# Dig Data Management Assignment 2

# Task2 - Word Count Task on a Serie of Novels Based on Hadoop MapReduce

In this task, we count the words in a series of Novels applying Hadoop MapReduce framework.

# **How We Implement**

In this task, we use Python to go through it. Hadoop provides a streaming method to make it possible to run MapReduce tasks for all programming languages.

Here shows the simplistic codes of mapper,

```
#!/usr/bin/python3.5
# mapper.py
import sys
import jieba

for line in sys.stdin:
words = jieba.cut(line.strip(), cut_all=False, HMM=False)

for word in words:
    if len(word) > 1:
        print ('%s\t%s' % (word, 1))
```

It's quite simple and straightforward. We can design any parsing pattern to build up our key-value pairs.

```
1 #!/usr/bin/python3.5
 2 # reducer.py
 3 import sys
 4
 5 current_word, current_count = None, 0
 6
 7
   for line in sys.stdin:
        word, count = line.strip().split('\t', 1)
 8
 9
        count = int(count)
10
      if current_word == word:
11
12
           current_count += count
13
       else:
           if current_count:
14
                print('%s\t%s' % (current_word, current_count))
15
            current_count, current_word = count, word
16
17 if current_count:
18
        print('%s\t%s' % (current_word, current_count))
```

We need to treat the key-value pairs as string and do customized parsing here, and then we do a simple counting task. Note that the standard input key-value pairs are sorted by key before passed into reducers, it

### **Result of Word Count**

#### Top 10

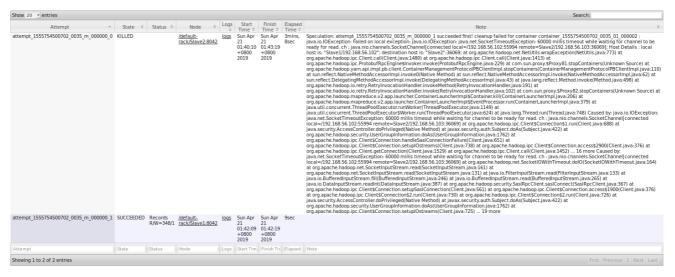
	337.334	
1	说道	13714
2	什么	12738
3	自己	10841
4	韦小宝	9920
5	一个	9361
6	咱们	7017
7	武功	6804
8	一声	6520
9	不是	6125
10	师父	6071

#### Last 10 (Chinese only)

1	一不怕苦	1
2	一下脸	1
3	一下手	1
4	一下子把	1
5	一万间	1
6	一万户	1
7	一万多	1
8	一万六千多	1
9	一万余	1
10	一万九千	1

# **Observations**

1. Kiled/Failed Tasks would be transfered to another nodes to run



2. Error caused from lack of RAM (exit code=137)

We need to set yarn-site.xml appropriately to solve the problem over lack of RAM. yarn.nodemanager.resource.memory-mb is set to be 1600 to solve the error this time.

- 3. Some files and packages(libraries) are required to be allocated and installed on every single nodes. In this task, bugs did torture us for a while for we didn't install package 'Jieba' for Python on Slave1 and Slave2 nodes.
- 4. The command format matters and diverses in different verions when you use streaming method to run MapReduce tasks. The final legal format turned out to be something like

```
hadoop jar /usr/local/hadoop/hadoop-2.7.7/share/hadoop/tools/lib/hadoop-streaming-2.7.7.jar
mapper user in the stream in th
```

# Reflections

- 1. This time, we're rather more familiar with the relation and involved items of 'NameNode', 'DataNode', 'SourceManager' and 'NodeManager'.
- 2. Trying to reset configuration such as yarn-site.xml to solve bugs help us be more into the structure of Hadoop.