

MPLab1-实验报告

Name: 傅小龙

Dept: CS

Grade: 3

ID: 191220029

实验目标

完成单机伪分布式Hadoop系统的安装，运行Hadoop自带的WordCount程序测试。

实验过程

1. 单机操作系统安装

在VMware中创建Linux(Ubuntu 20.02 2)虚拟机。

2. 安装SSH

步骤1中创建的虚拟机已经安装有SSH。

3. 安装JAVA

jdk1.8.0_321

4. 创建用户

添加hadoop-user用户组，组成员用户myhadoop。

5. 解压安装Hadoop

安装的Hadoop版本为2.7.1

6. 配置环境变量

```
1 PATH=$PATH:$HOME/bin
2 export JAVA_HOME=/usr/jdk1.8.0_321
3 export HADOOP_HOME=/home/myhadoop/hadoop-2.7.1
4 export PATH=$JAVA_HOME/bin:$HADOOP_HOME/bin:$PATH
5 export CLASSPATH=$JAVA_HOME/lib:.
```

7. 免密码SSH访问配置

生成SSH认证文件并将秘钥复制到 /.ssh/authorized_keys文件中。测试登陆结果见Figure1.

8. 修改Hadoop配置文件

具体过程参考教材《深入理解大数据》2.2.5节配置Hadoop环境。

9. 格式化NameNode 执行结果见Figure2.

10. 启动HDFS和MapReduce

执行start-all.sh后用jps指令查看进程信息,结果见Figure3.

```

root@ubuntu:/home/myhadoop/hadoop-2.7.1/sbin# ssh localhost
Welcome to Ubuntu 20.04.3 LTS (GNU/Linux 5.4.0-48-generic x86_64)

 * Documentation:  https://help.ubuntu.com
 * Management:    https://landscape.canonical.com
 * Support:       https://ubuntu.com/advantage

2 updates can be applied immediately.
To see these additional updates run: apt list --upgradable

Your Hardware Enablement Stack (HWE) is supported until April 2025.
Last login: Wed Mar 16 06:33:52 2022 from 127.0.0.1
root@ubuntu:~# exit
logout
Connection to localhost closed.

```

Figure 1: SSH访问配置

```

22/03/16 03:59:03 INFO namenode.FSNamesystem: dfs.namenode.safemode.extension = 30000
22/03/16 03:59:03 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.window.num.buckets = 10
22/03/16 03:59:03 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.num.users = 10
22/03/16 03:59:03 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.windows.minutes = 1,5,25
22/03/16 03:59:03 INFO namenode.FSNamesystem: Retry cache on namenode is enabled
22/03/16 03:59:03 INFO namenode.FSNamesystem: Retry cache will use 0.03 of total heap and retry cache entry expiry time is 600000 millis
22/03/16 03:59:03 INFO util.GSet: Computing capacity for map NameNodeRetryCache
22/03/16 03:59:03 INFO util.GSet: VM type = 64-bit
22/03/16 03:59:03 INFO util.GSet: 0.0299999999329447746% max memory 889 MB = 273.1 KB
22/03/16 03:59:03 INFO util.GSet: capacity = 2^15 = 32768 entries
22/03/16 03:59:03 INFO namenode.FSImage: Allocated new BlockPoolId: BP-1717107628-127.0.1.1-1047428343344
22/03/16 03:59:03 INFO common.Storage: Storage directory /home/myhadoop/hadoop_dir/dfs/name has been successfully formatted.
22/03/16 03:59:03 INFO namenode.NNStorageRetentionManager: Going to retain 1 images with txid >= 0
22/03/16 03:59:03 INFO util.ExitUtil: Exiting with status 0
22/03/16 03:59:03 INFO namenode.NameNode: SHUTDOWN_MSG:
/*****
SHUTDOWN_MSG: Shutting down NameNode at ubuntu/127.0.1.1
*****/

```

Figure 2: 格式化NameNode

11. 停止HDFS和MapReduce

执行stop-all.sh

12. 运行WordCount测试

重新执行步骤10.

用以下网页

- <https://github.com/malware-dev/MDK-SE>
- <https://github.com/malware-dev/MDK-SE/wiki>
- <https://github.com/malware-dev/MDK-SE/wiki/Api-Index>

作为输入数据进行词频统计。malware-dev/MDK-SE是github上的一个开源项目，旨在介绍太空工程师可编程模块中提供的C#接口。用到的这三个页面分别是项目主页、wiki手册主页以及手册中的API目录页面。

在dfs系统下新建lab1文件夹，然后将测试数据拷贝到该文件夹下的testfile文件夹内。以lab1/out作为输出文件夹，提交任务（见Figure4）：

运行结束后Hadoop Web作业状态查看界面见Figure5：

实验输出结果开头部分见Figure6:

实验体会


本次实验中完成了Hadoop单机伪分布式系统的安装。Hadoop系统的安装过程中需要编辑

```
root@ubuntu:/home/myhadoop/hadoop-2.7.1/sbin# jps
4212 DataNode
6615 Jps
4663 NodeManager
4412 SecondaryNameNode
4541 ResourceManager
4061 NameNode
```

Figure 3: jps

```
root@ubuntu:/home/myhadoop/hadoop-2.7.1/bin# ./hadoop jar ../share/hadoop/mapreduce/hadoop-mapreduce-examples-2.7.1.jar wordcount /lab1/testfile /lab1/out
```

Figure 4:

Application application_1647933712132_0003

Logged in as: dr.who

Cluster

About

Nodes

Node Labels

Applications

NEW

NEW SAVING

SUBMITTED

ACCEPTED

RUNNING

FINISHED

FAILED

KILLED

Scheduler

Tools

Kill Application

Application Overview

User: root

Name: word count

Application Type: MAPREDUCE

Application Tags:

YarnApplicationState: FINISHED

FinalStatus Reported by AM: SUCCEEDED

Started: Tue Mar 22 00:27:01 -0700 2022

Elapsed: 34sec

Tracking URL: History

Diagnostics:

Application Metrics

Total Resource Preempted: <memory:0, vCores:0>

Total Number of Non-AM Containers Preempted: 0

Total Number of AM Containers Preempted: 0

Resource Preempted from Current Attempt: <memory:0, vCores:0>

Number of Non-AM Containers Preempted from Current Attempt: 0

Aggregate Resource Allocation: 152461 MB-seconds, 105 vcore-seconds

Show 20 entries

Search:

Attempt ID

Started

Node

Logs

appattempt_1647933712132_0003_000001

Tue Mar 22 00:27:01 -0700 2022

http://ubuntu:8042

Logs

Showing 1 to 1 of 1 entries

First Previous 1 Next Last

Figure 5: 作业状态

大量的配置文件，这对用户不太友好，是否可以用脚本或设计一个UI界面来引导用户逐步完成Hadoop系统的配置来实现这一步呢？

在实验过程中主要遇到的问题为安装教程可能已经落后于要用到的版本: 2.7.1版本的Hadoop配置文件中并没有mapred-site.xml，而是mapred-site.xml.template，需要将改文件去除.template后缀才能使配置生效。在Web端查看提交作业的状态需要额外配置yarn。具体过程为在mapred-site.xml中添加如下部分：

```
1 <property>
2     <name>mapreduce.framework.name</name>
3     <value>yarn</value>
4 </property>
```

在yarn-site.xml中添加如下部分：

```
1 <property>
2     <name>yarn.nodemanager.aux-services</name>
3     <value>mapreduce_shuffle</value>
4 </property>
```

```
part-r-00000
~/BigData/lab1

1 !important;" 1
2 !important;;width: 2
3 " 47
4 "> 6300
5 ">Jump 2
6 "But 1
7 "Getting 2
8 "done". 1
9 # 6
10 #247 1
11 >; 12
12 "$opt_dummy_audience";, 1
13 "1096"}> 1
14 "10f8ab3fbc5ebe989a36a05f79d48f32";: 1
15 "16737760170";, 2
16 "16822470375";, 1
17 "1686089f6d540cd2deeaec60ee43ecf7";: 1
18 "17143601254";, 1
19 "17911811441";, 1
20 "18124116703";, 1
21 "18145892387";, 1
22 "18175660309";, 1
23 "18178755568";, 1
24 "18180553241";, 1
25 "18186103728";, 1
26 "18188530140";, 1
27 "18191963644";, 1
28 "18195612788";, 1
29 "18210945499";, 1
30 "18211063248";, 1
31 "18215721889";, 1
32 "18224360785";, 1
33 "18234832286";, 1
```

Figure 6: 输出结果的开头部分

本次实验中另外遇到的一个问题是在执行任务时进度卡在reduce 0%阶段。在网上查找原因后发现是虚拟机分配的内存太小。将虚拟机内存扩大后再次执行任务，成功完成。