

EJERCICIOS DE MAP REDUCE

Escribe las traducciones de la palabra PIG.

En primer lugar compruebo si están arrancados los demonios:

```
bigdata@bigdata:~/hadoop$ jps
9856 DataNode
10083 SecondaryNameNode
10235 ResourceManager
10366 NodeManager
24203 Jps
8267 JobHistoryServer
9720 NameNode
bigdata@bigdata:~/hadoop$
```

Veo que están arrancados.

Si no estuvieran tendría que arrancarlos con los siguientes comandos:

```
./sbin/start-dfs.sh
```

```
./sbin/start-yarn.sh
```

```
./sbin/mr-jobhistory-daemon.sh start historyserver
```

Descargo de los ficheros de diccionarios de la web con el comando wget:

```
wget http://www.ilovelanguages.com/IDP/files/Spanish.txt
```

```
wget http://www.ilovelanguages.com/IDP/files/Italian.txt
```

```
wget http://www.ilovelanguages.com/IDP/files/French.txt
```

```
wget http://www.ilovelanguages.com/IDP/files/German.txt
```

```
bigdata@bigdata:~$ mkdir ejercicioclase
bigdata@bigdata:~$ cd ejercicioclase
bigdata@bigdata:~/ejercicioclase$ wget http://www.ilovelanguages.com/IDP/files/Spanish.txt
--2016-07-09 13:53:12-- http://www.ilovelanguages.com/IDP/files/Spanish.txt
Resolviendo www.ilovelanguages.com (www.ilovelanguages.com)... 64.71.34.99
Conectando con www.ilovelanguages.com (www.ilovelanguages.com)[64.71.34.99]:80... conectado.
Petición HTTP enviada, esperando respuesta... 200 OK
Longitud: 171564 (168K) [text/plain]
Grabando a: "Spanish.txt"

100%[=====]

2016-07-09 13:53:13 (352 KB/s) - "Spanish.txt" guardado [171564/171564]

bigdata@bigdata:~/ejercicioclase$ wget http://www.ilovelanguages.com/IDP/files/Italian.txt
--2016-07-09 13:53:45-- http://www.ilovelanguages.com/IDP/files/Italian.txt
Resolviendo www.ilovelanguages.com (www.ilovelanguages.com)... 64.71.34.99
Conectando con www.ilovelanguages.com (www.ilovelanguages.com)[64.71.34.99]:80... conectado.
Petición HTTP enviada, esperando respuesta... 200 OK
Longitud: 128736 (126K) [text/plain]
Grabando a: "Italian.txt"

100%[=====]

2016-07-09 13:53:45 (313 KB/s) - "Italian.txt" guardado [128736/128736]

bigdata@bigdata:~/ejercicioclase$ wget http://www.ilovelanguages.com/IDP/files/French.txt
--2016-07-09 13:54:02-- http://www.ilovelanguages.com/IDP/files/French.txt
Resolviendo www.ilovelanguages.com (www.ilovelanguages.com)... 64.71.34.99
Conectando con www.ilovelanguages.com (www.ilovelanguages.com)[64.71.34.99]:80... conectado.
Petición HTTP enviada, esperando respuesta... 200 OK
Longitud: 87369 (85K) [text/plain]
Grabando a: "French.txt"

100%[=====]

2016-07-09 13:54:08 (351 KB/s) - "French.txt" guardado [87369/87369]

bigdata@bigdata:~/ejercicioclase$ wget http://www.ilovelanguages.com/IDP/files/German.txt
--2016-07-09 13:54:37-- http://www.ilovelanguages.com/IDP/files/German.txt
Resolviendo www.ilovelanguages.com (www.ilovelanguages.com)... 64.71.34.99
Conectando con www.ilovelanguages.com (www.ilovelanguages.com)[64.71.34.99]:80... conectado.
Petición HTTP enviada, esperando respuesta... 200 OK
Longitud: 211008 (206K) [text/plain]
Grabando a: "German.txt"

100%[=====]

2016-07-09 13:54:58 (373 KB/s) - "German.txt" guardado [211008/211008]

bigdata@bigdata:~/ejercicioclase$
```

Compruebo que se han cargado los cuatro ficheros:

```
bigdata@bigdata:~/ejercicioclase$ ls
French.txt German.txt Italian.txt Spanish.txt
```

Para trabajar con un solo fichero fusiono los 4 ficheros en uno con nombre dictionary.txt

```
cat French.txt > dictionary.txt
```

```
cat Spanish.txt >> dictionary.txt
```

```
cat Italian.txt >> dictionary.txt
```

```
cat German.txt >> dictionary.txt
```

```
bigdata@bigdata:~/ejercicioclase$ cat French.txt > dictionary.txt
bigdata@bigdata:~/ejercicioclase$ cat Spanish.txt >> dictionary.txt
bigdata@bigdata:~/ejercicioclase$ cat Italian.txt >> dictionary.txt
bigdata@bigdata:~/ejercicioclase$ cat German.txt >> dictionary.txt
bigdata@bigdata:~/ejercicioclase$ ls
dictionary.txt French.txt German.txt Italian.txt Spanish.txt
bigdata@bigdata:~/ejercicioclase$
```

Puedo ver que ha aparecido un fichero Nuevo dictionary.txt

En HDFS creo una carpeta nueva: mapreduce.

```
hdfs dfs -mkdir /user/bigdata/mapreduce
```

Copio de local a HDFS el fichero dictionary.txt y le doy nombre diccionario.txt

```
hdfs dfs -appendToFile dictionary.txt /user/bigdata/mapreduce/diccionario.txt
```

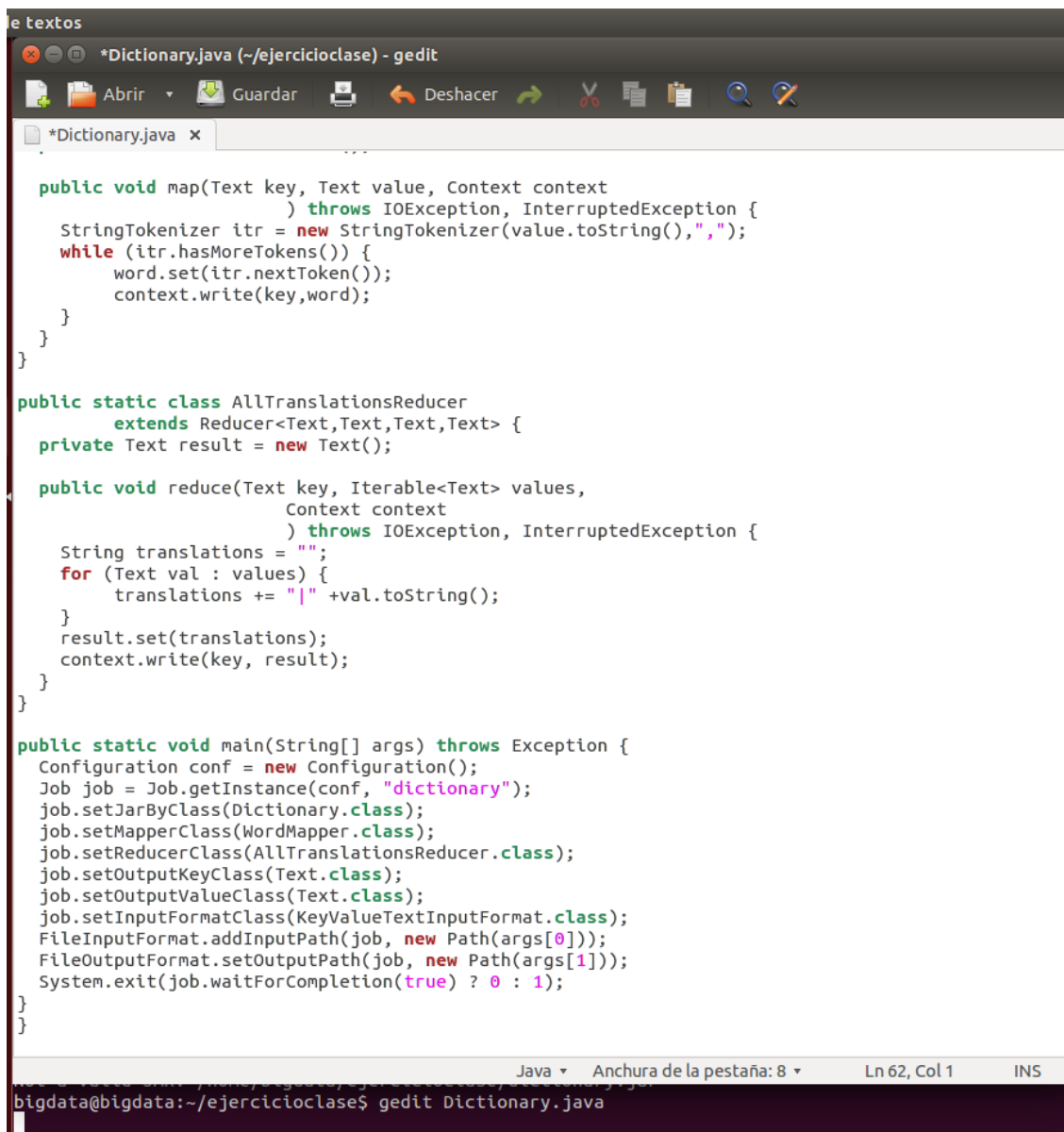
Compruebo si diccionario.txt está en HDFS:

```
hdfs dfs -ls /user/bigdata/mapreduce
```

```
bigdata@bigdata:~/ejercicioclase$ hdfs dfs -mkdir /user/bigdata/mapreduce
16/08/03 01:08:20 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using built
bigdata@bigdata:~/ejercicioclase$ hdfs dfs -appendToFile dictionary.txt /user/bigdata/mapreduce/diccionario.txt
16/08/03 01:08:53 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using built
bigdata@bigdata:~/ejercicioclase$ hdfs dfs -ls /user/bigdata/mapreduce
16/08/03 01:09:19 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using built
Found 1 items
-rw-r--r--  1 bigdata supergroup    598677 2016-08-03 01:08 /user/bigdata/mapreduce/diccionario.txt
```

Ahora voy a preparar un programa para usar los diccionarios con map reduce. EL código en java ya tengo escrito y solo creo un fichero .java pegando el código ahí.

gedit Dictionary.java



```
public void map(Text key, Text value, Context context
                ) throws IOException, InterruptedException {
    StringTokenizer itr = new StringTokenizer(value.toString(), ",");
    while (itr.hasMoreTokens()) {
        word.set(itr.nextToken());
        context.write(key,word);
    }
}

public static class AllTranslationsReducer
    extends Reducer<Text,Text,Text,Text> {
    private Text result = new Text();

    public void reduce(Text key, Iterable<Text> values,
                      Context context
                      ) throws IOException, InterruptedException {
        String translations = "";
        for (Text val : values) {
            translations += "|" +val.toString();
        }
        result.set(translations);
        context.write(key, result);
    }
}

public static void main(String[] args) throws Exception {
    Configuration conf = new Configuration();
    Job job = Job.getInstance(conf, "dictionary");
    job.setJarByClass(Dictionary.class);
    job.setMapperClass(WordMapper.class);
    job.setReducerClass(AllTranslationsReducer.class);
    job.setOutputKeyClass(Text.class);
    job.setOutputValueClass(Text.class);
    job.setInputFormatClass(KeyValueTextInputFormat.class);
    FileInputFormat.addInputPath(job, new Path(args[0]));
    FileOutputFormat.setOutputPath(job, new Path(args[1]));
    System.exit(job.waitForCompletion(true) ? 0 : 1);
}
```

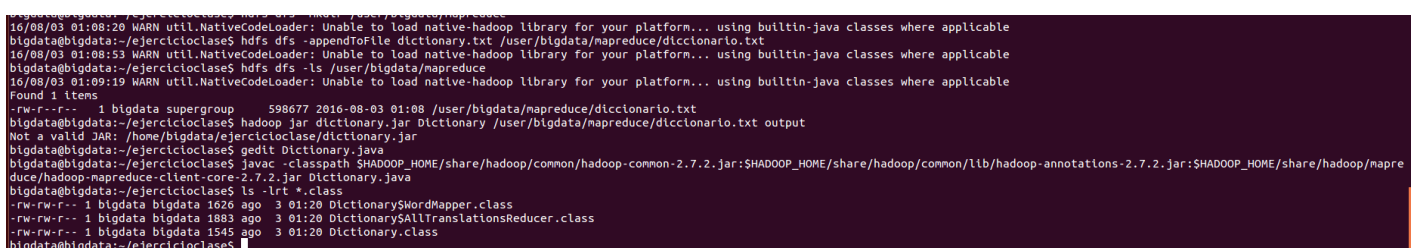
Tengo que compilar el código java para obtener fichers .class:

```
javac -classpath $HADOOP_HOME/share/hadoop/common/hadoop-common
```

```
2.7.2.jar:$HADOOP_HOME/share/hadoop/common/lib/hadoop-annotations-
```

```
2.7.2.jar:$HADOOP_HOME/share/hadoop/mapreduce/hadoop-mapreduce-client-core-2.7.2.jar Dictionary.java
```

Con ls -lrt *.class puedo ver los ficheros



```
16/08/03 01:08:20 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
bigdata@bigdata:~/ejercicioclase$ hdfs dfs -appendToFile dictionary.txt /user/bigdata/mapreduce/diccionario.txt
16/08/03 01:08:53 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
bigdata@bigdata:~/ejercicioclase$ hdfs dfs -ls /user/bigdata/mapreduce
16/08/03 01:09:19 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 1 items
-rw-r--r-- 1 bigdata supergroup 598677 2016-08-03 01:08 /user/bigdata/mapreduce/diccionario.txt
bigdata@bigdata:~/ejercicioclase$ hadoop jar dictionary.jar Dictionary /user/bigdata/mapreduce/diccionario.txt output
Not a valid JAR: /home/bigdata/ejercicioclase/dictionary.jar
bigdata@bigdata:~/ejercicioclase$ gedit Dictionary.java
bigdata@bigdata:~/ejercicioclase$ javac -classpath $HADOOP_HOME/share/hadoop/common/hadoop-common-2.7.2.jar:$HADOOP_HOME/share/hadoop/common/lib/hadoop-annotations-2.7.2.jar:$HADOOP_HOME/share/hadoop/mapre
duce/hadoop-mapreduce-client-core-2.7.2.jar Dictionary.java
bigdata@bigdata:~/ejercicioclase$ ls -lrt *.class
-rw-rw-r-- 1 bigdata bigdata 1626 ago 3 01:20 DictionaryWordMapper.class
-rw-rw-r-- 1 bigdata bigdata 1883 ago 3 01:20 Dictionary$AllTranslationsReducer.class
-rw-rw-r-- 1 bigdata bigdata 1545 ago 3 01:20 Dictionary.class
bigdata@bigdata:~/ejercicioclase$
```

Uso jar cf dictionary.jar Dictionary*.class para comprimir los ficheros class en un fichero .jar

Veo el resultado con ls -lrt dictionary.jar

```
bigdata@bigdata:~/ejercicioclase$ jar cf dictionary.jar Dictionary*.class
bigdata@bigdata:~/ejercicioclase$ ls -lrt dictionary.jar
-rw-rw-r-- 1 bigdata bigdata 3092 ago  3 01:24 dictionary.jar
bigdata@bigdata:~/ejercicioclase$
```

Ya puedo ejecutar el programa. Voy a escribir los resultados en carpeta output.

hadoop jar dictionary.jar Dictionary /user/bigdata/mapreduce/diccionario.txt output

```
bigdata@bigdata:~/ejercicioclase$ hadoop jar dictionary.jar Dictionary /user/bigdata/mapreduce/diccionario.txt output
16/08/03 01:28:42 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-
16/08/03 01:28:44 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
16/08/03 01:28:44 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the
16/08/03 01:28:45 INFO input.FileInputFormat: Total input paths to process : 1
16/08/03 01:28:45 INFO mapreduce.JobSubmitter: number of splits:1
16/08/03 01:28:45 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1469204298658_0013
16/08/03 01:28:46 INFO impl.YarnClientImpl: Submitted application application_1469204298658_0013
16/08/03 01:28:46 INFO mapreduce.Job: The url to track the job: http://bigdata:8088/proxy/application_1469204298658_00
16/08/03 01:28:46 INFO mapreduce.Job: Running job: job_1469204298658_0013
16/08/03 01:28:55 INFO mapreduce.Job: Job job_1469204298658_0013 running in uber mode : false
16/08/03 01:28:55 INFO mapreduce.Job:  map 0% reduce 0%
16/08/03 01:29:03 INFO mapreduce.Job:  map 100% reduce 0%
16/08/03 01:29:12 INFO mapreduce.Job:  map 100% reduce 100%
16/08/03 01:29:13 INFO mapreduce.Job: Job job_1469204298658_0013 completed successfully
16/08/03 01:29:13 INFO mapreduce.Job: Counters: 49
    File System Counters
      FILE: Number of bytes read=665645
      FILE: Number of bytes written=1566261
      FILE: Number of read operations=0
      FILE: Number of large read operations=0
      FILE: Number of write operations=0
      HDFS: Number of bytes read=598802
      HDFS: Number of bytes written=524677
      HDFS: Number of read operations=6
      HDFS: Number of large read operations=0
    Job Counters
      Launched map tasks=1
      Launched reduce tasks=1
      Data-local map tasks=1
      Total time spent by all maps in occupied slots (ms)=6252
      Total time spent by all reduces in occupied slots (ms)=6934
      Total time spent by all map tasks (ms)=6252
      Total time spent by all reduce tasks (ms)=6934
      Total vcore-milliseconds taken by all map tasks=6252
      Total vcore-milliseconds taken by all reduce tasks=6934
      Total megabyte-milliseconds taken by all map tasks=6402048
      Total megabyte-milliseconds taken by all reduce tasks=7100416
    Map-Reduce Framework
      Map input records=25647
      Map output records=27046
      Map output bytes=611547
      Map output materialized bytes=665645
      Input split bytes=125
      Combine input records=0
      Combine output records=0
      Reduce input groups=13901
      Reduce shuffle bytes=665645
      Reduce input records=27046
      Reduce output records=13901
      Spilled Records=54092
      Shuffled Maps =1
      Failed Shuffles=0
      Merged Map outputs=1
      GC time elapsed (ms)=113
      CPU time spent (ms)=3470
      Physical memory (bytes) snapshot=396959744
      Virtual memory (bytes) snapshot=1602519040
      Total committed heap usage (bytes)=305790976
    Shuffle Errors
      BAD_ID=0
      CONNECTION=0
      IO_ERROR=0
      WRONG_LENGTH=0
      WRONG_MAP=0
      WRONG_REDUCE=0
    File Input Format Counters
      Bytes Read=598677
    File Output Format Counters
      Bytes Written=524677
bigdata@bigdata:~/ejercicioclase$
```

Compruebo que se han guardado los resultados

```
hdfs dfs -ls -R /user/bigdata/output
```

```
bigdata@bigdata:~/ejercicioclase$ hdfs dfs -ls -R /user/bigdata/output
16/08/03 01:34:48 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java
-rw-r--r--  1 bigdata supergroup          0 2016-08-03 01:29 /user/bigdata/output/_SUCCESS
-rw-r--r--  1 bigdata supergroup    524677 2016-08-03 01:29 /user/bigdata/output/part-r-00000
bigdata@bigdata:~/ejercicioclase$
```

Copio los resultados a un fichero traducción.txt en local

```
hdfs dfs -cat /user/bigdata/output/* > traduccion.txt
```

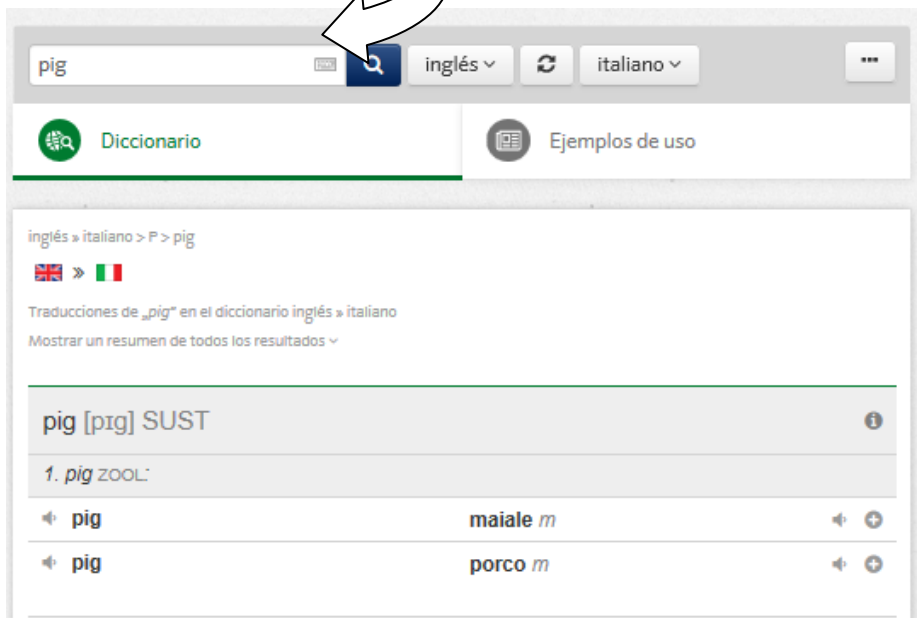
Ahora puedo hacer una consulta en el fichero y ver las traducciones de “pig”

```
grep 'pig\s' traduccion.txt
```

```
bigdata@bigdata:~/ejercicioclase$ grep 'pig\s' traduccion.txt
pig      |cochon[Noun]|Schwein (n)|el chancho[Noun]|el puerco
bigdata@bigdata:~/ejercicioclase$
```

Obtengo las traducciones:

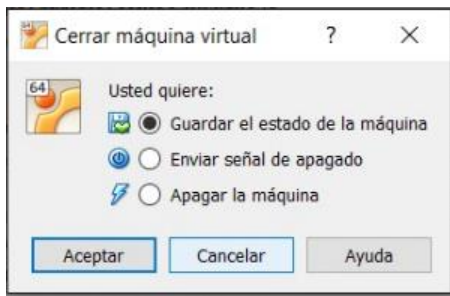
- ❖ Cochon (francés)
- ❖ Schwein (alemán)
- ❖ Chancho (italiano??)
- ❖ Puerco (Español)



The screenshot shows a web dictionary interface. At the top, there is a search bar with the word 'pig' entered. Below the search bar, there are tabs for 'Diccionario' and 'Ejemplos de uso'. The 'Diccionario' tab is selected. The interface shows the translation of 'pig' from English to Italian. The word 'pig' is shown in English, and the translations in Italian are 'maiale' and 'porco'. The interface also shows the word 'pig' in Italian and its translations in English.

| Word | Translations |
|----------------|--------------|
| pig [pig] SUST | |
| 1. pig ZOOL: | |
| pig | maiale m |
| pig | porco m |

Para finalizar cierre y guardo la máquina Archivo/cerrar y guardar el estado de la máquina.



Si no hago eso tendía que parar los demonios:

```
cd $HADOOP_HOME  
./sbin/stop-dfs.sh  
./sbin/stop-yarn.sh  
./sbin/mr-jobhistory-daemon.sh stop historyserver
```