Tarea 2 – RDDs

1. Para propósitos de esta tarea, trabajaremos únicamente en las variables “ent”, “sex”, “eda”, “clase1”, "e_con", "n_hij" e "ingocup". (inicial)

   Variables cuantitativas: “eda” e "ingocup".
   Variables cualitativas: “ent”, “sex”, “clase1”, "e_con" e "n_hij".

   Para poder tener un mejor y más fácil manejo de cada una de las variables, se anexa una columna “índice” a nuestros datos a observar; para el caso en que queramos seleccionar filas específicas. Nos aseguramos de que nuestra base de datos esté completa imprimiendo el número total de celdas.

```
+-----+---+---+---+------+-----+-----+-------+
|index|ent|sex|eda|clase1|e_con|n_hij|ingocup|
+-----+---+---+---+------+-----+-----+-------+
|    1|  1|  1|  0|  NULL| NULL| NULL|   NULL|
|    2|  1|  1|  0|  NULL| NULL| NULL|   NULL|
|    3|  1|  1|  0|  NULL| NULL| NULL|   NULL|
|    4|  1|  1|  0|  NULL| NULL| NULL|   NULL|
|    5|  1|  1|  0|  NULL| NULL| NULL|   NULL|
|    6|  1|  1|  0|  NULL| NULL| NULL|   NULL|
|    7|  1|  1|  0|  NULL| NULL| NULL|   NULL|
|    8|  1|  1|  0|  NULL| NULL| NULL|   NULL|
|    9|  1|  1|  0|  NULL| NULL| NULL|   NULL|
|   10|  1|  1|  0|  NULL| NULL| NULL|   NULL|
|   11|  1|  1|  0|  NULL| NULL| NULL|   NULL|
|   12|  1|  1|  0|  NULL| NULL| NULL|   NULL|
|   13|  1|  1|  0|  NULL| NULL| NULL|   NULL|
|   14|  1|  1|  0|  NULL| NULL| NULL|   NULL|
|   15|  1|  1|  0|  NULL| NULL| NULL|   NULL|
|   16|  1|  1|  0|  NULL| NULL| NULL|   NULL|
|   17|  1|  1|  0|  NULL| NULL| NULL|   NULL|
|   18|  1|  1|  0|  NULL| NULL| NULL|   NULL|
|   19|  1|  1|  0|  NULL| NULL| NULL|   NULL|
|   20|  1|  1|  0|  NULL| NULL| NULL|   NULL|
+-----+---+---+---+------+-----+-----+-------+
only showing top 20 rows

Total individuos: 415998
```

2. Toda nuestra base de datos podemos manejarla como un RDD. Aplicamos las funciones filtro-rdd a las columnas filtro-dataframe, de la imagen anterior, de la siguiente forma: (inicial)

   - Seleccionamos a la población económicamente activa ("clase1" == 1).
   - Seleccionamos a quienes tengan registrado su número de hijos ("n_hij" != None).
   - Seleccionamos a quienes tengan registrado sus ingresos mensuales ("ingocup" != None).
   - Seleccionamos a las entidades que sean de CDMX o Nuevo León ("ent" == 9 or "ent" == 19).
   - Imprimimos la información en formato lista (.collect()).

   Al trabajar con muchos datos, solo anexamos los primeros visualizados.

```
[Row(index=117738, ent=9, sex=2, eda=15, clase1=1, e_con=6, n_hij=0, ingocup=1032),
 Row(index=117756, ent=9, sex=2, eda=15, clase1=1, e_con=6, n_hij=0, ingocup=1290),
 Row(index=117765, ent=9, sex=2, eda=15, clase1=1, e_con=6, n_hij=0, ingocup=4300),
 Row(index=117825, ent=9, sex=2, eda=16, clase1=1, e_con=6, n_hij=0, ingocup=1935),
 Row(index=117838, ent=9, sex=2, eda=16, clase1=1, e_con=6, n_hij=0, ingocup=1935),
 Row(index=117863, ent=9, sex=2, eda=17, clase1=1, e_con=6, n_hij=0, ingocup=2150),
 Row(index=117878, ent=9, sex=2, eda=17, clase1=1, e_con=6, n_hij=0, ingocup=4300),
 Row(index=117910, ent=9, sex=2, eda=17, clase1=1, e_con=6, n_hij=1, ingocup=7740),
 Row(index=117933, ent=9, sex=2, eda=18, clase1=1, e_con=6, n_hij=0, ingocup=2580),
 Row(index=117947, ent=9, sex=2, eda=18, clase1=1, e_con=6, n_hij=0, ingocup=2580),
 Row(index=117953, ent=9, sex=2, eda=18, clase1=1, e_con=6, n_hij=0, ingocup=7740),
 Row(index=117954, ent=9, sex=2, eda=18, clase1=1, e_con=6, n_hij=1, ingocup=1720),
 Row(index=117956, ent=9, sex=2, eda=18, clase1=1, e_con=6, n_hij=0, ingocup=5160),
 Row(index=117980, ent=9, sex=2, eda=18, clase1=1, e_con=6, n_hij=0, ingocup=3870),
 Row(index=117989, ent=9, sex=2, eda=18, clase1=1, e_con=6, n_hij=0, ingocup=5400),
 Row(index=117997, ent=9, sex=2, eda=19, clase1=1, e_con=6, n_hij=0, ingocup=4300),
 Row(index=118007, ent=9, sex=2, eda=19, clase1=1, e_con=6, n_hij=1, ingocup=1032),
 Row(index=118021, ent=9, sex=2, eda=19, clase1=1, e_con=1, n_hij=0, ingocup=6880),
 Row(index=118025, ent=9, sex=2, eda=19, clase1=1, e_con=6, n_hij=1, ingocup=1720),
 Row(index=118027, ent=9, sex=2, eda=19, clase1=1, e_con=6, n_hij=0, ingocup=5160),
 Row(index=118036, ent=9, sex=2, eda=19, clase1=1, e_con=6, n_hij=0, ingocup=8200),
 Row(index=118039, ent=9, sex=2, eda=19, clase1=1, e_con=6, n_hij=0, ingocup=14000),
 Row(index=118047, ent=9, sex=2, eda=19, clase1=1, e_con=6, n_hij=0, ingocup=6450),
 Row(index=118048, ent=9, sex=2, eda=19, clase1=1, e_con=6, n_hij=1, ingocup=3655),
 Row(index=118049, ent=9, sex=2, eda=19, clase1=1, e_con=6, n_hij=0, ingocup=3870),
 Row(index=118050, ent=9, sex=2, eda=19, clase1=1, e_con=1, n_hij=1, ingocup=1720),
 Row(index=118055, ent=9, sex=2, eda=19, clase1=1, e_con=6, n_hij=0, ingocup=1720),
```

3. Para ver cuántos datos totales tenemos después de hacer las filtraciones con RDD, usamos la función count(). (inicial)

```
2895
```

Para los siguientes ejercicios, calcularemos como obtener promedios mediante el uso de RDD. Para ello usaremos nuestra base de datos antes de que le apliquemos RDD.

Estadísticas descriptivas básicas para CDMX

1. Realizamos la filtración-variables del punto 2 (inicial) sin haber convertido la información a RDD. Filtramos solo para la gente de la entidad 9.

```
+---+---+---+------+-----+-----+-------+
|ent|sex|eda|clase1|e_con|n_hij|ingocup|
+---+---+---+------+-----+-----+-------+
|  9|  2| 15|     1|    6|    0|   1032|
|  9|  2| 15|     1|    6|    0|   1290|
|  9|  2| 15|     1|    6|    0|   4300|
|  9|  2| 16|     1|    6|    0|   1935|
|  9|  2| 16|     1|    6|    0|   1935|
|  9|  2| 17|     1|    6|    0|   2150|
|  9|  2| 17|     1|    6|    0|   4300|
|  9|  2| 17|     1|    6|    1|   7740|
|  9|  2| 18|     1|    6|    0|   2580|
|  9|  2| 18|     1|    6|    0|   2580|
|  9|  2| 18|     1|    6|    0|   7740|
|  9|  2| 18|     1|    6|    1|   1720|
|  9|  2| 18|     1|    6|    0|   5160|
|  9|  2| 18|     1|    6|    0|   3870|
|  9|  2| 18|     1|    6|    0|   5400|
|  9|  2| 19|     1|    6|    0|   4300|
|  9|  2| 19|     1|    6|    1|   1032|
|  9|  2| 19|     1|    1|    0|   6880|
|  9|  2| 19|     1|    6|    1|   1720|
|  9|  2| 19|     1|    6|    0|   5160|
+---+---+---+------+-----+-----+-------+
only showing top 20 rows
```

Realizamos los siguientes pasos para calcular el número de hijos promedio que tienen las personas en CDMX:

- Seleccionamos la columna “n_hij” (select) y la convertimos en rdd.
- Convertimos los datos en un vector unitario (flatMap), y guardamos el resultado en una “variable 1”.
- Sumamos por pares los elementos guardados en “variable 1” (reduce), y guardamos el resultado en una “variable 2”.
- Dividimos el resultado “variable 2” entre “variable 1.count()”, e imprimimos el resultado.

```
1.6484962406015038
```

La gente de la CDMX tiene en promedio 2 hijos. **Nota:** después de analizar la información se observó que en dicha gente filtrada solamente están contempladas las mujeres. Por lo tanto, las mujeres de la CDMX tienen en promedio 2 hijos.

2. Hagamos un pequeño análisis descriptivo de las primeras 20 mujeres, que tienen 2 hijos, para ver lo que PySpark nos arroja

```
+------+---+---+---+------+-----+-----+--------+
| index|ent|sex|eda|clase1|e_con|n_hij|ingocup|
+------+---+---+---+------+-----+-----+--------+
|118073|  9|  2| 20|     1|    6|    2|  10000|
|118130|  9|  2| 20|     1|    2|    2|   3010|
|118307|  9|  2| 23|     1|    1|    2|   6880|
|118323|  9|  2| 23|     1|    2|    2|   3440|
|118336|  9|  2| 23|     1|    6|    2|   6450|
|118372|  9|  2| 24|     1|    6|    2|   3000|
|118374|  9|  2| 24|     1|    2|    2|   2580|
|118418|  9|  2| 24|     1|    1|    2|   6450|
|118474|  9|  2| 25|     1|    6|    2|   9030|
|118482|  9|  2| 25|     1|    6|    2|   3000|
|118483|  9|  2| 25|     1|    6|    2|   7000|
|118499|  9|  2| 25|     1|    1|    2|   3440|
|118547|  9|  2| 26|     1|    5|    2|   8000|
|118580|  9|  2| 26|     1|    1|    2|   4300|
|118605|  9|  2| 27|     1|    1|    2|   1500|
|118607|  9|  2| 27|     1|    2|    2|   3440|
|118615|  9|  2| 27|     1|    1|    2|   7000|
|118651|  9|  2| 28|     1|    1|    2|   6000|
|118653|  9|  2| 28|     1|    2|    2|   6000|
|118682|  9|  2| 28|     1|    4|    2|   2580|
+------+---+---+---+------+-----+-----+--------+
only showing top 20 rows
```

- La individua 118,073 tiene el mayor ingreso mensual ($10,000), está soltera (e_con = 6) y tiene 20 años.
- La individua 118,605 tiene el menor ingreso mensual ($1,500), vive con su pareja en unión libre (e_con = 1) y tiene 27 años.

Estadísticas descriptivas básicas para Nuevo León

1. Realizamos la filtración-variables del punto 2 (inicial) sin haber convertido la información a RDD. Filtramos solo para la gente de la entidad 19.

```
+---+---+---+------+-----+-----+-------+
|ent|sex|eda|clase1|e_con|n_hij|ingocup|
+---+---+---+------+-----+-----+-------+
| 19|  2| 13|     1|    6|    0|   3870|
| 19|  2| 14|     1|    6|    0|   6450|
| 19|  2| 14|     1|    6|    0|   4300|
| 19|  2| 14|     1|    6|    0|   3225|
| 19|  2| 14|     1|    6|    0|   6450|
| 19|  2| 15|     1|    6|    0|    860|
| 19|  2| 15|     1|    6|    0|    301|
| 19|  2| 15|     1|    6|    0|   1290|
| 19|  2| 16|     1|    6|    0|  12900|
| 19|  2| 16|     1|    6|    0|   2580|
| 19|  2| 16|     1|    6|    0|   2580|
| 19|  2| 16|     1|    6|    0|   6450|
| 19|  2| 16|     1|    6|    0|   6450|
| 19|  2| 16|     1|    6|    0|   5160|
| 19|  2| 17|     1|    6|    0|   7740|
| 19|  2| 17|     1|    6|    0|   6450|
| 19|  2| 17|     1|    6|    0|   7740|
| 19|  2| 17|     1|    6|    0|   9460|
| 19|  2| 17|     1|    6|    0|   1075|
| 19|  2| 17|     1|    6|    0|   7740|
+---+---+---+------+-----+-----+-------+
only showing top 20 rows
```

Realizamos los siguientes pasos para calcular el número de hijos promedio que tienen las personas en Nuevo León:

- Seleccionamos la columna “n_hij” (select) y la convertimos en rdd.
- Convertimos los datos en un vector unitario (flatMap), y guardamos el resultado en una “variable 3”.
- Sumamos por pares los elementos guardados en “variable 3” (reduce), y guardamos el resultado en una “variable 4”.
- Dividimos el resultado “variable 4” entre “variable 3.count()”, e imprimimos el resultado.

```
1.7749863462588749
```

La gente de Nuevo León tiene en promedio 2 hijos. **Nota:** después de analizar la información se observó que en dicha gente filtrada solamente están contempladas las mujeres. Por lo tanto, las mujeres de la Nuevo León tienen en promedio 2 hijos.

2. Hagamos un pequeño análisis descriptivo de las primeras 20 mujeres, que tienen 2 hijos, para ver lo que PySpark nos arroja

```
+------+---+---+---+------+-----+-----+-------+
| index|ent|sex|eda|clase1|e_con|n_hij|ingocup|
+------+---+---+---+------+-----+-----+-------+
|246599| 19|  2| 18|     1|    1|    2|    860|
|246854| 19|  2| 20|     1|    5|    2|   6000|
|246909| 19|  2| 20|     1|    6|    2|   7310|
|246912| 19|  2| 20|     1|    2|    2|   1935|
|246978| 19|  2| 21|     1|    1|    2|   4300|
|246996| 19|  2| 21|     1|    1|    2|   6600|
|247047| 19|  2| 21|     1|    1|    2|   2000|
|247094| 19|  2| 22|     1|    1|    2|   7310|
|247180| 19|  2| 22|     1|    5|    2|   8600|
|247312| 19|  2| 23|     1|    1|    2|  10750|
|247318| 19|  2| 23|     1|    2|    2|   9890|
|247346| 19|  2| 24|     1|    1|    2|   1290|
|247376| 19|  2| 24|     1|    5|    2|  21500|
|247380| 19|  2| 24|     1|    2|    2|   5590|
|247395| 19|  2| 24|     1|    1|    2|   3440|
|247403| 19|  2| 24|     1|    5|    2|   8000|
|247417| 19|  2| 24|     1|    2|    2|   4300|
|247428| 19|  2| 24|     1|    5|    2|   8600|
|247429| 19|  2| 24|     1|    1|    2|   8600|
|247438| 19|  2| 25|     1|    6|    2|   8600|
+------+---+---+---+------+-----+-----+-------+
only showing top 20 rows
```

- La individua 247,376 tiene el mayor ingreso mensual ($21,500), está casada (e_con = 5) y tiene 24 años.
- La individua 246,599 tiene el menor ingreso mensual ($860), vive con su pareja en unión libre (e_con = 1) y tiene 18 años.