

viu  
.es

2  
0  
2  
2  
-  
2  
0  
2  
3



# ACTIVIDAD 1: Programación con Hadoop MapReduce.

Máster en Big Data y Data Science

01MBID – Fundamentos de la Tecnología de Big Data

Autores: Adrián Hernández Padrón y Andrea San Blas Hernández.

Fecha: 24/06/2022

Curso 2022 – Ed. Abril

viu

Universidad  
Internacional

Para este ejercicio utilizarán el fichero de entrada cite75\_99.txt que puede ser descargado del National Bureau of Economic Research (NBER) de EEUU (<http://www.nber.org/patents/>).

Una descripción detallada de este fichero puede encontrarse en:

Hall, B. H., A. B. Jaffe, and M. Trajtenberg (2001). "The NBER Patent Citation Data File: Lessons, Insights and Methodological Tools." NBER Working Paper 8498.

Este fichero contiene citas de patentes emitidas entre 1975 y 1990 en los EEUU. Es un fichero CSV (comma-separated values) con más de 16,5 millones de filas, y las primeras líneas son como sigue:

**"CITING","CITED"**

**3858241,956203**

**3858241,1324234**

**3858241,3398406**

**3858241,3557384**

**3858241,3634889**

**3858242,1515701**

**3858242,3319261**

**3858242,3668705**

.....

La primera línea contiene una cabecera con la descripción de las columnas. Cada una de las otras líneas indica una cita que la patente con el número de la primera columna ha hecho a la patente con el número en la segunda. Por ejemplo, la segunda fila indica que la patente nº 3858241 ("citing" o citante) hace una cita a la patente nº 956203 ("cited" o citada).

El fichero está ordenado por las patentes citantes. Así podemos ver que la patente nº 3858241 cita a otras 5 patentes.

Deben implementar un programa MapReduce escrito en Java o Python (a elegir) que, para cada patente de cite75\_99.txt, obtenga la lista de las que la citan:

- Posible implementación:
  - El mapper obtiene cada línea del fichero de entrada, separar los campos y los invierte (para obtener como clave intermedia la patente citada y como valor intermedio la patente que la cita), por ejemplo:

**3858245, 3755824 → 3755824 3858245**

- El reducer, para cada patente recibe como valor una lista de las que la citan, ordena esa lista numéricamente y la convierte en un string de números separados por coma

**3755824 {3858245 3858247. . . } → 3755824 3858245, 3858247...**

- Formato de salida: patente patente1, patente2... (la separación entre la clave y los valores debe ser un **tabulado**)
  - La salida debe de estar **ordenada** por la clave (patentes citadas).
  - Los valores se deben guardar separados por coma, sin espacios en blanco entre ellos.
  - Deben tener en cuenta la cabecera, para que no aparezca en la salida (el fichero de entrada no debe modificarse de ninguna manera).

### Solución:

El programa 'CitantesPorPatentesCitadas' consta de dos archivos en lenguaje python. Veamos a continuación qué función cumplen cada uno de esos archivos.

- **Archivo 1:** CitantesPorPatentesCitadas\_mapper.py. Cada línea del fichero de entrada 'cite75\_99.txt' es procesada por un 'Map'. De ese string que recibe el 'Map' se construye un par <clave, valor>. Luego, este par es emitido por el 'Map' y recibido por el 'Reducer'. En este problema en concreto la clave es 'Cited' y el valor 'Citing'. Con otras palabras, el 'Mapper' separa los campos del fichero de entrada y emite los datos invirtiendo el orden de las columnas del archivo de entrada. De este modo, el 'Reducer' recibe como clave intermedia la patente citada y como valor intermedio la patente que la cita. Se ha seguido la posible implementación del enunciado del ejercicio.
- **Archivo 2:** CitantesPorPatentesCitadas\_reducer.py. Cada 'Reducer' recibe el par <clave, valor> emitido por el 'Mapper'. Y genera como único resultado cada patente citada con su lista de las que la citan, ordena esa lista en base a la clave y la convierte en un string de números separados por coma. De este modo el formato de salida se corresponde con:

**PatenteCitada1 PatenteCitante1.1, PatenteCitante1.2, ...**

**PatenteCitada2 PatenteCitante2.1, PatenteCitante2.2, PatenteCitante2.3, ...**

**PatenteCitada3 PatenteCitante3.1, PatenteCitante3.2, PatenteCitante3.3, ...**

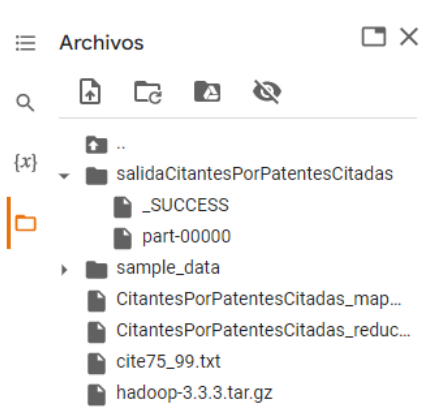
**PatenteCitada4 PatenteCitante4.1, ...**

•  
•  
•

**Donde la separación entre la clave y los valores debe ser un tabulado.**

Se tiene en cuenta que la explicación de cada uno de los códigos se encuentra explícita dentro de cada archivo.

Luego, se cargan los archivos en Google Colab:



Realizando una llamada al sistema damos los permisos de acceso a los ficheros:

```
[3] !chmod u+x ./CitantesPorPatentesCitadas_mapper.py
!chmod u+x ./CitantesPorPatentesCitadas_reducer.py
```

Ejecutamos en Hadoop el programa MapReduce:

```
!hadoop jar $HADOOP_HOME/share/hadoop/tools/lib/hadoop-streaming-3.3.3.jar
-files
./CitantesPorPatentesCitadas_mapper.py,./CitantesPorPatentesCitadas_reducer.py
-mapper ./CitantesPorPatentesCitadas_mapper.py
-reducer ./CitantesPorPatentesCitadas_reducer.py
-input cite75_99.txt -output ./salidaCitantesPorPatentesCitadas
```

Una explicación breve de este comando es la siguiente:

- `!hadoop jar`: Hadoop es una estructura de software de código abierto escrita en Java, por tanto debe ser ejecutado con `!hadoop jar`.
- `$HADOOP_HOME/share/hadoop/tools/lib/hadoop-streaming-3.3.3.jar`: Con esta línea de comando ejecutamos con `'hadoop-streaming'` y con la versión de Hadoop 3.3.3. El comando `'hadoop-streaming'` no quiere decir que los datos se procesan en tiempo real. Hadoop tiene procesamiento batch por defecto. Lo que indica este comando es la manera en la que se pasan los `'-input'` y los `'-output'`. Permite desarrollar ejecutables de MapReduce en lenguajes que no sean Java. En este caso, en lenguaje python.
- `-files ./CitantesPorPatentesCitadas_mapper.py, ./CitantesPorPatentesCitadas_reducer.py`: Indicamos cuales son los ficheros que vamos a utilizar. Estos ficheros se han subido previamente al entorno de trabajo.
- `-mapper ./CitantesPorPatentesCitadas_mapper.py`: Indicamos cual de esos archivos es el 'Mapper'.
- `-reducer ./CitantesPorPatentesCitadas_reducer.py`: Indicamos cual es el 'Reducer'.
- `-input cite75_99.txt`: Indicamos el archivo de entrada.
- `-output ./salidaCitantesPorPatentesCitadas`: Indicamos el archivo de salida.

Ejecutamos:

```
!hadoop jar $HADOOP_HOME/share/hadoop/tools/lib/hadoop-streaming-3.3.3.jar
-files ./CitantesPorPatentesCitadas_mapper.py,./CitantesPorPatentesCitadas_reducer.py
-mapper ./CitantesPorPatentesCitadas_mapper.py
-reducer ./CitantesPorPatentesCitadas_reducer.py
-input cite75_99.txt
-output ./salidaCitantesPorPatentesCitadas
```

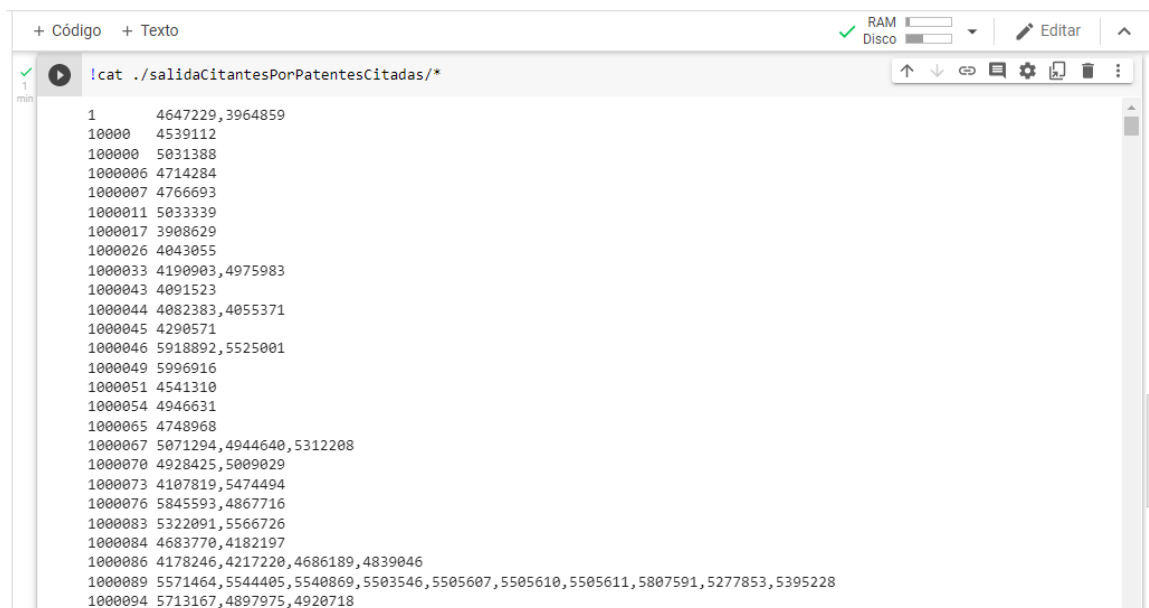
```
Reduce output records=3258983
Spilled Records=16522431
Shuffled Maps =8
Failed Shuffles=0
Merged Map outputs=8
GC time elapsed (ms)=15
Total committed heap usage (bytes)=382730240

Shuffle Errors
BAD_ID=0
CONNECTION=0
IO_ERROR=0
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0
File Output Format Counters
.....
```

Al finalizar la ejecución se obtiene como salida la creación de la carpeta: 'salidaCitantesPorPatentesCitadas'. Pedimos que se nos muestre la salida mediante el código:

```
!cat ./salidaCitantesPorPatentesCitadas/*
```

Finalmente, se muestra el resultado esperado para el fichero de entrada 'cite75\_99.txt' proporcionado. En la primera columna se tienen a los citados ordenados y en la segunda columna (y consecutivas) la lista de patentes citantes -separadas por comas- que corresponden a cada una de las patentes citadas.



```
+ Código + Texto
✓ 1 min
!cat ./salidaCitantesPorPatentesCitadas/*
1      4647229,3964859
10000  4539112
100000 5031388
1000006 4714284
1000007 4766693
1000011 5033339
1000017 3908629
1000026 4043055
1000033 4190903,4975983
1000043 4091523
1000044 4082383,4055371
1000045 4290571
1000046 5918892,5525001
1000049 5996916
1000051 4541310
1000054 4946631
1000065 4748968
1000067 5071294,4944640,5312208
1000070 4928425,5009029
1000073 4107819,5474494
1000076 5845593,4867716
1000083 5322091,5566726
1000084 4683770,4182197
1000086 4178246,4217220,4686189,4839046
1000089 5571464,5544405,5540069,5503546,5505607,5505610,5505611,5807591,5277853,5395228
1000094 5713167,4897975,4920718
```

Si observamos esta salida por pantalla se tiene, por ejemplo, que la patente citada número 1 tiene como citantes a las patentes número 4647229 y 3964859.