目录

[1. 实验一：Linux操作系统环境设置 2](#_Toc7838)

[1.1. 实验目的 2](#_Toc30237)

[1.2. 实验要求 2](#_Toc10462)

[1.3. 实验环境 2](#_Toc1739)

[1.4. 实验过程 2](#_Toc25763)

[1.4.1. 实验任务一：配置Linux系统基础环境 2](#_Toc5555)

[1.4.2. 实验任务二：安装JAVA环境 4](#_Toc20076)

[2. 实验二 安装Hadoop 软件 7](#_Toc32115)

[2.1. 实验目标 7](#_Toc28157)

[2.2. 实验要求 7](#_Toc18419)

[2.3. 实验环境 7](#_Toc2561)

[2.4. 实验过程 7](#_Toc7156)

[2.4.1. 实验任务一：获取Hadoop安装包 7](#_Toc17911)

[2.4.2. 实验任务二：安装Hadoop软件 7](#_Toc12236)

[3. 实验三 安装单机版Hadoop系统 10](#_Toc25697)

[3.1. 实验目标 10](#_Toc13143)

[3.2. 实验要求 10](#_Toc30600)

[3.3. 实验环境 10](#_Toc29224)

[3.4. 实验过程 10](#_Toc23899)

[3.4.1. 实验任务一：配置Hadoop配置文件 10](#_Toc14457)

[3.4.2. 实验任务二：测试Hadoop本地模式的运行 10](#_Toc1832)

# 实验一：Linux操作系统环境设置

## 实验目的

完成本实验，您应该能够：

* 掌握linux操作系统环境设置
* 掌握hadoop安装的环境要求

## 实验要求

* 熟悉常用Linux操作系统命令
* 熟悉hadoop安装的环境
* 了解linux修改系统变量命令及各参数

## 实验环境

本实验所需之主要资源环境如表1-1所示。

表1-1 资源环境

|  |  |
| --- | --- |
| **服务器集群** | **单节点，机器最低配置：双核CPU、8GB内存、100G硬盘** |
| **运行环境** | CentOS 7.4 |
| **大数据平台** | H3C教学与实践管理系统 |
| **服务和组件** | 服务和组件根据实验需求安装 |

## 实验过程

### 实验任务一：配置Linux系统基础环境

#### 步骤一：查看服务器的IP地址

查看服务器的IP地址

[root@VM-M-01597287170765 ~]# ip address show

#### 步骤二：设置服务器的主机名称

设置主机名称为master。

[root@VM-M-01597287170765 ~]# hostnamectl set-hostname master

[root@VM-M-01597287170765 ~]# bash

#### 步骤三：绑定主机名与IP地址

可以将主机名称与IP地址绑定，这样就可以通过主机名来访问主机，方便记忆。同时在后面的配置文件中也可以用主机名来代替IP地址表示主机，当IP地址改变时，只要修改主机名与IP地址的绑定文件，不用在多个配置文件中去修改。主机名与IP地址的绑定文件是本地名字解析文件hosts，在/etc目录中，使用vi对该文件进行编辑。

[root@master ~]# vi /etc/hosts

在其中增加如下一行内容：

192.168.1.6 master

#### 步骤四：查看SSH服务状态

SSH 为 Secure Shell 的缩写，是专为远程登录会话和其他网络服务提供安全性的协议。一般的用法是在本地计算机安装SSH客服端，在服务器端安装SSH服务，然后本地计算机利用SSH 协议远程登录服务器，对服务器进行管理。这样可以非常方便地对多台服务器进行管理。同时在Hadoop分布式环境下，集群中的各个节点之间（节点可以看作是一台主机）需要使用SSH协议进行通信。因此Linux系统必须安装并启用SSH服务。

CentOS 7默认安装SSH服务，可以使用如下命令查看SSH的状态。

[root@master ~]# systemctl status sshd

**sshd.service - OpenSSH server daemon**

**Loaded: loaded (/usr/lib/systemd/system/sshd.service; enabled; vendor preset: enabled)**

**Active: active (running) since 六 2020-05-02 17:24:51 CST; 30min ago**

**…… 省略…..**

看到active (running)就表示SSH已经安装并启用。

#### 步骤五：关闭防火墙

Hadoop可以使用Web页面进行管理，但需要关闭防火墙，否则打不开Web页面。同时不关闭防火墙也会造成Hadoop后台运行脚本出现莫名其妙的错误。关闭命令如下：

[root@master ~]# systemctl stop firewalld

关闭防火墙后要查看防火墙的状态，确认一下。

[root@master ~]# systemctl status firewalld

**● firewalld.service - firewalld - dynamic firewall daemon**

**Loaded: loaded (/usr/lib/systemd/system/firewalld.service; enabled; vendor preset: enabled)**

**Active: inactive (dead) since 六 2020-05-02 17:58:15 CST; 3s ago**

**…… 省略…..**

看到inactive (dead)就表示防火墙已经关闭。不过这样设置后，Linux系统如果重启，防火墙仍然会重新启动。执行如下命令可以永久关闭防火墙。

[root@master ~]# systemctl disable firewalld

#### 步骤六：创建hadoop 用户

在Linux系统中root用户为超级管理员，具有全部权限，使用root用户在Linux系统中进行操作，很可能因为误操作而对Linux系统造成损害。正常的作法是创建一个普通用户，平时使用普通用户在系统进行操作，当用户需要使用管理员权限，可以使用两种方法达到目的：一种方法是使用su命令，从普通用户切换到root用户，这需要知道root用户的密码。另一种方法是使用sudo命令。用户的sudo可以执行的命令由root用户事先设置好。

注意:本系统已经创建好hadoop用户,步骤六命令无需再次运行

在本教材中使用root用户来安装Hadoop的运行环境，当Hadoop运行环境都安装配置好后，使用hadoop用户（这只是一个用户名，也可以使用其他的用户名）来运行Hadoop，实际工作中也是这样操作的。因此需要创建一个hadoop用户来使用Hadoop。创建命令如下：

[root@master ~]# useradd hadoop

设置用户 hadoop 的密码为password，由于密码太简单需要输入两次。

[root@master ~]# passwd hadoop

**Changing password for user hadoop.**

**New password:**

**BAD PASSWORD: The password is shorter than 8 characters**

**Retype new password:**

**passwd: all authentication tokens updated successfully.**

### 实验任务二：安装JAVA环境

#### 步骤一：下载JDK安装包

JDK安装包需要在Oracle官网下载，下载地址为：https://www.oracle.com/java /technologies /javase-jdk8-downloads.html，本教材采用的Hadoop 2.7.1所需要的JDK版本为JDK7以上，这里采用的安装包为jdk-8u152-linux-x64.tar.gz。

在H3C教学与实践管理平台（参见附录）进行Linux系统环境配置实训时，平台会事先准备好JDK安装包和Hadoop安装包，不需要读者去下载。

#### 步骤二：卸载自带OpenJDK

键入命令

[root@master ~]# rpm -qa | grep java

**java-1.8.0-openjdk-headless-1.8.0.131-11.b12.el7.x86\_64**

**javapackages-tools-3.4.1-11.el7.noarch**

**tzdata-java-2017b-1.el7.noarch**

**java-1.7.0-openjdk-headless-1.7.0.141-2.6.10.5.el7.x86\_64**

**java-1.7.0-openjdk-1.7.0.141-2.6.10.5.el7.x86\_64**

**java-1.8.0-openjdk-1.8.0.131-11.b12.el7.x86\_64**

**python-javapackages-3.4.1-11.el7.noarch**

图3-0 查看java

删除相关文件，键入命令

[root@master ~]# rpm -e --nodeps java-1.8.0-openjdk-headless-1.8.0.131-11.b12.el7.x86\_64

[root@master ~]# rpm -e --nodeps java-1.8.0-openjdk-1.8.0.131-11.b12.el7.x86\_64

[root@master ~]# rpm -e --nodeps java-1.7.0-openjdk-headless-1.7.0.141-2.6.10.5.el7.x86\_64

[root@master ~]# rpm -e --nodeps java-1.7.0-openjdk-1.7.0.141-2.6.10.5.el7.x86\_64

查看删除结果再次键入命令 java -version 出现以下结果表示删除成功



图3-1 查看java版本

#### 步骤三：安装JDK

Hadoop 2.7.1要求JDK的版本为1.7以上，这里安装的是JDK1.8版（即JAVA 8）。

安装命令如下，将安装包解压到/usr/local/src目录下

[root@master ~]# tar -zxvf /opt/software/jdk-8u152-linux-x64.tar.gz -C /usr/local/src

解压完成后，查看目录确认一下。可以看出JDK安装在/usr/local/src/jdk1.8.0\_152目录中。

[root@master ~]# ll /usr/local/src

**总用量 0**

**drwxr-xr-x 8 10 143 255 9月 14 2017 jdk1.8.0\_152**

[root@master ~]# mv /usr/local/src/jdk1.8.0\_152/ /usr/local/src/java

[root@master ~]# ls /usr/local/src/

**java**

* + - 1. **步骤四：设置JAVA环境变量**

在Linux中设置环境变量的方法比较多，较常见的有两种：一是配置/etc/profile文件，配置结果对整个系统有效，系统所有用户都可以使用；二是配置~/bashrc文件，配置结果仅对当前用户有效。这里使用第一种方法。

[root@master ~]# vi /etc/profile

在文件的最后增加如下两行：

# set java environment

export JAVA\_HOME=/usr/local/src/java

export PATH=$PATH:$JAVA\_HOME/bin

执行source使设置生效：

[root@master ~]# source /etc/profile

检查JAVA是否可用。

[root@master ~]# echo $JAVA\_HOME

**/usr/local/src/java/**

说明JAVA\_HOME已指向JAVA安装目录。

[root@master ~]# java -version

**java version "1.8.0\_152"**

**Java(TM) SE Runtime Environment (build 1.8.0\_152-b16)**

**Java HotSpot(TM) 64-Bit Server VM (build 25.231-b16, mixed mode)**

能够正常显示Java版本则说明JDK安装并配置成功。

# 实验二 安装Hadoop 软件

## 实验目标

完成本实验，您应该能够：

* 掌握hadoop的安装步骤
* 掌握hadoop的运行原理
* 掌握运行hadoop的基本命令

## 实验要求

* 熟悉运行hadoop的基本命令
* 熟悉hadoop的运行原理
* 熟悉hadoop的安装步骤

## 实验环境

本实验所需之主要资源环境如表1-1所示。

表1-1 资源环境

|  |  |
| --- | --- |
| **服务器集群** | **单节点，机器最低配置：双核CPU、8GB内存、100G硬盘** |
| **运行环境** | CentOS 7.4 |
| **大数据平台** | H3C教学与实践管理系统 |
| **服务和组件** | 完成前面章节的实验，其他服务和组件根据实验需求安装 |

## 实验过程

### 实验任务一：获取Hadoop安装包

Apache Hadoop各个版本的下载网址：https://archive.apache.org/dist/hadoop /common/。本教材选用的是Hadoop 2.7.1版本，安装包为hadoop-2.7.1.tar.gz。

在H3C教学与实践管理平台（参见附录）进行Hadoop安装实训时，平台会事先准备好相关安装包，不需要读者去下载。但如果在自己的计算机上进行Hadoop实训，就需要先下载Hadoop安装包，再上传到Linux系统的/opt/software目录。具体的方法见前一节“实验一 Linux操作系统环境设置”，这里就不再赘述。

### 实验任务二：安装Hadoop软件

#### 步骤一：安装Hadoop软件

安装命令如下，将安装包解压到/usr/local/src/目录下

[root@master ~]# tar -zxvf /opt/software/hadoop-2.7.1.tar.gz -C /usr/local/src/

[root@master ~]# mv /usr/local/src/hadoop-2.7.1 /usr/local/src/hadoop

解压完成后，查看目录确认一下。可以看出Hadoop安装在/usr/local/src/hadoop目录中。

[root@master ~]# ll /usr/local/src/

**总用量 0**

**drwxr-xr-x 9 10021 10021 149 6月 29 2015 hadoop**

**drwxr-xr-x 8 10 143 255 9月 14 2017 java**

查看Hadoop目录,得知Hadoop目录内容如下:

[root@master ~]# ll /usr/local/src/hadoop/

**总用量 28**

**drwxr-xr-x 2 10021 10021 194 6月 29 2015 bin**

**drwxr-xr-x 3 10021 10021 20 6月 29 2015 etc**

**drwxr-xr-x 2 10021 10021 106 6月 29 2015 include**

**drwxr-xr-x 3 10021 10021 20 6月 29 2015 lib**

**drwxr-xr-x 2 10021 10021 239 6月 29 2015 libexec**

**-rw-r--r-- 1 10021 10021 15429 6月 29 2015 LICENSE.txt**

**-rw-r--r-- 1 10021 10021 101 6月 29 2015 NOTICE.txt**

**-rw-r--r-- 1 10021 10021 1366 6月 29 2015 README.txt**

**drwxr-xr-x 2 10021 10021 4096 6月 29 2015 sbin**

**drwxr-xr-x 4 10021 10021 31 6月 29 2015 share**

其中：

bin：此目录中存放Hadoop、HDFS、YARN和MapReduce运行程序和管理软件。

etc：存放Hadoop配置文件。

include: 类似C语言的头文件

lib：本地库文件，支持对数据进行压缩和解压。

libexe：同lib

sbin：Hadoop集群启动、停止命令

share：说明文档、案例和依赖jar包。

#### 步骤二：配置Hadoop环境变量

和设置JAVA环境变量类似，修改/etc/profile文件。

[root@master ~]# vi /etc/profile

在文件的最后增加如下两行：

# set Hadoop environment

export HADOOP\_HOME=/usr/local/src/hadoop

export PATH=$PATH:$HADOOP\_HOME/bin:$HADOOP\_HOME/sbin

执行source使用设置生效：

[root@master ~]# source /etc/profile

检查设置是否生效：

[root@master ~]# hadoop

**Usage: hadoop [--config confdir] [COMMAND | CLASSNAME]**

**CLASSNAME run the class named CLASSNAME**

**…… 省略 ……**

出现上述Hadoop帮助信息就说明Hadoop已经安装好了。

#### 步骤三：修改目录所有者和所有者组

上述安装完成的Hadoop软件只能让root用户使用，要让hadoop用户能够运行Hadoop软件，需要将目录/usr/local/src的所有者改为hadoop用户。

#查看java与hadoop的用户与用户组权限

[root@master ~]# ll /usr/local/src/

**总用量 0**

**drwxr-xr-x 9 10021 10021 149 6月 29 2015 hadoop**

**drwxr-xr-x 8 10 143 255 9月 14 2017 java**

#更改所有者

[root@master ~]# chown -R hadoop:hadoop /usr/local/src

[root@master ~]# ll /usr/local/src

**总用量 0**

**drwxr-xr-x 9 hadoop hadoop 149 6月 29 2015 hadoop**

**drwxr-xr-x 8 hadoop hadoop 255 9月 14 2017 java**

/usr/local/src目录的所有者已经改为hadoop了。

# 实验三 安装单机版Hadoop系统

## 实验目标

完成本实验，您应该能够：

* 掌握hadoop的安装步骤
* 掌握hadoop的运行原理
* 掌握运行hadoop的基本命令

## 实验要求

* 熟悉运行hadoop的基本命令
* 熟悉hadoop的运行原理
* 熟悉hadoop的安装步骤

## 实验环境

本实验所需之主要资源环境如表1-1所示。

表1-1 资源环境

|  |  |
| --- | --- |
| **服务器集群** | **单节点，机器最低配置：双核CPU、8GB内存、100G硬盘** |
| **运行环境** | CentOS 7.4 |
| **大数据平台** | H3C教学与实践管理系统 |
| **服务和组件** | 完成前面章节的实验，其他服务和组件根据实验需求安装 |

## 实验过程

### 实验任务一：配置Hadoop配置文件

进入Hadoop目录

[root@master ~]# cd /usr/local/src/hadoop/

配置hadoop-env.sh文件，目的是告诉Hadoop系统JDK的安装目录。

[root@master hadoop]# vi etc/hadoop/hadoop-env.sh

在文件中查找export JAVA\_HOME这行，将其改为如下所示内容。

export JAVA\_HOME=/usr/local/src/java

这样就设置好Hadoop的本地模式，下面使用官方案例来测试Hadoop是否运行正常。

### 实验任务二：测试Hadoop本地模式的运行

#### 步骤一: 切换到hadoop用户

使用hadoop这个用户来运行Hadoop软件。

[root@master hadoop]# su - hadoop

[hadoop@master ~]$

#### 步骤二: 创建输入数据存放目录

将输入数据存放在~/input目录（hadoop用户主目录下的input目录中）。

[hadoop@master ~]$ mkdir ~/input

#### 步骤三: 创建数据输入文件

创建数据文件data.txt，将要测试的数据内容输入到data.txt文件中。

[hadoop@master ~]$ vi ~/input/data.txt

输入如下内容，保存退出。

Hello World

Hello Hadoop

Hello Huasan

#### 步骤四: 测试MapReduce运行

运行WordCount官方案例，统计data.txt文件中单词的出现频度。这个案例可以用来统计年度十大热销产品、年度风云人物、年度最热名词等。命令如下：

[hadoop@master ~]$ hadoop jar /usr/local/src/hadoop/share/hadoop/mapreduce/hadoop-mapreduce-examples-2.7.1.jar wordcount ~/input/data.txt ~/output

WARN io.ReadaheadPool: Failed readahead on ifile

EBADF: Bad file descriptor

如果出现上面的警告，是因为快速读取文件的时候，文件被关闭引起，此处警告不影响单机mapreduce的运行。

运行结果保存在~/output目录中，命令执行后查看结果：

[hadoop@master ~]$ ll ~/output

**total 4**

**-rw-r--r--. 1 hadoop hadoop 34 May 2 12:08 part-r-00000**

**-rw-r--r--. 1 hadoop hadoop 0 May 2 12:08 \_SUCCESS**

文件\_SUCCESS表示处理成功，处理的结果存放在part-r-00000文件中，查看该文件。

[hadoop@master ~]$ cat ~/output/part-r-00000

**Hadoop 1**

**Hello 3**

**Huasan 1**

**World 1**

可以看出统计结果正确，说明Hadoop本地模式运行正常。读者可将这个运行结果与“3.2.3 MapReduce”中的WordCount案例运行过程进行对照，来加深对MapReduce框架的理解。

注意：输出目录不能事先创建，如果已经有~/output目录，就要选择另外的输出目录，或者将~/output目录先删除。删除命令如下所示。

[hadoop@master ~]$ rm -rf ~/output