网易云创学院

**Hadoop集群部署实验手册**

[一、实验介绍 3](#_Toc531630243)

[1.1 实验内容 3](#_Toc531630244)

[1.2 实验知识点 3](#_Toc531630245)

[1.3 实验环境 3](#_Toc531630246)

[1.4 实验资源 3](#_Toc531630247)

[1.5 适合人群 3](#_Toc531630248)

[1.6 实验步骤清单 4](#_Toc531630249)

[二、实验架构 4](#_Toc531630250)

[三、实验环境准备 4](#_Toc531630251)

[四、实验步骤 4](#_Toc531630252)

[4.1 查看环境 4](#_Toc531630253)

[4.2部署Hadoop集群 5](#_Toc531630254)

[4.2.1安装hadoop 5](#_Toc531630255)

[4.2.2创建hdfs数据文件存储目录 5](#_Toc531630256)

[4.2.3修改配置文件 6](#_Toc531630257)

[4.2.4主从节点同步 11](#_Toc531630258)

[4.3测试Hadoop集群 11](#_Toc531630259)

[4.3.1启动集群 11](#_Toc531630260)

[4.3.2验证Hadoop集群 12](#_Toc531630261)

[五、实验总结 14](#_Toc531630262)

[六、参考文档 14](#_Toc531630263)

# 一、实验介绍

## 1.1 实验内容

基于开源软件搭建满足企业需求的Hadoop生态系统,构建基础的大数据分析平台。

本实验采用3台机器搭建Hadoop完全分布式集群，其中1台机器作为Master节点，另外两台机器作为Slave节点，主机名分别为Slave1和Slave2。

## 1.2 实验知识点

* Hadoop集群部署
* HDFS分布式文件系统管理

## 1.3 实验环境

* Hadoop2.7.5
* VMware Workstation 15.1.0 Pro for Windows
* 虚拟机镜像
* PC电脑（内存16G以上、硬盘100G剩余空间、CPU CORE I3以上）
* 网络环境（单机、局域网）

## 1.4 实验资源

|  |  |
| --- | --- |
| 资源名称 | 存储目录 |
| hadoop安装包 | /opt/package/software |

## 1.5 适合人群

本课程难度为一般，属于初级课程，适合具有linux及hadoop基础的用户。

## 1.6 实验步骤清单

* 检查实验环境（防火墙、hosts配置、ssh互信）
* 部署hadoop集群（安装hadoop、创建hdfs数据文件、修改配置文件、主从节点同步）
* 测试hadoop集群（启动集群、验证集群）

# 二、实验架构

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **序号** | **IP地址** | **机器名** | **类型** | **核数/内存** | **目录** |
| 1 | 192.168.80.101 | master | NameNode  ResourceManager | 1核/3G | /usr/local/ |
| 2 | 192.168.80.102 | slave1 | DataNode  NodeManager | 1核/2G |
| 3 | 192.168.80.103 | slave2 | DataNode  NodeManager | 1核/2G |

# 三、实验环境准备

启动虚拟机master、slave1和slave2的快照“大数据平台-hadoop”。

点击“Not listed?”按钮，输入用户名root密码root登录系统。

# 四、实验步骤

## 4.1 查看环境

#检查防火墙是否关闭

[root@slave1 ~]# firewall-cmd --state

not running

#检查三台虚拟机hosts文件

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

127.0.0.1 localhost localhost.localdomain localhost4 localhost4.localdomain4

::1 localhost localhost.localdomain localhost6 localhost6.localdomain6

192.168.80.101 master

192.168.80.102 slave1

192.168.80.103 slave2

#检查ssh环境

[root@master ~]# ssh slave1 date

Mon Nov 19 10:23:43 CST 2018

[root@master ~]# ssh slave2 date

Mon Nov 19 10:23:52 CST 2018

## 4.2部署Hadoop集群

### 4.2.1安装hadoop

#解压安装包

[root@master ~]# tar zxvf /opt/package/software/hadoop-2.7.5.tar.gz -C /usr/local

#重命名Hadoop安装目录

[root@master ~]# mv /usr/local/hadoop-2.7.5 /usr/local/hadoop

### 4.2.2创建hdfs数据文件存储目录

#删除并创建hdfs数据文件存储目录

[root@master ~]# rm -rf /home/hadoopdir

[root@master ~]# mkdir /home/hadoopdir

#创建临时文件存储目录

[root@master ~]# mkdir /home/hadoopdir/tmp

#创建namenode数据目录

[root@master ~]# mkdir -p /home/hadoopdir/dfs/name

#创建datanode数据目录

[root@master ~]# mkdir /home/hadoopdir/dfs/data

### 4.2.3修改配置文件

1、配置环境变量

#检查环境变量

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

export HADOOP\_INSTALL=/usr/local/hadoop

export PATH=${HADOOP\_INSTALL}/bin:${HADOOP\_INSTALL}/sbin:${PATH}

#hadoop-env.sh配置JAVA\_HOME

[root@master ~]# vim /usr/local/hadoop/etc/hadoop/hadoop-env.sh

export JAVA\_HOME=/usr/lib/jvm/jre

#验证Hadoop版本

[root@master ~]# hadoop version

Hadoop 2.7.5

2、修改core-site.xml内容参考如下

[root@master ~]# vim /usr/local/hadoop/etc/hadoop/core-site.xml

<configuration>

<property>

<name>fs.defaultFS</name>

<value>hdfs://master:9000</value>

</property>

<property>

<name>io.file.buffer.size</name>

<value>131072</value>

</property>

<property>

<name>hadoop.tmp.dir</name>

<value>file:/home/hadoopdir/tmp/</value>

<description>A base for other temporary directories.</description>

</property>

<property>

<name>hadoop.proxyuser.hadoop.hosts</name>

<value>\*</value>

</property>

<property>

<name>hadoop.proxyuser.hadoop.groups</name>

<value>\*</value>

</property>

</configuration>

3、修改hdfs-site.xml文件

[root@master ~]# vim /usr/local/hadoop/etc/hadoop/hdfs-site.xml

<configuration>

<property>

<name>dfs.namenode.name.dir</name>

<value>file:///home/hadoopdir/dfs/name</value>

</property>

<property>

<name>dfs.datanode.data.dir</name>

<value>file:///home/hadoopdir/dfs/data</value>

</property>

<property>

<name>dfs.replication</name>

<value>2</value>

</property>

<property>

<name>dfs.webhdfs.enabled</name>

<value>true</value>

</property>

</configuration>

4、修改mapred-site.xml

#复制配置文件

[root@master ~]# cp /usr/local/hadoop/etc/hadoop/mapred-site.xml.template /usr/local/hadoop/etc/hadoop/mapred-site.xml

#修改配置文件

[root@master ~]# vim /usr/local/hadoop/etc/hadoop/mapred-site.xml

<configuration>

<property>

<name>mapreduce.framework.name</name>

<value>yarn</value>

</property>

<property>

<name>mapreduce.jobhistory.address</name>

<value>master:10020</value>

</property>

<property>

<name>mapreduce.jobhistory.webapp.address</name>

<value>master:19888</value>

</property>

<property>

<name>mapreduce.jobtracker.http.address</name>

<value>master:50030</value>

</property>

<property>

<name>mapred.job.tracker</name>

<value>master:9001</value>

</property>

</configuration>

5、修改 yarn-site.xml

[root@master ~]# vim /usr/local/hadoop/etc/hadoop/yarn-site.xml

<configuration>

<!-- Site specific YARN configuration properties -->

<property>

<name>yarn.nodemanager.aux-services</name>

<value>mapreduce\_shuffle</value>

</property>

<property>

<name>yarn.nodemanager.aux-services.mapreduce.shuffle.class</name>

<value>org.apache.hadoop.mapred.ShuffleHandler</value>

</property>

<property>

<name>yarn.resourcemanager.hostname</name>

<value>master</value>

</property>

<property>

<name>yarn.resourcemanager.address</name>

<value>master:8032</value>

</property>

<property>

<name>yarn.resourcemanager.scheduler.address</name>

<value>master:8030</value>

</property>

<property>

<name>yarn.resourcemanager.resource-tracker.address</name>

<value>master:8031</value>

</property>

<property>

<name>yarn.resourcemanager.admin.address</name>

<value>master:8033</value>

</property>

<property>

<name>yarn.resourcemanager.webapp.address</name>

<value>master:8088</value>

</property>

</configuration>

6、修改 slaves文件

[root@master ~]# vim /usr/local/hadoop/etc/hadoop/slaves

slave1

slave2

7、初始化HDFS

[root@master ~]# hadoop namenode -format

18/11/19 11:27:07 INFO util.ExitUtil: Exiting with status 0

18/11/19 11:27:07 INFO namenode.NameNode: SHUTDOWN\_MSG:

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

SHUTDOWN\_MSG: Shutting down NameNode at master/192.168.80.101

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

备注：最后出现“util.ExitUtil: Exiting with status 0”，表示成功。

### 4.2.4主从节点同步

1、同步/usr/local/hadoop目录文件到slave节点

[root@master ~]# scp -r /usr/local/hadoop slave1:/usr/local/

[root@master ~]# scp -r /usr/local/hadoop/ slave2:/usr/local/

2、同步/home/hadoopdir目录文件到slave节点

#删除目录

[root@master ~]# ssh slave1 rm -rf /home/hadoopdir

[root@master ~]# ssh slave2 rm -rf /home/hadoopdir

#同步目录

[root@master ~]# scp -r /home/hadoopdir slave1:/home/

[root@master ~]# scp -r /home/hadoopdir slave2:/home/

3、同步环境信息

[root@master ~]# scp /etc/profile slave1:/etc/profile

[root@master ~]# scp /etc/profile slave2:/etc/profile

[root@slave1 ~]# source /etc/profile

[root@slave2 ~]# source /etc/profile

## 4.3测试Hadoop集群

### 4.3.1启动集群

#启动hadoop集群

[root@master ~]# start-all.sh

This script is Deprecated. Instead use start-dfs.sh and start-yarn.sh

Starting namenodes on [master]

master: starting namenode, logging to /usr/local/hadoop/logs/hadoop-root-namenode-master.out

slave1: starting datanode, logging to /usr/local/hadoop/logs/hadoop-root-datanode-slave1.out

slave2: starting datanode, logging to /usr/local/hadoop/logs/hadoop-root-datanode-slave2.out

Starting secondary namenodes [0.0.0.0]

The authenticity of host '0.0.0.0 (0.0.0.0)' can't be established.

ECDSA key fingerprint is SHA256:lrhnnND23cf0F9Azp4qUwS+Ek6+LscJ28CRce/NofA0.

ECDSA key fingerprint is MD5:56:6b:86:5e:df:6f:4f:70:af:fc:3f:d2:81:c8:a8:e6.

Are you sure you want to continue connecting (yes/no)? yes

0.0.0.0: Warning: Permanently added '0.0.0.0' (ECDSA) to the list of known hosts.

0.0.0.0: starting secondarynamenode, logging to /usr/local/hadoop/logs/hadoop-root-secondarynamenode-master.out

starting yarn daemons

starting resourcemanager, logging to /usr/local/hadoop/logs/yarn-root-resourcemanager-master.out

slave1: starting nodemanager, logging to /usr/local/hadoop/logs/yarn-root-nodemanager-slave1.out

slave2: starting nodemanager, logging to /usr/local/hadoop/logs/yarn-root-nodemanager-slave2.out

### 4.3.2验证Hadoop集群

1、JPS查看Java进程

#master

[root@master ~]# jps

7779 Jps

7349 SecondaryNameNode

7499 ResourceManager

7134 NameNode

#slave1

[root@slave1 ~]# jps

3169 DataNode

3445 Jps

3277 NodeManager

#slave2

[root@slave2 ~]# jps

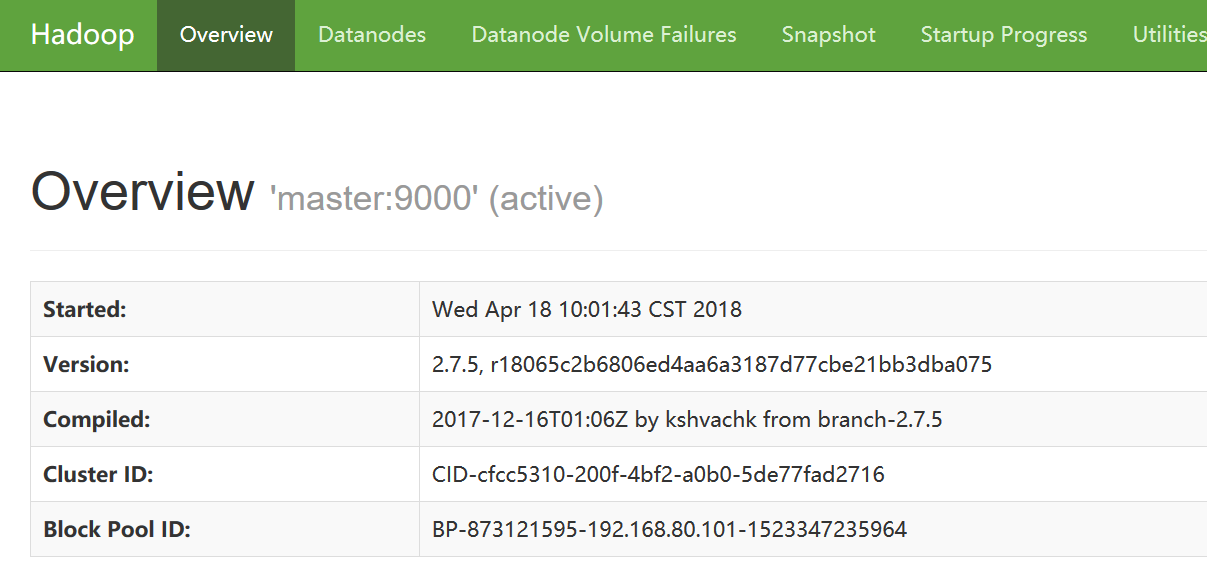
3270 NodeManager

3162 DataNode

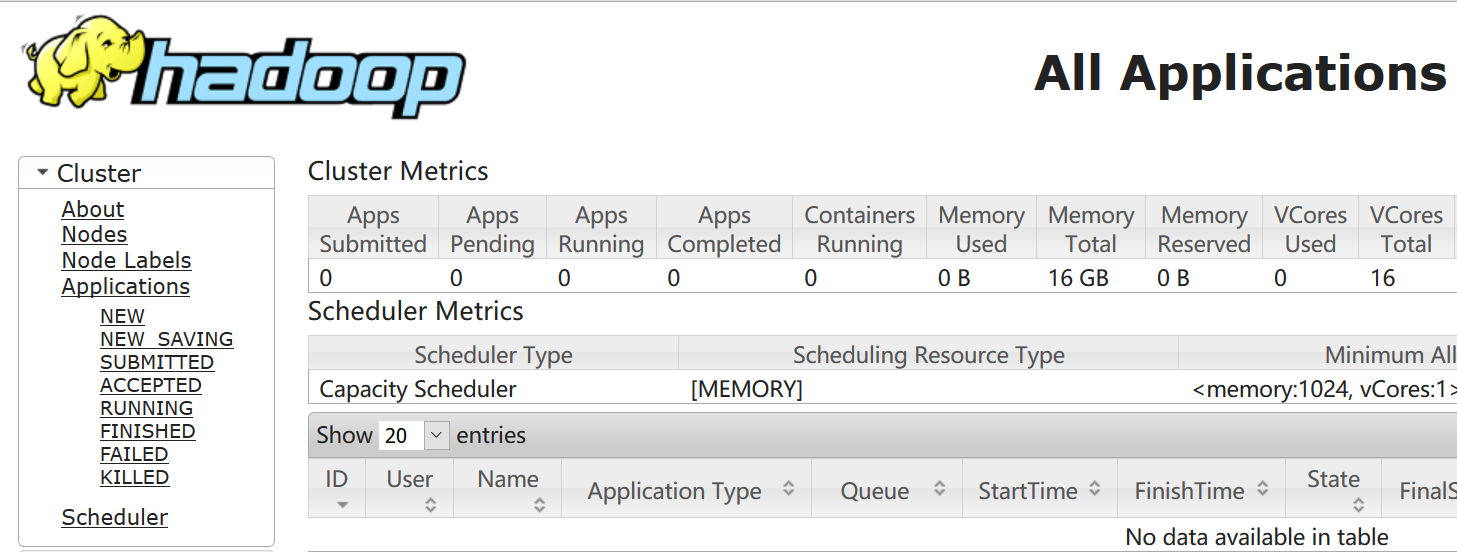
3391 Jps

2、登录网页查看

打开浏览器，登录http://master:50070



打开浏览器，查看yarn环境，登录<http://master:8088>



# 五、实验总结

本次实验采用完全分布式集群安装方式，需要提前部署JDK环境、SSH验证等过程。安装并启动后可以访问 Web 界面 [http://localhost:50070](http://localhost:50070/) 查看 NameNode 和 Datanode 信息，还可以在线查看 HDFS 中的文件。

# 六、参考文档

本实验参考下列文档内容制作：

* [http://hadoop.apache.org/docs/r2.7.5/hadoop-project-dist/hadoop-common/SingleCluster.html](http://hadoop.apache.org/docs/r2.6.0/hadoop-project-dist/hadoop-common/SingleCluster.html)