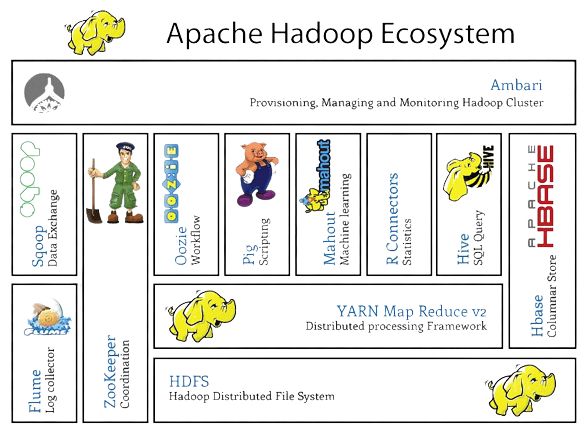
Configuração do Ecossistema Hadoop

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 |
| 1. Sistema Operacional CentOS 7.6 (64 bits) | completo |
| 1. MySQL | completo |
| 1. Shh (protocolo) | completo |
| 1. Java | completo |
| 1. Instalação do Hadoop | completo |
| 1. Java | completo |
| 1. Apache Hadoop | completo |
| 1. Apache Zookeeper | pendente |
| 1. Apache Hbase | pendente |
| 1. Apache Hive | pendente |
| 1. Apache Pig | pendente |
| 1. Apache Spark | pendente |
| 1. Apache Sqoop | pendente |
| 1. Apache Flume | pendente |
| 1. Apache Ambari | pendente |

**Etapas Principais**

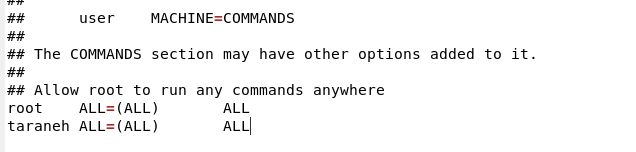
- Baixar o Virtual Box

- 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.

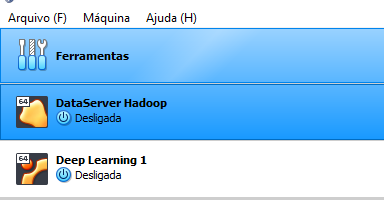
Em função de ter que executar algumas rotinas o meu usuário não era administrador. Logo acrescentei o usuário na linha do código.



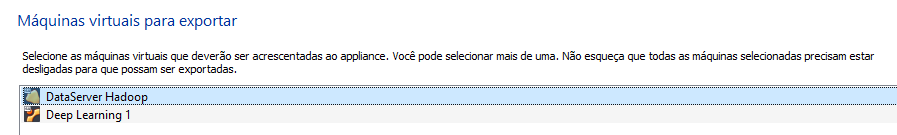
Agora eu consigo executar o comando: **sudo yum update kernel**, pois eu dei permissão para o usuário taraneh executar a atividade de atualização.

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.

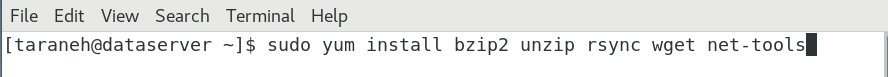


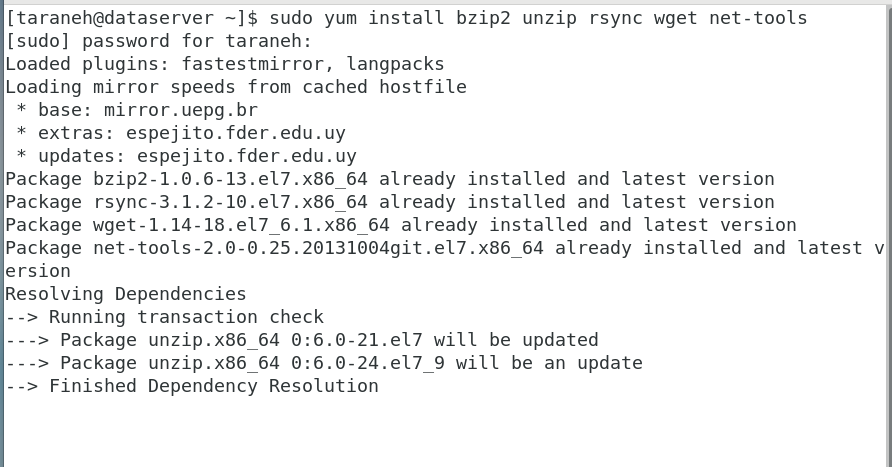
Selecione Arquivo > exportar appliance



Aproveito para instação de alguns utilitários.

sudo yum install bzip2 unzip rsync wget net-tools



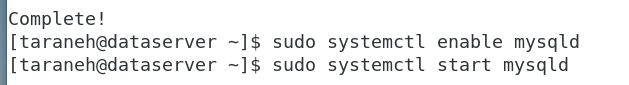


Instalação do MYSQL

- Baixei a versão;

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

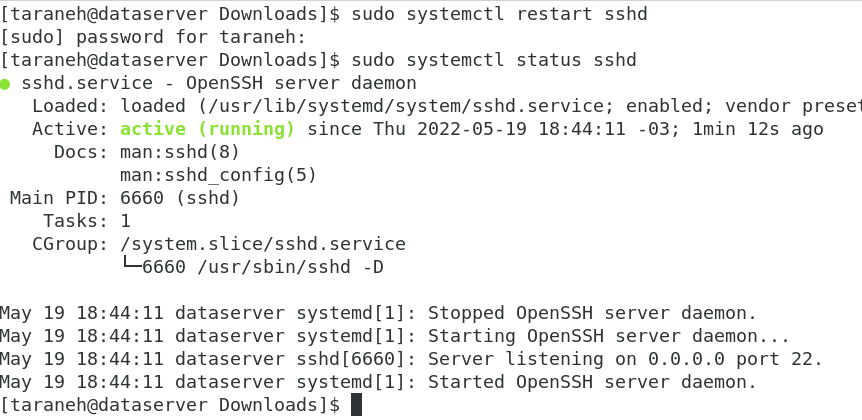
-Executei a instalação do MYSQL  
-Ativei o serviço do MySQL  
-Iniciei o MySQL



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.

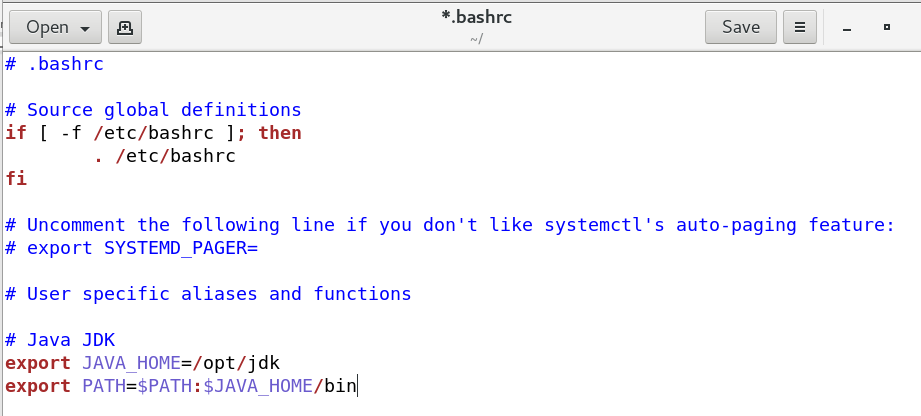
- Executei a instalação do ssh  
- Ativei o serviço do ssh   
- Iniciei o ssh



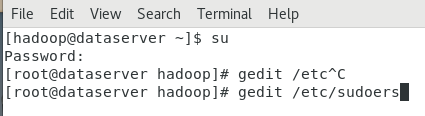
O Apache Hadoop é uma aplicação desenvolvida em Java, que roda sobre uma JVM (java virtual machine). No mercado existe a OpenJDk e o Oracle JDK, mas por padrão se usa mais o Oracle JDK. Logo, desinstalei a versão OpenJDk e instalei Oracle JDK.

Instalado o java

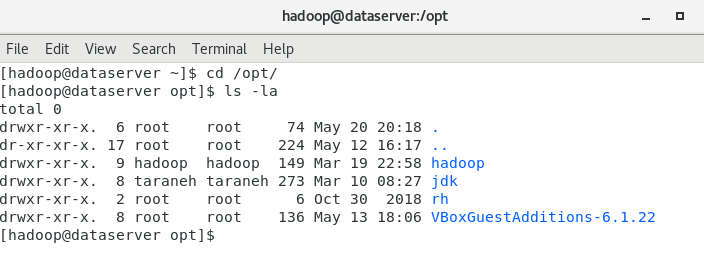
Configurei também as variáveis de ambiente (arquivo.bashrc)



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



**Instalação do Hadoop**



Configurado as variáveis de ambiente.bashrc .

Vamos validar se essas configurações estão ok.



Até aqui, toda a configuração realizada com sucesso. ☺

