《HBase 分布式数据库》实训指导书

院系名称：经管信息学院

课程代码：31091120

总学时数：64

适用专业：大数据技术与应用专业

编 制 人：韦 祥、孙 检

编制日期：2020 年 7 月

审 核 人：

审 定 人：

《HBase 分布式数据库》实训指导书

一、实训目的与要求

《HBase 分布式数据库》是一门分布式数据库，为学生搭建起通向“大数据知识空间”的桥梁和纽带，以“构建知识体系、阐明基本原理、引导初级实践、了解相关应用”为原则，为学生在大数据领域“深耕细作”奠定基础、指明方向。课程将系统讲授大数据的基本概念、HBase 数据模型、数据操纵语言数据可视化以及大数据在互联网、生物医学和物流等各个领域的应用。在 Hbase Shell 的使用、模式设计等重要章节，安排了 HBase 入门级的实践操作，让学生更好地学习和掌握大数据关键技术。

二、实训内容

（一）实例实训

以 HBase 项目案例讲解 HBase 数据表创建、数据创添加、修改、删除、查询等操作，让学生能够接触到真实的企业实例。

（二）项目实训

让学生根据 JJ 微博项目要求，完成相应的数据库设计和表创建。

（三）总结

对学生的全部作品进行考核，并选择典型的案例对实训的结果进行考核。

二、参考课时

标题	实训内容	实训课时
实训一	HBase 基础知识	4
实训二	HBase 安装	4
实训三	HBase Shell 操作	8
实训四	HBase 分布式部署	8
实训五	HBase 数据结构	4
实训六	HBase 原理	6
实训七	HBase 优化	4
实训八	HBase 实战之 JJ 微博	16

实训九	总结	4
总计		58

三、实训材料准备

（一）软件准备

序号	设备、软件名称	规格/技术参数、用途	备注
1	截图工具		系统自带截图工具
2	Hadoop2.6.0 或以上		选用 hadoop 生产环境稳定版本
3	JDK1.7 及以上		选用 openjdk
4	Hbase1.2 及以上		选用与 hadoop 版本兼容的 hbase

（二）硬件准备

序号	设备、软件名称	规格/技术参数、用途	备注
1	大数据技术实训机房	测试场地	保证参考人员有足够间距
2	计算机	CPU 奔腾4 以上，内存 2G 以上，linux 操作系统（ubuntu 或 centos）。	用于软件部署，每人三台。
3	交换机与网线	用于组建局域网	100 兆网络及以上

四、综合实训考核办法：

系统文档　　10 分

数据库设计　　20 分

数据表创建　　30 分

数据库调试　　10 分

实训出勤　　10 分

HBase 优化　　20 分

目　　录

实训一 HBase 基础知识

一、实训目的和要求

1. 了解 HBase 是什么；

2. 掌握 HBase 特点。

二、实训内容

HBase 概念、HBase 特点等。

三、实训准备

1. 操作系统：Linux（建议 Centos6.5 以上）；

2. Hadoop 版本：2.7.2；

3. JDK 版本：1.7 或以上版本；

4. Java IDE：Eclipse。

四、实训步骤

1.1 什么是 HBase

HBase 的原型是 Google 的 BigTable 论文，受到了该论文思想的启发，目前作为 Hadoop 的子项目来开发维护，用于支持结构化的数据存储。

官方网站：http://hbase.apache.org

-- 2006 年 Google 发表 BigTable 白皮书

--2008 年北京成功开奥运会，程序员默默地将 HBase 弄成了 Hadoop 的子项目

-- 2010 年 HBase 成为 Apache 顶级项目

-- 现在很多公司二次开发出了很多发行版本，你也开始使用了。

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，利用 HBASE 技术可在廉价 PC Server 上搭建起大规模结构化存储集群。

HBase 的目标是存储并处理大型的数据，更具体来说是仅需使用普通的硬件配置，就能够处理由成千上万的行和列所组成的大型数据。

HBase 是 Google Bigtable 的开源实现，但是也有很多不同之处。比如：Google Bigtable 利用 GFS 作为其文件存储系统，HBase 利用 Hadoop HDFS 作为其文件存储系统；Google 运行 MAPREDUCE 来处理 Bigtable 中的海量数据，HBase 同样利用 Hadoop MapReduce 来处理 HBase 中的海量数据；Google Bigtable 利用 Chubby 作为协同服务，HBase 利用 Zookeeper 作为对应。

1.2 HBase 特点

1）海量存储

Hbase 适合存储 PB 级别的海量数据，在 PB 级别的数据以及采用廉价 PC 存储的情况下，能在几十到百毫秒内返回数据。这与 Hbase 的极易扩展性息息相关。正式因为 Hbase 良好的扩展性，才为海量数据的存储提供了便利。

2）列式存储

这里的列式存储其实说的是列族存储，Hbase 是根据列族来存储数据的。列族下面可以有非常多的列，列族在创建表的时候就必须指定。

3）极易扩展

Hbase 的扩展性主要体现在两个方面，一个是基于上层处理能力（RegionServer）的扩展，一个是基于存储的扩展（HDFS）。

通过横向添加 RegionSever 的机器，进行水平扩展，提升 Hbase 上层的处理能力，提升 Hbsae 服务更多 Region 的能力。

备注：RegionServer 的作用是管理 region、承接业务的访问，这个后面会详细的介绍通过横向添加 Datanode 的机器，进行存储层扩容，提升 Hbase 的数据存储能力和提升后端存储的读写能力。

4）高并发

由于目前大部分使用 Hbase 的架构，都是采用的廉价 PC，因此单个 IO 的延迟其实并不小，一般在几十到上百 ms 之间。这里说的高并发，主要是在并发的情况下，Hbase 的单个 IO 延迟下降并不多。能获得高并发、低延迟的服务。

5）稀疏

稀疏主要是针对 Hbase 列的灵活性，在列族中，你可以指定任意多的列，在列数据为空的情况下，是不会占用存储空间的。

五、实训方法

机房利用本机软件完成。

六、考核办法

此实训以理解性记忆为主，故无需考核。

七、思考和练习

1. 什么是 HBase？
2. HBase 特点有哪些？

实训二 HBase 安装

一、实训目的和要求

1. 掌握 HBase 的安装；

2. 掌握 HBase 的部署、启动。

二、实训内容

HBase 安装、部署、启动测试等。

三、实训准备

1. 操作系统：Linux（建议 Centos6.5 以上）；

2. Hadoop 版本：2.7.2；

3. JDK 版本：1.7 或以上版本；

4. Java IDE：Eclipse。

四、实训步骤

1.1 Zookeeper 正常部署

首先保证 Zookeeper 集群的正常部署，并启动之：

```
[sun@hadoop102 zookeeper-3.4.10]$ bin/zkServer.sh start
[sun@hadoop103 zookeeper-3.4.10]$ bin/zkServer.sh start
[sun@hadoop104 zookeeper-3.4.10]$ bin/zkServer.sh start
```

1.2 Hadoop 正常部署

Hadoop 集群的正常部署并启动：

```
[sun@hadoop102 hadoop-2.7.2]$ sbin/start-dfs.sh
[sun@hadoop103 hadoop-2.7.2]$ sbin/start-yarn.sh
```

1.3 HBase 的解压

解压 HBase 到指定目录：

```
[sun@hadoop102 software]$ tar -zxvf hbase-1.3.1-bin.tar.gz -C /opt/module
```

1.4 HBase 的配置文件

修改 HBase 对应的配置文件。

1）hbase-env.sh(/conf 下)修改内容：

```
export JAVA_HOME=/opt/module/jdk1.8.0_144
export HBASE_MANAGES_ZK=false
```

2）hbase-site.xml 修改内容：

```
<configuration>
    <property>
        <name>hbase.rootdir</name>
        <value>hdfs://hadoop102:9000/hbase</value>
    </property>

    <property>
```

```
        <name>hbase.cluster.distributed</name>
        <value>true</value>
    </property>

    <!-- 0.98 后的新变动，之前版本没有.port,默认端口为
60000 -->
    <property>
        <name>hbase.master.port</name>
        <value>16000</value>
    </property>

    <property>
        <name>hbase.zookeeper.quorum</name>

<value>hadoop101:2181,hadoop102:2181,hadoop103:2181</va
lue>
    </property>

    <property><!--zkDate 是自定义 zookeeper 仓库-->
        <name>hbase.zookeeper.property.dataDir</name>

<value>/opt/module/zookeeper-3.4.10/zkData</value>
    </property>
</configuration>
```

3）regionservers：

```
hadoop101
hadoop102
hadoop103
```

4）软连接 hadoop 配置文件到 hbase：

```
[sun@hadoop102 module]$ ln -s
/opt/module/hadoop-2.7.2/etc/hadoop/core-site.xml
/opt/module/hbase/conf/core-site.xml
[sun@hadoop102 module]$ ln -s
/opt/module/hadoop-2.7.2/etc/hadoop/hdfs-site.xml
/opt/module/hbase/conf/hdfs-site.xml
```

1.5 HBase 远程发送到其他集群

```
[sun@hadoop102 module]$ xsync hbase/
```

1.6 HBase 服务的启动

1. 启动方式 1

```
[sun@hadoop102 hbase]$ bin/hbase-daemon.sh start master
[sun@hadoop102 hbase]$ bin/hbase-daemon.sh start regionserver
```

提示：如果集群之间的节点时间不同步，会导致 regionserver 无法启动，抛出 ClockOutOfSyncException 异常。

修复提示：

a、同步时间服务

b、属性：hbase.master.maxclockskew 设置更大的值

```
<property>
        <name>hbase.master.maxclockskew</name>
        <value>180000</value>
        <description>Time difference of regionserver from
master</description>
 </property>
```

2. 启动方式 2(HBase 没有配置环境变量)

```
[sun@hadoop102 hbase]$ bin/start-hbase.sh
```

对应的停止服务：

```
[sun@hadoop102 hbase]$ bin/stop-hbase.sh
```

1.7 查看 HBase 页面

启动成功后，可以通过“host:port”的方式来访问 HBase 管理页面，例如：

http://hadoop102:16010

五、实训方法

机房利用本机软件完成。

六、考核办法

1. 安装并部署 Hadoop、zookeeper（40 分）
2. 安装 HBase（40 分）
3. 启动 HBase（20 分）

七、思考和练习

1. HBase 安装过程中注意事项。
2. HBase 单机部署已经完成、分布式部署该如何进行？

实训三 HBase Shell 操作

一、实训目的和要求

1. 掌握启动/退出 HBase 客户端的方式；
2. 掌握 HBase Shell 操作。

二、实训内容

HBase 创建、删除表、HBase 数据增加、删除、修改、查询等。

三、实训准备

1. 操作系统：Linux（建议 Centos6.5 以上）；

2. Hadoop 版本：2.7.2；

3. JDK 版本：1.7 或以上版本；

4. Java IDE：Eclipse。

四、实训步骤

1.1 基本操作

1. 进入 HBase 客户端命令行

```
[atguigu@hadoop102 hbase]$ bin/hbase shell
```

2. 查看帮助命令

```
hbase(main):001:0> help
```

3. 查看当前数据库中有哪些表

```
hbase(main):002:0> list
```

1.2 表的操作

1. 创建表

```
hbase(main):002:0> create 'student','info'
```

2. 插入数据到表

```
hbase(main):003:0> put 'student','1001','info:sex','male'
hbase(main):004:0> put 'student','1001','info:age','18'
hbase(main):005:0> put 'student','1002','info:name','Janna'
hbase(main):006:0> put 'student','1002','info:sex','female'
hbase(main):007:0> put 'student','1002','info:age','20'
```

3. 扫描查看表数据

```
hbase(main):008:0> scan 'student'
hbase(main):009:0> scan 'student',{STARTROW => '1001', STOPROW  => '1001'}
hbase(main):010:0> scan 'student',{STARTROW => '1001'}
```

4. 查看表结构

```
hbase(main):011:0> describe  ‘student’
```

5. 更新指定字段的数据

```
hbase(main):012:0> put 'student','1001','info:name','Nick'
hbase(main):013:0> put 'student','1001','info:age','100'
```

6. 查看“指定行”或“指定列族:列”的数据

```
hbase(main):014:0> get 'student','1001'
hbase(main):015:0> get 'student','1001','info:name'
```

7. 统计表数据行数

```
hbase(main):021:0> count 'student'
```

8. 删除数据

删除某 rowkey 的全部数据：

```
hbase(main):016:0> deleteall 'student','1001'
```

删除某 rowkey 的某一列数据：

```
hbase(main):017:0> delete 'student','1002','info:sex'
```

9. 清空表数据

```
hbase(main):018:0> truncate 'student'
```

提示：清空表的操作顺序为先 disable，然后再 truncate。

10. 删除表

首先需要先让该表为 disable 状态：

hbase(main):019:0> disable 'student'

然后才能 drop 这个表：

hbase(main):020:0> drop 'student'

提示：如果直接 drop 表，会报错：ERROR: Table student is enabled. Disable it first.

11. 变更表信息

将 info 列族中的数据存放 3 个版本：

hbase(main):022:0> alter 'student',{NAME=>'info',VERSIONS=>3}

hbase(main):022:0> get 'student','1001',{COLUMN=>'info:name',VERSIONS=>3}

五、实训方法

机房利用本机软件完成。

六、考核办法

1. HBase 表创建（20 分）
2. HBase 数据添加、修改（30 分）
3. HBase 数据查询、删除（30 分）
4. Hbase 表结构修改、删除（20 分）

七、思考和练习

1. HBase 删除表时应注意什么？
2. HBase 进入/退出客户端的方式。

实训四 HBase 分布式部署

一、实训目的和要求

1. 了解分布式概念；
2. 掌握 HBase 分布式部署。

二、实训内容

HBase 分布式。

三、实训准备

5. 操作系统：Linux（建议 Centos6.5 以上）；
6. Hadoop 版本：2.7.2；

7. JDK 版本：1.7 或以上版本；

8. Java IDE：Eclipse。

四、实训步骤

1. 解压 Hbase 安装包到“/usr/local/src”路径，并修改解压后文件夹名为 hbase，截图并保存结果；

2. 设置 Hbase 环境变量，并使环境变量只对当前 root 用户生效，截图并保存结果；

3. 修改 Hbase 相应配置文件，截图并保存结果；

4. 把 Hadoop 的相应文件放到 hbase/conf 下，截图并保存结果；

5. 启动 Hbase 并保存命令输出结果，截图并保存结果；

6. 创建 Hbase 数据库表，截图并保存结果；

7. 将给定数据导入数据库表中，截图并保存结果；

8. 查看 Hbase 版本信息，截图并保存结果。

五、实训方法

机房利用本机软件完成。

六、考核办法

评分细则如下：

评价项	分值	评分细则
Hbase 解压	10 分	没有正确解压到指定位置扣 10 分
设置环境变量	10 分	环境变量不正确扣 10 分
Hbase 配置文件	20 分	环境变量每少一个扣 3 分
Hadoop 相关包	10 分	复制包每少一个扣 2 分
启动 hbase	10 分	不能正常启动扣 10 分
Hbase 数据库	10 分	未按要求建立表结构扣 4 分
导入数据	20 分	数据导入不成功每条数据扣 2 分
Hbase 版本信息	10 分	无法查看版本信息扣 10 分

七、思考和练习

1. HBase1.3 与 Hadoop 版本兼容性。

2. HBase 数据模型。

3. 思考 HBase 为什么被称为列族型数据库？

实训五 HBase 数据结构

一、实训目的和要求

1. 掌握 HBase 数据结构；

2. 了解 HBase 数据原理。

二、实训内容

RowKey、Column Family、Cell、Time Stamp、命名空间等。

三、实训准备

1. 操作系统：Linux（建议 Centos6.5 以上）；

2. Hadoop 版本：2.7.2；

3. JDK 版本：1.7 或以上版本；

4. Java IDE：Eclipse。

四、实训步骤

1.1 RowKey

与 nosql 数据库们一样,RowKey 是用来检索记录的主键。访问 HBASE table 中的行，只有三种方式：

1.通过单个 RowKey 访问

2.通过 RowKey 的 range（正则）

3.全表扫描

RowKey 行键 (RowKey)可以是任意字符串(最大长度是 64KB，实际应用中长度一般为 10-100bytes)，在 HBASE 内部，RowKey 保存为字节数组。存储时，数据按照 RowKey 的字典序(byte order)排序存储。设计 RowKey 时，要充分排序存储这个特性，将经常一起读取的行存储放到一起。(位置相关性)

1.2 Column Family

列族：HBASE 表中的每个列，都归属于某个列族。列族是表的 schema 的一部分(而列不是)，必须在使用表之前定义。列名都以列族作为前缀。例如 courses:history，courses:math 都属于 courses 这个列族。

1.3 Cell

由{rowkey, column Family:columu, version} 唯一确定的单元。cell 中的数据是没有类型的，全部是字节码形式存贮。

关键字：无类型、字节码

1.4 Time Stamp

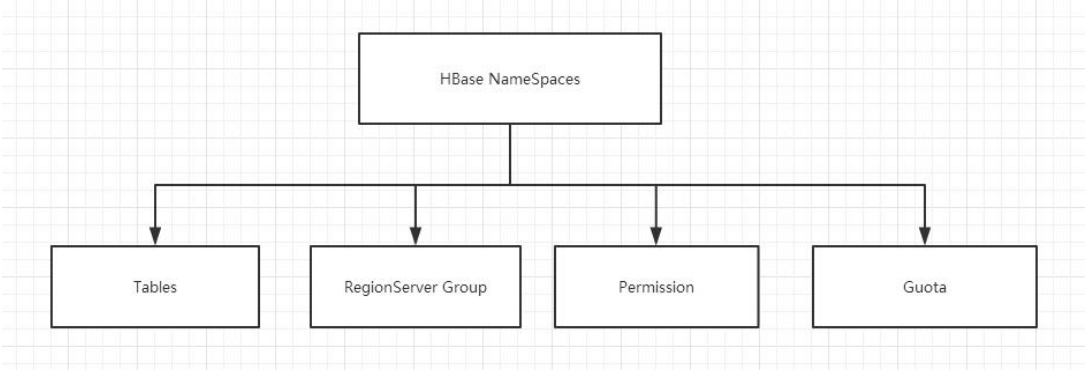
HBASE 中通过 rowkey 和 columns 确定的为一个存贮单元称为 cell。每个 cell 都保存 着同一份数据的多个版本。版本通过时间戳来索引。时间戳的类型是 64 位整型。时间戳可以由 HBASE(在数据写入时自动)赋值，此时时间戳是精确到毫秒 的当前系统时间。时间戳也可以由客户显式赋值。如果应用程序要避免数据版 本冲突，就必须自己生成具有唯一性的时间戳。每个 cell 中，不同版本的数据按照时间倒序排序，即最新的数据排在最前面。

为了避免数据存在过多版本造成的的管理 (包括存贮和索引)负担，HBASE

提供 了两种数据版本回收方式。一是保存数据的最后 n 个版本，二是保存最近一段 时间内的版本（比如最近七天）。用户可以针对每个列族进行设置。

1.5 命名空间

命名空间的结构:



1) Table：表，所有的表都是命名空间的成员，即表必属于某个命名空间，如果没有指定，则在 default 默认的命名空间中。

2) RegionServer group：一个命名空间包含了默认的 RegionServer Group。

3) Permission：权限，命名空间能够让我们来定义访问控制列表 ACL（Access Control List）。例如，创建表，读取表，删除，更新等等操作。

4) Quota：限额，可以强制一个命名空间可包含的 region 的数量。

五、实训方法

机房利用本机软件完成。

六、考核办法

1. 创建 Student 表（30 分）

2. 画出 Student 表数据结构（70 分）

七、思考和练习

1. 思考 HBase 表存储模型？

2. HBase 为什么被称为列族型数据库？

实训六 HBase 原理

一、实训目的和要求

1. 掌握 HBase 读、写流程；

2. 掌握数据 Flush 过程；

3. 掌握数据合并过程。

二、实训内容

HBase 读、写流程、数据 Flush、数据合并过程等。

三、实训准备

1. 操作系统：Linux（建议 Centos6.5 以上）；

2. Hadoop 版本：2.7.2；

3. JDK 版本：1.7 或以上版本；

4. Java IDE：Eclipse。

四、实训步骤

HBase 原理

1.1 读流程

HBase 读数据流程如图 3 所示

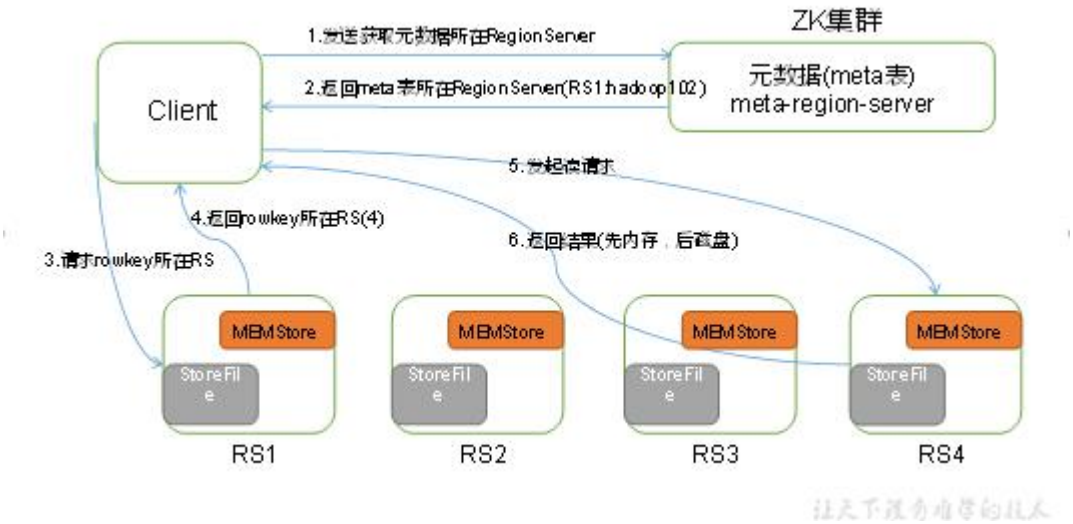


图 3 所示 HBase 读数据流程

1）Client 先访问 zookeeper，从 meta 表读取 region 的位置，然后读取 meta 表中的数据。meta 中又存储了用户表的 region 信息；

2）根据 namespace、表名和 rowkey 在 meta 表中找到对应的 region 信息；

3）找到这个 region 对应的 regionserver；

4）查找对应的 region；

5）先从 MemStore 找数据，如果没有，再到 BlockCache 里面读；

6）BlockCache 还没有，再到 StoreFile 上读(为了读取的效率)；

7）如果是从 StoreFile 里面读取的数据，不是直接返回给客户端，而是先写入 BlockCache，再返回给客户端。

1.2 写流程

Hbase 写流程如图 2 所示

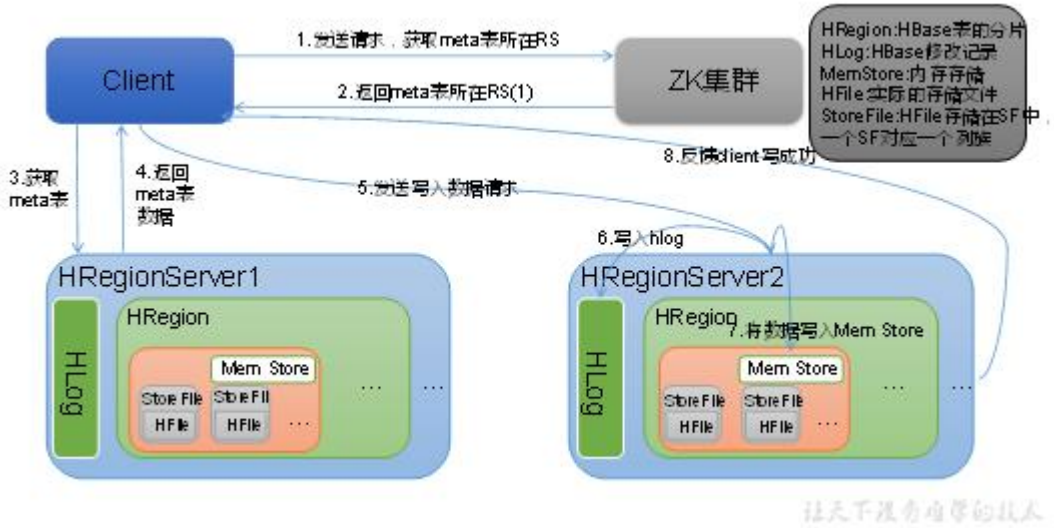


图 2　HBase 写数据流程

1）Client 向 HregionServer 发送写请求；

2）HregionServer 将数据写到 HLog（write ahead log）。为了数据的持久化和恢复；

3）HregionServer 将数据写到内存（MemStore）；

4）反馈 Client 写成功。

1.3 数据 Flush 过程

1）当 MemStore 数据达到阈值（默认是 128M，老版本是 64M），将数据刷到硬盘，将内存中的数据删除，同时删除 HLog 中的历史数据；

2）并将数据存储到 HDFS 中；

3）在 HLog 中做标记点。

1.4 数据合并过程

1）当数据块达到 4 块，Hmaster 触发合并操作，Region 将数据块加载到本地，进行合并；

2）当合并的数据超过 256M，进行拆分，将拆分后的 Region 分配给不同的 HregionServer 管理；

3）当 HregionServer 宕机后，将 HregionServer 上的 hlog 拆分，然后分配给不同的 HregionServer 加载，修改.META.；

4）注意：HLog 会同步到 HDFS。

五、实训方法

机房利用本机软件完成。

六、考核办法

1. 画出 HBase 读、写流程图（50 分）

2. 画出数据合并过程（50 分）

七、思考和练习

无

实训七 HBase 优化

一、实训目的和要求

1. 了解高可用、预分区概念；

2. 掌握 HBase 优化分类。

二、实训内容

高可用、预分区、RowKey 设计、内存优化、基础优化等。

三、实训准备

1. 操作系统：Linux（建议 Centos6.5 以上）；

2. Hadoop 版本：2.7.2；

3. JDK 版本：1.7 或以上版本；

4. Java IDE：Eclipse。

四、实训步骤

1.1 高可用

在 HBase 中 Hmaster 负责监控 RegionServer 的生命周期,均衡 RegionServer 的负载，如果 Hmaster 挂掉了，那么整个 HBase 集群将陷入不健康的状态，并且此时的工作状态并不会维持太久。所以 HBase 支持对 Hmaster 的高可用配置。

1．关闭 HBase 集群（如果没有开启则跳过此步）

2．在 conf 目录下创建 backup-masters 文件

3．在 backup-masters 文件中配置高可用 HMaster 节点

4．将整个 conf 目录 scp 到其他节点

5．打开页面测试查看

http://hadooo102:16010

1.2 预分区

每一个 region 维护着 startRow 与 endRowKey，如果加入的数据符合某个 region 维护的 rowKey 范围，则该数据交给这个 region 维护。那么依照这个原则，我们可以将数据所要投放的分区提前大致的规划好，以提高 HBase 性能。

1．手动设定预分区

2．生成 16 进制序列预分区

3．按照文件中设置的规则预分区

4．使用 JavaAPI 创建预分区

1.3 RowKey 设计

一条数据的唯一标识就是 rowkey，那么这条数据存储于哪个分区，取决于 rowkey 处于哪个一个预分区的区间内，设计 rowkey 的主要目的 ，就是让数据均匀的分布于所有的 region 中，在一定程度上防止数据倾斜。接下来我们就谈一谈 rowkey 常用的设计方案。

1．生成随机数、hash、散列值

2．字符串反转

3．字符串拼接

1.4 内存优化

HBase 操作过程中需要大量的内存开销，毕竟 Table 是可以缓存在内存中的，一般会分配整个可用内存的 70%给 HBase 的 Java 堆。但是不建议分配非常大的堆内存，因为 GC 过程持续太久会导致 RegionServer 处于长期不可用状态，一般 16~48G 内存就可以了，如果因为框架占用内存过高导致系统内存不足，框架一样会被系统服务拖死。

1.5 基础优化

1. 允许在 HDFS 的文件中追加内容
2. 优化 DataNode 允许的最大文件打开数
3. 优化延迟高的数据操作的等待时间
4. 优化数据的写入效率
5. 设置 RPC 监听数量
6. 优化 HStore 文件大小
7. 优化 hbase 客户端缓存
8. 指定 scan.next 扫描 HBase 所获取的行数
9. flush、compact、split 机制

当 MemStore 达到阈值，将 Memstore 中的数据 Flush 进 Storefile；compact 机制则是把 flush 出来的小文件合并成大的 Storefile 文件。split 则是当 Region 达到阈值，会把过大的 Region 一分为二。

五、实训方法

机房利用本机软件完成。

六、考核办法

1. 设计 HBase 表，设计可行性（30 分）
2. 进行正确的内存优化（30 分）
3. 进行正确的基础优化（40 分）

七、思考和练习

1. RowKey 设计的原则？
2. 优化数据的写入效率和哪些因素有关？

实训八 HBase 实战之 JJ 微博

一、实训目的和要求

在学生学习 HBase 数据库基础操作之后，进行一个 JJ 微博数据表设计，帮助学生总结和完善基础知识。

二、实训内容

JJ 微博数据表设计等。

三、实训准备

9. 操作系统：Linux（建议 Centos6.5 以上）；

10. Hadoop 版本：2.7.2；

11. JDK 版本：1.7 或以上版本；

12. Java IDE：Eclipse。

四、实训步骤

1.1 需求分析

1）微博内容的浏览，数据库表设计

2）用户社交体现：关注用户，取关用户

3）拉取关注的人的微博内容

1.2 HBase 表设计

1.2.1 数据库设计总览：

1）创建命名空间以及表名的定义

2）创建微博内容表

3）创建用户关系表

4）创建用户微博内容接收邮件表

5）发布微博内容

6）添加关注用户

7）移除（取关）用户

8）获取关注的人的微博内容

9）测试

1.2.3 创建微博内容表

表结构：

方法名	creatTableeContent
Table Name	weibo:content
RowKey	用户 ID_时间戳
ColumnFamily	info
ColumnLabel	标题,内容,图片
Version	1 个版本

1.2.4 创建用户关系表

表结构：

方法名	createTableRelations
Table Name	weibo:relations
RowKey	用户 ID
ColumnFamily	attends、fans
ColumnLabel	关注用户 ID，粉丝用户 ID
ColumnValue	用户 ID
Version	1 个版本

1.2.5 创建微博收件箱表
表结构：

方法名	createTableReceiveContentEmails
Table Name	weibo:receive_content_email
RowKey	用户 ID
ColumnFamily	info
ColumnLabel	用户 ID
ColumnValue	取微博内容的 RowKey
Version	1000

1.2.6 发布微博内容
a、微博内容表中添加 1 条数据
b、微博收件箱表对所有粉丝用户添加数据
1.2.7 添加关注用户
a、在微博用户关系表中，对当前主动操作的用户添加新关注的好友
b、在微博用户关系表中，对被关注的用户添加新的粉丝
c、微博收件箱表中添加所关注的用户发布的微博
1.2.8 移除（取关）用户
a、在微博用户关系表中，对当前主动操作的用户移除取关的好友(attends)
b、在微博用户关系表中，对被取关的用户移除粉丝
c、微博收件箱中删除取关的用户发布的微博
1.2.9 获取关注的人的微博内容
a、从微博收件箱中获取所关注的用户的微博 RowKey
b、根据获取的 RowKey，得到微博内容

五、实训方法

机房利用本机软件完成。

六、考核办法

1. 需求分析文档（20 分）
2. 数据库表设计（20 分）
3. 数据库表创建（30 分）
4. 数据库表设计合理性（30 分）

七、思考和练习

无

实训九 总结

一、实训目的和要求

将学生制作的作品进行综合的考核，并进行总结。

二、实训内容

1. 对学生作品进行考核。

2. 选择典型的（优秀的和劣质的）作品分别进行总结。

三、实训准备

1. 操作系统：Linux（建议 Centos6.5 以上）；

2. Hadoop 版本：2.7.2；

3. JDK 版本：1.7 或以上版本；

4. Java IDE：Eclipse。

四、实训步骤

1. 对学生的作品依次进行综合考核。

2. 抽取典型（优秀和劣质）的作品进行全面的解析。

五、实训方法

机房利用本机软件完成。

六、考核办法

系统文档	10 分
数据库设计	20 分
数据表创建	30 分
数据库调试	10 分
实训出勤	10 分
HBase 优化	20 分

七、思考和练习

无