

**大数据处理实验报告**

实验二：HBase的基本操作

专业班级： CS 2002

学 号： U202015324

姓 名： 屈绍博

指导教师： 石宣化

报告日期： 2022.4.5

**计算机科学与技术学院**

**《大数据处理》课程实验报告**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 实验地点 | | 南一楼804 | 课程名称 | | 大数据处理 | | |
| 实验题目 | | HBase的基本操作 | 成绩 |  | | 指导教师 |  |
| 教师评价 | □ 实验过程正确； □ 源程序/实验内容提交； □ 程序结构/实验步骤合理；  □ 实验结果正确； □ 语法、语义/命令正确； □ 报告规范；  其他： | | | | | | |
| **一、实验目的**   1. 了解HBase的用途 2. 掌握HBase的基本命令   **二、实验内容**   1. 实验环境配置 2. 准备数据 (20’) 3. 查询数据 (30’) 4. 修改数据 (20’) 5. Region初探 (20’) 6. Hive初探 (10’) 7. 实验总结 (10’)   **三、实验环境**  **（1）集群** 版本：MRS 1.9.2 类型：分析集群 组件选择：Hadoop 2.8.3, HBase 1.3.1, Hive 2.3.3 and Tez 0.9.1 **（2）CPU架构** Master节点 鲲鹏通用计算增强型： 4 vCPUs | 16 GB | kc1.xlarge.4 系统盘：通用型SSD 100 GB x 1 数据盘：通用型SSD 200 GB x 1 分析Core节点 鲲鹏通用计算增强型： 4 vCPUs | 16 GB | kc1.xlarge.4 系统盘： 通用型SSD 100 GB x 1 数据盘： 通用型SSD 100 GB x 1  在Windows11操作系统上，使用chrome浏览器登录网页端华为云CloudShell，使用控制台提供的VNC登录服务器，在linux环境下进行实验。  **四、实验过程或步骤（源程序）**  **1.1 环境配置**    **图1.1 配置环境变量并进入hbase shell**  配置好集群，登入并设置环境变量，如上图所示   * 1. **准备数据**  1. **进入 hbase shell**   利用hbase shell指令进入hbase shell    **图2.1 配置环境变量并进入hbase shell**   1. **创建一个表，表名为学号，列族名为 cf1**     **图2.2 创建表**  利用create指令创造一个表   1. **显示所有的表**     **图2.3 显示所有的表**  利用list指令，显示所有的表   1. **向表中增加两行数据**     **图2.4 向表中加入数据**  分多次利用put指令，向表中添加数据   * 1. **查询数据**  1. **查找表中，列族名为 cf1 的数据**     **图2.5 查询列族名为cf1的数据**  利用scan指令，并利用COLUMN进行限定，可以查找表中列族名为 cf1 的数据。   1. **查找表中，列族名为 cf1,列名为 name 的数据**     **图2.6 查询列族名为 cf1,列名为 name 的数据**  再次利用scan指令，并在COMLUMN中不仅对列族，也对列名进行限定，即可查找到列族名为 cf1列名为 name 的数据如上图所示。   1. **查找表中，行键为 20200001 的行**     **图2.7 查询行键为 20200001 的行**  利用get指令，查找表中行键为 20200001 的行，直接在表名后对行键名进行限定即可。   1. **查找表中，行键为 20200001，列族为 cf1，列名为 name**   **的数据**    **图2.8 查询结果**  利用get指令，依次在表名后对行键名，列族名和列名进行限定即可。   1. **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**     **图2.9 查询结果**  利用scan指令，并利用STARTROW对起始行键进行限定，LIMIT对所取的行键数进行限定即可。   1. **查看有数据值为 20 的行**     **图2.10 查询结果**  利用scan指令，并利用FILTER中的ValueFilter对数据值进行限定，将所取出的数据值限定在20，即可满足题目需求。   1. **查看有数据值为 tom 的行**     **图2.11 查询结果**  与上题类似，我们依然利用scan指令，并利用FILTER中的ValueFilter对数据值进行限定，将所取出的数据值限定为“TOM”，即可满足题目需求。   1. **查看列名为 gender 的列**     **图2.12 查询结果**  我们利用scan指令，并利用COLUMN限制列族和列名为gender，即可满足题目需求。   1. **查看列名为 name，值为 hanmeimei 的行**     **图2.13 查询结果**  我们利用scan指令，并利用COLUMN对列族和列进行限定，同时利用FILTER中的ValueFilter对值进行限制，限制为“hanmeimei”，即可满足题目需求。   1. **查看表的属性（desc）**     **图2.14 查询结果**  利用desc指令查看表的属性如上图所示。   * 1. **修改数据**  1. **改变表的 VERSIONS 为 5 以显示更多的历史版本**     **图2.15 历史版本限定转换**  利用alter指令对表进行操作，使其可以显示至多五个历史版本。   1. **添加行键 20200001，列族 cf1，列名 name 的多个历史版本**     **图2.16 修改数据**    **图2.17 修改数据**    **图2.18 修改数据**  利用上文提到的put方法对表中相应数据进行修改，并利用scan指令对修改结果进行验证，如上图所示。   1. **查看所有行键为 20200001，列簇为 cf1 的多版本数据**     **图2.19 查询结果**  利用scan指令，并用VERSION限定要查看的历史版本数，COLUMN限定要查看的历史版本的列族，即可满足题目要求。   1. **删除行键为 20200002，列名为 age，的数据**     **图2.20 删除数据**  利用delete指令删除指定行键的列。并利用get对删除结果进行验证。    **图2.21 查询结果**   1. **删除行键为 20200002 的行**     **图2.22 删除数据**  利用deleteall删除特定的行，在表名后对要删除的行进行说明即可，并利用scan进行验证删除后的表。   1. **删除整个表**     **图2.23 删除表**  依次利用disable和drop对表进行下线和删除操作。   * 1. **Region 初探**  1. **创建具有四个 region 的表，表名为”学号\_uniform”**     **图2.24 表的创建**  利用create指令，并利用NUMREGIONS指定region数，SPLITALGO指定分割算法。   1. **创建具有四个 region 的表，表名为”学号\_num”**     **图2.25 表的创建**   1. **在 Manager 中查看 HBase**     **图2.26 查询结果**  可以在manager中观察到我们刚刚创立的两个表。    **图2.27 查询结果**    **图2.28 查询结果**  并且两个表中均没有数据。   1. **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**   利用put指令，紧接表名跟随地址限定，即可实现向不同的region添加数据的功能。相应的指令以及对应region的request如下图所示。  向10000000地址添加数据    **图2.29 添加数据**  该地址对应的request增加    **图2.30 查询结果**  向20000000地址添加数据    **图2.31 添加数据**  该地址对应的request增加    **图2.32 查询结果**  向@\x00\x00\x00\x00\x00\x00\x00地址添加数据    **图2.33 添加数据**  该地址对应的request增加    **图2.34 查询结果**  向1地址添加数据    **图2.35 添加数据**  该地址对应的request增加    **图2.36 查询结果**  可以观察到，每次向相应的region中添加数据，相应的request都会加1。   1. **删除所有表**   利用disable和drop指令依次对两个表进行下线和删除操作。    **图2.37 表的删除**    **图2.38 表的删除**   * 1. **hive 初探**  1. **准备文件**   利用vim指令以及Linux指令添加文件并在文件中保存相应的值。    **图2.39 文件的准备**   1. **移动文件**   利用实验一中的方法，建立文件夹，并将刚刚创建的两个文件放到文件夹目录下。    **图2.40 文件的移动**  调用ls指令观察文件夹目录下文件，验证文件是否成功移动。    **图2.41 文件移动的验证**   1. **在 hive 中创建表**   首先输入hive指令进入hive，并调用create指令创建表名为学好的 表。    **图2.42 表的创建**   1. **加载 hdfs 中的数据到 hive 中**   调用下方指令将hdfs中的数据传入hive    **图2.43 数据的导入**   1. **通过 HiveQL 语句创建词频统计表**     **图2.44 词频统计表的创建**   1. **词频统计表的查看**     **图2.45 词频统计表的查看**  **五、出现的问题与解决方案**  在hive的后两步操作中，我通过hiveql语句建立词频表，但在最后一步操作对词频表进行查看时系统总是会报错提示表word\_count不存在，我重复尝试了几次这个问题还是没能得到解决，我便将命令复制下来一边后续使用。在我重启了集群后重新执行上面的hiveql语句，再次查看词频统计表，问题就得到了解决，我怀疑之前的错误时网络连接或是延迟的问题。  **六、实验总结**  这次实验的内容明显比第一次实验深入了许多，逐渐涉及到一些hbase以及hive具体指令的灵活使用，通过这次试验，我们不仅学会了使用实验任务书内要求的指令，也接触到了一些其他的指令，比如我在建立表的过程中经常会有一些误操作，而这时如何取出错误数据也成为了我必须要掌握的技能之一。在实验中遇到一些操作上的错误和突发情况，也能精进我对这个系统指令的理解。 | | | | | | | |