**Orientações para instalar o HADOOP no AWS**

**OBSERVAÇÃO IMPORTANTE**: Antes de executar cada passo abaixo leia seus subitens, pois eles contém informações importantes para cada passo. Só execute os passos após ler tudo a seu respeito.

**Passos**

1. Siga os passos do artigo <https://dzone.com/articles/how-set-multi-node-hadoop?mz=111665-bigdata>.
   1. Ao escolher a Amazon Machine Image (AMI) eu usei o Ubuntu 16.04 LTS 64-bit (x86), pois foi a mesma usada nas aulas em vídeo. Não esqueça de escolher uma instância “Free tier eligible”.
   2. No artigo, a orientação é configurar 4 instâncias (passo 1.5). Configure somente 1 (uma) para ficar igual ao ambiente disponibilizado pelo IGTI.
   3. No meu caso, eu não utilizei o WinSCP, mas o PSCP.EXE que vem com o PUTTY.ZIP (<https://www.chiark.greenend.org.uk/~sgtatham/putty/latest.html>). Apesar da interface ser através da linha de comando, é bem simples. Por exemplo, para enviar algum arquivo basta abrir o prompt do DOS e digitar: “***pscp -i <arquivo.ppk> <origem> ubuntu@<ip\_ou\_nomeDNS\_do servidor\_AWS:.***”. Lembrando que caminhos com espaço devem estar entre aspas e que o arquivo PPK foi gerado durante o passo 1.

Um exemplo de comando seria:

***D:\>*** ***pscp -i “D:\Tarefa 3\arquivo.ppk” “D:\Tarefa 3\covidData.txt” ubuntu@ec2-21-197-213-11.us-east-2.compute.amazonaws.com:.***

* 1. Não siga para a parte 2 do artigo, pois ela é para configurar um cluster com 4 nós e usa uma versão antiga do HADOOP.

1. Após terminar o passo 1 e conseguir se conectar ao servidor através do Putty (putty.exe), deve-se fazer algumas atualizações e configurações:
   1. Para atualizar packages e dependências execute o seguinte comando no terminal:

*sudo apt-get update*

1. Como o Java é um pré-requisito do HADOOP, é necessário instalar o JDK. De acordo com a documentação do Hadoop (https://cwiki.apache.org/confluence/display/HADOOP/Hadoop+Java+Versions), a versão 3.3.0 do Hadoop suporta a versão 8 do Java.
   1. Acesse o site de download do Java em <https://www.oracle.com/java/technologies/javase/javase-jdk8-downloads.html>
   2. A versão do Java disponível atualmente é a 8u271 (a que instalei foi a 8u261, mas isso não importa desde que seja versão 8). Na configuração do AWS escolhemos uma máquina 64-bit (x86) do Ubuntu, então o arquivo correto do Java é o “jdk-8u271-linux-x64.tar.gz” (a versão RPM é para sistemas baseados em Red Hat; não é o caso do Ubuntu). Clique no arquivo citado. Isso vai abrir uma pop-up para aceitar os termos de licenciamento da Oracle. IMPORTANTE: Você terá que completar um cadastro na Oracle para fazer o download. Faça o download do arquivo para seu computador no diretório de sua escolha (por exemplo: D:\Install).
   3. Após o download para seu computador você deverá transferir o arquivo para o servidor AWS. Para isso pode utilizar o WinSCP ou o PSCP (da suíte Putty). O WinSCP possui uma interface gráfica e seu uso foi explicado no passo 1. Caso deseje usar o pscp.exe, basta usar a linha de comando abaixo, modificada para sua situação (diretório, arquivo PPK e nome do servidor):

*pscp -i “D:\Tarefa 3\arquivo.ppk” “D:\Install\jdk-8u271-linux-x64.tar.gz”* [*ubuntu@ec2-21-197-213-11.us-east-2.compute.amazonaws.com*](mailto:ubuntu@ec2-21-197-213-11.us-east-2.compute.amazonaws.com)*:.*

* 1. A linha de comando acima vai fazer a transferência do arquivo de instalação do Java para a pasta */home/ubuntu* do seu AWS.
  2. Acesse o servidor AWS com o Putty (putty.exe) e faça login com o usuário “ubuntu” (sem as aspas). Se conseguiu completar a parte de conexão com o AWS do passo 1 você será levado ao prompt do terminal. Para saber a pasta que se encontra utilize o comando *pwd*. Para listar o conteúdo da pasta atual execute o comando *ls -sla*; o arquivo de instalação do Java deverá aparecer na listagem de arquivos e diretórios.
  3. Execute os seguintes comandos para instalar o Java no servidor:

*sudo mkdir /usr/lib/jvm*

*cd /usr/lib/jvm*

*sudo tar -xvzf /home/ubuntu/jdk-8u271-linux-x64.tar.gz*

* 1. Para configurar corretamente o Java é necessário editar o arquivo /etc/environment. Primeiro é necessário atualizar a variável PATH. Para isso execute o comando *sudo vi /etc/environment*.
     1. O conteúdo da variável PATH, inicialmente, é:

*PATH="/usr/local/sbin:/usr/local/bin:/usr/sbin:/usr/bin:/sbin:/bin:/usr/games”*

* + 1. O conteúdo após a atualização deve ser:

*PATH="/usr/local/sbin:/usr/local/bin:/usr/sbin:/usr/bin:/sbin:/bin:/usr/games:/usr/local/games:/usr/lib/jvm/jdk1.8.0\_271/bin:/usr/lib/jvm/jdk1.8.0\_271/db/bin:/usr/lib/jvm/jdk1.8.0\_271/jre/bin"*

* + 1. Importante destacar que o “:” é o separador de caminhos (paths).
    2. Atente para a versão do Java. Pode ser que o update da versão seja diferente quando você fizer o download, então você terá que se certificar que está usando o diretório correto. Por exemplo: versão 8 update 271 será jdk1.8.0\_271; versão 8 update 261 será jdk1.8.0\_261.
    3. Além de atualizar a variável PATH você também terá que adicionar 4 linhas ao final do arquivo. Novamente, atente para a versão/update do Java. As seguintes linhas devem ser adicionadas:

*J2SDKDIR="/usr/lib/jvm/jdk1.8.0\_271"*

*J2REDIR="/usr/lib/jvm/jdk1.8.0\_271/jre"*

*JAVA\_HOME="/usr/lib/jvm/jdk1.8.0\_271"*

*DERBY\_HOME="/usr/lib/jvm/jdk1.8.0\_271/db"*

* + 1. Grave e saia do arquivo. Lembre-se que para inserir caracteres no VI basta pressionar a tecla “i”; para sair do modo de inserção basta pressionar a tecla “Esc”; antes de gravar o arquivo é necessário sair do modo de inserção; para sair e gravar basta pressionar a seguinte sequência de teclas: “:wq!” (sem as aspas; não esqueça os dois pontos). Lembre-se que o Linux diferencia entre maiúsculas e minúsculas, inclusive para os comandos do VI!
    2. Para verificar se a alteração foi realizada com sucesso execute o comando *cat /etc/environment*. O resultado deve ser igual a (lembrando que pode ser diferente dependendo da versão do Java que você utilizou):

*PATH="/usr/local/sbin:/usr/local/bin:/usr/sbin:/usr/bin:/sbin:/bin:/usr/games:/usr/local/games:/usr/lib/jvm/jdk1.8.0\_271/bin:/usr/lib/jvm/jdk1.8.0\_271/db/bin:/usr/lib/jvm/jdk1.8.0\_271/jre/bin"*

*J2SDKDIR="/usr/lib/jvm/jdk1.8.0\_271"*

*J2REDIR="/usr/lib/jvm/jdk1.8.0\_271/jre"*

*JAVA\_HOME="/usr/lib/jvm/jdk1.8.0\_271"*

*DERBY\_HOME="/usr/lib/jvm/jdk1.8.0\_271/db"*

* + 1. Caso tudo esteja certo, execute o seguinte comando para atualizar as variáveis de ambiente na sua sessão: *source /etc/environment*
    2. Para informar o Ubuntu sobre o local de instalação do Java execute os seguintes comandos:

*sudo update-alternatives --install "/usr/bin/java" "java" "/usr/lib/jvm/jdk1.8.0\_271/bin/java" 0*

*sudo update-alternatives --install "/usr/bin/javac" "javac" "/usr/lib/jvm/jdk1.8.0\_271/bin/javac" 0*

*sudo update-alternatives --set java /usr/lib/jvm/jdk1.8.0\_271/bin/java*

*sudo update-alternatives --set javac /usr/lib/jvm/jdk1.8.0\_271/bin/javac*

* + 1. Para verificar a configuração acima execute os seguintes comandos:

*update-alternatives --list java*

*update-alternatives --list javac*

* + 1. Para verificar a versão do Java execute o comando (talvez seja necessário abrir outra janela de terminal):

*java -version*

A saída desse comando deve ser parecida com:

*java version "1.8.0\_271"*

*Java(TM) SE Runtime Environment (build 1.8.0\_271-b12)*

*Java HotSpot(TM) 64-Bit Server VM (build 25.271-b12, mixed mode)*

1. Agora é a hora de instalar e configurar o Hadoop. Primeiro temos que fazer o download do arquivo de instalação. Iremos instalar a versão 3.3.0 (conforme dito anteriormente), ao invés da versão 3.0.3 usada pelo IGTI. Não parece haver muita diferença entre elas, mas a escolhida é mais recente e corrige alguns bugs e traz melhorias, de acordo com o fornecedor (<http://hadoop.apache.org/releases.html#download>).
   1. Na janela do terminal (putty.exe) execute o seguinte comando para fazer o download do Hadoop versão 3.3.0 e arquivos de verificação:

*wget* [*https://downloads.apache.org/hadoop/common/hadoop-3.3.0/hadoop-3.3.0.tar.gz*](https://downloads.apache.org/hadoop/common/hadoop-3.3.0/hadoop-3.3.0.tar.gz)

*wget* [*https://downloads.apache.org/hadoop/common/hadoop-3.3.0/hadoop-3.3.0.tar.gz.asc*](https://downloads.apache.org/hadoop/common/hadoop-3.3.0/hadoop-3.3.0.tar.gz.asc)

*wget* [*https://downloads.apache.org/hadoop/common/hadoop-3.3.0/hadoop-3.3.0.tar.gz.sha512*](https://downloads.apache.org/hadoop/common/hadoop-3.3.0/hadoop-3.3.0.tar.gz.sha512)

*wget* [*https://downloads.apache.org/hadoop/common/KEYS*](https://downloads.apache.org/hadoop/common/KEYS)

* 1. Execute os seguintes comandos para verificar a integridade dos arquivos do Hadoop (esse passo é opcional, pois os arquivos foram baixados do site oficial e não de um site alternativo):

*gpg –import KEYS*

*gpg –verify hadoop-3.3.0.tar.gz.asc hadoop-3.3.0.tar.gz*

* 1. Descompacte o Hadoop com o comando *tar -xzvf hadoop-1.2.1.tar.gz*.
     1. Importante destacar que o arquivo foi baixado para a pasta */home/ubuntu* (pois não alteramos a pasta de trabalho). A instalação do Hadoop na imagem do IGTI está na pasta */usr/local*. É possível copiar, mas deve-se usar o comando SUDO.
     2. Na instalação que fiz não houve problema em usar o diretório do usuário. O único problema é que alguns caminhos precisarão ser alterados nos programas fornecidos pelo IGTI para que funcionem.
     3. Como o diretório de trabalho é */home/ubuntu*, após a descompactação os arquivos do Hadoop estarão em */home/ubuntu/hadoop-3.3.0*.
  2. Mova a pasta *hadoop-3.3.0* para a pasta *hadoop* com o comando *mv hadoop-3.3.0 hadoop*.
  3. Adicione as linhas abaixo no final do arquivo “.bashrc” com o comando *vi .bashrc*.

*export HADOOP\_CONF=/home/ubuntu/hadoop/etc/hadoop*

*export HADOOP\_HOME=/home/ubuntu/hadoop*

*#Set JAVA\_HOME*

*export JAVA\_HOME=/usr/lib/jvm/jdk1.8.0\_271*

*# Add Hadoop bin/ directory to path*

*export PATH=$PATH:$HADOOP\_HOME/bin*

*export PATH=$PATH:$HADOOP\_HOME/sbin*

*export HADOOP\_MAPRED\_HOME=${HADOOP\_HOME}*

*export HADOOP\_COMMON\_HOME=${HADOOP\_HOME}*

*export HADOOP\_HDFS\_HOME=${HADOOP\_HOME}*

*export YARN\_HOME=${HADOOP\_HOME}*

* + 1. O arquivo “.bashrc” encontra-se na pasta /home/ubuntu.
  1. Execute os seguintes comandos para atualizar as variáveis de ambiente e verificar se foram atualizadas:

*source ~/.bashrc*

*echo $HADOOP\_HOME*

*echo $HADOOP\_CONF*

* + 1. Os comandos “echo” devem exibir o seguinte, respectivamente: */home/ubuntu/hadoop* e */home/ubuntu/hadoop/etc/hadoop*.
  1. Adicione a variável de ambiente do Hadoop apontando para a instalação do Java com o comando *vi $HADOOP\_CONF/hadoop-env.sh*:

*export JAVA\_HOME=/usr/lib/jvm/jdk1.8.0\_271*

* + 1. O arquivo hadoop-env.sh possui diversas configurações e todas estão comentadas. Você pode descomentar a linha referente a essa variável ou adicionar uma nova linha no arquivo.
  1. Crie um diretório de trabalho (tmp) para o Hadoop com o comando *mkdir hdfstemp*.
  2. Adicione as linhas abaixo no arquivo core-site.xml com o comando *vi $HADOOP\_CONF/core-site.xml*.

*<configuration>*

*<property>*

*<name>fs.defaultFS</name>*

*<value>hdfs://nome\_dns\_do\_servidor:9000</value>*

*</property>*

*<property>*

*<name>hadoop.tmp.dir</name>*

*<value>/home/ubuntu/hdfstmp</value>*

*</property>*

*</configuration>*

* + 1. Não esqueça de substituir “nome\_dns\_do\_servidor” com o nome do servidor obtido no passo 1.
  1. Adicione as linhas abaixo no arquivo hdfs-site.xml com o comando *vi $HADOOP\_CONF/hdfs-site.xml*.

*<configuration>*

*<property>*

*<name>dfs.replication</name>*

*<value>1</value>*

*</property>*

*<property>*

*<name>dfs.permissions</name>*

*<value>true</value>*

*</property>*

*</configuration>*

* 1. Adicione as linhas abaixo no arquivo mapred-site.xml com o comando *vi $HADOOP\_CONF/mapred-site.xml*.

*<configuration>*

*<property>*

*<name>mapred.job.tracker</name>*

*<value> hdfs://nome\_dns\_do\_servidor:8021</value>*

*</property>*

*</configuration>*

* + 1. Não esqueça de substituir “nome\_dns\_do\_servidor” com o nome do servidor obtido no passo 1.
  1. Tente se conectar ao localhost sem uma senha com o comando *ssh localhost*. Caso não consiga, execute os comandos abaixo:

*ssh-keygen -t rsa -P '' -f ~/.ssh/id\_rsa*

*cat ~/.ssh/id\_rsa.pub >> ~/.ssh/authorized\_keys*

*chmod 0600 ~/.ssh/authorized\_keys*

1. Como a quantidade de RAM é insuficiente para rodar os programas no HADOOP, será necessário criar um arquivo SWAP. Na minha instalação só haviam 4 Gigabytes livres no disco, então eu criei um arquivo SWAP com 2 Gb. Eu segui as instruções do “modo mais rápido” em <https://www.digitalocean.com/community/tutorials/how-to-add-swap-on-ubuntu-14-04>. Abaixo os comandos que executei:

*sudo swapon -s*

*free -m*

*df -h*

*sudo fallocate -l 2G /swapfile*

*sudo chmod 600 /swapfile*

*sudo mkswap /swapfile*

*sudo swapon /swapfile*

*sudo swapon -s*

*free -m*

* 1. Para tornar o arquivo SWAP permanente deve-se editar o arquivo FSTAB com o comando *sudo nano /etc/fstab* e adicionar a seguinte linha (use um TAB entre cada informação da linha):

*/swapfile none swap sw 0 0*

1. Instale o ANT com os comandos:

*sudo apt update*

*sudo apt install ant*

1. Agora basta criar a mesma estrutura de diretórios dentro da pasta /home/ubuntu da mesma forma que foram exibidas nas aulas.
   1. Crie a pasta /home/ubuntu/hadoop/Dados usando o comando mkdir.
   2. Crie a pasta /home/ubuntu/hadoop/ExemploIGTI com o comando mkdir.
   3. Crie a pasta /home/ubuntu/hadoop/ExemploIGTI/src com o comando mkdir.
   4. Copie o arquivo covidData.txt para a pasta /home/ubuntu/hadoop/Dados.
   5. Copie o arquivo build\_ExemploIGTI.xml para a pasta /home/ubuntu/hadoop/ExemploIGTI.
   6. Copie o arquivo ExemploIGTI.java para a pasta /home/ubuntu/hadoop/ExemploIGTI/src.
      1. Os arquivos encontram-se no link <https://online.igti.com.br/courses/150/pages/arquivos-da-maquina-virtual-para-download?module_item_id=100222>
2. Altere o arquivo /home/ubuntu/hadoop/ExemploIGTI/build\_ExemploIGTI.xml: altere a versão do Hadoop e adicione as três linhas “pathelement” conforme modelo abaixo.

*<property name="hadoop.version" value="3.3.0"/>*

*<path id="build.classpath">*

*<pathelement location="${build.dir}"/>*

*<pathelement location="${hadoop.dir}/hadoop-core-${hadoop.version}.jar"/>*

*<!-- eu adicionei as 3 linhas abaixo -->*

*<pathelement location="${hadoop.dir}/share/hadoop/common/hadoop-common-3.3.0.jar"/>*

*<pathelement location="${hadoop.dir}/share/hadoop/mapreduce/hadoop-mapreduce-client-common-3.3.0.jar"/>*

*<pathelement location="${hadoop.dir}/share/hadoop/mapreduce/hadoop-mapreduce-client-core-3.3.0.jar"/>*

*</path>*

1. Agora é só executar os comandos apresentados em aula para inicializar o Hadoop, alterar o arquivo ExemploIGTI.java para realizar a tarefa, compilar e executar o arquivo. Partindo do diretório /home/ubuntu, executar:

*cd hadoop*

*bin/hdfs namenode -format*

*sbin/start-all.sh*

*bin/hdfs dfs -mkdir /user*

*bin/hdfs dfs -mkdir /user/ubuntu*

Espero que esse roteiro possa ajudar. Salvo algum esquecimento, esses foram os passos que segui para fazer o AWS rodar o programa da tarefa. Se der algum problema ou se tiverem dúvida podem me mandar uma mensagem pelo Canvas.

Boa sorte!