使用学院机器部署 Hadoop 单节点集群指南

1. 云平台登录与资源概览

1.1 登录云平台

a) 打开浏览器，使用校园网访问学院云平台地址：

```
http://172.19.240.2:5000
```

b) 点击"切换到用户登录"，选择"AD/LDAP用户"

- **账户**：您的学号（如 MF1932001）
- **密码**：南京大学统一身份认证密码

c) 登录后选择项目 `prj-MF1932XXX` 进入主界面

1.2 资源配额说明

- **CPU**：2核心
- **内存**：4GB
- **存储**：100GB

2. 创建云主机

2.1 创建步骤

a) 点击"云资源池" → "云主机" → "创建云主机"

b) 配置云主机参数：

- **名称**：自定义（如 hadoop-node）
- **计算规格**：选择 2C4G
- **镜像**：推荐选择 Ubuntu 服务器版
- **网络类型**：IPv4（校内地址）
- **控制台密码**：务必设置，确保主机安全

2.2 镜像选择建议

- **Ubuntu 服务器版**：推荐用于 Hadoop 部署
- **默认凭据**：
  - 用户名：`root`
  - 密码：`password`（请及时修改）

3. 云主机配置与管理

3.1 访问云主机

a) 在云主机列表中找到创建的实例，点击"打开控制台"
b) 使用控制台密码登录系统

3.2.1 方式一：控制台访问

- 

- 初始用户名：root
- 初始密码：password
- 

3.2.2 方式二：ssh访问

- 创建主机时 ssh 选择密码
-

- 方式一登录后（修改配置文件）：
  - sudo vim /etc/ssh/sshd_config
  - #PermitRootLogin prohibit-password 改为 PermitRootLogin yes
  - #PasswordAuthentication yes 改为 PasswordAuthentication yes
  - #PubkeyAuthentication yes 改为 PubkeyAuthentication yes
  - 重启ssh服务：sudo systemctl restart ssh
- 本地ssh连接：ssh root@云ip地址
  - 初始名：root
  - 初始密码：password

3.2 资源调整

a) **CPU/内存调整**：

- 在云主机详情页直接修改 CPU 核数和内存大小

b) **云盘扩容**：

- 点击"云主机操作" → "系统扩容"
- 输入新容量（注意：不支持缩小）

3.3 网络配置

a) **校园网认证**：

```
# 上网认证
curl -d '{"username":"学号", "password":"密码", "domain":"default"}'
https://p.nju.edu.cn/api/portal/v1/login

# 退出认证
curl -d '{"domain":"default"}' https://p.nju.edu.cn/api/portal/v1/logout
```

4. Hadoop 环境准备

4.1 系统更新

```
# 更新包列表
apt update

# 升级系统包
apt upgrade -y
```

4.2 安装必需软件

```
# 安装 OpenJDK 8
apt install openjdk-8-jdk -y

# 安装 SSH 和 pdsh
sudo apt-get install ssh -y
sudo apt-get install pdsh -y

# 启动 SSH 服务
sudo systemctl start ssh
sudo systemctl enable ssh

# 验证 SSH 服务状态
sudo systemctl status ssh

# 配置 PDSH 使用 SSH（重要：避免启动 Hadoop 服务时出错）
echo 'export PDSH_RCMD_TYPE=ssh' >> ~/.bashrc
```

```
source ~/.bashrc
```

4.3 配置 Java 环境

```
# 设置环境变量
echo 'export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64' >> /etc/environment
source /etc/environment

# 验证安装
java -version
```

5. Hadoop 安装与配置

5.1 创建 Hadoop 用户

```
# 创建 hadoop 用户
adduser hadoop
usermod -aG sudo hadoop

# 切换到 hadoop 用户
su - hadoop
```

5.2 下载和安装 Hadoop

```
# 切换到用户目录
cd /home/hadoop

# 下载 Hadoop（使用阿里云镜像加速）
wget https://mirrors.aliyun.com/apache/hadoop/common/hadoop-3.4.2/hadoop-3.4.2.tar.gz

# 解压并重命名
tar -xzf hadoop-3.4.2.tar.gz
mv hadoop-3.4.2 hadoop
rm hadoop-3.4.2.tar.gz

# 设置权限
sudo chown -R hadoop:hadoop ~/hadoop
```

5.3 配置环境变量

```
# 编辑 .bashrc
sudo nano ~/.bashrc

# 添加以下内容
export HADOOP_HOME=/home/hadoop/hadoop
export HADOOP_CONF_DIR=$HADOOP_HOME/etc/hadoop
export PATH=$PATH:$HADOOP_HOME/bin:$HADOOP_HOME/sbin
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64
export PDSH_RCMD_TYPE=ssh

# 重新加载配置
source ~/.bashrc
```

5.4 配置 Hadoop

5.4.1 配置 hadoop-env.sh

```
echo 'export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64' >>
$HADOOP_HOME/etc/hadoop/hadoop-env.sh
```

5.4.2 配置 core-site.xml

```
# 编辑 core-site.xml
nano $HADOOP_HOME/etc/hadoop/core-site.xml
```

添加以下配置内容：

```
<configuration>
    <property>
        <name>fs.defaultFS</name>
        <value>hdfs://localhost:9000</value>
        <description>默认文件系统 URI</description>
    </property>
    <property>
        <name>hadoop.tmp.dir</name>
        <value>/home/hadoop/hadoop-data/tmp</value>
        <description>Hadoop 临时目录（使用独立磁盘避免填满根分区）</description>
    </property>
</configuration>
```

5.4.3 配置 hdfs-site.xml

```
# 编辑 hdfs-site.xml
nano $HADOOP_HOME/etc/hadoop/hdfs-site.xml
```

添加以下配置内容：

```
<configuration>
    <property>
        <name>dfs.replication</name>
        <value>1</value>
        <description>数据块副本数量（单节点设置为1）</description>
    </property>
    <property>
        <name>dfs.namenode.name.dir</name>
        <value>/home/hadoop/hadoop-data/data/namenode</value>
        <description>NameNode 数据存储目录（使用独立磁盘）</description>
    </property>
    <property>
        <name>dfs.datanode.data.dir</name>
        <value>/home/hadoop/hadoop-data/data/datanode</value>
        <description>DataNode 数据存储目录（使用独立磁盘）</description>
    </property>
</configuration>
```

5.4.4配置 mapred-site.xml

```
# 编辑 mapred-site.xml
nano $HADOOP_HOME/etc/hadoop/mapred-site.xml
```

添加以下配置内容：

```
<configuration>
    <property>
        <name>mapreduce.framework.name</name>
        <value>yarn</value>
        <description>MapReduce 框架名称</description>
    </property>
    <property>
        <name>yarn.app.mapreduce.am.env</name>
        <value>HADOOP_MAPRED_HOME=/home/hadoop/hadoop</value>
        <description>ApplicationMaster 环境变量</description>
    </property>
    <property>
        <name>mapreduce.map.env</name>
        <value>HADOOP_MAPRED_HOME=/home/hadoop/hadoop</value>
        <description>Map 任务环境变量</description>
    </property>
    <property>
        <name>mapreduce.reduce.env</name>
```

```
        <value>HADOOP_MAPRED_HOME=/home/hadoop/hadoop</value>
        <description>Reduce 任务环境变量</description>
    </property>
</configuration>
```

重要提示：

- 配置文件中只能有一个 `<configuration>` 标签
- 请根据实际的 Hadoop 安装路径调整 `HADOOP_MAPRED_HOME` 的值
- 必须使用绝对路径

5.4.5 配置 yarn-site.xml

```
# 编辑 yarn-site.xml
nano $HADOOP_HOME/etc/hadoop/yarn-site.xml
```

添加以下配置内容：

```
<configuration>
    <property>
        <name>yarn.nodemanager.aux-services</name>
        <value>mapreduce_shuffle</value>
        <description>NodeManager 辅助服务</description>
    </property>
    <property>
        <name>yarn.resourcemanager.hostname</name>
        <value>localhost</value>
        <description>ResourceManager 主机名</description>
    </property>
</configuration>
```

5.5 创建数据目录

```
# 设置数据目录权限
sudo mkdir -p /home/hadoop/hadoop-data
sudo chown -R hadoop:hadoop /home/hadoop/hadoop-data
chmod -R 755 /home/hadoop/hadoop-data

# 创建子目录
mkdir -p /home/hadoop/hadoop-data/{namenode,datanode,tmp}
```

6. SSH 无密码登录配置

```
```

```
# 生成 SSH 密钥对
ssh-keygen -t rsa -P '' -f ~/.ssh/id_rsa

# 将公钥添加到授权文件
cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys

# 设置正确的权限
chmod 0600 ~/.ssh/authorized_keys
chmod 700 ~/.ssh

# 测试无密码 SSH 连接
ssh localhost

# 如果成功，应该能够无密码登录
# 退出 SSH 会话
exit
```

7. 启动 Hadoop 集群

7.1 格式化 HDFS

```
sudo hostnamectl set-hostname hadoop-node
sudo nano /etc/hosts
# 确保有 127.0.0.1 localhost 和 127.0.1.1 hadoop-node（没有就加进去）

hdfs namenode -format
```

7.2 启动服务

```
# 启动 HDFS
start-dfs.sh

# 启动 YARN
start-yarn.sh

# 验证进程
jps
```

预期输出应包含：

- NameNode
- DataNode
- SecondaryNameNode
- ResourceManager
- NodeManager

8. 验证与测试

8.1 Web UI 访问

由于学院网络限制，建议使用 SSH 隧道访问 Web 界面：

```
# 在本地机器执行（替换为实际IP）
ssh -L 9870:localhost:9870 -L 8088:localhost:8088 hadoop@云主机IP
```

访问地址：

- **HDFS Web UI**：http://localhost:9870
- **YARN Web UI**：http://localhost:8088

8.2 基本功能测试

```
# 创建 HDFS 目录
hdfs dfs -mkdir -p /user/hadoop/input

# 上传测试文件
hdfs dfs -put $HADOOP_HOME/etc/hadoop/*.xml /user/hadoop/input

# 运行 MapReduce 示例
hadoop jar $HADOOP_HOME/share/hadoop/mapreduce/hadoop-mapreduce-examples-3.4.2.jar \
grep /user/hadoop/input /user/hadoop/output 'dfs[a-z.]+'

# 查看结果
hdfs dfs -cat /user/hadoop/output/*
```

9. 资源管理建议

9.1 节约资源措施

- **关机释放资源**：不使用时关闭云主机，释放 CPU 和内存
- **及时清理**：删除不需要的云主机和快照
- **合理使用存储**：100GB 存储需合理规划

9.2 快照管理

- 创建重要节点快照备份
- 及时删除不必要的快照释放存储空间

10. 故障排除

10.1 常见问题

```
# 检查服务状态
jps

# 查看日志
tail -f $HADOOP_HOME/logs/hadoop-*-namenode-*.log

# 检查磁盘空间
df -h

# 检查网络连接
ping localhost
```

10.2 服务管理命令

```
# 停止所有服务
stop-all.sh

# 启动所有服务
start-all.sh

# 单独管理
start-dfs.sh
stop-dfs.sh
start-yarn.sh
stop-yarn.sh
```

11. 安全注意事项

- 及时修改默认密码
- 定期更新系统和软件
- 不使用时关闭云主机
- 重要数据及时备份

注意：学院资源有限，请合理使用，避免恶意占用资源。如有问题，请及时联系管理员。