

# WordCount - MapReduce

## Computer Specifications:

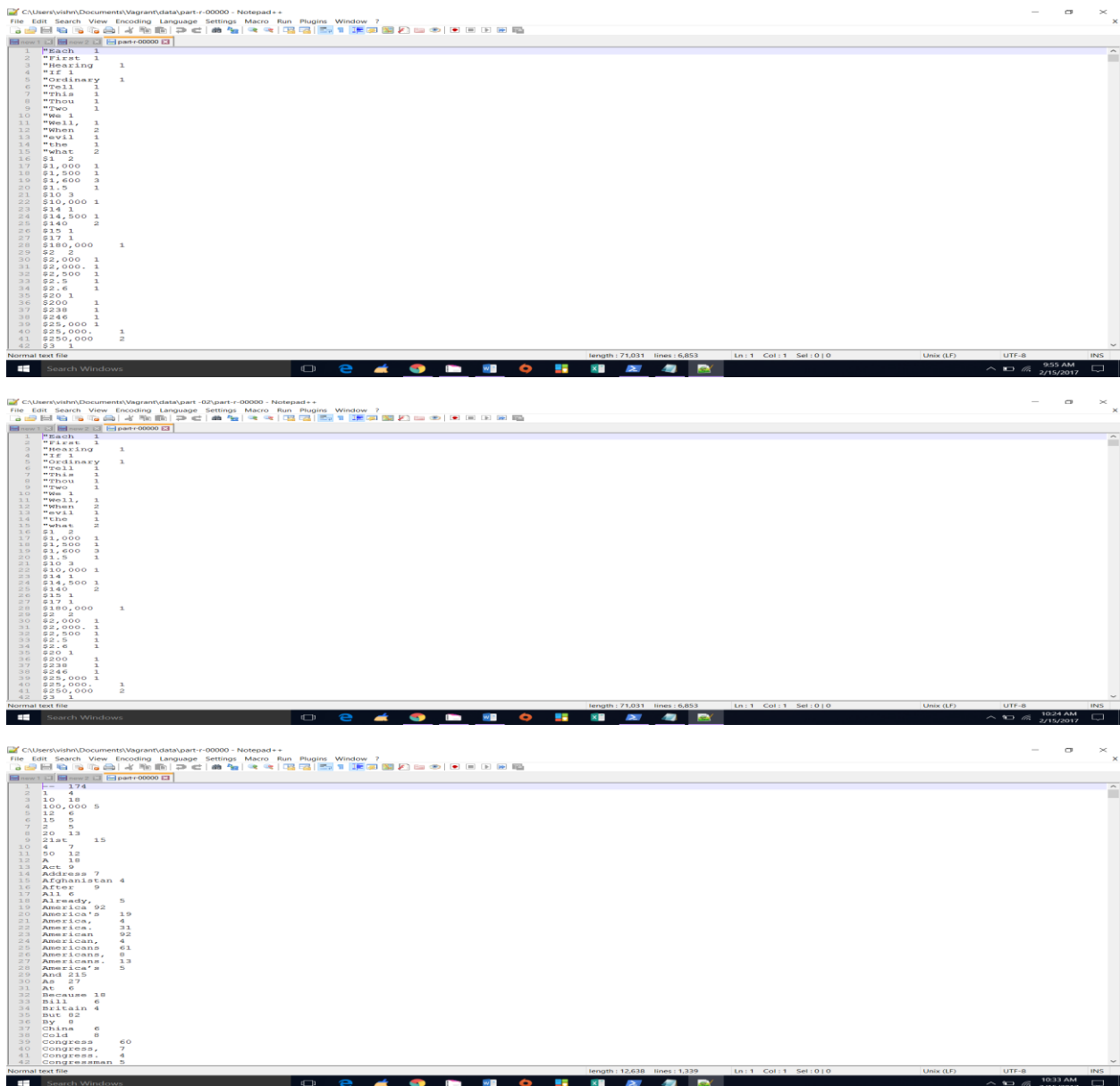
### Config:

- VRAM – 6144 MB
- PROCESSOR – 2 Core (Virtual machine)

### VMachine Used:

- ubuntu/trusty64

### Output Screenshot:



```
1 "Reach 1
2 "First 1
3 "Hearling 1
4 "if 1
5 "Ordinary 1
6 "Tell 1
7 "Thla 1
8 "Thou 1
9 "Two 1
10 "We 1
11 "Well 1
12 "When 2
13 "evil 1
14 "the 1
15 "What 2
16 $1 2
17 $1,000 1
18 $1,500 1
19 $1,500 3
20 $1.5 1
21 $10 3
22 $10,000 1
23 $14 1
24 $14,500 1
25 $145 2
26 $15 1
27 $17 1
28 $180,000 1
29 $2 2
30 $2,000 1
31 $2,000 1
32 $2,500 1
33 $2.5 1
34 $2.6 1
35 $20 1
36 $200 1
37 $238 1
38 $246 1
39 $25,000 1
40 $25,000 1
41 $250,000 2
42 $ 1
```

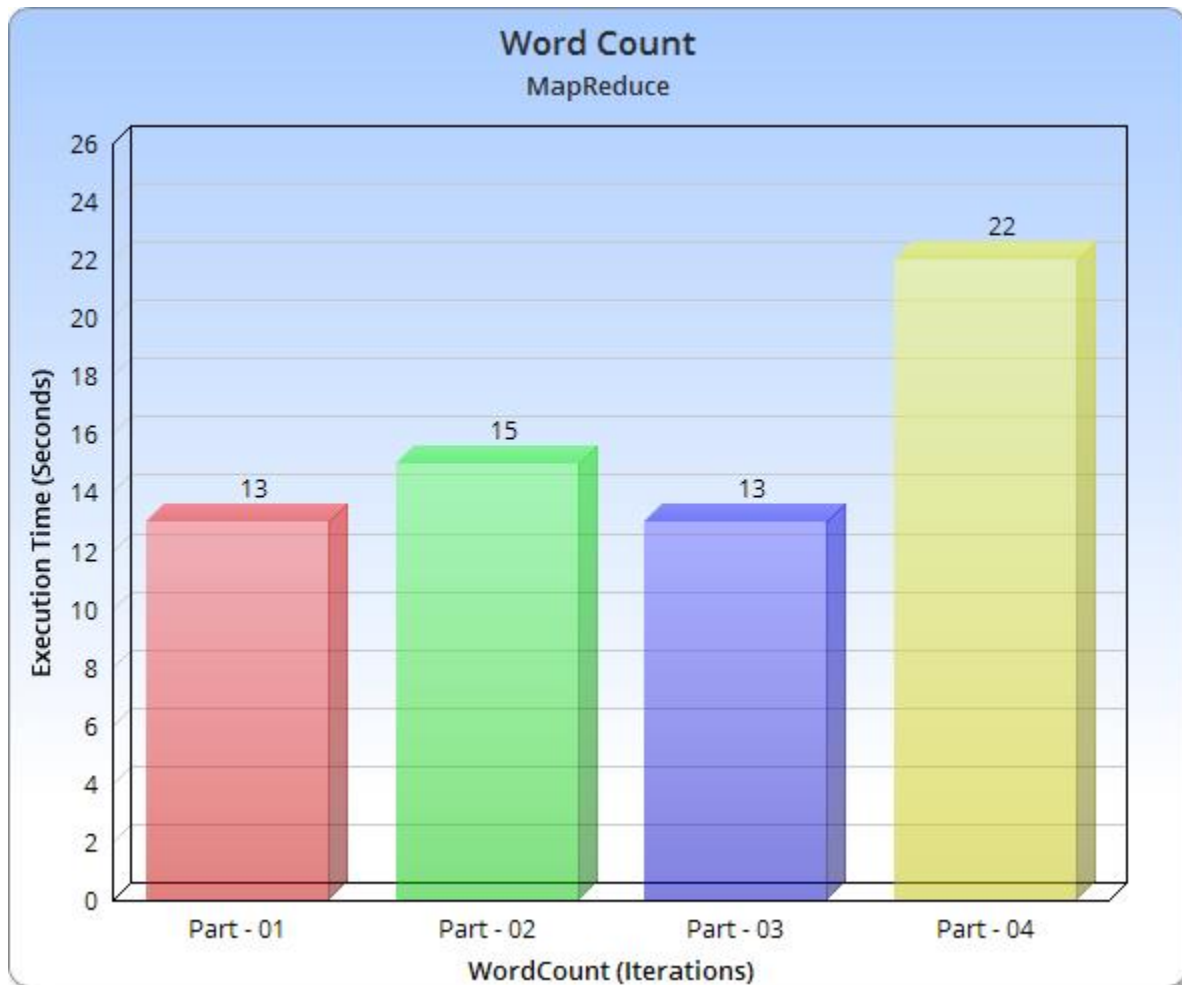
```
1 "Reach 1
2 "First 1
3 "Hearling 1
4 "if 1
5 "Ordinary 1
6 "Tell 1
7 "Thla 1
8 "Thou 1
9 "Two 1
10 "We 1
11 "Well 1
12 "When 2
13 "evil 1
14 "the 1
15 "What 2
16 $1 2
17 $1,000 1
18 $1,500 1
19 $1,500 3
20 $1.5 1
21 $10 3
22 $10,000 1
23 $14 1
24 $14,500 1
25 $145 2
26 $15 1
27 $17 1
28 $180,000 1
29 $2 2
30 $2,000 1
31 $2,000 1
32 $2,500 1
33 $2.5 1
34 $2.6 1
35 $20 1
36 $200 1
37 $238 1
38 $246 1
39 $25,000 1
40 $25,000 1
41 $250,000 2
42 $ 1
```

```
1 I- 174
2 1 4
3 10 10
4 100,000 5
5 12 6
6 15 5
7 2 5
8 20 13
9 21st 15
10 4 7
11 50 12
12 A 19
13 Ark 9
14 Address 7
15 Afghanistan 4
16 After 9
17 All 6
18 Already 5
19 America 92
20 America's 19
21 America, 4
22 American, 31
23 American 92
24 American, 4
25 American's 21
26 American's, 8
27 Americans, 13
28 America's 5
29 And 215
30 Ap 27
31 At 6
32 Because 18
33 Bill 6
34 Britain 4
35 But 82
36 By 8
37 China 6
38 Cold 8
39 Congress 60
40 Congress, 7
41 Congress, 8
42 Congressmen 5
```

## Chart & Graph:

S.No	teration	Total Maps
1	Part - 01 - Run Word Count 1 example on your local psudo-distributed system with supplied text files	1
2	Part - 02 - Run Word Count 2 example on your local psudo-distributed system with supplied text files	1
3	Part - 03 - Modify Wordcount 1 to look for only words that occur more than 4 times	1
4	Part - 04 - Modify Wordcount 2 to modify and use the -skip command	1

S.No	Start Time	End Time	Execution Time	Time Difference
1	15:48:35	15:48:48	<u>13s</u>	0s
2	16:20:19	16:20:34	<u>15s</u>	299s
3	16:32:23	16:32:36	<u>13s</u>	331s
4	16:42:20	16:42:42	<u>22s</u>	331s



### **Additional Runs:**

- WordCount1 was executed with words repeating more than 10 times.
- WordCount2 was executed with both Case-Sensitive: ON and Case-Sensitive: OFF. Both the execution took approx. ~14s to finish its run.

### **Analysis:**

- The run times of all the word count files took not more than 25 seconds to execute.
- Word Count 2 improves on Word Count by taking some of the features of MapReduce and implementing it.
- Most noticeable with Word Count 2 is that how DistributedCache can be used to distribute read-only data enhancing functions like case sensitivity, skipping selected patterns or words.
- It also makes use of Hadoop's inline commands providing a more specific result based on a given input.
- Larger datasets related to Word Count files can be set up in a cluster to improve efficiency in reading and parsing through data, provided the dataset is substantial (100GB or more).
- This exercise took just 1 map to run hence it can be seen that the complexity is really small.

### **Top 10 Words:**

The most frequently used words under all 4 tasks can be seen below.

- I 353
- a 757
- and 1217
- for 445
- in 640
- is 393
- of 1142
- our 657
- that 571
- the 1867
- to 1433
- we 560
- will 394