Turing AI 万维图灵 | 大数据系列课程

大数据技术-第二章：Hadoop运行及开发环境搭建

Hadoop伪分布式集群环境搭建（实验1）

CONTENTS

01. 项目目标

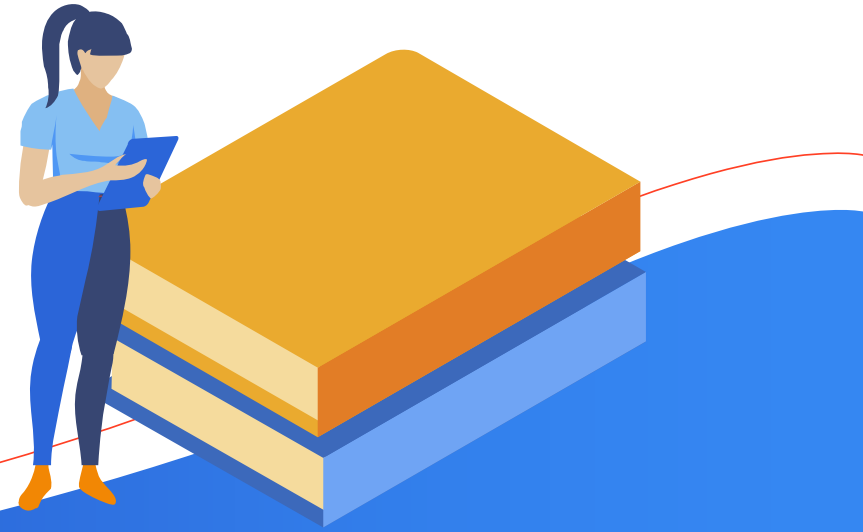
02. 项目描述

03. 知识准备

04. 项目实施

05. 知识拓展

06. 课后实训



01
项目目标

1 项目目标

01 掌握Linux操作系统环境搭建与配置

02 掌握Hadoop伪分布式搭建的步骤

03 掌握Hadoop伪分布式搭建的运行原理

02
项目描述

2 项目描述

Hadoop是一个能够对大量数据进行分布式处理的软件框架，包括并行计算模型Map/Reduce，分布式文件系统HDFS。Hadoop用于解决以下问题：

1. 海量数据的存储(HDFS)

2. 海量数据的分析(MapReduce)

3. 资源管理调度(YARN)

2 项目描述

Hadoop部署方式有以下三种：

1. 本地模式：在使用开发工具进行开发调试的时候使用的 只能启动一个map和一个reduce，适用于开发环境。
2. 伪分布式：通过一台机器模拟分布式环境，在开发和学习时使用。适用于测试环境。

3. 集群模式：真实的开发环境。

在大数据行业内Hadoop分布式处理框架主要用于对海量数据存储以及处理，本实验主要是介绍伪分布式测试环境的搭建，也是为了后续学习如何使用Hadoop奠定基础。

实验操作系统为CentOS-7-x86_64-DVD-1511，Hadoop版本为hadoop-2.6.5，JDK版本为jdk-8u201-linux-x64，远程连接工具为SecureCRTPortable，虚拟机工具为VMware-workstation-full-15.5.0-14665864。

本实验所用软件可根据文件名称在资源库内下载。

03
知识准备

为了完成本实验，学生需要提前掌握以下理论知识内容：

1. 虚拟机安装部署。
2. Linux运行环境部署。
3. Linux运行环境基础操作。
4. Hadoop安装配置。

04

项目实施

4.1 实施思路

基于项目描述与知识准备的内容，我们已经对Hadoop伪分布式集群环境搭建知识点有了一定的理解，现在我们回归Hadoop伪分布式集群搭建的案例，通过伪分布式集群搭建的知识点尝试着实现伪分布式集群搭建的案例。

我们将按照下述四个步骤来完成平台的搭建和训练。

1. Linux基础环境部署：在虚拟机中安装Linux操作系统并进行相关的配置，如Linux安装、关闭防火墙和禁用SELINUX、配置hostname与IP地址对应关系、创建用户和用户组、配置SSH免密码登录等。

2. JDK安装配置：完成JDK文件上传至系统并完成环境变量配置。

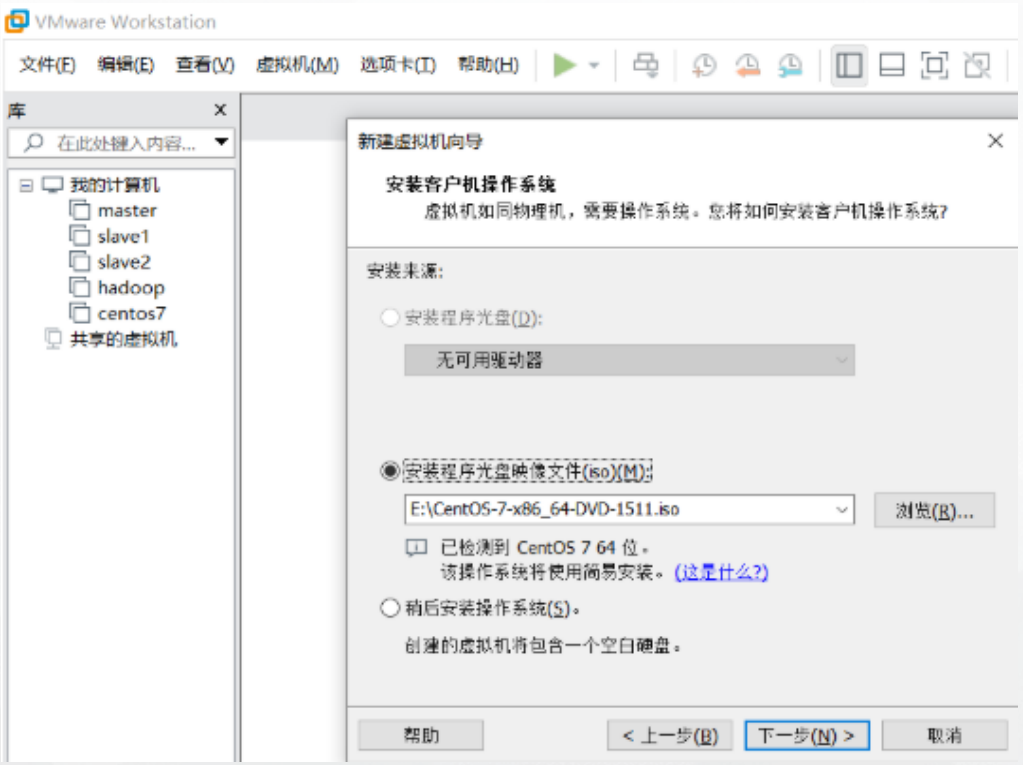
3. Hadoop安装配置：完成Hadoop安装包的上传以及相关配置文件的配置。

4. 测试运行：通过操作Hadoop平台测试平台的安装的正确性。

4.2 实施步骤

步骤一：安装配置Linux环境：

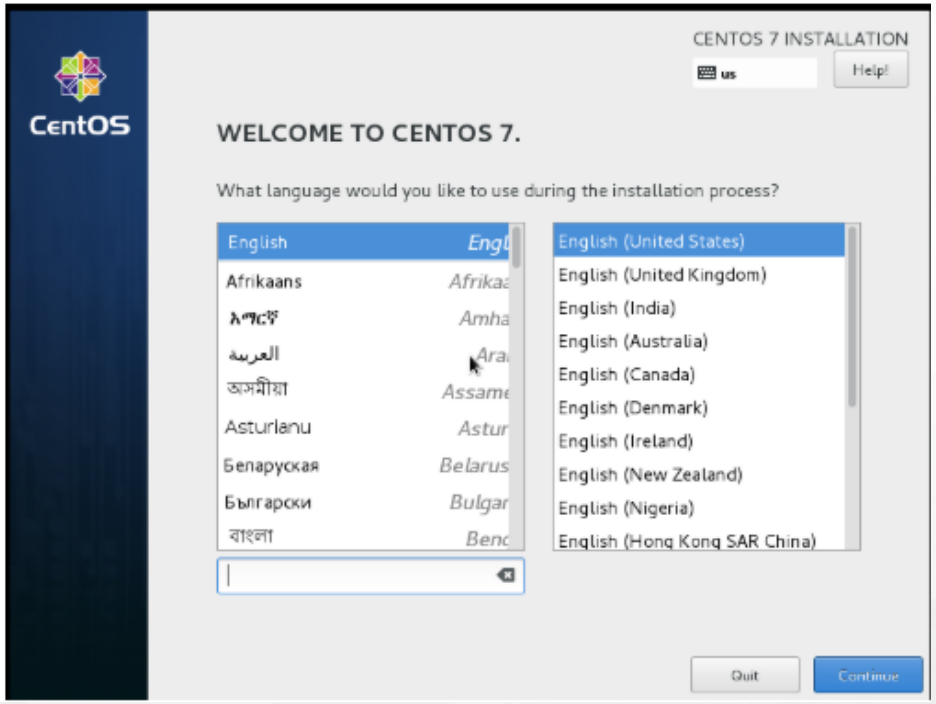
使用VMware Workstation Pro新建虚拟机安装CentOS-7-x86_64-DVD-1511系统,虚拟机配置根据自己本机配置建议设置即可



4.2 实施步骤

步骤一：安装配置Linux环境：

安装过程中选择系统语言为英语



4.2 实施步骤

步骤一：安装配置Linux环境：

执行如下操作，关闭防火墙

```
##关闭防火墙
[root@localhost ~]#  systemctl stop firewalld
##永久关闭防火墙
[root@localhost ~]#  systemctl disable firewalld
##查看防火墙状态，验证是否关闭成功
[root@localhost ~]#  systemctl status firewalld
```

执行结果如下图所示：

```
[root@master ~]#  systemctl status firewalld
● firewalld.service - firewalld - dynamic firewall daemon
   Loaded: loaded (/usr/lib/systemd/system/firewalld.service; disabled; vendor p
reset: enabled)
   Active: inactive (dead)

Feb 23 10:19:41 master systemd[1]: Starting firewalld - dynamic firewall da.....
Feb 23 10:19:42 master systemd[1]: Started firewalld - dynamic firewall daemon.
Feb 23 10:25:54 master systemd[1]: Stopping firewalld - dynamic firewall da.....
Feb 23 10:25:56 master systemd[1]: Stopped firewalld - dynamic firewall daemon.
Hint: Some lines were ellipsized, use -l to show in full.
```

4.2 实施步骤

步骤一：安装配置Linux环境：

配置网卡，进入/etc/sysconfig/network-scripts目录，根据以下内容编辑ifcfg-eno16777736文件

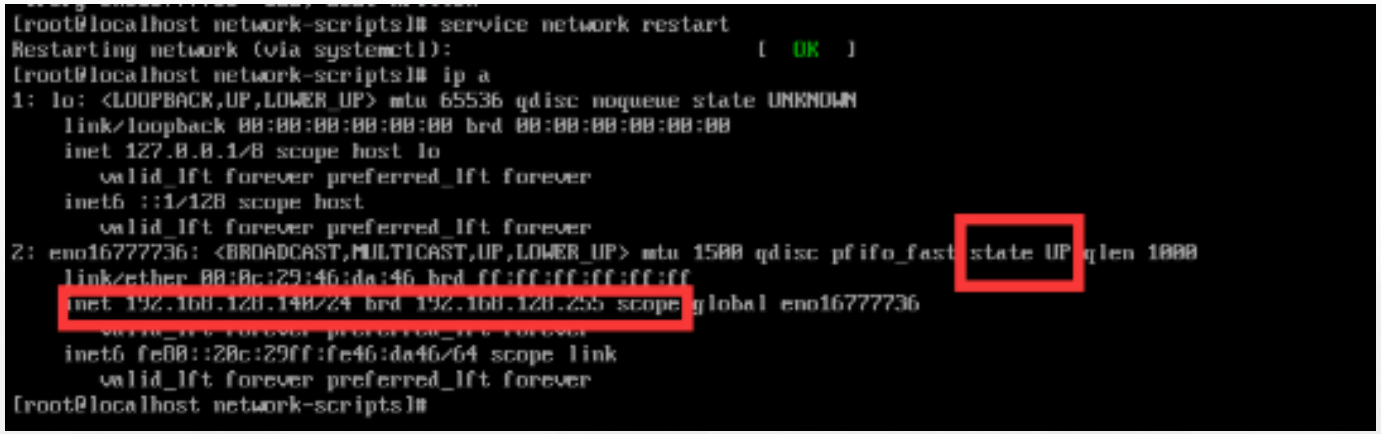
```
TYPE=Ethernet
BOOTPROTO=static
NAME=eno16777736
UUID=201e5b9d-fb4c-43a4-bf8d-084d47e5f588
DEVICE=eno16777736
ONBOOT=yes
NM_CONTTROLLED=yes
IPADDR=192.168.128.140
NETMASK=255.255.255.0
GATEWAY=192.168.128.2
DNS1=192.168.128.2
```

4.2 实施步骤

步骤一：安装配置Linux环境：

重启网络服务
[root@localhost ~]# service network restart

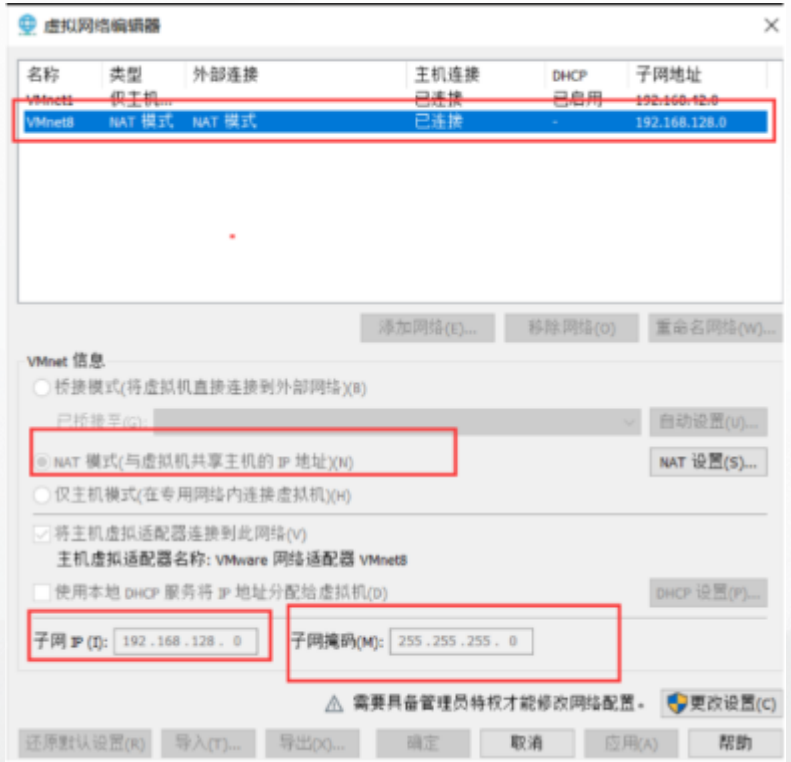
验证配置是否生效，出现下图内容表示配置成功
[root@localhost ~]# ip a



4.2 实施步骤

步骤一：安装配置Linux环境：

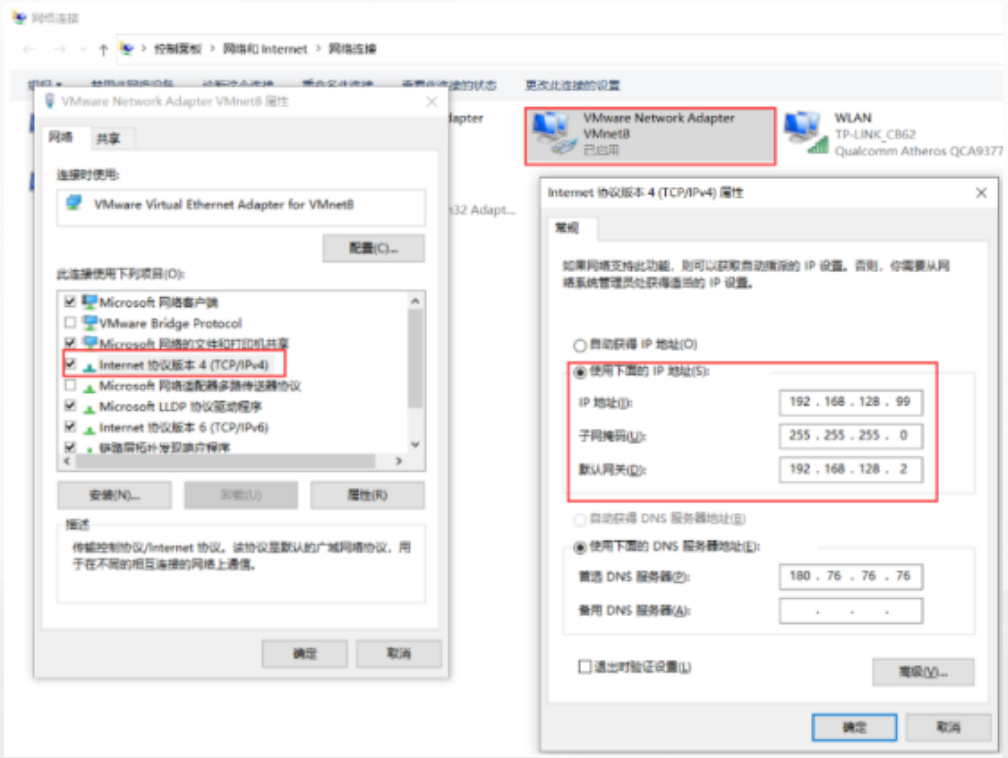
配置远程登录，设置虚拟机适配器信息，在VMware Workstation Pro菜单栏选择编辑，然后再选择虚拟网络编辑器，设置如下图



4.2 实施步骤

步骤一：安装配置Linux环境：

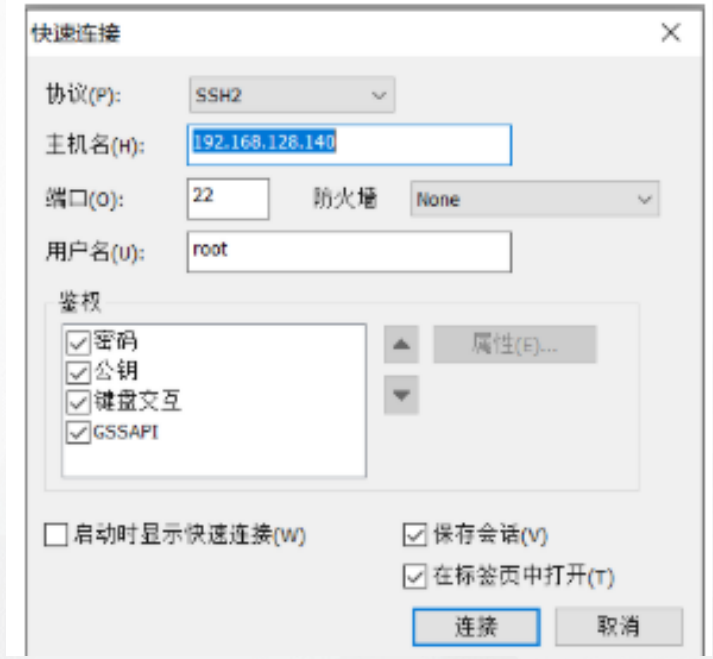
设置本机IP信息，在本机网络和Internet设置中选择VMnet8，然后根据下图设置相应内容，该处要注意IP地址必须与虚拟机网络在同一个网段内

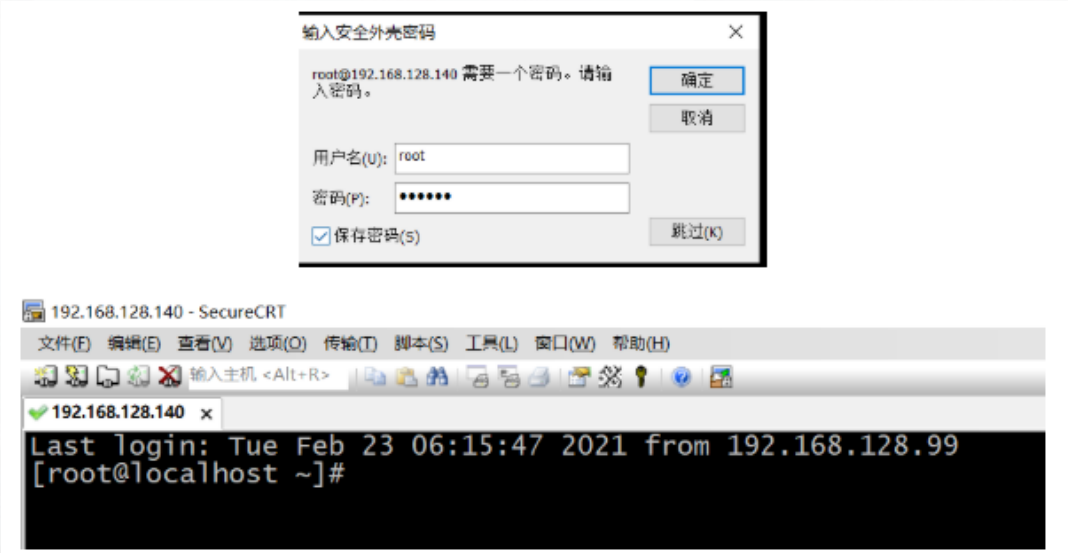


4.2 实施步骤

步骤一：安装配置Linux环境：

使用SecureCRTPortable远程登录工具连接服务器，在菜单栏选择文件内的快速连接选项填写服务器IP地址点击连接后设置用户名以及密码，点击确定登录服务器，如下图所示





4.2 实施步骤

步骤一：安装配置Linux环境：

设置hostname为master
[root@localhost ~]# hostnamectl set-hostname master

验证是否设置成功
[root@master /]# hostnamectl --static
Master

重新登录机器名已更改

进入/etc目录，编辑hosts文件，设置与IP地址映射关系

```
[root@master etc]# vi hosts
127.0.0.1   localhost localhost.localdomain localhost4 localhost4.localdomain4
::1         localhost localhost.localdomain localhost6 localhost6.localdomain6
192.168.128.140 master
```

4.2 实施步骤

步骤一：安装配置Linux环境：

创建用户

```
[root@master ~]#  useradd hadoop
[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.
```

4.2 实施步骤

步骤一：安装配置Linux环境：

配置SSH免密码登录，这里免密登陆指的是hadoop账户登陆的master，再ssh hadoop@master

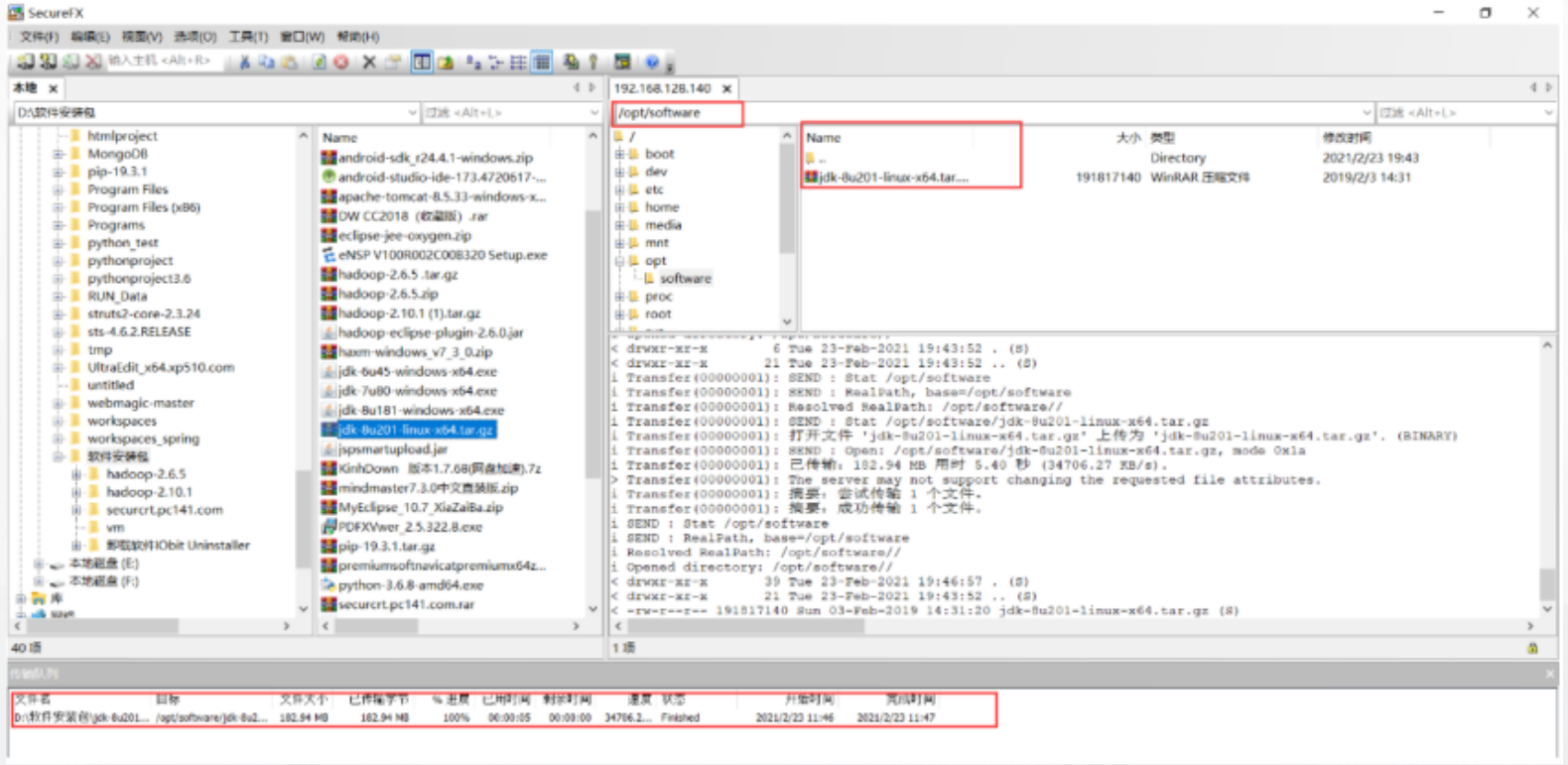
生成密钥
[hadoop@master ~]$ ssh-keygen -t rsa # 三次回车
[hadoop@master ~]$ ssh-copy-id hadoop@master # 输入密码

测试免密码登录是否成功
[hadoop@master /]$ ssh hadoop@master
Last login: Tue Feb 23 08:46:35 2021

4.2 实施步骤

步骤二：安装配置JDK：

上传解压JDK文件，使用root用户在服务器/opt目录下新建子目录software并使用SecureFXPortable工具将下载至本地的jdk-8u201-linux-x64.tar.gz安装包上传至服务器/opt/software目录，如下图所示



4.2 实施步骤

步骤二：安装配置JDK：

将上传后的压缩包进行解压,并更改文件名

```
[root@master software]# tar  -zxvf  /opt/software/jdk-8u201-linux-x64.tar.gz  -C
/usr/local/src/
[root@master src]# mv jdk1.8.0_201/ jdk8
```

配置JDK环境变量，编辑/etc/profile，添加如下内容

```
[root@master ~]# vi /etc/profile
# JAVA_HOME指向 JAVA安装目录
export JAVA_HOME=/usr/local/src/jdk8
export PATH=$PATH:$JAVA_HOME/bin  # 将 JAVA安装目录加入 PATH路径
```

执行 source 使设置生效：

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

4.2 实施步骤

步骤二：安装配置JDK：

验证安装结果

```
[root@master src]# java -version
java version "1.8.0_201"
Java(TM) SE Runtime Environment (build 1.8.0_201-b09)
Java HotSpot(TM) 64-Bit Server VM (build 25.201-b09, mixed mode)
```

能够正常显示 Java 版本则说明 JDK 安装并配置成功。

4.2 实施步骤

步骤三：安装配置Hadoop：

本次安装的是hadoop-2.6.6，使用hadoop用户安装，所以先以hadoop用户登陆。上传并解压Hadoop安装包，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-2.6.5.tar.gz -C /usr/local/src/

修改hadoop-env.sh

```
[hadoop@master /]$ cd usr/local/src/hadoop-2.6.5/etc/hadoop/
[hadoop@master hadoop]$ vi hadoop-env.sh
##修改export JAVA_HOME=${JAVA_HOME}为:
export JAVA_HOME=/usr/local/src/jdk8
```

4.2 实施步骤

步骤三：安装配置Hadoop：

修改core-site.xml配置文件

[hadoop@master hadoop]$ vi core-site.xml
添加如下配置

```
<configuration>
 <property>
                <name>fs.defaultFS</name>
                <value>hdfs://master:9000</value>
        </property>
        <property>
                <name>hadoop.tmp.dir</name>
                <value>/home/hadoop/data/hadoopdata</value>
        </property>
</configuration>
```

4.2 实施步骤

步骤三：安装配置Hadoop：

修改hdfs-site.xml配置文件配置文件

[hadoop@master hadoop]$ vi hdfs-site.xml
添加如下配置

```
<configuration>
        <property>
                <name>dfs.namenode.name.dir</name>
                <value>/home/hadoop/data/hadoopdata/name</value>
                <description>为了保证元数据的安全一般配置多个不同目录</descripti
on>
        </property>

        <property>
                <name>dfs.datanode.data.dir</name>
                <value>/home/hadoop/data/hadoopdata/data</value>
                <description>datanode 的数据存储目录</description>
        </property>

        <property>
                <name>dfs.replication</name>
                <value>2</value>
                <description>HDFS 的数据块的副本存储个数，默认是3</description>
        </property>

</configuration>
```

4.2 实施步骤

步骤三：安装配置Hadoop：

修改marped-site.xml配置文件

```
[hadoop@master hadoop]$ cp mapred-site.xml.template mapred-site.xml
[hadoop@master hadoop]$ vi mapred-site.xml
```

添加如下配置

```
<configuration>
  <property>
                <name>mapreduce.framework.name</name>
                <value>yarn</value>
        </property>
</configuration>
```

4.2 实施步骤

步骤三：安装配置Hadoop：

修改yarn-site.xml配置文件

```
[hadoop@master hadoop]$ vi yarn-site.xml
```

添加如下配置

```
<configuration>

<!-- Site specific YARN configuration properties -->
<property>
     <name>yarn.nodemanager.aux-services</name>
     <value>mapreduce_shuffle</value>
     <description>YARN 集群为 MapReduce 程序提供的 shuffle 服务</description>
</property>
</configuration>
```

4.2 实施步骤

步骤三：安装配置Hadoop：

修改slaves配置文件
在slaves文件添加节点机器名master

```
[hadoop@master hadoop]$ vi slaves
master
```

配置Hadoop环境变量并使其生效，在文件的最后增加如下两行：
[hadoop@master ~]# vi ~/.bashrc

```
# HADOOP_HOME
export HADOOP_HOME=/usr/local/src/hadoop-2.6.5
export PATH=$PATH:$HADOOP_HOME/bin:$HADOOP_HOME/sbin:
```

4.2 实施步骤

步骤三：安装配置Hadoop：

执行 source 使用设置生效

[hadoop@master ~]# source ~/.bashrc
验证版本检查设置是否生效，如下图

```
[hadoop@master hadoop]$ hadoop version
Hadoop 2.6.5
Subversion https://github.com/apache/hadoop.git -r e8c9fe0b4c252caf2ebf1464220599650f119997
Compiled by sjlee on 2016-10-02T23:43Z
Compiled with protoc 2.5.0
From source with checksum f05c9fa095a395faa9db9f7ba5d754
This command was run using /usr/local/src/hadoop-2.6.5/share/hadoop/common/hadoop-common-2.6.5.jar
```

出现上述 Hadoop 版本信息就说明 Hadoop 已经安装成功。

4.2 实施步骤

步骤三：安装配置Hadoop：

创建hdfs-site.xml里配置的路径

```
[hadoop@master hadoop]$ mkdir -p /home/hadoop/data/hadoopdata/name
[hadoop@master hadoop]$ mkdir -p /home/hadoop/data/hadoopdata/data
```

格式化namenode，在控制台输入bin/hadoop namenode –format命令格式化namenode

```
[hadoop@master hadoop]$ hadoop namenode -format
```

4.2 实施步骤

步骤三：安装配置Hadoop：

如图所示出现status 0即为初始化成功。

```
21/02/23 09:35:33 INFO common.Storage: Storage directory /home/hadoop/data/hadoopdata/name has been successfully formatted.
21/02/23 09:35:33 INFO namenode.FSImageFormatProtobuf: Saving image file /home/hadoop/data/hadoopdata/name/current/fsimage.ckpt_0000000000000000000 using no compression
21/02/23 09:35:33 INFO namenode.FSImageFormatProtobuf: Image file /home/hadoop/data/hadoopdata/name/current/fsimage.ckpt_0000000000000000000 of size 323 bytes saved in 0 seconds.
21/02/23 09:35:33 INFO namenode.NNStorageRetentionManager: Going to retain 1 images with txid >= 0
21/02/23 09:35:33 INFO util.ExitUtil: Exiting with status 0
21/02/23 09:35:33 INFO namenode.NameNode: SHUTDOWN_MSG:
/************************************************************
SHUTDOWN_MSG: Shutting down NameNode at master/192.168.128.140
************************************************************/
```

启动Hadoop伪分布式集群并查看启动进程，在控制台输入启动命令：sbin/start-all.sh，启动Hadoop伪分布式集群

```
[hadoop@master /]$ cd /usr/local/src/hadoop-2.6.5/
[hadoop@master hadoop-2.5.6]$ sbin/start-all.sh
```

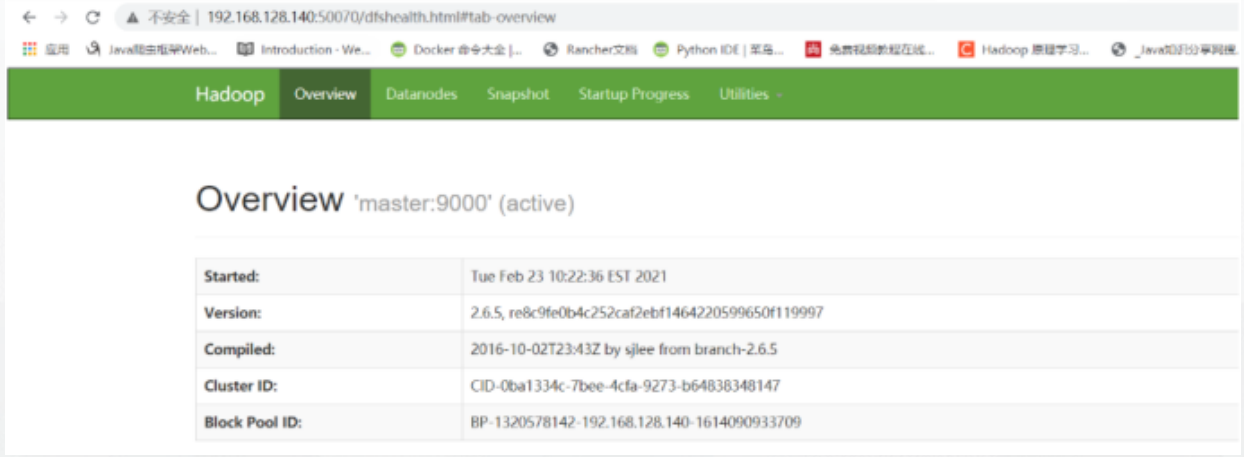
4.2 实施步骤

步骤三：安装配置Hadoop：

查看启动进程是否成果，运行jps命令，出现以下进程表示启动成功

```
[hadoop@master hadoop-2.6.5]$ jps
3715 NodeManager
3622 ResourceManager
4023 Jps
3225 NameNode
3339 DataNode
3485 SecondaryNameNode
```

浏览器访问http:/ip:50070



4.2 实施步骤

步骤四：测试运行：

新建测试文件，并输入以下内容

```
[hadoop@master ~]$ vi test.txt
hadoop  test
hadoop  test
hadoop  test
```

将文件上传至HDFS文件系统

```
[hadoop@master ~]$ hdfs dfs  -put test.txt /test
```

4.2 实施步骤

步骤四：测试运行：

运行Hadoop自带的Wordcount程序，在命令行窗口运行以下命令执行分词例子程序

```
[hadoop@master hadoop-2.6.5]$ hadoop jar share/hadoop/mapreduce/hadoop-mapreduce-examples-2.6.5.jar
wordcount /test/test.txt /test/output
```

运行结束后查看执行结果,出现以下结果表示程序执行成功

```
[hadoop@master hadoop-2.6.5]$ hdfs dfs -cat /test/output/part-r-00000
hadoop  3
test    3
```

05
项目拓展

5 知识拓展

在前面的任务中，我们利学习了如何搭建Hadoop伪分布式集群环境，而对于在日常使用过程中平台会出现一些异常问题，这时需要我们对Hadoop的配置文件参数具有一定的理解，通过配置参数来优化和维护平台正常运行 。接下来让我们来了解下Hadoop相关配置文件的参数含义（拓展知识点）。

core-site.xml文件参数

序号	参数名	默认值	参数解释
1	fs.defaultFS	file:///	文件系统主机和端口
2	io.file.buffer.size	4096	流文件的缓冲区大小
3	hadoop.tmp.dir	/tmp/hadoop-${user.name}	临时文件夹

5 知识拓展

hdfs-site.xml文件参数

序号	参数名	默认值	参数解释
1	dfs.namenode.secondary.http-address	0.0.0.0:50090	HDFS 对应的 HTTP 服务器地址和端口
2	dfs.namenode.name.dir	file://${hadoop.tmp.dir}/dfs/name	DFS 的名称节点在本地文件系统的位置
3	dfs.datanode.data.dir	file://${hadoop.tmp.dir}/dfs/data	DFS 数据节点数据块存储在本地文件系统的位置
4	dfs.replication	3	缺省的块复制数量
5	dfs.webhdfs.enabled	true	是否通过 http 协议读取 hdfs 文件

5 知识拓展

mapred-site.xml文件参数

序号	参数名	默认值	参数解释
1	mapreduce.framework.name	local	取值 local、classic 或yarn 其中之一，如果不是yarn，则不会使用 YARN 集群来实现资源的分配
2	mapreduce.jobhistory.address	0.0.0.0:10020	历史服务器的地址和端口，通过历史服务器查看已经运行完的 Mapreduce 作业记录
3	mapreduce.jobhistory.webapp.addres	0.0.0.0:19888	历史服务器 web 应用访问的地址和端口

5 知识拓展

yarn-site.xml文件参数

序号	参数名	默认值	参数解释
1	yarn.resourcemanager.address	0.0.0.0:8032	ResourceManager 提供给客户端访问的地址。客户端通过该地址向 RM 提交应用程序，杀死应用程序等
2	yarn.resourcemanager.scheduler.address	0.0.0.0:8030	定义历史服务器的地址和端口，通过历史服务器查看已经运行完的Mapreduce 作业记录
3	yarn.resourcemanager.resource-tracker.address	0.0.0.0:8031	ResourceManager 提供给NodeManager的地址。NodeManager 通过该地址向RM 汇报心跳，领取任务等

5 知识拓展

4	yarn.resourcemanager.admin.address	0.0.0.0:8033	ResourceManager 提供给管理员的访问地址。管理员通过该地址向 RM 发送管理命令等
5	yarn.resourcemanager.webapp.address	0.0.0.0:8088	ResourceManager 对 web 服务提供地址。用户可通过该地址在浏览器中查看集群各类信息
6	yarn.nodemanager.aux-services	org.apache.hadoop.mapred.ShuffleHandler	通过该配置项，用户可以自定义一些服务，例如 Map-Reduce 的 shuffle 功能就是采用这种方式实现的，这样就可以在 NodeManager 上扩展自己的服务。

06

项目实训

课后练习

6.1 实训习题

- 简答题：如何修改HDFS的数据块的副本数量？

Turing AI 万维图灵 | 大数据系列课程

大数据

BIG DATA

智 / 能 / 科 / 技　　放 / 眼 / 未 / 来