UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

Departamento Acadêmico de Informática

Big Data

Lab 1

Hadoop – Instalação em modo Single Node

Professor:	Leandro Batista de Almeida
Data:	24 de junho de 2025
Número de páginas:	7

Recursos:

- Referências
  - Hadoop - site:
    - `https://hadoop.apache.org/`
  - Hadoop - documentation site:
    - `https://hadoop.apache.org/docs/current/`
  - Hadoop - documentation site:
    - `https://hadoop.apache.org/docs/stable/hadoop-project-dist/hadoop-common/SingleCluster.html`

Ambiente da Atividade de Laboratório:

Esta atividade usará uma máquina virtual pré-configurada para VirtualBox, já disponibilizado para os alunos. A máquina virtual usa um sistema operacional Linux (distribuição LUbuntu), com um JDK Java instalado, bem como o servidor ssh também instalado.
-Usuário e senha
Usuário Linux: *bigdata*, senha: *hadoop*

Conteúdo:

1 Instalação Hadoop Single Node
2 Preparar o Ambiente
3 Preparando arquivos Hadoop
4 Configurando .bashrc
5 Configurando arquivos do Hadoop
6 Preparar o sistema de arquivos para o namenode e datanode
7 Iniciar o cluster e testar

1 Instalação Hadoop Single Node

Nesta atividade de laboratório, você vai instalar um servidor Hadoop single-node. Nesse tipo de instalação, todos os serviços do hadoop serão executados em um único computador. Esta instalação é indicada para testes e desenvolvimento, mas não para um ambiente de produção.

Os procedimentos de instalação tem os seguintes passos:

1. Instalar um ambiente Java (JVM e SDK)
2. Instalar SSH
3. Gerar chave própria
4. Preparar arquivos da distribuição Hadoop
5. Editar arquivo .bashrc file (variáveis de ambiente)
6. Atualizar arquivos de configuração do Hadoop
7. Criar diretórios para namenode e datanode
8. Formatar namenode
9. Iniciar deamons hadoop e testar

2 Preparar o ambiente

Ambiente Java

Você precisa instalar uma JVM e compilar para poder executar o Hadoop. Qualquer distribuição pode ser escolhida (OpenJDK, Oracle, IBM), mas você precisa se certificar que um compilador está também disponível.

A máquina virtual deste lab já possui um Java Development Kit instalado.

SSH

Hadoop usa ssh para comunicação inter-deamon (e inter-node), então cada nó deve ter o servidor e o cliente ssh instalados. Além disso, alguma forma de autenticação ssh é necessária. Você pode usar a geração e distribuição de chaves, ou qualquer outro método.

A máquina virtual para este lab já possui ssh instalado.

- Gerando chave ssh:

```
$ ssh-keygen -t rsa -P ""
$ cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
```

- Testando login ssh sem senha (depois disso feito, é necessário sair do shell remoto com *exit*):

```
$ ssh localhost
```

3 Preparando arquivos Hadoop

Você deve baixar os arquivos de instalação do Hadoop do site oficial (hadoop.apache.org).

Os arquivos devem ser extraídos do pacote e a estrutura de diretórios deve ser movida para /usr/local/hadoop. Antes de mover o conteúdo, o diretório deverá ser criado. A criação do diretório e a movimentação deverá ser feita com autorização de root (sudo).

O restante deste lab assume que todos os arquivos do hadoop estão no diretório /usr/local/hadoop

4 Configurando .bashrc

Você precisa editar o arquivo .bashrc, no home directory do usuário, adicionando as seguintes linhas:

```
#variaveis Hadoop
export JAVA_HOME=”/usr/lib/jvm/java-8-openjdk-amd64”
export HADOOP_HOME=/usr/local/hadoop
export PATH=$PATH:$HADOOP_HOME/bin
export PATH=$PATH:$HADOOP_HOME/sbin
export HADOOP_COMMON_LIB_NATIVE_DIR=$HADOOP_HOME/lib/native
export HADOOP_OPTS="-Djava.library.path=$HADOOP_HOME/lib"
unset JAVA_TOOL_OPTIONS
export PDSH_RCMD_TYPE=ssh
```

Após alterar o arquivo .bashrc, para ativar as alterações, é necessário se logar novamente ou executar `source ~./bashrc`.

5 Configurando arquivos do hadoop:

No diretório $HADOOP_HOME/etc/hadoop, altere os seguintes arquivos segundo as instruções.

hadoop-env.sh file (ao final do arquivo):

```
export JAVA_HOME="/usr/lib/jvm/java-8-openjdk-amd64"
export HADOOP_OPTS=-Djava.net.preferIPv4Stack=true
```

core-site.xml file (dentro da tag configuration):

```
<property>
   <name>fs.defaultFS</name>
   <value>hdfs://localhost:9000</value>
</property>
```

mapred-site.xml (dentro da tag configuration):

```
<property>
   <name>mapreduce.framework.name</name>
   <value>yarn</value>
</property>
<property>
   <name>mapreduce.application.classpath</name>
   <value>$HADOOP_HOME/share/hadoop/mapreduce/*:$HADOOP_HOME/share/hadoop/
mapreduce/lib/*</value>
</property>
```

hdfs-site.xml (dentro da tag configuration):

```
<property>
   <name>dfs.replication</name>
   <value>1</value>
</property>
<property>
   <name>dfs.namenode.name.dir</name>
   <value>file:/usr/local/hadoop/hadoop_data/hdfs/namenode</value>
</property>
<property>
   <name>dfs.datanode.data.dir</name>
   <value>file:/usr/local/hadoop/hadoop_data/hdfs/datanode</value>
</property>
```

yarn-site.xml (dentro da tag 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.nodemanager.env-whitelist</name>
<value>JAVA_HOME,HADOOP_COMMON_HOME,HADOOP_HDFS_HOME,HADOOP_CONF_DIR,CLAS
SPATH_PREPEND_DISTCACHE,HADOOP_YARN_HOME,HADOOP_HOME,PATH,LANG,TZ,HADOOP_
MAPRED_HOME</value>
</property>
```

6 Preparar o sistema de arquivos para o namenode e datanode

Criar os diretórios para namenode e datanode:

```
$ mkdir -p /usr/local/hadoop/hadoop_data/hdfs/namenode
$ mkdir -p /usr/local/hadoop/hadoop_data/hdfs/datanode
$ sudo chown bigdata -R /usr/local/hadoop
```

Formatar sistema de arquivos HDFS:

```
$ hdfs namenode -format
```

7 Iniciar o cluster e testar

Iniciar os deamons (HDFS e YARN):

```
$ start-dfs.sh
$ start-yarn.sh
```

O HistoryServer é iniciado por:

```
$ mapred historyserver
```

É preciso ter certeza de que todos os serviços estão executando. Para fazer isso, use a ferramenta jps, e verifique os resultados. Você deve ver 6 tarefas executando.

Acesse as ferramentas web de gerenciamento:

- HDFS admin
  - http://localhost:9870
- YARN admin
  - http://localhost:8088
- History Server
  - http://localhost:19888

Testar a operação do cluster:

Usar TestDFSIO para verificar se o cluster está funcionando adequadamente. Você vai encontrar o arquivo jar no diretório $HADOOP_HOME/share/hadoop/mapreduce . O número de versão do arquivo jar é igual ao da versão do hadoop (3.3.6, neste exemplo).

```
$ hadoop jar hadoop-mapreduce-client-jobclient-3.3.6-tests.jar TestDFSIO
-write -nrFiles 5 -fileSize 10
```

Leia a documentação e localize o job WordCount. Baixe arquivos em formato de texto puro (txt) de alguma fonte (do Project Gutenbert, por exemplo) e use o WordCount para contar as palavras dos arquivos no Hadoop.
Maiores informações na documentação do Hadoop.

Lembre-se de desligar os serviços antes de desligar a VM. Isso pode ser feito com os scripts contrários aos da inicialização (stop-yarn.sh e stop-dfs.sh).