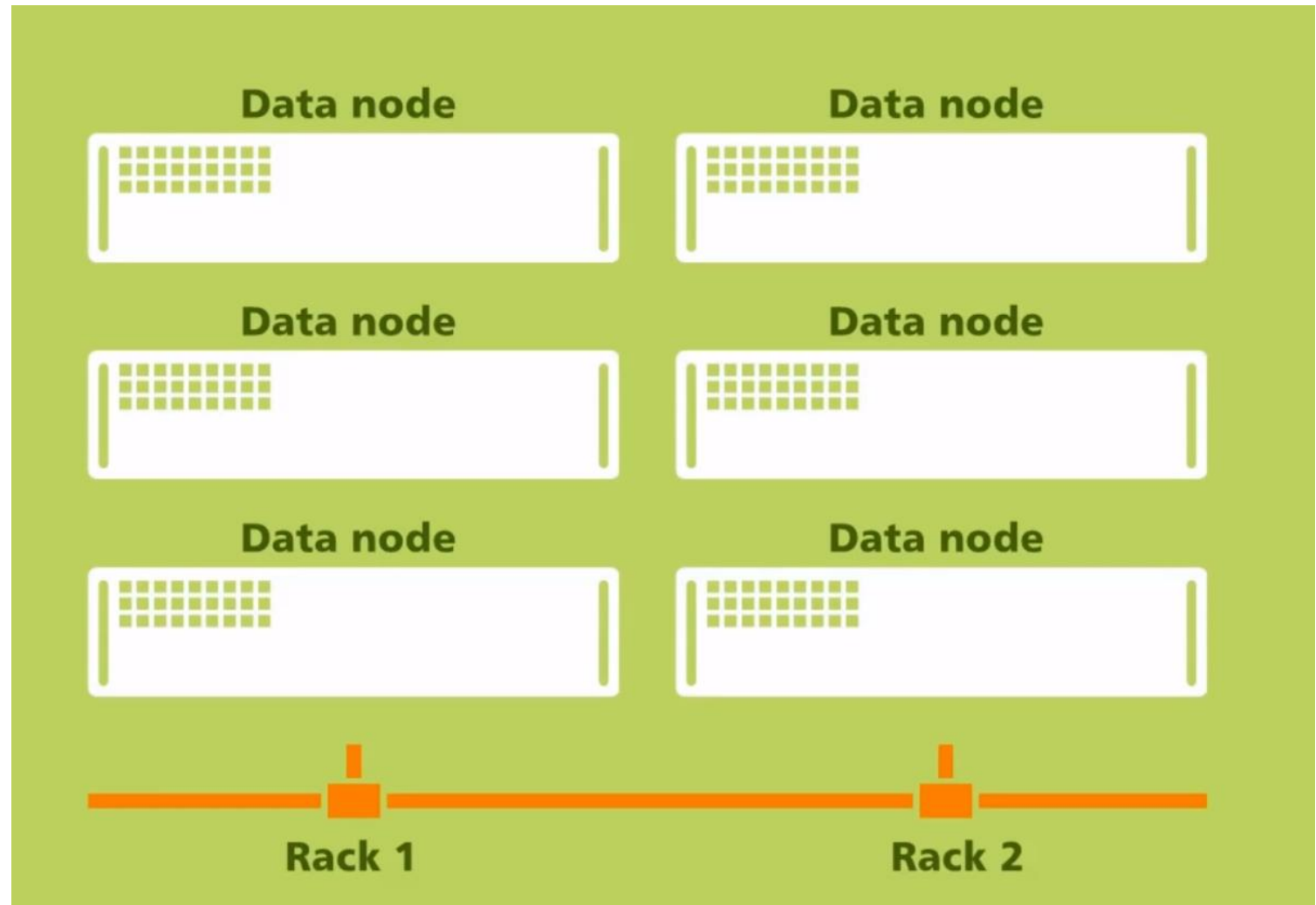


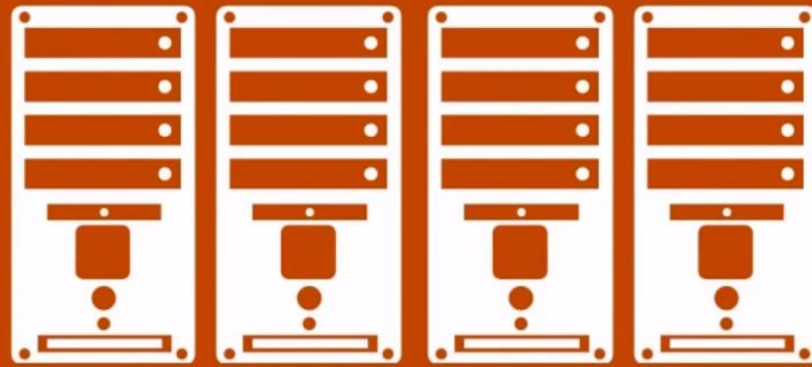
Apache Hadoop

Jorge Siqueira

William Thiago

O que é
Hadoop?





**Escalabilidade Horizontal
(scale out)**

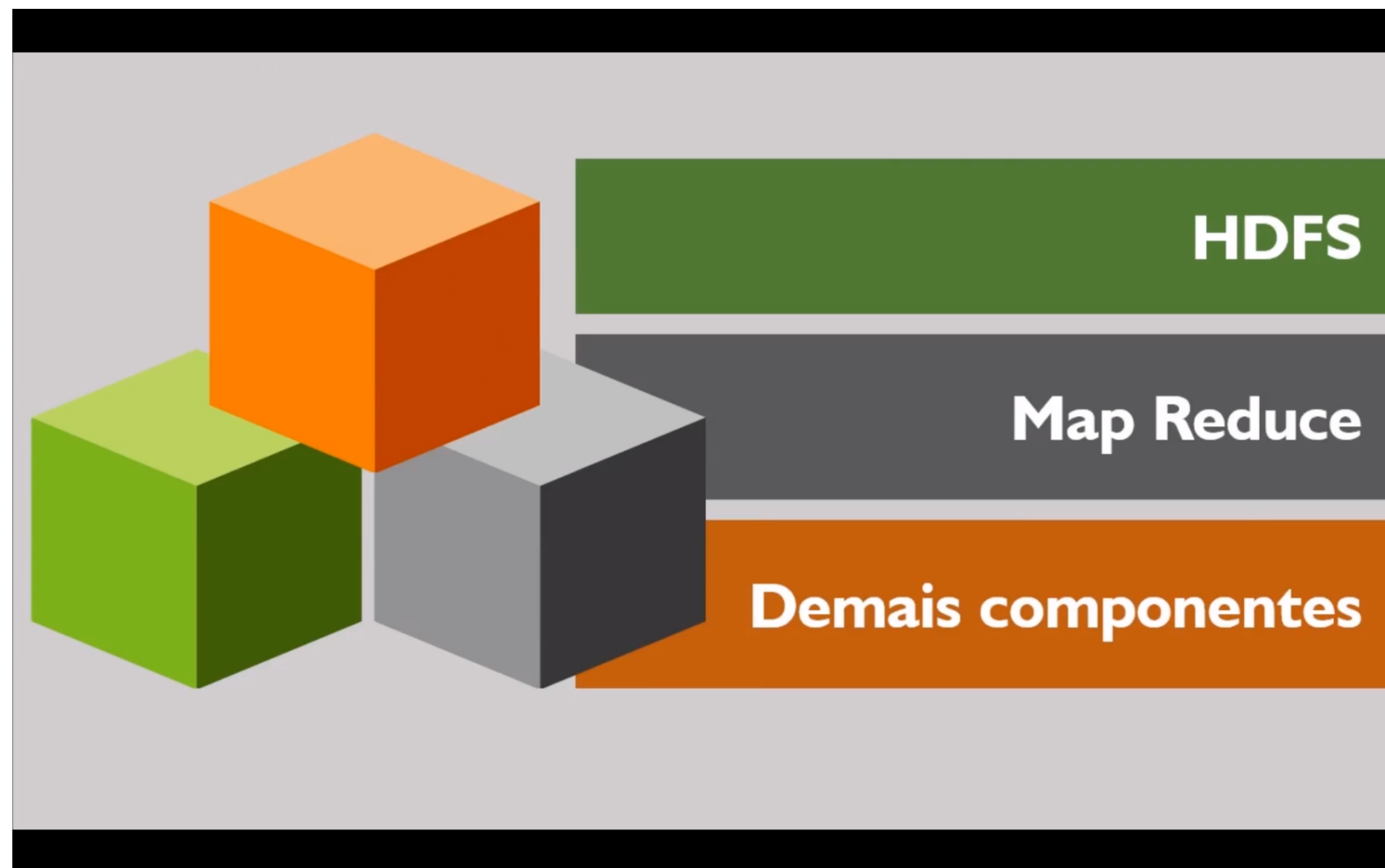
Ecossistema Hadoop

**Família de
projetos
relacionados**

**Infra
estrutura de
computação
distribuída**

**Processamento
de dados de
larga escala**

O
Ecossistema
se baseia
em:

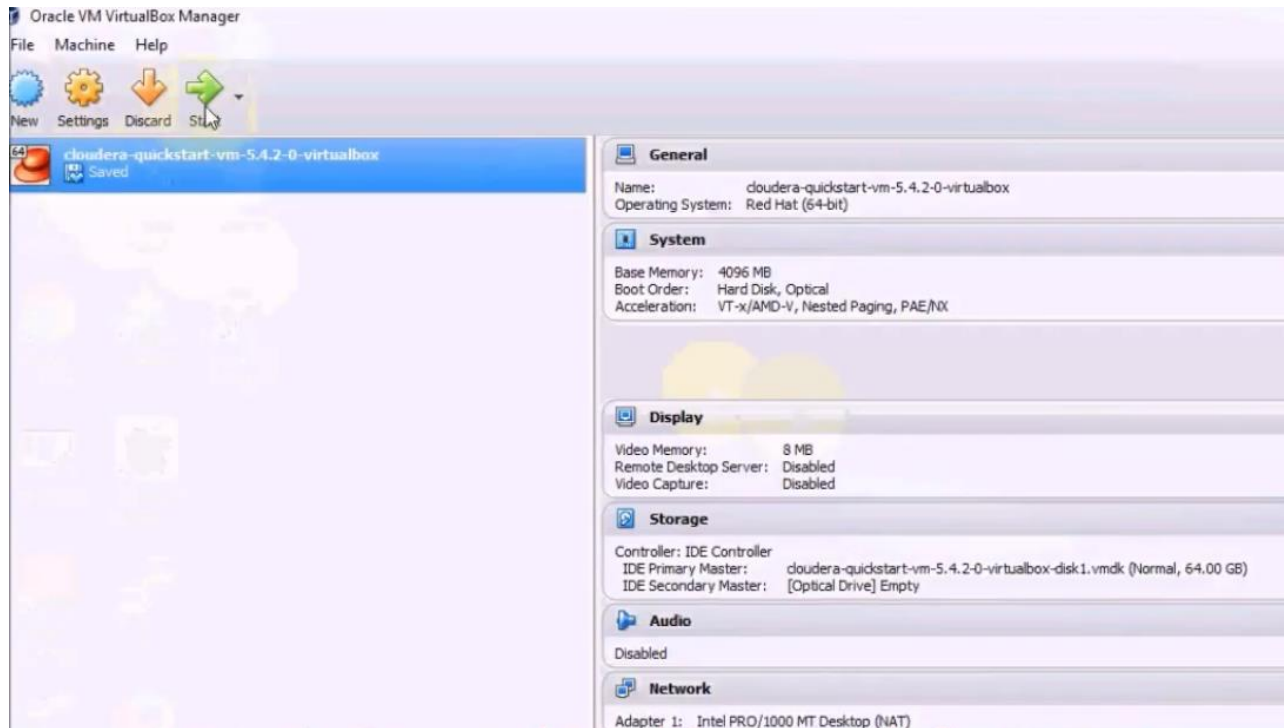


Fazer uma
consulta de
dados em
clusters



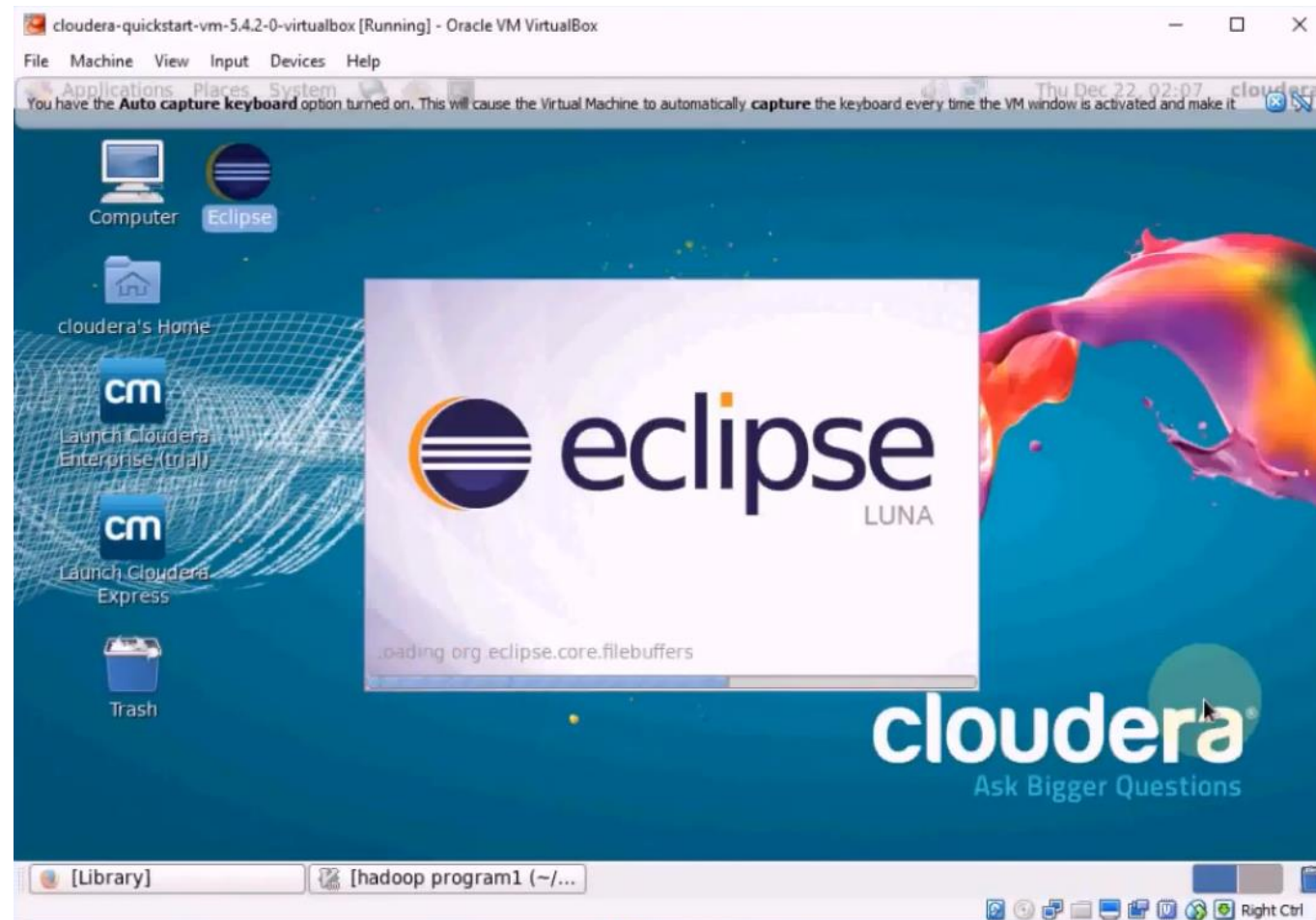
Rodando programas Hadoop

Primeiro baixamos a Oracle virtual machine (virtual box) para rodar o nosso framework hadoop, quem já vem instalado no S.O linux da empresa Cloudera, disponível em: <https://www.cloudera.com/>

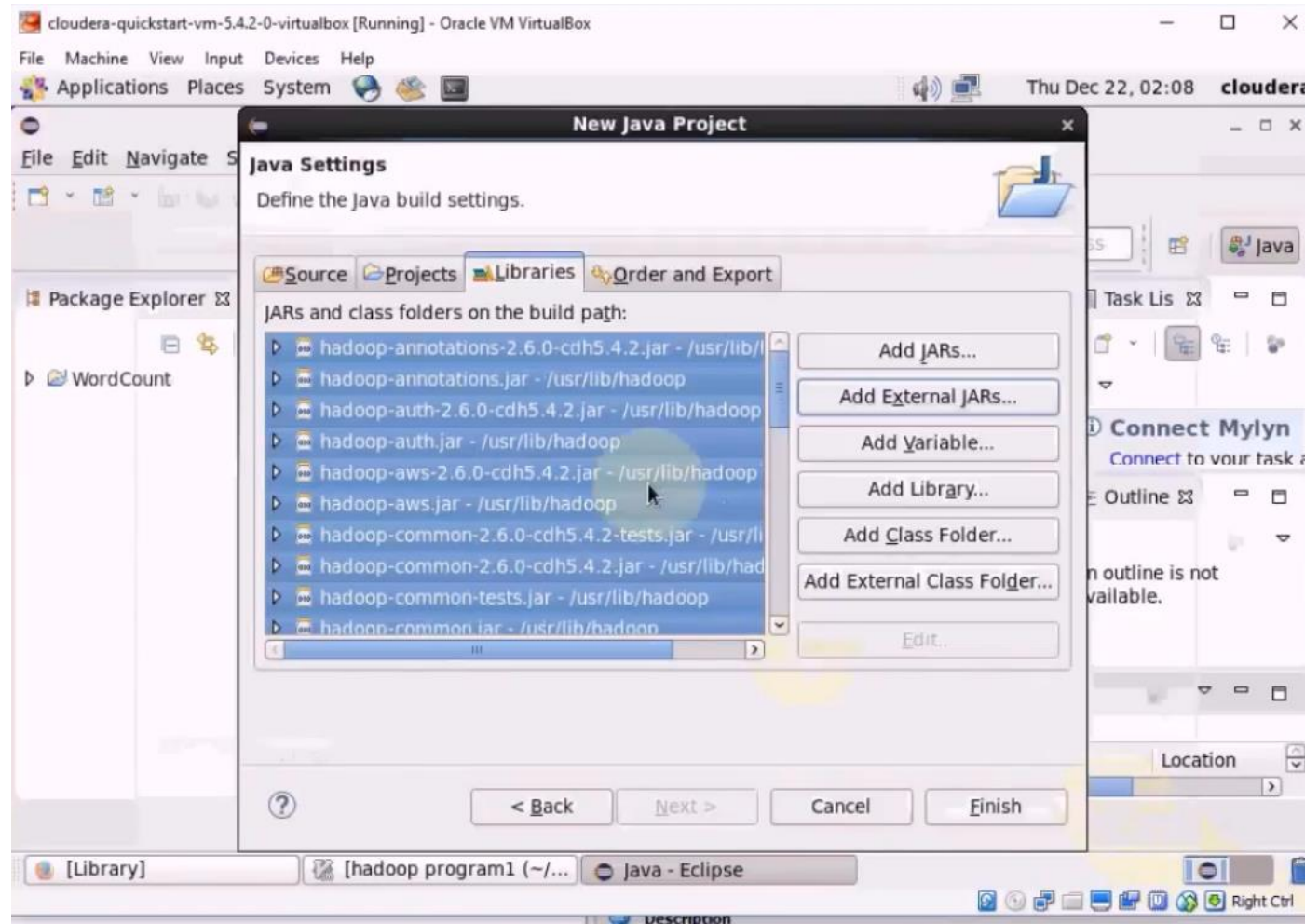


Após isso realizamos a instalação da máquina virtual e executamos.

Após isso, já dentro do Cloudera, abrimos a IDE eclipse, que já veio instalada na VM, para podermos criar o jar file, que será o map reduce do nosso exemplo.



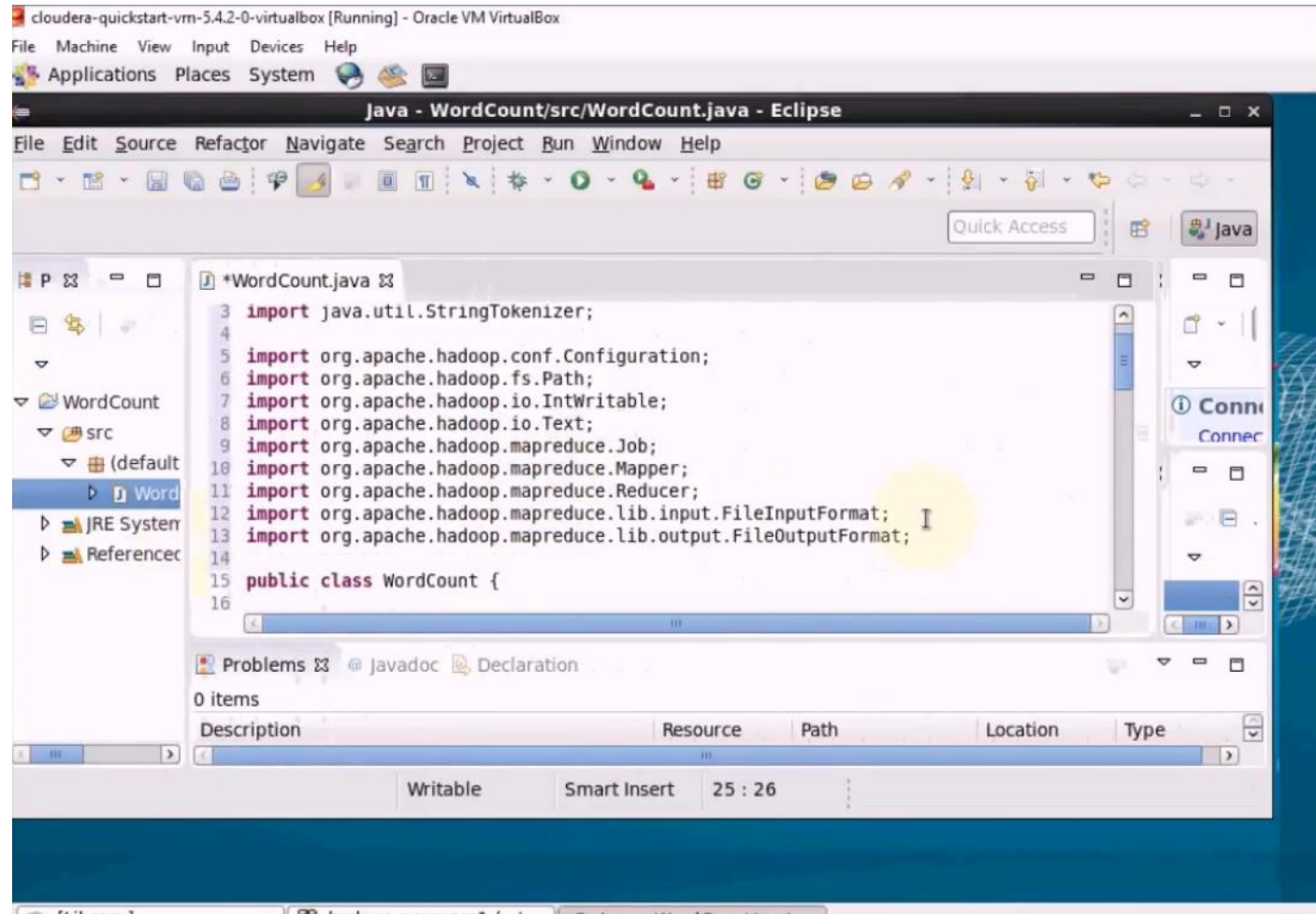
Após isso, criamos um novo arquivo java em branco, e importamos todas as bibliotecas hadoop, e as bibliotecas externas do hadoop também.



Após isso, copiamos e colamos um script java, que é o nosso map reduce. Esse script irá receber uma lista de palavras, e irá contar quantas vezes cada palavra se repete na lista.

O script completo está disponível em:

<https://github.com/dbtsai/hadoop-word-count/blob/master/mapreduce/src/main/java/com/dbtsai/hadoop/mapreduce/WordCountMR.java>



```
cloudera-quickstart-vm-5.4.2-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System

Java - WordCount/src/WordCount.java - Eclipse
File Edit Source Refactor Navigate Search Project Run Window Help

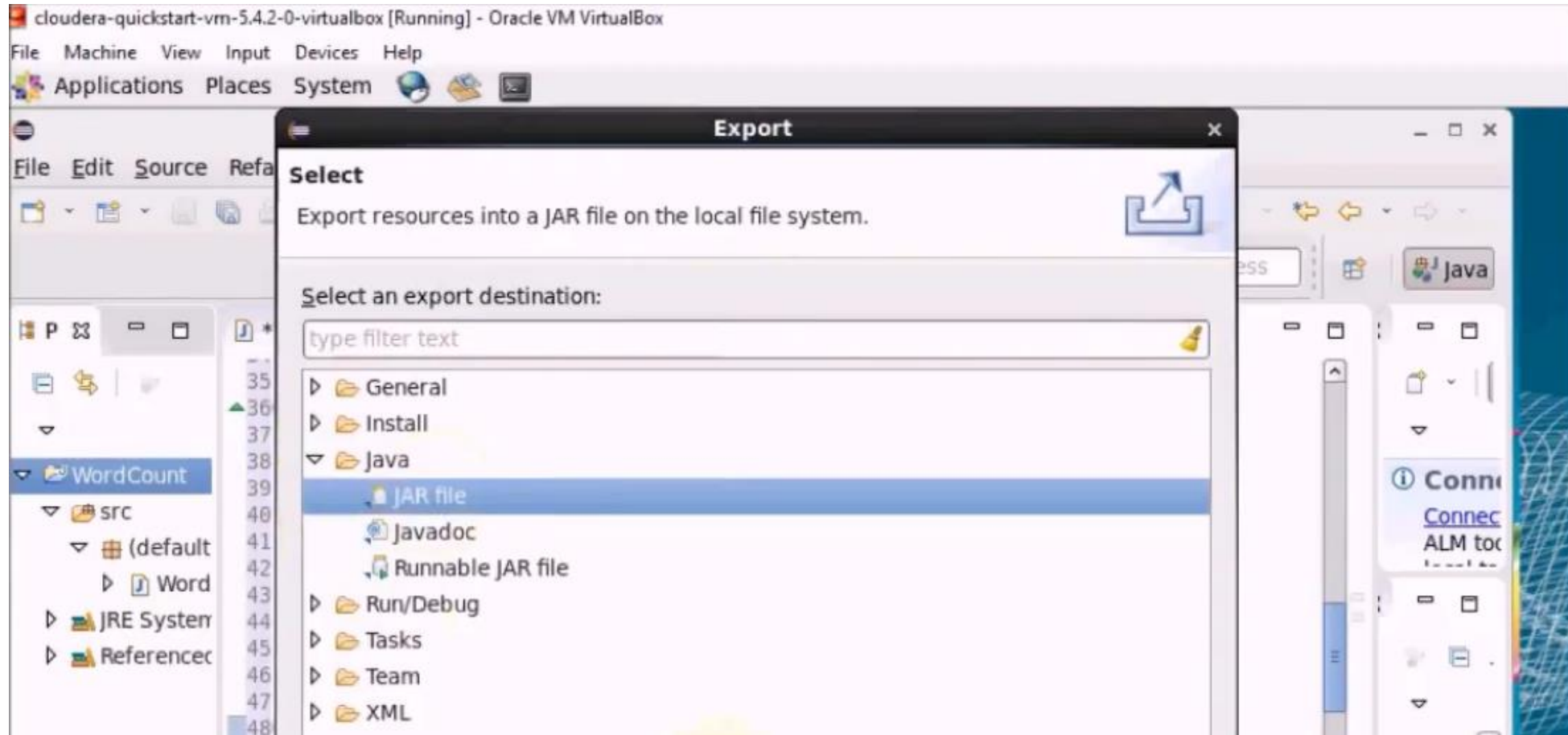
Quick Access

WordCount
  src
    (default)
      Word

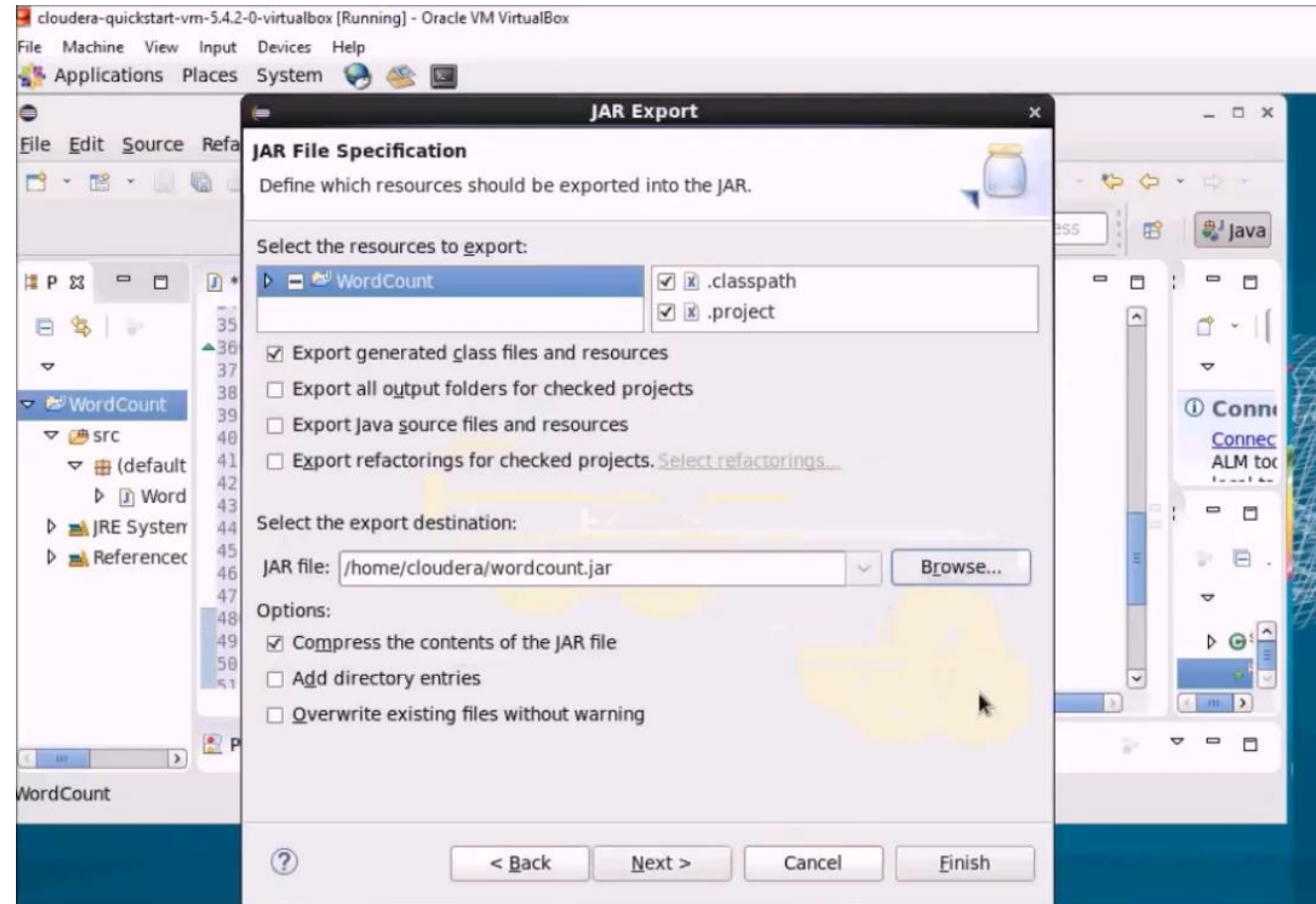
JRE System
References

WordCount.java
3 import java.util.StringTokenizer;
4
5 import org.apache.hadoop.conf.Configuration;
6 import org.apache.hadoop.fs.Path;
7 import org.apache.hadoop.io.IntWritable;
8 import org.apache.hadoop.io.Text;
9 import org.apache.hadoop.mapreduce.Job;
10 import org.apache.hadoop.mapreduce.Mapper;
11 import org.apache.hadoop.mapreduce.Reducer;
12 import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
13 import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
14
15 public class WordCount {
16
```

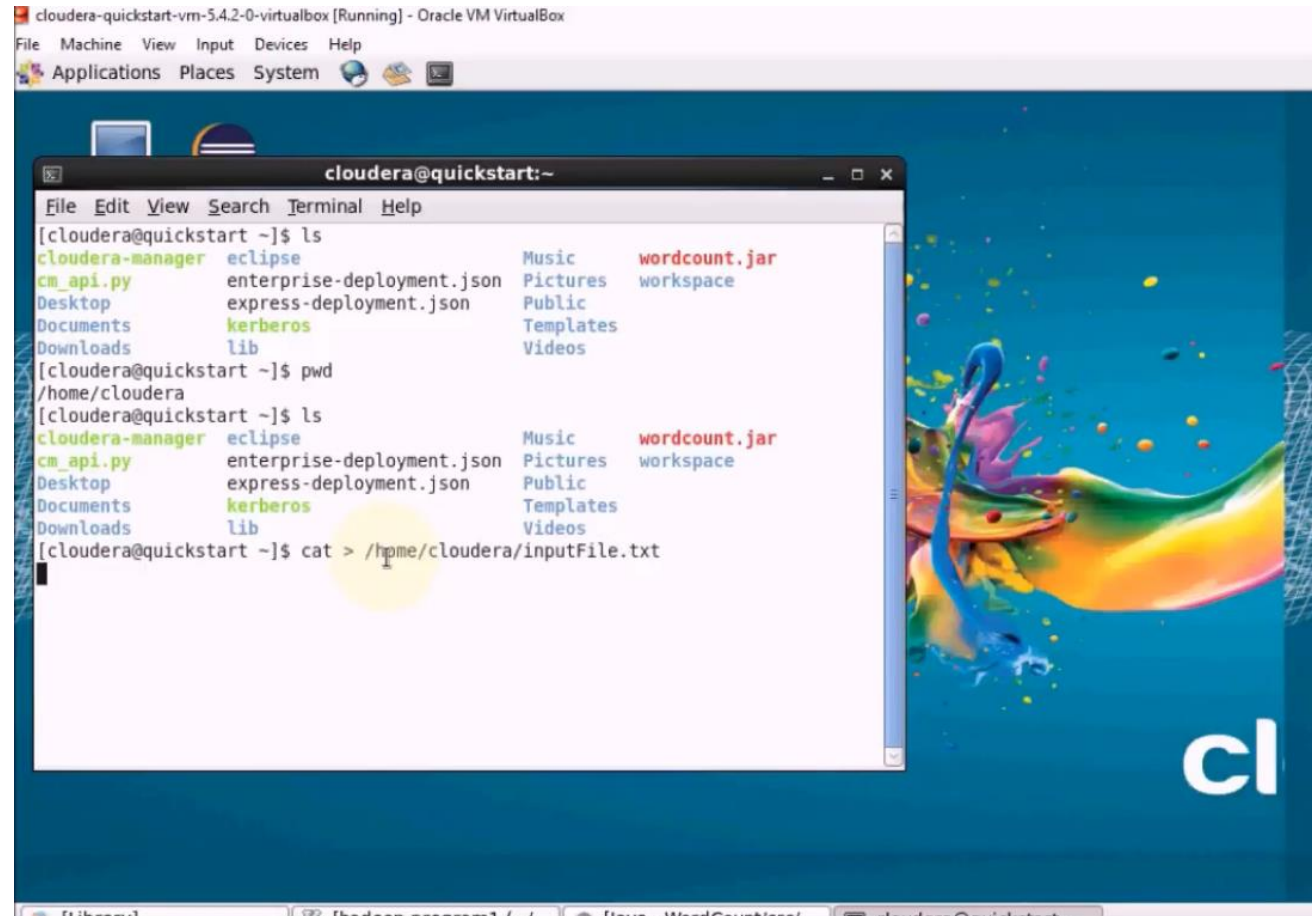
Após isso, clicamos com o botão direito sobre o projeto do eclipse, e exportamos o projeto no formato .jar file, que é um arquivo reconhecido pelo hadoop para realizar o mapreduce



Após isso, indicamos o local de salvamento do arquivo gerado, que é na pasta raiz do cloudera (HOME).



Após isso, usando o terminal do cloudera, criamos um arquivo de texto, que será o nosso input file, nele escrevemos as palavras que queremos contar.



The screenshot shows a terminal window titled "cloudera@quickstart:~" within an Oracle VM VirtualBox environment. The terminal displays the following commands and output:

```
[cloudera@quickstart ~]$ ls
cloudera-manager  eclipse          Music      wordcount.jar
cm_api.py         enterprise-deployment.json Pictures    workspace
Desktop          express-deployment.json Public
Documents        kerberos        Templates
Downloads        lib             Videos

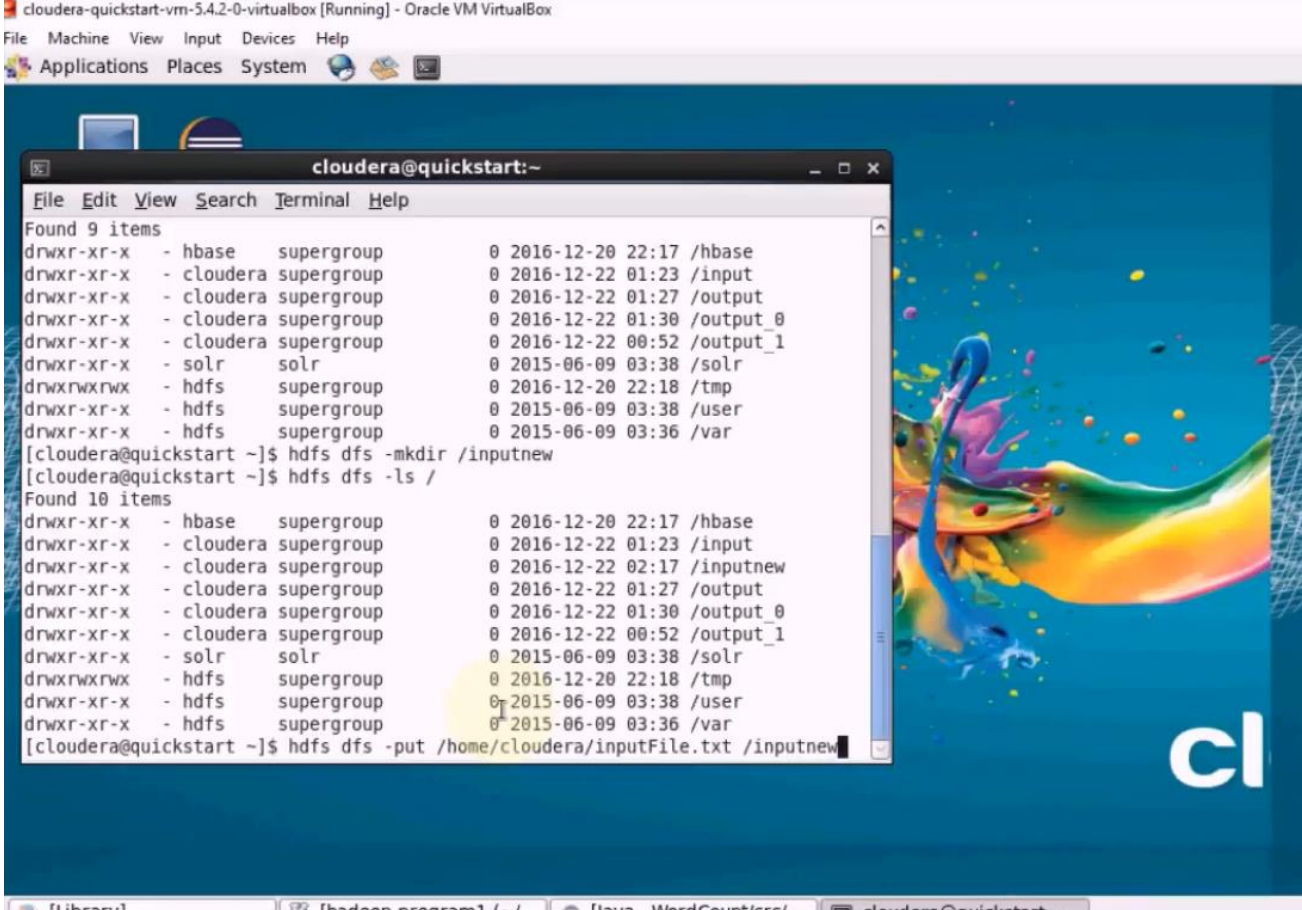
[cloudera@quickstart ~]$ pwd
/home/cloudera

[cloudera@quickstart ~]$ ls
cloudera-manager  eclipse          Music      wordcount.jar
cm_api.py         enterprise-deployment.json Pictures    workspace
Desktop          express-deployment.json Public
Documents        kerberos        Templates
Downloads        lib             Videos

[cloudera@quickstart ~]$ cat > /home/cloudera/inputFile.txt
```

The background of the VM desktop features a blue wall with a colorful abstract splash and the Cloudera logo in the bottom right corner.

Após isso, salvamos o inputfile.txt na pasta raíz.



The screenshot shows a terminal window titled 'cloudera@quickstart:~' within an Oracle VM VirtualBox environment. The terminal displays the output of the 'ls -l' command, showing file permissions, owners, groups, sizes, and timestamps for various files and directories. The output is as follows:

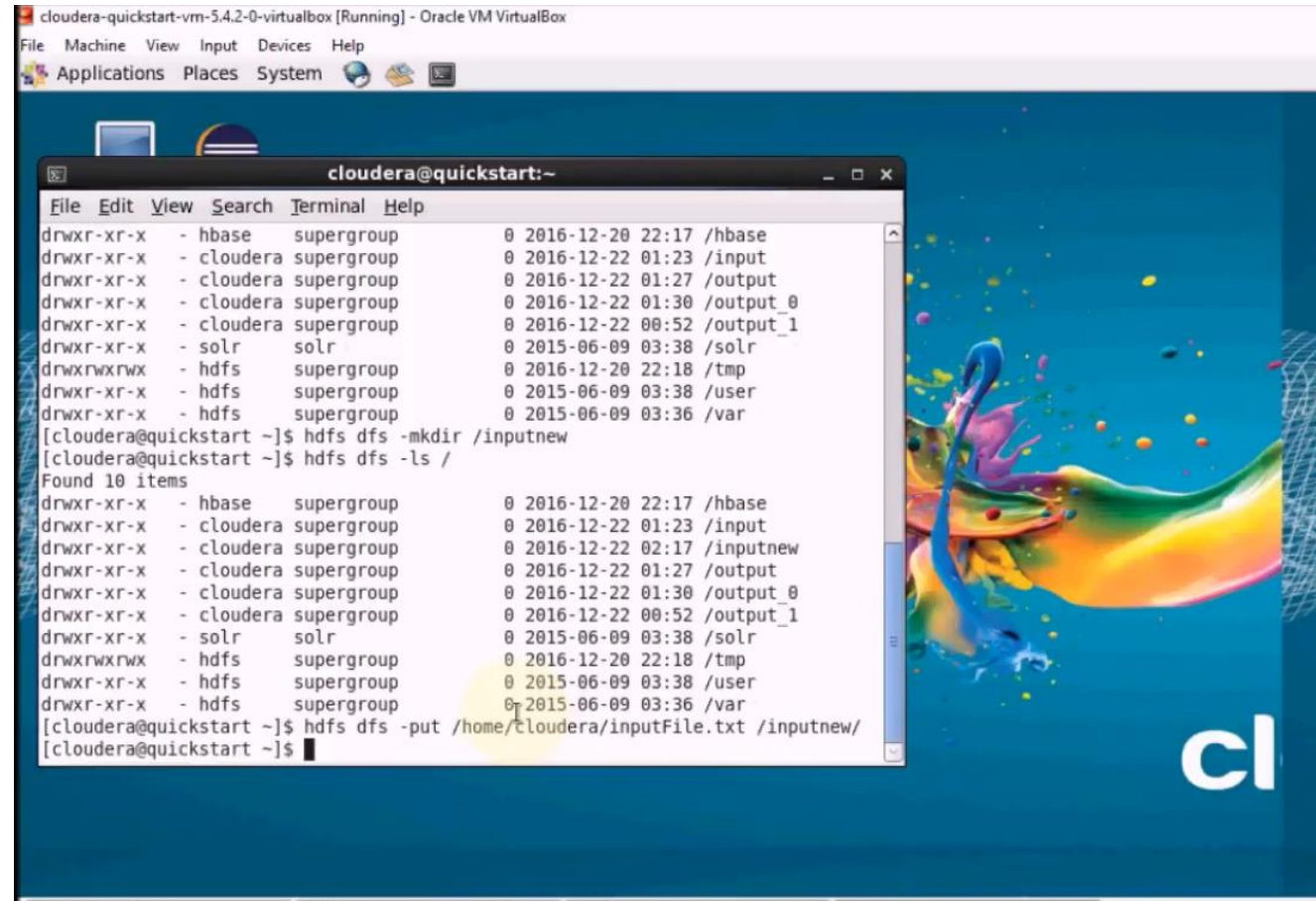
```
Found 9 items
drwxr-xr-x - hbase supergroup 0 2016-12-20 22:17 /hbase
drwxr-xr-x - cloudera supergroup 0 2016-12-22 01:23 /input
drwxr-xr-x - cloudera supergroup 0 2016-12-22 01:27 /output
drwxr-xr-x - cloudera supergroup 0 2016-12-22 01:30 /output_0
drwxr-xr-x - cloudera supergroup 0 2016-12-22 00:52 /output_1
drwxr-xr-x - solr solr 0 2015-06-09 03:38 /solr
drwxrwxrwx - hdfs supergroup 0 2016-12-20 22:18 /tmp
drwxr-xr-x - hdfs supergroup 0 2015-06-09 03:38 /user
drwxr-xr-x - hdfs supergroup 0 2015-06-09 03:36 /var

[cloudera@quickstart ~]$ hdfs dfs -mkdir /inputnew
[cloudera@quickstart ~]$ hdfs dfs -ls /
Found 10 items
drwxr-xr-x - hbase supergroup 0 2016-12-20 22:17 /hbase
drwxr-xr-x - cloudera supergroup 0 2016-12-22 01:23 /input
drwxr-xr-x - cloudera supergroup 0 2016-12-22 02:17 /inputnew
drwxr-xr-x - cloudera supergroup 0 2016-12-22 01:27 /output
drwxr-xr-x - cloudera supergroup 0 2016-12-22 01:30 /output_0
drwxr-xr-x - cloudera supergroup 0 2016-12-22 00:52 /output_1
drwxr-xr-x - solr solr 0 2015-06-09 03:38 /solr
drwxrwxrwx - hdfs supergroup 0 2016-12-20 22:18 /tmp
drwxr-xr-x - hdfs supergroup 0 2015-06-09 03:38 /user
drwxr-xr-x - hdfs supergroup 0 2015-06-09 03:36 /var

[cloudera@quickstart ~]$ hdfs dfs -put /home/cloudera/inputFile.txt /inputnew
```

The terminal window is part of a larger application window titled 'cloudera-quickstart-vm-5.4.2-0-virtualbox [Running] - Oracle VM VirtualBox'. The background of the terminal window features a colorful abstract design with a large 'cl' logo in the bottom right corner.

Após isso, por meio do hadoop executamos o arquivo jar, e indicamos onde está o script java, a classe, o input e a saída que queremos.

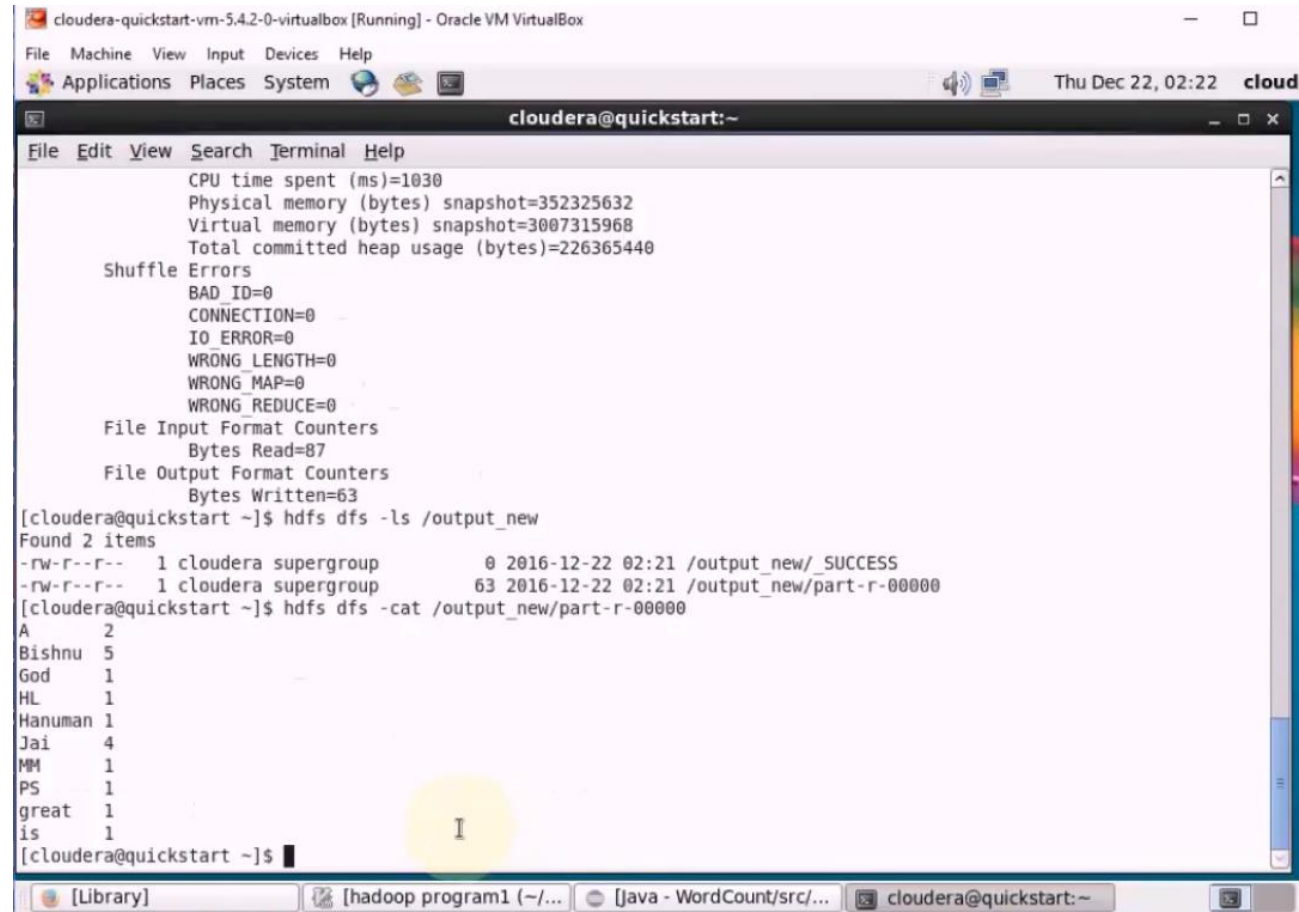


The screenshot shows a terminal window titled 'cloudera@quickstart:~' within an Oracle VM VirtualBox environment. The terminal displays the output of several HDFS commands. First, 'hdfs dfs -ls /' lists the root directory contents, showing directories like hbase, cloudera, solr, tmp, user, and var. Then, 'hdfs dfs -mkdir /inputnew' is executed. Finally, 'hdfs dfs -ls /' is run again, showing the new 'inputnew' directory has been added to the root. The background of the VM desktop features a colorful abstract splash graphic and the Cloudera logo.

```
cloudera-quickstart-vm-5.4.2-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System

cloudera@quickstart:~
File Edit View Search Terminal Help
drwxr-xr-x - hbase supergroup 0 2016-12-20 22:17 /hbase
drwxr-xr-x - cloudera supergroup 0 2016-12-22 01:23 /input
drwxr-xr-x - cloudera supergroup 0 2016-12-22 01:27 /output
drwxr-xr-x - cloudera supergroup 0 2016-12-22 01:30 /output_0
drwxr-xr-x - cloudera supergroup 0 2016-12-22 00:52 /output_1
drwxr-xr-x - solr solr 0 2015-06-09 03:38 /solr
drwxrwxrwx - hdfs supergroup 0 2016-12-20 22:18 /tmp
drwxr-xr-x - hdfs supergroup 0 2015-06-09 03:38 /user
drwxr-xr-x - hdfs supergroup 0 2015-06-09 03:36 /var
[cloudera@quickstart ~]$ hdfs dfs -mkdir /inputnew
[cloudera@quickstart ~]$ hdfs dfs -ls /
Found 10 items
drwxr-xr-x - hbase supergroup 0 2016-12-20 22:17 /hbase
drwxr-xr-x - cloudera supergroup 0 2016-12-22 01:23 /input
drwxr-xr-x - cloudera supergroup 0 2016-12-22 02:17 /inputnew
drwxr-xr-x - cloudera supergroup 0 2016-12-22 01:27 /output
drwxr-xr-x - cloudera supergroup 0 2016-12-22 01:30 /output_0
drwxr-xr-x - cloudera supergroup 0 2016-12-22 00:52 /output_1
drwxr-xr-x - solr solr 0 2015-06-09 03:38 /solr
drwxrwxrwx - hdfs supergroup 0 2016-12-20 22:18 /tmp
drwxr-xr-x - hdfs supergroup 0 2015-06-09 03:38 /user
drwxr-xr-x - hdfs supergroup 0 2015-06-09 03:36 /var
[cloudera@quickstart ~]$ hdfs dfs -put /home/cloudera/inputFile.txt /inputnew/
[cloudera@quickstart ~]$
```

Após isso, analisamos o outputfile, que é o resultado que buscamos.



The screenshot shows a terminal window titled "cloudera-quickstart-vm-5.4.2-0-virtualbox [Running] - Oracle VM VirtualBox". The terminal prompt is "cloudera@quickstart:~". The output of the commands is as follows:

```
File Edit View Search Terminal Help
CPU time spent (ms)=1030
Physical memory (bytes) snapshot=352325632
Virtual memory (bytes) snapshot=3007315968
Total committed heap usage (bytes)=226365440
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=87
File Output Format Counters
Bytes Written=63
[cloudera@quickstart ~]$ hdfs dfs -ls /output_new
Found 2 items
-rw-r--r-- 1 cloudera supergroup 0 2016-12-22 02:21 /output_new/_SUCCESS
-rw-r--r-- 1 cloudera supergroup 63 2016-12-22 02:21 /output_new/part-r-00000
[cloudera@quickstart ~]$ hdfs dfs -cat /output_new/part-r-00000
A 2
Bishnu 5
God 1
HL 1
Hanuman 1
Jai 4
MM 1
PS 1
great 1
is 1
[cloudera@quickstart ~]$
```

The terminal window has a taskbar at the bottom with icons for [Library], [hadoop program1 (-/...], [java - WordCount/src/...], and cloudera@quickstart:~.