

hadoop分布式计算平台使用指南

1 商品说明

Hadoop 是一个开源的分布式计算平台，用于处理大规模数据的存储和计算。它最初由 Apache 基金会开发，灵感来源于 Google 的分布式系统论文（如 GFS、MapReduce），旨在通过普通硬件构建分布式集群，实现高可靠性、高扩展性、低成本的数据处理能力

本商品通过鲲鹏服务器安装部署的 hadoop3.3.6 版本。

2 商品购买

您可以在云商店搜索“hadoop分布式计算平台”。

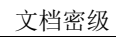
其中，地域、规格、推荐配置使用默认，购买方式根据您的需求选择按需/按月/按年，短期使用推荐按需，长期使用推荐按月/按年，确认配置后点击“立即购买”。

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

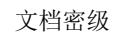


必填项填写后，点击 下一步



创建直接计划后，点击 确定

点击 部署



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：

购买弹性云服务器

返回

完成购买

基础配置

网络配置

高级配置

确认配置

区域

华北-北京四

选择区域

华北-乌兰察布一

华北-北京四

西南-贵阳一

华南-广州

华东-上海一

不同区域的云服务器产品之间网络互不连通。请就近选择就近业务部署的区域，可减少网络时延，提高访问速度。如何选择区域

计费模式

包年包月

按量计费

竞价计费

可用区

随机分配

可用区1

可用区2

可用区3

可用区7

实例规格

规格类型选型

业务场景选型

CPU架构

x86计算

鲲鹏计算

规格

最新系列

vCPUs

请选择vCPUs

内存

请选择内存

规格名称

清除规格的名称

通用计算增强型

通用计算型

内存优化型

超大内存型

磁盘增强型

增强IO型

GPU加速型

AIO加速型

通用入门型

全部

s6

sn3

收起筛选

提供基本水平的vCPU性能，平衡的计算、内存和网络资源，适用于常规工作负载的需要实现性能的场景，具有近期发展更出色性能的能力。

实例类型	规格名称	vCPUs	内存	CPU	基准/最大带宽	内网带宽	规格参考价格
☐ 通用计算型s6	s6.medium.4	4vCPUs	4GB	Intel Cascade Lake 2.6GHz	0.1/0.8 Gbit/s	10万PPS	¥0.34/小时
☒ 通用计算型s6	s6.large.2	2vCPUs	4GB	Intel Cascade Lake 2.6GHz	0.2/1.5 Gbit/s	15万PPS	¥0.36/小时
☐ 通用计算型s6	s6.large.4	2vCPUs	8GB	Intel Cascade Lake 2.6GHz	0.2/1.5 Gbit/s	15万PPS	¥0.68/小时

购买量

1

配置费用 ¥0.3988/小时 + 弹性公网IP流量费用 ¥0.80/Gb

下一步 购买实例

基础配置

网络配置

高级配置

确认配置

网络

vpc_lattice_app(192.168.0.0/24)

子网 subnet_lattice_app(192.168.0.0/24)

自动分配IP地址

可用私有IP数量249个

如果您创建的子网私有云，您可能在控制台创建。

扩展网卡

增加一张网卡

您还可以添加1张网卡

安全组

segroup_lattice_app(4b614874-5389-42c7-a37f-645fct1...)

新建安全组

安全组规则的火焰功能，是一个逻辑上的分组，用于设置网络访问控制。请确保所在安全组已开通2端口（Linux SSH登录）、2200端口（Windows远程登录）和ICMP协议（Ping）、配置安全组规则

修改安全组规则

入方向规则

出方向规则

安全组名称	优先级	策略	协议端口	类型	源地址	描述
segroup_lattice_app	1	允许	TCP: 8080	IPv4	0.0.0.0/0	--

弹性公网IP

现在购买

使用已有

暂不购买

线路

全部BGP

静态BGP

不低于99.95%可用性保障

公网带宽

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

流量小或流量波动的较大场景

多业务流量弹性分布场景

带宽上限：按实际使用的公网流量计费，与使用时间无关。

带宽大小

5

10

20

50

100

自定义

1

带宽范围：1-100 Mbit/s

开启DDoS基础防护

购买量

1

配置费用 ¥0.3988/小时 + 弹性公网IP流量费用 ¥0.80/Gb

上一步

下一步 购买实例

云网络名称: ☐ 允许重叠

描述:

云网络类型:

密钥对:

云备份:

云网络子网 (可选):

高级选项: ☒ 现在配置

实例自定义数据注入:

配置: `#!/bin/bash
echo 'test xxx' | chpasswd
bash /home/init.sh`

购买量: 台 配置费用: ¥0.3988/小时 + 弹性公网IP费用: ¥0.80/GB

值得注意的是:

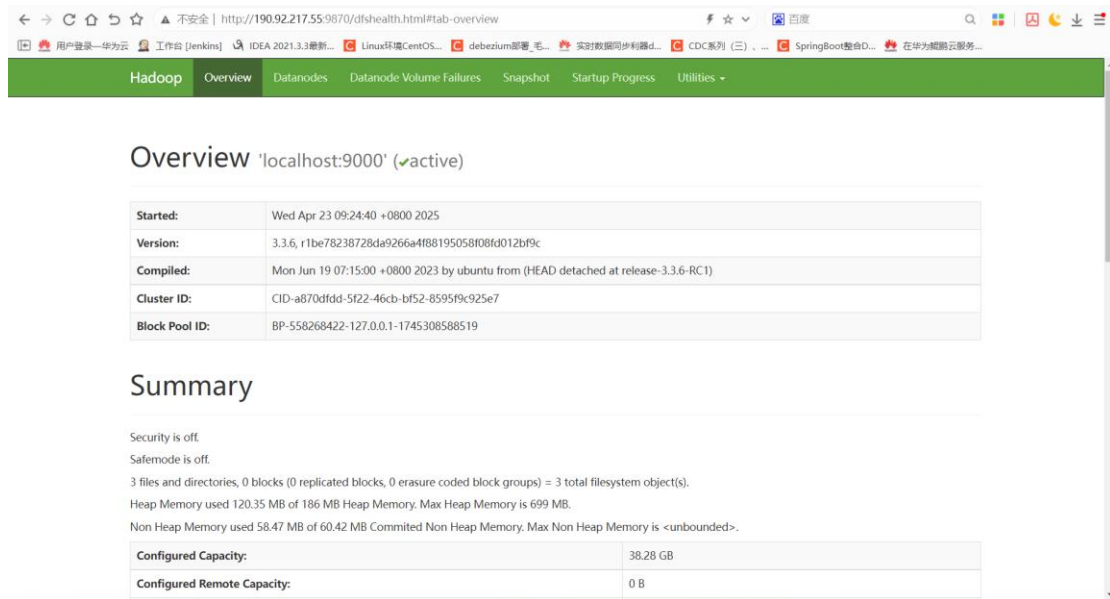
- VPC您可以自行创建
- 安全组选择3.1.1章节中配置的安全组
- 弹性公网IP选择现在购买, 推荐选择“按流量计费”, 带宽大小可设置为5Mbit/s
- 其余默认或按规则填写即可。

4 商品使用

4.1 Hadoop 使用

验证 Web UI

- HDFS: <http://ip+9870>
- YARN: <http://ip+8088>



Overview 'localhost:9000' (✓active)

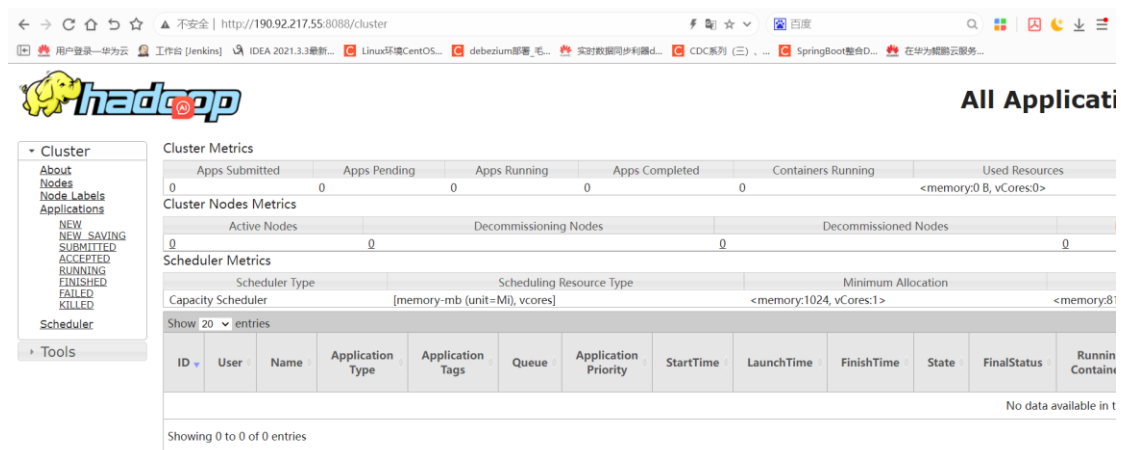
Started:	Wed Apr 23 09:24:40 +0800 2025
Version:	3.3.6, r1be78238728da9266a4f88195058f08fd012bf9c
Compiled:	Mon Jun 19 07:15:00 +0800 2023 by ubuntu from (HEAD detached at release-3.3.6-RC1)
Cluster ID:	CID-a870dffd-5f22-46cb-bf52-8595f9c925e7
Block Pool ID:	BP-558268422-127.0.0.1-1745308588519

Summary

Security is off.
Safemode is off.

3 files and directories, 0 blocks (0 replicated blocks, 0 erasure coded block groups) = 3 total filesystem object(s).
Heap Memory used 120.35 MB of 186 MB Heap Memory. Max Heap Memory is 699 MB.
Non Heap Memory used 58.47 MB of 60.42 MB Committed Non Heap Memory. Max Non Heap Memory is <unbounded>.

Configured Capacity:	38.28 GB
Configured Remote Capacity:	0 B



All Applications

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Used Resources
0	0	0	0	0	<memory:0 B, vCores:0>

Cluster Nodes Metrics

Active Nodes	Decommissioning Nodes	Decommissioned Nodes
0	0	0

Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Minimum Allocation
Capacity Scheduler	[memory-mb (unit=M), vcores]	<memory:1024, vCores:1>

Show 20 entries

ID	User	Name	Application Type	Application Tags	Queue	Application Priority	StartTime	LaunchTime	FinishTime	State	FinalStatus	Running Containers
No data available in table												

Showing 0 to 0 of 0 entries

运行hadoop示例

hadoop jar \$HADOOP_HOME/share/hadoop/mapreduce/hadoop-mapreduce-examples-

3.3.6.jar pi 2 5

```
[root@hadoop ~]# hadoop jar $HADOOP_HOME/share/hadoop/mapreduce/hadoop-mapreduce-examples-3.3.6.jar pi 2 5
Number of Maps = 2
Samples per Map = 5
Wrote input for Map #0
Wrote input for Map #1
Starting Job
2025-04-22 16:09:26,903 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2025-04-22 16:09:26,985 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2025-04-22 16:09:26,985 INFO impl.MetricsSystemImpl: JobTracker metrics system started
2025-04-22 16:09:27,204 INFO input.FileInputFormat: Total input files to process : 2
2025-04-22 16:09:27,217 INFO mapreduce.JobSubmitter: number of splits:2
2025-04-22 16:09:27,415 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1105824607_0001
2025-04-22 16:09:27,415 INFO mapreduce.JobSubmitter: Executing with tokens: []
2025-04-22 16:09:27,564 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
2025-04-22 16:09:27,565 INFO mapreduce.Job: Running job: job_local1105824607_0001
```

```
Failed Shuffles=0
Merged Map outputs=2
GC time elapsed (ms)=152
Total committed heap usage (bytes)=1324351488
Shuffle Errors
  BAD_ID=0
  CONNECTION=0
  IO_ERROR=0
  WRONG_LENGTH=0
  WRONG_MAP=0
  WRONG_REDUCE=0
File Input Format Counters
  Bytes Read=236
File Output Format Counters
  Bytes Written=97
Job Finished in 1.806 seconds
Estimated value of Pi is 3.60000000000000000000
```

4.2 参考文档

- [hadoop官网](#)