湘潭大学

分布式框架搭建 2

讲课：季俊豪

专　　业：数据科学与大数据技术

内容纲要

伪分布的配置

在本地模式的基础上，配置四个配置文件。
hadoop 的配置文件参考目录，在/opt/hadoop*/etc/hadoop 下：

- hdfs-site.xml
- core-site.xml
- mapred-site.xml
- yarn-site.xml

伪分布的配置

其中 mapred-site.xml 比较特殊，由于配置文件中默认没有该文件 mapred-site.xml，在配置前需要根据模板复制一份
cp mapred-site.xml.template mapred-site.xml

伪分布的配置

在各个文件的 configuration 里面配置文件

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
  Licensed under the Apache License, Version 2.0 (the "License");
  you may not use this file except in compliance with the License.
  You may obtain a copy of the License at

    http://www.apache.org/licenses/LICENSE-2.0

  Unless required by applicable law or agreed to in writing, software
  distributed under the License is distributed on an "AS IS" BASIS,
  WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or impli
  See the License for the specific language governing permissions and
  limitations under the License. See accompanying LICENSE file.
-->

<!-- Put site-specific property overrides in this file. -->

<configuration>

</configuration>
~
~
~
~
~
~
~
~
```

hdfs-site.xml

```
<configuration>
<!– hdfs 保存副本数据的数量，包括自己，默认为 3，伪分布配置成 1，全分布按照连接的数量来 –>
<property>
<name>dfs.replication</name>
<value>1</value>
</property>
<!– hadoop 权限管理，false 表示任何用户都可以在 hdfs 上操作文件 –>
<property>
<name>dfs.permissions</name>
<value>false</value>
</property>
<!– 指定元信息的存储路径 /tmp –>
<property>
<name>dfs.namenode.name.dir</name>
<value>file:/data/name/data</value>
</property>
<!– 指定数据的存储路径 /tmp –>
<property>
<name>dfs.datanode.data.dir</name>
<value>file:/data/dfs/data</value>
</property>
</configuration>
```

core-site.xml

```
<configuration>
<!– hdfs 文件系统的老大, 指定 namenode 的节点位置 –>
<property>
<name>fs.defaultFS</name>
<value>hdfs://192.168.59.132:9000</value>
</property>
<!– hdfs 运行时产生的临时文件的存储路径，一定是本地真实存在的路径 –>
<property>
<name>hadoop.tmp.dir</name>
<value>/opt/hadoop-2.6.0-cdh5.14.0/tmp</value>
</property>
</configuration>
```

mapred-site.xml

<configuration>
<!– 指定 mapreduce 的处理框架 yarn, 在 yarn 上运行–>
<property>
<name>mapreduce.framework.name</name>
<value>yarn</value>
</property>
</configuration>

yarn-site.xml

```
<configuration>
<!– yarn 所在的地址 –>
<property>
<name>yarn.resourcemanager.hostname</name>
<value>192.168.59.132</value>
</property>
<!– nodemanager 获取数据的方式，mapreduce 方式通过 shuffle –>
<property>
<name>yarn.nodemanager.aux-services</name>
<value>mapreduce_shuffle</value>
</property>
</configuration>
```

注意事项

- 配置的 ip 地址一定要是自己的 ip
- 一定要保证 tmp 路径真实存在
- 上述配置，其中的 <!– –> 是备注

创建文件夹

1. tmp 文件
   cd /opt/hadoop*
   mkdir tmp
2. 创建/下的 data 文件夹存数据
   mkidr /data

格式化 hdfs

hdfs namenode -format 或者 hadoop namenode -format

启动和关闭

start-all.sh 启动

stop-all.sh 关闭

命令 jps 查看和 java 相关的进程出现 6 个节点说明配置成功

```
[root@hadoop1 hadoop-2.6.0-cdh5.14.0]# start-all.sh
This script is Deprecated. Instead use start-dfs.sh and start-yarn.sh
20/12/12 22:55:59 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platfo
Starting namenodes on [hadoop1]
hadoop1: starting namenode, logging to /opt/hadoop-2.6.0-cdh5.14.0/logs/hadoop-root-namenode-hadoo
localhost: starting datanode, logging to /opt/hadoop-2.6.0-cdh5.14.0/logs/hadoop-root-datanode-hao
Starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to /opt/hadoop-2.6.0-cdh5.14.0/logs/hadoop-root-secon
20/12/12 22:56:15 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platfo
starting yarn daemons
starting resourcemanager, logging to /opt/hadoop-2.6.0-cdh5.14.0/logs/yarn-root-resourcemanager-ha
localhost: starting nodemanager, logging to /opt/hadoop-2.6.0-cdh5.14.0/logs/yarn-root-nodemanager
[root@hadoop1 hadoop-2.6.0-cdh5.14.0]# jps
16065 ResourceManager
16355 NodeManager
16472 Jps
15755 DataNode
15915 SecondaryNameNode
15629 NameNode
[root@hadoop1 hadoop-2.6.0-cdh5.14.0]#
```

注意事项

jps 进程数量一定要是六节点，不是 6 节点的，自己排错

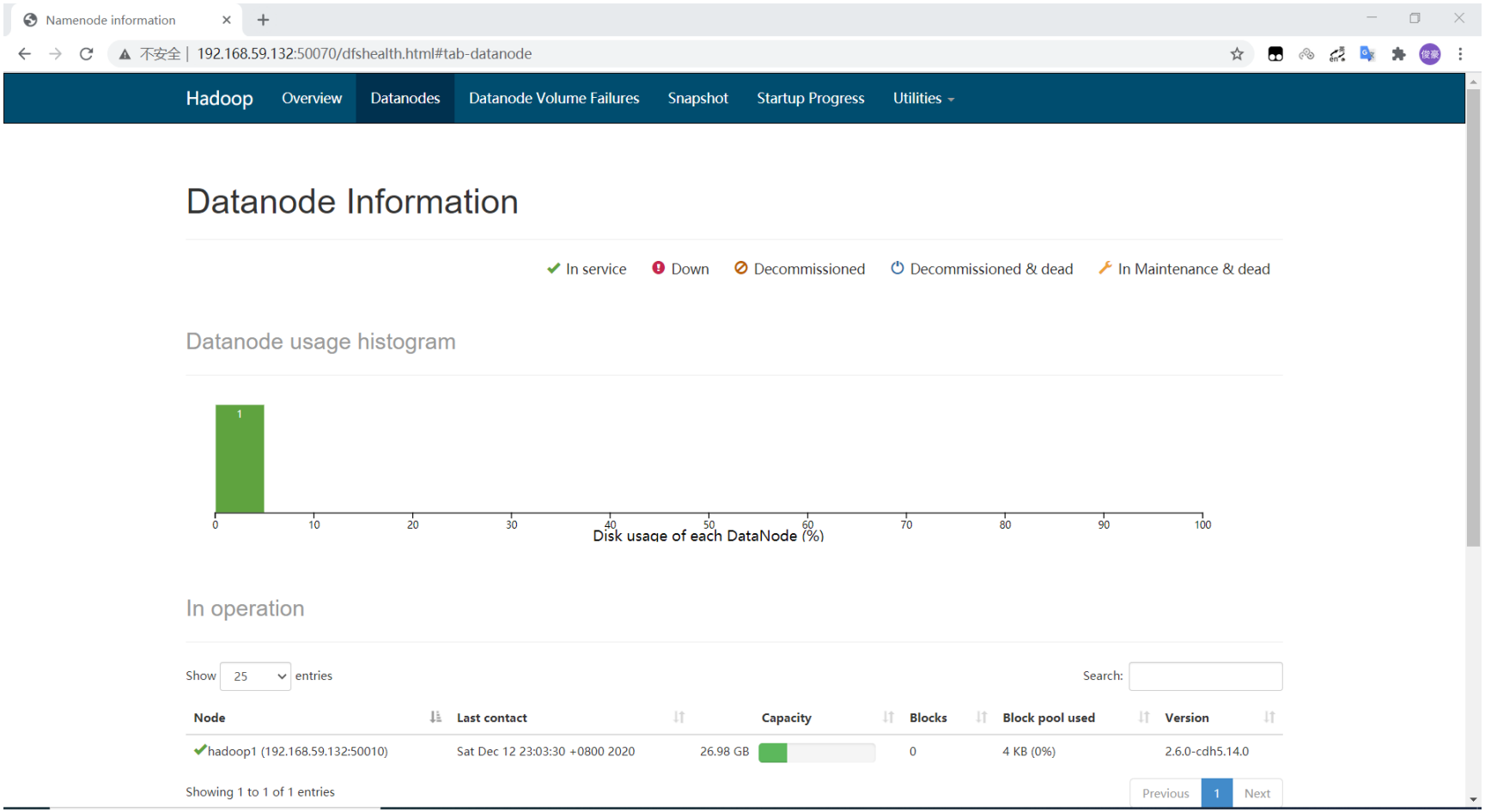
1. stop-all.sh
2. 清除数据
   cd /data
   rm -rf *
   cd /opt/hadoop-2.6.0-cdh5.14.0/tmp
   rm -rf *
3. 查看四个配置文件
4. 查看主机 ip 是否变化
5. 重新格式化 (重新格式化前一定要清除数据)
6. 启动

浏览器查看

http://192.168.59.132:50070

自己的 ip 地址:50070

输入英文:，不要输入中文



浏览器查看

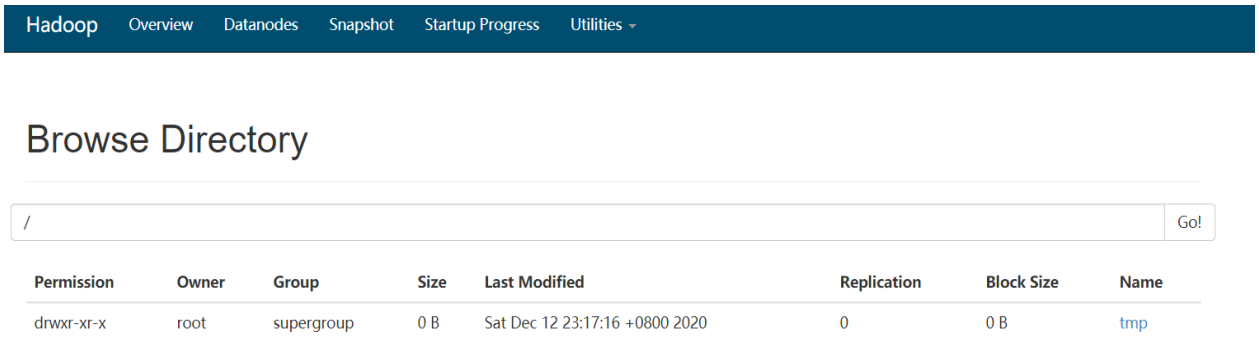
- overview 里面是设备信息等
- Datanodes 里面是你的数据存储信息
- startup progress 里面是正在运行的程序和占用的信息
- utilities 里面是 hdfs 文件系统的整个目录

内容纲要

hdfs 创建文件夹

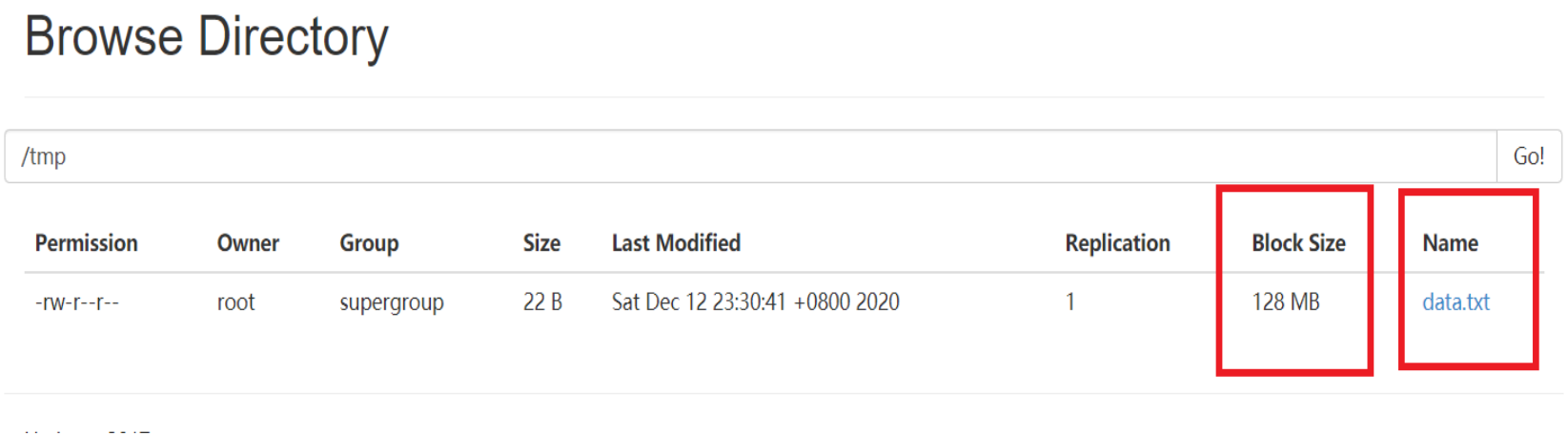
hdfs dfs -ls /
hdfs dfs -mkdir /tmp
hdfs dfs -ls /
同时在浏览器查看文件系统，发现其实和 linux 命令差不多，实际上 80% 相似

```
[root@hadoop1 hadoop-2.6.0-cdh5.14.0]# hdfs dfs -ls /
20/12/12 23:17:01 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
 using builtin-java classes where applicable
[root@hadoop1 hadoop-2.6.0-cdh5.14.0]# hdfs dfs -mkdir /tmp
20/12/12 23:17:16 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
 using builtin-java classes where applicable
[root@hadoop1 hadoop-2.6.0-cdh5.14.0]# hdfs dfs -ls /
20/12/12 23:17:21 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
 using builtin-java classes where applicable
Found 1 items
drwxr-xr-x   - root supergroup          0 2020-12-12 23:17 /tmp
[root@hadoop1 hadoop-2.6.0-cdh5.14.0]#
```



hdfs 上传下载文件

- 上传文件到 hadoop
  hdfs dfs -put /data.txt /tmp
- 下载到 linux
  hdfs dfs -get /tmp/data.txt /data/



一个奇怪的现象，输入是个小文件，为什么文件大小是 128MB???

内容纲要

全分布的配置

首先需要三台 Centos，其次需要配置 hosts 和五个配置文件。

- hdfs-site.xml
- core-site.xml
- mapred-site.xml
- yarn-site.xml
- slaves

修改主机主机名和映射

- 分别修改主机名
- 在 hosts 里面输入三台的映射, 例如
  192.168.59.132 hadoop132
  192.168.59.133 hadoop133
  192.168.59.134 hadoop134

hdfs-site.xml

dfs.replication 改为 3，也就是 1 台 namenode，下面三个 datanode，包括自己。

corn-site.xml 和 yarn-site.html

ip 地址，改成 namenode 的 ip 地址，即统一

slaves

字面意思，奴隶，输入三台主机的名字，都用作 datanode，即
hadoop132
hadoop133
hadoop134

启动

三台主机都格式化，对 namenode 的主机 start-all.sh, 查看 jps

谢 谢 大 家!