

# 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。*

1. 完成初始化网络、将 Java 安装包发送到子节点上等前置步骤：



```
root@master-0:~  
文件(F) 编辑(E) 查看(V) 搜索(S) 终端(T) 帮助(H)  
[root@master-0 ~]# initnetwork  
正在初始化网络  
Warning: Permanently added the ECDSA host key for IP address '172.20.248.61' to  
the list of known hosts.  
Warning: Permanently added the ECDSA host key for IP address '172.20.85.242' to  
the list of known hosts.  
初始化网络完成！
```

```

实训2-1 Java安装及Hadoop完全分布式集群搭建
实训用时剩余: 694分28秒 延时 隐藏 同屏 全屏
应用程序 位置 终端 英 星期日 16:31
root@master-0:~
文件(E) 编辑(E) 查看(V) 搜索(S) 终端(T) 帮助(H)
初始化网络完成!
[root@master-0 ~]#
[root@master-0 ~]# wget -P /data http://datasrc.tipdm.net:81/bigdata/SocialCompete/jdk-8u281-linux-x64.rpm
--2024-09-29 16:12:47-- http://datasrc.tipdm.net:81/bigdata/SocialCompete/jdk-8u281-linux-x64.rpm
正在解析主机 datasrc.tipdm.net (datasrc.tipdm.net)... 203.88.218.216
正在连接 datasrc.tipdm.net (datasrc.tipdm.net)|203.88.218.216|:81... 已连接。
已发出 HTTP 请求, 正在等待回应... 200 OK
长度: 113304268 (108M)
正在保存至: "/data/jdk-8u281-linux-x64.rpm"

100%[====>] 113,304,268 791KB/s 用时 2m 15s

2024-09-29 16:15:02 (819 KB/s) - 已保存 "/data/jdk-8u281-linux-x64.rpm" [113304268/113304268]

[root@master-0 ~]# wget -P /data http://datasrc.tipdm.net:81/bigdata/SocialCompete/hadoop-3.1.4.tar.gz
--2024-09-29 16:17:20-- http://datasrc.tipdm.net:81/bigdata/SocialCompete/hadoop-3.1.4.tar.gz
正在解析主机 datasrc.tipdm.net (datasrc.tipdm.net)... 203.88.218.216
正在连接 datasrc.tipdm.net (datasrc.tipdm.net)|203.88.218.216|:81... 已连接。
已发出 HTTP 请求, 正在等待回应... 200 OK
长度: 348326890 (332M) [application/x-gzip]
正在保存至: "/data/hadoop-3.1.4.tar.gz"

100%[====>] 348,326,890 316KB/s 用时 11m 54s

2024-09-29 16:29:15 (476 KB/s) - 已保存 "/data/hadoop-3.1.4.tar.gz" [348326890/348326890]

[root@master-0 ~]# ssh slave1 "mkdir -p /data"
[root@master-0 ~]# ssh slave2 "mkdir -p /data"
[root@master-0 ~]# scp /data/jdk-8u281-linux-x64.rpm slave1:/data/
jdk-8u281-linux-x64.rpm 100% 108MB 68.9MB/s 00:01
[root@master-0 ~]# scp /data/jdk-8u281-linux-x64.rpm slave2:/data/
jdk-8u281-linux-x64.rpm 100% 108MB 72.0MB/s 00:01
[root@master-0 ~]#

```

## 2. 在 Linux 下安装 Java

```

root@slave1-0:/data
文件(F) 编辑(E) 查看(V) 搜索(S) 终端(T) 帮助(H)
Java(TM) SE Runtime Environment (build 1.8.0_281-b09)
Java HotSpot(TM) 64-Bit Server VM (build 25.281-b09, mixed mode)
[root@master-0 data]# ssh slave1
System is booting up. See pam_nologin(8)
Last login: Sun Oct 9 09:36:24 2022 from master
[root@slave1-0 ~]# java -version
-bash: java: command not found
[root@slave1-0 ~]# cd /data
[root@slave1-0 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@slave1-0 data]#
bash: java: 未找到命令
[root@master-0 ~]# cd /data
[root@master-0 data]# rpm -ivh jdk-8u281-linux-x64.rpm
警告: jdk-8u281-linux-x64.rpm: 头V3 RSA/SHA256 Signature, 密钥 ID ec551f03: NOKEY
Y
准备中... ##### [100%]
正在升级/安装...
 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@master-0 data]# vi /etc/profile
[root@master-0 data]# source /etc/profile
[root@master-0 data]# java -version
java version "1.8.0_281"
Java(TM) SE Runtime Environment (build 1.8.0_281-b09)
Java HotSpot(TM) 64-Bit Server VM (build 25.281-b09, mixed mode)

```

进入节点 slave1 配置 Java:

```

root@master-0:/data
文件(F) 编辑(E) 查看(V) 搜索(S) 终端(T) 帮助(H)
warning: jdk-8u281-linux-x64.rpm: Header V3 RSA/SHA256 Signature, key ID ec551f0
3: 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...
  localdata.jar...
[root@slave1-0 data]# vi /etc/profile
[root@slave1-0 data]# source /etc/profile
[root@slave1-0 data]# java -version
java version "1.8.0_281"
Java(TM) SE Runtime Environment (build 1.8.0_281-b09)
Java HotSpot(TM) 64-Bit Server VM (build 25.281-b09, mixed mode)
[root@slave1-0 data]# exit
logout
Connection to slave1 closed.
[root@master-0 data]#

```

进入节点 slave2 配置 Java, 并在配置完成后通过命令“exit”退出 slave2, 回到 master 节点:

```

-bash: java: command not found
[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 ec551f0
3: 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...
  localdata.jar...
[root@slave2-0 data]# vi /etc/profile
[root@slave2-0 data]# source /etc/profile
[root@slave2-0 data]# java -version
java version "1.8.0_281"
Java(TM) SE Runtime Environment (build 1.8.0_281-b09)
Java HotSpot(TM) 64-Bit Server VM (build 25.281-b09, mixed mode)
[root@slave2-0 data]#

```

3. 在 master 节点上修改 Hadoop 配置文件:

```

Connection to slave2 closed.
[root@master-0 data]# tar -zxf /data/hadoop-3.1.4.tar.gz -C /usr/local
[root@master-0 data]# cd /usr/local/hadoop-3.1.4/etc/hadoop/
[root@master-0 hadoop]# vi
[root@master-0 hadoop]# vi core-site.xml
[root@master-0 hadoop]# export JAVA_HOME=/usr/java/jdk1.8.0_281-amd64
[root@master-0 hadoop]# vi mapred-site.xml
[root@master-0 hadoop]# vi yarn-site.xml
[root@master-0 hadoop]# vi workers
[root@master-0 hadoop]# vi hdfs-site.xml
[root@master-0 hadoop]# cd /usr/local/hadoop-3.1.4/sbin
[root@master-0 sbin]# vi start-dfs.sh
[root@master-0 sbin]# vi stop-dfs.sh
[root@master-0 sbin]# cd /usr/local/hadoop-3.1.4/sbin
[root@master-0 sbin]# vi start-yarn.sh
[root@master-0 sbin]# vi stop-yarn.sh
[root@master-0 sbin]#

```

## 4. 启动关闭集群:

配置环境变量:

```
[root@master-0 sbin] # vi /etc/profile
[root@master-0 sbin] # vi /etc/profile
[root@master-0 sbin] # source /etc/profile
[root@master-0 sbin] #
```

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

```
root@master-0:/usr/local/hadoop-3.1.4/sbin
文件(E) 编辑(E) 查看(V) 搜索(S) 终端(T) 帮助(H)
hadoop-archive-logs.sh 100% 1404 675.4KB/s 00:00
hadoop-streaming.sh 100% 1639 750.6KB/s 00:00
hadoop-openstack.sh 100% 1171 462.3KB/s 00:00
hadoop-aws.sh 100% 1272 583.7KB/s 00:00
hadoop-archives.sh 100% 1062 454.3KB/s 00:00
hadoop-sls.sh 100% 1252 648.5KB/s 00:00
hadoop-gridmix.sh 100% 1060 522.3KB/s 00:00
hadoop-distcp.sh 100% 1058 430.7KB/s 00:00
hadoop-resourceestimator.sh 100% 1266 565.0KB/s 00:00
hadoop-rumen.sh 100% 1056 411.4KB/s 00:00
hadoop-extras.sh 100% 1058 538.6KB/s 00:00
hadoop-archive-logs.sh 100% 1276 628.8KB/s 00:00
hadoop-streaming.sh 100% 1064 484.1KB/s 00:00
yarn-config.cmd 100% 2132 918.7KB/s 00:00
mapred-config.sh 100% 2808 1.0MB/s 00:00
hadoop-config.cmd 100% 8486 3.8MB/s 00:00
hdfs-config.cmd 100% 1640 697.0KB/s 00:00
[root@master-0 sbin] # scp /etc/profile slave1:/etc/profile
profile 100% 1963 804.6KB/s 00:00
[root@master-0 sbin] # scp /etc/profile slave2:/etc/profile
profile 100% 1963 938.0KB/s 00:00
[root@master-0 sbin] # ssh slave1 "source /etc/profile"
[root@master-0 sbin] # ssh slave2 "source /etc/profile"
[root@master-0 sbin] #
```

格式化:

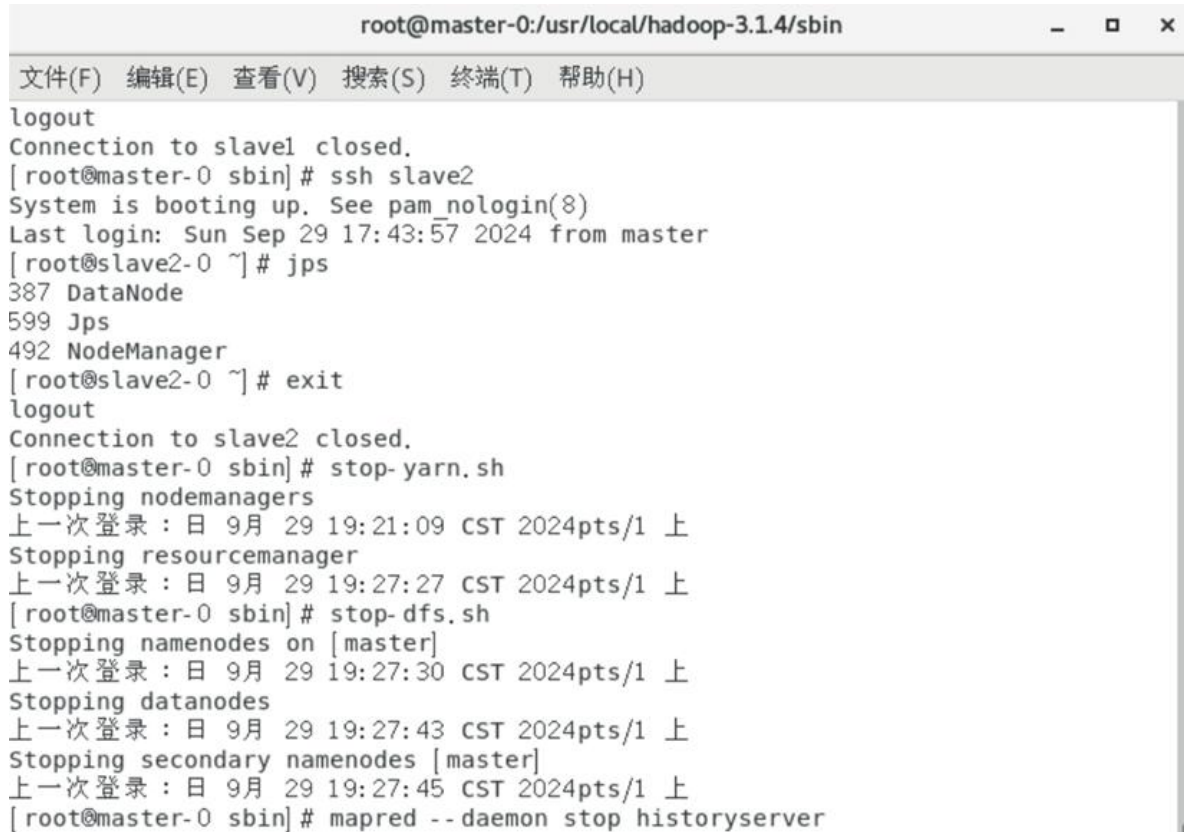
```
root@master-0:/usr/local/hadoop-3.1.4/sbin
文件(F) 编辑(E) 查看(V) 搜索(S) 终端(T) 帮助(H)
2024-09-29 09:23:31,855 INFO namenode.NameNode: Caching file names occurring more than 10 times
2024-09-29 09:23:31,903 INFO snapshot.SnapshotManager: Loaded config captureOpenFiles: false, skipCaptureAccessTi
meOnlyChange: false, snapshotDiffAllowSnapRootDescendant: true, maxSnapshotLimit: 65536
2024-09-29 09:23:31,906 INFO snapshot.SnapshotManager: SkipList is disabled
2024-09-29 09:23:31,912 INFO util.GSet: Computing capacity for map cachedBlocks
2024-09-29 09:23:31,912 INFO util.GSet: VM type = 64-bit
2024-09-29 09:23:31,913 INFO util.GSet: 0.25% max memory 910.5 MB = 2.3 MB
2024-09-29 09:23:31,913 INFO util.GSet: capacity = 2^18 = 262144 entries
2024-09-29 09:23:31,921 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.window.num.buckets = 10
2024-09-29 09:23:31,921 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.num.users = 10
2024-09-29 09:23:31,921 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.windows.minutes = 1,5,25
2024-09-29 09:23:31,925 INFO namenode.FSNamesystem: Retry cache on namenode is enabled
2024-09-29 09:23:31,925 INFO namenode.FSNamesystem: Retry cache will use 0.03 of total heap and retry cache entry
expiry time is 600000 millis
2024-09-29 09:23:31,926 INFO util.GSet: Computing capacity for map NameNodeRetryCache
2024-09-29 09:23:31,927 INFO util.GSet: VM type = 64-bit
2024-09-29 09:23:31,927 INFO util.GSet: 0.0299999999329447746% max memory 910.5 MB = 279.7 KB
2024-09-29 09:23:31,927 INFO util.GSet: capacity = 2^15 = 32768 entries
2024-09-29 09:23:31,947 INFO namenode.FSImage: Allocated new BlockPoolId: BP-839358983-172.20.58.216-172760181194
0
2024-09-29 09:23:32,007 INFO common.Storage: Storage directory /usr/local/hadoop-3.1.4/hdfs/name has been success
fully formatted.
2024-09-29 09:23:32,097 INFO namenode.FSImageFormatProtobuf: Saving image file /usr/local/hadoop-3.1.4/hdfs/name/
current/fsimage.ckpt_00000000000000000000 using no compression
2024-09-29 09:23:32,252 INFO namenode.FSImageFormatProtobuf: Image file /usr/local/hadoop-3.1.4/hdfs/name/current
/fsimage.ckpt_00000000000000000000 of size 391 bytes saved in 0 seconds.
2024-09-29 09:23:32,275 INFO namenode.NNStorageRetentionManager: Going to retain 1 images with txid >= 0
2024-09-29 09:23:32,282 INFO namenode.FSImage: FSImageSaver clean checkpoint: txid = 0 when meet shutdown.
2024-09-29 09:23:32,283 INFO namenode.NameNode: SHUTDOWN_MSG:
/*****
SHUTDOWN_MSG: Shutting down NameNode at master/172.20.58.216
*****/
[root@master-0 sbin] #
```



启动集群：

```
[root@master-0 sbin]# jps
62656 NameNode
64581 Jps
64202 JobHistoryServer
62955 SecondaryNameNode
63583 ResourceManager
[root@master-0 sbin]# ssh slavel
System is booting up. See pam_nologin(8)
Last login: Sun Sep 29 17:35:00 2024 from master
[root@slavel-0 ~]# jps
401 DataNode
613 Jps
506 NodeManager
[root@slavel-0 ~]# exit
logout
Connection to slavel closed.
[root@master-0 sbin]# ssh slave2
System is booting up. See pam_nologin(8)
Last login: Sun Sep 29 17:43:57 2024 from master
[root@slave2-0 ~]# jps
387 DataNode
599 Jps
492 NodeManager
```

关闭集群：



```
root@master-0:/usr/local/hadoop-3.1.4/sbin
文件(F) 编辑(E) 查看(V) 搜索(S) 终端(T) 帮助(H)
logout
Connection to slavel closed.
[root@master-0 sbin]# ssh slave2
System is booting up. See pam_nologin(8)
Last login: Sun Sep 29 17:43:57 2024 from master
[root@slave2-0 ~]# jps
387 DataNode
599 Jps
492 NodeManager
[root@slave2-0 ~]# exit
logout
Connection to slave2 closed.
[root@master-0 sbin]# stop-yarn.sh
Stopping nodemanagers
上一次登录：日 9月 29 19:21:09 CST 2024pts/1 上
Stopping resourcemanager
上一次登录：日 9月 29 19:27:27 CST 2024pts/1 上
[root@master-0 sbin]# stop-dfs.sh
Stopping namenodes on [master]
上一次登录：日 9月 29 19:27:30 CST 2024pts/1 上
Stopping datanodes
上一次登录：日 9月 29 19:27:43 CST 2024pts/1 上
Stopping secondary namenodes [master]
上一次登录：日 9月 29 19:27:45 CST 2024pts/1 上
[root@master-0 sbin]# mapred --daemon stop historyserver
```

## 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。

(1) 获取词频统计 Top 20 关键词

**Mapper 处理逻辑：**从输入中读取一行文本数据，转换为字符串 `line`。使用空格作为分隔符，将 `line` 切割成一个包含所有单词的数组 `words`。遍历 `words` 数组，对于每个单词，将单词封装为键 `k`（类型为 `Text`）。设置值 `v` 为整数 1（类型为 `IntWritable`）。使用 `context.write(k, v)` 输出键值对。

**Reducer 处理逻辑：**

在 `reduce` 方法中，对于每个单词（键 `key`），累加其对应的所有值（即该单词的出现次数）。累加结果不直接输出，而是存储在 `wordCounts` 的 `Map` 中，键为单词，值为累加次数。在 `cleanup` 方法中（`Reducer` 在所有 `reduce` 方法执行完毕后会调用此方法），将存储了所有单词及其出现次数的 `Map` 转换为一个 `List`，再对列表按照单词出现次数从高到低排序。遍历排序后的列表，输出出现次数最高的前 20 个单词及其对应的次数到上下文（`context`）。

```
[root@master-0 ~]# cd /data
[root@master-0 data]# cat Top20_result
电影      92611
世界      60862
生活      56578
蝶衣      55246
人生      35662
影片      35019
希望      34904
小楼      33783
阿甘      32801
现实      31463
这部      31036
自由      29582
楚门      28795
程蝶衣    27748
霸王      27483
故事      27023
监狱      25580
梦境      25057
安迪      24696
爱情      23799
```

(2) 获取成绩表的最高分记录

**Mapper 处理逻辑：**从输入文件中逐行读取数据，使用空白字符将每一行切割成两个部分：科目名称（`subject`）和分数（`score`）。将科目名称作为键（`Text` 类型），

分数作为值（IntWritable 类型）输出。相同科目的成绩被发送到同一个 Reducer 进行处理。

**Reducer 处理逻辑：**接收以科目名称为键，多个分数值为值的 Iterable 集合。初始化一个变量 maxScore，设置为最小整数值。遍历 Iterable 集合中的所有分数，使用 Math.max 函数比较当前值和 maxScore，更新 maxScore 为其中的最大值。最终将科目名称和对应的最高成绩作为键值对输出到结果文件（成绩表 B）。键为科目名称（Text），值为最高分数（IntWritable）。

```
[root@master-0 ~]# wget -P /opt/data http://datasrc.tipdm.net:81/bigdata/hadoop/data/subject_score.txt
--2024-10-07 21:43:21-- http://datasrc.tipdm.net:81/bigdata/hadoop/data/subject_score.txt
正在解析主机 datasrc.tipdm.net (datasrc.tipdm.net)... 192.168.1.158
正在连接 datasrc.tipdm.net (datasrc.tipdm.net)|192.168.1.158|:81... 已连接。
已发出 HTTP 请求，正在等待回应... 200 OK
长度：662601 (647K) [text/plain]
正在保存至: "/opt/data/subject_score.txt"

100%[=====>] 662,601 --.-K/s 用时 0.007s

2024-10-07 21:43:21 (87.3 MB/s) - 已保存 "/opt/data/subject_score.txt" [662601/662601])

[root@master-0 ~]# hdfs dfs -get /user/root/SubjectScoreResult/data
[root@master-0 ~]#
[root@master-0 ~]# cd /data
[root@master-0 data]# cd SubjectScoreResult/
[root@master-0 SubjectScoreResult]# ls -l
总用量 4
-rw-r--r-- 1 root root 63 10月 7 22:25 part-r-00000
-rw-r--r-- 1 root root 0 10月 7 22:25 _SUCCESS
[root@master-0 SubjectScoreResult]# cat part-r-00000
化学      99
数学      149
物理      99
生物      99
英语      144
语文      114
[root@master-0 SubjectScoreResult]#
```

### （3）统计网站每日的访问次数

**Mapper 模块处理逻辑：**从输入文件 raceData.csv 中逐行读取数据，使用逗号（,）作为分隔符，将每行数据拆分成一个字符串数组 fields。获取第 5 个字段 fields[4]，使用空白字符（如空格、制表符）将日期和时间拆分成 dateTimeParts 数组，从中提取日期部分，得到日期字符串。将提取的日期作为键（Text 类型），值为 1（IntWritable 类型），表示一次访问记录。调用 context.write(k, v) 将键值对发送到 Reducer。

**Reducer 模块处理逻辑：**接收按日期分组的键值对，键是日期，值是对应日期的所有访问次数（每个值都是 1）。对于每个日期，遍历其值列表 values，将所有值累加得到该日期的总访问次数 sum。将日期和总访问次数作为键值对输出。日期作为键（Text 类型），总访问次数作为值（IntWritable 类型）。调用 context.write(key, v) 将结果写入输出文件。

```
[root@master-0 ~]# wget -P /opt/data http://datasrc.tipdm.net:81/bigdata/hadoop/data/raceData.csv
--2024-10-07 22:37:45-- http://datasrc.tipdm.net:81/bigdata/hadoop/data/raceData.csv
正在解析主机 datasrc.tipdm.net (datasrc.tipdm.net)... 192.168.1.158
正在连接 datasrc.tipdm.net (datasrc.tipdm.net)|192.168.1.158|:81... 已连接。
已发出 HTTP 请求，正在等待回应... 200 OK
长度：51640901 (49M) [text/csv]
正在保存至: "/opt/data/raceData.csv"

100%|=====>| 51,640,901 112MB/s 用时 0.4s

2024-10-07 22:37:46 (112 MB/s) - 已保存 "/opt/data/raceData.csv" [51640901/51640901]

[root@master-0 ~]# mkdir -p /opt/code/jars
[root@master-0 ~]# cd /opt/code/jars
[root@master-0 jars]# hadoop jar wordcount.jar WordCountDriver /user/root/raceData.csv /user/root/DailyWebAccessCount
2024-10-07 16:13:15,994 INFO client.RMPProxy: Connecting to ResourceManager at master/172.20.217.189:8032
2024-10-07 16:13:16,799 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your ap
plication with ToolRunner to remedy this.
2024-10-07 16:13:16,819 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/root/.staging/job_1728308390091_0003
...
[root@master-0 jars]# hdfs dfs -get /user/root/DailyWebAccessCount /data
[root@master-0 jars]#
[root@master-0 jars]# cd /data
[root@master-0 data]# cd DailyWebAccessCount/
[root@master-0 DailyWebAccessCount]# ls -l
总用量 8
-rw-r--r-- 1 root root 4159 10月 8 00:17 part-r-00000
-rw-r--r-- 1 root root 0 10月 8 00:17 _SUCCESS
[root@master-0 DailyWebAccessCount]#
```

个人签名：

2024 年 10 月 7 日