T2.3: Wordcount no CESGA. Introducción a HDFS

Big Data Aplicado

11/12/23 – IES Fernando Wirtz

David Fernández Reboredo

Índice

Wordcount no CESGA. Introducción a HDFS

Creación de un documento lista_compra

Entramos no CESGA no hadoop en hadoop.cesga.es

```
(base) PS C:\Users\david.fernandezrebor> ssh xuwira02@hadoop.cesga.es
Enter passphrase for key 'C:\Users\david.fernandezrebor/.ssh/id_rsa':
Last login: Thu Dec  1 20:42:45 2022 from 10.121.0.2
*******************************************************************************
*                       BIENVENIDO A LA PLATAFORMA HADOOP3                    *
*                                                                             *
*  Financiada por el Ministerio de Economia y Competitividad,                 *
*  la Xunta de Galicia y FEDER (Fondo Europeo de Desarrollo Regional)         *
*******************************************************************************
*      En caso de problemas/dudas: telefono ---> 981 56 98 10, ext. 814       *
*                                  web      ---> https://bigdata.cesga.es     *
*                                  e-mail   ---> helpdesk_bigdata@cesga.es    *
*                                                                             *
*      Documentacion de la plataforma:                                        *
*        - Guia de usuario:  https://bigdata.cesga.es/user-guide/             *
*        - Tutoriales:  http://bigdata.cesga.es/#tutorials                    *
*                                                                             *
*******************************************************************************
```

Logo de iniciar sesión crearemos co editor nano unha lista da compra:

```
[xuwira02@cdh61-login6 ~]$ nano lista_compra.txt
```

Introduciremos os alimentos que conformarán a nosa lista da compra:

Premeremos agora sair con CRTL+X e seleccionaremos a opción de gardar os cambios do ficheiro de texto.

Creamos no HDFS un directorio chamado compras.

E logo diso introduciremos no hdfs no directorio compras a nosa lista da compra:

```
[xuwira02@cdh61-login6 ~]$ hdfs dfs -mkdir compras
[xuwira02@cdh61-login6 ~]$ hdfs dfs -put lista_compra.txt compras/
```

Agora cun LS poderemos observar os distintos permisos que presenta o directorio

compras xunto con algunha función relevante como cando foi creado:

```
[xuwira02@cdh61-login6 ~]$ hdfs dfs -ls compras
Found 1 items
-rw-r--r--   3 xuwira02 xunta        140 2023-12-11 20:36 compras/lista_compra.txt
```

Lanzamento dun traballo a yarn

Co comando yarn podemos lanzar un conxunto de archivos jar:

```
[xuwira02@cdh61-login6 ~]$ yarn jar /opt/cloudera/parcels/CDH-6.1.1-1.cdh6.1.1.p0.875
250/jars/hadoop-mapreduce-examples-3.0.0-cdh6.1.1.jar wordcount compras resumo_compra
s
```

O resultado gárdase en HDFS e a continuación traeremos o resultado no noso sistema de ficheiros:

```
[xuwira02@cdh61-login6 ~]$ hdfs dfs -get resumo_compras
```

Agora cun cat poderemos observar a lista traida en HDFS.

HDFS para Libros.

SCP para pasar o arquivo libros.zip a HADOOP

Para pasar o libros.zip a HADOOP debemos de empregar o comando SCP especificando o arquivo no directorio que estaba o arquivo libros.zip e a nosa dirección co usuario do Cesga(pediranos a contraseña de inicio de sesión):

```
(base) PS C:\Users\david.fernandezrebor\Downloads> scp libros.zip xuwira02@hadoop.ces
ga.es:~
Enter passphrase for key 'C:\Users\david.fernandezrebor/.ssh/id_rsa':
libros.zip                                    100% 2841KB   1.7MB/s   00:01
```

Agora facemos un unzip para descomprimir o arquivo:

```
[xuwira02@cdh61-login3 ~]$ unzip libros.zip
Archive:  libros.zip
   creating: libros/
  extracting: libros/contra_el_copyright.txt
  extracting: libros/copia_este_libro.txt
  extracting: libros/cuando_los_administradores_de_sistemas.txt
  extracting: libros/elcodigo20.txt
  extracting: libros/en_el_principio_fue_la_linea_de_comandos.txt
  extracting: libros/softwarelibre_sociedadlibre.txt
[xuwira02@cdh61-login3 ~]$ ls
libros  libros.zip  lista_compra.txt  notebook  resumo_compras
```

Agora con HDFS debemos crear un directorio e pasar libros ao directorio novo:

```
[xuwira02@cdh61-login3 ~]$ hdfs dfs -mkdir libros
[xuwira02@cdh61-login3 ~]$ hdfs dfs -put libros libros/
```

```
[xuwira02@cdh61-login3 ~]$ hdfs dfs -ls libros
Found 1 items
drwxr-xr-x   - xuwira02 xunta          0 2023-12-11 21:23 libros/libros
[xuwira02@cdh61-login3 ~]$ hdfs dfs -ls libros/libros
Found 6 items
-rw-r--r--   3 xuwira02 xunta     110046 2023-12-11 21:23 libros/libros/contra_el_copyright.txt
-rw-r--r--   3 xuwira02 xunta     405839 2023-12-11 21:23 libros/libros/copia_este_libro.txt
-rw-r--r--   3 xuwira02 xunta      90063 2023-12-11 21:23 libros/libros/cuando_los_administradores_de_sistemas.txt
-rw-r--r--   3 xuwira02 xunta    1402812 2023-12-11 21:23 libros/libros/elcodigo20.txt
-rw-r--r--   3 xuwira02 xunta     244499 2023-12-11 21:23 libros/libros/en_el_principio_fue_la_linea_de_comandos.txt
-rw-r--r--   3 xuwira02 xunta     654893 2023-12-11 21:23 libros/libros/softwarelibre_sociedadlibre.txt
```

Agora, executaremos o yarn jar sobre o directorio Libros/Libros

```
[xuwira02@cdh61-login3 ~]$ yarn jar /opt/cloudera/parcels/CDH-6.1.1-1.cdh6.1.1.p0.875250/jars/hadoop-mapreduce-examples-3.0.
0-cdh6.1.1.jar wordcount libros/libros contaje_libros
WARNING: YARN_OPTS has been replaced by HADOOP_OPTS. Using value of YARN_OPTS.
23/12/11 21:30:34 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /user/xuwira02/.staging/job_1678696
618277_9795
23/12/11 21:30:34 INFO input.FileInputFormat: Total input files to process : 6
23/12/11 21:30:34 INFO mapreduce.JobSubmitter: number of splits:6
23/12/11 21:30:34 INFO Configuration.deprecation: yarn.resourcemanager.zk-address is deprecated. Instead, use hadoop.zk.addr
ess
23/12/11 21:30:34 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Inste
ad, use yarn.system-metrics-publisher.enabled
23/12/11 21:30:34 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1678696618277_9795
23/12/11 21:30:34 INFO mapreduce.JobSubmitter: Executing with tokens: []
23/12/11 21:30:34 INFO conf.Configuration: resource-types.xml not found
23/12/11 21:30:34 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
23/12/11 21:30:34 INFO impl.YarnClientImpl: Submitted application application_1678696618277_9795
23/12/11 21:30:35 INFO mapreduce.Job: The url to track the job: https://c14-18.bd.cluster.cesga.es:8090/proxy/application_16
78696618277_9795/
23/12/11 21:30:35 INFO mapreduce.Job: Running job: job_1678696618277_9795
23/12/11 21:30:40 INFO mapreduce.Job: Job job_1678696618277_9795 running in uber mode : false
23/12/11 21:30:40 INFO mapreduce.Job:  map 0% reduce 0%
23/12/11 21:30:46 INFO mapreduce.Job:  map 100% reduce 0%
23/12/11 21:30:51 INFO mapreduce.Job:  map 100% reduce 2%
23/12/11 21:30:52 INFO mapreduce.Job:  map 100% reduce 6%
23/12/11 21:30:53 INFO mapreduce.Job:  map 100% reduce 96%
23/12/11 21:30:54 INFO mapreduce.Job:  map 100% reduce 100%
23/12/11 21:30:54 INFO mapreduce.Job: Job job_1678696618277_9795 completed successfully
23/12/11 21:30:54 INFO mapreduce.Job: Counters: 55
```

Agora traeremos o noso resultado ao sistema de ficheiros local:

```
[xuwira02@cdh61-login3 ~]$ hdfs dfs -get contaje_libros
```

Finalmente poderemos comprobar con un cat que o contaje se realizou correctamente

```
[xuwira02@cdh61-login3 ~]$ cat contaje_libros/*
```

```
mante-nerlas.   1
mantequilla.    1
material        64
materializa     1
matizaciones,   1
mayor í a       21
m á gico 1
mechero,        1
medidas 39
mejorado        2
mejorarlo       1
mejorarse):     1
mejoras...      1
memento 1
menores.        6
mercado:        2
mercados)       1
meridianamente  4
mesiánica.      1
metiéndolo      1
mezclaban       1
migró   1
mini-campaña    1
ministra        8
mismo   360
```

```
pisotear        2
plana?  1
planean 4
planeta 12
planteamiento.  2
plásticas       1
pod í as 2
popular 20
portador        2
posibilitaban   1
posici ó n.     1
precio? 1
precisan        1
prescindir.     1
presentarnos?   1
preservaron     1
presidente.     1
pretendo.       1
pre ff eren     1
principiantes-  1
probabilidad    9
```

```
sangrante       1
sanitaria;      1
secretario      5
secreto-        2
secuestrados    1
segregados.34   1
seguirá 13
seguridad»      1
seguro- 1
seguían 5
semi ó tica     1
sentarnos       1
servidor),      1
servidores;     1
sheriff.        1
sido:   1
silencio»,      1
similares       11
simulación,     1
sincronicen     1
singular        5
sino    431
```

Estes son algúns exemplos nos que poderiamos ver cantas veces se repiten as palabras.