

# Engenharia de Dados com Hadoop e Spark

## Sumário

1	Versão.....	3
2	Configuração do Ambiente.....	4
2.1	Criação da Máquina Virtual no VirtualBox .....	5
2.2	Instalação do Sistema Operacional .....	10
2.3	Instalação de Utilitários do Sistema Operacional .....	28
3	Instalação do servidor ssh .....	34
4	Instalação do Java 8 .....	43
5	Instalação e Configuração do Hadoop.....	49
5.1	Criando o usuário hadoop.....	49
5.2	Configuração do ssh sem senha .....	53
5.3	Download e Instalação do Hadoop .....	66
5.3.1	Editando o arquivo hosts .....	66
5.3.2	Download do Hadoop .....	68
5.4	Configuração do Hadoop .....	76
5.4.1	Editar arquivos de configuração do Hadoop .....	76
5.4.2	Formatando o Namenode .....	81
5.4.3	Iniciando o Hadoop.....	83
5.4.4	Iniciando o Yarn .....	86
5.5	Processando Big Data.....	90
6	Instalação e Configuração do Spark .....	105
6.1	Download e Instalação do Spark .....	105
7	Processando Big Data com Spark .....	115
8	Referências .....	119

## 1 Versão

Este documento foi criado por Cleber Zumba de Souza e pode ser distribuído livremente, desde que se faça menção à fonte.

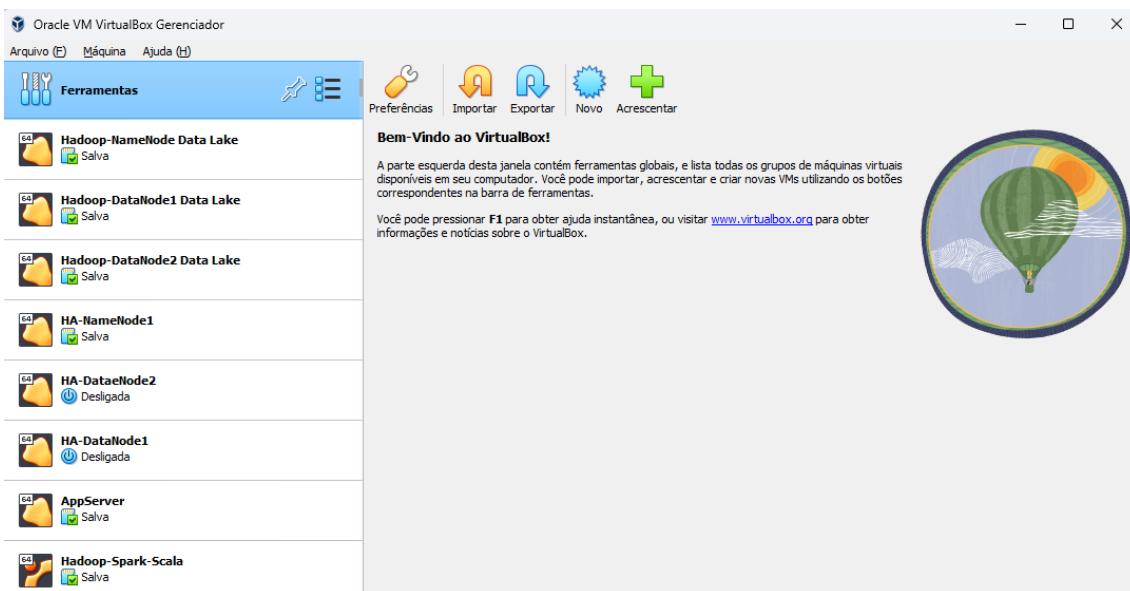
Versão	Ação	Data
1.0	Criação do documento	21/04/2024

## 2 Configuração do Ambiente

Item	Versão
Virtual Box	7.0
Sistema Operacional	Centos 7 (64 bits)
Interface Gráfica	Gnome
Java	1.8
Apache Hadoop	3.3.6
Apache Spark	3.4.3

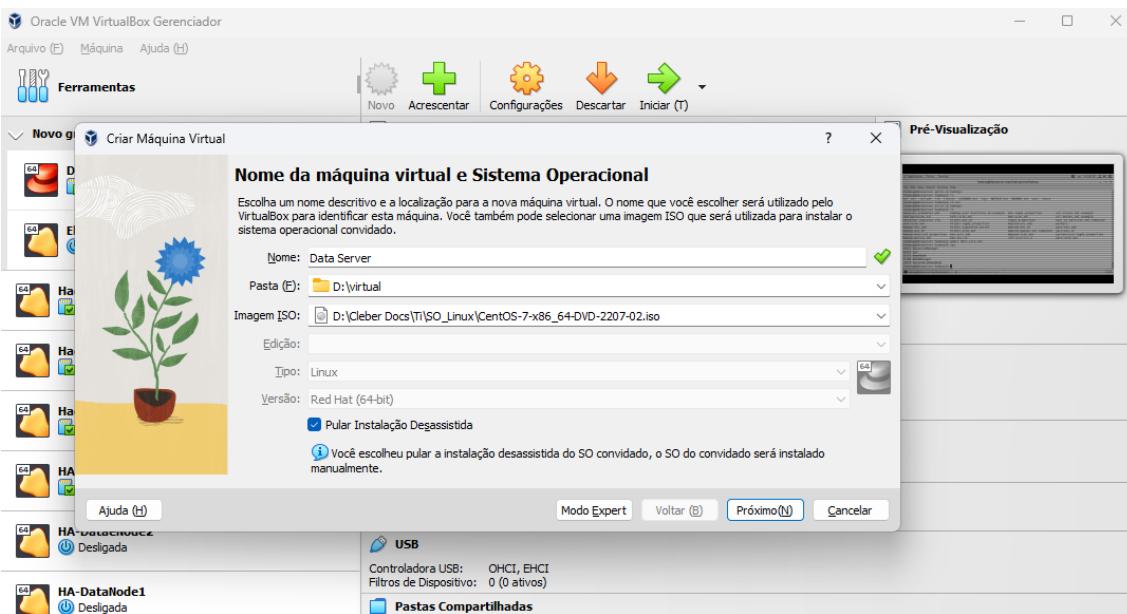
## 2.1 Criação da Máquina Virtual no VirtualBox

O Oracle VM Virtual Box é gratuito e pode ser baixado em <https://www.virtualbox.org>. Aqui utilizei a versão 7.0.



Abrindo o Gerenciador do Oracle Virtual Box

# Instalação e Configuração do Ambiente Big Data com Hadoop e Spark

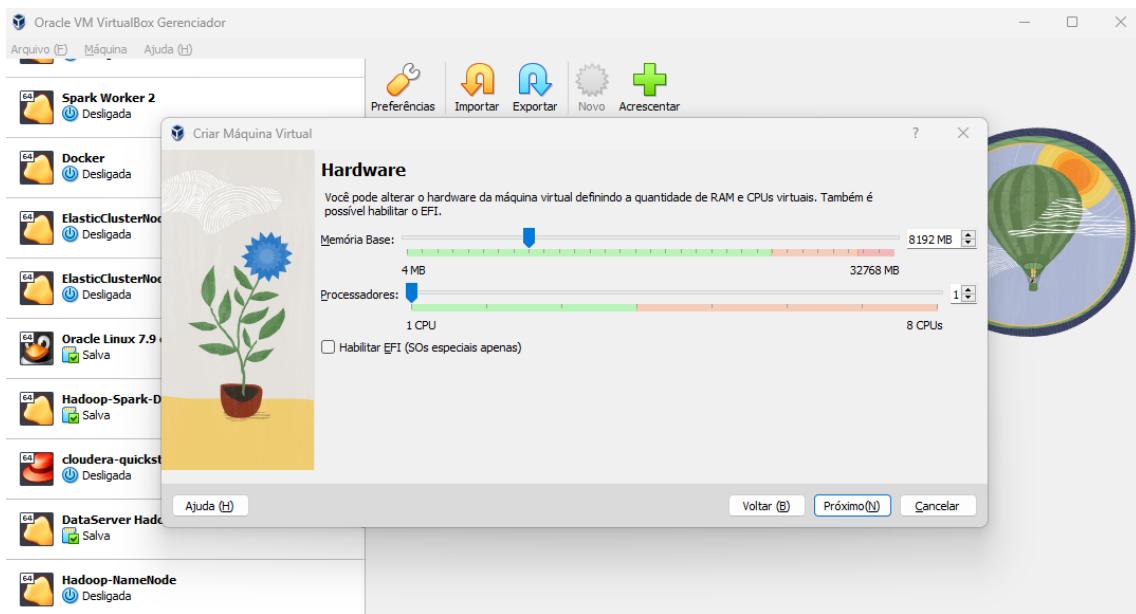


Definindo o nome da máquina virtual e a versão do sistema operacional

Selecione a mídia de instalação do sistema operacional

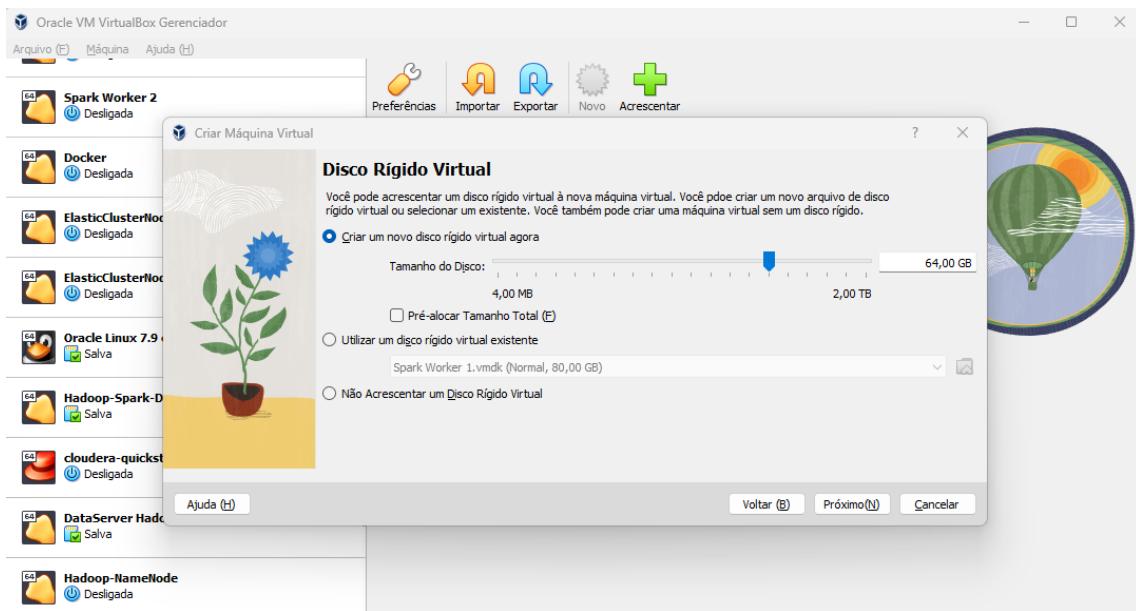
CentOS 64 bits (versão 7)

## Instalação e Configuração do Ambiente Big Data com Hadoop e Spark



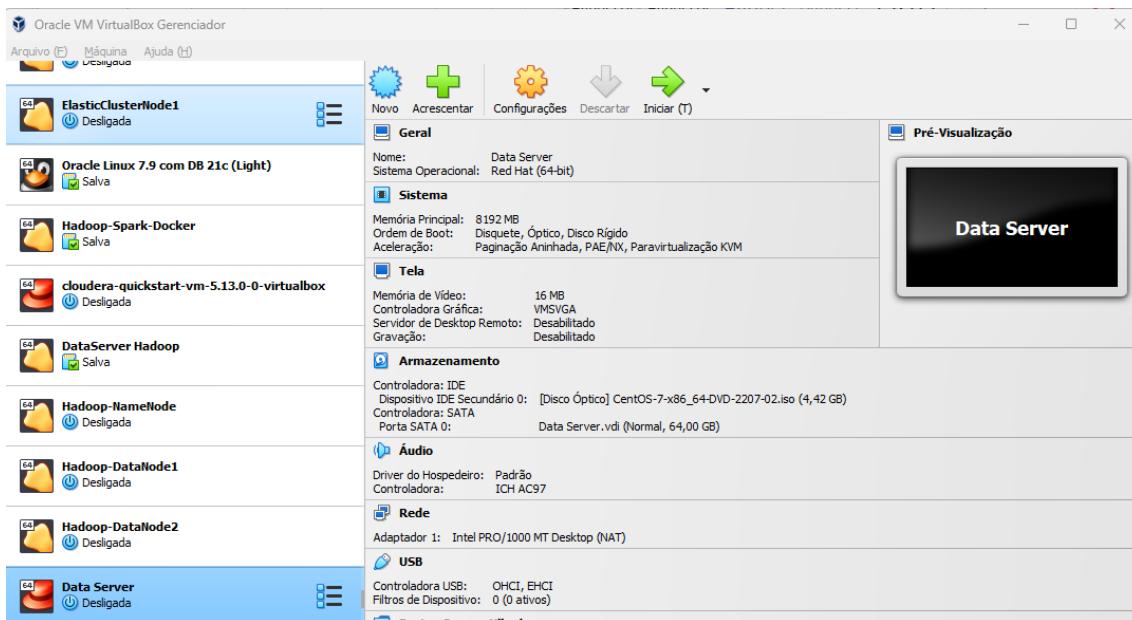
Configure metade da memória física do seu computador para a VM

# Instalação e Configuração do Ambiente Big Data com Hadoop e Spark



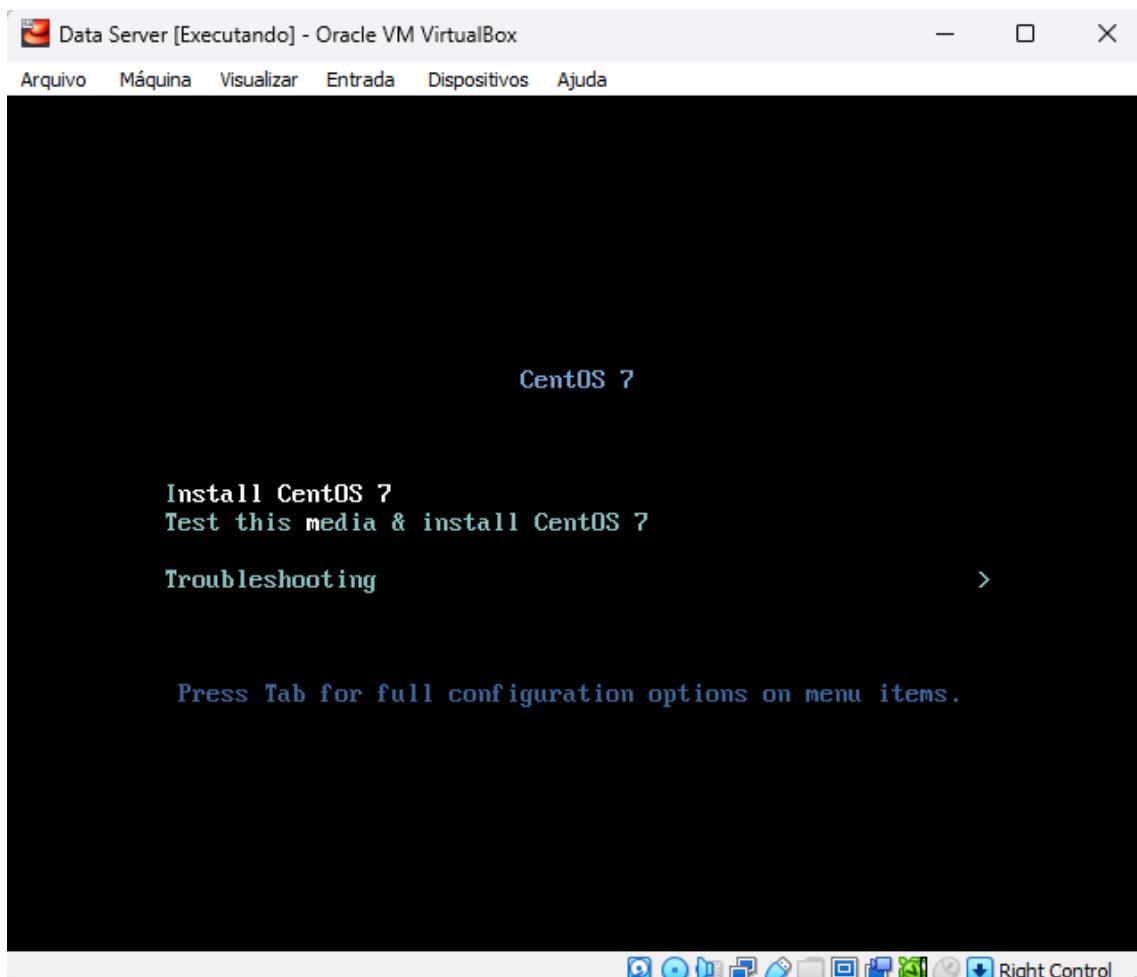
Selecione 64 GB para o disco virtual

## Instalação e Configuração do Ambiente Big Data com Hadoop e Spark



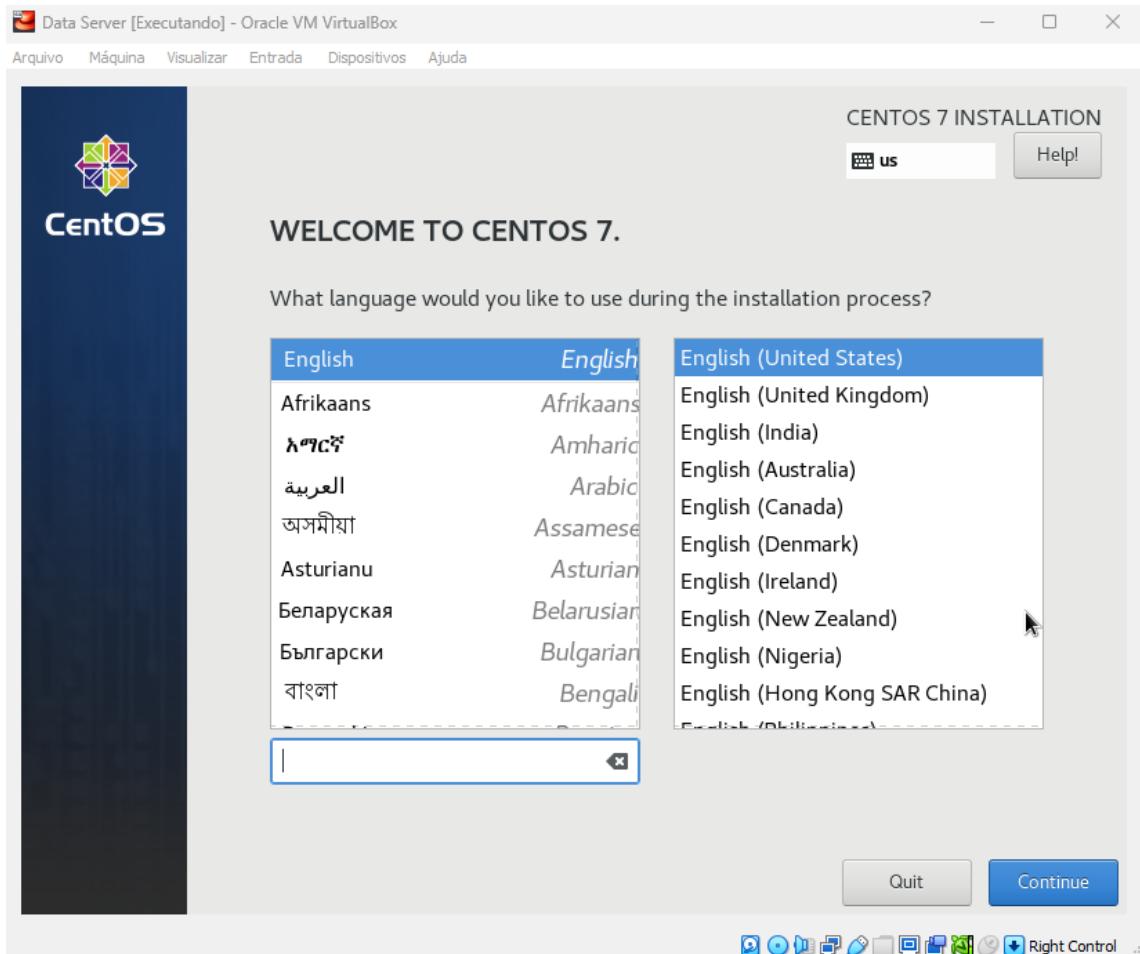
Máquina virtual criada. Selecione a VM e clique no botão Iniciar para inicializar a VM

## 2.2 Instalação do Sistema Operacional



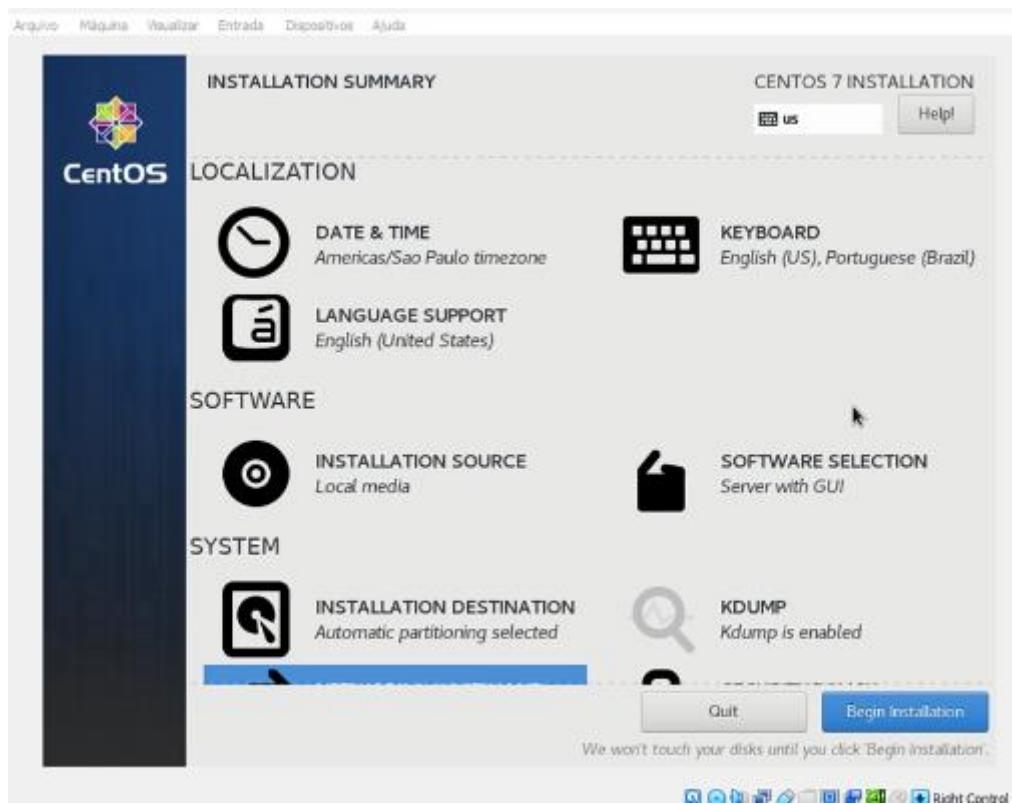
Selecione a opção de Instalação do Sistema Operacional CentOS 7

## Instalação e Configuração do Ambiente Big Data com Hadoop e Spark



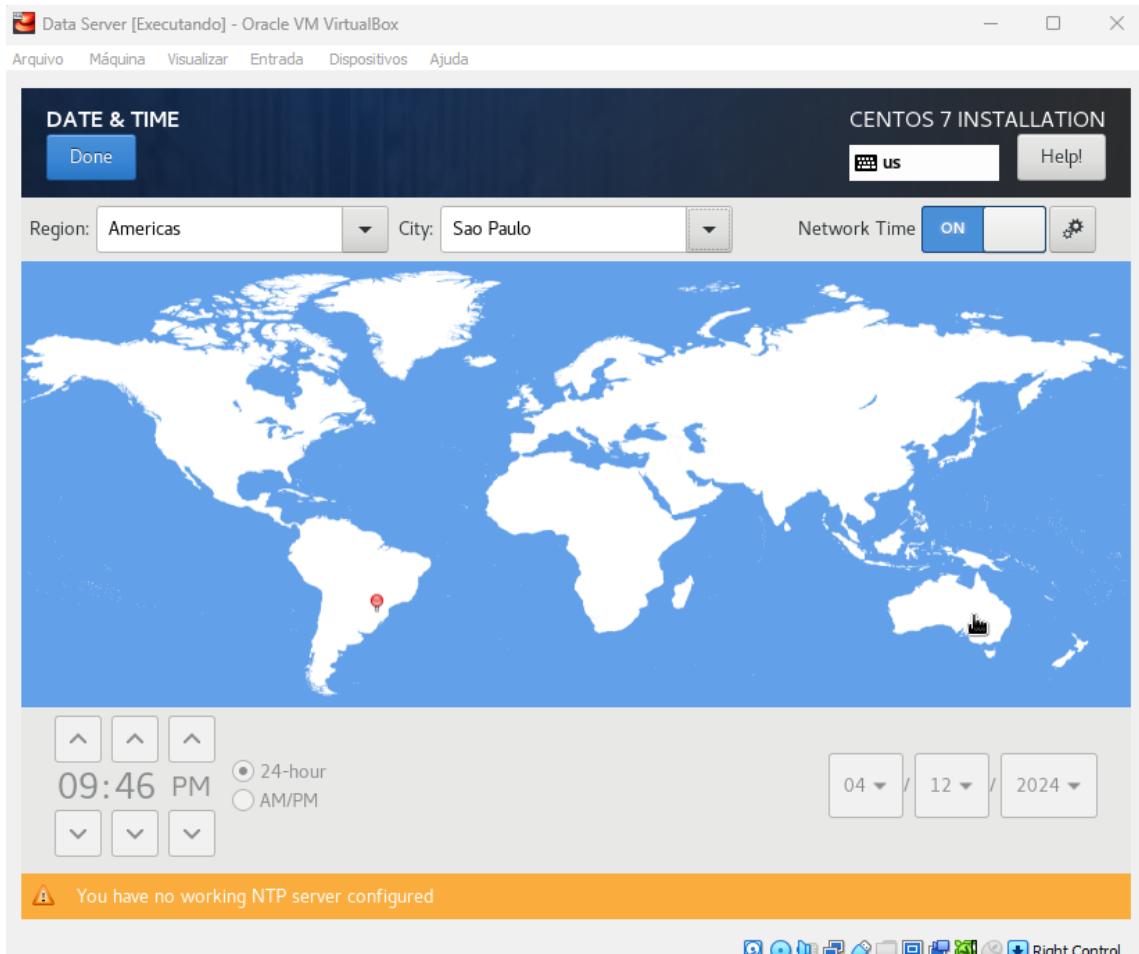
Seleção do idioma usado na instalação

## Instalação e Configuração do Ambiente Big Data com Hadoop e Spark



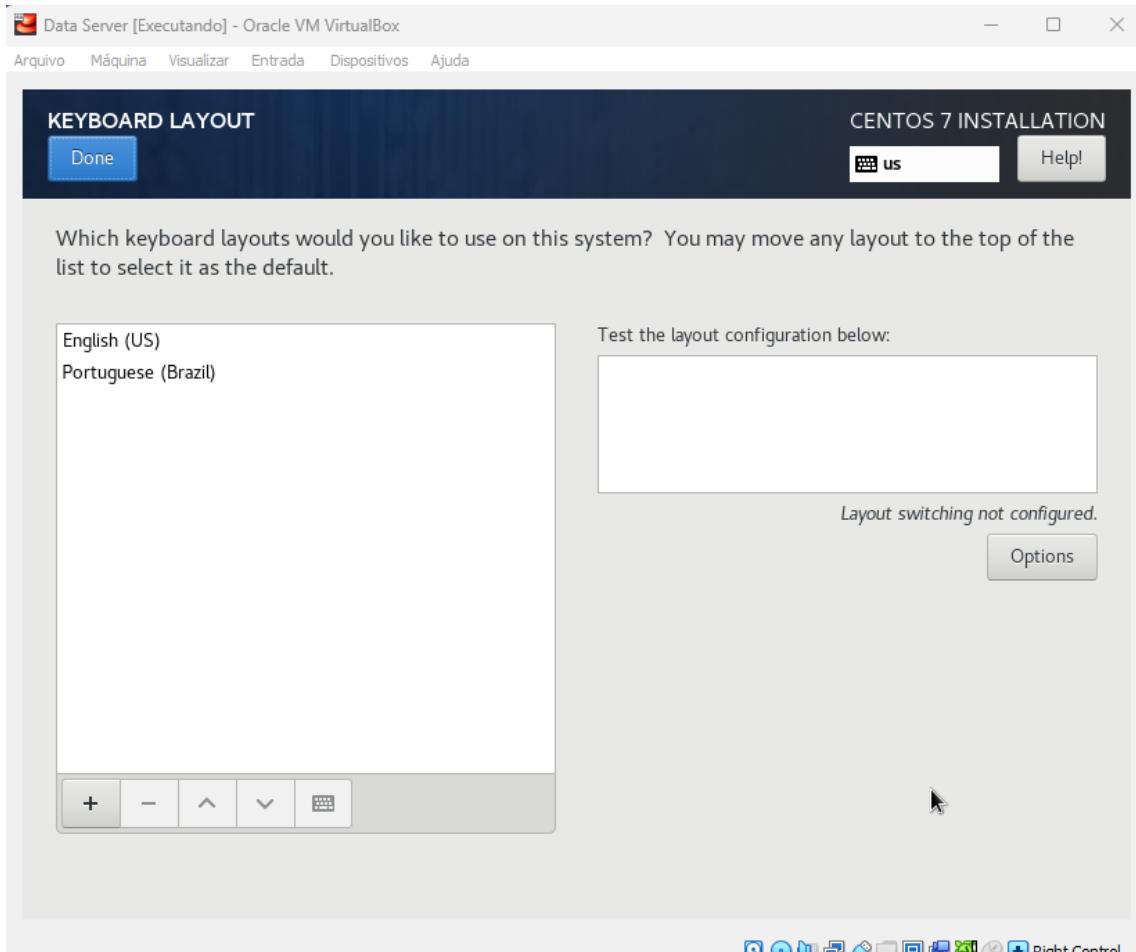
Opções de configuração

## Instalação e Configuração do Ambiente Big Data com Hadoop e Spark



### Timezone

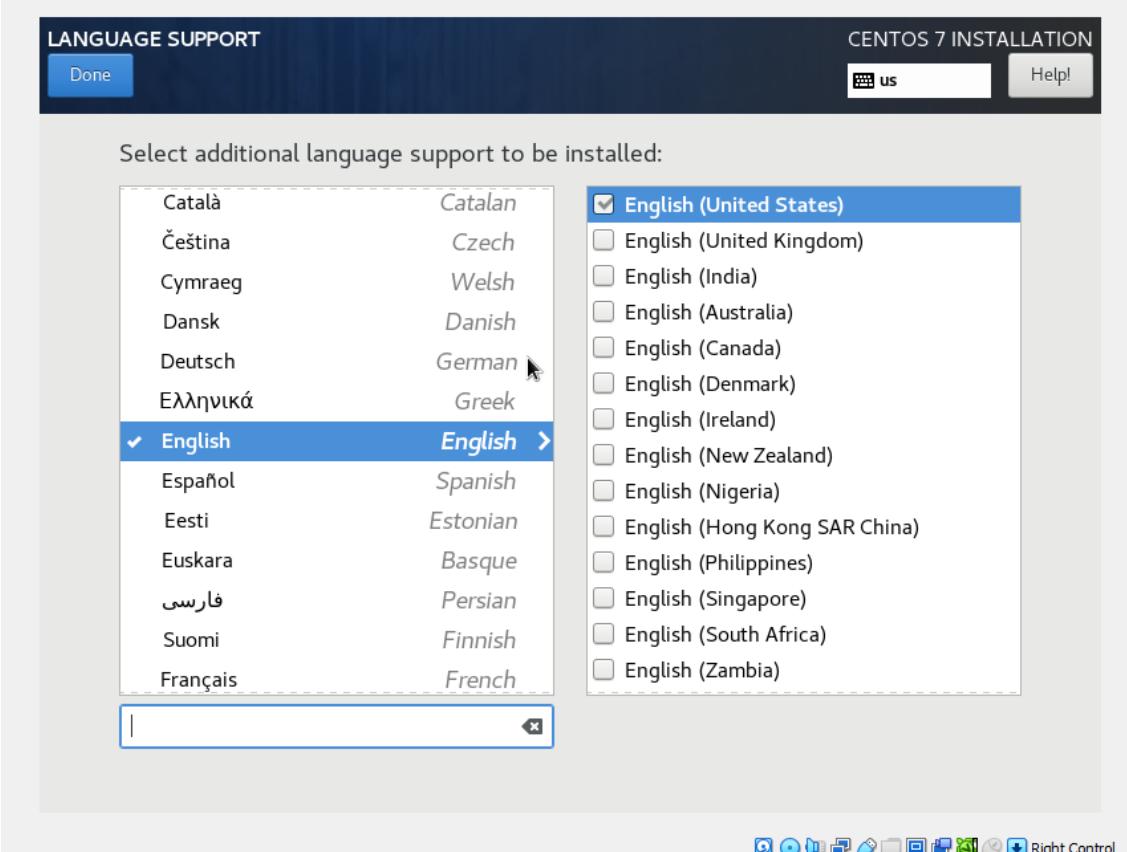
## Instalação e Configuração do Ambiente Big Data com Hadoop e Spark



Layout do teclado

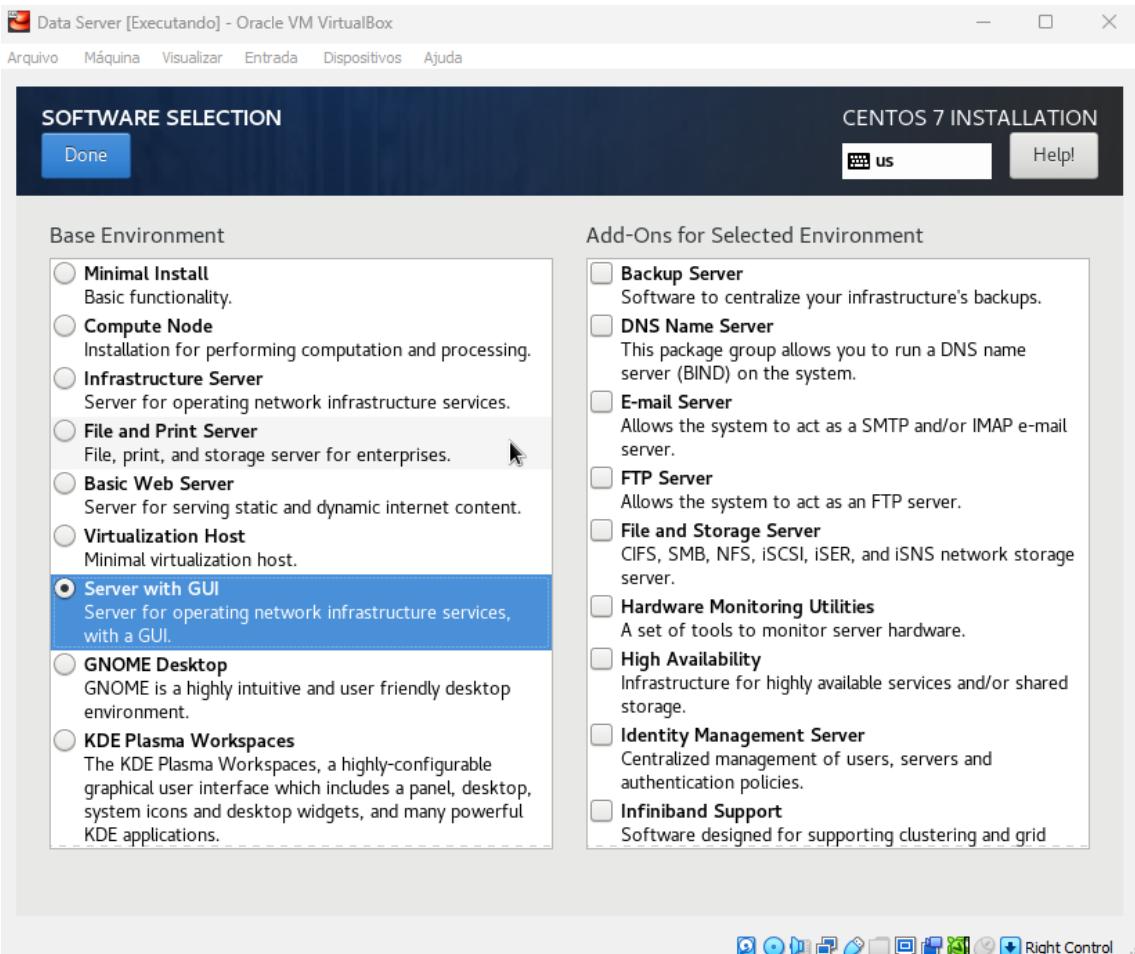
# Instalação e Configuração do Ambiente Big Data com Hadoop e Spark

Arquivo Máquina Visualizar Entrada Dispositivos Ajuda



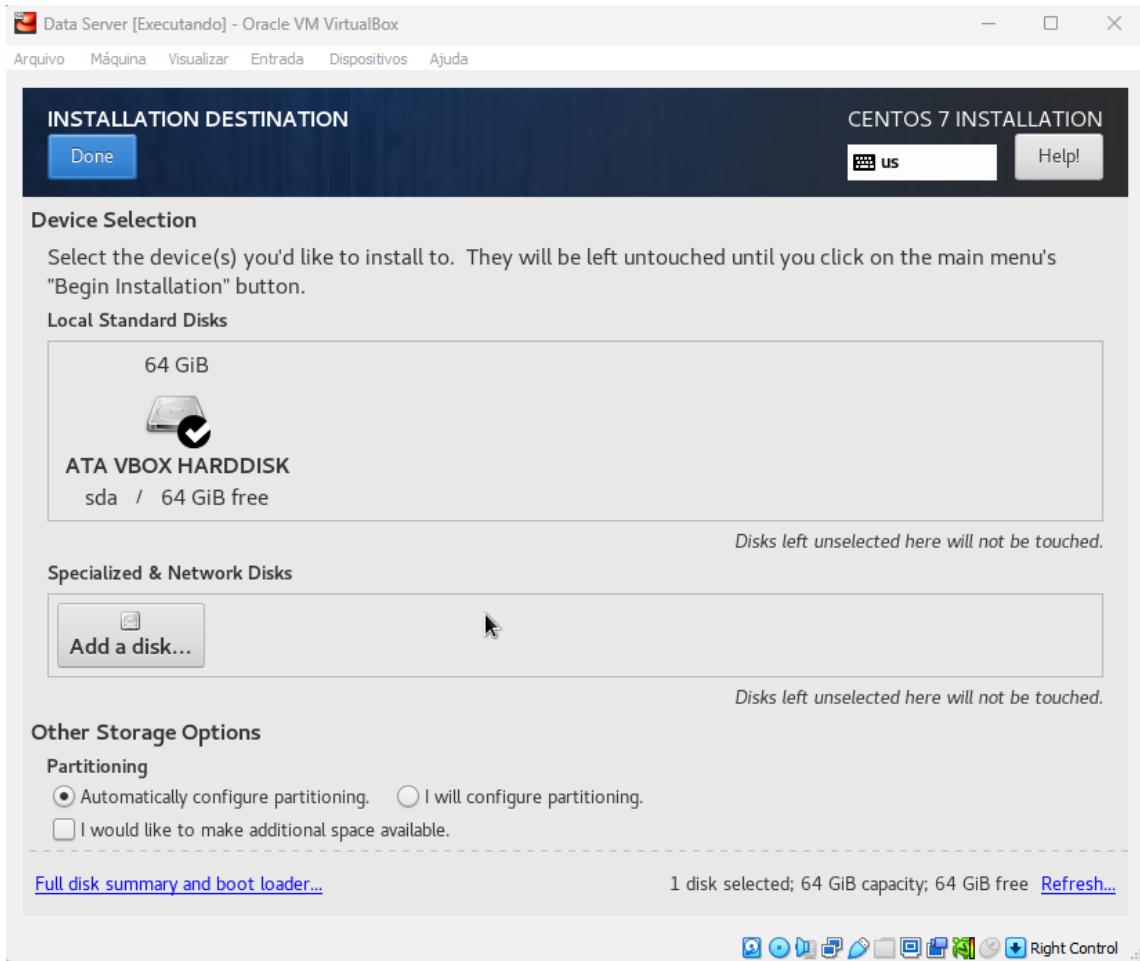
Idioma do sistema operacional

# Instalação e Configuração do Ambiente Big Data com Hadoop e Spark



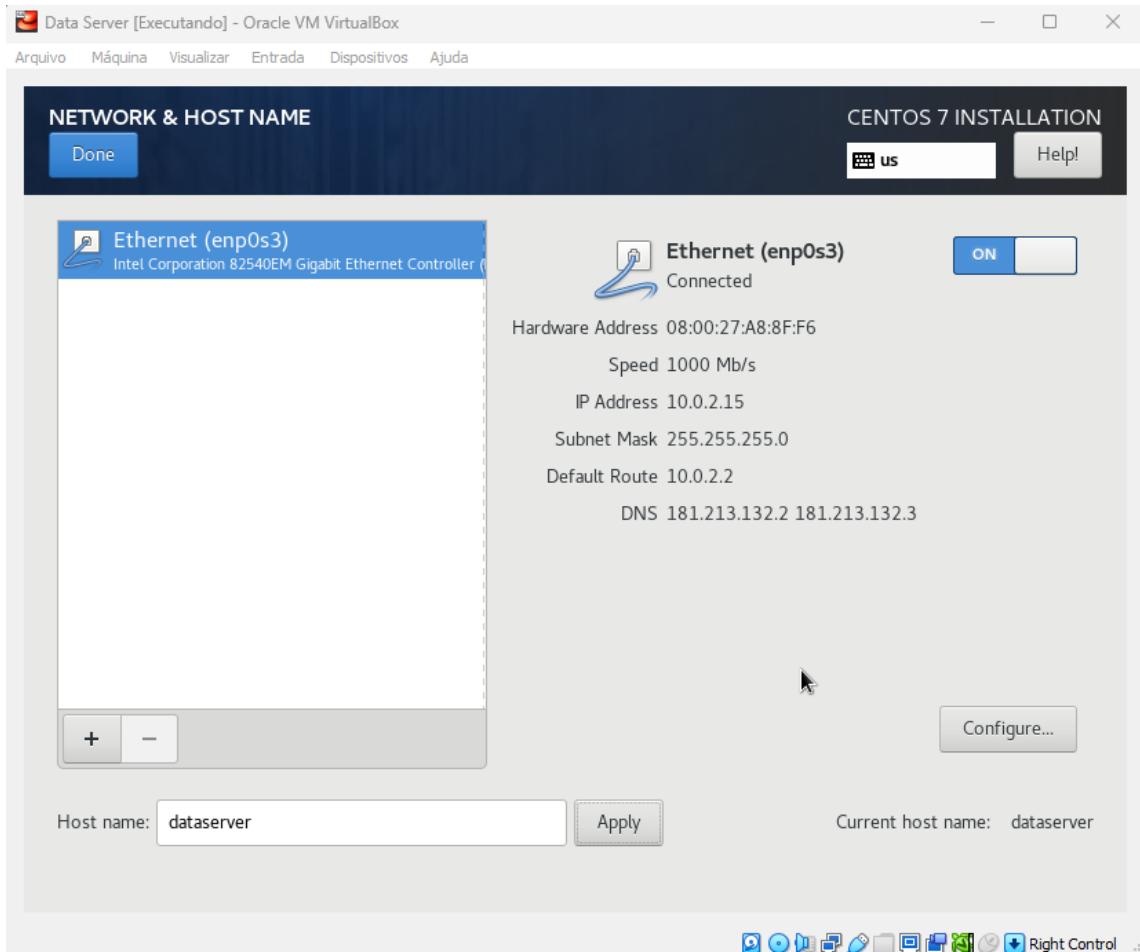
Selecione a opção Server with GUI

# Instalação e Configuração do Ambiente Big Data com Hadoop e Spark



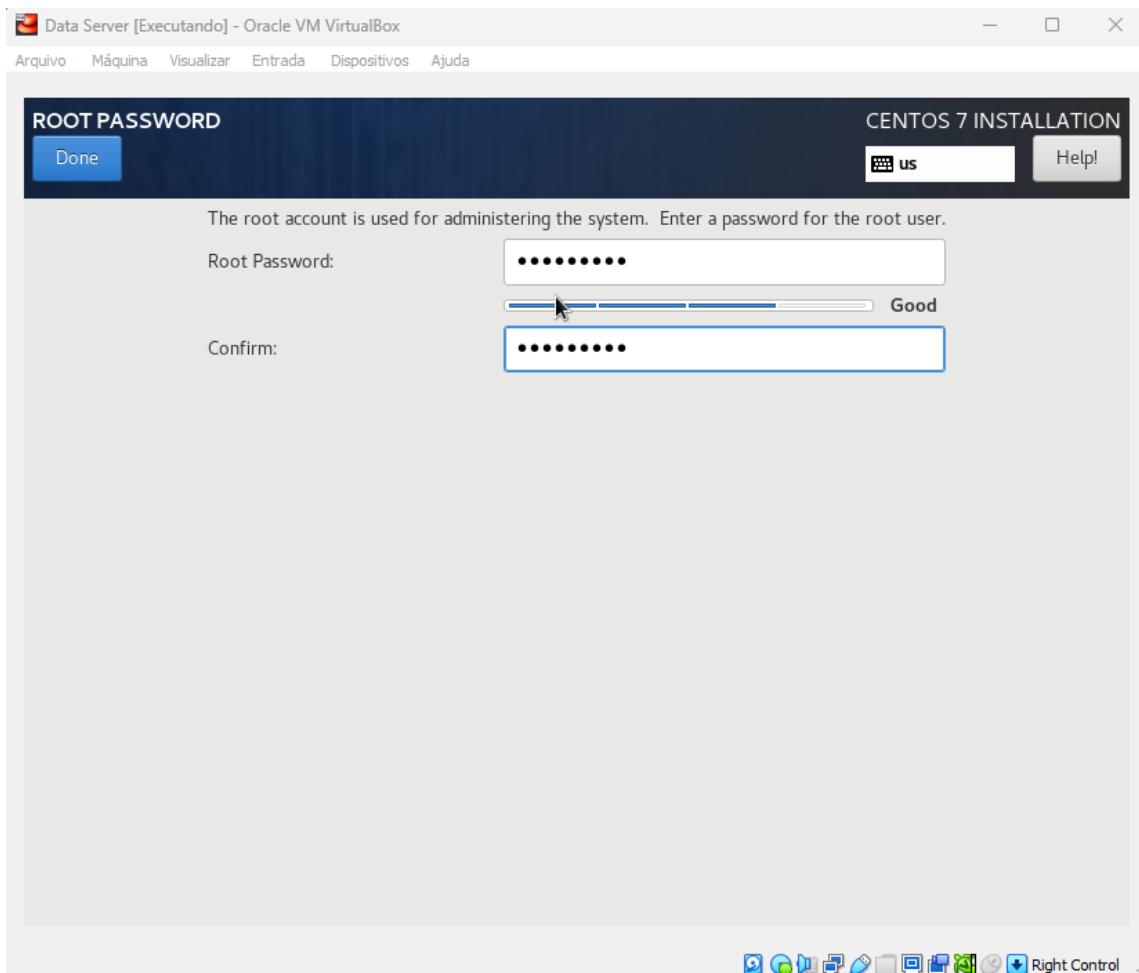
Disco

## Instalação e Configuração do Ambiente Big Data com Hadoop e Spark



Configuração de rede e nome do servidor – **dataserver** – Clique em Apply  
Certifique-se de habilitar a opção de ativar a Ethernet (botão on)

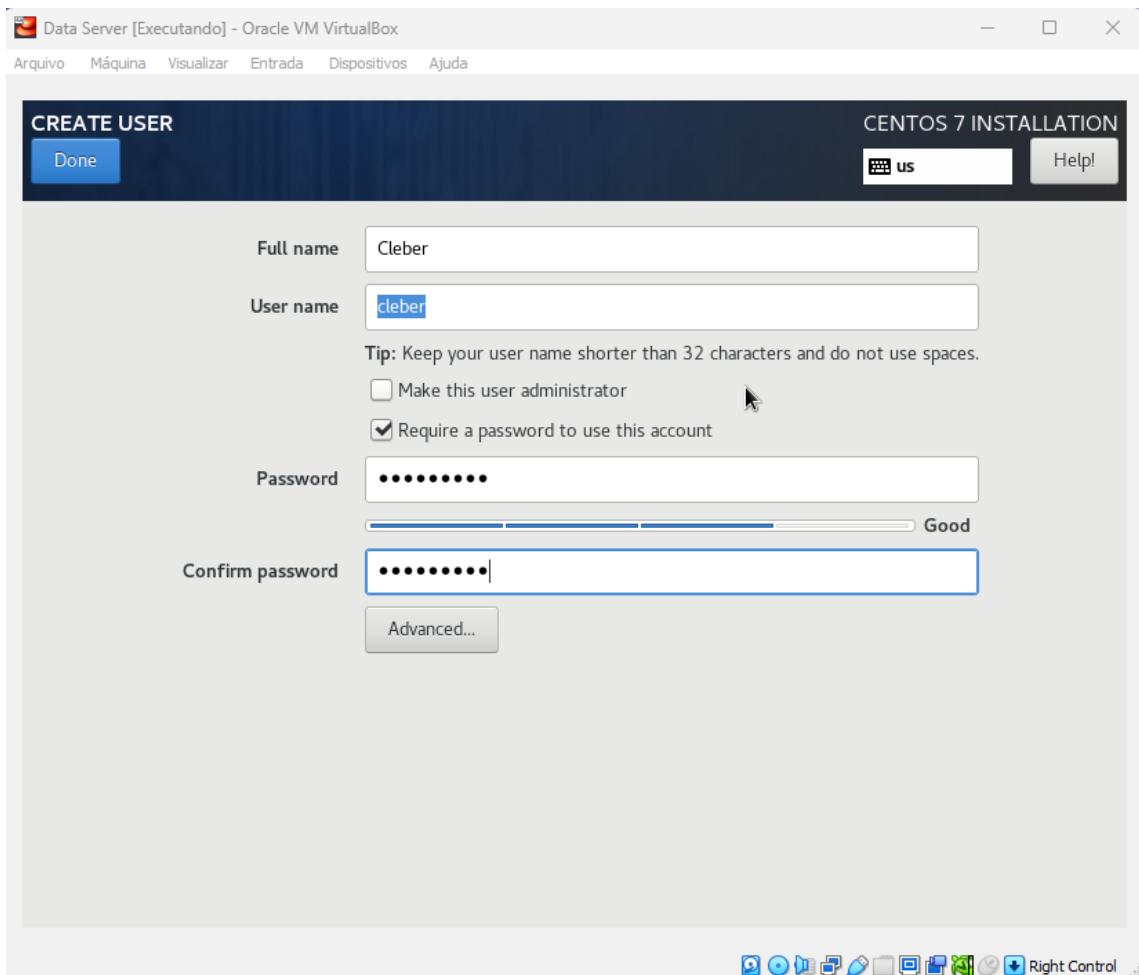
## Instalação e Configuração do Ambiente Big Data com Hadoop e Spark



Definir senha do root – usuário administrador

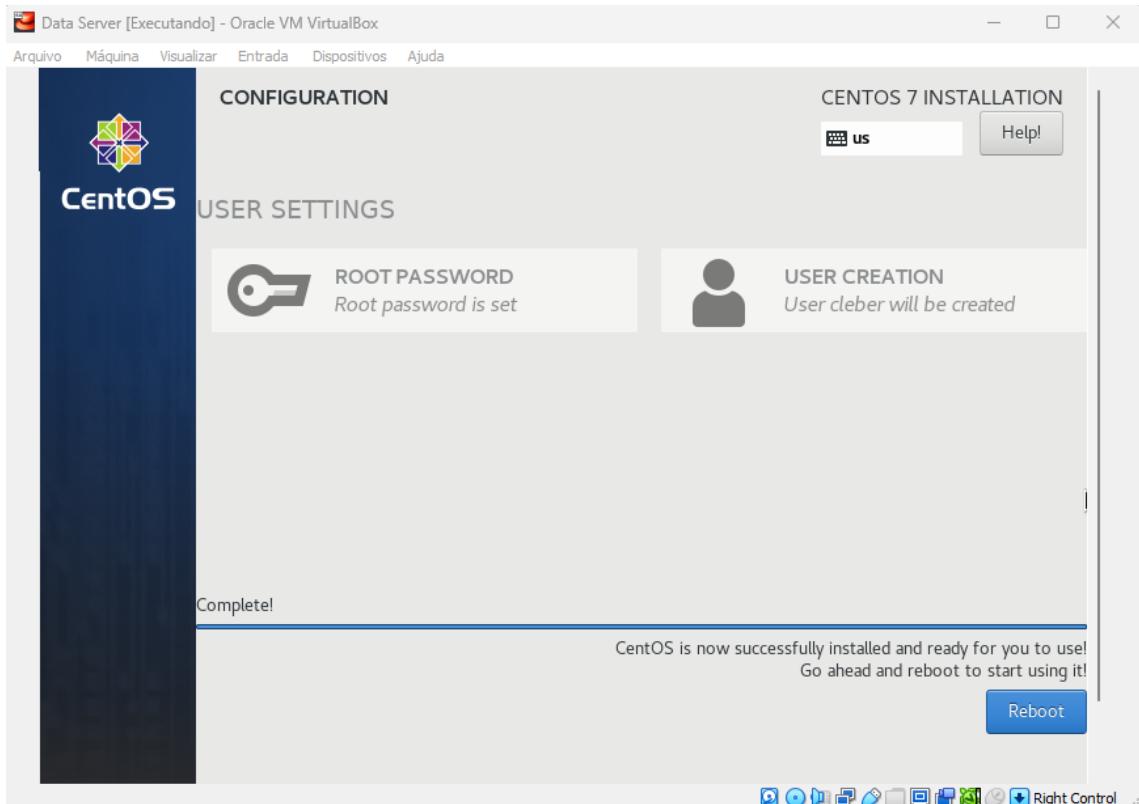
Senha: **hadoop123**

## Instalação e Configuração do Ambiente Big Data com Hadoop e Spark



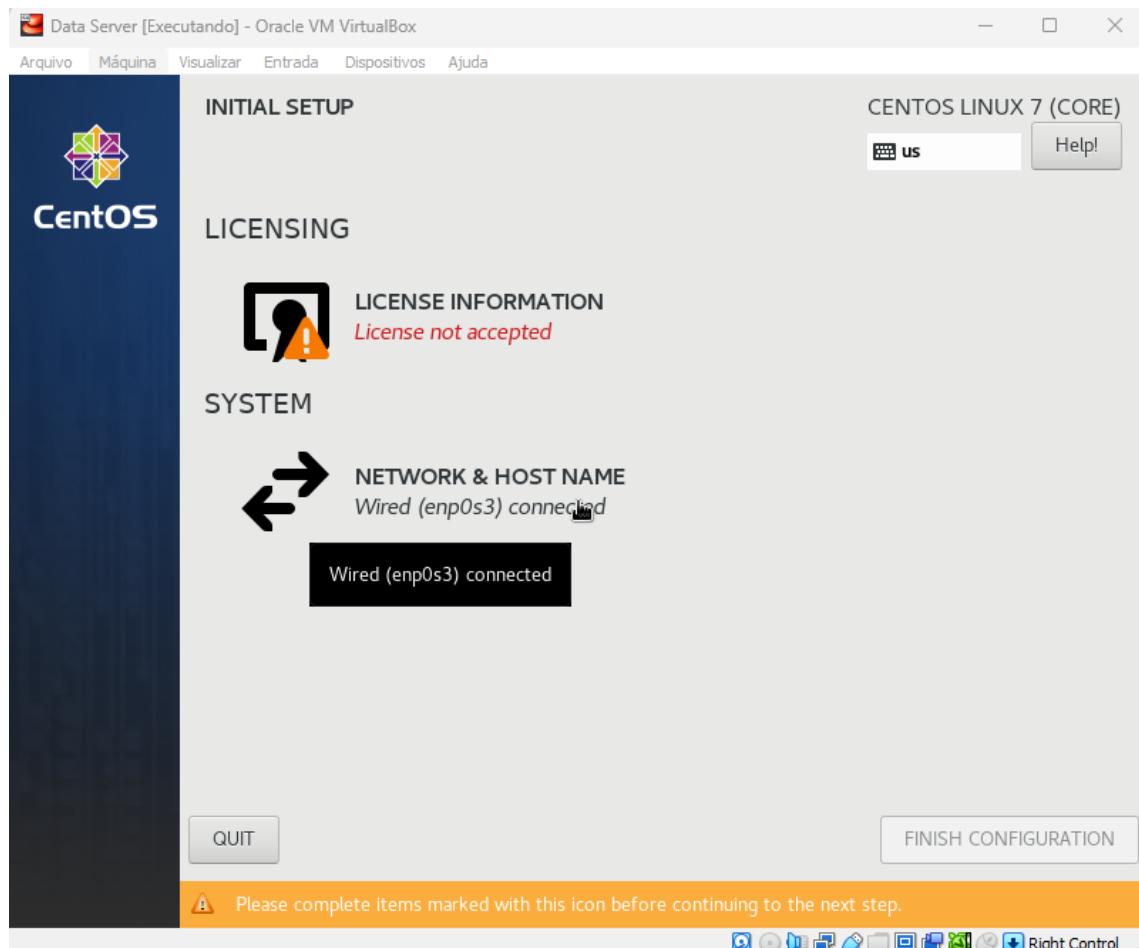
Criação de um usuário – Cleber  
(username: cleber, senha: **hadoop123**)

## Instalação e Configuração do Ambiente Big Data com Hadoop e Spark



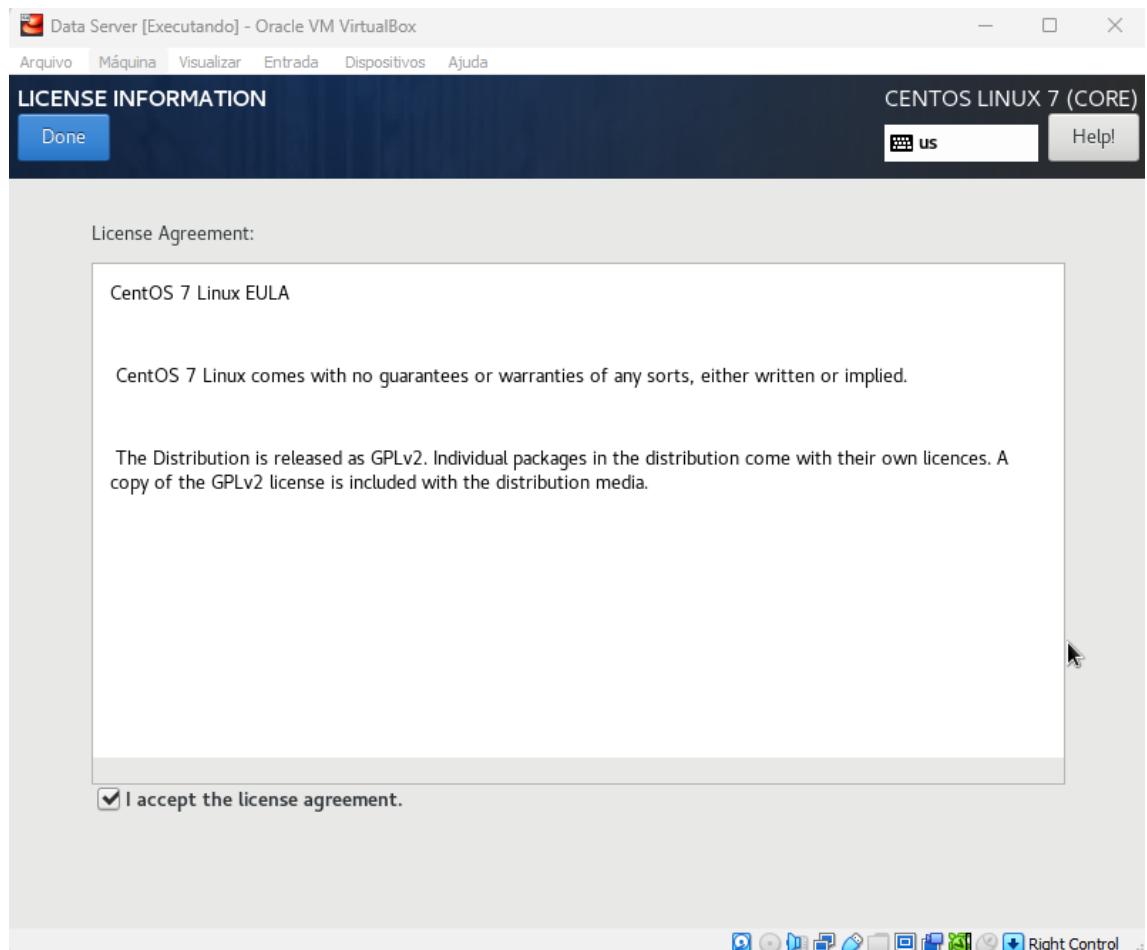
Conclusão da instalação. Clique no botão Reboot.

## Instalação e Configuração do Ambiente Big Data com Hadoop e Spark



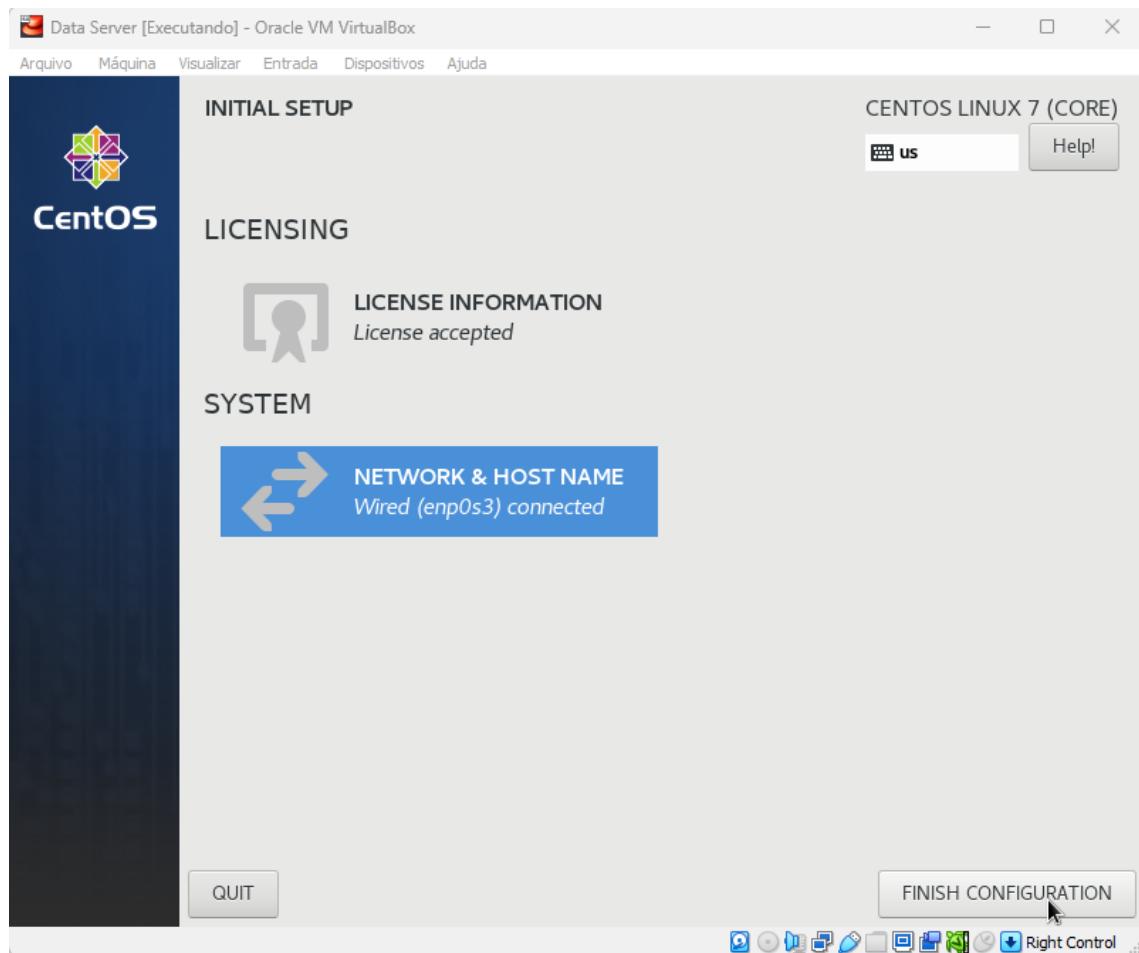
Clique em License Information para aceitar os termos de uso

## Instalação e Configuração do Ambiente Big Data com Hadoop e Spark



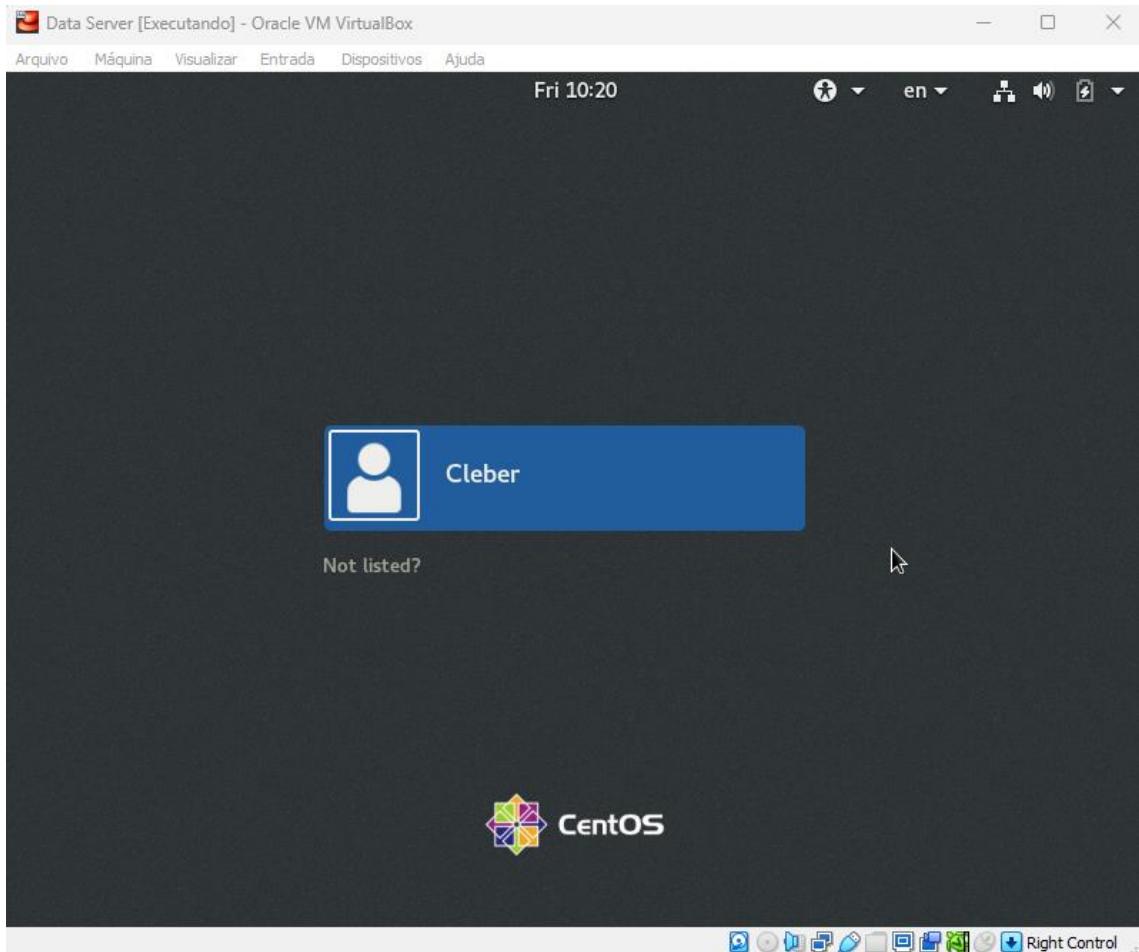
Marque a caixa e pressione Done.

## Instalação e Configuração do Ambiente Big Data com Hadoop e Spark



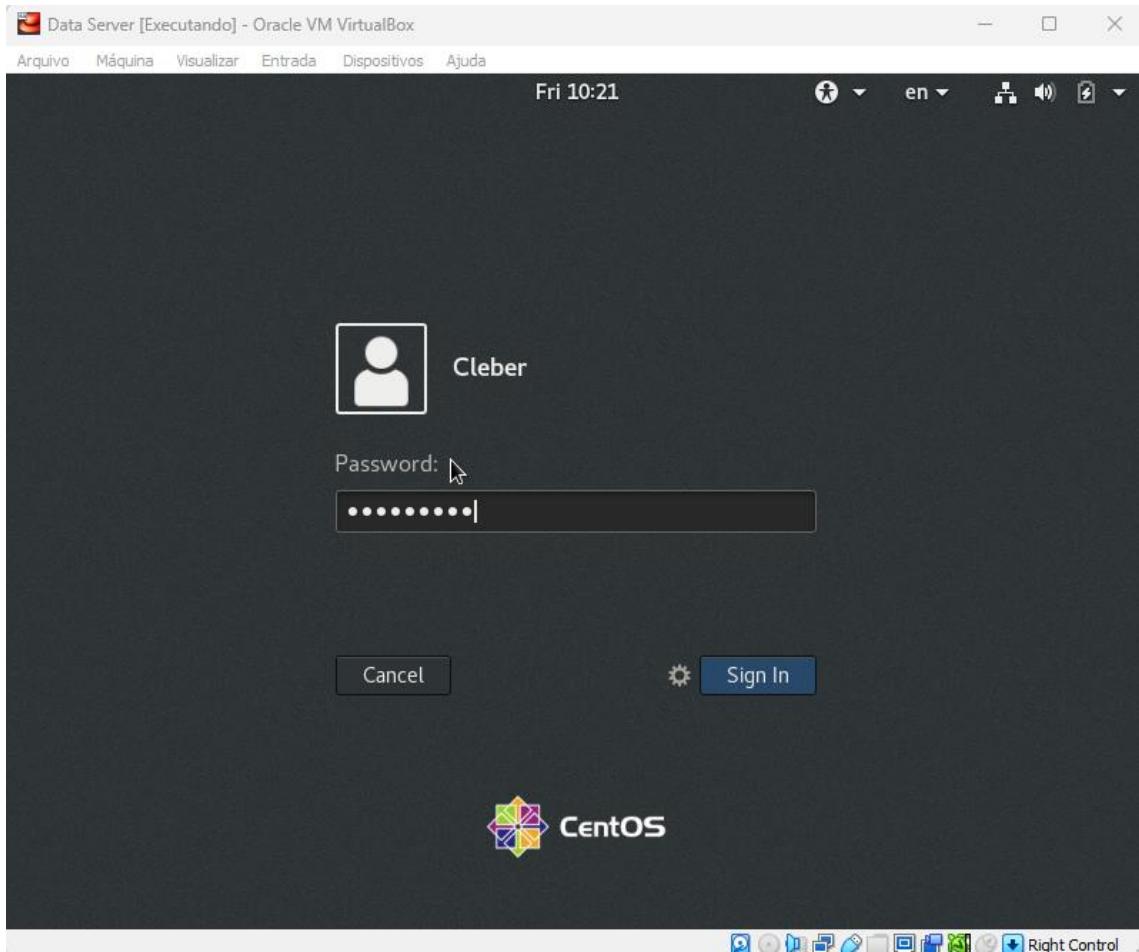
Se necessário, revise se o nome da máquina está correto e se a rede está ativada e pressione “Finish Configuration”

## Instalação e Configuração do Ambiente Big Data com Hadoop e Spark



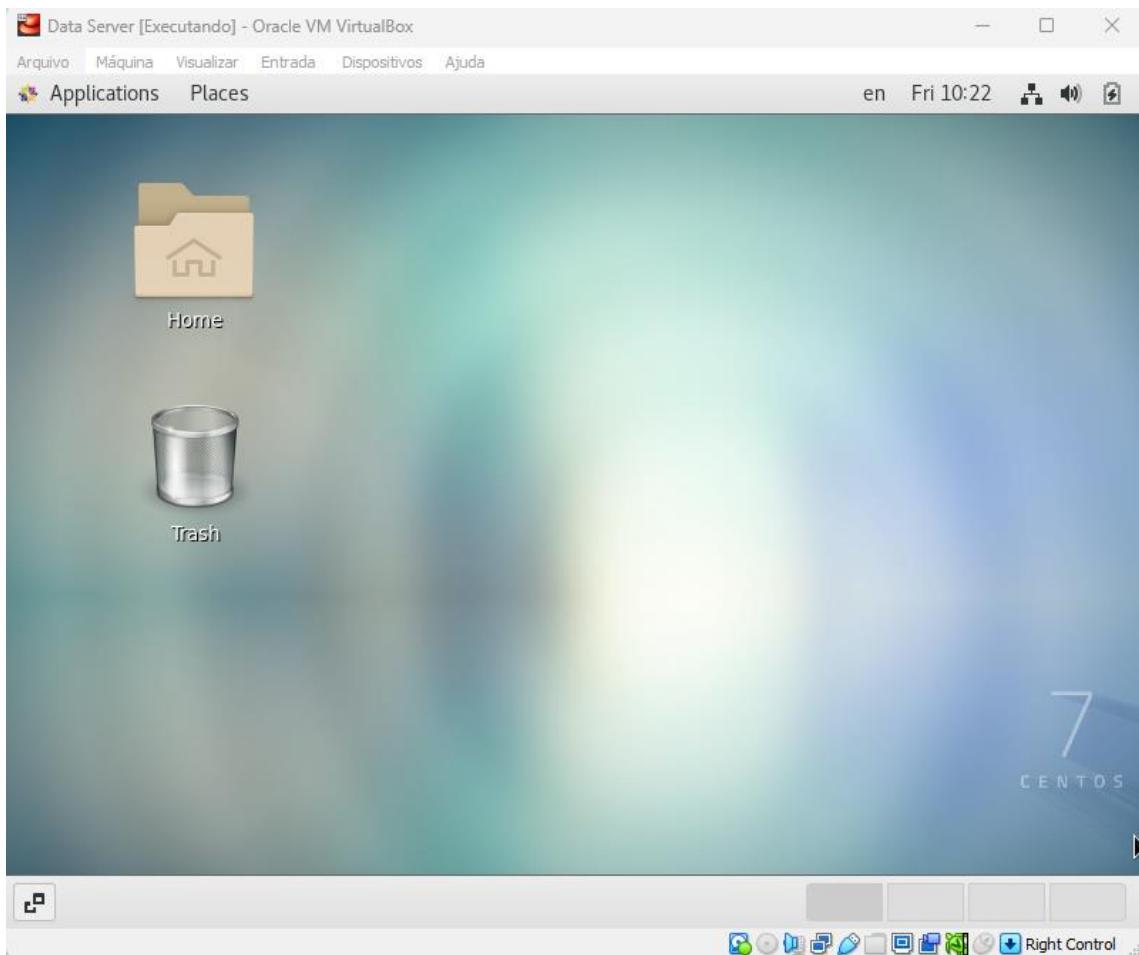
Clique no usuário Cleber

## Instalação e Configuração do Ambiente Big Data com Hadoop e Spark



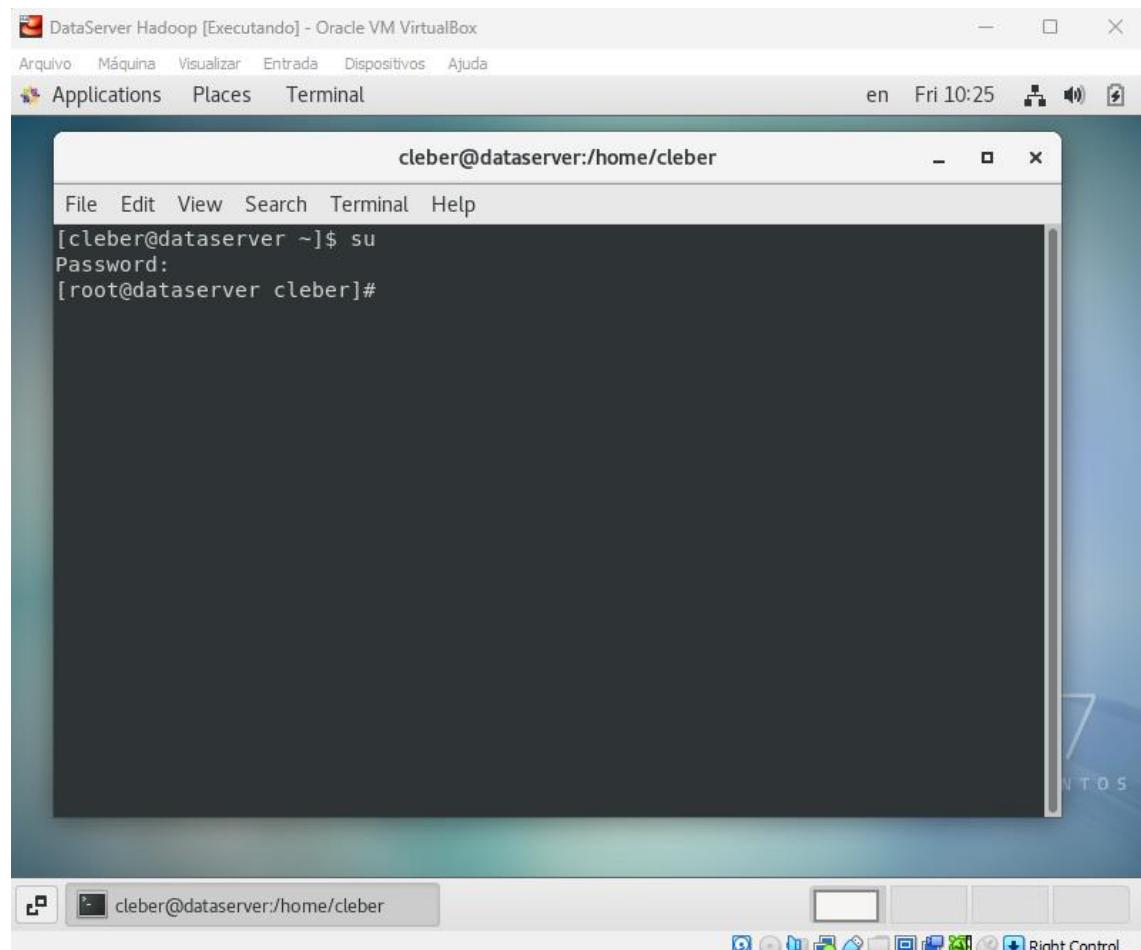
Usuário/Senha (**hadoop123**)

## Instalação e Configuração do Ambiente Big Data com Hadoop e Spark



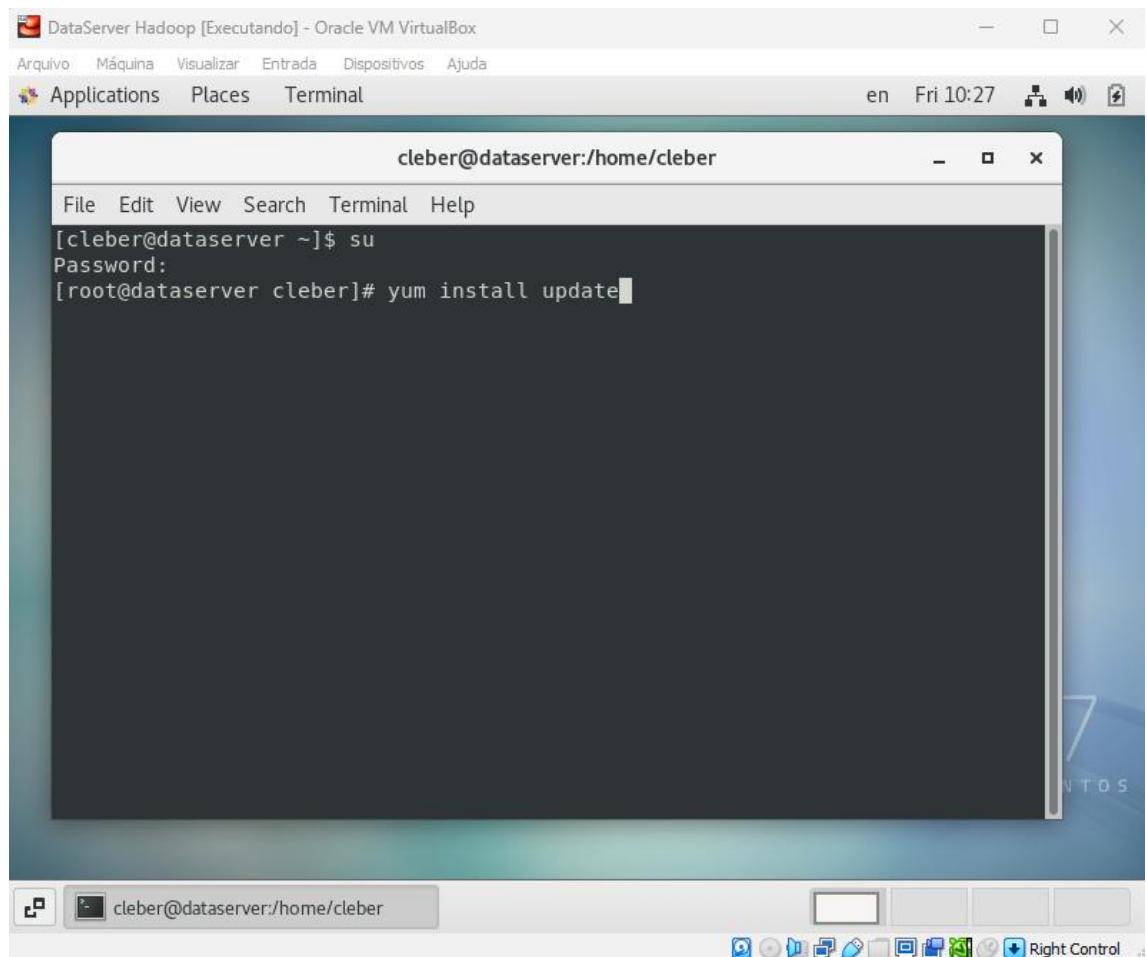
Instalação concluída com sucesso

## 2.3 Instalação de Utilitários do Sistema Operacional



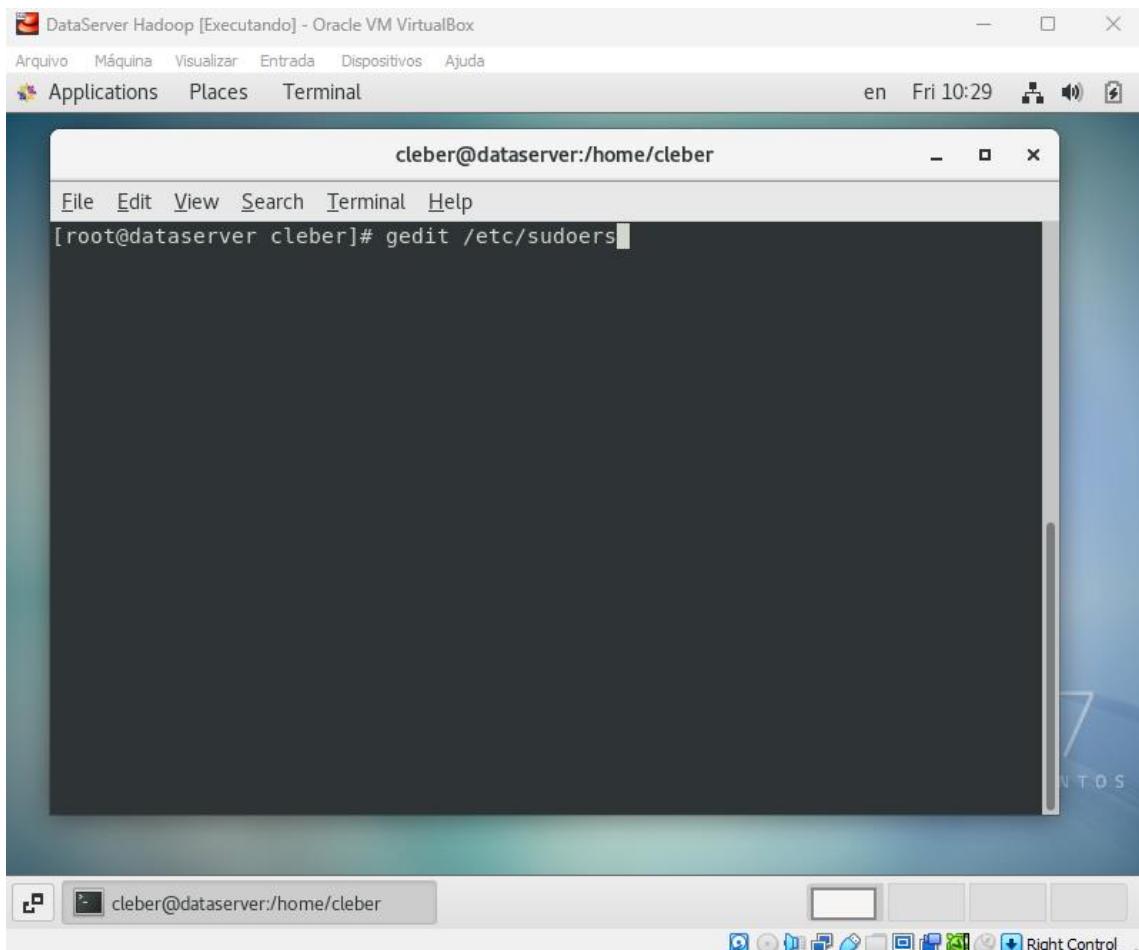
Efetuar login como root, usando o comando su. Senha: **hadoop123**

## Instalação e Configuração do Ambiente Big Data com Hadoop e Spark



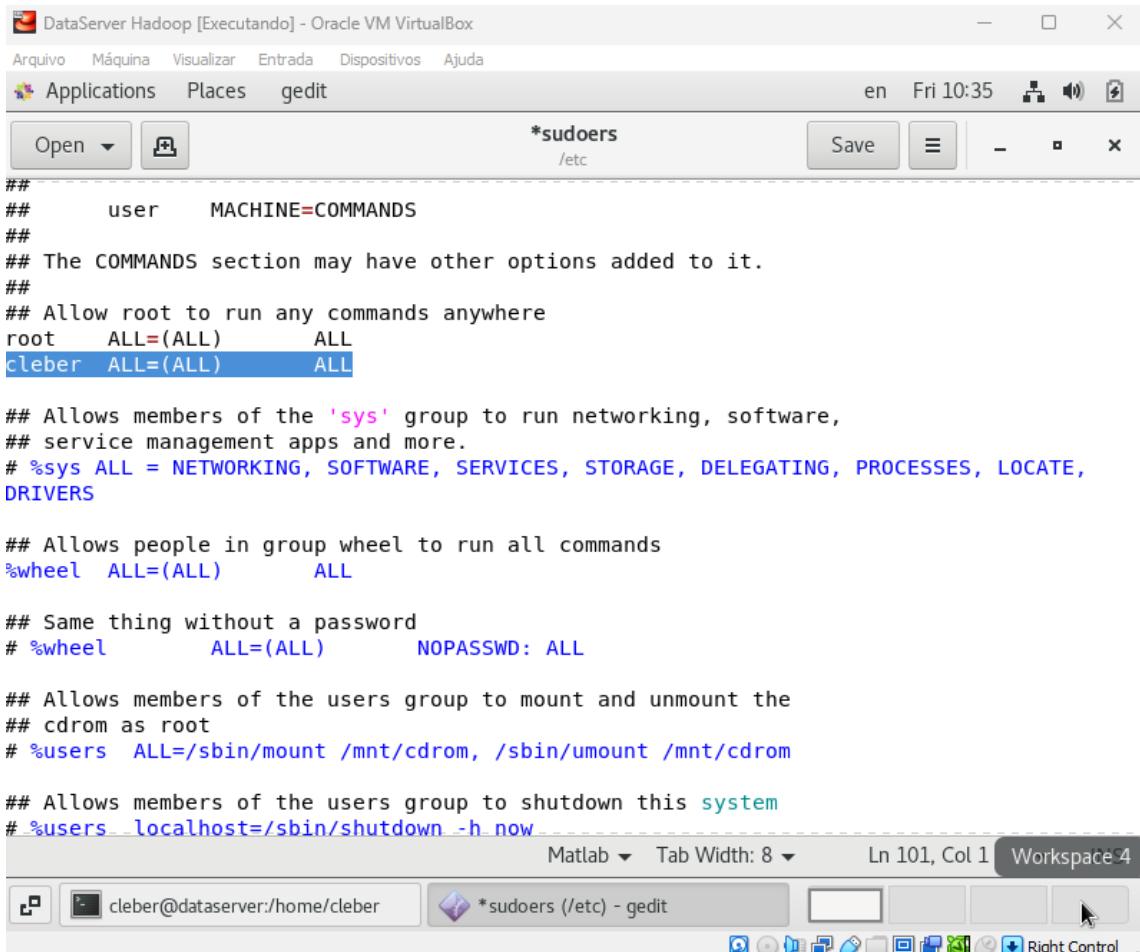
Atualizando o Sistema Operacional

## Instalação e Configuração do Ambiente Big Data com Hadoop e Spark



Editar o arquivo /etc/sudoers usando o gedit

## Instalação e Configuração do Ambiente Big Data com Hadoop e Spark



```
##      user      MACHINE=COMMANDS
##
## The COMMANDS section may have other options added to it.
##
## Allow root to run any commands anywhere
root    ALL=(ALL)        ALL
cleber  ALL=(ALL)        ALL

## Allows members of the 'sys' group to run networking, software,
## service management apps and more.
# %sys  ALL = NETWORKING, SOFTWARE, SERVICES, STORAGE, DELEGATING, PROCESSES, LOCATE,
DRIVERS

## Allows people in group wheel to run all commands
%wheel  ALL=(ALL)        ALL

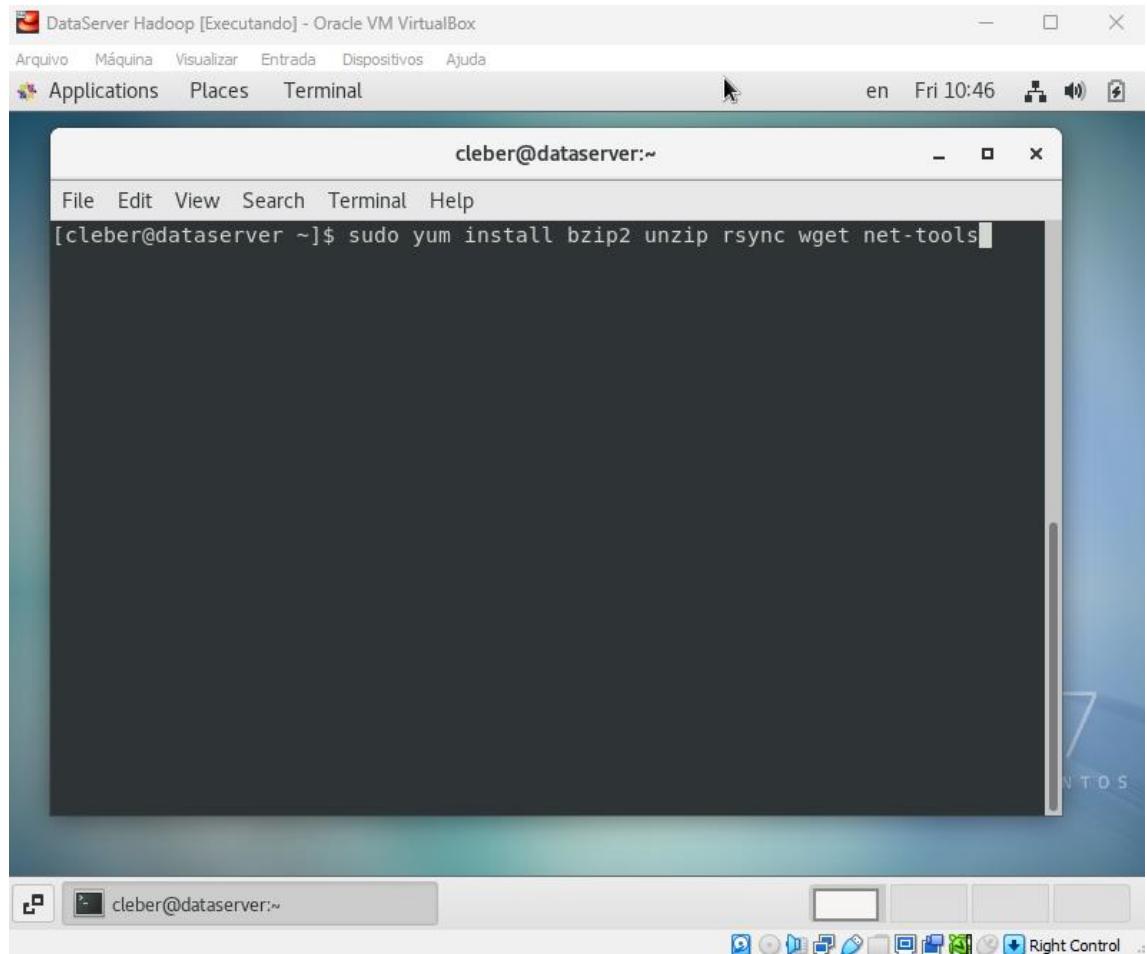
## Same thing without a password
# %wheel      ALL=(ALL)        NOPASSWD: ALL

## Allows members of the users group to mount and umount the
## cdrom as root
# %users  ALL=/sbin/mount /mnt/cdrom, /sbin/umount /mnt/cdrom

## Allows members of the users group to shutdown this system
# %users  localhost=/sbin/shutdown -h now
```

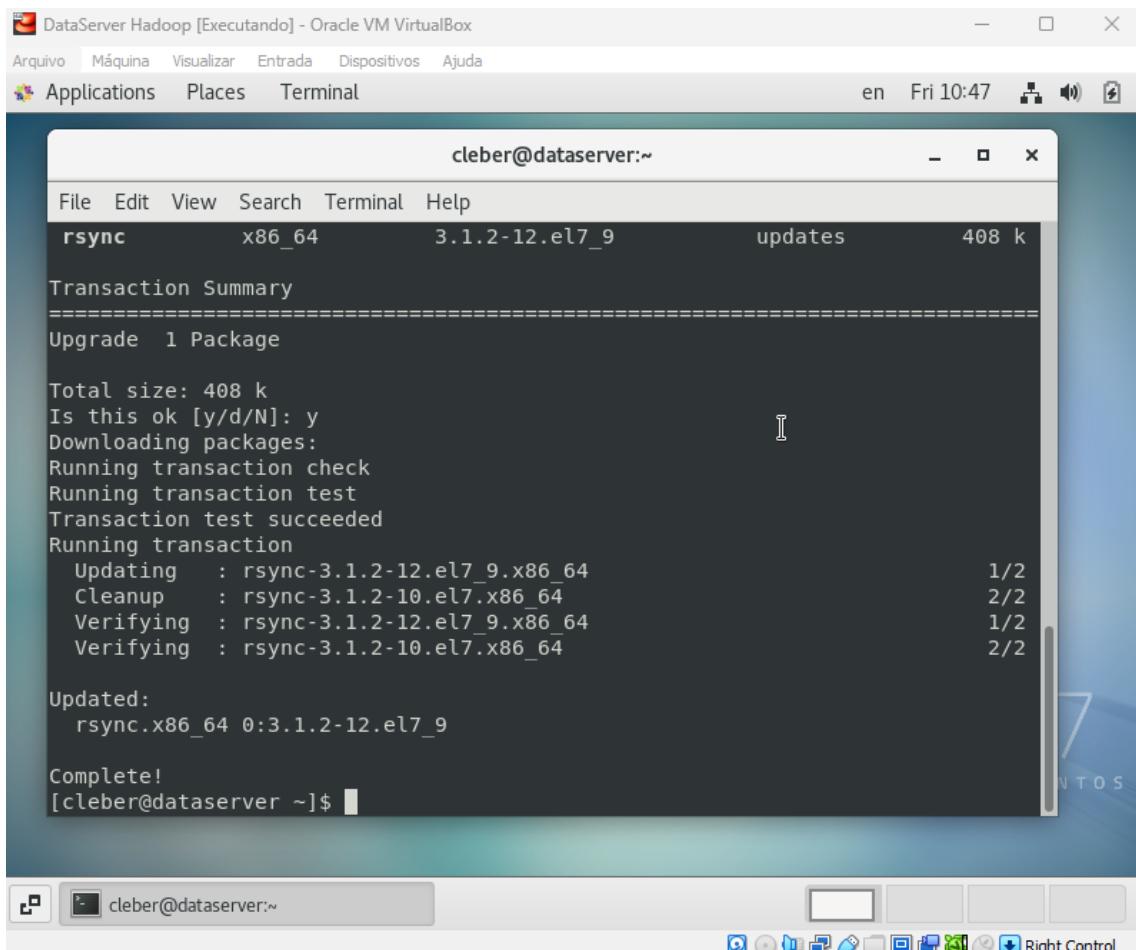
Incluir no arquivo a linha marcada acima e salvar o arquivo. Isso permitirá o usuário cleber executar comandos de administrador (root).

## Instalação e Configuração do Ambiente Big Data com Hadoop e Spark



Instalar outros aplicativos: bzip2, unzip, rsync, wget e net-tools

## Instalação e Configuração do Ambiente Big Data com Hadoop e Spark



A screenshot of a Linux desktop environment within an Oracle VM VirtualBox window. The desktop has a blue and white theme with a 'CENTOS' wallpaper. A terminal window titled 'cleber@dataserver:~' is open, showing the output of a 'yum update' command. The terminal window has a dark background with light-colored text. The desktop bar at the top includes icons for Applications, Places, Terminal, and system status (en Fri 10:47). The taskbar at the bottom shows several application icons.

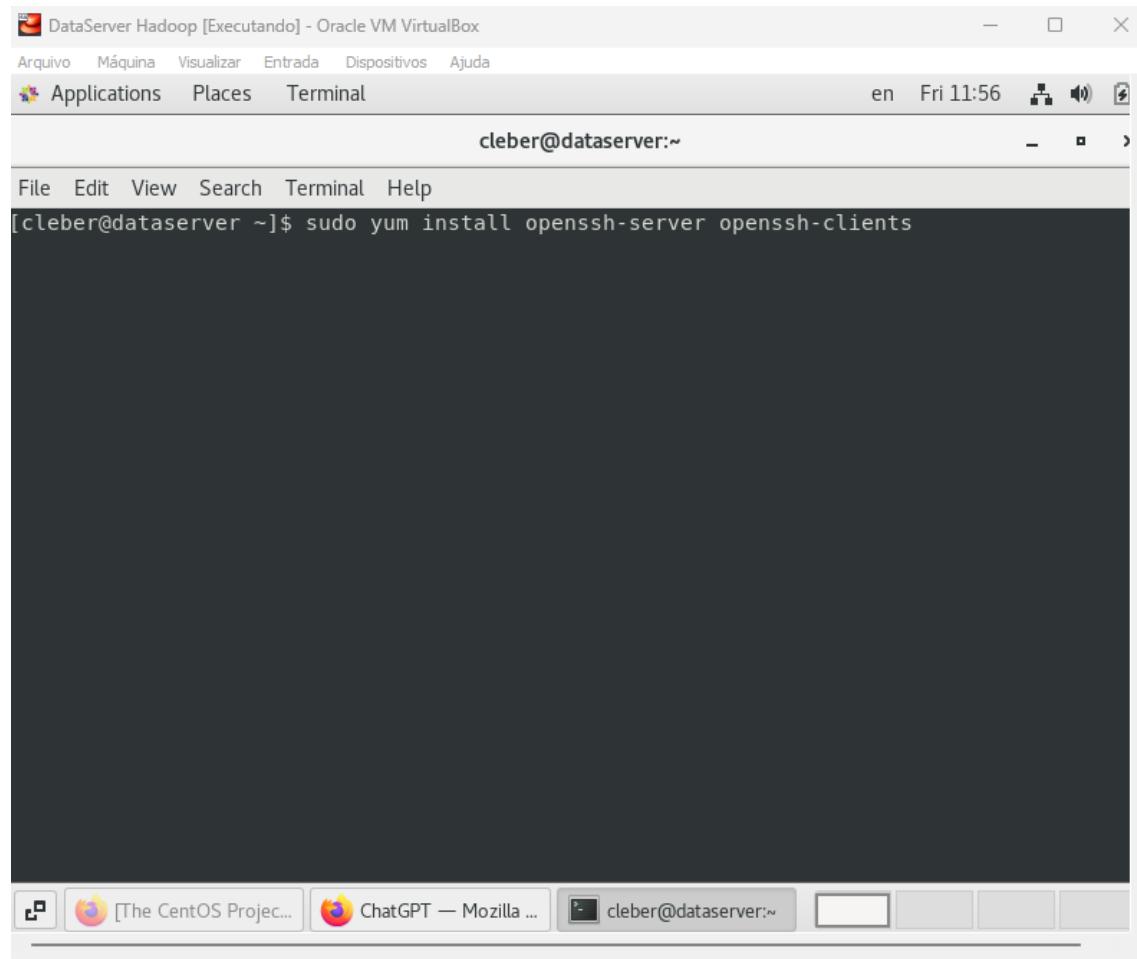
```
cleber@dataserver:~$ yum update
[sudo] password for cleber:
Loading repository information...
Resolving Dependencies...
--> Running transaction check
--> Updating   : rsync-3.1.2-12.el7_9.x86_64                                1/2
     Cleanup    : rsync-3.1.2-10.el7.x86_64                                    2/2
     Verifying  : rsync-3.1.2-12.el7_9.x86_64                                1/2
     Verifying  : rsync-3.1.2-10.el7.x86_64                                2/2
Total size: 408 k
Is this ok [y/d/N]: y
Downloading packages:
=====
Upgrading 1 Package

Total size: 408 k
Is this ok [y/d/N]: y
Downloading packages:
=====
Running transaction check
Running transaction test
Transaction test succeeded
Running transaction
  Updating   : rsync-3.1.2-12.el7_9.x86_64                                1/2
  Cleanup    : rsync-3.1.2-10.el7.x86_64                                    2/2
  Verifying  : rsync-3.1.2-12.el7_9.x86_64                                1/2
  Verifying  : rsync-3.1.2-10.el7.x86_64                                2/2
=====
Updated:
  rsync.x86_64 0:3.1.2-12.el7_9

Complete!
[cleber@dataserver ~]$
```

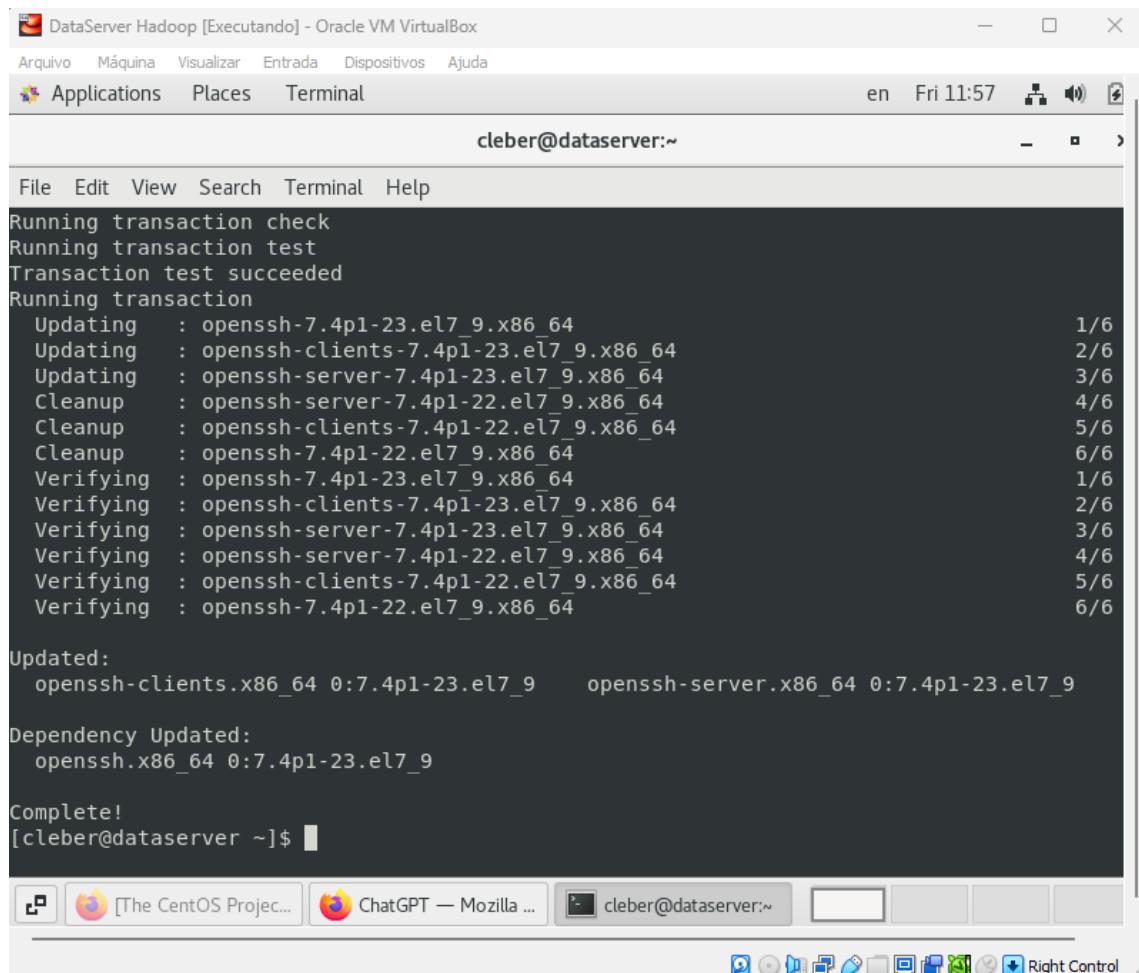
Aplicativos instalados

### 3 Instalação do servidor ssh



`sudo yum install openssh-server openssh-clients`

# Instalação e Configuração do Ambiente Big Data com Hadoop e Spark



```
cleber@dataserver:~$ apt update
Running transaction check
Running transaction test
Transaction test succeeded
Running transaction
Upgrading  : openssh-7.4p1-23.el7_9.x86_64 1/6
Upgrading  : openssh-clients-7.4p1-23.el7_9.x86_64 2/6
Upgrading  : openssh-server-7.4p1-23.el7_9.x86_64 3/6
Cleanup    : openssh-server-7.4p1-22.el7_9.x86_64 4/6
Cleanup    : openssh-clients-7.4p1-22.el7_9.x86_64 5/6
Cleanup    : openssh-7.4p1-22.el7_9.x86_64 6/6
Verifying   : openssh-7.4p1-23.el7_9.x86_64 1/6
Verifying   : openssh-clients-7.4p1-23.el7_9.x86_64 2/6
Verifying   : openssh-server-7.4p1-23.el7_9.x86_64 3/6
Verifying   : openssh-server-7.4p1-22.el7_9.x86_64 4/6
Verifying   : openssh-clients-7.4p1-22.el7_9.x86_64 5/6
Verifying   : openssh-7.4p1-22.el7_9.x86_64 6/6

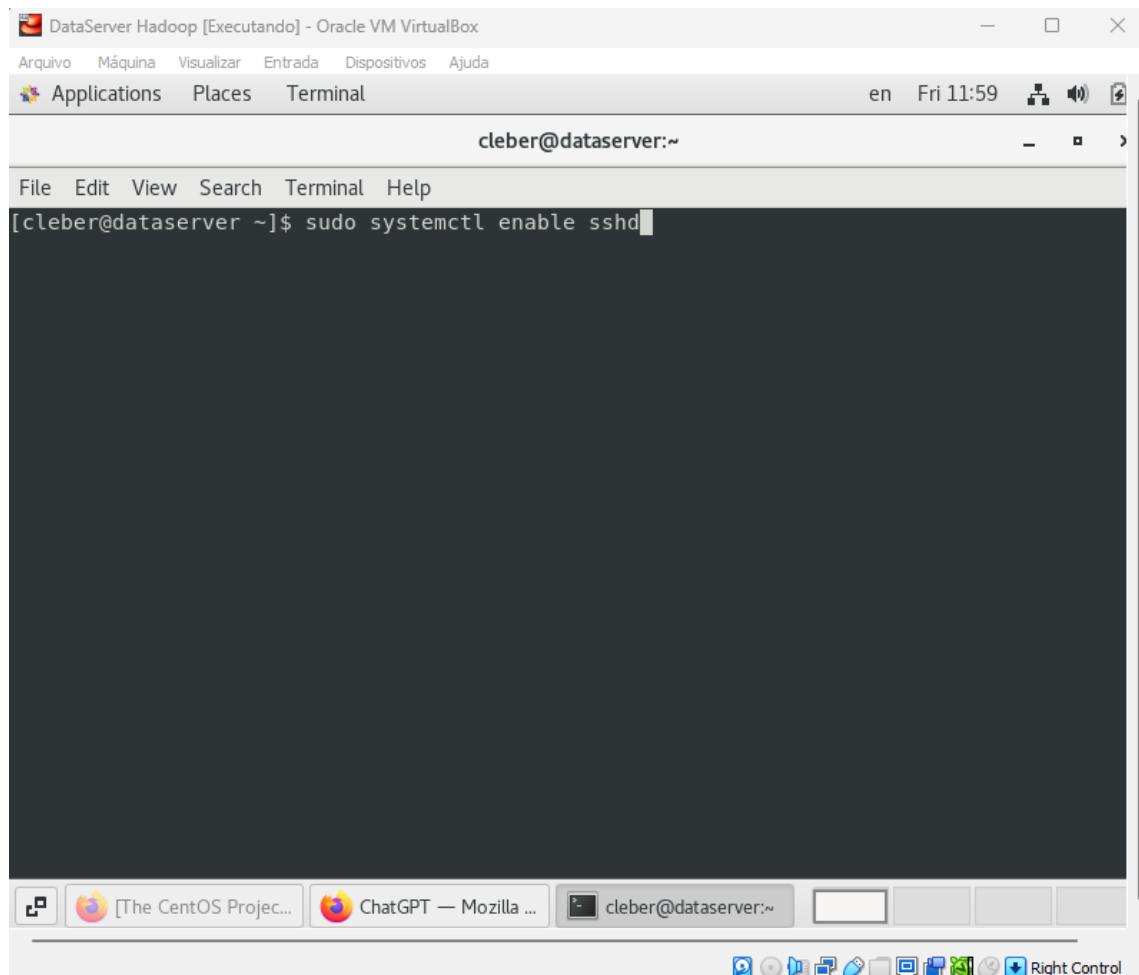
Updated:
  openssh-clients.x86_64 0:7.4p1-23.el7_9      openssh-server.x86_64 0:7.4p1-23.el7_9

Dependency Updated:
  openssh.x86_64 0:7.4p1-23.el7_9

Complete!
[cleber@dataserver ~]$
```

Concluído

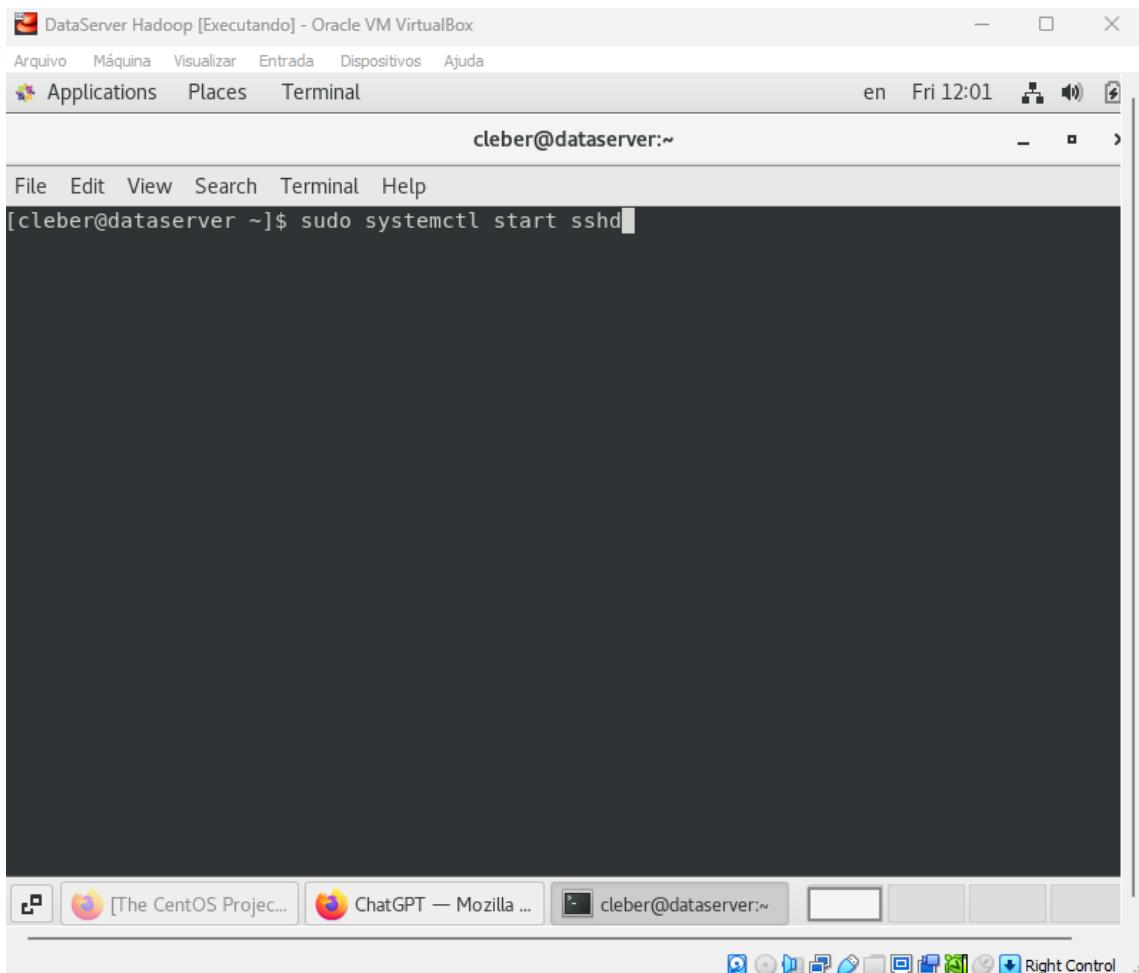
## Instalação e Configuração do Ambiente Big Data com Hadoop e Spark



Habilitando o serviço

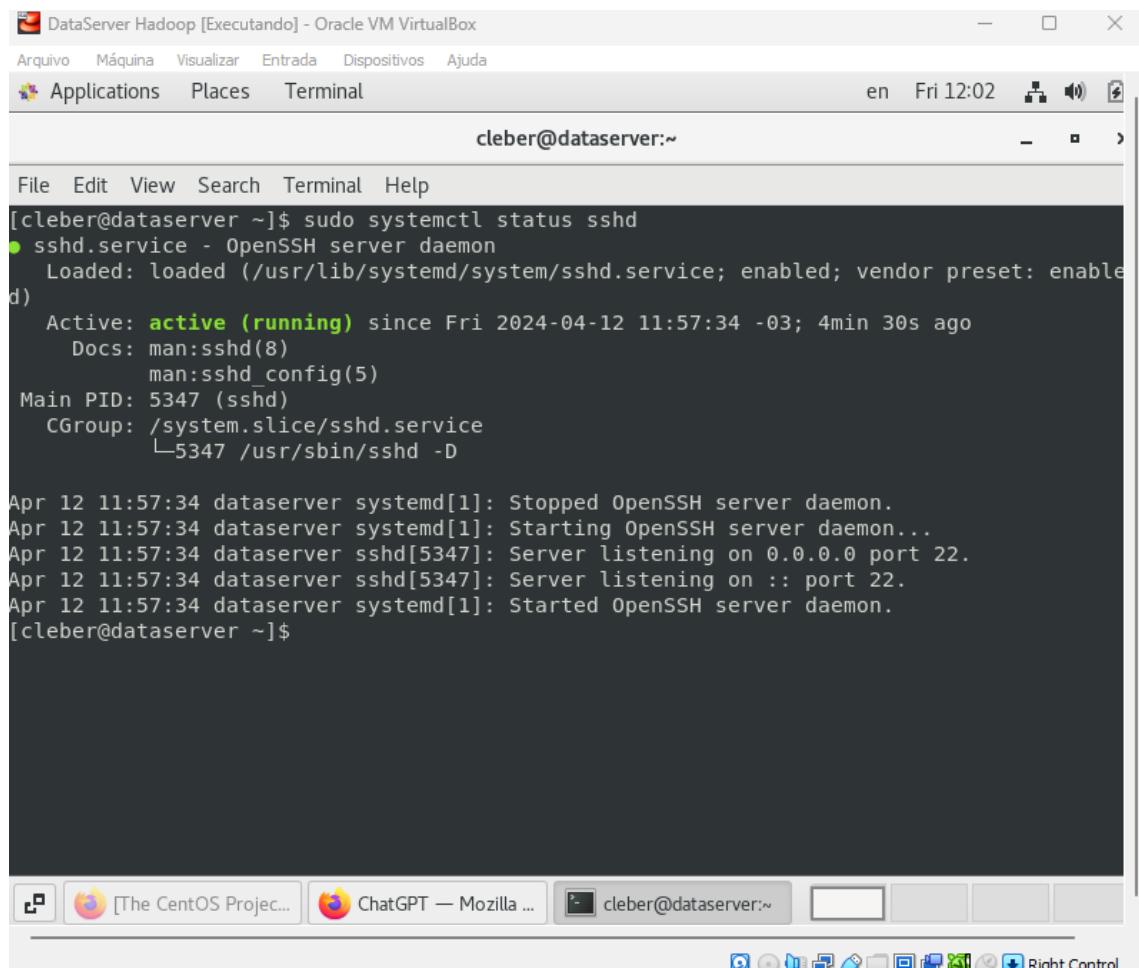
`sudo systemctl enable sshd`

## Instalação e Configuração do Ambiente Big Data com Hadoop e Spark



Iniciando o serviço  
sudo systemctl start sshd

## Instalação e Configuração do Ambiente Big Data com Hadoop e Spark



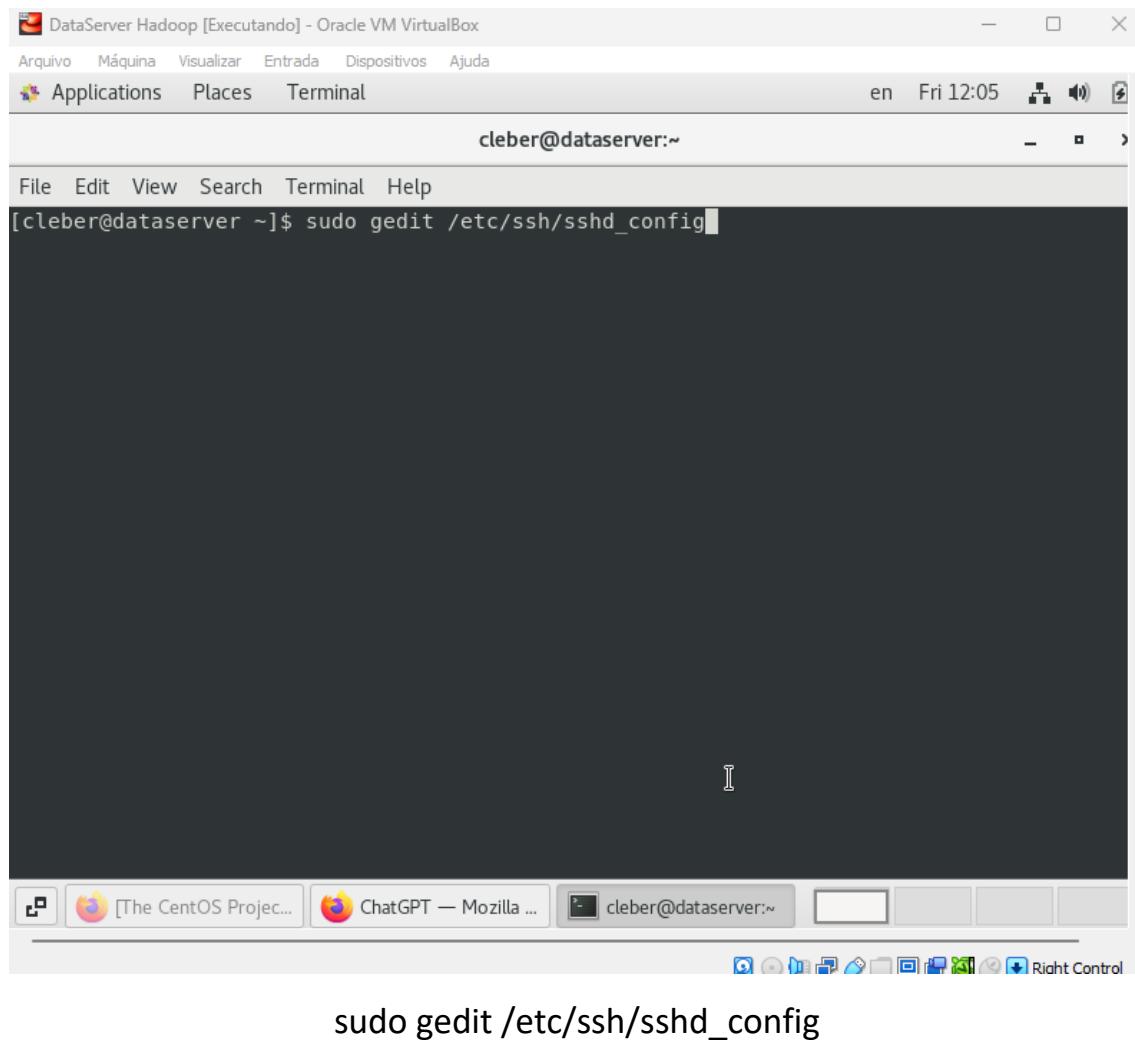
A screenshot of a Linux desktop environment, likely CentOS, showing a terminal window titled "cleber@dataserver:~". The terminal displays the output of the command "sudo systemctl status sshd". The output shows that the sshd service is active and running. Below this, several log entries from the system log (syslog) are shown, indicating the start and stop of the sshd service at 11:57:34 on April 12, 2024.

```
[cleber@dataserver ~]$ sudo systemctl status sshd
● sshd.service - OpenSSH server daemon
   Loaded: loaded (/usr/lib/systemd/system/sshd.service; enabled; vendor preset: enabled)
   Active: active (running) since Fri 2024-04-12 11:57:34 -03; 4min 30s ago
     Docs: man:sshd(8)
           man:sshd_config(5)
   Main PID: 5347 (sshd)
      CGroup: /system.slice/sshd.service
              └─5347 /usr/sbin/sshd -D

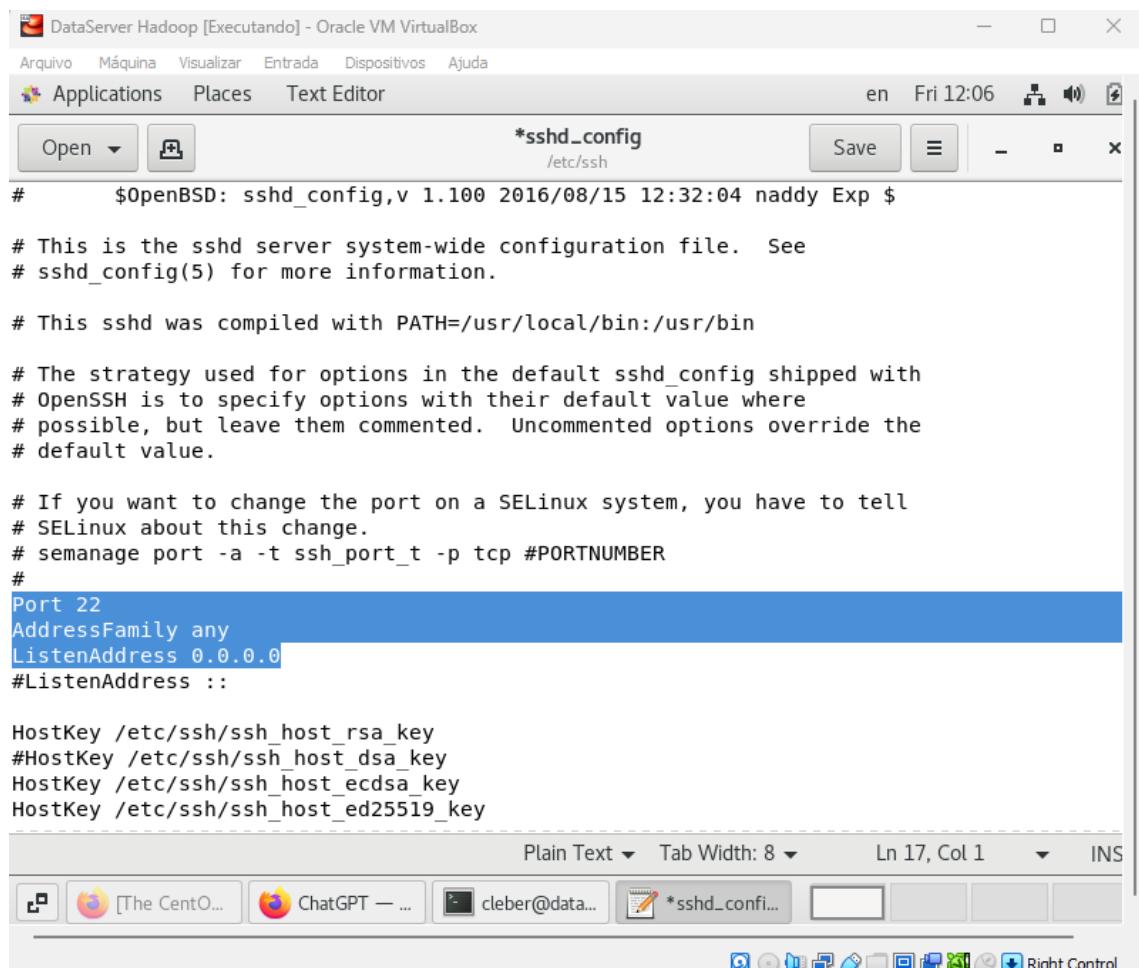
Apr 12 11:57:34 dataserver systemd[1]: Stopped OpenSSH server daemon.
Apr 12 11:57:34 dataserver systemd[1]: Starting OpenSSH server daemon...
Apr 12 11:57:34 dataserver sshd[5347]: Server listening on 0.0.0.0 port 22.
Apr 12 11:57:34 dataserver sshd[5347]: Server listening on :: port 22.
Apr 12 11:57:34 dataserver systemd[1]: Started OpenSSH server daemon.
[cleber@dataserver ~]$
```

Serviço em execução

## Instalação e Configuração do Ambiente Big Data com Hadoop e Spark



## Instalação e Configuração do Ambiente Big Data com Hadoop e Spark

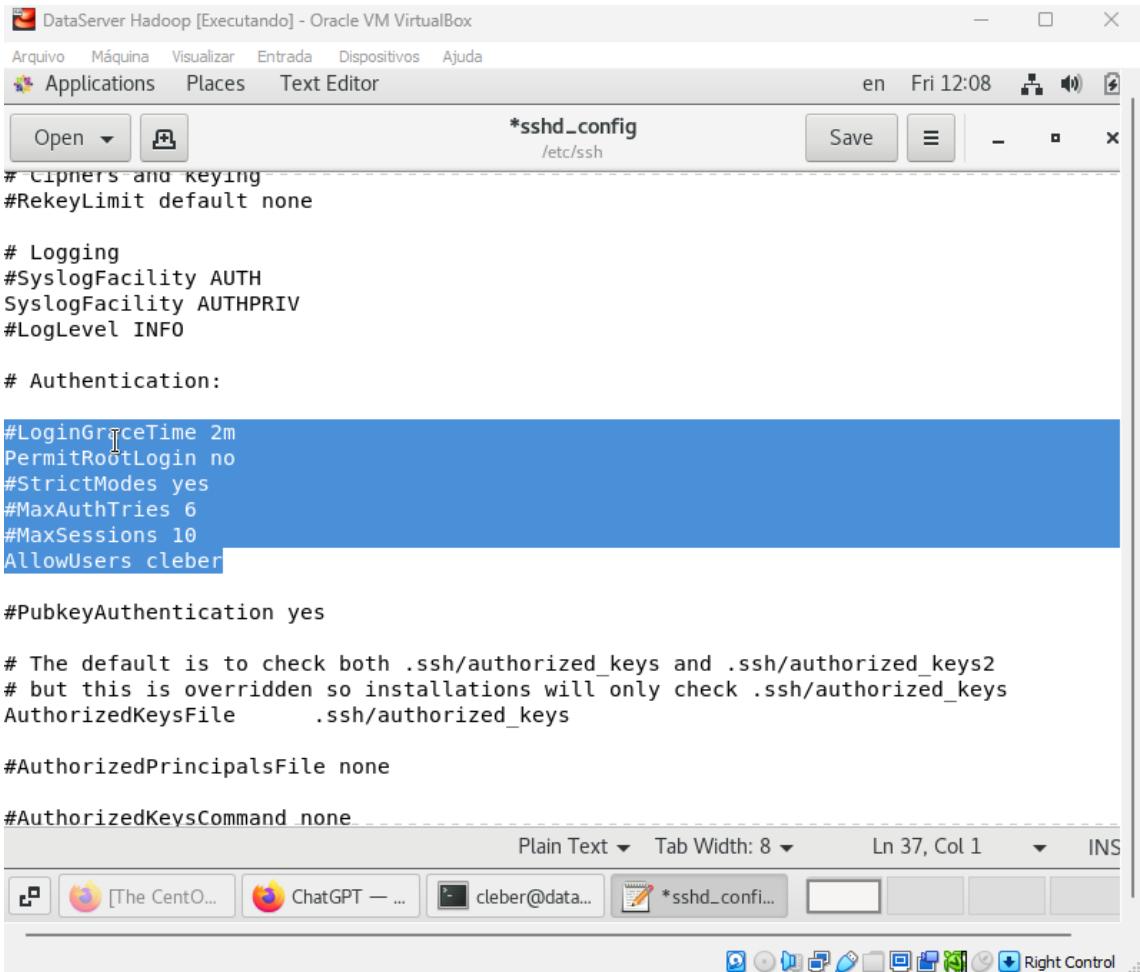


```
*sshd_config  
/etc/ssh  
# $OpenBSD: sshd_config,v 1.100 2016/08/15 12:32:04 naddy Exp $  
  
# This is the sshd server system-wide configuration file. See  
# sshd_config(5) for more information.  
  
# This sshd was compiled with PATH=/usr/local/bin:/usr/bin  
  
# The strategy used for options in the default sshd_config shipped with  
# OpenSSH is to specify options with their default value where  
# possible, but leave them commented. Uncommented options override the  
# default value.  
  
# If you want to change the port on a SELinux system, you have to tell  
# SELinux about this change.  
# semanage port -a -t ssh_port_t -p tcp #PORTNUMBER  
#  
Port 22  
AddressFamily any  
ListenAddress 0.0.0.0  
#ListenAddress ::  
  
HostKey /etc/ssh/ssh_host_rsa_key  
#HostKey /etc/ssh/ssh_host_dsa_key  
HostKey /etc/ssh/ssh_host_ecdsa_key  
HostKey /etc/ssh/ssh_host_ed25519_key
```

Primeira parte da configuração ssh.

Remova o símbolo (#) de comentário das 3 linhas marcadas acima.

## Instalação e Configuração do Ambiente Big Data com Hadoop e Spark



The screenshot shows a terminal window titled "DataServer Hadoop [Executando] - Oracle VM VirtualBox". The window contains the configuration file for the SSH daemon, specifically the "/etc/ssh/sshd\_config" file. The configuration includes sections for ciphers and keying, logging, authentication (allowing root login, specifying LogLevel as INFO, and permitting public key authentication), and authorized keys. The file ends with a note about checking both .ssh/authorized\_keys and .ssh/authorized\_keys2. The terminal interface includes standard Linux navigation keys at the bottom.

```
*sshd_config
/etc/ssh
#Ciphers and Keying
#RekeyLimit default none

# Logging
#SyslogFacility AUTH
SyslogFacility AUTHPRIV
#LogLevel INFO

# Authentication:
#LoginGraceTime 2m
PermitRootLogin no
#StrictModes yes
#MaxAuthTries 6
#MaxSessions 10
AllowUsers cleber

#PubkeyAuthentication yes

# The default is to check both .ssh/authorized_keys and .ssh/authorized_keys2
# but this is overridden so installations will only check .ssh/authorized_keys
AuthorizedKeysFile      .ssh/authorized_keys

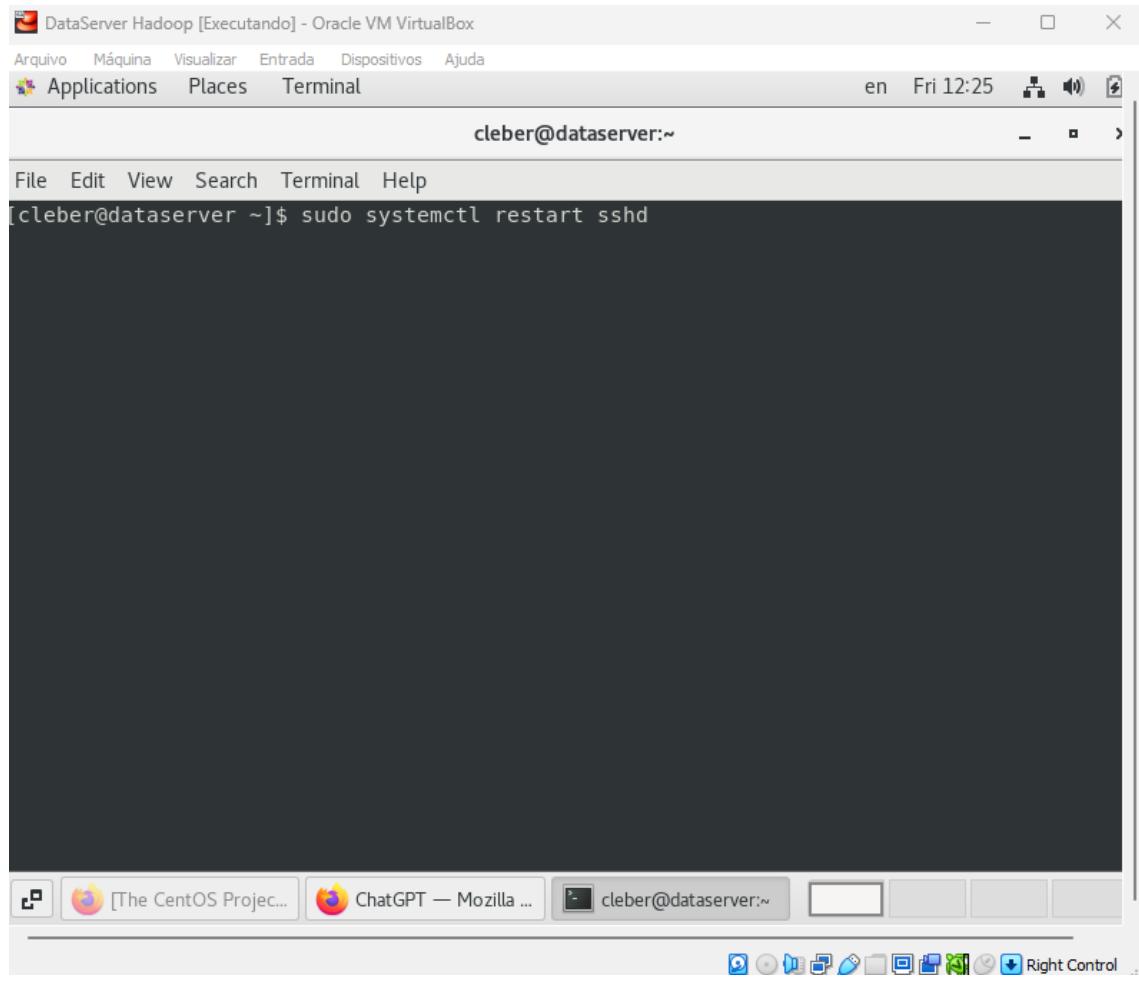
#AuthorizedPrincipalsFile none

#AuthorizedKeysCommand none
```

Segunda parte da configuração do ssh

## Instalação e Configuração do Ambiente Big Data com Hadoop e Spark

---



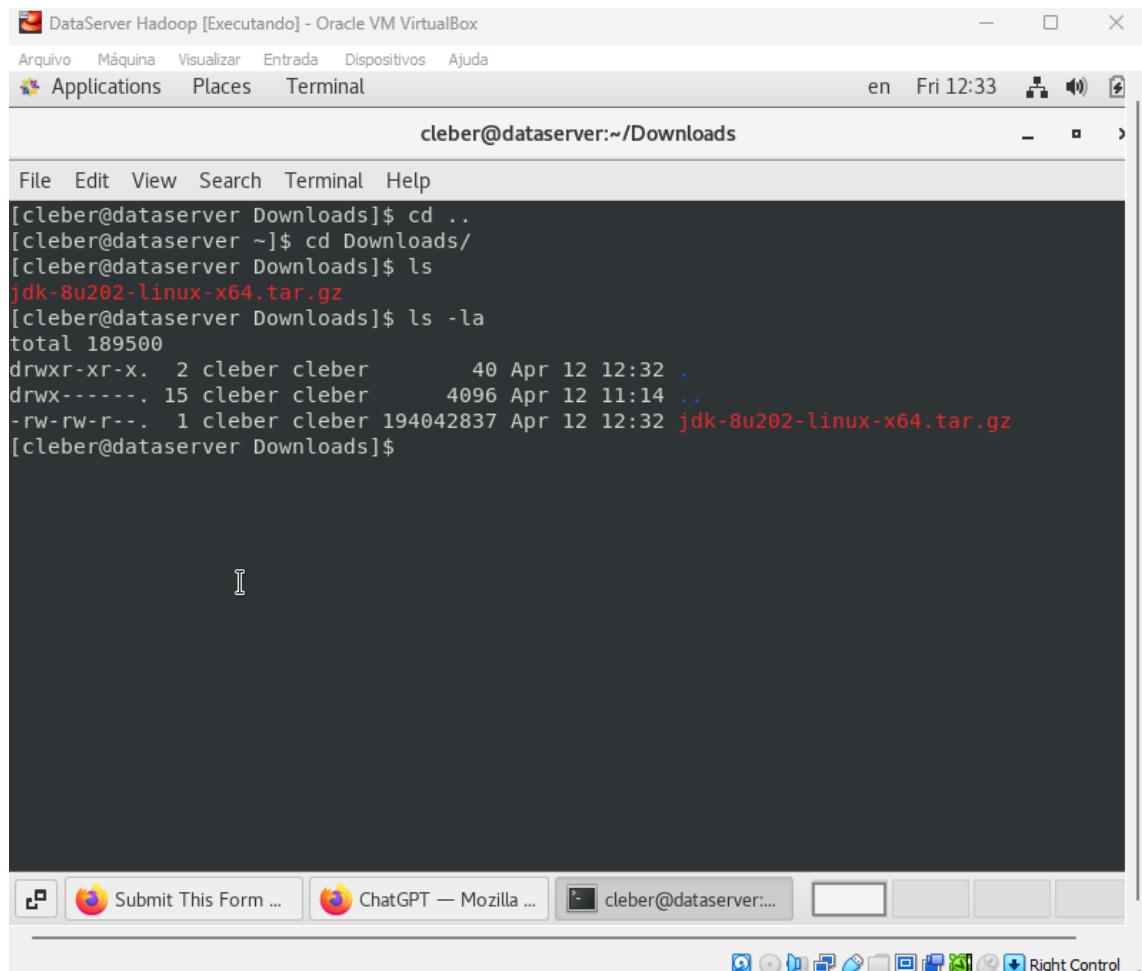
A screenshot of a Linux desktop environment, likely CentOS, running in Oracle VM VirtualBox. The terminal window shows the command `sudo systemctl restart sshd` being run by the user `cleber`. The taskbar at the bottom shows several open applications, including a browser tab for 'The CentOS Project' and another for 'ChatGPT — Mozilla ...'. The desktop background is a standard blue gradient.

```
cleber@dataserver:~$ sudo systemctl restart sshd
```

`sudo systemctl restart sshd`

## 4 Instalação do Java 8

Acesse o site da Oracle e faça download do Java JDK 1.8 para Linux

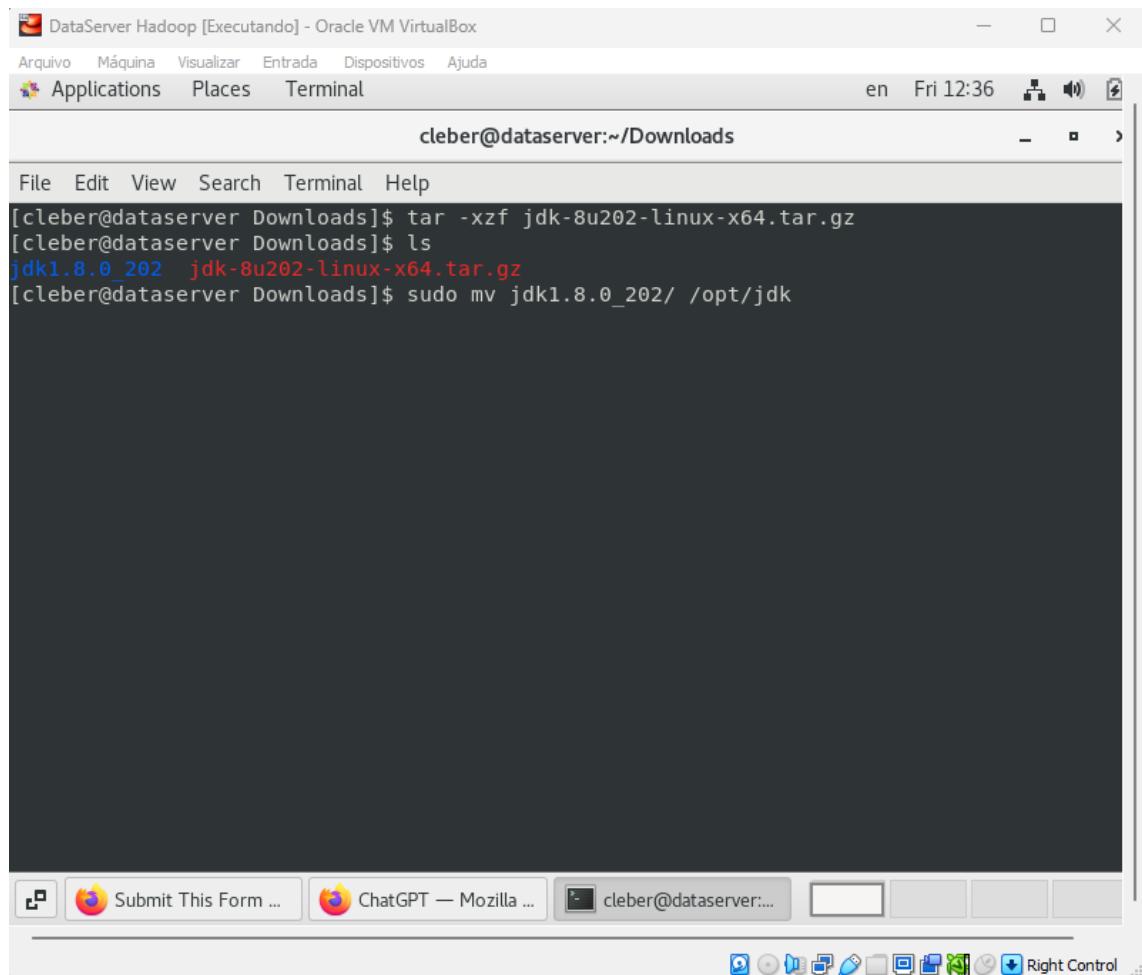


The screenshot shows a terminal window titled "DataServer Hadoop [Executando] - Oracle VM VirtualBox". The window includes a menu bar with "Arquivo", "Máquina", "Visualizar", "Entrada", "Dispositivos", and "Ajuda". Below the menu is a toolbar with "Applications", "Places", and "Terminal". The status bar at the bottom right shows "en Fri 12:33". The terminal session shows the user navigating to the Downloads directory and listing its contents. A file named "jdk-8u202-linux-x64.tar.gz" is visible. The terminal prompt is "[cleber@dataserver Downloads]\$".

```
[cleber@dataserver Downloads]$ cd ..
[cleber@dataserver ~]$ cd Downloads/
[cleber@dataserver Downloads]$ ls
[jdk-8u202-linux-x64.tar.gz]
[cleber@dataserver Downloads]$ ls -la
total 189500
drwxr-xr-x. 2 cleber cleber      40 Apr 12 12:32 .
drwx----- 15 cleber cleber    4096 Apr 12 11:14 ..
-rw-rw-r--.  1 cleber cleber 194042837 Apr 12 12:32 jdk-8u202-linux-x64.tar.gz
[cleber@dataserver Downloads]$
```

Executar o comando tar para descompactar o arquivo: **tar -xzf jdk-8u212-linux-x64.tar.gz**

## Instalação e Configuração do Ambiente Big Data com Hadoop e Spark



The screenshot shows a terminal window titled "DataServer Hadoop [Executando] - Oracle VM VirtualBox". The window contains the following terminal session:

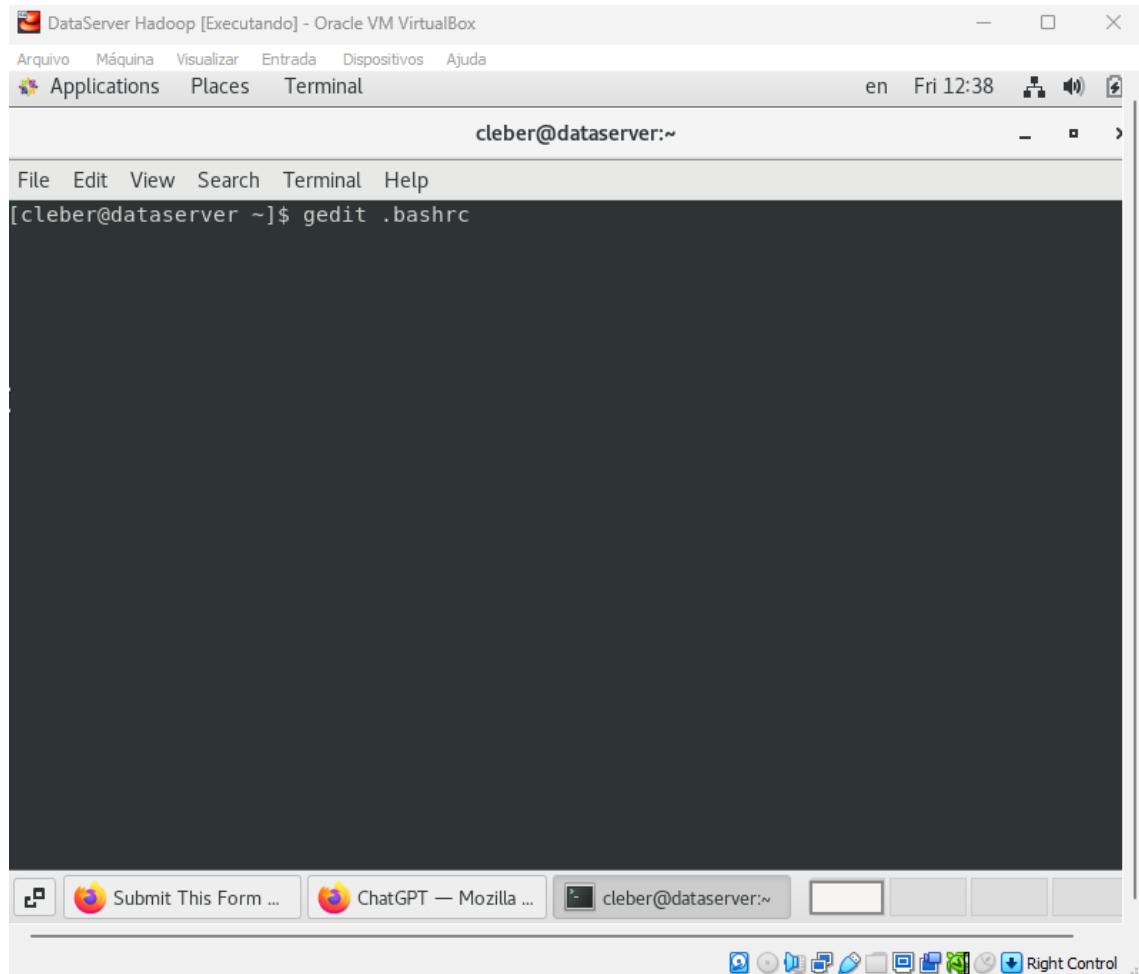
```
[cleber@dataserver Downloads]$ tar -xzf jdk-8u202-linux-x64.tar.gz
[cleber@dataserver Downloads]$ ls
jdk1.8.0_202  jdk-8u202-linux-x64.tar.gz
[cleber@dataserver Downloads]$ sudo mv jdk1.8.0_202/ /opt/jdk
```

The terminal window has a menu bar with "Arquivo", "Máquina", "Visualizar", "Entrada", "Dispositivos", and "Ajuda". It also has tabs for "Applications", "Places", and "Terminal". The status bar at the bottom shows "en Fri 12:36". Below the terminal window, the desktop environment is visible with icons for "Submit This Form ...", "ChatGPT — Mozilla ...", and "cleber@dataserver:...". A dock at the bottom contains various application icons.

Mover o diretório do JDK

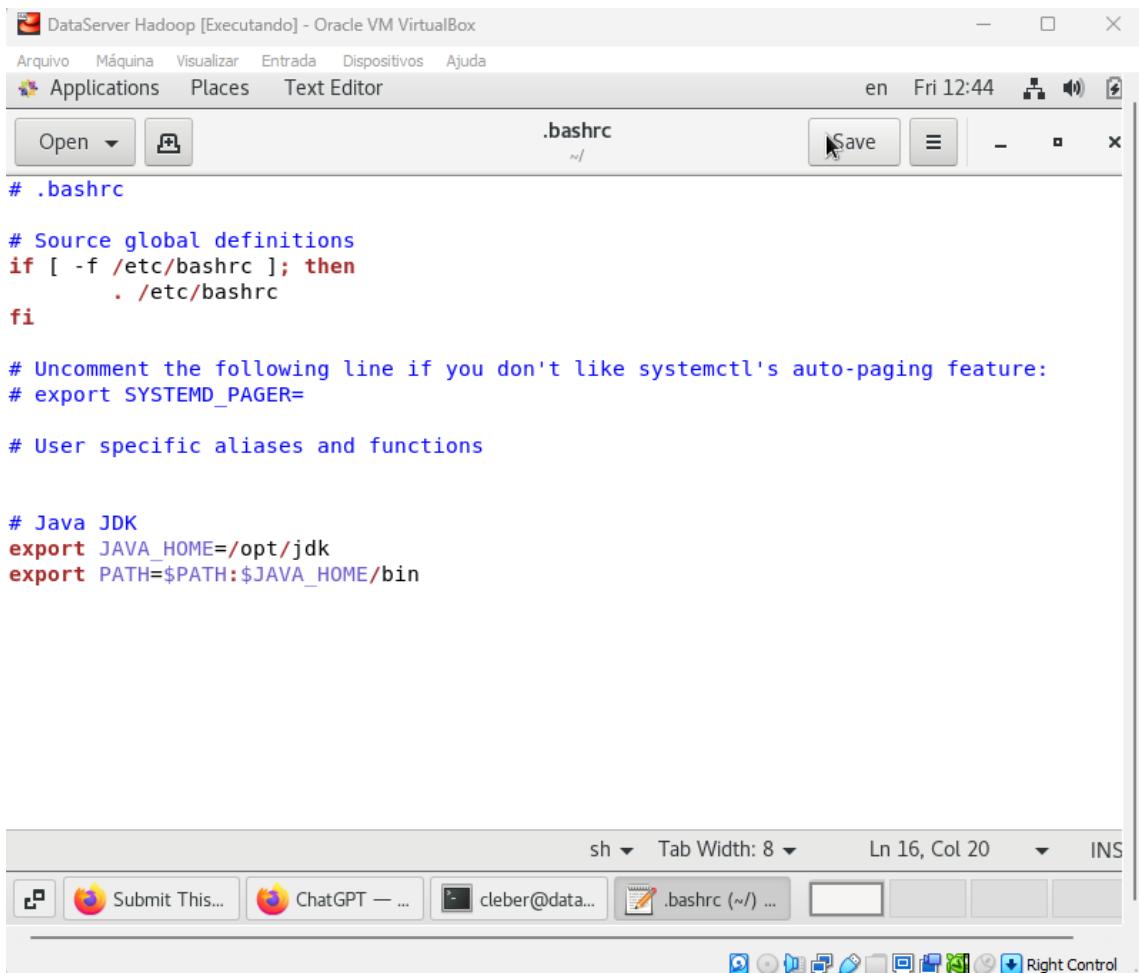
## Instalação e Configuração do Ambiente Big Data com Hadoop e Spark

---



gedit .bashrc

## Instalação e Configuração do Ambiente Big Data com Hadoop e Spark



The screenshot shows a Linux desktop environment with a terminal window and a text editor window.

The terminal window at the bottom has the following status bar text:

- sh ▾ Tab Width: 8 ▾ Ln 16, Col 20 ▾ INS
- Submit This... ChatGPT — ... cleber@data... .bashrc (~) ...
- Right Control

The text editor window at the top displays the contents of the `.bashrc` file:

```
# .bashrc

# Source global definitions
if [ -f /etc/bashrc ]; then
    . /etc/bashrc

# Uncomment the following line if you don't like systemctl's auto-paging feature:
# export SYSTEMD_PAGER=

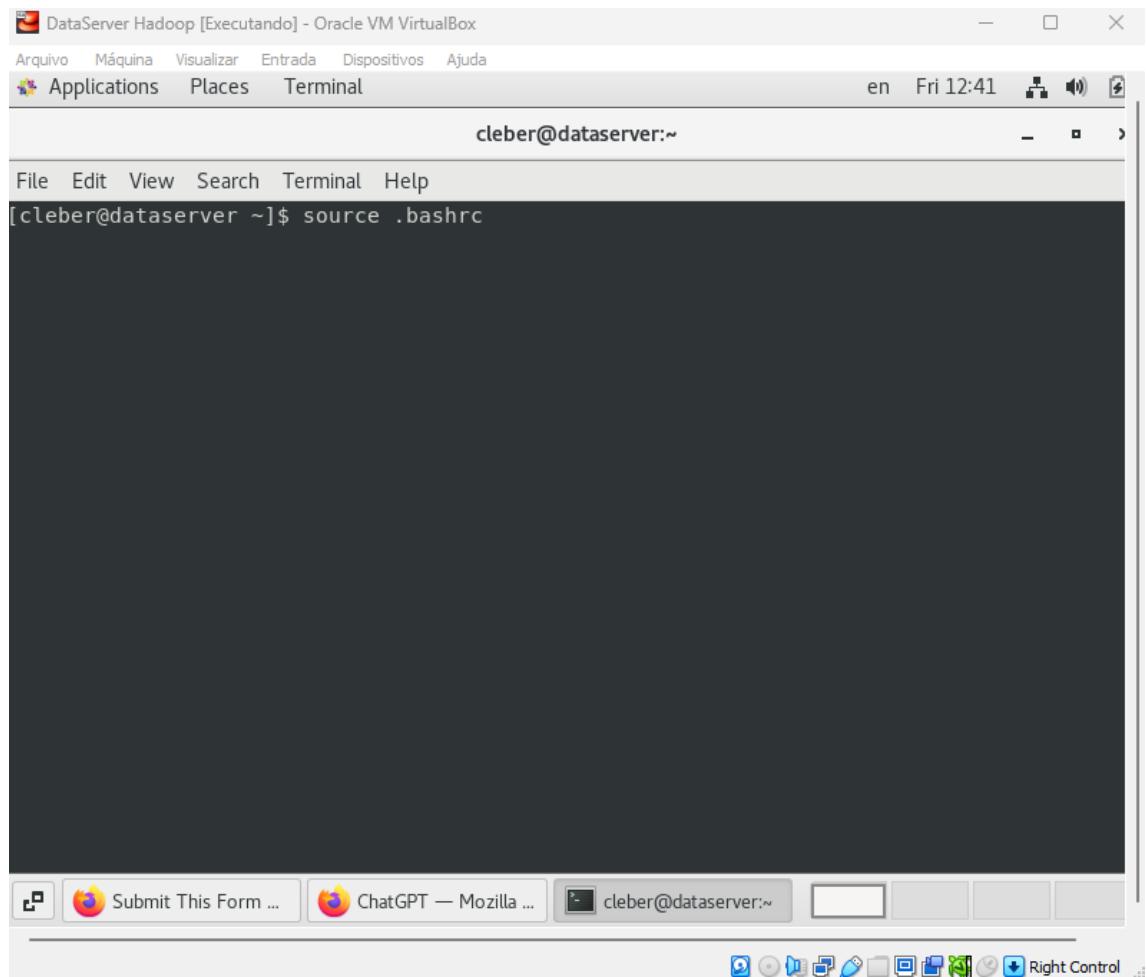
# User specific aliases and functions

# Java JDK
export JAVA_HOME=/opt/jdk
export PATH=$PATH:$JAVA_HOME/bin
```

Editar as variáveis de ambiente conforme acima e salvar o arquivo

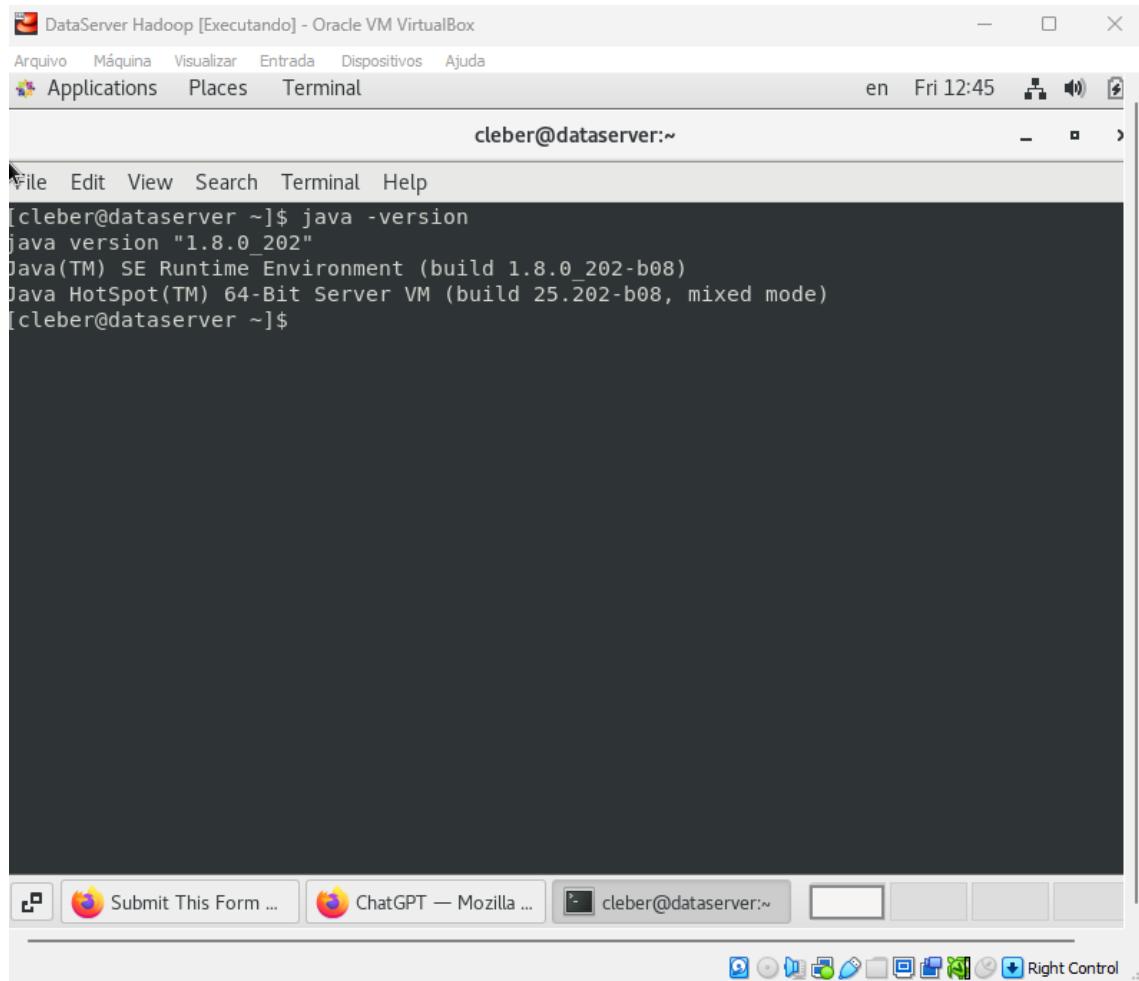
## Instalação e Configuração do Ambiente Big Data com Hadoop e Spark

---



source .bashrc

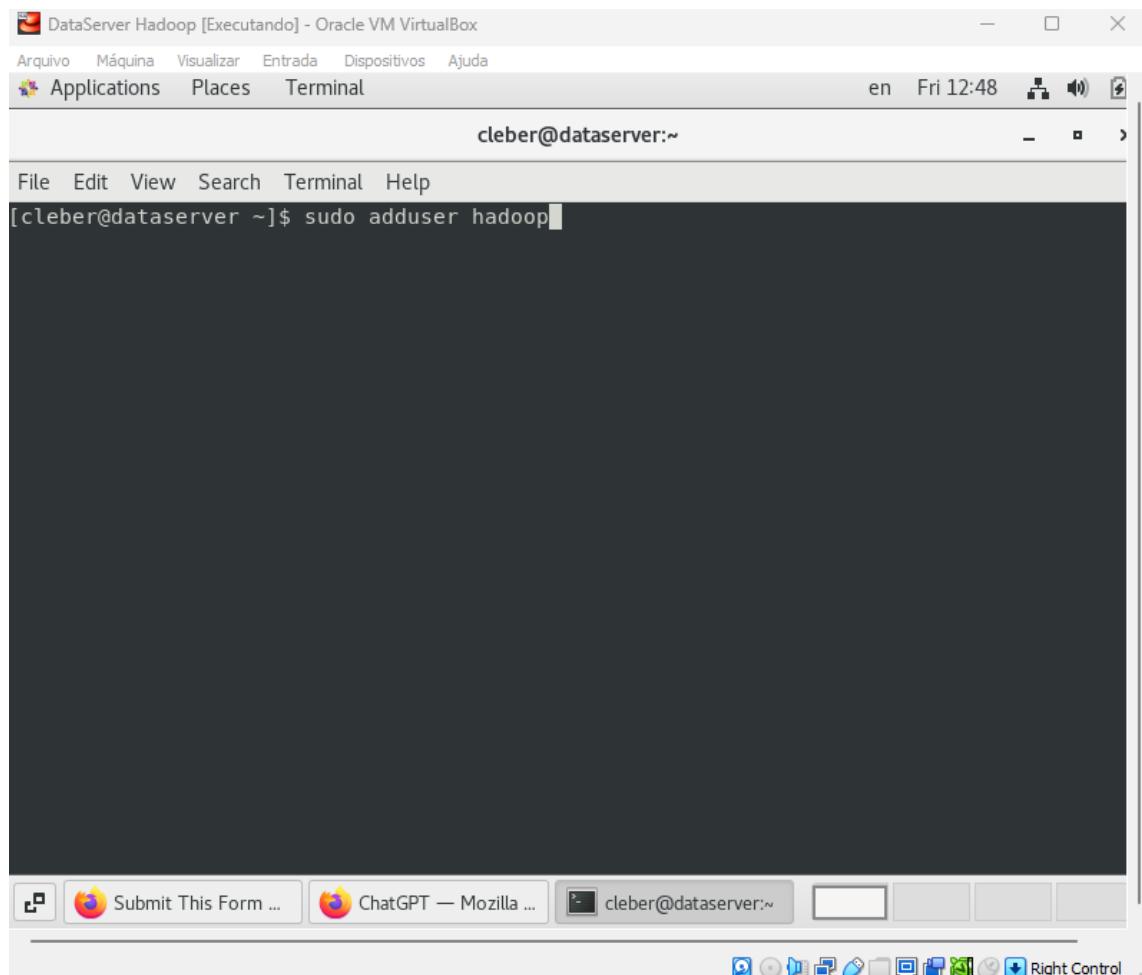
## Instalação e Configuração do Ambiente Big Data com Hadoop e Spark



Verificando a versão do Java JDK

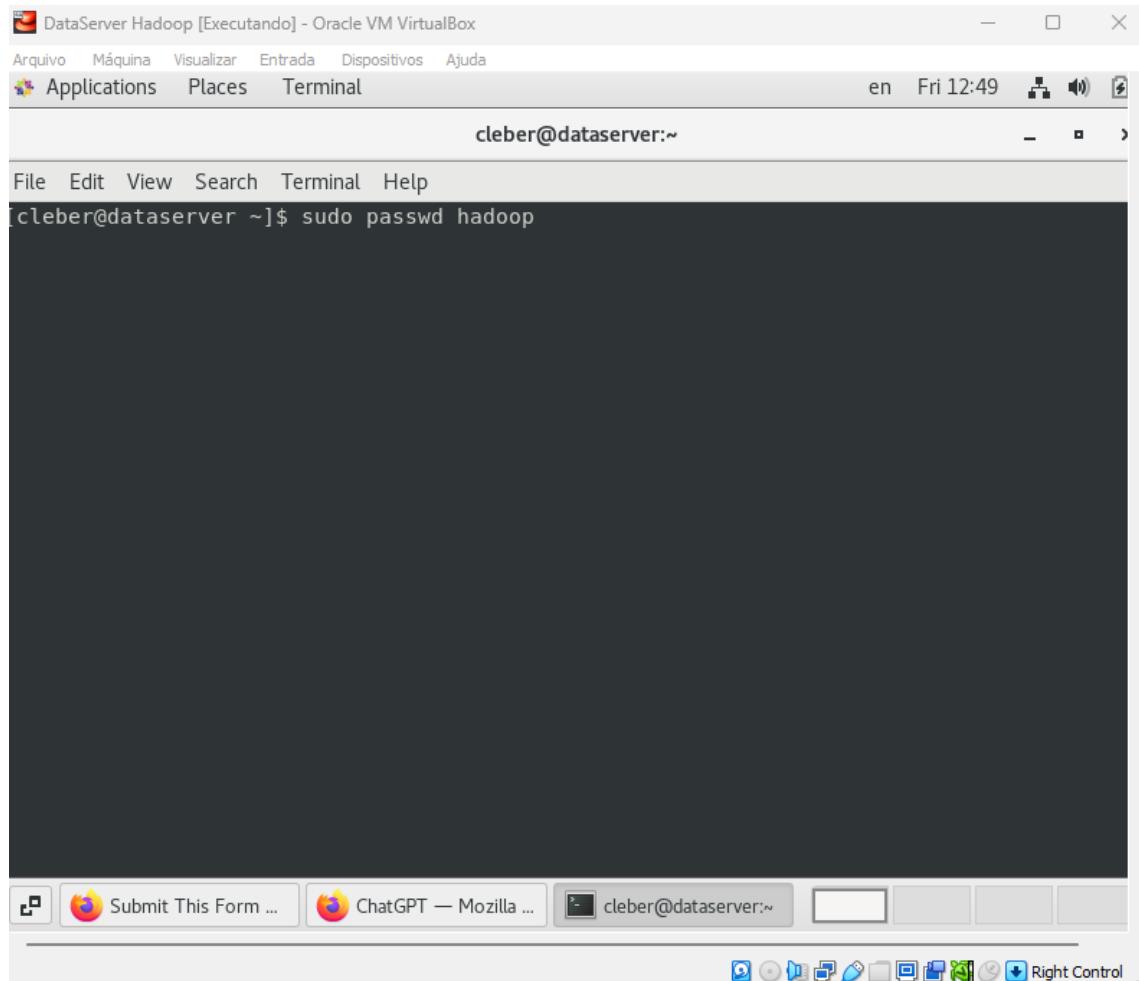
## 5 Instalação e Configuração do Hadoop

### 5.1 Criando o usuário hadoop



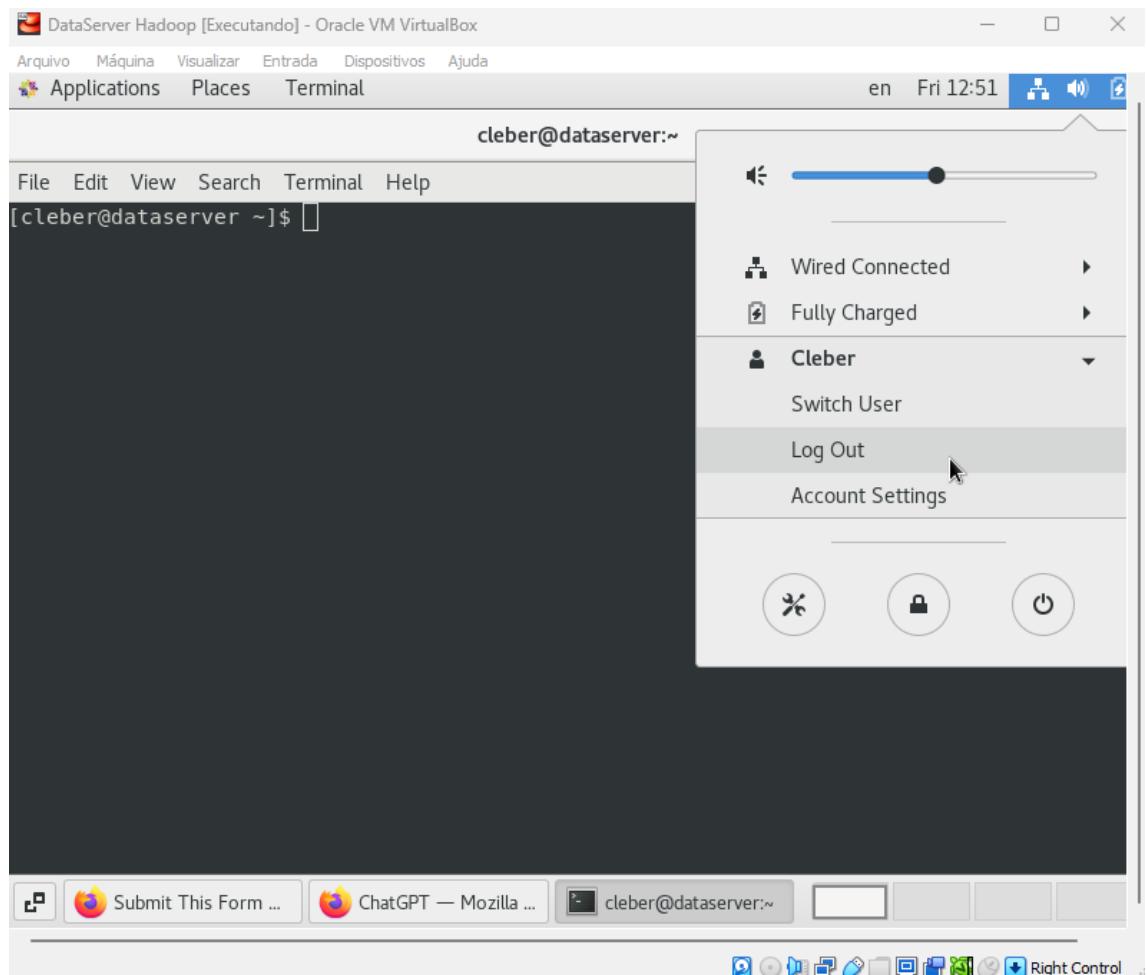
`sudo adduser hadoop` – para criar o usuário hadoop

## Instalação e Configuração do Ambiente Big Data com Hadoop e Spark



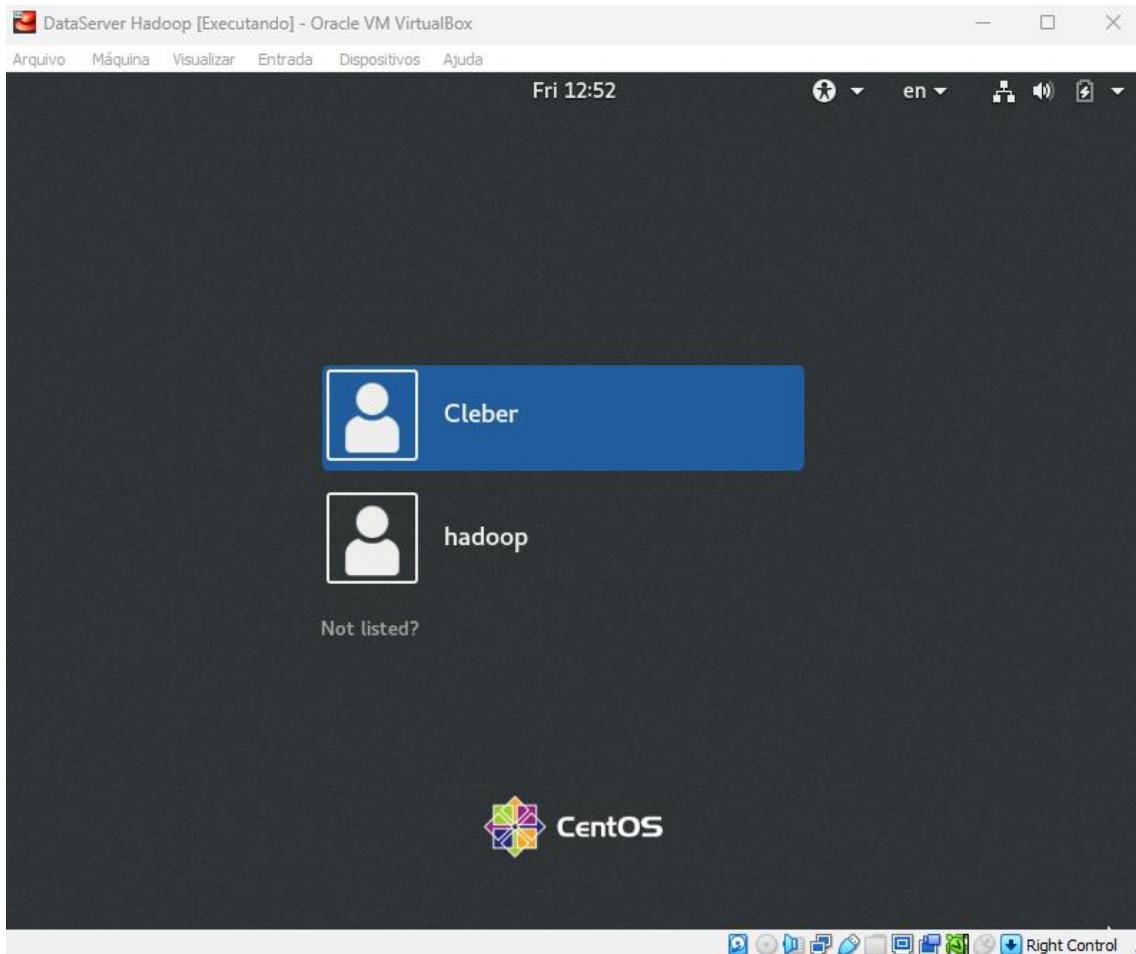
**sudo passwd hadoop – para definir a senha do usuário hadoop  
(hadoop123)**

## Instalação e Configuração do Ambiente Big Data com Hadoop e Spark



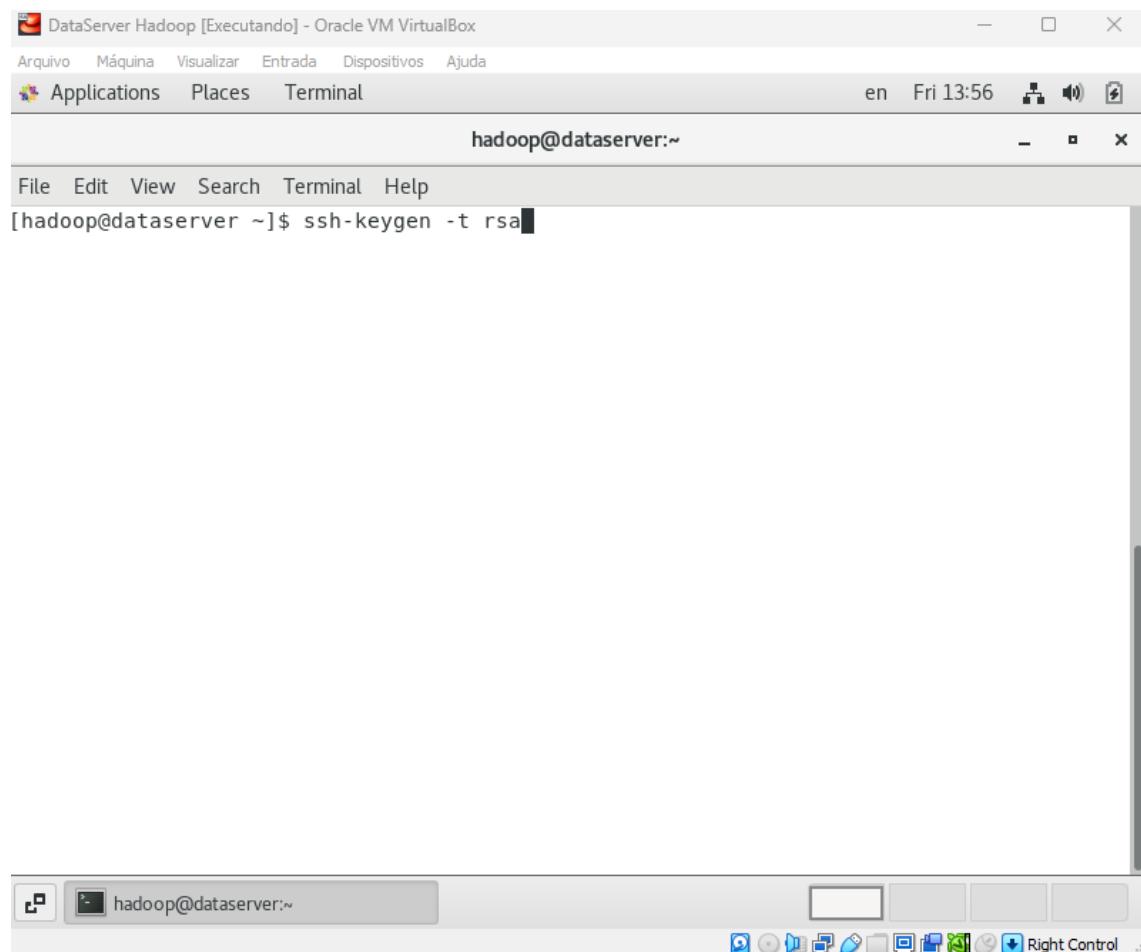
Efetue logout como usuário cleber

## Instalação e Configuração do Ambiente Big Data com Hadoop e Spark



Efetue login como usuário hadoop  
Adicione o usuário hadoop no arquivo /etc/sudoers conforme foi feito  
com o usuário cleber

## 5.2 Configuração do ssh sem senha

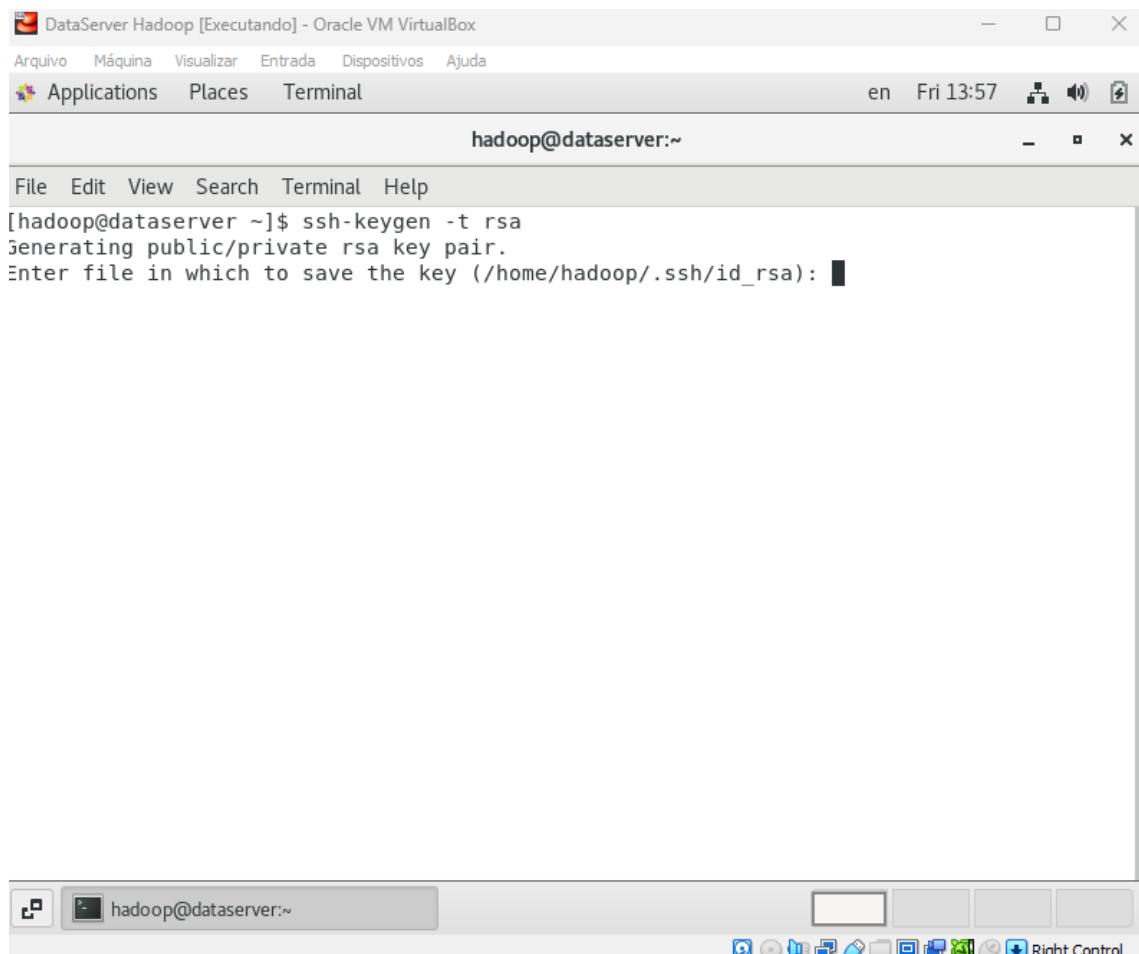


A screenshot of a Linux desktop environment showing a terminal window. The window title is "DataServer Hadoop [Executando] - Oracle VM VirtualBox". The menu bar includes "Arquivo", "Máquina", "Visualizar", "Entrada", "Dispositivos", "Ajuda", "Applications", "Places", and "Terminal". The system tray shows "en Fri 13:56". The terminal prompt is "hadoop@dataserver:~". The command "ssh-keygen -t rsa" is being typed into the terminal. The desktop interface includes a dock with icons for Home, Applications, Places, Terminal, and Right Control.

ssh-keygen -t rsa

## Instalação e Configuração do Ambiente Big Data com Hadoop e Spark

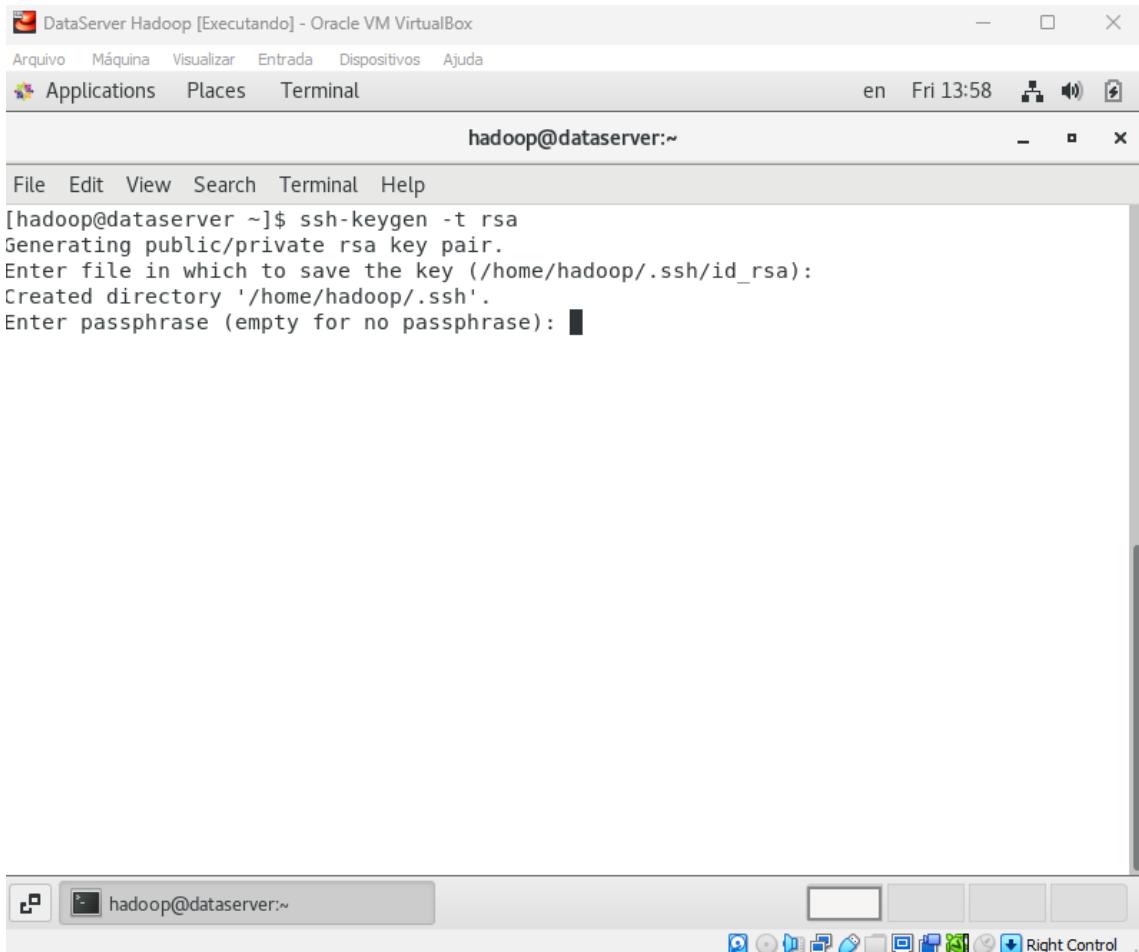
---



```
DataServer Hadoop [Executando] - Oracle VM VirtualBox
Arquivo Máquina Visualizar Entrada Dispositivos Ajuda
Applications Places Terminal en Fri 13:57
hadoop@dataserver:~_
File Edit View Search Terminal Help
[hadoop@dataserver ~]$ ssh-keygen -t rsa
Generating public/private rsa key pair.
Enter file in which to save the key (/home/hadoop/.ssh/id_rsa):
```

Pressionar Enter para confirmar o diretório onde as chaves serão geradas

## Instalação e Configuração do Ambiente Big Data com Hadoop e Spark



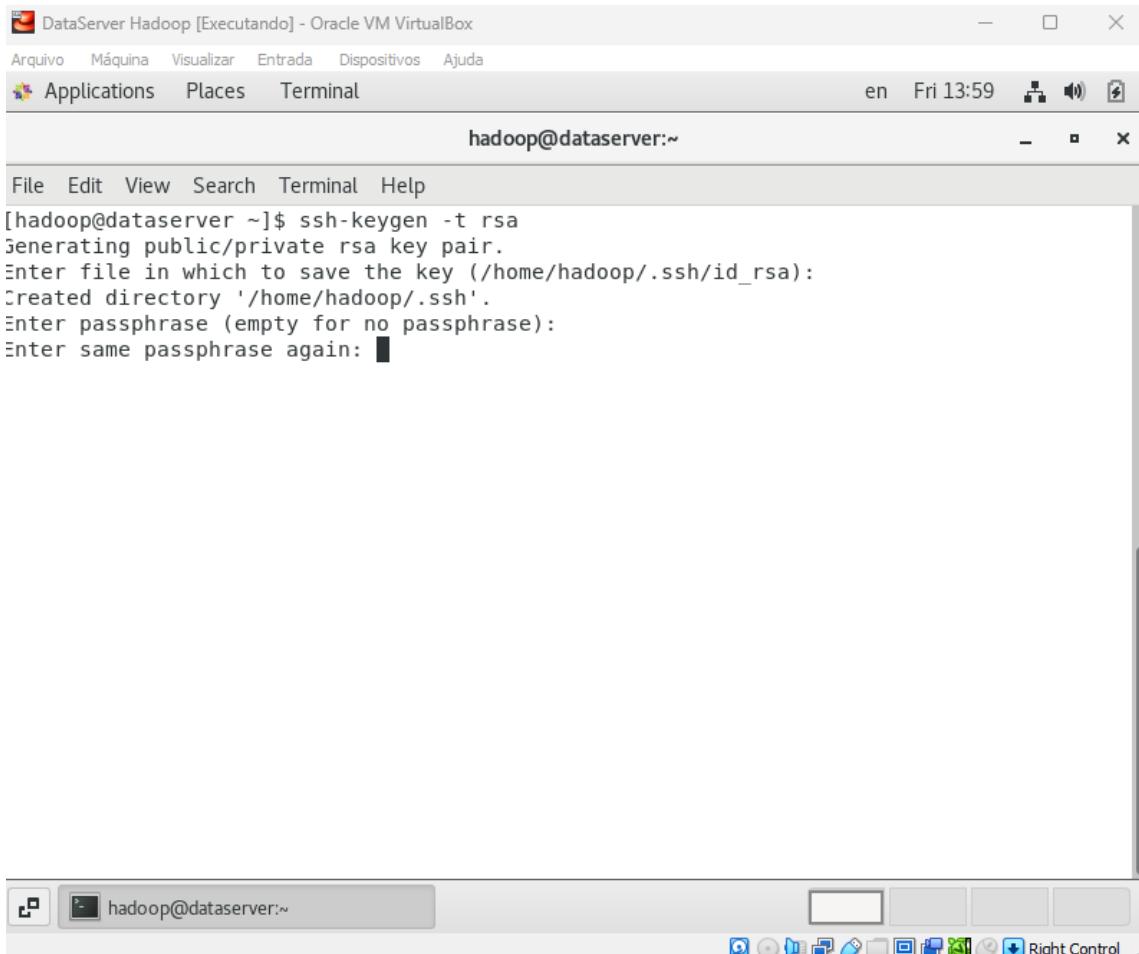
A screenshot of a Linux desktop environment showing a terminal window titled "hadoop@dataserver:~". The window contains the following text:

```
[hadoop@dataserver ~]$ ssh-keygen -t rsa
Generating public/private rsa key pair.
Enter file in which to save the key (/home/hadoop/.ssh/id_rsa):
Created directory '/home/hadoop/.ssh'.
Enter passphrase (empty for no passphrase): █
```

The terminal window has a standard Linux-style title bar with icons for minimize, maximize, and close. Below the title bar is a menu bar with "File", "Edit", "View", "Search", "Terminal", and "Help". The main area of the window shows the command-line interaction. At the bottom of the window is a toolbar with various icons.

Pressionar Enter

## Instalação e Configuração do Ambiente Big Data com Hadoop e Spark



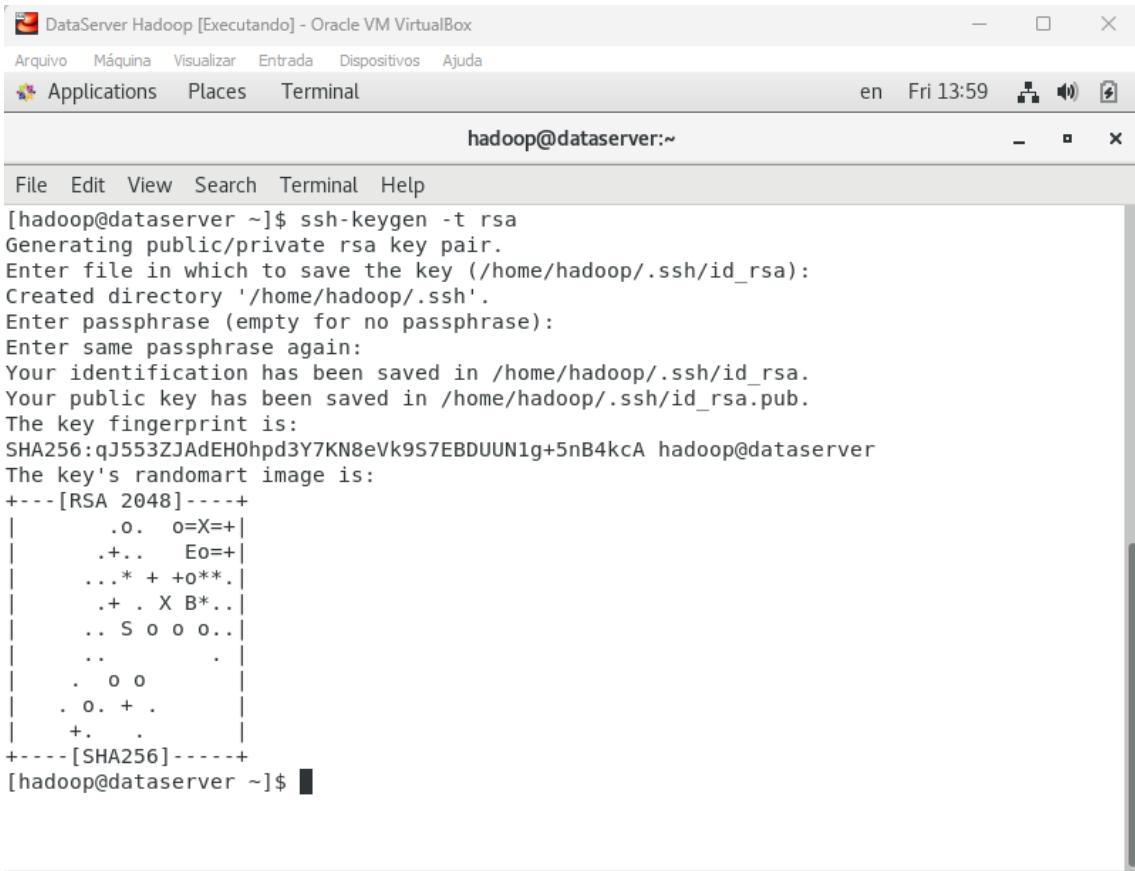
A screenshot of a Linux desktop environment showing a terminal window titled "hadoop@dataserver:~". The window contains the following command and its output:

```
[hadoop@dataserver ~]$ ssh-keygen -t rsa
Generating public/private rsa key pair.
Enter file in which to save the key (/home/hadoop/.ssh/id_rsa):
Created directory '/home/hadoop/.ssh'.
Enter passphrase (empty for no passphrase):
Enter same passphrase again: █
```

The terminal window has a standard Linux-style interface with a menu bar, application menu, and system status indicators at the top. The bottom of the window shows the desktop environment's taskbar with various icons.

Pressionar Enter novamente

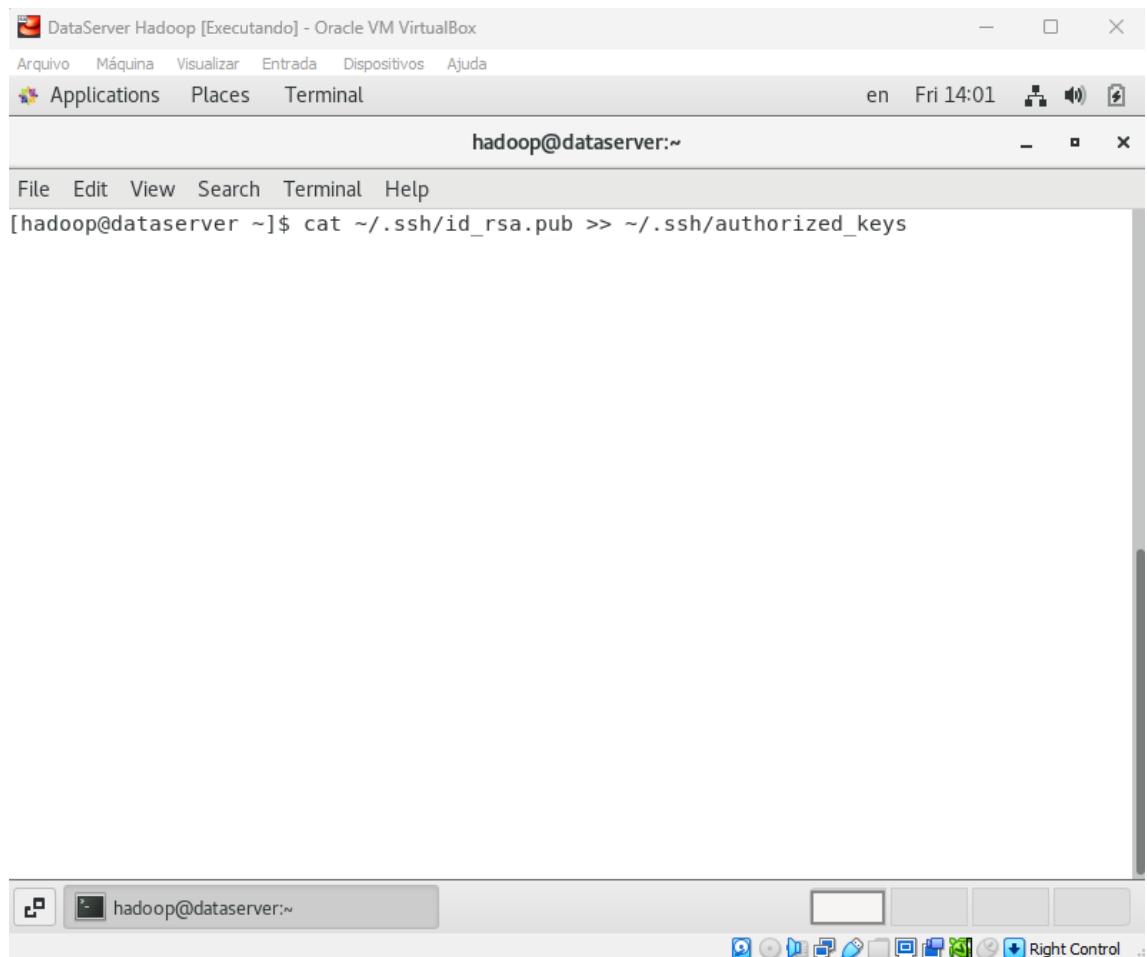
## Instalação e Configuração do Ambiente Big Data com Hadoop e Spark



```
[hadoop@dataserver ~]$ ssh-keygen -t rsa
Generating public/private rsa key pair.
Enter file in which to save the key (/home/hadoop/.ssh/id_rsa):
Created directory '/home/hadoop/.ssh'.
Enter passphrase (empty for no passphrase):
Enter same passphrase again:
Your identification has been saved in /home/hadoop/.ssh/id_rsa.
Your public key has been saved in /home/hadoop/.ssh/id_rsa.pub.
The key fingerprint is:
SHA256:qJ553ZJAdEH0hpD3Y7KN8eVkJ9S7EBDUUN1g+5nB4kcA hadoop@dataserver
The key's randomart image is:
+---[RSA 2048]---+
| .o. o=X=+|
| .+.. Eo=+|
| ...* ++o**.|
| .+ . X B*..|
| .. S o o o..|
| .. . . . . .|
| . o o |
| . o. + . |
| +. . |
+---[SHA256]---+
[hadoop@dataserver ~]$
```

Chaves de segurança geradas

## Instalação e Configuração do Ambiente Big Data com Hadoop e Spark

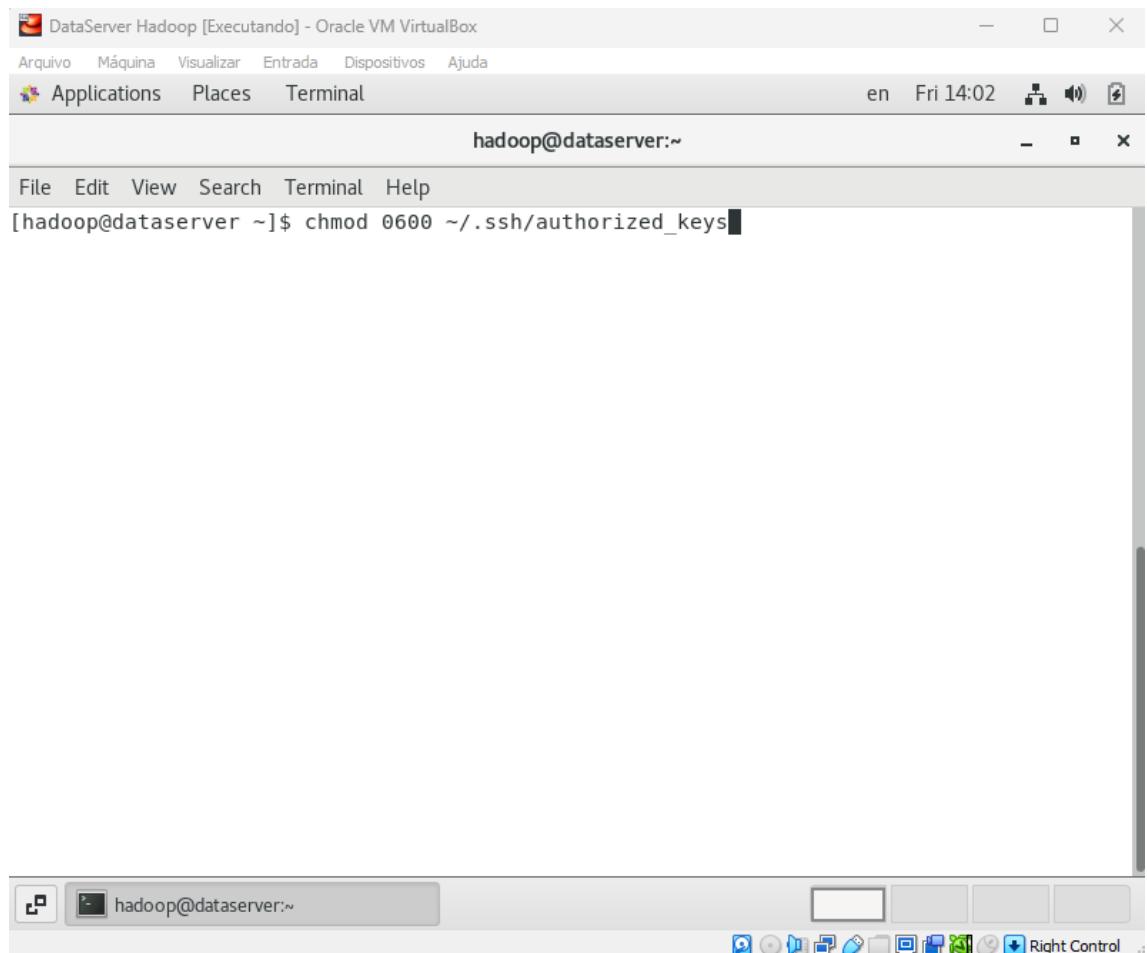


```
[hadoop@dataserver ~]$ cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
```

`cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys`

Esse comando copia a chave pública para o arquivo `authorized_keys` do ssh

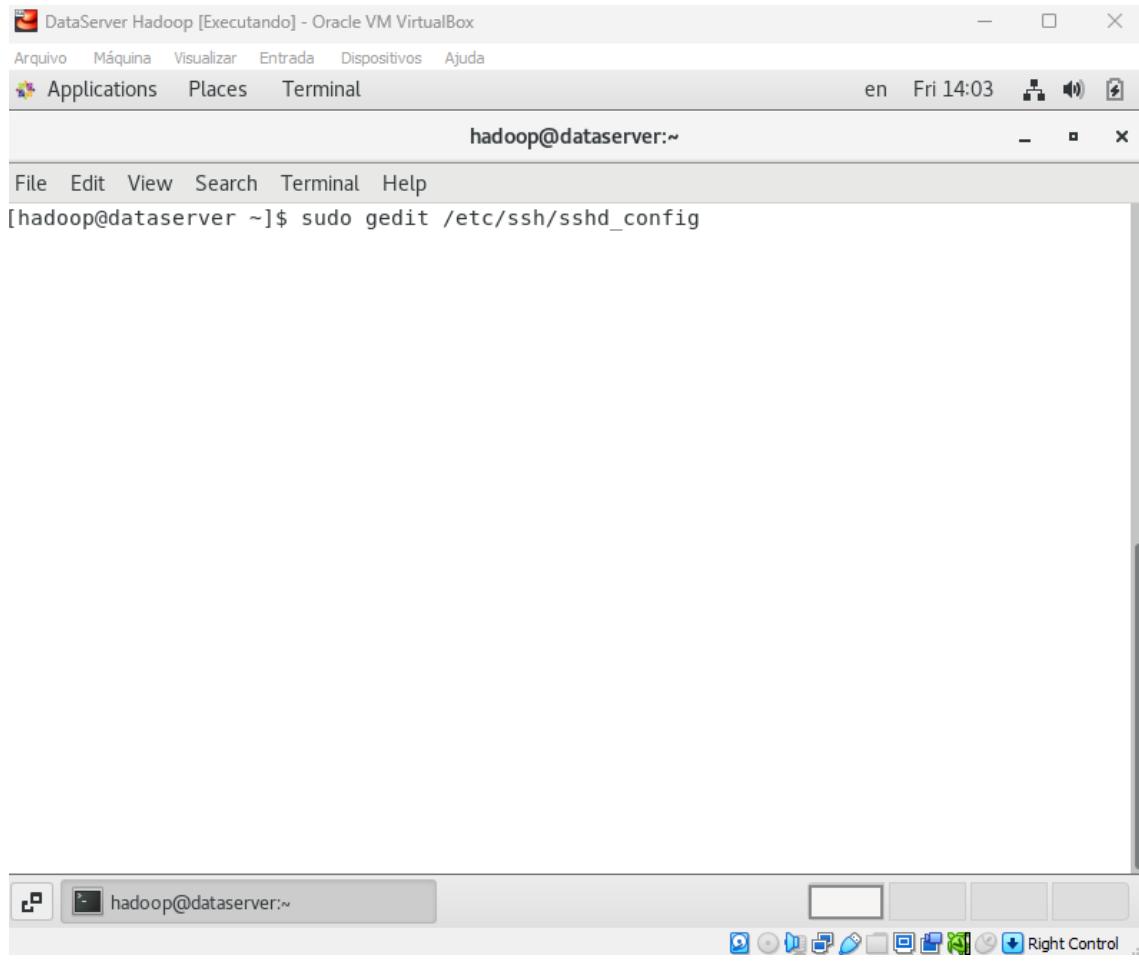
## Instalação e Configuração do Ambiente Big Data com Hadoop e Spark



```
hadoop@dataserver:~$ chmod 0600 ~/.ssh/authorized_keys
```

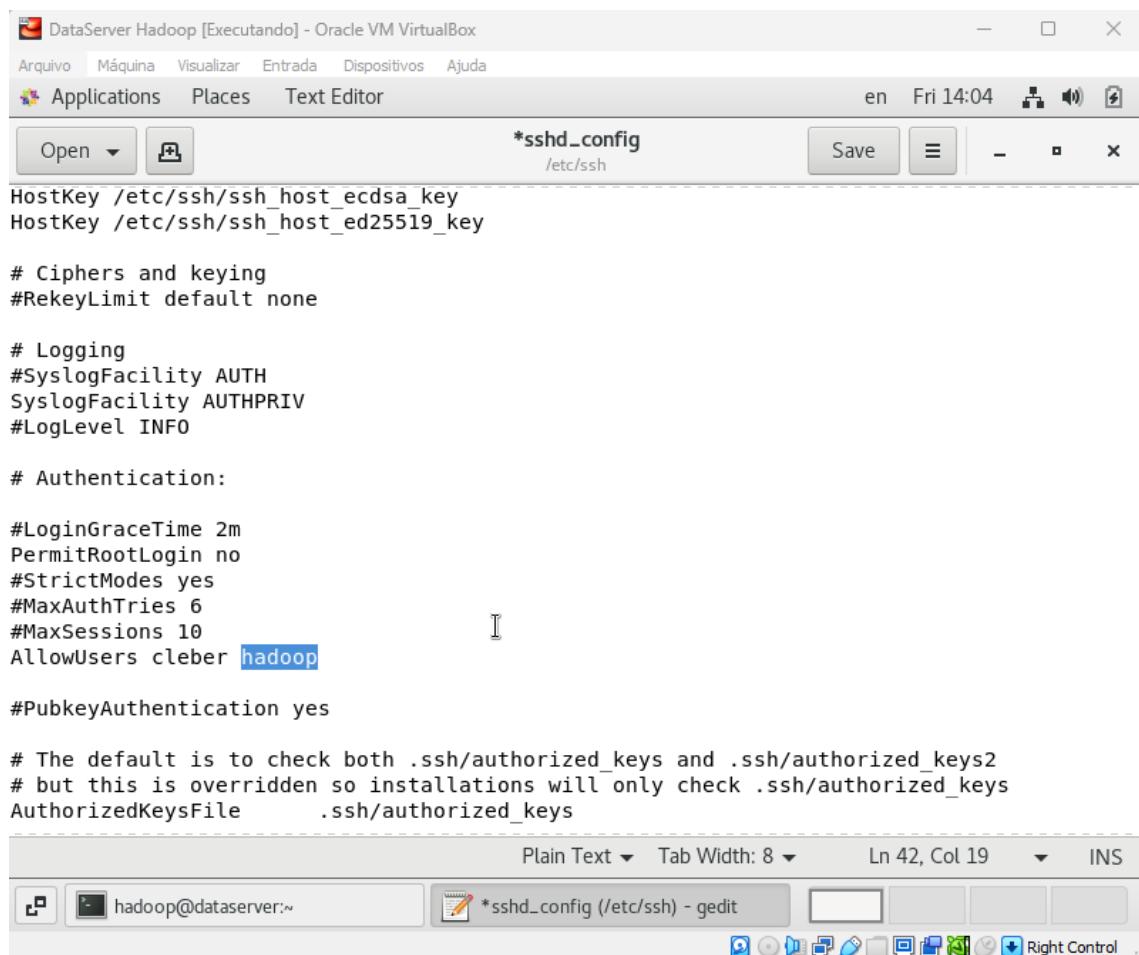
chmod 0600 ~/.ssh/authorized\_keys  
Esse comando define a permissão do arquivo authorized\_keys

## Instalação e Configuração do Ambiente Big Data com Hadoop e Spark



sudo gedit /etc/ssh/sshd\_config  
Edite o arquivo de configuração do ssh

## Instalação e Configuração do Ambiente Big Data com Hadoop e Spark



```
*sshd_config
/etc/ssh

HostKey /etc/ssh/ssh_host_ecdsa_key
HostKey /etc/ssh/ssh_host_ed25519_key

# Ciphers and keying
#RekeyLimit default none

# Logging
#SyslogFacility AUTH
SyslogFacility AUTHPRIV
#LogLevel INFO

# Authentication:

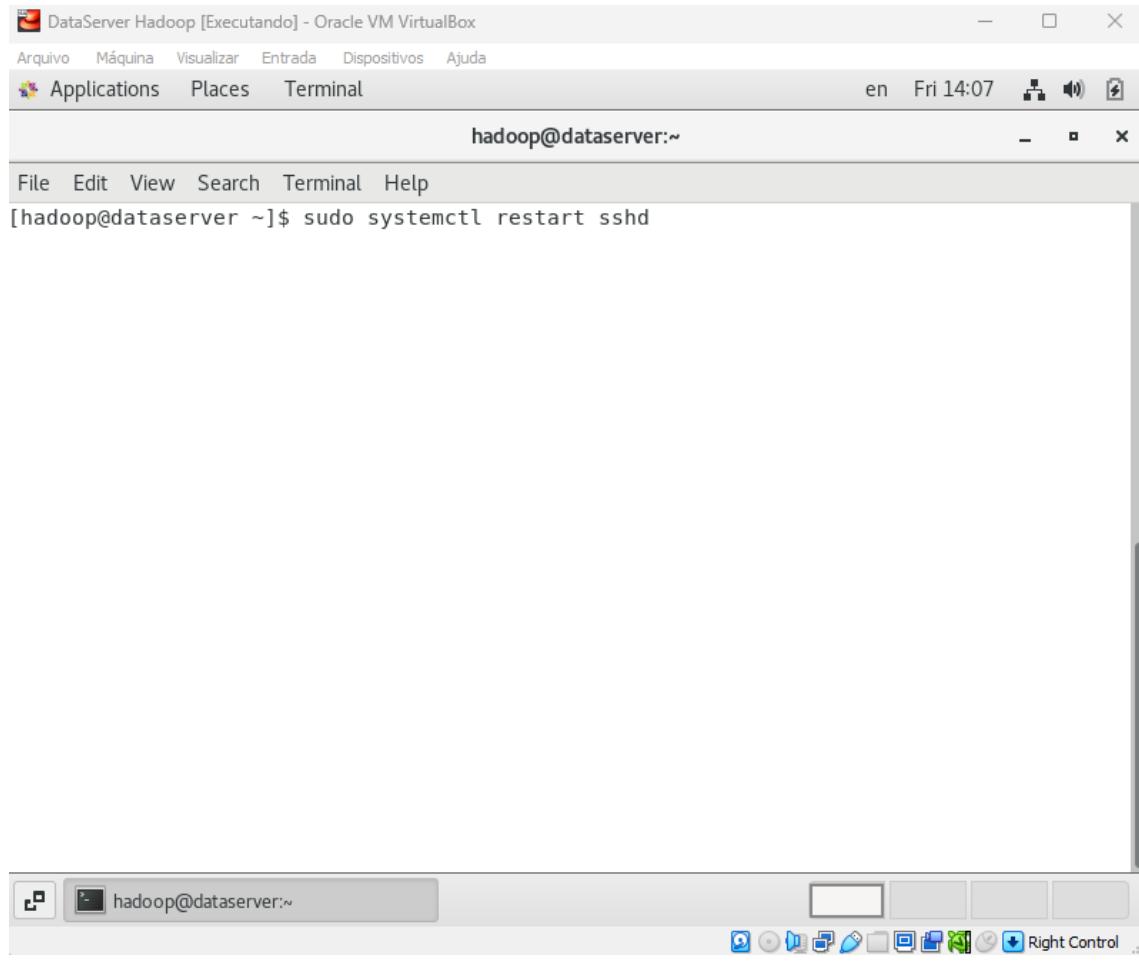
#LoginGraceTime 2m
PermitRootLogin no
#StrictModes yes
#MaxAuthTries 6
#MaxSessions 10
AllowUsers cleber hadoop

#PubkeyAuthentication yes

# The default is to check both .ssh/authorized_keys and .ssh/authorized_keys2
# but this is overridden so installations will only check .ssh/authorized_keys
AuthorizedKeysFile      .ssh/authorized_keys
```

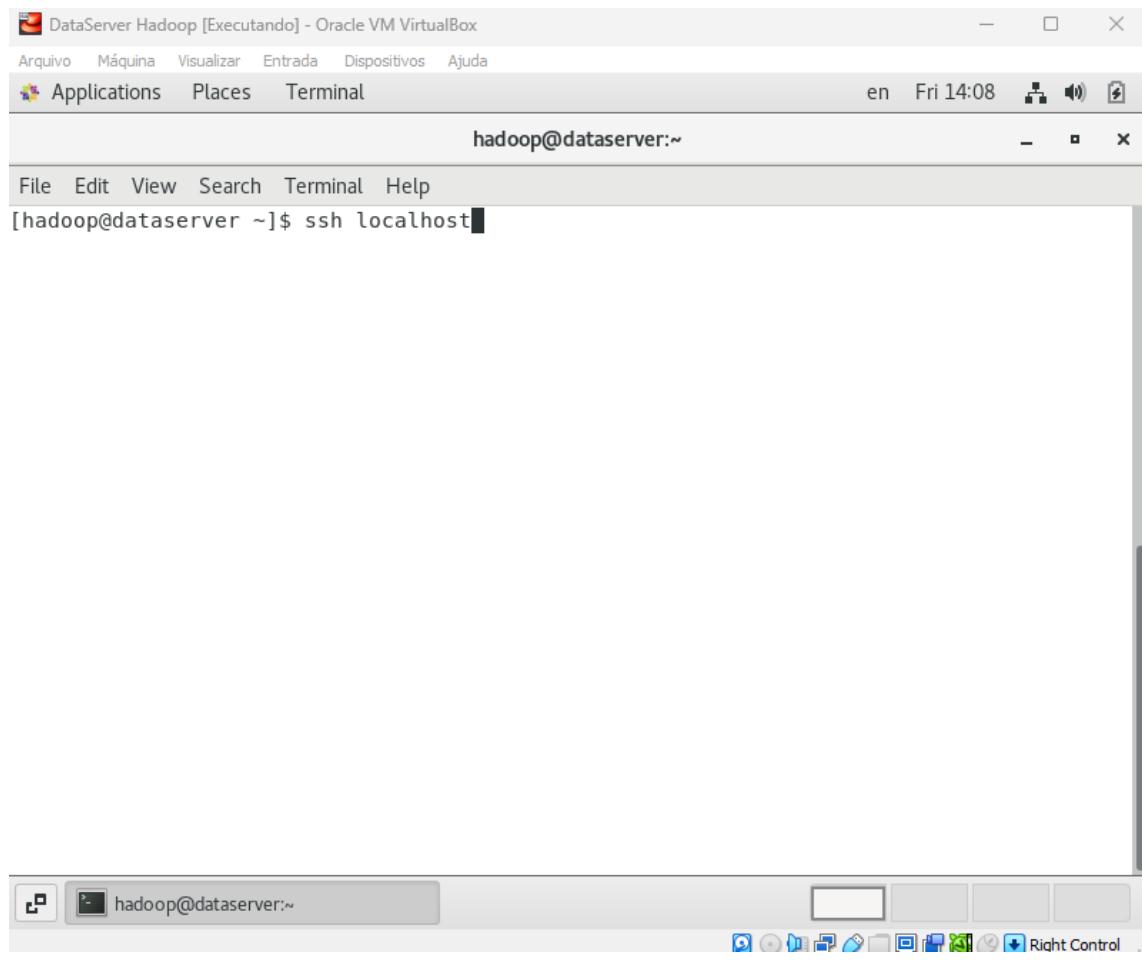
Inclua o usuário hadoop na linha AllowUsers, salve o arquivo e feche-o

## Instalação e Configuração do Ambiente Big Data com Hadoop e Spark



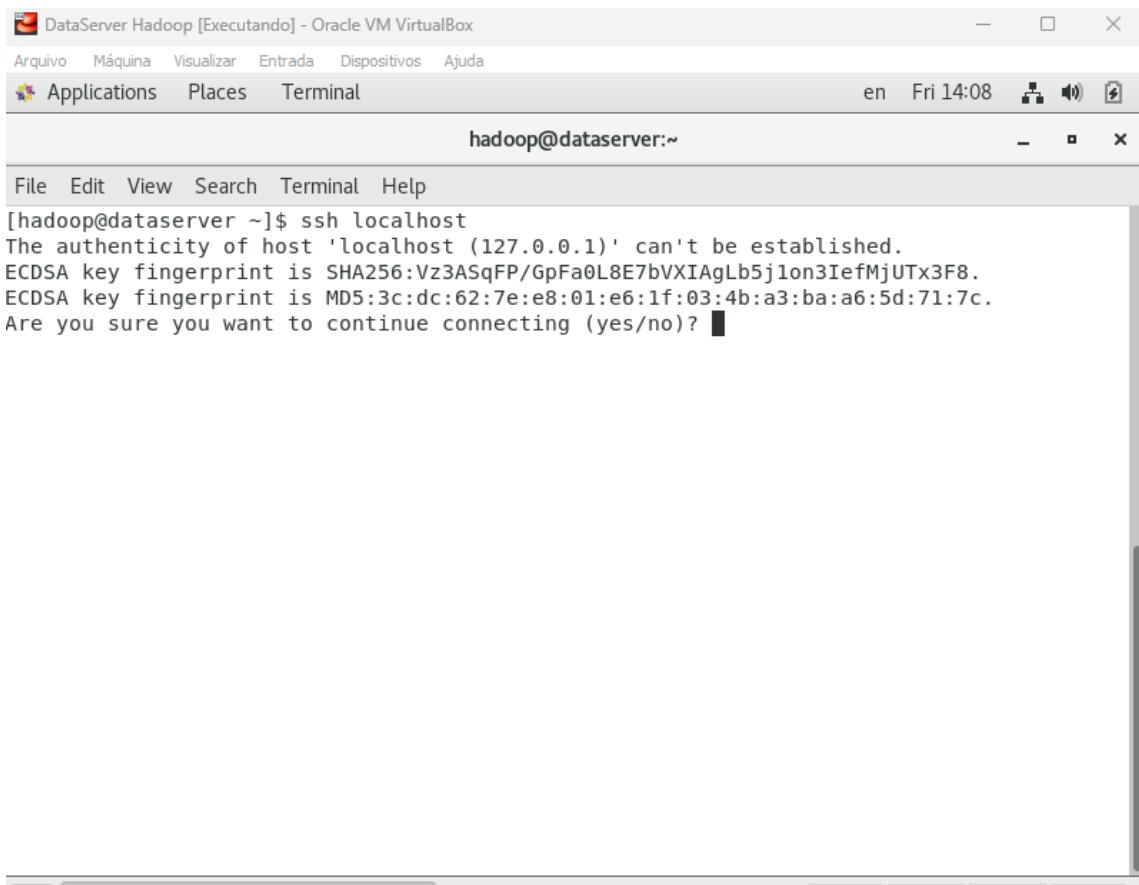
**sudo systemctl restart sshd**  
Reinic peace o serviço ssh

## Instalação e Configuração do Ambiente Big Data com Hadoop e Spark



ssh localhost

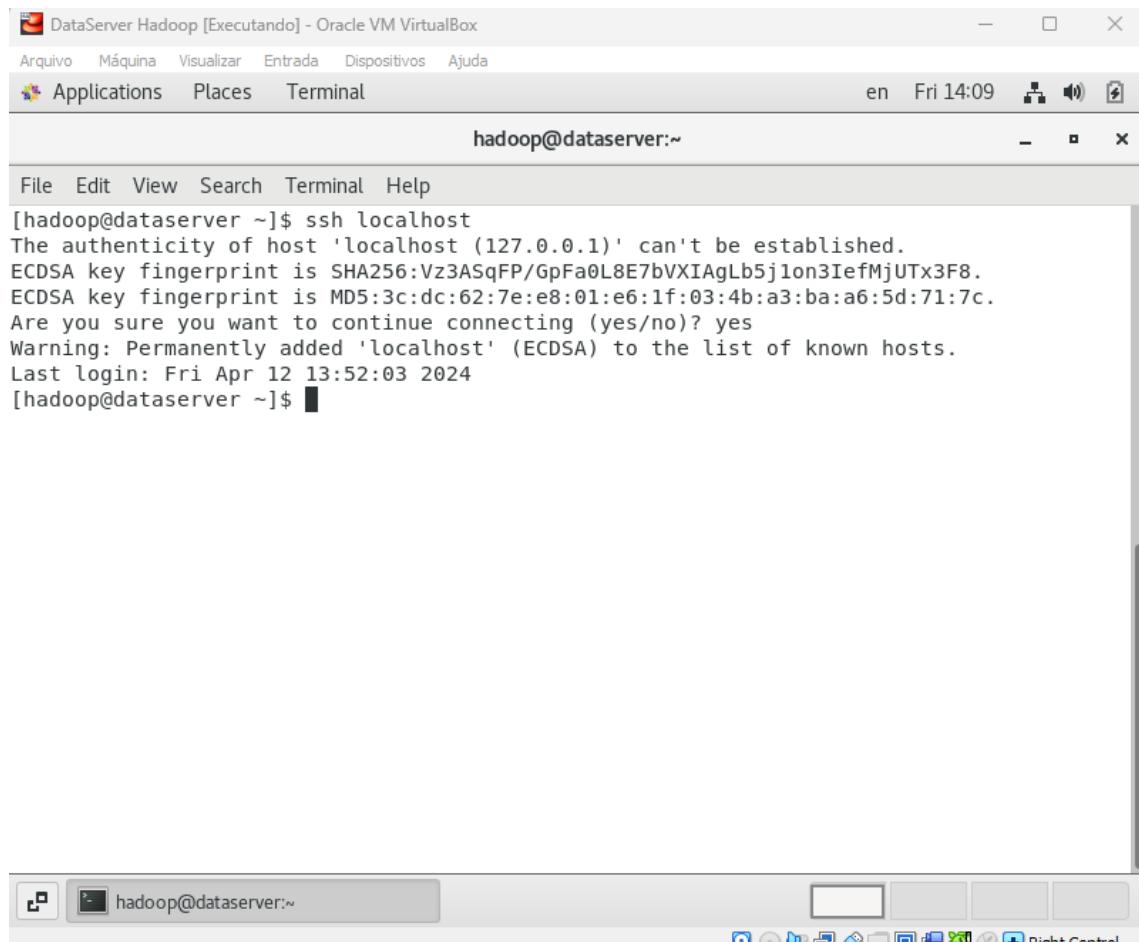
## Instalação e Configuração do Ambiente Big Data com Hadoop e Spark



hadoop@dataserver:~\$ ssh localhost  
The authenticity of host 'localhost (127.0.0.1)' can't be established.  
ECDSA key fingerprint is SHA256:Vz3ASqFP/GpFa0L8E7bVXIAgLb5j1on3IefMjUTx3F8.  
ECDSA key fingerprint is MD5:3c:dc:62:7e:e8:01:e6:1f:03:4b:a3:ba:a6:5d:71:7c.  
Are you sure you want to continue connecting (yes/no)?

Yes

## Instalação e Configuração do Ambiente Big Data com Hadoop e Spark



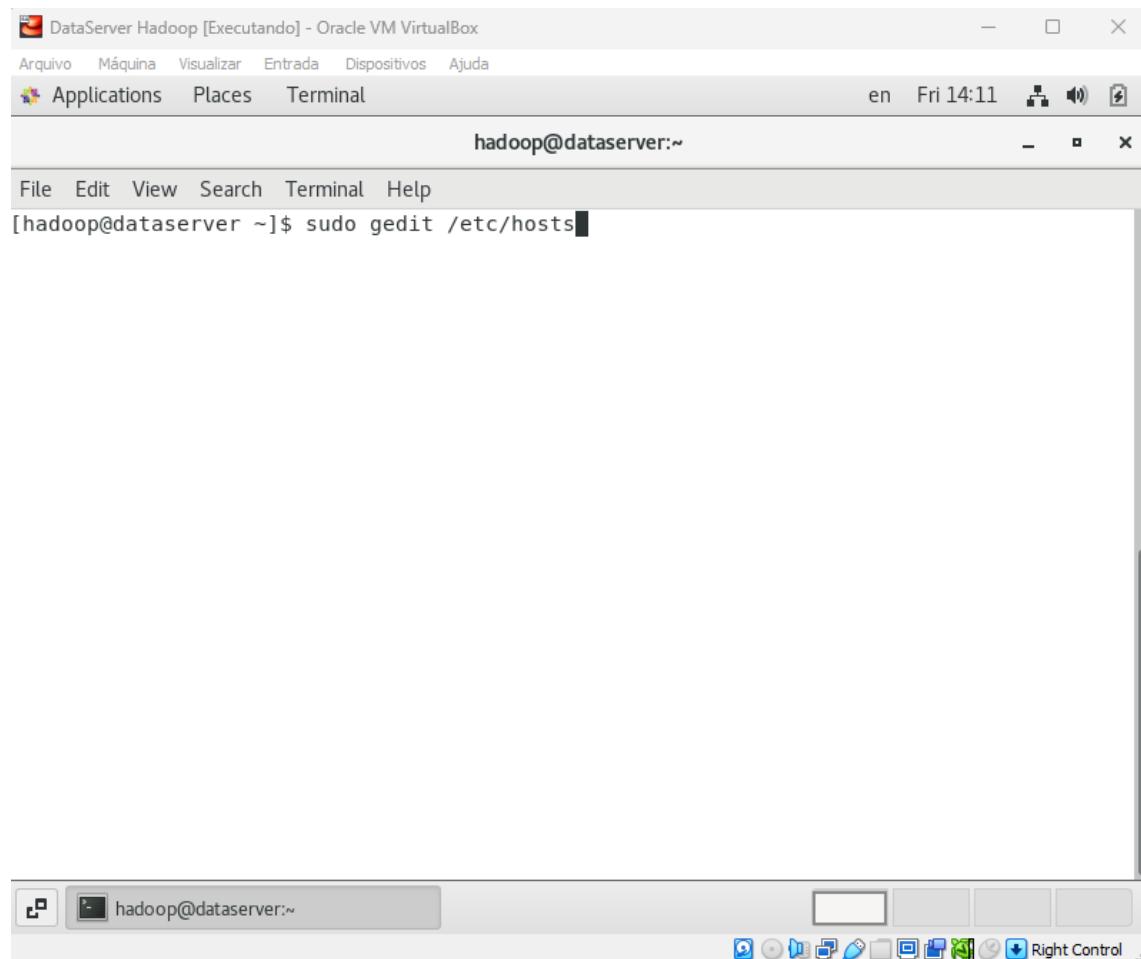
```
[hadoop@dataserver ~]$ ssh localhost
The authenticity of host 'localhost (127.0.0.1)' can't be established.
ECDSA key fingerprint is SHA256:Vz3ASqFP/GpFa0L8E7bVXIAgLb5j1on3IefMjUTx3F8.
ECDSA key fingerprint is MD5:3c:dc:62:7e:e8:01:e6:1f:03:4b:a3:ba:a6:5d:71:7c.
Are you sure you want to continue connecting (yes/no)? yes
Warning: Permanently added 'localhost' (ECDSA) to the list of known hosts.
Last login: Fri Apr 12 13:52:03 2024
[hadoop@dataserver ~]$
```

Conexão ssh sem senha efetuada com sucesso. Digite exit e pressione Enter.

O ambiente está pronto para receber o Hadoop!!

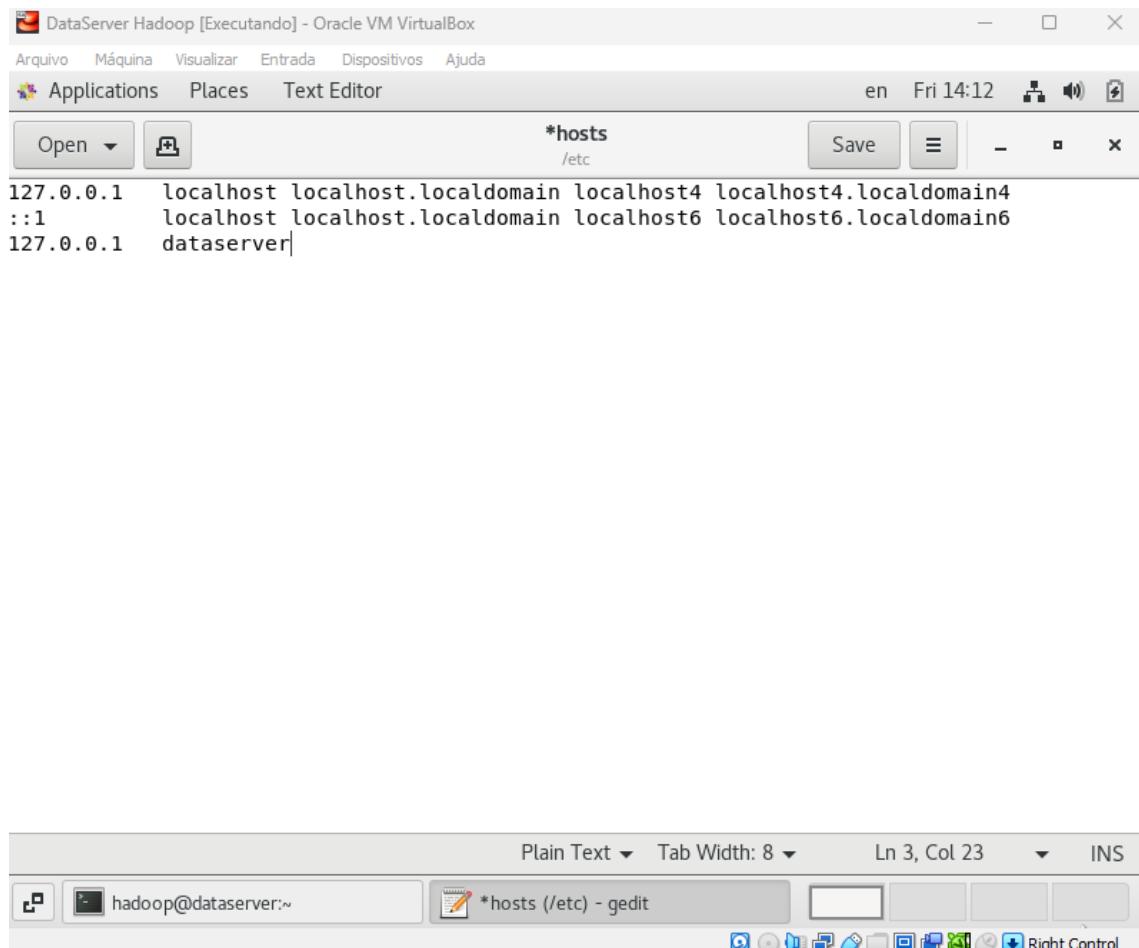
## 5.3 Download e Instalação do Hadoop

### 5.3.1 Editando o arquivo hosts



Editar o arquivo hosts

## Instalação e Configuração do Ambiente Big Data com Hadoop e Spark



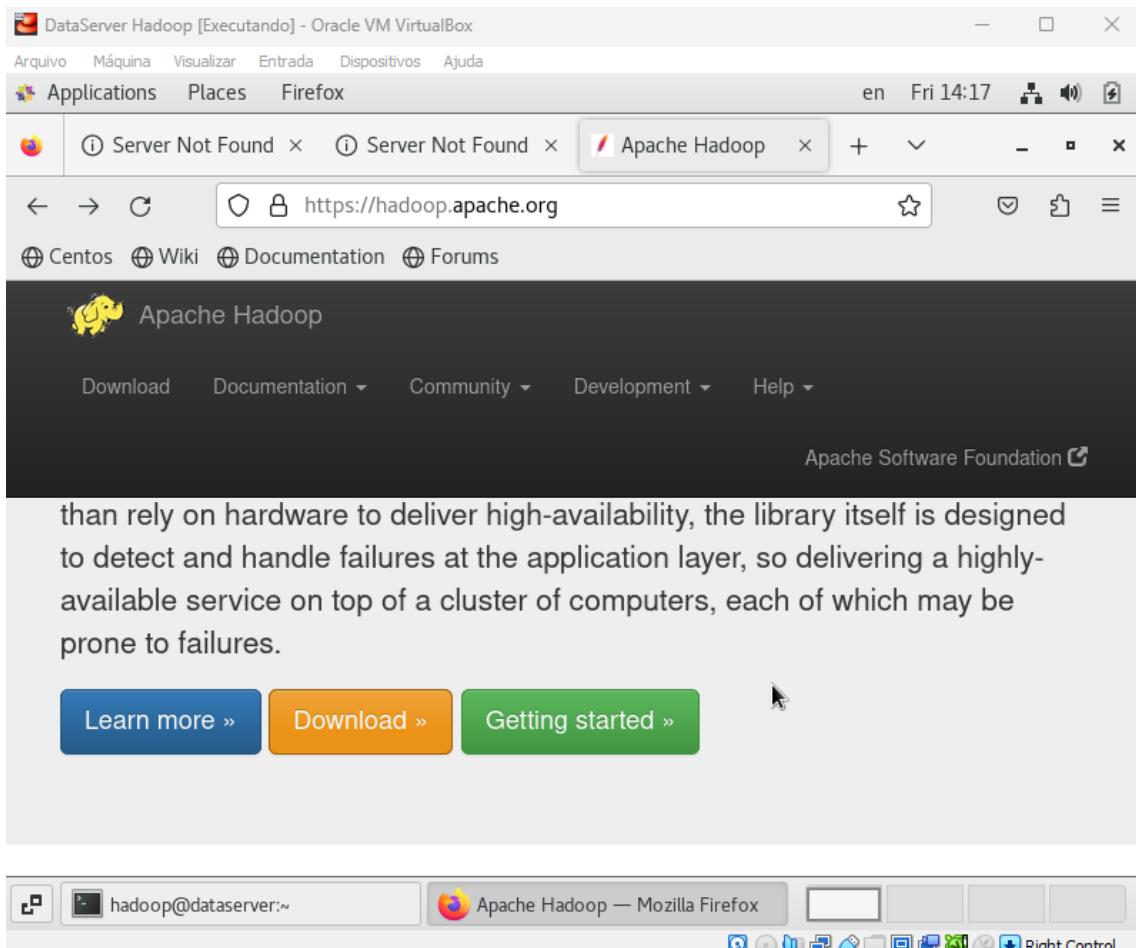
The screenshot shows a window titled "DataServer Hadoop [Executando] - Oracle VM VirtualBox". The menu bar includes "Arquivo", "Máquina", "Visualizar", "Entrada", "Dispositivos", and "Ajuda". The top right shows "en Fri 14:12". The window title bar says "Applications Places Text Editor". The main area is a text editor for the "/etc/hosts" file. It contains the following entries:

```
127.0.0.1 localhost localhost.localdomain localhost4 localhost4.localdomain4
::1 localhost localhost.localdomain localhost6 localhost6.localdomain6
127.0.0.1 dataserver|
```

Below this is a smaller window titled "Plain Text" showing the same content. The status bar at the bottom indicates "Ln 3, Col 23" and "INS".

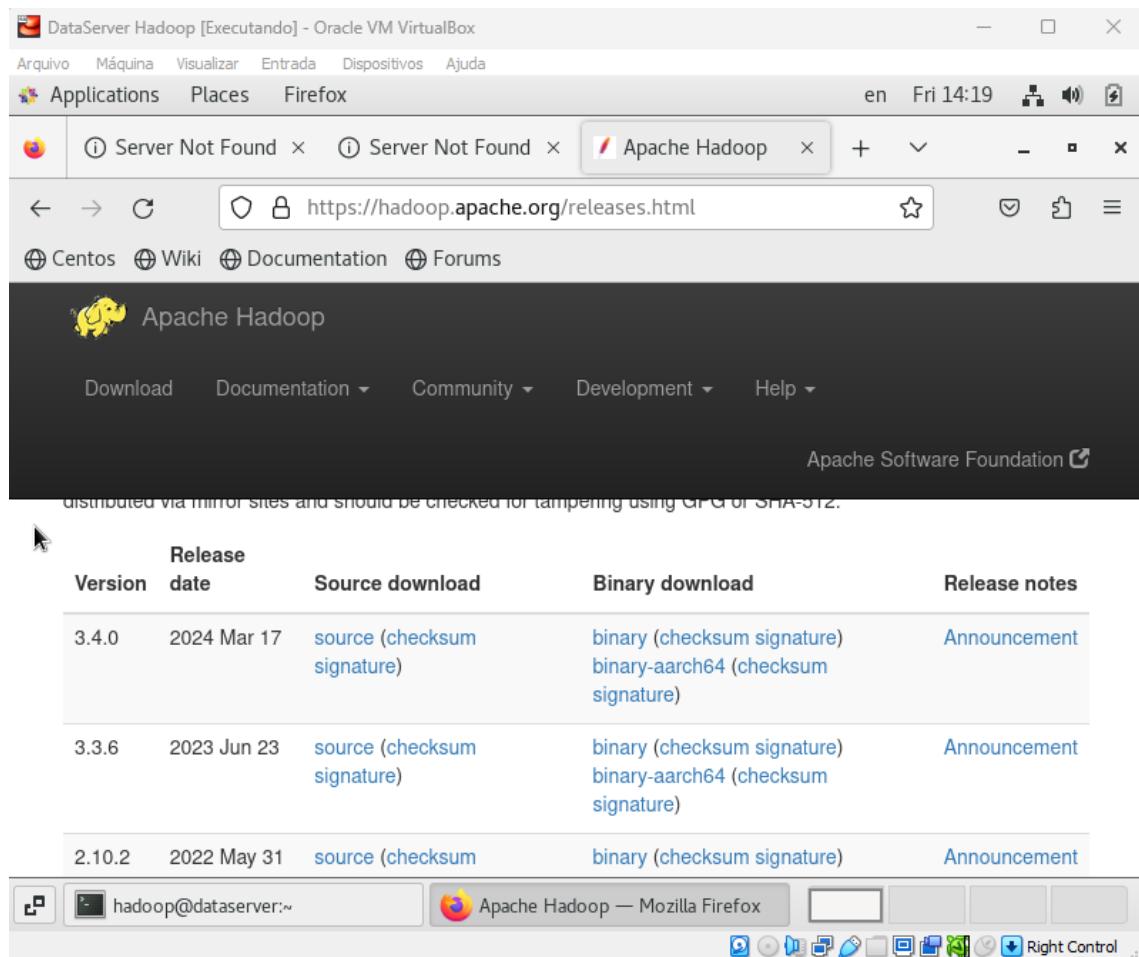
Incluir a linha 127.0.0.1 dataserver conforme acima

### 5.3.2 Download do Hadoop



Acesse a página do Hadoop e clique em Download

# Instalação e Configuração do Ambiente Big Data com Hadoop e Spark

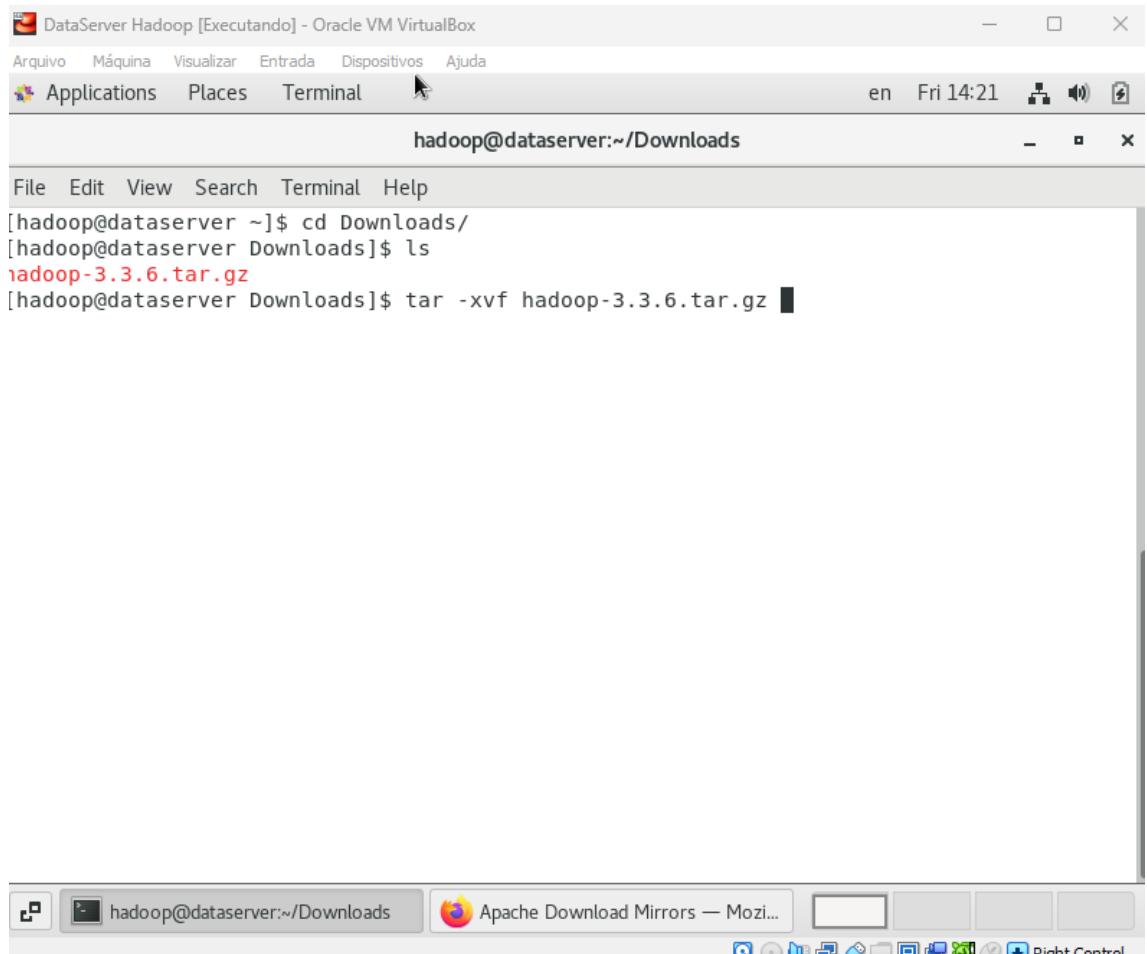


The screenshot shows a Firefox browser window titled "DataServer Hadoop [Executando] - Oracle VM VirtualBox". The address bar displays "https://hadoop.apache.org/releases.html". The page content is the Apache Hadoop releases page, featuring the Apache Software Foundation logo at the top right. Below it, a note states: "distributed via mirror sites and should be checked for tampering using GPG or SHA-512." A table lists three releases:

Version	Release date	Source download	Binary download	Release notes
3.4.0	2024 Mar 17	<a href="#">source (checksum signature)</a>	<a href="#">binary (checksum signature)</a> <a href="#">binary-aarch64 (checksum signature)</a>	<a href="#">Announcement</a>
3.3.6	2023 Jun 23	<a href="#">source (checksum signature)</a>	<a href="#">binary (checksum signature)</a> <a href="#">binary-aarch64 (checksum signature)</a>	<a href="#">Announcement</a>
2.10.2	2022 May 31	<a href="#">source (checksum)</a>	<a href="#">binary (checksum signature)</a>	<a href="#">Announcement</a>

Faça o download da versão 3.3, opção binary.  
O arquivo será baixado no diretório /home/hadoop/Downloads

## Instalação e Configuração do Ambiente Big Data com Hadoop e Spark



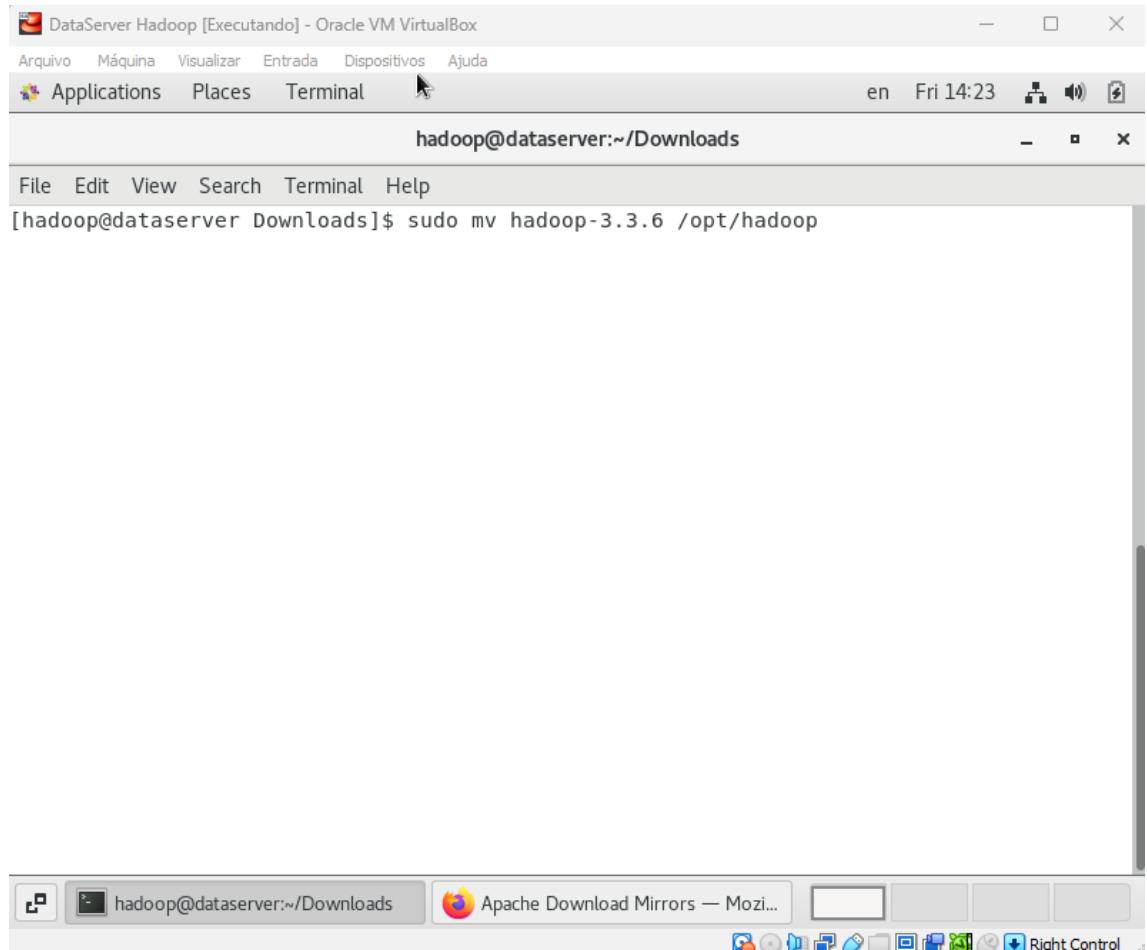
The screenshot shows a Linux desktop environment with a terminal window open. The terminal window title is "hadoop@dataserver:~/Downloads". The terminal content shows the command sequence:

```
[hadoop@dataserver ~]$ cd Downloads/  
[hadoop@dataserver Downloads]$ ls  
hadoop-3.3.6.tar.gz  
[hadoop@dataserver Downloads]$ tar -xvf hadoop-3.3.6.tar.gz
```

Below the terminal, a file manager window titled "Apache Download Mirrors — Mozilla Firefox" is visible, showing a list of download links. The desktop interface includes a menu bar with "Arquivo", "Máquina", "Visualizar", "Entrada", "Dispositivos", and "Ajuda". The taskbar at the bottom shows icons for Applications, Places, Terminal, and a file manager.

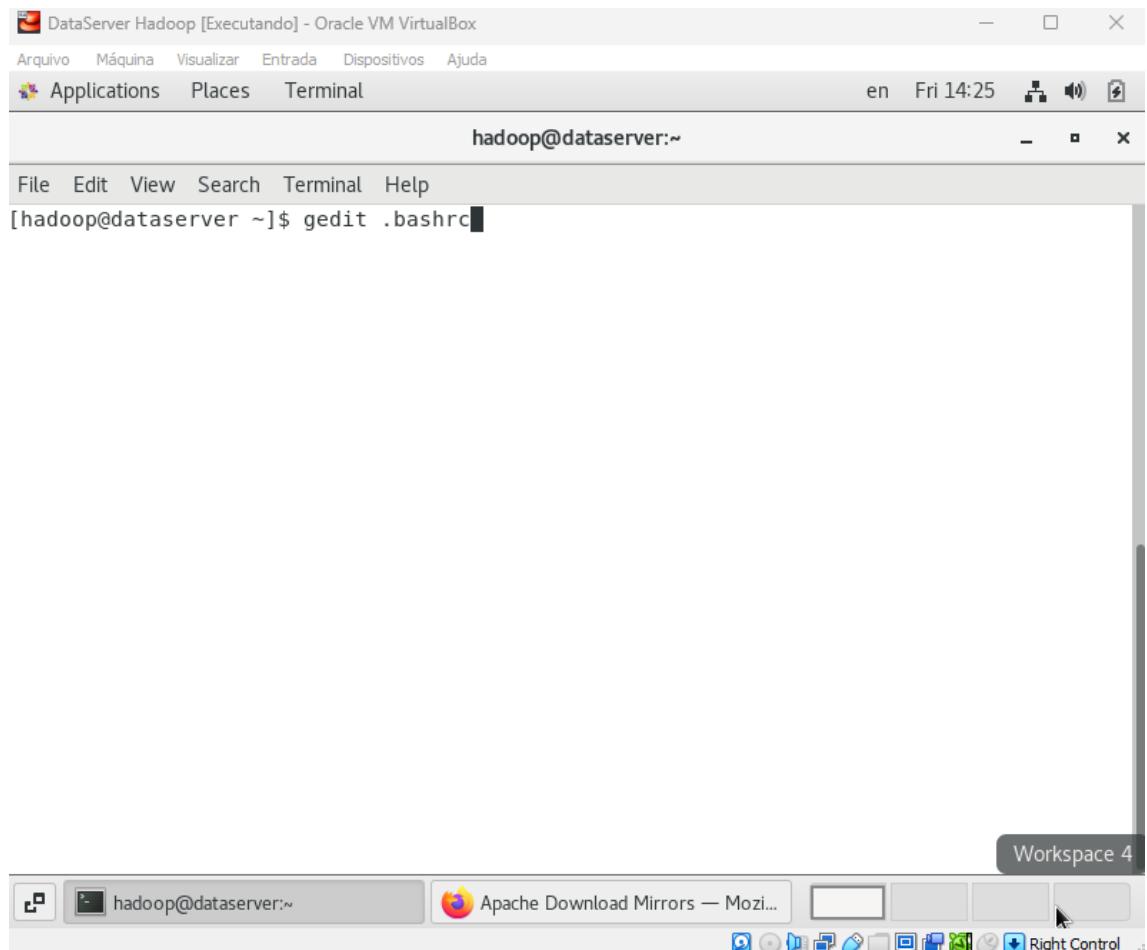
Descompacte o arquivo

## Instalação e Configuração do Ambiente Big Data com Hadoop e Spark



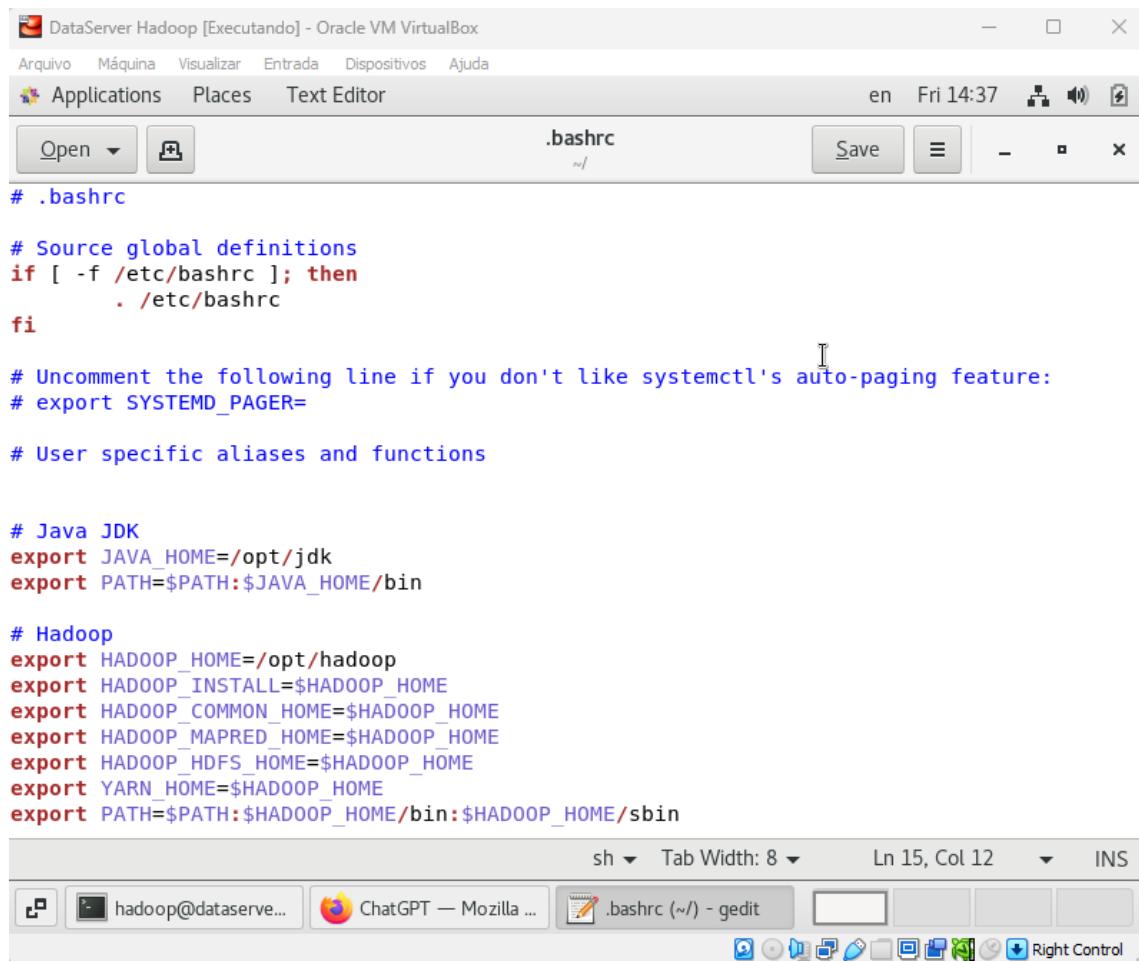
Mover o diretório para /opt/hadoop

## Instalação e Configuração do Ambiente Big Data com Hadoop e Spark



Abrir o arquivo de profile do usuário hadoop

## Instalação e Configuração do Ambiente Big Data com Hadoop e Spark



The screenshot shows a terminal window titled "DataServer Hadoop [Executando] - Oracle VM VirtualBox". The window contains the contents of the .bashrc file. The code includes global definitions for Java and Hadoop, and specific configurations for HDFS and YARN. The terminal interface includes tabs for "sh", "Tab Width: 8", "Ln 15, Col 12", and "INS". Below the terminal are several icons for file operations like copy, paste, and save.

```
# .bashrc

# Source global definitions
if [ -f /etc/bashrc ]; then
    . /etc/bashrc
fi

# Uncomment the following line if you don't like systemctl's auto-paging feature:
# export SYSTEMD_PAGER=

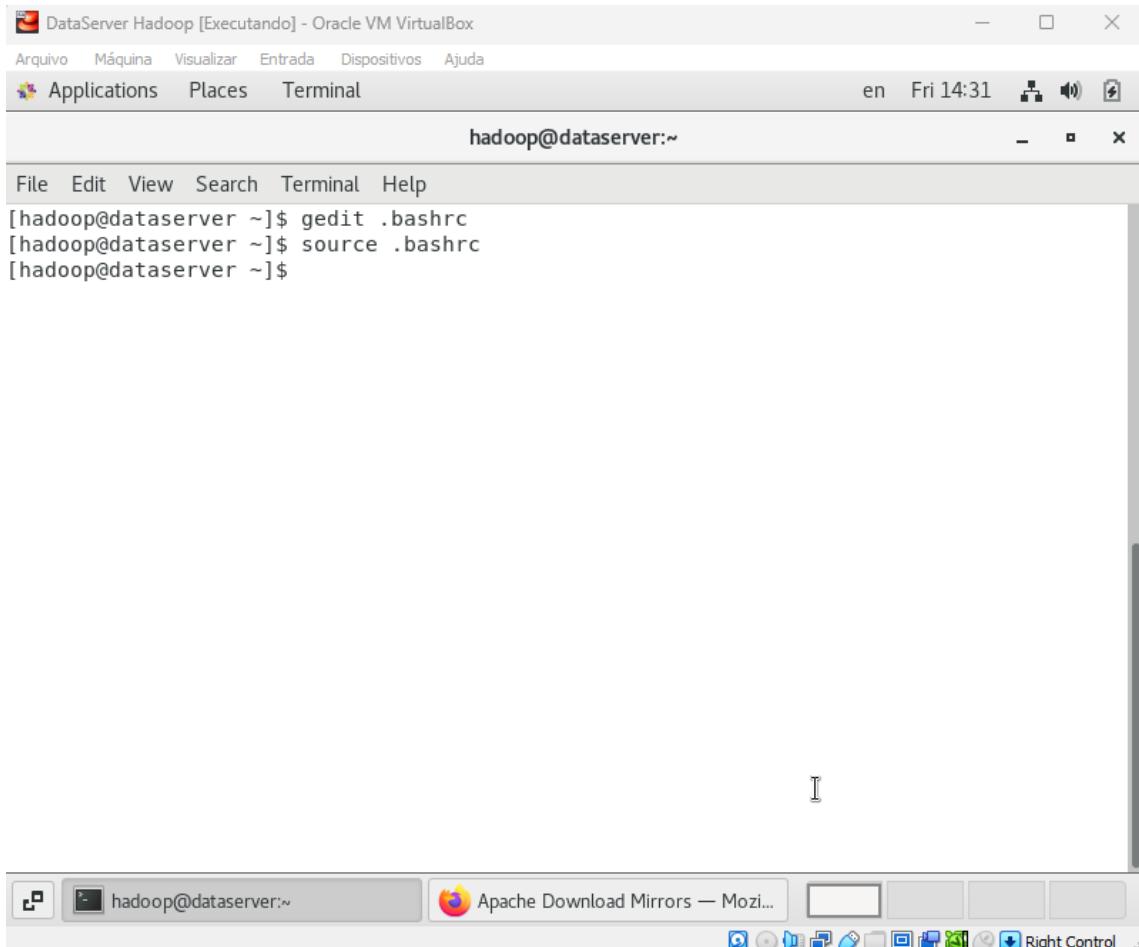
# User specific aliases and functions

# Java JDK
export JAVA_HOME=/opt/jdk
export PATH=$PATH:$JAVA_HOME/bin

# Hadoop
export HADOOP_HOME=/opt/hadoop
export HADOOP_INSTALL=$HADOOP_HOME
export HADOOP_COMMON_HOME=$HADOOP_HOME
export HADOOP_MAPRED_HOME=$HADOOP_HOME
export HADOOP_HDFS_HOME=$HADOOP_HOME
export YARN_HOME=$HADOOP_HOME
export PATH=$PATH:$HADOOP_HOME/bin:$HADOOP_HOME/sbin
```

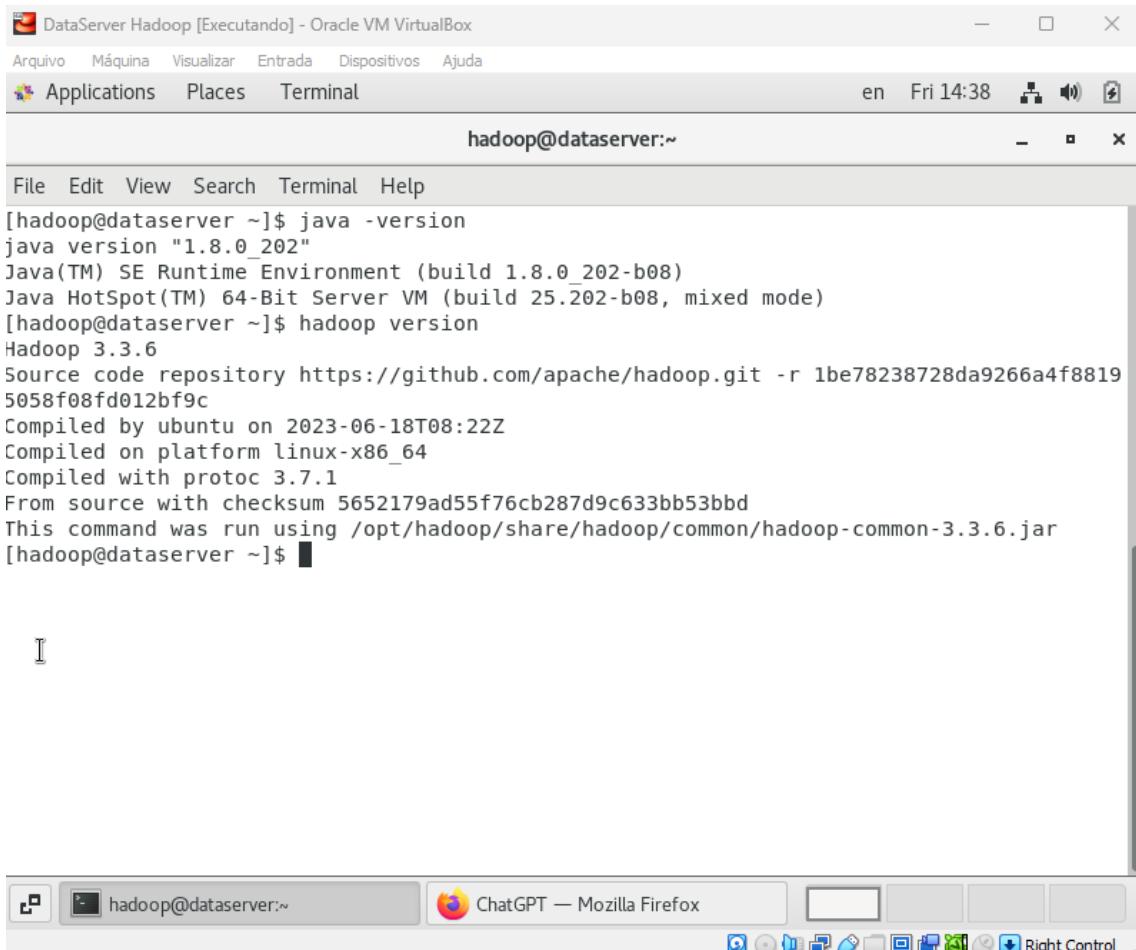
Configure as variáveis de ambiente conforme mostrado na imagem acima  
e salve o arquivo

Instalação e Configuração do Ambiente Big Data com Hadoop e Spark



Execute source .bashrc para efetivar as mudanças das variáveis no SO

# Instalação e Configuração do Ambiente Big Data com Hadoop e Spark



The screenshot shows a Linux desktop environment with a terminal window open. The terminal window title is "DataServer Hadoop [Executando] - Oracle VM VirtualBox". The window contains the following command-line session:

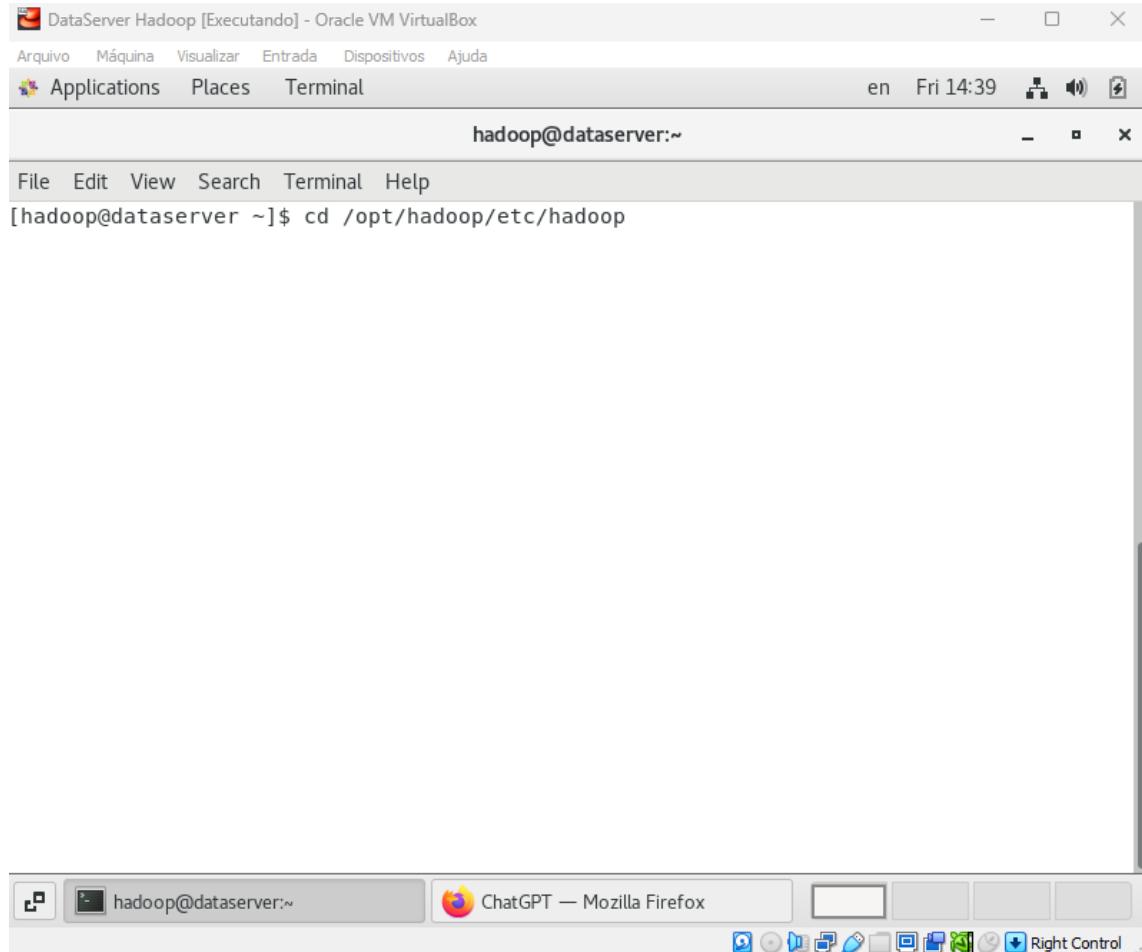
```
[hadoop@dataserver ~]$ java -version
java version "1.8.0_202"
Java(TM) SE Runtime Environment (build 1.8.0_202-b08)
Java HotSpot(TM) 64-Bit Server VM (build 25.202-b08, mixed mode)
[hadoop@dataserver ~]$ hadoop version
Hadoop 3.3.6
Source code repository https://github.com/apache/hadoop.git -r 1be78238728da9266a4f8819
5058f08fd012bf9c
Compiled by ubuntu on 2023-06-18T08:22Z
Compiled on platform linux-x86_64
Compiled with protoc 3.7.1
From source with checksum 5652179ad55f76cb287d9c633bb53bbd
This command was run using /opt/hadoop/share/hadoop/common/hadoop-common-3.3.6.jar
[hadoop@dataserver ~]$ █
```

Below the terminal window, a dock displays several icons, including a terminal icon, a ChatGPT icon, and a Right Control icon.

Java e Hadoop instalados e configurados com sucesso!!!

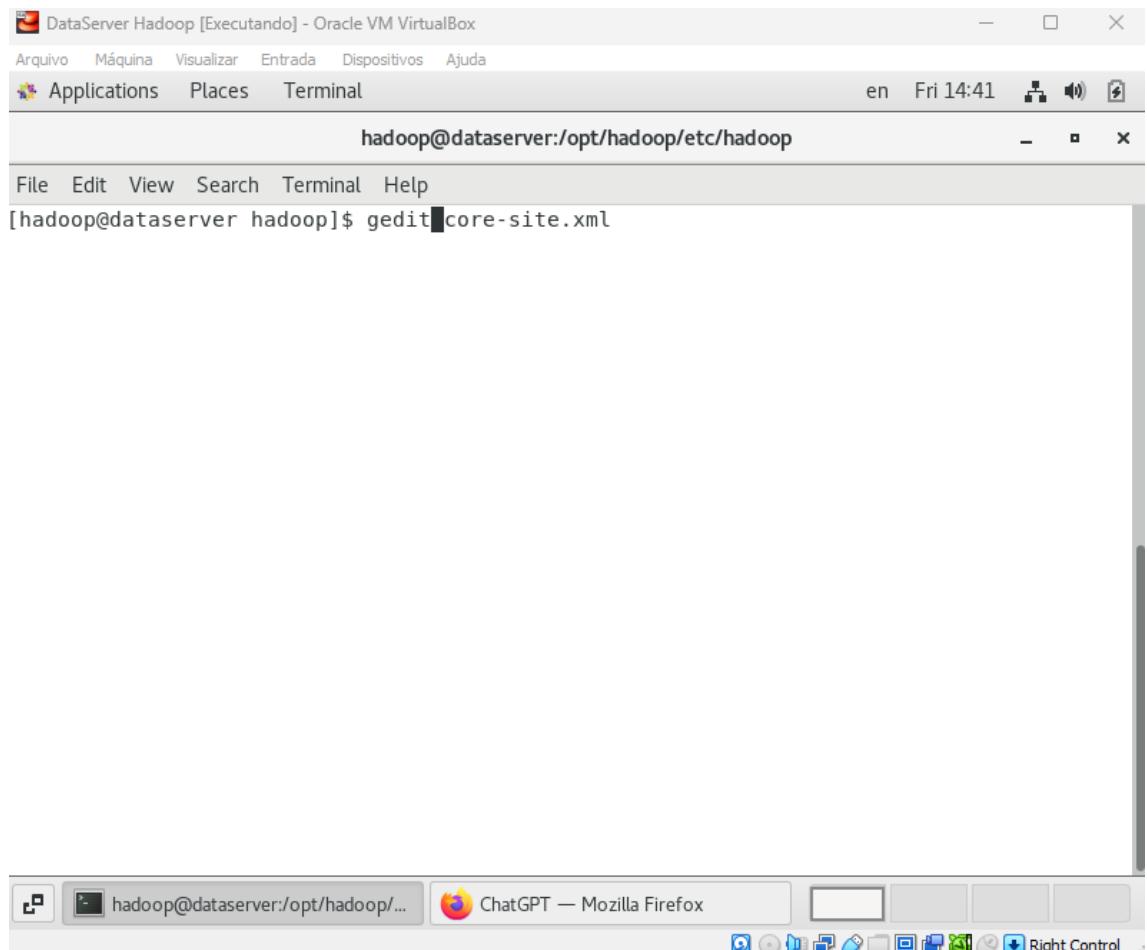
## 5.4 Configuração do Hadoop

### 5.4.1 Editar arquivos de configuração do Hadoop



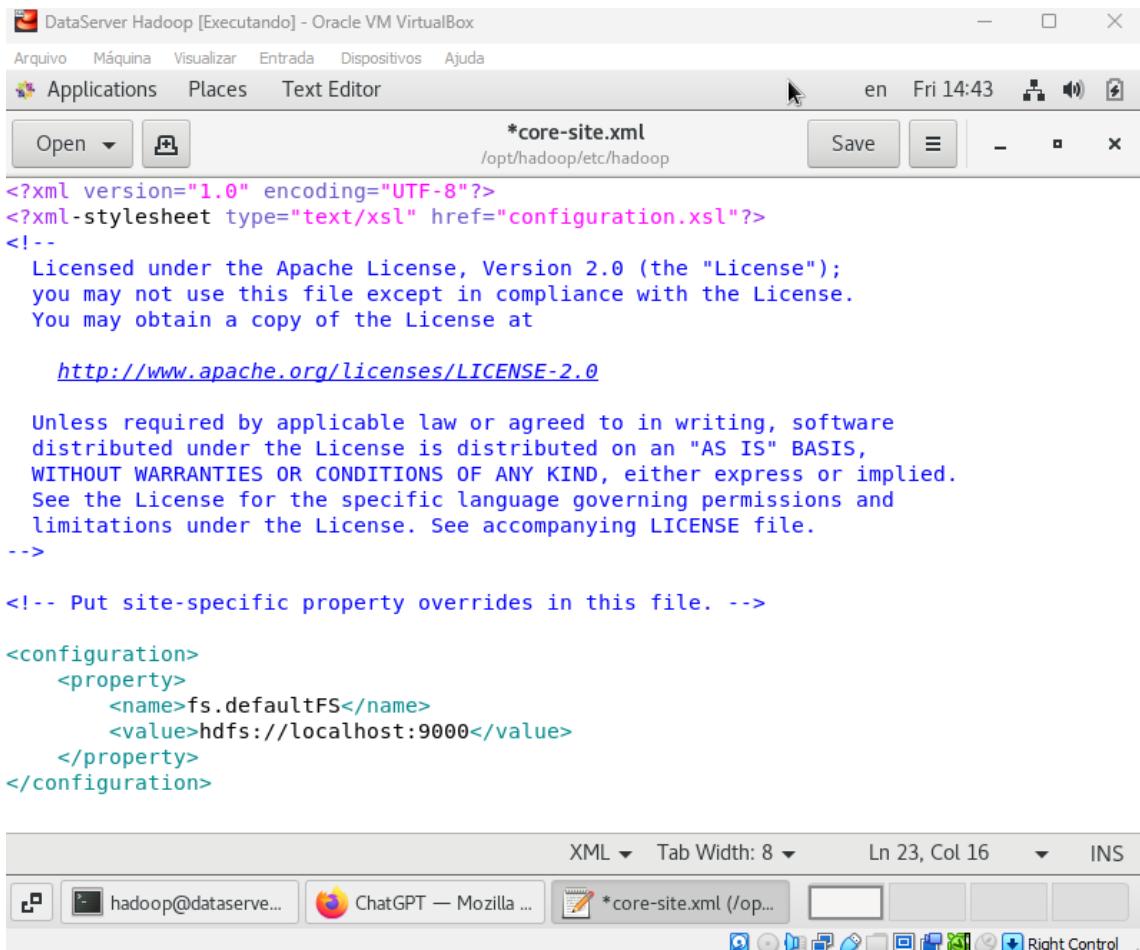
Os arquivos de configuração do Hadoop estão em  
\$HADOOP\_HOME/etc/hadoop  
Nesse caso: /opt/hadoop/etc/hadoop

## Instalação e Configuração do Ambiente Big Data com Hadoop e Spark



Editar o arquivo core-site.xml

## Instalação e Configuração do Ambiente Big Data com Hadoop e Spark



```
*core-site.xml
/opt/hadoop/etc/hadoop
Save
en Fri 14:43
Open    Save
*core-site.xml
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
Licensed under the Apache License, Version 2.0 (the "License");
you may not use this file except in compliance with the License.
You may obtain a copy of the License at

http://www.apache.org/licenses/LICENSE-2.0

Unless required by applicable law or agreed to in writing, software
distributed under the License is distributed on an "AS IS" BASIS,
WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
See the License for the specific language governing permissions and
limitations under the License. See accompanying LICENSE file.
-->

<!-- Put site-specific property overrides in this file. -->

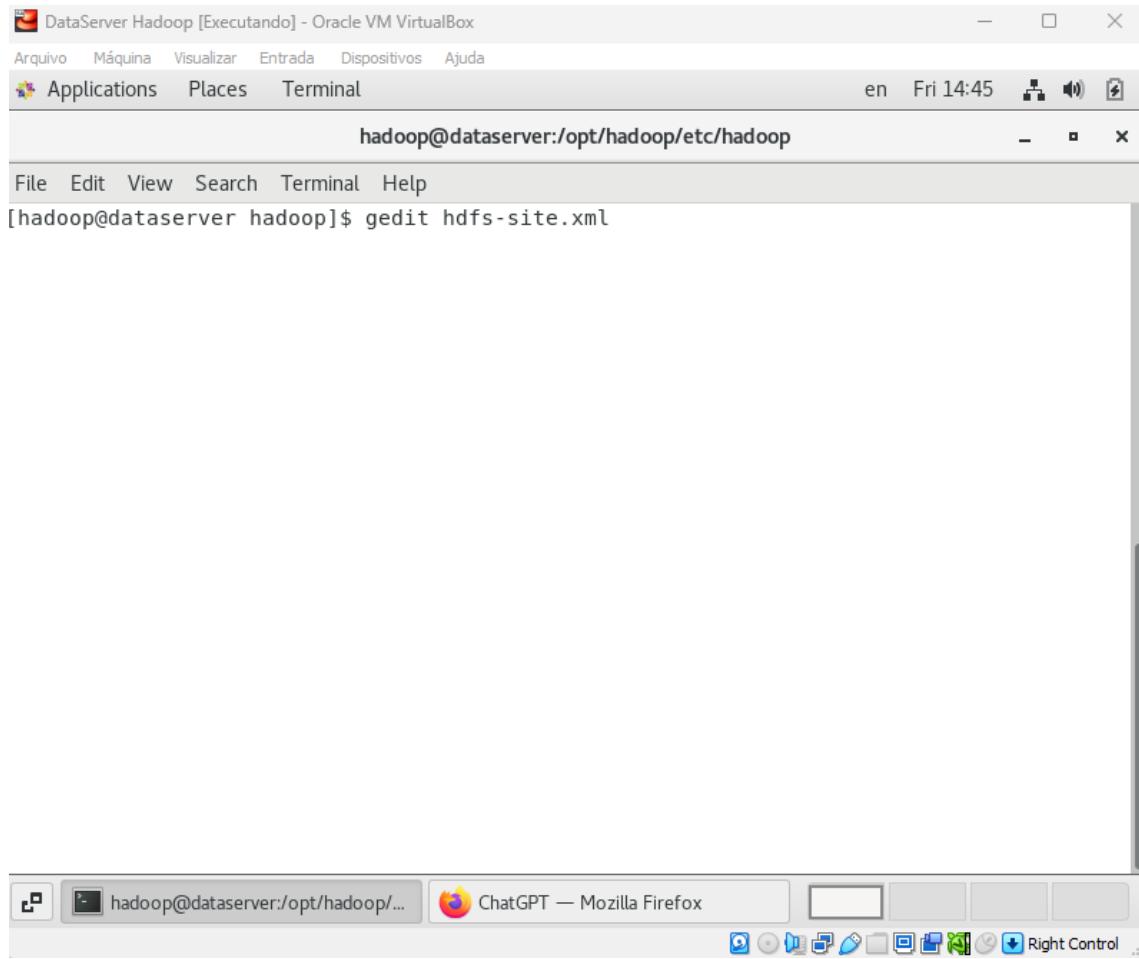
<configuration>
  <property>
    <name>fs.defaultFS</name>
    <value>hdfs://localhost:9000</value>
  </property>
</configuration>
```

XML Tab Width: 8 Ln 23, Col 16 INS

hadoop@dataserve... ChatGPT — Mozilla ... \*core-site.xml (/op... Right Control

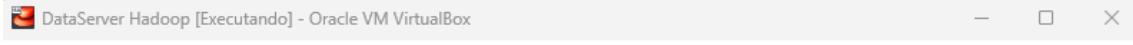
Acrescentar as propriedades conforme acima e salvar o arquivo.  
Essa propriedade indica o endereço do HDFS.

## Instalação e Configuração do Ambiente Big Data com Hadoop e Spark



Editar o arquivo hdfs-site.xml

## Instalação e Configuração do Ambiente Big Data com Hadoop e Spark



DataServer Hadoop [Executando] - Oracle VM VirtualBox

Arquivo Máquina Visualizar Entrada Dispositivos Ajuda

Applications Places Text Editor en Fri 14:47

Open Save \*hdfs-site.xml /opt/hadoop/etc/hadoop

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
Licensed under the Apache License, Version 2.0 (the "License");
you may not use this file except in compliance with the License.
You may obtain a copy of the License at

http://www.apache.org/licenses/LICENSE-2.0

Unless required by applicable law or agreed to in writing, software
distributed under the License is distributed on an "AS IS" BASIS,
WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
See the License for the specific language governing permissions and
limitations under the License. See accompanying LICENSE file.
-->

<!-- Put site-specific property overrides in this file. -->

<configuration>
  <property>
    <name>dfs.replication</name>
    <value>1</value>
  </property>
</configuration>
```

XML Tab Width: 8 Ln 23, Col 14 INS

hadoop@dataserver... ChatGPT — Mozilla ... \*hdfs-site.xml (/op... Right Control

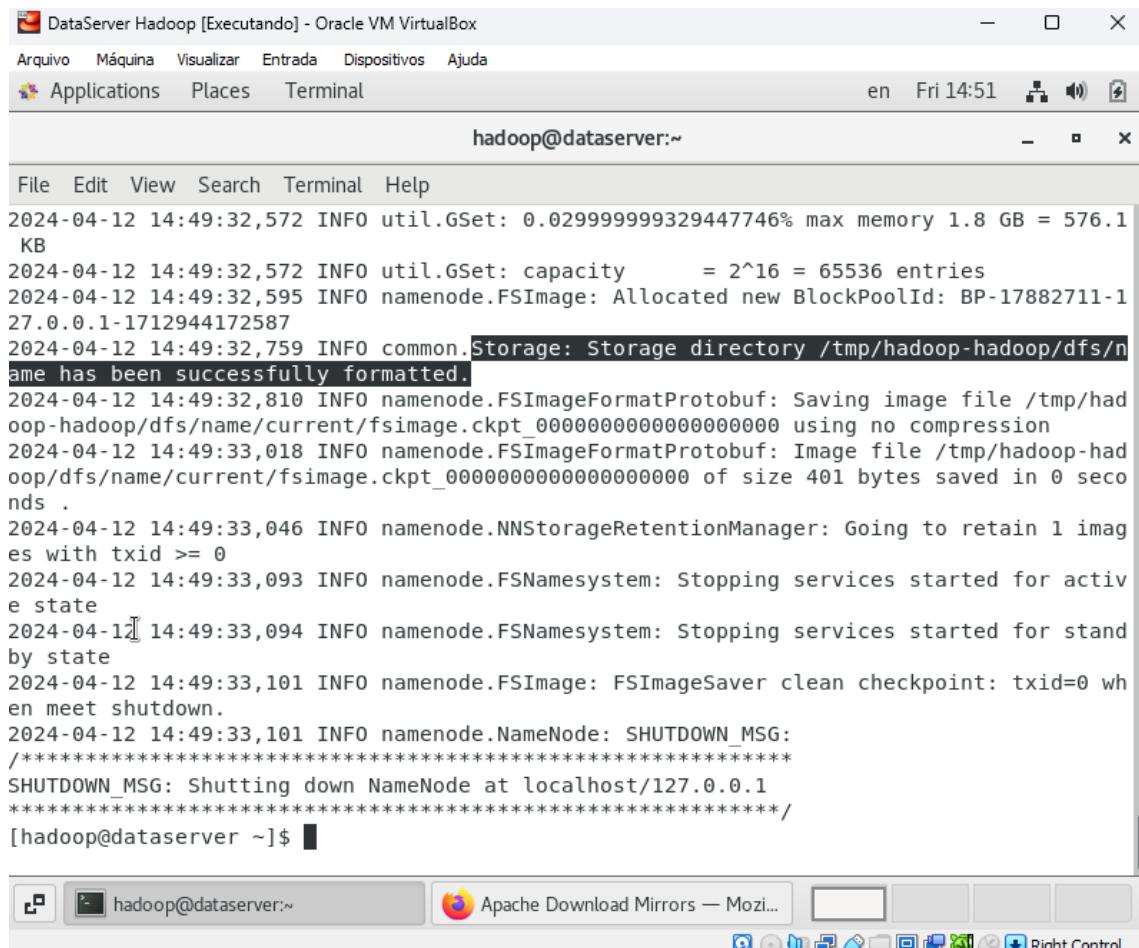
Acrescentar as propriedades conforme acima e salvar o arquivo.

## 5.4.2 Formatando o Namenode

The screenshot shows a Linux desktop environment with a window titled "DataServer Hadoop [Executando] - Oracle VM VirtualBox". The window contains a terminal session for the user "hadoop@dataserver:~". The terminal menu bar includes "File", "Edit", "View", "Search", "Terminal", and "Help". The command entered in the terminal is "hdfs namenode -format". The desktop taskbar at the bottom shows the terminal window and a browser window titled "Apache Download Mirrors — Mozilla Firefox".

hdfs namenode –format

## Instalação e Configuração do Ambiente Big Data com Hadoop e Spark

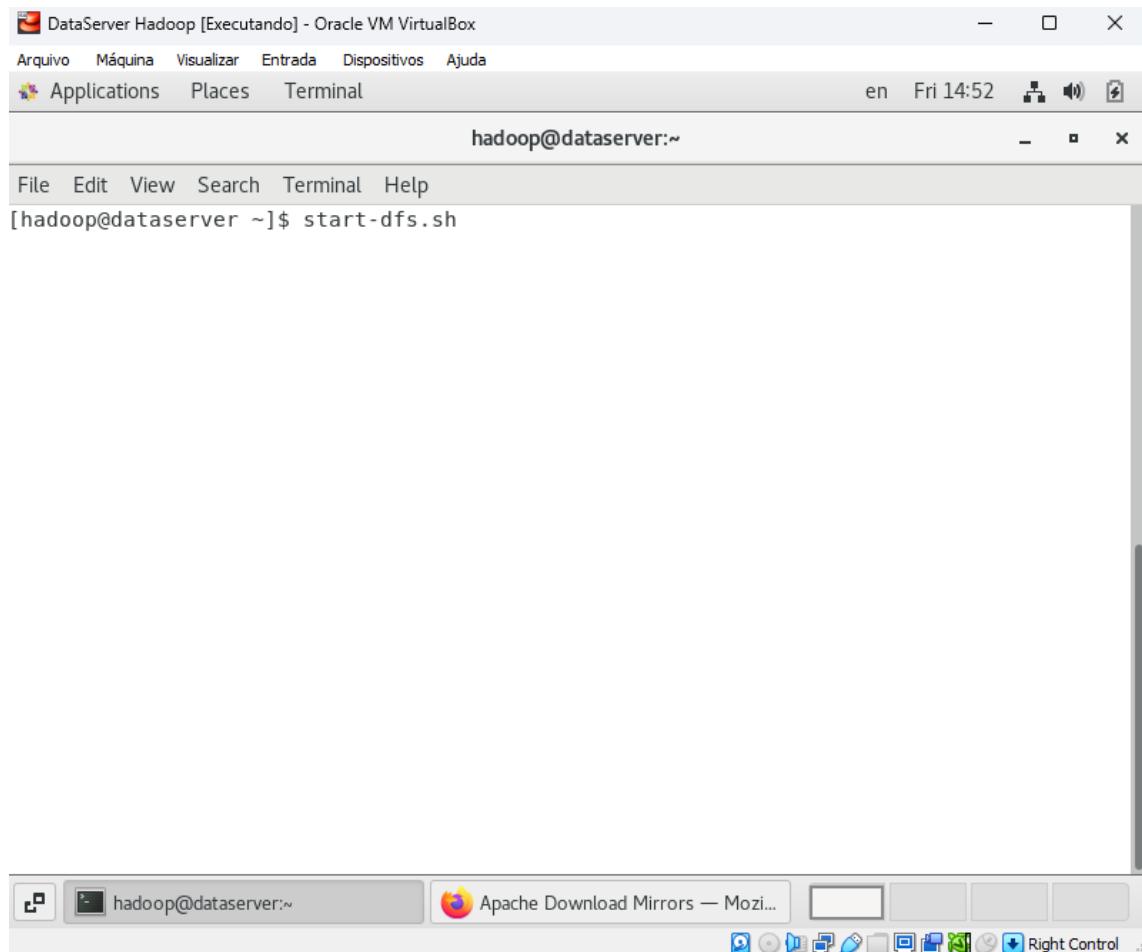


The screenshot shows a terminal window titled "DataServer Hadoop [Executando] - Oracle VM VirtualBox". The window contains a log of Hadoop operations. Key lines from the log include:

```
2024-04-12 14:49:32,572 INFO util.GSet: 0.029999999329447746% max memory 1.8 GB = 576.1 KB
2024-04-12 14:49:32,572 INFO util.GSet: capacity      = 2^16 = 65536 entries
2024-04-12 14:49:32,595 INFO namenode.FSImage: Allocated new BlockPoolId: BP-17882711-1
27.0.0.1-1712944172587
2024-04-12 14:49:32,759 INFO common.Storage: Storage directory /tmp/hadoop-hadoop/dfs/name has been successfully formatted.
2024-04-12 14:49:32,810 INFO namenode.FSImageFormatProtobuf: Saving image file /tmp/hadoop-hadoop/dfs/name/current/fsimage.ckpt_00000000000000000000 using no compression
2024-04-12 14:49:33,018 INFO namenode.FSImageFormatProtobuf: Image file /tmp/hadoop-hadoop/dfs/name/current/fsimage.ckpt_00000000000000000000 of size 401 bytes saved in 0 seconds.
2024-04-12 14:49:33,046 INFO namenode.NNStorageRetentionManager: Going to retain 1 images with txid >= 0
2024-04-12 14:49:33,093 INFO namenode.FSNamesystem: Stopping services started for active state
2024-04-12 14:49:33,094 INFO namenode.FSNamesystem: Stopping services started for standby state
2024-04-12 14:49:33,101 INFO namenode.FSImage: FSImageSaver clean checkpoint: txid=0 when meet shutdown.
2024-04-12 14:49:33,101 INFO namenode.NameNode: SHUTDOWN_MSG:
/*****
SHUTDOWN_MSG: Shutting down NameNode at localhost/127.0.0.1
*****/
[hadoop@dataserver ~]$
```

Formatação realizada com sucesso

### 5.4.3 Iniciando o Hadoop



start-dfs.sh

Instalação e Configuração do Ambiente Big Data com Hadoop e Spark

DataServer Hadoop [Executando] - Oracle VM VirtualBox

Arquivo Máquina Visualizar Entrada Dispositivos Ajuda

Applications Places Terminal en Fri 14:52

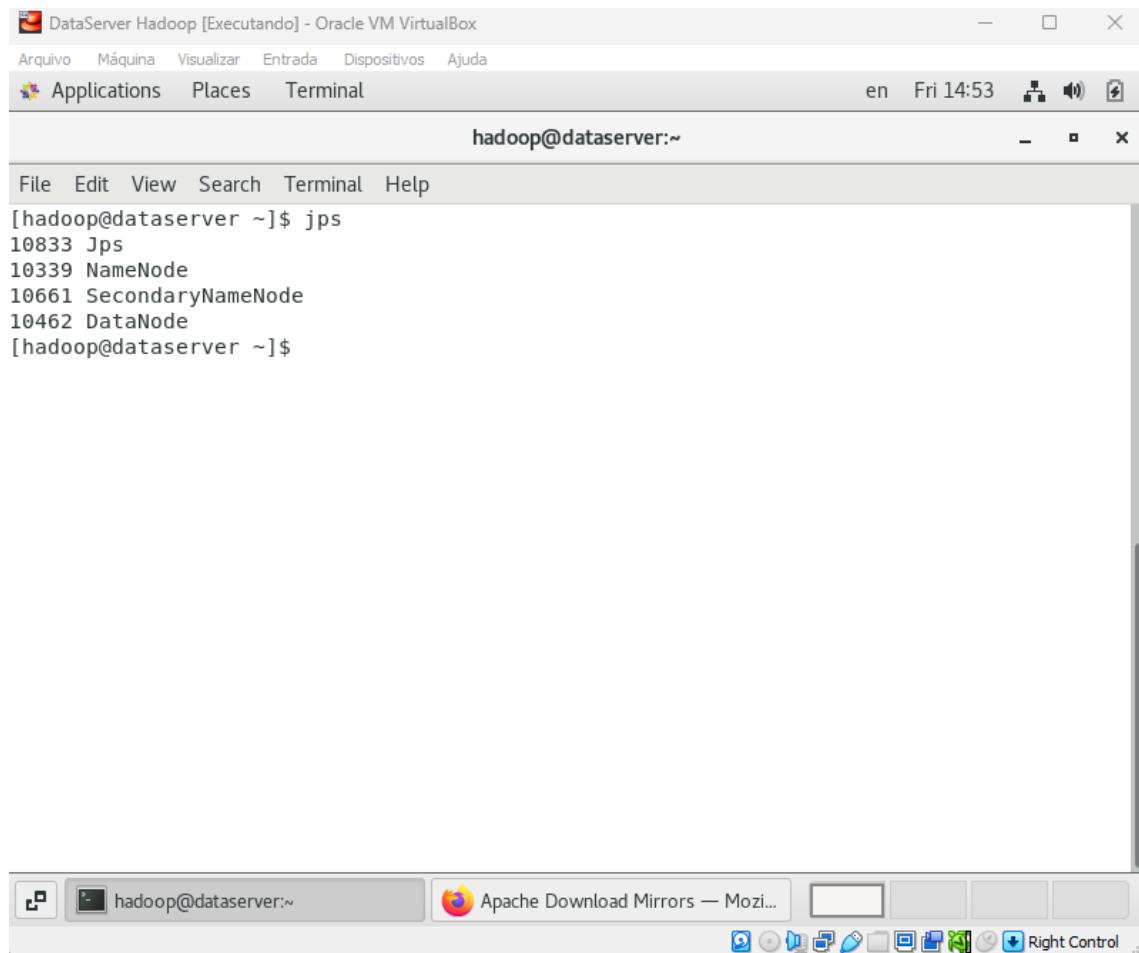
hadoop@dataserver:~

File Edit View Search Terminal Help

```
[hadoop@dataserver ~]$ start-dfs.sh
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [dataserver]
dataserver: Warning: Permanently added 'dataserver' (ECDSA) to the list of known hosts.
[hadoop@dataserver ~]$
```

## Hadoop iniciado

## Instalação e Configuração do Ambiente Big Data com Hadoop e Spark



The screenshot shows a Linux desktop environment with a terminal window and a browser window.

**Terminal Window:**

- Title bar: DataServer Hadoop [Executando] - Oracle VM VirtualBox
- Menu bar: Arquivo, Máquina, Visualizar, Entrada, Dispositivos, Ajuda
- Toolbar: Applications, Places, Terminal
- User: hadoop@dataserver:~
- Date/Time: en Fri 14:53
- Terminal content:

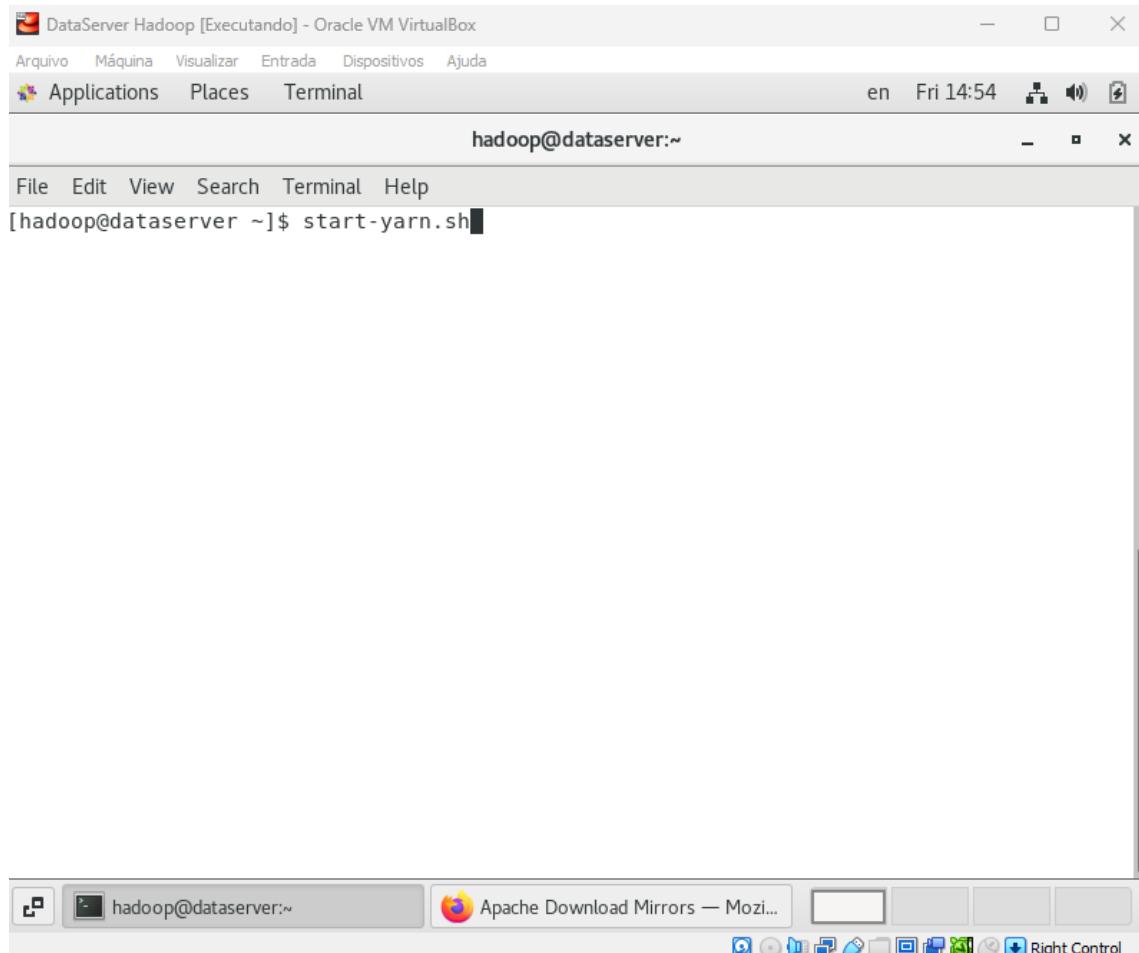
```
File Edit View Search Terminal Help
[hadoop@dataserver ~]$ jps
10833 Jps
10339 NameNode
10661 SecondaryNameNode
10462 DataNode
[hadoop@dataserver ~]$
```

**Browser Window:**

- Title bar: hadoop@dataserver:~
- Address bar: Apache Download Mirrors — Mozilla Firefox
- Toolbar icons: Back, Forward, Stop, Refresh, Home, etc.
- Status bar: Right Control

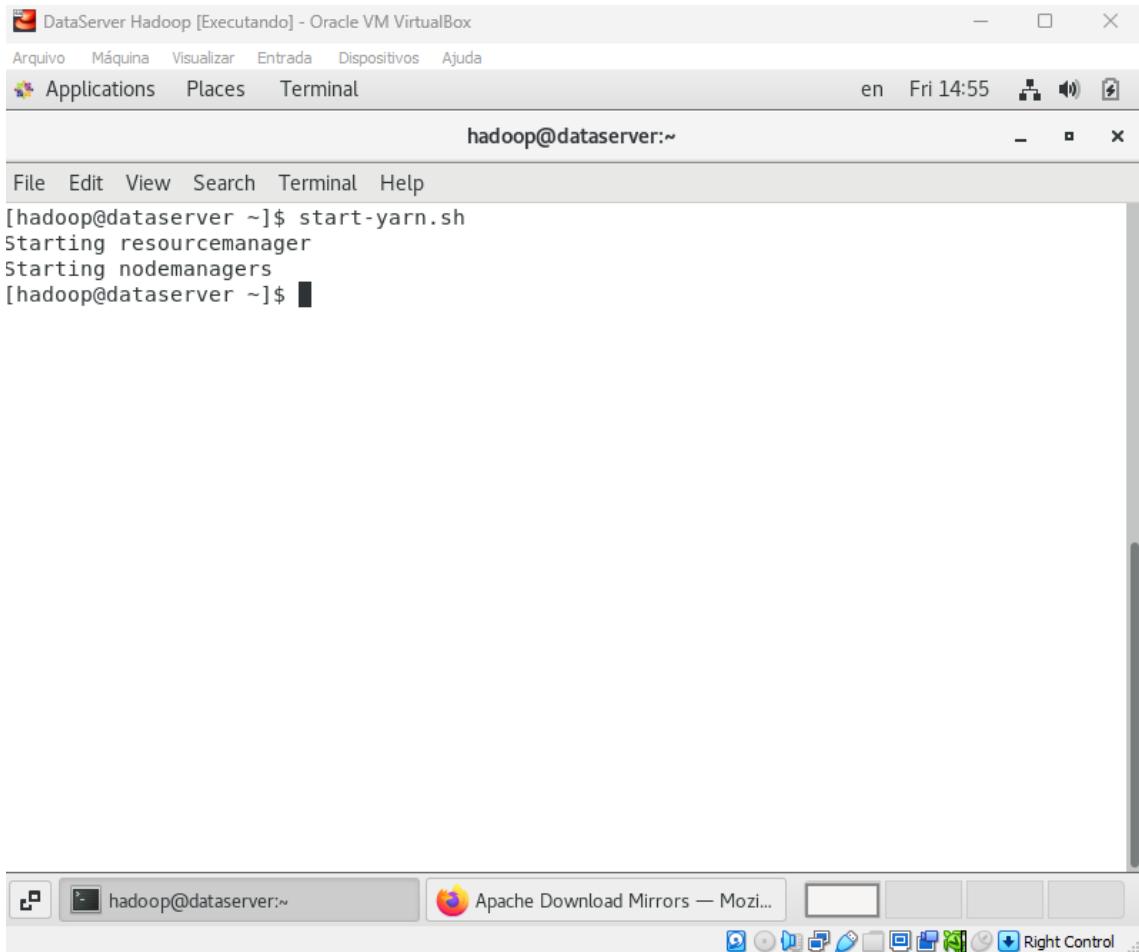
Verificando os serviços inicializados com o comando **jps**

#### 5.4.4 Iniciando o Yarn



start-yarn.sh

## Instalação e Configuração do Ambiente Big Data com Hadoop e Spark



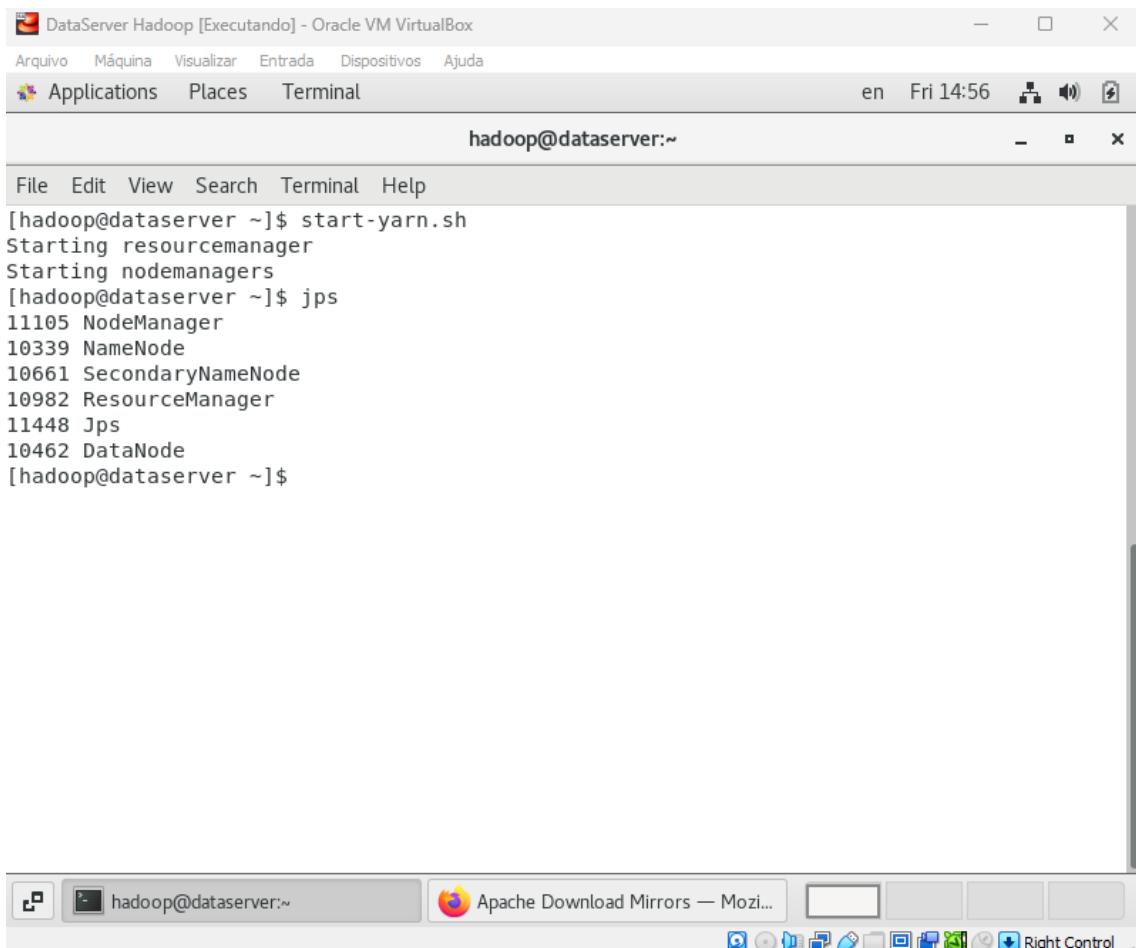
A screenshot of a Linux desktop environment. At the top, there is a menu bar with options like Arquivo, Máquina, Visualizar, Entrada, Dispositivos, Ajuda, Applications, Places, Terminal, and a language indicator (en). The system tray shows the date (Fri 14:55) and some icons. Below the menu is a terminal window titled "hadoop@dataserver:~". The terminal contains the following command and output:

```
File Edit View Search Terminal Help
[hadoop@dataserver ~]$ start-yarn.sh
Starting resourcemanager
Starting nodemanagers
[hadoop@dataserver ~]$
```

Below the terminal is a browser window titled "Apache Download Mirrors — Mozi...". The browser interface includes tabs, a search bar, and various icons.

Yarn iniciado

## Instalação e Configuração do Ambiente Big Data com Hadoop e Spark



The screenshot shows a Linux desktop environment with a terminal window and a browser window.

**Terminal Window:**

```
[hadoop@dataserver ~]$ start-yarn.sh
Starting resourcemanager
Starting nodemanagers
[hadoop@dataserver ~]$ jps
11105 NodeManager
10339 NameNode
10661 SecondaryNameNode
10982 ResourceManager
11448 Jps
10462 DataNode
[hadoop@dataserver ~]$
```

**Browser Window:**

Apache Download Mirrors — Mozilla Firefox

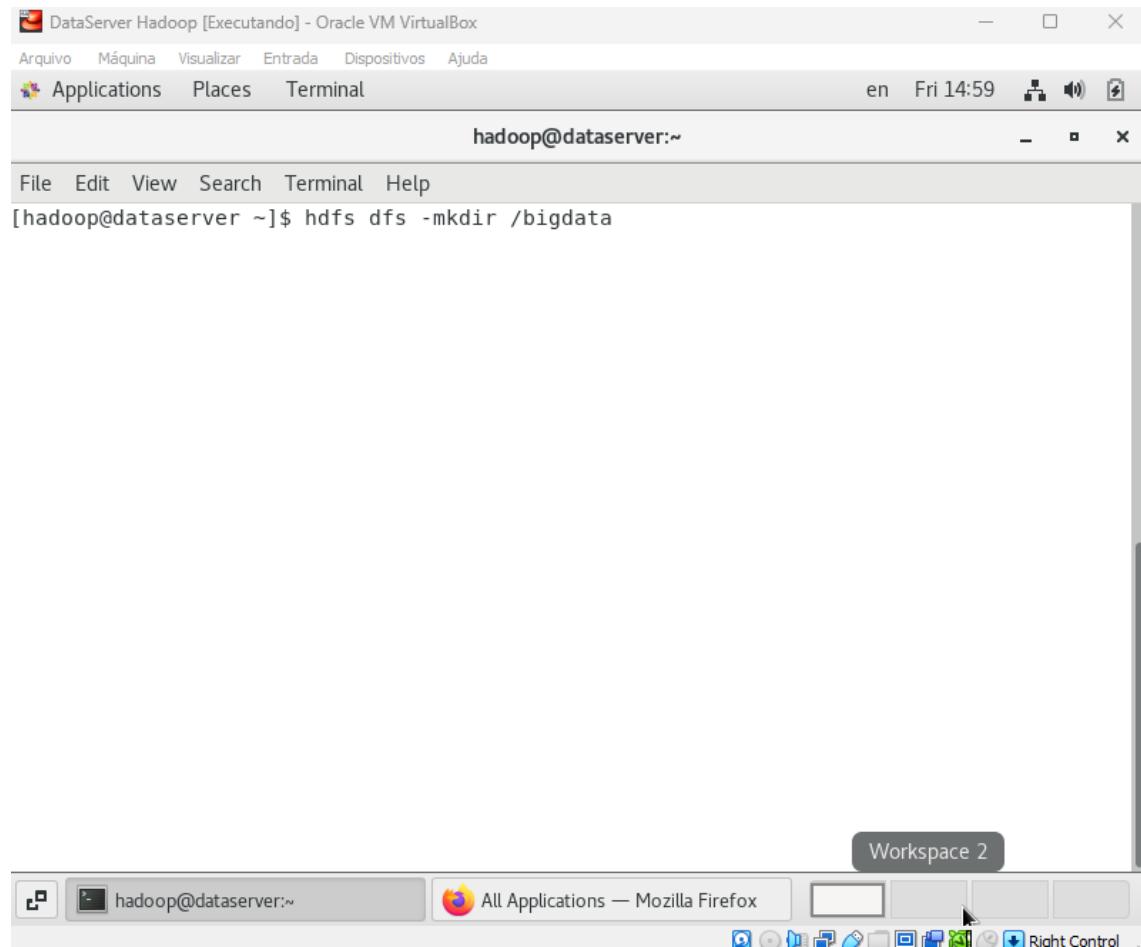
Verificando os serviços com o comando **jps**

## Instalação e Configuração do Ambiente Big Data com Hadoop e Spark

The screenshot shows a Firefox browser window titled "DataServer Hadoop [Executando] - Oracle VM VirtualBox". The address bar displays "localhost:8088/cluster". The page content is the Hadoop cluster metrics interface. On the left, there is a sidebar with a tree view under "Cluster" containing "About", "Nodes", "Node Labels", "Applications" (with sub-options: NEW, NEW\_SAVING, SUBMITTED, ACCEPTED, RUNNING, FINISHED, FAILED, KILLED), and "Scheduler". Below this is a "Tools" button. The main area has three sections: "Cluster Metrics" (showing 0 Apps Submitted, 0 Apps Pending, 0 Apps Running), "Cluster Nodes Metrics" (showing 1 Active Nodes, 0 Decommissioning Nodes), and "Scheduler Metrics" (showing Capacity Scheduler with scheduling results in memory-mb (unit=Mi), vcores). A "Show 20 entries" dropdown is present. At the bottom, the status bar shows "hadoop@dataserver:~" and "All Applications — Mozilla Firefox".

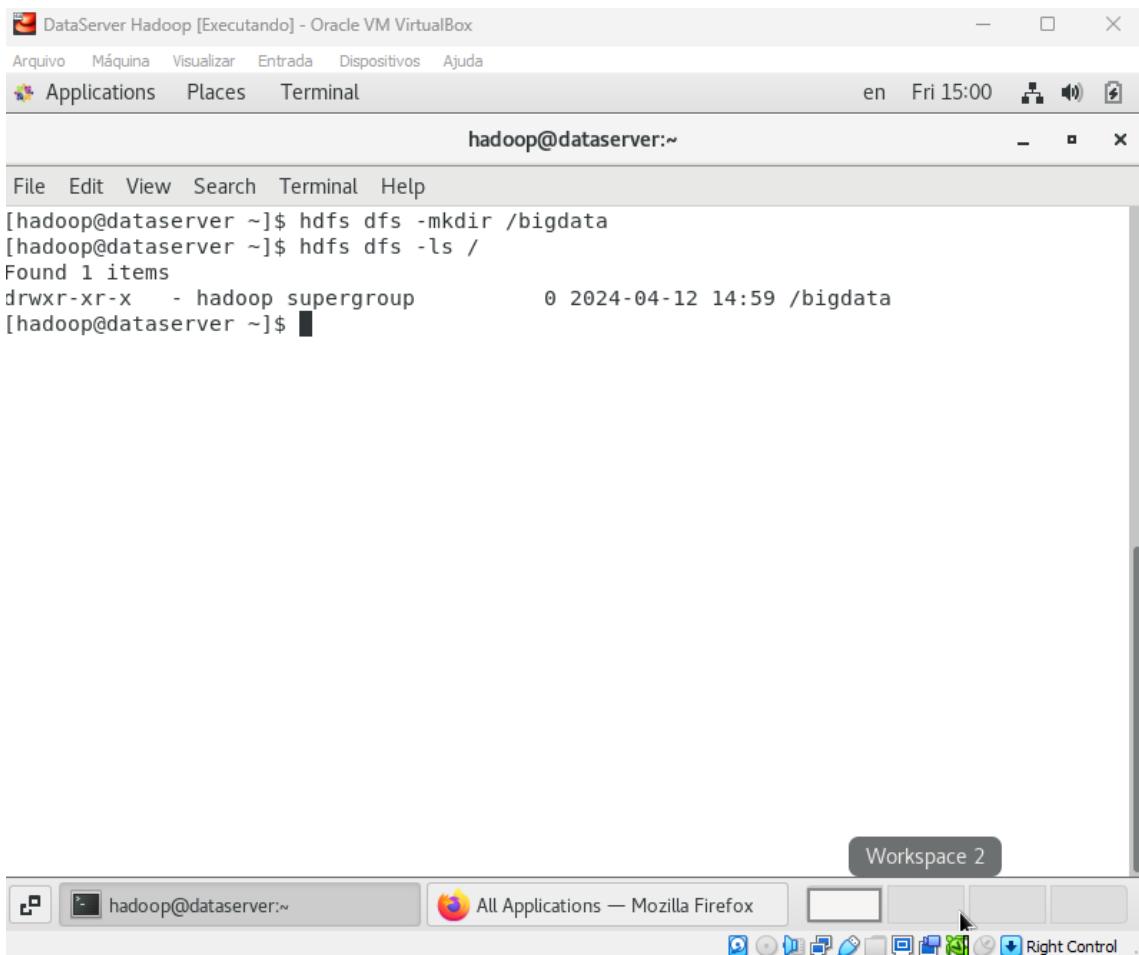
Visualizando jobs – <http://localhost:8088>

## 5.5 Processando Big Data



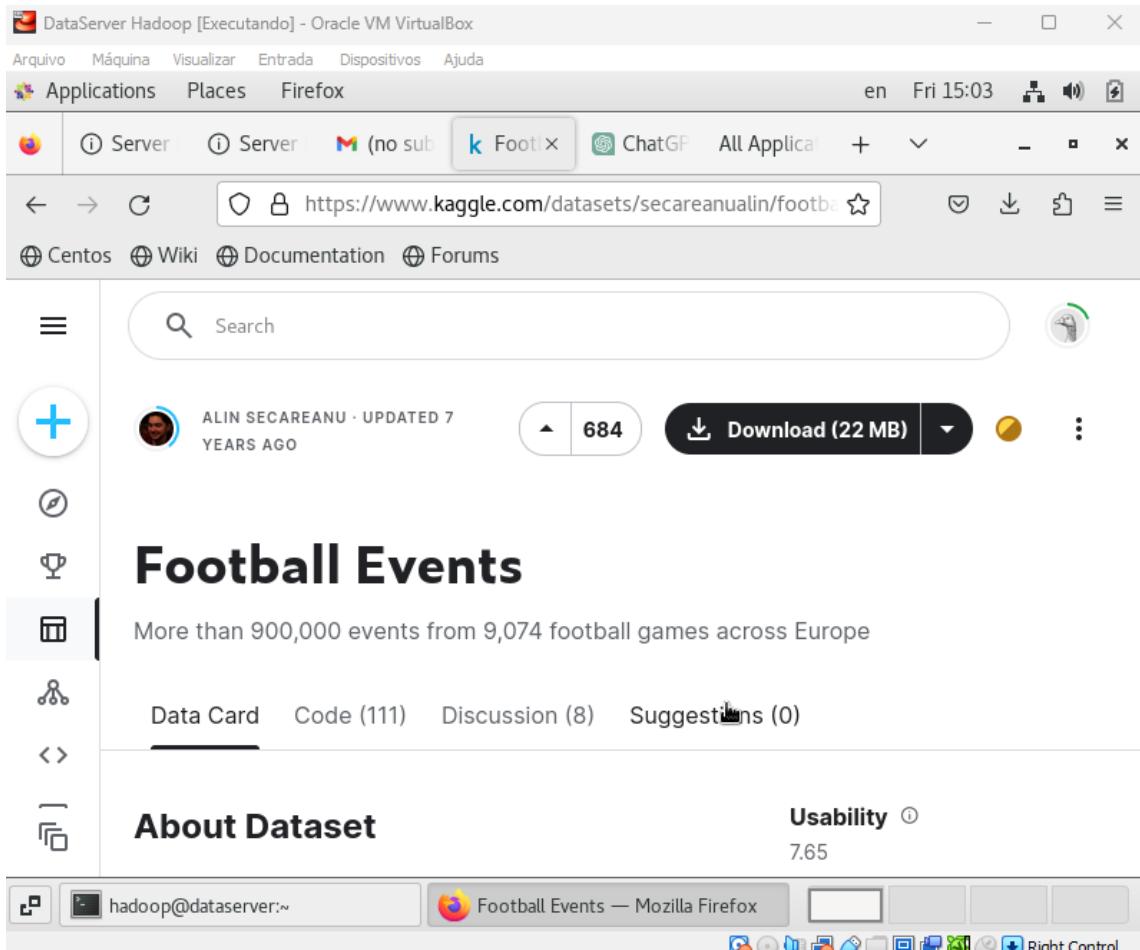
Criar o diretório **bigdata** no HDFS

## Instalação e Configuração do Ambiente Big Data com Hadoop e Spark



Listar o HDFS e verificar o diretório criado

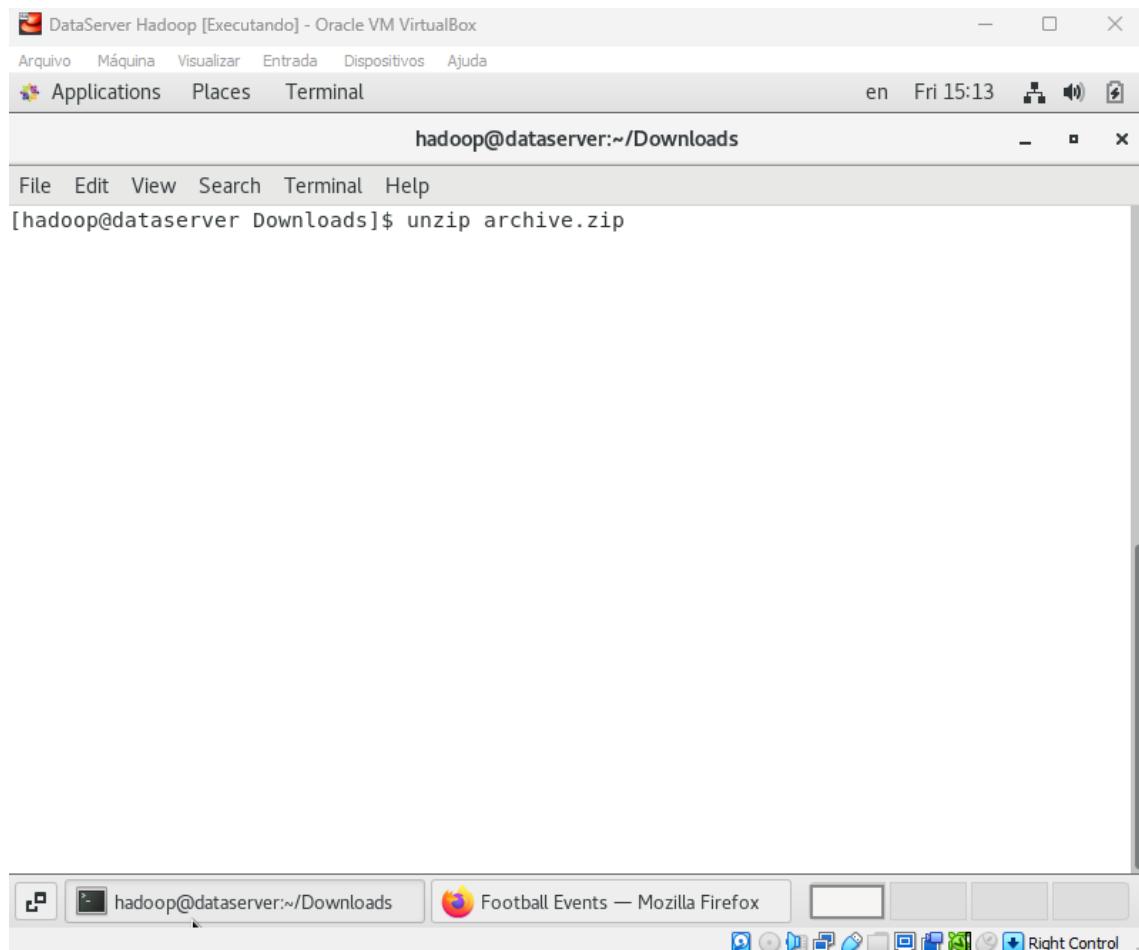
## Instalação e Configuração do Ambiente Big Data com Hadoop e Spark



Acessar o portal Kaggle e faça download do arquivo archive.zip

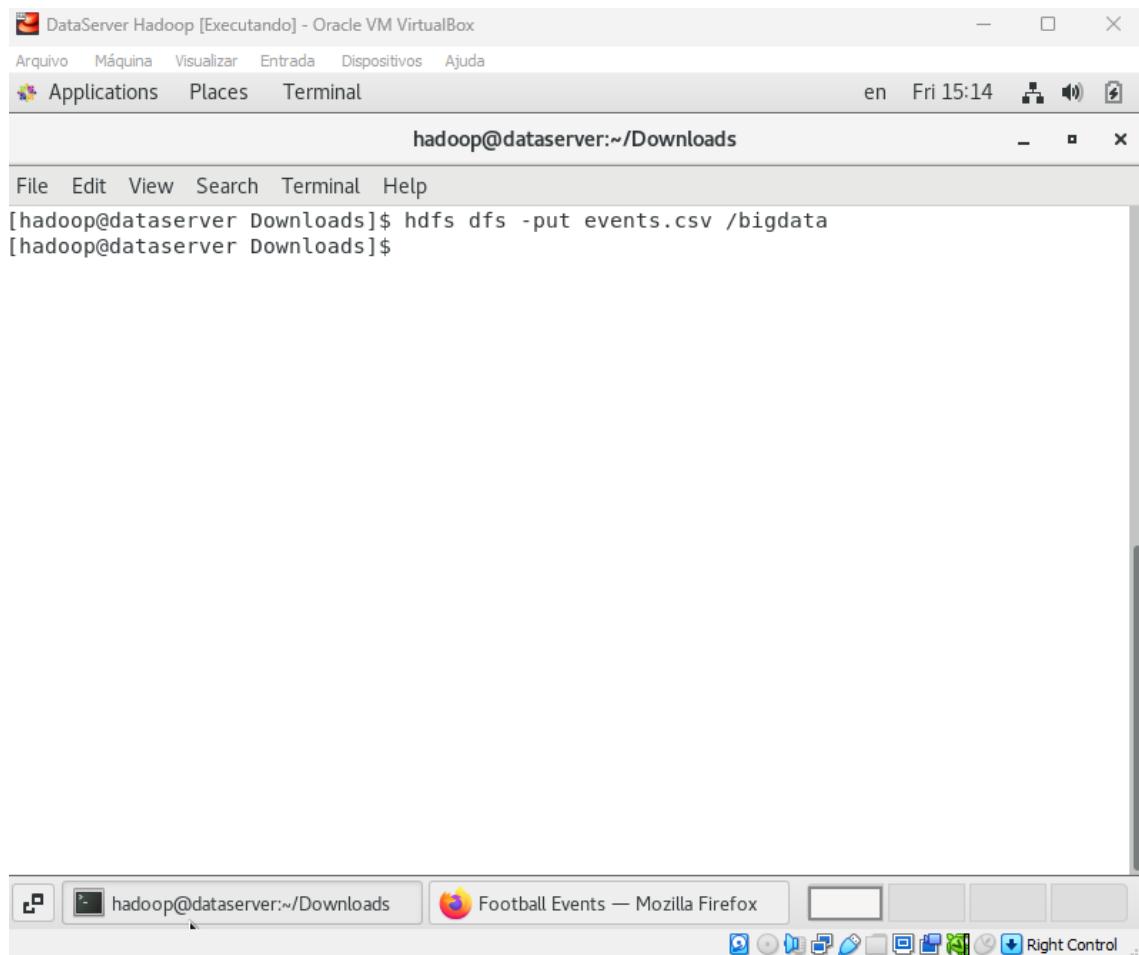
<https://www.kaggle.com/datasets/secareanulin/football-events?resource=download>

## Instalação e Configuração do Ambiente Big Data com Hadoop e Spark



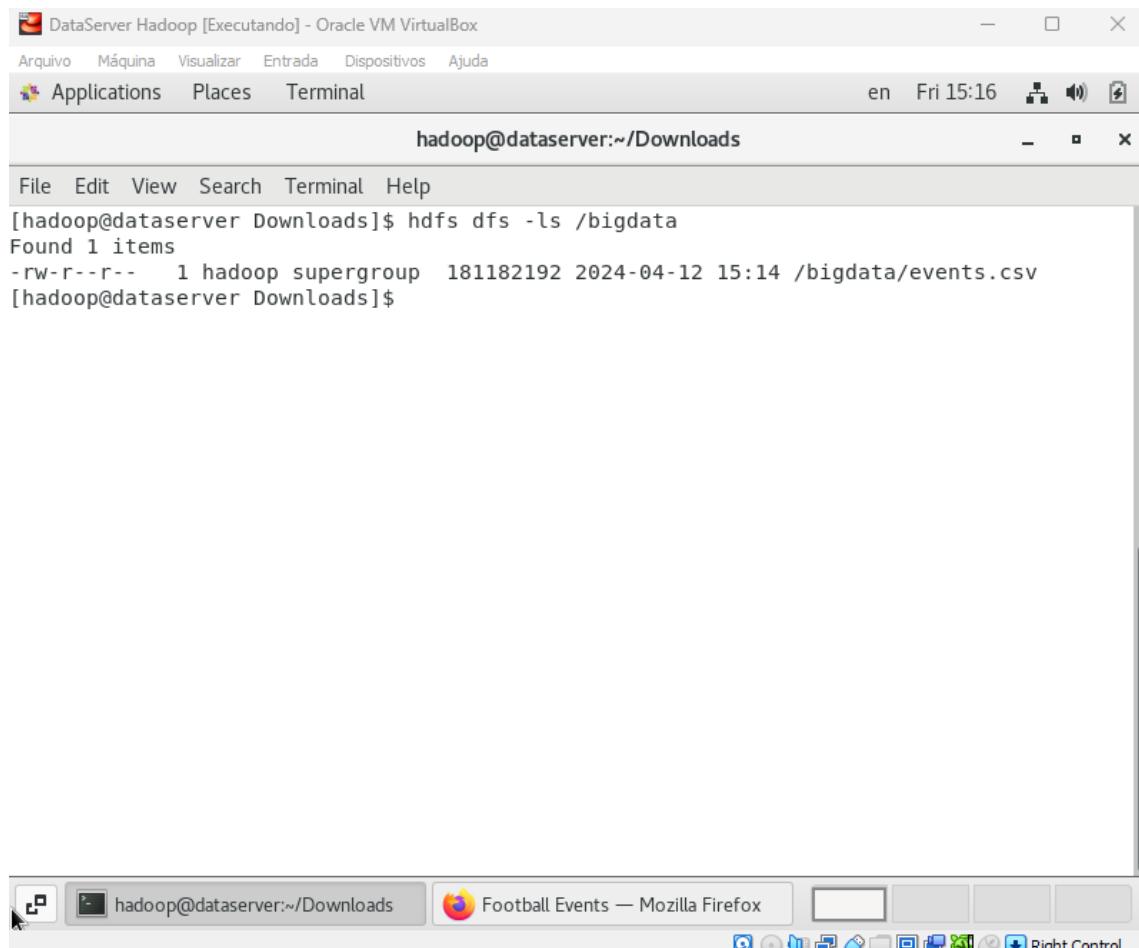
Descompacte-o

## Instalação e Configuração do Ambiente Big Data com Hadoop e Spark



Copiar o arquivo para a pasta diretório no HDFS

## Instalação e Configuração do Ambiente Big Data com Hadoop e Spark

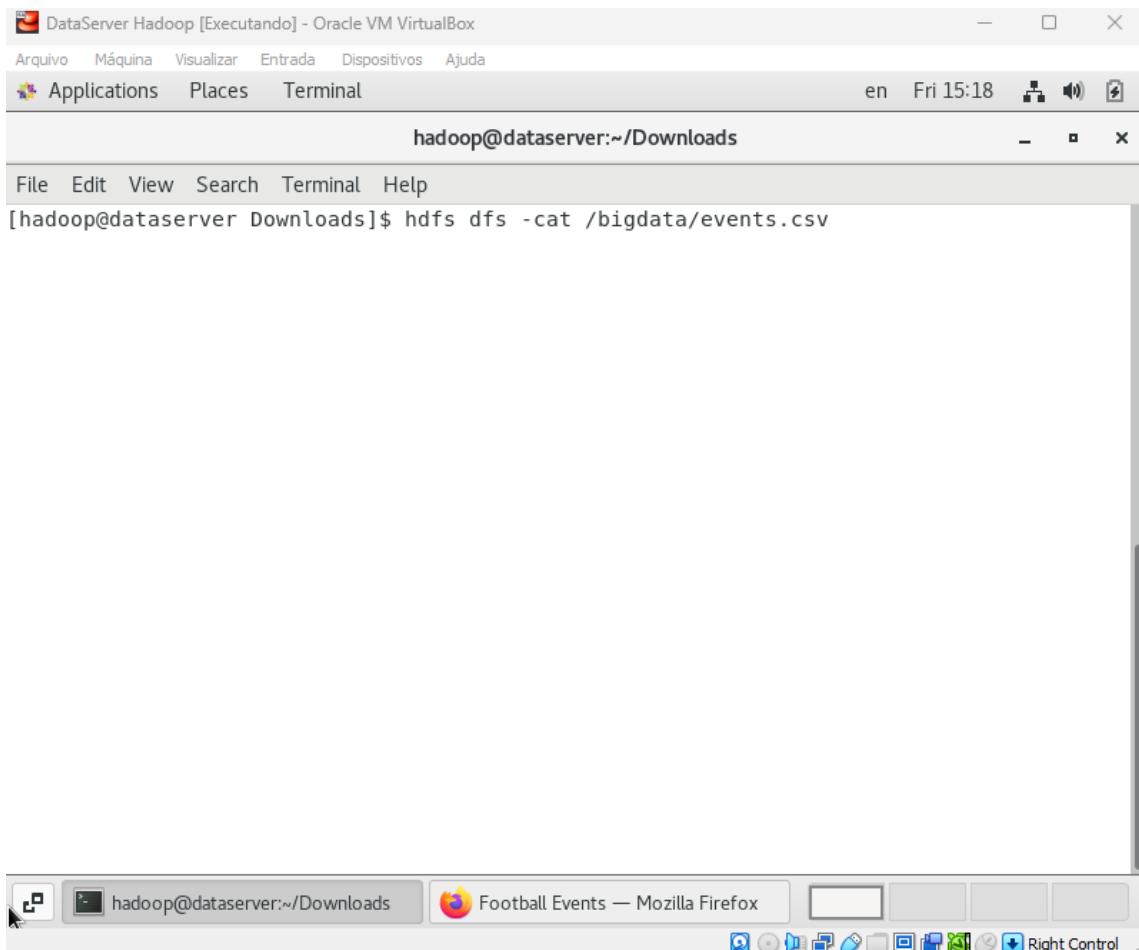


A screenshot of a Linux desktop environment. At the top, there is a menu bar with options like Arquivo, Máquina, Visualizar, Entrada, Dispositivos, and Ajuda. Below the menu bar is a toolbar with Applications, Places, and Terminal. The system tray shows language (en), date (Fri 15:16), and other icons. The main window is a terminal titled "hadoop@dataserver:~/Downloads". The terminal displays the command `hdfs dfs -ls /bigdata` and its output: "Found 1 items" followed by a file entry. The taskbar at the bottom has several icons, including one for the terminal window and another for Mozilla Firefox.

```
[hadoop@dataserver Downloads]$ hdfs dfs -ls /bigdata
Found 1 items
-rw-r--r-- 1 hadoop supergroup 181182192 2024-04-12 15:14 /bigdata/events.csv
[hadoop@dataserver Downloads]$
```

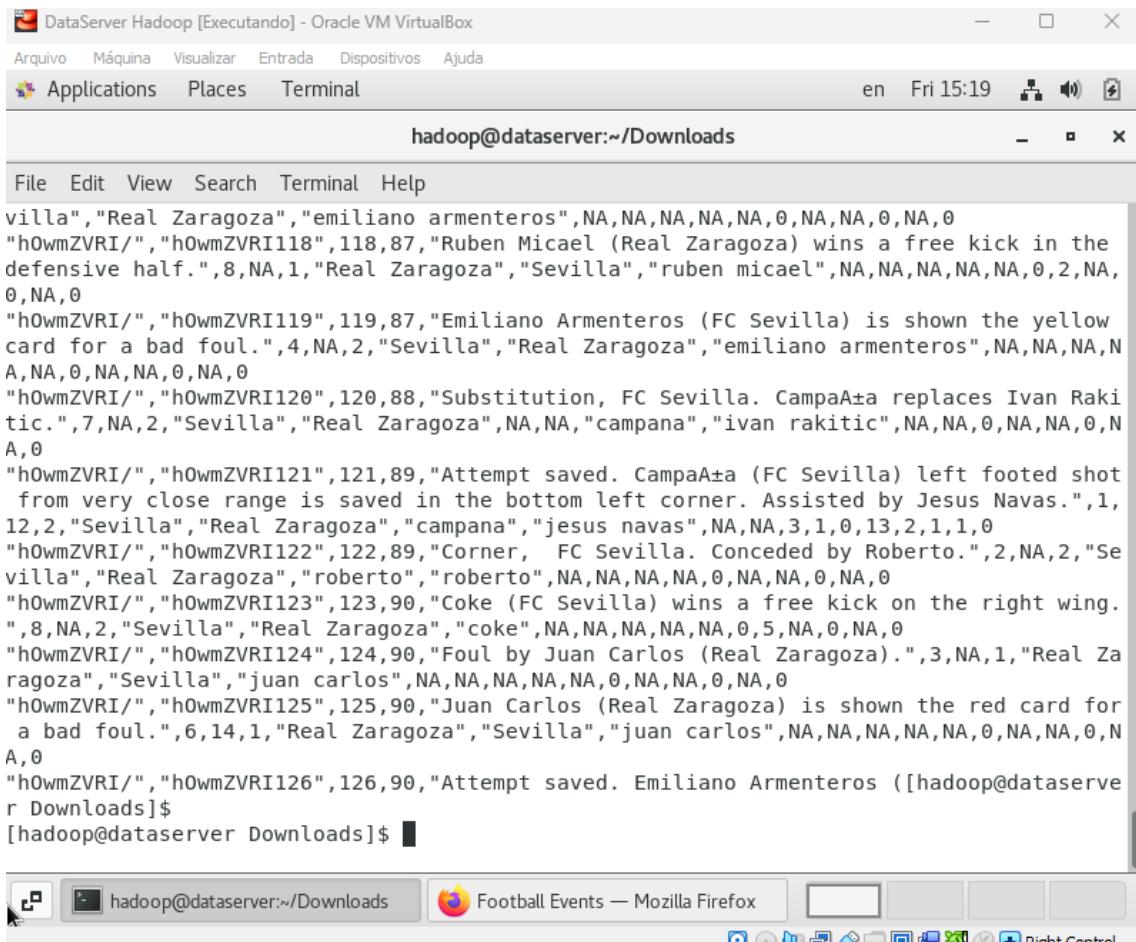
Listar o diretório bigdata

## Instalação e Configuração do Ambiente Big Data com Hadoop e Spark



Ver o conteúdo do arquivo

## Instalação e Configuração do Ambiente Big Data com Hadoop e Spark



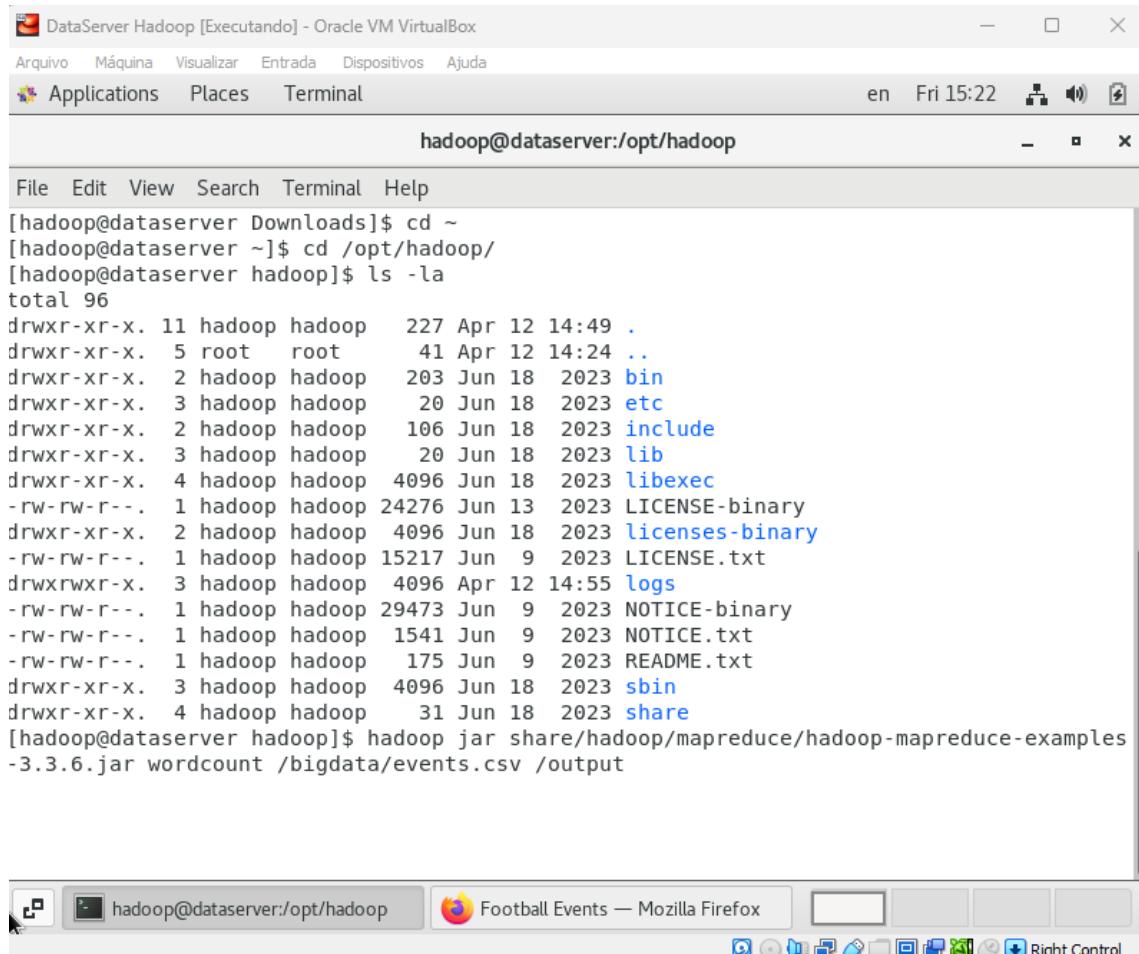
The screenshot shows a Linux desktop environment. In the foreground, a terminal window titled "hadoop@dataserver:~/Downloads" displays a log of football events. The log entries are as follows:

```
villa", "Real Zaragoza", "emiliano armenteros", NA, NA, NA, NA, NA, 0, NA, NA, 0, NA, 0  
"hOwmZVRI/", "hOwmZVRI118", 118, 87, "Ruben Micael (Real Zaragoza) wins a free kick in the  
defensive half.", 8, NA, 1, "Real Zaragoza", "Sevilla", "ruben micael", NA, NA, NA, NA, 0, 2, NA,  
0, NA, 0  
"hOwmZVRI/", "hOwmZVRI119", 119, 87, "Emiliano Armenteros (FC Sevilla) is shown the yellow  
card for a bad foul.", 4, NA, 2, "Sevilla", "Real Zaragoza", "emiliano armenteros", NA, NA, NA, N  
A, NA, 0, NA, NA, 0, NA, 0  
"hOwmZVRI/", "hOwmZVRI120", 120, 88, "Substitution, FC Sevilla. CampaA±a replaces Ivan Rakitic.", 7, NA, 2, "Sevilla", "Real Zaragoza", NA, NA, "campana", "ivan rakitic", NA, NA, 0, NA, NA, 0, N  
A, 0  
"hOwmZVRI/", "hOwmZVRI121", 121, 89, "Attempt saved. CampaA±a (FC Sevilla) left footed shot  
from very close range is saved in the bottom left corner. Assisted by Jesus Navas.", 1, 12, 2, "Sevilla", "Real Zaragoza", "campana", "jesus navas", NA, NA, 3, 1, 0, 13, 2, 1, 1, 0  
"hOwmZVRI/", "hOwmZVRI122", 122, 89, "Corner, FC Sevilla. Conceded by Roberto.", 2, NA, 2, "Se  
villa", "Real Zaragoza", "roberto", "roberto", NA, NA, NA, NA, 0, NA, NA, 0, NA, 0  
"hOwmZVRI/", "hOwmZVRI123", 123, 90, "Coke (FC Sevilla) wins a free kick on the right wing.  
, 8, NA, 2, "Sevilla", "Real Zaragoza", "coke", NA, NA, NA, NA, 0, 5, NA, 0, NA, 0  
"hOwmZVRI/", "hOwmZVRI124", 124, 90, "Foul by Juan Carlos (Real Zaragoza).", 3, NA, 1, "Real Za  
ragoza", "Sevilla", "juan carlos", NA, NA, NA, NA, 0, NA, NA, 0, NA, 0  
"hOwmZVRI/", "hOwmZVRI125", 125, 90, "Juan Carlos (Real Zaragoza) is shown the red card for  
a bad foul.", 6, 14, 1, "Real Zaragoza", "Sevilla", "juan carlos", NA, NA, NA, NA, 0, NA, NA, 0, N  
A, 0  
"hOwmZVRI/", "hOwmZVRI126", 126, 90, "Attempt saved. Emiliano Armenteros ([hadoop@dataserv  
er Downloads]$  
[hadoop@dataserver Downloads]$
```

In the background, a Firefox browser window titled "Football Events — Mozilla Firefox" is visible, showing a blank page.

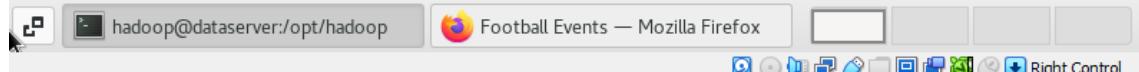
Conteúdo do arquivo já gravado no HDFS

## Instalação e Configuração do Ambiente Big Data com Hadoop e Spark



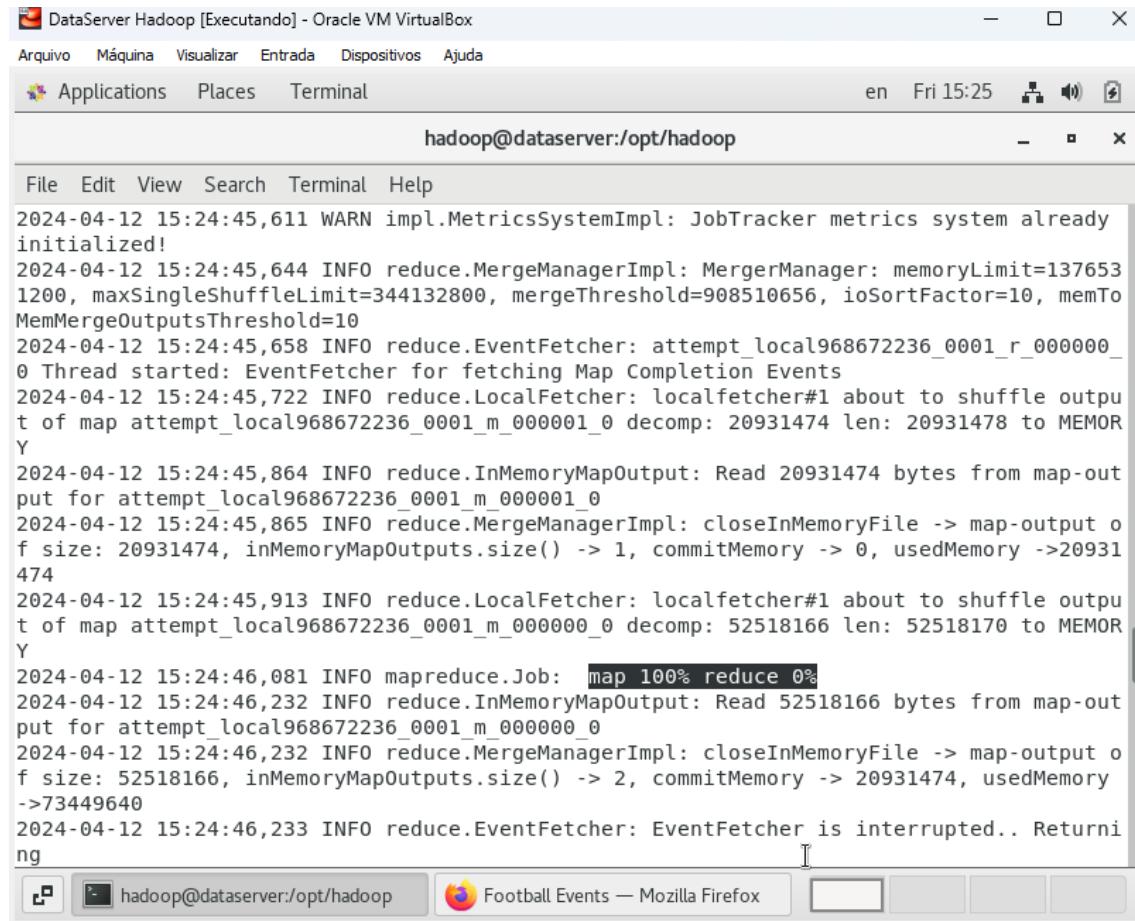
The screenshot shows a terminal window titled "DataServer Hadoop [Executando] - Oracle VM VirtualBox". The window contains the following command-line session:

```
[hadoop@dataserver Downloads]$ cd ~  
[hadoop@dataserver ~]$ cd /opt/hadoop/  
[hadoop@dataserver hadoop]$ ls -la  
total 96  
drwxr-xr-x. 11 hadoop hadoop 227 Apr 12 14:49 .  
drwxr-xr-x.  5 root  root   41 Apr 12 14:24 ..  
drwxr-xr-x.  2 hadoop hadoop 203 Jun 18 2023 bin  
drwxr-xr-x.  3 hadoop hadoop  20 Jun 18 2023 etc  
drwxr-xr-x.  2 hadoop hadoop 106 Jun 18 2023 include  
drwxr-xr-x.  3 hadoop hadoop  20 Jun 18 2023 lib  
drwxr-xr-x.  4 hadoop hadoop 4096 Jun 18 2023 libexec  
-rw-rw-r--.  1 hadoop hadoop 24276 Jun 13 2023 LICENSE-binary  
drwxr-xr-x.  2 hadoop hadoop 4096 Jun 18 2023 licenses-binary  
-rw-rw-r--.  1 hadoop hadoop 15217 Jun  9 2023 LICENSE.txt  
drwxrwxr-x.  3 hadoop hadoop 4096 Apr 12 14:55 logs  
-rw-rw-r--.  1 hadoop hadoop 29473 Jun  9 2023 NOTICE-binary  
-rw-rw-r--.  1 hadoop hadoop 1541 Jun  9 2023 NOTICE.txt  
-rw-rw-r--.  1 hadoop hadoop 175 Jun  9 2023 README.txt  
drwxr-xr-x.  3 hadoop hadoop 4096 Jun 18 2023 sbin  
drwxr-xr-x.  4 hadoop hadoop 31 Jun 18 2023 share  
[hadoop@dataserver hadoop]$ hadoop jar share/hadoop/mapreduce/hadoop-mapreduce-examples-3.3.6.jar wordcount /bigdata/events.csv /output
```



A instalação do Hadoop possui um job chamado wordcount, que pode ser usado como exemplo para processamento de Big Data. Basicamente, o job conta a ocorrência de cada palavra no arquivo. Vou executar com o comando acima (a versão do arquivo .jar é a mesma versão do Hadoop que eu instalei).

# Instalação e Configuração do Ambiente Big Data com Hadoop e Spark

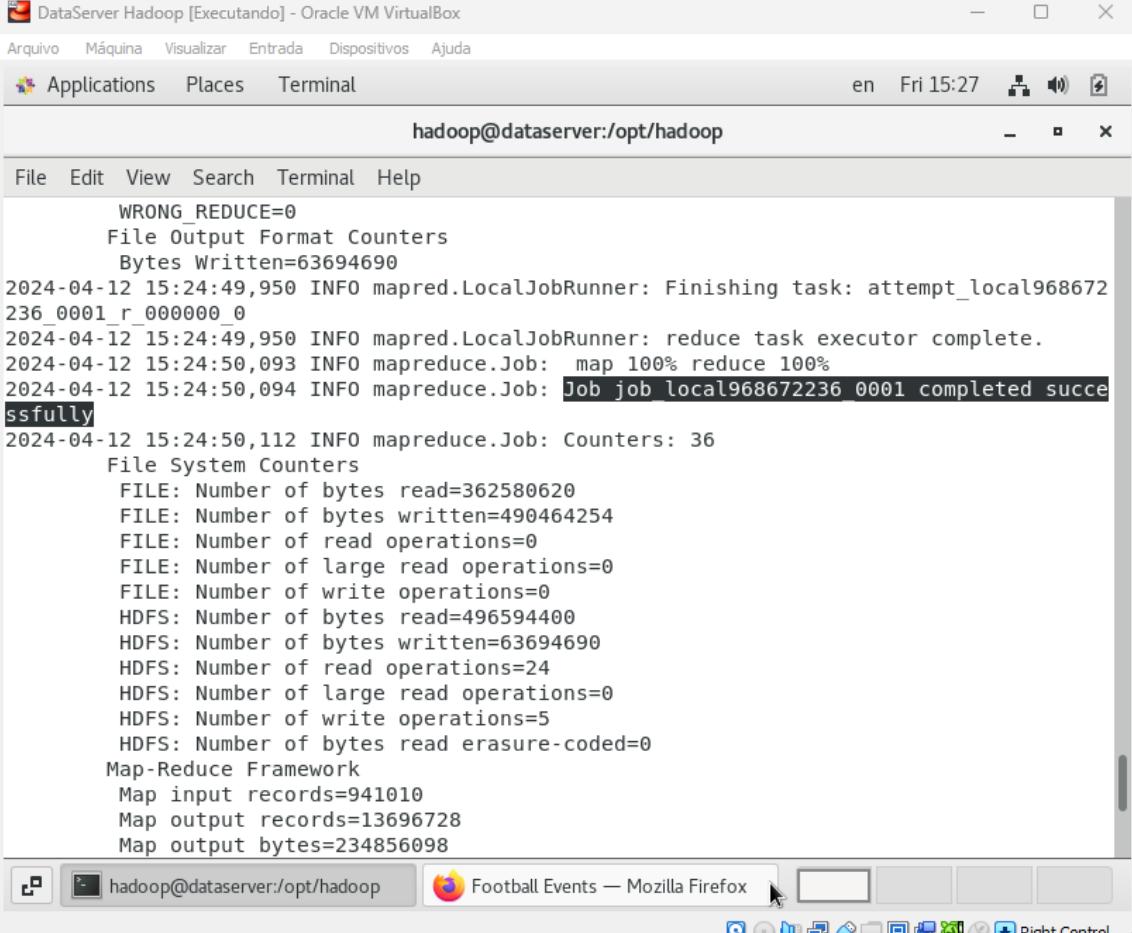


The screenshot shows a terminal window titled "hadoop@dataserver:/opt/hadoop". The window contains a log of Hadoop operations. Key log entries include:

- 2024-04-12 15:24:45,611 WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
- 2024-04-12 15:24:45,644 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=1376531200, maxSingleShuffleLimit=344132800, mergeThreshold=908510656, ioSortFactor=10, memToMemMergeOutputsThreshold=10
- 2024-04-12 15:24:45,658 INFO reduce.EventFetcher: attempt\_local968672236\_0001\_r\_000000\_0 Thread started: EventFetcher for fetching Map Completion Events
- 2024-04-12 15:24:45,722 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt\_local968672236\_0001\_m\_000001\_0 decompress: 20931474 len: 20931478 to MEMORY
- 2024-04-12 15:24:45,864 INFO reduce.InMemoryMapOutput: Read 20931474 bytes from map-output for attempt\_local968672236\_0001\_m\_000001\_0
- 2024-04-12 15:24:45,865 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 20931474, inMemoryMapOutputs.size() -> 1, commitMemory -> 0, usedMemory -> 20931474
- 2024-04-12 15:24:45,913 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt\_local968672236\_0001\_m\_000000\_0 decompress: 52518166 len: 52518170 to MEMORY
- 2024-04-12 15:24:46,081 INFO mapreduce.Job: map 100% reduce 0%
- 2024-04-12 15:24:46,232 INFO reduce.InMemoryMapOutput: Read 52518166 bytes from map-output for attempt\_local968672236\_0001\_m\_000000\_0
- 2024-04-12 15:24:46,232 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 52518166, inMemoryMapOutputs.size() -> 2, commitMemory -> 20931474, usedMemory -> 73449640
- 2024-04-12 15:24:46,233 INFO reduce.EventFetcher: EventFetcher is interrupted.. Returning

Job sendo processado

## Instalação e Configuração do Ambiente Big Data com Hadoop e Spark



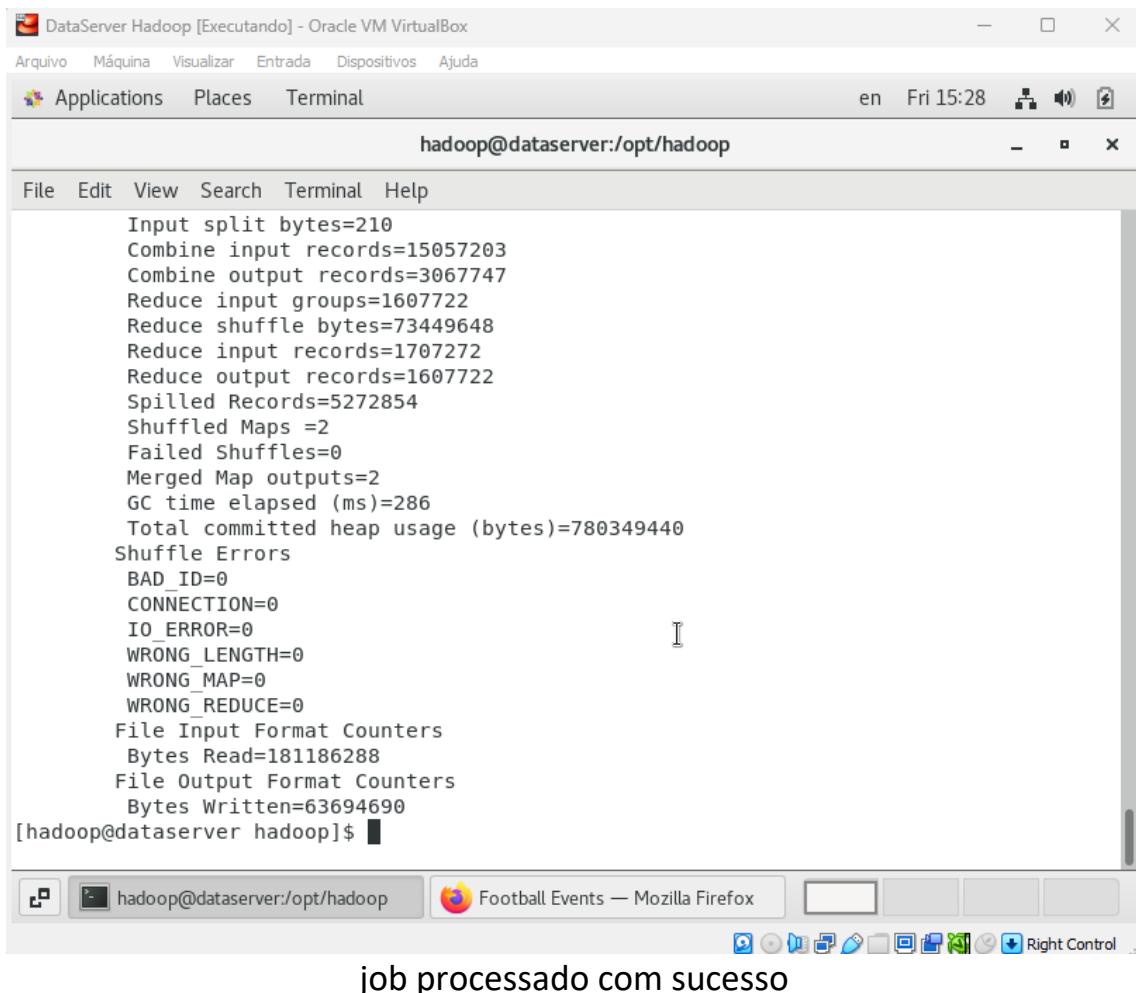
The screenshot shows a terminal window titled "hadoop@dataserver:/opt/hadoop". The window contains the following log output from a Hadoop job:

```
WRONG_REDUCE=0
File Output Format Counters
Bytes Written=63694690
2024-04-12 15:24:49,950 INFO mapred.LocalJobRunner: Finishing task: attempt_local968672
236_0001_r_000000_0
2024-04-12 15:24:49,950 INFO mapred.LocalJobRunner: reduce task executor complete.
2024-04-12 15:24:50,093 INFO mapreduce.Job: map 100% reduce 100%
2024-04-12 15:24:50,094 INFO mapreduce.Job: Job job_local968672236_0001 completed successfully
2024-04-12 15:24:50,112 INFO mapreduce.Job: Counters: 36
  File System Counters
    FILE: Number of bytes read=362580620
    FILE: Number of bytes written=490464254
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=496594400
    HDFS: Number of bytes written=63694690
    HDFS: Number of read operations=24
    HDFS: Number of large read operations=0
    HDFS: Number of write operations=5
    HDFS: Number of bytes read erasure-coded=0
  Map-Reduce Framework
    Map input records=941010
    Map output records=13696728
    Map output bytes=234856098
```

The terminal window has a menu bar with "File", "Edit", "View", "Search", "Terminal", and "Help". The status bar at the bottom right shows "en Fri 15:27". Below the terminal window, the desktop taskbar shows two open applications: "hadoop@dataserver:/opt/hadoop" and "Football Events — Mozilla Firefox".

job processado com sucesso

## Instalação e Configuração do Ambiente Big Data com Hadoop e Spark

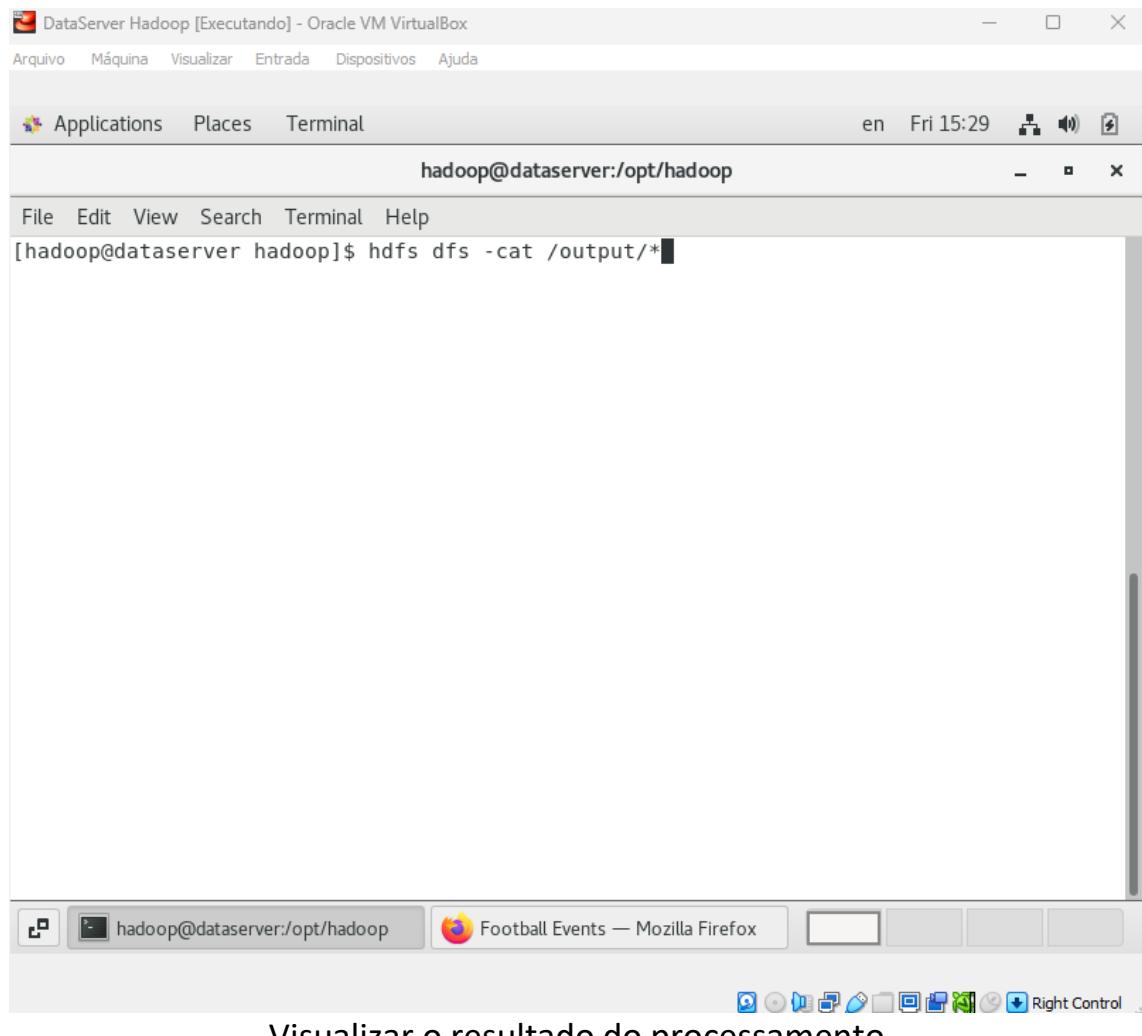


A screenshot of a Linux desktop environment. At the top, there's a window titled "DataServer Hadoop [Executando] - Oracle VM VirtualBox". Below it is a menu bar with "Arquivo", "Máquina", "Visualizar", "Entrada", "Dispositivos", and "Ajuda". The desktop has icons for "Applications", "Places", and "Terminal". The system tray shows "en", "Fri 15:28", and some icons. The terminal window is open at the command prompt "hadoop@dataserver:/opt/hadoop". It displays the output of a Hadoop job, including counters for input splits, combine input records, combine output records, reduce input groups, reduce shuffle bytes, reduce input records, reduce output records, spilled records, shuffled maps, failed shuffles, merged map outputs, GC time elapsed, total committed heap usage, and various shuffle errors. The job summary shows 181186288 bytes read and 63694690 bytes written. The terminal ends with "[hadoop@dataserver hadoop]\$".

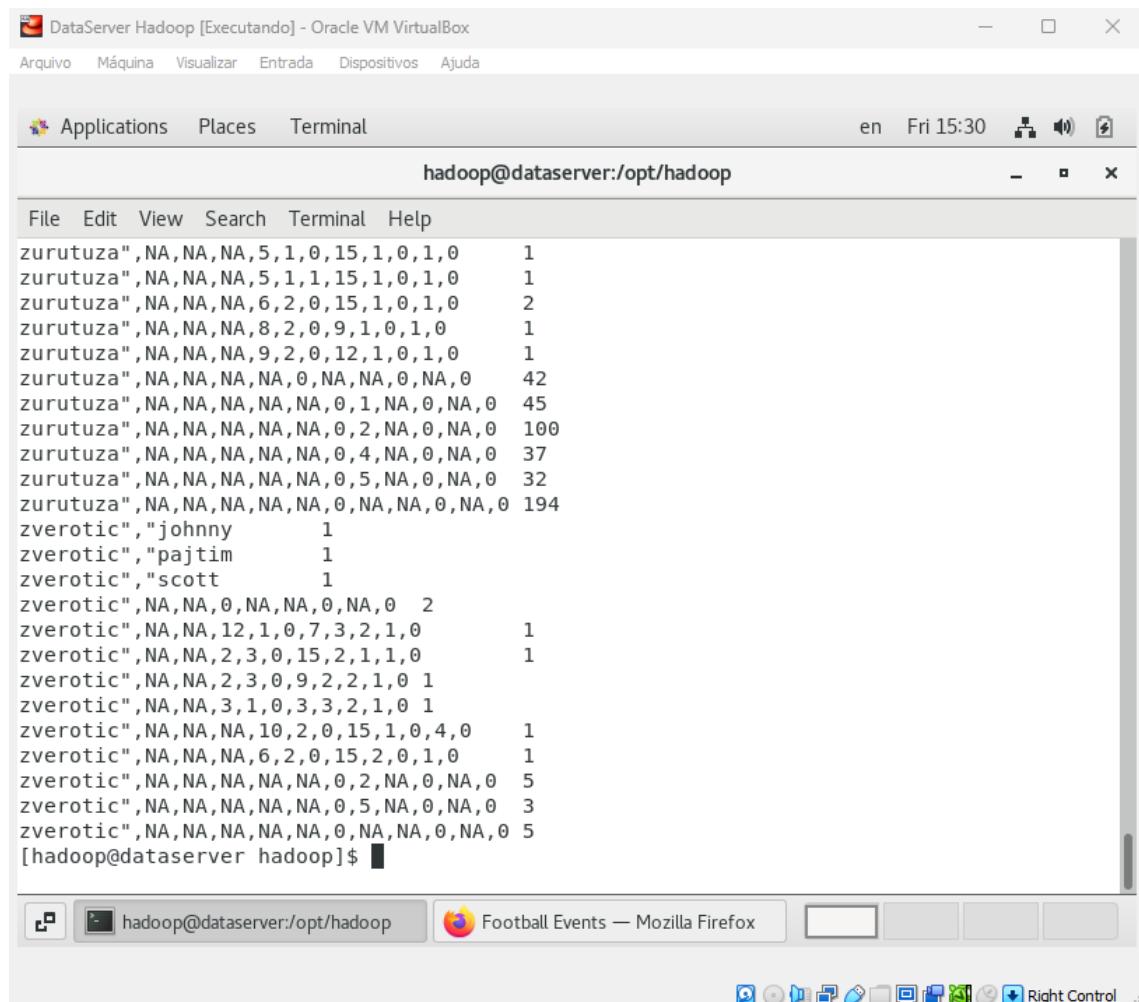
```
Input split bytes=210
Combine input records=15057203
Combine output records=3067747
Reduce input groups=1607722
Reduce shuffle bytes=73449648
Reduce input records=1707272
Reduce output records=1607722
Spilled Records=5272854
Shuffled Maps =2
Failed Shuffles=0
Merged Map outputs=2
GC time elapsed (ms)=286
Total committed heap usage (bytes)=780349440
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=181186288
File Output Format Counters
Bytes Written=63694690
[hadoop@dataserver hadoop]$
```

job processado com sucesso

## Instalação e Configuração do Ambiente Big Data com Hadoop e Spark



## Instalação e Configuração do Ambiente Big Data com Hadoop e Spark



The screenshot shows a Linux desktop environment with a terminal window and a Firefox browser window.

**Terminal Window:**

```
zurutuza",NA,NA,NA,5,1,0,15,1,0,1,0      1
zurutuza",NA,NA,NA,5,1,1,15,1,0,1,0      1
zurutuza",NA,NA,NA,6,2,0,15,1,0,1,0      2
zurutuza",NA,NA,NA,8,2,0,9,1,0,1,0      1
zurutuza",NA,NA,NA,9,2,0,12,1,0,1,0      1
zurutuza",NA,NA,NA,NA,0,NA,NA,0,NA,0      42
zurutuza",NA,NA,NA,NA,0,1,NA,0,NA,0      45
zurutuza",NA,NA,NA,NA,0,2,NA,0,NA,0      100
zurutuza",NA,NA,NA,NA,0,4,NA,0,NA,0      37
zurutuza",NA,NA,NA,NA,0,5,NA,0,NA,0      32
zurutuza",NA,NA,NA,NA,NA,0,NA,NA,0,NA,0  194
zverotic","johnny           1
zverotic","pajtim           1
zverotic","scott            1
zverotic",NA,NA,0,NA,NA,0,NA,0           2
zverotic",NA,NA,12,1,0,7,3,2,1,0        1
zverotic",NA,NA,2,3,0,15,2,1,1,0        1
zverotic",NA,NA,2,3,0,9,2,2,1,0          1
zverotic",NA,NA,3,1,0,3,3,2,1,0          1
zverotic",NA,NA,NA,10,2,0,15,1,0,4,0    1
zverotic",NA,NA,NA,6,2,0,15,2,0,1,0    1
zverotic",NA,NA,NA,NA,0,2,NA,0,NA,0      5
zverotic",NA,NA,NA,NA,0,5,NA,0,NA,0      3
zverotic",NA,NA,NA,NA,NA,0,NA,NA,0,NA,0  5
[hadoop@dataserver hadoop]$ █
```

**Firefox Browser Window:**

Football Events — Mozilla Firefox

Right Control ..

Arquivo processado. Número de ocorrência de cada palavra/termo no arquivo.

# Instalação e Configuração do Ambiente Big Data com Hadoop e Spark

The screenshot shows a Firefox browser window titled "Namenode information - Mozilla Firefox". The address bar indicates the URL is "localhost:9870/dfshealth.html#tab-overview". The main content area displays the "Overview" tab for a cluster, with the identifier "localhost:9000" and the status "(active)". Below this, there is a table with the following data:

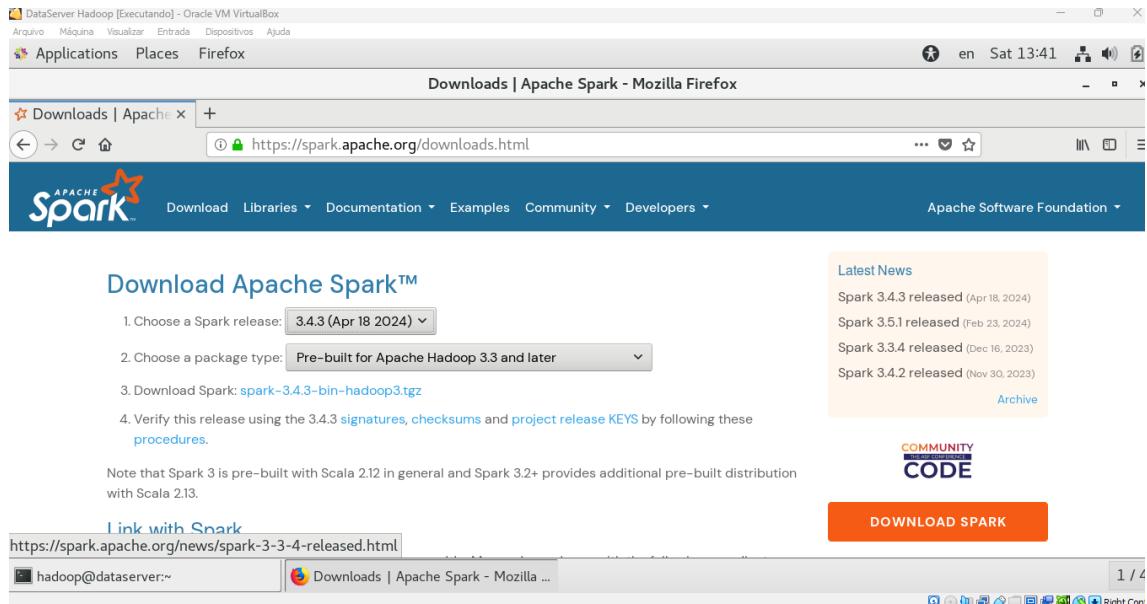
<b>Started:</b>	Fri Apr 19 13:35:16 -0700 2024
<b>Version:</b>	3.2.0, re97acb3bd8f3befd27418996fa5d4b50bf2e17bf
<b>Compiled:</b>	Mon Jan 07 22:08:00 -0800 2019 by sunilg from branch-3.2.0
<b>Cluster ID:</b>	CID-5fc51e33-7f93-40c0-b911-f1ac1f138657
<b>Block Pool ID:</b>	BP-2232935-127.0.0.1-1713558895722

At the bottom of the browser window, the status bar shows "hadoop@dataserver:~" and "Namenode information - Mozilla Fire...". The bottom right corner of the browser window shows "1 / 4".

Acesso ao Hadoop pelo browser <http://localhost:9870>

## 6 Instalação e Configuração do Spark

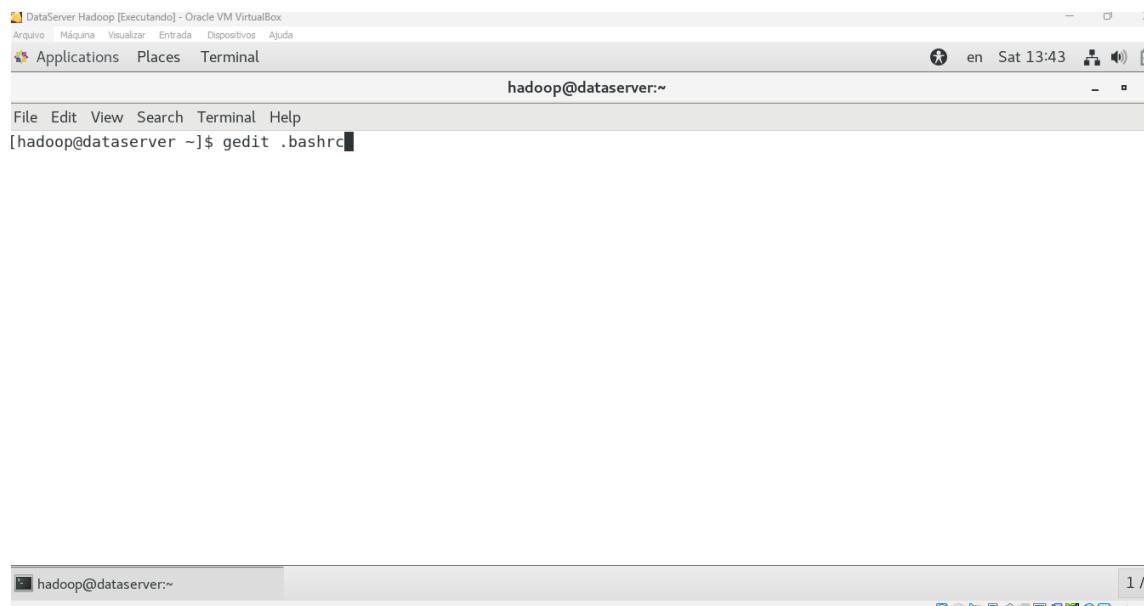
### 6.1 Download e Instalação do Spark



#### Download do Spark – Versão 3.4.3

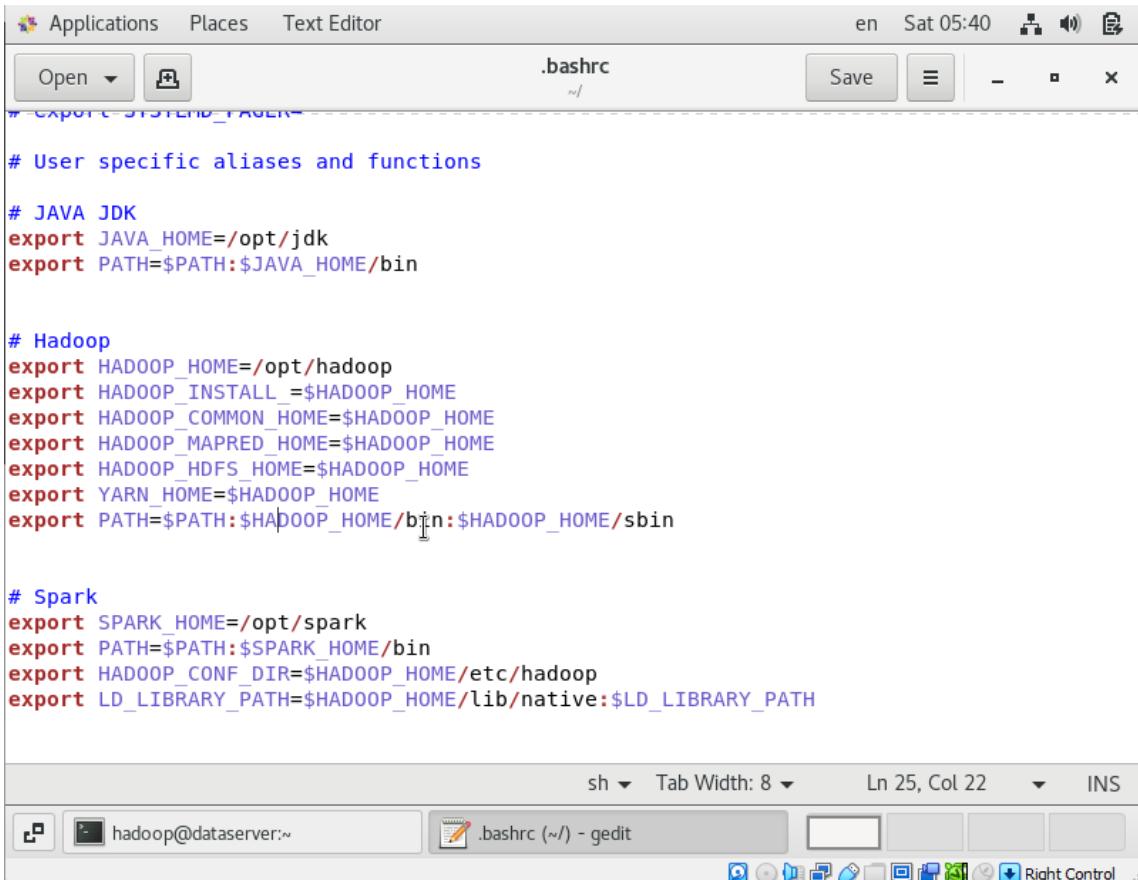
Faça o download, descompacte o arquivo e mova o diretório para /opt/spark da mesma forma como foi feito com o Java JDK e com o Hadoop.

## Instalação e Configuração do Ambiente Big Data com Hadoop e Spark



Editando o arquivo .bashrc

## Instalação e Configuração do Ambiente Big Data com Hadoop e Spark



The screenshot shows a Linux desktop interface with a terminal window and a text editor window.

The terminal window at the bottom has the following status bar text:

- sh ▾ Tab Width: 8 ▾ Ln 25, Col 22 ▾ INS
- hadoop@dataserver:~

The text editor window above it is titled ".bashrc" and contains the following configuration script:

```
# User specific aliases and functions

# JAVA JDK
export JAVA_HOME=/opt/jdk
export PATH=$PATH:$JAVA_HOME/bin

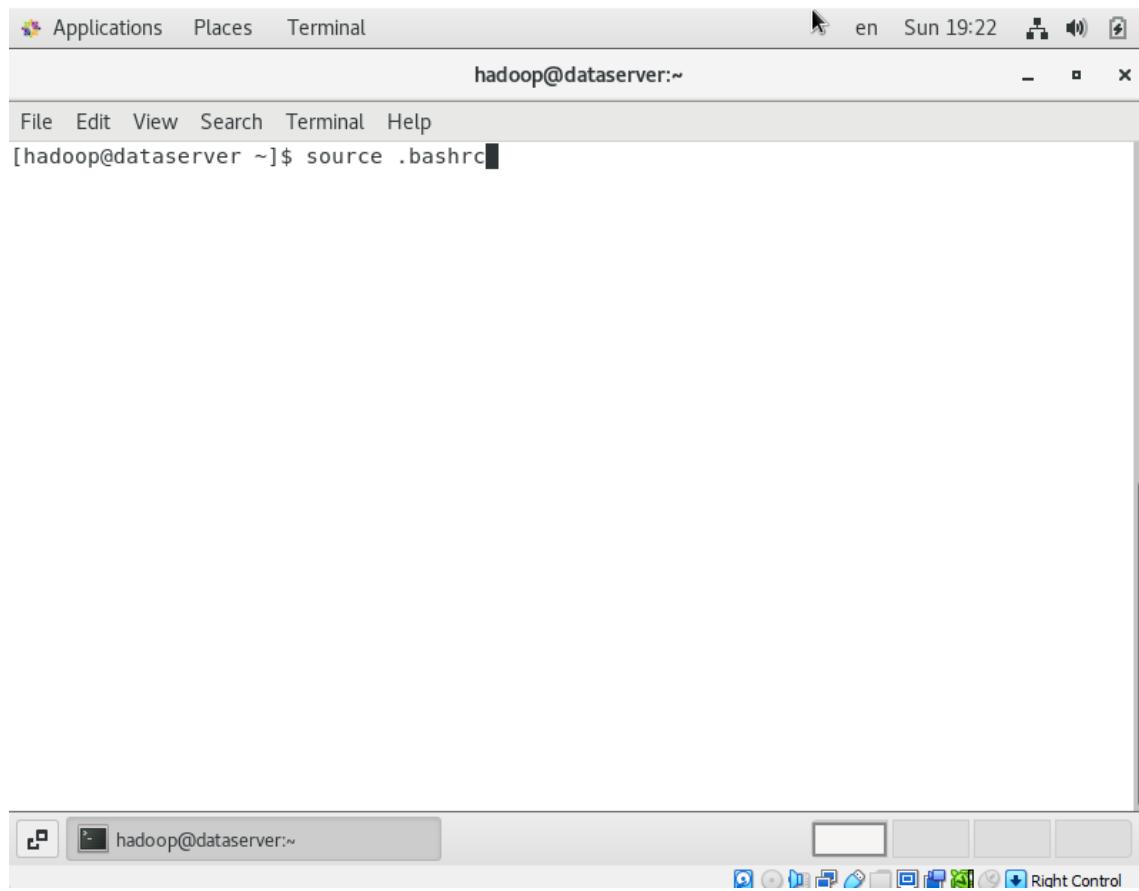
# Hadoop
export HADOOP_HOME=/opt/hadoop
export HADOOP_INSTALL_=$HADOOP_HOME
export HADOOP_COMMON_HOME=$HADOOP_HOME
export HADOOP_MAPRED_HOME=$HADOOP_HOME
export HADOOP_HDFS_HOME=$HADOOP_HOME
export YARN_HOME=$HADOOP_HOME
export PATH=$PATH:$HADOOP_HOME/bin:$HADOOP_HOME/sbin

# Spark
export SPARK_HOME=/opt/spark
export PATH=$PATH:$SPARK_HOME/bin
export HADOOP_CONF_DIR=$HADOOP_HOME/etc/hadoop
export LD_LIBRARY_PATH=$HADOOP_HOME/lib/native:$LD_LIBRARY_PATH
```

Incluir variáveis Spark

## Instalação e Configuração do Ambiente Big Data com Hadoop e Spark

---



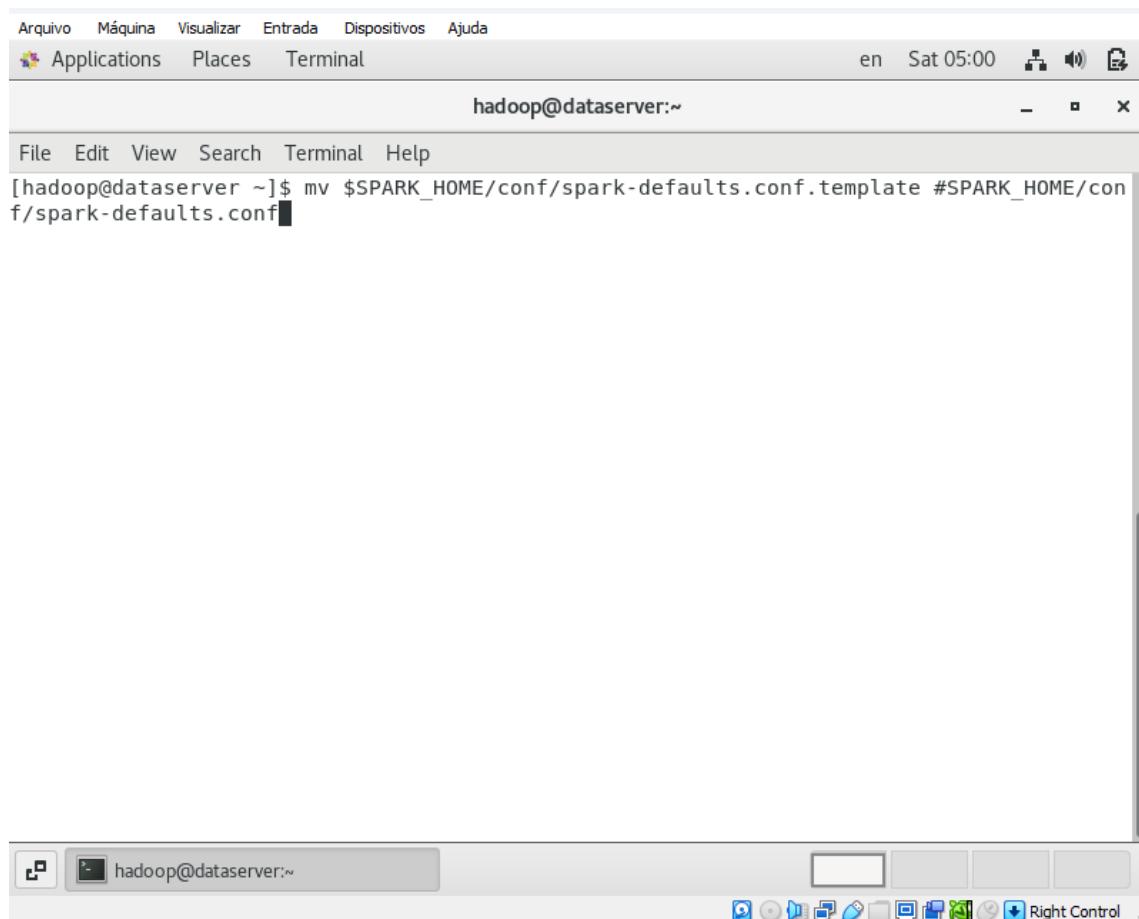
A screenshot of a Linux desktop environment. At the top, there is a panel with icons for Applications, Places, Terminal, and system status (en, Sun 19:22). Below the panel is a terminal window titled "hadoop@dataserver:~". The terminal shows the command "[hadoop@dataserver ~]\$ source .bashrc" being typed. The desktop background is visible behind the terminal window.

source .bashrc

Instalação e Configuração do Ambiente Big Data com Hadoop e Spark

Spark shell - Digite :q para sair do shell

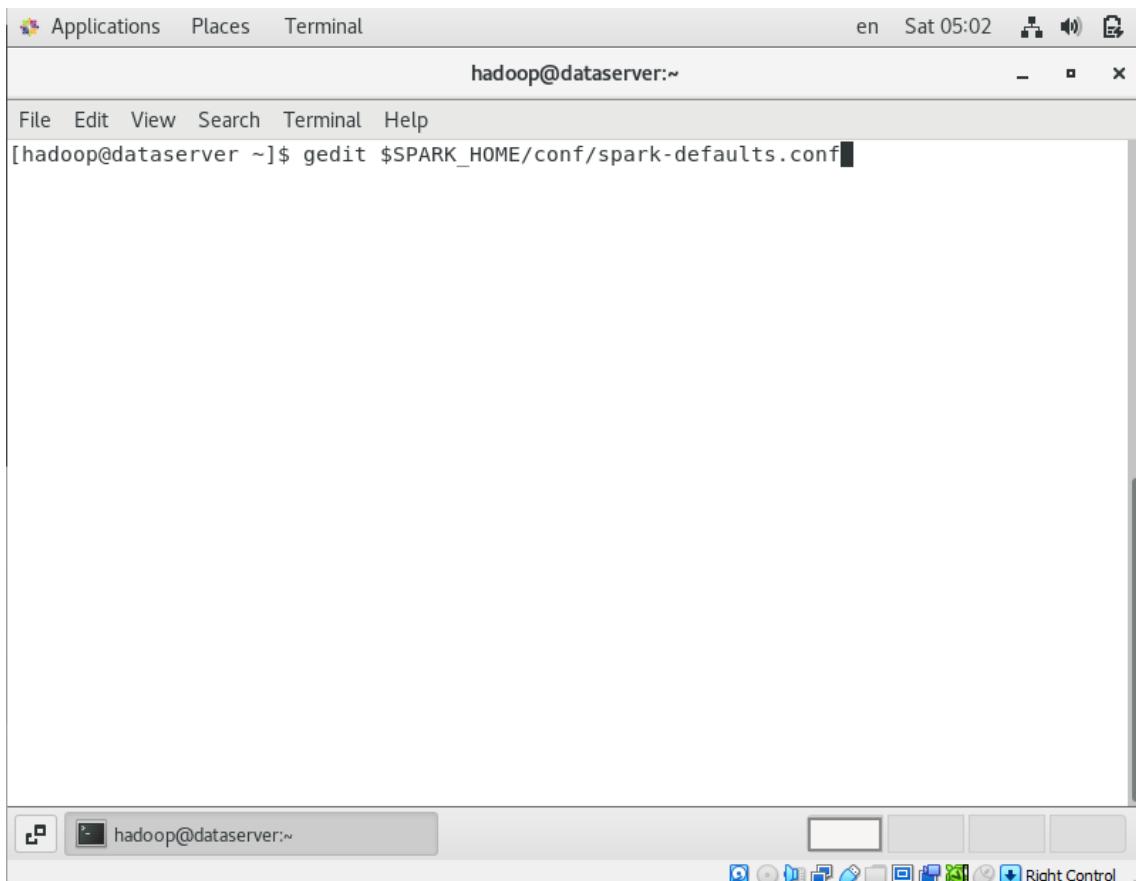
## Instalação e Configuração do Ambiente Big Data com Hadoop e Spark



```
Arquivo Máquina Visualizar Entrada Dispositivos Ajuda
Applications Places Terminal en Sat 05:00
hadoop@dataserver:~ - x
File Edit View Search Terminal Help
[hadoop@dataserver ~]$ mv $SPARK_HOME/conf/spark-defaults.conf.template #SPARK_HOME/conf/spark-defaults.conf
```

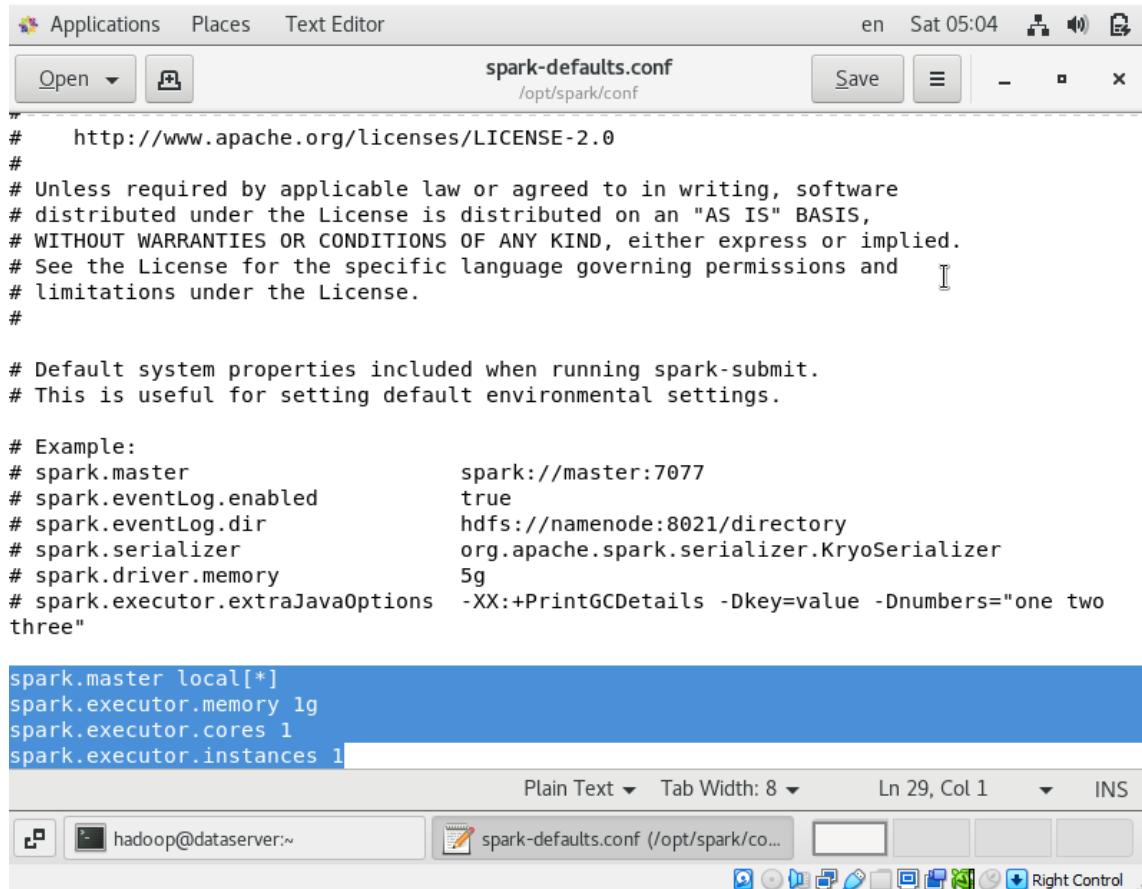
Renomeie o arquivo spark-defaults.conf.templates

## Instalação e Configuração do Ambiente Big Data com Hadoop e Spark



Edite o arquivo \$SPARK\_HOME/conf/spark-defaults.conf

## Instalação e Configuração do Ambiente Big Data com Hadoop e Spark



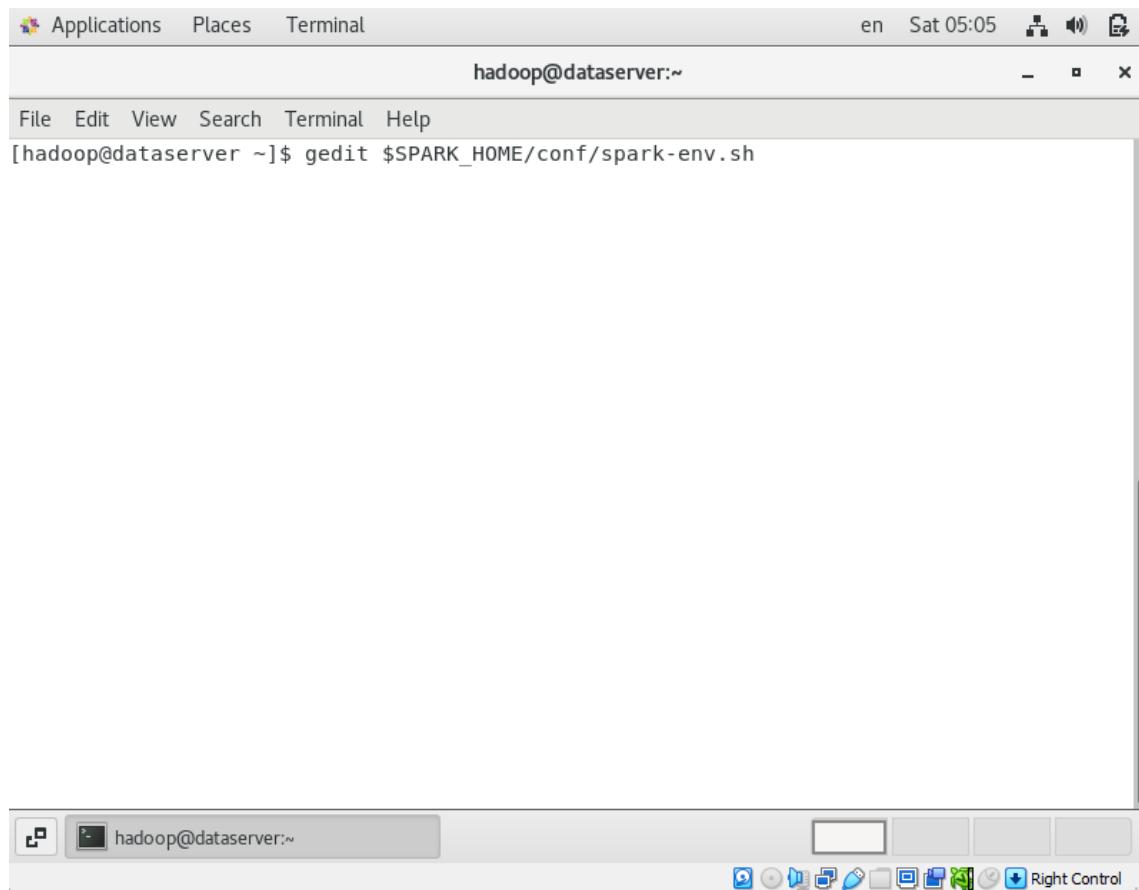
The screenshot shows a desktop environment with a terminal window open. The terminal title is 'spark-defaults.conf /opt/spark/conf'. The window contains the configuration file content, with several lines highlighted in blue. The highlighted lines are:

```
# spark.master local[*]
spark.executor.memory 1g
spark.executor.cores 1
spark.executor.instances 1
```

Below the terminal window, the desktop taskbar shows icons for various applications like Applications, Places, Text Editor, and a file manager.

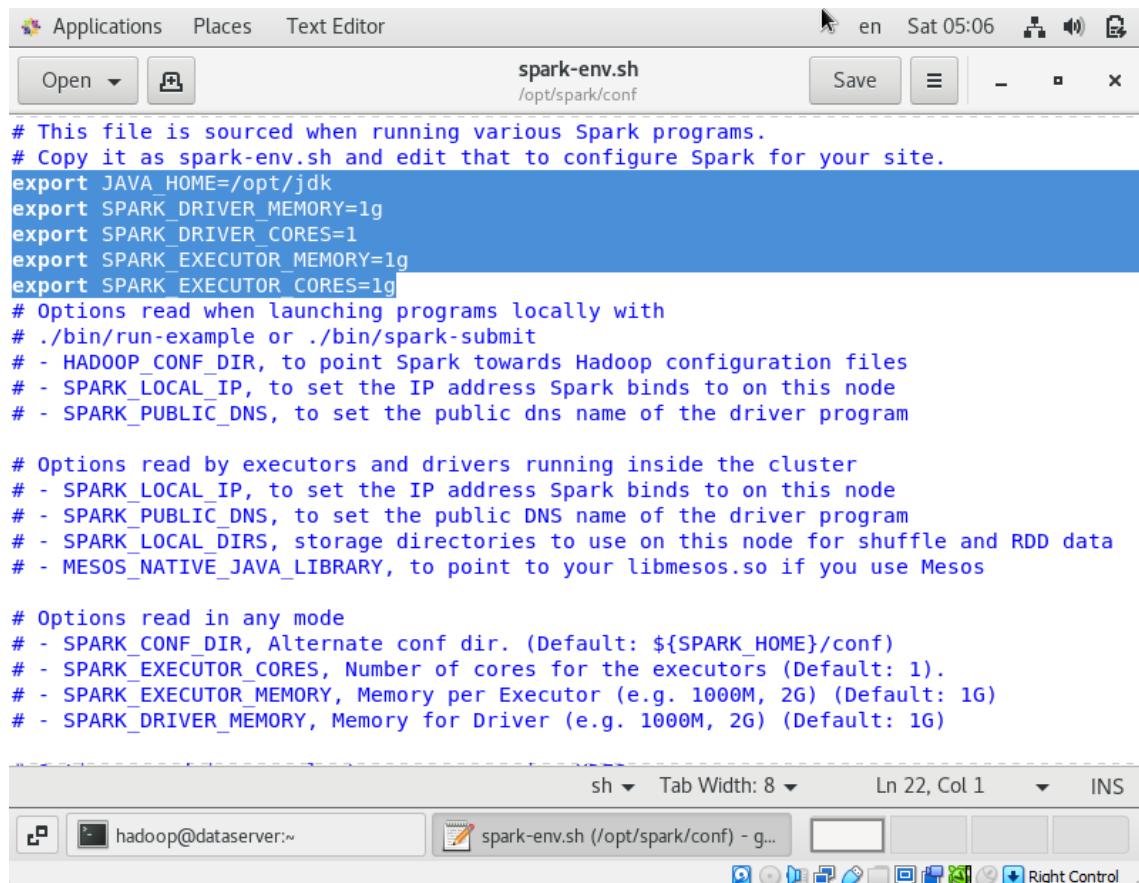
Acrescente as linhas selecionadas

## Instalação e Configuração do Ambiente Big Data com Hadoop e Spark



Edite o arquivo

## Instalação e Configuração do Ambiente Big Data com Hadoop e Spark

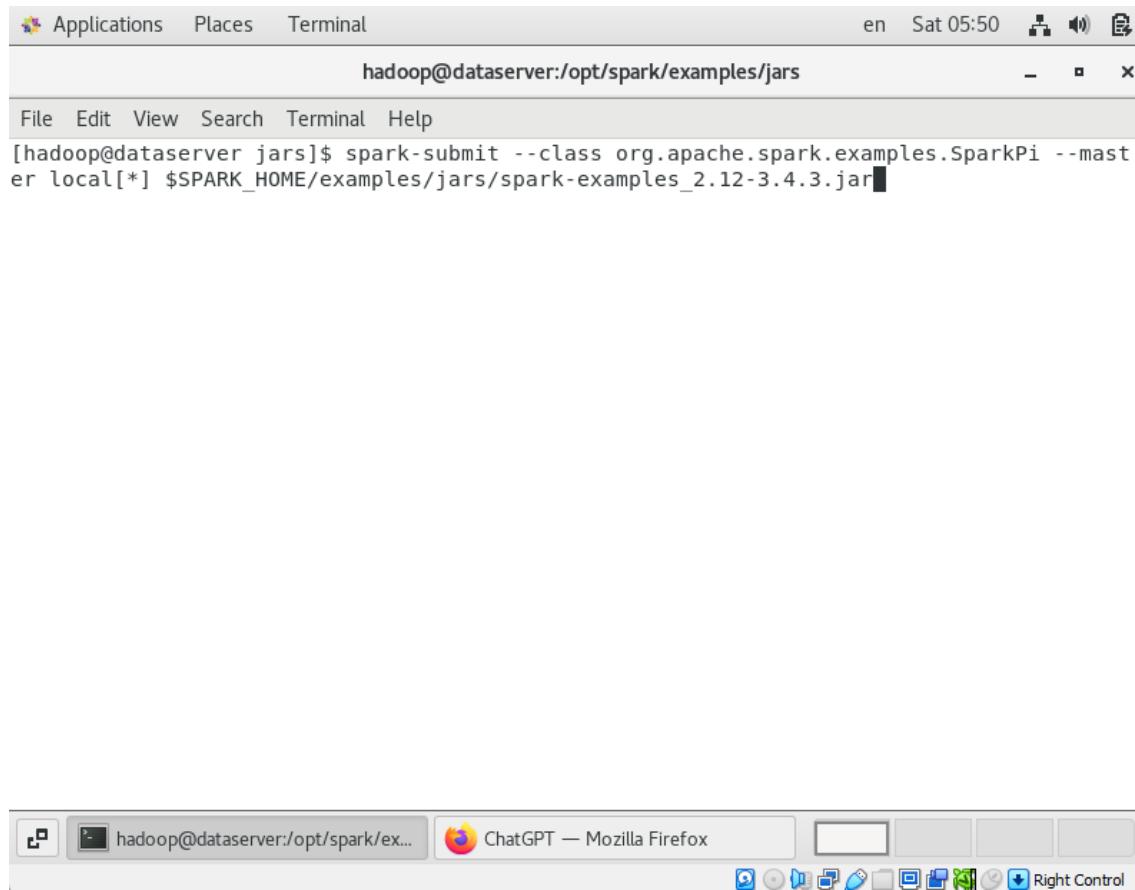


The screenshot shows a desktop environment with a terminal window titled "spark-env.sh" open in a text editor. The file path is "/opt/spark/conf". The terminal window has a blue selection bar highlighting several lines of code. The code is a shell script with comments explaining various environment variables and options for running Spark programs.

```
# This file is sourced when running various Spark programs.  
# Copy it as spark-env.sh and edit that to configure Spark for your site.  
export JAVA_HOME=/opt/jdk  
export SPARK_DRIVER_MEMORY=lg  
export SPARK_DRIVER_CORES=1  
export SPARK_EXECUTOR_MEMORY=lg  
export SPARK_EXECUTOR_CORES=1  
# Options read when launching programs locally with  
# ./bin/run-example or ./bin/spark-submit  
# - HADOOP_CONF_DIR, to point Spark towards Hadoop configuration files  
# - SPARK_LOCAL_IP, to set the IP address Spark binds to on this node  
# - SPARK_PUBLIC_DNS, to set the public dns name of the driver program  
  
# Options read by executors and drivers running inside the cluster  
# - SPARK_LOCAL_IP, to set the IP address Spark binds to on this node  
# - SPARK_PUBLIC_DNS, to set the public DNS name of the driver program  
# - SPARK_LOCAL_DIRS, storage directories to use on this node for shuffle and RDD data  
# - MESOS_NATIVE_JAVA_LIBRARY, to point to your libmesos.so if you use Mesos  
  
# Options read in any mode  
# - SPARK_CONF_DIR, Alternate conf dir. (Default: ${SPARK_HOME}/conf)  
# - SPARK_EXECUTOR_CORES, Number of cores for the executors (Default: 1).  
# - SPARK_EXECUTOR_MEMORY, Memory per Executor (e.g. 1000M, 2G) (Default: 1G)  
# - SPARK_DRIVER_MEMORY, Memory for Driver (e.g. 1000M, 2G) (Default: 1G)
```

Acrescente as linhas selecionadas.

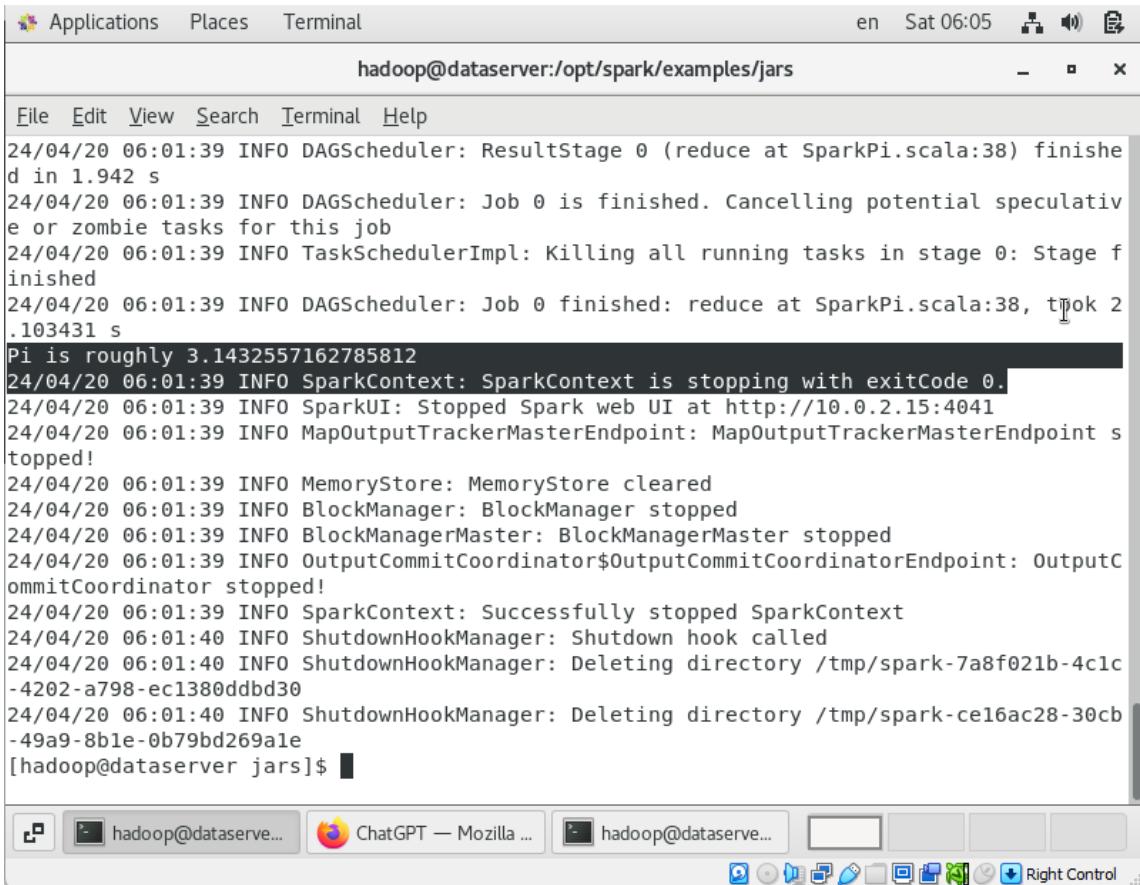
## 7 Processando Big Data com Spark



A screenshot of a Linux desktop environment. At the top, there is a panel with icons for Applications, Places, Terminal, language (en), date (Sat 05:50), and system status. Below the panel is a terminal window titled "hadoop@dataserver:/opt/spark/examples/jars". The terminal shows the command: [hadoop@dataserver jars]\$ spark-submit --class org.apache.spark.examples.SparkPi --master local[\*] \$SPARK\_HOME/examples/jars/spark-examples\_2.12-3.4.3.jar. The desktop dock at the bottom contains icons for the terminal, file manager, browser, and other applications.

Submeter um job Spark de processamento de dados. Execute um job com Spark-Submit. Esse job pode ser qualquer aplicação de processamento de dados. Neste exemplo, usamos o Spark examples. Após a execução deve ver o SparkContext encerrado com o código de saída 0, isso indica que a execução foi bem sucedida.

## Instalação e Configuração do Ambiente Big Data com Hadoop e Spark



The screenshot shows a Linux desktop environment with a terminal window open. The terminal title is "hadoop@dataserver:/opt/spark/examples/jars". The window contains the output of a Spark application. The output includes log messages from the DAGScheduler, TaskSchedulerImpl, and MemoryStore, indicating the completion of a job. A key line in the output is "Pi is roughly 3.1432557162785812". The terminal prompt "[hadoop@dataserver jars]\$" is visible at the bottom. The desktop interface includes a top bar with icons for Applications, Places, Terminal, and system status (en Sat 06:05). Below the terminal are several application icons in the dock, including "hadoop@dataserve...", "ChatGPT — Mozilla ...", and "Right Control".

```
hadoop@dataserver:/opt/spark/examples/jars
File Edit View Search Terminal Help
24/04/20 06:01:39 INFO DAGScheduler: ResultStage 0 (reduce at SparkPi.scala:38) finished in 1.942 s
24/04/20 06:01:39 INFO DAGScheduler: Job 0 is finished. Cancelling potential speculative or zombie tasks for this job
24/04/20 06:01:39 INFO TaskSchedulerImpl: Killing all running tasks in stage 0: Stage finished
24/04/20 06:01:39 INFO DAGScheduler: Job 0 finished: reduce at SparkPi.scala:38, took 2.103431 s
Pi is roughly 3.1432557162785812
24/04/20 06:01:39 INFO SparkContext: SparkContext is stopping with exitCode 0.
24/04/20 06:01:39 INFO SparkUI: Stopped Spark web UI at http://10.0.2.15:4041
24/04/20 06:01:39 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
24/04/20 06:01:39 INFO MemoryStore: MemoryStore cleared
24/04/20 06:01:39 INFO BlockManager: BlockManager stopped
24/04/20 06:01:39 INFO BlockManagerMaster: BlockManagerMaster stopped
24/04/20 06:01:39 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
24/04/20 06:01:39 INFO SparkContext: Successfully stopped SparkContext
24/04/20 06:01:40 INFO ShutdownHookManager: Shutdown hook called
24/04/20 06:01:40 INFO ShutdownHookManager: Deleting directory /tmp/spark-7a8f021b-4c1c-4202-a798-ec1380ddbd30
24/04/20 06:01:40 INFO ShutdownHookManager: Deleting directory /tmp/spark-ce16ac28-30cb-49a9-8b1e-0b79bd269a1e
[hadoop@dataserver jars]$
```

Job processado e informando o valor de Pi. Job submetido com sucesso!!

## Instalação e Configuração do Ambiente Big Data com Hadoop e Spark

The screenshot shows a terminal window titled "hadoop@dataserver:~/Downloads". The window contains the following text:

```
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
24/04/20 05:29:06 WARN Utils: Service 'SparkUI' could not bind on port 4040. Attempting port 4041.
Spark context Web UI available at http://10.0.2.15:4041
Spark context available as 'sc' (master = local[*], app id = local-1713601746175).
Spark session available as 'spark'.
Welcome to

    / \   / \
   / \ / \ / \
  / \ / \ / \ / \
 / \ / \ / \ / \ / \
version 3.4.3

Using Scala version 2.12.17 (Java HotSpot(TM) 64-Bit Server VM, Java 1.8.0_202)
Type in expressions to have them evaluated.
Type :help for more information.

scala> var input = spark.read.csv("file:///home/hadoop/Downloads/events.csv")
input: org.apache.spark.sql.DataFrame = [_c0: string, _c1: string ... 20 more fields]

scala> val totalLinhas = input.count()
totalLinhas: Long = 941010

scala> █
```

Execute o script acima para contar o número de linhas no arquivo. O script Scala está fazendo a leitura de um arquivo no HDFS e processando em memória com o Spark. O Spark está funcionado!!

# Instalação e Configuração do Ambiente Big Data com Hadoop e Spark

User: hadoop  
Total Uptime: 1.6 min  
Scheduling Mode: FIFO

Event Timeline

Executors

- Added
- Removed

Jobs

- Succeeded
- Failed
- Running

Sun 21 Mon 22 Tue 23 Wed 24 Thu 25 Fri 26 Sat 27

April 2024

Acessando o Apache Spark pelo browser em <http://localhost:4040>

## 8 Referências

WHITE, Tom. **Hadoop The Definitive Guide**. Newton, Massachusetts, United States: O'Reilly Media, 2015.

DAMJI, Jules S. at al. **Learning Spark Lightning**. Newton, Massachusetts, United States: O'Reilly Media, 2020.

Um ambiente de testes para  
armazenar e processar Big Data!