****

**实验六：KV 查询**

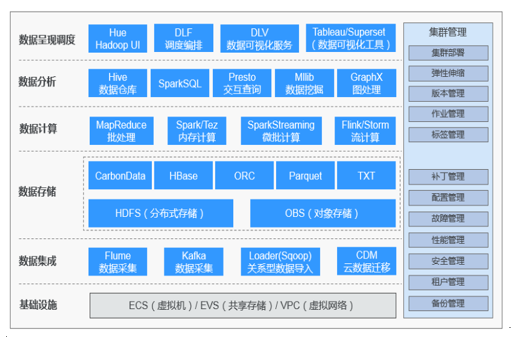
**实验报告**

|  |  |
| --- | --- |
| **学 期：** | 2023-2024学年 第二学期 |
| **专 业：** | 计算机科学与技术（实验班） |
| **学号姓名：** | 21071003 高立扬 |
| **完成日期：** | 2024.05 |

### 问题

1. 关于MRS

大数据是人类进入互联网时代以来面临的一个巨大问题：社会生产生活产生的数据量越来越大数据种类越来越多，数据产生的速度越来越快。传统的数据处理技术，比如说单机存储，关系数据库已经无法解决这些新的大数据问题。为解决以上大数据处理问题，Apache基金会推出了Hadoop大数据处理的开源解决方案。Hadoop是一个开源分布式计算平台，可以充分利用集群的计算和存储能力，完成海量数据的处理。企业自行部署Hadoop系统有成本高，周期长，难运维和不灵活等问题。针对上述问题，华为云提供了大数据服务MapReduce(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一站式企业级大数据集群云服务，完全兼容开源接口，结合华为云计算、存储优势及大数据行业经验，为客户提供高性能、低成本、灵活易用的全栈大数据平台，轻松运行Hadoop、Spark、HBase、Kafka、Storm等大数据组件，并具备在后续根据业务需要进行定制开发的能力，帮助企业快速构建海量数据信息处理系统，并通过对海量信息数据实时与非实时的分析挖掘，发现全新价值点和企业商机。



MRS架构包括了基础设施和大数据处理流程各个阶段的能力。

基础设施

MRS基于华为云弹性云服务器ECS构建的大数据集群，充分利用了其虚拟化层的高可靠、高安全的能力。

虚拟私有云(VPC)为每个租户提供的虚拟内部网络，默认与其他网络隔离。

云硬盘(EVS)提供高可靠、高性能的存储。

弹性云服务器(ECS)提供的弹性可扩展虚拟机，结合VPC、安全组、EVS数据多副本等能力打造一个高效、可靠、安全的计算环境。

数据集成

数据集成层提供了数据接入到MRS集群的能力，包括Flume(数据采集)、Loader(关系型数据导入)、Kafka(高可靠消息队列)，支持各种数据源导入数据到大数据集群中。

数据存储

MRS支持结构化和非结构化数据在集群中的存储，并且支持多种高效的格式来满足不同计算引擎的要求。

HDFS是大数据上通用的分布式文件系统。

OBS是对象存储服务，具有高可用低成本的特点。

HBase支持带索引的数据存储，适合高性能基于索引查询的场景。

数据计算

MRS提供多种主流计算引擎：MapReduce(批处理)、Tez(DAG模型)、Spark(内存计算)、SparkStreaming(微批流计算)、Storm(流计算)、Flink(流计算)，满足多种大数据应用场景，将数据进行结构和逻辑的转换，转化成满足业务目标的数据模型

数据分析

基于预设的数据模型，使用易用SQL的数据分析，用户可以选择Hive(数据仓库)，SparkSQL以及Presto交互式查询引擎。

数据呈现调度

用于数据分析结果的呈现，并与数据湖工厂(DLF)集成，提供一站式的大数据协同开发平台，帮助用户轻松完成数据建模、数据集成、脚本开发、作业调度、运维监控等多项任务，可以极大降低用户使用大数据的门槛，帮助用户快速构建大数据处理中心。

集群管理

以Hadoop为基础的大数据生态的各种组件均是以分布式的方式进行部署，其部署、管理和运维复杂度较高。

MRS集群管理提供了统一的运维管理平台，包括一键式部署集群能力，并提供多版本选择，支持运行过程中集群在无业务中断条件下，进行扩缩容、弹性伸缩。同时MRS集群管理还提供了作业管理、资源标签管理，以及对上述数据处理各层组件的运维，并提供监控、告警、配置、补丁升级等一站式运维能力。

产品优势

MRS服务拥有强大的Hadoop内核团队，基于华为FusionInsight大数据企业级平台构筑。历经行业数万节点部署量的考验，提供多级用户SLA保障。

MRS具有如下优势：

高性能

MRS支持自研的CarbonData存储技术。CarbonData是一种高性能大数据存储方案，以一份数据同时支持多种应用场景，并通过多级索引、字典编码、预聚合、动态Partition、准实时数据查询等特性提升了IO扫描和计算性能，实现万亿数据分析秒级响应。同时MRS支持自研增强型调度器Superior，突破单集群规模瓶颈，单集群调度能力超10000节点。

低成本

基于多样化的云基础设施，提供了丰富的计算、存储设施的选择，同时计算存储分离，提供了低成本海量数据存储方案。MRS可以按业务峰谷，自动弹性伸缩，帮助客户节省大数据平台闲时资源。MRS集群可以用时再创建、用时再扩容，用完就可以销毁、缩容，确保成本最优。

高安全

MRS服务支持Kerberos安全认证，实现了基于角色的安全控制及完善的审计功能。MRS支持在华为云的公共资源区，资源专属区、客户机房的HCS Online上为客户不同物理隔离方式的一站式大数据平台。集群内支持逻辑多租户，通过权限隔离，对集群的计算、存储、表格等资源按租户划分。

易运维

MRS提供可视化大数据集群管理平台，提高运维效率。并支持滚动补丁升级，可视化补丁发布信息，一键式补丁安装，无需人工干预，不停业务，保障用户集群长期稳定。

高可靠

MRS服务支持全节点高可用，实时短信/邮件通知。

1. 关于本实验

本课程为《大数据》实验课程，模拟高校学生系统，在华为MRS上进行HBase操作，通过创建表、对表进行CURD操作、创建预分region表及Filter过滤器的使用，让学员达到在MRS上熟练掌握KV查询，为了更好地掌握本书内容，阅读本书的读者应首先具备以下基本条件：

具有基本的大数据常用概念。

熟悉HBase。

有一定的逻辑性。

### 解决方法

1. 华为云服务器配置

为了满足基于HBase-MRS实验需要，建议每位学员提前注册华为云账号并充值，确保通过PUTTY访问外网畅通。

检查设备

实验开始之前请每位学员检查自己的实验设备是否齐全，实验清单如下。

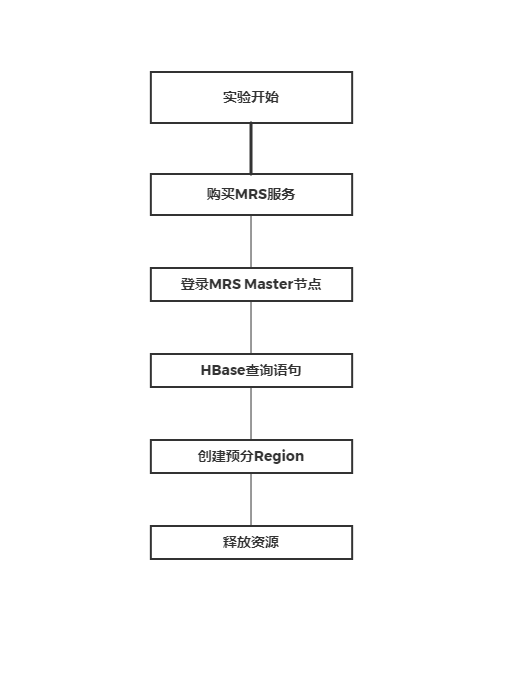
|  |  |  |
| --- | --- | --- |
| **设备名称** | **数量** | **备注** |
| MapReduce | 1个 | 选择“鲲鹏计算” |
| 弹性公网IP | 1个 | 需购买弹性公网IP |

每组检查自己的设备列表如下：

购买MapReduce 1个

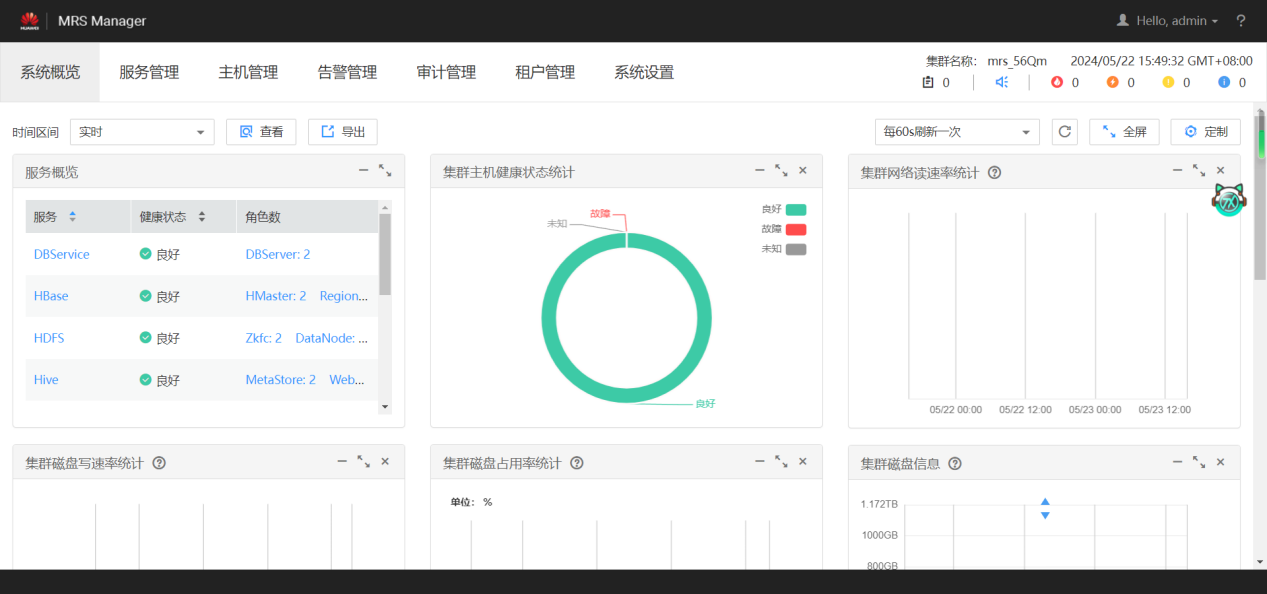
购买弹性IP 1个

1. 工作流程



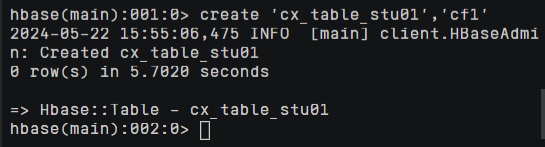
### 实验结果

1. 购买MapReduce服务

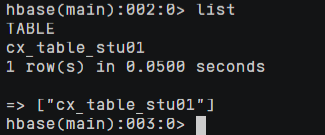


1. 远程登录云服务器操作HBase

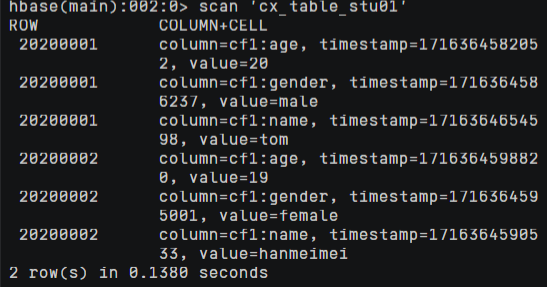
创建基表



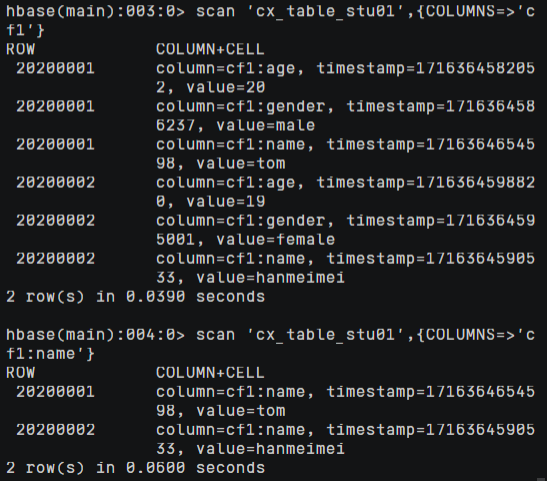
list显示所有表



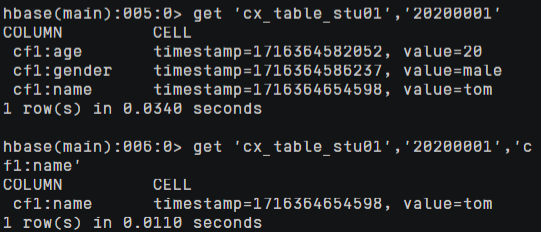
增加数据



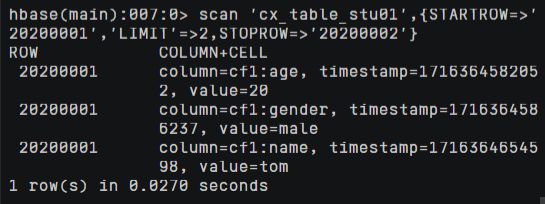
Scan方式查询数据



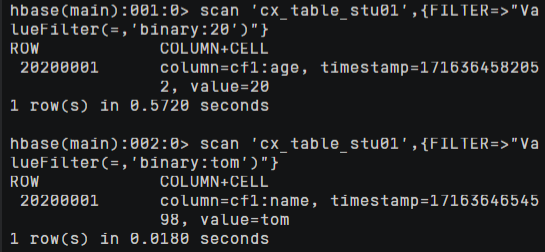
Get方式查询数据

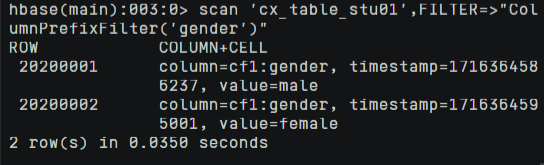


指定条件查询数据

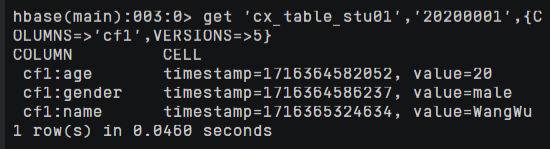
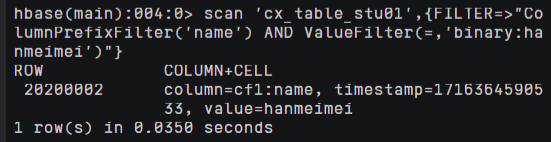


过滤器查询

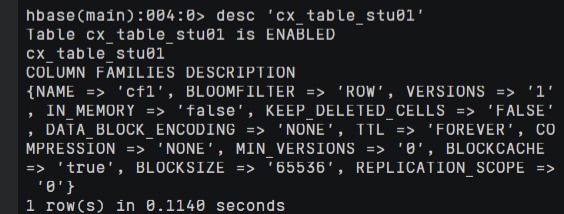




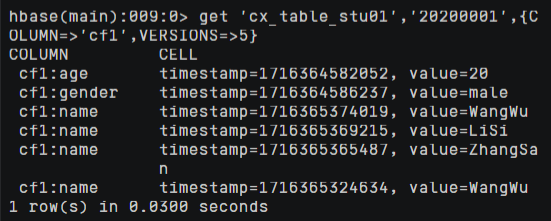
查询多版本数据



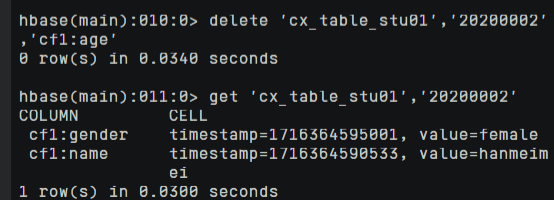
查看表的属性，执行语句

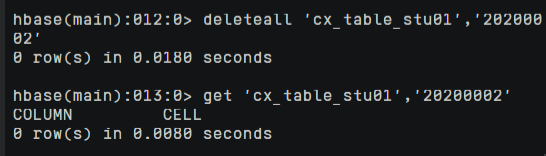


查看多个版本的数据

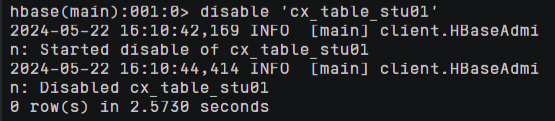


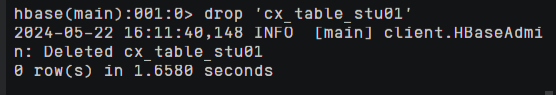
删除数据





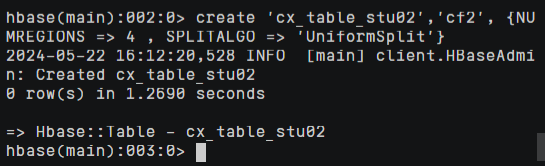
删除表



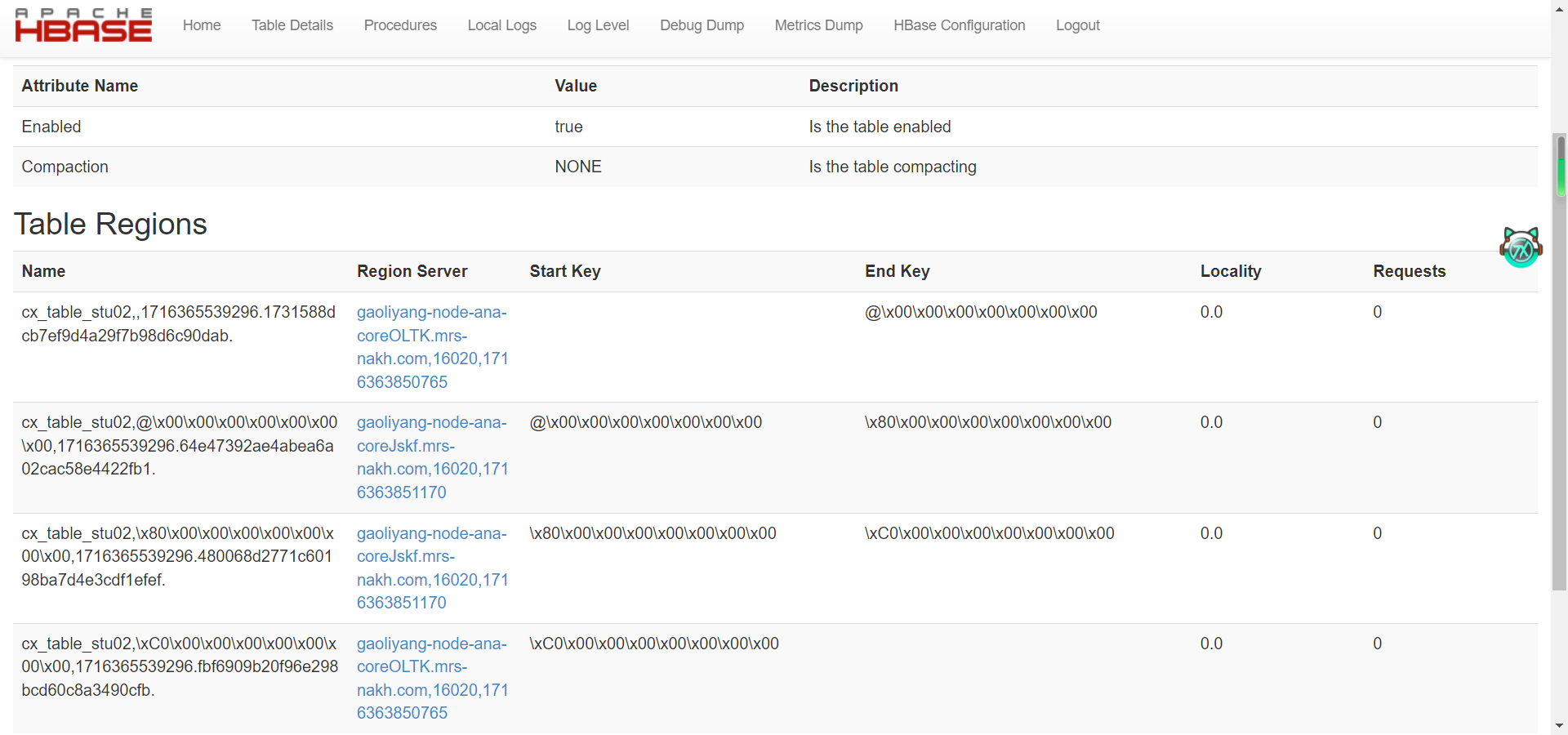


1. 创建预分region表

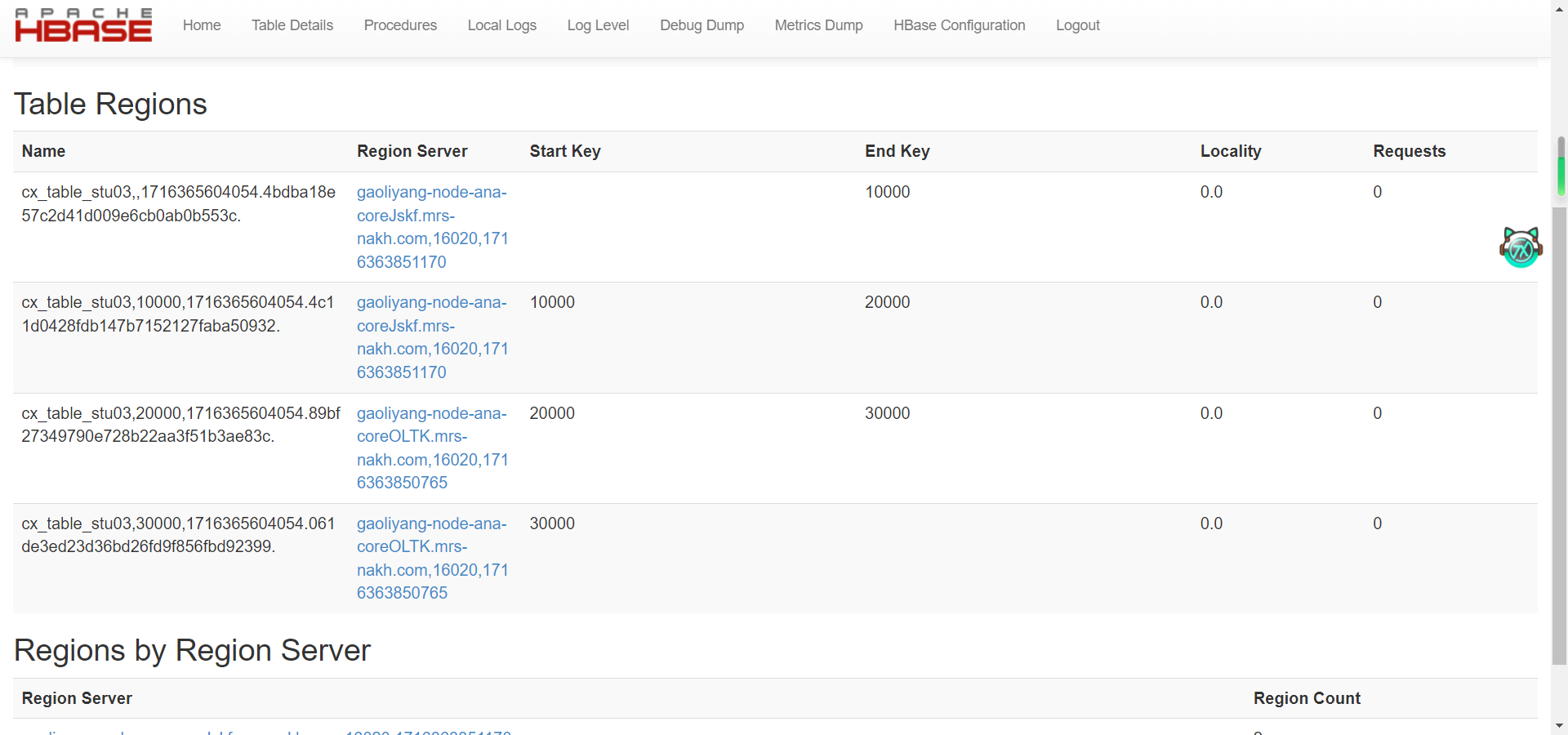
创建表执行命令



登录HBase WebUI查看表的分区情况



指定region的startKey和endKey



### 经验和技巧

tips1.指导书第20页的scan ‘cx\_table\_stu01’，如果直接粘贴进华为云hbase shell内，会出现错误，这是因为单引号为中文符号，需要用户手动改为英文符号

tips2.指导书第21页的scan 'cx\_table\_stu01',FILTER=>"ColumnPrefixFilter('gender')“，同样也不能直接粘贴，最后一个双引号应该改为英文双引号

tips3.指导书第22页的put 'cx\_table\_stu01','20200001','cf1:name','ZhangSan'scan

，结尾的scan是多余的

tips3.指导书第23页的disable ‘cx\_table\_stu01’和drop ‘cx\_table\_stu01’同样是中文单引号问题