

DC-02 MAP REDUCE

DC-02 TRABAJO PRÁCTICO 01 MAP REDUCE

Programa: Máster Executive en Big Data, Cloud & Analytics

Periodo académico: 2019-2020

Autor/es: CARLOS ALFONSEL JAÉN

OBJETIVO DEL TRABAJO PRÁCTICO:

- Recordar y fijar los conceptos vistos en las sesiones de clase sobre Hadoop HDFS y Map Reduce.
- Evaluar un 25% de la nota final del sub-módulo.

DESCRIPCIÓN:

- Debemos obtener un repositorio de términos para poder traducir a diferentes idiomas.
- Disponemos de unos diccionarios de inglés a diferentes idiomas.
- Cada fichero contiene términos y su traducción a un determinado idioma, separados por un tabulador.
- Para evitar complejidad: no nos importa si todos los términos figuran en todos los idiomas. Tampoco si un término tiene varias acepciones en un mismo idioma.
- Entregar en un PDF los pasos, comandos y los pantallazos del proceso realizado contra el clúster de Dataproc de Google Cloud.

SE PIDE:

1. **Descargar los ficheros de idiomas de la web.**
Usar el comando (`wget http://www.ilovelanguages.com/IDP/files/German.txt`, etc.)
2. **Subir a HDFS.**
Ya sabemos que Hadoop está especialmente diseñado para trabajar con ficheros muy grandes. Por ese motivo vamos a agregar los ficheros de forma que generaremos un fichero `dictionary.txt` con el contenido de cada uno de los ficheros individuales descargados (usar el comando `cat [nombre fichero] >> dictionary.txt`)
3. **Ejecutar los procesos Map Reduce que hay en los scripts de Python adjuntos.**
Se pueden sugerir mejoras al código actual.
4. **Mostrar el resultado del término "House".**

RESPUESTAS:

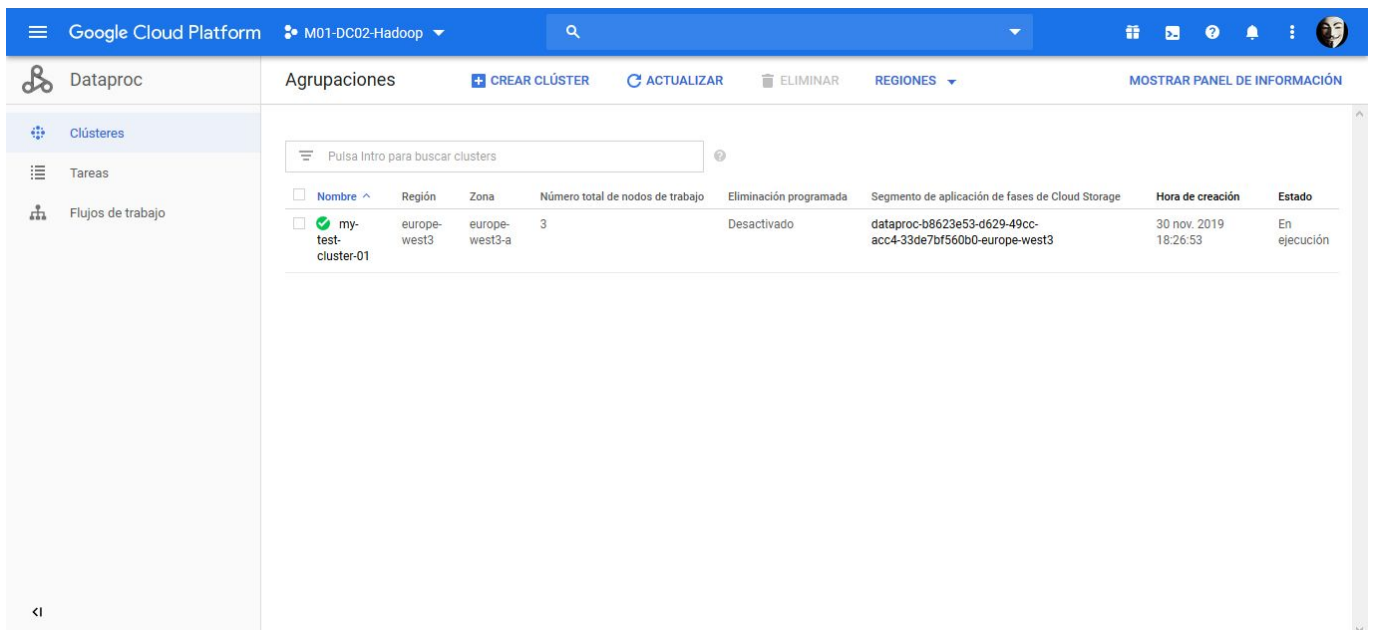
1. El primer paso es crear un clúster en Google Cloud Platform (GCP):
 - a. Desde la Google Cloud SDK Shell:

```
> gcloud beta dataproc clusters create my-test-cluster-01 --  
enable-component-gateway --region Europe-west3 --subnet default  
--zone Europe-west3-a --master-machine-type n1-standard-2 --  
master-boot-disk-size 50 --num-workers 3 --worker-machine-type  
n1-standard-2 --worker-boot-disk-size 50 --image-version 1.3-  
deb9 --project m01-dc02-hadoop
```

```

Google Cloud SDK Shell
Welcome to the Google Cloud SDK! Run "gcloud -h" to get the list of available commands.
---
C:\Users\calfo\AppData\Local\Google\Cloud SDK>gcloud beta dataproc clusters create my-test-cluster-01 --enable-component-gateway --region europe-west3 --subnet default --zone europe-west3-a --master-machine-type n1-standard-2 --master-boot-disk-size 50 --num-workers 3 --worker-machine-type n1-standard-2 --worker-boot-disk-size 50 --image-version 1.3-deb9 --project m01-dc02-hadoop
  
```

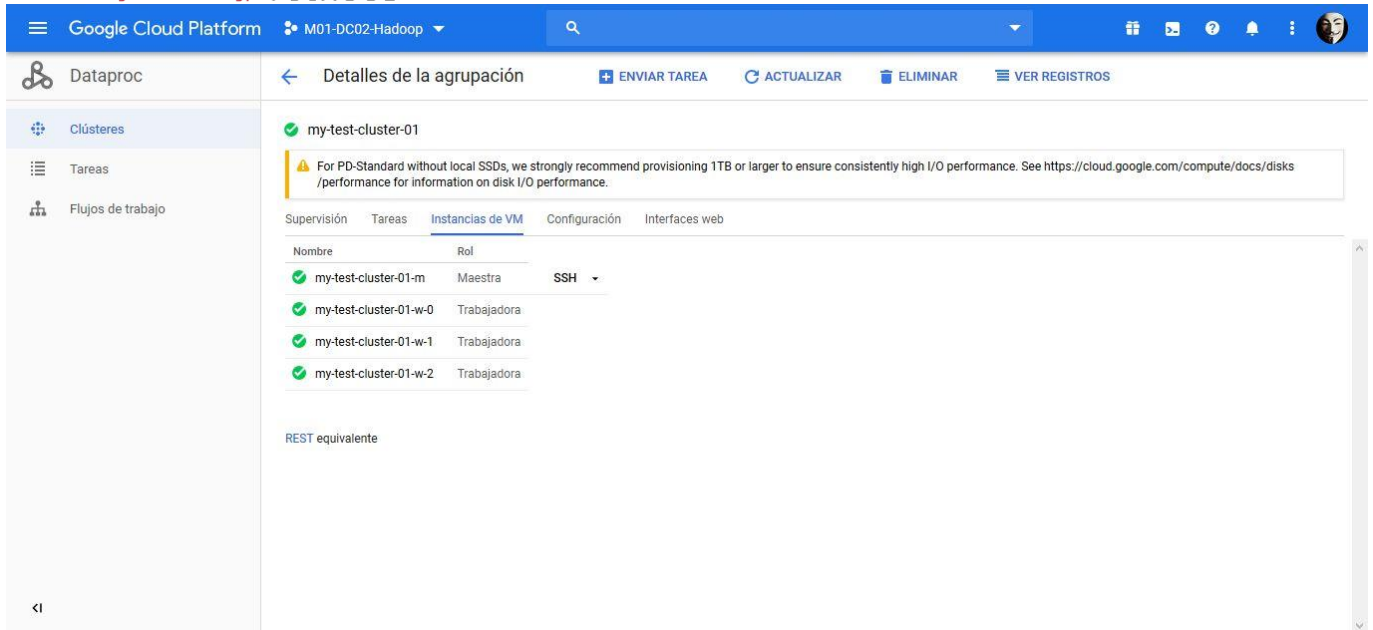
b. Comprobamos en GCP que el clúster ha arrancado correctamente:



The screenshot shows the Google Cloud Platform console for the project 'M01-DC02-Hadoop'. The 'Dataproc' section is active, and the 'Agrupaciones' (Clusters) tab is selected. A table lists the clusters, with one cluster 'my-test-cluster-01' shown in a state of 'En ejecución' (Running).

Nombre	Región	Zona	Número total de nodos de trabajo	Eliminación programada	Segmento de aplicación de fases de Cloud Storage	Hora de creación	Estado
my-test-cluster-01	europe-west3	europe-west3-a	3	Desactivado	dataproc-b8623e53-d629-49cc-acc4-33de7bf560b0-europe-west3	30 nov. 2019 18:26:53	En ejecución

c. El clúster consta de un nodo maestro y tres trabajadores, todos de tipo n1-standard-2 (2 vCPU y 7.5 GB de memoria cada uno) y 50 GB de espacio en disco.



Google Cloud Platform M01-DC02-Hadoop

Dataproc Detalles de la agrupación

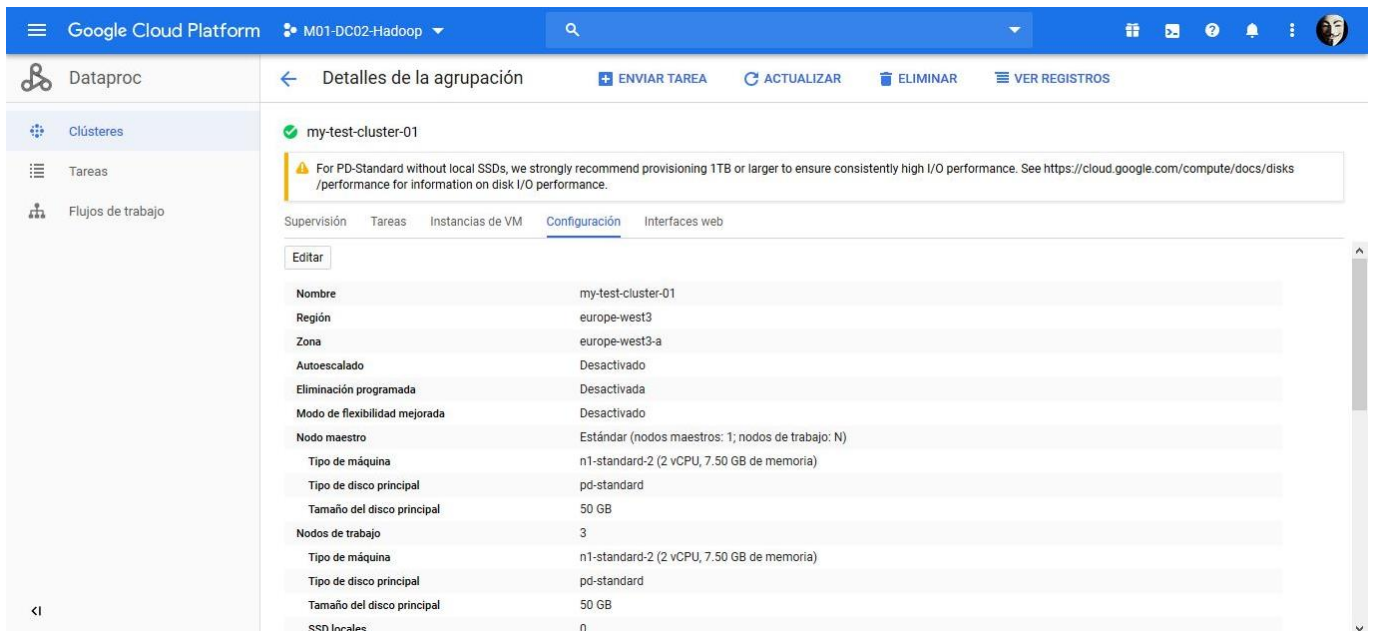
my-test-cluster-01

For PD-Standard without local SSDs, we strongly recommend provisioning 1TB or larger to ensure consistently high I/O performance. See <https://cloud.google.com/compute/docs/disks/performance> for information on disk I/O performance.

Supervisión Tareas **Instancias de VM** Configuración Interfaces web

Nombre	Rol	SSH
my-test-cluster-01-m	Maestra	SSH
my-test-cluster-01-w-0	Trabajadora	
my-test-cluster-01-w-1	Trabajadora	
my-test-cluster-01-w-2	Trabajadora	

REST equivalente



Google Cloud Platform M01-DC02-Hadoop

Dataproc Detalles de la agrupación

my-test-cluster-01

For PD-Standard without local SSDs, we strongly recommend provisioning 1TB or larger to ensure consistently high I/O performance. See <https://cloud.google.com/compute/docs/disks/performance> for information on disk I/O performance.

Supervisión Tareas Instancias de VM **Configuración** Interfaces web

Editar

Nombre	my-test-cluster-01
Región	europa-west3
Zona	europa-west3-a
Autoescalado	Desactivado
Eliminación programada	Desactivada
Modo de flexibilidad mejorada	Desactivado
Nodo maestro	Estándar (nodos maestros: 1; nodos de trabajo: N)
Tipo de máquina	n1-standard-2 (2 vCPU, 7.50 GB de memoria)
Tipo de disco principal	pd-standard
Tamaño del disco principal	50 GB
Nodos de trabajo	3
Tipo de máquina	n1-standard-2 (2 vCPU, 7.50 GB de memoria)
Tipo de disco principal	pd-standard
Tamaño del disco principal	50 GB
SSD locales	0

- A continuación, desde la consola del clúster importamos los 6 diccionarios que vamos a utilizar para la práctica (francés, alemán, italiano, latín, portugués y español), ejecutando los siguientes comandos:

```
$ wget http://www.ilovelanguages.com/IDP/files/French.txt
$ wget http://www.ilovelanguages.com/IDP/files/German.txt
$ wget http://www.ilovelanguages.com/IDP/files/Italian.txt
$ wget http://www.ilovelanguages.com/IDP/files/Latin.txt
$ wget http://www.ilovelanguages.com/IDP/files/Portuguese.txt
$ wget http://www.ilovelanguages.com/IDP/files/Spanish.txt
```

```
cajaenh@my-test-cluster-01-m: ~ - Mozilla Firefox
https://ssh.cloud.google.com/projects/m01-dc02-hadoop/zones/europe-west3-a/instances/my-test-cluster-01-m?ai...

Italian.txt          100%[=====>] 125.72K  322KB/s  in 0.4s

2019-11-30 17:41:12 (322 KB/s) - 'Italian.txt' saved [128736/128736]

cajaenh@my-test-cluster-01-m:~$ wget http://www.ilovelanguages.com/IDP/files/Portuguese.txt
--2019-11-30 17:41:24-- http://www.ilovelanguages.com/IDP/files/Portuguese.txt
Resolving www.ilovelanguages.com (www.ilovelanguages.com)... 64.71.34.99
Connecting to www.ilovelanguages.com (www.ilovelanguages.com)|64.71.34.99|:80... connected.
HTTP request sent, awaiting response... 200 OK
Length: 37076 (36K) [text/plain]
Saving to: 'Portuguese.txt'

Portuguese.txt       100%[=====>] 36.21K  206KB/s  in 0.2s

2019-11-30 17:41:25 (206 KB/s) - 'Portuguese.txt' saved [37076/37076]

cajaenh@my-test-cluster-01-m:~$ wget http://www.ilovelanguages.com/IDP/files/Spanish.txt
--2019-11-30 17:41:37-- http://www.ilovelanguages.com/IDP/files/Spanish.txt
Resolving www.ilovelanguages.com (www.ilovelanguages.com)... 64.71.34.99
Connecting to www.ilovelanguages.com (www.ilovelanguages.com)|64.71.34.99|:80... connected.
HTTP request sent, awaiting response... 200 OK
Length: 171564 (168K) [text/plain]
Saving to: 'Spanish.txt'

Spanish.txt          100%[=====>] 167.54K  373KB/s  in 0.4s

2019-11-30 17:41:38 (373 KB/s) - 'Spanish.txt' saved [171564/171564]

cajaenh@my-test-cluster-01-m:~$ ls -la
total 656
drwxr-xr-x 3 cajaenh cajaenh 4096 Nov 30 17:41 .
drwxr-xr-x 3 root      root    4096 Nov 30 17:33 ..
-rw-r--r-- 1 cajaenh cajaenh 220 May 15 2017 .bash_logout
-rw-r--r-- 1 cajaenh cajaenh 3526 May 15 2017 .bashrc
-rw-r--r-- 1 cajaenh cajaenh 87369 Apr 22 2001 French.txt
-rw-r--r-- 1 cajaenh cajaenh 211008 Apr 22 2001 German.txt
-rw-r--r-- 1 cajaenh cajaenh 128736 Apr 22 2001 Italian.txt
-rw-r--r-- 1 cajaenh cajaenh 37076 Apr 22 2001 Portuguese.txt
-rw-r--r-- 1 cajaenh cajaenh 675 May 15 2017 .profile
-rw-r--r-- 1 cajaenh cajaenh 171564 Apr 22 2001 Spanish.txt
drwx----- 2 cajaenh cajaenh 4096 Nov 30 17:37 .ssh
cajaenh@my-test-cluster-01-m:~$
```

Y antes de subirlos a HDFS los combinamos en un único fichero **dictionary.txt** que además se ordena alfabéticamente:

```
$ cat French.txt German.txt Italian.txt Latin.txt Portuguese.txt
Spanish.txt | sort > dictionary.txt
```

Para poder practicar más tarde con un fichero que supere el tamaño mínimo de bloque (1048576 bytes), y al no haber más diccionarios disponibles, concateno dos veces el archivo existente y lo llamo **dictionaryx2.txt**.

```
$ cat dictionary.txt dictionary.txt > dictionaryx2.txt
```



```

cajaenh@my-test-cluster-01-m: ~ - Mozilla Firefox
https://ssh.cloud.google.com/projects/m01-dc02-hadoop/zones/europe-west3-a/instances/my-test-cluster-01-m?auth=...

cajaenh@my-test-cluster-01-m:~$ wget http://www.ilovelanguages.com/IDP/files/Spanish.txt
--2019-11-30 17:41:37-- http://www.ilovelanguages.com/IDP/files/Spanish.txt
Resolving www.ilovelanguages.com (www.ilovelanguages.com)... 64.71.34.99
Connecting to www.ilovelanguages.com (www.ilovelanguages.com)|64.71.34.99|:80... connected.
HTTP request sent, awaiting response... 200 OK
Length: 171564 (168K) [text/plain]
Saving to: 'Spanish.txt'

Spanish.txt          100%[=====>] 167.54K  373KB/s   in 0.4s

2019-11-30 17:41:38 (373 KB/s) - 'Spanish.txt' saved [171564/171564]

cajaenh@my-test-cluster-01-m:~$ ls -la
total 656
drwxr-xr-x 3 cajaenh cajaenh 4096 Nov 30 17:41 .
drwxr-xr-x 3 root     root     4096 Nov 30 17:33 ..
-rw-r--r-- 1 cajaenh cajaenh 220 May 15 2017 .bash_logout
-rw-r--r-- 1 cajaenh cajaenh 3526 May 15 2017 .bashrc
-rw-r--r-- 1 cajaenh cajaenh 87369 Apr 22 2001 French.txt
-rw-r--r-- 1 cajaenh cajaenh 211008 Apr 22 2001 German.txt
-rw-r--r-- 1 cajaenh cajaenh 128736 Apr 22 2001 Italian.txt
-rw-r--r-- 1 cajaenh cajaenh 37076 Apr 22 2001 Portuguese.txt
-rw-r--r-- 1 cajaenh cajaenh 675 May 15 2017 .profile
-rw-r--r-- 1 cajaenh cajaenh 171564 Apr 22 2001 Spanish.txt
drwx----- 2 cajaenh cajaenh 4096 Nov 30 17:37 .ssh
cajaenh@my-test-cluster-01-m:~$ cat French.txt German.txt Italian.txt Portuguese.txt Spanish.txt | sort > joinDictio.txt
cajaenh@my-test-cluster-01-m:~$ ls -la
total 1280
drwxr-xr-x 3 cajaenh cajaenh 4096 Nov 30 17:44 .
drwxr-xr-x 3 root     root     4096 Nov 30 17:33 ..
-rw-r--r-- 1 cajaenh cajaenh 220 May 15 2017 .bash_logout
-rw-r--r-- 1 cajaenh cajaenh 3526 May 15 2017 .bashrc
-rw-r--r-- 1 cajaenh cajaenh 87369 Apr 22 2001 French.txt
-rw-r--r-- 1 cajaenh cajaenh 211008 Apr 22 2001 German.txt
-rw-r--r-- 1 cajaenh cajaenh 128736 Apr 22 2001 Italian.txt
-rw-r--r-- 1 cajaenh cajaenh 635753 Nov 30 17:44 joinDictio.txt
-rw-r--r-- 1 cajaenh cajaenh 37076 Apr 22 2001 Portuguese.txt
-rw-r--r-- 1 cajaenh cajaenh 675 May 15 2017 .profile
-rw-r--r-- 1 cajaenh cajaenh 171564 Apr 22 2001 Spanish.txt
drwx----- 2 cajaenh cajaenh 4096 Nov 30 17:37 .ssh
cajaenh@my-test-cluster-01-m:~$

```

3. Editamos los scripts de Python **mapper.py** y **reducer.py** utilizando el comando nano:

```

$ nano mapper.py
$ nano reducer.py

```

```

cajaenh@my-test-cluster-01-m: ~ - Mozilla Firefox
https://ssh.cloud.google.com/projects/m01-dc02-hadoop/zones/europe-west3-a/instances/my-test-cluster-01-m?auth=...

cajaenh@my-test-cluster-01-m:~$ cat mapper.py
#!/usr/bin/env python
# -*- coding: utf-8 -*-
import sys

# input comes from STDIN (standard input)
for line in sys.stdin:
    # Limpiamos espacimientos, buscamos el tabulador y separamos en 2 elementos (tupla)
    terms = line.strip().split('\t')
    # Volcamos la salida por consola
    print '\t'.join(terms)

cajaenh@my-test-cluster-01-m:~$ cat reducer.py
#!/usr/bin/env python
from operator import itemgetter
import sys

current_word = None
word = None
trad_complete = None
# input comes from STDIN
for line in sys.stdin:
    # remove leading and trailing whitespace
    line = line.strip().split('\t')

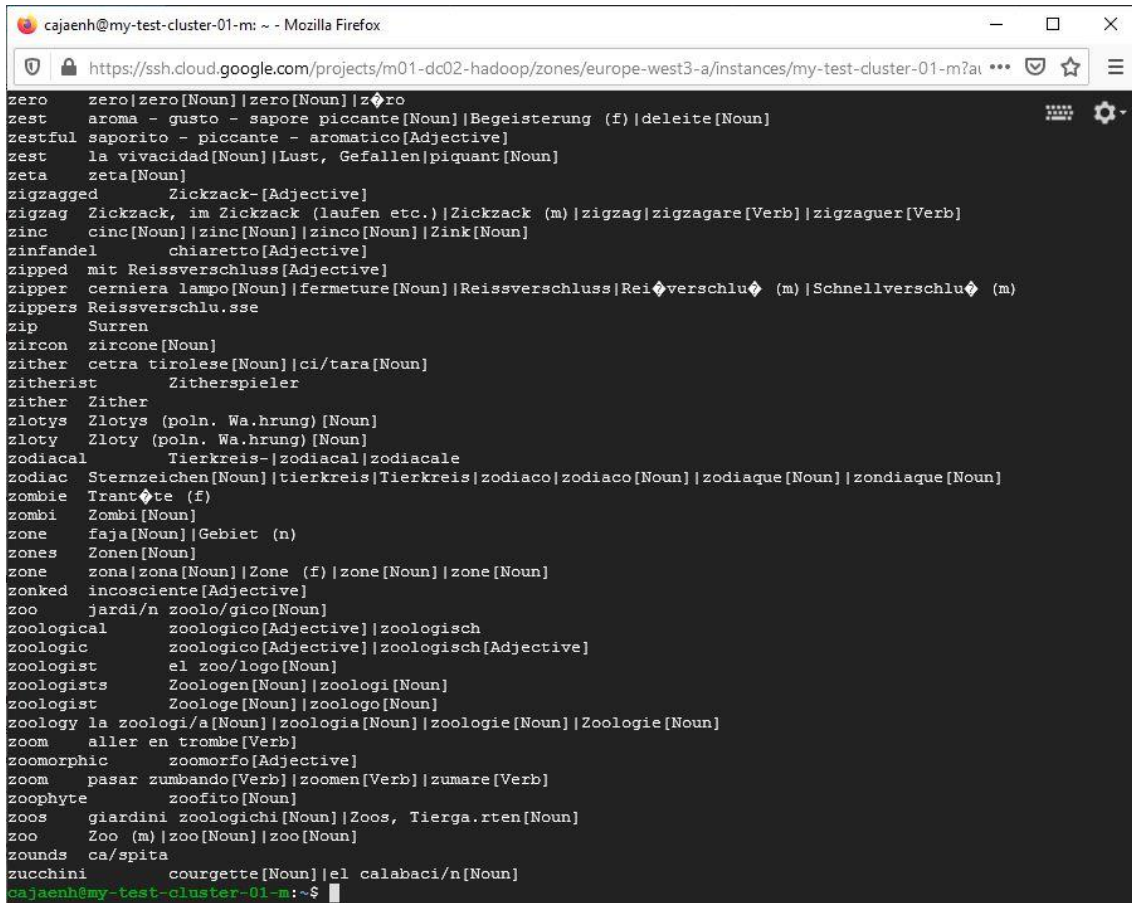
    # parse the input we got from mapper.py
    word = line[0]
    try:
        trad = line[1]
    except:
        trad = ''
    pass
    if current_word == word:
        trad_complete = trad_complete + "|" + trad
    else:
        if current_word:
            print '%s\t%s' % (current_word, trad_complete)
            trad_complete = trad
            current_word = word

# do not forget to output the last word if needed!
if current_word == word:
    print '%s\t%s' % (current_word, trad_complete)
cajaenh@my-test-cluster-01-m:~$

```

Se prueba en local que los procesos Map Reduce funcionan correctamente:

```
$ cat dictionary.txt | python mapper.py | sort | python reducer.py
```



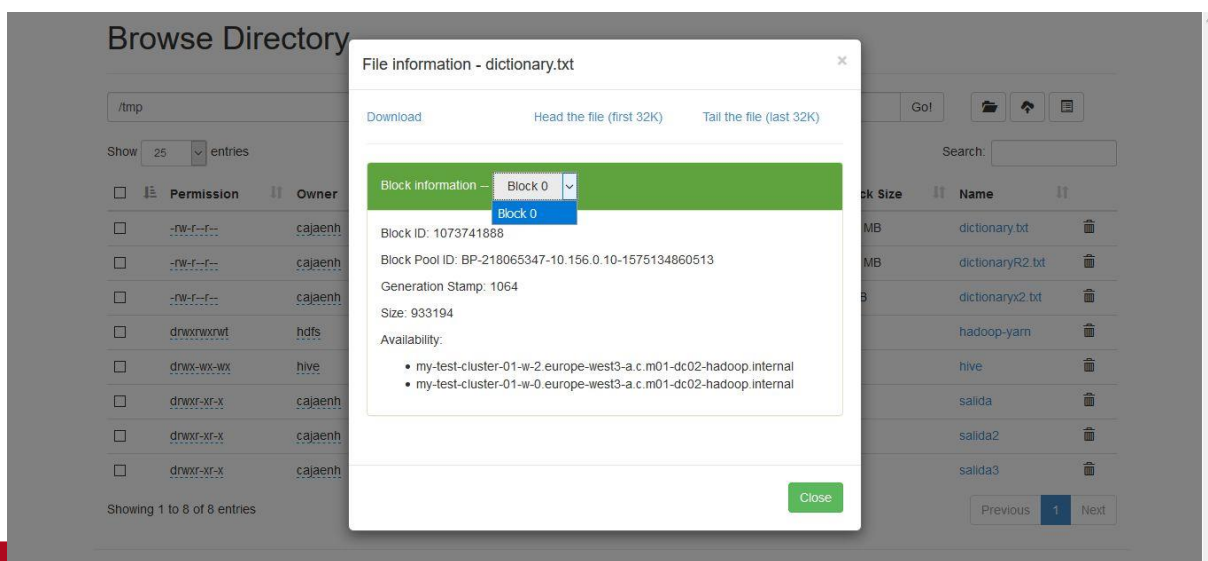
```

zero zero|zero[Noun]|zero[Noun]|zero
zest aroma - gusto - sapore piccante[Noun]|Begeisterung (f)|deleite[Noun]
zestful saporito - piccante - aromatico[Adjective]
zest la vivacidad[Noun]|Lust, Gefallen|piquant[Noun]
zeta zeta[Noun]
zigzagged Zickzack- [Adjective]
zigzag Zickzack, im Zickzack (laufen etc.)|Zickzack (m)|zigzag|zigzagare[Verb]|zigzaguer[Verb]
zinc cinc[Noun]|zinc[Noun]|zinco[Noun]|Zink[Noun]
zinfandel chiaretto[Adjective]
zipped mit Reissverschluss[Adjective]
zipper cerniera lampo[Noun]|fermeture[Noun]|Reissverschluss|Reißverschluss (m)|Schnellverschlus (m)
zippers Reissverschluss
zip Surren
zircon zircone[Noun]
zither cetra tirolese[Noun]|ci/tara[Noun]
zitherist Zitherspieler
zither Zither
zlotys Zlotys (poln. Wa.hrung) [Noun]
zloty Zloty (poln. Wa.hrung) [Noun]
zodiacal Tierkreis|zodiacal|zodiacale
zodiac Sternzeichen[Noun]|tierkreis|Tierkreis|zodiaco|zodiaco[Noun]|zodiaque[Noun]|zondiaque[Noun]
zombie Trantöte (f)
zombi Zombi[Noun]
zone faja[Noun]|Gebiet (n)
zones Zonen[Noun]
zone zona|zona[Noun]|Zone (f)|zone[Noun]|zone[Noun]
zonked incosciente[Adjective]
zoo jardi/n zoolo/gico[Noun]
zoological zoologico[Adjective]|zoologisch
zoologic zoologico[Adjective]|zoologisch[Adjective]
zoologist el zoo/logo[Noun]
zoologists Zoologen[Noun]|zoologi[Noun]
zoologist Zoologe[Noun]|zoologo[Noun]
zoology la zoologi/a[Noun]|zoologia[Noun]|zoologie[Noun]|Zoologie[Noun]
zoom aller en trombe[Verb]
zoomorphic zoomorfo[Adjective]
zoom pasar zumbando[Verb]|zoomen[Verb]|zumare[Verb]
zoophyte zoofito[Noun]
zoos giardini zoologici[Noun]|Zoos, Tiergarten[Noun]
zoo Zoo (m)|zoo[Noun]|zoo[Noun]
zounds ca/spita
zucchini courgette[Noun]|el calabaci/n[Noun]
cajaanh@my-test-cluster-01-m:~$

```

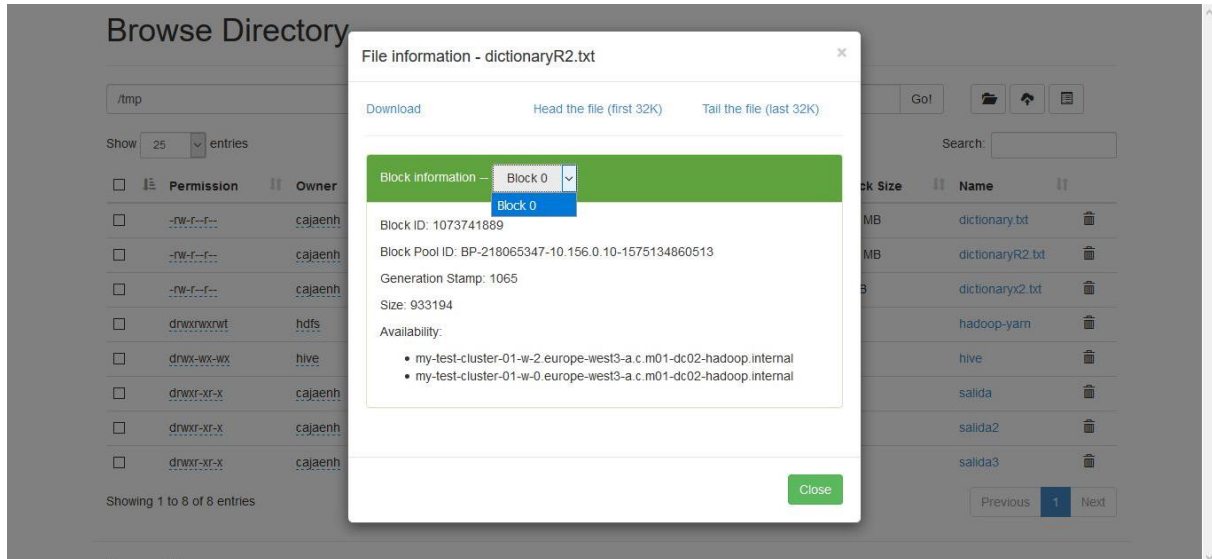
4. Subimos los diccionarios a HDFS con el comando **hadoop fs -put**:

```
$ hadoop fs -put dictionary.txt /tmp/dictionary.txt
```



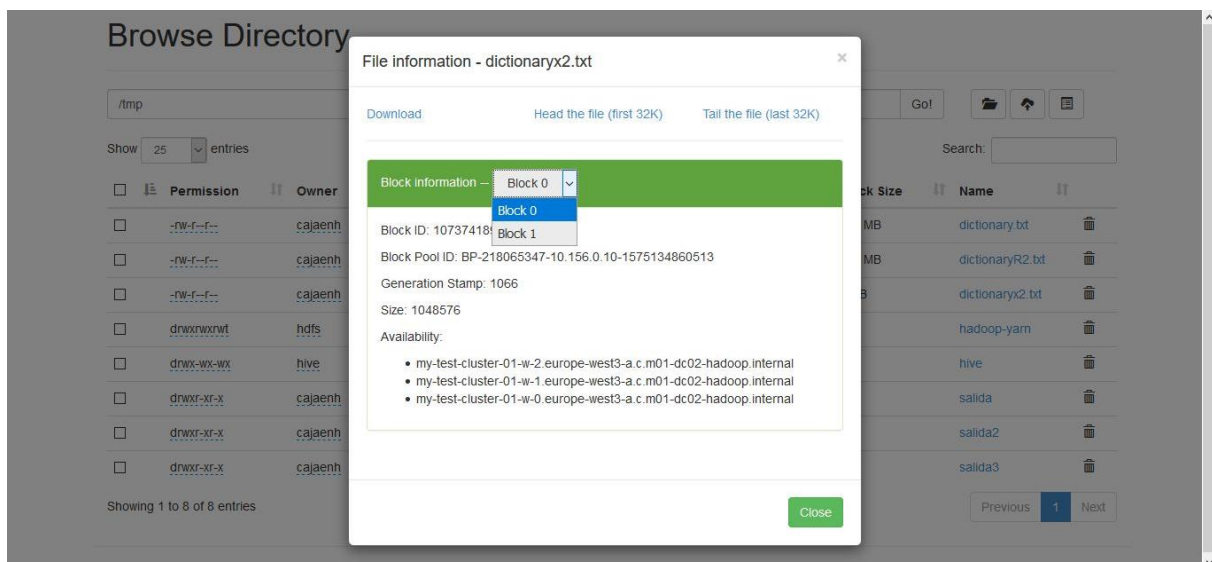
Con factor de replicación 2:

```
$ hadoop fs -D dfs.replication=2 -put dictionary.txt /tmp/dictionaryR2.txt
```



Con factor de replicación 3 y tamaño de bloque 1048576:

```
$ hadoop fs -D dfs.replication=3 -D dfs.block.size=1048576 -put dictionaryx2.txt /tmp/dictionaryx2.txt
```



5. Aplicamos la función Map Reduce a los tres diccionarios alojados en el HDFS del clúster:

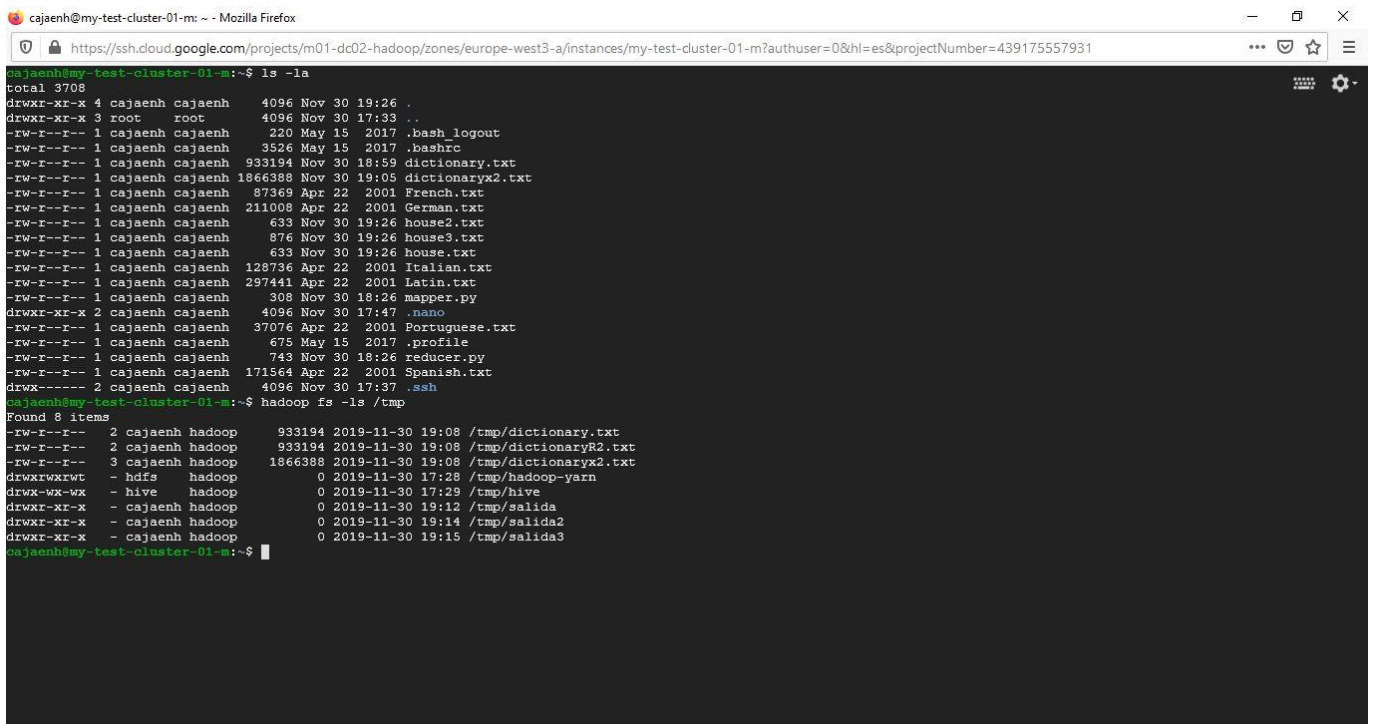
```
$ hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming.jar \
> -D mapred.reduce.tasks=1 \
> -files mapper.py,Reducer.py \
> -mapper mapper.py \
```



```
> -reducer reducer.py \
> -input /tmp/dictionary.txt \
> -output /tmp/salida
```

```
$ hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming.jar \
> -D mapred.reduce.tasks=1 \
> -files mapper.py,reducer.py \
> -mapper mapper.py \
> -reducer reducer.py \
> -input /tmp/dictionaryR2.txt \
> -output /tmp/salida2
```

```
$ hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming.jar \
> -D mapred.reduce.tasks=1 \
> -files mapper.py,reducer.py \
> -mapper mapper.py \
> -reducer reducer.py \
> -input /tmp/dictionaryx2.txt \
> -output /tmp/salida3
```



```
cajaenh@my-test-cluster-01-m: ~ - Mozilla Firefox
https://ssh.cloud.google.com/projects/m01-dc02-hadoop/zones/europe-west3-a/instances/my-test-cluster-01-m?authuser=0&hl=es&projectNumber=439175557931

cajaenh@my-test-cluster-01-m:~$ ls -la
total 3708
drwxr-xr-x 4 cajaenh cajaenh 4096 Nov 30 19:26 .
drwxr-xr-x 3 root root 4096 Nov 30 17:33 ..
-rw-r--r-- 1 cajaenh cajaenh 220 May 15 2017 .bash_logout
-rw-r--r-- 1 cajaenh cajaenh 3526 May 15 2017 .bashrc
-rw-r--r-- 1 cajaenh cajaenh 933194 Nov 30 18:59 dictionary.txt
-rw-r--r-- 1 cajaenh cajaenh 1866388 Nov 30 19:05 dictionaryx2.txt
-rw-r--r-- 1 cajaenh cajaenh 87369 Apr 22 2001 French.txt
-rw-r--r-- 1 cajaenh cajaenh 211008 Apr 22 2001 German.txt
-rw-r--r-- 1 cajaenh cajaenh 633 Nov 30 19:26 house2.txt
-rw-r--r-- 1 cajaenh cajaenh 876 Nov 30 19:26 house3.txt
-rw-r--r-- 1 cajaenh cajaenh 633 Nov 30 19:26 house.txt
-rw-r--r-- 1 cajaenh cajaenh 128736 Apr 22 2001 Italian.txt
-rw-r--r-- 1 cajaenh cajaenh 297441 Apr 22 2001 Latin.txt
-rw-r--r-- 1 cajaenh cajaenh 308 Nov 30 18:26 mapper.py
drwxr-xr-x 2 cajaenh cajaenh 4096 Nov 30 17:47 .nano
-rw-r--r-- 1 cajaenh cajaenh 37076 Apr 22 2001 Portuguese.txt
-rw-r--r-- 1 cajaenh cajaenh 675 May 15 2017 .profile
-rw-r--r-- 1 cajaenh cajaenh 743 Nov 30 18:26 reducer.py
-rw-r--r-- 1 cajaenh cajaenh 171564 Apr 22 2001 Spanish.txt
drwxr-xr-x 2 cajaenh cajaenh 4096 Nov 30 17:37 .ssh
cajaenh@my-test-cluster-01-m:~$ hadoop fs -ls /tmp
Found 8 items
-rw-r--r-- 2 cajaenh hadoop 933194 2019-11-30 19:08 /tmp/dictionary.txt
-rw-r--r-- 2 cajaenh hadoop 933194 2019-11-30 19:08 /tmp/dictionaryR2.txt
-rw-r--r-- 3 cajaenh hadoop 1866388 2019-11-30 19:08 /tmp/dictionaryx2.txt
drwxrwxrwt - hdfs hadoop 0 2019-11-30 17:28 /tmp/hadoop-yarn
drwx-wx-wx - hive hadoop 0 2019-11-30 17:29 /tmp/hive
drwxr-xr-x - cajaenh hadoop 0 2019-11-30 19:12 /tmp/salida
drwxr-xr-x - cajaenh hadoop 0 2019-11-30 19:14 /tmp/salida2
drwxr-xr-x - cajaenh hadoop 0 2019-11-30 19:15 /tmp/salida3
cajaenh@my-test-cluster-01-m:~$
```

Al ejecutarse **hadoop fs -ls /tmp** pueden verse los ficheros de salida que se han generado en el HDFS.

En la pestaña **Hadoop JobHistory** pueden verse los procesos ejecutados y su estado (SUCCEEDED o FAILED):



JobHistory

Logged in

Application

About Jobs

Tools

Retired Jobs

Submit Time	Start Time	Finish Time	Job ID	Name	User	Queue	State	Maps Total	Maps Completed	Reduces Total	Reduces Completed
2019.11.30 19:14:47 UTC	2019.11.30 19:14:53 UTC	2019.11.30 19:15:37 UTC	job_1575134861452_0007	streamjob5821381651553386572.jar	cajaenh	default	SUCCEEDED	24	24	1	1
2019.11.30 19:13:18 UTC	2019.11.30 19:13:23 UTC	2019.11.30 19:14:08 UTC	job_1575134861452_0006	streamjob2991487431454781366.jar	cajaenh	default	SUCCEEDED	24	24	1	1
2019.11.30 19:11:57 UTC	2019.11.30 19:12:04 UTC	2019.11.30 19:12:48 UTC	job_1575134861452_0005	streamjob8237191398451268464.jar	cajaenh	default	SUCCEEDED	24	24	1	1
2019.11.30 18:42:59 UTC	2019.11.30 18:43:06 UTC	2019.11.30 18:43:51 UTC	job_1575134861452_0004	streamjob2389816293434595533.jar	cajaenh	default	SUCCEEDED	24	24	1	1
2019.11.30 18:36:05 UTC	2019.11.30 18:36:11 UTC	2019.11.30 18:36:52 UTC	job_1575134861452_0003	streamjob3852836084302957766.jar	cajaenh	default	SUCCEEDED	24	24	1	1
2019.11.30 18:32:39 UTC	2019.11.30 18:32:46 UTC	2019.11.30 18:33:40 UTC	job_1575134861452_0002	streamjob3901597924196659040.jar	cajaenh	default	SUCCEEDED	24	24	8	8
2019.11.30 18:22:45 UTC	2019.11.30 18:22:55 UTC	2019.11.30 18:23:36 UTC	job_1575134861452_0001	streamjob3175217803930185763.jar	cajaenh	default	FAILED	0	0	0	0

6. En el navegador web del HDFS de nuestro clúster podemos ver su contenido: los tres diccionarios subidos, más las tres carpetas de salida con el resultado de aplicar la función Map Reduce a esos ficheros:

Hadoop Overview Datanodes Datanode Volume Failures Snapshot Startup Progress Utilities

Browse Directory

/tmp/ Go!

Show 25 entries Search:




Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	cajaenh	hadoop	911.32 KB	Nov 30 20:08	2	128 MB	dictionary.txt
-rw-r--r--	cajaenh	hadoop	911.32 KB	Nov 30 20:08	2	128 MB	dictionaryR2.txt
-rw-r--r--	cajaenh	hadoop	1.78 MB	Nov 30 20:08	3	1 MB	dictionaryx2.txt
drwxrwxrwt	hdfs	hadoop	0 B	Nov 30 18:28	0	0 B	hadoop-yarn
drwx-wx-wx	hive	hadoop	0 B	Nov 30 18:29	0	0 B	hive
drwxr-xr-x	cajaenh	hadoop	0 B	Nov 30 20:12	0	0 B	salida
drwxr-xr-x	cajaenh	hadoop	0 B	Nov 30 20:14	0	0 B	salida2
drwxr-xr-x	cajaenh	hadoop	0 B	Nov 30 20:15	0	0 B	salida3

Showing 1 to 8 of 8 entries



Previous 1 Next

- a. En la carpeta **salida** se encuentra el resultado (satisfactorio) de aplicar el Map Reduce al fichero **dictionary.txt** (se adjunta el **archivo salida_part-00000.txt**):

Browse Directory

/tmp/salida Go!   

Show 25 entries Search:




<input type="checkbox"/>	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name	
<input type="checkbox"/>	-rw-r--r--	cajaenh	hadoop	0 B	Nov 30 20:12	2	128 MB	_SUCCESS	
<input type="checkbox"/>	-rw-r--r--	cajaenh	hadoop	807.22 KB	Nov 30 20:12	2	128 MB	part-00000	

Showing 1 to 2 of 2 entries Previous 1 Next



Hadoop, 2018.

- b. En la carpeta **salida2** se encuentra el resultado (satisfactorio) de aplicar el Map Reduce al fichero **dictionaryR2.txt** (se adjunta el **archivo salida2_part-00000.txt**):

Browse Directory

/tmp/salida2 Go!   

Show 25 entries Search:

<input type="checkbox"/>	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name	
<input type="checkbox"/>	-rw-r--r--	cajaenh	hadoop	0 B	Nov 30 20:14	2	128 MB	_SUCCESS	
<input type="checkbox"/>	-rw-r--r--	cajaenh	hadoop	807.22 KB	Nov 30 20:14	2	128 MB	part-00000	

Showing 1 to 2 of 2 entries Previous 1 Next

Hadoop, 2018.

- c. En la carpeta **salida3** se encuentra el resultado (satisfactorio) de aplicar el Map Reduce al fichero **dictionaryx2.txt** (se adjunta el **archivo salida3_part-00000.txt**):

Browse Directory

/tmp/salida3 Go!

Show 25 entries Search:

<input type="checkbox"/>	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
<input type="checkbox"/>	-rw-r--r--	cajaenh	hadoop	0 B	Nov 30 20:15	2	128 MB	_SUCCESS
<input type="checkbox"/>	-rw-r--r--	cajaenh	hadoop	1.25 MB	Nov 30 20:15	2	128 MB	part-00000

Showing 1 to 2 of 2 entries

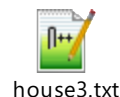
Previous 1 Next

Hadoop, 2018.

7. Por último, para mostrar el resultado de la palabra **House** utilizamos el comando **grep**:

```
$ hadoop fs -cat /tmp/salida/part-00000 | sort | grep -i -w House
> house.txt
$ hadoop fs -cat /tmp/salida2/part-00000 | sort | grep -i -w House
> house2.txt
$ hadoop fs -cat /tmp/salida3/part-00000 | sort | grep -i -w House
> house3.txt
```

Los resultados se guardan en ficheros de texto **house.txt**, **house2.txt** y **house3.txt**.



Si se quiere ver el contenido de cualquiera de ellos utilizamos el comando **cat**:

```
$ cat house.txt
$ cat house2.txt
$ cat house3.txt
```

El resultado es el siguiente (el de **house3.txt** duplicando las líneas de los otros dos):

```
butcher, slaughter-house. macellarius
casa house[Conjunction]|casa (house)
chapter, chapter meeting, chapter house. capitulus
country house, country estate /(med.) manor, village. villa
guest house Pension (f)
house casa[Noun]|casa|CASA|casa[Noun]|CASITA|das Haus|Haus
(n)|la casa|maison[Noun]|Unterkunft (f)|Rente (f)
household, house, abode. domus
```

house, home, residence. domus
house wine Hauswein (m)
lady of the house Hausfrau (f)
(masc. nom. sing.) THAT (house) is filthy. ille
(masc. nom. sing.) THIS (house) is filthy. hic
(prep. + acc.) among, in the presence of, at, at the house of.
apud
wall (of a house). paries parietis