

## 4.Hadoop单机版环境搭建

---

### 一. 案例信息

---

#### 1. 实验内容

Hadoop的安装部署的模式一共有三种：

- 本地模式，默认的模式，无需运行任何守护进程（daemon），所有程序都在单个JVM上执行。由于在本机模式下测试和调试MapReduce程序较为方便，因此，这种模式适宜用在开发阶段。使用本地文件系统，而不是分布式文件系统。
- 伪分布模式，在一台主机模拟多主机。即，Hadoop的守护程序在本地计算机上运行，模拟集群环境，并且是相互独立的Java进程。在这种模式下，Hadoop使用的是分布式文件系统，各个作业也是由JobTraker服务，来管理的独立进程。在单机模式之上增加了代码调试功能，允许检查内存使用情况，HDFS输入输出，以及其他的守护进程交互。类似于完全分布式模式，因此，这种模式常用来开发测试Hadoop程序的执行是否正确。
- 全分布模式，完全分布模式的守护进程运行在由多台主机搭建的集群上，是真正的生产环境。在所有的主机上安装JDK和Hadoop，组成相互连通的网络。

本案例采用伪分布模式搭建Hadoop，在一台主机模拟多主机，用于后续的程序开发。

#### 2. 实验目的

- 掌握Hadoop单机版的搭建及配置方法
- 掌握HDFS文件系统的开启及关闭方法
- 掌握Yarn的开启及关闭方法
- 掌握Hadoop平台的基本使用

#### 3. 实验环境

- hadoop == 3.1.0
- CentOS == 7.3
- jdk == 1.8

---

## 二. 实验指导

---

#### 1. 关联技术

- 环境准备
  - 文件解包解压
  - JDK安装配置
  - VIM文本编辑命令
  - 文件上传
  - 系统环境配置
- 配置Hadoop
  - XML配置
  - 防火墙配置

- jps命令
- 使用内置程序计算PI值
  - jar包执行

## 2. 实验步骤

- 环境准备
- 配置Hadoop
- 初始化并启动Hadoop
- Hadoop(YARN)环境搭建
- 使用内置程序计算PI值

## 3. 实验效果

# 三. 实验操作

## 01. 步骤一：环境准备

### 步骤操作说明

### 1. 配置JDK

- 下载JDK，登录官方<https://www.oracle.com/java/technologies/downloads/#java8> 下载所需版本的JDK，版本为JDK 1.8

JDK 8u311 checksum

**Linux** macOS Solaris Windows

Product/file description	File size	Download
ARM 64 RPM Package	59.25 MB	<a href="#">jdk-8u311-linux-aarch64.rpm</a>
ARM 64 Compressed Archive	71 MB	<a href="#">jdk-8u311-linux-aarch64.tar.gz</a>
ARM 32 Hard Float ABI	73.69 MB	<a href="#">jdk-8u311-linux-arm32-vfp-hflt.tar.gz</a>
x86 RPM Package	110.22 MB	<a href="#">jdk-8u311-linux-i586.rpm</a>
x86 Compressed Archive	139.61 MB	<a href="#">jdk-8u311-linux-i586.tar.gz</a>
x64 RPM Package	109.97 MB	<a href="#">jdk-8u311-linux-x64.rpm</a>
x64 Compressed Archive	140 MB	<a href="#">jdk-8u311-linux-x64.tar.gz</a>

- 上传JDK至服务器
    - 使用yum下载时，可以先修改软件源（非必须）：  
[https://help.aliyun.com/document\\_detail/405635.html?spm=5176.smart-service\\_service\\_chat.0.0.712c3f1bBoZ19l](https://help.aliyun.com/document_detail/405635.html?spm=5176.smart-service_service_chat.0.0.712c3f1bBoZ19l)
    - 创建tools目录，用于存放文件
- ```
mkdir /opt/tools
```
- 切换至tools目录，上传JDK安装包
  - 解压JDK安装包
    - 创建server目录，用于存放JDK解压后的文件

```
mkdir /opt/server
```

- 解压至server目录

```
tar -zxvf jdk-8u131-linux-x64.tar.gz -C /opt/server
```

- 配置环境变量

- 编辑 /etc/profile 文件

```
vim /etc/profile  
# 文件末尾增加  
export JAVA_HOME=/opt/server/jdk1.8.0_131  
export PATH=${JAVA_HOME}/bin:$PATH
```

- 执行source命令，使配置立即生效

```
source /etc/profile
```

- 检查是否安装成功

```
java -version
```

```
[root@node01 jdk1.8.0_131]# java -version  
java version "1.8.0_131"  
Java(TM) SE Runtime Environment (build 1.8.0_131-b11)  
Java HotSpot(TM) 64-Bit Server VM (build 25.131-b11, mixed mode)
```

## 2. 配置免密登录

Hadoop 组件之间需要基于 SSH 进行通讯，配置免密登录后不需要每次都输入密码。

- 配置映射，配置 ip 地址和主机名映射

```
vim /etc/hosts  
# 文件末尾增加  
192.168.80.100 server
```

- 生成公钥私钥

```
ssh-keygen -t rsa
```



- 授权，进入 ~/.ssh 目录下，查看生成的公匙和私匙，并将公匙写入到授权文件：

```
cd ~/.ssh  
cat id_rsa.pub >> authorized_keys  
chmod 600 authorized_keys
```

### 3. 下载解压Hadoop

- 访问<http://archive.apache.org/dist/hadoop/core/hadoop-3.1.0/> 下载Hadoop



|                                                                                   |                                             |                  |      |  |
|-----------------------------------------------------------------------------------|---------------------------------------------|------------------|------|--|
|  | <a href="#">Parent Directory</a>            | -                |      |  |
|  | <a href="#">hadoop-3.1.0-src.tar.gz</a>     | 2018-04-05 20:19 | 26M  |  |
|  | <a href="#">hadoop-3.1.0-src.tar.gz.asc</a> | 2018-04-05 20:19 | 819  |  |
|  | <a href="#">hadoop-3.1.0-src.tar.gz.mds</a> | 2018-04-05 20:19 | 1.0K |  |
|  | <a href="#">hadoop-3.1.0.tar.gz</a>         | 2018-04-05 20:19 | 311M |  |
|  | <a href="#">hadoop-3.1.0.tar.gz.asc</a>     | 2018-04-05 20:19 | 819  |  |
|  | <a href="#">hadoop-3.1.0.tar.gz.mds</a>     | 2018-04-05 20:19 | 1.0K |  |

- 切换至tools目录，上传Hadoop安装包
- 解压Hadoop至server目录

```
tar -zxvf hadoop-3.1.0.tar.gz -C /opt/server/
```

### 4. 步骤效果

无

## 02. 步骤二：配置Hadoop

### 步骤操作说明

#### 1. 修改配置文件

进入/opt/server/hadoop-3.1.0/etc/hadoop 目录下，修改以下配置：

- 修改hadoop-env.sh文件，设置JDK的安装路径

```
vim hadoop-env.sh
export JAVA_HOME=/opt/server/jdk1.8.0_131
```

- 修改core-site.xml文件，分别指定hdfs 协议文件系统的通信地址及hadoop 存储临时文件的目录（此目录不需要手动创建）

```
<configuration>
  <property>
    <!--指定 namenode 的 hdfs 协议文件系统的通信地址-->
    <name>fs.defaultFS</name>
    <value>hdfs://server:8020</value>
  </property>
  <property>
    <!--指定 hadoop 数据文件存储目录-->
    <name>hadoop.tmp.dir</name>
    <value>/home/hadoop/data</value>
  </property>
</configuration>
```

- 修改hdfs-site.xml, 指定 dfs 的副本系数

```
<configuration>
  <property>
    <!-- 由于我们这里搭建是单机版本, 所以指定 dfs 的副本系数为 1-->
    <name>dfs.replication</name>
    <value>1</value>
  </property>
</configuration>
```

- 修改workers文件, 配置所有从属节点

```
vim workers
# 配置所有从属节点的主机名或 IP 地址, 由于是单机版本, 所以指定本机即可:
server
```

## 2. 初始化并启动HDFS

- 关闭防火墙, 不关闭防火墙可能导致无法访问 Hadoop 的 Web UI 界面

```
# 查看防火墙状态
sudo firewall-cmd --state
# 关闭防火墙:
sudo systemctl stop firewalld
# 禁止开机启动
sudo systemctl disable firewalld
```

- 初始化, 第一次启动 Hadoop 时需要进行初始化, 进入 /opt/server/hadoop-3.1.0/bin目录下, 执行以下命令:

```
cd /opt/server/hadoop-3.1.0/bin
./hdfs namenode -format
```

```
2021-11-09 10:02:48,450 INFO namenode.NameNode: Caching file names occurring more than 10 times
2021-11-09 10:02:48,453 INFO snapshot.SnapshotManager: Loaded config captureOpenFiles: false, skipCaptureAccessTimeOnlyChange
: false, snapshotDiffAllowSnapRootDescendant: true, maxSnapshotLimit: 65536
2021-11-09 10:02:48,454 INFO snapshot.SnapshotManager: SkipList is disabled
2021-11-09 10:02:48,457 INFO util.GSet: Computing capacity for map cachedBlocks
2021-11-09 10:02:48,457 INFO util.GSet: VM type = 64-bit
2021-11-09 10:02:48,457 INFO util.GSet: 0.25% max memory 839.5 MB = 2.1 MB
2021-11-09 10:02:48,457 INFO util.GSet: capacity = 2^18 = 262144 entries
2021-11-09 10:02:48,461 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.window.num.buckets = 10
2021-11-09 10:02:48,461 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.num.users = 10
2021-11-09 10:02:48,461 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.windows.minutes = 1,5,25
2021-11-09 10:02:48,463 INFO namenode.FSNamesystem: Retry cache on namenode is enabled
2021-11-09 10:02:48,464 INFO namenode.FSNamesystem: Retry cache will use 0.03 of total heap and retry cache entry expiry time
is 600000 millis
2021-11-09 10:02:48,464 INFO util.GSet: Computing capacity for map NameNodeRetryCache
2021-11-09 10:02:48,465 INFO util.GSet: VM type = 64-bit
2021-11-09 10:02:48,465 INFO util.GSet: 0.0299999999329447746% max memory 839.5 MB = 257.9 KB
2021-11-09 10:02:48,465 INFO util.GSet: capacity = 2^15 = 32768 entries
2021-11-09 10:02:48,479 INFO namenode.FSImage: Allocated new BlockPoolId: BP-672993686-192.168.40.100-1636423368475
2021-11-09 10:02:48,487 INFO common.Storage: Storage directory /home/hadoop/tmp/dfs/name has been successfully formatted.
2021-11-09 10:02:48,492 INFO namenode.FSImageFormatProtobuf: Saving image file /home/hadoop/tmp/dfs/name/current/fsimage.ckpt
_00000000000000000000 using no compression
2021-11-09 10:02:48,544 INFO namenode.FSImageFormatProtobuf: Image file /home/hadoop/tmp/dfs/name/current/fsimage.ckpt_000000
00000000000000 of size 389 bytes saved in 0 seconds .
2021-11-09 10:02:48,552 INFO namenode.NNStorageRetentionManager: Going to retain 1 images with txid >= 0
2021-11-09 10:02:48,556 INFO namenode.NameNode: SHUTDOWN_MSG:
/*****
SHUTDOWN_MSG: Shutting down NameNode at node01/192.168.40.100
*****/
[root@node01 bin]#
```

- Hadoop 3中不允许使用root用户来一键启动集群, 需要配置启动用户

```
cd /opt/server/hadoop-3.1.0/sbin/
# 编辑start-dfs.sh、stop-dfs.sh,在顶部加入以下内容
HDFS_DATANODE_USER=root
HDFS_DATANODE_SECURE_USER=hdfs
HDFS_NAMENODE_USER=root
HDFS_SECONDARYNAMENODE_USER=root
```

- 启动HDFS，进入/opt/server/hadoop-3.1.0/sbin/ 目录下，启动 HDFS：

```
cd /opt/server/hadoop-3.1.0/sbin/
./start-dfs.sh
```

- 验证是否启动
  - 方式一：执行 jps 查看 NameNode 和 DataNode 服务是否已经启动：

```
[root@server bin]# jps
41032 DataNode
41368 Jps
40862 NameNode
41246 SecondaryNameNode
```

- 方式二：查看 Web UI 界面，端口为 9870：

▲ 不安全 192.168.40.100:9870/dfshealth.html#tab-overview

Non Heap Memory used 45.18 MB of 46.34 MB Committed Non Heap Memory. Max Non Heap Memory is <unbounded>.

Configured Capacity:	35.1 GB
Configured Remote Capacity:	0 B
DFS Used:	4 KB (0%)
Non DFS Used:	2.95 GB
DFS Remaining:	32.15 GB (91.59%)
Block Pool Used:	4 KB (0%)
DataNodes usages% (Min/Median/Max/stdDev):	0.00% / 0.00% / 0.00% / 0.00%
Live Nodes	1 (Decommissioned: 0, In Maintenance: 0)
Dead Nodes	0 (Decommissioned: 0, In Maintenance: 0)
Decommissioning Nodes	0
Entering Maintenance Nodes	0
Total Datanode Volume Failures	0 (0 B)
Number of Under-Replicated Blocks	0
Number of Blocks Pending Deletion	0
Block Deletion Start Time	Tue Nov 09 10:32:48 +0800 2021
Last Checkpoint Time	Tue Nov 09 10:02:48 +0800 2021

- 配置环境变量，方便启动

```
export HADOOP_HOME=/opt/server/hadoop-3.1.0
export PATH=$PATH:${HADOOP_HOME}/bin:${HADOOP_HOME}/sbin
source /etc/profile
```

### 3. 步骤效果

无

## 03. 步骤三：Hadoop(YARN)环境搭建

### 步骤操作说明

#### 1. 修改配置文件

进入/opt/server/hadoop-3.1.0/etc/hadoop 目录下，修改以下配置：

- 修改mapred-site.xml文件

```
<configuration>
  <property>
    <name>mapreduce.framework.name</name>
    <value>yarn</value>
  </property>
  <property>
    <name>yarn.app.mapreduce.am.env</name>
    <value>HADOOP_MAPRED_HOME=${HADOOP_HOME}</value>
  </property>
  <property>
    <name>mapreduce.map.env</name>
    <value>HADOOP_MAPRED_HOME=${HADOOP_HOME}</value>
  </property>
  <property>
    <name>mapreduce.reduce.env</name>
    <value>HADOOP_MAPRED_HOME=${HADOOP_HOME}</value>
  </property>
</configuration>
```

- 修改yarn-site.xml文件，配置 NodeManager 上运行的附属服务

```
<configuration>
  <property>
    <!--配置 NodeManager 上运行的附属服务。需要配置成 mapreduce_shuffle 后才可以在
Yarn 上运行 MapReduce 程序。-->
    <name>yarn.nodemanager.aux-services</name>
    <value>mapreduce_shuffle</value>
  </property>
</configuration>
```

#### 2. 启动服务

- Hadoop 3中不允许使用root用户来一键启动集群，需要配置启动用户

```
# start-yarn.sh stop-yarn.sh在两个文件顶部添加以下内容
YARN_RESOURCEMANAGER_USER=root
HADOOP_SECURE_DN_USER=yarn
YARN_NODEMANAGER_USER=root
```

- 进入 \${HADOOP\_HOME}/sbin/ 目录下，启动 YARN：

```
./start-yarn.sh
```

- 验证是否启动成功
  - 方式一：执行 jps 命令查看 NodeManager 和 ResourceManager 服务是否已经启动

```
[root@server bin]# jps
41655 ResourceManager
41032 DataNode
42125 Jps
40862 NameNode
41246 SecondaryNameNode
41983 NodeManager
```

- 方式二：查看 Web UI 界面，端口为 8088

## 04. 步骤四：使用内置程序计算PI值

Hadoop自带的hadoop-mapreduce-examples-x.jar中包含一些示例程序，位于 `${HADOOP_HOME}/share/hadoop/mapreduce` 目录。

### 步骤操作说明

#### 1. 运行示例程序

- 进入 `${HADOOP_HOME}/bin/` 目录下，执行以下命令

```
hadoop jar /opt/server/hadoop-3.1.0/share/hadoop/mapreduce/hadoop-mapreduce-examples-3.1.0.jar pi 2 10
```

#### 2. 查看运行效果

```
Job Finished in 14.85 seconds
Estimated value of Pi is 3.800000000000000000000000
[root@node01 bin]#
```

## 四. 实验扩展

无



## 1. 创新扩展点

无

## 2. 创新实现步骤

01. 步骤一: 无

02. 步骤二:无

---

## 五. 附录

### 技术资料

- 官方文档:<https://hadoop.apache.org/docs/r1.0.4/cn/quickstart.html>
-