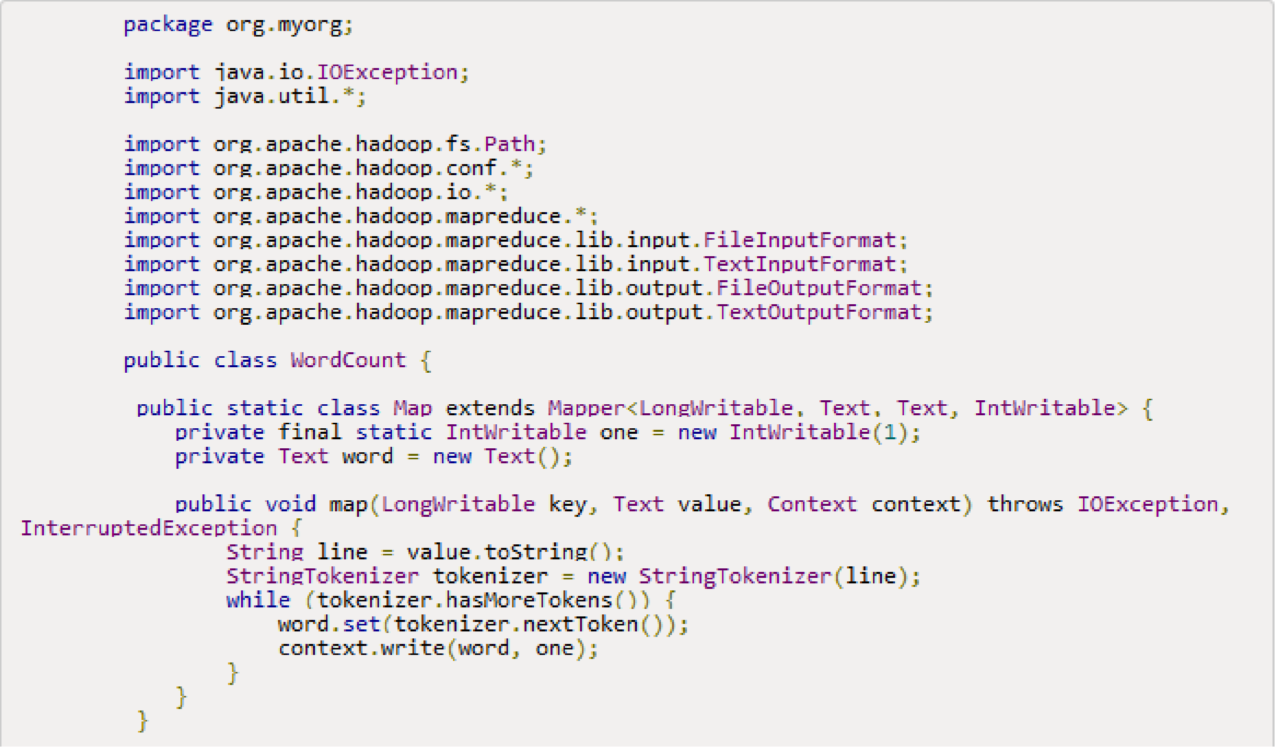
Curso: Ciência de dados e Big Data

Professor: Cláudio Lúcio
Atividade Prática sobre Map Reduce

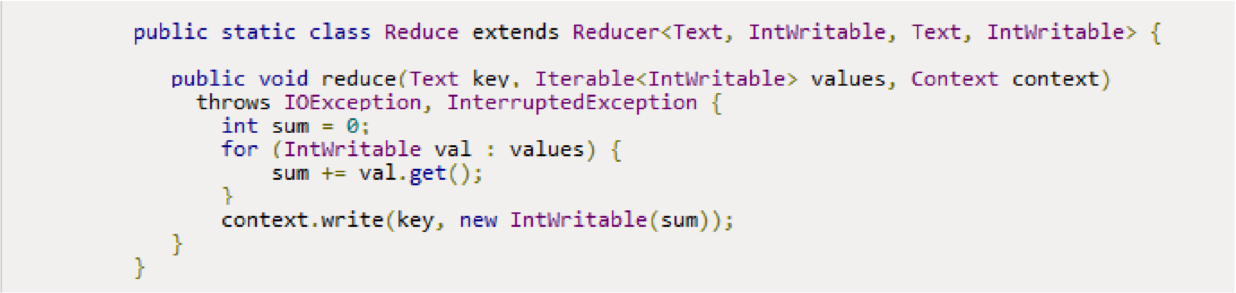
Neste caso vamos analisar uma implementação de map reduce no Hadoop.
Vamos utilizar os um exemplo de contagem de palavras já existente na sandbox da HortonWorks. Veja abaixo o programa que será executado (obviamente, implemetanto em Java):

1. Porção relativa ao Map



```
        package org.myorg;

        import java.io.IOException;
        import java.util.*;

        import org.apache.hadoop.fs.Path;
        import org.apache.hadoop.conf.*;
        import org.apache.hadoop.io.*;
        import org.apache.hadoop.mapreduce.*;
        import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
        import org.apache.hadoop.mapreduce.lib.input.TextInputFormat;
        import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
        import org.apache.hadoop.mapreduce.lib.output.TextOutputFormat;

        public class WordCount {

         public static class Map extends Mapper<LongWritable, Text, Text, IntWritable> {
            private final static IntWritable one = new IntWritable(1);
            private Text word = new Text();

            public void map(LongWritable key, Text value, Context context) throws IOException,
InterruptedException {
                String line = value.toString();
                StringTokenizer tokenizer = new StringTokenizer(line);
                while (tokenizer.hasMoreTokens()) {
                    word.set(tokenizer.nextToken());
                    context.write(word, one);
                }
            }
         }
```

2. Porção relativa ao Reduce

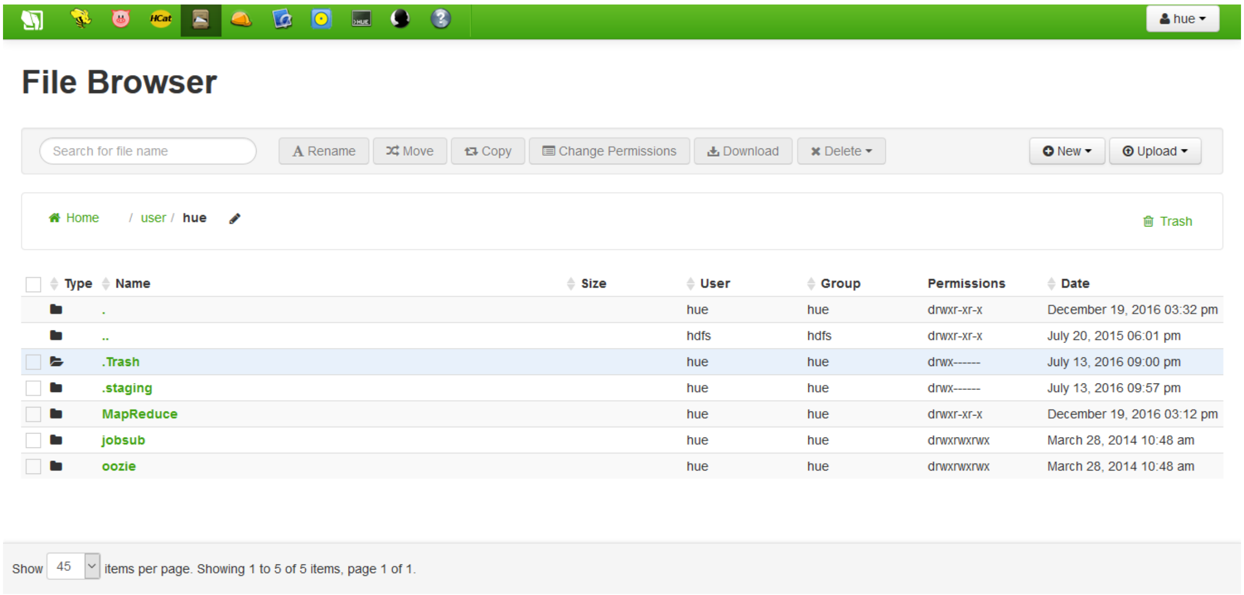


```
 public static class Reduce extends Reducer<Text, IntWritable, Text, IntWritable> {

    public void reduce(Text key, Iterable<IntWritable> values, Context context)
      throws IOException, InterruptedException {
        int sum = 0;
        for (IntWritable val : values) {
            sum += val.get();
        }
        context.write(key, new IntWritable(sum));
    }
 }
```

3. Porção relativa ao programa principal

```
        public static void main(String[] args) throws Exception {
           Configuration conf = new Configuration();

               Job job = new Job(conf, "wordcount");

           job.setOutputKeyClass(Text.class);
           job.setOutputValueClass(IntWritable.class);

           job.setMapperClass(Map.class);
           job.setReducerClass(Reduce.class);

           job.setInputFormatClass(TextInputFormat.class);
           job.setOutputFormatClass(TextOutputFormat.class);

           FileInputFormat.addInputPath(job, new Path(args[0]));
           FileOutputFormat.setOutputPath(job, new Path(args[1]));

           job.waitForCompletion(true);
        }

       }
```

4. Para executar este programa vamos antes criar dois diretórios no usuário Hue, veja a tela a seguir:
   1. MapReduce



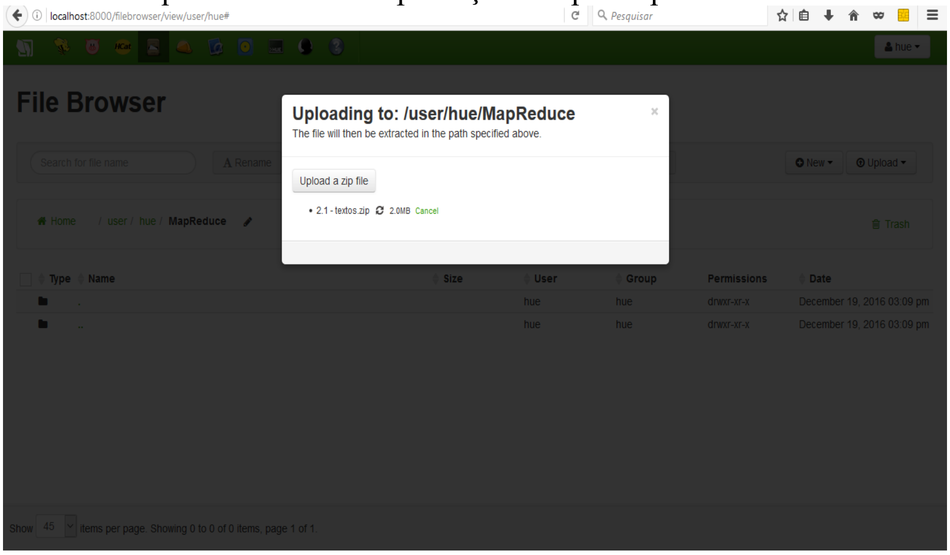
File Browser

Home / user / hue

Type	Name	Size	User	Group	Permissions	Date
	.		hue	hue	drwxr-xr-x	December 19, 2016 03:32 pm
	..		hdfs	hdfs	drwxr-xr-x	July 20, 2015 06:01 pm
	.Trash		hue	hue	drwx------	July 13, 2016 09:00 pm
	.staging		hue	hue	drwx------	July 13, 2016 09:57 pm
	MapReduce		hue	hue	drwxr-xr-x	December 19, 2016 03:12 pm
	jobsub		hue	hue	drwxrwxrwx	March 28, 2014 10:48 am
	oozie		hue	hue	drwxrwxrwx	March 28, 2014 10:48 am

Show 45 items per page. Showing 1 to 5 of 5 items, page 1 of 1.

5. Agora utilize o arquivo 2.1 Textos.zip e faça seu upload para o HDFS.



File Browser

Uploading to: /user/hue/MapReduce

The file will then be extracted in the path specified above.

Upload a zip file

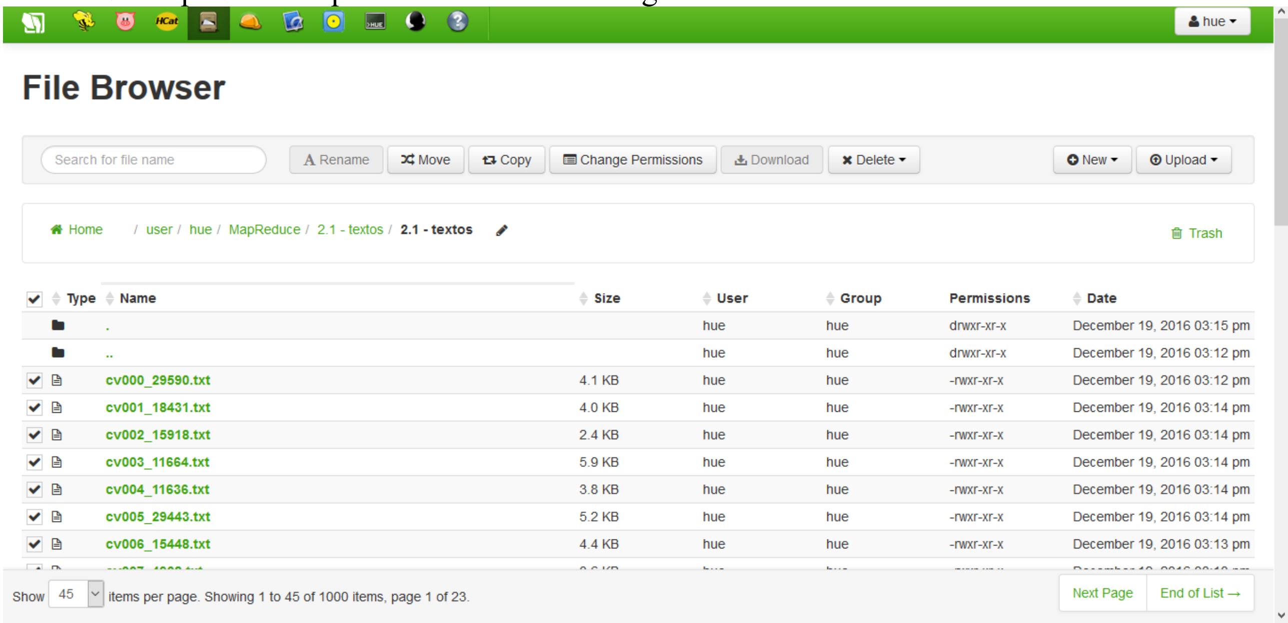
- 2.1 - textos.zip 2.0MB Cancel

Home / user / hue / MapReduce

Type	Name	Size	User	Group	Permissions	Date
	.		hue	hue	drwxr-xr-x	December 19, 2016 03:09 pm
	..		hue	hue	drwxr-xr-x	December 19, 2016 03:09 pm

Show 45 items per page. Showing 0 to 0 of 0 items, page 1 of 1

6. Verifique se os arquivos foram todos carregados



File Browser

/ user / hue / MapReduce / 2.1 - textos / 2.1 - textos

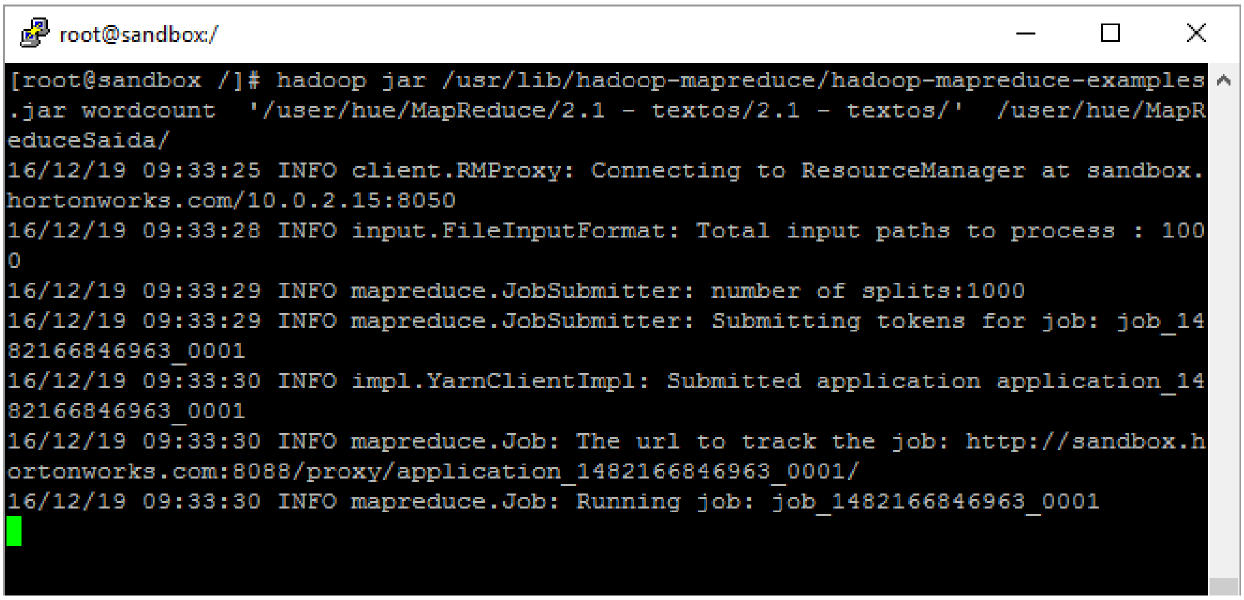
Type	Name	Size	User	Group	Permissions	Date
	.		hue	hue	drwxr-xr-x	December 19, 2016 03:15 pm
	..		hue	hue	drwxr-xr-x	December 19, 2016 03:12 pm
☑	cv000_29590.txt	4.1 KB	hue	hue	-rwxr-xr-x	December 19, 2016 03:12 pm
☑	cv001_18431.txt	4.0 KB	hue	hue	-rwxr-xr-x	December 19, 2016 03:14 pm
☑	cv002_15918.txt	2.4 KB	hue	hue	-rwxr-xr-x	December 19, 2016 03:14 pm
☑	cv003_11664.txt	5.9 KB	hue	hue	-rwxr-xr-x	December 19, 2016 03:14 pm
☑	cv004_11636.txt	3.8 KB	hue	hue	-rwxr-xr-x	December 19, 2016 03:14 pm
☑	cv005_29443.txt	5.2 KB	hue	hue	-rwxr-xr-x	December 19, 2016 03:14 pm
☑	cv006_15448.txt	4.4 KB	hue	hue	-rwxr-xr-x	December 19, 2016 03:13 pm

Show 45 items per page. Showing 1 to 45 of 1000 items, page 1 of 23.

Next Page End of List →

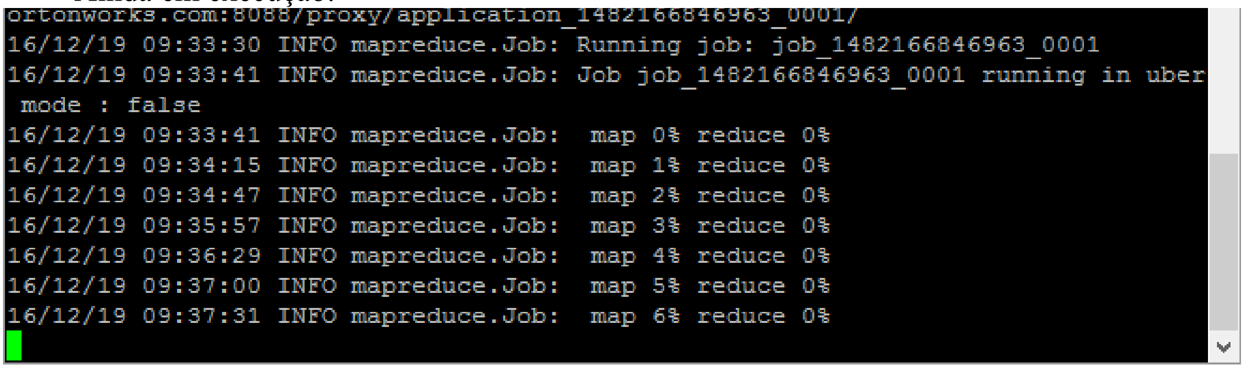
7. Com os arquivos vamos abrir uma linha de comando:

```
su hdfs
hadoop  jar  /usr/hdp/2.3.2.0-2950/hadoop-mapreduce/hadoop-mapreduce-examples.jar  wordcount
'/user/hue/MapReduce/2.1 - textos/2.1 - textos/' /user/hue/MapReduceSaida/
```



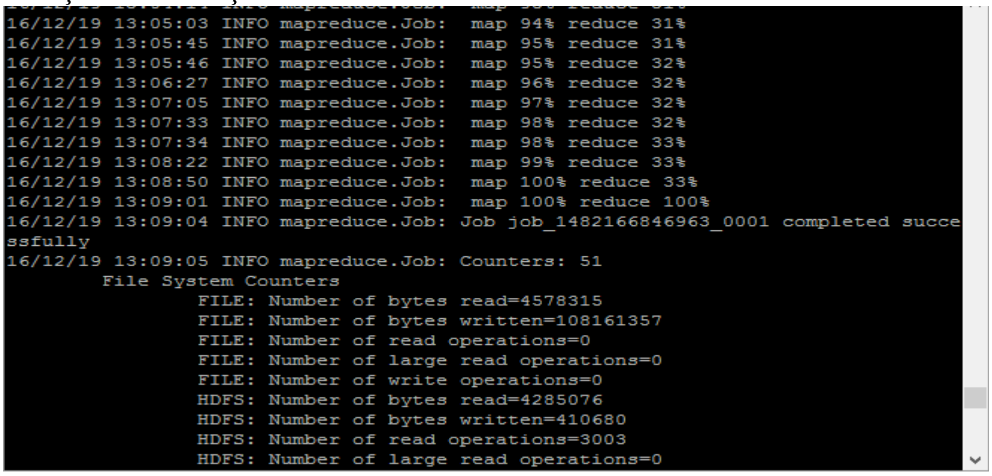
```
[root@sandbox /]# hadoop jar /usr/lib/hadoop-mapreduce/hadoop-mapreduce-examples
.jar wordcount  '/user/hue/MapReduce/2.1 - textos/2.1 - textos/'  /user/hue/MapR
educeSaida/
16/12/19 09:33:25 INFO client.RMProxy: Connecting to ResourceManager at sandbox.
hortonworks.com/10.0.2.15:8050
16/12/19 09:33:28 INFO input.FileInputFormat: Total input paths to process : 100
0
16/12/19 09:33:29 INFO mapreduce.JobSubmitter: number of splits:1000
16/12/19 09:33:29 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_14
82166846963_0001
16/12/19 09:33:30 INFO impl.YarnClientImpl: Submitted application application_14
82166846963_0001
16/12/19 09:33:30 INFO mapreduce.Job: The url to track the job: http://sandbox.h
ortonworks.com:8088/proxy/application_1482166846963_0001/
16/12/19 09:33:30 INFO mapreduce.Job: Running job: job_1482166846963_0001
```

Ainda em execução:

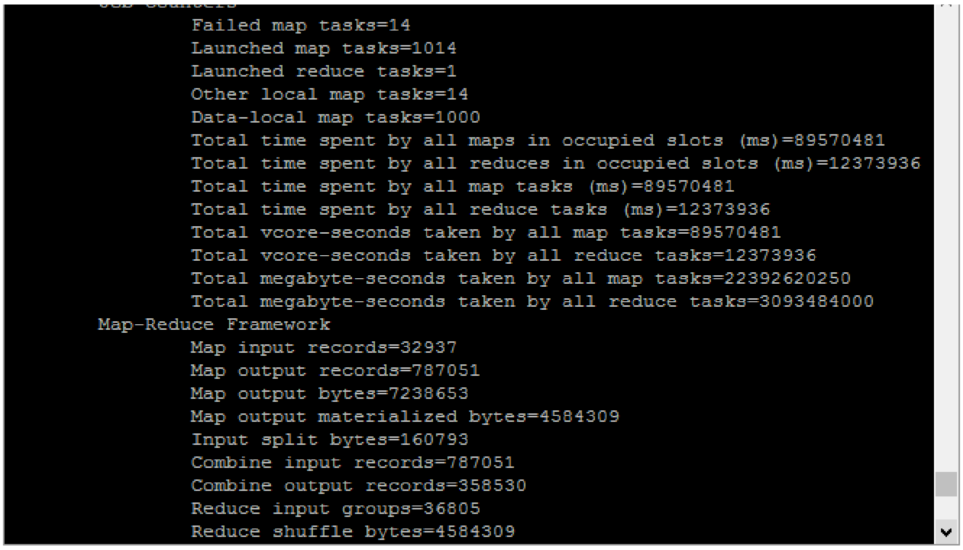


```
ortonworks.com:8088/proxy/application_1482166846963_0001/
16/12/19 09:33:30 INFO mapreduce.Job: Running job: job_1482166846963_0001
16/12/19 09:33:41 INFO mapreduce.Job: Job job_1482166846963_0001 running in uber
 mode : false
16/12/19 09:33:41 INFO mapreduce.Job:  map 0% reduce 0%
16/12/19 09:34:15 INFO mapreduce.Job:  map 1% reduce 0%
16/12/19 09:34:47 INFO mapreduce.Job:  map 2% reduce 0%
16/12/19 09:35:57 INFO mapreduce.Job:  map 3% reduce 0%
16/12/19 09:36:29 INFO mapreduce.Job:  map 4% reduce 0%
16/12/19 09:37:00 INFO mapreduce.Job:  map 5% reduce 0%
16/12/19 09:37:31 INFO mapreduce.Job:  map 6% reduce 0%
```

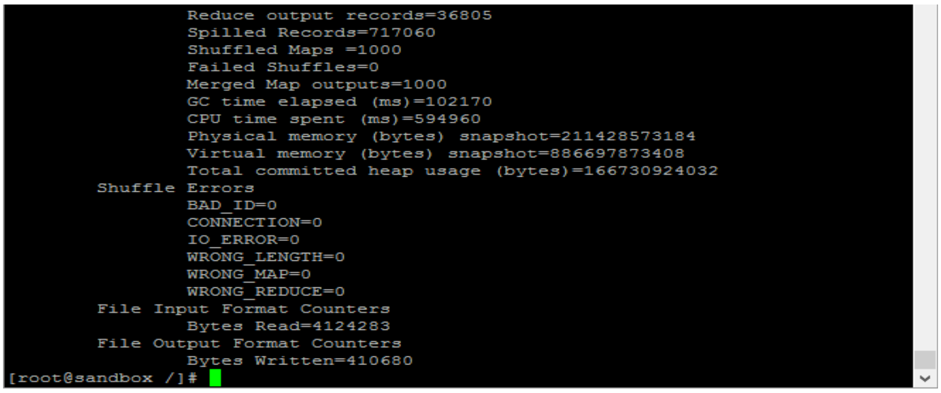
Finlalização da execução:



```
16/12/19 13:05:03 INFO mapreduce.Job:  map 94% reduce 31%
16/12/19 13:05:45 INFO mapreduce.Job:  map 95% reduce 31%
16/12/19 13:05:46 INFO mapreduce.Job:  map 95% reduce 32%
16/12/19 13:06:27 INFO mapreduce.Job:  map 96% reduce 32%
16/12/19 13:07:05 INFO mapreduce.Job:  map 97% reduce 32%
16/12/19 13:07:33 INFO mapreduce.Job:  map 98% reduce 32%
16/12/19 13:07:34 INFO mapreduce.Job:  map 98% reduce 33%
16/12/19 13:08:22 INFO mapreduce.Job:  map 99% reduce 33%
16/12/19 13:08:50 INFO mapreduce.Job:  map 100% reduce 33%
16/12/19 13:09:01 INFO mapreduce.Job:  map 100% reduce 100%
16/12/19 13:09:04 INFO mapreduce.Job: Job job_1482166846963_0001 completed succe
ssfully
16/12/19 13:09:05 INFO mapreduce.Job: Counters: 51
        File System Counters
                FILE: Number of bytes read=4578315
                FILE: Number of bytes written=108161357
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=4285076
                HDFS: Number of bytes written=410680
                HDFS: Number of read operations=3003
                HDFS: Number of large read operations=0
```

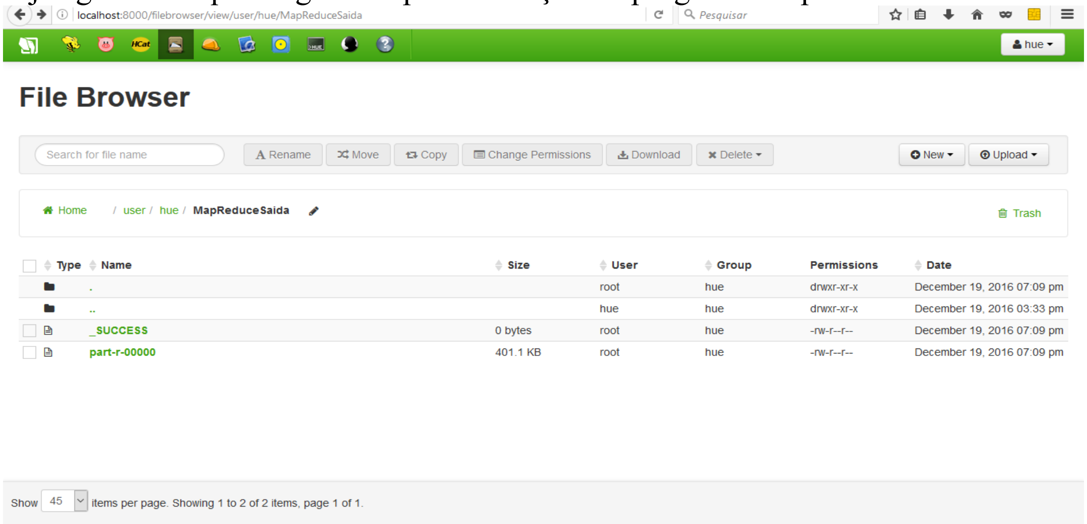


```
                Failed map tasks=14
                Launched map tasks=1014
                Launched reduce tasks=1
                Other local map tasks=14
                Data-local map tasks=1000
                Total time spent by all maps in occupied slots (ms)=89570481
                Total time spent by all reduces in occupied slots (ms)=12373936
                Total time spent by all map tasks (ms)=89570481
                Total time spent by all reduce tasks (ms)=12373936
                Total vcore-seconds taken by all map tasks=89570481
                Total vcore-seconds taken by all reduce tasks=12373936
                Total megabyte-seconds taken by all map tasks=22392620250
                Total megabyte-seconds taken by all reduce tasks=3093484000
        Map-Reduce Framework
                Map input records=32937
                Map output records=787051
                Map output bytes=7238653
                Map output materialized bytes=4584309
                Input split bytes=160793
                Combine input records=787051
                Combine output records=358530
                Reduce input groups=36805
                Reduce shuffle bytes=4584309
```



```
                Reduce output records=36805
                Spilled Records=717060
                Shuffled Maps =1000
                Failed Shuffles=0
                Merged Map outputs=1000
                GC time elapsed (ms)=102170
                CPU time spent (ms)=594960
                Physical memory (bytes) snapshot=211428573184
                Virtual memory (bytes) snapshot=886697873408
                Total committed heap usage (bytes)=166730924032
        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=4124283
        File Output Format Counters
                Bytes Written=410680
[root@sandbox /]#
```

8. Veja agora os arquivos gerados pela execução do programa map reduce:



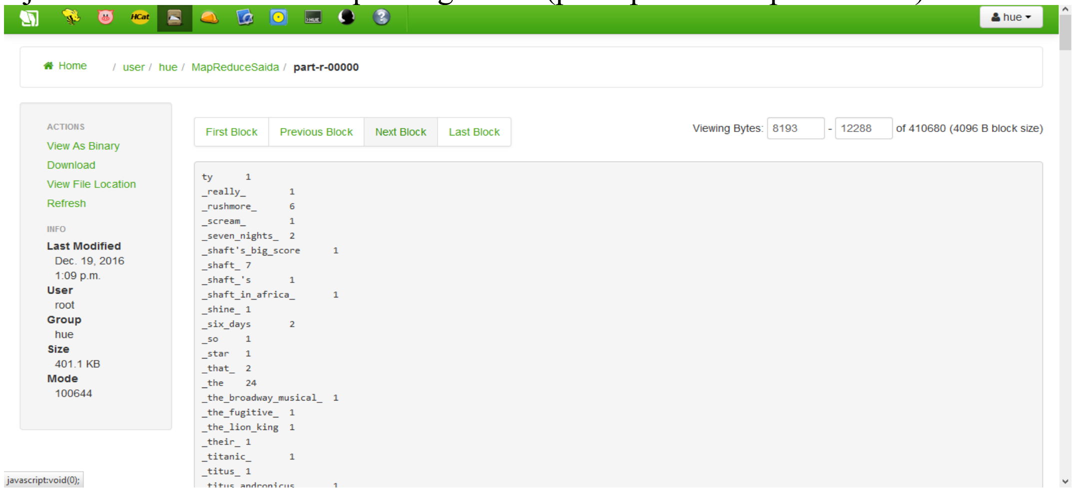
File Browser

/ user / hue / MapReduceSaida

Type	Name	Size	User	Group	Permissions	Date
	.		root	hue	drwxr-xr-x	December 19, 2016 07:09 pm
	..		hue	hue	drwxr-xr-x	December 19, 2016 03:33 pm
	_SUCCESS	0 bytes	root	hue	-rw-r--r--	December 19, 2016 07:09 pm
	part-r-00000	401.1 KB	root	hue	-rw-r--r--	December 19, 2016 07:09 pm

Show 45 items per page. Showing 1 to 2 of 2 items, page 1 of 1.

9. Veja o conteúdo dos dois arquivos gerados (principalemente part-r-0000):



/ user / hue / MapReduceSaida / part-r-00000

Last Modified
Dec. 19, 2016
1:09 p.m.
User
root
Group
hue
Size
401.1 KB
Mode
100644

Viewing Bytes: 8193 - 12288 of 410680 (4096 B block size)

```
ty	1
_really_	1
_rushmore_	6
_scream_	1
_seven_nights_	2
_shaft's_big_score	1
_shaft_	7
_shaft_'s	1
_shaft_in_africa_	1
_shine_	1
_six_days	2
_so	1
_star	1
_that_	2
_the	24
_the_broadway_musical_	1
_the_fugitive_	1
_the_lion_king	1
_their_	1
_titanic_	1
_titus_	1
_titus_andronicus	1
```