

# Отчёт

Студент: Алиев Ф. Р. Группа: M16-522

Ссылка на проект: [https://github.com/FirdavsRobertovich/DSBDA\\_ALIEV\\_HW1](https://github.com/FirdavsRobertovich/DSBDA_ALIEV_HW1)

Задача.

Необходимо найти самое длинное слово в .txt файле. Формат выходного файла – CSV.

Реализация.

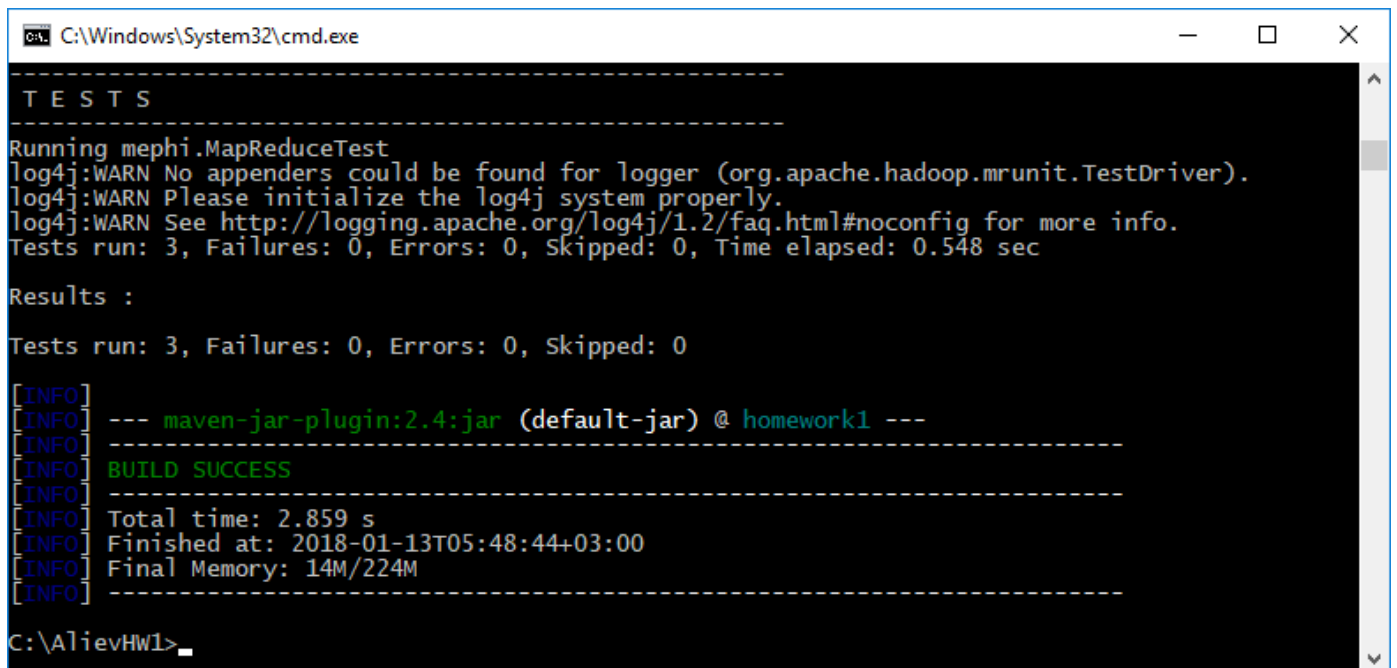
Задача была реализована на ОС Windows 10x64. Использовался hadoop 2.6.2, который был установлен в соответствии с инструкцией:

[http://www.ics.uci.edu/~shantas/Install\\_Hadoop2.6.0\\_on\\_Windows10.pdf](http://www.ics.uci.edu/~shantas/Install_Hadoop2.6.0_on_Windows10.pdf)

Для сборки проекта использовалась IDE agnostic build Maven. pom.xml файл находится в репозитории. Для создания файловой системы HDFS и копирования данных был написан .bat файл hdfshw1.bat (в директории scripts). Скриншоты результатов тестов представлены ниже.

Сборка проекта производится из командной строки командой `<mvn package>`

Файлы с данными для тестов test1.txt и test2.txt



```
-----
T E S T S
-----
Running mephi.MapReduceTest
log4j:WARN No appenders could be found for logger (org.apache.hadoop.mrunit.TestDriver).
log4j:WARN Please initialize the log4j system properly.
log4j:WARN See http://logging.apache.org/log4j/1.2/faq.html#noconfig for more info.
Tests run: 3, Failures: 0, Errors: 0, Skipped: 0, Time elapsed: 0.548 sec

Results :

Tests run: 3, Failures: 0, Errors: 0, Skipped: 0

[INFO]
[INFO] --- maven-jar-plugin:2.4:jar (default-jar) @ homework1 ---
[INFO]
[INFO] BUILD SUCCESS
[INFO]
[INFO] Total time: 2.859 s
[INFO] Finished at: 2018-01-13T05:48:44+03:00
[INFO] Final Memory: 14M/224M
[INFO]
C:\AlievHW1>
```

Рисунок 1 – Выполнение тестов.

```
C:\Windows\System32\cmd.exe

C:\AlievHW1>cd scripts
C:\AlievHW1\scripts>hdfshw1.bat
C:\AlievHW1\scripts>call hadoop fs -mkdir /input
Found 1 items
-rw-r--r--  1 User supergroup      134635 2018-01-13 06:57 /input/inputfile1.txt
C:\AlievHW1\scripts>
```

Рисунок 2 – Запуск скрипта hdfshw1.bat. (Копирование файла в HDFS)

Запуск производится командой:

C:\AlievHW1>hadoop jar target/homework1-0.0.1.jar /input /output

```
C:\Windows\System32\cmd.exe

18/01/13 06:08:27 INFO mapreduce.Job: Job job_local1740646710_0001 completed successfully
18/01/13 06:08:27 INFO mapreduce.Job: Counters: 38
  File System Counters
    FILE: Number of bytes read=524248
    FILE: Number of bytes written=1290424
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=269270
    HDFS: Number of bytes written=24
    HDFS: Number of read operations=15
    HDFS: Number of large read operations=0
    HDFS: Number of write operations=4
  Map-Reduce Framework
    Map input records=3491
    Map output records=22812
    Map output bytes=210890
    Map output materialized bytes=256520
    Input split bytes=108
    Combine input records=0
    Combine output records=0
    Reduce input groups=4798
    Reduce shuffle bytes=256520
    Reduce input records=22812
    Reduce output records=1
    Spilled Records=45624
    Shuffled Maps =1
    Failed Shuffles=0
    Merged Map outputs=1
    GC time elapsed (ms)=4
    CPU time spent (ms)=0
    Physical memory (bytes) snapshot=0
    Virtual memory (bytes) snapshot=0
    Total committed heap usage (bytes)=541065216
  Shuffle Errors
    BAD_ID=0
    CONNECTION=0
    IO_ERROR=0
    WRONG_LENGTH=0
    WRONG_MAP=0
    WRONG_REDUCE=0
  File Input Format Counters
    Bytes Read=134635
  File Output Format Counters
    Bytes Written=24
Job was successful!

C:\AlievHW1>hadoop fs -cat /output/part-r-*
longest_word_in_text,20
C:\AlievHW1>
```

Рисунок 3 – Результат выполнения программы.