UNIVERSITAT DE BARCELONA

Módulo I:

Tarea – Procesamiento de datos con Spark 2.X. Natalidad EE.UU.

Estudiante:

Heiner Romero Leiva

Máster en Big Data & Data Science

Noviembre, 2021

Descripción de la tarea:

Has sido contratado por una empresa consultora como Data Engineer y te proporcionan un fichero CSV con datos reales sobre la natalidad en EE. UU.

El esquema del fichero es el siguiente:

Field name	Type	Mode	Description
source_year	INTEGER	REQUIRED	Four-digit year of the birth. Example: 1975.
year	INTEGER	NULLABLE	Four-digit year of the birth. Example: 1975.
month	INTEGER	NULLABLE	Month index of the date of birth, where 1=January.
day	INTEGER	NULLABLE	Day of birth, starting from 1.
wday	INTEGER	NULLABLE	Day of the week, where 1 is Sunday and 7 is Saturday.
state	STRING	NULLABLE	The two character postal code for the state. Entries after 2004 do not include this value.
is_male	BOOLEAN	REQUIRED	TRUE if the child is male, FALSE if female.
child_race	INTEGER	NULLABLE	The race of the child. One of the following numbers: 1 - White 2 - Black 3 - American Indian 4 - Chinese 5 - Japanese 6 - Hawaiian 7 - Filipino 9 - Unknown/Other 18 - Asian Indian 28 - Korean 39 - Samoan 48 - Vietnamese
weight_pounds	FLOAT	NULLABLE	Weight of the child, in pounds.
plurality	INTEGER	NULLABLE	How many children were born as a result of this pregnancy. twins=2, triplets=3, and so on.
apgar_1min	INTEGER	NULLABLE	Apgar scores measure the health of a newborn child on a scale from 0-10. Value after 1 minute. Available from 1978-2002.
apgar_5min	INTEGER	NULLABLE	Apgar scores measure the health of a newborn child on a scale from 0-10. Value after 5 minutes. Available from 1978-2002.
mother_residence_state	STRING	NULLABLE	The two-letter postal code of the mother's state of residence when the child was born.
mother_race	INTEGER	NULLABLE	Race of the mother. Same values as child_race.
mother_age	INTEGER	NULLABLE	Reported age of the mother when giving birth.
gestation_weeks	INTEGER	NULLABLE	The number of weeks of the pregnancy.
lmp	STRING	NULLABLE	Date of the last menstrual period in the format MMDDYYYY. Unknown values are recorded as "99" or "9999".
mother_married	BOOLEAN	NULLABLE	True if the mother was married when she gave birth.
mother_birth_state	STRING	NULLABLE	The two-letter postal code of the mother's birth state.
cigarette_use	BOOLEAN	NULLABLE	True if the mother smoked cigarettes. Available starting 2003.
cigarettes_per_day	INTEGER	NULLABLE	Number of cigarettes smoked by the mother per day. Available starting 2003.
alcohol_use	BOOLEAN	NULLABLE	True if the mother used alcohol. Available starting 1989.
drinks_per_week	INTEGER	NULLABLE	Number of drinks per week consumed by the mother. Available starting 1989.
weight_gain_pounds	INTEGER	NULLABLE	Number of pounds gained by the mother during pregnancy.
born_alive_alive	INTEGER	NULLABLE	Number of children previously born to the mother who are now living.
born_alive_dead	INTEGER	NULLABLE	Number of children previously born to the mother who are now dead.
born_dead	INTEGER	NULLABLE	Number of children who were born dead (i.e. miscarriages)
ever_born	INTEGER	NULLABLE	Total number of children to whom the woman has ever given birth (includes the current birth).
father_race	INTEGER	NULLABLE	Race of the father. Same values as child_race.
father_age	INTEGER	NULLABLE	Age of the father when the child was born.
record_weight	INTEGER	NULLABLE	1 or 2, where 1 is a row from a full-reporting area, and 2 is a row from a 50% sample area.

Debes crear un programa **Python 3** y utilizar el API de **Spark 2.x**.

Tu punto de partida es la creación de un DataFrame a partir del fichero **natality.csv**.

IMPORTANTE

No utilices el API RDD de bajo nivel. Utiliza SparkSession, DataFrame y DataSet que proporciona Spark 2.x.

Debes realizar las siguientes tareas:

- Utilizando el API DataFrame:
  - Obtén en qué 10 estados nacieron más niños y niñas en 2003.
  - Obtén la media de peso de los niños y niñas por año y estado.
  - Evolución por año y por mes del número de niños y niñas nacidas (Resultado por separado con una sola consulta).
  - Obtén los tres meses de 2005 en que nacieron más niños y niñas.
  - Obtén los estados donde las semanas de gestación son superiores a la media de EE. UU.
  - Obtén los cinco estados donde la media de edad de las madres ha sido mayor.
  - Indica cómo influye en el peso del bebé y las semanas de gestación que la madre haya tenido un parto múltiple (campo plurality) a las que no lo han tenido.
- Utilizando el lenguaje SQL: Responde a las mismas pregunas enumeradas arriba.

Para la presente tarea se utiliza un archivo de Python, en el que, se han detallado todos los pasos para obtener los resultados, por lo que, es necesario abrir dicho archivo. Para facilitar su lectura se ha adjuntado un archivo con extensión HTML de fácil lectura que es muy recomendable abrir junto con este documento. En todos los archivos los ejercicios se resuelven primero usando PySpark y luego se hace lo mismo con los demás usando Spark-SQL. Se adjunta tres documentos: uno en formato py, ipynb y HTML.

Ejercicios:

1. **Obtén en qué 10 estados nacieron más niños y niñas en 2003.**

Para el año 2003 no se cuenta con 10 estados, solo con dos en los que nacieron más infantes (de ahora en adelante se referirá a infantes para incluir tanto a los niños como a las niñas). Los estados corresponden a Pensilvania (PA) y Washington (WA) respectivamente.

```
+-----+-----------+-----+
|state|source_year|count|
+-----+-----------+-----+
|   PA|       2003|   39|
|   WA|       2003|   20|
+-----+-----------+-----+
```

2. **Obtén la media de peso de los niños y niñas por año y estado.**

Se adjunta el siguiente resultado ordenado por la media de peso en orden descendente, se puede ver como en el estado de New Hampshire (NH) en el 2004 es en donde se obtiene la media de peso mayor con 8.24 libras mientras que en el estado de Idaho (ID) es donde se obtiene la media más baja con 6.34 libras.

```
+----+-----+-----------------+
|year|state|       media_peso|
+----+-----+-----------------+
|2004|   NH|    8.24969784404|
|2004|   KY|7.708830427933333|
|2004|   TN|   7.506188865445|
|2003|   WA|7.220139080499999|
|2004|   NY| 7.21329067831872|
|2004|   WA|7.211556799586427|
|2003|   PA|7.205867049854738|
|2008| null|7.134678552179328|
|2006| null|7.107366397303514|
|2004|   PA|7.104341277384499|
|2004|   FL|7.099987147710001|
|2007| null|7.098195196629874|
|2005| null|7.080841872104706|
|2004|   SC|     6.4705673897|
|2004|   ID|   6.340053730596|
+----+-----+-----------------+
```

3. **Evolución por año y por mes del número de niños y niñas nacidas (Resultado por separado con una sola consulta).**

Se ordena dicho resultado por año, en orden descendente y se divide la columna en niños y niñas para realizar el comparativo.

year	month	Ninos	Ninas
2008	5	14	18
2008	4	31	10
2008	10	23	18
2008	3	16	26
2008	2	20	28
2008	6	17	14
2008	9	20	28
2008	11	16	25
2008	8	28	18
2008	7	26	19
2008	1	32	19
2008	12	25	23
2007	11	294	266
2007	4	182	157
2007	5	168	142
2007	2	203	214
2007	12	287	303
2007	7	166	155
2007	3	228	225
2007	6	143	170
2007	1	249	242
2007	10	299	240
2007	8	164	138
2007	9	165	168
2006	12	283	251
2006	11	256	260
2006	4	254	229
2006	6	264	238
2006	10	302	238
2006	8	295	261
2006	2	263	253
2006	7	283	279
2006	9	272	272
2006	3	273	276
2006	1	242	242
2006	5	257	238
2005	11	178	167
2005	1	172	172
2005	7	162	180
2005	5	157	128
2005	8	177	173
2005	12	181	171
2005	2	151	145
2005	9	175	171

```
|2005|   10|  159|  182|
|2005|    3|  174|  186|
|2005|    4|  153|  141|
|2005|    6|  165|  170|
|2004|   10|    6|   11|
|2004|    7|    9|    6|
|2004|   12|   11|   13|
|2004|    6|   10|   10|
|2004|    4|   10|    6|
|2004|    2|    6|    5|
|2004|    1|    4|   10|
|2004|    5|    8|    5|
|2004|    8|    4|    9|
|2004|    9|    6|    5|
|2004|    3|    9|    6|
|2004|   11|   11|    8|
|2003|   10|    5|    2|
|2003|   12|    3|    1|
|2003|    3|    0|    4|
|2003|    1|    3|    1|
|2003|    9|    4|    3|
|2003|   11|    2|    3|
|2003|    5|    2|    6|
|2003|    7|    2|    4|
|2003|    8|    2|    3|
|2003|    6|    4|    5|
+----+-----+-----+-----+
```

4. Obtén los tres meses de 2005 en que nacieron más niños y niñas.

Los meses en que nacieron más infantes para 2005 corresponden a marzo, diciembre y Agosto con 360, 352 y 350 nacimientos respectivamente. Los datos parecen bajos para ser de Estados Unidos, pero se sobreentiende que este dataset es una muestra.

```
+-----+-----------+--------------------+
|month|source_year|cantidad_nacimientos|
+-----+-----------+--------------------+
|    3|       2005|                 360|
|   12|       2005|                 352|
|    8|       2005|                 350|
+-----+-----------+--------------------+
```

5. Obtén los estados donde las semanas de gestación son superiores a la media de EE. UU.

Para este ejercicio primero es importante conocer cuál es la media de gestación para todo EE.UU., y es de **38.65** semanas. Mediante la consulta se puede ver que los primeros estados con las semanas de gestación mayores se encuentran nulos pero el tercero corresponde a Washington (WA) con 45 semanas (está bastante por arriba de lo que dura un embarazo ya que generalmente duran 40 semanas).

Por último tenemos el caso de Idaho (ID) con 39 semanas.

Importante: *se opta por no limpiar los datos nulos, ya que estos arrojan información muy valiosa en el análisis, porque un auditor se puede preguntar cuáles serán esos estados en los que se tarda hasta 47 y 46 semanas de gestación y eso permite una exploración posterior de los datos; por ejemplo: ¿Por qué tengo datos nulos? ¿Por qué esos registros no existen? ¿Debo implementar una mejor manera de ingresar mis datos al Sistema? ¿Pasó algo para no tener esos datos (migración de base de datos, archivo plano dañado, mis datos se perdieron, etc.)? Si los nulos se borran o no se contemplan se estaría engañando al consumidor final ya que no se le está dando el panorama actual de sus datos y no se podrían tomar medidas para corregir esos errores.*

state	gestation_weeks
null	47
null	46
WA	45
null	45
null	44
null	43
FL	43
NY	42
PA	42
null	42
null	41
WA	41
FL	41
NY	41
PA	41
SC	40
TN	40
null	40
WA	40
FL	40
NY	40
PA	40
NY	39
PA	39
null	39
WA	39
TN	39
FL	39
KY	39
ID	39

6. Obtén los cinco estados donde la media de edad de las madres ha sido mayor.

Primero es importante saber cuál es la edad media de las madres para todo EE.UU., y esta corresponde a **25.45** años, seguidamente los cinco estados corresponden a un estado nulo como el que presenta la edad mayor, el segundo corresponde a Nueva York (NY) con 48 años de edad y los otros 3 estados se encuentran nulos. *Como se mencionó más arriba se opta por no borrar los nulos para futuros análisis.*

El query para obtener solamente los 5 estados sería:

```
data.select("state", "mother_age").\
    distinct().where(F.col("mother_age") > data.select(F.avg("mother_age")
).\
    head()[0]).\
    orderBy(data.mother_age.desc()).\
    show(5)
```

```
+-----+----------+
|state|mother_age|
+-----+----------+
| null|        50|
|   NY|        48|
| null|        48|
| null|        47|
| null|        46|
+-----+----------+
```

Se incluye la lista completa aquí para poder hacer una comparación:

```
+-----+----------+
|state|mother_age|
+-----+----------+
| null|        50|
|   NY|        48|
| null|        48|
| null|        47|
| null|        46|
| null|        45|
|   PA|        44|
| null|        44|
|   WA|        44|
|   WA|        43|
| null|        43|
|   NY|        42|
|   FL|        42|
|   ID|        42|
|   PA|        42|
| null|        42|
|   TN|        42|
|   WA|        41|
```

PA	41
null	41
FL	41
FL	40
PA	40
null	40
WA	40
ID	40
null	39
FL	39
PA	39
TN	39
WA	39
KY	39
NY	39
PA	38
FL	38
NY	38
null	38
WA	38
null	37
FL	37
WA	37
PA	37
NY	37
ID	37
NY	36
WA	36
TN	36
PA	36
FL	36
null	36
FL	35
NY	35
SC	35
null	35
WA	35
PA	35
null	34
PA	34
TN	34
NY	34
WA	34
WA	33
null	33
NY	33
SC	33
PA	33
FL	32
ID	32
NY	32
null	32
PA	32
NH	32
TN	32
KY	31
WA	31
NY	31
null	31

```
|   TN|        31|
|   PA|        31|
|   PA|        30|
|   TN|        30|
|   NY|        30|
|   FL|        30|
|   KY|        30|
| null|        30|
|   WA|        30|
|   FL|        29|
|   PA|        29|
|   WA|        29|
|   NY|        29|
| null|        29|
| null|        28|
|   NY|        28|
|   TN|        28|
|   FL|        28|
|   PA|        28|
|   WA|        27|
|   FL|        27|
| null|        27|
|   TN|        27|
+-----+----------+
```

7. **Indica cómo influye en el peso del bebé y las semanas de gestación que la madre haya tenido un parto múltiple (campo plurality) a las que no lo han tenido.**

Para realizar este ejercicio se opta por utilizar la media, ya qua es el estadístico más fácil de interpreter y se obtiene la siguiente tabla:

```
+------------------------+-----------------+--------------------------+
|Qty_nacimientos_por_parto|       Peso_medio|Media_de_semanas_gestacion|
+------------------------+-----------------+--------------------------+
|                       1|7.152029712115611|        38.748935895782274|
|                       2|4.842073024972972|        34.708708708708706|
|                       3|   3.814878981648|        31.733333333333334|
|                       4|    4.06311948866|                      32.5|
+------------------------+-----------------+--------------------------+
```

Haciendo un rápido análisis se puede ver que entre más hijos tienen las madres en un mismo parto el peso medio va disminuyendo (con excepción de tener un parto de 4 bebés ya que en este se tiene un peso medio mayor que dando a luz trillizos) asimismo pasa con la media de las semanas de gestación, entre más hijos de a luz una madre en un mismo parto las semanas de gestación van disminuyendo con excepción de tener 4 hijos, ya que tanto las semanas de gestación así como el peso medio son mayores a tener trillizos. Sería interesante ver porqué tener 4 hijos aumenta el peso medio y la media de semanas de gestación de los infantes significativamente comparado con las madres que dan a luz a trillizos.

---------- FIN ----------