

**大数据处理实验报告**

实验二：HBase的基本操作

姓 名：刘日星

学 号：X2020I1007

院 系：计算机科学与技术

专 业：计算机科学&金融

年 级：CS1804 交换生

指导教师：石宣化

2021年 12 月 8 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pdf版本命名为：大数据实验二+姓名+学号.pdf，并在这周五前,发到邮箱:

[aaaaltaaaa@126.com(石老师班)](mailto:aaaaltaaaa@126.com(石老师班)，)

798792873@qq.com(郑老师班)

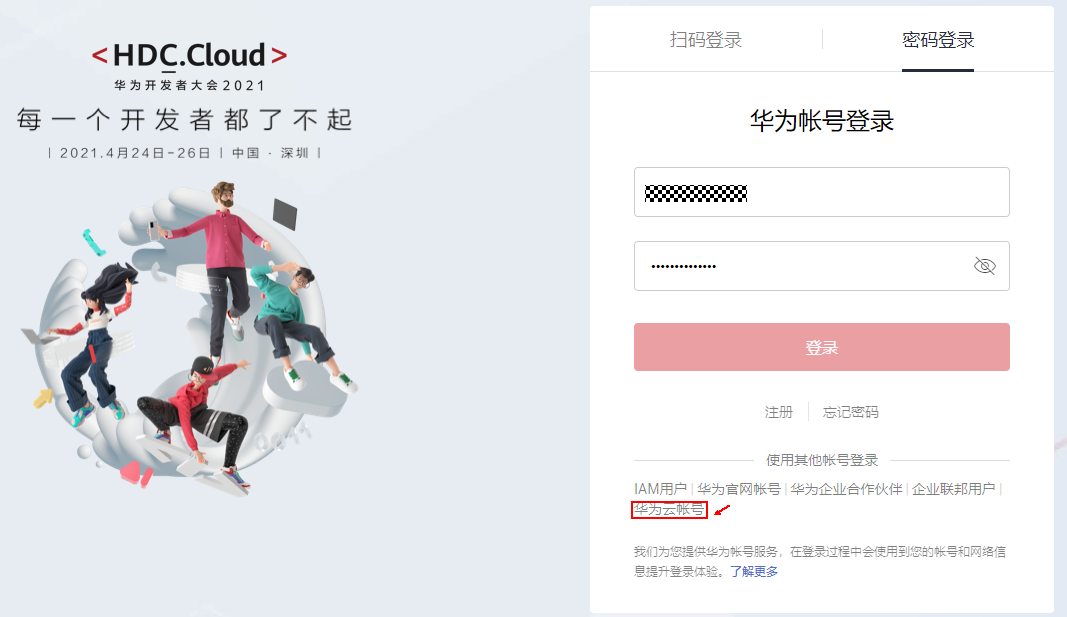
三：实验环境配置

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。

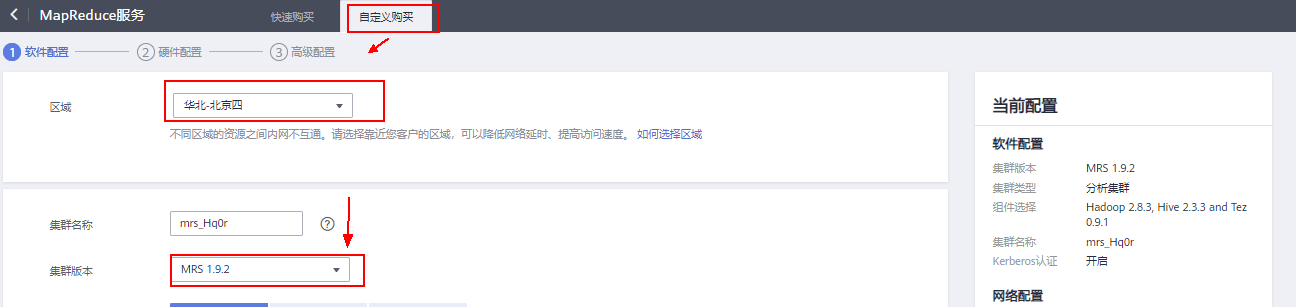
点击“EI企业智能”选择“MapReduce服务”



点击“立即购买”



选择“自定义购买”





点击下一步，进入硬件配置

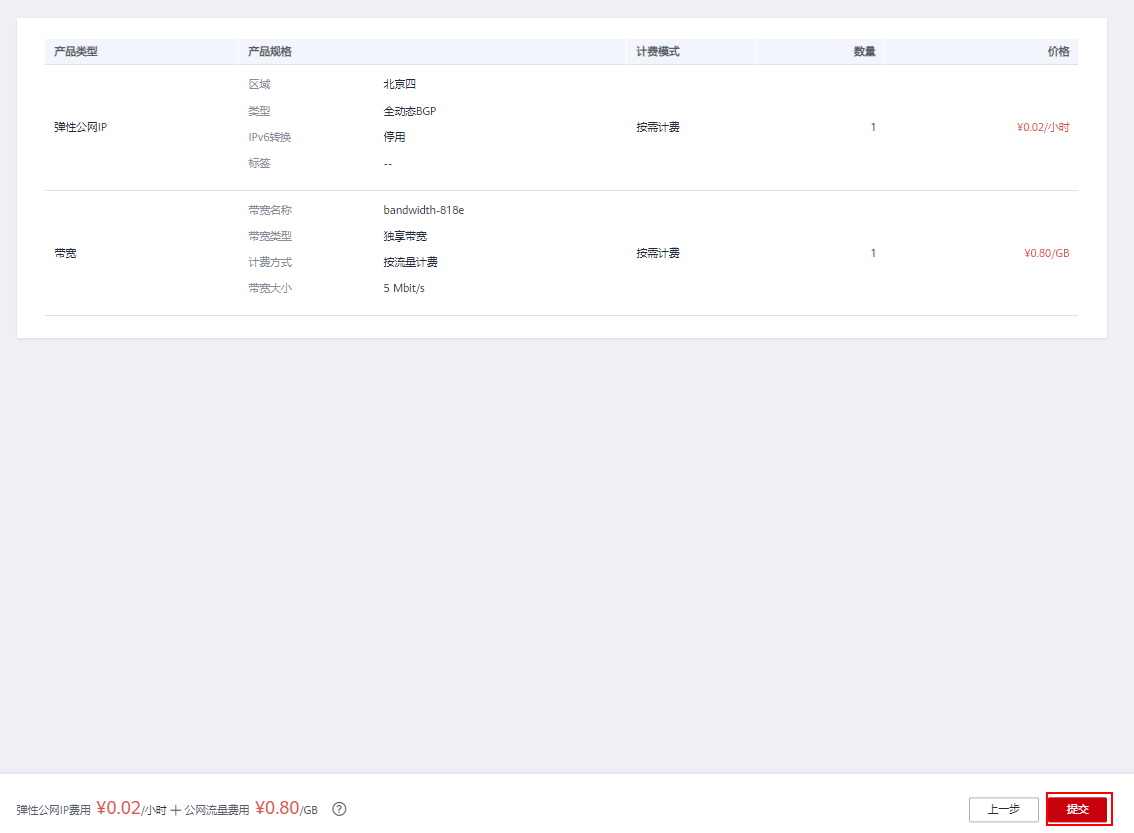
选择“按需计费”，“可用区2”，点击“弹性公网IP”，如下图：



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：



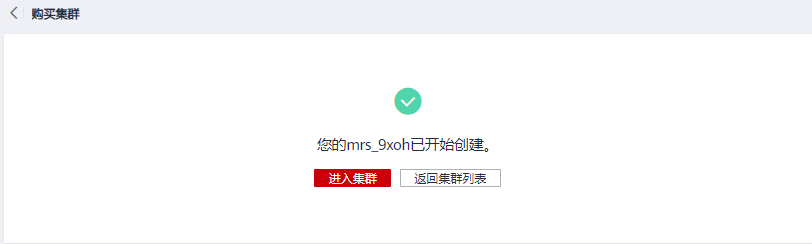


点击“确认授权”

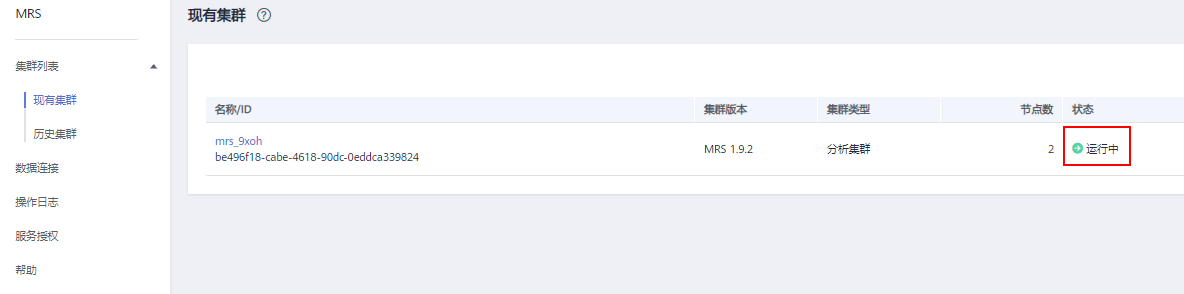


点击“立即购买”

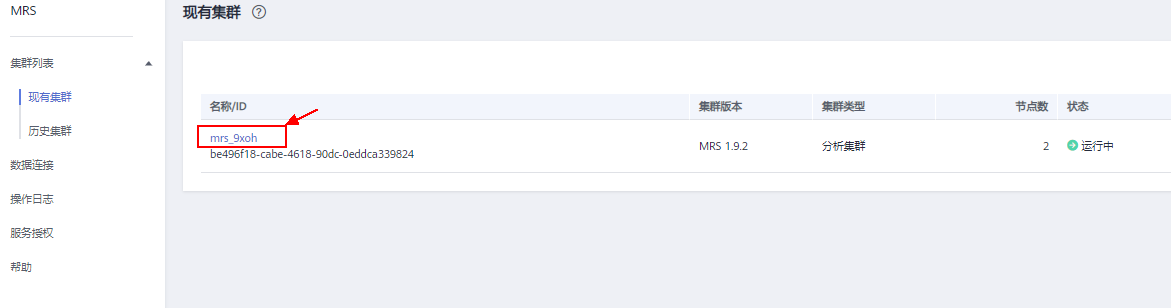
点击“返回集群列表”，如下图：



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



点击集群名称



点击“前往Manager”



参照下图进入Manager



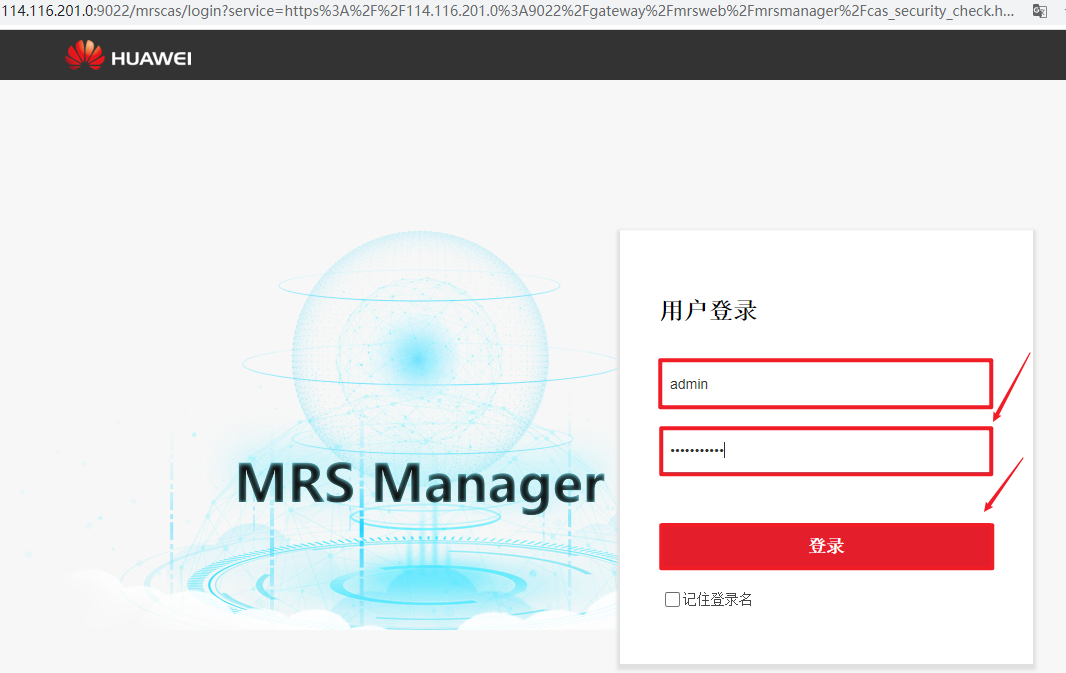
点击“高级”,如下图：

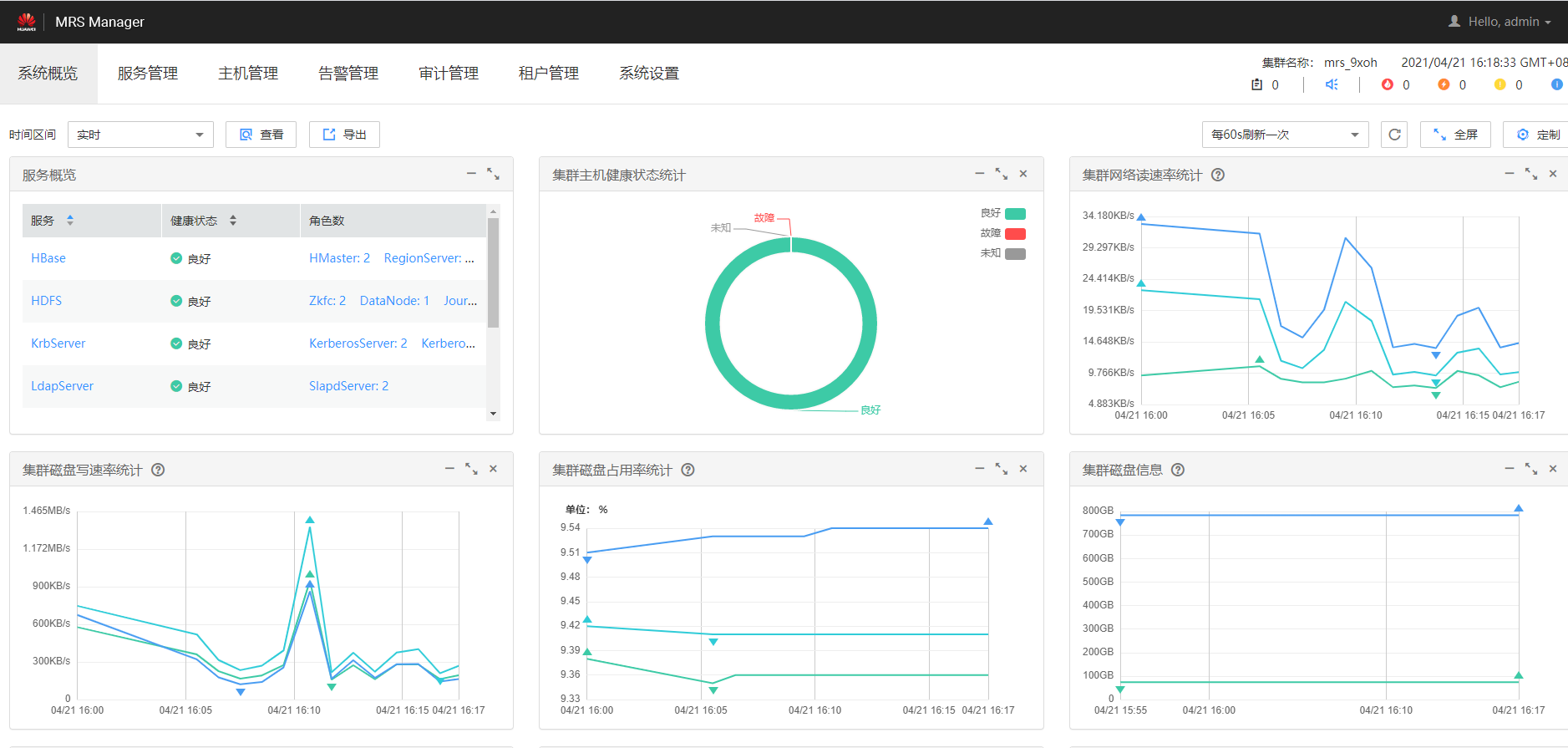


点击“继续向前”，如下图：



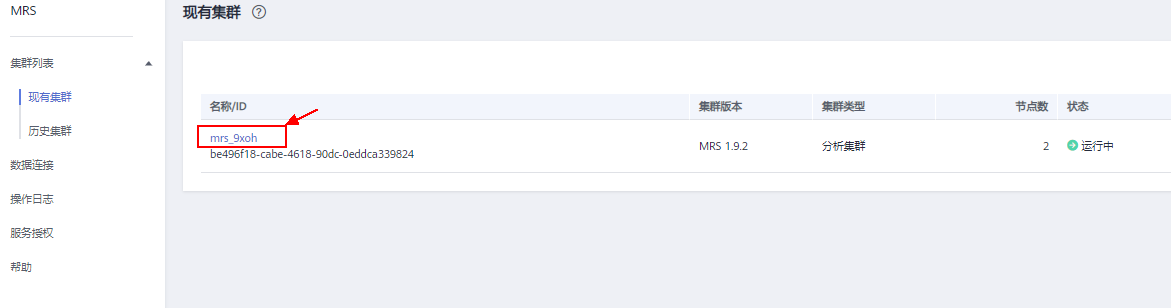
输入用户名admin及密码,点击“登录”，进入MRS Manager





配置安全组

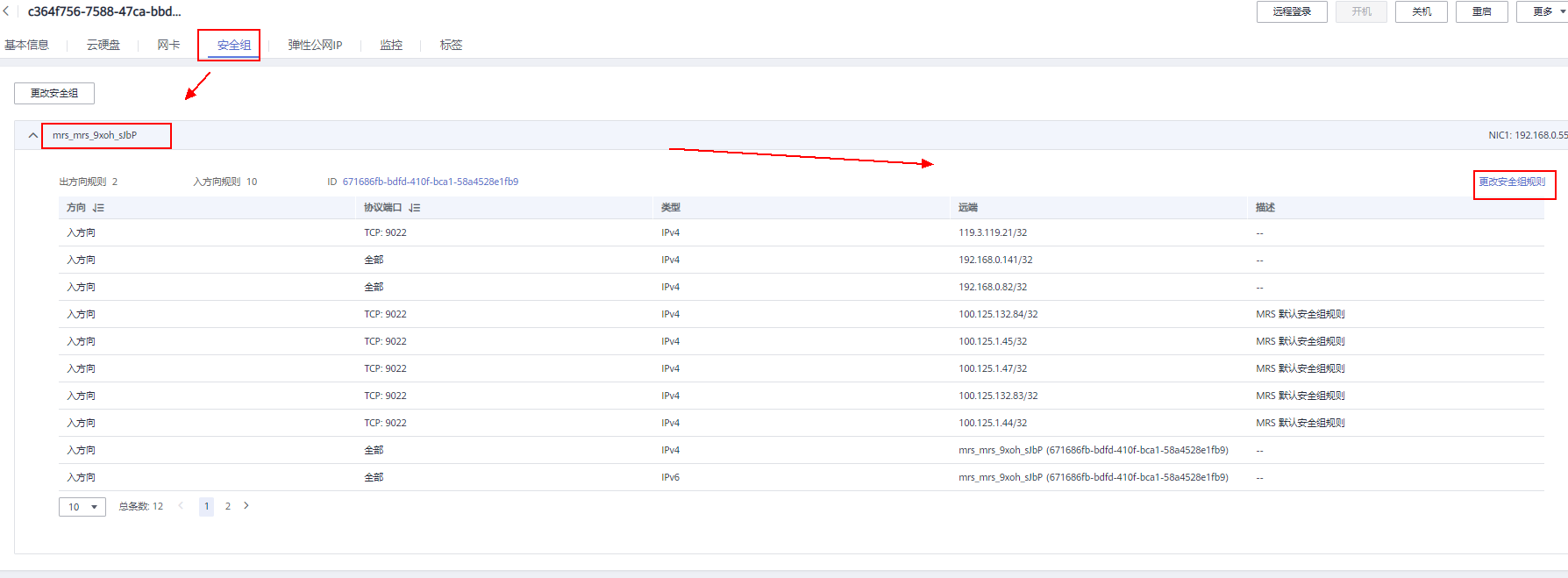
点击集群名称



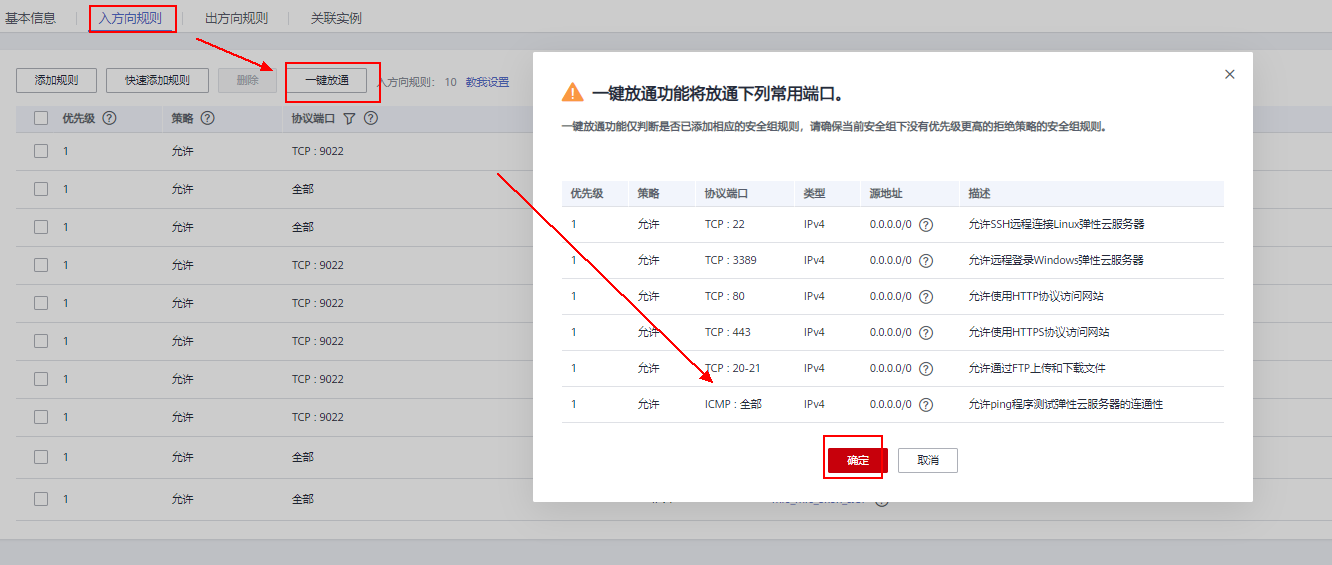
选择“节点管理”，点击含有“master1”的节点



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：

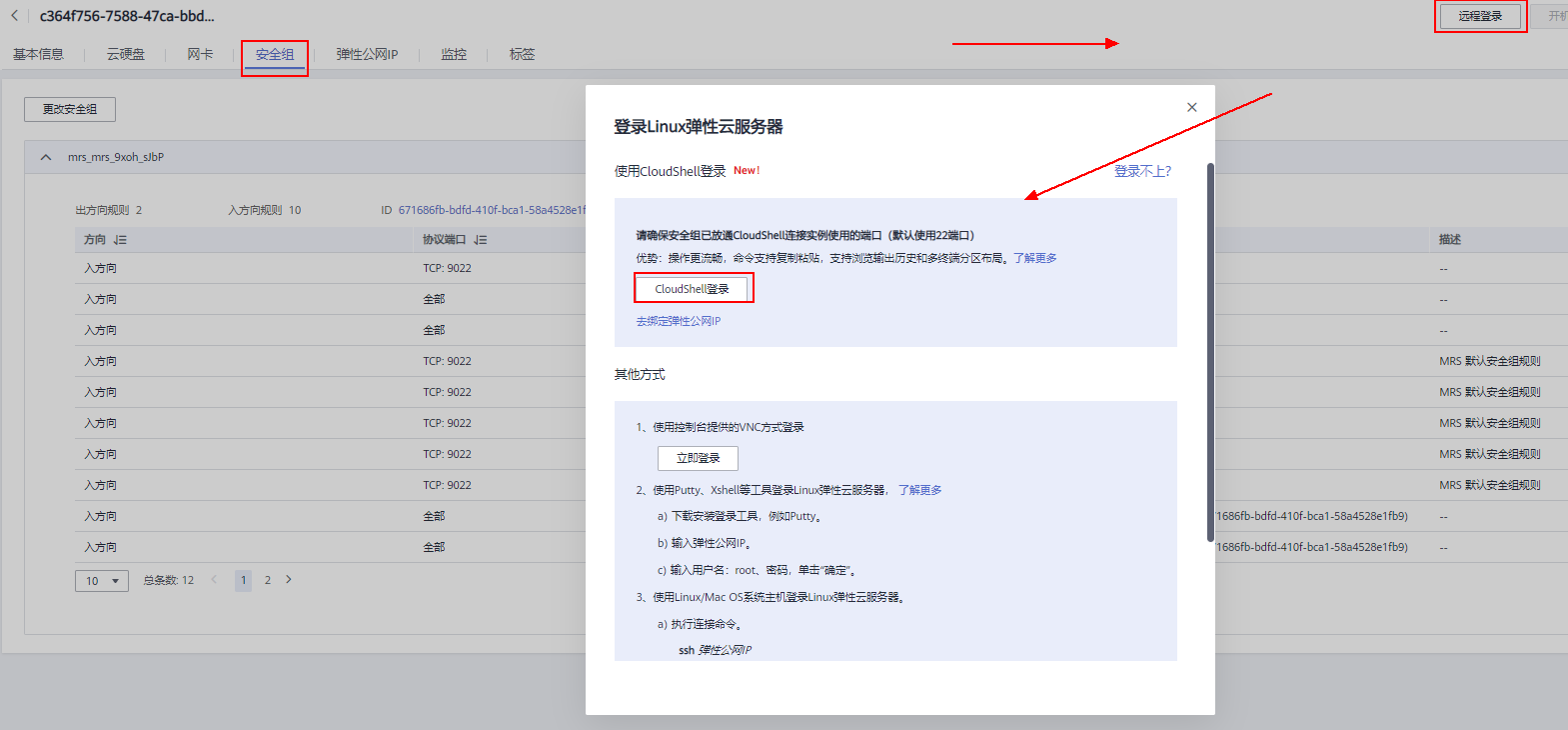


选择“入方向规则”，点击“一键放通”，确认即可。



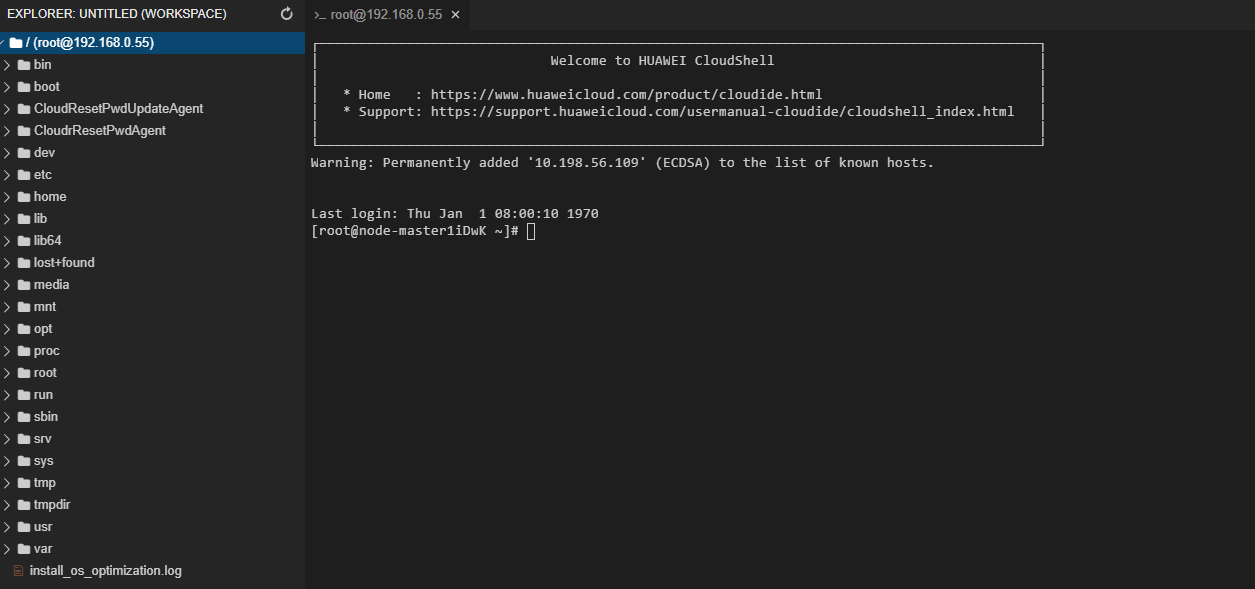
远程登录master节点

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。



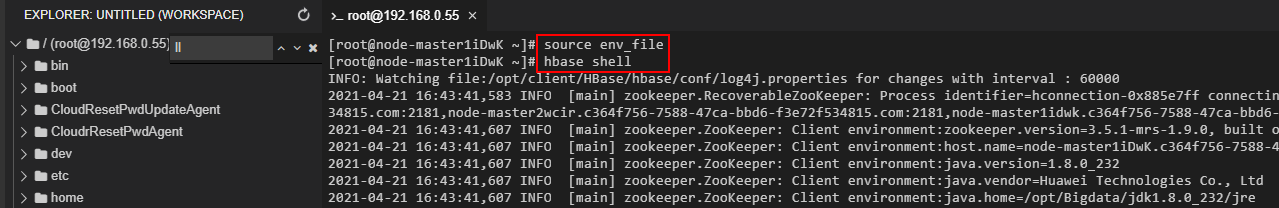


设置环境变量

执行命令：

# source env\_file

# hbase shell

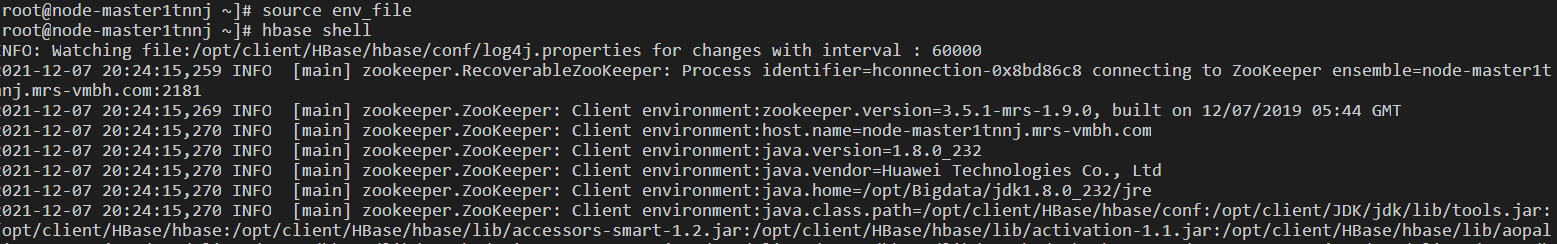


四：实验内容及步骤、实验的详细记录、实验结果分析

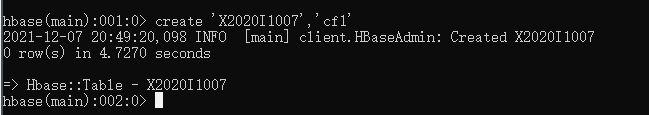
**请附上实验过程截图（截图需包含指令）以及必要的文字分析**

## 准备数据(20’)

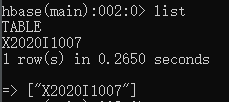
### 进入hbase shell(5’)



### 创建一个表，表名为学号，列族名为cf1（create）(5’)

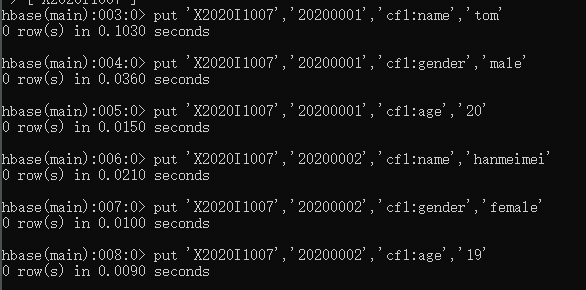


### 显示所有的表（list）(5’)

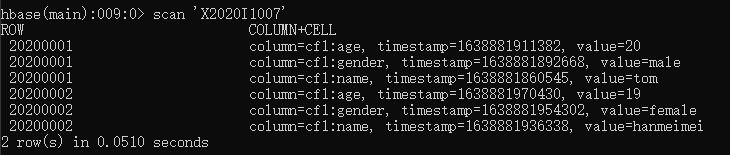


### 向表中增加两行数据（put）(5’)

|  |  |  |  |
| --- | --- | --- | --- |
| **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** |



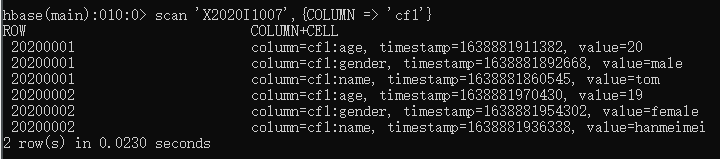
输入表格内容4.1.1



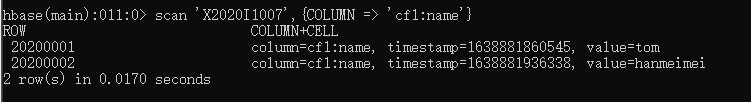
显示表格内容4.1.1(1)

## 查询数据(30’)

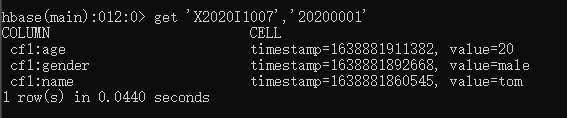
### 查找表中，列族名为cf1的数据（scan）(3’)



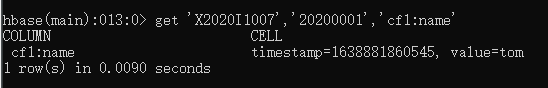
### 查找表中，列族名为cf1,列名为name的数据（scan）(3’)



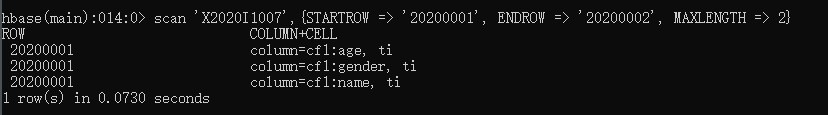
### 查找表中，行键为20200001的行（get）(3’)



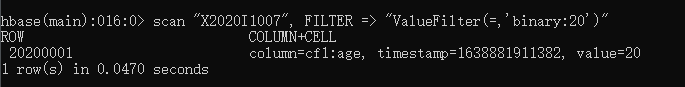
### 查找表中，行键为20200001，列族为cf1，列名为name的数据（get）(3’)



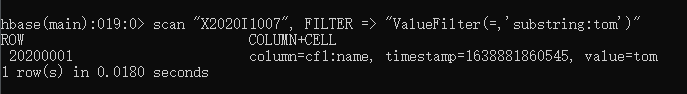
### 查看起始行键为20200001，终止行键为20200002(不包括)，限制长度为2的数据(scan)(3’)



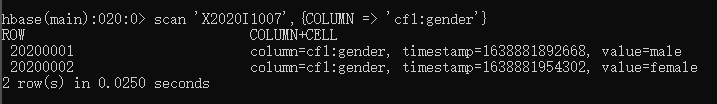
### 查看有数据值为20的行(scan)(3’)



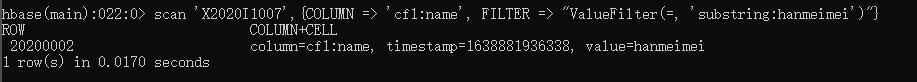
### 查看有数据值为tom的行(scan)(3’)



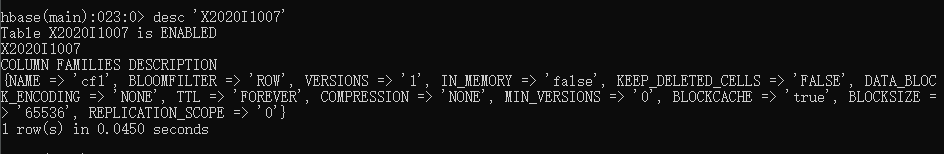
### 查看列名为gender的列(scan)(3’)



### 查看列名为name，值为hanmeimei的行(scan)(3’)

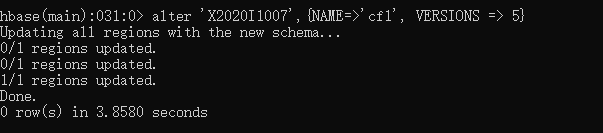


### 查看表的属性（desc）(3’)



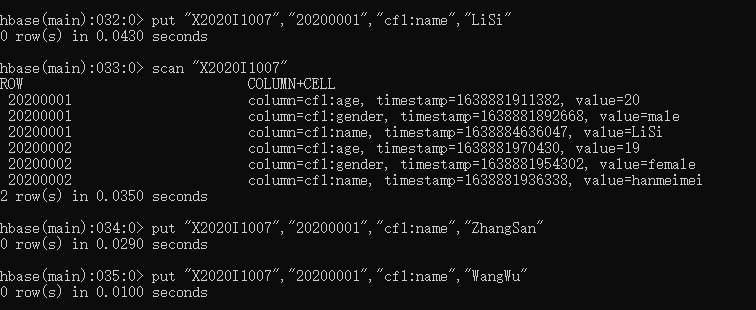
## 修改数据(20’)

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

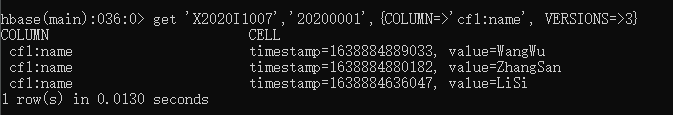


|  |  |  |  |
| --- | --- | --- | --- |
| **行键** | **列族** | **列名** | **单元格的值** |
| **20200001** | **cf1** | **name** | **LiSi** |
| **20200001** | **cf1** | **name** | **ZhangSan'** |
| **20200002** | **cf1** | **name** | **WangWu** |

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



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



这里说明一下，因手快了直接查看了name部分的多版本数据

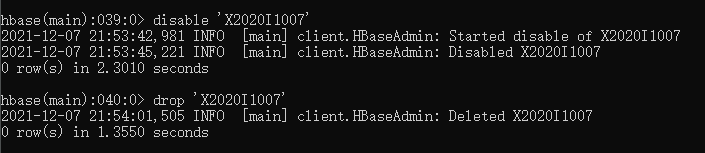
### 删除行键为20200002，列名为age，的数据（delete）(3’)



### 删除行键为20200002的行（deleteall）(4’)



### 删除整个表（disable，drop）(4’)



## Region初探(20’)

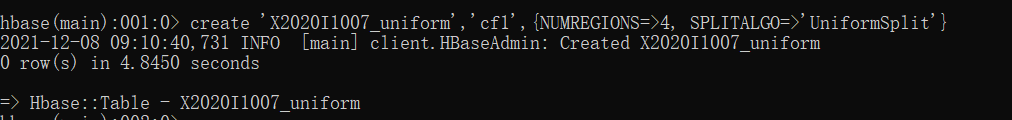
HBase默认建表时只有一个region，这个region的rowkey是没有边界的，即没有startkey，也没有endkey。在数据写入时，所有数据都会写入这个默认的region，随着数据量的不断增加，此region已经不能承受不断增长的数据量，会进行split，分成2个region。在此过程中，会产生两个问题：

1.数据往一个region上写,会有写热点问题。

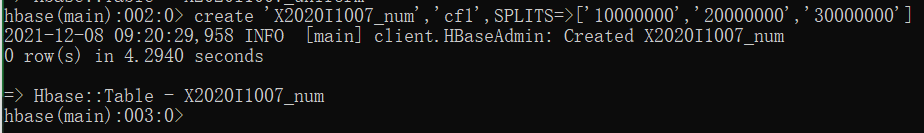
2.region split会消耗宝贵的集群I/O资源。

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。

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



### 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’)

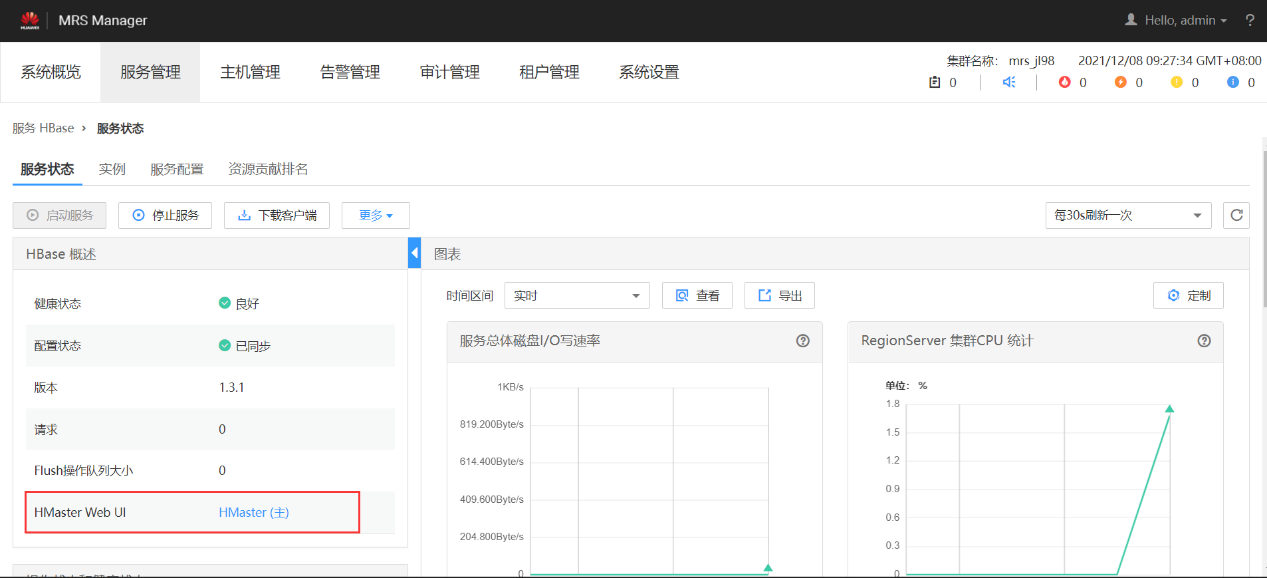


### 在Manager中查看HBase(5’)

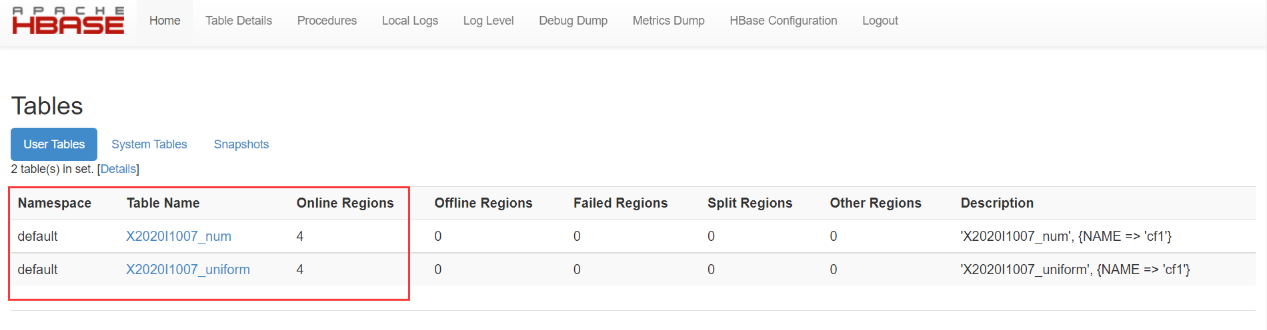
MRS Manager界面，点击“HBase”服务



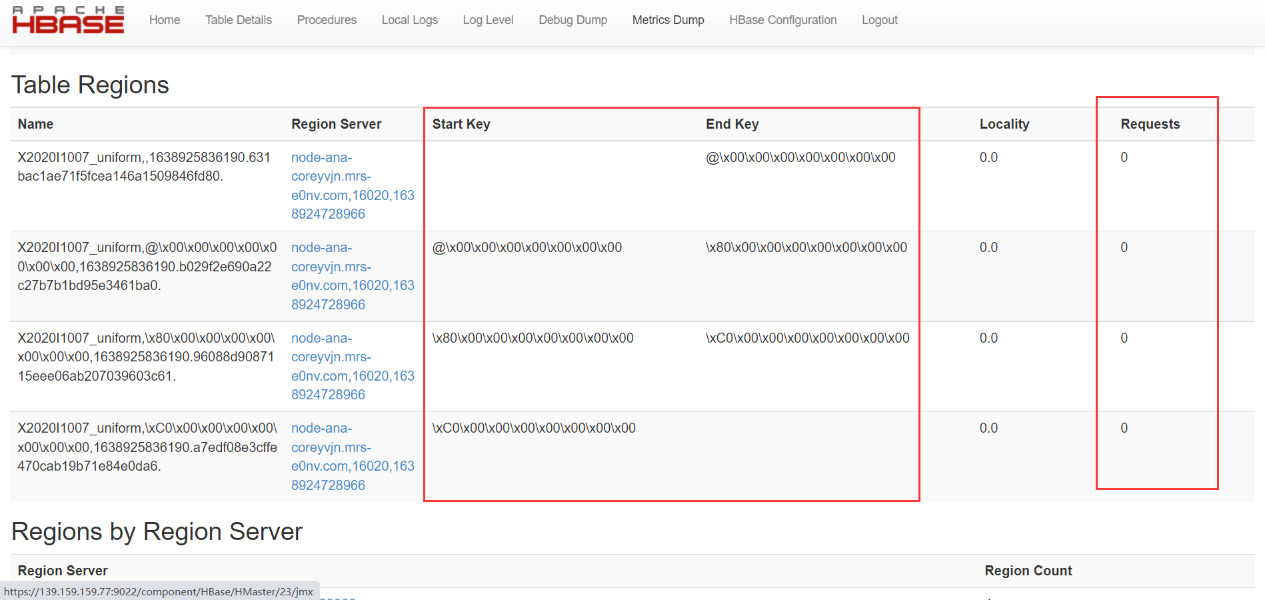
点击HMaster(主)进入HBase UI

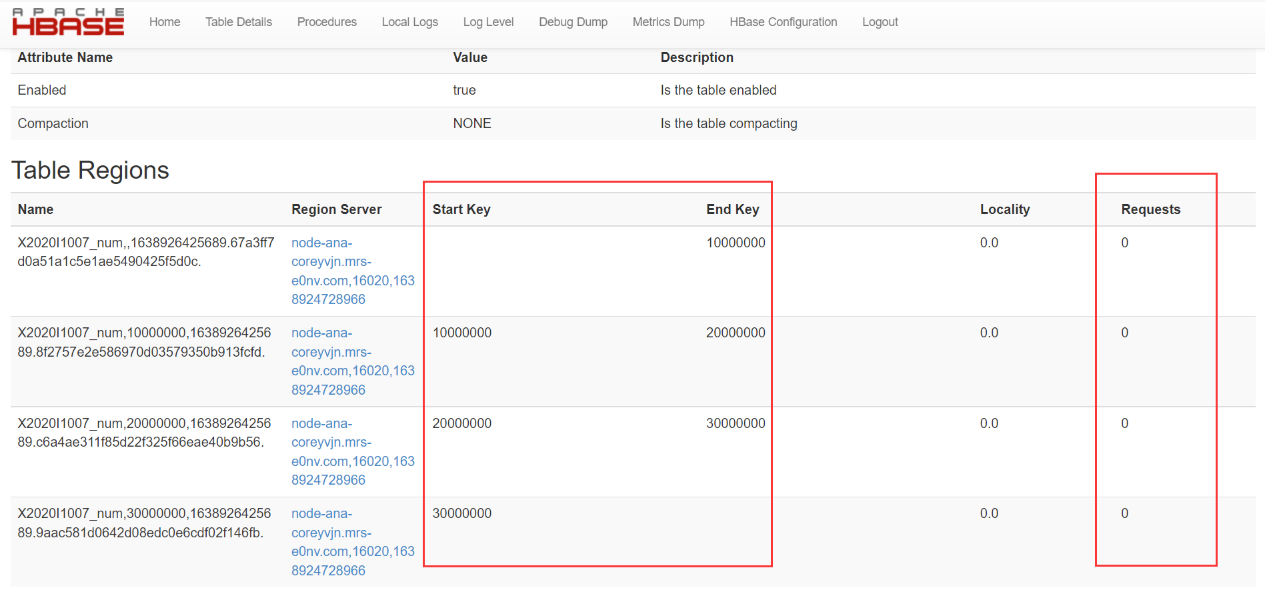


“User Tables”下点击创建好的表名“cx\_table\_stu02”，如“下图”：

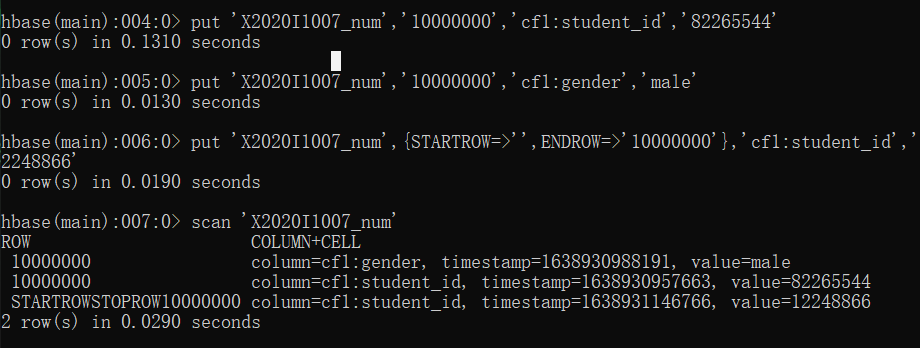


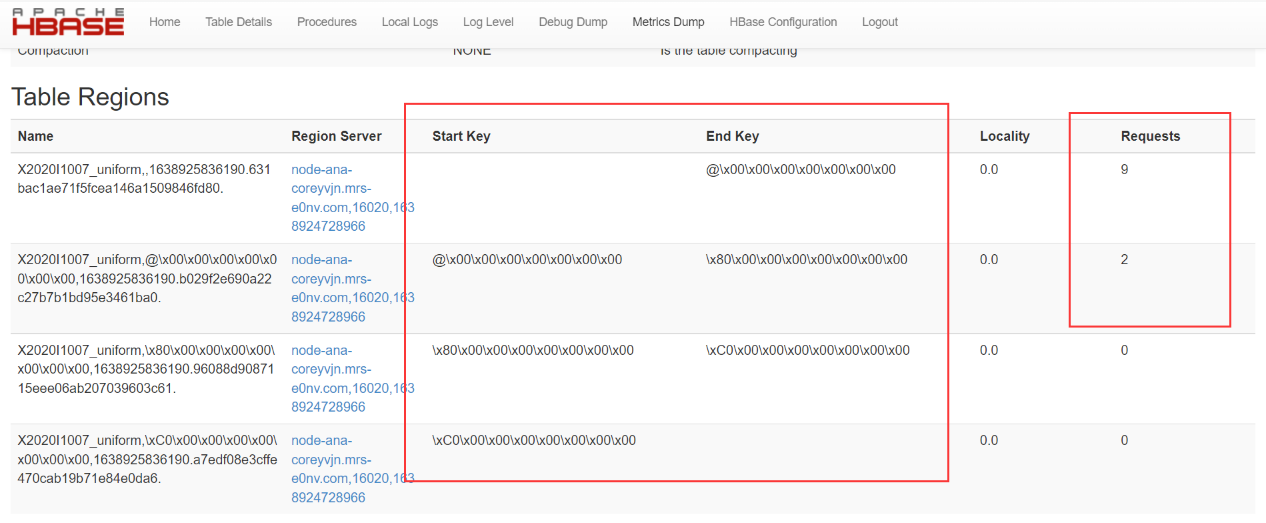
查看该表分区情况，如下图：

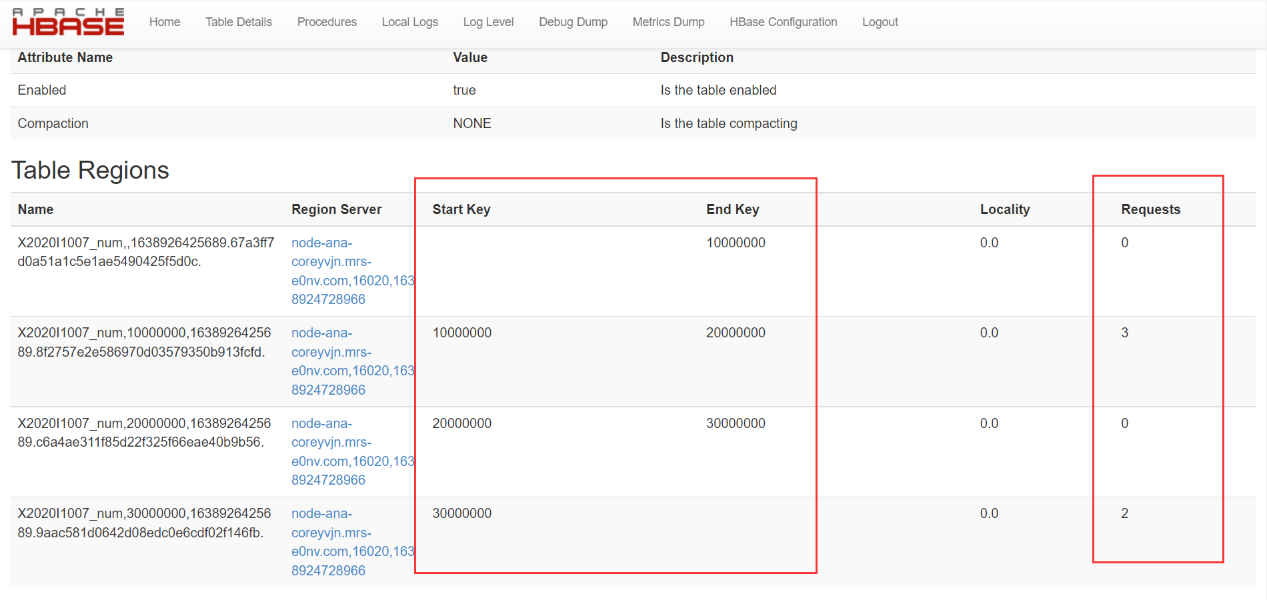




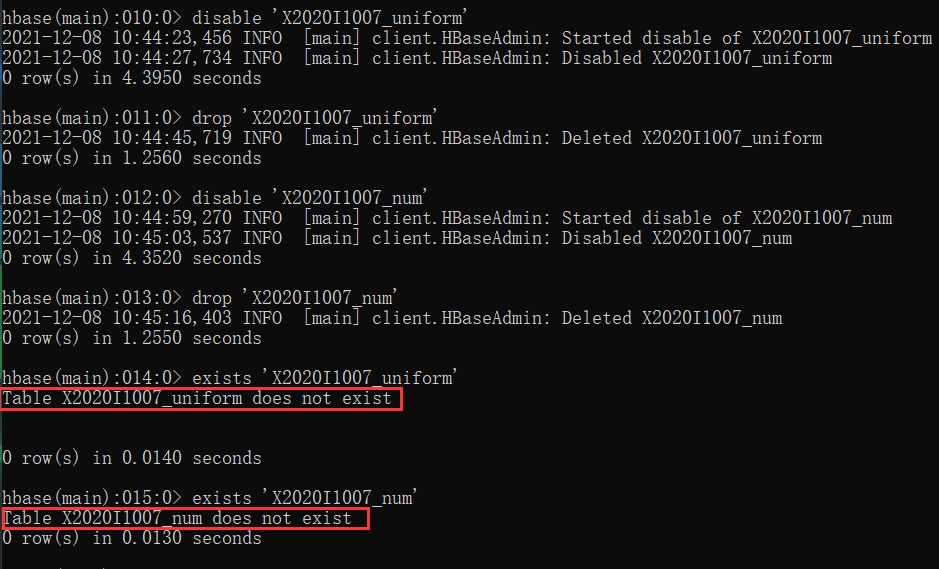
### 根据两个表的End key和Start Key，选择适当的行键往两个表的不同region中添加任意两个数据，使得每个表至少有两个不同region中Requests不为0（put）(5’)







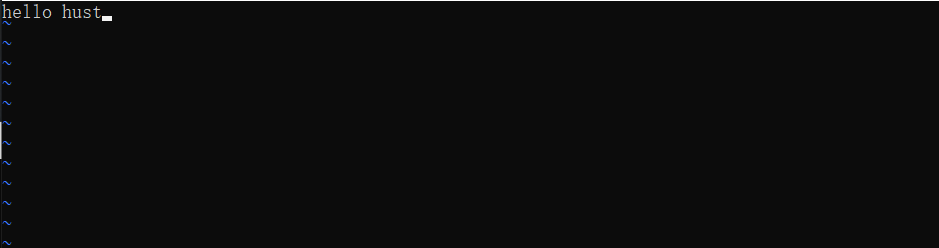
### 删除所有表(5’)



## hive初探(10’)

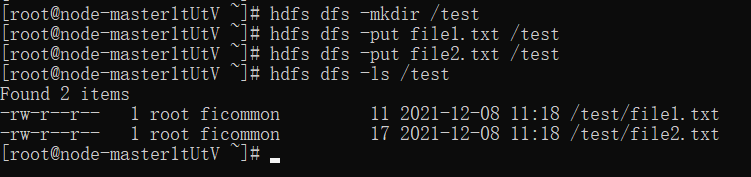
### 准备file1.txt, 内容为”hello hust”，file2.txt，内容为”hello 学号”（vim）(3’)





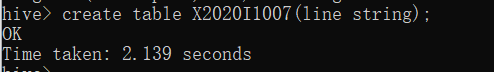


### 将创建的文件移动到HDFS中/test文件夹内（见上次实验）(3’)



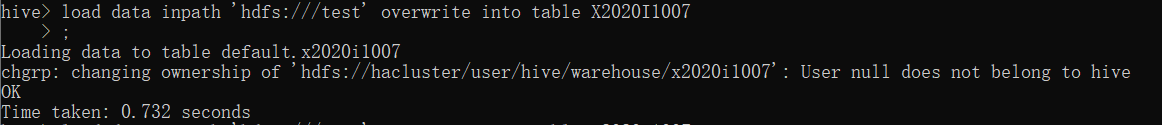
### 在hive中创建表，tablename替换为学号(1’)

create table tablename(line string);



### 加载hdfs中的数据到hive中(1’)

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



### 通过HiveQL语句创建词频统计表(1’)

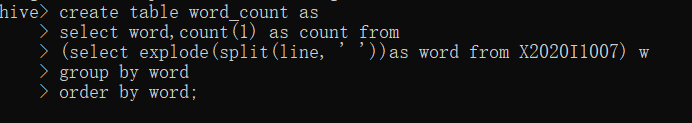
create table word\_count as

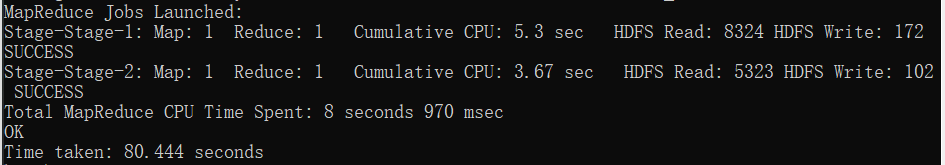
select word,count(1) as count from

(select explode(split(line, ' '))as word from tablename) w

group by word

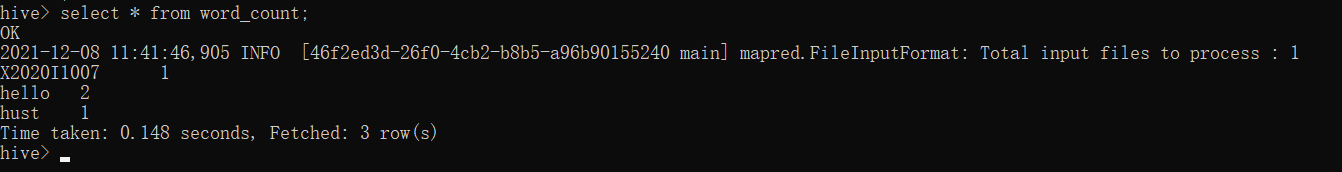
order by word;





### 通过HiveQL语句创建词频统计表(1’)

select \* from word\_count;



五：实验总结(10’)

本次实验目的在于实现MapReduce中的Hbase分布式数据库的功能和基于Hadoop的Hive数据仓库分析系统的操作。实验前期，总是因为华为云的MapReduce大数据并行处理平台出问题，一是卡在集群创建安装时候；二是卡在无法打开MapReduce MRS的集群管理manager；三是进入Cloud Shell云平台后，在Hbase Shell模式下创建表报error。到后面使用命令提示符才能顺利进入Hbase Shell开始实验内容。后面就是抓紧时间把实验进度赶上来。希望华为云以后不要再出各种各样的问题，太闹心和浪费时间以及金钱了。总的来说，在完成本次实验后，我感到受益匪浅，不仅对Hbase分布式存储系统的运用和运作原理有了更深入的了解，同时还学到了新的知识“Hive数据仓库分析系统操作和运作模式”，受益匪浅。我会在接下来的实验里再接再厉。也感谢助教的指导，否则我是不可能自己找到方法去解决华为云各种出错的问题。

**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