和浩特6-eks-准生产-分布式部署 | 198.19.0.2.61

公网地址

aa | 内网呼和浩特6-eks-准生产-分布式部署

创建时间

2025-12-11 14:43:19

更新时间

2025-12-11 14:43:26

实例管理

批量新建

搜索容器关键词

实例名称	状态	CPU使用率	内存使用率	GPU使用率	创建时间	操作
app0-0ap0oc0-subset-ent-huheha	等待中	—	—	—	2025-12-11 14:43:26	日志 终端 事件 更多

监控

新建

3.5 启动服务

启动应用

1. 登录应用托管控制台，在左侧导航栏单击【应用实例】，在列表页，找到需要启动的实例，在操作栏点击【启动】按钮。

应用实例管理

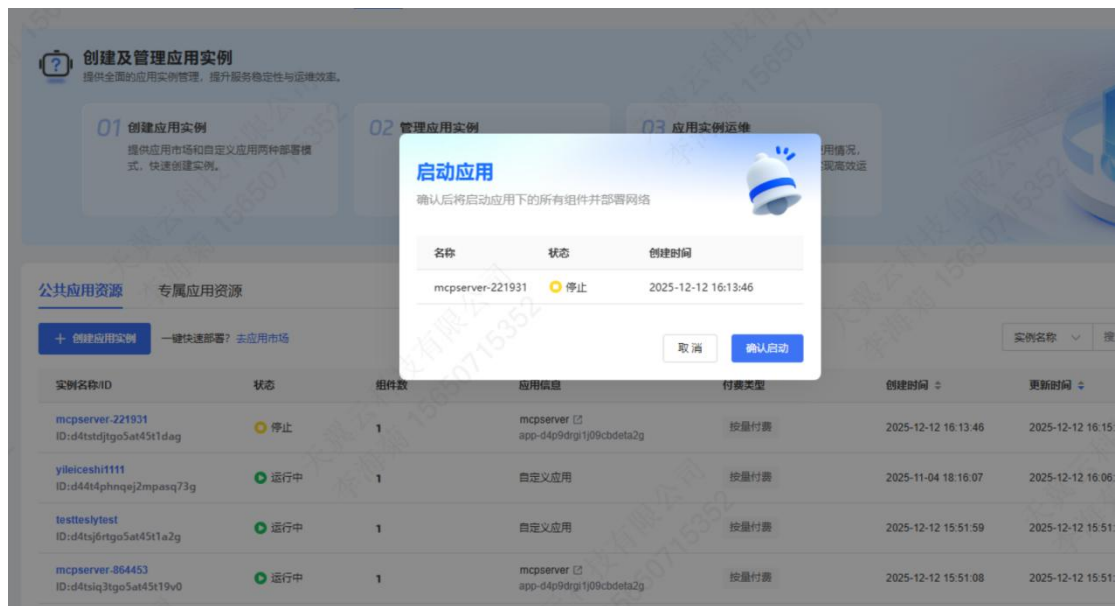
新增应用实例

实例名称

搜索关键词

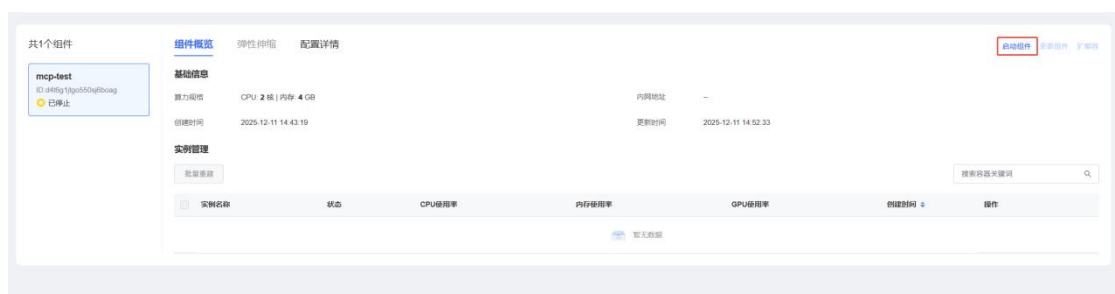
实例名称	实例ID	实例规格	实例状态	创建时间	更新时间	操作	
mcp-cjc-394891	d486g15gr550qkbaa0	已停止	9/11	mcp-计算节点 12 app-470k3m5c3dpg14g	2025-09-29 17:46:52	2025-07-22 10:38:55	<div>启动</div> 更多

2. 在进入启动服务操作弹窗后，点击【确认】完成启动操作。



启动组件

1.当进入应用实例详情，选择停止状态组件，在操作栏点击【启动组件】操作。



2.在进入启动组件操作弹窗后，点击【确认】完成启动操作。



3.6 停止服务

停止应用

1. 登录应用托管控制台，在左侧导航栏单击【应用实例】，在列表页，找到需要停止的实例，在操作栏点击【停用】按钮。



2. 在进入停止服务操作弹窗后，点击【确认】完成停止操作。



停止组件

1.当进入应用实例详情，选择运行中状态组件，在操作栏点击【停止组件】操作。



2.在进入停止组件操作弹窗后，点击【确认】完成启动操作。



3.7 更新服务

前提条件

确认应用实例中欲升级的组件处于可升级的状态，包括：服务启动中、 运行中 、运行失败。

3.7.1 更新应用市场/MCP 市场部署的应用

1.登录应用托管控制台，在左侧导航栏单击【应用实例】，在列表页，找到需要升级的服务，进入实例详情，选择要更新的组件，点击【更新组件】操作。



2.通过应用/MCP 市场部署的应用，与第一次应用部署时参数一致，更新参数如下。

配置文件	文件路径	操作
	暂无数据	
	添加配置文件	

* 选择MCP

全部

MCP应用分类

请输入关键词



fetch

使大型语言模型能够检索和处理网页内容，将HTML转换为markdown格式，以便于更轻...

参数配置

bing

bing

计算器MCP

一种模型上下文协议服务器，通过评估数学表达式，使大型语言模型能够进行精确的数值...

Airbnb民宿搜索

搜索Airbnb房源并获取有关特定房产的详细信息。通过结构化数据轻松规划您的下一次旅...

参数配置

Time时间服务

一个提供时间和时区转换功能的模型上下文协议服务器。该服务器使 LLM 能够获取当前...

参数配置

已选 1 / 5 个MCP

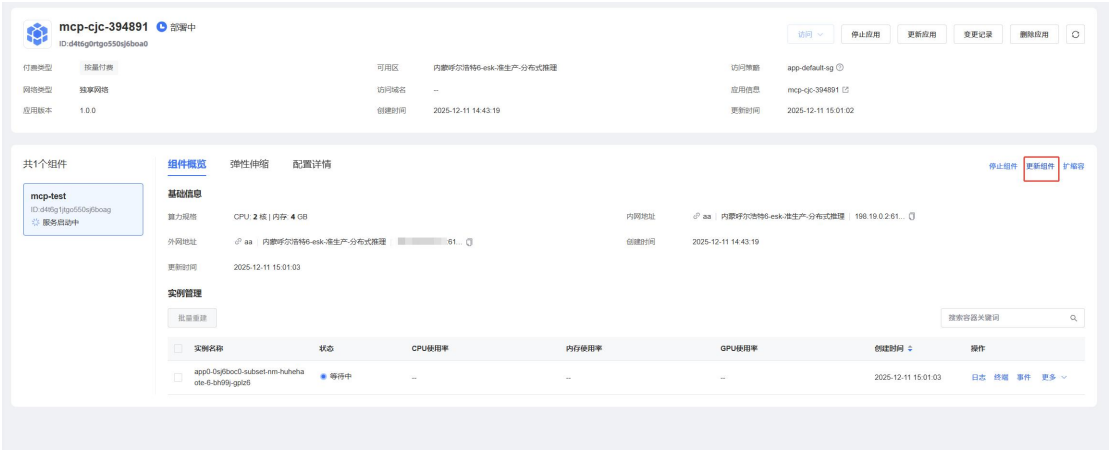
取消

确定



3.7.2 更新自定义应用（公共资源）

1.登录应用托管控制台，在左侧导航栏单击【应用实例】，在列表页，找到需要升级的服务，进入实例详情，选择要更新的组件，点击【更新组件】操作。



2.通过【应用实例】创建的自定义应用，资源类型为公共资源的应用，具体更新参数如下。

参数名称	参数说明
镜像	填写镜像地址（私有仓库需提供授权）
运行命令	配置服务的启动命令
环境变量	配置服务的环境变量
网络变量	配置中文名称、英文名称、容器端口、url 地址、是否对外服务
配置文件	挂载到容器中的指定文件
存储	挂载到容器中的指令目录

更新组件

X

* 镜像地址

ehub.ctcdn.cn/public-zz/nginx-stress-ping:latest

☐私有化镜像仓库授权

* 网络变量

共享网络

中文名称	* 英文名称	* 容器端口	* uri地址	对外服务	操作
zhongwen	zhongwen	80	/*	<input type="checkbox"/> 关闭	删除

添加网络变量

websocket

☐关闭

运行命令

请输入Shell启动命令
例如: /webui.sh --listen --port 8000 --skip-version-check --no-hashing --no-download-sd-model --skip-prepare-environment --api --filebrowser

环境变量

支持批量输入, 格式为key=value, 多变量用英文分号,或换行分隔。点击「内容识别」或输入框内任意位置, 系统将自动解析并填充环境变量
例如: mongoUrl=127.0.0.1:8000;redisUrl=127.0.0.1:8001;env1=test

内容识别

添加环境变量

配置文件

文件路径	操作
暂无数据	

添加配置文件

存储

存储	创建新存储	选择文件路径	挂载路径	操作
				暂无数据

添加存储

取消

确定

3.7.3 更新自定义应用（独享资源）

1.登录应用托管控制台，在左侧导航栏单击【应用实例】，在列表页，找到需要升级的服务，在操作栏点击【更新】按钮。

应用实例管理						
实例名称ID	状态	实例数	应用信息	创建时间	更新时间	操作
City-V1.0.0-20250505 id:42e4e4e4e4e4e4e4	运行中	1/1	City-V1.0.0-20250505 app-42e4e4e4e4e4e4e4	2025-07-30 15:24:31	2025-07-30 15:28:23	详情 日志 更多
22q123123 id:42e4e4e4e4e4e4e4	运行中	1/1	应用 MCP 部署模板 app-42e4e4e4e4e4e4e4	2025-07-24 15:22:20	2025-07-30 15:15:40	详情 日志 更多

2.通过【应用实例】创建的自定义应用，资源类型为独享资源的应用，具体更新参数如下。

参数名称	参数说明
------	------

参数名称	参数说明
Helm chart 地址	指定 Helm Chart 的存储位置。如果您的 Helm Chart 存放在私有仓库中，需要进行授权设置
命名空间	制定应用部署的命名空间
chart values	自定义 Helm Chart 的配置参数，根据实际需求调整各个参数，如镜像版本、副本数量、服务类型等
超时等待时间	定义 Helm 部署过程中拉取 Chart 的最大等待时间

更新实例

X

* Helm chart地址

oci://harbor.ctyuncdn.cn/esx-dev/chart-test/http-echo:0.1.3

获取chart values

☐ 私有仓库授权

* 命名空间

dadada

chart values

1

Default values for lzy-test.

2

This is a YAML-formatted file.

3

Declare variables to be passed into your templates.

4

5

This will set the replicaset count more information can be found here: <https://kubernetes.io/docs/concepts/workloads/controllers/replicaset/>

6

replicaCount: 1

7

8

This sets the container image more information can be found here: <https://kubernetes.io/docs/concepts/containers/images/>

9

image:

10

repository: harbor.ctyuncdn.cn/esx/hashicorp/http-echo

11

tag: latest

12

13

args:

14

- "-text=Hello from Helm chart!"

15

- "-listen=:8080"

16

请根据需要修改配置值，未修改的值将使用默认设置

* 超时等待时间

-

300

+

秒

Helm chart说明

规范

示例

1.templates模版文件中每一个资源都必须包含metadata, metadata.labels, metadata.namespace

2.helm默认会对资源的metadata.labels添加[- include "\$名称.labels" - | nindent 4]], 请保留或按第三点添加

3.metadata.labels需包含helm内置的`app.kubernetes.io/instance` 标签, 值为release名称(例:[.Release.Name])

4.metadata.namespace的值需指定为release.namespace(例:[.Release.Namespace])

5.工作负载类资源(deployment statefulset 等), 需要在spec.template.metadata中添加`app.kubernetes.io/instance` 标签

3.完成修改后，点击【确认】即可启动更新流程。

3.7.4 更新应用网络

1.登录应用托管控制台，在左侧导航栏单击【应用实例】，在列表页，找到需要升级的服务，进入实例详情，点击【更新应用】操作。



2.调整对外组件对外暴露域名、url、是否开启 websocket，点击确定，更新网络配置。

更新应用×

* 网络配置

组件名称	端口	* url地址	对外服务	websocket
nginx-nports-1lei	80	<input type="text" value="/a"/>	<div>启用</div>	<div>关闭</div>
	8081	<input type="text" value="/b"/>	<div>启用</div>	
ng-1con-1port-1lei	80	<input type="text" value="/"/>	<div>启用</div>	<div>关闭</div>

* 访问域名

* 可见范围

仅自己可见

空间内公开可见

取消

确认更新

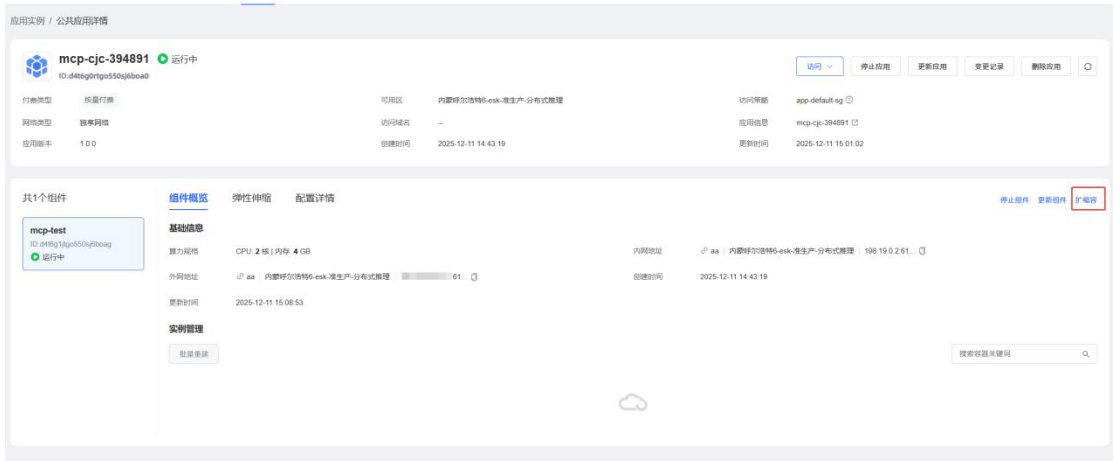
3.8 扩缩容

前提条件

确认应用实例中欲扩缩容的组件状态处于可扩缩容状态，包括：运行中、服务启动中。

操作步骤

- 1.登录应用托管控制台，在左侧导航栏单击【应用实例】，在列表页，找到需要扩缩容的实例，进入实例详情，选择组件，在操作栏点击【扩缩容】按钮。



2.在进入扩缩容操作页面后，可以调整的字段如下：实例数量：增加/减少实例数量，实例数量最小为 1。



3.完成修改后，点击【确认】即可启动扩缩容流程。在组件详情中的可查询扩缩容实例信息。



3.9 弹性伸缩

概述

应用托管平台支持通过监控指标自动对组件的部署资源进行弹性伸缩,在满足业务需求的同时优化算力成本。适用于突发流量和典型周期性流量的应用场景。

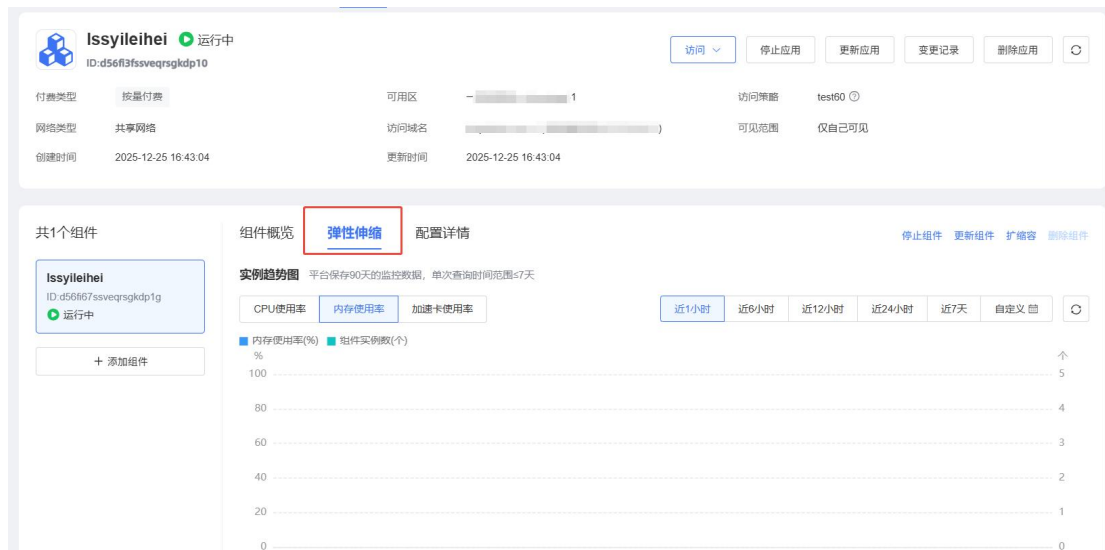
说明：当前支持 CPU 使用率、内存使用率和加速卡使用率三种监控指标。

前提条件

已在平台部署应用实例，且目标应用实例当前处于正常运行状态（非异常/故障状态）。

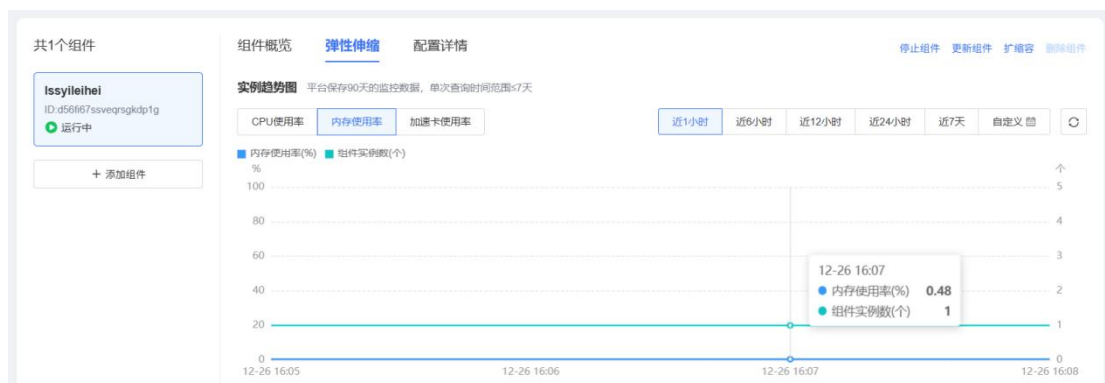
操作步骤

- 1.登录应用托管控制台，在左侧导航栏单击【应用实例】，在应用实例列表页，找到需要配置弹性伸缩策略的目标应用实例，进入应用详情，在详情页左侧选中需要配置弹性伸缩的组件，单击顶部【弹性伸缩】页签。



2.在进入弹性伸缩操作页面后，可以查询实例的各指标变化趋势。

说明：平台保存 90 天的监控数据，单次查询时间范围不超过 7 天。



3.根据实例历史监控指标，配置弹性伸缩策略，单击目标弹性伸缩策略的【编辑】按钮。



4.在弹出的弹性伸缩面板，配置以下参数信息。

参数	说明	示例
策略名称	自定义策略名称	default-policy
触发条件	<div>选择指标类型：</div> <div><div><div></div><div>CPU 使用率</div></div><div><div></div><div>内存使用率</div></div><div><div></div><div>加速卡使用率</div></div></div> <div>说明：</div> <div><div><div></div><div>指标类型的聚合方式，请参见控制台说明</div></div><div><div></div><div>您可以同时设置多个指标类型</div></div></div>	CPU 使用率

	设置指标触发阈值：当指标实际值持续达到 / 超过（或低于）该阈值时触发策略	80%
最小实例数	弹性伸缩的最低保障实例量，缩容时实例数不会低于此值	最小实例数：1
最大实例数	弹性伸缩的最高可扩容实例量，扩容时实例数不会超过此值	最大实例数：10

编辑弹性伸缩策略

×

* 策略类型

监控指标策略

* 策略名称

default-policy

* 触发条件

当设置多个监测指标时，按各个指标分别计算出预期的可用区实例数并取最大值

指标	目标值	操作
CPU使用率	80 %	删除
内存使用率	80 %	删除
加速卡使用率	85 %	删除

取消

确认编辑

5.策略配置完成后，单击弹性伸缩策略的【启用】按钮，启用策略。



6.可查看扩缩容事件，验证弹性策略效果。



7.停用弹性伸缩策略，单击弹性伸缩策略的【停用】按钮，停用策略。



3.10 增减组件

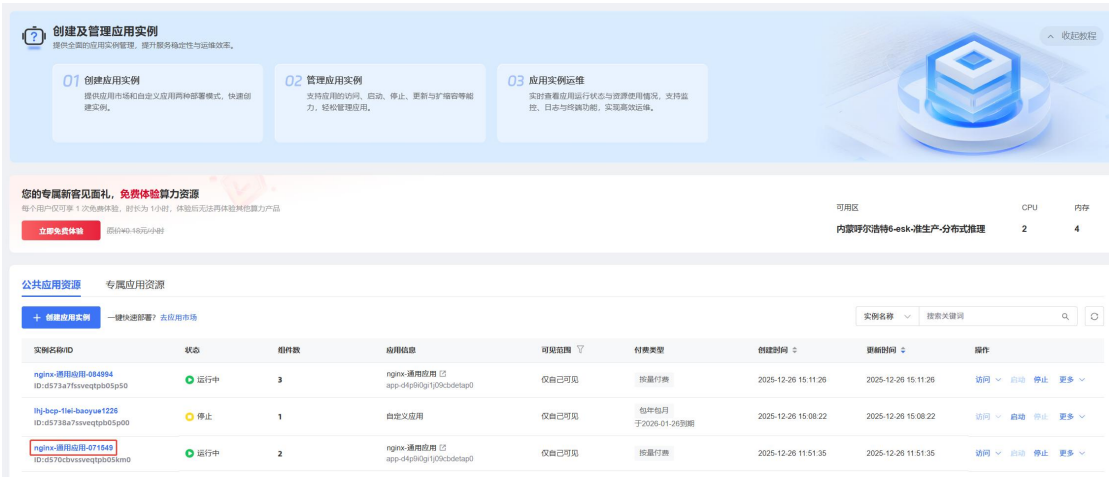
前提条件



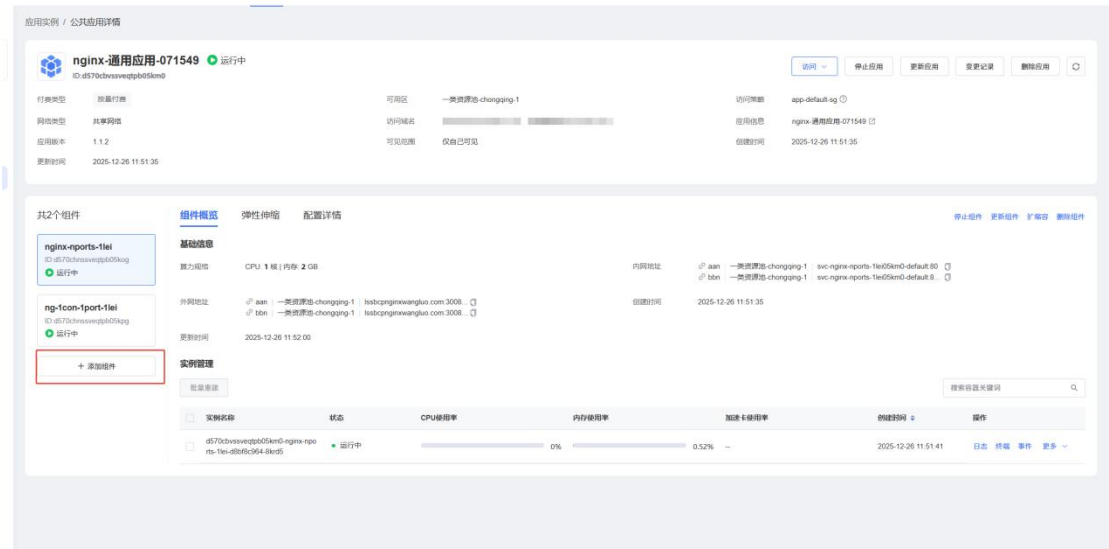
- 1. 应用实例处于部署中、运行中。
- 2. 应用实例为按需计费类型,使用共享网络模式。
- 3. 应用实例不是 MCP Servers。

增加组件

1.登录应用托管控制台，在左侧导航栏单击【应用实例】，在列表页，找到需要增减组件的实例，点击实例名称。



2.点击组件列表下方【添加组件】操作。



3.输入组件名称。

所属应用

nginx-通用应用-071549

付费类型: 按需计费 | 访问策略: app-default-sg

* 组件名称

请输入内容

长度为4-32字符，支持大小写字母、数字及分隔符(-)

4.组件配置配置：填入镜像地址，选择算力规格，数量。

组件配置

* 镜像地址

请输入镜像仓库地址/镜像名/版本

☐ 私有化镜像仓库授权

可用区

一类可用区-chongqing-1

* 算力规格

CPU: 1核 | 内存: 208M | GPU: 1*10

选择实例

* 数量

-

1

+

5.网络配置：配置容器端口及是否开启对外服务。访问策略为服务的公网访问提供安全防护，可根据需要调整访问策略内容。一个应用只能绑定一个访问策略。在共享网络模式下，可配置是否需要开启 websocket。

- 中文名称：用于直观展示端口用途，在应用实例列表中可下拉展示。
- 英文名称：标识端口。
- 容器端口：应用组件需要暴露的服务端口，系统会根据网络类型不同自动分配端口。
- url：指应用组件对应端口访问的 url 表达式，支持正则表达式。

网络配置

* 网络类型

共享网络

中文名称	英文名称	容器端口	url地址	对外服务	操作
<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input checked="" type="checkbox"/>	删除
添加网络变量					

访问域名

testcngnmxwz.com

websocket

☒

6.高级配置：

参数名称	参数说明
运行命令	配置服务的运行命令
环境变量	配置服务的环境变量
存储	挂载持久化存储
配置文件	将文件内容挂载到容器中指定文件

高级配置

运行命令

请输入Shell启动命令
例如: webui.sh --listen --port 8000 --skip-version-check --no-flushing --no-download-sd-model --skip-prepare-environment --api --filebrowser

环境变量

支持批量输入，格式为key=value，多变量用英文分号分隔，点击「内容识别」或输入框外任意位置，系统将自动解析并填充环境变量
例如: mongoUri=127.0.0.1:8000;redisUri=127.0.0.1:8001;env1=test

内容识别

添加环境变量

配置文件

文件路径	操作
	删除
	添加配置文件

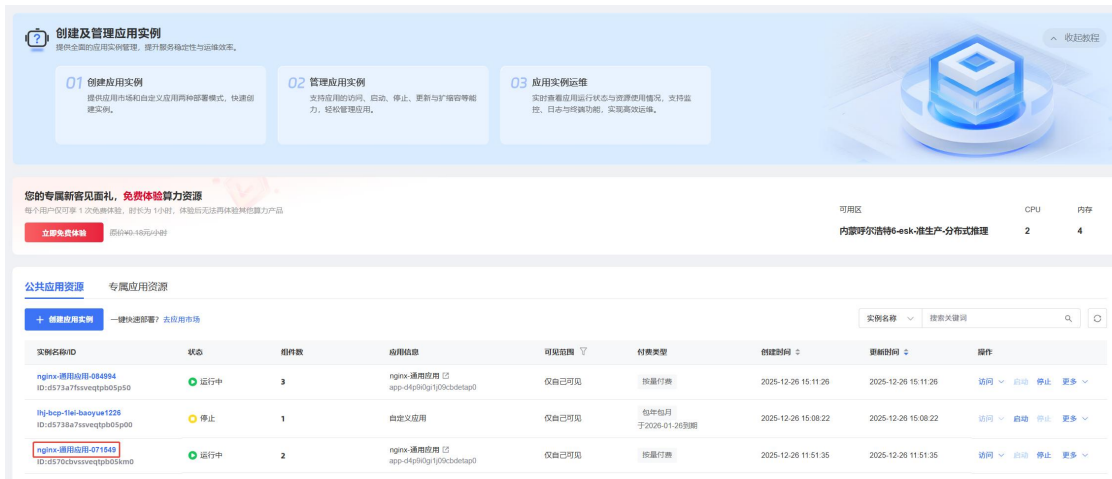
存储

存储	选择文件路径	挂载路径	操作
请选择存储	请选择	默认 /tmp/username	删除
			添加存储


7.完成以上配置，确认部署组件的资源配置费用，并勾选协议点击【添加组件】后，完成组件的添加。


删除组件

1.登录应用托管控制台，在左侧导航栏单击【应用实例】，在列表页，找到需要增减组件的实例，点击实例名称。



应用实例 / 公共应用实例

 nginx-通用应用-071549

 运行中

ID: d570cbssveqpb05km0

行集类型

扩展行集

网络类型

共享网络

应用版本

1.1.2

更新时间

2025-12-26 11:51:35

可用区

一类资源池-chongang-1

访问策略

app-default-sg

访问策略

nginx-通用应用-071549

创建时间

2025-12-26 11:51:35

实例名称

nginx-通用应用-071549

可见范围

仅自己可见

共2个组件

组件概述

弹性伸缩

配置详情

停止组件


更新组件

扩容组件

删除组件


nginx-ports-1lei

ID: d570cbssveqpb05kg

 运行中

ng-1con-1port-1lei

ID: d570cbssveqpb05kpg

 运行中

算力规格

CPU 1核 | 内存 2 GB

内网地址

ip: aan 一类资源池-chongang-1 svc: nginx-ports-1lei05km0-default-60

ip: bbn 一类资源池-chongang-1 svc: nginx-ports-1lei05km0-default-8

公网地址

ip: aan 一类资源池-chongang-1 lsb: cngmwxwangleu.com.3008

ip: bbn 一类资源池-chongang-1 lsb: cngmwxwangleu.com.3008

更新时间

2025-12-26 11:52:00

创建时间


2025-12-26 11:51:35

实例管理

登录连接

搜索容器关键词

Q

实例名称	状态	CPU使用率	内存使用率	加速卡使用率	创建时间	操作
d570cbssveqpb05km0-nginx-ports-1lei-d8bfc964-8x05	 运行中	<div></div>	<div>0%</div>	<div>0.52%</div>	2025-12-26 11:51:41	<div>日志</div> <div>详情</div> <div>事件</div> <div>更多</div>

删除组件

将删除组件及其关联的网络，无法恢复，请谨慎操作！

组件名称	状态	创建时间
nginx-nports-1lei	运行中	2025-12-26 11:5

取消确认删除



我已知晓操作风险取消确认删除

3.11 删除服务

前提条件

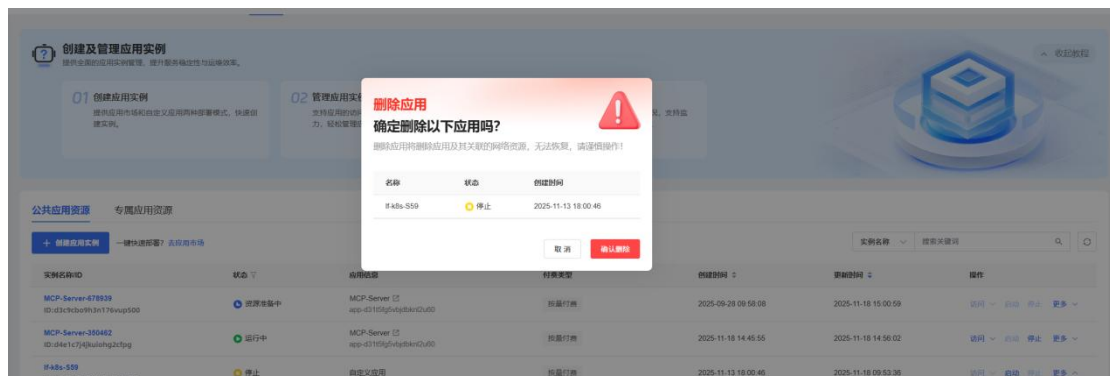
确认已完成服务的迁移，删除后该实例所有资源将会清空且不可恢复。

删除应用

1. 登录应用托管控制台，在左侧导航栏单击【应用实例】，在列表页，找到需要删除的实例，在操作栏点击【删除】按钮。



2. 二次弹框删除服务，点击【确认删除】，删除应用。



4.应用访问策略

4.1 创建访问策略

操作步骤

1.登录应用托管控制台，在左侧导航栏单击【应用】->【应用访问策略】， 点击【创建访问策略】按钮。



2.填写基础信息：策略名称、选择策略类型、IP 地址/网段、描述。

创建访问控制策略



* 策略名称

演示策略

* 策略类型?

☒ 白名单

☐ 黑名单

* IP地址/网段

1.0.0.0

描述

- ✓ 每行一个IP地址或网段，以回车键换行
- ✓ 每个IP地址或网段都可以用“|”分隔添加描述 如 “10.0.0.0/24|办公网出口”
- ✓ 描述长度范围是0到255字符，不能包含<>

0/255

确认

取消



注意：访问控制策略对关联的应用生效。

3.完成表单填写后，点击【确认】按钮，完成访问策略创建。

创建访问控制策略



* 策略名称

演示策略

* 策略类型?

☒ 白名单

☐ 黑名单

* IP地址/网段

1.0.0.0

描述

- ✓ 每行一个IP地址或网段，以回车键换行
- ✓ 每个IP地址或网段都可以用“|”分隔添加描述 如 “10.0.0.0/24|办公网出口”
- ✓ 描述长度范围是0到255字符，不能包含<>

0/255

确认

取消

4.您可在访问控制列表查看您创建的访问控制策略，并对访问控制策略进行管理。

访问控制策略						
策略名称	策略类型	包含IP地址范围	关联服务	描述	创建时间	操作
演示策略	白名单	1.0.0.0	所有服务	--	2025-05-15 11:33:37	修改 删除

共1条 10 < 1 >

4.2 访问策略管理

前提条件

已创建访问控制策略。

4.2.1 修改访问控制策略

1.登录应用托管控制台，在左侧导航栏单击【应用】->【应用访问策略】，进入访问控制策略列表页，单击【修改】。

应用访问策略					
+ 新建访问策略		请选择策略类型		请输入策略名称	
策略信息	IP地址网站	关联应用数（运行中/总数）	描述	更新时间	操作
 演示策略 白名单	1.1.1.1	0/0	--	2025-07-30 15:39:30	修改 删除

2.填写基础信息：策略名称、选择策略类型、IP 地址/网段、描述。

修改访问控制策略

×

* 策略名称

演示策略

* 策略类型?

☒ 白名单

☐ 黑名单

* IP地址/网段

1.0.0.0

描述

0/255

确认

取消

3.完成表单填写后，点击【确认】按钮，完成访问策略修改。

修改访问控制策略



* 策略名称 演示策略

* 策略类型 ☒ 白名单 ☐ 黑名单

* IP地址/网段 1.0.0.0

描述

0/255

确认

取消

4.2.2 删除访问控制策略

1. 登录应用托管控制台，在左侧导航栏单击【应用】->【应用访问策略】，进入访问控制策略列表页，单击【删除】。

应用访问策略					
+ 新建访问策略		请选择策略类型		请输入策略名称	
策略名称	IP地址/网段	关联应用数 (运行中/总数)	描述	更新时间	操作
演示策略 白名单	1.1.1.1	0/0	--	2025-07-30 15:39:30	详情 删除

2.点击弹窗【删除】按钮，即可删除访问控制策略



注意：系统默认策略无法删除。有关联应用的访问策略无法删除。

5.项目空间

5.1 创建项目空间

概述

项目空间是应用托管的顶层概念，为企业和团队提供统一的资源管控及人员权限管理能力，同时可用于不同环境的资源隔离，支持灵活资源分配与团队协作。资源创建后不支持更改其所属的项目空间，请提前做好规划。

使用说明

项目空间内创建的资源归属项目空间所属人，项目成员仅作为空间的操作员。

角色定义

为满足团队管理的需求，应用通过预定义了以下多种系统角色，以便团队根据分工为成员分配合适的角色：

- 管理员：项目空间所属人，拥有空间所有权限，包括增删工作空间。
- 项目管理员：拥有空间所有权限，不可删除空间。
- 资产管理员：拥有项目资产管理的全部权限。
- 项目成员：拥有应用实例管理全部权限。
- 访客：可查看项目资产信息，但不可操作。

操作步骤

1. 登录应用托管控制台，在左侧导航栏单击【资产】->【项目空间】，点击【创建项目空间】按钮。

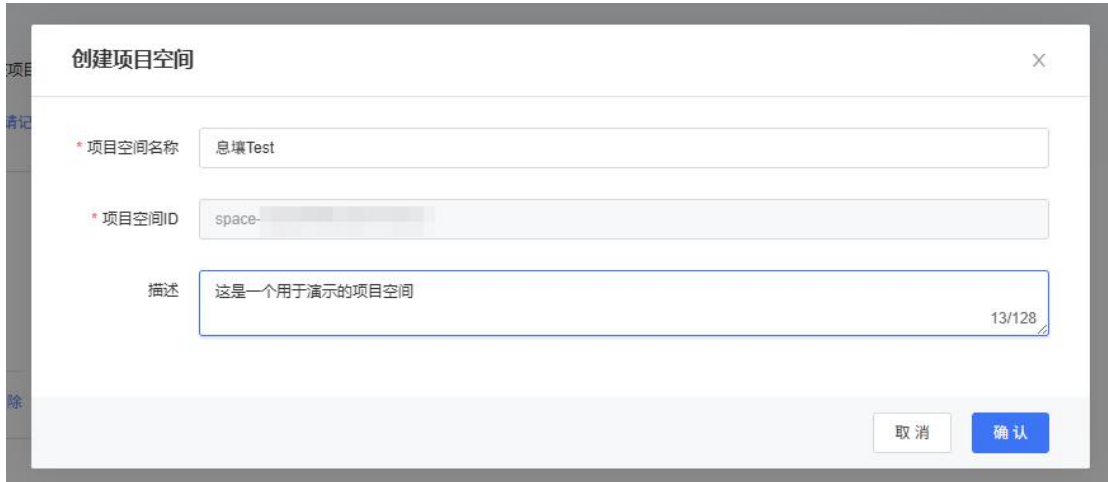


说明

租户首次使用应用托管控制台，平台会自动为用户创建默认空间并初始化。

若项目空间初始化失败导致服务不可用，请刷新页面或重新进入应用托管控制台。

2.填写基础信息：项目空间名称、描述。

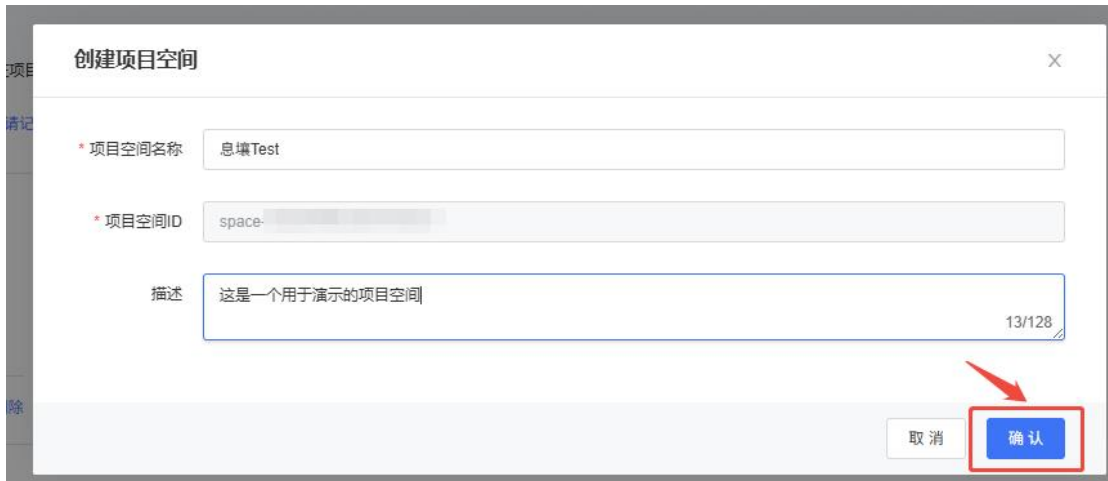


说明

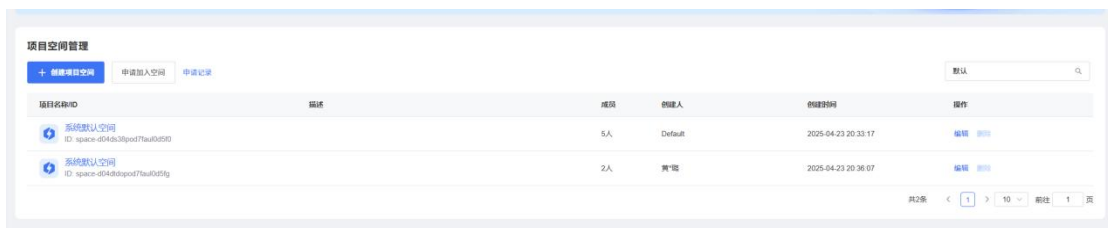
项目空间名称具有租户下唯一性，同一个租户下不可创建同名项目空间。

项目空间 ID 由平台自动生成，项目空间 ID 全局唯一，不可修改。

3.完成表单填写后，点击【确认】按钮，完成项目空间创建。



4.您可在项目空间栏目查看您创建或已加入的项目空间，并对项目空间进行管理。



项目空间名称/ID	描述	成员	创建人	创建时间	操作
系统默认空间 ID: space-d048b39pod77au0b50		5人	Default	2025-04-23 20:33:17	编辑
系统默认空间 ID: space-d048b39pod77au0b50		2人	黄**	2025-04-23 20:36:07	编辑

5. 点击目标项目空间操作栏的【编辑】按钮，可对项目空间进行管理。支持开启“允许自主申请加入空间”，其他用户可通过项目空间 ID 搜索加入空间。

项目空间 / 编辑

88 基本信息

成员管理

空间审批

角色管理

资源管理

* 项目空间名称

演示项目空间

* 项目空间ID

space-d31mu3nievrrfc0ep4g

描述

请输入项目描述

0/128

* 允许自主申请加入空间

开

开启该功能后，其他用户可通过项目ID搜索申请加入空间

5.2 项目空间管理

前提条件

已开通项目空间并完成项目空间初始化。

5.2.1 查询项目空间列表

1. 登录应用托管控制台，在左侧导航栏单击【资产】->【项目空间】，查询项目空间列表。

支持输入项目空间名称关键字进行模糊搜索。

创建及管理项目空间

项目空间为开发人员提供团队协作的支持，可在项目空间内管理计算资源和人员权限。

01 创建项目空间

根据您所需的项目信息填写，也可以将企业项目中已有的信息同步。

02 管理项目空间

您可以在项目空间中添加项目成员，并对使用资源进行预算设置。

03 监控资源使用情况

在项目空间详情页，您可以监控资源消耗趋势，并查看每位成员消耗的资源金额。

项目空间管理

+ 新建项目空间

申请加入空间

申请记录

默认

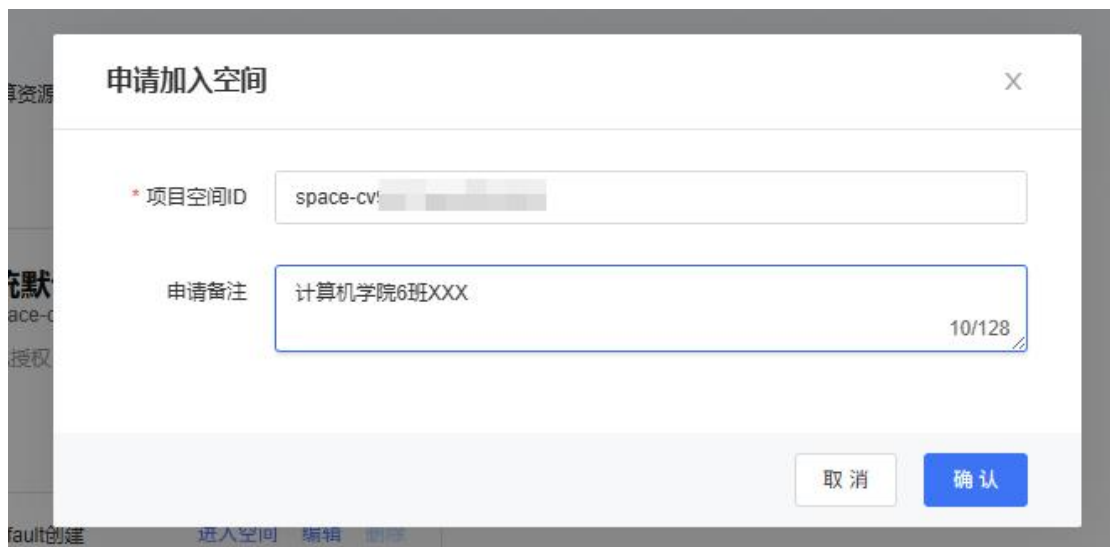
搜索

项目空间ID	描述	成员	创建人	创建时间	操作
系统默认空间 ID: space-d048c3gpd77au0c050		5人	Default	2025-04-23 20:33:17	编辑
系统默认空间 ID: space-d048d0poc77au0c05g		2人	黄飞	2025-04-23 20:36:07	编辑

共2条 < 1 > 10 页 前往 1 页

5.2.2 申请加入空间

1. 点击【项目空间】进入项目空间列表页，点击【申请加入空间】，填写项目空间 ID 和申请备注，点击【确认】即可发送申请。

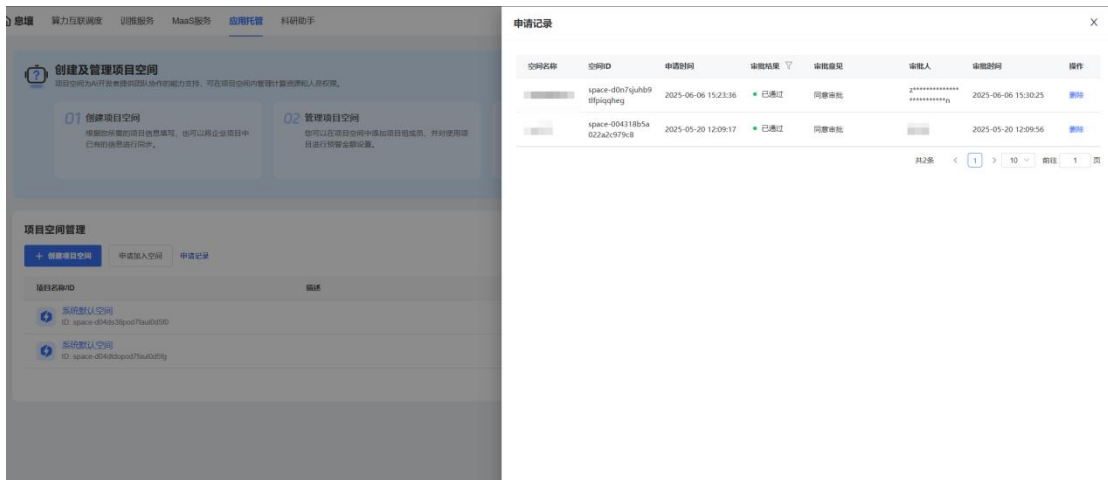


说明

项目空间 ID 可联系对应的项目空间创建人或管理员获取。

若项目空间未开启【允许自主申请加入】或项目空间 ID 不存在会导致申请失败，申请前请确认项目空间 ID 是否正确及已开启【允许自动申请加入】

2. 点击【申请记录】可查看已提交的项目空间申请及审批状态，支持撤回申请。



空间名称	空间ID	申请时间	审批结果	审批意见	审批人	审批时间	操作
空间名称	space-d8a7qjh89 8fplqplqeg	2025-06-06 15:23:36	已通过	同意审批	***** *****	2025-06-06 15:30:25	撤回
空间名称	space-00431855a 022a8-c979a8	2025-05-20 12:09:17	已通过	同意审批		2025-05-20 12:09:56	撤回

共2条 < 1 > 10 条 1 页

5.2.3 进入/切换项目空间

1. 在项目空间列表，找到想要使用的项目空间，点击【空间名称】按钮，即可切换到指定工作空间，系统会自动跳转【概览】页面。



空间名称/ID	描述	成员	创建人	创建时间	操作
系统默认空间 ID: space-d8a7qjh89 8fplqplqeg		5人	Default	2025-04-23 20:33:17	编辑 删除
系统默认空间 ID: space-d8a7qjh89 8fplqplqeg		2人	黄**	2025-04-23 20:36:07	编辑 删除

共2条 < 1 > 10 条 1 页

2. 可以在菜单切换项目空间，也可以在【概览】页顶部查看当前的项目空间信息及点击下拉列表进行切换。项目空间切换后系统会自动更新【概览】页面数据。

应用托管

系统默认空间

项目管理员

02377a12

a2c979c8

670b8da5

系统默认空间(@Default创建)

ID: space-d04ds38pod7faul0d5f0

374e097

项目

项目空间: 系统默认空间(@Default创建)

空间ID: space-d04ds38pod7faul0d5f0

Hi, 欢迎使用应用托管

为您提供更丰富、更优质的应用产品, 助力您的业务轻松部署与应用便捷管理

找应用

丰富的云应用和工具, 助力企业数字化转型

LLaMA Factory

LLAMA FACTORY是一个旨在普及LLMs微调的框架。它通过可扩展的模块统一了多种高效微调方法, 使得教...

DeepSeek-R1-7B (1卡进阶型)

AnythingLLM 是助力大模型应用开发的平台。它有强大模型适配能力, 支持各类开源、闭源大模型, 可按需灵...

DeepSeek-R1知识库

一站式大模型应用开发平台。它能便捷接入各类数据并管理, 支持模型微调与多模型调用。

查看全部 >

找MCP

提供优质的MCP应用资源, 赋能AI多元服务

中文趋势聚合

基于 Model Context Protocol (MCP) 协议的全网热点趋势一站式聚合服务

Wikipedia MCP 服务器

这是一个 Model Context Protocol (MCP) 服务器, 可以从维基百科检索信息, 为大型语言模型 (LLMs) 提供上...

mcp-server-tmdb

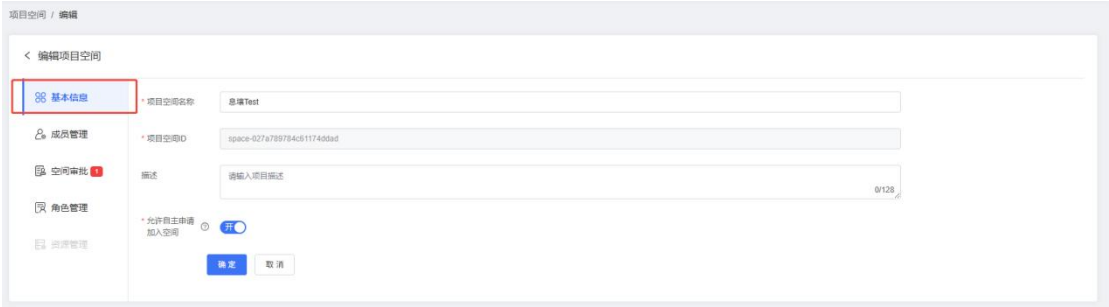
此MCP服务器与The Movie Database (TMDB) API集成, 提供电影信息、搜索功能和推荐。

查看全部 >

5.2.4 编辑项目空间基本信息

1.在项目空间列表，点击项目空间的【编辑】按钮，进入编辑页面。

2.在【基本信息】栏目，可以修改项目空间名称、项目描述等信息，点击【确定】按钮保存修改。若您希望其他用户可以自主申请加入您的项目空间，可以打开允许自主申请加入项目空间按钮。空间加入申请可在【空间审批】中查看。



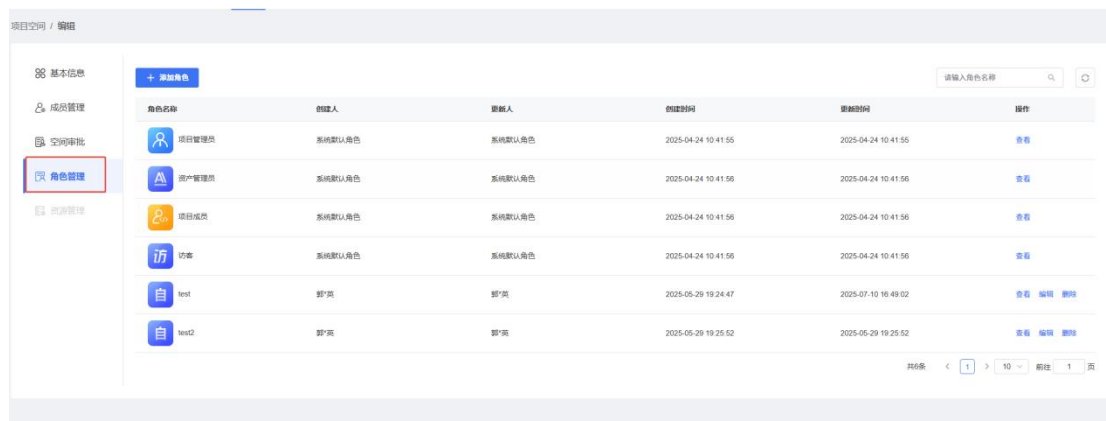
说明

同一个租户下的项目空间不允许重名。

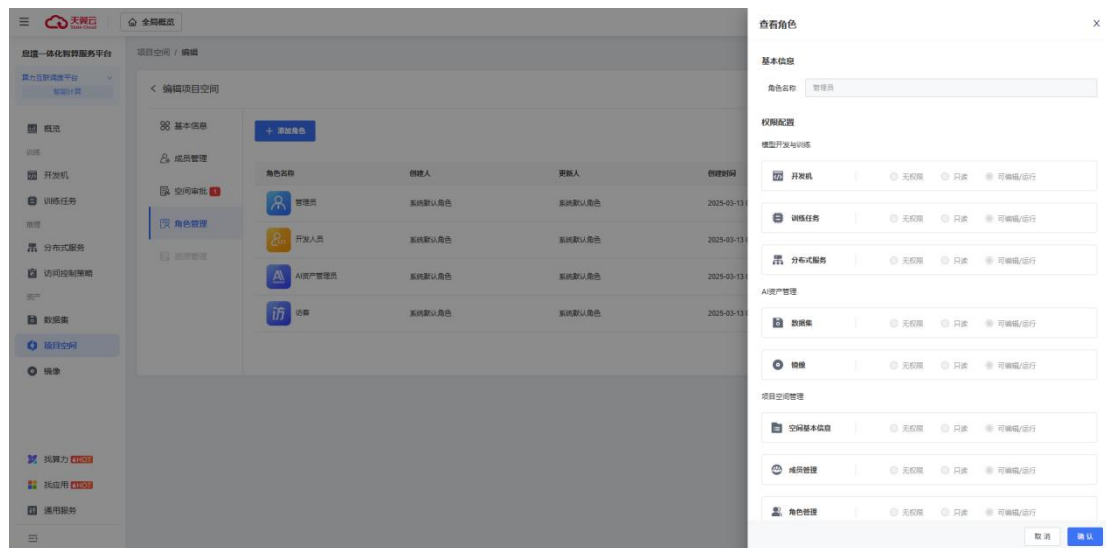
项目空间 ID 由平台自动生成，不可修改。

5.2.5 项目空间角色管理

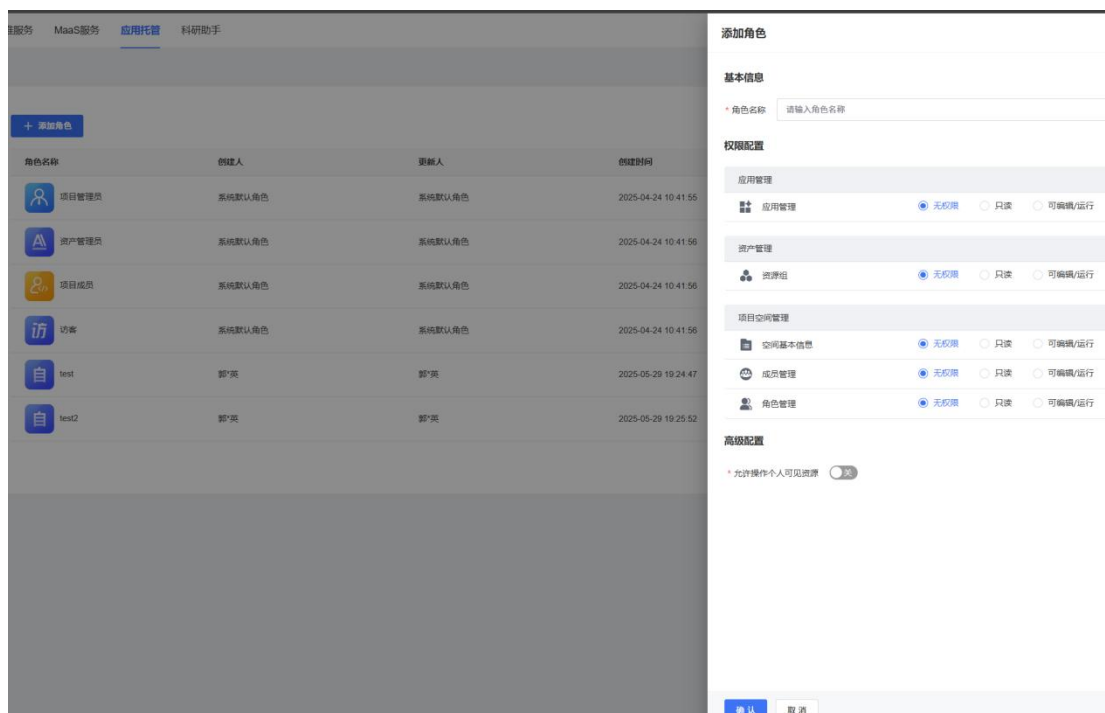
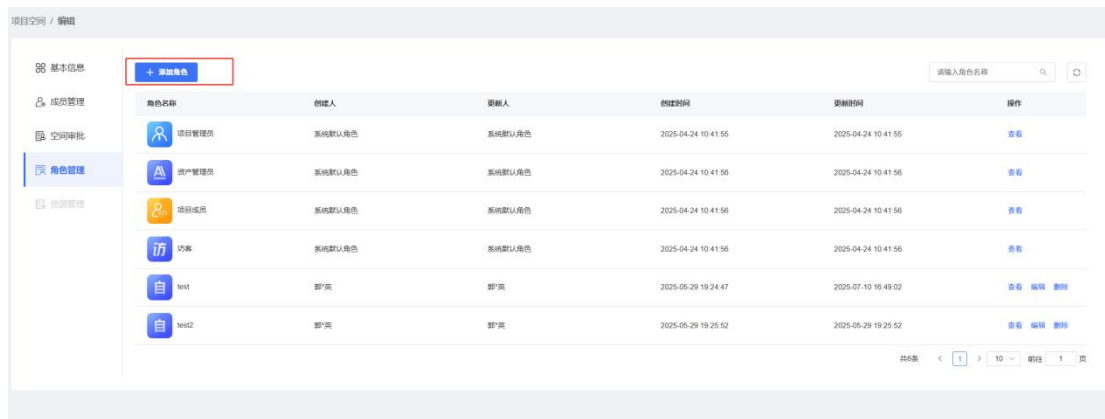
1.在项目空间编辑页面，点击【角色管理】进入项目空间角色管理页面，您可在【角色管理】添加和管理项目空间角色。



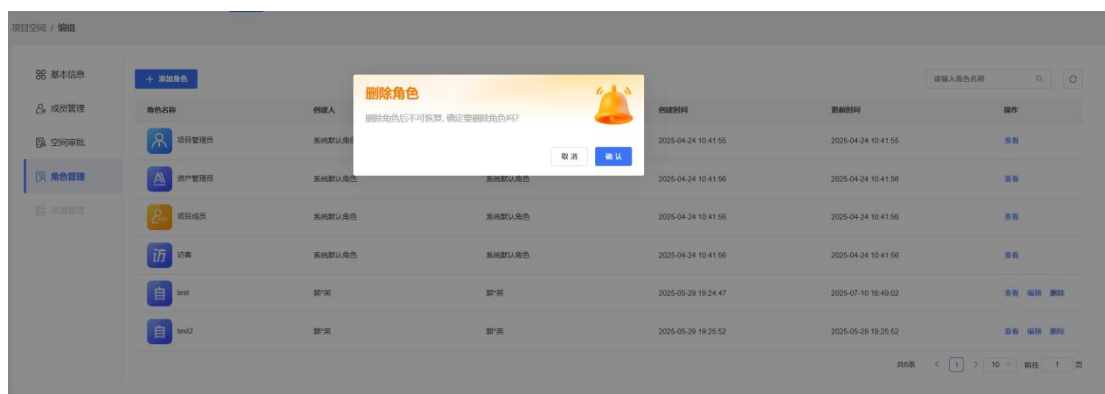
2.系统默认角色：系统会自动为项目空间创建【管理员】【项目成员】【资产管理员】【访客】四种默认角色。可以点击【查看】系统默认角色基本信息和权限信息，系统默认角色不支持编辑和删除。



3.添加角色：支持用户添加自定义角色，点击【添加角色】按钮，编辑角色名称和配置对应权限，点击【确认】即可完成添加自定义角色。自定义角色支持二次编辑和删除。

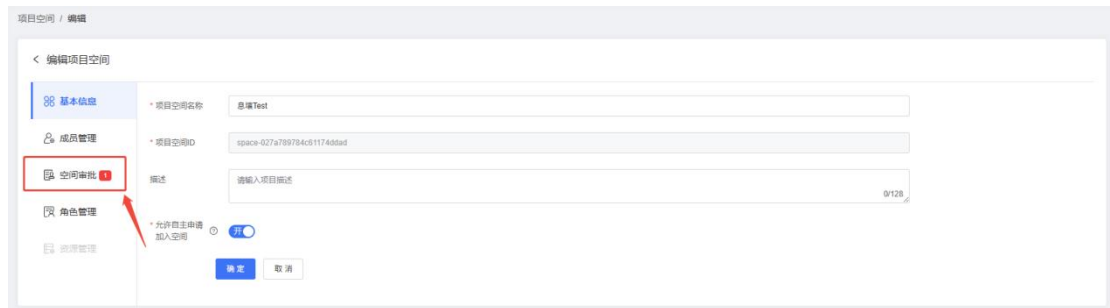


4.删除角色：支持用户删除项目空间自定义角色，点击【删除】按钮并进行二次确认即可删除角色。当角色下存在已关联项目空间成员时不可删除。



5.2.6 空间审批

1.在项目空间编辑页面，点击【空间审批】进入项目审批管理页面，您可在【空间审批】查看其他用户的空间加入申请并进行处理。



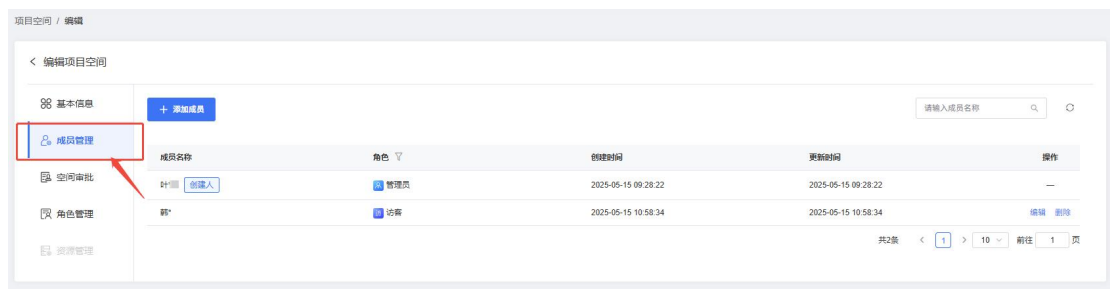
2.在项目空间审批列表，找到需要处理的申请消息，为用户选择项目空间角色，并点击【同意】即可允许该用户加入该项目空间。点击【拒绝】则会驳回该用户的申请。

3.可在【已审批】列表查看已处理的空间申请。



5.2.7 项目空间成员管理

1.在项目空间编辑页面，点击【成员管理】进入空间成员管理页面，您可在【成员管理】查看已加入空间的成员并进行管理。



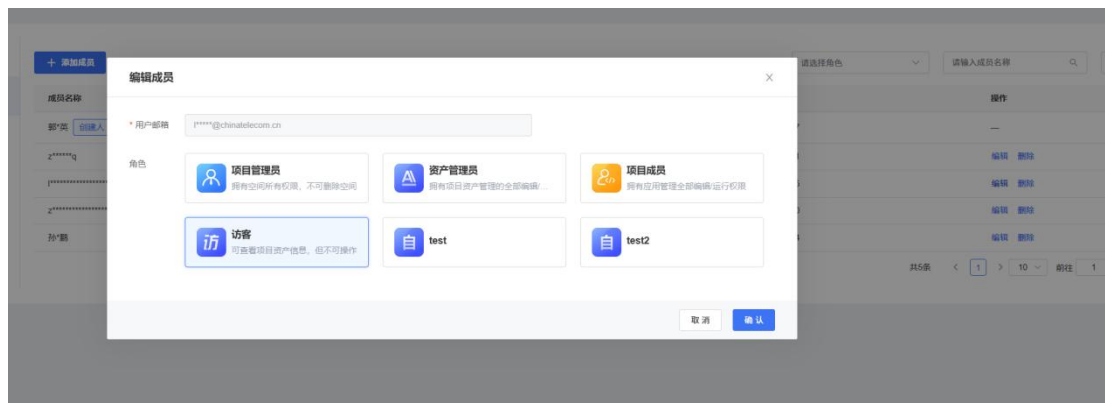
2.添加成员：点击【添加成员】按钮，输入用户邮箱并选择角色，点击【确认】即可邀请指定用户加入空间。



注意

项目空间内的成员可共享资源 and 数据, 请勿随意添加成员并为所有成员设置合理的项目空间权限，避免造成资产损失。

3.编辑成员：点击【编辑】按钮，可修改成员角色。



4.删除： 点击【删除】并输入确认信息可删除项目空间成员。



5.2.8 删除项目空间

1.在项目空间列表页， 点击项目空间卡片下的【删除】按钮。



2.系统会弹框确认信息进行二次确认， 点击【确认】按钮，删除项目空间



注意

- 项目空间删除后不可恢复，请谨慎操作。
- 系统默认空间不允许删除。

6.存储管理

6.1 创建存储

操作步骤

1.登录应用托管控制台，在左侧导航栏单击【资产】->【存储】，单击【创建存储】按钮。



2.在创建存储表单中，填写存储信息：存储名称、存储配置和可见范围。

字段	说明
存储名称	存储名称，名称长度为 4-20 个字符，支持小写字母、数字、中划线（-）和小数点（.）
付费类	当前为用户提供包年包月、按量付费

字段	说明
型	
可用区	选择存储所在的可用区
性能规格	选择所需要的性能规格
存储容量	需填写所需存储容量，存储创建完成后即按容量计费
可见范围	<p>存储实例的可见范围：</p> <p>仅自己可见：存储的创建人、项目空间所属人和开启“允许操作个人可见资源”权限的成员 可以查询到该资源</p> <p>空间内公开可见：项目空间内所有成员都可以查询到该存储</p>



3.完成表单填写后，单击按钮【确认创建】，完成存储创建。在存储列表可查询到已创建的存储。



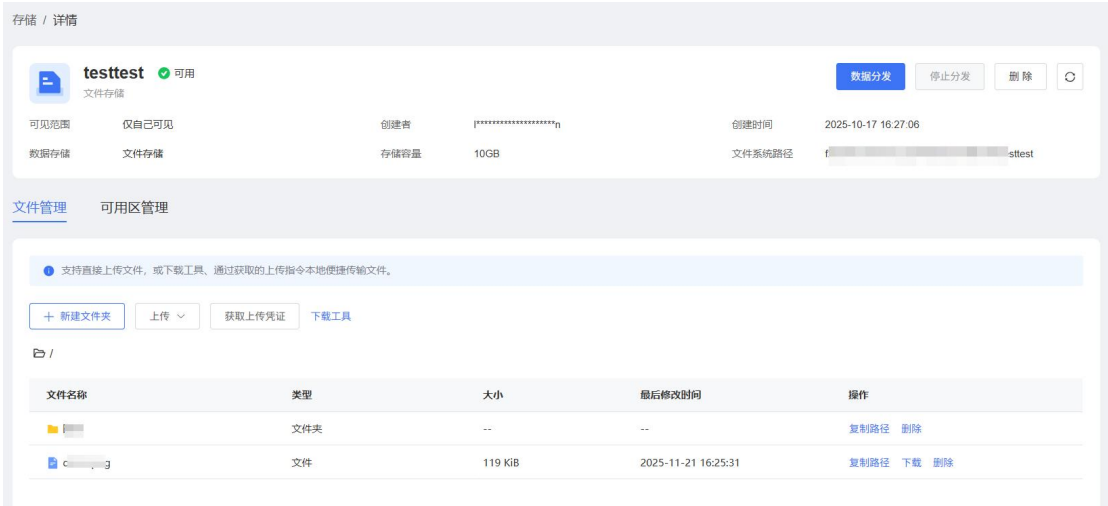
6.2 存储生命周期管理

查询存储详情

1.登录应用托管控制台，在左侧导航栏单击【资产】->【存储】， 查询存储列表。



2.单击存储“名称”进入存储详情。

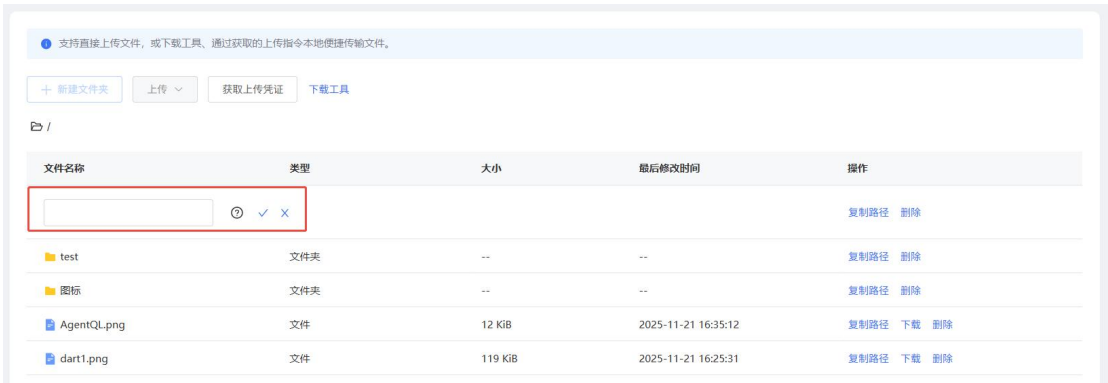


在存储内创建文件夹

1.在存储详情->文件管理，单击按钮【新建文件夹】。



2.对文件夹命名，单击右侧按钮【√】，完成文件夹新建。

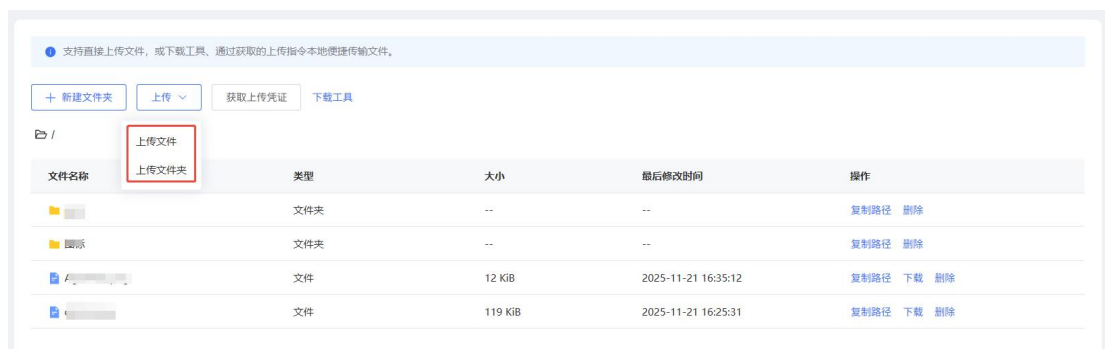


说明

- 1.不允许使用表情,请使用符合要求的 UTF-8 字符。
- 2.文件夹名称中不能使用/, 不要以\开头。
- 3.不能以英文句号.开头或结尾, 不允许出现..的文件夹名称。
- 4.总长度控制在 1-254 个字符。

存储内上传文件

- 1.在存储详情->文件管理，单击按钮【上传】->【上传文件】。

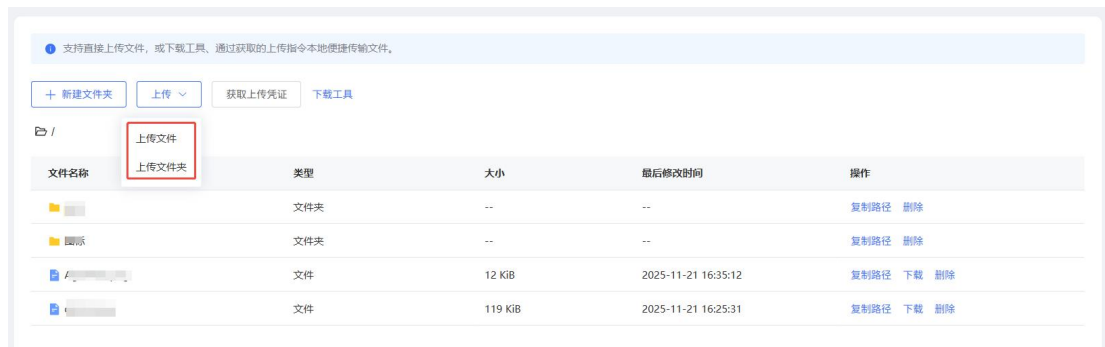


- 2.选择文件，单击按钮【打开】，页面右下角会自动显示“上传列表”弹窗。

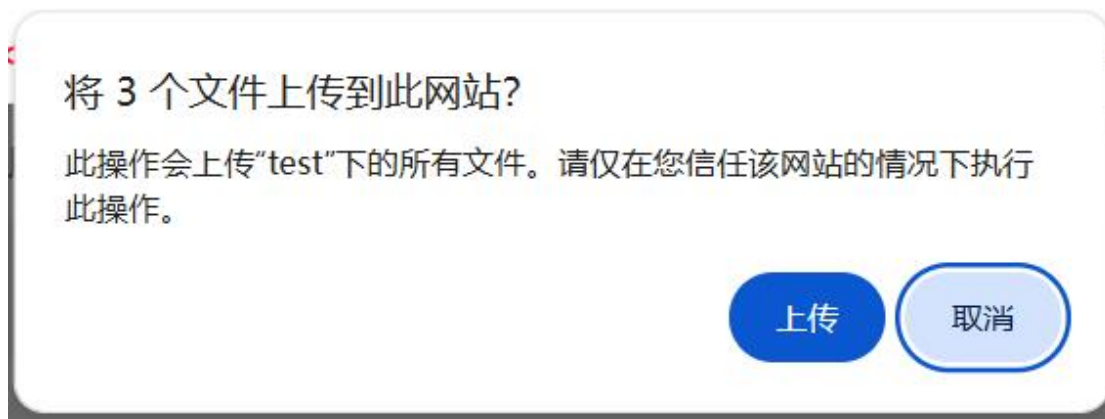


存储内上传文件夹

1.在存储详情->文件管理，单击按钮【上传】->【上传文件夹】。



2.选择文件夹，单击按钮【上传】，会弹框两次确认是否上传文件到网站，单击按钮【上传】。



3.页面右下角会自动显示“上传列表”弹窗。

上传列表

X

1

离开网站将会清空上传列表，任何正在上传的进程都将会被终止。

上传完毕：4/4

文件名	路径	进度	操作
service.yaml	test/service.yaml	已上传(275B)	
ng.yaml	test/ng.yaml	已上传(1.14KB)	
in.yaml	test/in.yaml	已上传(1.15KB)	
AgentQL.png	AgentQL.png	已上传(12.54KB)	

下载存储内文件

1.在存储详情->文件管理，单击指定文件的【下载】按钮。

支持直接上传文件，或下载工具、通过获取的上传指令本地便捷传输文件。

+ 新建文件夹

上传

获取上传凭证

下载工具

📁 /

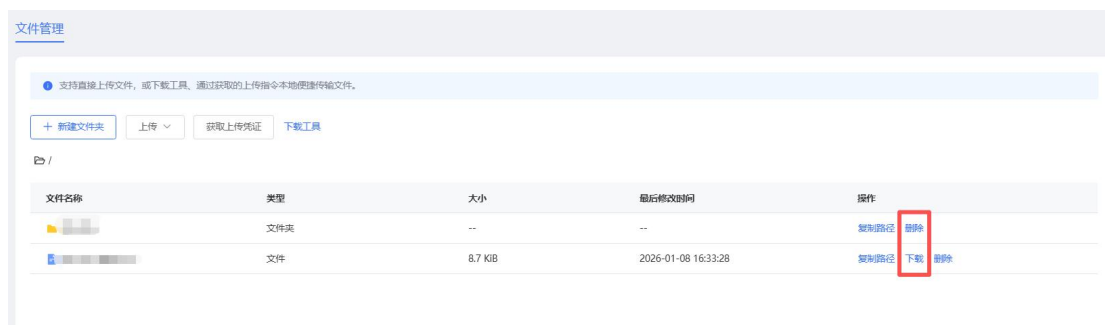
文件名称	类型	大小	最后修改时间	操作
test	文件夹	--	--	复制路径 删除
	文件夹	--	--	复制路径 删除
	文件	12 KiB	2025-11-21 16:35:12	复制路径 下载 删除
	文件	119 KiB	2025-11-21 16:25:31	复制路径 下载 删除

2.下载完成后可在本地设备中查询到下载的指定文件。



删除存储内的文件

1.在存储详情->文件管理，单击指定文件的【删除】按钮。



2.弹框的确认信息，单击按钮二次【确认删除】。



删除存储内的文件夹

在存储详情->文件管理，单击指定文件夹的【删除】按钮。

弹框的确认信息，单击按钮二次【确认删除】。



编辑存储

在存储列表，单击指定存储操作栏的【编辑】按钮。



在弹窗中可编辑存储的可见范围。

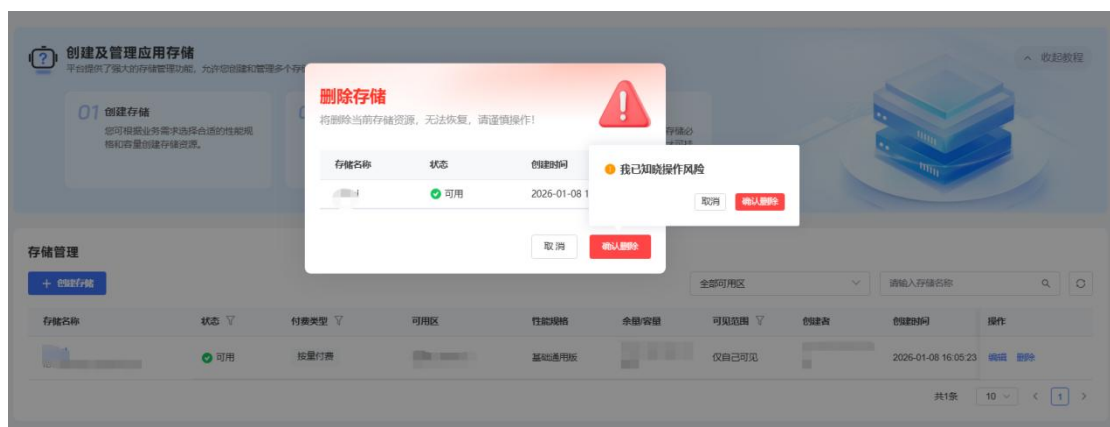


删除存储

在存储列表，单击指定存储操作栏的【删除】按钮。



弹框的确认信息，单击按钮二次【确认删除】。



6.3 文件上传工具

前提条件

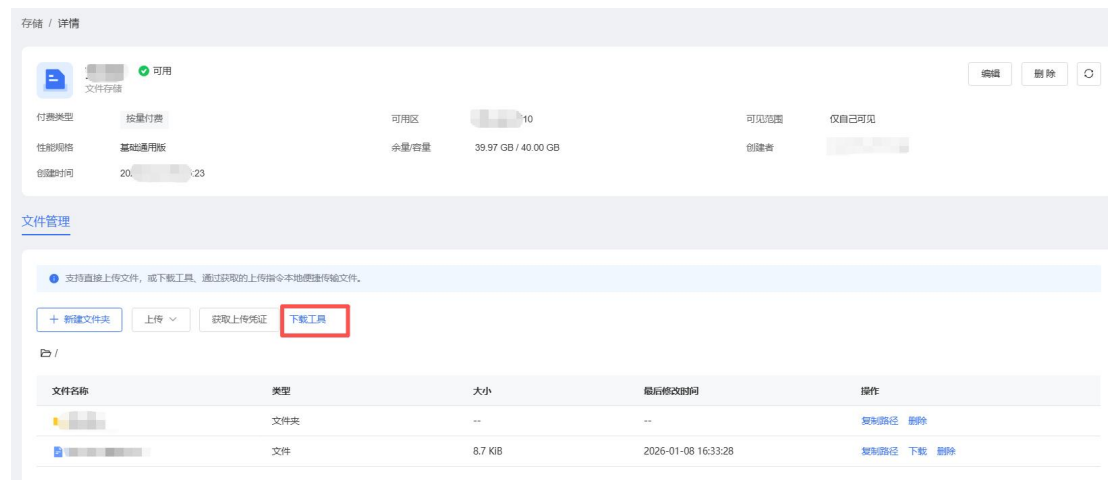
已创建应用托管存储实例。

拥有一台 Windows 或 Linux 系统的服务器。

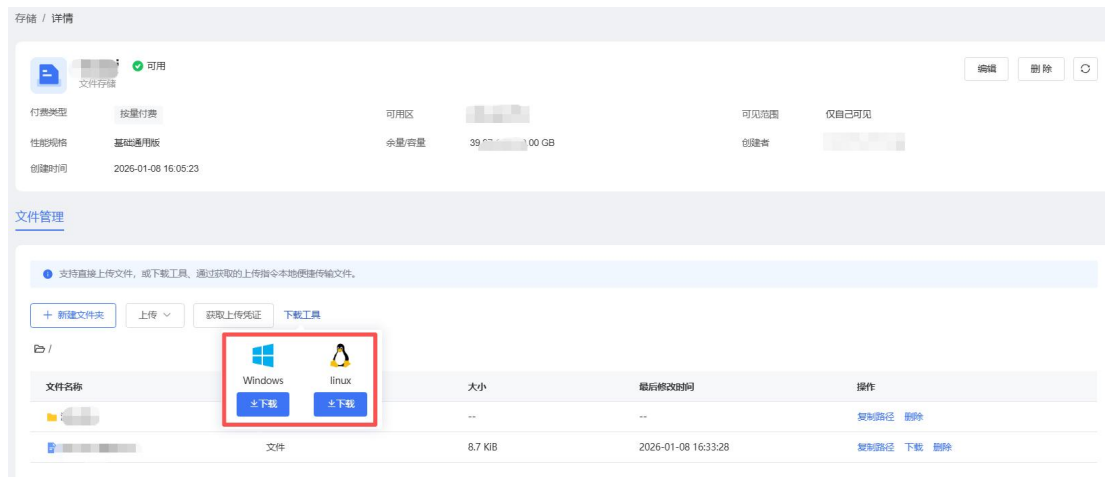
下载与安装

软件下载

登录应用托管控制台，在左侧导航栏单击【资产】->【存储】，单击存储实例名称进入详情。在详情页单击【下载工具】按钮。



根据不同平台需求，下载不同的软件。当前支持 Windows 和 Linux。

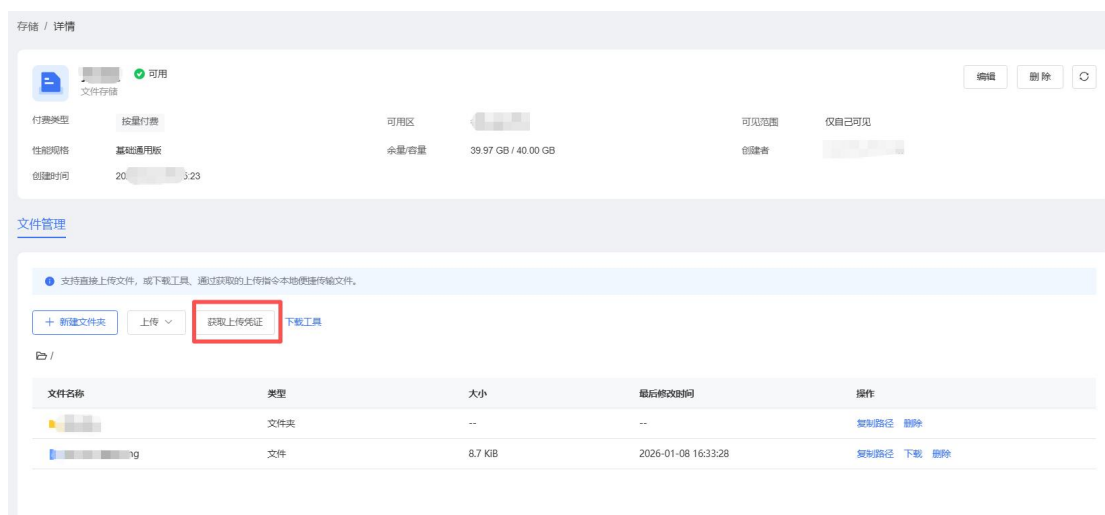


软件安装

用户通过程序包在本地完成软件安装。

获取上传凭证

登录应用托管控制台，在左侧导航栏单击【资产】->【存储】，单击存储实例名称进入详情。在详情页单击【获取上传凭证】按钮。



在“获取上传凭证”弹窗中查看获取上传凭证的操作说明。



天翼云账号中心获取 AccessKey

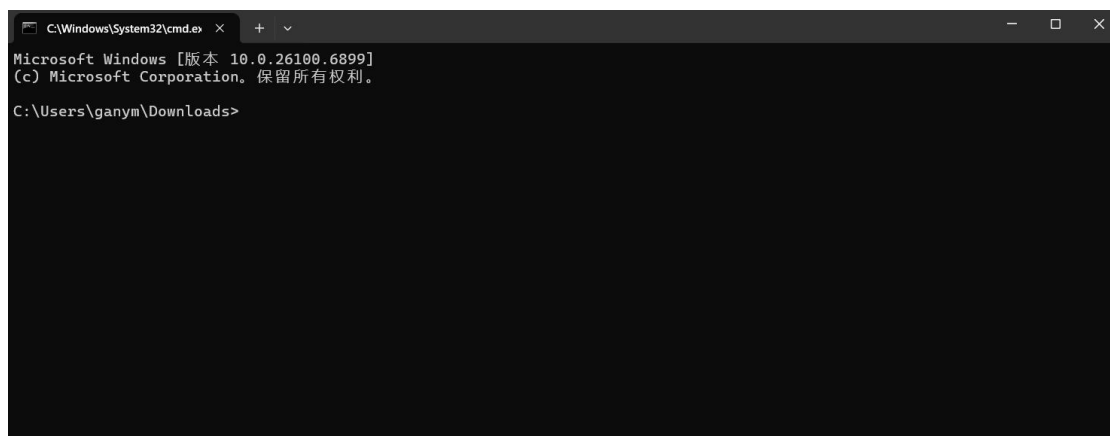
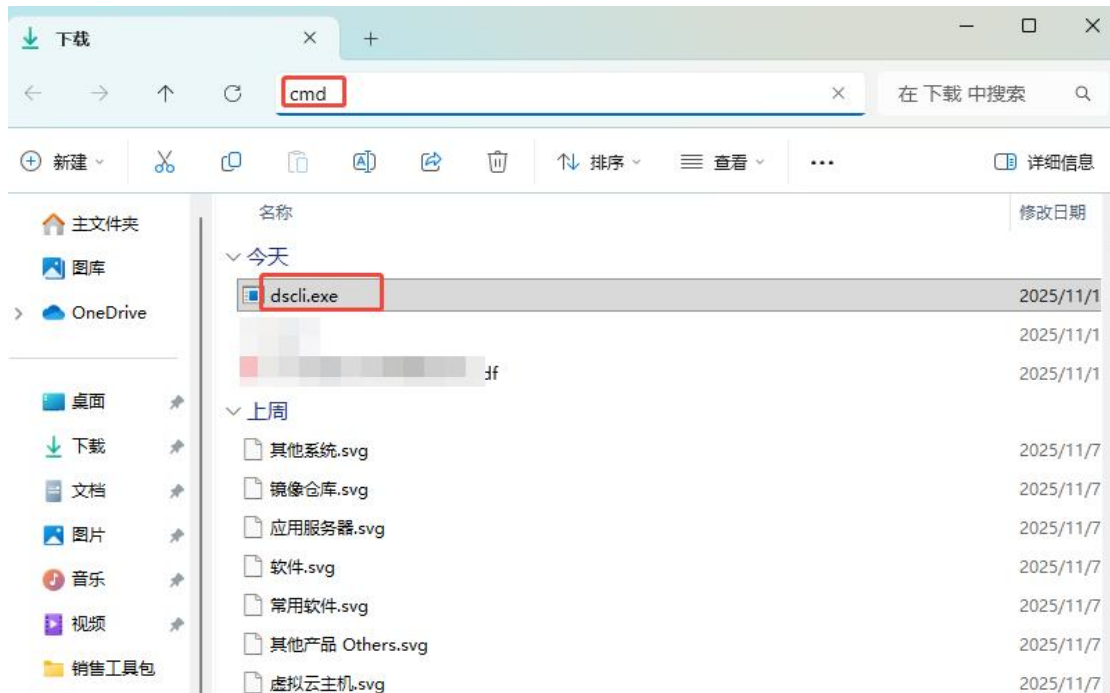
登录天翼云账号中心，在左侧导航栏单击【安全设置】进入安全设置，单击【登录保护】在“用户 AccessKey” 可查看和管理 AccessKey。



配置凭证

Windows

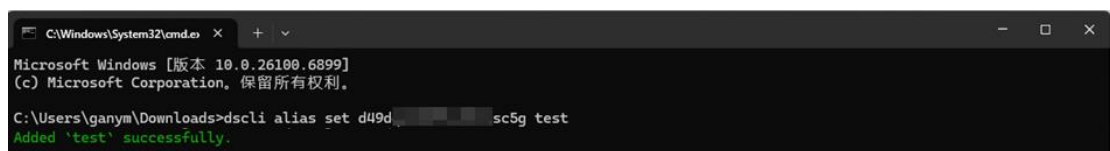
进入工具目录，输入 cmd，回车进入终端



在工具中执行以下指令，完成凭证配置。

```
dscli config <ak> <sk>
```

```
dscli alias set {存储 id} {别名}
```



Linux

在 Linux 服务器上，输入以下指令，赋予工具执行权限。

```
chmod +x dscli
```



从控制台获取上传凭证，执行以下指令，完成凭证配置。

```
dscli config <ak> <sk>

dscli alias set {存储 id} {别名}
```

上传文件

在已连接的工具，执行以下指令，实现上传文件。

```
dscli cp tmp.txt {别名}/tmp/tmp.txt
```

```
C:\Users\ganym\Downloads>dscli cp tmp.txt ganym/tmp/tmp.txt
...Downloads\tmp.txt: 0 B / ? [
```

查看文件列表

在已连接的工具，执行以下指令，实现查看文件列表。

```
dscli ls {别名}/{文件目录}
```

```
C:\Users\ganym\Downloads>dscli ls ganym/
[2025-11-11 14:52:34 CST]      0B STANDARD /
[2025-11-11 17:29:49 CST]      0B ganym/
[2025-11-11 17:29:49 CST]      0B tmp/
```

查看帮助

在已连接的工具，执行以下指令，实现查看帮助说明。

```
dscli alias set --help
```

```
C:\Users\ganym\Downloads>dscli alias set --help
NAME:
  dscli alias set - set a new alias to configuration file

USAGE:
  dscli alias set alias set <dataset_uuid> <alias_name>

FLAGS:
  --path value          bucket path lookup supported by the server. Valid options are '[auto, on, off]'
  (default: "auto")
  --api value           API signature. Valid options are '[S3v4, S3v2]'
  --cred value          credential from console
  --config-dir value, -C value path to configuration folder (default: "C:\\Users\\ganym\\dscli") [%MC_CONFIG_DIR%]
  --quiet, -q          disable progress bar display [%MC_QUIET%]
  --disable-pager, --dp disable mc internal pager and print to raw stdout [%MC_DISABLE_PAGER%]
  --no-color           disable color theme [%MC_NO_COLOR%]
  --json              enable JSON lines formatted output [%MC_JSON%]
  --debug             enable debug output [%MC_DEBUG%]
  --resolve value      resolves HOST[:PORT] to an IP address. Example: minio.local:9000=10.10.75.1
  [%MC_RESOLVE%]
  --insecure           disable SSL certificate verification [%MC_INSECURE%]
  --limit-upload value limits uploads to a maximum rate in KiB/s, MiB/s, GiB/s. (default: unlimited)
  [%MC_LIMIT_UPLOAD%]
  --limit-download value limits downloads to a maximum rate in KiB/s, MiB/s, GiB/s. (default: unlimited)
  [%MC_LIMIT_DOWNLOAD%]
  --custom-header value, -H value add custom HTTP header to the request. 'key:value' format.
  --help, -h          show help
```

7.应用门户

概述

平台为用户提供企业级应用门户，作为企业应用的统一访问入口，整合部署完成的各类应用
和工具，赋能业务高效推进。用户可将正常运行的应用访问方式添加到应用门户，实现快捷
访问。

说明：该功能目前正处于内测阶段，暂未面向所有用户开放。如果您想成为内测用户，提前
体验这些新功能，请联系您的客户经理或通过工单提交申请。

上架门户应用

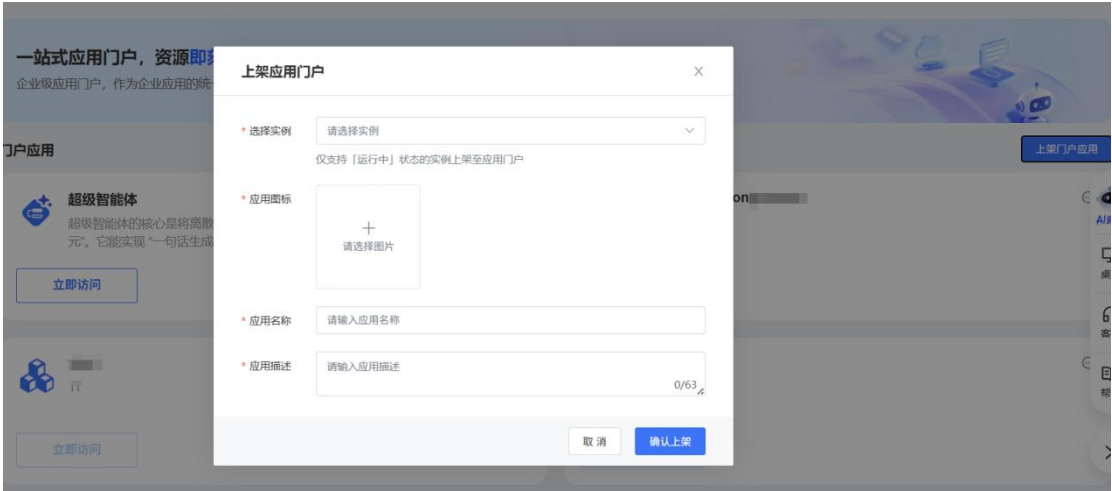
登录应用托管控制台，在左侧导航栏单击【应用】->【应用门户】，单击【上架门户应用】
按钮。



在“上架门户应用”弹窗表单中，填写上架应用信息：应用实例、应用图标、应用名称和应
用描述。

字段	说明
选择实例	仅支持「运行中」状态的应用实例上架至应用门户

字段	说明
应用图标	默认使用应用实例图标，支持修改
应用名称	默认使用应用实例名称，支持修改
应用描述	输入应用相关说明



完成表单填写后，单击按钮【确认上架】，完成应用上架。在应用门户可查询到已上架的应用。



访问应用



登录应用托管控制台，在左侧导航栏单击【应用】->【应用门户】，单击目标应用的【立即访问】按钮，即可访问应用。



注意：当部署的应用实例出现异常，门户应用将不可访问。如应用不可访问，请检测实例状态。

下架门户应用

登录应用托管控制台，在左侧导航栏单击【应用】->【应用门户】，单击目标应用的【下架门户应用】按钮。



在“下架门户应用”弹窗中，单击【确认下架】按钮，完成应用下架。



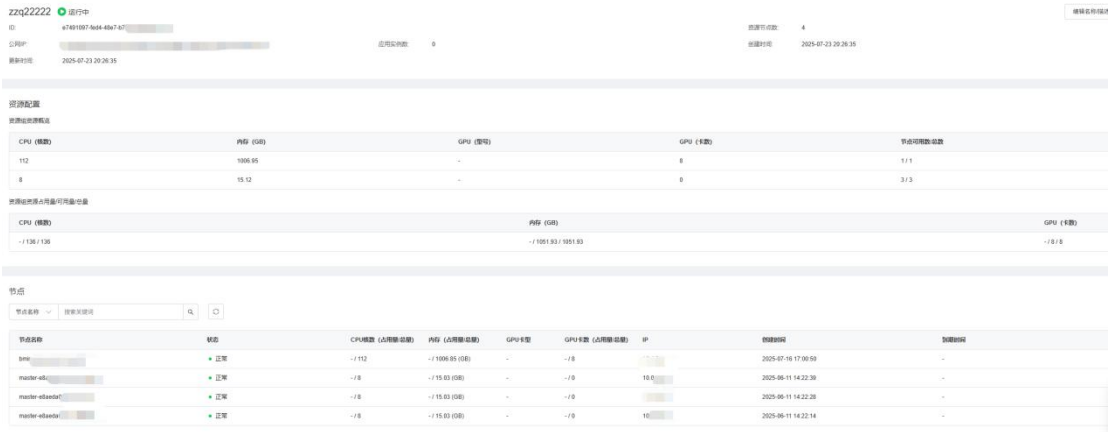
8.资源组管理

8.1 查看资源组

1.登录应用托管控制台，在左侧导航栏单击【资源组管理】，进入资源组列表，点击资源组名称或详情进入详情页。



2.支持查看集群的基础信息、资源配置消息和节点信息



8.2 管理资源组

当前尚未提供自助资源组新建能力，如果您想基于资源组部署独享资源应用，可通过联系客服经理或工单申请。

五、常见问题

1. 计费类

支持哪些资源池？

公共资源池：目前控制台开放了多个可用区和多种实例规格，如规格或型号不满足您的需求，

您可联系您的客户经理，或线上咨询或拨打客服热线电话寻求帮助，热线电话：

400-810-9889 转 1。后续会随业务需求和资源布局会逐步开放自助区域。

是否支持试用？

可针对客户提供试用，如有您试用需求请联系您的客户经理进行申请。

支持哪些计费方式？

- 按需计费
- 包周期

应用实例停止是否需要收费？

应用实例停止后释放占用的算力资源，不收取费用；存储和网络收费。

收取哪些费用？

- 在应用托管部署应用创建应用实例后，会根据申请的资源规格内容进行收费，例如一个实例会包含 CPU、内存、GPU、网络 and 存储等费用。
- 如用到天翼云其他产品，需另外支付对应产品费用。

2. 操作类

如何部署应用？

方式 1：通过应用市场或 MCP 市场浏览选择所需的应用，一键部署应用，支持为部署的应用提供公网访问能力。

方式 2：通过应用实例菜单，创建自定义应用，自定义部署参数后完成应用部署。

如何管理应用？

通过应用实例管理已部署的应用，支持应用的访问、启停、更新、扩缩容等操作，支持资源实例的终端登录，支持查看日志、事件和监控。

购买算力时，资源已经售罄怎么办？

如果在部署应用时，所需算力资源时遇到资源售罄，建议您根据需要调整需求：

- 更换可用区
- 更换算力规格

平台会视情况补充资源，如没有合适的算力可联系您的客户经理，或者提交客服工单，亦或直接拨打客服热线。

为什么我能查询到的算力比较少？

算力规格和资源池未针对用户全网开放，如没有合适的算力请联系您的客户经理，或者提交客服工单，亦或直接拨打客服热线。

3. 技术运维类

3.1 应用部署常见问题

使用自定义镜像部署应用，应用长时间处于创建中

1. 用户自有镜像可能容量大，需要较长的镜像拉取时间，请耐心等待，可在实例事件中查询部署情况。
2. 用户镜像地址填写错误，拉取不到镜像，在事件中会展示镜像拉取失败，请检查镜像地址并更新。
3. 用户使用私有镜像仓库，没有填写访问凭证或访问凭证填写错误，拉取不到镜像，在事件中会展示镜像拉取失败，请更新正确的访问凭证。

使用自定义镜像部署应用，应用实例不断重启

1. 用户没有填写启动命令，可在实例日志中查看错误日志信息，更新组件填写正确的启动命令。
2. 启动命令填写错误，可在实例日志中查看错误日志信息，更新组件填写正确的启动命令。

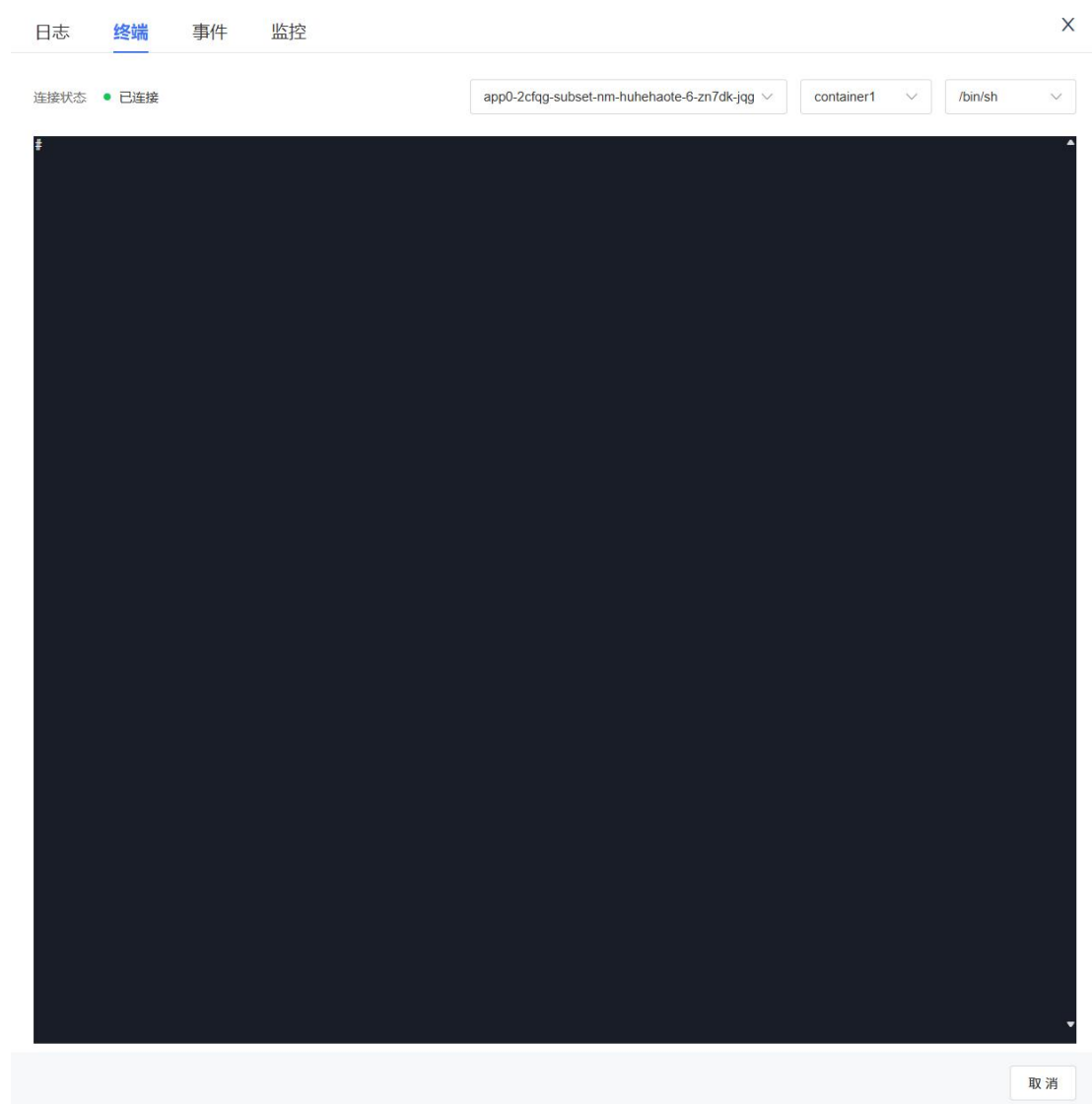
部署的应用状态运行中，却无法通过公网访问

1. 用户未开启公网访问能力，请更新应用的网络配置，开放公网。
2. 用户部署的应用非 web 类应用（如 MCP Server），无法通过浏览器访问，非 web 类应用请通过其他方式访问。
3. 应用访问策略未放通访问 IP，请在应用访问策略放通访问 IP。
4. 用户容器端口填写错误，请更新应用的网络配置，修正错误端口。

3.2 调试和排查常见问题的方法

进入终端调试

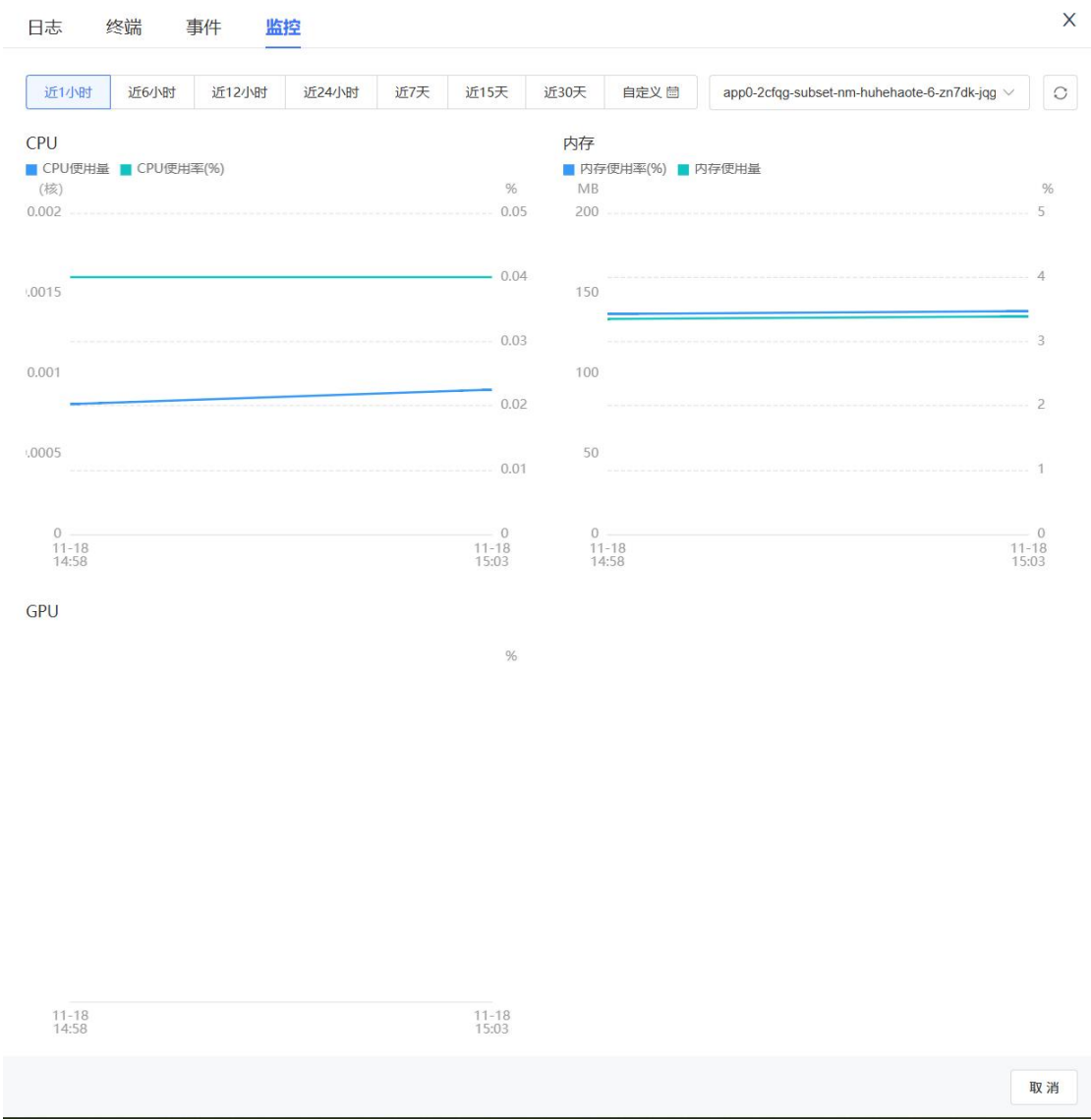
- 1.选择目标应用和组件，进入组件详情。
- 2.点击实例的【终端】进入到资源的控制台页面，支持输入命令进行故障的排查。



查看实例监控

登录应用托管控制台，在左侧导航栏单击【应用实例】，点击应用实例名称进入详情页。

选择组件，点击【监控】，可查看该应用实例的监控信息。



查看实例日志

登录应用托管控制台，在左侧导航栏单击【应用实例】，点击应用实例名称进入详情页。

选择组件，点击【日志】，可查看该应用实例的日志信息，支持切换实例。

连接状态 ● 正在连接中

app0-2cfs0-subset-nm-huhehaote-6-6cfkh-xdh

container1

最近500条

☐

容器退出日志

☒

自动滚动

☒

自动刷新(每5秒)



1 加载中...



取消

查看实例事件

登录应用托管控制台，在左侧导航栏单击【应用实例】，点击应用实例名称进入详情页。

选择组件，点击【事件】，可查看该应用实例的最近一个小时事件信息，支持切换实例。

展示最近 1 小时产生的事件数据

app0-2cfqg-subset-nm-huhehaote-6-zn7dk-jqg

搜索关键词



事件名称	类型	事件描述	发生次数	最近发生时间	首次发生时间
容器调度完成	正常	Successfully assigned cps-23/app0-2cfqg-subset-nm-huhehaote-6-zn7dk-jqgqn to vk-nm-huhehaote-6-f26e787a911c4cceb168b7307a66776	--	2025-11-18 14:46:01	2025-11-18 14:46:01
资源准备完成	正常	Create pod in provider successfully	1	2025-11-18 14:46:08	2025-11-18 14:46:08
拉取镜像中	正常	Pulling image ***	1	2025-11-18 14:54:03	2025-11-18 14:54:03
镜像拉取完成	正常	Successfully pulled image *** in 1m28.991613488s	1	2025-11-18 14:55:32	2025-11-18 14:55:32
容器创建中	正常	Created container container1	1	2025-11-18 14:55:37	2025-11-18 14:55:37
容器启动	正常	Started container container1	1	2025-11-18 14:55:37	2025-11-18 14:55:37

取消

查看变更记录

登录应用托管控制台，在左侧导航栏单击【应用实例】， 点击应用实例名称进入详情页。

点击【变更记录】按钮，可查看该应用实例的操作变更记录。

应用实例 / 公共应用详情

mcp-cjc-394891

ID: d4tsg6nqgo550j6bxa0

部署中

访问策略

停止应用

更新应用

变更记录

删除应用

付费类型

按量计费

可用区

内蒙呼和浩特5-esk-准生产-分布式推理

访问策略

app-default-sg

网络类型

独享网络

访问域名

—

应用信息

mcp-cjc-394891

应用版本

1.0.0

创建时间

2025-12-11 14:43:19

更新时间

2025-12-11 14:43:19

共1个组件

mcp-test

ID: d4tsg6nqgo550j6bxa0

查看启动中

基础信息

弹性伸缩

配置详情

算力规格

CPU 2 核 | 内存 4 GB

内网地址

aa | 内蒙呼和浩特5-esk-准生产-分布式推理 | 198.19.0.261...

外网地址

aa | 内蒙呼和浩特5-esk-准生产-分布式推理 | 182.42.228.171.61...

创建时间

2025-12-11 14:43:19

更新时间

2025-12-11 14:43:26

实例管理

批量重建

搜索容器关键词

实例名称	状态	CPU使用率	内存使用率	GPU使用率	创建时间	操作
app0-0sglbox0-subset-nm-huhehaote-6-bh69g-r7g6d	等待中	—	—	—	2025-12-11 14:43:26	日志 终端 事件 更多

监控 重建

重建实例

选择目标应用和组件，可查看该组件实例的部署实例资源信息，确认实例状态。

应用实例 / 公共应用详情

mcp-cjc-394891

ID: d4tsg6nqgo550j6bxa0

部署中

访问策略

停止应用

更新应用

变更记录

删除应用

付费类型

按量计费

可用区

内蒙呼和浩特5-esk-准生产-分布式推理

访问策略

app-default-sg

网络类型

独享网络

访问域名

—

应用信息

mcp-cjc-394891

应用版本

1.0.0

创建时间

2025-12-11 14:43:19

更新时间

2025-12-11 14:43:19

共1个组件

mcp-test

ID: d4tsg6nqgo550j6bxa0

查看启动中

基础信息

弹性伸缩

配置详情

算力规格

CPU 2 核 | 内存 4 GB

内网地址

aa | 内蒙呼和浩特5-esk-准生产-分布式推理 | 198.19.0.261...

外网地址

aa | 内蒙呼和浩特5-esk-准生产-分布式推理 | 182.42.228.171.61...

创建时间

2025-12-11 14:43:19

更新时间

2025-12-11 14:43:26

实例管理

批量重建

搜索容器关键词


实例名称	状态	CPU使用率	内存使用率	GPU使用率	创建时间	操作
app0-0sglbox0-subset-nm-huhehaote-6-bh69g-r7g6d	等待中	—	—	—	2025-12-11 14:43:26	日志 终端 事件 更多

监控 重建

点击【重建】，可手动对实例进行重启操作。

重建确认

重建实例，可能会导致系统盘中临时数据丢失。



取消

确认

六、最佳实践

1. 基于 Dify 和 DeepSeek 构建本地知识库的最佳实践

在本教程中, 我们将为您详细介绍基于 Dify 和 DeepSeek-R1 模型搭建私有知识库的方案。

利用该方案, 您可以在本地便捷地搭建起相应的私有知识库, 大幅提升文本检索能力。

概述

前言

大模型私有知识库, 作为大语言模型技术与企业、组织自有数据深度融合的创新知识管理及应用解决方案, 能够为特定用户群体提供更为精准、专业且安全的知识服务。具体来讲, 它是借助大语言模型搭建而成, 专门为特定组织或个人定制的知识存储与检索系统。此系统会对组织内部的专业知识、业务数据、历史文档等各类信息进行深度整合与精细化处理, 从而构建出独一无二的专属知识集合。依托大模型强大的语言理解与生成能力, 用户能够在此基础上实现高效的知识查询与问答交互。

在本教程中, 我们将为您详细介绍基于 Dify 和 DeepSeek-R1 模型搭建私有知识库的方案。

利用该方案, 您可以在本地便捷地搭建起相应的私有知识库, 大幅提升文本检索能力。

DeepSeek-R1 介绍

DeepSeek-R1 是幻方量化旗下 AI 公司深度求索 (DeepSeek) 研发的一款高性能推理模型。该模型使用强化学习技术进行后训练, 专注于提升在数学、代码和自然语言推理等复杂任务上的表现。



本最佳实践在算力互联调度平台的应用市场中，您可快速部署应用，并对外提供 API 调用服务，方便您即刻体验，开箱即用。

Dify 介绍

Dify 是一站式大模型应用开发平台。它能便捷接入各类数据并管理，支持模型微调与多模型调用。通过可视化界面和低代码工具，用户无需深厚技术背景，就能快速搭建如聊天机器人、智能客服等应用。可用于客户服务、内容创作等场景，具有降低开发门槛、提升效率、灵活可扩展等优势，助力企业高效落地大模型应用。



前置说明

该文档为应用托管控制台通过应用市场体验 Dify+DeepSeek-R1 模型搭建私有知识库的说明，通过应用托管控制台，打开应用市场，创建对应的 Dify 应用，使用 Dify 构建本地知识库。



本产品中的模型由第三方主体提供，尽管云公司已尽最大努力进行识别和维护，但仍无法保证模型的可用性。请客户按照该产品的服务协议使用该产品，做好甄别工作，并对自行选择的服务承担相应责任。

使用前准备

天翼云账号注册：使用应用托管平台须具备天翼云官网账号。已有天翼云账号的直接登录即可，如无天翼云账号需先注册，注册流程可参考：注册账号。

使用前提：如需使用服务请先完成实名认证，请参考账号中心-实名认证。如需使用按需服务，请确认账号余额 ≥ 100 元。

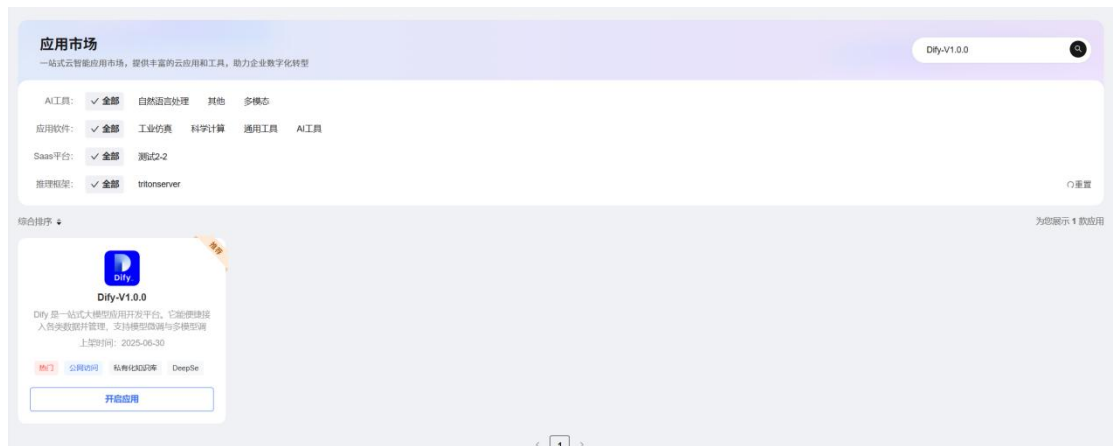
操作步骤

部署 Dify 应用

订购应用

登录应用托管控制台，选择左侧的应用市场点击进入页面。

在应用市场页面，选择目标应用，点击【开启应用】，进入应用部署页面。



填写管理密码，选择访问策略及公网访问，勾选同意用户协议，点击【开启应用】，支付订单，即可完成 Dify 应用服务部署。



服务详情查看

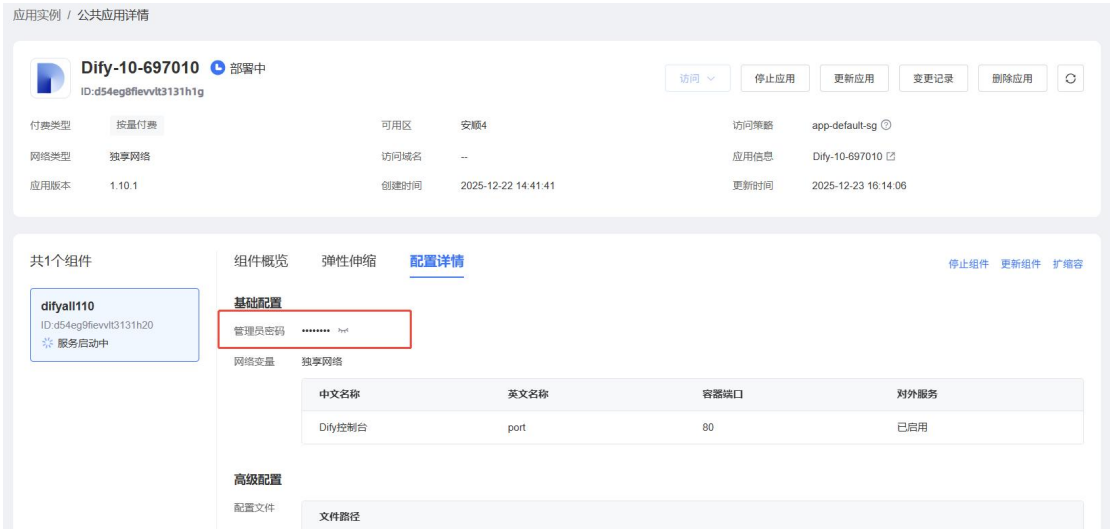
进入应用托管租户控制台。

选择【应用实例】，进入应用实例列表页面。

刷新应用实例列表，等待新创建的应用进入到【运行中】状态后点击右侧操作栏【访问】，点击下拉访问链接。



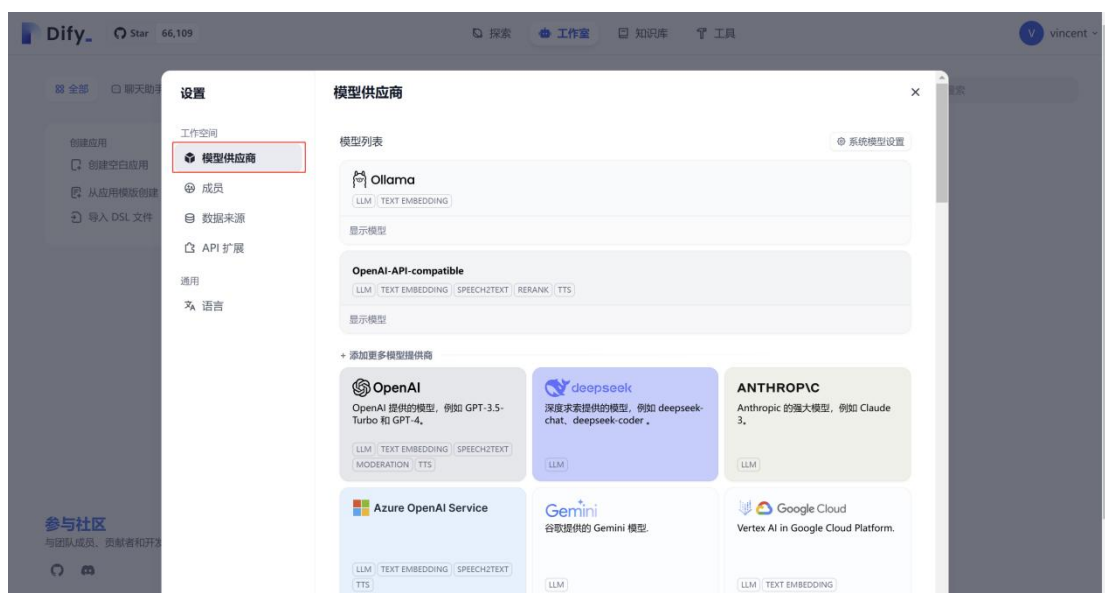
进入 Dify 应用界面。可使用管理员账号和设置的密码登录 Dify，登陆后用户可自行进行更改密码。账号：admin@admin.com；密码：可点击“应用实例”进入详情页，在“配置详情”中的“管理员密码”查看。



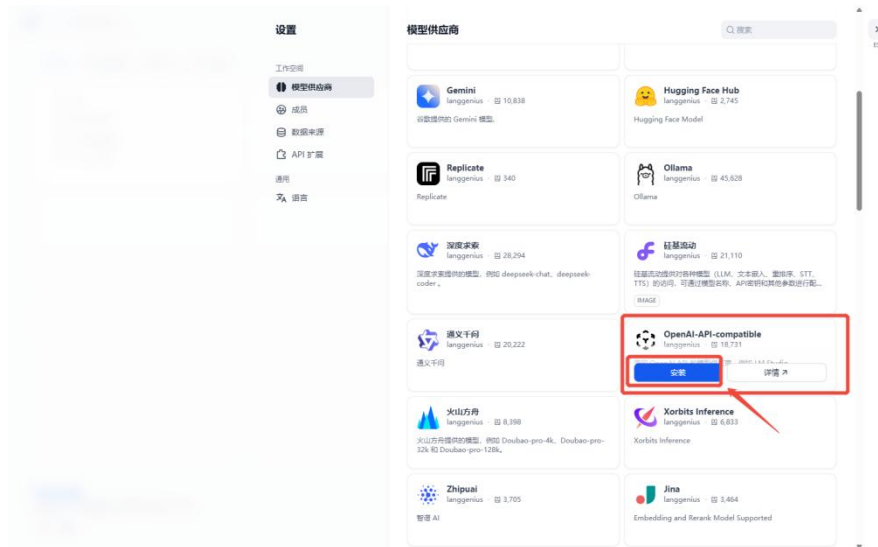
在 Dify 中添加 DeepSeek 模型

Dify 支持多模型管理与调用，可以便捷的添加我们所需的模型，我们这里以使用 ollama 添加 DeepSeek 模型为例。

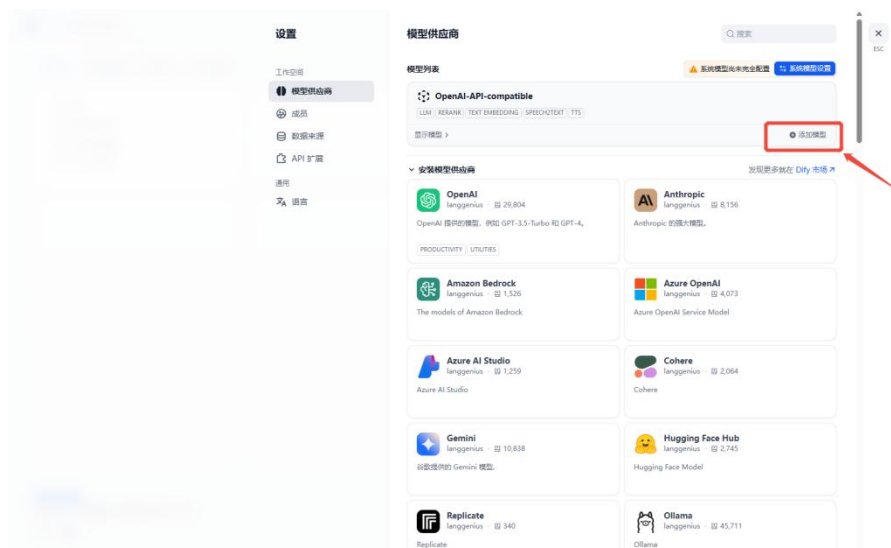
在 Dify 控制台点击设置进入模型供应商编辑页面。

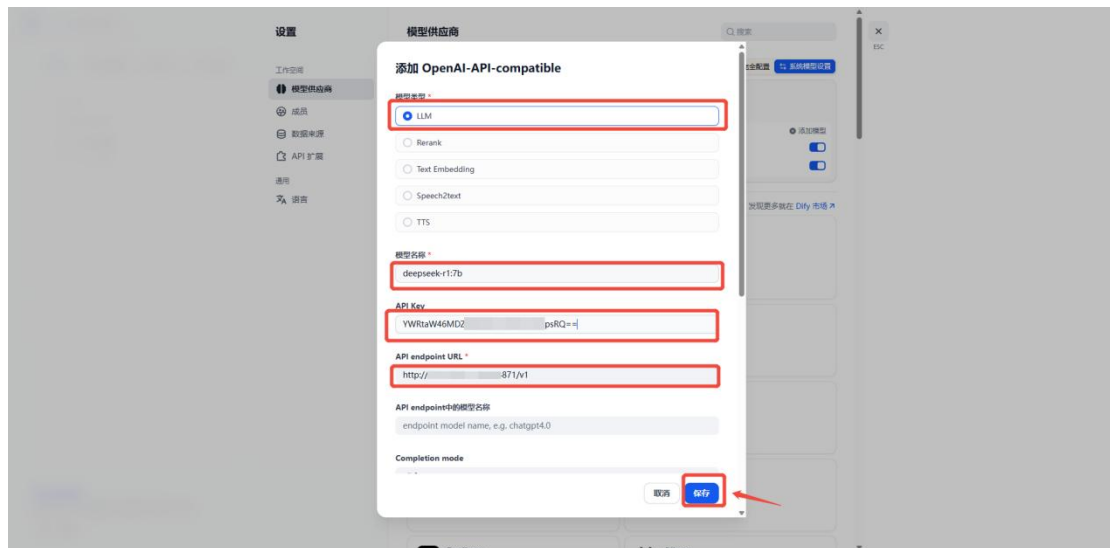


通过 OpenAI-API-compatible 分别添加一个 LLM 模型和 Text Embedding 模型，Text Embedding 模型用于知识库的向量化处理。

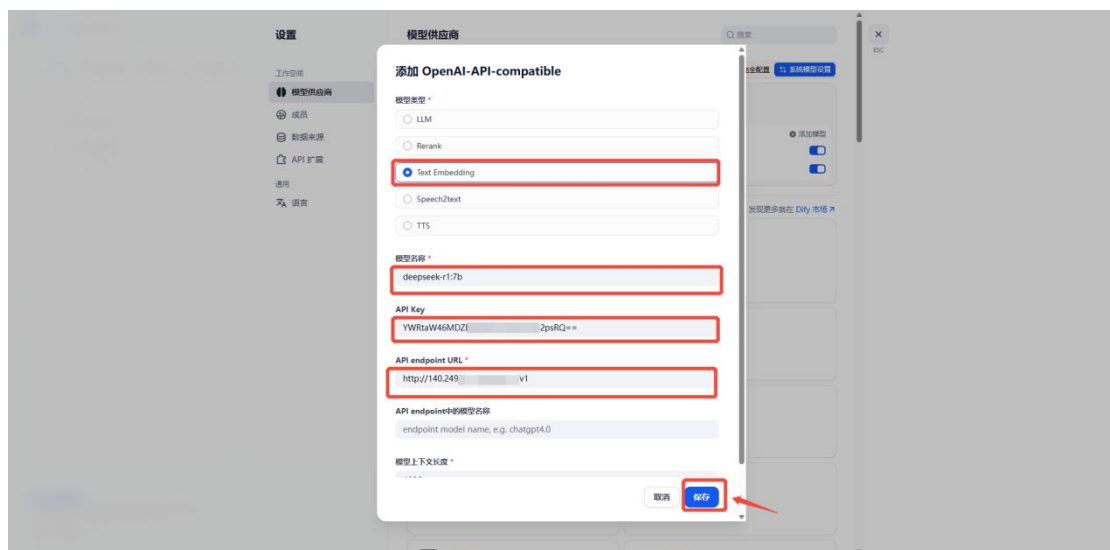


添加 LLM 模型，输入 ollama 部署的 DeepSeek 模型名称、基础 URL 和 API KEY，点击保存添加模型。输入示例，实际使用时请按情况替换：模型名称：deepseek-r1:7b；API KEY：YWRta****dPb2psRQ==；基础 URL：需在调用地址后增加后缀/v1。





同样的方式，添加一个 Text Embedding 模型用于文档向量化处理。Text Embedding 模型可以与 LLM 模型使用同一个模型和基础 URL。



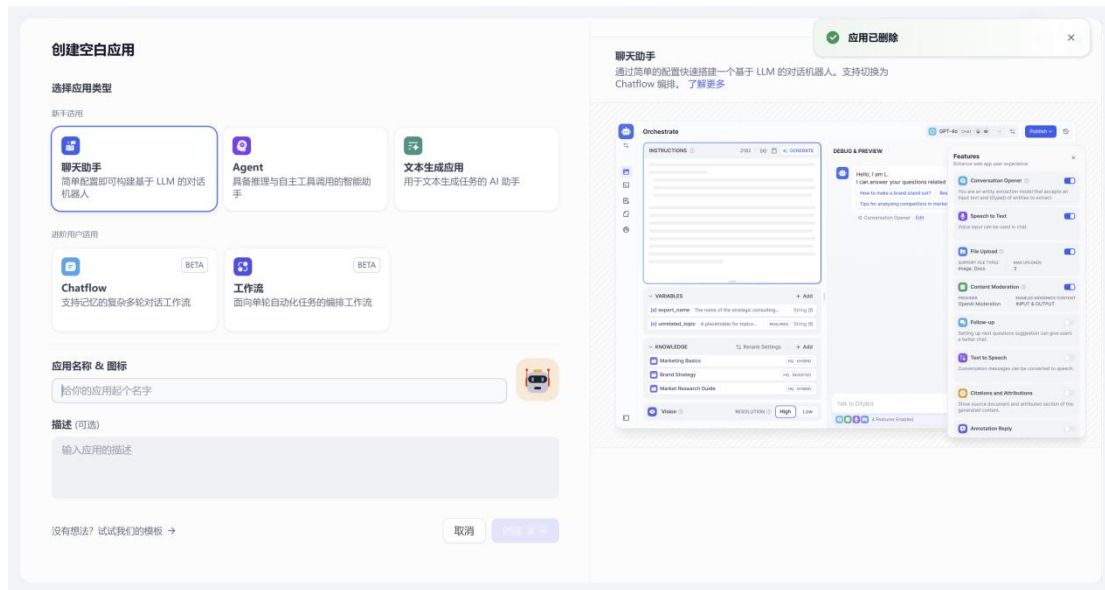
在系统模型配置页面设置系统推理模型和 Embedding 模型，一般系统会默认选择第一个可用模型，若模型显示为空，请刷新页面后重试。

系统模型配置完毕后，即可在应用编排界面配置模型参数和使用模型。模型参数会影响模型的输出效果，用户可自行探索，这里不再详细描述。

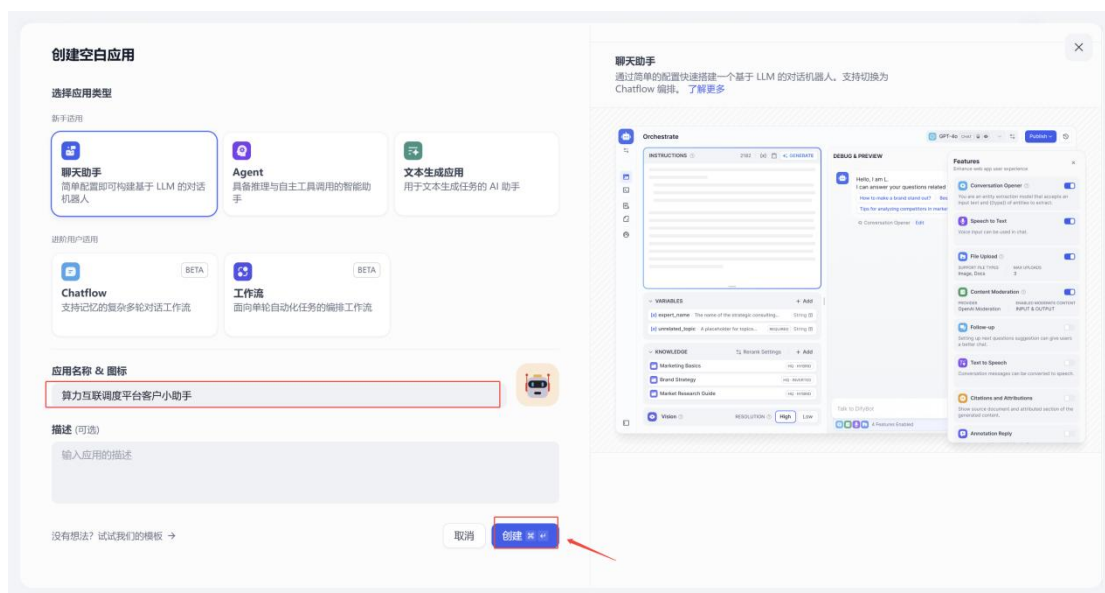


在 Dify 中搭建本地知识库

使用 Dify 的 RAG 管道功能，企业可以快速搭建本地知识库，系统会基于本地知识库返回准确的答案。搭建本地知识库前请先完成前置步骤：使用 Ollama 添加 DeepSeek 模型。

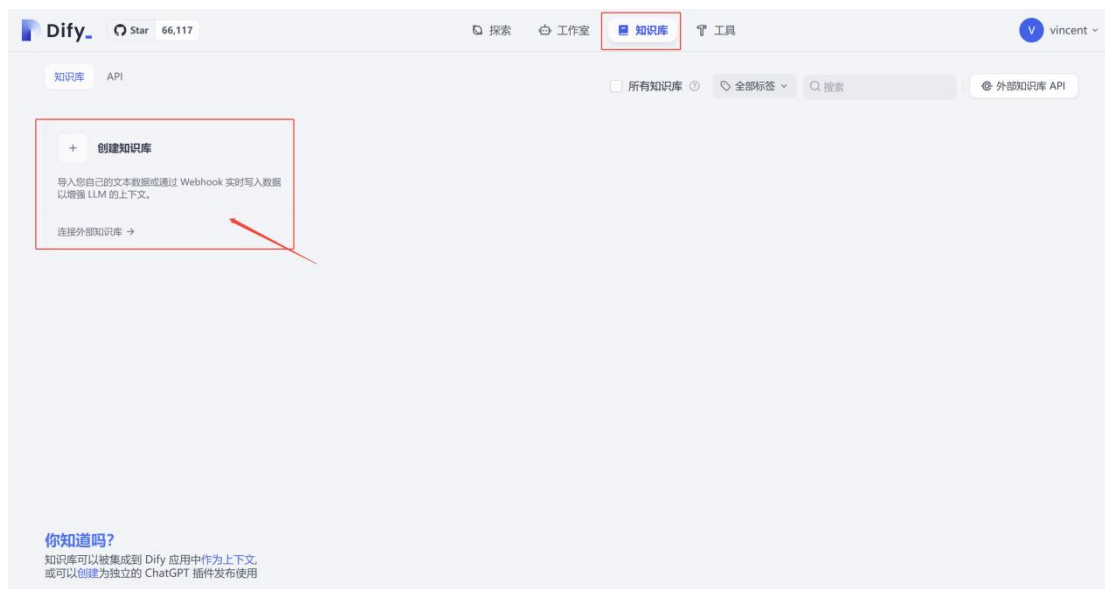


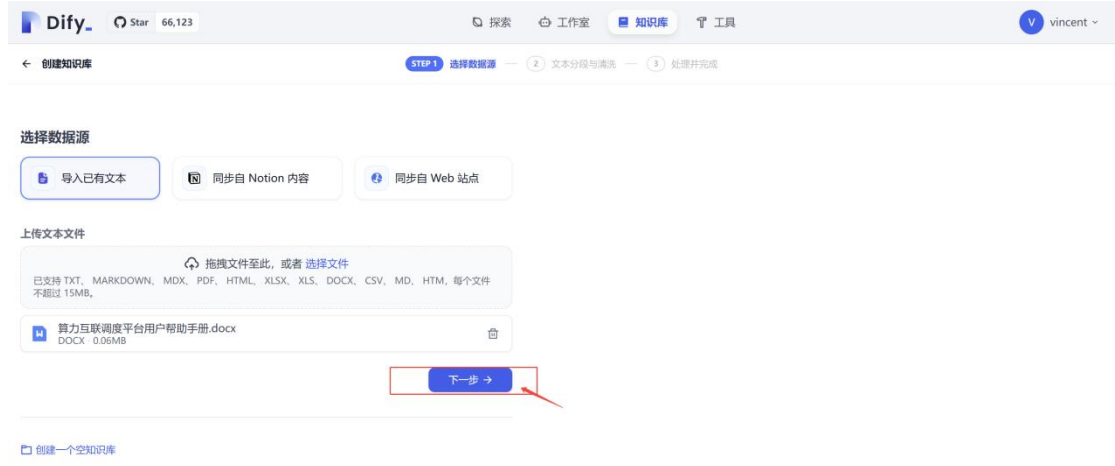
创建聊天助手：选择聊天助手，输入应用名称，点击创建，进入应用编排页面。



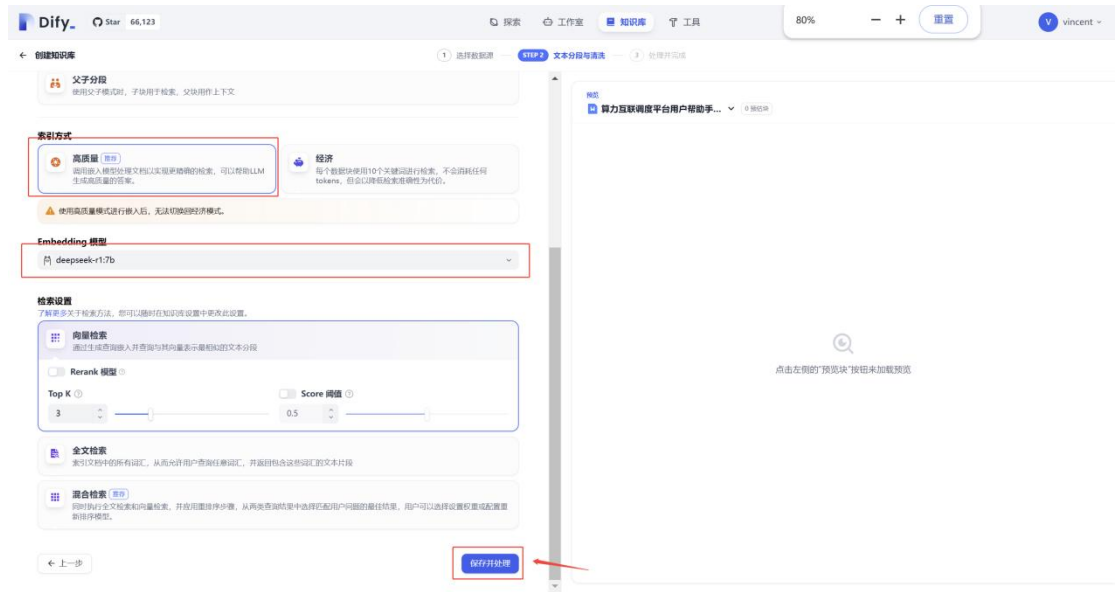


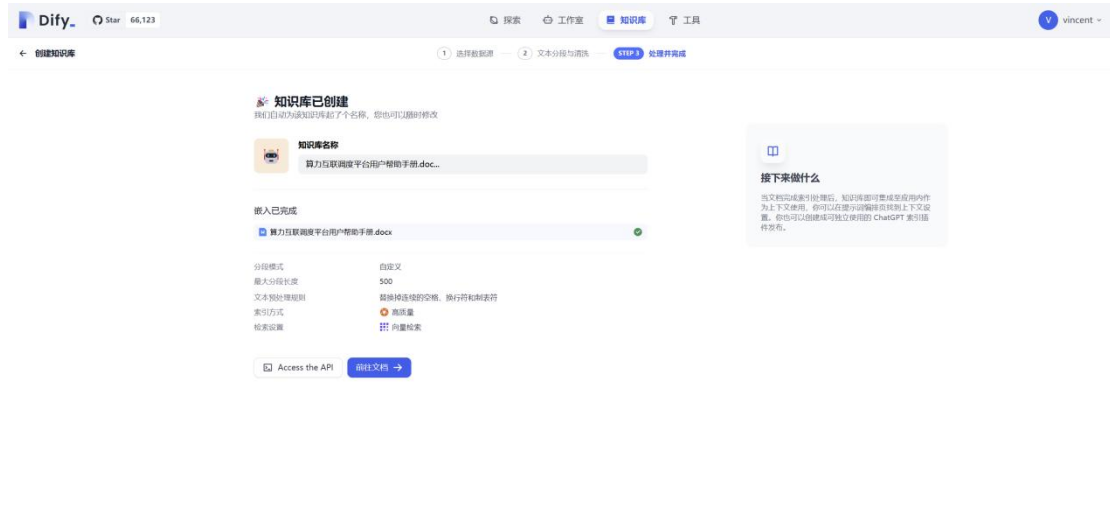
在 Dify 顶部导航栏找到“知识库”按钮，点击进入知识库编辑页面，上传需要处理的文档（支持 TXT、MARKDOWN、MDX、PDF、HTML、XLSX、XLS、DOCX、CSV、MD、HTM，每个文件不超过 15MB）。



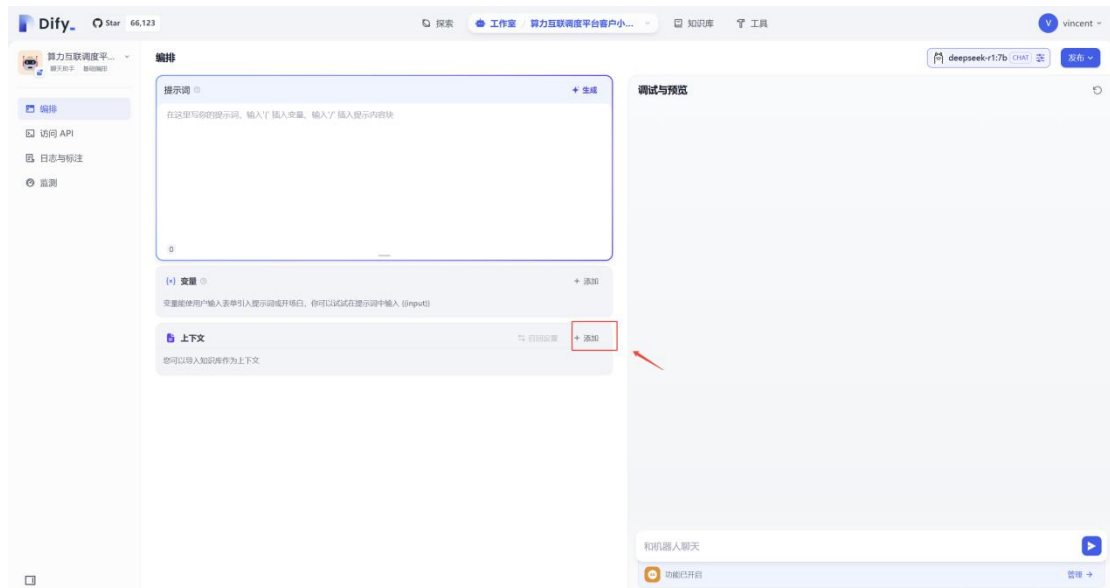


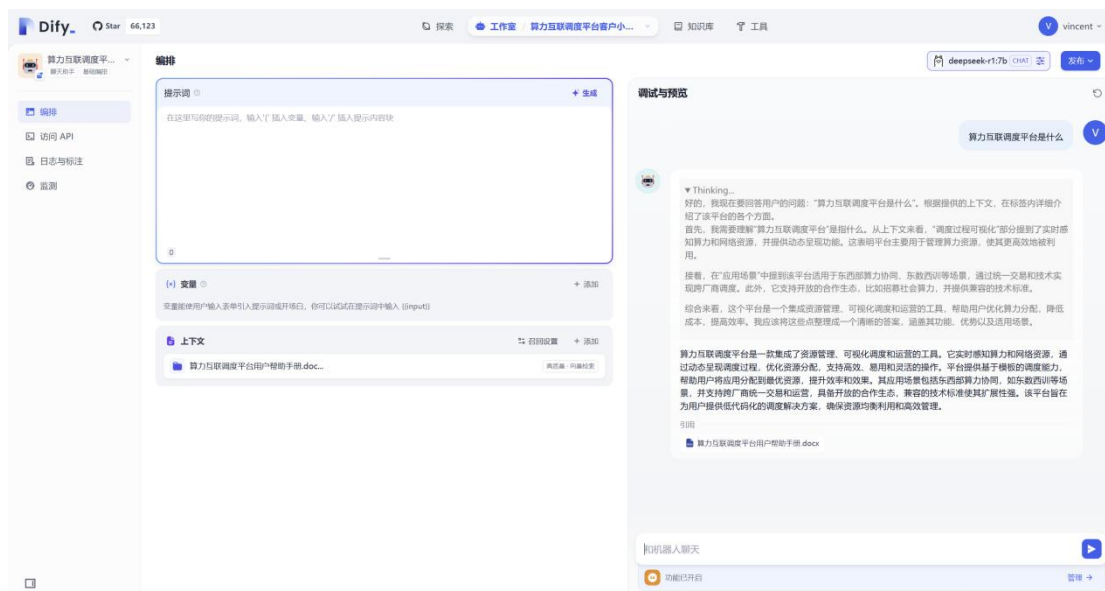
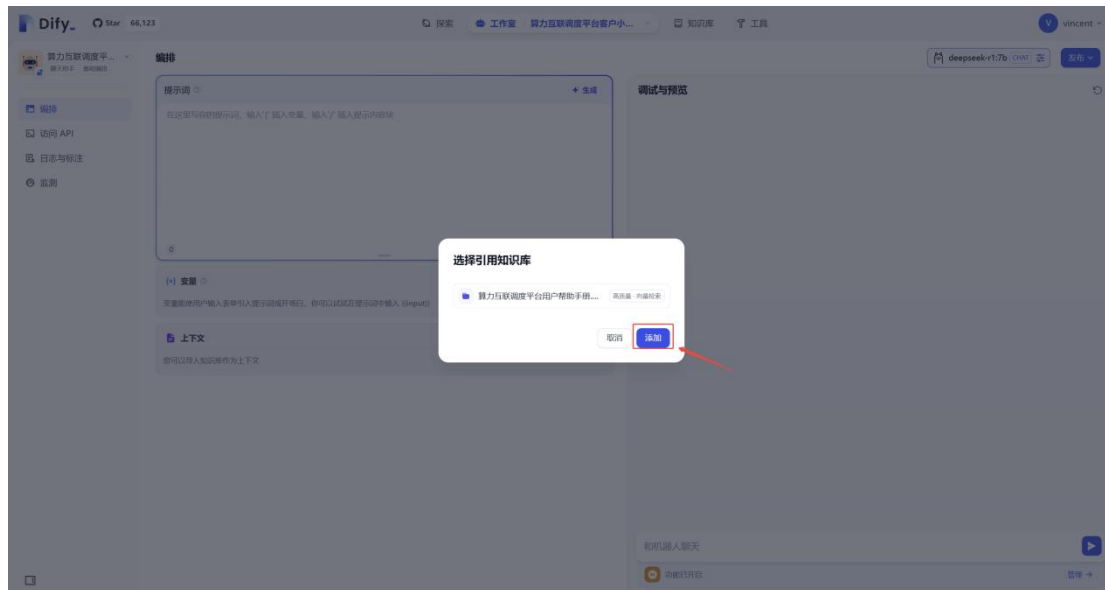
进入文本分段与清洗页面，选择我们刚才配置的 Embedding 模型并配置其他参数。完成配置后，系统会自动对上传的文档进行解释和向量化处理。





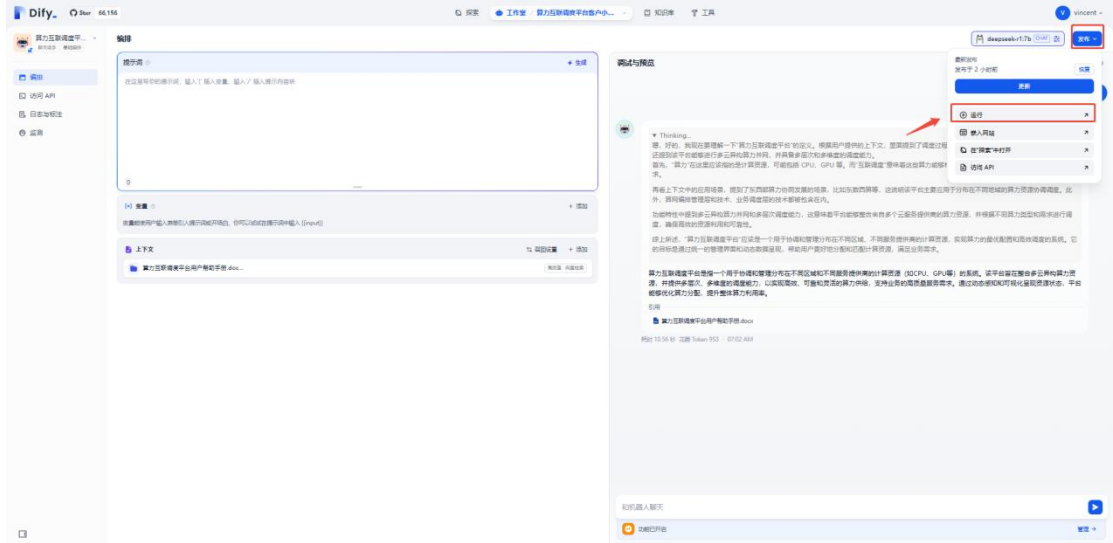
返回 Dify 聊天应用编排页面，在上下文中添加我们刚才导入的知识库。现在我们可以测试页面中输入问题进行提问，系统将会基于本地知识库返回答案。至此，我们已成功使用 Dify+DeepSeek 模型构建私有知识库，Dify 还具备工具、函数等强大能力，请用户自行探索相关功能。



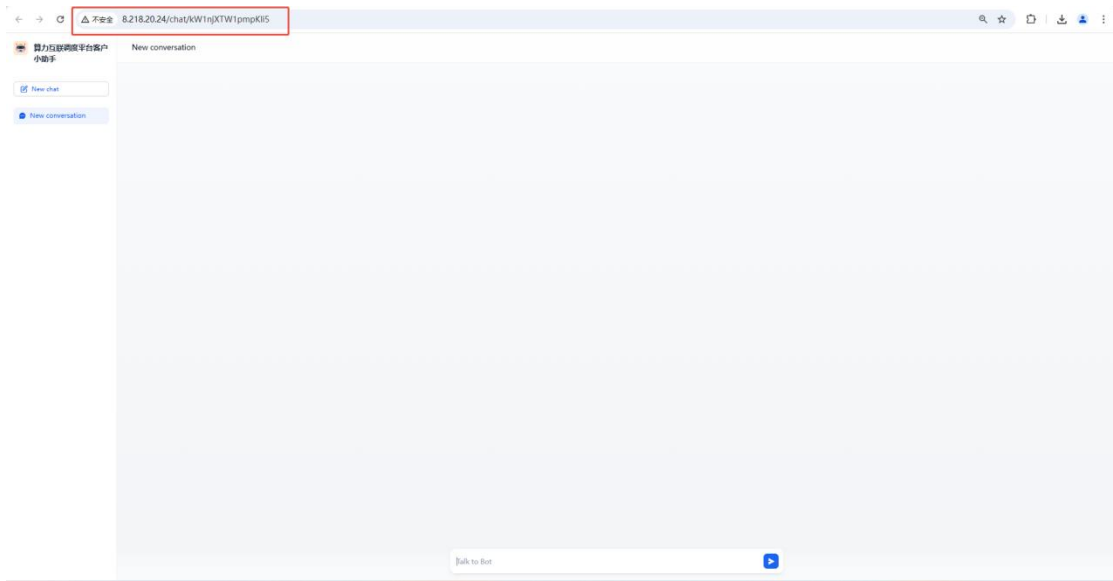


在 Dify 中发布编排好的应用，对外开放使用

在应用编排界面，点击发布-运行，即可生成网站应用。



复制生成的应用链接，其他用户可通过链接登录进行体验。请勿大量用户同时使用，否则可能会造成访问失败。





2.通过应用市场部署 Qwen3 推理 api 服务

概述

概述 阿里巴巴开源的新一代通义千问模型 Qwen2.5，在部分公开基准测试中表现优异。其创新的“混合推理”架构，将“快思考”与“慢思考”集成于同一个模型。对于简单问题，模型可以低算力快速响应；对于复杂问题，则可启动多步骤深度思考，旨在实现算力消耗与回答质量的平衡。

本最佳实践将指导您如何在应用托管平台的应用市场中，快速部署应用，并对外提供 API 调用服务，方便您即刻体验，开箱即用。

前置说明

该文档为应用托管平台控制台通过应用市场以推理 api 形态体验 Qwen3 的说明，通过应用托管平台控制台，打开应用市场，创建对应的 Qwen3 推理 api 应用，使用 web 界面和 Qwen3 推理对话，Qwen3 推理 api 应用支持 8b、14b、32b 不同参数量的模型体验。



本产品中的模型由第三方主体提供，尽管云公司已尽最大努力进行识别和维护，但仍无法保证模型的可用性。请客户按照该产品的服务协议使用该产品，做好甄别并对自行选择的服务负责。

使用前准备

天翼云账号注册：使用应用托管平台须具备天翼云官网账号。已有天翼云账号的直接登录即可，如无天翼云账号需先注册，注册流程可参考：注册账号。

使用前提：如需使用服务请先完成实名认证，请参考账号中心-实名认证。如需使用按需服务，请确认账号余额 ≥ 100 元。

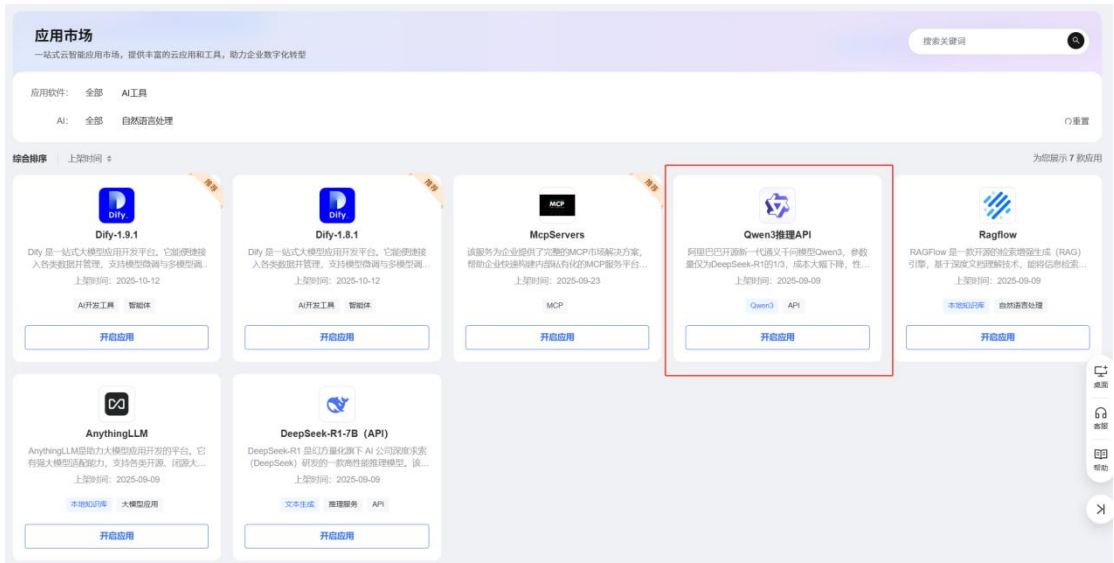
操作指引

部署 Qwen3 应用

订购应用

登录应用托管控制台，选择左侧的应用市场点击进入页面。

在应用市场页面，选择目标应用，点击【开启应用】，进入应用部署页面。



输入 API_KEY，选择访问策略及公网访问，勾选同意用户协议，点击【开启应用】，支付订单，即可完成 Qwen3 服务部署。



服务详情查看

进入应用托管租户控制台。

选择【应用实例】，进入应用实例列表页面。



服务调用

API 调用地址 (base_url)：在应用托管租户控制台，进入【应用实例】，进入应用详情，在组件概览模块，获取“外网地址”，API 调用地址格式为 “http://{外网地址}/v1/”

应用实例 / 公共应用详情

Qwen3-api-064384

ID:d5551s7levvt3133e40

部署中

访问

停止应用

更新应用

变更记录

删除应用

付费类型	按量付费	可用区	昆明2	访问策略	app-default-sg
网络类型	独享网络	访问域名	--	应用信息	Qwen3-api-064384
应用版本	1.0.0	创建时间	2025-12-23 16:21:09	更新时间	2025-12-23 16:21:09

共1个组件

qwen3

ID:d5551f7revvlt3133e4g

服务启动中

组件概览

弹性伸缩

配置详情

停止组件

更新组件

扩容组件

基础信息

算力规格CPU: 16核 | 内存 32 GB | GPU: NVIDIA-A10 * 1

内网地址api | 昆明2 | 198.19.0.2:17...

外网地址api | 昆明2 | ...17...

更新时间2025-12-23 16:21:13

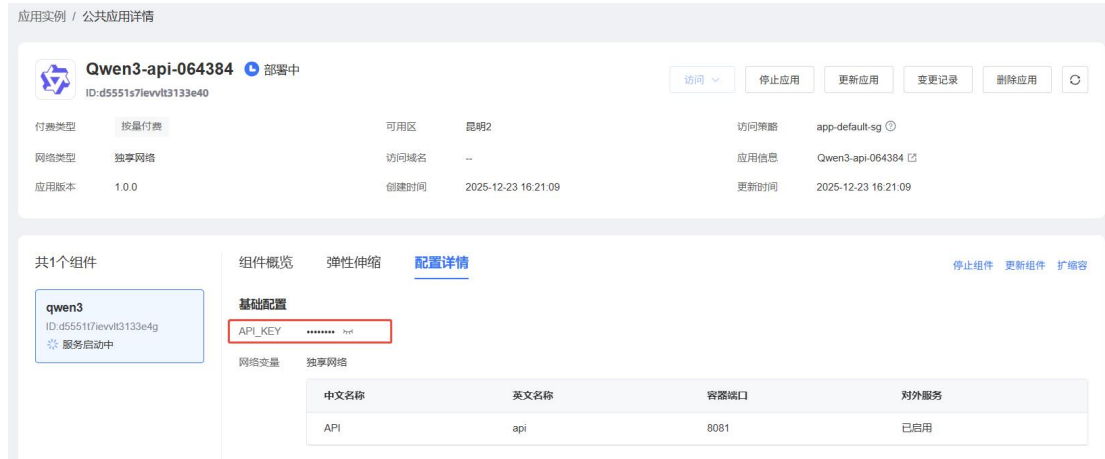
实例管理

批量重建

搜索容器关键词



API 认证 (api_key)：在应用托管租户控制台，进入【应用实例】，进入应用详情，点击查看配置，获取的 api_key。



模型名称: qwen3:32b, qwen3:8b

使用 python 代码调用 api 接口

准备 python 代码

```
from openai import OpenAIimport httpx
```

```
c = httpx.Client()
```

```
client = OpenAI(
```

```
    base_url="", # 准备工作的 base_url
```

```
    api_key="", # 准备工作的 api_key,
```

```
    http_client=c
```



```
)
```

```
stream=False
```

```
completion = client.chat.completions.create(
```

```
    messages=[
```

```
        {
```

```
            'role': 'user',
```

```
            'content': '你是谁',
```

```
        }
```

```
    ],
```

```
    model="",准备工作的模型名称
```

```
    stream=stream
```

```
)
```

```
if stream:
```

```
    output_text = "
```

```
    for chunk in completion:
```



```
# print(chunk.choices[0].delta.content,end='')
```

```
print(chunk.choices[0].delta.content or "", end="")
```

```
content= chunk.choices[0].delta.content
```

```
if content: # 检查 content 是否为 None
```

```
output text +=content
```

```
print('\n','*' * 100)
```

```
print('接收到内容: \n',output_text)
```

```
else:
```

```
print(completion)
```

```
print(completion.choices[0].message)
```

执行 python 代码，Qwen3 将给出答复

```
messages=[
    {
        'role': 'user',
        'content': '你是谁',
    },
],

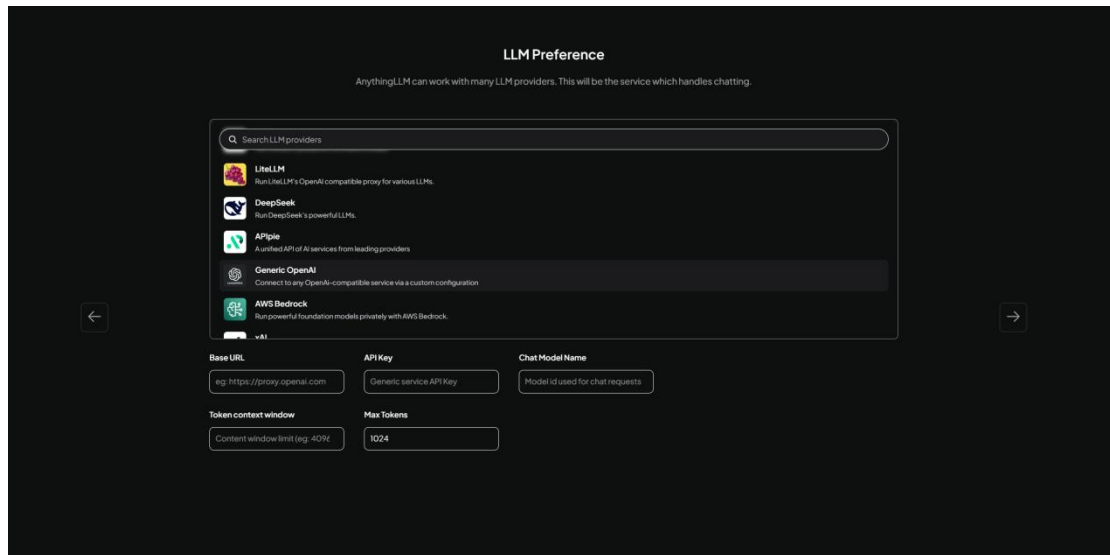
运行: openai-1 <
D:\python3\python.exe D:\workcode\test-work-platform\backend\openai-1.py
ChatCompletion(id='chatcmpl-230', choices=[Choice(finish_reason='stop', index=0, logprobs=None, message=ChatCompletionMessage(content='<think>\n嗯，用户问“你是谁”，我应该先确认他们是
否知道我的中文名。可能他们不太熟悉我的名字，或者是想确认一下，我应该用亲切友好的语气回答。同时，作为一个AI，我需要说明自己是由通义实验室研发的，这样可以展现专业性，也能让用户对我的来源有所了解。此外，可以提到可以回答
各类问题，提供帮助。\\n\\n用户可能还想知道具体有哪些能力，比如能做什么，这时候可以适当列举几个应用场景作为参考。不过也要注意用词简洁，不能太复杂。同时要让用户感觉友好，没有压力，最后应该邀请用户提问，这样可以开始互动。
比如他们有什么问題或需要帮助的，这样不仅回答了用户的问题，还鼓励他们进一步交流。\\n</think>\\n\\n你好！我是Qwen3，是阿里巴巴集团旗下的通义实验室自主研发的超大规模语言模型。我能够回答问题、创作文字、进行逻辑推理和
编程等任务。如果你有任何问題或需要帮助，随时告诉我，我会尽力协助你！', 'role='assistant', function_call=None, tool_calls=None]], created=1745919824, model='qwen3:32b', object='chat
.completion', system_fingerprint='fp_ollama', usage=CompletionUsage(completion_tokens=216, prompt_tokens=10, total_tokens=226))
ChatCompletionMessage(content='<think>\n嗯，用户问“你是谁”，我应该先确认他们是否知道我的中文名。可能他们不太熟悉我的名字，或者是想确认一下，我应该用亲切友好的语气回答。同时，作为一个AI，我需要说明自己是由通义
实验室研发的，这样可以展现专业性，也能让用户对我的来源有所了解。此外，可以提到可以回答各类问题，提供帮助。\\n\\n用户可能还想知道具体有哪些能力，比如能做什么，这时候可以适当列举几个应用场景作为参考。不过也要注意用词简洁
，不能太复杂。同时要让用户感觉友好，没有压力，最后应该邀请用户提问，这样可以开始互动。比如他们有什么问題或需要帮助的，这样不仅回答了用户的问题，还鼓励他们进一步交流。\\n</think>\\n\\n你好！我是Qwen3，是阿里巴巴
集团下的通义实验室自主研发的超大规模语言模型。我能够回答问题、创作文字、进行逻辑推理和编程等任务。如果你有任何问題或需要帮助，随时告诉我，我会尽力协助你！', 'role='assistant', function_call=None,
tool_calls=None)

进程已结束,退出代码0
```

使用 anythingllm 连接 api 服务

下载 anythingllm 桌面端程序并安装。

打开 anythingllm 程序，选择模型提供方【Generic OpenAI】。



输入 baseURL: 准备工作的 base_url, APIKey: 准备工作的 api_key, Chat Model Name:

准备工作的模型名称, Token context windows 大小, 完成模型提供方配置。

输入问题, 并发送问题, Qwen3 将给出答复。

🗨️ 嗯，用户问的是鸡兔同笼的问题，有35个头，96条腿，问鸡和兔子各有多少只。首先，我得确认鸡和兔子的...

🔊 设鸡有 x 只，兔有 y 只。根据题意可得方程组：

$$\begin{cases} x + y = 35 & \text{(头的总数)} \\ 2x + 4y = 96 & \text{(腿的总数)} \end{cases}$$

解法步骤：

01. 由第一个方程得 $x = 35 - y$ ，代入第二个方程：

$$2(35 - y) + 4y = 96$$

02. 展开并化简：

$$70 - 2y + 4y = 96 \Rightarrow 2y = 26 \Rightarrow y = 13$$

03. 代入 $x = 35 - y$ 得：

$$x = 35 - 13 = 22$$

答案：
鸡有22只，兔有13只。

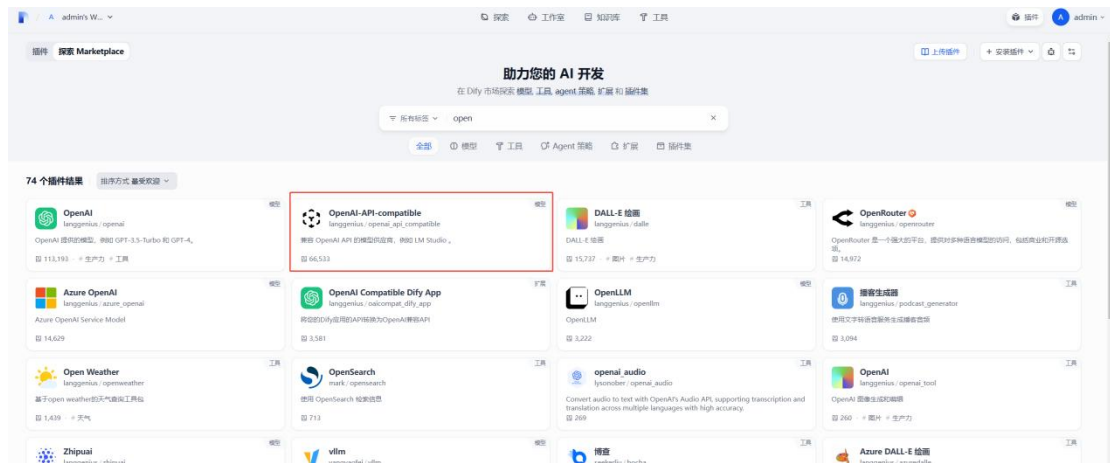
验证：
鸡腿数 $22 \times 2 = 44$ ，兔腿数 $13 \times 4 = 52$ ，总腿数 $44 + 52 = 96$ ，符合题意。

📄 📌 🗑️ :

32.674s (19.04 tok)

使用 Dify v1.8.1 连接 api 服务

部署 dify v1.8.1，选择 OpenAI-API-compatible 插件进行安装。



点击系统配置，添加模型。

设置

工作空间

模型供应商

成员

数据来源

API 扩展

通用

语言

模型供应商

模型列表

待配置

安装模型供应商

搜索

系统模型尚未完全配置

系统模型设置

尚未安装模型供应商
请安装模型供应商。

OpenAI-API-compatible

LLM | RERANK | TEXT EMBEDDING | SPEECH2TEXT | TTS

请配置 API 密钥，添加模型。

添加模型

发现更多就在 Dify 市场

OpenAI

langgenius · 113,218

OpenAI 提供的模型，例如 GPT-3.5-Turbo 和 GPT-4。

PRODUCTIVITY | UTILITIES

Anthropic

langgenius · 26,904

Anthropic 的强大模型。

Amazon Bedrock

langgenius · 5,945

The models of Amazon Bedrock

Azure OpenAI

langgenius · 14,630

Azure OpenAI Service Model

输入 API endpoint URL：准备工作的 base_url，API Key：准备工作的 api_key，API endpoint 中的模型名称：准备工作的模型名称，完成模型提供方配置。

添加 OpenAI-API-compatible

模型类型 *

☒ LLM

☐ Rerank

☐ Text Embedding

☐ Speech2text

☐ TTS

模型名称 *

qwen3:32b

模型显示名称

qwen3:32b

API Key

[REDACTED]

API endpoint URL *

http://[REDACTED]/v1/

API endpoint中的模型名称

qwen3:32b

Completion mode

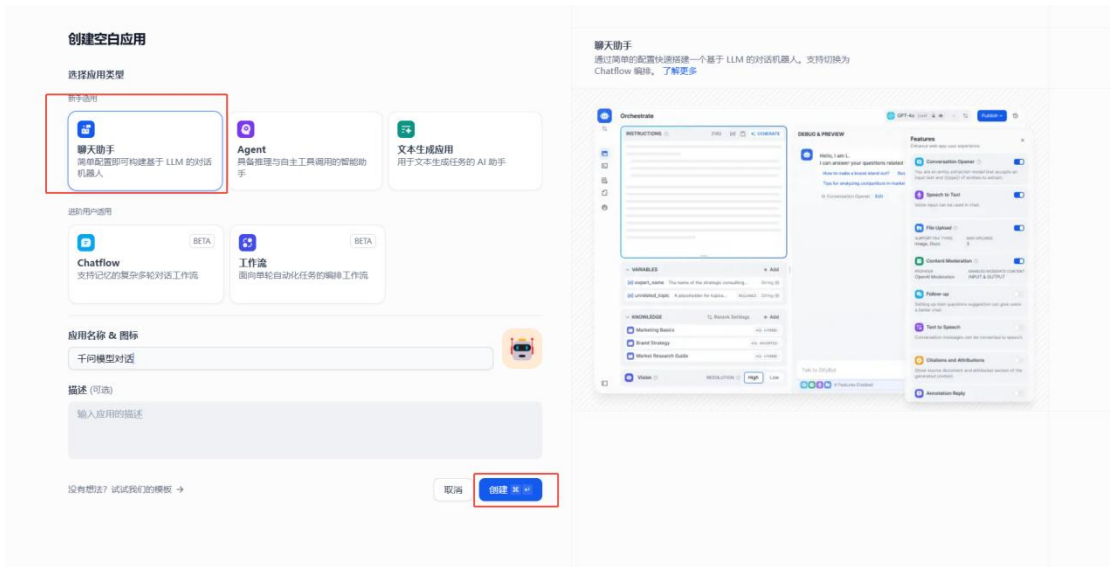
取消

保存

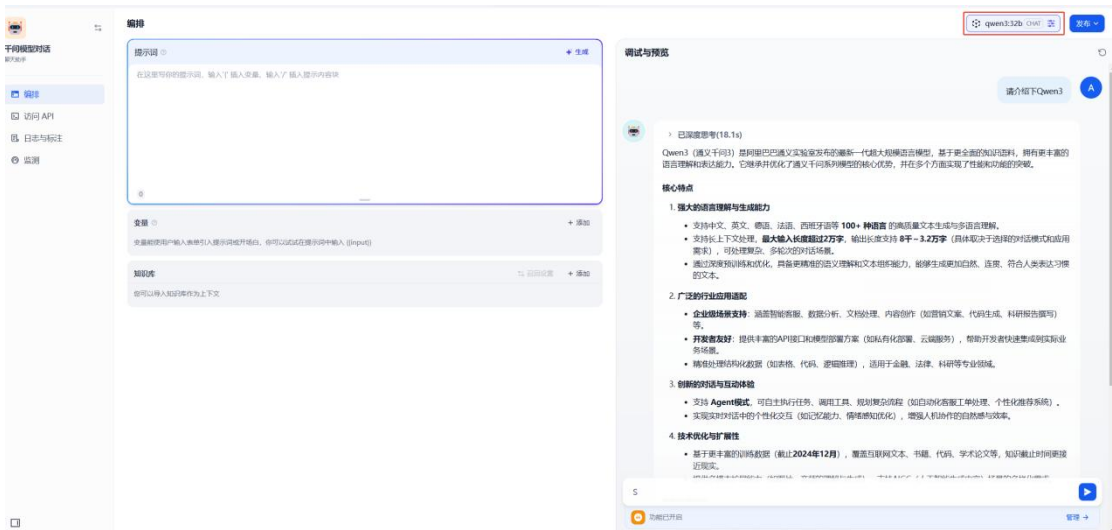
返回工作室，创建空白应用。



选择聊天助手，输入应用名称，提交创建。



输入问题，并发送问题，Qwen3 将给出答复。



3.在 Dify 中集成 MCP Server

本文介绍在应用托管中搭建 Dify 和部署 MCP Server,以及在 Dify 中集成基于部署的 MCP Server。

概述

Dify 介绍

Dify 是一站式大模型应用开发平台。它能便捷接入各类数据并管理，支持模型微调与多模型调用。通过可视化界面和低代码工具，用户无需深厚技术背景，就能快速搭建如聊天机器人、智能客服等应用。可用于客户服务、内容创作等场景，具有降低开发门槛、提升效率、灵活可扩展等优势，助力企业高效落地大模型应用。



前置说明



本文档旨在指导用户如何通过应用托管控制台，体验将 MCP Server 集成至 Dify。

本产品中的模型由第三方主体提供，尽管云公司已尽最大努力进行识别和维护，但仍无法保证模型的可用性。请客户按照该产品的服务协议使用该产品，做好甄别工作，并对自行选择的服务承担相应责任。

使用前准备

天翼云账号注册：使用应用托管平台须具备天翼云官网账号。已有天翼云账号的直接登录即可，如无天翼云账号需先注册，注册流程可参考：注册账号。

使用前提：如需使用服务请先完成实名认证，请参考账号中心-实名认证。如需使用按需服务，请确认账号余额 ≥ 100 元。

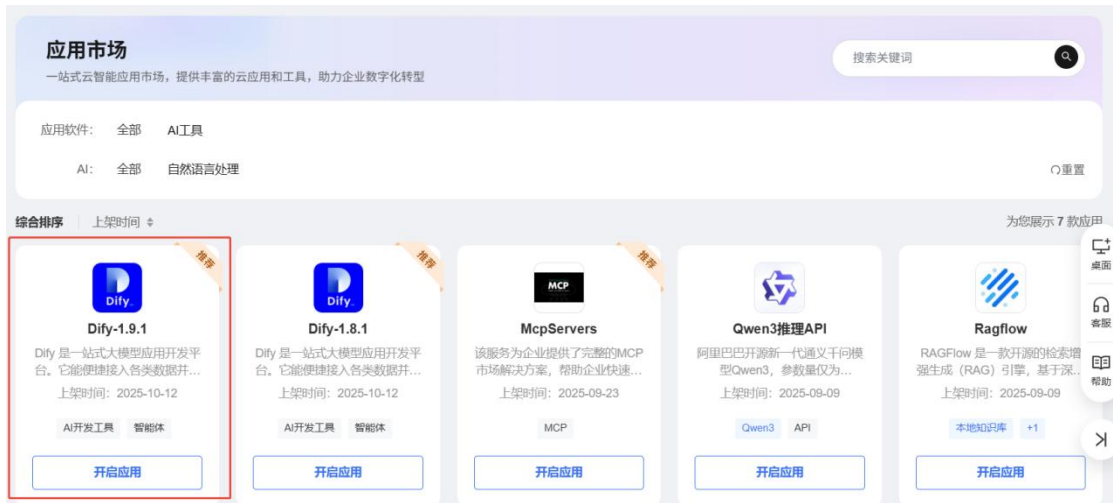
操作步骤

部署 Dify 应用

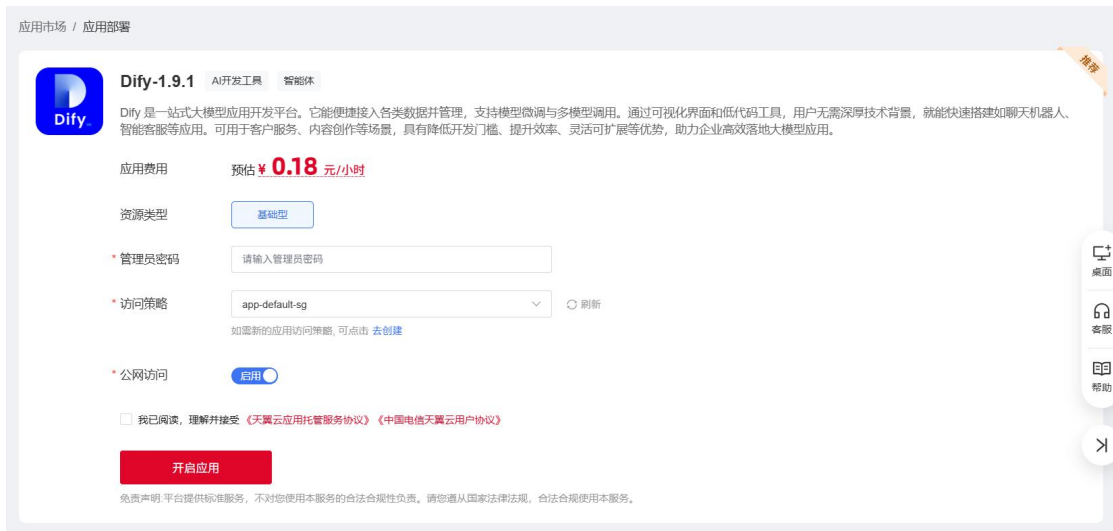
订购应用

登录应用托管控制台，选择左侧的应用市场点击进入页面。

在应用市场页面，选择目标应用，点击【开启应用】，进入应用部署页面。



填写管理密码，选择访问策略及公网访问，勾选同意用户协议，点击【开启应用】，支付订单，即可完成 Dify 应用服务部署。



服务详情查看

进入应用托管租户控制台。

选择【应用实例】，进入应用实例列表页面。

刷新应用实例列表，等待新创建的应用进入到【运行中】状态后，点击右侧操作栏【访问】，
点击出现的访问链接，即可进入 Dify 使用界面。



部署 MCP Server

订购应用

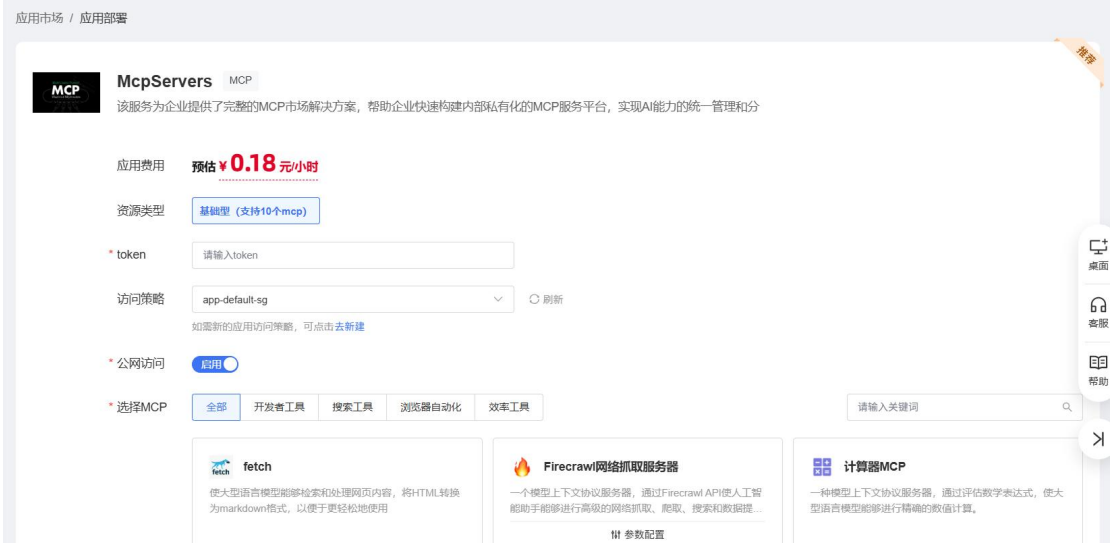
登录应用托管控制台，选择左侧的应用市场点击进入页面。

在应用市场页面，选择目标应用 “McpServers”，点击【开启应用】，进入应用部署页面。





填写认证 token，选择访问策略及公网访问，选择需要的 MCP 服务，勾选同意用户协议，
点击【开启应用】，支付订单，即可完成服务部署。



服务详情查看

进入应用托管租户控制台。

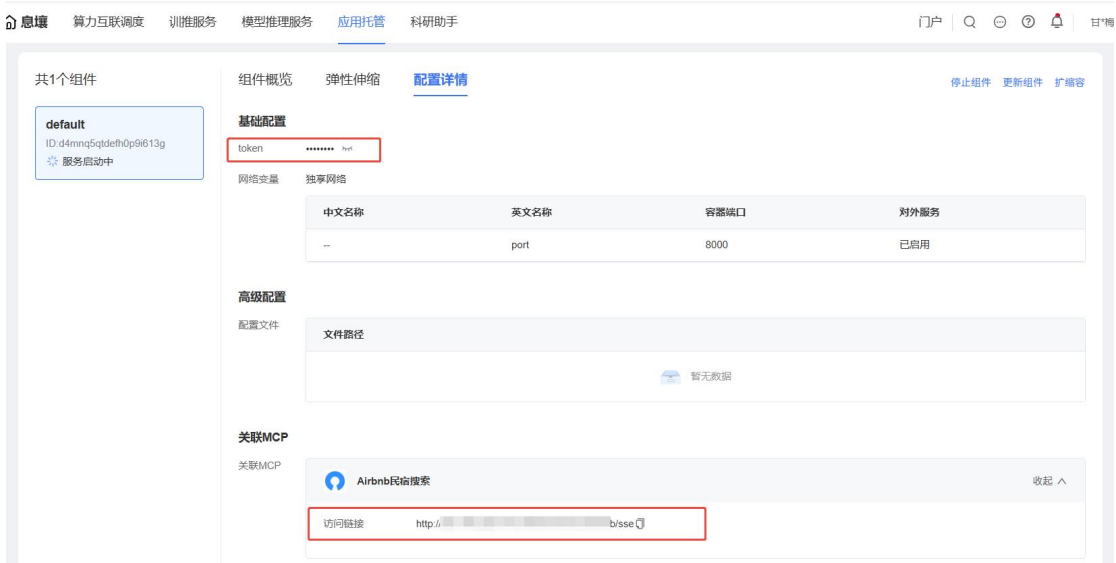
选择【应用实例】，进入应用实例列表页面。



点击 实例名称，进入应用详情。



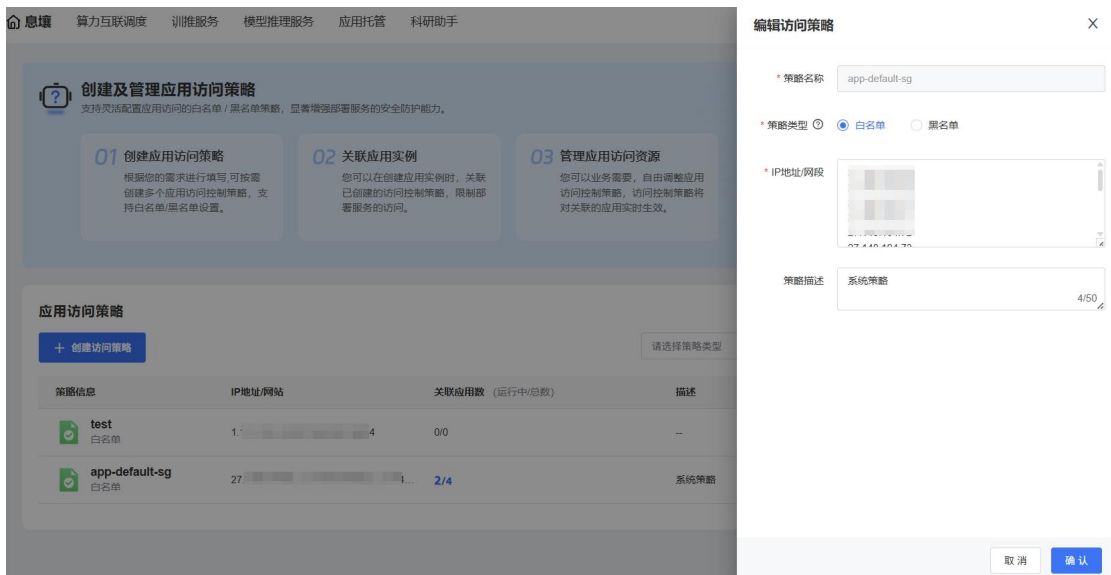
点击详情中的【配置详情】TAB，可查询 token 和 mcp 服务的 URL。



MCP 服务放通 Dify 访问

进入应用托管租户控制台。选择【应用访问策略】，进入应用访问策略列表页面。

点击 MCP 服务绑定的策略操作栏的【修改】，在白名单策略中增加 Dify 的 IP 地址。



在 Dify 中添加 MCP Server

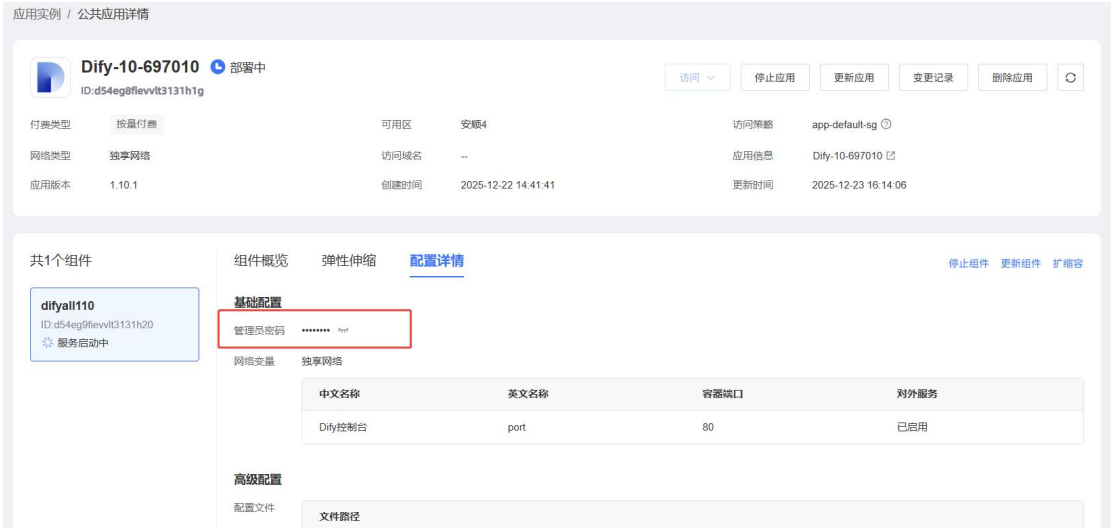


进入应用托管租户控制台。选择【应用实例】，进入应用实例列表页面。点击右侧操作栏【访问】，点击下拉访问链接。

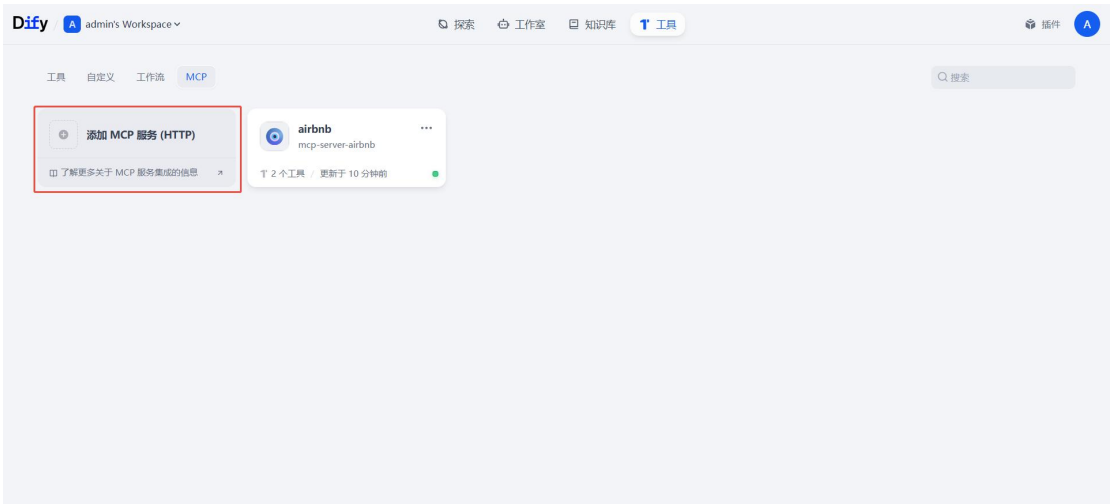


进入 Dify 应用界面。可使用管理员账号和设置的密码登录 Dify，登录后用户可自行进行更改密码。账号：admin@admin.com；密码：可点击“应用实例”进入详情页，在“配置详情”中的“管理员密码”查看。





在工具 > MCP 中添加 MCP 服务。



添加页面中服务端点 URL 填写部署的 MCP Server 的 URL, 请求头的请求头值填写 token。

请求头名称	请求头值	备注
Authorization	Bearer token	请将 token 替换为应用托管控制台中 MCP server 的配置的真实 token



添加后等待自动授权完成。如出现以下信息，表示在 Dify 中已成功添加 MCP Server。



测试 MCP 功能

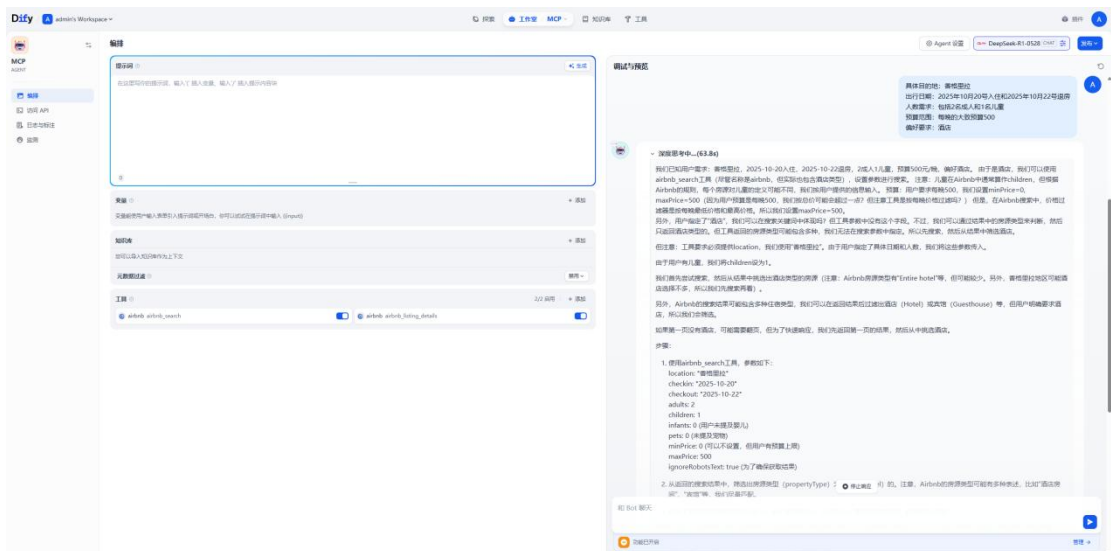
通过搭建一个简单的 Agent 应用测试 MCP Server 的能力。

创建一个 Agent 应用，请注意应用类型选择 Agent。

在 Agent 应用中添加工具，选择已添加到 Dify 的 MCP server。



在调试与预览界面输入自然语言，在 Bot 处理过程中将会调用 MCP Server。



4. 云主机通过内网访问应用实例

本文介绍云主机等公有云资源如何实现内网访问相同地域的应用实例。

前提条件

云主机等资源与应用托管应用实例处于相同地域，且应用实例使用共享网络。

已申请获取应用托管的 VPC 终端节点服务的服务实例 ID（请通过您的客户经理或工单提交申请）。

云资源访问应用实例与应用实例访问云资源需分别申请，应用实例访问云资源详见最佳实践【应用实例通过内网访问数据库等云资源】。

创建 VPC 终端节点

为云主机等资源创建 VPC 终端节点。登录天翼云网络控制台，在左侧导航栏单击【VPC 终端节点】-【终端节点】，单击【创建终端节点】按钮。



在创建终端节点页面，配置相关信息。

字段	说明
服务类型	请选择“按服务实例 ID 查找服务”
可用服务	请填入已获取的“应用托管的 VPC 终端节点服务的服务实例 ID”
虚拟私有云	请选择云主机等资源所在的虚拟私有云



提供账号邮箱，由应用托管后台添加终端节点服务白名单，并审核连接申请（请通过您的客户经理或工单提交申请）。

确定连接状态

应用托管平台侧通过连接申请后，可在终端节点列表查看到状态为“已连接”。

<input type="checkbox"/>	名称/ID	标签	企业项目	IP地址...	终端节...	名称/服务ID	VPC	状态	创建时间	操作
<input type="checkbox"/>	VPC-1254 vpc-9652		default	IPv4	接口	VPCE-1254 vpce-1254	vpc-9652	已连接	2025-11...	删除

验证云主机可内网访问应用实例

在开通终端节点的 VPC 下创建一台云主机。

如果开启了终端节点的访问白名单，则编辑终端节点添加白名单(所创建云主机的内网网段)。

终端节点

实例ID

end-hw

实例名称

VF-测试

终端节点类型

接口

服务名称

vpce-应用托管

服务ID

状态

已连接

节点VPC

vpc-

节点子网

subnet-hsl-test

节点IP

192.168.0.6

IP地址类型

IPv4

创建时间

2025-11-03 15:58:08

私网域名

私网域名名称

-

描述

--

删除保护

企业项目

default

访问控制

访问白名单: 关闭

开启

添加白名单

请输入白名单网段

最多允许添加20个网段，反向时填入的时容许被服务访问的网段

白名单网段

操作

192.168.0/24

删除

在应用托管控制台部署应用实例，启用对外服务，绑定域名。并修改应用访问策略，放行198.19.128.0/20 网段。

网络变量 HTTP

url地址

本地监听端口

名称

对外服务

操作

/*

80

请输入名称

启用

删除

添加

* 域名绑定

ctyun..yz

* 镜像类型

镜像地址

docker.m.daocloud.io/library/nginx:1.17.6

私有化镜像仓库授权

启动命令

请输入Shell启动命令

例如: /webui.sh --listen --port 8000 --skip-version-check --no-hashing --no-download-sd-model --skip-prepare-environment

编辑访问策略X

* 策略名称

app-default-sg

* 策略类型 ?

☒ 白名单 ☐ 黑名单

* IP地址/网段

183.71.238.0/24
~~106.86.111.233/32~~
198.19.128.0/20

策略描述

☒ 每行一个IP地址或网段，以回车键换行

☒ 每个IP地址或网段都可以用“|”分隔添加描述 如“10.0.0.0/24|办公网出口”

9/50

在 VPC 终端节点详情获取应用实例访问的内网 IP（即终端节点的节点 IP），并在云主机配置应用域名 host。

终端节点

实例ID	enc-...	实例名称	VPCE-应用托管测试	终端节点类型	接口
服务名称	vpce-应用托管	服务ID	er-...	状态	已连接
节点VPC	vpc-...	节点子网	subn-...	节点IP	192.168.0.6
IP地址类型	IPv4	创建时间	2025-11-03 15:58:08	私网域名	<input type="checkbox"/>
私网域名名称	-	描述	--	删除保护	<input type="checkbox"/>
企业项目	default				

```
<center><h1>502 Bad Gateway</h1></center>
<hr><center>openresty</center>
<p><em>Powered by <a href="https://apisix.apache.org/">APISIX</a>.</em></p></body>
</html>
(base) root@ecm-hsl-test:~# echo "192.168.0.6 ctyu[REDACTED].xyz" | sudo tee -a /etc/hosts
```

IP 地址	域名
VPC 终端节点详情中的节点 IP	应用访问域名

登录云主机，通过应用域名内网访问应用实例。

```
[root@ecm-vpc ~]# curl http://[REDACTED]:80/
<!DOCTYPE html>
<html>
<head>
<title>Welcome to nginx!</title>
<style>
html { color-scheme: light dark; }
body { width: 35em; margin: 0 auto;
font-family: Tahoma, Verdana, Arial, sans-serif; }
</style>
</head>
<body>
<h1>Welcome to nginx!</h1>
<p>If you see this page, the nginx web server is successfully installed and
working. Further configuration is required.</p>

<p>For online documentation and support please refer to
<a href="http://nginx.org/">nginx.org</a>.<br>
Commercial support is available at
<a href="http://nginx.com/">nginx.com</a>.</p>

<p><em>Thank you for using nginx.</em></p>
</body>
</html>
```

5. 应用实例通过内网访问数据库等云资源

前提条件

应用托管应用实例与数据库等云资源处于相同地域，且应用实例使用共享网络。



已申请获取应用托管的 VPC 终端节点服务的服务实例 ID（请通过您的客户经理或工单提交申请）。

应用实例访问云资源与云资源访问应用实例需分别申请，云资源访问应用实例详见最佳实践【云主机通过内网访问应用实例】。

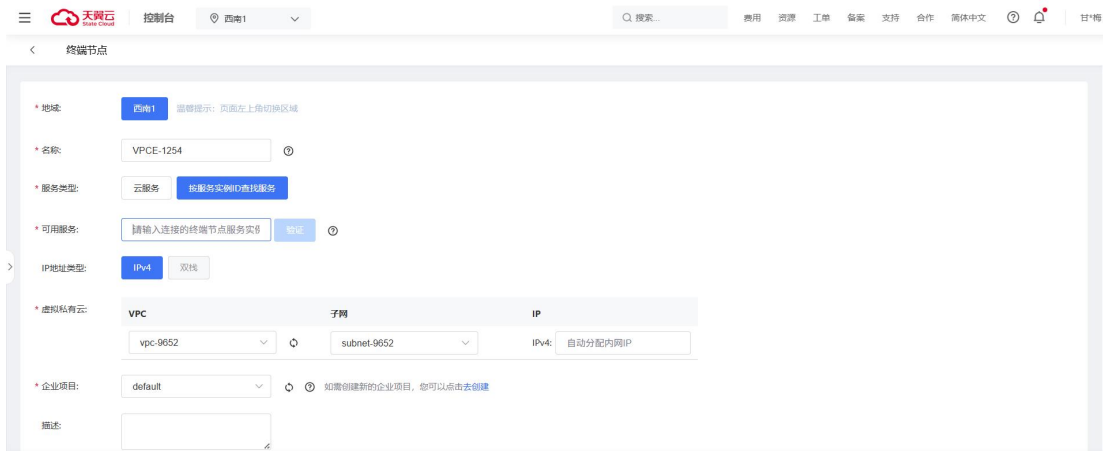
创建 VPC 终端节点

为数据库等云资源创建 VPC 终端节点。登录天翼云网络控制台，在左侧导航栏单击【VPC 终端节点】-【终端节点】，单击【创建终端节点】按钮。



在创建终端节点页面，配置相关信息。

字段	说明
服务类型	请选择“按服务实例 ID 查找服务”
可用服务	请填入已获取的“应用托管的 VPC 终端节点服务的服务实例 ID”
虚拟私有云	请选择数据库等云资源所在的虚拟私有云



提供账号邮箱，由应用托管后台添加终端节点服务白名单，并审核连接申请（请通过您的客户经理或工单提交申请）。

确定连接状态

应用托管平台侧通过连接申请后，可在终端节点列表查看到状态为“已连接”。

<input type="checkbox"/>	名称/ID	标签	企业项目	IP地址...	终端节...	名称/服务ID	VPC	状态	创建时间	操作
<input type="checkbox"/>	VPC-1254 vpce-1254-3vf6p		default	IPv4	接口	vpce-1254-3vf6p-7xzc	vpc-1254	已连接	2025-11...	删除

验证应用实例可内网访问数据库

在开通终端节点的 VPC 下创建数据库服务，这里以创建关系型数据库（MySQL）为例。



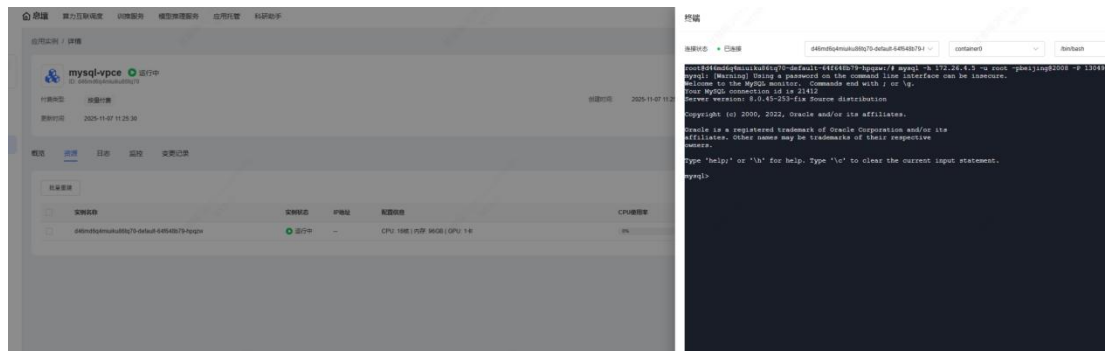
基本信息	实例信息			
备份恢复	实例名称	mysql-168 编辑	IPV4地址	连接地址: 192.168.0.15 访问名称: 设置
日志管理	实例ID	7481a*****1184c0 设置数据库只读	IPV6地址	连接地址: --
参数设置	数据库端口	13049 修改端口	计费方式	按量
SQL审计	模式	非MGR	实例大小写	不分大小写
账号管理	时区	UTC +8:00	实例类型	单机
数据安全	创建时间	2025-11-07 14:58:54	到期时间	--
数据库代理	管理员用户名	root 修改密码 重置帐号权限 禁用帐号	订单号	358096
数据安全	可用区信息 迁移可用区			
SQL主备	主节点可用区	可用区1	主节点ID	6935
SQL读写	配置信息			
死锁分析	CPU类型	海光	性能类型	海光-计算增强型
	实例规格	2C4G 规格扩容	数据库版本	8.0

The screenshot shows the '白名单与安全组' (Whitelist and Security Group) page in the Alibaba Cloud console. The left sidebar has '白名单与安全组' highlighted with a red box. The main content area shows a list of security groups. The first group is 'vpc_default | IPv4' with a status of '正常' (Normal). The second group is 'slb | IPv4' with a status of '192.19.13%'. Red arrows point to the status of the first group and the IP address of the second group.

在 VPC 终端节点详情获取节点 IP，同时提供数据库连接信息，应用托管后台添加数据库访问规则（请通过您的客户经理或工单提交申请）。

终端节点					
实例ID	enc- /vsw	实例名称	VPCE-应用托管测试	终端节点类型	接口
服务名称	vpce-应用托管	服务ID	er- -c	状态	已连接
节点VPC	vpc- -it	节点子网	subn- 	节点IP	192.168.0.6
IP地址类型	IPv4	创建时间	2025-11-03 15:58:08	私网域名	<input type="checkbox"/>
私网域名名称	-	描述	--	删除保护	<input type="checkbox"/>
企业项目	default				

访问规则添加完成后，获取到数据库访问地址，在应用托管平台部署数据库客户端，成功通过内网连接数据库。



6.使用容器镜像服务部署应用

前提条件

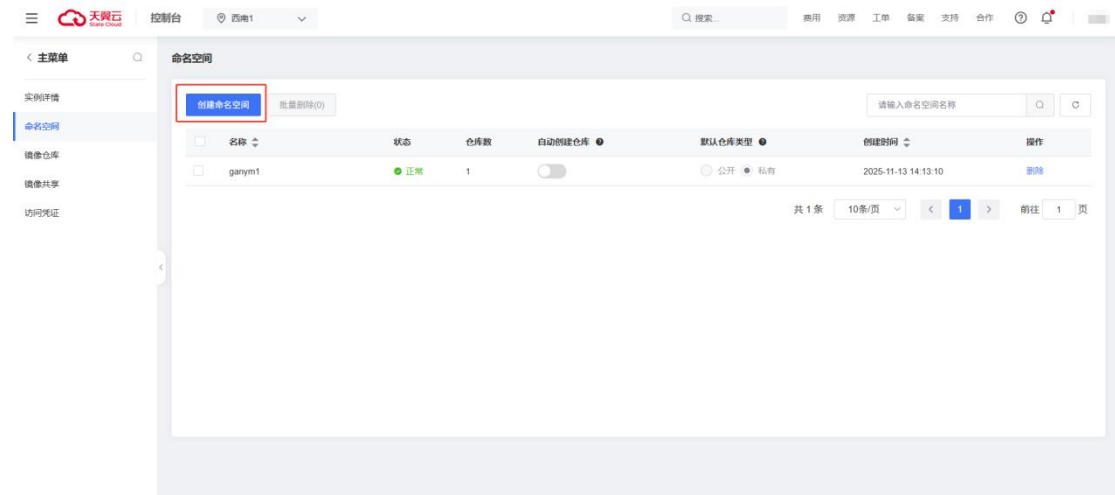
已开通个人版容器镜像服务实例。

如需使用内网拉取容器镜像，应用托管应用实例与容器镜像服务实例需处于相同地域，且应用实例使用共享网络。

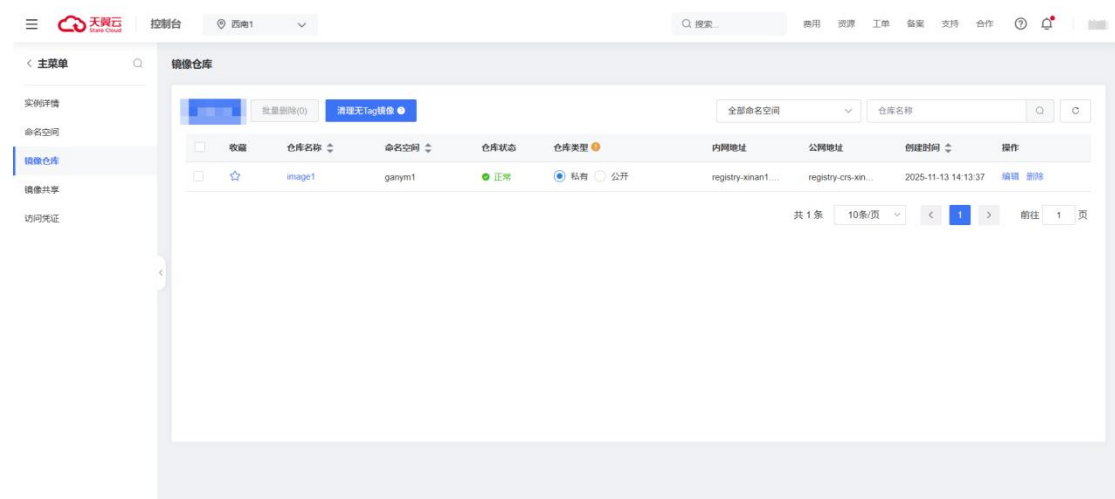
在容器镜像服务创建镜像仓库



登录容器镜像服务控制台，在实例列表点击【实例名称】进入指定的实例。点击左侧菜单栏的【命名空间】，再点击【创建命名空间】，在容器镜像服务实例中创建命名空间（也可以使用已经创建好的命名空间）。



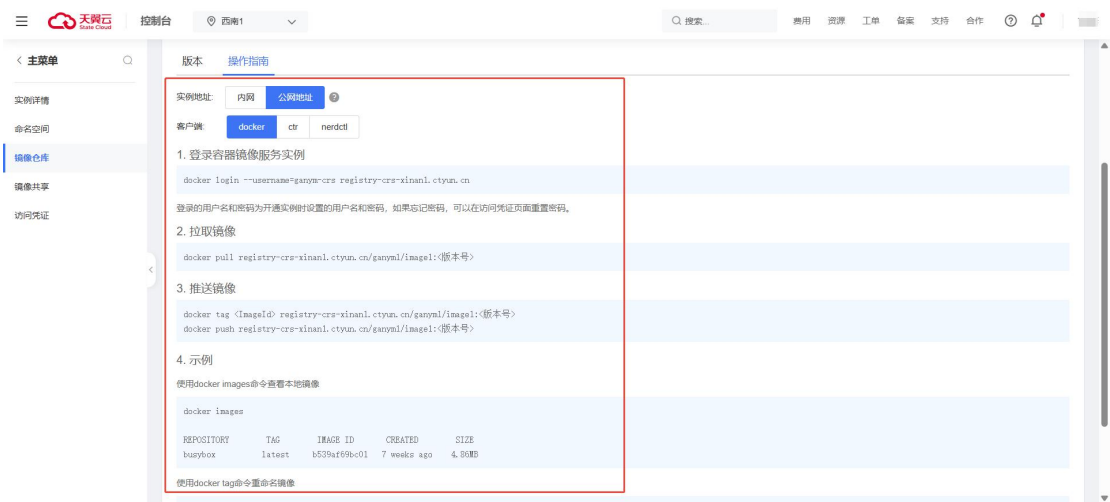
点击【镜像仓库】-【创建仓库】，在容器镜像服务中创建镜像仓库（当命名空间设置为允许自动创建仓库时，可以直接通过推送镜像来创建镜像仓库）。



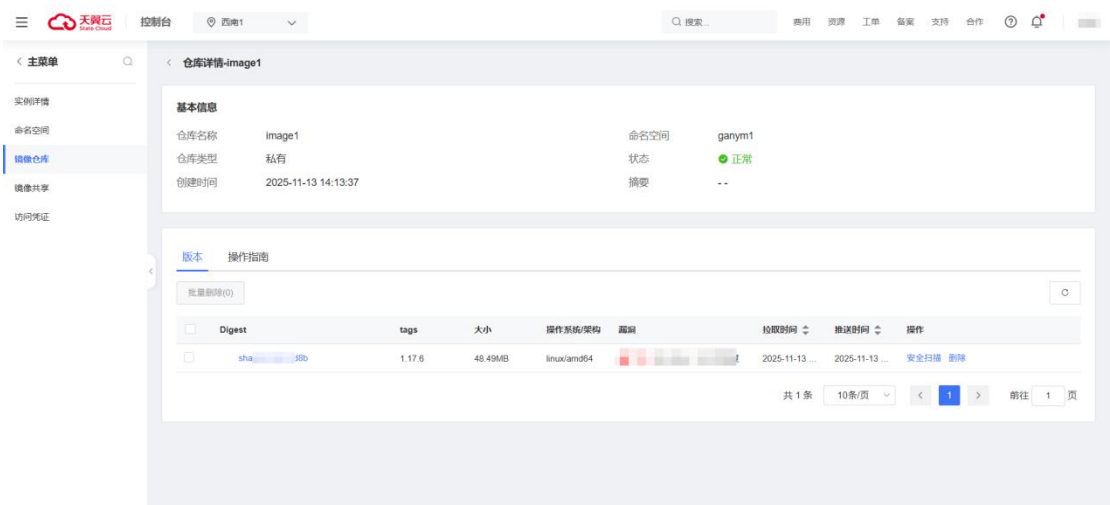
将容器镜像推送到容器镜像服务实例

用户可以根据自身的业务需求，通过 Dockerfile 构建镜像或者使用其它已经构建好的镜像。

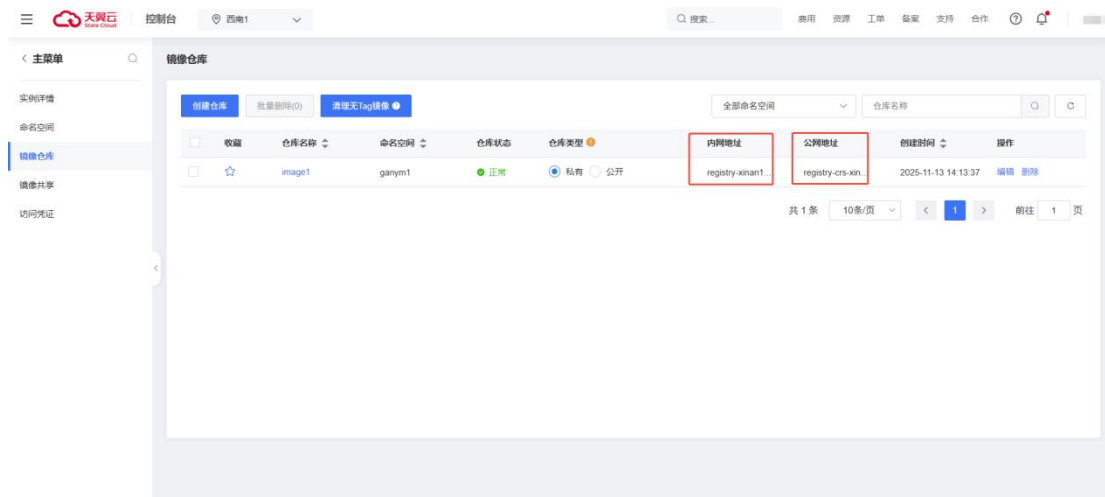
通过客户端工具将构建好的镜像推送至容器镜像服务。



推送成功后，可以在容器镜像服务控制台看到镜像的版本。



在镜像仓库中获取内网/公网镜像拉取地址。



应用托管中部署自定义应用

登录应用托管控制台，在左侧导航栏单击【应用实例】，点击【创建应用实例】按钮。

配置应用的镜像，填写容器镜像服务中目标镜像的内网 / 公网地址”（私有仓库需提供账号密码授权；使用内网地址需确认应用托管应用实例与容器镜像服务实例处于相同地域，产品间已内网互通）。

参数配置

* 镜像类型

镜像地址

请输入镜像仓库地址/镜像名/版本

☐ 私有化镜像仓库授权

完成全部应用配置，确认部署应用的资源配置费用，并勾选协议点击【部署应用】后，完成应用实例创建。