RWKV使用指南

1 商品说明

RWKV（Receptance Weighted Key Value）是一种结合RNN和Transformer优势的新型大语言模型架构，由元始智能团队主导开发并开源。

本商品通过kunpeng服务器进行安装部署的RWKV-4-Pile-1B5-EngChn-test4-20230115.pth模型

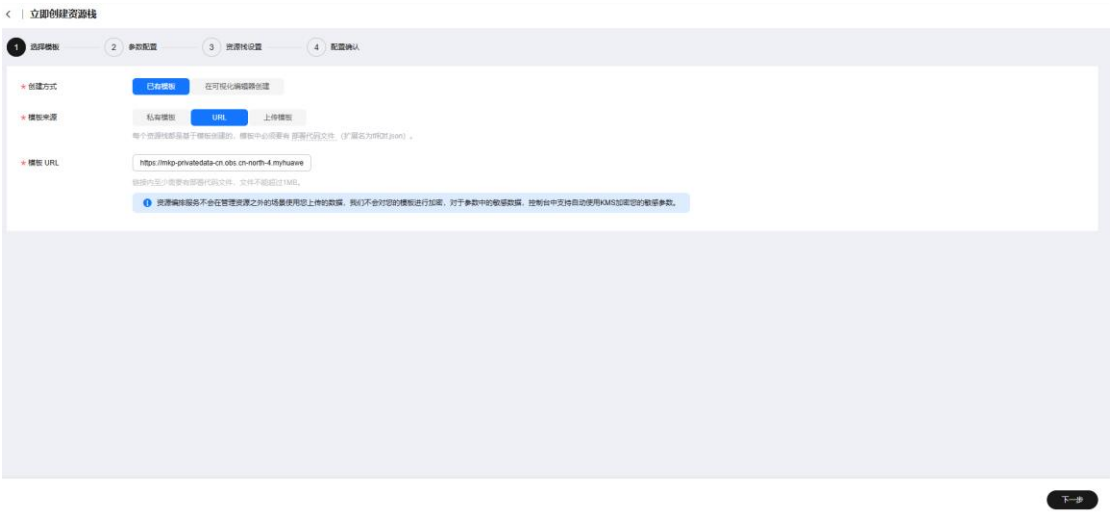
2 商品购买

您可以在云商店搜索“RWKV”。

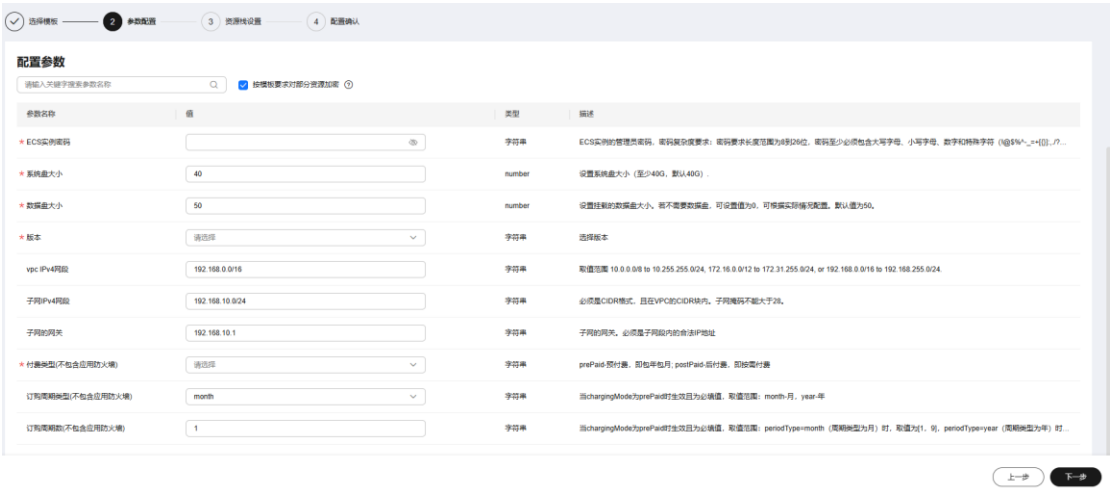
其中，地域、规格、推荐配置使用默认，购买方式根据您的需求选择按需/按月/按年，短期使用推荐按需，长期使用推荐按月/按年，确认配置后点击“立即购买”。

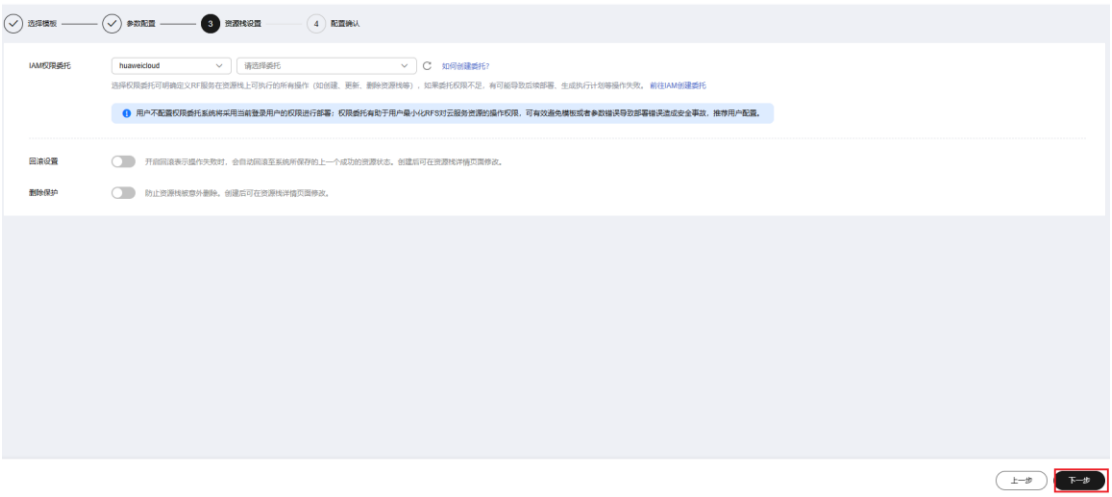
2.1 商品支持自定义 ECS 购买，具体见章节 3.1.1

2.2 使用 RFS 模板直接部署

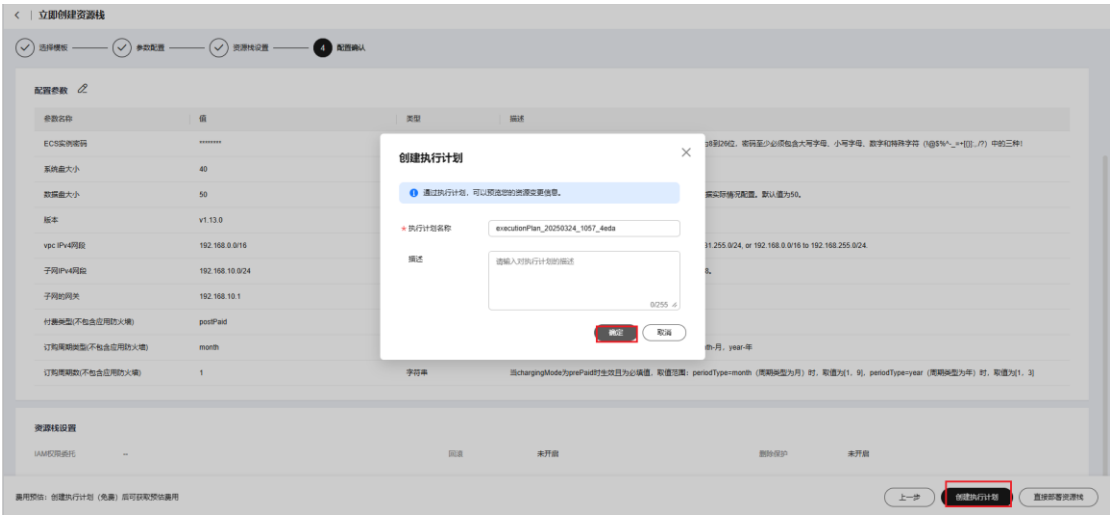


必填项填写后，点击 下一步

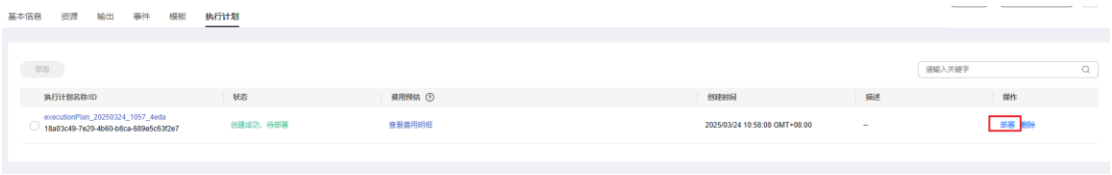




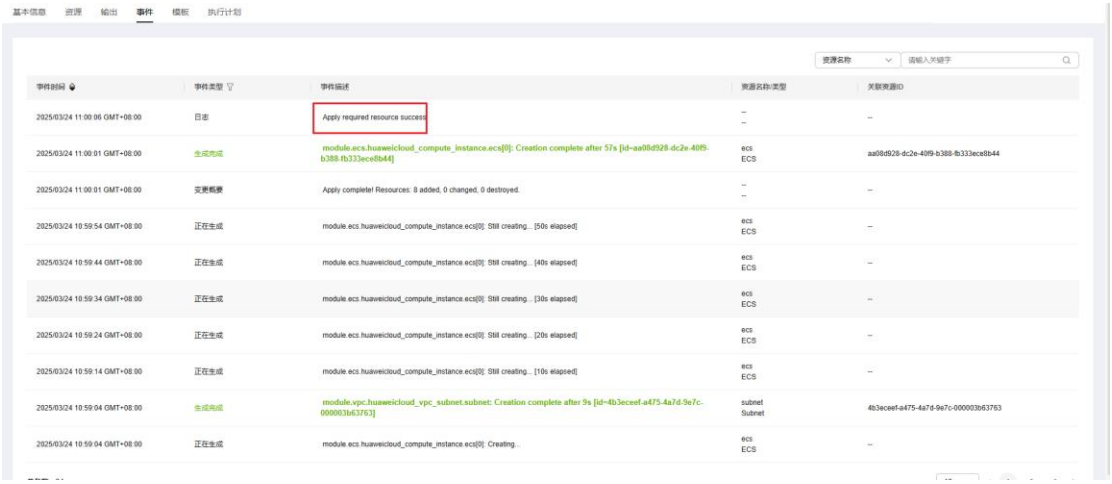
创建直接计划后，点击 确定



点击 部署



如下图“Apply required resource success. ”即为资源创建完成



3 商品资源配置

商品支持**ECS控制台配置**，下面对资源配置的方式进行介绍。

3.1 ECS 控制台配置

3.1.1 准备工作

在使用ECS控制台配置前，需要您提前配置好**安全组规则**。

安全组规则的配置如下：

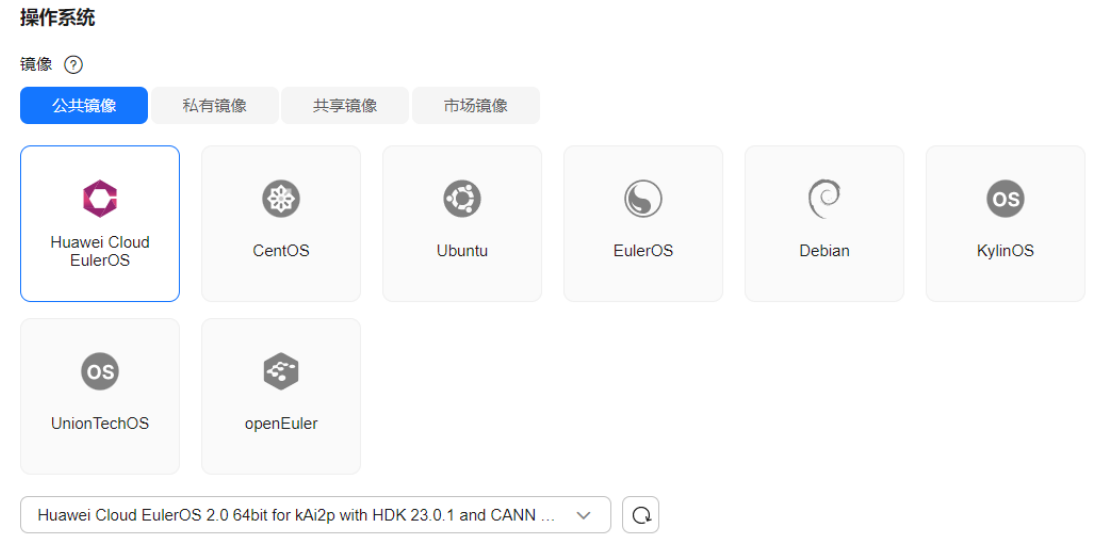
- 入方向规则放通CloudShell连接实例使用的端口22，以便在控制台登录调试。
- 出方向规则一键放通

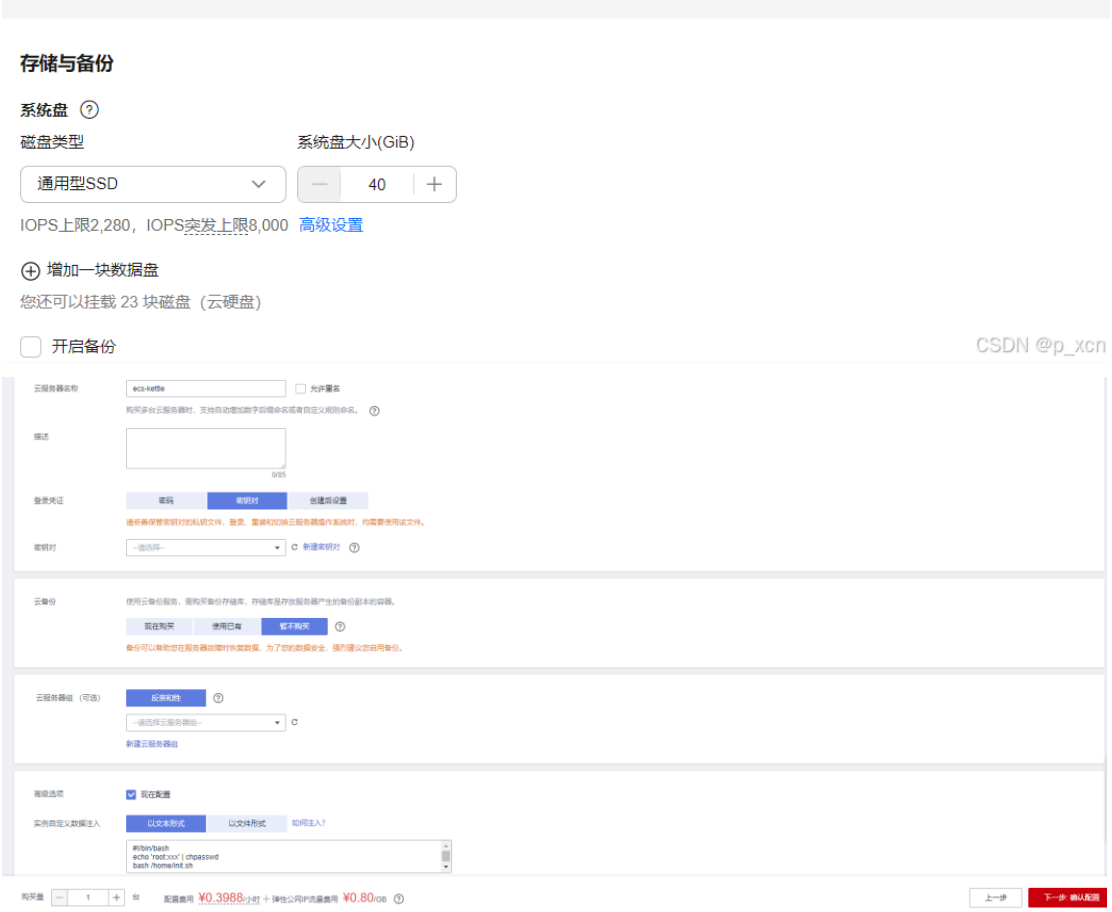
3.1.2 创建 ECS

前提工作准备好后，选择ECS控制台配置跳转到购买ECS页面，ECS资源的配置如下图所

示：


基础配置
计费模式
包年/包月
按需计费
竞价计费
按需计费实例不支持备案。了解备案限制
区域
华北-北京四
推荐区域 华北-北京四 华南-广州 华东-上海一 华北-乌兰察布一 西南-贵阳一
云服务器创建后无法更改区域；不同区域之间内网互不相通，请就近选择靠近您业务的区域，减少网络时延。如何选择区域
可用区
随机分配
可用区1
可用区2
可用区3
可用区7
随机至多可用区
实例
规格类型选型
业务场景选型
CPU架构
x86计算
鲲鹏计算
实例筛选
--请选择vCPUs--
--请选择内存--
请输入规格名称模糊搜索
隐藏售罄的规格
鲲鹏通用计算增强型
鲲鹏内存优化型
鲲鹏超高I/O型
CSDN @p_xcn





值得注意的是：

- VPC您可以自行创建
- 安全组选择3.1.1章节中配置的安全组
- 弹性公网IP选择现在购买，推荐选择“按流量计费”，带宽大小可设置为5Mbit/s
- 高级配置需要在高级选项支持注入自定义数据，所以登录凭证不能选择“密码”，选

择创建后设置

- 其余默认或按规则填写即可。

4 商品使用

4.1 RWKV 使用

4.1.1 激活环境

RWKV模型主要的通途包含文本生成、对话系统等，测试使用的模型文件主要是用来和机器人进行聊天，会获得对应的回答。

登录到服务器上运行conda activate py39

在/opt/ChatRWKV下执行python v2/app.py 运行之后先选择语言，然后选择要使用的处理器，CPU推理就选择CPU，最后选择需要进行推理的模型。

```
(base) [root@rwkv ~]# conda activate py39
(py39) [root@rwkv ~]# cd /opt/ChatRWKV
(py39) [root@rwkv ChatRWKV]# python v2/app.py

Please choose the language.
0 = English
1 = 简体中文
Waiting for the language ref (etc. 0): 1

Please choose the mode.
0 = GPU
1 = CPU
Waiting for the mode ref (etc. 0): 1

Please choose the model.
0 = .cache
1 = RWKV-4-Pile-1B5-EngChn-test4-20230115.pth
Waiting for the model ref (etc. 0): 1

Loading ChatRWKV - Chinese - cpu - fp32 - QA_PROMPT False

RWKV_JIT_ON 1

Loading model - /opt/models/RWKV-4-Pile-1B5-EngChn-test4-20230115
blocks.0.ln1.weight             float32    cpu    2048
blocks.0.ln1.bias               float32    cpu    2048
blocks.0.ln2.weight             float32    cpu    2048
blocks.0.ln2.bias               float32    cpu    2048
blocks.0.att.time_decay         float32    cpu    2048
```

4.1.2 推理结果

```
Run prompt...
### prompt ###
[
The following is a verbose and detailed conversation between an AI assistant called Bot, and a human user called User. Bot is intelligent, knowledgeable, wise and polite.

User: french revolution what year

Bot: The French Revolution started in 1789, and lasted 10 years until 1799.

User: 3+5=?

Bot: The answer is 8.

User: guess i marry who ?

Bot: Only if you tell me more about yourself - what are your interests?

User: solve for a: 9-a=2

Bot: The answer is a = 7, because 9 - 7 = 2.

User: wat is lhc

Bot: LHC is a high-energy particle collider, built by CERN, and completed in 2008. They used it to confirm the existence of the Higgs boson in 2012.

]

User: wat is nlp
User: wat is nlp
Bot: the natural language processing language of smart phones, and its use on the internet. It is the main form of communication for modern day computers and smartphones.
```

4.2 参考文档

- RWKV参考文档