

FastGPT企业级智能AI模型问答知识库使用指南

1 商品说明

FastGPT 是一个 AI Agent 构建平台，提供开箱即用的数据处理、模型调用等能力，同时可以通过 Flow 可视化进行工作流编排，从而实现复杂的应用场景！

本商品通过鲲鹏服务器基于Huawei Cloud EulerOS 2.0 标准版 64位进行安装部署

2 商品购买

您可以在云商店搜索“FastGPT 企业级智能 AI 模型问答知识库”，此商品相关资源配置提供模板化部署，也可根据用户需求自定义购买。

（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



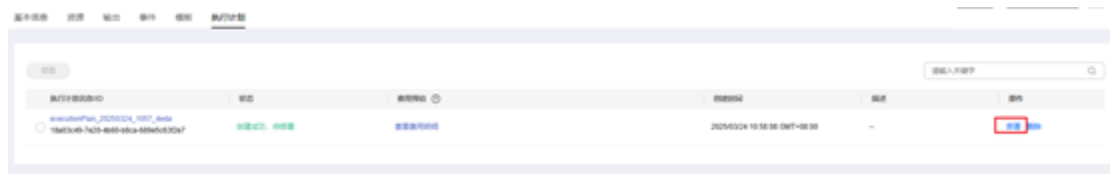
必填项填写后，点击 下一步



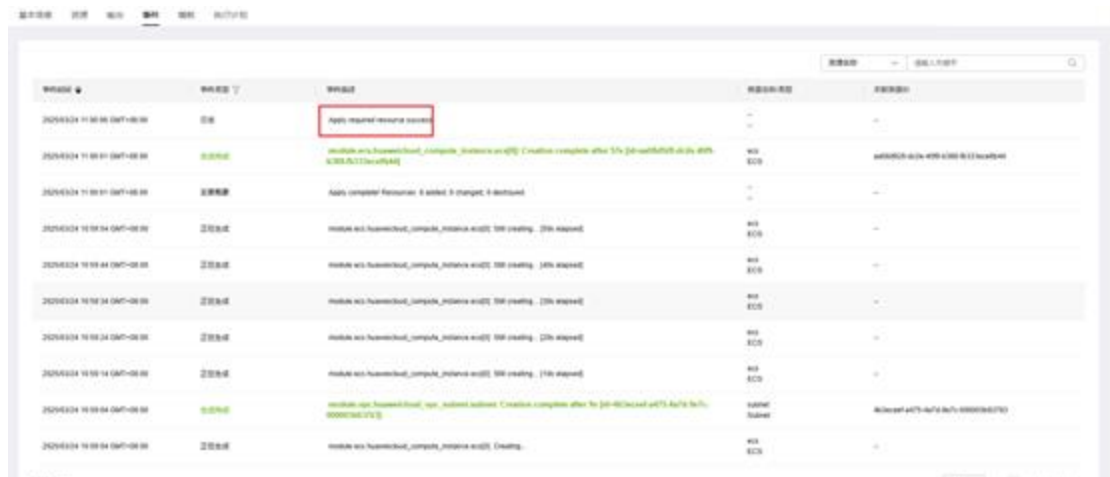
创建直接计划后，点击 确定

[illegible]

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：

- 入方向规则放通端口3000，源地址内必须包含您的客户端ip，否则无法访问
- 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



示:

基础配置

计费模式 ^①

包年包月 [🔥]

按量计费

竞价计费

按量计费实例不支持备机。 [了解详情限制](#)

区域 ^②

华东-上海一

推荐区域

华北-北京四

华南-广州

华东-上海一

西南-贵阳一

华北-乌兰察布一

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 ^③

随机分配

可用区1

可用区4

可用区3

可用区2

☐ 随机至多可用区

x86计算

鲲鹏计算

实例规格 ^④

--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 ☒ 仅显示最新系列 ☐ 仅显示裸金属类型的规格 ^⑤

鲲鹏通用计算增强型

鲲鹏内存优化型

鲲鹏超密HIO型

全部

kc1

kc2

搭载鲲鹏920处理器及25GE智能高速网卡，提供强劲鲲鹏算力和高性能网络，更好满足政府、互联网等各类企业对云上业务高性价比、安全可靠等诉求。

实例类型	规格名称	vCPUs ^⑥	内存 ^⑦	CPU ^⑧	基准 / 最大带宽 ^{⑨ ⑩}	内网收发包 ^⑪	IPv6	规格参考价 ^⑫
☐ 鲲鹏通用计算增强型k...	kc1.small.1	1vCPUs	1GB	Huawei Kunpeng ...	0.5 / 2 Gbit/s	20万PPS	否	¥0.12/小时
☒ 鲲鹏通用计算增强型k...	kc1.large.2	2vCPUs	4GB	Huawei Kunpeng ...	0.8 / 3 Gbit/s	30万PPS	否	¥0.30/小时
☐ 鲲鹏通用计算增强型k...	kc1.large.4	2vCPUs	8GB	Huawei Kunpeng ...	0.8 / 3 Gbit/s	30万PPS	否	¥0.41/小时
☐ 鲲鹏通用计算增强型k...	kc1.xlarge.2	4vCPUs	8GB	Huawei Kunpeng ...	1.5 / 5 Gbit/s	50万PPS	否	¥0.60/小时
☐ 鲲鹏通用计算增强型k...	kc1.xlarge.4	4vCPUs	16GB	Huawei Kunpeng ...	1.5 / 5 Gbit/s	50万PPS	否	¥0.81/小时
☐ 鲲鹏通用计算增强型k...	kc1.2xlarge.2	8vCPUs	16GB	Huawei Kunpeng ...	3 / 7 Gbit/s	80万PPS	否	¥1.20/小时
☐ 鲲鹏通用计算增强型k...	kc1.7xlarge.4	8vCPUs	32GB	Huawei Kunpeng ...	3 / 7 Gbit/s	80万PPS	否	¥1.63/小时

当前规格 鲲鹏通用计算增强型 | kc1.large.2 | 2vCPUs | 4GB

操作系统

镜像 ^⑬

公共镜像

私有镜像

共享镜像

市场镜像

Flink1.13.0-arm-v1.0(Flink-流数据工具)(v1.13.0)

请选择镜像

使用在云市场购买的第三方镜像需要遵守第三方服务协议。请确保您已阅读并同意《华为云商店通用商品用户协议》《Flink-流数据工具服务协议》。

存储与备份

系统盘 ^⑭

磁盘类型

系统盘大小(GIB)

通用型SSD

40

+

IOPS上限2,280，IOPS突发上限8,000 加密：未启用 高级设置

⑮ 增加一块数据盘

您还可以挂载2块VBD磁盘或2块SCSI磁盘。

☐ 开启备份

云备份服务可以帮助您在协助灾、数据误删、合规审查、服务器故障场景恢复数据，为了您的数据安全，建议您开启备份。

公网访问

弹性公网IP ^⑯

现在购买

使用已有

暂不购买

线路 ^⑰

全动态BGP

静态BGP

^⑱ 不低于99.95%可用性保障

公网带宽 ^⑲

按带宽计费 [🔥]

按流量计费

加入共享带宽

流量较大或较稳定的场景 流量小或流量波动较大场景 多业务流量错峰分布场景

指定带宽上限，按实际使用的出公网流量计费，与使用时间无关。

带宽大小 (Mbit/s)

5

10

20

50

100

自定义

带宽范围：1-300 Mbit/s；开启DDoS基础防护 ^⑳ 免费

释放行为

☐ 随实例释放

对于设置了随实例释放的弹性公网IP，将在删除云服务器同时执行删除。



云服务器管理

云服务器名称

ecs-1007 ☐ 允许重名

购买多台云服务器时，支持自动增加数字后缀命名或者自定义规则命名。 [了解更多](#)

登录凭证 ^①

请牢记密码，如忘记密码可登录ECS控制台重置密码。

用户名 密码 确认密码

root *****

标签 ^②

如果您需要使用同一标签标识多种云资源，即所有服务均可在标签输入框下拉选择同一标签，建议在TMS中创建预定义标签。 [查看预定义标签](#)

[+ 添加标签](#)

您还可以添加20个标签。

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。

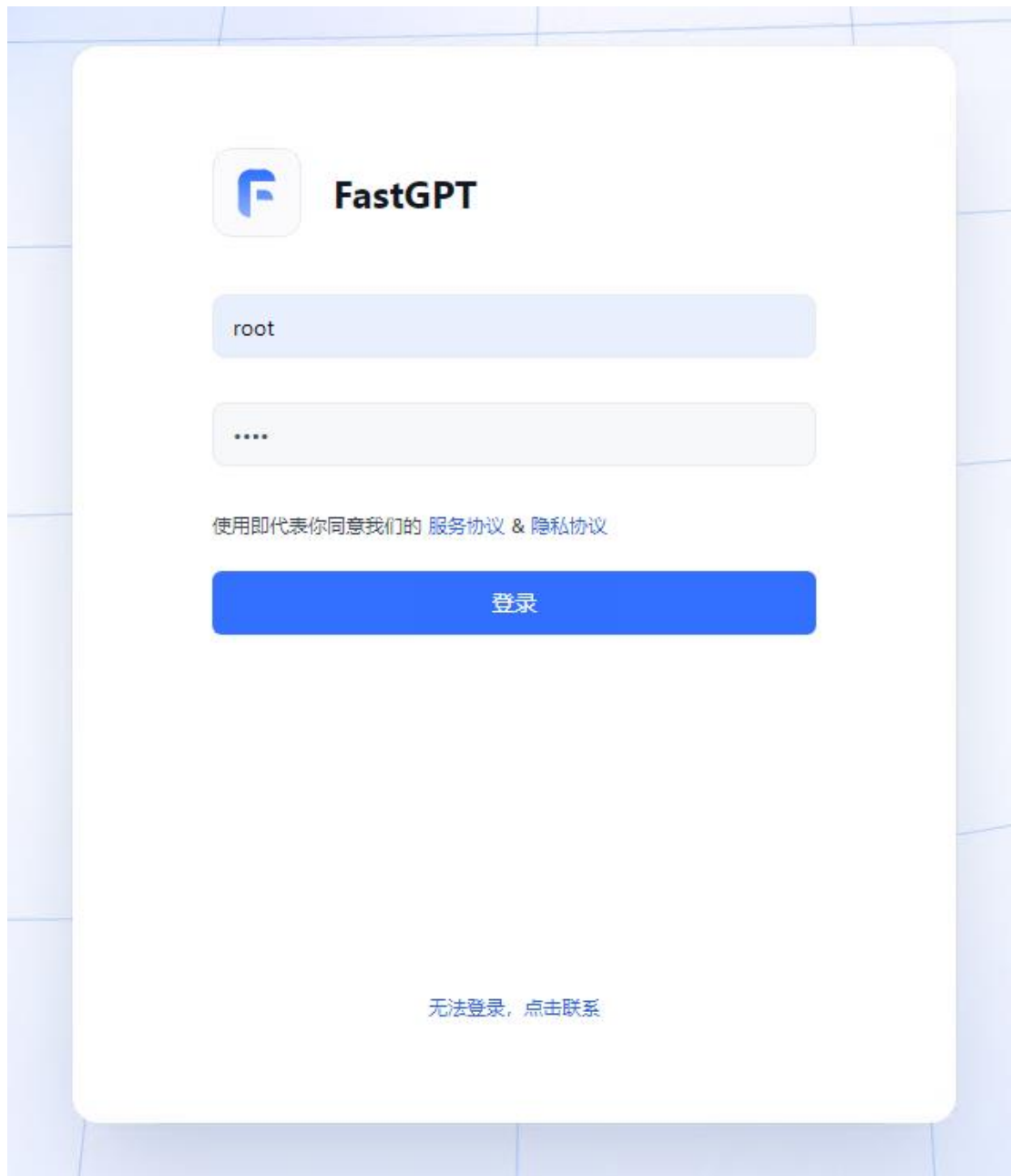
4 商品使用

4.1 登录

地址：http://弹性公网IP:3000

用户名：root

密码：1234



4.2 新增模型

可用模型 模型配置 模型渠道 调用日志		
模型提供商	全部	模型类型
全部	全部	
模型名	模型类型	模型计费
 DeepSeek-V3	语言模型	0 积分 / 1K Tokens
 bce_embedding_base_v1	索引模型	输入 0 积分 / 1K Tokens
 bce-reranker-base_v1	重排模型	-

遵循OpenAI规范的模型服务，选择OpenAI和自定义请求地址

4.3 新建知识库

创建一个通用知识库

取个名字

测试

索引模型 ?

bce_embedding_base_v1

文本理解模型 ?

DeepSeek-V3

图片理解模型

未配置相关模型

关闭

确认创建

4.3.1 导入文档

✓ 选择文件

2 参数设置

3 数据预览

4 确认上传

数据处理方式设置

^

处理方式

☒ 分块存储 ^①

☐ 问答对提取 ^①

分块条件 ^①

原文长度大于

▼

500

^

▼

索引增强

☐ 自动生成补充索引 ^①

☐ 图片自动索引 ^①

分块处理参数

☒ 默认
使用系统默认的参数和规则

☐ 自定义
自定义设置数据处理规则

下一步

4.3.2 创建应用



4.4 参考文档

- [官网文档](#)