Desarrollar aplicación wordcount

1. Abrimos la terminal
2. Nos cambiamos al escritorio: cd /home/moranjesus/Desktop

```
[moranjesus@localhost ~]$ cd /home/moranjesus/Desktop/
[moranjesus@localhost Desktop]$
```

3. Creamos una carpeta: mkdir wordcountSpark

```
[moranjesus@localhost Desktop]$ mkdir wordcountSpark
[moranjesus@localhost Desktop]$
```

4. Nos cambiamos a la carpeta: cd wordcountSpark

```
[moranjesus@localhost Desktop]$ cd wordcountSpark/
[moranjesus@localhost wordcountSpark]$
```

5. Creamos el archivo del programa Spark, para ello: touch miPrograma.py

```
[moranjesus@localhost wordcountSpark]$ touch miPrograma.py
[moranjesus@localhost wordcountSpark]$
```

6. Damos doble click en el documento y escribimos el siguiente código

```
#!/usr/bin/python

import sys

from pyspark import SparkContext, SparkConf

'''
Programa creado por Jesus Moran
Este programa cuenta el numero de apariciones de cada palabra
'''

#inicializacion
conf = SparkConf().setMaster("local").setAppName("mi programa")
sc = SparkContext(conf = conf)

entrada = sys.argv[1]
salida = sys.argv[2]
print salida

#cargamos los datos de entrada
datosEntrada = sc.textFile(entrada)

#hacemos el conteo de cada palabra
conteo = datosEntrada.flatMap(lambda linea: linea.split("
")).map(lambda palabra: (palabra, 1)).reduceByKey(lambda x, y: x +
y)

#guardamos la salida
conteo.saveAsTextFile(salida)
```

Open — miPrograma.py — ~/Desktop/wordcountSpark — Save

```
#!/usr/bin/python

import sys

from pyspark import SparkContext, SparkConf

'''
Programa creado por Jesus Moran
Este programa cuenta el numero de apariciones de cada palabra
'''

#inicializacion
conf = SparkConf().setMaster("local").setAppName("mi programa")
sc = SparkContext(conf = conf)

entrada = sys.argv[1]
salida = sys.argv[2]
print salida

#cargamos los datos de entrada
datosEntrada = sc.textFile(entrada)

#hacemos el conteo de cada palabra
conteo = datosEntrada.flatMap(lambda linea: linea.split(" ")).map(lambda palabra: (palabra, 1)).reduceByKey
(lambda x, y: x + y)

#guardamos la salida
conteo.saveAsTextFile(salida)
```

7. Le damos permisos de ejecución al programa: chmod u+x ./miPrograma.py

```
[moranjesus@localhost wordcountSpark]$ chmod u+x ./miPrograma.py
[moranjesus@localhost wordcountSpark]$
```

8. Creamos una carpeta un archivo para la entrada: touch miEntrada
9. Le damos doble click y escribimos varias líneas, por ejemplo:

   Esto es una prueba para contar palabras

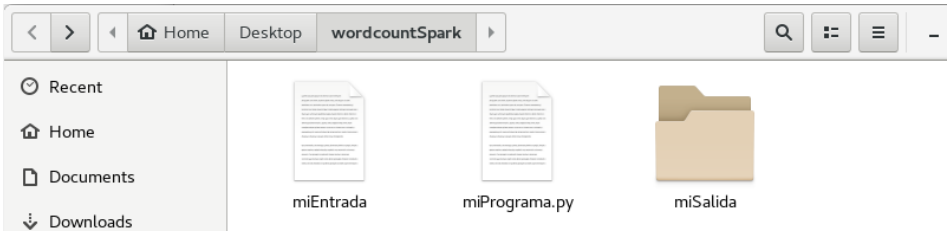
   El archivo tiene varias líneas

   El programa cuenta el numero de apariciones de cada palabra

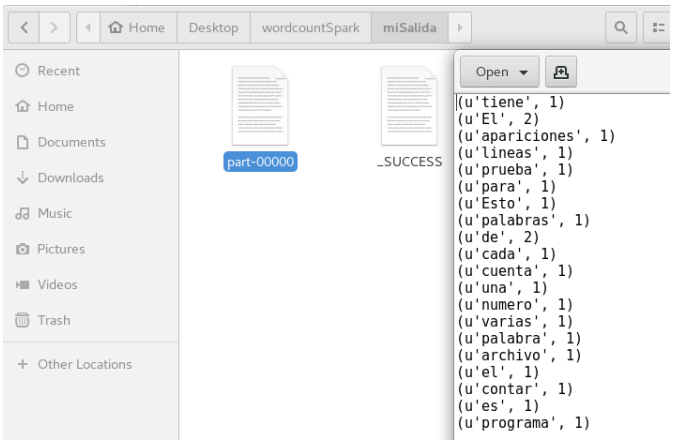
   (no dejamos una línea en blanco en el archivo)
10. Ejecutamos el programa, para ello: spark-submit miPrograma.py file:/home/moranjesus/Desktop/wordcountSpark/miEntrada file:/home/moranjesus/Desktop/wordcountSpark/miSalida

```
[moranjesus@localhost wordcountSpark]$ spark-submit miPrograma.py file:/home/mor
anjesus/Desktop/wordcountSpark/miEntrada file:/home/moranjesus/Desktop/wordcount
Spark/miSalida
```

Tendremos



11. En miSalida tenemos:



Desarrollar programa wordcount sin funciones lambda

1. Creamos un archivo: touch miProgramaSinLambda.py
2. Damos doble click en el documento y escribimos el siguiente código

```
#!/usr/bin/python

import sys

from pyspark import SparkContext, SparkConf

'''
Programa creado por Jesus Moran
Este programa cuenta el numero de apariciones de cada palabra
'''

#funcion que utilizo en el map: por cada palabra emite <palabra,
1>
def obtenerPalabrasUno(linea):
      paresClaveValor = []

      #genero los pares <clave, valor>
      palabras = linea.split(" ")
      for palabra in palabras:
            claveValor = (palabra, 1)
            paresClaveValor.append(claveValor)

      #emito los pares <clave, valor>
      return paresClaveValor

#funcion que utilizo en reduce: sumo cada valor de la clave
def obtenerSuma(valor1, valor2):
      return valor1 + valor2


#inicializacion
conf = SparkConf().setMaster("local").setAppName("mi programa")
sc = SparkContext(conf = conf)
```

```
entrada = sys.argv[1]
salida = sys.argv[2]

#cargamos los datos de entrada
datosEntrada = sc.textFile(entrada)

#hacemos el conteo de cada palabra
conteo =
datosEntrada.flatMap(obtenerPalabrasUno).reduceByKey(obtenerSuma
)

#guardamos la salida
conteo.saveAsTextFile(salida)
```

3. Le damos permisos de ejecución al programa: chmod u+x ./miProgramaSinLambda.py

```
[moranjesus@localhost wordcountSpark]$ chmod u+x ./miProgramaSinLambda.py
[moranjesus@localhost wordcountSpark]$
```

4. Ejecutamos el programa, para ello: spark-submit miProgramaSinLambda.py
   file:/home/moranjesus/Desktop/wordcountSpark/miEntrada
   file:/home/moranjesus/Desktop/wordcountSpark/miSalida2