Práctica Spark

Gabriela Ballesteros Gómez

César de Diego Morales

Programación Paralela 2022-2023

El objetivo de esta práctica es trabajar con un conjunto de datos correspondiente a la empresa bicimad de la EMT y hacer varios análisis sobre estos utilizando las librerías de Spark adaptadas a python, pues el ambiente habitual de este es el lenguaje Java. Mencionar además que el formato habitual de estas bases de datos es .json.

1. Datos

Esta base de datos cuenta con información sobre los distintos movimientos realizados por los usuarios de Bicimad, así como algunas de sus características más significativas a lo largo de los años 2017 hasta la actualizada.

Hemos decidido centrarnos en el primer semestre de 2021, un año particular altamente influido por la situación de emergencia sanitaria ocasionada por la COVID-19. Por aquel entonces las restricciones marcaban el día a día de la sociedad, no solo por las medidas de distanciamiento social y sanitario si no también por las restricciones

geográficas definidas por el consejo de gobierno de la Comunidad de Madrid que variaban según la incidencia del virus en cada localidad.

El conjunto de datos está formado por distintas variables que describimos a continuación:

- *_id*: identificador del movimiento.
- *user_day_code*: código del usuario.
- *idunplug_station*: número de la estación de la que se desengancha la bicicleta.
- *idplug_station*: número de la estación en la que se engancha la bicicleta.
- *idunplug_base*: número de la base de la que se desengancha la bicicleta.
- *idplug_base*: número de la base en la que se engancha la bicicleta.
- *user_type*: número que indica el tipo de usuario que ha realizado del movimiento. Tiene distintos valores:
  - 0: no se ha podido determinar el tipo de usuario.
  - 1: usuario anual (poseedor de un pase anual)
  - 2: usuario ocasional
  - 3: trabajador de la empresa
- *ageRange*: número que determina el rango de edad. Sus posibles valores son:
  - 0: no se ha podido determinar el rango de edad del usuario.
  - 1: el usuario tiene entre 0 y 16 años.
  - 2: el usuario tiene entre 17 y 18 años.
  - 3: el usuario tiene entre 19 y 26 años.
  - 4: el usuario tiene entre 27 y 40 años.
  - 5: el usuario tiene entre 41 y 65 años.
  - 6: el usuario tiene 66 años o más.
- *travel_time*: tiempo total en segundos entre el enganche y el desenganche de la bicicleta.

- *unplug_hour_time*: muestra el día, mes y año en el que ha sucedido dicho desenganche.
- *zip_code*: texto que indica el código postal del usuario que ha realizado el movimiento

2. Método/objetivo

Tras un largo análisis de los datos disponibles y tras comprender el funcionamiento de bicimad, hasta entonces desconocido para nosotros, hemos optado por analizar distintas componentes que nos resultaban curiosas, así como relaciones entre ellas.

Para poder comenzar el proceso de análisis necesitábamos compactar los datos en un mismo dataFrame utilizando la librería Spark. Además, Para no sobrecargar la máquina y nuestro dataFrame seleccionamos únciamente las componentes de información que nos han resultado más relevantes, estas son: *travel_time, idunplug_station, idplug_station, ageRange, unplug_hourTime* y *user_type*.

Nuestra motivación, en primer lugar, fue conocer según grupo de edad la media recorrida por los usuarios según esta clasificación. Tras esto nos dimos cuenta de que debíamos distinguir entre los distintos tipos de usuario para poder ver la influencia ue esto tenía. Poco a poco descrubrimos otros aspectos interesantes, como las estaciones más y menos transitadas desde el punto de vista de partida y de llegada.

3. Resultados

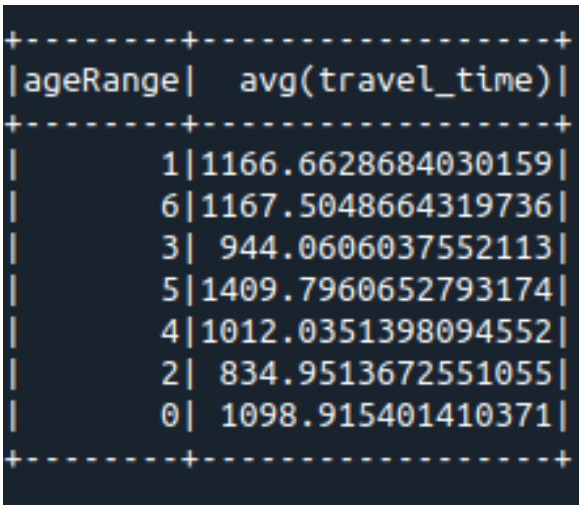
Veamos, en primer lugar, un pequeño boceto de la estructura del dataFrame tomando una muestra de los 20 primeros usuarios.

travel_time	idunplug_station	idplug_station	ageRange	unplug_hourTime	user_type
306	33	128	5	2021-01-01T00:00:00Z	1
305	166	114	3	2021-01-01T00:00:00Z	1
481	163	153	4	2021-01-01T00:00:00Z	1
378	9	198	0	2021-01-01T00:00:00Z	1
381	9	198	0	2021-01-01T00:00:00Z	1
269	221	182	0	2021-01-01T00:00:00Z	1
530	46	80	4	2021-01-01T00:00:00Z	1
375	206	153	3	2021-01-01T00:00:00Z	1
576	125	169	3	2021-01-01T00:00:00Z	1
673	187	136	5	2021-01-01T00:00:00Z	1
582	31	81	3	2021-01-01T00:00:00Z	1
239	253	14	4	2021-01-01T00:00:00Z	1
320	85	79	4	2021-01-01T00:00:00Z	1
728	134	224	4	2021-01-01T00:00:00Z	1
664	17	175	4	2021-01-01T00:00:00Z	1
365	155	200	3	2021-01-01T00:00:00Z	1
353	90	83	0	2021-01-01T00:00:00Z	1
797	17	175	5	2021-01-01T00:00:00Z	1
360	175	36	0	2021-01-01T00:00:00Z	1
288	83	89	5	2021-01-01T00:00:00Z	1

Nuestro primer análisis ha sido cuántas bicis se cogen por rango de edad.

ageRange	count
1	23344
6	9658
3	65722
5	356315
4	400116
2	5778

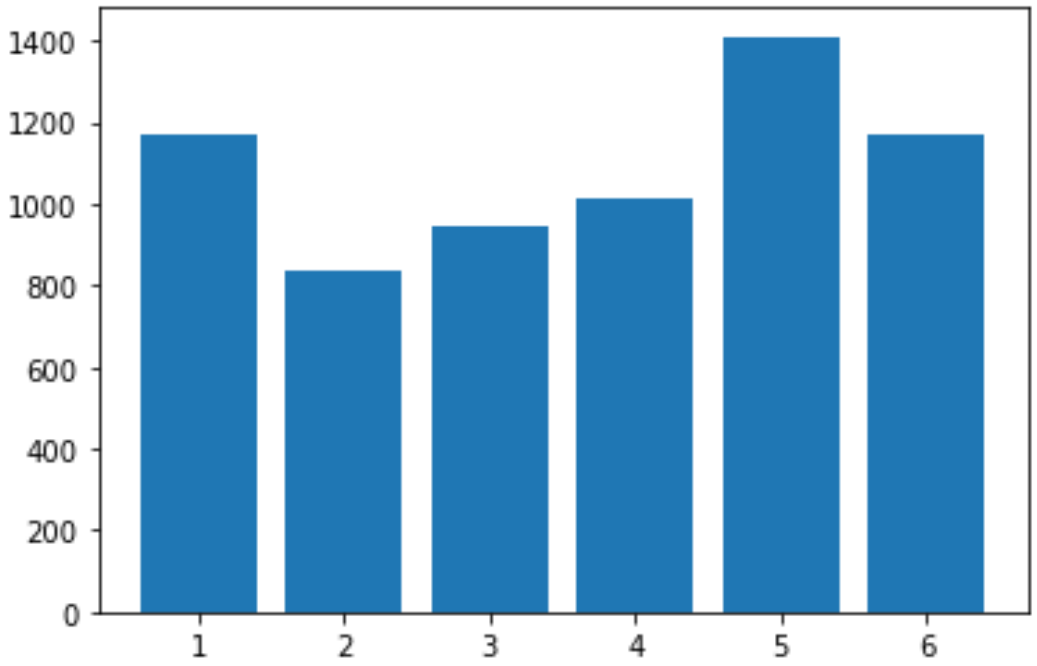
Como podemos observar, los usuarios que más uso hacen del servicio Bicimad fueron los grupos 4 y 5, formados mayoritariamente por trabajadores que utilizaron este sistema como alternativa más segura al transporte público. Esto se ve reforzado en nuestro siguiente análisis, donde calculamos el tiempo medio de uso por rango de edad y tenemos de nuevo al grupo 5 en el ranking más alto.



ageRange	avg(travel_time)
1	1166.6628684030159
6	1167.5048664319736
3	944.0606037552113
5	1409.7960652793174
4	1012.0351398094552
2	834.9513672551055
0	1098.915401410371

Los siguientes grupos con mayor tiempo medio de uso de bicis fueron el 1 y el 6, los grupos más extremos en cuanto a edades. Inferimos que esto se debe a que una gran cantidad de bicis de este servicio están disponibles en parques públicos y como es de esperar, es un rato de paseo y juego al aire libre.

Optamos graficar este resultado para tener una aproximación más visual.



Cabe destacar, que la media de cada viaje sin distinción de edades está alrededor de los 20 minutos.

Como es de esperar, no se tiene la misma afluencia de viajeros en todas las estaciones. Es por ello que hemos calculado la estación de salida más concurrida. Los resultados son los siguientes:

```
+----------------+-----+
|idunplug_station|count|
+----------------+-----+
|              43|22329|
+----------------+-----+
```

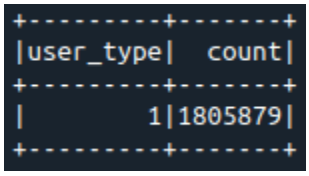
Estación donde más viajes comienzan

La estación 43 está ubicada en Lavapies. Coincide además con la estación destino más frecuente:

```
+--------------+-----+
|idplug_station|count|
+--------------+-----+
|            43|22591|
+--------------+-----+
```

Estación donde más viajes finalizan

Finalmente observamos que el tipo de usuario más habitual es el usuario anual como era de esperar.



```
+---------+-------+
|user_type|  count|
+---------+-------+
|        1|1805879|
+---------+-------+
```

4. Conclusiones

Hemos llegado a la conclusión de que bicimad fue una alternativa real al transporte público durante la primera etapa post-pandemia. Tuvo grandes beneficios, no sólo en situación de riesgo sanitario, pues la posibilidad de contagio con buenas medidas higiénicas era menor que usando el transporte público habitual. Además, se trata de un medio de transporte beneficioso para el medio ambiente y nuestra propia salud, por lo que nos ha agradado observar que tiene un alto número de usuarios en todos los rangos de edad.