



哈爾濱工業大學 (深圳)
HARBIN INSTITUTE OF TECHNOLOGY

实验报告

开课学期: 2024 秋季

课程名称: 大数据导论

实验名称: _____

实验性质: 设计型

实验学时: 2 地点: T2608

学生班级: 22 级 5 班

学生学号: 220110519

学生姓名: 邢瑞龙

评阅教师: _____

报告成绩: _____

实验与创新实践教育中心制

2024 年 9 月

1. 实验目的

1. 熟悉 Hadoop 分布式集群的配置方法和基本操作；
2. 理解 MapReduce 的基本原理和框架；
3. 掌握 MapReduce 的基础编程方法，操作和运行 Mapreduce 作业。

2. 实验内容

1. 在大数据教学管理平台完成 Hadoop 完全分布式集群搭建；
2. 在 IntelliJ IDEA 中创建 MapReduce 工程，编码解决以下 3 个问题：
 - (1) 获取词频统计 Top 20 关键词
 - (2) 获取成绩表的最高分记录
 - (3) 统计网站每日的访问次数

3. 实验环境

- ✓ CentOS 7.9
- ✓ JDK 1.8
- ✓ Hadoop3.1.4
- ✓ IntelliJ IDEA 2022.2

4. 实验过程及结果

4.1 Hadoop 集群环境搭建

Hadoop 分布式环境搭建过程中的关键步骤截图，如命令运行结果，修改后的配置文件（使用 `cat` 命令查看）等。词频统计的结果文件需提交。

为 `etc/profile` 添加 java 的环境变量：

```
# You could check uidgid reservation validity in
# /usr/share/doc/setup-*/uidgid file
if [ $UID -gt 199 ] && [ "`/usr/bin/id -gn`" = "`/usr/bin/id -un`" ]; then
    umask 002
else
    umask 022
fi

for i in /etc/profile.d/*.sh /etc/profile.d/sh.local ; do
    if [ -r "$i" ]; then
        if [ "${-#*i}" != "$-" ]; then
            . "$i"
        else
            . "$i" >/dev/null
        fi
    fi
done

unset i
unset -f pathmunge
export JAVA_HOME=/usr/java/jdk1.8.0_281-amd64

export PATH=$PATH:$JAVA_HOME/bin
```

安装 JDK：

```
root@slave2-0 ~]# cd /data
[root@slave2-0 data]# rpm -ivh jdk-8u281-linux-x64.rpm
warning: jdk-8u281-linux-x64.rpm: Header V3 RSA/SHA256 Signature, key ID ec551f03: NOKEY
Preparing...                               ##### [100%]
Updating / installing...
 1: jdk1.8-2000:1.8.0_281-fcs               ##### [100%]
Unpacking JAR files...
  tools.jar...
  plugin.jar...
  javaws.jar...
  deploy.jar...
  rt.jar...
  jsse.jar...
  charsets.jar...
  localedata.jar...
[root@slave2-0 data]# vi /etc/profile
```

修改 Hadoop 的配置文件：（图为 `yarn-site.xml` 实例）：

```

</property>
<property>
  <name>yarn.nodemanager.resource.cpu-vcores</name>
  <value>1</value>
</property>
<property>
  <name>yarn.application.classpath </name>
  <value>
/usr/local/hadoop-3.1.4/etc/hadoop: /usr/local/hadoop-3.1.4/share/hadoop/common/lib/*: /usr/local/hadoop-3.1.4/share/hadoop/common/*: /usr/local/hadoop-3.1.4/share/hadoop/hdfs: /usr/local/hadoop-3.1.4/share/hadoop/hdfs/lib/*: /usr/local/hadoop-3.1.4/share/hadoop/hdfs/*: /usr/local/hadoop-3.1.4/share/hadoop/mapreduce/lib/*: /usr/local/hadoop-3.1.4/share/hadoop/mapreduce/*: /usr/local/hadoop-3.1.4/share/hadoop/yarn: /usr/local/hadoop-3.1.4/share/hadoop/yarn/lib/*: /usr/local/hadoop-3.1.4/share/hadoop/yarn/*
  </value>
</property>
</configuration>
[root@master-0 hadoop] #

```

修改 workers 文件:

```

[root@master-0 hadoop] # cat workers
slave1

slave2

```

修改 HDFS 的启动脚本 start-dfs.sh 与停止脚本 stop-dfs.sh:

```

[root@master-0 sbin] # head -15 start-dfs.sh
#!/usr/bin/env bash
HDFS_DATANODE_USER=root

HDFS_DATANODE_SECURE_USER=root

HDFS_NAMENODE_USER=root

HDFS_SECONDARYNAMENODE_USER=root

# 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

```

```
[root@master-0 sbin] # vi stop-dfs.sh
[root@master-0 sbin] # head -15 stop-dfs.sh
#!/usr/bin/env bash
HDFS_DATANODE_USER=root

HDFS_DATANODE_SECURE_USER=root

HDFS_NAMENODE_USER=root

HDFS_SECONDARYNAMENODE_USER=root
```

master 节点已经部署好的 Hadoop 与/etc/profile 文件复制传输到 slave1、slave2 节点：

BlacklistedNodesInfo.html	100%	7663	3.3MB/s	00:00
TasksInfo.html	100%	7891	3.9MB/s	00:00
JobTaskAttemptCounterInfo.html	100%	8439	3.8MB/s	00:00
TaskCounterInfo.html	100%	7769	3.1MB/s	00:00
ConfEntryInfo.html	100%	10KB	4.2MB/s	00:00
TaskAttemptsInfo.html	100%	8044	3.1MB/s	00:00
JobCounterInfo.html	100%	7810	285.9KB/s	00:00
AMAttemptsInfo.html	100%	7710	3.5MB/s	00:00
JobTaskAttemptState.html	100%	10KB	4.0MB/s	00:00
JobInfo.html	100%	12KB	5.0MB/s	00:00
JobsInfo.html	100%	7506	3.3MB/s	00:00
ReduceTaskAttemptInfo.html	100%	5198	2.4MB/s	00:00
JobTaskCounterInfo.html	100%	8092	3.0MB/s	00:00
ConfInfo.html	100%	7690	3.0MB/s	00:00
TaskCounterGroupInfo.html	100%	8608	3.7MB/s	00:00
CounterGroupInfo.html	100%	7785	3.0MB/s	00:00
AMAttemptInfo.html	100%	10KB	4.1MB/s	00:00
TasksInfo.html	100%	14KB	5.8MB/s	00:00
JobTaskAttemptCounterInfo.html	100%	14KB	5.7MB/s	00:00
TaskCounterInfo.html	100%	15KB	6.5MB/s	00:00
ConfEntryInfo.html	100%	19KB	8.1MB/s	00:00
TaskAttemptsInfo.html	100%	15KB	6.3MB/s	00:00
JobCounterInfo.html	100%	15KB	5.4MB/s	00:00

格式化成功：

```

2024-09-29 09:19:24,733 INFO util.GSet: capacity = 2^18 = 262144 entries
2024-09-29 09:19:24,741 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.window.num.buckets = 10
2024-09-29 09:19:24,741 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.num.users = 10
2024-09-29 09:19:24,741 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.windows.minutes = 1,5,25
2024-09-29 09:19:24,745 INFO namenode.FSNamesystem: Retry cache on namenode is enabled
2024-09-29 09:19:24,746 INFO namenode.FSNamesystem: Retry cache will use 0.03 of total heap and retry cache entry expiry time is 600000
    millis
2024-09-29 09:19:24,748 INFO util.GSet: Computing capacity for map NameNodeRetryCache
2024-09-29 09:19:24,748 INFO util.GSet: VM type = 64-bit
2024-09-29 09:19:24,748 INFO util.GSet: 0.029999999329447746% max memory 910.5 MB = 279.7 KB
2024-09-29 09:19:24,748 INFO util.GSet: capacity = 2^15 = 32768 entries
2024-09-29 09:19:24,835 INFO namenode.FSImage: Allocated new BlockPoolId: BP-1038372589-172.20.58.255-1727601564826
2024-09-29 09:19:24,854 INFO common.Storage: Storage directory /usr/local/hadoop-3.1.4/hdfs/name has been successfully formatted.
2024-09-29 09:19:24,928 INFO namenode.FSImageFormatProtobuf: Saving image file /usr/local/hadoop-3.1.4/hdfs/name/current/fsimage.ckpt_
3000000000000000000 using no compression
2024-09-29 09:19:25,052 INFO namenode.FSImageFormatProtobuf: Image file /usr/local/hadoop-3.1.4/hdfs/name/current/fsimage.ckpt_000000000
300000000000 of size 391 bytes saved in 0 seconds .
2024-09-29 09:19:25,079 INFO namenode.NNStorageRetentionManager: Going to retain 1 images with txid >= 0
2024-09-29 09:19:25,105 INFO namenode.FSImage: FSImageSaver clean checkpoint: txid = 0 when meet shutdown.
2024-09-29 09:19:25,106 INFO namenode.NameNode: SHUTDOWN_MSG:
/*****
SHUTDOWN_MSG: Shutting down NameNode at master/172.20.58.255
*****/

```

启动集群:

```

[ root@master-0 sbin] # start-dfs.sh
Starting namenodes on [master]
上一次登录: 三 9月 15 14:59:01 CST 2021
最后一次失败的登录: 四 4月 14 10:35:36 CST 2022从 localhostssh: notty 上
最有一次成功登录后有 1 次失败的登录尝试。
master: Warning: Permanently added 'master,172.20.58.255' (ECDSA) to the list of known hosts.
Starting datanodes
上一次登录: 日 9月 29 17:21:08 CST 2024pts/0 上
slave1: WARNING: /usr/local/hadoop-3.1.4/logs does not exist. Creating.
slave2: WARNING: /usr/local/hadoop-3.1.4/logs does not exist. Creating.
Starting secondary namenodes [master]
上一次登录: 日 9月 29 17:21:10 CST 2024pts/0 上

[ root@master-0 sbin] #
[ root@master-0 sbin] # jps
21077 NameNode
21382 SecondaryNameNode
23687 JobHistoryServer
23802 Jps
23148 ResourceManager

[ root@slave1-0 ~] # jps
589 Jps
534 NodeManager
429 DataNode

[ root@slave2-0 ~] # jps
401 DataNode
663 Jps
506 NodeManager

```

关闭集群:

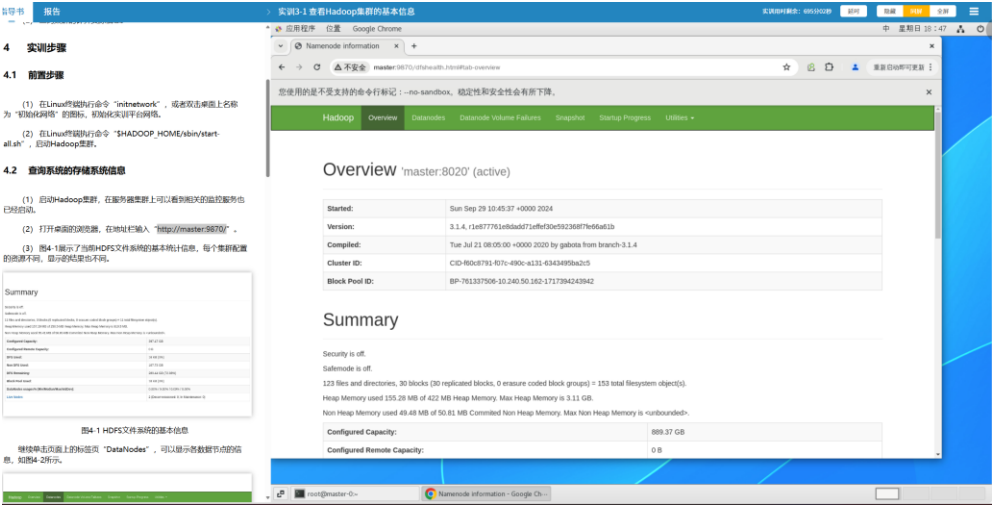
```

[ root@master-0 sbin] # stop-yarn.sh
Stopping nodemanagers
上一次登录: 日 9月 29 17:24:55 CST 2024pts/0 上
Stopping resourcemanager
上一次登录: 日 9月 29 17:27:41 CST 2024pts/0 上

```

```
[root@master-0 sbin]# stop-dfs.sh
Stopping namenodes on [master]
上一次登录：日 9月 29 17:27:44 CST 2024pts/0 上
Stopping datanodes
上一次登录：日 9月 29 17:29:19 CST 2024pts/0 上
Stopping secondary namenodes [master]
上一次登录：日 9月 29 17:29:20 CST 2024pts/0 上
```

查看集群信息：



Security is off.
Safemode is off.
123 files and directories, 30 blocks (30 replicated blocks, 0 erasure coded block groups) = 153 total filesystem object(s).
Heap Memory used 155.28 MB of 422 MB Heap Memory. Max Heap Memory is 3.11 GB.
Non Heap Memory used 49.48 MB of 50.81 MB Committed Non Heap Memory. Max Non Heap Memory is <unbounded>.

Configured Capacity:	889.37 GB
Configured Remote Capacity:	0 B
DFS Used:	3.05 MB (0%)
Non DFS Used:	242.39 GB
DFS Remaining:	646.98 GB (72.75%)
Block Pool Used:	3.05 MB (0%)
DataNodes usages% (Min/Median/Max/stdDev):	0.00% / 0.00% / 0.00% / 0.00%
Live Nodes	2 (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 (including replicas)	0
Block Deletion Start Time	Sun Sep 29 10:45:37 +0000 2024
Last Checkpoint Time	Sun Sep 29 10:45:38 +0000 2024

查看 hadoop 节点的计算资源：

实训3-1 查看Hadoop集群的基本信息

实训用时剩余: 68分20秒

超时

隐藏

打印

全屏

应用程序

位置

Google Chrome

中

星期日 18:59

Nodes of the cluster

不安全 master:8088/cluster/nodes

您使用的是不受支持的命令行标记: --no-sandbox。稳定性和安全性会有所下降。

Logged in as: root

Cluster

About

Nodes

Node Labels

Applications

NEW

NEW_SAVING

SUBMITTED

ACCEPTED

RUNNING

FINISHED

FAILED

KILLED

Scheduler

Tools

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memory Total	Memory Reserved	VCores Used	VCores Total	VCores Reserved
0	0	0	0	0	0 B	4 GB	0 B	0	16	0

Cluster Nodes Metrics

Active Nodes	Decommissioning Nodes	Decommissioned Nodes	Lost Nodes	Unhealthy Nodes	Rebooted Nodes	Shutdown Nodes
2	0	0	0	0	0	0

Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation	Maximum Cluster Application Priority
Capacity Scheduler	[memory-mb (unit=M), vcores]	<memory:1024, vCores:1>	<memory:2048, vCores:4>	0

Show 20 entries

Node Labels	Rack	Node State	Node Address	Node HTTP Address	Last health-update	Health-report	Containers	Allocation Tags	Mem Used	Mem Avail	VCores Used	VCores Avail	Version
/default-rack		RUNNING	slave1:41295	slave1:8042	星期日 九月 29 10:57:59 +0000 2024		0		0 B	2 GB	0	8	3.1.4
/default-rack		RUNNING	slave2:41591	slave2:8042	星期日 九月 29 10:57:59 +0000 2024		0		0 B	2 GB	0	8	3.1.4

Showing 1 to 2 of 2 entries

First

Previous

1

Next

Last

实训3-1 查看Hadoop集群的基本信息

实训用时剩余: 68分15秒

超时

隐藏

打印

全屏

应用程序

位置

Google Chrome

中

星期日 18:58

slave1:8042/node

不安全 slave1:8042/node

您使用的是不受支持的命令行标记: --no-sandbox。稳定性和安全性会有所下降。

ResourceManager

NodeManager

Node Information

List of Applications

List of Containers

Tools

NodeManager information

Total Vmem allocated for Containers	4.20 GB
Vmem enforcement enabled	false
Total Pmem allocated for Container	2 GB
Pmem enforcement enabled	true
Total VCoers allocated for Containers	8
Resource types	memory-mb (unit=M), vcores
NodeHealthyStatus	true
LastNodeHealthTime	Sun Sep 29 10:57:59 UTC 2024
NodeHealthReport	
NodeManager started on	Sun Sep 29 10:45:55 UTC 2024
NodeManager Version:	3.1.4 from 1e877761e8dadd71effef30e592368f7fe66a61b by gabota source checksum c366d34f26916ba3cdf69aef06f5fbb on 2020-07-21T08:10Z
Hadoop Version:	3.1.4 from 1e877761e8dadd71effef30e592368f7fe66a61b by gabota source checksum 38405c63945c88fd7a6fe391494799b on 2020-07-21T08:05Z

.8.

实训3-1 查看Hadoop集群的基本信息

应用程序 位置 终端

文件(F) 编辑(E) 查看(V) 搜索(S) 终端(T) 帮助(H)

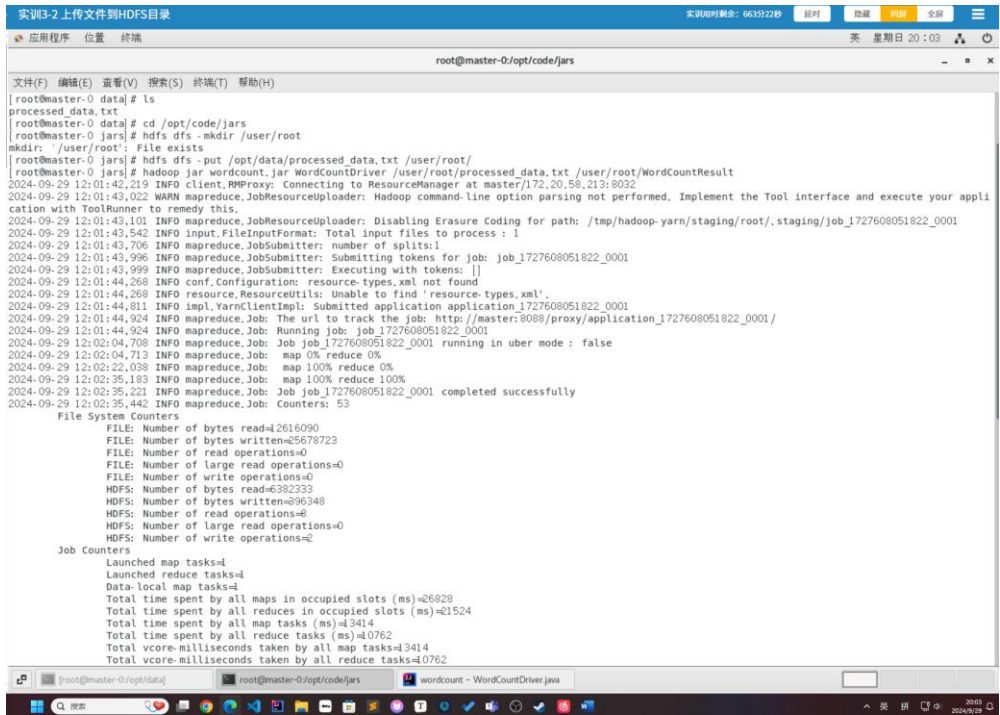
```
Starting nodemanagers
上一次登录：日 9月 29 18:45:50 CST 2024pts/0 上
[root@master-0 ~]# hdfs dfsadmin -report
Configured Capacity: 954953998336 (889.37 GB)
Present Capacity: 694621978624 (646.92 GB)
DFS Remaining: 694618783744 (646.91 GB)
DFS Used: 3194880 (3.05 MB)
DFS Used%: 0.00%
Replicated Blocks:
    Under replicated blocks: 0
    Blocks with corrupt replicas: 0
    Missing blocks: 0
    Missing blocks (with replication factor 1): 0
    Low redundancy blocks with highest priority to recover: 0
    Pending deletion blocks: 0
Erasure Coded Block Groups:
    Low redundancy block groups: 0
    Block groups with corrupt internal blocks: 0
    Missing block groups: 0
    Low redundancy blocks with highest priority to recover: 0
    Pending deletion blocks: 0
-----
Live datanodes (2):

Name: 172.20.85.202:9866 (slave1)
Hostname: slave1
Decommission Status : Normal
Configured Capacity: 477476999168 (444.69 GB)
DFS Used: 1597440 (1.52 MB)
Non DFS Used: 130166009856 (121.23 GB)
DFS Remaining: 347309391872 (323.46 GB)
DFS Used%: 0.00%
DFS Remaining%: 72.74%
Configured Cache Capacity: 0 (0 B)
Cache Used: 0 (0 B)
Cache Remaining: 0 (0 B)
Cache Used%: 100.00%
Cache Remaining%: 0.00%
Xceivers: 1
Last contact: Sun Sep 29 10:49:01 UTC 2024
Last Block Report: Sun Sep 29 10:45:46 UTC 2024
Num of Blocks: 30
```

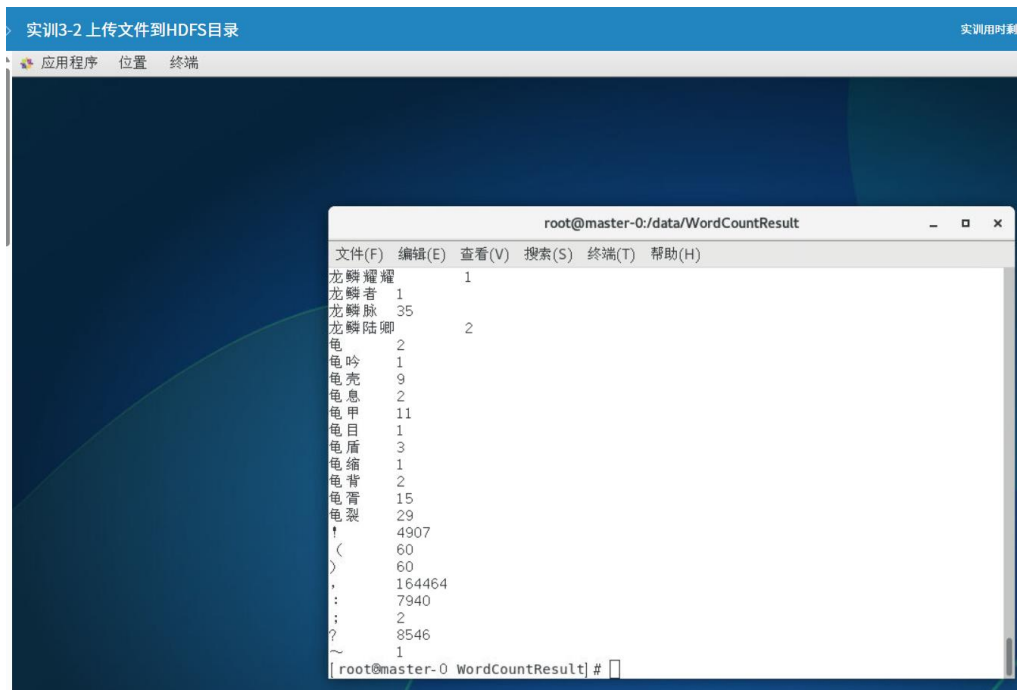
4.2 MapReduce 编程

针对以下每个问题，简单描述你的 Mapper 和 Reducer 模块的处理逻辑，并截图部分运行结果。每个项目 Mapper, Reducer 和 Driver 模块的代码文件 (*.java) 以及完整的运行结果也需在作业平台提交。

统计词频：



```
root@master-0:/opt/code/jars
[root@master-0 data] # ls
processed_data.txt
[root@master-0 data] # cd /opt/code/jars
[root@master-0 jars] # hdfs dfs -mkdir /user/root
mkdir: /user/root: File exists
[root@master-0 jars] # hdfs dfs -put /opt/data/processed_data.txt /user/root/
[root@master-0 jars] # hadoop jar wordcount.jar WordCountDriver /user/root/processed_data.txt /user/root/WordCountResult
2024-09-29 12:01:42,219 INFO client.RMProxy: Connecting to ResourceManager at master/172.20.58.213:8032
2024-09-29 12:01:43,022 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
2024-09-29 12:01:43,101 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/root/.staging/job_1727608051822_0001
2024-09-29 12:01:43,542 INFO input.FileInputFormat: Total input files to process : 1
2024-09-29 12:01:43,706 INFO mapreduce.JobSubmitter: number of splits:1
2024-09-29 12:01:43,996 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1727608051822_0001
2024-09-29 12:01:43,999 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-09-29 12:01:44,268 INFO conf.Configuration: resource-types.xml not found
2024-09-29 12:01:44,268 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2024-09-29 12:01:44,811 INFO impl.YarnClientImpl: Submitted application application_1727608051822_0001
2024-09-29 12:01:44,924 INFO mapreduce.Job: The url to track the job: http://master:8088/proxy/application_1727608051822_0001/
2024-09-29 12:01:44,924 INFO mapreduce.Job: Running job: job_1727608051822_0001
2024-09-29 12:02:04,708 INFO mapreduce.Job: Job job_1727608051822_0001 running in uber mode : false
2024-09-29 12:02:04,713 INFO mapreduce.Job: map 0% reduce 0%
2024-09-29 12:02:22,038 INFO mapreduce.Job: map 100% reduce 0%
2024-09-29 12:02:35,183 INFO mapreduce.Job: map 100% reduce 100%
2024-09-29 12:02:35,221 INFO mapreduce.Job: Job job_1727608051822_0001 completed successfully
2024-09-29 12:02:35,442 INFO mapreduce.Job: Counters: 53
File System Counters
  FILE: Number of bytes read=42616090
  FILE: Number of bytes written=2678723
  FILE: Number of read operations=0
  FILE: Number of large read operations=0
  FILE: Number of write operations=0
  HDFS: Number of bytes read=4382333
  HDFS: Number of bytes written=96348
  HDFS: Number of read operations=8
  HDFS: Number of large read operations=0
  HDFS: Number of write operations=2
Job Counters
  Launched map tasks=1
  Data-local map tasks=1
  Total time spent by all maps in occupied slots (ms)=26828
  Total time spent by all reduces in occupied slots (ms)=21524
  Total time spent by all map tasks (ms)=13414
  Total time spent by all reduce tasks (ms)=10762
  Total vcore-milliseconds taken by all map tasks=13414
  Total vcore-milliseconds taken by all reduce tasks=10762
```



```
root@master-0:/data/WordCountResult
文件(F) 编辑(E) 查看(V) 搜索(S) 终端(T) 帮助(H)
龙鳞耀耀 1
龙鳞者 1
龙鳞脉 35
龙鳞陆卿 2
龟 2
龟吟 1
龟壳 9
龟息 2
龟甲 11
龟目 1
龟盾 3
龟缩 1
龟背 2
龟膏 15
龟裂 29
! 4907
( 60
) 60
, 164464
: 7940
; 2
? 8546
~ 1
[root@master-0 WordCountResult] #
```

(1) 获取词频统计 Top 20 关键词

Map 按行处理每个单词的出现次数。

Reduce 按一个 key 一个 key 的来处理，以单词作为 key 值，汇总单个 key 对应的所有 value 值求和便是该单词出现的总次数。

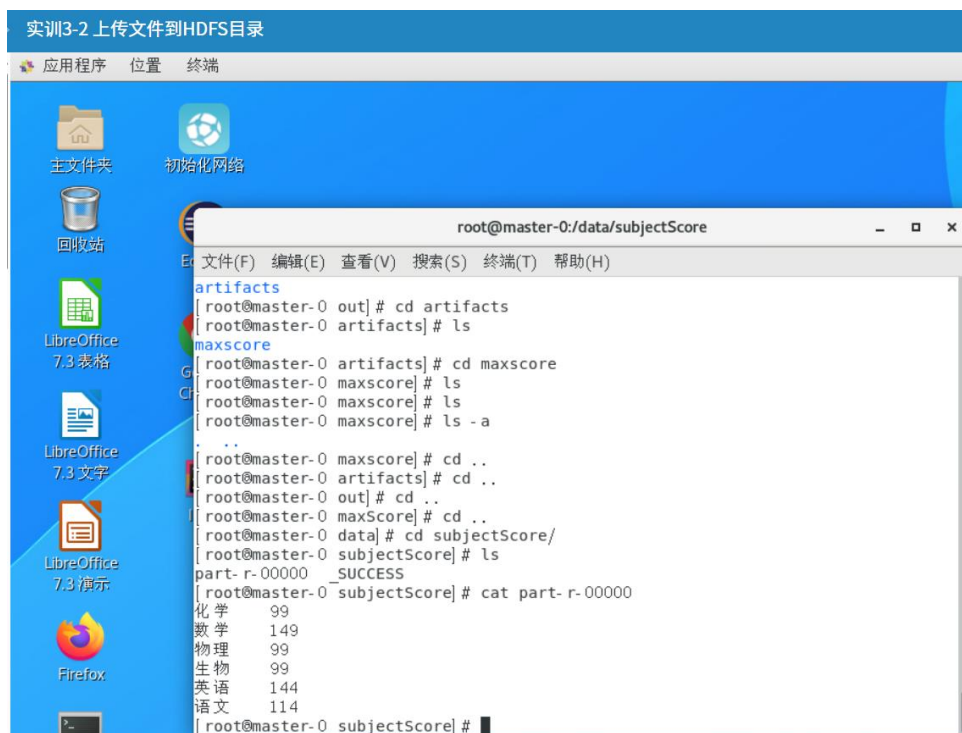
最后为了得到前 20 个词频最高的单词，在 reduce 中重写 clean_up, 使用优先队列的方式，优先队列从头到尾词频依次递增。若当前队列长度>20，则弹出队首。



(2) 获取成绩表的最高分记录

Map 按行处理，一次只处理一条记录，以课程作为 key，成绩为其 value。

Reduce 按 key 处理，取当前 key 中 value 值最高的，则为该门课程的最高成绩。



（3）统计网站每日的访问次数

Map 按行处理，一次只处理一条记录，将第五列中的字符串以空格作为分隔符得到访问日期，将此作为 key，value 为 1 对应一次访问。

Reduce 按 key 处理，累计 value 出现的个数，便是当前日期对应的访问次数。

实训3-2 上传文件到HDFS目录

应用程序 位置 终端

```
root@master-0:/data/visitCount
文件(F) 编辑(E) 查看(V) 搜索(S) 终端(T) 帮助(H)
[ root@master-0 visitCount] # cat p
part-r-00000 pom.xml
[ root@master-0 visitCount] # cat part-r-00000
2020/10/1 552
2020/10/10 1583
2020/10/11 1583
2020/10/12 1947
2020/10/13 1948
2020/10/14 2125
2020/10/15 1594
2020/10/16 1691
2020/10/17 1517
2020/10/18 1778
2020/10/19 1698
2020/10/2 583
2020/10/20 1785
2020/10/21 1759
2020/10/22 1948
2020/10/23 2196
2020/10/24 1993
2020/10/25 3246
2020/10/26 1445
2020/10/27 1271
2020/10/28 1338
```

个人签名： 邢瑞龙

2024 年 9 月 29 日