1_1

August 7, 2020

[1]:
```
# Using the textFile method to create an RDD from a text file
rdd = sc.textFile("/data/students/bigdata_internet/lab1/lab1_dataset.txt")
# Question: Where is the input file? On which file system?
# The data is stored in the hdfs (as seen from hue)
```

[2]:
```
# the map transformation is used to create a new RDD by applying
# a function f on each element of the input RDD
# for each element of the input RDD there is a corresponding element
# in the output RDD
fields_rdd = rdd.map(lambda line: line.split(","))
```

[3]:
```
# a second map transformation takes the RDD that was given as
# output from the previous map transformation
# we take the second element of the line from the input RDD
value_rdd = fields_rdd.map(lambda l: int(l[1]))
```

[4]:
```
# reduce is an action so it will output a single python object
# obtained by combining all the objects of the input RDD
value_sum = value_rdd.reduce(lambda v1, v2: v1+v2)
```

[5]:
```
print("The sum is:", value_sum)
```

```
The sum is: 46
```

[6]:
```
# Question: Which value is printed by the print statement?
# The value given by the print statement is 46, which corresponds
# to the sum of all the second elements from the tuples of the input RDD
# Which is the purpose of each line of code?
# See line by line the explenations
```

[ ]:
```
# Question: Where is the input file? On which file system?
# The input file is stored in the hdfs
```

[1]:
```
# 1.2

# s274990@jupyter-s274990:~/newLabs/lab1$ pyspark --master local --deploy-mode␣
 ↪client
```

```
# WARNING: User-defined SPARK_HOME (/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/spark) overrides detected (/opt/cloudera/parcels/CDH/lib/spark).
# WARNING: Running pyspark from user-defined location.
# Python 3.7.5 (default, Oct 25 2019, 15:51:11)
# [GCC 7.3.0] :: Anaconda, Inc. on linux
# Type "help", "copyright", "credits" or "license" for more information.
# 20/08/04 16:23:26 WARN util.Utils: Service 'SparkUI' could not bind on port 4040. Attempting port 4041.
# 20/08/04 16:23:26 WARN util.Utils: Service 'SparkUI' could not bind on port 4041. Attempting port 4042.
# 20/08/04 16:23:26 WARN util.Utils: Service 'SparkUI' could not bind on port 4042. Attempting port 4043.
# 20/08/04 16:23:26 WARN util.Utils: Service 'SparkUI' could not bind on port 4043. Attempting port 4044.
# 20/08/04 16:23:26 WARN util.Utils: Service 'SparkUI' could not bind on port 4044. Attempting port 4045.
# Welcome to
#       ____              __
#      / __/__  ___ _____/ /__
#     _\ \/ _ \/ _ `/ __/  '_/
#    /__ / .__/\_,_/_/ /_/\_\   version 2.4.0-cdh6.2.1
#       /_/

# Using Python version 3.7.5 (default, Oct 25 2019 15:51:11)
# SparkSession available as 'spark'.
# >>> rdd = sc.textFile("/data/students/bigdata_internet/lab1/lab1_dataset.txt")
# >>> fields_rdd = rdd.map(lambda line: line.split(","))
# >>> value_rdd = fields_rdd.map(lambda l: int(l[1]))
# >>> value_sum = value_rdd.reduce(lambda v1, v2:v1+v2)
# >>> print("the sum is:", value_sum)
# the sum is: 46
```

[2]:

```
# The --master option specifies the master URL for a distributed cluster
#  local to run locally with one thread
# --deploy-mode client
# Whether to deploy your driver on the worker nodes (cluster) or locally as an external client (client)
```

[3]:

```
#1.3
```

[ ]:

```python
from pyspark import SparkConf, SparkContext
conf = SparkConf().setAppName("My app")
sc = SparkContext(conf=conf)
rdd = sc.textFile("/data/students/bigdata_internet/lab1/lab1_dataset.txt")
fields_rdd = rdd.map(lambda line: line.split(","))
```

```
value_rdd = fields_rdd.map(lambda l: int(l[1]))
value_sum = value_rdd.reduce(lambda v1, v2: v1 + v2)
print("The sum is:", value_sum)
# Question: In which file system are located your script and the /data/students/
 ↪bigdata_internet/lab1/lab1_dataset.txt files?
# Are they on the same file system?
# The script is in the local file system and the dataset is stored in the hdfs
```

[ ]:

```
# 2
```

[ ]:

```
hdfs dfs -ls /data/students/bigdata_internet/lab1
Found 1 items
-rwxrwx---+  3 trevisan students         62 2019-09-06 12:15 /data/students/
 ↪bigdata_internet/lab1/lab1_dataset.txt

hdfs dfs -usage
...
...

Now copy the HDFS file /data/students/bigdata_internet/lab1/lab1_dataset.txt in␣
 ↪the local file system.
Question: if you modify the local file, does the modifications automatically␣
 ↪affect also the HDFS file?
No it does not modify the file in the hdfs, but only that which is stored in␣
 ↪the local directory
```

[ ]:

```

```

3

August 7, 2020

[13]:
```
rdd = sc.textFile("/data/students/bigdata_internet/lab1/lab1_dataset.txt")
outputPath = "lab1/ex3/"
# Defining the input file with which to create the RDD to work with
# Defining the output path where to save the file
```

[14]:
```
#fields_rdd = rdd.map(lambda line: line.split(","))
```

[15]:
```
#print(fields_rdd.first())
# ['alice', '4']
```

[16]:
```
#name_total_rdd = fields_rdd.reduceByKey(lambda accum,n:accum+n)
```

[17]:
```
#print(name_total_rdd.first())
# ('bob', '533')
```

[23]:
```
# Function to pass to the map transformation in order to obtain a tuple of␣
 ↪(string,int)
def fromnamevaltotuplevalint(line):
    k,v = line.split(",")
    return (k,int(v))
```

[24]:
```
#tuples_rdd = rdd.map(lambda line: tuple(line.split(",")))
# applying the map funcion and passing the previously defined function to␣
 ↪create a new RDD
# that will contain the key, value pairs needed for the reduce operation
tuples_rdd = rdd.map(fromnamevaltotuplevalint)
```

[25]:
```
# print(tuples_rdd.first())
# ('alice', 4)
```

[26]:
```
# applying the reduce by key operation in order to obtain as output
# an RDD of pairs containing one pair for each key of the input RDD
name_total_rdd = tuples_rdd.reduceByKey(lambda accum,n:accum+n)
```

[27]:
```
print(name_total_rdd.first())
# ('bob', 11)
```

```
('bob', 11)
```

[ ]:

```
# saving the RDD to a text file
#name_total_rdd.saveAsTextFile(outputPath)
```

4

August 7, 2020

[41]:
```
# Always considering /data/students/bigdata_internet/lab1/lab1_dataset.txt ,
# write a script that reads the file, and concatenates all values for a name,
# separating them by : . Then, it saves the output in a HDFS file.
```

[42]:
```
rdd = sc.textFile("/data/students/bigdata_internet/lab1/lab1_dataset.txt")
outputPath = "lab1/ex4/"
```

[43]:
```
print(rdd.first())
```

```
alice,4
```

[44]:
```
# define a function that given a line of the RDD as input return a tuple
# of key value
def createkvpair(line):
    k,v = line.split(",")
    #v = int(v)
    return (k,v)
```

[45]:
```
# create an RDD of key value pairs by applying a map transformation and passing
# the user defined function
kv_rdd = rdd.map(createkvpair)
```

[46]:
```
print(kv_rdd.first())
```

```
('alice', '4')
```

[47]:
```
# defining the function that will perform the concatenation of values for a␣
 ↪given key
# of the input (key,value) RDD
def concatvalues(val1,val2):
    return val1 + ':' + val2
```

[48]:
```
# apply the reduce by key transformation in order to
# associate with each key of the input RDD one value
# the function must be associative and commutative
concat_rdd = kv_rdd.reduceByKey(concatvalues)
```

[49]:
```
print(concat_rdd.first())
```

```
('bob', '5:3:3')
```

[ ]: