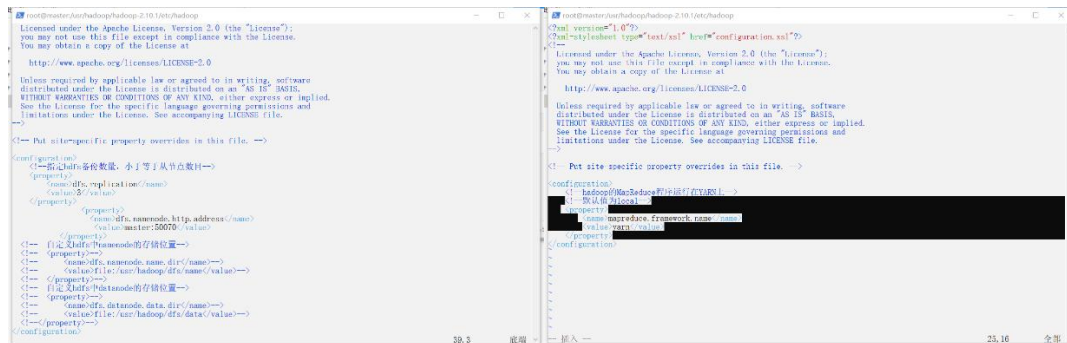


IEMS 5730 Spring 2021 Homework #0

Q1:Hadoop Cluster Setup

I used the VM Workstation to set up four virtual machines and build a multi-node Hadoop cluster. (1 Master and 3Slaves)

Step1: setup the Hadoop environment for the master

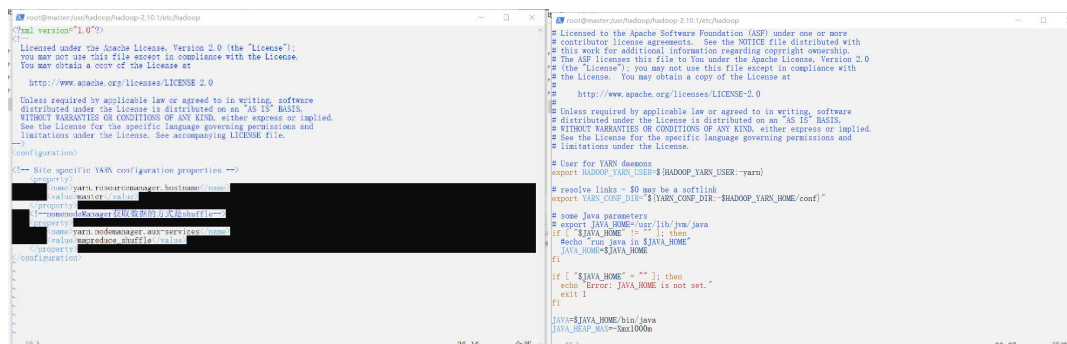


```
root@master:/usr/hadoop/hadoop-2.10.1/etc/hadoop
Licensed under the Apache License, Version 2.0 (the "license");
you may not use this file except in compliance with the license.
You may obtain a copy of the license at
http://www.apache.org/licenses/LICENSE-2.0
Unless required by applicable law or 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 implied.
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>
  <!--高hdfs名称数量，小j等于从什么数H-->
  <property>
    <name>dfs.replication</name>
    <value>3</value>
  </property>
  <property>
    <name>dfs.nameNode.http.address</name>
    <value>master:50070</value>
  </property>
  <!-- hdfs名称的存储位置 -->
  <property>
    <name>dfs.nameNode.dir</name>
    <value>file:/usr/hadoop/dfs/name</value>
  </property>
  <!-- hdfs数据块的存储位置 -->
  <property>
    <name>dfs.dataNode.dir</name>
    <value>file:/usr/hadoop/dfs/data</value>
  </property>
</configuration>
```

```
root@master:/usr/hadoop/hadoop-2.10.1/etc/hadoop
<?xml version="1.0"?>
<?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 implied.
  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>
  <property>
    <name>mapred.master.address</name>
    <value>master</value>
  </property>
  <property>
    <name>mapred.framework.name</name>
    <value>org.apache.hadoop.mapred.YarnPlatform</value>
  </property>
</configuration>
```

P1: hdfs configuration

P2: mapred-site.xml configuration



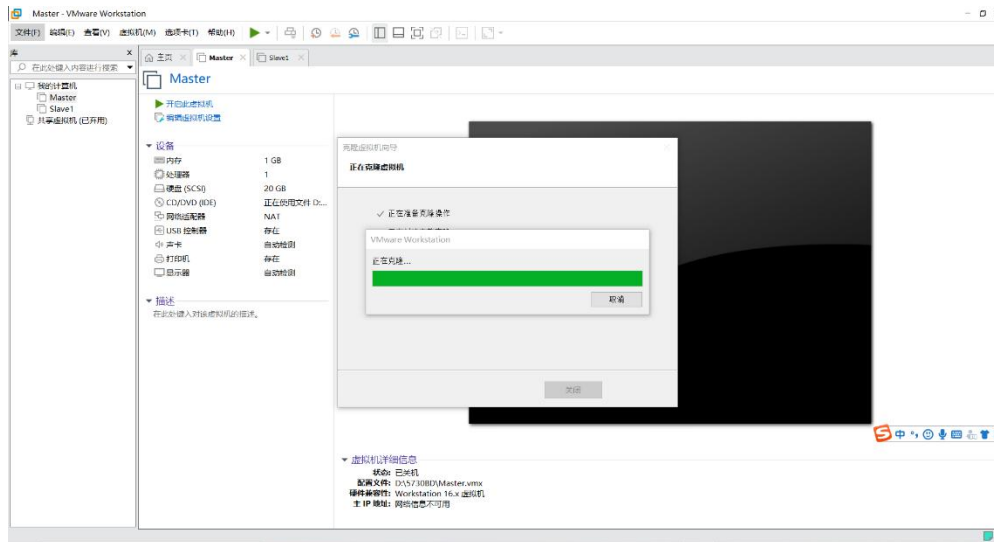
```
root@master:/usr/hadoop/hadoop-2.10.1/etc/hadoop
<?xml version="1.0"?>
<!--
  Licensed under the Apache License, Version 2.0 (the "License");
  you may not use this file except in compliance with the License.
  You may obtain a copy of the License at
    http://www.apache.org/licenses/LICENSE-2.0
  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 implied.
  See the License for the specific language governing permissions and
  limitations under the License. See accompanying LICENSE file.
-->
<configuration>
  <!-- Site specific YARN configuration properties -->
  <property>
    <name>yarn.resourcemanager.hostname</name>
    <value>master</value>
  </property>
  <property>
    <name>yarn.resourcemanager.address</name>
    <value>master:10255</value>
  </property>
  <property>
    <name>yarn.resourcemanager.aux-services</name>
    <value>mapreduce_shuffle</value>
  </property>
</configuration>
```

```
root@master:/usr/hadoop/hadoop-2.10.1/etc/hadoop
# Licensed to the Apache Software Foundation (ASF) under one or more
# contributor license agreements. See the NOTICE file distributed with
# this work for additional information regarding copyright ownership.
# The ASF licenses this file to You 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 implied.
# See the License for the specific language governing permissions and
# limitations under the License.
# User for YARN daemons
export HADOOP_YARN_USER=${HADOOP_YARN_USER:-yarn}
# resolve links - $0 may be a softlink
export YARN_CONF_DIR=${YARN_CONF_DIR:-$HADOOP_YARN_HOME/conf}
# some Java parameters
export JAVA_HOME=/usr/lib/jvm/java
if [ "$JAVA_HOME" != "" ]; then
  echo "You have JAVA_HOME = $JAVA_HOME"
  JAVA_HOME=$JAVA_HOME
fi
if [ "$JAVA_HOME" != "" ]; then
  echo "Error: JAVA_HOME is not set."
  exit 1
fi
JAVA=$JAVA_HOME/bin/java
JAVA_HEAP_MAX=32m
```

P3: yarn-site.xml configuration

P4: Set the Hadoop path

Step2: clone the master virtual machine



P5: Clone the master(three times)

```
127.0.0.1 localhost localhost.loc
:1 localhost localhost.loc
192.168.2.130 master
192.168.2.131 slave1
192.168.2.132 slave2
192.168.2.133 slave3
```

P6: Modify the hosts

```
[root@master network-scripts]# hostname
slave3
[root@master network-scripts]# ssh master
The authenticity of host 'master (192.168.2.130)' can't be established.
ECDSA key fingerprint is SHA256:SXw9MXyGDOP/w6ALh0IT6ZINPPCWiwuslWXwSQLKd0.
ECDSA key fingerprint is MD5:2b:a5:dd:c8:d0:72:1b:1a:5c:ab:6e:87:df:85:bf:82.
Are you sure you want to continue connecting (yes/no)? yes
Warning: Permanently added 'master,192.168.2.130' (ECDSA) to the list of known hosts.
root@master's password:
Last login: Mon Jan 25 00:42:14 2021 from master
[root@master ~]# hostname
master
[root@master ~]# exit
logout
Connection to master closed.
[root@master network-scripts]# hostname
slave3
[root@master network-scripts]# ping master
PING master (192.168.2.130) 56(84) bytes of data.
64 bytes from master (192.168.2.130): icmp_seq=1 ttl=64 time=0.193 ms
64 bytes from master (192.168.2.130): icmp_seq=2 ttl=64 time=0.295 ms
64 bytes from master (192.168.2.130): icmp_seq=3 ttl=64 time=0.378 ms
^C
--- master ping statistics ---
3 packets transmitted, 3 received, 0% packet loss, time 2002ms
rtt min/avg/max/mdev = 0.193/0.288/0.378/0.078 ms
[root@master network-scripts]# _
```

P7: Slave tries to connect to the host

```
root@master:~# ssh root@192.168.2.130
root@192.168.2.130's password:
Warning: Permanently added '192.168.2.130' (ECDSA) to the list of known hosts.
root@slave2:~#

root@master:~# ssh root@192.168.2.131
root@192.168.2.131's password:
Warning: Permanently added '192.168.2.131' (ECDSA) to the list of known hosts.
root@slave1:~#

root@master:~# ssh root@192.168.2.132
root@192.168.2.132's password:
Warning: Permanently added '192.168.2.132' (ECDSA) to the list of known hosts.
root@slave2:~#

root@master:~# ssh root@192.168.2.133
root@192.168.2.133's password:
Warning: Permanently added '192.168.2.133' (ECDSA) to the list of known hosts.
root@slave3:~#
```

P8: All four hosts are connected

```
root@master:~# ssh root@192.168.2.130
root@192.168.2.130's password:
Warning: Permanently added '192.168.2.130' (ECDSA) to the list of known hosts.
root@slave2:~#

root@master:~# ssh root@192.168.2.131
root@192.168.2.131's password:
Warning: Permanently added '192.168.2.131' (ECDSA) to the list of known hosts.
root@slave1:~#

root@master:~# ssh root@192.168.2.132
root@192.168.2.132's password:
Warning: Permanently added '192.168.2.132' (ECDSA) to the list of known hosts.
root@slave2:~#

root@master:~# ssh root@192.168.2.133
root@192.168.2.133's password:
Warning: Permanently added '192.168.2.133' (ECDSA) to the list of known hosts.
root@slave3:~#
```

P9: Generate a public key and send it to each slave for easy connection

Step3: Hadoop Test

```
root@master:~# cd /usr/hadoop/hadoop-2.10.1/sbin
root@master/sbin# ls
container-executor  hadoop.cmd  hdfs.cmd  mapred.cmd  test-container-executor  yarn.cmd
hadoop             hdfs        mapred     rcc          yarn
root@master/sbin# cd ../
root@master/sbin# cd sbin/
root@master/sbin# ls
distributed-exclude.sh  httpfs.sh  start-all.sh  start-yarn.sh  stop-secure-dns.sh
FederationStateStore  kms.sh     start-balancer.sh  stop-all.cmd  stop-yarn.cmd
hadoop-daemon.sh       mr-jobhistory-daemon.sh  start-dfs.cmd  stop-all.sh  stop-yarn.sh
hadoop-daemons.sh     refresh-namenodes.sh  start-dfs.sh  stop-balancer.sh  yarn-daemon.sh
hdfs-config.cmd        slaves.sh  start-secure-dns.sh  stop-dfs.cmd  yarn-daemons.sh
hdfs-config.sh         start-all.cmd  start-yarn.cmd  stop-dfs.sh
root@master/sbin# start-dfs.sh
Starting namenodes on [master]
master: starting namenode, logging to /usr/hadoop/hadoop-2.10.1/logs/hadoop-root-namenode-master.out
slave2: starting datanode, logging to /usr/hadoop/hadoop-2.10.1/logs/hadoop-root-datanode-slave2.out
slave1: starting datanode, logging to /usr/hadoop/hadoop-2.10.1/logs/hadoop-root-datanode-slave1.out
slave3: starting datanode, logging to /usr/hadoop/hadoop-2.10.1/logs/hadoop-root-datanode-slave3.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:SW9MXyGDOP/w6ALh0IT6Z1NPPCWiwusWXwvSQLKd0.
ECDSA key fingerprint is MD5:2b:a5:dd:c8:d0:72:1b:1a:5c:ab:6e:87:df:85:bf:82.
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.
root@master/sbin# start-yarn.sh
Starting yarn daemons
Starting resourcemanager, logging to /usr/hadoop/hadoop-2.10.1/logs/yarn-root-resourcemanager-master.out
slave2: starting nodemanager, logging to /usr/hadoop/hadoop-2.10.1/logs/yarn-root-nodemanager-slave2.out
slave3: starting nodemanager, logging to /usr/hadoop/hadoop-2.10.1/logs/yarn-root-nodemanager-slave3.out
slave1: starting nodemanager, logging to /usr/hadoop/hadoop-2.10.1/logs/yarn-root-nodemanager-slave1.out
root@master/sbin#
```

P10: Start the Hadoop (using start-dfs.sh and start-yarn.sh)

```

[root@master sbin]# jps
1634 SecondaryNameNode
1444 NameNode
1781 ResourceManager
2047 Jps

```

P11: Checking the Hadoop process (using jps)

All Applications

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Used Resources	Total Resources
0	0	0	0	0	<memory:0, vCores:0>	<memory:24576, vCores:1>

Cluster Nodes Metrics

Active Nodes	Decommissioning Nodes	Decommissioned Nodes	Lost Nodes	Unhealthy Nodes	Reserved Nodes
3	0	0	0	0	0

Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation
Capacity Scheduler	[<name=memory-mb default-unit=Mi type=COUNTABLE>, <name=vcores default-unit= type=COUNTABLE>]	<memory:1024, vCores:1>	<memory:8192, vCores:1>

Show 20 entries

ID	User	Name	Application Type	Queue	Application Priority	StartTime	LaunchTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU VCoers	Allocated Memory MB	Allocated GPUs	Reserved CPU VCoers	Reserved Memory MB	Reserved GPUs
No data available in table																	

Showing 0 to 0 of 0 entries

P12: Yarn interface display (port 8088)

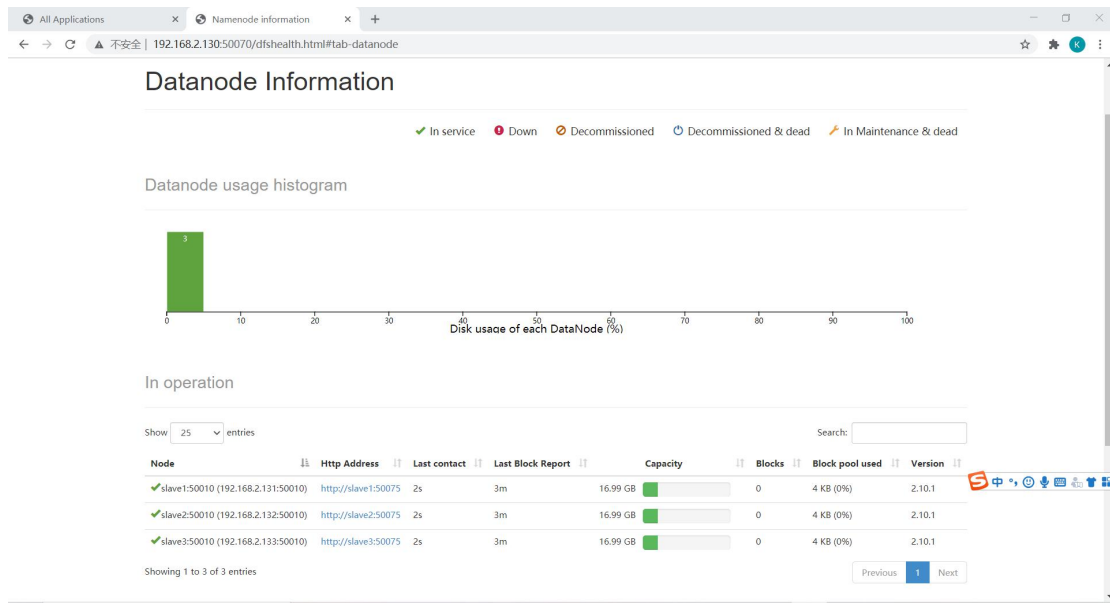
Overview 'master:8020' (active)

Started:	Mon Jan 25 01:04:19 +0800 2021
Version:	2.10.1, r1827467c9a56f133025f28557bfc2c562d78e816
Compiled:	Mon Sep 14 21:17:00 +0800 2020 by centos from branch-2.10.1
Cluster ID:	CID-343bdb67-e6d1-4f4e-84bb-b873fa1486bd
Block Pool ID:	BP-1371969055-192.168.2.130-1611507590583

Summary

Security is off.
Safemode is off.
1 files and directories, 0 blocks = 1 total filesystem object(s).
Heap Memory used 29 MB of 49.38 MB Heap Memory. Max Heap Memory is 966.69 MB.
Non Heap Memory used 42.48 MB of 43.44 MB Committed Non Heap Memory. Max Non Heap Memory is <unbounded>.

Configured Capacity:	50.96 GB
DFS Used:	12 KB (0%)
Non DFS Used:	8.55 GB



P12 & P13: Namenode information (port 50070)

Step4: Terasort Test

```
root@master:usr/hadoop/hadoop-2.10.1/sbin
[root@master sbin]# hadoop jar /usr/hadoop/hadoop-2.10.1/share/hadoop/mapreduce/hadoop-mapreduce-examples-2.10.1.jar ter
agen 10737418 /terasort/1G-input
21/01/25 01:18:45 INFO client.RMProxy: Connecting to ResourceManager at master/192.168.2.130:8032
21/01/25 01:18:46 INFO terasort.TeraGen: Generating 10737418 using 2
21/01/25 01:18:46 INFO mapreduce.JobSubmitter: number of splits:2
21/01/25 01:18:46 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1611507897104_0001
21/01/25 01:18:47 INFO conf.Configuration: resource-types.xml not found
21/01/25 01:18:47 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
21/01/25 01:18:47 INFO resource.ResourceUtils: Adding resource type - name = memory-mb, units = Mi, type = COUNTABLE
21/01/25 01:18:47 INFO resource.ResourceUtils: Adding resource type - name = vcores, units = , type = COUNTABLE
21/01/25 01:18:47 INFO impl.YarnClientImpl: Submitted application application_1611507897104_0001
21/01/25 01:18:47 INFO mapreduce.Job: The url to track the job: http://master:8088/proxy/application_1611507897104_0001/
21/01/25 01:18:47 INFO mapreduce.Job: Running job: job_1611507897104_0001
21/01/25 01:18:57 INFO mapreduce.Job: Job job_1611507897104_0001 running in uber mode : false
21/01/25 01:18:57 INFO mapreduce.Job: map 0% reduce 0%
21/01/25 01:19:17 INFO mapreduce.Job: map 19% reduce 0%
21/01/25 01:19:18 INFO mapreduce.Job: map 38% reduce 0%
21/01/25 01:19:23 INFO mapreduce.Job: map 47% reduce 0%
21/01/25 01:19:24 INFO mapreduce.Job: map 56% reduce 0%
21/01/25 01:19:30 INFO mapreduce.Job: map 65% reduce 0%
21/01/25 01:19:31 INFO mapreduce.Job: map 76% reduce 0%
21/01/25 01:19:36 INFO mapreduce.Job: map 86% reduce 0%
21/01/25 01:19:37 INFO mapreduce.Job: map 97% reduce 0%
21/01/25 01:19:38 INFO mapreduce.Job: map 100% reduce 0%
21/01/25 01:19:39 INFO mapreduce.Job: Job job_1611507897104_0001 completed successfully
21/01/25 01:19:39 INFO mapreduce.Job: Counters: 31
File System Counters
  FILE: Number of bytes read=0
  FILE: Number of bytes written=416530
  FILE: Number of read operations=0
  FILE: Number of large read operations=0
  FILE: Number of write operations=0
  HDFS: Number of bytes read=167
  HDFS: Number of bytes written=1073741800
  HDFS: Number of read operations=8
  HDFS: Number of large read operations=0
  HDFS: Number of write operations=4
Job Counters
  Launched map tasks=2
  Other local map tasks=2
```

P14: map/reduce process of teragen

```
[root@master ~]# hadoop jar /usr/hadoop/hadoop-2.10.1/share/hadoop/mapreduce/hadoop-mapreduce-examples-2.10.1.jar terage
n 10737418 /terasort/2G-input
```

P15: Generate 2G input

```
[root@master ~]# hadoop jar /usr/hadoop/hadoop-2.10.1/share/hadoop/mapreduce/hadoop-mapreduce-examples-2.10.1.jar teraso
rt /terasort/2G-input/terasort/2G-output
```

P16: Terasort process

When I am dealing with the terasort process, an error happened(shown in P17) and I can't find a solution for a long time. After checking, I found that I forgot to type a space before “hadoop” so the process can't find the correct path.

```
[root@master hadoop-2.10.1]# bin/hadoop jar /usr/hadoop/hadoop-2.10.1/share/hadoop/mapreduce/hadoop-mapreduce-examples-
10.1.jar terasort terasort/2G-input terasort/2G-output
21/01/25 03:10:45 INFO terasort.TeraSort: starting
21/01/25 03:10:46 ERROR terasort.TeraSort: Input path does not exist: hdfs://master:8020/user/root/terasort/2G-input
```

P17: Error

```
[root@master mapreduce]# hadoop jar /usr/hadoop/hadoop-2.10.1/share/hadoop/mapreduce/hadoop-mapreduce-examples-2.10.1.ja
r terasort /terasort/2G-input /terasort/2G-output
21/01/25 03:36:27 INFO terasort.TeraSort: starting
21/01/25 03:36:28 INFO input.FileInputFormat: Total input files to process : 2
Spent 162ms computing base-splits.
Spent 2ms computing TeraScheduler splits.
Computing input splits took 164ms
Sampling 8 splits of 8
Making 1 from 100000 sampled records
Computing partitions took 865ms
Spent 1031ms computing partitions.
21/01/25 03:36:29 INFO client.RMProxy: Connecting to ResourceManager at master/192.168.2.130:8032
21/01/25 03:36:30 INFO mapreduce.JobSubmitter: number of splits:8
21/01/25 03:36:31 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1611507897104_0003
21/01/25 03:36:31 INFO conf.Configuration: resource-types.xml not found
21/01/25 03:36:31 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
21/01/25 03:36:31 INFO resource.ResourceUtils: Adding resource type - name = memory-mb, units = Mi, type = COUNTABLE
21/01/25 03:36:31 INFO resource.ResourceUtils: Adding resource type - name = vcores, units = , type = COUNTABLE
21/01/25 03:36:31 INFO impl.YarnClientImpl: Submitted application application_1611507897104_0003
21/01/25 03:36:31 INFO mapreduce.Job: The url to track the job: http://master:8088/proxy/application_1611507897104_0003/
21/01/25 03:36:31 INFO mapreduce.Job: Running job: job_1611507897104_0003
21/01/25 03:36:41 INFO mapreduce.Job: Job job_1611507897104_0003 running in uber mode : false
21/01/25 03:36:41 INFO mapreduce.Job: map 0% reduce 0%
```

P18: Problem solved

```
GC time elapsed (ms)=14024
CPU time spent (ms)=78590
```

P19: Terasort running time

```
[root@master mapreduce]# hdfs dfs -ls /terasort/
Found 4 items
drwxr-xr-x - root supergroup          0 2021-01-25 01:19 /terasort/1G-input
drwxr-xr-x - root supergroup          0 2021-01-25 02:46 /terasort/2G-input
drwxr-xr-x - root supergroup          0 2021-01-25 03:38 /terasort/2G-output
drwxr-xr-x - root supergroup          0 2021-01-25 03:42 /terasort/2G-validate
[root@master mapreduce]#
```

P20: Teravalidate succeeded

NEW
NEW SAVING
SUBMITTED
ACCEPTED
RUNNING
FINISHED
FAILED
KILLED

Scheduler

Tools

Active Nodes0

Decommissioning Nodes0

Decommissioned Nodes0

Lost Nodes0

Un0

Scheduler Metrics

Scheduler Type

Capacity Scheduler

Scheduling Resource Type

[<name=memory-mb default-unit=Mi type=COUNTABLE>, <name=vcores default-unit= type=COUNTABLE>]

Minimur

<memory:1024

Show 20 entries

ID	User	Name	Application Type	Queue	Application Priority	StartTime	LaunchTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU VCoers	Allic Me
application_1611507897104_0004	root	TeraValidate	MAPREDUCE	default	0	Mon Jan 25 03:42:03 +0800 2021	Mon Jan 25 03:42:03 +0800 2021	Mon Jan 25 03:42:31 +0800 2021	FINISHED	SUCCEEDED	N/A	N/A	N/A
application_1611507897104_0003	root	TeraSort	MAPREDUCE	default	0	Mon Jan 25 03:36:31 +0800 2021	Mon Jan 25 03:36:31 +0800 2021	Mon Jan 25 03:38:29 +0800 2021	FINISHED	SUCCEEDED	N/A	N/A	N/A
application_1611507897104_0002	root	TeraGen	MAPREDUCE	default	0	Mon Jan 25 02:46:08 +0800 2021	Mon Jan 25 02:46:08 +0800 2021	Mon Jan 25 02:46:54 +0800 2021	FINISHED	SUCCEEDED	N/A	N/A	N/A
application_1611507897104_0001	root	TeraGen	MAPREDUCE	default	0	Mon Jan 25 01:18:47 +0800 2021	Mon Jan 25 01:18:48 +0800 2021	Mon Jan 25 01:19:37 +0800 2021	FINISHED	SUCCEEDED	N/A	N/A	N/A

Showing 1 to 4 of 4 entries

P21: Three test process in yarn

Step5: Running python on Hadoop

In this part, I first create a folder named “WordCount” in hdfs, and then use the command “Hadoop fs – put” to put the two Shakespeare-basket into the folder, making it easier for me to count the word.

```
[root@master hadoop]# hdfs fs -put /usr/hadoop/shakespeare-basket1 /WordCount
错误: 找不到或无法加载主类 fs
[root@master hadoop]# hadoop fs -put /usr/hadoop/shakespeare-basket1 /WordCount
[root@master hadoop]# hadoop fs -ls /WordCount
Found 1 items
-rw-r--r-- 3 root supergroup 124620254 2021-01-25 04:26 /WordCount/shakespeare-basket1
[root@master hadoop]#
```

P22: Put Shakespeare-basket1 and Shakespeare-basket2 into WordCount


```

[root@master hduser]# echo "foo foo quux labs foo bar quux" | /home/hduser/mapper.py
/home/hduser/mapper.py:行4: import: 未找到命令
/home/hduser/mapper.py:行9: 未预期的符号 `line' 附近有语法错误
/home/hduser/mapper.py:行9:         line = line.strip()
[root@master hduser]# python --version
Python 2.7.5
[root@master hduser]# vim mapper.py
[root@master hduser]# echo "foo foo quux labs foo bar quux" | /home/hduser/mapper.py
foo      1
foo      1
quux     1
labs     1
foo      1
bar      1
quux     1

```

P23: Test mapper.py and reducer.py

```

root@master:/home/hduser
21/01/25 07:46:24 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
21/01/25 07:46:24 INFO resource.ResourceUtils: Adding resource type - name = memory-mb, units
= Mi, type = COUNTABLE
21/01/25 07:46:24 INFO resource.ResourceUtils: Adding resource type - name = vcores, units =
, type = COUNTABLE
21/01/25 07:46:24 INFO impl.YarnClientImpl: Submitted application application_1611507897104_0
010
21/01/25 07:46:24 INFO mapreduce.Job: The url to track the job: http://master:8088/proxy/appl
ication_1611507897104_0010/
21/01/25 07:46:24 INFO mapreduce.Job: Running job: job_1611507897104_0010
21/01/25 07:46:34 INFO mapreduce.Job: Job job_1611507897104_0010 running in uber mode : false
21/01/25 07:46:34 INFO mapreduce.Job: map 0% reduce 0%
21/01/25 07:46:59 INFO mapreduce.Job: map 28% reduce 0%
21/01/25 07:47:06 INFO mapreduce.Job: map 33% reduce 0%
21/01/25 07:47:12 INFO mapreduce.Job: map 44% reduce 0%
21/01/25 07:47:18 INFO mapreduce.Job: map 46% reduce 0%
21/01/25 07:47:19 INFO mapreduce.Job: map 50% reduce 0%
21/01/25 07:47:24 INFO mapreduce.Job: map 52% reduce 0%
21/01/25 07:47:25 INFO mapreduce.Job: map 53% reduce 0%
21/01/25 07:47:43 INFO mapreduce.Job: map 54% reduce 0%
21/01/25 07:47:47 INFO mapreduce.Job: map 54% reduce 11%
21/01/25 07:47:49 INFO mapreduce.Job: map 59% reduce 11%
21/01/25 07:47:50 INFO mapreduce.Job: map 60% reduce 11%
21/01/25 07:47:55 INFO mapreduce.Job: map 62% reduce 11%
21/01/25 07:47:56 INFO mapreduce.Job: map 66% reduce 11%
21/01/25 07:48:02 INFO mapreduce.Job: map 70% reduce 11%
21/01/25 07:48:08 INFO mapreduce.Job: map 75% reduce 11%
21/01/25 07:48:14 INFO mapreduce.Job: map 85% reduce 11%
21/01/25 07:48:17 INFO mapreduce.Job: map 85% reduce 22%
21/01/25 07:48:26 INFO mapreduce.Job: map 89% reduce 22%
21/01/25 07:48:40 INFO mapreduce.Job: map 92% reduce 22%
21/01/25 07:48:46 INFO mapreduce.Job: map 95% reduce 22%
21/01/25 07:48:52 INFO mapreduce.Job: map 100% reduce 22%
21/01/25 07:48:59 INFO mapreduce.Job: map 100% reduce 68%
21/01/25 07:49:05 INFO mapreduce.Job: map 100% reduce 72%
21/01/25 07:49:11 INFO mapreduce.Job: map 100% reduce 76%
21/01/25 07:49:17 INFO mapreduce.Job: map 100% reduce 80%
21/01/25 07:49:23 INFO mapreduce.Job: map 100% reduce 84%
21/01/25 07:49:30 INFO mapreduce.Job: map 100% reduce 87%
21/01/25 07:49:36 INFO mapreduce.Job: map 100% reduce 92%
21/01/25 07:49:42 INFO mapreduce.Job: map 100% reduce 95%
21/01/25 07:49:48 INFO mapreduce.Job: map 100% reduce 99%
21/01/25 07:49:49 INFO mapreduce.Job: map 100% reduce 100%
21/01/25 07:49:49 INFO mapreduce.Job: Job job_1611507897104_0010 completed successfully
21/01/25 07:49:49 INFO mapreduce.Job: Counters: 50
File System Counters
FILE: Number of bytes read=715074630
FILE: Number of bytes written=1083568264
FILE: Number of read operations=0

```

P24: Map/reduce process

Map-Reduce Framework

```
Map input records=4340061
Map output records=37015545
Map output bytes=293614077
Map output materialized bytes=367645185
Input split bytes=300
Combine input records=0
Combine output records=0
Reduce input groups=93358
Reduce shuffle bytes=367645185
Reduce input records=37015545
Reduce output records=93358
Spilled Records=109134525
Shuffled Maps =3
Failed Shuffles=0
Merged Map outputs=3
GC time elapsed (ms)=4744
CPU time spent (ms)=131100
Physical memory (bytes) snapshot=666931200
Virtual memory (bytes) snapshot=8372662272
Total committed heap usage (bytes)=447262720
```

P25: Running time of the python wordcount

```
warranteth      42
warranties      70
warrantise      6
warrantize      80
warrantor       1
warrants        248
warranty        331
warre 3
warred 16
warren 111
warren's      82
warrener      40
warrenton     4
warring 46
warrington    24
warrior 792
warrior''     4
warrior's     26
warriors      1046
warriorship   1
wars 6274
wars' 43
warsaw 20
```

P26: Part of the count results

Step6: Running java on Hadoop

(bonus)

```
root@master:usr/hadoop/hadoop-2.10.1
package org.myorg;

import java.io.IOException;
import java.util.*;

import org.apache.hadoop.fs.Path;
import org.apache.hadoop.conf.*;
import org.apache.hadoop.io.*;
import org.apache.hadoop.mapred.*;
import org.apache.hadoop.util.*;

public class WordCount {

    public static class Map extends MapReduceBase implements Mapper<LongWritable, Text, Text, IntWritable> {
        private final static IntWritable one = new IntWritable(1);
        private Text word = new Text();

        public void map(LongWritable key, Text value, OutputCollector<Text, IntWritable> output, Reporter reporter) throws IOException {
            String line = value.toString();
            StringTokenizer tokenizer = new StringTokenizer(line);
            while (tokenizer.hasMoreTokens()) {
                word.set(tokenizer.nextToken());
                output.collect(word, one);
            }
        }
    }

    public static class Reduce extends MapReduceBase implements Reducer<Text, IntWritable, Text, IntWritable> {
        public void reduce(Text key, Iterator<IntWritable> values, OutputCollector<Text, IntWritable> output, Reporter reporter) throws IOException {
            int sum = 0;
            while (values.hasNext()) {
                sum += values.next().get();
            }
            output.collect(key, new IntWritable(sum));
        }
    }
}
```

P29: Write java wordcount program

```
[root@master hadoop-2.10.1]# javac -classpath /usr/hadoop/hadoop-2.10.1/share/hadoop/common/hadoop-common-2.10.1.jar:/usr/hadoop/hadoop-2.10.1/share/hadoop/common/hadoop-hdfs-2.10.1.jar:/usr/hadoop/hadoop-2.10.1/share/hadoop/common/hadoop-client-2.10.1.jar:/usr/hadoop/hadoop-2.10.1/share/hadoop/common/hadoop-mapreduce-client-core-2.10.1.jar -d wordcount_classes/ WordCount.java
```


P30: Compile the java program we write

```
[root@master hadoop-2.10.1]# jar -cvf WordCount.jar -C wordcount_classes ./
已添加清单
正在添加: org/(输入 = 0) (输出 = 0) (存储了 0%)
正在添加: org/myorg/(输入 = 0) (输出 = 0) (存储了 0%)
正在添加: org/myorg/WordCount$Map.class(输入 = 1938) (输出 = 799) (压缩了 58%)
正在添加: org/myorg/WordCount$Reduce.class(输入 = 1611) (输出 = 649) (压缩了 59%)
正在添加: org/myorg/WordCount.class(输入 = 1534) (输出 = 752) (压缩了 50%)
正在添加: org/myorg/WordCount.jar(输入 = 338) (输出 = 183) (压缩了 45%)
[root@master hadoop-2.10.1]# ls
```

P31: Pack the jar package for wordcount

```
root@master:usr/hadoop/hadoop-2.10.1
21/01/26 12:39:34 INFO mapreduce.Job: map 68% reduce 11%
21/01/26 12:39:37 INFO mapreduce.Job: map 86% reduce 11%
21/01/26 12:39:41 INFO mapreduce.Job: map 100% reduce 22%
21/01/26 12:39:43 INFO mapreduce.Job: map 100% reduce 100%
21/01/26 12:39:43 INFO mapreduce.Job: Job job_1611627390332_0002 completed successfully
21/01/26 12:39:43 INFO mapreduce.Job: Counters: 50
  File System Counters
    FILE: Number of bytes read=5227603
    FILE: Number of bytes written=7902101
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=223927444
    HDFS: Number of bytes written=1092600
    HDFS: Number of read operations=12
    HDFS: Number of large read operations=0
    HDFS: Number of write operations=2
  Job Counters
    Killed map tasks=2
    Launched map tasks=5
    Launched reduce tasks=1
    Data-local map tasks=5
    Total time spent by all maps in occupied slots (ms)=171876
    Total time spent by all reduces in occupied slots (ms)=30992
    Total time spent by all map tasks (ms)=171876
    Total time spent by all reduce tasks (ms)=30992
    Total vcore-milliseconds taken by all map tasks=171876
    Total vcore-milliseconds taken by all reduce tasks=30992
    Total megabyte-milliseconds taken by all map tasks=176001024
    Total megabyte-milliseconds taken by all reduce tasks=31735808
  Map-Reduce Framework
    Map input records=4340061
    Map output records=37015545
    Map output bytes=367645167
    Map output materialized bytes=1839369
    Input split bytes=300
    Combine input records=37251276
    Combine output records=361259
    Reduce input groups=93358
    Reduce shuffle bytes=1839369
    Reduce input records=125528
    Reduce output records=93358
    Spilled Records=486787
    Shuffled Maps =3
    Failed Shuffles=0
```

P32: Map/reduce process and running details



All Applications

Cluster

About

Nodes

Node Labels

Applications

NEW

NEW SAVING

ACCEPTED

RUNNING

FINISHED

FAILED

KILLED

Scheduler

Tools

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Used Resources
1	0	0	1	0	<memory:0, vCores:0>

Cluster Nodes Metrics

Active Nodes	Decommissioning Nodes	Decommissioned Nodes	Lost Nodes	Unhealthy Nodes
3	0	0	0	0

Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Minimum Allocated
Capacity Scheduler	[<name=memory-mb default-unit=Mi type=COUNTABLE>, <name=vcores default-unit= type=COUNTABLE>]	<memory:1024, vCores:1>

Show 20 entries

ID	User	Name	Application Type	Queue	Application Priority	StartTime	LaunchTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU VCoers	Allocated Memory MB
application_1611627390332_0002	root	wordcount	MAPREDUCE	default	0	Tue Jan 26 12:38:32 +0800 2021	Tue Jan 26 12:38:34 +0800 2021	Tue Jan 26 12:39:42 +0800 2021	FINISHED	SUCCEEDED	N/A	N/A	N/A

Showing 1 to 1 of 1 entries

P33: Process shown on yarn

We can see the time spent on this wordcount program is 70 seconds (from start time to finish time), while the time spent on the python program is 204 seconds. So in my test, using java can take less time. But I don't know the principle, is it because the Hadoop framework is written in Java?