

# Practica 05.01 Apache Spark

## 1. Algoritmo de Euclides

### Enunciado:

Investiga el algoritmo de Euclides, explica para qué puede usarse en diferentes situaciones e impleméntalo en código de tal manera que desde Apache Spark se pueda ejecutar de manera distribuida. Describe un ejemplo de ejecución del algoritmo en notación matemática (Euclid).

El algoritmo de Euclides es un método clásico y muy eficiente para calcular el **máximo común divisor (MCD)** de dos números enteros positivos. Se basa en la siguiente propiedad fundamental:

**Propiedad.** Para  $a, b \in \mathbb{Z}^+$  con  $a \geq b$ :

$$\gcd(a, b) = \gcd(b, a \bmod b).$$

Iterando este paso (reemplazando  $(a, b)$  por  $(b, a \bmod b)$ ) hasta que el segundo término sea cero, el primer término restante es el MCD.

### ¿Para qué puede usarse?

- **Simplificación de fracciones.** Para reducir  $\frac{a}{b}$  a su forma irreducible dividiendo numerador y denominador por  $\gcd(a, b)$ .
- **Criptografía.** En algoritmos como RSA, para calcular inversos modulares y en pruebas de primalidad.
- **Ecuaciones diofánticas lineales.** Para resolver  $ax + by = c$  sólo si  $\gcd(a, b) \mid c$ .
- **Geometría computacional y gráficos.** Para determinar pasos enteros de un segmento de recta o en rasterización (algoritmo de Bresenham).
- **Teoría de números y análisis algorítmico.** Sirve de bloque básico en muchos teoremas y algoritmos que manipulan divisores.

### Ejemplo de ejecución en notación matemática

Calculemos  $\gcd(48, 18)$  paso a paso:

```
\begin{aligned}
48 &= 18 \times 2 \ ; \ + \ ; \ 12, & \quad & \gcd(48, 18) = \gcd(18, 12) \\
18 &= 12 \times 1 \ ; \ + \ ; \ 6, & \quad & \gcd(18, 12) = \gcd(12, 6) \\
12 &= 6 \times 2 \ ; \ + \ ; \ 0, & \quad & \gcd(12, 6) = 6 \\
\end{aligned}
```

Como el resto llega a cero, el **MCD** es el último divisor distinto de cero:

$$\gcd(48, 18) = 6.$$

## 2. Configuración y ejecución en Docker

A continuación se detallan los pasos para crear la red Docker, construir la imagen personalizada de Spark, levantar el contenedor maestro y ejecutar el script de Euclides. Cada comando incluye una breve explicación.

```
# Crea la red Docker para que Spark Master y posibles Workers puedan comunicarse.
docker network create spark-network

# Construye la imagen Docker con OpenJDK 17 y Spark 3.3.2, etiquetada como spark-custom:3.3.2.
docker build -t spark-custom:3.3.2 .

# Inicia el contenedor maestro en segundo plano, exponiendo puertos y montando el directorio de datos.
docker run -d --name spark-master-custom --network spark-network \
  -p 8080:8080 -p 7077:7077 \
  -v "${PWD}:/opt/spark/data" \
  spark-custom:3.3.2 tail -f /dev/null

# Arranca manualmente el servicio Spark Master dentro del contenedor.
docker exec -d spark-master-custom bash -c "/opt/spark/bin/spark-class \
  org.apache.spark.deploy.master.Master --host 0.0.0.0 --port 7077 \
  --webui-port 8080"

# Ejecuta el script Euclid en modo local[*], filtrando solo las líneas que empiezan con 'gcd'.
docker exec spark-master-custom bash -c "/opt/spark/bin/spark-submit \
  --master local[*] data/euclid.py 2>&1 | grep '^gcd'"
```

```
C:\Users\Miguel\Desktop\Tema 3 Práctica 5.1\spark-docker > docker network rm spark-network
spark-network
C:\Users\Miguel\Desktop\Tema 3 Práctica 5.1\spark-docker > docker network create spark-network
17ecc06ba915adef5d992acdfc0726ac613eae2b3608e53544b73a74d7eb809f
C:\Users\Miguel\Desktop\Tema 3 Práctica 5.1\spark-docker > docker build -t spark-custom:3.3.2 .
[+] Building 4.7s (8/8) FINISHED
=> [internal] load build definition from Dockerfile
=> => transferring dockerfile: 661B
=> [internal] load metadata for docker.io/library/openjdk:17-jdk-slim
=> [auth] library/openjdk:pull token for registry-1.docker.io
=> [internal] load .dockerignore
=> => transferring context: 2B
=> [1/3] FROM docker.io/library/openjdk:17-jdk-slim@sha256:aaa3b3cb27e3e520b8f116863d0580c438ed55ecfa0bc126b41f68c3f62f9774
=> => resolve docker.io/library/openjdk:17-jdk-slim@sha256:aaa3b3cb27e3e520b8f116863d0580c438ed55ecfa0bc126b41f68c3f62f9774
=> CACHED [2/3] RUN apt-get update && apt-get install -y wget procps python3 python3-pip && wget -qO /tmp/spark.tgz https://archive.apache.org/dist/spark/spark-3.3.2/spark-3.3.
=> CACHED [3/3] WORKDIR /opt/spark
=> => exporting layers
=> => exporting manifest sha256:c0ecb39cb6e5d1305fc7feda8da8b93cc53a14f8aa04e35b25e8d1d754b09c2c
=> => exporting config sha256:d6ca1d00f3712d3406824f765447d25809e80d057c25dc250d5ec68cc7575212
=> => exporting attestation manifest sha256:608c8de4d7f42c61e7a973dc37af966c828f4bbb4886704e833a12446539ce6b
=> => exporting manifest list sha256:7d991acc43dbde35ff59356c088cd1c531fdb1ee0340d150478a403172125f57
=> => naming to docker.io/library/spark-custom:3.3.2
=> => unpacking to docker.io/library/spark-custom:3.3.2

View build details: docker-desktop://dashboard/build/desktop-linux/desktop-linux/ddrro41gor7otugc3ssfbf31l
C:\Users\Miguel\Desktop\Tema 3 Práctica 5.1\spark-docker > docker run -d --name spark-master-custom --network spark-network -p 8080:8080 -p 7077:7077 -v "${PWD}:/opt/spark/data" spark-cust
om:3.3.2 tail -f /dev/null
bf6f0ce27df908cf2c3ed09c2db64bd00f23f9946c1558fe19c8e460b02a8e
C:\Users\Miguel\Desktop\Tema 3 Práctica 5.1\spark-docker > docker exec -d spark-master-custom bash -c "/opt/spark/bin/spark-class org.apache.spark.deploy.master.Master --host 0.0.0.0 --por
t 7077 --webui-port 8080"
C:\Users\Miguel\Desktop\Tema 3 Práctica 5.1\spark-docker > docker exec spark-master-custom bash -c "/opt/spark/bin/spark-submit --master local[*] data/euclid.py 2>&1 | grep '^gcd'"
gcd(1193, 661) = 1
gcd(3408, 824291) = 1
gcd(1193, 3408) = 1
C:\Users\Miguel\Desktop\Tema 3 Práctica 5.1\spark-docker > |
```

```
C:\Users\Miguel\Desktop\T
gcd(1193, 661) = 1
gcd(3408, 824291) = 1
gcd(1193, 3408) = 1
C:\Users\Miguel\Desktop\T
```

## 3. Cálculo de notas medias

### Enunciado:

Teniendo como entrada el fichero `ratings.txt`, crear dos scripts en Python que se ejecuten en Spark para obtener:

- **movies\_each\_average**: nota media por película.
- **average\_movie\_mark**: películas con nota media > 3.

### 3.1 Ejecución de `movies_each_average`

```
# Calcula la nota media de todas las votaciones de cada película.  
docker exec spark-master-custom bash -c "/opt/spark/bin/spark-submit \  
--master local[*] data/movies_each_average.py 2>/dev/null"
```

1444	3.29
1850	2.78
769	3.25
985	3.00
1675	3.29
2127	3.50
2831	3.00
1121	3.33
1181	2.33
2562	3.50
3644	3.57
401	3.00
2221	3.17
3131	3.50
3228	1.00
864	3.50
3294	2.50
2632	3.00
3888	4.00
2619	3.00
2563	3.00
3905	3.33
1548	2.00
1878	2.00
666	2.50
1360	2.00
696	3.25
3601	4.00
3322	2.75
3887	3.00
789	3.00
3216	3.33
2277	2.00
1630	2.00
977	2.00
3621	1.00
3493	1.00
3641	2.33
1115	1.00
1118	2.00
601	3.00
651	2.00
3876	2.00
1510	2.00
1386	1.00
859	2.00
1553	2.00
3166	1.00
142	1.00

## 3.2 Ejecución de `average_movie_mark`

```
# Lista solo las películas cuya nota media supera el umbral de 3.0.  
docker exec spark-master-custom bash -c "/opt/spark/bin/spark-submit \  
--master local[*] data/average_movie_mark.py 2>/dev/null"
```

3847	3.21
563	3.44
729	3.67
758	4.00
2441	3.25
245	3.10
1116	3.33
90	3.25
567	3.25
2209	3.80
3003	3.17
967	3.44
1444	3.29
769	3.25
1675	3.29
2127	3.50
1121	3.33
2562	3.50
3644	3.57
2221	3.17
3131	3.50
864	3.50
3888	4.00
3905	3.33
696	3.25
3601	4.00
3216	3.33
1685	3.25
2299	4.00
1534	4.00
2444	5.00
398	4.00
989	5.00
732	3.33
3057	4.00
1915	4.00
3242	3.67
1851	4.00
3881	5.00
3443	3.67
1470	4.00
981	5.00
3381	4.00
3233	5.00
3172	5.00
1139	4.00
2623	3.50
396	4.00

---

## 4. DataFrame MapReduce

### Enunciado:

Realiza el ejercicio de MapReduce de la práctica anterior usando DataFrames de Spark.

```
# Ejecuta el script que implementa el ejercicio de MapReduce con DataFrames.  
docker exec spark-master-custom bash -c '/opt/spark/bin/spark-submit \  
--master local[*] data/mapreduce.py'
```

```
≡≡≡ Clasificación de temperaturas ≡≡≡  
+-----+-----+  
|category|count|  
+-----+-----+  
|cold    |809  |  
|hot     |2    |  
|neutral |950  |  
+-----+-----+
```

---

## 5. RDD sobre películas

### Enunciado:

Crea un RDD con la tabla de películas y muestra las que tienen 4 o más vocales en su nombre.

```
# Filtra títulos con al menos 4 vocales mediante la API RDD.  
docker exec spark-master-custom bash -c "/opt/spark/bin/spark-submit \  
--master local[*] data/movies_rdd.py"
```

978 Amadeus Night  
979 It's Only the End of the World  
980 The Skin I Live In  
981 Miracles from Heaven  
983 Across the Universe  
986 Your Highness  
987 Final Destination 5  
988 Endless Love  
991 Underworld: Rise of the Lycans  
992 Taare Zameen Par  
993 Take Me Home Tonight  
994 Resident Evil: Afterlife  
996 Secret in Their Eyes  
997 Hostel: Part II  
998 Step Up 2: The Streets  
1000 Nine Lives

11 Fantastic Beasts and Where to Find Them  
25 Independence Day: Resurgence  
35 Resident Evil: The Final Chapter  
36 Captain America: Civil War  
46 Pirates of the Caribbean: On Stranger Tides  
51 Star Wars: Episode VII - The Force Awakens  
61 Batman v Superman: Dawn of Justice  
76 Pirates of the Caribbean: At World's End  
79 Pirates of the Caribbean: Dead Man's Chest  
115 Harry Potter and the Deathly Hallows: Part 2  
121 Miss Peregrine's Home for Peculiar Children  
123 Mike and Dave Need Wedding Dates  
127 Transformers: Age of Extinction  
156 Aliens vs Predator - Requiem  
196 Captain America: The First Avenger  
217 Captain America: The Winter Soldier  
234 Mission: Impossible - Rogue Nation  
245 Percy Jackson & the Olympians: The Lightning Thief  
258 Rise of the Planet of the Apes  
260 Norman: The Moderate Rise and Tragic Fall of a New York Fi  
261 Birdman or (The Unexpected Virtue of Ignorance)  
265 Teenage Mutant Ninja Turtles: Out of the Shadows  
272 The Hobbit: An Unexpected Journey  
297 Ice Age: Collision Course  
307 Alice Through the Looking Glass  
309 The Fast and the Furious: Tokyo Drift  
314 The Hobbit: The Battle of the Five Armies  
315 Harry Potter and the Order of the Phoenix  
367 The Twilight Saga: Breaking Dawn - Part 2  
419 Harry Potter and the Deathly Hallows: Part 1  
423 Pride and Prejudice and Zombies  
426 The Curious Case of Benjamin Button  
434 Mission: Impossible - Ghost Protocol  
466 Sherlock Holmes: A Game of Shadows  
474 Harry Potter and the Half-Blood Prince  
519 The Hobbit: The Desolation of Smaug  
584 What to Expect When You're Expecting  
614 The Assassination of Jesse James by the Coward Robert Ford  
630 Night at the Museum: Secret of the Tomb  
651 Sweeney Todd: The Demon Barber of Fleet Street  
658 Teenage Mutant Ninja Turtles  
682 We Need to Talk About Kevin  
684 The Mortal Instruments: City of Bones  
691 Resident Evil: Retribution  
711 Transformers: Revenge of the Fallen  
740 The Mummy: Tomb of the Dragon Emperor  
746 A Million Ways to Die in the West  
749 Ouija: Origin of Evil  
768 Indiana Jones and the Kingdom of the Crystal Skull  
807 Prince of Persia: The Sands of Time  
810 USS Indianapolis: Men of Courage

---

## 6. RDD de El Quijote

### Enunciado:

Procesa el texto de El Quijote con RDDs para:

1. Crear una lista con todas las palabras.
2. Contar cuántas veces aparece una palabra elegida.



3. Ordenar y guardar el recuento de todas las palabras en HDFS.

## 6.1 Listado de palabras

```
# Genera la lista completa de palabras en el documento.
```

```
docker exec spark-master-custom bash -c "/opt/spark/bin/spark-submit \
--master local[*] data/quijote_all_words_list.py"
```

```
canal, y, vámonos, al, campo, vestidos, de, pastores, como, tenemos, concertado, quizá, tras, de, alguna, mata, hallaremos, a, la, señora, doña, dulcinea, desecnan
tada, que, no, haya, más, que, ver, si, es, que, se, muere, de, pesar, de, verse, vencido, écheme, a, mi, la, culpa, diciendo, que, por, haber, yo, cinchad
o, mal, a, rocante, le, derribaron, cuanto, más, que, vuestra, merced, habrá, visto, en, sus, libros, de, caballerías, ser, cosa, ordinaria, derribarse, unos, ca
balleros, a, otros, y, el, que, es, vencido, hoy, ser, vencedor, mañana, así, es, dijo, sánson, y, el, buen, sancho, panza, está, muy, en, la, verdad, destos
casos, señores, dijo, don, quijote, vámonos, poco, a, poco, pues, ya, en, los, nidos, de, antaño, no, hay, pájaros, hogano, yo, fui, loco, y, ya, soy, cu
erdo, fui, don, quijote, de, la, mancha, y, soy, agora, como, he, dicho, alonso, quijano, el, bueno, pueda, con, vuestras, mercedes, mi, arrepentimiento, y, mi,
verdad, volverme, a, la, estimación, que, de, mi, se, tenía, y, prosiga, adelante, el, señor, escribano, ítem, mando, toda, mi, hacienda, a, puerta, cerrada, a
, antonia, quijana, mi, sobrina, que, está, presente, habiendo, sacado, primero, de, lo, más, bien, parado, della, lo, que, fuere, menester, para, cumplir, las,
mandas, que, de, hechas, y, la, primera, satisfacción, que, se, haga, quiero, que, sea, pagar, el, salario, que, debo, del, tiempo, que, mi, ama, me, ha, se
rido, y, más, veinte, ducados, para, un, vestido, de,jo, por, mis, albacas, al, señor, cura, y, al, señor, bachiller, sánson, carrasco, que, están, presentes, i
tem, es, mi, voluntad, que, si, antonia, quijana, mi, sobrina, quisiere, casarse, se, case, con, hombre, de, quien, primero, se, haya, hecho, información, que, n
o, sabe, qué, cosas, sean, libros, de, caballerías, y, en, caso, que, se, averiguar, que, lo, sabe, y, con, todo, eso, mi, sobrina, quisiere, casarse, con,
él, y, se, casare, pierda, todo, lo, que, le, he, mandado, lo, cual, puedan, mis, albacas, distribuir, en, obras, pias, a, su, voluntad, ítem, suplico, a, l
os, dichos, señores, mis, albacas, que, si, la, buena, suerte, les, trujere, a, conocer, al, autor, que, dicen, que, compuso, una, historia, que, anda, por, a
hi, con, el, título, de, segunda, parte, de, las, hazañas, de, don, quijote, de, la, mancha, de, mi, parte, le, pidan, cuan, encarecidamente, ser, pueda, perdo
ne, la, ocasión, que, sin, yo, pensarlo, le, di, de, haber, escrito, tantos, y, tan, grandes, disparates, como, en, ella, escribe, porque, parto, desta, vida,
con, escrupulo, de, haberle, dado, motivo, para, escribirlos, cerró, con, esto, el, testamento, y, comándole, un, desmayo, se, tendió, de, largo, a, largo, en, l
a, cama, alborotáronse, todos, y, acudieron, a, su, remedio, y, en, tres, días, que, vivió, después, deste, donde, hizo, el, testamento, se, desmayaba, muy, a,
menudo, andaba, la, casa, alborotada, pero, con, todo, comía, la, sobrina, brindaba, el, ama, y, se, regocijaba, sancho, panza, que, esto, del, heredar, algo, b
orra, o, templa, en, el, heredero, la, memoria, de, la, pena, que, es, razón, que, deje, el, muerto, en, fin, llegó, el, último, de, don, quijote, después,
de, recibidos, todos, los, sacramentos, y, después, de, haber, aboninado, con, muchas, y, eficaces, razones, de, los, libros, de, caballerías, hallóse, el, escriban
o, presente, y, dijo, que, nunca, había, leído, en, ningún, libro, de, caballerías, que, algún, caballero, andante, hubiese, muerto, en, su, techo, tan, sosgadame
nte, y, tan, cristiano, como, don, quijote, el, cual, entre, compasiones, y, lágrimas, de, los, que, allí, se, hallaron, dio, su, espíritu, quiero, decir, que,
se, murió, viendo, lo, cual, el, cura, pidió, al, escribano, le, diese, por, testimonio, como, alonso, quijano, el, bueno, llamado, comúnmente, don, quijote, de,
la, mancha, había, pasado, desta, presente, vida, y, muerto, naturalmente, y, que, el, tal, testimonio, pedia, para, quitar, la, ocasión, de, algún, otro, autor,
que, cide, hamete, benengeli, le, resucitase, falsamente, y, hiciese, increíbles, historias, de, sus, hazañas, este, fin, tuvo, el, ingenioso, hidalgo, de, la, m
ancha, cuyo, lugar, no, quiso, poner, cide, hamete, puntualmente, por, dejar, que, todas, las, villas, y, lugares, de, la, mancha, contendiesen, entre, si, por,
ahijárselo, y, tenérselo, por, suyo, como, contendieron, las, siete, ciudades, de, grecia, por, homero, déjanse, de, poner, aquí, los, llantos, de, sancho, sobrina,
y, ama, de, don, quijote, los, nuevos, epitafios, de, su, sepultura, aunque, sánson, carrasco, le, puso, éste, yace, aquí, el, hidalgo, fuerte, que, a, tanto,
estreño, llegó, de, valiente, que, se, advierte, que, la, muerte, no, triunfó, de, su, vida, con, su, muerte, tuvo, a, todo, el, mundo, en, poco, fue, el,
espantajo, y, el, coco, del, mundo, en, tal, coyuntura, que, acreditó, su, ventura, morir, cuerdo, y, vivir, loco, y, el, prudentísimo, cide, hamete, dijo, a,
su, pluma, aquí, quedarás, colgada, desta, espetera, y, deste, hilo, de, alambre, ni, sé, si, bien, cortada, o, mal, tajada, peñola, mia, adonde, vivirás, luen
gos, siglos, si, presuntuosos, y, malandrines, historiadores, no, te, descuelgan, para, profanarte, pero, antes, que, a, ti, lleguen, les, puedes, advertir, y, decir
les, en, el, mejor, modo, que, pudieres, tate, tate, follonicos, de, ninguno, sea, tocada, porque, esta, impresa, buen, rey, para, mi, estaba, guardada, para, m
i, sola, nació, don, quijote, y, yo, para, él, el, supo, obrar, y, yo, escribir, solos, los, dos, somos, para, en, uno, a, despecho, y, pesar, del, escrito
o, fingido, y, el, tordisillesco, que, se, atrevió, o, se, ha, de, atrever, a, escribir, con, pluma, de, avestruz, grosera, y, mal, deliada, las, hazañas, de, mi
, valeroso, caballero, porque, no, es, carga, de, sus, hombros, ni, asunto, de, su, resfriado, ingenio, a, quien, advertirás, si, acaso, llegas, a, conocerle, que
de,je, reposar, en, la, sepultura, los, cansados, y, ya, podridos, huesos, de, don, quijote, y, no, le, quiera, llevar, contra, todos, los, fueros, de, la,
muerte, a, castilla, la, vieja, haciéndole, salir, de, la, fuesa, donde, real, y, verdaderamente, yace, tendido, de, largo, a, largo, imposibilitado, de, hacer, te
ncera, jornada, y, si, nuevo, que, para, hacer, burta, de, tantas, como, hicieron, tantos, andantes, caballeros, bastan, las, dos, que, el, hizo, tan, a, gus
to, y, beneplácito, de, las, gentes, a, cuya, noticia, llegaron, así, en, éstos, como, en, los, extraños, reinos, y, con, esto, cumplirás, con, tu, cristiana,
profesión, aconsejando, bien, a, quien, mal, te, quiere, y, yo, quedaré, satisfecho, y, ufano, de, haber, sido, el, primero, que, gozó, el, fruto, de, sus, esc
ritos, enteramente, como, deseaba, pues, no, ha, sido, otro, mi, deseo, que, poner, en, aborrecimiento, de, los, hombres, las, fingidas, y, disparatadas, historias,
de, los, libros, de, caballerías, que, por, las, de, mi, verdadero, don, quijote, van, ya, tropezando, y, han, de, caer, del, todo, sin, duda, alguna, vale,
fin"
```

## 6.2 Conteo de palabra específica

```
# Cuenta cuántas veces aparece 'dichoso' (puedes cambiar la palabra).
```

```
docker exec spark-master-custom bash -c "/opt/spark/bin/spark-submit \
--master local[*] data/quijote_word_count.py dichoso"
```

La palabra 'dichoso' aparece 14 veces en El Quijote.

## 6.3 Guardado en HDFS

```
# Envía el recuento completo de palabras a HDFS.
```

```
docker exec spark-master-custom bash -c "/opt/spark/bin/spark-submit \
data/quijote_wordcount_to_hdfs.py"
```

```
C:\Users\Miguel>
que 20628
de 18214
y 18189
la 10363
a 9824
en 8242
el 8210
no 6335
los 4748
se 4691
```