Hadoop Distributed File System - HDFS

Eduardo Mendes
github.com/z4r4tu5tr4

z4r4tu5tr4@babbge: screenfetch

Nome: Eduardo Mendes

Instituição: Fatec Americana

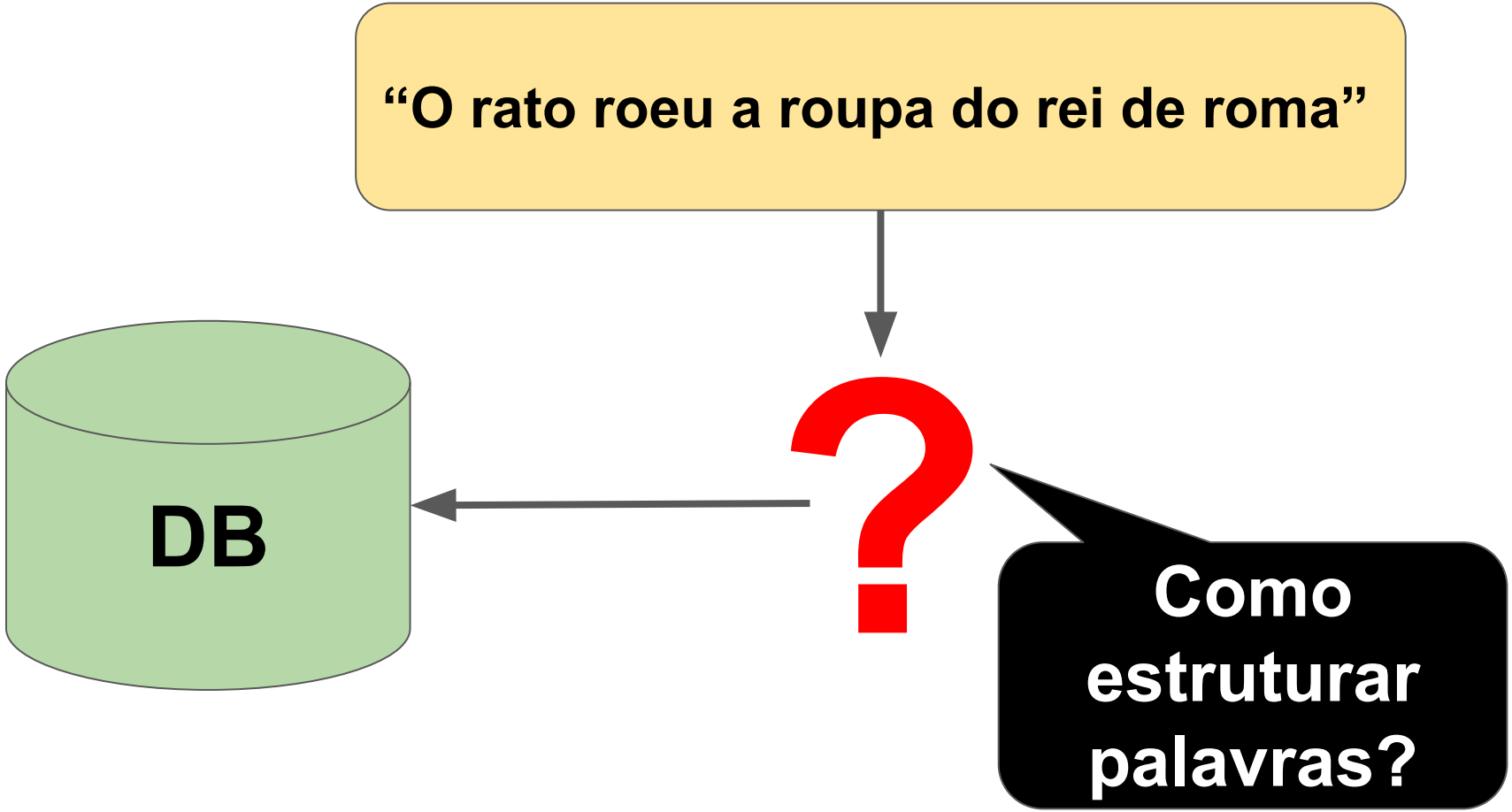
Uptime: 12097080s

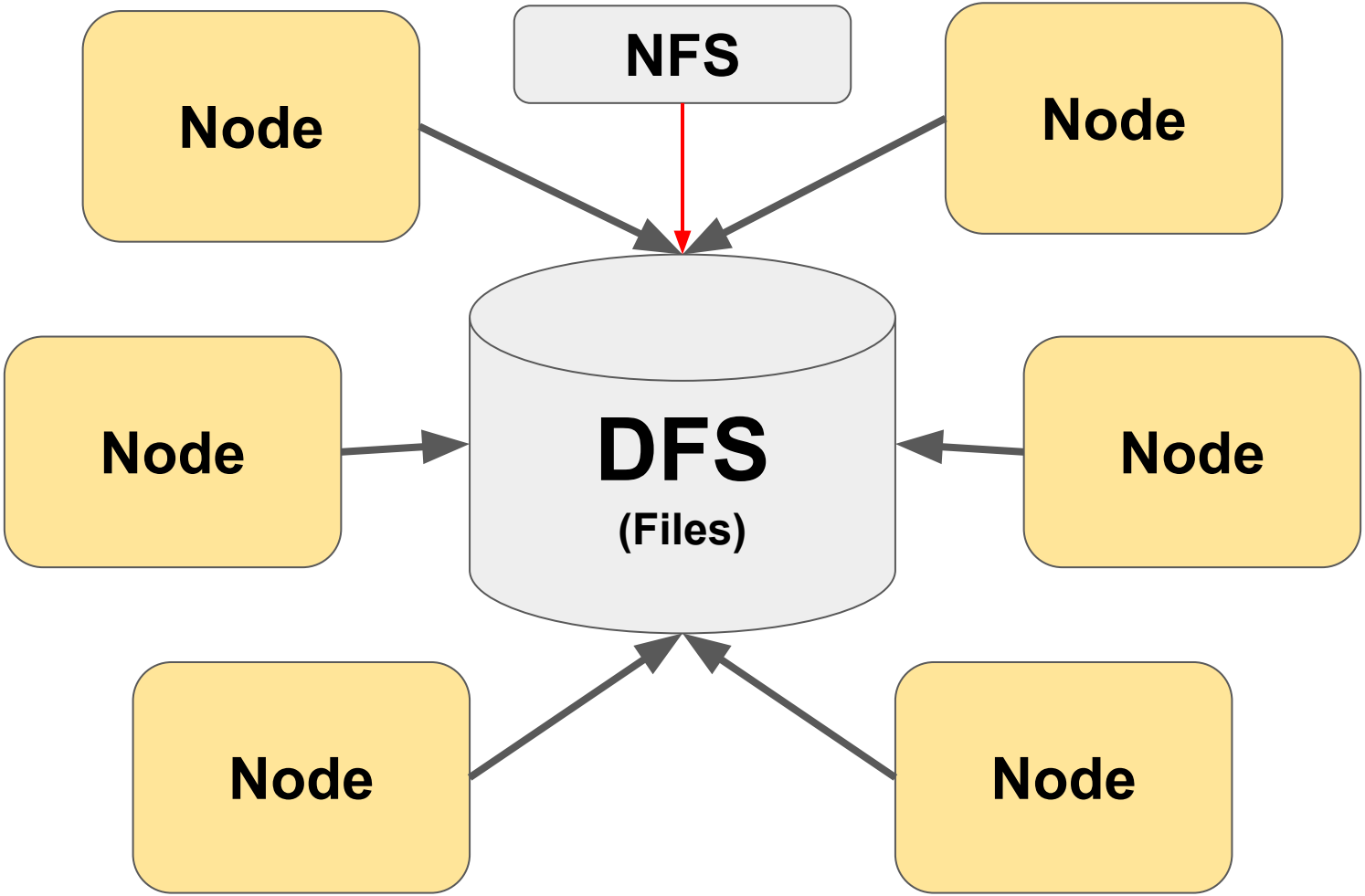
Email: mendexeduardo@gmail.com

git: github.com/z4r4tu5tr4

Por que precisamos de um **DFS**?

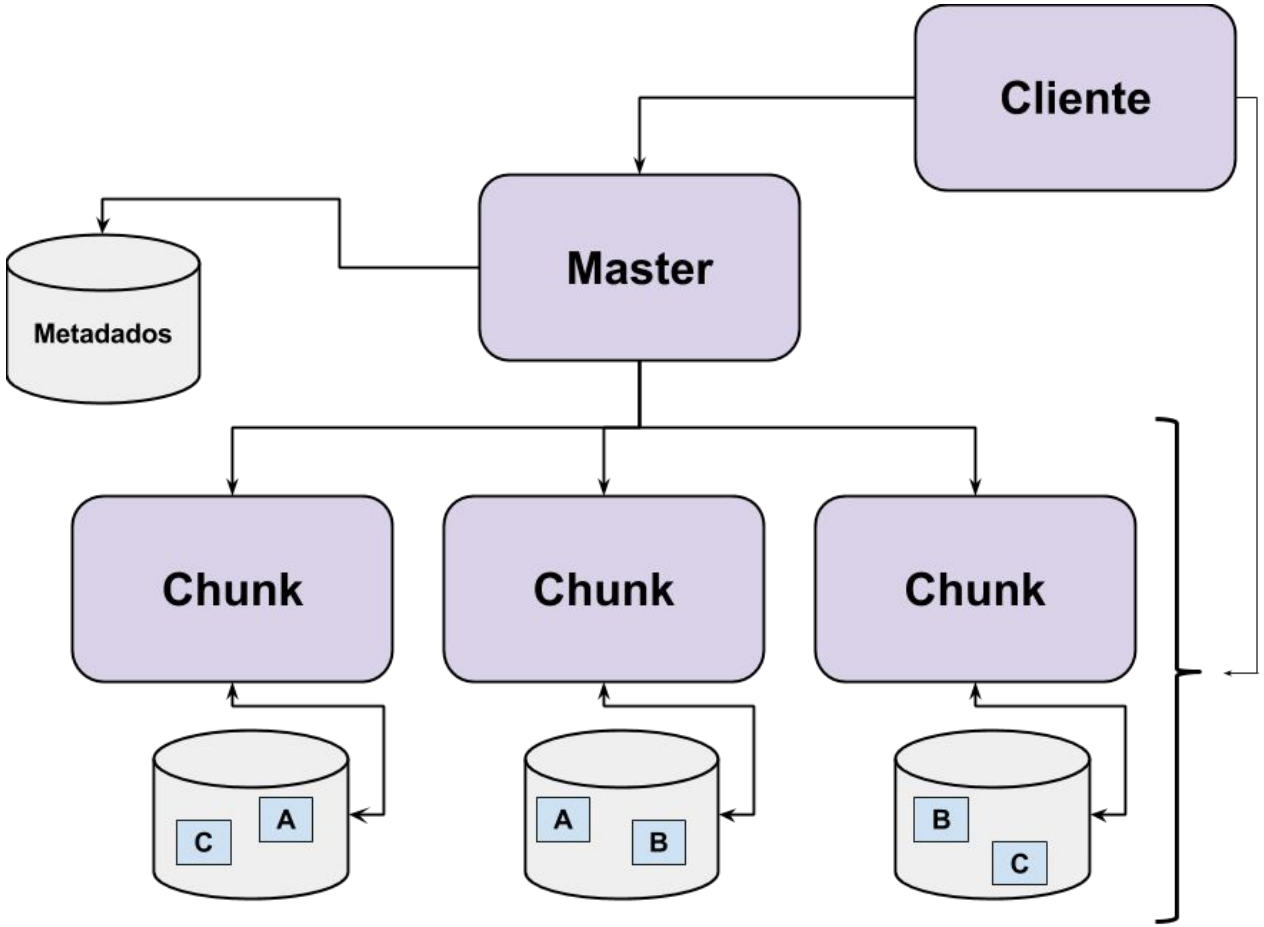
Ou quando precisamos


“O rato roeu a roupa do rei de roma”
DB
?
Como estruturar palavras?


NFS
Node
Node
Node
DFS
(Files)
Node
Node
Node
Node

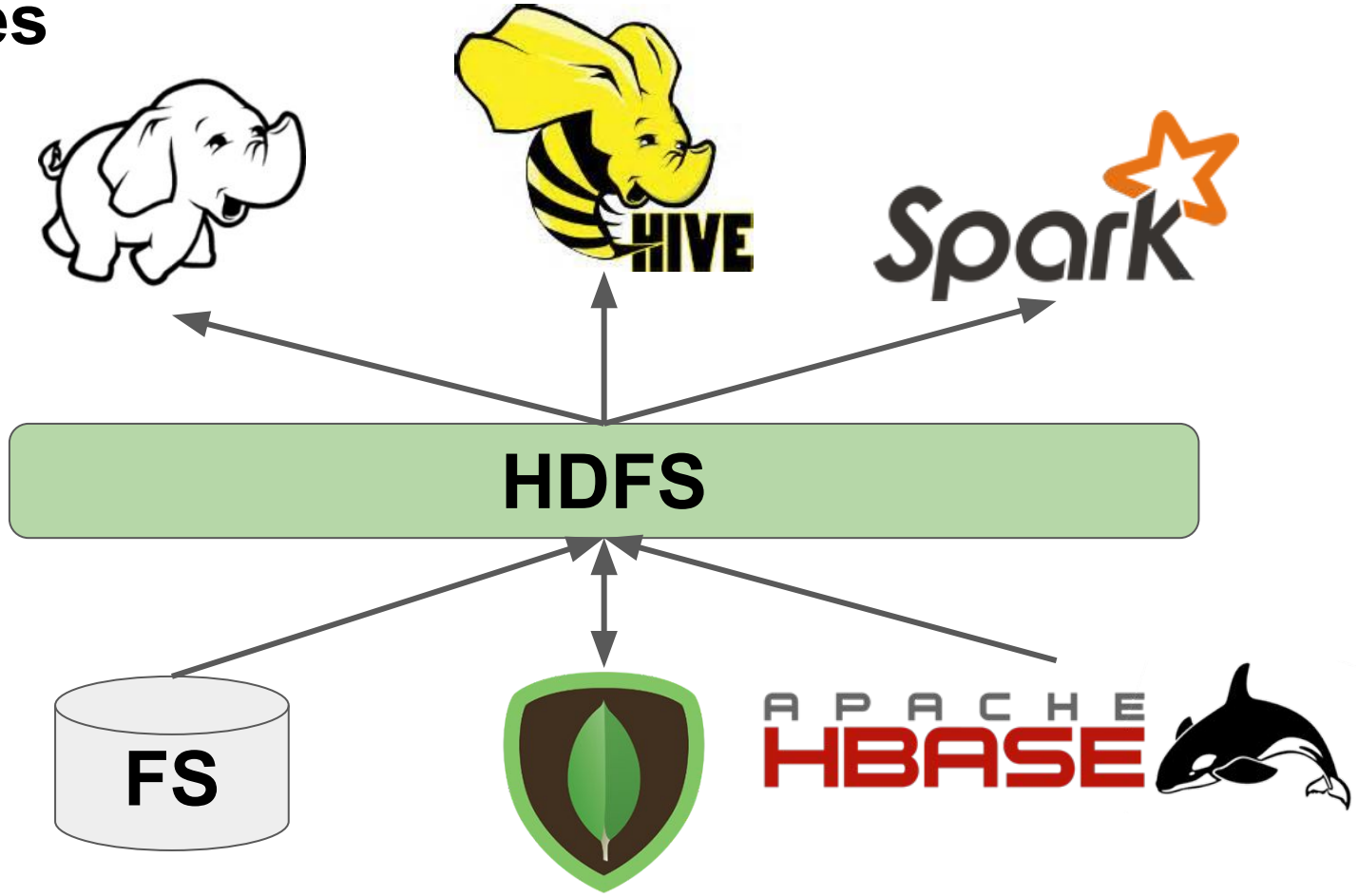
HDFS

- Diretórios
- “POSIX”
- Balanceamento
- HASH



Comando	FS
POSIX	
ls	Lista diretório
cd	Muda de diretório
mkdir	Cria diretório
Hadoop	
put	Insere no HDFS
get	Transfere para p FS

Clientes



MapReduceLib

github.com/z4r4tu5tr4/mapreducelib

MapReduceLib

- Uma maneira de interagir com o terminal e vendo suas respostas (Diferente da API disponibilizada pelo HDFS)
- Código simples para qualquer modificação sem muito esforço
- Só uma ponte entre o bash e o terminal interativo

```
In [1]: from mapreducelib import Hadoop

In [2]: node = Hadoop()

In [3]: node.dfs_start()
```

```
Starting namenodes on [localhost]
localhost: starting namenode, logging to /usr/local/hadoop/logs/hadoop-z4r4tu5t
4-namenode-Babbage.out
localhost: starting datanode, logging to /usr/local/hadoop/logs/hadoop-z4r4tu5t
4-datanode-Babbage.out
Starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to /usr/local/hadoop/logs/hadoop-z
r4tu5tr4-secondarynamenode-Babbage.out
```

```
In [4]: from mapreducelib import Hdfs

In [5]: dfs = Hdfs()

In [6]: dfs.ls()
```

```
Found 1 items
-rw-r--r--   1 z4r4tu5tr4 supergroup          0 2016-04-02 00:39 /xxx.txt
```

```
In [8]: dfs.get("/xxx.txt")
16/04/04 23:18:28 WARN hdfs.DFSClient: DFSInputStream has been closed already
```

```
In [10]: system("ls")
Contador_de_palavras.py  noticias.txt  pyspark-note.sh  Untitled.ipynb
mapreducelib.py          __pycache__   Untitled1.ipynb  xxx.txt
Out[10]: 0
```

```
In [11]: dfs.put("noticias.txt", "/")
```

```
In [12]: dfs.ls()
Found 2 items
-rw-r--r--   1 z4r4tu5tr4 supergroup   32662350 2016-04-04 23:38 /noticias.txt
-rw-r--r--   1 z4r4tu5tr4 supergroup          0 2016-04-02 00:39 /xxx.txt
```

```
In [13]: dfs.mkdir("/lulamolusco")

In [14]: dfs.ls()
Found 3 items
drwxr-xr-x   - z4r4tu5tr4 supergroup          0 2016-04-04 23:39 /lulamolusco
-rw-r--r--   1 z4r4tu5tr4 supergroup   32662350 2016-04-04 23:38 /noticias.txt
-rw-r--r--   1 z4r4tu5tr4 supergroup          0 2016-04-02 00:39 /xxx.txt
```

```
In [16]: dfs.hdfs = '/usr/local/hadoop/bin/hadoop fs'
```

```
In [17]: dfs.rm_dir("/")
16/04/04 23:43:45 INFO fs.TrashPolicyDefault: Namenode trash configuration: Dele
tion interval = 0 minutes, Emptier interval = 0 minutes.
rm: `/': Input/output error
```

```
In [18]: dfs.rm_dir("/lulamolusco")
16/04/04 23:44:09 INFO fs.TrashPolicyDefault: Namenode trash configuration: Dele
tion interval = 0 minutes, Emptier interval = 0 minutes.
Deleted /lulamolusco
```

```
In [19]: dfs.ls()
Found 2 items
-rw-r--r--   1 z4r4tu5tr4 supergroup   32662350 2016-04-04 23:38 /noticias.txt
-rw-r--r--   1 z4r4tu5tr4 supergroup          0 2016-04-02 00:39 /xxx.txt
```

XOXO

Dúvidas?

mendesxeduardo@gmail.com