

Milvus使用指南

1 商品说明

Milvus 是一款全球领先的开源向量数据库，赋能 AI 应用和向量相似度搜索，加速非结构化数据检索。作为一个专门设计用于处理输入向量查询的数据库，它能够处理万亿级别的向量索引。与现有的关系型数据库主要处理遵循预定义模式的结构化数据不同，Milvus 从底层设计用于处理从非结构化数据转换而来的嵌入向量。

本商品通过鲲鹏服务器+EulerOS2.0进行安装部署

2 商品购买

您可以在云商店搜索“Milvus”。

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





返回

配置参数

名称：云原生应用部署环境

新增配置项并默认使用默认值

参数名称	值	类型	描述
* 主机操作系统		字符串	ECI实例的操作系统，该值为必填项。默认值为alinux3.90。该值以公共仓库为准，小写字母、数字和特殊字符（正则表达式为[a-z0-9_+~@.-]*）。
* 系统盘大小	40	number	设置系统盘大小（范围40G~160G）。
* 数据盘大小	50	number	设置数据盘的容量大小，但不需要数据盘，可设置为0G，可根据实际需求调整，最小单位为1G。
* 版本	最新	字符串	应用版本
cpu核数限制	102.100.0.0/16	字符串	配置范围 10.0.0.0/16至10.200.0.0/24、172.16.0.0/12至172.31.255.0/24、或102.100.0.0/16至102.100.255.0/24
子网vpc网络	102.100.10.0/24	字符串	必须为CIDR格式，符合VPC/CIDR规范，子网掩码不能大于31。
子网路由表	102.100.10.1	字符串	子网路由表，必须为子网范围内的路由标识
* 容器镜像（不由应用注入）	最新	字符串	profnet探针镜像，即电声元月_profnet探针，探针探针
(有策略限制)不由应用注入	month	字符串	当chargeableType为profnet时生效且必选；取值范围：month-月、year-年
(有策略限制)不由应用注入	1	字符串	当chargeableType为profnet时生效且必选；取值范围：periodType-month（周期类型为月）、周、quarter-year（周期类型为季）、年、

上一步

下一步

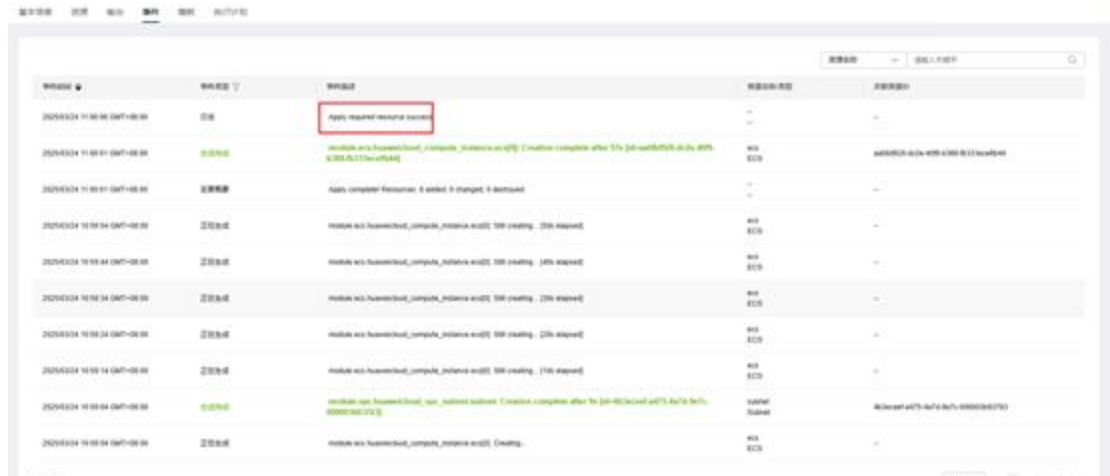
[illegible]

The screenshot shows the 'Configuration Center' (配置中心) interface. A modal dialog titled '创建执行计划' (Create Execution Plan) is open. The dialog contains a blue tip bar with the text: '通过执行计划，可以预发您的配置变更信息。' (Through the execution plan, you can pre-release your configuration change information.). Below this, there is a red label '执行计划名称' (Execution Plan Name) and a text input field containing 'executionPlan_20250214_1007_test'. To the right of the input field is a '清除' (Clear) button. Below the name field is a label '描述' (Description) and a larger text area containing '测试执行计划执行计划描述'. To the right of the text area is a '清除' (Clear) button. At the bottom of the dialog are two buttons: '确定' (Confirm) and '取消' (Cancel). The background shows a table with configuration items and a sidebar with navigation options like '配置项列表' (Configuration Item List) and '配置项详情' (Configuration Item Details).

第2页, 共7页



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9091、19530，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：

弹性云服务器 / 购买弹性云服务器

<  购买弹性云服务器

1 购买页全面升级，诚邀您参与问卷调查，精选反馈，尊享好礼回馈！[立即参与](#)

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架。[了解备架限制](#)

区域 

华北-北京四

推荐区域 华北-北京四 (31) 华南-广州 (54) 华东-上海一 (6) 西南-贵阳一 (11) 华东-上海二

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 ☐ 仅显示黄金类型的规格 

鲲鹏通用计算增强型 鲲鹏内存优化型 鲲鹏超密IO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 	内存 	CPU 	基准 / 最大带宽  	内网收发包  	IPv6	规格参考价  
☐ 鲲鹏通用计算增强型k...	kc1.small.1	1vCPUs	1GiB	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	4GiB	Huawei Kunpeng ...	0.8 / 3 Gbit/s	30万PPS	是	¥0.30/小时
☐ 鲲鹏通用计算增强型k...	kc1.large.4	2vCPUs	8GiB	Huawei Kunpeng ...	0.8 / 3 Gbit/s	30万PPS	是	¥0.41/小时
☒ 鲲鹏通用计算增强型k...	kc1.xlarge.2	4vCPUs	8GiB	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	16GiB	Huawei Kunpeng ...	1.5 / 5 Gbit/s	50万PPS	是	¥0.81/小时
☐ 鲲鹏通用计算增强型k...	kc1.2xlarge.2	8vCPUs	16GiB	Huawei Kunpeng ...	3 / 7 Gbit/s	80万PPS	否	¥1.20/小时
☐ 鲲鹏通用计算增强型k...	kc1.2xlarge.4	8vCPUs	32GiB	Huawei Kunpeng ...	3 / 7 Gbit/s	80万PPS	是	¥1.63/小时



操作系统

镜像 ②

公共镜像

私有镜像

共享镜像

市场镜像

Milvus2.5.10-arm64-v1.0-HCE2.0(40GiB)



新建私有镜像 ②

使用私有镜像创建弹性云服务器前，请查看[操作系统已知问题](#) ②☒ 开启安全防护 ②

主机安全为您提供风险预防、入侵检测、高级防御、安全运营、网页防篡改等安全防护，构建云服务器安全体系。

基础防护

口令检测、漏洞检测等基础防护

免费一个月

企业版

漏洞修复、病毒查杀、等保必备

¥0.18/台/小时

试用1个月后，主机安全服务自动停止服务。

存储与备份

系统盘 ②

磁盘类型

通用型SSD

系统盘大小(GiB)

— 40 +

IOPS上限2,280，IOPS突发上限8,000

加密：未启用

[高级设置](#)

⊕ 增加一块数据盘

您还可以挂载22块VBD磁盘或22块SCSI磁盘。

网络

虚拟私有云 ②

ecs-vpc-default(192.168.0.0/16)



新建虚拟私有云 ②

主网卡

ecs-subnet-default(192.168.10.0/24)



自动分配IP地址



可用私有IP数量250个

⊕ 新增扩展网卡

您还可以增加 2 块网卡

源/目的检查 ②



安全组

选择安全组 ②

default(cd587ac7-a9f8-4840-a56d-561fbc7acce) X



新建安全组

请确保所选安全组已放通22端口（Linux SSH登录），3389端口（Windows远程登录）和ICMP协议（Ping）。[配置安全组规则](#) ②建议将安全组入方向规则中高危端口的源地址设置为已知IP地址、安全组或IP地址组，避免因网络入侵出现业务中断、数据泄露或数据勒索等严重后果。[查看高危端口](#) ②

展开安全组规则 ▾

公网访问

弹性公网IP ①

现在购买

使用已有

暂不购买

线路 ②

全动态BGP

静态BGP

☒ 不低于99.95%可用性保障

公网带宽 ③

**按带宽计费**
流量较大或较稳定的场景**按流量计费**
流量小或流量波动较大场景**加入共享带宽**
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 ④ 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。

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10Mbit/s
- 其余默认或按规则填写即可。

4 商品使用

4.1 Milvus 使用

访问后台使用 <http://IP:9091/webui>程序及客户端连接使用 <http://IP:19530>

4.1.1 创建数据库

创建数据库

```
from pymilvus import MilvusClient

client = MilvusClient(
    uri="http://localhost:19530",
    token="root:Milvus"
)

client.create_database(
    db_name="my_db"
)

✓ 3.2s
```

4.2 参考文档

[官方文档](#)