In [ ]:

```python
from pyspark.sql import SparkSession

from pyspark.sql.functions import *
from pyspark.sql.types import *

spark = SparkSession \
    .builder \
    .appName("how to read csv file") \
    .getOrCreate()
```

Cargo los datos en un dataframe de spark

In [ ]:

```python
users_refined = spark.read.csv('/user/ort/obligatorio/users.csv', header=True, encoding="ISO-8859-1")
```

Vista previa de las primeras filas

In [ ]:

```python
users_refined.show()
```

```
+-------+----------+-----------+--------------------+-----------+---+---------------+------------------------+-------
------------------+
|user_id|first_name|  last_name|               email|     gender|age|principal_sport|other_sports_of_interest|princip
al_sport_frequency|
+-------+----------+-----------+--------------------+-----------+---+---------------+------------------------+-------
------------------+
|      1|      Alic|    Kaemena|akaemena0@4shared...|       Male| 18|     Senderismo|                Handball|
Yearly|
|      2|    Dorrie|   Fountian|dfountian1@craigs...|     Female| 51|        Running|              Senderismo|
Weekly|
|      3|   Spenser|  Fernandes|sfernandes2@devhu...|       Male| 31|       Natación|                  Tennis|
Once|
|      4|    Andris|    Moakson|amoakson3@rediff.com|       Male| 22|         Tennis|                  Volley|
Weekly|
|      5|     Elsey|   Rathbone|erathbone4@bloglo...|     Female| 23|       Natación|                  Tennis|
Seldom|
|      6|  Vivianna|     Naerup|vnaerup5@sourcefo...|     Female| 46|        Rafting|                  Fútbol|
Monthly|
|      7|    Jobina|     Toller|jtoller6@cbslocal...|     Female| 56|       Ciclismo|                  Tennis|
Never|
|      8|    Debbie|    Ambrodi|dambrodi7@seattle...|     Female| 51|       Trecking|                 Rafting|
Often|
|      9|   Kathryn|  Lissenden|klissenden8@quant...|     Female| 53|         Tennis|                 Running|
Monthly|
|     10|   Jeannie|Prettejohns|jprettejohns9@the...| Non-binary| 24|       Trecking|                 Running|
Once|
|     11|      Nike|     Clelle|nclellea@csmonito...|Genderfluid| 33|       Natación|                  Fútbol|
Monthly|
|     12|  Shoshana|    Tenaunt|stenauntb@mozilla...|     Female| 49|        Running|              Senderismo|
Monthly|
|     13|        Em|    Albrook|ealbrookc@netscap...|     Female| 29|        Running|                Ciclismo|
Seldom|
|     14|     Judah|     Fearon| jfearond@phpbb.com|       Male| 36|       Handball|                  Fútbol|
Often|
|     15|      Jock|Mothersdale|jmothersdalee@wik...|       Male| 58|       Handball|                  Volley|
Weekly|
|     16| Kristofor|   Holstein|kholsteinf@nature...|       Male| 51|       Natación|                  Fútbol|
Once|
|     17|     Edwin|    Battams|ebattamsg@pagineg...|       Male| 38|         Fútbol|                Handball|
Often|
|     18|   Stefano|     Truman|strumanh@xinhuane...|       Male| 24|       Natación|                  Volley|
Once|
|     19|  Mohammed|    Kincaid|mkincaidi@columbi...|       Male| 69|         Volley|                 Running|
Weekly|
|     20|      Tova|    Chazier|  tchazierj@nasa.gov|     Female| 65|       Natación|                  Fútbol|
Once|
+-------+----------+-----------+--------------------+-----------+---+---------------+------------------------+-------
------------------+
only showing top 20 rows
```

Cantidad de columnas del dataframe monthly_pays

In [ ]:

```
num_columns=len(users_refined.columns)
num_columns
```

Out[ ]:

```
9
```

Nombre de las columnas de monthly_pays

In [ ]:

```
users_refined.columns
```

Out[ ]:

```
['user_id',
 'first_name',
 'last_name',
 'email',
 'gender',
 'age',
 'principal_sport',
 'other_sports_of_interest',
 'principal_sport_frequency']
```

Descripción de los datos de la tabla

In [ ]:

```
users_refined.describe
```

Out[ ]:

```
<bound method DataFrame.describe of DataFrame[user_id: string, first_name: string, last_name: string, email: string,
gender: string, age: string, principal_sport: string, other_sports_of_interest: string, principal_sport_frequency: s
tring]>
```

Schema de la tabla

In [ ]:

```
users_refined.printSchema()
```

```
root
 |-- user_id: string (nullable = true)
 |-- first_name: string (nullable = true)
 |-- last_name: string (nullable = true)
 |-- email: string (nullable = true)
 |-- gender: string (nullable = true)
 |-- age: string (nullable = true)
 |-- principal_sport: string (nullable = true)
 |-- other_sports_of_interest: string (nullable = true)
 |-- principal_sport_frequency: string (nullable = true)
```

Valores Nulos o Faltantes

In [ ]:

```
total_nulos = users_refined.select([sum(col(c).isNull().cast("int")).alias(c) for c in users_refined.columns])

total_nulos.show()
```

```
+-------+----------+---------+-----+------+---+---------------+------------------------+-------------------------+
|user_id|first_name|last_name|email|gender|age|principal_sport|other_sports_of_interest|principal_sport_frequency|
+-------+----------+---------+-----+------+---+---------------+------------------------+-------------------------+
|      0|         0|        0|    0|     0|  0|              0|                       5|                        0|
+-------+----------+---------+-----+------+---+---------------+------------------------+-------------------------+
```

In [ ]:

```
hdfs_path = "/user/ort/obligatorio/refined/refined_users/"
users_refined.write.csv(hdfs_path, header=False, mode="overwrite")
```