

Apache Hadoop – Hands-On

Objetivo: O objetivo deste hands-on é criar uma máquina virtual Linux e configurar o Apache Hadoop no modo pseudo-distribuído.

Passos

1. Crie uma máquina virtual Ubuntu 16.04 – 64 bits (Xenial) usando VirtualBox e Vagrant (<https://www.vagrantup.com/>)

```
$ mkdir ~/HadoopVM
$ cd ~/HadoopVM
$ vagrant init ubuntu/xenial64
```

2. Editar o arquivo Vagrantfile criado na pasta local. Incluir as linhas abaixo antes do “end” final.

```
# Forward ports
# HDFS Namenode Web UI
config.vm.network "forwarded_port", guest: 50070, host: 50070, host_ip: "127.0.0.1"
# YARN Resource Manager Web UI
config.vm.network "forwarded_port", guest: 8088, host: 8088, host_ip: "127.0.0.1"

config.vm.provider "virtualbox" do |vb|
  # 2 cores
  vb.cpus = 2
  # 4GB of RAM
  vb.memory = 4096
end
config.vm.provision "shell", privileged:true, path: "setupadmin.sh"
config.vm.provision "shell", privileged:false, path: "setuphadoop.sh"
config.vm.provision "shell", privileged:false, path: "startdaemons.sh", run: "always"
config.vm.provision "shell", privileged:false, path: "confighdfsdirs.sh"
```

3. Baixar os arquivos setupadmin.sh, setuphadoop.sh, startdaemons.sh e confighdfsdirs.sh da pasta “02_Hadoop/HadoopVM” do repositório do GitHub (<https://github.com/GerlInfraBigDataPUCRS/CursoDataScience2018/>) na mesma pasta do arquivo Vagrantfile.
4. Iniciar a máquina virtual e esperar completar o provisionamento.

```
$ vagrant up
```

5. Verificar se todos os daemons estão executando (ResourceManager, NodeManager, DataNode, Jps, NameNode)

```
$ vagrant ssh -c jps
```

6. Acessar as interfaces do NameNode e ResourceManager usando navegador da máquina local

```
NameNode - http://localhost:50070
ResourceManager - http://localhost:8088
```

7. Executar a aplicação Pi usando MapReduce dentro da VM

```
$ vagrant ssh
$ $HADOOP_HOME/bin/hadoop jar $HADOOP_HOME/share/hadoop/mapreduce/hadoop-mapreduce-examples-2.9.1.jar pi 16 1000
```