List 2 - Introduction to Big Data Analytics

Wojciech Korczyński, 229949

Every task was solved using map and reduce approach. There were utilized both serial and parallel processing. The both modes were compared.

Task 1

Below are presented results for one run of the program from task 1. Computational time for parallel processing is longer than for serial one.

```
n_rows = 100000
n_cols = 250
min_value = 3
max_value = 7

Serialized
Result: 12499374
Elapsed time: 1.851048231124878 s

Parallel (multiprocessing)
Result: 12499374
Elapsed time: 2.1980035305023193 s
```

In figure 1 is shown plot comparing time of executing the task for different number of rows in matrix. We see that for very large matrices serial processing is slower than parallel one.

Task 2

Below is presented the result of running the program for task 2. It counts words from 160 books. The working is quicker for parallel processing. There is also a list of top20 popular words in the books.

```
Books number: 160
SERIAL PROCESSING
Elapsed time: 17.70080327987671 s

Top 20 counts
1.one: 25199
2.said: 20829
3.would: 18893
4.man: 14622
5.little: 13077
6.could: 12884
7.upon: 12398
8.like: 11820
9.time: 11645
10.see: 11021
11.well: 10494
12.may: 10416
13.great: 9948
14.us: 9890
15.sir: 9729
```

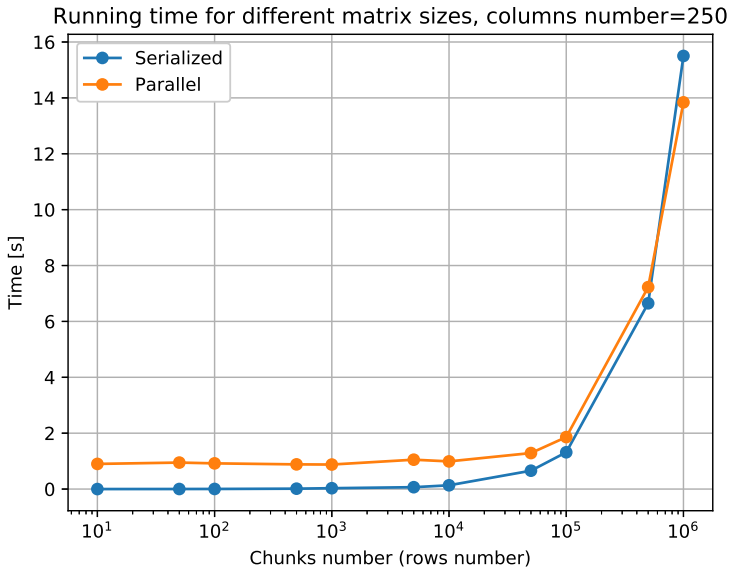

Figure 1: Comparison of running time for task 1

```
16.know: 9626
17.must: 9600
18.good: 9505
19.two: 9204
20.mr: 8940


PARALLEL PROCESSING
Elapsed time: 10.203799486160278 s

Top 20 counts
1.one: 25199
2.said: 20829
3.would: 18893
4.man: 14622
5.little: 13077
6.could: 12884
7.upon: 12398
8.like: 11820
9.time: 11645
10.see: 11021
11.well: 10494
12.may: 10416
13.great: 9948
14.us: 9890
15.sir: 9729
16.know: 9626
17.must: 9600
18.good: 9505
19.two: 9204
20.mr: 8940
```

In figure 2 is presented the histogram of top20 words. It shows that the most popular words are: one, said, would.

In figure 3 is shown plot comparing time of executing the task for different number of lines from books. The more lines is taken into account the better results are obtained by parallel processing. We see that for number of lines greater than 300000 it overcomes serial processing.

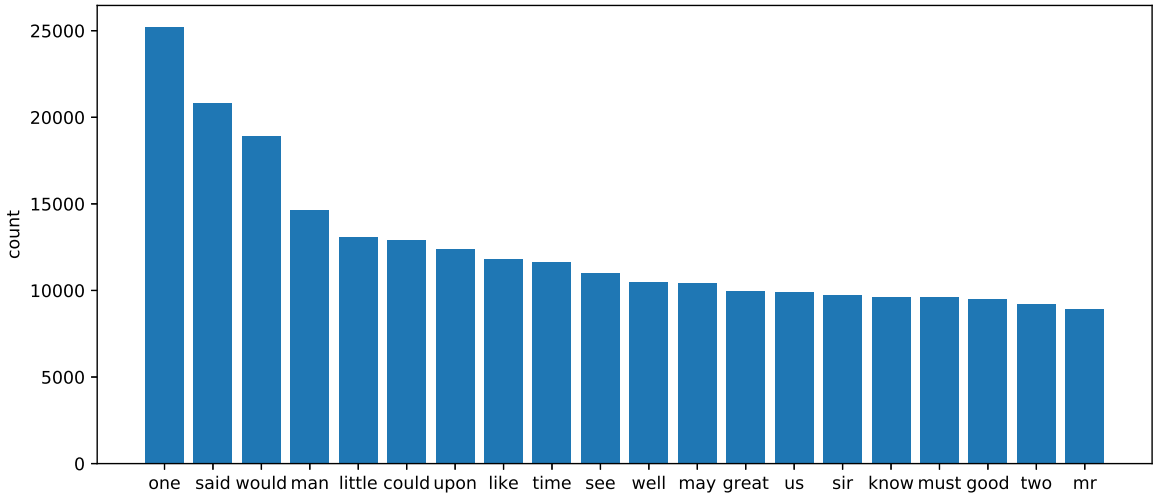

Figure 2: Words count histogram of top20 popular words

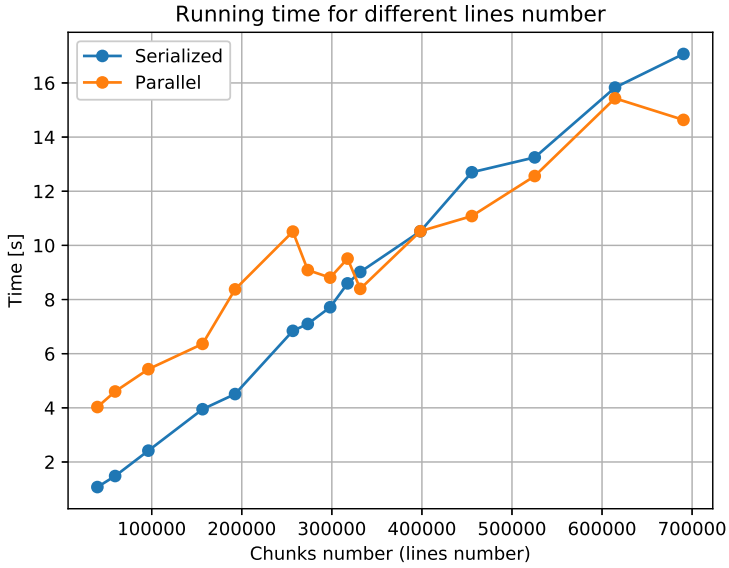

Figure 3: Comparison of running time for task 2

Task 3

Program in the task 3 determines digits of π number in hexadecimal system. The result of its working is presented below. The results are the time and the time for parallel solving is longer.

```
Digits number: 30
SERIAL PROCESSING
Elapsed time: 0.0019948482513427734 s

3.243f6a8885a308d313198a2e03707
```

```
#########################

PARALLEL PROCESSING
Elapsed time: 1.1191227436065674 s

3.243f6a8885a308d313198a2e03707
```

Time of executing the task for different manner and different number of digits is presented In figure 4. Parallel processing acts quicker when number of digits is greater than 700. The difference of times is considerable for larger numbers of digits.

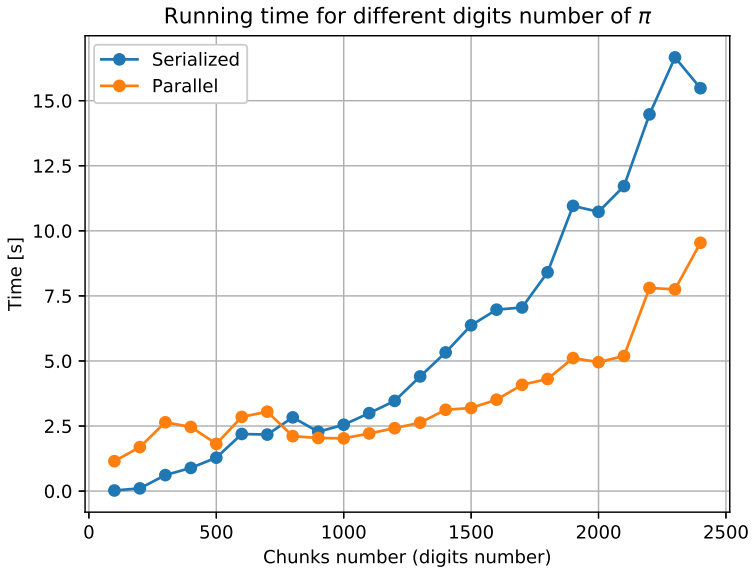

Figure 4: Comparison of running time for task 3

Task 4

Below are presented results obtained by program for small values of `m`, `n` and `p`. The calculation by serial manner are quicker than by parallel one. We see that the results are the same

```
Matrix M: m by n
Matrix N: n by p
m = 5
n = 10
p = 6

SERIAL PROCESSING
Elapsed time: 0.005982637405395508 s
```

```
[[2.57128946 1.51606242 1.70109693 2.01234079 1.52709571 2.10624917]
 [3.03178124 1.50608367 1.87539867 2.5081463  2.42226483 2.53311597]
 [2.84532629 1.62797553 1.52370304 2.37486291 2.35279493 2.24003674]
 [2.1205462  1.44278819 1.28043423 1.68366984 1.6840366  1.99467547]
 [3.61051389 2.03450888 2.23501249 1.91066708 2.28958498 2.55739363]]

#########################

PARALLEL PROCESSING
Elapsed time: 1.0362296104431152 s

[[2.57128946 1.51606242 1.70109693 2.01234079 1.52709571 2.10624917]
 [3.03178124 1.50608367 1.87539867 2.5081463  2.42226483 2.53311597]
 [2.84532629 1.62797553 1.52370304 2.37486291 2.35279493 2.24003674]
 [2.1205462  1.44278819 1.28043423 1.68366984 1.6840366  1.99467547]
 [3.61051389 2.03450888 2.23501249 1.91066708 2.28958498 2.55739363]]
```

Below are presented results obtained by program for larger values of `m`, `n` and `p`. The calculation time by parallel manner is slightly longer than by serial processing.

```
Matrix M: m by n
Matrix N: n by p
m = 100
n = 1500
p = 75

SERIAL PROCESSING
Elapsed time: 29.52592968940735 s

PARALLEL PROCESSING
Elapsed time: 44.67364764213562 s
```

Conclusions

Running of various type of task in the map and reduce manner showed that for larger number of chunks it is preferable to use parallel processing. In that case calculation time is less than for normal serial manner.