Importaciones

In [17]:
```
import findspark
findspark.init()

import pandas as pd
import pyspark
```

In [18]:
```
from pyspark.sql import SparkSession
from pyspark.sql.functions import *

spark = SparkSession.builder\
        .master("local[*]")\
        .appName('PySpark_Tutorial')\
        .getOrCreate()
```

Lectura de CSV

In [19]:
```
world_happ_data_2021 = spark.read.csv(
    'C:/Users/sonia/Desktop/big-data-processing/Proyecto-Final/Datasets/world-happiness-report-2021.csv',
    sep = ',',
    header = True,
    inferSchema = True
    )
```

In [20]:
```
world_happ_data = spark.read.csv(
    'C:/Users/sonia/Desktop/big-data-processing/Proyecto-Final/Datasets/world-happiness-report.csv',
    sep = ',',
    header = True,
    inferSchema = True
    )
```

Cambio de variables

In [21]:
```
# world-happiness-report-2021.csv
world_happ_data_2021 = world_happ_data_2021.withColumnRenamed('Country name', 'Country_name')
world_happ_data_2021 = world_happ_data_2021.withColumnRenamed('Ladder score', 'Ladder_score')
world_happ_data_2021 = world_happ_data_2021.withColumnRenamed('Regional indicator', 'Regional_indicator')

# world-happiness-report.csv
world_happ_data = world_happ_data.withColumnRenamed('Country name', 'Country_name')
world_happ_data = world_happ_data.withColumnRenamed('Life Ladder', 'Life_Ladder')
world_happ_data = world_happ_data.withColumnRenamed('Log GDP per capita', 'Log_GDP_per_capita')
world_happ_data = world_happ_data.withColumnRenamed('Healthy life expectancy at birth', 'Healthy_life_expectancy_at_birth')
```

Ejercicio 1 ¿Cuál es el país más “feliz” del 2021?

In [22]:
```
# tabla temporal
world_happ_data_2021.createOrReplaceTempView("temp_table_2021")

# Query
spark.sql("select Country_name, Ladder_score from temp_table_2021 where Ladder_score = (select max(Ladder_score) from temp_table_2021 )").show()
```

```
+------------+-------------+
|Country_name|Ladder_score|
+------------+-------------+
|     Finland|        7.842|
+------------+-------------+
```

Ejercicio 2 ¿Cuál es el país más “feliz” del 2021 por continente?

In [23]:
```
df = pd.read_csv('C:/Users/sonia/Desktop/big-data-processing/Proyecto-Final/Datasets/world-happiness-report-2021.csv')

# Agrupa los datos por la columna 'Regional indicator' y obtiene el máximo valor de la columna 'Ladder score' para cada grupo
max_values = df.groupby('Regional indicator')['Ladder score','Country name'].max()

print(max_values)

# da un Warning pero ejecuta de forma correcta
```

```
                                 Ladder score              Country name
Regional indicator
Central and Eastern Europe              6.965                  Slovenia
Commonwealth of Independent States      6.179                Uzbekistan
East Asia                               6.584  Taiwan Province of China
Latin America and Caribbean             7.069                 Venezuela
Middle East and North Africa            7.157                     Yemen
North America and ANZ                   7.277             United States
South Asia                              5.269                 Sri Lanka
Southeast Asia                          6.377                   Vietnam
Sub-Saharan Africa                      6.049                  Zimbabwe
Western Europe                          7.842            United Kingdom
C:\Users\sonia\AppData\Local\Temp\ipykernel_17320\3983244100.py:4: FutureWarning: Indexing with multiple keys (implicitly converted to a tuple of keys) wil
l be deprecated, use a list instead.
  max_values = df.groupby('Regional indicator')['Ladder score','Country name'].max()
```

Ejercicio 3 ¿Cuál es el país que más veces ocupó el primer lugar en todos los años?</h1

In [24]:
```
# tabla temporal
world_happ_data.createOrReplaceTempView("temp_table")

# Query
spark.sql("SELECT Country_name, COUNT(*) as veces_primero FROM (SELECT *, ROW_NUMBER() OVER(PARTITION BY year ORDER by Life_Ladder DESC) as fila FROM temp_
```

```
+------------+-------------+
|Country_name|veces_primero|
+------------+-------------+
|     Denmark|            7|
|     Finland|            6|
|      Norway|            1|
| Switzerland|            1|
|      Canada|            1|
+------------+-------------+
```

Ejercicio 4 ¿Qué puesto de Felicidad tiene el país con mayor GDP del 2020?</h1

In [25]:
```
# tabla temporal
world_happ_data.createOrReplaceTempView("temp_table")

# Query
spark.sql("select Country_name,ROW_NUMBER() over (order by Log_GDP_per_capita DESC) as posicion_de_GDP, Log_GDP_per_capita, ROW_NUMBER() over (order by Life

# podemos ponerle LIMIT 1 para ver solo Ireland
```

```
+--------------------+--------------+------------------+---------------------+-----------+
|        Country_name|posicion_de_GDP|Log_GDP_per_capita|posicion_de_Felicidad|Life_Ladder|
+--------------------+--------------+------------------+---------------------+-----------+
|             Ireland|             1|            11.323|                   13|      7.035|
|         Switzerland|             2|            11.081|                    4|      7.508|
|United Arab Emirates|             3|            11.053|                   26|      6.458|
|              Norway|             4|            11.042|                    8|       7.29|
|       United States|             5|            11.001|                   14|      7.028|
|             Denmark|             6|             10.91|                    3|      7.515|
|         Netherlands|             7|            10.901|                    5|      7.504|
|             Austria|             8|            10.851|                   10|      7.213|
|              Sweden|             9|            10.838|                    6|      7.314|
|             Germany|            10|            10.833|                    7|      7.312|
|             Iceland|            11|            10.824|                    2|      7.575|
|             Belgium|            12|            10.771|                   17|      6.839|
|           Australia|            13|             10.76|                   12|      7.137|
|             Finland|            14|             10.75|                    1|      7.889|
|              Canada|            15|             10.73|                   15|      7.025|
|        Saudi Arabia|            16|            10.701|                   20|       6.56|
|         South Korea|            17|            10.648|                   46|      5.793|
|              France|            18|            10.643|                   19|      6.714|
|      United Kingdom|            19|            10.626|                   18|      6.798|
|             Bahrain|            20|             10.62|                   32|      6.173|
+--------------------+--------------+------------------+---------------------+-----------+
only showing top 20 rows
```

Ejercicio 5 ¿En que porcentaje a variado a nivel mundial el GDP promedio del 2020 respecto al 2021? ¿Aumentó o disminuyó?

In [26]:
```
df_all = pd.read_csv('C:/Users/sonia/Desktop/big-data-processing/Proyecto-Final/Datasets/world-happiness-report.csv')

df_2020 = df_all[df_all['year'] == 2020]

porcentaje_2020 = df_2020['Log GDP per capita'].mean()

print(f"GPD promedio del 2020 : {porcentaje_2020}\n")

df_2021 = pd.read_csv('C:/Users/sonia/Desktop/big-data-processing/Proyecto-Final/Datasets/world-happiness-report-2021.csv')

porcentaje_2021 = df_2021['Logged GDP per capita'].mean()

print(f"GPD promedio del 2021 : {porcentaje_2021}\n")

porcentaje_total = ((porcentaje_2021/ porcentaje_2020) - 1) * 100

print(f"El GPD promedio del 2020 al 2021 disminuyó en : {porcentaje_total}")
```

```
GPD promedio del 2020 : 9.751329545454546

GPD promedio del 2021 : 9.432208053691273

El GPD promedio del 2020 al 2021 disminuyó en : -3.272594678251106
```

Ejercicio 6 ¿Cuál es el país con mayor expectativa de vida ?

In [27]:
```
# tabla temporal
world_happ_data.createOrReplaceTempView("temp_table")

# Query
spark.sql("select Country_name, Healthy_life_expectancy_at_birth from temp_table where Healthy_life_expectancy_at_birth = (select max(Healthy_life_expectanc
```

```
+------------+-----------------------------------+
|Country_name|Healthy_life_expectancy_at_birth|
+------------+-----------------------------------+
|   Singapore|                               77.1|
+------------+-----------------------------------+
```

Ejercicio 6 bis ¿Cuánto tenia en ese indicador en el 2019?

In [28]:
```
# tabla temporal
world_happ_data.createOrReplaceTempView("temp_table")

# Query
spark.sql("select Country_name, Healthy_life_expectancy_at_birth, year from temp_table where Country_name = 'Singapore' and year = 2019").show()
```

```
+------------+-----------------------------------+----+
|Country_name|Healthy_life_expectancy_at_birth|year|
+------------+-----------------------------------+----+
|   Singapore|                               77.1|2019|
+------------+-----------------------------------+----+
```