Recomendaciones para desarrollar las prácticas 1-3:

el curso cuenta con una máquina virtual preparada sobre la distribución Ubuntu 20.04, lo cual puede diferir de las capturas mostradas en la plataforma.

Usuario: bigdata
Password: bigdata

Para interactuar de mejor manera con esta máquina puede instalar Guest Additions

3.2. Ejecutar el programa WordCount (MapReduce)

1. Abrir la terminal de comandos

Para ejecutar WordCount (MapReduce), primero crearemos un archivo con un texto en nuestra máquina local y luego lo cargaremos a HDFS:

```
bigdata@bigdata-VirtualBox:~$ echo "Salve salve Señor del Olimpo, salve Rey de todos los tiempos" > palabras.txt
```

Nótese que el símbolo “>” redirecciona la salida al archivo palabras.txt y dado que no existe, lo crea.

Lo podemos ver con ls:

```
bigdata@bigdata-VirtualBox:~$ ls -l
total 60
-rw-rw-r--  1 bigdata bigdata   38 mar  4 15:25 datos1.txt
-rw-r--r--  1 bigdata bigdata   38 mar  4 15:30 datos2.txt
-rw-rw-r--  1 bigdata bigdata    0 mar  3 23:59 demo.csv
drwxr-xr-x  2 bigdata bigdata 4096 jul  9 22:46 Desktop
drwxr-xr-x  2 bigdata bigdata 4096 mar  4 17:28 Documents
drwxr-xr-x  2 bigdata bigdata 4096 mar  4 17:20 Downloads
drwxr-xr-x 11 bigdata bigdata 4096 mar  3 23:50 hadoop-3.4.1
drwxr-xr-x  2 bigdata bigdata 4096 mar  3 17:16 Music
-rw-------  1 bigdata bigdata    1 mar  3 22:20 nano.save
-rw-rw-r--  1 bigdata bigdata   62 jul 29 02:20 palabras.txt
drwxr-xr-x  2 bigdata bigdata 4096 mar  3 17:16 Pictures
drwxr-xr-x  2 bigdata bigdata 4096 mar  3 17:16 Public
drwx------  3 bigdata bigdata 4096 mar  3 17:25 snap
drwxr-xr-x  2 bigdata bigdata 4096 mar  3 17:16 Templates
drwxr-xr-x  2 bigdata bigdata 4096 mar  3 17:16 Videos
-rw-rw-r--  1 bigdata bigdata 2513 mar  4 17:24 words-laodisea.txt
-rw-r--r--  1 bigdata bigdata    0 jul 29 01:54 words.txt
bigdata@bigdata-VirtualBox:~$
```

Subamos ese archivo a HDFS y listemos el contenido de la carpeta bigdata:

```
bigdata@bigdata-VirtualBox:~$ hadoop fs -put palabras.txt /bigdata/palabras.txt
```

```
bigdata@bigdata-VirtualBox:~$ hadoop fs -ls /bigdata
Found 2 items
-rw-r--r--   1 bigdata supergroup         62 2025-07-29 02:24 /bigdata/palabras.txt
-rw-r--r--   1 bigdata supergroup       2513 2025-07-29 01:42 /bigdata/words-laodisea.txt
bigdata@bigdata-VirtualBox:~$
```

2. Mira los programas de ejemplo de MapReduce

```
bigdata@bigdata-VirtualBox:~$ hadoop jar /usr/jar/hadoop-examples-1.2.0.jar
An example program must be given as the first argument.
Valid program names are:
  aggregatewordcount: An Aggregate based map/reduce program that counts the words in the input files.
  aggregatewordhist: An Aggregate based map/reduce program that computes the histogram of the words in the input files.
  bbp: A map/reduce program that uses Bailey-Borwein-Plouffe to compute exact digits of Pi.
  dbcount: An example job that count the pageview counts from a database.
  distbbp: A map/reduce program that uses a BBP-type formula to compute exact bits of Pi.
  grep: A map/reduce program that counts the matches of a regex in the input.
  join: A job that effects a join over sorted, equally partitioned datasets
  multifilewc: A job that counts words from several files.
  pentomino: A map/reduce tile laying program to find solutions to pentomino problems.
  pi: A map/reduce program that estimates Pi using a quasi-Monte Carlo method.
  randomtextwriter: A map/reduce program that writes 10GB of random textual data per node.
  randomwriter: A map/reduce program that writes 10GB of random data per node.
  secondarysort: An example defining a secondary sort to the reduce.
  sort: A map/reduce program that sorts the data written by the random writer.
  sudoku: A sudoku solver.
  teragen: Generate data for the terasort
  terasort: Run the terasort
  teravalidate: Checking results of terasort
  wordcount: A map/reduce program that counts the words in the input files.
  wordmean: A map/reduce program that counts the average length of the words in the input files.
  wordmedian: A map/reduce program that counts the median length of the words in the input files.
  wordstandarddeviation: A map/reduce program that counts the standard deviation of the length of the words in the input files.
```

Es de interés en esta parte del curso realizar ejercicios con Mapreduce, por lo usaremos wordcount.

3. Verificar la existencia de los archivos de entrada:

```
bigdata@bigdata-VirtualBox:~$ hadoop fs -ls /bigdata
Found 2 items
-rw-r--r--   1 bigdata supergroup         62 2025-07-29 02:24 /bigdata/palabras.txt
-rw-r--r--   1 bigdata supergroup       2513 2025-07-29 01:42 /bigdata/words-laodisea.txt
bigdata@bigdata-VirtualBox:~$
```

4. Ver los argumentos de comandos de wordcount

```
bigdata@bigdata-VirtualBox:~$ hadoop jar /usr/jar/hadoop-examples-1.2.0.jar wordcount
Usage: wordcount <in> [<in>...] <out>
```

5. Ejecutar wordcount

```
bigdata@bigdata-VirtualBox:~$ hadoop jar /usr/jar/hadoop-examples-1.2.0.jar wordcount /bigdata/palabras.txt out
```

A medida que se ejecute mostrara mensajes en el terminal:

```
2025-07-29 02:36:39,570 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/bigdata
/.staging/job_1753762773560_0002
2025-07-29 02:36:40,021 INFO input.FileInputFormat: Total input files to process : 1
2025-07-29 02:36:40,191 INFO mapreduce.JobSubmitter: number of splits:1
2025-07-29 02:36:40,430 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1753762773560_0002
2025-07-29 02:36:40,431 INFO mapreduce.JobSubmitter: Executing with tokens: []
2025-07-29 02:36:40,734 INFO conf.Configuration: resource-types.xml not found
2025-07-29 02:36:40,735 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2025-07-29 02:36:41,199 INFO impl.YarnClientImpl: Submitted application application_1753762773560_0002
2025-07-29 02:36:41,316 INFO mapreduce.Job: The url to track the job: http://bigdata-VirtualBox:8088/proxy/application_17537627
73560_0002/
2025-07-29 02:36:41,317 INFO mapreduce.Job: Running job: job_1753762773560_0002
2025-07-29 02:36:51,852 INFO mapreduce.Job: Job job_1753762773560_0002 running in uber mode : false
2025-07-29 02:36:51,861 INFO mapreduce.Job:  map 0% reduce 0%
```

….

hasta terminar el proceso:

```
                Merged Map outputs=1
                GC time elapsed (ms)=196
                CPU time spent (ms)=890
                Physical memory (bytes) snapshot=365989888
                Virtual memory (bytes) snapshot=4978073600
                Total committed heap usage (bytes)=230821888
                Peak Map Physical memory (bytes)=234090496
                Peak Map Virtual memory (bytes)=2486403072
                Peak Reduce Physical memory (bytes)=131964928
                Peak Reduce Virtual memory (bytes)=2491670528
        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=62
        File Output Format Counters
                Bytes Written=76
bigdata@bigdata-VirtualBox:~$
```

6. Revisar el directorio de salida de wordcount

```
bigdata@bigdata-VirtualBox:~$ hadoop fs -ls
Found 1 items
drwxr-xr-x   - bigdata supergroup          0 2025-07-29 02:37 out
```

7. Inspeccionar dentro del directorio de salida

```
bigdata@bigdata-VirtualBox:~$ hadoop fs -ls out
Found 2 items
-rw-r--r--   1 bigdata supergroup          0 2025-07-29 02:37 out/_SUCCESS
-rw-r--r--   1 bigdata supergroup         76 2025-07-29 02:37 out/part-r-00000
bigdata@bigdata-VirtualBox:~$
```

El archivo part-r-00000 contiene el resultado de wordcount.

8. Copiar los resultados al sistema local de ficheros de la máquina virtual.

```
bigdata@bigdata-VirtualBox:~$ hadoop fs -copyToLocal out/part-r-00000 local.txt
```

para ver los resultados puede usar el comando cat:

```
bigdata@bigdata-VirtualBox:~$ cat local.txt
Olimpo, 1
Rey     1
Salve   1
Señor   1
de      1
del     1
los     1
salve   2
tiempos 1
todos   1
```

Observe que en este caso la palabra salve tiene dos ocurrencias, mientras que la palabra Salve tiene una ocurrencia. Esto debido a la diferencia entre mayúsculas y minúsculas.

Nota para la tarea:

Para el desarrollo de la tarea, observe que la carpeta de trabajo es /user/bigdata , por tanto en HDFS (hdfs fs -mkdir …) deberá crear primero la carpeta user y dentro de ella la carpeta bigdata