

Министерство науки и высшего образования РФ
Федеральное государственное автономное образовательное
учреждение высшего образования «Национальный
исследовательский ядерный университет «МИФИ»

Институт ИИКС
Кафедра компьютерных систем и технологий

Лабораторная работа №1:

«Hadoop»

по дисциплине

«Наука о данных и анализ больших данных»

Выполнил:
студент гр. М21-502
Корнилов А. Н.

Проверил:
Синельников Дмитрий Михайлович

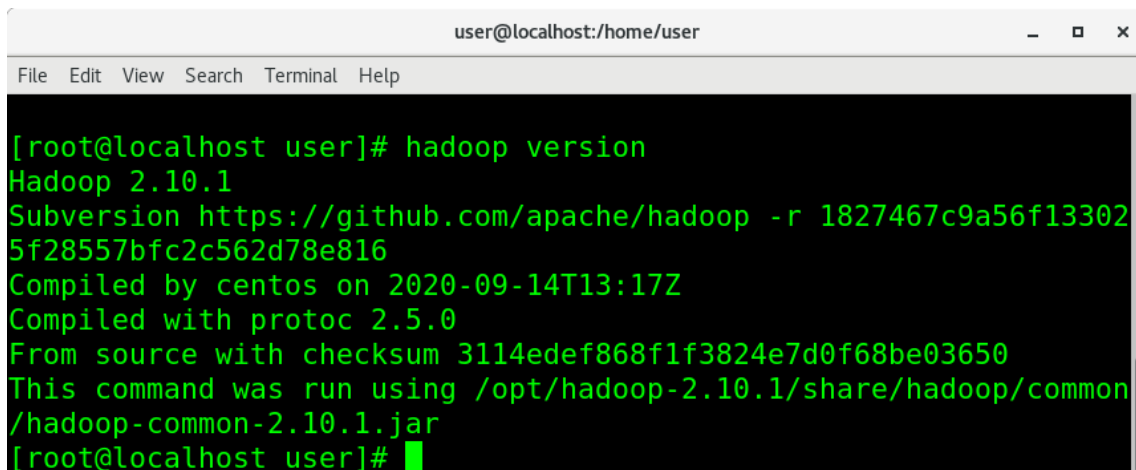
Москва, 2022 г.

Оглавление

Программные средства.....	3
Тесты.....	3
Выполненные действия	4
Результаты.....	6

Программные средства

Разработка велась с использованием Hadoop (см. рисунок 1) на OS Centos. Для сборки использовался maven плагин для IntelliJ IDEA.



```
user@localhost:/home/user
File Edit View Search Terminal Help

[root@localhost user]# hadoop version
Hadoop 2.10.1
Subversion https://github.com/apache/hadoop -r 1827467c9a56f13302
5f28557bfc2c562d78e816
Compiled by centos on 2020-09-14T13:17Z
Compiled with protoc 2.5.0
From source with checksum 3114edef868f1f3824e7d0f68be03650
This command was run using /opt/hadoop-2.10.1/share/hadoop/common
/hadoop-common-2.10.1.jar
[root@localhost user]#
```

Рисунок 1 – скриншот версии Hadoop

Тесты

Результат успешного прохождения тестов представлен на рисунке 2.

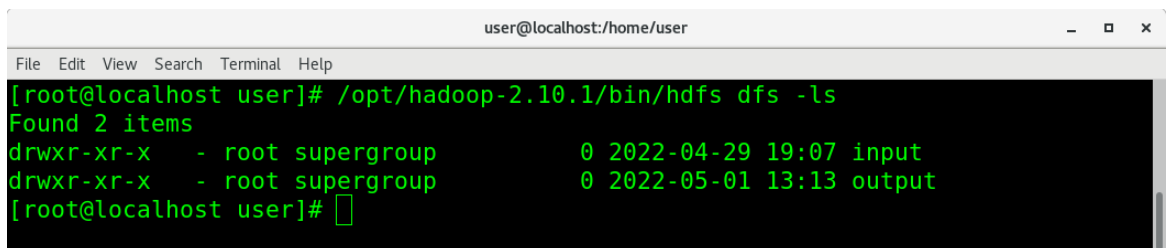


Рисунок 2 – скриншот результата успешного прохождения тестов

Выполненные действия

Последовательность выполнения:

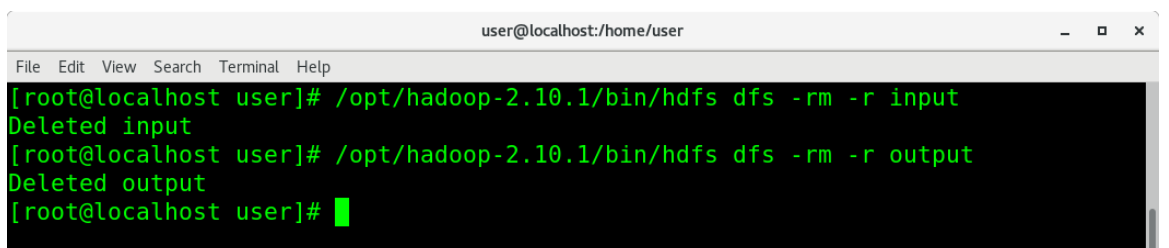
1. Генерация данных с помощью скрипта generateInputData.sh
2. Build-программы
3. Проверка dfs на наличие конечной и начальной директории (обнулить при необходимости). На рисунке 3 представлена команда просмотра содержимого dfs.



```
user@localhost:/home/user
File Edit View Search Terminal Help
[root@localhost user]# /opt/hadoop-2.10.1/bin/hdfs dfs -ls
Found 2 items
drwxr-xr-x  - root supergroup          0 2022-04-29 19:07 input
drwxr-xr-x  - root supergroup          0 2022-05-01 13:13 output
[root@localhost user]#
```

Рисунок 4 – скриншот содержимого dfs

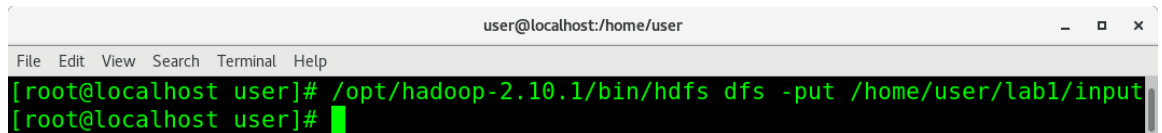
4. Удаление начальных и конечных директорий dfs (см. рисунок 5).



```
user@localhost:/home/user
File Edit View Search Terminal Help
[root@localhost user]# /opt/hadoop-2.10.1/bin/hdfs dfs -rm -r input
Deleted input
[root@localhost user]# /opt/hadoop-2.10.1/bin/hdfs dfs -rm -r output
Deleted output
[root@localhost user]#
```

Рисунок 5 – скриншот команд удаления начальной и конечной директорий

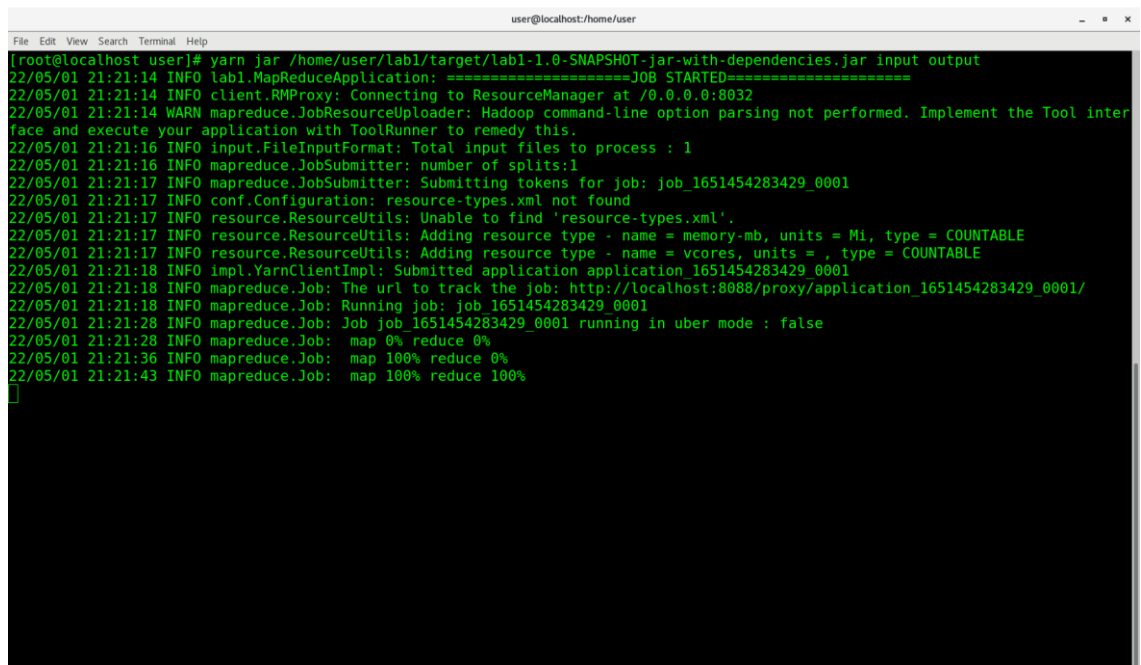
5. Помещение директории со сгенерированным скриптом generateInputData.sh файлом начальных данных в dfs (см. рисунок 6).



```
user@localhost:/home/user
File Edit View Search Terminal Help
[root@localhost user]# /opt/hadoop-2.10.1/bin/hdfs dfs -put /home/user/lab1/input
[root@localhost user]#
```

Рисунок 6 – команды помещения директории со сгенерированным скриптом generateInputData.sh файлом начальных данных в dfs

6. Запуск обработки (см. рисунок 7)

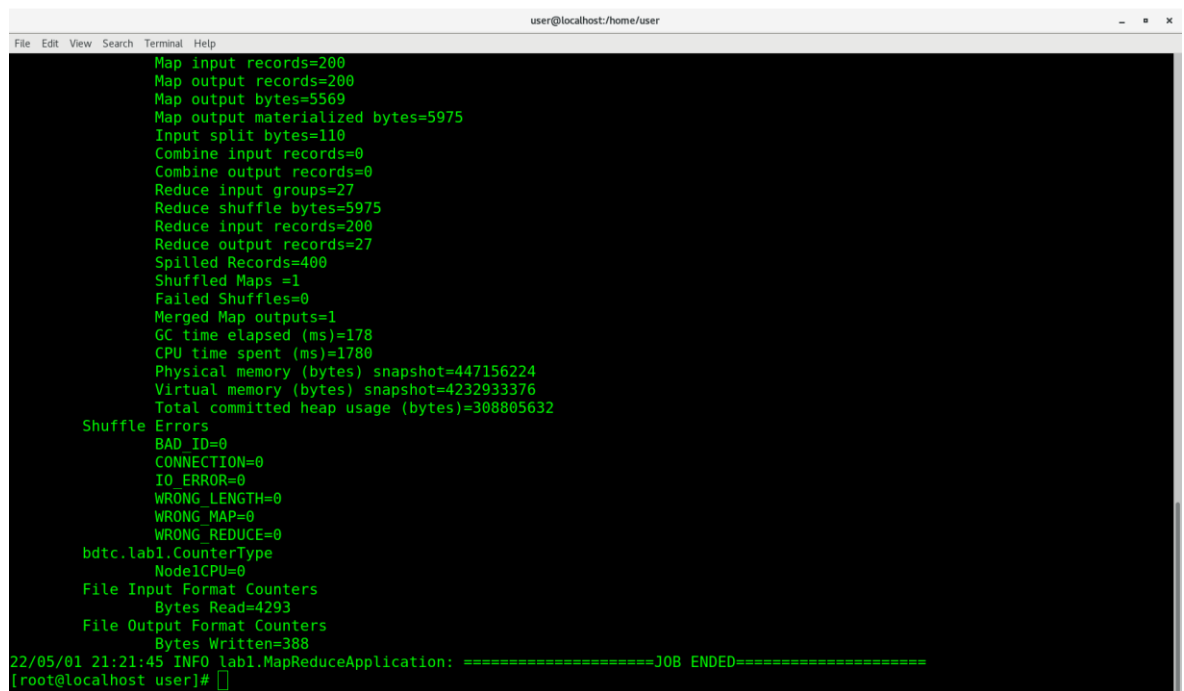


```
user@localhost:/home/user
File Edit View Search Terminal Help
[root@localhost user]# yarn jar /home/user/lab1/target/lab1-1.0-SNAPSHOT-jar-with-dependencies.jar input output
22/05/01 21:21:14 INFO lab1.MapReduceApplication: =====JOB STARTED=====
22/05/01 21:21:14 INFO client.RMPProxy: Connecting to ResourceManager at /0.0.0.0:8032
22/05/01 21:21:14 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
22/05/01 21:21:16 INFO input.FileInputFormat: Total input files to process : 1
22/05/01 21:21:16 INFO mapreduce.JobSubmitter: number of splits:1
22/05/01 21:21:17 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1651454283429_0001
22/05/01 21:21:17 INFO conf.Configuration: resource-types.xml not found
22/05/01 21:21:17 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
22/05/01 21:21:17 INFO resource.ResourceUtils: Adding resource type - name = memory-mb, units = Mi, type = COUNTABLE
22/05/01 21:21:17 INFO resource.ResourceUtils: Adding resource type - name = vcores, units = , type = COUNTABLE
22/05/01 21:21:18 INFO impl.YarnClientImpl: Submitted application application_1651454283429_0001
22/05/01 21:21:18 INFO mapreduce.Job: The url to track the job: http://localhost:8088/proxy/application_1651454283429_0001/
22/05/01 21:21:18 INFO mapreduce.Job: Running job: job_1651454283429_0001
22/05/01 21:21:28 INFO mapreduce.Job: Job job_1651454283429_0001 running in uber mode : false
22/05/01 21:21:28 INFO mapreduce.Job: map 0% reduce 0%
22/05/01 21:21:36 INFO mapreduce.Job: map 100% reduce 0%
22/05/01 21:21:43 INFO mapreduce.Job: map 100% reduce 100%
```

Рисунок 7 – скриншот запуска обработки

Результаты

Результат успешного выполнения MapReduce Job представлен на рисунке 8.



```
user@localhost:/home/user
Map input records=200
Map output records=200
Map output bytes=5569
Map output materialized bytes=5975
Input split bytes=110
Combine input records=0
Combine output records=0
Reduce input groups=27
Reduce shuffle bytes=5975
Reduce input records=200
Reduce output records=27
Spilled Records=400
Shuffled Maps =1
Failed Shuffles=0
Merged Map outputs=1
GC time elapsed (ms)=178
CPU time spent (ms)=1780
Physical memory (bytes) snapshot=447156224
Virtual memory (bytes) snapshot=423293376
Total committed heap usage (bytes)=308805632

Shuffle Errors
BAD_ID=0
CONNECTION=0
IO_ERROR=0
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0

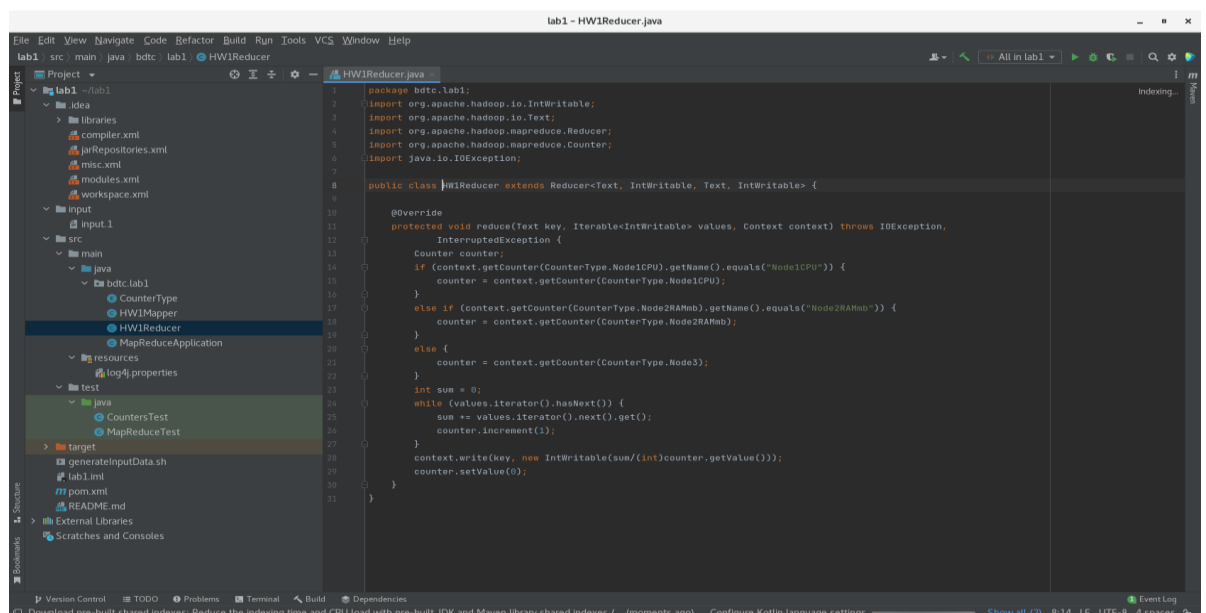
bdtc.lab1.CounterType
Node1CPU=0

File Input Format Counters
Bytes Read=4293
File Output Format Counters
Bytes Written=388

22/05/01 21:21:45 INFO lab1.MapReduceApplication: =====JOB ENDED=====
[root@localhost user]#
```

Рисунок 8 – скриншот успешного выполнения MapReduce Job

На рисунке 9 представлен скриншот использования счётчиков.



```
lab1 - HW1Reducer.java
File Edit View Navigate Code Refactor Build Run Tools VCS Window Help
lab1 src main java bdtc lab1 HW1Reducer
Project
lab1
  libraries
  compiler.xml
  jarRepositories.xml
  misc.xml
  modules.xml
  workspace.xml
  input
  input.1
  src
  main
  java
  bdtc.lab1
    CounterType
    HW1Mapper
    HW1Reducer
    MapReduceApplication
  resources
  log4j.properties
  test
  java
    CountersTest
    MapReduceTest
  target
  generateInputData.sh
  lab1.iml
  pom.xml
  README.md
  External Libraries
  Scratches and Consoles

package bdtc.lab1;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Reducer;
import org.apache.hadoop.mapreduce.Counter;
import java.io.IOException;

public class HW1Reducer extends Reducer<Text, IntWritable, Text, IntWritable> {

    @Override
    protected void reduce(Text key, Iterable<IntWritable> values, Context context) throws IOException,
        InterruptedException {
        Counter counter;
        if (context.getCounter(CounterType.Node1CPU).getName().equals("Node1CPU")) {
            counter = context.getCounter(CounterType.Node1CPU);
        }
        else if (context.getCounter(CounterType.Node2RAMb).getName().equals("Node2RAMb")) {
            counter = context.getCounter(CounterType.Node2RAMb);
        }
        else {
            counter = context.getCounter(CounterType.Node3);
        }

        int sum = 0;
        while (values.iterator().hasNext()) {
            sum += values.iterator().next().get();
            counter.increment(1);
        }

        context.write(key, new IntWritable(sum/(int)counter.getValue()));
        counter.setValue(0);
    }
}
```

Рисунок 9 – скриншот использования счётчиков

На рисунке 10 представлено содержимое конечной директории dfs.

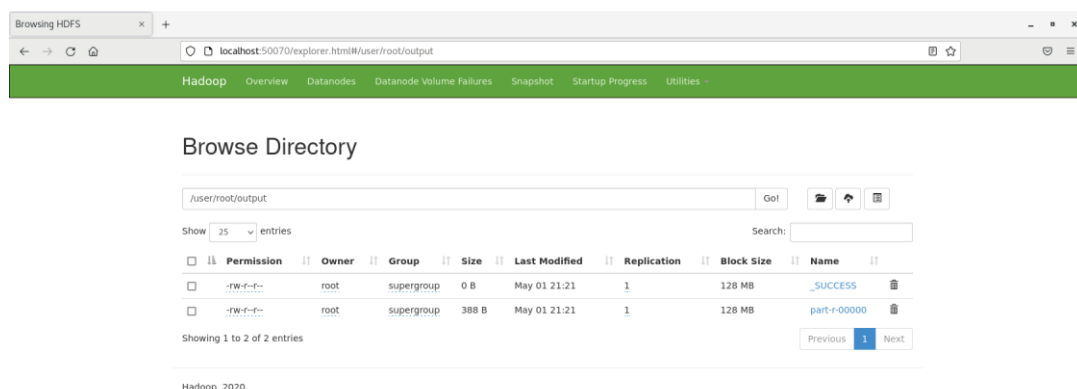


Рисунок 10 – скриншот содержимого конечной директории dfs

На рисунке 11 представлен скриншот содержимого файла с результатами.

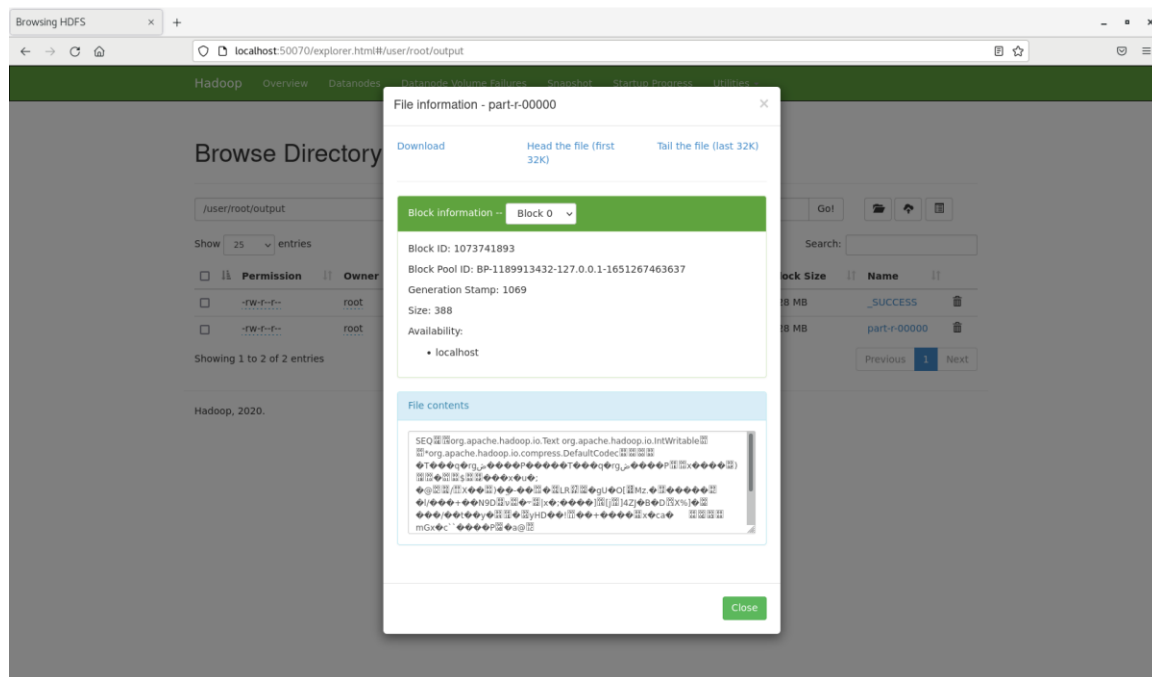
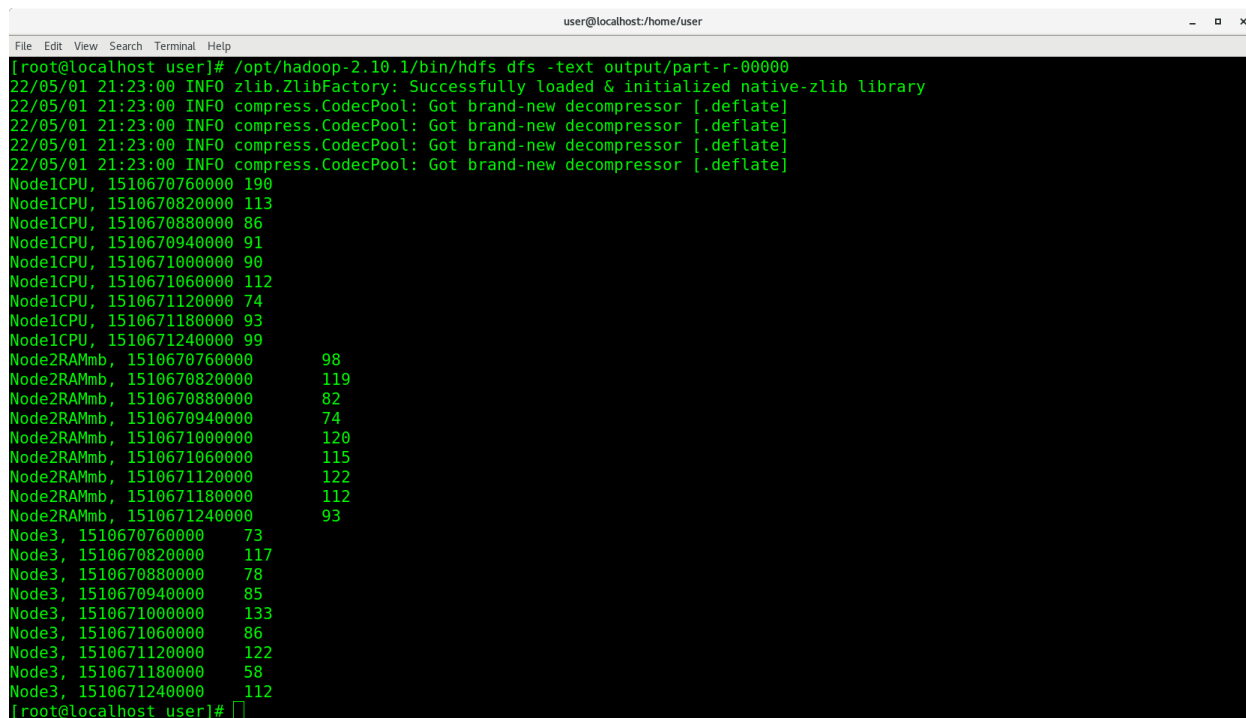


Рисунок 11 – скриншот содержимого файла с результатами

На рисунке 12 представлено содержимое распакованного файла с результатами.



```
user@localhost/home/user
File Edit View Search Terminal Help
[root@localhost user]# /opt/hadoop-2.10.1/bin/hdfs dfs -text output/part-r-00000
22/05/01 21:23:00 INFO zlib.ZlibFactory: Successfully loaded & initialized native-zlib library
22/05/01 21:23:00 INFO compress.CodecPool: Got brand-new decompressor [.deflate]
22/05/01 21:23:00 INFO compress.CodecPool: Got brand-new decompressor [.deflate]
22/05/01 21:23:00 INFO compress.CodecPool: Got brand-new decompressor [.deflate]
22/05/01 21:23:00 INFO compress.CodecPool: Got brand-new decompressor [.deflate]
Node1CPU, 1510670760000 190
Node1CPU, 1510670820000 113
Node1CPU, 1510670880000 86
Node1CPU, 1510670940000 91
Node1CPU, 1510671000000 90
Node1CPU, 1510671060000 112
Node1CPU, 1510671120000 74
Node1CPU, 1510671180000 93
Node1CPU, 1510671240000 99
Node2RAMmb, 1510670760000 98
Node2RAMmb, 1510670820000 119
Node2RAMmb, 1510670880000 82
Node2RAMmb, 1510670940000 74
Node2RAMmb, 1510671000000 120
Node2RAMmb, 1510671060000 115
Node2RAMmb, 1510671120000 122
Node2RAMmb, 1510671180000 112
Node2RAMmb, 1510671240000 93
Node3, 1510670760000 73
Node3, 1510670820000 117
Node3, 1510670880000 78
Node3, 1510670940000 85
Node3, 1510671000000 133
Node3, 1510671060000 86
Node3, 1510671120000 122
Node3, 1510671180000 58
Node3, 1510671240000 112
[root@localhost user]#
```

Рисунок 12 – скриншот содержимого распакованного файла с результатами