第3章Hadoop基础操作实训

(源自:https://biglab.site)

(版本:Ver2.1-20230926)

CH03实训1统计文件中所有单词的平均长度

训练要点

1. 掌握HDFS的基本操作
2. 掌握提交MapReduce 任务
3. 掌握对MapReduce任务的查询

需求说明

将集群服务器的master目录/usr/local/hadoop-3.1.4/logs/hadoop-root-namenode-master.log上传到hdfs的 /user/myname目录下，然后对文件中的单词进行平均长度统计，将统计结果输出到/user/myname/output_namenode_wordmean目录下。

实现思路及步骤

1. 上传文件/usr/local/hadoop-3.1.4/logs/hadoop-root-namenode-master.log到 hdfs 的:/user/myname 目录下
2. 使用官方的/hadoop-mapreduce-examples-3.1.4jar提交MR任务，将结果输出到hdfs:/user/myname/output_namenode_wordmean目录下
3. 通过http://master:9870查看输出结果:查看hdfs:/user/myname/output_namenode_wordmean目录下结果

作业要求

1. 展示网络拓朴图。图中需说明master,slave的名称、IP、JPS进程名。
2. 在http://master:9870上拍照截取本小组集群中本成员目录下/user/myname中上传的文件，包含hadoop-root-namenode-master.log
3. 在linux的虚拟机中包含运行mr任务的命令行开始截图，以及运行结束时末尾截图
4. 在http://master:8088上截取与命令行对应的任务ID的记录的截图，和任务详细信息界面的截图
5. 截取在http://master:9870上/user/myname/output_namenode_wordmean/part-r-00000文件内容的截图

脚本参考

以下myname 要换成本人姓名的拼音或简拼； 以下命令在master上执行都可以。

```
hdfs dfs -mkdir -p /user/myname/
hdfs dfs -put /usr/local/hadoop-3.1.4/logs/hadoop-root-namenode-master.log \
/user/myname/
cd $HADOOP_HOME/share/hadoop/mapreduce/
hadoop jar ./hadoop-mapreduce-examples-3.1.4.jar wordmean \
/user/myname/hadoop-root-namenode-master.log \
/user/myname/output_namenode_wordmean
```

运行结果参考：

```
[root@master logs]# hdfs dfs -put /usr/local/hadoop-3.1.4/logs/hadoop-root-namenode-master.log /user/myname/
cd $HADOOP_HOME/share/hadoop/mapreduce/
hadoop jar ./hadoop-mapreduce-examples-3.1.4.jar wordmean /user/myname/hadoop-root-namenode-master.log /user/myname/output_namenode_wordmean
[root@master logs]# cd $HADOOP_HOME/share/hadoop/mapreduce/
[root@master mapreduce]# hadoop jar ./hadoop-mapreduce-examples-3.1.4.jar wordmean /user/myname/hadoop-root-namenode-master.log /user/myname/output_namenode_wordmean
2023-09-28 09:29:50,660 INFO client.RMProxy: Connecting to ResourceManager at master/192.168.128.130:8032
2023-09-28 09:29:51,617 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/root/.staging/job_1695864204511_0001
2023-09-28 09:29:52,095 INFO input.FileInputFormat: Total input files to process : 1
2023-09-28 09:29:52,264 INFO mapreduce.JobSubmitter: number of splits:1
```

```
2023-09-28 09:29:52,547 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1695864204511_0001
2023-09-28 09:29:52,549 INFO mapreduce.JobSubmitter: Executing with tokens: []
2023-09-28 09:29:52,951 INFO conf.Configuration: resource-types.xml not found
2023-09-28 09:29:52,951 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2023-09-28 09:29:53,422 INFO impl.YarnClientImpl: Submitted application application_1695864204511_0001
2023-09-28 09:29:53,490 INFO mapreduce.Job: The url to track the job: http://master:8088/proxy/application_1695864204511_0001/
2023-09-28 09:29:53,491 INFO mapreduce.Job: Running job: job_1695864204511_0001
2023-09-28 09:30:05,851 INFO mapreduce.Job: Job job_1695864204511_0001 running in uber mode : false
2023-09-28 09:30:05,853 INFO mapreduce.Job:  map 0% reduce 0%
2023-09-28 09:30:15,000 INFO mapreduce.Job:  map 100% reduce 0%
2023-09-28 09:30:22,083 INFO mapreduce.Job:  map 100% reduce 100%
2023-09-28 09:30:23,110 INFO mapreduce.Job: Job job_1695864204511_0001 completed successfully
2023-09-28 09:30:23,255 INFO mapreduce.Job: Counters: 53
        File System Counters
                FILE: Number of bytes read=39
                FILE: Number of bytes written=444095
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=535471
                HDFS: Number of bytes written=26
                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
                Launched reduce tasks=1
                Data-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=26292
                Total time spent by all reduces in occupied slots (ms)=17452
                Total time spent by all map tasks (ms)=6573
                Total time spent by all reduce tasks (ms)=4363
                Total vcore-milliseconds taken by all map tasks=6573
                Total vcore-milliseconds taken by all reduce tasks=4363
                Total megabyte-milliseconds taken by all map tasks=13461504
                Total megabyte-milliseconds taken by all reduce tasks=8935424
        Map-Reduce Framework
                Map input records=2268
                Map output records=49474
                Map output bytes=717373
                Map output materialized bytes=39
                Input split bytes=127
                Combine input records=49474
                Combine output records=2
                Reduce input groups=2
```

```
                Reduce shuffle bytes=39
                Reduce input records=2
                Reduce output records=2
                Spilled Records=4
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=253
                CPU time spent (ms)=3190
                Physical memory (bytes) snapshot=319643648
                Virtual memory (bytes) snapshot=7198687232
                Total committed heap usage (bytes)=140873728
                Peak Map Physical memory (bytes)=208363520
                Peak Map Virtual memory (bytes)=3594727424
                Peak Reduce Physical memory (bytes)=111280128
                Peak Reduce Virtual memory (bytes)=3603959808
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=535344
        File Output Format Counters
                Bytes Written=26
The mean is: 20.587338804220398
[root@master mapreduce]#
```

答题评分参考：

```
- 1，展示网络拓朴图。图中需说明master,slave的名称、IP、JPS进程名。 --20分
- 2，在http://master:9870上拍照截取本小组集群中本成员目录下/user/myname中上传的文件，包含hadoop-root-namenode-master.log --20分
- 3，在linux的虚拟机中包含运行mr任务的命令行开始截图，以及运行结束时末尾截图 --20分
- 4，在http://master:8088上截取与命令行对应的任务ID的记录的截图，和任务详细信息界面的截图 --20分
- 5，截取在http://master:9870上/user/myname/output_namenode_wordmean/part-r-00000文件内容的截图 --20分
```

CH03实训2查找大文件并进词频统计

训练要点

1. 掌握HDFS的操作方法
2. 掌握hadoop jar命令提交MapReduce任务的方法

需求说明

熟悉HDFS操作，包括上传文件； 熟悉hadoop jar进行词频统计操作

操作步骤

查大文件

```
find / -size -5000M -size +500M
/var/log/messages-20220208
```

上传文件

```
hdfs dfs -put /var/log/messages-20220208 /user/root/message.txt
```

词频统计

```
cd /usr/local/hadoop-3.3.1/share/hadoop/mapreduce
hadoop jar ./hadoop-mapreduce-examples-3.3.1.jar wordcount
/user/root/message.txt /user/root/output-message
```

使用hadoop日志

```
cd /usr/local/hadoop-3.3.1/logs
ls -lrt
```

上传Hadoop日志

```
hdfs dfs -put ./hadoop-root-nodemanager-c22.log /user/myname/
```

c22要改为具体的文件名

词频统计

```
cd /usr/local/hadoop-3.3.1/share/hadoop/mapreduce
hadoop jar ./hadoop-mapreduce-examples-3.3.1.jar wordcount /user/limm/hadoop-
root-nodemanager-c22.log /user/limm/output-nodemanager_log
```

CH03实训3查询与中断MapReduce任务

训练要点

1. 掌握查询 MapReduce任务信息
2. 掌握查询集群的计算资源信息
3. 掌握中断执行中的MapReduce任务

需求说明

在集群服务器的目录:本地目录/usr/local/hadoop-3.3.1/logs/下，将.log的文件上传集群hdfs的/user/myname/logs目录下
依次提交词频统计任务(wordcount)、计算平均长度任务(wordmean)、计算单词长度中位数任务(wordmedian)，查看当前集群的计算资源使用情况，以及任务列表信息，中断第2个任务(wordmean)，观察后续任务的执行情况

实现思路及步骤

- 上传日志文件hadoop-root-*.log到 hdfs的:/user/myname/logs 目录下
- 使用CRT分3次打开本组员服务器，这样有3个Tab页同时连接到同个服务器，分别运行wordcount 、wordmean 、wordmedian
- 进入http://master:8088站点，打开MapReduce任务列表，查看任务详细信息
- 找到wordmean的任务，进入该任务详细信息，然后中断它

作业要求

1. 环境说明:网络拓朴图
2. 在linux本组员的虚拟机上，截图运行 mr任务(wordcount、wordmean、wordmedian)的包含任务号的截图(至少3张)
   任务号如：job: http://master:8088/proxy/application_1651126254203_0008/
3. 在http://master:8088上拍照截取本组员运行的3个任务记录详细信息，其中wordcount(FINISHED)、wordmean(Killed)、wordmedian(FINISHED)

实现参考

准备数据

```
hdfs dfs -put /usr/local/hadoop-3.3.1/logs/* /user/myname/logs/
```

运行MR任务

```
cd /usr/local/hadoop-3.3.1/share/hadoop/mapreduce/

hadoop jar ./hadoop-mapreduce-examples-3.3.1.jar wordcount /user/myname/logs/* /user/myname/output_logs_wordcount
hadoop jar ./hadoop-mapreduce-examples-3.3.1.jar wordmean /user/myname/logs/* /user/myname/output_logs_wordmean
hadoop jar ./hadoop-mapreduce-examples-3.3.1.jar wordmedian /user/myname/logs/* /user/myname/output_logs_wordmedian
```

观察MR任务

进入http://master:8088站点，打开MapReduce任务列表

观察任务(wordcount、wordmean、wordmedian)任务运行情况，查看任务分别是由哪些节点来完成的，截图

找到第任务wordmean的任务，进入该任务详细信息，然后中断它，截图