Big Data Aplicado

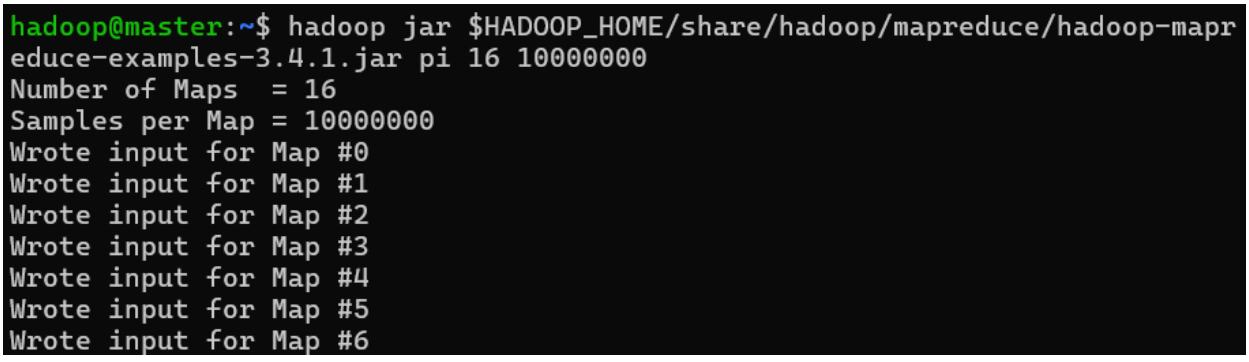
UD 4 - Apache Hadoop

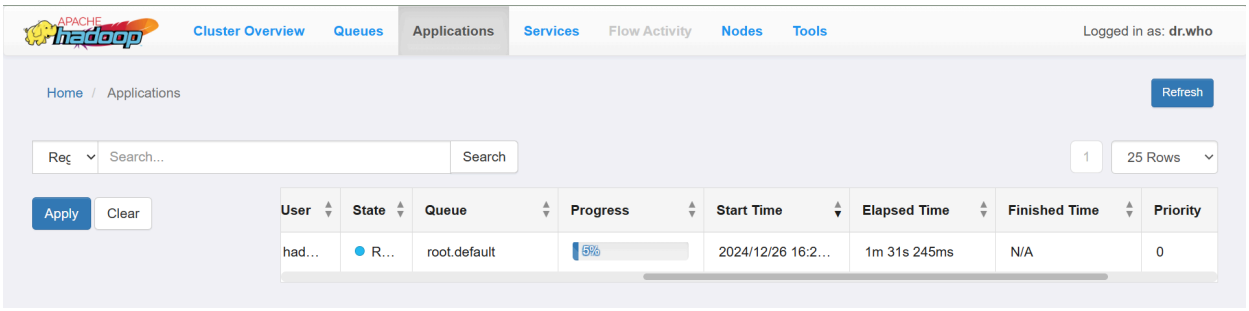
Práctica 2 MapReduce-Yarn

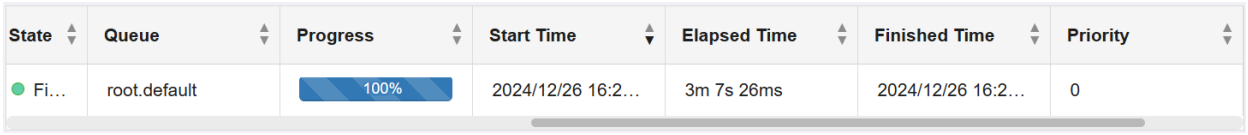
> Para los siguientes ejercicios, copia el comando y haz una captura de pantalla donde se muestre el resultado de cada acción Debes entregar los correspondientes comandos y capturas. Recuerda que tienes que tener correctamente configurado Apache Hadoop (HDFS, MapReduce y Yarn). Si no es así, consulta la documentación del módulo

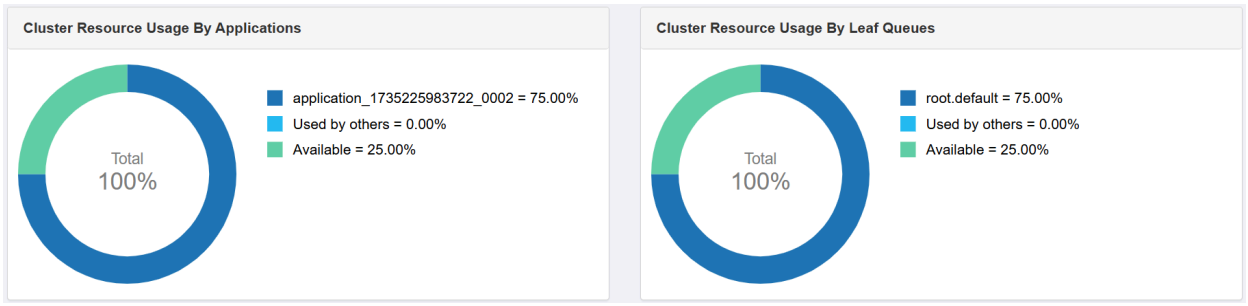
1. Calcula el del número π, con 16 maps y 10000000 muestras. *(RA5075.2 / CE.2b)*

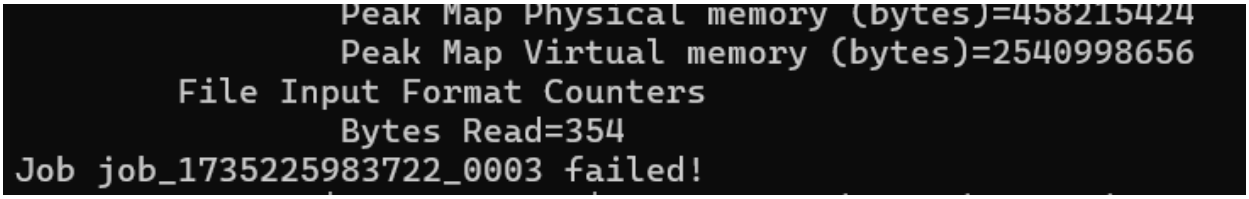
```
hadoop jar $HADOOP_HOME/share/hadoop/mapreduce/hadoop-mapreduce-examples-3.4.1.jar pi 16 10000000
```



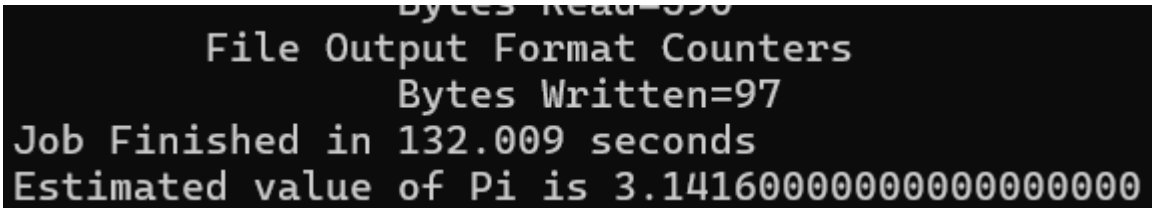








Debido a la falta de recursos, hemos tenido que reducir los parámetros a 5 Maps y 1000 Samples.



2. Cambia el cálculo con 4 maps y las mismas muestras *(RA5075.2 / CE.2b)*

```
hadoop jar $HADOOP_HOME/share/hadoop/mapreduce/hadoop-mapreduce-examples-3.4.1.jar pi 4 10000000
```

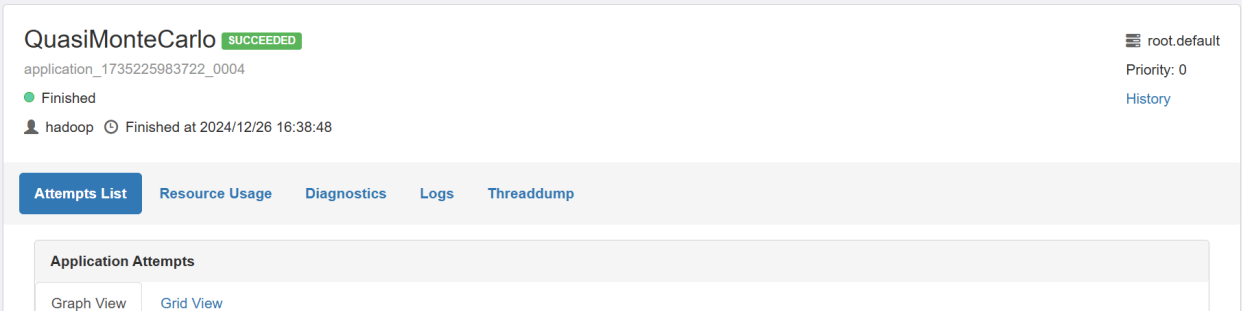
```
hadoop@master:~$ hadoop jar $HADOOP_HOME/share/hadoop/mapreduce/hadoop-mapr
educe-examples-3.4.1.jar pi 4 100000
Number of Maps  = 4
Samples per Map = 100000
Wrote input for Map #0
Wrote input for Map #1
Wrote input for Map #2
Wrote input for Map #3
Starting Job
2024-12-26 15:50:33,831 INFO client.DefaultNoHARMFailoverProxyProvider: Con
necting to ResourceManager at cluster-bda/192.168.18.8:8032
2024-12-26 15:50:36,775 INFO mapreduce.JobResourceUploader: Disabling Erasu
re Coding for path: /tmp/hadoop-yarn/staging/hadoop/.staging/job_1735225983
722_0005
2024-12-26 15:50:38,149 INFO input.FileInputFormat: Total input files to pr
```

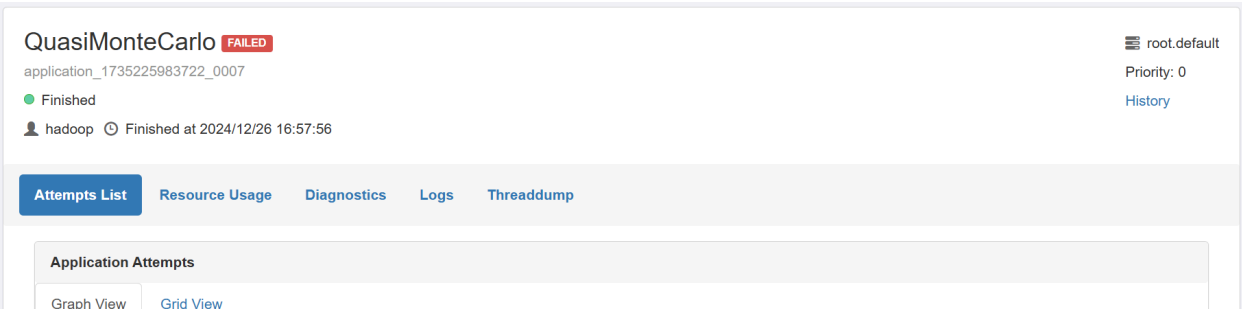
En este caso, se han bajado las muestras y no parece que de resultados solamente falla.

```
                Peak Map Virtual memory (bytes)=2545627136
        File Input Format Counters
                Bytes Read=354
Job job_1735225983722_0008 failed!
```

3. Compara los resultados estudiando cada "Jobs" en Yarn a través de su WebUI y analiza las diferencias. ¿Qué conclusiones obtienes? (Añade las correspondientes capturas) *(RA5075.4 / CE.4a)*

En la WebUI de Yarn vemos que la app que si funcionó indica su estado `SUCCEEDED` mientras que el resto pone `FAILED` . Y en los diagnósticos podemos ver por qué se originó el error.





Diagnostics

Task failed task_1735225983722_0007_m_000002
Job failed as tasks failed. failedMaps:1 failedReduces:0 killedMaps:0 killedReduces: 0

Outstanding Resource Requests

Priority	Resource Name	Capability	# Containers	Relax Locality	Node Label Expression
No data available!					

Scheduling Info

Allocated Resource	Running Containers	Preempted Resource	Num Non-AM container preempted	Num AM container preempted	Aggregated Resource Usage
0 MBs, 0 VCores	0	0 MBs, 0 VCores	0	0	543509 MBs, 399 VCores (× Secs)

4. Resuelve el siguiente sudoku *(RA5075.2 / CE.2b)*

```
8 5 ? 3 9 ? ? ? ?
? ? 2 ? ? ? ? ? ?
? ? 6 ? 1 ? ? ? 2
? ? 4 ? ? 3 ? 5 9
? ? 8 9 ? 1 4 ? ?
3 2 ? 4 ? ? 8 ? ?
9 ? ? ? 8 ? 5 ? ?
? ? ? ? ? ? 2 ? ?
? ? ? ? 4 5 ? 7 8
```

Guardamos el sudoku en un archivo llamado `sudoku.txt` y ejecutamos el comando:

```
hadoop jar $HADOOP_HOME/share/hadoop/mapreduce/hadoop-mapreduce-examples-3.4.1.jar sudoku sudoku.txt
```

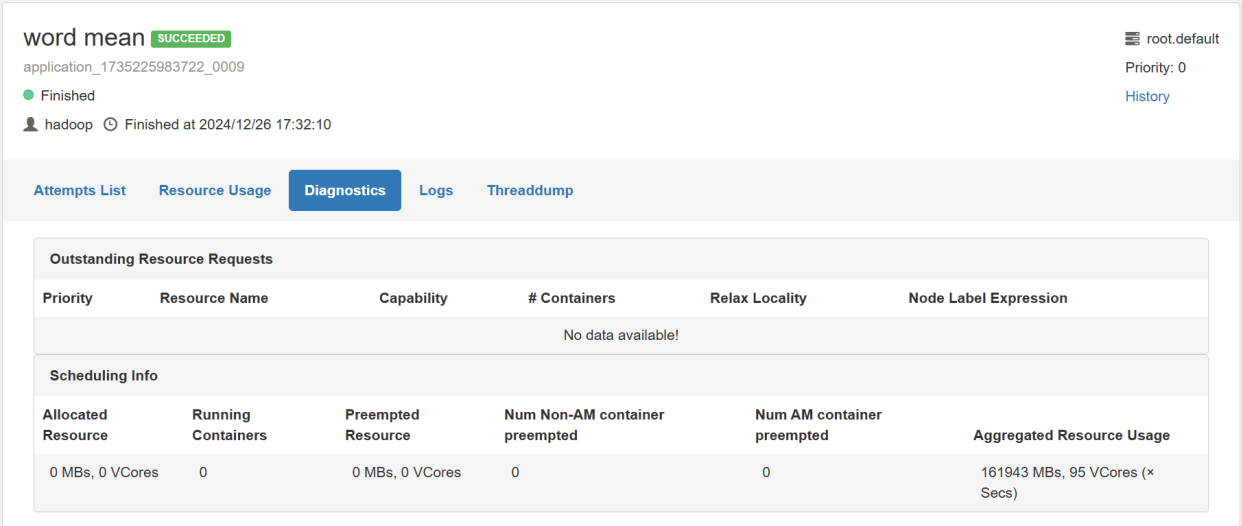
```
hadoop@master:~$ hadoop jar $HADOOP_HOME/share/hadoop/mapreduce/hadoop-mapr
educe-examples-3.4.1.jar sudoku sudoku.txt
Solving sudoku.txt
8 5 1 3 9 2 6 4 7
4 3 2 6 7 8 1 9 5
7 9 6 5 1 4 3 8 2
6 1 4 8 2 3 7 5 9
5 7 8 9 6 1 4 2 3
3 2 9 4 5 7 8 1 6
9 4 7 2 8 6 5 3 1
1 8 5 7 3 9 2 6 4
2 6 3 1 4 5 9 7 8

Found 1 solutions
hadoop@master:~$
```

5. Usando un fichero de texto de gran volumen, realiza el cálculo de la media, mediana y desviación estándar del tamaño de las palabras del texto. Compara los jobs de cada uno de ellos observando su coste en recursos, rendimiento y tiempo. Detalla la conclusión que puedes sacar de estos datos. *(RA5075.2 / CE.2b y RA5075.4 / CE.4a)*

```
hadoop jar $HADOOP_HOME/share/hadoop/mapreduce/hadoop-mapreduce-examples-3.4.1.jar wordmean /bda/mapreduce/ejercicios/El_Quijote.txt /bda/mapreduce/ejercicios/salida_quijote_media
```

```
hadoop@master:~$ hadoop jar $HADOOP_HOME/share/hadoop/mapreduce/hadoop-mapreduce-examples-3.4.1.jar wordmean /bda/mapreduce/ejercicios/El_Quijote.txt /bda/mapreduce/ejercicios/salida_quijote_media
2024-12-26 16:31:12,912 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at cluster-bda/192.168.18.8:8032
2024-12-26 16:31:13,763 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hadoop/.staging/job_1735225983722_0009
2024-12-26 16:31:14,654 INFO input.FileInputFormat: Total input files to process : 1
2024-12-26 16:31:15,082 INFO mapreduce.JobSubmitter: number of splits:1
2024-12-26 16:31:15,856 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1735225983722_0009
```

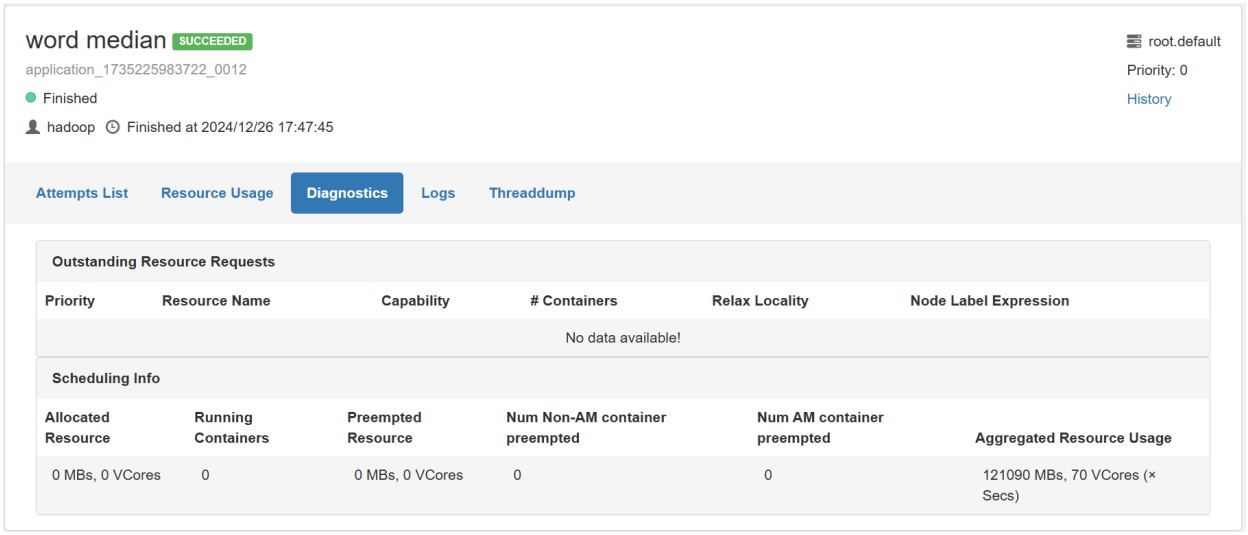
```
        File Input Format Counters
                Bytes Read=2161175
        File Output Format Counters
                Bytes Written=28
The mean is: 4.485917637108841
```




```
hadoop jar $HADOOP_HOME/share/hadoop/mapreduce/hadoop-mapreduce-examples-3.4.1.jar wordmedian /bda/mapreduce/ejercicios/El_Quijote.txt /bda/mapreduce/ejercicios/salida_quijote_mediana
```

```
hadoop@master:~$ hadoop jar $HADOOP_HOME/share/hadoop/mapreduce/hadoop-mapreduce-examples-3.4.1.jar wordmedian /bda/mapreduce/ejercicios/El_Quijote.txt /bda/mapreduce/ejercicios/salida_quijote_mediana
2024-12-26 16:35:07,981 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at cluster-bda/192.168.18.8:8032
2024-12-26 16:35:08,738 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hadoop/.staging/job_1735225983722_0010
2024-12-26 16:35:09,671 INFO input.FileInputFormat: Total input files to process : 1
2024-12-26 16:35:10,088 INFO mapreduce.JobSubmitter: number of splits:1
```

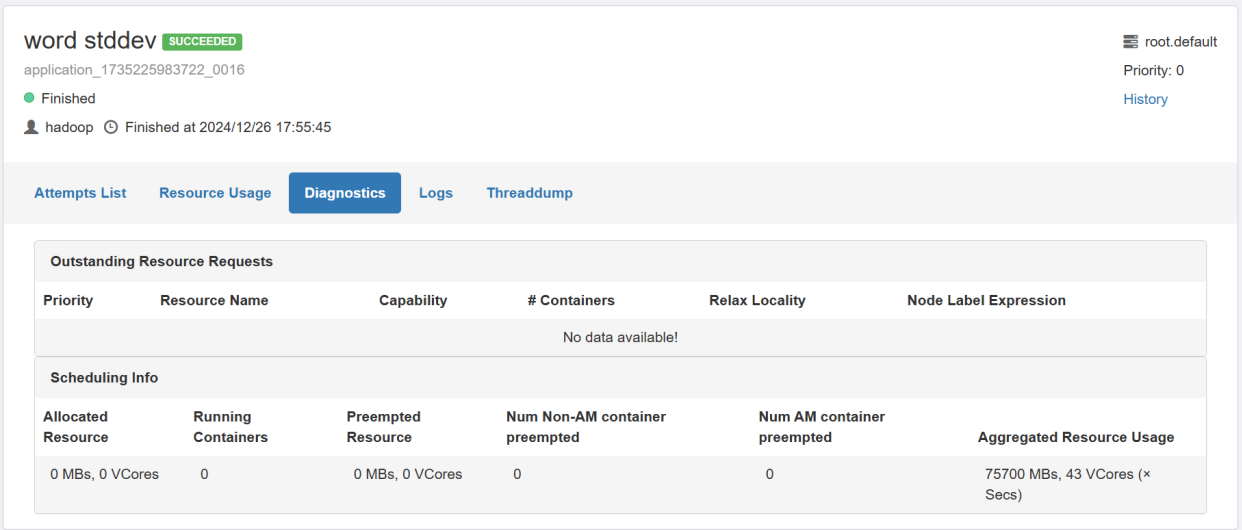
```
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=2161175
        File Output Format Counters
                Bytes Written=180
The median is: 4
```




```
hadoop jar $HADOOP_HOME/share/hadoop/mapreduce/hadoop-mapreduce-examples-3.4.1.jar wordstandarddeviation /bda/mapreduce/ejercicios/El_Quijote.txt /bda/mapreduce/ejercicios/salida_quijote_std
```

```
hadoop@master:~$ hadoop jar $HADOOP_HOME/share/hadoop/mapreduce/hadoop-mapr
educe-examples-3.4.1.jar wordstandarddeviation /bda/mapreduce/ejercicios/El
_Quijote.txt /bda/mapreduce/ejercicios/salida_quijote_std3
2024-12-26 16:55:19,368 INFO client.DefaultNoHARMFailoverProxyProvider: Con
necting to ResourceManager at cluster-bda/192.168.18.8:8032
2024-12-26 16:55:20,137 INFO mapreduce.JobResourceUploader: Disabling Erasu
re Coding for path: /tmp/hadoop-yarn/staging/hadoop/.staging/job_1735225983
722_0016
2024-12-26 16:55:20,707 INFO input.FileInputFormat: Total input files to pr
ocess : 1
2024-12-26 16:55:20,996 INFO mapreduce.JobSubmitter: number of splits:1
2024-12-26 16:55:21,282 INFO mapreduce.JobSubmitter: Submitting tokens for
job: job_1735225983722_0016
2024-12-26 16:55:21,282 INFO mapreduce.JobSubmitter: Executing with tokens:
```

```
        File Input Format Counters
                Bytes Read=2161175
        File Output Format Counters
                Bytes Written=44
The standard deviation is: 2.669806280162475
```



Podemos aquí el tiempo que han tardado cada aplicación:

Application Name	User	State	Queue	Progress	Start Time	Elapsed Time
word stddev	had…	F…	root.default	100%	2024/12/26 17:5…	24s 52ms
word stddev	had…	F…	root.default	100%	2024/12/26 17:5…	1m 20s 390ms
word stddev	had…	F…	root.default	100%	2024/12/26 17:5…	1m 5s 892ms
word stddev	had…	F…	root.default	100%	2024/12/26 17:5…	1m 7s 710ms
word median	had…	F…	root.default	100%	2024/12/26 17:4…	39s 627ms
word median	had…	F…	root.default	100%	2024/12/26 17:3…	1m 12s 245ms
word median	had…	F…	root.default	100%	2024/12/26 17:3…	1m 22s 191ms
word mean	had…	F…	root.default	100%	2024/12/26 17:3…	54s 129ms

Podemos apreciar que los resultados que han salido correctamente son los que han tardado menos tiempo en completarse. Podemos apreciar que el proceso más pesado es la desviación estandar y el que más VCores ha usado es la mediana.

6. Vamos a seguir utilizando los ejemplos que nos facilita MapReduce. Para ello vamos a usar un conjunto de 3 aplicaciones(teragen, terasort y teravalidate). Estos se basan en la implementación de MapReduce desarrolladas por Owen O'Malley y Arun Murthy. Estas aplicaciones ganaron el estándar de comparación anual de ordenación de terabytes de uso general ("Daytona") en 2009 con una velocidad de 0,578 TB/min (100 TB en 173 minutos). Para obtener más información sobre este y otros estándares de comparación de ordenación, consulte el sitio Sort Benchmark. ***(Opcional)***

Este ejemplo utiliza tres conjuntos de programas de MapReduce:

- TeraGen: programa de MapReduce que genera filas de datos que se van a ordenar.
- TeraSort: toma una muestra de los datos de entrada y usa MapReduce para ordenar los datos de manera absoluta.

TeraSort es una ordenación MapReduce estándar, salvo por el particionador personalizado. El particionador usa una lista ordenada de N-1 claves de muestra que definen el intervalo de claves para cada reducción. En concreto, todas las claves, como esa sample[i-1] <= key < sample[i] se envían para reducir i. Este particionador garantiza que las salidas de la reducción i sean todas menores que la salida de la reducción i+1.

- TeraValidate: programa de MapReduce que valida que la salida se ordene de manera global.

  Crea una asignación por archivo en el directorio de salida y cada asignación asegura que cada clave es menor o igual que la anterior. La función de asignación genera registros de la primera y última clave de cada archivo. La función de reducción se asegura de que la primera clave del archivo i es mayor que la última clave del archivo i-1. Los problemas se notifican como una salida de la fase de reducción con las claves que no están en orden.

7. Crea un fichero a través de la aplicación `Teragen` de 2GB *(Observa primero si tienes espacio suficiente en el cluster (al menos 10GB) http://bda-iesgrancapitan:9870/. Puedes hacerlo con menos para el ejercicio)*. El formato del fichero debe ser `Apellido1Nombre_teragen` *(RA5075.2 / CE.2b)* ***(Opcional)***

8. Muestra el fichero generado incluyendo permisos y tamaño en formato de GB. Indica que observas *(RA5075.2 / CE.2b)* ***(Opcional)***

9. Ejecuta la siguiente aplicación `terasort` para ordenar los datos generados anteriormente. *Ponte cómodo, dependiendo de tu máquina, tardará un ratito ;p*. Observa en la WebUI de Yarn el correspondiente *JOB* y en HDFS como se añaden los datos(bloques), y el espacio ocupado. El formato del fichero debe ser `Apellido1Nombre_terasort` *(RA5075.2 / CE.2b)* ***(Opcional)***

10. Una vez finalizado, observa y compara el resultado. El formato del fichero debe ser `Apellido1Nombre_teravalidate` *(RA5075.2 / CE.2b)* ***(Opcional)***

11. Finalmente, valida el resultado obtenido con `teravalidate` . El formato del fichero debe ser `Apellido1Nombre_teravalidate` *(RA5075.2 / CE.2b)* ***(Opcional)***

12. Busca el checksum resultante de la operación e indica cuál es y donde lo has encontrado *(RA5075.2 / CE.2b)* ***(Opcional)***

13. Investiga el ejemplo `pentomino` . Deberás explicar cómo funciona, realizar un ejemplo con él y explicarlo de forma detallada. Si lo prefieres, cambia la investigación del ejemplo del `pentomino` por cualquier otro de la lista de los ejemplos que proporciona hadoop *(RA5075.2 / CE.2b)*

Los pentominós son figuras planas formadas por la unión de cinco cuadrados iguales, donde cada cuadrado comparte al menos un lado con otro. Existen 12 formas diferentes de pentominós, cada una con su propia forma única.

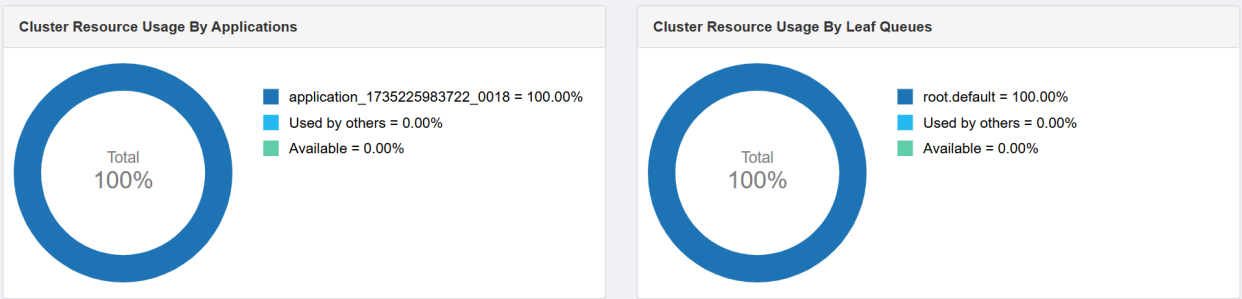
El problema de los pentominós consiste en utilizar todas las piezas de pentominó para cubrir una determinada área, como un tablero de ajedrez o un rectángulo, sin que las piezas se superpongan. También existen los problemas en tres dimensiones en donde se debe armar un cubo con las piezas.

De hecho, el programa de hadoop funciona con tres dimensiones.

```
hadoop jar $HADOOP_HOME/share/hadoop/mapreduce/hadoop-mapreduce-examples-3.4.1.jar pentomino /bda/pentomino 2 5 6
```

```
hadoop@master:~$ hadoop jar $HADOOP_HOME/share/hadoop/mapreduce/hadoop-mapr
educe-examples-3.4.1.jar pentomino /bda/pentomino2 -depth 2 -height 5 -widt
h 6
2024-12-26 17:31:57,267 INFO client.DefaultNoHARMFailoverProxyProvider: Con
necting to ResourceManager at cluster-bda/192.168.18.8:8032
2024-12-26 17:31:58,043 INFO mapreduce.JobResourceUploader: Disabling Erasu
re Coding for path: /tmp/hadoop-yarn/staging/hadoop/.staging/job_1735225983
722_0018
2024-12-26 17:31:58,495 INFO input.FileInputFormat: Total input files to pr
ocess : 1
2024-12-26 17:31:59,291 INFO mapreduce.JobSubmitter: number of splits:28
2024-12-26 17:31:59,617 INFO mapreduce.JobSubmitter: Submitting tokens for
job: job_1735225983722_0018
2024-12-26 17:31:59,618 INFO mapreduce.JobSubmitter: Executing with tokens:
 []
```

```
_0018_m_000021_2, Status : FAILED
2024-12-26 17:34:36,834 INFO mapreduce.Job:  map 39% reduce 0%
2024-12-26 17:34:40,809 INFO mapreduce.Job:  map 100% reduce 100%
2024-12-26 17:34:44,164 INFO mapreduce.Job: Job job_1735225983722_0018 fail
ed with state FAILED due to: Task failed task_1735225983722_0018_m_000008
Job failed as tasks failed. failedMaps:1 failedReduces:0 killedMaps:0 kille
dReduces: 0

2024-12-26 17:34:44,830 INFO mapreduce.Job: Counters: 44
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=1858200
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=843
                HDFS: Number of bytes written=0
                HDFS: Number of read operations=18
```



Solo ha salido un caso que es con altura 3 y anchura 20. Pero no tenemos nada en el archivo de salida.

```
hadoop@master:~$ hadoop jar $HADOOP_HOME/share/hadoop/mapreduce/hadoop-mapr
educe-examples-3.4.1.jar pentomino /bda/pentomino5 -depth 0 -height 3 -widt
h 20
2024-12-26 17:52:02,221 INFO client.DefaultNoHARMFailoverProxyProvider: Con
necting to ResourceManager at cluster-bda/192.168.18.8:8032
2024-12-26 17:52:02,798 INFO mapreduce.JobResourceUploader: Disabling Erasu
re Coding for path: /tmp/hadoop-yarn/staging/hadoop/.staging/job_1735225983
722_0021
2024-12-26 17:52:03,159 INFO input.FileInputFormat: Total input files to pr
ocess : 1
2024-12-26 17:52:03,469 INFO mapreduce.JobSubmitter: number of splits:1
2024-12-26 17:52:03,837 INFO mapreduce.JobSubmitter: Submitting tokens for
job: job_1735225983722_0021
2024-12-26 17:52:03,838 INFO mapreduce.JobSubmitter: Executing with tokens:
 []
2024-12-26 17:52:04,214 INFO conf.Configuration: resource-types.xml not fou
nd
```

Browse Directory