

第 5 章

HBase 的安装和基础编程

HBase 是一个高可靠、高性能、面向列、可伸缩的分布式数据库,是谷歌公司 BigTable 的开源实现,主要用来存储非结构化和半结构化的松散数据。HBase 可用于处理非常庞大的表,可以通过水平扩展的方式,利用廉价计算机集群处理超过 10 亿行数据和数百万列元素组成的数据表。Hadoop 安装以后,本身并不包含 HBase,需要另外安装。

本章首先介绍 HBase 的安装方法,并介绍 HBase 的两种不同模式的配置方法,包括单机模式和伪分布式模式;然后介绍一些操作 HBase 的常用的 Shell 命令;最后,介绍如何使用 Eclipse 开发可以操作 HBase 数据库的 Java 应用程序。

5.1 安装 HBase

本节介绍 HBase 的安装方法,包括下载安装文件、配置环境变量、添加用户权限等。

5.1.1 下载安装文件

HBase 是 Hadoop 生态系统中的一个组件,但是,Hadoop 安装以后,本身并不包含 HBase,因此需要单独安装 HBase。登录 Linux 系统,在 Linux 系统中(不是 Windows 系统)中打开火狐浏览器,访问本书官网的“下载专区”,“软件”目录下下载 HBase 安装文件 hbase-2.5.4-bin.tar.gz。火狐浏览器会默认把下载文件都保存到当前用户的下载目录,由于本书全部采用 hadoop 用户登录 Linux 系统,因此 hbase-2.5.4-bin.tar.gz 文件会被保存到/home/hadoop/Downloads/目录下。

需要注意的是,如果是在 Windows 系统下面下载安装文件 hbase-2.5.4-bin.tar.gz,则需要上传到 Linux 系统的/home/hadoop/Downloads/目录下,这个目录是本书所有安装文件的中转站。

下载完安装文件以后,需要对文件进行解压。按照 Linux 系统使用的默认规范,用户安装的软件一般都是存放在/usr/local/目录下。使用 hadoop 用户登录 Linux 系统,打开一个终端,执行如下命令:

```
$ sudo tar -zxvf ~/Downloads/hbase-2.5.4-bin.tar.gz -C /usr/local
```

将解压的文件名 hbase-2.5.4 改为 hbase,以方便使用,命令如下:

```
$ cd /usr/local
$ sudo mv hbase-2.5.4 hbase
```

5.1.2 配置环境变量

将 HBase 安装目录下的 bin 目录(即/usr/local/hbase/bin)添加到系统的 PATH 环境变量中,这样,每次启动 HBase 时就不需要到/usr/local/hbase 目录下执行启动命令,方便 HBase 的使用。使用 vim 编辑器打开 ~/.bashrc 文件,命令如下:

```
$ vim ~/.bashrc
```

打开.bashrc 文件以后,可以看到,已经存在如下所示的 PATH 环境变量的配置信息,因为之前在第3章安装配置 Hadoop 时,已经为 Hadoop 添加了 PATH 环境变量的配置信息:

```
export PATH=$PATH:/usr/local/hadoop/sbin:/usr/local/hadoop/bin
```

这里,需要把 HBase 的 bin 目录/usr/local/hbase/bin 追加到 PATH 中。当要在 PATH 中继续加入新的路径时,只要用英文冒号(:)隔开,把新的路径加到后面即可,追加后的结果如下:

```
export PATH = $PATH:/usr/local/hadoop/sbin:/usr/local/hadoop/bin:/usr/local/hbase/bin
```

添加后,执行如下命令使设置生效:

```
$ source ~/.bashrc
```

5.1.3 添加用户权限

需要为当前登录 Linux 系统的 hadoop 用户添加访问 HBase 目录的权限,将 HBase 安装目录下的所有文件的所有者改为 hadoop,命令如下:

```
$ cd /usr/local
$ sudo chown -R hadoop:hadoop ./hbase
```

5.1.4 查看 HBase 版本信息

可以通过如下命令查看 HBase 版本信息,以确认 HBase 已经安装成功:

```
$ /usr/local/hbase/bin/hbase version
```

执行上述命令以后,如果出现图 5-1 所示信息,则说明安装成功。

```
hadoop@ubuntu:/usr/local/hbase$ /usr/local/hbase/bin/hbase version
HBase 2.5.4
Source code repository git://buildbox.localdomain/home/apurtell/tmp/RM/hbase revision=2
e426ab69d126e683577b6e94f890800c5122910
Compiled by apurtell on Thu Apr 6 09:11:53 PDT 2023
From source with checksum 74cef211e0f229be66c12578a49953af5d5a91d7bfaa3fe59fe7db81d6a21
7c007f3387640891a59dcaee40b0667b2ed7dbb2c73ddea80129128f6d2ea05fdbb
```

图 5-1 查看 HBase 版本信息

5.2 HBase 的配置

HBase 有 3 种运行模式，即单机模式、伪分布式模式和分布式模式。

(1) 单机模式：采用本地文件系统存储数据。

(2) 伪分布式模式：采用伪分布式模式的 HDFS 存储数据。

(3) 分布式模式：采用分布式模式的 HDFS 存储数据。

本节仅介绍单机模式和伪分布式模式，分布式模式在下一节介绍。

在进行 HBase 配置之前，需要确认已经安装了 3 个组件：JDK、Hadoop、SSH。HBase 单机模式不需要安装 Hadoop，伪分布式模式和分布式模式需要安装 Hadoop。JDK、Hadoop 和 SSH 的安装方法，已经在第 3 章中做了详细介绍，如果已经按照第 3 章的方法安装了 Hadoop，则这里不需要另外安装 JDK、Hadoop 和 SSH。

5.2.1 单机模式配置

1. 配置 hbase-env.sh 文件

使用 vim 编辑器打开 /usr/local/hbase/conf/hbase-env.sh，命令如下：

```
$ vim /usr/local/hbase/conf/hbase-env.sh
```

打开 hbase-env.sh 文件以后，需要在 hbase-env.sh 文件中配置 Java 环境变量，在第 3 章中，已经配置了 JAVA_HOME=/usr/lib/jvm/jdk1.8.0_371，这里可以直接复制该配置信息到 hbase-env.sh 文件中。此外，还需要添加 ZooKeeper 配置信息，配置 HBASE_MANAGES_ZK 为 true，表示由 HBase 自己管理 ZooKeeper，不需要单独的 ZooKeeper，由于 hbase-env.sh 文件中本来就存在这些变量的配置，因此只需要删除前面的注释 # 符号并修改配置内容即可，修改后的 hbase-env.sh 文件应该包含如下两行信息：

```
export JAVA_HOME=/usr/lib/jvm/jdk1.8.0_371
export HBASE_MANAGES_ZK=true
```

修改完成以后，保存 hbase-env.sh 文件并退出 vim 编辑器。

2. 配置 hbase-site.xml 文件

使用 vim 编辑器打开并编辑 /usr/local/hbase/conf/hbase-site.xml 文件，命令如下：

```
$ vim /usr/local/hbase/conf/hbase-site.xml
```

在 hbase-site.xml 文件中，需要设置属性 hbase.rootdir，用于指定 HBase 数据的存储位置，如果没有设置，则 hbase.rootdir 默认为 /tmp/hbase- \${user.name}，这意味着每次重启系统都会丢失数据。这里把 hbase.rootdir 设置为 HBase 安装目录下的 hbase-tmp 文件夹，即 /usr/local/hbase/hbase-tmp，修改后的 hbase-site.xml 文件中的配置信息如下：

```
<configuration>
  <property>
    <name>hbase.rootdir</name>
    <value>file:///usr/local/hbase/hbase-tmp</value>
  </property>
</configuration>
```

保存 hbase-site.xml 文件,并退出 vim 编辑器。

3. 启动运行 HBase

现在就可以测试运行 HBase,命令如下:

```
$ cd /usr/local/hbase
$ bin/start-hbase.sh          #启动 HBase
$ bin/hbase shell             #进入 HBase Shell 命令行模式
```

进入 HBase Shell 命令行模式以后,用户可以通过输入 Shell 命令操作 HBase 数据库。成功启动 HBase 后会出现图 5-2 所示的界面。

```
hadoop@hadoop01:/usr/local/hbase$ bin/hbase shell
HBase Shell
Use "help" to get list of supported commands.
Use "exit" to quit this interactive shell.
For Reference, please visit: http://hbase.apache.org/2.0/book.html#shell
Version 2.5.4, r2e426ab69d126e683577b6e94f890800c5122910, Thu Apr 6 09:11:53 PDT 2023
Took 0.0081 seconds
hbase:001:0>
```

图 5-2 进入 HBase Shell 模式

最后,可以使用如下命令停止 HBase 运行:

```
$ bin/stop-hbase.sh
```

需要说明的是,如果在操作 HBase 的过程中发生错误,可以查看 {HBASE_HOME} 目录(即 /usr/local/hbase)下的 logs 子目录中的日志文件,查找可能的错误原因,然后搜索网络资料寻找相关解决方案。

5.2.2 伪分布式配置

HBase 的底层数据存储需要借助于 Hadoop 的 HDFS。当 HBase 采用伪分布式模式时,底层的 Hadoop 也要采用伪分布式模式。

1. 配置 hbase-env.sh 文件

使用 vim 编辑器打开 /usr/local/hbase/conf/hbase-env.sh,命令如下:

```
$ vim /usr/local/hbase/conf/hbase-env.sh
```

打开 hbase-env.sh 文件以后,需要在 hbase-env.sh 文件中配置 JAVA_HOME、HBASE_CLASSPATH 和 HBASE_MANAGES_ZK。其中,HBASE_CLASSPATH 设置为本机

Hadoop 安装目录下的 conf 目录(即/usr/local/hadoop/conf)。JAVA_HOME 和 HBASE_MANAGES_ZK 的配置方法和上面单机模式的配置方法相同。修改后的 hbase-env.sh 文件应该包含如下 3 行信息：

```
export JAVA_HOME=/usr/lib/jvm/jdk1.8.0_371
export HBASE_CLASSPATH=/usr/local/hbase/conf
export HBASE_MANAGES_ZK=true
```

修改完成以后,保存 hbase-env.sh 文件并退出 vim 编辑器。

2. 配置 hbase-site.xml 文件

使用 vim 编辑器打开并编辑/usr/local/hbase/conf/hbase-site.xml 文件,命令如下：

```
$ vim /usr/local/hbase/conf/hbase-site.xml
```

在 hbase-site.xml 文件中,需要设置属性 hbase.rootdir,用于指定 HBase 数据的存储位置。在 HBase 伪分布式模式中,是使用伪分布式模式的 HDFS 存储数据,因此需要把 hbase.rootdir 设置为 HBase 在 HDFS 上的存储路径,根据第 3 章 Hadoop 伪分布式模式的配置可以指导,HDFS 的访问路径为 hdfs://localhost:9000/,因为这里设置 hbase.rootdir 为 hdfs://localhost:9000/hbase。此外,由于采用了伪分布式模式,因此还需要将属性 hbase.cluter.distributed 设置为 true。修改后的 hbase-site.xml 文件中的配置信息如下：

```
<configuration>
  <property>
    <name>hbase.rootdir</name>
    <value>hdfs://localhost:9000/hbase</value>
  </property>
  <property>
    <name>hbase.cluster.distributed</name>
    <value>true</value>
  </property>
  <property>
    <name>hbase.unsafe.stream.capability.enforce</name>
    <value>>false</value>
  </property>
</configuration>
```

保存 hbase-site.xml 文件,并退出 vim 编辑器。

3. 启动运行 HBase

首先登录 SSH,由于之前在第 3 章中已经设置了无密码登录,因此这里不需要输入密码;然后,切换至/usr/local/hadoop,启动 Hadoop,让 HDFS 进入运行状态,从而可以为 HBase 存储数据,具体命令如下：

```
$ ssh localhost
```

```
$ cd /usr/local/hadoop
$ ./sbin/start-dfs.sh
```

输入命令 `jps`, 如果能够看到 `NameNode`、`DataNode` 和 `SecondaryNameNode` 这 3 个进程, 则表示已经成功启动 Hadoop。

然后, 启动 HBase, 命令如下:

```
$ cd /usr/local/hbase
$ bin/start-hbase.sh
```

输入命令 `jps`, 如果出现以下进程, 则说明 HBase 启动成功:

```
Jps
HMaster
HQuorumPeer
NameNode
HRegionServer
SecondaryNameNode
DataNode
```

现在就可以进入 HBase Shell 模式, 命令如下:

```
$ bin/hbase shell #进入 HBase Shell 命令行模式
```

进入 HBase Shell 命令行模式以后, 用户可以通过输入 Shell 命令操作 HBase 数据库。

4. 停止运行 HBase

最后, 可以使用如下命令停止 HBase 运行:

```
$ bin/stop-hbase.sh
```

如果在操作 HBase 的过程中发生错误, 可以查看 `{HBASE_HOME}` 目录 (即 `/usr/local/hbase`) 下的 `logs` 子目录中的日志文件, 以寻找可能的错误原因。

关闭 HBase 以后, 如果不再使用 Hadoop, 就可以运行如下命令关闭 Hadoop:

```
$ cd /usr/local/hadoop
$ ./sbin/stop-dfs.sh
```

最后需要注意的是, 启动关闭 Hadoop 和 HBase 的顺序一定是: 启动 Hadoop→启动 HBase→关闭 HBase→关闭 Hadoop。

5.3 HBase 集群搭建

本节介绍 HBase 集群搭建的具体方法, 包括下载安装文件、配置环境变量和分发安装文件。HBase 的底层数据存储需要借助于 Hadoop 的 HDFS。当 HBase 采用分布式模式

时,底层的 Hadoop 也要采用分布式模式。这里假设已经完成了 Hadoop 集群的搭建。

5.3.1 下载安装文件

访问 HBase 官网 (<https://archive.apache.org/dist/hbase/2.5.4/>), 下载安装文件 hbase-2.5.4-bin.tar.gz, 保存到 Linux 系统的 /home/hadoop/Downloads/ 目录下。

下载完安装文件以后,需要对文件进行解压。按照 Linux 系统使用的默认规范,用户安装的软件一般都是存放在 /usr/local/ 目录下。使用 hadoop 用户登录 Linux 系统,打开一个终端,执行如下命令:

```
$ sudo tar -zxvf ~/Downloads/hbase-2.5.4-bin.tar.gz -C /usr/local
```

将解压的文件名 hbase-2.5.4 改为 hbase, 以方便使用, 命令如下:

```
$ cd /usr/local
$ sudo mv hbase-2.5.4 hbase
```

需要为当前登录 Linux 系统的 hadoop 用户添加访问 HBase 目录的权限, 将 HBase 安装目录下的所有文件的所有者改为 hadoop, 命令如下:

```
$ cd /usr/local
$ sudo chown -R hadoop ./hbase
```

5.3.2 配置环境变量

1. 配置 hbase-env.sh 文件

在 hadoop01 节点上使用 vim 编辑器打开 /usr/local/hbase/conf/hbase-env.sh, 命令如下:

```
$ vim /usr/local/hbase/conf/hbase-env.sh
```

在 hbase-env.sh 文件中添加如下信息:

```
export JAVA_HOME=/usr/lib/jvm/jdk1.8.0_371
export HBASE_CLASSPATH=/usr/local/hbase/conf
export HBASE_MANAGES_ZK=true
export HBASE_DISABLE_HADOOP_CLASSPATH_LOOKUP=true
```

修改完成以后,保存 hbase-env.sh 文件并退出 vim 编辑器。

在 hadoop01、hadoop02 和 hadoop03 节点上分别执行如下操作:

```
$ vim ~/.bashrc
```

在 .bashrc 文件中添加如下一行配置信息:

```
export HBASE_HOME=/usr/local/hbase
```

然后执行如下命令让配置生效：

```
$ source ~/.bashrc
```

2. 配置 hbase-site.xml 文件

在 hadoop01 节点上使用 vim 编辑器打开并编辑 /usr/local/hbase/conf/hbase-site.xml 文件, 命令如下：

```
$ vim /usr/local/hbase/conf/hbase-site.xml
```

hbase-site.xml 文件中的配置信息设置如下：

```
<configuration>
  <property>
    <!-- Hbase Web UI 登录 -->
    <name>hbase.master.info.port</name>
    <value>60010</value>
  </property>
  <property>
    <name>hbase.tmp.dir</name>
    <value>/usr/local/hbase/tmp</value>
  </property>
  <property>
    <name>hbase.unsafe.stream.capability.enforce</name>
    <value>false</value>
  </property>
  <property>
    <!-- Hbase 写入数据的目录 -->
    <name>hbase.rootdir</name>
    <value>hdfs://hadoop01:9000/hbase </value>
  </property>
  <property>
    <name>hbase.cluster.distributed</name>
    <value>true</value>
  </property>
  <property>
    <!-- list of zookeeper -->
    <name>hbase.zookeeper.quorum</name>
    <value>hadoop01,hadoop02,hadoop03</value>
  </property>
  <property>
    <!-- zookeeper 配置、日志等的存储位置 -->
    <name>hbase.zookeeper.property.dataDir</name>
    <value>/usr/local/hbase/zookeeper</value>
  </property>
</configuration>
```


3. 配置 regionservers 文件

在 hadoop01 节点上使用 vim 编辑器打开 /usr/local/hbase/conf/regionservers, 命令如下:

```
$ vim /usr/local/hbase/conf/regionservers
```

在 regionservers 文件中添加如下信息:

```
hadoop01  
hadoop02  
hadoop03
```

5.3.3 分发安装文件

在 hadoop01 节点上执行如下命令, 将 Master 节点上的 /usr/local/spark 文件夹复制到各个 Worker 节点上(即 hadoop02 和 hadoop03 节点):

```
$ cd /usr/local/  
$ tar -zcf ~/hbase.master.tar.gz ./hbase  
$ cd ~  
$ scp ./hbase.master.tar.gz hadoop02:/home/hadoop  
$ scp ./hbase.master.tar.gz hadoop03:/home/hadoop
```

在 hadoop02 和 hadoop03 节点上分别执行如下操作:

```
$ cd ~  
$ sudo rm -rf /usr/local/hbase/  
$ sudo tar -zxvf ~/hbase.master.tar.gz -C /usr/local  
$ sudo chown -R hadoop /usr/local/hbase
```

5.3.4 启动集群

HBase 的存储依赖于 HDFS, 所以要首先在 hadoop01 节点上启动 Hadoop 集群, 命令如下:

```
$ cd /usr/local/hadoop  
$ ./sbin/start-all.sh
```

然后, 在 hadoop01 节点上执行如下命令启动 HBase 集群:

```
$ cd /usr/local/hbase  
$ ./bin/start-hbase.sh
```

在 hadoop01 节点上输入 jps 命令, 如果能够看到如下进程, 说明 HBase 启动成功:

```
HMaster  
HQuorumPeer  
HRegionServer
```

现在就可以进入 HBase Shell 模式,在 hadoop01 节点执行如下命令:

```
$ bin/hbase shell #进入 HBase Shell 命令行模式
```

进入 HBase Shell 命令行模式以后,用户可以通过输入 Shell 命令操作 HBase 数据库。输入 exit 命令即可退出命令行模式,命令如下:

```
hbase> exit
```

注意: 这里退出 HBase 数据库是退出 HBase Shell,而不是停止 HBase 数据库后台运行,执行 exit 后,HBase 仍然在后台运行,如果要停止 HBase 运行,需要使用如下命令:

```
$ bin/stop-hbase.sh
```

关闭 HBase 以后,如果不再使用 Hadoop,就可以运行如下命令关闭 Hadoop:

```
$ cd /usr/local/hadoop  
$ ./sbin/stop-dfs.sh
```

最后需要注意的是,启动关闭 Hadoop 和 HBase 的顺序一定是:启动 Hadoop→启动 HBase→关闭 HBase→关闭 Hadoop。

5.4 HBase 常用 Shell 命令

在使用具体的 Shell 命令操作 HBase 数据之前,需要首先启动 Hadoop,然后再启动 HBase,并且启动 HBase Shell,进入 Shell 命令提示符状态,具体命令如下:

```
$ cd /usr/local/hadoop  
$ ./sbin/start-dfs.sh  
$ cd /usr/local/hbase  
$ ./bin/start-hbase.sh  
$ ./bin/hbase shell
```

需要注意的是,如果 HBase 采用分布式模式,则 Hadoop 也要采用分布式模式;如果 HBase 采用伪分布式模式,则 Hadoop 也要采用伪分布式模式。这里建议采用伪分布式模式,这样可以节省计算机系统资源,确保实验顺利开展。

5.4.1 在 HBase 中创建表

假设这里要创建一个表 student,该表包含 Sname、Ssex、Sage、Sdept、course 等字段。需要注意的是,在关系型数据库(如 MySQL)中,需要首先创建数据库,然后再创建表,但