華中科技大學

大数据处理实验任务书

实验二：HBase 的基本操作

计算机科学与技术学院

2022 年 3 月 24 日

一：实验目的

1、了解 HBase 的用途

2、掌握 HBase 的基本命令

二：实验要求

1、第四节中的实验内容要附上完整的实验过程截图以及必要的文字说明，每个人的 IP 地址等不同，不能直接套用样例的截图。

2、请同学们在完成报告后，将报告的 word 版本命名为：姓名_学号_实验二.docx，并在第四次实验课之前,发到邮箱:

aaaaltaaaa@126.com(石老师班)

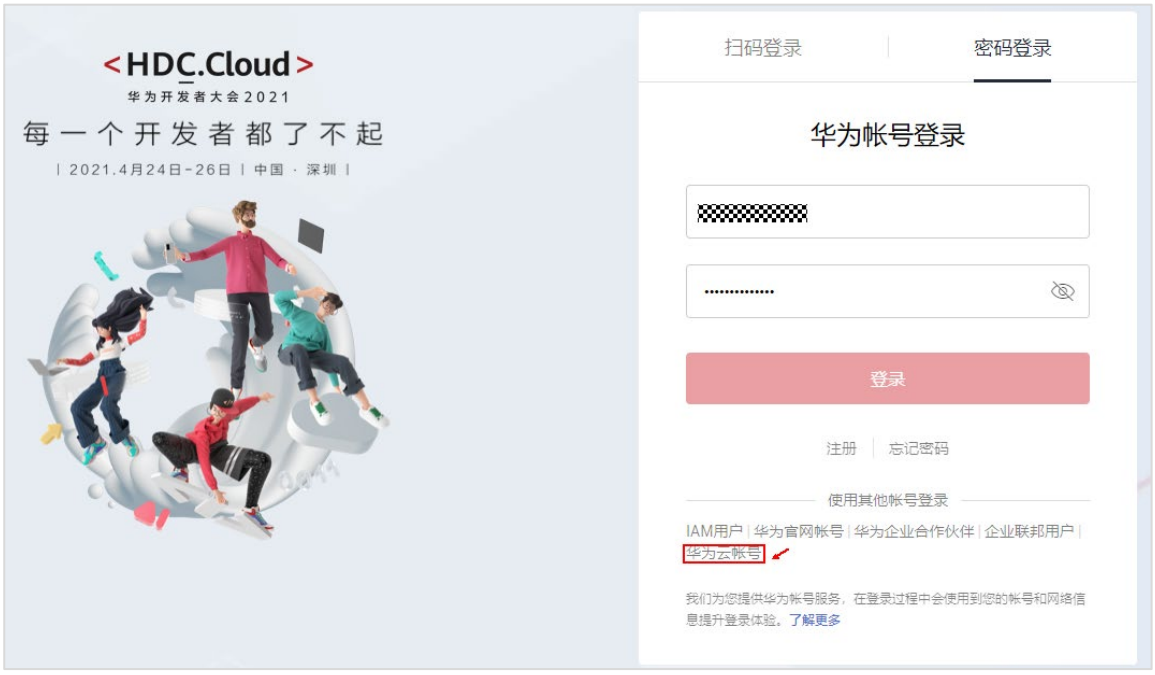
970623990@qq.com (郑老师班)

三：实验环境配置

步骤 1　登录华为云网站

https://www.huaweicloud.com

点击右上角登录，输入账号和密码

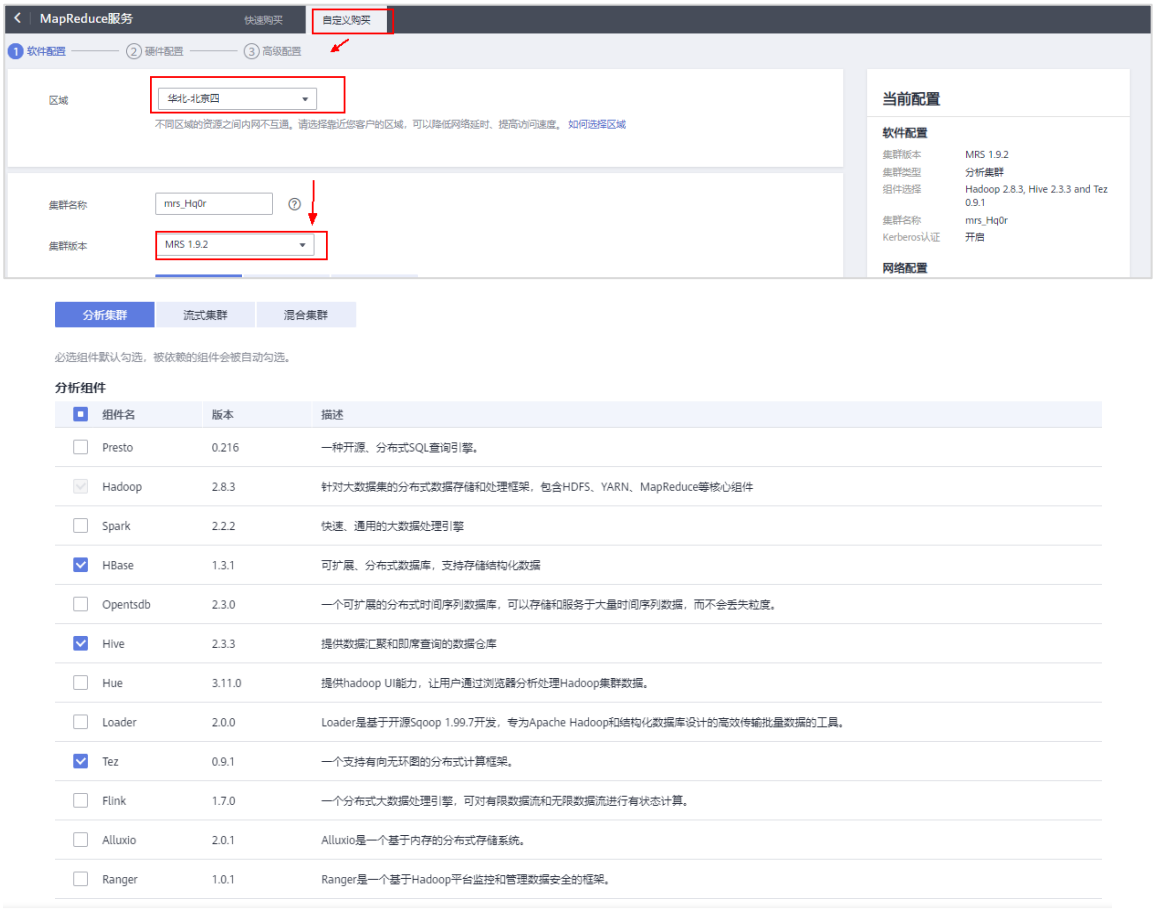


注意：华为云已统一登录入口，若仍不能登录则点击下方华为云账号进行登录。

步骤 2　点击“EI 企业智能”选择“MapReduce 服务”

步骤 3　点击“立即购买”

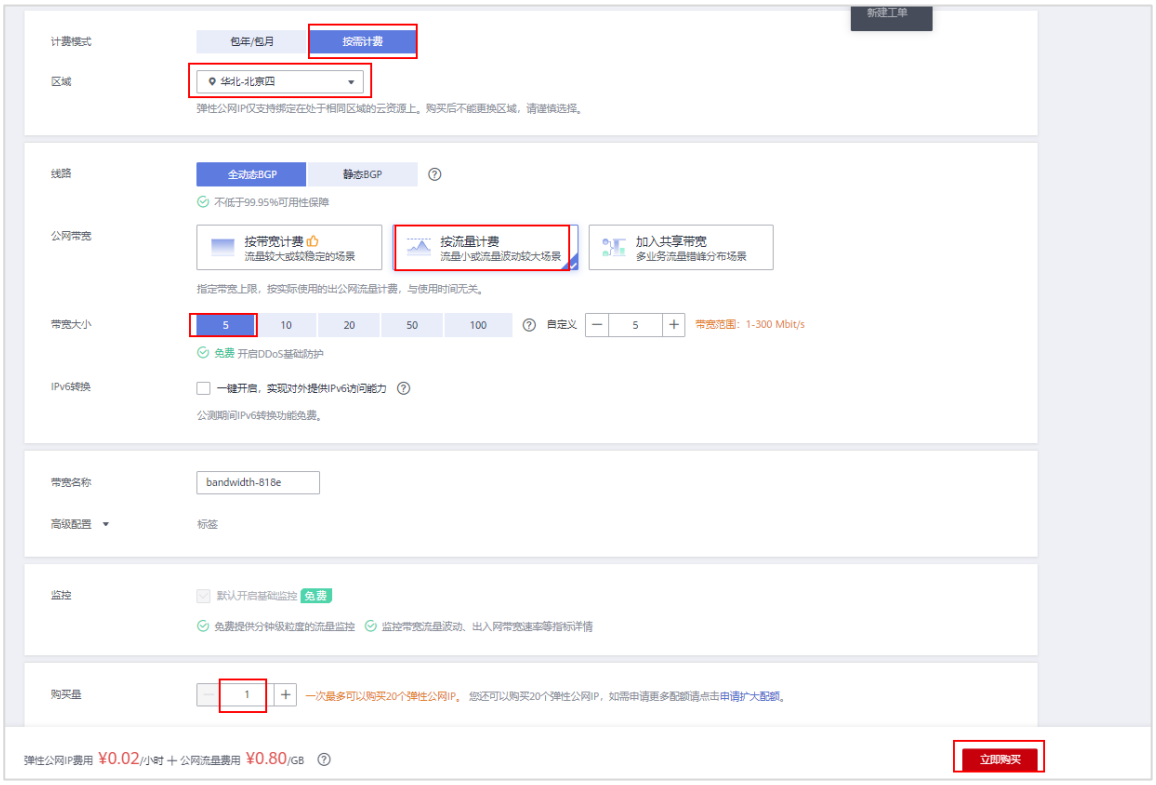
选择“自定义购买”



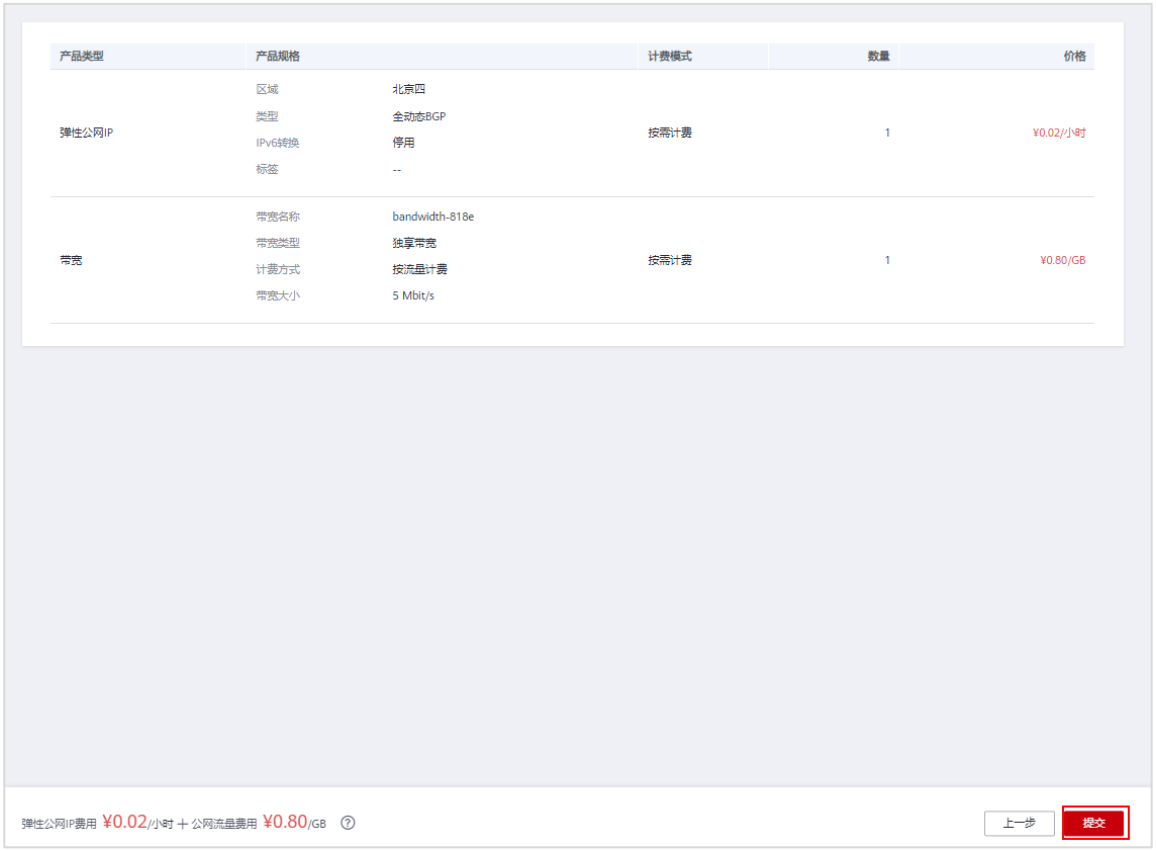
点击下一步，进入硬件配置

选择“按需计费”，“可用区 2”，点击“弹性公网 IP”，如下图：

点击“购买弹性公网 IP”，选择“按需计费”，“按流量计费”，“5M”,点击“立即购买”，如下图：



点击“提交”，如下图：



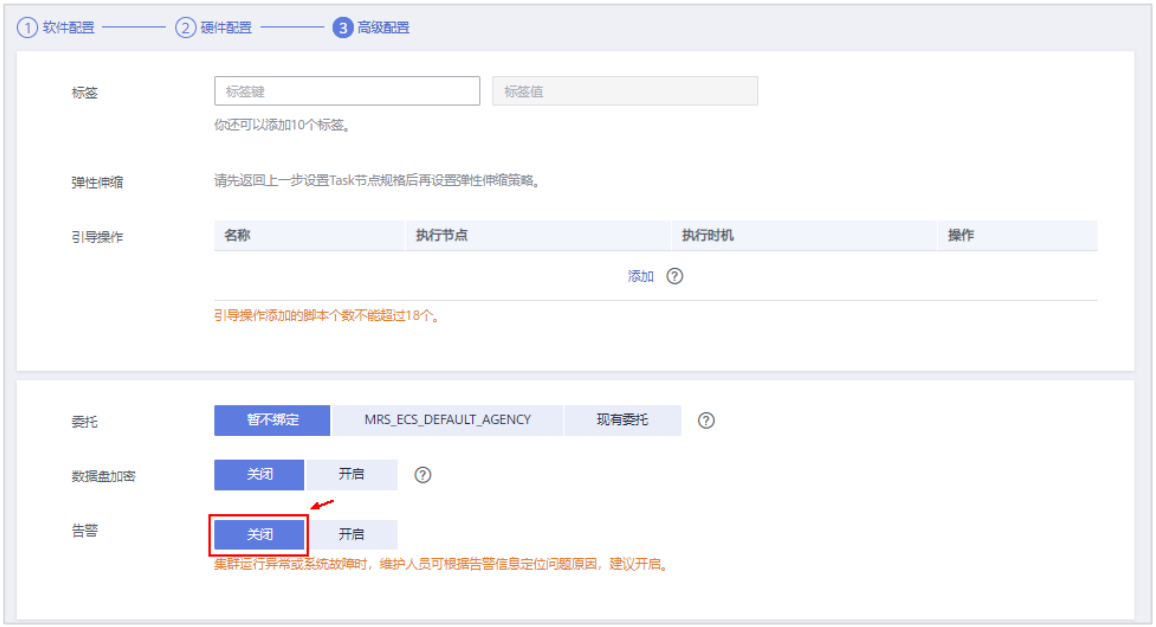
购买成功，如下图：

返回 MapReduce 服务自定义购买界面绑定 EIP

选择“鲲鹏计算”，关闭高可用，调整 core 节点数为 1,如下图：

点击“下一步”

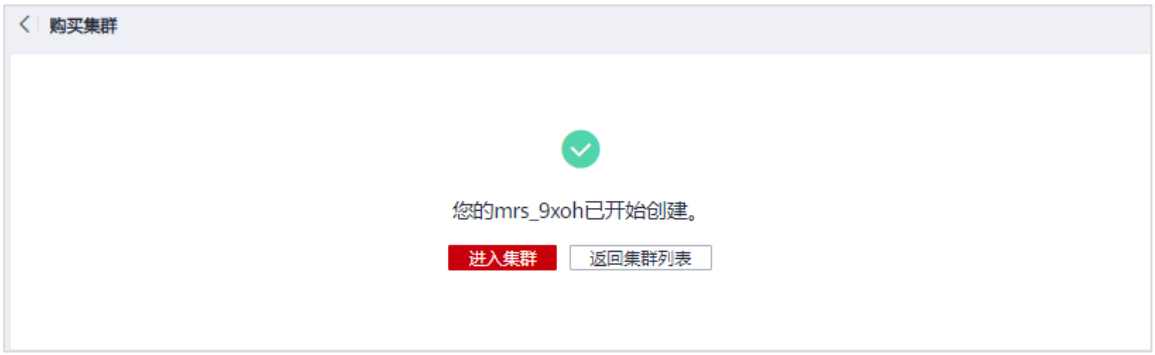
高级配置项参考如下：



点击“确认授权”

点击“立即购买”

步骤 4 点击“返回集群列表”，如下图：



创建过程需要等待几分钟，待状态变为“运行中”集群创建完成

步骤 5 点击集群名称

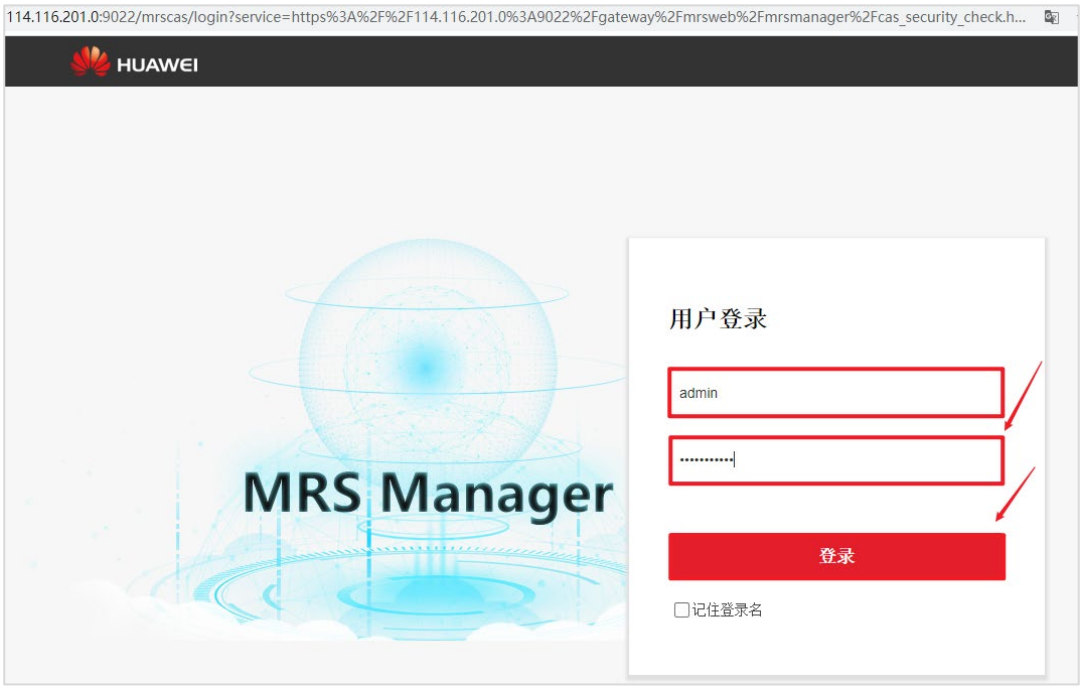
步骤 6　点击“前往 Manager”

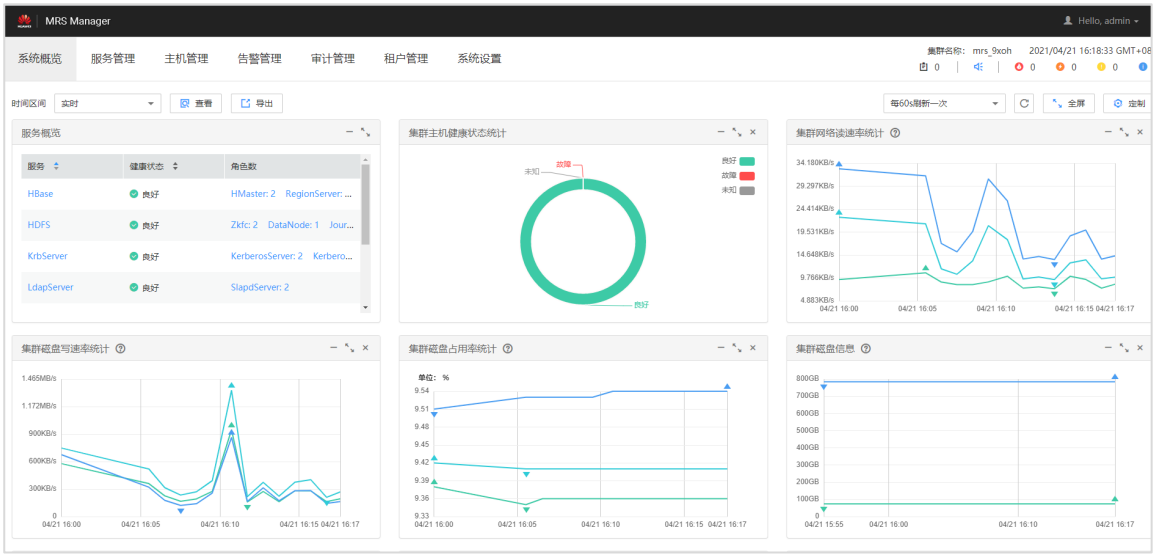
参照下图进入 Manager

点击“高级”,如下图:

点击“继续向前”，如下图：

输入用户名 admin 及密码,点击“登录”，进入 MRS Manager



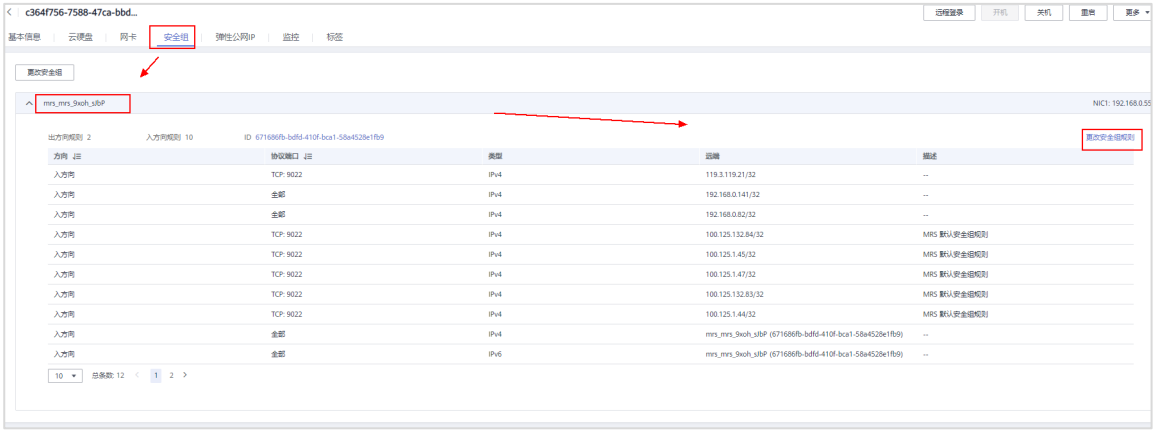


步骤 7　配置安全组

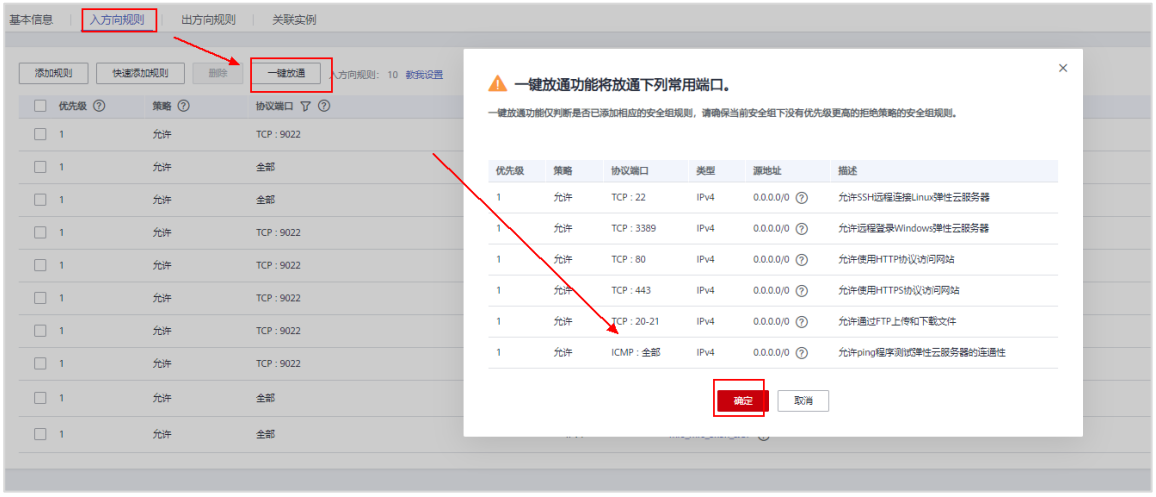
点击集群名称

选择“节点管理”，点击含有“master1”的节点

在弹出页面中选择“安全组”，点击“更改安全组规则”，如下图所示：



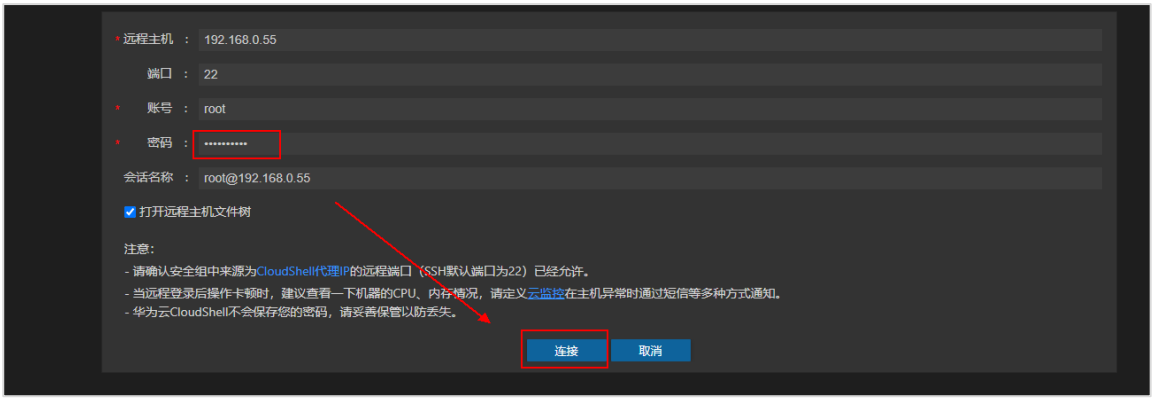
选择“入方向规则”，点击“一键放通”，确认即可。

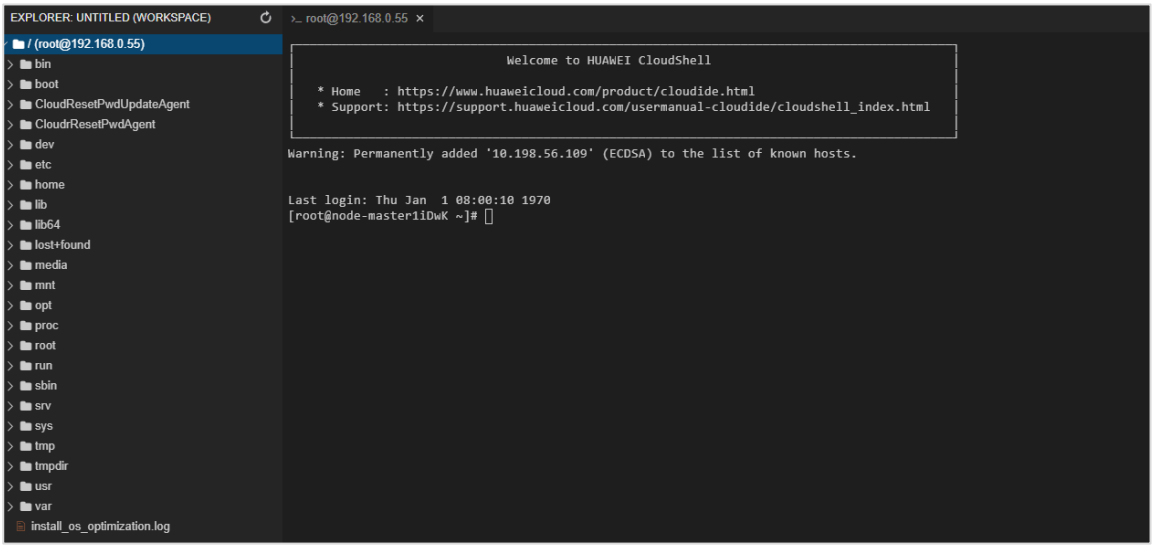


步骤 8　远程登录 master 节点

在安全组配置项，点击右上方“远程登录”，选择 cloudshell 登录。

输入密码，点击连接即可。



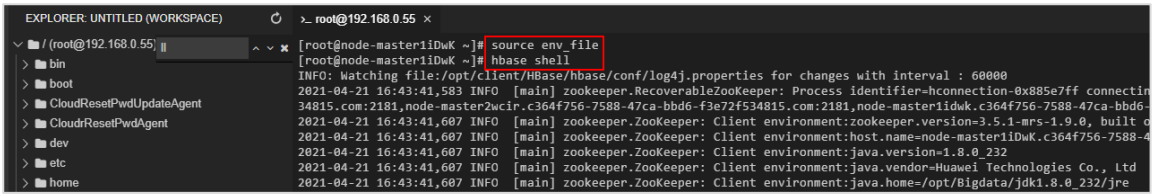


步骤 9 设置环境变量

执行命令：

source env_file

hbase shell



四：实验内容及步骤、实验的详细记录、实验结果分析

请附上实验过程截图（截图需包含指令）以及必要的文字分析

4.1 准备数据(20’)

4.1.1 进入 hbase shell(5’)

4.1.2 创建一个表，表名为学号，列族名为 cf1（create）(5’)

4.1.3 显示所有的表（list）(5’)

4.1.4 向表中增加两行数据（put）(5’)

行键	列族	列名	单元格的值
20200001	cf1	name	tom
20200001	cf1	gender	male
20200001	cf1	age	20
20200002	cf1	name	hanmeimei
20200002	cf1	gender	female
20200002	cf1	age	19

4.2 查询数据(30’)

4.2.1 查找表中，列族名为 cf1 的数据（scan）(3’)

4.2.2 查找表中，列族名为 cf1,列名为 name 的数据（scan）(3’)

4.2.3 查找表中，行键为 20200001 的行（get）(3’)

4.2.4 查找表中，行键为 20200001，列族为 cf1，列名为 name 的数据（get）(3’)

4.2.5 查看起始行键为 20200001，终止行键为 20200002(不包括)，限制长度为 2 的数据(scan)(3’)

4.2.6 查看有数据值为 20 的行(scan)(3’)

4.2.7 查看有数据值为 tom 的行(scan)(3’)

4.2.8 查看列名为 gender 的列(scan)(3’)

4.2.9 查看列名为 name，值为 hanmeimei 的行(scan)(3’)

4.2.10 查看表的属性（desc）(3’)

4.3 修改数据(20’)

4.3.1 改变表的 VERSIONS 为 5 以显示更多的历史版本(alter)(3’)

4.3.2 添加行键 20200001，列族 cf1，列名 name 的多个历史版本（put）(3’)

行键	列族	列名	单元格的值
20200001	cf1	name	LiSi
20200001	cf1	name	ZhangSan'
20200002	cf1	name	WangWu

4.3.3 查看所有行键为 20200001，列簇为 cf1 的多版本数据（get）(3’)

4.3.4 删除行键为 20200002，列名为 age，的数据（delete）(3’)

4.3.5 删除行键为 20200002 的行（deleteall）(4’)

4.3.6 删除整个表（disable，drop）(4’)

4.4 Region 初探(20’)

HBase 默认建表时只有一个 region，这个 region 的 rowkey 是没有边界的，即没有 startkey，也没有 endkey。在数据写入时，所有数据都会写入这个默认的 region，随着数据量的不断增加，此 region 已经不能承受不断增长的数据量，会进行 split，分成 2 个 region。在此过程中，会产生两个问题：

1.数据往一个 region 上写,会有写热点问题。

2.region split 会消耗宝贵的集群 I/O 资源。

基于此我们可以在建表的时候，创建多个空 region，并确定每个 region 的起始和终止

rowky，这样只要我们的 rowkey 设计能均匀的命中各个 region，就不会存在写热点问题，自然 split 的几率也会大大降低。hbase 提供了两种 pre-split 算法：HexStringSplit 和 UniformSplit，前者适用于十六进制字符的 rowkey，后者适用于随机字节数组的 rowkey。以 rowkey 切分，随机分为 4 个 region。

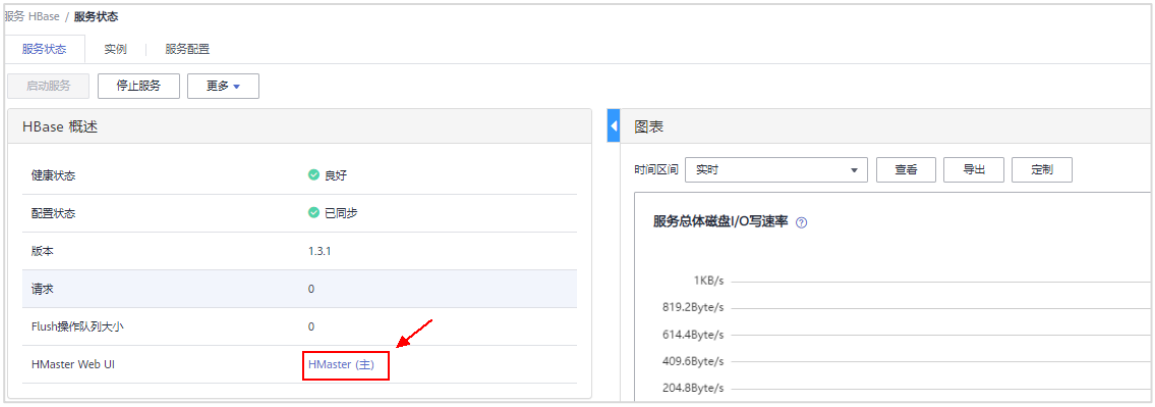
4.4.1 创建具有四个 region 的表，表名为”学号_uniform”，pre-split 算法选择 UniformSplit（create）(5’)

4.4.2 创建具有四个 region 的表，表名为”学号_num”指定 region 以行键 10000000,20000000,30000000 划分（create）(5’)

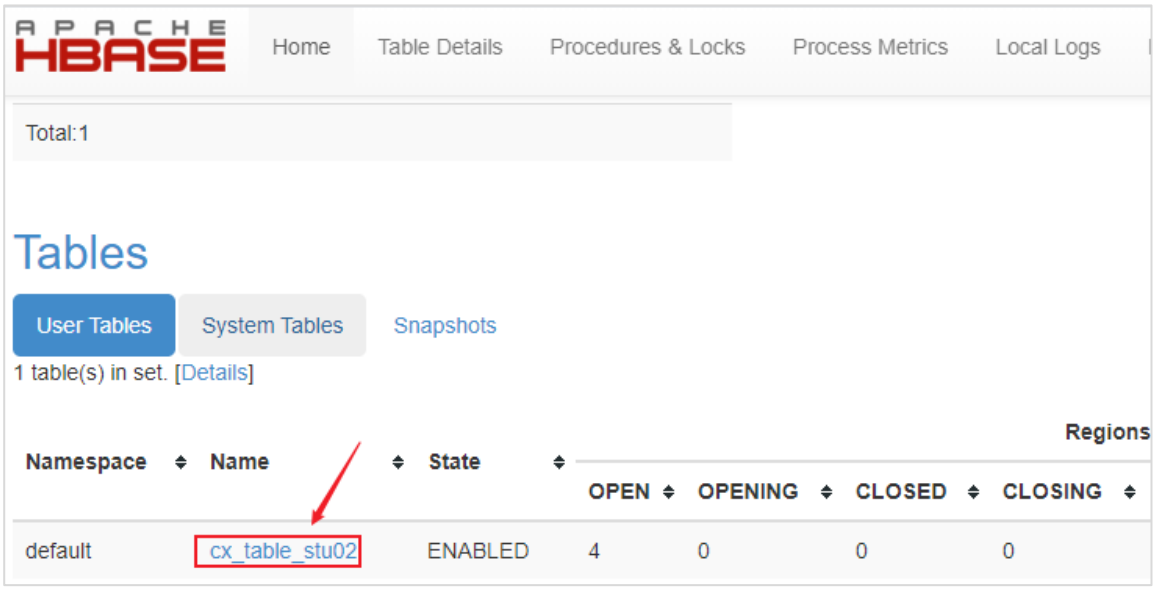
4.4.3 在 Manager 中查看 HBase(5’)

MRS Manager 界面，点击“HBase”服务

点击 HMaster(主)进入 HBase UI



“User Tables”下点击创建好的表名“cx_table_stu02”，如“下图”：



查看该表分区情况，如下图：

Table Regions

Name	Region Server	Start Key	End Key	Locality	Requests
cx_table_stu03,,1638685607047.3c187187dcb7a16281bcaa9da2244a8b.	node-ana-coreVePS.mrs-ntz2.com,16020,1638672259344		@\x00\x00\x00\x00\x00\x00\x00	0.0	1
cx_table_stu03,@\x00\x00\x00\x00\x00\x00\x00,1638685607047.b357e5ad15128dc51c9b67a3f1fc1b97.	node-ana-coreVePS.mrs-ntz2.com,16020,1638672259344	@\x00\x00\x00\x00\x00\x00\x00	\x80\x00\x00\x00\x00\x00\x00\x00	0.0	1
cx_table_stu03,\x80\x00\x00\x00\x00\x00\x00\x00,1638685607047.5caf2e6650e019e6357aec2100f2a7d9.	node-ana-coreVePS.mrs-ntz2.com,16020,1638672259344	\x80\x00\x00\x00\x00\x00\x00\x00	\xC0\x00\x00\x00\x00\x00\x00\x00	0.0	0
cx_table_stu03,\xC0\x00\x00\x00\x00\x00\x00\x00,1638685607047.a4f5082728ce616507907c8bd13211ad.	node-ana-coreVePS.mrs-ntz2.com,16020,1638672259344	\xC0\x00\x00\x00\x00\x00\x00\x00		0.0	0

4.4.4 根据两个表的 End key 和 Start Key，选择适当的行键往两个表的不同 region 中添加任意两个数据，使得每个表至少有两个不同 region 中 Requests 不为 0（put）(5’)

4.4.5 删除所有表(5’)

4.5 hive 初探(10’)

4.5.1 准备 file1.txt，内容为”hello hust”，file2.txt，内容为”hello 学号”（vim）(3’)

4.5.2 将创建的文件移动到 HDFS 中/test 文件夹内（见上次实验）(3’)

4.5.3 在 hive 中创建表，tablename 替换为学号(1’)

```
create table tablename(line string);
```

4.5.4 加载 hdfs 中的数据到 hive 中(1’)

```
load data inpath 'hdfs:///test' overwrite into table tablename;
```

4.5.5 通过 HiveQL 语句创建词频统计表(1’)

```
create table word_count as
select word,count(1) as count from
    (select explode(split(line, ' '))as word from tablename) w
    group by word
    order by word;
```

4.5.6 通过 HiveQL 语句创建词频统计表(1’)

```
select * from word_count;
```

五：实验总结（10’）

附录：

1 容易出错的地方:多打或少打空格；英文引号打成成文分号。

2 HBase 基本命令:http://c.biancheng.net/view/3587.html

3 牢记删除资源