



Engenharia de Dados com Hadoop e Spark 3.0

Engenharia de Dados com Hadoop e Spark Versão 3.0

Instalação e Configuração do Ecossistema Hadoop

Sumário

1. Versão.....	4
2. Configuração do Ambiente.....	5
2.1. Criação da Máquina Virtual no VirtualBox.....	6
2.2. Instalação do Sistema Operacional.....	15
2.3. Instalação de Utilitários do Sistema Operacional.....	35
2.4. Instalação do MySQL.....	46
3. Instalação do servidor ssh.....	55
4. Instalação do Java 8.....	65
4.1. Removendo o OpenJDK.....	65
4.2. Instalação do JDK.....	67
5. Instalação e Configuração do Hadoop.....	75
5.1. Criando o usuário hadoop.....	75
5.2. Configuração do ssh sem senha.....	78
5.3. Download e Instalação do Hadoop.....	86
5.3.1. Editando o arquivo hosts.....	86
5.3.2. Download do Hadoop.....	88
5.4. Configuração do Hadoop.....	93
5.4.1. Editar arquivos de configuração do Hadoop.....	93
5.4.2. Formatando o Namenode.....	96
5.4.3. Iniciando o Hadoop.....	97
5.4.4. Iniciando o Yarn.....	99
5.5. Processando Big Data.....	103
6. Instalação e Configuração do Zookeeper.....	121
6.1. Download e Instalação do Zookeeper.....	121
6.2. Configurando do Zookeeper.....	122
7. Instalação e Configuração do HBase.....	134
7.1. Download e Instalação do HBase.....	134
7.2. Configurando o HBase.....	135
8. Instalação e Configuração do Hive.....	146
8.1. Download e Instalação do Hive.....	146
8.2. Configurando o Hive.....	147
9. Instalação e Configuração do Pig.....	158
9.1. Download e Instalação do Pig.....	158
9.2. Configurando do Pig.....	159
10. Instalação e Configuração do Spark.....	166
10.1. Download e Instalação do Spark.....	166



11.	Instalação e Configuração do Sqoop.....	173
11.1.	Download do Sqoop.....	173
11.2.	Configuração do Sqoop.....	174
12.	Instalação e Configuração do Apache Flume.....	182
13.	Instalação e Configuração do Ambari (Opcional).....	189

1. Versão

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

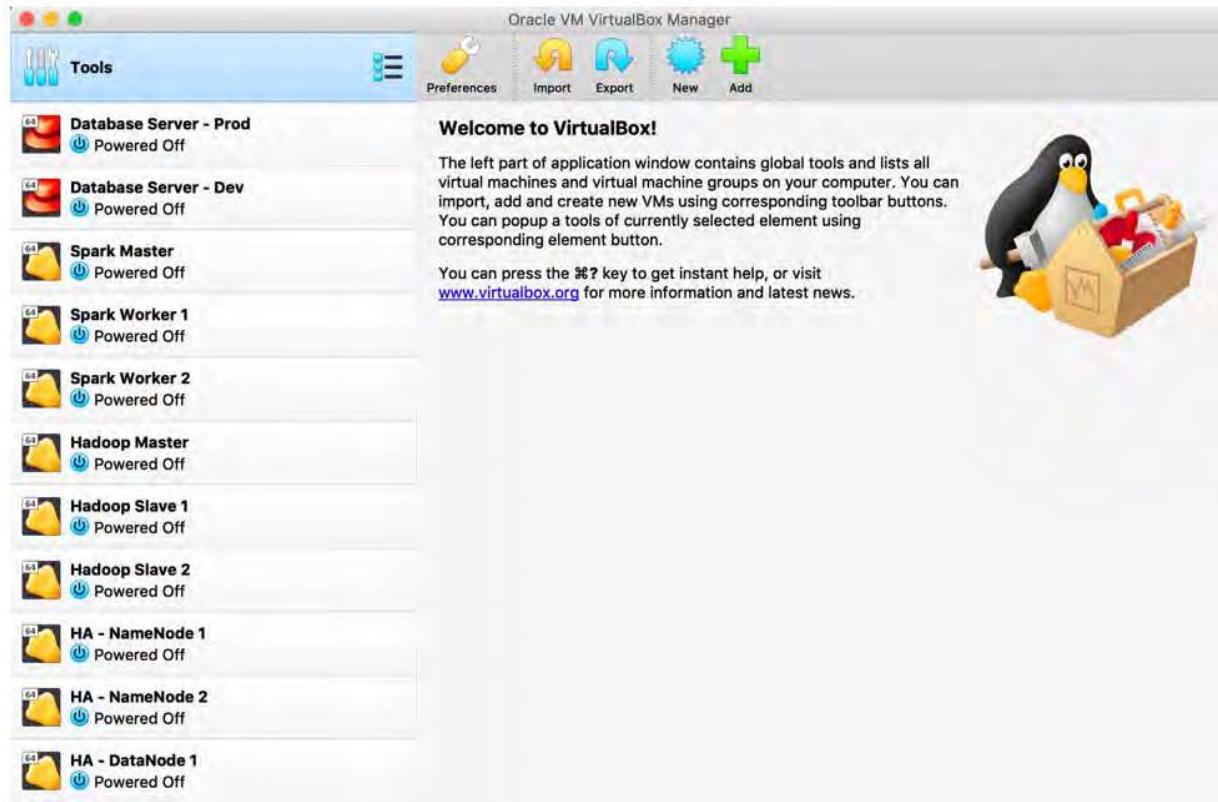
Versão	Ação	Data
1.0	Criação do documento	28/07/2019

2. Configuração do Ambiente

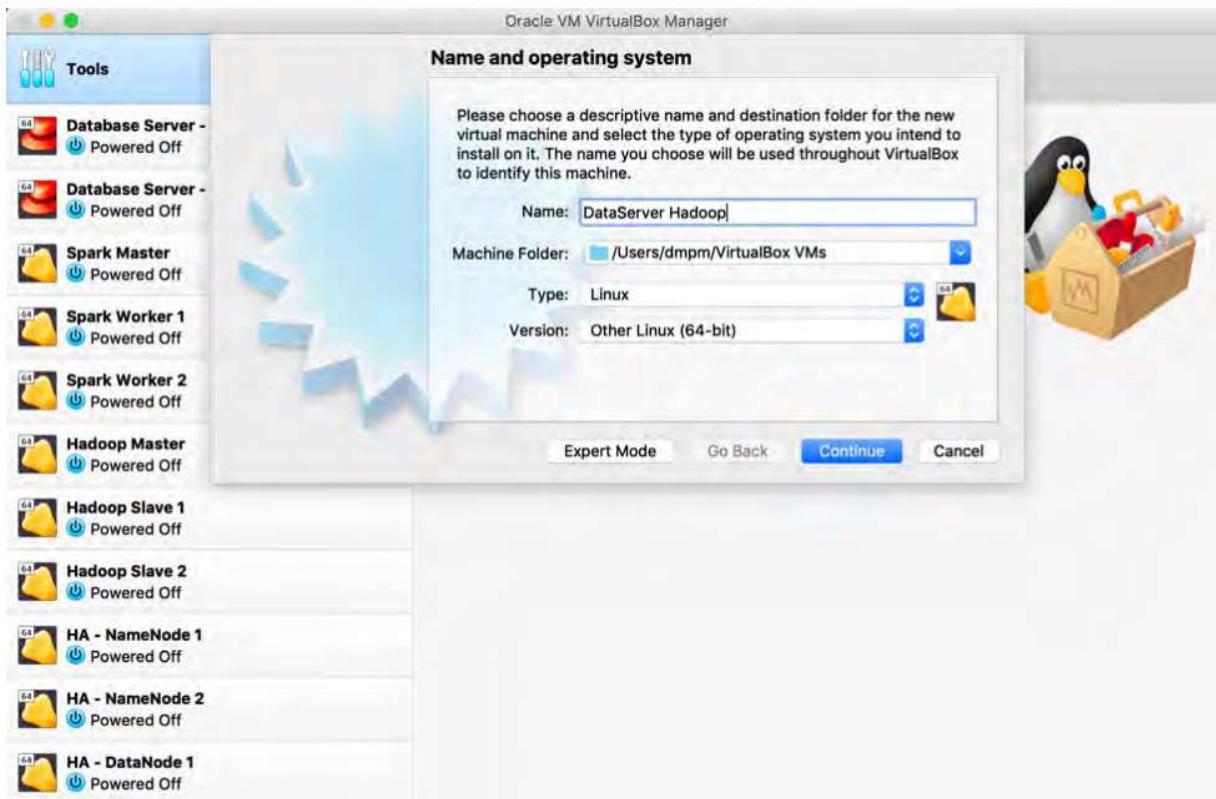
Item	Versão
Virtual Box	6.0.8
Sistema Operacional	CentOS 7.6 (64 bits) ou CentOS 6.8 (32 bits)
Interface Gráfica	Gnome
Firefox Web Browser	60.7
Java	1.8
Apache Hadoop	3.2.0
Apache Zookeeper	3.5.5
Apache Hbase	2.2.0
Apache Hive	3.1.1
Apache Pig	0.17.0
Apache Spark	2.4.3
Apache Sqoop	1.4.7
Apache Flume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

2.1. Criação da Máquina Virtual no VirtualBox

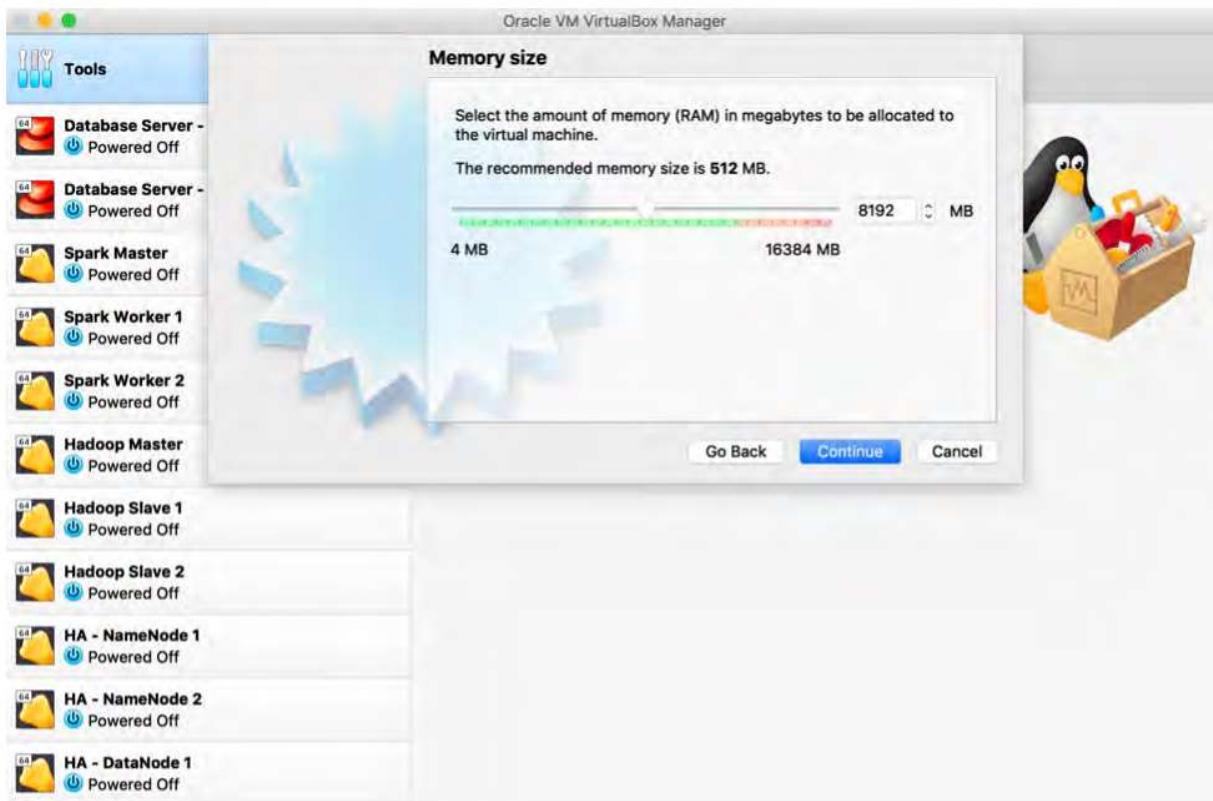
O Oracle VM Virtual Box é gratuito e pode ser baixado em <https://www.virtualbox.org>. Existem versões disponíveis para Windows, MAC, Linux e Solaris. Aqui utilizaremos a versão 6.0 e o tutorial será o mesmo independente do sistema operacional do seu computador. O uso do VirtualBox não é obrigatório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



Abrindo o Gerenciador do Oracle Virtual Box

*Engenharia de Dados com Hadoop e Spark 3.0*

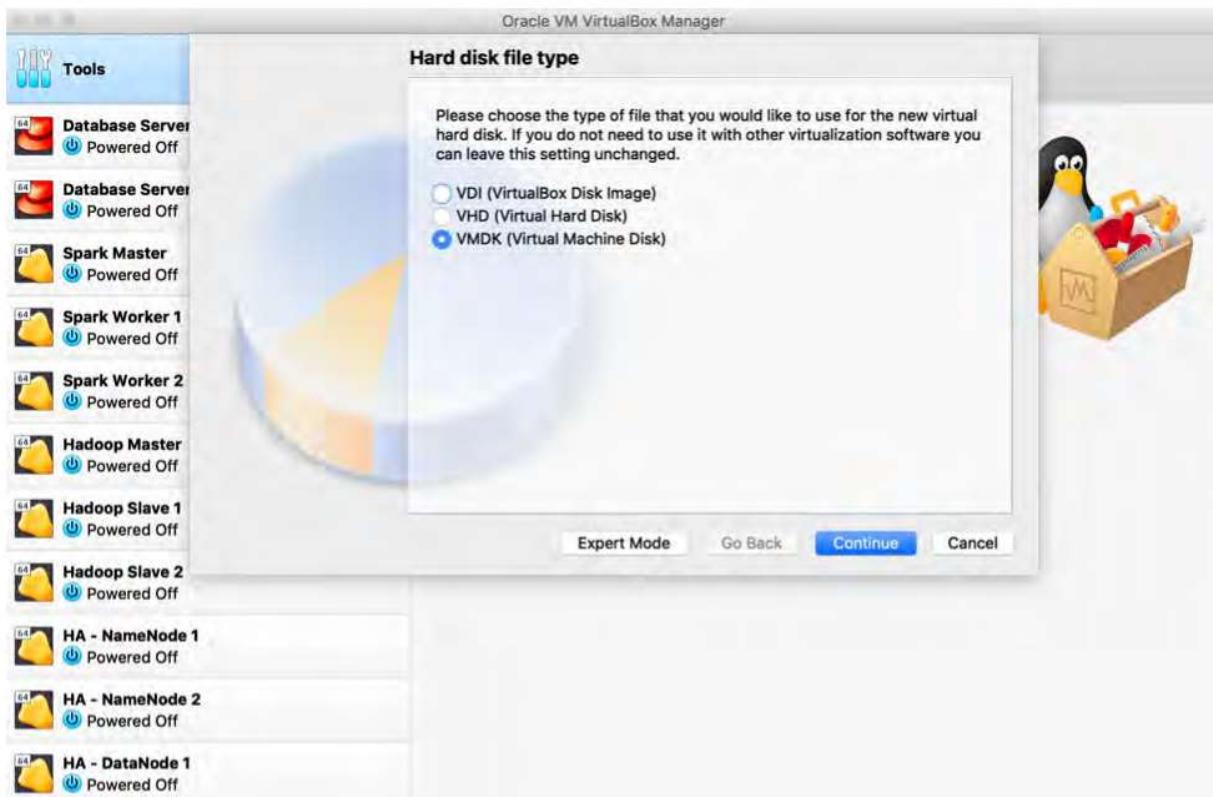
Definindo o nome da máquina virtual e a versão do sistema operacional

*Engenharia de Dados com Hadoop e Spark 3.0*

Configure metade da memória física do seu computador para a VM

*Engenharia de Dados com Hadoop e Spark 3.0*

Criar um novo disco rígido virtual

*Engenharia de Dados com Hadoop e Spark 3.0*

Selecione a opção VMDK

*Engenharia de Dados com Hadoop e Spark 3.0*

O disco deve ser alocado dinamicamente

*Engenharia de Dados com Hadoop e Spark 3.0*

Oracle VM VirtualBox Manager

File location and size

Please type the name of the new virtual hard disk file into the box below or click on the folder icon to select a different folder to create the file in.

/dmpm/VirtualBox VMs/DataServer Hadoop/DataServer Hadoop.vmdk

Select the size of the virtual hard disk in megabytes. This size is the limit on the amount of file data that a virtual machine will be able to store on the hard disk.

64.00 GB

4.00 MB 2.00 TB

Go Back Create Cancel

The screenshot shows the Oracle VM VirtualBox Manager interface. On the left, a sidebar lists various virtual machines: Database Server (x2), Spark Master, Spark Worker 1, Spark Worker 2, Hadoop Master, Hadoop Slave 1, Hadoop Slave 2, HA - NameNode 1, HA - NameNode 2, and HA - DataNode 1. All are currently powered off. The main window displays a configuration dialog for creating a new virtual hard disk. It asks for the file location and size. The location is set to '/dmpm/VirtualBox VMs/DataServer Hadoop/DataServer Hadoop.vmdk'. The size is set to 64.00 GB, with options for 4.00 MB and 2.00 TB available. Below the size slider are 'Go Back', 'Create', and 'Cancel' buttons. A decorative penguin icon is visible on the right side of the dialog.

Selecione 64 GB para o disco virtual

*Engenharia de Dados com Hadoop e Spark 3.0*

Oracle VM VirtualBox Manager

Tools

- DataServer Hadoop** Powered Off
- Database Server - Prod** Powered Off
- Database Server - Dev** Powered Off
- Spark Master** Powered Off
- Spark Worker 1** Powered Off
- Spark Worker 2** Powered Off
- Hadoop Master** Powered Off
- Hadoop Slave 1** Powered Off
- Hadoop Slave 2** Powered Off
- HA - NameNode 1** Powered Off
- HA - NameNode 2** Powered Off

New **Settings** **Start**

General

Name: DataServer Hadoop
Operating System: Other Linux (64-bit)
Settings File Location: /Users/dmpm/VirtualBox VMs/DataServer Hadoop

System

Base Memory: 8192 MB
Boot Order: Floppy, Optical, Hard Disk
Acceleration: VT-x/AMD-V, Nested Paging, PAE/NX, KVM
Paravirtualization

Display

Video Memory: 16 MB
Graphics Controller: VMSVGA
Remote Desktop Server: Disabled
Recording: Disabled

Storage

Controller: IDE
IDE Primary Master: DataServer Hadoop.vmdk (Normal, 64.00 GB)
IDE Secondary Master: [Optical Drive] Empty

Audio

Host Driver: CoreAudio
Controller: ICH AC97

Network

Adapter 1: Intel PRO/1000 MT Desktop (NAT)

USB

USB Controller: OHCI, EHCI
Device Filters: 0 (0 active)

Shared folders

Preview

DataServer Hadoop

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

*Engenharia de Dados com Hadoop e Spark 3.0*

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

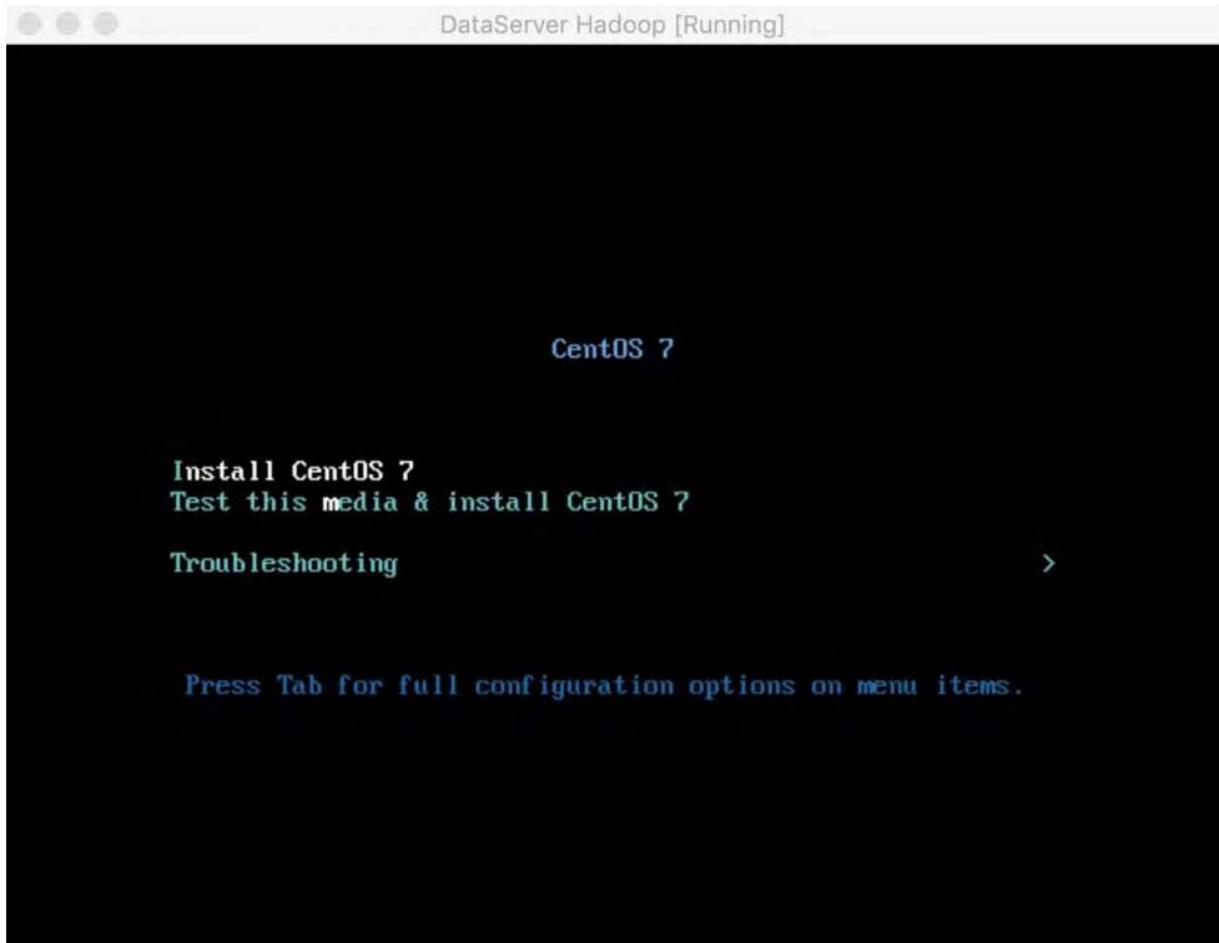
CentOS 64 bits (versão 7):

http://isoredirect.centos.org/centos/7/isos/x86_64/

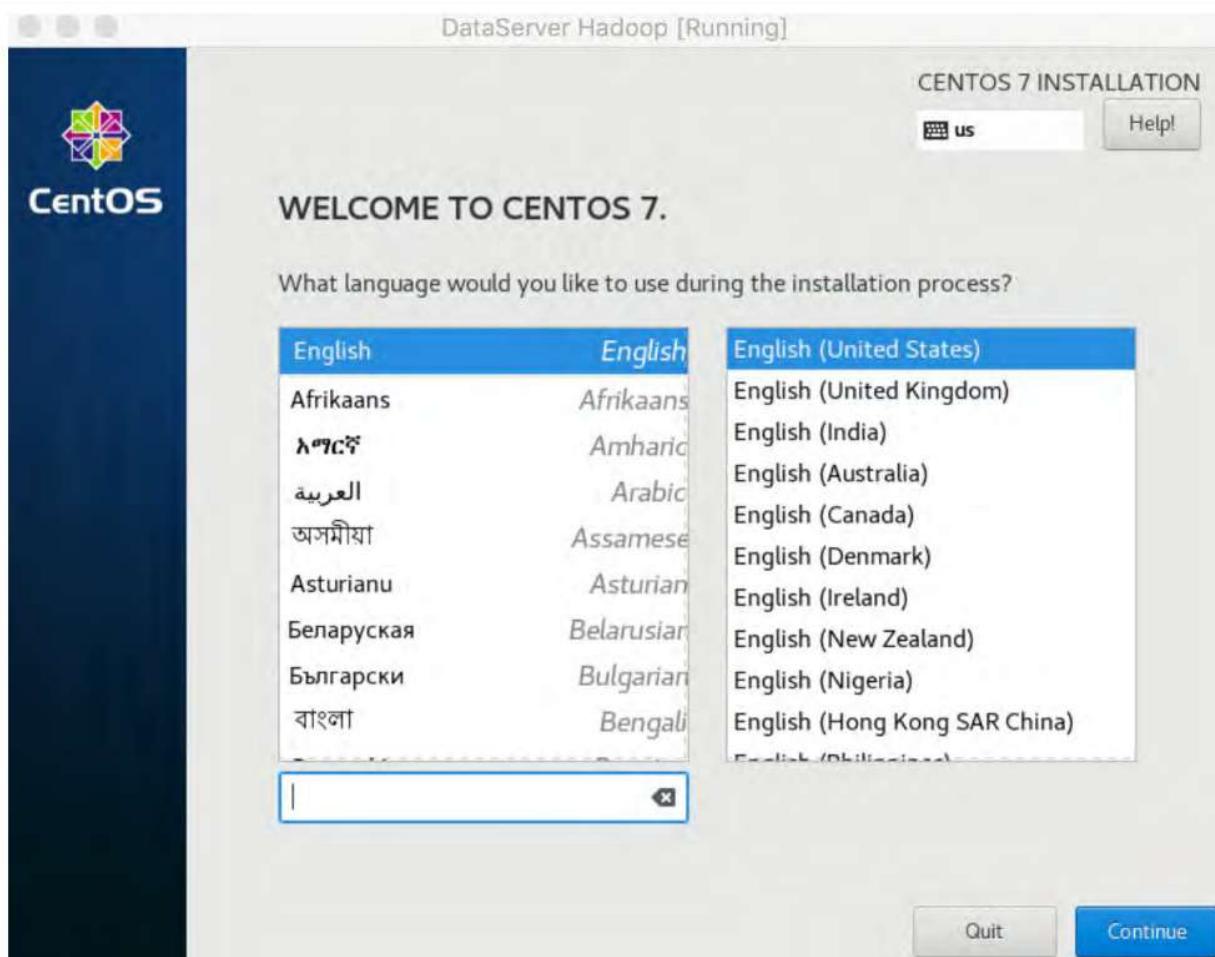
CentOS 32 bits (versão 6.8):

<http://centos.mirror.netelligent.ca/centos/>

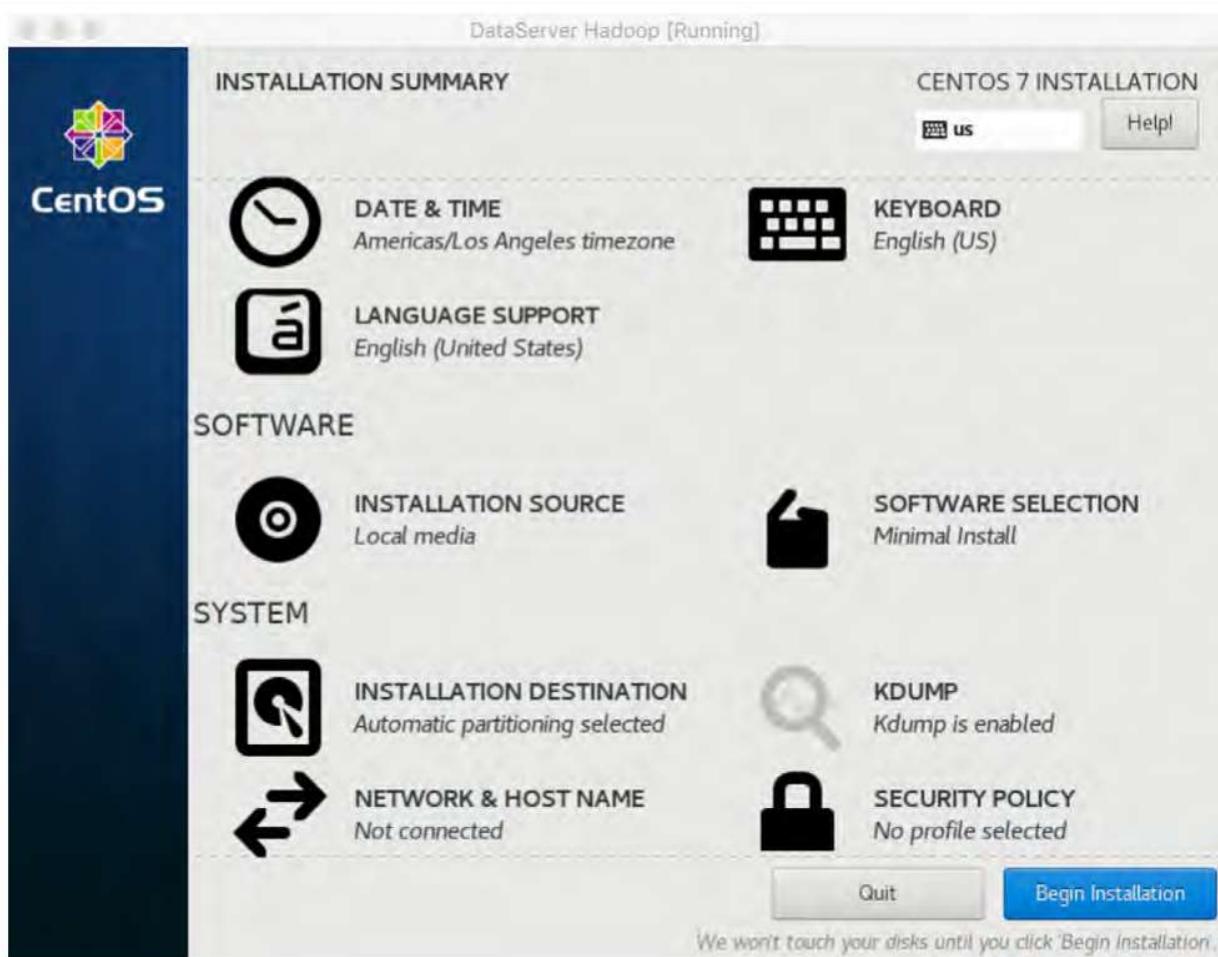
2.2. Instalação do Sistema Operacional



Selecione a opção de Instalação do Sistema Operacional CentOS 7



Seleção do idioma usado na instalação



Opções de configuração

*Engenharia de Dados com Hadoop e Spark 3.0*

DataServer Hadoop [Running]

DATE & TIME

Done

CENTOS 7 INSTALLATION

us Help!

Region: Americas City: Sao Paulo Network Time OFF

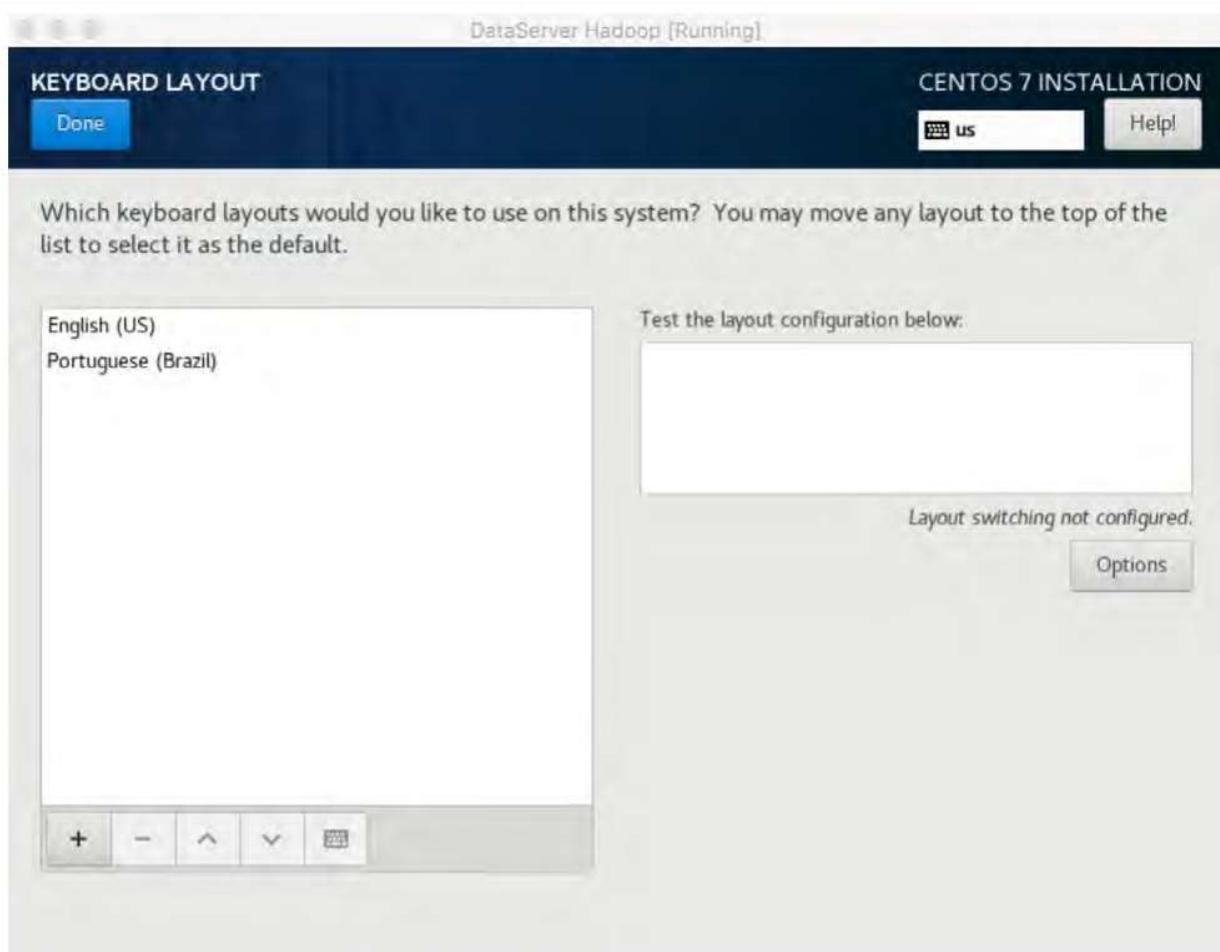
A world map with green highlighted regions representing different time zones. A red dot marks the location of Sao Paulo, Brazil, which is located in the green-highlighted South America time zone.

23:08 PM 24-hour
 AM/PM

06 / 21 / 2019

⚠ You need to set up networking first if you want to use NTP

Timezone

*Engenharia de Dados com Hadoop e Spark 3.0*

Layout do teclado

LANGUAGE SUPPORT CENTOS 7 INSTALLATION

Done us Help!

Select additional language support to be installed:

বাংলা	Bengali
Bosanski	Bosnian
Català	Catalan
Čeština	Czech
Cymraeg	Welsh
Dansk	Danish
Deutsch	German
Ελληνικά	Greek
✓ English	English ➤
Español	Spanish
Eesti	Estonian
Euskara	Basque
فارسی	Persian

Type here to search:

English (United States)

English (United Kingdom)

English (India)

English (Australia)

English (Canada)

English (Denmark)

English (Ireland)

English (New Zealand)

English (Nigeria)

English (Hong Kong SAR China)

English (Philippines)

English (Singapore)

English (South Africa)

English (Zambia)

Idioma do sistema operacional



SECURITY POLICY

Finalizado

INSTALAÇÃO DO CENTOS 7

br Help

Change content Apply security policy: ON

Choose profile below:

Default
The implicit XCCDF profile. Usually, the default contains no rules.

Standard System Security Profile
This profile contains rules to ensure standard security base of CentOS Linux 7 system.

Draft PCI-DSS v3 Control Baseline for CentOS Linux 7
This is a *draft* profile for PCI-DSS v3

CentOS Profile for Cloud Providers (CPCP)
This is a *draft* SCAP profile for CentOS Cloud Providers

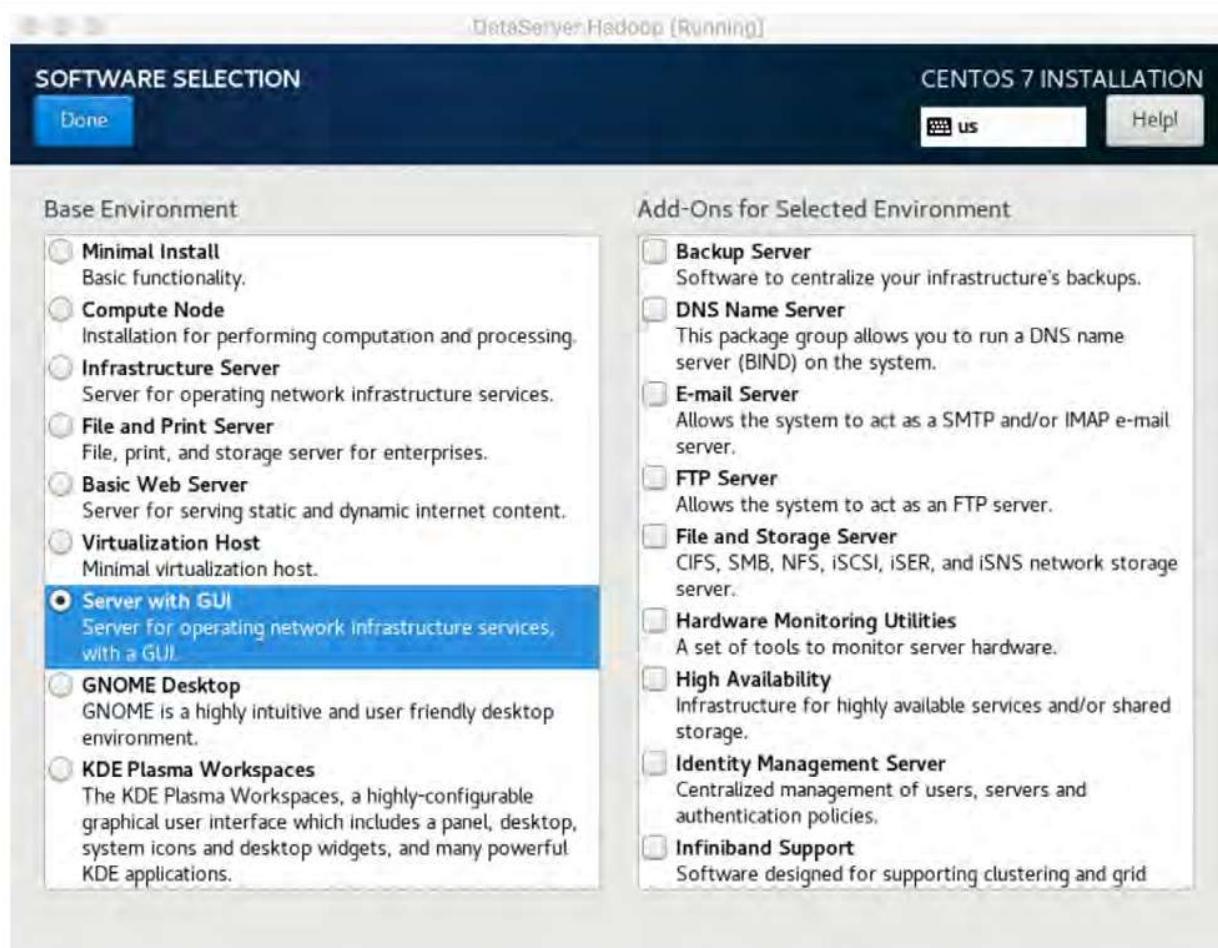
Common Profile for General-Purpose Systems
This profile contains items common to general-purpose desktop and server installations.

Pre-release Draft STIG for CentOS Linux 7 Server
This profile is being developed under the DoD consensus model to become a STIG in coordination with DISA FSO.

Select profile

Changes that were done or need to be done:
Não há regras para a fase de pré-instalação

Política de segurança padrão



The screenshot shows the 'Software Selection' step of the CentOS 7 Installation process. The title bar says 'DataServer-Hadoop (Running)' and 'CENTOS 7 INSTALLATION'. The 'Done' button is visible in the top left, and 'Help!' in the top right.

Base Environment

- Minimal Install**
Basic functionality.
- Compute Node**
Installation for performing computation and processing.
- Infrastructure Server**
Server for operating network infrastructure services.
- File and Print Server**
File, print, and storage server for enterprises.
- Basic Web Server**
Server for serving static and dynamic internet content.
- Virtualization Host**
Minimal virtualization host.
- Server with GUI**
Server for operating network infrastructure services, with a GUI.
- GNOME Desktop**
GNOME is a highly intuitive and user friendly desktop environment.
- KDE Plasma Workspaces**
The KDE Plasma Workspaces, a highly-configurable graphical user interface which includes a panel, desktop, system icons and desktop widgets, and many powerful KDE applications.

Add-Ons for Selected Environment

- Backup Server**
Software to centralize your infrastructure's backups.
- DNS Name Server**
This package group allows you to run a DNS name server (BIND) on the system.
- E-mail Server**
Allows the system to act as a SMTP and/or IMAP e-mail server.
- FTP Server**
Allows the system to act as an FTP server.
- File and Storage Server**
CIFS, SMB, NFS, iSCSI, iSER, and iSNS network storage server.
- Hardware Monitoring Utilities**
A set of tools to monitor server hardware.
- High Availability**
Infrastructure for highly available services and/or shared storage.
- Identity Management Server**
Centralized management of users, servers and authentication policies.
- Infiniband Support**
Software designed for supporting clustering and grid

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



Engenharia de Dados com Hadoop e Spark 3.0

DataServer Hadoop [Running]

NETWORK & HOST NAME

Done

CENTOS 7 INSTALLATION

us Help!

Ethernet (enp0s3)
Intel Corporation 82540EM Gigabit Ethernet Controller

Connected

ON

Hardware Address 08:00:27:99:4F:56

Speed 1000 Mb/s

IP Address 10.0.2.15

Subnet Mask 255.255.255.0

Default Route 10.0.2.2

DNS 192.168.1.1

Configure...

Host name: Apply

Current host name: dataserver

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



The screenshot shows the 'User Configuration' step of the CentOS 7 installation process. On the left, there's a vertical blue sidebar with the 'CentOS' logo. The main window has a light gray header with 'CONFIGURAÇÃO' on the left and 'INSTALAÇÃO DO CENTOS 7' on the right, with a 'us' button and a 'Help!' link. Below the header, the title 'CONFIGURAÇÕES DE USUÁRIO' is centered. Two sections are shown: 'SENHA RAIZ' (Root Password) with the note 'A senha root não foi determinada' (The root password was not determined), and 'CRIAÇÃO DE USUÁRIO' (User Creation) with the note 'Nenhum usuário será criado' (No user will be created). At the bottom, a progress bar indicates 'Criando xfs em /dev/mapper/centos_dataserver-root'. A banner for 'CentOS Virtualization SIG' is visible, along with a warning message: 'Por favor complete os itens marcados com esse ícone antes de seguir para o próximo passo.' (Please complete the items marked with this icon before proceeding to the next step.).

Definir senha do root e criar um novo usuário



DataServer Hadoop [Running]

ROOT PASSWORD

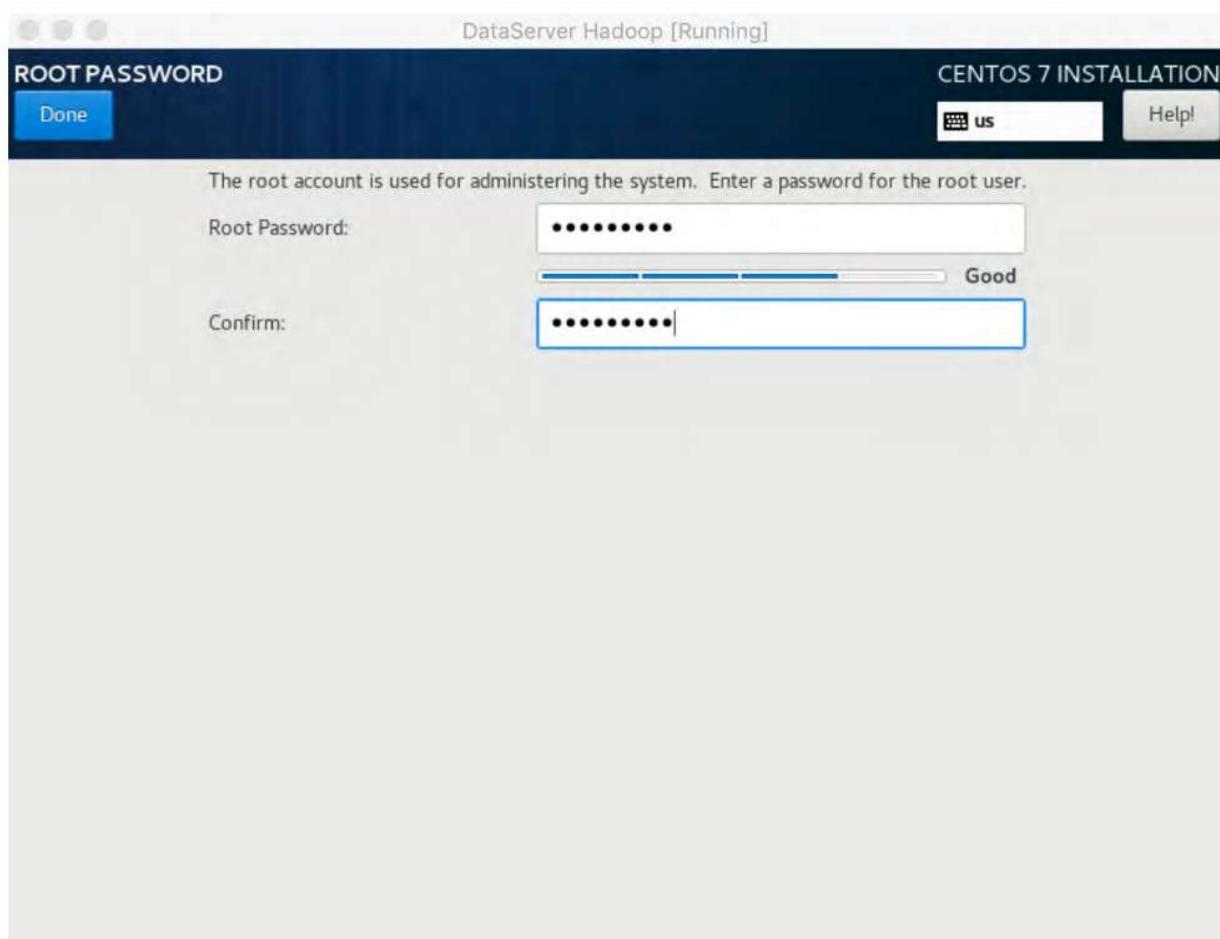
CENTOS 7 INSTALLATION

Done us Help!

The root account is used for administering the system. Enter a password for the root user.

Root Password: ······ Good

Confirm: ······|



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



DataServer Hadoop [Running]

CENTOS 7 INSTALLATION

Done

us Help!

CREATE USER

Full name: Aluno

User name: aluno

Tip: Keep your user name shorter than 32 characters and do not use spaces.

Make this user administrator

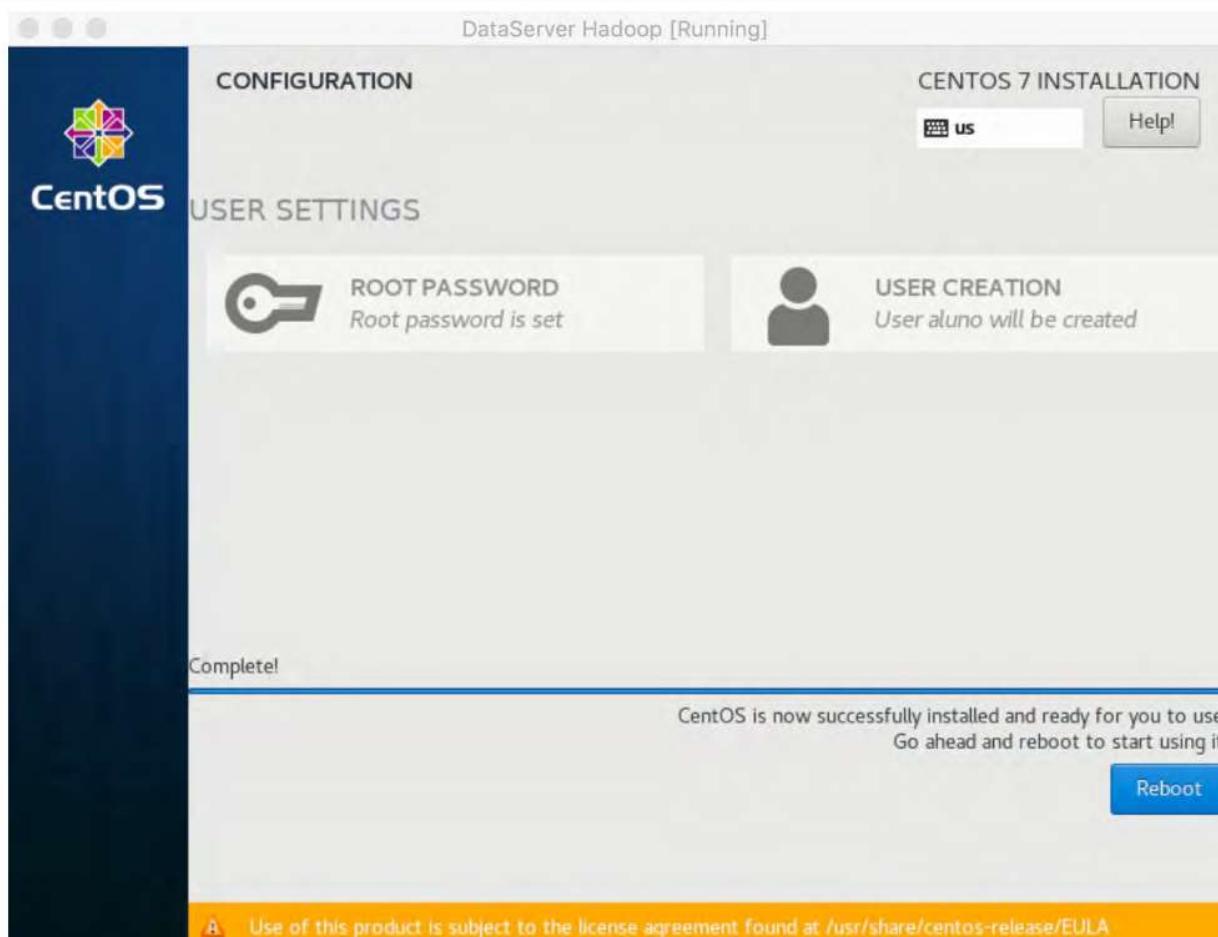
Require a password to use this account

Password: ······ Good

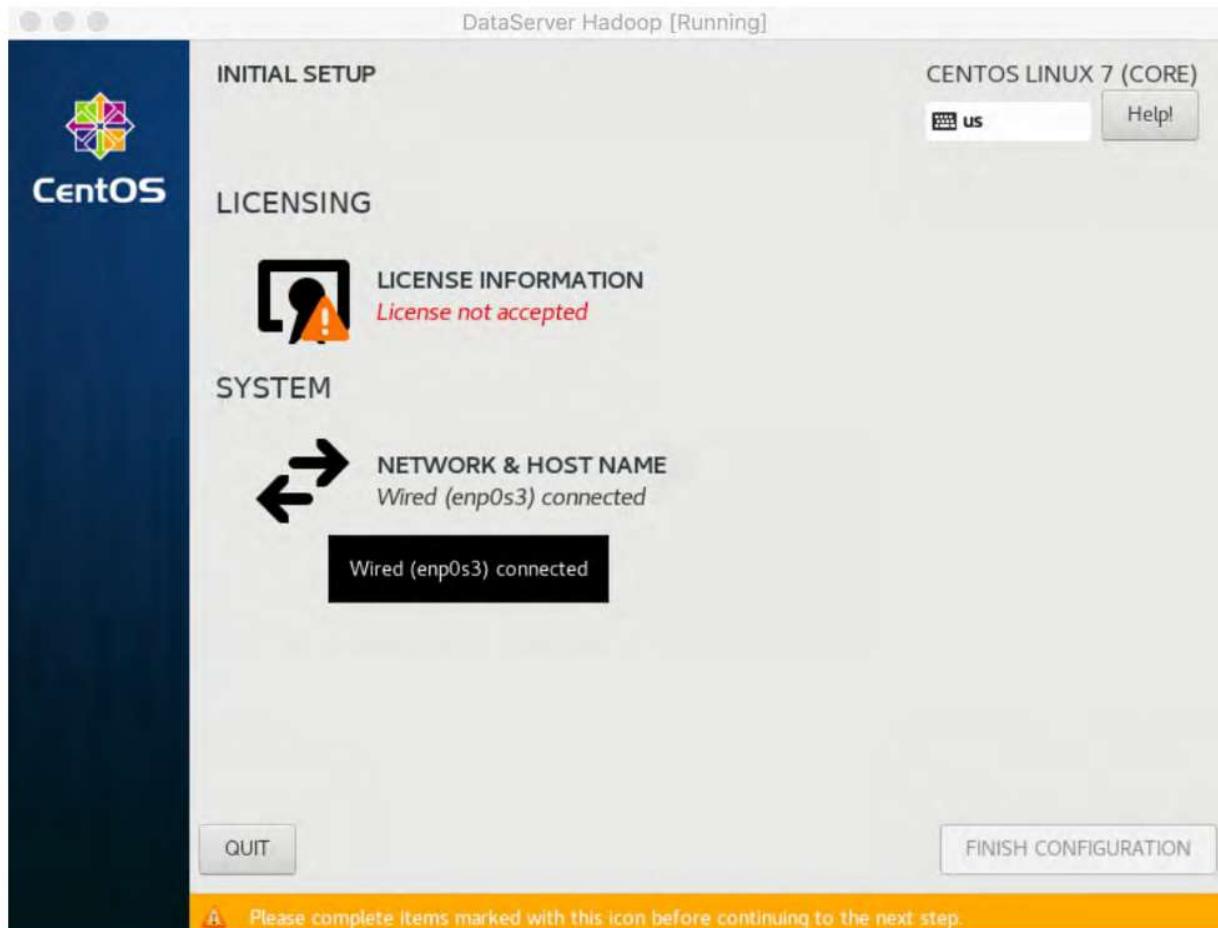
Confirm password: ······|

Advanced...

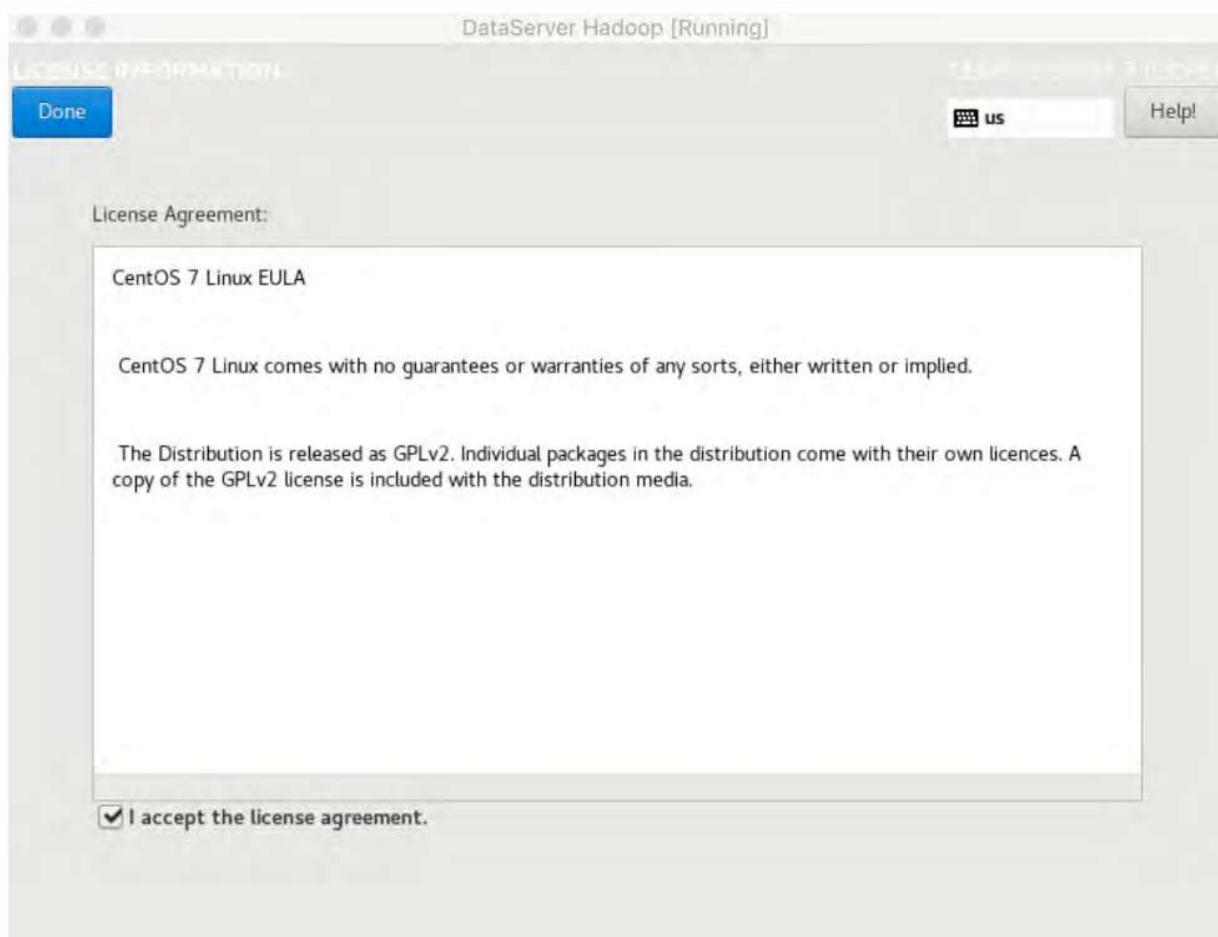
Criação de um usuário – Aluno
(username: aluno, senha: **dsahadoop**)



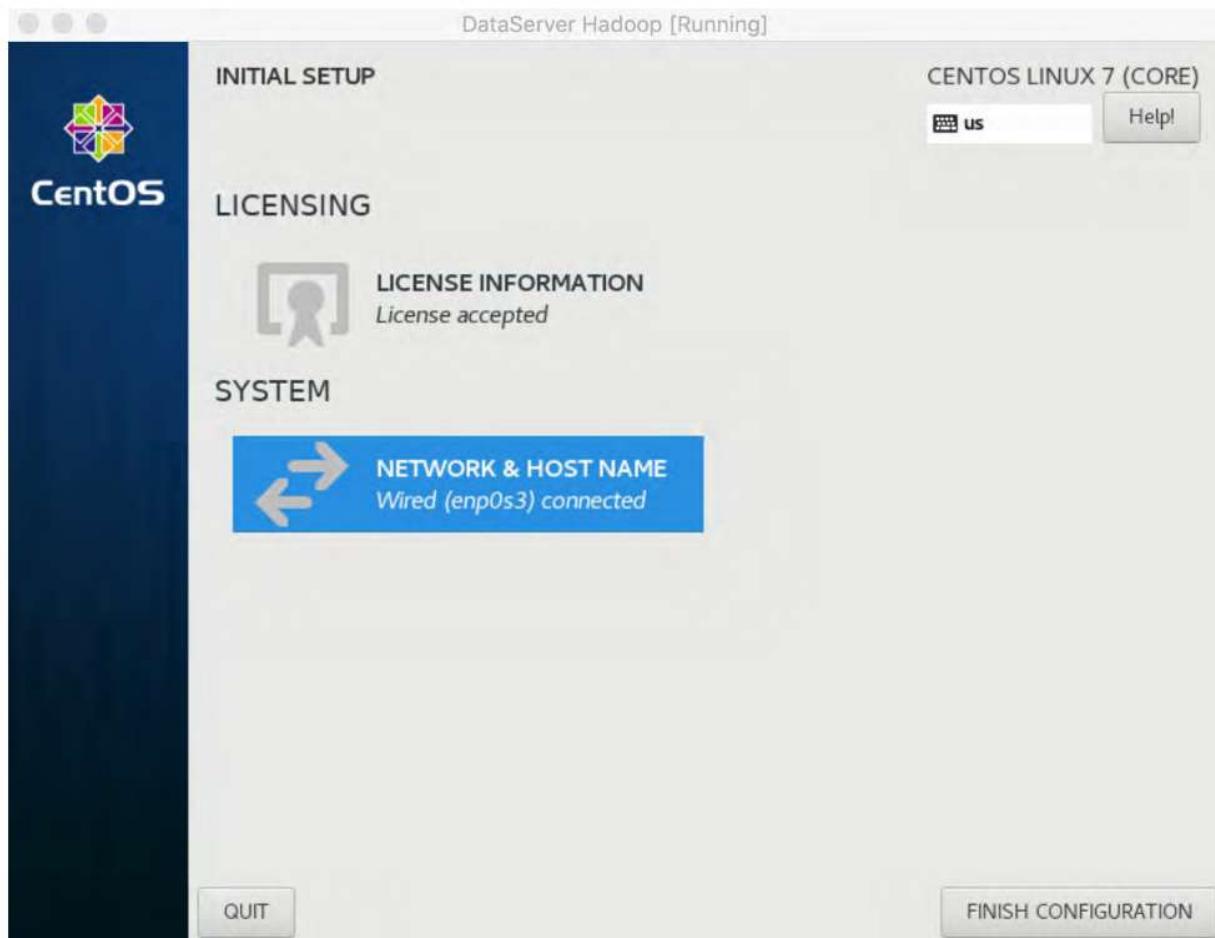
Conclusão da instalação. Clique no botão Reboot.



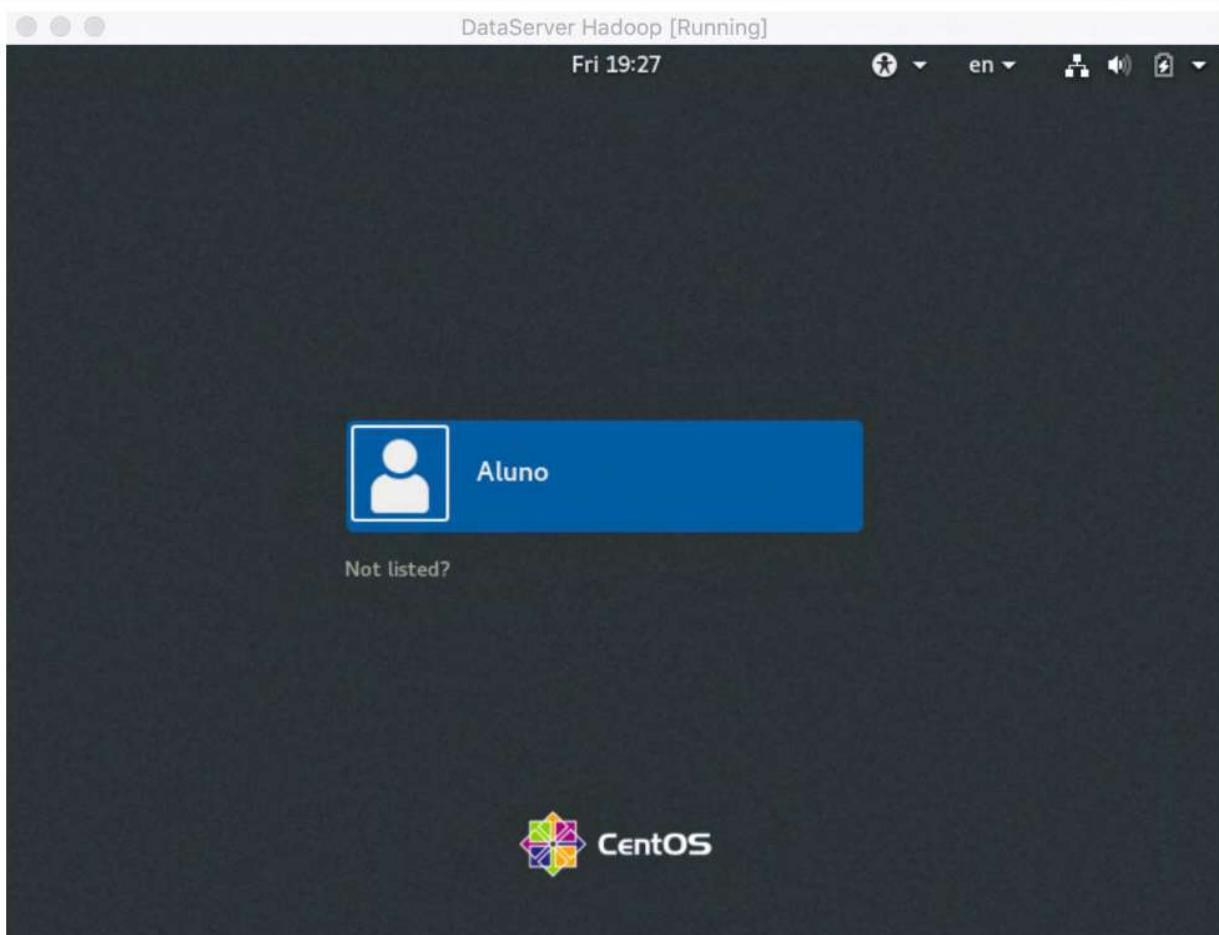
Clique em License Information para aceitar os termos de uso

*Engenharia de Dados com Hadoop e Spark 3.0*

Marque a caixa e pressione Done.

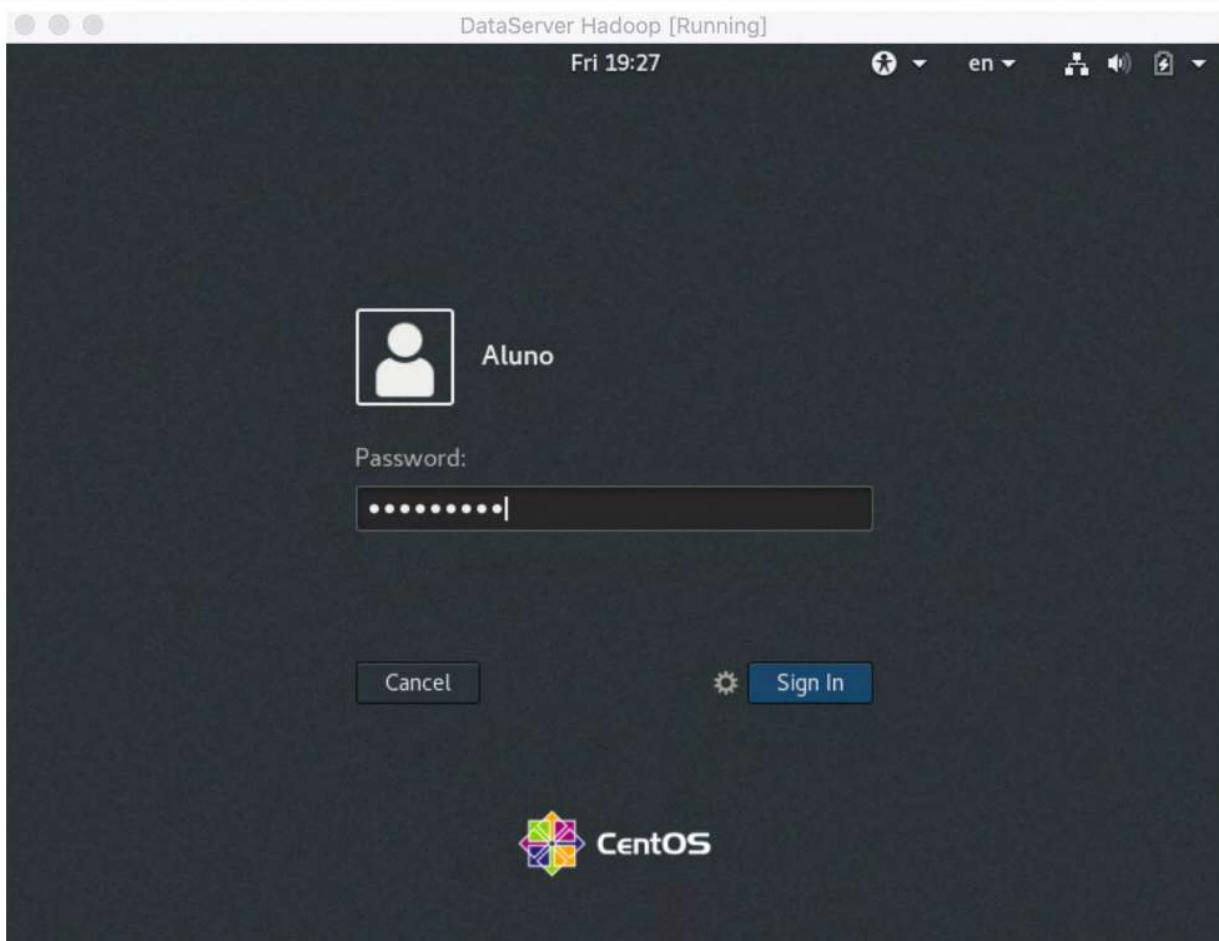


Se necessário, revise se o nome da máquina está correto e se a rede está ativada e pressione
“Finish Configuration”

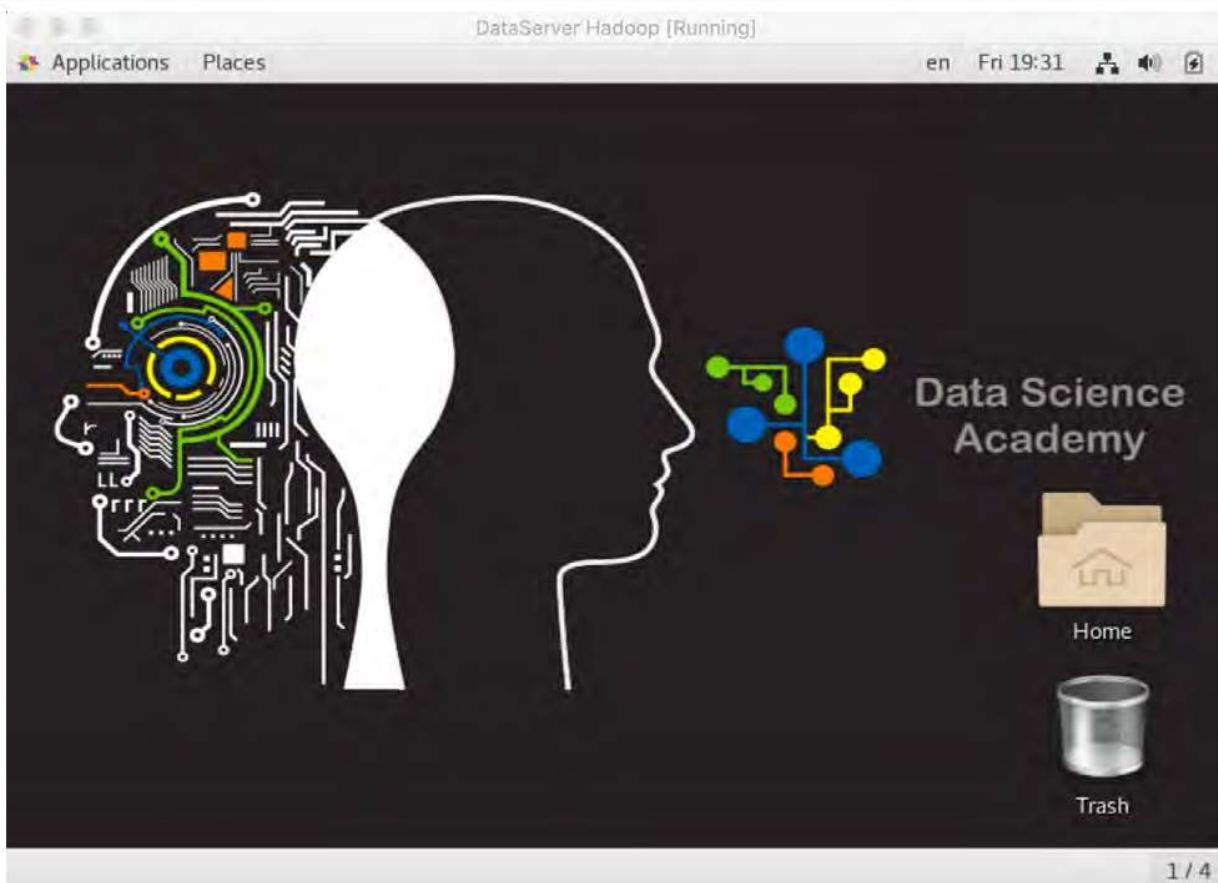
*Engenharia de Dados com Hadoop e Spark 3.0*



Data Science Academy phelipe.utsemprboni@outlook.com 5c8a62005e4cde1acb8b45a3

Engenharia de Dados com Hadoop e Spark 3.0

Usuário/Senha (dsahadoop)



Instalação concluída com sucesso

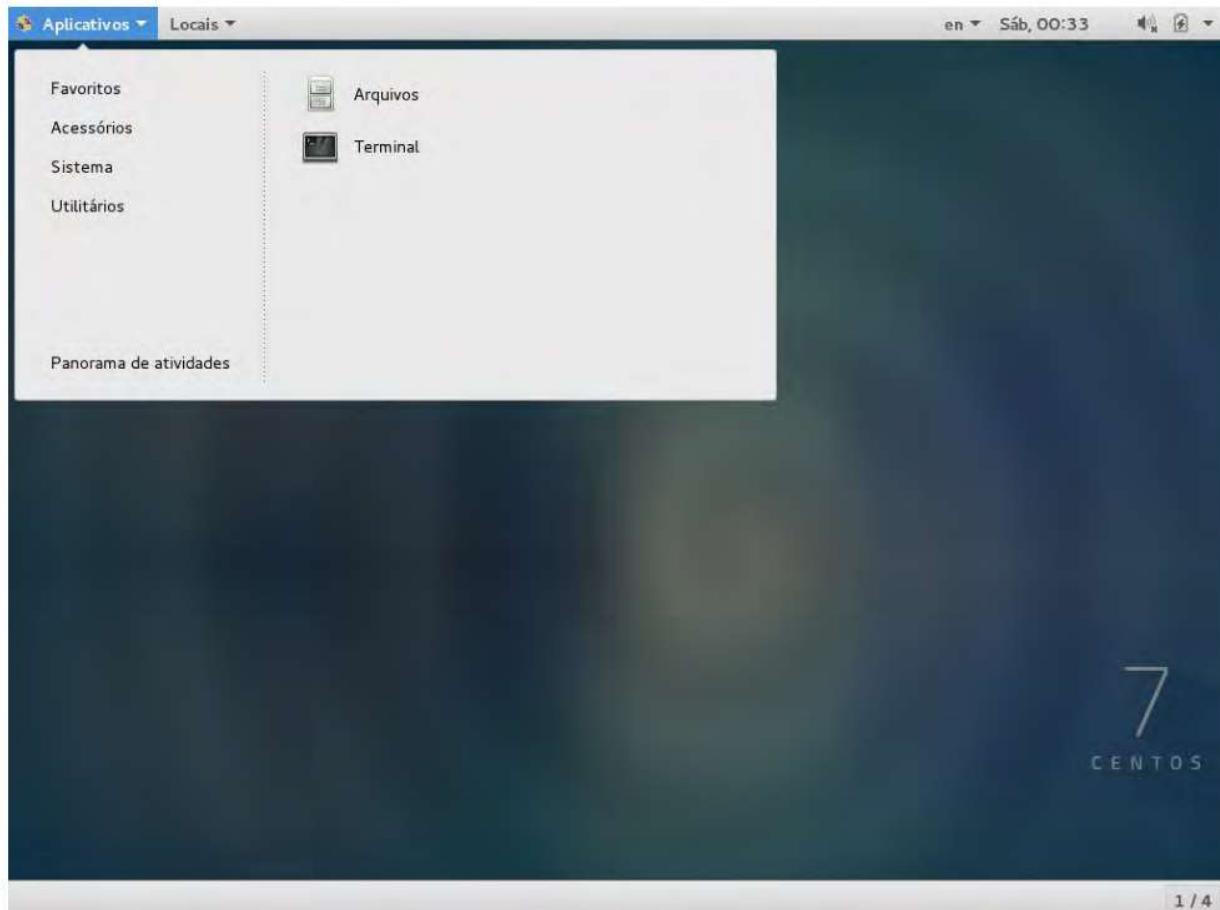
Primeiro checkpoint:

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

*Engenharia de Dados com Hadoop e Spark 3.0*

A screenshot of a Linux desktop environment showing a terminal window. The window title bar says "Terminal". The terminal prompt is "aluno@dataserver:/home/aluno". The user types "[aluno@dataserver ~]\$ su" followed by "Senha:" and then "[root@dataserver aluno]#". The window has a standard window manager interface with icons for "Aplicativos" and "Locais" in the top left, and "en" and "Sáb, 00:34" in the top right. The bottom status bar shows "aluno@dataserver:/home/aluno" and "1 / 4".

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

*Engenharia de Dados com Hadoop e Spark 3.0*A screenshot of a Linux terminal window titled "Terminal". The window has a title bar with "Aplicativos", "Locais", and "Terminal". The status bar shows "en Seg, 01:12". The terminal itself shows the following session:

```
aluno@dataserver:~$ su
Senha:
[root@dataserver aluno]# yum install gedit
```

The terminal is running on a server named "dataserver" under user "aluno". The command "su" is used to switch to root privileges, and "yum install gedit" is run to install the text editor "gedit".

1 / 4

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

*Engenharia de Dados com Hadoop e Spark 3.0*

```

Aplicativos Locais Terminal en Sáb, 00:37
aluno@dataserver:/home/aluno

Arquivo Editar Ver Pesquisar Terminal Ajuda
=====
Instalando:
gedit           x86_64      2:3.14.3-9.el7      base      2,5 M
Instalando para as dependências:
gtksourceview3  x86_64      3.14.3-1.el7      base      946 k
libpeas         x86_64      1.12.1-1.el7      base      119 k

Resumo da transação
=====
Instalar 1 Package (+2 Dependent packages)

Tamanho total do download: 3,6 M
Tamanho depois de instalado: 19 M
Is this ok [y/d/N]: y
Downloading packages:
(1/3): libpeas-1.12.1-1.el7.x86_64.rpm | 119 KB  00:00:00
(2/3): gtksourceview3-3.14.3-1.el7.x86_64.rpm | 946 KB  00:00:01
(3/3): gedit-3.14.3-9.el7.x86_64.rpm | 2,5 MB   00:00:02

Total                                         1,8 MB/s | 3,6 MB  00:00:02
Running transaction check
Running transaction test
Transaction test succeeded
Running transaction
  Instalando : libpeas-1.12.1-1.el7.x86_64          1/3
  Instalando : gtksourceview3-3.14.3-1.el7.x86_64    2/3
  Instalando : 2:gedit-3.14.3-9.el7.x86_64          3/3
  Verificando: gtksourceview3-3.14.3-1.el7.x86_64    1/3
  Verificando: 2:gedit-3.14.3-9.el7.x86_64          2/3
  Verificando: libpeas-1.12.1-1.el7.x86_64          3/3

Instalados:
gedit.x86_64 2:3.14.3-9.el7

Dependência(s) instalada(s):
gtksourceview3.x86_64 0:3.14.3-1.el7               libpeas.x86_64 0:1.12.1-1.el7

Concluído!
[root@dataserver aluno]# █

```

1 / 4

gedit instalado

*Engenharia de Dados com Hadoop e Spark 3.0*A screenshot of a Linux desktop environment showing a terminal window. The window title is 'Terminal'. The terminal prompt shows 'aluno@dataserver:~\$'. Below the prompt, the command '[root@dataserver aluno]# gedit /etc/sudoers' is visible. The window has standard window controls (minimize, maximize, close) at the top right. The background of the desktop is white.

Editar o arquivo /etc/sudoers usando o gedit

1 / 4

*Engenharia de Dados com Hadoop e Spark 3.0*

Aplicativos Locais gedit

sudoers /etc

Abrir Salvar

```

Defaults env_reset
Defaults env_keep += "COLORS DISPLAY HOSTNAME HISTSIZE INPUTRC KDEDIR LS_COLORS"
Defaults env_keep += "MAIL PS1 PS2 QTDIR USERNAME LANG LC_ADDRESS LC_CTYPE"
Defaults env_keep += "LC_COLLATE LC_IDENTIFICATION LC_MEASUREMENT LC_MESSAGES"
Defaults env_keep += "LC_MONETARY LC_NAME LC_NUMERIC LC_PAPER LC_TELEPHONE"
Defaults env_keep += "LC_TIME LC_ALL LANGUAGE LINGUAS _XKB_CHARSET XAUTHORITY"

#
# Adding HOME to env_keep may enable a user to run unrestricted
# commands via sudo.
#
# Defaults env_keep += "HOME"

Defaults secure_path = /sbin:/bin:/usr/sbin:/usr/bin

## Next comes the main part: which users can run what software on
## which machines (the sudoers file can be shared between multiple
## systems).
## Syntax:
##
##     user    MACHINE=COMMANDS
##
## The COMMANDS section may have other options added to it.
##
## Allow root to run any commands anywhere
root    ALL=(ALL)        ALL
aluno   ALL=(ALL)        ALL

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

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

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

```

Matlab Largura da tabulação: 8 Lin 99, Col 28 INS

aluno@dataserver:~\$ sudoers (/etc) - gedit 1 / 4

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

*Engenharia de Dados com Hadoop e Spark 3.0*A screenshot of a Linux desktop environment showing a terminal window. The window title is 'Terminal'. The terminal prompt shows 'aluno@dataserver:~'. Below the prompt, the command '[aluno@dataserver ~]\$ sudo yum install firefox' is visible. The background of the desktop shows a blurred image of a presentation slide.

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

*Engenharia de Dados com Hadoop e Spark 3.0*

```

Aplicativos Locais Terminal en Sáb, 00:45
aluno@dataserver:~>

Arquivo Editar Ver Pesquisar Terminal Ajuda
liberation-sans-fonts noarch 1:1.07.2-15.el7 base 279 k
libvpx x86_64 1.3.0-5.el7_0 base 498 k

Resumo da transação
=====
Instalar 1 Package (+3 Dependent packages)

Tamanho total do download: 72 M
Tamanho depois de instalado: 133 M
Is this ok [y/d/N]: y
Downloading packages:
(1/4): centos-indexhtml-7-9.el7.centos.noarch.rpm | 92 kB 00:00:00
(2/4): liberation-sans-fonts-1.07.2-15.el7.noarch.rpm | 279 kB 00:00:00
(3/4): libvpx-1.3.0-5.el7_0.x86_64.rpm | 498 kB 00:00:01
(4/4): firefox-88.6.0-1.el7.centos.x86_64.rpm | 72 MB 00:00:25
Total 2.9 MB/s | 72 MB 00:00:25

Running transaction check
Running transaction test
Transaction test succeeded
Running transaction
  Instalando : 1:liberation-sans-fonts-1.07.2-15.el7.noarch 1/4
  Instalando : centos-indexhtml-7-9.el7.centos.noarch 2/4
  Instalando : libvpx-1.3.0-5.el7_0.x86_64 3/4
  Instalando : firefox-88.6.0-1.el7.centos.x86_64 4/4
  verifying   : libvpx-1.3.0-5.el7_0.x86_64 1/4
  verifying   : centos-indexhtml-7-9.el7.centos.noarch 2/4
  verifying   : firefox-88.6.0-1.el7.centos.x86_64 3/4
  verifying   : 1:liberation-sans-fonts-1.07.2-15.el7.noarch 4/4

Instalados:
firefox.x86_64 0:88.6.0-1.el7.centos

Dependência(s) instalada(s):
centos-indexhtml.noarch 0:7-9.el7.centos           liberation-sans-fonts.noarch 1:1.07.2-15.el7
libvpx.x86_64 0:1.3.0-5.el7_0

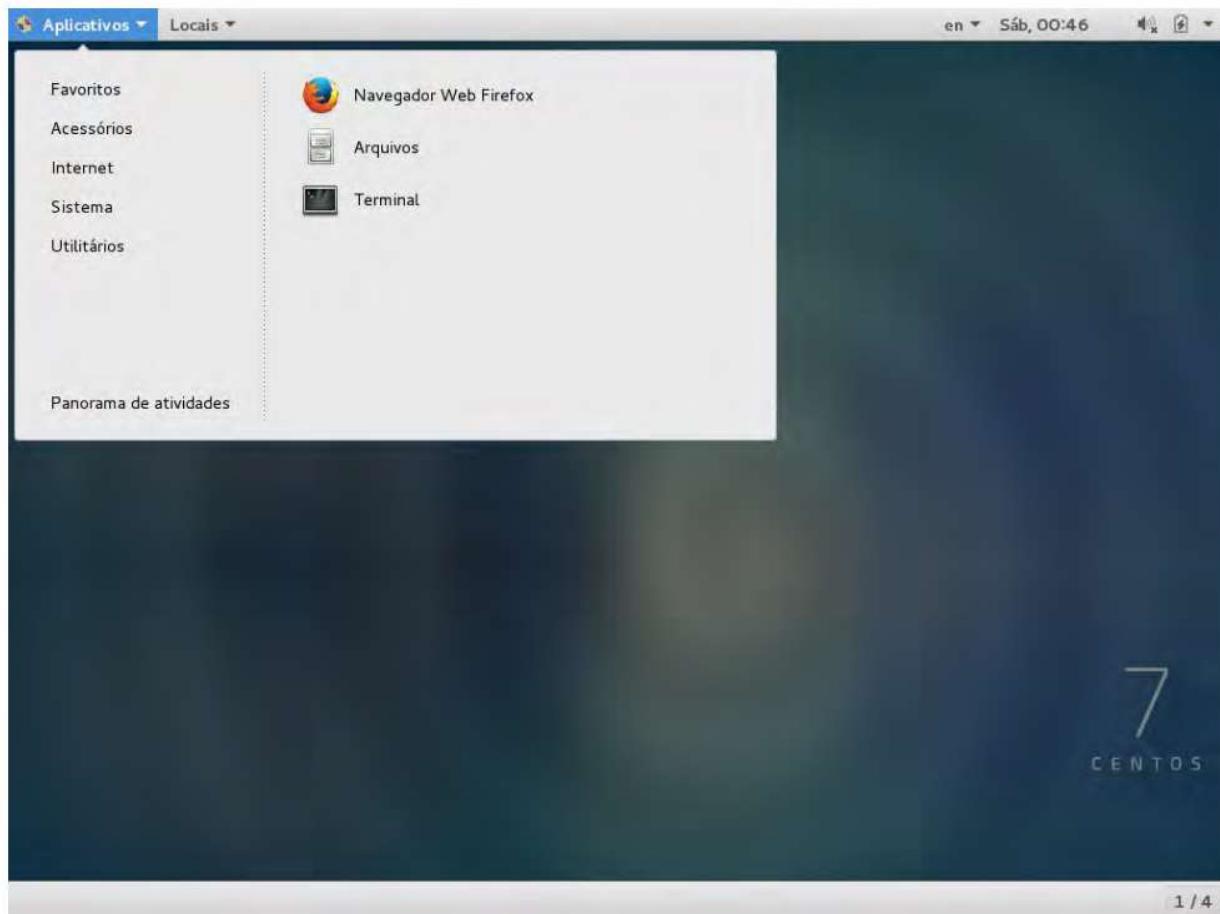
Concluído!
[aluno@dataserver ~]$ 1 / 4

```

Instalação concluída

Data Science
Academy

Data Science Academy phelipe.utsemprboni@outlook.com 5c8a62005e4cde1acb8b45a3

Engenharia de Dados com Hadoop e Spark 3.0

Firefox instalado

*Engenharia de Dados com Hadoop e Spark 3.0*

A screenshot of a Linux terminal window titled "Terminal". The window has a menu bar with "Aplicativos", "Locais", and "Terminal". The status bar shows "en Dom, 07:04". The terminal prompt is "aluno@dataserver:~\$". Below the prompt, the command "sudo yum install bzip2 unzip rsync wget net-tools" is visible. The window has a standard title bar with close, minimize, and maximize buttons. A vertical scroll bar is on the right side of the terminal window.

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

1 / 4

*Engenharia de Dados com Hadoop e Spark 3.0*

```

Aplicativos Locais Terminal en Dom, 07:04
aluno@dataserver:~>

Arquivo Editar Ver Pesquisar Terminal Ajuda
unzip x86_64 6.0-15.el7 base 166 k
wget x86_64 1.14-10.el7_0.1 base 545 k

Resumo da transação
=====
Instalar 5 Packages

Tamanho total do download: 1.4 M
Tamanho depois de instalado: 4.0 M
Is this ok [y/d/N]: y
Downloading packages:
(1/5): bzip2-1.0.6-13.el7.x86_64.rpm | 52 kB 00:00:00
(2/5): wget-1.14-10.el7_0.1.x86_64.rpm | 545 kB 00:00:00
(3/5): unzip-6.0-15.el7.x86_64.rpm | 166 kB 00:00:00
(4/5): net-tools-2.0-0.17.20131004git.el7.x86_64.rpm | 304 kB 00:00:01
(5/5): rsync-3.0.9-17.el7.x86_64.rpm | 360 kB 00:00:02
-----
Total 666 kB/s | 1.4 MB 00:00:02

Running transaction check
Running transaction test
Transaction test succeeded
Running transaction
  Instalando : bzip2-1.0.6-13.el7.x86_64 1/5
  Instalando : net-tools-2.0-0.17.20131004git.el7.x86_64 2/5
  Instalando : wget-1.14-10.el7_0.1.x86_64 3/5
  Instalando : rsync-3.0.9-17.el7.x86_64 4/5
  Instalando : unzip-6.0-15.el7.x86_64 5/5
  Verifying  : bzip2-1.0.6-13.el7.x86_64 1/5
  Verifying  : net-tools-2.0-0.17.20131004git.el7.x86_64 2/5
  Verifying  : wget-1.14-10.el7_0.1.x86_64 3/5
  Verifying  : rsync-3.0.9-17.el7.x86_64 4/5
  Verifying  : unzip-6.0-15.el7.x86_64 5/5

Instalados:
  bzip2.x86_64 0:1.0.6-13.el7    net-tools.x86_64 0:2.0-0.17.20131004git.el7    rsync.x86_64 0:3.0.9-17.el7
  unzip.x86_64 0:6.0-15.el7      wget.x86_64 0:1.14-10.el7_0.1

Concluído!
[aluno@dataserver ~]$ 1 / 4

```

Aplicativos instalados



2.4. Instalação do MySQL

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

A screenshot of a Linux terminal window titled "DataServer Hadoop [Running]". The window has a standard Gnome-style title bar with icons for minimize, maximize, and close. The terminal itself shows the command "sudo yum localinstall https://dev.mysql.com/get/mysql80-community-release-el7-1.noarch.rpm" being typed in. The user is "aluno" and the host is "databserver". The terminal window is part of a desktop environment with other windows visible in the background.

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
```

*Engenharia de Dados com Hadoop e Spark 3.0*

```
Dataserver Hadoop (Running)
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
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]: █
```

Pressione y

*Engenharia de Dados com Hadoop e Spark 3.0*

```
aluno@datavir: Hadoop [Running]
File Edit View Search Terminal Help
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]: y
Downloading packages:
Running transaction check
Running transaction test
Transaction test succeeded
Running transaction
  Installing : mysql80-community-release-el7-1.noarch 1/1
  Verifying  : mysql80-community-release-el7-1.noarch 1/1

Installed:
  mysql80-community-release.noarch 0:el7-1

Complete!
[aluno@datavir ~]$
```

Download do pacote concluído

*Engenharia de Dados com Hadoop e Spark 3.0*

The screenshot shows a terminal window titled "DataServer Hadoop (Running)". The window has a standard Linux desktop interface at the top with icons for Applications, Places, Terminal, and a user icon. The status bar at the bottom right shows the date as Fri 21:15 and the battery level as Left 96%. The terminal itself has a light gray background and a dark gray border. It displays the command "aluno@dataserver:~\$ sudo yum install mysql-community-server" in white text. The cursor is positioned at the end of the command line.

Esse comando inicia a instalação do MySQL:

`sudo yum install mysql-community-server`

*Engenharia de Dados com Hadoop e Spark 3.0*

```

aluno@dataserver:~$ sudo yum install mysql-community-server
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]: 

```

Pressione y



Engenharia de Dados com Hadoop e Spark 3.0

```
Fingerprint: a4a9 4068 76fc bd3c 4567 78c8 8c71 8d3b 5072 e1f5
Package      : mysql80-community-release-el7-1.noarch (installed)
From        : /etc/pki/rpm-gpg/RPM-GPG-KEY-mysql
Is this ok [y/N]: y
Running transaction check
Running transaction test
Transaction test succeeded
Running transaction
  Installing : mysql-community-common-8.0.16-2.el7.x86_64          1/6
  Installing : mysql-community-libs-8.0.16-2.el7.x86_64          2/6
  Installing : mysql-community-client-8.0.16-2.el7.x86_64         3/6
  Installing : mysql-community-server-8.0.16-2.el7.x86_64         4/6
  Installing : mysql-community-libs-compat-8.0.16-2.el7.x86_64     5/6
  Erasing   : 1:mariadb-libs-5.5.60-1.el7_5.x86_64               6/6
  Verifying  : mysql-community-libs-8.0.16-2.el7.x86_64          1/6
  Verifying  : mysql-community-libs-compat-8.0.16-2.el7.x86_64     2/6
  Verifying  : mysql-community-client-8.0.16-2.el7.x86_64         3/6
  Verifying  : mysql-community-common-8.0.16-2.el7.x86_64         4/6
  Verifying  : mysql-community-server-8.0.16-2.el7.x86_64         5/6
  Verifying  : 1:mariadb-libs-5.5.60-1.el7_5.x86_64               6/6

Installed:
  mysql-community-libs.x86_64 0:8.0.16-2.el7                      mysql-community-libs-compat.x86_64 0:8.0.16-2.el7
  mysql-community-server.x86_64 0:8.0.16-2.el7

Dependency Installed:
  mysql-community-client.x86_64 0:8.0.16-2.el7                  mysql-community-common.x86_64 0:8.0.16-2.el7

Replaced:
  mariadb-libs.x86_64 1:5.5.60-1.el7_5

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

Instalação concluída

*Engenharia de Dados com Hadoop e Spark 3.0*A screenshot of a Linux desktop environment showing a terminal window titled "DataServer Hadoop [Running]". The terminal window has a title bar with the title and the user "aluno@dataserver:~". The main area of the terminal shows the command "[aluno@dataserver ~]\$ sudo systemctl enable mysqld" entered by the user. The background of the desktop shows icons for various applications like a browser, file manager, and system tools.

Ativando o serviço do MySQL

`sudo systemctl enable mysqld`

*Engenharia de Dados com Hadoop e Spark 3.0*A screenshot of a Linux desktop environment showing a terminal window titled "DataServer-Hadoop [Running]". The terminal window has a title bar, a menu bar with "File Edit View Search Terminal Help", and a status bar showing "aluno@dataserver:~" and the date/time "Fri 21:22". The main area of the terminal shows the command "[aluno@dataserver ~]\$ sudo systemctl start mysqld" being typed. The background of the desktop shows icons for various applications like a browser, file manager, and system tools.

```
[aluno@dataserver ~]$ sudo systemctl start mysqld
```

Inicia o MySQL

sudo systemctl start mysqld

*Engenharia de Dados com Hadoop e Spark 3.0*

The screenshot shows a terminal window titled "DataServer/Hadoop [Running]". The window title bar includes "File Edit View Search Terminal Help" and the user "aluno@dataserver:~". The terminal content displays the following command and its output:

```
[aluno@dataserver ~]$ sudo systemctl start mysqld
[aluno@dataserver ~]$ sudo systemctl status mysqld
● mysqld.service - MySQL Server
  Loaded: loaded (/usr/lib/systemd/system/mysqld.service; enabled; vendor preset: disabled)
  Active: active (running) since Fri 2019-06-21 21:23:06 PDT; 5s ago
    Docs: man:mysqld(8)
          http://dev.mysql.com/doc/refman/en/using-systemd.html
  Process: 6169 ExecStartPre=/usr/bin/mysqld_pre_systemd (code=exited, status=0/SUCCESS)
 Main PID: 6252 (mysqld)
   Status: "SERVER_OPERATING"
  Tasks: 38
 CGroup: /system.slice/mysqld.service
         └─6252 /usr/sbin/mysqld

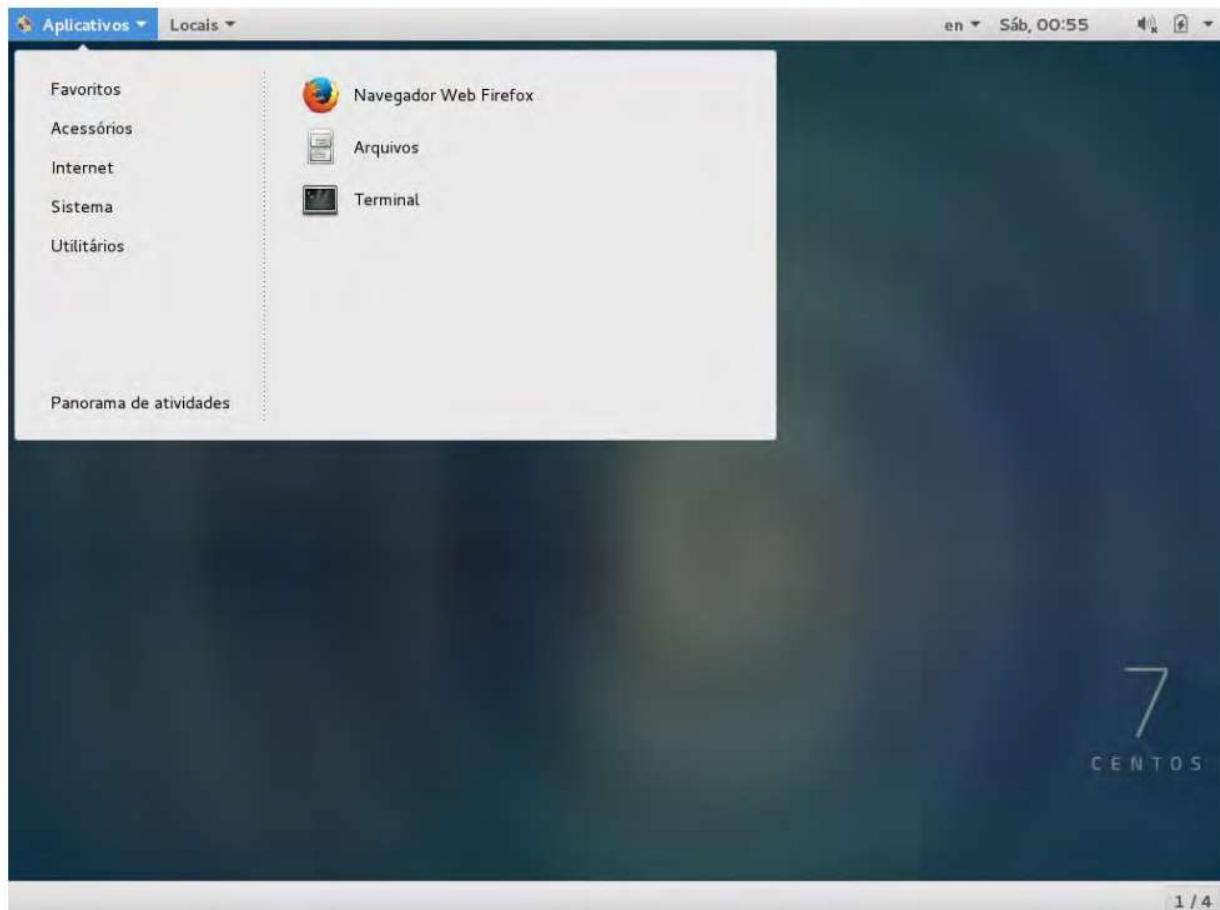
Jun 21 21:23:00 dataserver systemd[1]: Starting MySQL Server...
Jun 21 21:23:06 dataserver systemd[1]: Started MySQL Server.
[aluno@dataserver ~]$
```

Below the terminal window, the desktop environment's taskbar is visible, showing icons for various applications like a browser, file manager, and system tools.

MySQL em execução



3. Instalação do servidor ssh



Abrindo o terminal

*Engenharia de Dados com Hadoop e Spark 3.0*A screenshot of a Linux terminal window titled "Terminal". The window has a menu bar with "Aplicativos", "Locais", and "Terminal". The status bar shows "en Ter, 05:09". The terminal prompt is "aluno@dataserver:~". Below the prompt, the command "[aluno@dataserver ~]\$ sudo yum install openssh-server openssh-clients" is visible. The window has a scroll bar on the right side.

sudo yum install openssh-server openssh-clients

1 / 4

*Engenharia de Dados com Hadoop e Spark 3.0*

```

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-6.6.1pl-23.el7_2.x86_64.rpm | 435 KB 00:00:00
(2/3): openssh-server-6.6.1pl-23.el7_2.x86_64.rpm | 436 KB 00:00:01
(3/3): openssh-clients-6.6.1pl-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-6.6.1pl-23.el7_2.x86_64 1/6
  Atualizando : openssh-server-6.6.1pl-23.el7_2.x86_64 2/6
  Atualizando : openssh-clients-6.6.1pl-23.el7_2.x86_64 3/6
  Limpeza    : openssh-clients-6.6.1pl-22.el7.x86_64 4/6
  Limpeza    : openssh-server-6.6.1pl-22.el7.x86_64 5/6
  Limpeza    : openssh-6.6.1pl-22.el7.x86_64 6/6
  verifying   : openssh-server-6.6.1pl-23.el7_2.x86_64 1/6
  verifying   : openssh-clients-6.6.1pl-23.el7_2.x86_64 2/6
  verifying   : openssh-6.6.1pl-23.el7_2.x86_64 3/6
  verifying   : openssh-clients-6.6.1pl-22.el7.x86_64 4/6
  verifying   : openssh-6.6.1pl-22.el7.x86_64 5/6
  verifying   : openssh-server-6.6.1pl-22.el7.x86_64 6/6

Atualizados:
  openssh-clients.x86_64 0:6.6.1pl-23.el7_2           openssh-server.x86_64 0:6.6.1pl-23.el7_2

Dependência(s) atualizada(s):
  openssh.x86_64 0:6.6.1pl-23.el7_2

Concluído!
[aluno@dataserver ~]$ 1 / 4

```

Concluído

*Engenharia de Dados com Hadoop e Spark 3.0*A screenshot of a Linux desktop environment showing a terminal window titled "DataServer Hadoop [Running]". The window title bar also includes "aluno@dataserver ~". The terminal shows the command "[aluno@dataserver ~]\$ sudo systemctl enable sshd" being typed. The desktop background is visible at the bottom, showing icons for various applications like a browser, file manager, and system tools.

```
aluno@dataserver ~$ sudo systemctl enable sshd
```

Habilitando o serviço

`sudo systemctl enable sshd`

*Engenharia de Dados com Hadoop e Spark 3.0*

```
[aluno@dataserver ~]$ sudo systemctl start sshd
```

Iniciando o serviço

`sudo systemctl start sshd`

*Engenharia de Dados com Hadoop e Spark 3.0*

```
[aluno@dataserver ~]$ sudo systemctl start sshd
[aluno@dataserver ~]$ sudo systemctl status sshd
● sshd.service - OpenSSH server daemon
  Loaded: loaded (/usr/lib/systemd/system/sshd.service; enabled; vendor preset: enabled)
  Active: active (running) since Fri 2019-06-21 21:02:24 PDT; 32min ago
    Docs: man:sshd(8)
          man:sshd_config(5)
  Main PID: 3593 (sshd)
     CGroup: /system.slice/sshd.service
             └─3593 /usr/sbin/sshd -D

Jun 21 21:02:24 dataserver systemd[1]: Starting OpenSSH server daemon...
Jun 21 21:02:24 dataserver sshd[3593]: Server listening on 0.0.0.0 port 22.
Jun 21 21:02:24 dataserver sshd[3593]: Server listening on :: port 22.
Jun 21 21:02:24 dataserver systemd[1]: Started OpenSSH server daemon.
[aluno@dataserver ~]$
```

Serviço em execução

*Engenharia de Dados com Hadoop e Spark 3.0*A screenshot of a Linux desktop environment showing a terminal window. The window title is 'Terminal'. The terminal prompt shows 'aluno@dataserver:~'. Below the prompt, the command '[aluno@dataserver ~]\$ sudo gedit /etc/ssh/sshd_config' is visible. The window has a standard title bar with icons for minimize, maximize, and close, and a status bar at the bottom showing '1 / 4'.

```
aluno@dataserver:~$ sudo gedit /etc/ssh/sshd_config
```



Engenharia de Dados com Hadoop e Spark 3.0

The screenshot shows a terminal window titled "aluno@dataserver:~" with a command prompt. Below it is a gedit window titled "*sshd_config /etc/ssh". The file content is the SSH configuration file. A blue selection bar highlights three lines of code:

```
# AddressFamily any
#ListenAddress 0.0.0.0
#ListenAddress ::
```

These three lines are preceded by a single "#". The rest of the file contains standard SSH configuration options like HostKey, Protocol, and Ciphers.

Primeira parte da configuração ssh.

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



Engenharia de Dados com Hadoop e Spark 3.0

```
HostKey /etc/ssh/ssh_host_rsa_key
#HostKey /etc/ssh/ssh_host_dsa_key
HostKey /etc/ssh/ssh_host_ecdsa_key
HostKey /etc/ssh/ssh_host_ed25519_key

# Lifetime and size of ephemeral version 1 server key
#KeyRegenerationInterval 1h
#ServerKeyBits 1024

# Ciphers and keying
#RekeyLimit default none

# Logging
# obsoletes QuietMode and FascistLogging
#SyslogFacility AUTH
SyslogFacility AUTHPRIV
#LogLevel INFO

# Authentication:

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

#RSAAuthentication yes
#PubkeyAuthentication yes

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

#AuthorizedPrincipalsFile none

#AuthorizedKeysCommand none
#AuthorizedKeysCommandUser nobody
```

Segunda parte da configuração do ssh

*Engenharia de Dados com Hadoop e Spark 3.0*A screenshot of a Linux desktop environment showing a terminal window titled "DataServer-Hadoop [Running]". The window title bar also includes "Applications", "Places", "Terminal", and the date/time "Fri 21:38". The terminal window has a dark background and contains the command "[aluno@dataserver ~]\$ sudo systemctl restart sshd". Below the terminal window, the desktop environment's taskbar is visible, showing various icons and the status "Left 36%".

```
[aluno@dataserver ~]$ sudo systemctl restart sshd
```



4. Instalação do Java 8

4.1. Removendo o OpenJDK

A screenshot of a Linux desktop environment showing a terminal window titled "DataServer Hadoop [Running]". The terminal window has a title bar with "Applications", "Places", "Terminal", and the date "Fri 21/41". The user is logged in as "aluno@dataserver". The command "sudo yum -y remove java*" is being typed into the terminal. The desktop background shows a grid of icons, and the taskbar at the bottom includes icons for file operations like cut, copy, paste, and save.

Removendo o OpenJDK

`sudo yum -y remove java*`

*Engenharia de Dados com Hadoop e Spark 3.0*

```

DataServer Hadoop (running)
Applications Places Terminal
aluno@dataserver:~ 
File Edit View Search Terminal Help
=====
Remove 3 Packages (+4 Dependent packages)

Installed size: 108 M
Downloading packages:
Running transaction check
Running transaction test
Transaction test succeeded
Running transaction
Erasing : icedtea-web-1.7.1-1.el7.x86_64 1/7
Erasing : rhino-1.7R5-1.el7.noarch 2/7
Erasing : jline-1.0-8.el7.noarch 3/7
Erasing : tagsoup-1.2.1-8.el7.noarch 4/7
Erasing : 1:java-1.8.0-openjdk-1.8.0.181-7.b13.x86_64 5/7
Erasing : 1:java-1.8.0-openjdk-headless-1.8.0.181-7.b13.el7.x86_64 6/7
Erasing : javapackages-tools-3.4.1-11.el7.noarch 7/7
Verifying : tagsoup-1.2.1-8.el7.noarch 1/7
Verifying : 1:java-1.8.0-openjdk-1.8.0.181-7.b13.el7.x86_64 2/7
Verifying : javapackages-tools-3.4.1-11.el7.noarch 3/7
Verifying : icedtea-web-1.7.1-1.el7.x86_64 4/7
Verifying : jline-1.0-8.el7.noarch 5/7
Verifying : rhino-1.7R5-1.el7.noarch 6/7
Verifying : 1:java-1.8.0-openjdk-headless-1.8.0.181-7.b13.el7.x86_64 7/7

Removed:
java-1.8.0-openjdk.x86_64 1:1.8.0.181-7.b13.el7      java-1.8.0-openjdk-headless.x86_64 1:1.8.0.181-7.b13.el7
javapackages-tools.noarch 0:3.4.1-11.el7

Dependency Removed:
icedtea-web.x86_64 0:1.7.1-1.el7    jline.noarch 0:1.0-8.el7    rhino.noarch 0:1.7R5-1.el7    tagsoup.noarch 0:1.2.1-8.el7

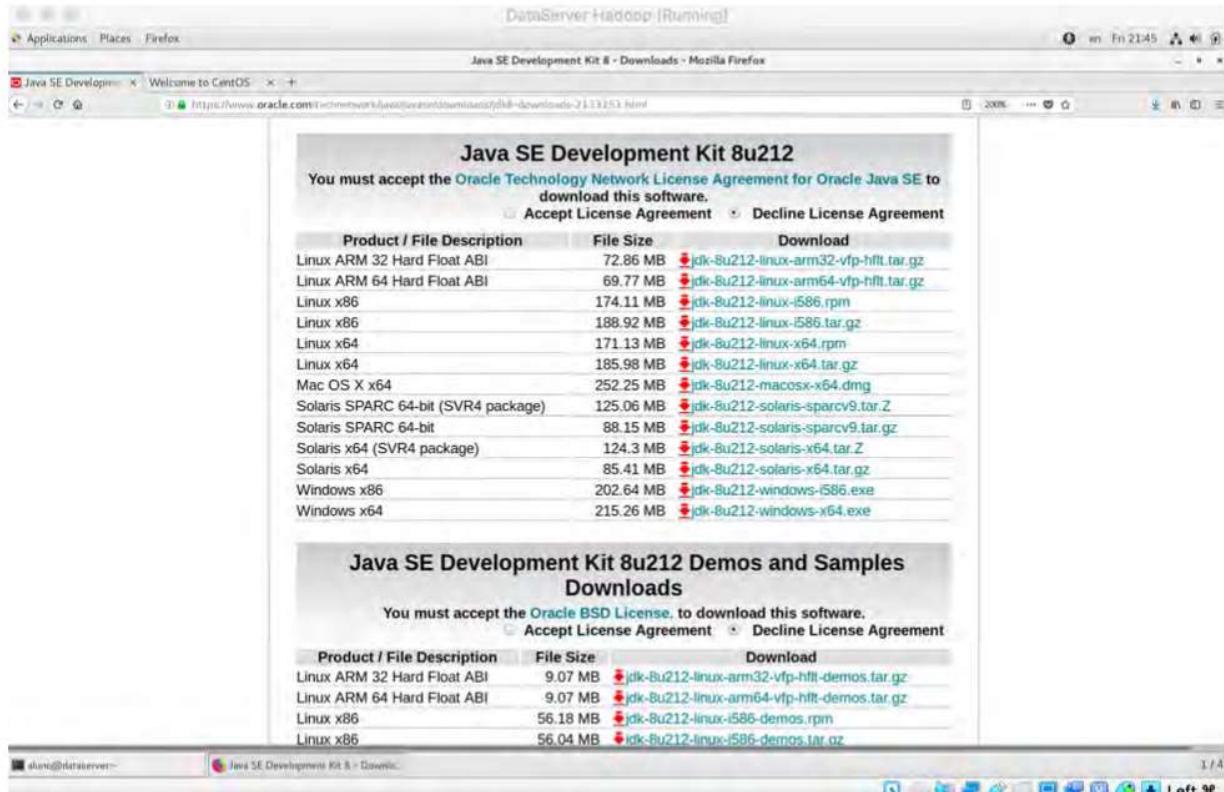
Complete!
[aluno@dataserver ~]$ 

```

Concluído

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

4.2. Instalação do JDK



Java SE Development Kit 8u212

You must accept the [Oracle Technology Network License Agreement](#) for Oracle Java SE to download this software.

Accept License Agreement Decline License Agreement

Product / File Description	File Size	Download
Linux ARM 32 Hard Float ABI	72.86 MB	jdk-8u212-linux-arm32-vfp-hflt.tar.gz
Linux ARM 64 Hard Float ABI	69.77 MB	jdk-8u212-linux-arm64-vfp-hflt.tar.gz
Linux x86	174.11 MB	jdk-8u212-linux-i586.rpm
Linux x86	188.92 MB	jdk-8u212-linux-i586.tar.gz
Linux x64	171.13 MB	jdk-8u212-linux-x64.rpm
Linux x64	185.98 MB	jdk-8u212-linux-x64.tar.gz
Mac OS X x64	252.25 MB	jdk-8u212-macosx-x64.dmg
Solaris SPARC 64-bit (SVR4 package)	125.06 MB	jdk-8u212-solaris-sparcv9.tar.Z
Solaris SPARC 64-bit	88.15 MB	jdk-8u212-solaris-sparcv9.tar.gz
Solaris x64 (SVR4 package)	124.3 MB	jdk-8u212-solaris-x64.tar.Z
Solaris x64	85.41 MB	jdk-8u212-solaris-x64.tar.gz
Windows x86	202.64 MB	jdk-8u212-windows-i586.exe
Windows x64	215.26 MB	jdk-8u212-windows-x64.exe

Java SE Development Kit 8u212 Demos and Samples Downloads

You must accept the [Oracle BSD License](#), to download this software.

Accept License Agreement Decline License Agreement

Product / File Description	File Size	Download
Linux ARM 32 Hard Float ABI	9.07 MB	jdk-8u212-linux-arm32-vfp-hflt-demos.tar.gz
Linux ARM 64 Hard Float ABI	9.07 MB	jdk-8u212-linux-arm64-vfp-hflt-demos.tar.gz
Linux x86	56.18 MB	jdk-8u212-linux-i586-demos.rpm
Linux x86	56.04 MB	jdk-8u212-linux-i586-demos.tar.gz

No site da Oracle, fazer o download do JDK

*Engenharia de Dados com Hadoop e Spark 3.0*

The screenshot shows a terminal window titled "DataServer Hadoop [Running]". The window has a standard Linux desktop interface at the top with icons for Applications, Places, Terminal, and a clock showing Fri 21:47. The terminal itself has a light gray background and a dark gray border. It displays the following command sequence:

```
[aluno@dataserver ~]$ cd Downloads/  
[aluno@dataserver Downloads]$ ls -l  
total 256320  
-rw-rw-r--. 1 aluno aluno 195013152 Jun 21 21:44 jdk-8u212-linux-x64.tar.gz  
[aluno@dataserver Downloads]$ tar -xzf jdk-8u212-linux-x64.tar.gz
```

Below the terminal window, the desktop environment is visible, showing a taskbar with several icons and windows. One window is titled "Java Development Kit 8 - Download" and another is titled "aluno@dataserver ~/Downloads". The desktop has a light blue background with a grid pattern.

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

*Engenharia de Dados com Hadoop e Spark 3.0*

```
[aluno@dataserver ~]$ cd Downloads/
[aluno@dataserver Downloads]$ ls -l
total 256320
-rw-rw-r--. 1 aluno aluno 195013152 Jun 21 21:44 jdk-8u212-linux-x64.tar.gz
[aluno@dataserver Downloads]$ tar -xzf jdk-8u212-linux-x64.tar.gz
[aluno@dataserver Downloads]$ sudo mv jdk1.8.0_212/ /opt/jdk
```

The screenshot shows a terminal window titled "DataServer Hadoop [Running]". The terminal session starts with the user navigating to the "Downloads" directory, listing its contents, and then extracting a Java Development Kit (JDK) archive. Finally, the user runs a "sudo mv" command to move the extracted directory to the "/opt/jdk" location. The terminal window has a standard Linux-style interface with a menu bar and a toolbar at the top.

Mover o diretório do JDK

*Engenharia de Dados com Hadoop e Spark 3.0*A screenshot of a Linux terminal window titled "Terminal". The window has a menu bar with "Aplicativos", "Locais", and "Terminal". The status bar shows "en Ter, 03:31". The terminal prompt is "aluno@dataserver:~". Below the prompt, the user has typed the command "cd ~".

```
Arquivo Editar Ver Pesquisar Terminal Ajuda
[aluno@dataserver opt]$ cd ~
[aluno@dataserver ~]$
```

cd ~

1 / 4

*Engenharia de Dados com Hadoop e Spark 3.0*

A screenshot of a terminal window titled "Terminal" at "aluno@dataserver:~". The window shows a file listing with the command "ls -la" and a command to edit ".bashrc" with "gedit .bashrc".

```
Arquivo Editar Ver Pesquisar Terminal Ajuda
[aluno@dataserver opt]$ cd ~
[aluno@dataserver ~]$ ls -la
total 40
drwx----- 8 aluno aluno 4096 Abr  5 03:24 .
drwxr-xr-x  3 root  root  18 Abr  3 03:00 ..
-rw-------  1 aluno aluno 1229 Abr  3 03:54 .bash_history
-rw-r--r--  1 aluno aluno  18 Nov 20 03:02 .bash_logout
-rw-r--r--  1 aluno aluno 193 Nov 20 03:02 .bash_profile
-rw-r--r--  1 aluno aluno 231 Nov 20 03:02 .bashrc
drwx----- 8 aluno aluno 4096 Abr  3 03:43 .cache
drwxr-xr-x 11 aluno aluno 4096 Abr  3 03:16 .config
drwxr-xr-x  2 aluno aluno   6 Abr  3 03:16 Desktop
drwx----- 2 aluno aluno  38 Abr  5 03:28 Downloads
-rw-------  1 aluno aluno  16 Abr  3 03:16 .esd_auth
-rw-------  1 aluno aluno 1240 Abr  5 03:24 .ICEAuthority
drwx----- 3 aluno aluno  18 Abr  3 03:16 .local
drwxrwxr-x  4 aluno aluno  37 Abr  3 03:27 .mozilla
-rw-------  1 aluno aluno  91 Abr  3 03:36 .mysql_history
[aluno@dataserver ~]$ gedit .bashrc
```

gedit .bashrc

1 / 4

*Engenharia de Dados com Hadoop e Spark 3.0*

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

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

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

Editar as variáveis de ambiente conforme acima e salvar o arquivo

*Engenharia de Dados com Hadoop e Spark 3.0*A screenshot of a Linux terminal window titled "Terminal". The window has a menu bar with "Aplicativos", "Locais", and "Terminal". The status bar shows "en Ter, 03:34". The terminal prompt is "aluno@dataserver:~". The user has typed the command "[aluno@dataserver ~]\$ source .bashrc" and pressed Enter. The window title bar also displays "aluno@dataserver:~".

```
aluno@dataserver:~$ source .bashrc
```

source .bashrc

1 / 4

*Engenharia de Dados com Hadoop e Spark 3.0*

```
[aluno@dataserver ~]$ java -version
java version "1.8.0_212"
Java(TM) SE Runtime Environment (build 1.8.0_212-b10)
Java HotSpot(TM) 64-Bit Server VM (build 25.212-b10, mixed mode)
[aluno@dataserver ~]$
```

Java(TM) SE Development Kit X (Java(TM) 8) - Overview

aluno@dataserver:~\$

1 / 4

Left %

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

A screenshot of a terminal window titled "DataServer Hadoop (Running)". The window shows a command-line interface with the prompt "aluno@dataserver:~\$". A user is typing the command "sudo adduser hadoop" into the terminal. The window has a standard Linux-style header with icons for Applications, Places, Terminal, and a date/time indicator "Fri 21/01".

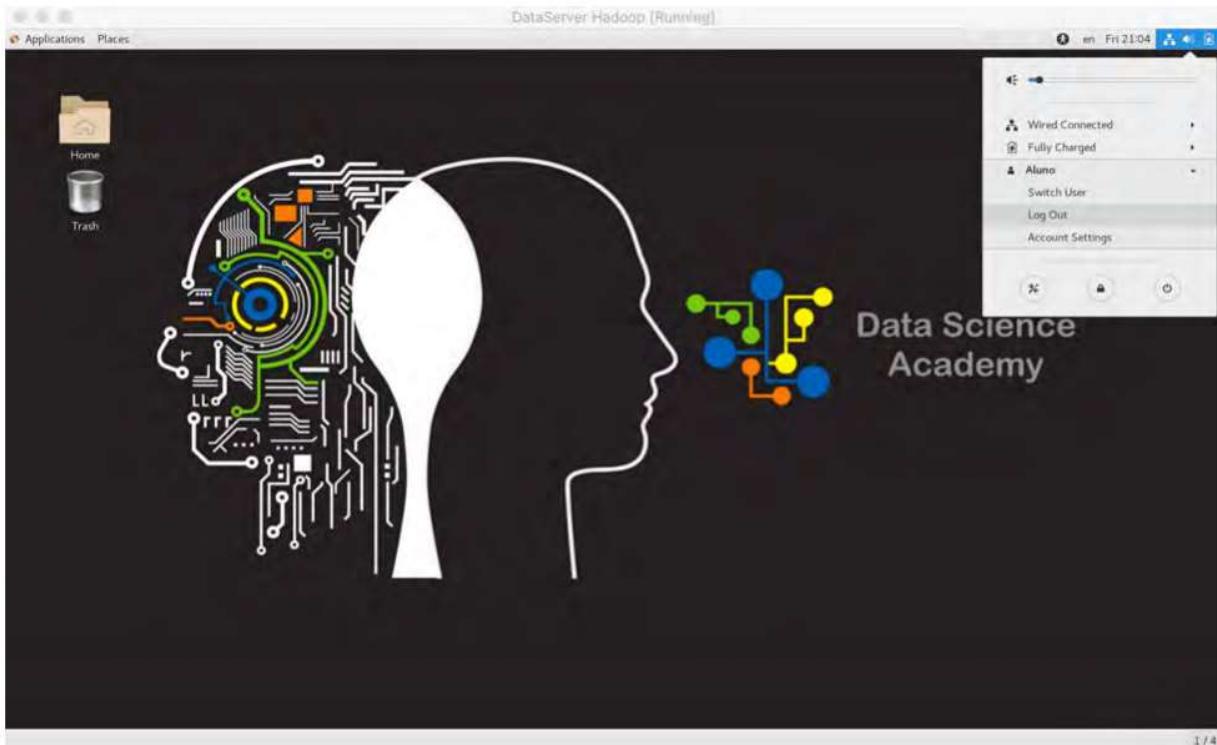
```
aluno@dataserver ~$ sudo adduser hadoop
```

sudo adduser hadoop – para criar o usuário hadoop

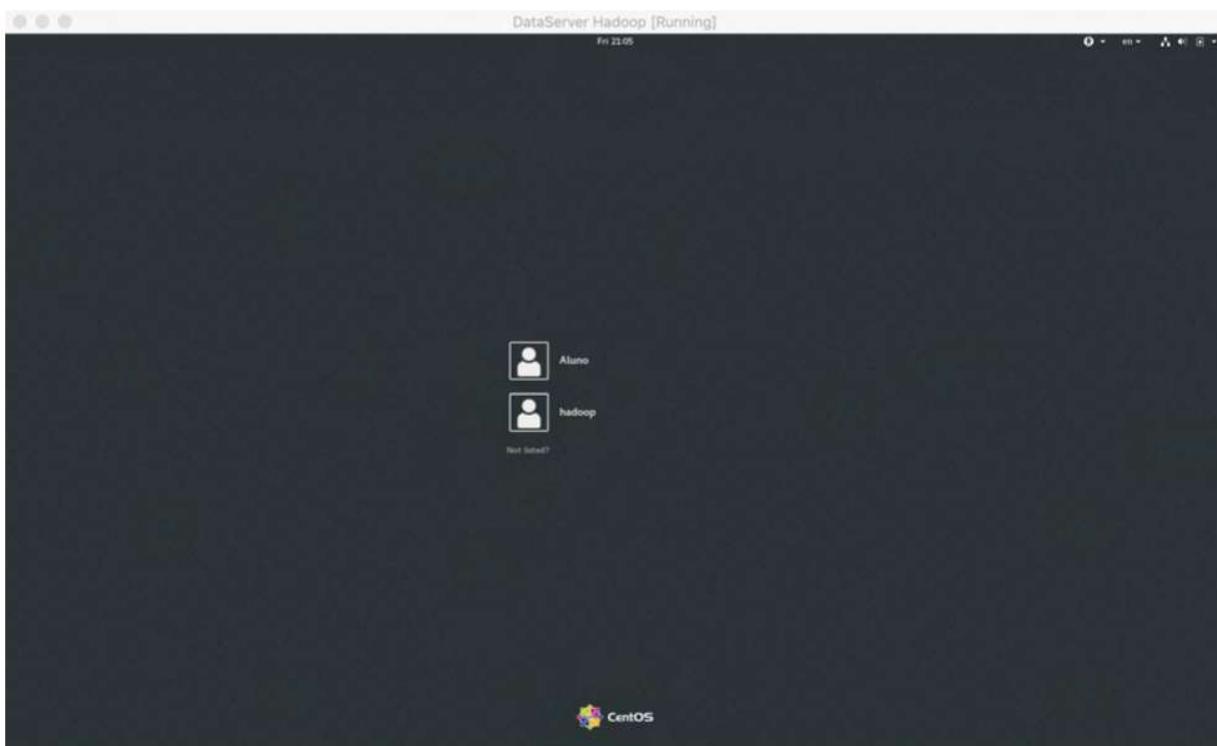
*Engenharia de Dados com Hadoop e Spark 3.0*A screenshot of a terminal window titled "DataServer Hadoop [Running]". The window is part of a desktop environment with a menu bar at the top showing "Applications", "Places", and "Terminal". The terminal itself has a title bar with "aluno@dataserver:~>". The command "sudo passwd hadoop" is visible in the terminal window. The background shows a blurred view of the desktop environment.

```
File Edit View Search Terminal Help
[aluno@dataserver ~]$ sudo passwd hadoop
```

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

*Engenharia de Dados com Hadoop e Spark 3.0*

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

The screenshot shows a terminal window titled "DataServer Hadoop [Running]". The command entered is "[hadoop@dataserver ~]\$ ssh-keygen -t rsa". The terminal is running on a Linux desktop environment with a window title bar showing "hadoop@dataserver ~" and a status bar indicating "Fri 21:12".

The screenshot shows a terminal window titled "DataServer Hadoop [Running]". The command entered is "[hadoop@dataserver ~]\$ ssh-keygen -t rsa". The output shows the generation of an RSA key pair and prompts the user to enter a file name to save the key. The terminal window has a window title bar showing "hadoop@dataserver ~" and a status bar indicating "Fri 21:12".

The screenshot shows a terminal window titled "DataServer Hadoop [Running]". The command entered is "[hadoop@dataserver ~]\$ ssh-keygen -t rsa". The output shows the generation of an RSA key pair and prompts the user to enter a file name to save the key. The terminal window has a window title bar showing "hadoop@dataserver ~" and a status bar indicating "Fri 21:12".



Engenharia de Dados com Hadoop e Spark 3.0

Pressionar Enter para confirmar o diretório onde as chaves serão geradas

A screenshot of a terminal window titled "DataServer Hadoop (Ubuntu)". The window shows the command [hadoop@datavserver ~]\$ ssh-keygen -t rsa being run. The output indicates that a public/private RSA key pair is being generated, a directory for saving the key is chosen, and a passphrase is requested.

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

Pressionar Enter

A second screenshot of a terminal window titled "DataServer Hadoop (Ubuntu)". The command [hadoop@datavserver ~]\$ ssh-keygen -t rsa is run again. The output shows the key generation process, including the creation of a directory and the request for a passphrase.

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

Pressionar Enter novamente



```
File Edit View Search Terminal Help
[hadoop@dataserver ~]$ ssh-keygen -t rsa
Generating public/private rsa key pair.
Enter file in which to save the key (/home/hadoop/.ssh/id_rsa):
Created directory '/home/hadoop/.ssh'.
Enter passphrase (empty for no passphrase):
Enter same passphrase again:
Your identification has been saved in /home/hadoop/.ssh/id_rsa.
Your public key has been saved in /home/hadoop/.ssh/id_rsa.pub.
The key fingerprint is:
SHA256:8aiwd7k6fxuy25ljIBB4tY0Fqef23V5es+/n8u47CFM hadoop@dataserver
The key's randomart image is:
+---[RSA 2048]---+
|   .oo. |
|   o .= |
|   .oo o |
|   o . + E |
|   .+ S .. |
|   o+...o |
|   ..oo+o..o.. |
|   ....==+.+o+ |
|   .+=+=+.X%|
+---[SHA256]---+
[hadoop@dataserver ~]$
```

Chaves de segurança geradas

```
File Edit View Search Terminal Help
[hadoop@dataserver ~]$ cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
```

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



Engenharia de Dados com Hadoop e Spark 3.0

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

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

```
[hadoop@dataserver ~]$ sudo gedit /etc/ssh/sshd_config
```

*Engenharia de Dados com Hadoop e Spark 3.0*

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

```

# This sshd was compiled with PATH=/usr/local/bin:/usr/bin

# The strategy used for options in the default sshd_config shipped with
# OpenSSH is to specify options with their default value where
# possible, but leave them commented. Uncommented options override the
# default value.

# If you want to change the port on a SELinux system, you have to tell
# SELinux about this change.
# semanage port -a -t ssh_port_t -p tcp #PORTNUMBER
#
Port 22
AddressFamily any
ListenAddress 0.0.0.0
#ListenAddress ::

HostKey /etc/ssh/ssh_host_rsa_key
#HostKey /etc/ssh/ssh_host_dsa_key
HostKey /etc/ssh/ssh_host_ecdsa_key
HostKey /etc/ssh/ssh_host_ed25519_key

# Ciphers and keying
#RekeyLimit default none

# Logging
#SyslogFacility AUTH
#SyslogFacility AUTHPRIV
#LogLevel INFO

# Authentication:

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

#PubkeyAuthentication yes

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

#AuthorizedPrincipalsFile none

```

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

```

[hadoop@dataserver ~]$ sudo systemctl restart sshd

```

*Engenharia de Dados com Hadoop e Spark 3.0*

Reinic peace o serviço ssh

A screenshot of a Linux desktop environment showing a terminal window titled "DataServer Hadoop (Running)". The window is a standard Xfce-style terminal. The title bar includes the window name and the user's name "hadoop@dataserver:~". The menu bar has options like File, Edit, View, Search, Terminal, Help. The main pane shows a command-line interface where the user has just typed "[hadoop@dataserver ~]\$ ssh localhost" and pressed Enter. The cursor is positioned at the end of the command. The bottom status bar shows the same information as the title bar: "hadoop@dataserver:~".

ssh localhost

*Engenharia de Dados com Hadoop e Spark 3.0*

```
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:f6xshsQ3TV0jHuqn5c4eoJg7wm2t04Uo5lUK8BMzIPo.
ECDSA key fingerprint is MD5:8b:8e:f5:5c:cf:89:30:69:c5:17:e7:39:9a:5f:2a:c6.
Are you sure you want to continue connecting (yes/no)?
```

Yes

```
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:f6xshsQ3TV0jHuqn5c4eoJg7wm2t04Uo5lUK8BMzIPo.
ECDSA key fingerprint is MD5:8b:8e:f5:5c:cf:89:30:69:c5:17:e7:39:9a:5f:2a:c6.
Are you sure you want to continue connecting (yes/no)? yes
Warning: Permanently added 'localhost' (ECDSA) to the list of known hosts.
Last login: Fri Jun 28 21:09:29 2019
[hadoop@dataserver ~]$
```

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

A screenshot of a terminal window titled "DataServer Hadoop (Running)". The window shows the command "[hadoop@dataserver ~]\$ sudo gedit /etc/hosts" being typed. The terminal is part of a desktop environment with a menu bar at the top. The title bar also includes "Fri 21/25".

[hadoop@dataserver ~]\$ sudo gedit /etc/hosts

Editar o arquivo hosts

*Engenharia de Dados com Hadoop e Spark 3.0*A screenshot of a terminal window titled "Aplicativos" with "Locais" selected. The window shows a file named "hosts" in the "/etc" directory. The content of the file is:

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

Buttons for "Abrir", "Salvar", and other file operations are visible at the top.A screenshot of a terminal window titled "aluno@dataserver:~" with the current directory being "/etc". The window shows a file named "hosts" in the "/etc" directory. The content of the file is:

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

At the bottom of the terminal, the text "Incluir a última linha conforme acima" is displayed in bold black font.



5.3.2. Download do Hadoop

The screenshot shows a Firefox browser window displaying the Apache Hadoop homepage. The URL in the address bar is <https://www.apache.org>. The page features the Apache Hadoop logo and a brief description of the project. Below the description are three buttons: "Learn more", "Download", and "Getting started". The "Download" button is highlighted with a yellow background. At the bottom of the page, there are sections for "Latest news", "Modules", and "Related projects".

Acesse a página do Hadoop e clique em Download



Engenharia de Dados com Hadoop e Spark 3.0

DataServer Hadoop (Running)

Applications Places Firefox Fri 21:33

Apache Hadoop - Mozilla Firefox https://hadoop.apache.org/releases.html

Apache Hadoop Download Documentação Community Development Help Old site Apache Software Foundation

Download

Hadoop is released as source code tarballs with corresponding binary tarballs for convenience. The downloads are distributed via mirror sites and should be checked for tampering using GPG or SHA-256.

Version	Release date	Source download	Binary download	Release notes
3.1.2	2019 Feb 6	source (checksum signature)	binary (checksum signature)	Announcement
3.2.0	2019 Jan 16	source (checksum signature)	binary (checksum signature)	Announcement
2.9.2	2018 Nov 19	source (checksum signature)	binary (checksum signature)	Announcement
2.8.5	2018 Sep 15	source (checksum signature)	binary (checksum signature)	Announcement
2.7.7	2018 May 31	source (checksum signature)	binary (checksum signature)	Announcement

To verify Hadoop releases using GPG:

1. Download the release hadoop-X.Y.Z-src.tar.gz from a [mirror site](#).
2. Download the signature file hadoop-X.Y.Z-src.tar.gz.asc from Apache.
3. Download the Hadoop KEYS file.
4. gpg --import KEYS
5. gpg --verify hadoop-X.Y.Z-src.tar.gz.asc

To perform a quick check using SHA-256:

1. Download the release hadoop-X.Y.Z-src.tar.gz from a [mirror site](#).
2. Download the checksum hadoop-X.Y.Z-src.tar.gz.md5 from Apache.
3. shasum -a 256 hadoop-X.Y.Z-src.tar.gz

All previous releases of Hadoop are available from the [Apache release archive site](#).

Many third parties distribute products that include Apache Hadoop and related tools. Some of these are listed on the [Distributions wiki page](#).

License

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

DataServer Hadoop (Running)

Applications Places Terminal Fri 21:55

hadoop@dataserver:~/Downloads

```
[hadoop@dataserver ~]$ cd Downloads/
[hadoop@dataserver Downloads]$ tar -xvf hadoop-3.2.0.tar.gz
```

Descompacte o arquivo

*Engenharia de Dados com Hadoop e Spark 3.0*

```
DataServer Hadoop [Running]
Applications Places Terminal Fri 21:57
hadoop@dataserver:~/Downloads
File Edit View Search Terminal Help
[hadoop@dataserver Downloads]$ sudo mv hadoop-3.2.0 /opt/hadoop
```

Mover o diretório para /opt/hadoop

```
DataServer Hadoop [Running]
Applications Places Terminal Fri 21:58
hadoop@dataserver:-
File Edit View Search Terminal Help
[hadoop@dataserver ~]$ gedit .bashrc
```

Abrir o arquivo de profile do usuário hadoop

*Engenharia de Dados com Hadoop e Spark 3.0*

DataServer Hadoop (Running)

```

Applications Places Text Editor
Open .bashrc
Save Fri 22:03
# .bashrc

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

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

# User specific aliases and functions

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

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

```

Saving file "/home/hadoop/.bashrc"...

hadoop@dataserver: ~ bashrc: l-ri - sedt: sh Tab Width: 8 Lr 24, Col 53 INS 1 / 4

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

DataServer Hadoop (Running)

File Edit View Search Terminal Help

[hadoop@dataserver ~]\$ gedit .bashrc

[hadoop@dataserver ~]\$ source .bashrc

[hadoop@dataserver ~]\$

hadoop@dataserver: ~ bashrc: l-ri - sedt: sh Tab Width: 8 Lr 24, Col 53 INS 1 / 4

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

*Engenharia de Dados com Hadoop e Spark 3.0*A screenshot of a terminal window titled "DataServer Historico [Riminiagi]". The window shows the following command-line session:

```
[hadoop@dataserver ~]$ java -version
java version "1.8.0_212"
Java(TM) SE Runtime Environment (build 1.8.0_212-b10)
Java HotSpot(TM) 64-Bit Server VM (build 25.212-b10, mixed mode)
[hadoop@dataserver ~]$ hadoop version
Hadoop 3.2.0
Source code repository https://github.com/apache/hadoop.git -r e97acb3bd8f3befd27418996fa5d4b50bf2e17bf
Compiled by sunilg on 2019-01-08T06:08Z
Compiled with protoc 2.5.0
From source with checksum d3f0795ed0d9dc378e2c785d3668f39
This command was run using /opt/hadoop/share/hadoop/common/hadoop-common-3.2.0.jar
[hadoop@dataserver ~]$
```

Java e Hadoop instalados e configurados com sucesso!!!



5.4. Configuração do Hadoop

5.4.1. Editar arquivos de configuração do Hadoop

A screenshot of a Linux terminal window titled "DataServer Hadoop [Running]". The window shows a terminal session with the prompt "hadoop@dataserver:~\$". The user has just typed the command "cd /opt/hadoop/etc/hadoop" and is awaiting the results.

```
[hadoop@dataserver ~]$ cd /opt/hadoop/etc/hadoop/
```

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



Engenharia de Dados com Hadoop e Spark 3.0

```

DataServer-Hadoop [Running]
File Edit View Search Terminal Help
[hadoop@dataserver hadoop]$ gedit core-site.xml

[hadoop@dataserver /opt/hadoop/etc...]

```

Editar o arquivo core-site.xml

```

DataServer-Hadoop [Running]
File Edit View Search Terminal Help
Open core-site.xml /opt/hadoop/etc/hadoop Save
<?xml version="1.0" encoding="UTF-8"?>
<xsl:stylesheet type="text/xsl" href="configuration.xsl">
<!--
Licensed under the Apache License, Version 2.0 (the "License");
you may not use this file except in compliance with the License.
You may obtain a copy of the License at

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

Unless required by applicable law or agreed to in writing, software
distributed under the License is distributed on an "AS IS" BASIS,
WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
See the License for the specific language governing permissions and
limitations under the License. See accompanying LICENSE file.
-->
<!-- Put site-specific property overrides in this file. -->
<configuration>
<property>
<name>fs.defaultFS</name>
<value>hdfs://localhost:9000</value>
</property>
</configuration>

```

Saving file "/opt/hadoop/etc/hadoop/core-site.xml..."

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

*Engenharia de Dados com Hadoop e Spark 3.0*

DataServer Hadoop [Running]

File Edit View Search Terminal Help

[hadoop@dataserver hadoop]\$ gedit hdfs-site.xml

1 / 4

Editar o arquivo hdfs-site.xml

DataServer Hadoop [Running]

File Edit View Search Terminal Help

Open ... Save ...

hdfs-site.xml
/opt/hadoop/etc/hadoop/

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

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

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

<!-- Put site-specific property overrides in this file. -->
<configuration>
<property>
<name>dfs.replication</name>
<value>1</value>
</property>
</configuration>
```

Saving file "/opt/hadoop/etc/hadoop/hdfs-site.xml".

hadoop@dataserver:/opt/hadoop/etc/

1 / 4

Acrescentar as propriedades conforme acima e salvar o arquivo.

5.4.2. Formatando o Namenode

```
File Edit View Search Terminal Help
[hadoop@dataserver ~]$ hdfs namenode -format
```

hdfs namenode –format

```
File Edit View Search Terminal Help
[hadoop@dataserver ~]$ hdfs namenode -format
2019-06-28 22:13:00,489 INFO namenode.FSDirectory: ACLs enabled? false
2019-06-28 22:13:00,489 INFO namenode.FSDirectory: POSIX ACL inheritance enabled? true
2019-06-28 22:13:00,490 INFO namenode.FSDirectory: XAttrs enabled? true
2019-06-28 22:13:00,490 INFO namenode.NameNode: Caching file names occurring more than 10 times
2019-06-28 22:13:00,495 INFO snapshot.SnapshotManager: Loaded config captureOpenFiles: false, skipCaptureAccessTimeOnlyChange: false, snapshotDiffAllowSnapRootDescendant: true, maxSnapshotLimit: 65536
2019-06-28 22:13:00,496 INFO snapshot.SnapshotManager: Skiplist is disabled
2019-06-28 22:13:00,525 INFO util.GSet: Computing capacity for map cachedBlocks
2019-06-28 22:13:00,525 INFO util.GSet: VM type          = 64-bit
2019-06-28 22:13:00,526 INFO util.GSet: 0.25% max memory 1.8 GB = 4.7 MB
2019-06-28 22:13:00,526 INFO util.GSet: capacity        = 2^19 = 524288 entries
2019-06-28 22:13:00,532 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.window.num.buckets = 10
2019-06-28 22:13:00,532 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.num.users = 10
2019-06-28 22:13:00,532 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.windows.minutes = 1,5,25
2019-06-28 22:13:00,563 INFO namenode.FSNamesystem: Retry cache on namenode is enabled
2019-06-28 22:13:00,564 INFO namenode.FSNamesystem: Retry cache will use 0.03 of total heap and retry cache entry expiry time is 600000 millis
2019-06-28 22:13:00,566 INFO util.GSet: Computing capacity for map NameNodeRetryCache
2019-06-28 22:13:00,566 INFO util.GSet: VM type          = 64-bit
2019-06-28 22:13:00,566 INFO util.GSet: 0.02999999932944774% max memory 1.8 GB = 580.9 KB
2019-06-28 22:13:00,566 INFO util.GSet: capacity        = 2^16 = 65536 entries
2019-06-28 22:13:00,647 INFO namenode.FSIImage: Allocated new BlockPoolId: BP-1867265039-127.0.0.1-1561785180618
2019-06-28 22:13:00,734 INFO common.Storage: Storage directory /tmp/hadoop-hadoop/dfs/name has been successfully formatted.
2019-06-28 22:13:00,744 INFO namenode.FSIImageFormatProtobuf: Saving image file /tmp/hadoop-hadoop/dfs/name/current/fsimage.ckpt_00000000000000000000 using no compression
2019-06-28 22:13:00,902 INFO namenode.FSIImageFormatProtobuf: Image file /tmp/hadoop-hadoop/dfs/name/current/fsimage.ckpt_0000000000000000 of size 401 bytes saved in 0 seconds .
2019-06-28 22:13:00,914 INFO namenode.NNStorageRetentionManager: Going to retain 1 images with txid >= 0
2019-06-28 22:13:00,927 INFO namenode.NameNode: SHUTDOWN MSG:
*****
SHUTDOWN MSG: Shutting down NameNode at localhost/127.0.0.1
*****
```

Formatação realizada com sucesso



5.4.3. Iniciando o Hadoop

```
[hadoop@datavserver ~]$ start-dfs.sh
```

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

```
[hadoop@datavserver ~]$
```

*Engenharia de Dados com Hadoop e Spark 3.0*

Hadoop iniciado

A screenshot of a terminal window titled "DataServer Hadoop [Running]". The window shows a Linux desktop environment with a menu bar at the top. The terminal itself has a header "hadoop@dataserver:~\$". The command "jps" is run, and the output shows four processes: NameNode (8805), Jps (9293), SecondaryNameNode (9150), and DataNode (8959).

```
File Edit View Search Terminal Help  
[hadoop@dataserver ~]$ jps  
8805 NameNode  
9293 Jps  
9150 SecondaryNameNode  
8959 DataNode  
[hadoop@dataserver ~]$
```

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



5.4.4. Iniciando o Yarn

A screenshot of a terminal window titled "Terminal". The window shows the command `[aluno@dataserver ~]$ start-yarn.sh` being typed. The terminal is located on a desktop interface with other windows visible in the background.

start-yarn.sh

1 / 4

*Engenharia de Dados com Hadoop e Spark 3.0*A screenshot of a terminal window titled "Terminal". The window shows the command "start-yarn.sh" being run and its output. The output indicates that the yarn daemons are starting, including the resource manager and node manager, with logs directed to specific paths. The terminal window has a standard Linux-style interface with tabs for "Aplicativos" and "Locais", and status indicators at the top right.

```
Arquivo Editar Ver Pesquisar Terminal Ajuda
[aluno@dataserver ~]$ start-yarn.sh
starting yarn daemons
starting resourcemanager, logging to /opt/hadoop/logs/yarn-aluno-resourcemanager-dataserver.out
localhost: starting nodemanager, logging to /opt/hadoop/logs/yarn-aluno-nodemanager-dataserver.out
[aluno@dataserver ~]$
```

Yarn iniciado

1 / 4

*Engenharia de Dados com Hadoop e Spark 3.0*A screenshot of a terminal window titled "Terminal". The window shows the following command-line session:

```
aluno@dataserver:~$ start-yarn.sh
starting yarn daemons
starting resourcemanager, logging to /opt/hadoop/logs/yarn-aluno-resourcemanager-dataserver.out
localhost: starting nodemanager, logging to /opt/hadoop/logs/yarn-aluno-nodemanager-dataserver.out
[aluno@dataserver ~]$ jps
13329 NameNode
14037 ResourceManager
13782 SecondaryNameNode
13527 DataNode
14524 Jps
14189 NodeManager
[aluno@dataserver ~]$
```

The terminal window has a standard Linux-style interface with a menu bar at the top.

Checando os serviços com o comando **jps**



Engenharia de Dados com Hadoop e Spark 3.0

The screenshot shows the Hadoop Web UI interface. On the left, there's a sidebar with a navigation tree under 'Cluster' and a 'Scheduler' section. The main area displays 'Cluster Metrics' and 'Scheduler Metrics'. The 'Cluster Metrics' table has the following data:

	Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memory Total	Memory Reserved	Vcores Used	Vcores Total	Vcores Reserved
0	0	0	0	0	0 B	8 GB	0 B	0	8	8	0

The 'Scheduler Metrics' table shows the 'Capacity Scheduler' configuration with 'MEMORY' as the resource type and a minimum allocation of '<memory:1024, vcores:1>'. Below these tables is a table titled 'Show 20 entries' with columns: ID, User, Name, Application Type, Queue, StartTime, FinishTime, State, and FinalStatus. A message 'No data available in table' is displayed below the table.

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



5.5. Processando Big Data

A screenshot of a Linux terminal window titled 'Terminal'. The window shows the command 'hdfs dfs -mkdir /bigdata' being typed at the prompt 'aluno@dataserver:~\$'. The terminal interface includes a menu bar with 'Arquivo', 'Editar', 'Ver', 'Pesquisar', 'Terminal', and 'Ajuda'. The status bar at the bottom right shows 'en' and 'Sáb, 03:30'.

aluno@dataserver:~\$ hdfs dfs -mkdir /bigdata

Criar o diretório **bigdata** no HDFS

*Engenharia de Dados com Hadoop e Spark 3.0*A screenshot of a terminal window titled "Terminal". The window shows the command "hadoop fs -ls /" being run, which lists a single item: "/bigdata" with permissions "drwxr-xr-x", owner "aluno", group "supergroup", and creation date "0 2016-01-30 03:30".

```
Arquivo Editar Ver Pesquisar Terminal Ajuda
[aluno@dataserver ~]$ hadoop fs -ls /
Found 1 items
drwxr-xr-x  - aluno supergroup          0 2016-01-30 03:30 /bigdata
[aluno@dataserver ~]$
```

Listar o HDFS e checar o diretório criado

1 / 4

*Engenharia de Dados com Hadoop e Spark 3.0*

Bem vindo - Portal Brasileiro de Dados Abertos - Mozilla Firefox

dados.gov.br
PORTAL BRASILEIRO DE DADOS ABERTOS

Pesquisa ... PESQUISAR

em 1068 conjuntos de dados com 8685 recursos (o que é isto?)

Dados em destaque

- Compras públicas do governo federal**
Dados Abertos do Sistema Integrado de Administração e Serviços Gerais - SIASG. O SIASG é o sistema onde se operacionaliza as compras do Governo ...
- Ocorrências Aeronáuticas na Aviação Civil Brasileira**
A base de dados de ocorrências aeronáuticas é gerenciada pelo Centro de Investigação e Prevenção de Acidentes aeronáuticos (CENIPA). Constam nesta ...
- Lista de Eleitores Filados aos Partidos Políticos**
Conforme [provimento nº 04/2012](http://www.justicaeleitoral.jus.br/arquivos

Publicações mais recentes

Conjunto de dados	Autor	Quando
Aeroporto - 1º Balanço do PAC 2015	Sem Responsável	25 Jan
Cidades Digitais - 1º Balanço do PAC 2015	Sem Responsável	25 Jan
Centro de Iniciação ao Esporte - 1º ...	Sem Responsável	25 Jan
Base Cartográfica Contínua do Brasil - ...	Diretoria de ...	15 Dez
REGIÃO DE INFORMAÇÃO DE VOO - FIR	Divisão de Operações	15 Dez

aluno@dataserver:~ Bem vindo - Portal Brasileiro de D... 1 / 4

Acessar o portal de dados abertos do governo federal



Engenharia de Dados com Hadoop e Spark 3.0

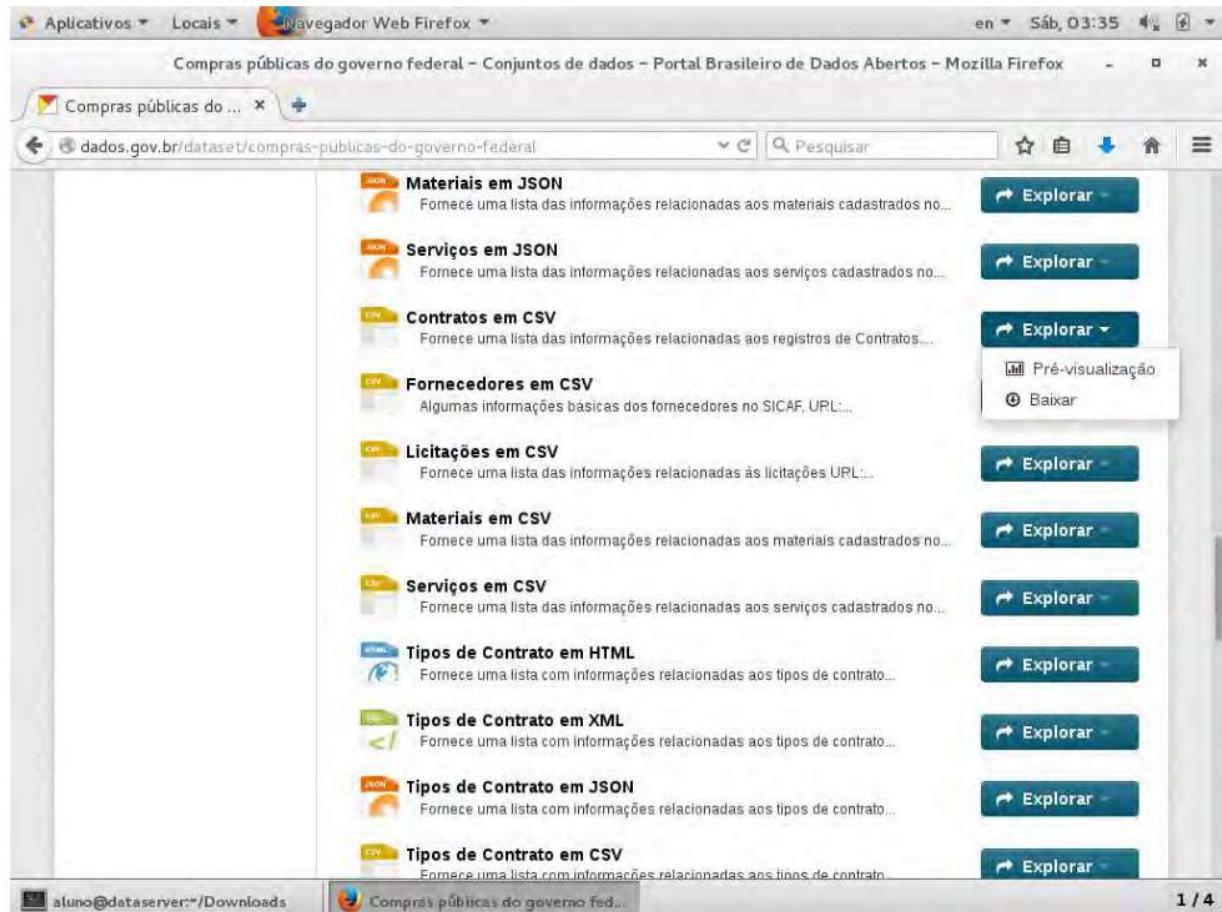
The screenshot shows the homepage of the Brazilian Open Data Portal (dados.gov.br). The page has a green header with the portal's name and navigation links like 'BRASIL', 'Acesso à informação', 'Participe', 'Serviços', 'Legislação', and 'Canais'. Below the header, there's a search bar and a 'Pesquisar' button. A sidebar on the left lists 'Dados em destaque' with links to 'Compras públicas do governo federal', 'Ocorrências Aeronáuticas na Aviação Civil Brasileira', and 'Lista de Eleitores Filiados aos Partidos Políticos'. The main content area features a section titled 'Publicações mais recentes' with a table showing five recent publications:

Conjunto de dados	Autor	Quando
Aeroporto - 1º Balanço do PAC 2015	Sem Responsável	25 Jan
Cidades Digitais - 1º Balanço do PAC 2015	Sem Responsável	25 Jan
Centro de Iniciação ao Esporte - 1º ...	Sem Responsável	25 Jan
Base Cartográfica Contínua do Brasil - ...	Diretoria de ...	15 Dez
REGIÃO DE INFORMAÇÃO DE VOO - FIR	Divisão de Operações	15 Dez

At the bottom, there's a footer with user information ('aluno@dataserver...') and a link to the portal ('Bem Vindo - Portal Brasileiro de D...').

Clicar no link de compras públicas

Engenharia de Dados com Hadoop e Spark 3.0



Compras públicas do governo federal – Conjuntos de dados – Portal Brasileiro de Dados Abertos – Mozilla Firefox

dados.gov.br/dataset/compras-publicas-do-governo-federal

Materiais em JSON
Fornecce uma lista das informações relacionadas aos materiais cadastrados no...

Serviços em JSON
Fornecce uma lista das informações relacionadas aos serviços cadastrados no...

Contratos em CSV
Fornecce uma lista das informações relacionadas aos registros de Contratos....

Fornecedores em CSV
Algumas informações básicas dos fornecedores no SICAF. URL:...

Licitações em CSV
Fornecce uma lista das informações relacionadas às licitações URL:...

Materiais em CSV
Fornecce uma lista das informações relacionadas aos materiais cadastrados no...

Serviços em CSV
Fornecce uma lista das informações relacionadas aos serviços cadastrados no...

Tipos de Contrato em HTML
Fornecce uma lista com informações relacionadas aos tipos de contrato...

Tipos de Contrato em XML
Fornecce uma lista com informações relacionadas aos tipos de contrato...

Tipos de Contrato em JSON
Fornecce uma lista com informações relacionadas aos tipos de contrato...

Tipos de Contrato em CSV
Fornecce uma lista com informações relacionadas aos tipos de contrato...

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

*Engenharia de Dados com Hadoop e Spark 3.0*

```
aluno@dataserver:~/Downloads$ cd Downloads/
[aluno@dataserver Downloads]$ ls -la
total 384432
drwx----- 2 aluno aluno 84 Jan 30 03:35 ..
drwx----- 10 aluno aluno 4096 Jan 30 03:09 ..
-rw-rw-r-- 1 aluno aluno 303071 Jan 30 03:33 contratos.csv
-rw-rw-r-- 1 aluno aluno 212046774 Jan 25 23:29 hadoop-2.7.2.tar.gz
-rw-rw-r-- 1 aluno aluno 181299489 Jan 30 01:18 jdk-8u71-linux-x64.tar.gz
[aluno@dataserver Downloads]$ hadoop fs -copyFromLocal contratos.csv /bigdata
```



Copiar o arquivo para a pasta bigdata no HDFS

*Engenharia de Dados com Hadoop e Spark 3.0*A screenshot of a Linux desktop environment showing a terminal window. The window title is "Terminal". The terminal prompt is "aluno@dataserver:~/Downloads". The user has run several commands to copy a CSV file from their local Downloads directory to a HDFS directory named "bigdata".

```
Arquivo Editar Ver Pesquisar Terminal Ajuda
[aluno@dataserver ~]$ cd Downloads/
[aluno@dataserver Downloads]$ ls -la
total 384432
drwx----- 2 aluno aluno 84 Jan 30 03:35 ..
drwx----- 10 aluno aluno 4096 Jan 30 03:09 ..
-rw-rw-r-- 1 aluno aluno 303071 Jan 30 03:33 contratos.csv
-rw-rw-r-- 1 aluno aluno 212046774 Jan 25 23:29 hadoop-2.7.2.tar.gz
-rw-rw-r-- 1 aluno aluno 181299489 Jan 30 01:18 jdk-8u71-linux-x64.tar.gz
[aluno@dataserver Downloads]$ hadoop fs -copyFromLocal contratos.csv /bigdata
[aluno@dataserver Downloads]$ hadoop fs -ls /bigdata
```

Listar o diretório bigdata

1 / 4

*Engenharia de Dados com Hadoop e Spark 3.0*A screenshot of a Linux terminal window titled "Terminal". The window shows a command-line session where a user named "aluno" is navigating through their Downloads directory. They run an "ls -la" command to list files, then use "hadoop fs -copyFromLocal" to move a CSV file named "contratos.csv" to a directory named "bigdata". Finally, they use "hadoop fs -cat" to view the contents of the file. The terminal interface includes a menu bar with "Arquivo", "Editar", "Ver", "Pesquisar", "Terminal", and "Ajuda".

```
aluno@dataserver:~/Downloads
[aluno@dataserver ~]$ cd Downloads/
[aluno@dataserver Downloads]$ ls -la
total 384432
drwx----- 2 aluno aluno 84 Jan 30 03:35 ..
drwx----- 10 aluno aluno 4096 Jan 30 03:09 ..
-rw-rw-r-- 1 aluno aluno 303071 Jan 30 03:33 contratos.csv
-rw-rw-r-- 1 aluno aluno 212046774 Jan 25 23:29 hadoop-2.7.2.tar.gz
-rw-rw-r-- 1 aluno aluno 181299489 Jan 30 01:18 jdk-8u71-linux-x64.tar.gz
[aluno@dataserver Downloads]$ hadoop fs -copyFromLocal contratos.csv /bigdata
[aluno@dataserver Downloads]$ hadoop fs -ls /bigdata
Found 1 items
-rw-r--r-- 1 aluno supergroup 303071 2016-01-30 03:36 /bigdata/contratos.csv
[aluno@dataserver Downloads]$ hadoop fs -cat /bigdata/contratos.csv
```



Engenharia de Dados com Hadoop e Spark 3.0

```

Aplicativos Locais Terminal en Sáb, 03:38
aluno@dataserver:~/Downloads

Arquivo Editar Ver Pesquisar Terminal Ajuda

25 distrito do departamento nacional de producao mineral no estado de alagoas.,1,48425000080/97,,Fornecedor 70.0
05.657/0001-27: DINAMICA SERVICOS GERAIS LTDA,03/11/1997,lei/8.666/93.,03/11/1997,02/11/1998,"R$ 9.708,96",/cont
ratos/id/contrato/32302750000011997/aditivos,/contratos/id/contrato/32302750000011997/apostilamentos,/contratos/
id/contrato/32302750000011997/eventos
20006350000011997,200063: MJ-DPF-SUPERINTENDENCIA REGIONAL/RS,2: TOMADA DE PREÇOS,000019/1997,50; CONTRATO,Licit
ação 20006302000191997,,00001/1997,Contratação de empresa para realização de serviço de manutenção pre ventiva e
corretiva nos equipamentos de informática pertencentes a SR/DPF/R S e suas Delegacias descentralizadas,2,084300
11006/97-04,,Fornecedor 82.885,112/0001-31; VR COMPUTADORES LTDA,01/01/1998,"Art. 22, II da Lei 8.666/93",01/01/
1998,31/12/1998,"R$ 132,000,00",/contratos/id/contrato/20006350000011997/aditivos,/contratos/id/contrato/2000635
0000011997/apostilamentos,/contratos/id/contrato/20006350000011997/eventos
20009257000011997,200092: SUPERINTENDENCIA REG.DEF.POLICIA FEDERAL- FE,,0,57; CONVÉNIO,Licitação 200092null00000
0000,,00001/1997,"Prorrogação da vigência do Convênio de Cooperação recíproca entre as partes conveniadas, visan
do o desenvolvimento de atividades conjuntas relacionadas ao estágio de estudantes",0,082000141519643,,,24/05/1
999,"Decreto nº 87.497/82 e suas alterações; IN/SAF nº 07/92, alterada pelas IN/SAF nºs 01/93 e 06/94 e Lei nº 8.
666/93.",,24/05/1999,23/05/2000,"R$ 150,000,00",/contratos/id/contrato/20009257000011997/aditivos,/contratos/id/c
ontrato/20009257000011997/apostilamentos,/contratos/id/contrato/20009257000011997/eventos
15301054000011997,153010: MEC-CEFET-CENT.FED.ED.TEC.CELSO S.FONSECA/RJ,3; CONCORRENÇIA,00003/1997,54; CONCESSÃO,
Licitação 15301003000031997,,00001/1997,"Concessão de uso para instalação de 12 (doze ) outdoors, no tamanho 3m
x 9m mediante remuneração na testada dos muros existentes não incluindo a ocupação interna.",5,23063001327/97-84
,,Fornecedor 29.248.390/0001-08: Klimes Rio Propaganda au Ar LivreLtda,15/12/1997,lei 8987/85 e lei 8666/93 e su
as atualizações,15/12/1997,15/12/1998,"R$ 3.000,00",/contratos/id/contrato/15301054000011997/aditivos,/contratos
/id/contrato/15301054000011997/apostilamentos,/contratos/id/contrato/15301054000011997/eventos
25442050000011997,254420: FUNDACAO OSWALDO CRUZ/RJ,4: CONCORRÊNCIA INTERNACIONAL,00008/1996,50: CONTRATO,Licitaç
ão 25442004000081996,,00001/1997,Pretação de serviços de operação do Espaço Museu da vida da Fiocru z,4,25380011
6889663,,Fornecedor 31.880.164/0001-84: HOFE-CONSULTORIA DE RECURSOS HUMANOS LTDA,14/01/1997,Artigo 62 da Lei 8
.666/93,14/01/1997,14/01/1998,"R$ 1.485,188,05",/contratos/id/contrato/25442050000011997/aditivos,/contratos/id/
contrato/25442050000011997/apostilamentos,/contratos/id/contrato/25442050000011997/eventos
19311250000011997,193112: IBAMA-SUPERINTENDENCIA ESTADUAL/MS,6: DISPENSA DE LICITAÇÃO,00001081/1997,50; CONTRATO
,Licitação 19311205010811997,,00001/1997,"Locação de imóvel situado à Rua Paranaiba, 272, centro, Três Lagoas MS
, que a LOCADORA entregará ao LOCATÁRIO em perfeito estado de conservação e asseio, livre e desembaraçado de qua
lquer ônus judicial ou extrajudicial, para sua utilização.",4,02014001081/97-74,***546088**,,01/08/1997,Inciso X
do Art. 24 da Lei 8.666/93,01/08/1997,31/07/1998,"R$ 8.400,00",/contratos/id/contrato/19311250000011997/aditivo
s,/contratos/id/contrato/19311250000011997/apostilamentos,/contratos/id/contrato/19311250000011997/eventos
25003850000011997,250038: GERENCIA ESTADUAL EM SERGIP/MS/SE,2: TOMADA DE PREÇOS,00001/1997,50: CONTRATO,Licitaç
ão 25003802000011997,,00001/1997,Prestacao de servicos de fornecimento de passagens aereas domesticas,7,333591/
0009141/97,,Fornecedor 32.705.949/0001-83: PONTAL TURISMO LTDA,23/05/1997,"Lei 8666/93, alterada pela lei 8883/9
3.",,23/05/1997,31/12/1997,"R$ 53.353,60",/contratos/id/contrato/25003850000011997/aditivos,/contratos/id/contrat
o/25003850000011997/apostilamentos,/contratos/id/contrato/25003850000011997/eventos
[aluno@dataserver Downloads]$
```



Conteúdo do arquivo já gravado no HDFS



Engenharia de Dados com Hadoop e Spark 3.0

```
aluno@dataserver:~$ cd ~
aluno@dataserver ~]$ cd /opt/hadoop/
aluno@dataserver hadoop]$ ls -la
total 48
drwxr-xr-x. 10 aluno aluno 4096 Jan 30 03:22 .
drwxr-xr-x. 14 root root 4096 Jan 30 02:36 ..
drwxr-xr-x. 2 aluno aluno 4096 Jan 25 22:20 bin
drwxr-xr-x. 3 aluno aluno 19 Jan 25 22:20 etc
drwxr-xr-x. 2 aluno aluno 101 Jan 25 22:20 include
drwxr-xr-x. 3 aluno aluno 19 Jan 25 22:20 lib
drwxr-xr-x. 2 aluno aluno 4096 Jan 25 22:20 libexec
-rw-r--r--. 1 aluno aluno 15429 Jan 25 22:20 LICENSE.txt
drwxrwxr-x. 3 aluno aluno 4096 Jan 30 03:25 logs
-rw-r--r--. 1 aluno aluno 101 Jan 25 22:20 NOTICE.txt
-rw-r--r--. 1 aluno aluno 1366 Jan 25 22:20 README.txt
drwxr-xr-x. 2 aluno aluno 4096 Jan 25 22:20 sbin
drwxr-xr-x. 4 aluno aluno 29 Jan 25 22:20 share
[aluno@dataserver hadoop]$ hadoop jar share/hadoop/mapreduce/hadoop-mapreduce-examples-2.7.2.jar wordcount /bigdata/contratos.csv /output
```

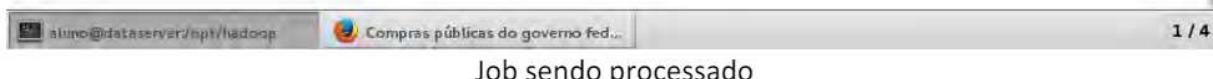
A instalação do Hadoop possui um job chamado wordcount, que pode ser usado como exemplo para processamento de Big Data. Basicamente, o job conta a ocorrência de cada palavra no arquivo. Vamos executar com o comando acima (a versão do arquivo .jar é a mesma versão do Hadoop que você instalou).

*Engenharia de Dados com Hadoop e Spark 3.0*

```

Arquivo Editar Ver Pesquisar Terminal Ajuda
[aluno@dataserver hadoop]$ cd ~
[aluno@dataserver ~]$ cd /opt/hadoop/
[aluno@dataserver hadoop]$ ls -la
total 48
drwxr-xr-x. 10 aluno aluno 4096 Jan 30 03:22 .
drwxr-xr-x. 14 root root 4096 Jan 30 02:36 ..
drwxr-xr-x. 2 aluno aluno 4096 Jan 25 22:20 bin
drwxr-xr-x. 3 aluno aluno 19 Jan 25 22:20 etc
drwxr-xr-x. 2 aluno aluno 101 Jan 25 22:20 include
drwxr-xr-x. 3 aluno aluno 19 Jan 25 22:20 lib
drwxr-xr-x. 2 aluno aluno 4096 Jan 25 22:20 libexec
-rw-r--r--. 1 aluno aluno 15429 Jan 25 22:20 LICENSE.txt
drwxrwxr-x. 3 aluno aluno 4096 Jan 30 03:25 logs
-rw-r--r--. 1 aluno aluno 101 Jan 25 22:20 NOTICE.txt
-rw-r--r--. 1 aluno aluno 1366 Jan 25 22:20 README.txt
drwxr-xr-x. 2 aluno aluno 4096 Jan 25 22:20 sbin
drwxr-xr-x. 4 aluno aluno 29 Jan 25 22:20 share
[aluno@dataserver hadoop]$ hadoop jar share/hadoop/mapreduce/hadoop-mapreduce-examples-2.7.2.jar wordcount /bigdata/contratos.csv /output
16/01/30 03:40:53 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
16/01/30 03:40:54 INFO input.FileInputFormat: Total input paths to process : 1
16/01/30 03:40:54 INFO mapreduce.JobSubmitter: number of splits:1
16/01/30 03:40:54 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1454131523182_0001
16/01/30 03:40:55 INFO impl.YarnClientImpl: Submitted application application_1454131523182_0001
16/01/30 03:40:55 INFO mapreduce.Job: The url to track the job: http://localhost:8088/proxy/application_1454131523182_0001/
16/01/30 03:40:55 INFO mapreduce.Job: Running job: job_1454131523182_0001
16/01/30 03:41:04 INFO mapreduce.Job: Job job_1454131523182_0001 running in uber mode : false
16/01/30 03:41:04 INFO mapreduce.Job: map 0% reduce 0%
16/01/30 03:41:10 INFO mapreduce.Job: map 100% reduce 0%

```



*Engenharia de Dados com Hadoop e Spark 3.0*

```

Aplicativos Locais Terminal en Sáb, 03:41
aluno@dataserver:~/opt/hadoop

Arquivo Editar Ver Pesquisar Terminal Ajuda
16/01/30 03:40:55 INFO impl.YarnClientImpl: Submitted application application_1454131523182_0001
16/01/30 03:40:55 INFO mapreduce.Job: The url to track the job: http://localhost:8088/proxy/application_1454131523182_0001/
16/01/30 03:40:55 INFO mapreduce.Job: Running job: job_1454131523182_0001
16/01/30 03:41:04 INFO mapreduce.Job: Job job_1454131523182_0001 running in uber mode : false
16/01/30 03:41:04 INFO mapreduce.Job: map 0% reduce 0%
16/01/30 03:41:10 INFO mapreduce.Job: map 100% reduce 0%
16/01/30 03:41:16 INFO mapreduce.Job: map 100% reduce 100%
16/01/30 03:41:18 INFO mapreduce.Job: Job job_1454131523182_0001 completed successfully
16/01/30 03:41:18 INFO mapreduce.Job: Counters: 49
  File System Counters
    FILE: Number of bytes read=266936
    FILE: Number of bytes written=768711
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=308179
    HDFS: Number of bytes written=234650
    HDFS: Number of read operations=6
    HDFS: Number of large read operations=0
    HDFS: Number of write operations=2
  Job Counters
    Launched map tasks=1
    Launched reduce tasks=1
    Data-local map tasks=1
    Total time spent by all maps in occupied slots (ms)=3572
    Total time spent by all reduces in occupied slots (ms)=3848
    Total time spent by all map tasks (ms)=3572
    Total time spent by all reduce tasks (ms)=3848
    Total vcore-milliseconds taken by all map tasks=3572
    Total vcore-milliseconds taken by all reduce tasks=3848
    Total megabyte-milliseconds taken by all map tasks=3657728
    Total megabyte-milliseconds taken by all reduce tasks=3940352
  Map-Reduce Framework
    Map input records=501
    Map output records=19752
    Map output bytes=382199
    Map output materialized bytes=266936
    Input split bytes=108

```

aluno@dataserver:~/opt/hadoop Compras públicas do governo fed... 1 / 4

job processado com sucesso



Engenharia de Dados com Hadoop e Spark 3.0

```
A Aplicativos Locais Terminal en Sáb, 03:41
aluno@dataserver:/opt/hadoop
Arquivo Editar Ver Pesquisar Terminal Ajuda
Total time spent by all map tasks (ms)=3572
Total time spent by all reduce tasks (ms)=3848
Total vcore-milliseconds taken by all map tasks=3572
Total vcore-milliseconds taken by all reduce tasks=3848
Total megabyte-milliseconds taken by all map tasks=3657728
Total megabyte-milliseconds taken by all reduce tasks=3940352
Map-Reduce Framework
Map input records=501
Map output records=19752
Map output bytes=382199
Map output materialized bytes=266936
Input split bytes=108
Combine input records=19752
Combine output records=7874
Reduce input groups=7874
Reduce shuffle bytes=266936
Reduce input records=7874
Reduce output records=7874
Spilled Records=15748
Shuffled Maps =1
Failed Shuffles=0
Merged Map outputs=1
GC time elapsed (ms)=115
CPU time spent (ms)=1280
Physical memory (bytes) snapshot=315006976
virtual memory (bytes) snapshot=4161437696
Total committed heap usage (bytes)=219676672
Shuffle Errors
BAD_ID=0
CONNECTION=0
IO_ERROR=0
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0
File Input Format Counters
Bytes Read=303071
File Output Format Counters
Bytes Written=234650
[aluno@dataserver hadoop]$
```

job processado com sucesso

*Engenharia de Dados com Hadoop e Spark 3.0*A screenshot of a terminal window titled "Terminal". The window shows the command "hdfs dfs -cat /output/*" being run by the user "aluno@dataserver:opt/hadoop". The terminal interface includes a menu bar with "Arquivo", "Editar", "Ver", "Pesquisar", "Terminal", and "Ajuda". The status bar at the bottom right indicates "en" and "Sáb, 03:42".

```
aluno@dataserver:opt/hadoop
[aluno@dataserver hadoop]$ hdfs dfs -cat /output/*
```

Vamos ver o resultado do processamento



Engenharia de Dados com Hadoop e Spark 3.0

```
A Aplicativos Locais Terminal en Sáb, 03:43
aluno@dataserver:~/opt/hadoop
Arquivo Editar Ver Pesquisar Terminal Ajuda
x" ,06/08/1992,05/08/1993,"R$      1
x-5021, 1
xerograficas, 1
xerox 1
xerox.",2,53670.000313/96,,Fornecedor 1
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx",3,10469004898/94-13,,Fornecedor 1
z,4,253800116589663,,Fornecedor 1
zada 1
zo 1
zona 1
S2º.",02/03/1996,01/03/2001,"R$ 1
'b' 1
Á 2
ÁREA 3
Áreas.",4,37041000389/97-87,,Fornecedor 1
Órgãos 2
Único.",05/12/1997,28/02/1998,"R$      1
Único.",06/11/1996,06/11/1997,"R$      1
Útil 1
à 36
ás 6
água 5
área 15
áreas 9
âmbito 10
ão 1
ças 2
ças,5,46217.00280/97,,Fornecedor 1
ção 3
ção, 1
óleo 1
Órgãos 1
ônibus 1
ônus 1
Único, 1
útil 1
[aluno@dataserver hadoop]$
```

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

*Engenharia de Dados com Hadoop e Spark 3.0*

Aplicativos Locais Navegador Web Firefox

Namenode information – Mozilla Firefox

dataserver:50070/dfshealth.html#tab-overview

Pesquisar

Hadoop Overview DataNodes DataNode Volume Failures Snapshot Startup Progress Utilities

Overview 'localhost:9000' (active)

Started:	Sat Jan 30 03:22:42 BRST 2016
Version:	2.7.2, rb165c4fe8a74265c792ce23f546c64604acf0e41
Compiled:	2016-01-26T00:08Z by jenkins from (detached from b165c4f)
Cluster ID:	CID-45278722-9d78-4cc1-a53a-03d72c15a266
Block Pool ID:	BP-1791171697-127.0.0.1-1454181094108

Security is off.
Safemode is off.
17 files and directories, 5 blocks = 22 total filesystem object(s).

aluno@dataserver:/opt/hadoop

Namenode information - Mozilla F...

1 / 4

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

*Engenharia de Dados com Hadoop e Spark 3.0*

A screenshot of a Mozilla Firefox browser window. The title bar says "DataNode Information - Mozilla Firefox". The address bar shows "dataserver:50075". The main content area has a green header bar with the text "Hadoop" and "Overview". Below the header, the text "DataNode on dataserver:50075" is displayed.

DataNode on `dataserver:50075`

Hadoop, 2015.

A screenshot showing two windows side-by-side. On the left is a terminal window titled "aluno@dataserver:~\$" with the command "ls" entered. On the right is a Mozilla Firefox browser window titled "DataNode Information - Mozilla F...". The address bar shows "dataserver:50075". The main content area displays the same "DataNode on dataserver:50075" information as the previous screenshot. A status bar at the bottom indicates "1 / 4".

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



SecondaryNamenode Information – Mozilla Firefox

Screenshot of Mozilla Firefox browser window showing the SecondaryNamenode Information page for Hadoop. The URL in the address bar is <http://dataserver:50090/status.html>. The page title is "SecondaryNamenode Information". The main content area shows a table with the following data:

Version	2.7.2
Compiled	2016-01-26T00:08Z by jenkins from (detached from b165c4f)
NameNode Address	localhost:9000
Started	30/01/2016 03:22:54
Last Checkpoint	31/12/1969 22:00:17
Checkpoint Period	3600 seconds
Checkpoint Transactions	1000000

Below the table, there is a section titled "Checkpoint Image URI" with the value "file:///tmp/hadoop-aluno/dfs/namesecondary".

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



The screenshot shows a Firefox browser window with the title "Apache ZooKeeper - Releases - Mozilla Firefox". The address bar displays "zookeeper.apache.org/releases.html". The main content area shows the Apache ZooKeeper™ Releases page. At the top, there is a cartoon illustration of a character holding a broom next to a large feather. Below the illustration, the text "Apache ZooKeeper™ Releases" is prominently displayed. To the right of the title, there is a search bar with the placeholder "Search with Apache Solr" and a "Search" button. On the left side of the page, there is a sidebar with a "Project" section containing links to News, Releases, Wiki, Credits, Bylaws, License, Privacy Policy, Sponsorship, Security, and Thanks. There is also a "Subprojects" section with a link to BookKeeper (with Hedwig). On the right side, there is a "Documentation" section with links to Release 3.5.2-alpha, Release 3.5.1-alpha, Release 3.5.0-alpha, Release 3.4.9(stable), Release 3.4.9(current), and Release 3.4.8. The bottom of the page shows a navigation bar with the text "Apache ZooKeeper - Releases - Mozilla Firefox" and "1 / 4".

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

A screenshot of a terminal window titled "Terminal". The window shows the command "mkdir /opt/zookeeper/data" being typed by the user "aluno@dataserver:~".

```
aluno@dataserver:~$ mkdir /opt/zookeeper/data
```

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

*Engenharia de Dados com Hadoop e Spark 3.0*

```
aluno@dataserver:~$ cd /opt/zookeeper/conf/
[aluno@dataserver conf]$
```

Acessar o diretório /opt/zookeeper/conf

1 / 4

*Engenharia de Dados com Hadoop e Spark 3.0*

```
aluno@dataserver:~$ cd /opt/zookeeper/conf/  
[aluno@dataserver conf]$ cp zoo_sample.cfg zoo.cfg
```

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

*Engenharia de Dados com Hadoop e Spark 3.0*

```
aluno@dataserver:/opt/zookeeper/conf$ cd /opt/zookeeper/conf/
[aluno@dataserver conf]$ cp zoo_sample.cfg zoo.cfg
[aluno@dataserver conf]$ gedit zoo.cfg
```

The terminal window shows the user navigating to the configuration directory, copying the sample configuration file, and then opening it with the gedit text editor.

1 / 4

Editar o arquivo zoo.cfg

The browser window displays the contents of the 'zoo.cfg' file, which is currently being edited.

*Engenharia de Dados com Hadoop e Spark 3.0*

A Aplicativos Locais gedit

zoo.cfg
/opt/zookeeper/conf

en Ter, 00:46 Salvar

```
# The number of milliseconds of each tick
tickTime=2000

# The number of ticks that the initial
# synchronization phase can take
initLimit=5

# The number of ticks that can pass between
# sending a request and getting an acknowledgement
syncLimit=2

# the directory where the snapshot is stored,
# do not use /tmp for storage, /tmp here is just
# example sakes.
dataDir=/opt/zookeeper/data

# the port at which the clients will connect
clientPort=2181

# the maximum number of client connections.
# increase this if you need to handle more clients
#maxClientCnxns=60
#
# Be sure to read the maintenance section of the
# administrator guide before turning on autopurge.
#
# http://zookeeper.apache.org/doc/current/zookeeperAdmin.html#sc_maintenance
#
# The number of snapshots to retain in dataDir
#autopurge.snapRetainCount=3
# Purge task interval in hours
# Set to "0" to disable auto purge feature
#autopurge.purgeInterval=1
```

aluno@dataserver:/opt/zookeeper... Index of /zookeeper/stable - Mozilla/5.0... zoo.cfg /opt/zookeeper/conf - g...

1 / 4

Editar o arquivo conforme tela acima

*Engenharia de Dados com Hadoop e Spark 3.0*

```
aluno@dataserver:~$ mkdir /opt/zookeeper/data
[aluno@dataserver ~]$ gedit .bashrc
```



Incluir variáveis Zookeeper no /home/hadoop/.bashrc

*Engenharia de Dados com Hadoop e Spark 3.0*

Aplicativos Locais gedit * .bashrc

Abrir Salvar - x

```
# .bashrc

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

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

# User specific aliases and functions

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

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

# Zookeeper
export ZOOKEEPER_HOME=/opt/zookeeper
export PATH=$PATH:$ZOOKEEPER_HOME/bin
```

sh Largura da tabulação: 8 Lin 31, Col 1 INS

aluno@dataserver:~ Mozilla Firefox * .bashrc (~) - gedit 1 / 4

Variáveis Zookeeper

*Engenharia de Dados com Hadoop e Spark 3.0*

A screenshot of a Linux desktop environment showing a terminal window. The terminal title bar says "Terminal". The terminal window shows the following command-line session:

```
aluno@dataserver:~$ mkdir /opt/zookeeper/data
[aluno@dataserver ~]$ gedit .bashrc
[aluno@dataserver ~]$ source .bashrc
[aluno@dataserver ~]$
```

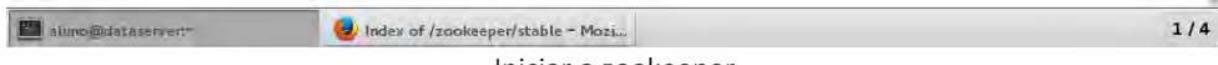


Data Science
Academy

Data Science Academy phelipe.utsemprboni@outlook.com 5c8a62005e4cde1acb8b45a3

Engenharia de Dados com Hadoop e Spark 3.0

A screenshot of a terminal window titled "Terminal". The window has a menu bar with "Aplicativos", "Locais", and "Terminal". The status bar shows "en Ter, 00:49". The terminal prompt is "aluno@dataserver:~". Below the prompt, the command "[aluno@dataserver ~]\$ zkServer.sh start" is visible. The window is set against a light gray background.



*Engenharia de Dados com Hadoop e Spark 3.0*A screenshot of a terminal window titled "Terminal". The window shows the command "zkServer.sh start" being run by the user "aluno@dataserver". The output indicates that JMX is enabled by default, using configuration file "/opt/zookeeper/bin/../conf/zoo.cfg", and the zookeeper service has started successfully.

```
Arquivo Editar Ver Pesquisar Terminal Ajuda
[aluno@dataserver ~]$ zkServer.sh start
JMX enabled by default
Using config: /opt/zookeeper/bin/../conf/zoo.cfg
Starting zookeeper ... STARTED
[aluno@dataserver ~]$
```



Serviço iniciado

*Engenharia de Dados com Hadoop e Spark 3.0*

A screenshot of a terminal window titled "Terminal". The window has a menu bar with "Aplicativos", "Locais", and "Terminal". The status bar shows "en Ter, 00:50". The terminal prompt is "aluno@dataserver:~". Below the prompt, the text "[aluno@dataserver ~]\$ zkCli.sh" is visible. The rest of the window is blank, indicating no output from the command.

Iniciar o Zookeeper Command Line Interface (CLI)

*Engenharia de Dados com Hadoop e Spark 3.0*

```

Arquivo Editar Ver Pesquisar Terminal Ajuda
.jar:/opt/hive/lib/geronimo-jaspic_1.0_spec-1.0.jar:/opt/hive/lib/geronimo-annotation_1.0_spec-1.1.1.jar:/opt/hive/lib/asm-commons-3.1.jar:/opt/hive/lib/asm-tree-3.1.jar:/opt/hive/lib/curator-recipes-2.6.0.jar:/opt/hive/lib/hive-jdbc-1.2.1.jar:/opt/hive/lib/hive-jdbc-1.2.1-standalone.jar:/opt/hive/lib/hive-beeline-1.2.1.jar:/opt/hive/lib/super-csv-2.2.0.jar:/opt/hive/lib/hive-cli-1.2.1.jar:/opt/hive/lib/hive-contrib-1.2.1.jar:/opt/hive/lib/hive-hbase-handler-1.2.1.jar:/opt/hive/lib/hive-hwi-1.2.1.jar:/opt/hive/lib/jetty-all-server-7.6.0.v20120127.jar:/opt/hive/lib/hive/hive-accumulo-handler-1.2.1.jar:/opt/hive/lib/accumulo-core-1.6.0.jar:/opt/hive/lib/jcommander-1.32.jar:/opt/hive/lib/commons-configuration-1.6.jar:/opt/hive/lib/commons-digester-1.8.jar:/opt/hive/lib/commons-beanutils-1.7.0.jar:/opt/hive/lib/commons-beanutils-core-1.8.0.jar:/opt/hive/lib/accumulo-fate-1.6.0.jar:/opt/hive/lib/accumulo-start-1.6.0.jar:/opt/hive/lib/commons-vfs2-2.0.jar:/opt/hive/lib/maven-scm-api-1.4.jar:/opt/hive/lib/plexus-utils-1.5.6.jar:/opt/hive/lib/maven-scm-provider-svnexe-1.4.jar:/opt/hive/lib/maven-scm-provider-svn-commons-1.4.jar:/opt/hive/lib/regexp-1.3.jar:/opt/hive/lib/accumulo-trace-1.6.0.jar:/opt/hive/lib/commons-math-2.1.jar:
2016-02-02 00:50:53,863 [myid:] - INFO [main:Environment@100] - Client environment:java.library.path=/usr/java/packages/lib/amd64:/usr/lib64:/lib64:/lib:/usr/lib
2016-02-02 00:50:53,863 [myid:] - INFO [main:Environment@100] - Client environment:java.io.tmpdir=/tmp
2016-02-02 00:50:53,864 [myid:] - INFO [main:Environment@100] - Client environment:java.compiler=<NA>
2016-02-02 00:50:53,864 [myid:] - INFO [main:Environment@100] - Client environment:os.name=Linux
2016-02-02 00:50:53,864 [myid:] - INFO [main:Environment@100] - Client environment:os.arch=amd64
2016-02-02 00:50:53,864 [myid:] - INFO [main:Environment@100] - Client environment:os.version=3.10.0-327.4.5.el7.x86_64
2016-02-02 00:50:53,864 [myid:] - INFO [main:Environment@100] - Client environment:user.name=aluno
2016-02-02 00:50:53,864 [myid:] - INFO [main:Environment@100] - Client environment:user.home=/home/aluno
2016-02-02 00:50:53,864 [myid:] - INFO [main:Environment@100] - Client environment:user.dir=/home/aluno
2016-02-02 00:50:53,865 [myid:] - INFO [main:ZooKeeper@438] - Initiating client connection, connectString=localhost:2181 sessionTimeout=30000 watcher=org.apache.zookeeper.ZooKeeperMain$MyWatcher@4cc77c2e
2016-02-02 00:50:53,921 [myid:] - INFO [main-SendThread(localhost:2181):ClientCnxn$SendThread@975] - Opening socket connection to server localhost/127.0.0.1:2181. Will not attempt to authenticate using SASL (unknown error)
Welcome to ZooKeeper!
JLine support is enabled
2016-02-02 00:50:54,081 [myid:] - INFO [main-SendThread(localhost:2181):ClientCnxn$SendThread@852] - Socket connection established to localhost/127.0.0.1:2181, initiating session
2016-02-02 00:50:54,116 [myid:] - INFO [main-SendThread(localhost:2181):ClientCnxn$SendThread@1235] - Session establishment complete on server localhost/127.0.0.1:2181, sessionid = 0x1529fe2f6d40001, negotiated timeout = 30 000
WATCHER:::
WatchedEvent state:SyncConnected type:None path:null
[zk: localhost:2181(CONNECTED) 0]

```

CLI iniciado

1 / 4



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

Welcome to Apache HBase™

Apache HBase™ is the Hadoop database, a distributed, scalable, big data store.

Use Apache HBase™ when you need random, realtime read/write access to your Big Data. This project's goal is the hosting of very large tables -- billions of rows X millions of columns -- atop clusters of commodity hardware. Apache HBase is an open-source, distributed, versioned, non-relational database modeled after Google's [BigTable: A Distributed Storage System for Structured Data](#) by Chang et al. Just as BigTable leverages the distributed data storage provided by the Google File System, Apache HBase provides Bigtable-like capabilities on top of Hadoop and HDFS.

Download

Click [here](#) to download Apache HBase™.

Features

- Linear and modular scalability,
- Strictly consistent reads and writes,
- Automatic and configurable sharding of tables

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

A screenshot of a Linux terminal window titled "Terminal". The window shows the following command sequence:

```
aluno@dataserver:/opt/hbase/conf
Arquivo Editar Ver Pesquisar Terminal Ajuda
[aluno@dataserver ~]$ cd /opt/hbase/
[aluno@dataserver hbase]$ cd conf
[aluno@dataserver conf]$ gedit hbase-env.sh
```

The terminal window has a standard Linux interface with a menu bar at the top and a scroll bar on the right side.

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



Engenharia de Dados com Hadoop e Spark 3.0

```

Aplicativos Locais gedit
* *hbase-env.sh
/opt/hbase/conf

# Set environment variables here.

# This script sets variables multiple times over the course of starting an hbase process,
# so try to keep things idempotent unless you want to take an even deeper look
# into the startup scripts (bin/hbase, etc.)

# The java implementation to use. Java 1.7+ required.
export JAVA_HOME=/opt/jdk

# Extra Java CLASSPATH elements. Optional.
# export HBASE_CLASSPATH=

# The maximum amount of heap to use. Default is left to JVM default.
# export HBASE_HEAPSIZE=1G

# Uncomment below if you intend to use off heap cache. For example, to allocate 8G of
# offheap, set the value to "8G".
# export HBASE_OFFHEAPSIZE=1G

# Extra Java runtime options.
# Below are what we set by default. May only work with SUN JVM.
# For more on why as well as other possible settings,
# see http://wiki.apache.org/hadoop/PerformanceTuning
export HBASE_OPTS="-XX:+UseConcMarkSweepGC"

# Configure PermSize. Only needed in JDK7. You can safely remove it for JDK8+
#export HBASE_MASTER_OPTS="$HBASE_MASTER_OPTS -XX:PermSize=128m -XX:MaxPermSize=128m"
#export HBASE_REGIONSERVER_OPTS="$HBASE_REGIONSERVER_OPTS -XX:PermSize=128m -XX:MaxPermSize=128m"
#
# Uncomment one of the below three options to enable java garbage collection logging for the server-side
processes.

# This enables basic gc logging to the .out file.
# export SERVER_GC_OPTS="-verbose:gc -XX:+PrintGCDetails -XX:+PrintGCDateStamps"

# This enables basic gc logging to its own file.
# If FILE-PATH is not replaced, the log file(.gc) would still be generated in the HBASE_LOG_DIR .
# export SERVER_GC_OPTS="-verbose:gc -XX:+PrintGCDetails -XX:+PrintGCDateStamps -Xloggc:<FILE-PATH>"
```

sh Largura da tabulação: 8 Lin 48, Col 1 1 / 4

Editar o PATH do Java e comentar as linhas do PermSize

*Engenharia de Dados com Hadoop e Spark 3.0*

A screenshot of a terminal window titled "Terminal". The window shows a command-line session:

```
aluno@dataserver:/opt/hbase/conf
Arquivo Editar Ver Pesquisar Terminal Ajuda
[aluno@dataserver ~]$ cd /opt/hbase/
[aluno@dataserver hbase]$ cd conf
[aluno@dataserver conf]$ gedit hbase-env.sh
[aluno@dataserver conf]$ gedit hbase-site.xml
```

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

1 / 4

*Engenharia de Dados com Hadoop e Spark 3.0*

gedit Fri 17:08

hbase-site.xml /opt/hbase/conf

```
<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
  *
  * Licensed to the Apache Software Foundation (ASF) under one
  * or more contributor license agreements. See the NOTICE file
  * distributed with this work for additional information
  * regarding copyright ownership. The ASF licenses this file
  * to you under the Apache License, Version 2.0 (the
  * "License"); you may not use this file except in compliance
  * with the License. You may obtain a copy of the License at
  *
  *     http://www.apache.org/licenses/LICENSE-2.0
  *
  * Unless required by applicable law or agreed to in writing, software
  * distributed under the License is distributed on an "AS IS" BASIS,
  * WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
  * See the License for the specific language governing permissions and
  * limitations under the License.
  */
-->
<configuration>
  <property>
    <name>hbase.rootdir</name>
    <value>file:///opt/hbase/hfiles</value>
  </property>
  <property>
    <name>hbase.zookeeper.property.dataDir</name>
    <value>/opt/zookeeper/data</value>
  </property>
</configuration>
```

XML Tab Width: 8 Ln 33, Col 17 INS

Index of /zookeeper/zookeeper-3.... aluno@dataserver:/opt/hbase/conf hbase-site.xml (/opt/hbase/conf) ... 1 / 4

Incluir as linhas entre as tags <configuration>

*Engenharia de Dados com Hadoop e Spark 3.0*A screenshot of a Linux desktop environment showing a terminal window. The window title is 'Terminal'. The terminal prompt shows 'aluno@dataserver:~\$'. Below the prompt, the command 'gedit .bashrc' is visible. The window has a standard OS X-style title bar with icons for application, location, and system status. A scroll bar is visible on the right side of the terminal window.

```
aluno@dataserver:~$ gedit .bashrc
```

Editar o arquivo .bashrc

1 / 4

*Engenharia de Dados com Hadoop e Spark 3.0*

Aplicativos Locais gedit * .bashrc

en Ter, 02:13 Salvar

```
# .bashrc

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

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

# User specific aliases and functions

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

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

# Zookeeper
export ZOOKEEPER_HOME=/opt/zookeeper
export PATH=$PATH:$ZOOKEEPER_HOME/bin

# HBase
export HBASE_HOME=/opt/hbase
export PATH=$PATH:$HBASE_HOME/bin
```

sh Largura da tabulação: 8 Lin 35, Col 1 INS

aluno@dataserver:~ Index of /hbase/1.1.3 – Mozilla Fir... *.bashrc ("") – gedit 1 / 4

Variáveis HBase

*Engenharia de Dados com Hadoop e Spark 3.0*A screenshot of a Linux terminal window titled "Terminal". The window shows the command history:

```
aluno@dataserver:~$ Arquivo Editar Ver Pesquisar Terminal Ajuda
[aluno@dataserver ~]$ gedit .bashrc
[aluno@dataserver ~]$ source .bashrc
[aluno@dataserver ~]$
```

The window has a standard OS X-style title bar with icons for "Aplicativos", "Locais", and "Terminal". The status bar at the bottom right shows "en" and the date "Ter, 02:13".

A screenshot of a Mozilla Firefox browser window. The address bar shows "aluno@dataserver:~". The main content area displays the "Index of /hbase/1.1.3" page, which includes a table of contents for various files and sub-directories. The status bar at the bottom right shows "1 / 4".

*Engenharia de Dados com Hadoop e Spark 3.0*A screenshot of a terminal window titled "Terminal". The window has a menu bar with "Aplicativos", "Locais", and "Terminal". The status bar shows "en Ter, 02:13". The terminal prompt is "aluno@dataserver:~". Below the prompt, the command "[aluno@dataserver ~]\$ start-hbase.sh" is visible. The window is set against a dark background.

Iniciar o Hbase - start-hbase.sh

1 / 4

*Engenharia de Dados com Hadoop e Spark 3.0*A screenshot of a terminal window titled "Terminal". The window shows the command "start-hbase.sh" being run, which starts an Hbase master process. The output indicates that the master is starting and logging to a specific file path.

```
aluno@dataserver:~$ Arquivo Editar Ver Pesquisar Terminal Ajuda [aluno@dataserver ~]$ start-hbase.sh starting master, logging to /opt/hbase/logs/hbase-aluno-master-dataserver.out [aluno@dataserver ~]$
```

Hbase iniciado

*Engenharia de Dados com Hadoop e Spark 3.0*A screenshot of a terminal window titled "Terminal". The window shows the command "start-hbase.sh" being run, which starts a master process and begins logging to a file. The user then runs "hbase shell", which is currently active and awaiting input.

```
Arquivo Editar Ver Pesquisar Terminal Ajuda
[aluno@dataserver ~]$ start-hbase.sh
starting master, logging to /opt/hbase/logs/hbase-aluno-master-dataserver.out
[aluno@dataserver ~]$ hbase shell
```

Abrir o shell do Hbase

1 / 4

*Engenharia de Dados com Hadoop e Spark 3.0*

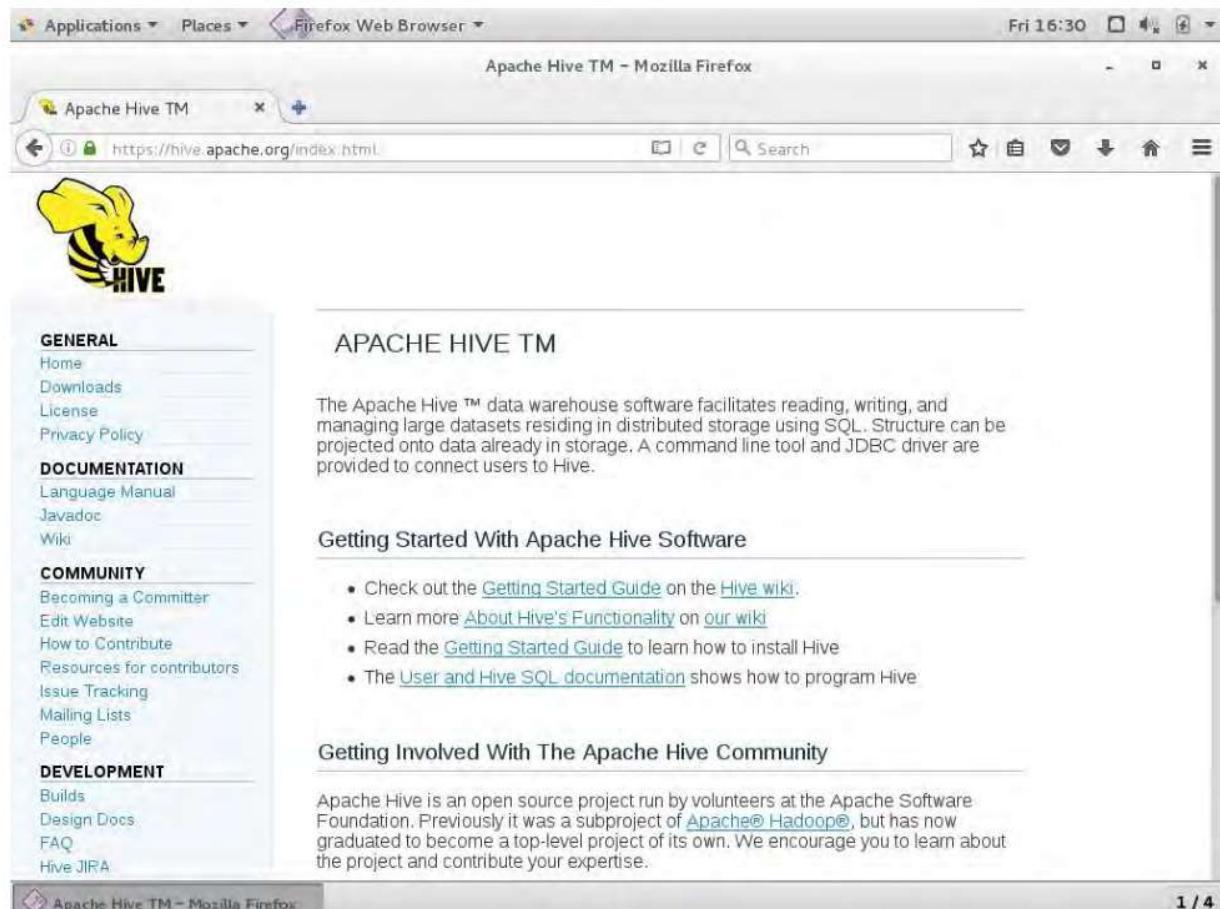
A screenshot of a terminal window titled "Terminal". The window shows the command "start-hbase.sh" being run, which starts an HBase master server. The output includes logs about SLF4J binding issues and the version of the HBase shell.

```
Arquivo Editar Ver Pesquisar Terminal Ajuda
[aluno@dataserver ~]$ start-hbase.sh
starting master, logging to /opt/hbase/logs/hbase-aluno-master-dataserver.out
[aluno@dataserver ~]$ hbase shell
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/hbase-1.1.3/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLogge
rBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hadoop-2.7.2/share/hadoop/common/lib/slf4j-log4j12-1.7.10.jar!/org/
slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
HBase Shell; enter 'help<RETURN>' for list of supported commands.
Type "exit<RETURN>" to leave the HBase Shell
Version 1.1.3, r72bc50f5fafecb105b2139e42bbe3d61ca724989, Sat Jan 16 18:29:00 PST 2016
hbase(main):001:0> ■
```

Shell iniciado

8. Instalação e Configuração do Hive

8.1. Download e Instalação do Hive



The screenshot shows a Firefox browser window with the title "Apache Hive TM - Mozilla Firefox". The address bar displays the URL <https://hive.apache.org/index.html>. The page content is the Apache Hive homepage, featuring the Hive logo (a yellow elephant) and the text "APACHE HIVE TM". The left sidebar contains navigation links for "GENERAL" (Home, Downloads, License, Privacy Policy), "DOCUMENTATION" (Language Manual, Javadoc, Wiki), "COMMUNITY" (Becoming a Committer, Edit Website, How to Contribute, Resources for contributors, Issue Tracking, Mailing Lists, People), and "DEVELOPMENT" (Builds, Design Docs, FAQ, Hive JIRA). The main content area includes a brief description of what Apache Hive is, a "Getting Started With Apache Hive Software" section with a bulleted list of links, and a "Getting Involved With The Apache Hive Community" section with information about the project's history and contribution guidelines. The bottom right corner of the browser window shows "1 / 4".

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

A screenshot of a terminal window titled "Terminal". The window shows the command "aluno@dataserver:~\$ gedit .bashrc" being typed. The terminal is running on a Linux system, indicated by the window title bar which includes "Aplicativos", "Locais", and "Terminal". The status bar at the bottom right shows "en Seg, 00:21".

Arquivo Editar Ver Pesquisar Terminal Ajuda
[aluno@dataserver ~]\$ gedit .bashrc

Editando o arquivo .bashrc

*Engenharia de Dados com Hadoop e Spark 3.0*

Aplicativos Locais gedit

bashrc

.bashrc

```

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

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

# User specific aliases and functions

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

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

# Hive
export HIVE_HOME=/opt/hive
export PATH=$PATH:$HIVE_HOME/bin
export CLASSPATH=$CLASSPATH:$HADOOP_HOME/lib/*
export CLASSPATH=$CLASSPATH:$HIVE_HOME/lib/*

```

sh Largura da tabulação: 8 Lin 32, Col 1 INS

aluno@dataserver:~/.bashrc (~) - gedit 1 / 4

Variáveis de ambiente do Hive

*Engenharia de Dados com Hadoop e Spark 3.0*A screenshot of a Linux terminal window titled "Terminal". The window has a menu bar with "Aplicativos", "Locais", and "Terminal". The status bar shows "en Seg, 00:25". The terminal prompt is "aluno@dataserver:~". Below the prompt, the command "[aluno@dataserver ~]\$ source .bashrc" is visible. The window has a standard title bar with minimize, maximize, and close buttons. A vertical scroll bar is on the right side of the terminal area. The bottom of the window shows a toolbar with icons for file operations and a status bar with "1 / 4".

```
source .bashrc
```

*Engenharia de Dados com Hadoop e Spark 3.0*A screenshot of a terminal window titled "Terminal". The window shows a command-line session:

```
aluno@dataserver:/opt/hive/conf
[aluno@dataserver ~]$ cd $HIVE_HOME/conf
[aluno@dataserver conf]$ cp hive-env.sh.template hive-env.sh
```

The terminal has a standard window title bar with "Aplicativos", "Locais", and "Terminal". The status bar at the bottom right shows "en Seg, 00:28".

1 / 4

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

*Engenharia de Dados com Hadoop e Spark 3.0*

A screenshot of a terminal window titled "Terminal". The window shows a command-line session:

```
aluno@dataserver:~/opt/hive/conf
[aluno@dataserver ~]$ cd $HIVE_HOME/conf
[aluno@dataserver conf]$ cp hive-env.sh.template hive-env.sh
[aluno@dataserver conf]$ gedit hive-env.sh
```

The terminal window has a standard Linux-style interface with a title bar, menu bar, and status bar indicating the user is at "aluno@dataserver:~/opt/hive/conf" and it's Seg, 00:28.

Editar o arquivo

1 / 4

*Engenharia de Dados com Hadoop e Spark 3.0*

Aplicativos Locais gedit

*hive-env.sh
/opt/hive/conf

```
# to control the execution of Hive. It should be used by admins to configure
# the Hive installation (so that users do not have to set environment variables
# or set command line parameters to get correct behavior).
#
# The hive service being invoked (CLI/HWI etc.) is available via the environment
# variable SERVICE

# Hive Client memory usage can be an issue if a large number of clients
# are running at the same time. The flags below have been useful in
# reducing memory usage:
#
# if [ "$SERVICE" = "cli" ]; then
#   if [ -z "$DEBUG" ]; then
#     export HADOOP_OPTS="$HADOOP_OPTS -XX:NewRatio=12 -Xms10m -XX:MaxHeapFreeRatio=40 -XX:MinHeapFreeRatio=15
# -XX:+UseParNewGC -XX:-UseGCOverheadLimit"
#   else
#     export HADOOP_OPTS="$HADOOP_OPTS -XX:NewRatio=12 -Xms10m -XX:MaxHeapFreeRatio=40 -XX:MinHeapFreeRatio=15
# -XX:-UseGCOverheadLimit"
#   fi
# fi

# The heap size of the jvm started by hive shell script can be controlled via:
#
# export HADOOP_HEAPSIZE=1024
#
# Larger heap size may be required when running queries over large number of files or partitions.
# By default hive shell scripts use a heap size of 256 (MB). Larger heap size would also be
# appropriate for hive server (hwi etc).

# Set HADOOP_HOME to point to a specific hadoop install directory
export HADOOP_HOME=/opt/hadoop

# Hive Configuration Directory can be controlled by:
# export HIVE_CONF_DIR=

# Folder containing extra libraries required for hive compilation/execution can be controlled by:
# export HIVE_AUX_JARS_PATH=
```

sh ▾ Largura da tabulação: 8 ▾ Lin 48, Col 31 ▾ INS

aluno@dataserver:/opt/hive/conf *hive-env.sh (/opt/hive/conf) - gedit 1 / 4

Incluir PATH do Hadoop, conforme tela acima

*Engenharia de Dados com Hadoop e Spark 3.0*

```
aluno@dataserver:/opt/hive/conf
Arquivo Editar Ver Pesquisar Terminal Ajuda
[aluno@dataserver conf]$ cd $HIVE_HOME/conf
[aluno@dataserver conf]$ cp hive-default.xml.template hive-default.xml
[aluno@dataserver conf]$ gedit hive-default.xml
```

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

*Engenharia de Dados com Hadoop e Spark 3.0*

Aplicativos Locais gedit

hive-default.xml
/opt/hive/conf

Two supported values are : kryo and javaXML. Kryo is default.

```

</description>
</property>
<property>
  <name>hive.exec.stagingdir</name>
  <value>.hive-staging</value>
  <description>Directory name that will be created inside table locations in order to support HDFS encryption. This is replaces ${hive.exec.scratchdir} for query results with the exception of read-only tables. In all cases ${hive.exec.scratchdir} is still used for other temporary files, such as job plans.</description>
</property>
<property>
  <name>hive.exec.scratchdir</name>
  <value>/tmp/hive</value>
  <description>HDFS root scratch dir for Hive jobs which gets created with write all (733) permission. For each connecting user, an HDFS scratch dir: ${hive.exec.scratchdir}/&lt;username&gt; is created, with ${hive.scratch.dir.permission}.</description>
</property>
<property>
  <name>hive.exec.local.scratchdir</name>
  <value>/tmp/hive</value>
  <description>Local scratch space for Hive jobs</description>
</property>
<property>
  <name>hive.downloaded.resources.dir</name>
  <value>/tmp/hive</value>
  <description>Temporary local directory for added resources in the remote file system.</description>
</property>
<property>
  <name>hive.scratch.dir.permission</name>
  <value>700</value>
  <description>The permission for the user specific scratch directories that get created.</description>
</property>
<property>
  <name>hive.exec.submitviachild</name>
  <value>false</value>
  <description/>
</property>
<property>
```

aluno@dataserver:/opt/hive/conf

hive-default.xml (/opt/hive/conf) -

XML Largura da tabulação: 8 Lin 61, Col 45 INS 1 / 4

Editar as linhas conforme cima

*Engenharia de Dados com Hadoop e Spark 3.0*

The screenshot shows a terminal window titled "Terminal" with the command "aluno@dataserver:~". The terminal output is as follows:

```
File Edit View Search Terminal Help
[aluno@dataserver conf]$ cd ~
[aluno@dataserver ~]$ schematool -initSchema -dbType derby
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/apache-hive-2.1.0-bin/lib/log4j-slf4j-impl-2.4.1.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hadoop-2.7.3/share/hadoop/common/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
Metastore connection URL: jdbc:derby:;databaseName=metastore_db;create=true
Metastore Connection Driver : org.apache.derby.jdbc.EmbeddedDriver
Metastore Connection User: APP
Starting metastore schema initialization to 2.1.0
Initialization script hive-schema-2.1.0.derby.sql
Initialization script completed
schemaTool completed
[aluno@dataserver ~]$
```


The screenshot shows a Firefox browser window with a slide from a presentation. The title of the slide is "Inicializar o schema do Hive" and the content is "schematool -initSchema -dbType derby". The browser address bar shows "index.html/base - Mozilla Firefox".

*Engenharia de Dados com Hadoop e Spark 3.0*A screenshot of a terminal window titled "Terminal" with the command "aluno@dataserver:~". The window shows the output of the "hive" command. It includes several SLF4J binding messages, a logging initialization message, and a warning about Hive-on-MR being deprecated. The session ends with the prompt "hive>".

```
File Edit View Search Terminal Help
[aluno@dataserver ~]$ hive
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/apache-hive-2.1.0-bin/lib/log4j-slf4j-impl-2.4.1.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hadoop-2.7.3/share/hadoop/common/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]

Logging initialized using configuration in jar:file:/usr/local/apache-hive-2.1.0-bin/lib/hive-common-2.1.0.jar!/hive-log4j2.properties Async: true
Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
hive>
```

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



Engenharia de Dados com Hadoop e Spark 3.0

A screenshot of a terminal window titled "Terminal". The window shows the command "hive" being run, followed by several SLF4J binding logs. Then, it shows the configuration of the logging system using "hive-log4j2.properties". Finally, it displays the result of the "show tables;" command, which is "OK".

```
File Edit View Search Terminal Help
[aluno@dataserver ~]$ hive
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/apache-hive-2.1.0-bin/lib/log4j-slf4j-impl-2.4.1.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hadoop-2.7.3/share/hadoop/common/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]

Logging initialized using configuration in jar:file:/usr/local/apache-hive-2.1.0-bin/lib/hive-common-2.1.0.jar!/hive-log4j2.properties Async: true
Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
hive> show tables;
OK
Time taken: 1,352 seconds
hive>
```

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



Welcome to Apache Pig! – Mozilla Firefox

Welcome to Apache ...

https://pig.apache.org

Apache > Pig >

hadoop

Welcome to Apache Pig!

News

- Apache Pig 0.16.0 is released!
- Getting Started
- Getting Involved

Apache Pig is a platform for analyzing large data sets that consists of a high-level language for expressing data analysis programs, coupled with infrastructure for evaluating these programs. The salient property of Pig programs is that their structure is amenable to substantial parallelization, which in turns enables them to handle very large data sets.

At the present time, Pig's infrastructure layer consists of a compiler that produces sequences of Map-Reduce programs, for which large-scale parallel implementations already exist (e.g., the Hadoop subproject). Pig's language layer currently consists of a textual language called Pig Latin, which has the following key properties:

- Ease of programming.** It is trivial to achieve parallel execution of simple, "embarrassingly parallel" data analysis tasks. Complex tasks comprised of multiple interrelated data transformations are explicitly encoded as data flow sequences, making them easy to write, understand, and maintain.
- optimization opportunities.** The way in which tasks are encoded permits the system to optimize their execution automatically, allowing the user to focus on semantics rather than

1 / 4

Download do Pig – Versão 0.17.0

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



9.2. Configurando do Pig

A screenshot of a terminal window titled 'Terminal'. The window shows the command 'aluno@dataserver:~\$ gedit .bashrc' being typed. The terminal interface includes a menu bar with 'Arquivo', 'Editar', 'Ver', 'Pesquisar', 'Terminal', and 'Ajuda'. The status bar at the bottom right shows 'en Seg, 01:30'.

aluno@dataserver:~\$ gedit .bashrc

Editando o arquivo .bashrc



Engenharia de Dados com Hadoop e Spark 3.0

```
# export SYSTEMD_PAGER=

# User specific aliases and functions

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

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

# Zookeeper
export ZOOKEEPER_HOME=/opt/zookeeper
export PATH=$PATH:$ZOOKEEPER_HOME/bin

# HBase
export HBASE_HOME=/opt/hbase
export PATH=$PATH:$HBASE_HOME/bin

# Hive
export HIVE_HOME=/opt/hive
export PATH=$PATH:$HIVE_HOME/bin
export CLASSPATH=$CLASSPATH:$HADOOP_HOME/lib/*:.
export CLASSPATH=$CLASSPATH:$HIVE_HOME/lib/*:.

# Pig
export PIG_HOME=/opt/pig
export PATH=$PATH:$PIG_HOME/bin
export PIG_CLASSPATH=$HADOOP_HOME/conf|
```

sh ▾ Tab Width: 8 ▾ Ln 44, Col 39 ▾ INS
aluno@dataserver:~/.bashrc (~/) - gedit 1 / 4

Inserir variáveis de ambiente do Pig

*Engenharia de Dados com Hadoop e Spark 3.0*A screenshot of a terminal window titled "Terminal". The window has a menu bar with "Aplicativos", "Locais", and "Terminal". The status bar shows "en Seg, 01:33". The terminal prompt is "aluno@dataserver:~". Below the prompt, the command "[aluno@dataserver ~]\$ source .bashrc" is visible. The window is set against a dark background.

source .bashrc

1 / 4

*Engenharia de Dados com Hadoop e Spark 3.0*

A screenshot of a terminal window titled "Terminal". The window shows a command-line session where the user runs "source .bashrc" and then "pig". The terminal displays a series of INFO log messages from the Apache Pig 0.15.0 version, indicating the configuration of the execution environment (LOCAL vs MAPREDUCE), the logging of error messages to a file, and the connection to a local HDFS system at port 9000. The session ends with the prompt "grunt>".

```
Arquivo Editar Ver Pesquisar Terminal Ajuda
[aluno@dataserver ~]$ source .bashrc
[aluno@dataserver ~]$ pig
16/02/01 01:33:45 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
16/02/01 01:33:45 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
16/02/01 01:33:45 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2016-02-01 01:33:45,239 [main] INFO org.apache.pig.Main - Apache Pig version 0.15.0 (r1682971) compiled Jun 01
2015, 11:44:35
2016-02-01 01:33:45,239 [main] INFO org.apache.pig.Main - Logging error messages to: /home/aluno/pig_1454297625
238.log
2016-02-01 01:33:45,270 [main] INFO org.apache.pig.impl.util.Utils - Default bootup file /home/aluno/.pigbootup
not found
2016-02-01 01:33:45,869 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is de
precated. Instead, use mapreduce.jobtracker.address
2016-02-01 01:33:45,869 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is depre
cated. Instead, use fs.defaultFS
2016-02-01 01:33:45,869 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting
to hadoop file system at: hdfs://localhost:9000
2016-02-01 01:33:46,777 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is depre
cated. Instead, use fs.defaultFS
grunt>
```

Pig instalado com sucesso

1 / 4

*Engenharia de Dados com Hadoop e Spark 3.0*

A screenshot of a terminal window titled "Terminal". The window shows the command [aluno@dataserver ~]\$ pig -h properties being run. The output of the command is displayed below the command line, indicating that no properties are listed.

```
[aluno@dataserver ~]$ pig -h properties
```

1 / 4

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

*Engenharia de Dados com Hadoop e Spark 3.0*

```

Aplicativos Locais Terminal en Seg, 01:36
aluno@dataserver:~$

Arquivo Editar Ver Pesquisar Terminal Ajuda

Logging:
    verbose=true|false; default is false. This property is the same as -v switch
    brief=true|false; default is false. This property is the same as -b switch
    debug=OFF|ERROR|WARN|INFO|DEBUG; default is INFO. This property is the same as -d switch
    aggregate.warning=true|false; default is true. If true, prints count of warnings
        of each type rather than logging each warning.

Performance tuning:
    pig.cachedbag.memusage=<mem fraction>; default is 0.2 (20% of all memory).
        Note that this memory is shared across all large bags used by the application.
    pig.skewedjoin.reduce.memusage=<mem fraction>; default is 0.3 (30% of all memory).
        Specifies the fraction of heap available for the reducer to perform the join.
    pig.exec.no combiner=true|false; default is false.
        Only disable combiner as a temporary workaround for problems.
    opt.multiquery=true|false; multiquery is on by default.
        Only disable multiquery as a temporary workaround for problems.
    opt.fetch=true|false; fetch is on by default.
        Scripts containing Filter, Foreach, Limit, Stream, and Union can be dumped without MR jobs.
    pig.tmpfilecompression=true|false; compression is off by default.
        Determines whether output of intermediate jobs is compressed.
    pig.tmpfilecompression.codec=lzo|gzip; default is gzip.
        Used in conjunction with pig.tmpfilecompression. Defines compression type.
    pig.noSplitCombination=true|false. Split combination is on by default.
        Determines if multiple small files are combined into a single map.
    pig.exec.mapPartAgg=true|false. Default is false.
        Determines if partial aggregation is done within map phase,
        before records are sent to combiner.
    pig.exec.mapPartAgg.minReduction=<min aggregation factor>. Default is 10.
        If the in-map partial aggregation does not reduce the output num records
        by this factor, it gets disabled.

Miscellaneous:
    exec.type=mapreduce|tez|local; default is mapreduce. This property is the same as -X switch
    pig.additional.jars.uris=<comma separated list of jars>. Used in place of register command.
    udf.import.list=<comma separated list of imports>. Used to avoid package names in UDF.
    stop.on.failure=true|false; default is false. Set to true to terminate on the first error.
    pig.datetime.default.tz=<UTC time offset>. e.g. +08:00. Default is the default timezone of the host.
        Determines the timezone used to handle datetime datatype and UDFs.

Additionally, any Hadoop property can be specified.
16/02/01 01:36:00 INFO pig.Main: Pig script completed in 224 milliseconds (224 ms)
[aluno@dataserver ~]$
```

1 / 4

Variáveis Pig

*Engenharia de Dados com Hadoop e Spark 3.0*A screenshot of a terminal window titled "Terminal". The window shows the command "pig -version" being run and its output: "Apache Pig version 0.15.0 (r1682971) compiled Jun 01 2015, 11:44:35".

```
aluno@dataserver:~$ pig -version
Apache Pig version 0.15.0 (r1682971)
compiled Jun 01 2015, 11:44:35
[aluno@dataserver ~]$
```

Verificar a versão do Pig

1 / 4



10. Instalação e Configuração do Spark

10.1. Download e Instalação do Spark

Screenshot of the Apache Spark website (<https://spark.apache.org>) showing the homepage content.

The page features the Apache Spark logo and tagline "Lightning-fast unified analytics engine". It highlights "Speed" (running workloads 100x faster) and "Ease of Use" (writing applications quickly in Java, Scala, Python, R, and SQL). A chart compares the running time of logistic regression between Hadoop and Spark, showing Spark is significantly faster (0.9 seconds vs 110 seconds).

The "Documentation" section includes a snippet of Python code for reading JSON files:

```
df = spark.read.json("logs.json")
df.where("age > 21")
.select("name.first").show()
```

The "Latest News" sidebar lists recent releases and events, such as the plan for dropping Python 2 support and the APACHECON conference.

A prominent green button at the bottom right says "Download 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.

*Engenharia de Dados com Hadoop e Spark 3.0*A screenshot of a Linux desktop environment showing a terminal window titled "Terminal". The window title bar also includes "Applications", "Places", and "Fri 18:25". The terminal window shows the command line interface with the user "aluno" at "databserver". The commands entered are "pwd" (showing the current directory as "/home/aluno") and "gedit .bashrc" (opening the file ".bashrc" for editing).

```
File Edit View Search Terminal Help
[aluno@databserver ~]$ pwd
/home/aluno
[aluno@databserver ~]$ gedit .bashrc
```

Editando o arquivo .bashrc



The screenshot shows a Linux desktop environment. In the top panel, there are application icons for Applications, Places, and gedit. The gedit window is open, showing the file `.bashrc`. The terminal window below it has tabs for "aluno@dataserver:" and "Apache Pig Releases - Mozilla Fire...". The terminal content is a series of `export` commands for setting up the paths for Java, Hadoop, Zookeeper, HBase, Hive, Pig, and Spark.

```
# User specific aliases and functions

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

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

# Zookeeper
export ZOOKEEPER_HOME=/opt/zookeeper
export PATH=$PATH:$ZOOKEEPER_HOME/bin

# HBase
export HBASE_HOME=/opt/hbase
export PATH=$PATH:$HBASE_HOME/bin

# Hive
export HIVE_HOME=/opt/hive
export PATH=$PATH:$HIVE_HOME/bin
export CLASSPATH=$CLASSPATH:$HADOOP_HOME/lib/*:.
export CLASSPATH=$CLASSPATH:$HIVE_HOME/lib/*:.

# Pig
export PIG_HOME=/opt/pig
export PATH=$PATH:$PIG_HOME/bin
export PIG_CLASSPATH=$HADOOP_HOME/conf

# Spark
export SPARK_HOME=/opt/spark
export PATH=$PATH:$SPARK_HOME/bin
```

Incluir variáveis Spark

*Engenharia de Dados com Hadoop e Spark 3.0*A screenshot of a Linux terminal window titled "Terminal". The window shows a command-line interface with the following session:

```
File Edit View Search Terminal Help  
[aluno@dataserver ~]$ pwd  
/home/aluno  
[aluno@dataserver ~]$ source .bashrc
```

The terminal is running on a server named "dataserver" under user "aluno". The window has a standard title bar with icons for Applications, Places, and Terminal, and a status bar at the bottom right showing the date and time: Fri 18:26.

source .bashrc

*Engenharia de Dados com Hadoop e Spark 3.0*A screenshot of a Linux desktop environment showing a terminal window. The window title is "Terminal". The terminal prompt is "aluno@dataserver:~". The user has run the commands "mkdir /tmp/hive" and "chmod 777 /tmp/hive".

```
File Edit View Search Terminal Help
[aluno@dataserver ~]$ mkdir /tmp/hive
[aluno@dataserver ~]$ chmod 777 /tmp/hive
[aluno@dataserver ~]$
```

Se necessário chmod 777 /tmp/hive



Engenharia de Dados com Hadoop e Spark 3.0

Spark shell

*Engenharia de Dados com Hadoop e Spark 3.0*

Applications ▾ Places ▾ Firefox Web Browser ▾ Fri 18:28

Spark shell - Spark Jobs - Mozilla Firefox

Spark shell - Spark Jobs × +

localhost:4040/jobs/ Search

Spark 2.0.0 Jobs Stages Storage Environment Executors SQL Spark shell application UI

Spark Jobs (?)

User: aluno
Total Uptime: 38 s
Scheduling Mode: FIFO

Event Timeline Enable zooming

Executors							
■ Added							
■ Removed							
	river added						

Jobs							
■ Succeeded							
■ Failed							
■ Running							
	Sat 1	Sun 2	Mon 3	Tue 4	Wed 5	Thu 6	Fri 7
	October 2016						

Spark shell - Spark Jobs - Mozilla... aluno@dataserver:~ 1 / 4

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



11. Instalação e Configuração do Sqoop

11.1. Download do Sqoop

The screenshot shows a Firefox browser window with the title "Scoop - Mozilla Firefox". The address bar contains "sqoop.apache.org". The main content area displays the Apache Sqoop project page. At the top, there's a logo for "The Apache Software Foundation" and a link to "http://www.apache.org/". Below the header, the text "Apache Sqoop" is prominently displayed. A paragraph describes Sqoop as a tool for transferring bulk data between Hadoop and relational databases. It mentions that Sqoop graduated from the Incubator in March 2012 and is now a Top-Level Apache project. The latest stable release is 1.4.6, and the latest cut of Sqoop2 is 1.99.7. A note states that 1.99.7 is not compatible with 1.4.6 and is not feature complete, so it's not intended for production deployment. A section titled "Download" provides instructions for downloading from mirrors or cloning the Git repository. A specific link for "Download do Sqoop – Versão 1.4.7-hadoop-2.0.4-alpha" is highlighted. The bottom of the page shows a footer with the Apache Software Foundation logo and a copyright notice.

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

A screenshot of a Linux desktop environment showing a terminal window titled "Terminal". The window title bar also includes "Aplicativos" and "Locais". The status bar at the top right shows "en Dom, 05:08". The terminal window contains the command "[aluno@dataserver ~]\$ gedit .bashrc" which is being typed by the user. The background of the desktop shows a blurred image of a person working on a computer.

Editar arquivo .bashrc

*Engenharia de Dados com Hadoop e Spark 3.0*

Applications ▾ Places ▾ gedit ▾

.bashrc

Fri 18:48

Save

```

export JRE_HOME=/opt/jre
export PATH=$PATH:$JRE_HOME/bin:$JAVA_HOME/bin

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

# Zookeeper
export ZOOKEEPER_HOME=/opt/zookeeper
export PATH=$PATH:$ZOOKEEPER_HOME/bin

# HBase
export HBASE_HOME=/opt/hbase
export PATH=$PATH:$HBASE_HOME/bin

# Hive
export HIVE_HOME=/opt/hive
export PATH=$PATH:$HIVE_HOME/bin
export CLASSPATH=$CLASSPATH:$HADOOP_HOME/lib/*:.
export CLASSPATH=$CLASSPATH:$HIVE_HOME/lib/*:.

# Pig
export PIG_HOME=/opt/pig
export PATH=$PATH:$PIG_HOME/bin
export PIG_CLASSPATH=$HADOOP_HOME/conf

# Spark
export SPARK_HOME=/opt/spark
export PATH=$PATH:$SPARK_HOME/bin

# Sqoop
export SQOOP_HOME=/opt/sqoop
export PATH=$PATH:$SQOOP_HOME/bin

```

sh ▾ Tab Width: 8 ▾ Ln 54, Col 1 ▾ INS

aluno@dataserver:~/.bashrc (~) - gedit

1 / 4

Incluir variáveis Sqoop

*Engenharia de Dados com Hadoop e Spark 3.0*A screenshot of a Linux terminal window titled "Terminal". The window has a menu bar with "Aplicativos", "Locais", and "Terminal". The status bar shows "en Dom, 05:10". The terminal prompt is "aluno@dataserver:~". The user has run the commands "gedit .bashrc" and "source .bashrc".

```
Arquivo Editar Ver Pesquisar Terminal Ajuda
[aluno@dataserver ~]$ gedit .bashrc
[aluno@dataserver ~]$ source .bashrc
[aluno@dataserver ~]$
```

source .bashrc

1 / 4

*Engenharia de Dados com Hadoop e Spark 3.0*

```
aluno@dataserver:~$ cd $SQOOP_HOME/conf
[aluno@dataserver conf]$ cp sqoop-env-template.sh sqoop-env.sh
[aluno@dataserver conf]$
```

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

*Engenharia de Dados com Hadoop e Spark 3.0*

```
aluno@dataserver:~$ cd $SQOOP_HOME/conf
[aluno@dataserver conf]$ cp sqoop-env-template.sh sqoop-env.sh
[aluno@dataserver conf]$ gedit sqoop-env.sh
```

The screenshot shows a terminal window titled "Terminal" with the command history displayed. The user is navigating to the \$SQOOP_HOME/conf directory, copying the sqoop-env-template.sh file to sqoop-env.sh, and then opening it in a text editor (gedit). The window title bar also shows the path "aluno@dataserver:~/opt/sqoop/conf".

Editar o arquivo

*Engenharia de Dados com Hadoop e Spark 3.0*

Applications ▾ Places ▾ _gedit ▾

Fri 19:04 Save ▾ - ▾ ×

sqoop-env.sh
/opt/sqoop/conf

```
# Licensed to the Apache Software Foundation (ASF) under one or more
# contributor license agreements. See the NOTICE file distributed with
# this work for additional information regarding copyright ownership.
# The ASF licenses this file to You under the Apache License, Version 2.0
# (the "License"); you may not use this file except in compliance with
# the License. You may obtain a copy of the License at
#
#     http://www.apache.org/licenses/LICENSE-2.0
#
# Unless required by applicable law or agreed to in writing, software
# distributed under the License is distributed on an "AS IS" BASIS,
# WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
# See the License for the specific language governing permissions and
# limitations under the License.

# included in all the hadoop scripts with source command
# should not be executable directly
# also should not be passed any arguments, since we need original $*

# Set Hadoop-specific environment variables here.

#Set path to where bin/hadoop is available
export HADOOP_COMMON_HOME=/opt/hadoop

#Set path to where hadoop-*core.jar is available
export HADOOP_MAPRED_HOME=/opt/hadoop

#set the path to where bin/hbase is available
export HBASE_HOME=/opt/hbase

#Set the path to where bin/hive is available
export HIVE_HOME=/opt/hive

#Set the path for where zookeper config dir is
export ZOOCFGDIR=/opt/zookeeper/conf]
```

sh ▾ Tab Width: 8 ▾ Ln 35, Col 37 ▾ INS

aluno@dataserver:/opt/sqoop/conf | Index of /sqoop/1.4.6 - Mozilla F... | sqoop-env.sh (/opt/sqoop/conf) ~ ... 1 / 4

Editar variáveis conforme tela acima

*Engenharia de Dados com Hadoop e Spark 3.0*A screenshot of a terminal window titled "Terminal". The window has a menu bar with "Aplicativos", "Locais", and "Terminal". The status bar shows "en Dom, 16:49". The terminal prompt is "aluno@dataserver:~". Below the prompt, the command "[aluno@dataserver ~]\$ sqoop version" is visible, with the cursor at the end of "version".

```
aluno@dataserver:~$ sqoop version
```

A screenshot of a Mozilla Firefox browser window. The address bar shows "aluno@dataserver:~". The main content area displays the command "sqoop version".

```
sqoop version
```

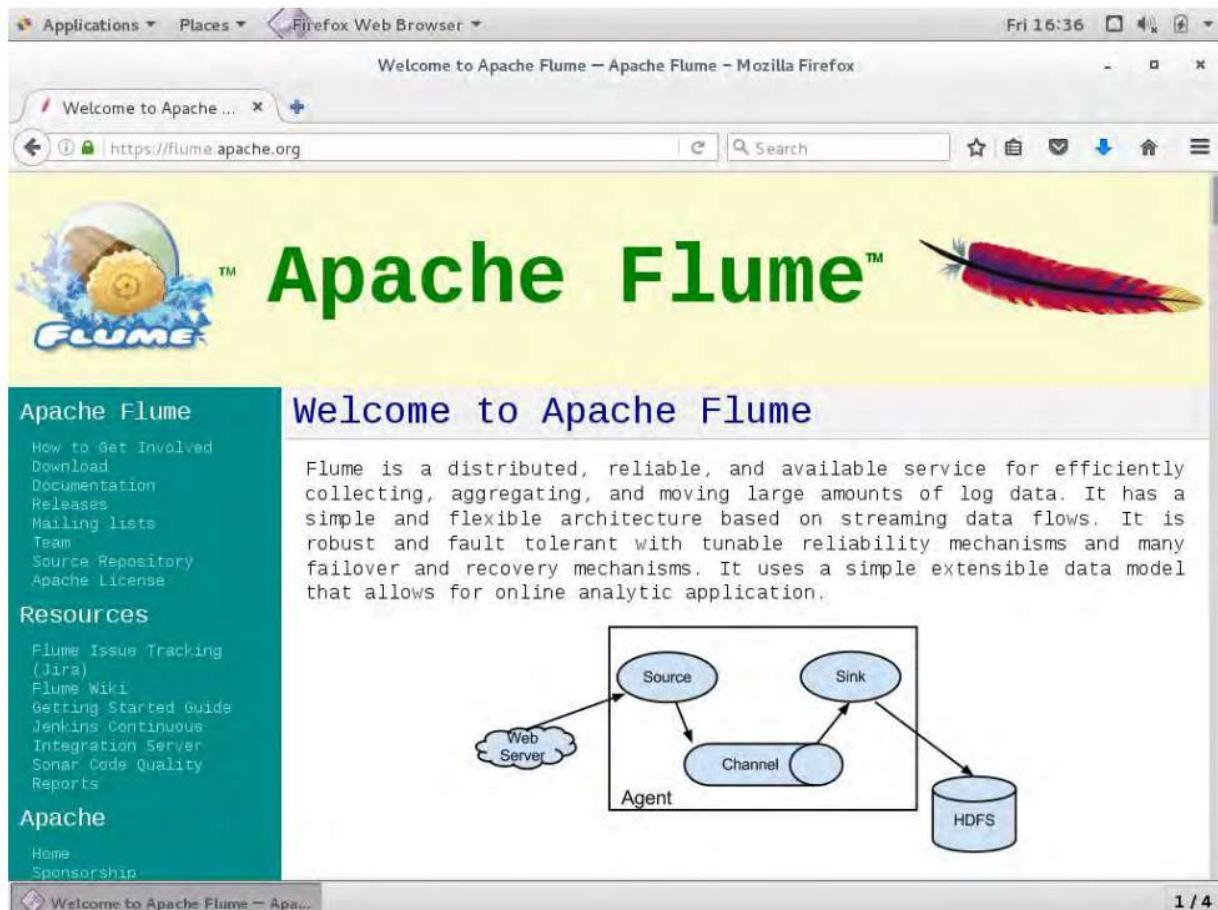
*Engenharia de Dados com Hadoop e Spark 3.0*A screenshot of a terminal window titled "Terminal". The window shows the command "sqoop version" being run and its output. The output indicates that Sqoop version 1.4.6 was running on April 27, 2015, at 14:38:36 CST, with a git commit ID of c0c5a81723759fa575844a0a1eae8f510fa32c25.

```
File Edit View Search Terminal Help
[aluno@dataserver ~]$ sqoop version
16/09/30 19:13:13 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6
Sqoop 1.4.6
git commit id c0c5a81723759fa575844a0a1eae8f510fa32c25
Compiled by root on Mon Apr 27 14:38:36 CST 2015
[aluno@dataserver ~]$
```

Sqoop version

1 / 4

12. Instalação e Configuração do Apache Flume



Welcome to Apache Flume – Apache Flume – Mozilla Firefox

Fri 16:36

Welcome to Apache ...

https://flume.apache.org

Search

Apache Flume™

Welcome to Apache Flume

Flume is a distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of log data. It has a simple and flexible architecture based on streaming data flows. It is robust and fault tolerant with tunable reliability mechanisms and many failover and recovery mechanisms. It uses a simple extensible data model that allows for online analytic application.

Apache Flume

- How to Get Involved
- Download
- Documentation
- Releases
- Mailing lists
- Team
- Source Repository
- Apache License

Resources

- Flume Issue Tracking (Jira)
- Flume Wiki
- Getting Started Guide
- Jenkins Continuous Integration Server
- Sonar Code Quality Reports

Apache

- Home
- Sponsorship

1 / 4

Download do Apache Flume – Versão 1.9

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

*Engenharia de Dados com Hadoop e Spark 3.0*A screenshot of a Linux desktop environment showing a terminal window titled "Terminal". The window title bar also includes "Applications", "Places", and "Fri 19:41". The terminal prompt shows "aluno@dataserver:~". The user has run the command "gedit .bashrc" in the terminal. The window has a standard Linux-style border with minimize, maximize, and close buttons.

Editar as variáveis de ambiente

1 / 4

*Engenharia de Dados com Hadoop e Spark 3.0*

Applications ▾ Places ▾ gedit Fri 19:41

bashrc Save

```

export HADOOP_HDFS_HOME=$HADOOP_HOME
export YARN_HOME=$HADOOP_HOME
export PATH=$PATH:$HADOOP_HOME/bin:$HADOOP_HOME/sbin

# Zookeeper
export ZOOKEEPER_HOME=/opt/zookeeper
export PATH=$PATH:$ZOOKEEPER_HOME/bin

# HBase
export HBASE_HOME=/opt/hbase
export PATH=$PATH:$HBASE_HOME/bin

# Hive
export HIVE_HOME=/opt/hive
export PATH=$PATH:$HIVE_HOME/bin
export CLASSPATH=$CLASSPATH:$HADOOP_HOME/lib/*:.
export CLASSPATH=$CLASSPATH:$HIVE_HOME/lib/*:.

# Pig
export PIG_HOME=/opt/pig
export PATH=$PATH:$PIG_HOME/bin
export PIG_CLASSPATH=$HADOOP_HOME/conf

# Spark
export SPARK_HOME=/opt/spark
export PATH=$PATH:$SPARK_HOME/bin

# Sqoop
export SQOOP_HOME=/opt/sqoop
export PATH=$PATH:$SQOOP_HOME/bin
export ACCUMULO_HOME=/opt/sqoop/accumulo
export HCAT_HOME=/opt/sqoop/hcatalog

# Flume
export FLUME_HOME=/opt/flume
export PATH=$PATH:$FLUME_HOME/bin
export CLASSPATH=$CLASSPATH:$FLUME_HOME/lib/*

```

sh ▾ Tab Width: 8 ▾ Ln 59, Col 46 ▾ INS

aluno@dataserver:~/.bashrc (~) - gedit 1 / 4

Variáveis de ambiente para o Flume

*Engenharia de Dados com Hadoop e Spark 3.0*

A screenshot of a terminal window titled "Terminal" at "aluno@dataserver:/opt/flume/conf". The window shows a file listing from the command "ls -la" and a command to edit "flume-env.sh" with "gedit".

```
File Edit View Search Terminal Help
[aluno@dataserver ~]$ cd /opt/flume/conf/
[aluno@dataserver conf]$ ls -la
total 32
drwxr-xr-x, 2 aluno aluno 4096 Sep 30 19:29 .
drwxrwxr-x, 7 aluno aluno 4096 Sep 30 19:23 ..
-rw-r--r--, 1 aluno aluno 1661 Sep 30 19:28 flume-conf.properties
-rw-r--r--, 1 aluno aluno 1661 May  8 2015 flume-conf.properties.template
-rw-r--r--, 1 aluno aluno 1110 May  8 2015 flume-env.ps1.template
-rw-r--r--, 1 aluno aluno 1197 Sep 30 19:29 flume-env.sh
-rw-r--r--, 1 aluno aluno 1214 May  8 2015 flume-env.sh.template
-rw-r--r--, 1 aluno aluno 3107 May  8 2015 log4j.properties
[aluno@dataserver conf]$ gedit flume-env.sh
```

Editar o arquivo flume-env.sh

1 / 4

*Engenharia de Dados com Hadoop e Spark 3.0*

Applications ▾ Places ▾ _gedit ▾

flume-env.sh
/opt/flume/conf

Fri 19:42 Save - X

```
# Licensed to the Apache Software Foundation (ASF) under one
# or more contributor license agreements. See the NOTICE file
# distributed with this work for additional information
# regarding copyright ownership. The ASF licenses this file
# to you under the Apache License, Version 2.0 (the
# "License"); you may not use this file except in compliance
# with the License. You may obtain a copy of the License at
#
#     http://www.apache.org/licenses/LICENSE-2.0
#
# Unless required by applicable law or agreed to in writing, software
# distributed under the License is distributed on an "AS IS" BASIS,
# WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
# See the License for the specific language governing permissions and
# limitations under the License.

# If this file is placed at FLUME_CONF_DIR/flume-env.sh, it will be sourced
# during Flume startup.

# Environment variables can be set here.

export JAVA_HOME=/opt/jdk

# Give Flume more memory and pre-allocate, enable remote monitoring via JMX
# export JAVA_OPTS="-Xms100m -Xmx2000m -Dcom.sun.management.jmxremote"

# Note that the Flume conf directory is always included in the classpath.
#FLUME_CLASSPATH=""

|
```

sh ▾ Tab Width: 8 ▾ Ln 29, Col 1 ▾ INS

aluno@dataserver:/opt/flume/conf flume-env.sh (/opt/flume/conf) ~ ...

1 / 4

Acrescentar o JAVA_HOME

*Engenharia de Dados com Hadoop e Spark 3.0*A screenshot of a terminal window titled "Terminal". The window shows the command "flume-ng" being typed at the prompt "[aluno@dataserver conf]\$". The window has a standard Linux-style title bar with "Applications", "Places", and "Terminal" buttons, and a status bar at the bottom right showing "Fri 19:43".

```
File Edit View Search Terminal Help  
[aluno@dataserver conf]$ flume-ng
```

Testar a instalação

1 / 4

*Engenharia de Dados com Hadoop e Spark 3.0*

aluno@dataserver:/opt/flume/conf Fri 19:44

```

File Edit View Search Terminal Help
commands:
  help           display this help text
  agent          run a Flume agent
  avro-client    run an avro Flume client
  version        show Flume version info

global options:
  --conf,-c <conf>      use configs in <conf> directory
  --classpath,-C <cp>    append to the classpath
  --dryrun,-d            do not actually start Flume, just print the command
  --plugins-path <dirs>  colon-separated list of plugins.d directories. See the
                        plugins.d section in the user guide for more details.
                        Default: $FLUME_HOME/plugins.d
  -Dproperty=value     sets a Java system property value
  -Xproperty=value     sets a Java -X option

agent options:
  --name,-n <name>       the name of this agent (required)
  --conf-file,-f <file>   specify a config file (required if -z missing)
  --zkConnString,-z <str> specify the ZooKeeper connection to use (required if -f missing)
  --zkBasePath,-p <path>  specify the base path in ZooKeeper for agent configs
  --no-reload-conf       do not reload config file if changed
  --help,-h               display help text

avro-client options:
  --rpcProps,-P <file>   RPC client properties file with server connection params
  --host,-H <host>        hostname to which events will be sent
  --port,-p <port>        port of the avro source
  --dirname <dir>         directory to stream to avro source
  --filename,-F <file>    text file to stream to avro source (default: std input)
  --headerFile,-R <file>  File containing event headers as key/value pairs on each new line
  --help,-h               display help text

Either --rpcProps or both --host and --port must be specified.

Note that if <conf> directory is specified, then it is always included first
in the classpath.

[aluno@dataserver conf]$ ■

```

1 / 4

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.

A screenshot of a terminal window titled "Terminal". The window shows a command-line session:

```
File Edit View Search Terminal Help
[aluno@dataserver ~]$ su
Password:
[root@dataserver alumno]# cd /etc/yum.repos.d/
[root@dataserver yum.repos.d]#
```

The terminal is running on a Linux system with a desktop environment, as evidenced by the window title bar which includes "Applications", "Places", and "Terminal". The status bar at the bottom right indicates the date and time: "Fri 19:57".

Conectado como root, acessar o diretório de repositórios do CentOS

*Engenharia de Dados com Hadoop e Spark 3.0*A screenshot of a terminal window titled "Terminal" at "aluno@dataserver:/etc/yum.repos.d". The window shows a command-line session where a user is downloading the Ambari repository. The command used is "wget http://public-repo-1.hortonworks.com/ambari/centos7/2.x/updates/2.2.0.0/ambari.repo". The progress bar indicates the download is at 100% completion, with a speed of 288 KB/s and a duration of 0s. The download file is named "ambari.repo".

```
File Edit View Search Terminal Help
[aluno@dataserver ~]$ su
Password:
[root@dataserver alumno]# cd /etc/yum.repos.d/
[root@dataserver yum.repos.d]# wget http://public-repo-1.hortonworks.com/ambari/centos7/2.x/updates/2.2.0.0/ambari.repo
--2016-09-30 19:58:45 - http://public-repo-1.hortonworks.com/ambari/centos7/2.x/updates/2.2.0.0/ambari.repo
Resolving public-repo-1.hortonworks.com (public-repo-1.hortonworks.com)... 52.84.16.206, 52.84.16.45, 52.84.16.1
13, ...
Connecting to public-repo-1.hortonworks.com (public-repo-1.hortonworks.com)|52.84.16.206|:80... connected.
HTTP request sent, awaiting response... 200 OK
Length: 288 [binary/octet-stream]
Saving to: 'ambari.repo'

100%[=====] 288 --.-K/s in 0s

2016-09-30 19:58:45 (37.9 MB/s) - 'ambari.repo' saved [288/288]

[root@dataserver yum.repos.d]#
```

Download do arquivo de repositório do Ambari

1 / 4

*Engenharia de Dados com Hadoop e Spark 3.0*A screenshot of a Linux desktop environment showing a terminal window. The window title is "Terminal". The terminal prompt shows "aluno@dataserver:/etc/yum.repos.d" and the command "[root@dataserver yum.repos.d]# yum install ambari-server" is being typed. The status bar at the bottom of the terminal window shows "Fri 19:59".

File Edit View Search Terminal Help
[root@dataserver yum.repos.d]# yum install ambari-server

Como root, executar: yum install ambari-server

*Engenharia de Dados com Hadoop e Spark 3.0*

```

Applications Places Terminal Fri 19:59
aluno@dataserver:/etc/yum.repos.d

File Edit View Search Terminal Help
Loading mirror speeds from cached hostfile
 * base: mirror.its.sfu.ca
 * extras: centos.mirror.rafael.ca
 * updates: centos.mirror.rafael.ca
Resolving Dependencies
--> Running transaction check
--> Package ambari-server.x86_64 0:2.2.0.0-1310 will be installed
--> Processing Dependency: postgresql-server >= 8.1 for package: ambari-server-2.2.0.0-1310.x86_64
--> Running transaction check
--> Package postgresql-server.x86_64 0:9.2.15-1.el7_2 will be installed
--> Processing Dependency: postgresql-libs(x86-64) = 9.2.15-1.el7_2 for package: postgresql-server-9.2.15-1.el7_2.x86_64
--> Processing Dependency: postgresql(x86-64) = 9.2.15-1.el7_2 for package: postgresql-server-9.2.15-1.el7_2.x86_64
--> Processing Dependency: libpq.so.5()(64bit) for package: postgresql-server-9.2.15-1.el7_2.x86_64
--> Running transaction check
--> Package postgresql.x86_64 0:9.2.15-1.el7_2 will be installed
--> Package postgresql-libs.x86_64 0:9.2.15-1.el7_2 will be installed
--> Finished Dependency Resolution

Dependencies Resolved

=====
Package           Arch      Version       Repository      Size
=====
Installing:
ambari-server    x86_64   2.2.0.0-1310   Updates-ambari-2.2.0.0   406 M
Installing for dependencies:
postgresql        x86_64   9.2.15-1.el7_2   updates          3.0 M
postgresql-libs   x86_64   9.2.15-1.el7_2   updates          231 k
postgresql-server x86_64   9.2.15-1.el7_2   updates          3.8 M

Transaction Summary
=====
Install 1 Package (+3 Dependent packages)

Total download size: 418 M
Installed size: 465 M
Is this ok [y/d/N]: ■

```

Instalação do Ambari

1 / 4

*Engenharia de Dados com Hadoop e Spark 3.0*

```

File Edit View Search Terminal Help
Is this ok [y/d/N]: y
Downloading packages:
(1/4): postgresql-libs-9.2.15-1.el7_2.x86_64.rpm | 231 KB 00:00:00
(2/4): postgresql-server-9.2.15-1.el7_2.x86_64.rpm | 3.8 MB 00:00:02
(3/4): postgresql-9.2.15-1.el7_2.x86_64.rpm | 3.0 MB 00:00:02
warning: /var/cache/yum/x86_64/7/Updates-ambari-2.2.0.0/packages/ambari-server-2.2.0.0-1310.x86_64.rpm: Header V4 RSA/SHA1 Signature, key ID 07513cad; NOKEY
Public key for ambari-server-2.2.0.0-1310.x86_64.rpm is not installed
(4/4): ambari-server-2.2.0.0-1310.x86_64.rpm | 406 MB 00:03:52
-----
Total 1.8 MB/s | 413 MB 00:03:52
Retrieving key from http://public-repo-1.hortonworks.com/ambari/centos7/RPM-GPG-KEY/RPM-GPG-KEY-Jenkins
Importing GPG key 0x07513CAD:
  Userid : "Jenkins (HDP Builds) <jenkin@hortonworks.com>"
  Fingerprint: df52 ed4f 7ada 5882 c099 4c66 b973 3a7a 0751 3cad
  From   : http://public-repo-1.hortonworks.com/ambari/centos7/RPM-GPG-KEY/RPM-GPG-KEY-Jenkins
Is this ok [y/N]: y
Running transaction check
Running transaction test
Transaction test succeeded
Running transaction
  Installing : postgresql-libs-9.2.15-1.el7_2.x86_64 1/4
  Installing : postgresql-9.2.15-1.el7_2.x86_64 2/4
  Installing : postgresql-server-9.2.15-1.el7_2.x86_64 3/4
  Installing : ambari-server-2.2.0.0-1310.x86_64 4/4
  verifying   : postgresql-libs-9.2.15-1.el7_2.x86_64 1/4
  verifying   : postgresql-server-9.2.15-1.el7_2.x86_64 2/4
  verifying   : ambari-server-2.2.0.0-1310.x86_64 3/4
  verifying   : postgresql-9.2.15-1.el7_2.x86_64 4/4
Installed:
  ambari-server.x86_64 0:2.2.0.0-1310

Dependency Installed:
  postgresql.x86_64 0:9.2.15-1.el7_2           postgresql-libs.x86_64 0:9.2.15-1.el7_2
  postgresql-server.x86_64 0:9.2.15-1.el7_2

Complete!
[root@dataserver yum.repos.d]#

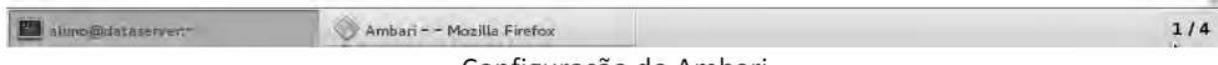
```

Instalação concluída com sucesso

*Engenharia de Dados com Hadoop e Spark 3.0*

A screenshot of a terminal window titled "Terminal". The window shows the command `sudo ambari-server setup` being run by the user `aluno`. The terminal output includes messages about SELinux status and mode, and a warning about SELinux being set to 'permissive' mode. The terminal interface includes standard menu options like File, Edit, View, Search, Terminal, Help, and system status indicators at the top right.

```
File Edit View Search Terminal Help
[aluno@dataserver ~]$ sudo ambari-server setup
[sudo] password for aluno:
Using python /usr/bin/python2
Setup ambari-server
Checking SELinux...
SELinux status is 'enabled'
SELinux mode is "enforcing"
Temporarily disabling SELinux
WARNING: SELinux is set to 'permissive' mode and temporarily disabled.
OK to continue [y/n] (y)? ■
```



*Engenharia de Dados com Hadoop e Spark 3.0*

```
File Edit View Search Terminal Help
[aluno@dataserver ~]$ sudo ambari-server setup
Using python /usr/bin/python2
Setup ambari-server
Checking SELinux...
SELinux status is 'enabled'
SELinux mode is 'permissive'
WARNING: SELinux is set to 'permissive' mode and temporarily disabled.
OK to continue [y/n] (y)? y
Customize user account for ambari-server daemon [y/n] (n)? y
Enter user account for ambari-server daemon (root);aluno
Adjusting ambari-server permissions and ownership...
Checking firewall status...
Redirecting to /bin/systemctl status iptables.service

Checking JDK...
[1] Oracle JDK 1.8 + Java Cryptography Extension (JCE) Policy Files 8
[2] Oracle JDK 1.7 + Java Cryptography Extension (JCE) Policy Files 7
[3] Custom JDK
=====
Enter choice (1); 3
WARNING: JDK must be installed on all hosts and JAVA_HOME must be valid on all hosts.
WARNING: JCE Policy files are required for configuring Kerberos security. If you plan to use Kerberos, please make sure JCE Unlimited Strength Jurisdiction Policy Files are valid on all hosts.
Path to JAVA_HOME: /opt/jdk
validating JDK on Ambari Server...done.
Completing setup...
Configuring database...
Enter advanced database configuration [y/n] (n)?
```



Configuração do Ambari

1 / 4

*Engenharia de Dados com Hadoop e Spark 3.0*

File Edit View Search Terminal Help
[aluno@dataserver ~]\$ sudo ambari-server setup
Using python /usr/bin/python2
Setup ambari-server
Checking SELinux...
SELinux status is 'enabled'
SELinux mode is 'permissive'
WARNING: SELinux is set to 'permissive' mode and temporarily disabled.
OK to continue [y/n] (y)? y
Customize user account for ambari-server daemon [y/n] (n)? y
Enter user account for ambari-server daemon (root);aluno
Adjusting ambari-server permissions and ownership...
Checking firewall status...
Redirecting to /bin/systemctl status iptables.service

Checking JDK...
[1] Oracle JDK 1.8 + Java Cryptography Extension (JCE) Policy Files 8
[2] Oracle JDK 1.7 + Java Cryptography Extension (JCE) Policy Files 7
[3] Custom JDK
=====
Enter choice (1); 3
WARNING: JDK must be installed on all hosts and JAVA_HOME must be valid on all hosts.
WARNING: JCE Policy files are required for configuring Kerberos security. If you plan to use Kerberos, please make sure JCE Unlimited Strength Jurisdiction Policy Files are valid on all hosts.
Path to JAVA_HOME: /opt/jdk
validating JDK on Ambari Server...done.
Completing setup...
Configuring database...
Enter advanced database configuration [y/n] (n)? y
Configuring database...
=====
Choose one of the following options:
[1] - PostgreSQL (Embedded)
[2] - Oracle
[3] - MySQL
[4] - PostgreSQL
[5] - Microsoft SQL Server (Tech Preview)
[6] - SQL Anywhere
=====
Enter choice (1); 1

Configuração do Ambari

1 / 4

*Engenharia de Dados com Hadoop e Spark 3.0*

```

Applications Places Terminal Fri 20:09
aluno@dataserver:~>

File Edit View Search Terminal Help
[1] Oracle JDK 1.8 + Java Cryptography Extension (JCE) Policy Files 8
[2] Oracle JDK 1.7 + Java Cryptography Extension (JCE) Policy Files 7
[3] Custom JDK
=====
Enter choice (1): 3
WARNING: JDK must be installed on all hosts and JAVA_HOME must be valid on all hosts.
WARNING: JCE Policy files are required for configuring Kerberos security. If you plan to use Kerberos, please make sure JCE Unlimited Strength Jurisdiction Policy Files are valid on all hosts.
Path to JAVA_HOME: /opt/jdk
Validating JDK on Ambari Server...done.
Completing setup...
Configuring database...
Enter advanced database configuration [y/n] (n)? y
Configuring database...
=====
Choose one of the following options:
[1] - PostgreSQL (Embedded)
[2] - Oracle
[3] - MySQL
[4] - PostgreSQL
[5] - Microsoft SQL Server (Tech Preview)
[6] - SQL Anywhere
=====
Enter choice (1): 1
Database name (ambari): ambari
Postgres schema (ambari): ambari
Username (ambari): ambari
Enter Database Password (bigdata):
Re-enter password:
Default properties detected. Using built-in database.
Configuring ambari database...
Checking PostgreSQL...
Running initdb: This may take upto a minute.
Initializing database ... OK

About to start PostgreSQL
Configuring local database...

```

■

Configuração em andamento

*Engenharia de Dados com Hadoop e Spark 3.0*

```
File Edit View Search Terminal Help
Path to JAVA_HOME: /opt/jdk
validating JDK on Ambari Server...done.
Completing setup...
Configuring database...
Enter advanced database configuration [y/n] (n)? y
Configuring database...
=====
Choose one of the following options:
[1] - PostgreSQL (Embedded)
[2] - Oracle
[3] - MySQL
[4] - PostgreSQL
[5] - Microsoft SQL Server (Tech Preview)
[6] - SQL Anywhere
=====
Enter choice (1); 1
Database name (ambari); ambari
Postgres schema (ambari); ambari
Username (ambari); ambari
Enter Database Password (bigdata);
Re-enter password:
Default properties detected. Using built-in database.
Configuring ambari database...
Checking PostgreSQL...
Running initdb; This may take upto a minute.
Initializing database ... OK

About to start PostgreSQL
Configuring local database...
Connecting to local database...done.
Configuring PostgreSQL...
Restarting PostgreSQL
Extracting system views...
ambari-admin-2.2.0.0.1310.jar
...
Adjusting ambari-server permissions and ownership...
Ambari Server 'setup' completed successfully.
[aluno@dataserver ~]$
```

1 / 4

Configuração concluída

*Engenharia de Dados com Hadoop e Spark 3.0*A screenshot of a terminal window titled "Terminal". The window has a title bar with "Applications", "Places", and "Terminal". The status bar shows "Fri 20:10". The terminal prompt is "aluno@dataserver:~". Below the prompt, the command "[aluno@dataserver ~]\$ ambari-server start" is visible. The window is set against a dark background.

*Engenharia de Dados com Hadoop e Spark 3.0*

The screenshot shows a terminal window titled 'Terminal'. The title bar includes 'Applications', 'Places', and 'Terminal'. The status bar at the top right shows 'Fri 20:11'. The terminal prompt is 'aluno@dataserver:~\$'. The output of the command 'ambari-server start' is displayed:

```
File Edit View Search Terminal Help
[aluno@dataserver ~]$ ambari-server start
Using python /usr/bin/python2
Starting ambari-server
Unable to check PostgreSQL server status when starting without root privileges.
Please do not forget to start PostgreSQL server.
Organizing resource files at /var/lib/ambari-server/resources...
Unable to check firewall status when starting without root privileges.
Please do not forget to disable or adjust firewall if needed
/usr/bin/sh: line 0: ulimit: open files: cannot modify limit: Operation not permitted
WARNING: setpgid(11188, 0) failed - [Errno 13] Permission denied
Server PID at: /var/run/ambari-server/ambari-server.pid
Server out at: /var/log/ambari-server/ambari-server.out
Server log at: /var/log/ambari-server/ambari-server.log
Waiting for server start.....
Ambari Server 'start' completed successfully.
[aluno@dataserver ~]$
```

Inicializado

1 / 4



Data Science Academy phelipe.utsemprboni@outlook.com 5c8a62005e4cde1acb8b45a3

Engenharia de Dados com Hadoop e Spark 3.0

A screenshot of a Mozilla Firefox browser window. The title bar says "Ambari - Mozilla Firefox". The address bar shows "localhost:8080/#/Login". The main content area displays the Ambari "Sign in" page. It has fields for "Username" (admin) and "Password" (*****), and a green "Sign in" button. At the bottom of the page, there is a note about Apache License 2.0 and third-party tools/resources used by Ambari.

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

*Engenharia de Dados com Hadoop e Spark 3.0*

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!