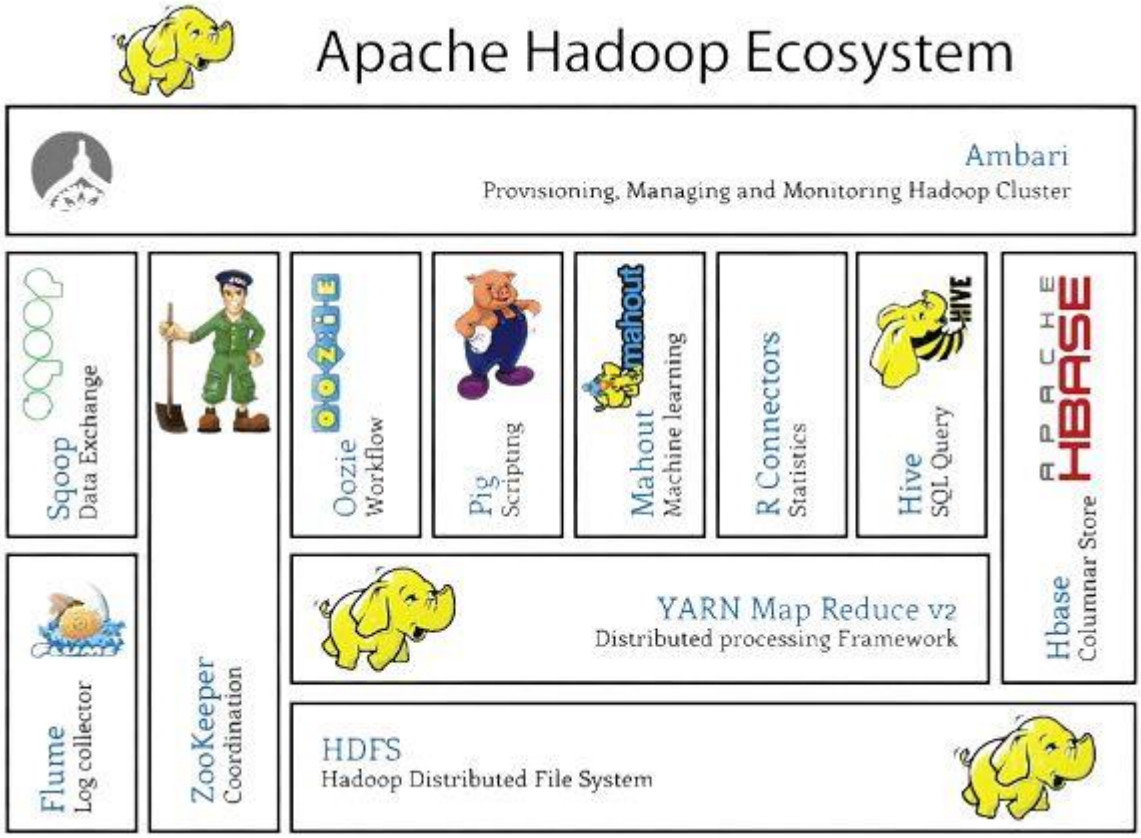
Configuração do Ecossistema Hadoop

Sempre ouvi dizer que um Engenheiro de Dados tem que saber montar um ambiente com o ecossistema Hadoop. E nesse momento, como eu estou montando um ambiente de testes na minha máquina e, aproveito para consolidar os conhecimentos adquiridos.

Em uma publicação recente que fiz no meu linkedin, falei dos principais motivos de aprender esse Framework. Dê uma olhada na publicação:

https://www.linkedin.com/feed/update/urn:li:activity:6924423796195999744/



Vamos ao trabalho porque a atividade é longa!!

***Essa documentação está em constante atualização uma vez que estou em processo de instalação.**

Itens a serem instalados	Instalação
1- Virtual Box	completo
2- Sistema Operacional CentOS 7.6 (64 bits)	completo
3- MySQL	completo
4- Shh (protocolo)	completo
5- Java	completo
6- Instalação do Hadoop	completo
7- Java	completo
8- Apache Hadoop	completo
9- Apache Zookeeper	pendente
10- Apache Hbase	pendente
11- Apache Hive	pendente
12- Apache Pig	pendente
13- Apache Spark	pendente
14- Apache Sqoop	pendente
15- Apache Flume	pendente
16- Apache Ambari	pendente

Etapas Principais

- Baixar o Virtual Box
- Baixar o CentOS (SO)

A versão de Sistema Operacional mais usado no ambiente corporativo é o RED HAT. Voltado para grandes volumes de Dados. Como é para uso doméstico usei o CentOS além de ser o mais indicado para ambiente de servidores

No Virtual Box, alocar uma quantidade de memória significativa uma vez que o Haddop requer bastante memória. Acabei instalando uma série de ferramentas do Virtual Box para otimizar e melhorar a integração da máquina física com a máquina virtual.

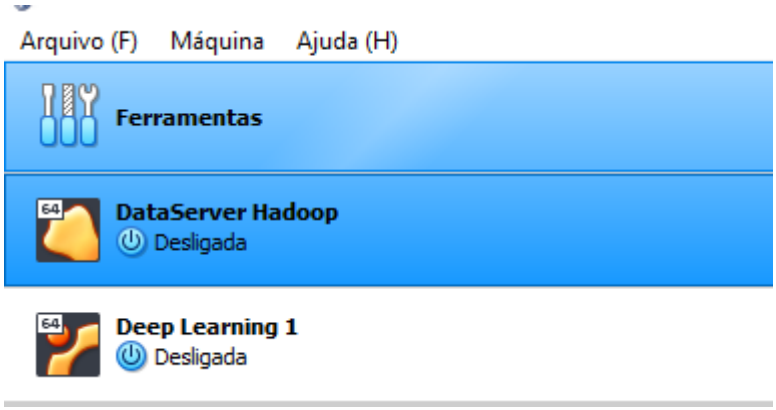
Em função de ter que executar algumas rotinas o meu usuário não era administrador. Logo acrescentei o usuário na linha do código.

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

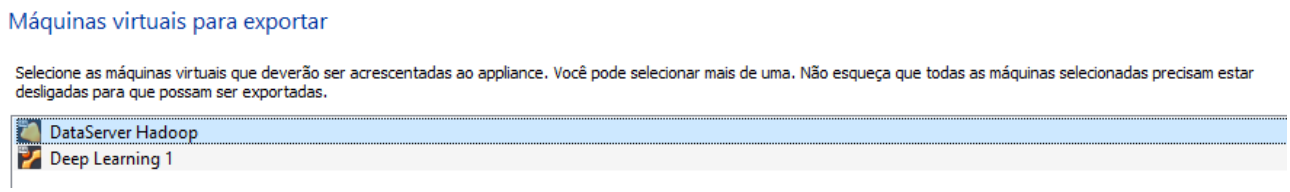
Agora eu consigo executar o comando: **sudo yum update kernel**, pois eu dei permissão para o usuário taraneh executar a atividade de atualização.

Ainda faltam 03 pacotes; **sudo yum install gcc make perl**

Momento de fazer um backup da instalação do SO para o apache Hadoop. Se algo der errado na instalação adiante, começo a partir daqui.

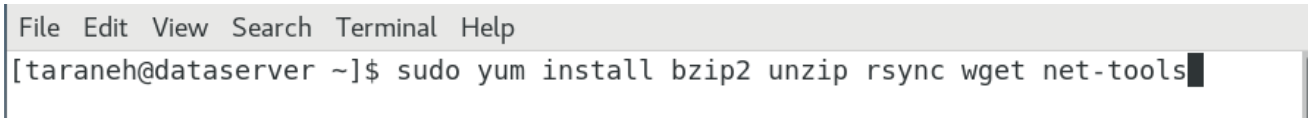


Selecione Arquivo > exportar appliance



Aproveito para instação de alguns utilitários.

sudo yum install bzip2 unzip rsync wget net-tools



```
[taraneh@dataserver ~]$ sudo yum install bzip2 unzip rsync wget net-tools
[sudo] password for taraneh:
Loaded plugins: fastestmirror, langpacks
Loading mirror speeds from cached hostfile
 * base: mirror.uepg.br
 * extras: espejito.fder.edu.uy
 * updates: espejito.fder.edu.uy
Package bzip2-1.0.6-13.el7.x86_64 already installed and latest version
Package rsync-3.1.2-10.el7.x86_64 already installed and latest version
Package wget-1.14-18.el7_6.1.x86_64 already installed and latest version
Package net-tools-2.0-0.25.20131004git.el7.x86_64 already installed and latest v
ersion
Resolving Dependencies
--> Running transaction check
---> Package unzip.x86_64 0:6.0-21.el7 will be updated
---> Package unzip.x86_64 0:6.0-24.el7_9 will be an update
--> Finished Dependency Resolution
```

Instalação do MYSQL

- Baixei a versão;

sudo yum localinstall https://dev.mysql.com/get/mysql80-community-release-el8-3.noarch.rpm

-Executei a instalação do MYSQL
-Ativei o serviço do MySQL
-Iniciei o MySQL

```
Complete!
[taraneh@dataserver ~]$ sudo systemctl enable mysqld
[taraneh@dataserver ~]$ sudo systemctl start mysqld
```

Instalação do servidor ssh (secure shell)

Instalei esse importante protocolo de comunicação criptografada entre as máquinas do cluster. Nesse ambiente que estou mandando teremos um cluster de uma máquina só. Como eu estou montando um cluster pseudo- distribuído, ou seja, um name-node e data-node na mesma máquina e, portanto, irão se comunicar via ssh.

- Executei a instalação do ssh

- Ativei o serviço do ssh

- Iniciei o ssh

```
[taraneh@dataserver Downloads]$ sudo systemctl restart sshd
[sudo] password for taraneh:
[taraneh@dataserver Downloads]$ sudo systemctl status sshd
● sshd.service - OpenSSH server daemon
   Loaded: loaded (/usr/lib/systemd/system/sshd.service; enabled; vendor preset
   Active: active (running) since Thu 2022-05-19 18:44:11 -03; 1min 12s ago
     Docs: man:sshd(8)
           man:sshd_config(5)
 Main PID: 6660 (sshd)
    Tasks: 1
   CGroup: /system.slice/sshd.service
           └─6660 /usr/sbin/sshd -D

May 19 18:44:11 dataserver systemd[1]: Stopped OpenSSH server daemon.
May 19 18:44:11 dataserver systemd[1]: Starting OpenSSH server daemon...
May 19 18:44:11 dataserver sshd[6660]: Server listening on 0.0.0.0 port 22.
May 19 18:44:11 dataserver systemd[1]: Started OpenSSH server daemon.
[taraneh@dataserver Downloads]$
```

O Apache Hadoop é uma aplicação desenvolvida em Java, que roda sobre uma JVM (java virtual machine). No mercado existe a OpenJDk e o Oracle JDK, mas por padrão se usa mais o Oracle JDK. Logo, desinstalei a versão OpenJDk e instalei Oracle JDK.

Instalado o java

```
[taraneh@dataserver ~]$ source .bashrc
[taraneh@dataserver ~]$ java -version
java version "1.8.0_331"
Java(TM) SE Runtime Environment (build 1.8.0_331-b09)
Java HotSpot(TM) 64-Bit Server VM (build 25.331-b09, mixed mode)
[taraneh@dataserver ~]$
```

Configurei também as variáveis de ambiente (arquivo.bashrc)

```
Open   *.bashrc   Save
~/
# .bashrc

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

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

# User specific aliases and functions

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

Para a criação do apache Hadoop criei o usuário hadoop. Por questões de boas práticas, não utilizei o usuário root para esta instalação.

```
File  Edit  View  Search  Terminal  Help
[hadoop@dataserver ~]$ su
Password:
[root@dataserver hadoop]# gedit /etc^C
[root@dataserver hadoop]# gedit /etc/sudoers
```

Instalação do Hadoop

```
hadoop@dataserver:/opt
File  Edit  View  Search  Terminal  Help
[hadoop@dataserver ~]$ cd /opt/
[hadoop@dataserver opt]$ ls -la
total 0
drwxr-xr-x.  6 root    root     74 May 20 20:18 .
dr-xr-xr-x. 17 root    root    224 May 12 16:17 ..
drwxr-xr-x.  9 hadoop  hadoop  149 Mar 19 22:58 hadoop
drwxr-xr-x.  8 taraneh taraneh 273 Mar 10 08:27 jdk
drwxr-xr-x.  2 root    root      6 Oct 30  2018 rh
drwxr-xr-x.  8 root    root    136 May 13 18:06 VBoxGuestAdditions-6.1.22
[hadoop@dataserver opt]$
```

Configurado as variáveis de ambiente.bashrc .

Vamos validar se essas configurações estão ok.

```
[hadoop@dataserver ~]$ source .bashrc
[hadoop@dataserver ~]$
```

Até aqui, toda a configuração realizada com sucesso. ☺

```
[hadoop@dataserver ~]$ hadoop version
Hadoop 3.2.3
Source code repository https://github.com/apache/hadoop -r abe5358143720085498613d399be3bbf01e0f131
Compiled by ubuntu on 2022-03-20T01:18Z
Compiled with protoc 2.5.0
From source with checksum 39bb14faec14b3aa25388a6d7c345fe8
This command was run using /opt/hadoop/share/hadoop/common/hadoop-common-3.2.3.jar
[hadoop@dataserver ~]$
```