

Отчёт

Студент: Турко С.А.

Группа: M16-522

Link: https://github.com/sergeyturko/DSBDA_HW1

Задача.

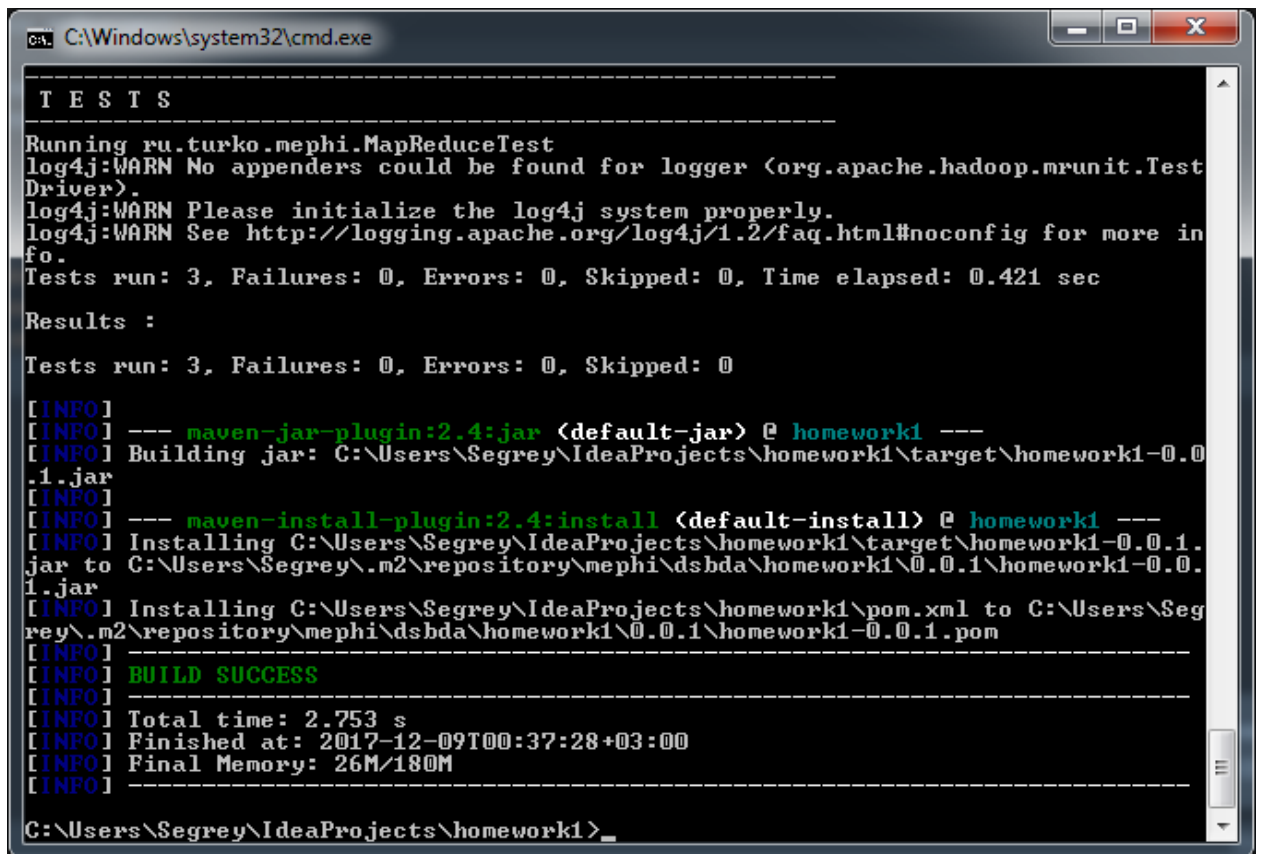
Необходимо найти самое длинное слово в .txt файле. Формат выходного файла – CSV.

Реализация.

Реализация задачи осуществлялась на ОС Windows 7x64. Использовался hadoop 2.6.5, установлен в соответствии с инструкцией [http://www.ics.uci.edu/~shantas/Install_Hadoop-2.6.0 on Windows10.pdf](http://www.ics.uci.edu/~shantas/Install_Hadoop-2.6.0_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 hdfs.bat (в директории scripts). Скриншоты результатов представлены ниже.

Сборка производится из командной строки командой <mvn package>



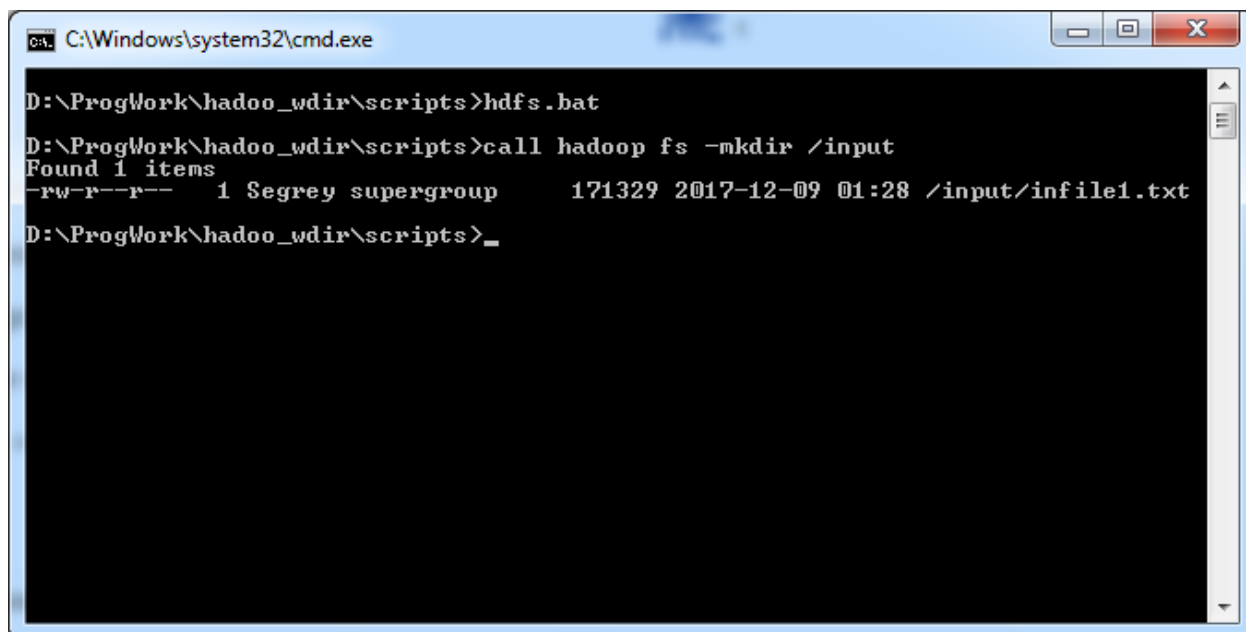
```
-----
T E S T S
-----
Running ru.turko.mephi.MapReduceTest
log4j:WARN No appenders could be found for logger <org.apache.hadoop.mrunit.Test
Driver>.
log4j:WARN Please initialize the log4j system properly.
log4j:WARN See http://logging.apache.org/log4j/1.2/faq.html#noconfig for more in
fo.
Tests run: 3, Failures: 0, Errors: 0, Skipped: 0, Time elapsed: 0.421 sec

Results :

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

[INFO] --- maven-jar-plugin:2.4:jar <default-jar> @ homework1 ---
[INFO] Building jar: C:\Users\Segrey\IdeaProjects\homework1\target\homework1-0.0
.1.jar
[INFO] --- maven-install-plugin:2.4:install <default-install> @ homework1 ---
[INFO] Installing C:\Users\Segrey\IdeaProjects\homework1\target\homework1-0.0.1
.jar to C:\Users\Segrey\.m2\repository\mephi\dsbda\homework1\0.0.1\homework1-0.0
.1.jar
[INFO] Installing C:\Users\Segrey\IdeaProjects\homework1\pom.xml to C:\Users\Seg
rey\.m2\repository\mephi\dsbda\homework1\0.0.1\homework1-0.0.1.pom
[INFO] BUILD SUCCESS
[INFO]
[INFO] Total time: 2.753 s
[INFO] Finished at: 2017-12-09T00:37:28+03:00
[INFO] Final Memory: 26M/180M
[INFO] -----
C:\Users\Segrey\IdeaProjects\homework1>_
```

Рисунок 1 – Успешное выполнение тестов.



A screenshot of a Windows command prompt window. The title bar shows the path `C:\Windows\system32\cmd.exe`. The command prompt is open at the directory `D:\ProgWork\hadoo_wdir\scripts`. The user has entered the command `hdfs.bat`, which has been executed. The output shows the command `call hadoop fs -mkdir /input` and the result `Found 1 items`. Below this, a file listing is shown: `-rw-r--r-- 1 Segrey supergroup 171329 2017-12-09 01:28 /input/infile1.txt`. The prompt is now `D:\ProgWork\hadoo_wdir\scripts>_`.

```
C:\Windows\system32\cmd.exe
D:\ProgWork\hadoo_wdir\scripts>hdfs.bat
D:\ProgWork\hadoo_wdir\scripts>call hadoop fs -mkdir /input
Found 1 items
-rw-r--r-- 1 Segrey supergroup 171329 2017-12-09 01:28 /input/infile1.txt
D:\ProgWork\hadoo_wdir\scripts>_
```

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

Запуск производится командой:

```
hadoop jar homework1-0.0.1.jar homework1 /input /output
```



```
mode : false
17/12/09 01:49:59 INFO mapreduce.Job: map 0% reduce 0%
17/12/09 01:50:06 INFO mapreduce.Job: map 100% reduce 0%
17/12/09 01:50:13 INFO mapreduce.Job: map 100% reduce 100%
17/12/09 01:50:15 INFO mapreduce.Job: Job job_1512763206212_0007 completed successfully
17/12/09 01:50:15 INFO mapreduce.Job: Counters: 49
  File System Counters
    FILE: Number of bytes read=256470
    FILE: Number of bytes written=732091
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=138199
    HDFS: Number of bytes written=22
    HDFS: Number of read operations=6
    HDFS: Number of large read operations=0
    HDFS: Number of write operations=2
  Job Counters
    Launched map tasks=1
    Launched reduce tasks=1
    Rack-local map tasks=1
    Total time spent by all maps in occupied slots (ms)=2879
    Total time spent by all reduces in occupied slots (ms)=3484
    Total time spent by all map tasks (ms)=2879
    Total time spent by all reduce tasks (ms)=3484
    Total vcore-milliseconds taken by all map tasks=2879
    Total vcore-milliseconds taken by all reduce tasks=3484
    Total megabyte-milliseconds taken by all map tasks=2948096
    Total megabyte-milliseconds taken by all reduce tasks=3567616
  Map-Reduce Framework
    Map input records=3490
    Map output records=22809
    Map output bytes=210846
    Map output materialized bytes=256470
    Input split bytes=105
    Combine input records=0
    Combine output records=0
    Reduce input groups=4794
    Reduce shuffle bytes=256470
    Reduce input records=22809
    Reduce output records=1
    Spilled Records=45618
    Shuffled Maps =1
    Failed Shuffles=0
    Merged Map outputs=1
    GC time elapsed (ms)=88
    CPU time spent (ms)=2043
    Physical memory (bytes) snapshot=454688768
    Virtual memory (bytes) snapshot=541540352
    Total committed heap usage (bytes)=300941312
  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=138094
  File Output Format Counters
    Bytes Written=22
Job was successful!
D:\>hadoop fs -text /output/*
freestone-colour'd,18
D:\>
```

Рисунок 3 – Результат выполнения.