Engenharia de Dados com Hadoop e Spark

Cleber Zumba
Engenheiro de Dados

Sumário

1 Versão

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

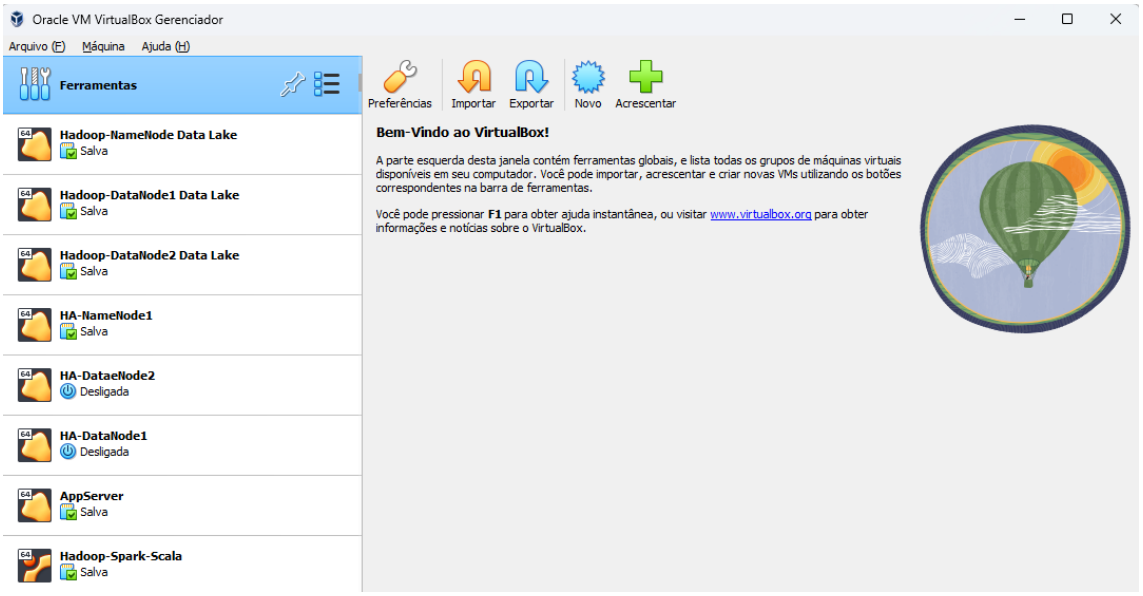
Versão	Ação	Data
1.0	Criação do documento	21/04/2024

2 Configuração do Ambiente

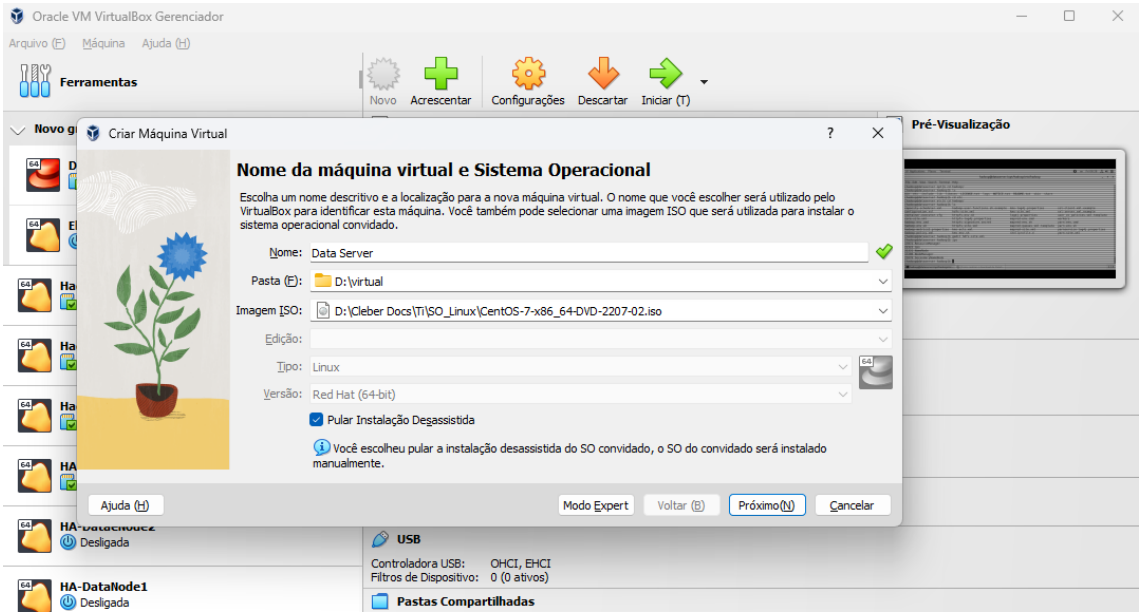
Item	Versão
Virtual Box	7.0
Sistema Operacional	Centos 7 (64 bits)
Interface Gráfica	Gnome
Java	1.8
Apache Hadoop	3.3.6
Apache Spark	3.4.3

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.



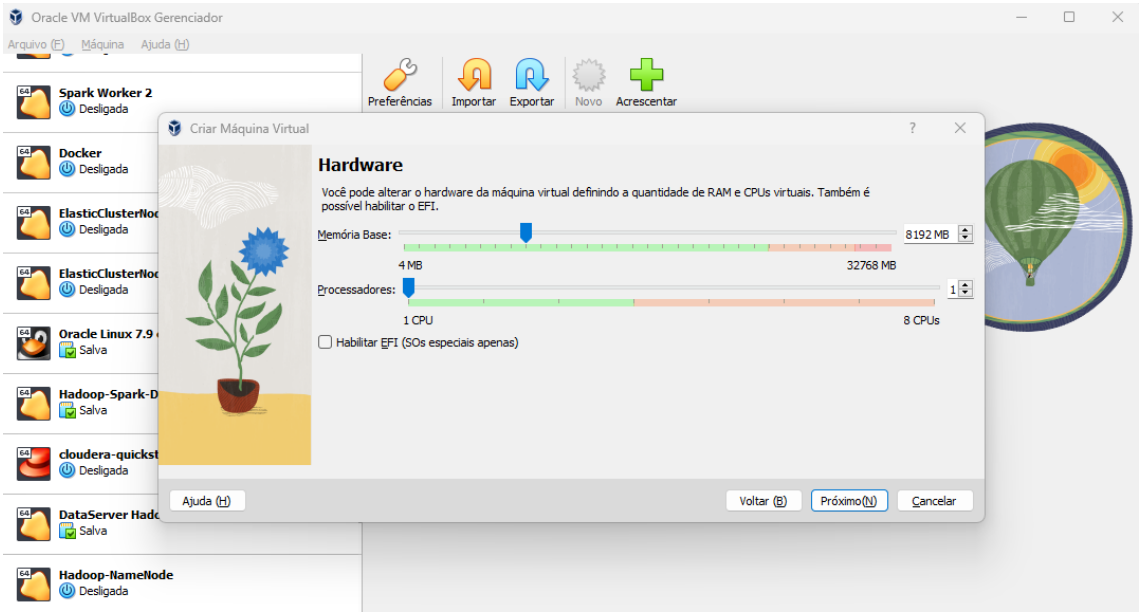
Abrindo o Gerenciador do Oracle Virtual Box



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

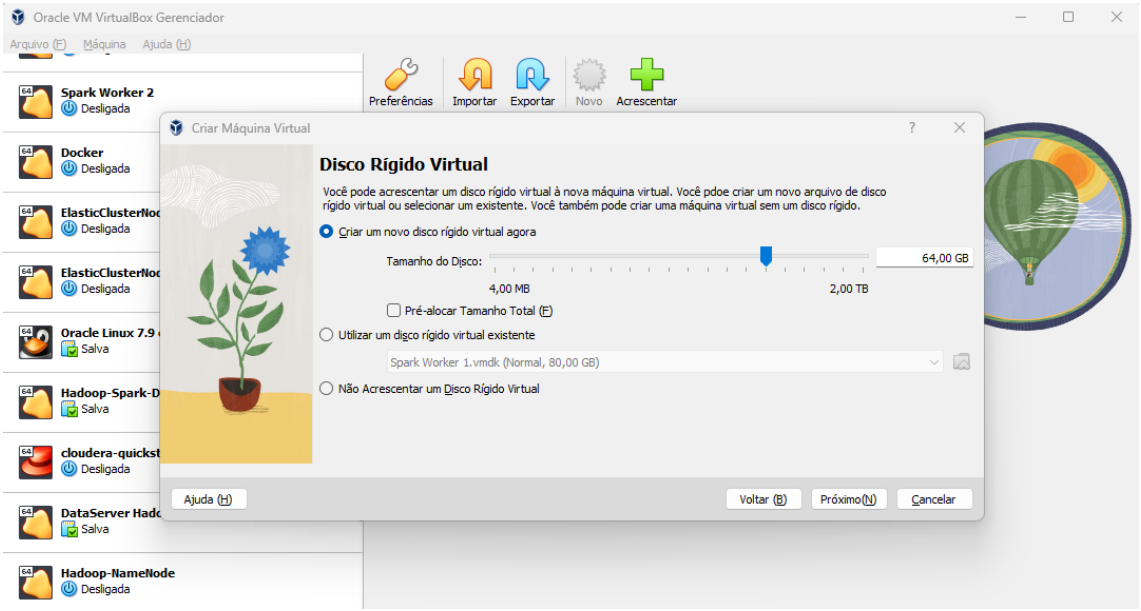
Selecione a mídia de instalação do sistema operacional

CentOS 64 bits (versão 7)

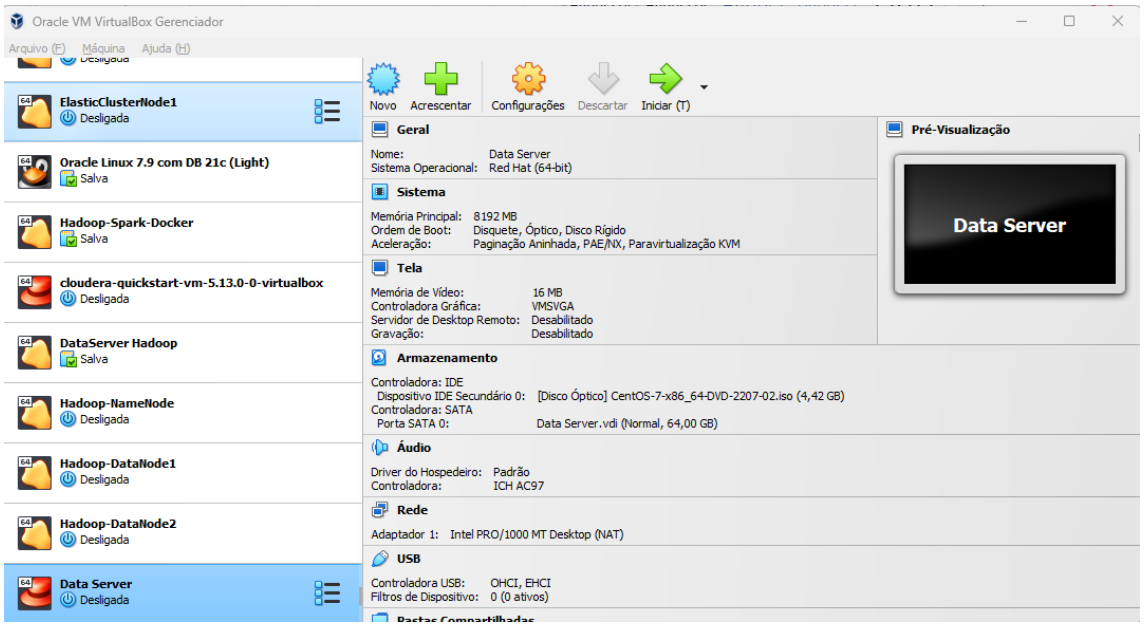


Configure metade da memória física do seu computador para a VM

Cleber Zumba
Engenheiro de Dados

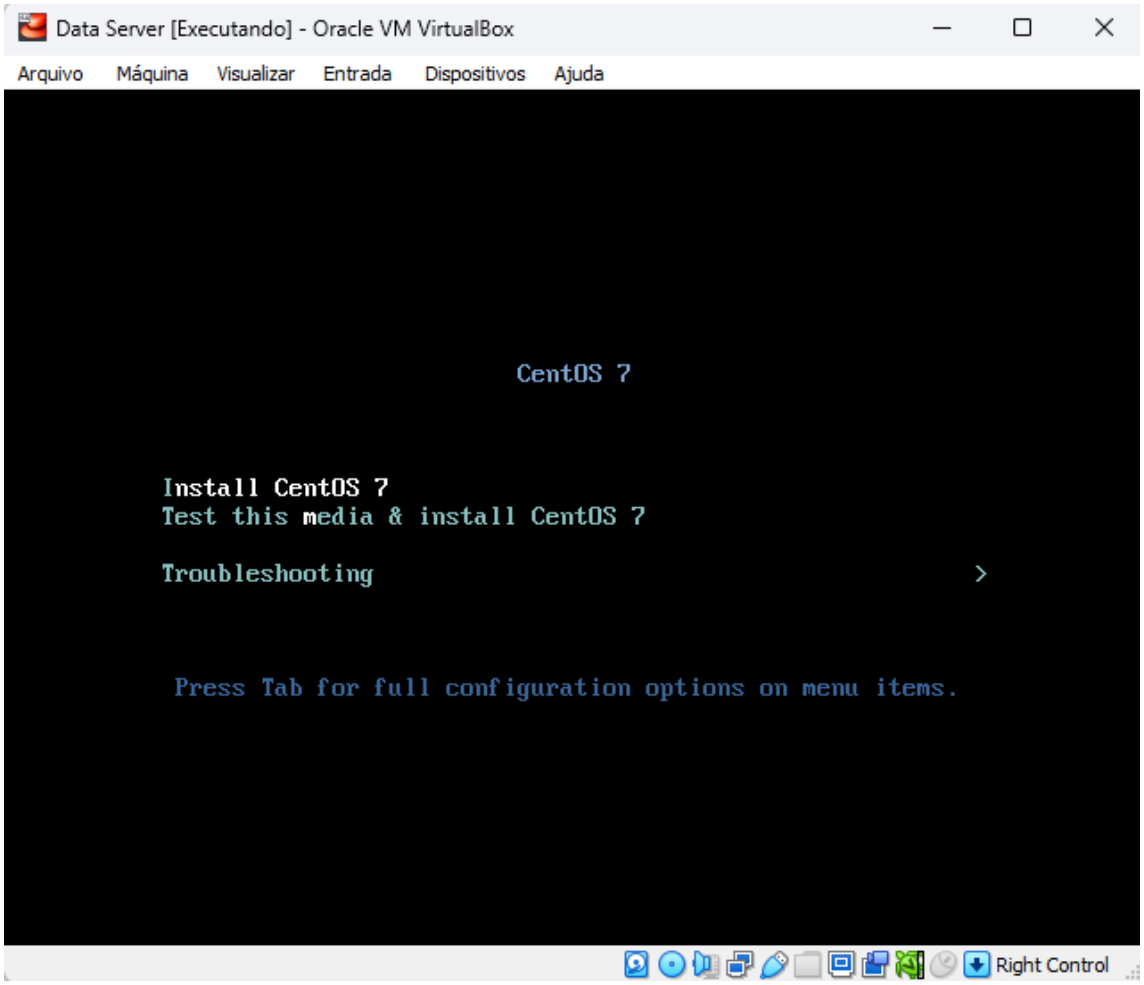


Selecione 64 GB para o disco virtual

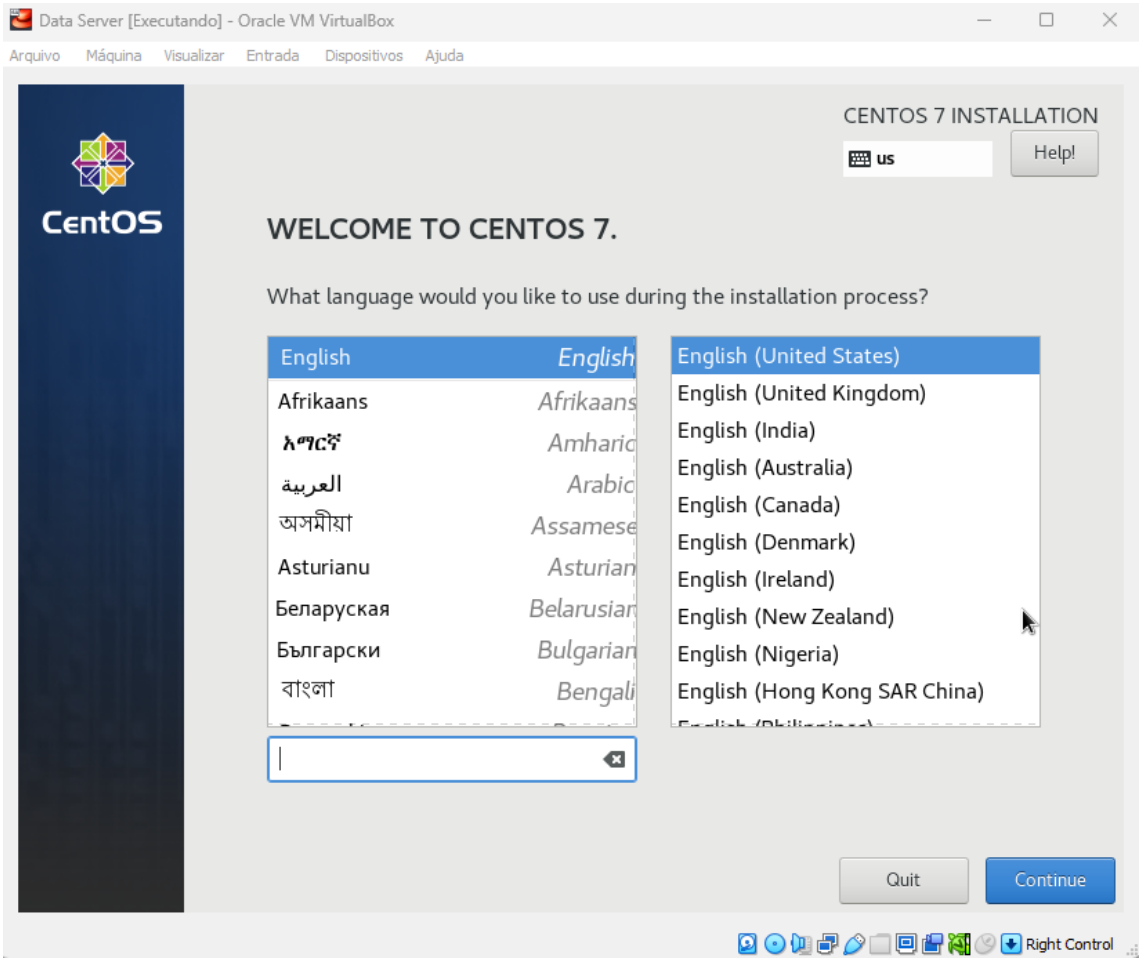


Máquina virtual criada. Selecione a VM e clique no botão Iniciar para inicializar a VM

2.2 Instalação do Sistema Operacional

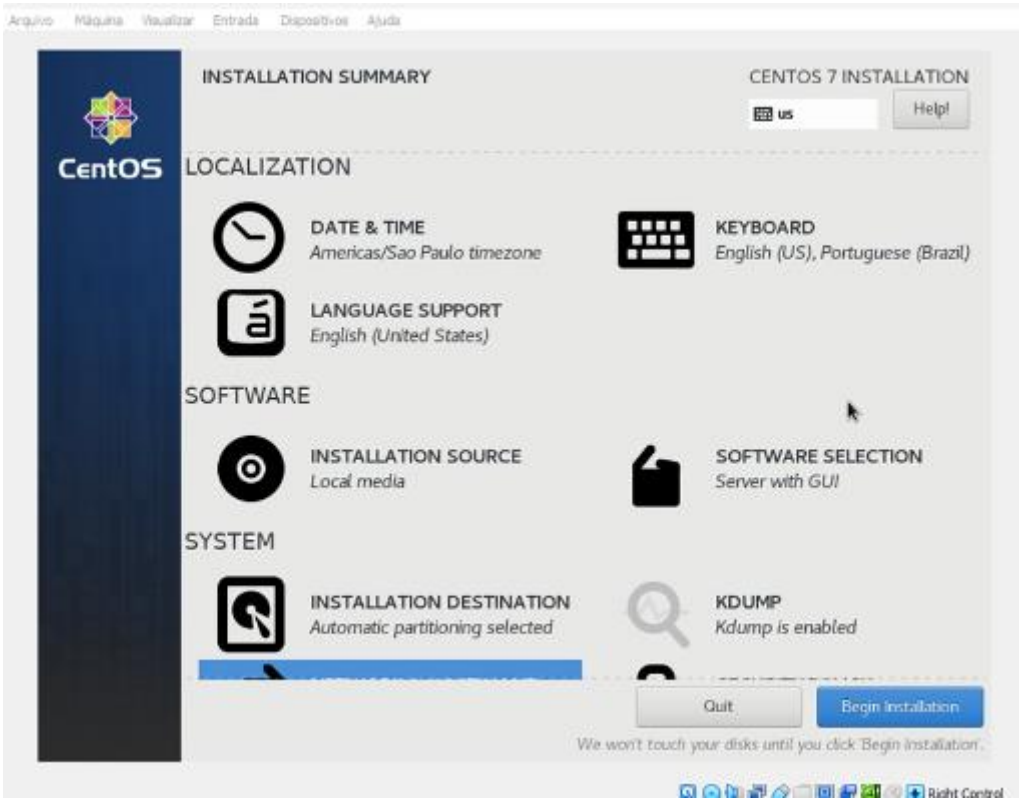


Selecione a opção de Instalação do Sistema Operacional CentOS 7

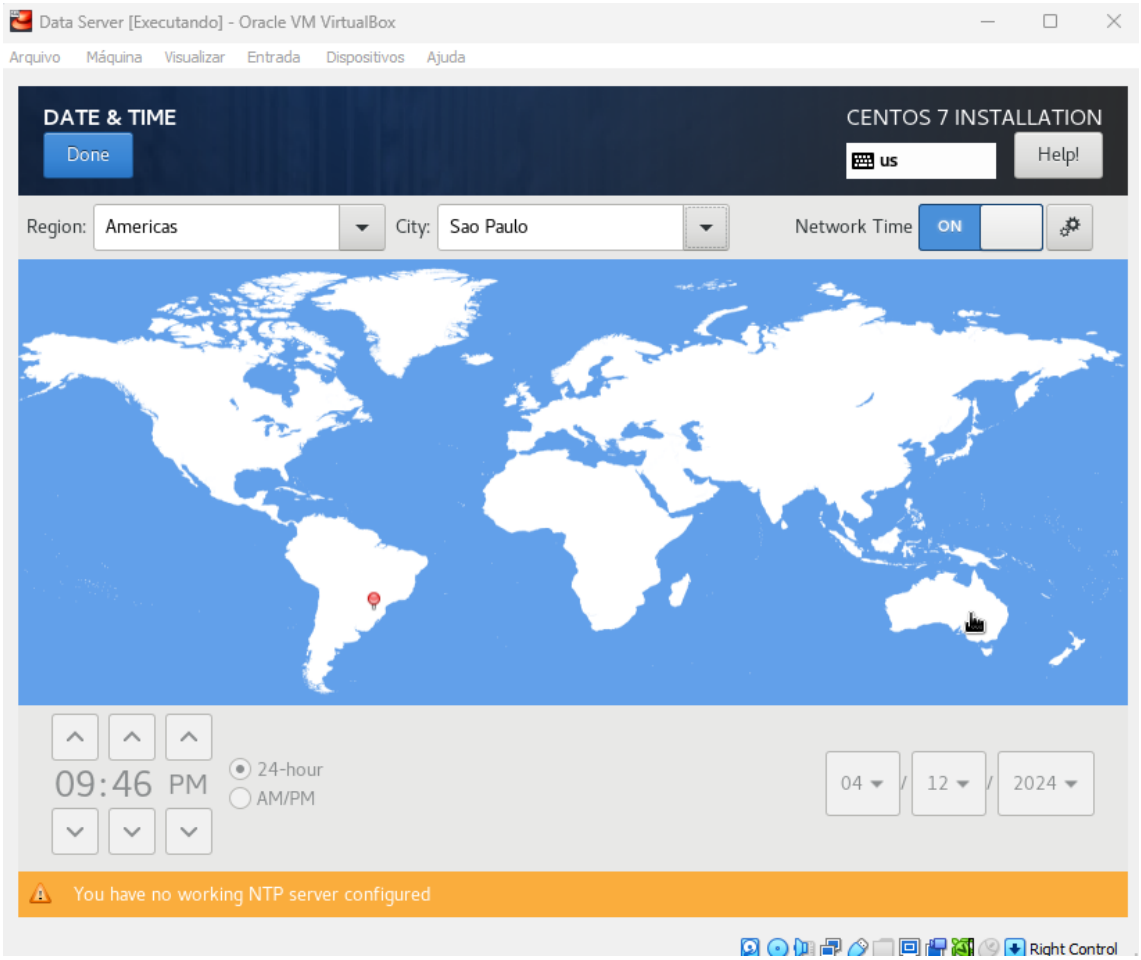


Seleção do idioma usado na instalação

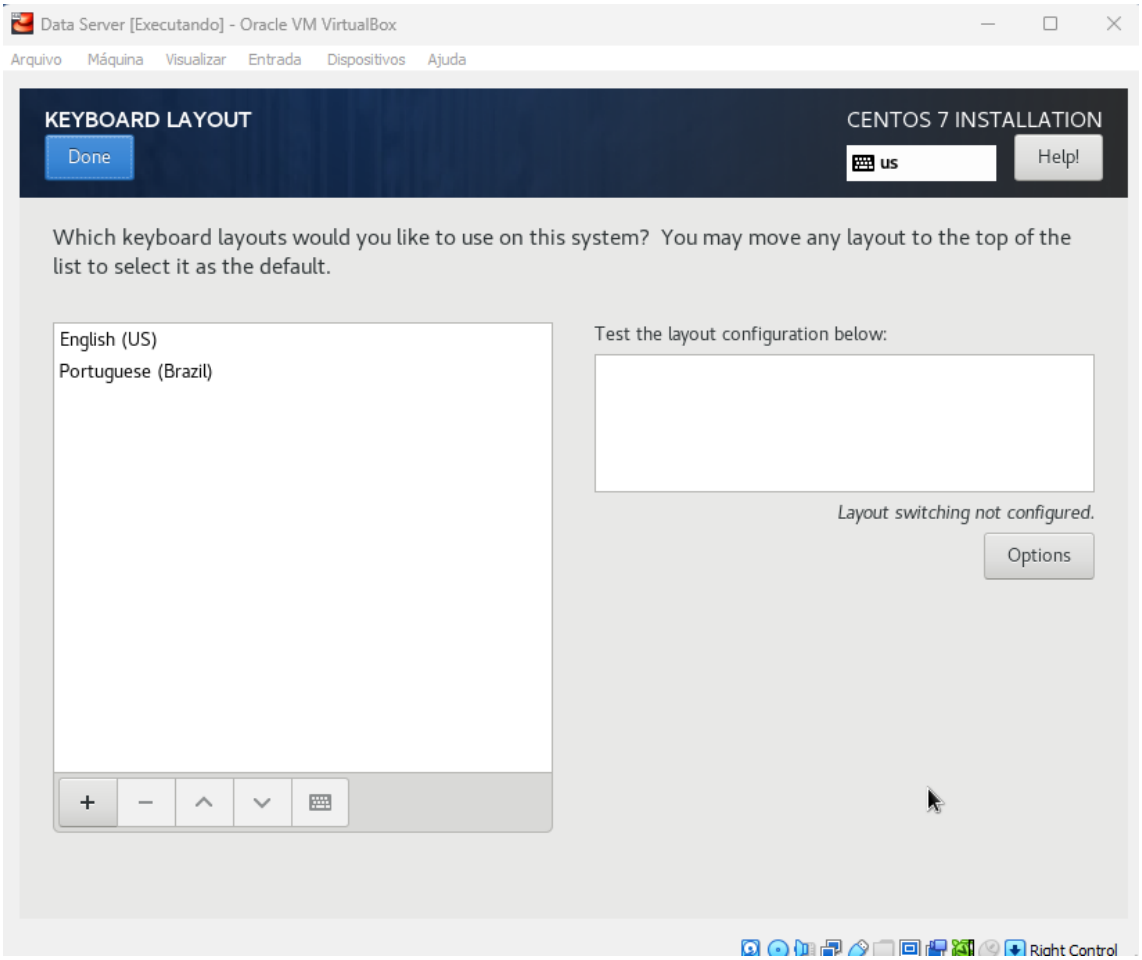
Cleber Zumba
Engenheiro de Dados



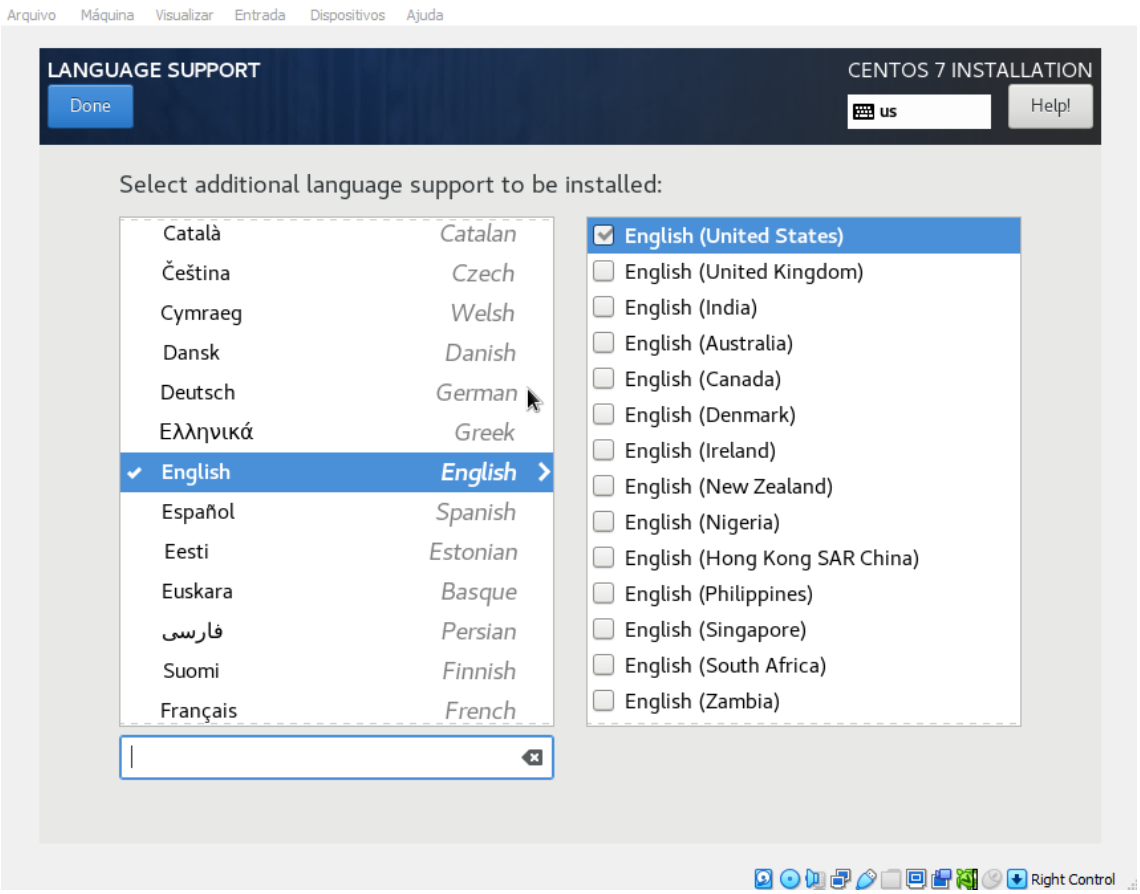
Opções de configuração


Data Server [Executando] - Oracle VM VirtualBox
Arquivo Máquina Visualizar Entrada Dispositivos Ajuda
DATE & TIME
Done
CENTOS 7 INSTALLATION
us
Help!
Region: Americas
City: Sao Paulo
Network Time
ON
09:46 PM
24-hour
AM/PM
04 / 12 / 2024
You have no working NTP server configured
Right Control

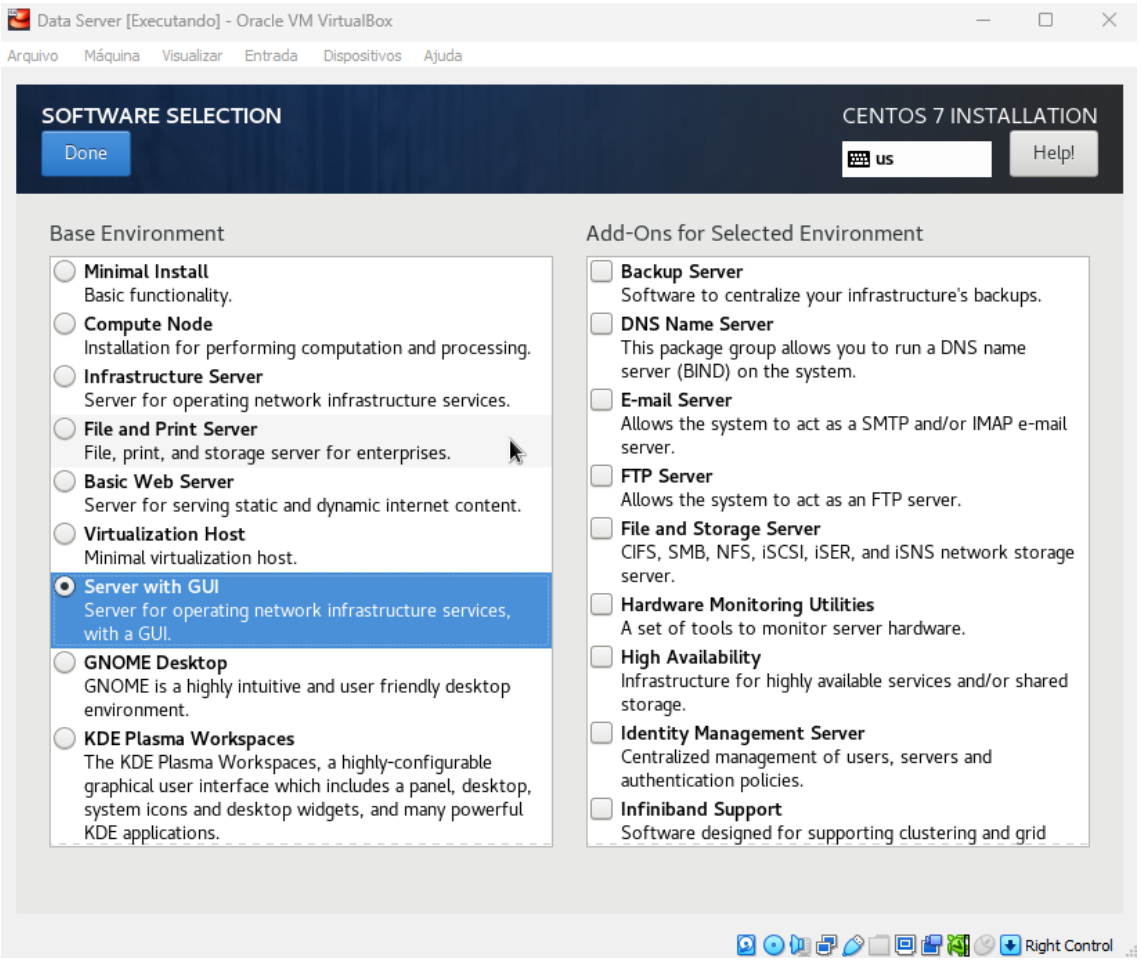
Timezone



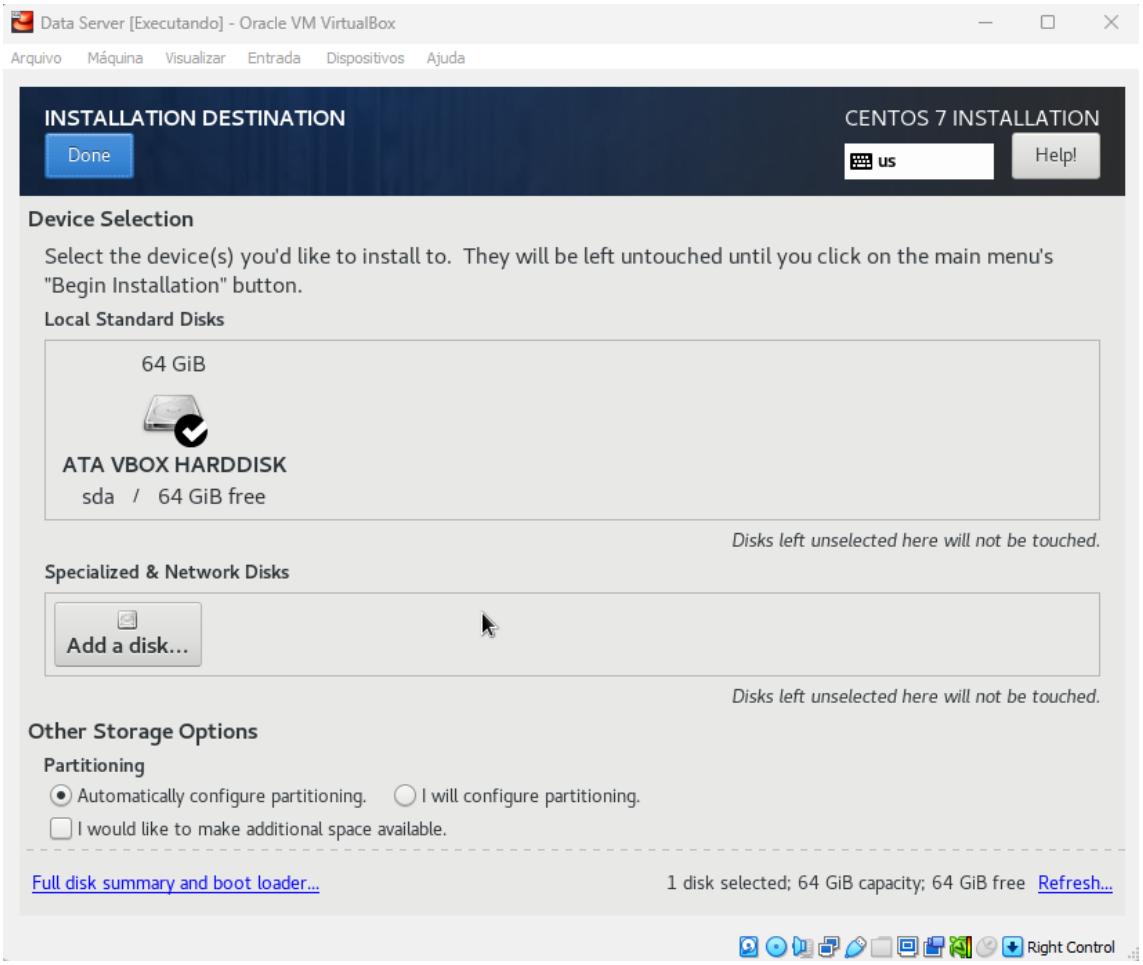
Layout do teclado



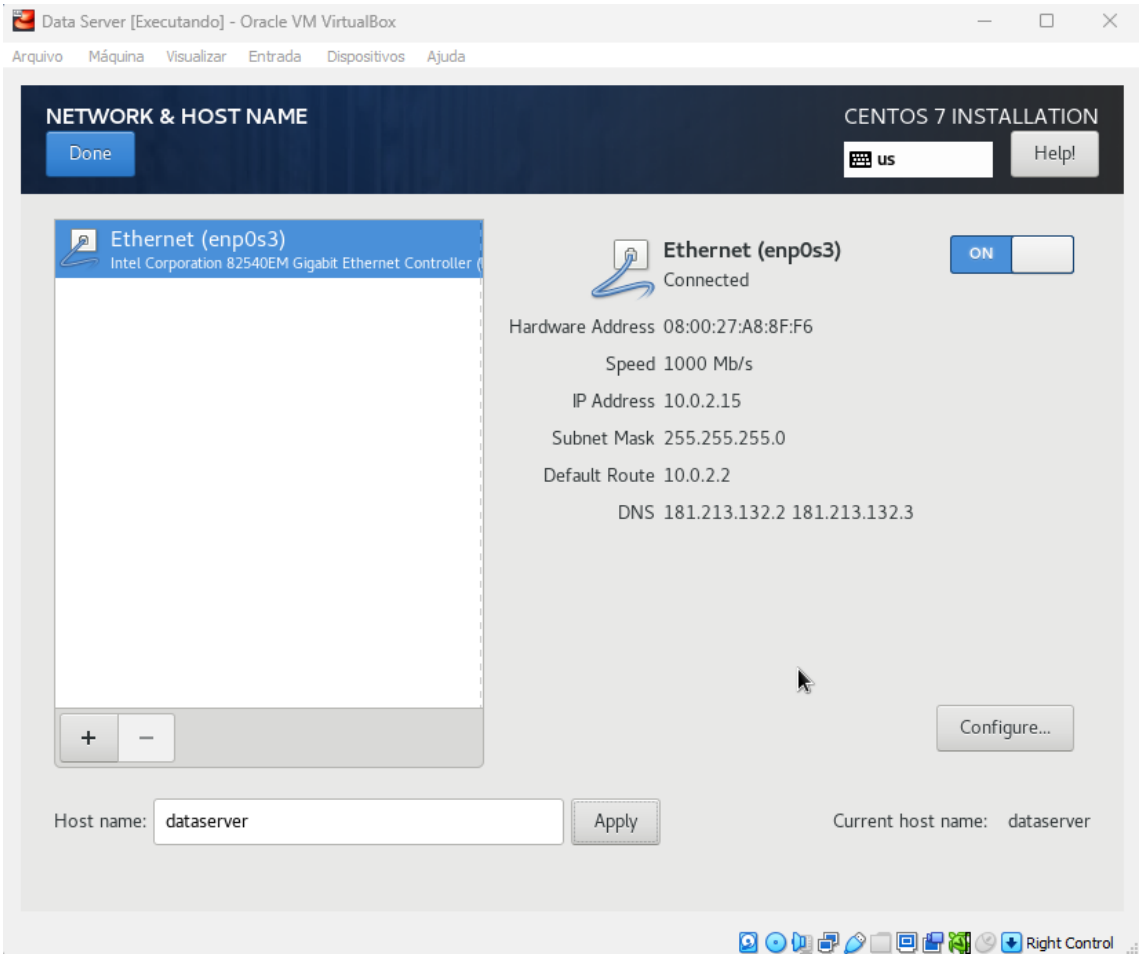
Idioma do sistema operacional



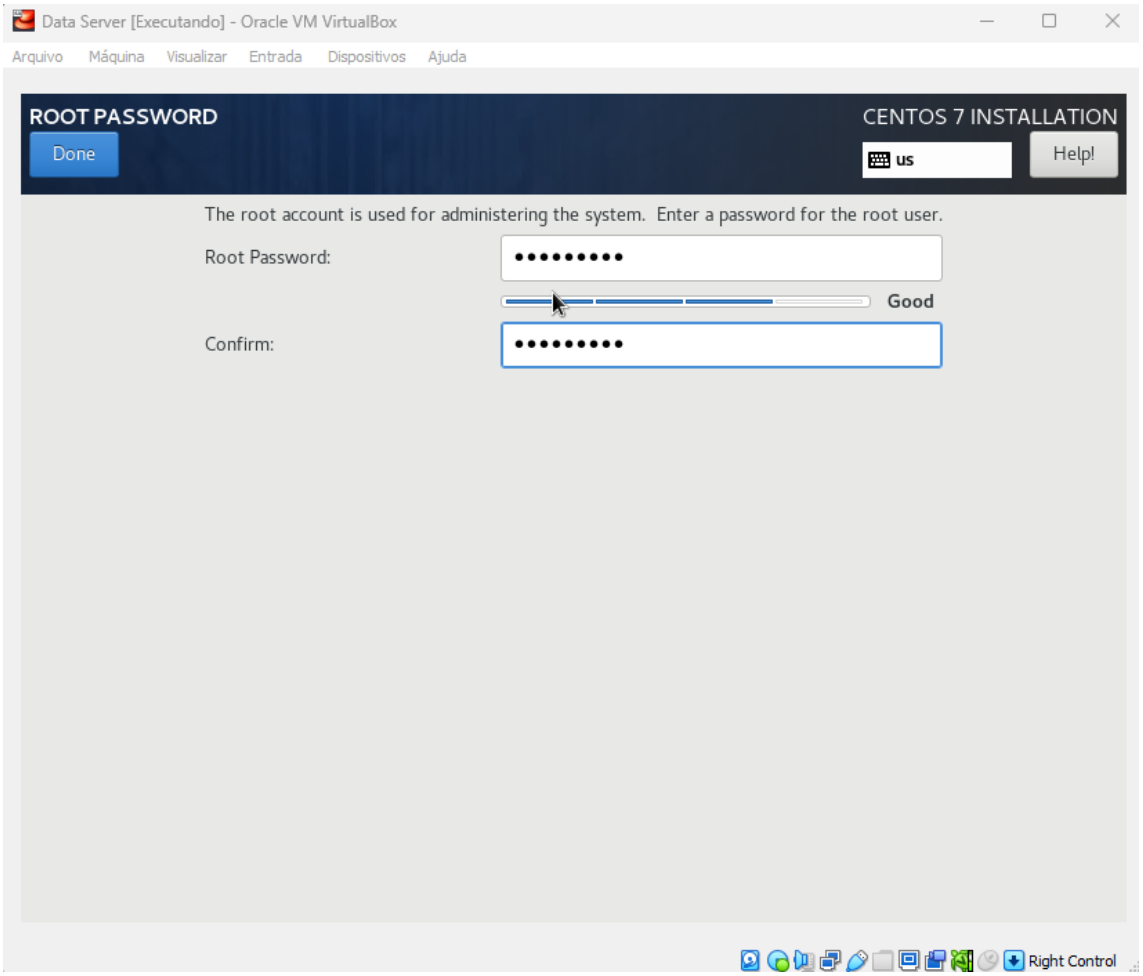
Selecione a opção Server with GUI


Data Server [Executando] - Oracle VM VirtualBox
Arquivo Máquina Visualizar Entrada Dispositivos Ajuda
INSTALLATION DESTINATION
Done
CENTOS 7 INSTALLATION
us
Help!
Device Selection
Select the device(s) you'd like to install to. They will be left untouched until you click on the main menu's "Begin Installation" button.
Local Standard Disks
64 GiB
ATA VBOX HARDDISK
sda / 64 GiB free
Disks left unselected here will not be touched.
Specialized & Network Disks
Add a disk...
Disks left unselected here will not be touched.
Other Storage Options
Partitioning
Automatically configure partitioning.
I will configure partitioning.
I would like to make additional space available.
Full disk summary and boot loader...
1 disk selected; 64 GiB capacity; 64 GiB free
Refresh...
Right Control

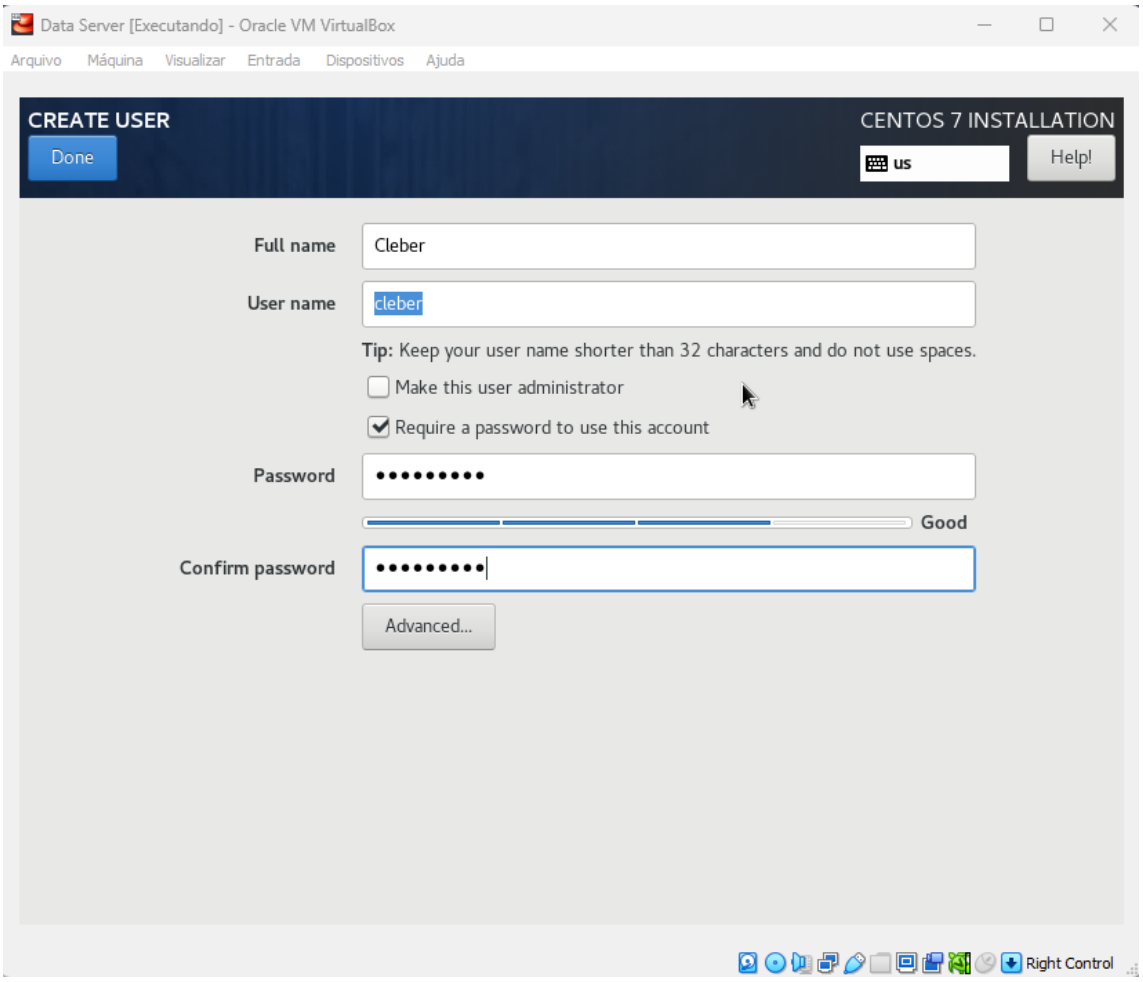
Disco



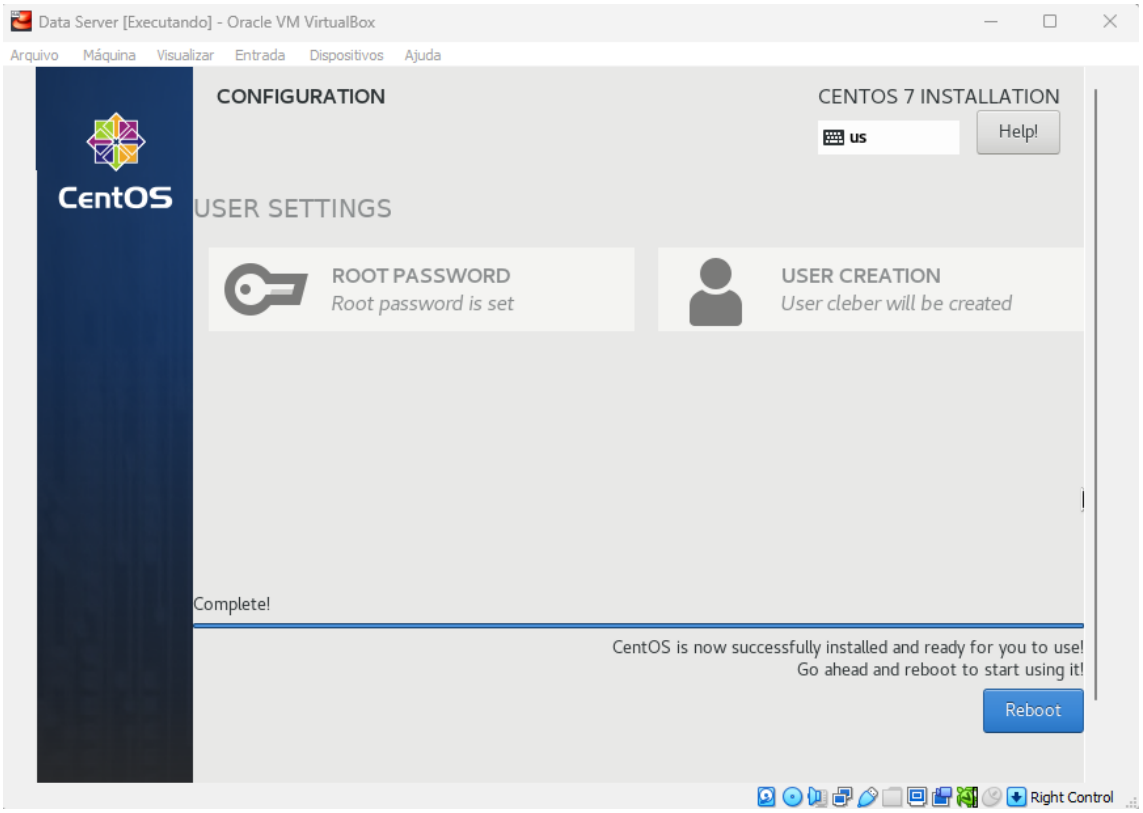
Configuração de rede e nome do servidor – **dataserver –** Clique em Apply
Certifique-se de habilitar a opção de ativar a Ethernet (botão on)



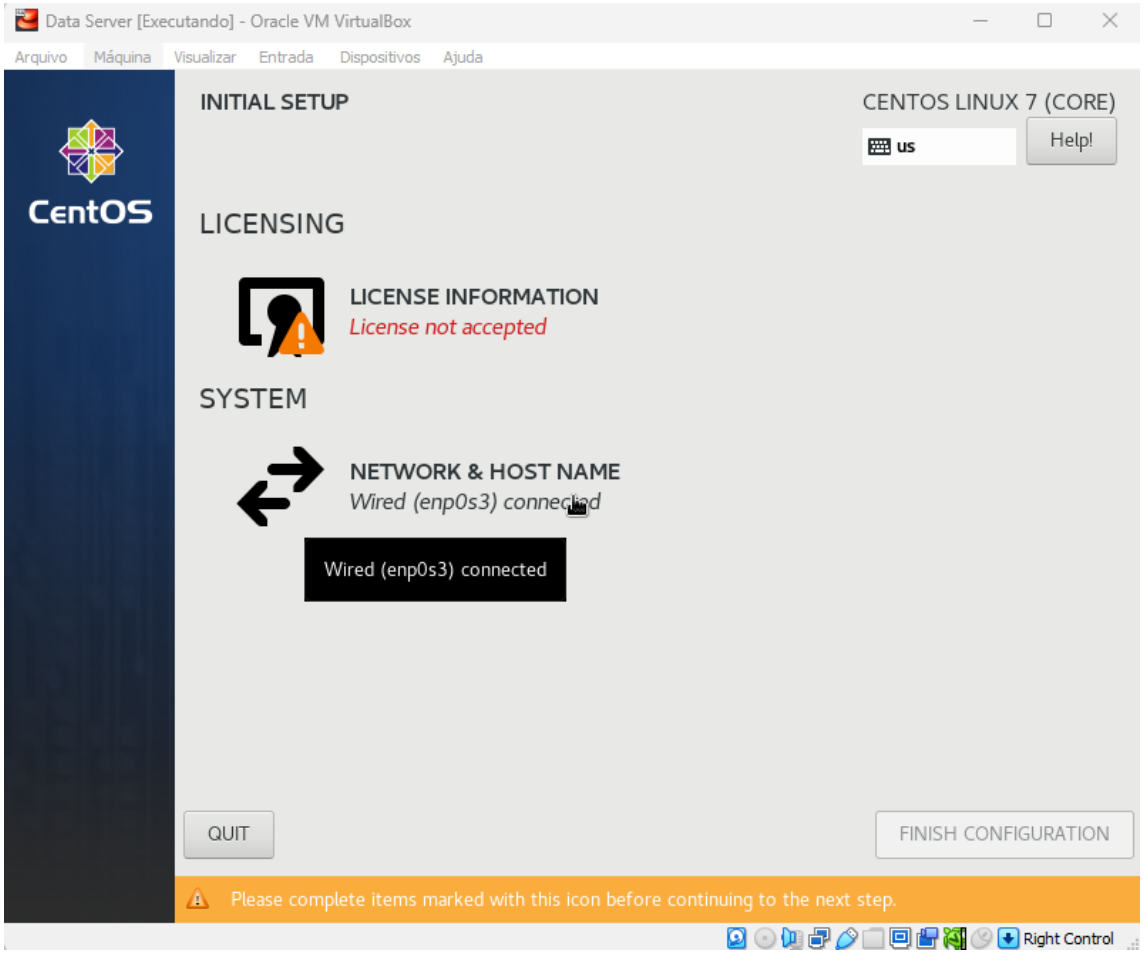
Definir senha do root – usuário administrador
Senha: **hadoop123**



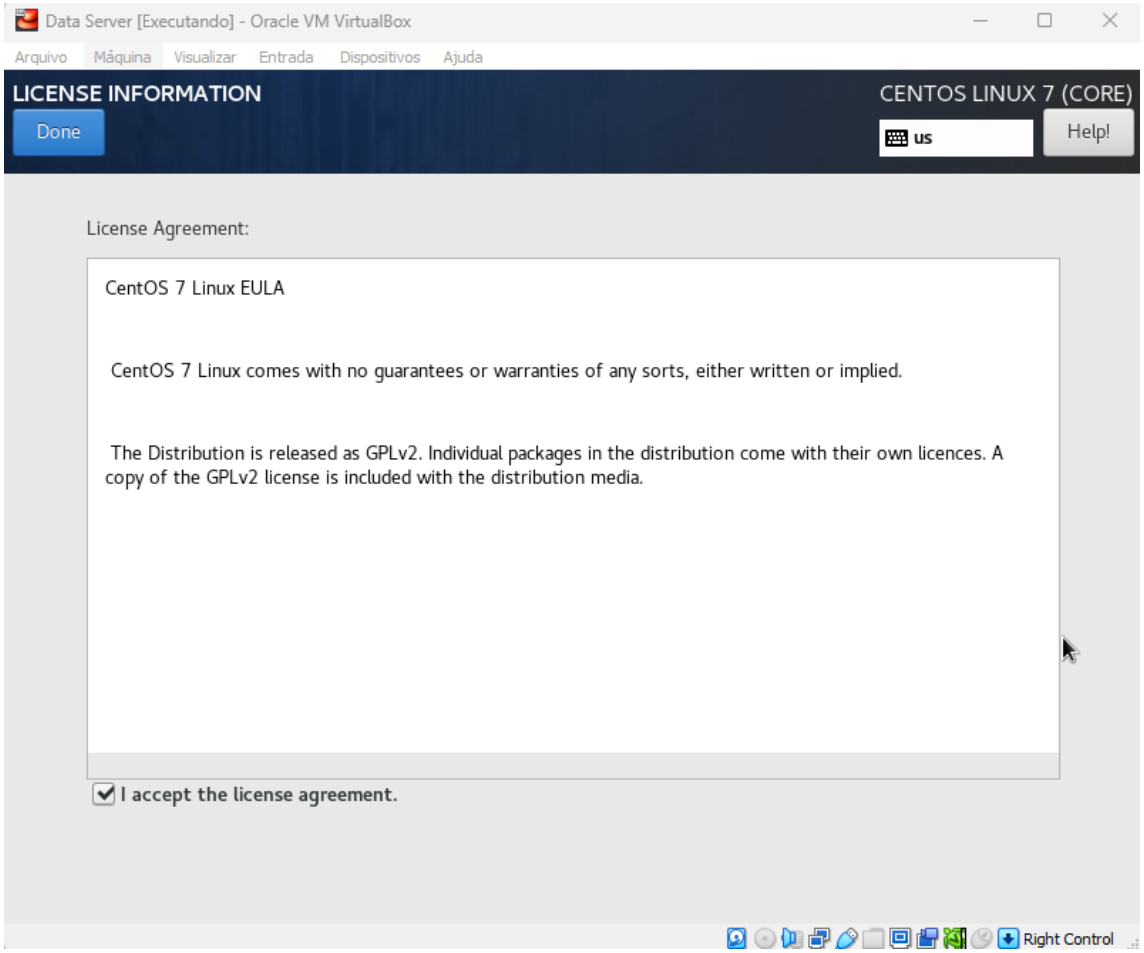
Criação de um usuário – Cleber
(username: cleber, senha: **hadoop123**)



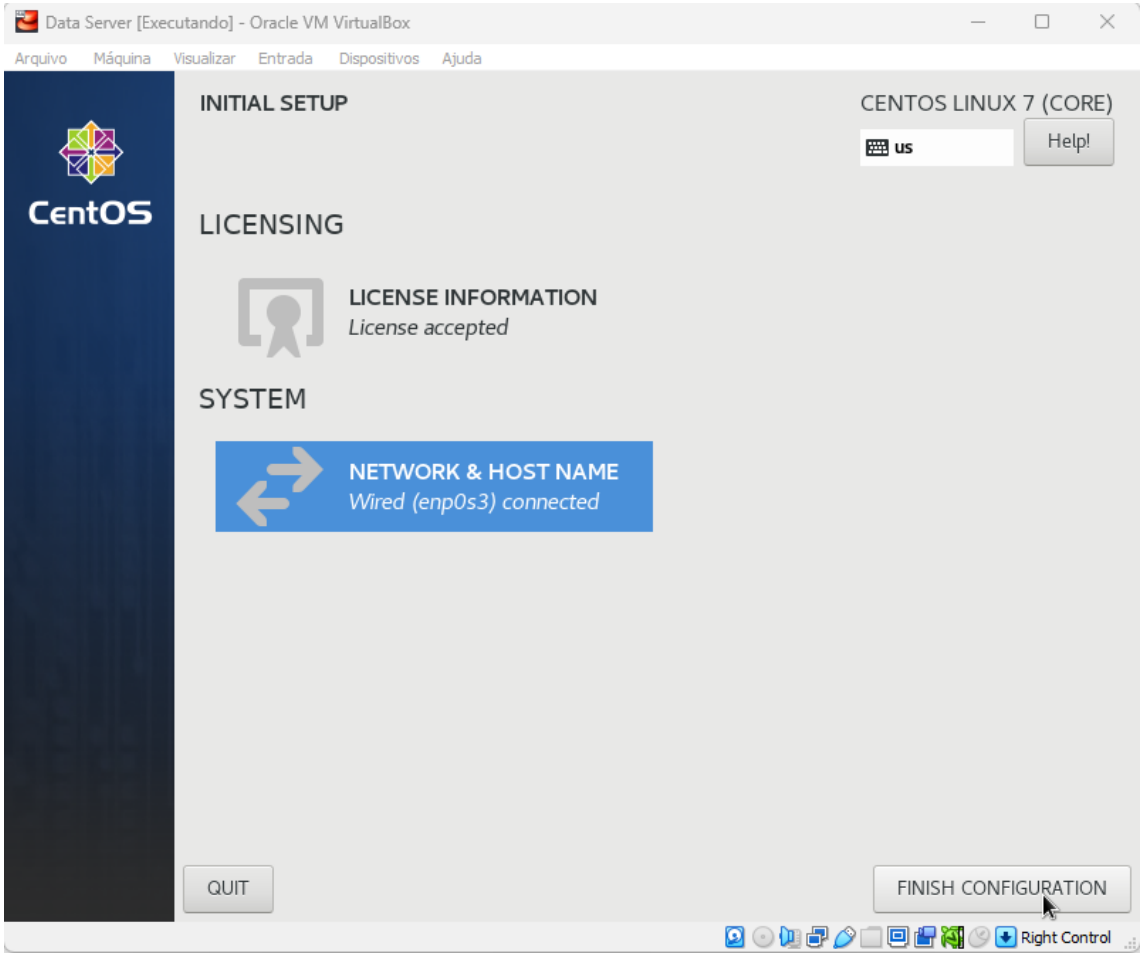
Conclusão da instalação. Clique no botão Reboot.



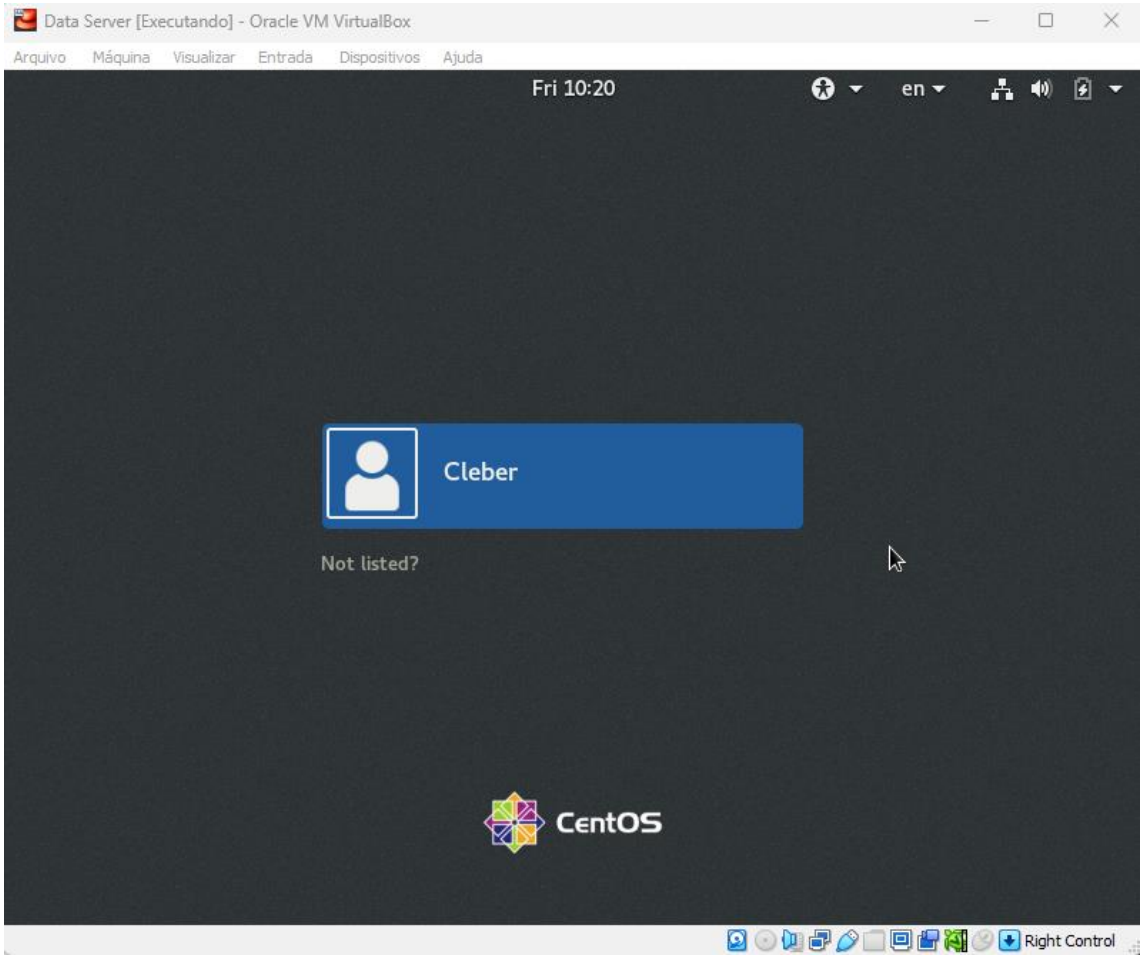
Clique em License Information para aceitar os termos de uso



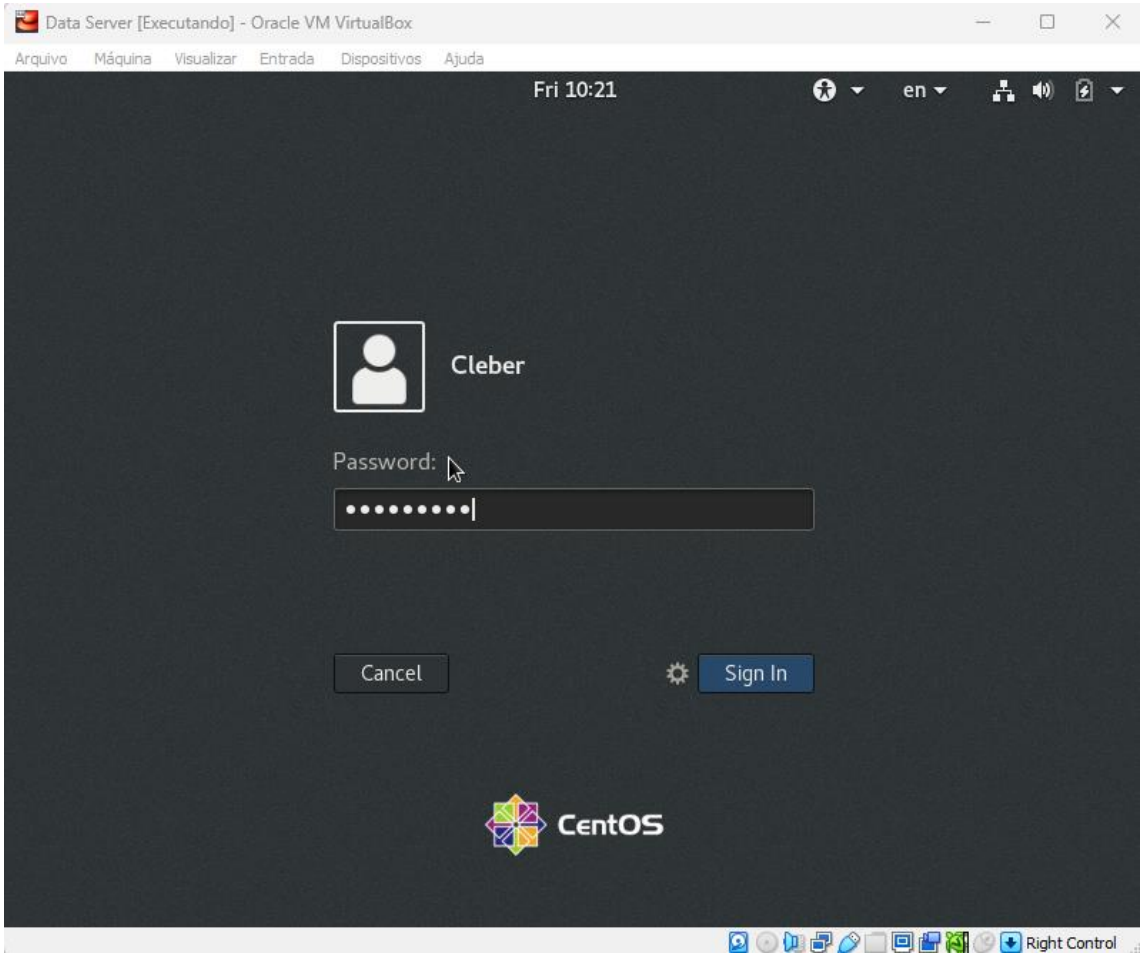
Marque a caixa e pressione Done.



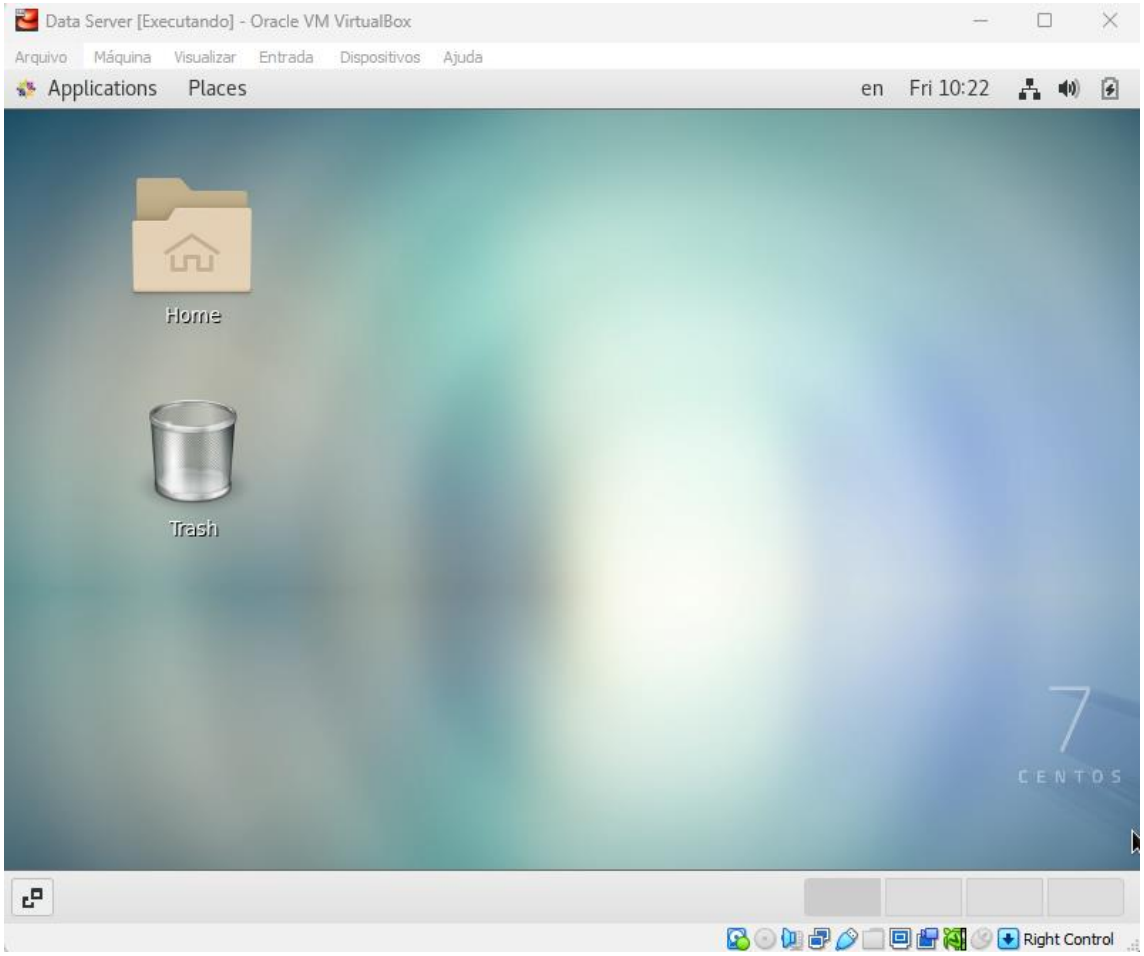
Se necessário, revise se o nome da máquina está correto e se a rede está ativada e pressione “Finish Configuration”



Clique no usuário Cleber

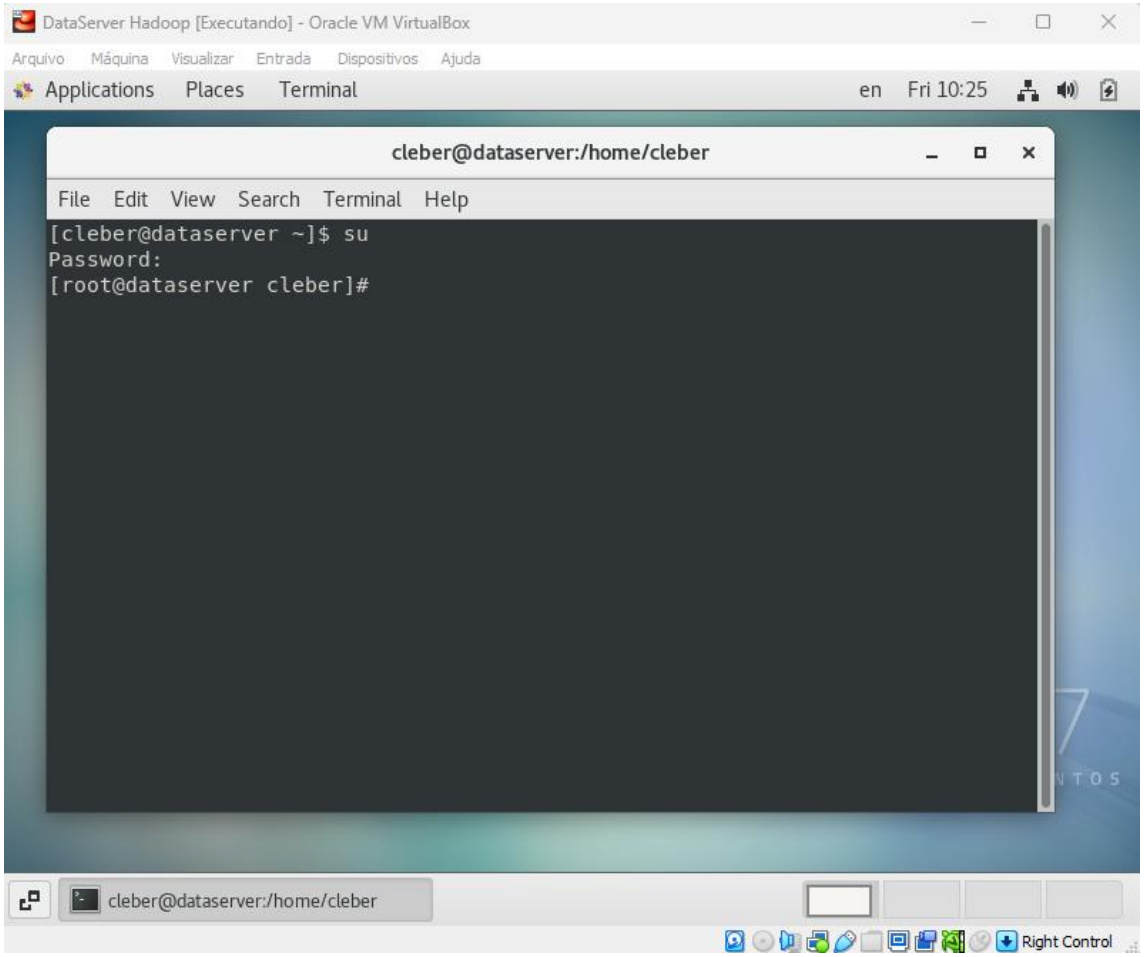


Usuário/Senha (**hadoop123**)

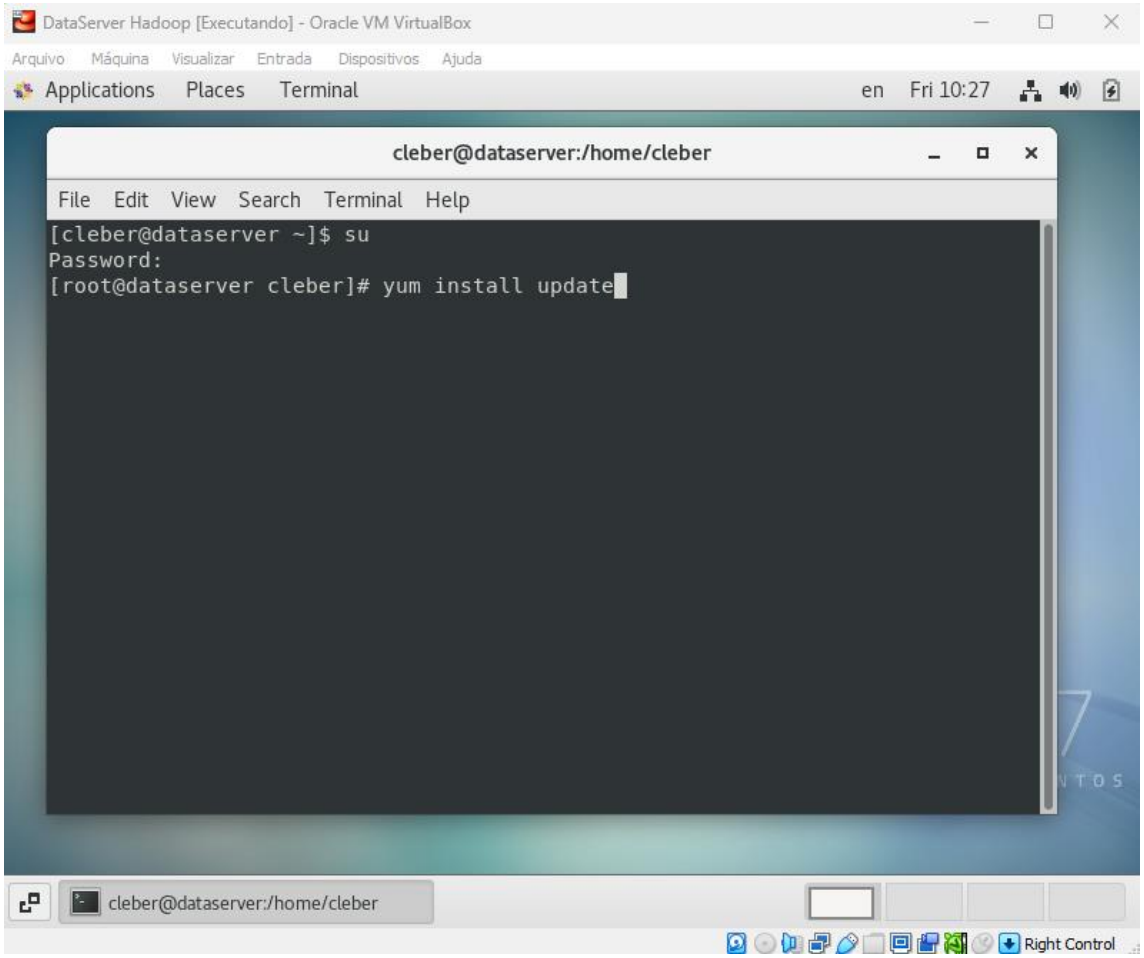


Instalação concluída com sucesso

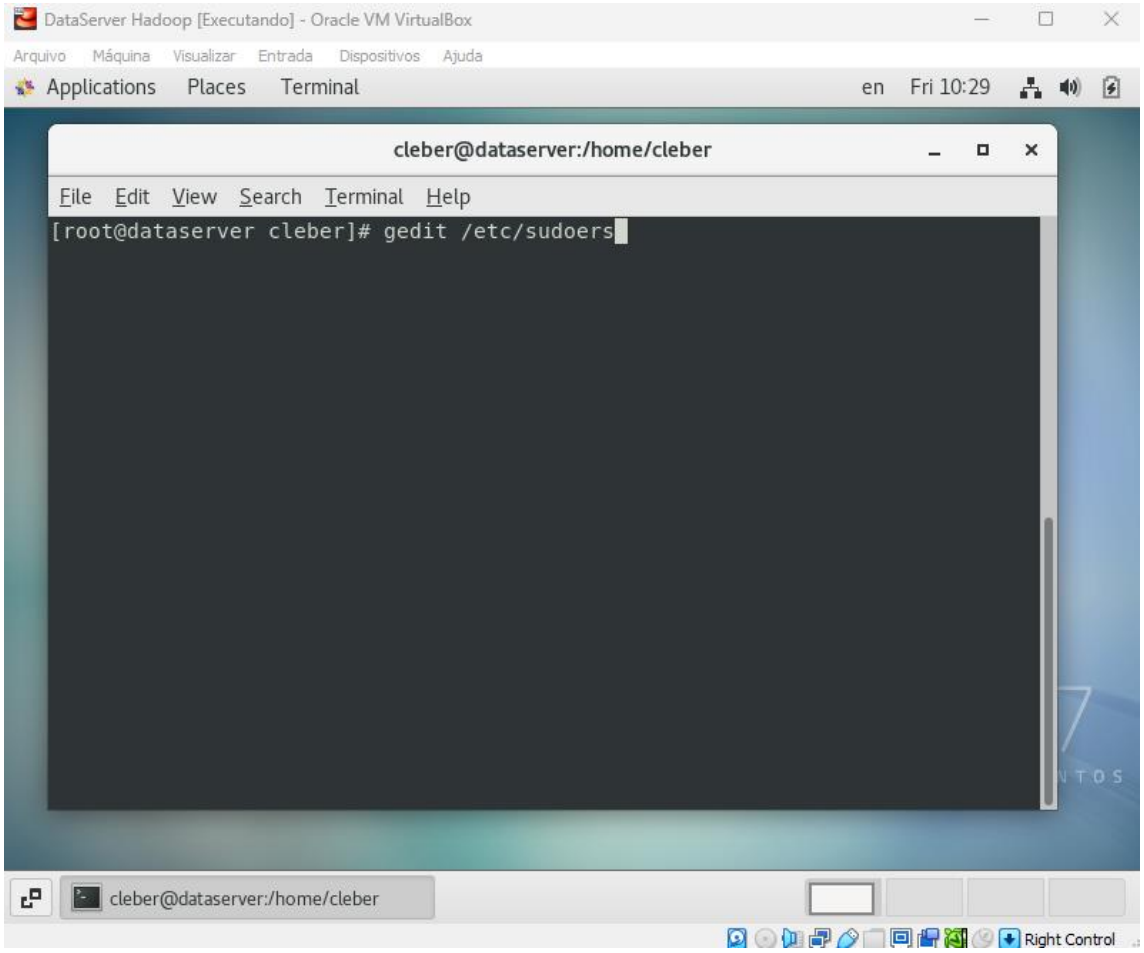
2.3 Instalação de Utilitários do Sistema Operacional



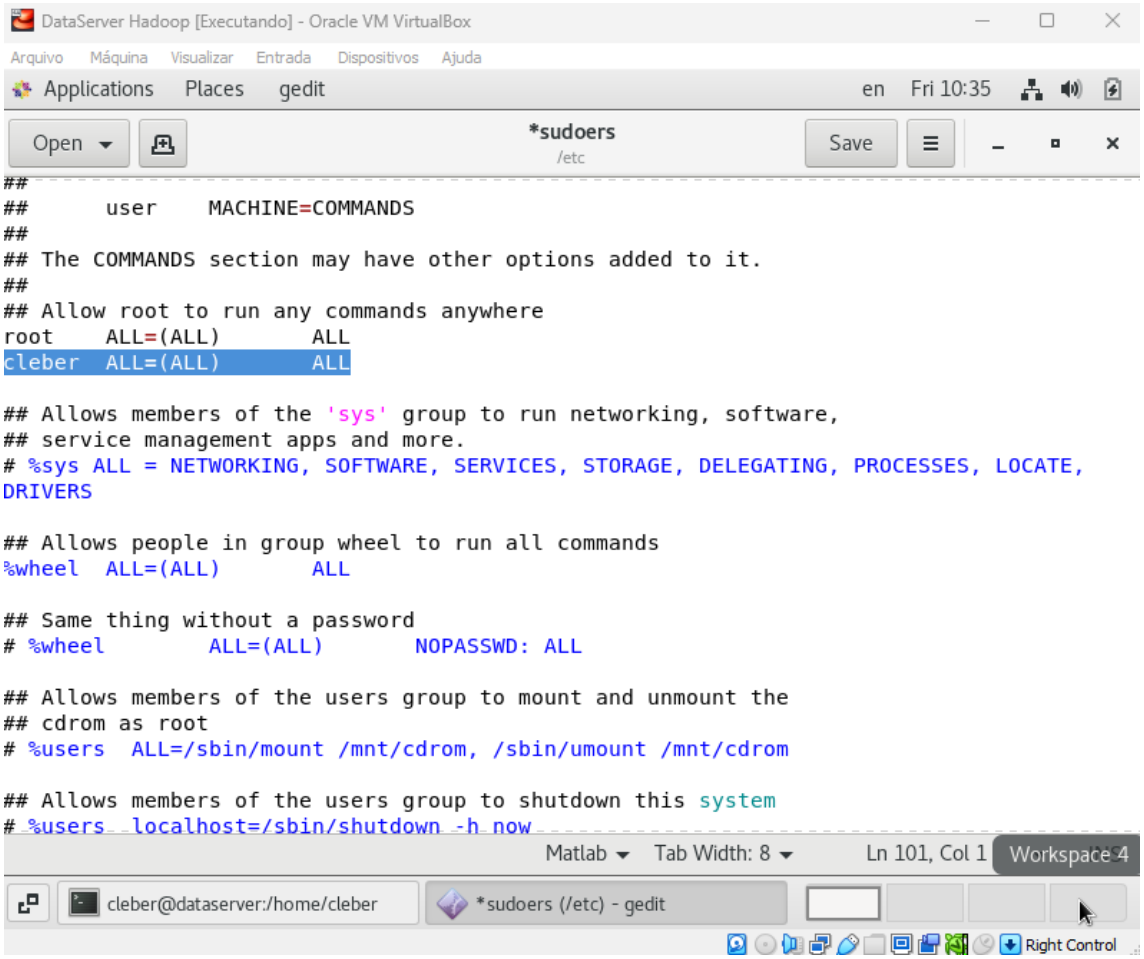
Efetuar login como root, usando o comando su. Senha: **hadoop123**



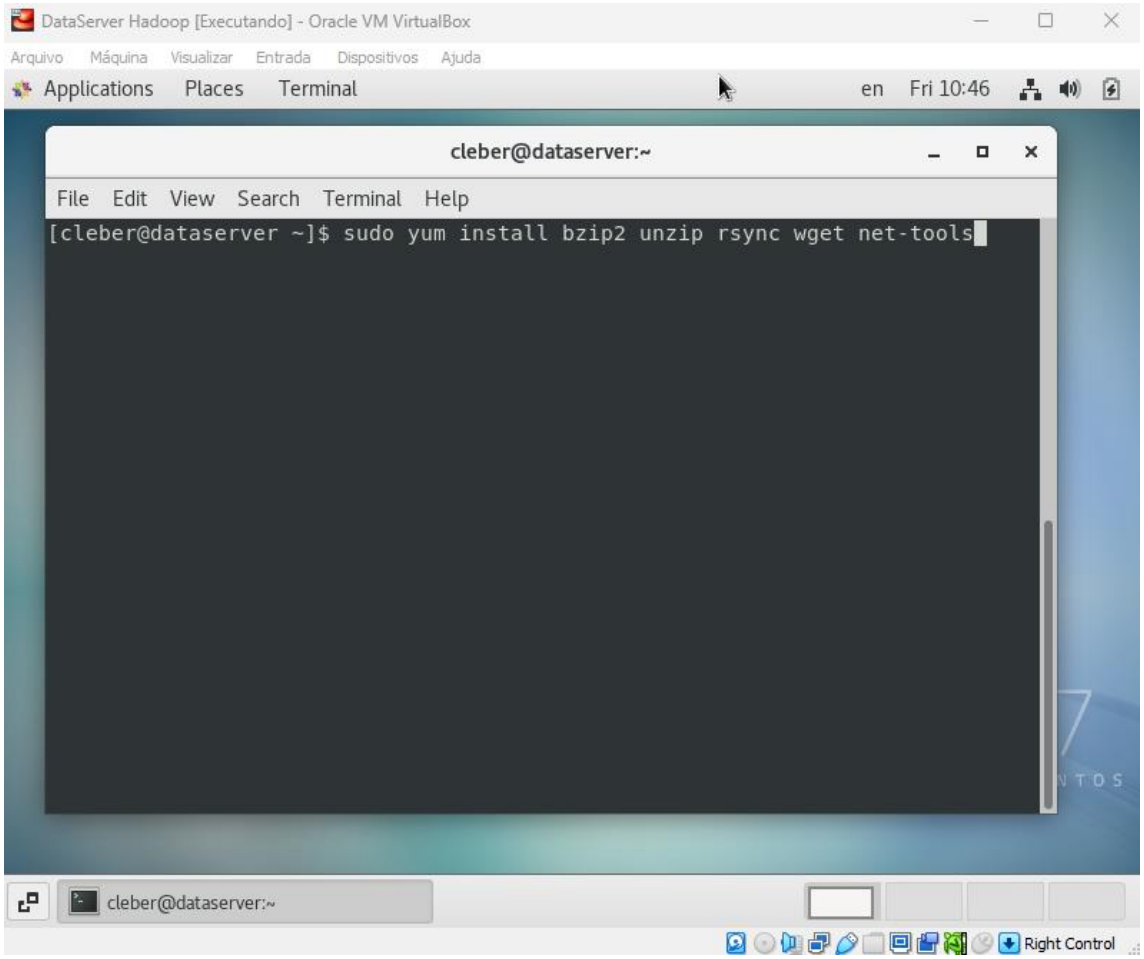
Atualizando o Sistema Operacional



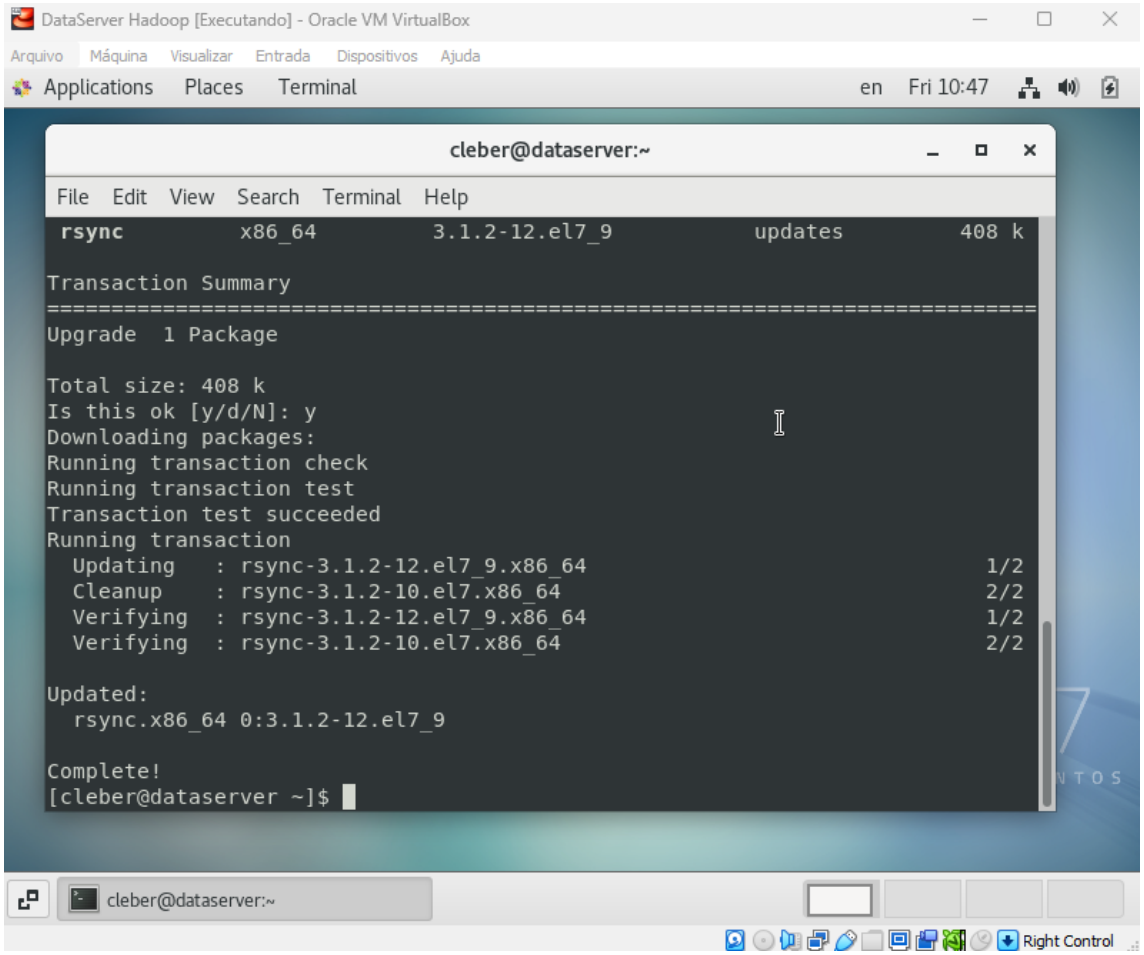
Editar o arquivo /etc/sudoers usando o gedit



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

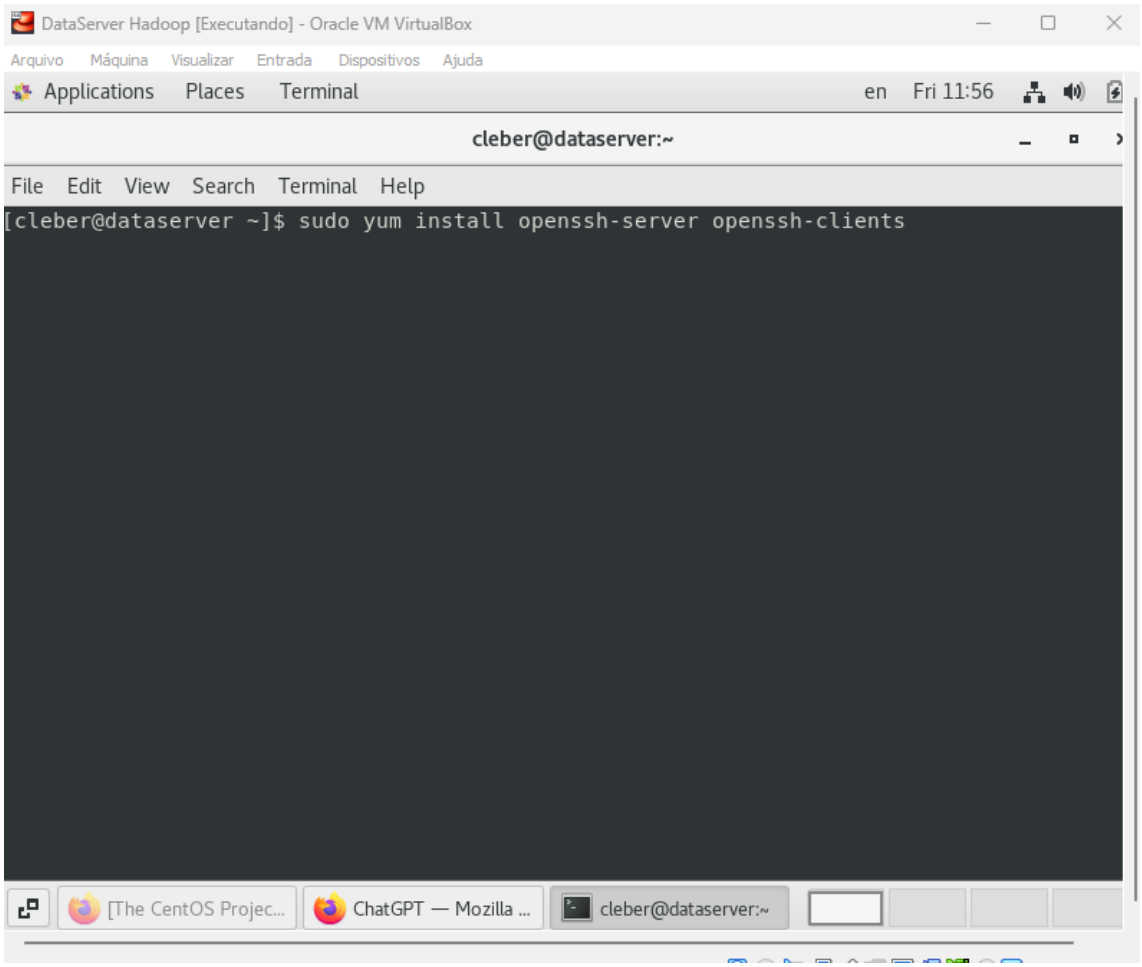


Instalar outros aplicativos: bzip2, unzip, rsync, wget e net-tools

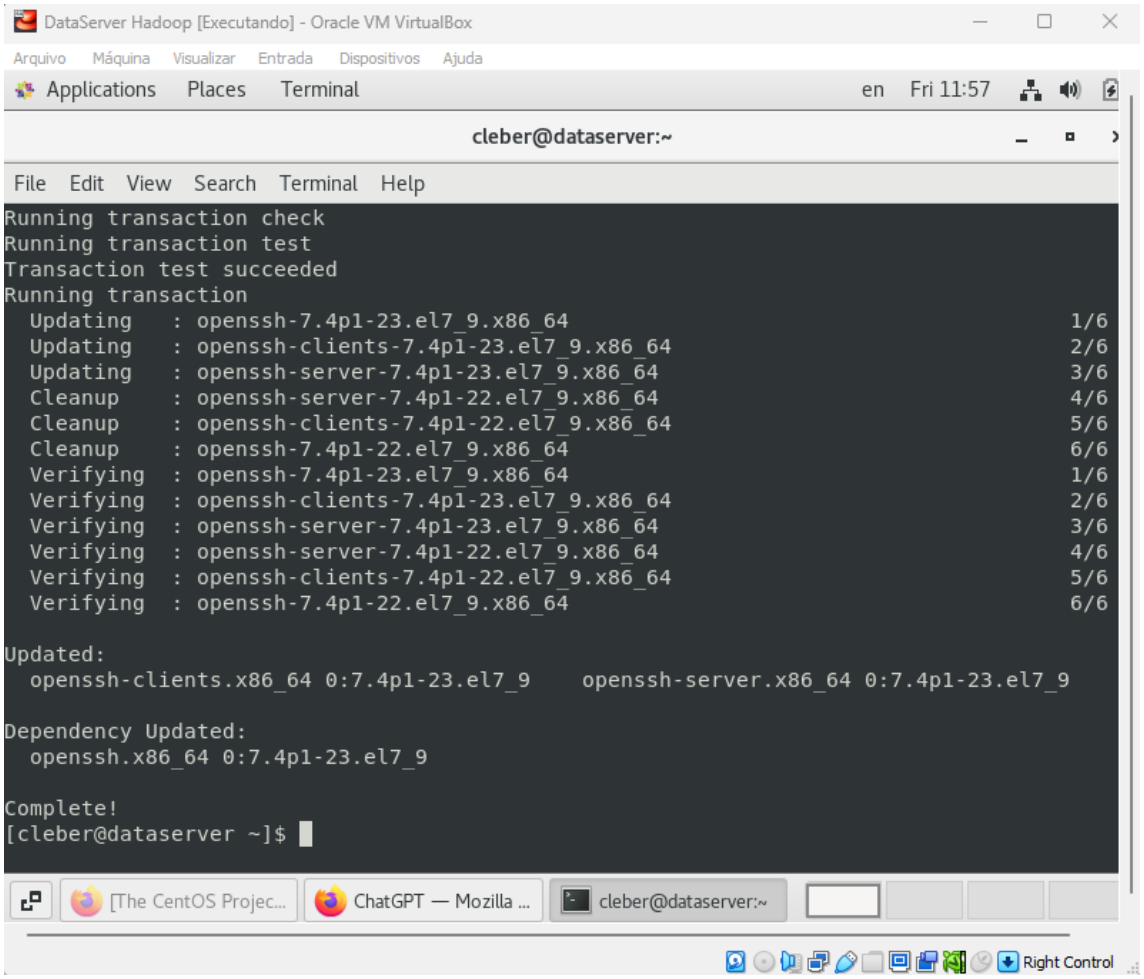


Aplicativos instalados

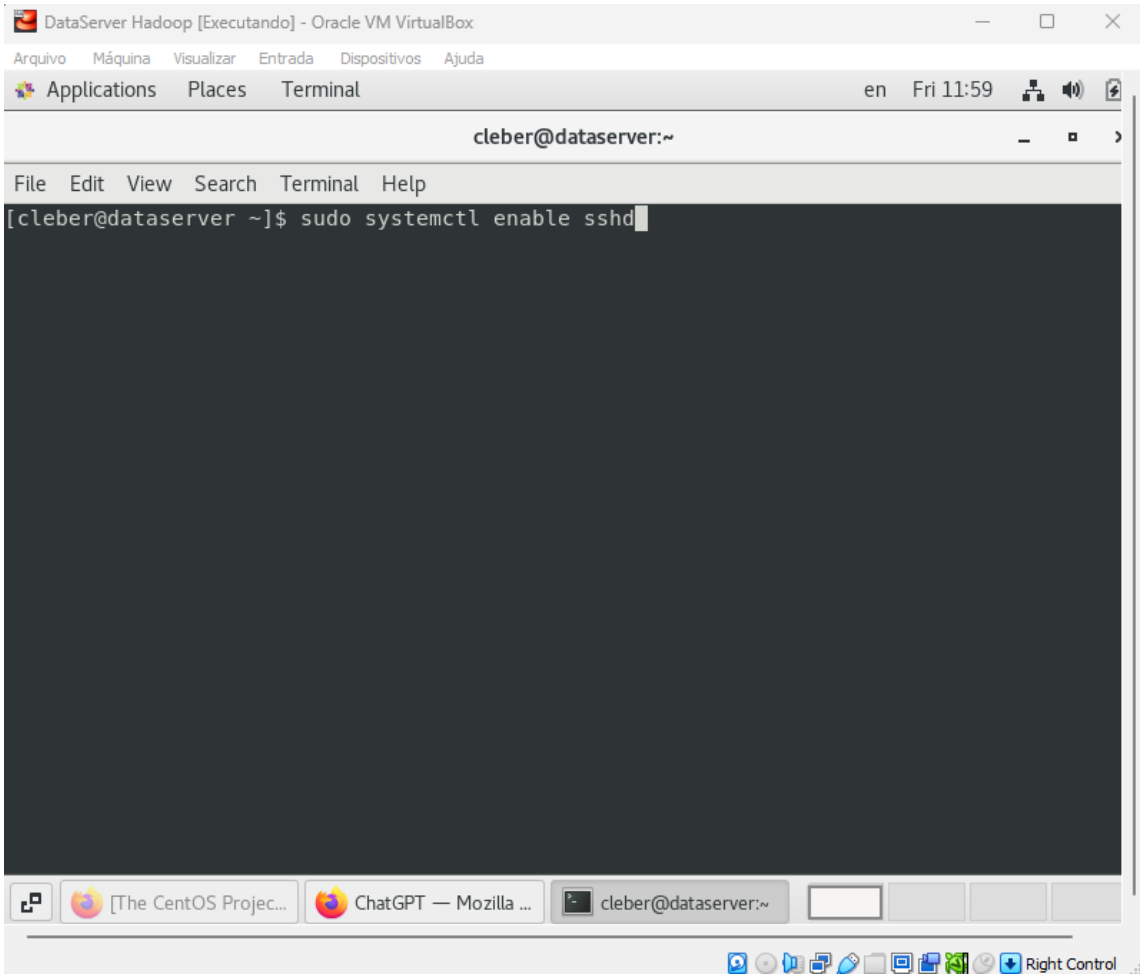
3 Instalação do servidor ssh



sudo yum install openssh-server openssh-clients

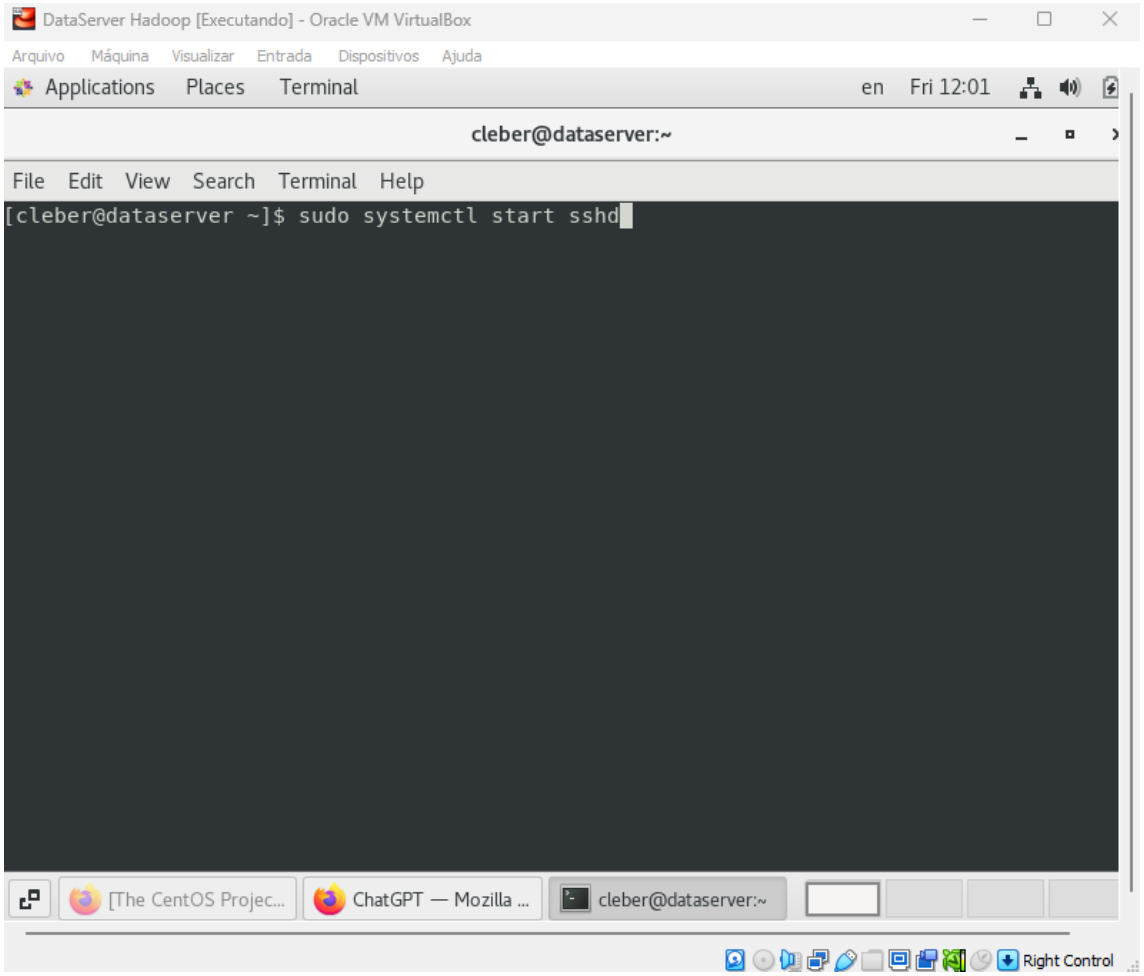

DataServer Hadoop [Executando] - Oracle VM VirtualBox
Arquivo Máquina Visualizar Entrada Dispositivos Ajuda
Applications Places Terminal en Fri 11:57
cleber@dataserver:~
File Edit View Search Terminal Help
Running transaction check
Running transaction test
Transaction test succeeded
Running transaction
Updating : openssh-7.4p1-23.el7_9.x86_64 1/6
Updating : openssh-clients-7.4p1-23.el7_9.x86_64 2/6
Updating : 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 ~]$
[The CentOS Projec...
ChatGPT — Mozilla ...
cleber@dataserver:~
Right Control

Concluído

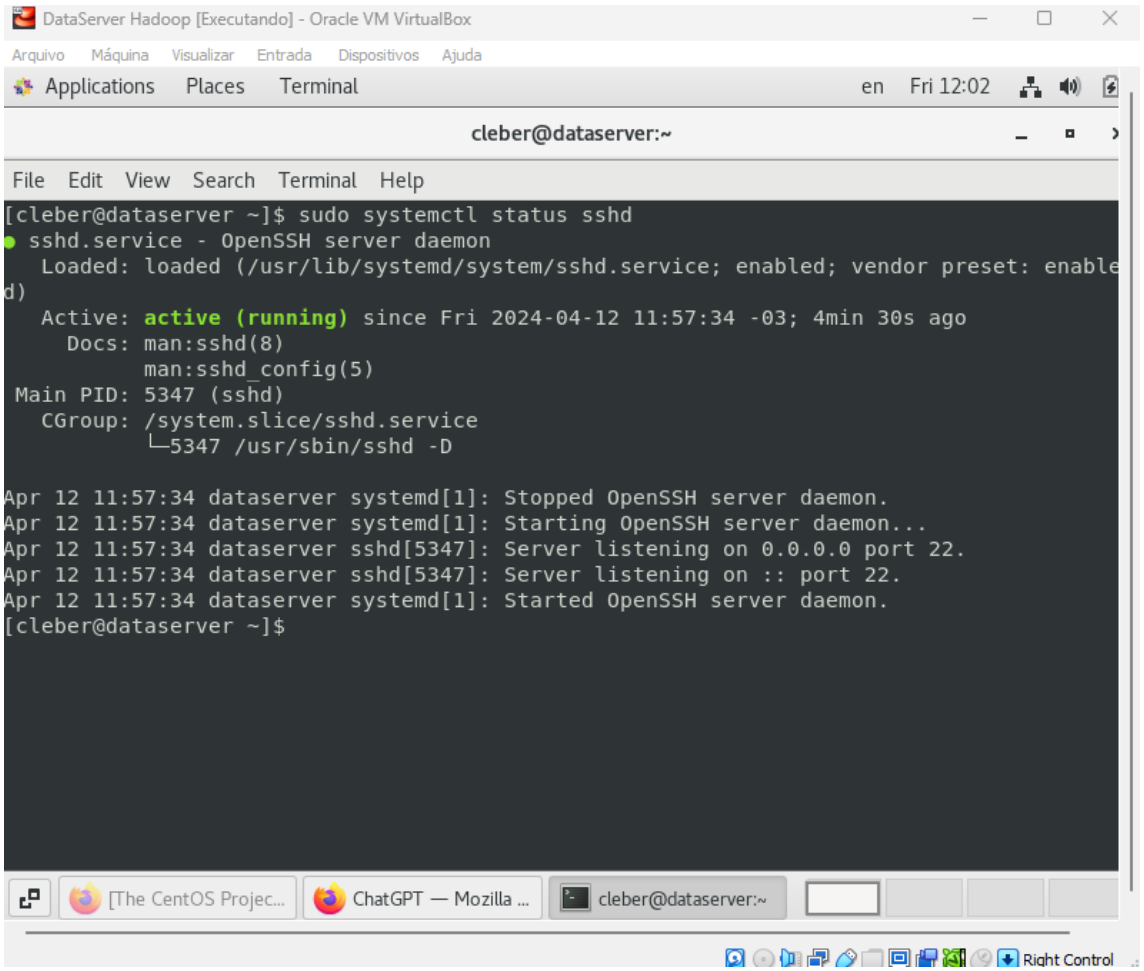


Habilitando o serviço

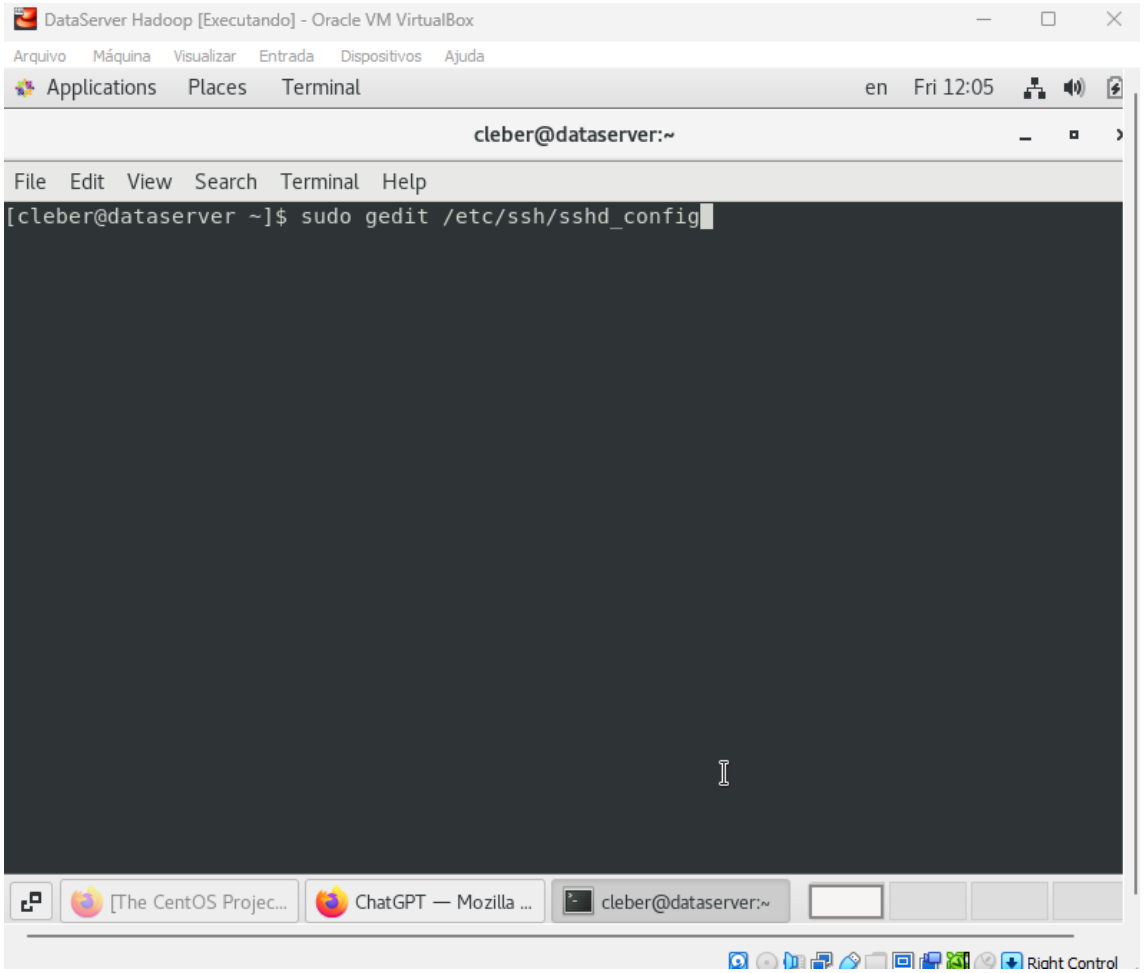
sudo systemctl enable sshd



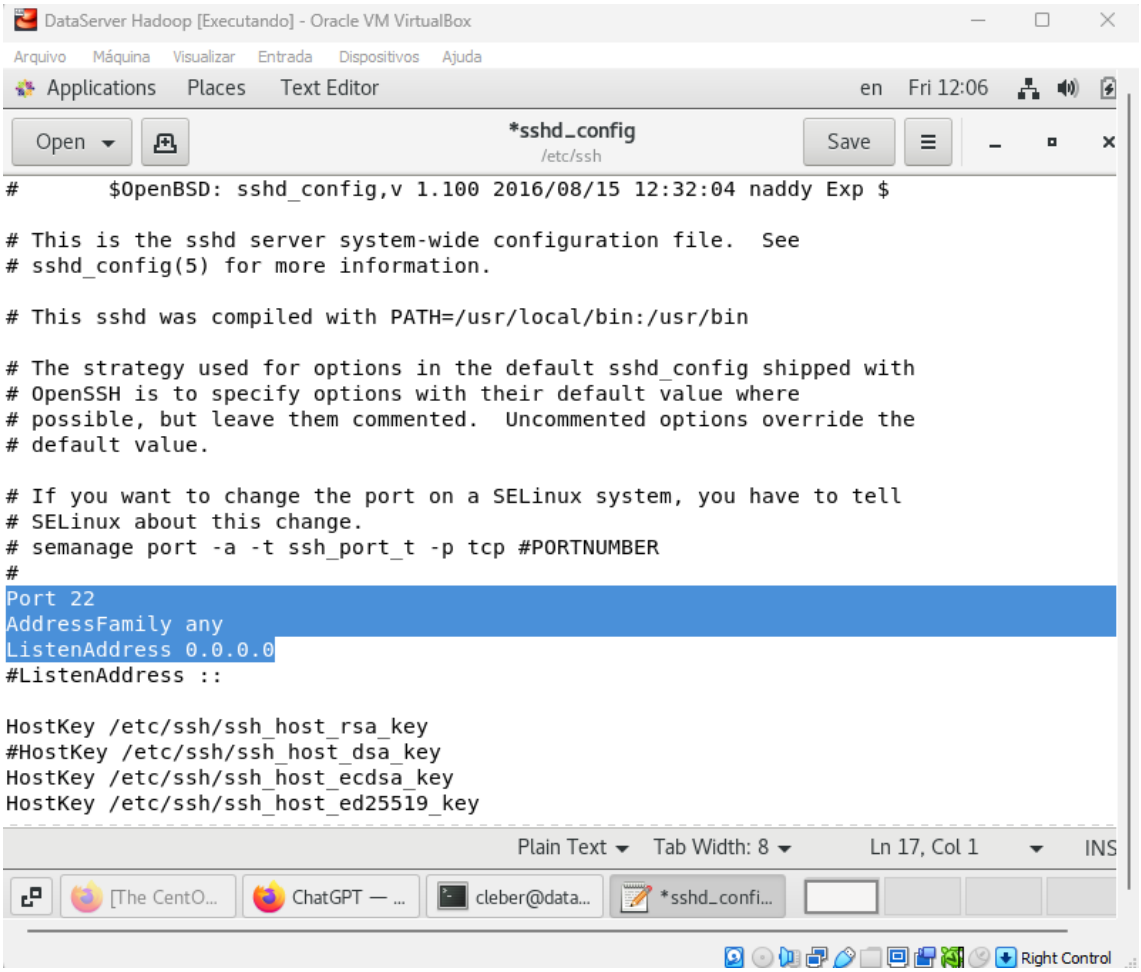
Iniciando o serviço
sudo systemctl start sshd



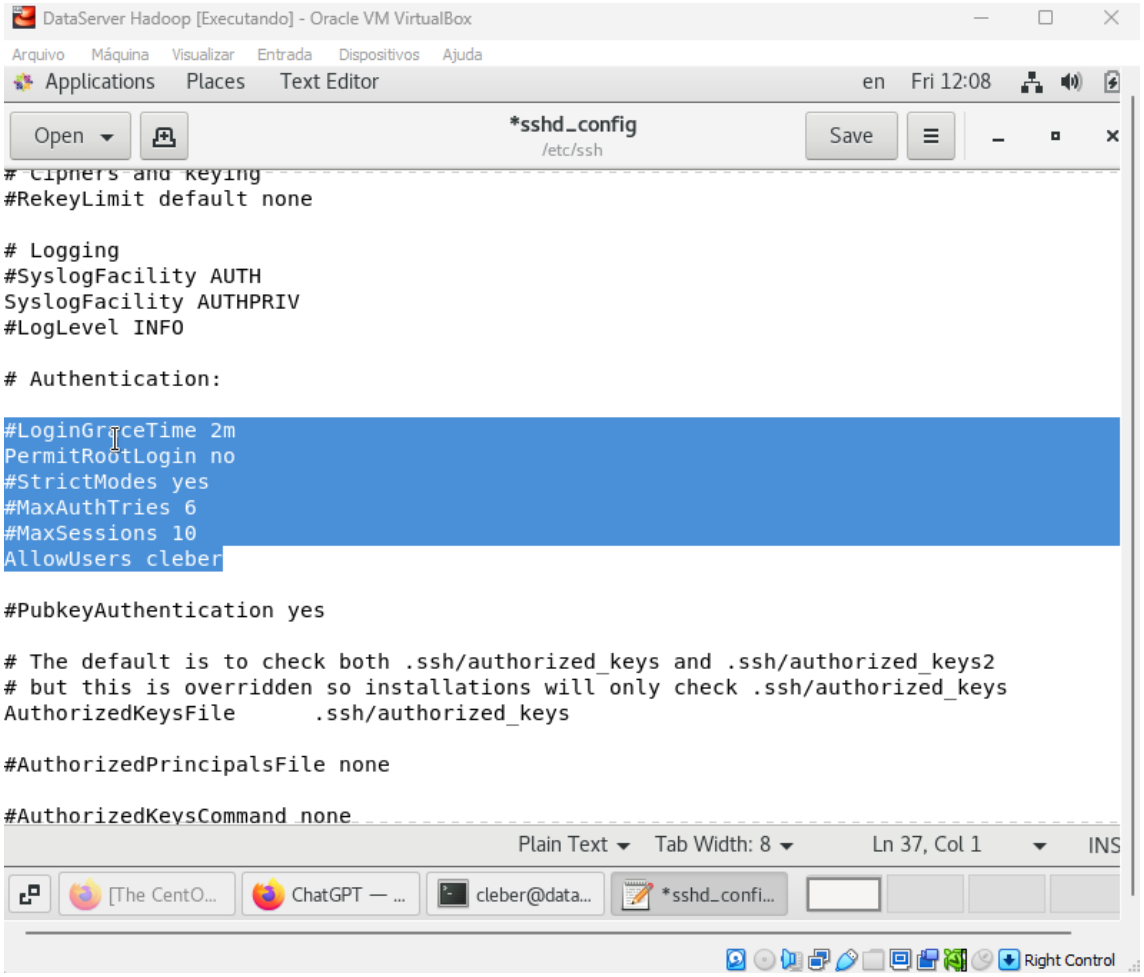
Serviço em execução



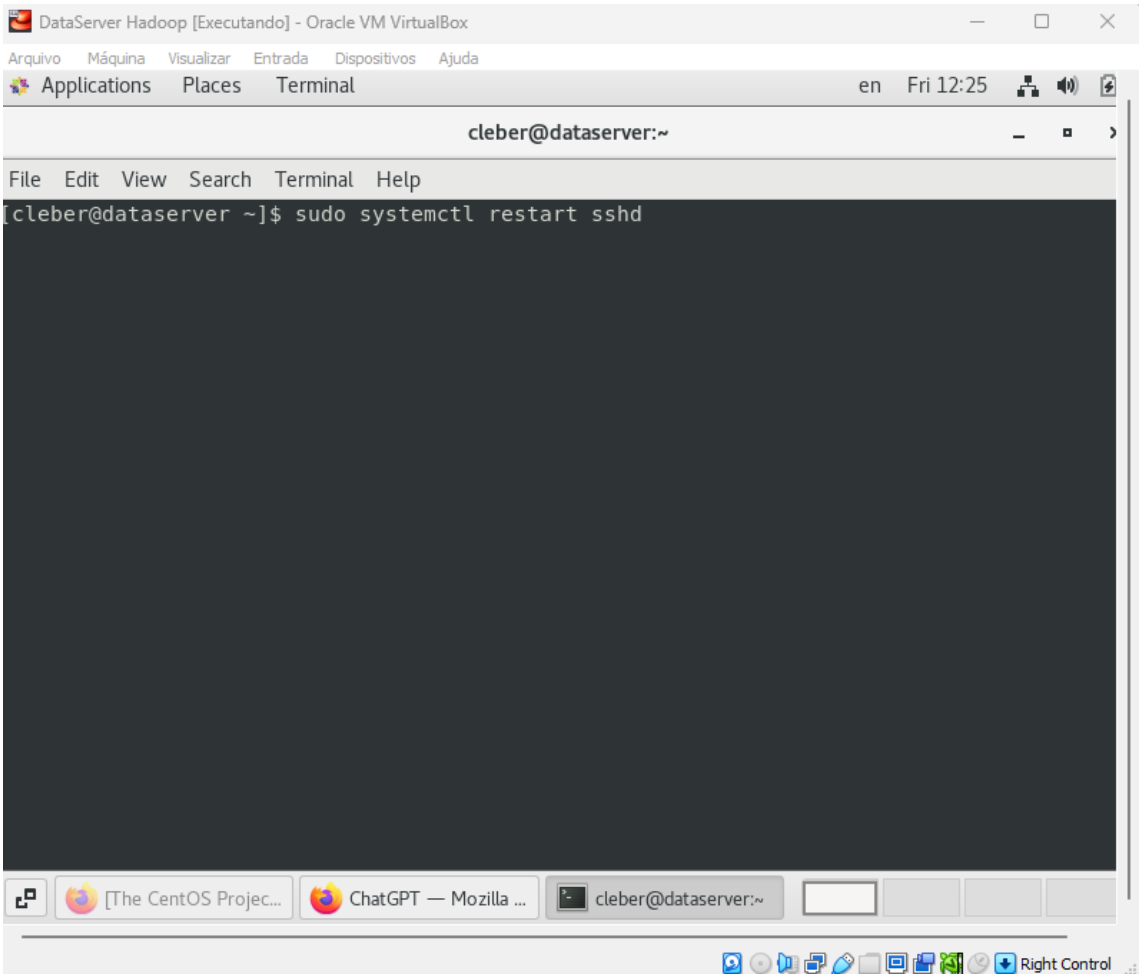
sudo gedit /etc/ssh/sshd_config



Primeira parte da configuração ssh.
Remova o símbolo (#) de comentário das 3 linhas marcadas acima.



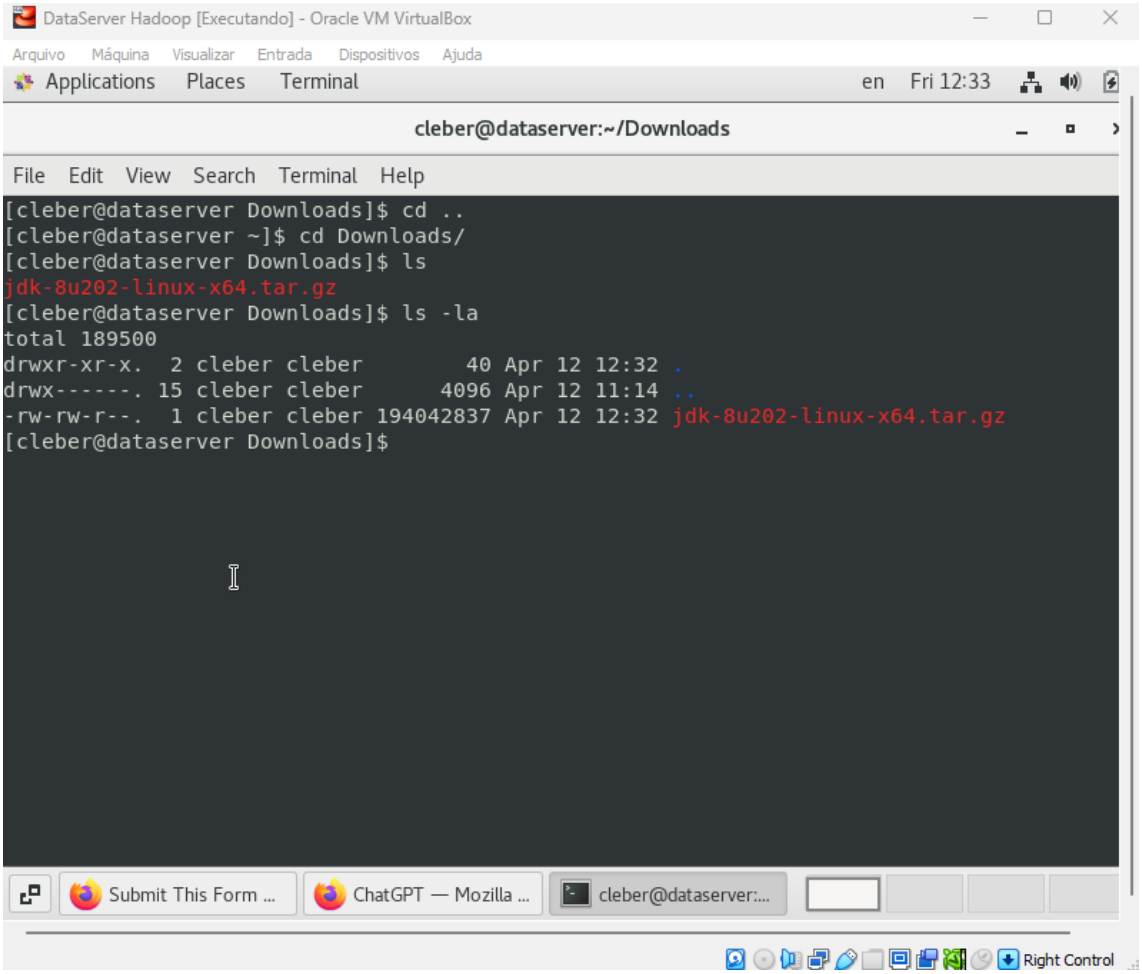
Segunda parte da configuração do ssh



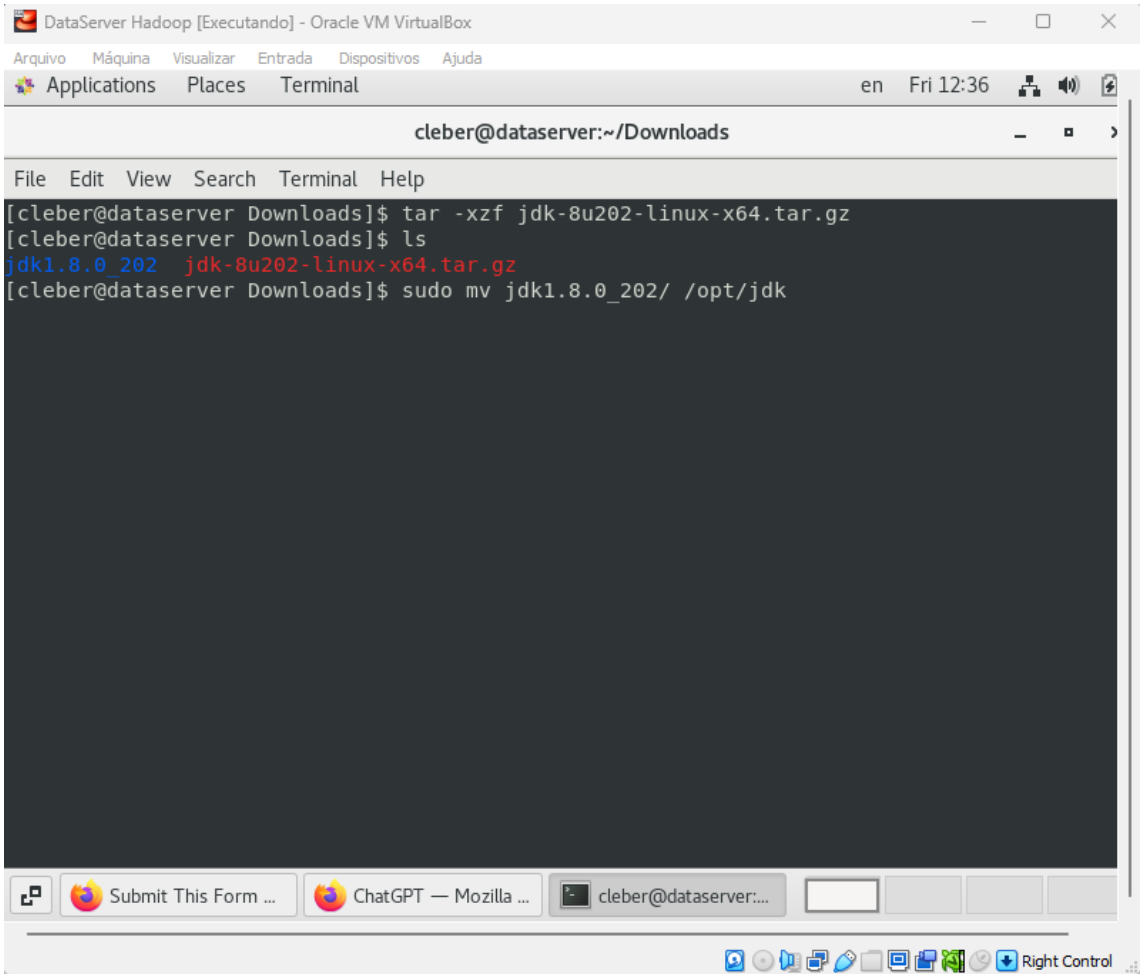
sudo systemctl restart sshd

4 Instalação do Java 8

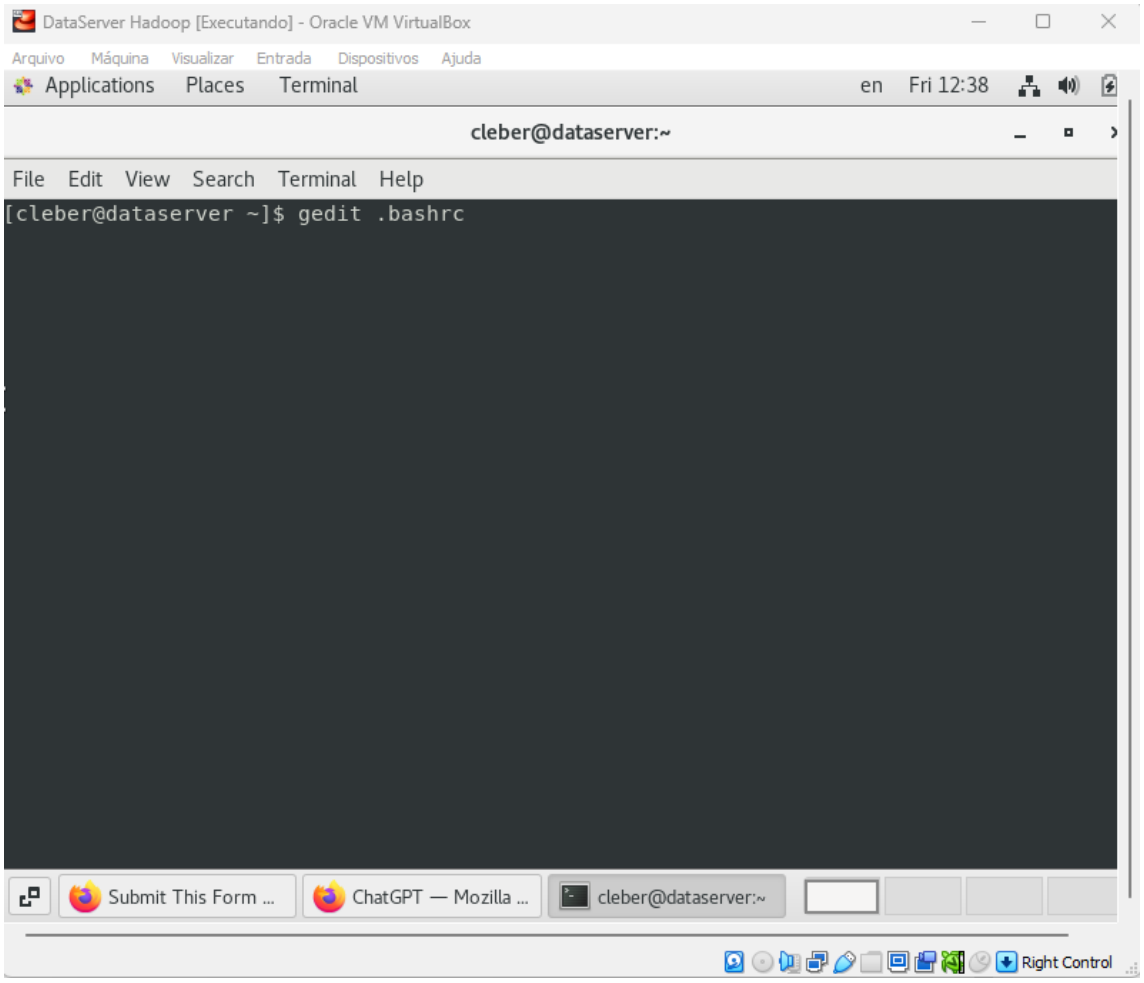
Acesse o site da Oracle e faça download do Java JDK 1.8 para Linux



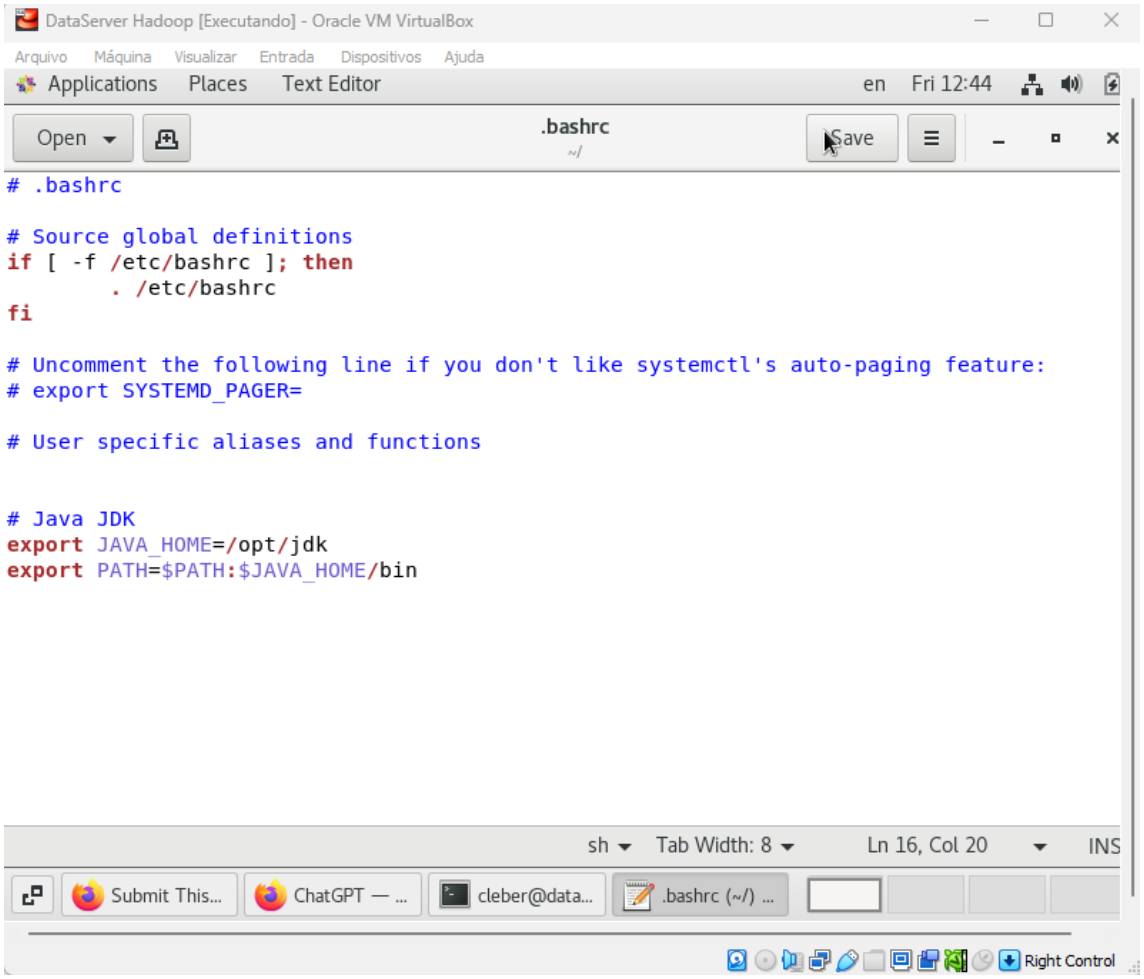
Executar o comando tar para descompactar o arquivo: **tar -xzf jdk-8u212-linux-x64.tar.gz**



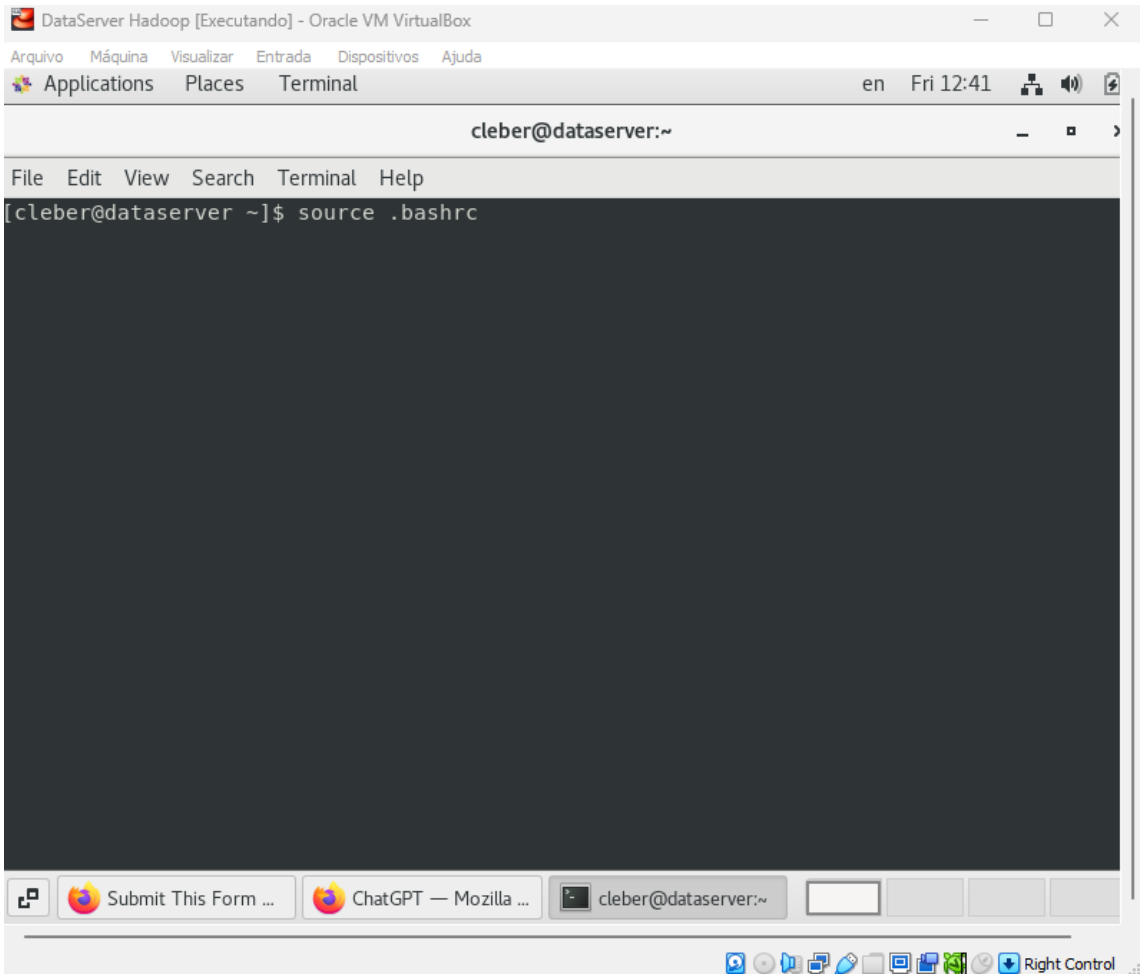
Mover o diretório do JDK



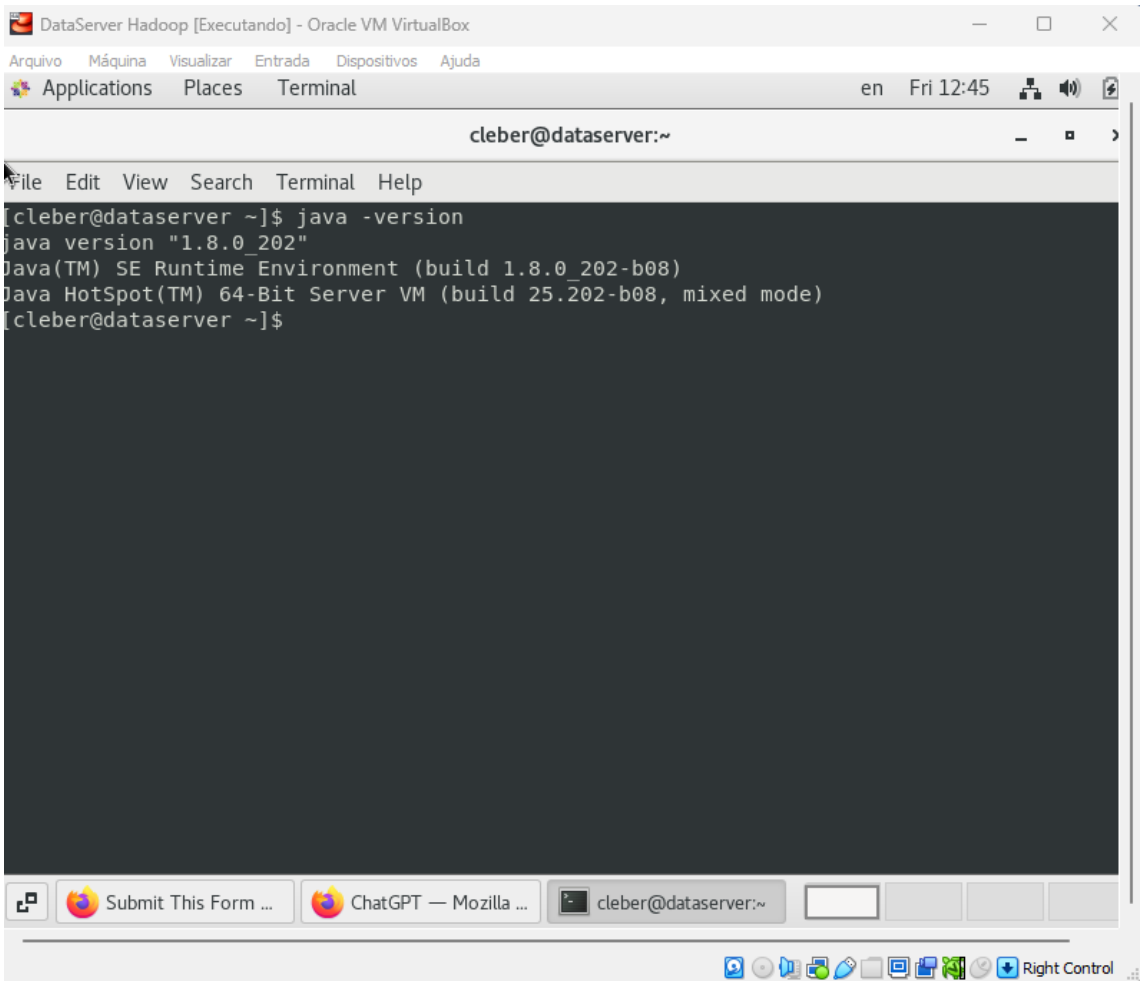
gedit .bashrc



Editar as variáveis de ambiente conforme acima e salvar o arquivo



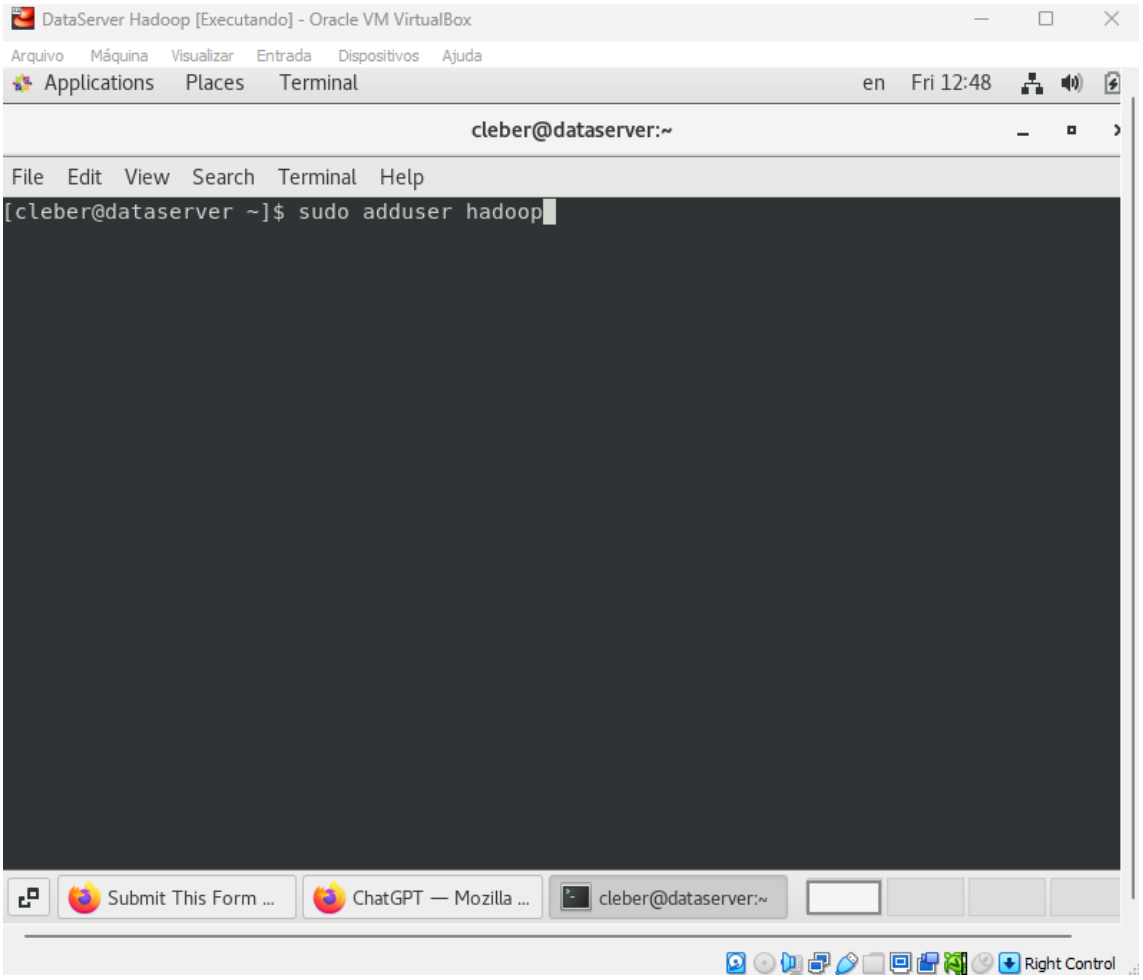
source .bashrc



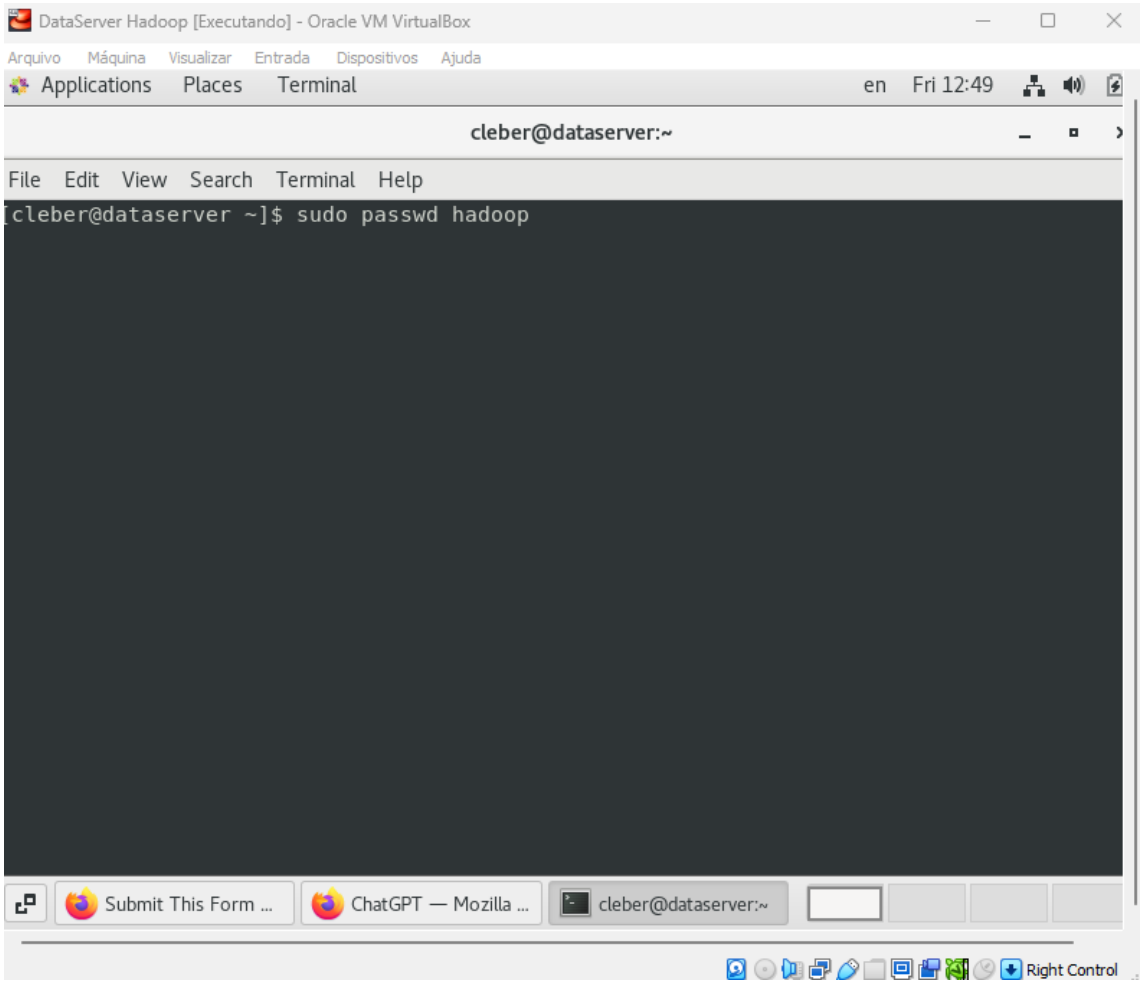
Verificando a versão do Java JDK

5 Instalação e Configuração do Hadoop

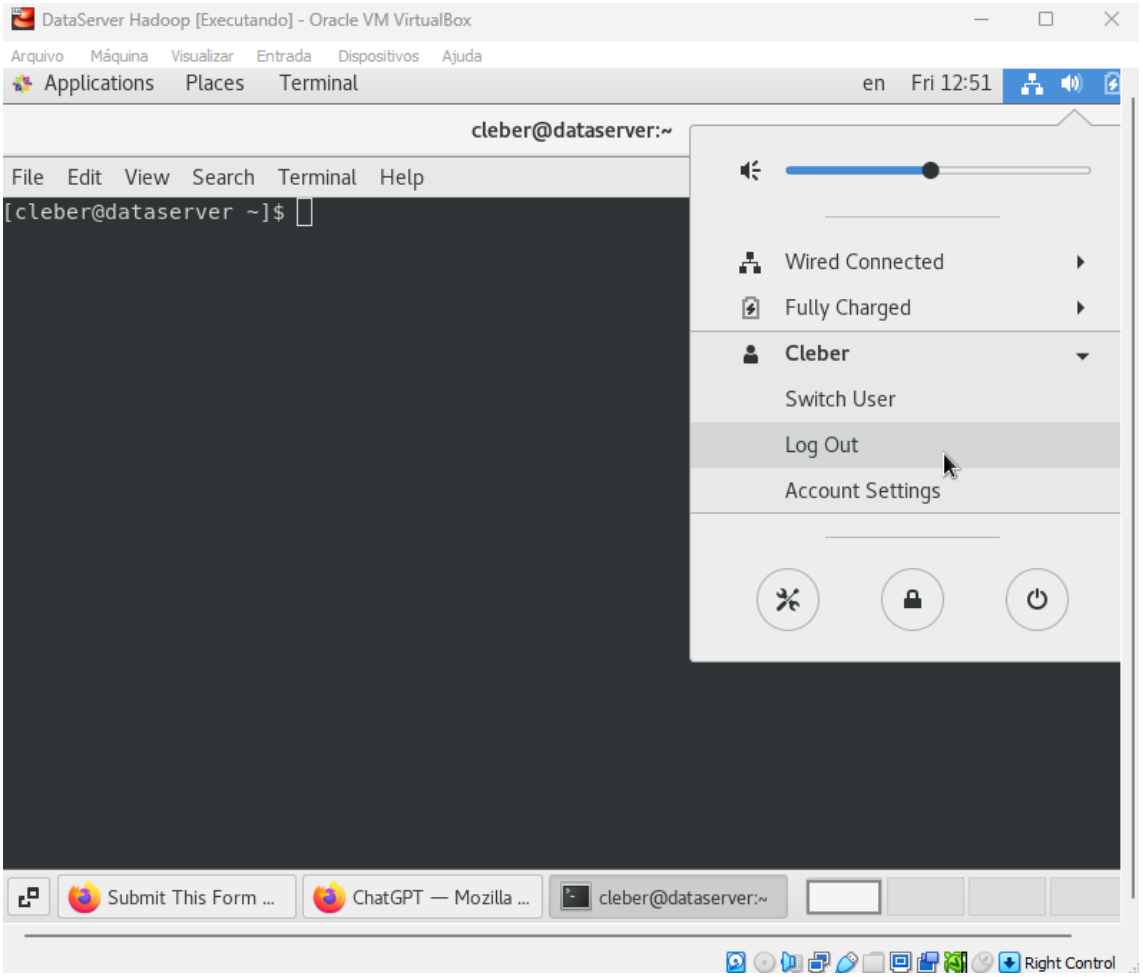
5.1 Criando o usuário hadoop



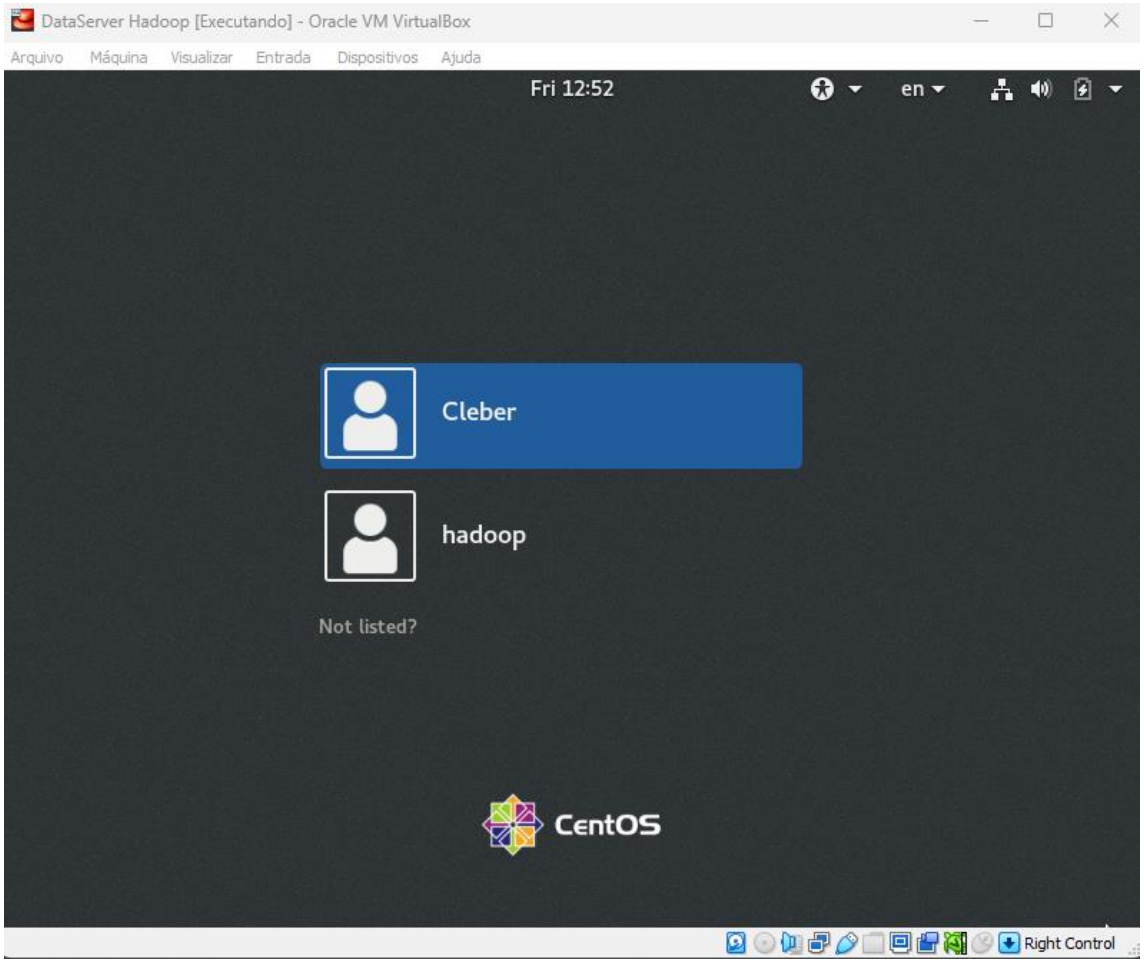
sudo adduser hadoop – para criar o usuário hadoop



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

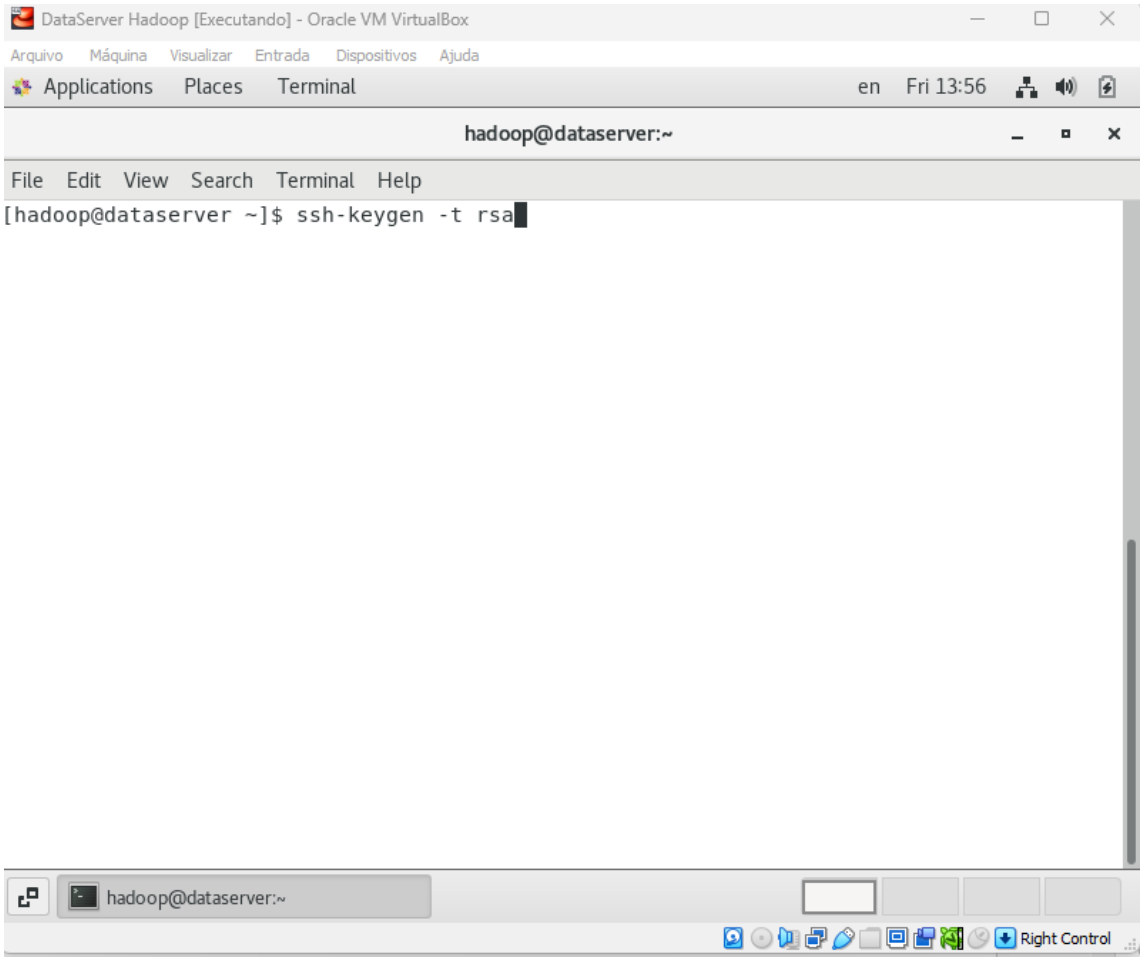


Efetue logout como usuário cleber

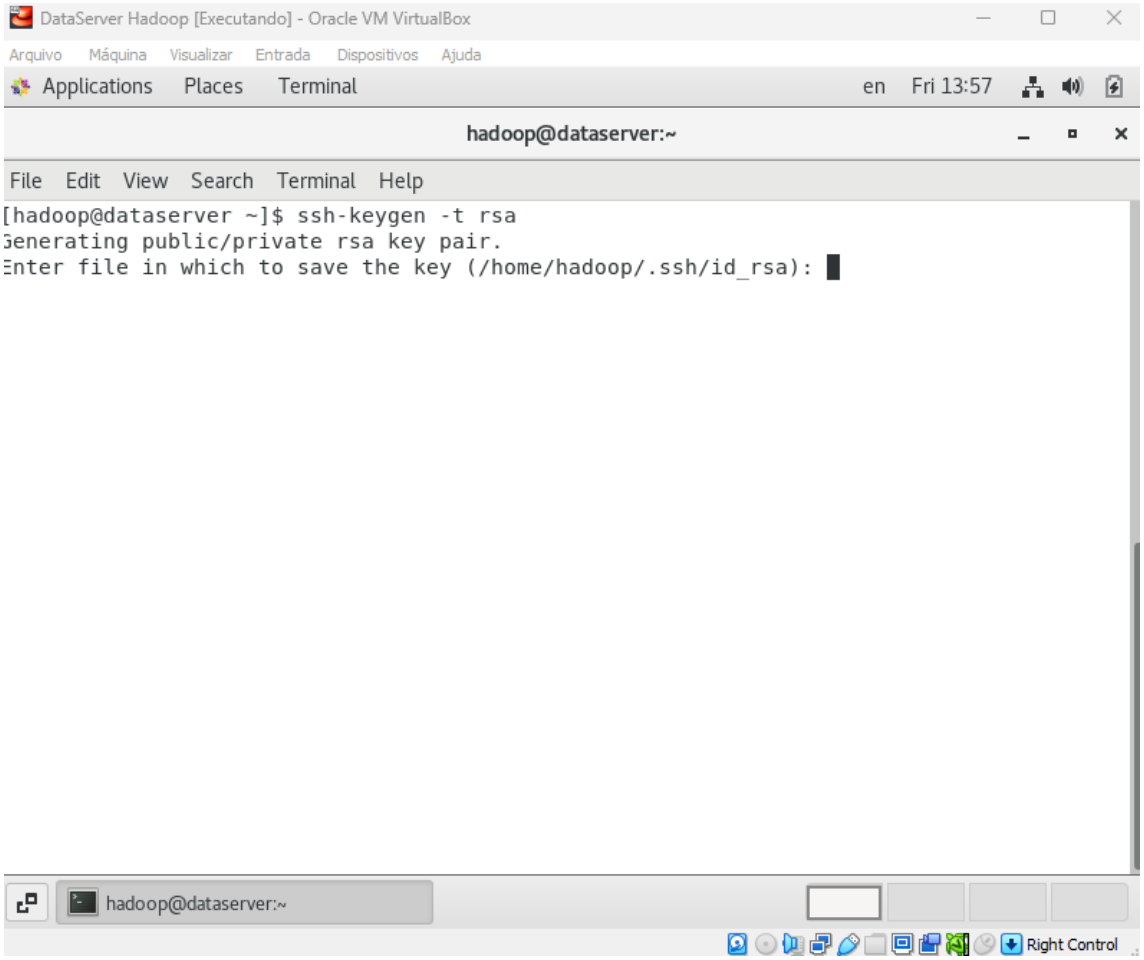


Efetue login como usuário hadoop
Adicione o usuário hadoop no arquivo /etc/sudoers conforme foi feito com o usuário cleber

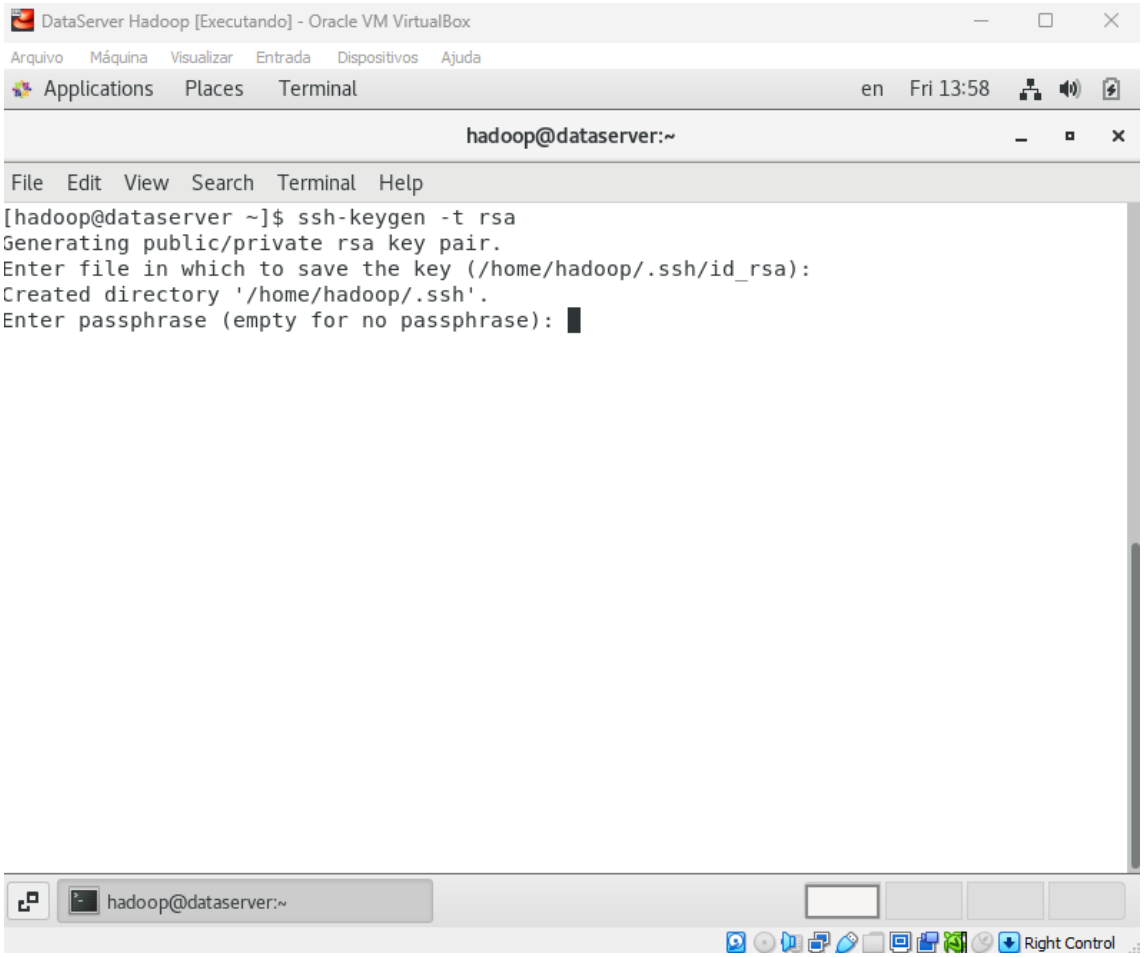
5.2 Configuração do ssh sem senha



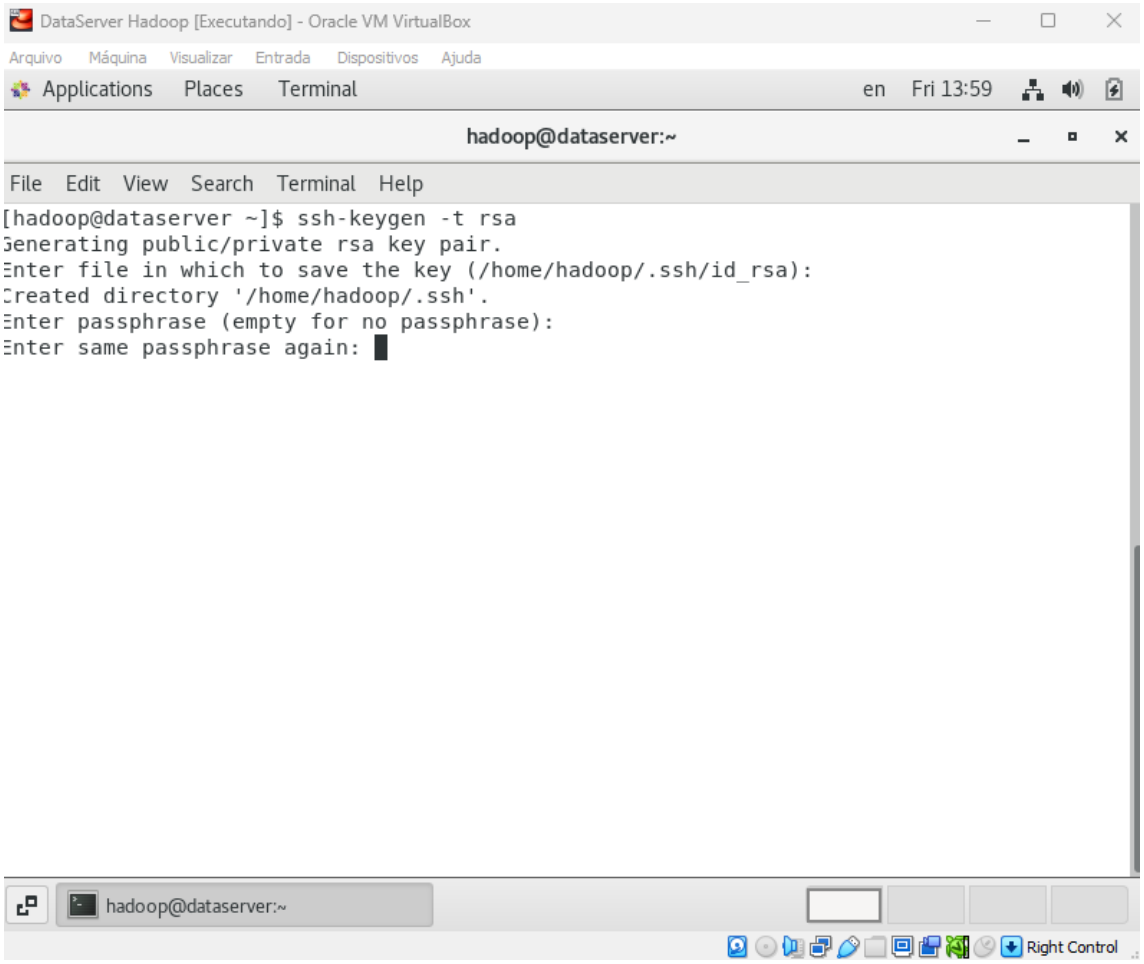
ssh-keygen –t rsa



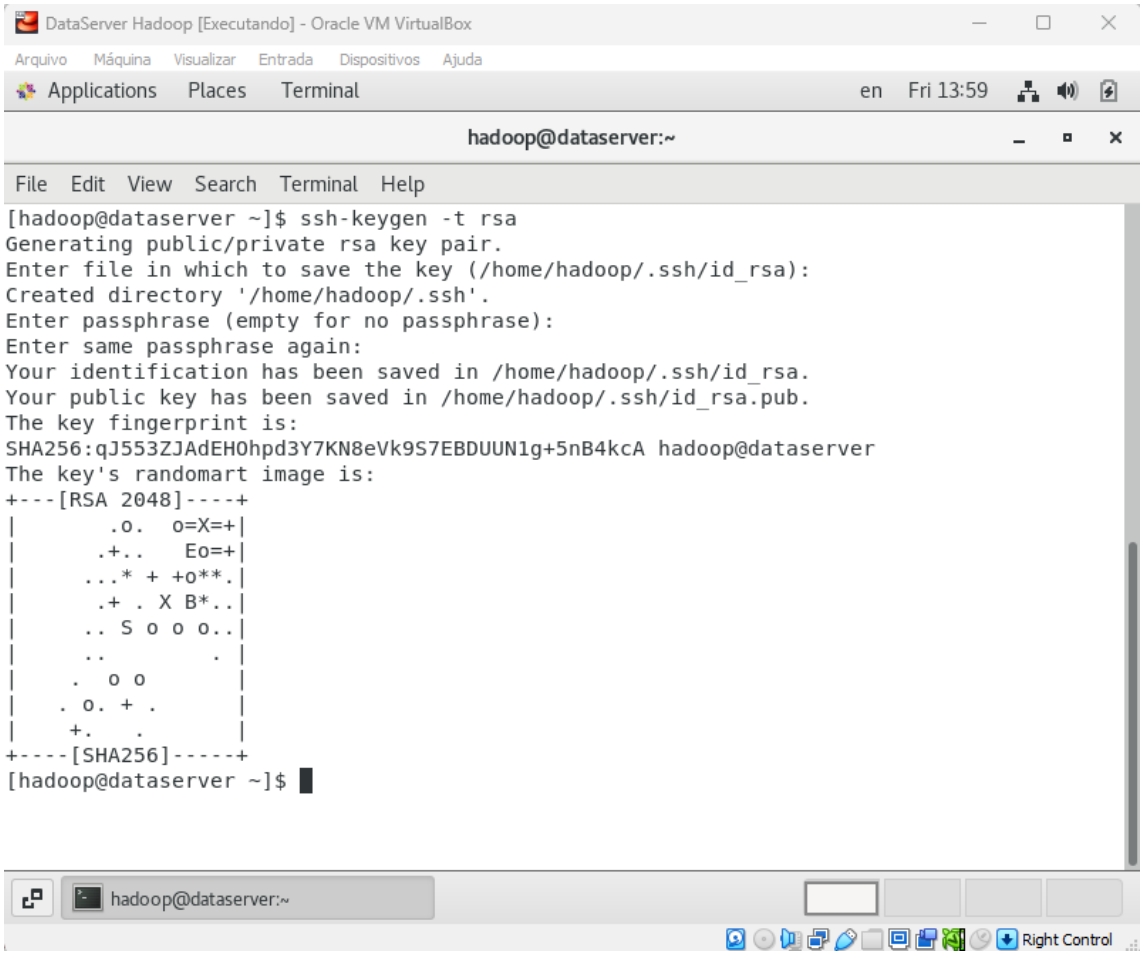
Pressionar Enter para confirmar o diretório onde as chaves serão geradas



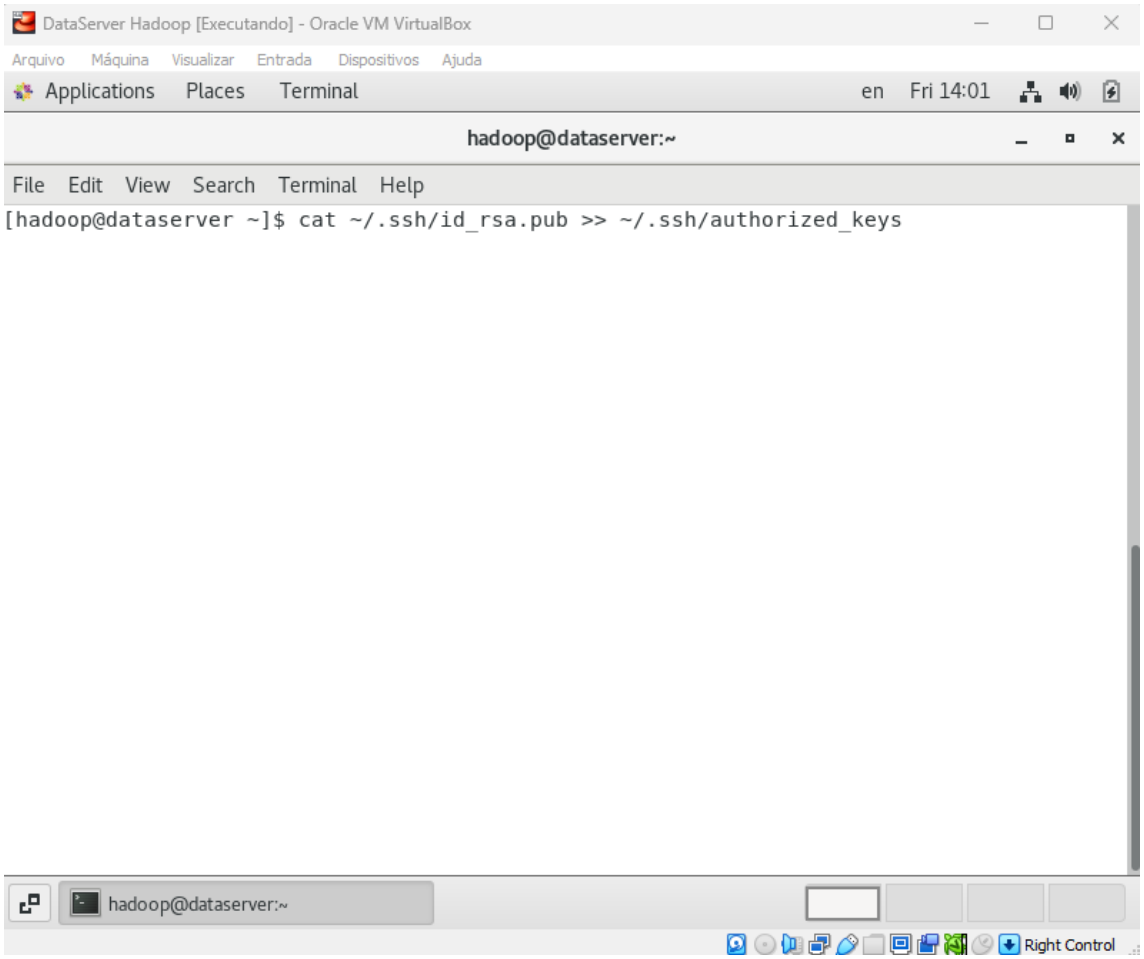
Pressionar Enter



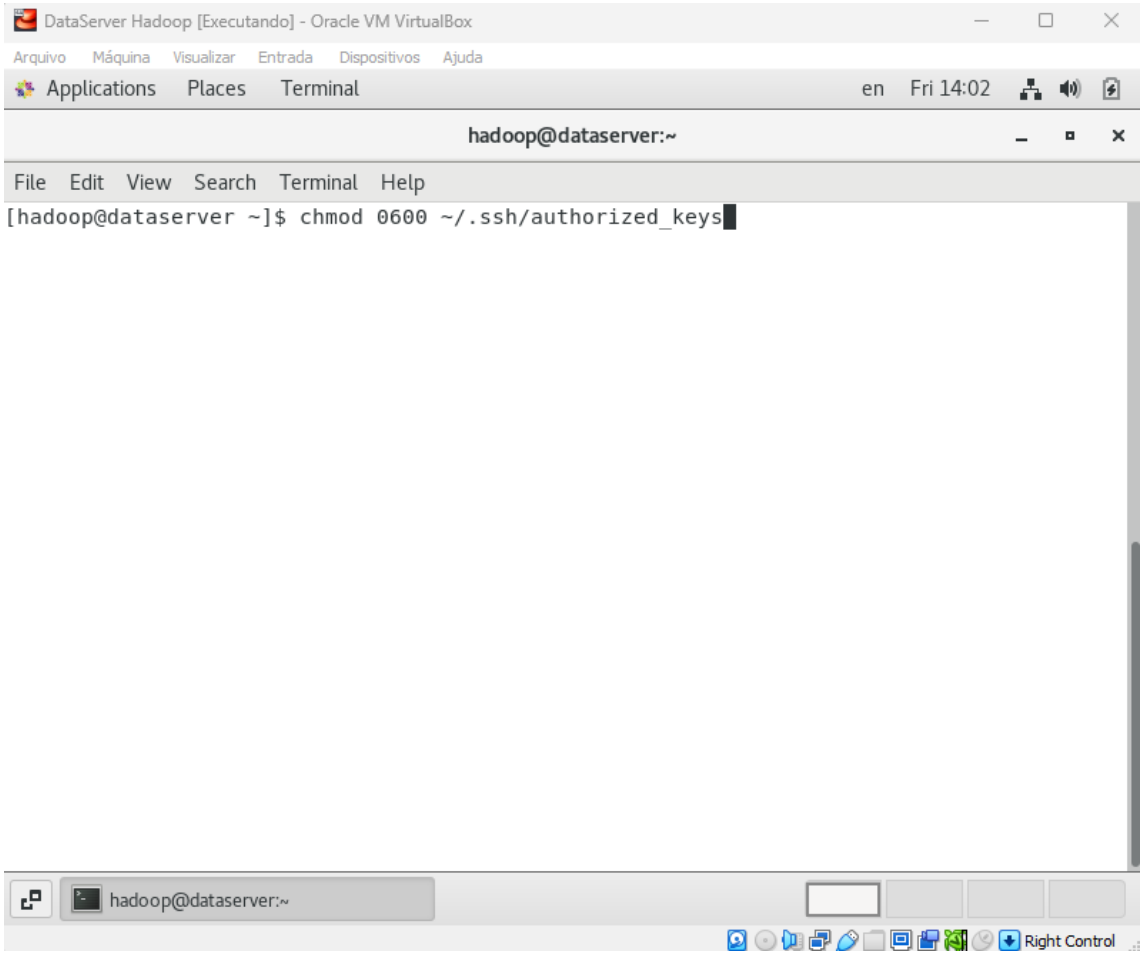
Pressionar Enter novamente



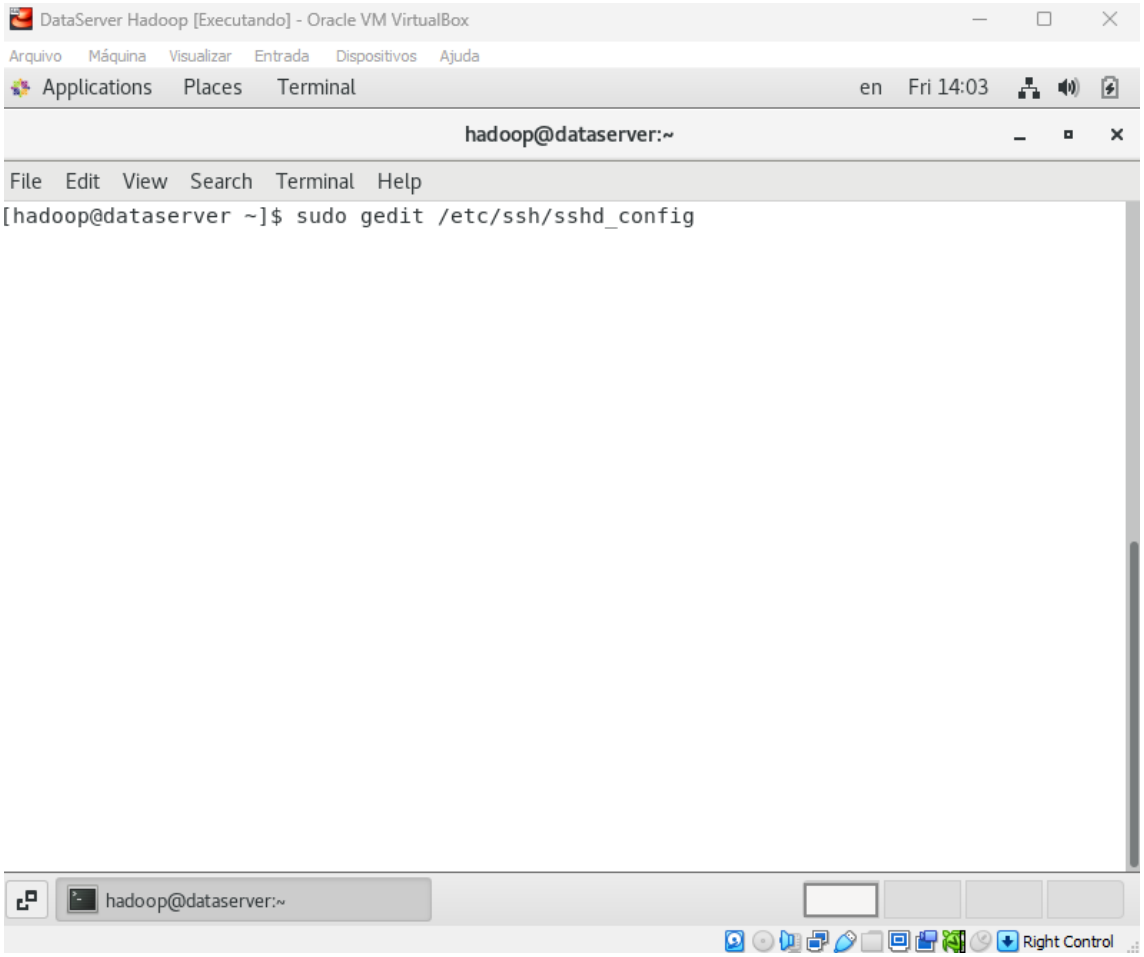
Chaves de segurança geradas



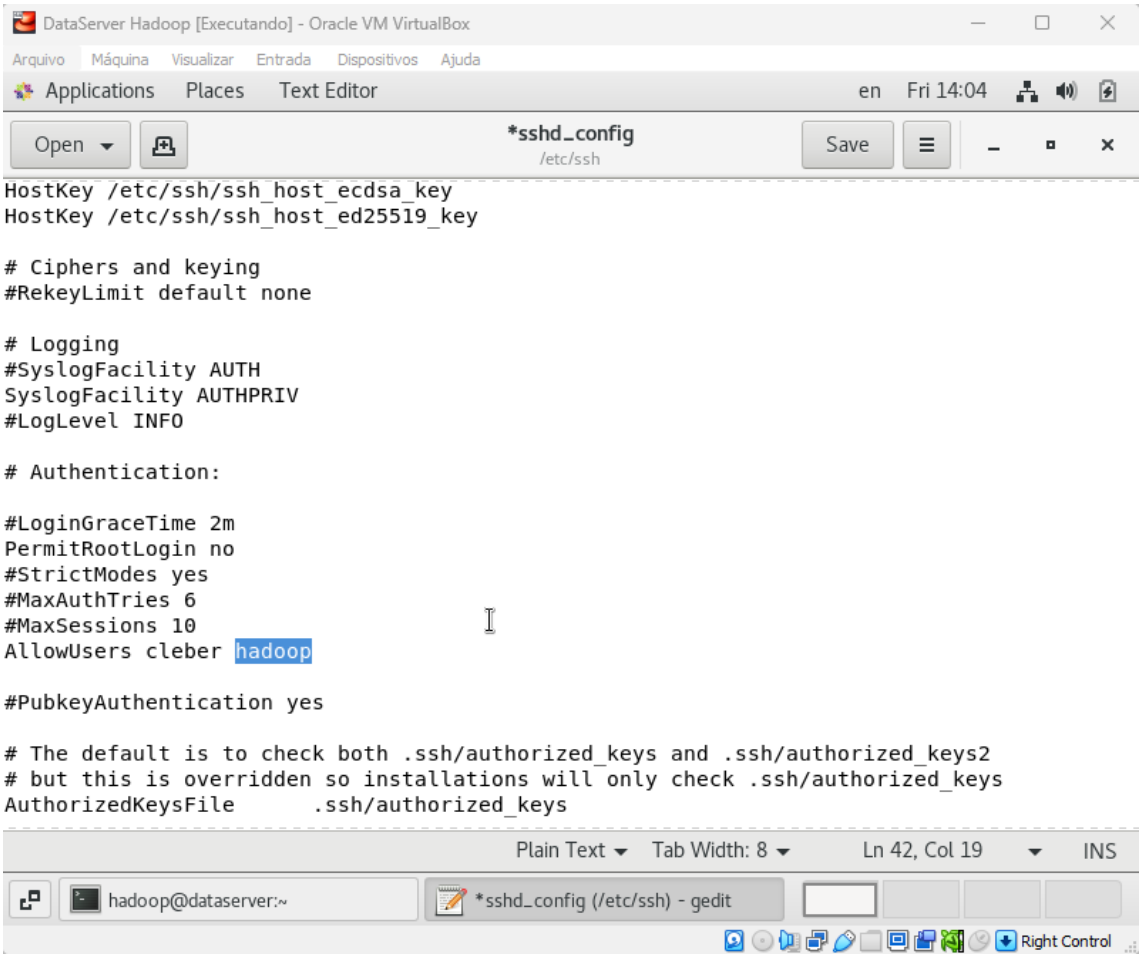
cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
Esse comando copia a chave pública para o arquivo authorized_keys do ssh



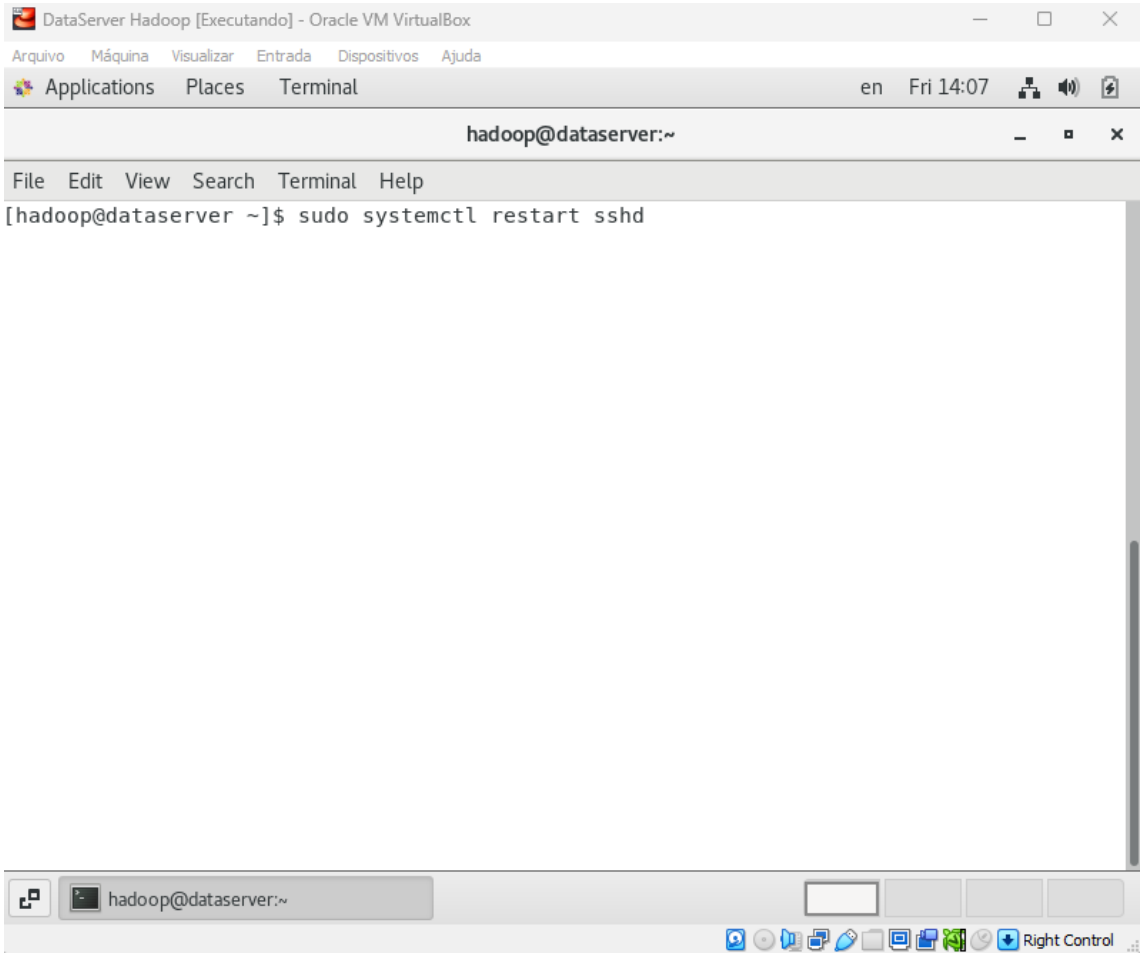
chmod 0600 ~/.ssh/authorized_keys
Esse comando define a permissão do arquivo authorized_keys



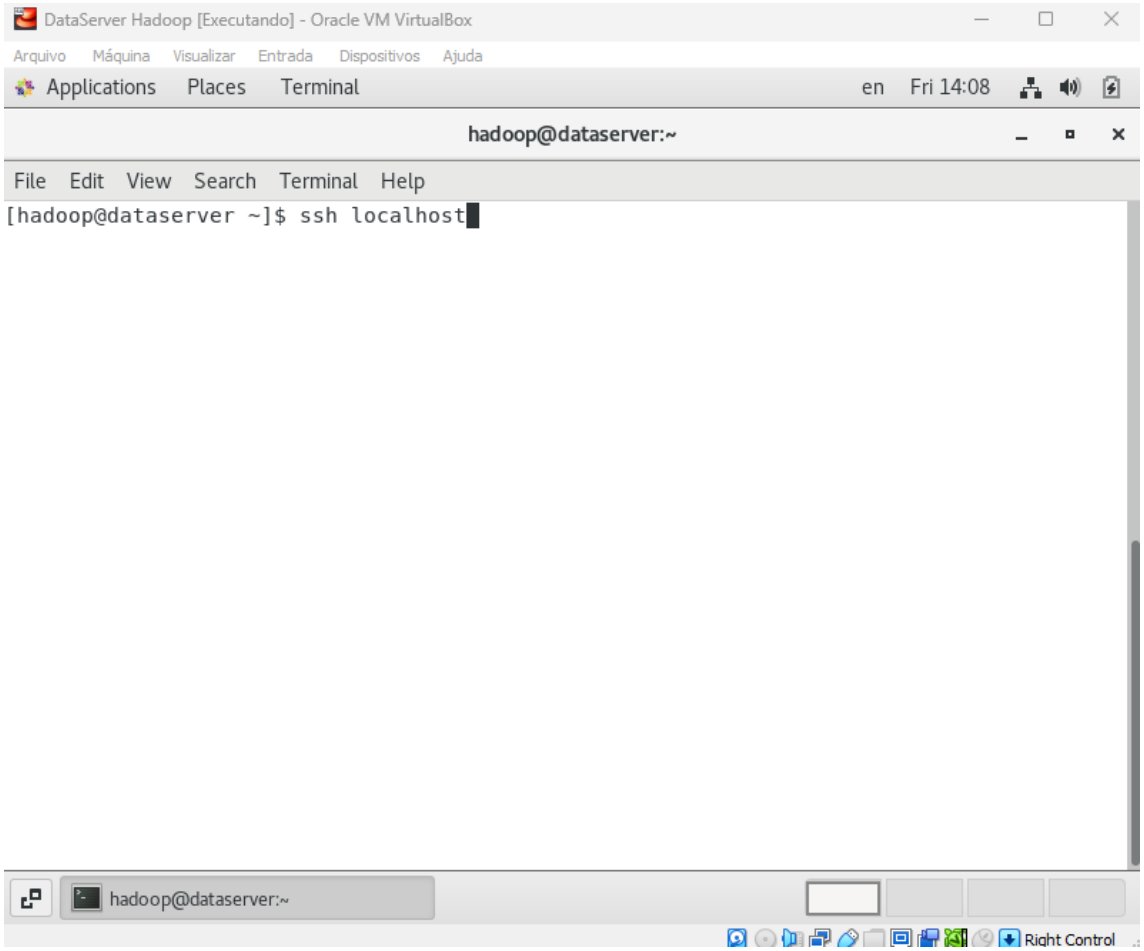
sudo gedit /etc/ssh/sshd_config
Edite o arquivo de configuração do ssh



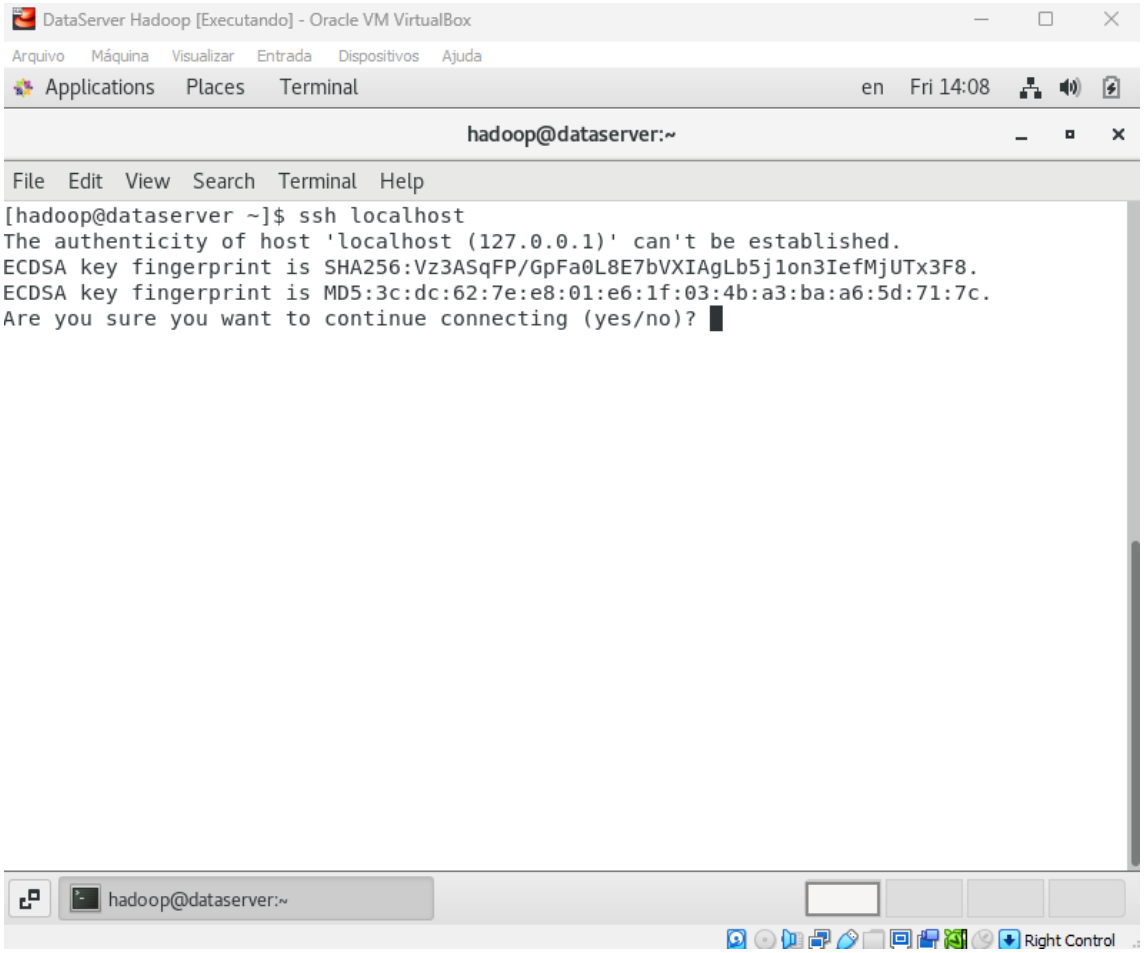
Inclua o usuário hadoop na linha AllowUsers, salve o arquivo e feche-o



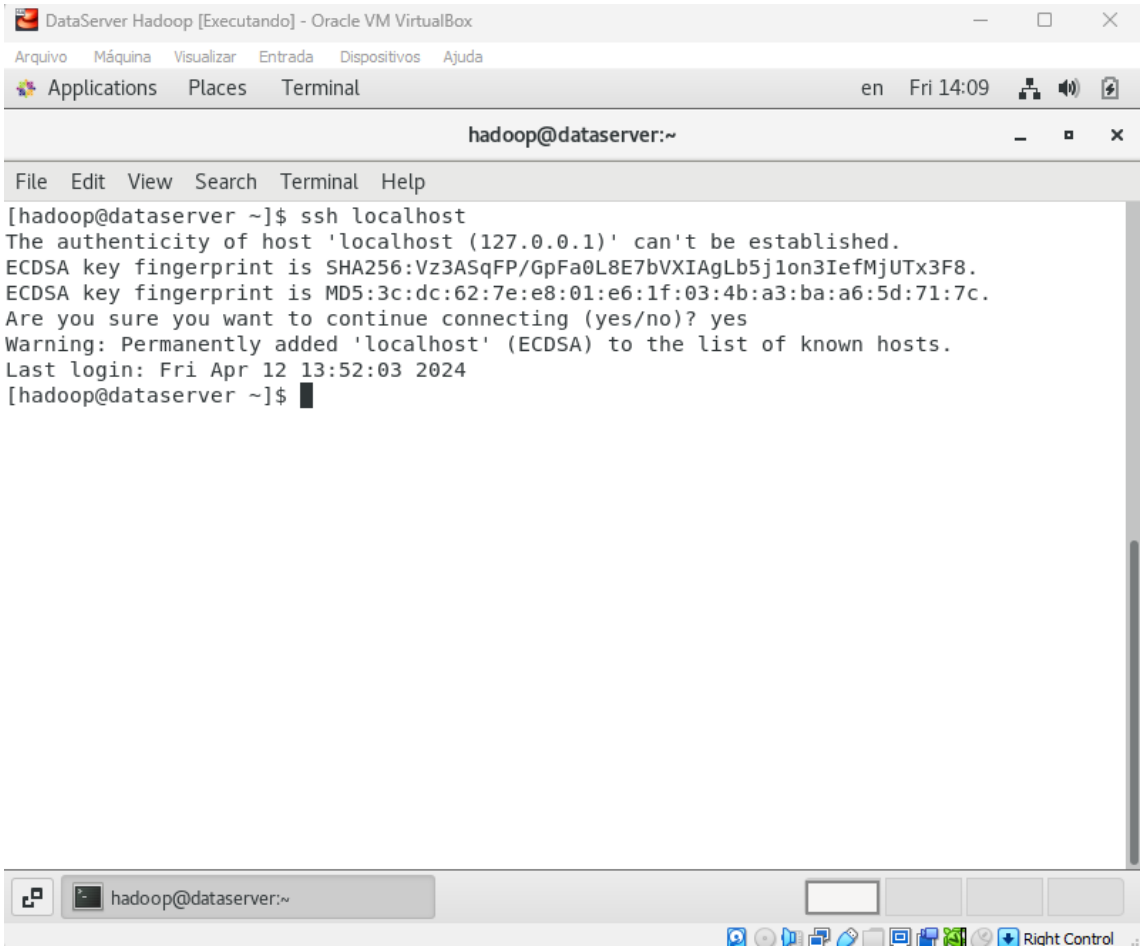
sudo systemctl restart sshd
Reinicie o serviço ssh



ssh localhost


DataServer Hadoop [Executando] - Oracle VM VirtualBox
Arquivo Máquina Visualizar Entrada Dispositivos Ajuda
Applications Places Terminal en Fri 14:08
hadoop@dataserver:~
File Edit View Search Terminal Help
[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)?
hadoop@dataserver:~
Right Control

Yes

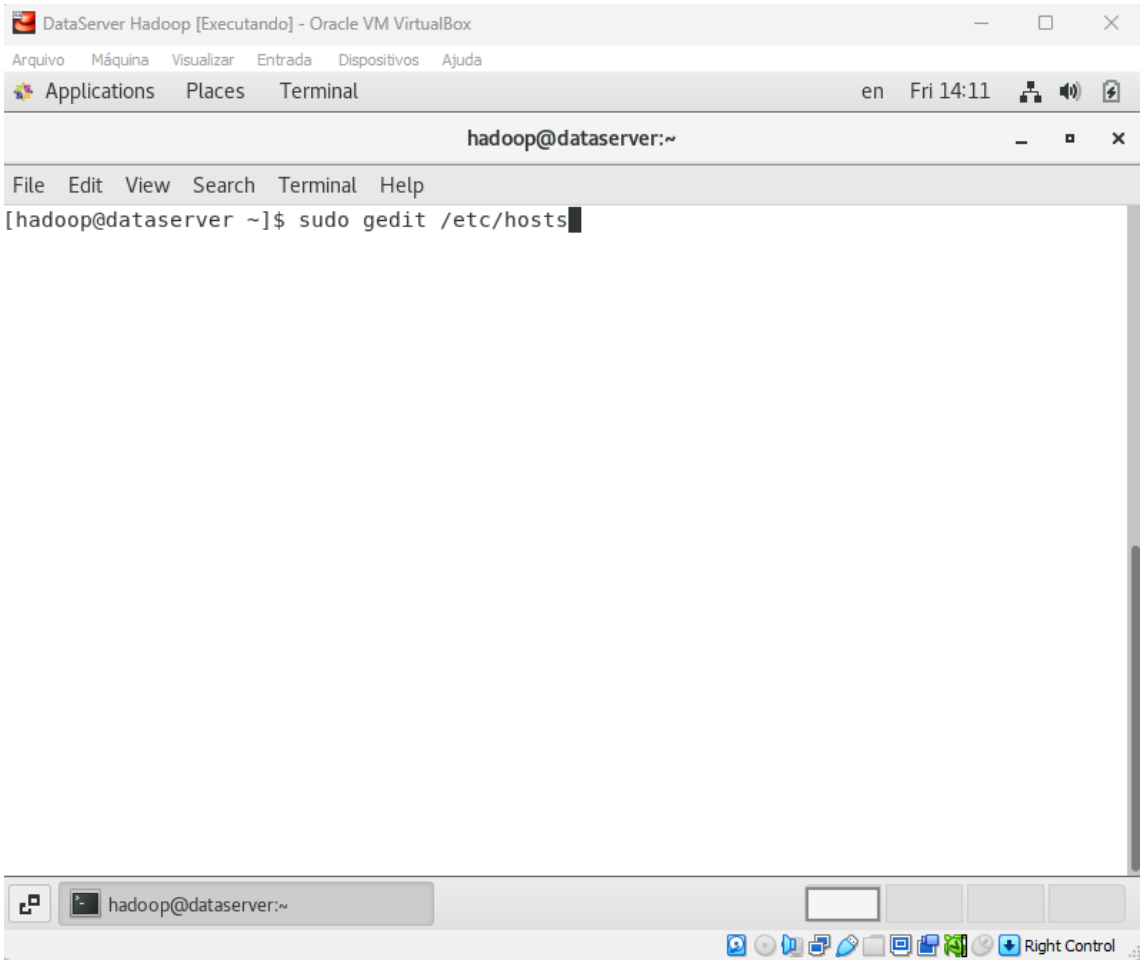


Conexão ssh sem senha efetuada com sucesso. Digite exit e pressione Enter.

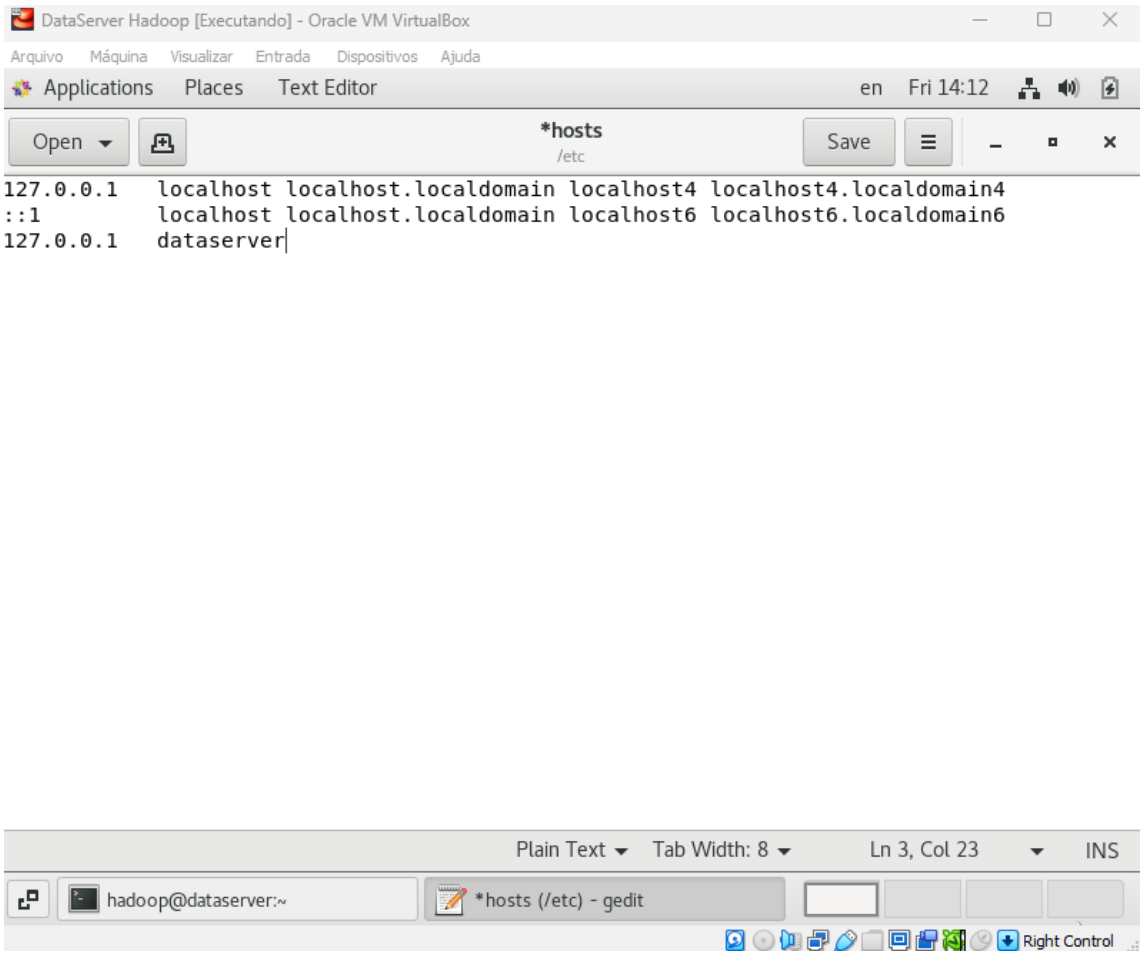
O ambiente está pronto para receber o Hadoop!!

5.3 Download e Instalação do Hadoop

5.3.1 Editando o arquivo hosts

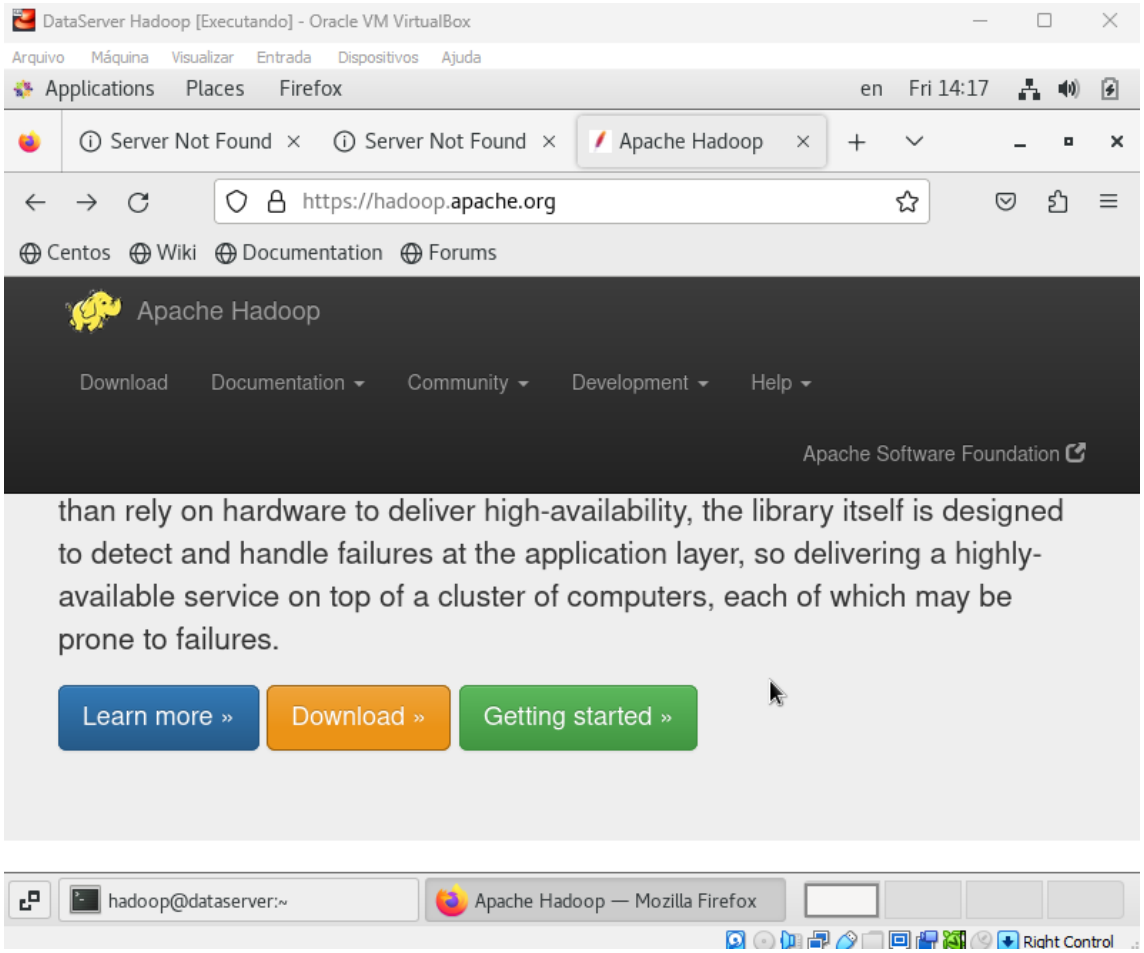


Editar o arquivo hosts

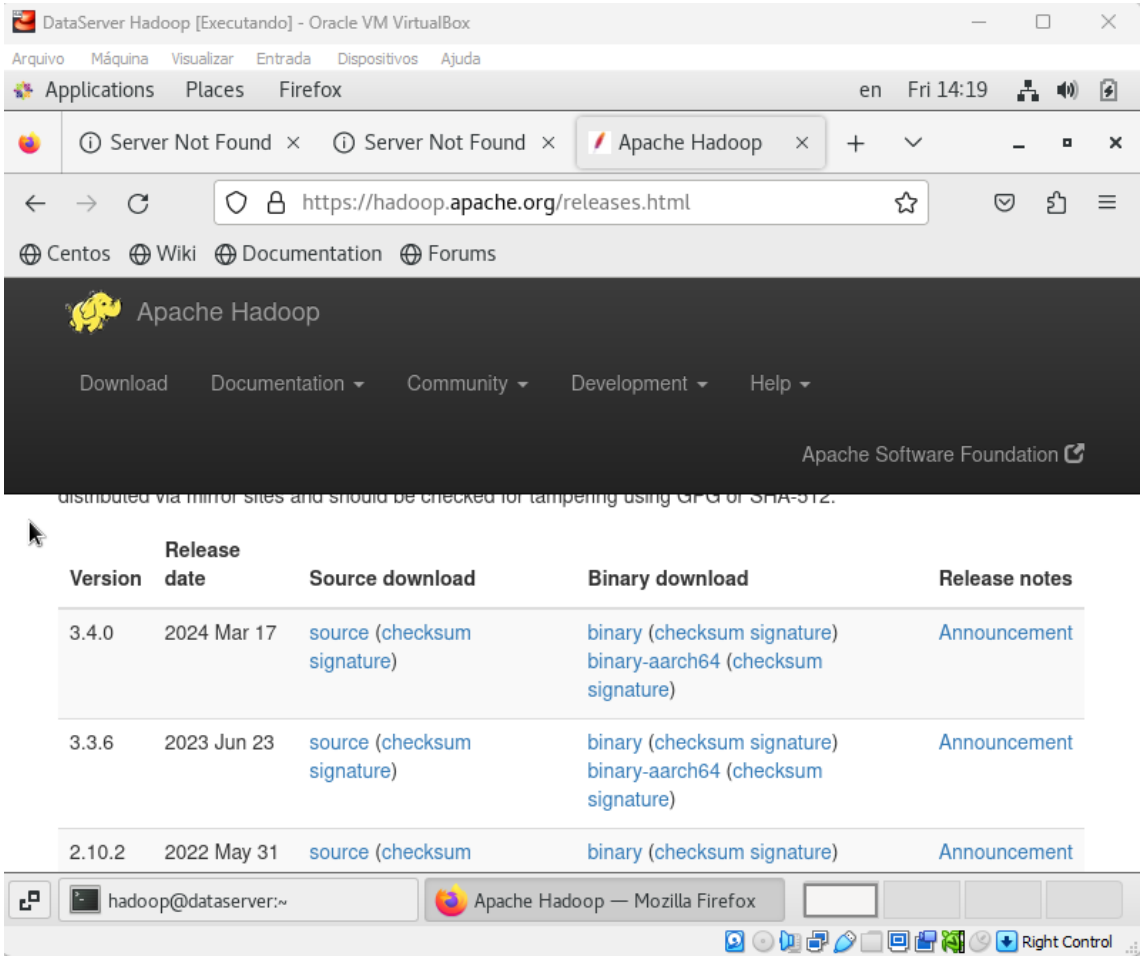


Incluir a linha 127.0.0.1 dataserver conforme acima

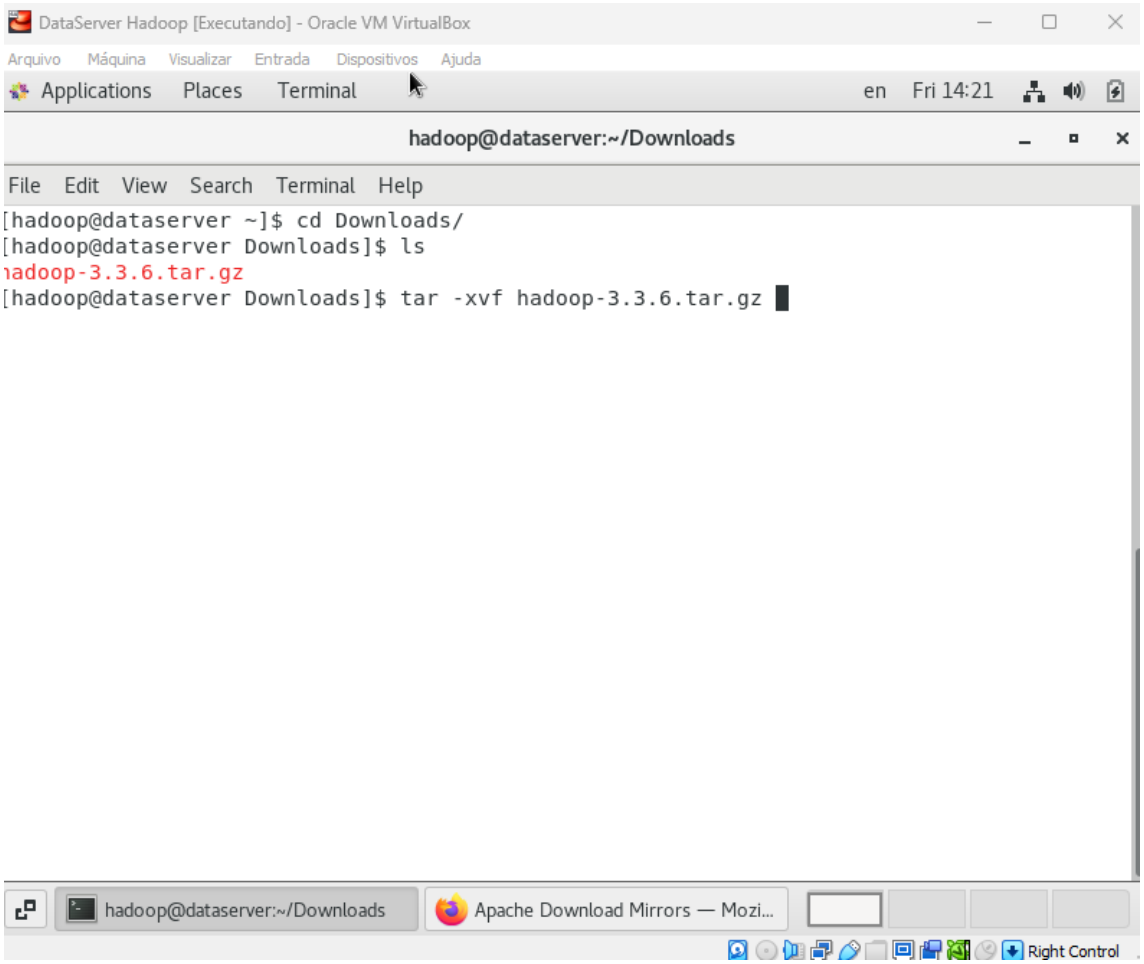
5.3.2 Download do Hadoop



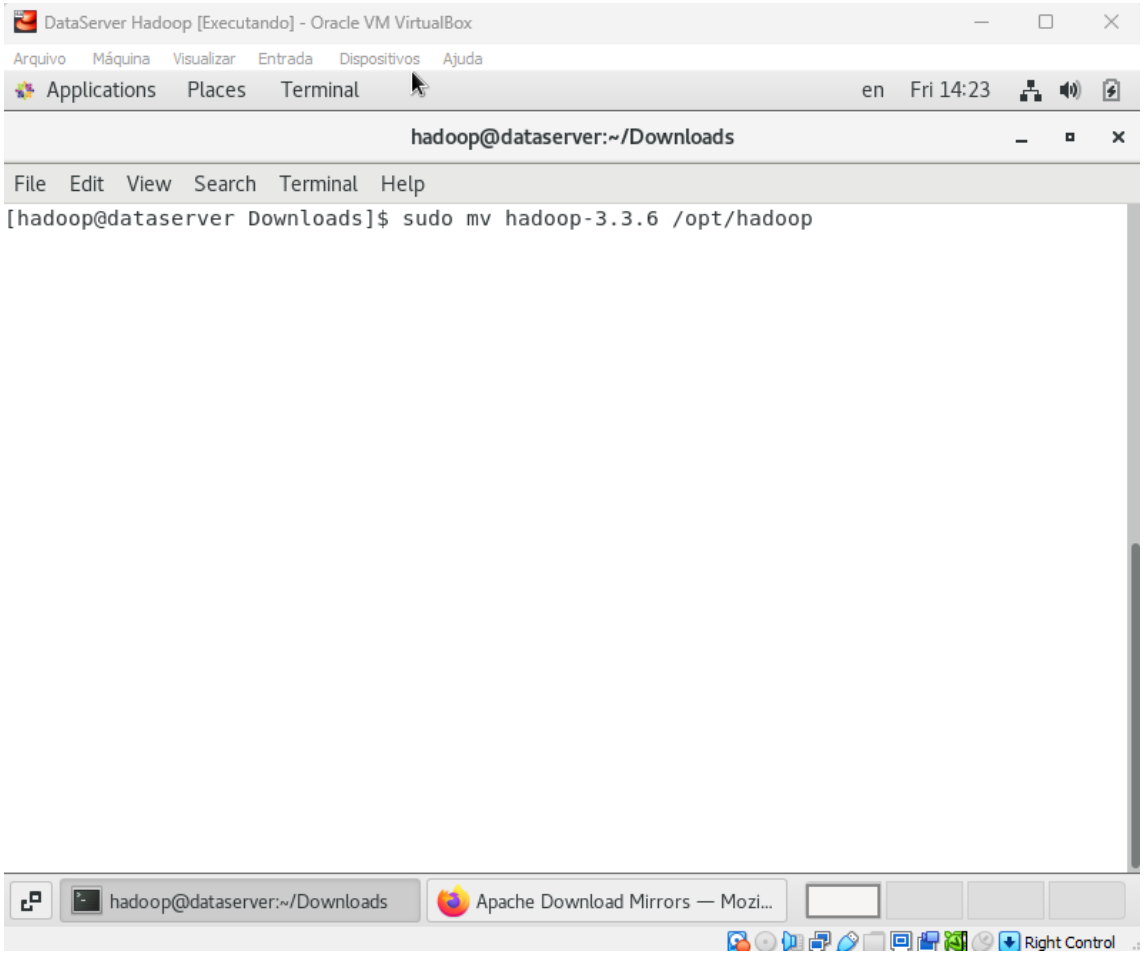
Acesse a página do Hadoop e clique em Download



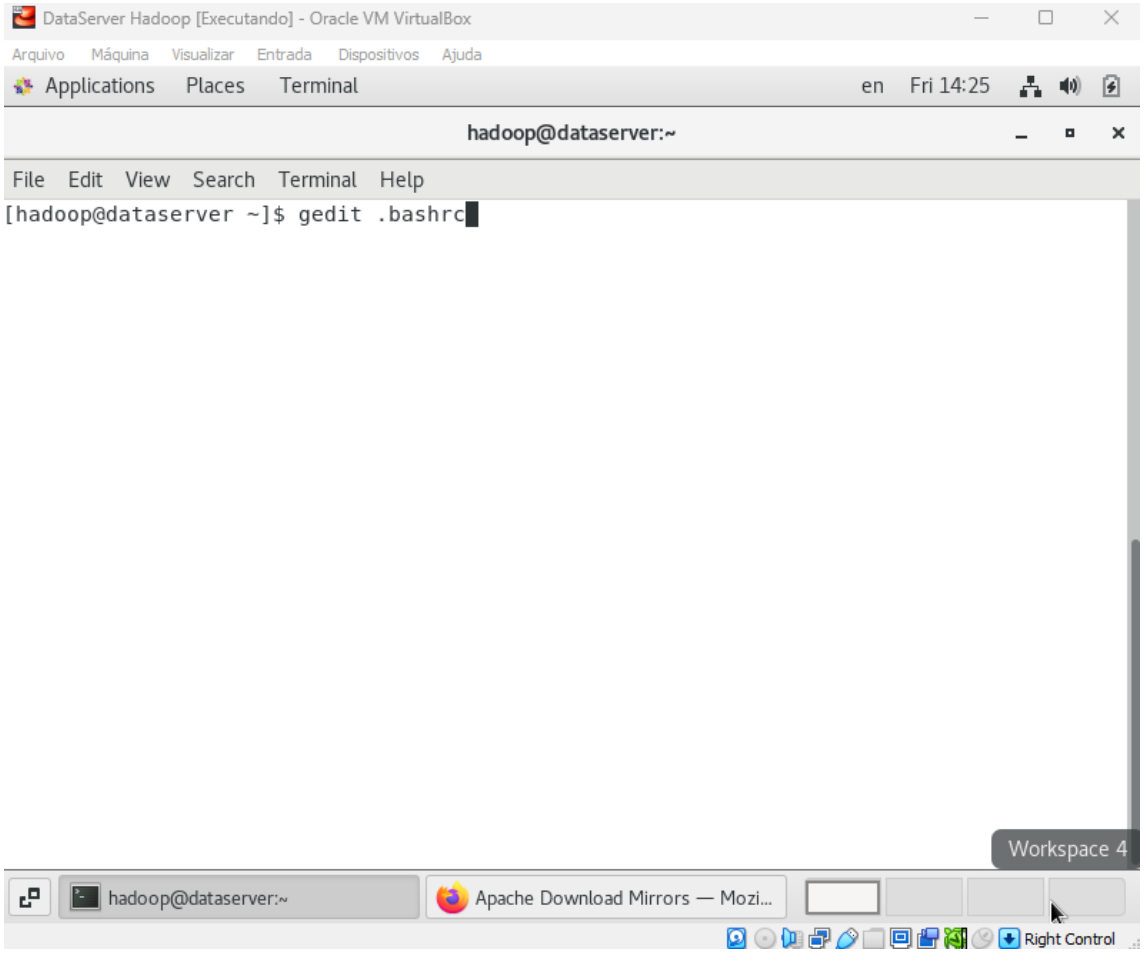
Faça o download da versão 3.3, opção binary.
O arquivo será baixado no diretório /home/hadoop/Downloads



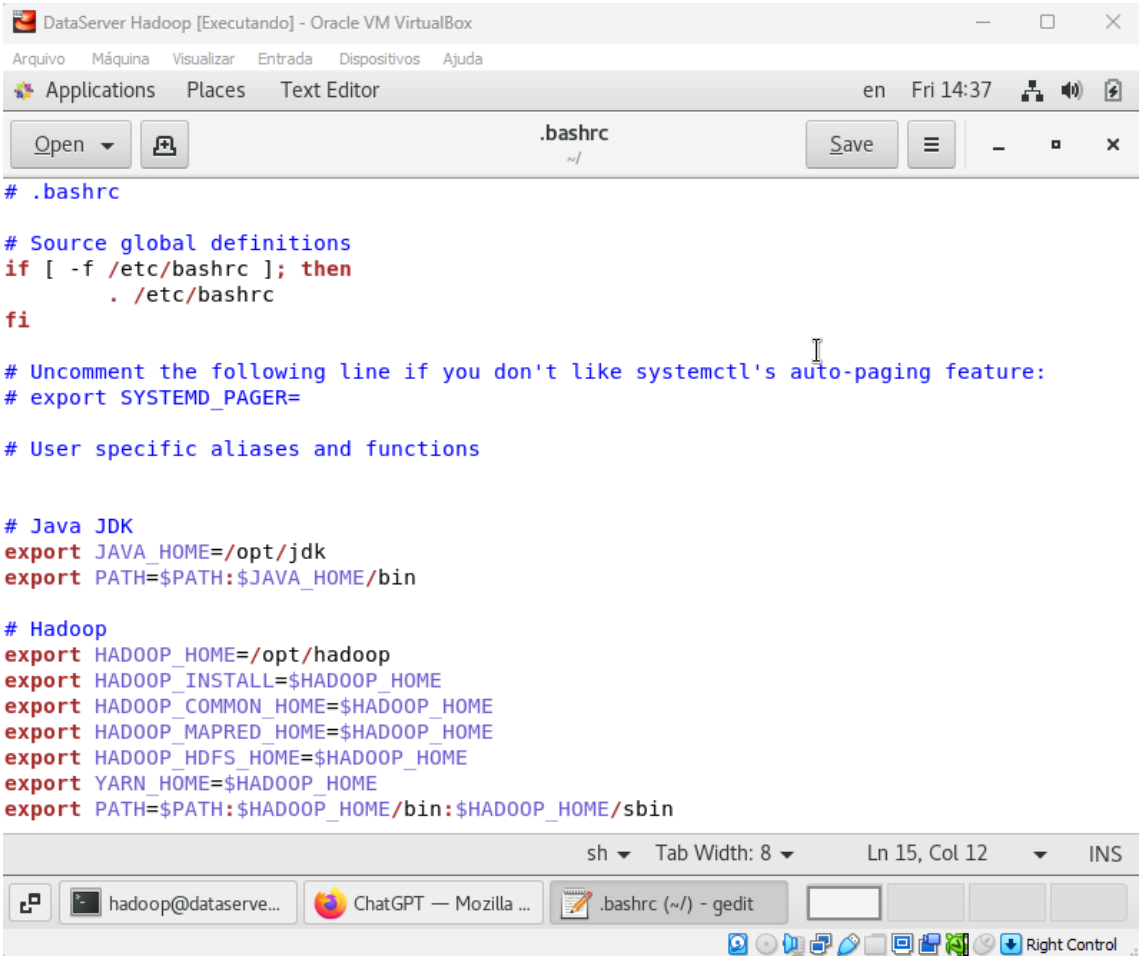
Descompacte o arquivo



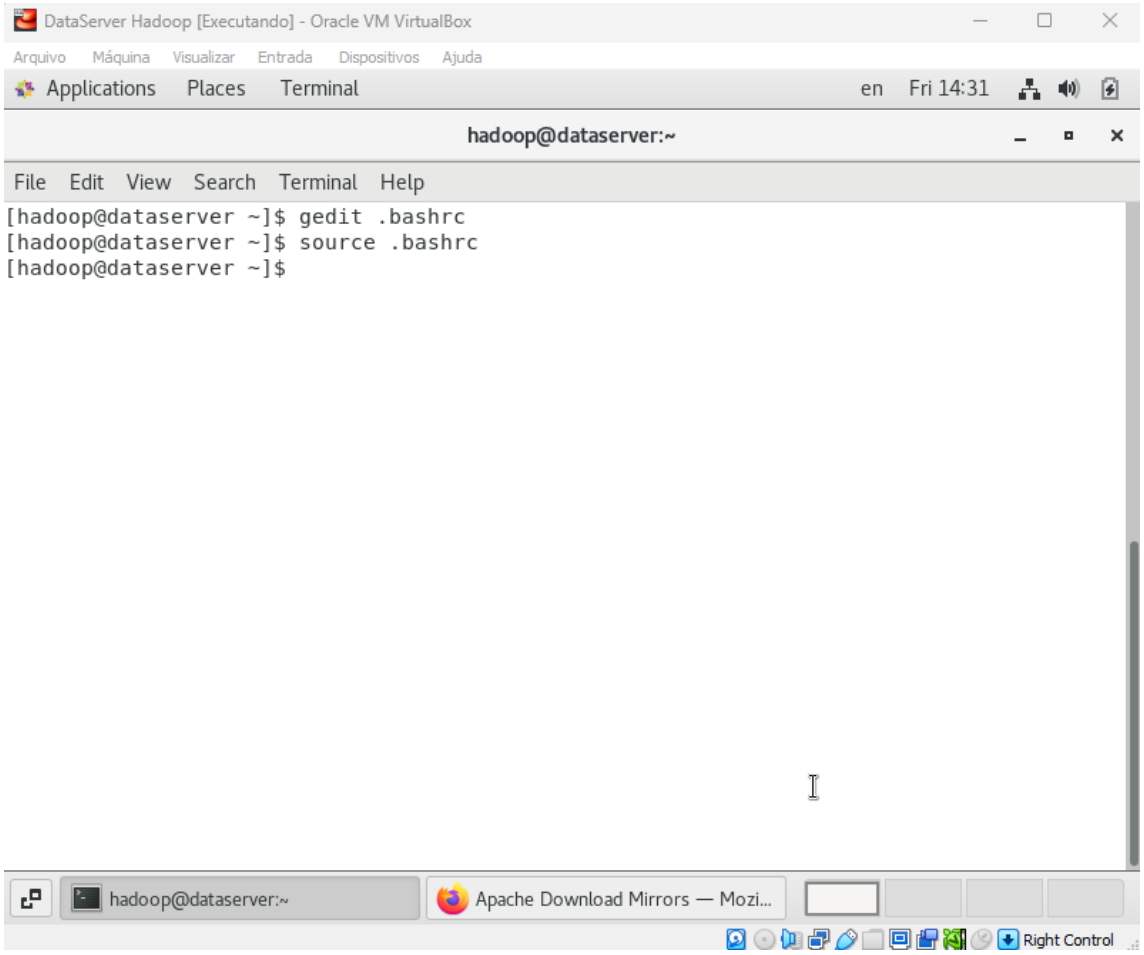
Mover o diretório para /opt/hadoop



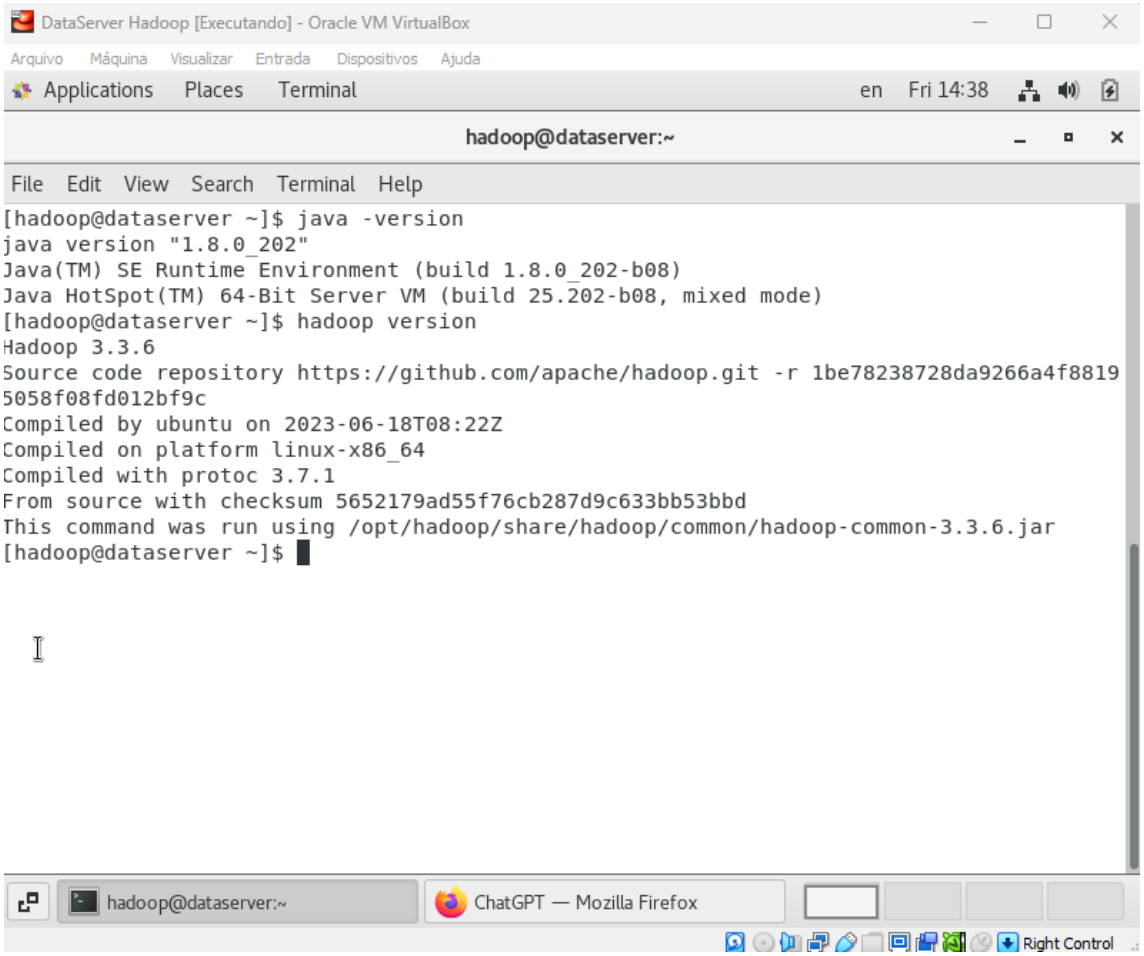
Abrir o arquivo de profile do usuário hadoop



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



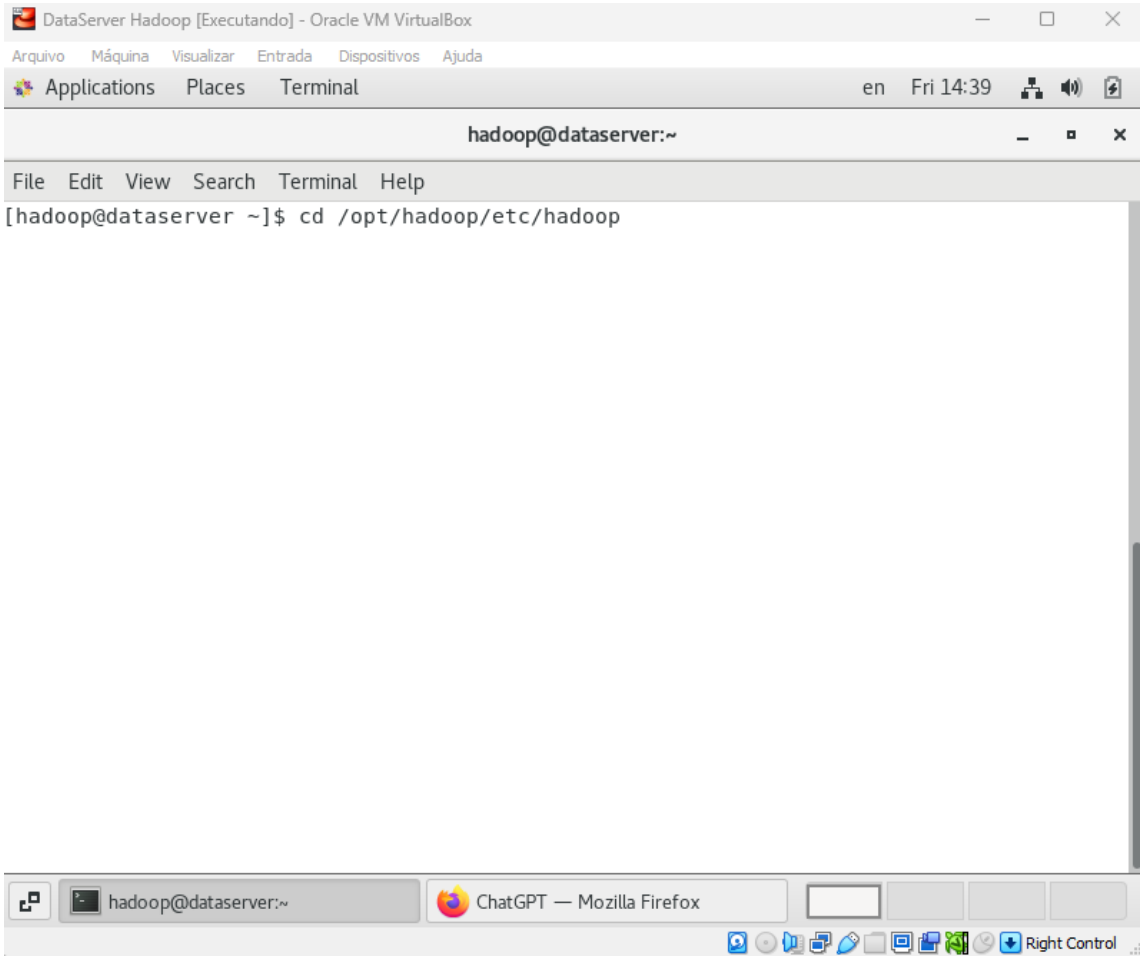
Execute source .bashrc para efetivar as mudanças das variáveis no SO



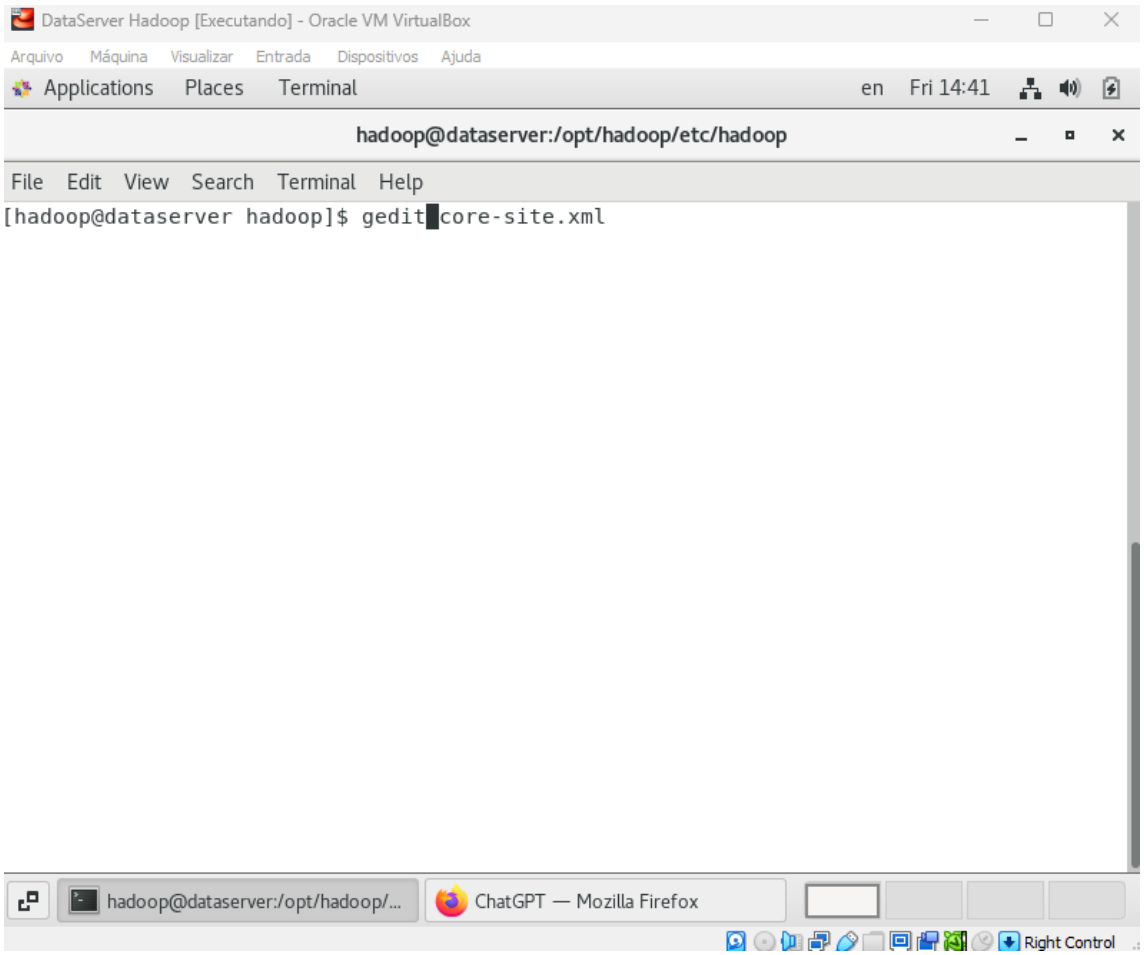
Java e Hadoop instalados e configurados com sucesso!!!

5.4 Configuração do Hadoop

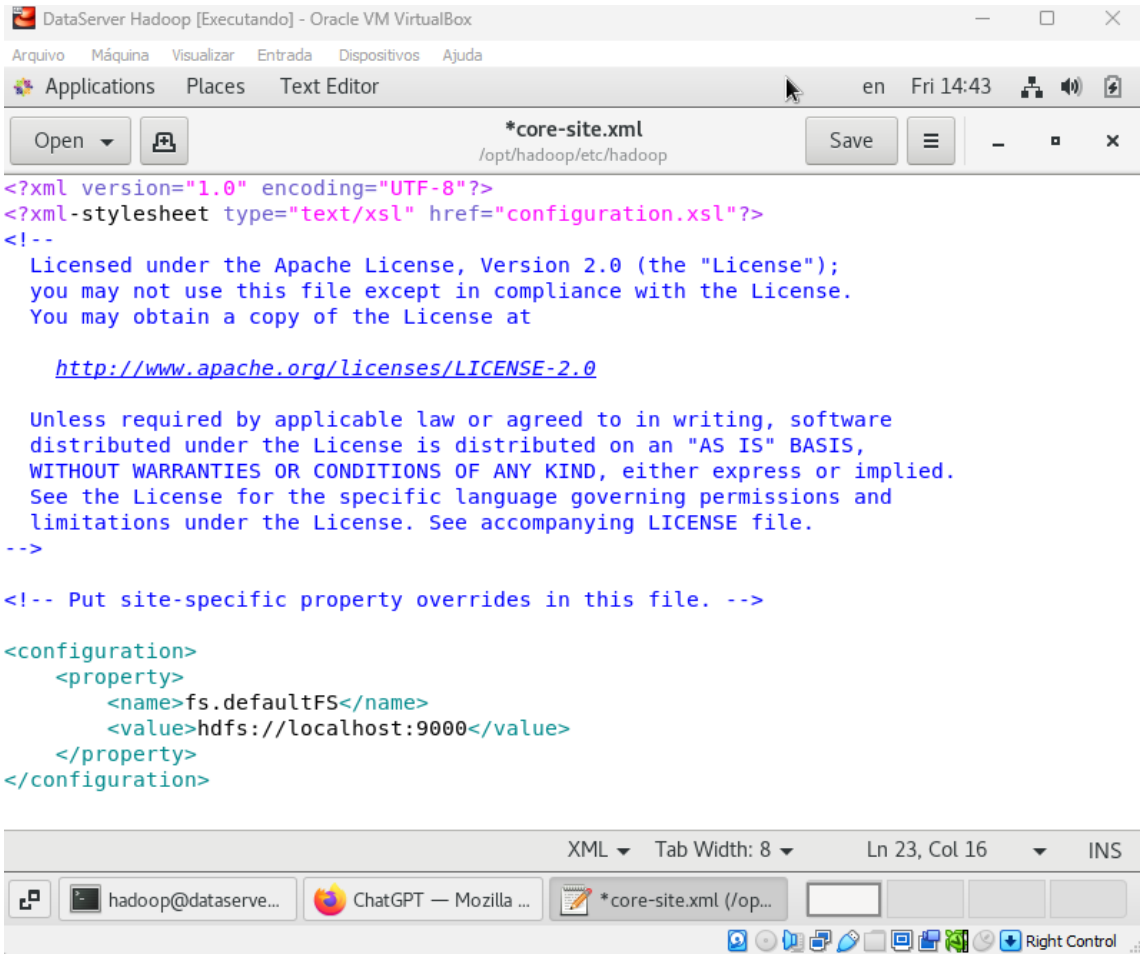
5.4.1 Editar arquivos de configuração do Hadoop



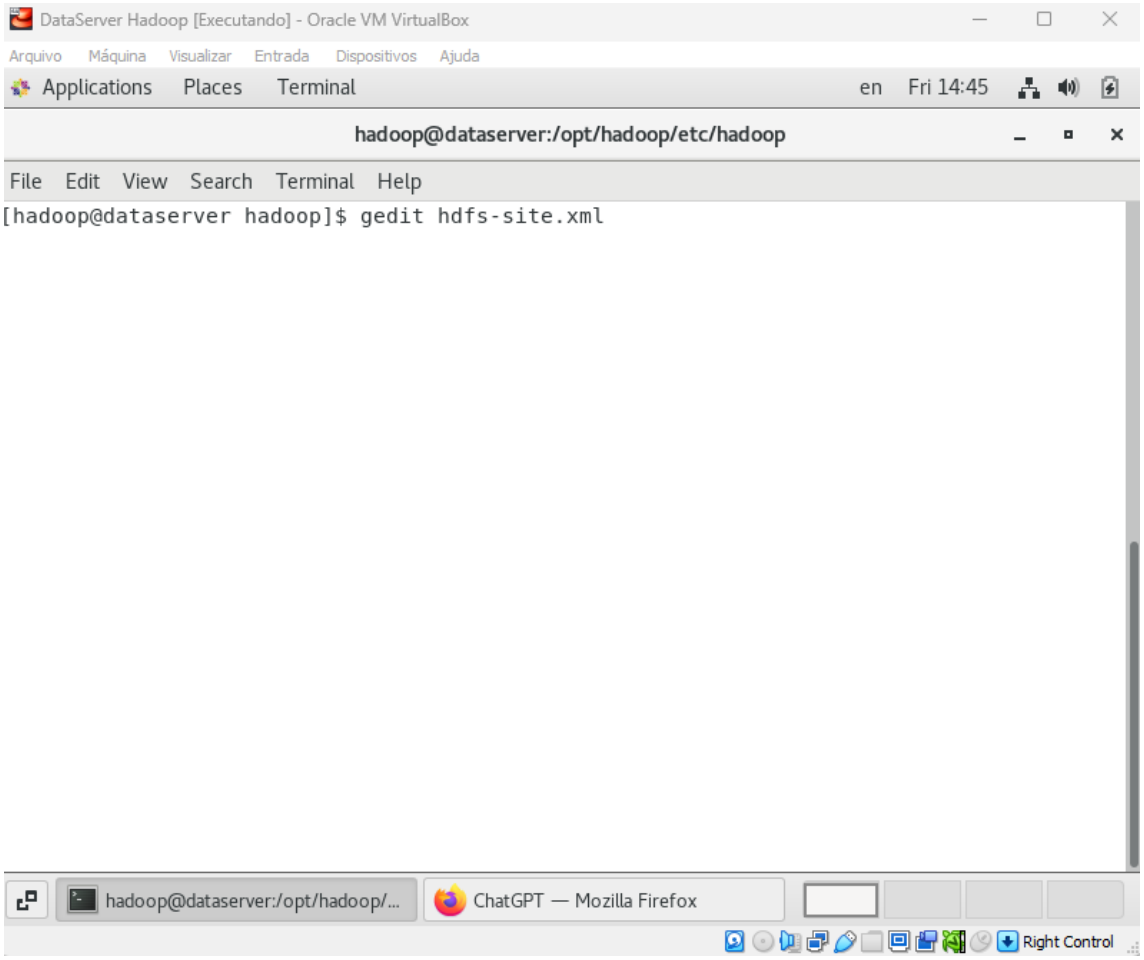
Os arquivos de configuração do Hadoop estão em
$HADOOP_HOME/etc/hadoop
Nesse caso: /opt/hadoop/etc/hadoop



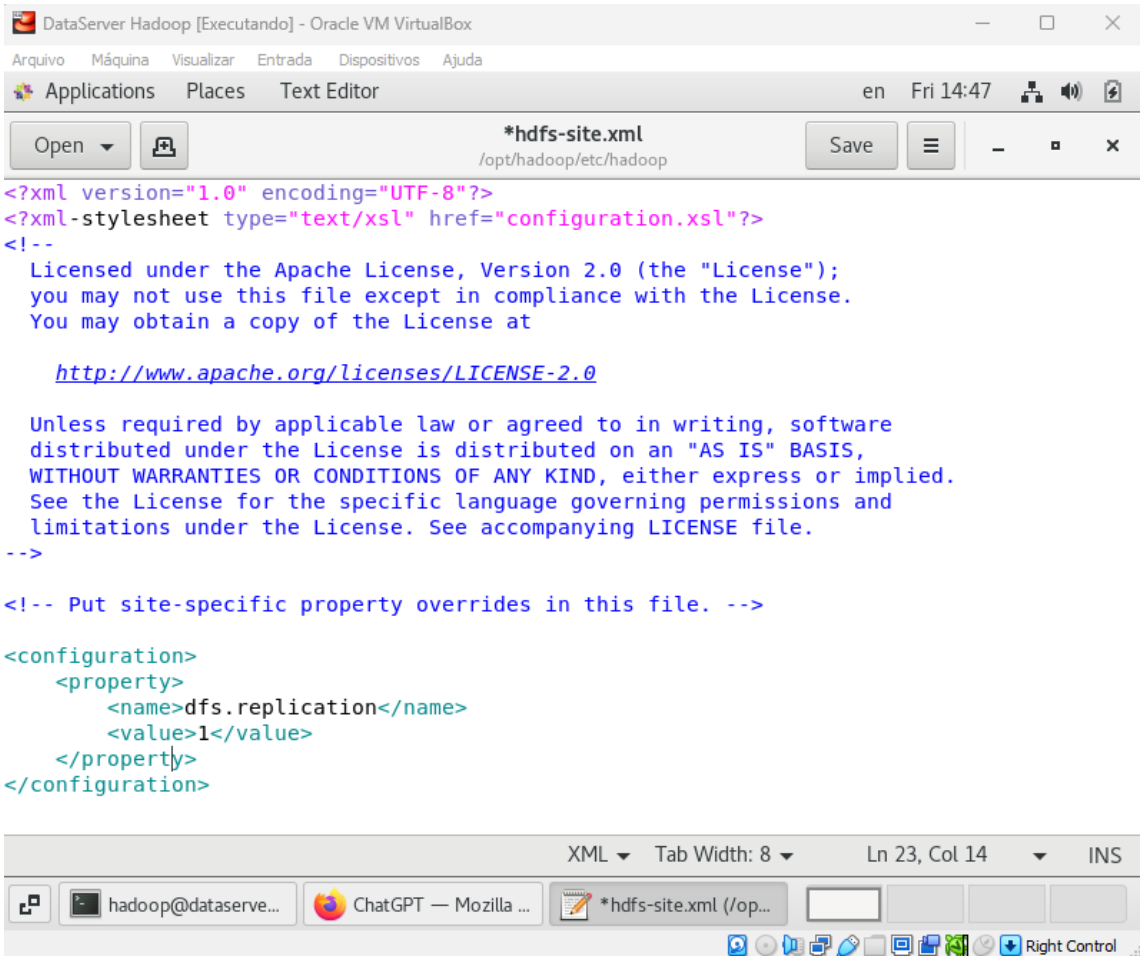
Editar o arquivo core-site.xml



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

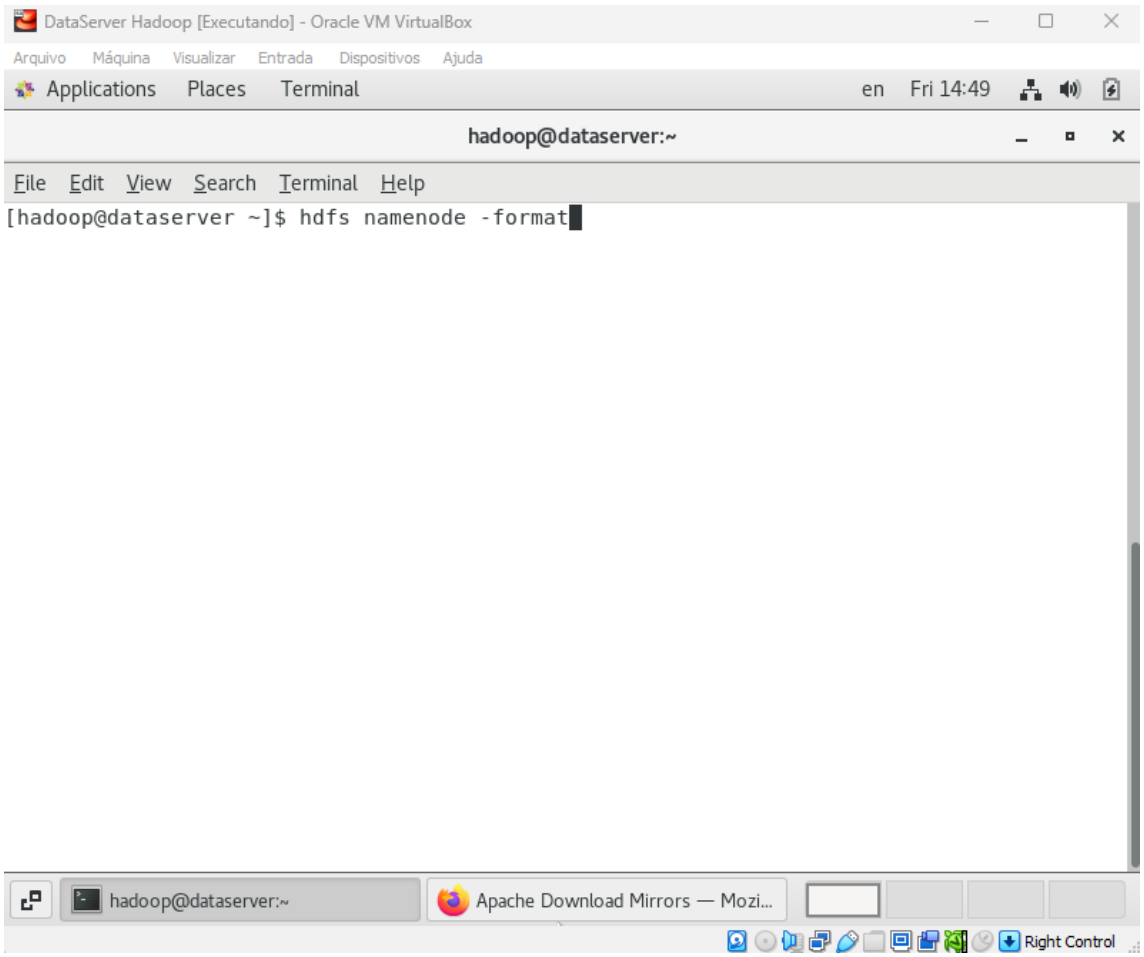


Editar o arquivo hdfs-site.xml

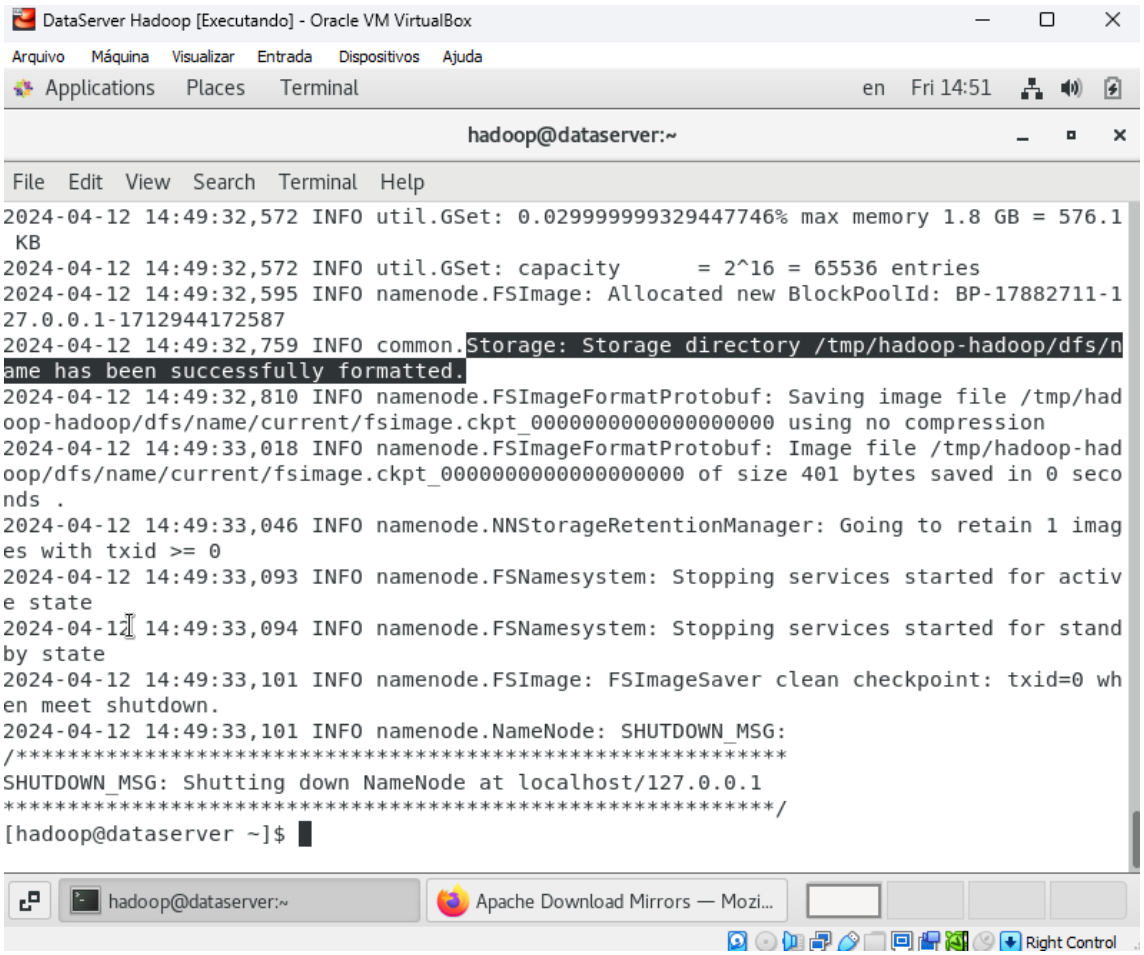


Acrescentar as propriedades conforme acima e salvar o arquivo.

5.4.2 Formatando o Namenode

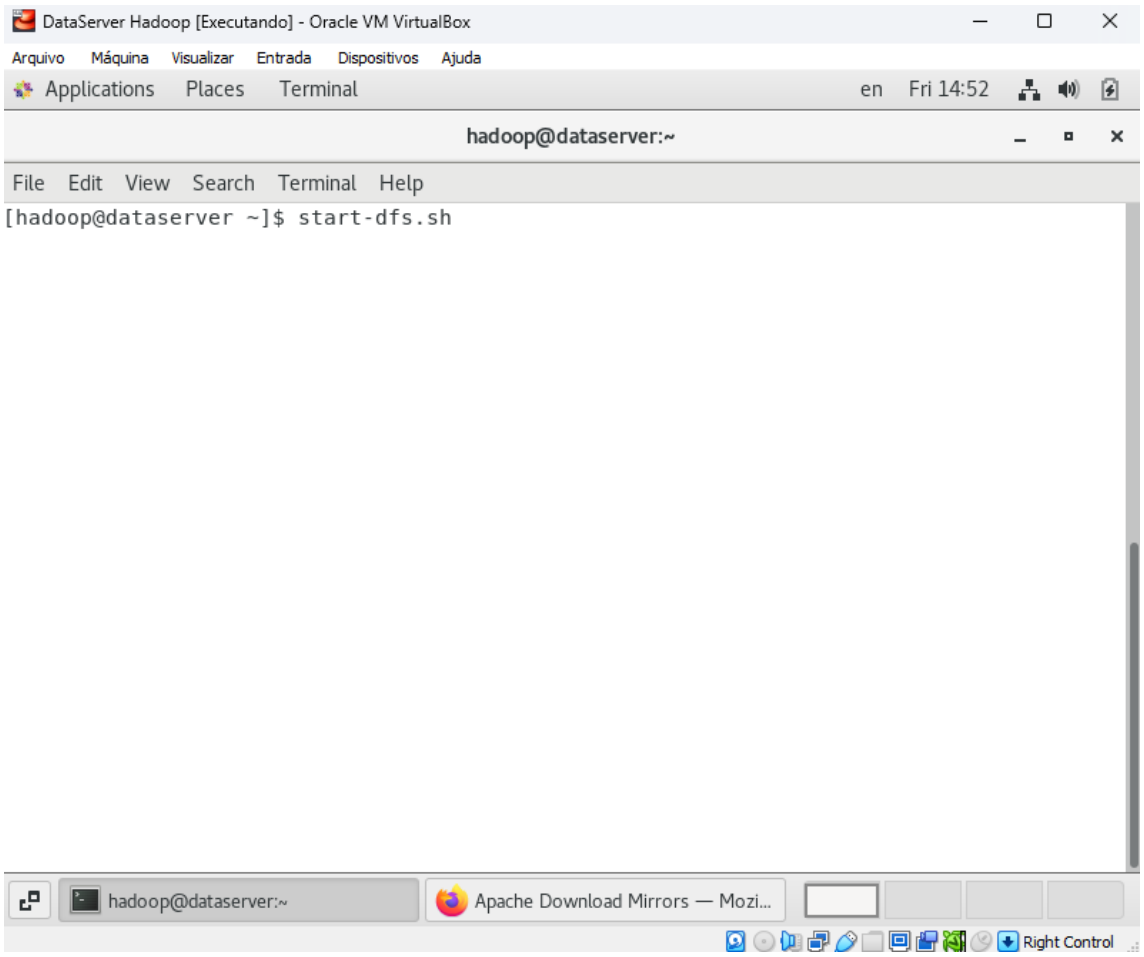


hdfs namenode –format

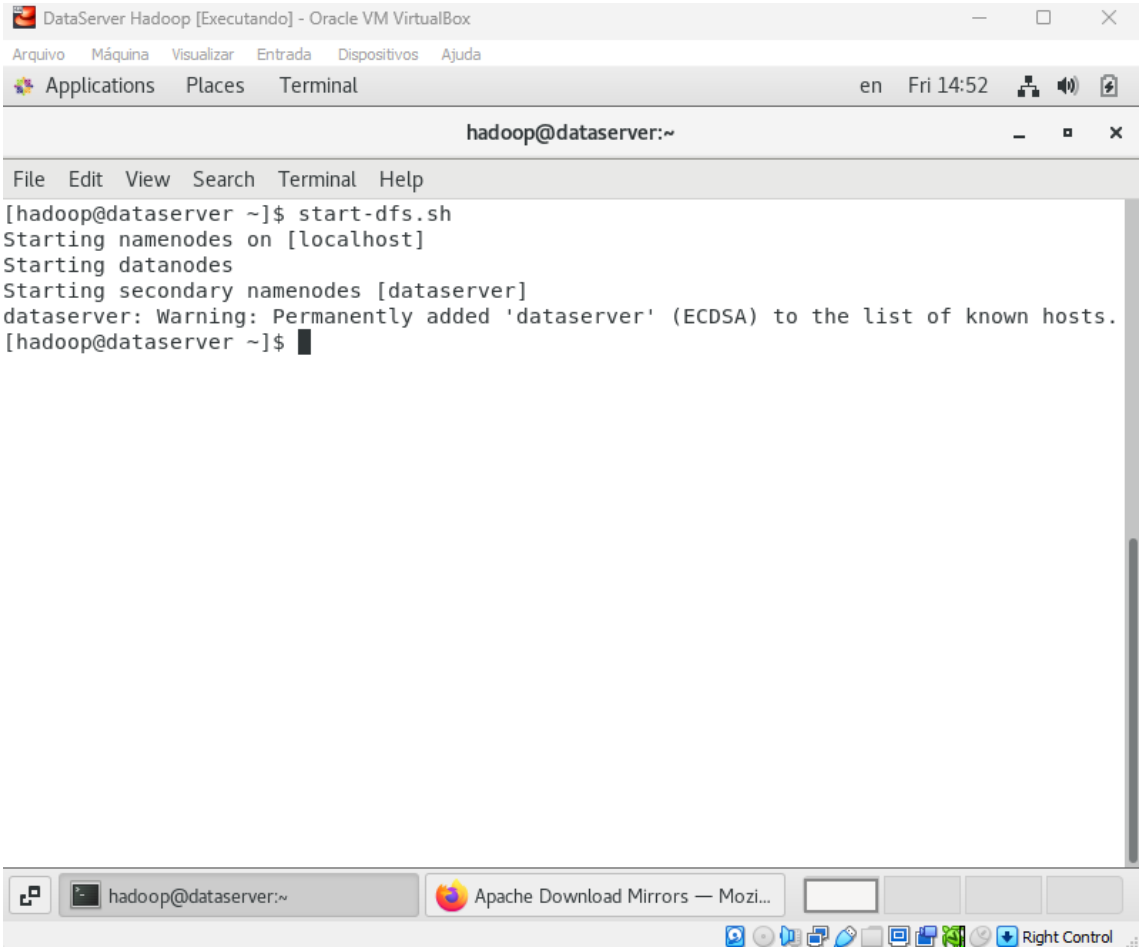


Formatação realizada com sucesso

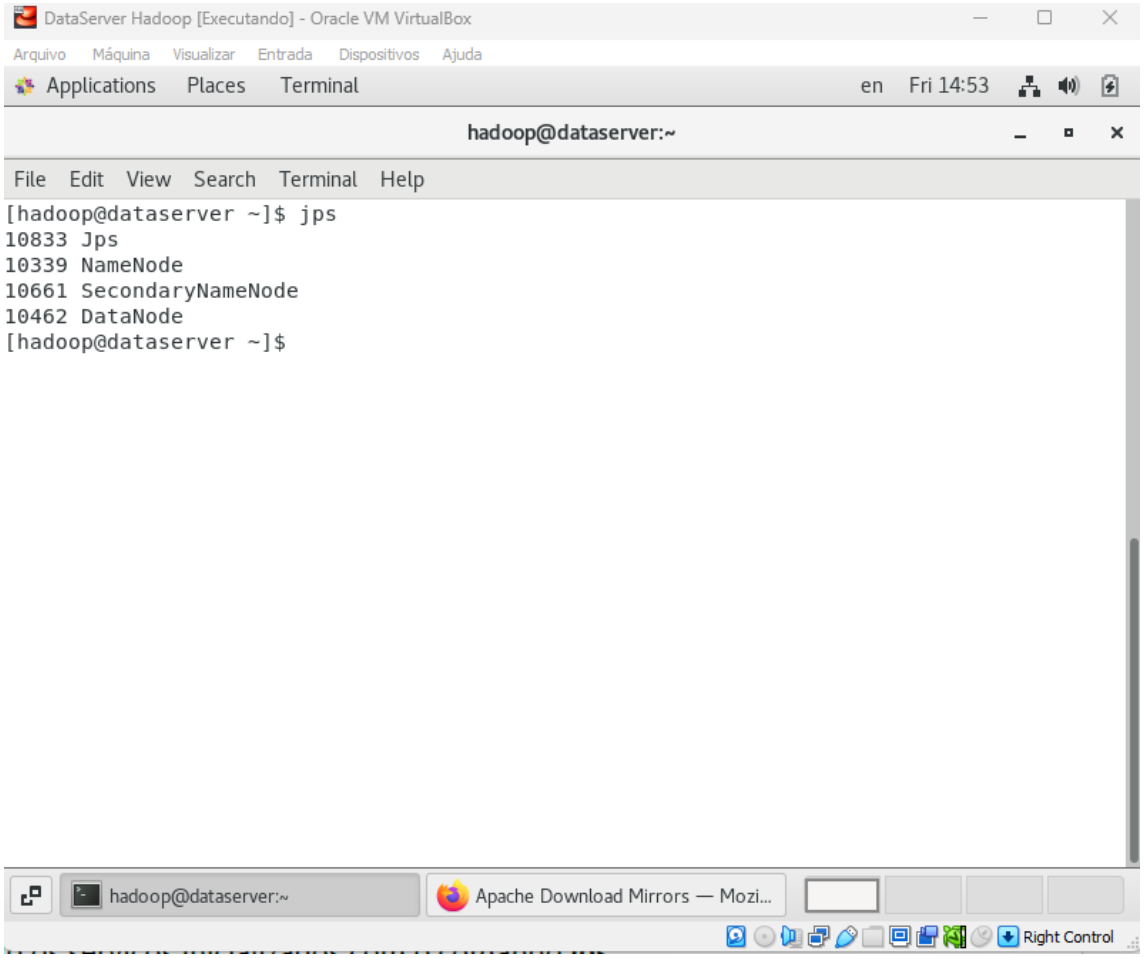
5.4.3 Iniciando o Hadoop



start-dfs.sh

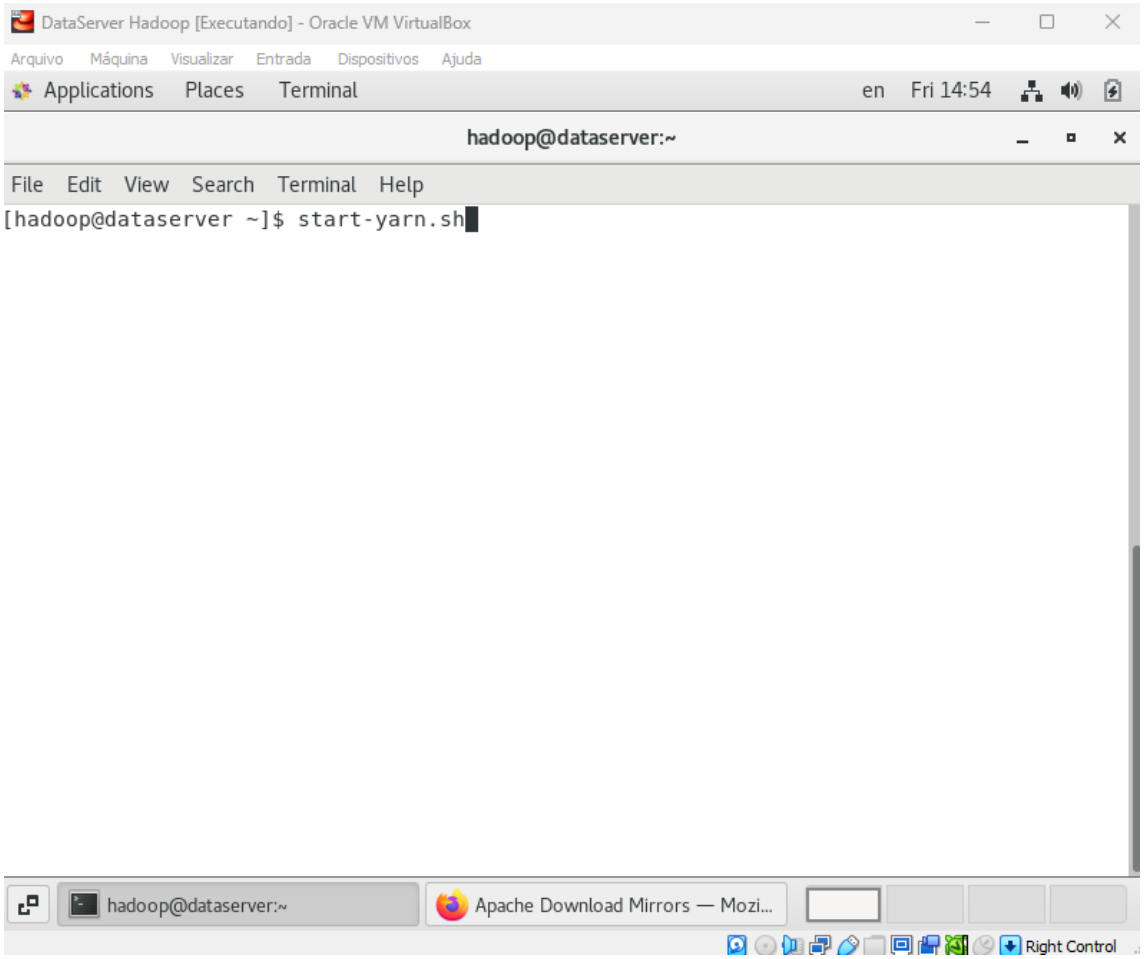


Hadoop iniciado

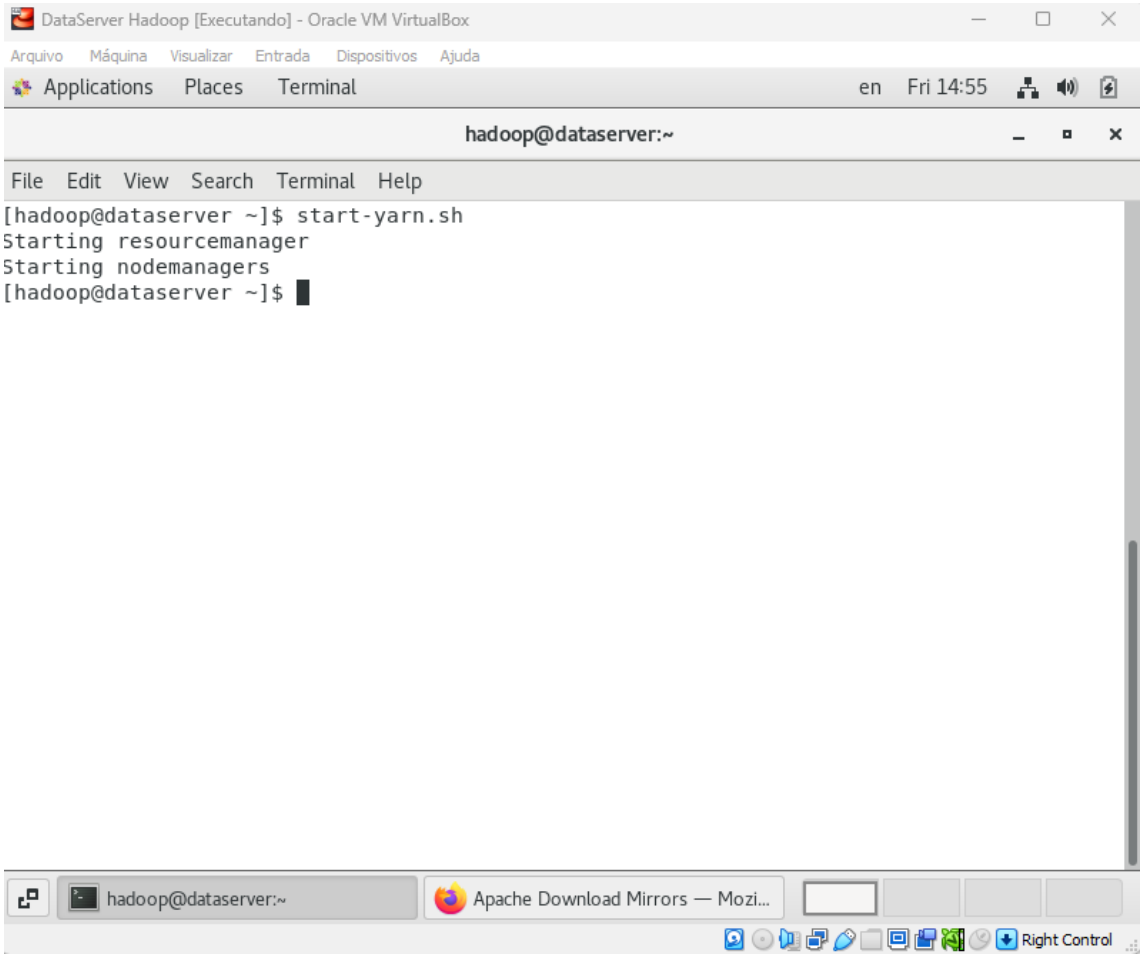


Verificando os serviços inicializados com o comando **jps**

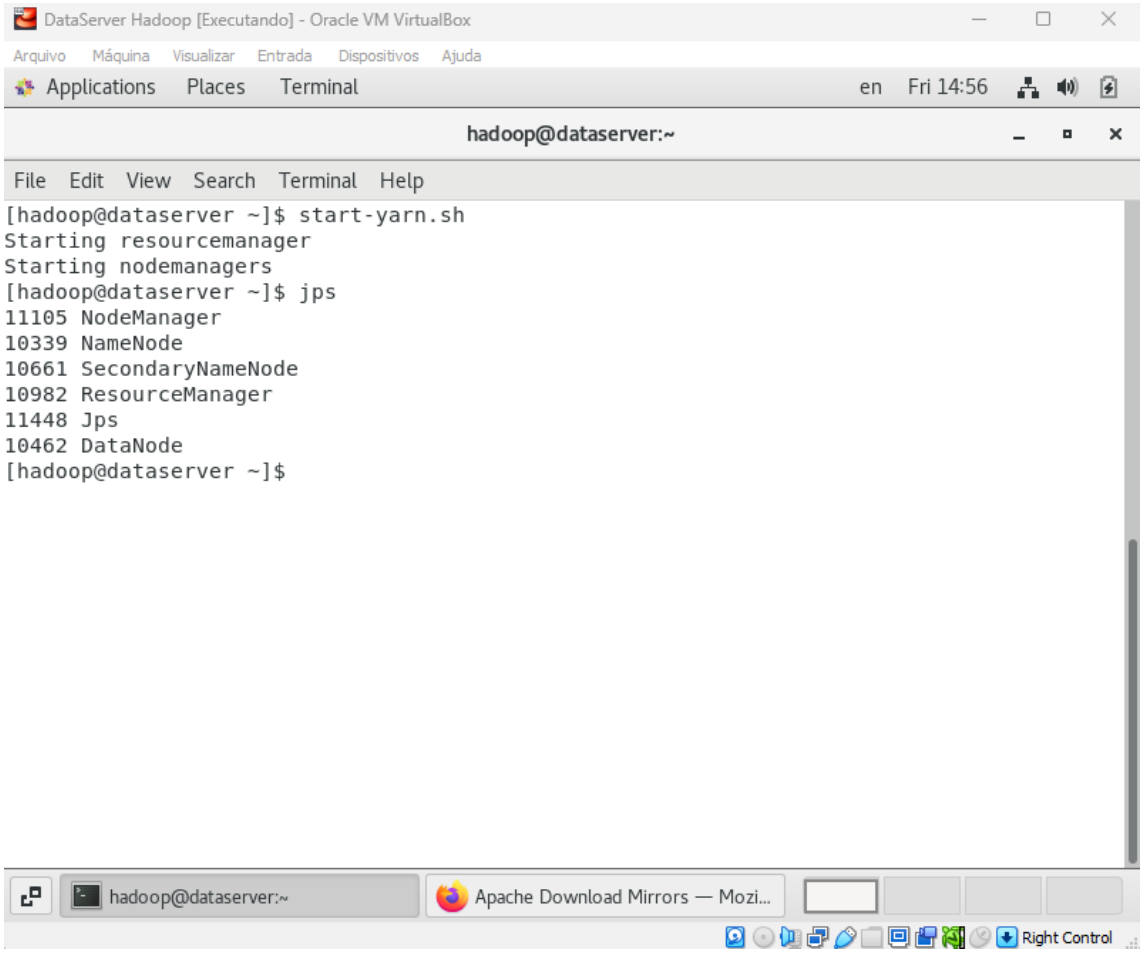
5.4.4 Iniciando o Yarn



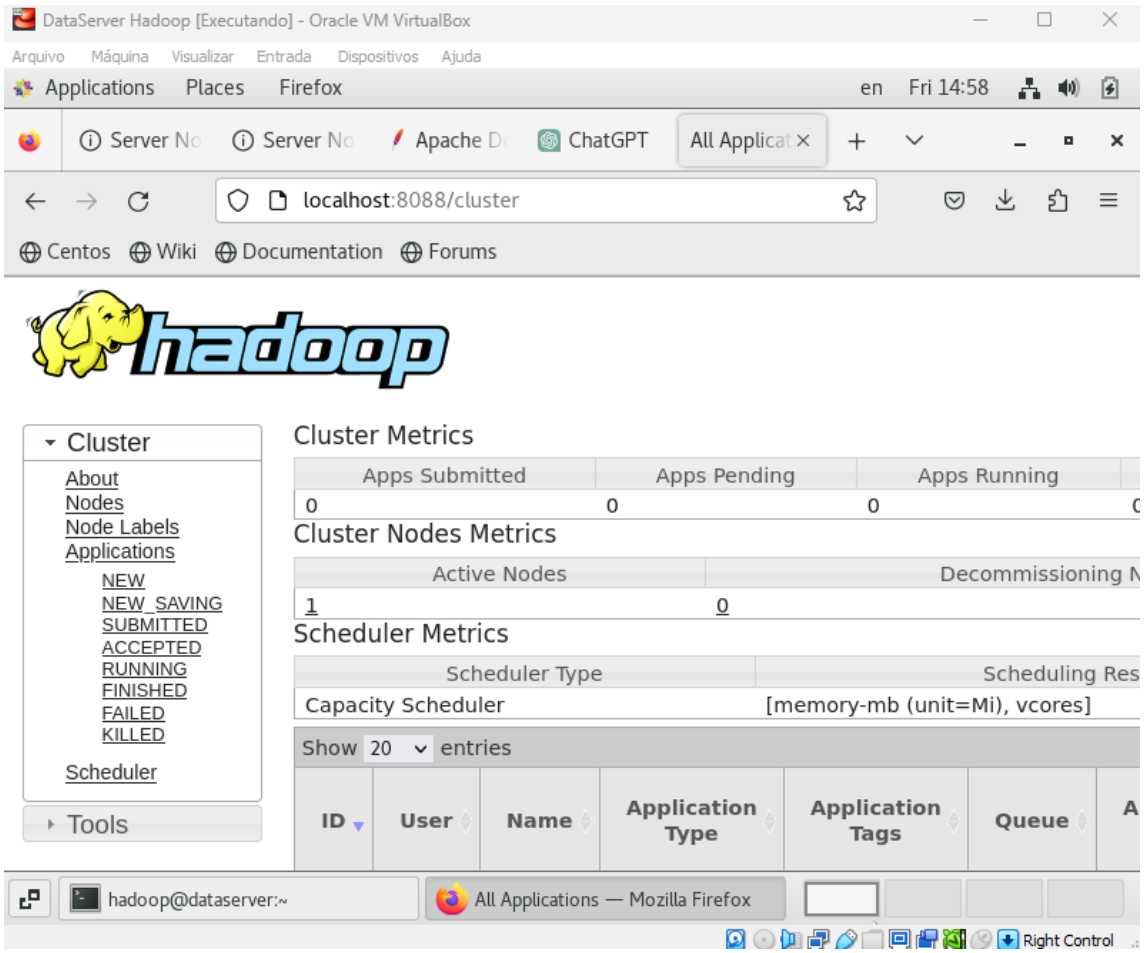
start-yarn.sh



Yarn iniciado

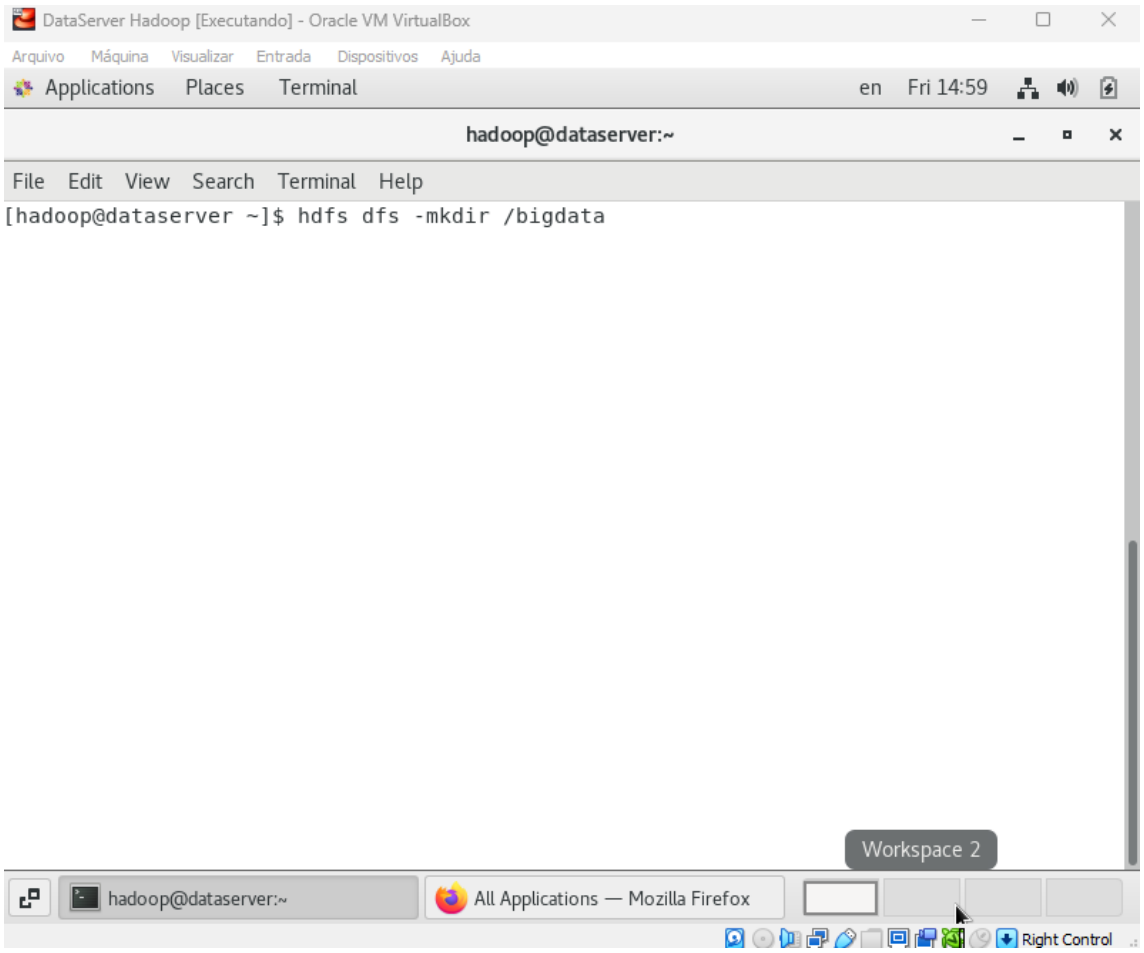


Verificando os serviços com o comando **jps**

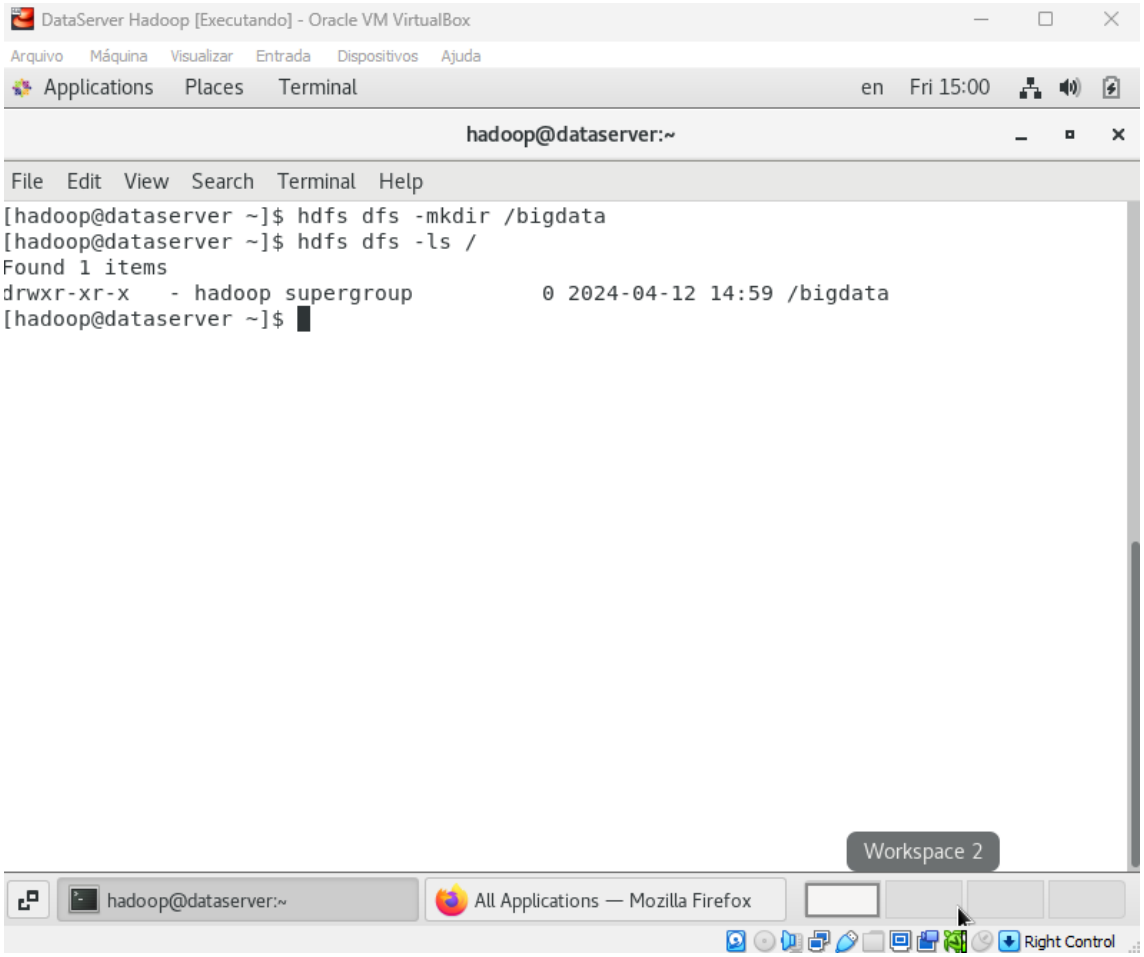


Visualizando jobs – http://localhost:8088

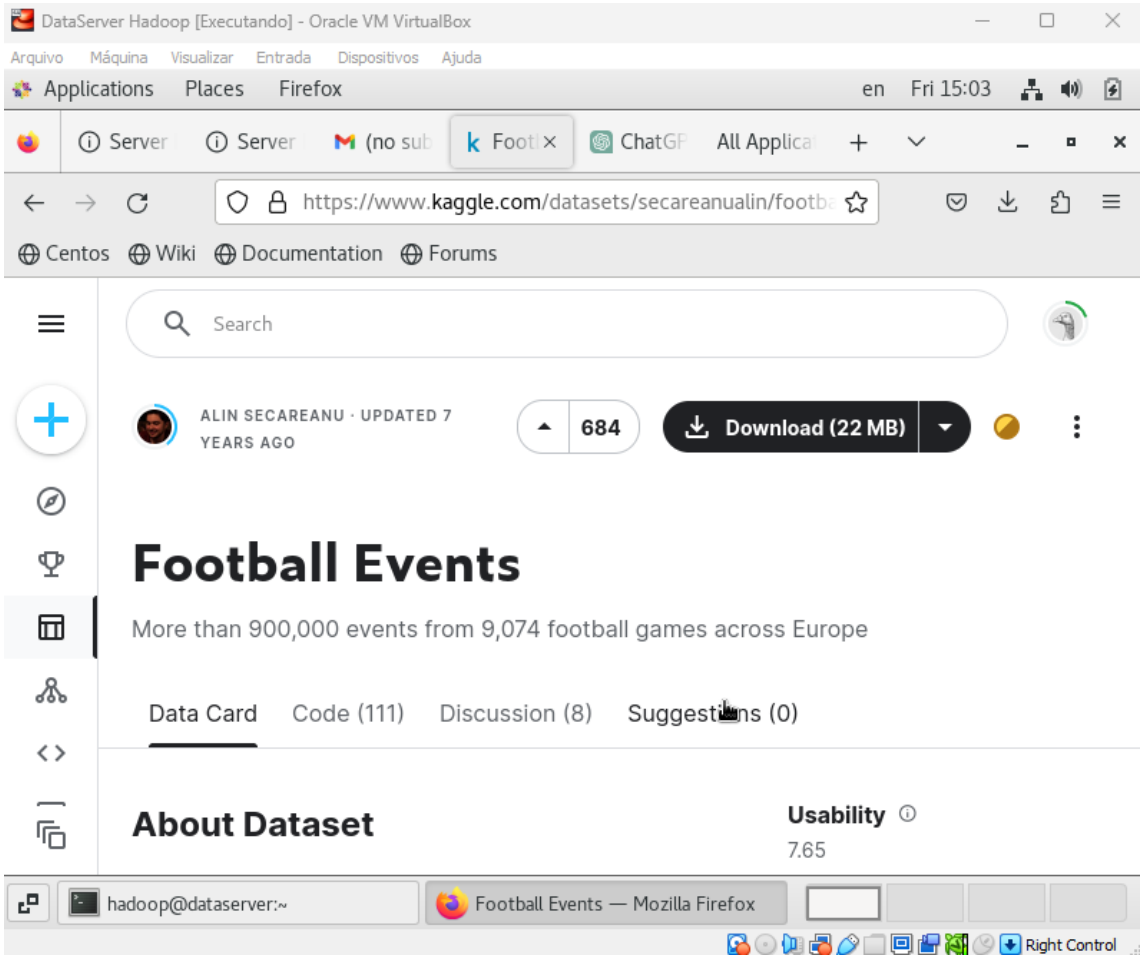
5.5 Processando Big Data



Criar o diretório **bigdata** no HDFS

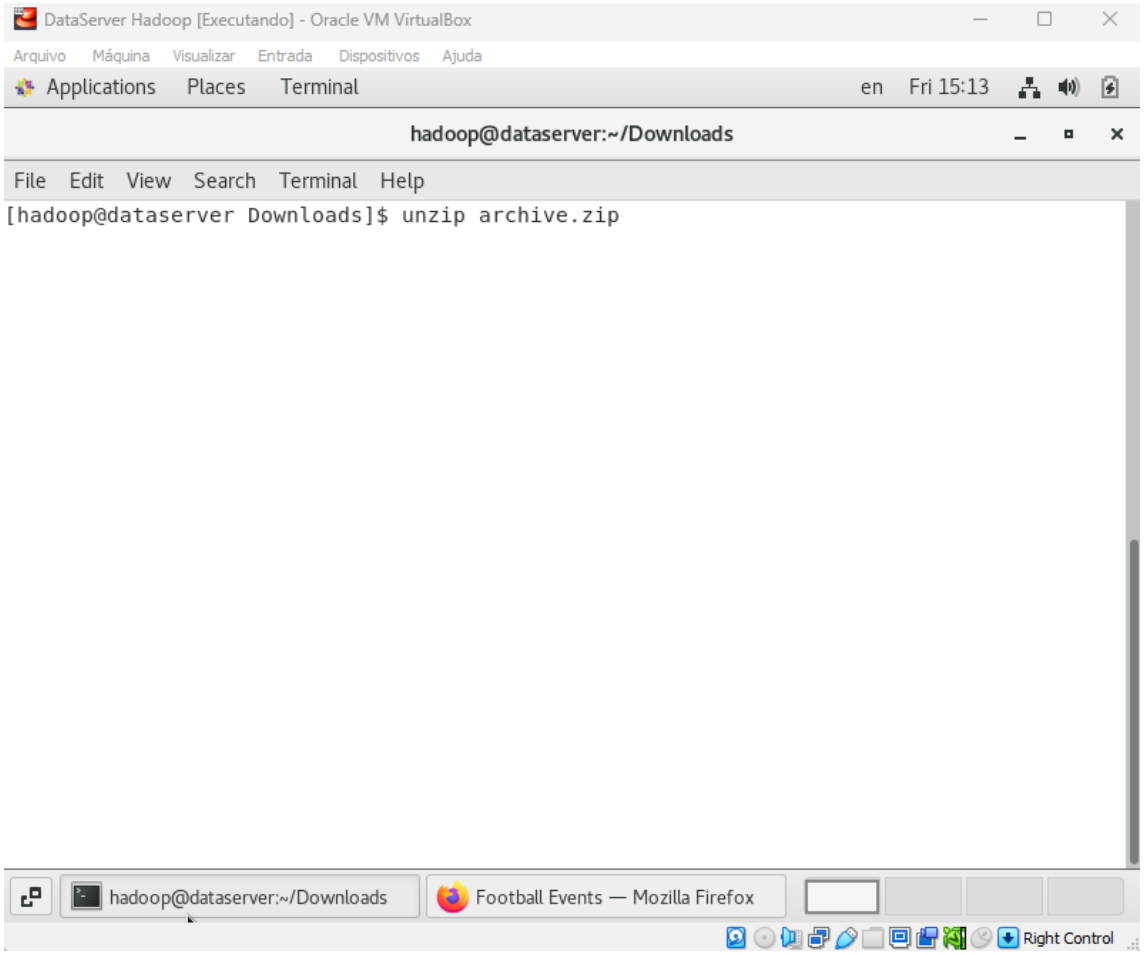


Listar o HDFS e verificar o diretório criado

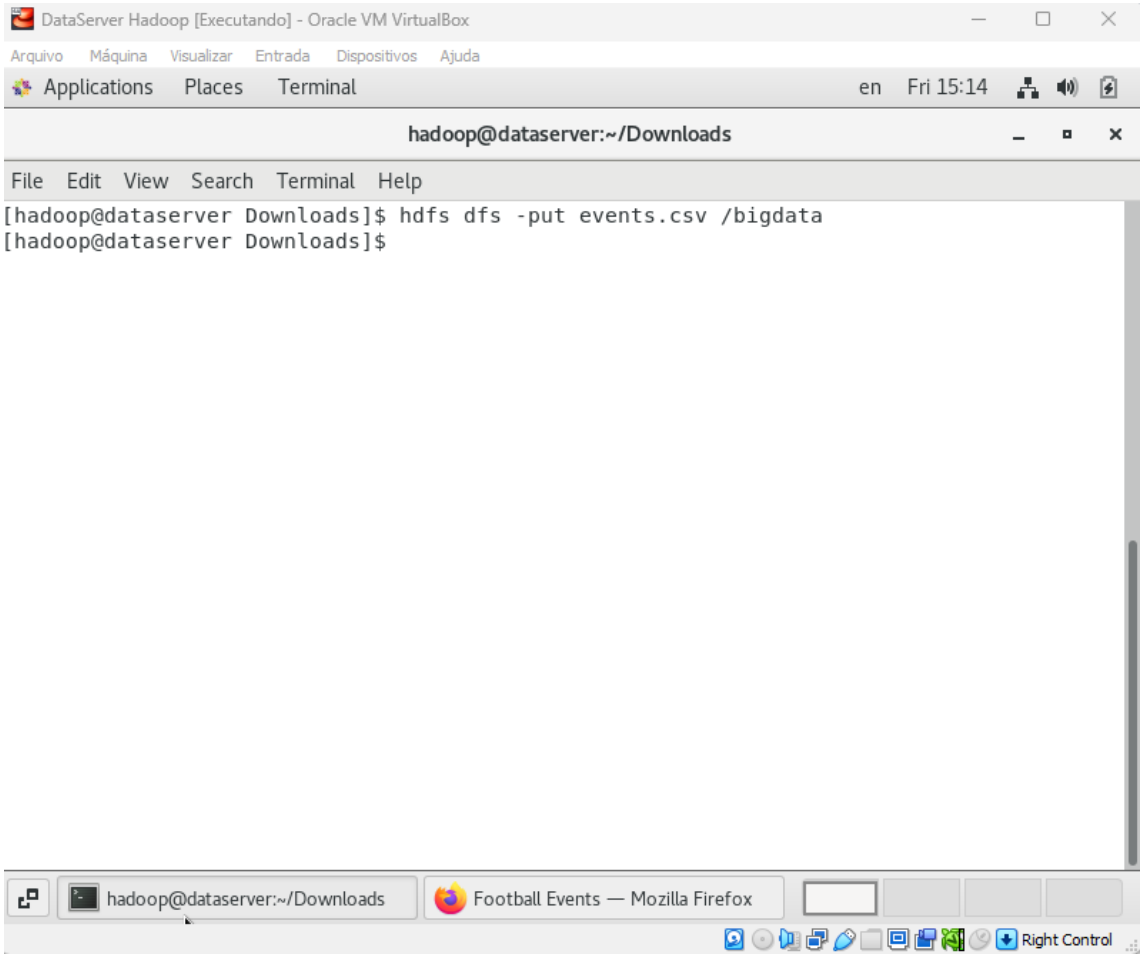


Acessar o portal Kaggle e faça download do arquivo archive.zip

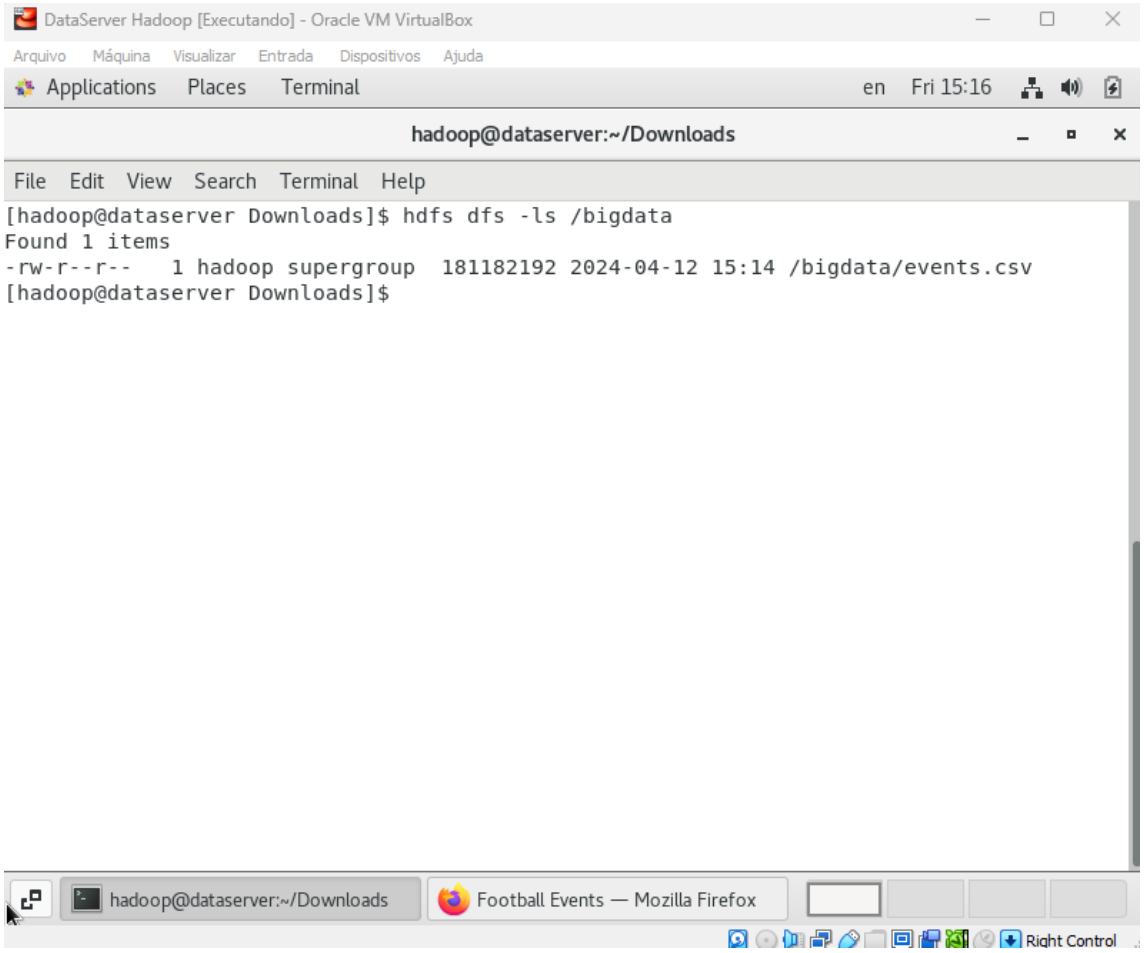
https://www.kaggle.com/datasets/secareanualin/football-events?resource=download



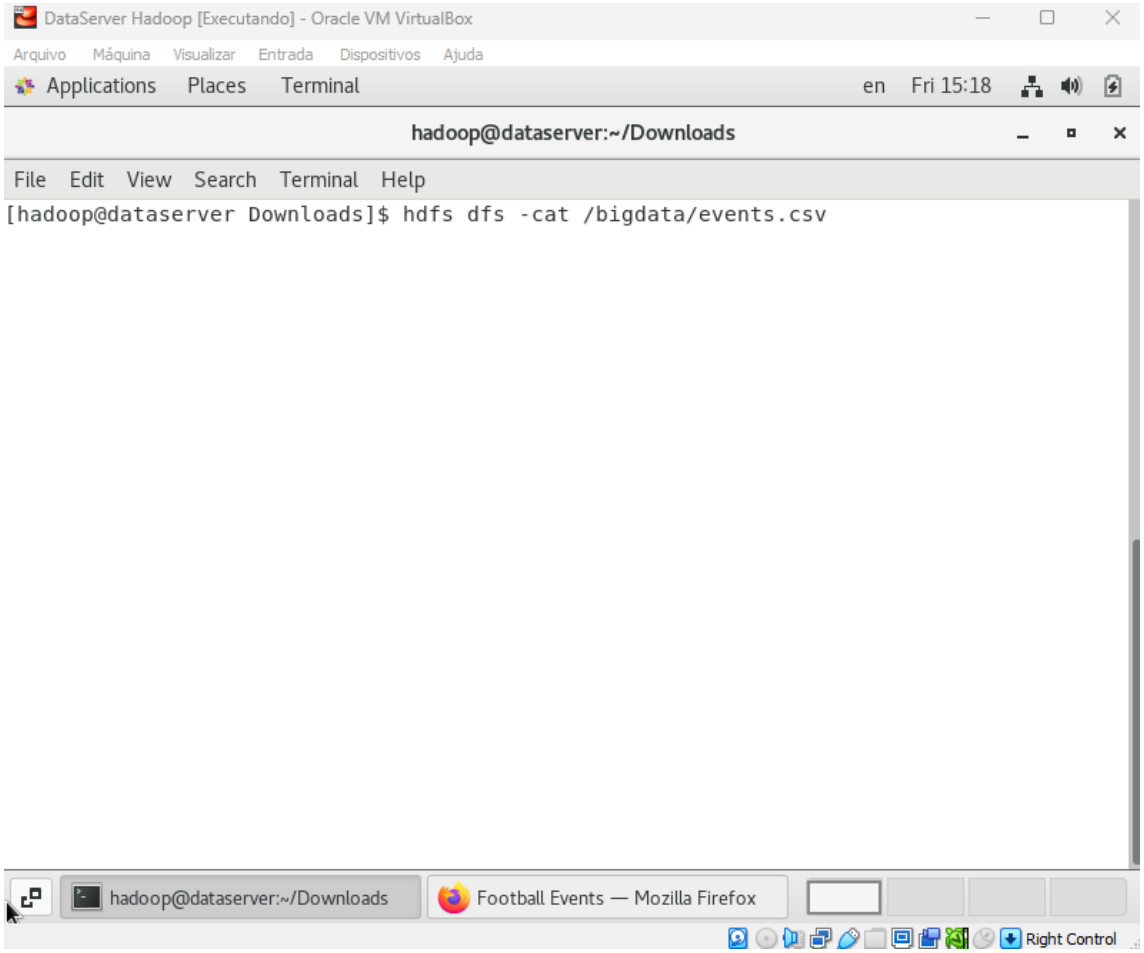
Descompacte-o



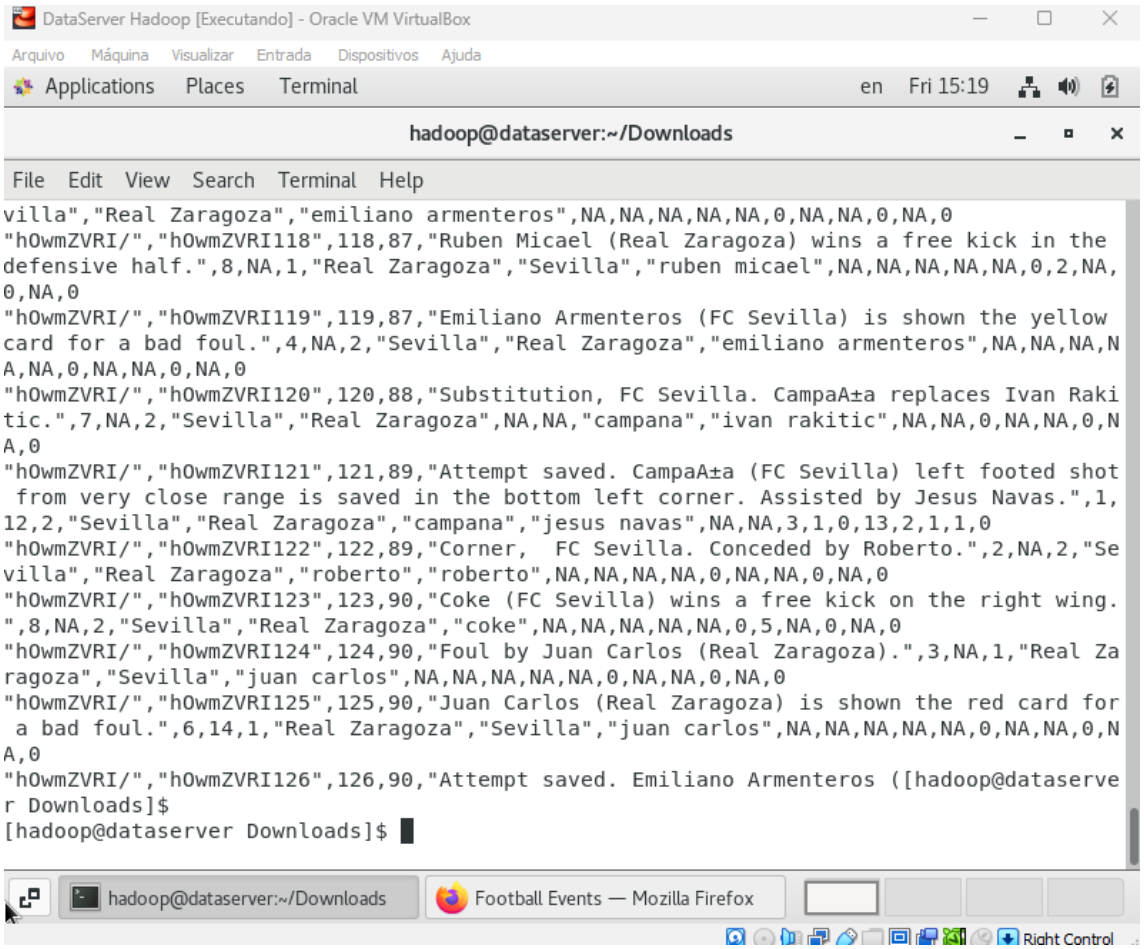
Copiar o arquivo para a pasta diretório no HDFS



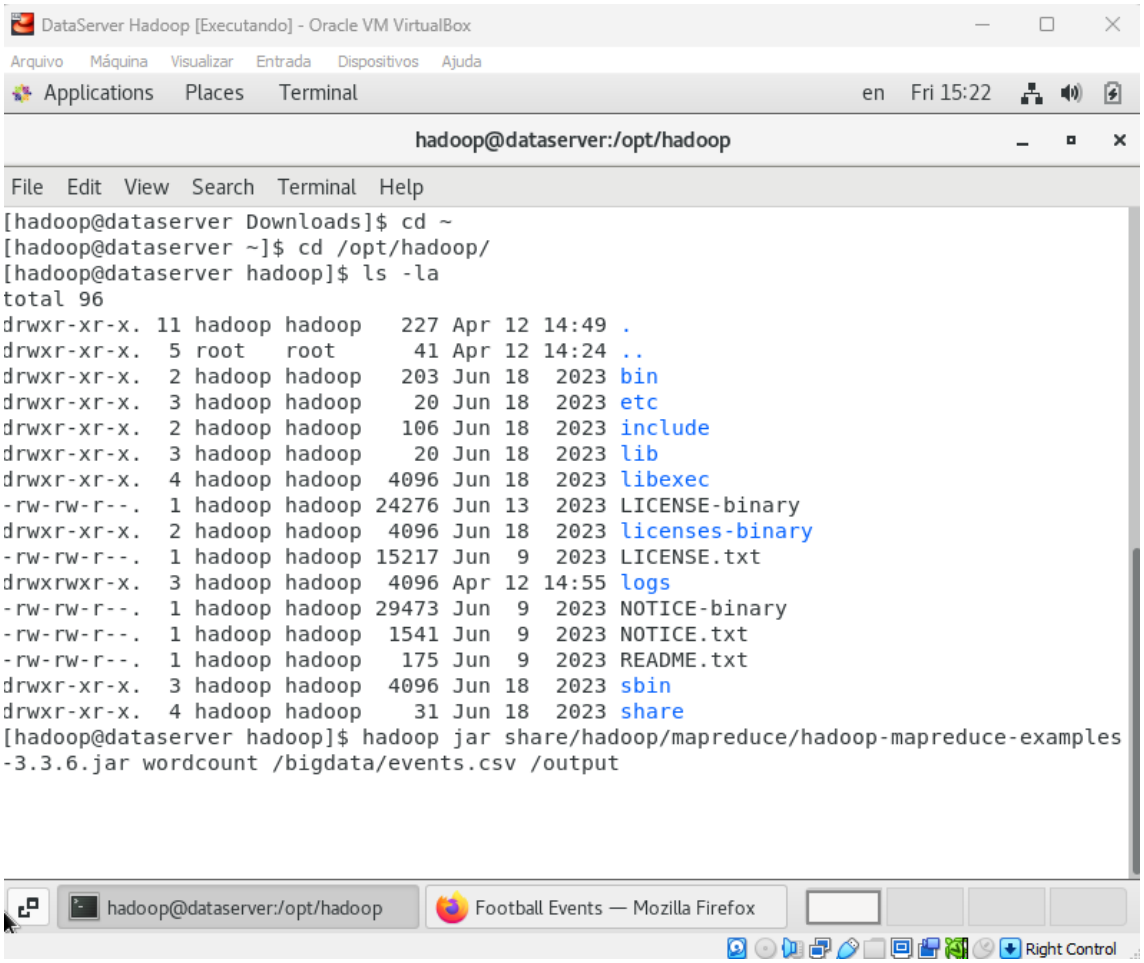
Listar o diretório bigdata



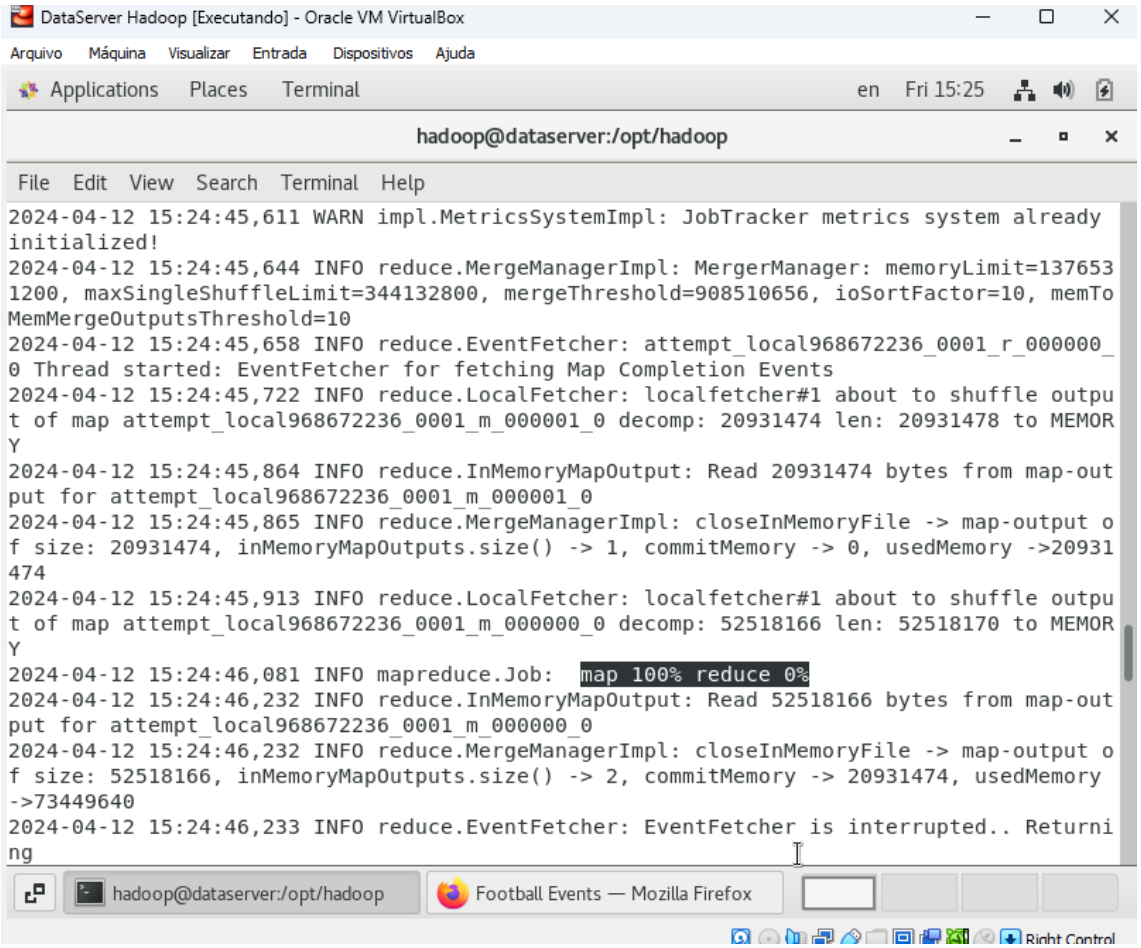
Ver o conteúdo do arquivo



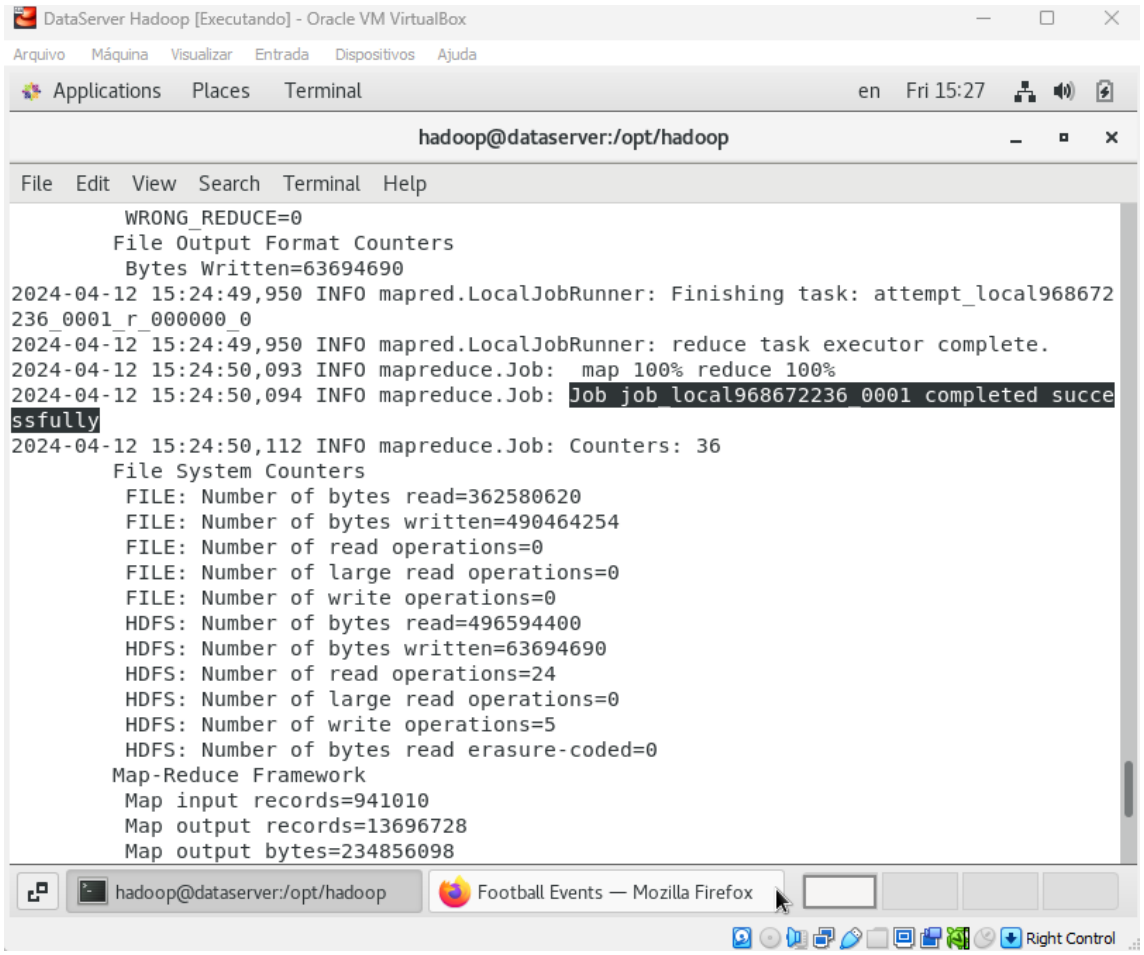
Conteúdo do arquivo já gravado no HDFS



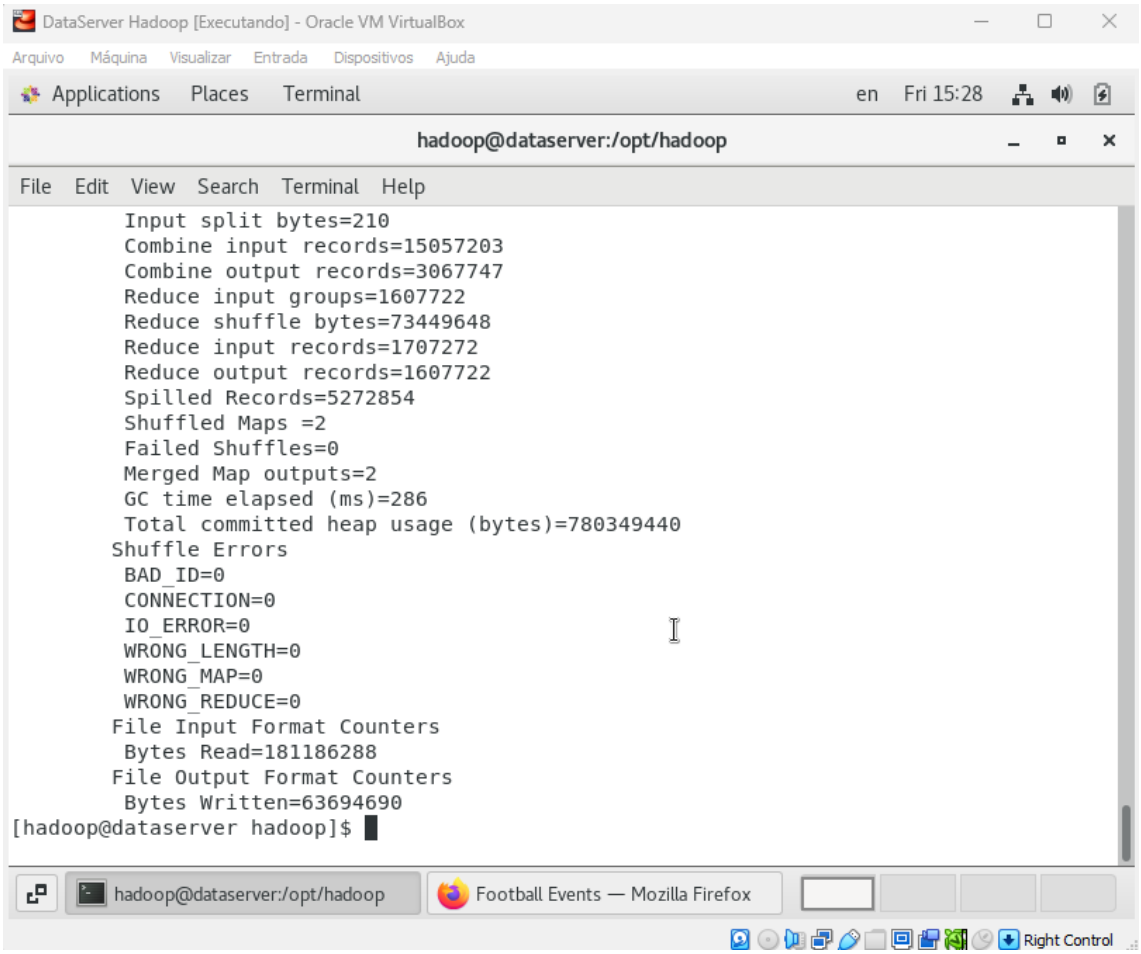
A instalação do Hadoop possui um job chamado wordcount, que pode ser usado como exemplo para processamento de Big Data. Basicamente, o job conta a ocorrência de cada palavra no arquivo. Vou executar com o comando acima (a versão do arquivo .jar é a mesma versão do Hadoop que eu instalei).



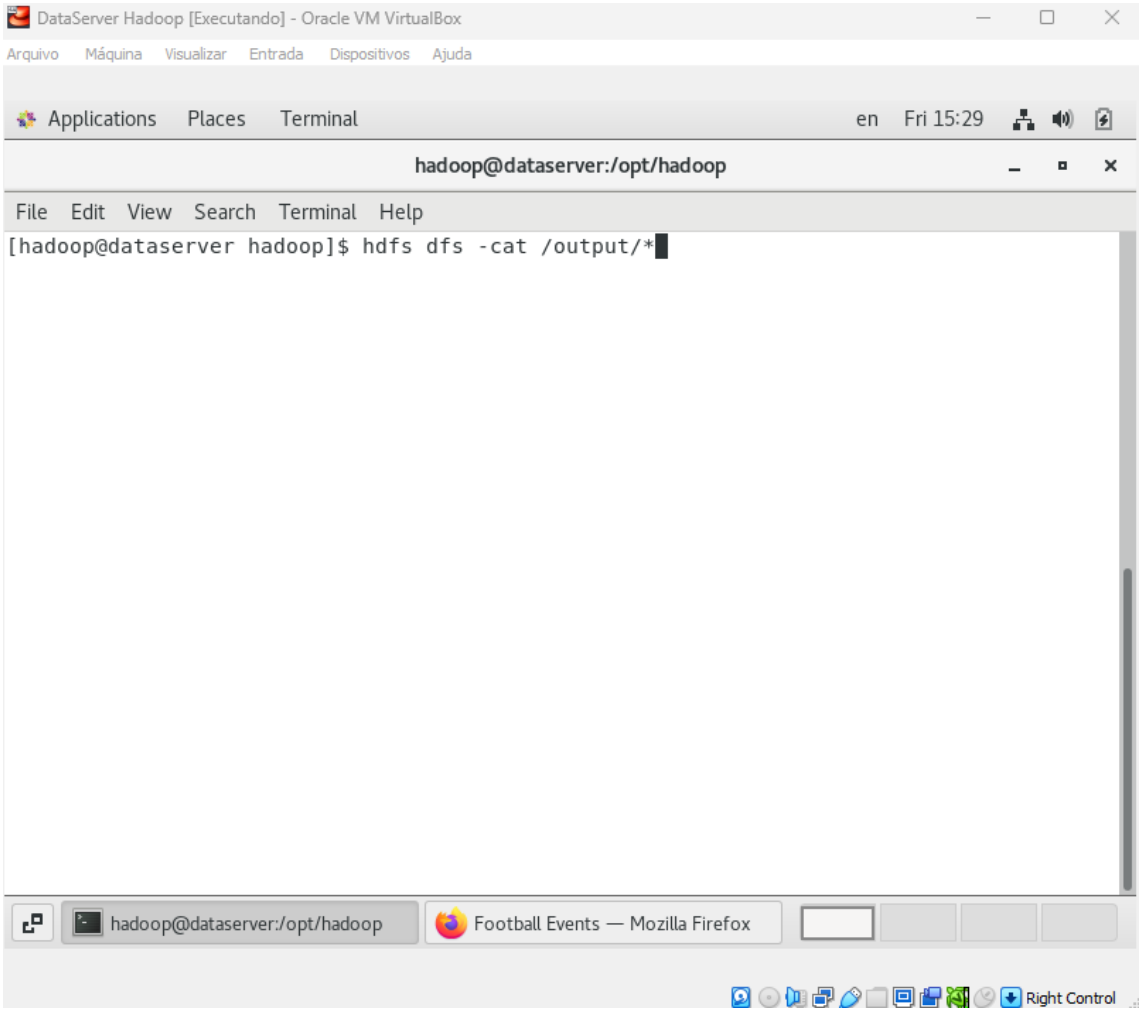
Job sendo processado



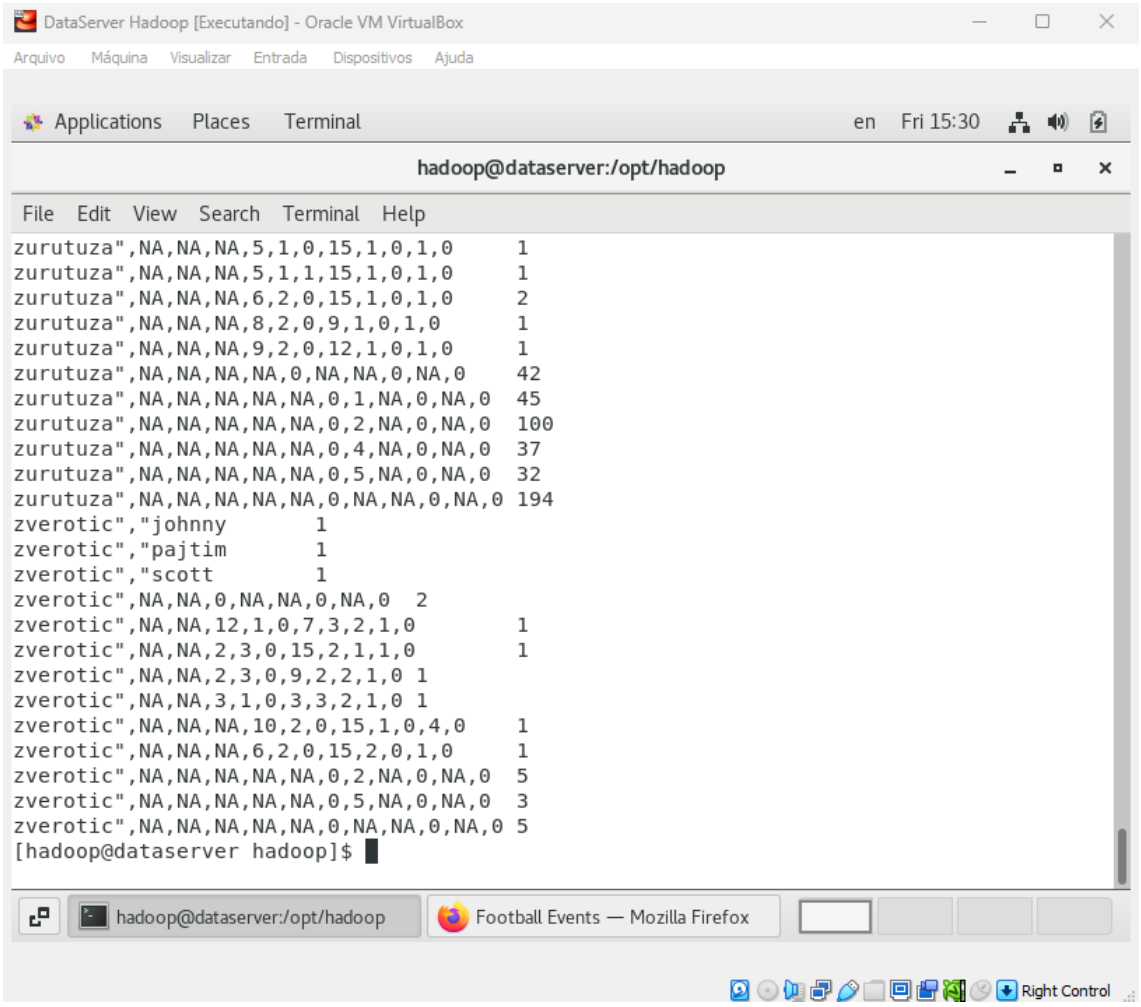
job processado com sucesso



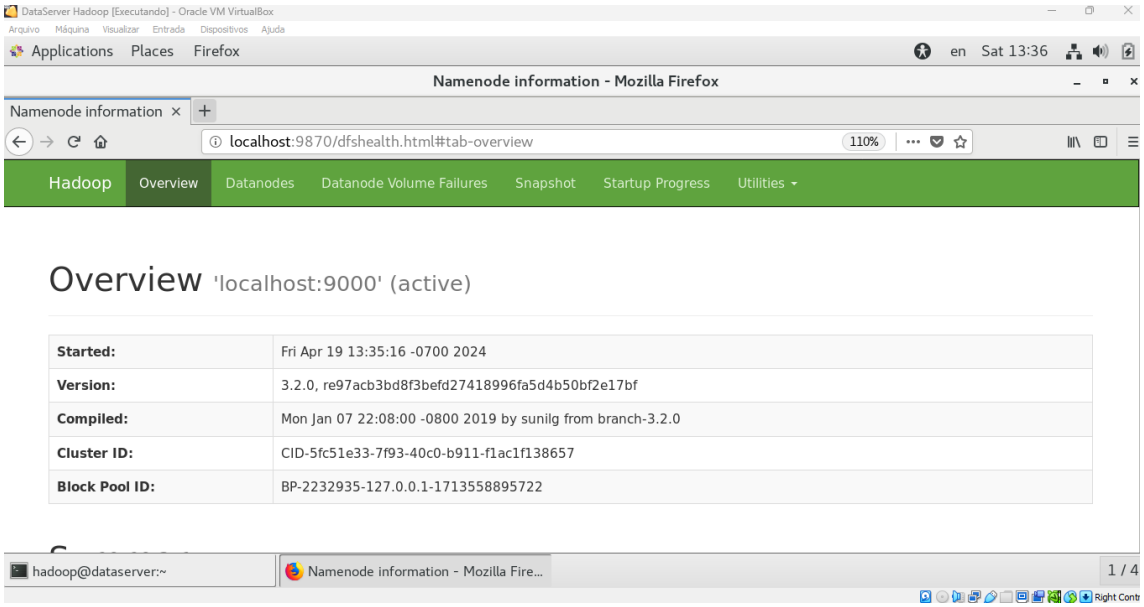
job processado com sucesso



Visualizar o resultado do processamento



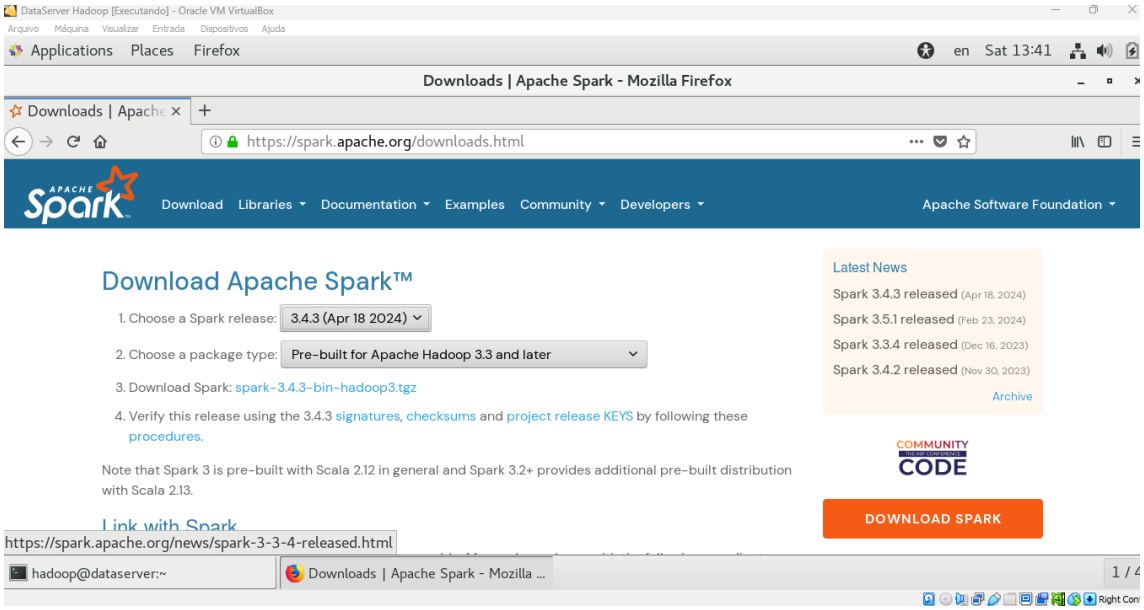
Arquivo processado. Número de ocorrência de cada palavra/termo no arquivo.



Acesso ao Hadoop pelo browser hhtp://localhost:9870

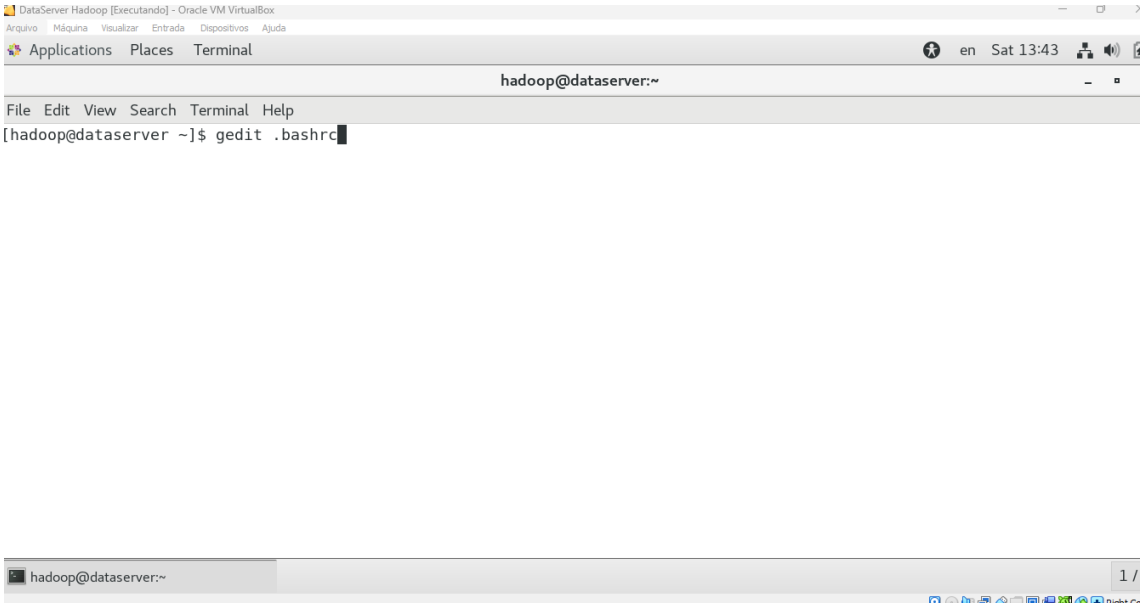
6 Instalação e Configuração do Spark

6.1 Download e Instalação do Spark

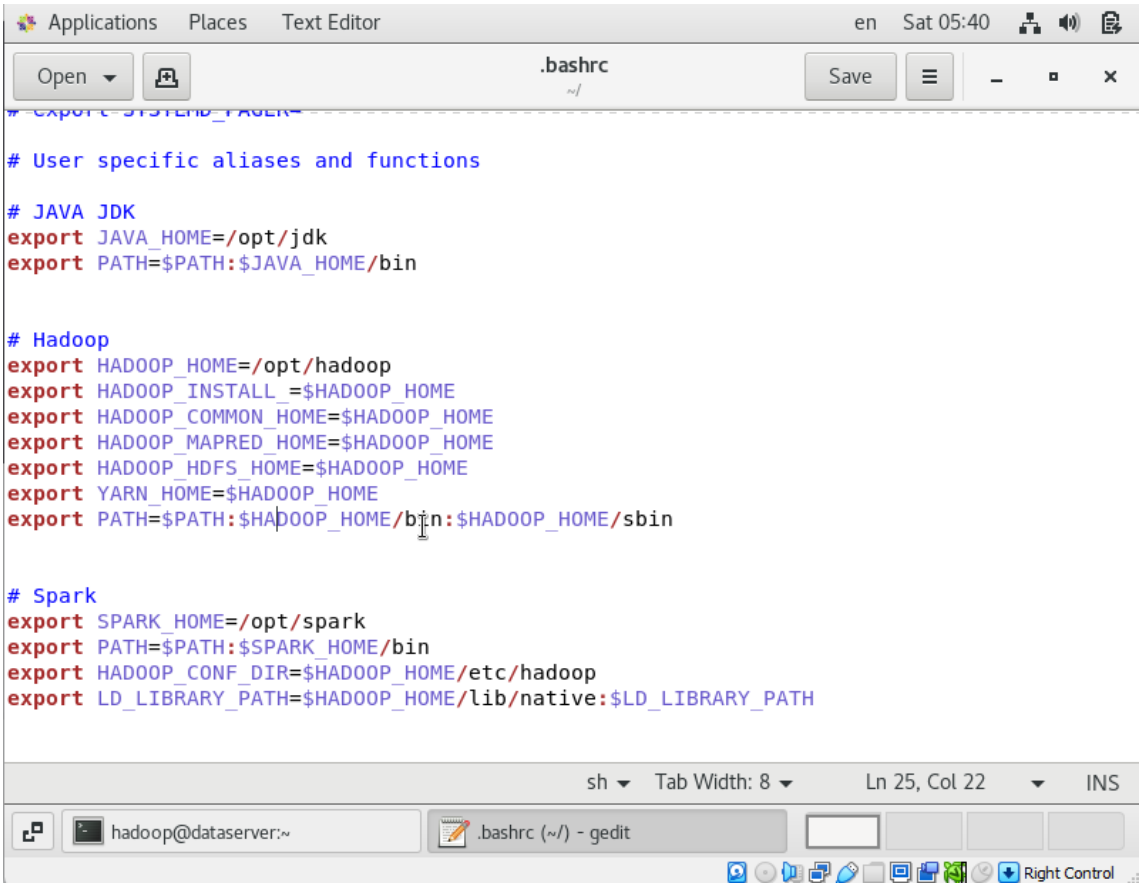


Download do Spark – Versão 3.4.3

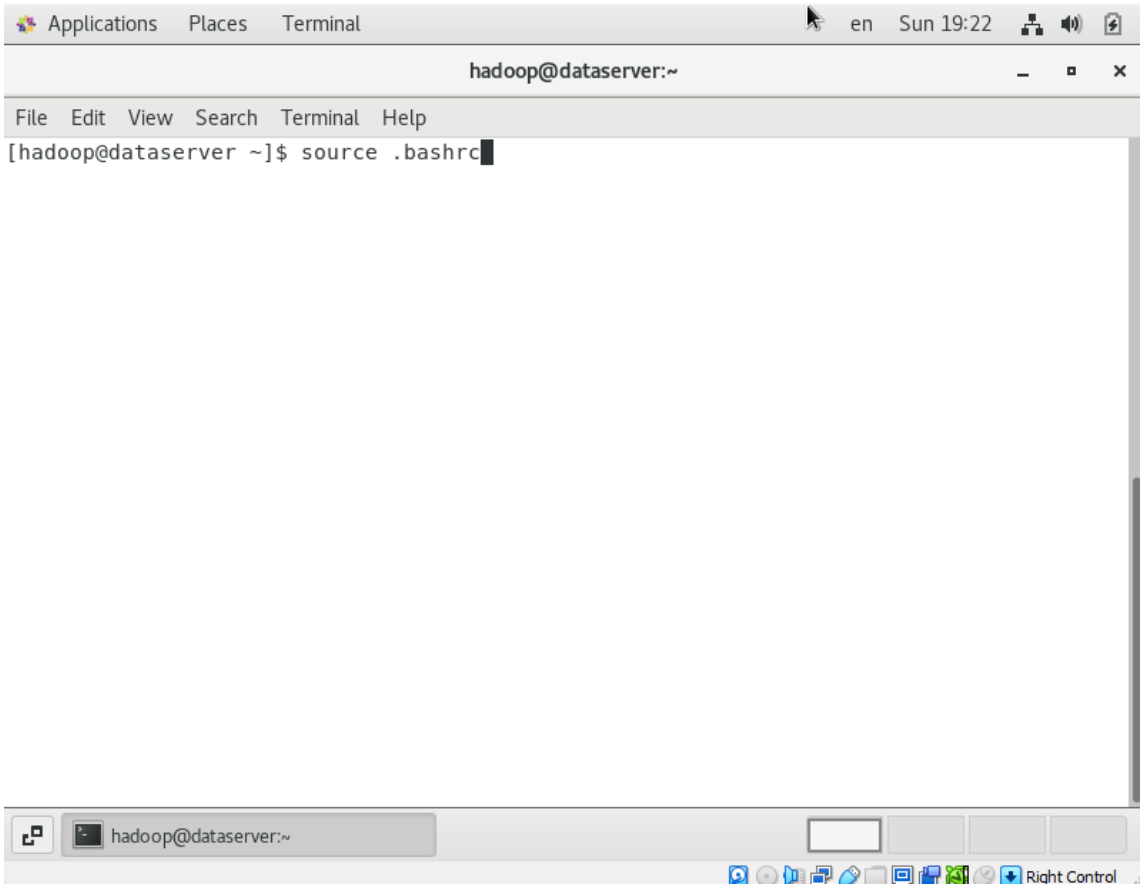
Faça o download, descompacte o arquivo e mova o diretório para /opt/spark da mesma forma como foi feito com o Java JDK e com o Hadoop.



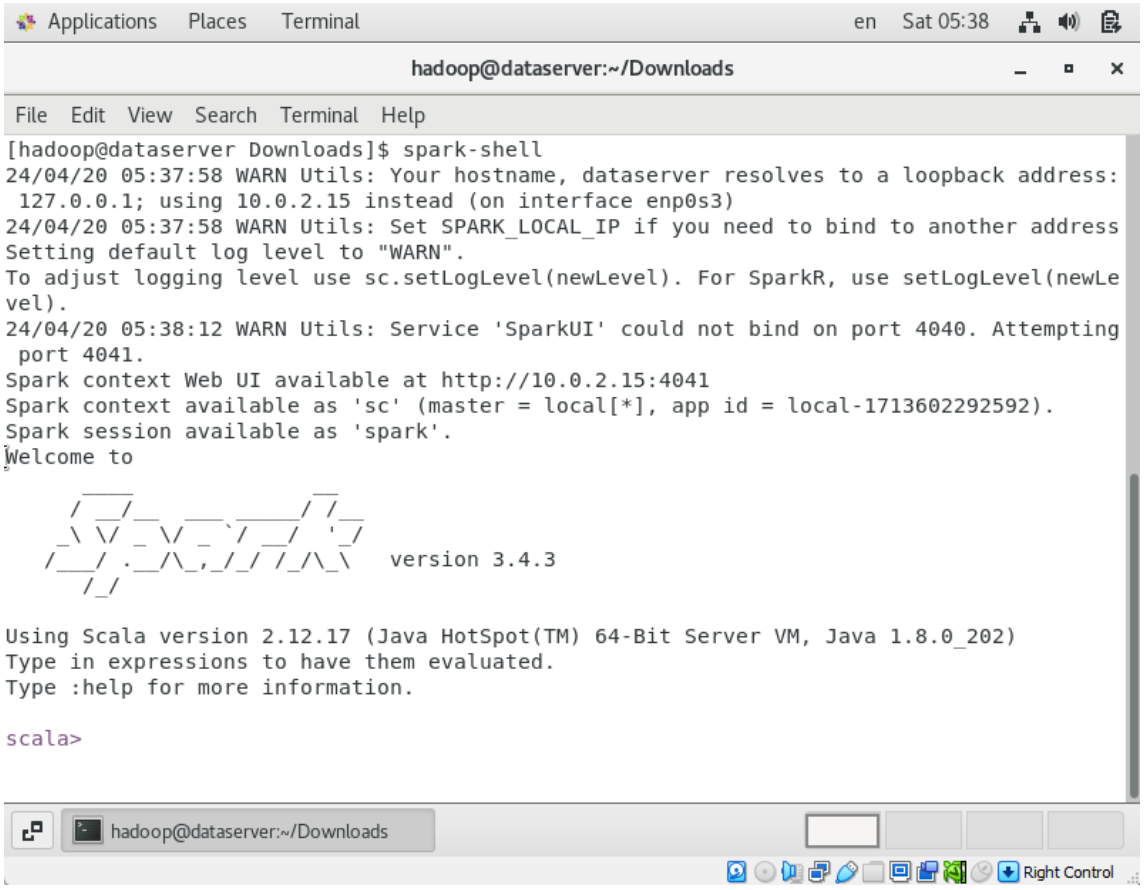
Editando o arquivo .bashrc



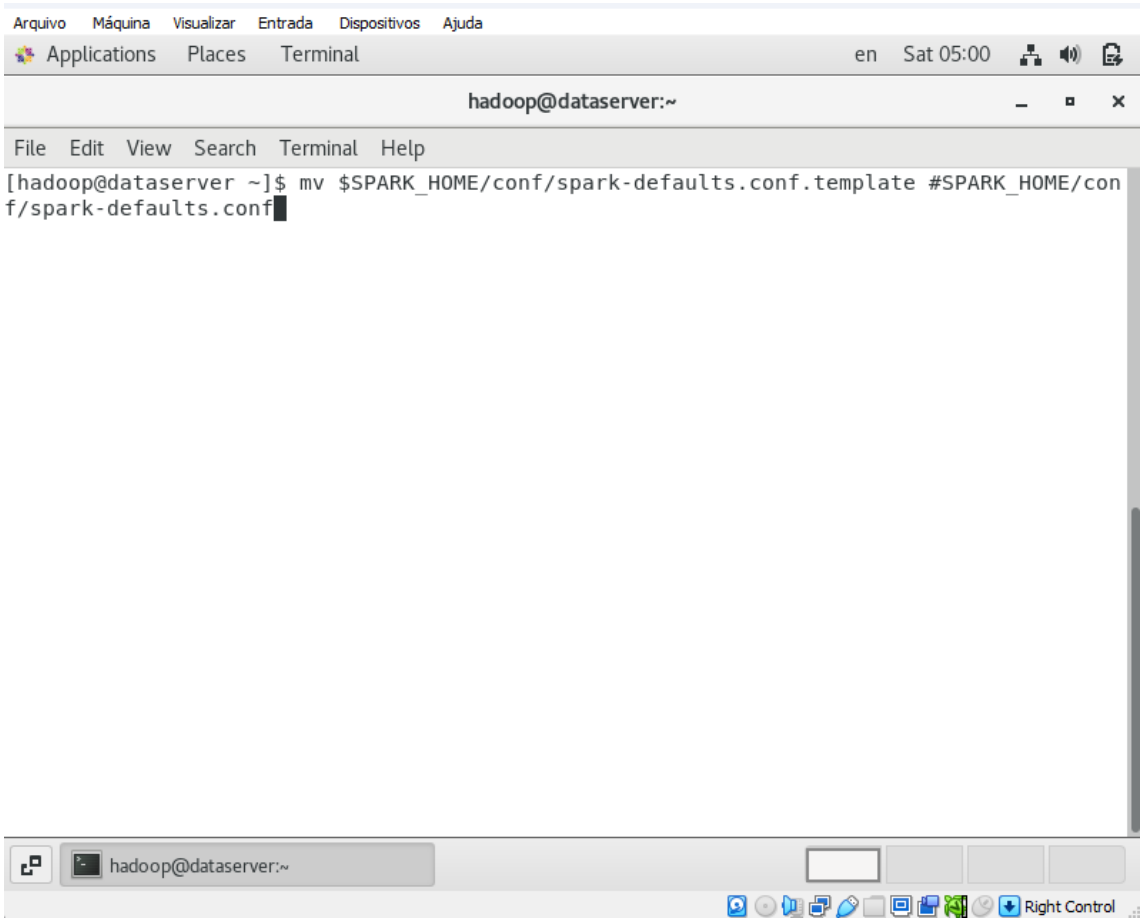
Incluir variáveis Spark



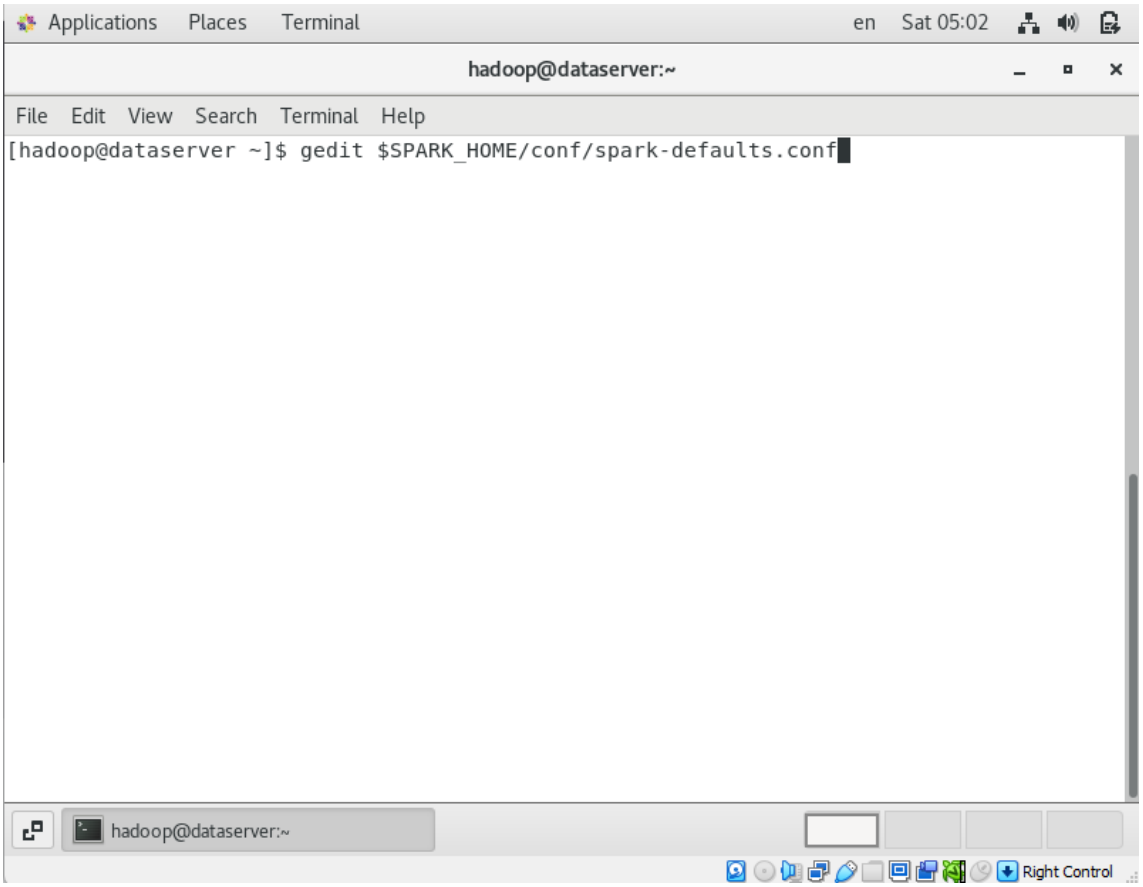
source .bashrc



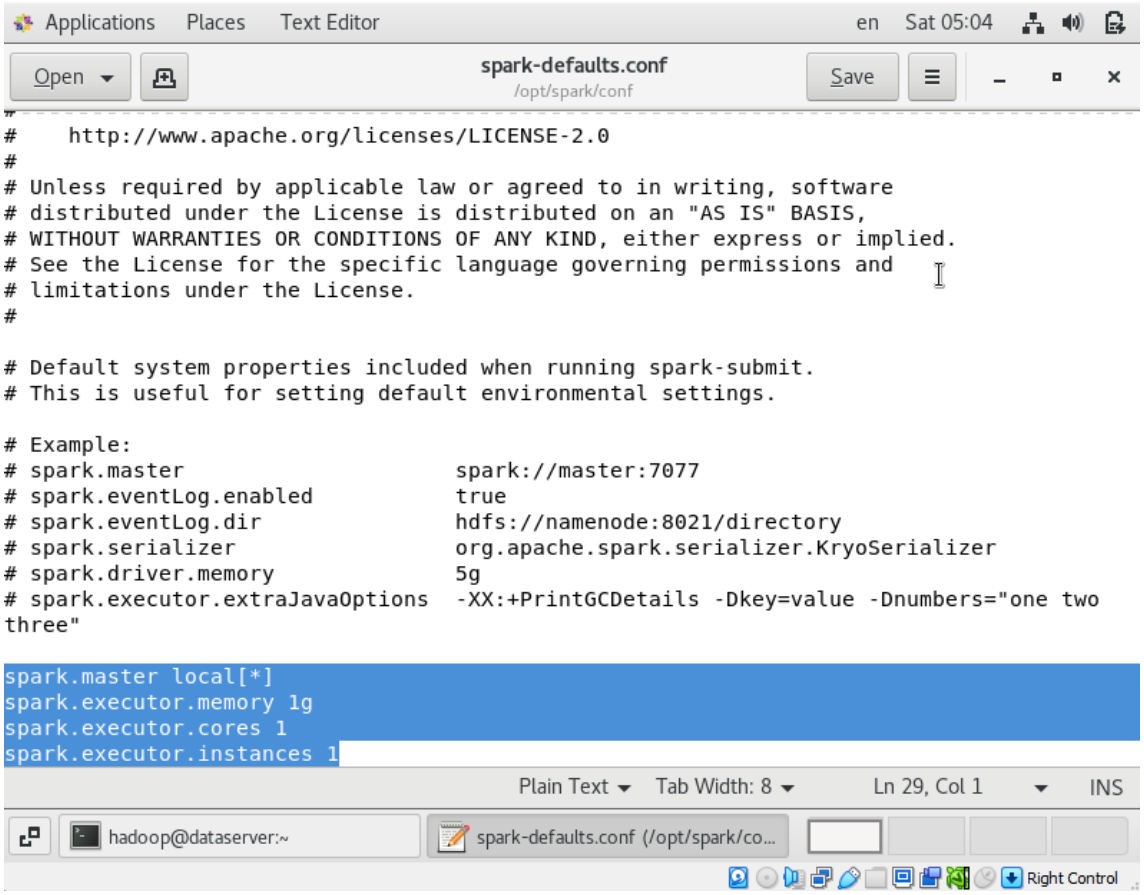
Spark shell - Digite :q para sair do shell



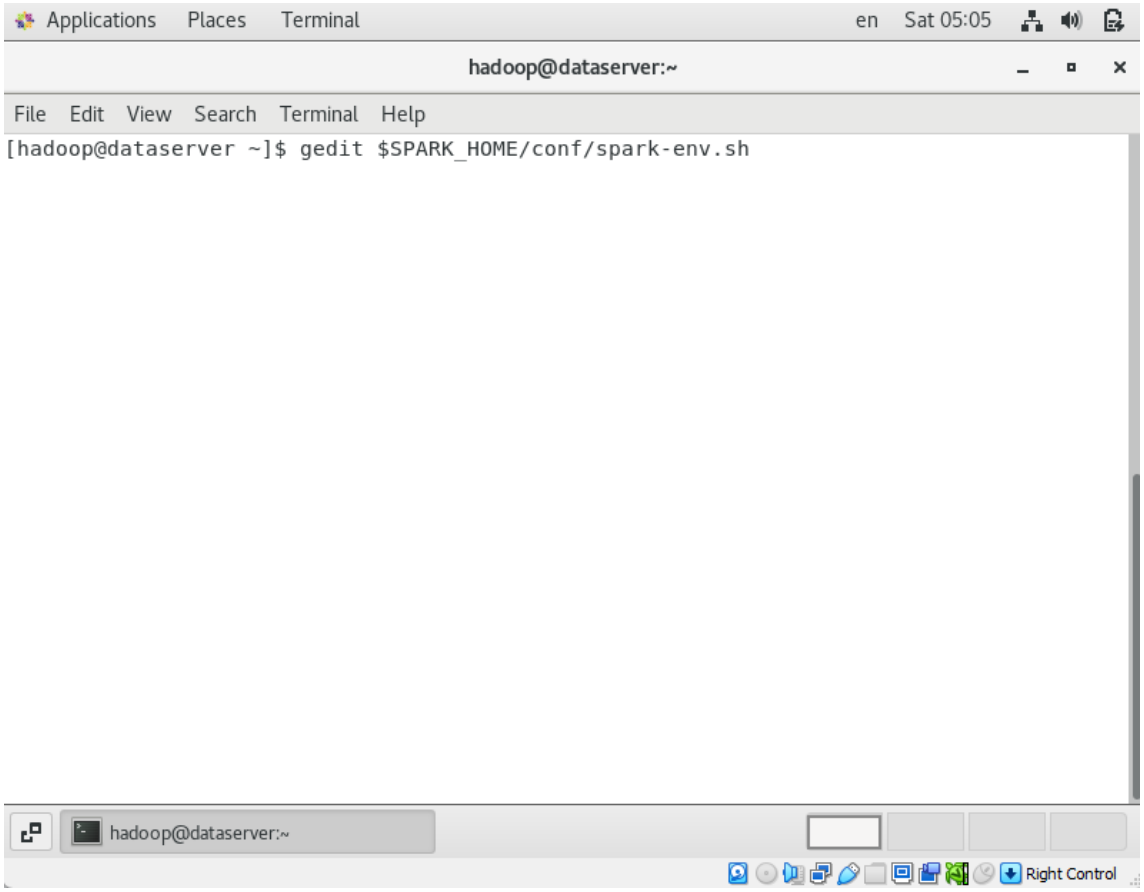
Renomeie o arquivo spark-defaults.conf.templates



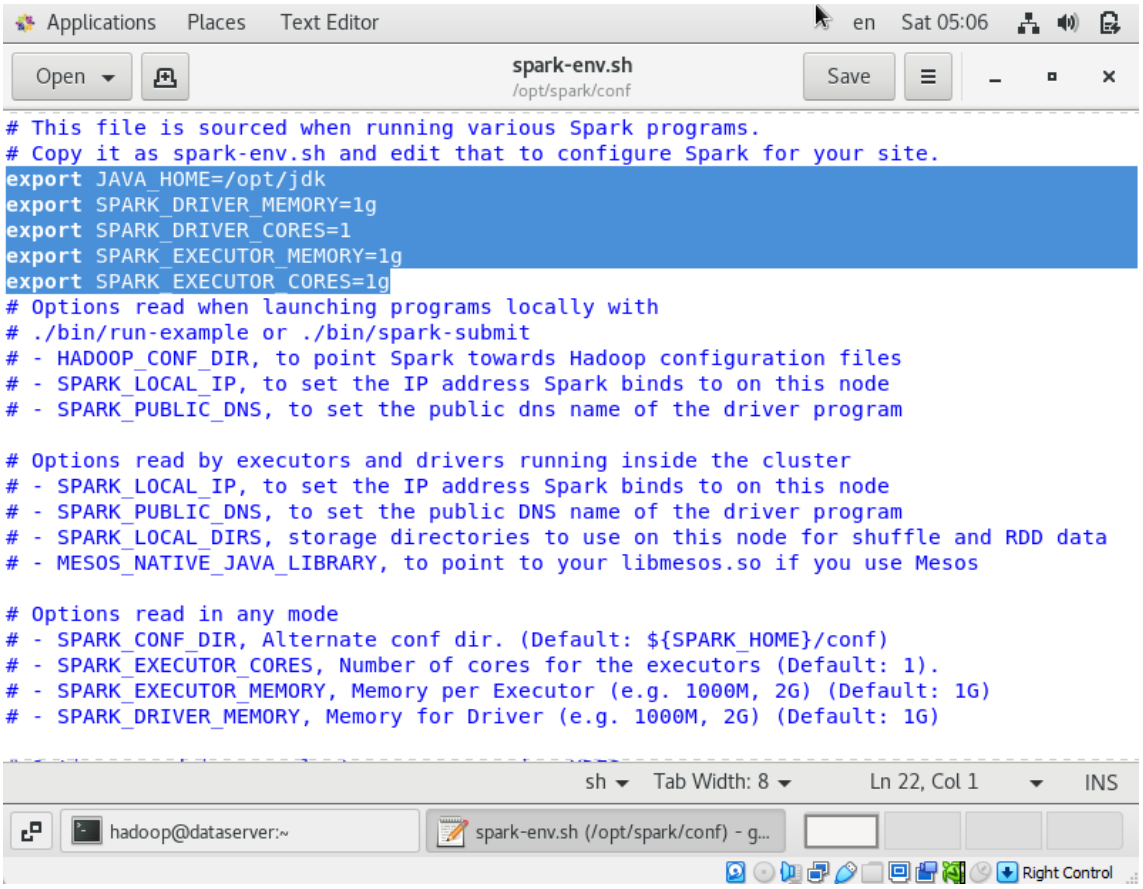
Edite o arquivo $SPARK_HOME/conf/spark-defaults.conf



Acrescente as linhas selecionadas

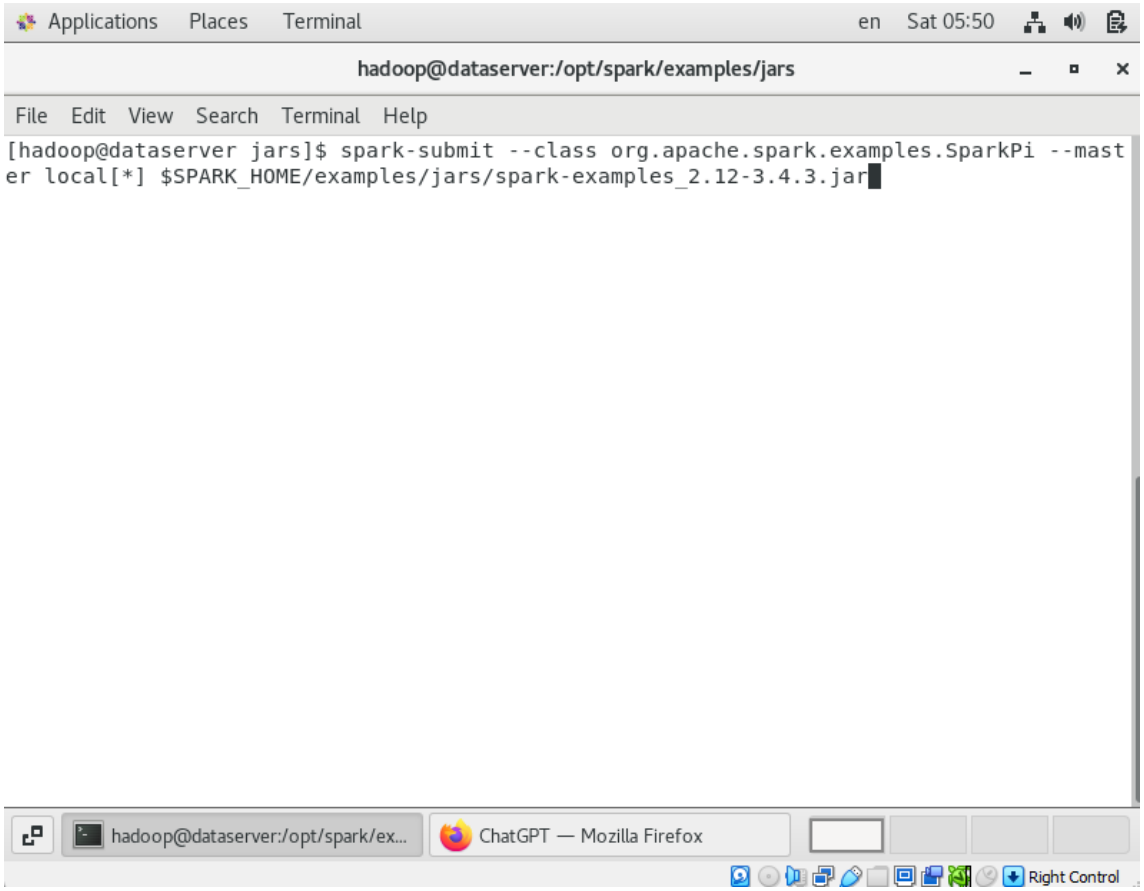


Edite o arquivo

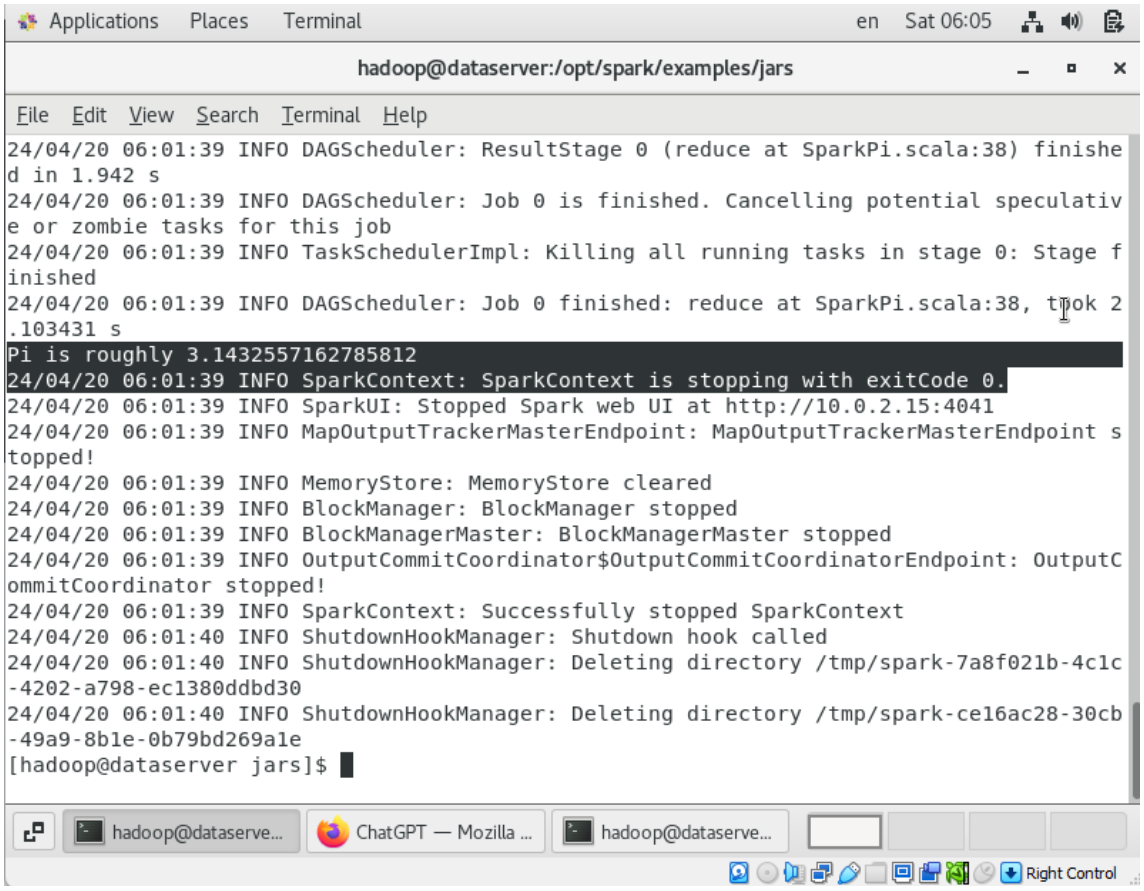


Acrescente as linhas selecionadas.

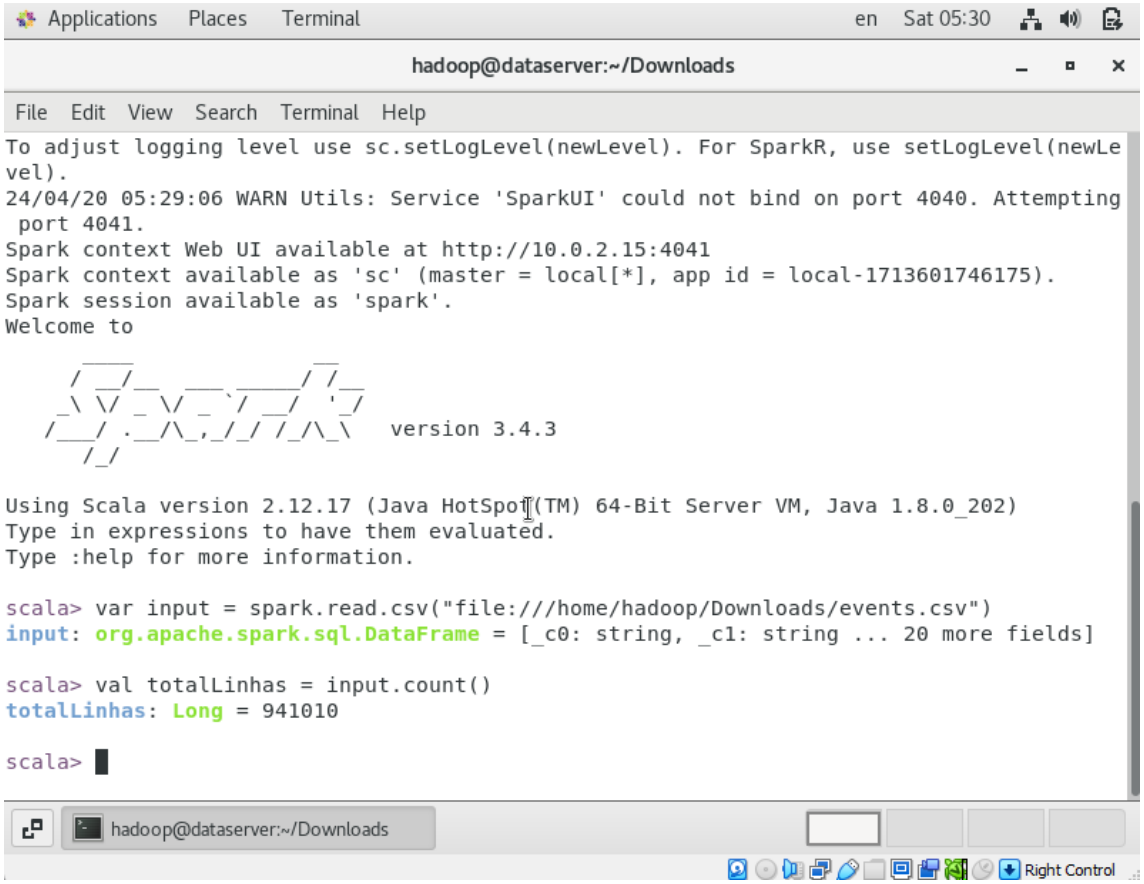
7 Processando Big Data com Spark



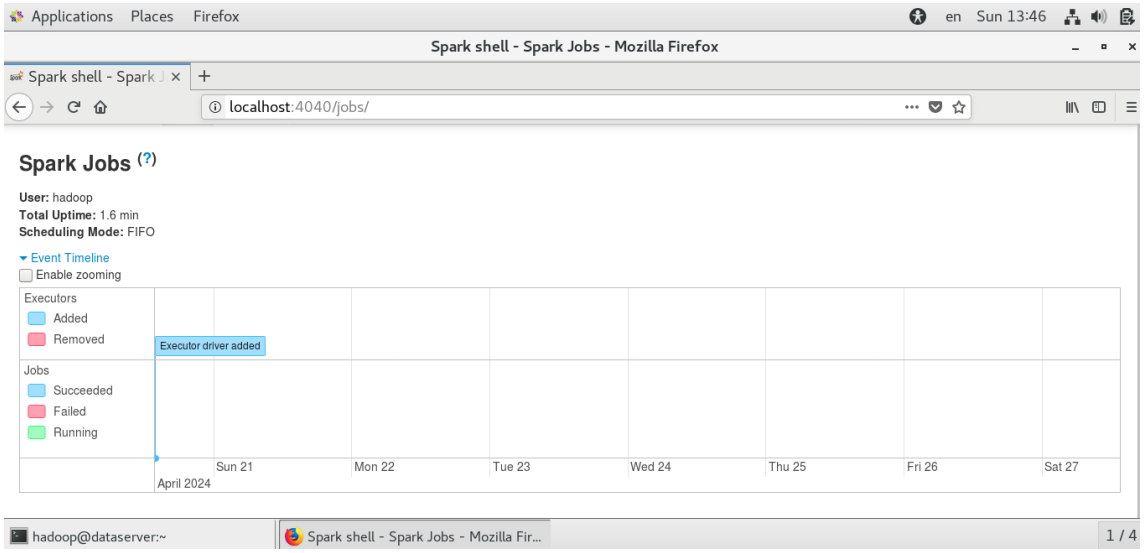
Submeter um job Spark de processamento de dados. Execute um job com Spark-Submit. Esse job pode ser qualquer aplicação de processamento de dados. Neste exemplo, usamos o Spark examples. Após a execução deve ver o SparkContext encerrado com o código de saída 0, isso indica que a execução foi bem sucedida.



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



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!!



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!