

TCDM Práctica 4

José Antonio

- Introducción

Si quisiéramos ejecutar pyspark aprovechando YARN para que los trabajos se distribuyan en los diferentes nodos del cluster habría que cambiar alguna configuración. Pero como nos piden ejecutar con este comando *spark-submit --master 'local[*]'* no hace falta configurar nada, vamos a ejecutar pyspark en local, sin cluster manager.

Lo único a tener en cuenta es que al tener una versión antigua de JAVA, tenemos que descargarnos una versión anterior de pyspark (la 3.5.3).

- Ejercicio 1

Se puede ver como se crean los ficheros de preguntas y respuestas

```
(.venv) luser@namenode:~/practica4$ vim ej1.py
(.venv) luser@namenode:~/practica4$ ls -ltr
total 230728
-rw-rw-r-- 1 luser luser 12057586 dic  6 12:51 Users.parquet
-rw-rw-r-- 1 luser luser 224198312 dic  6 12:51 Posts.parquet
-rw-rw-r-- 1 luser luser 5267 dic 18 13:30 ej1.py
(.venv) luser@namenode:~/practica4$ spark-submit --master 'local[*]' --num-executors 4
25/12/18 13:36:36 INFO SparkContext: Running Spark version 3.5.3
25/12/18 13:36:36 INFO SparkContext: OS info Linux, 6.6.87.2-microsoft-standard-WSL2, 
25/12/18 13:36:36 INFO SparkContext: Java version 1.8.0_462
25/12/18 13:36:36 WARN NativeCodeLoader: Unable to load native-hadoop library for your
25/12/18 13:36:36 INFO ResourceUtils: =====
25/12/18 13:36:36 INFO ResourceUtils: No custom resources configured for spark.driver.
25/12/18 13:36:36 INFO ResourceUtils: =====
25/12/18 13:36:36 INFO SparkContext: Submitted application: Ejercicio 1 de Diego
25/12/18 13:36:36 INFO ResourceProfile: Default ResourceProfile created, executor reso
mount: 0, script: , vendor: ), task resources: Map(cpus -> name: cpus, amount: 1.0)
25/12/18 13:36:37 INFO ResourceProfile: Limiting resource is cpu
25/12/18 13:36:37 INFO ResourceProfileManager: Added ResourceProfile id: 0
25/12/18 13:36:37 INFO SecurityManager: Changing view acls to: luser
25/12/18 13:36:37 INFO SecurityManager: Changing modify acls to: luser
25/12/18 13:36:37 INFO SecurityManager: Changing view acls groups to:
25/12/18 13:36:37 INFO SecurityManager: Changing modify acls groups to:
25/12/18 13:36:37 INFO SecurityManager: SecurityManager: authentication disabled; ui a
ssions: EMPTY
25/12/18 13:36:37 INFO Utils: Successfully started service 'sparkDriver' on port 42987
25/12/18 13:36:37 INFO SparkEnv: Registering MapOutputTracker
25/12/18 13:36:37 INFO SparkEnv: Registering BlockManagerMaster
25/12/18 13:36:37 INFO BlockManagerMasterEndpoint: Using org.apache.spark.storage.Defa
25/12/18 13:36:37 INFO BlockManagerMasterEndpoint: BlockManagerMasterEndpoint up
25/12/18 13:36:37 INFO SparkEnv: Registering BlockManagerMasterHeartbeat
25/12/18 13:36:37 INFO DiskBlockManager: Created local directory at /tmp/blockmgr-60c4
25/12/18 13:36:37 INFO MemoryStore: MemoryStore started with capacity 2.1 GiB
25/12/18 13:36:37 INFO SparkEnv: Registering OutputCommitCoordinator
25/12/18 13:36:37 INFO JettyUtils: Start Jetty 0.0.0.0:4040 for SparkUI
25/12/18 13:36:37 INFO Utils: Successfully started service 'SparkUI' on port 4040.
25/12/18 13:36:37 INFO Executor: Starting executor ID driver on host namenode
25/12/18 13:36:37 INFO Executor: OS info Linux, 6.6.87.2-microsoft-standard-WSL2, amd6
25/12/18 13:36:37 INFO Executor: Java version 1.8.0_462
25/12/18 13:36:37 INFO Executor: Starting executor with user classpath (userClassPathF
25/12/18 13:36:37 INFO Executor: Created or updated repl class loader org.apache.spark
25/12/18 13:36:38 INFO Utils: Successfully started service 'org.apache.spark.network.n
25/12/18 13:36:38 INFO NettyBlockTransferService: Server created on namenode:43235
25/12/18 13:36:38 INFO BlockManager: Using org.apache.spark.storage.RandomBlockReplica
25/12/18 13:36:38 INFO BlockManagerMaster: Registering BlockManager BlockManagerId(dri
25/12/18 13:36:38 INFO BlockManagerMasterEndpoint: Registering block manager namenode:
25/12/18 13:36:38 INFO BlockManagerMaster: Registered BlockManager BlockManagerId(driv
25/12/18 13:36:38 INFO BlockManager: Initialized BlockManager: BlockManagerId(driver,
(.venv) luser@namenode:~/practica4$ ls -ltr
total 233240
-rw-rw-r-- 1 luser luser 12057586 dic  6 12:51 Users.parquet
-rw-rw-r-- 1 luser luser 224198312 dic  6 12:51 Posts.parquet
-rw-rw-r-- 1 luser luser 5267 dic 18 13:30 ej1.py
-rw-r--r-- 1 luser luser 1154827 dic 18 13:36 dfRespuestas.parquet
-rw-r--r-- 1 luser luser 1416126 dic 18 13:36 dfPreguntas.parquet
(.venv) luser@namenode:~/practica4$ |
```

- Ejercicio 2

```
(.venv) luser@namenode:~/practica4$ ls -ltr
total 233244
-rw-rw-r-- 1 luser luser 12057586 dic  6 12:51 Users.parquet
-rw-rw-r-- 1 luser luser 224198312 dic  6 12:51 Posts.parquet
-rw-r--r-- 1 luser luser 1154827 dic 18 13:36 dfRespuestas.parquet
-rw-r--r-- 1 luser luser 1416126 dic 18 13:36 dfPreguntas.parquet
-rw-rw-r-- 1 luser luser 5459 dic 18 13:45 ej1.py
-rw-rw-r-- 1 luser luser 3402 dic 18 13:53 ej2.py
(.venv) luser@namenode:~/practica4$ spark-submit --master 'local[*]' --num
25/12/18 13:54:23 INFO SparkContext: Running Spark version 3.5.3
25/12/18 13:54:23 INFO SparkContext: OS info Linux, 6.6.87.2-microsoft-sta
25/12/18 13:54:23 INFO SparkContext: Java version 1.8.0_462
25/12/18 13:54:23 WARN NativeCodeLoader: Unable to load native-hadoop libr
25/12/18 13:54:23 INFO ResourceUtils: =====
25/12/18 13:54:23 INFO ResourceUtils: No custom resources configured for s
25/12/18 13:54:23 INFO ResourceUtils: =====
25/12/18 13:54:23 INFO SparkContext: Submitted application: Ejercicio 1 de
25/12/18 13:54:23 INFO ResourceProfile: Default ResourceProfile created, e
mount: 0, script: , vendor: ), task resources: Map(cpus -> name: cpus, amo
25/12/18 13:54:23 INFO ResourceProfile: Limiting resource is cpu
25/12/18 13:54:23 INFO ResourceProfileManager: Added ResourceProfile id: 0
25/12/18 13:54:23 INFO SecurityManager: Changing view acls to: luser
25/12/18 13:54:23 INFO SecurityManager: Changing modify acls to: luser
25/12/18 13:54:23 INFO SecurityManager: Changing view acls groups to:
25/12/18 13:54:23 INFO SecurityManager: Changing modify acls groups to:
25/12/18 13:54:23 INFO SecurityManager: SecurityManager: authentication di
ssions: EMPTY
25/12/18 13:54:23 INFO Utils: Successfully started service 'sparkDriver' o
25/12/18 13:54:23 INFO SparkEnv: Registering MapOutputTracker
25/12/18 13:54:24 INFO SparkEnv: Registering BlockManagerMaster
25/12/18 13:54:24 INFO BlockManagerMasterEndpoint: Using org.apache.spark.
25/12/18 13:54:24 INFO BlockManagerMasterEndpoint: BlockManagerMasterEndpo
25/12/18 13:54:24 INFO SparkEnv: Registering BlockManagerMasterHeartbeat
25/12/18 13:54:24 INFO DiskBlockManager: Created local directory at /tmp/b
25/12/18 13:54:24 INFO MemoryStore: MemoryStore started with capacity 2.1
25/12/18 13:54:24 INFO SparkEnv: Registering OutputCommitCoordinator
25/12/18 13:54:24 INFO JettyUtils: Start Jetty 0.0.0.0:4040 for SparkUI
25/12/18 13:54:24 INFO Utils: Successfully started service 'SparkUI' on po
25/12/18 13:54:24 INFO Executor: Starting executor ID driver on host namen
25/12/18 13:54:24 INFO Executor: OS info Linux, 6.6.87.2-microsoft-standar
25/12/18 13:54:24 INFO Executor: Java version 1.8.0_462
25/12/18 13:54:24 INFO Executor: Starting executor with user classpath (us
25/12/18 13:54:24 INFO Executor: Created or updated repl class loader org.
25/12/18 13:54:24 INFO Utils: Successfully started service 'org.apache.spa
25/12/18 13:54:24 INFO NettyBlockTransferService: Server created on nameno
25/12/18 13:54:24 INFO BlockManager: Using org.apache.spark.storage.Random
25/12/18 13:54:24 INFO BlockManagerMaster: Registering BlockManager BlockM
25/12/18 13:54:24 INFO BlockManagerMasterEndpoint: Registering block manag
25/12/18 13:54:24 INFO BlockManagerMaster: Registered BlockManager BlockMa
25/12/18 13:54:24 INFO BlockManager: Initialized BlockManager: BlockManage
(.venv) luser@namenode:~/practica4$ ls -ltr
total 237252
-rw-rw-r-- 1 luser luser 12057586 dic  6 12:51 Users.parquet
-rw-rw-r-- 1 luser luser 224198312 dic  6 12:51 Posts.parquet
-rw-r--r-- 1 luser luser 1154827 dic 18 13:36 dfRespuestas.parquet
-rw-r--r-- 1 luser luser 1416126 dic 18 13:36 dfPreguntas.parquet
-rw-rw-r-- 1 luser luser 5459 dic 18 13:45 ej1.py
-rw-rw-r-- 1 luser luser 3402 dic 18 13:53 ej2.py
-rw-r--r-- 1 luser luser 4101826 dic 18 13:54 resultado_e2.csv
```

```
(.venv) luser@namenode:~/practica4$ head resultado_e2.csv
Usuario,Año,NumPreguntas,TotalRespuestas,TotalComentarios,MediaRespuestas,MaxRespuestas
Ricardo,2016,7,12,15,1.71,4
Ignacio,2016,1,1,0,1.0,1
Norma.P,2016,1,1,5,1.0,1
Cris Valdez,2016,14,17,37,1.21,3
Franxo Bass,2016,2,2,2,1.0,1
edgargrs,2016,1,1,0,1.0,1
brahim,2016,1,2,8,2.0,2
Alberto Rodriguez,2016,1,1,1,1.0,1
Javier MCar Mcar,2016,1,0,4,0.0,0
```

- Ejercicio 3

```
(.venv) user@namenode:~/practica4$ vim ej3.py
(.venv) user@namenode:~/practica4$ spark-submit --master 'local[*]' --num-executors 4 --driver-memory 4g ej3.py dfRespuestas
.parquet dfPreguntas.parquet Posts.parquet lista,tags,separados resultado_e3.csv
25/12/18 14:38:11 INFO SparkContext: Running Spark version 3.5.3
25/12/18 14:38:11 INFO SparkContext: OS info Linux, 6.6.87.2-microsoft-standard-WSL2, amd64
25/12/18 14:38:11 INFO SparkContext: Java version 1.8.0_462
25/12/18 14:38:11 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes
where applicable
25/12/18 14:38:11 INFO ResourceUtils: =====
25/12/18 14:38:11 INFO ResourceUtils: No custom resources configured for spark.driver.
25/12/18 14:38:11 INFO ResourceUtils: =====
25/12/18 14:38:11 INFO SparkContext: Submitted application: Ejercicio 1 de Diego
25/12/18 14:38:11 INFO ResourceProfile: Default ResourceProfile created, executor resources: Map(cores -> name: cores, amount
: 1, script: , vendor: , memory -> name: memory, amount: 1024, script: , vendor: , offHeap -> name: offHeap, amount: 0, scrip
t: , vendor: ), task resources: Map(cpus -> name: cpus, amount: 1.0)
25/12/18 14:38:11 INFO ResourceProfile: Limiting resource is cpu
```

```
25/12/18 14:38:12 INFO Executor: Java version 1.8.0_462
25/12/18 14:38:12 INFO Executor: Starting executor with user classpath (userC
25/12/18 14:38:12 INFO Executor: Created or updated repl class loader org.apa
25/12/18 14:38:12 INFO Utils: Successfully started service 'org.apache.spark.
25/12/18 14:38:12 INFO NettyBlockTransferService: Server created on namenode:
25/12/18 14:38:12 INFO BlockManager: Using org.apache.spark.storage.RandomBlo
25/12/18 14:38:12 INFO BlockManagerMaster: Registering BlockManager BlockMana
25/12/18 14:38:12 INFO BlockManagerMasterEndpoint: Registering block manager
25/12/18 14:38:12 INFO BlockManagerMaster: Registered BlockManager BlockManag
25/12/18 14:38:12 INFO BlockManager: Initialized BlockManager: BlockManagerId
25/12/18 14:38:13 INFO SharedState: Setting hive.metastore.warehouse.dir ('nu
25/12/18 14:38:13 INFO SharedState: Warehouse path is 'file:/home/user/pract
(.venv) user@namenode:~/practica4$ ls -ltr
total 235508
-rw-rw-r-- 1 user user 12057586 dic  6 12:51 Users.parquet
-rw-rw-r-- 1 user user 224198312 dic  6 12:51 Posts.parquet
-rw-r--r-- 1 user user 1154827 dic 18 13:36 dfRespuestas.parquet
-rw-r--r-- 1 user user 1416126 dic 18 13:36 dfPreguntas.parquet
-rw-rw-r-- 1 user user 5459 dic 18 13:45 ej1.py
-rw-r--r-- 1 user user 2270822 dic 18 14:00 resultado_e2.csv
-rw-rw-r-- 1 user user 3513 dic 18 14:01 ej2.py
-rw-rw-r-- 1 user user 5918 dic 18 14:37 ej3.py
-rw-r--r-- 1 user user 35607 dic 18 14:38 resultado_e3.csv
(.venv) user@namenode:~/practica4$ head resultado_e3.csv
tag,Año,QuestionId,NRespuestas,Rango
lista,2015,1218,2,1
lista,2015,1245,2,2
lista,2015,24,1,3
lista,2016,36846,4,1
lista,2016,23335,3,2
lista,2016,27074,3,3
lista,2016,32867,3,4
lista,2016,39239,3,5
lista,2016,39726,3,6
(.venv) user@namenode:~/practica4$ |
```

- Ejercicio 4

```
(.venv) luser@namenode:~/practica4$ vim ej4.py
(.venv) luser@namenode:~/practica4$ spark-submit --master 'local[*]' --num-executors 4 --driver-memory 4g ej4.py dfPreguntas.
parquet Posts.parquet Users.parquet resultado_e4.csv
25/12/18 14:56:07 INFO SparkContext: Running Spark version 3.5.3
25/12/18 14:56:07 INFO SparkContext: OS info Linux, 6.6.87.2-microsoft-standard-WSL2, amd64
25/12/18 14:56:07 INFO SparkContext: Java version 1.8.0_462
25/12/18 14:56:07 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes
where applicable
25/12/18 14:56:07 INFO ResourceUtils: =====
25/12/18 14:56:07 INFO ResourceUtils: No custom resources configured for spark.driver.
25/12/18 14:56:07 INFO ResourceUtils: =====
25/12/18 14:56:07 INFO SparkContext: Submitted application: Ejercicio 1 de Diego
25/12/18 14:56:07 INFO ResourceProfile: Default ResourceProfile created, executor resources: Map(cores -> name: cores, amount
```

```
Usuario,Tag,Año,NPreguntas,Dif
Luxifrido Saturnino,c#,2023,1,0
Luxifrido Saturnino,consola,2023,1,0
Luxifrido Saturnino,parámetros,2023,1,0
Alejandro Jose,django,2020,1,0
Alejandro Jose,django-models,2020,1,0
Alejandro Jose,mysql,2020,1,0
Alejandro Jose,python,2020,1,0
Alejandro Jose,sql,2020,1,0
Cristhian Castro,imagen,2022,1,0
```