

BIG DATA

SECTION D

Fall'2023

Yogya Sharma

ys5250

Practice Assignment : Jupyter Notebook

Exercise 1:

For 100 strings:

```
list_of_strings = ['abc', 'bigData', 'Students']
large_list_of_strings = list_of_strings*100
%time max_length = print(find_longest_string(list_of_strings))
```

```
Students
CPU times: user 367 µs, sys: 1.02 ms, total: 1.38 ms
Wall time: 1.39 ms
```

For 1000 strings:

```
list_of_strings = ['abc', 'bigData', 'Students']
large_list_of_strings = list_of_strings*1000
%time max_length = max(large_list_of_strings, key=len)
```

```
CPU times: user 65 µs, sys: 0 ns, total: 65 µs
Wall time: 68.9 µs
```

For 1 million strings:

```
: large_list_of_strings = list_of_strings*100000000
%time max_length = max(large_list_of_strings, key=len)
```

```
CPU times: user 4.93 s, sys: 56.7 ms, total: 4.98 s
Wall time: 5 s
```

For 1 million strings using map reduce:

```
In [5]: import functools
|
| list_of_strings = ['abc', 'bigData', 'Students']
| large_list_of_strings = list_of_strings*100000000
| mapper = len
| def reducer(p, c):
|     if p[1] > c[1]:
|         return p
|     return c
| mapped = map(mapper, list_of_strings)
| mapped = zip(list_of_strings, mapped)
| #step 2:
| %time reduced = functools.reduce(reducer, mapped)
| print(reduced)
```

```
CPU times: user 8 µs, sys: 7 µs, total: 15 µs
Wall time: 18.8 µs
('Students', 8)
```

Exercise 2:

Word count on Jupyter:

```
In [7]: import functools
        from collections import Counter

        # Define the mapper function to split text into words
        def mapper(text):
            words = text.split()
            return Counter(words)

        def reducer(d1, d2):
            return d1 + d2

        # Read input text from a file
        with open('text2.txt', 'r') as file:
            text = file.read()

        %time word_counts = functools.reduce(reducer, map(mapper, text.splitlines()), Counter())

        for word, frequency in word_counts.items():
            print(f'Word: {word}, Frequency: {frequency}')
```

```
CPU times: user 54.7 s, sys: 309 ms, total: 55 s
Wall time: 55.7 s
Word: Warning, Frequency: 123553
Word: bells, Frequency: 179712
Word: sounded, Frequency: 179712
Word: fairly, Frequency: 179712
Word: early, Frequency: 179712
Word: with, Frequency: 359424
Word: one, Frequency: 179712
Word: of, Frequency: 718848
Word: the, Frequency: 898560
Word: participants,, Frequency: 179712
Word: who, Frequency: 359424
```

Word count using map reduce Hadoop:

SSH-in-browser

UPLOAD FILE DOWNLOAD FILE

```
ys5250_nyu_edu@nyu-dataproc-m1:~$ hadoop jar $HADOOP_HOME/hadoop-streaming-3.2.2.jar -input text2.txt -output outputpython -mapper "python mapper.py" -reducer "python reducer.py" -file
mapper.py -file reducer.py
2023-11-15 23:51:39,262 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packagel.jar: [mapper.py, reducer.py] [/usr/lib/hadoop/hadoop-streaming-3.2.2.jar] /tmp/streamjob7501967169589011822.jar tmpDir=null
2023-11-15 23:51:39,232 INFO client.RMProxy: Connecting to ResourceManager at nyu-dataproc-m/192.168.1.38:8032
2023-11-15 23:51:39,430 INFO client.AHSProxy: Connecting to Application History server at nyu-dataproc-m/192.168.1.38:10200
2023-11-15 23:51:39,471 INFO client.RMProxy: Connecting to ResourceManager at nyu-dataproc-m/192.168.1.38:8032
2023-11-15 23:51:39,471 INFO client.AHSProxy: Connecting to Application History server at nyu-dataproc-m/192.168.1.38:10200
2023-11-15 23:51:40,252 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/ys5250_nyu_edu/.staging/job_1691775874963_32133
2023-11-15 23:51:40,529 INFO mapred.FileInputFormat: Total input files to process : 1
2023-11-15 23:51:40,586 INFO mapreduce.JobSubmitter: number of splits:141
2023-11-15 23:51:40,737 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1691775874963_32133
2023-11-15 23:51:40,739 INFO mapreduce.JobSubmitter: Executing with tokens: []
2023-11-15 23:51:40,905 INFO conf.Configuration: resource-types.xml not found
2023-11-15 23:51:40,906 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2023-11-15 23:51:41,064 INFO impl.YarnClientImpl: Submitted application application_1691775874963_32133
2023-11-15 23:51:41,098 INFO mapreduce.Job: The url to track the job: http://nyu-dataproc-m:8088/proxy/application_1691775874963_32133/
2023-11-15 23:51:41,099 INFO mapreduce.Job: Running job: job_1691775874963_32133
2023-11-15 23:51:49,330 INFO mapreduce.Job: Job job_1691775874963_32133 running in uber mode : false
2023-11-15 23:51:49,331 INFO mapreduce.Job: map 0% reduce 0%
2023-11-15 23:51:59,444 INFO mapreduce.Job: map 7% reduce 0%
2023-11-15 23:52:00,451 INFO mapreduce.Job: map 13% reduce 0%
2023-11-15 23:52:01,462 INFO mapreduce.Job: map 24% reduce 0%
2023-11-15 23:52:08,518 INFO mapreduce.Job: map 28% reduce 0%
2023-11-15 23:52:09,525 INFO mapreduce.Job: map 39% reduce 0%
2023-11-15 23:52:10,544 INFO mapreduce.Job: map 41% reduce 0%
2023-11-15 23:52:11,551 INFO mapreduce.Job: map 46% reduce 0%
2023-11-15 23:52:12,556 INFO mapreduce.Job: map 48% reduce 0%
2023-11-15 23:52:17,595 INFO mapreduce.Job: map 51% reduce 0%
2023-11-15 23:52:18,614 INFO mapreduce.Job: map 57% reduce 0%
2023-11-15 23:52:19,619 INFO mapreduce.Job: map 60% reduce 0%
2023-11-15 23:52:20,634 INFO mapreduce.Job: map 65% reduce 0%
2023-11-15 23:52:21,639 INFO mapreduce.Job: map 69% reduce 0%
2023-11-15 23:52:22,645 INFO mapreduce.Job: map 72% reduce 0%
2023-11-15 23:52:26,670 INFO mapreduce.Job: map 75% reduce 0%
2023-11-15 23:52:27,676 INFO mapreduce.Job: map 81% reduce 0%
2023-11-15 23:52:28,681 INFO mapreduce.Job: map 82% reduce 0%
2023-11-15 23:52:29,689 INFO mapreduce.Job: map 87% reduce 0%
2023-11-15 23:52:30,694 INFO mapreduce.Job: map 90% reduce 0%
2023-11-15 23:52:31,699 INFO mapreduce.Job: map 92% reduce 0%
2023-11-15 23:52:32,705 INFO mapreduce.Job: map 100% reduce 0%
2023-11-15 23:52:39,737 INFO mapreduce.Job: map 100% reduce 2%
2023-11-15 23:52:40,742 INFO mapreduce.Job: map 100% reduce 13%
2023-11-15 23:52:41,751 INFO mapreduce.Job: map 100% reduce 17%
2023-11-15 23:52:42,756 INFO mapreduce.Job: map 100% reduce 23%
2023-11-15 23:52:43,760 INFO mapreduce.Job: map 100% reduce 28%
2023-11-15 23:52:44,765 INFO mapreduce.Job: map 100% reduce 30%
2023-11-15 23:52:45,780 INFO mapreduce.Job: map 100% reduce 34%
2023-11-15 23:52:46,785 INFO mapreduce.Job: map 100% reduce 40%
2023-11-15 23:52:47,790 INFO mapreduce.Job: map 100% reduce 45%
2023-11-15 23:52:48,795 INFO mapreduce.Job: map 100% reduce 53%
2023-11-15 23:52:49,801 INFO mapreduce.Job: map 100% reduce 62%
```

Transferred 3 items

Transferred 3 items

2023.11.15 23:51:40 UTC	2023.11.15 23:51:47 UTC	2023.11.15 23:52:53 UTC	job_1691775874963_32133	streamjob7501967169589011822.jar	ys5250_nyu_edu	nyu-dataproc-	default	SUCCEEDED	141	141	47	47	00hrs, 01mins, 05sec
-------------------------------	-------------------------------	-------------------------------	-------------------------	----------------------------------	----------------	---------------	---------	-----------	-----	-----	----	----	----------------------------

Here the elapsed time for Hadoop job is : 1 min 5 seconds

And for JupyterHub is: 56 seconds