Máster en Big Data

Tecnologías de Almacenamiento

6. Hands-On: Ejecución Apache Spark

Presentado por:

José David Angulo y Albert Ripoll

Índice

1. Introducción

El objetivo de este Hands-On es aprender a ejecutar un job de Spark mediante un ejemplo precompilado.

2. Entorno

Para la realización de este hands-on, vamos a desplegar un sandbox de MapR en VMWare.

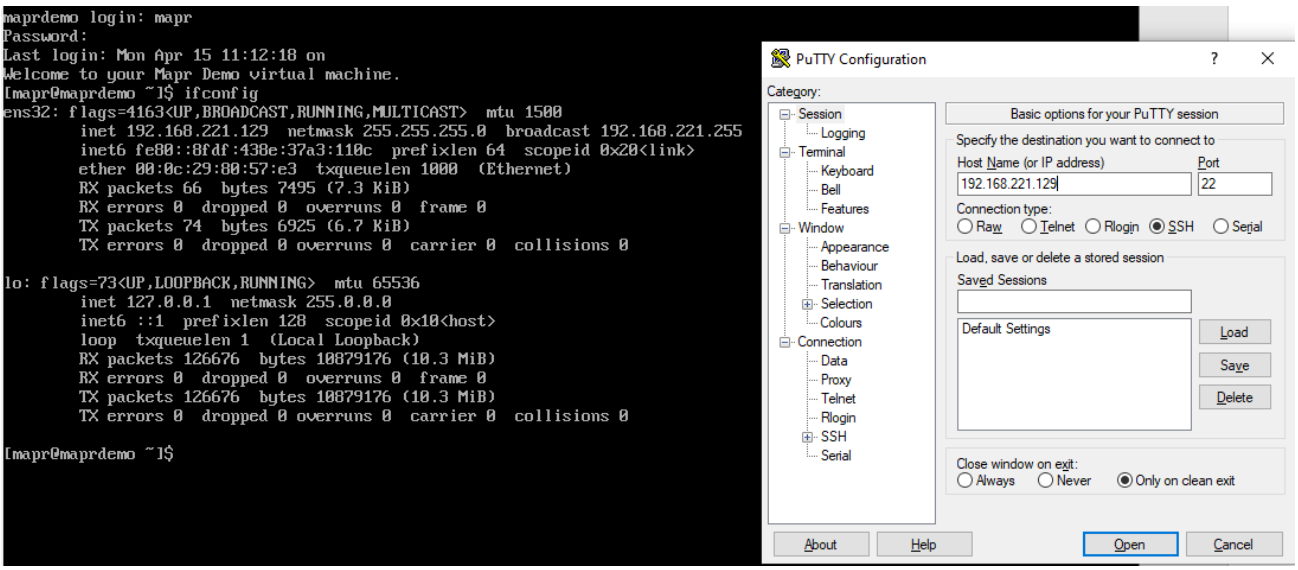
Una vez arrancado el nodo, las credenciales son:

Usuario: *mapr*; Password: *mapr*

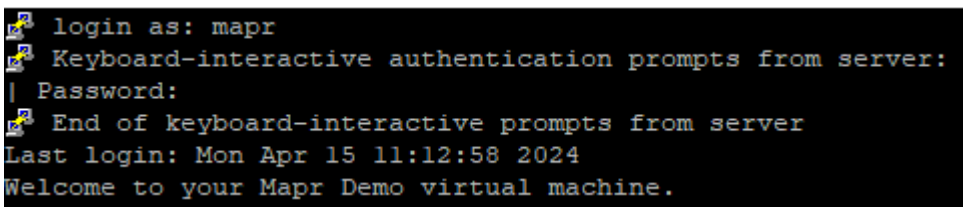
Busca la IP de la máquina virtual y accede a HUE http://<ip>:8888/ y abre una conexión ssh para más comodidad (puedes añadir la ip en hosts con el nombre *maprdemo*).

Usuario: *mapr*; Password: *mapr*

ifconfig



Después de ifconfig → la ip de inet la escribimos en host Name or IP address de PuTTY configuration.



Volvemos a poner el loging y pass mapr, mapr

3. Ejecución Spark

La ubicación de Spark dentro de la distribución MapR la podemos encontrar en:

/opt/mapr/spark/spark-2.1.0/

Se continua después del loging

cd /opt/mapr/spark/spark-2.1.0/

```
Last login: Mon Apr 15 11:12:58 2024
Welcome to your Mapr Demo virtual machine.
[mapr@maprdemo ~]$ /opt/mapr/spark/spark-2.1.0/
-bash: /opt/mapr/spark/spark-2.1.0/: Is a directory
[mapr@maprdemo ~]$ ls
[mapr@maprdemo ~]$ cd /opt/mapr/spark/spark-2.1.0/
[mapr@maprdemo spark-2.1.0]$
```

Se pide ejecutar el programa SparkPi, contra YARN en modo cliente, con una memoria por ejecutor de 1g.

El jar ya compilado lo podemos encontrar en: /opt/mapr/spark/spark-2.1.0/examples/jars/spark-examples_2.11-2.1.0-mapr-1710.jar

Y el package y nombre de la clase es: *org.apache.spark.examples.SparkPi*

El programa acepta por parámetro el número de paralelización que queremos darle, vamos a pasarle, por ejemplo, 4.

Engancha aquí el comando para enviar el Job y una captura de pantalla con la salida.

• En la diapositiva 36 del powerpoint "3. Apache spark" vemos:

```
./bin/spark-submit \
  --class org.apache.spark.examples.SparkPi \
  --master local[8] \
  /path/to/examples.jar \
  100
```

```
export HADOOP_CONF_DIR=XXX
./bin/spark-submit \
  --class org.apache.spark.examples.SparkPi \
  --master yarn \
  --deploy-mode cluster \  # can be client for client mode
  --executor-memory 20G \
  --num-executors 50 \
  /path/to/examples.jar \
  1000
```

• En la diapositiva 37 del powerpoint "3. Apache spark" vemos:

spark-sumit --jars lib/json-simple-1.1.1.jar -- class com.pragsis.ccma.hbbtv.HbbtcProcess

--master yarn-client --executor-memory 15H ccma.jar -i /user/devel/data/pending/ -o /user/devel/data/

• Por lo que para responder a las condiciones de la práctica se elabora el siguiente código para ejecutar.

./bin/spark-submit \

--class org.apache.spark.examples.SparkPi \
--master yarn-client \
/opt/mapr/spark/spark-2.1.0/examples/jars/spark-examples_2.11-2.1.0-mapr-1710.jar \
4

```
[mapr@maprdemo spark-2.1.0]$ ./bin/spark-submit \
> --class org.apache.spark.examples.SparkPi \
> --master yarn-client \
> /opt/mapr/spark/spark-2.1.0/examples/jars/spark-examples_2.11-2.1.0-mapr-1710.
jar \
> 4
Warning: Unable to determine $DRILL_HOME
Warning: Master yarn-client is deprecated since 2.0. Please use master "yarn" wi
th specified deploy mode instead.
24/04/15 11:39:59 WARN Utils: Your hostname, maprdemo.local resolves to a loopba
ck address: 127.0.0.1; using 192.168.221.129 instead (on interface ens32)
24/04/15 11:39:59 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another
address
24/04/15 11:40:01 WARN Client: Neither spark.yarn.jars nor spark.yarn.archive is
 set, falling back to uploading libraries under SPARK_HOME.
Pi is roughly 3.143077857694644
[mapr@maprdemo spark-2.1.0]$
```

Después se cuelga con la ip original que era 192.168.221.129 y se vuelve a repetir todo otra vez con lo que cambia la ip a 192.168.221.131 y sale

```
[mapr@maprdemo spark-2.1.0]$ ./bin/spark-submit \
> --class org.apache.spark.examples.SparkPi \
> --master yarn-client \
> /opt/mapr/spark/spark-2.1.0/examples/jars/spark-examples_2.11-2.1.0-mapr-1710.
jar \
> 4
Warning: Unable to determine $DRILL_HOME
Warning: Master yarn-client is deprecated since 2.0. Please use master "yarn" wi
th specified deploy mode instead.
24/04/15 12:16:00 WARN Utils: Your hostname, maprdemo.local resolves to a loopba
ck address: 127.0.0.1; using 192.168.221.131 instead (on interface ens32)
24/04/15 12:16:00 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another
address
24/04/15 12:16:02 WARN Client: Neither spark.yarn.jars nor spark.yarn.archive is
 set, falling back to uploading libraries under SPARK_HOME.
[Stage 0:>                                                          (0 + 0) / 4]
24/04/15 12:17:03 WARN YarnScheduler: Initial job has not accepted any resources
; check your cluster UI to ensure that workers are registered and have sufficien
t resources
Pi is roughly 3.142577856444641
[mapr@maprdemo spark-2.1.0]$
```

Después de colgarse otra vez se repite el proceso y la nueva ip es a 192.168.221.132

Se tiene que repetir para 2, 4 y 8 jobs.

4. SparkUI

Consulta la *SparkUI* después de ejecutar el Job y responde a las siguientes preguntas.

Requisito: No accedas directamente a la interficie del job history de spark, hazlo mediante HUE (acceso user:*mapr* / pass: *mapr*)

¿Cuántas tasks se han ejecutado?

¿Cuántos executors se han ejecutado?

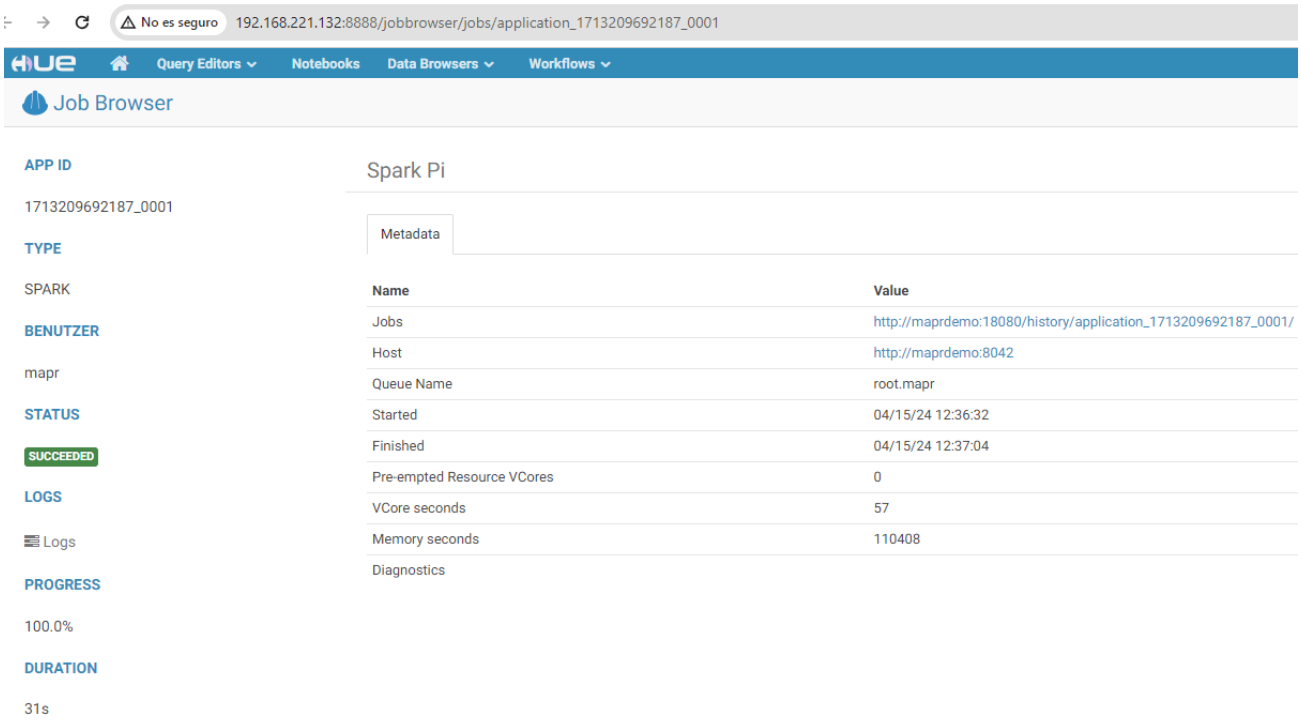
Para el job 4: ¿Cuántos RDDs se han creado y que transformaciones los han desencadenado?

Ejecuta el job con 2, 4 y 10 tasks y compara el rendimiento, ¿Que sucede en relación con la performance? ¿Porque se produce este comportamiento?

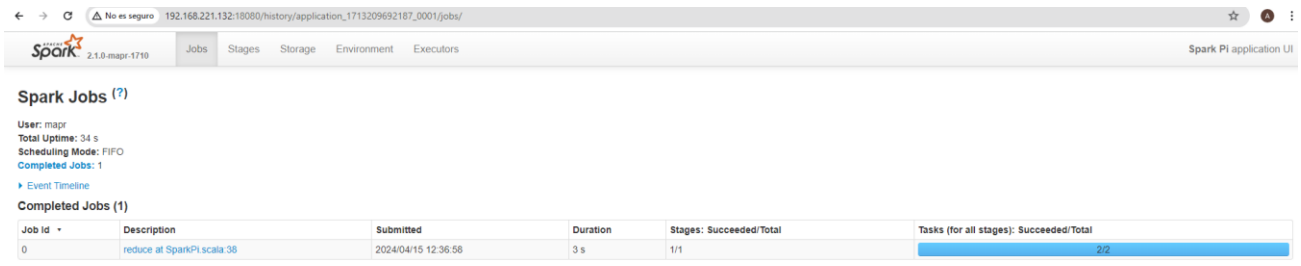
DETALLES CON Jobs = 2

```
./bin/spark-submit \
--class org.apache.spark.examples.SparkPi \
--master yarn-client \
/opt/mapr/spark/spark-2.1.0/examples/jars/spark-examples_2.11-2.1.0-mapr-1710.jar \
2
```

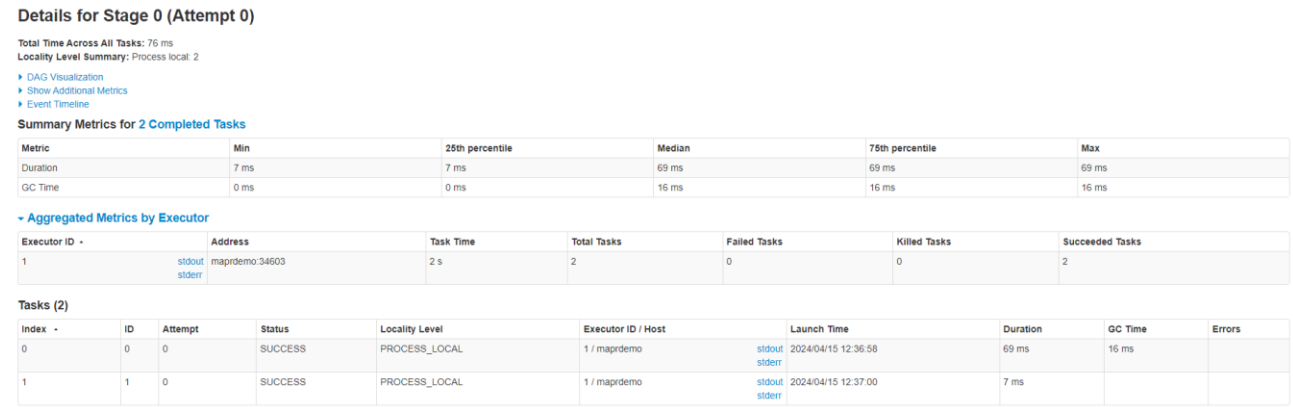
```
[mapr@maprdemo ~]$ cd /opt/mapr/spark/spark-2.1.0/
[mapr@maprdemo spark-2.1.0]$
[mapr@maprdemo spark-2.1.0]$ ./bin/spark-submit \
> --class org.apache.spark.examples.SparkPi \
> --master yarn-client \
> /opt/mapr/spark/spark-2.1.0/examples/jars/spark-examples_2.11-2.1.0-mapr-1710.
jar \
> 2
Warning: Unable to determine $DRILL_HOME
Warning: Master yarn-client is deprecated since 2.0. Please use master "yarn" wi
th specified deploy mode instead.
24/04/15 12:36:27 WARN Utils: Your hostname, maprdemo.local resolves to a loopba
ck address: 127.0.0.1; using 192.168.221.132 instead (on interface ens32)
24/04/15 12:36:27 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another
address
24/04/15 12:36:28 WARN Client: Neither spark.yarn.jars nor spark.yarn.archive is
 set, falling back to uploading libraries under SPARK_HOME.
Pi is roughly 3.1460957304786525
```

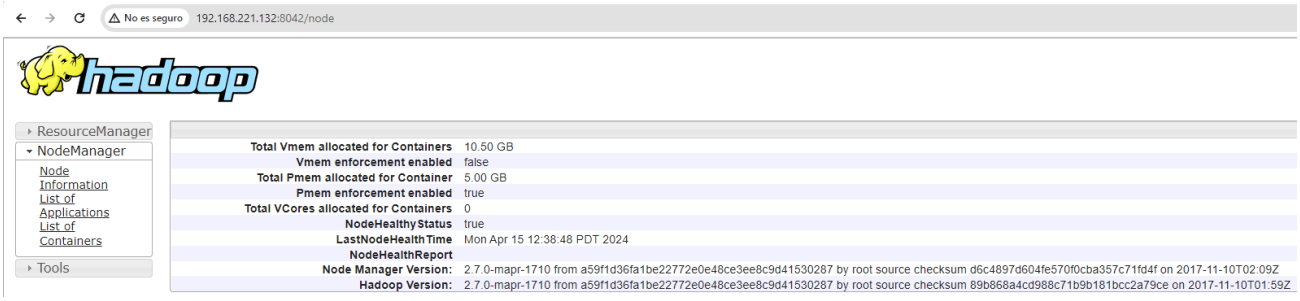


Para el job 2: ¿Cuántas tasks se han ejecutado? Dos

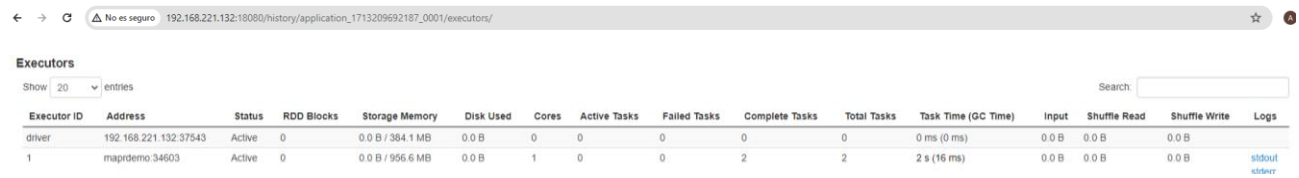


Al darle a reduce at SparkPi.scala:38

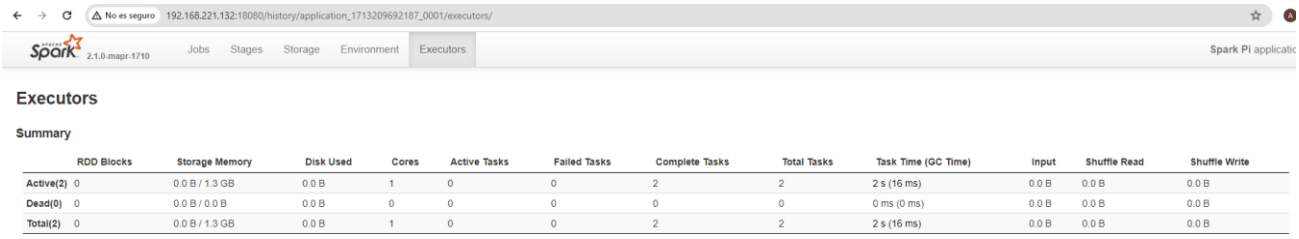




Para el job 2: ¿Cuántos executors se han ejecutado?



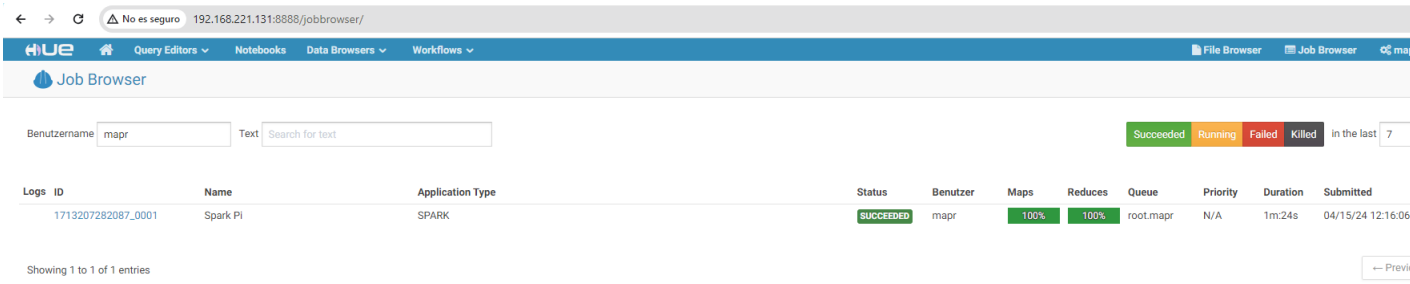
Para el job 2: ¿Cuántos RDDs se han creado y que transformaciones los han desencadenado?



DETALLES CON Jobs = 4

Entonces en el Google Chrome se escribe http://192.168.221.131:8888/

Y en job browser vemos el job.



En view job vemos:

APP ID

1713207282087_0001

TYPE

SPARK

BENUTZER

mapr

STATUS

SUCCEEDED

LOGS

Logs

PROGRESS

100.0%

DURATION

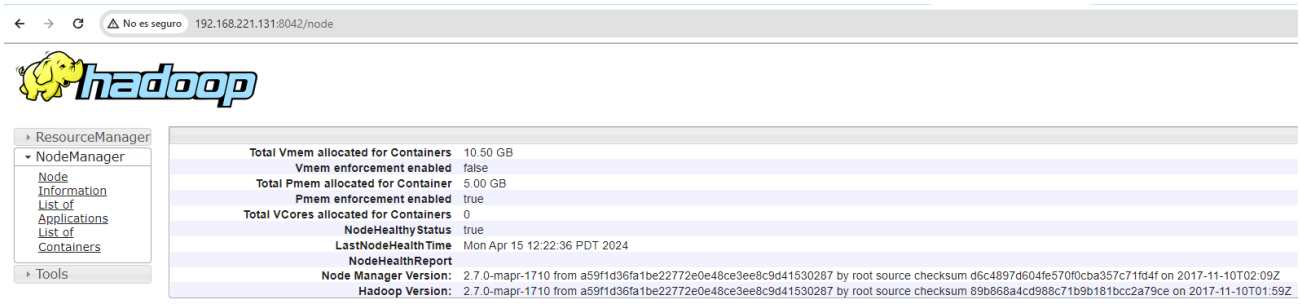
1m:24s

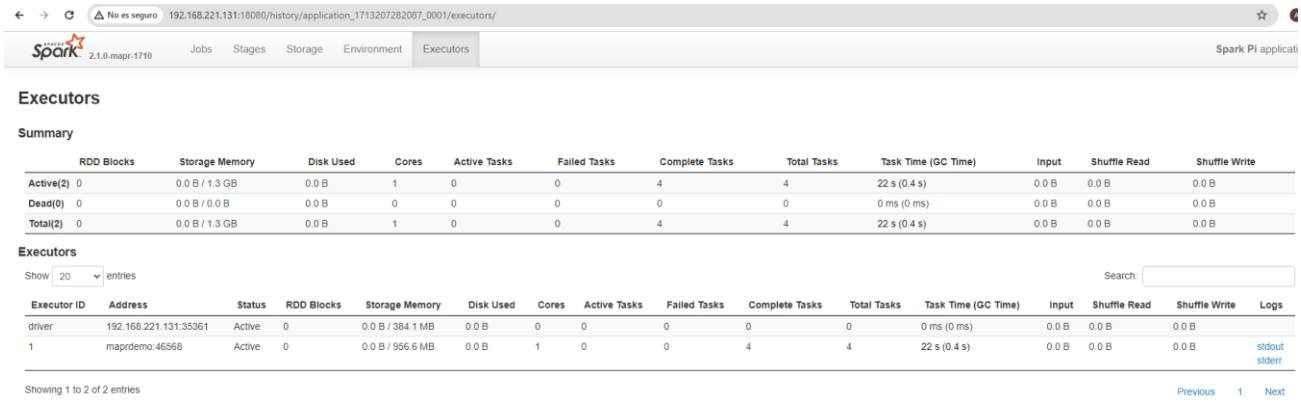
Spark Pi

Metadata

Name	Value
Jobs	http://maprdemo:18080/history/application_1713207282087_0001/
Host	http://maprdemo:8042
Queue Name	root.mapr
Started	04/15/24 12:16:06
Finished	04/15/24 12:17:31
Pre-empted Resource VCores	0
VCore seconds	133
Memory seconds	239991
Diagnostics	

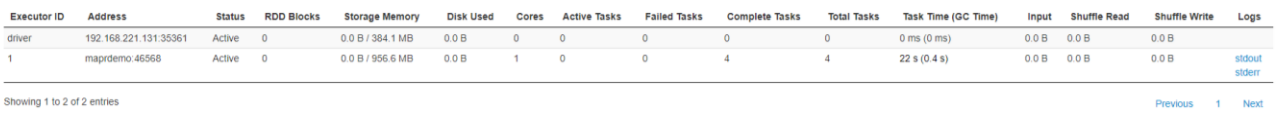
Y su acedemos a los jobs cambiando maprdemo con la ip.





Para el job 4: ¿Cuántas tasks se han ejecutado? Una

Para el job 4: ¿Cuántos executors se han ejecutado? Se han ejecutado dos. Un driver y un maprdemo



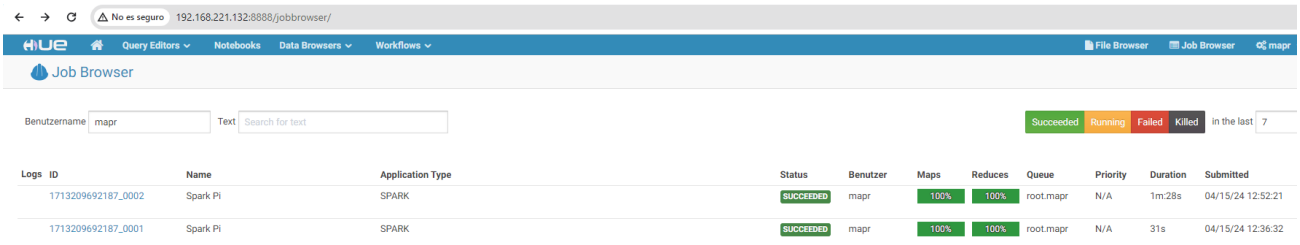
Para el job 4: ¿Cuántos RDDs se han creado y que transformaciones los han desencadenado? Se han creado dos que están en activo.

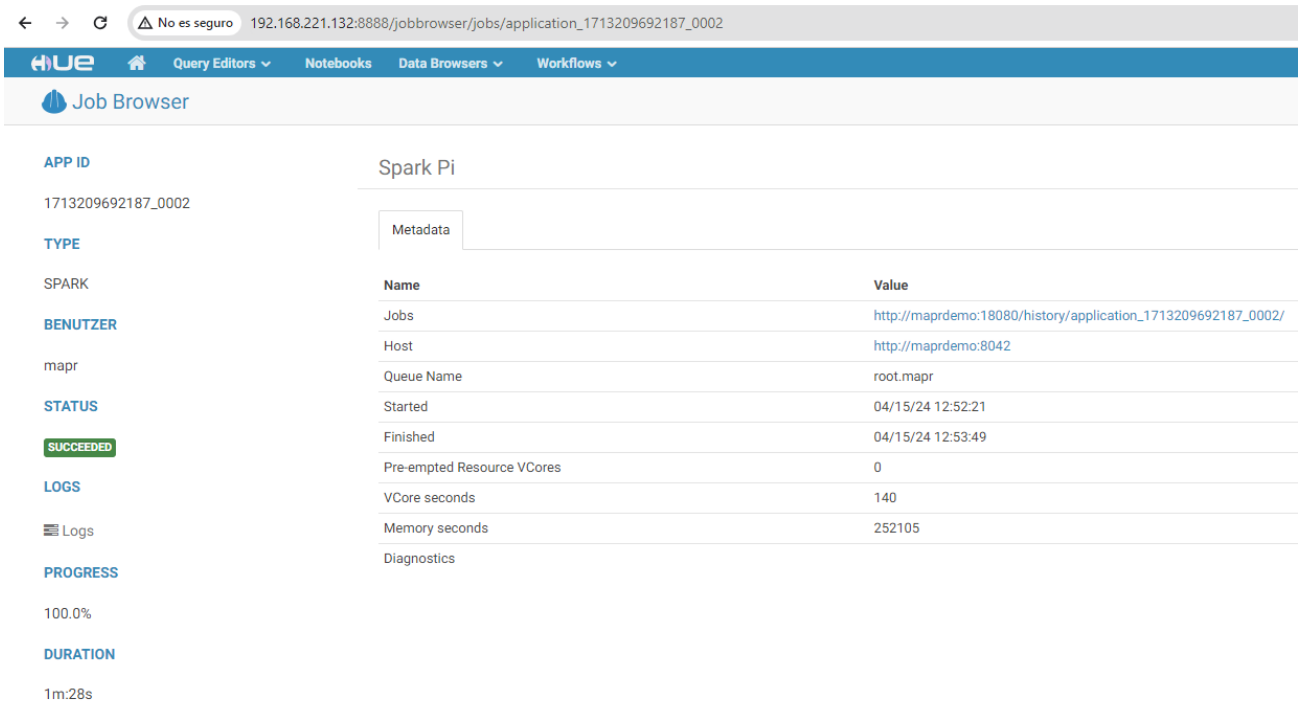
	RDD Blocks	Storage Memory	Disk Used	Cores	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time (GC Time)	Input	Shuffle Read	Shuffle Write
Active(2)	0	0.0 B / 1.3 GB	0.0 B	1	0	0	4	4	22 s (0.4 s)	0.0 B	0.0 B	0.0 B
Dead(0)	0	0.0 B / 0.0 B	0.0 B	0	0	0	0	0	0 ms (0 ms)	0.0 B	0.0 B	0.0 B
Total(2)	0	0.0 B / 1.3 GB	0.0 B	1	0	0	4	4	22 s (0.4 s)	0.0 B	0.0 B	0.0 B

DETALLES CON Jobs = 10

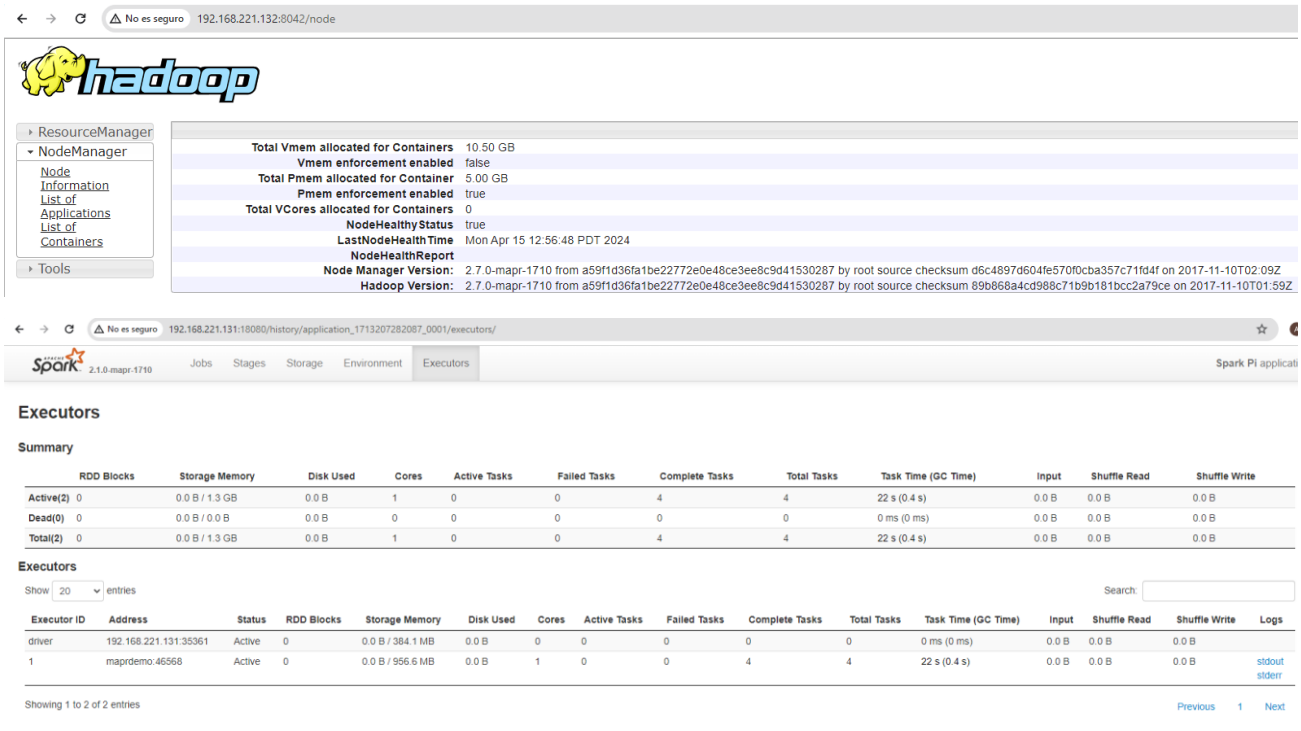
./bin/spark-submit \
--class org.apache.spark.examples.SparkPi \
--master yarn-client \
/opt/mapr/spark/spark-2.1.0/examples/jars/spark-examples_2.11-2.1.0-mapr-1710.jar \
10

```
[mapr@maprdemo spark-2.1.0]$ ./bin/spark-submit \
> --class org.apache.spark.examples.SparkPi \
> --master yarn-client \
> /opt/mapr/spark/spark-2.1.0/examples/jars/spark-examples_2.11-2.1.0-mapr-1710.
jar \
> 10
Warning: Unable to determine $DRILL_HOME
Warning: Master yarn-client is deprecated since 2.0. Please use master "yarn" wi
th specified deploy mode instead.
24/04/15 12:52:13 WARN Utils: Your hostname, maprdemo.local resolves to a loopba
ck address: 127.0.0.1; using 192.168.221.132 instead (on interface ens32)
24/04/15 12:52:13 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another
address
24/04/15 12:52:16 WARN Client: Neither spark.yarn.jars nor spark.yarn.archive is
 set, falling back to uploading libraries under SPARK_HOME.
[Stage 0:>                                                          (0 + 0) / 10]
24/04/15 12:53:28 WARN YarnScheduler: Initial job has not accepted any resources
; check your cluster UI to ensure that workers are registered and have sufficien
t resources
Pi is roughly 3.141339141339141
[mapr@maprdemo spark-2.1.0]$
```

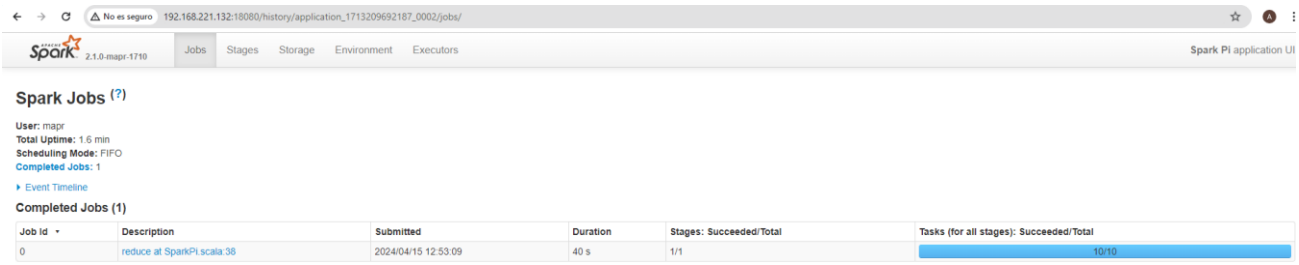




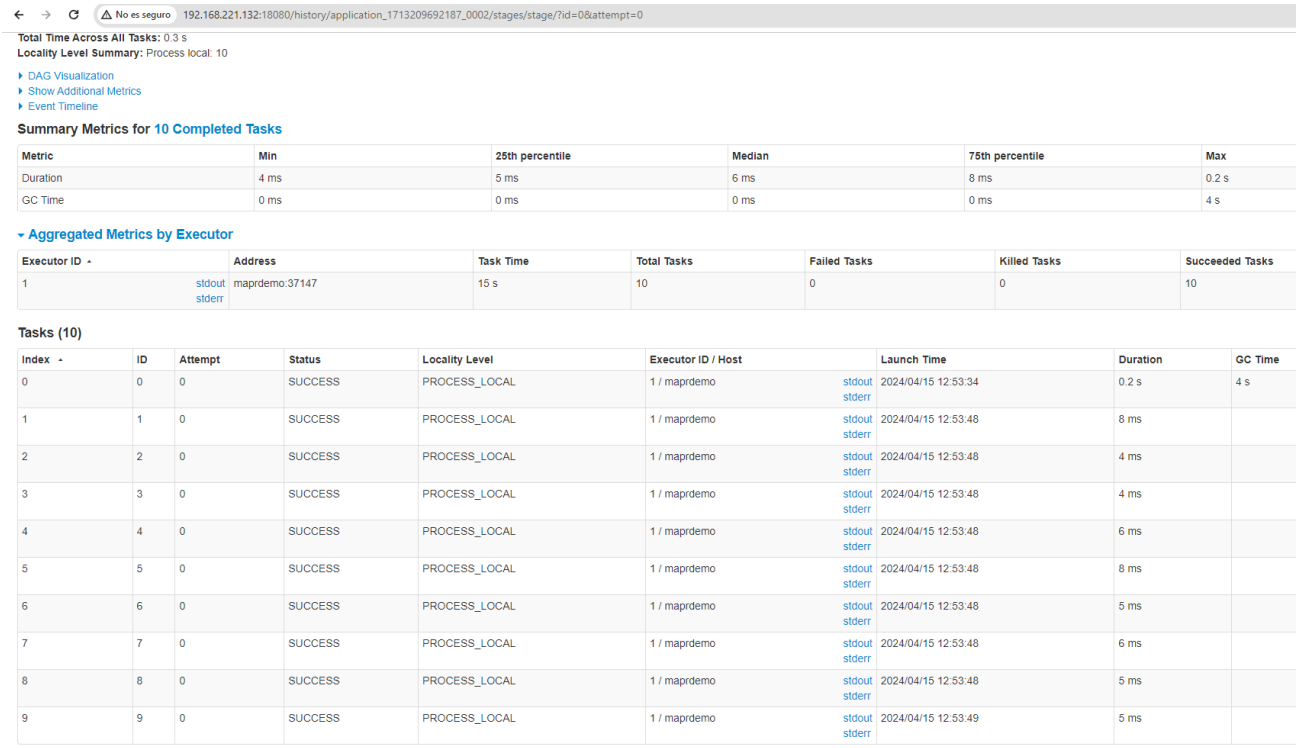
Y su acedemos a los jobs cambiando maprdemo con la ip.



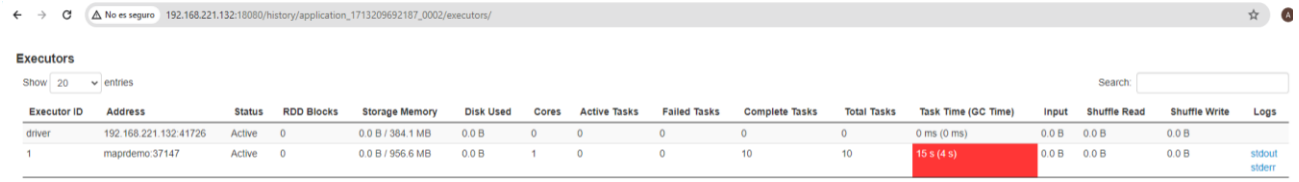
Para el job 10: ¿Cuántas tasks se han ejecutado? Diez



Detalles



Para el job 10: ¿Cuántos executors se han ejecutado?



Para el job 10: ¿Cuántos RDDs se han creado y que transformaciones los han desencadenado?