



Engenharia de Computação



Especialização Lato Sensu em Ciência de Dados e Analytics

Aula Prática

{ HDFS }

Prof. Jairson Rodrigues
jairson.rodrigues@univasf.edu.br

{ roteiro }

- Interface Web
- Acesso remoto ao Namenode
- Principais comandos HDFS Shell
- Manipulação de arquivos
- Gerenciamento de réplicas
- Gerenciamento de tamanhos de bloco

{ nota }

Esta aula prática será concentrada na distribuição padrão **Apache Hadoop 2.7 com acesso SSH via Namenode**. Mas o aluno deve ter em mente que há muitas outras formas de acesso, por exemplo: WebHDFS, Ambari ou opções proprietárias. Há produtos de mercado, tais como Cloudera, MapR e Hortonworks Sandbox, dentre outros, que fornecem clusters em VirtualBox/VMWARE/KVM ou até mesmo interfaces para clusters reais em provedores de computação em nuvem. Cada solução possui sua interface e forma de acesso próprios.

{ namenode - configuração e acesso remoto }

1. `vagrant up`
2. `vagrant ssh node-1`
3. `sh admin/start-all.sh # inicializa hadoop/yarn/spark`
4. `sh admin/stop-all.sh # para hadoop/yarn/spark`

IMPORTANTE: os comandos 3 e 4 são scripts bash personalizados para a aula e não fazem parte da distribuição Hadoop ou Spark.

{ interface web }

<http://192.168.100.101:50070>

The screenshot shows the Hadoop DFS Health web interface. The browser address bar displays the URL `192.168.100.101:50070/dfshealth.html#tab-overview`. The interface has a green header with the title "Hadoop" and a navigation menu with tabs: "Overview" (selected), "Datanodes", "Datanode Volume Failures", "Snapshot", "Startup Progress", and "Utilities".

Overview 'hadoop-master:8020' (active)

| | |
|----------------|--|
| Started: | Wed Sep 04 20:22:32 UTC 2019 |
| Version: | 2.7.4, rcd915e1e8d9d0131462a0b7301586c175728a282 |
| Compiled: | 2017-08-01T00:29Z by kshvachk from branch-2.7.4 |
| Cluster ID: | CID-bf769198-6df7-48b8-9bab-637220f7652f |
| Block Pool ID: | BP-1190021583-10.0.2.15-1567628485237 |

Summary

Security is off.

Safemode is off.

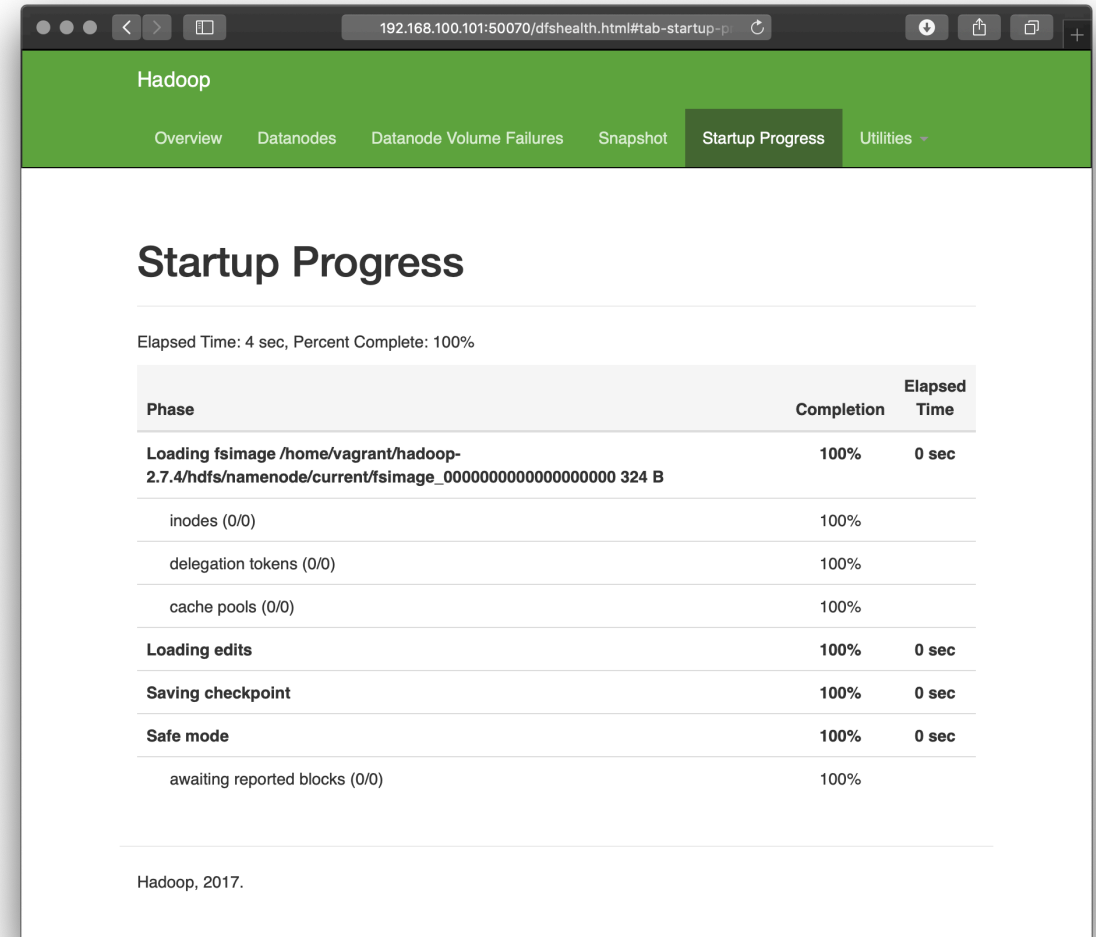
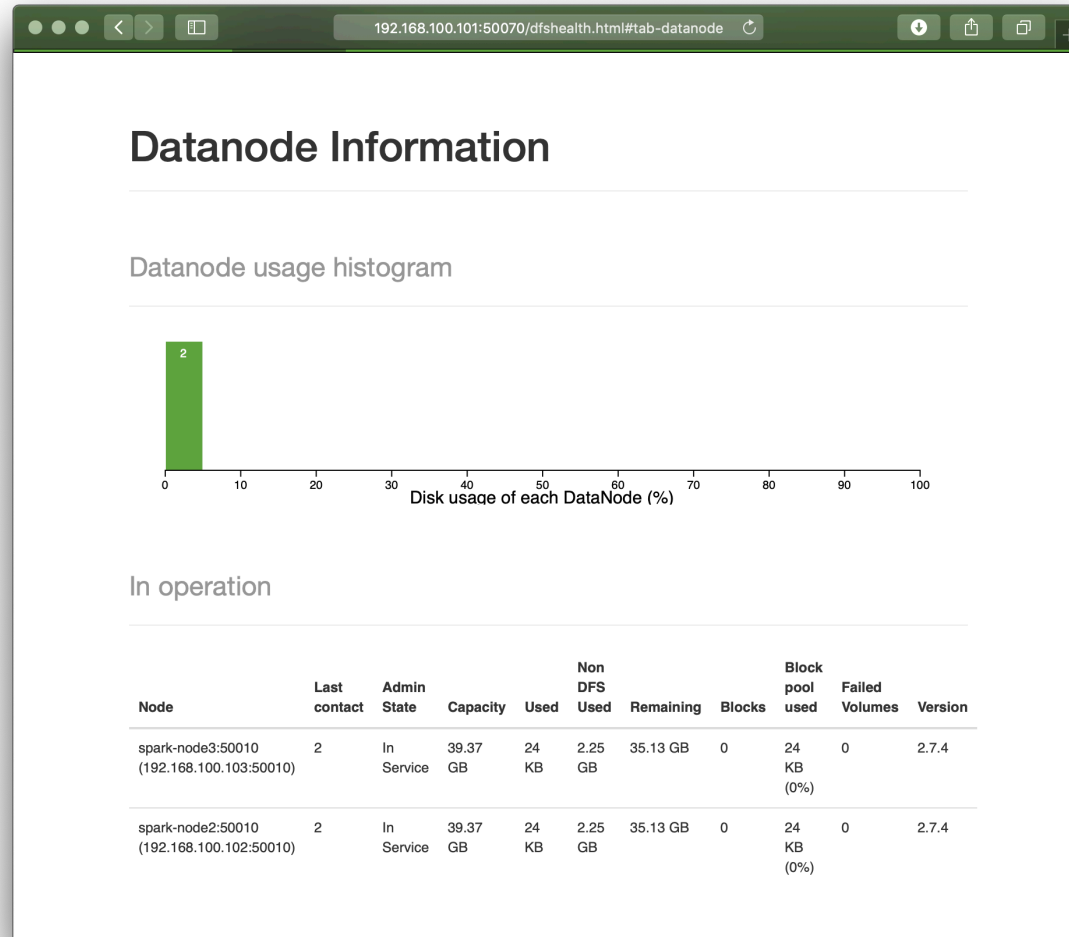
1 files and directories, 0 blocks = 1 total filesystem object(s).

Heap Memory used 32.47 MB of 48.79 MB Heap Memory. Max Heap Memory is 966.69 MB.

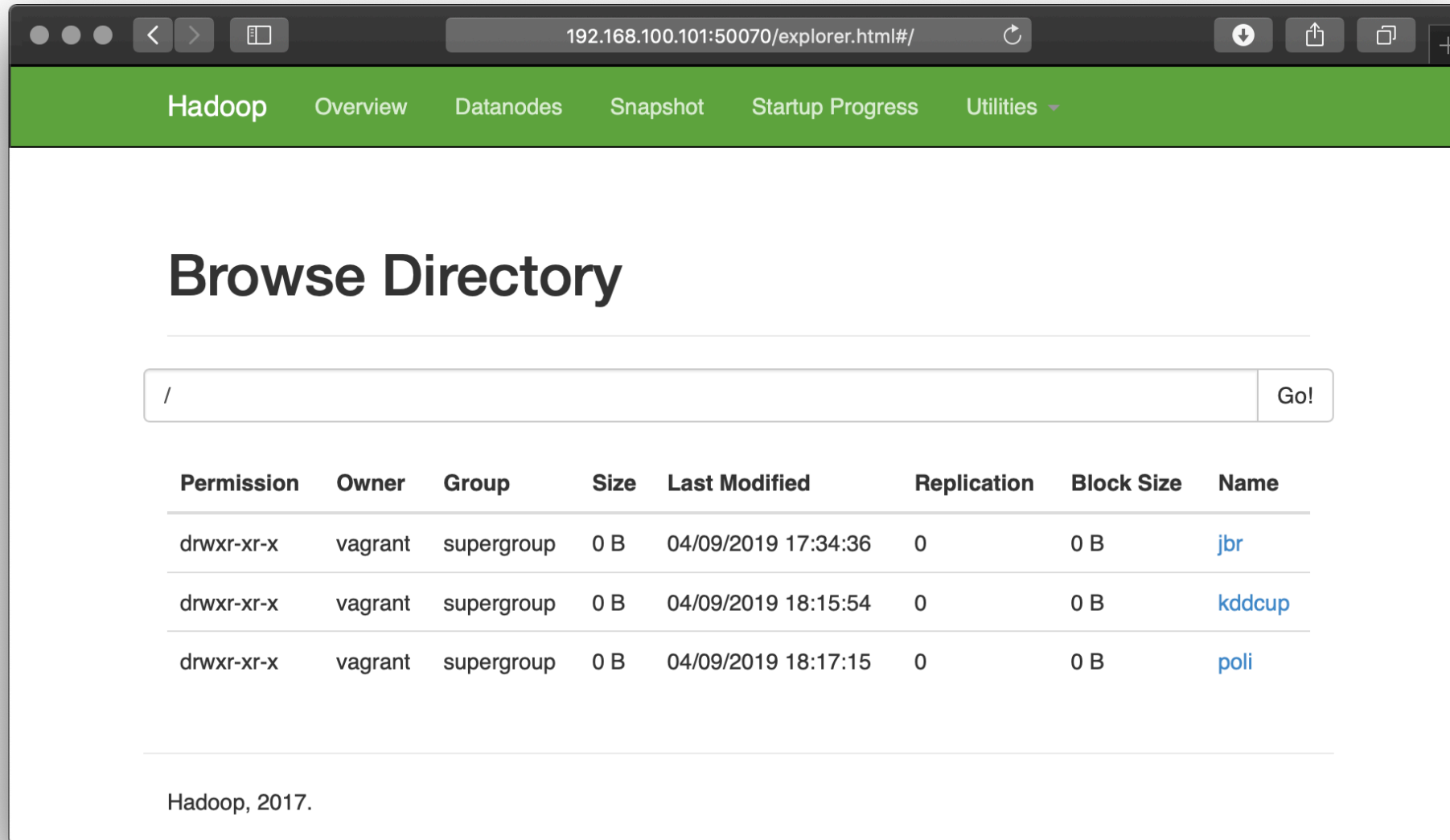
Non Heap Memory used 35.96 MB of 36.69 MB Committed Non Heap Memory. Max Non Heap Memory is -1 B.

| | |
|----------------------|------------|
| Configured Capacity: | 78.74 GB |
| DFS Used: | 48 KB (0%) |

{ interface web }



{ interface web }



192.168.100.101:50070/explorer.html#/

Hadoop Overview Datanodes Snapshot Startup Progress Utilities

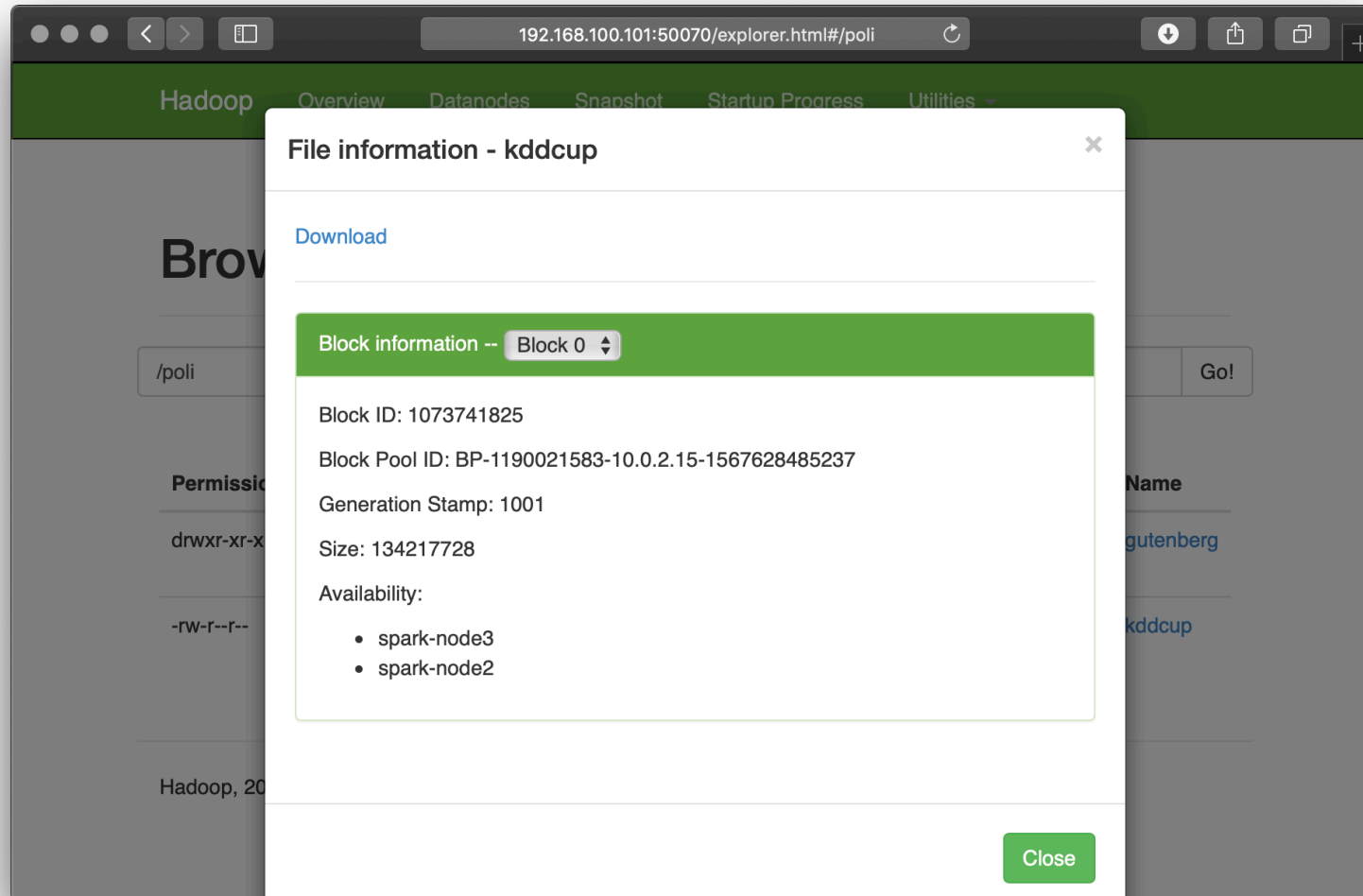
Browse Directory

/ Go!

| Permission | Owner | Group | Size | Last Modified | Replication | Block Size | Name |
|------------|---------|------------|------|---------------------|-------------|------------|------------------------|
| drwxr-xr-x | vagrant | supergroup | 0 B | 04/09/2019 17:34:36 | 0 | 0 B | jbr |
| drwxr-xr-x | vagrant | supergroup | 0 B | 04/09/2019 18:15:54 | 0 | 0 B | kddcup |
| drwxr-xr-x | vagrant | supergroup | 0 B | 04/09/2019 18:17:15 | 0 | 0 B | poli |

Hadoop, 2017.

{ hadoop web interface }



{ vamos praticar? }

- Crie uma pasta com suas iniciais
- `mkdir <suas_iniciais>`
- `cd <suas_iniciais>`

{ **hadoop fs** }

- Interage com diversos sistemas de arquivo
 - HDFS, HFTP, S3 FS etc
- **hdfs** **dfs** => sinônimo de **hadoop fs** para armazenamento HDFS
- Nesta aula, serão utilizados sem distinção

hadoop fs -rm /poli/input/big.txt

hdfs dfs -rm /poli/input/big.txt

hdfs dfs -rm hdfs:///poli/input/big.txt

hdfs dfs -rm hdfs://namenode/poli/input/big.txt

...

*** { uma nota sobre bug do hadoop 2.7 }

- No curso utilizamos a versão 2.7, que pode gerar o erro abaixo:

WARN hdfs.DFSClient: Caught exception

java.lang.InterruptedException

at java.lang.Object.wait(Native Method)

at java.lang.Thread.join(Thread.java:1252)

at java.lang.Thread.join(Thread.java:1326)

at org.apache.hadoop.hdfs.DFSOutputStream\$DataStreamer.closeResponder(DFSOutputStream.java:716)

at org.apache.hadoop.hdfs.DFSOutputStream\$DataStreamer.closeInternal(DFSOutputStream.java:684)

at org.apache.hadoop.hdfs.DFSOutputStream\$DataStreamer.run(DFSOutputStream.java:680)

- No escopo desta aula prática o aluno de ignorar o erro. Maiores detalhes podem ser encontrados da descrição do bug em <https://issues.apache.org/jira/browse/HDFS-10429>

{ criando pastas }

mkdir

```
hadoop fs -mkdir [-p] <paths>
```

-d: cria pastas pai ao longo do caminho

```
$ hdfs dfs -mkdir -p /<sua_pasta>/data/
```

{ copiando dados no cluster }

cp

```
hadoop fs -cp [-f] URI [URI ...] <dest>
```

-f: sobrescreve o destino, se existir

```
$ hdfs dfs -cp /poli/ /<sua_pasta>
```

{ listagem de arquivos }

ls

```
hadoop fs -ls [-d] [-h] [-R] <args>
```

-d: apenas o diretório

-h: formata o tamanho do arquivo

-R: busca recursiva

```
$ hdfs dfs -ls -h -R /<sua_pasta>/
```

{ enviando arquivos para o cluster }

copyFromLocal

```
hadoop fs -copyFromLocal <localsrc> URI
```

```
$ hdfs dfs -copyFromLocal gutenber/*.*txt /<sua_pasta>/data
```

```
$ hdfs dfs -ls -h -R /<sua_pasta>/data
```

{ recuperando arquivos do cluster }

copyToLocal

```
hadoop fs -copyToLocal <localsrc> URI
```

```
$ hdfs dfs -copyToLocal /poli/kddcup/kddcup.data
```

CUIDADO: a máquina destino pode não ter capacidade para receber o **BIG** arquivo.

{ espaço livre no cluster }

df

```
hadoop fs -df -h
```

```
$ hdfs dfs -df -h
```

{ modificando o tamanho padrão do bloco }

put

```
hadoop fs -D dfs.block.size=xxx -put <localsrc> URI
```

```
$ hdfs dfs -D dfs.block.size=4194304 -put  
kddcup.data /<sua_pasta>/poli/kddcup/kddcup4M.data
```

{ modificando o número de réplicas }

setrep

```
hadoop dfs -setrep -w -R xxx URI
```

w - prompt espera até o final da operação

```
$ hdfs dfs -setrep -w 2 /<sua_pasta>/poli/kddcup/kddcup.data
```

Obs: a diminuição de réplicas é um processo mais moroso que a adição.

{ referência completa }

- Shell
 - <http://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-common/FileSystemShell.html>
- Outros Comandos
 - <http://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-hdfs/HDFSCommands.html>
- Guia do Usuário HDFS
 - <http://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-hdfs/HdfsUserGuide.html>

