

DeepSeek-R1-32b社区版使用指南

1 DeepSeek-R1-32b-Ascend 版本介绍

昇腾版本需要部署在昇腾云服务器上，需要先联系客户经理（如果没有客户经理，直接在华为云管理平台提交工单）申请白名单才可以开通。

白名单申请成功后，可按照下面的指导文档部署DeepSeek-R1 32b蒸馏版的模型。

2 云资源消耗说明

云资源类型	使用说明
ECS	部署工具，用于上传模型权重至OBS、上传部署所需的Docker镜像至CCE Node节点、执行部署命令等。部署结束后，会自动删除。
OBS	存储模型权重
CCE	部署模型
ModelArts AI专属资源池	昇腾云资源，作为CCE集群的Node节点使用，为DeepSeek-R1模型提供算力。

3 部署准备工作

1) 配置ModelArts的授权

由于大模型即服务平台的数据存储、模型导入以及部署上线等功能依赖OBS、SWR等服务，需获取依赖服务授权后才能正常使用相关功能。先授权新增委托，使用需要的权限模板进行创建即可，点击“此处”，然后按照如下截图中的配置进行授权即可。

⚠ 由于大模型即服务平台的数据存储、模型导入以及部署上线等功能依赖OBS、SWR等服务，需获取依赖服务授权后才能正常使用相关功能。点击 [此处](#) 获取依赖服务授权，查看 [配置ModelArts委托授权](#)。

授权配置

授权对象类型

授权对象

hid_skp-c45ajs06ntn

委托选择

查看和配置更多权限，请访问IAM服务进行配置。[立即前往](#)

委托名称

modelarts_agency

您最多可以创建50个委托，您还可以创建46个委托。

权限配置

普通模式

推荐使用，该模式可针对用户业务场景进行自由定制，并保持最小授权，安全可靠。

高权限模式

对高权限有特殊需求的用户，可使用该模式，建议管理员谨慎配置该模式下的权限。

权限模板 可以选择预设的模板快速完成权限配置

服务列表

- 弹性云服务器 (ECS) (13/13)
- MapReduce服务 (MR) (0/0)
- 应用性能管理 (APM) (1/1)
- 云硬盘服务 (EVS) (4/4)
- 数据仓库服务 (DWS) (0/3)
- 密钥管理 (DEW) (1/4)
- ModelArts服务 (ModelArts) (0/0)
- 应用运维管理 (AOM) (0/7)
- 裸金属服务器 (BMS) (2/2)
- 镜像服务 (IMS) (2/2)
- 数据湖探索 (DLI) (0/14)
- 消息通知服务 (SMN) (0/5)
- 云容灾引擎 (CCE) (12/12)
- 对象存储服务 (OBS) (0/23)
- 云监控服务 (CES) (0/1)
- 虚拟私有云 (VPC) (0/11)

功能权限

使用便捷 弹性集群 Cluster

- 全选 (共4项权限, 已选择4项)
- sf-turbo-shares:showShareNics 查询sf-turbo的网卡详情
- sf-turbo-shares:listShareNics 查询sf-turbo的网卡列表
- sf-turbo-shares:addShareNics 添加网卡
- sf-turbo-shares:deleteShareNics 删除网卡

4 云资源开通

1) 进入商品详情页，计费方式选择“按需”，点击“立即购买”，如下图所示。

2) 选择“模板配置开通”，然后直接点击“下一步”，进入参数配置页面，如下图所示。

参数名称	值	类型	描述
CCE Node节点密码	*****	字符串	CCE Node节点密码的管理员密码，密码复杂度要求：密码要求长度范围为8到26位，密码至少必须包含大...
ModelArts Lite集群规格	modelarts.bm.npu.arm.8ant9b1	字符串	指定MA Lite Cluster节点flavor ID。
ModelArts Lite集群创建节...	1	整型	指定对应flavor的资源数量。
部署工具实例密码	*****	字符串	部署工具实例的管理员密码，密码复杂度要求：密码要求长度范围为8到26位，密码至少必须包含大写字...
OBS桶名称	ds-file-bucket-app-0820	字符串	指定OBS桶名称。
VPC IPv4网段	192.168.0.0/16	字符串	取值范围：10.0.0.0/8 - 10.255.255.0/24, 172.16.0.0/12 - 172.31.255.0/24, 或者 192.168.0.0/16 - 192.1...
子网 IPv4网段	192.168.10.0/24	字符串	必须是CIDR格式，且在VPC的CIDR块内，子网掩码不能大于28。
子网网关	192.168.10.1	字符串	子网的网关，必须是子网段内的合法IP地址。
付费类型	postPaid	字符串	prePaid-预付费，即包年包月；postPaid-后付费，即按需付费。ModelArts Lite集群仅支持包年包月！
订购周期类型	month	字符串	当chargingMode为prePaid时生效且为必填值，取值范围：month-月，year-年。
订购周期数	1	整型	当chargingMode为prePaid时生效且为必填值，取值范围：periodType=month (周期类型为月) 时，取值...

3) 首次开通时，需要您授权RFS创建并使用密钥加密敏感参数，点击：“确定”

配置参数

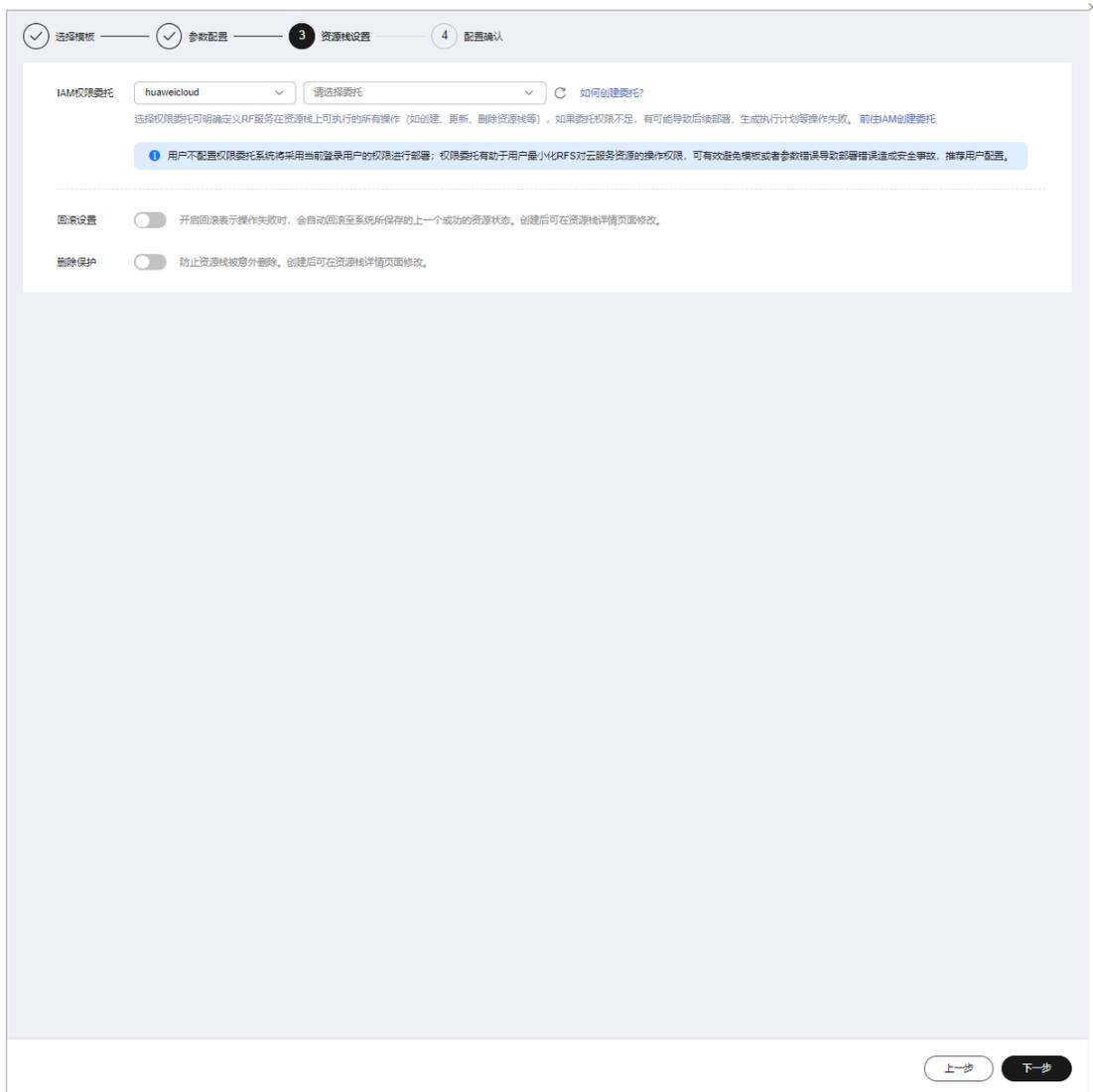
请输入关键字搜索参数名称 按模板要求对部分资源加密 

参数名称	值	类型	描述
CCE Node节点密码	*****	字符串	CCE Node节点密码的管理员密码。密码复杂度要求：密码要求长度范围为8到26位，密码至少必须包含大...
ModelArts Lite集群规格	modelarts_bm_npu_arm_8snf9b1	字符串	指定MA Lite Cluster节点flavor ID。
ModelArts Lite集群创建节...	1	整型	指定对应flavor的资源数量。
部署工具实例密码	*****	字符串	部署工具实例密码。密码复杂度要求：密码要求长度范围为8到26位，密码至少必须包含大写字...
OBS桶名称	ds-file-bucket-app-0820	字符串	桶名称。桶名称只能包含小写字母、数字、短划线（-）和点（.），且必须以字母或数字开头。桶名称长度不能超过63个字符。
VPC IPv4网段	192.168.0.0/16	字符串	VPC的IPv4网段。网段必须是CIDR格式，且在VPC的CIDR块内。子网掩码不能大于28。
子网 IPv4网段	192.168.10.0/24	字符串	子网的IPv4网段。网段必须是CIDR格式，且在VPC的CIDR块内。子网掩码不能大于28。
子网网关	192.168.10.1	字符串	子网的网关。必须是子网段内的合法IP地址。
付费类型	postPaid	字符串	prePaid-预付费，即包年包月；postPaid-后付费，即按带付费。ModelArts Lite集群仅支持包年包月！
订购周期类型	month	字符串	当chargingMode为prePaid时生效且为必填值。取值范围：month-月，year-年。
订购周期数	1	整型	当chargingMode为prePaid时生效且为必填值。取值范围：periodType=month（周期类型为月）时，取值...

开启加密

注意！初次开启加密功能会自动为您创建aos/default默认密钥，请确定是否授权资源编排服务创建并使用密钥？

4) 进入“资源栈设置”页面，然后点击“下一步”



选择模板 — 参数配置 — **3 资源线设置** — 4 配置确认

IAM权限策略 huaweicloud 请选择策略 C 如何创建策略?

选择权限策略可明确定义RFS服务在资源线上可执行的所有操作（如创建、更新、删除资源线等），如果策略权限不足，有可能导致后续部署、生成执行计划等操作失败。前往IAM创建策略

用户不配置权限策略系统将采用当前登录用户的权限进行部署，权限策略有助于用户最小化RFS对云资源的操作权限，可有效避免模板或者参数错误导致部署错误造成安全事故，推荐用户配置。

回滚设置 开启回滚表示操作失败时，会自动回滚至系统所保存的上一个成功的资源状态。创建后可在资源线详情页面修改。

删除保护 防止资源线被意外删除。创建后可在资源线详情页面修改。

上一步 下一步

5) 进入配置确认页面，需要您确认下配置参数的值，如果没有问题点击“创建执行计划”。

选择模板

资源栈名称: mlkp_stack_20250310_203... 描述: --

配置参数

参数名称	值	类型	描述
OCE Node节点密码	*****	字符串	OCE Node节点密码的管理员密码, 密码复杂程度要求: 密码要求长度范围为8到24位, 密码至少必须包含大写字母、小...
ModelArts Lite集群规格	modelarts_bm_npu_arm_8x19b1	字符串	指定MA Lite Cluster节点flavor ID。
ModelArts Lite集群创建节...	1	整型	指定对应flavor的资源数量。
部署工具实例密码	*****	字符串	部署工具实例的管理员密码, 密码复杂程度要求: 密码要求长度范围为8到24位, 密码至少必须包含大写字母、小写字母...
OBS桶名称	ds-file-bucket-app-0820	字符串	指定OBS桶名称。
VPC IPv4网段	192.168.0.0/16	字符串	取值范围: 10.0.0.0/8 - 10.255.255.0/24, 172.16.0.0/12 - 172.31.255.0/24, 或者 192.168.0.0/16 - 192.168.255.0/24,
子网 IPv4网段	192.168.10.0/24	字符串	必须是CIDR格式, 且在VPC的CIDR块内, 子网掩码不能大于28。
子网网关	192.168.10.1	字符串	子网的网关, 必须是子网范围内的合法IP地址。
付费类型	postPaid	字符串	prePaid-预付费, 即包年包月, postPaid-后付费, 即按量付费, ModelArts Lite集群仅支持包年包月!
订购周期类型	month	字符串	当chargingMode为prePaid时生效且为必填值, 取值范围: month-月, year-年。

资源栈设置

IAM权限委托	回收	未开启	删除保护	未开启
--	--	--	--	--

费用预估: 创建执行计划 (免费) 后可获取预估费用

上一步 创建执行计划

6) 点击“查看费用明细”，确认待创建的资源 and 费用信息，如果没有问题，点击“部署”。

费用明细

以下费用为参考价格, 具体扣费以账单为准了解计费详情, 其中部分资源暂不支持定价, 请您前往价格计算器计算费用。

费用总计 包年/包月模式费用预估: ¥88,061.30 | 按需计费模式预估: ¥0.72/小时 + ¥0.80/GB

其中部分资源暂不支持定价, 具体请查看如下待支持表格

包年/包月	按需计费	免费	待支持				
云产品名称	逻辑名称	区域	购买时长	数	原价	优惠详情	预估优惠后价格
云引擎引擎	this	西南-贵阳一	1个月	1	¥4,671.30	¥0.00	¥4,671.30
AI开发平台	resource_pool	西南-贵阳一	1个月	1	¥83,390.00	¥0.00	¥83,390.00

导出价格清单

关闭

7) 在“事件”的TAB页，您可以查看部署的过程

时间	操作	描述	资源名称	资源ID
2025-03-12 20:47:59 (GMT+08:00)	成功	创建 vpc huaweicloud_vpc_vpc: Creation complete after 1s [id=9364830c-2956-4734-800c-2035148701f1]	vpc	9364830c-2956-4734-800c-2035148701f1
2025-03-12 20:47:59 (GMT+08:00)	成功	创建 huaweicloud_security_group.huaweicloud_networking_vpc_subnet_group_vpc_subnet: Creation complete after 1s [id=28844776-3275-4d73-822a-23885020264f]	security_group	28844776-3275-4d73-822a-23885020264f
2025-03-12 20:47:59 (GMT+08:00)	正在	创建 huaweicloud_vpc_peering_connection_vpc_peering_connection: Creating.	vpc_peering_connection	-
2025-03-12 20:47:59 (GMT+08:00)	正在	创建 huaweicloud_vpc_peering_connection_vpc_peering_connection: Creating.	vpc_peering_connection	-
2025-03-12 20:47:59 (GMT+08:00)	正在	创建 huaweicloud_vpc_vpc: Creating.	vpc	-
2025-03-12 20:47:59 (GMT+08:00)	正在	创建 huaweicloud_vpc_vpc: Creating.	vpc	-
2025-03-12 20:47:59 (GMT+08:00)	正在	创建 huaweicloud_vpc_vpc: Creating.	vpc	-
2025-03-12 20:47:59 (GMT+08:00)	成功	创建 huaweicloud_security_group.huaweicloud_networking_vpc_subnet_group_vpc_subnet: Creation complete after 3s [id=0822588c-6386-4853-8482-82727a6c2262]	security_group	0822588c-6386-4853-8482-82727a6c2262
2025-03-12 20:47:59 (GMT+08:00)	正在	创建 huaweicloud_networking_vpc_subnet_group_vpc_subnet: Creating.	networking_vpc_subnet_group	-
2025-03-12 20:47:59 (GMT+08:00)	正在	创建 huaweicloud_networking_vpc_subnet_group_vpc_subnet: Creating.	networking_vpc_subnet_group	-
2025-03-12 20:47:59 (GMT+08:00)	正在	创建 huaweicloud_networking_vpc_subnet_group_vpc_subnet: Creating.	networking_vpc_subnet_group	-
2025-03-12 20:47:59 (GMT+08:00)	正在	创建 huaweicloud_networking_vpc_subnet_group_vpc_subnet: Creating.	networking_vpc_subnet_group	-

8) 部署成功后，您可以在“资源”的TAB页查看部署过程中创建的云资源，如下图所示。如果您需要查看云资源的详情，可直接点击云资源名称跳转到云服务管理台查看。

云产品名称	物理资源名称	逻辑名称	资源类型	资源状态
统一身份认证服务	iam_agency_1247	agency	huaweicloud_iam_iam_iam_iam	成功
虚拟私有云	vpc-1247-vpc	vpc	huaweicloud_networking_vpc_vpc	成功
虚拟私有云	vpc-1247-vpc-subnet	vpc_subnet	huaweicloud_networking_vpc_vpc_subnet	成功
虚拟私有云	vpc-1247-vpc-peering-connection	vpc_peering_connection	huaweicloud_networking_vpc_vpc_peering_connection	成功
虚拟私有云	vpc-1247-vpc-subnet-group	vpc_subnet_group	huaweicloud_networking_vpc_vpc_subnet_group	成功
弹性云服务器	ecs-1247-ecs	ecs	huaweicloud_compute_compute_compute	成功
虚拟私有云	vpc-1247-vpc-subnet	vpc_subnet	huaweicloud_networking_vpc_vpc_subnet	成功

5 模型权重上传

云资源创建成功后，会自动将模型的权重上传到您在“云资源开通”章节中创建的OBS桶，上传过程大概耗时40分钟左右。您可以登录部署工具实例查看部署结果。查看部署结果的命令：`cat /home/ds-deploy/tmp/obs_upload_stat.info`，结果如下图所示：

```
root@app-1026-ecs-mkp:/home/ds-deploy/tmp# cat obs_upload_stat.info
Start at 2025-03-11 10:37:39.728122787 +0000 UTC

Bucket:
  obs://ds-file-bucket-app-1026
StorageClass:
  standard
Location:
  cn-southwest-2
ObsVersion:
  3.0
AvailableZone:
  multi-az
BucketType:
  OBJECT
ObjectNumber:
  25
Size:
  38.75GB
Quota:
  0
root@app-1026-ecs-mkp:/home/ds-deploy/tmp#
```

同时可以在OBS管理台确认上传的模型权重，如下图所示。

名称	存储类别	大小
<input type="checkbox"/> modelscope_test	--	38.75 GB 2025/03/11 18:40:59 GMT+08:00

6 容器镜像上传和加载

云资源创建成功后，会自动将部署所需的docker镜像从部署工具实例发送到CCE的Node节点，然后做自动化的加载。发送时间约10分钟左右，加载时间约30分钟。您可以直接登录CCE Node节点查看docker镜像加载过程，运行docker images命令查看docker镜像加载结果，如下图所示。

```
root@app-0905-cce-mkp-10774:~# docker images
REPOSITORY                                TAG                IMAGE ID           CREATED           SIZE
swr.cn-southwest-2.myhuaweicloud.com/marketplace-ds/pytorch_2_1_ascend  pytorch_2.1.0    59c630b40f77      2 months ago    44.5GB
```

7 模型部署

进入/home/ds-deploy/scripts目录下，执行deploy_ds.sh，下载部署所需的helm charts，如下图所示：

```
root@app-1026-ecs-mkp:/home/ds-deploy/scripts# ./deploy_ds.sh
--2025-03-11 19:44:45-- https://ad-deepseek.obs.cn-southwest-2.myhuaweicloud.com/ds-deploy/mkp-ds-deploy-chart.tar
Resolving ad-deepseek.obs.cn-southwest-2.myhuaweicloud.com (ad-deepseek.obs.cn-southwest-2.myhuaweicloud.com)... 100.125.81.35,
100.125.81.3, 100.125.81.67
Connecting to ad-deepseek.obs.cn-southwest-2.myhuaweicloud.com (ad-deepseek.obs.cn-southwest-2.myhuaweicloud.com)|100.125.81.35:
443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 30720 (30K) [application/x-tar]
Saving to: 'mkp-ds-deploy-chart.tar'

mkp-ds-deploy-chart.tar      100%[=====] 30.00K  --.-KB/s  in 0.001s

2025-03-11 19:44:45 (24.1 MB/s) - 'mkp-ds-deploy-chart.tar' saved [30720/30720]

--2025-03-11 19:44:45-- https://ad-deepseek.obs.cn-southwest-2.myhuaweicloud.com/ds-deploy/values.yaml
Resolving ad-deepseek.obs.cn-southwest-2.myhuaweicloud.com (ad-deepseek.obs.cn-southwest-2.myhuaweicloud.com)... 100.125.81.67,
100.125.81.3, 100.125.81.35
Connecting to ad-deepseek.obs.cn-southwest-2.myhuaweicloud.com (ad-deepseek.obs.cn-southwest-2.myhuaweicloud.com)|100.125.81.67:
443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 205 [text/yaml]
Saving to: 'values.yaml'

values.yaml                  100%[=====] 205  --.-KB/s  in 0s

2025-03-11 19:44:45 (175 MB/s) - 'values.yaml' saved [205/205]

./mkp-ds-deploy-chart/
./mkp-ds-deploy-chart/Dchart.yaml
./mkp-ds-deploy-chart/.helmignore
./mkp-ds-deploy-chart/values_bak.yaml
./mkp-ds-deploy-chart/charts/
./mkp-ds-deploy-chart/templates/
./mkp-ds-deploy-chart/templates/deploy_pv.yaml
./mkp-ds-deploy-chart/templates/deploy_service.yaml
./mkp-ds-deploy-chart/templates/deploy_ds.yaml
./mkp-ds-deploy-chart/templates/_helpers.tpl
./mkp-ds-deploy-chart/templates/deploy_pvc.yaml
./mkp-ds-deploy-chart/templates/deploy_secret.yaml
./mkp-ds-deploy-chart/values.yaml
root@app-1026-ecs-mkp:/home/ds-deploy/scripts# ls
deploy_ds.sh  mkp-ds-deploy-chart  mkp-ds-deploy-chart.tar  myValues.yaml
root@app-1026-ecs-mkp:/home/ds-deploy/scripts#
```

修改myValues.yaml，在其中填写您部署的区域、账号的ak、sk等信息，如下图所示，AK和SK的获取可参考：https://support.huaweicloud.com/usermanual-ca/ca_01_0003.html

```
# values.yaml
region: region

access:
  key: myak
secret:
  key: mysk

pv:
  obsbucket: obsbucket

deployment:
  replicas: 1
```

然后运行helm install ds-deploy ./mkp-ds-deploy-chart -f myValues.yaml，如下图所示。整个的部署过程大概耗时10分钟。

```
root@ecs-6d88:/home/ds-deploy/scripts# helm install ds-deploy ./mkp-ds-deploy-chart -f myValues.yaml
NAME: ds-deploy
LAST DEPLOYED: Tue Mar 11 20:10:15 2025
NAMESPACE: default
STATUS: deployed
REVISION: 1
TEST SUITE: None
```

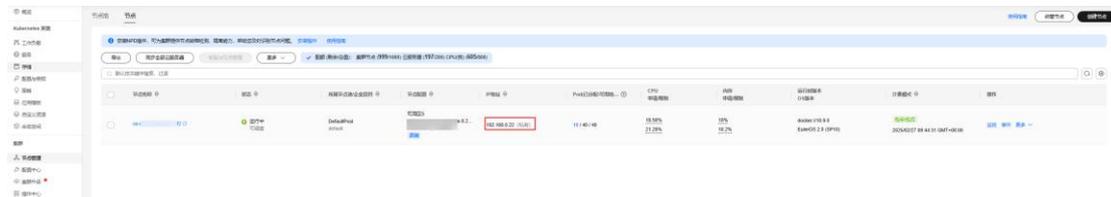
您可以进入CCE管理台，查看创建的工作负载、服务、存储卷、配置与密钥等。后续服务的启动、停止、扩容、升级等操作可直接在CCE管理台操作，详情可参考CCE的文档：

https://support.huaweicloud.com/productdesc-cce/cce_productdesc_0022.html



8 部署结果验证

部署成功后，进入CCE管理台，Node节点的内网IP，如下图所示。



您可以直接使用NodeIP直接访问，完成的API URL：

<http://NodeIp:30000/v1/chat/completions>，示例命令如下：

```
curl -X POST http://NodeIp:30000/v1/chat/completions --header 'Content-Type: application/json' --data '{"model":"DS-Distill-Qwen-32B","messages":[{"role":"user","content":"你好"}],"max_tokens":1000,"temperature":0.6,"top_p":0.95,"stream":false}'
```

测试结果如下，代表部署成功

```
l-Qwen-32B":{"messages":[{"role":"user","content":"你好"}],"max_tokens":1000,"temperature":0.6,"top_p":0.95,"stream":false} [{"role":"assistant","content":"您好！很高兴为您服务。有什么我可以帮您解答的吗？\n\n"}]
```

如果您需要在集群内访问，可以直接使用Service的内部域名替代NodeIp,如下图所示。



9 云资源清理

如果您不在需要使用，可直接进入RFS管理平台删除资源栈，自动部署产生的云资源会自动删除，如下图所示，详情可参考：https://support.huaweicloud.com/usermanual-aos/af_04_0007.html

