

Big Data

Hadoop

Prof. Jean Paul Barddal

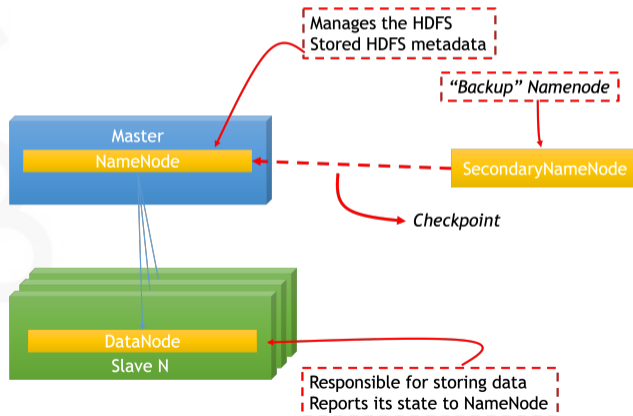


Agenda

- 1 Hadoop Distributed File System (HDFS)
- 2 Rack-Awareness

HDFS

■ Arquitetura master-slave



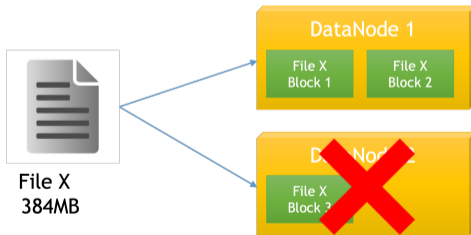
HDFS

- No HDFS, arquivos são divididos em blocos
- Um bloco é a unidade básica de leitura e escrita
- Tamanho default de 128MB, (costumava ser de 64 MB)
- Cada bloco pode ser replicado e distribuído em diferentes computadores
- Escalabilidade
- NameNode gerencia os blocos associados a cada arquivo
- Torna o HDFS tolerante a falhas



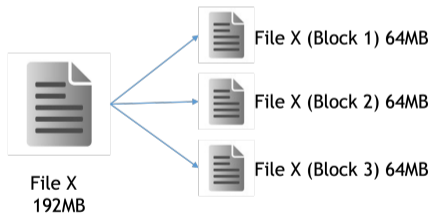
HDFS

- HDFS divide arquivos em blocos que são distribuídos em DataNodes
- E se um DataNode falhar?
- Dado que os blocos são distribuídos e um dos blocos é perdido, perdemos o arquivo todo?



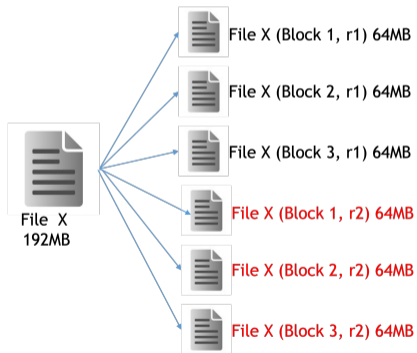
Fator de Replicação

- HDFS usa replicação para garantir tolerância a falhas
- Fator de replicação = 1

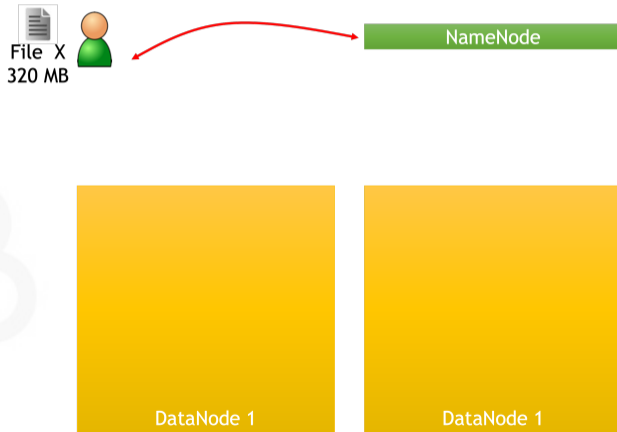


Fator de Replicação

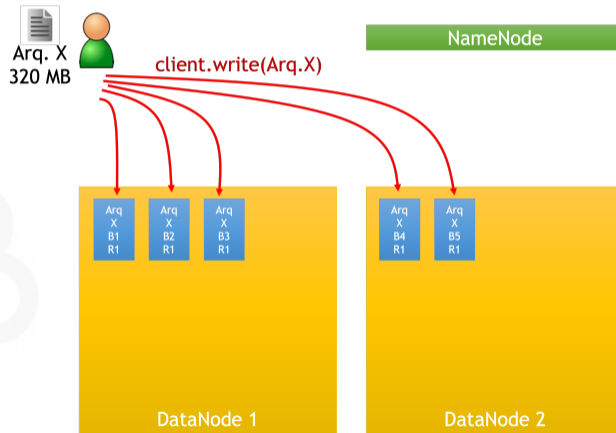
- Fator de replicação = 2



Salvando e Lendo Dados do HDFS



Salvando e Lendo Dados do HDFS

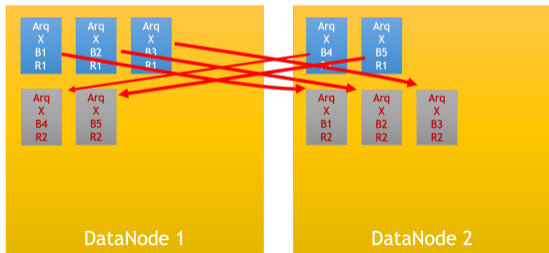


Salvando e Lendo Dados do HDFS

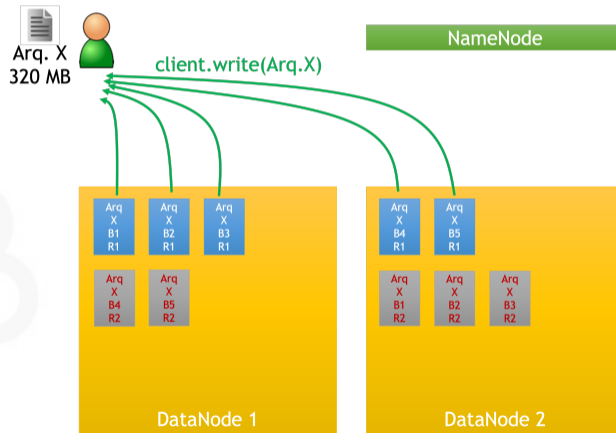


`client.write(Arq.X)`

NameNode



Salvando e Lendo Dados do HDFS



Salvando e Lendo Dados do HDFS



NameNode

Arq X B1 R1	Arq X B2 R1	Arq X B3 R1
----------------------	----------------------	----------------------

Arq X B4 R2	Arq X B5 R2
----------------------	----------------------

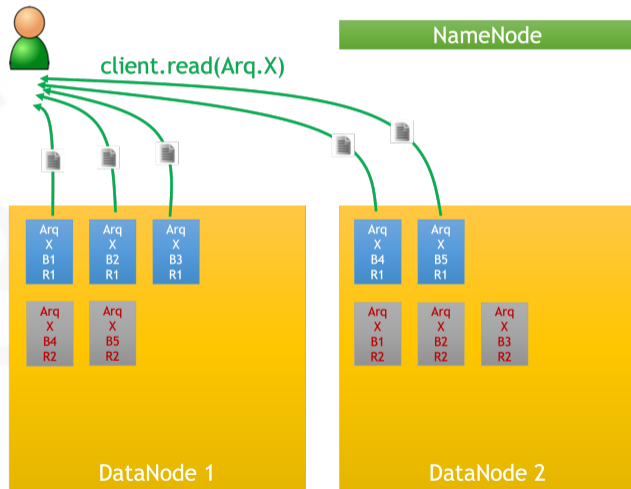
DataNode 1

Arq X B4 R1	Arq X B5 R1
----------------------	----------------------

Arq X B1 R2	Arq X B2 R2	Arq X B3 R2
----------------------	----------------------	----------------------

DataNode 2

Salvando e Lendo Dados do HDFS



Atividade prática

- Vamos realizar testes com o HDFS e compreender os principais comandos

Primeiros passos

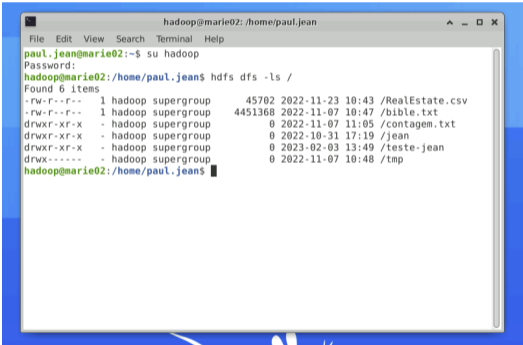
- Acesse a VPN e a máquina virtual que dispõe de Hadoop
- Este passo a passo está disposto em outro conjunto de slides

Comandos do HDFS

- Por ser um sistema de arquivos, os principais utilitários disponíveis em sistemas operacionais estão disponíveis também no HDFS
- Principais comandos:
 - `dfs -mkdir`: criação de diretório
 - `dfs -touchz`: criação de arquivo vazio
 - `dfs -copyFromLocal`: cópia de arquivo local para o HDFS
 - `dfs -copyToLocal` (ou `-get`): cópia de arquivo do HDFS para o host
 - `dfs -cat`: print do arquivo desejado
 - `dfs -cp`: cópia de arquivos dentro do HDFS
 - `dfs -mv`: movimentação de arquivos dentro do HDFS
 - `dfs -rmr`: apaga arquivos de forma recursiva dentro do HDFS
 - `dfs -du`: fornece o tamanho de cada arquivo em um diretório
 - `dfs -dus`: fornece o tamanho total de um arquivo/diretório
 - `dfs -getmerge`: copia o arquivo e o retorna ao host concatenado

Exemplo - Acessando o HDFS

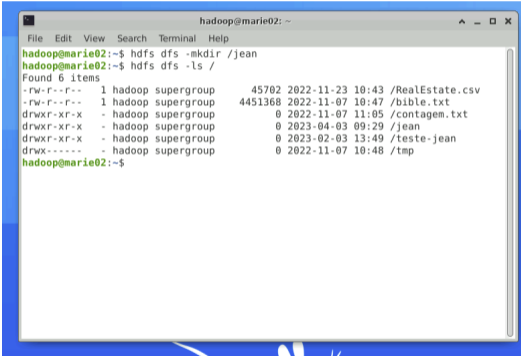
- Na VM, usar o terminal e usar o comando **su hadoop** (senha: hadoop)
- Para verificar os diretórios, usar **hdfs dfs -ls /**



```
hadoop@marie02: /home/paul.jean
File Edit View Search Terminal Help
paul.jean@marie02:~$ su hadoop
Password:
hadoop@marie02:/home/paul.jean$ hdfs dfs -ls /
Found 6 items
-rw-r--r-- 1 hadoop supergroup 45702 2022-11-23 10:43 /RealEstate.csv
-rw-r--r-- 1 hadoop supergroup 4451368 2022-11-07 10:47 /bible.txt
drwxr-xr-x - hadoop supergroup 0 2022-11-07 11:05 /contagem.txt
drwxr-xr-x - hadoop supergroup 0 2022-10-31 17:19 /jean
drwxr-xr-x - hadoop supergroup 0 2023-02-03 13:49 /teste-jean
drwx----- - hadoop supergroup 0 2022-11-07 10:48 /tmp
hadoop@marie02:/home/paul.jean$
```

Exemplo - Criando diretório

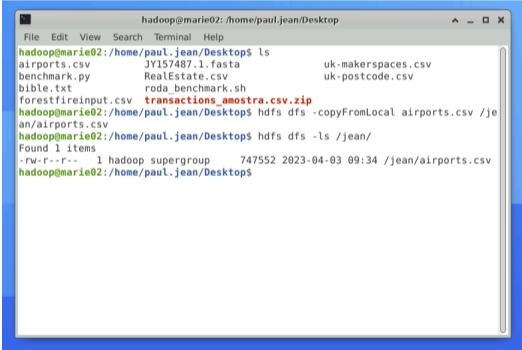
- Para criar um diretório, usamos **hdfs dfs -mkdir /nome-do-diretorio**



```
hadoop@marie02: ~  
File Edit View Search Terminal Help  
hadoop@marie02:~$ hdfs dfs -mkdir /jean  
hadoop@marie02:~$ hdfs dfs -ls /  
Found 6 items  
-rw-r--r-- 1 hadoop supergroup      45702 2022-11-23 10:43 /RealEstate.csv  
-rw-r--r-- 1 hadoop supergroup  4451368 2022-11-07 10:47 /bible.txt  
drwxr-xr-x - hadoop supergroup      0 2022-11-07 11:05 /contagem.txt  
drwxr-xr-x - hadoop supergroup      0 2023-04-03 09:29 /jean  
drwxr-xr-x - hadoop supergroup      0 2023-02-03 13:49 /teste-jean  
drwx----- - hadoop supergroup      0 2022-11-07 10:48 /tmp  
hadoop@marie02:~$
```

Exemplo - Copiando arquivo para o HDFS

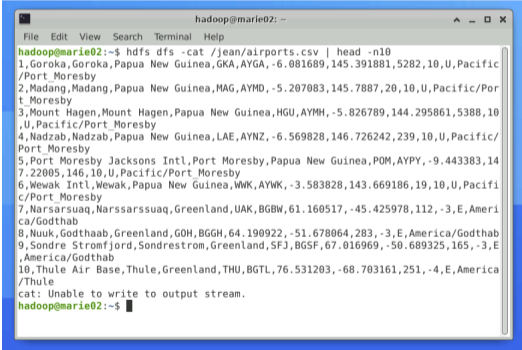
- Para copiar um arquivo local para o HDFS, usamos o comando **hdfs dfs -copyFromLocal**



```
hadoop@marie02: /home/paul.jean/Desktop
File Edit View Search Terminal Help
hadoop@marie02:/home/paul.jean/Desktop$ ls
airports.csv      JY157487.1.fasta      uk-makerspaces.csv
benchmark.py     RealEstate.csv        uk-postcode.csv
bible.txt        roda_benchmark.sh
forestfireinput.csv transactions_amostra.csv.zip
hadoop@marie02:/home/paul.jean/Desktop$ hdfs dfs -copyFromLocal airports.csv /jean/airports.csv
hadoop@marie02:/home/paul.jean/Desktop$ hdfs dfs -ls /jean/
Found 1 items
-rw-r--r--  1 hadoop supergroup  747552 2023-04-03 09:34 /jean/airports.csv
hadoop@marie02:/home/paul.jean/Desktop$
```

Exemplo - Visualizando o arquivo no HDFS

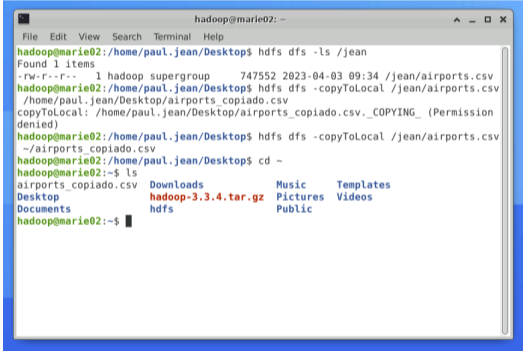
- Para verificar o conteúdo de um arquivo, podemos usar o comando **hdfs dfs -cat**



```
hadoop@marie02: ~  
File Edit View Search Terminal Help  
hadoop@marie02:~$ hdfs dfs -cat /jean/airports.csv | head -n10  
1,Goroka,Goroka,Papua New Guinea,GKA,AYGA,-6.081689,145.391881,5282,10,U,Pacific  
/Port_Moresby  
2,Madang,Madang,Papua New Guinea,MAG,AYMD,-5.207083,145.7887,20,10,U,Pacific/Port  
_Moresby  
3,Mount Hagen,Mount Hagen,Papua New Guinea,HGU,AYMH,-5.826789,144.295861,5388,10  
,U,Pacific/Port_Moresby  
4,Nadzab,Nadzab,Papua New Guinea,LAE,AYNZ,-6.569828,146.726242,239,10,U,Pacific/  
Port_Moresby  
5,Port Moresby Jacksons Intl,Port Moresby,Papua New Guinea,POM,AYPY,-9.443383,14  
7.22005,146,10,U,Pacific/Port_Moresby  
6,Wewak Intl,Wewak,Papua New Guinea,MWK,AYWK,-3.583828,143.669186,19,10,U,Pacifi  
c/Port_Moresby  
7,Narsarsuaq,Narsarsuaq,Greenland,UAK,BGBW,61.160517,-45.425978,112,-3,E,Ameri  
ca/Godthab  
8,Nuuk,Godthaab,Greenland,GOH,BGGH,64.190922,-51.678064,283,-3,E,America/Godthab  
9,Sondre Stromfjord,Sondrestrom,Greenland,SFJ,BGSF,67.016969,-50.689325,165,-3,E  
,America/Godthab  
10,Thule Air Base,Thule,Greenland,THU,BGTL,76.531203,-68.703161,251,-4,E,America  
/Thule  
cat: Unable to write to output stream.  
hadoop@marie02:~$
```

Exemplo - Copiando arquivo do HDFS

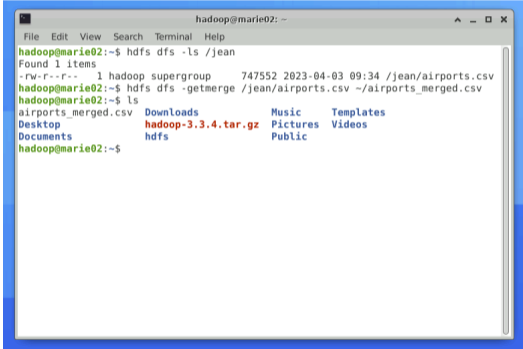
- Para copiar um arquivo do HDFS para o computador local, usamos o comando **hdfs dfs -copyToLocal**



```
hadoop@marie02: ~  
File Edit View Search Terminal Help  
hadoop@marie02:/home/paul.jean/Desktop$ hdfs dfs -ls /jean  
Found 1 items  
-rw-r--r-- 1 hadoop supergroup 747552 2023-04-03 09:34 /jean/airports.csv  
hadoop@marie02:/home/paul.jean/Desktop$ hdfs dfs -copyToLocal /jean/airports.csv  
/home/paul.jean/Desktop/airports_copiado.csv  
copyToLocal: /home/paul.jean/Desktop/airports_copiado.csv._COPYING_ (Permission  
denied)  
hadoop@marie02:/home/paul.jean/Desktop$ hdfs dfs -copyToLocal /jean/airports.csv  
~/airports_copiado.csv  
hadoop@marie02:/home/paul.jean/Desktop$ cd ~  
hadoop@marie02:~$ ls  
airports_copiado.csv Downloads Music Templates  
Desktop hadoop-3.3.4.tar.gz Pictures Videos  
Documents hdfs Public  
hadoop@marie02:~$
```

Exemplo - Copiando arquivo do HDFS (merge)

- Análogo ao comando anterior, mas os blocos (partes) de um arquivo são concatenados usando **hdfs dfs -getmerge**

A terminal window titled 'hadoop@marie02: ~' with a menu bar (File, Edit, View, Search, Terminal, Help). The terminal shows the following commands and output:

```
hadoop@marie02:~$ hdfs dfs -ls /jean
Found 1 items
-rw-r--r-- 1 hadoop supergroup 747552 2023-04-03 09:34 /jean/airports.csv
hadoop@marie02:~$ hdfs dfs -getmerge /jean/airports.csv ~/airports_merged.csv
hadoop@marie02:~$ ls
airports_merged.csv  Downloads  Music  Templates
Desktop              hadoop-3.3.4.tar.gz  Pictures  Videos
Documents            hdfs       Public
```


Agenda

1 Hadoop Distributed File System (HDFS)

2 Rack-Awareness

Rack-awareness

- Tenha em mente que Hadoop, HDFS e MapReduce foram projetados para computação distribuída e de larga escala
- Isso significa que Hadoop pode ser usado em clusters de diferentes tamanhos

Como os clusters deveriam ser



Como muitos clusters são



Rack-awareness

- Em cenários onde o cluster possui muitos computadores, o Hadoop define a “localização” e “proximidade” de cada computador em relação aos demais
- Replicação garante que os blocos de um mesmo arquivo estejam em racks diferentes (nenhum rack mantém mais de duas réplicas de um mesmo bloco)
- Neste caso, o acesso às réplicas se dá entre os Datanodes mais próximos de onde o processamento ocorrerá

