

MapReduce 服务

# 快速入门

文档版本 01  
发布日期 2022-03-01



**版权所有 © 华为技术有限公司 2022。保留一切权利。**

非经本公司书面许可，任何单位和个人不得擅自摘抄、复制本文档内容的部分或全部，并不得以任何形式传播。

## **商标声明**



HUAWEI和其他华为商标均为华为技术有限公司的商标。

本文档提及的其他所有商标或注册商标，由各自的所有人拥有。

## **注意**

您购买的产品、服务或特性等应受华为公司商业合同和条款的约束，本文档中描述的全部或部分产品、服务或特性可能不在您的购买或使用范围之内。除非合同另有约定，华为公司对本文档内容不做任何明示或暗示的声明或保证。

由于产品版本升级或其他原因，本文档内容会不定期进行更新。除非另有约定，本文档仅作为使用指导，本文档中的所有陈述、信息和建议不构成任何明示或暗示的担保。

## 目录

<b>1 购买并使用 MRS 集群</b>	<b>1</b>
1.1 如何快速使用 MRS	1
1.2 创建集群	1
1.3 上传数据	4
1.4 添加作业	7
1.5 删除集群	9
<b>2 安装并使用集群客户端</b>	<b>11</b>
<b>3 快速使用 Kerberos 认证集群</b>	<b>15</b>
<b>4 从零开始使用 Hadoop</b>	<b>24</b>
<b>5 从零开始使用 Kafka</b>	<b>25</b>
<b>6 从零开始使用 HBase</b>	<b>26</b>
<b>7 修改 MRS 服务配置参数</b>	<b>33</b>
<b>8 配置 MRS 集群弹性伸缩</b>	<b>37</b>
<b>9 配置 Hive 存算分离</b>	<b>47</b>
<b>10 提交 Spark 任务到新增 Task 节点</b>	<b>52</b>

# 1 购买并使用 MRS 集群

## 1.1 如何快速使用 MRS

MRS是一个在华为云上部署和管理Hadoop系统的服务，一键即可部署Hadoop集群。MRS提供租户完全可控的企业级大数据集群云服务，轻松运行Hadoop、Spark、HBase、Kafka等大数据组件。

MRS使用简单，通过使用在集群中连接在一起的多台计算机，您可以运行各种任务，处理或者存储（PB级）巨量数据。

MRS的基本使用流程如下：

1. 通过MRS管理控制台购买集群，用户可以指定集群类型用于离线数据分析和流处理任务，也可以指定集群中预置的弹性云服务器实例规格、实例数量、数据盘类型（普通IO、高IO、超高IO）、要安装的组件等。
2. 开发数据处理程序，MRS的[开发指南](#)为用户提供代码示例和教程，帮助您快速开始开发自己的程序并正常运行。
3. 上传准备好的程序和数据文件到对象存储服务（OBS）或者集群内的HDFS文件系统中。
4. 集群创建成功后，可直接添加作业，执行由用户自身开发的程序或者SQL语句，进行数据的处理与分析。
5. MRS为用户提供企业级的大数据集群的统一管理平台，帮助用户快速掌握服务及主机的健康状态，通过图形化的指标监控及定制及时的获取系统的关键信息，根据实际业务的性能需求修改服务属性的配置，对集群、服务、角色实例等实现一键启停等操作。
6. 如果作业执行结束后不再需要集群，可以快速删除MRS集群。集群删除后不再产生费用。

## 1.2 创建集群

使用MRS的首要操作就是购买一个集群，本章节为您介绍如何在MRS管理控制台快速创建一个新的集群。

## 操作步骤

**步骤1** 登录MRS管理控制台。

**步骤2** 单击“购买集群”，进入“购买集群”页面。

### 说明

创建集群时需要注意配额提醒。当资源配额不足时，建议按照提示申请足够的资源，再创建集群。

**步骤3** 在购买集群页面，选择“自定义购买”页签。

**步骤4** 配置集群软件信息。

- 区域：默认即可。
- 集群名称：可以设置为系统默认名称，但为了区分和记忆，建议带上项目拼音缩写或者日期等。例如：“mrs\_20180321”。
- 集群版本：默认最新版本即可。
- 集群类型：默认选择“分析集群”即可。
- 组件选择：分析集群勾选Spark2x、HBase和Hive等组件。流式集群勾选Kafka和Storm等组件。混合集群可同时勾选分析集群流式集群的组件。
- 元数据：默认即可。

区域

不同区域的资源之间内网不互通。请选择靠近您客户的区域。可以降低网络延时、提高访问速度。[如何选择区域](#)

集群名称

mrs\_1v4V

?

集群版本

MRS 3.1.0

集群类型

分析集群


流式集群

混合集群

自定义

必选组件默认勾选，被依赖的组件会被自动勾选。

分析组件

	组件名	版本	描述
<input checked="" type="checkbox"/>	Hadoop	3.1.1	针对大数据集的分布式数据处理框架。
<input checked="" type="checkbox"/>	Spark2x	2.4.5	Spark2x是一个对大规模数据处理的快速通用引擎,基于开源Spark2.x版本开发。
<input checked="" type="checkbox"/>	HBase	2.2.3	HBase是一个高可靠性、高性能、面向列、可伸缩的分布式存储系统。
<input checked="" type="checkbox"/>	Hive	3.1.0	方便查询、管理存储在分布式存储系统上的大数据集的数据仓库软件。
<input type="checkbox"/>	Hue	4.7.0	Apache Hadoop的UI界面。
<input type="checkbox"/>	Flink	1.12.0	Flink是一个开源的批流合一的消息处理系统。
<input type="checkbox"/>	Oozie	5.1.0	Hadoop作业调度系统。
<input checked="" type="checkbox"/>	ZooKeeper	3.5.6	Zookeeper是一个集中的服务，它用于维护配置信息、命名、提供分布式的同步和提供
<input checked="" type="checkbox"/>	Ranger	2.0.0	RANGER是一个基于Hadoop平台，旨在为复杂数据提供监控、操作、管理等权限管控接口的集中式框架。
<input type="checkbox"/>	Tez	0.9.2	一个支持有向无环图的分布式计算框架。
<input type="checkbox"/>	Impala	3.4.0	一种处理大量数据的SQL查询引擎。
<input type="checkbox"/>	Presto	333	一种开源、分布式SQL查询引擎。
<input type="checkbox"/>	Kudu	1.12.1	一个能对快速数据进行快速分析的Hadoop存储引擎
<input type="checkbox"/>	Sqoop	1.4.7	Sqoop是专为Apache Hadoop和结构化数据库（如关系型数据库）设计的高效传输大量数据的工具。

元数据

本地元数据

数据连接

?

## 说明

针对MRS 3.x之前版本，分析集群勾选Spark、HBase和Hive等组件。

### 步骤5 单击“下一步”。

- 计费模式：默认即可。
- 可用区：默认即可。
- 虚拟私有云：默认即可。如果没有虚拟私有云，请单击“查看虚拟私有云”进入虚拟私有云，创建一个新的虚拟私有云。
- 子网：默认即可。
- 安全组：选择“自动创建”。
- 弹性公网IP：选择“暂不绑定”。
- 企业项目：默认即可。
- 实例规格：Master和Core节点都选择“通用计算型S3->8核16GB(s3.2xlarge.2)”。
- 系统盘：存储类型选择“普通IO”，存储空间默认即可。
- 数据盘：存储类型选择“普通IO”，存储空间默认即可，数据盘数量默认即可。
- 实例数量：Master节点数量默认为2，Core节点数量配置为3。

The screenshot displays the MRS cluster configuration interface. At the top, there are tabs for '包年/包月' (Pay-as-you-go) and '按需计费' (Pay-per-use). Below this, a '购买时长' (Purchase duration) section shows options from 1 to 12 months. The main configuration area includes dropdowns for '可用区' (Availability Zone), '虚拟私有云' (VPC), '子网' (Subnet), '安全组' (Security Group), '弹性公网IP' (Elastic Public IP), and '企业项目' (Enterprise Project). A table at the bottom lists the node configurations:

集群节点	节点类型	计费模式	实例规格	实例数量
Master节点	通用计算型S3	包年/包月	16 vCPUs   64 GB   c6.4xlarge.4 系统盘 通用型SSD 480 GB x 1 数据盘 通用型SSD 600 GB x 1	2
分析Core节点	通用计算型S3	包年/包月	16 vCPUs   64 GB   c6.4xlarge.4 系统盘 通用型SSD 480 GB x 1 数据盘 通用型SSD 600 GB x 1	3
分析Task节点		按需计费		

### 步骤6 单击“下一步”进入高级配置页签，配置参数，其他参数保持默认。

- Kerberos认证：
  - Kerberos认证：关闭Kerberos认证。
  - 用户名：Manager管理员用户，目前默认为admin用户。
  - 密码：Manager管理员用户的密码。
- 登录方式：选择登录ECS节点的登录方式。
  - 密码：设置登录ECS节点的登录密码。

- 密钥对：从下拉框中选择密钥对，如果已获取私钥文件，请勾选“我确认已获取该密钥对中的私钥文件 *SSHkey-xxx*，否则无法登录弹性云服务器”。如果没有创建密钥对，请单击“查看密钥对”创建或导入密钥，然后再获取私钥文件。
- 通信安全授权：勾选确认授权。
- 服务协议：勾选“我已阅读并同意《华为MRS服务声明》”。

**步骤7** 单击“立即购买”。

当集群开启Kerberos认证时，需要确认是否需要开启Kerberos认证，若确认开启请单击“继续”，若无需开启Kerberos认证请单击“返回”关闭Kerberos认证后再创建集群。

**步骤8** 单击“返回集群列表”，可以查看到集群创建的状态。

集群创建需要时间，所创集群的初始状态为“启动中”，创建成功后状态更新为“运行中”，请您耐心等待。

----结束

### 1.3 上传数据

集群创建成功后，用户通过“文件管理”页面可以在分析集群进行HDFS目录的创建、删除，文件的导入、导出、删除等操作。

若集群为开启Kerberos认证的安全集群，请在使用“文件管理”前，先完成IAM用户同步（在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“同步”进行IAM用户同步）。

### 背景信息

MRS集群处理的数据源通常来源于OBS或HDFS，OBS为客户提供海量、安全、高可靠、低成本的数据存储能力。MRS可以直接处理OBS中的数据，客户可以基于管理控制台Web界面和OBS客户端对数据进行浏览、管理和使用。如果需要将OBS的数据导入集群的HDFS系统后进行处理，可参考以下步骤进行操作。

### 导入数据

MRS目前支持将OBS上的数据导入至HDFS中。上传文件速率会随着文件大小的增大而变慢，适合数据量小的场景下使用。

支持导入文件和目录，操作方法如下：

1. 登录MRS管理控制台。
2. 选择“集群列表 > 现有集群”，选中一集群并单击集群名进入集群信息页面。
3. 单击“文件管理”，进入“文件管理”页面。
4. 选择“HDFS文件列表”。

HDFS文件列表

文件操作记录

请从左侧窗口的HDFS中查看统计信息。

/ user /

新增

导入数据

导出数据

C

文件名	大小	时间	操作
-	-	-	-
hive	-	2021/09/15 10:42:03 GMT+08:00	删除
loader	-	2021/09/15 10:42:03 GMT+08:00	删除
magrad	-	2021/09/15 10:45:21 GMT+08:00	删除
oozie	-	2021/09/15 10:42:03 GMT+08:00	删除
yarn	-	2021/09/15 10:42:03 GMT+08:00	删除

5. 进入数据存储目录，如“bd\_app1”。  
“bd\_app1”目录仅为示例，可以是界面上的任何目录，也可以通过“新建”创建新的文件夹。  
新建文件夹时需要满足以下要求：
  - 文件夹名称小于等于255字符。
  - 不允许为空。
  - 不能包含：/\*?"<>|\\;&,'!{}[]\$%+特殊字符。
  - 不能以“.”开头或结尾。
  - 开头和末尾的空格会被忽略。
6. 单击“导入数据”，正确配置HDFS和OBS路径。配置OBS或者HDFS路径时，单击“浏览”并选择文件目录，然后单击“是”。

图 1-1 导入数据

### 从OBS导入数据至HDFS

OBS路径 

HDFS路径

- OBS路径
  - 必须以“obs://”开头。
  - 不支持导入KMS加密的文件或程序。
  - 不支持导入空的文件夹。
  - 目录和文件名称可以包含中文、字母、数字、中划线和下划线，但不能包含|&>,<'\$\*?\\特殊字符。
  - 目录和文件名称不能以空格开头或结尾，中间可以包含空格。
  - OBS全路径长度小于等于255字符。
- HDFS路径
  - 默认以“/user”开头。
  - 目录和文件名称可以包含中文、字母、数字、中划线和下划线，但不能包含|&>,<'\$\*?\\特殊字符。



- 目录和文件名称不能以空格开头或结尾，中间可以包含空格。
  - HDFS全路径长度小于等于255字符。
7. 单击“确定”。
- 文件上传进度可在“文件操作记录”中查看。MRS将数据导入操作当做Distcp作业处理，也可在“作业管理”中查看Distcp作业是否执行成功。

## 导出数据

数据完成处理和分析后，您可以将数据存储存储在HDFS中，也可以将集群中的数据导出至OBS系统。

支持导出文件和目录，操作方法如下：

1. 登录MRS管理控制台。
2. 选择“集群列表 > 现有集群”，选中一集群并单击集群名进入集群基本信息页面。
3. 单击“文件管理”，进入“文件管理”页面。
4. 选择“HDFS文件列表”。
5. 进入数据存储目录，如“bd\_app1”。
6. 单击“导出数据”，配置OBS和HDFS路径。配置OBS或者HDFS路径时，单击“浏览”并选择文件目录，然后单击“是”。

图 1-2 导出数据

### 从HDFS导出数据至OBS

HDFS路径  浏览

OBS路径  浏览

确定 取消

- OBS路径
  - 必须以“obs://”开头。
  - 目录和文件名称可以包含中文、字母、数字、中划线和下划线，但不能包含|&>,<'\$\*?\特殊字符。

- 目录和文件名称不能以空格开头或结尾，中间可以包含空格。
- OBS全路径长度小于等于255字符。
- HDFS路径
  - 默认以“/user”开头。
  - 目录和文件名称可以包含中文、字母、数字、中划线和下划线，但不能包含;|&>,<'\$\*?\\:特殊字符。
  - 目录和文件名称不能以空格开头或结尾，中间可以包含空格。
  - HDFS全路径长度小于等于255字符。

#### 说明

当导出文件夹到OBS系统时，在OBS路径下，将增加一个标签文件，文件命名为“folder name\_\$folder\$”。请确保导出的文件夹为非空文件夹，如果导出的文件夹为空文件夹，OBS无法显示该文件夹，仅生成一个命名为“folder name\_\$folder\$”的文件。

#### 7. 单击“确定”。

文件上传进度可在“文件操作记录”中查看。MRS将数据导出操作当做Distcp作业处理，也可在“作业管理”中查看Distcp作业是否执行成功。

## 1.4 添加作业

用户可将自己开发的程序提交到MRS中，执行程序并获取结果。

本章节以MapReduce作业为例指导您在MRS集群页面如何提交一个新的作业。MapReduce作业用于提交jar程序快速并行处理大量数据，是一种分布式数据处理模式和执行环境。

若在集群详情页面不支持“作业管理”和“文件管理”功能，请通过后台功能来提交作业。

用户创建作业前需要将本地数据上传至OBS系统用于计算分析。当然MRS也支持将OBS中的数据导入至HDFS中，并使用HDFS中的数据进行计算分析。数据完成处理和分析后，您可以将数据存储在HDFS中，也可以将集群中的数据导出至OBS系统。需要注意，HDFS和OBS也支持存储压缩格式的数据，目前支持存储bz2、gz压缩格式的数据。

### 通过界面提交作业

**步骤1** 登录MRS管理控制台。

**步骤2** 选择“集群列表 > 现有集群”，选中一个运行中的集群并单击集群名称，进入集群信息页面。

**步骤3** 若集群开启Kerberos认证时执行该步骤，若集群未开启Kerberos认证，请无需执行该步骤。

在“概览”页签的基本信息区域，单击“IAM用户同步”右侧的“单击同步”进行IAM用户同步。

**步骤4** 单击“作业管理”，进入“作业管理”页签。

**步骤5** 单击“添加”，进入“添加作业”页面。

**步骤6** “作业类型”选择“MapReduce”，并配置其他作业信息。

添加作业

\* 作业类型

MapReduce

\* 作业名称

请输入作业名称

\* 执行程序路径

obs://bucket/program/xx.jar

HDFS

OBS

执行程序参数 ?

HDFS

OBS

服务配置参数 ?

参数

值

+

命令参考

yarn jar

确定

取消

表 1-1 作业配置信息

参数	参数说明
作业名称	作业名称，只能由字母、数字、中划线和下划线组成，并且长度为1~64个字符。 <b>说明</b> 建议不同的作业设置不同的名称。
执行程序路径	待执行程序包地址，需要满足如下要求： <ul style="list-style-type: none"><li>最多为1023字符，不能包含 &amp;&gt;,&lt;'\$特殊字符，且不可为空或全空格。</li><li>执行程序路径可存储于HDFS或者OBS中，不同的文件系统对应的路径存在差异。<ul style="list-style-type: none"><li>OBS：以“obs://”开头。示例：obs://wordcount/program/xxx.jar。</li><li>HDFS：以“/user”开头。</li></ul></li><li>SparkScript和HiveScript需要以“.sql”结尾，MapReduce需要以“.jar”结尾，Flink和SparkSubmit需要以“.jar”或“.py”结尾。sql、jar、py不区分大小写。</li></ul>

参数	参数说明
执行程序参数	<p>可选参数，程序执行的关键参数。多个参数间使用空格隔开。</p> <p>配置方法：程序类名 数据输入路径 数据输出路径</p> <ul style="list-style-type: none"><li>程序类名：由用户程序内的函数指定，MRS只负责参数的传入。</li><li>数据输入路径：通过单击“HDFS”或者“OBS”选择或者直接手动输入正确路径。</li><li>数据输出路径：输出路径请手动输入一个不存在的目录。最多为150000字符，不能包含 &amp;&gt;&lt;'\$特殊字符，可为空。</li></ul> <p><b>注意</b></p> <p>若输入带有敏感信息（如登录密码）的参数可能在作业详情展示和日志打印中存在暴露的风险，请谨慎操作。</p>
服务配置参数	<p>可选参数，用于为本次执行的作业修改服务配置参数，该参数的修改仅适用于本次执行的作业。</p> <p>如需添加多个参数，请单击右侧⊕增加，如需删除参数，请单击右侧“删除”。</p> <p>常用服务配置参数如表1-2。</p>
命令参考	用于展示提交作业时提交到后台执行的命令。

表 1-2 服务配置参数

参数	参数说明	取值样例
fs.obs.access.key	访问OBS的密钥ID。	-
fs.obs.secret.key	访问OBS与密钥ID对应的密钥。	-

**步骤7** 确认作业配置信息，单击“确定”，完成作业的新增。

作业新增完成后，可对作业进行管理。

----结束

## 1.5 删除集群

如果作业执行结束后不需要集群，可以删除MRS集群。集群删除或退订后不再产生费用。

### 背景信息

一般在数据完成分析和存储后或集群异常无法提供服务时才执行集群删除操作。当MRS集群部署失败时，集群会被自动删除。

## 操作步骤

**步骤1** 登录MRS管理控制台。

**步骤2** 在左侧导航栏中选择“集群列表 > 现有集群”。

**步骤3** 在需要删除的集群对应的“操作”列中，单击“删除”。

集群状态由“运行中”更新为“删除中”，待集群删除成功后，集群状态更新为“已删除”，并且显示在“历史集群”中。集群删除后不再产生费用。

----结束

# 2 安装并使用集群客户端

本快速入门指导您在创建了MRS 3.x及之后版本集群后快速安装MRS集群所有服务的客户端并使用。

客户端可以安装在集群内节点，也可以安装在集群外的节点，本示例为您介绍在集群内的节点上安装及使用客户端的操作流程。

## 说明

集群安装有Flume组件时，Flume客户端需要单独安装才可以使用。Flume客户端安装请参见[安装Flume客户端](#)。

本指导的基本内容如下所示：

1. [下载客户端](#)
2. [安装客户端](#)
3. [使用客户端](#)

## 下载客户端

**步骤1** 参考[访问FusionInsight Manager（MRS 3.x及之后版本）](#)登录集群的FusionInsight Manager界面。

**步骤2** 下载集群客户端软件包到待安装的节点上。


在“主页”页签的集群名称后单击 ，单击“下载客户端”下载集群客户端。

图 2-1 下载客户端



步骤3 在“下载集群客户端”弹窗中填写集群客户端下载信息。

图 2-2 下载集群客户端提示框



- “选择客户端类型”中选择“完整客户端”
- “选择平台类型”必须与待安装节点的架构匹配，以“x86\_64”为例。
- 勾选“仅保存到如下路径”，填写下载路径，本示例以“/opt/Bigdata/client”为例，需确保omm用户对该路径有操作权限。

#### 说明

集群支持下载x86\_64和aarch64两种类型客户端，但是客户端类型必须与待安装节点的架构匹配，否则客户端会安装失败。

步骤4 客户端软件包下载完成后，以root用户登录集群的主OMS节点。

客户端软件包默认下载至集群的主OMS节点（可通过FusionInsight Manager的“主机”界面查看带有★标识的节点），如需要在集群内其他节点上安装，执行以下命令将软件包传输至其他节点，否则本步骤可忽略。

在MRS服务管理控制台的集群列表中，单击集群名称，在集群的“节点管理”页签，单击要登录的节点名称，在ECS详情页面可通过远程登录方式登录对应节点。

概览   <b>节点管理</b>   组件管理   告警管理   文件管理   作业管理	
新增节点组	
节点组名称	节点类型
^ master_node_default_group	Master
节点名称 展开	IP 展开
71b38c44-a170-4f0a-a5d4-82a071dc7594_node_master3qxew	192.168.189.159
71b38c44-a170-4f0a-a5d4-82a071dc7594_node_master2KSgs	192.168.125.251
71b38c44-a170-4f0a-a5d4-82a071dc7594_node_master1cElQ	192.168.146.102

```
scp -p /opt/Bigdata/client/FusionInsight_Cluster_1_Services_Client.tar 待安装客  
户端节点的IP地址:/opt/Bigdata/client
```

----结束

## 安装客户端

**步骤1** 以待安装客户端的用户（例如root用户）登录客户端软件包的节点，执行以下命令解压软件包。

```
cd /opt/Bigdata/client
```

```
tar -xvf FusionInsight_Cluster_1_Services_Client.tar
```

**步骤2** 执行sha256sum命令校验解压得到的文件。

```
sha256sum -c FusionInsight_Cluster_1_Services_ClientConfig.tar.sha256
```

```
FusionInsight_Cluster_1_Services_Client.tar: OK
```

**步骤3** 解压获取的安装文件。

```
tar -xvf FusionInsight_Cluster_1_Services_ClientConfig.tar
```

**步骤4** 进入安装包所在目录，安装客户端。

```
cd /opt/Bigdata/client/FusionInsight_Cluster_1_Services_ClientConfig
```

执行如下命令安装客户端到指定目录（绝对路径），例如安装到“/opt/hadoopclient”目录，等待客户端安装完成。

```
./install.sh /opt/hadoopclient
```

```
...  
The component client is installed successfully
```



### 📖 说明

- 如果已经安装的全部服务或某个服务的客户端使用了“/opt/hadoopclient”目录，再安装其他服务的客户端时，需要使用不同的目录。
- 卸载客户端请删除客户端安装目录。
- 如果要求安装后的客户端仅能被该安装用户使用，请在安装时加“-o”参数，即执行./install.sh /opt/hadoopclient -o命令安装客户端。
- 由于HBase使用的Ruby语法限制，如果安装的客户端中包含了HBase客户端，建议客户端安装目录路径只包含大写字母、小写字母、数字以及\_?!.@+=字符。
- 如果安装NTP服务器为chrony模式，请在安装时加“chrony”参数，即执行./install.sh /opt/hadoopclient -o chrony命令安装客户端。

----结束

## 使用客户端

**步骤1** 使用客户端安装用户登录已安装了客户端的节点，执行以下命令切换到客户端目录。

```
cd /opt/hadoopclient
```

**步骤2** 执行以下命令加载环境变量。

```
source bigdata_env
```

**步骤3** 如果当前集群已启用Kerberos认证，执行以下命令认证当前用户。如果当前集群未启用Kerberos认证，则通常情况下无需认证。

```
kinit MRS集群用户
```

例如：

```
kinit admin
```

**步骤4** 直接执行组件的客户端命令。

例如：

使用HDFS客户端命令查看HDFS根目录文件。

```
hdfs dfs -ls /
```

```
Found 15 items
drwxrwx--x - hive      hive      0 2021-10-26 16:30 /apps
drwxr-xr-x - hdfs     hadoop   0 2021-10-18 20:54 /datasets
drwxr-xr-x - hdfs     hadoop   0 2021-10-18 20:54 /datastore
drwxrwx---+ - flink   hadoop   0 2021-10-18 21:10 /flink
drwxr-x--- - flume    hadoop   0 2021-10-18 20:54 /flume
drwxrwx--x - hbase    hadoop   0 2021-10-30 07:31 /hbase
...
```

----结束

# 3 快速使用 Kerberos 认证集群

本章节提供从零开始使用安全集群并执行MapReduce程序、Spark程序和Hive程序的操作指导。

MRS 3.x版本Presto组件暂不支持开启Kerberos认证。

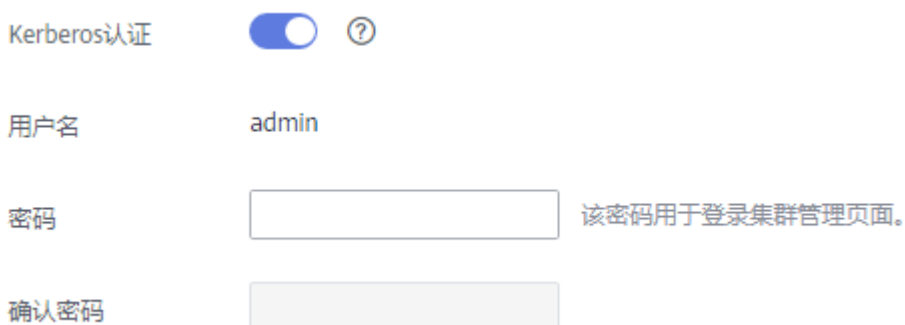
本指导的基本内容如下所示：

1. [创建安全集群并登录其Manager](#)
2. [创建角色和用户](#)
3. [执行MapReduce程序](#)
4. [执行Spark程序](#)
5. [执行Hive程序](#)

## 创建安全集群并登录其 Manager

**步骤1** 创建安全集群，请参见[创建集群](#)页面，开启“Kerberos认证”参数开关，并配置“密码”、“确认密码”参数。该密码用于登录Manager，请妥善保管。

图 3-1 安全集群参数配置



该图展示了安全集群参数配置界面。界面包含以下元素：

- Kerberos认证**：右侧有一个蓝色的开关按钮，处于开启状态，旁边有一个问号图标。
- 用户名**：显示为“admin”。
- 密码**：右侧有一个输入框，旁边有提示文字“该密码用于登录集群管理页面。”。
- 确认密码**：右侧有一个灰色的输入框。

**步骤2** 登录MRS管理控制台页面。

**步骤3** 单击“集群列表”，在“现有集群”列表，单击指定的集群名称，进入集群信息页面。

**步骤4** 单击“集群管理页面”后的“前往Manager”，打开Manager页面。

- 若用户创建集群时已经绑定弹性公网IP，如图3-2所示。
  - a. 添加安全组规则，默认填充的是用户访问公网IP地址9022端口的规则。如需对安全组规则进行查看，修改和删除操作，请单击“管理安全组规则”。

#### 说明

- 自动获取的访问公网IP与用户本机IP不一致，属于正常现象，无需处理。
- 9022端口为knox的端口，需要开启访问knox的9022端口权限，才能访问Manager服务。
- b. 勾选“我确认xx.xx.xx.xx为可信任的公网访问IP，并允许从该IP访问MRS Manager页面。”

图 3-2 访问 Manager 页面

### 访问MRS Manager页面

访问MRS Manager页面需绑定弹性公网IP以及添加安全组规则。 [了解更多](#)

弹性公网IP ?

[管理弹性公网IP](#)

安全组

添加安全组规则 ?

[管理安全组规则](#)

☒ 我确认1.1.1.1为可信任的公网访问IP，并允许从该IP访问MRS Manager页面。

确定

取消

- 若用户创建集群时暂未绑定弹性公网IP，如图3-3所示。
  - a. 在弹性公网IP下拉框中选择可用的弹性公网IP或单击“管理弹性公网IP”购买弹性公网IP。
  - b. 添加安全组规则，默认填充的是用户访问公网IP地址9022端口的规则。如需对安全组规则进行查看，修改和删除操作，请点击“管理安全组规则”。

#### 说明

- 自动获取的访问公网IP与用户本机IP不一致，属于正常现象，无需处理。
- 9022端口为knox的端口，需要开启访问knox的9022端口权限，才能访问Manager服务。
- c. 勾选“我确认xx.xx.xx.xx为可信任的公网访问IP，并允许从该IP访问MRS Manager页面。”

图 3-3 访问 Manager 页面设置

### 访问MRS Manager页面

访问MRS Manager页面需绑定弹性公网IP以及添加安全组规则。 [了解更多](#)

弹性公网IP ?

[管理弹性公网IP](#)

安全组

添加安全组规则 ?

[管理安全组规则](#)

☒ 我确认  为可信任的公网访问IP，并允许从该IP访问MRS Manager页面。

确定

取消

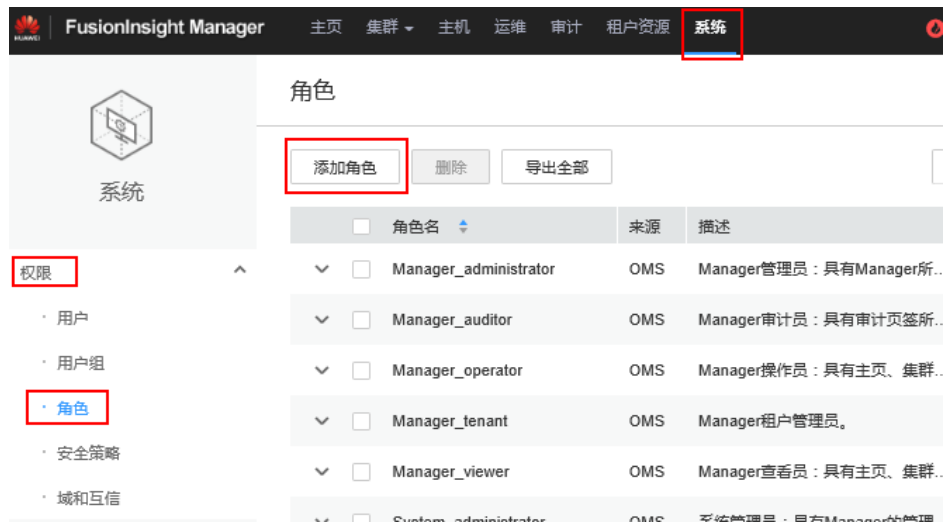
- 步骤5 单击“确定”，进入Manager登录页面，如需给其他用户开通访问Manager的权限，请参见[访问Manager](#)章节，添加对应用户访问公网的IP地址为可信范围。
- 步骤6 输入创建集群时默认的用户名“admin”及设置的密码，单击“登录”进入Manager页面。
- 结束

创建角色和用户

开启Kerberos认证的集群，必须通过以下步骤创建一个用户并分配相应权限来允许用户执行程序。

- 步骤1 在Manager界面选择“系统 > 权限 > 角色”。

图 3-4 角色



- 步骤2 单击“添加角色”，详情请参见[创建角色](#)。

图 3-5 添加角色

角色 > 添加角色

\* 角色名称:

配置资源权限:

所有资源

所有资源

描述

Manager

集群管理

mrs\_

描述:

确定

取消

- 填写如下信息：
- 填写角色的名称，例如mrrole。
  - 在“配置资源权限”选择待操作的集群，然后选择“Yarn > 调度队列 > root”，勾选“权限”列中的“提交”和“管理”，勾选完全后，不要单击确认，要单击如下图的待操作的集群名，再进行后面权限的选择。

图 3-6 配置 Yarn 服务权限

配置资源权限: 所有资源 **mrs\_** Yarn > 调度队列 > root

资源名称	资源类型	提交	管理
launcher-job	子队列	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
default	子队列	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>

- 选择“HBase > HBase Scope”，勾选global的“权限”列的“创建”、“读”、“写”和“执行”，勾选完全后，不要单击确认，要单击如下图的待操作的集群名，再进行后面权限的选择。

图 3-7 配置 HBase 服务权限

配置资源权限: 所有资源 **mrs\_** HBase > HBase Scope

资源名称	资源类型	管理	创建	读	写	执行
global	全局	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>

- 选择“HDFS > 文件系统 > hdfs://hacluster/”，勾选“权限”列的“读”、“写”和“执行”，勾选完全后，不要单击确认，要单击如下图的待操作的集群名，再进行后面权限的选择。

图 3-8 配置 HDFS 服务权限

配置资源权限: 所有资源 **mis\_集群资源** > HDFS > 文件系统

资源名称	资源类型	权限		
		读	写	执行
hdfs://hacluster/	目录	<input checked="" type="checkbox"/> ⓘ	<input checked="" type="checkbox"/> ⓘ	<input checked="" type="checkbox"/> ⓘ
viewfs://ClusterXX/	目录	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

- 选择“Hive > Hive读写权限”，勾选“权限”列的“查询”、“删除”、“插入”和“建表”，单击“确定”，完成角色的创建。

图 3-9 配置 Hive 服务权限

配置资源权限: 所有资源 > mis\_集群资源 > Hive > Hive读写权限

资源名称	资源类型	权限			
		查询	删除	插入	建表
default	数据库	<input checked="" type="checkbox"/> ⓘ	<input checked="" type="checkbox"/> ⓘ	<input checked="" type="checkbox"/> ⓘ	<input checked="" type="checkbox"/> ⓘ
test	数据库	<input checked="" type="checkbox"/> ⓘ	<input checked="" type="checkbox"/> ⓘ	<input checked="" type="checkbox"/> ⓘ	<input checked="" type="checkbox"/> ⓘ

描述:

确定 取消

**步骤3** 选择“系统 > 权限 > 用户组 > 添加用户组”，为样例工程创建一个用户组，例如 mrgroup，详情请参见[创建用户组](#)。

图 3-10 添加用户组

FusionInsight Manager 主页 集群 主机 运维 审计 租户资源 **系统**

系统

权限

用户组

添加用户组

\* 组名: mrgroup

角色: 添加 清除全部

用户: 添加 清除全部

描述:

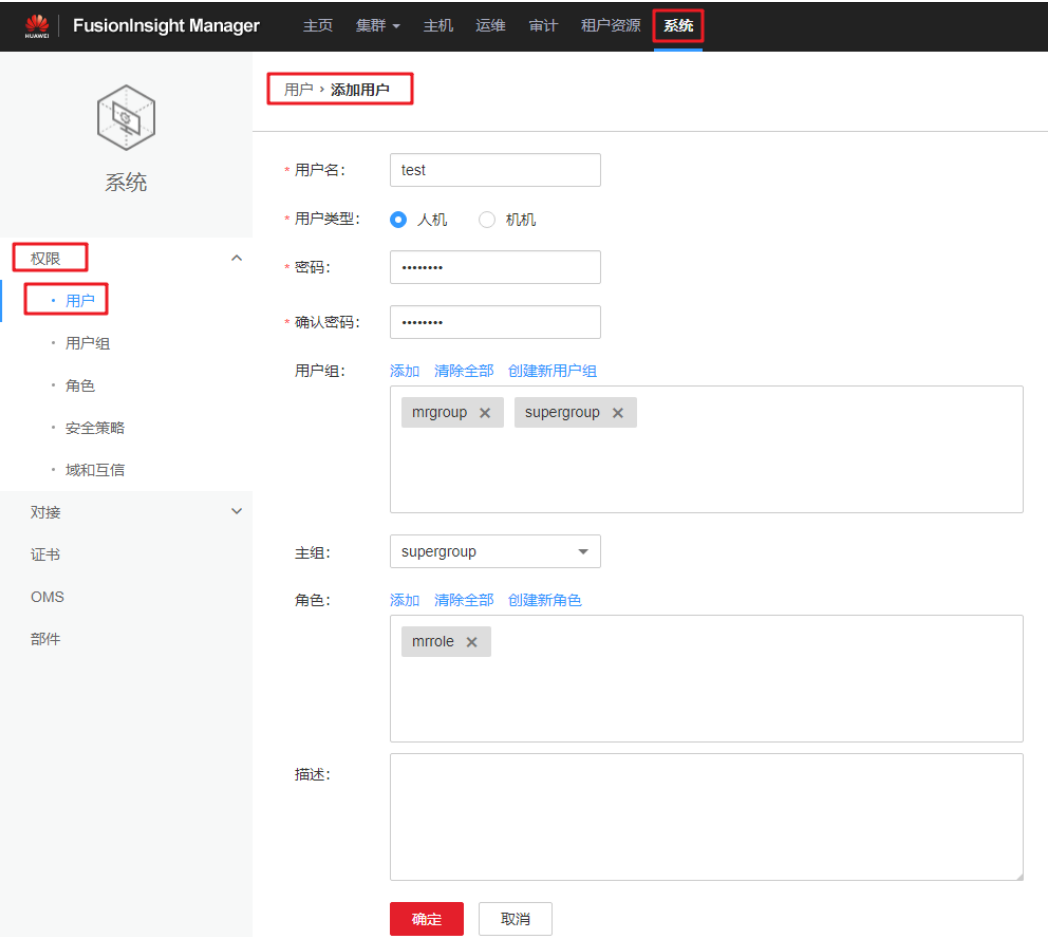
确定 取消

**步骤4** 选择“系统 > 权限 > 用户 > 添加用户”，为样例工程创建一个用户，详情请参见[创建用户](#)。

- 填写用户名，例如test，当需要执行Hive程序时，请设置用户名为“hiveuser”。
- 用户类型为“人机”用户。

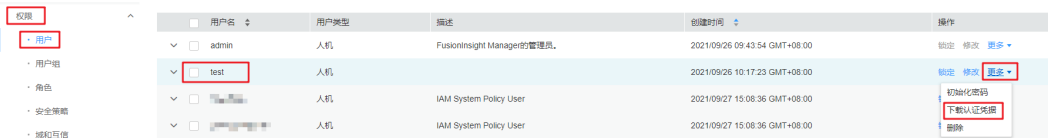
- 输入密码（特别注意该密码在后面运行程序时要用到）。
- 加入用户组mrgroup和supergroup。
- 设置其“主组”为supergroup，并绑定角色mrrole取得权限。  
单击“确定”完成用户创建。

图 3-11 添加用户



**步骤5** 选择“系统 > 权限 > 用户”，选择新建用户**test**，选择“更多 > 下载认证凭据”，保存后解压得到用户的keytab文件与krb5.conf文件。

图 3-12 下载认证凭据



----结束

执行 MapReduce 程序

本小节提供执行MapReduce程序的操作指导，旨在指导用户在安全集群模式下运行程序。

前提条件

已编译好待运行的程序及对应的数据文件，如mapreduce-examples-1.0.jar、input\_data1.txt和input\_data2.txt，MapReduce程序开发及数据准备请参见[MapReduce应用开发](#)。

### 操作步骤

**步骤1** 采用远程登录软件（比如：MobaXterm）通过ssh登录（使用集群弹性IP登录）到安全集群的master节点。

**步骤2** 登录成功后分别执行下列命令，在/opt/Bigdata/client目录下创建test文件夹，在test目录下创建conf文件夹：

```
cd /opt/Bigdata/client
mkdir test
cd test
mkdir conf
```

**步骤3** 使用上传工具（比如：WinScp）将mapreduce-examples-1.0.jar、input\_data1.txt和input\_data2.txt复制到test目录下，将“创建角色和用户”中的步骤[步骤5](#)获得的keytab文件和krb5.conf文件复制到conf目录。

**步骤4** 执行如下命令配置环境变量并认证已创建用户，例如test。

```
cd /opt/Bigdata/client
source bigdata_env
export YARN_USER_CLASSPATH=/opt/Bigdata/client/test/conf/
kinit test
```

然后按照提示输入密码，无异常提示返回（首次登录需按照系统提示修改密码），则完成了用户的kerberos认证。

**步骤5** 执行如下命令将数据导入到HDFS中：

```
cd test
hdfs dfs -mkdir /tmp/input
hdfs dfs -put input_data* /tmp/input
```

**步骤6** 执行如下命令运行程序：

```
yarn jar mapreduce-examples-1.0.jar com.huawei.bigdata.mapreduce.examples.FemaleInfoCollector /tmp/
input /tmp/mapreduce_output
```

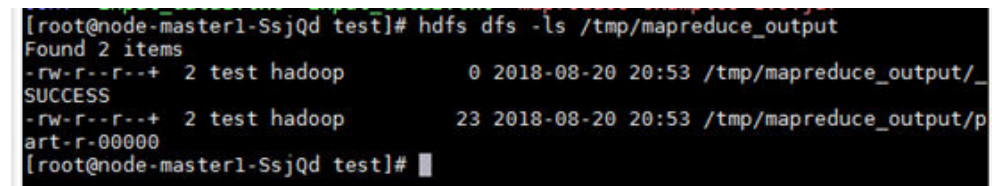
其中：

/tmp/input指HDFS文件系统中input的路径。

/tmp/mapreduce\_output指HDFS文件系统中output的路径，该目录必须不存在，否则会报错。

**步骤7** 程序运行成功后，执行 hdfs dfs -ls /tmp/mapreduce\_output会显示如下：

图 3-13 查看程序运行结果



```
[root@node-master1-SsjQd test]# hdfs dfs -ls /tmp/mapreduce_output
Found 2 items
-rw-r--r--+ 2 test hadoop          0 2018-08-20 20:53 /tmp/mapreduce_output/_
SUCCESS
-rw-r--r--+ 2 test hadoop         23 2018-08-20 20:53 /tmp/mapreduce_output/p
art-r-00000
[root@node-master1-SsjQd test]#
```

----结束

## 执行 Spark 程序

本小节提供执行Spark程序的操作指导，旨在指导用户在安全集群模式下运行程序。



### 前提条件

已编译好待运行的程序及对应的数据文件，如FemaleInfoCollection.jar、input\_data1.txt和input\_data2.txt，Spark程序开发及数据准备请参见[Spark应用开发](#)。

### 操作步骤

**步骤1** 采用远程登录软件（比如：MobaXterm）通过ssh登录（使用集群弹性IP登录）到安全集群的master节点。

**步骤2** 登录成功后分别执行下列命令，在/opt/Bigdata/client目录下创建test文件夹，在test目录下创建conf文件夹：

```
cd /opt/Bigdata/client
mkdir test
cd test
mkdir conf
```

**步骤3** 使用上传工具（比如：WinScp）将FemaleInfoCollection.jar、input\_data1.txt和input\_data2.txt复制到test目录下，将“创建角色和用户”中的步骤[步骤5](#)获得的keytab文件和krb5.conf文件复制到conf目录。

**步骤4** 执行如下命令配置环境变量并认证已创建用户，例如test。

```
cd /opt/Bigdata/client
source bigdata_env
export YARN_USER_CLASSPATH=/opt/Bigdata/client/test/conf/
kinit test
```

然后按照提示输入密码，无异常提示返回，则完成了用户的kerberos认证。

**步骤5** 执行如下命令将数据导入到HDFS中：

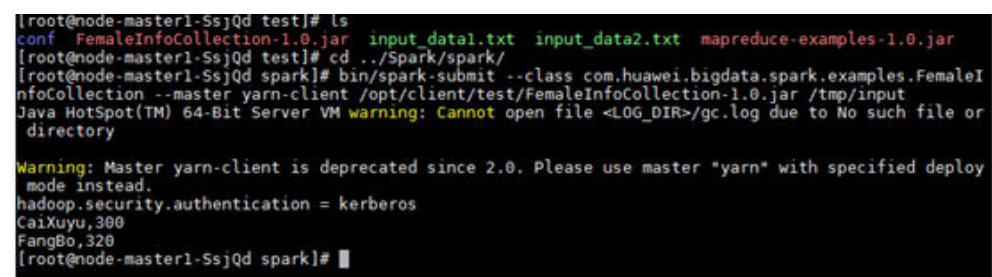
```
cd test
hdfs dfs -mkdir /tmp/input
hdfs dfs -put input_data* /tmp/input
```

**步骤6** 执行如下命令运行程序：

```
cd /opt/Bigdata/client/Spark/spark
bin/spark-submit --class com.huawei.bigdata.spark.examples.FemaleInfoCollection --master yarn-client /opt/Bigdata/client/test/FemaleInfoCollection-1.0.jar /tmp/input
```

**步骤7** 程序运行成功后，会显示如下：

图 3-14 程序运行结果



```
[root@node-master1-SsjQd test]# ls
conf  FemaleInfoCollection-1.0.jar  input_data1.txt  input_data2.txt  mapreduce-examples-1.0.jar
[root@node-master1-SsjQd test]# cd ../Spark/spark/
[root@node-master1-SsjQd spark]# bin/spark-submit --class com.huawei.bigdata.spark.examples.FemaleInfoCollection --master yarn-client /opt/client/test/FemaleInfoCollection-1.0.jar /tmp/input
Java HotSpot(TM) 64-Bit Server VM warning: Cannot open file <LOG_DIR>/gc.log due to No such file or directory

Warning: Master yarn-client is deprecated since 2.0. Please use master "yarn" with specified deploy mode instead.
hadoop.security.authentication = kerberos
CaiXuyu,300
FangBo,320
[root@node-master1-SsjQd spark]#
```

----结束

## 执行 Hive 程序

本小节提供执行Hive程序的操作指导，旨在指导用户在安全集群模式下运行程序。

### 前提条件

已编译好待运行的程序及对应的数据文件，如hive-examples-1.0.jar、input\_data1.txt和input\_data2.txt，Hive程序开发及数据准备请参见[Hive应用开发](#)。

### 操作步骤

**步骤1** 采用远程登录软件（比如：MobaXterm）通过ssh登录（使用集群弹性IP登录）到安全集群的master节点。

**步骤2** 登录成功后分别执行下列命令，在/opt/Bigdata/client目录下创建test文件夹，在test目录下创建conf文件夹：

```
cd /opt/Bigdata/client
mkdir test
cd test
mkdir conf
```

**步骤3** 使用上传工具（比如：WinScp）将样FemaleInfoCollection.jar、input\_data1.txt和input\_data2.txt复制到test目录下，将“创建角色和用户”中的步骤[步骤5](#)获得的keytab文件和krb5.conf文件复制到conf目录。

**步骤4** 执行如下命令配置环境变量并认证已创建用户，例如test。

```
cd /opt/Bigdata/client
source bigdata_env
export YARN_USER_CLASSPATH=/opt/Bigdata/client/test/conf/
kinit test
```

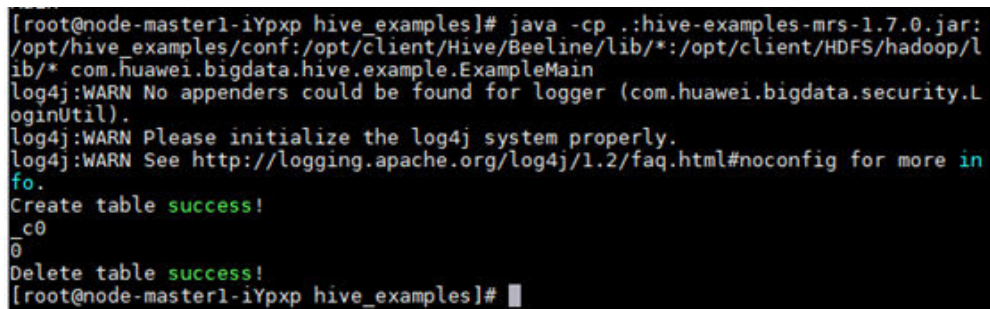
然后按照提示输入密码，无异常提示返回，则完成了用户的kerberos认证。

**步骤5** 执行如下命令运行程序：

```
chmod +x /opt/hive_examples -R cd /opt/hive_examples java -cp ./hive-examples-1.0.jar:/opt/hive_examples/conf:/opt/Bigdata/client/Hive/Beeline/lib/*:/opt/Bigdata/client/HDFS/hadoop/lib/* com.huawei.bigdata.hive.example.ExampleMain
```

**步骤6** 程序运行成功后，会显示如下：

图 3-15 程序运行的结果



```
[root@node-master1-iYpxp hive_examples]# java -cp ./hive-examples-mrs-1.7.0.jar:/opt/hive_examples/conf:/opt/client/Hive/Beeline/lib/*:/opt/client/HDFS/hadoop/lib/* com.huawei.bigdata.hive.example.ExampleMain
log4j:WARN No appenders could be found for logger (com.huawei.bigdata.security.LoginUtil).
log4j:WARN Please initialize the log4j system properly.
log4j:WARN See http://logging.apache.org/log4j/1.2/faq.html#noconfig for more info.
Create table success!
_c0
0
Delete table success!
[root@node-master1-iYpxp hive_examples]#
```

----结束

# 4 从零开始使用 Hadoop

---

图文版。

# 5 从零开始使用 Kafka

---

图文版。

# 6 从零开始使用 HBase

MapReduce服务（MapReduce Service）提供租户完全可控的企业级大数据集群云服务，轻松运行Hadoop、Spark、HBase、Kafka等大数据组件。

本入门以不开启Kerberos认证的集群为例提供从零开始使用HBase的操作指导，通过登录HBase客户端后创建表，往表中插入数据并修改表数据。

本指导的基本内容如下所示：

1. [准备MRS集群](#)
2. [安装HBase客户端](#)
3. [使用HBase客户端创建表](#)

## 准备 MRS 集群

**步骤1** 购买集群。

1. 登录华为云管理控制台。
2. 选择“大数据 > MapReduce服务 MRS”，进入MapReduce MRS服务管理控制台。



3. 选择“集群列表 > 现有集群”，单击“购买集群”，进入快速购买页面，选择“自定义购买”。



**步骤2** 参考下列软件配置参数说明填写配置，并单击“下一步”。

- “区域”请根据需要选择。
- “集群名称”填写“mrs\_demo”或按命名规范命名。
- “集群版本”选择“MRS 3.1.0”。

区域

不同区域的资源之间内网不互通。请选择靠近您客户的区域，可以降低网络延时、提高访问速度。[如何选择区域](#)

集群名称

mrs\_demo

?

集群版本

集群类型

分析集群

流式集群

混合集群

自定义

- “集群类型”选择“分析集群”，并勾选HBase组件。

集群类型

分析集群

流式集群

混合集群

自定义

必选组件默认勾选，被依赖的组件会被自动勾选。

分析组件

<input checked="" type="checkbox"/>	组件名	版本	描述
<input checked="" type="checkbox"/>	Hadoop	3.1.1	针对大数据集的分布式数据处理框架。
<input type="checkbox"/>	Spark2x	2.4.5	Spark2x是一个对大规模数据处理的快速和通用引擎,基于开源Spark2.x版本开发。
<input checked="" type="checkbox"/>	HBase	2.2.3	HBase是一个高可靠性、高性能、面向列、可伸缩的分布式存储系统。
<input type="checkbox"/>	Hive	3.1.0	方便查询、管理存储在分布式存储系统上的大数据集的数据仓库软件。
<input type="checkbox"/>	Hue	4.7.0	Apache Hadoop的UI界面。
<input type="checkbox"/>	Flink	1.12.0	Flink是一个开源的批流合一的消息处理系统。
<input type="checkbox"/>	Oozie	5.1.0	Hadoop作业调度系统。
<input checked="" type="checkbox"/>	ZooKeeper	3.5.6	Zookeeper是一个集中的服务，它用于维护配置信息、命名、提供分布式的同步和提供分组服务。
<input checked="" type="checkbox"/>	Ranger	2.0.0	RANGER是一个基于Hadoop平台，旨在为复杂数据提供监控、操作、管理等权限管控接口的集中式框架。
<input type="checkbox"/>	Tez	0.9.2	一个支持有向无环图的分布式计算框架。
<input type="checkbox"/>	Impala	3.4.0	一种处理大量数据的SQL查询引擎。
<input type="checkbox"/>	Presto	333	一种开源、分布式SQL查询引擎。
<input type="checkbox"/>	Kudu	1.12.1	一个能对快速数据进行快速分析的Hadoop存储引擎
<input type="checkbox"/>	Sqoop	1.4.7	Sqoop是专为Apache Hadoop和结构化数据库（如关系型数据库）设计的高效传输大量数据的工具。

**步骤3** 在“硬件配置”页面，参考表6-1填写配置。单击“下一步”。

**表 6-1** MRS 集群硬件配置

参数名称	示例
计费模式	按需计费
可用区	可用区2
虚拟私有云	保持默认不修改，也可单击“查看虚拟私有云”重新创建。

参数名称	示例
弹性公网IP	可选择下拉框中已有的弹性公网IP进行绑定。若下拉框中没有可选的弹性公网IP，可以单击“管理弹性公网IP”进入弹性公网IP服务进行创建。
企业项目	default

图 6-1 硬件配置

计费模式

包年/包月

按需计费

可用区

可用区1

可用区2

可用区3

可用区7

虚拟私有云

查看虚拟私有云

子网

查看子网

当前子网剩余可用IP数: 224

安全组

自动创建

管理安全组

弹性公网IP

管理弹性公网IP

企业项目

default

新建企业项目

集群节点

节点类型

计费模式

实例规格

实例数量

Master节点

按需计费

通用计算增强型  
16 vCPUs | 64 GB | c6.4xlarge.4  
系统盘  
通用型SSD 480 GB x 1  
数据盘  
通用型SSD 600 GB x 1

2

分析Core节点

按需计费

通用计算增强型  
16 vCPUs | 64 GB | c6.4xlarge.4  
系统盘  
通用型SSD 480 GB x 1  
数据盘  
通用型SSD 600 GB x 1

3

分析Task节点

按需计费

步骤4 高级设置。

1. 在“高级配置”页面，参考表6-2填写配置。

表 6-2 MRS 集群高级配置拓扑

参数名称	示例
Kerberos认证	不开启
密码	Test@!123456
确认密码	Test@!123456
登录方式	密码
密码	Test@#123456
确认密码	Test@#123456
通信安全授权	勾选“确认授权”

图 6-2 高级配置

Kerberos认证

☐

?

用户名

admin

密码

该密码用于登录集群管理页面。

确认密码

登录方式

密码

密钥对

用户名

root

密码

该密码用于远程登录ECS机器或BMS机器。

确认密码

通信安全授权

☒ 确认授权

授权MRS集群创建和切换子网时开通相应的访问控制规则，从而使得用户可以通过MRS管理控制台进行大数据组件部署和后续集群的使用、运维和管理等操作，此时不授权将无法创建集群。[了解更多](#)

需要开通的访问控制规则

协议端口	类型	源地址	描述
TCP : 9022	IPv4		MRS 默认访问控制规则
TCP : 9022	IPv4		MRS 默认访问控制规则
TCP : 9022	IPv4		MRS 默认访问控制规则
TCP : 9022	IPv4		MRS 默认访问控制规则
TCP : 9022	IPv4		MRS 默认访问控制规则

2. 单击“立即购买”，进入任务提交成功页面。
3. 单击“返回集群列表”，在“现有集群”列表中可以查看到集群创建的状态。
4. 集群创建需要时间，所创集群的初始状态为“启动中”，创建成功后状态更新为“运行中”，请您耐心等待。

----结束

安装 HBase 客户端

- 步骤1 在“集群列表 > 现有集群”列表中，单击名称“mrs\_demo”，进入集群信息页面。
- 步骤2 单击“集群管理页面 ”后的“前往 Manager”，在弹出的窗口中配置弹性IP信息，单击“确定”，输入用户名和密码进入Manager界面。

概览 | 节点管理 | 组件管理 | 告警管理 | 文件管理 | 作业管理 | 租户管理 | 标签管理

基本信息 [了解更多](#)

集群名称

集群状态 运行中

集群管理页面 ? [前往 Manager](#)

- 步骤3 在Manager界面，选择“集群 > 服务 > HBase > 更多 > 下载客户端”，选择“完整客户端”、对应的平台类型，勾选“仅保存到如下路径”，单击“确定”。





**步骤4** 以root用户登录主节点。

**步骤5** 进入安装包所在目录，执行如下命令解压、校验安装包。并解压获取的安装文件。

```
cd /tmp/FusionInsight-Client
```

```
tar -xvf FusionInsight_Cluster_1_HBase_Client.tar
```

```
sha256sum -c FusionInsight_Cluster_1_HBase_ClientConfig.tar.sha256
```

```
tar -xvf FusionInsight_Cluster_1_HBase_ClientConfig.tar
```

**步骤6** 进入安装包所在目录，执行如下命令安装客户端到指定目录（绝对路径），例如安装到“/opt/hbaseclient”目录。

```
cd /tmp/FusionInsight-Client/FusionInsight_Cluster_1_HBase_ClientConfig
```

执行./install.sh /opt/hbaseclient命令，等待客户端安装完成。

**步骤7** 检查客户端是否安装成功。

```
cd /opt/hbaseclient
```

```
source bigdata_env
```

```
hbase shell
```

执行成功则说明HBase客户端安装成功。

----结束

## 使用 HBase 客户端创建表

**步骤1** 登录Master节点（VNC方式）。

1. 在MRS控制台界面，选择“集群列表 > 现有集群”，在列表中选择“mrs\_demo”名称，选择“节点管理”，单击Master节点类型中名称包含“master1”的节点，跳转至该节点的弹性云服务器详情页面。



2. 单击页面右上角的“远程登录”，远程登录Master节点。使用root用户登录，密码为购买集群时设置的密码。



**步骤2** 执行以下命令切换到客户端目录。

```
cd /opt/hbaseclient
```

**步骤3** 执行以下命令配置环境变量。

```
source bigdata_env
```

**说明**

若集群开启Kerberos认证，需执行以下命令认证当前用户，当前用户需要具有创建HBase表的权限。

例如：

```
kinit hbaseuser
```

**步骤4** 执行以下命令进入HBase Shell命令行。

```
hbase shell
```

**步骤5** 运行HBase客户端命令，创建表“user\_info”。

1. 创建表“user\_info”并添加相关数据。

```
create 'user_info',{NAME => 'i'}  
put 'user_info','12005000201','i:name','A'  
put 'user_info','12005000201','i:gender','Male'  
put 'user_info','12005000201','i:age','19'  
put 'user_info','12005000201','i:address','City A'
```
2. 在用户信息表“user\_info”中新增用户的学历、职称信息。

```
put 'user_info','12005000201','i:degree','master'
```

```
put 'user_info','12005000201','i:pose','manager'
```

3. 根据用户编号查询用户姓名和地址。

```
scan 'user_info',
```

```
{STARTROW=>'12005000201',STOPROW=>'12005000201',COLUMNS=>['i:name','i:address']}
```

ROW	COLUMN
+CELL	
12005000201	column=i:address, timestamp=2021-10-30T10:21:42.196, value=City
A	
12005000201	column=i:name, timestamp=2021-10-30T10:21:18.594,
value=A	
1 row(s)	
Took 0.0996 seconds	

4. 根据用户姓名进行查询。

```
scan 'user_info',{FILTER=>"SingleColumnValueFilter('i','name',=,'binary:A')"}]
```

ROW	COLUMN
+CELL	
12005000201	column=i:address, timestamp=2021-10-30T10:21:42.196, value=City
A	
12005000201	column=i:age, timestamp=2021-10-30T10:21:30.777,
value=19	
12005000201	column=i:degree, timestamp=2021-10-30T10:21:53.284,
value=master	
12005000201	column=i:gender, timestamp=2021-10-30T10:21:18.711,
value=Male	
12005000201	column=i:name, timestamp=2021-10-30T10:21:18.594,
value=A	
12005000201	column=i:pose, timestamp=2021-10-30T10:22:07.152,
value=manager	
1 row(s)	
Took 0.2158 seconds	

5. 删除用户信息表中该用户的数据。

```
delete 'user_info','12005000201','i'
```

6. 删除用户信息表。

```
disable 'user_info'
```

```
drop 'user_info'
```

----结束

# 7 修改 MRS 服务配置参数

MRS集群创建后，在日常使用中，您可以通过MRS管理控制台或者集群Manager界面对集群内服务的相关配置参数进行修改。

本入门以修改HBase服务的日志文件数量参数“hbase.log.maxbackupindex”参数为例，提供从零开始修改MRS服务配置参数的操作指导。

本指导的基本内容如下所示：

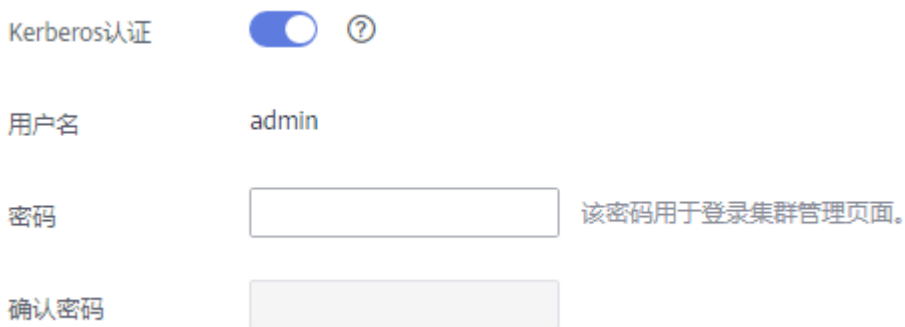
1. [通过MRS管理控制台修改服务参数](#)
2. [通过FusionInsight Manager界面修改服务参数](#)

视频帮助指导请参见：[https://support.huaweicloud.com/mrs\\_video/index.html](https://support.huaweicloud.com/mrs_video/index.html)

## 通过 MRS 管理控制台修改服务参数

**步骤1** 创建安全集群，请参见[创建集群](#)页面，开启“Kerberos认证”参数开关，并配置“密码”、“确认密码”参数。该密码用于登录Manager，请妥善保管。

图 7-1 安全集群参数配置



该图展示了安全集群参数配置界面。界面包含以下元素：

- Kerberos认证**：右侧有一个蓝色的开关按钮，处于开启状态，旁边有一个问号图标。
- 用户名**：输入框中显示为“admin”。
- 密码**：右侧有一个输入框，旁边有提示文字“该密码用于登录集群管理页面。”。
- 确认密码**：下方有一个对应的输入框。

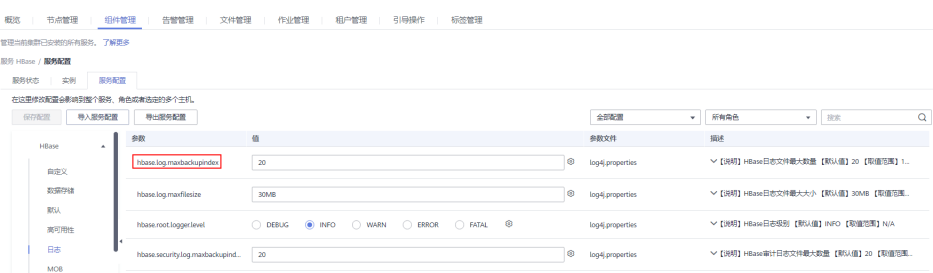
**步骤2** 登录MRS控制台，在左侧导航栏选择“集群列表 > 现有集群”，单击集群名称。

图 7-2 单击集群名称



- 步骤3** 选择“组件管理 > HBase > 服务配置”，在页面右上角选择“全部配置”。
- 步骤4** 在左侧参数导航树中选择“HBase > 日志”。
- 步骤5** 找到参数“hbase.log.maxbackupindex”，根据业务需求修改“值”。

图 7-3 修改参数值



- 步骤6** 修改完成后单击“保存配置”，在弹出框中确认已修改的参数值，单击“是”，等待系统保存并更新配置，单击“完成”。

图 7-4 确认修改参数值



- 步骤7** 查看当前服务配置状态。
- 单击“服务状态”，查看当前服务配置状态，若存在配置过期，选择“更多 > 重启服务”，在弹出框中确认重启服务并单击“是”，等待服务重启。

图 7-5 重启服务



**步骤8** 查看相关服务配置状态。

返回组件管理页面，查看其他相关服务配置状态，若存在配置过期的服务，单击对应“操作”列的“重启”，在弹出框中确认重启服务并单击“是”，等待服务重启。

图 7-6 重启服务

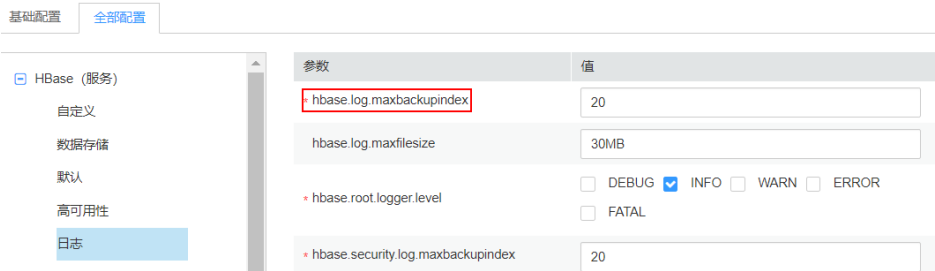


----结束

通过 FusionInsight Manager 界面修改服务参数

- 步骤1** 参考[创建安全集群并登录其Manager](#)创建集群并登录FusionInsight Manager。
- 步骤2** 选择“集群 > 服务 > HBase > 配置 > 全部配置”。
- 步骤3** 选择“HBase（服务）> 日志”。
- 步骤4** 找到参数“hbase.log.maxbackupindex”，根据业务需求修改“值”。

图 7-7 修改参数值



**步骤5** 修改完成后单击“保存”，在弹出框中确认已修改的参数值，单击“确定”，等待系统保存并更新配置，单击“完成”。

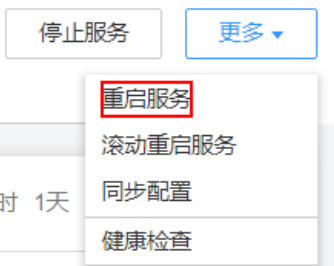
图 7-8 确认修改参数值



步骤6 查看当前服务配置状态。

单击“概览”，查看当前服务配置状态，若存在配置过期，选择“更多 > 重启服务”，输入密码，单击“确定”，等待服务重启。

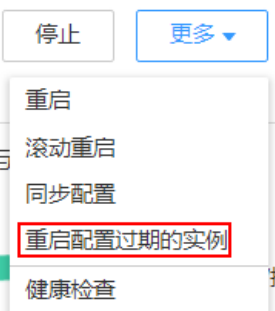
图 7-9 重启服务



步骤7 查看相关服务配置状态。

选择“集群 > 服务”，查看其他相关服务配置状态，若存在配置过期的服务，选择“集群 > 概览 > 更多 > 重启配置过期的实例”，输入密码，单击“确定”可重启所有配置过期的实例。

图 7-10 重启配置过期的实例



----结束

# 8 配置 MRS 集群弹性伸缩

在大数据应用，尤其是实时分析处理数据的场景中，常常需要根据数据量的变化动态调整集群节点数量以增减资源。MRS的弹性伸缩规则功能支持根据集群负载对集群进行弹性伸缩。

- 弹性伸缩规则：根据集群实时负载对Task节点数量进行调整，数据量变化后触发扩缩容，有一定的延后性。
- 资源计划（按时间段设置Task节点数量范围）：若数据量变化存在周期性规律，则可通过资源计划在数据量变化前提前完成集群的扩缩容，避免出现增加或减少资源的延后。

弹性伸缩规则与资源计划均可触发弹性伸缩，两者至少配置其中一种，也可以叠加使用。本入门指导您快速根据实际业务场景配置MRS集群的弹性伸缩规则。

本指导的基本内容如下所示：

1. [创建集群并配置Task节点。](#)
2. [场景一：单独使用弹性伸缩规则。](#)
3. [场景二：单独使用资源计划。](#)
4. [场景三：弹性伸缩规则与资源计划叠加使用。](#)

## 创建集群并配置 Task 节点

### 说明

- 本操作以快速购买一个MRS 3.1.0版本的Hadoop分析集群为例进行说明。
- 仅Task类型节点组支持配置弹性伸缩，在配置弹性伸缩前请先检查当前集群是否有Task节点。

**步骤1** 登录华为云管理控制台，选择“大数据 > MapReduce服务 MRS”，单击“购买集群”，选择“快速购买”，填写软件配置参数，单击“立即购买”。

**表 8-1** 参数配置（以下参数仅供参考，可根据实际情况调整）

参数项	取值
区域	根据实际情况选择
计费模式	按需计费



参数项	取值
集群名称	MRS_demo
集群版本	MRS 3.1.0
组件选择	Hadoop分析集群
可用区	可用区2
虚拟私有云	vpc-gggg
子网	subnet-64db
企业项目	default
Kerberos认证	不开启
用户名	root/admin
密码	设置密码登录集群管理页面及ECS节点用户的密码，例如：Test!@12345。
确认密码	再次输入设置用户密码
通信安全授权	勾选“确认授权”

区域

不同区域的资源之间内网不互通。请选择靠近您客户的区域，可以降低网络延时、提高访问速度。[如何选择区域](#)

计费模式

包年/包月

按需计费

集群名称

mrs\_demo

集群版本

MRS 3.1.0

组件选择

Hadoop分析集群

Hadoop 3.1.1, Hive 3.1.0, Spark2x 2.4.5, Flink 1.12.0, ZooKeeper 3.5.6, Ranger 2.0.0, Tez 0.9.2 and Presto 3...

海量数据分析与查询

HBase查询集群

Hadoop 3.1.1, HBase 2.2.3, ZooKeeper 3.5.6 and Ranger 2.0.0

海量数据存储，实现毫秒级数据查询...

ClickHouse集群

ClickHouse 21.3.4.25 and ZooKeeper 3.5.6

用于联机分析(OLAP)的列式数据库管理系统(DBMS...

实时分析集群

Hadoop 3.1.1, Kafka 2.11-2.4.0, Flink 1.12.0, ClickHouse 21.3.4.25, ZooKeeper 3.5.6 and Ranger 2.0.0

海量数据采集，数据实时分析与查...

可用区

可用区1

可用区2

可用区3

可用区7

**步骤2** 等待集群创建成功后，进入集群，选择“节点管理”页签，查看当前集群是否有Task节点。

- 是，操作完成。
- 否，请执行**步骤3**。

**步骤3** 配置Task节点。

1. 在“节点管理”页签，单击“配置Task节点”，进入“配置Task节点”页面。

说明

- 对于MRS 3.x及之后版本，“配置Task节点”仅适用于分析集群、流式集群和混合集群。
2. 根据需要配置相关参数。

配置Task节点

Task节点是处理数据的实例，不存储集群数据（如HDFS数据）。

节点类型

分析Task节点

节点规格

12 vCPUs 96 GB | m6.3xlarge.8

节点数量

—

1

+

系统盘

通用型SSD

—

480

+

添加数据盘

☒

数据盘

通用型SSD

—

600

+

数据盘数量

—

1

+

确定


取消

3. 单击“确定”。
- 结束

场景一：单独使用弹性伸缩规则

例如业务场景如下：

需要根据Yarn资源使用情况动态调整节点数，在Yarn可用内存低于20%时扩容5个节点，可用内存高于70%时缩容5个节点。Task节点组最高不超过10个节点，最低不少于1个节点。

- 步骤1** 参考[创建集群并配置Task节点](#)创建集群并配置Task节点。
- 步骤2** 在MRS管理控制台，选择“集群列表 > 现有集群”，单击待操作的集群名称，进入集群详情页面。
- 步骤3** 单击“节点管理”页签，在Task节点组的“操作”列单击“弹性伸缩”，进入“弹性伸缩”页面。
- 步骤4** 在“弹性伸缩”界面单击 开启弹性伸缩，并配置“节点数量范围”为“1 - 10”。

弹性伸缩

弹性伸缩规则会影响节点数量的使用变化，因此会对实际价格产生影响。当弹性伸缩开启时，MRS服务会依次检测配置的弹性伸缩规则，由第一个满足条件的规则触发弹性伸缩。

弹性伸缩

节点数量范围

默认范围

1

-

10

配置指定时间段的节点数量范围

您还可以添加5个资源计划。

伸缩规则

扩容

添加规则

缩容

添加规则

策略配置

策略配置

策略配置

我同意授权MRS服务根据以上策略自动进行节点扩容/缩容操作。

确定

取消

步骤5 配置扩容规则。

1. 勾选“伸缩规则”区域的“扩容”。

2. 单击“扩容”后方的“添加规则”，弹出“添加规则”窗口。

3. 在“添加规则”窗口中配置相关参数。

规则名称：保持默认，例如“default-expand-2”

如果：YARNMemoryAvailablePercentage 小于 20%（相关指标项对应含义请参考表8-2）

持续：1个五分钟

添加：5个节点

冷却时间：20分钟

4. 单击“确定”。

添加规则

规则

YARNMemoryAvailablePercentage

如果

YARNMemoryAvail...

小于

20

%

持续

1

个五分钟

添加

5

个节点

冷却时间

20

分钟

确定

取消

步骤6 配置缩容规则。

1. 勾选“伸缩规则”区域的“缩容”。

2. 单击“缩容”后方的“添加规则”，弹出“添加规则”窗口。

文档版本 01 (2022-03-01)

版权所有 © 华为技术有限公司

40

3. 在“添加规则”窗口中配置相关参数。
  - “规则名称”：保持默认，例如“default-shrink-2”
  - “如果”：YARNMemoryAvailablePercentage 大于 70%（相关指标项对应含义请参考表8-2）
  - “持续”：1个五分钟
  - “终止”：5个节点
  - “冷却时间”：20分钟
4. 单击“确定”。

#### 添加规则

规则 YARNMemoryAvailablePercentage

如果 YARNMemoryAvail... 大于 70 %

持续 1 个五分钟

终止 5 个节点

冷却时间 20 分钟

确定 取消

**步骤7** 勾选“我同意授权MRS服务根据以上策略自动进行节点扩容/缩容操作。”。

**步骤8** 单击“确定”，完成弹性伸缩集群设置。

----结束

## 场景二：单独使用资源计划

例如业务场景如下：

需要定时增加或减少节点数，在10:00~22:00由于业务量较大，需要使用10个节点，其余时间使用5个节点。

**步骤1** 参考[创建集群并配置Task节点](#)创建集群并配置Task节点。

**步骤2** 在MRS管理控制台，选择“集群列表 > 现有集群”，单击待操作的集群名称，进入集群详情页面。

**步骤3** 单击“节点管理”页签，在Task节点组的“操作”列单击“弹性伸缩”，进入“弹性伸缩”页面。

**步骤4** 在“弹性伸缩”界面开启弹性伸缩及配置“节点数量范围”。

- “弹性伸缩”：开启
- “节点数量范围”：5 - 5

**步骤5** 单击“默认范围”下方的“配置指定时间段的节点数量范围”，配置相关参数。

- “时间范围”：10:00 - 22:00
- “节点数量范围”：10 - 10

### 弹性伸缩

弹性伸缩规则会影响节点数量的使用变化，因此会对实际价格产生影响。当弹性伸缩开启时，MRS服务会依次检测配置的弹性伸缩规则，由第一个满足条件的规则触发弹性伸缩。

弹性伸缩 ☒

节点数量范围 ? 默认范围  -

时间范围

-

节点数量范围

-  删除

+ 配置指定时间段的节点数量范围 ? 您还可以添加3个资源计划。

伸缩规则 ?

☐ 扩容

添加规则

☒ 我同意授权MRS服务根据以上策略自动进行节点扩容/缩容操作。

确定

取消

**步骤6** 勾选“我同意授权MRS服务根据以上策略自动进行节点扩容/缩容操作。”。

**步骤7** 单击“确定”，完成弹性伸缩集群设置。

----结束

## 场景三：弹性伸缩规则与资源计划叠加使用

本操作以如下为例，配置弹性伸缩规则与资源计划叠加使用操作。

例如业务场景如下：

某项实时处理业务数据量在每天7:00-13:00出现规律性变化，但是数据量变化并非非常平稳。假设在7:00-13:00期间，需要Task节点的数量范围是5-8个，其他时间需要根据YARN组件运行中的任务数动态伸缩Task节点范围为2-4个。

**步骤1** 参考[创建集群并配置Task节点](#)创建集群并配置Task节点。

**步骤2** 在MRS管理控制台，选择“集群列表 > 现有集群”，单击待操作的集群名称，进入集群详情页面。

**步骤3** 单击“节点管理”页签，在Task节点组的“操作”列单击“弹性伸缩”，进入“弹性伸缩”页面。

**步骤4** 在“弹性伸缩”界面单击 ☒ 开启弹性伸缩，并配置“节点数量范围”为“2 - 4”。

图 8-1 配置弹性伸缩

弹性伸缩

弹性伸缩规则会影响节点数量的使用变化，因此会对实际价格产生影响。当弹性伸缩开启时，MRS服务会依次检测配置的弹性伸缩规则，由第一个满足条件的规则触发弹性伸缩。

弹性伸缩

节点数量范围 默认范围 2 - 4

配置指定时间段的节点数量范围 您还可以添加5个资源计划。

伸缩规则

扩容

添加规则

缩容

添加规则

我同意授权MRS服务根据以上策略自动进行节点扩容/缩容操作。

确定

取消

步骤5 配置资源计划。

1. 在弹性伸缩页面单击默认范围下方的“配置指定时间段的节点数量范围”。
2. 配置“时间范围”和“节点数量范围”。

图 8-2 弹性伸缩

弹性伸缩

弹性伸缩规则会影响节点数量的使用变化，因此会对实际价格产生影响。当弹性伸缩开启时，MRS服务会依次检测配置的弹性伸缩规则，由第一个满足条件的规则触发弹性伸缩。

弹性伸缩

节点数量范围 默认范围 2 - 4

时间范围

节点数量范围

07:00 - 13:00 5 - 8 删除

配置指定时间段的节点数量范围 您还可以添加4个资源计划。

伸缩规则

扩容

添加规则

我同意授权MRS服务根据以上策略自动进行节点扩容/缩容操作。

确定

取消

“时间范围”：“07:00-13:00”

“节点数量范围”：“5-8”

步骤6 配置弹性伸缩规则。

1. 勾选“扩容”。

文档版本 01 (2022-03-01)

版权所有 © 华为技术有限公司

43

2. 单击右侧“添加规则”，进入“添加规则”页面。

图 8-3 添加规则

添加规则

规则名称

default-expand-2

如果

YARNAppRunning

?

大于

75

?

持续

1

个五分钟

?

添加

1

个节点

?

冷却时间

20

分钟

?

确定

取消

“规则名称”：default-expand-2

“如果”：在下拉框中选择规则对象及约束要求，例如YARNAppRunning 大于 75

“持续”：1个五分钟

“添加”：1个节点

“冷却时间”：20分钟

3. 单击“确定”。

**步骤7** 勾选“我同意授权MRS服务根据以上策略自动进行节点扩容/缩容操作。”。

**步骤8** 单击“确定”，完成弹性伸缩集群设置。

----结束

参考信息

在添加规则时，可以参考表8-2配置相应的指标。

说明

混合集群的支持分析集群和流式集群的所有指标。

表 8-2 弹性伸缩指标列表

集群类型	指标名称	数值类型	说明
流式集群	StormSlotAvailabl e	整型	Storm组件的可用slot数。 取值范围为[0 ~ 2147483646]。

集群类型	指标名称	数值类型	说明
	StormSlotAvailablePercentage	百分比	Storm组件可用slot百分比。是可用slot数与总slot数的比值。 取值范围为[0 ~ 100]。
	StormSlotUsed	整型	Storm组件的已用slot数。 取值范围为[0 ~ 2147483646]。
	StormSlotUsedPercentage	百分比	Storm组件已用slot百分比。是已用slot数与总slot数的比值。 取值范围为[0 ~ 100]。
	StormSupervisorMemAverageUsage	整形	Storm组件Supervisor的内存平均使用量。 取值范围为[0 ~ 2147483646]。
	StormSupervisorMemAverageUsagePercentage	百分比	Storm组件Supervisor进程使用的内存占系统总内存的平均百分比。 取值范围[0 ~ 100]。
	StormSupervisorCPUAverageUsagePercentage	百分比	Storm组件Supervisor进程使用的CPU占系统总CPU的平均百分比。 取值范围[0 ~ 6000]。
分析集群	YARNAppPending	整型	YARN组件挂起的任务数。 取值范围为[0 ~ 2147483646]。
	YARNAppPendingRatio	比率	YARN组件挂起的任务数比例。是YARN挂起的任务数与YARN运行中的任务数比值。 取值范围为[0 ~ 2147483646]。
	YARNAppRunning	整型	YARN组件运行中的任务数。 取值范围为[0 ~ 2147483646]。
	YARNContainerAllocated	整型	YARN组件中已分配的container个数。 取值范围为[0 ~ 2147483646]。
	YARNContainerPending	整型	YARN组件挂起的container个数。 取值范围为[0 ~ 2147483646]。
	YARNContainerPendingRatio	比率	YARN组件挂起的container比率。是挂起的container数与运行中的container数的比值。 取值范围为[0 ~ 2147483646]。
	YARNCPUAllocated	整型	YARN组件已分配的虚拟CPU核心数。 取值范围为[0 ~ 2147483646]。



集群类型	指标名称	数值类型	说明
	YARNCPUAvailable	整型	YARN组件可用的虚拟CPU核心数。 取值范围为[0 ~ 2147483646]。
	YARNCPUAvailablePercentage	百分比	YARN组件可用虚拟CPU核心数百分比。 是可用虚拟CPU核心数与总虚拟CPU核心数比值。 取值范围为[0 ~ 100]。
	YARNCPUPending	整型	YARN组件挂起的虚拟CPU核心数。 取值范围为[0 ~ 2147483646]。
	YARNMemoryAllocated	整型	YARN组件已分配内存大小。单位为MB。 取值范围为[0 ~ 2147483646]。
	YARNMemoryAvailable	整型	YARN组件可用内存大小。单位为MB。 取值范围为[0 ~ 2147483646]。
	YARNMemoryAvailablePercentage	百分比	YARN组件可用内存百分比。是YARN组件可用内存大小与YARN组件总内存大小的比值。 取值范围为[0 ~ 100]。
	YARNMemoryPending	整型	YARN组件挂起的内存大小。 取值范围为[0 ~ 2147483646]。

在添加资源计划时，可以参考[表8-3](#)配置相应的参数。

**表 8-3** 资源计划配置项说明

配置项	说明
时间范围	资源计划的起始时间和结束时间，精确到分钟，取值范围[00:00, 23:59]。例如资源计划开始于早上8:00，结束于10:00，则配置为8:00-10:00。结束时间必须晚于开始时间至少30分钟。
节点数量范围	资源计划内的节点数量上下限，取值范围[0,500]，在资源计划时间内，集群Task节点数量小于最小节点数时，弹性伸缩会将集群Task节点一次性扩容到最小节点数。在资源计划时间内，集群Task节点数量大于最大节点数时，弹性伸缩会将集群Task节点一次性缩容到最大节点数。最小节点数必须小于或等于最大节点数。

# 9 配置 Hive 存算分离

MRS支持用户将数据存储在OBS服务中，使用MRS集群仅做数据计算处理的存算分离场景。用户通过IAM服务的“委托”机制进行简单配置，即可实现OBS的访问。

本章节指导用户创建Hive表存放数据到OBS，基本内容如下所示：

1. [创建ECS委托](#)
2. [为MRS集群配置委托](#)
3. [创建OBS文件系统](#)
4. [Hive访问OBS文件系统](#)

## 创建 ECS 委托

1. 登录华为云管理控制台。
2. 在服务列表中选择“管理与监管 > 统一身份认证服务”。
3. 选择“委托 > 创建委托”。
4. 设置“委托名称”。例如：mrs\_ecs\_obs。
5. “委托类型”选择“云服务”，在“云服务”中选择“弹性云服务器ECS 裸金属服务器BMS”，授权ECS或BMS调用OBS服务，如[图9-1](#)所示。
6. “持续时间”选择“永久”并单击“下一步”。

图 9-1 创建委托

\*

委托名称

mrs\_ecs\_obs

\*

委托类型

普通帐号

将帐号内资源的操作权限委托给其他华为云帐号。

云服务

将帐号内资源的操作权限委托给华为云服务。

\*

云服务

弹性云服务器 ECS 裸金属服务器 BMS

\*

持续时间

永久

描述

请输入委托信息。

下一步

取消

7. 在弹出页面中作用范围选择“全局服务”，在权限中搜索“OBS OperateAccess”策略，勾选“OBS OperateAccess”策略如图9-2所示。

图 9-2 配置权限

给 mrs\_ecs\_obs 授权

基于区域授权

基于区域授权是指对指定区域内资源的操作权限

在以下作用范围

全局服务

跨所有区域授权，不需要切换区域。授权后在全局生效。如对象存储服务（OBS），[了解更多](#)。

区域授权

仅对指定区域授权，需要切换至授权区域生效。如弹性云服务器（ECS），[了解更多](#)。

拥有以下权限

无法找到指定策略的权限？

搜索已选(1)

全部策略

所有云服务

OBS OperateAccess

X

Q

名称	描述	类型
<div><div></div>OBS OperateAccess</div>	具有对象存储服务（OBS）策略授权，获取桶元数据、列举桶内对象、查询桶位置、上传对象、获取对象、删除对象、获取对象ACL、系统策略	

10

总数: 1

1

8. 单击“确定”完成委托创建。

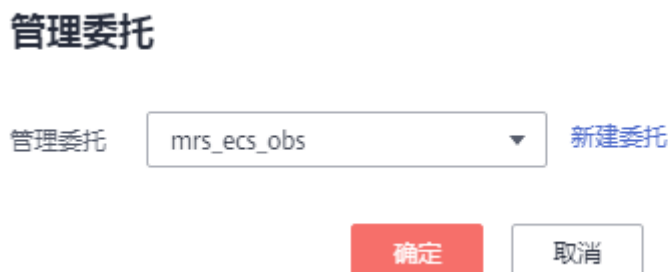
为 MRS 集群配置委托

配置存算分离支持在新建集群中配置委托实现，也可以通过为已有集群绑定委托实现。本示例以为已有集群配置委托为例介绍。

1. 登录MRS控制台，在导航栏选择“集群列表 > 现有集群”。
2. 单击集群名称，进入集群详情页面。
3. 在集群详情页的“概览”页签，单击“IAM用户同步”右侧的“单击同步”进行IAM用户同步。

- 在集群详情页的“概览”页签，单击“委托”右侧的“管理委托”选择[创建ECS委托](#)的委托并单击“确定”进行绑定，或单击“新建委托”进入IAM控制台进行创建后再在此处进行绑定。

图 9-3 绑定委托



## 创建 OBS 文件系统

- 登录OBS控制台。
- 选择“并行文件系统 > 创建并行文件系统”。
- 填写文件系统名称，例如“mrs-demo01”。

其他参数请根据需要填写。

- 单击“立即创建”。
- 在OBS控制台并行文件系统列表中，单击文件系统名称进入详情页面。
- 在左侧导航栏选择“文件”，新建program、input文件夹。
  - program：请上传程序包到该文件夹。
  - input：请上传输入数据到该文件夹。

## Hive 访问 OBS 文件系统

- 用root用户登录集群Master节点，具体请参见[登录集群节点](#)。
- 验证Hive访问OBS。
  - 用root用户登录集群Master节点，执行如下命令：

```
cd /opt/Bigdata/client
source bigdata_env
source Hive/component_env
```

- b. 查看文件系统mrs-demo01下面的文件列表。

```
hadoop fs -ls obs://mrs-demo01/
```

- c. 返回文件列表即表示访问OBS成功。

```
Found 2 items
drwxrwxrwx - hive hive          0 2021-10-22 10:08 obs://mrs-demo01/input
drwxrwxrwx - hive hive          0 2021-10-22 10:08 obs://mrs-demo01/program
```

- d. 执行以下命令进行用户认证（普通模式即未开启Kerberos认证无需执行此步骤）。

```
kinit hive
```

输入用户hive密码，默认密码为Hive@123，第一次使用需要修改密码。

- e. 执行Hive组件的客户端命令。

```
beeline
```

- f. 在beeline中直接使用OBS的目录进行访问。例如，执行如下命令创建Hive表并指定数据存储在mrs-demo01文件系统的test\_demo01目录中。

```
create table test_demo01(name string) location "obs://mrs-demo01/
test_demo01";
```

- g. 执行如下命令查询所有表，返回结果中存在表test\_demo01，即表示访问OBS成功。

```
show tables;
```

```
-----+-----+
| tab_name |
+-----+-----+
| test_demo01 |
+-----+-----+
1 row selected (0.301 seconds)
```

- h. 查看表的Location。

```
show create table test_demo01;
```

查看表的Location是否为“obs://OBS桶名/”开头。

```
serialization.format = ,
STORED AS INPUTFORMAT
'org.apache.hadoop.mapred.TextInputFormat'
OUTPUTFORMAT
'org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat'
LOCATION
'obs://mrs-demo01/test_demo01'
TBLPROPERTIES (
  'bucketing_version'='2',
  'transient_lastDdlTime'='1634872329')
```

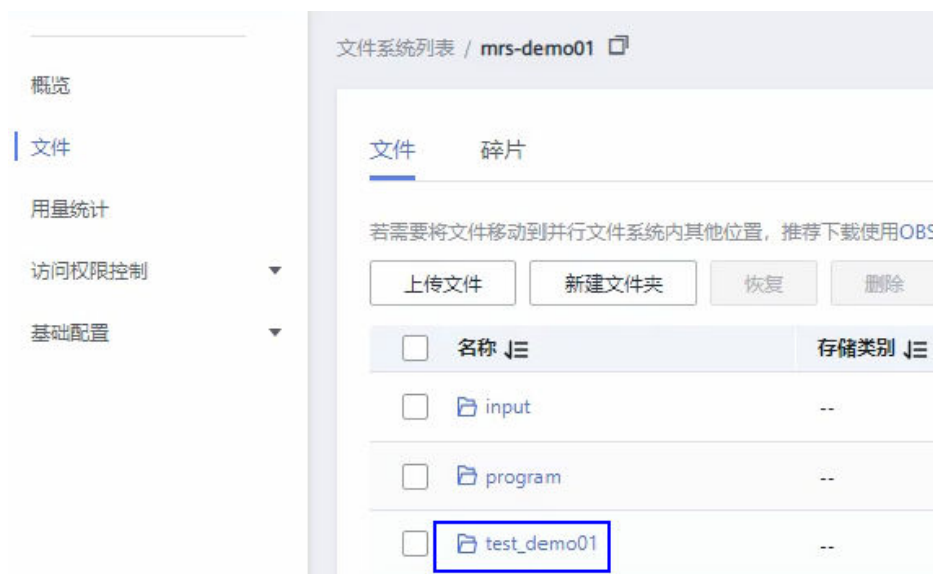
- i. 写入数据。

```
insert into test_demo01 values('mm'),('ww'),('ww');
```

执行select \* from test\_demo01;查询是否写入成功。

```
+-----+
| test_demo01.name |
+-----+
| mm               |
| ww               |
| ww               |
+-----+
```

- j. 执行命令!**q**退出beeline客户端。
- k. 重新登录OBS控制台。
- l. 单击“并行文件系统”，选择创建的文件系统名称。
- m. 单击“文件”，查看是否存在创建的数据。



# 10 提交 Spark 任务到新增 Task 节点

MRS集群可以通过增加Task节点，提升计算能力。集群Task节点主要用于处理数据，不存放持久数据。

本章节指导用户通过租户资源绑定新增的Task节点，并提交Spark任务到新增的Task节点。基本内容如下所示：

1. [添加Task节点](#)
2. [添加资源池](#)
3. [添加租户](#)
4. [队列配置](#)
5. [配置资源分布策略](#)
6. [创建用户](#)
7. [使用spark-submit提交任务](#)
8. [删除Task节点](#)

## 添加 Task 节点

1. 在集群详情页面，选择“节点管理”页签，单击“新增节点组”，进入“新增节点组”页面。
2. 根据需求配置参数。

表 10-1 新增节点组参数说明

参数名称	描述
节点规格	选择节点组内主机的规格类型。
节点数量	设置新增节点组内的节点数量。
系统盘	设置新增节点的系统盘的规格与容量。
数据盘/数据盘数量	设置新增节点的数据盘的规格与容量及数量。
部署角色	添加“NodeManager”角色。

3. 单击“确定”。

## 添加资源池

- 步骤1** 在集群详情页，单击“租户管理”。
- 步骤2** 单击“资源池”页签。
- 步骤3** 单击“添加资源池”。
- 步骤4** 在“添加资源池”设置资源池的属性。
- “名称”：填写资源池的名称，例如“test1”。
  - “资源标签”：填写资源池的标签。例如“1”。
  - “可用主机”：选择[添加Task节点](#)添加的节点。
- 步骤5** 单击“确定”保存。

----结束

## 添加租户

- 步骤1** 在集群详情页，单击“租户管理”。
- 步骤2** 单击“添加租户”，打开添加租户的配置页面，参见以下表格内容为租户配置属性。

表 10-2 租户参数一览表

参数名	描述
名称	例如：tenant_spark
租户类型	选择“叶子租户”。当选中“叶子租户”时表示当前租户为叶子租户，无法再添加子租户。当选中“非叶子租户”时表示当前租户可以再添加子租户。
动态资源	选择“Yarn”，系统将自动在Yarn中以租户名称创建任务队列。动态资源不选择“Yarn”时，系统不会自动创建任务队列。
默认资源池容量 (%)	配置当前租户在“default”资源池中使用的计算资源百分比，例如“20%”。
默认资源池最大容量 (%)	配置当前租户在“default”资源池中使用的最大计算资源百分比，例如“80%”。
储存资源	选择“HDFS”，第一次创建租户时，系统自动在HDFS根目录创建“/tenant”目录。存储资源不选择“HDFS”时，系统不会在HDFS中创建存储目录。
文件/目录数上线	例如：100000000000



参数名	描述
存储空间配额 (MB)	例如：50000，单位为MB。此参数值表示租户可使用的HDFS存储空间上限，不代表一定使用了这么多空间。如果参数值大于HDFS物理磁盘大小，实际最多使用全部的HDFS物理磁盘空间。 <b>说明</b> 为了保证数据的可靠性，HDFS中每保存一个文件则自动生成1个备份文件，即默认共2个副本。HDFS存储空间表示所有副本文件在HDFS中占用的磁盘空间大小总和。例如“存储空间配额”设置为“500”，则实际只能保存约500/2=250MB大小的文件。
存储路径	例如：“tenant/spark_test”，系统默认将自动在“/tenant”目录中以租户名称创建文件夹。例如租户“spark_test”，默认HDFS存储目录为“tenant/spark_test”。第一次创建租户时，系统自动在HDFS根目录创建“/tenant”目录。支持自定义存储路径。
服务	配置当前租户关联使用的其他服务资源，支持HBase。单击“关联服务”，在“服务”选择“HBase”。在“关联类型”选择“独占”表示独占服务资源，选择“共享”表示共享服务资源。
描述	配置当前租户的描述信息。

**步骤3** 单击“确定”保存，完成租户添加。

保存配置需要等待一段时间，界面右上角弹出提示“租户创建成功。”，租户成功添加。

#### 说明

- 创建租户时将自动创建租户对应的角色、计算资源和存储资源。
- 新角色包含计算资源和存储资源的权限。此角色及其权限由系统自动控制，不支持通过“角色管理”进行手动管理。
- 使用此租户时，请创建一个系统用户，并分配Manager\_tenant角色以及租户对应的角色。

----结束


## 队列配置

**步骤1** 在集群详情页，单击“租户管理”。

**步骤2** 单击“队列配置”页签。

**步骤3** 在租户队列表格，指定租户队列的“操作”列，单击“修改”。

#### 说明

- 在“租户管理”页签左侧租户列表，单击目标的租户，切换到“资源”页签，单击 也能打开修改队列配置页面。
- 一个队列只能绑定一个非default资源池。

默认资源标签选择[添加资源池](#)时填写的标签，其他参数请根据实际情况填写。

**步骤4** 单击“确定”。

----结束

## 配置资源分布策略

**步骤1** 在集群详情页，单击“租户管理”。

**步骤2** 单击“资源分布策略”页签，资源池选择[添加资源池](#)创建的资源池。

**步骤3** 在租户tenant\_spark“操作”列，单击“修改”。

- 权重：20
- 最小资源：20
- 最大资源：80
- 预留资源：10

**步骤4** 单击“确定”。

----结束

## 创建用户

**步骤1** 登录FusionInsight Manager管理界面，具体请参考[访问FusionInsight Manager](#)。

**步骤2** 选择“系统 > 权限 > 用户”，单击“添加用户”。

- 用户名：spark\_test
- 用户类型：人机
- 用户组：hadoop、hive
- 主组：hadoop
- 角色：tenant\_spark

**步骤3** 单击“确定”，完成用户添加。

----结束

## 使用 spark-submit 提交任务

1. 使用root用户登录客户端节点，执行如下命令：

```
cd 客户端安装目录
```

```
source bigdata_env
```

```
source Spark2x/component_env
```

安全集群（开启kerberos认证）执行命令**kinit spark\_test**，普通模式（关闭kerberos认证）无需执行。

输入密码，完成认证（第一次登录需要修改密码）。

```
cd Spark2x/spark/bin
```

```
sh spark-submit --queue tenant_spark --class
```

```
org.apache.spark.examples.SparkPi --master yarn-client ../examples/jars/  
spark-examples_2.11-2.4.5-hw-ei-310012.jar
```

## 删除 Task 节点

1. 在集群详情页面，单击“节点管理”页签。
2. 在Task节点组的“操作”列，选择“缩容”。
3. 缩容类型选择“指定节点缩容”，勾选需要缩容的节点。

### 说明

缩容指定的节点需要该节点处理关机状态。

4. 勾选“我已经阅读上述信息，并了解具体影响。”，并单击“确定”。