# 作业 4

钟诚 16307110259

March 31, 2020

## 1. 统计其中各类文件的数量（按文件名后缀区分类型）

首先我们开启 HDFS，并将输入文件传入 HDFS 系统中

```
cd $HADOOP_HOME
./sbin/start-dfs.sh
./bin/hdfs dfs -mkdir -p /user/hadoop
./bin/hdfs dfs -mkdir input
./bin/hdfs dfs -put /home/hadoop/hadoop_test/sample_utf8.txt input
```

之后我们可以通过 jps 和 ls 指令查看 Datanode，Namenode 是否成功开启，输入文件是否成功传入，显示如下，则表示准备工作完成:

```
hadoop@dblab-VirtualBox:/usr/local/hadoop$ jps
2432 ResourceManager
2642 NodeManager
11364 Jps
7911 SecondaryNameNode
7528 NameNode
7688 DataNode
```

Figure 1: JPS

```
hadoop@dblab-VirtualBox:/usr/local/hadoop$ ./bin/hdfs dfs -ls input
Found 2 items
-rw-r--r--   1 hadoop supergroup       4725 2020-03-31 17:31 input/sample.txt
-rw-r--r--   1 hadoop supergroup       5000 2020-03-31 11:56 input/sample_utf8.txt
```

Figure 2: Input

随后我们对 mapper 和 reducer 程序进行编写

mapper.py:

```
#!/usr/bin python3
import sys
for line in sys.stdin:
    wtype =line.rstrip().split('.')[1]
    print("%s\t%s" % (wtype, 1))
```

reducer.py:

```
#!/usr/bin python3
from operator import itemgetter
import sys

current_key =None
total_count =0
total_number =0
for line in sys.stdin:
    line =line.strip()
    wtype, count, number =line.split('\t', 2)
```

```
    try:
        count =int(count)
        number =int(number)
    except ValueError:
        continue #如果无法取整，则不处理这一行
    if current_key ==wtype:
        #类别相同则累加
        total_count +=count
        total_number +=number
    else:
        if current_key: #如果不相同（证明上一个类型已经结束），则输出并更新
        print("%s\t%s\t%s" % (current_key, total_count, total_number))
            total_count =count
            total_number =number
            current_key =wtype
if word ==current_key: #最后输出当前类别
    print("%s\t%s\t%s" % (current_key, total_count, total_number))
```

最后，我们通过以下指令执行 MapReduce 进程并查看结果

```
./bin/hadoop jar /usr/local/hadoop/share/hadoop/tools/lib/hadoop-streaming-3.2.1.jar \
-file "/home/hadoop/hadoop_test/mapper.py" \
-file "/home/hadoop/hadoop_test/reducer.py" \
-mapper "python /home/hadoop/hadoop_test/mapper.py" \
-reducer "python /home/hadoop/hadoop_test/reducer.py" \
-input input/sample_utf8.txt -output Solution1
```

得到结果如下

```
hadoop@dblab-VirtualBox:/usr/local/hadoop$ ./bin/hdfs dfs -cat Solution1/*
2020-03-31 13:01:00,666 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted =
 false, remoteHostTrusted = false
docx    1
dwg     32
jpg     9
pdf     43
```

Figure 3: Question1 结果

**值得注意的是**:

1. Python 的头文件意味着这个文件的解释器，在我的 linux 系统的/!/usr/bin python3 中才有 python 编译器，所以我对 PPT 中头文件进行了修改
2. 在调用 Hadoop 进程时，"-mapper" 和"-reducer" 的后面都需要跟一个可执行对象，在使用"mapper.py" 时可能会出现权限问题，在经过"chmod +x"，调整无效后，使用 "python mapper.py" 的方法能够输出正确结果

## 2. 按文件的字节数大小降序排序输出文件名

因为在第一题中已经完成了准备工作，所以我们可以直接编写 mapper 和 reducer 程序

mapper2.py

```
#!/usr/bin python3
import sys
import re
for line in sys.stdin:
    l =re.sub('+ ',' ', line) #去掉多余的空格
    l =l.strip().split(' ')
    pic =l[-1]
    number =l[2].replace(',','')
    print("%s\t%s" % (number.pic)) # 这里为了后面排序将数字第一个输出
```

reducer2.py

```
#!/usr/bin python3
import sys
for line in sys.stdin:
```

```
line =line.strip().split('\t')
wtype =line[1]
count =line[0]
print("%s\t%s" % (wtype , count)) #再正序输出
```

同样的，我们通过以下指令执行 MapReduce 进程，因为我们需要将数值降序输出，所以再调用进程时需要做一些修改

```
./bin/hadoop jar /usr/local/hadoop/share/hadoop/tools/lib/hadoop-streaming-3.2.1.jar \
-D mapred.output.key.comparator.class=org.apache.hadoop.mapred.lib.KeyFieldBasedComparator \
-D mapred.text.key.comparator.options=-k1nr \
-file "/home/hadoop/hadoop_test/mapper.py" \
-file "/home/hadoop/hadoop_test/reducer.py" \
-mapper "python /home/hadoop/hadoop_test/mapper2.py" \
-reducer "python /home/hadoop/hadoop_test/reducer2.py" \
-input input/sample_utf8.txt -output Solution1
```

**值得注意的是**:

1. -D 是提交功能参数的一种方式，但使用-D 时必须将参数放置在所有参数之首
2. org.apache.hadoop.mapred.lib.KeyFieldBasedComparator 是按键来对 map 结果进行排序的一种方式
3. 在参数中，-k1nr 指对第 1 列按数值形式降序排列，所以在 mapper 的输出中要将数值输出在前面
4. 在本实验中，KeyFieldBasedComparator 隐含了以下假设：默认的分隔符是"\t"，默认的键为分隔符分隔结果的第一位，如不符合这一假设，则需要添加更多的参数

同样的，查看相关结果（篇幅所限只展示部分结果）:

```
hadoop@dblab-VirtualBox:/usr/local/hadoop$ ./bin/hdfs dfs -cat Solution2/*
2020-03-31 19:23:46,804 INFO sasl.SaslDataTransferClient: SASL encryption trust check:
总体-弱施-00-02.pdf     5272668
总体-弱施-00-01.pdf     5203305
典型幕墙详图.dwg        4520750
一层平面图.dwg  3395145
核心筒详图（一）.dwg    3389704
核心筒详图（二）.dwg    3389704
核心筒详图（三）.dwg    3389704
核心筒详图（四）.dwg    3389704
核心筒详图（五）.dwg    3389704
核心筒详图（六）.dwg    3389704
核心筒详图（八）.dwg    3389704
核心筒详图（七）.dwg    3389704
00-04.pdf       2605370
00-05.pdf       2263846
00-06.pdf       2238600
1-1剖面.dwg     2201666
00-11.pdf       2088121
00-12.pdf       1940285
00-08.pdf       1904441
00-09.pdf       1824144
00-13.pdf       1746362
00-03.pdf       1652761
21-15立面图.dwg 1607088
15-21立面图.dwg 1607088
A-N立面图.dwg   1607088
N-A立面图.dwg   1607088
00-10.pdf       1511442
00-14.pdf       1482151
03-01.pdf       1326463
一层防火分区图.dwg      1326450
00-07.pdf       1003206
标准层防火分区图.dwg    992840
屋面层平面图.dwg        989338
机房层、屋顶平面图.dwg  989338
```

Figure 4: Question2 结果