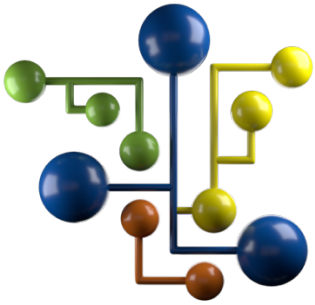


Data Science Academy

www.datascienceacademy.com.br

Engenharia de Dados com Hadoop e Spark

Instalação e Configuração do Ecossistema Hadoop

Sumário

1. Versão

Este documento foi criado pela equipe Data Science Academy e pode ser distribuído livremente, desde que se faça menção à fonte.

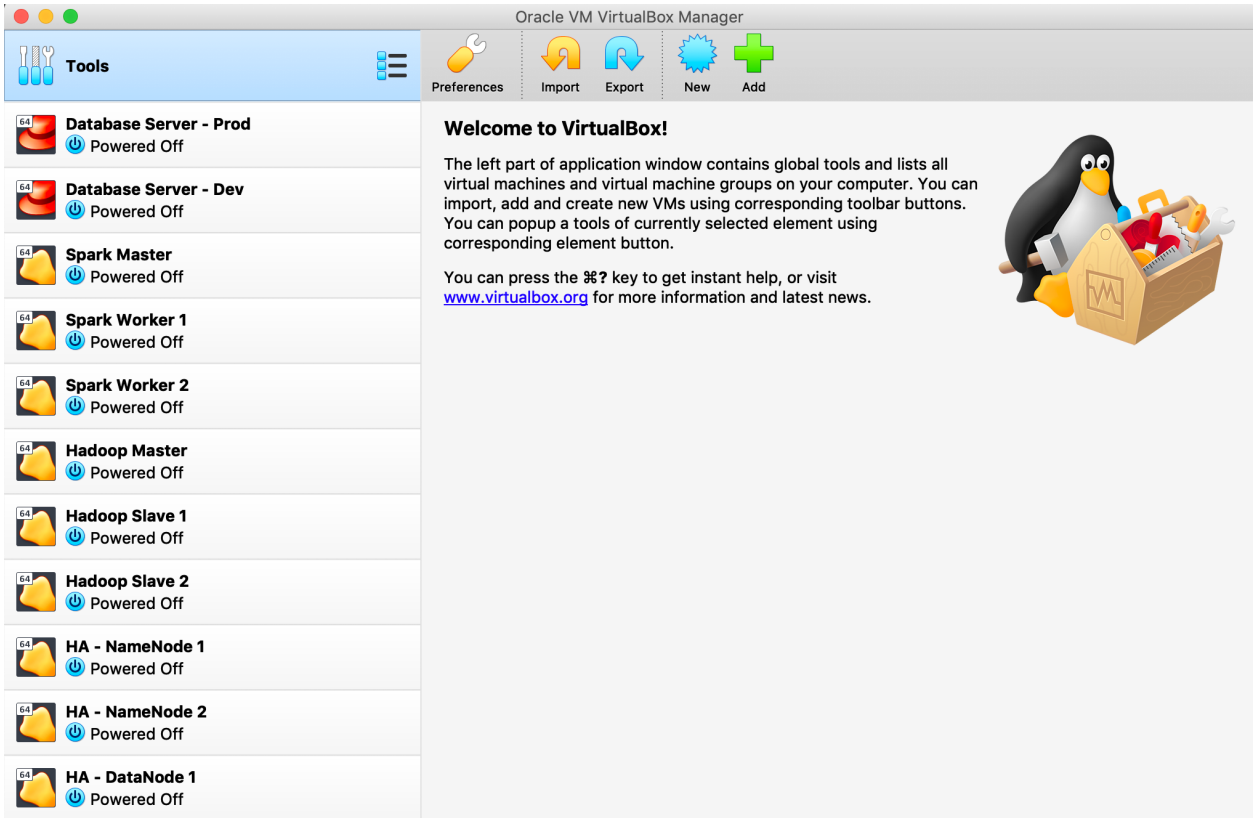
Versão	Ação	Data
1.0	Criação do documento	28/07/2019

2. Configuração do Ambiente

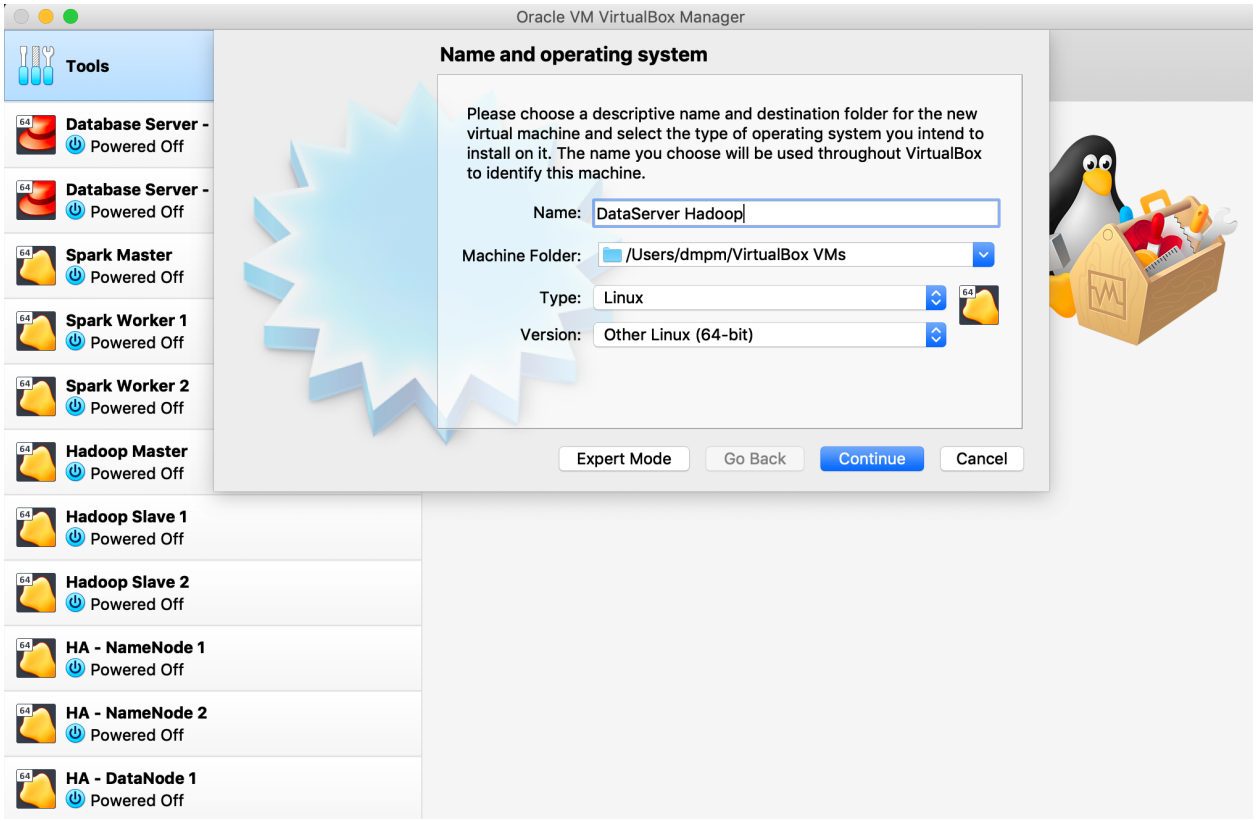
Item	Versão
Virtual Box	6.0.8
Sistema Operacional	CentOS 7.6 (64 bits) ou CentOS 6.8 (32 bits)
Interface Gráfica	Gnome
Firefox Web Browser	60.7
Java	1.8
Apache Hadoop	3.2.0
Apache Zookeeper	3.5.5
Apache Hbase	2.2.0
Apache Hive	3.1.1
Apache Pig	0.17.0
Apache Spark	2.4.3
Apache Sqoop	1.4.7
Apache Flume	1.9.0
Apache Ambari	2.4.1

2.1. Criação da Máquina Virtual no VirtualBox

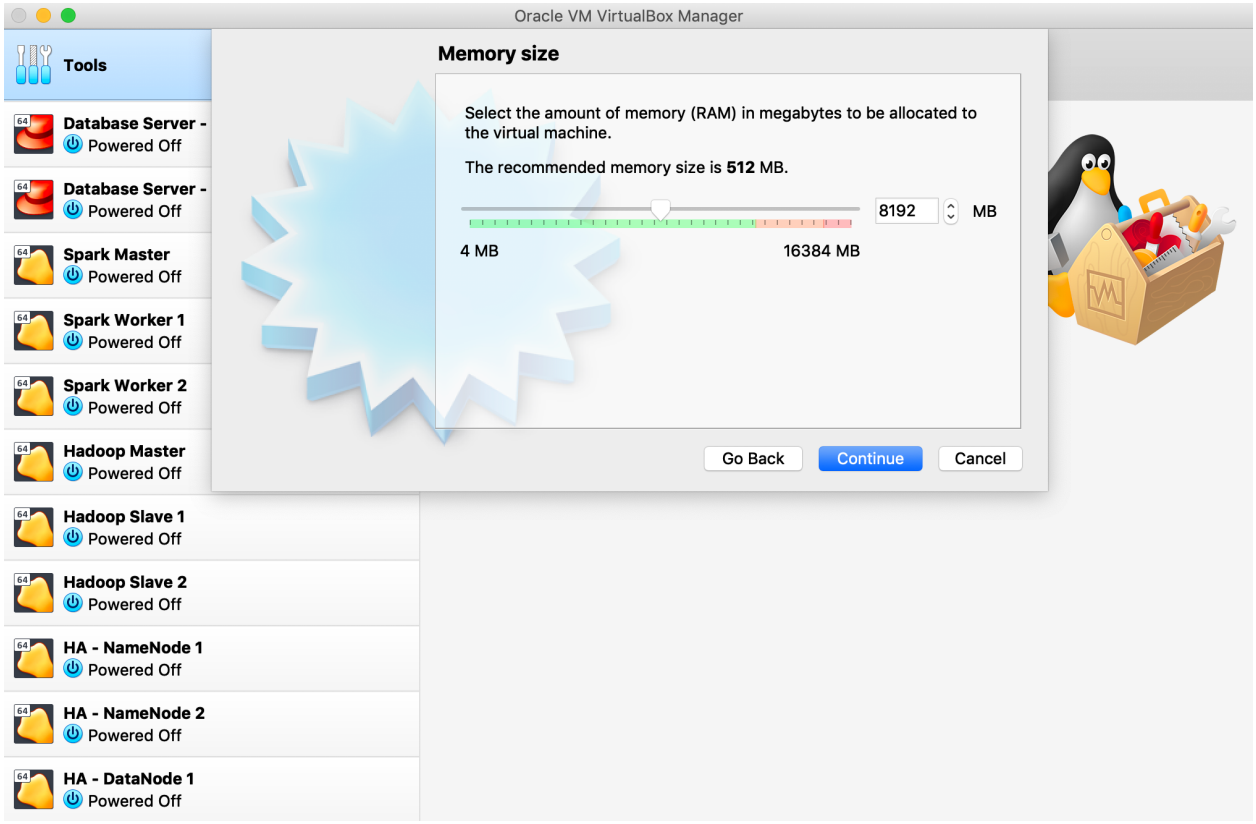
O Oracle VM Virtual Box é gratuito e pode ser baixado em https://www.virtualbox.org. Existem versões disponíveis para Windows, MAC, Linux e Solaris. Aqui utilizaremos a versão 6.0 e o tutorial será o mesmo independente do sistema operacional do seu computador. O uso do VirtualBox não é obrigatório e você pode instalar em uma máquina física se desejar.



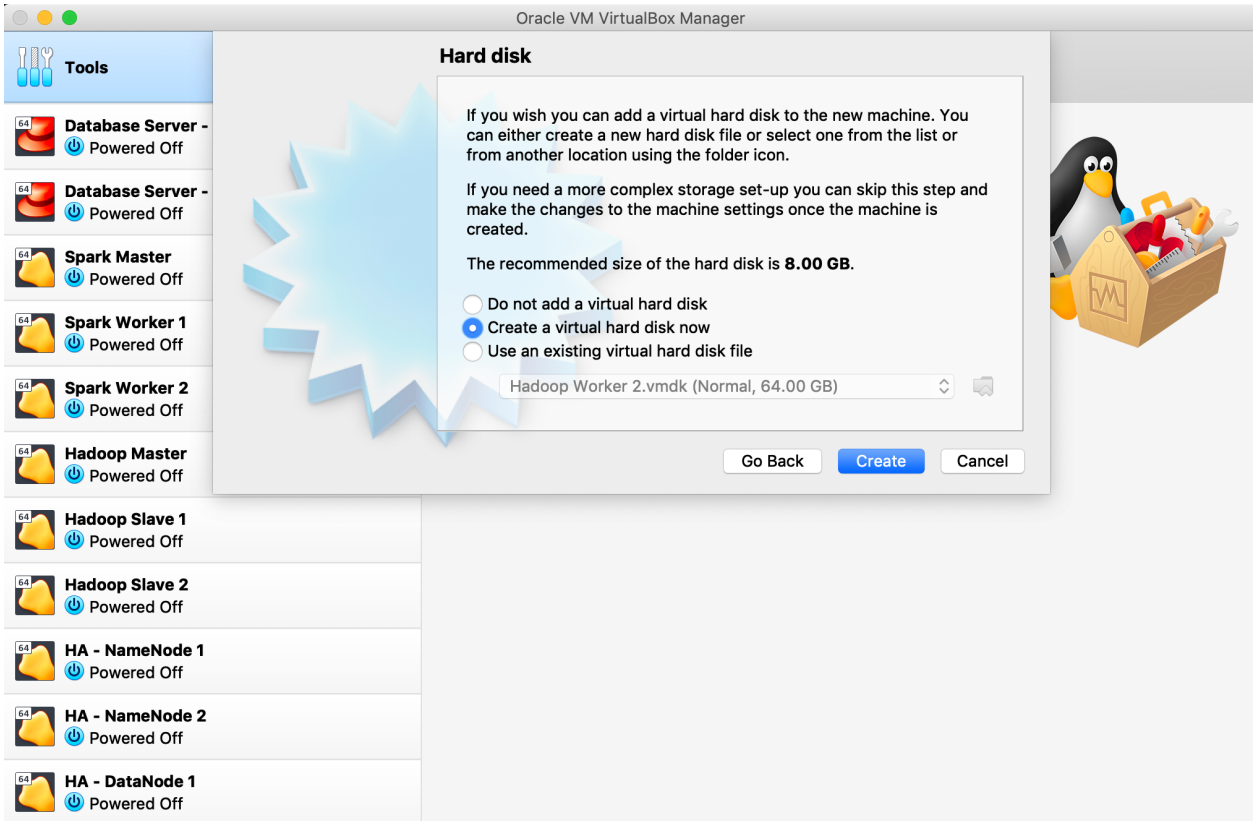
Abrindo o Gerenciador do Oracle Virtual Box



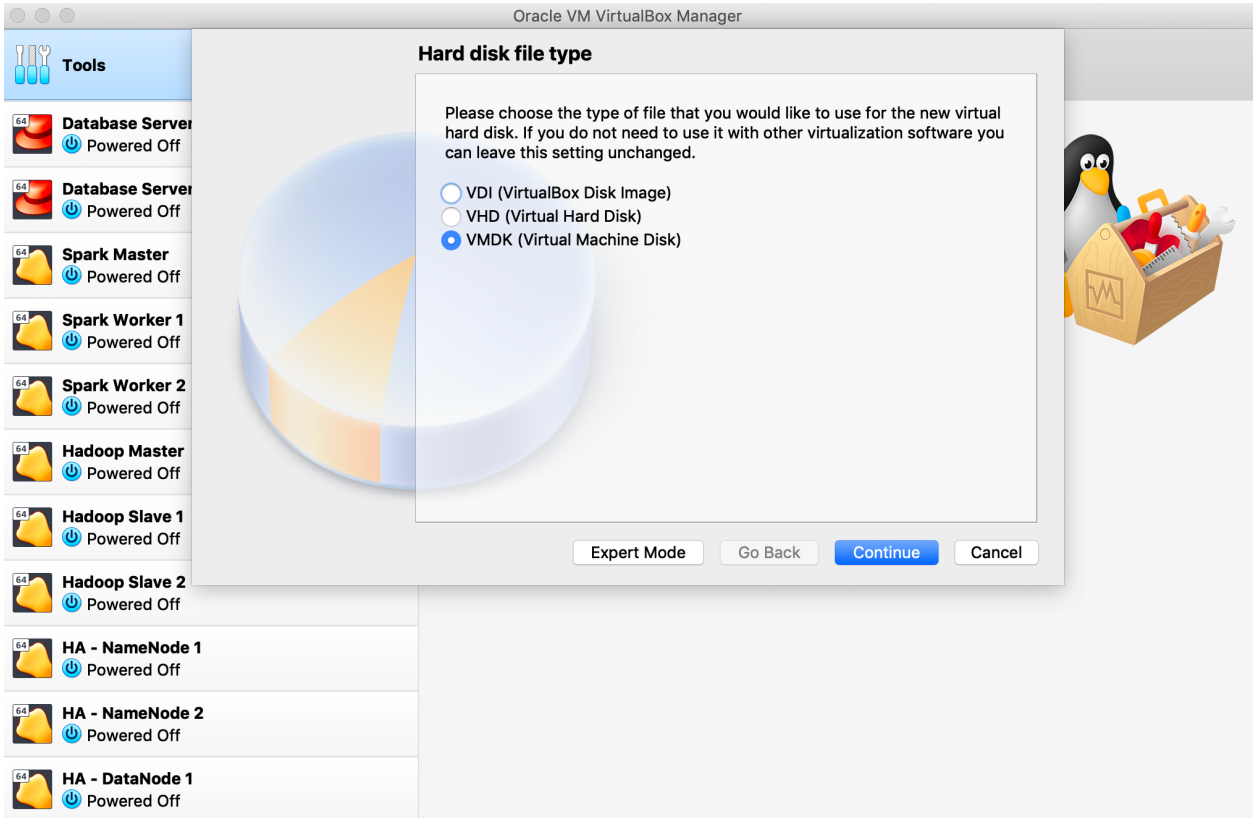
Definindo o nome da máquina virtual e a versão do sistema operacional



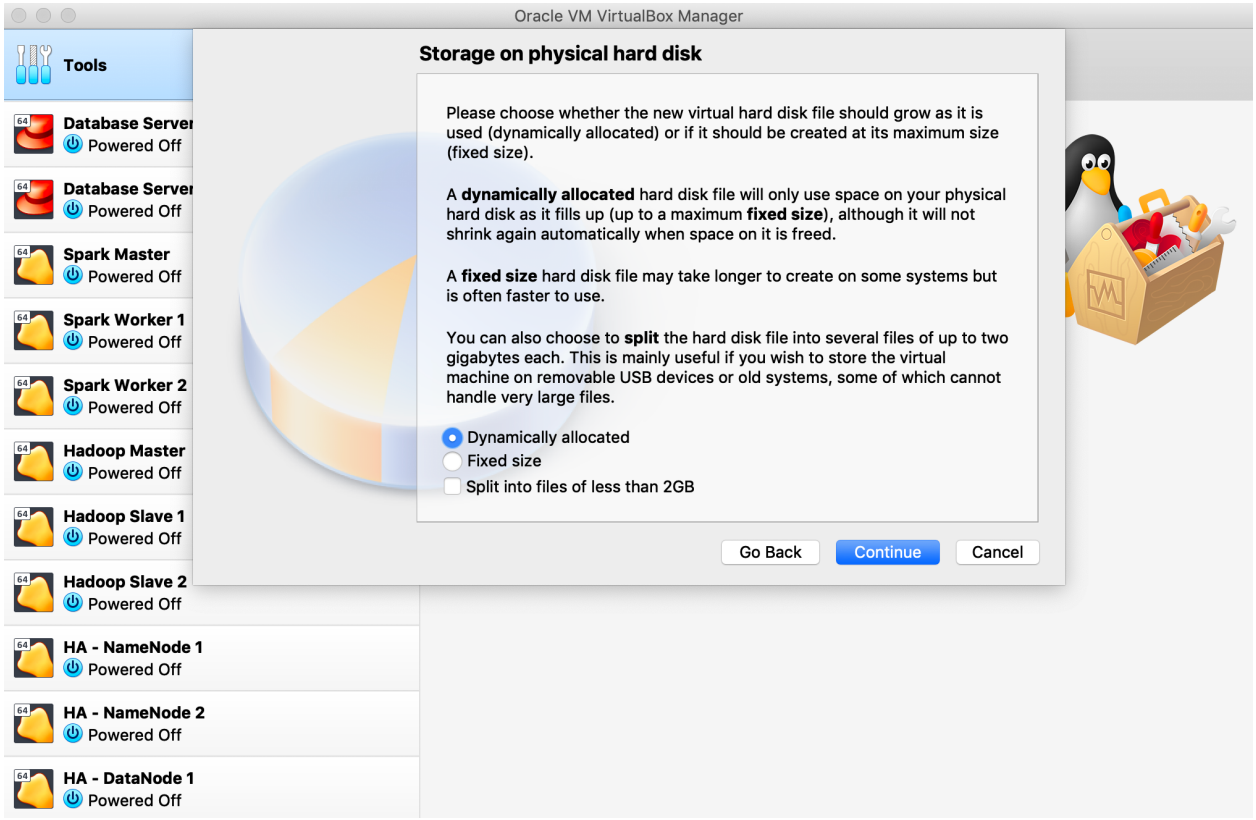
Configure metade da memória física do seu computador para a VM



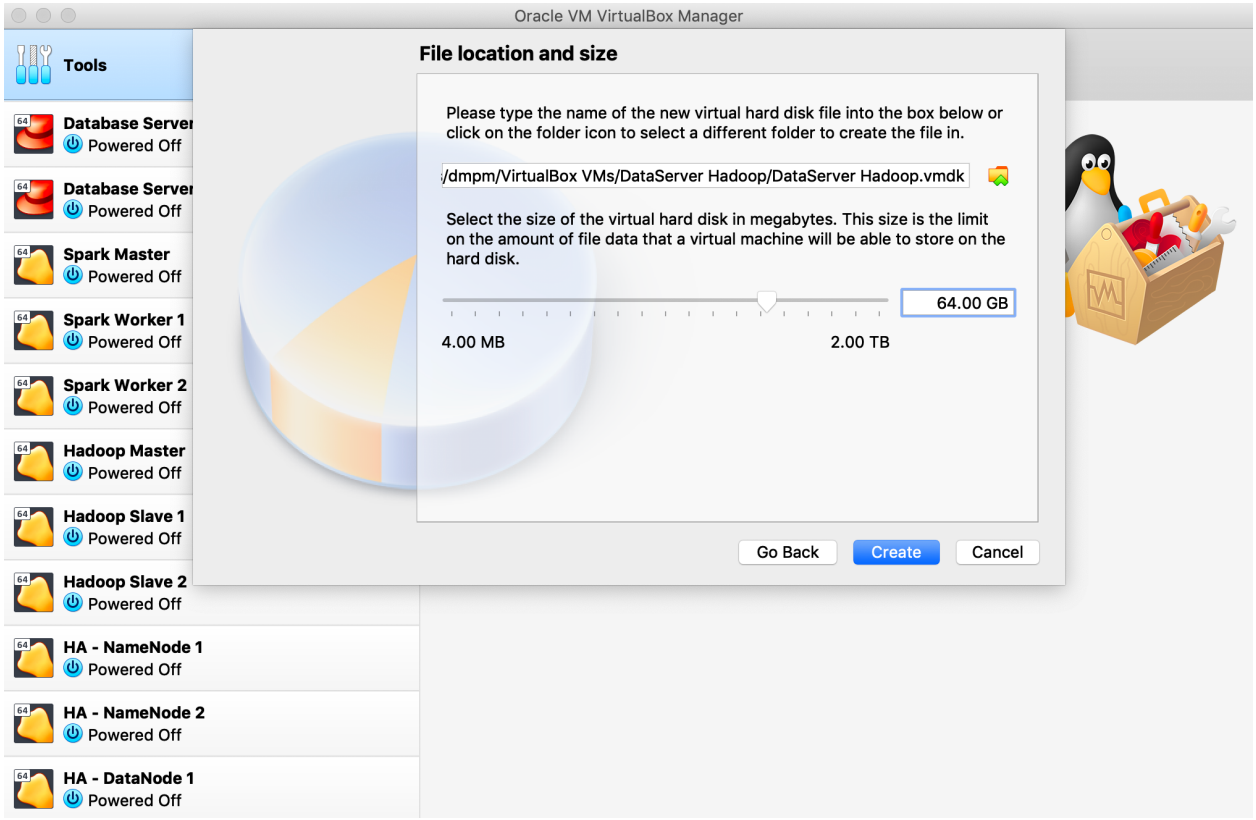
Criar um novo disco rígido virtual



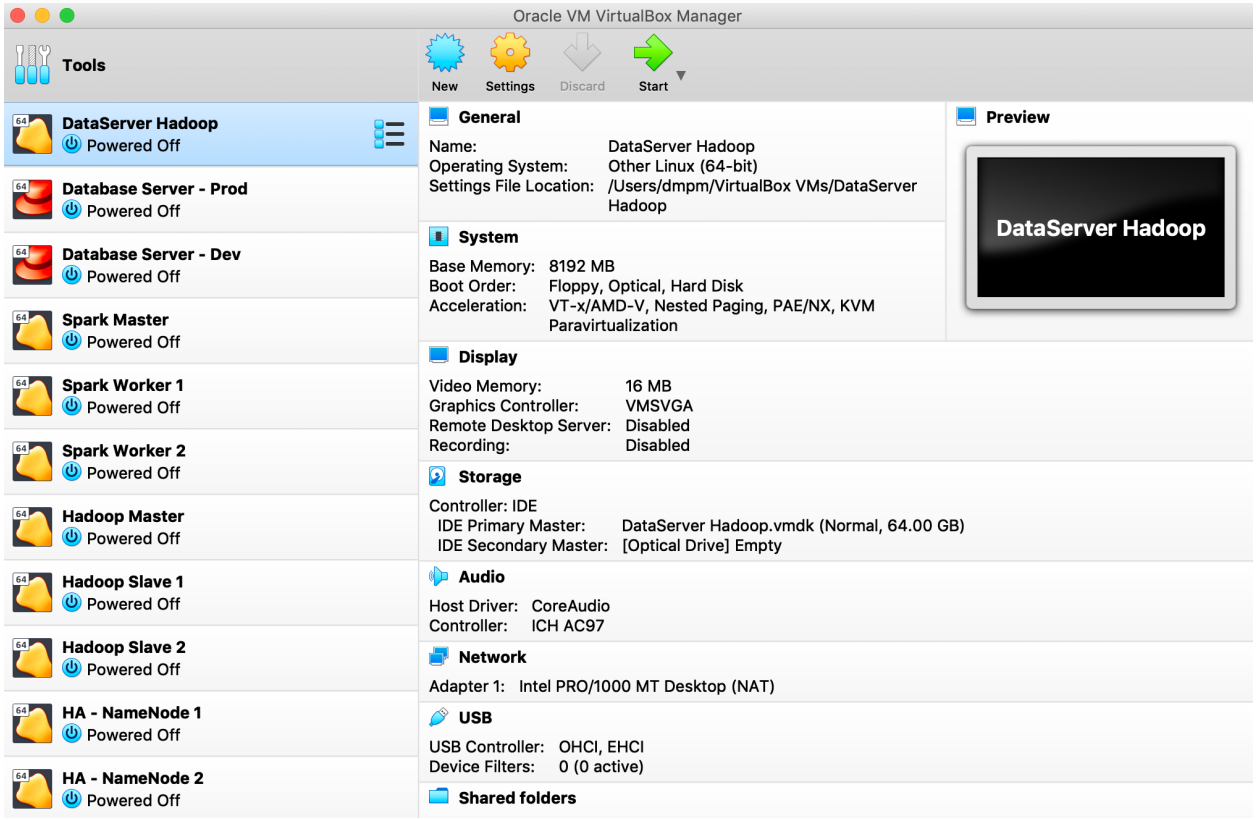
Selecione a opção VMDK



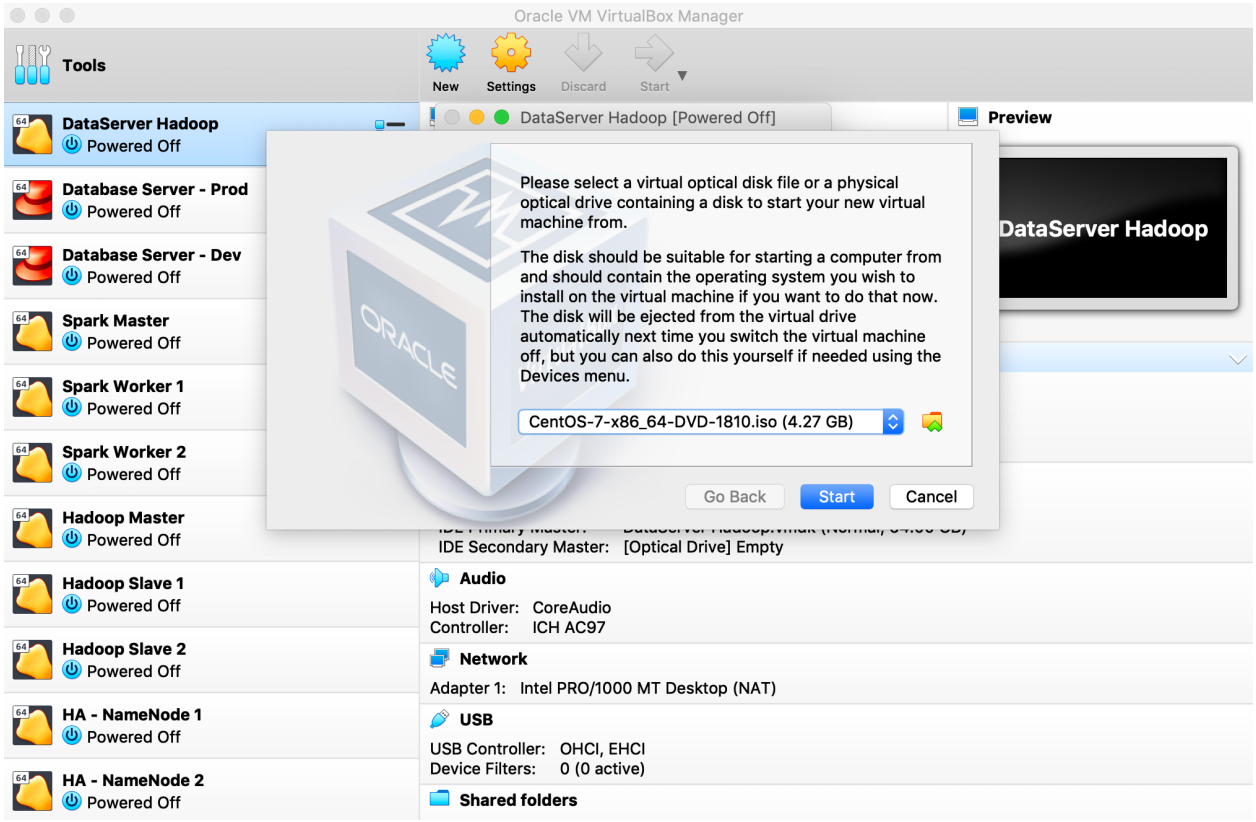
O disco deve ser alocado dinamicamente



Selecione 64 GB para o disco virtual



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



Selecione a mídia de instalação do sistema operacional

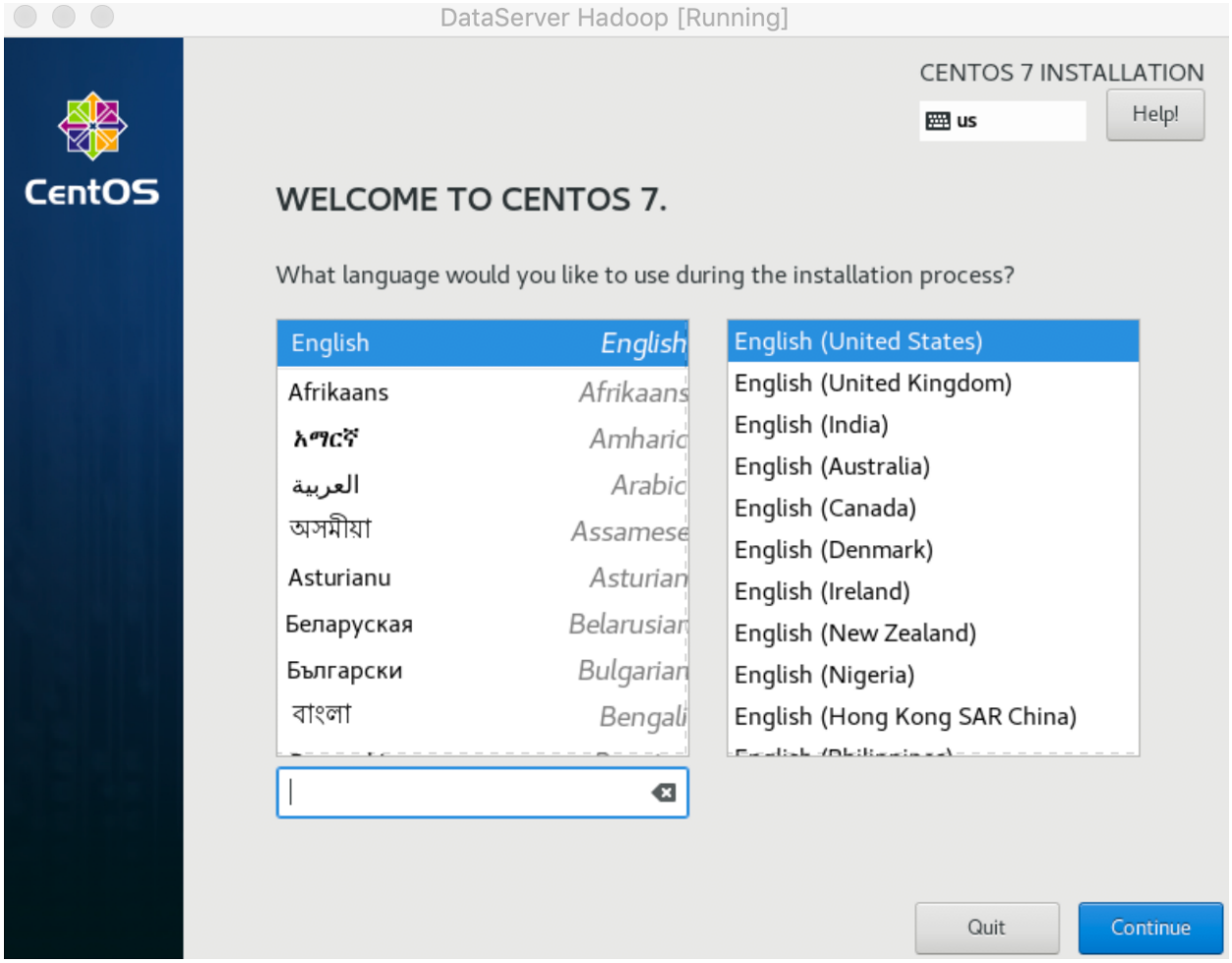
Utilizaremos o CentOS versão 7 64 bits. Caso sua máquina seja 32 bits você deve usar CentOS versão 6.8. Em ambos os casos faça o download do DVD de instalação como imagem .iso

CentOS 64 bits (versão 7):
http://isoredirect.centos.org/centos/7/isos/x86_64/

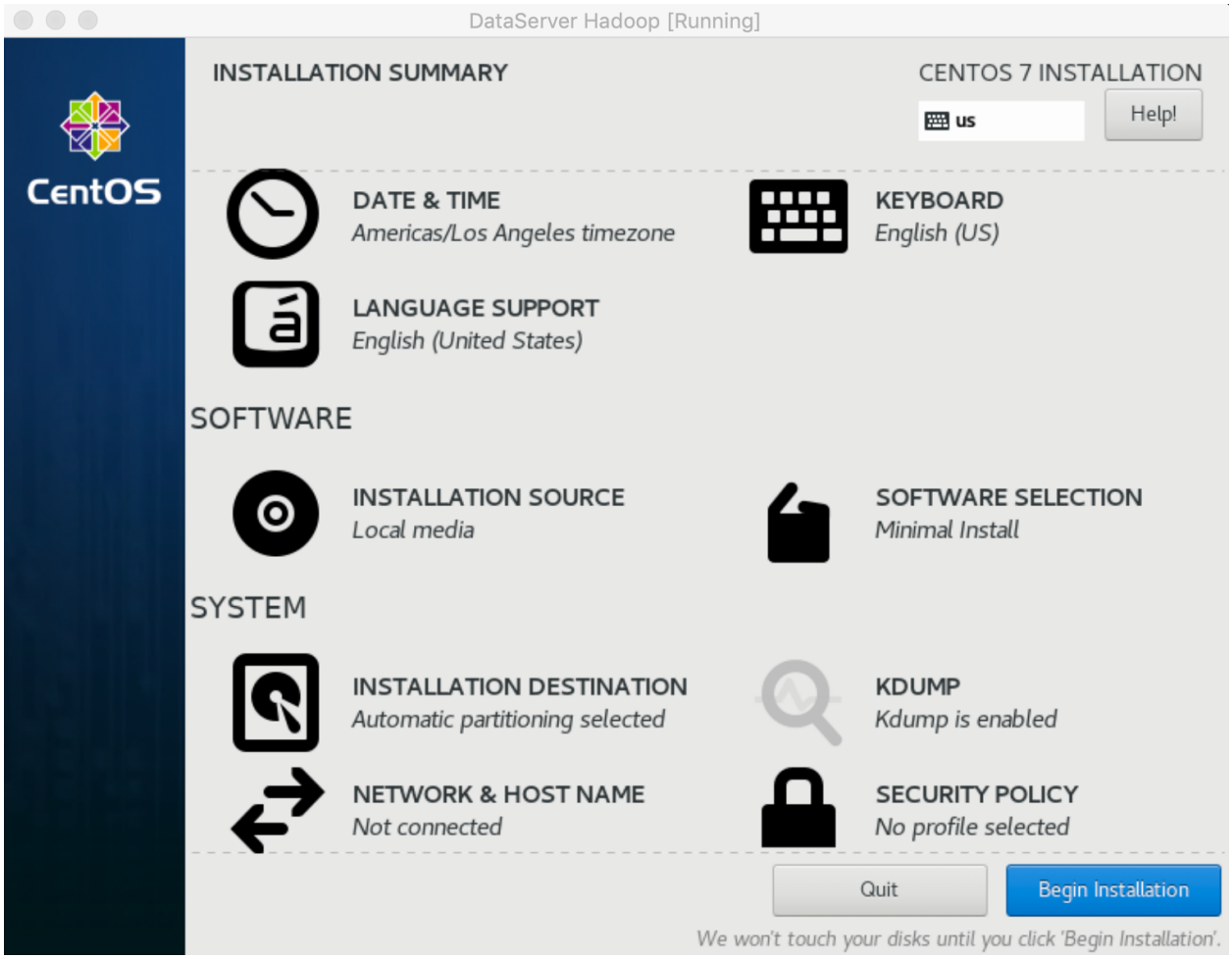
CentOS 32 bits (versão 6.8):
http://centos.mirror.netelligent.ca/centos/

2.2. Instalação do Sistema Operacional

Selecione a opção de Instalação do Sistema Operacional CentOS 7



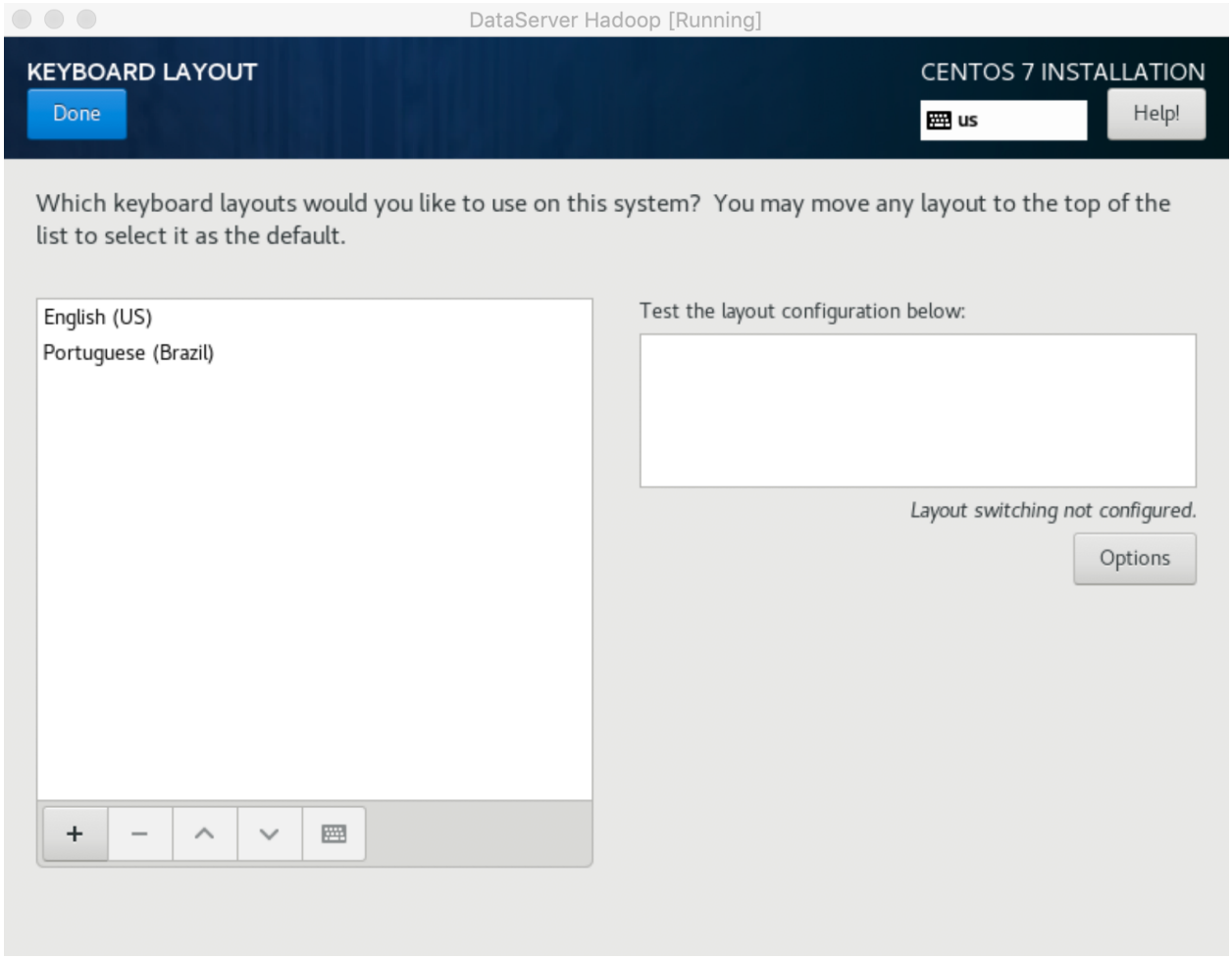
Seleção do idioma usado na instalação



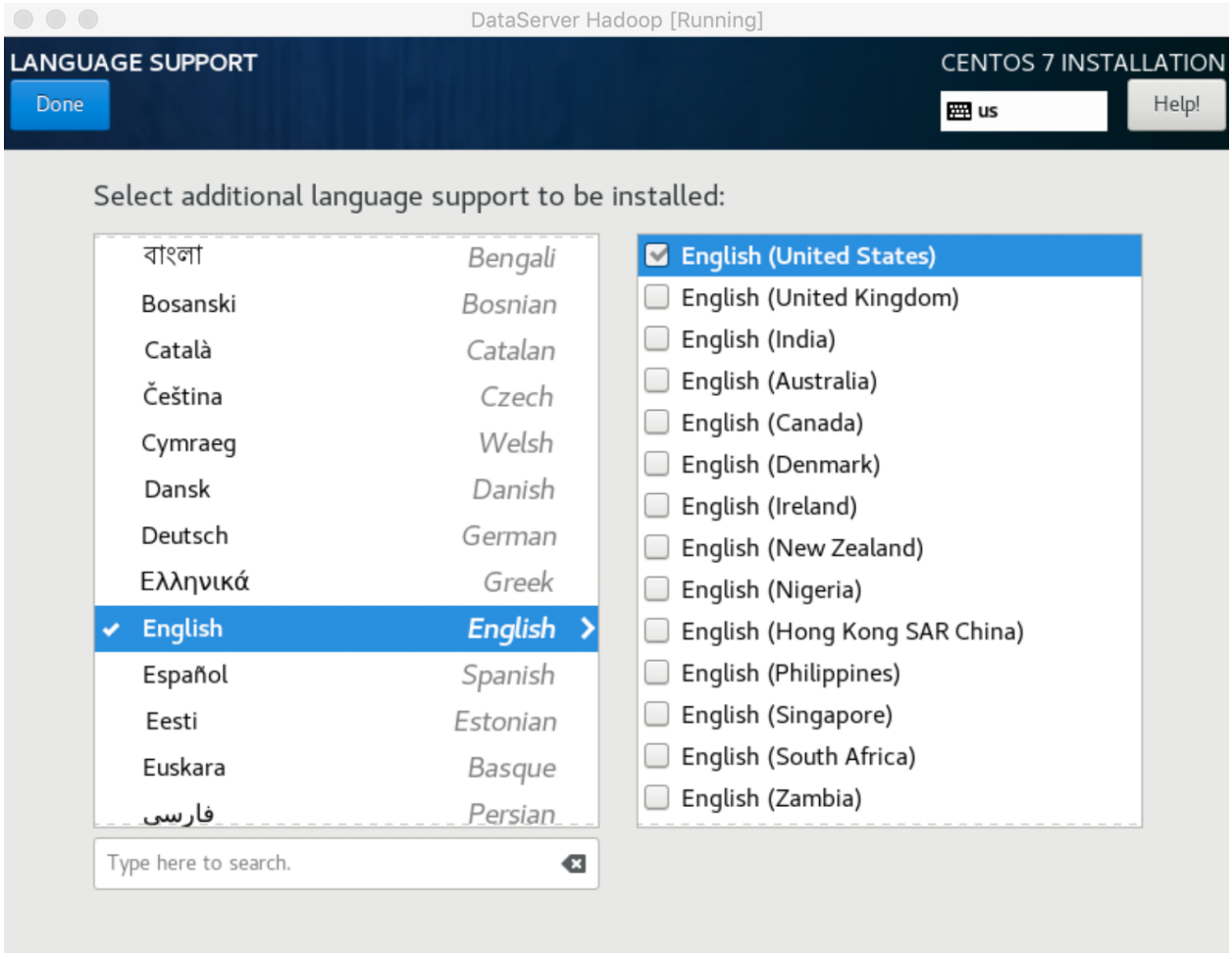
Opções de configuração


DataServer Hadoop [Running]
DATE & TIME
Done
CENTOS 7 INSTALLATION
us
Help!
Region: Americas
City: Sao Paulo
Network Time
OFF
23:08 PM
24-hour
AM/PM
06 / 21 / 2019
You need to set up networking first if you want to use NTP

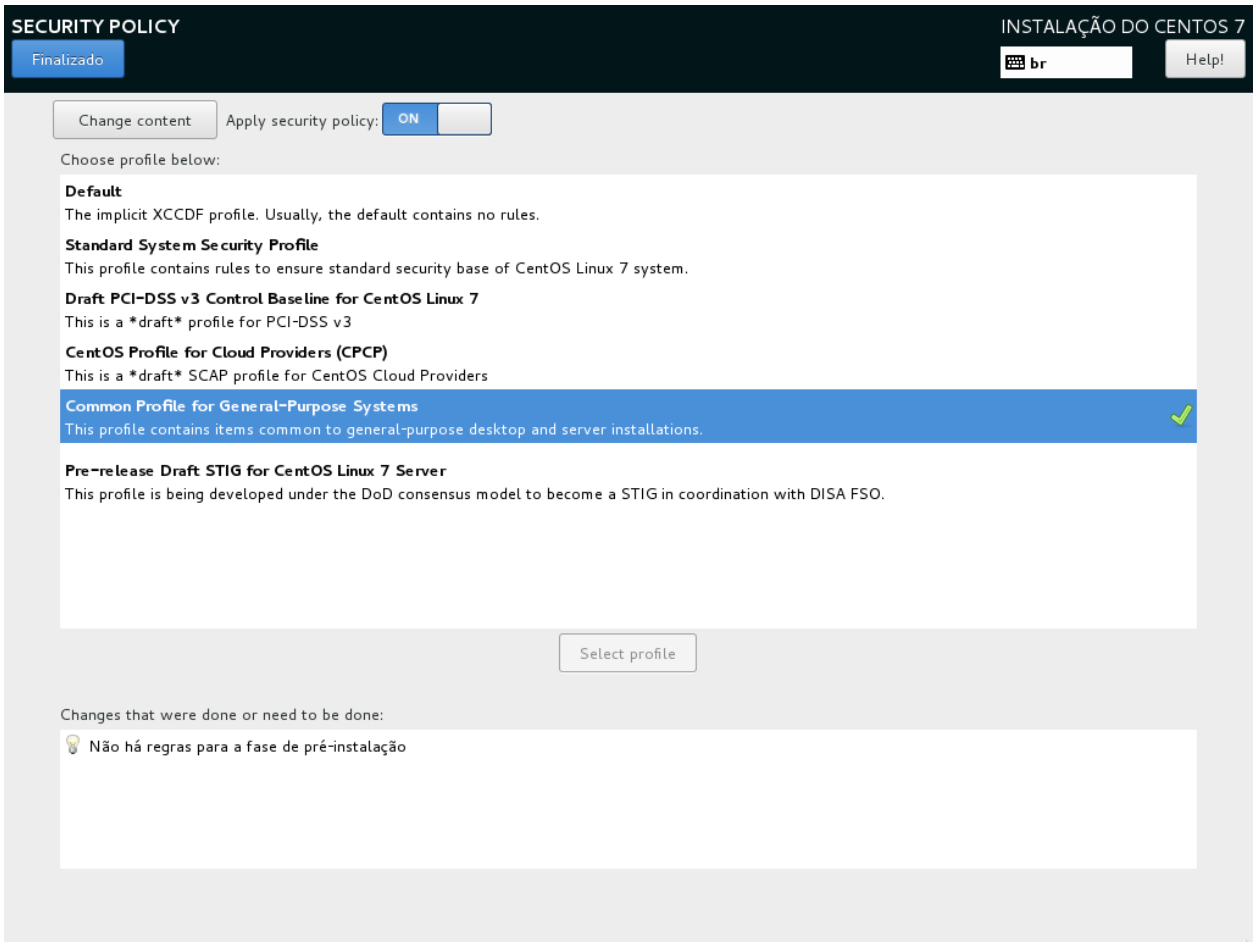
Timezone



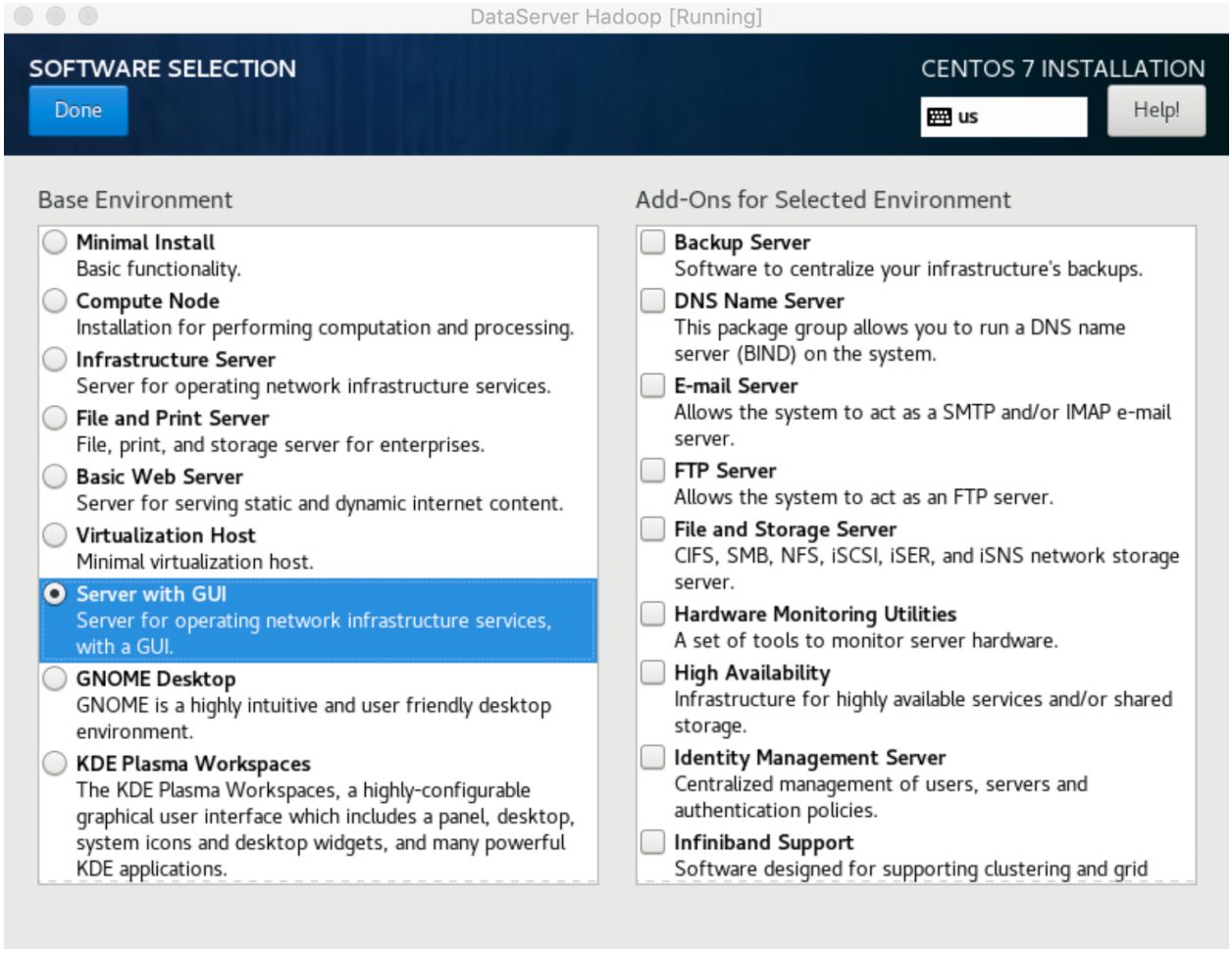
Layout do teclado



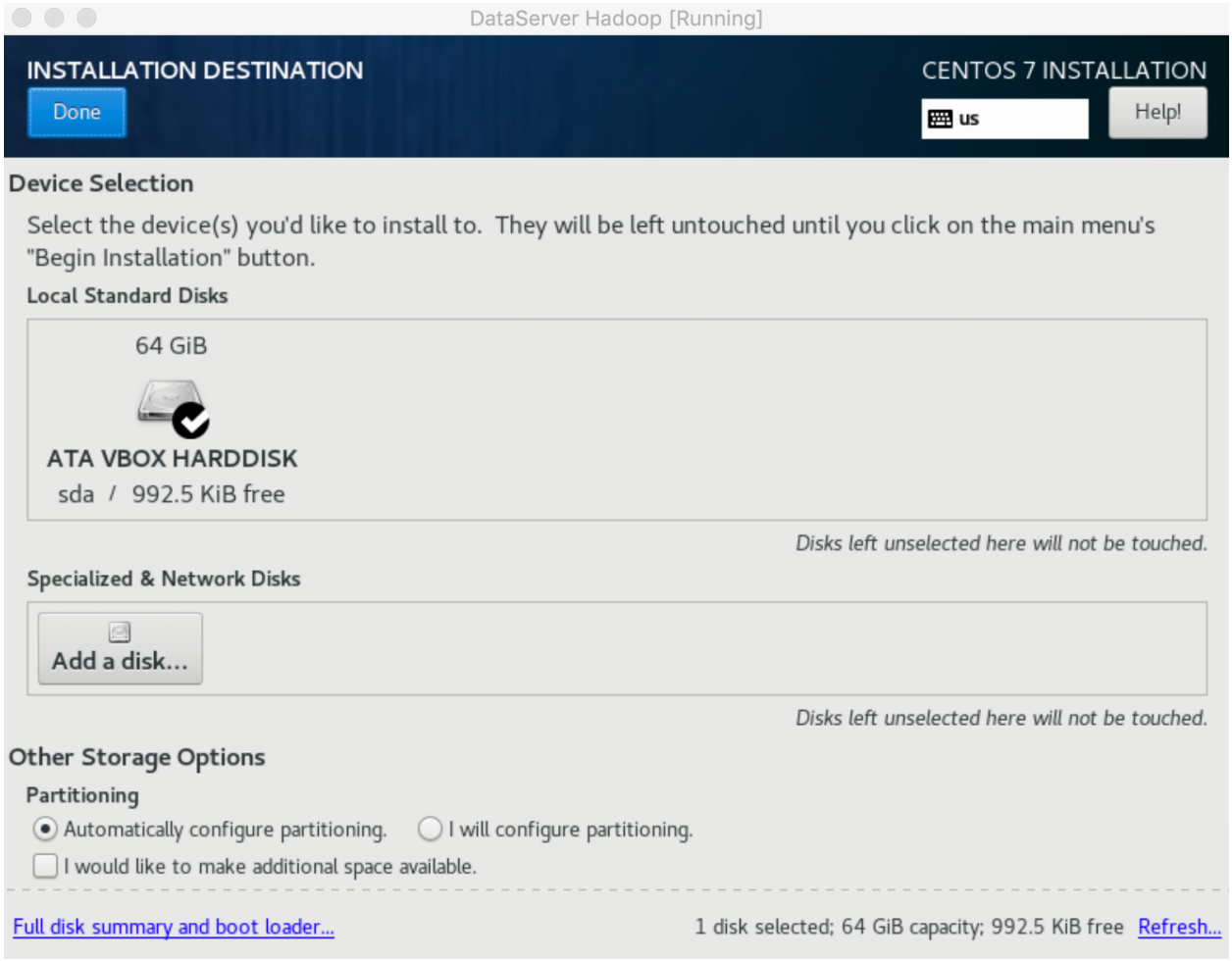
Idioma do sistema operacional



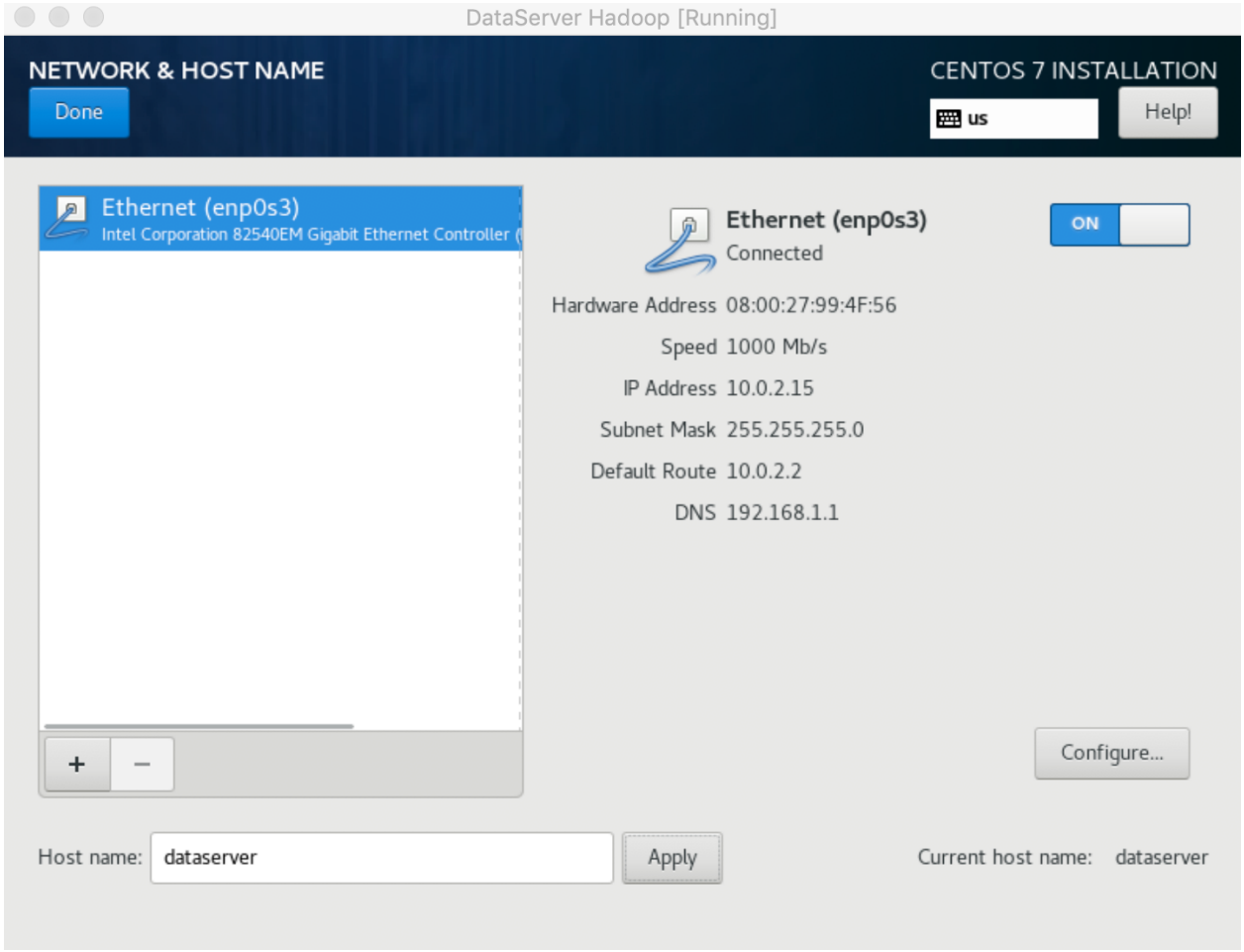
Política de segurança padrão



Selecione a opção Server with GUI

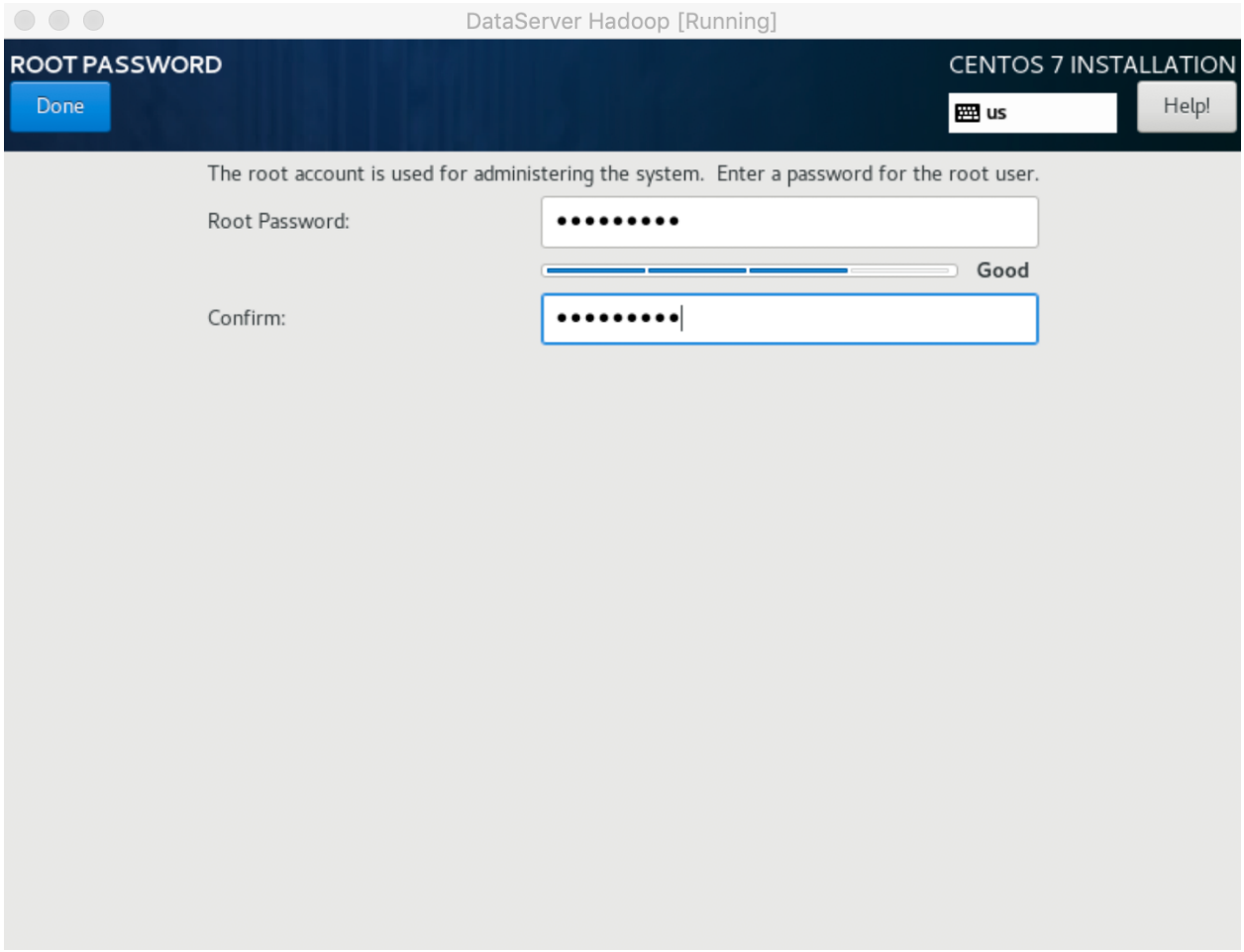

DataServer Hadoop [Running]
INSTALLATION DESTINATION
Done
CENTOS 7 INSTALLATION
us
Help!
Device Selection
Select the device(s) you'd like to install to. They will be left untouched until you click on the main menu's "Begin Installation" button.
Local Standard Disks
64 GiB
ATA VBOX HARDDISK
sda / 992.5 KiB 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; 992.5 KiB free
Refresh...

Disco

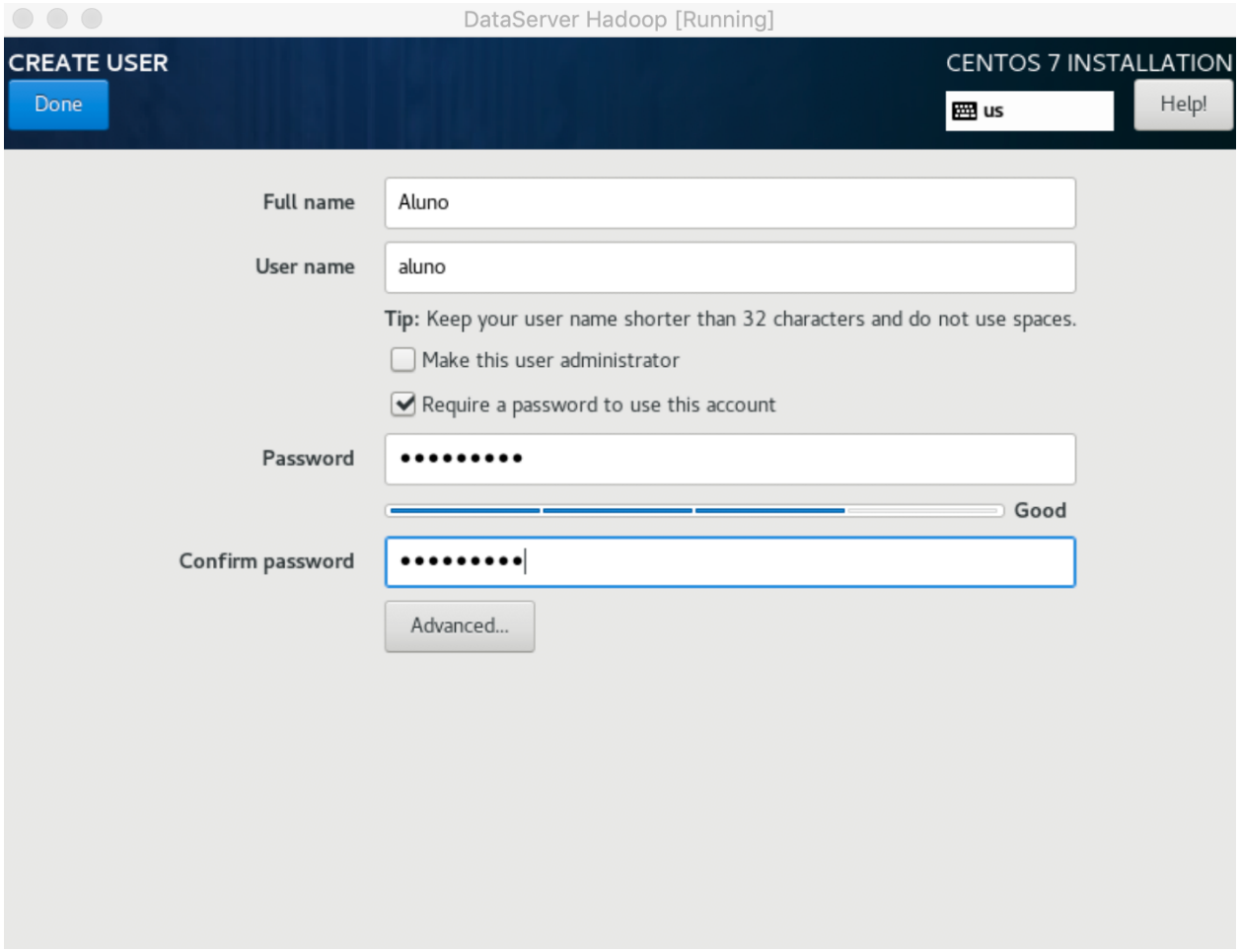


Configuração de rede e nome do servidor – **dataserver –** Clique em Apply
Certifique-se de habilitar a opção de ativar a Ethernet (botão on)

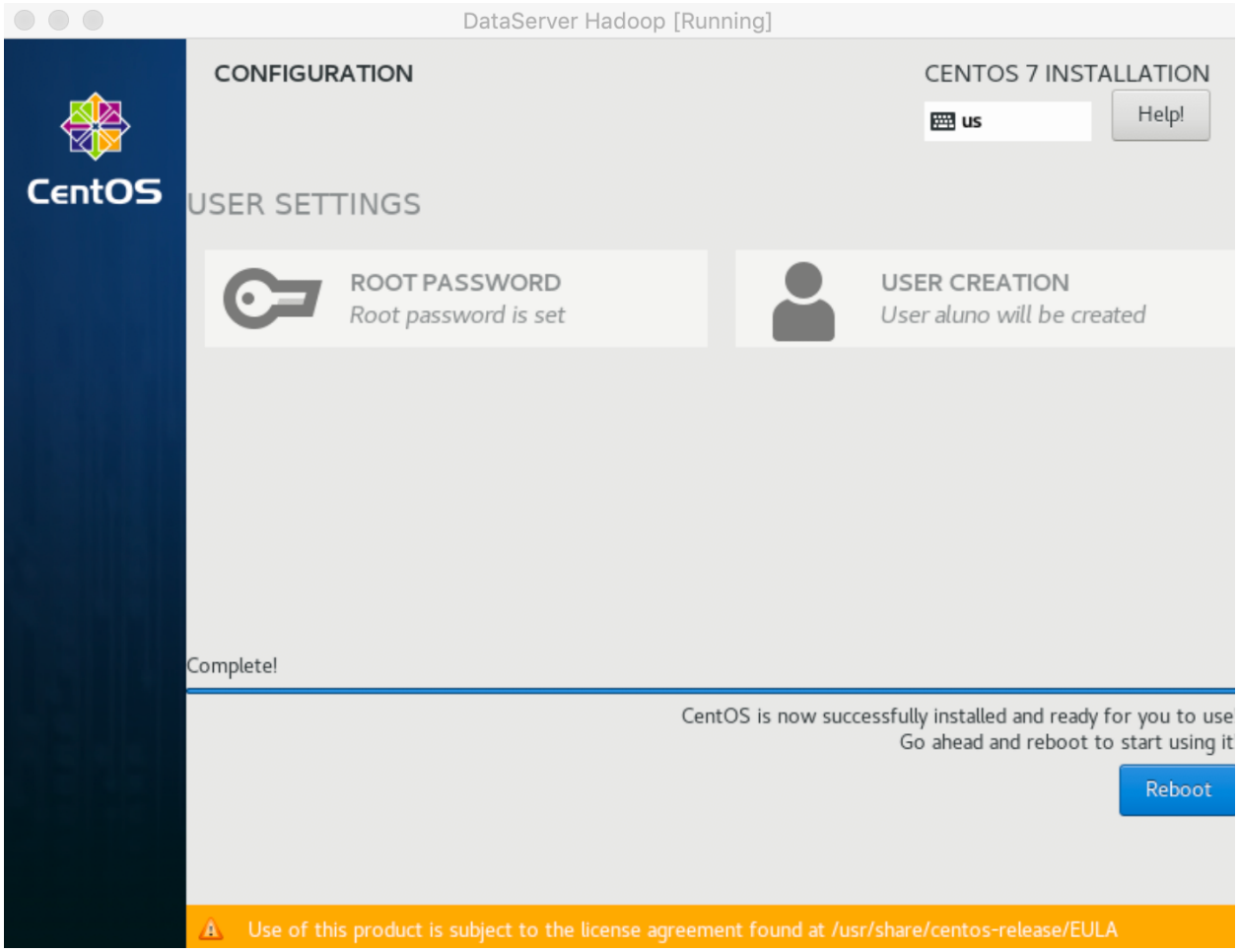
Definir senha do root e criar um novo usuário



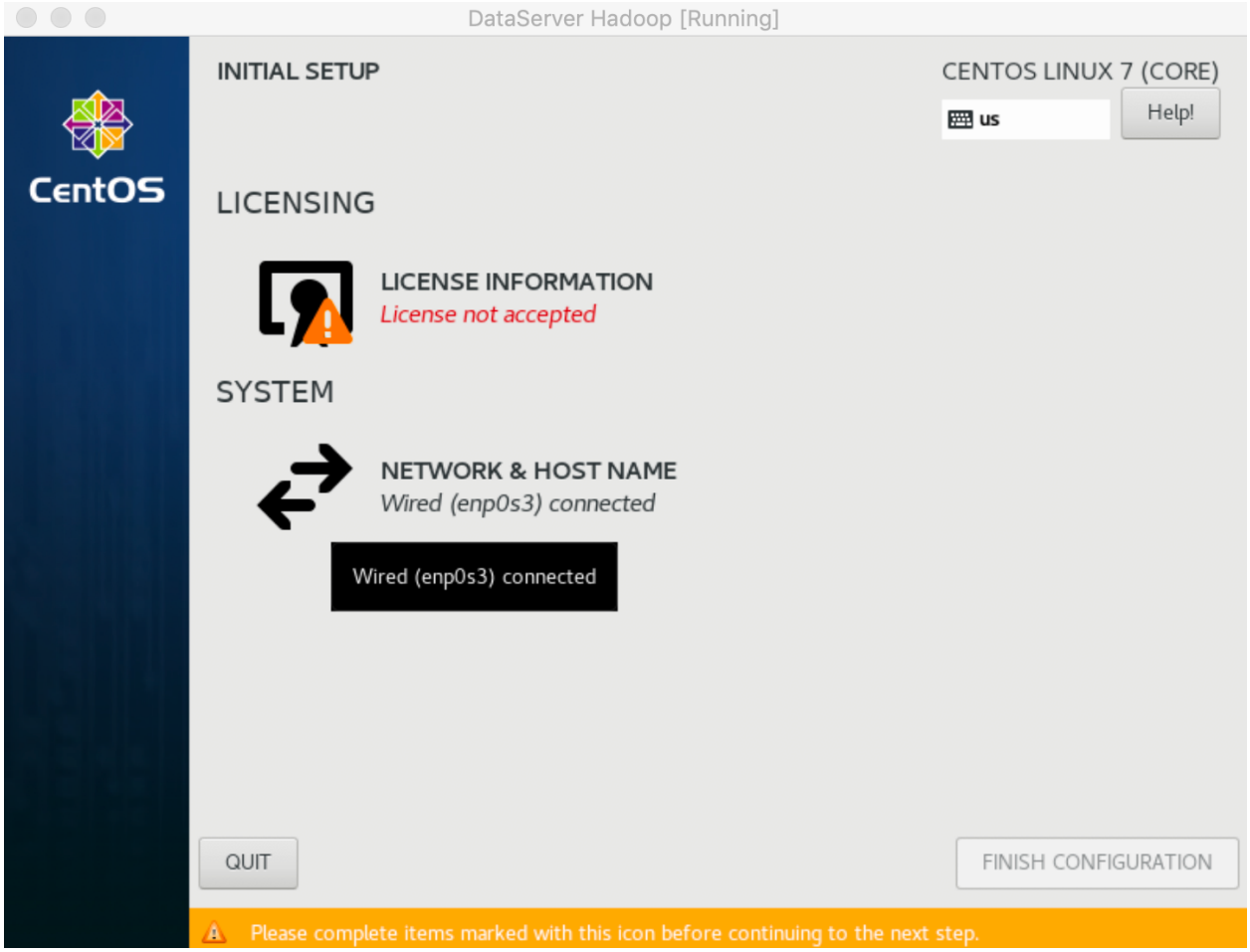
Definir senha do root – usuário administrador
Senha: **dsahadoop**



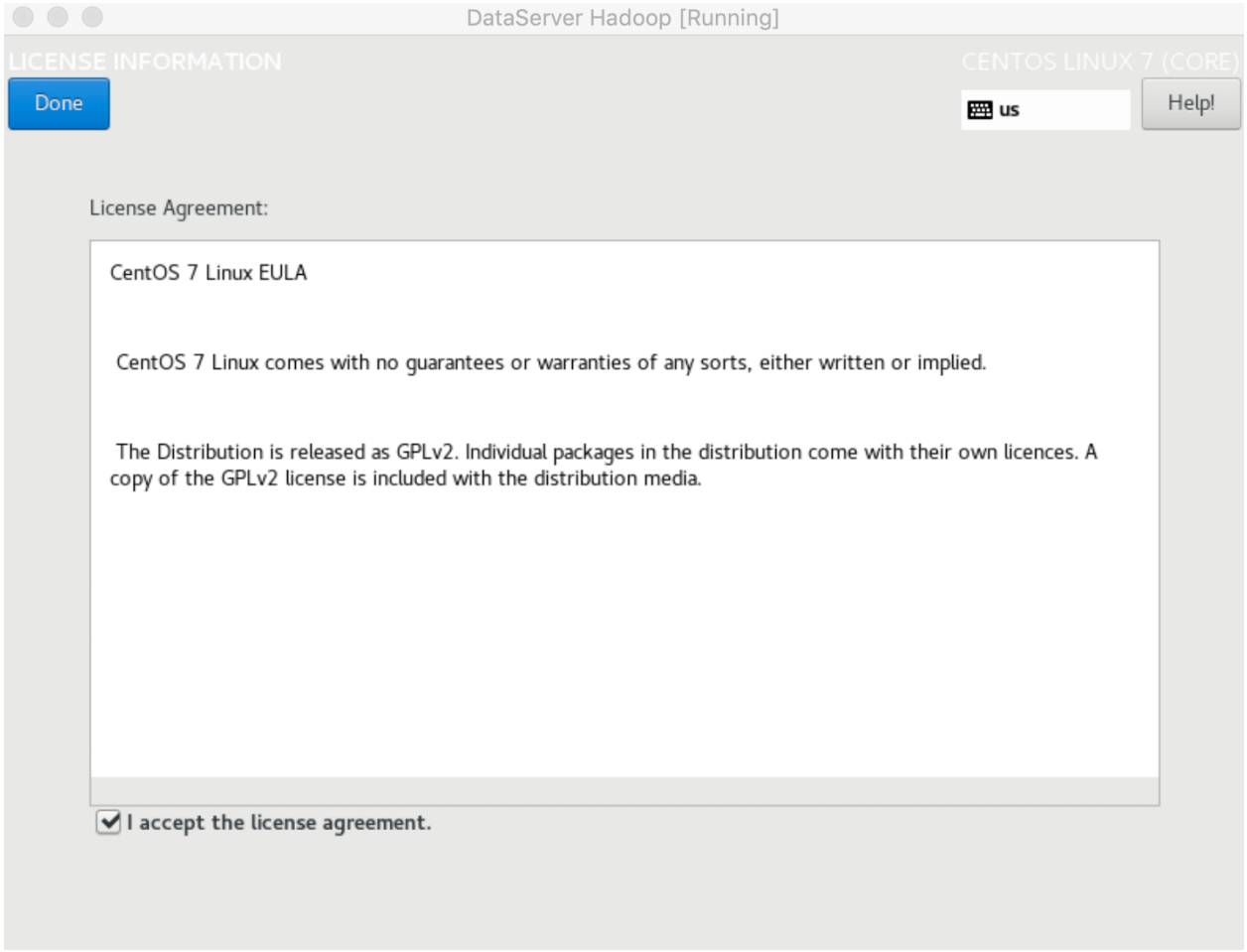
Criação de um usuário – Aluno
(username: aluno, senha: **dsahadoop**)



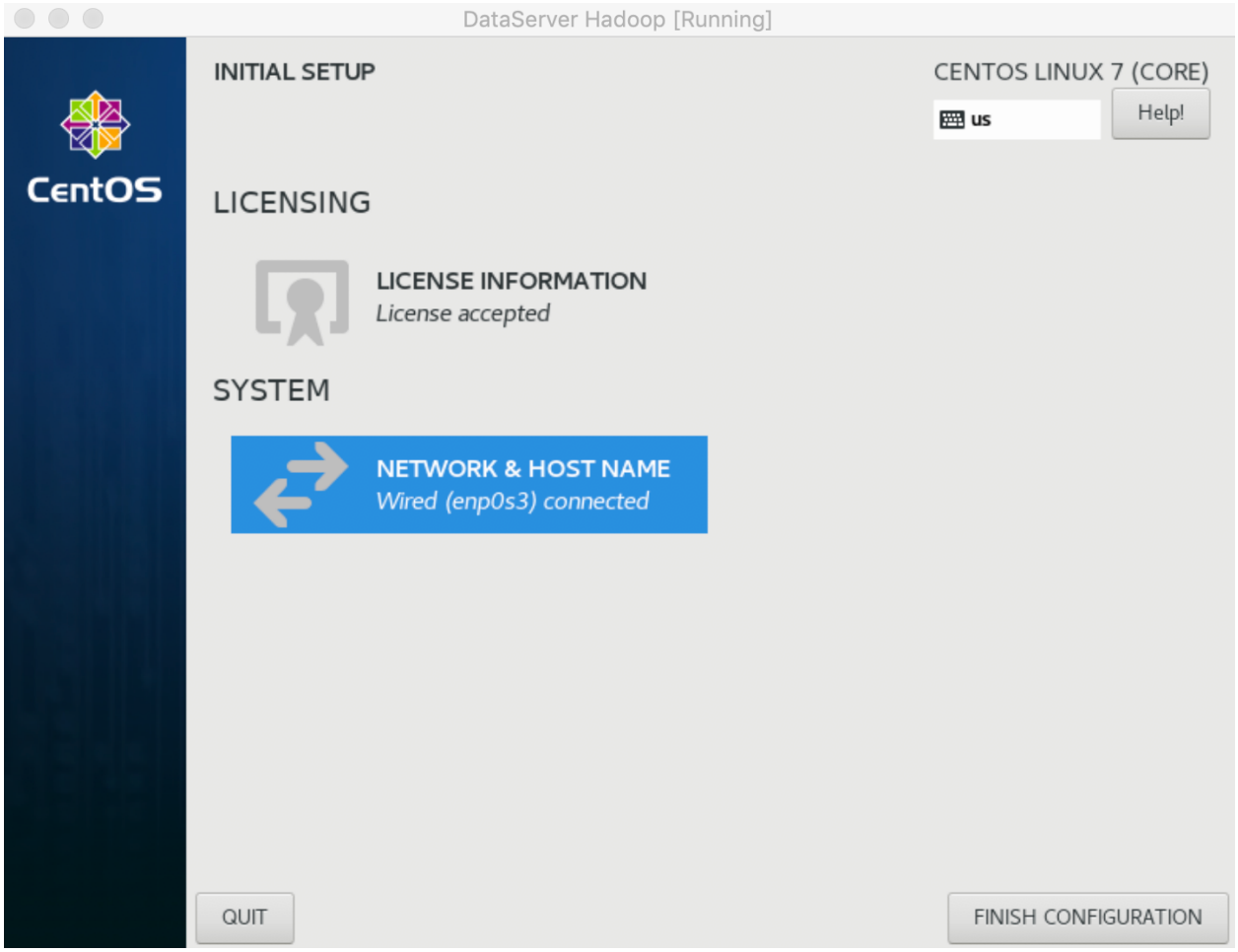
Conclusão da instalação. Clique no botão Reboot.



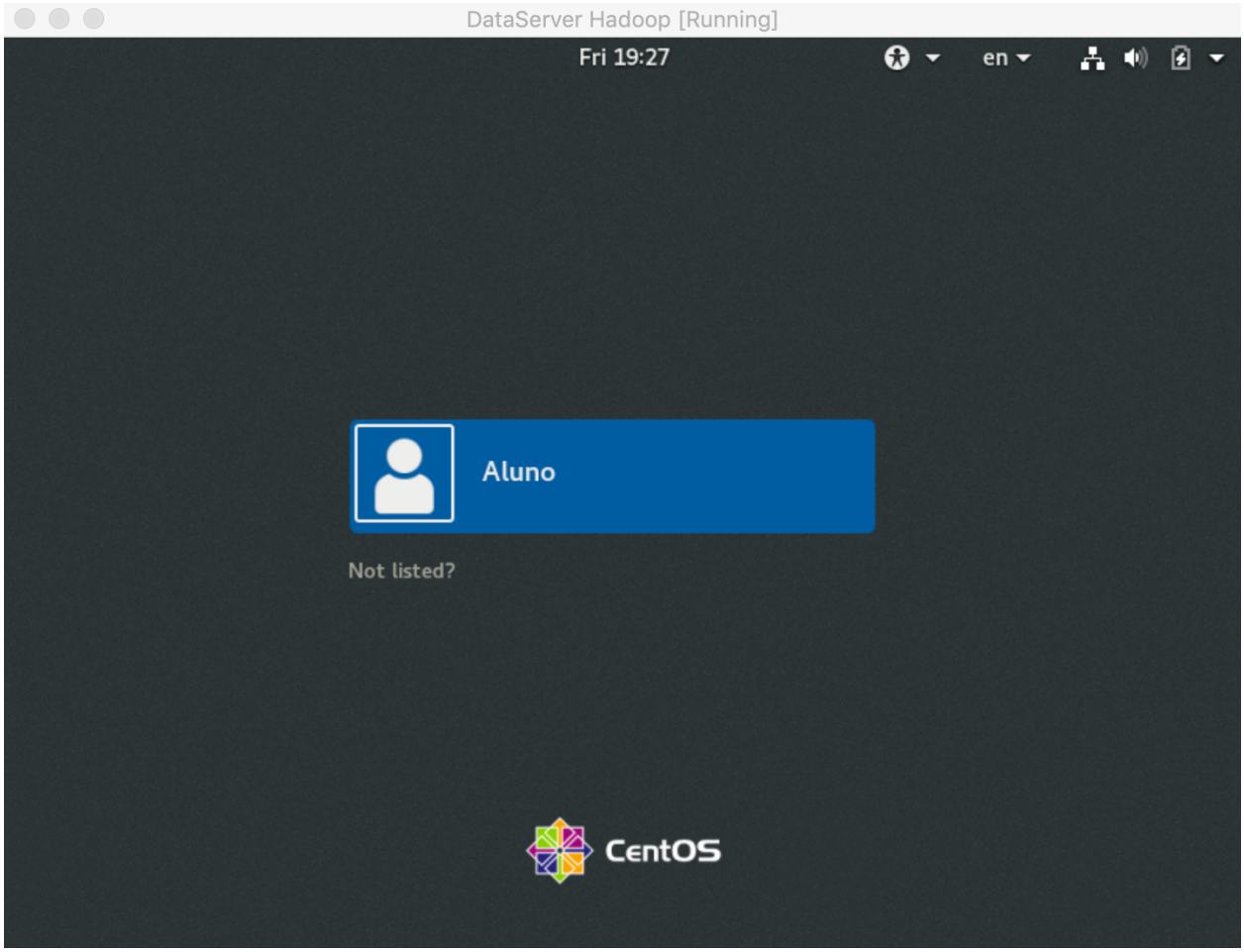
Clique em License Information para aceitar os termos de uso



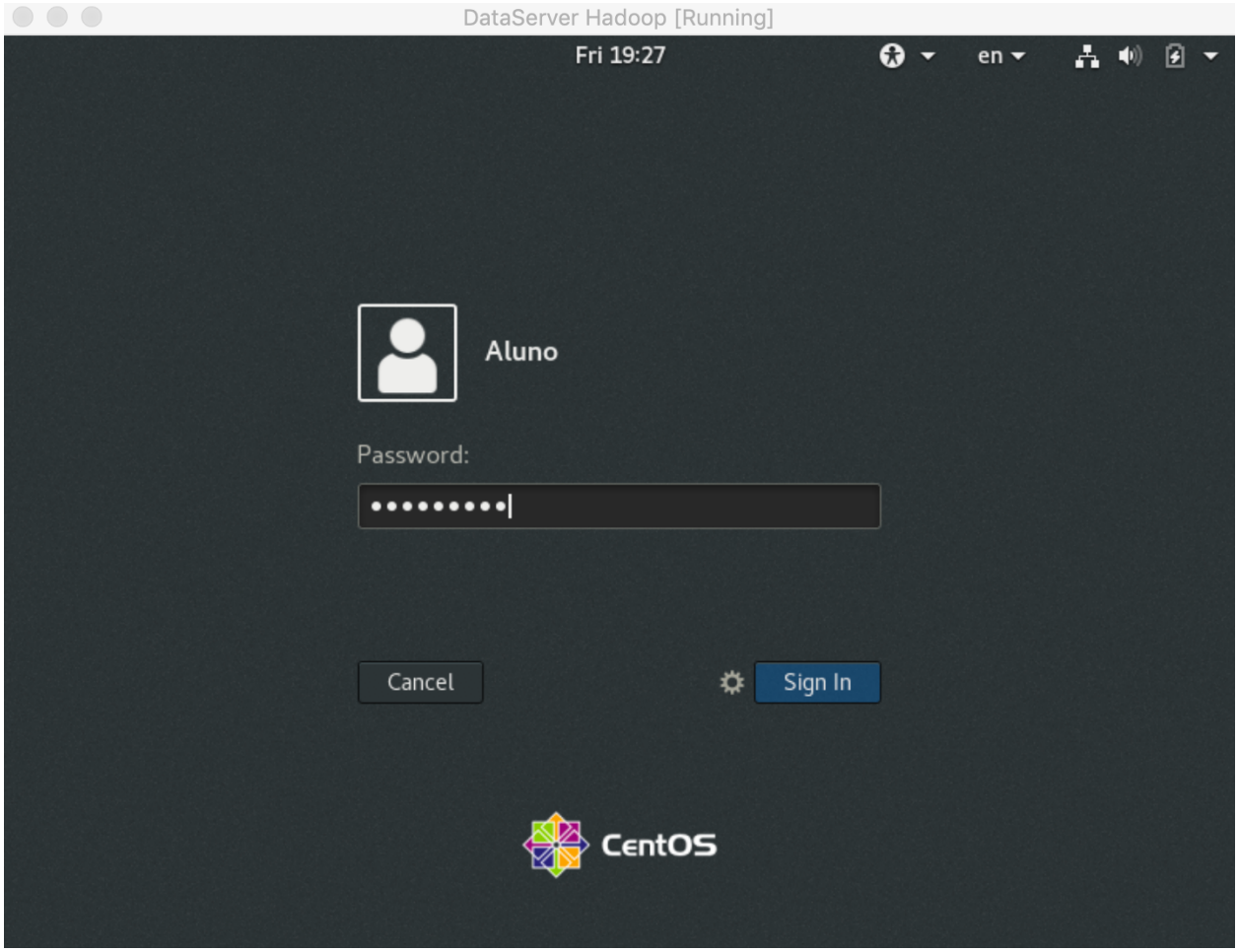
Marque a caixa e pressione Done.



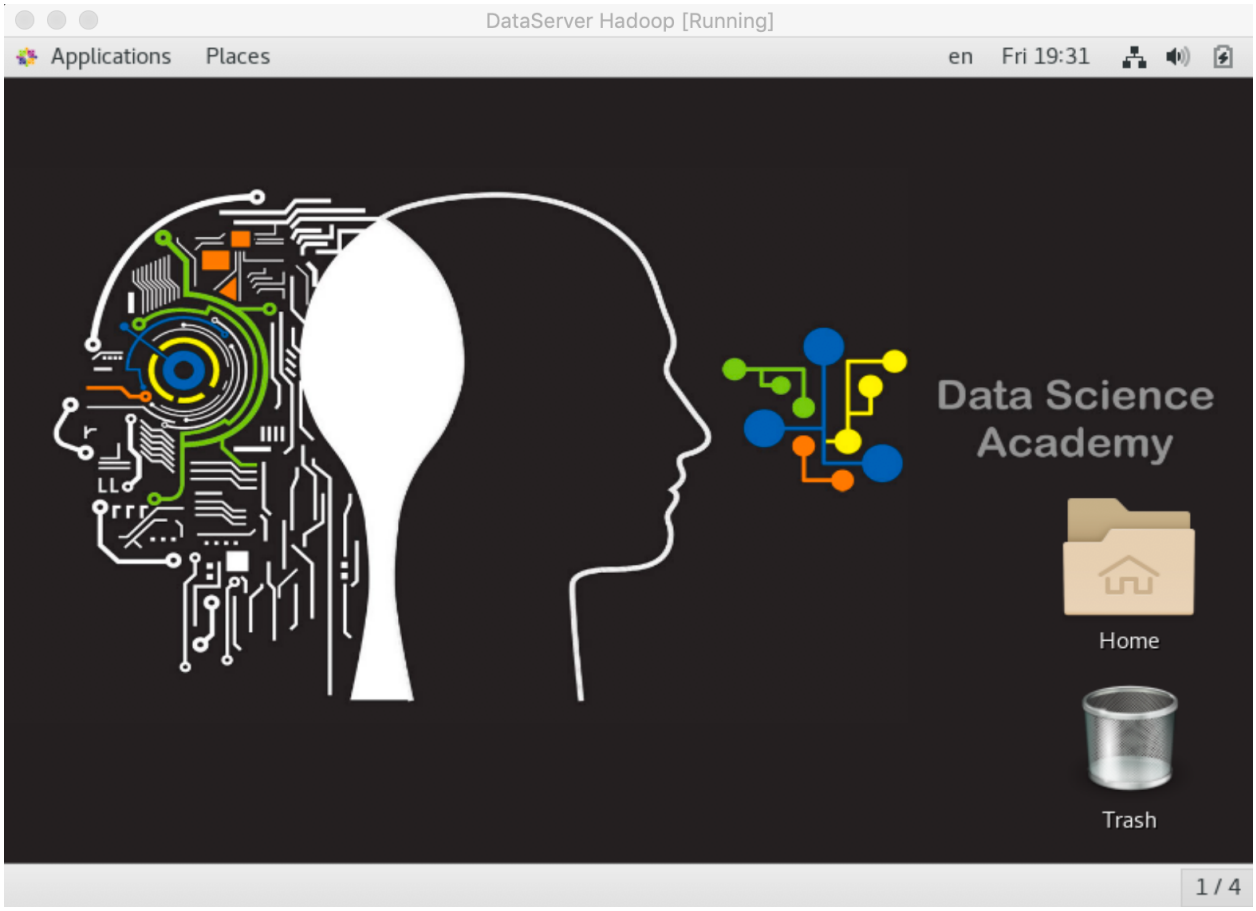
Se necessário, revise se o nome da máquina está correto e se a rede está ativada e pressione "Finish Configuration"



Clique no usuário Aluno



Usuário/Senha (**dsahadoop**)



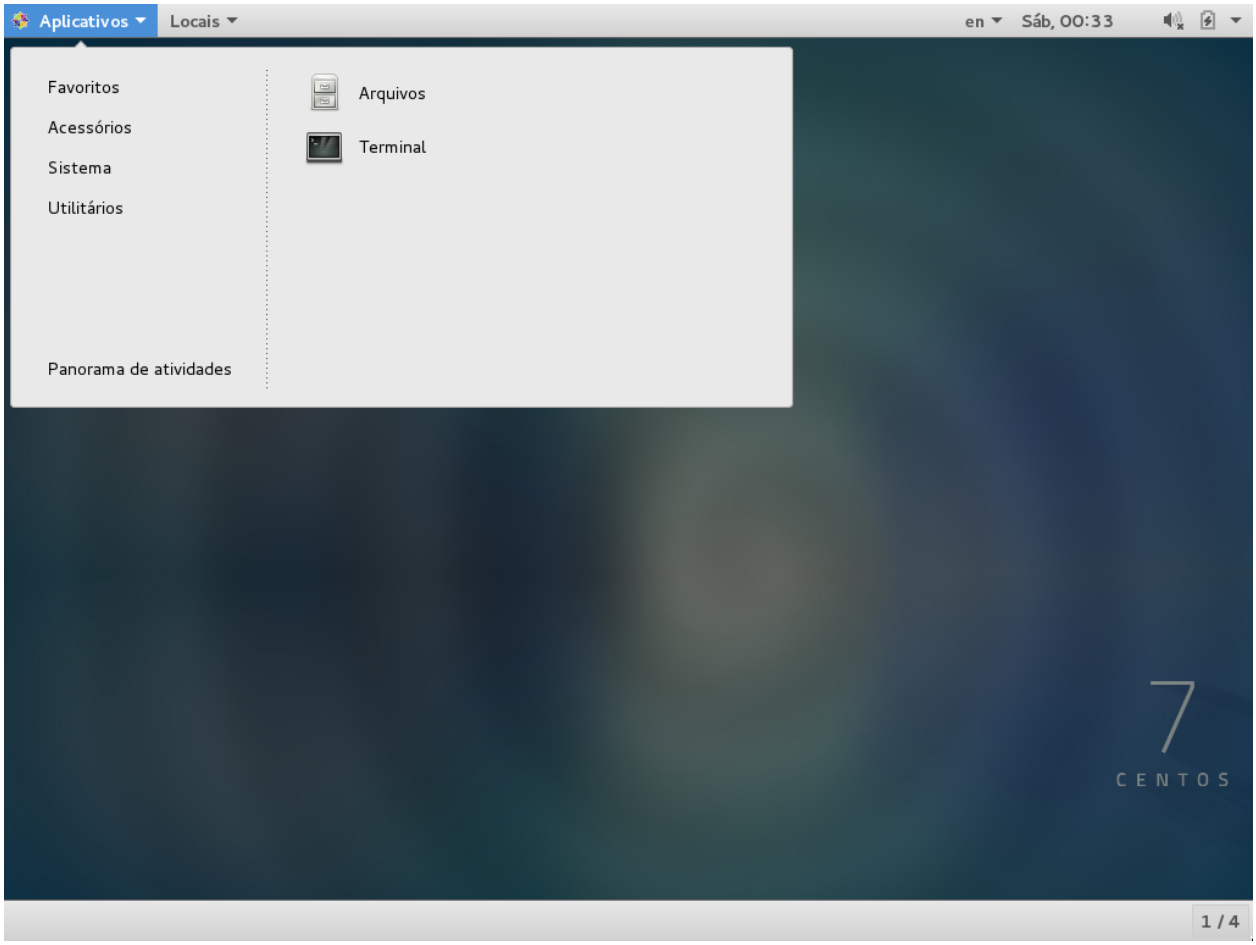
Instalação concluída com sucesso

Primeiro checkpoint:

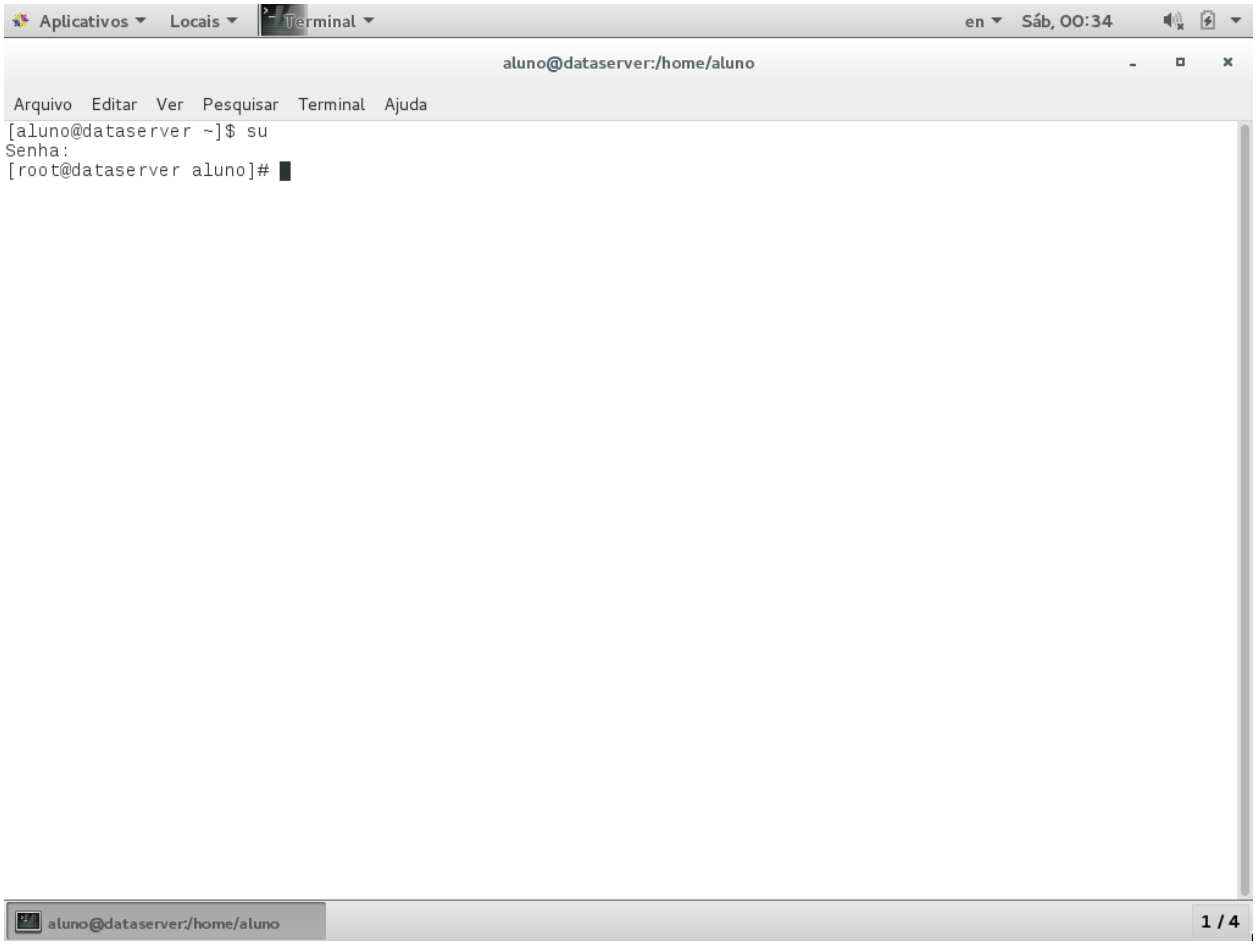
Desligue a VM (clique no ícone da bateria e então em desligar).
Clique no menu File do VirtualBox e clique em Export Appliance.
Será gerada uma cópia de segurança da sua máquina virtual.

→ **VM: DataServer-Hadoop-v1.0.ova (Apenas SO)**

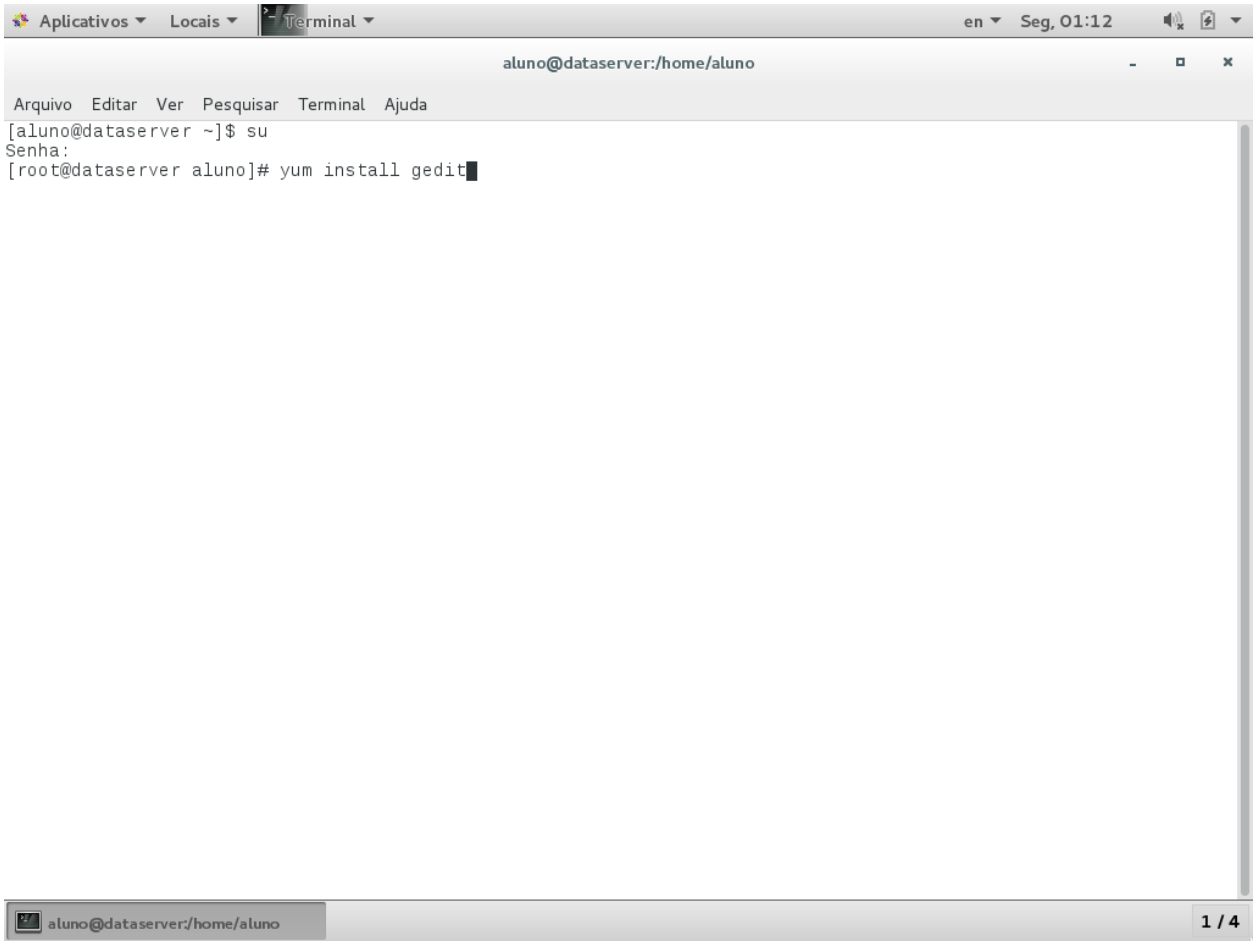
2.3. Instalação de Utilitários do Sistema Operacional



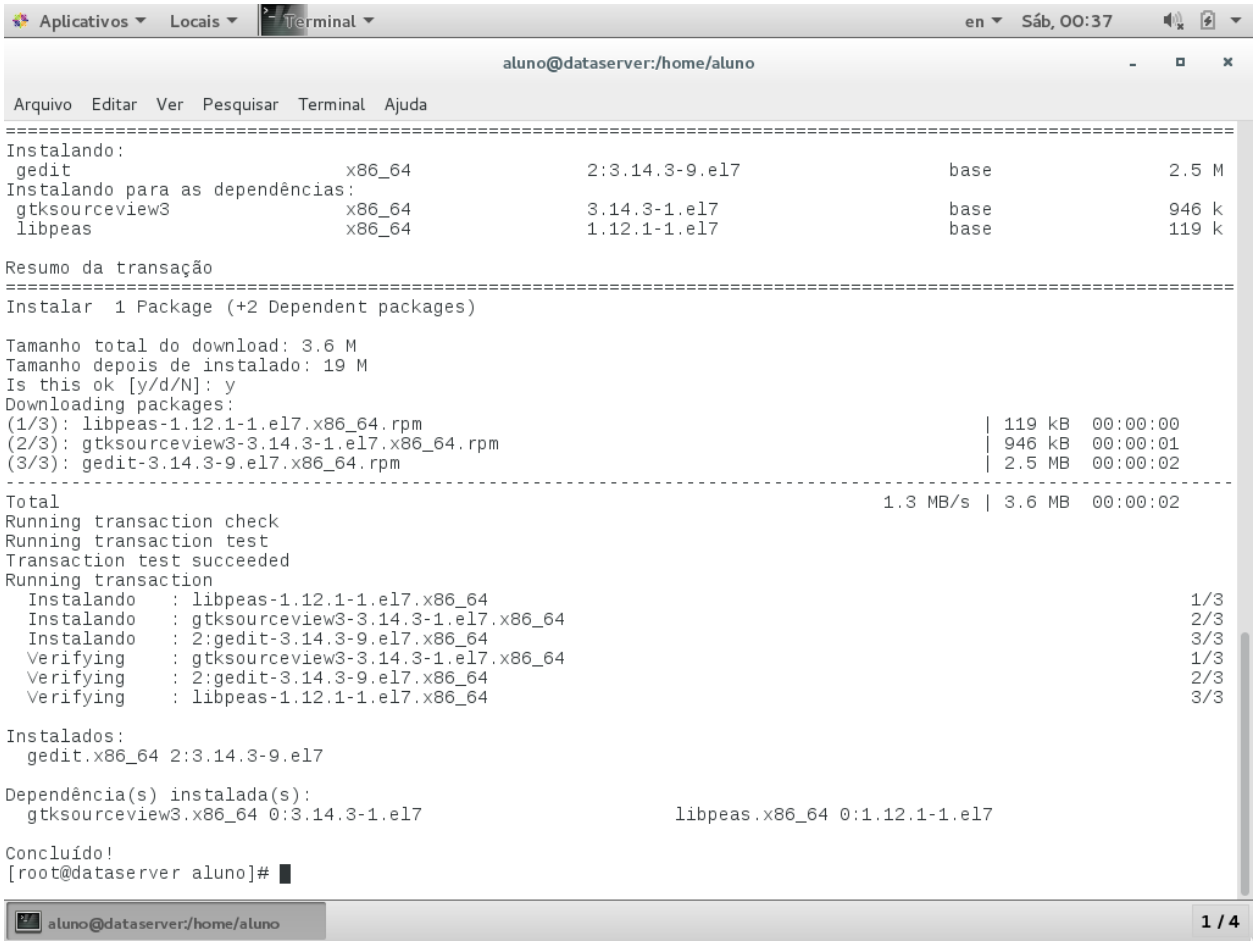
Abrindo o terminal



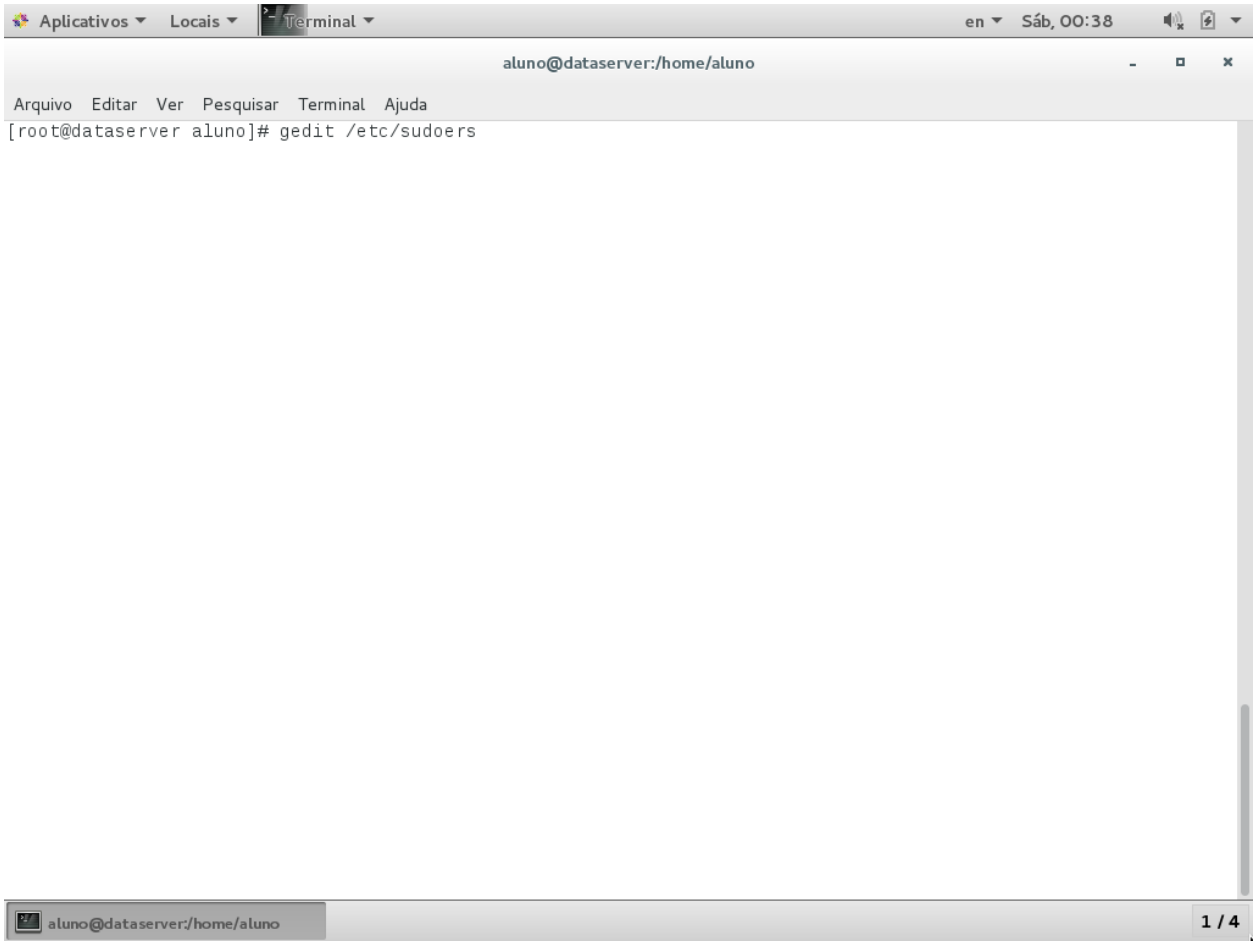
Efetuar login como root, usando o comando su. Senha: **dsahadoop**



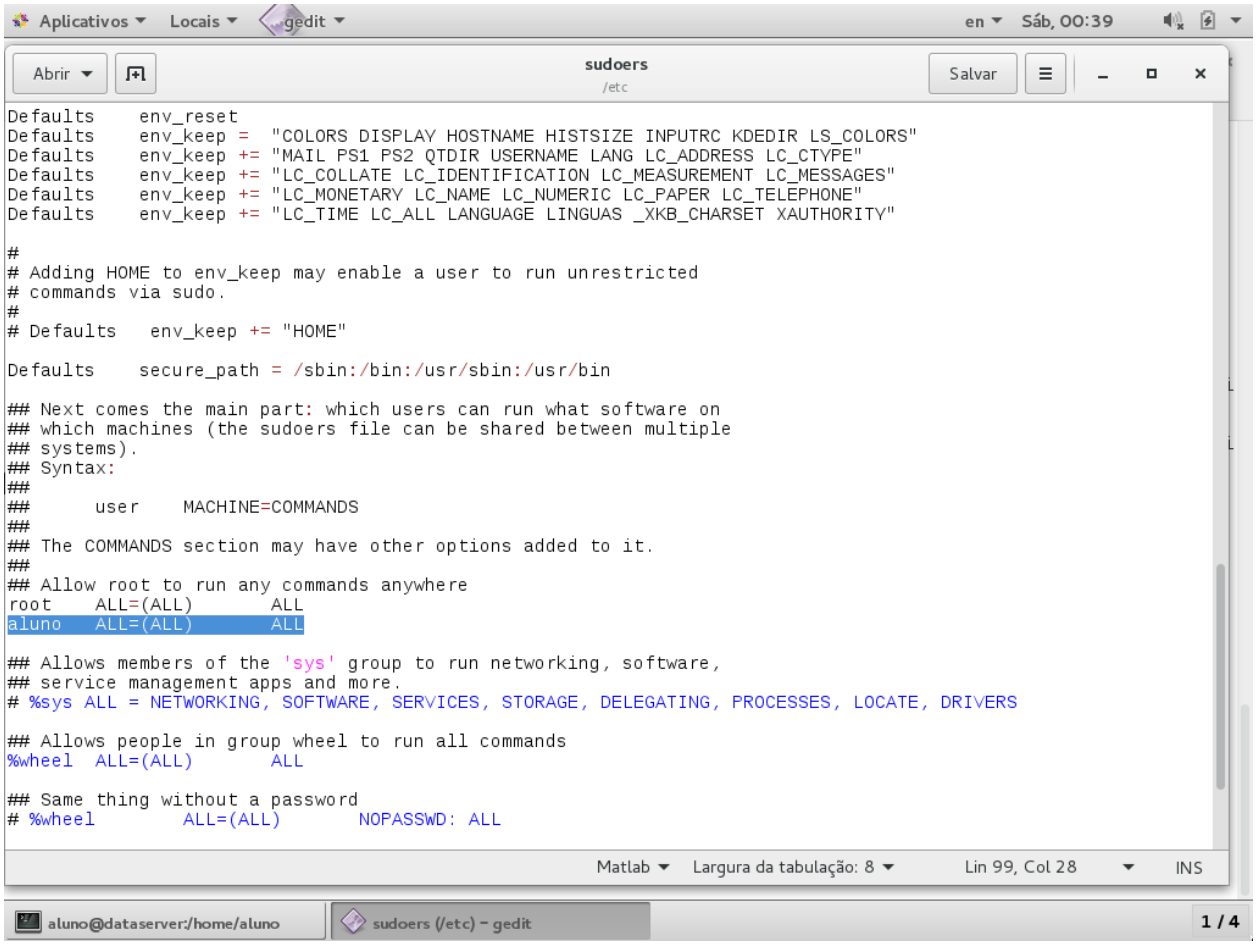
Instalar o editor de texto gedit, com o comando **yum install gedit**



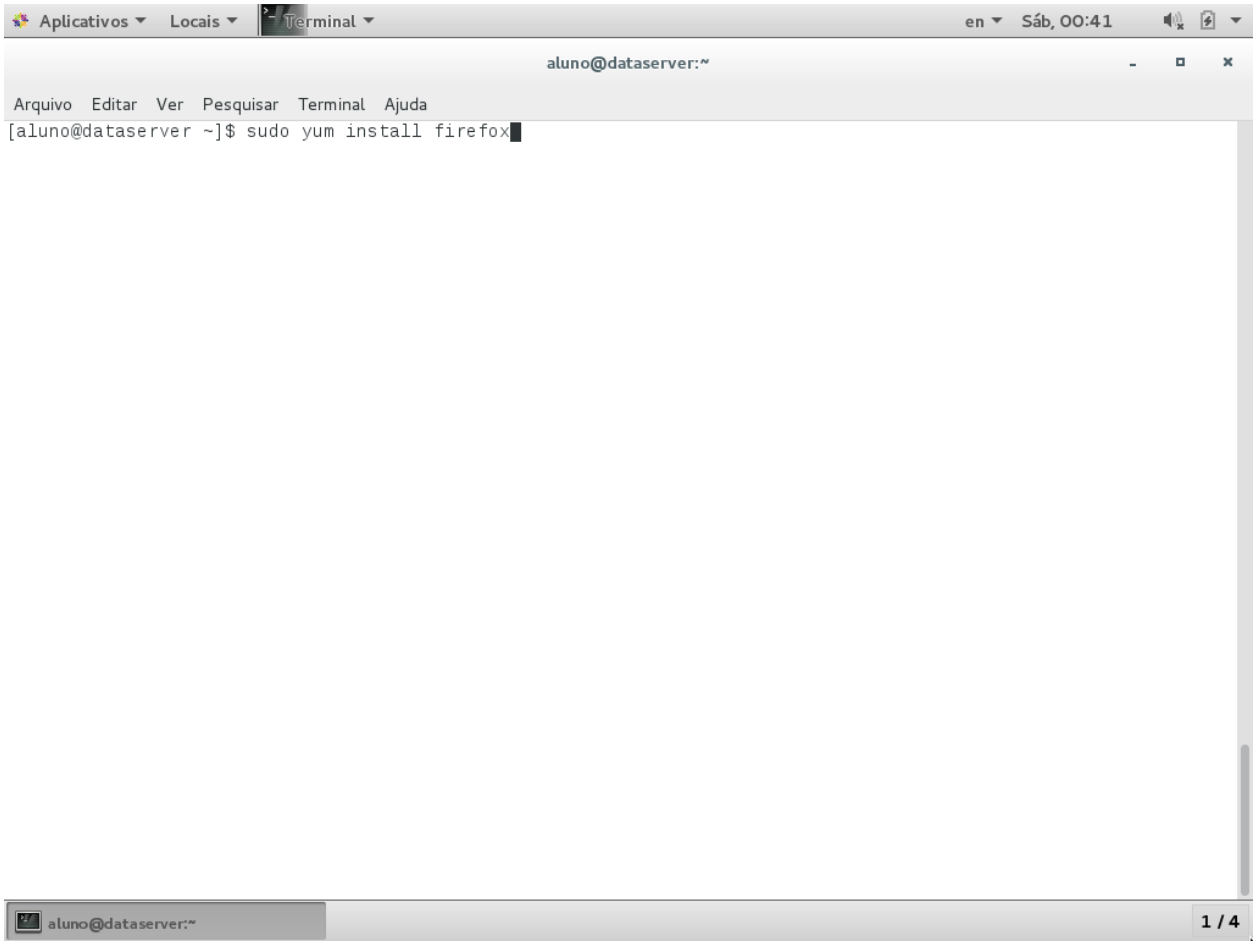
gedit instalado



Editar o arquivo /etc/sudoers usando o gedit

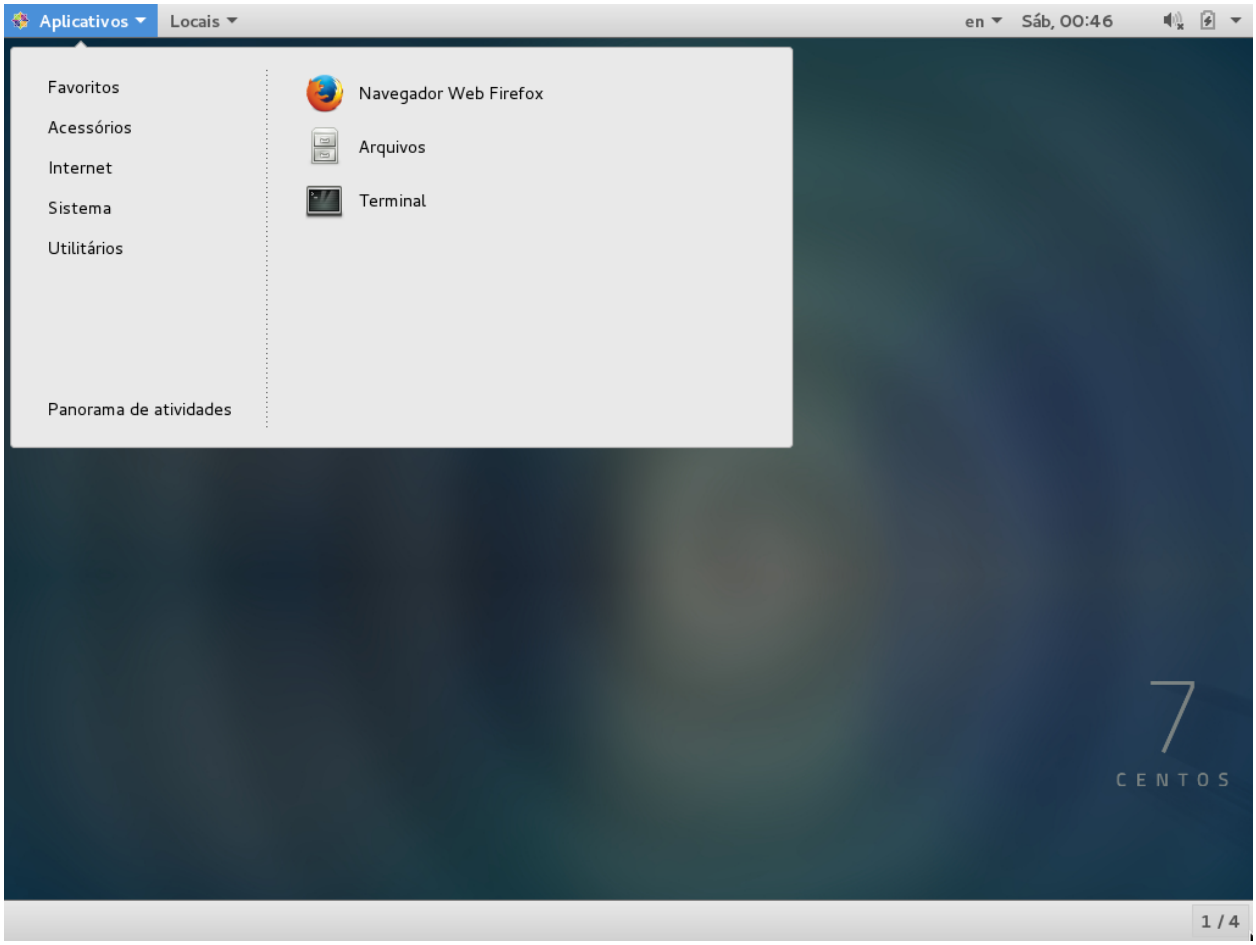


Incluir no arquivo a linha marcada acima e salvar o arquivo. Isso permitirá o usuário aluno executar comandos de administrador (root).

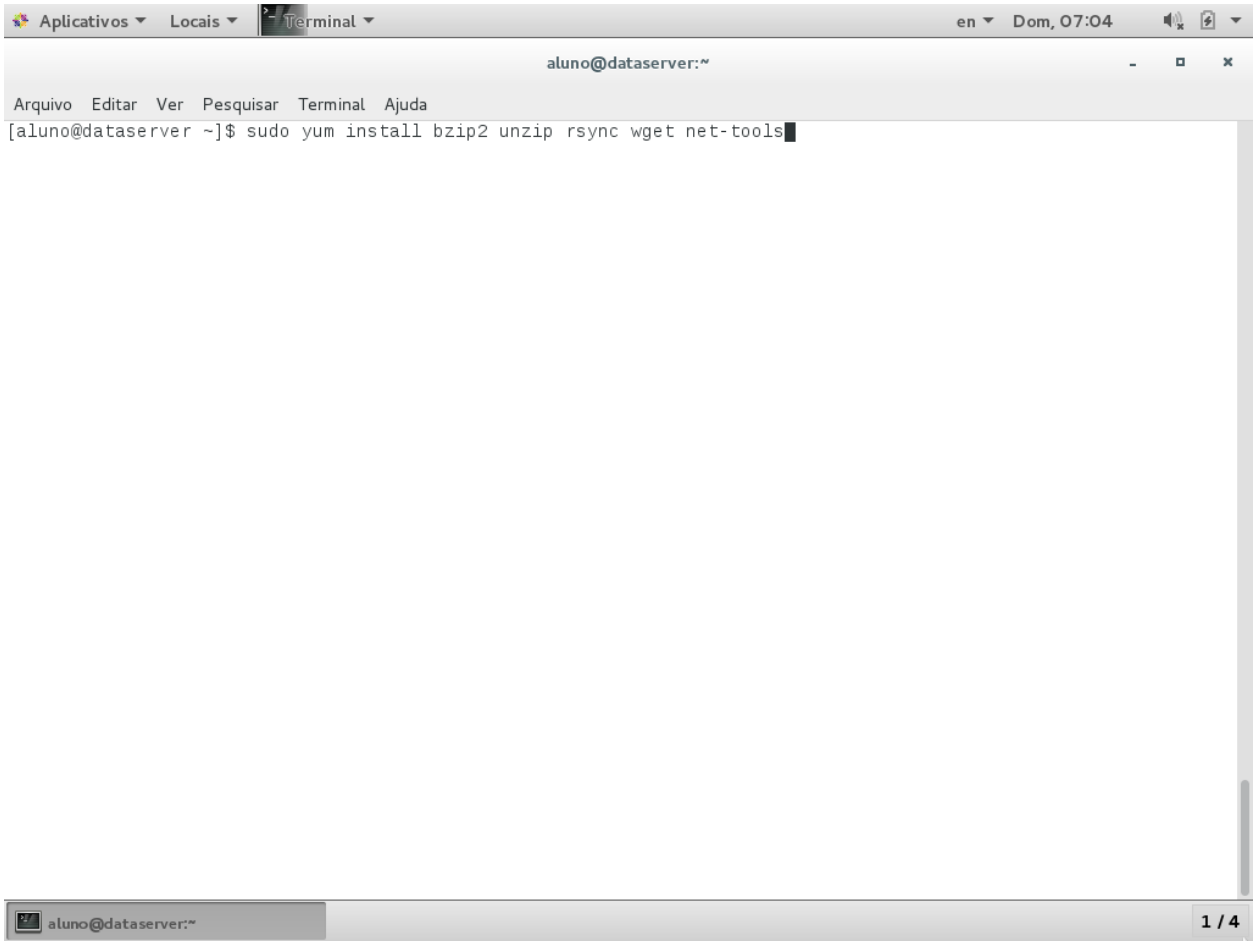


Conectado como usuário aluno, instalar o Firefox com o comando: **sudo yum install firefox**

Instalação concluída



Firefox instalado

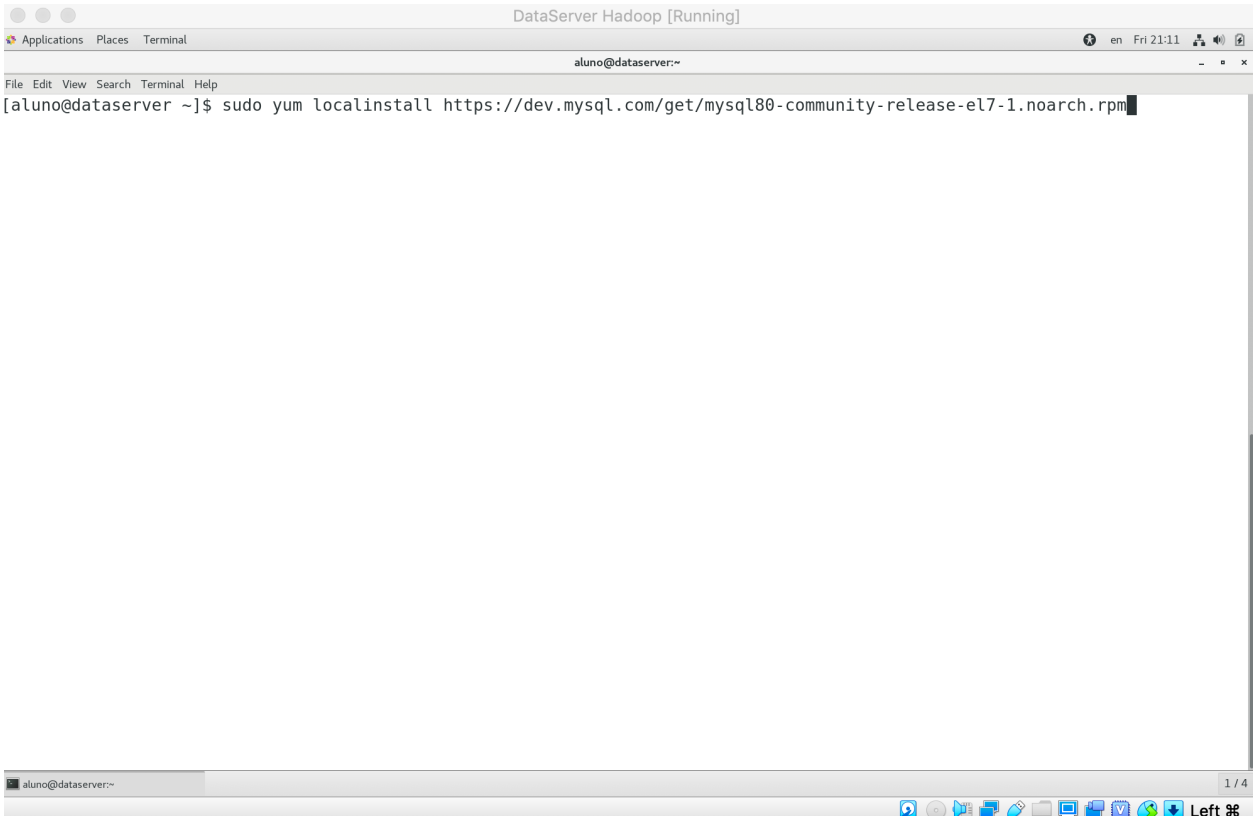


Instalar outros aplicativos: bzip2, unzip, rsync, wget e net-tools

Aplicativos instalados

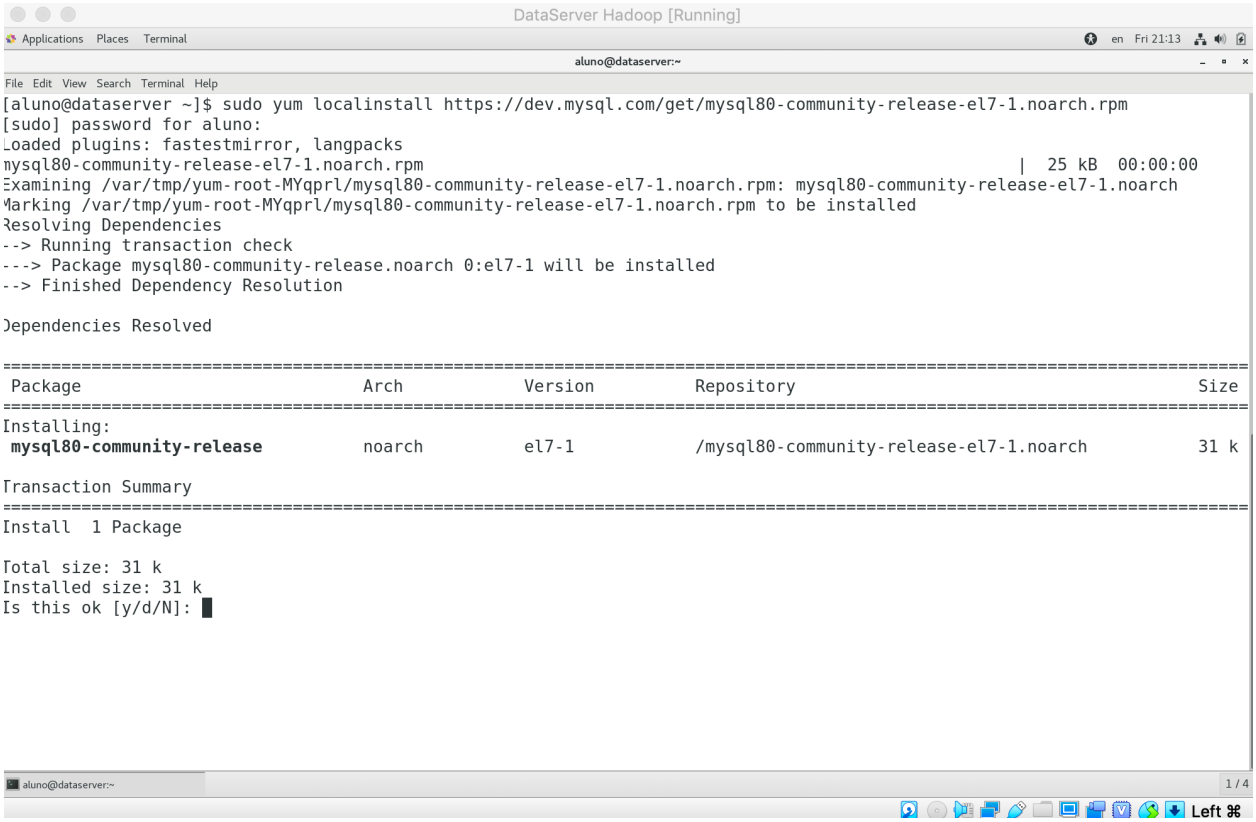
2.4. Instalação do MySQL

A instalação do MySQL pode ser feita via linha de comando com o seguinte procedimento:

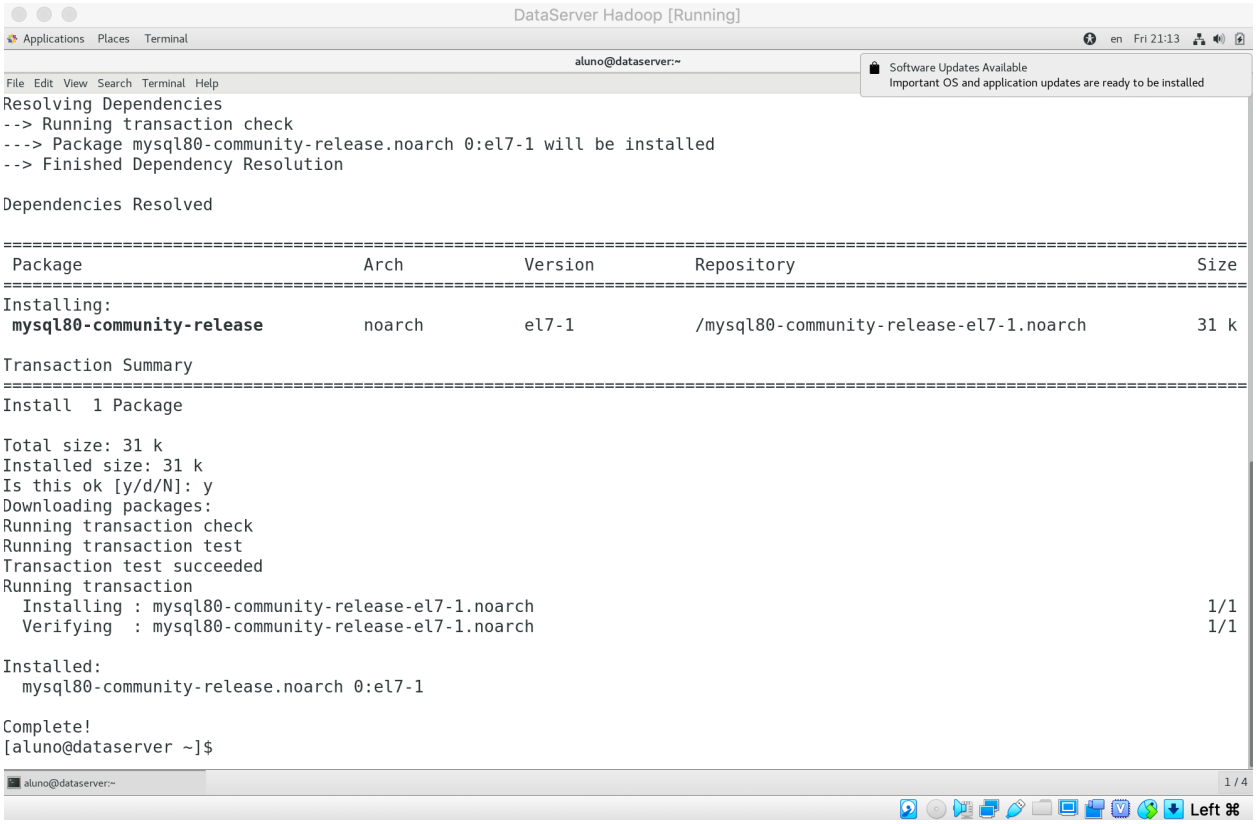


Execute o comando abaixo para baixar o pacote de instalação do MySQL para o CentOS:

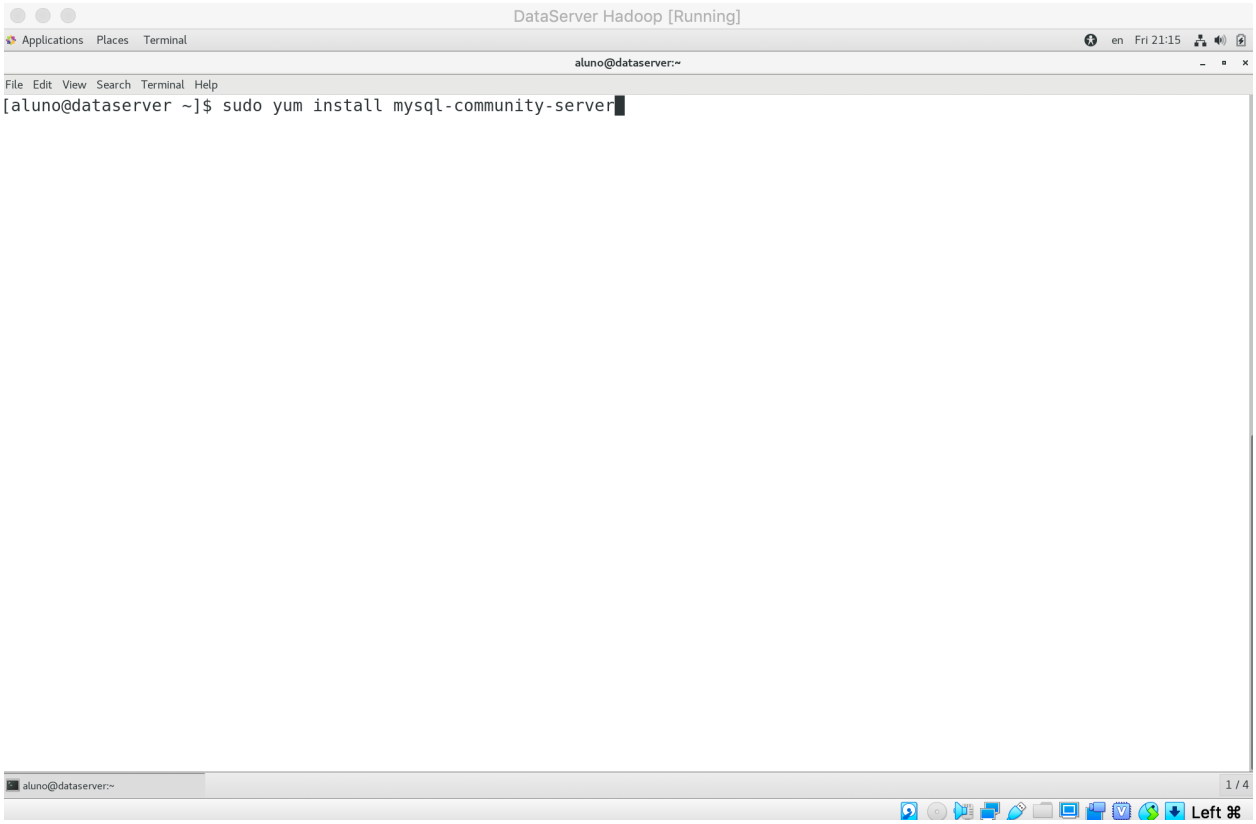
sudo yum localinstall https://dev.mysql.com/get/mysql80-community-release-el7-1.noarch.rpm


DataServer Hadoop [Running]
Applications Places Terminal
en Fri 21:13
aluno@dataserver:~
File Edit View Search Terminal Help
[aluno@dataserver ~]$ sudo yum localinstall https://dev.mysql.com/get/mysql80-community-release-el7-1.noarch.rpm
[sudo] password for aluno:
Loaded plugins: fastestmirror, langpacks
mysql80-community-release-el7-1.noarch.rpm | 25 kB 00:00:00
Examining /var/tmp/yum-root-MYqprl/mysql80-community-release-el7-1.noarch.rpm: mysql80-community-release-el7-1.noarch
Marking /var/tmp/yum-root-MYqprl/mysql80-community-release-el7-1.noarch.rpm to be installed
Resolving Dependencies
--> Running transaction check
---> Package mysql80-community-release.noarch 0:el7-1 will be installed
--> Finished Dependency Resolution
Dependencies Resolved
Package Arch Version Repository Size
Installing:
mysql80-community-release noarch el7-1 /mysql80-community-release-el7-1.noarch 31 k
Transaction Summary
Install 1 Package
Total size: 31 k
Installed size: 31 k
Is this ok [y/d/N]:
aluno@dataserver:~
1 / 4
Left ⌘

Pressione y

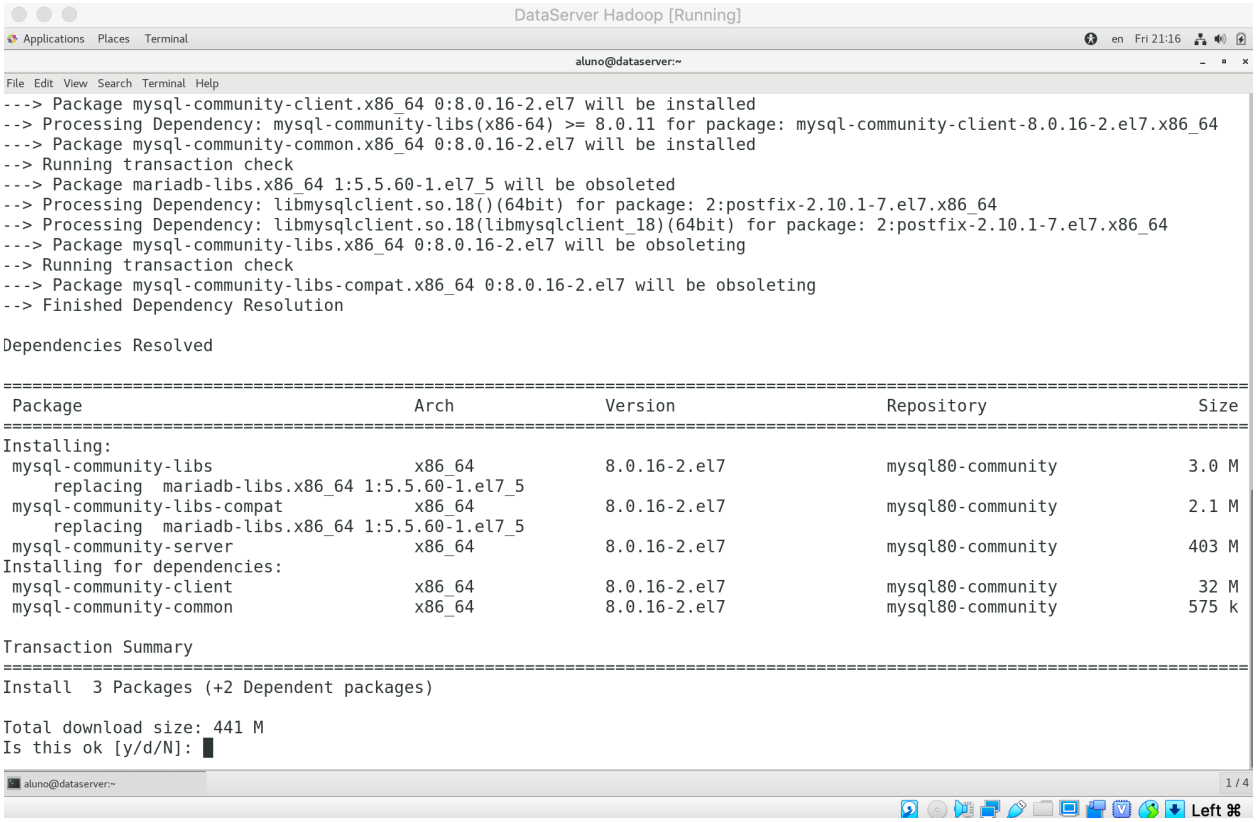


Download do pacote concluído

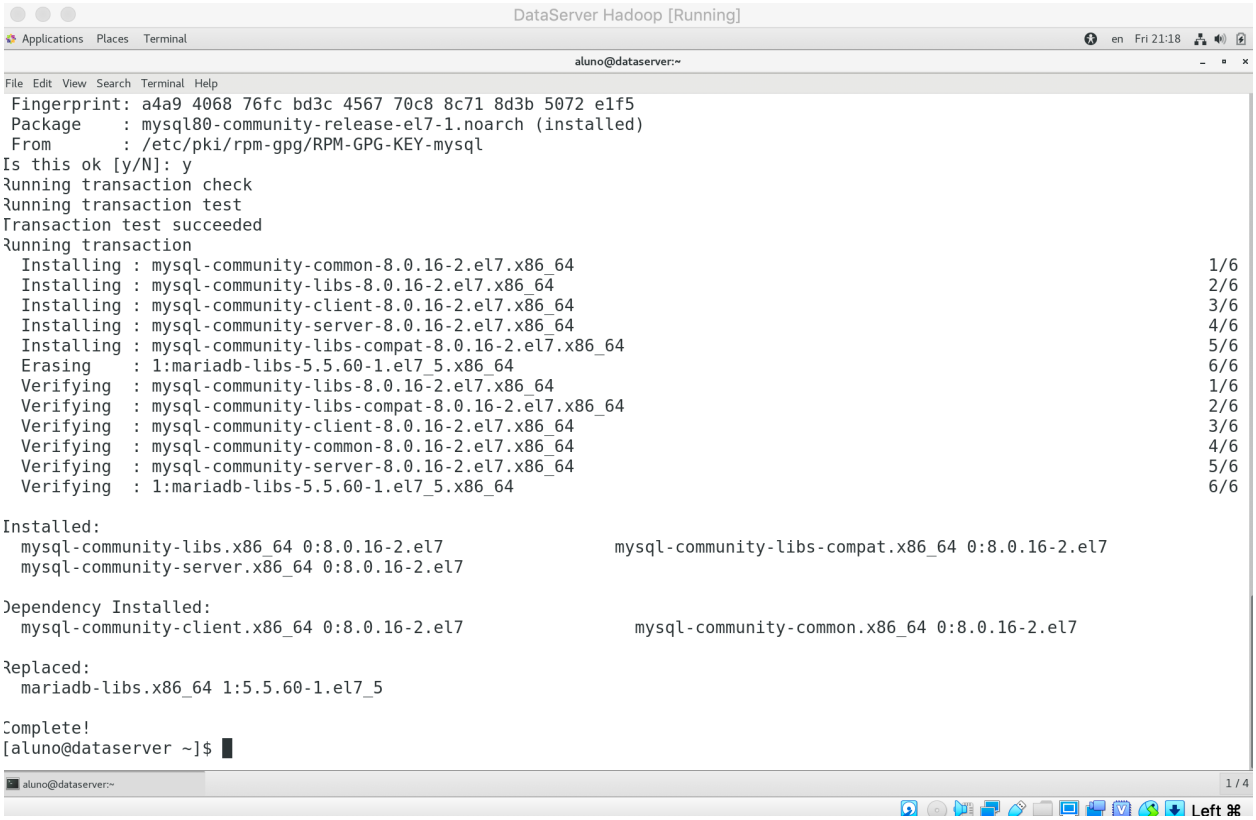


Esse comando inicia a instalação do MySQL:

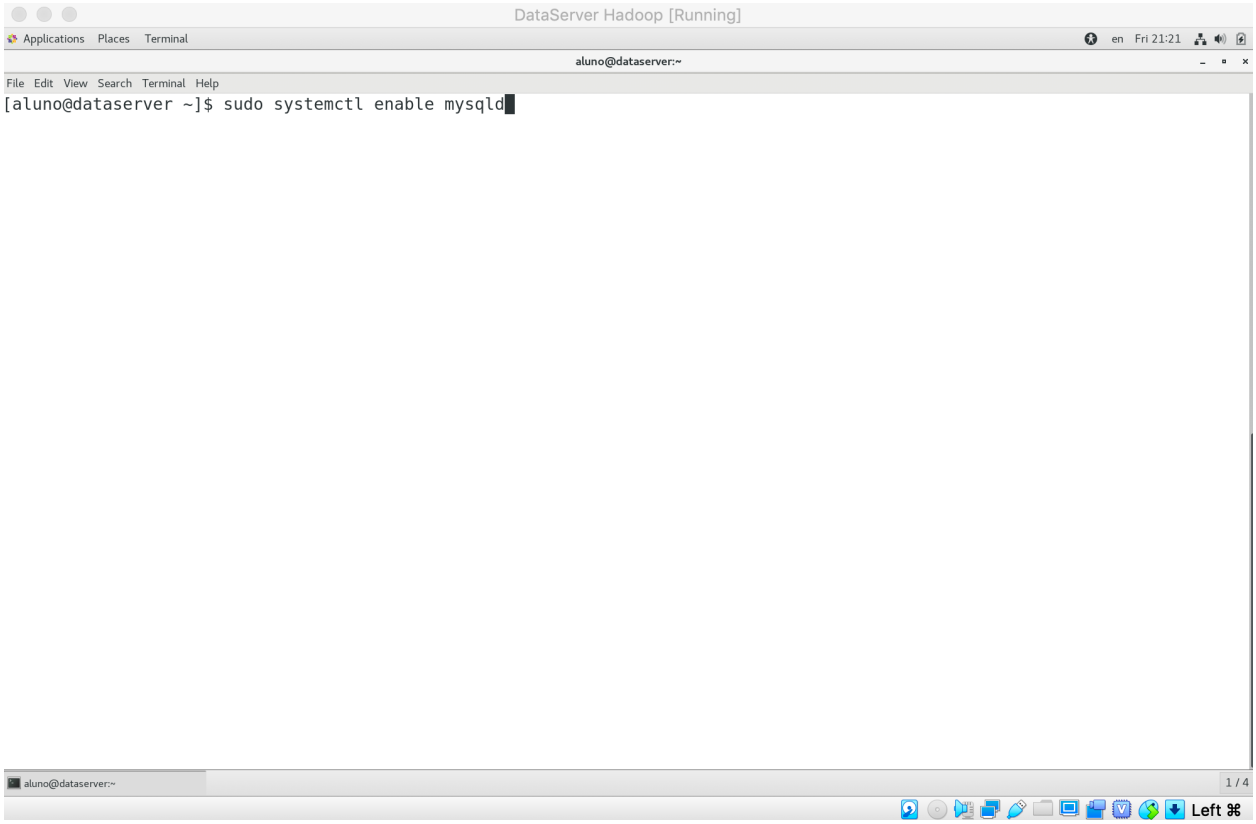
sudo yum install mysql-community-server


DataServer Hadoop [Running]
Applications Places Terminal
en Fri 21:16
aluno@dataserver:~
File Edit View Search Terminal Help
---> Package mysql-community-client.x86_64 0:8.0.16-2.el7 will be installed
--> Processing Dependency: mysql-community-libs(x86-64) >= 8.0.11 for package: mysql-community-client-8.0.16-2.el7.x86_64
---> Package mysql-community-common.x86_64 0:8.0.16-2.el7 will be installed
--> Running transaction check
---> Package mariadb-libs.x86_64 1:5.5.60-1.el7_5 will be obsoleted
--> Processing Dependency: libmysqlclient.so.18()(64bit) for package: 2:postfix-2.10.1-7.el7.x86_64
--> Processing Dependency: libmysqlclient.so.18(libmysqlclient_18)(64bit) for package: 2:postfix-2.10.1-7.el7.x86_64
---> Package mysql-community-libs.x86_64 0:8.0.16-2.el7 will be obsoleting
--> Running transaction check
---> Package mysql-community-libs-compat.x86_64 0:8.0.16-2.el7 will be obsoleting
--> Finished Dependency Resolution
Dependencies Resolved
Package Arch Version Repository Size
Installing:
mysql-community-libs x86_64 8.0.16-2.el7 mysql80-community 3.0 M
replacing mariadb-libs.x86_64 1:5.5.60-1.el7_5
mysql-community-libs-compat x86_64 8.0.16-2.el7 mysql80-community 2.1 M
replacing mariadb-libs.x86_64 1:5.5.60-1.el7_5
mysql-community-server x86_64 8.0.16-2.el7 mysql80-community 403 M
Installing for dependencies:
mysql-community-client x86_64 8.0.16-2.el7 mysql80-community 32 M
mysql-community-common x86_64 8.0.16-2.el7 mysql80-community 575 k
Transaction Summary
Install 3 Packages (+2 Dependent packages)
Total download size: 441 M
Is this ok [y/d/N]:
aluno@dataserver:~
1 / 4
Left ⌘

Pressione y

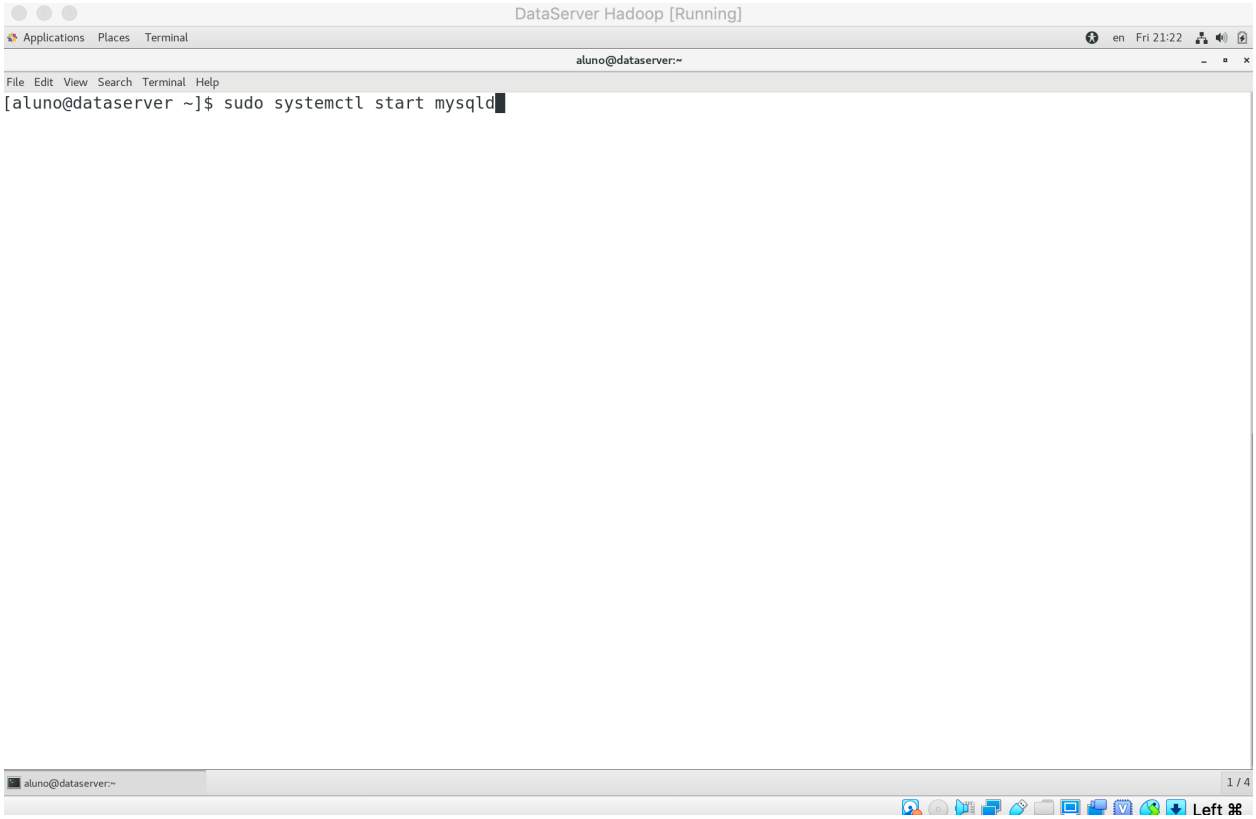


Instalação concluída



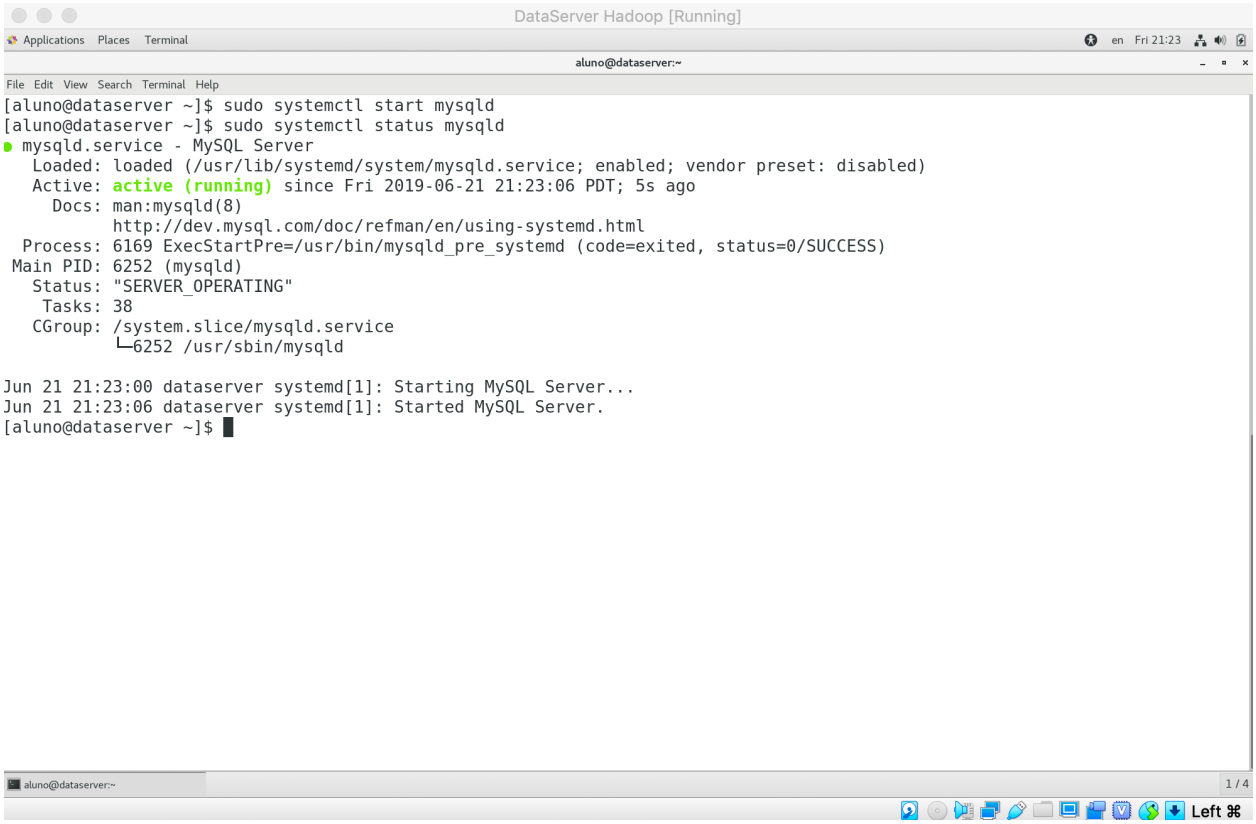
Ativando o serviço do MySQL

sudo systemctl enable mysqld



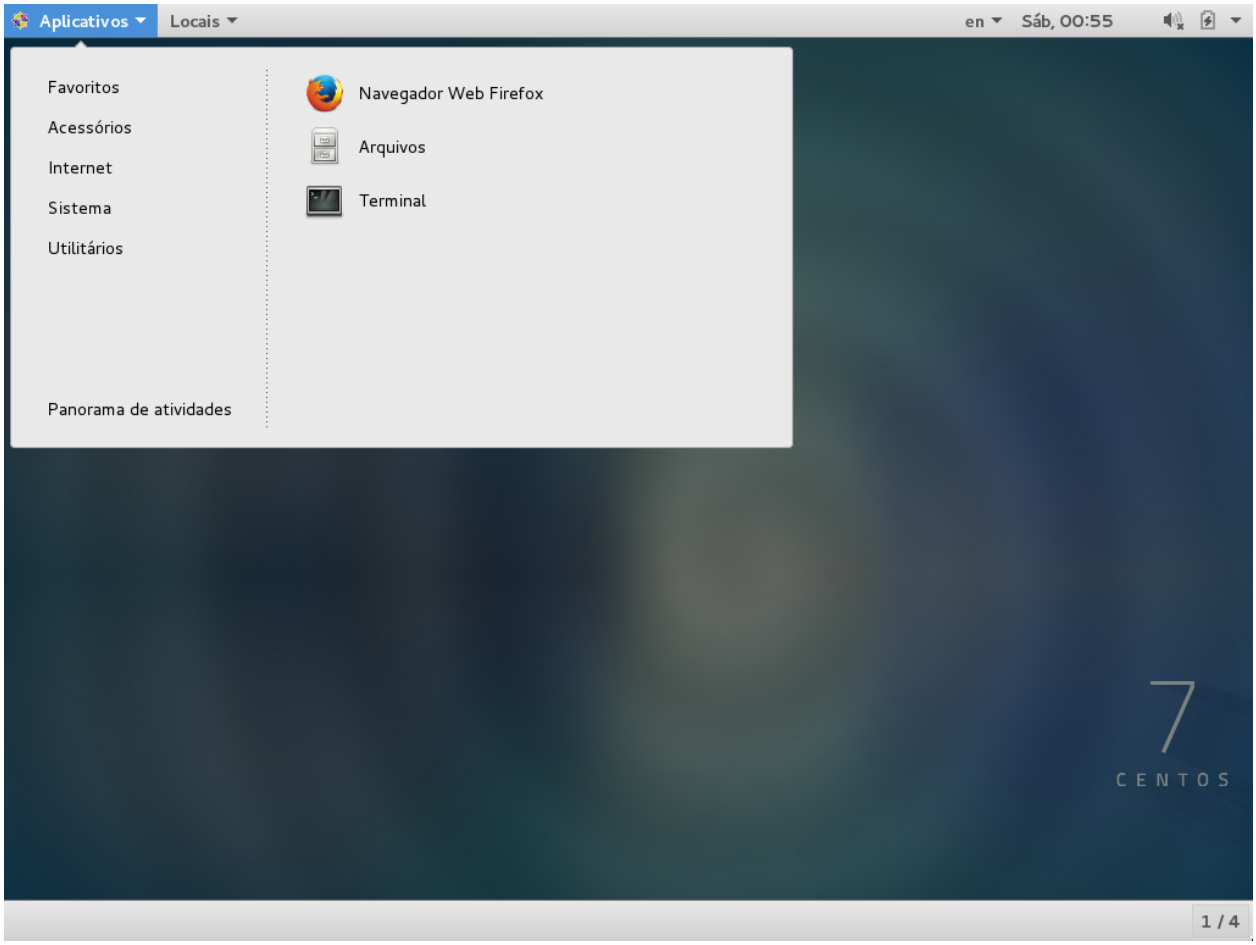
Inicia o MySQL

sudo systemctl start mysqld

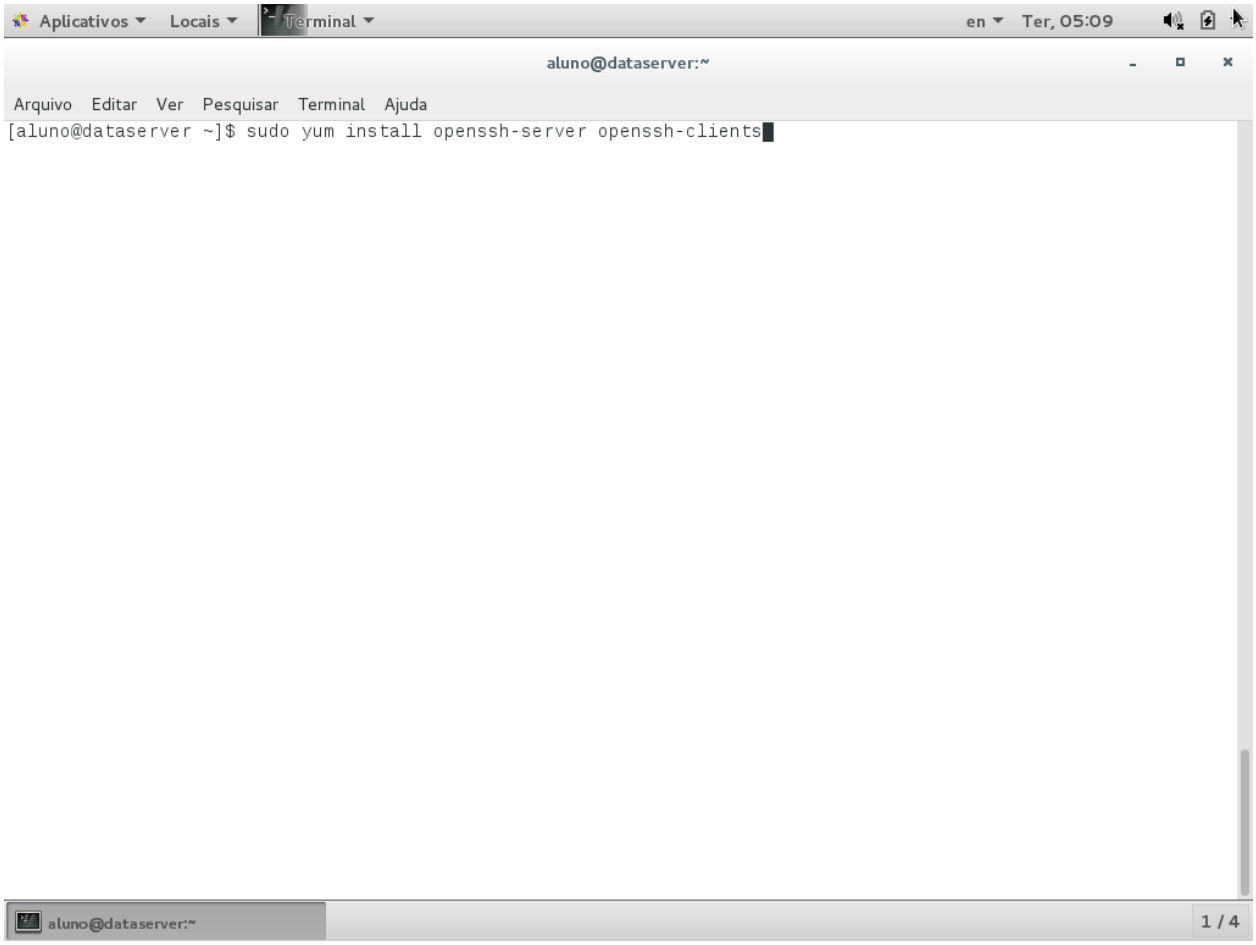


MySQL em execução

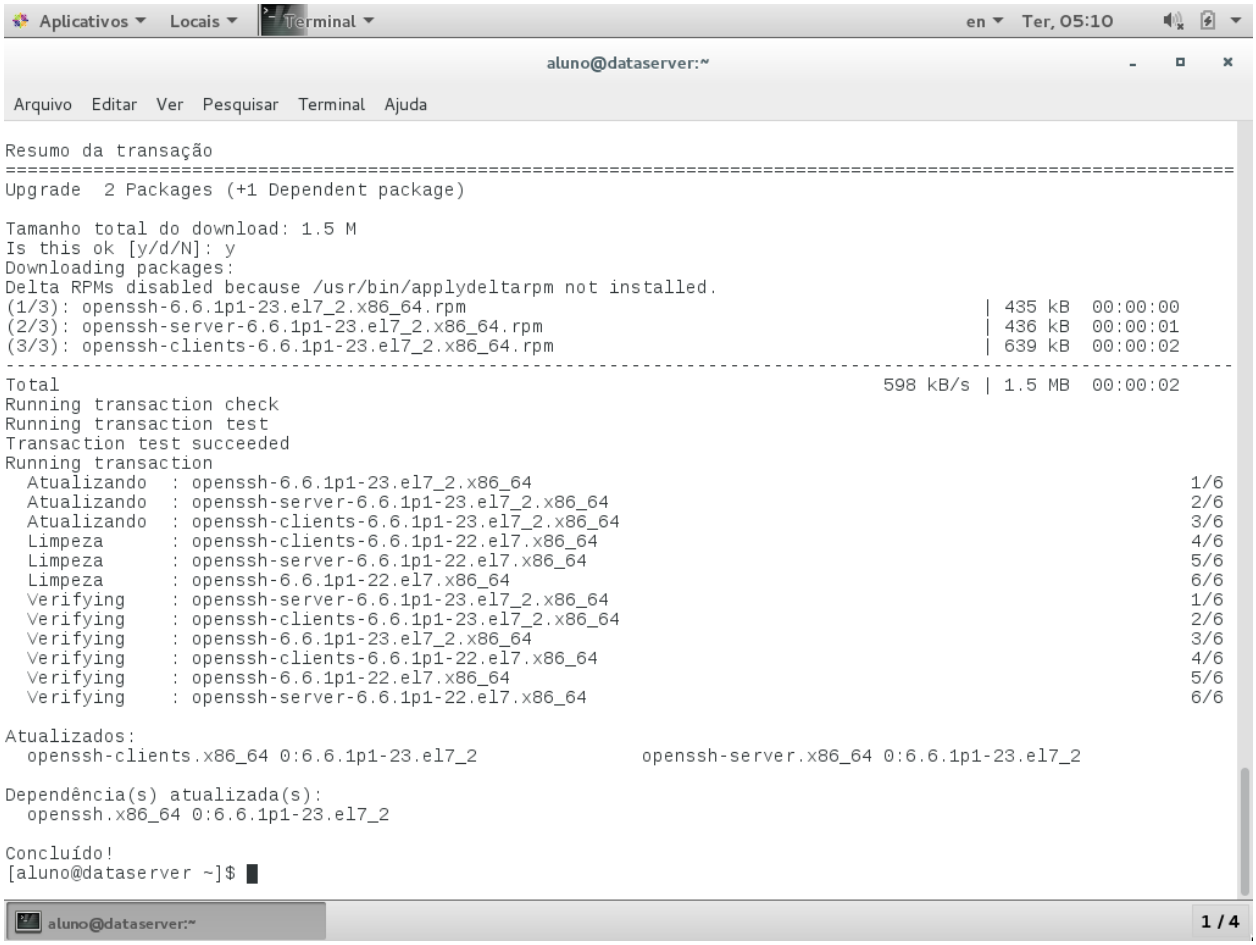
3. Instalação do servidor ssh



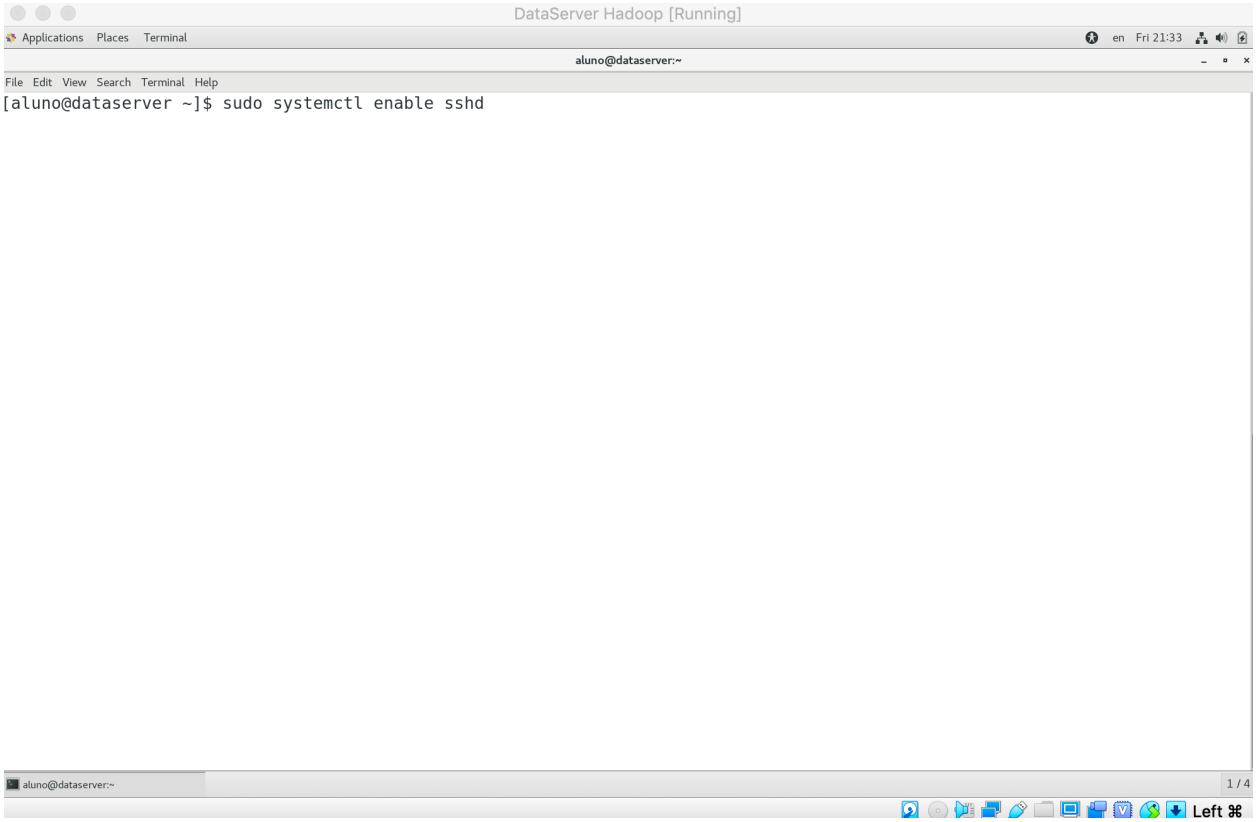
Abrindo o terminal



sudo yum install openssh-server openssh-clients

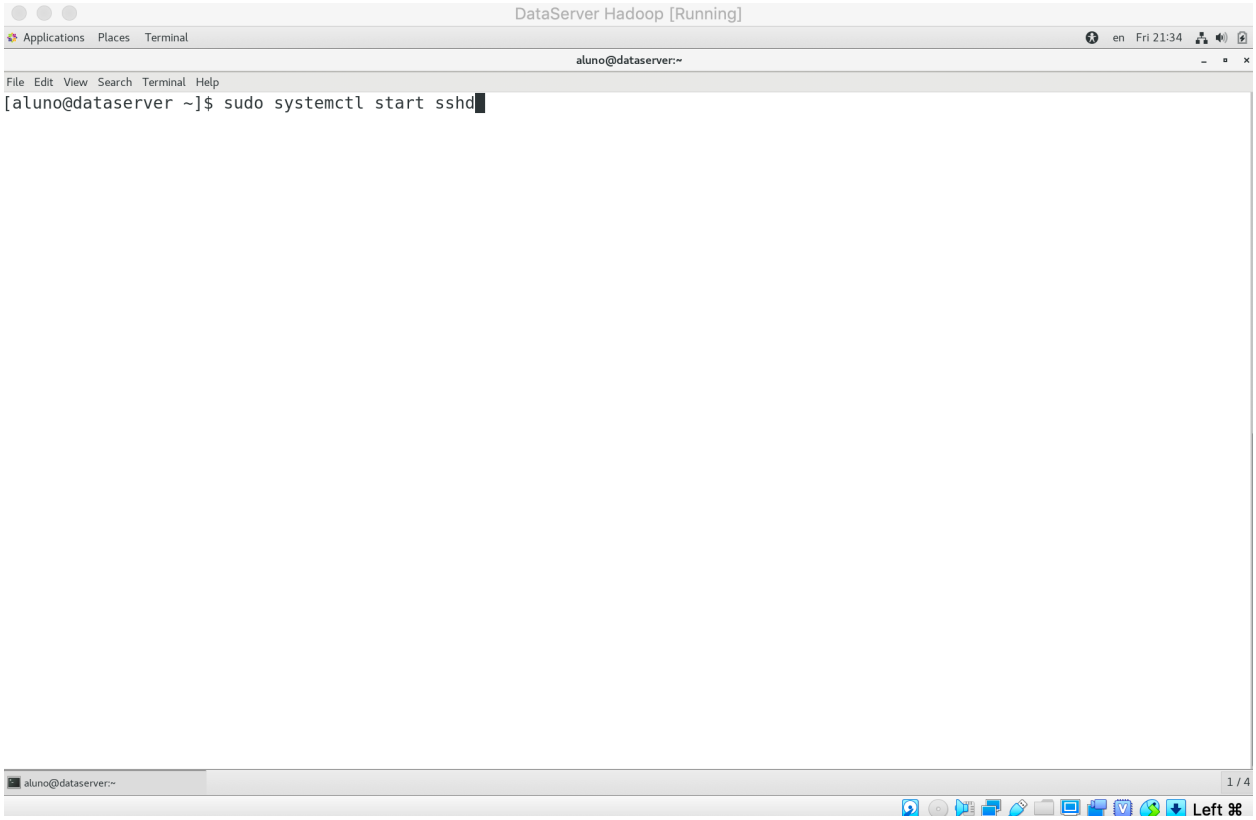

Aplicativos Locais Terminal en Ter, 05:10
aluno@dataserver:~
Arquivo Editar Ver Pesquisar Terminal Ajuda
Resumo da transação
Upgrade 2 Packages (+1 Dependent package)
Tamanho total do download: 1.5 M
Is this ok [y/d/N]: y
Downloading packages:
Delta RPMs disabled because /usr/bin/applydeltarpm not installed.
(1/3): openssh-6.6.1p1-23.el7_2.x86_64.rpm | 435 kB 00:00:00
(2/3): openssh-server-6.6.1p1-23.el7_2.x86_64.rpm | 436 kB 00:00:01
(3/3): openssh-clients-6.6.1p1-23.el7_2.x86_64.rpm | 639 kB 00:00:02
Total 598 kB/s | 1.5 MB 00:00:02
Running transaction check
Running transaction test
Transaction test succeeded
Running transaction
Atualizando : openssh-6.6.1p1-23.el7_2.x86_64 1/6
Atualizando : openssh-server-6.6.1p1-23.el7_2.x86_64 2/6
Atualizando : openssh-clients-6.6.1p1-23.el7_2.x86_64 3/6
Limpeza : openssh-clients-6.6.1p1-22.el7.x86_64 4/6
Limpeza : openssh-server-6.6.1p1-22.el7.x86_64 5/6
Limpeza : openssh-6.6.1p1-22.el7.x86_64 6/6
Verifying : openssh-server-6.6.1p1-23.el7_2.x86_64 1/6
Verifying : openssh-clients-6.6.1p1-23.el7_2.x86_64 2/6
Verifying : openssh-6.6.1p1-23.el7_2.x86_64 3/6
Verifying : openssh-clients-6.6.1p1-22.el7.x86_64 4/6
Verifying : openssh-6.6.1p1-22.el7.x86_64 5/6
Verifying : openssh-server-6.6.1p1-22.el7.x86_64 6/6
Atualizados:
openssh-clients.x86_64 0:6.6.1p1-23.el7_2 openssh-server.x86_64 0:6.6.1p1-23.el7_2
Dependência(s) atualizada(s):
openssh.x86_64 0:6.6.1p1-23.el7_2
Concluído!
[aluno@dataserver ~]$
aluno@dataserver:~
1 / 4

Concluído



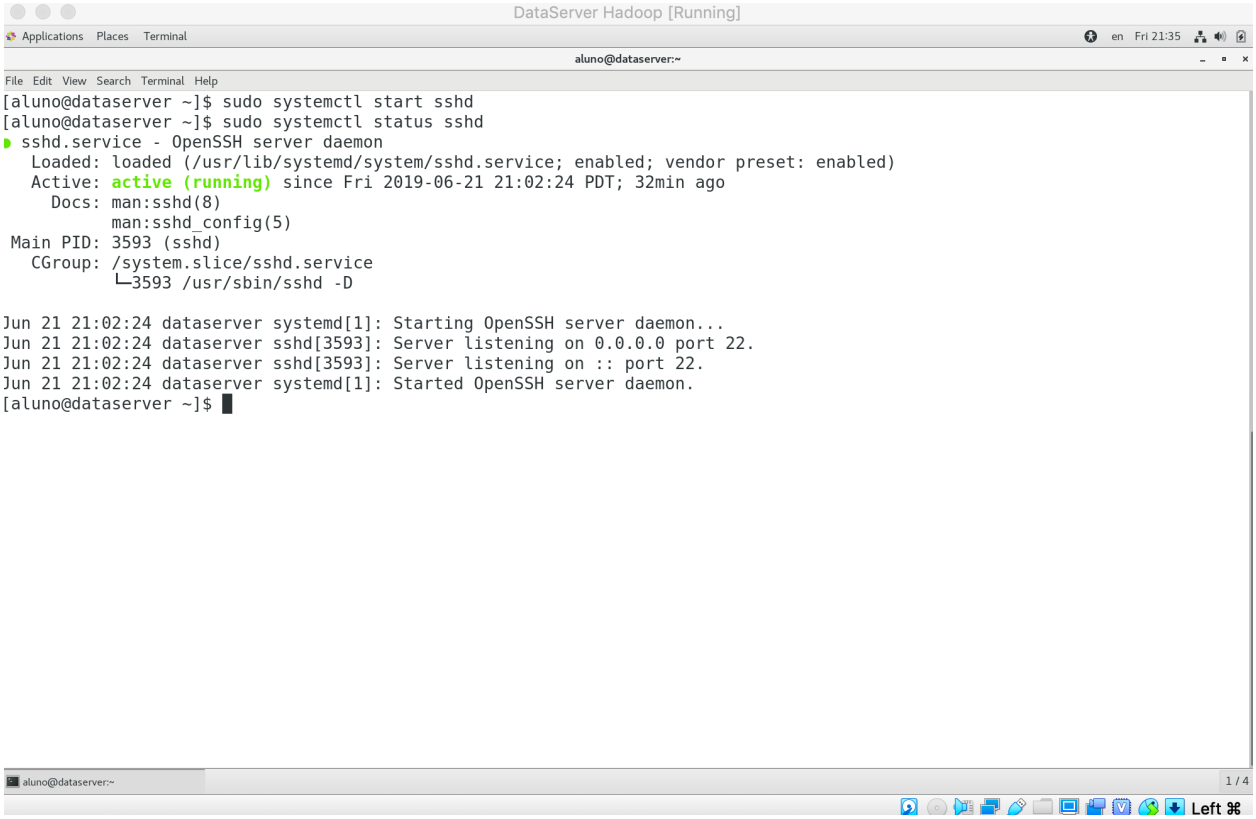
Habilitando o serviço

sudo systemctl enable sshd

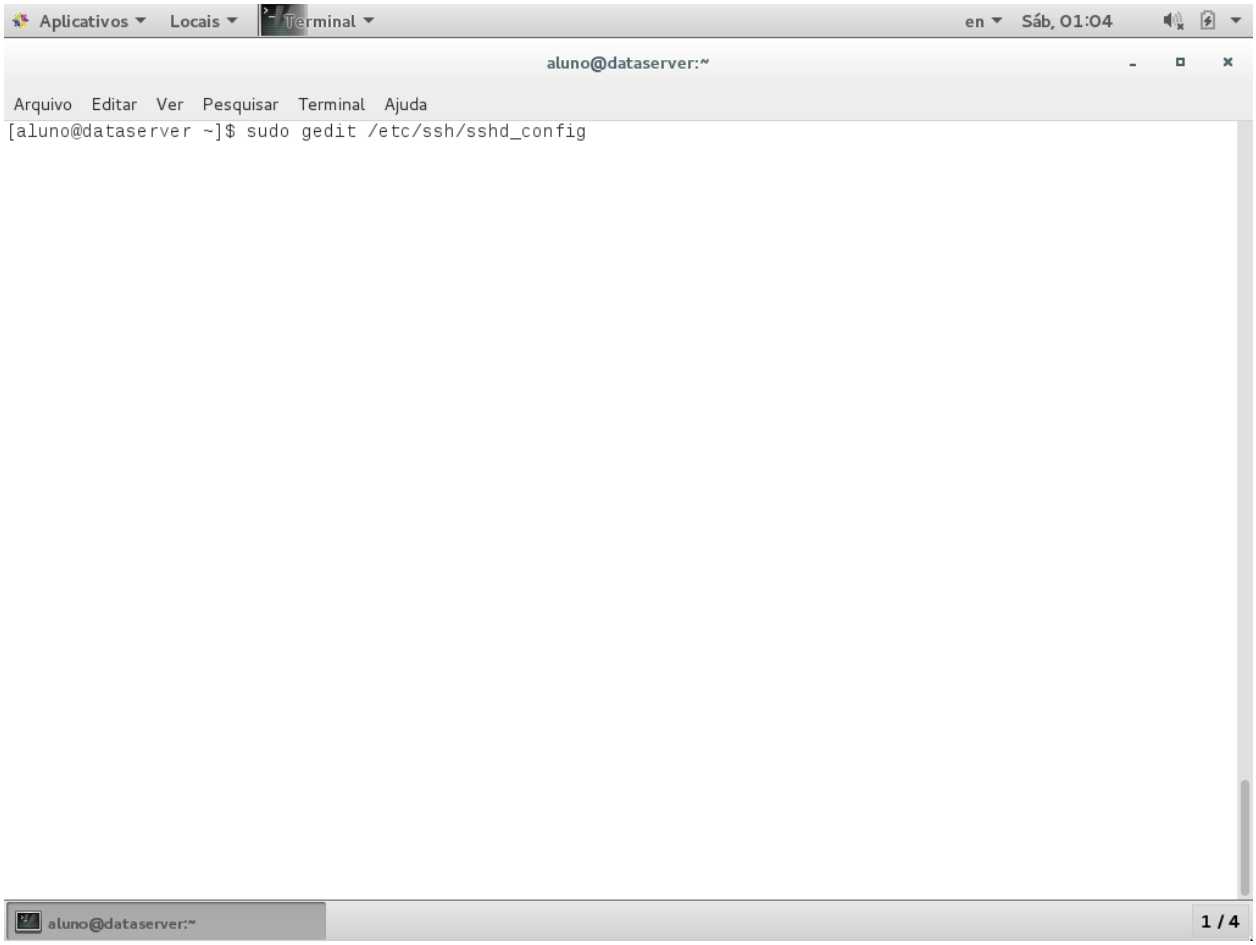


Iniciando o serviço

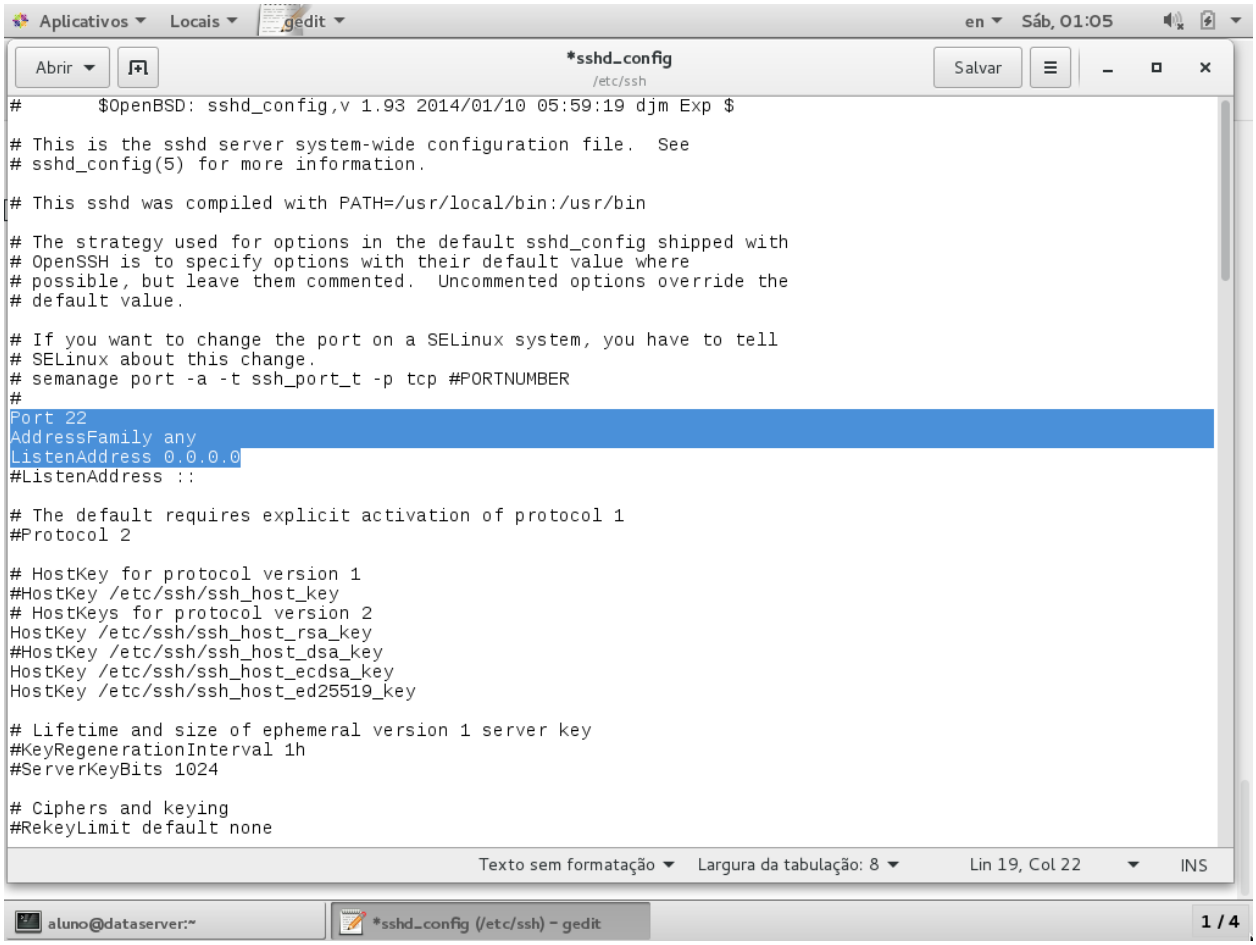
sudo systemctl start sshd



Serviço em execução

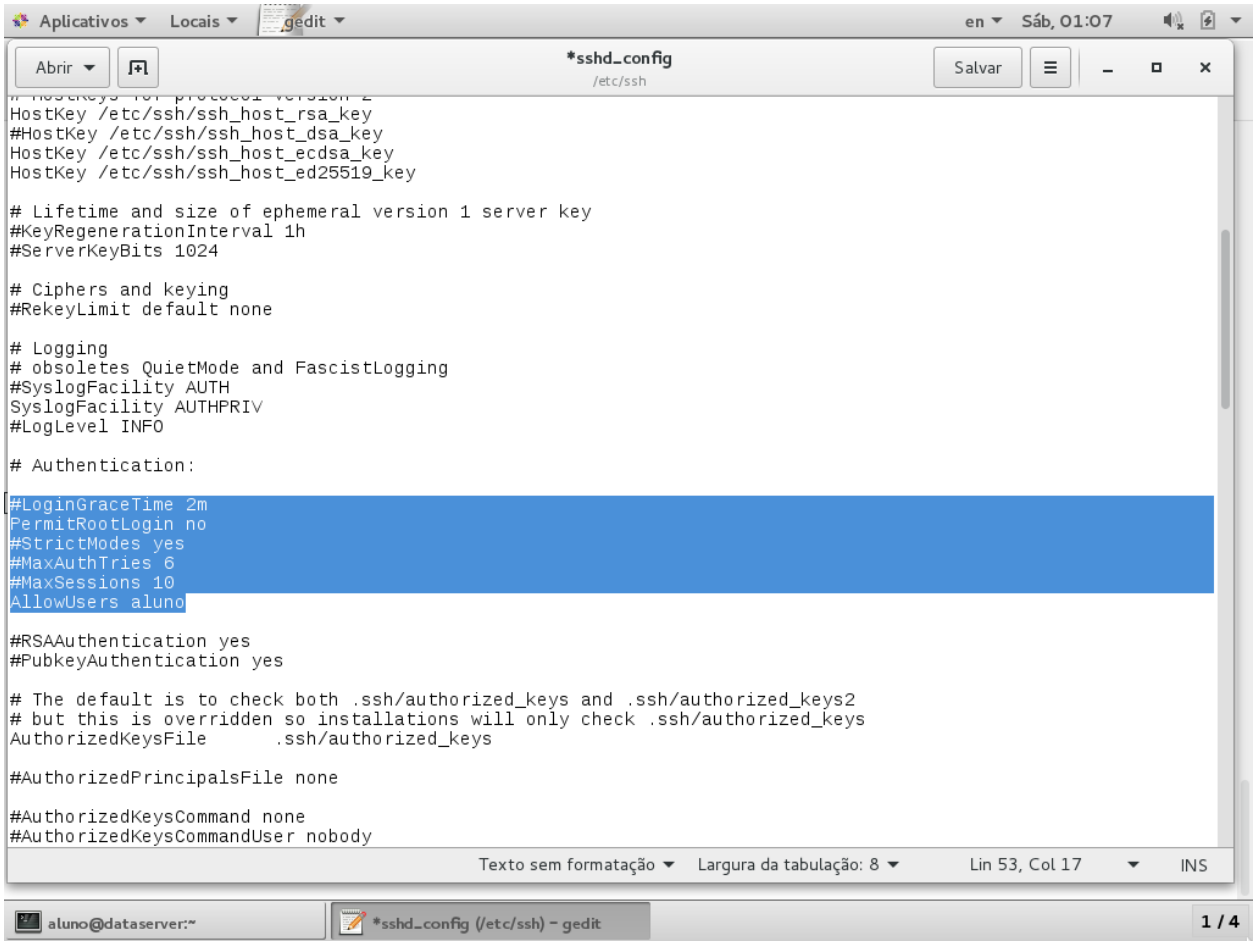


sudo gedit /etc/ssh/sshd_config

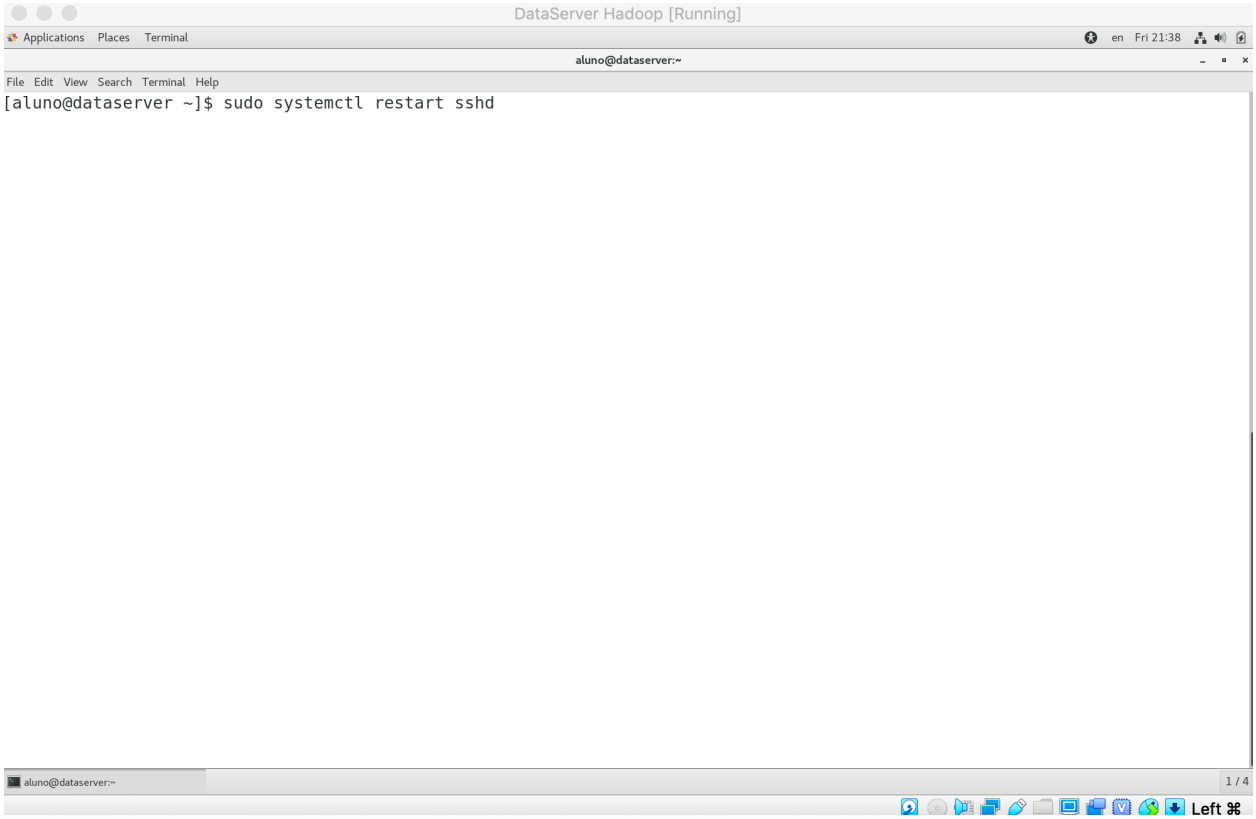


Primeira parte da configuração ssh.

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



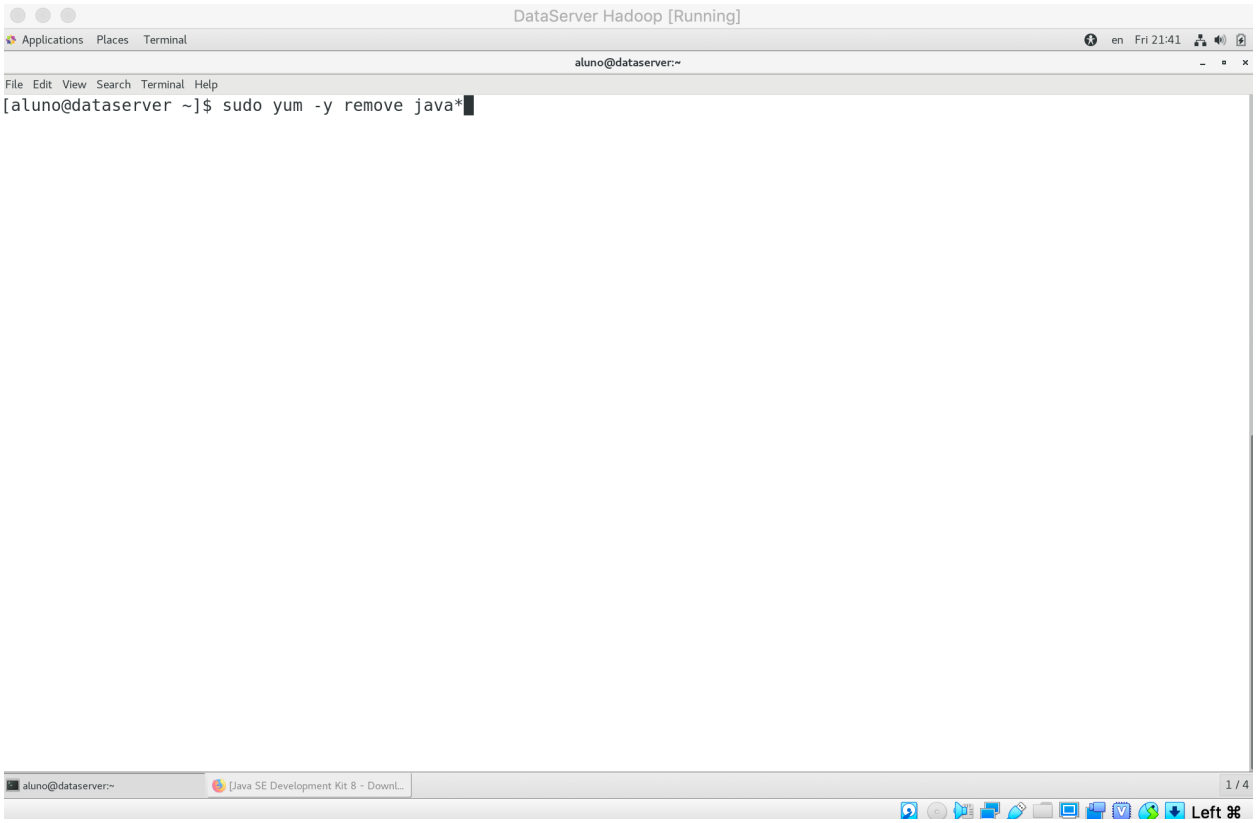
Segunda parte da configuração do ssh



sudo systemctl restart sshd

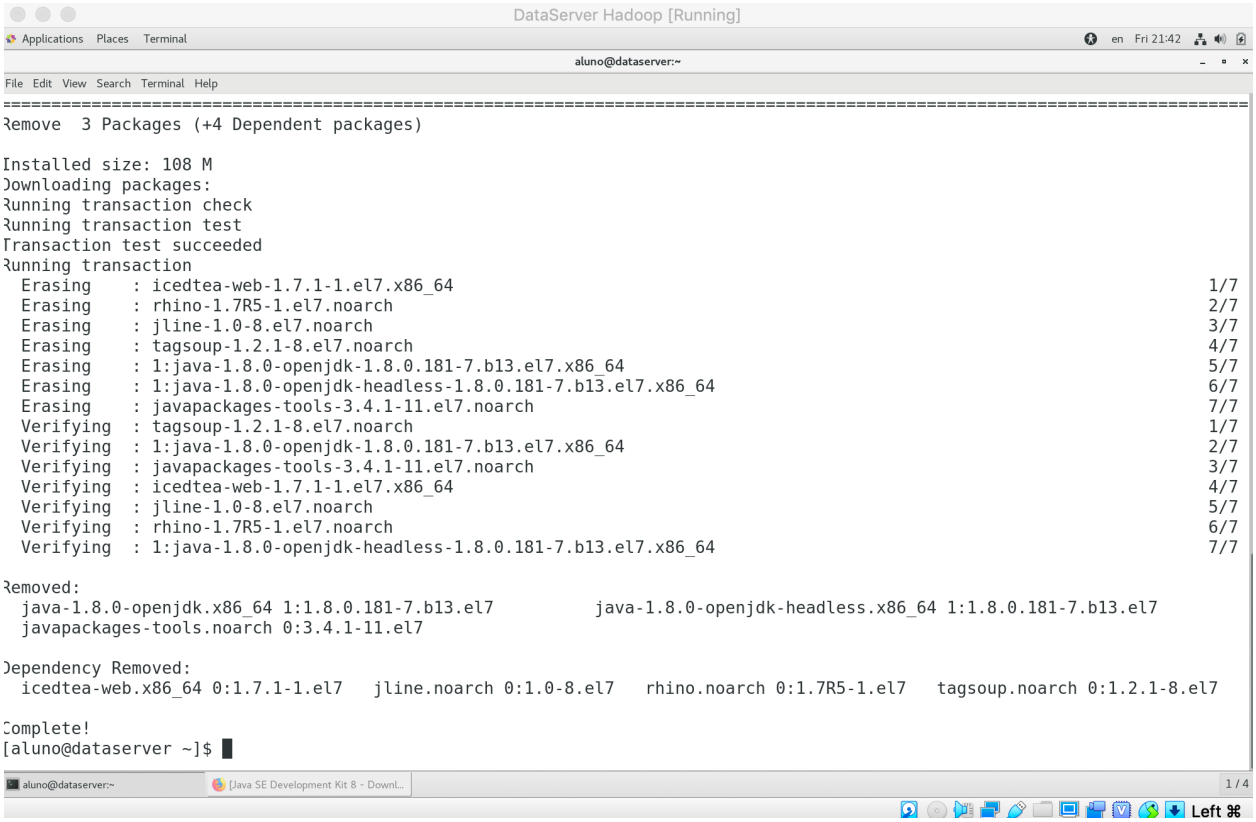
4. Instalação do Java 8

4.1. Removendo o OpenJDK



Removendo o OpenJDK

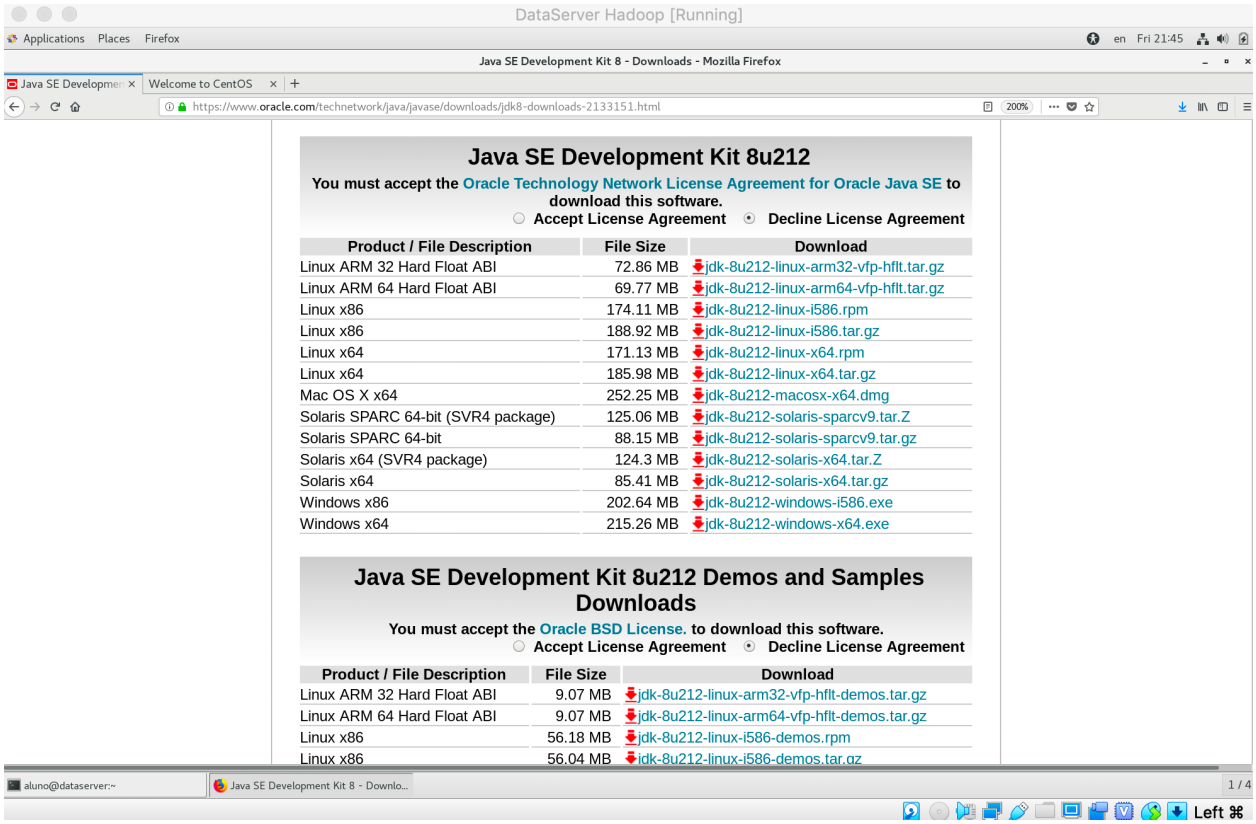
sudo yum -y remove java*



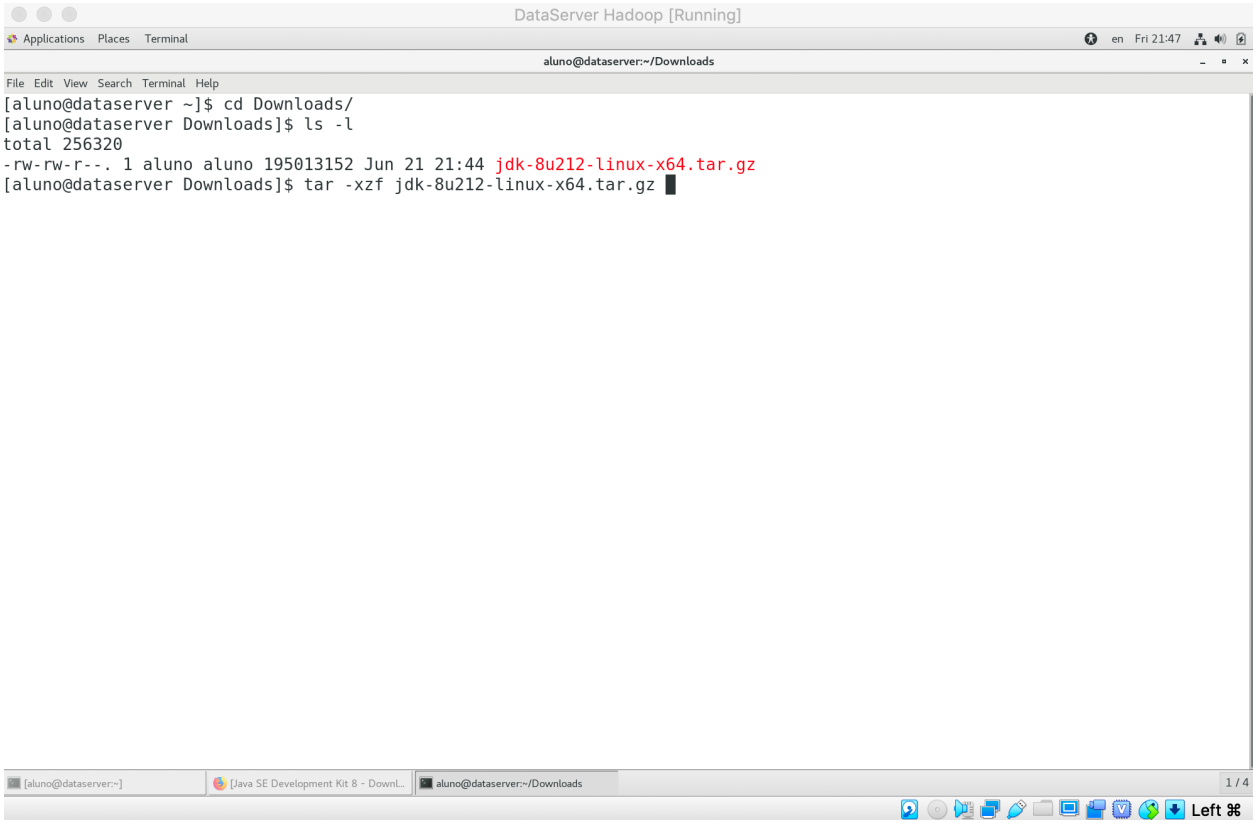
Concluído

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

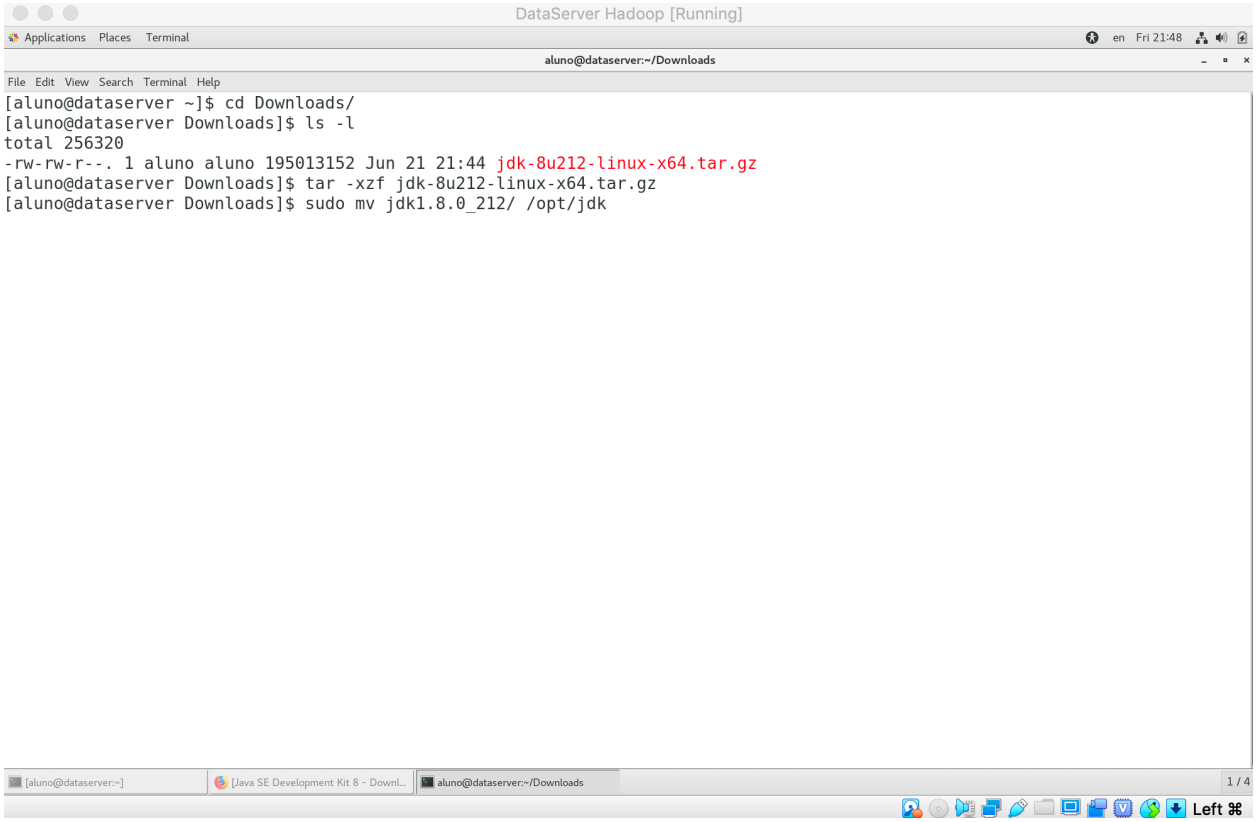
4.2. Instalação do JDK



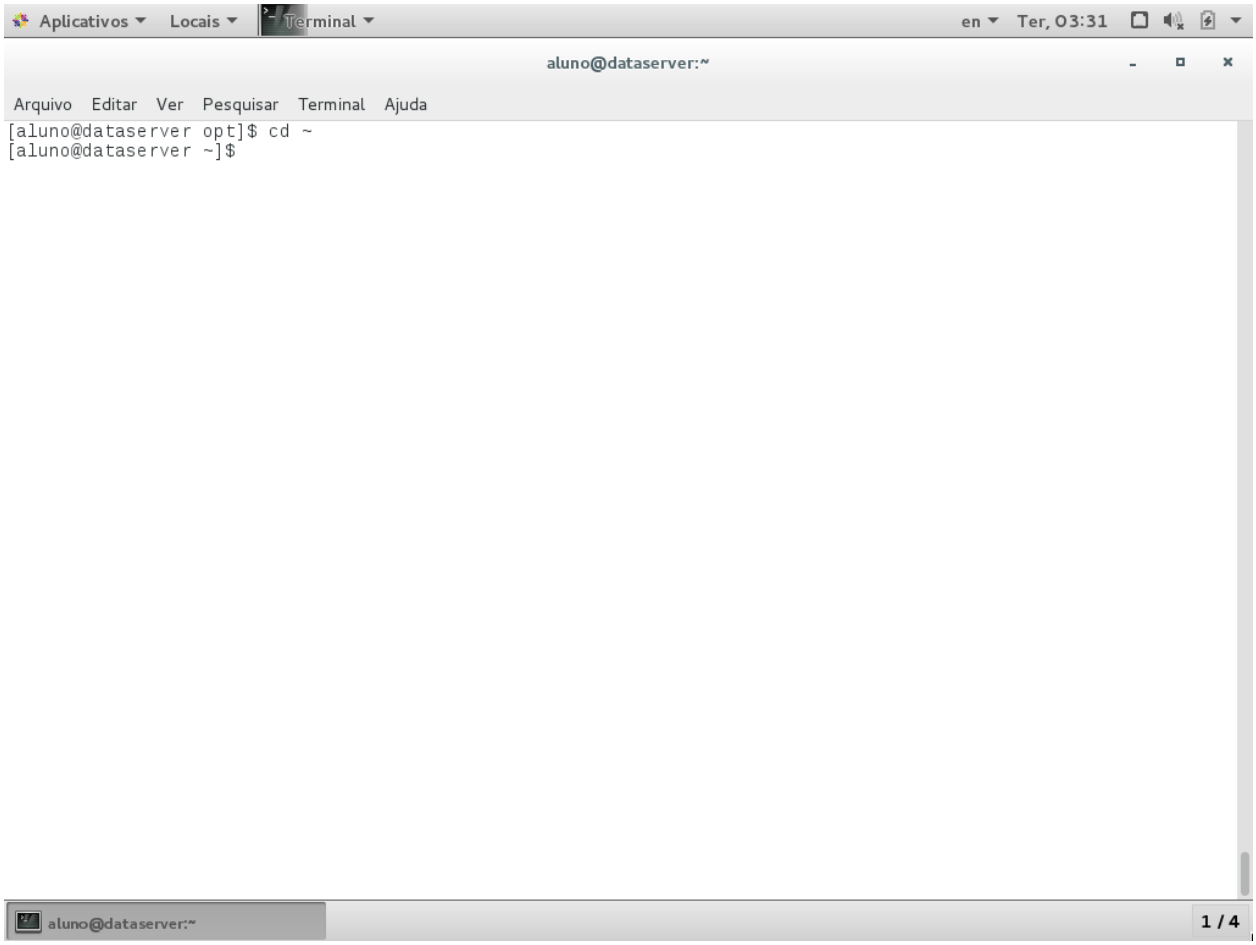
No site da Oracle, fazer o download do JDK



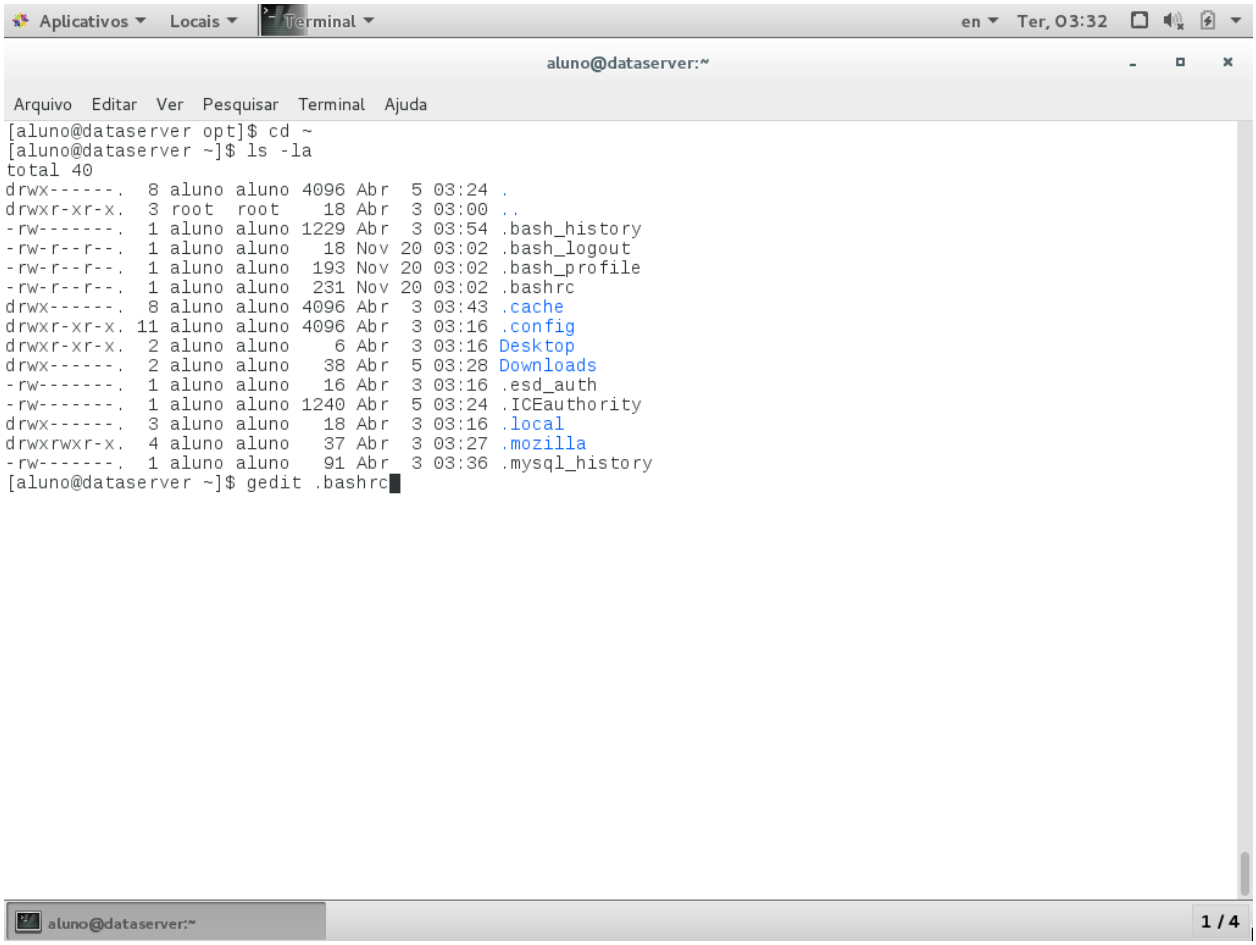
Executar o comando tar para descompactar o arquivo: **tar -xzf jdk-8u212-linux-x64.tar.gz**



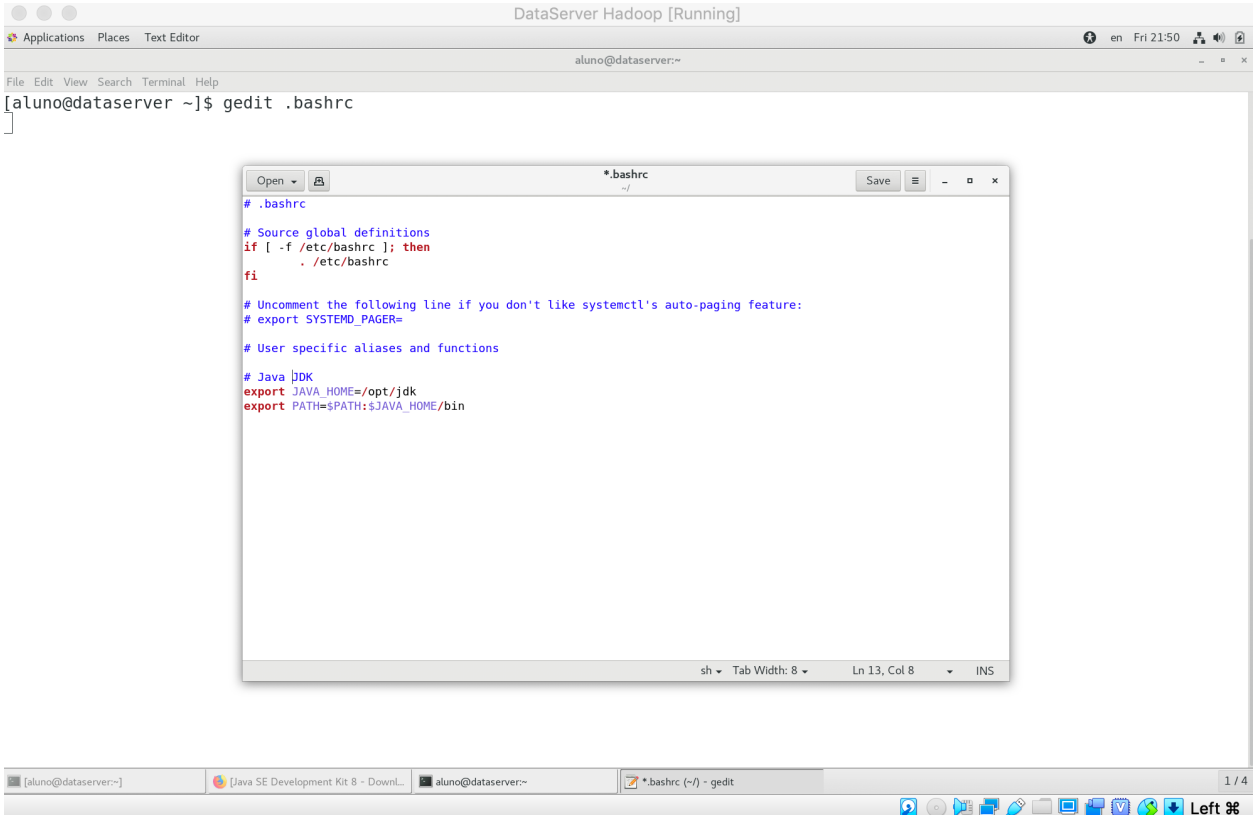
Mover o diretório do JDK


Aplicativos
Locais
Terminal
en
Ter, 03:31
aluno@dataserver:~
Arquivo Editar Ver Pesquisar Terminal Ajuda
[aluno@dataserver opt]$ cd ~
[aluno@dataserver ~]$
aluno@dataserver:~
1 / 4

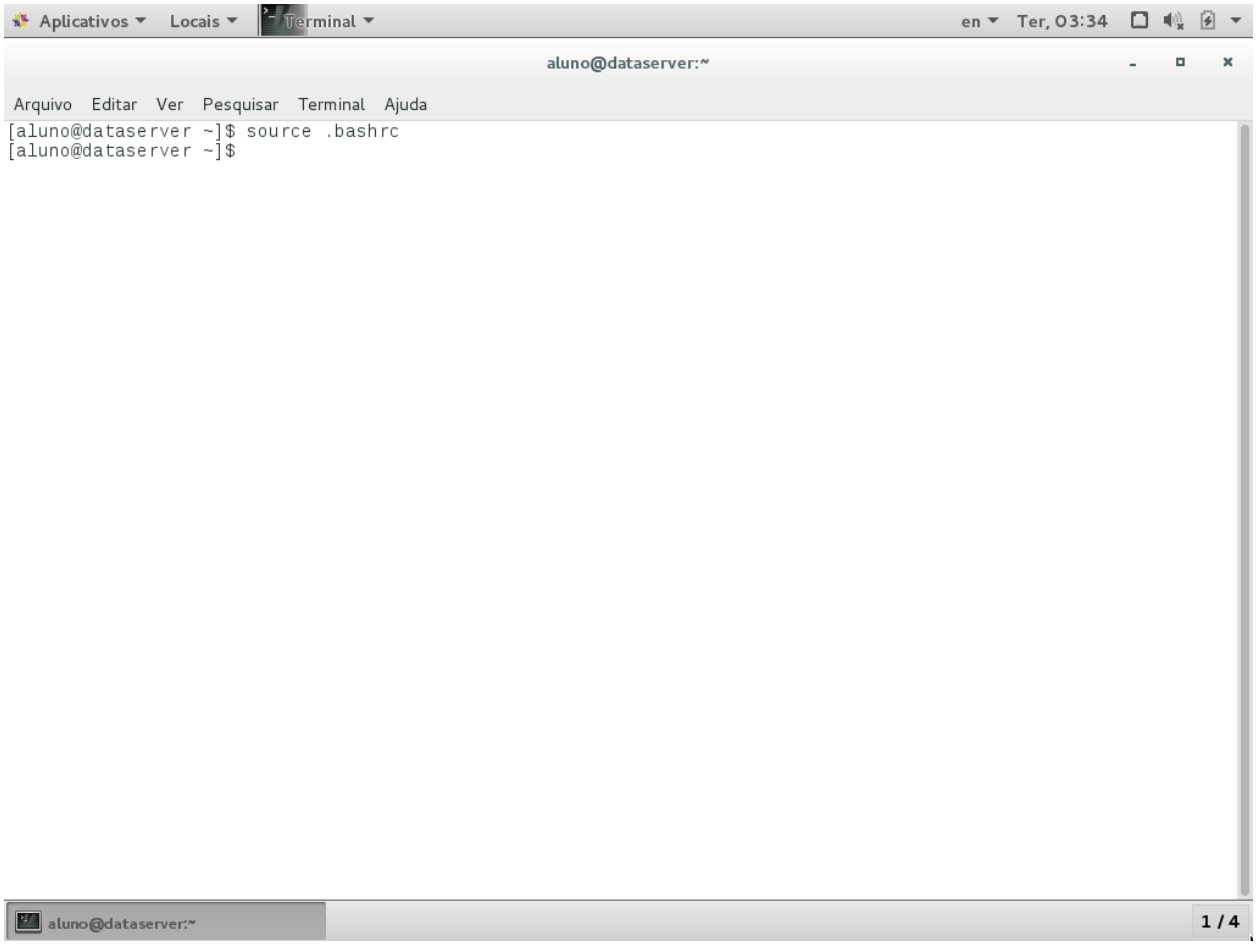
cd ~



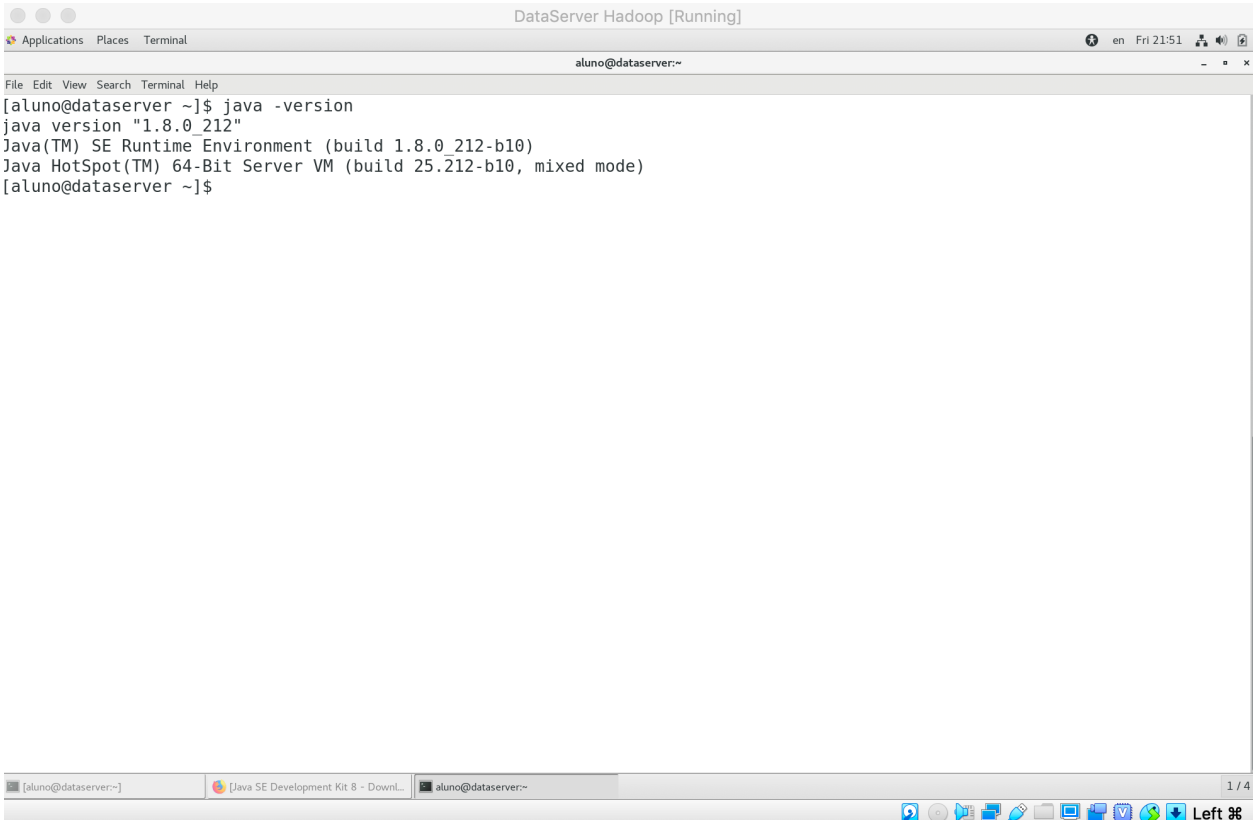
gedit .bashrc



Editar as variáveis de ambiente conforme acima e salvar o arquivo



source .bashrc



Checando a versão do Java JDK

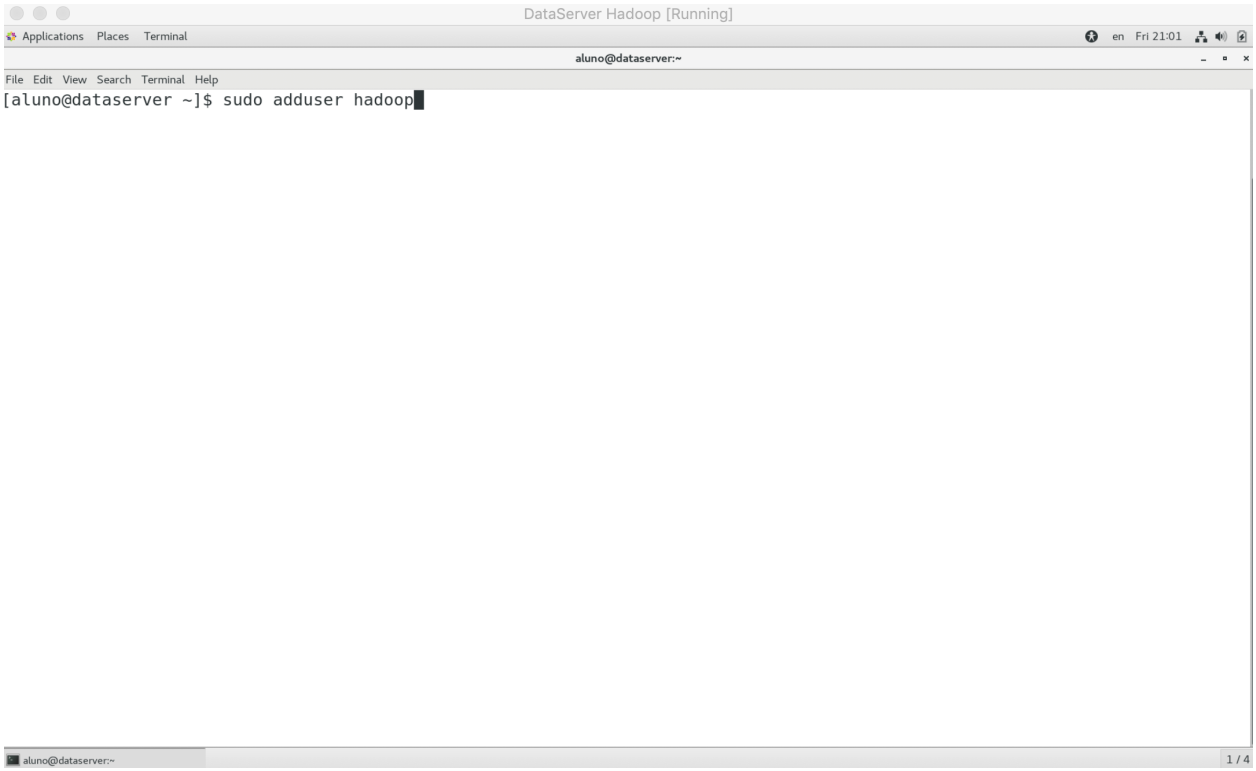
Segundo checkpoint:

Desligue a VM (clique no ícone da bateria e então em desligar).
Clique no menu File do VirtualBox e clique em Export Appliance.
Será gerada uma cópia de segurança da sua máquina virtual.

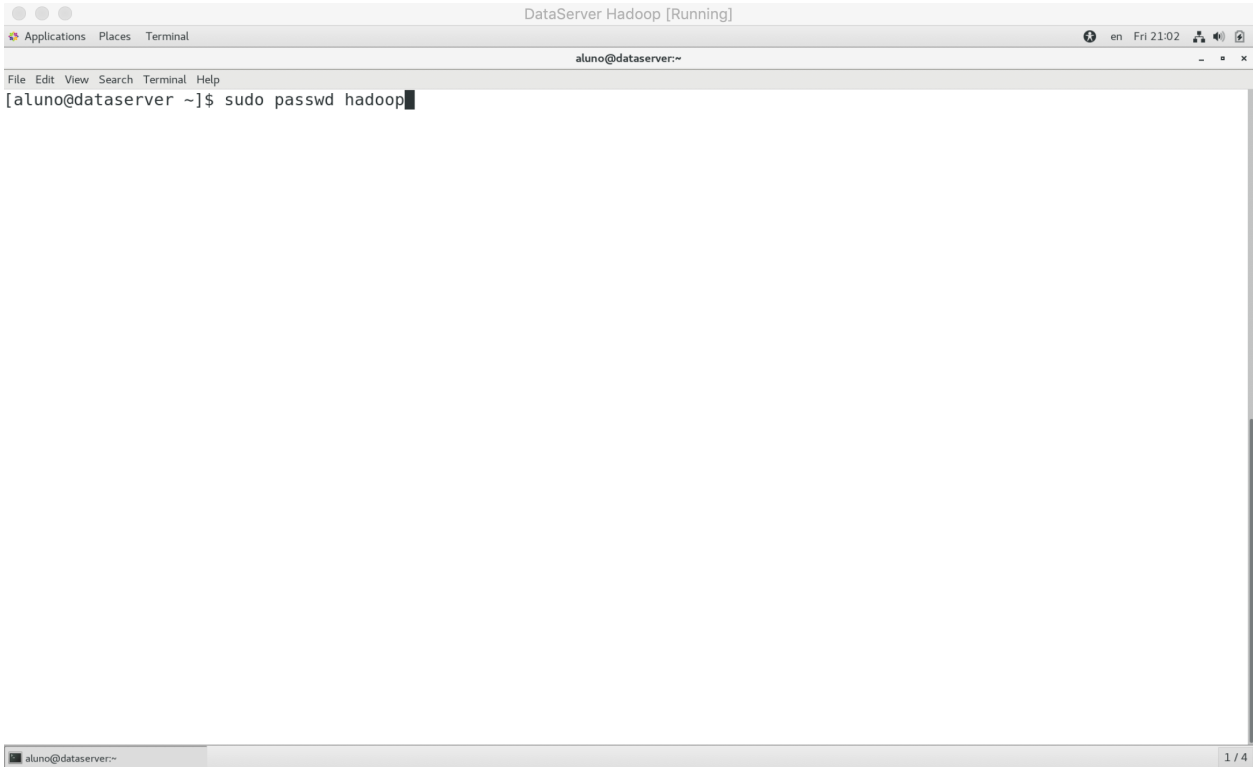
→ VM: DataServer-Hadoop-v2.0.ova

5. Instalação e Configuração do Hadoop

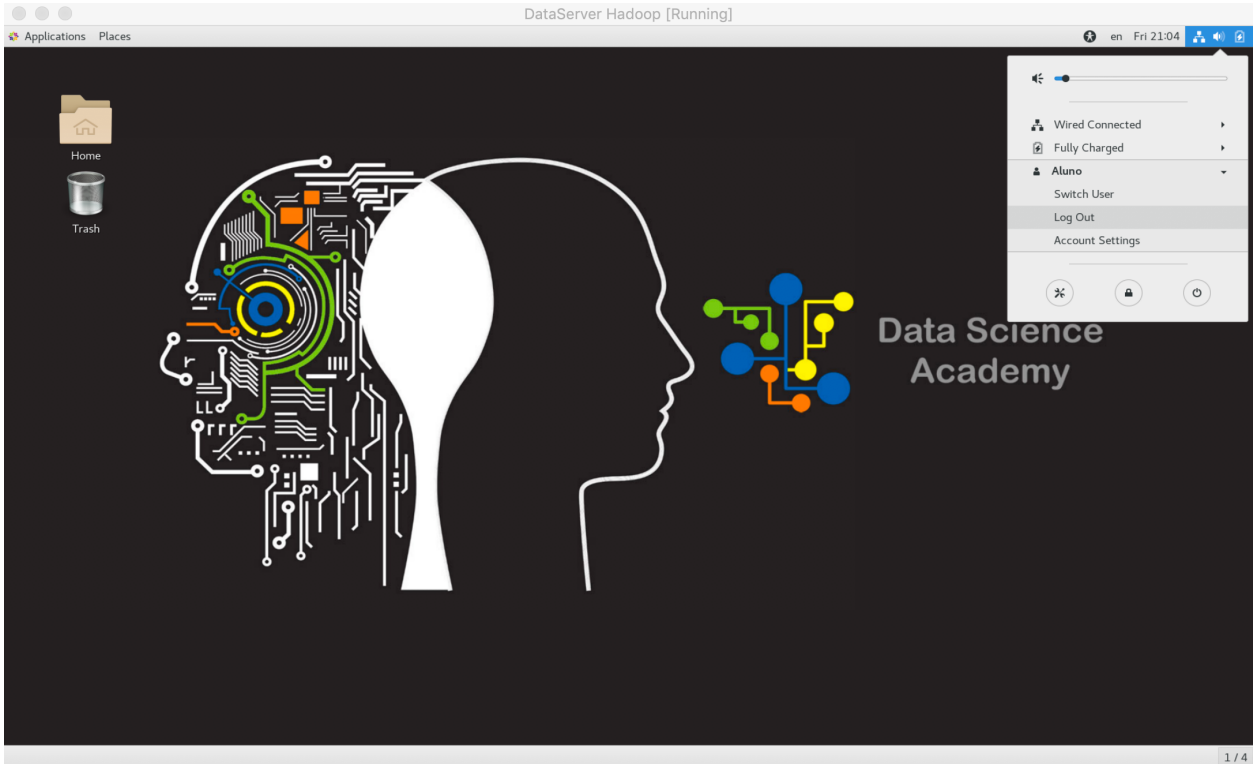
5.1. Criando o usuário hadoop



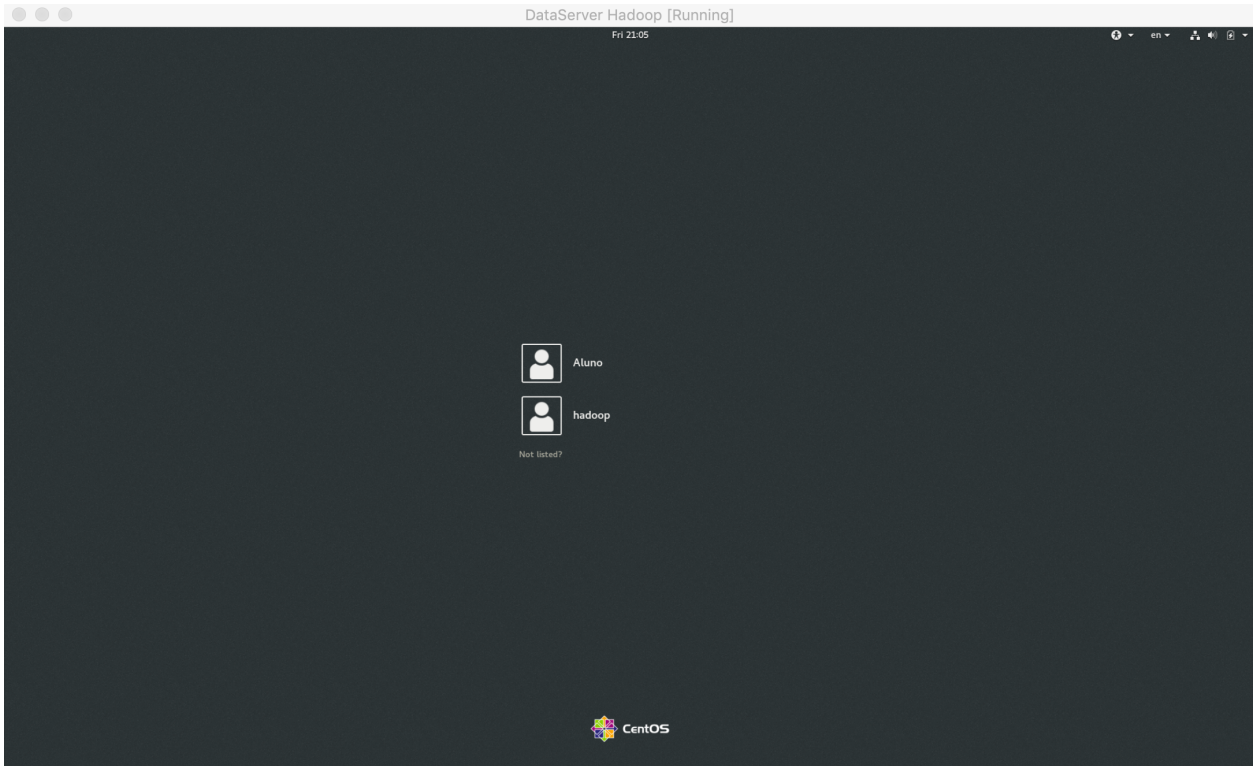
sudo adduser hadoop – para criar o usuário hadoop



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



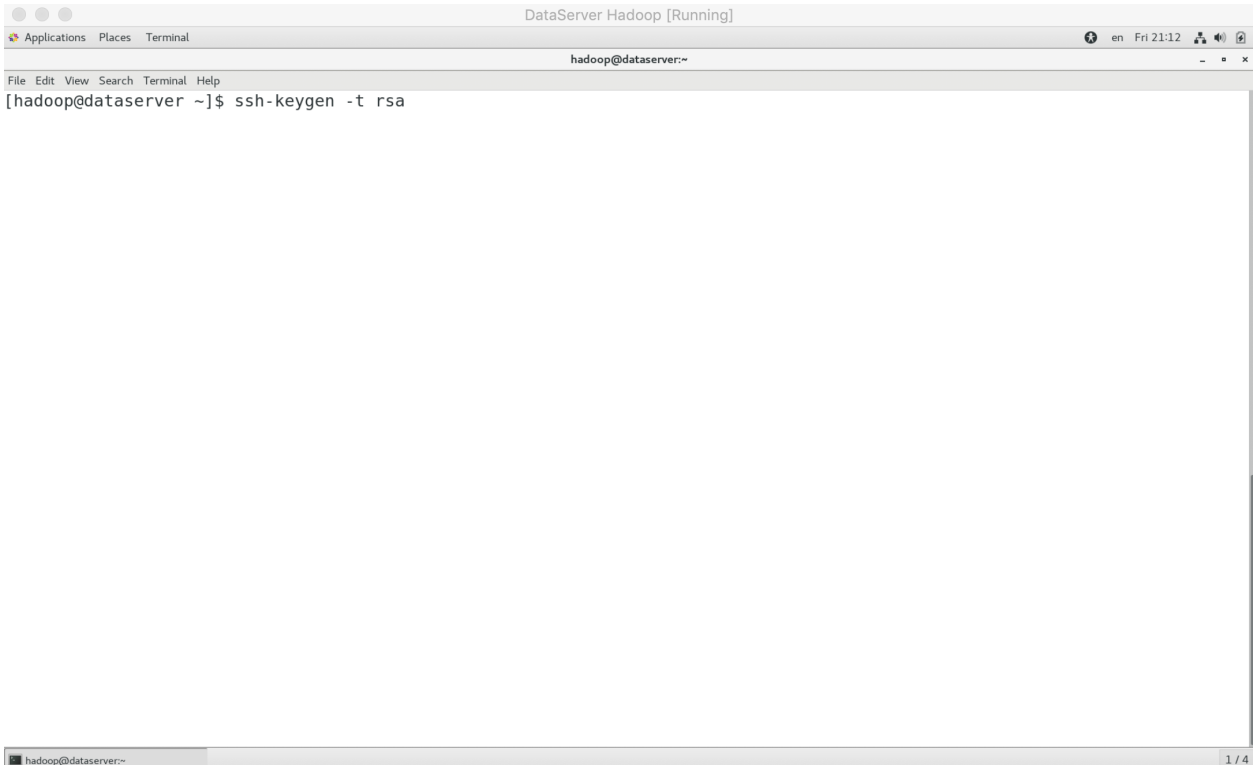
Efetue logout como usuário aluno



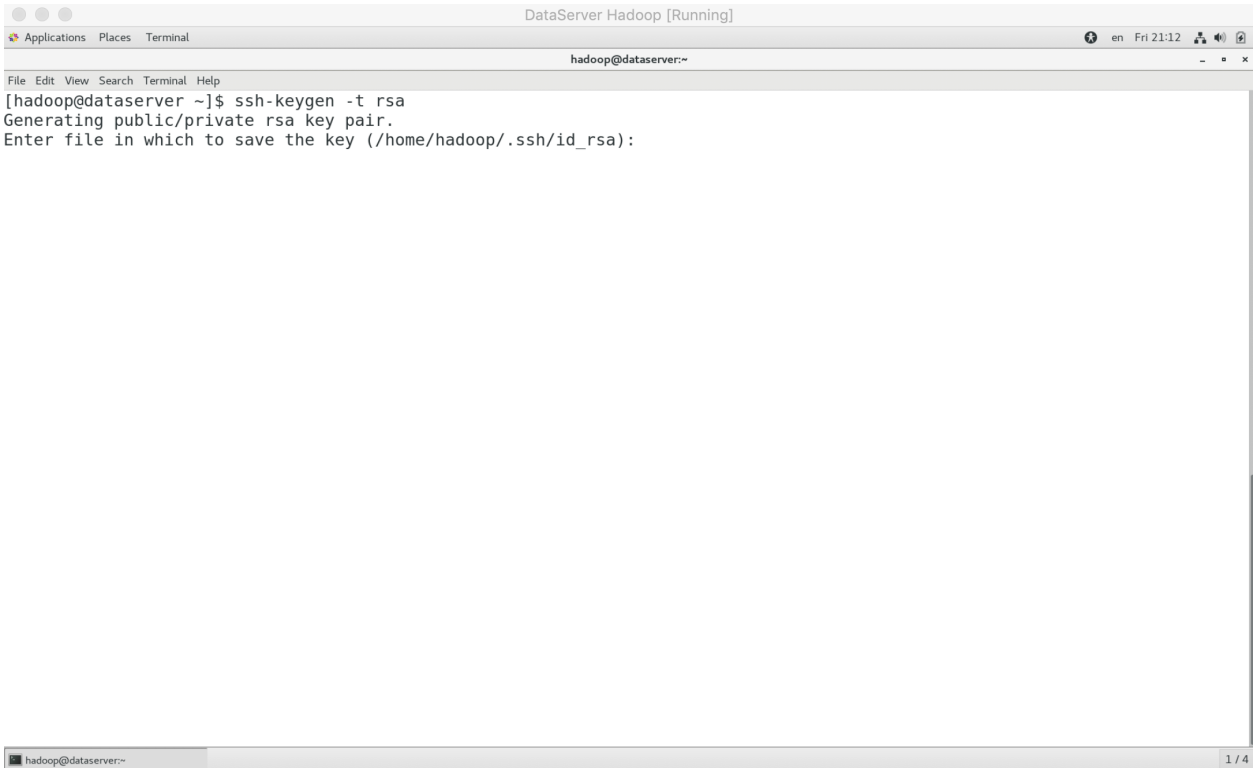
E efetue login como usuário hadoop

Adicione o usuário hadoop no arquivo /etc/sudoers conforme você fez com o usuário aluno

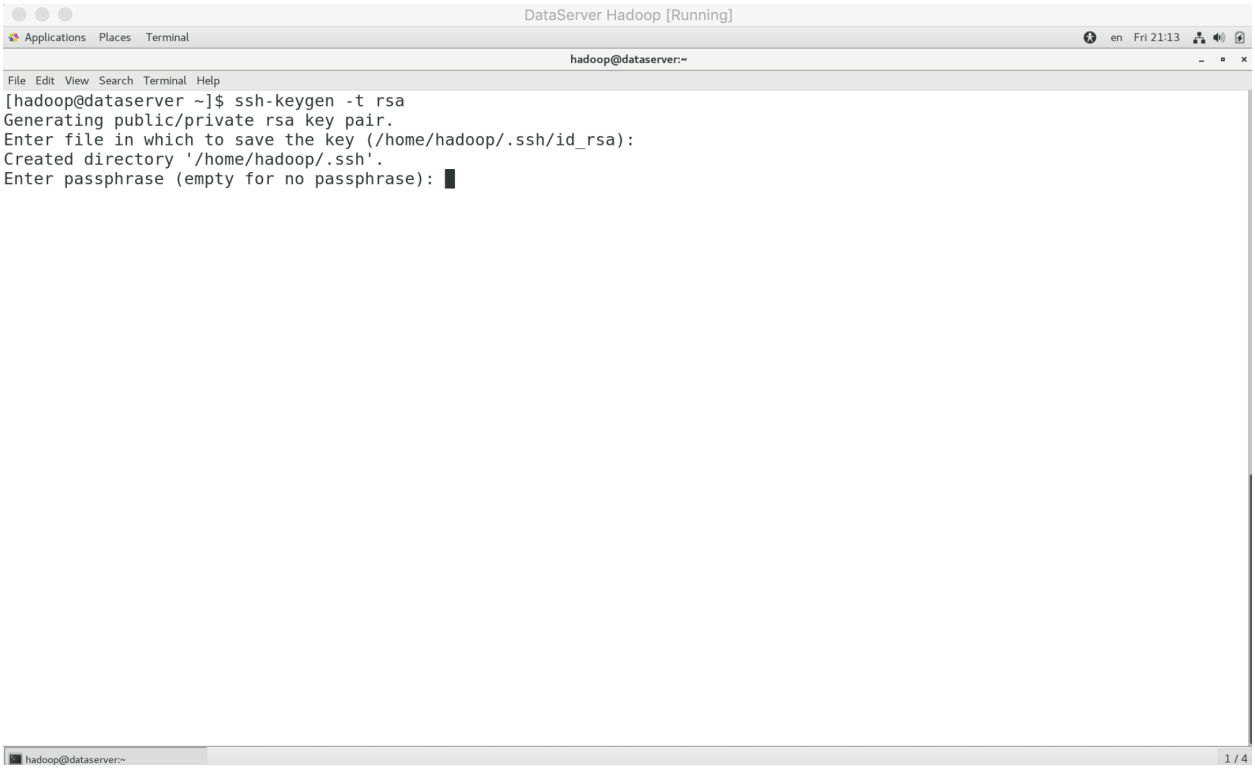
5.2. Configuração do ssh sem senha



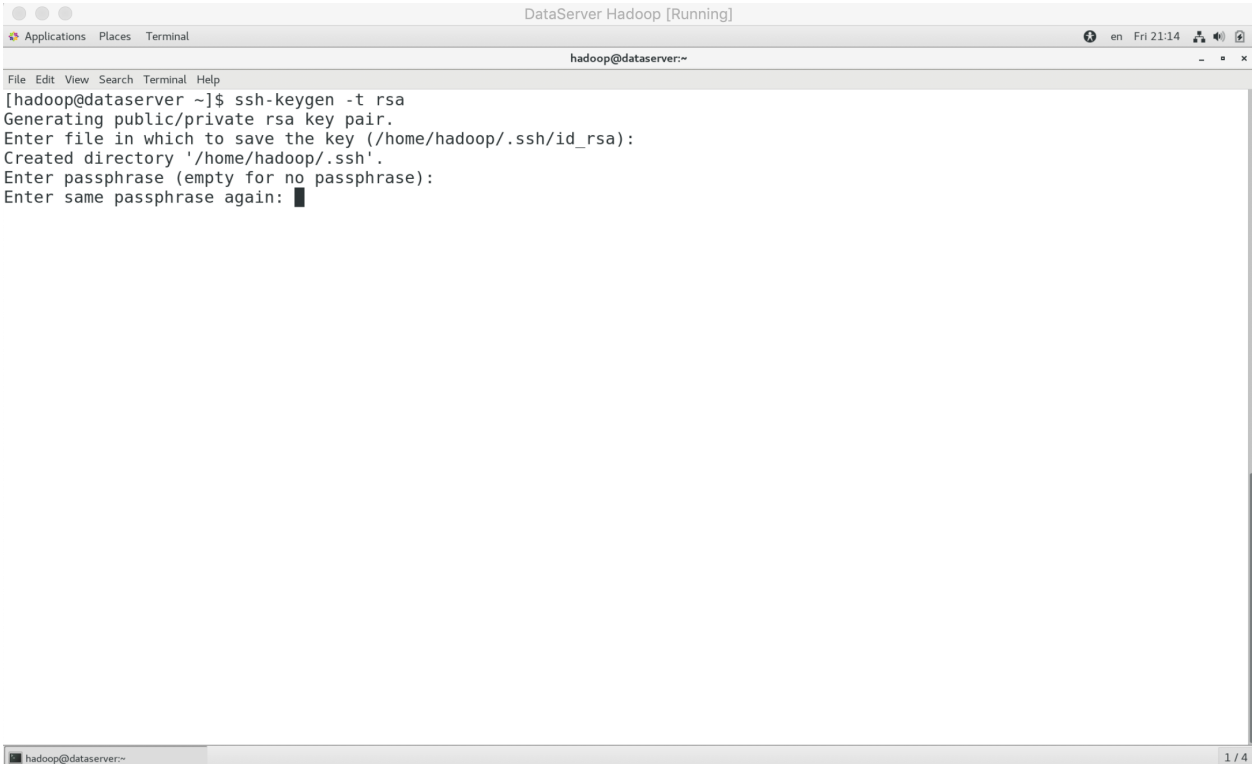
ssh-keygen –t rsa



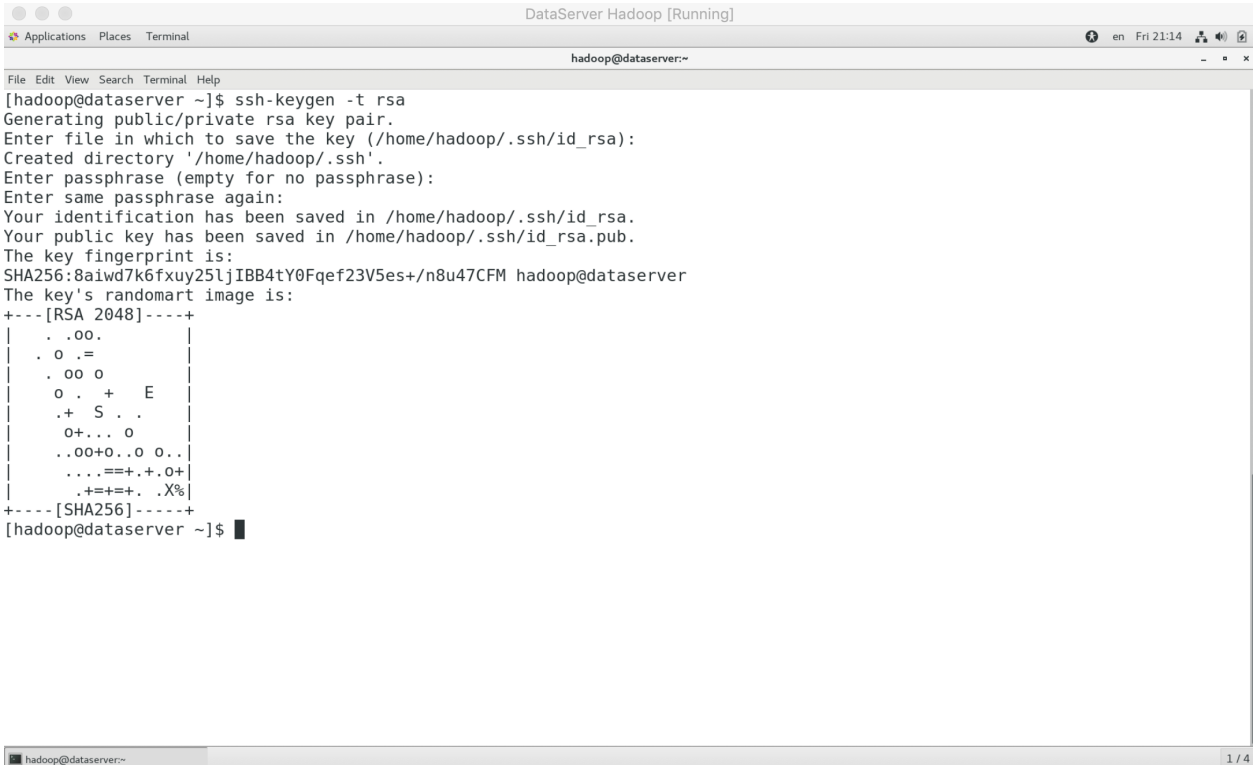
Pressionar Enter para confirmar o diretório onde as chaves serão geradas



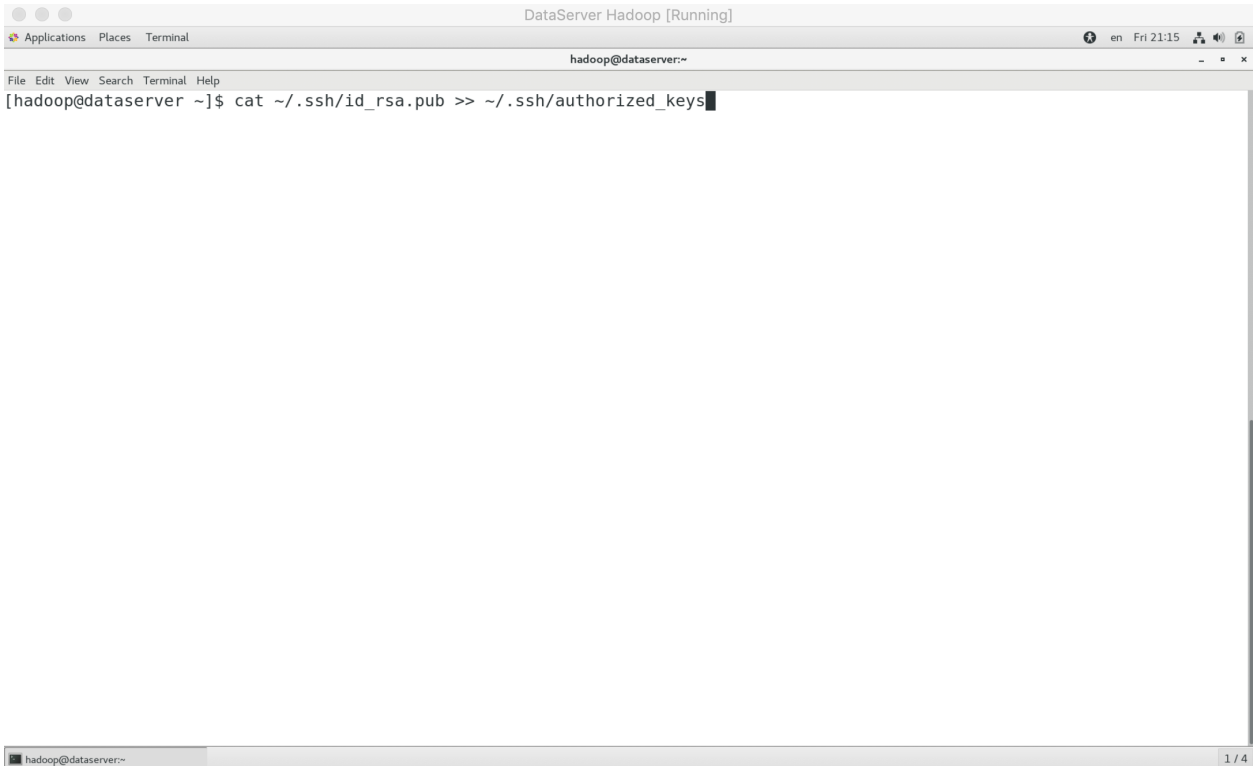
Pressionar Enter



Pressionar Enter novamente

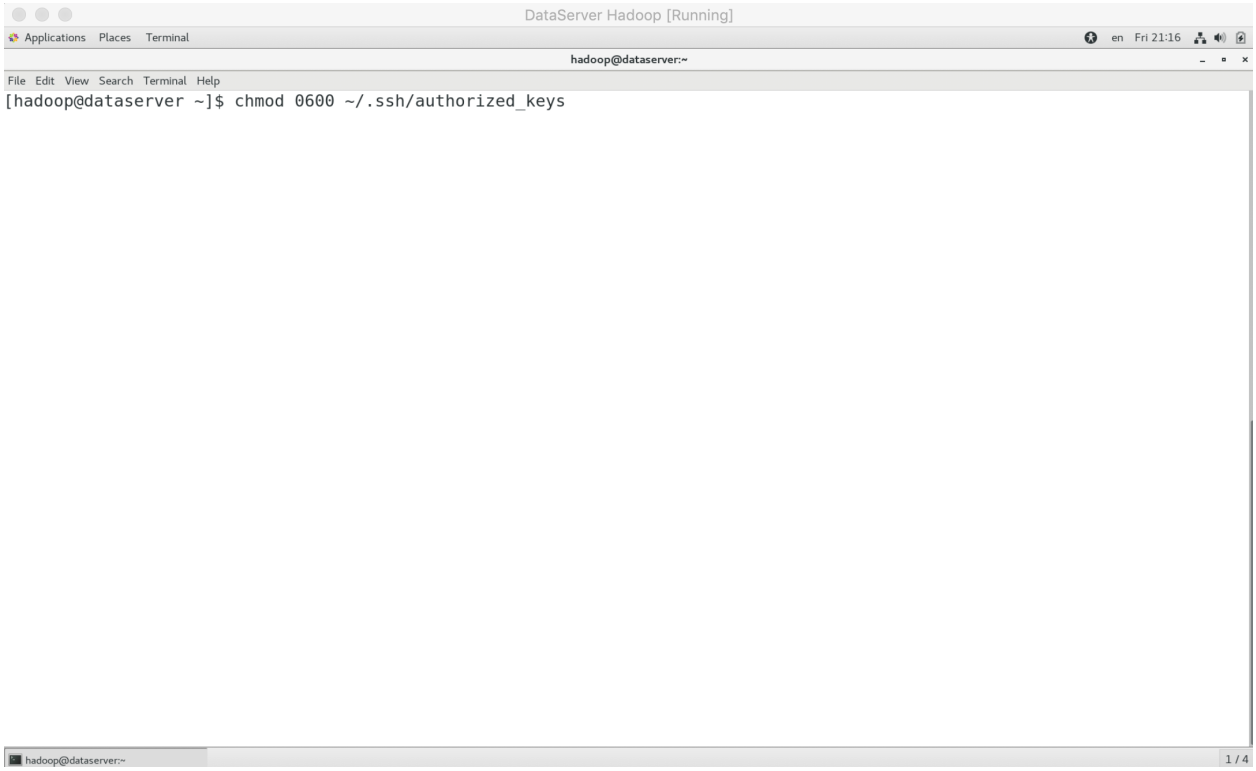


Chaves de segurança geradas



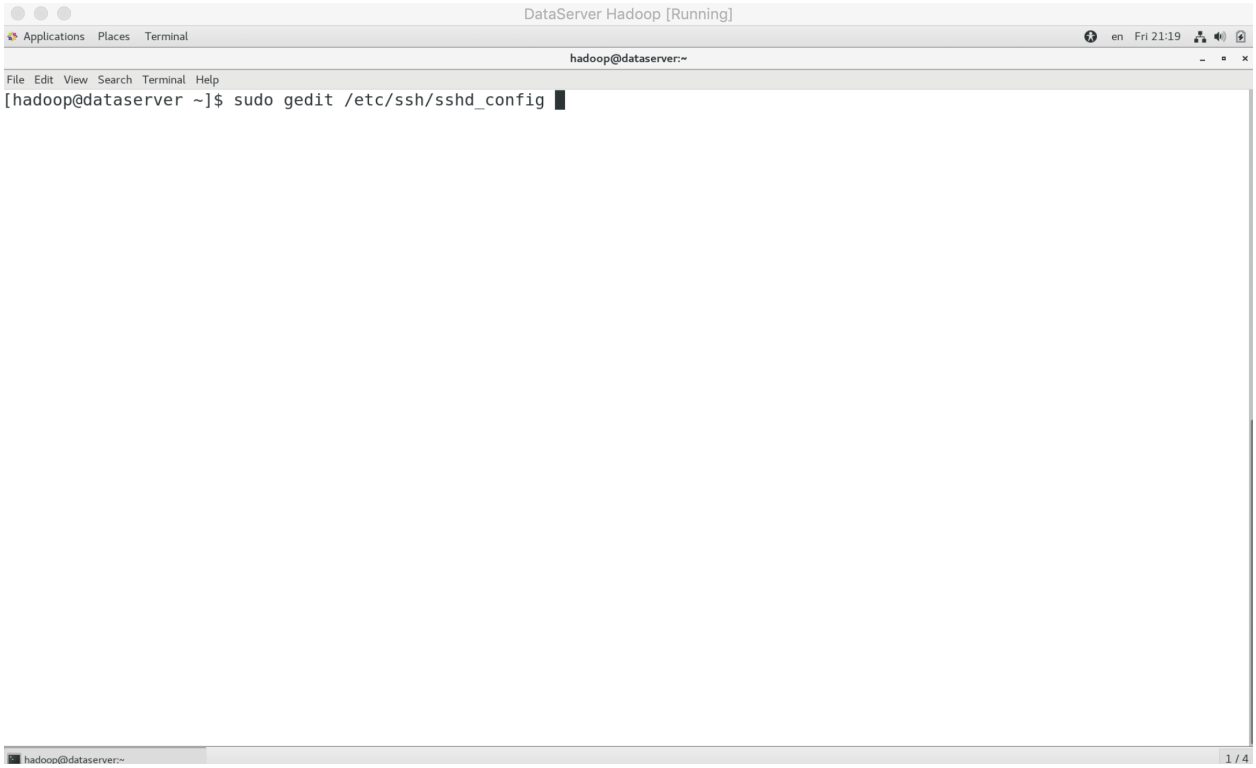
cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys

Esse comando cópia a chave pública para o arquivo authorized_keys do ssh



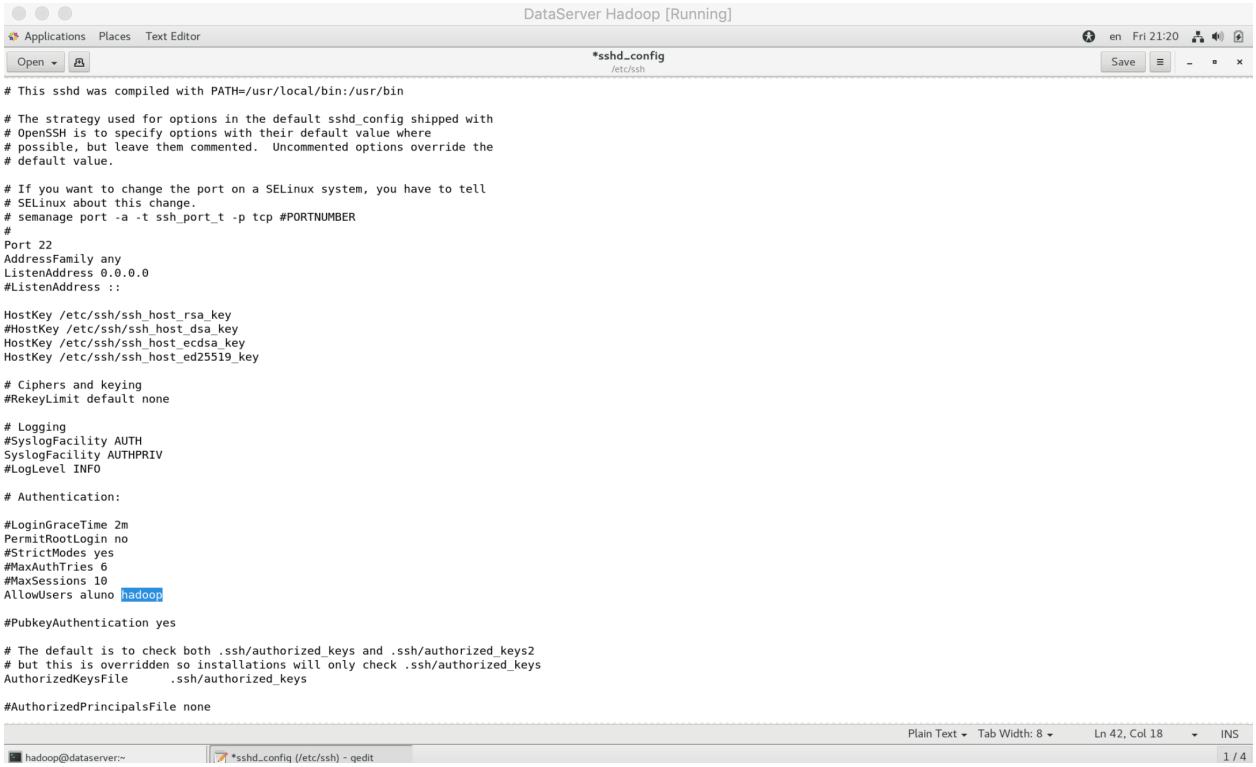
chmod 0600 ~/.ssh/authorized_keys

Esse comando define a permissão do arquivo authorized_keys

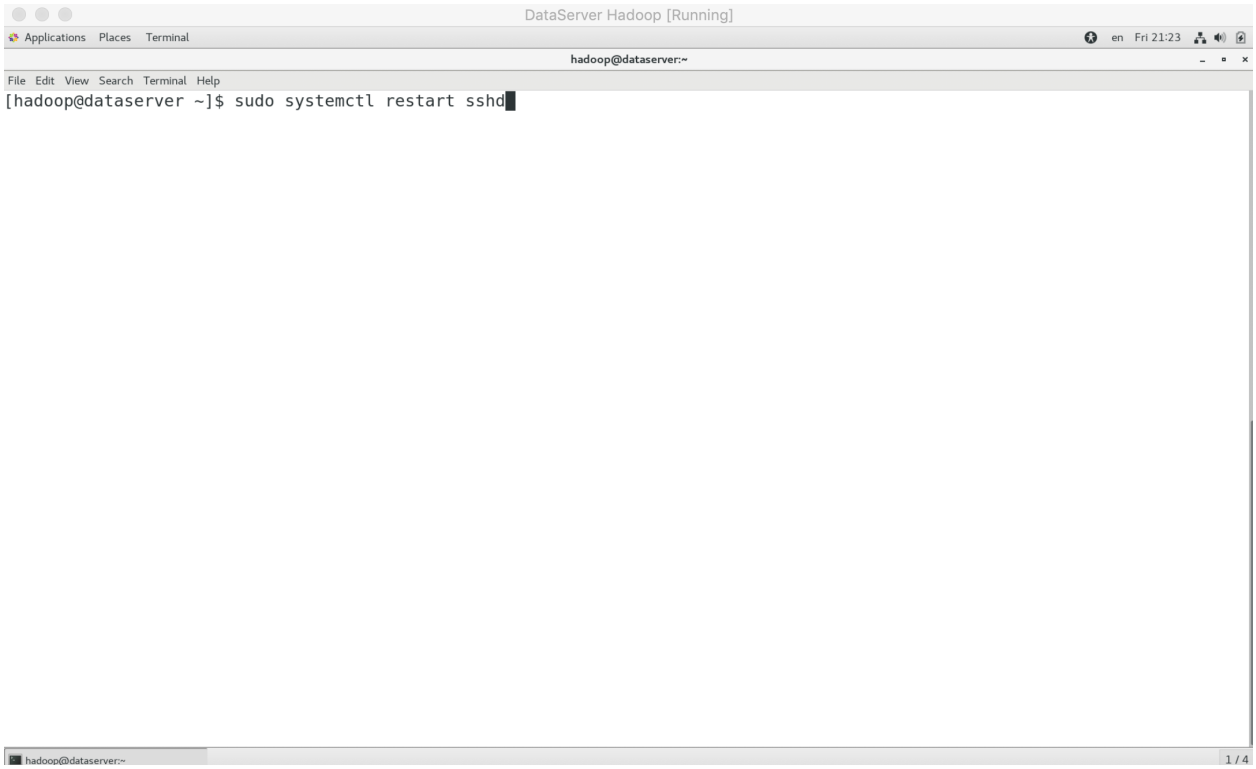


sudo gedit /etc/ssh/sshd_config

Edite o arquivo de configuração do ssh

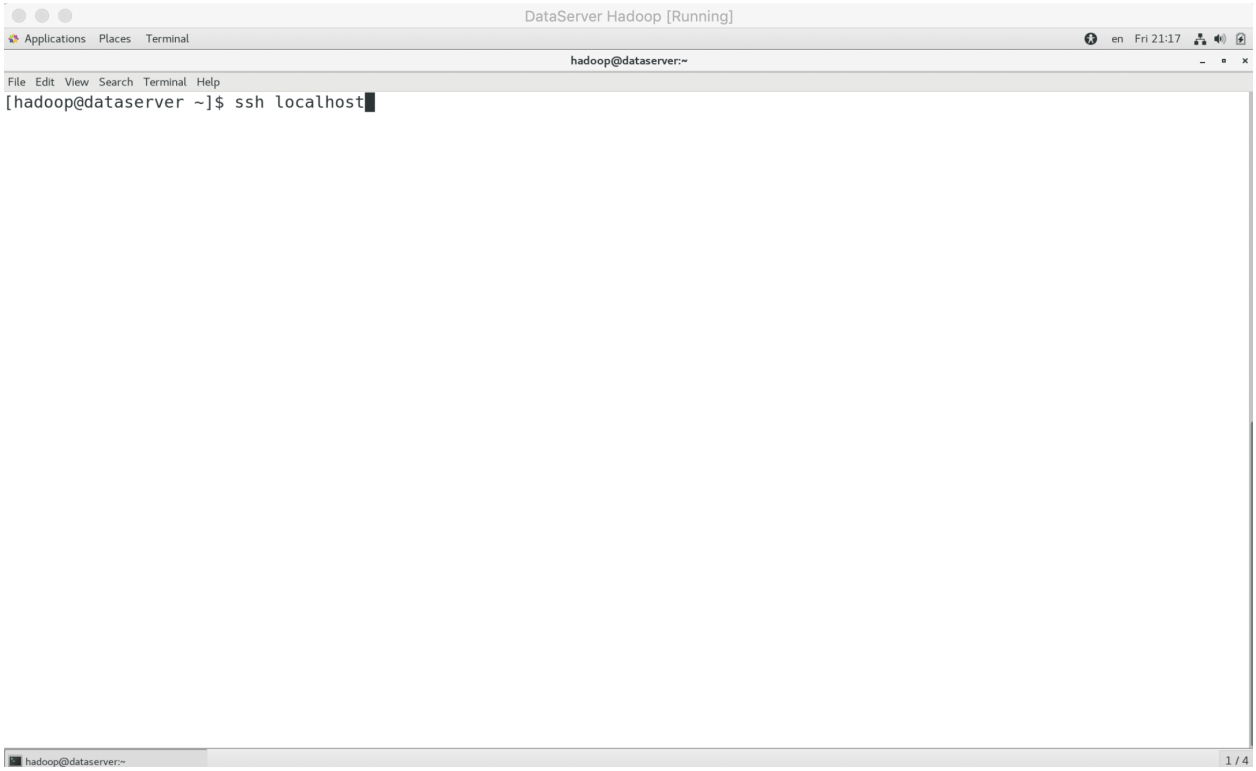


Inclua o usuário hadoop na linha AllowUsers, salve o arquivo e feche-o

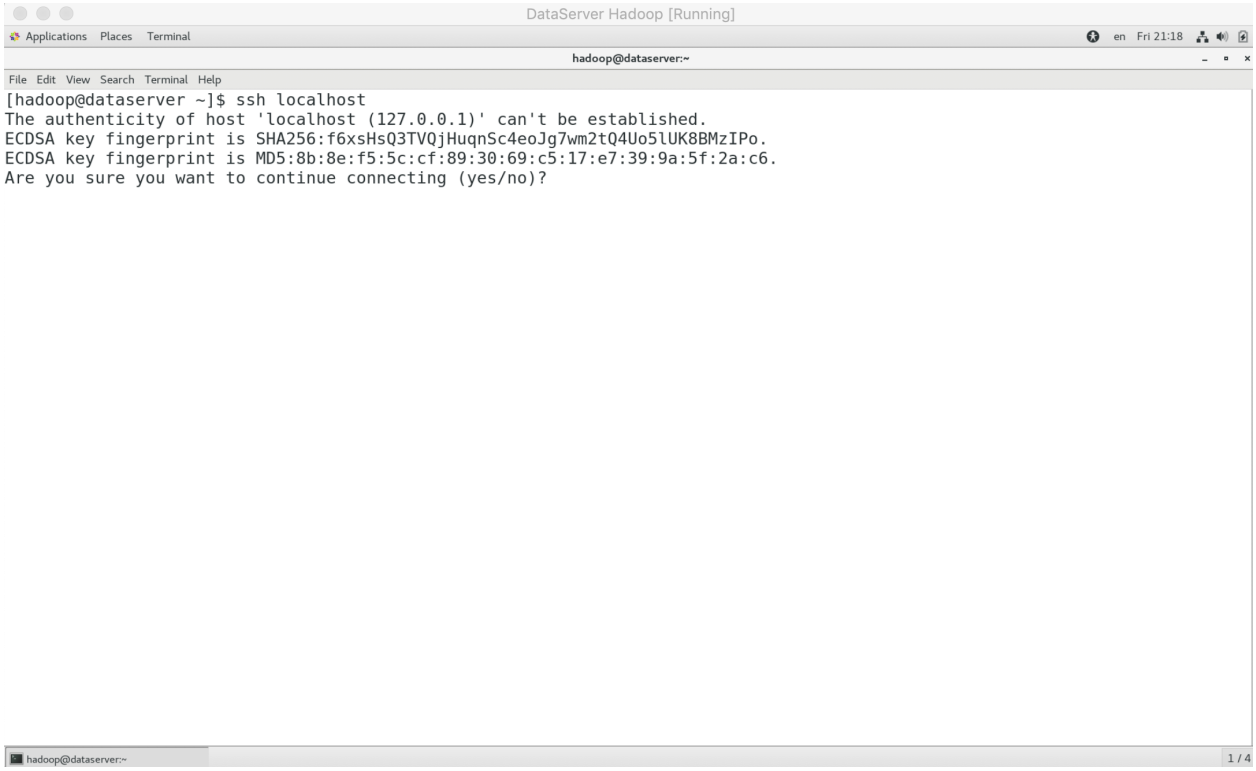


sudo systemctl restart sshd

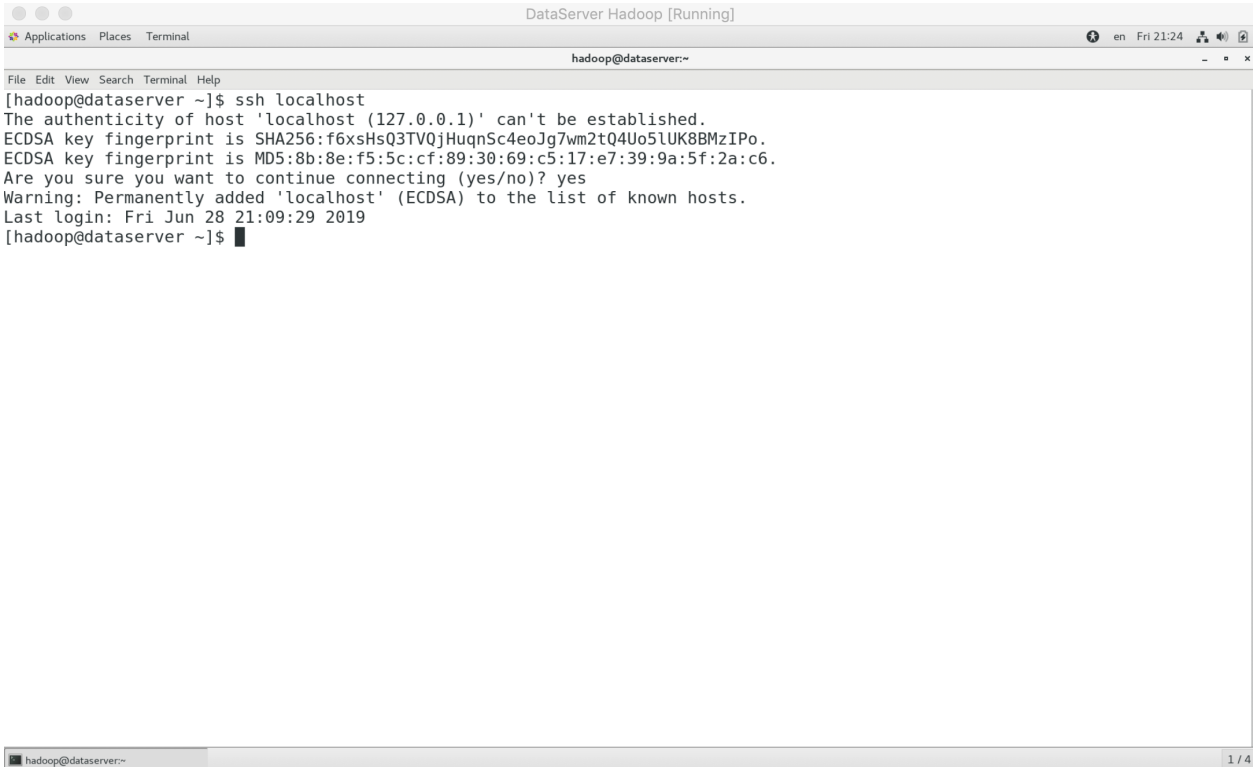
Reinicie o serviço ssh



ssh localhost



Yes

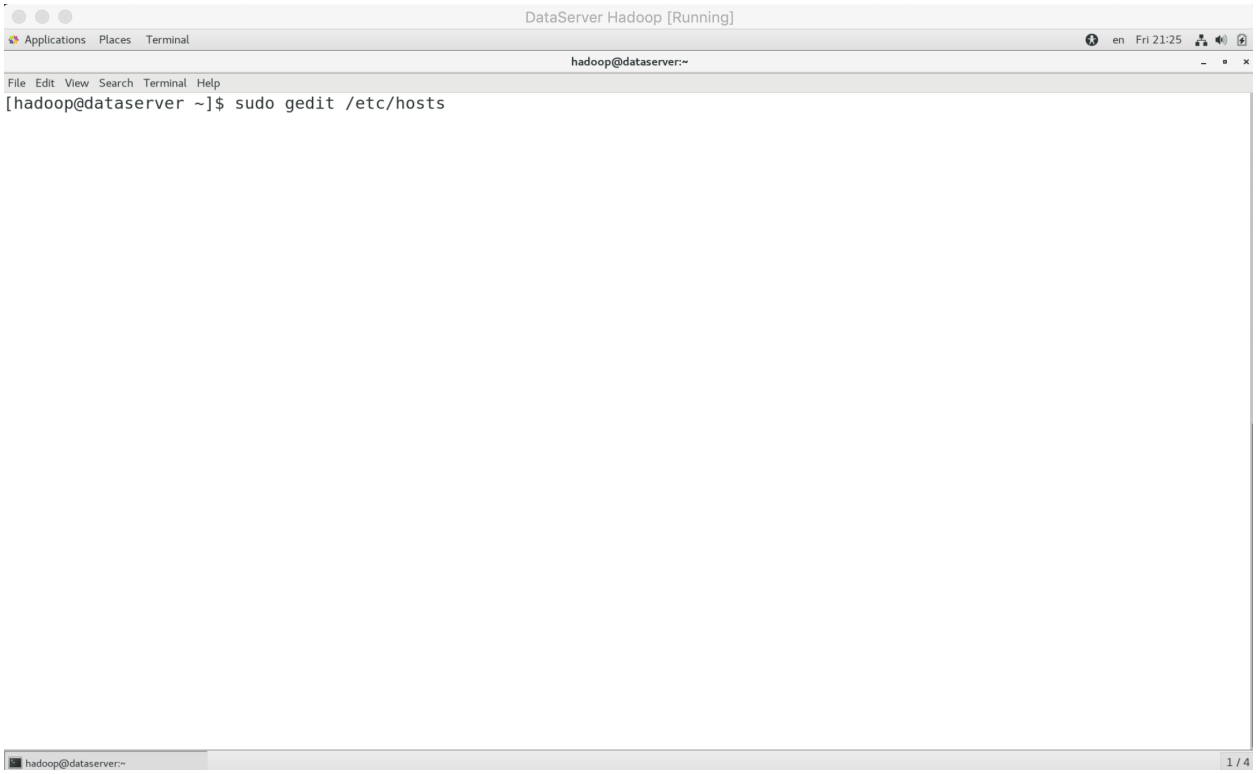


Conexão ssh sem senha efetuada com sucesso. Digite exit e pressione Enter.

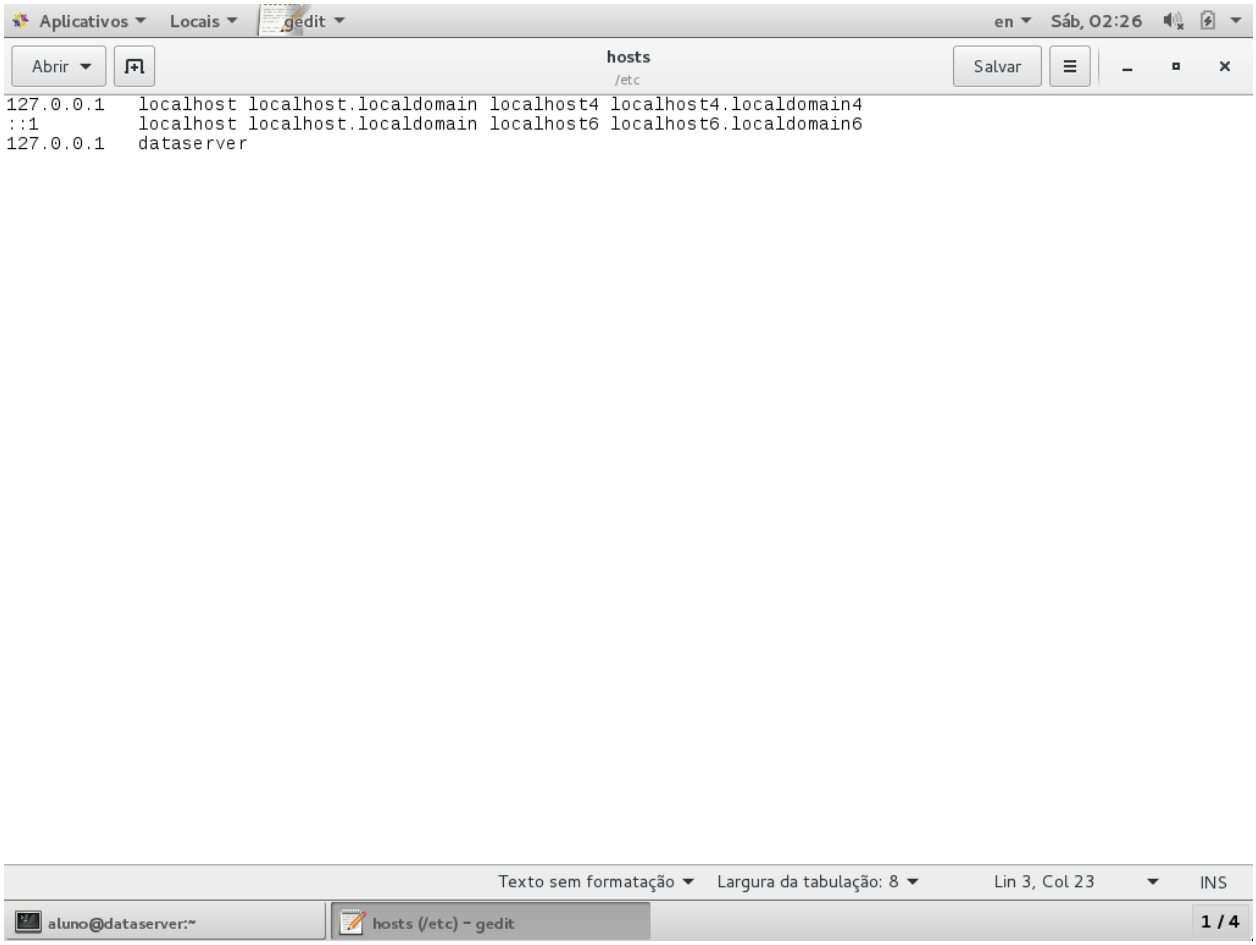
Parabéns, seu ambiente está pronto para receber o Hadoop!!

5.3. Download e Instalação do Hadoop

5.3.1. Editando o arquivo hosts

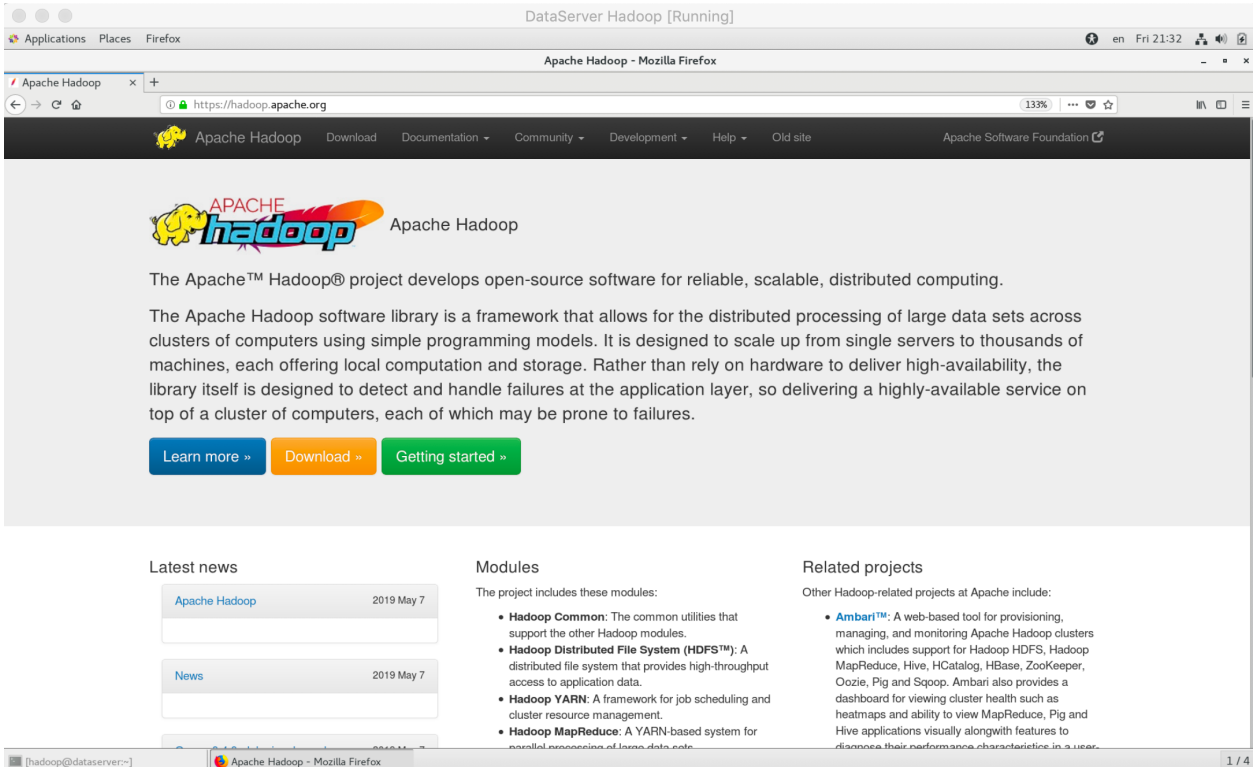


Editar o arquivo hosts

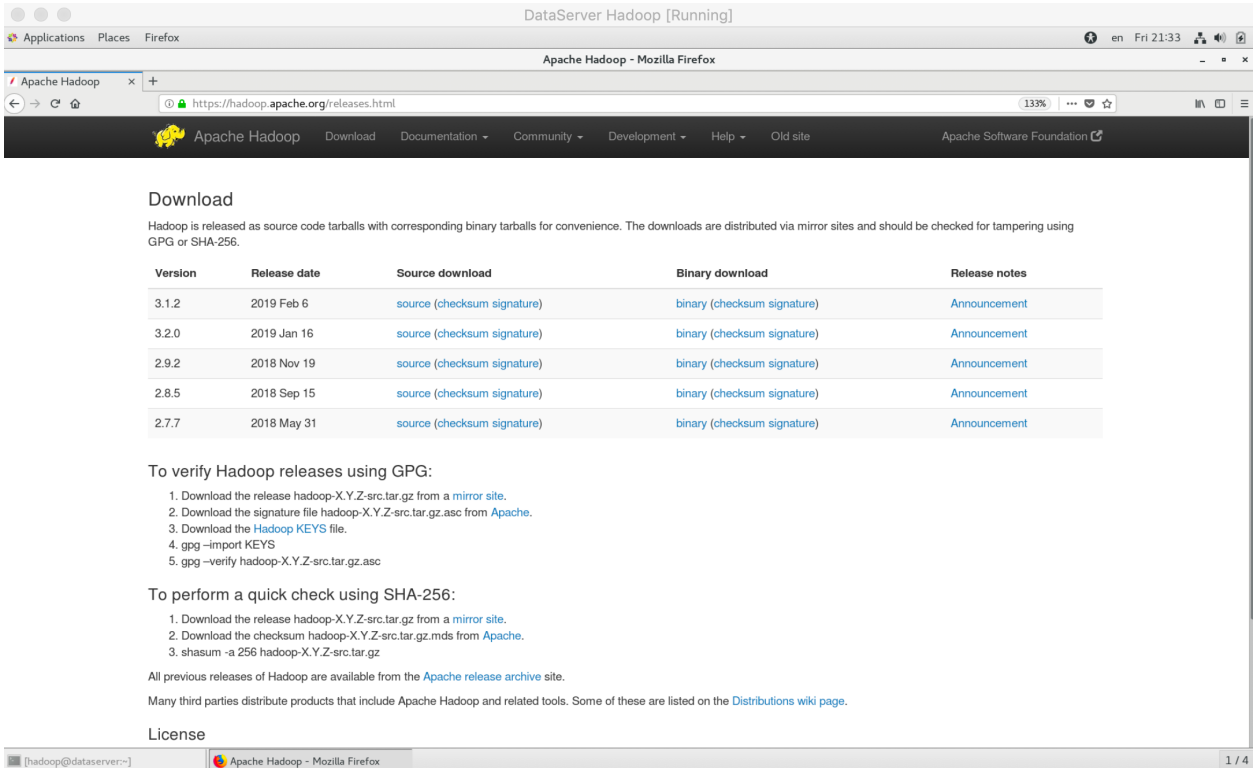


Incluir a última linha conforme acima

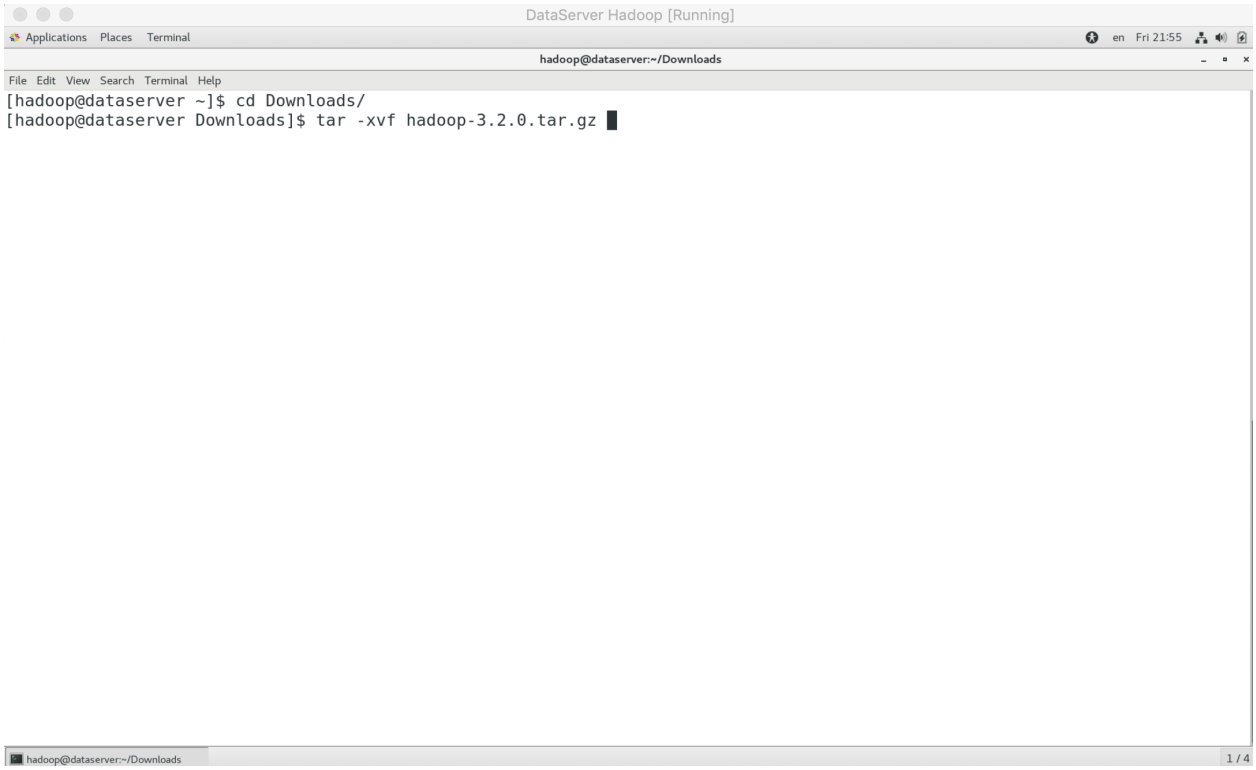
5.3.2. Download do Hadoop



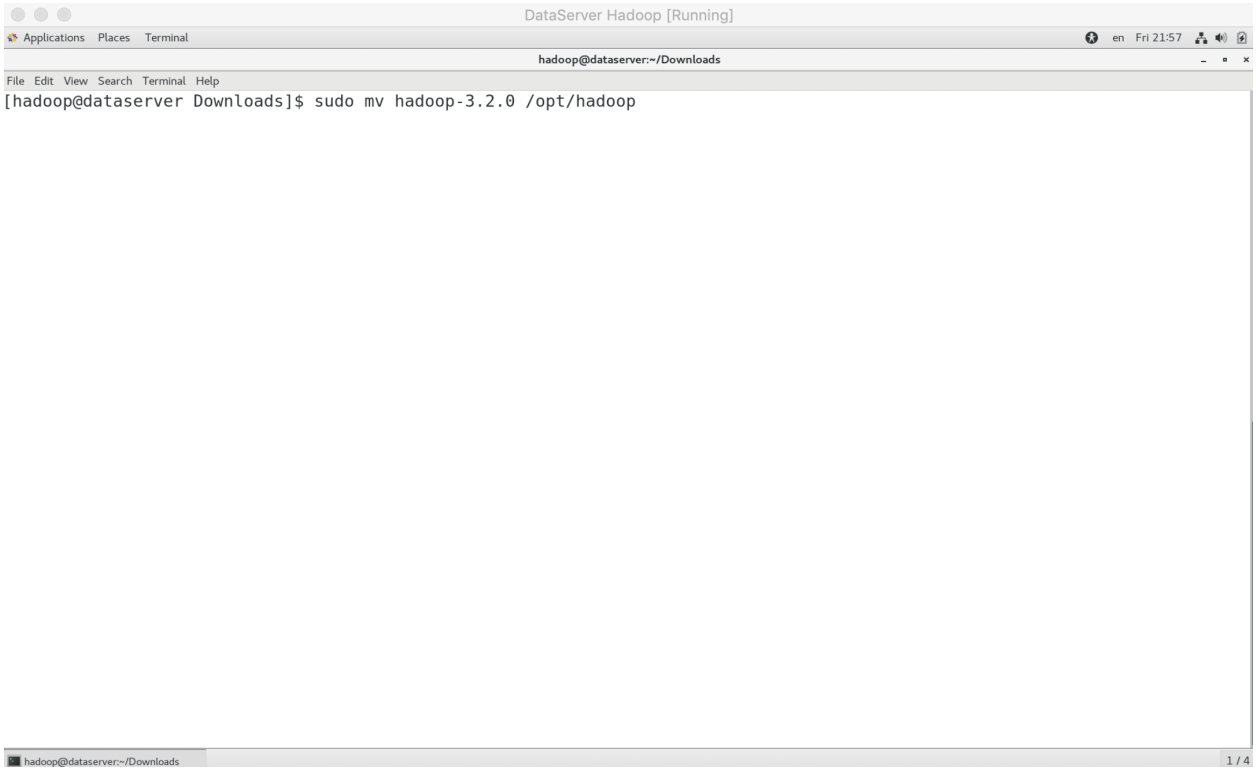
Acesse a página do Hadoop e clique em Download



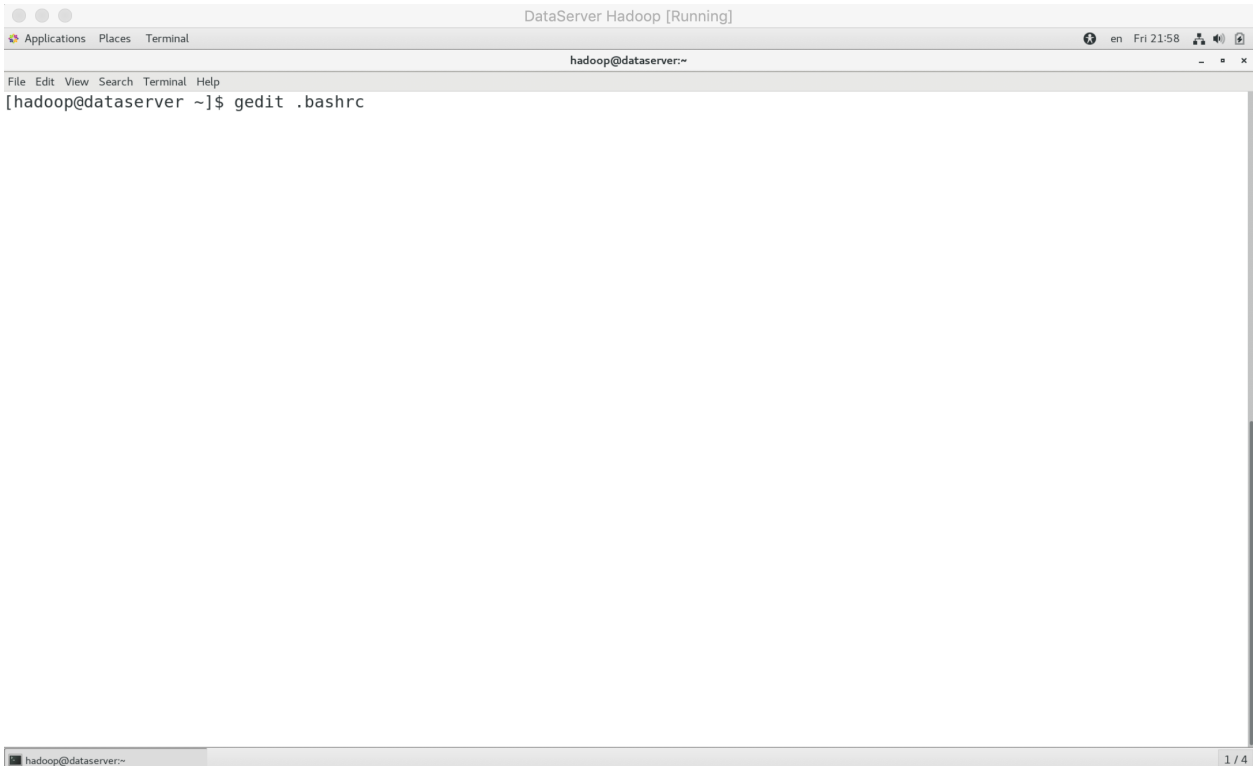
Faça o download da versão 3.2, opção binary.
O arquivo será baixado no diretório /home/hadoop/Downloads



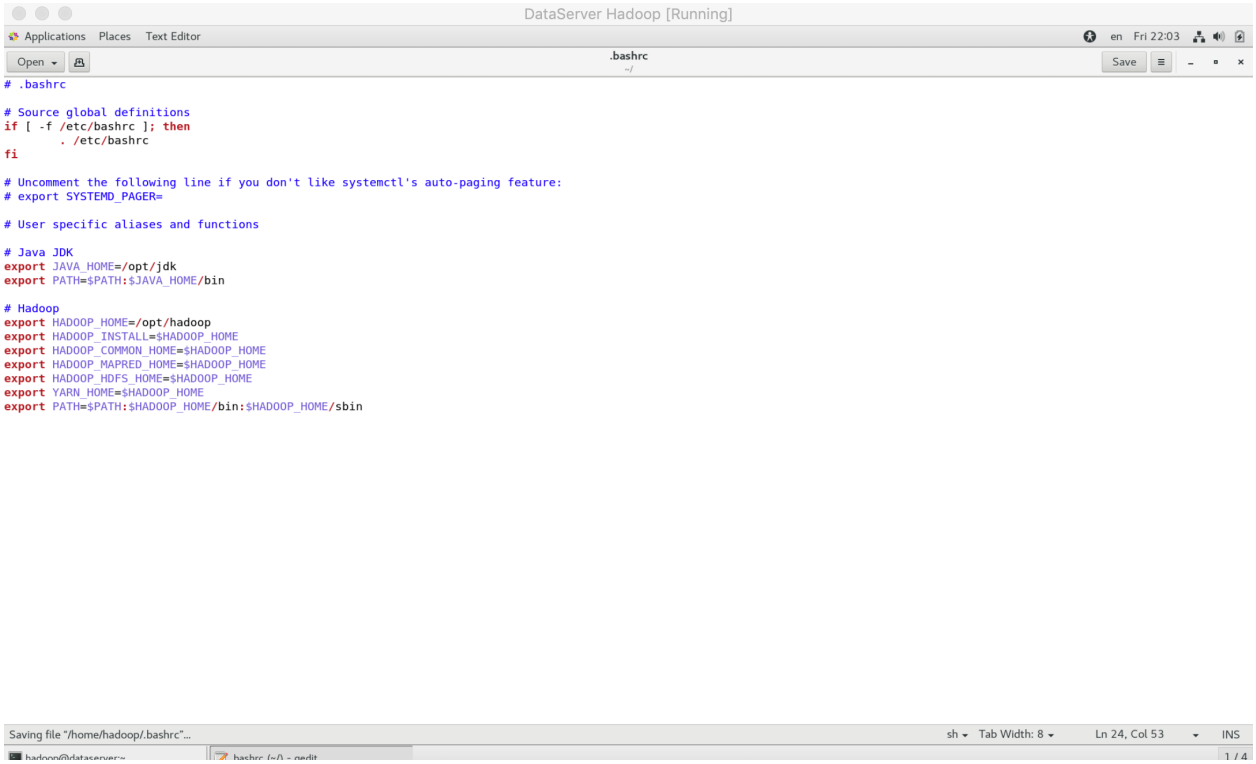
Descompacte o arquivo



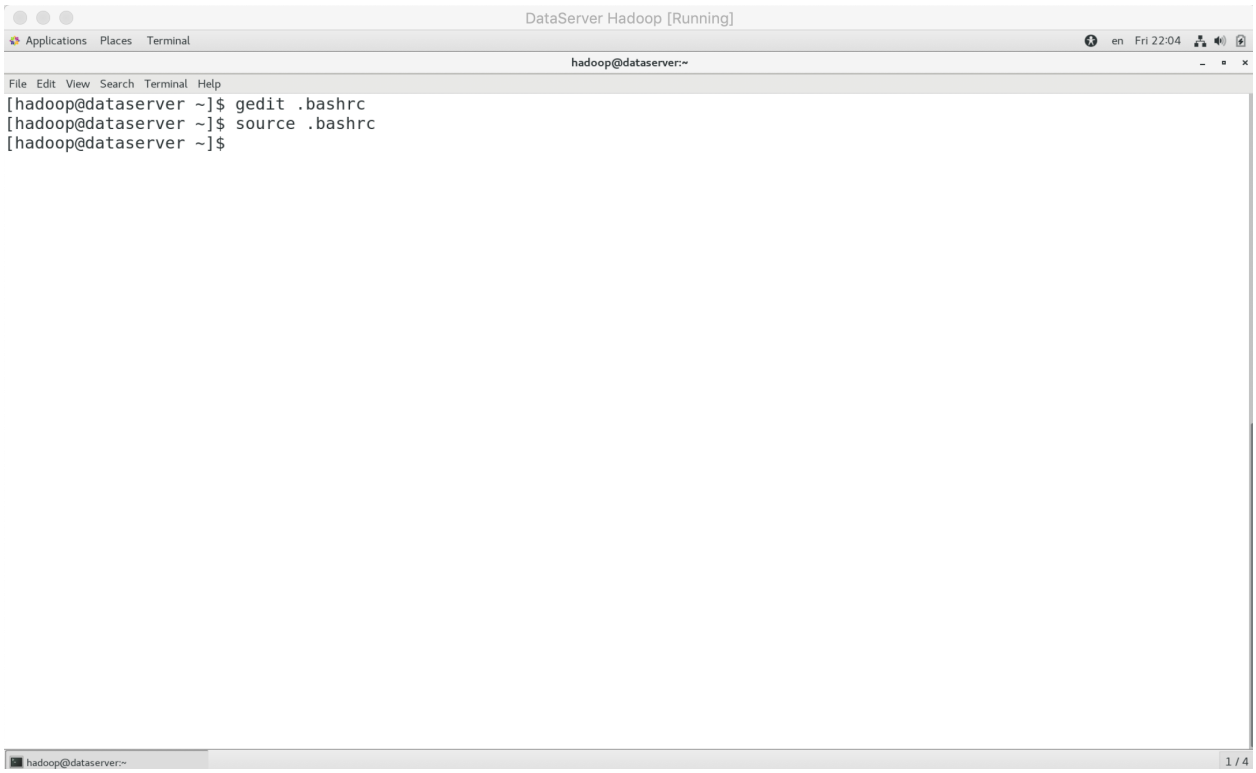
Mover o diretório para /opt/hadoop



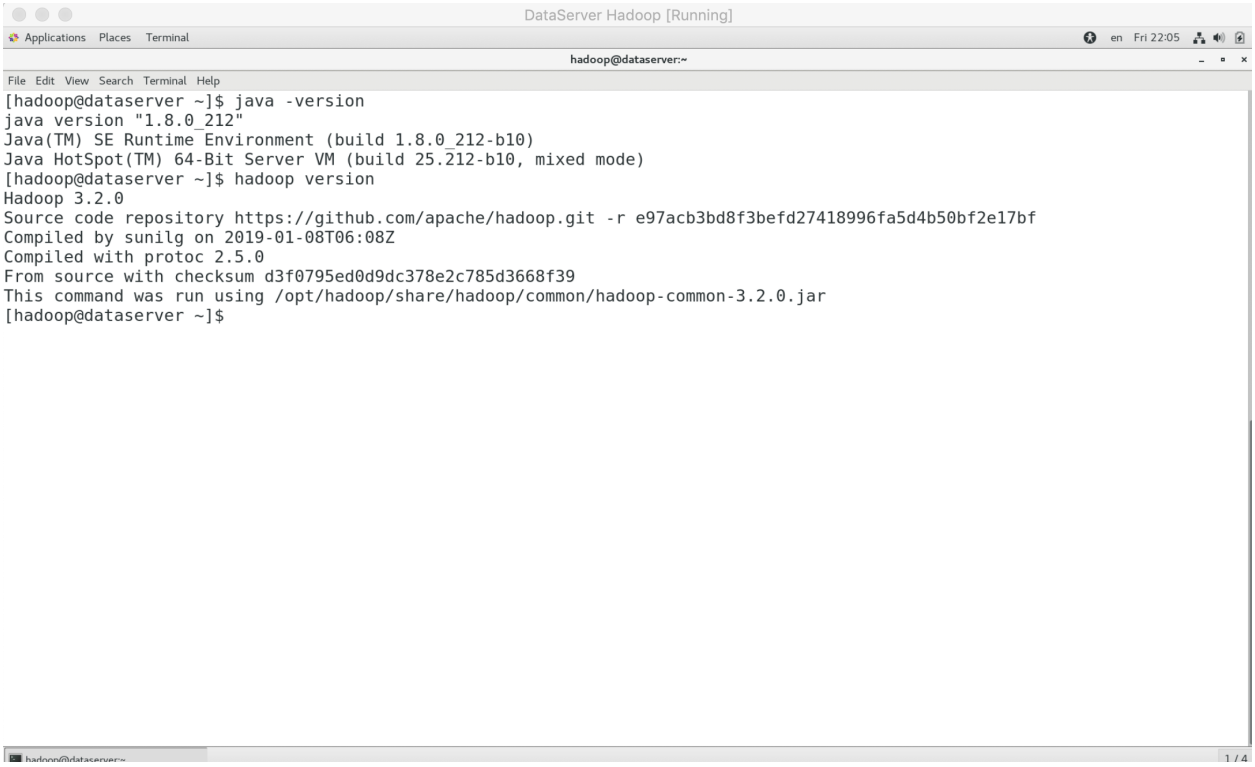
Abrir o arquivo de profile do usuário hadoop



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



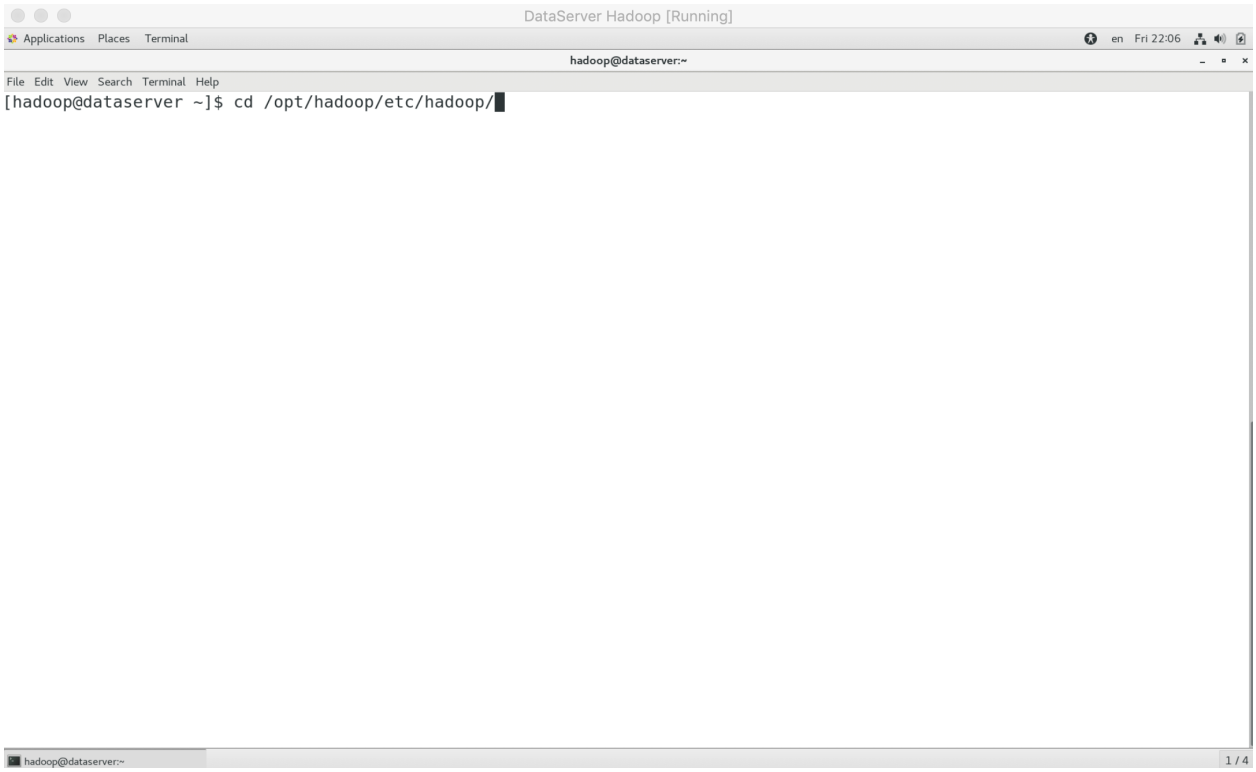
Execute source .bashrc para efetivar as mudanças das variáveis no SO



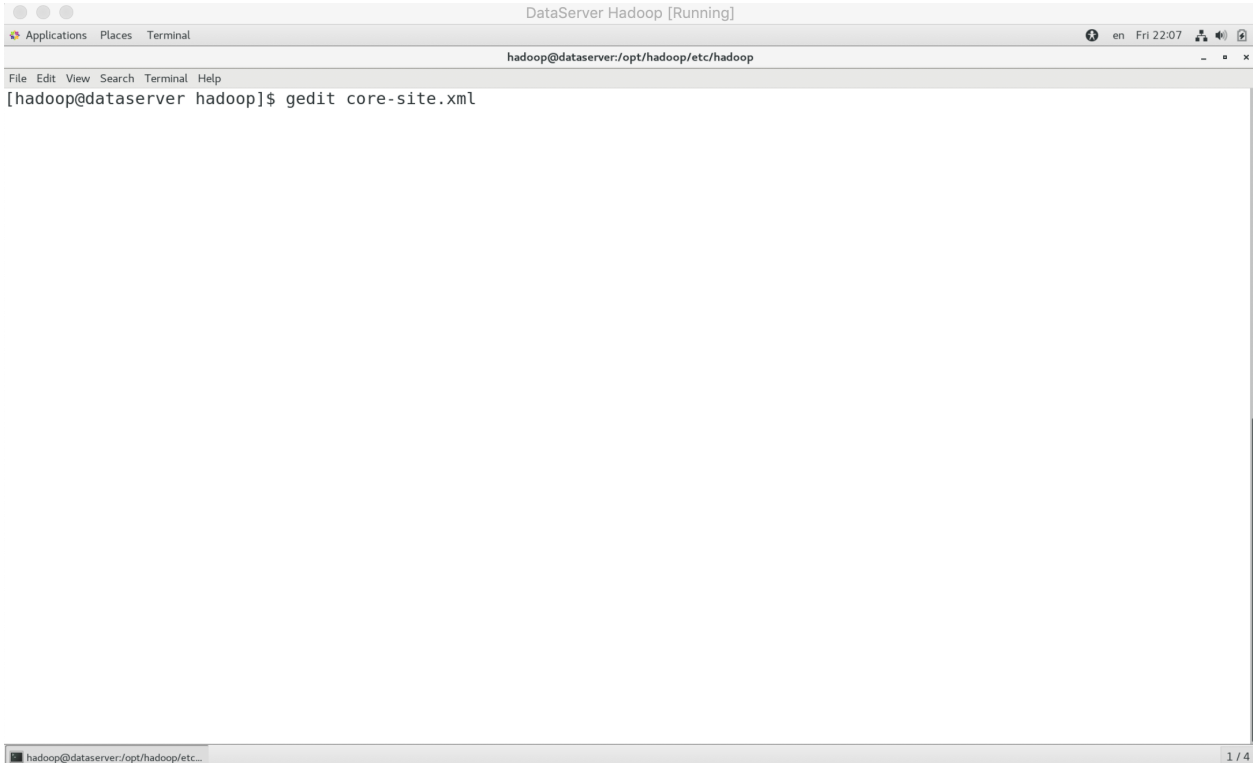
Java e Hadoop instalados e configurados com sucesso!!!

5.4. Configuração do Hadoop

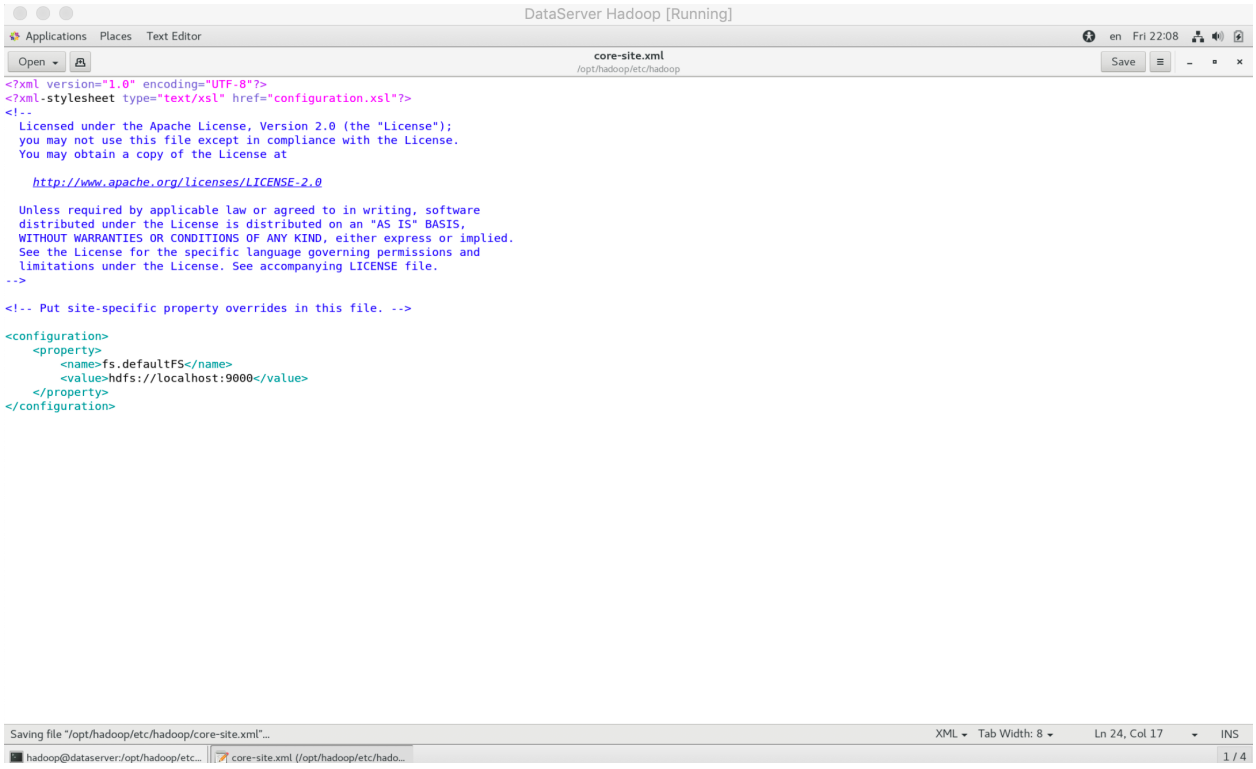
5.4.1. Editar arquivos de configuração do Hadoop



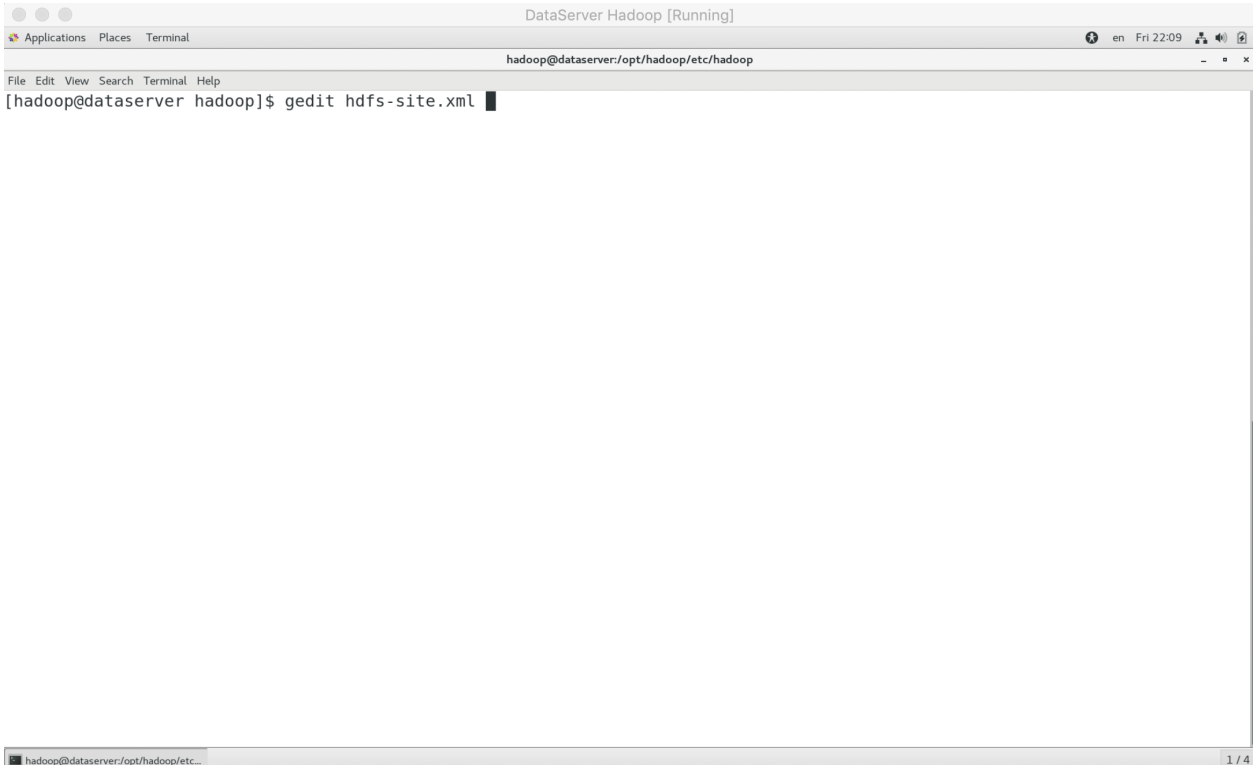
Os arquivos de configuração do Hadoop estão em
[Diretório de instalação do Hadoop]/etc/hadoop
Nesse caso: /opt/hadoop/etc/hadoop



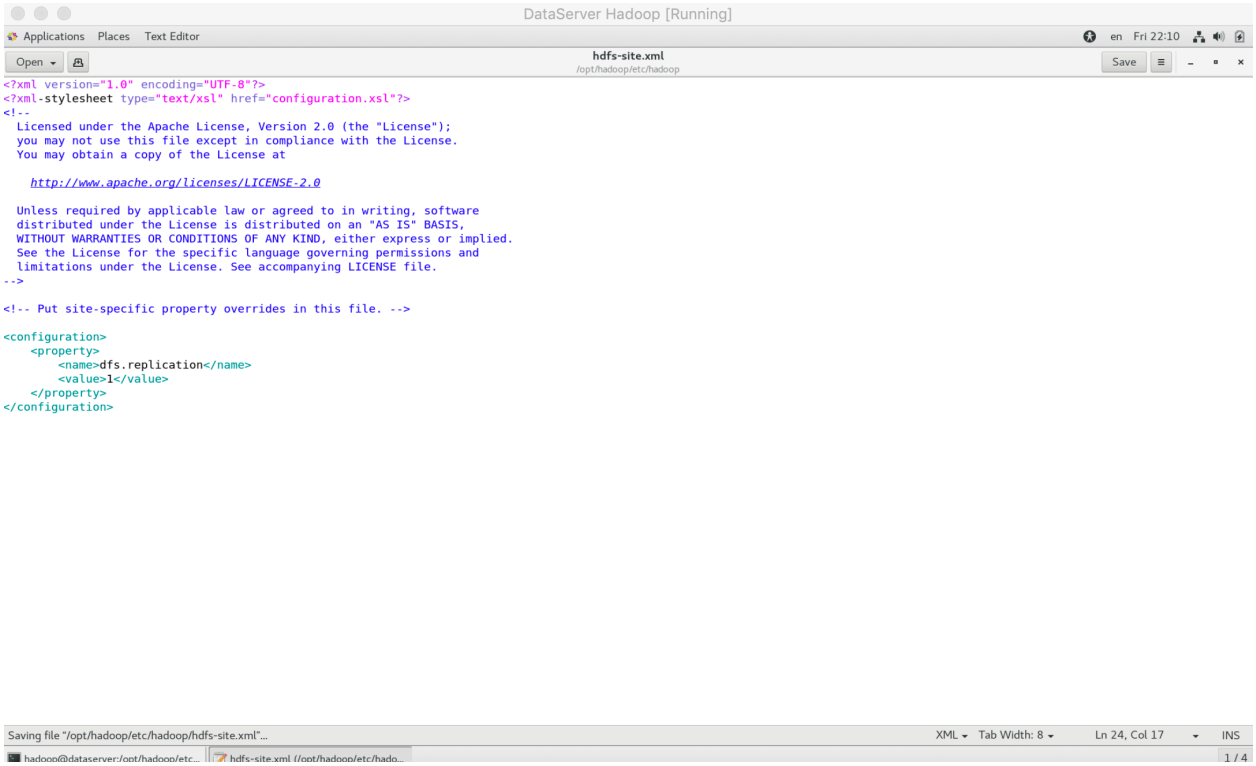
Editar o arquivo core-site.xml



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

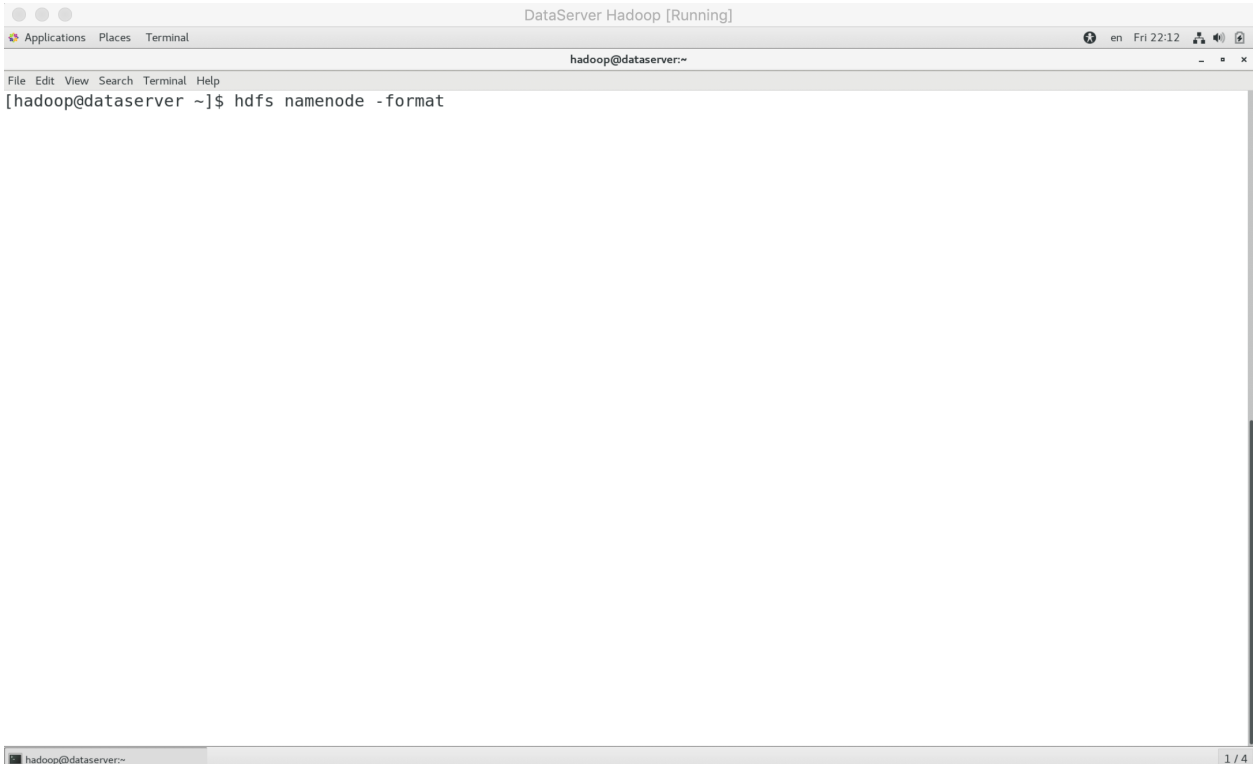


Editar o arquivo hdfs-site.xml

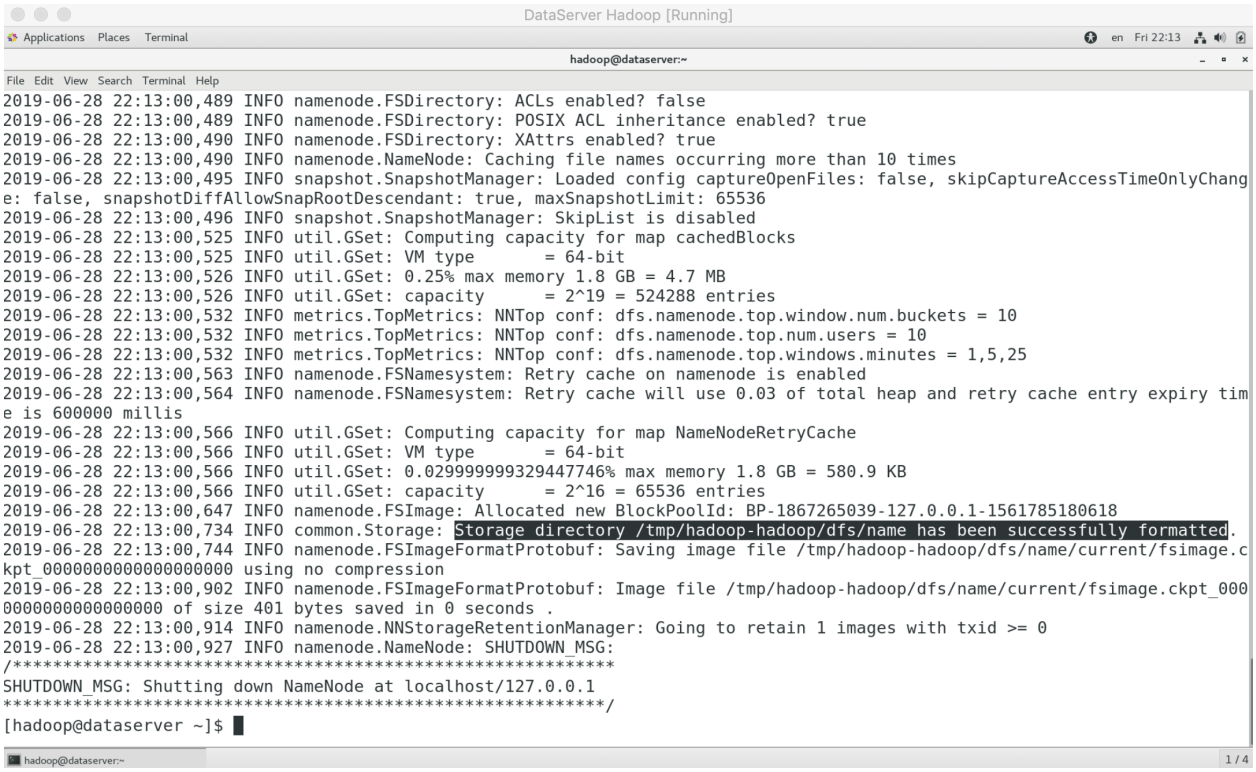


Acrescentar as propriedades conforme acima e salvar o arquivo.

5.4.2. Formatando o Namenode

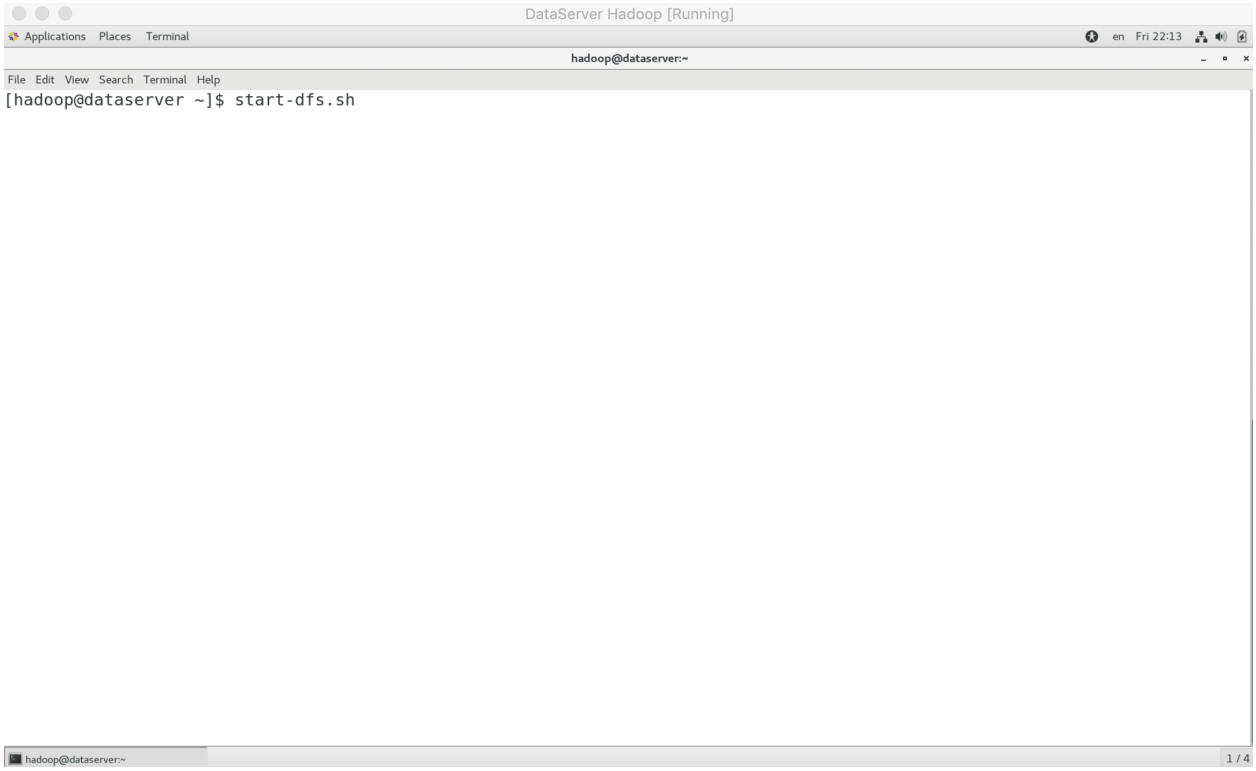


hdfs namenode –format

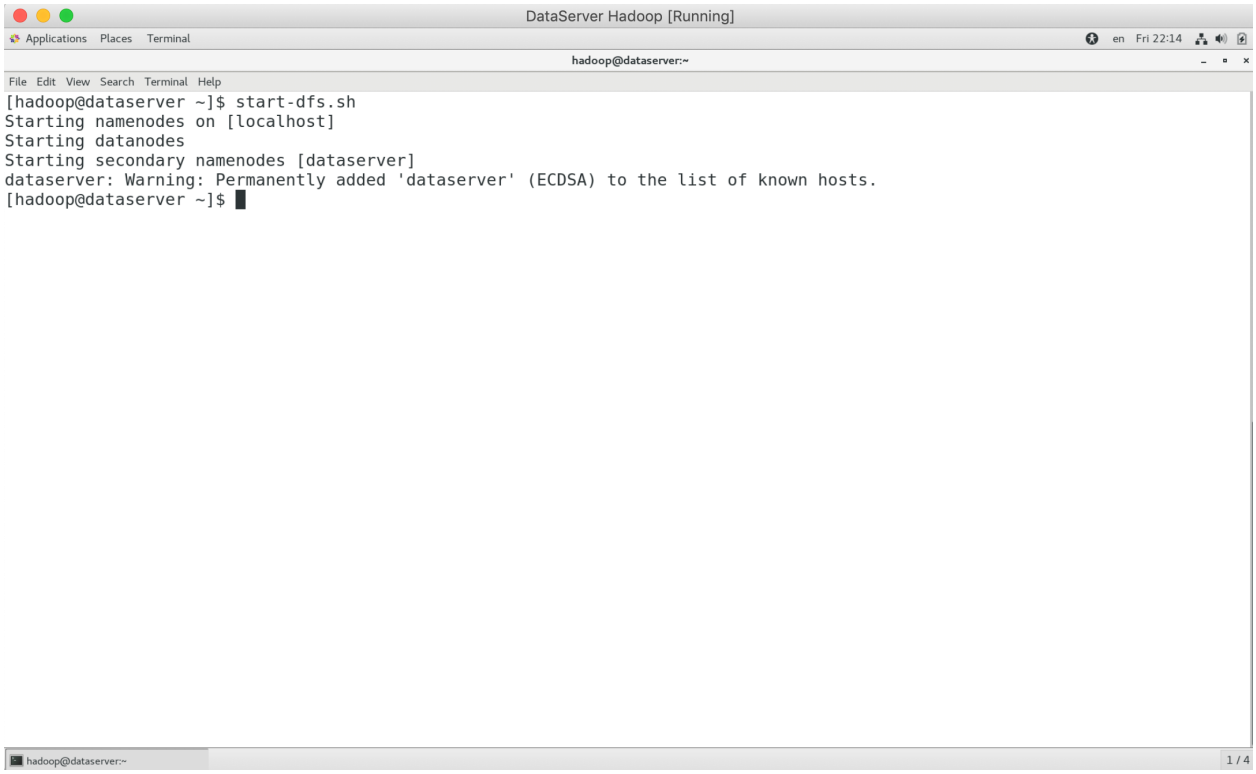


Formatação realizada com sucesso

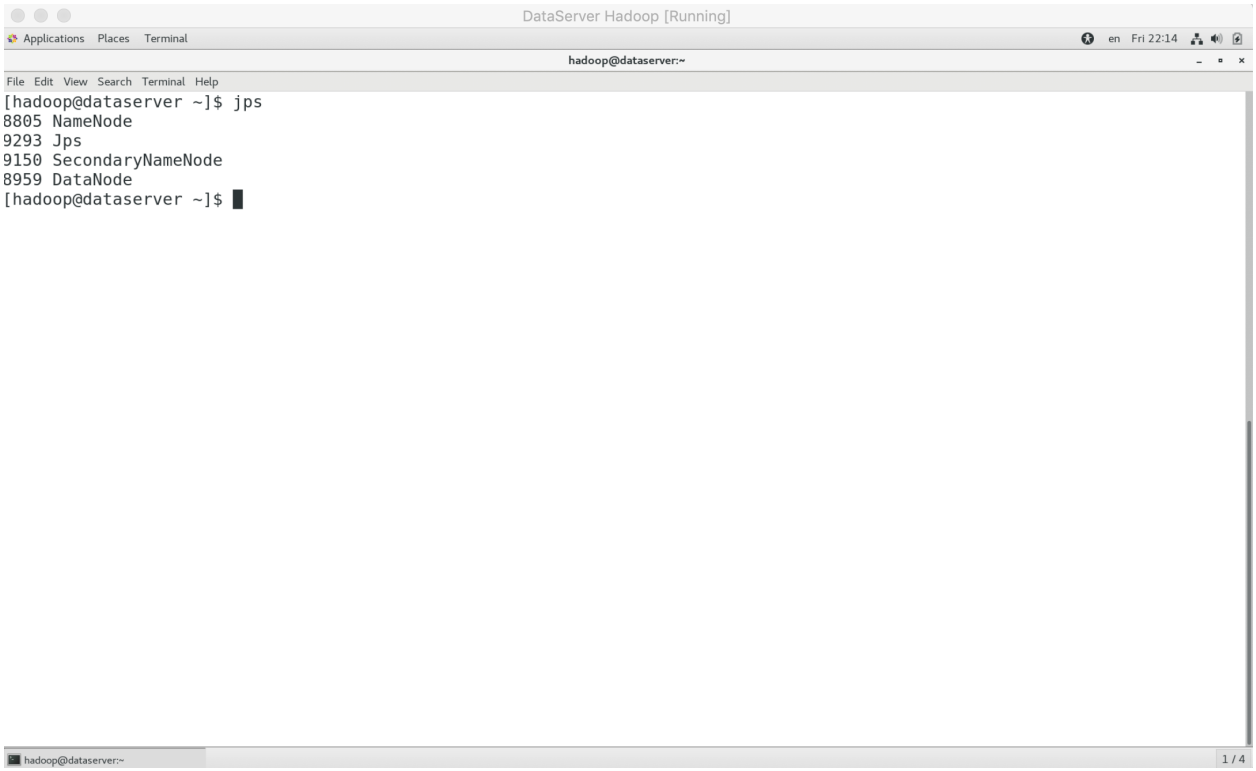
5.4.3. Iniciando o Hadoop



start-dfs.sh

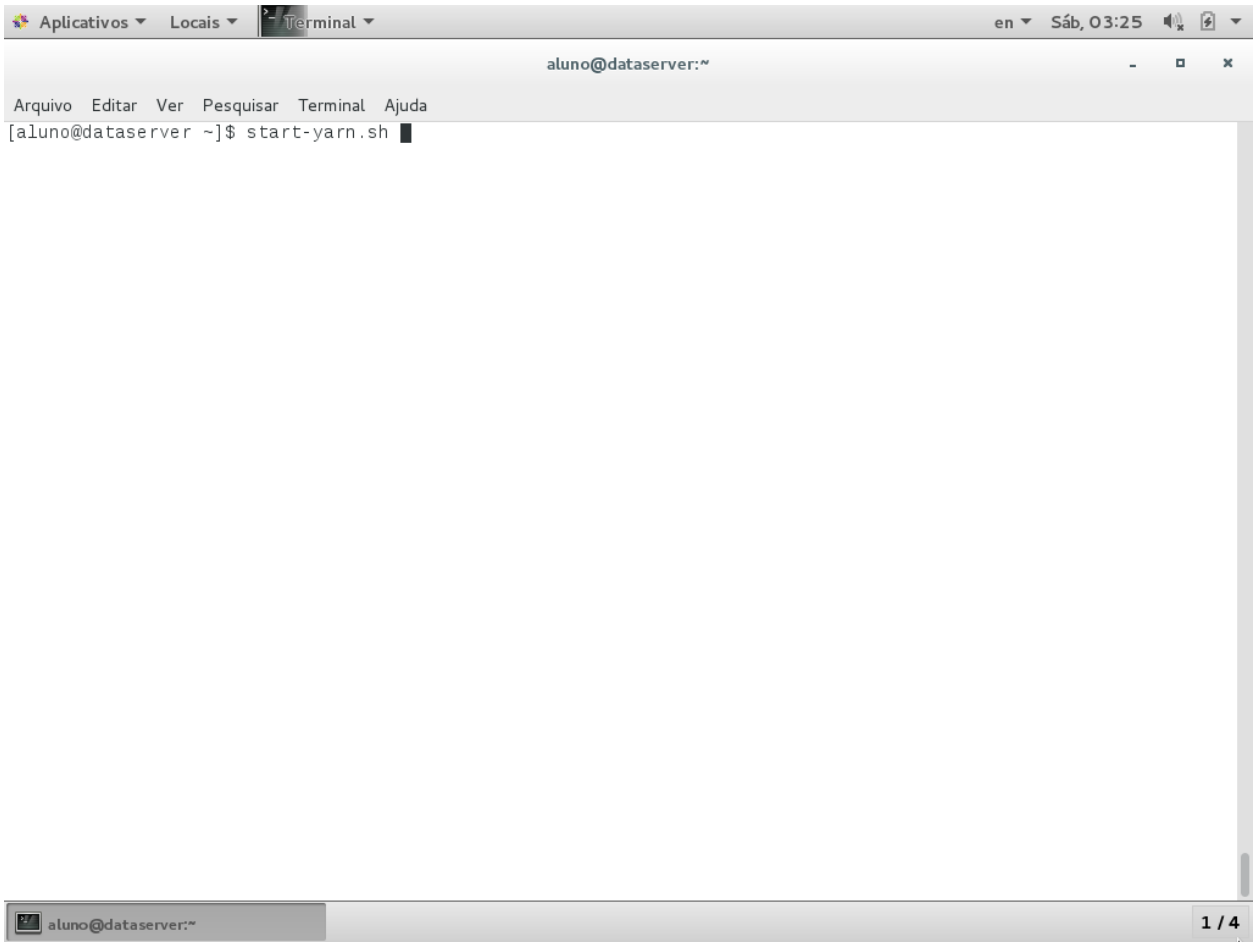


Hadoop iniciado

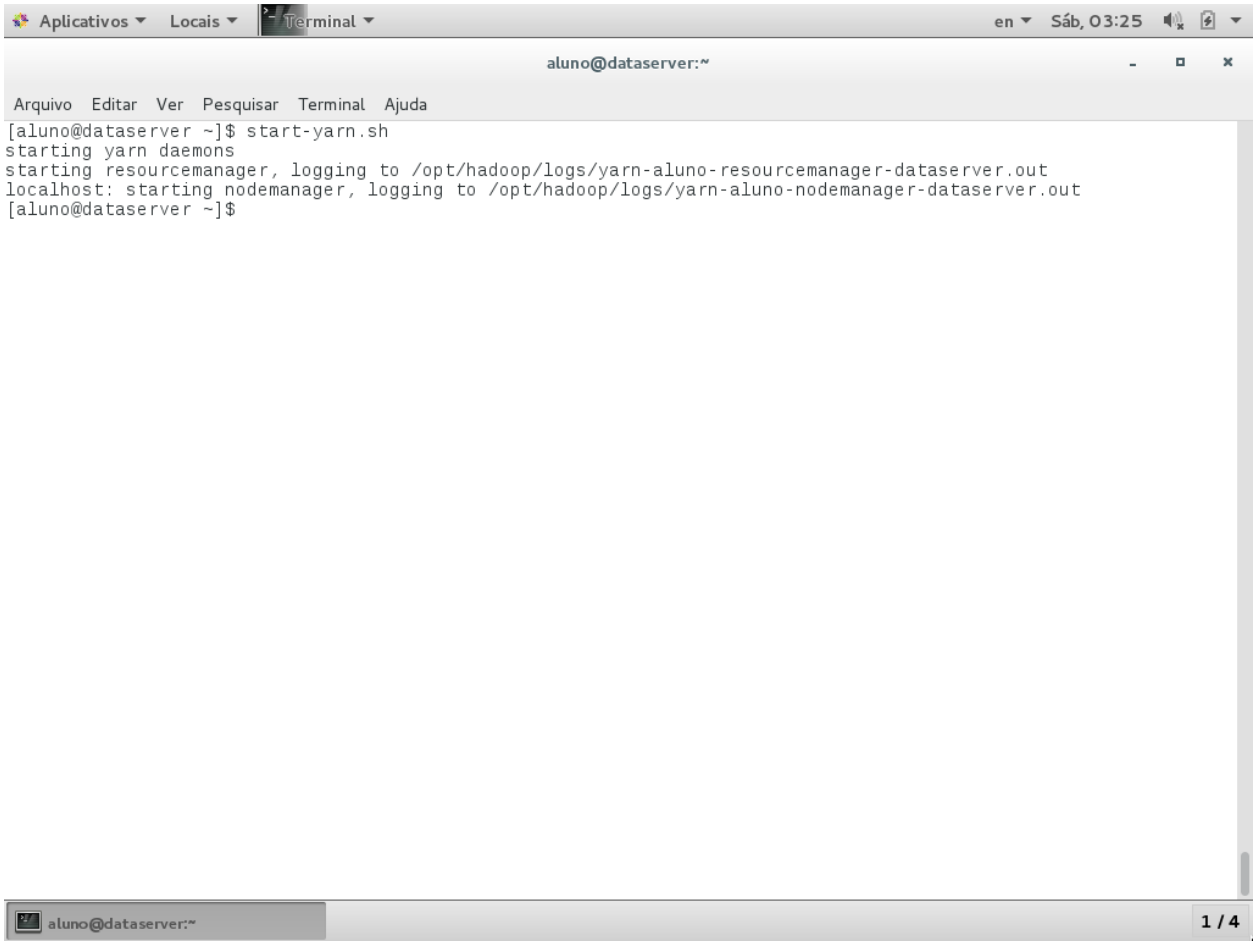


Checando os serviços inicializados com o comando **jps**

5.4.4. Iniciando o Yarn

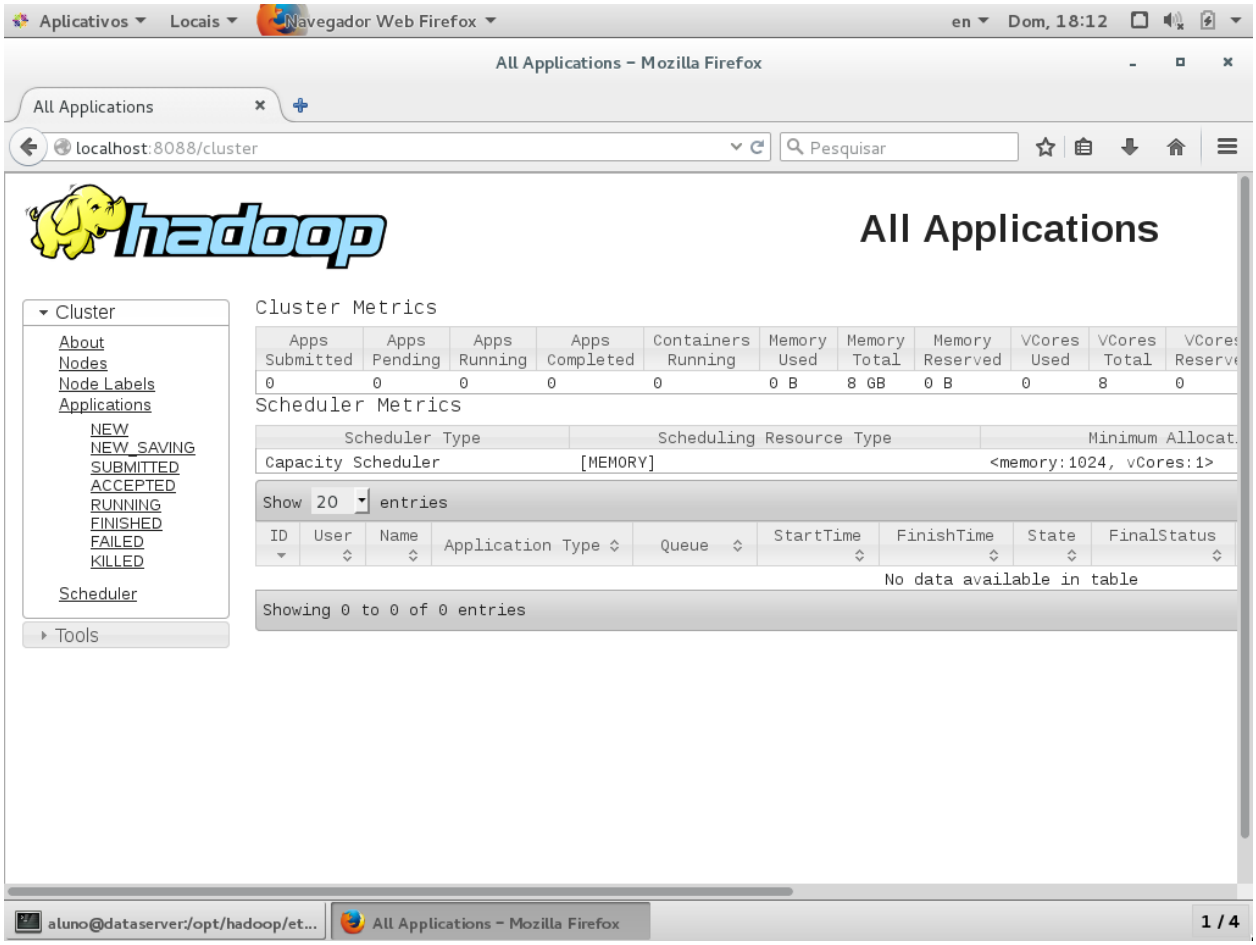


start-yarn.sh



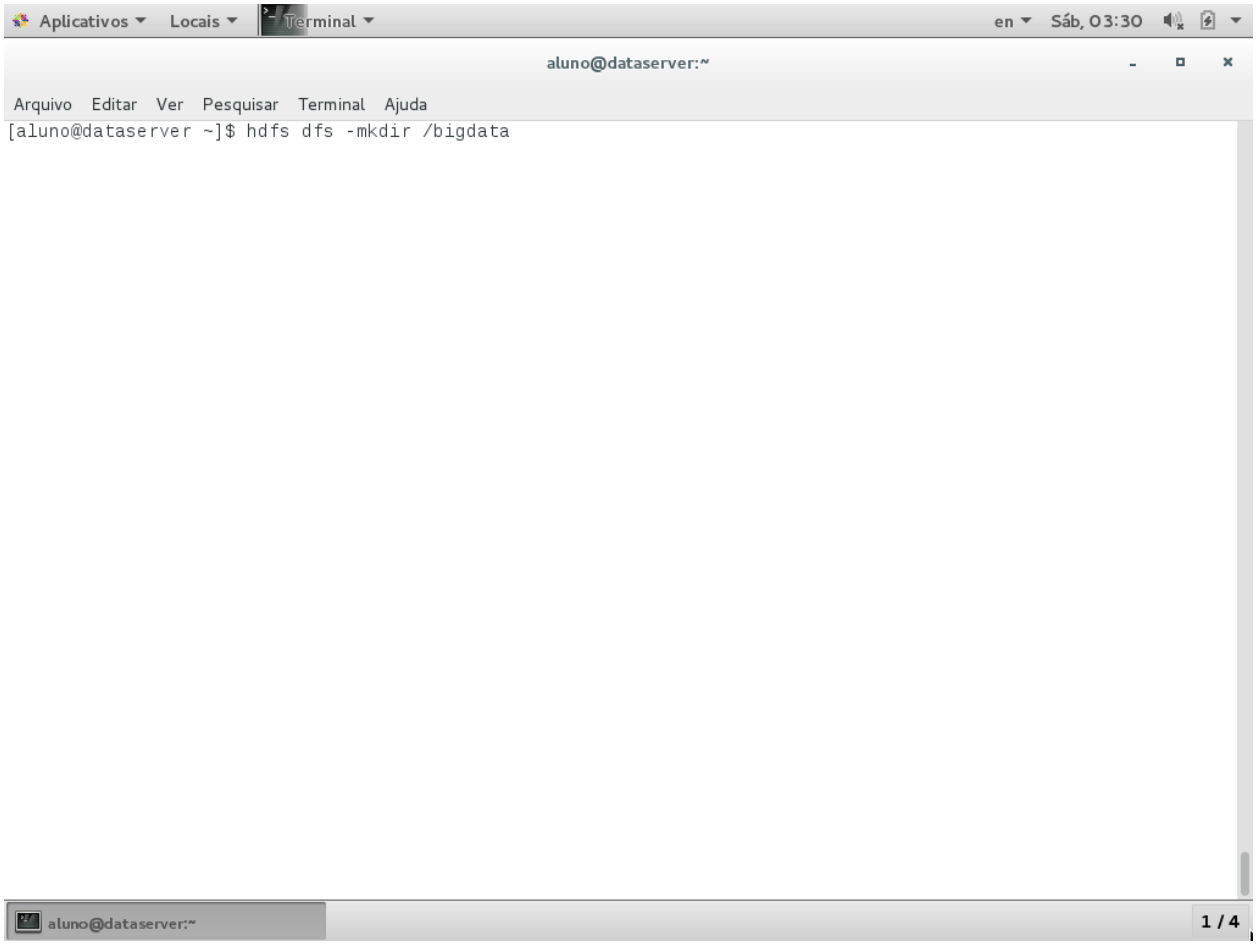
Yarn iniciado

Checando os serviços com o comando **jps**

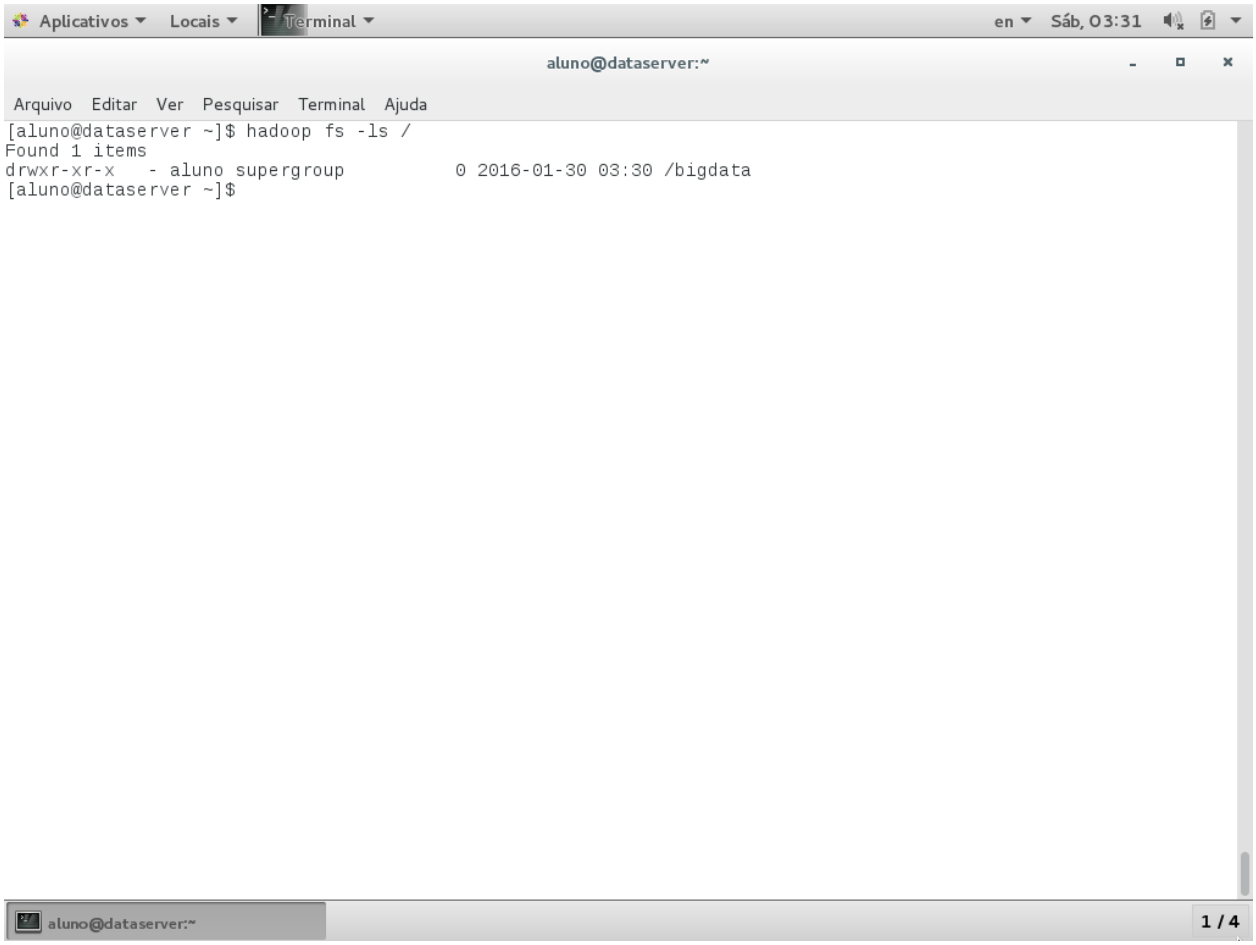


Visualizando jobs – http://localhost:8088

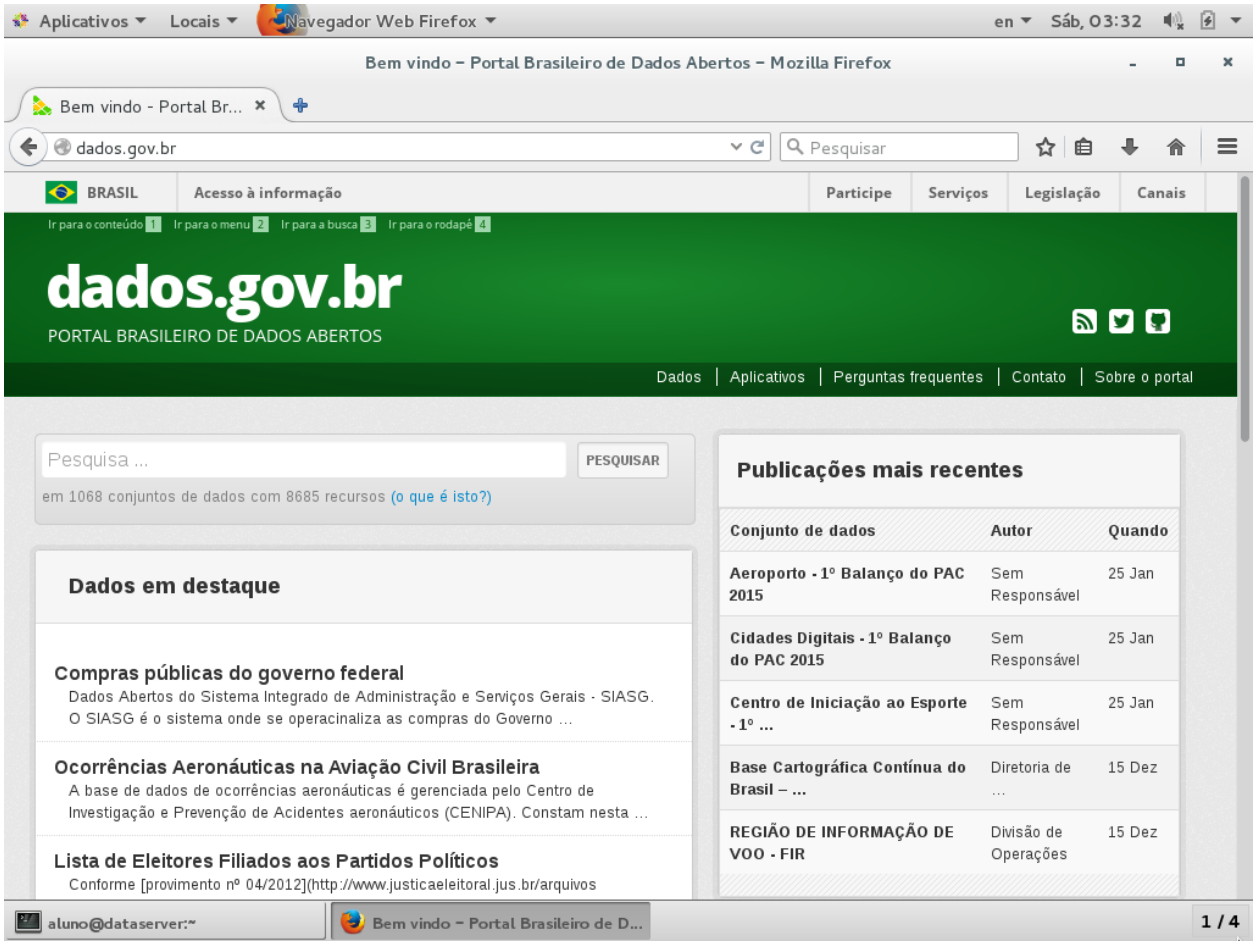
5.5. Processando Big Data



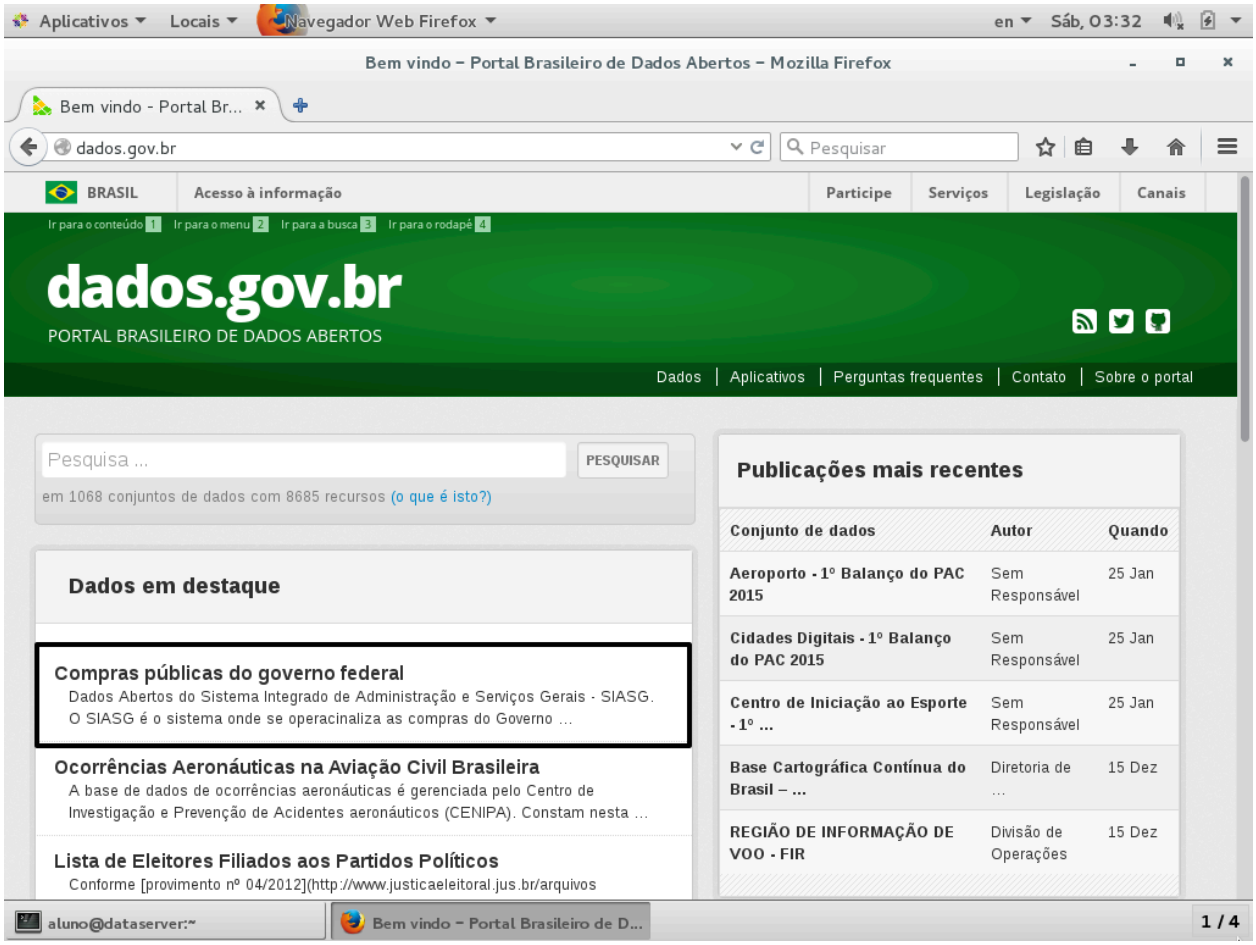
Criar o diretório **bigdata** no HDFS



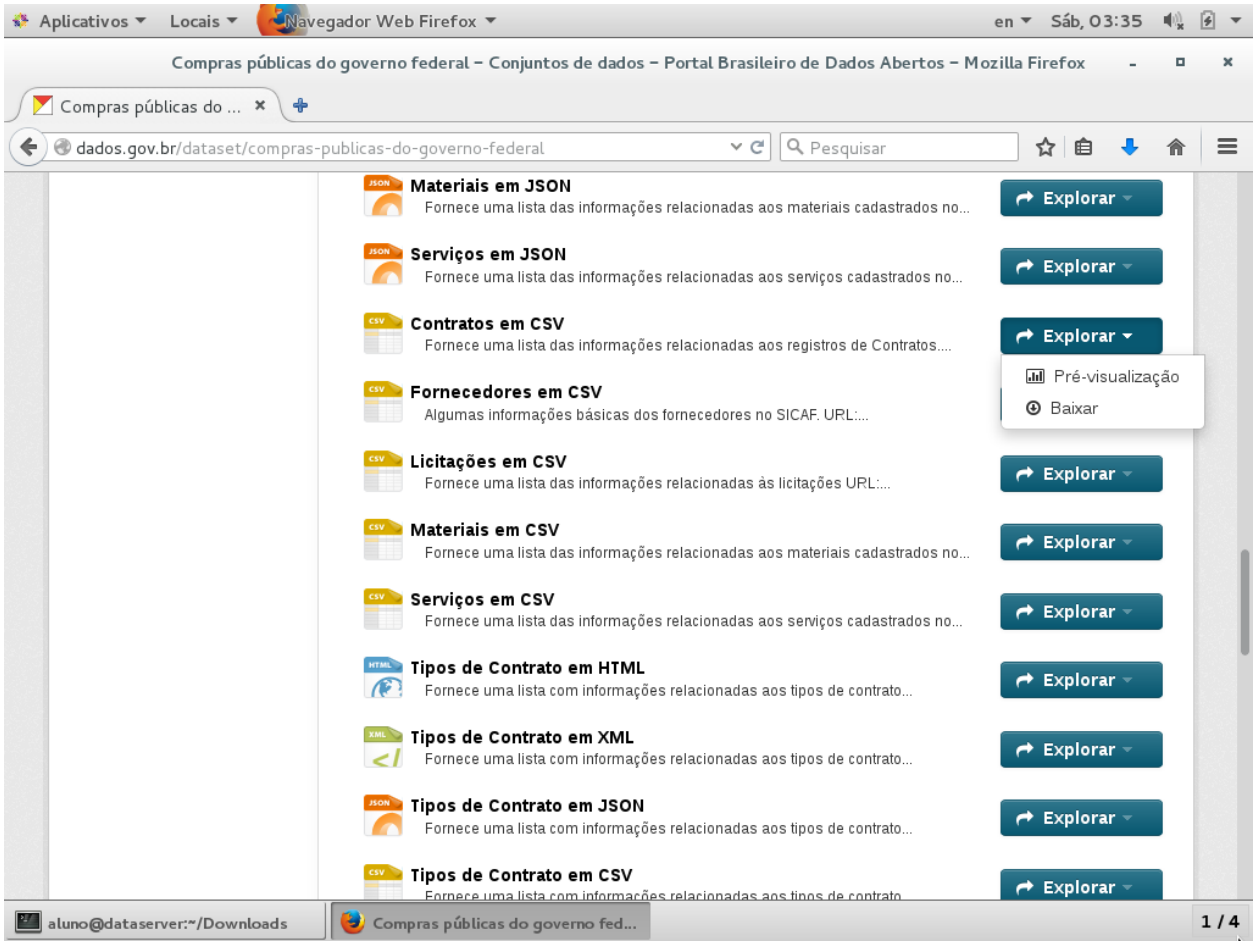
Listar o HDFS e checar o diretório criado



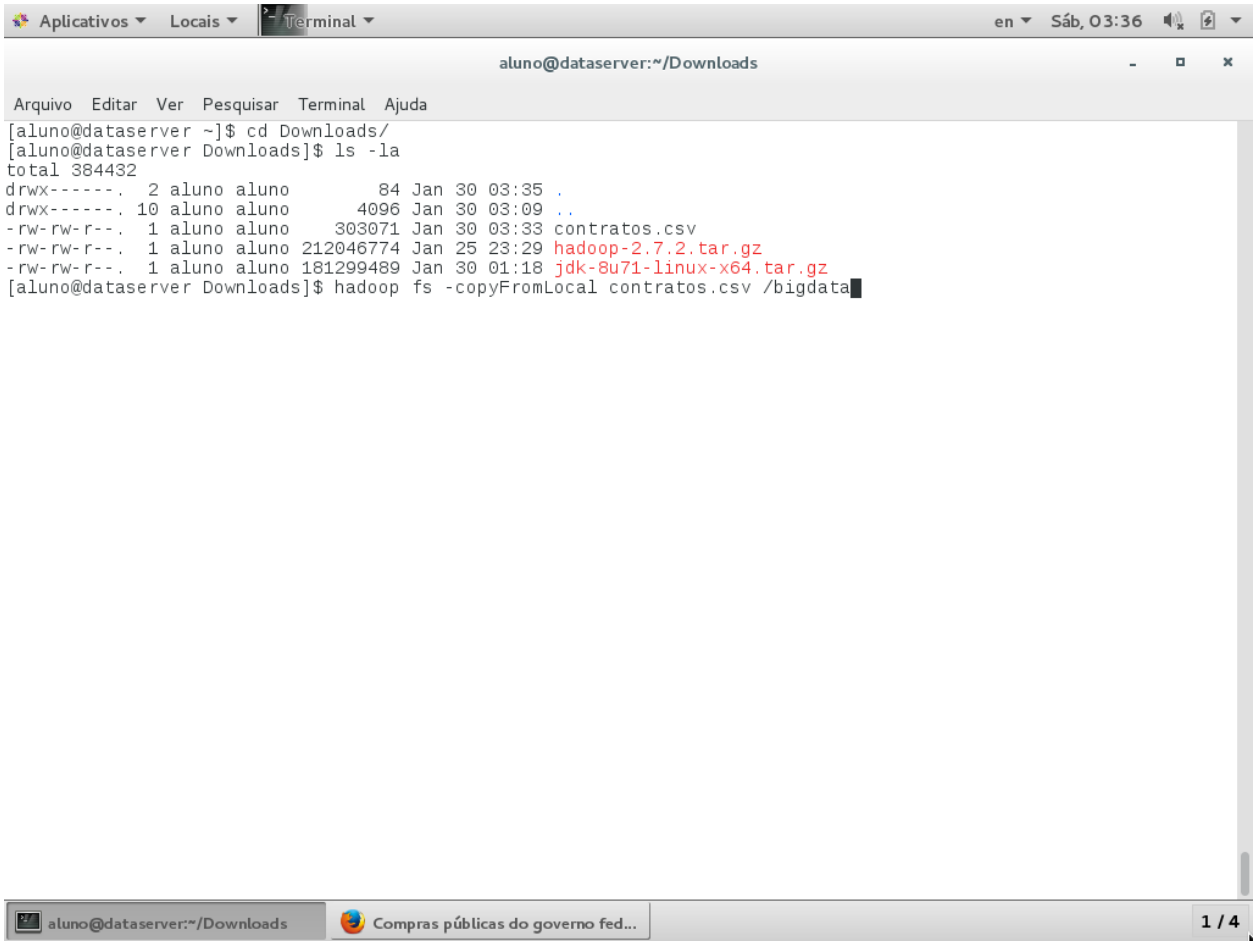
Acessar o portal de dados abertos do governo federal



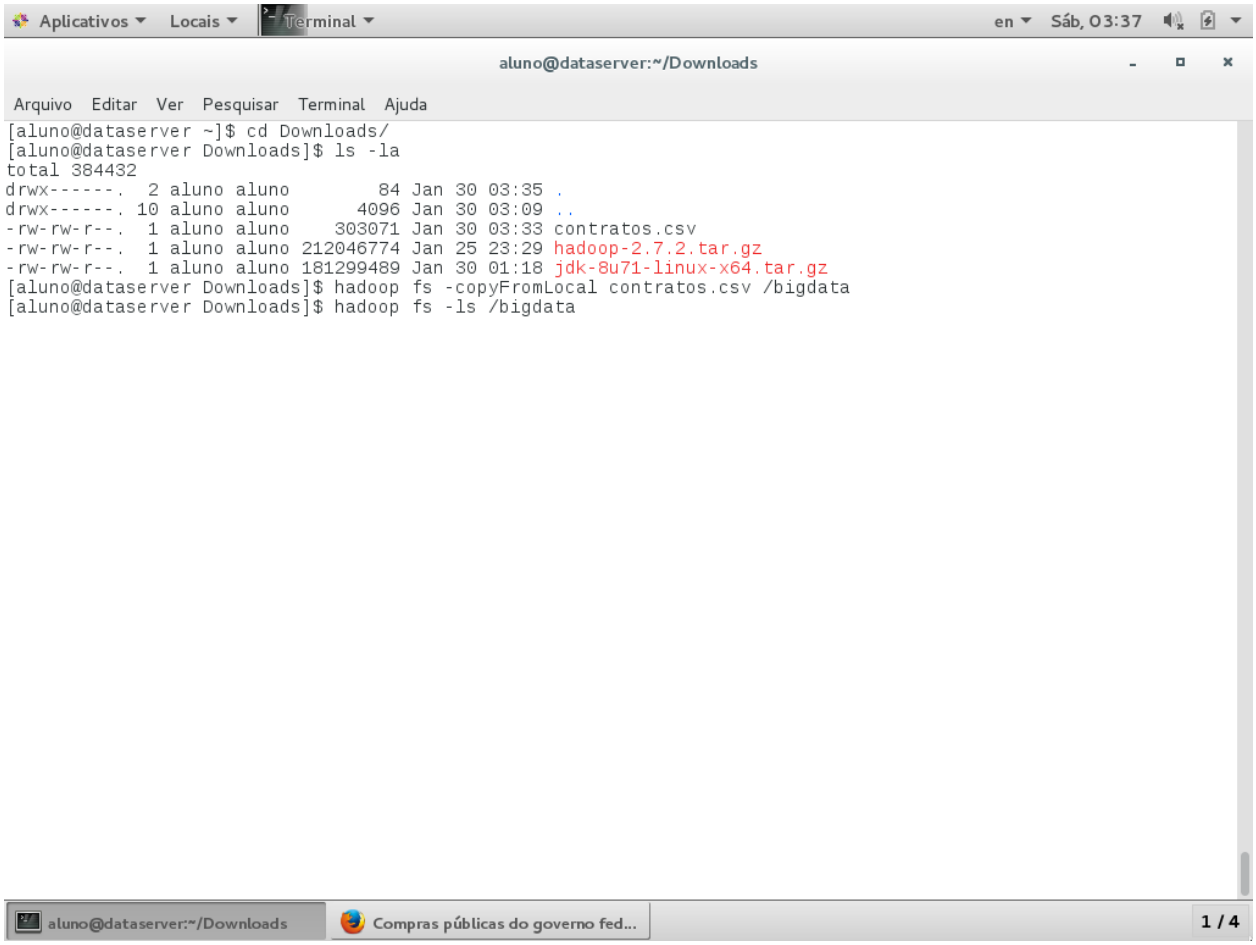
Clicar no link de compras públicas



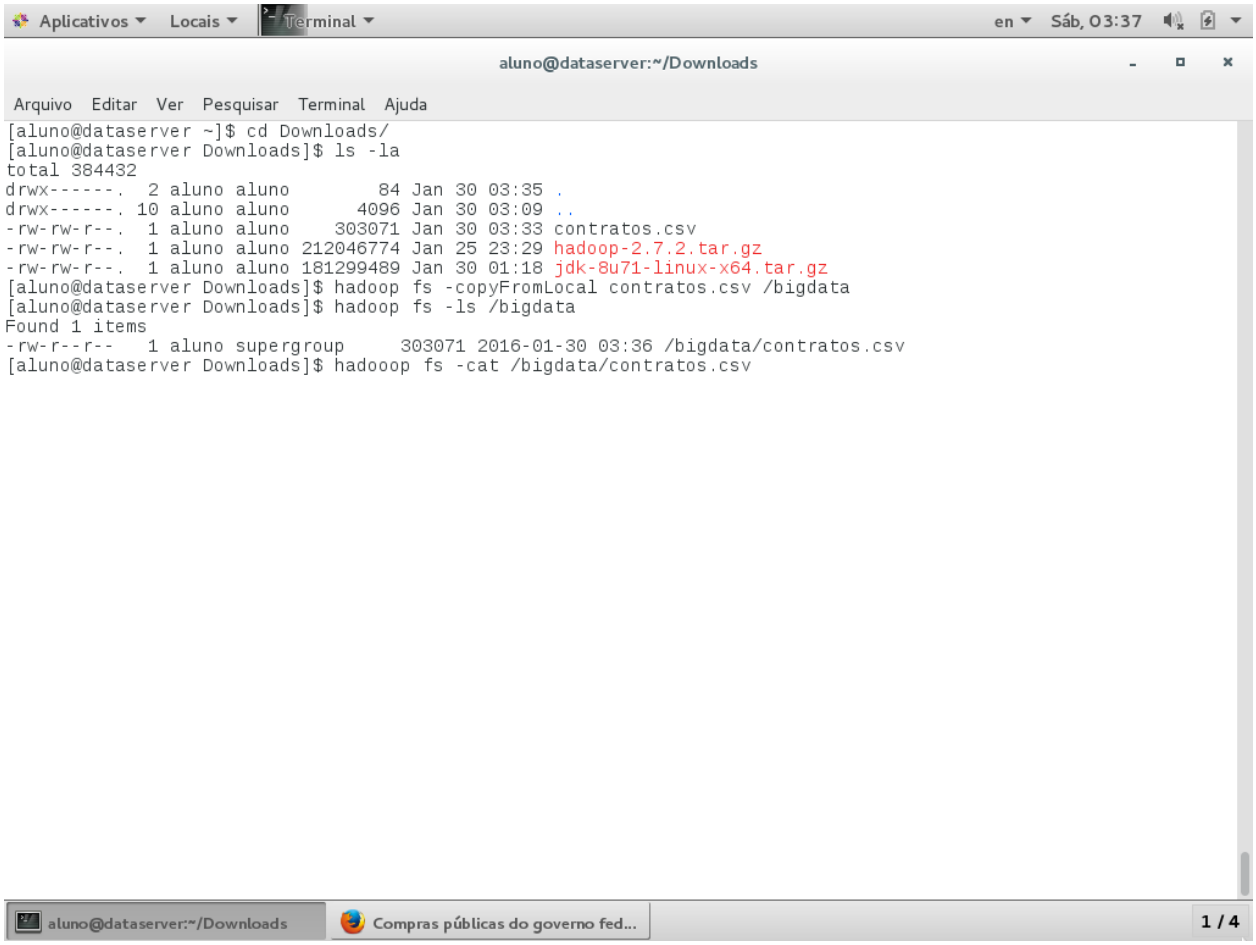
Baixar o arquivo de contratos em formato csv (pode ser qualquer um)



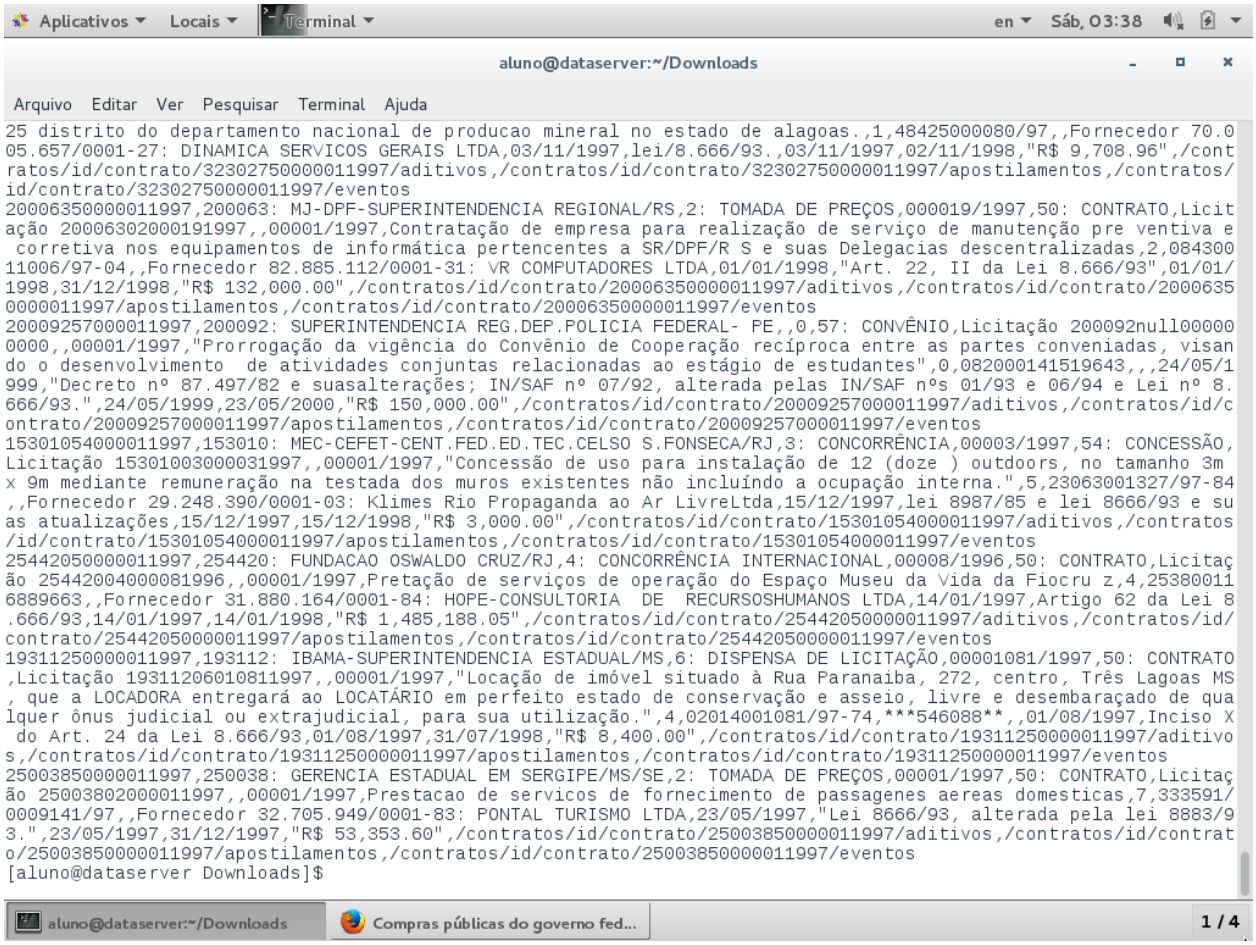
Copiar o arquivo para a pasta bigdata no HDFS



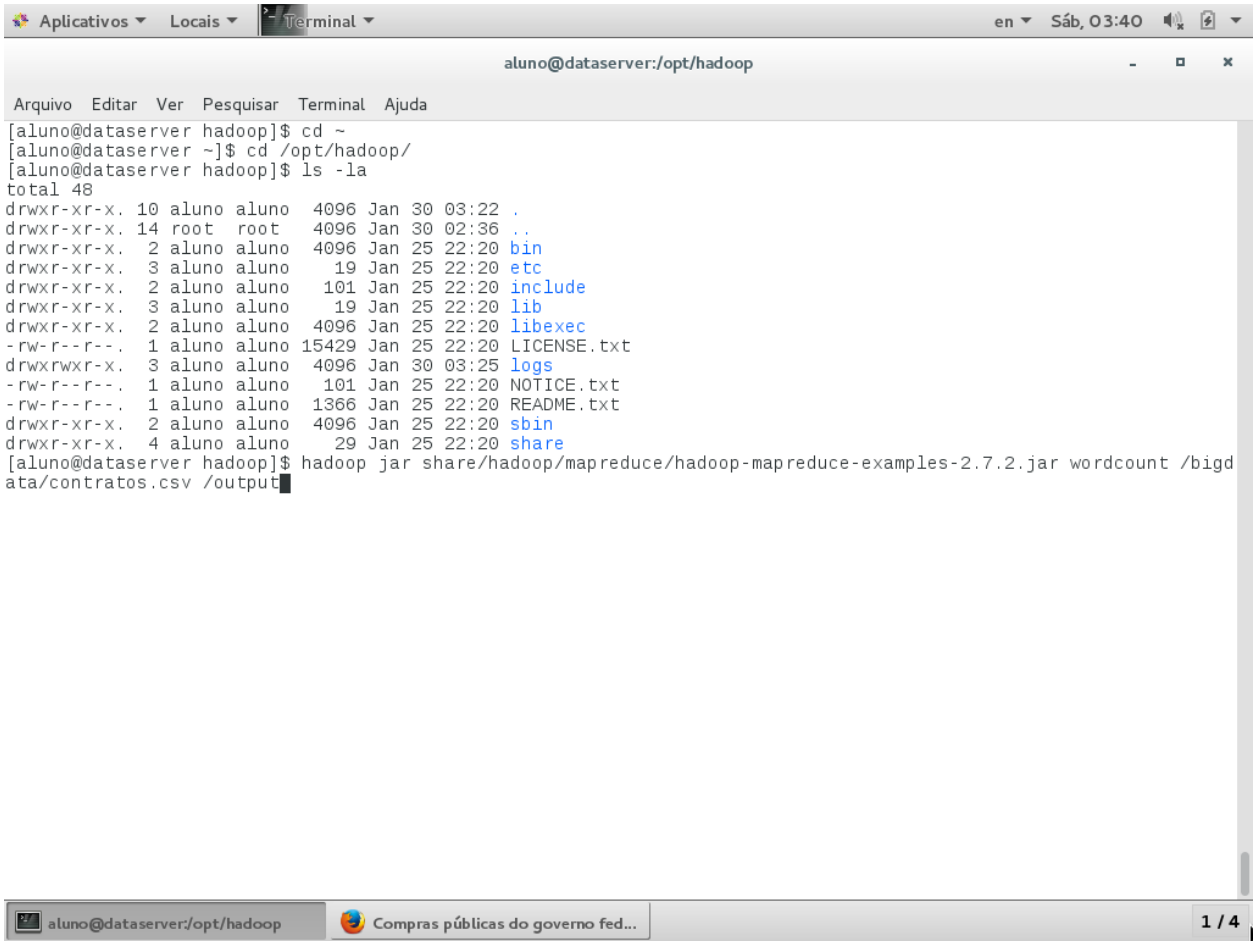
Listar o diretório bigdata



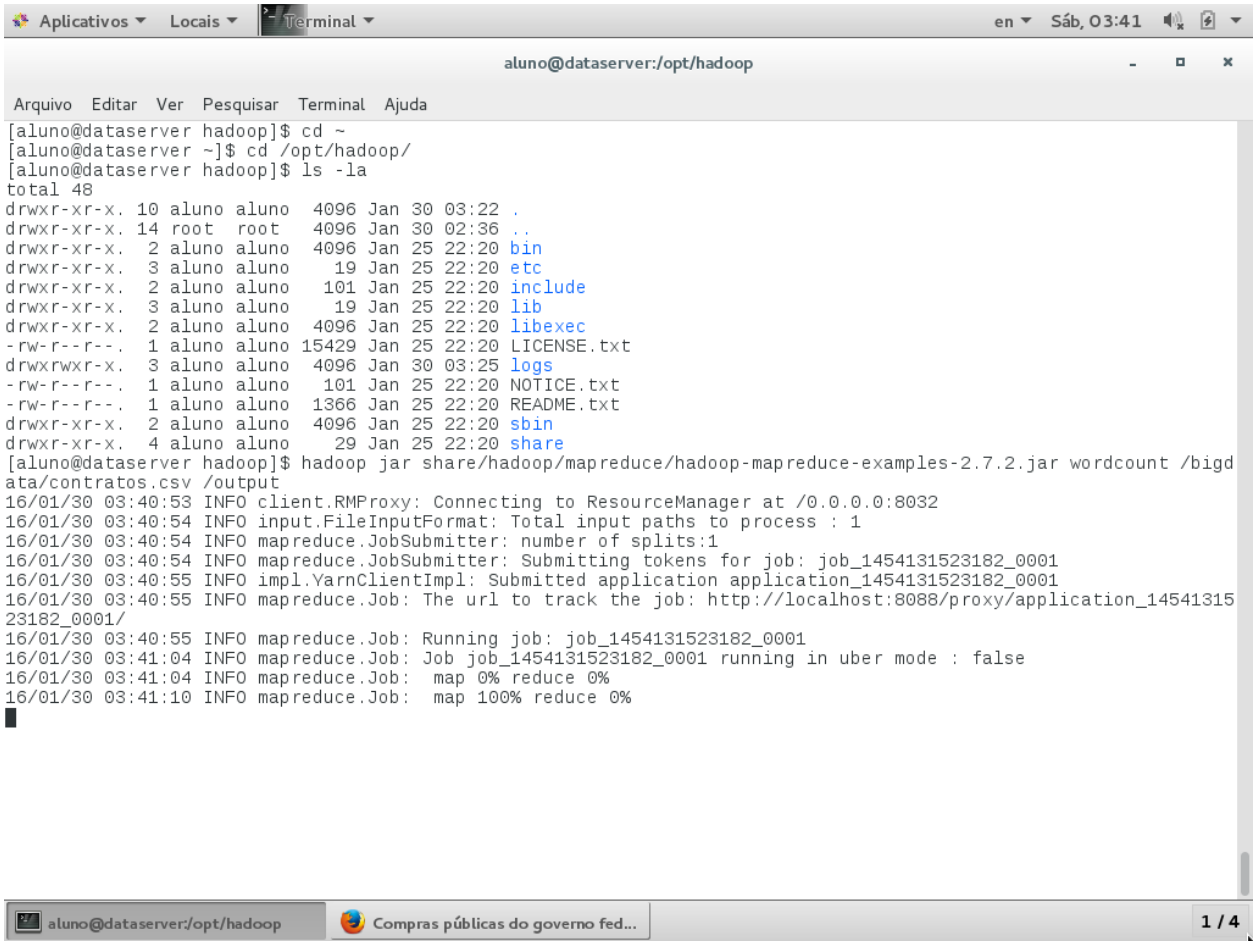
Ver o conteúdo do arquivo



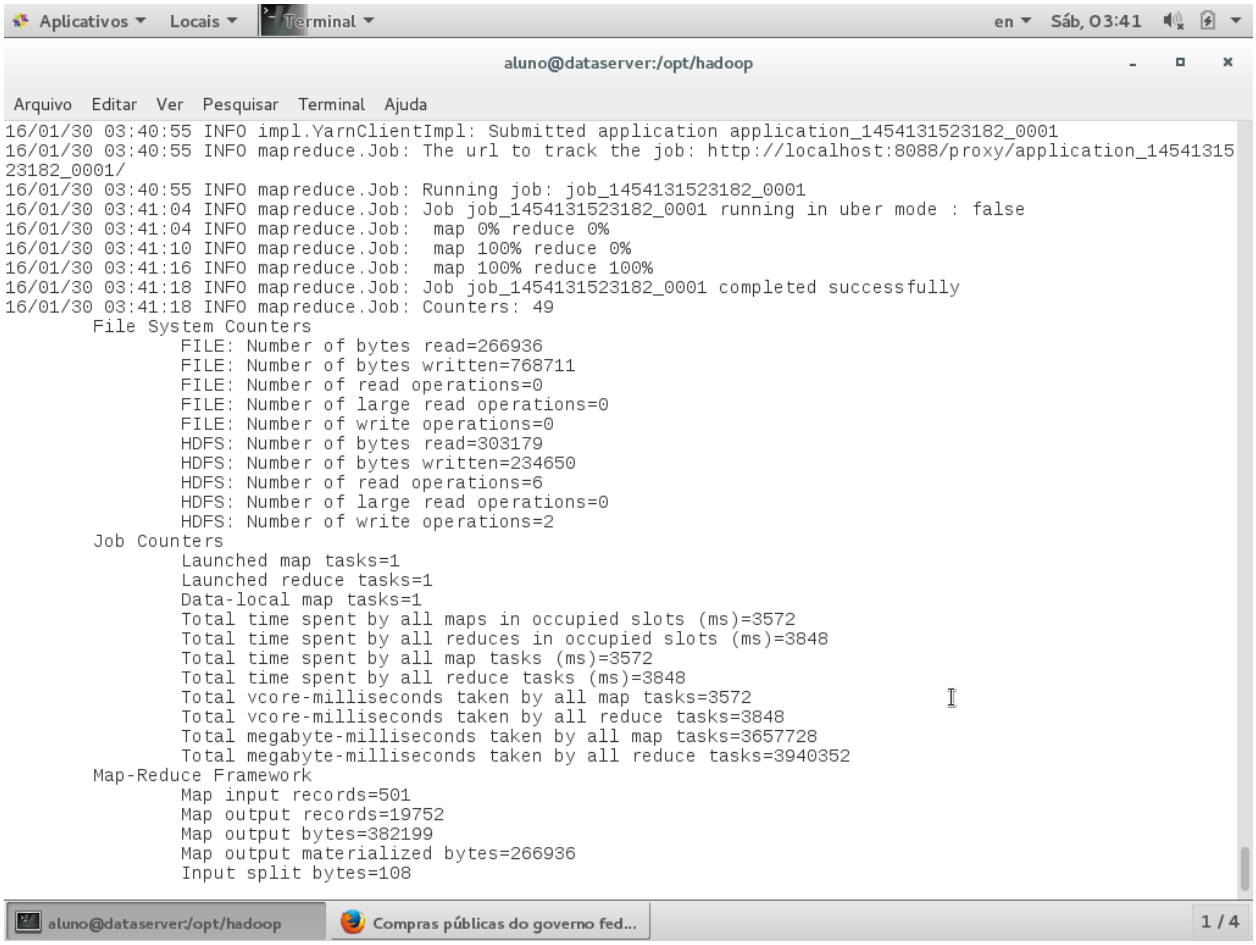
Conteúdo do arquivo já gravado no HDFS



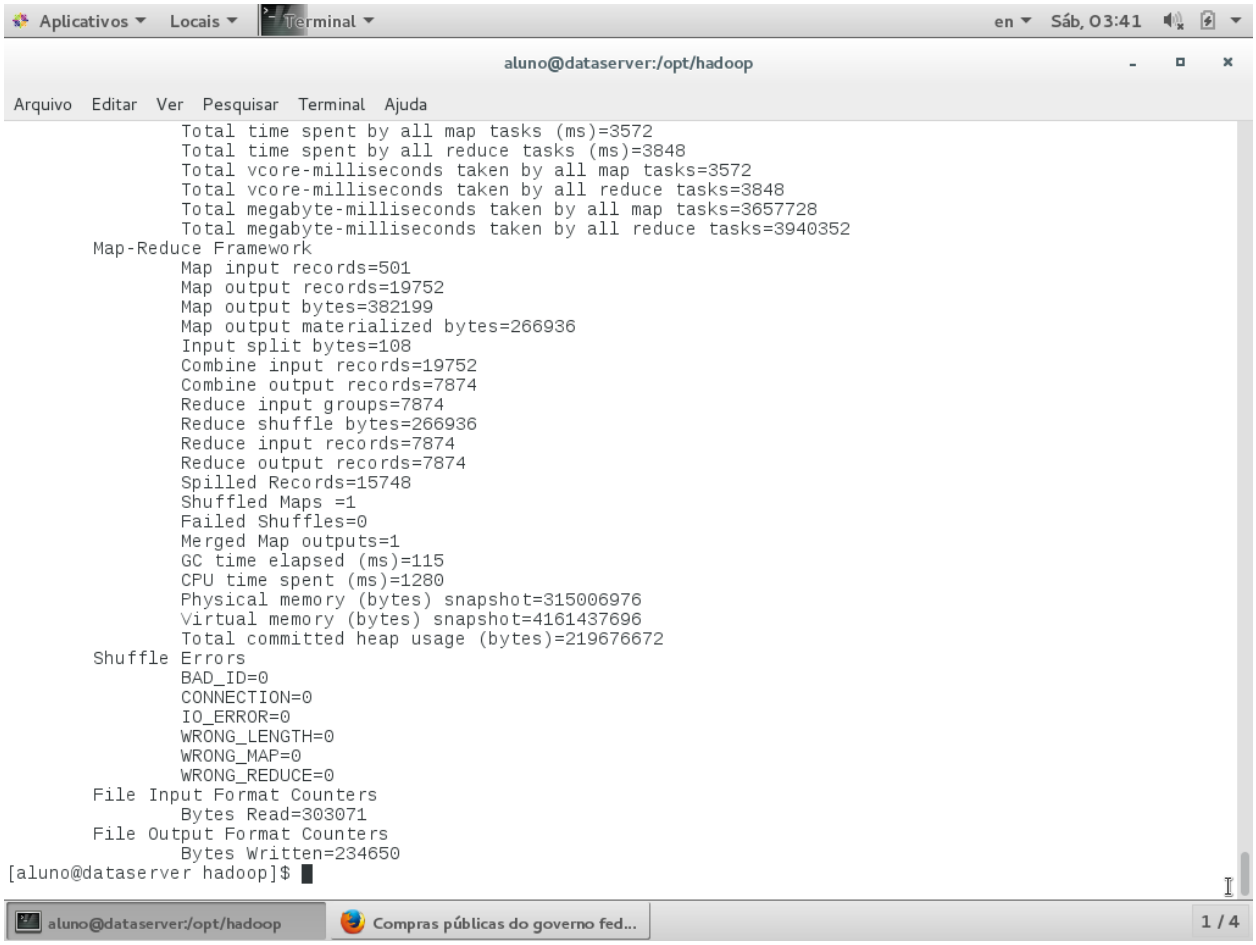
A instalação do Hadoop possui um job chamado wordcount, que pode ser usado como exemplo para processamento de Big Data. Basicamente, o job conta a ocorrência de cada palavra no arquivo. Vamos executar com o comando acima (a versão do arquivo .jar é a mesma versão do Hadoop que você instalou).



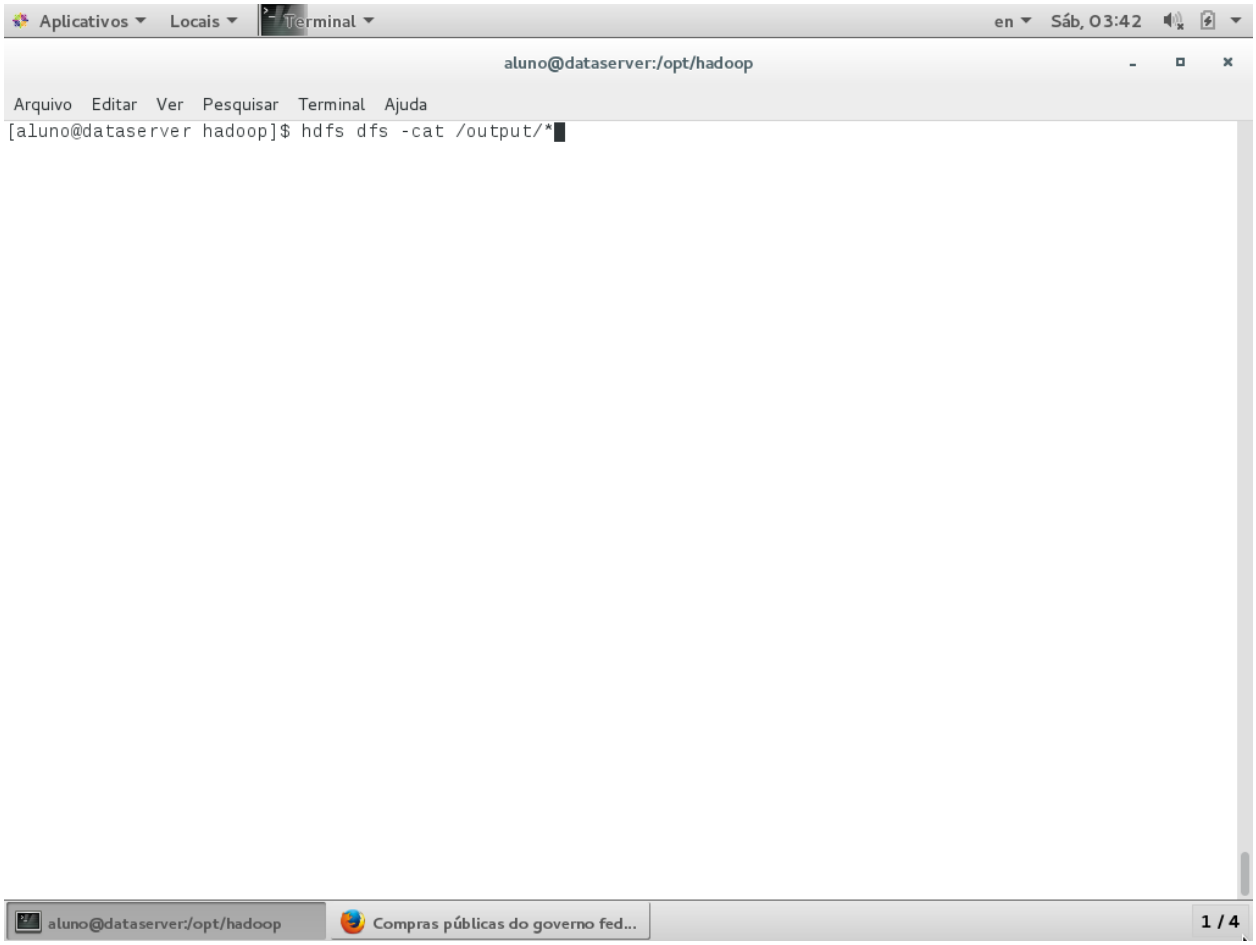
Job sendo processado



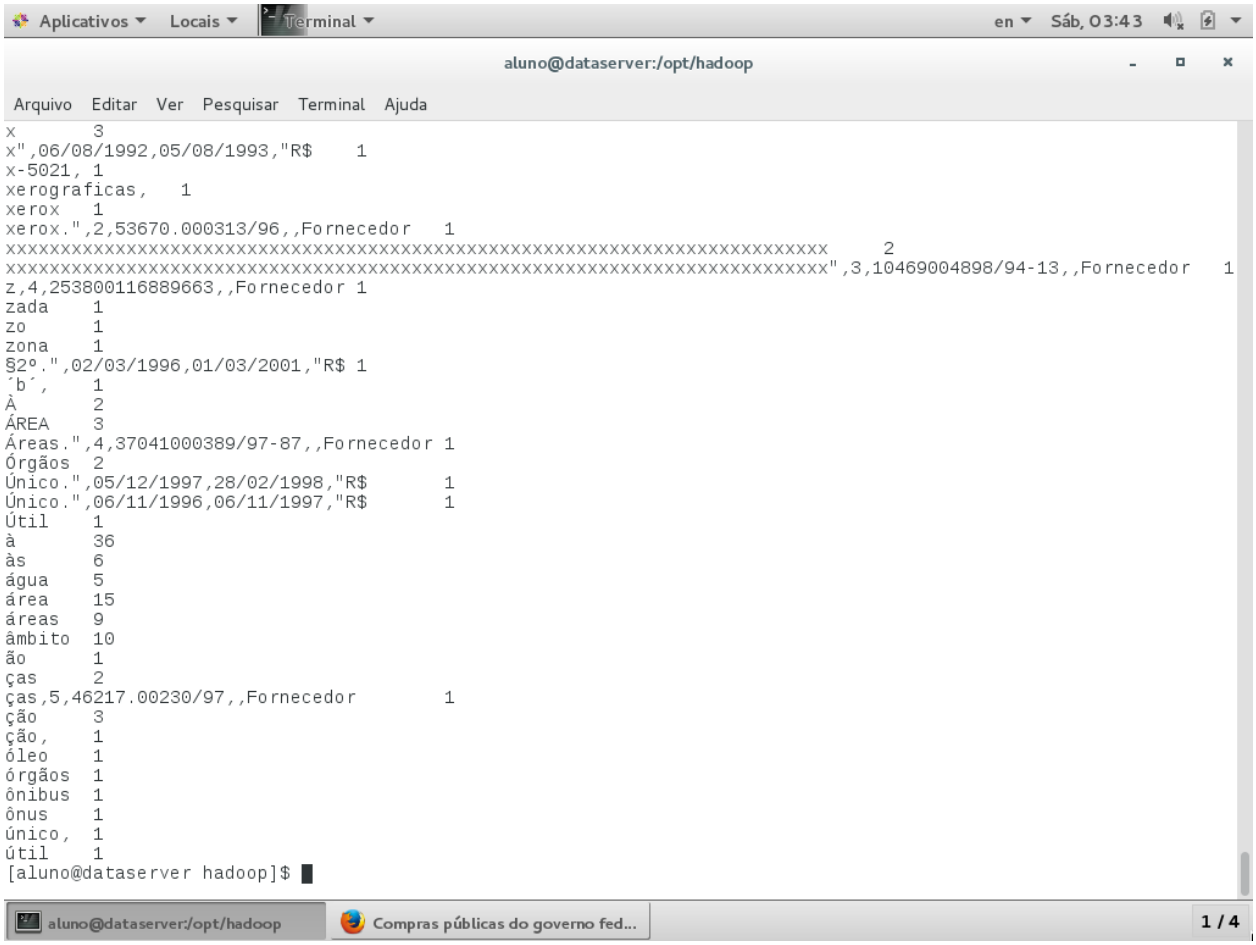
job processado com sucesso



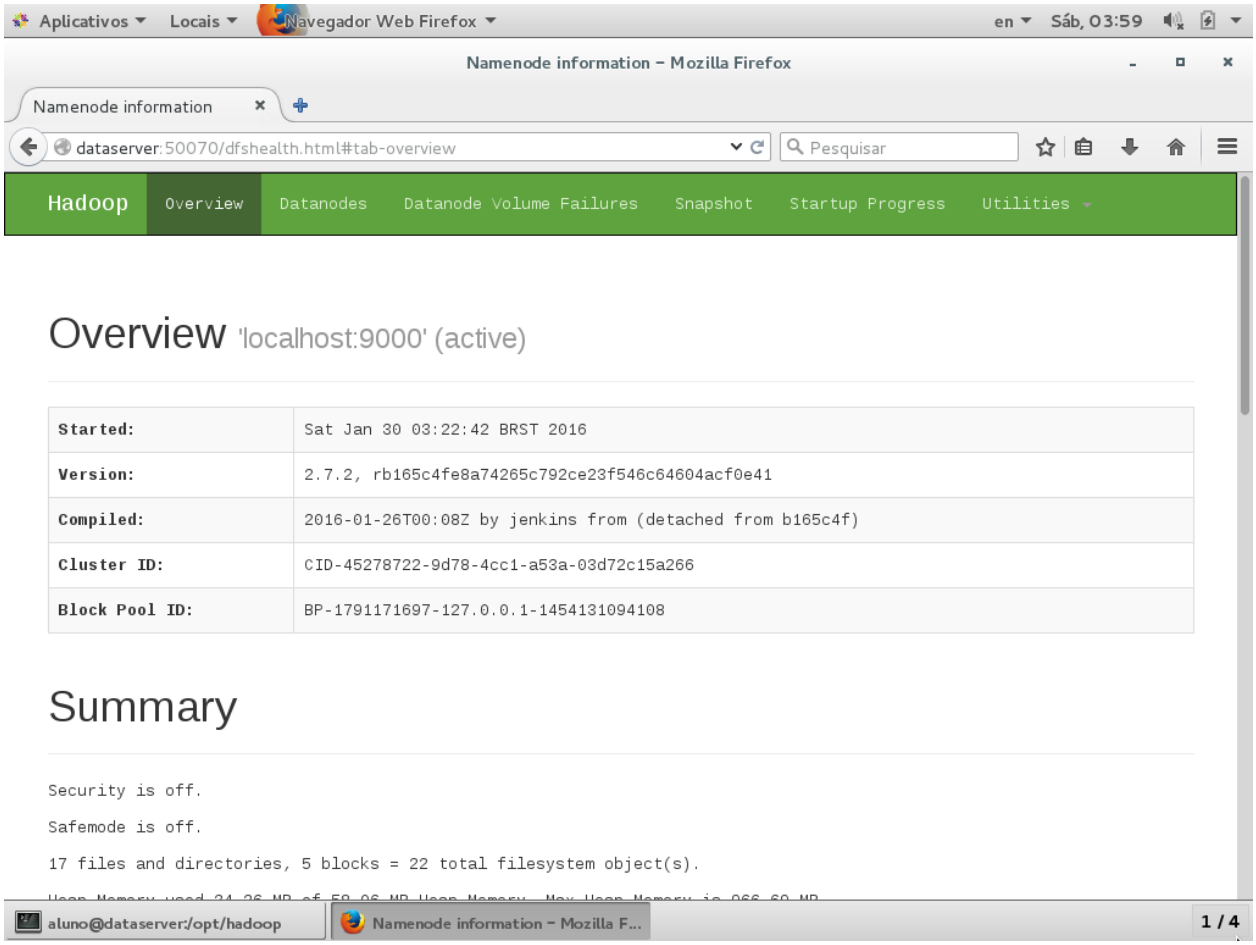
job processado com sucesso



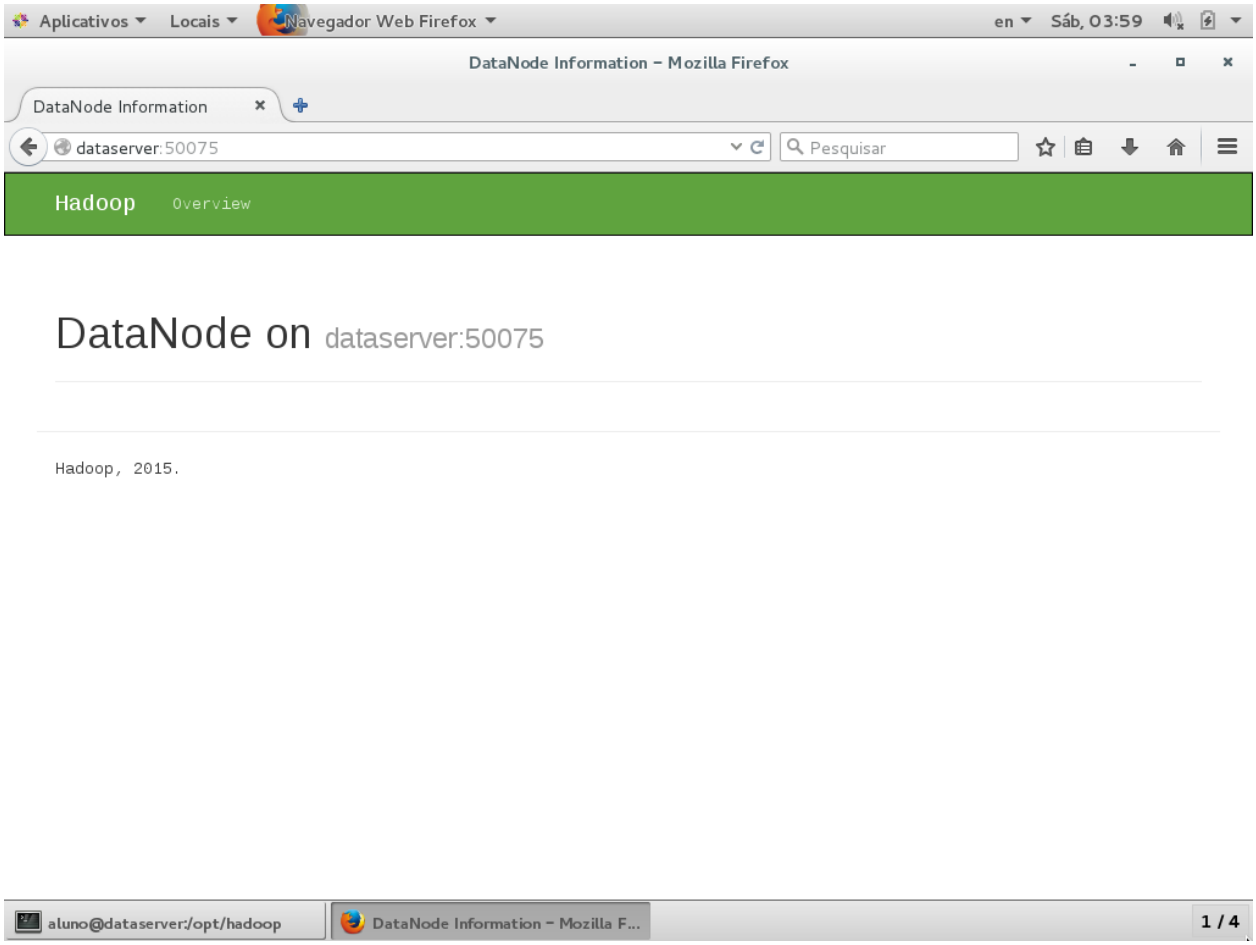
Vamos ver o resultado do processamento



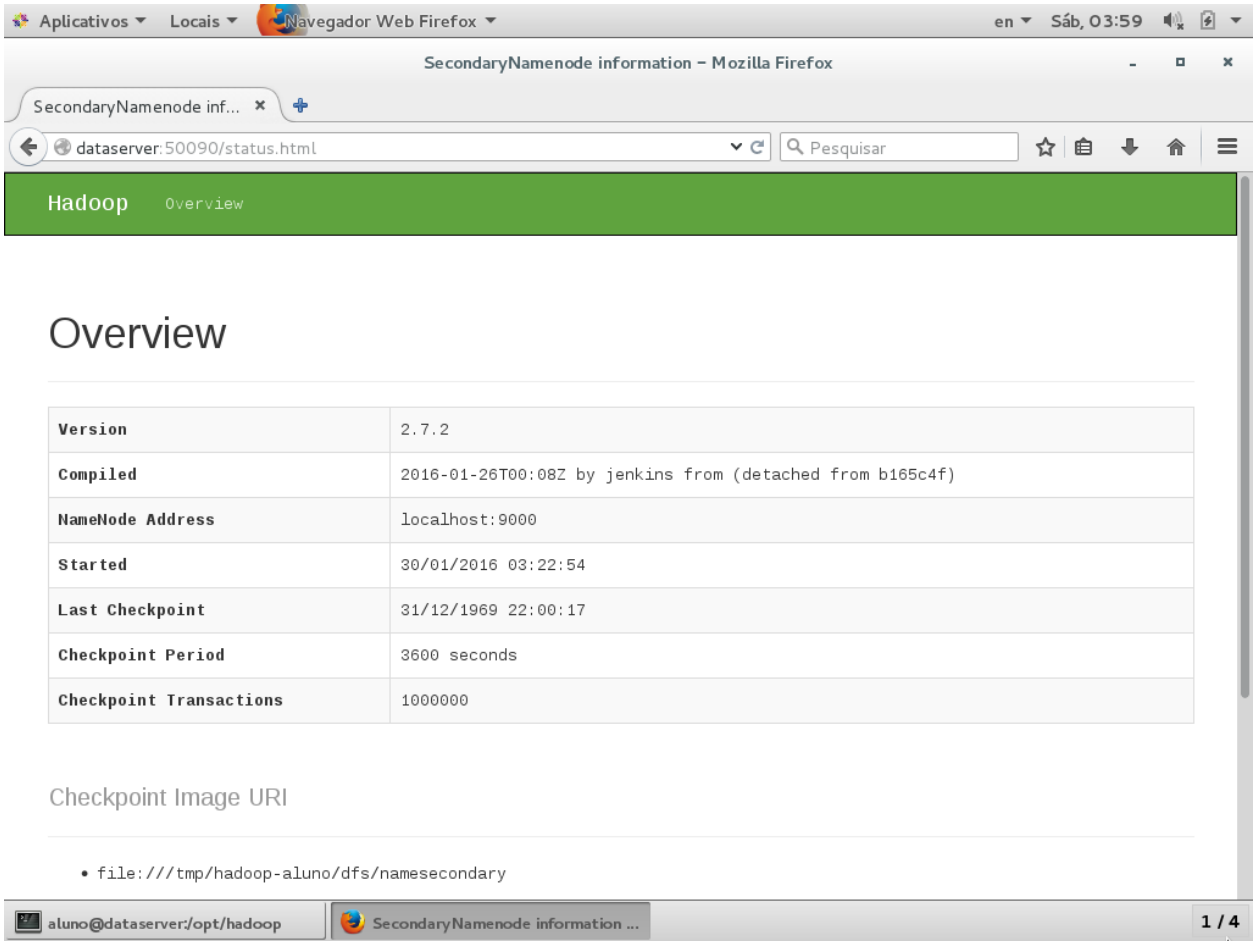
Arquivo processado. Número de ocorrência de cada palavra/termo no arquivo.



Acesso ao Hadoop pelo browser: http://dataserver:50070



Acesso ao Hadoop pelo browser: http://dataserver:50075



Acesso ao Hadoop pelo browser: http://dataserver:50090

Terceiro checkpoint:

Clique no meu File – Export Appliance.
Será gerada uma cópia de segurança da sua máquina virtual.

→ VM: DataServer-3.0.ova (Hadoop)

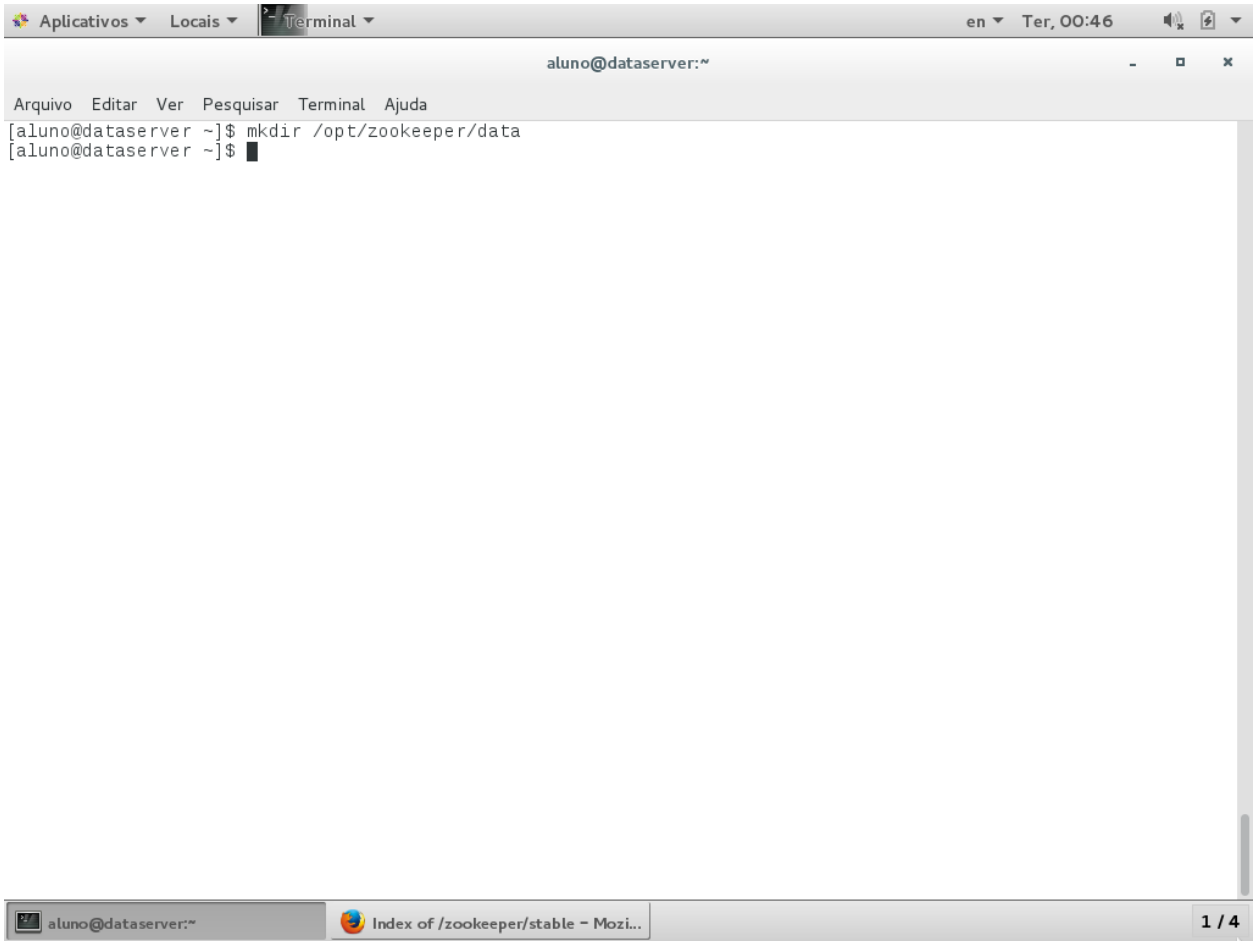
6. Instalação e Configuração do Zookeeper

6.1. Download e Instalação do Zookeeper

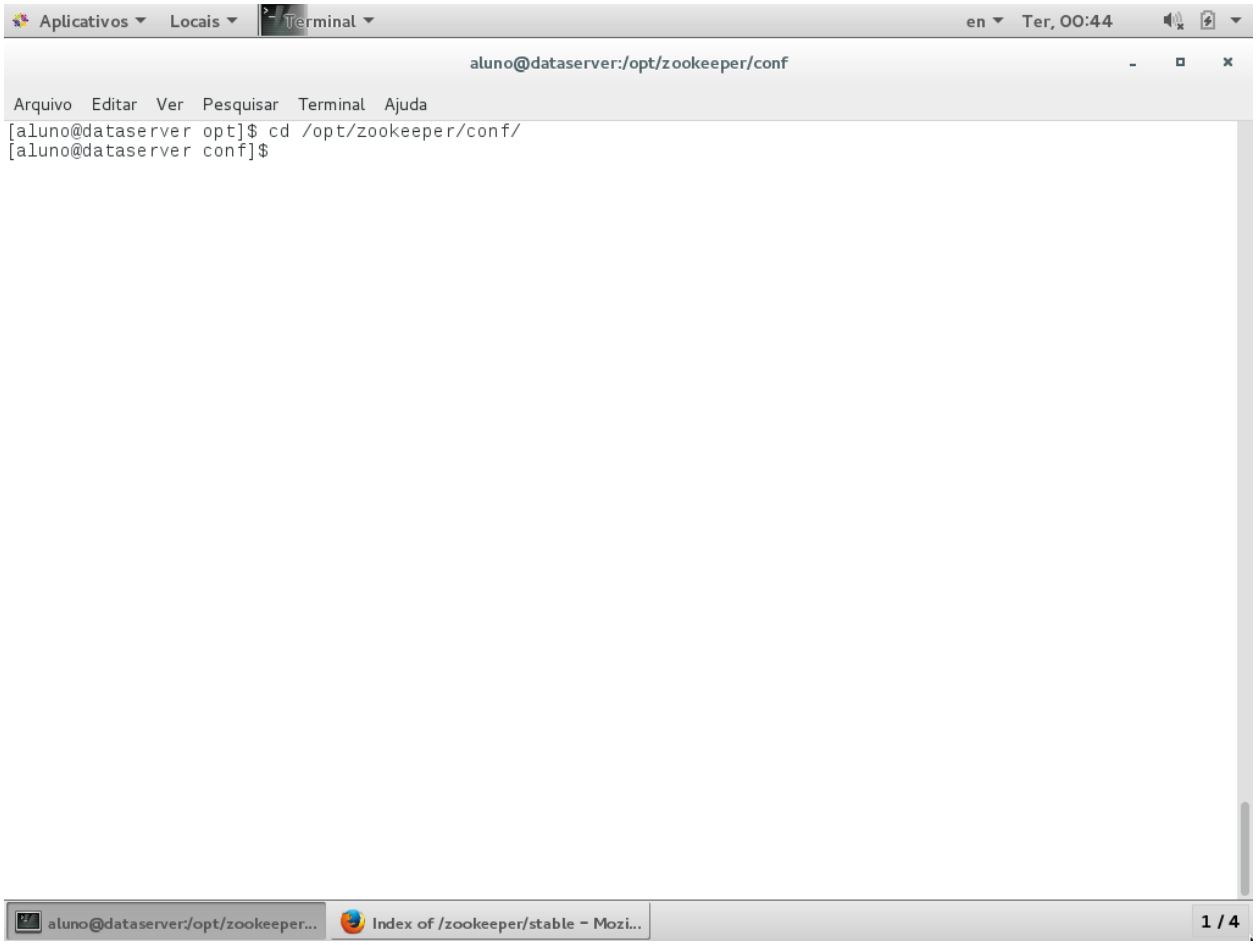
Download do Zookeeper – Versão 3.5.5

Faça o download, descompacte o arquivo e mova o diretório para /opt/zookeeper da mesma forma como você fez com o Java JDK e com o Hadoop.

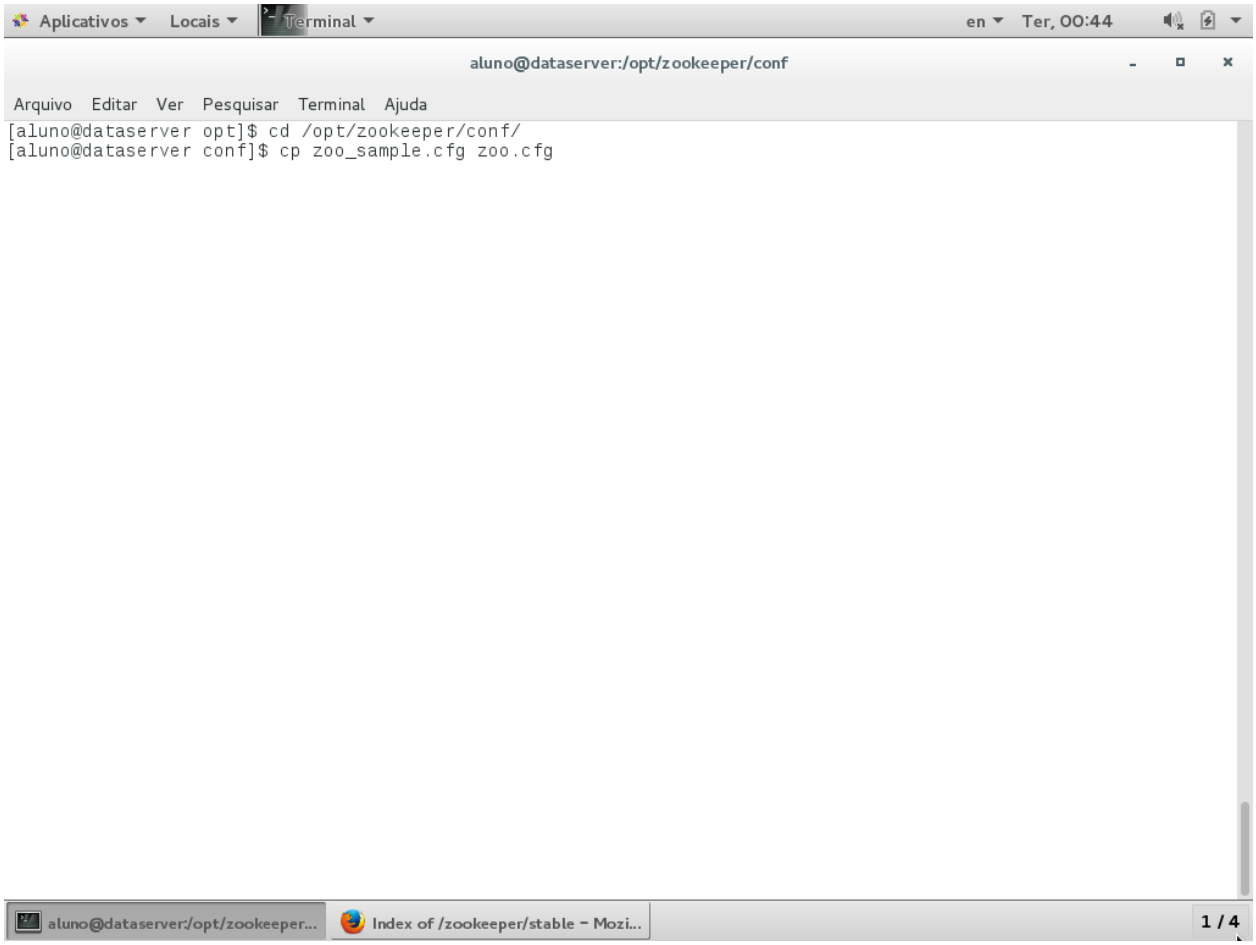
6.2. Configurando do Zookeeper



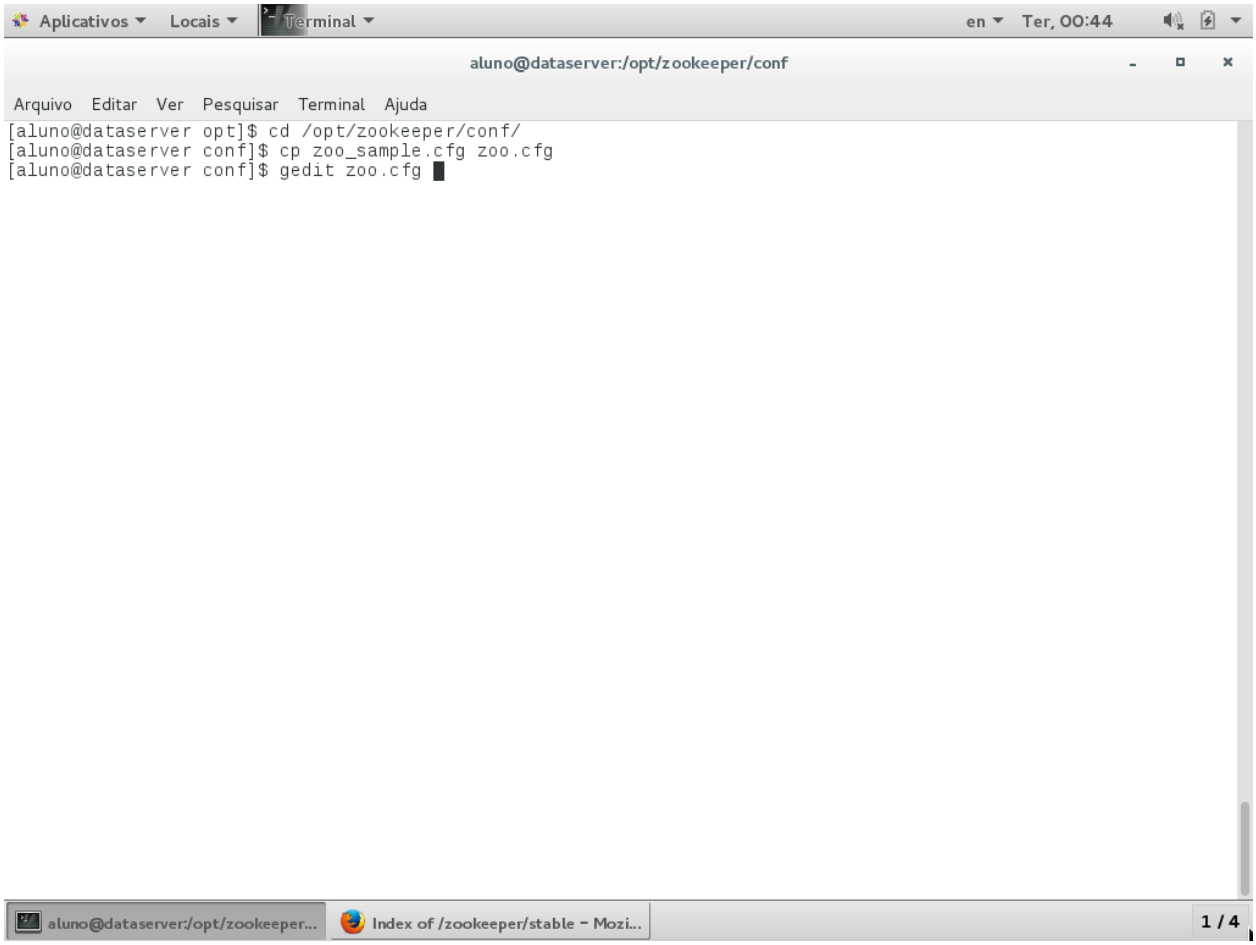
Criar o diretório **data** dentro de /opt/zookeeper



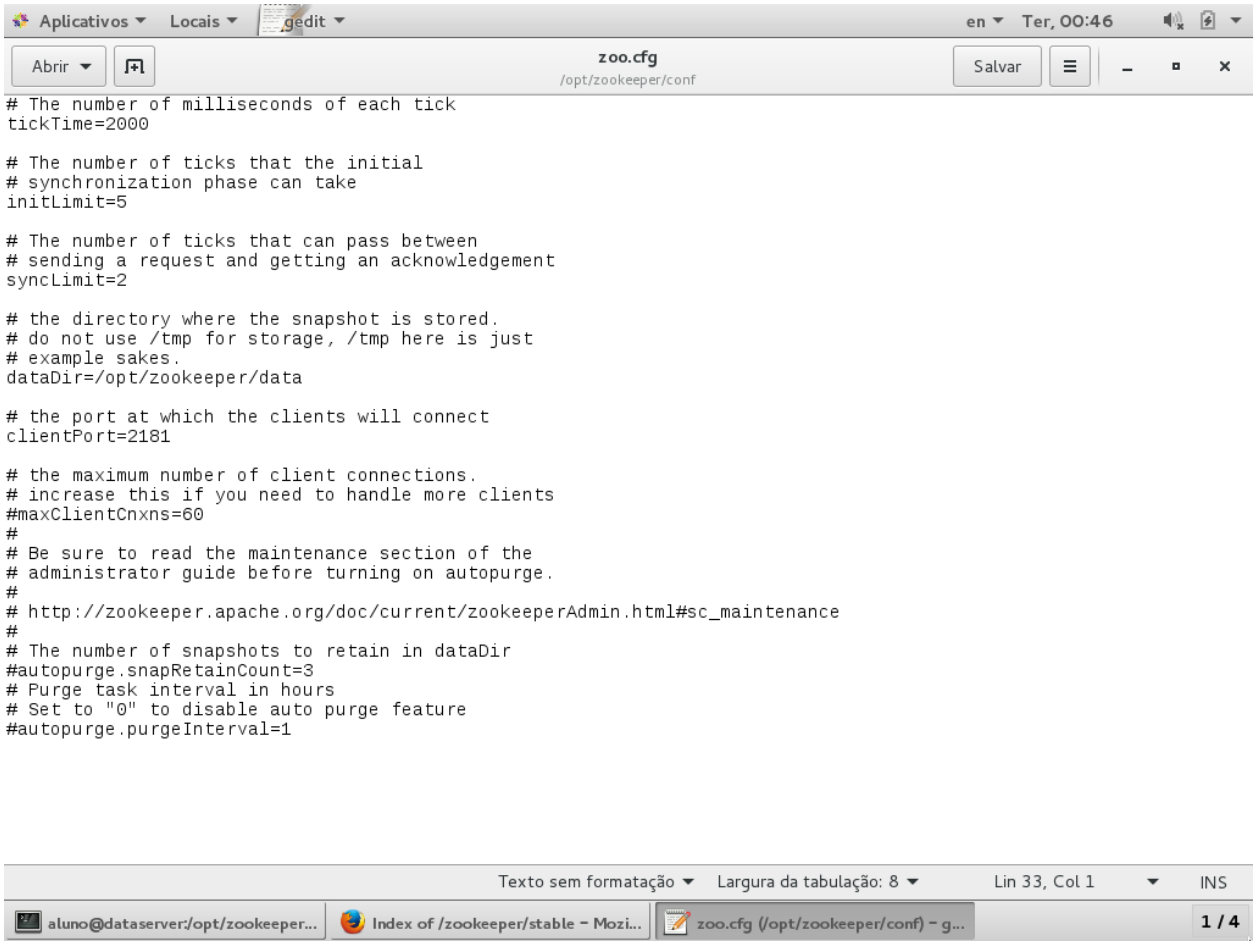
Acessar o diretório /opt/zookeeper/conf



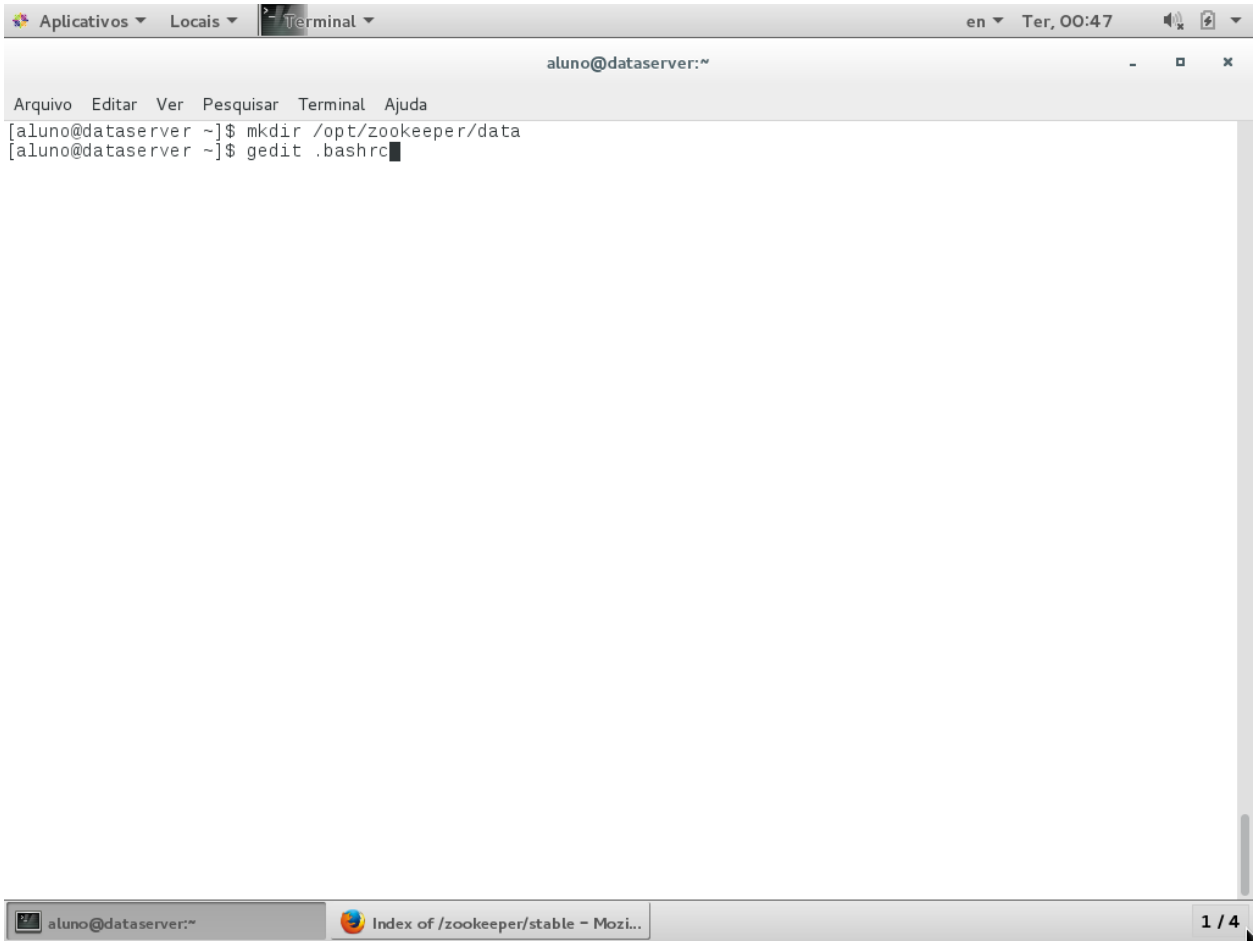
A partir do arquivo template, gerar o arquivo zoo.cfg



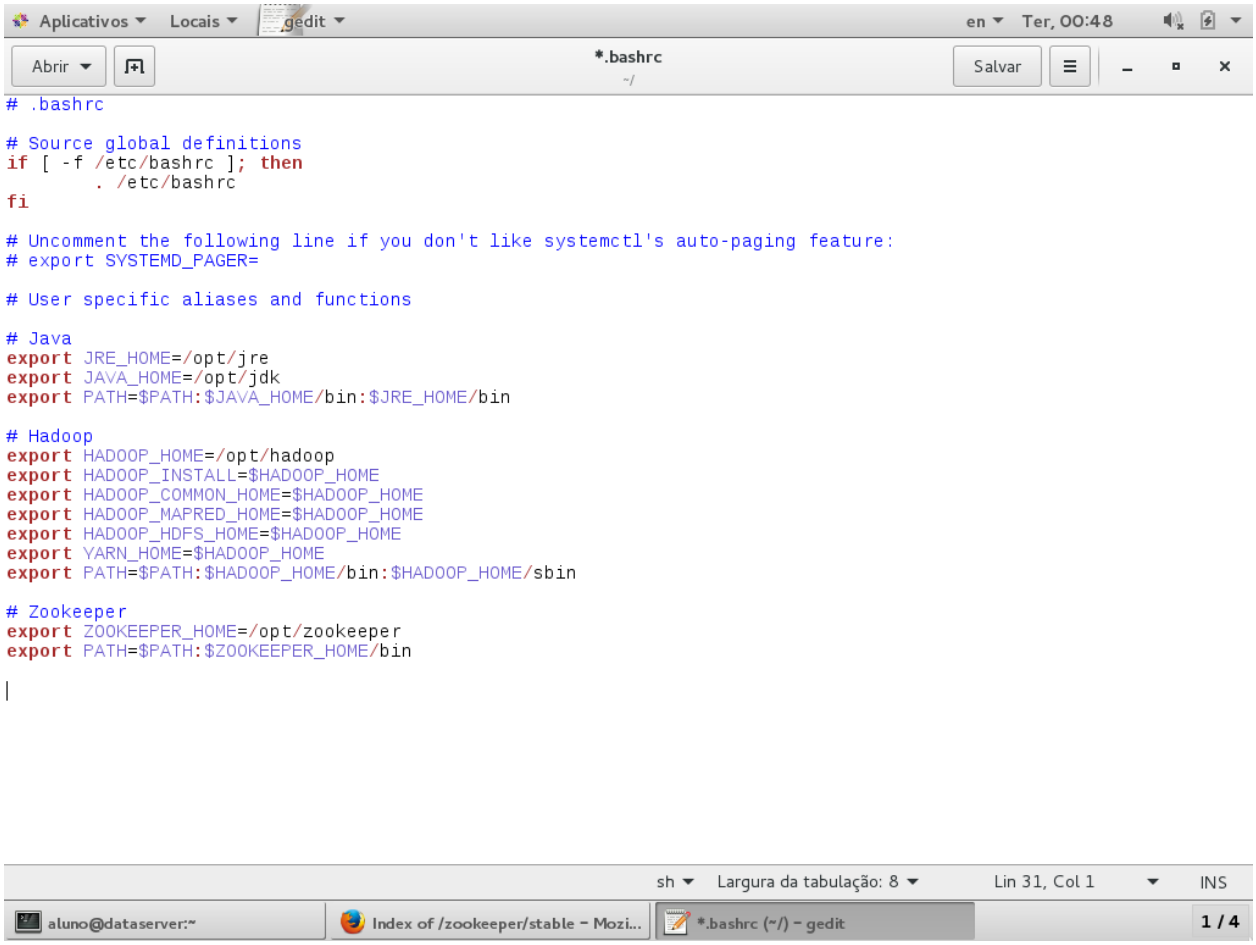
Editar o arquivo zoo.cfg



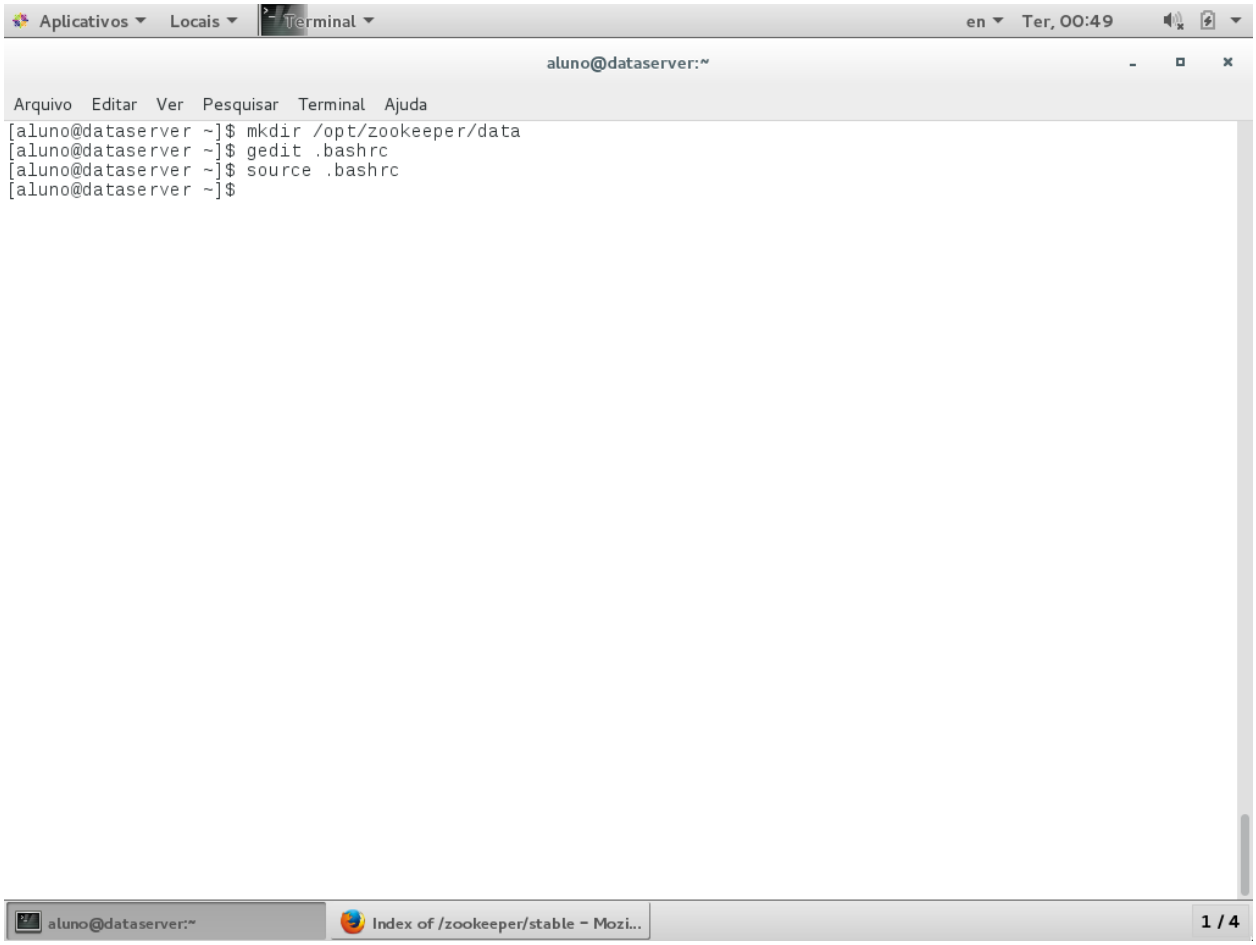
Editar o arquivo conforme tela acima



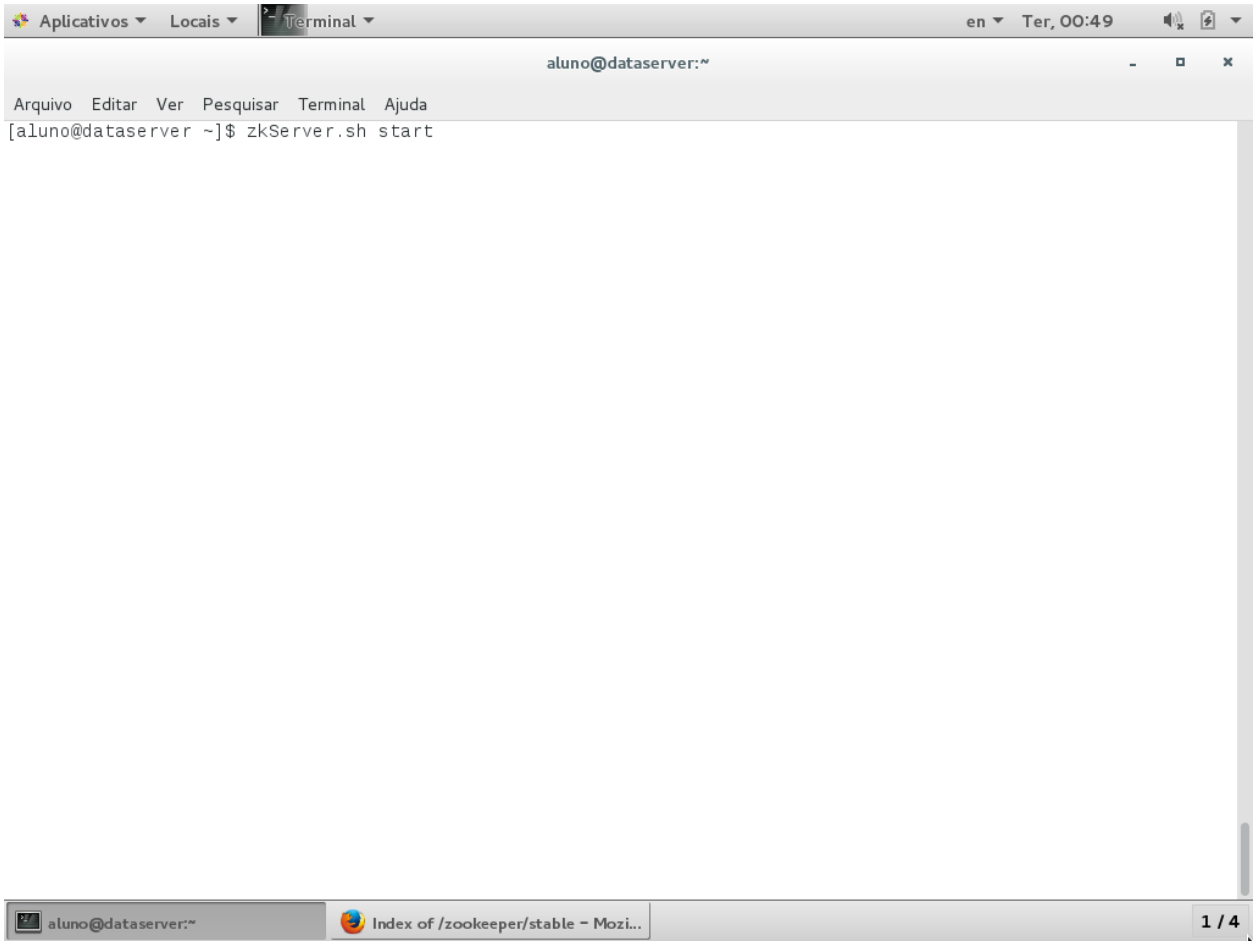
Incluir variáveis Zookeeper no /home/hadoop/.bashrc



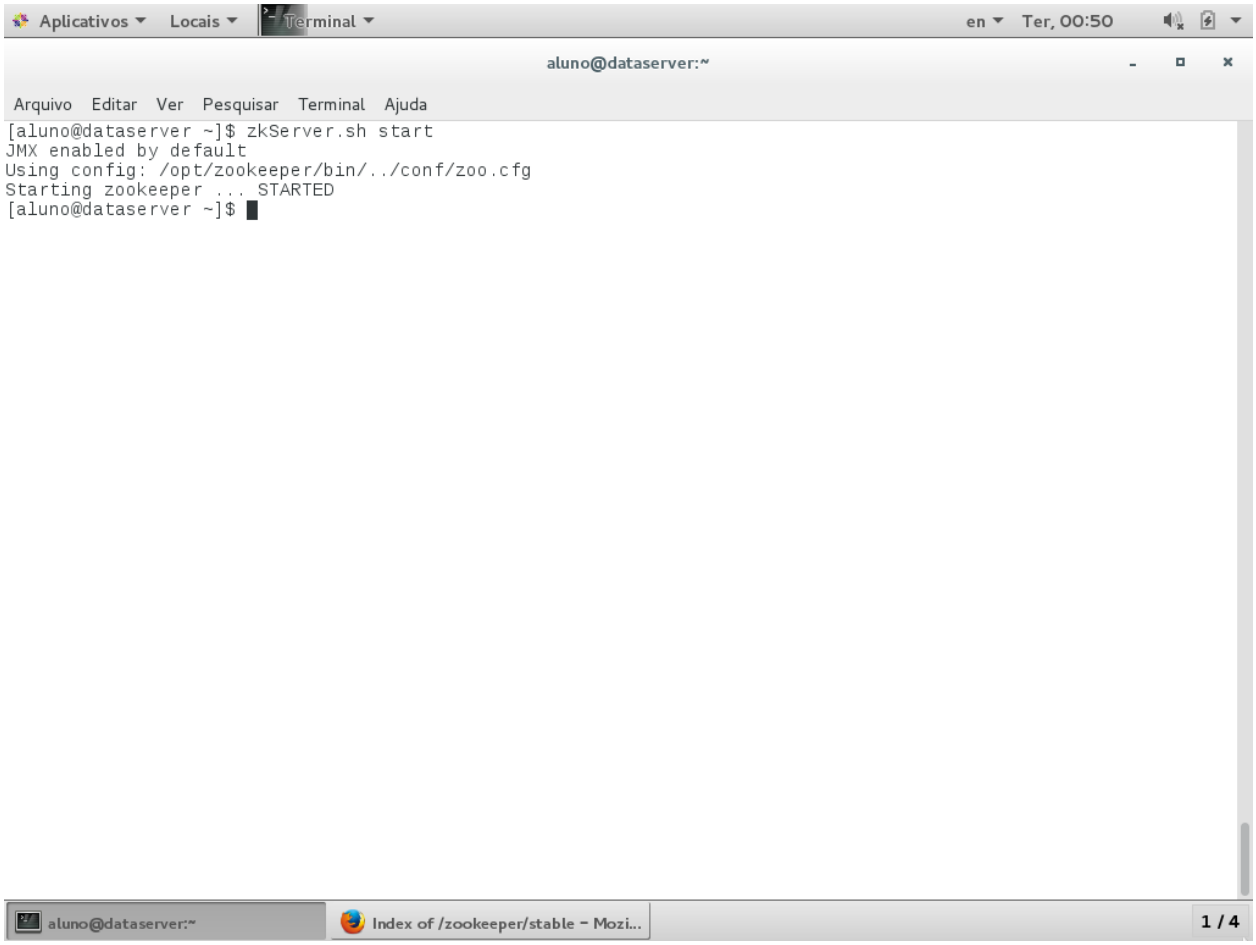
Variáveis Zookeeper



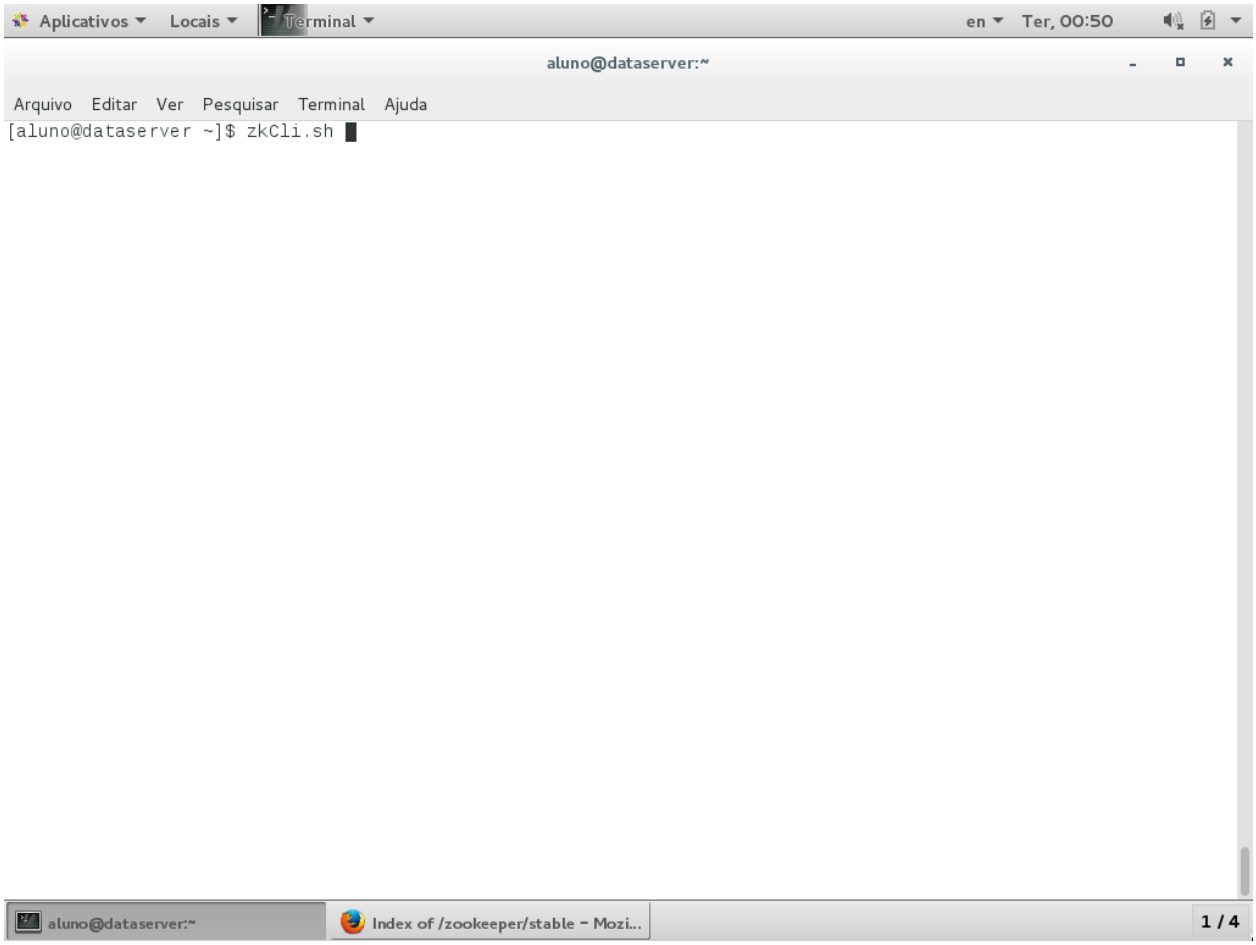
source .bashrc



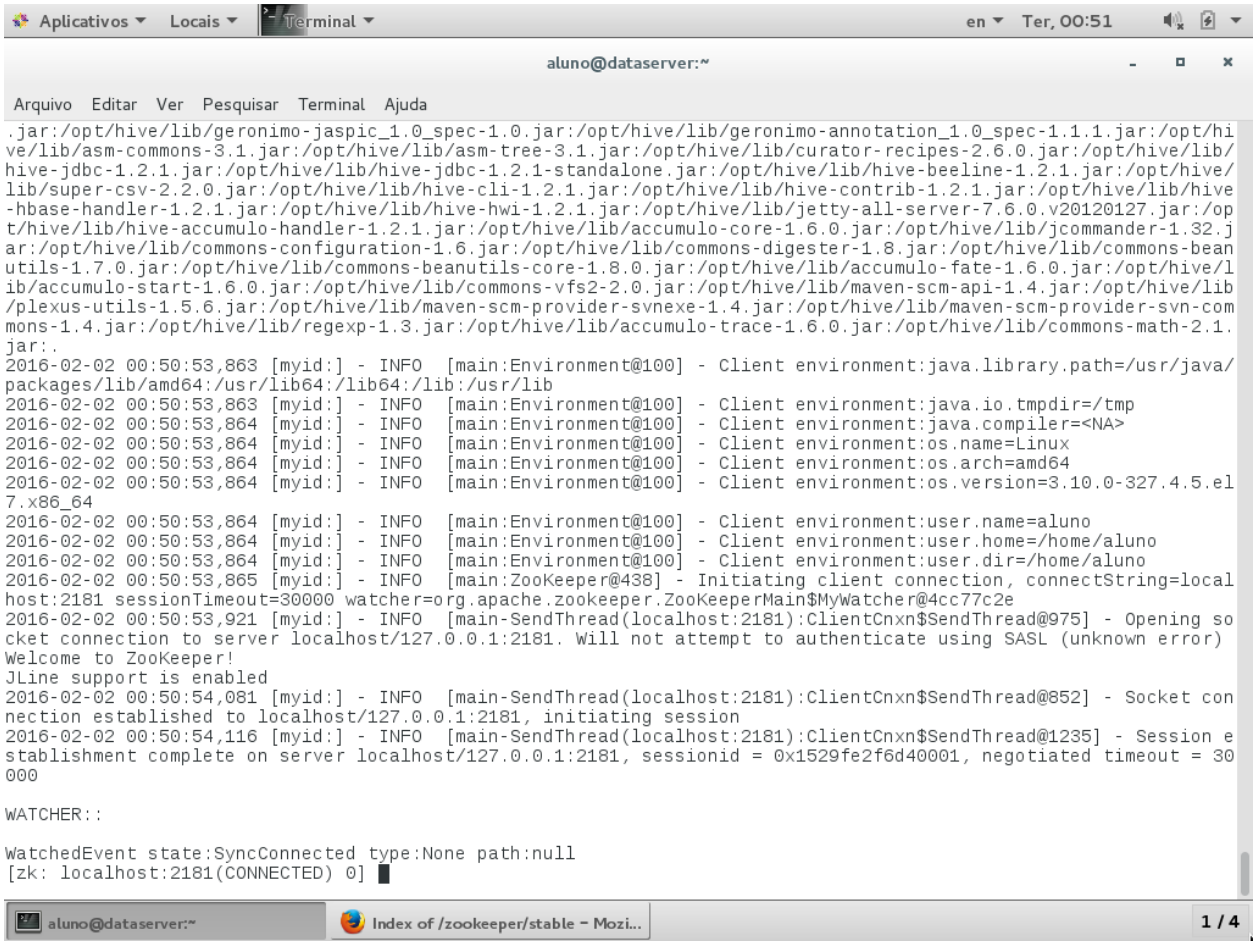
Iniciar o zookeeper



Serviço iniciado



Iniciar o Zookeeper Command Line Interface (CLI)

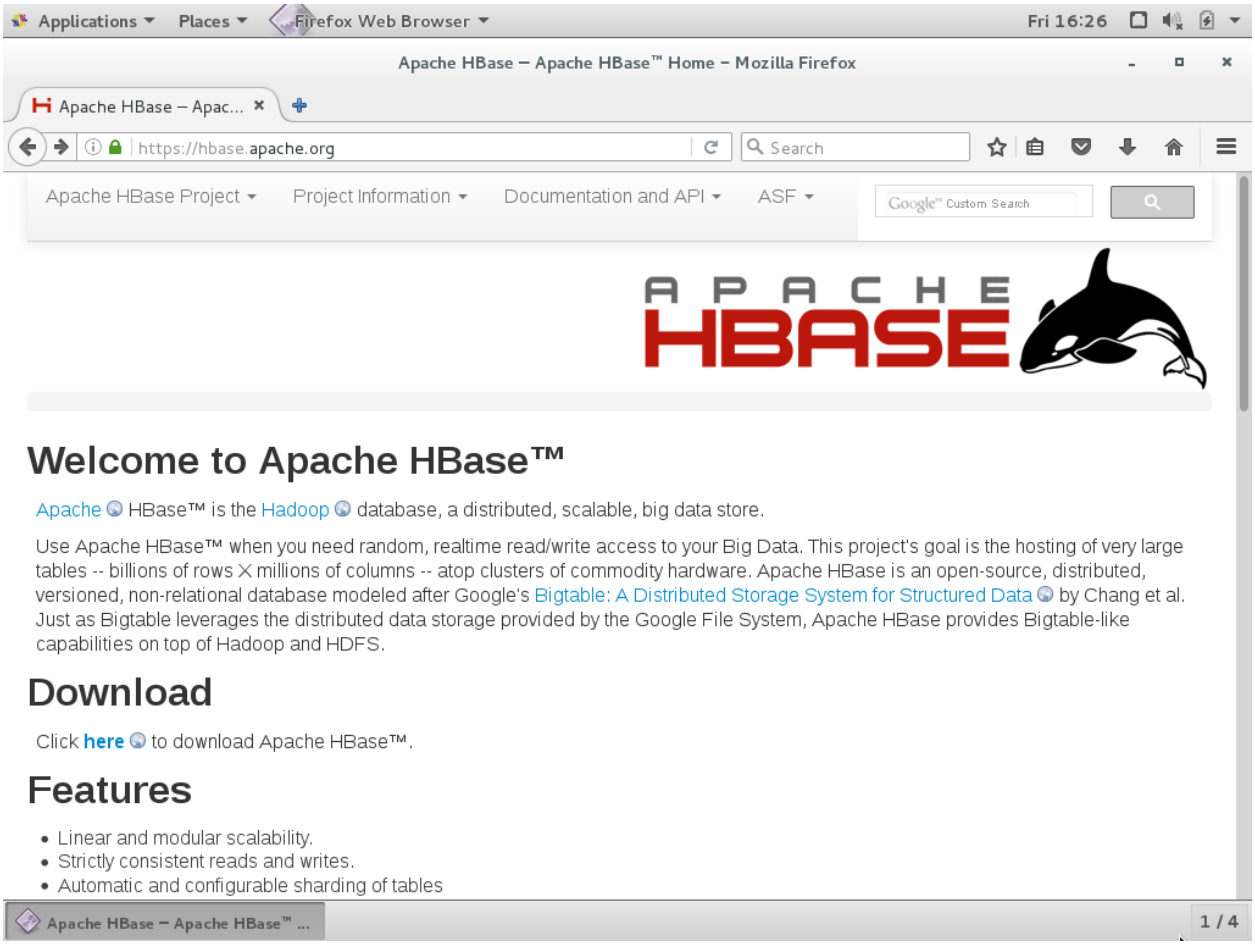


CLI iniciado

7. Instalação e Configuração do HBase

Podemos instalar HBase em qualquer um dos três modos: Standalone mode, Pseudo Distributed mode e Fully Distributed mode.

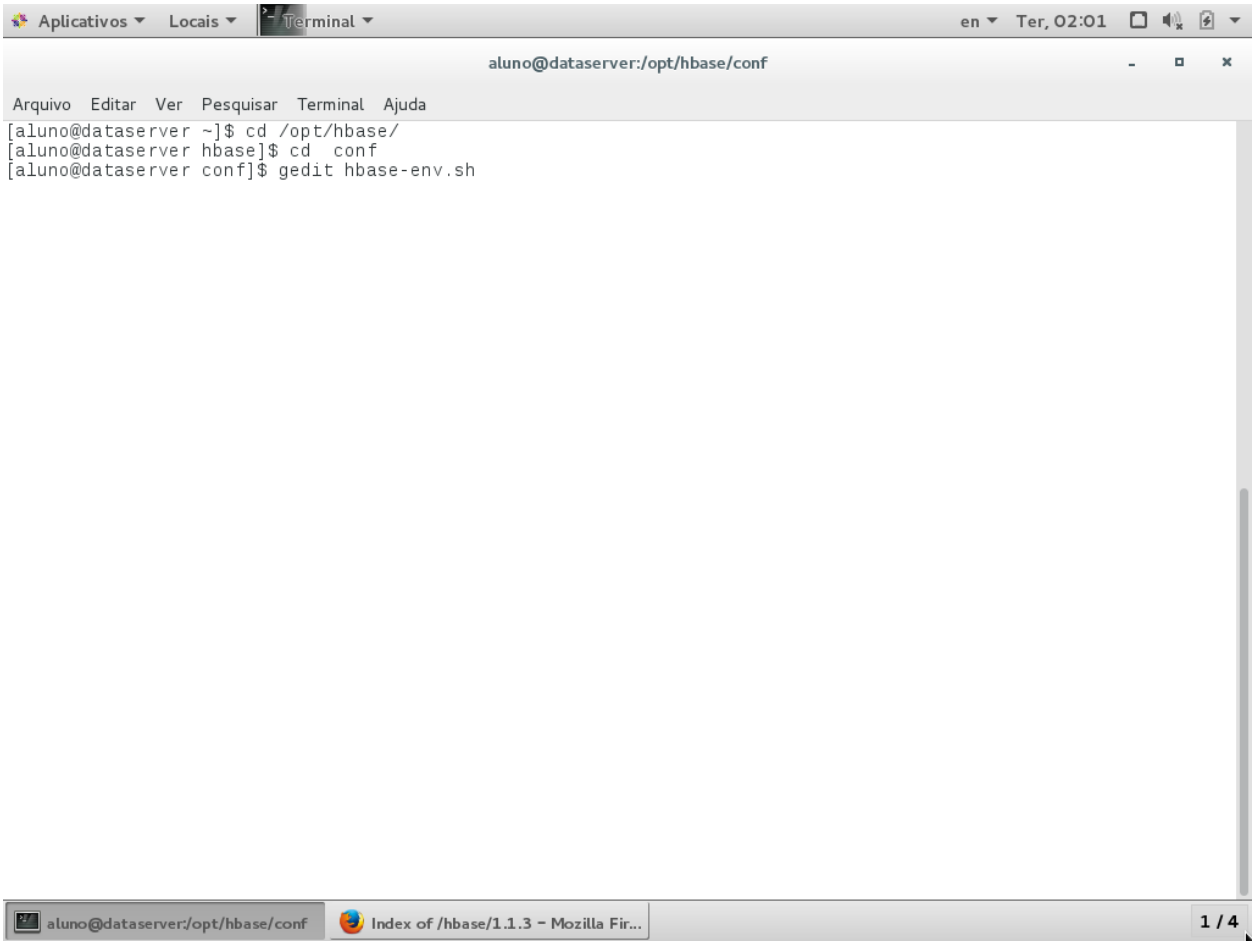
7.1. Download e Instalação do HBase



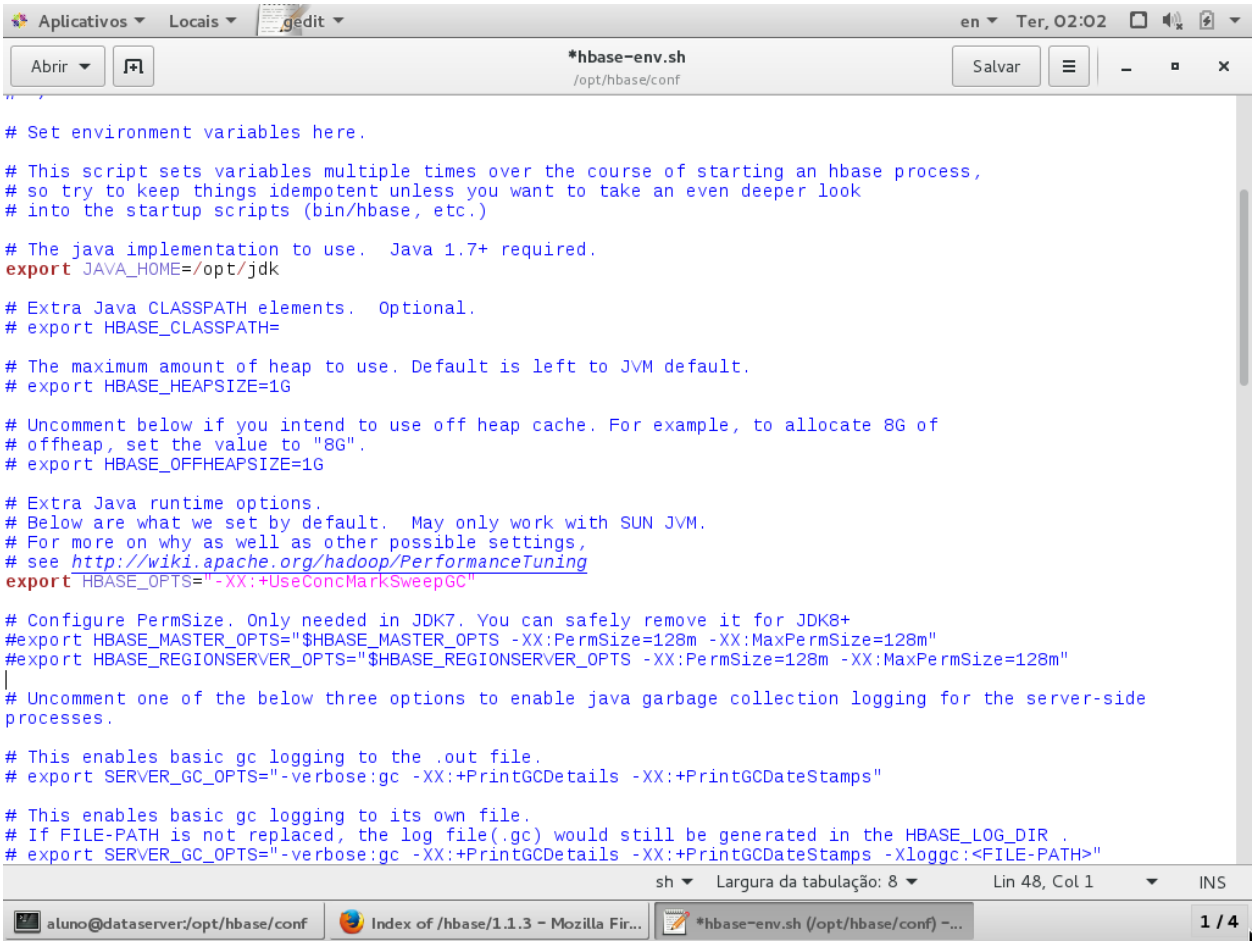
Download do Hbase – Versão 2.2.0

Faça o download, descompacte o arquivo e mova o diretório para /opt/hbase da mesma forma como você fez com o Java JDK e com o Hadoop.

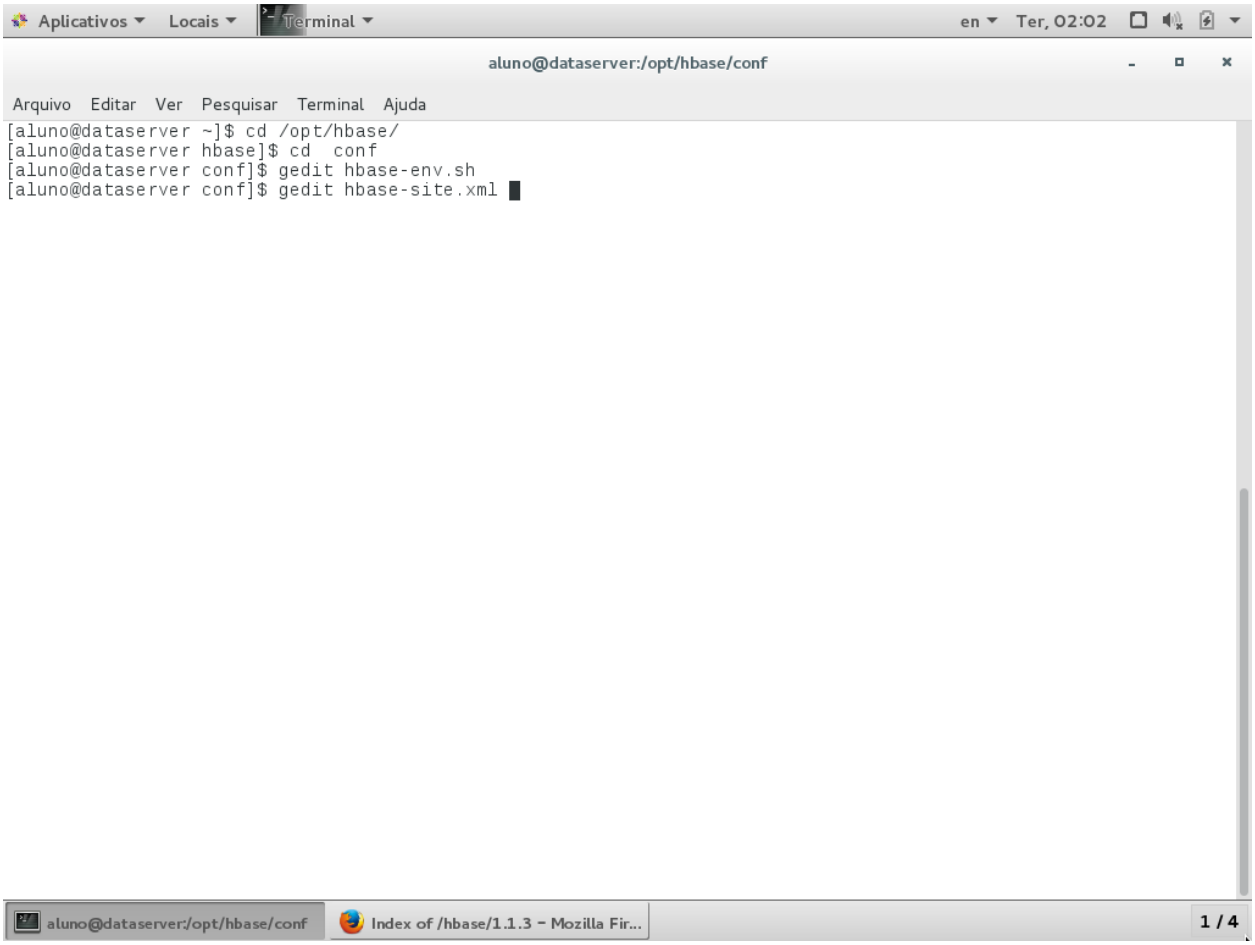
7.2. Configurando o HBase



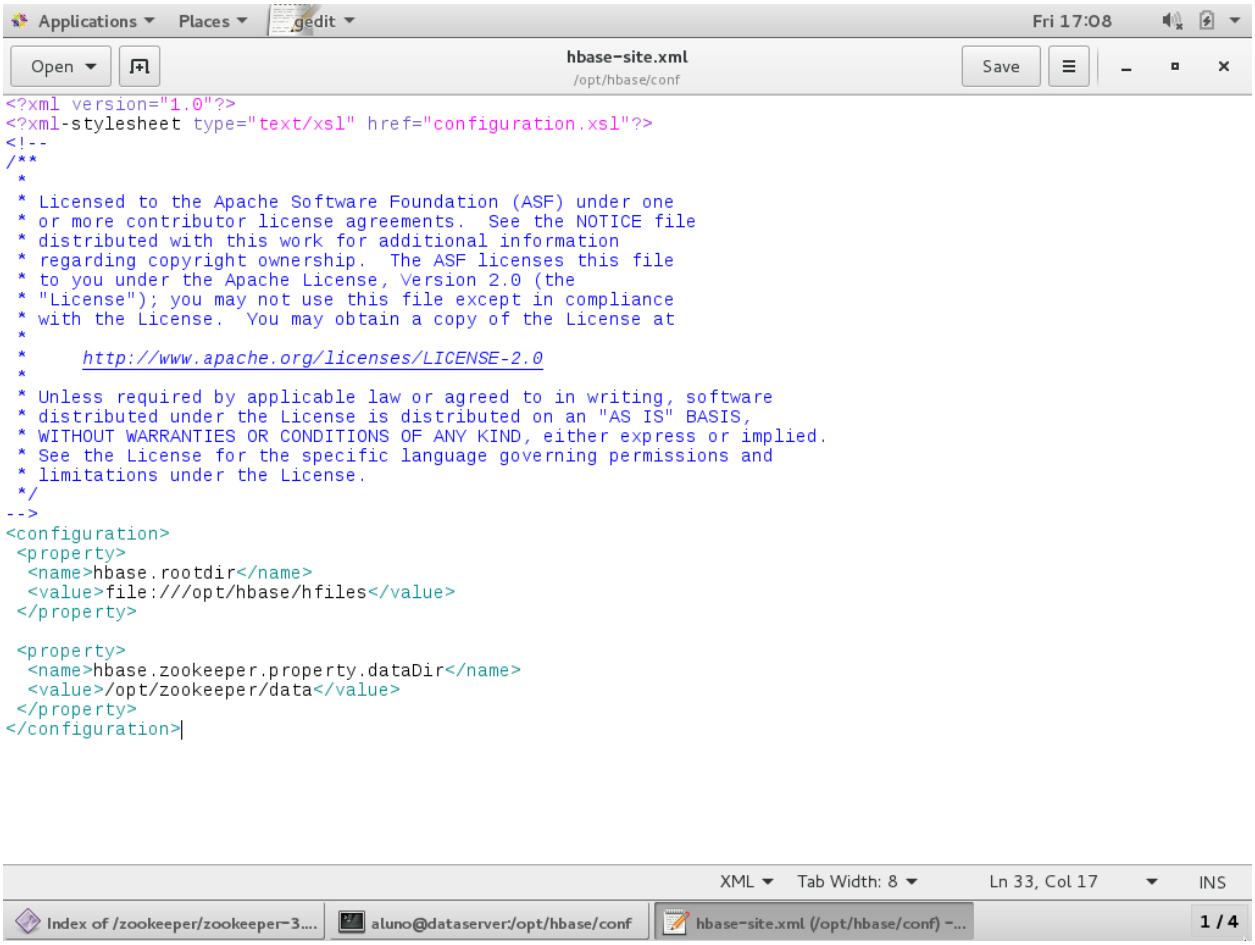
No diretório /opt/hbase/conf, editar o arquivo hbase-env.sh



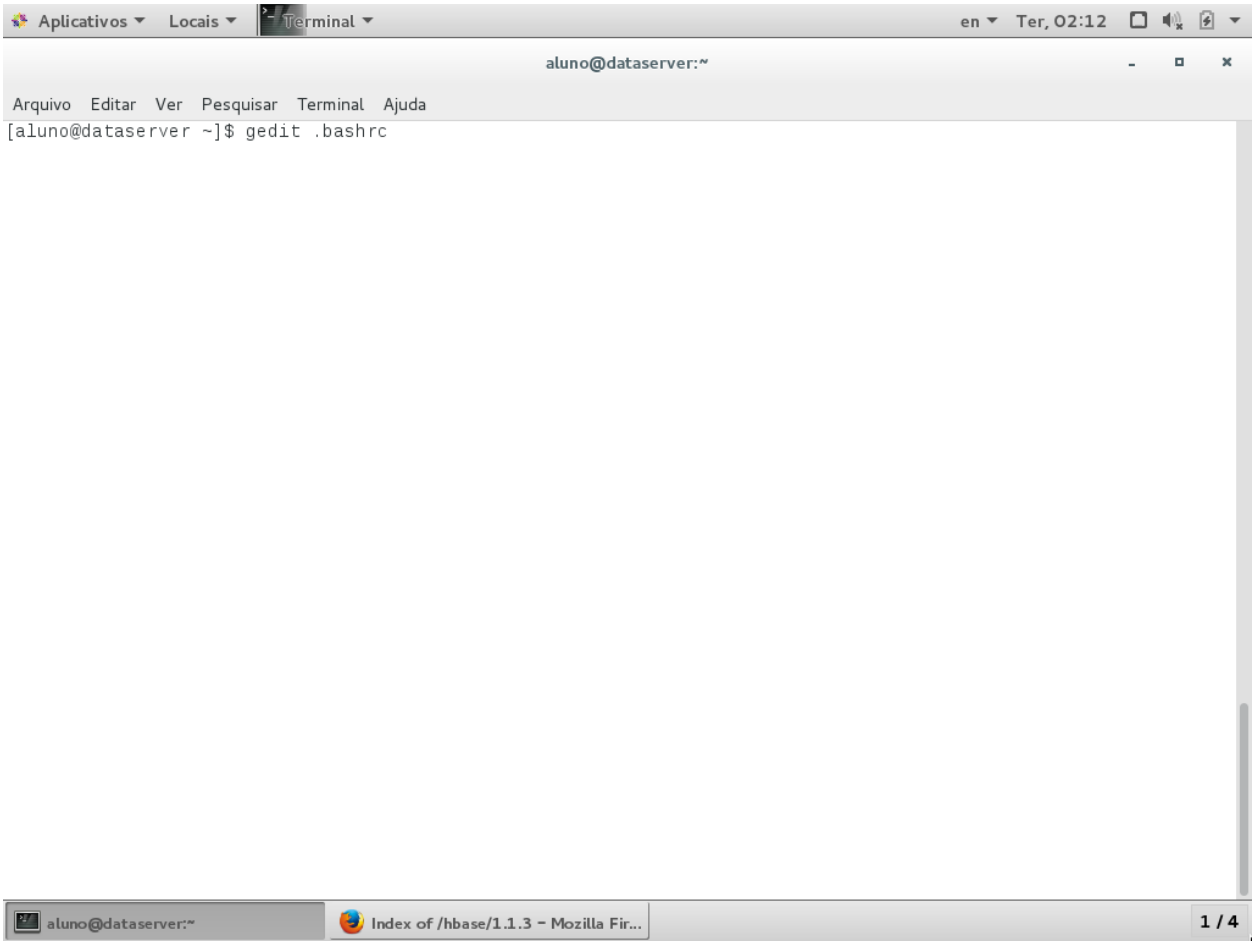
Editar o PATH do Java e comentar as linhas do PermSize



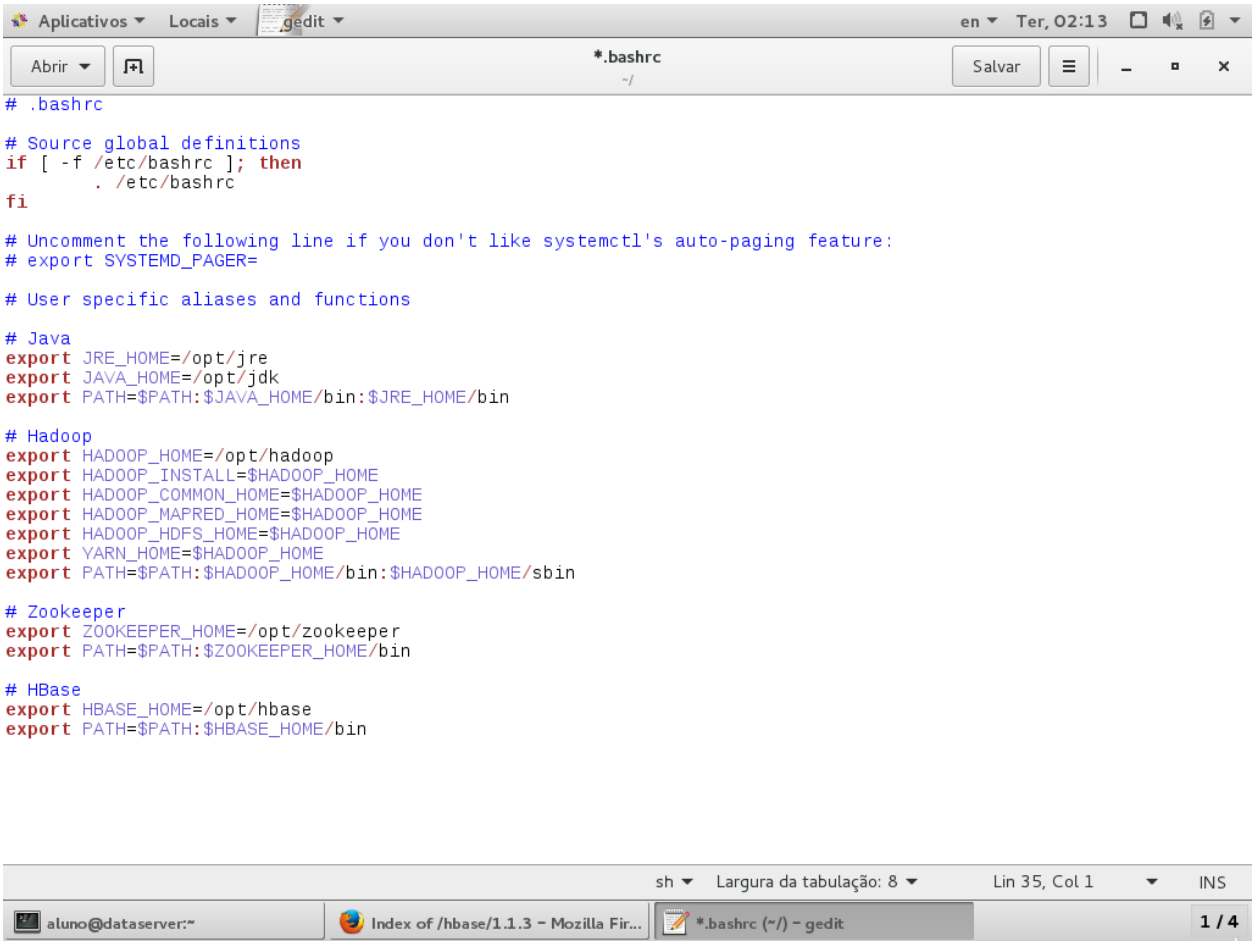
No mesmo diretório conf, editar o arquivo hbase-site.xml



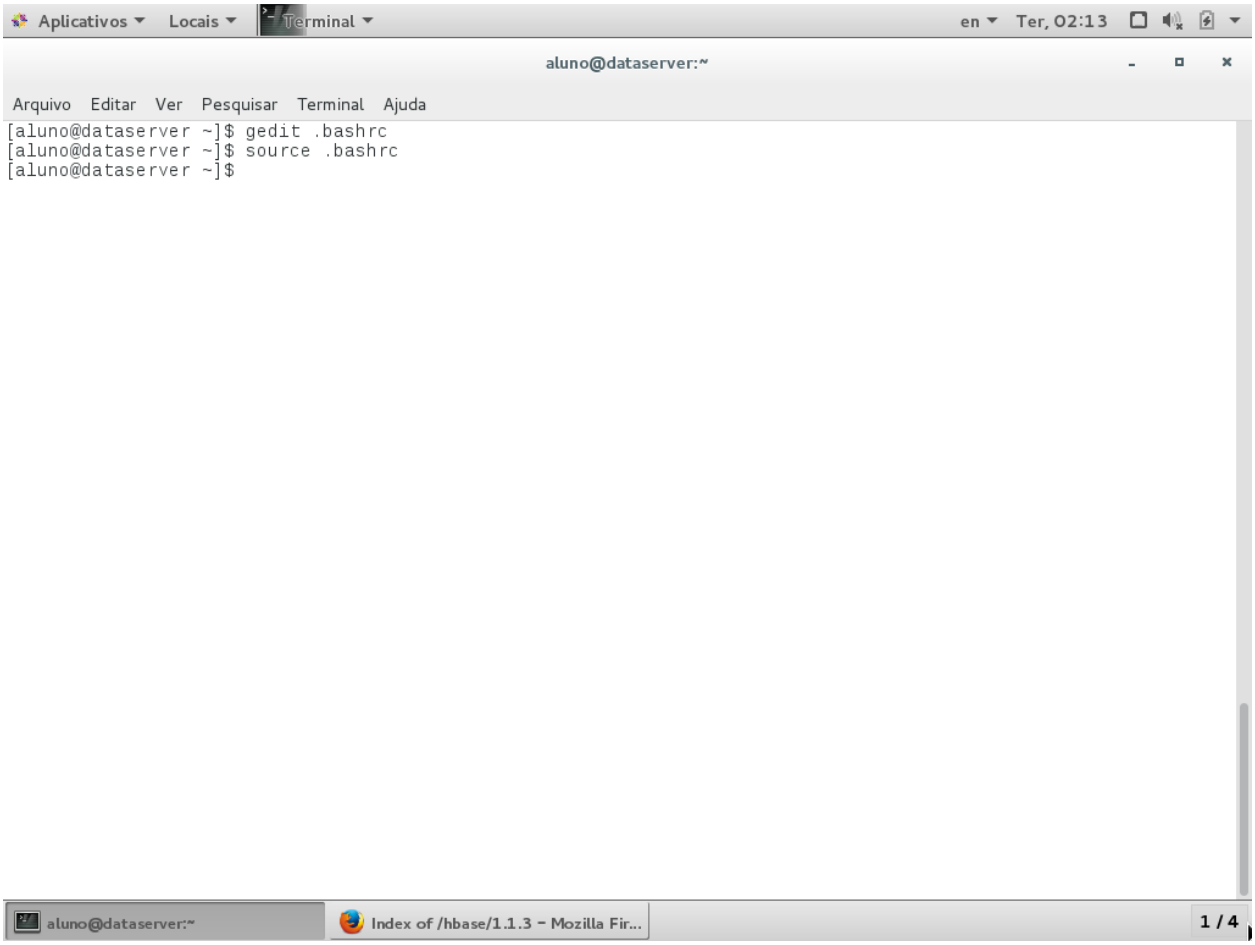
Incluir as linhas entre as tags <configuration>



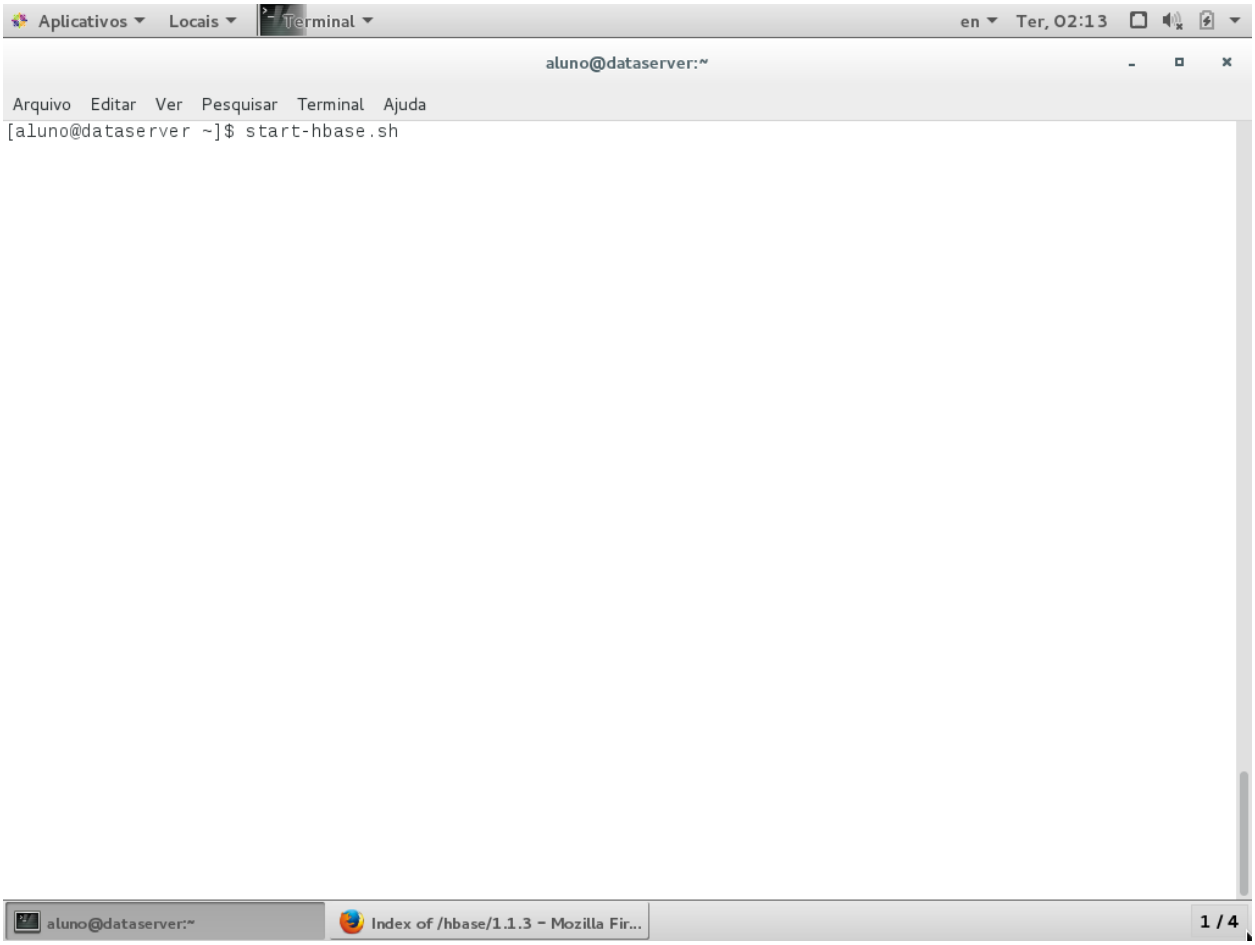
Editar o arquivo .bashrc



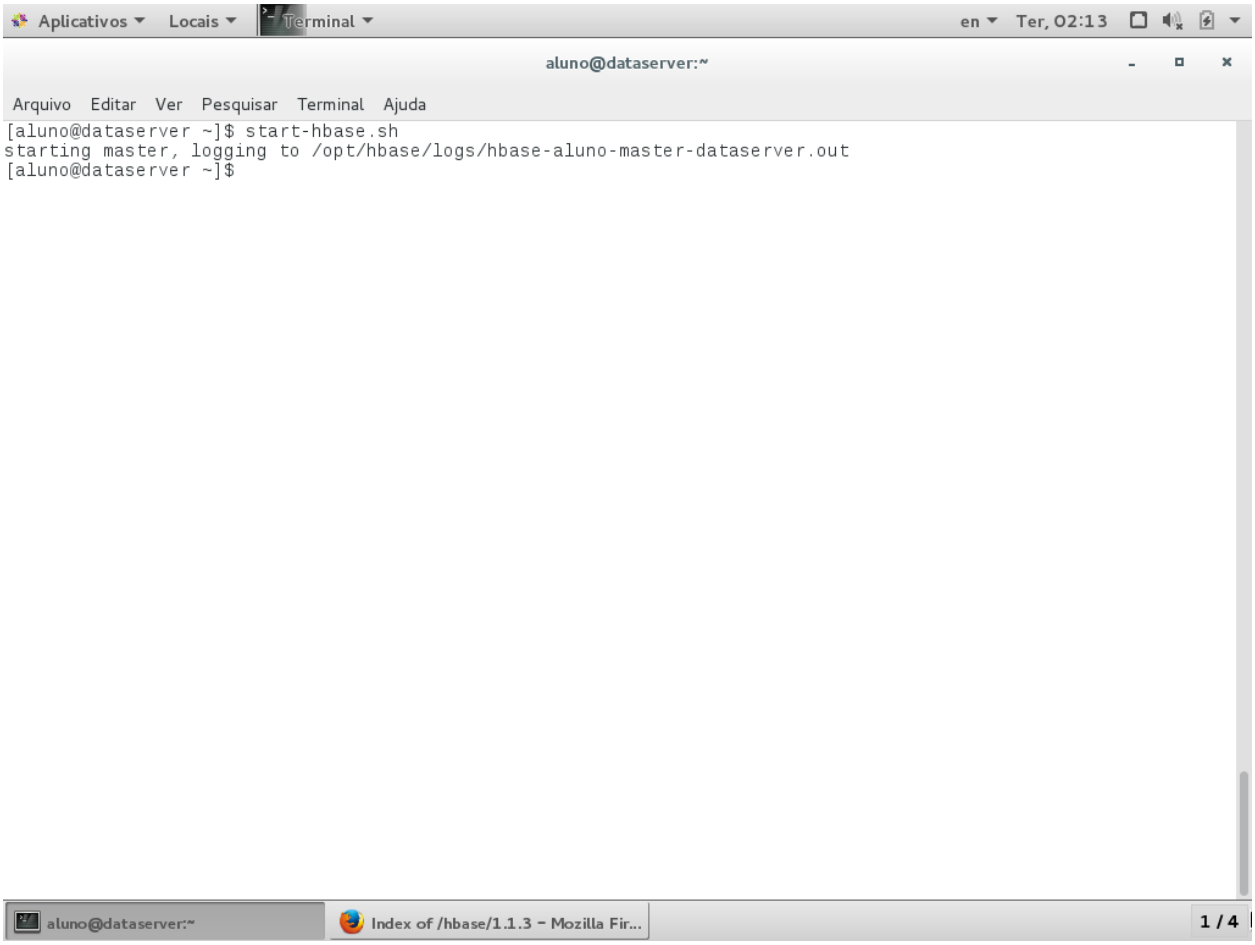
Variáveis HBase



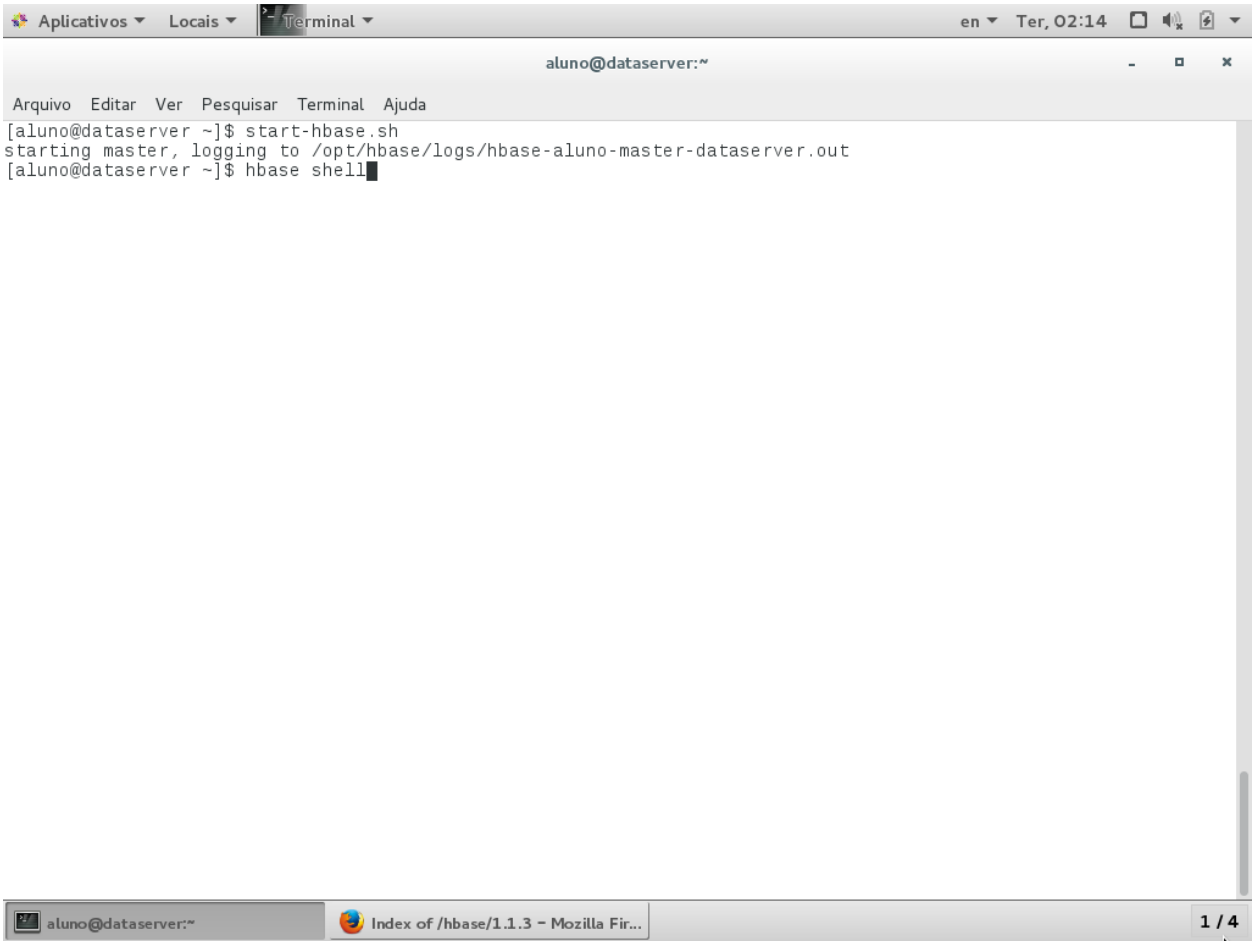
source .bashrc



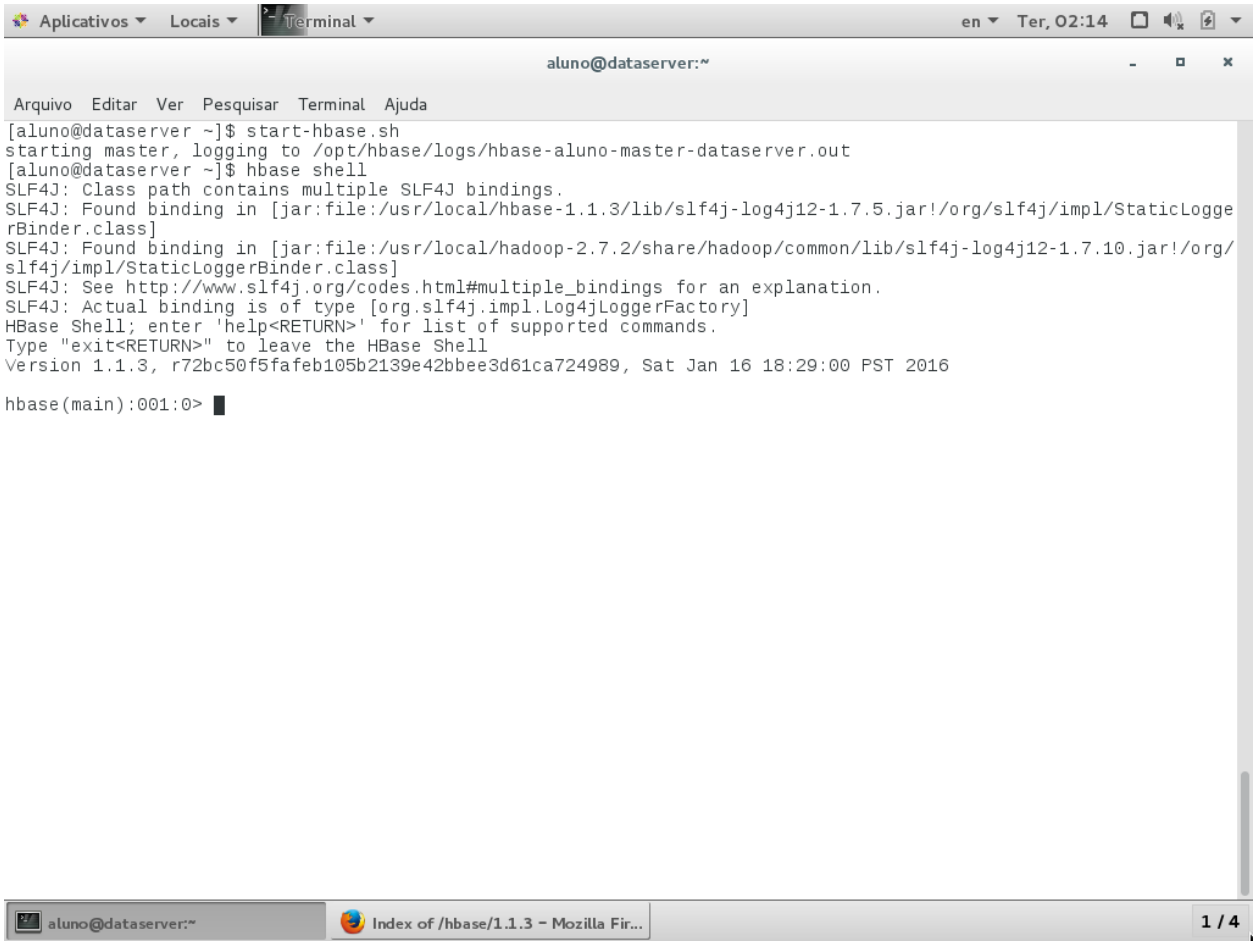
Iniciar o Hbase - start-hbase.sh



Hbase iniciado



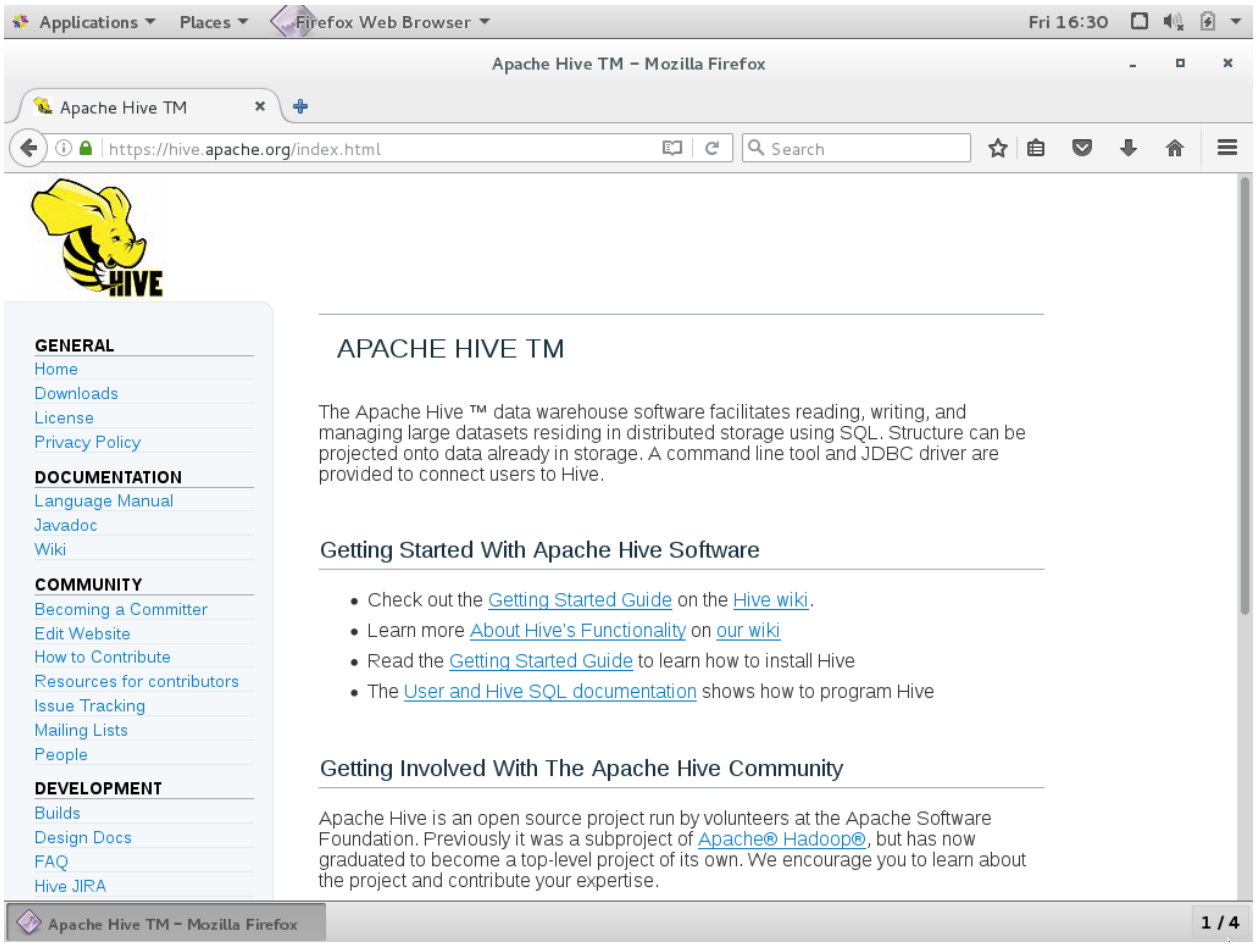
Abrir o shell do Hbase



Shell iniciado

8. Instalação e Configuração do Hive

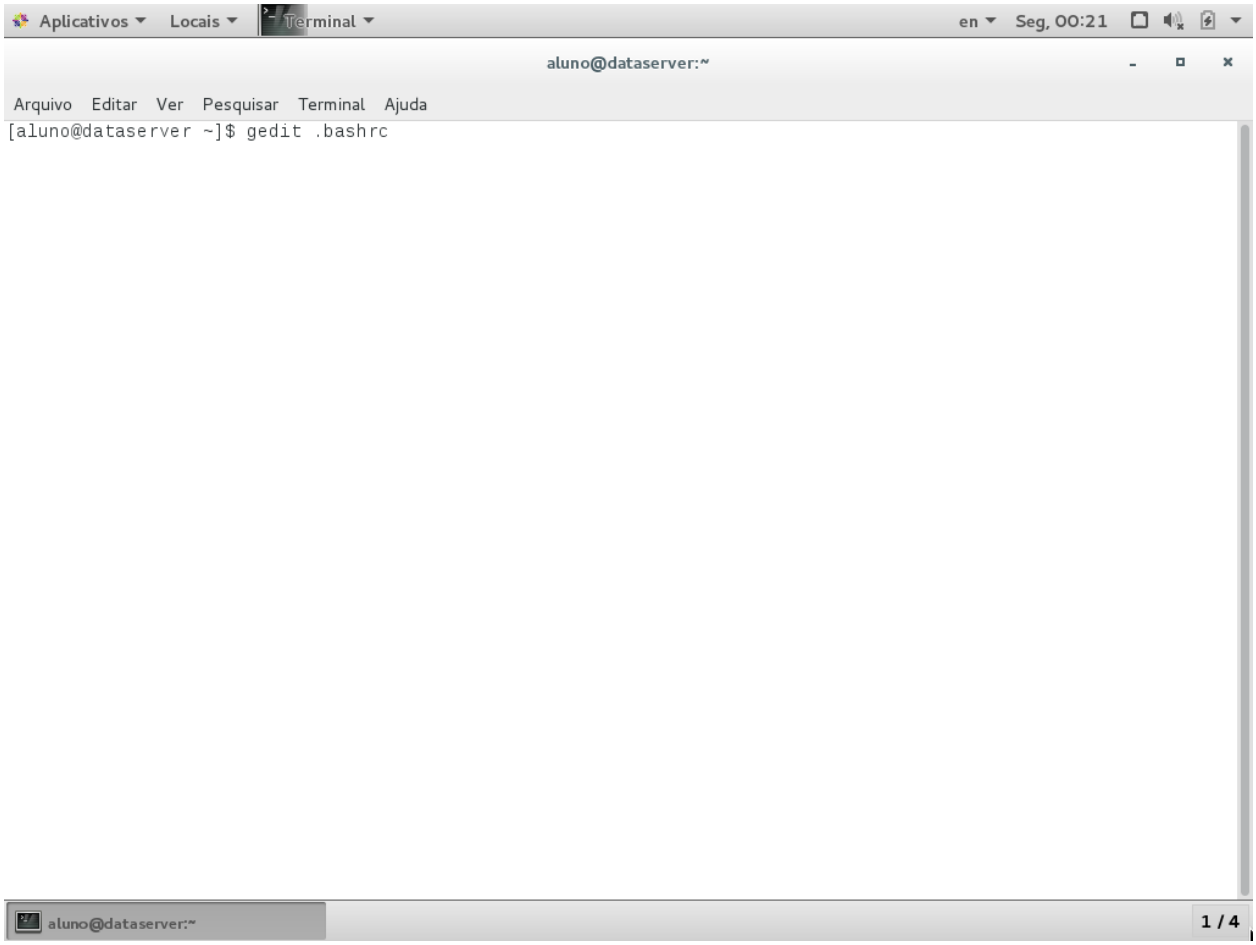
8.1. Download e Instalação do Hive



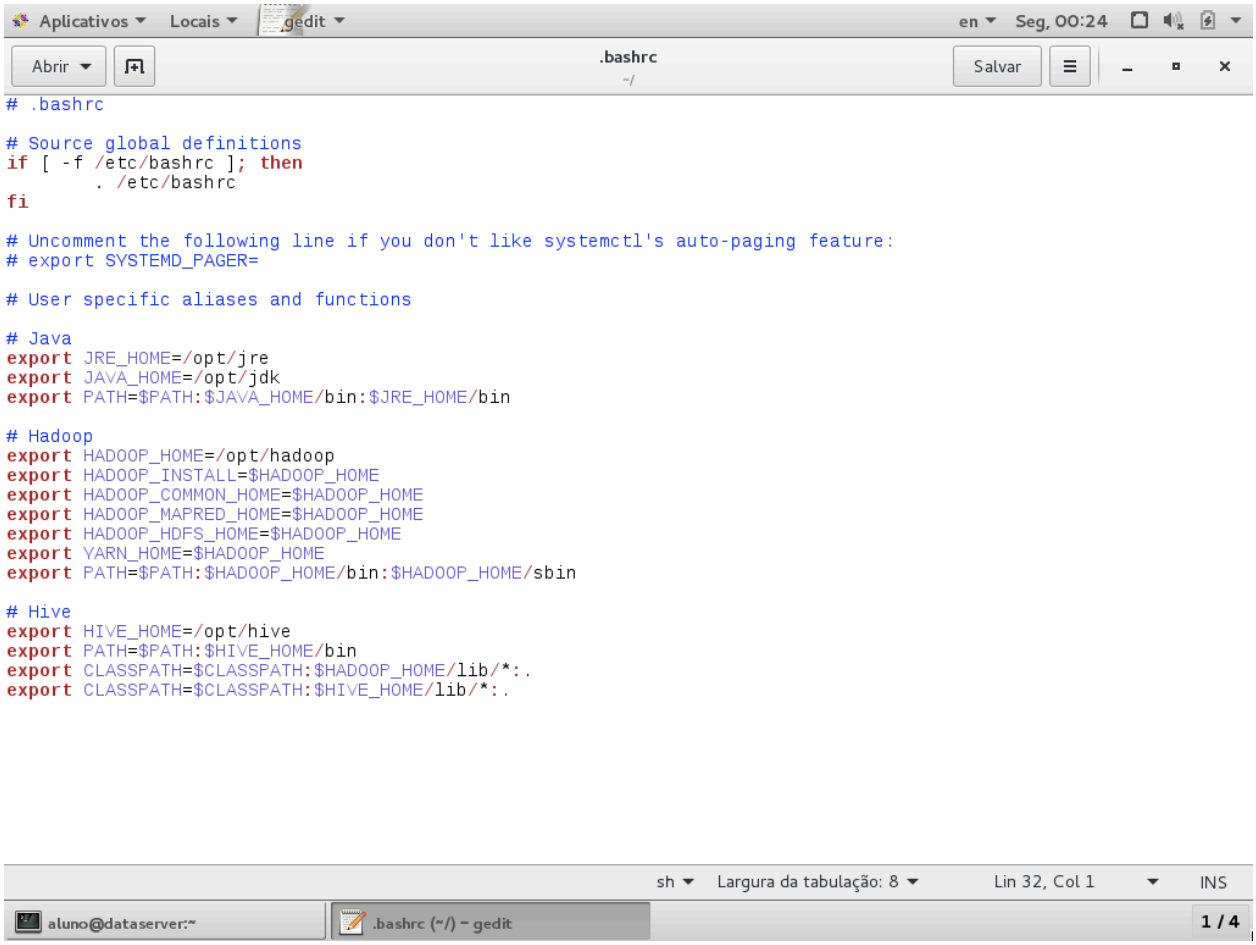
Download do Hive – Versão 3.1.1

Faça o download, descompacte o arquivo e mova o diretório para /opt/hive da mesma forma como você fez com o Java JDK e com o Hadoop.

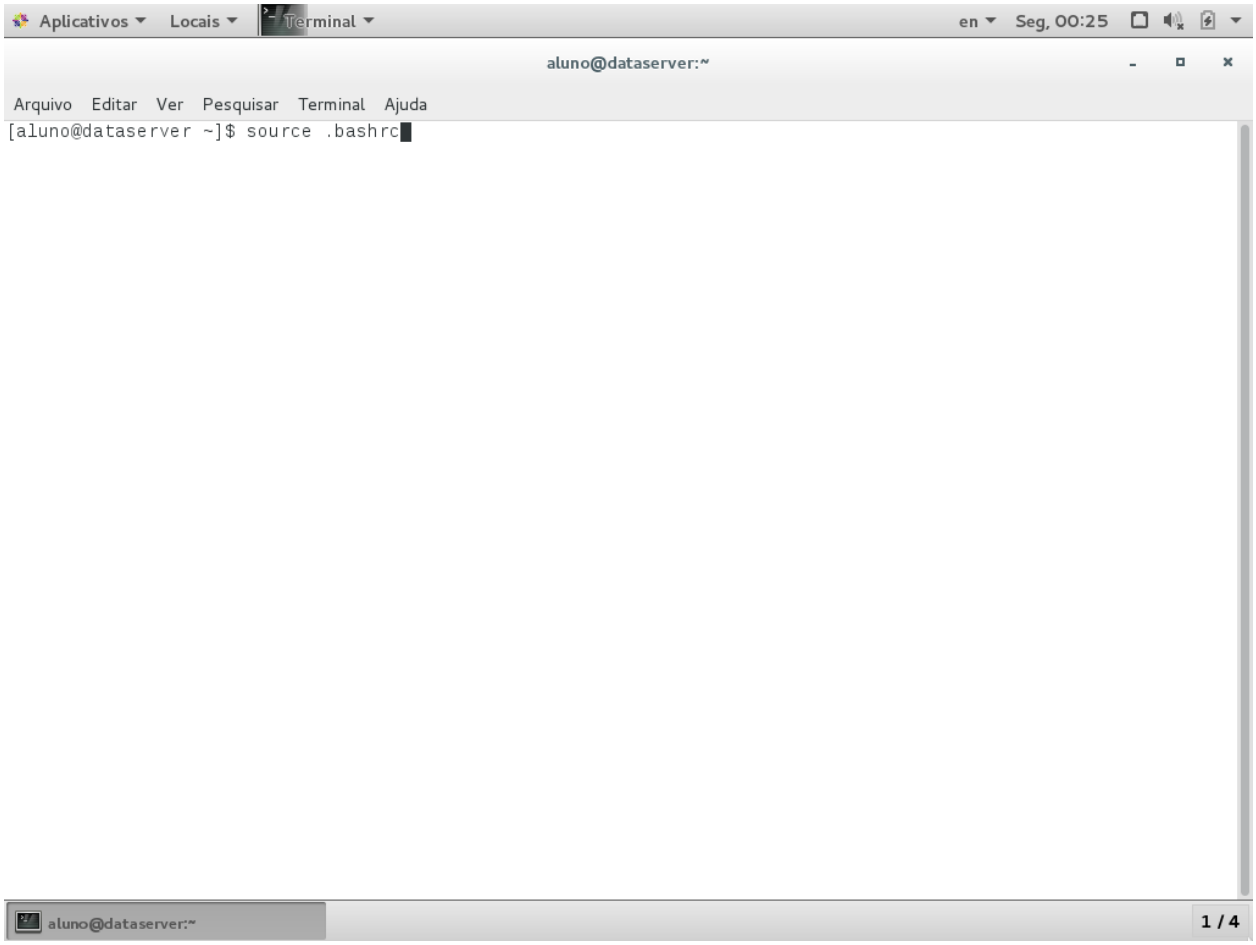
8.2. Configurando o Hive



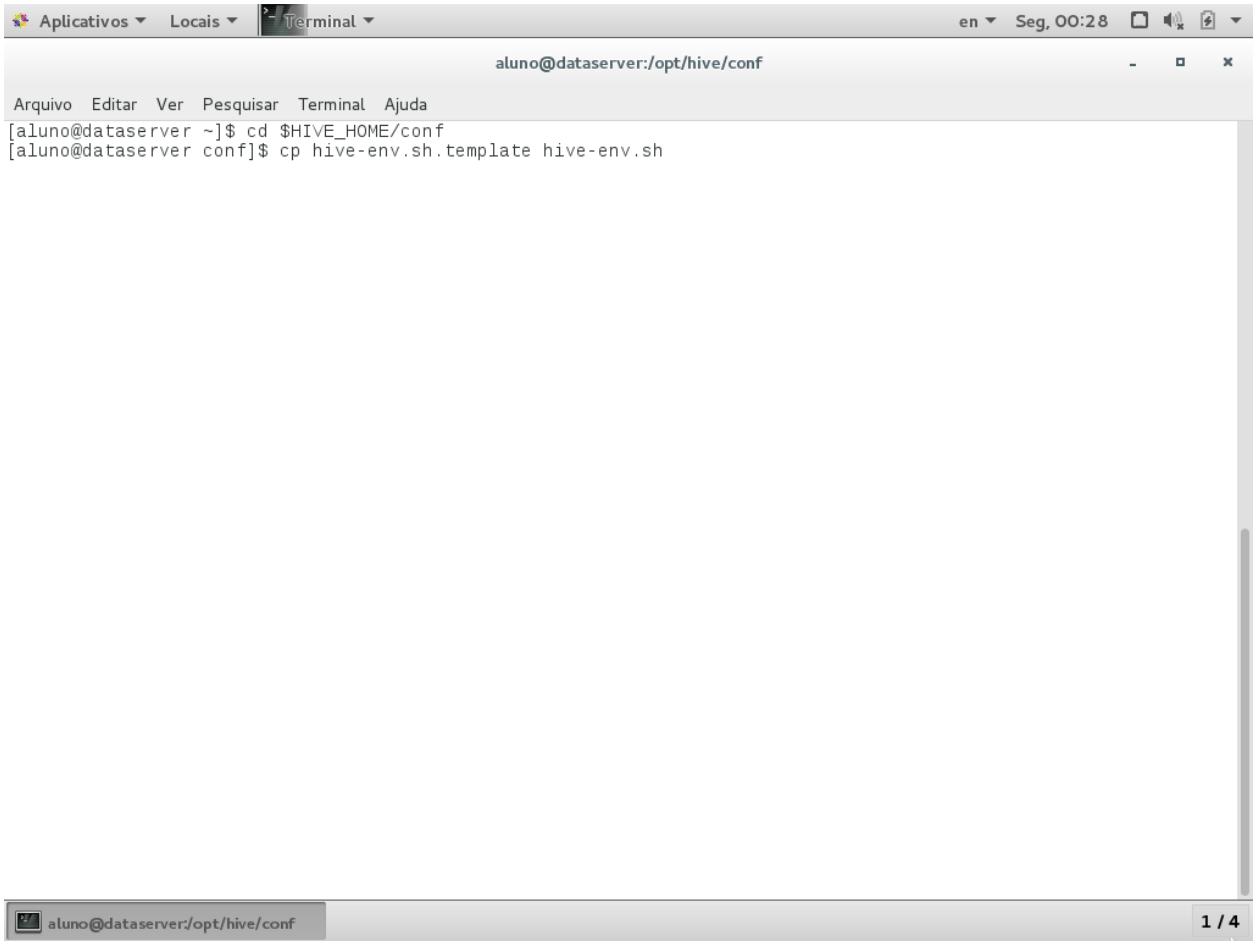
Editando o arquivo .bashrc



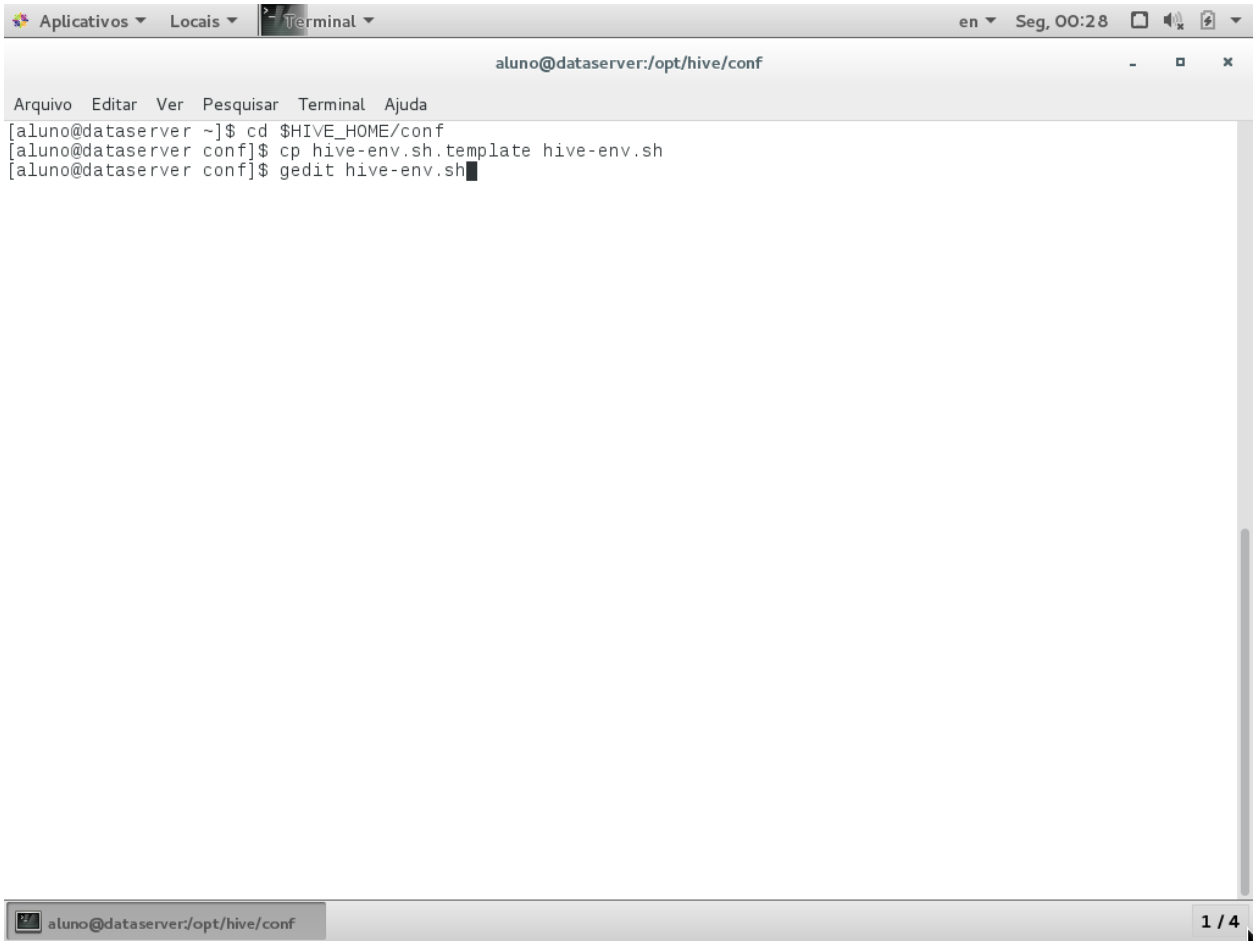
Variáveis de ambiente do Hive



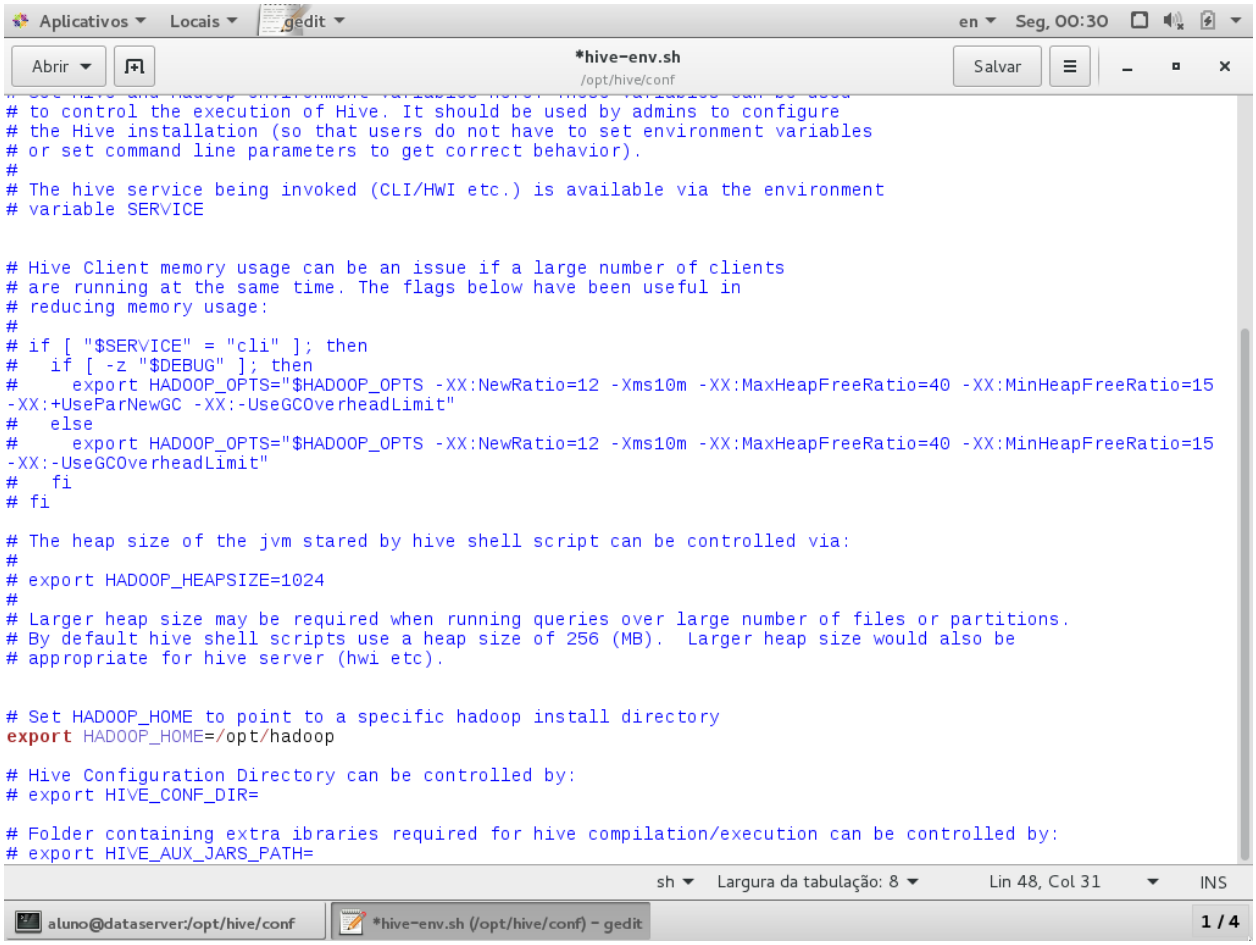
source .bashrc



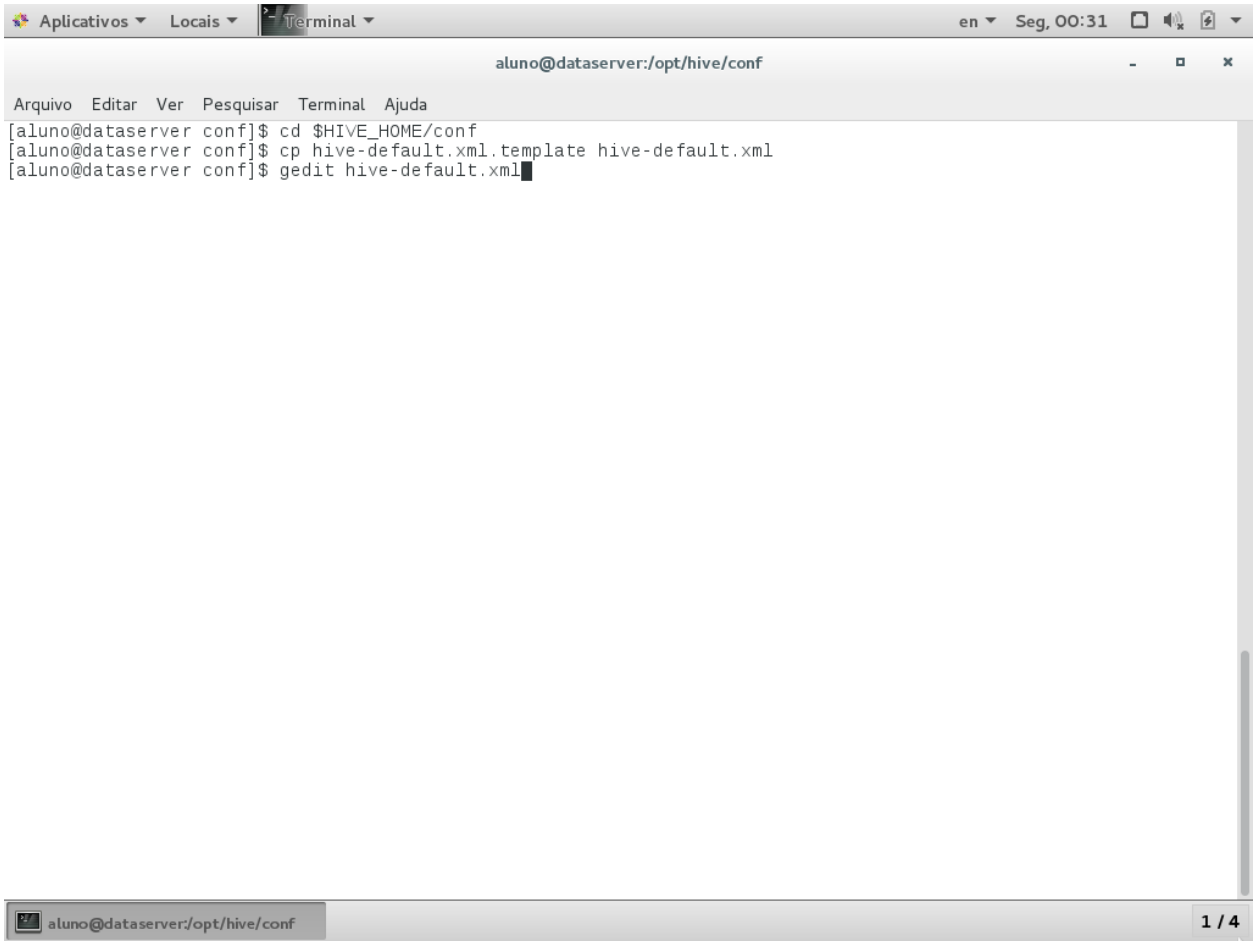
A partir do arquivo template, gerar o arquivo hive-env.sh



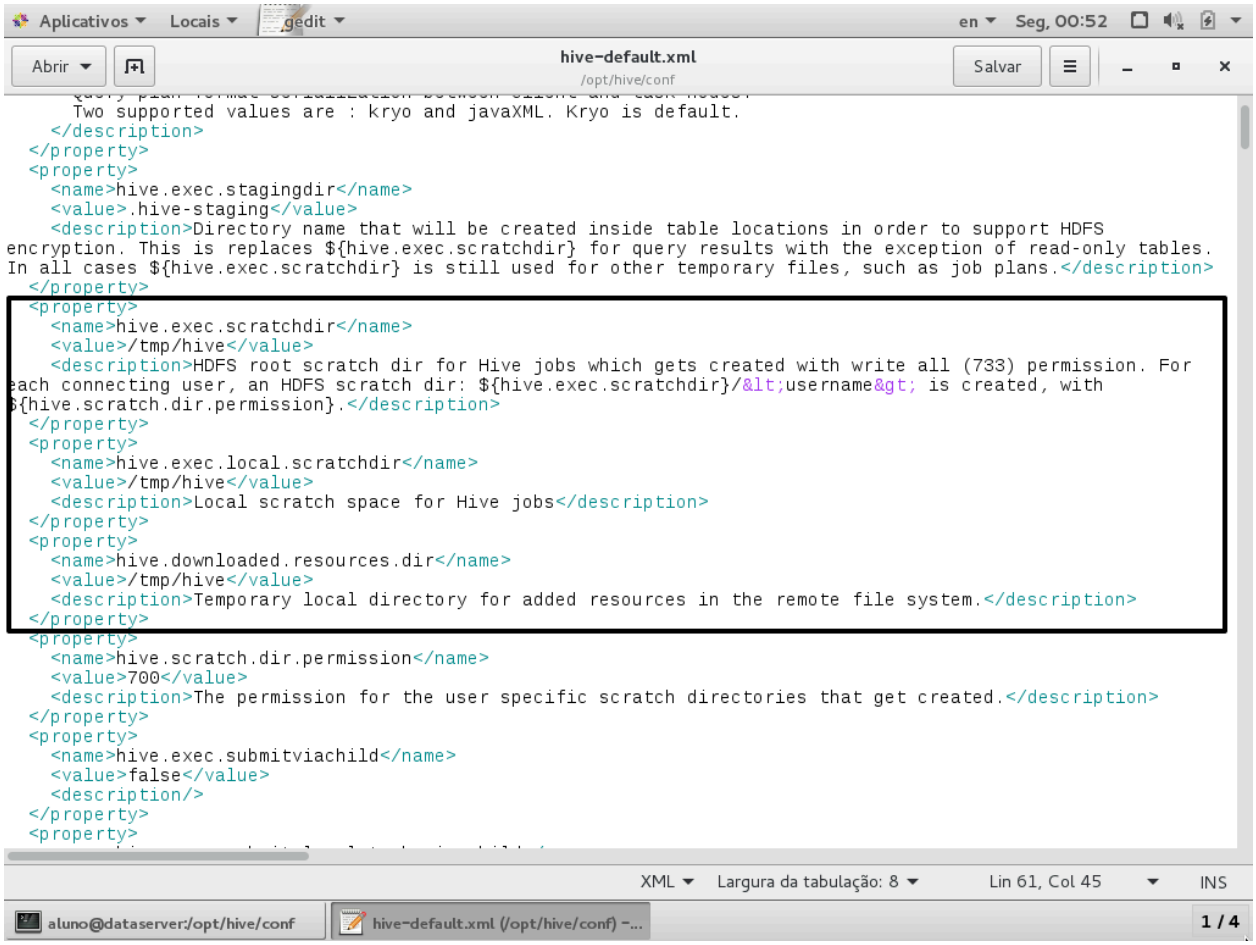
Editar o arquivo



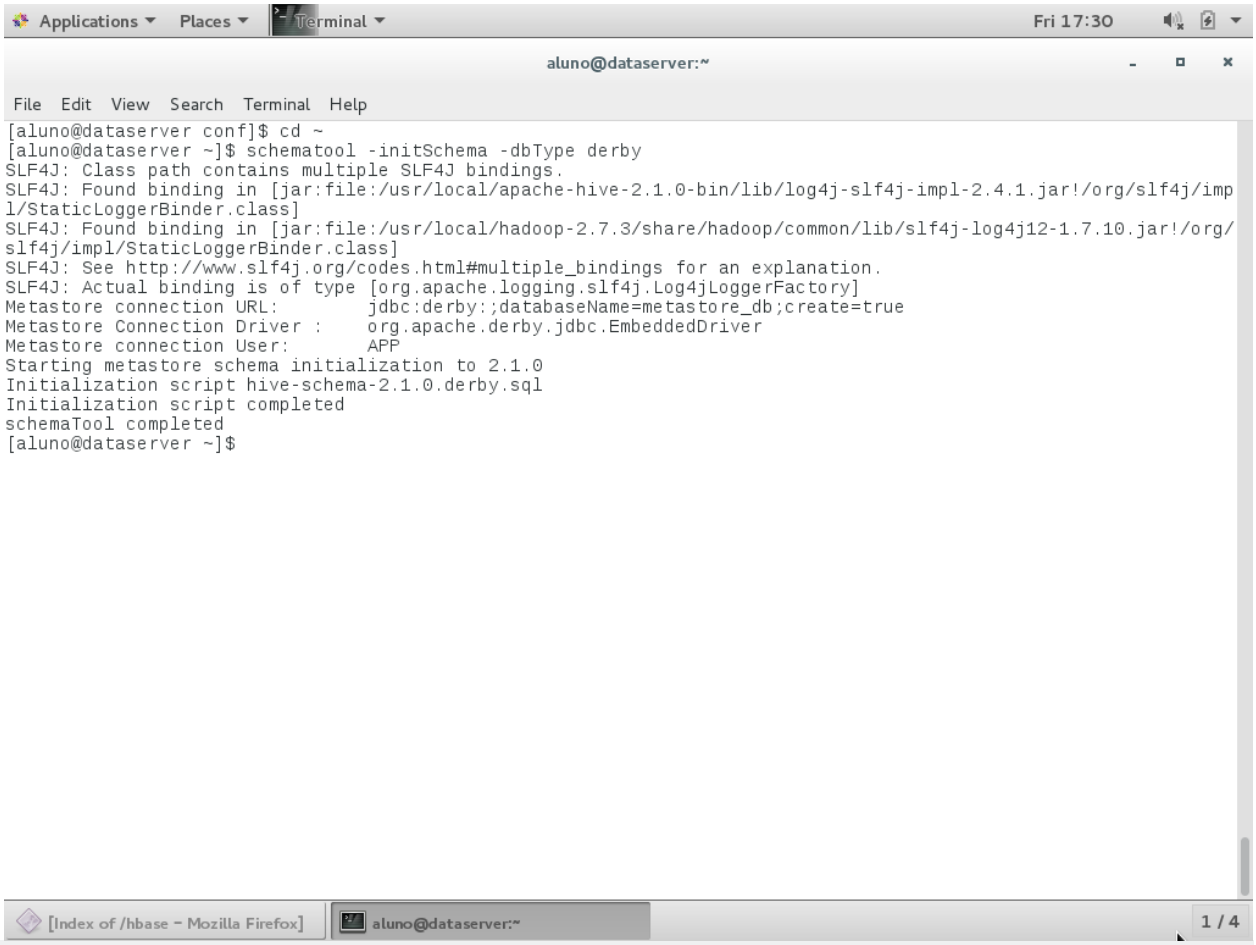
Incluir PATH do Hadoop, conforme tela acima



A partir do template, gerar o arquivo hive-site.xml



Editar as linhas conforme cima



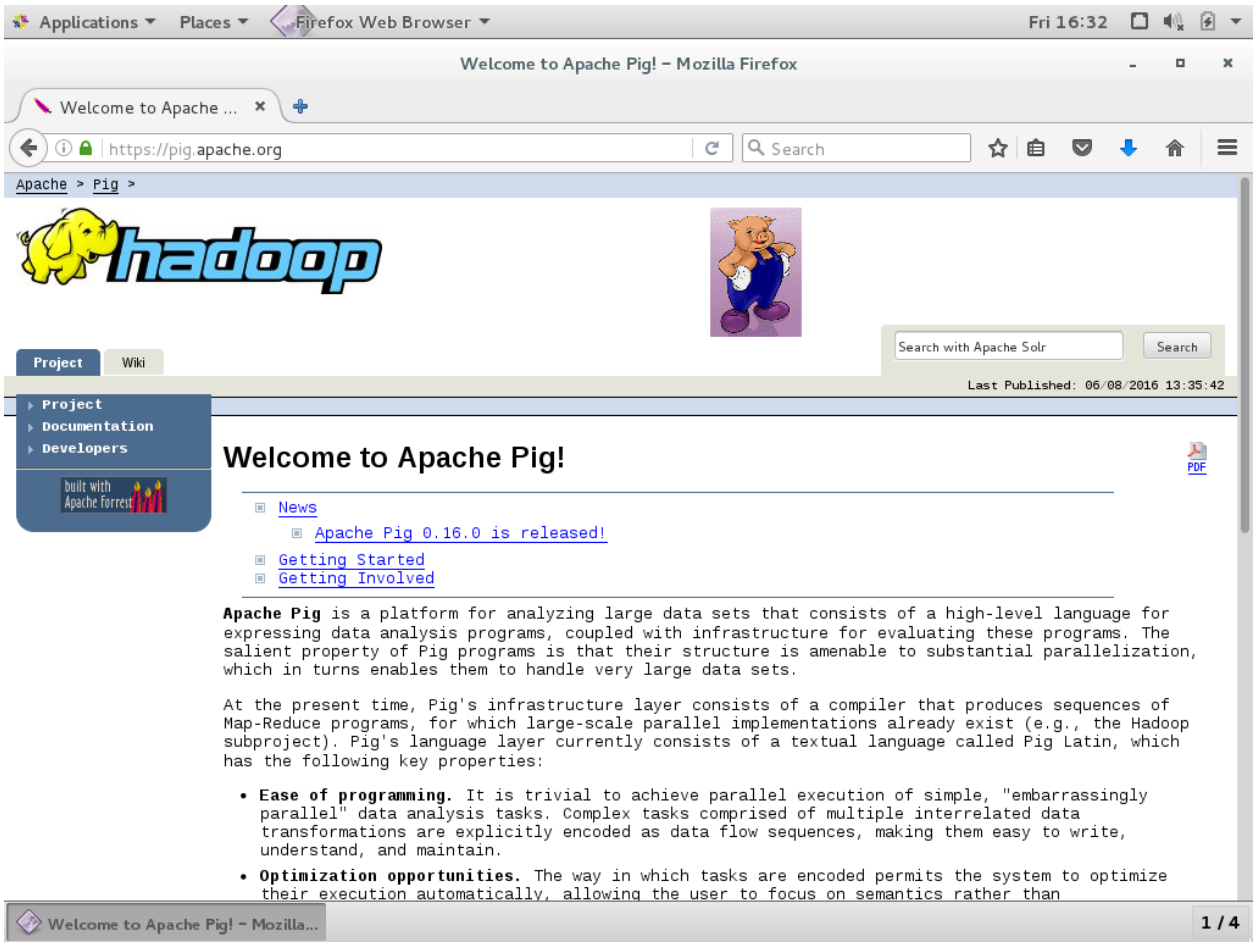
Inicializar o schema do Hive
schematool -initSchema -dbType derby

Executando o Hive (execute o comando jps para se certificar que o Hadoop está ativo)

O comando “show tables;” demonstra que o Hive foi instalado com sucesso

9. Instalação e Configuração do Pig

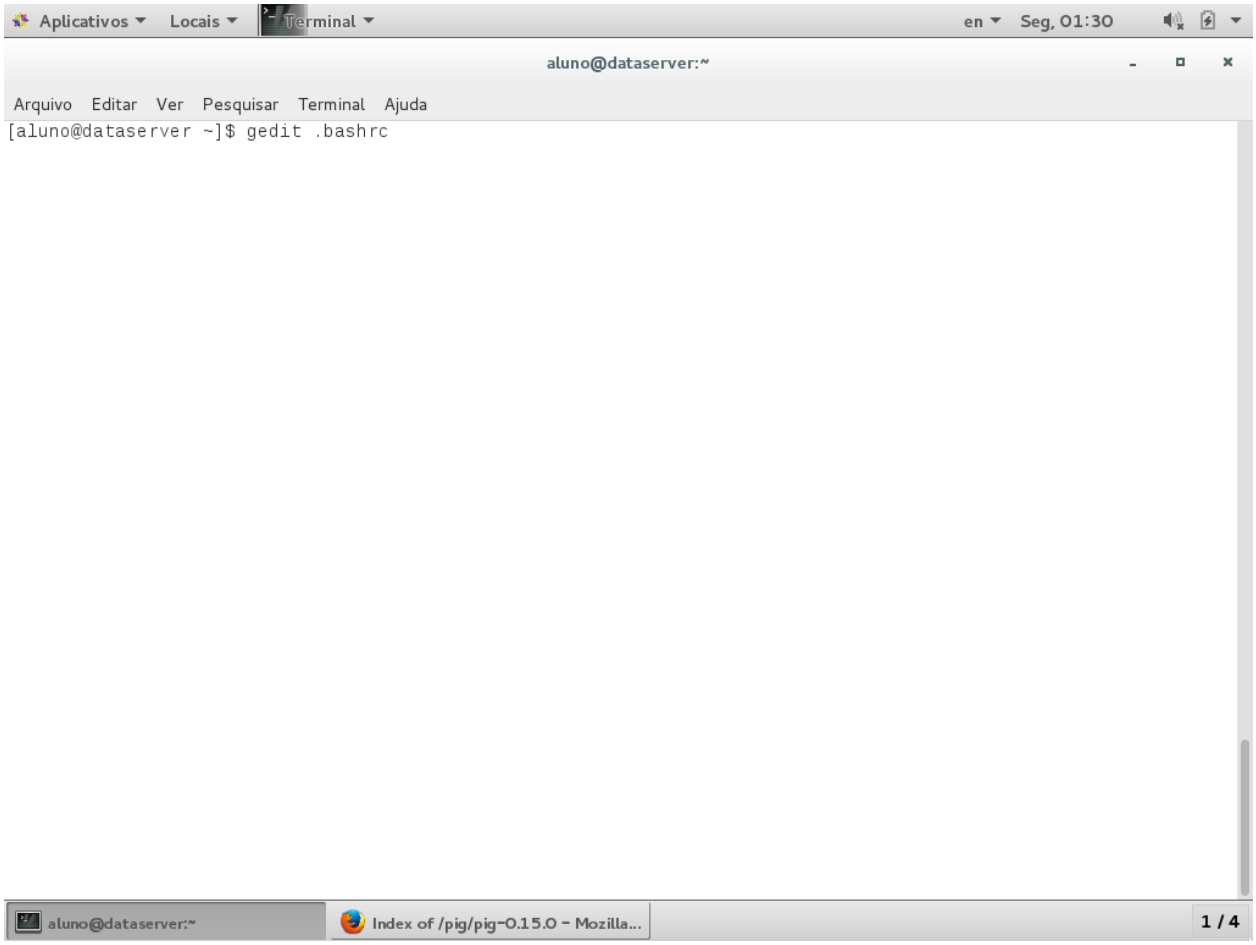
9.1. Download e Instalação do Pig



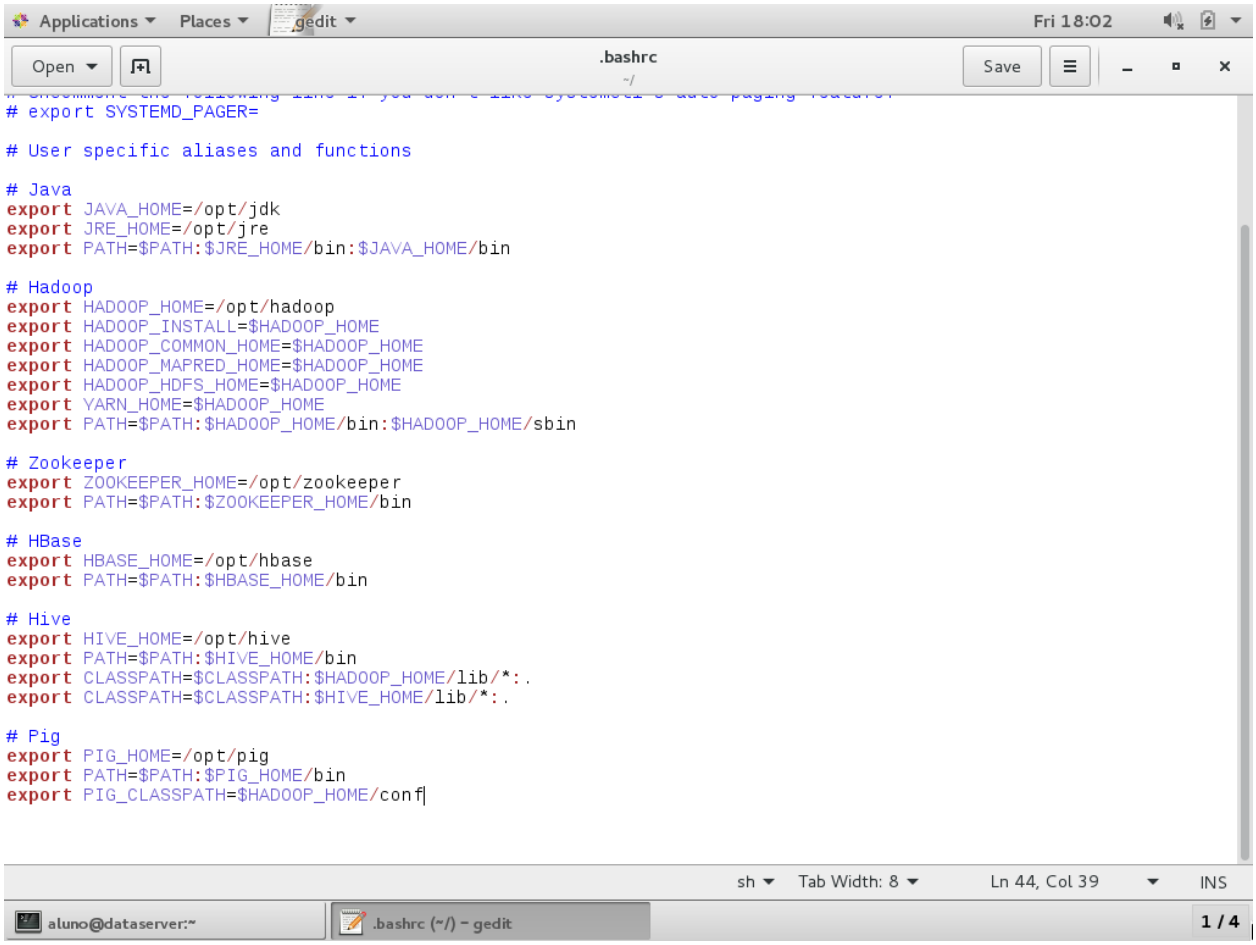
Download do Pig – Versão 0.17.0

Faça o download, descompacte o arquivo e mova o diretório para /opt/pig da mesma forma como você fez com o Java JDK e com o Hadoop.

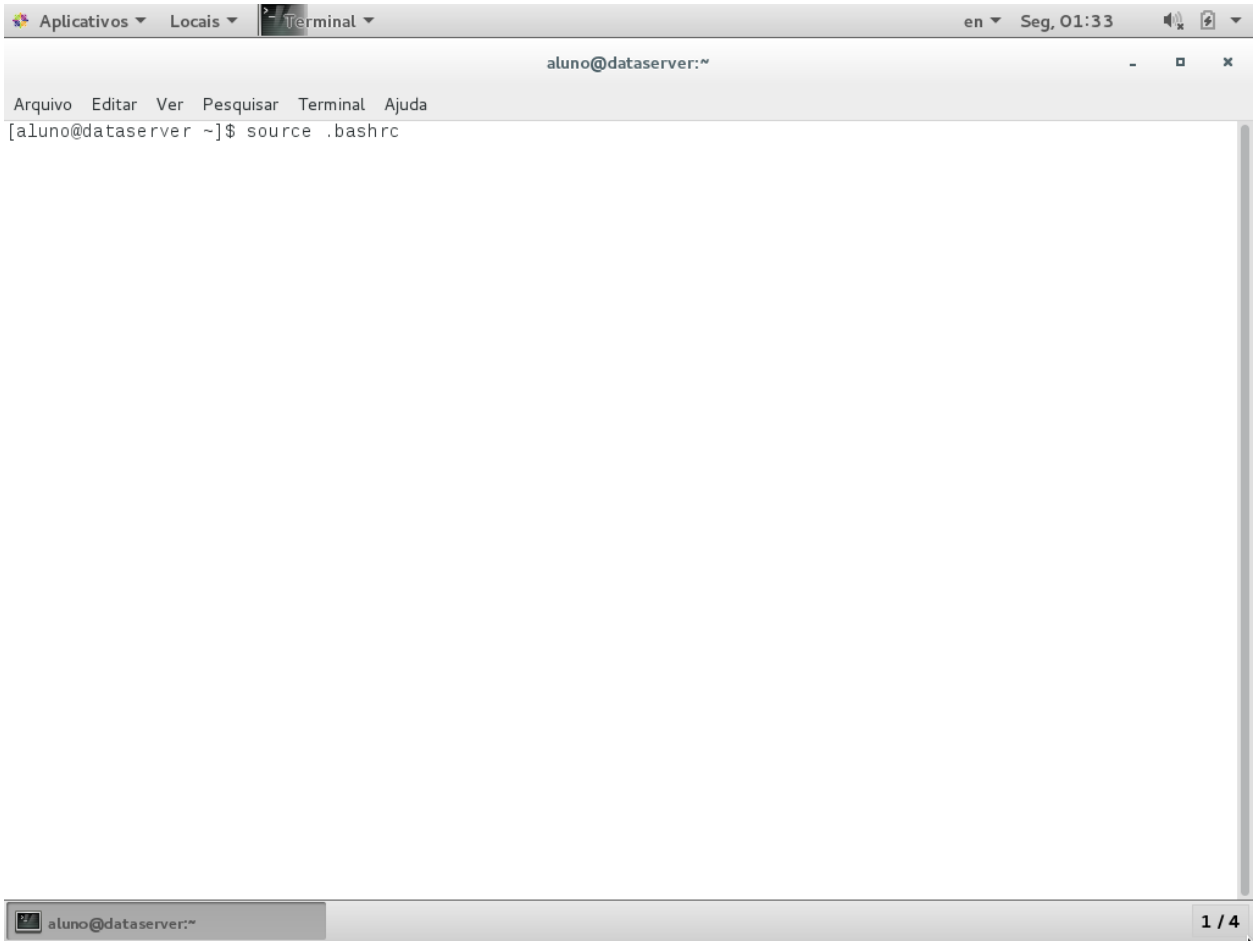
9.2. Configurando do Pig



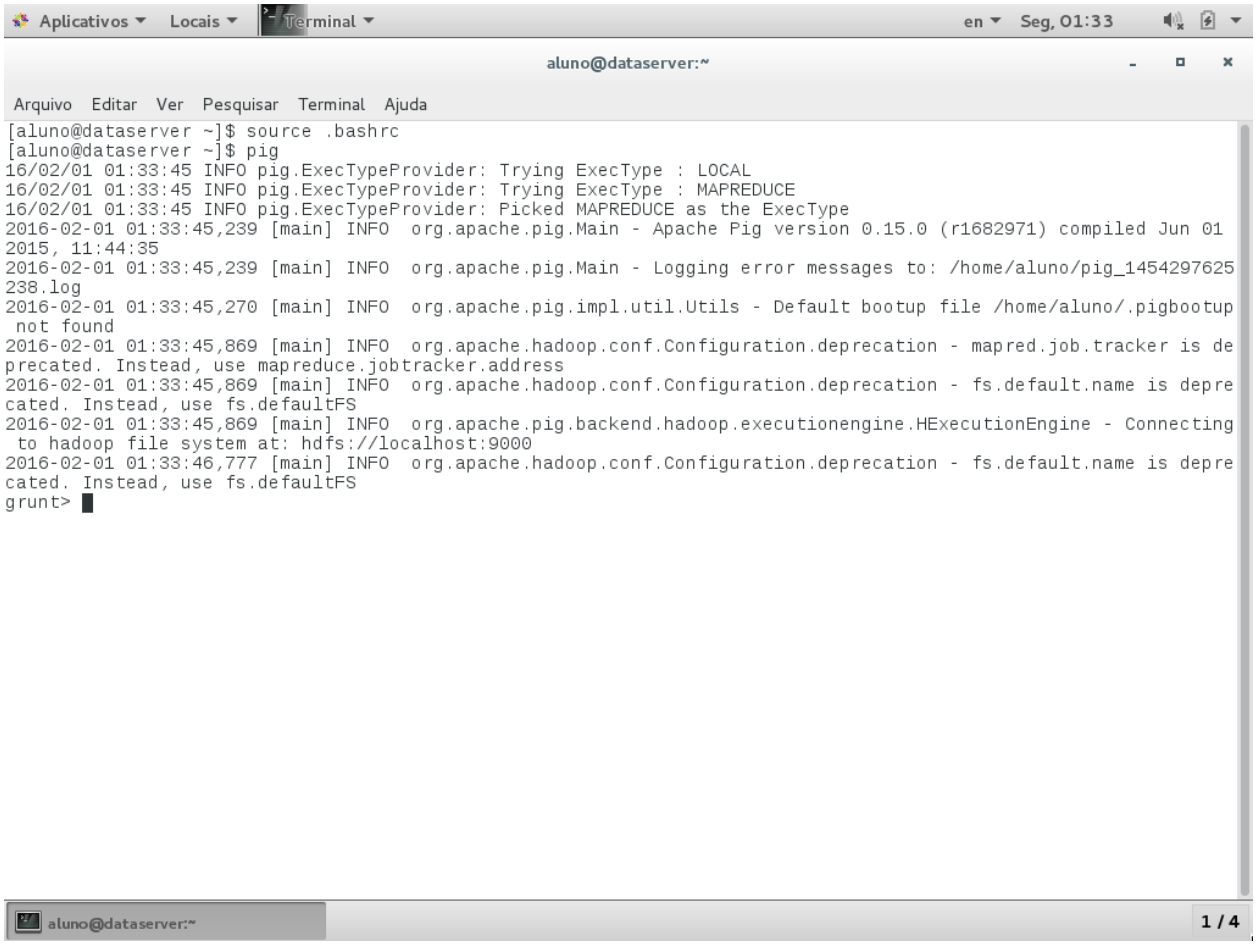
Editando o arquivo .bashrc



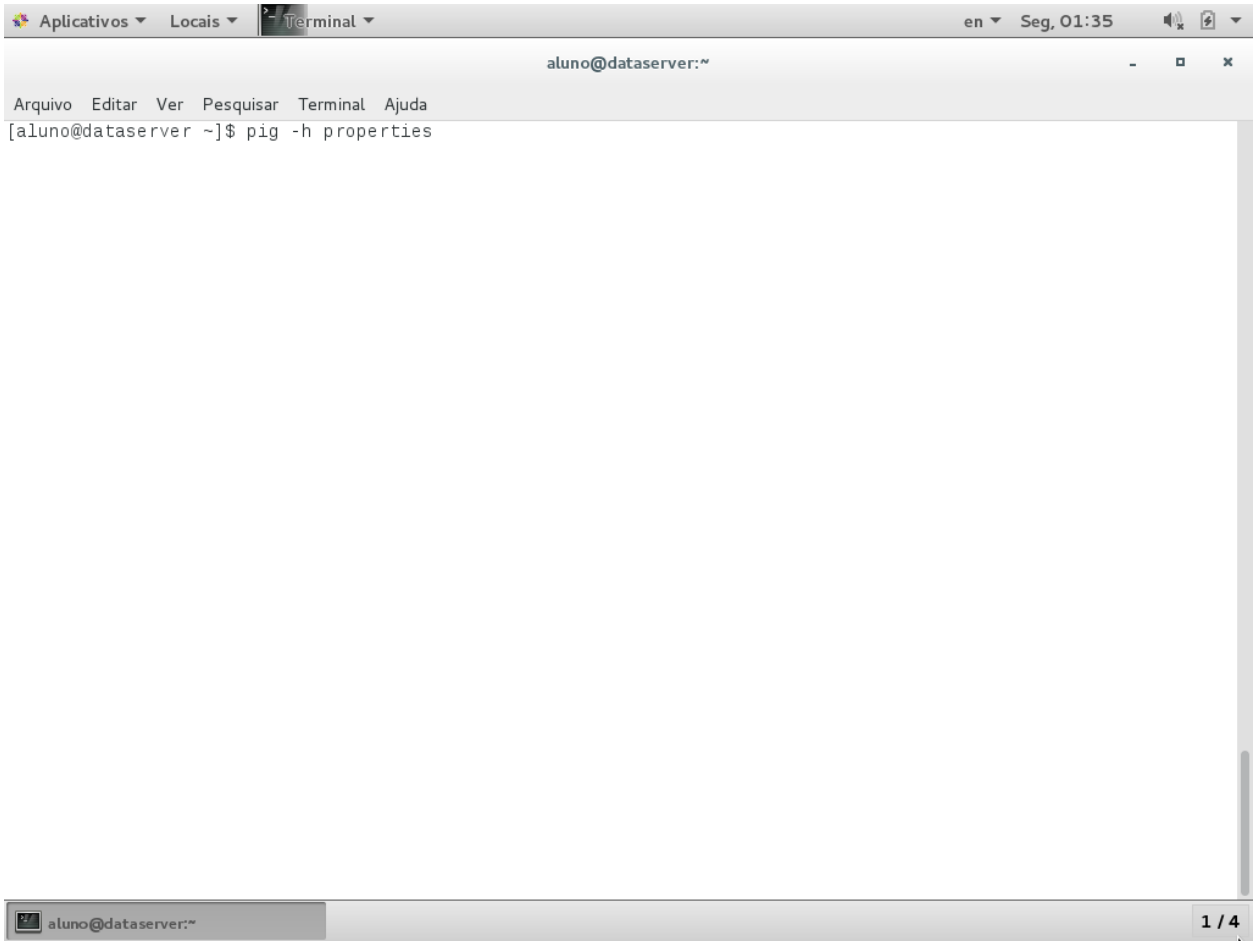
Inserir variáveis de ambiente do Pig



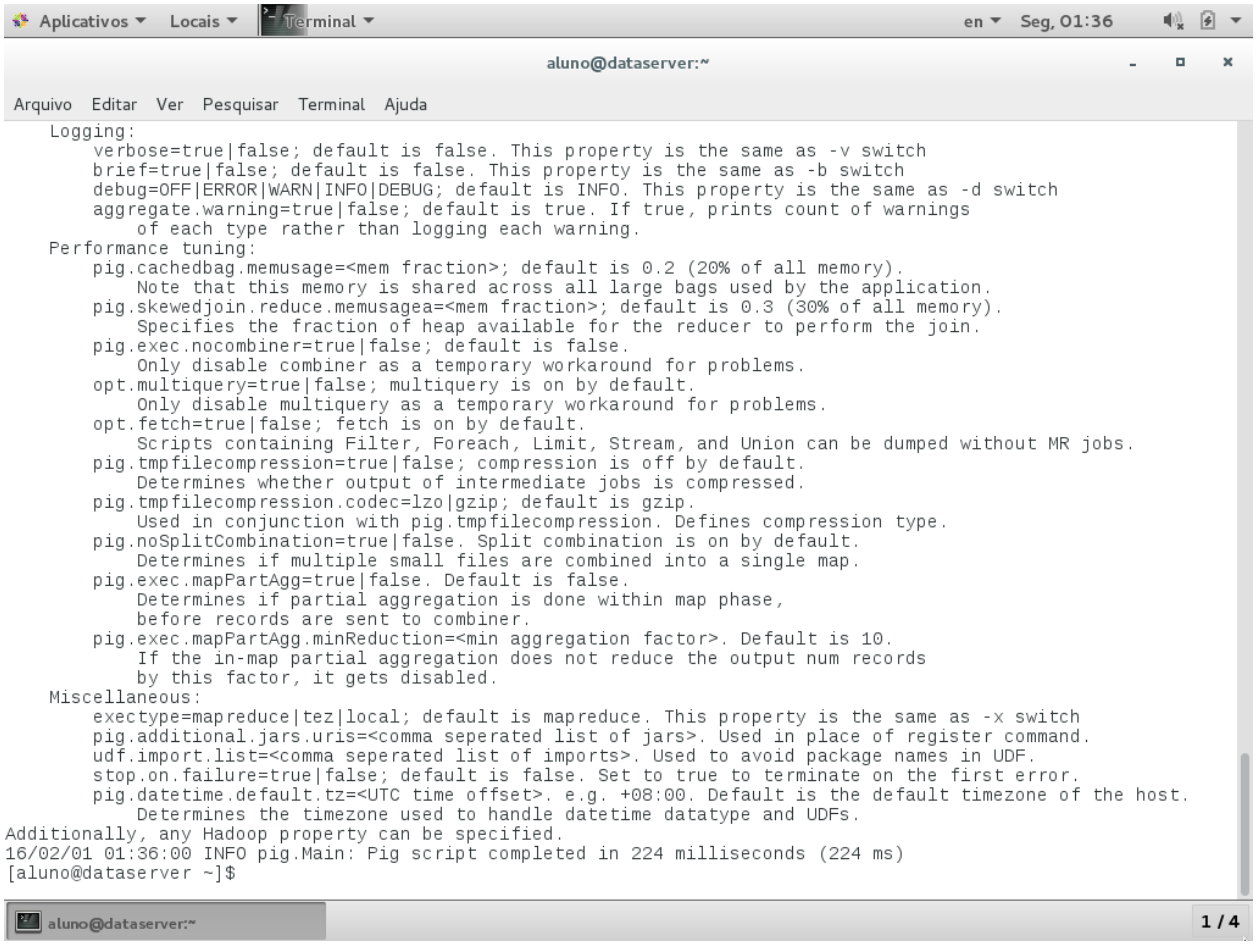
source .bashrc



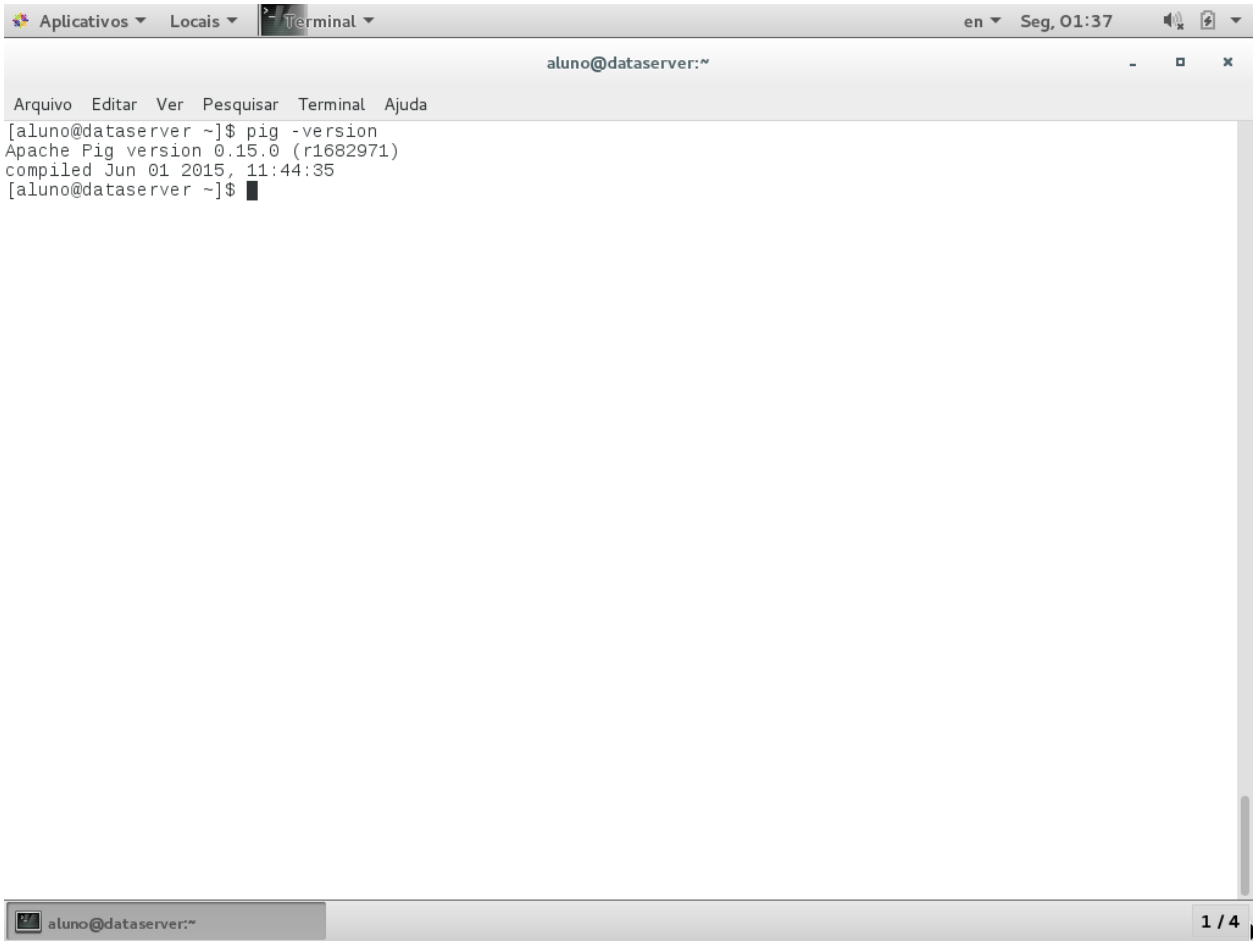
Pig instalado com sucesso



O comando **pig –h properties** lista as variáveis configuradas



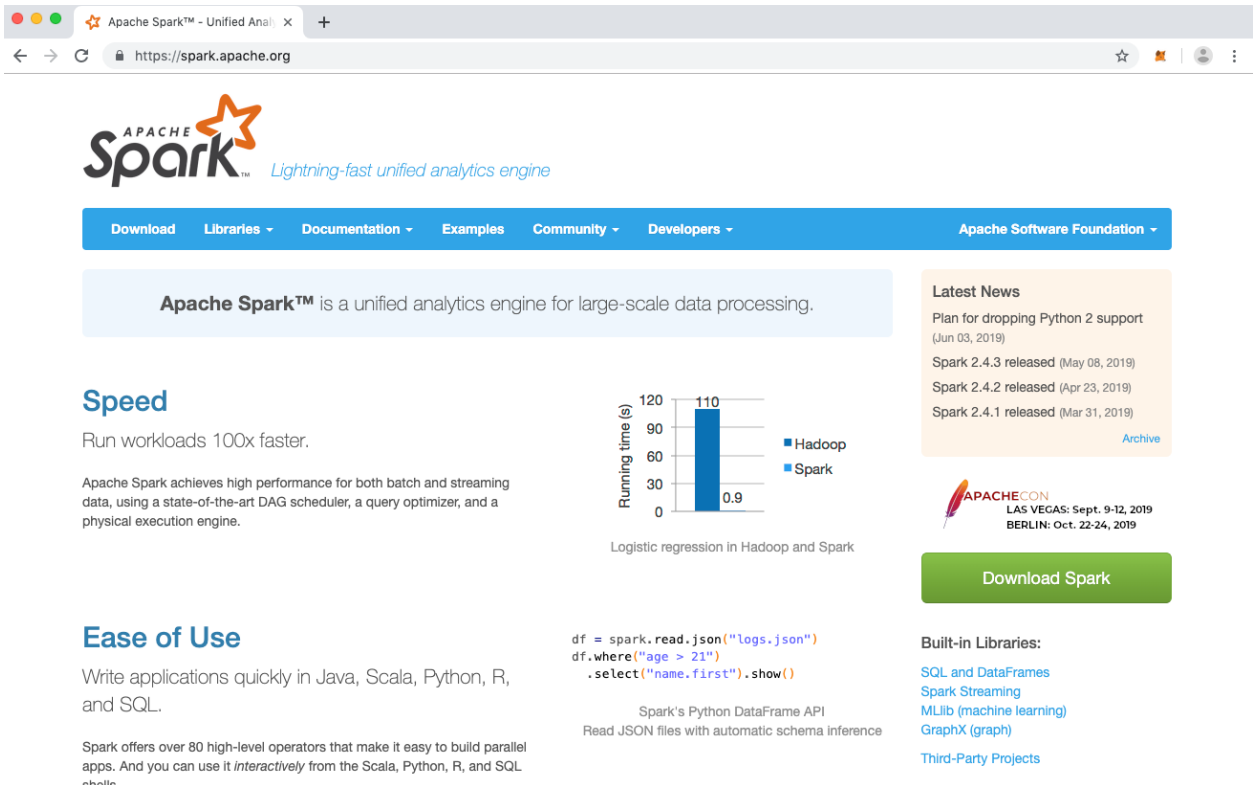
Variáveis Pig



Verificar a versão do Pig

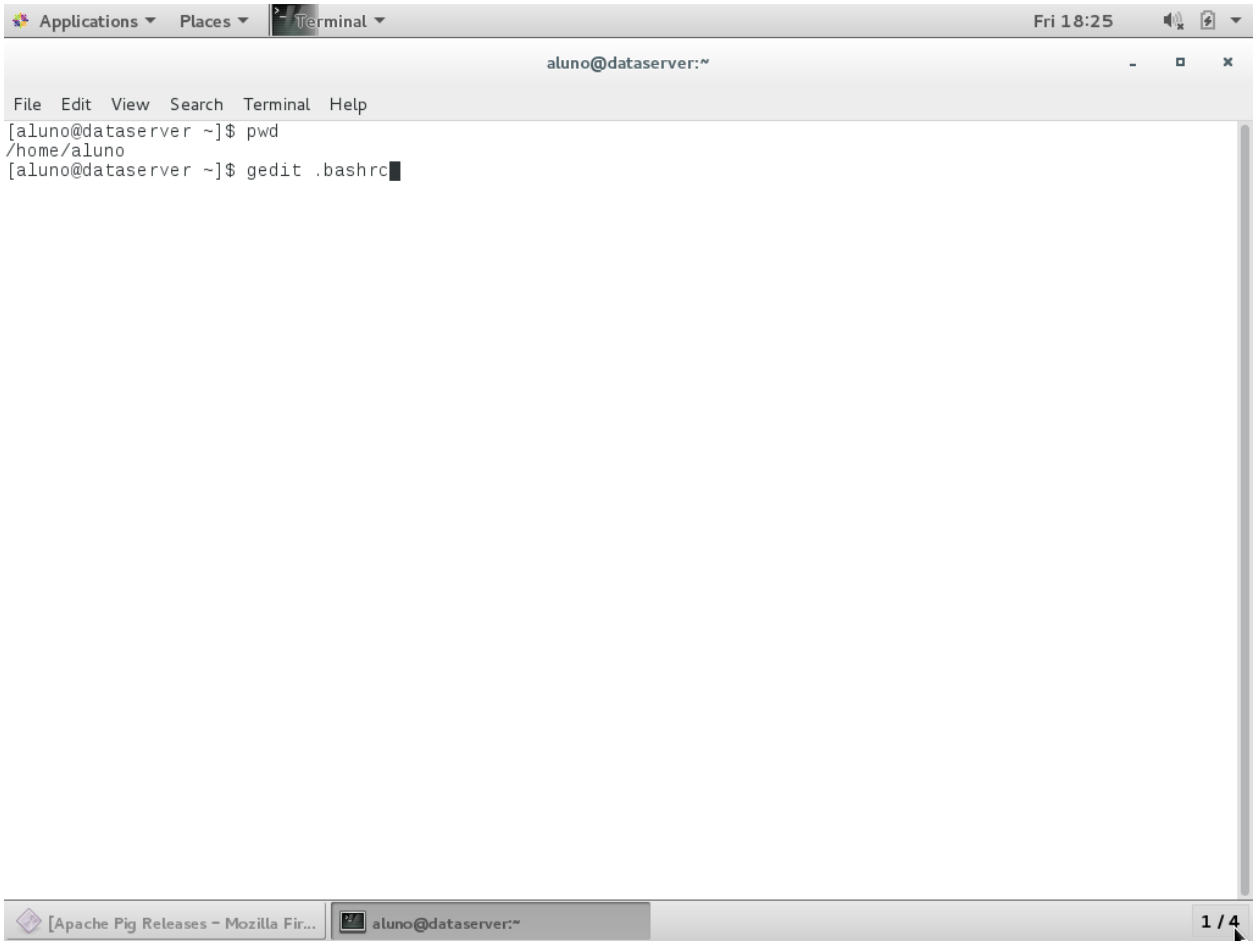
10. Instalação e Configuração do Spark

10.1. Download e Instalação do Spark

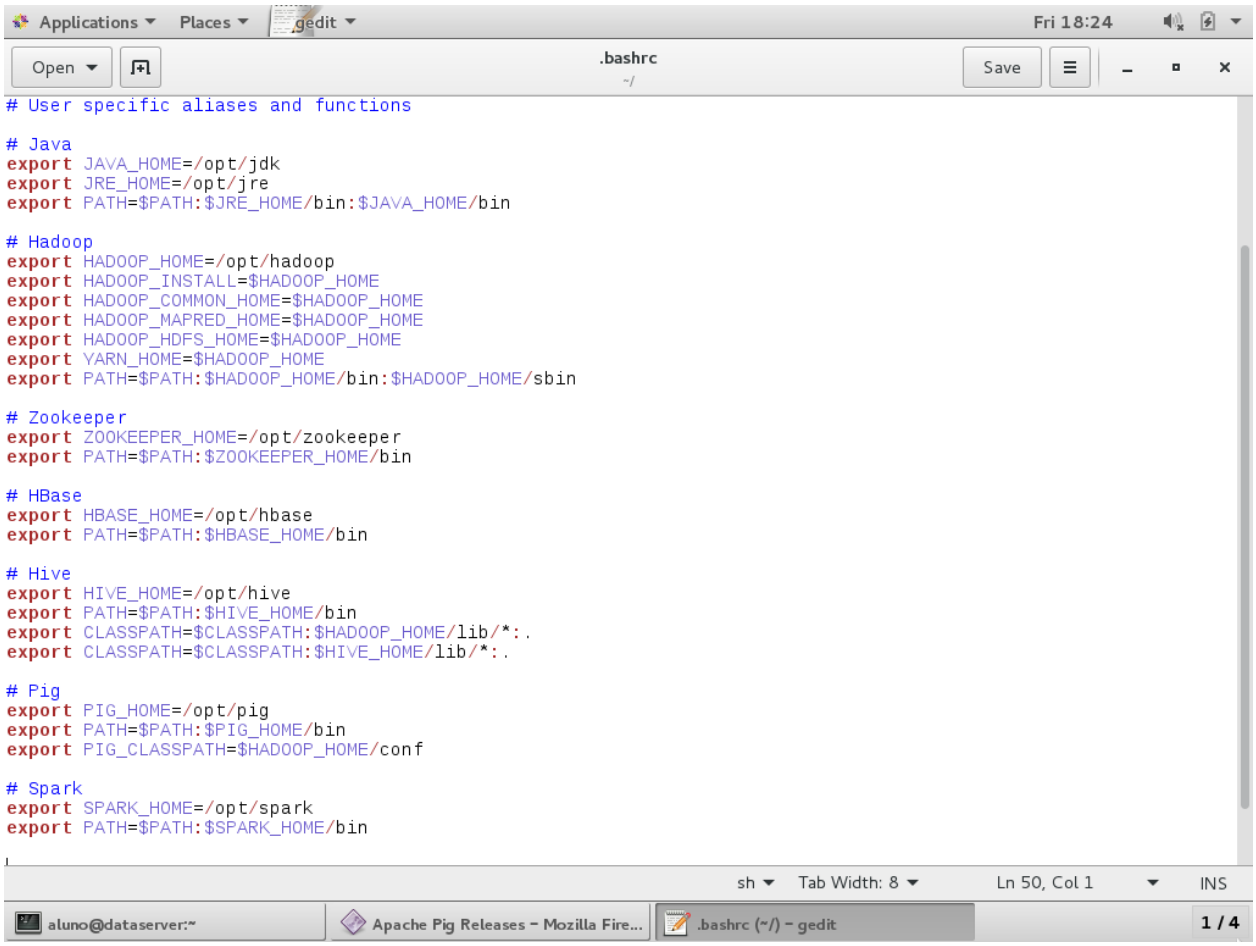


Download do Spark – Versão 2.4.3

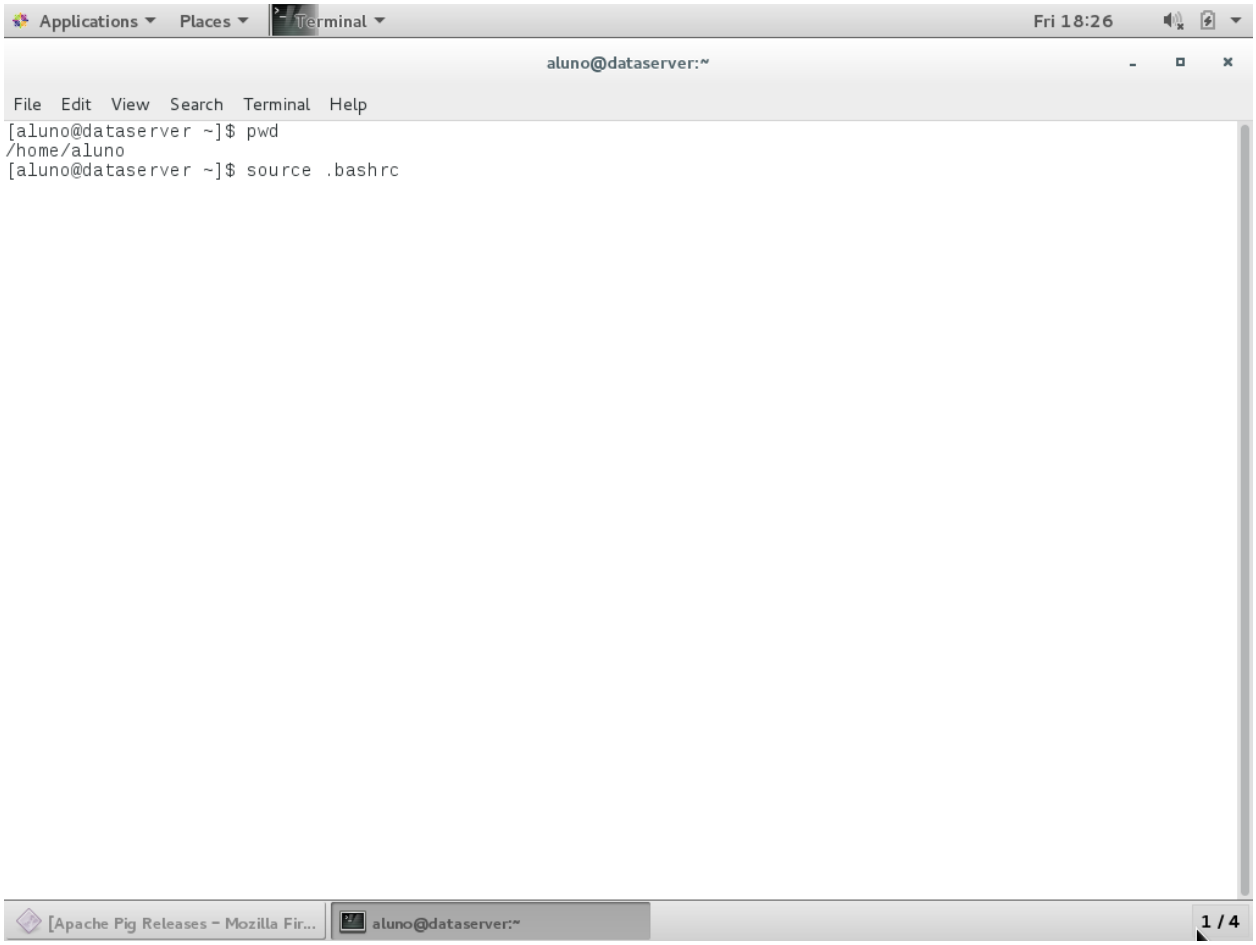
Faça o download, descompacte o arquivo e mova o diretório para /opt/spark da mesma forma como você fez com o Java JDK e com o Hadoop.



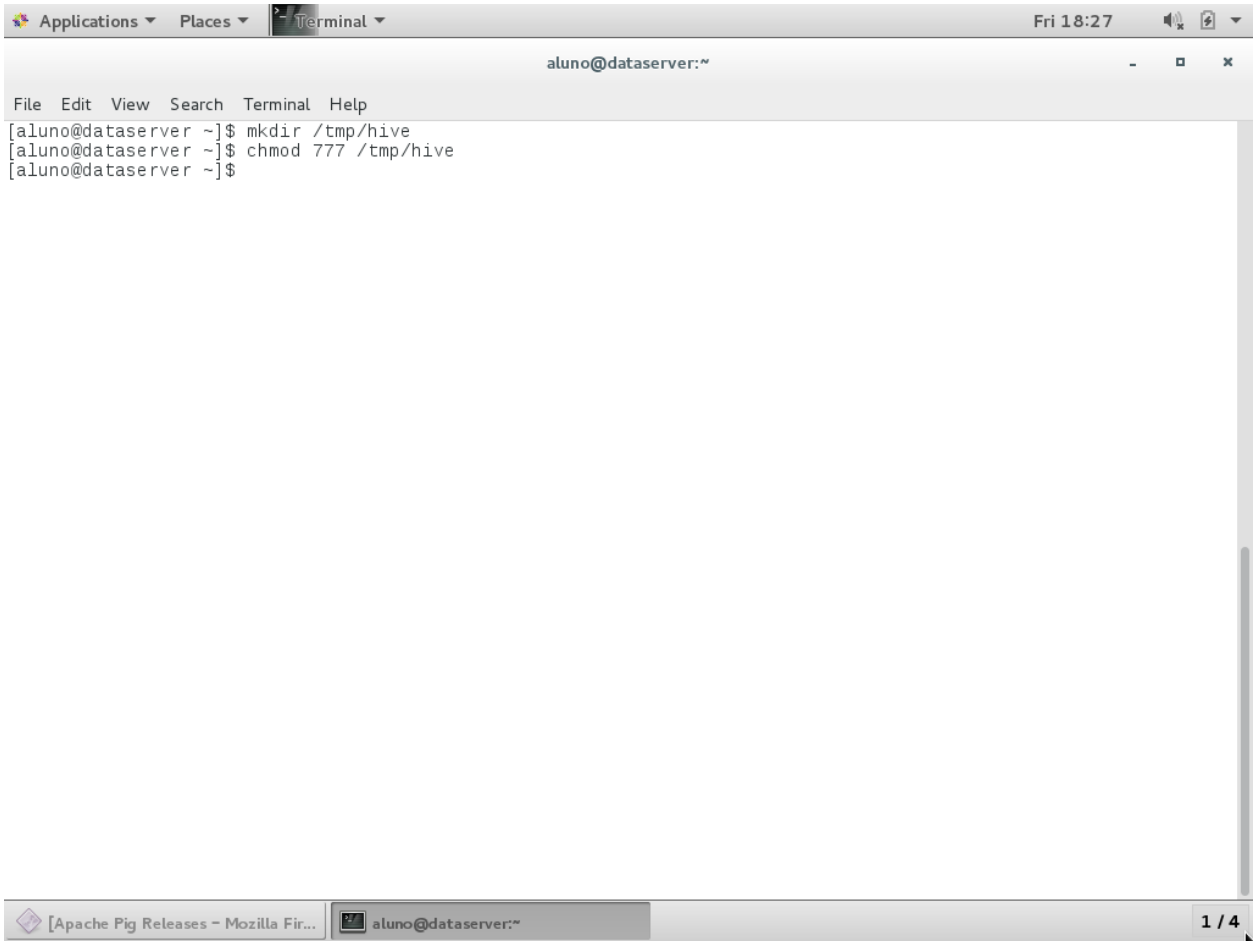
Editando o arquivo .bashrc



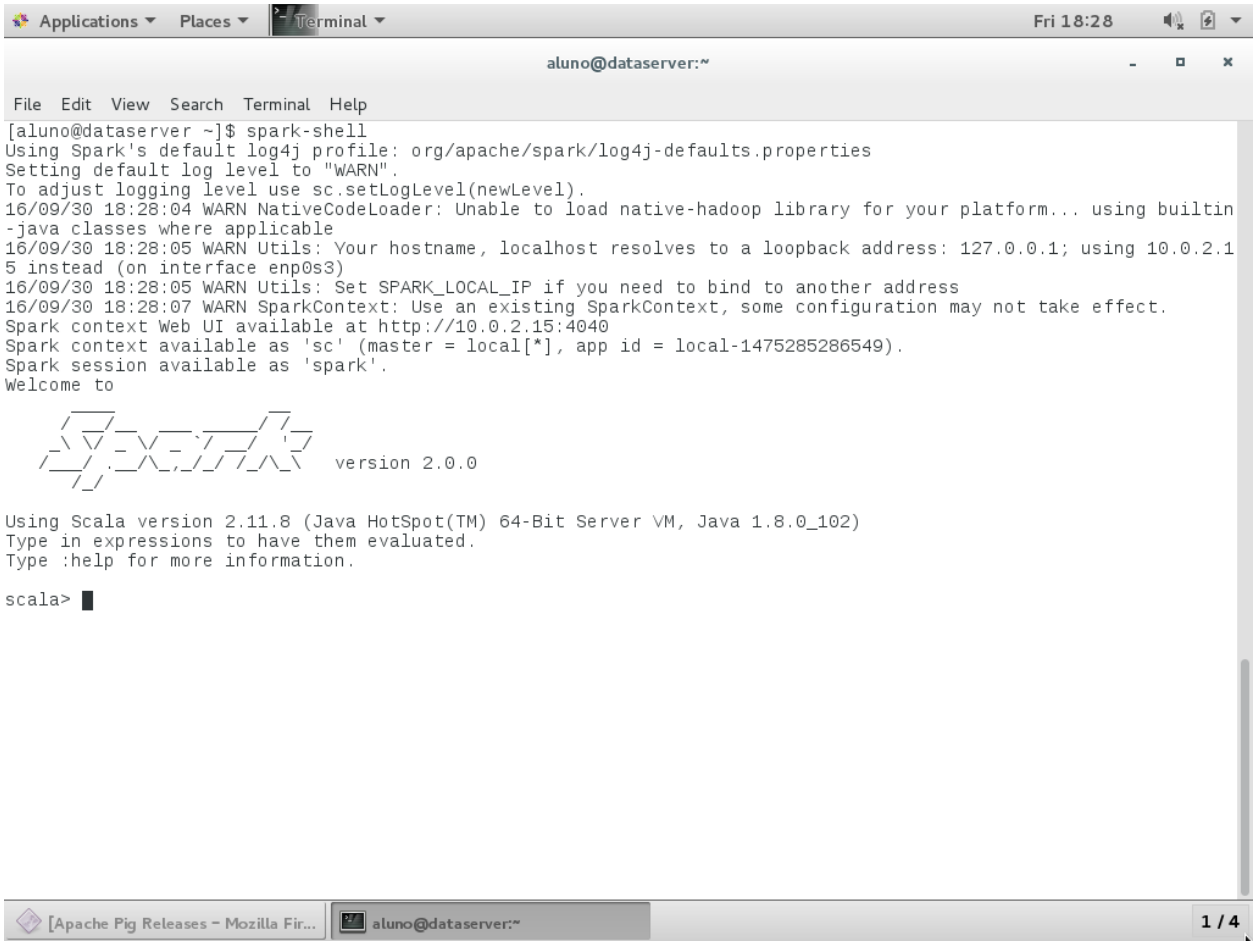
Incluir variáveis Spark



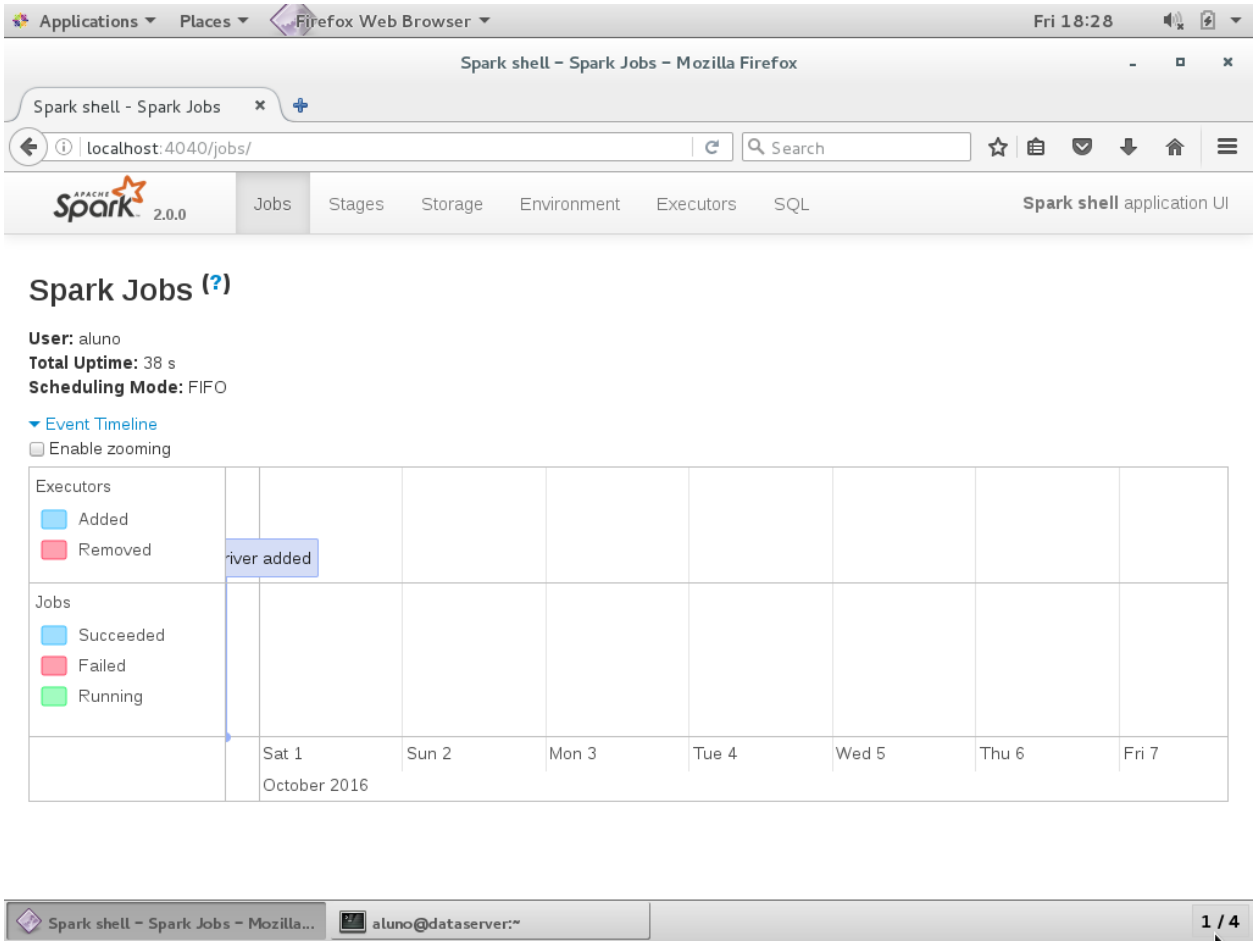
source .bashrc



Se necessário chmod 777 /tmp/hive


Applications
Places
Terminal
Fri 18:28
aluno@dataserver:~
File Edit View Search Terminal Help
[aluno@dataserver ~]$ spark-shell
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel).
16/09/30 18:28:04 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin
-java classes where applicable
16/09/30 18:28:05 WARN Utils: Your hostname, localhost resolves to a loopback address: 127.0.0.1; using 10.0.2.1
5 instead (on interface enp0s3)
16/09/30 18:28:05 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
16/09/30 18:28:07 WARN SparkContext: Use an existing SparkContext, some configuration may not take effect.
Spark context Web UI available at http://10.0.2.15:4040
Spark context available as 'sc' (master = local[*], app id = local-1475285286549).
Spark session available as 'spark'.
Welcome to
version 2.0.0
Using Scala version 2.11.8 (Java HotSpot(TM) 64-Bit Server VM, Java 1.8.0_102)
Type in expressions to have them evaluated.
Type :help for more information.
scala>
[Apache Pig Releases – Mozilla Fir...
aluno@dataserver:~
1 / 4

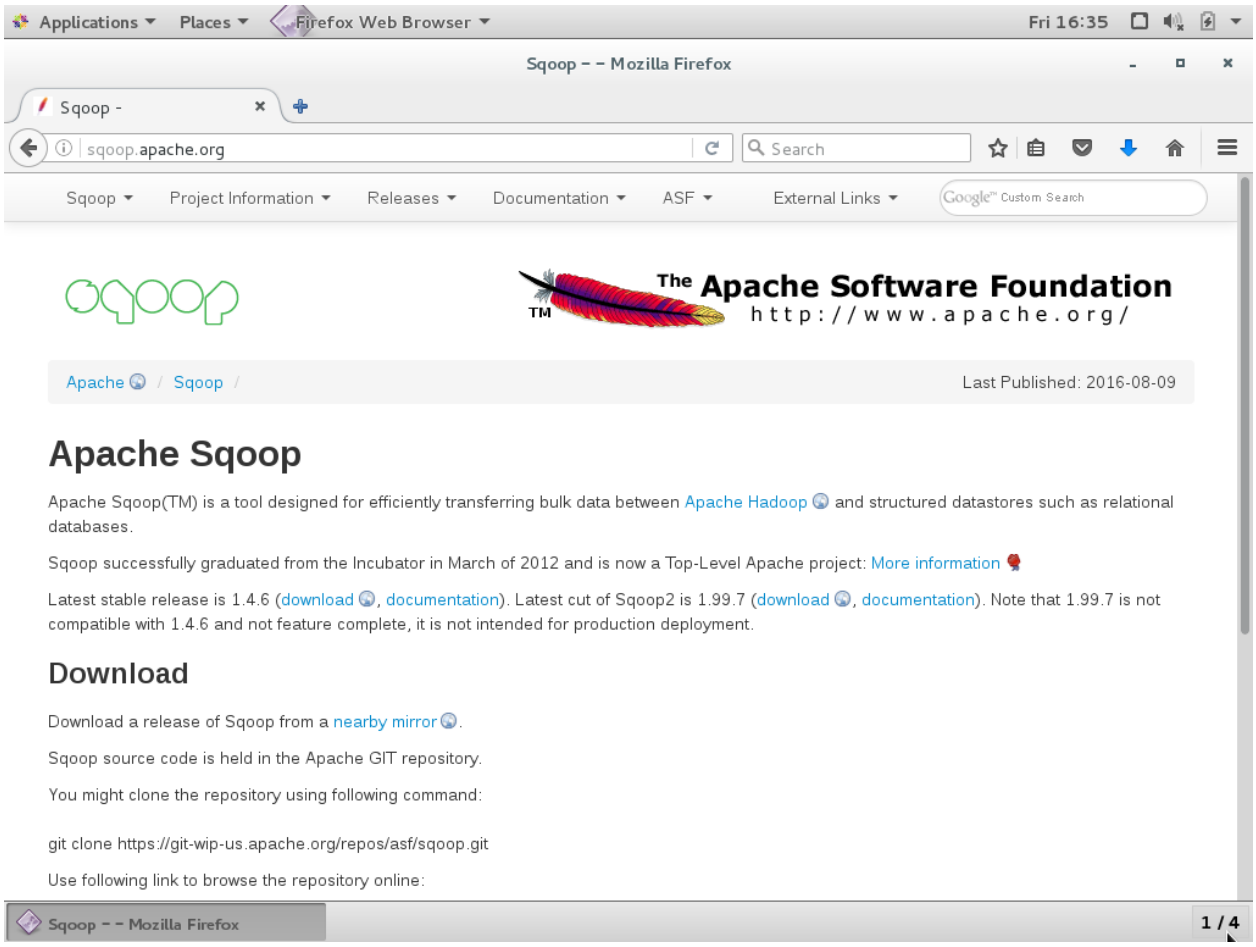
Spark shell



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

11. Instalação e Configuração do Sqoop

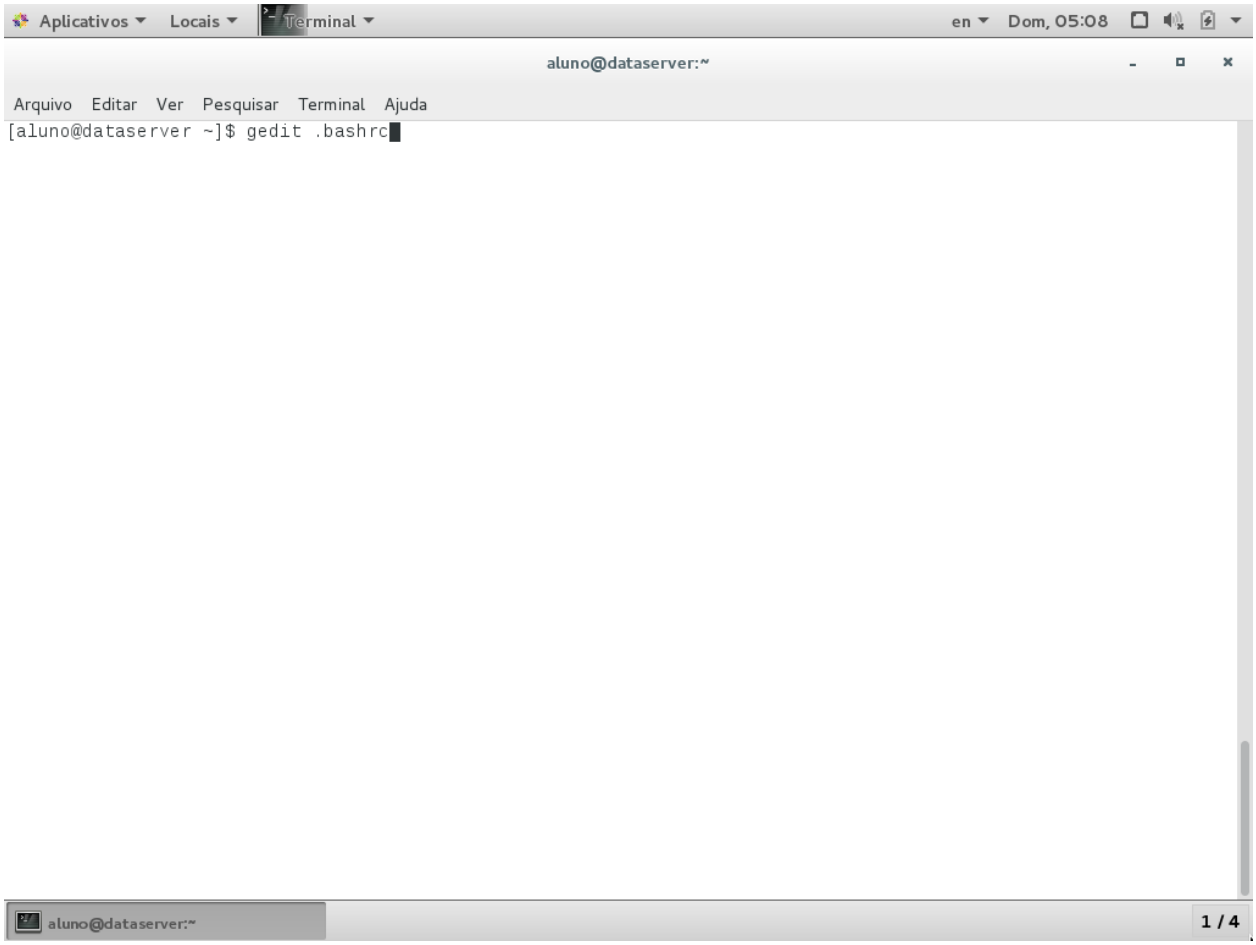
11.1. Download do Sqoop



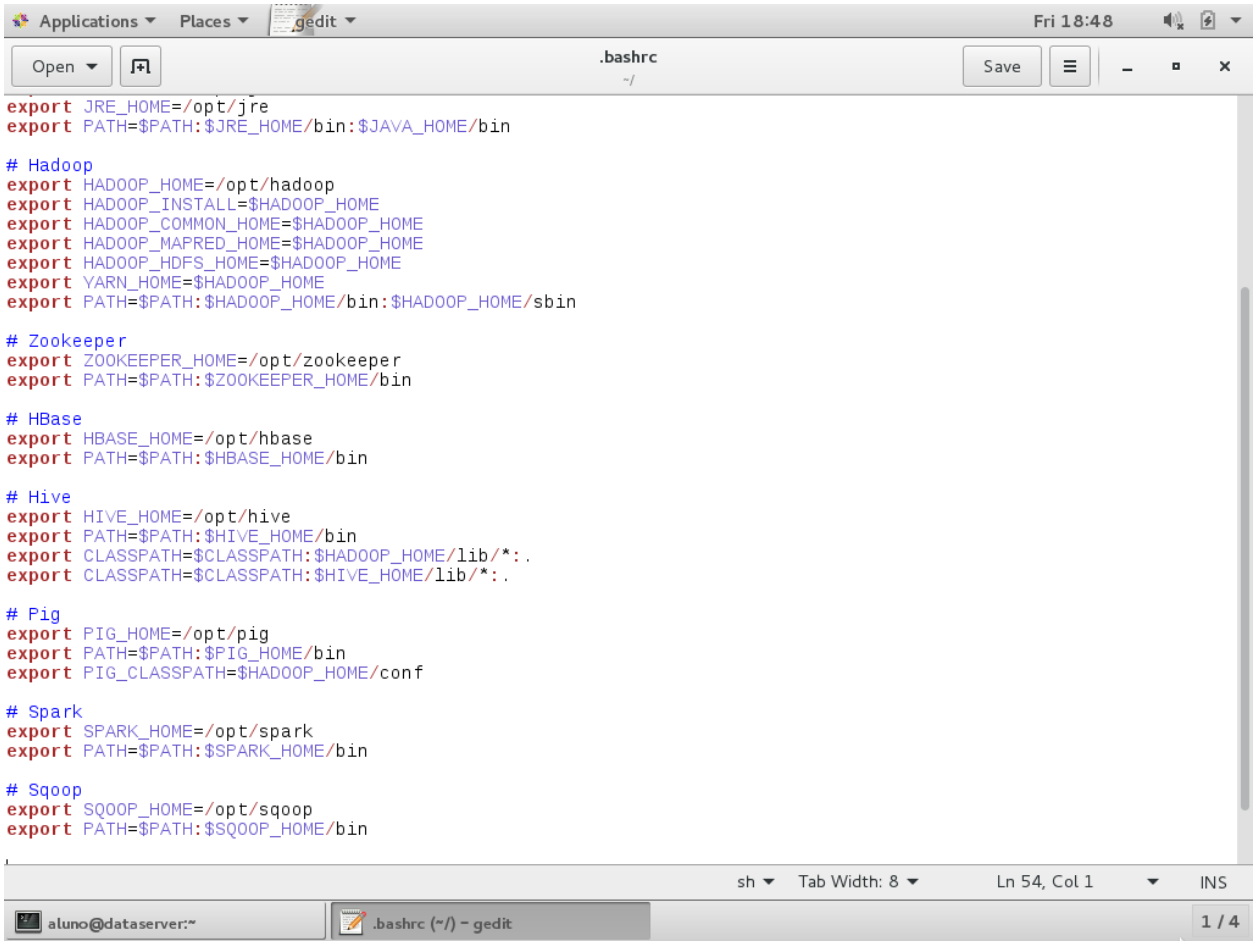
Download do Sqoop – Versão 1.4.7-hadoop-2.0.4-alpha

Faça o download, descompacte o arquivo e mova o diretório para /opt/sqoop da mesma forma como você fez com o Java JDK e com o Hadoop.

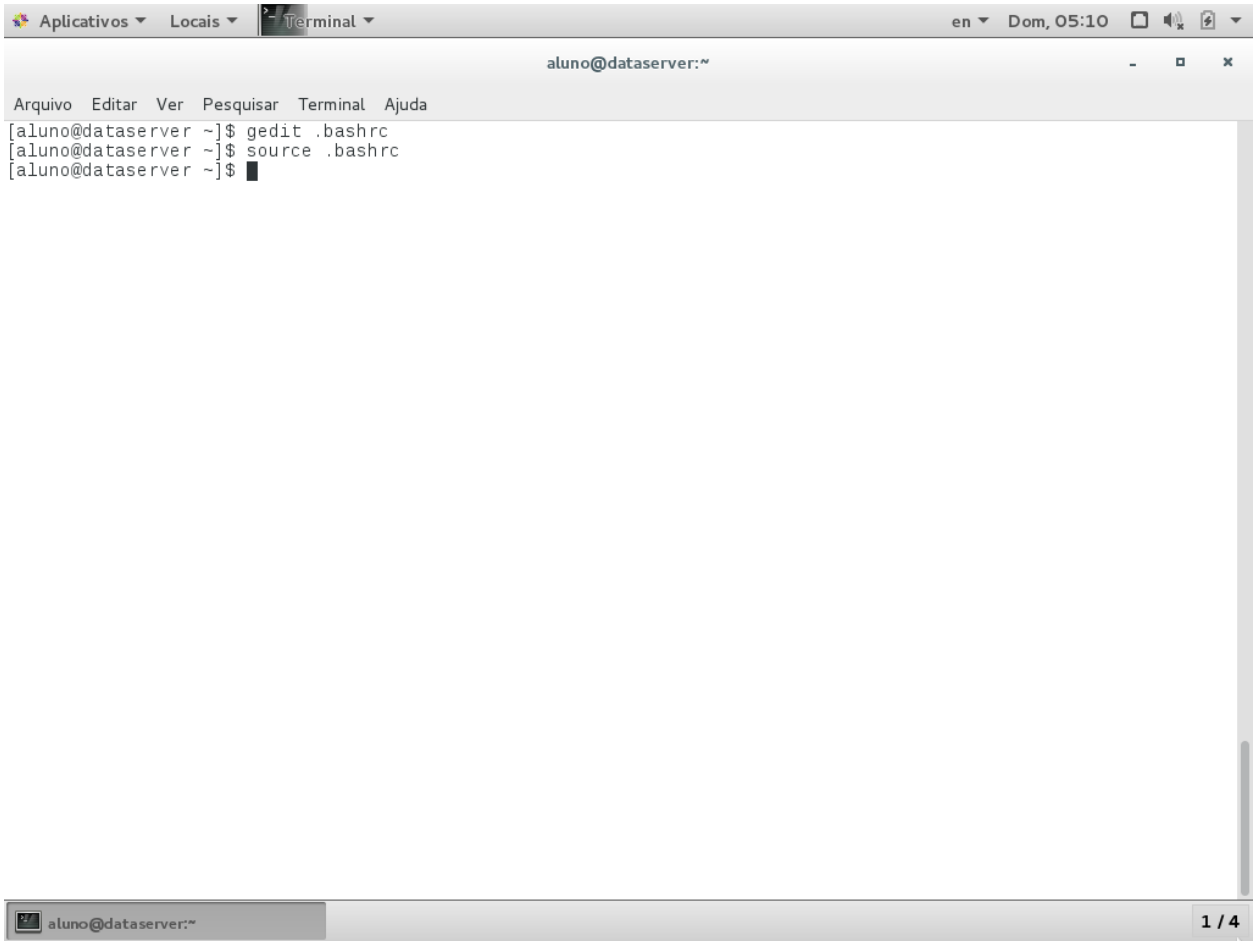
11.2. Configuração do Sqoop



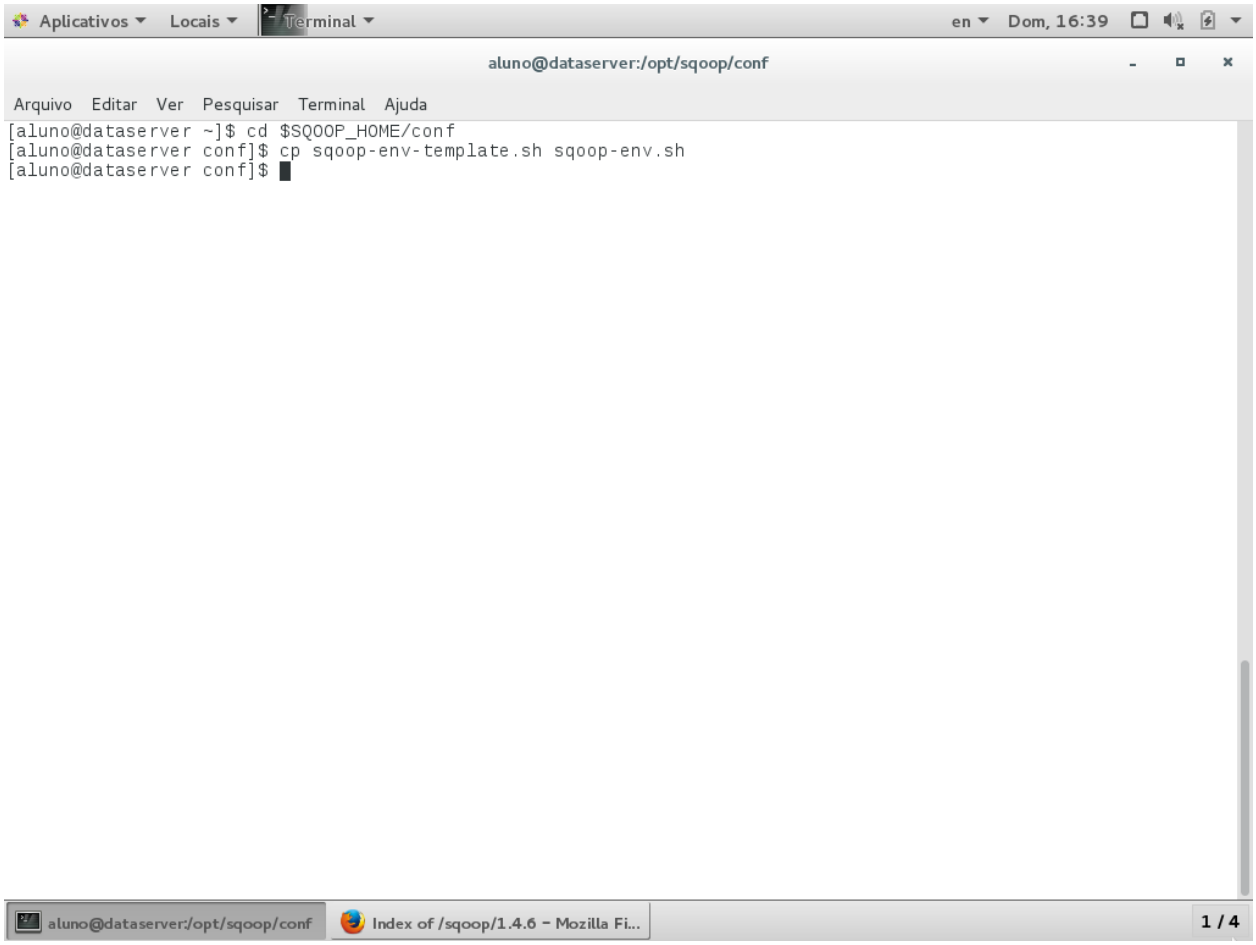
Editar arquivo .bashrc



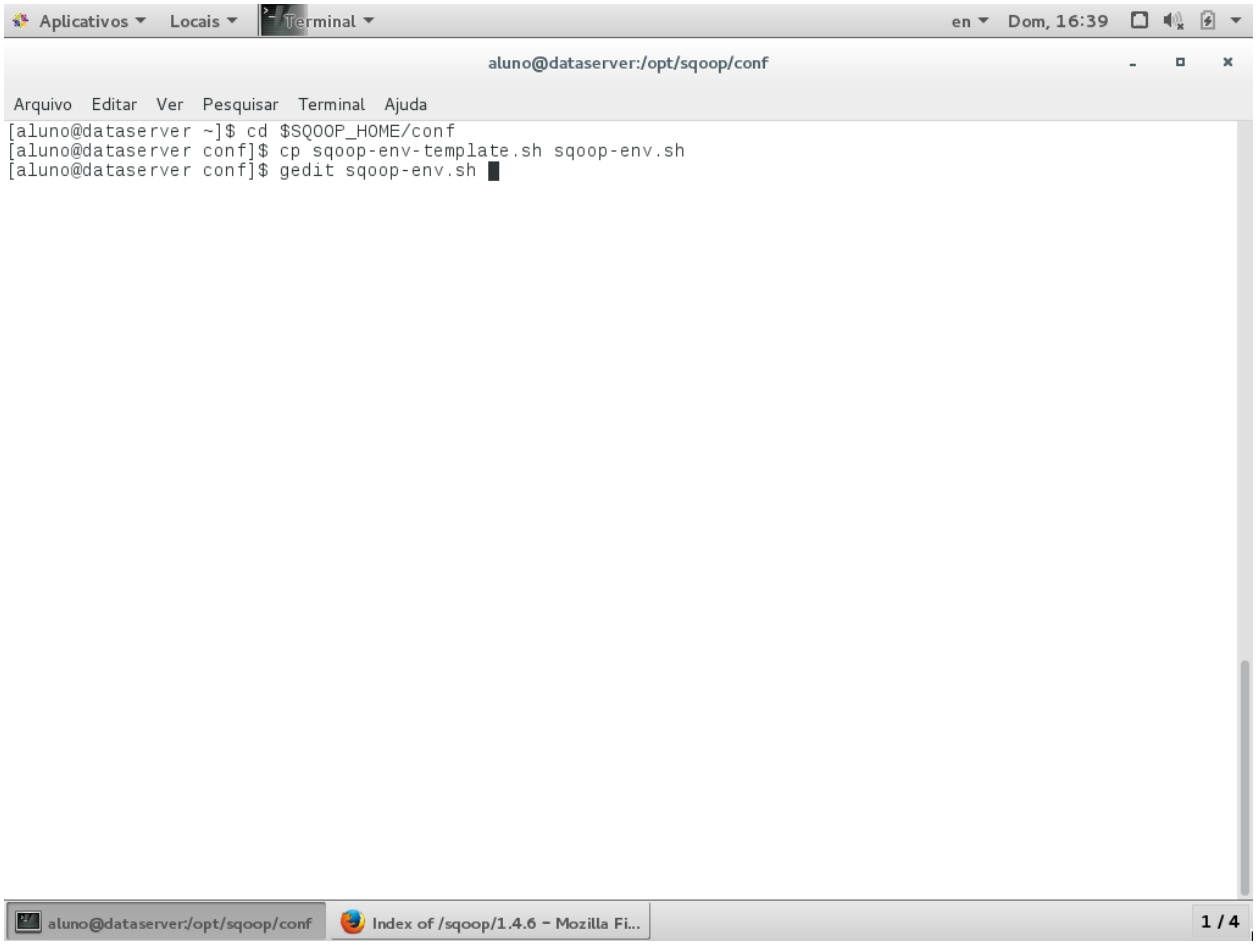
Incluir variáveis Sqoop



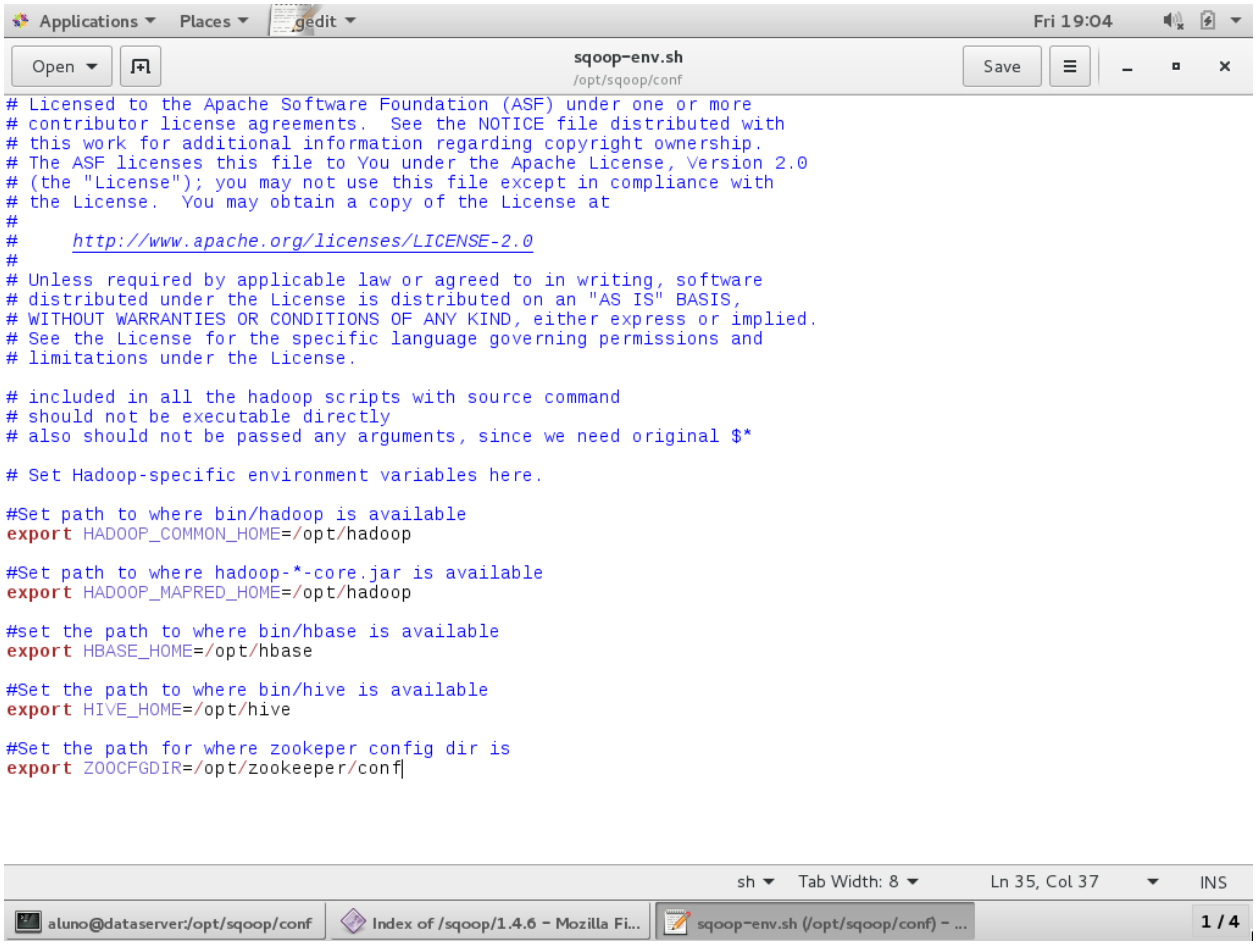
source .bashrc



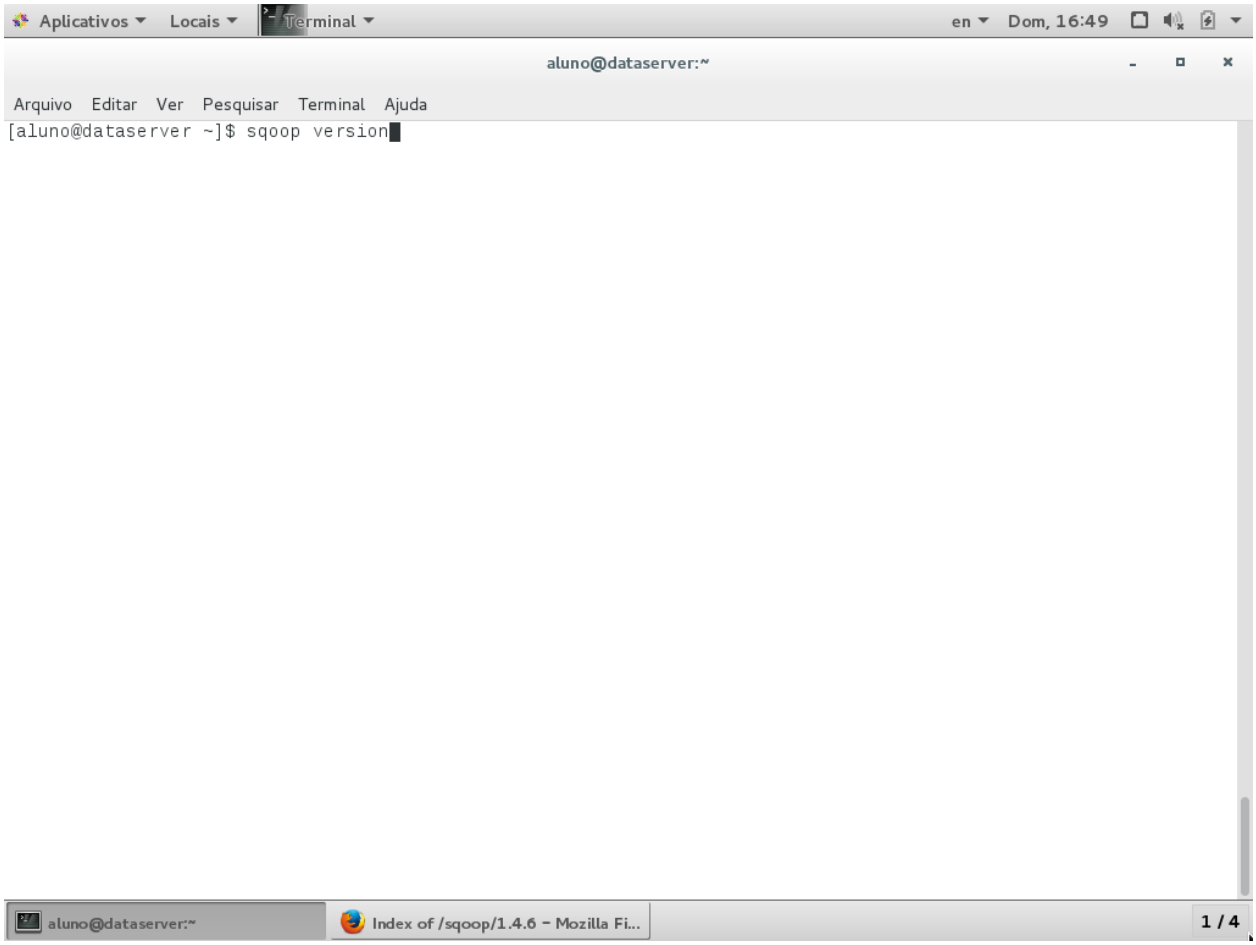
A partir do template, criar o arquivo sqoop-env.sh e editá-lo



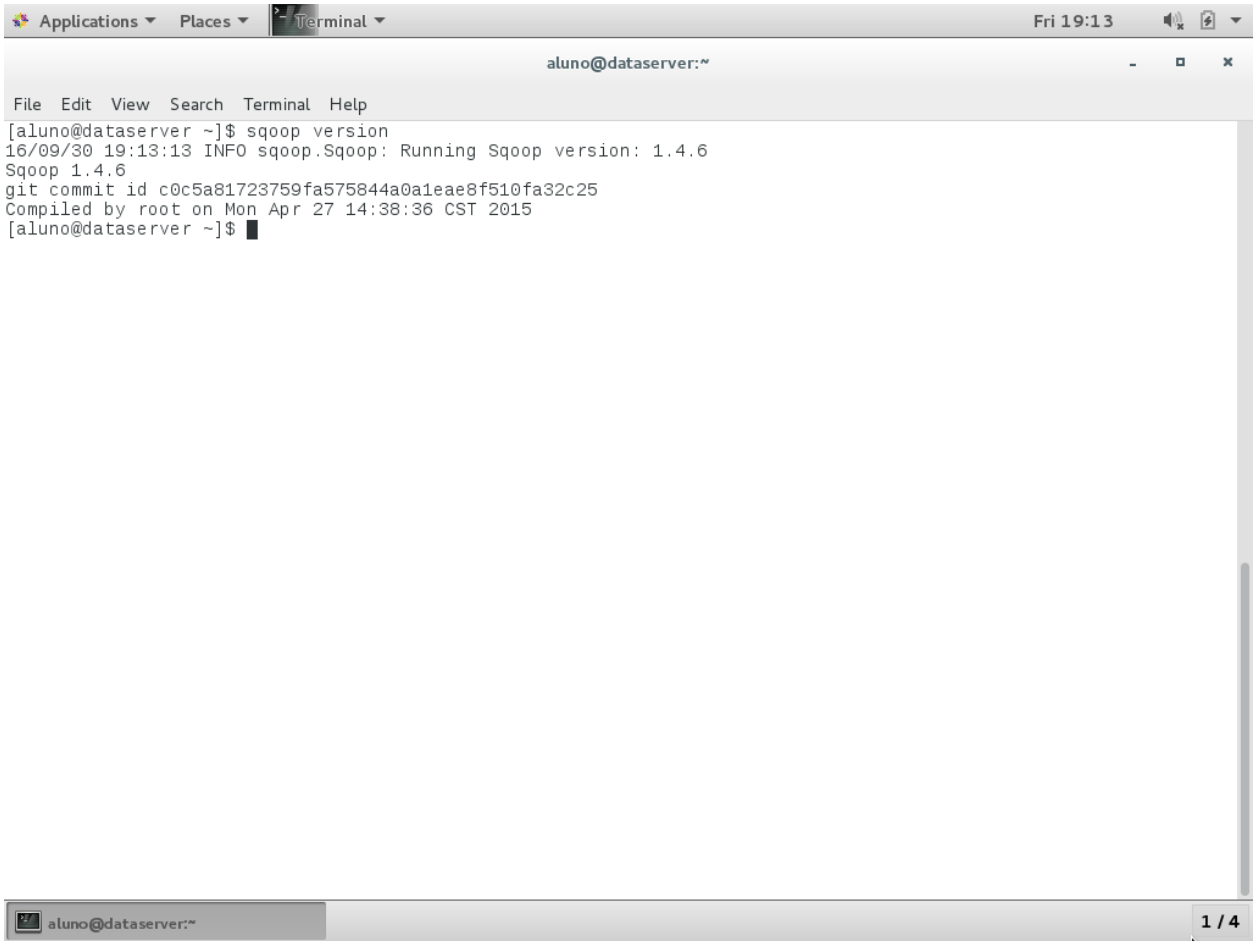
Editar o arquivo



Editar variáveis conforme tela acima

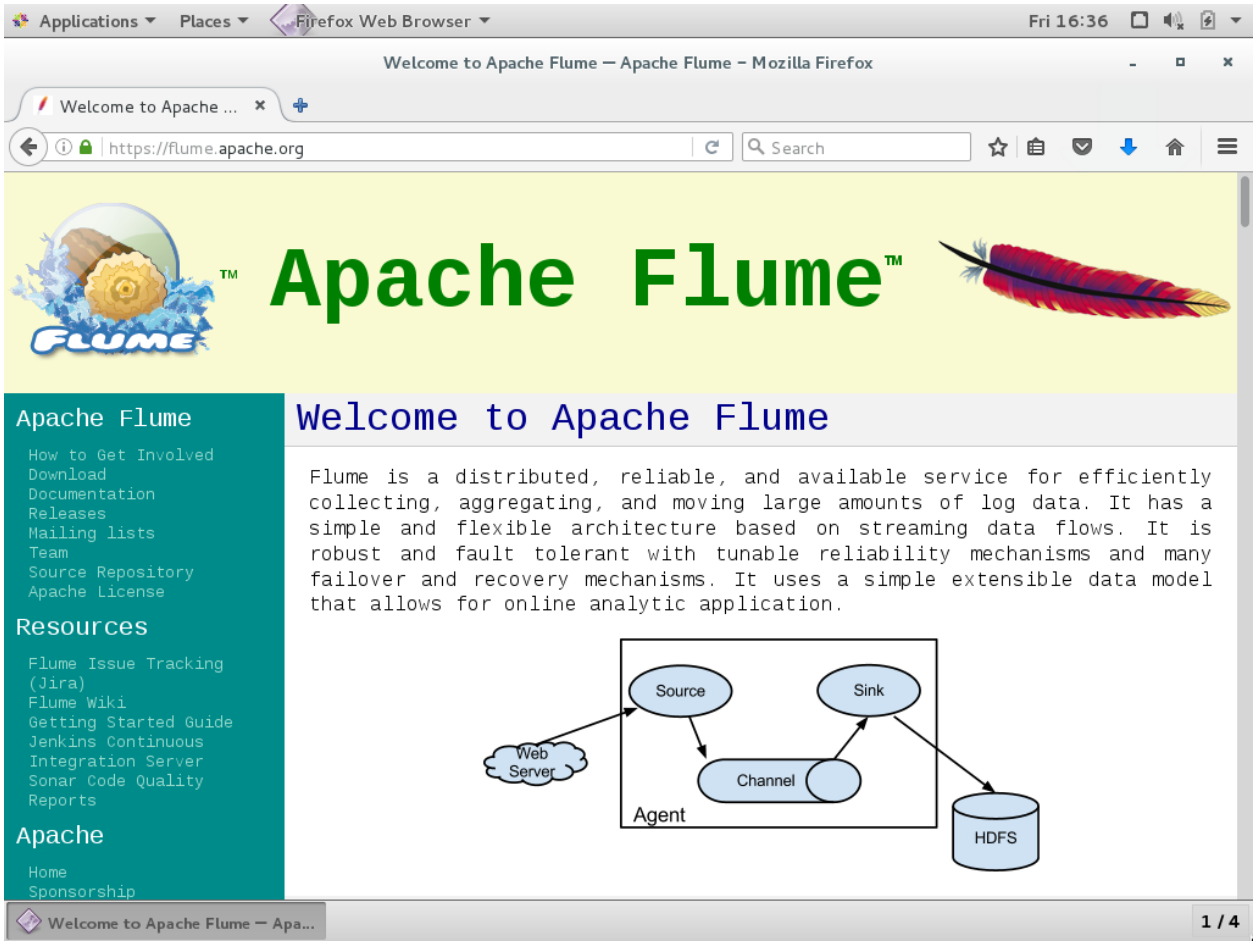


sqoop version



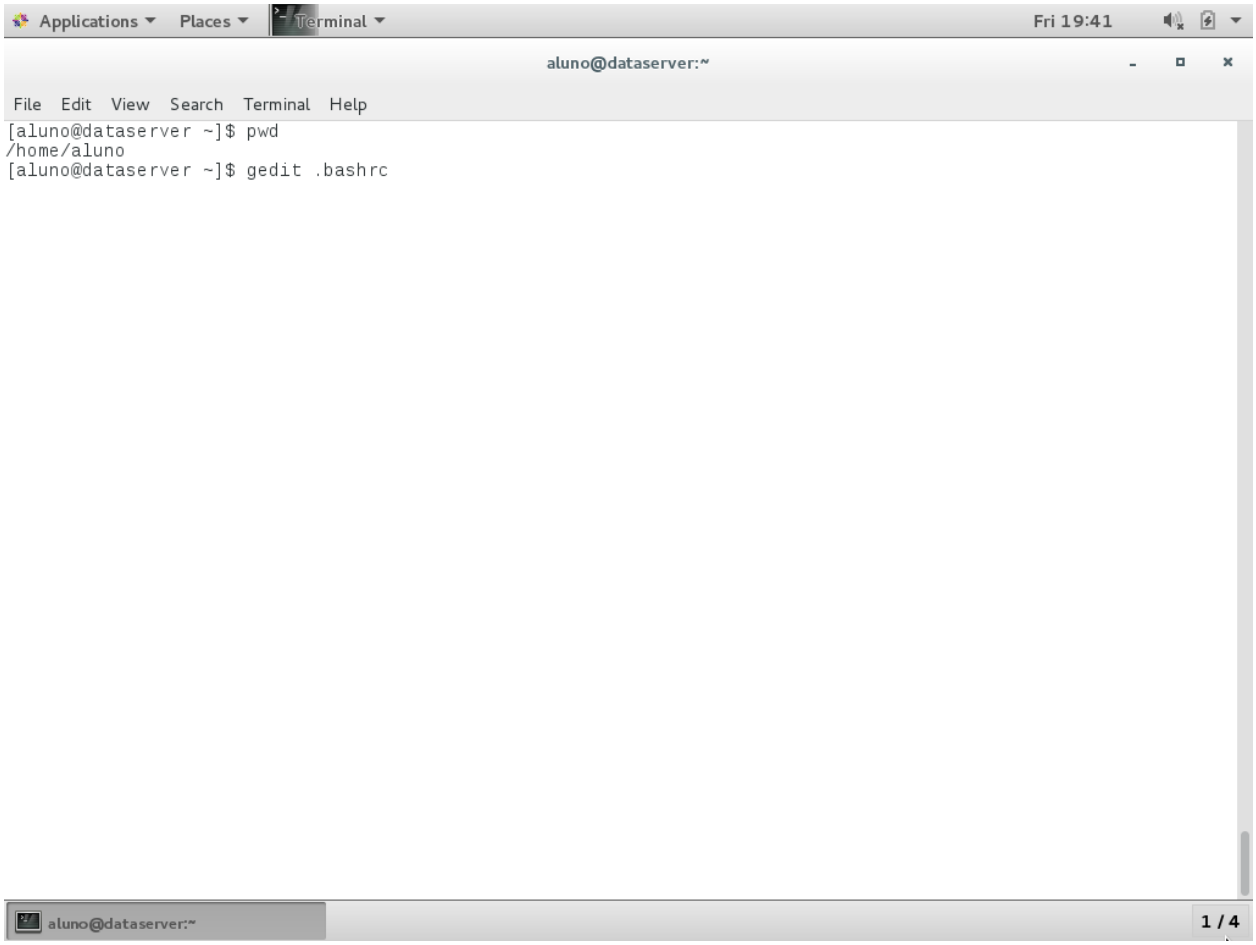
Sqoop version

12. Instalação e Configuração do Apache Flume

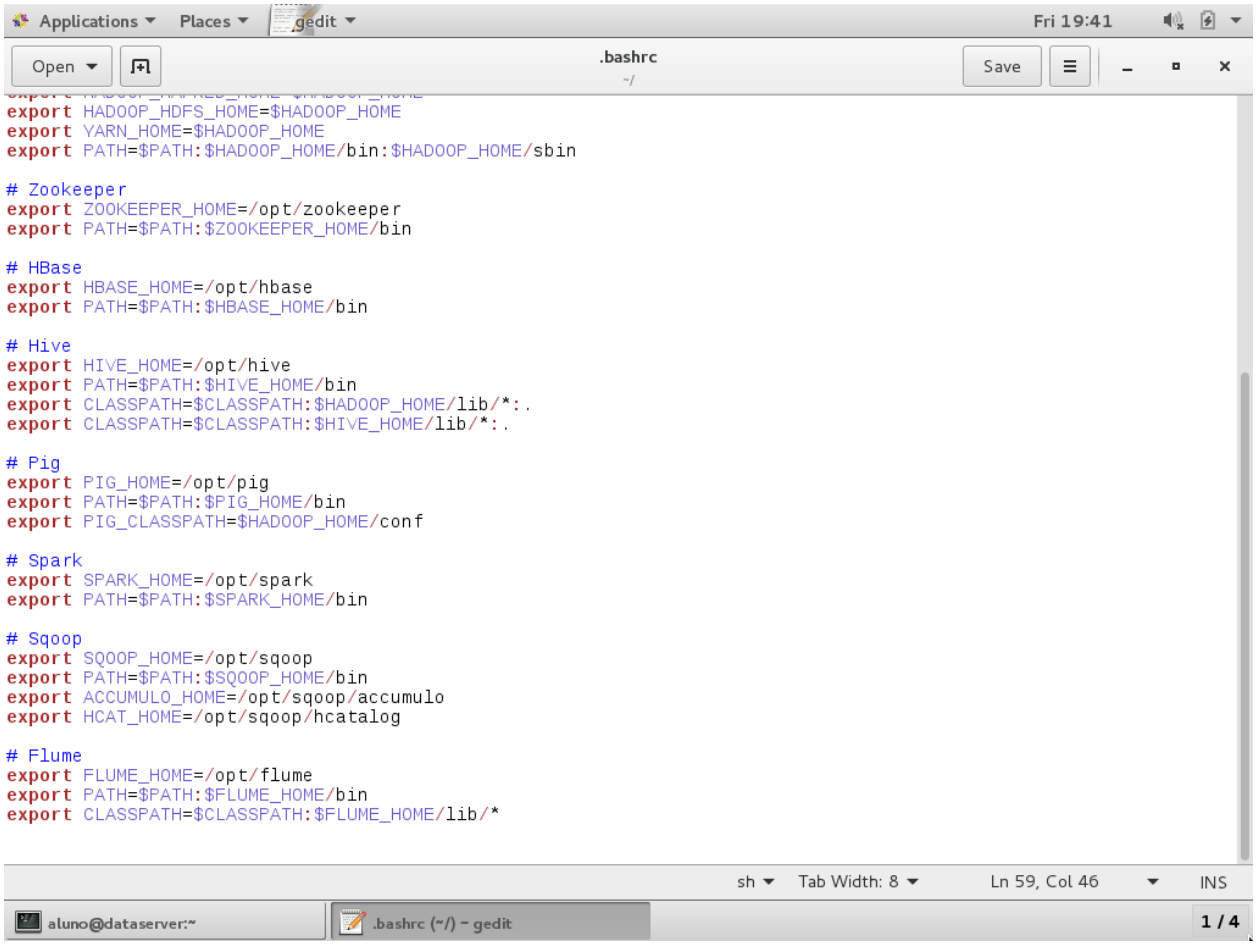


Download do Apache Flume – Versão 1.9

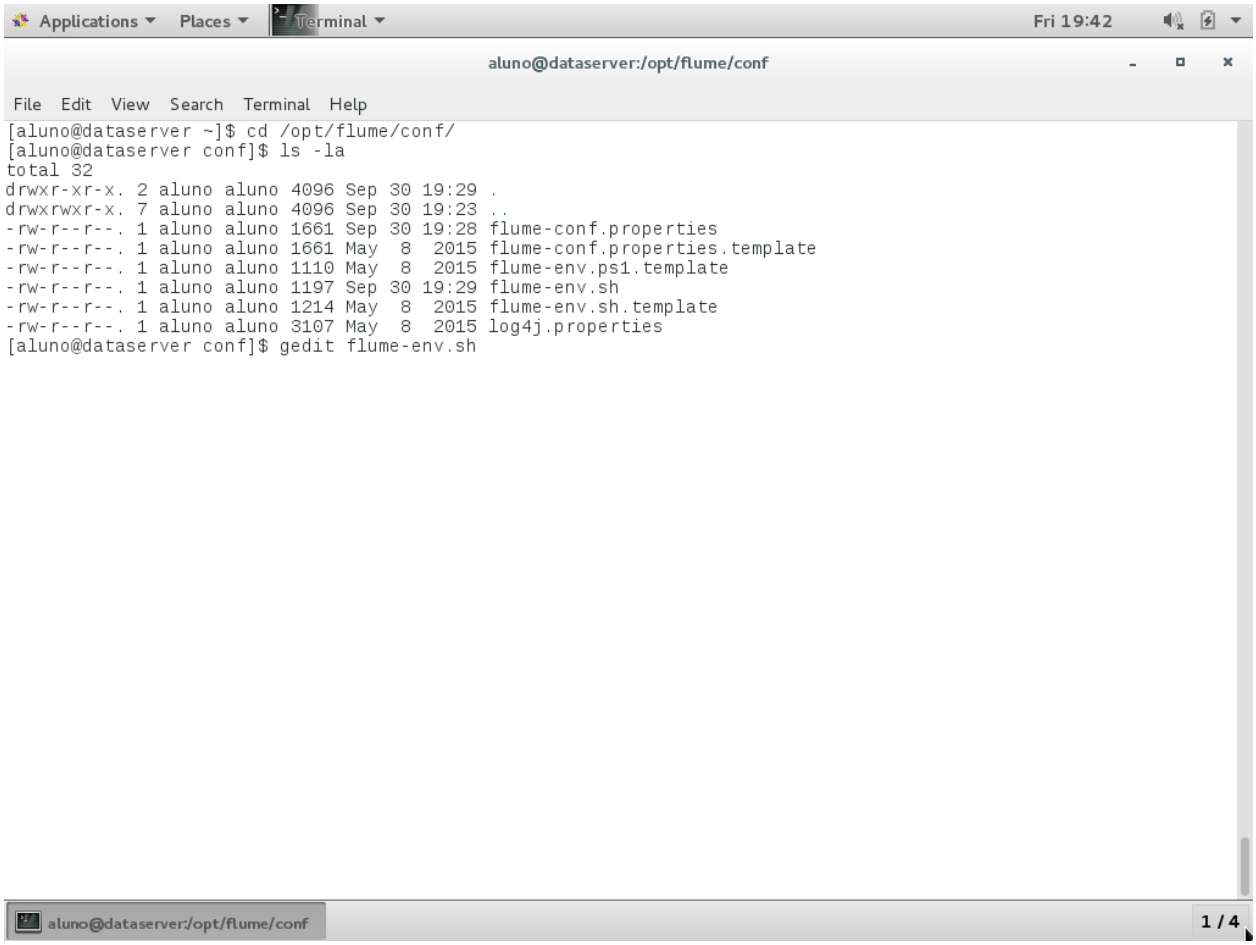
Faça o download, descompacte o arquivo e mova o diretório para /opt/flume da mesma forma como você fez com o Java JDK e com o Hadoop.



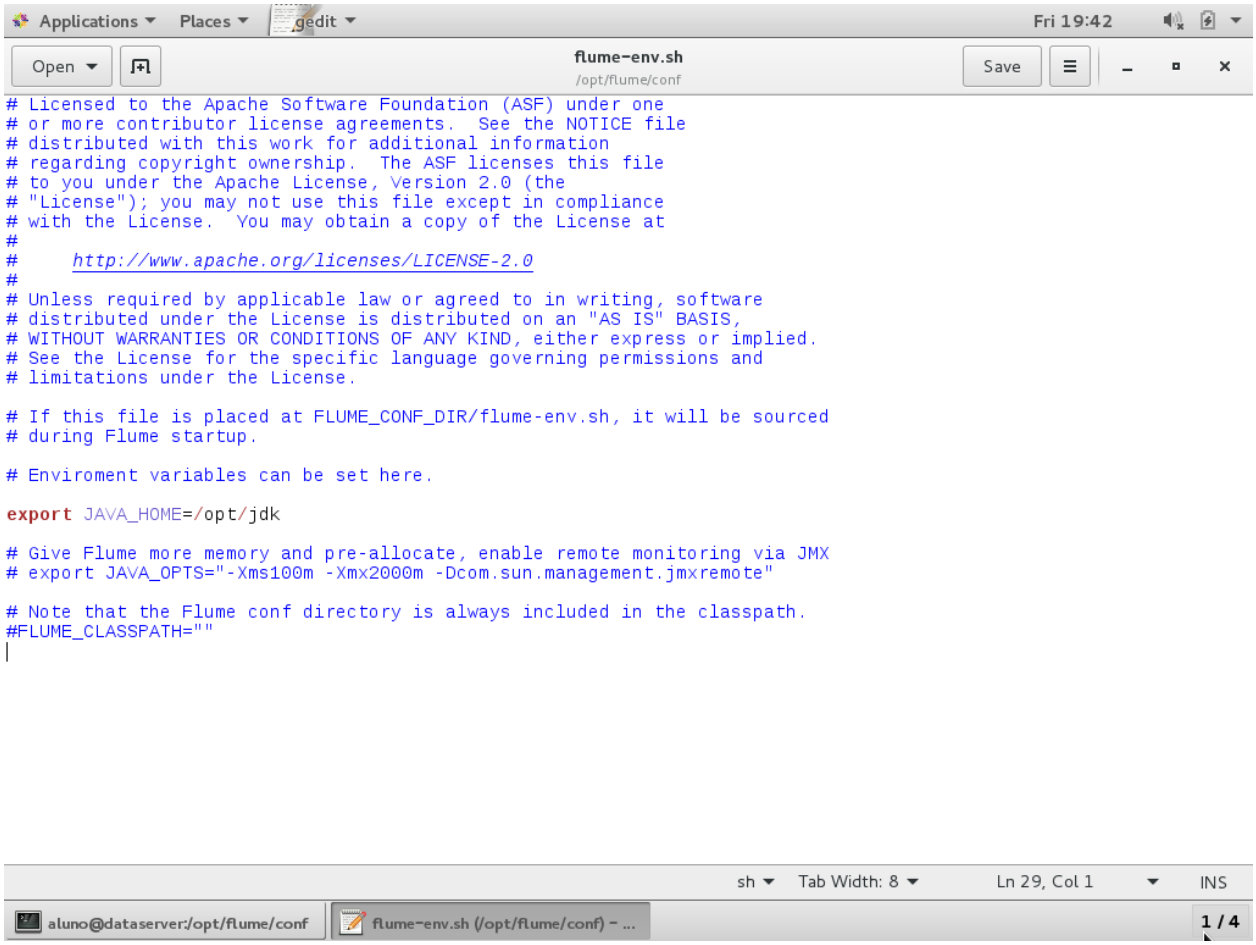
Editar as variáveis de ambiente



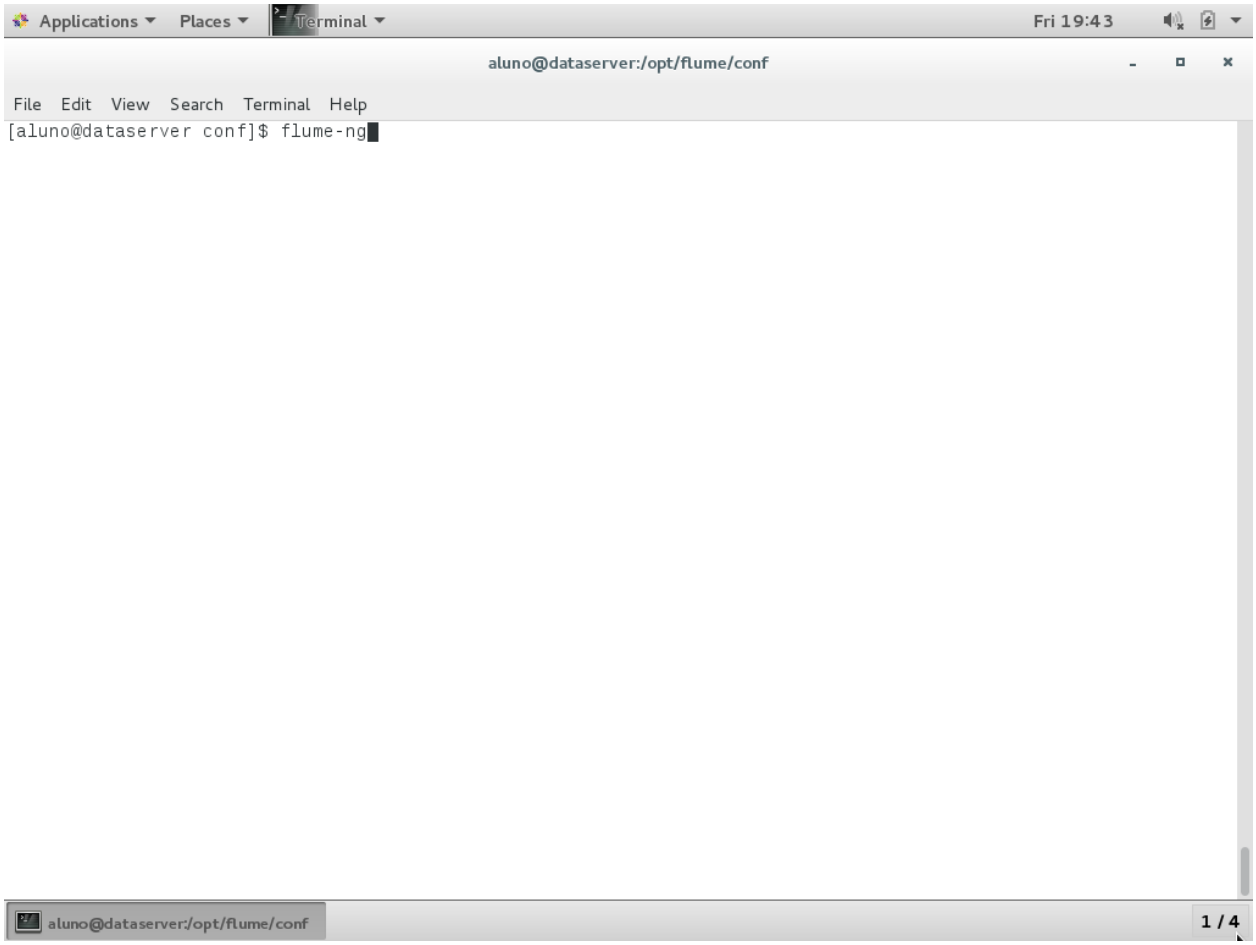
Variáveis de ambiente para o Flume



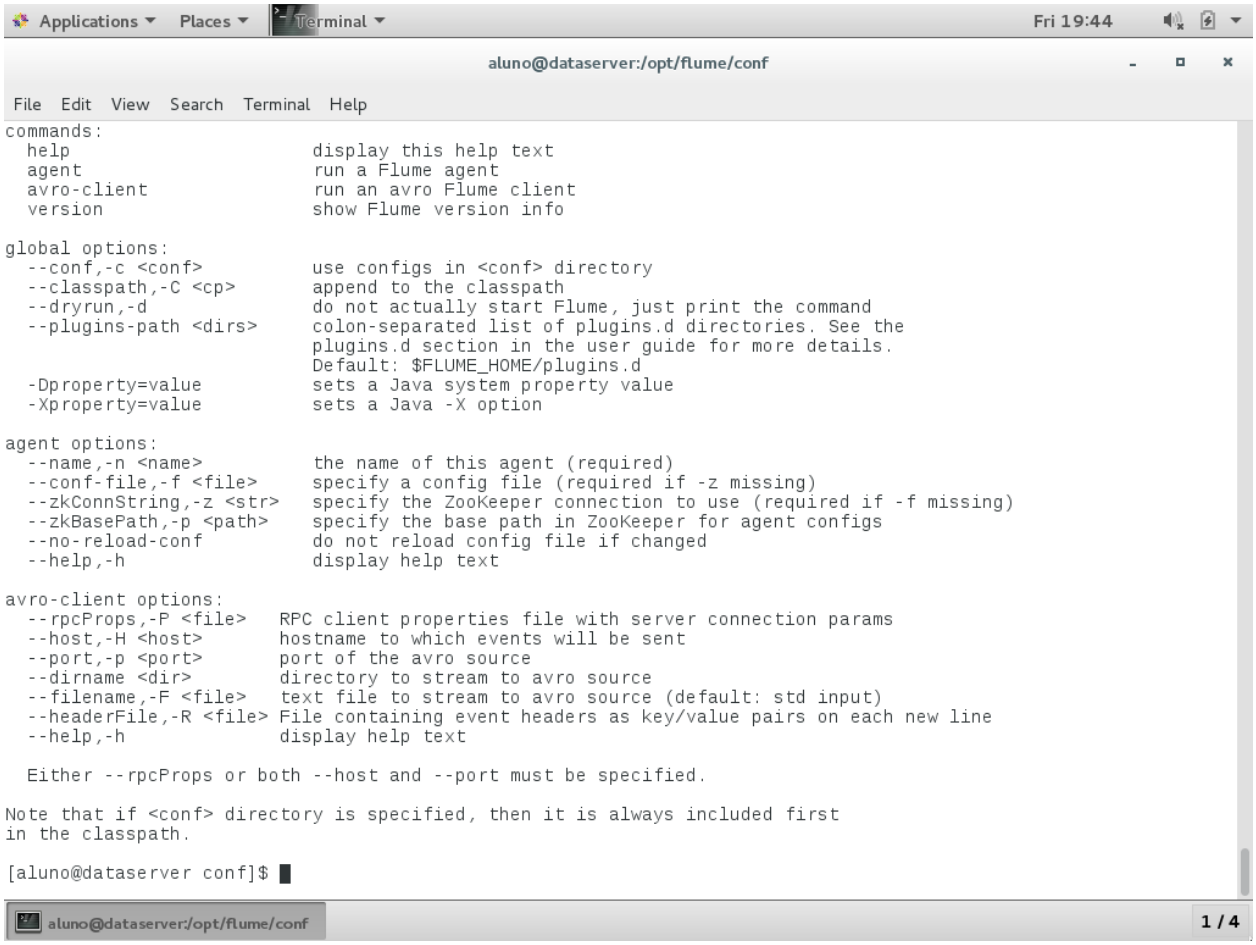
Editar o arquivo flume-env.sh



Acrescentar o JAVA_HOME



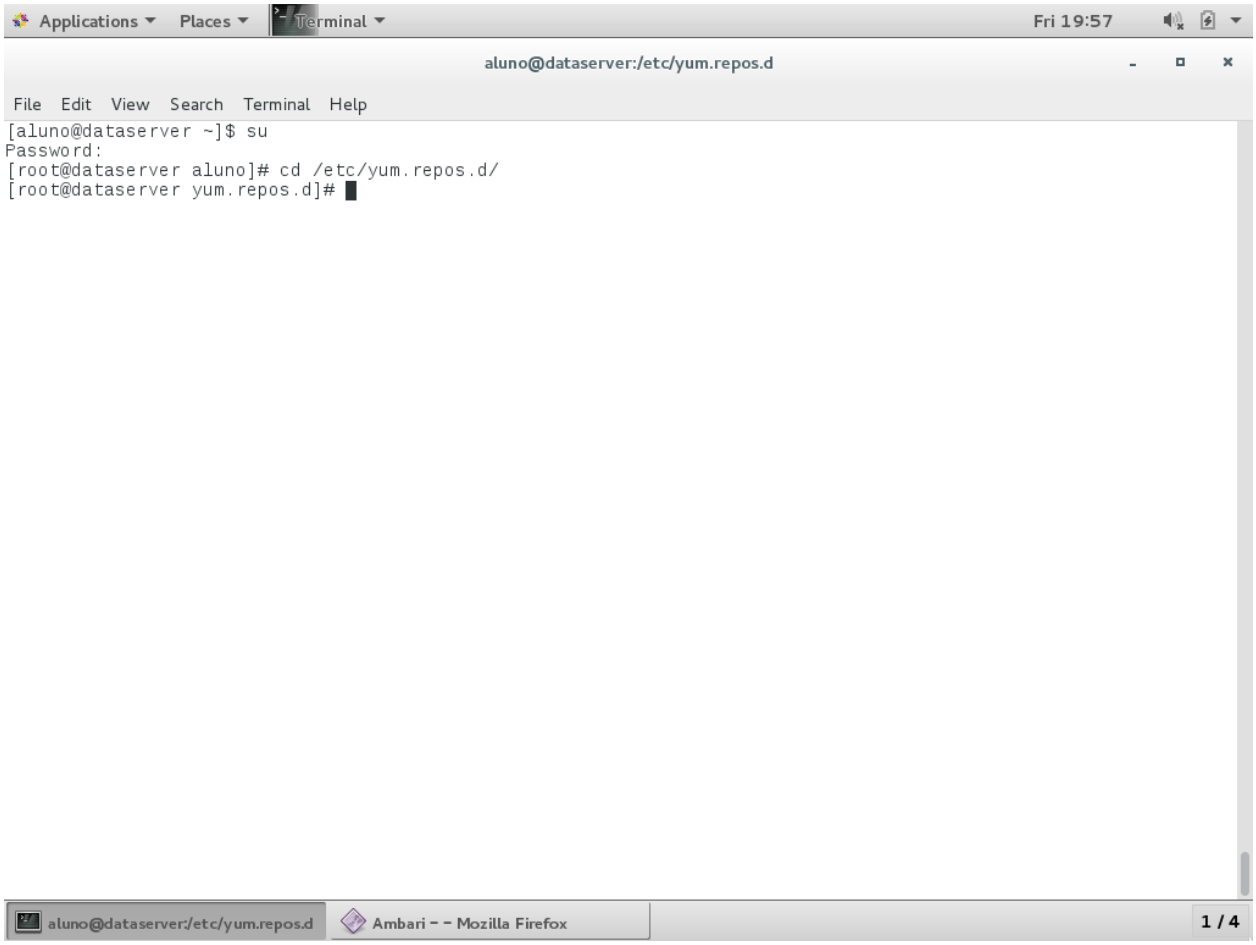
Testar a instalação



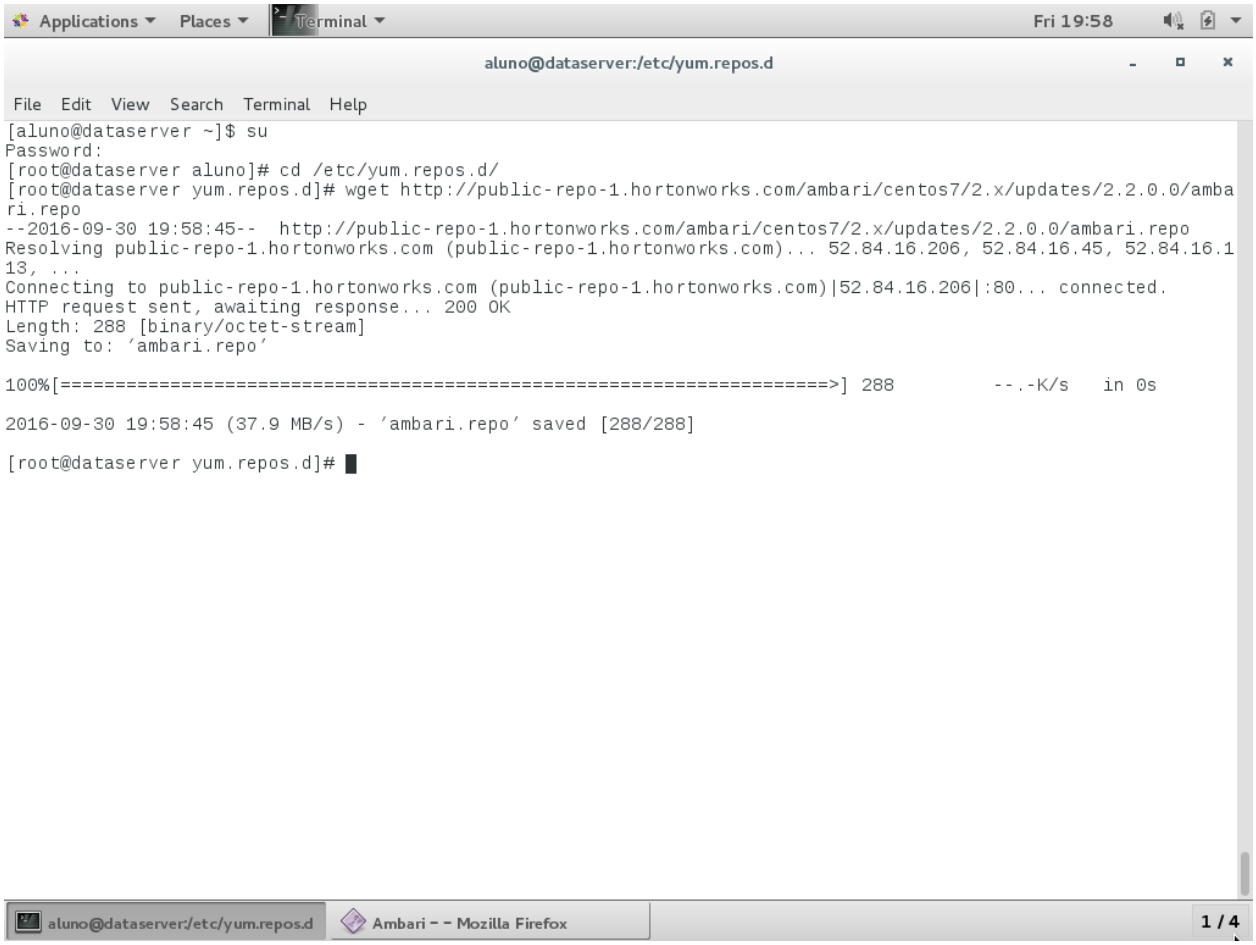
Flume instalado com sucesso

13. Instalação e Configuração do Ambari (Opcional)

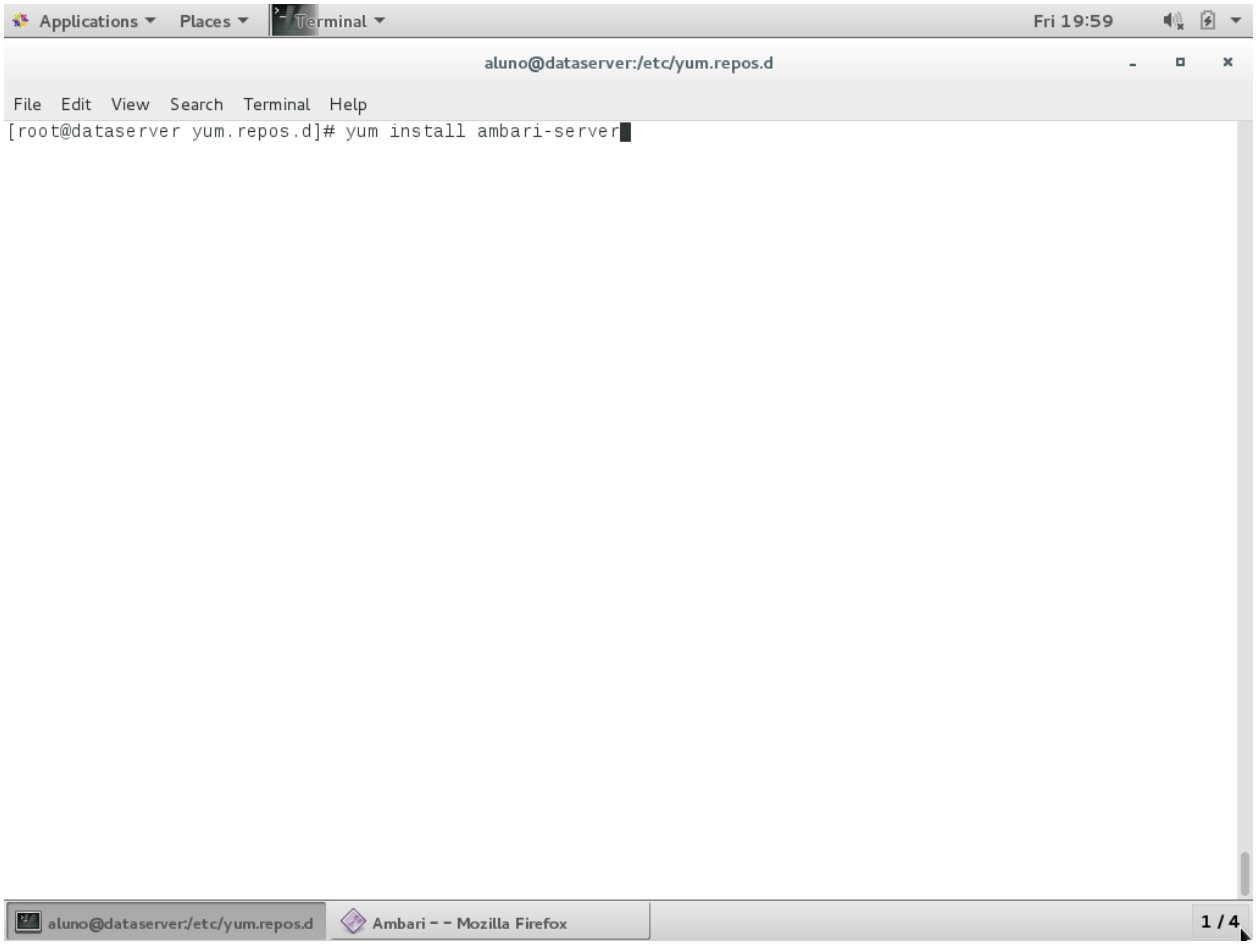
Nota: No CentOS, o Ambari pode ser instalado mais facilmente através do gerenciador de pacotes yum.



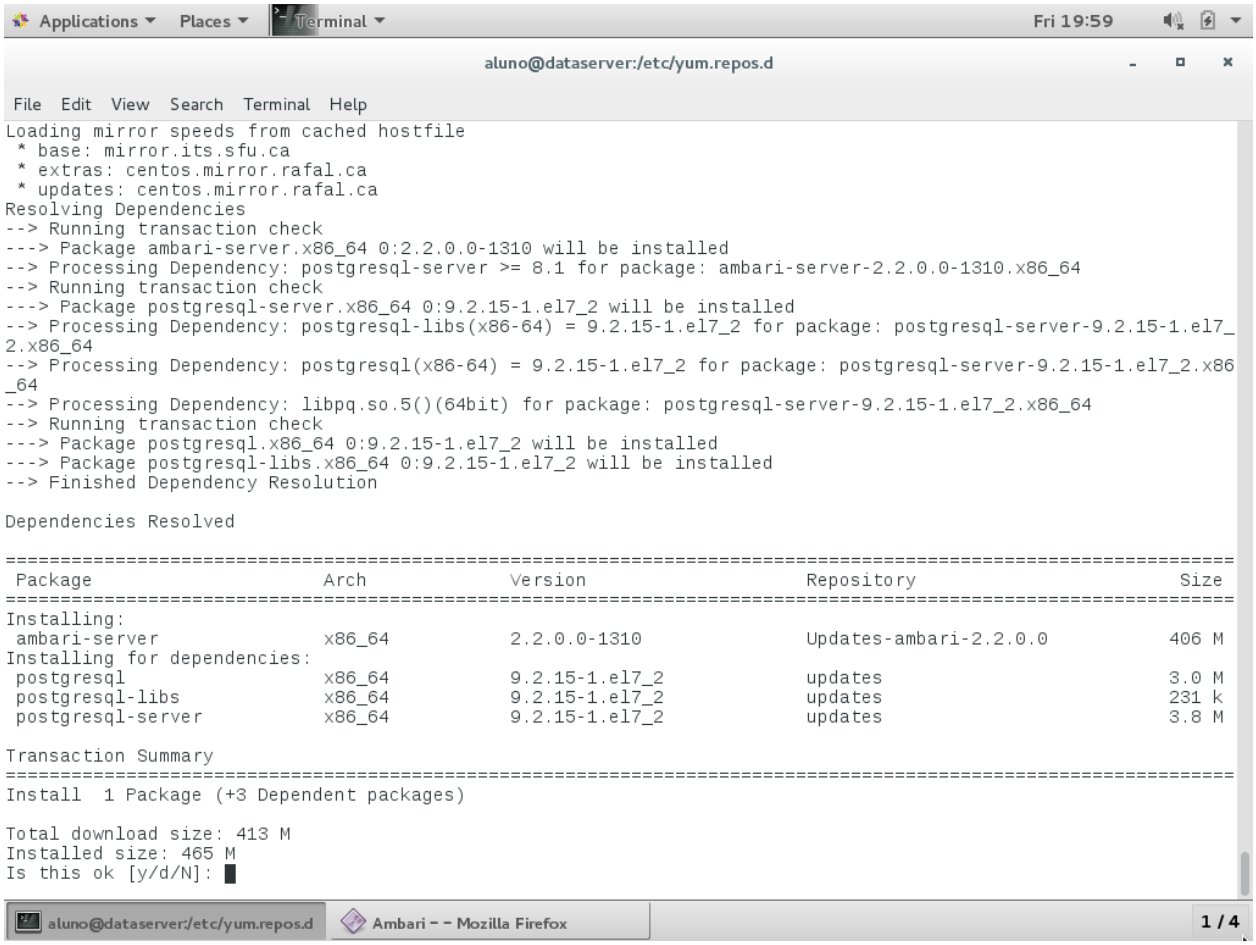
Conectado como root, acessar o diretório de repositórios do CentOS



Download do arquivo de repositório do Ambari

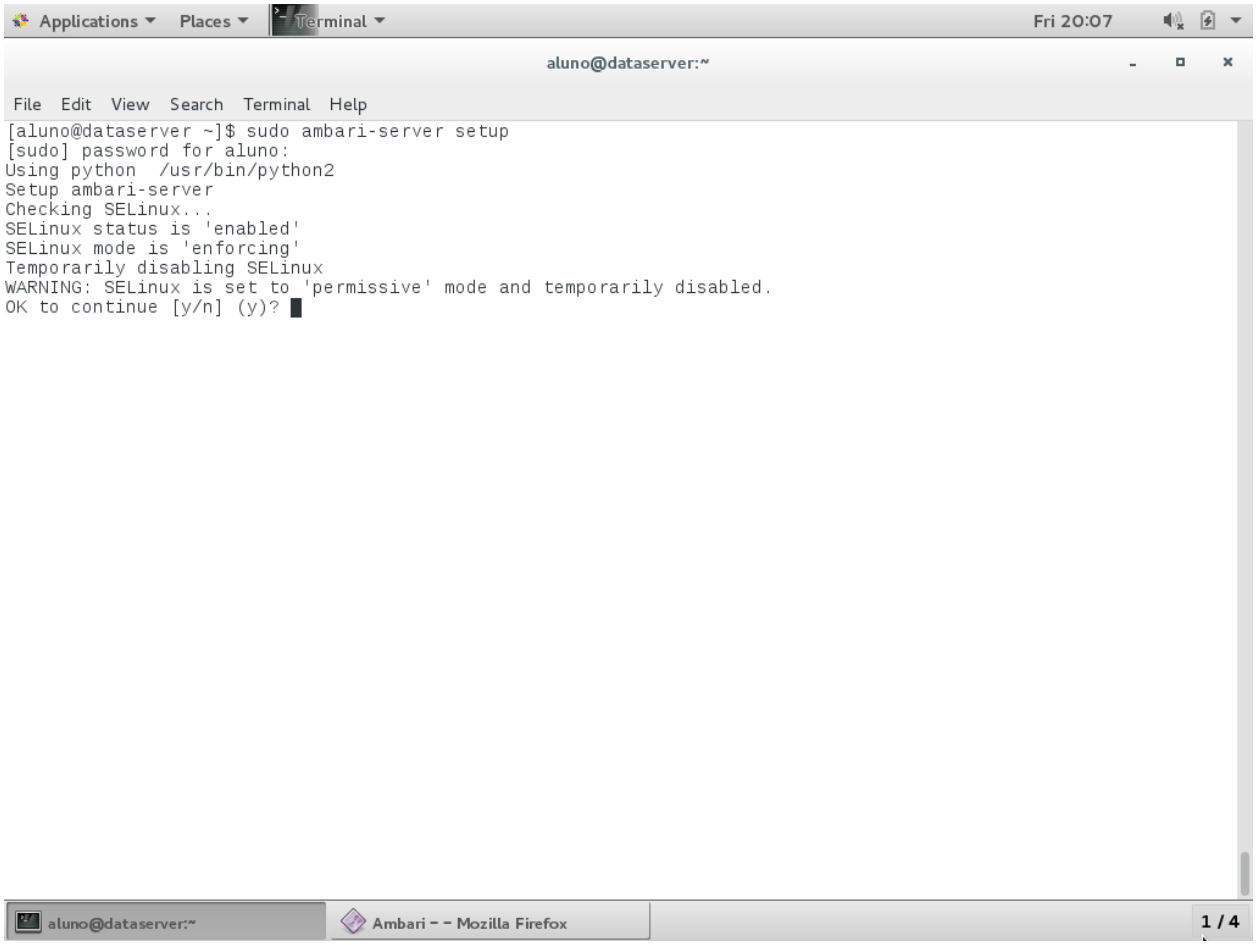


Como root, executar: yum install ambari-server

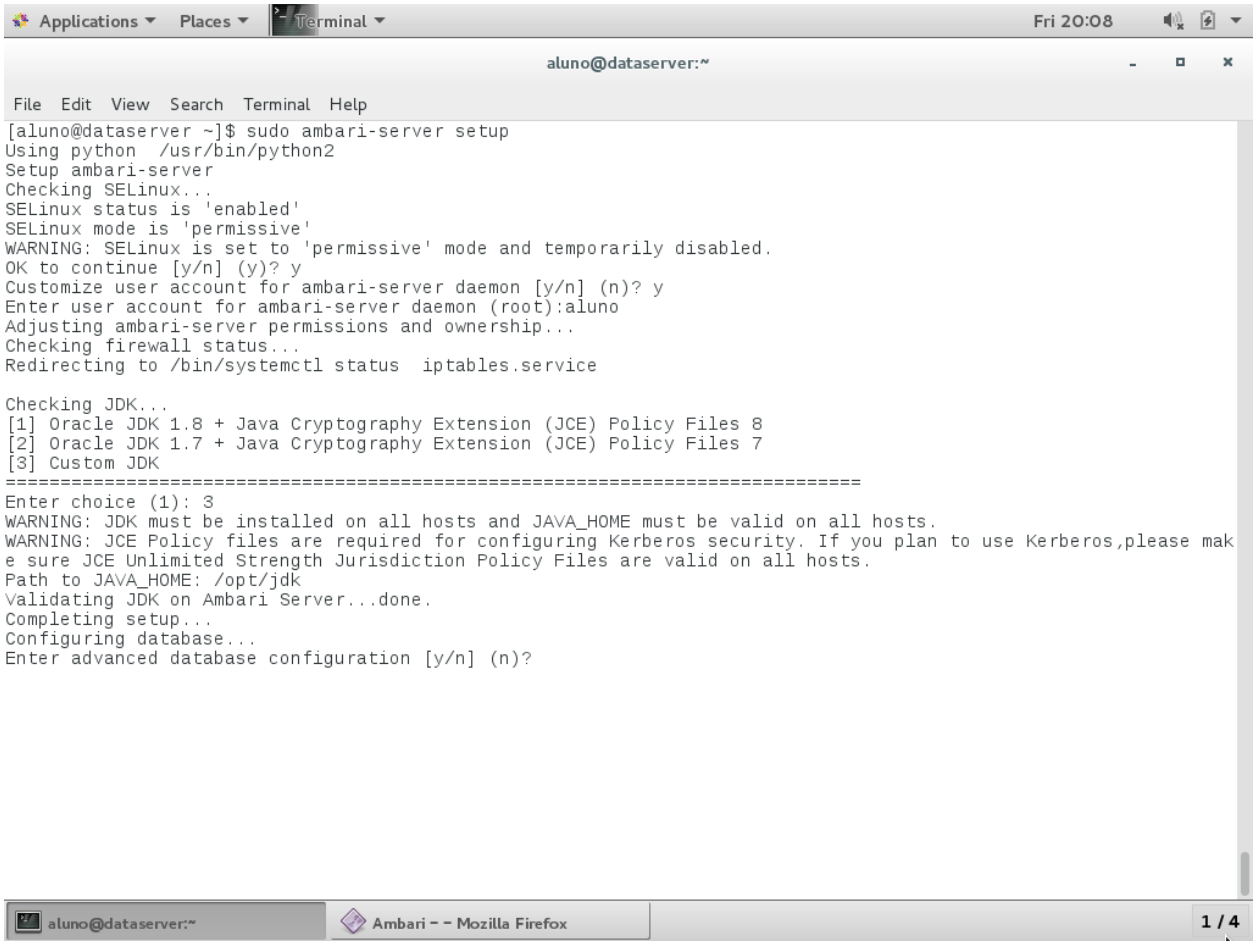


Instalação do Ambari

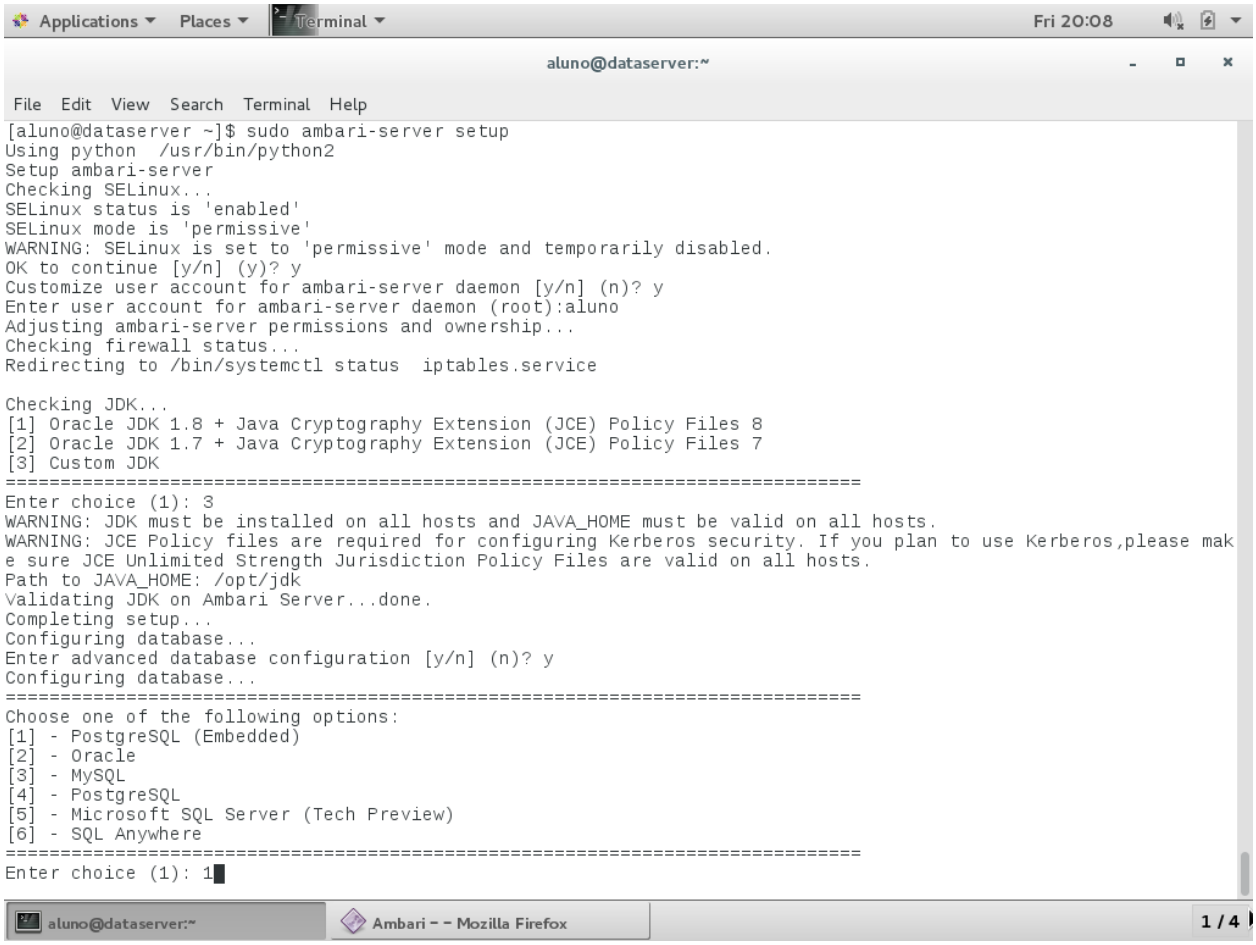
Instalação concluída com sucesso



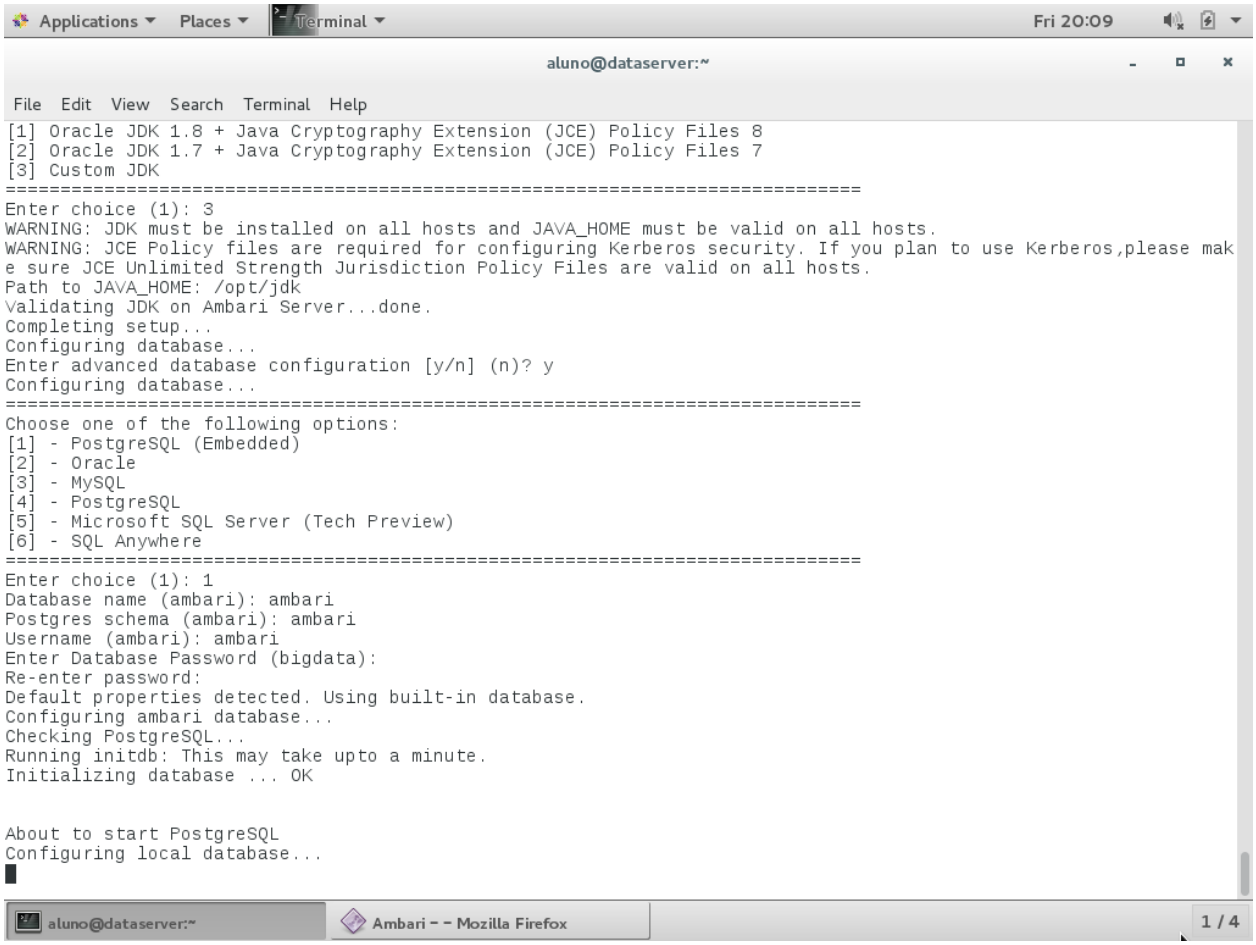
Configuração do Ambari



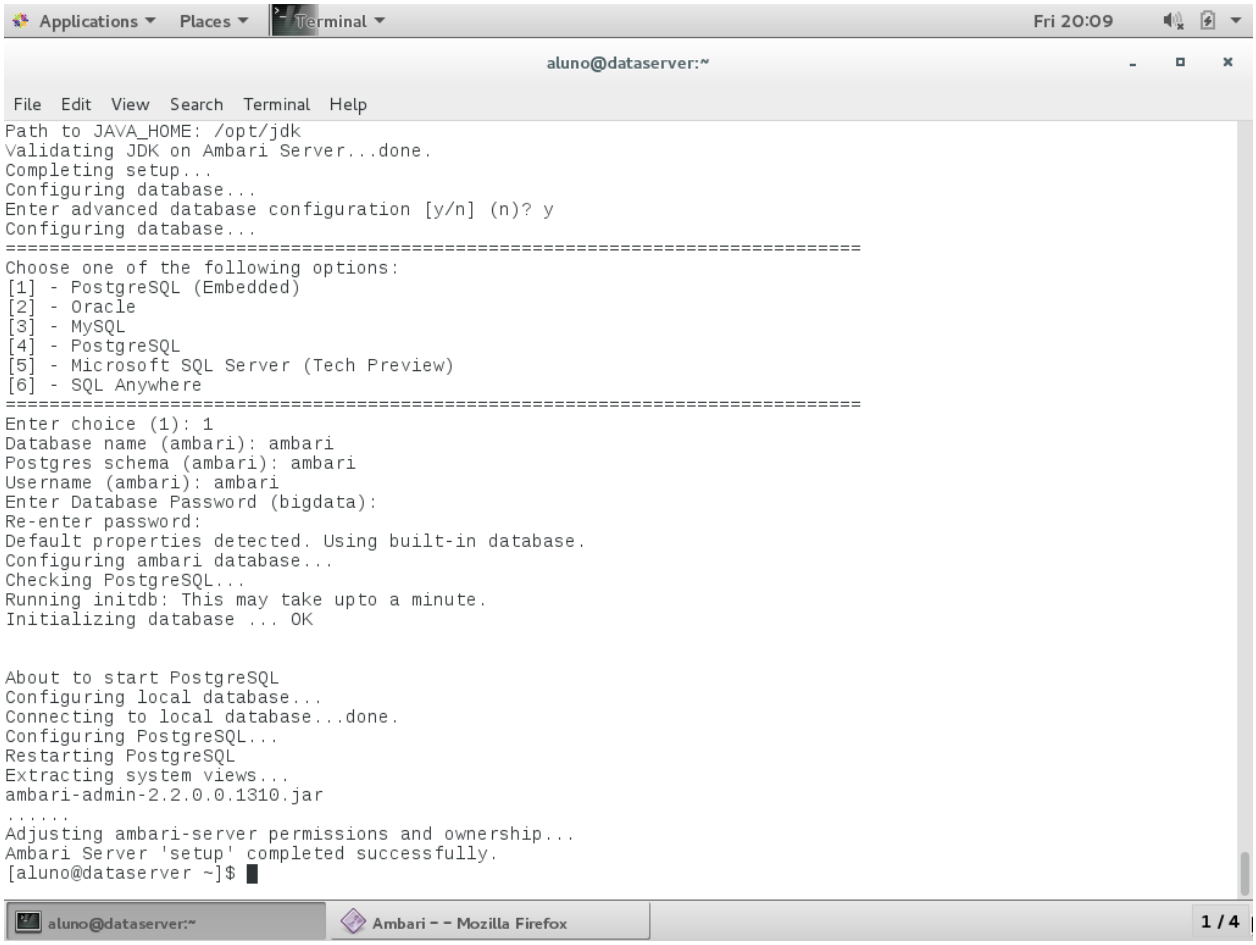
Configuração do Ambari



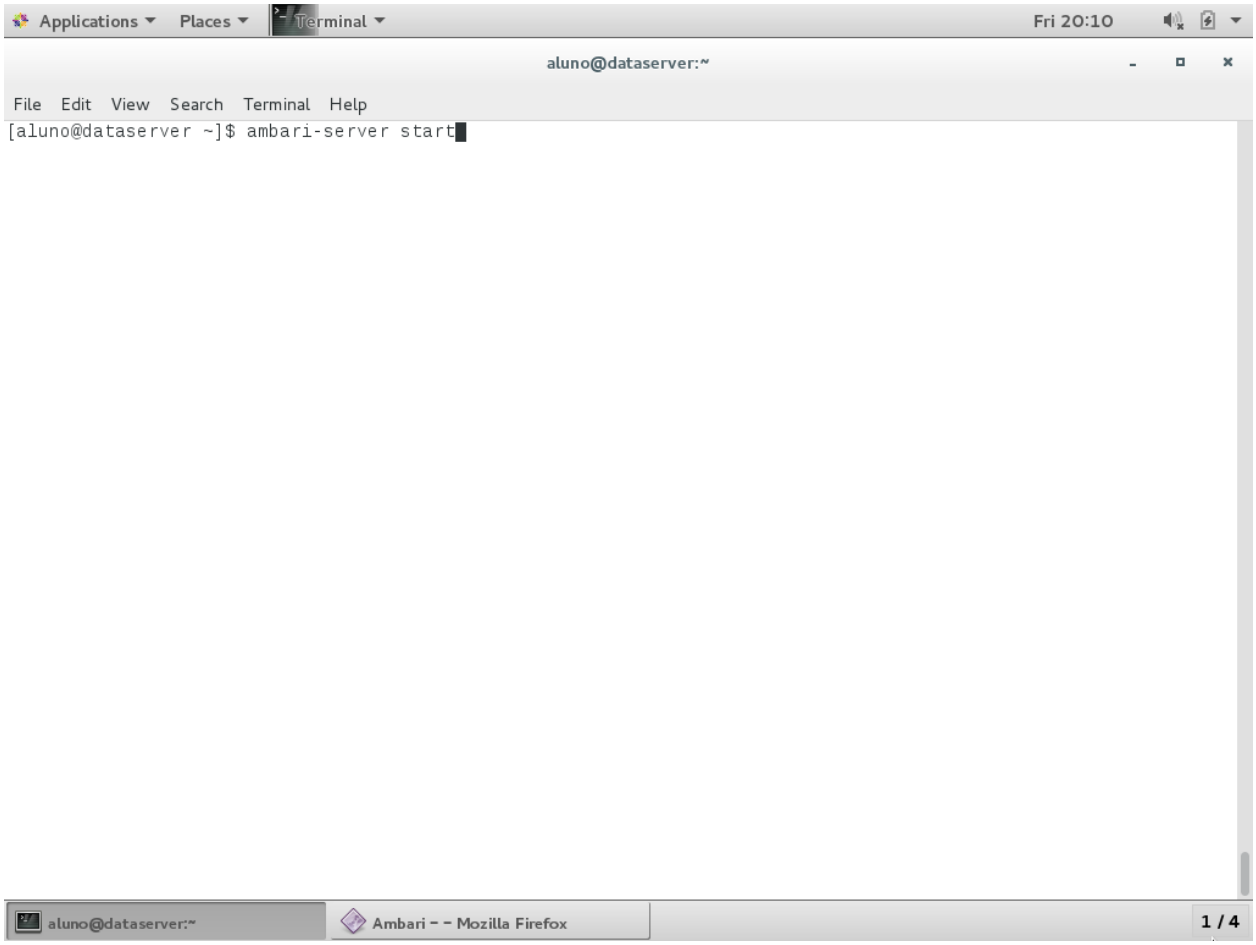
Configuração do Ambari



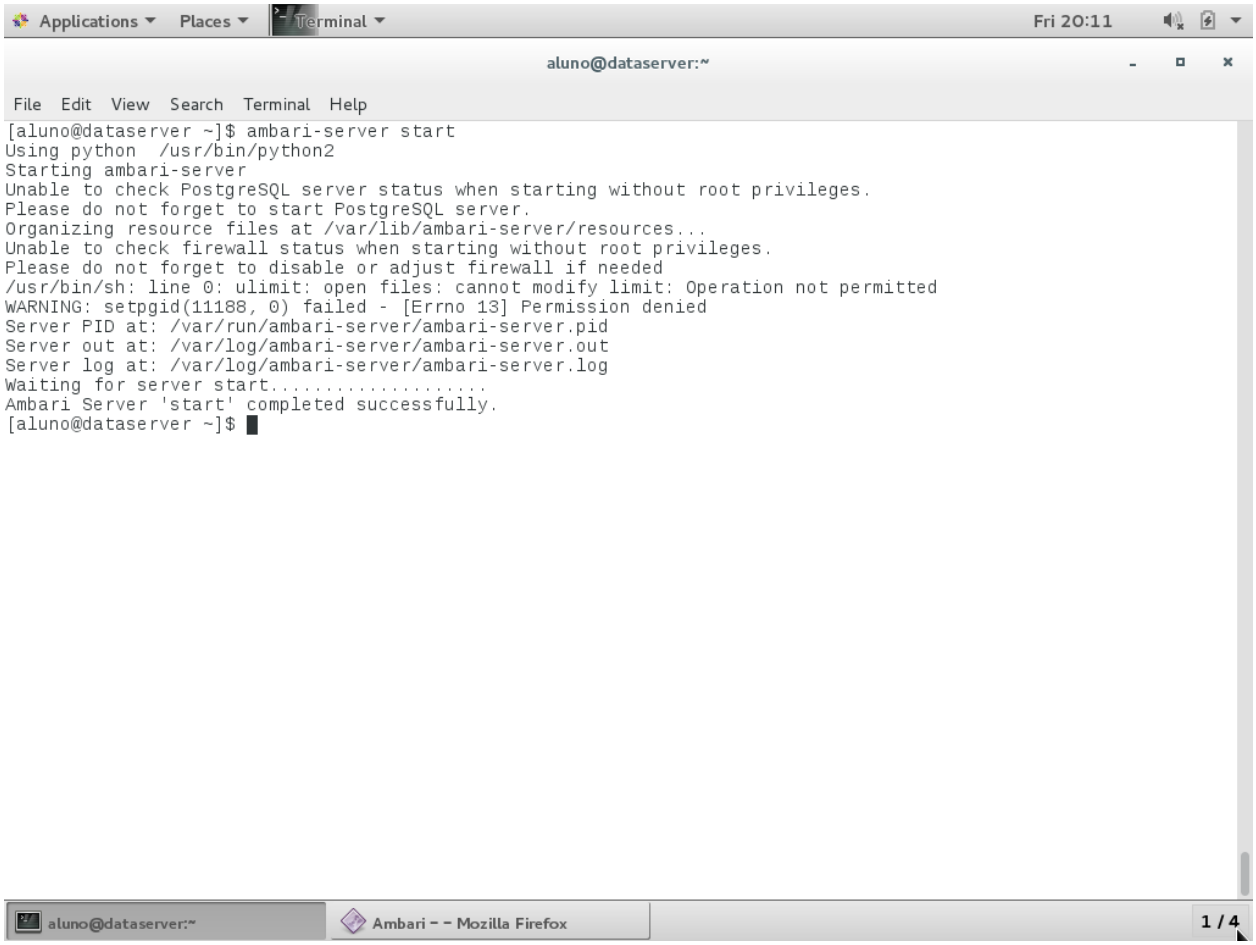
Configuração em andamento



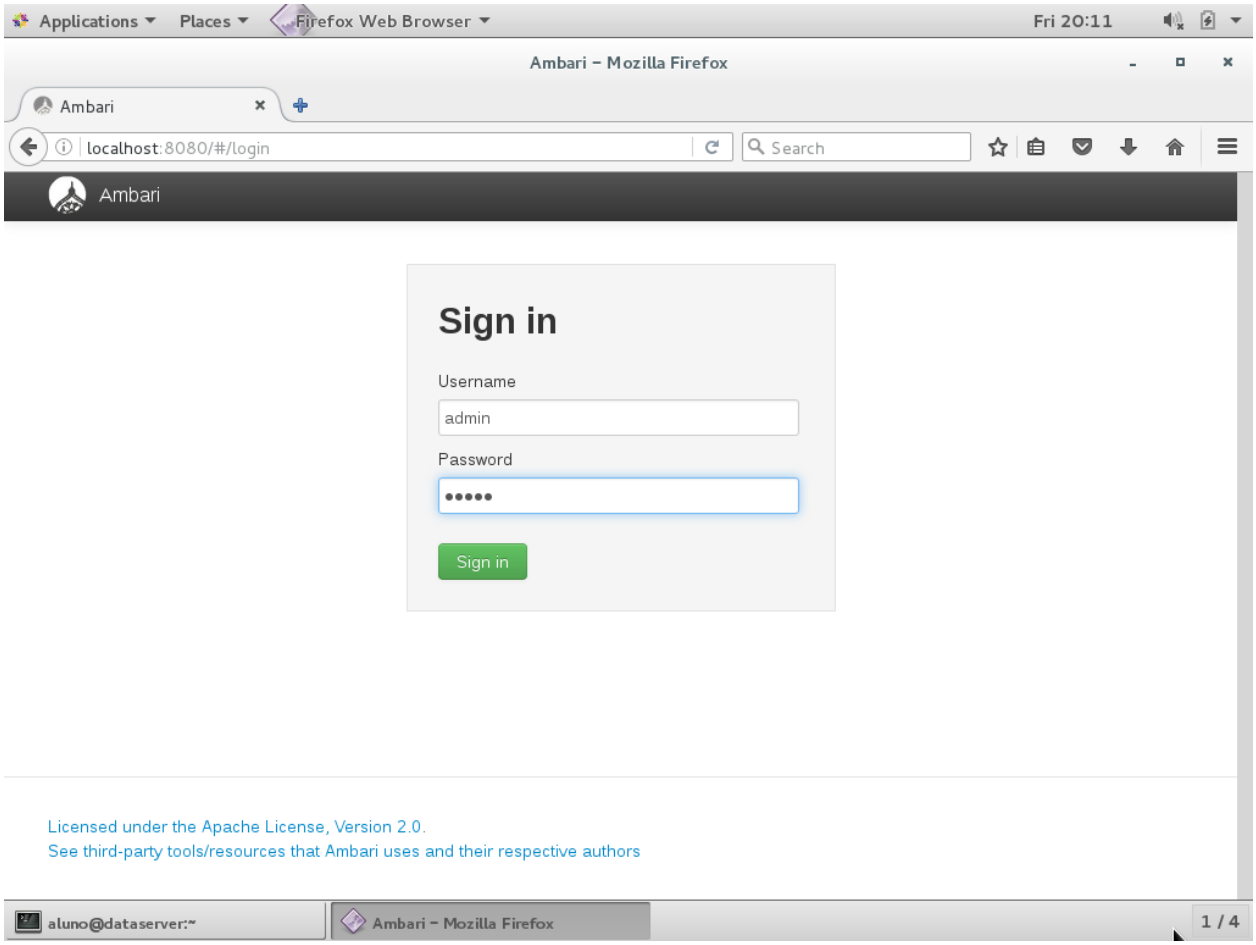
Configuração concluída



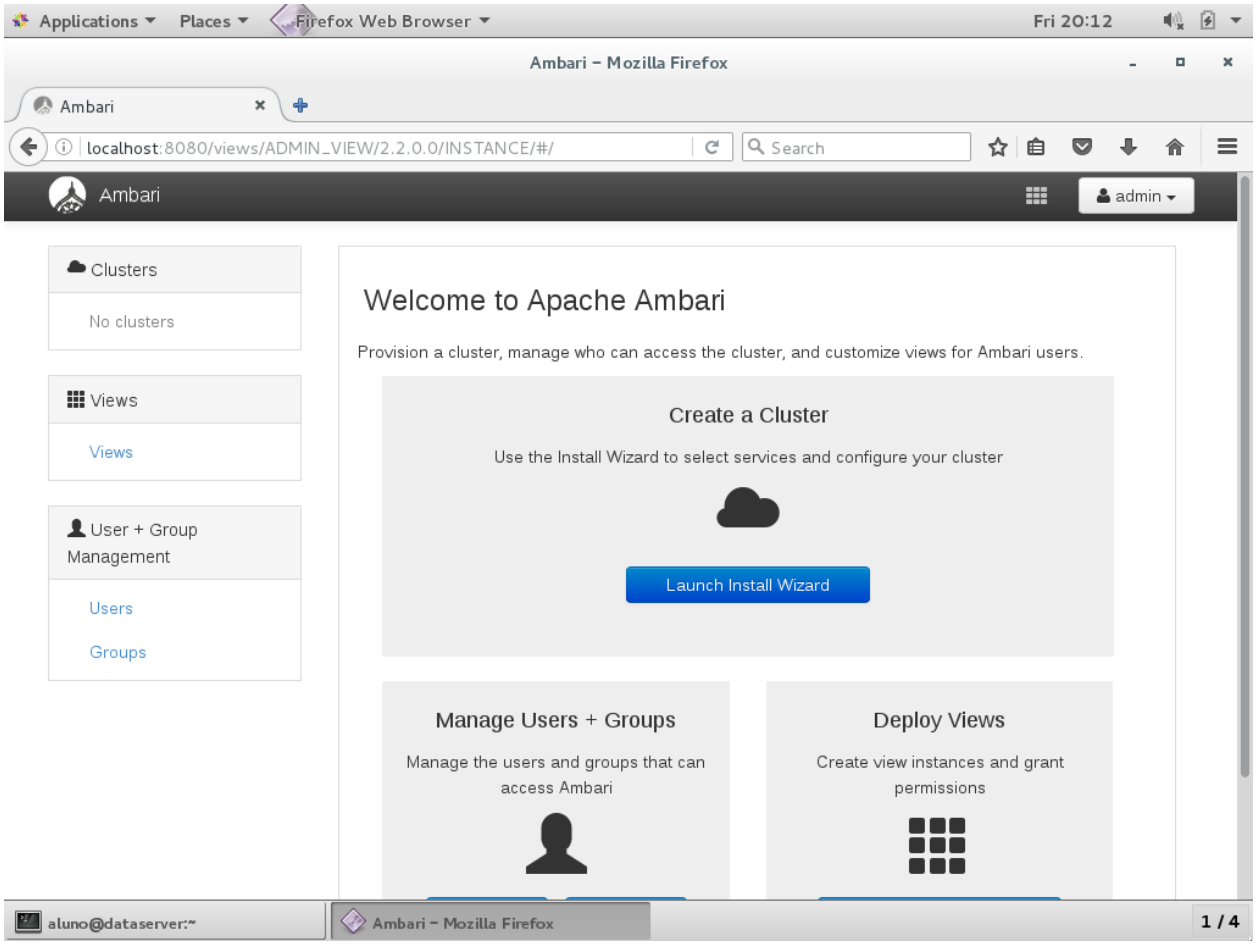
Inicializar o Ambari



Inicializado



Acessar o browser – http://dataserver:8080 - usuário: admin / senha: admin



Pronto para configuração do cluster

Quarto checkpoint:

Clique no meu File – Export Appliance.
Será gerada uma cópia de segurança da sua máquina virtual.

→ VM: DataServer-vFinal.ova (Completa)

Parabéns!

Você tem um ambiente de testes para armazenar e processar Big Data!