弹性 MapReduce

快速入门

产品文档

腾讯云

文档目录

快速入门
业务评估

最近更新时间：2023-12-27 09:58:07

选择集群类型

EMR 集群提供六种集群类型，可根据实际业务需要选择集群类型：
Hadoop集群：基于开源 Hadoop 及其周边生态组件，提供了5种应用场景：默认场景、Zookeeper、HBase、Presto、Kudu；满足海量数据存储、离线/实时数据分析、流式数据计算、机器学习等场景的大数据解决方案。
Druid 集群：高性能实时分析数据库，提供了大数据查询毫秒级延迟，支持多种数据摄入方式，适用于大数据实时查询场景。
ClickHouse 集群：列式数据库管理系统，适用于大宽表实时分析、实时 BI 报表分析、用户行为分析等高性能数仓分析业务场景。
Doris 集群：MPP分析型数据库产品，对于 PB 数量级、结构化数据可以做到亚秒级查询响应，使用上兼容 MySQL 协议，语法是标准的 SQL。适用于固定历史报表分析、实时数据分析、交互式数据分析等场景。
Kafka 集群：是一个分布式、分区的、多副本的、多订阅者，基于 Zookeeper 协调的消息处理系统，主要适用于异步处理，消息通讯以及流式数据接收和分发场景。
StarRocks 集群：采用了全面向量化技术，支持极速统一的 OLAP 分析数据库，适用多维分析，实时分析，高并发等场景等多种数据分析场景。

选择计费模式

EMR 集群提供的计费模式：
按量计费集群：集群的全部节点计费模式均为按量计费，适用于短时间存在或周期性存在的集群。
注意
在 CVM 控制台对 EMR 集群按量计费节点进行关机操作时，请谨慎选择关机模式，EMR 节点不支持关机不收费模式。

选择机型规格

EMR 提供了多种云服务器机型，包括 EMR 标准型、EMR 计算型、EMR 高 IO 型、EMR 内存型及 EMR 大数据型（若您需要黑石机型，请 提交工单 联系我们）。
您可以根据自身的业务需要及成本考量，进行机型的选择。
如您对离线计算的时延有一定的要求，我们建议您选择本地盘或大数据机型。

如您需要使用实时数据库 HBase，我们建议您选择 EMR 高 IO 型，并选择本地 SSD 盘，以实现最高的性能。

节点规格推荐

EMR 定义了5种节点类型，您可以根据集群类型进行选择：

集群类型	应用场景	节点类型	推荐规格
Hadoop	默认场景	Master	Master 节点：建议选择内存较大的实例规格，推荐内存大小至少8G。磁盘建议选择云盘可以让集群获得更高的稳定性。
		Core	若您的大部分数据在 COS 对象存储上，Core 节点与 Task 节点的功能则类似，大小不少于500G。Core 节点不具备弹性功能。 若您的架构未使用 COS 对象存储，则 Core 节点负责集群的计算与存储任务，EMR 默认开启三备份，在做数据盘大小预估时需考虑三备份空间，推荐使用大数据机型。
		Task	若您的架构未使用 COS 对象存储，则可以不使用 Task 节点。 若您的大部分数据在 COS 对象存储上，则 Task 节点可用作弹性计算资源，按需获取。
		Common	common 节点：节点主要做 zk 节点使用，最低选择2C4G 云盘100G 的规格可满足需求。
		Router	Router 节点：主要用于缓解主节点负载和用作任务提交机，因此建议选择较大内存的机型，最好不低于 Master 规格。
	Zookeeper	Common	common 节点：主要做 zk 节点使用，最低选择2C4G 云盘100G 的规格即可满足需求。
	HBase	Master	Master 节点：建议选择内存较大的实例规格，推荐内存大小至少8G。磁盘建议选择云盘可以让集群获得更高的稳定性。
		Core	若您的大部分数据在 COS 对象存储上，Core 节点与 Task 节点的功能则类似，大小不少于500G。 注意，Core 节点不具备弹性功能。 若您的架构未使用 COS 对象存储，则 Core 节点负责集群的计算与存储任务。
		Task	若您的架构未使用 COS 对象存储，则可以不使用 Task 节点。 若您的大部分数据在 COS 对象存储上，则 Task 节点可用作弹性计算资源，按需获取。
		Common	common 节点：主要做 zk 节点使用，最低选择2C4G 云盘100G 的规格即可满足需求。
		Router	Router 节点：主要用于缓解主节点负载和用作任务提交机，因此建议选择较大内存的机型，最好不低于 Master 规格。

	kudu	Master	Master 节点：建议选择内存较大的实例规格，推荐内存大小至少8G。磁盘建议选择云盘可以让集群获得更高的稳定性。
		Core	若您的大部分数据在 COS 对象存储上，Core 节点与 Task 节点的功能则类似，大小不少于500G。 注意：Core 节点不具备弹性功能。 若您的架构未使用 COS 对象存储，则 Core 节点负责集群的计算与存储任务，EMR 默认开启三备份，在做数据盘大小预估时需考虑三备份空间，推荐使用大数据机型。
		Task	若您的架构未使用COS对象存储，则可以不使用Task节点。 若您的大部分数据在 COS 对象存储上，则Task节点可用作弹性计算资源，按需获取。
		Common	common 节点：主要做 zk 节点使用，最低选择2C4G 云盘100G 的规格即可满足需求。
		Router	Router 节点：主要用于缓解主节点负载和用作任务提交机，因此建议选择较大内存的机型，最好不低于 Master 规格。
	presto	Master	Master 节点：建议选择内存较大的实例规格，推荐内存大小至少8G。磁盘建议选择云盘可以让集群获得更高的稳定性。
		Core	若您的大部分数据在 COS 对象存储上，Core 节点与 Task 节点的功能则类似，大小不少于500G。 注意：Core 节点不具备弹性功能。 若您的架构未使用 COS 对象存储，则 Core 节点负责集群的计算与存储任务，EMR 默认开启三备份，在做数据盘大小预估时需考虑三备份空间，推荐使用大数据机型。
		Task	若您的架构未使用COS对象存储，则可以不使用Task节点。 若您的大部分数据在 COS 对象存储上，则Task节点可用作弹性计算资源，按需获取。
		Common	common 节点：主要做 zk 节点使用，最低选择2C4G 云盘100G 的规格即可满足需求。
		Router	Router 节点：主要用于缓解主节点负载和用作任务提交机，因此建议选择较大内存的机型，最好不低于 Master 规格。
ClickHouse	默认场景	Core	Core 节点：建议选择 CPU 和内存较高的机型，由于本地磁盘遇到坏盘情况存在数据丢失风险，磁盘建议选择云硬盘。
		Common	common 节点：建议 CPU 和内存最小配置不低于4C16G。
Kafka	默认场景	Core	Core 节点：建议选择 CPU 和内存较高的机型，由于本地磁盘遇到坏盘情况存在数据丢失风险，磁盘建议选择云硬盘。

		Common	common 节点：建议 CPU 和内存最小配置不低于4C16G。
Doris	默认场景	Master	Master 节点：建议选择内存较大的实例规格，推荐内存大小至少8G，Master 节点上元数据全部存储在内存中。
		Core	Core 节点：建议选择内存较大的实例规格，推荐内存大小至少8G。磁盘推荐使用云 SSD 盘以获得更好的 IO 性能及稳定性。
		Router	Router 节点：部署 Frontend 模块，实现读写高可用，因此建议选择较大内存的机型，最好不低于 Master 规格。
Druid	默认场景	Master	Master 节点：建议选择内存较大的实例规格，推荐内存不低于16G。磁盘推荐使用 SSD 盘，可以获得更好的 IO 性能。
		Core	Core 节点：建议选择内存较大的实例规格，推荐内存大小至少8G。磁盘推荐使用云 SSD 盘以获得更好的 IO 性能及稳定性。
		Task	若您的架构未使用 COS 对象存储，则可以不使用 Task 节点。 若您的大部分数据在 COS 对象存储上，则 Task 节点可用作弹性计算资源，按需获取。
		Common	common 节点：主要做 zk 节点使用，建议选择2C4G云盘100G的规格即可满足需求。
		Router	Router节点：主要用于缓解主节点负载和用作任务提交机，因此建议选择较大内存的机型，最好不低于 Master 规格。
StarRocks	默认场景	Master	Master 节点：建议选择内存较大的实例规格，推荐内存大小至少8G，Master 节点上元数据全部存储在内存中。
		Core	Core 节点：建议选择内存较大的实例规格，推荐内存大小至少8G。磁盘推荐使用云 SSD 盘以获得更好的 IO 性能及稳定性。
		Router	Router 节点：部署 Frontend 模块，实现读写高可用，因此建议选择较大内存的机型，最好不低于 Master 规格。

注意

不同集群类型对节点规格要求不同，目前系统将默认推荐满足集群要求的配置，您可以根据业务需求调整机型规格，推荐机型仅作参考。

Core 节点不具备弹性功能。若您的架构未使用 COS 对象存储，则 Core 节点负责集群的计算与存储任务，EMR 默认开启三备份，在做数据盘大小预估时需考虑三备份空间，推荐使用大数据机型。

网络及安全

为保证集群的网络安全，EMR 集群将会被放置在一个 VPC 中，我们会给该 VPC 增加一个安全组策略。同时为了保证 Hadoop 生态组件的 WebUI 能够便捷访问，我们为其中一个 Master 节点开启了外网 IP，按照流量计费的模式；Router 节点默认不开通外网 IP，如需开通，可以在 CVM 控制台自由绑定弹性公网 IP。

注意

Master 节点在创建集群时默认开启外网 IP，但用户可根据情况选择不开启外网 IP。

开启集群 Master 节点公网，主要用于 ssh 登录和组件 WebUI 查看。

主节点 Master 节点会开启外网，按流量付费，带宽上限为5M。创建集群后，您可在控制台对该网络进行调整。

创建 EMR 集群

最近更新时间：2023-12-27 09:58:25

操作场景

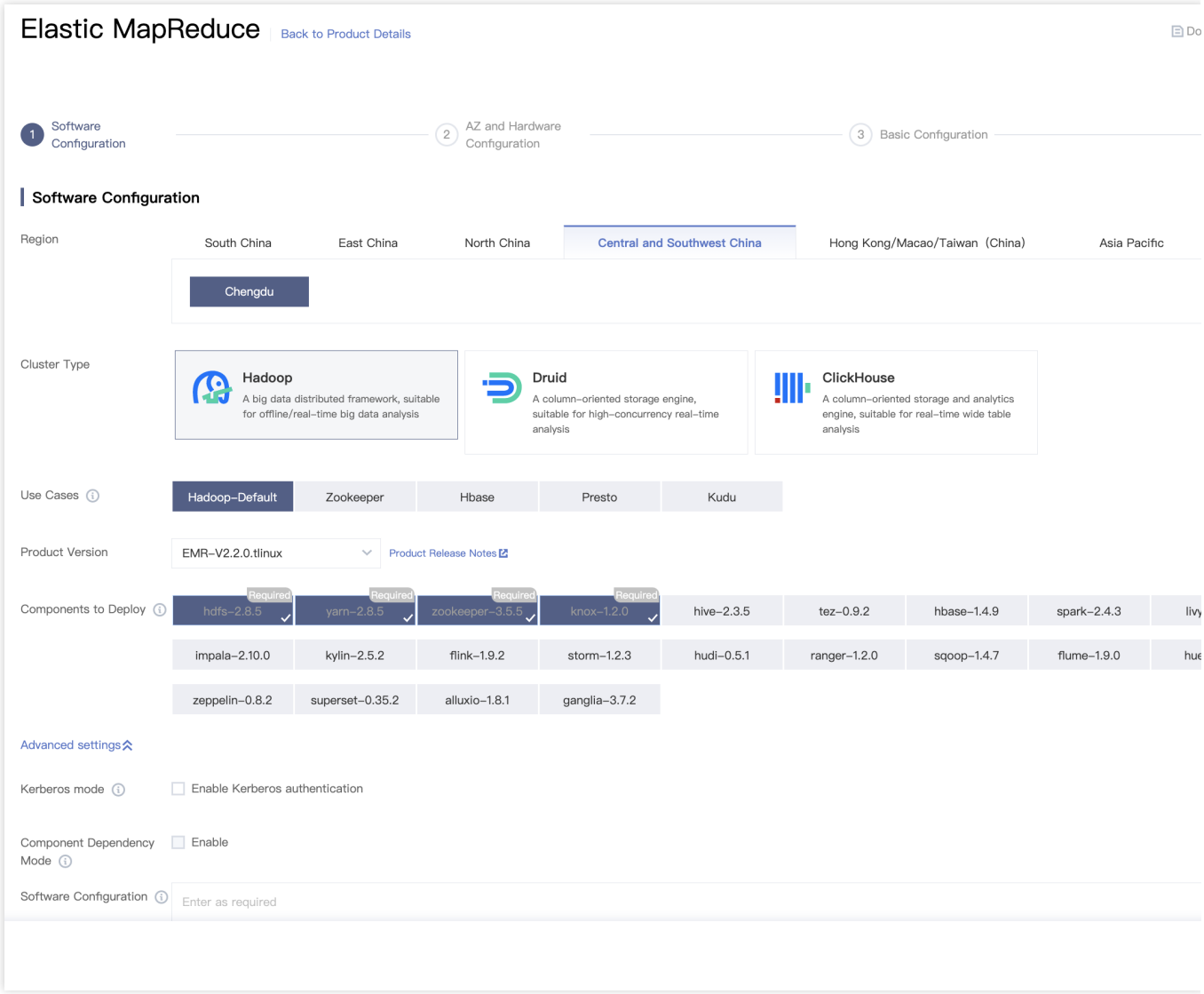
本文为您介绍通过 EMR 控制台创建一个 EMR 集群的操作。

操作步骤

登录 EMR 控制台，在集群列表页单击**创建集群**。

1. 软件配置

地域：地域（Region）是指物理的数据中心的地理区域，支持地域有：广州、上海、北京、新加坡、硅谷、成都、南京、孟买等；不同地域的云产品之间内网不互通。
集群类型：EMR 目前支持六种集群部署方式，分别为 Hadoop 集群、ClickHouse 集群、Druid 集群、Doris 集群、Kafka 集群、StarRocks 集群；需根据实际业务需要选择集群类型进行部署。
应用场景：基于 Hadoop 集群类型支持五种应用场景，分别为：默认场景、zookeeper、HBase、Presto、Kudu；根据实际业务需要选择相应的应用场景进行部署。
产品版本和部署组件：EMR 推荐了一些常用的 Hadoop 组件搭配，您也可以根据自身需求组合各组件。
Kerberos 安全集群：是否开启集群的 Kerberos 认证功能，一般的个人用户集群无需该功能，默认关闭。
软件配置：按照要求填写参数可实现自定义软件参数创建集群，同时兼容访问外部集群功能，在参数中正确配置访问地址信息即可读写外部集群的数据。



2. 区域与硬件配置

计费模式：支持按量计费模式。

按量计费：按照使用时长付费，需对账户进行实名认证，在开通时需冻结2小时的费用（代金券不可用作冻结凭证），销毁时退还冻结资源费用。

可用区：同一地域下不同可用区支持机型规格不同，建议选择最新可用区；处在不同地域的云产品内网不通，购买后不能更换。建议选择靠近业务数据的地域可用区，以降低访问延迟、提高下载速度。

集群网络：为保证EMR集群的安全性，我们会将集群各节点放入一个私有网络中，您需要设置一个私有网络以保证EMR 集群的正确创建。

安全组：安全组具有防火墙功能，用于设置云服务器 CVM 的网络访问控制。如果没有安全组，EMR 会自动帮您新建一个安全组。若已经有在使用的安全组可以直接选择使用。若安全组数量已达到上限无法新建，可删除部分不再使用的安全组。查看已在使用的安全组。

创建安全组：EMR 帮助用户创建一个安全组，开启22和30001端口及必要的内网通信网段。

已有EMR安全组：选择已创建的 EMR 安全组作为当前实例的安全组，开启22和30001端口及必要的内网通信网段。

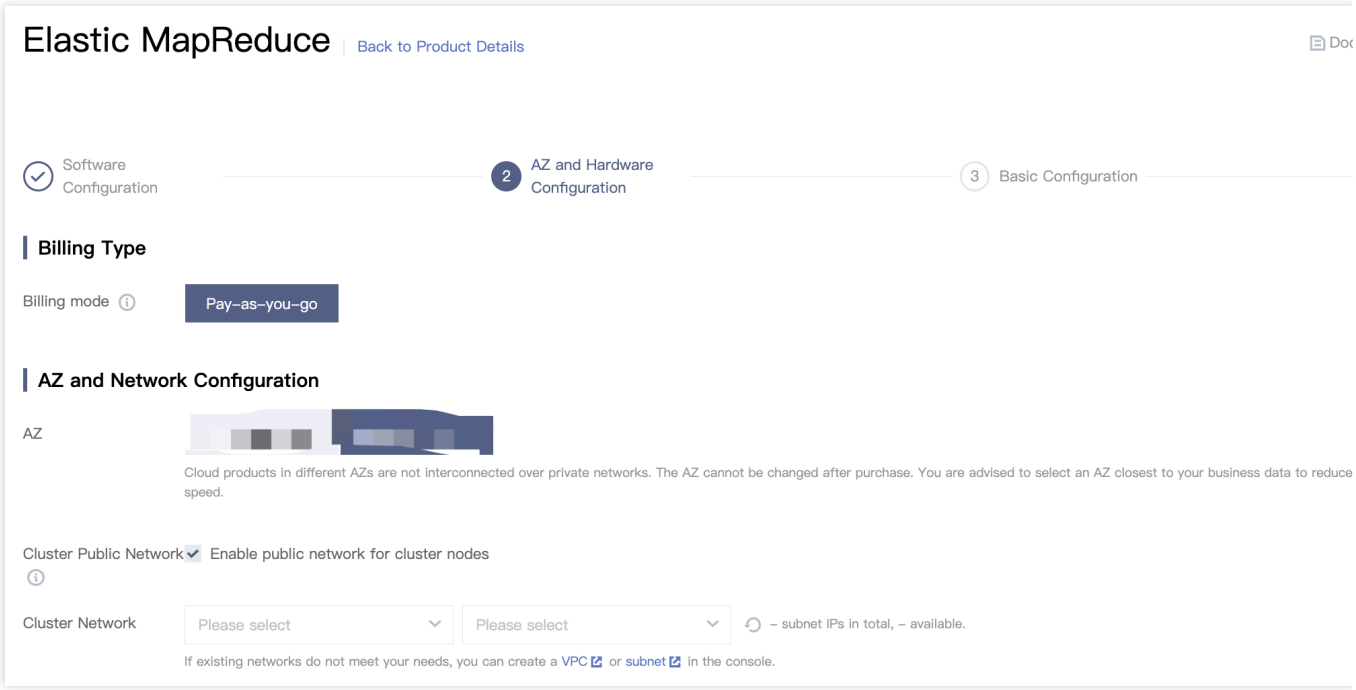
远程登录：22端口常用于远程登录，新建安全组将默认开启，您可以根据业务需要关闭该端口。

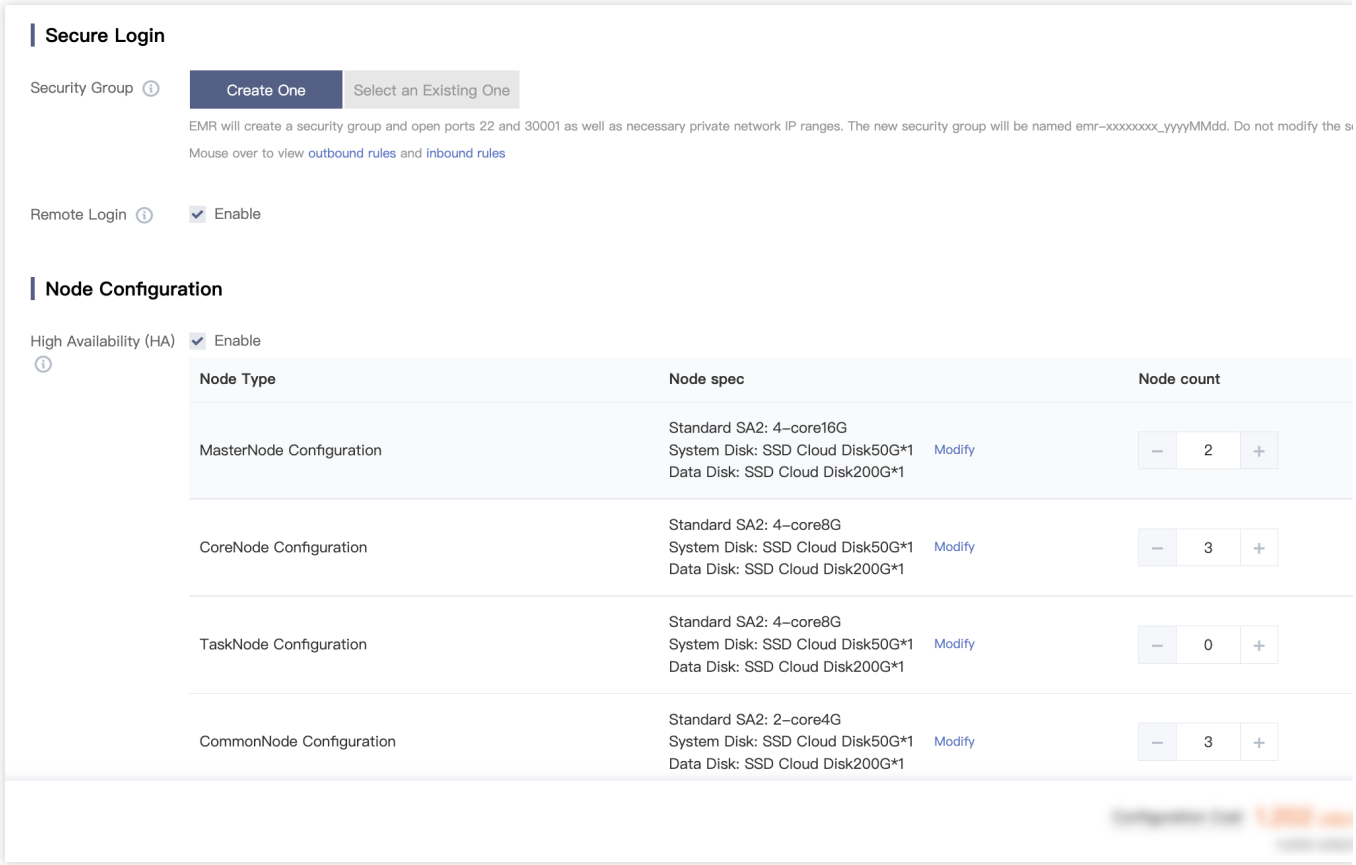
高可用（HA）：默认启动高可用，不同集群类型和应用场景在 HA 或非 HA 场景下，不同节点类型部署数量不同，详见 集群类型。

节点配置：EMR提供了多种节点类型，可以根据业务需要为不同节点类型选择合适机型配置。

说明

目前支持 Core 节点、Task 节点和 Router 节点挂载多种云盘类型（每种云盘类型最多只能选择1次）和多块云盘（最多15块）。





置放群组：放置群组是云服务器实例在底层硬件上分布放置的策略，可参考 置放群组。

Hive 元数据库：如果选择了 Hive 组件，Hive 元数据库提供了两种存储方式：第一种集群默认，Hive 元数据存储于集群独立购买的MetaDB；第二种是关联外部 Hive 元数据库，可选择关联 EMR-MetaDB 或自建 MySQL 数据库，元数据将存储于关联的数据库中，不随集群销毁而销毁。

3. 基础配置

所属项目：将当前集群分配给不同的项目组，如果您要将实例分配至新的项目，请先新建项目。具体操作请参考新建项目

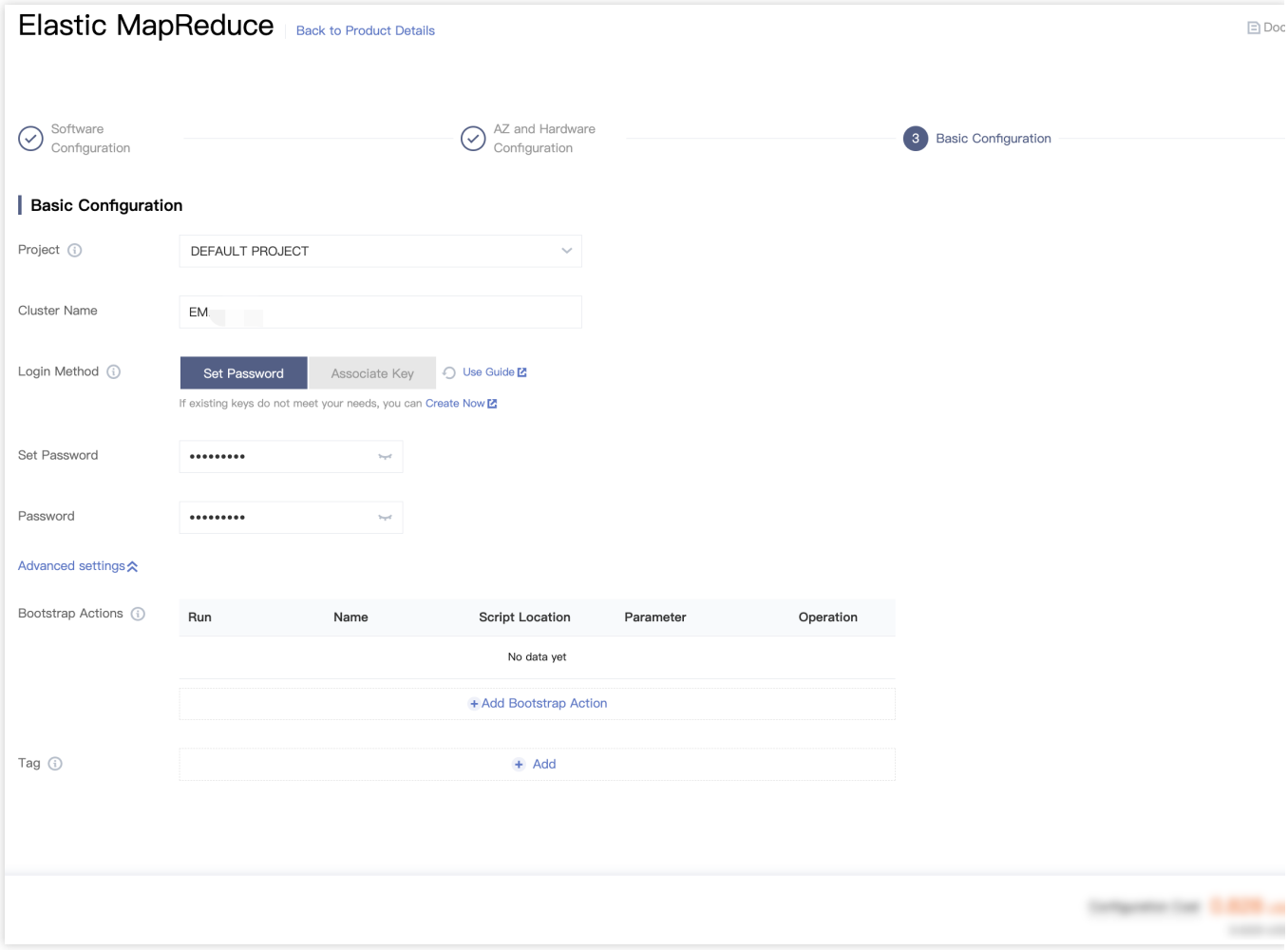
集群名称：通过设置集群名称，来区分不同的EMR集群。

登录方式：目前 EMR 提供两种登录集群服务、节点、MetaDB 的方式，自定义设置密码方式和关联密钥方式；SSH 密钥仅用于 EMR-UI 快捷入口登录。其中，用户名默认为“root”，superset 组件 webUI 快捷入口的用户名为“admin”。

高级设置：

引导操作：引导脚本操作方便您在创建集群的过程中执行自定义脚本，以便您修改集群环境、安装第三方软件和使用自有数据。

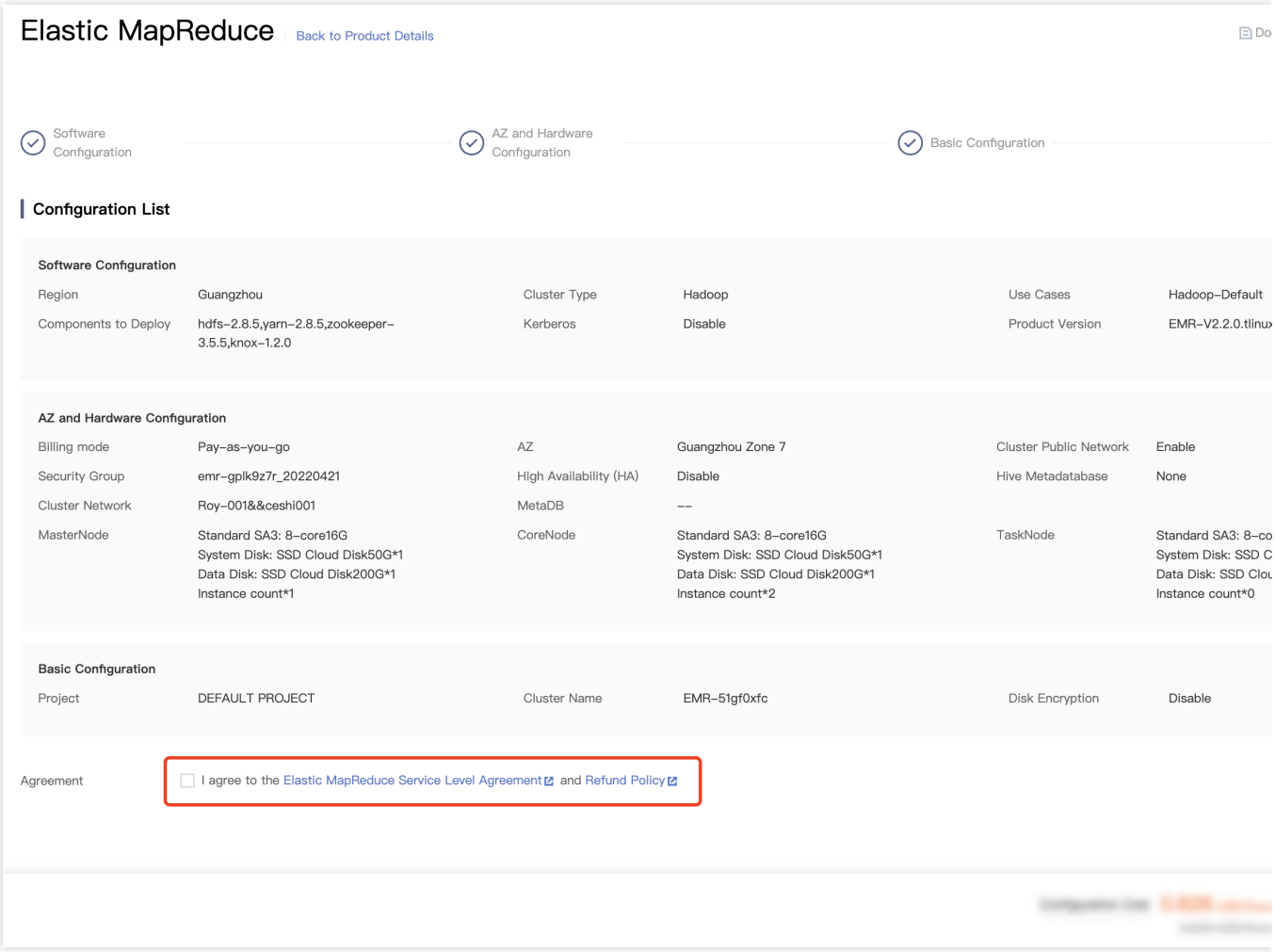
标签：您在创建时对集群或节点资源添加标签，以便于管理集群和节点资源，最多可绑定5条，标签键不可重复。



4. 确认配置信息

自动续费：系统将在集群到期前7天，每天检测用户账户上的可用余额是否充足，设置为自动续费的集群资源进行续费。

完成以上配置后，单击购买进行支付，支付成功后EMR集群进入创建过程，大约10分钟后即可在EMR控制台中找到新建的集群。



注意

您可以在 CVM 控制台中查看各节点的实例信息，为保证 EMR 集群的正常运行，请不要在 CVM 控制台中更改这些实例的配置信息。

后续步骤

集群创建成功后，您可根据自身情况登录集群后，对集群进行进一步的配置等操作，具体操作可参考如下文档：

登录集群

配置集群：软件配置、挂载 CHDFS、统一 HIVE 元数据

管理集群：设置标签、设置引导操作、集群销毁

登录集群

最近更新时间：2023-12-27 09:58:42

远程登录软件登录（本地系统为 Windows）

以 Xshell 为例，介绍本地为 Windows 系统的电脑如何使用远程登录软件通过密码登录 EMR 集群。

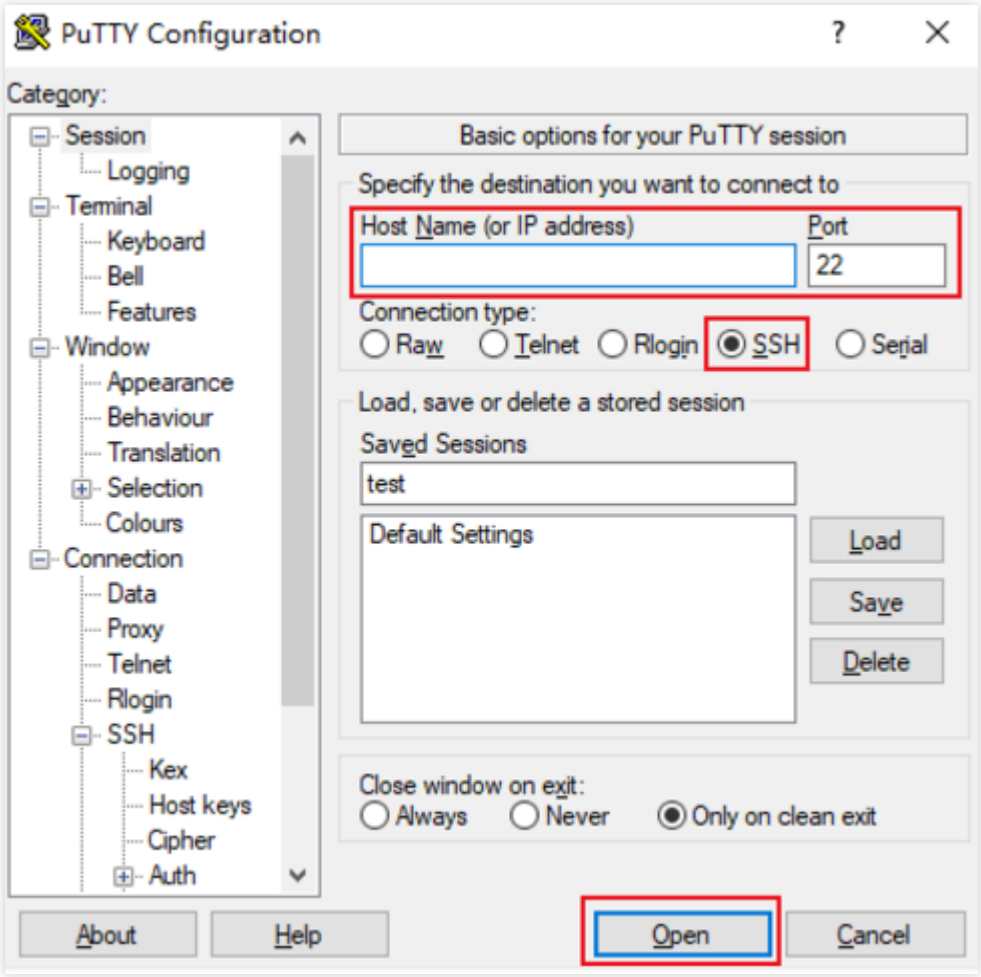
适用本地操作系统

Windows

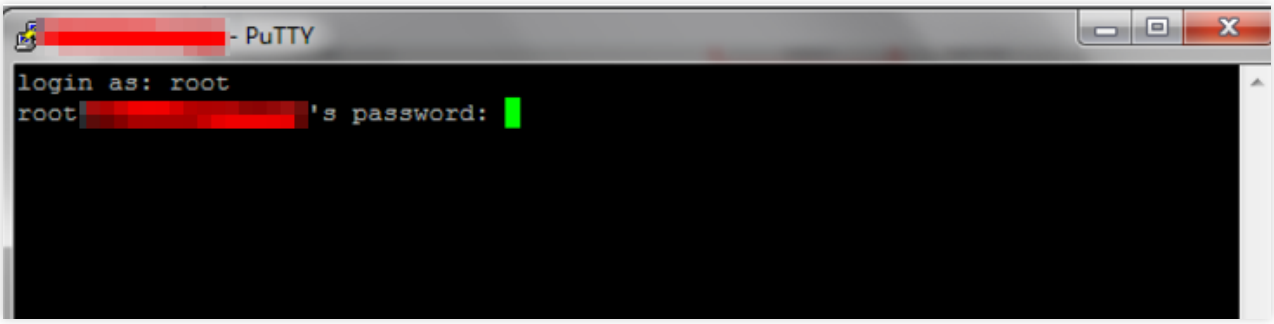
使用密码登录

1. 安装 Windows 远程登录软件，即 PuTTY。PuTTY 的获取方式参见 获取链接。
2. 打开 PuTTY 客户端，在 PuTTY Configuration 窗口中输入以下内容，并单击**Open**，创建一个新对话。如下图所示：

Host Name：EMR 集群的公网 IP，登录 EMR 控制台，可在列表页及详情页中获取集群公网 IP。
Port：云服务器的端口，必须填22。请确保云服务器22端口已开放，详情可参见 安全组 及 网络 ACL
Connection type：选择**SSH**。



3. 在 PuTTY 会话窗口中，输入已获取的管理员帐号，按 Enter 键。
4. 输入已获取的登录密码，按 Enter 键，即可完成登录。如下图所示：



使用 SSH 登录（本地系统为 Linux/Mac OS）

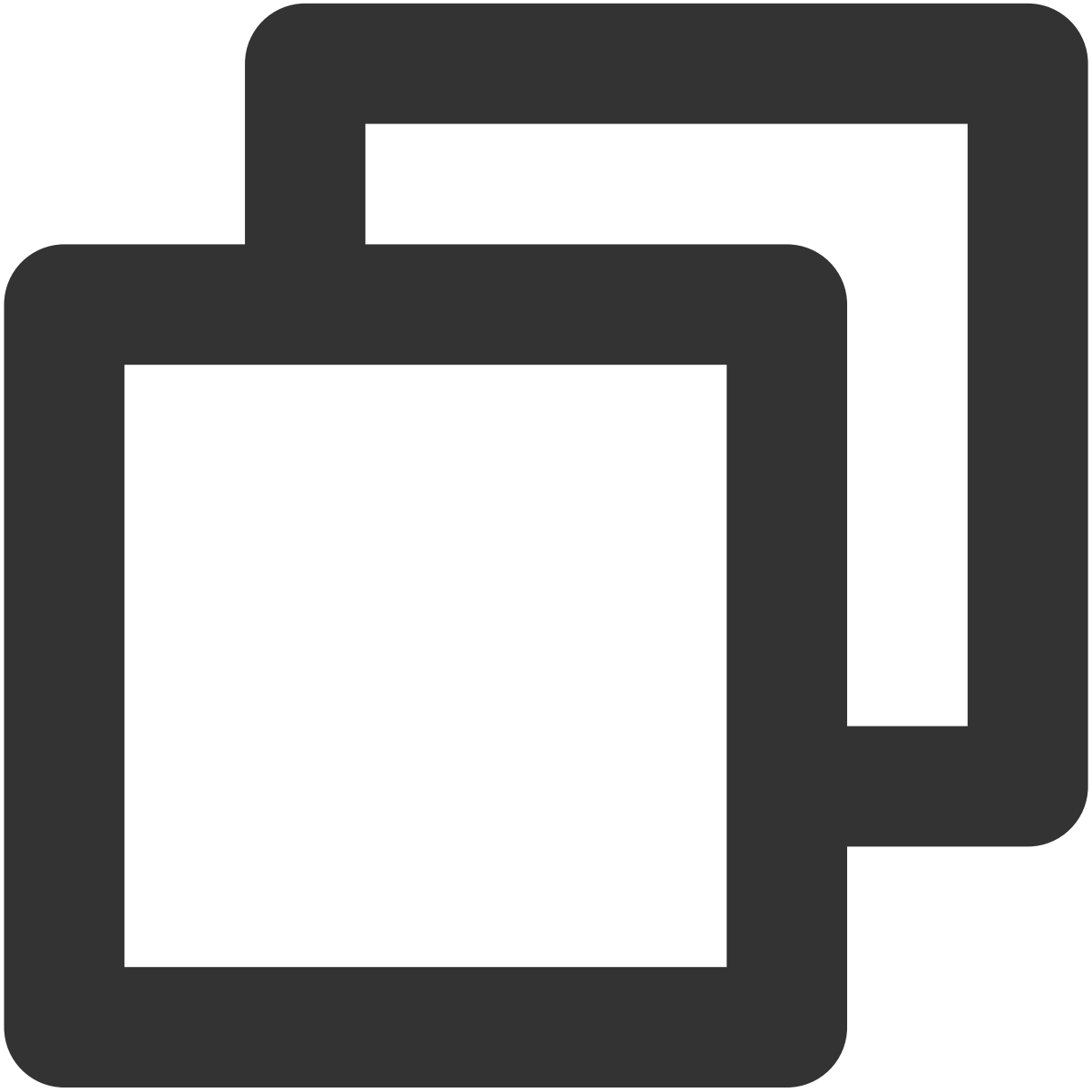
介绍本地为 Linux/Mac OS 系统的电脑通过 SSH 登录 EMR 集群。

适用本地操作系统

Linux 或 Mac OS

使用密码登录

1. Mac OS 用户请打开系统自带的终端（Terminal）并执行以下命令，Linux 用户请直接执行以下命令：



```
ssh <username>@<hostname or IP address>
```

username：即管理员帐号，例如 root。

hostname or IP address：为您的 EMR 实例公网 IP 或自定义域名。

2. 输入已获取的密码（此时仅有输入没有显示输出），按 Enter 键，即可完成登录。