

Alunos:

João Pedro Moura Oliveira - 190030879

Matheus Calixto Vaz Pinheiro - 190055201

Instalação Hadoop

1 - Instalação Pré-requisitos

- Openssh
 - Ubuntu

```
sudo apt-get install ssh
sudo service ssh start
```
 - Arch

```
sudo pacman -S openssh
sudo systemctl enable sshd.service
sudo systemctl start sshd.service
```
- Java JDK 8
 - Ubuntu

```
sudo apt-get install openjdk-8-jdk
```
 - Arch

```
sudo pacman -S jdk8-openjdk
```
- Executável Hadoop

```
wget https://dlcdn.apache.org/hadoop/common/stable/hadoop-3.3.2.tar.gz
tar xzf hadoop-3.3.2.tar.gz
```
- Hadoop-streaming
Baixe no link: <https://jar-download.com/artifact-search/hadoop-streaming>
unzip jar_files.zip

2 - Configurar SSH

- Permitindo localhost sem senha

```
ssh-keygen -t rsa -P "" -f ~/.ssh/id_rsa
cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
chmod 0600 ~/.ssh/authorized_keys
```
- Testando conexão

```
ssh localhost
```

3 - Configura bashrc ou zshrc

- Adicionar a variável do ambiente

```
export HADOOP_HOME=local_salvo_do_executavel_do_hadoop
```
- Caso já tenha o PATH setado

```
export PATH=$PATH:$HADOOP_HOME/sbin:$HADOOP_HOME/bin (bashrc)
export PATH="${PATH}:${HADOOP_HOME}/sbin:${HADOOP_HOME}/bin"
(zshrc)
```
- Aplicar alterações

```
source ~/.bashrc ou source ~/.zshrc
```

4 - Configurações Hadoop

4.1 - **hadoop-env.sh**

- Descomentar a linha onde se encontra o export do binário JAVA_HOME
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk

4.2 - **core-site.xml**

- Adicionar a seguinte configuração para o hdfs

```
<configuration>
<property>
  <name>hadoop.tmp.dir</name>
  <value>pasta_home_para_o_hdfs</value>
</property>
<property>
  <name>fs.default.name</name>
  <value>hdfs://127.0.0.1:9000</value>
</property>
</configuration>
```

4.3 - **hdfs-site.xml**

- Adicionar as seguintes configurações para os namenodes e datanodes

```
<configuration>
<property>
  <name>dfs.data.dir</name>
  <value>diretório_do_namenode</value>
</property>
<property>
  <name>dfs.data.dir</name>
  <value>diretório_do_datanode</value>
</property>
<property>
  <name>dfs.replication</name>
  <value>quantidade_de_replicação</value>
</property>
</configuration>
```

4.4 - **mapred-site.xml**

- Adicionar configuração para o paradigma MapReduce

```
<configuration>
<property>
  <name>mapreduce.framework.name</name>
  <value>yarn</value>
```

```
</property>
</configuration>
```

4.5 - yarn-site.xml

- Adicionar configuração para o YARN, com as configurações do NodeManager, ResourceManager, Containers e ApplicationManager

```
<configuration>
<property>
  <name>yarn.nodemanager.aux-services</name>
  <value>mapreduce_shuffle</value>
</property>
<property>
<name>yarn.nodemanager.aux-services.mapreduce.shuffle.class</name>
  <value>org.apache.hadoop.mapred.ShuffleHandler</value>
</property>
<property>
  <name>yarn.resourcemanager.hostname</name>
  <value>127.0.0.1</value>
</property>
<property>
  <name>yarn.acl.enable</name>
  <value>0</value>
</property>
<property>
  <name>yarn.nodemanager.env-whitelist</name>
  <value>JAVA_HOME,HADOOP_HOME,HADOOP_HDFS_HOME,HADOOP_CONF_DIR,CLASSPATH_PERPEND_DISTCACHE,HADOOP_YARN_HOME,HADOOP_MAPRED_HOME</value>
</property>
</configuration>
```

5 - Formatação no Namenode HDFS

- Formate o diretório HDFS do namenode
hdfs namenode -format

6 - Inicie os nodes ou clusters

- Inicialize o NameNode e DataNode com o seguinte script
start-dfs.sh
- Inicialize os recursos YARN e o NodeManager
start-yarn.sh
- Cheque se todos os processos estão rodando
jps

7 - Crie um arquivo texto para contagem de palavras

- Para esse exemplo foi utilizado um Lorem Ipsum

8 - Transfira o arquivo para para o diretório hdfs

- `hdfs dfs -copyFromLocal arquivo.txt /`

9 - Execute o streaming com o input enviado para o hdfs, com o mapper feito em python e o reduce também em python, também disponibilizados juntos com esse tutorial

```
hadoop jar local_do_streaming.jar \  
-input input_texto_no_hdfs.txt \  
-mapper local_do_mapper.py \  
-reducer local_do_reducer.py
```

10 - Para checar a saída, deve-se utilizar o comando cat no hdfs

```
hdfs dfs -cat pasta_hdfs/pasta_de_output/part-XXXX
```