



Semantix

Hadoop - Básico

Aula 3

Quem sou eu?

Eu sou Rodrigo Augusto Rebouças.

Engenheiro de dados da Semantix
Instrutor do Semantix Mentoring Academy

Você pode me encontrar em:
rodrigo.augusto@semantix.com.br





Comandos HDFS

Sistema HDFS e diretórios

Comandos Sistema HDFS

- Similar ao Linux, mas inicia com “hadoop fs” ou “hdfs dfs (usado atualmente)”
- `hadoop fs -<comando> [argumentos]`
 - Diferentes sistemas
 - HDFS, Local FS, WebHDFS, S3 FS e outros
- `hdfs dfs -<comando>[argumentos]`
 - Sistema HDFS
- Ex. comando help
 - `$ hadoop fs -help`
 - `$ hdfs dfs -help`
 - `$ hdfs dfs -help ls`

Comandos Diretórios

- Criar diretório
 - mkdir <diretório>
 - Criar estrutura de diretórios
 - mkdir -p <diretório>/<diretório>/<diretório>
- Listar diretório
 - ls <diretório>
 - Recursivo: -R
- Remoção arquivos e diretórios
 - rm <src>
 - Argumentos
 - -r: Deletar diretório
 - -skipTrash: Remover permanentemente

Exemplo Diretórios Cloudera CDH 5.13

- Listar diretórios HDFS

```
Applications Places System
Browse and run installed applications
File Edit View Search Terminal Help
[cloudera@quickstart ~]$ hdfs dfs -ls /
Found 6 items
drwxrwxrwx - hdfs supergroup 0 2017-10-23 09:15 /benchmarks
drwxr-xr-x - hbase supergroup 0 2019-02-08 02:28 /hbase
drwxr-xr-x - solr solr 0 2017-10-23 09:18 /solr
drwxrwxrwt - hdfs supergroup 0 2019-02-05 11:03 /tmp
drwxr-xr-x - hdfs supergroup 0 2019-02-08 02:37 /user
drwxr-xr-x - hdfs supergroup 0 2017-10-23 09:17 /var
[cloudera@quickstart ~]$ hdfs dfs -ls /user/
Found 8 items
drwxr-xr-x - cloudera cloudera 0 2019-02-05 11:57 /user/cloudera
drwxr-xr-x - mapred hadoop 0 2017-10-23 09:15 /user/history
drwxrwxrwx - hive supergroup 0 2017-10-23 09:17 /user/hive
drwxrwxrwx - hue supergroup 0 2017-10-23 09:16 /user/hue
drwxrwxrwx - jenkins supergroup 0 2017-10-23 09:15 /user/jenkins
drwxrwxrwx - oozie supergroup 0 2017-10-23 09:16 /user/oozie
drwxrwxrwx - root supergroup 0 2017-10-23 09:16 /user/root
drwxr-xr-x - hdfs supergroup 0 2017-10-23 09:17 /user/spark
```



Comandos HDFS

Manipulação de dados

Comandos Enviar dados

- Local /HDFS
- Enviar arquivo ou diretório
 - put <src> <dst> (mais usado)
 - Argumentos
 - f: Sobrescreve o destino, se já existir.
 - l: Força um fator de replicação de
 - copyFromLocal <src> <dst>
- Mover arquivo ou diretório
 - Put que deleta do local
 - moveFromLocal <src> <dst>

Comandos Receber dados

- HDFS/Local
- Receber arquivo ou diretório
 - `get <src> <dst>`
 - Argumento - f
 - Cria um único arquivo mesclado
 - `getmerge <src> <dst>`
- Mover para o local
 - Get que deleta a cópia do HDFS
 - `moveToLocal <src> <localdst>`

Comandos copiar e mover dados

- HDFS/HDFS
- Copiar arquivo ou diretório
 - `cp <src> <dst>`
 - Argumento `-f`
- Mover arquivo ou diretório
 - `mv <src> <dst>`
 - `mv <arquivo1> <arquivo2> <arquivo3> <dst>`
- Não é permitido copiar ou mover arquivos entre sistemas de arquivos



Comandos HDFS

Informações de arquivos e sistema

Comandos Informações

- Mostrar o uso do disco
 - `du -h <diretório>`
- Exibir o espaço livre
 - `df -h <diretório>`
- Mostrar as informações do diretório
 - `stat <diretório>`
 - `hdfs dfs -stat %r name.txt #fator de replicação`
 - `hdfs dfs -stat %o name.txt #tamanho do bloco`
- Contar o número de diretórios, número de arquivos e tamanho do arquivo especificado
 - `count -h <diretório>`
- Esvaziar a lixeira
 - `expunge`

Comandos Arquivos

- Ver conteúdo do arquivo
 - `cat <arquivo>`
- Alterar o fator de replicação do arquivo
 - `setrep <nº repetições> <arquivo>`
- Criar um arquivo de registro com data e hora
 - `touchz <diretório>`
- Retornar as informações da soma de verificação de um arquivo
 - `checksum <arquivo>`
- Mostra o último 1KB do arquivo no console:
 - `tail [-f] <arquivo>`
 - `hdfs dfs -tail name.txt`
 - `hdfs dfs -cat name.txt | head -n 1`

Comandos Localização

- Localiza todos os arquivos que correspondem à expressão

- `find <caminho> <procura> -print`

- Exemplos:

- `hdfs dfs -find / -name data`

- `hdfs dfs -find / -iname Data -print`

- `/user/semantix/input/data`

- `hdfs dfs -find input/ -name *.txt -print`

- `input/teste1.txt, input/teste2.txt`

Exercícios Comandos HDFS

1. Iniciar o cluster de Big Data
 - `cd docker-bigdata`
 - `docker-compose up -d`
2. Baixar os dados dos exercícios do treinamento
 - `cd input`
 - `sudo git clone https://github.com/rodrigo-reboucas/exercises-data.git`
3. Acessar o container do namenode
4. Criar a estrutura de pastas apresentada ao lado pelo comando: `$ hdfs dfs -ls -R /`
5. Enviar a pasta `"/input/exercises-data/escola"` e o arquivo `"/input/exercises-data/entrada1.txt"` para `data`
6. Mover o arquivo `"entrada1.txt"` para `recover`
7. Baixar o arquivo do hdfs `"escola/alunos.json"` para o sistema local /
8. Deletar a pasta `recover`
9. Deletar permanentemente o `delete`
10. Procurar o arquivo `"alunos.csv"` dentro do `/user`
11. Mostrar o último 1KB do arquivo `"alunos.csv"`
12. Mostrar as 2 primeiras linhas do arquivo `"alunos.csv"`
13. Verificação de soma das informações do arquivo `"alunos.csv"`
14. Criar um arquivo em branco com o nome de `"test"` no `data`
15. Alterar o fator de replicação do arquivo `"test"` para 2

```
hdfs dfs -ls -R /
```

```
1. user/aluno
```

```
i. <nome>
```

```
1. data
```

```
2. recover
```

```
3. delete
```

16. Ver as informações do arquivo `"alunos.csv"`

17. Exibir o espaço livre do `data` e o uso do disco



Semantix

Obrigado!

Alguma pergunta?



Você pode me encontrar em:
rodrigo.augusto@semantix.com.br

GET SMARTER