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

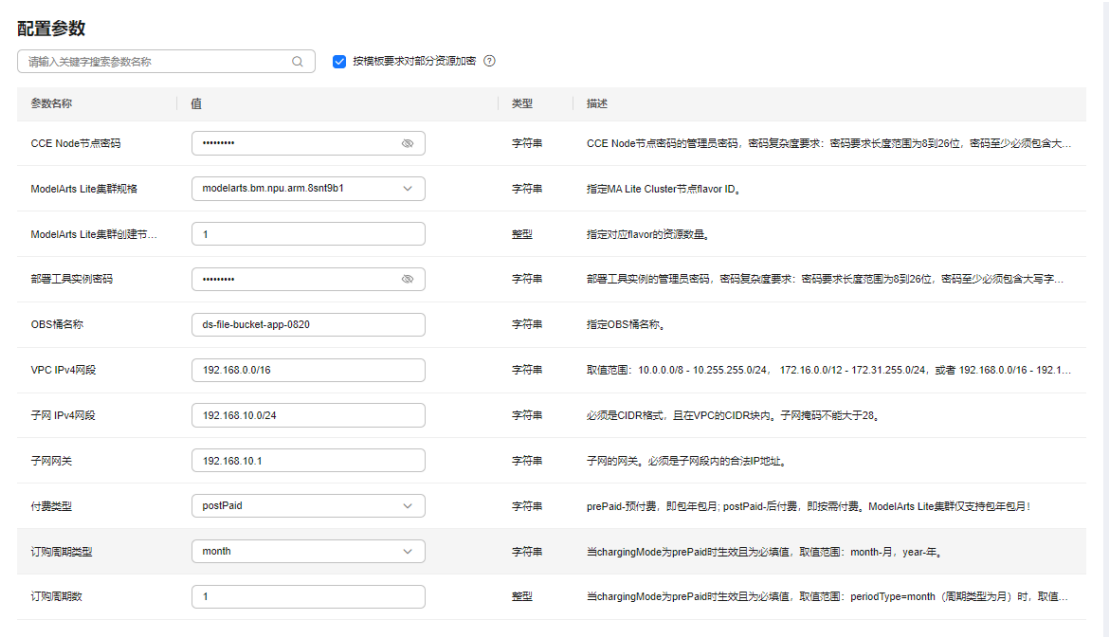
1）配置ModelArts的授权

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。

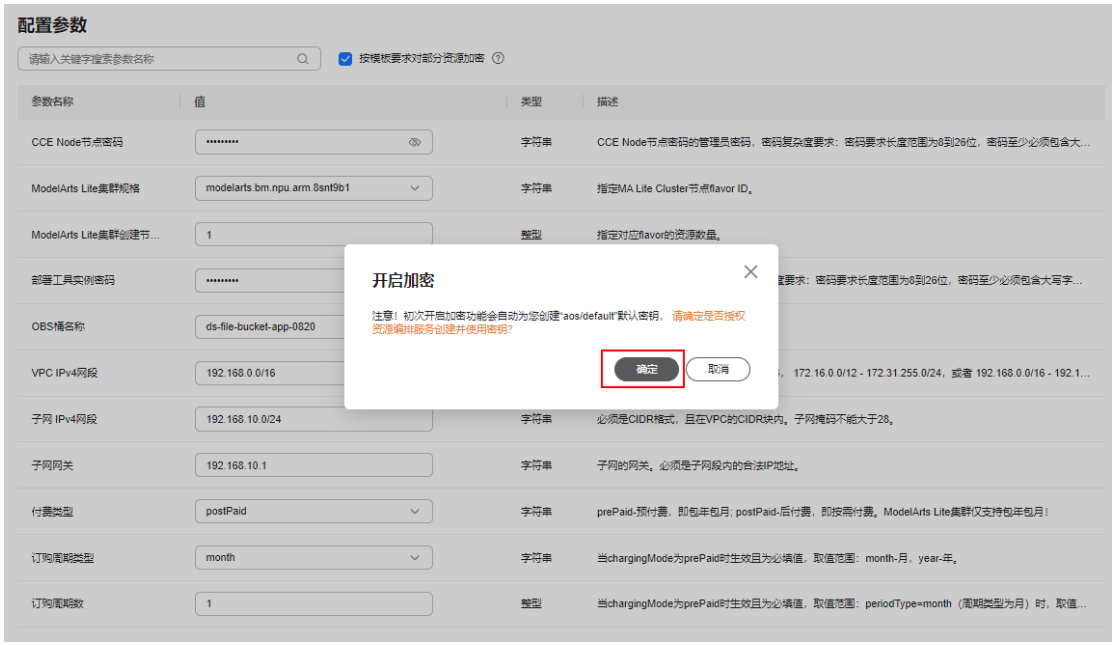
4 云资源开通

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。

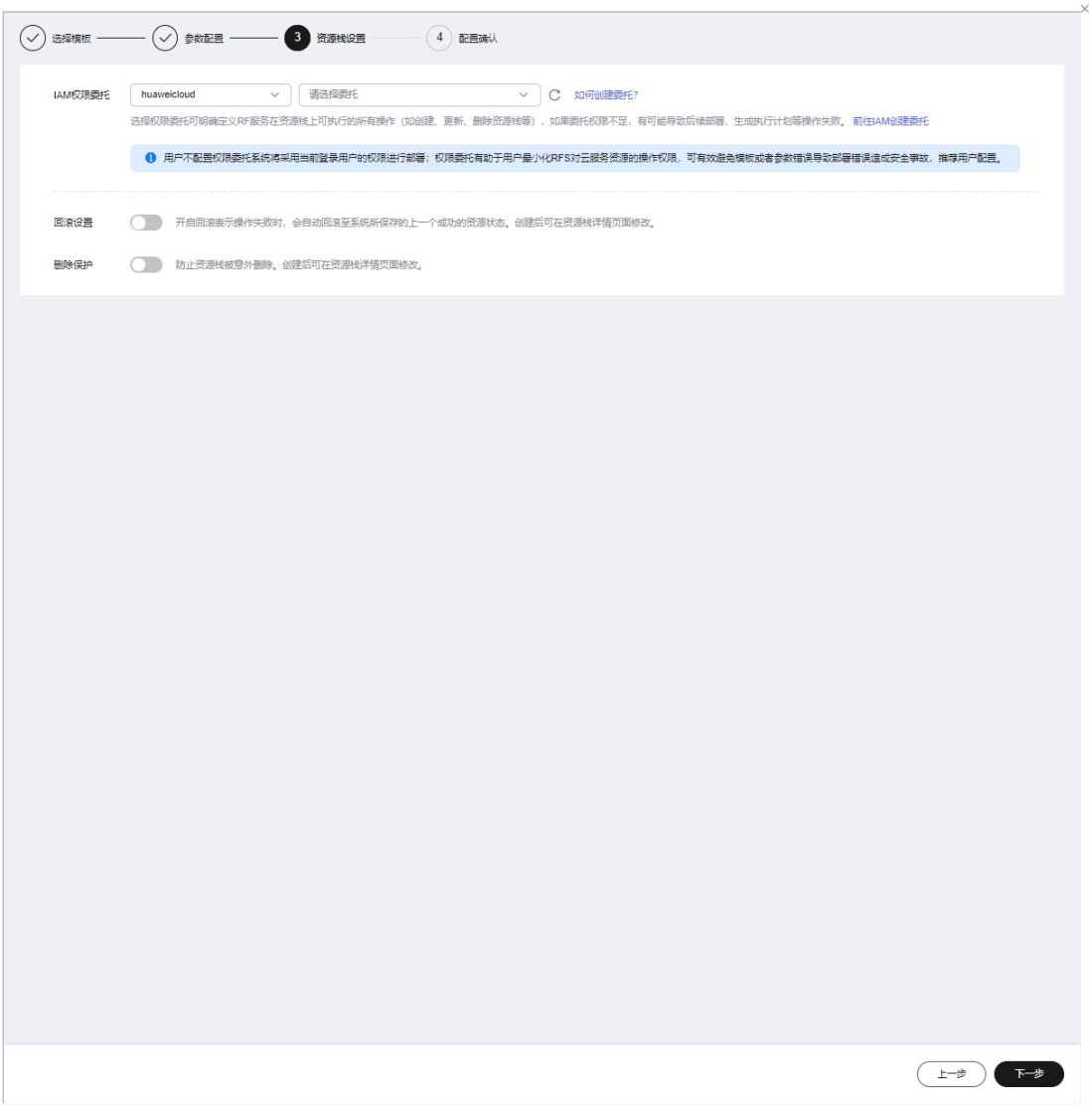
2）选择“模板配置开通”，然后直接点击“下一步”，进入参数配置页面，如下图所示。



3）首次开通时，需要您授权RFS创建并使用秘钥加密敏感参数，点击：“确定”



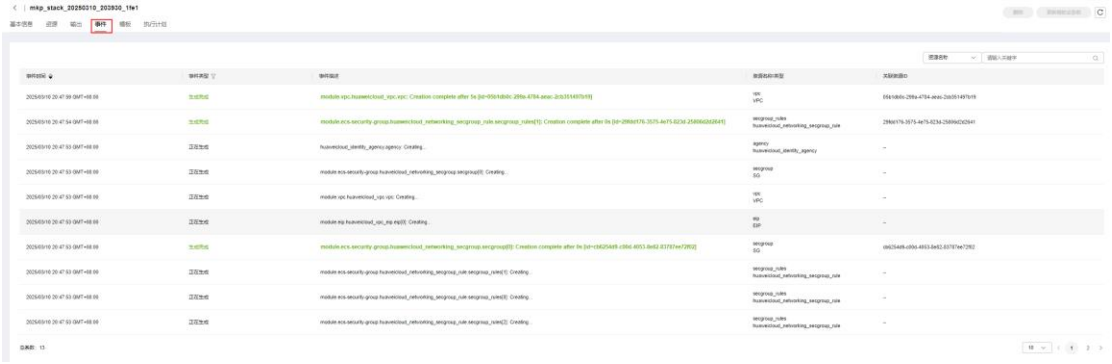
4）进入“资源栈设置”页面，然后点击“下一步”



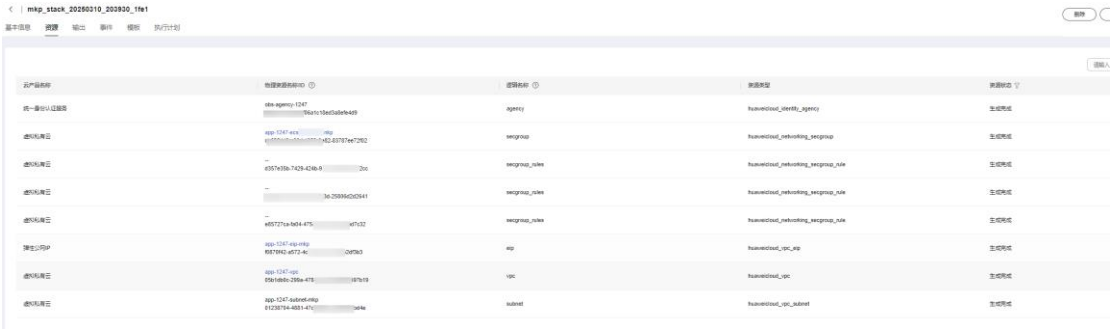
5）进入配置确认页面，需要您确认下配置参数的值，如果没有问题点击“创建执行计划”。

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

7）在“事件”的TAB页，您可以查看部署的过程



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。



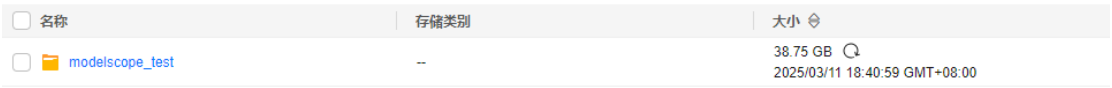
5 模型权重上传

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。



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。

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/Chart.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

```yaml
# 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。

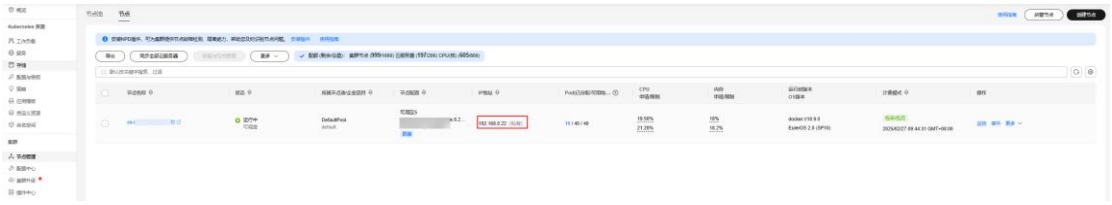
```
root@ecs-6d88:/home/ds-deploy/scripts# helm install ds-deploy ./mkp-ds-deploy-chart -f myValues.yaml
NAME: ds-deploy
LAST DEPLOYED: Tue Mar 11 20:10:15 2025
NAMESPACE: default
STATUS: deployed
REVISION: 1
TEST SUITE: None
```

您可以进入CCE管理台，查看创建的工作负载、服务、存储卷、配置与秘钥等。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

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/rf_04_0007.html