Python- Split Rating Assignment

isra@MILE-BL-4782-LAP:/mnt/c/Users/Miles/Downloads/ml-latest-small/ml-latest-small$ cp ratings.csv ~/ratings.csv

(puts the file to hdfs)

isra@MILE-BL-4782-LAP:~$ hadoop fs -put ratings.csv hdfs://localhost:9000/user/training

(to view the files that are present in hdfs)

hadoop fs -ls hdfs://localhost:9000/user/training

Found 5 items

-rw-r--r-- 1 isra supergroup 2483723 2023-02-15 14:42 hdfs://localhost:9000/user/training/Rating

-rwxrwxrwx 1 root root 187 2023-02-14 12:48 hdfs://localhost:9000/user/training/hadoop.txt

-rw-r--r-- 1 isra supergroup 187 2023-02-14 12:45 hdfs://localhost:9000/user/training/hadoop1.txt

-rw-r--r-- 1 isra supergroup 2483723 2023-02-15 15:35 hdfs://localhost:9000/user/training/ratings.csv

-rw-r--r-- 1 isra supergroup 41 2023-02-14 17:43 hdfs://localhost:9000/user/training/wordcount-input.txt

isra@MILE-BL-4782-LAP:~$ cd futurense\_hadoop-pyspark/labs/examples/python-mr/

isra@MILE-BL-4782-LAP:~/futurense\_hadoop-pyspark/labs/examples/python-mr$ ls -l

total 16

-rw-r--r-- 1 isra isra 924 Feb 15 15:27 RatingSplit.py

-rw-r--r-- 1 isra isra 843 Feb 14 16:28 WordCount.py

-rw-r--r-- 1 isra isra 4136 Feb 14 16:28 python-wordcount-mr-execution-steps

isra@MILE-BL-4782-LAP:~/futurense\_hadoop-pyspark/labs/examples/python-mr$ python3 RatingSplit.py -r hadoop hdfs://localhost:9000/user/training/ratings.csv

RatingSplit.py

from mrjob.job import MRJob

class Count(MRJob):

""" The below mapper() function defines the mapper for MapReduce and takes

key value argument and generates the output in tuple format .

The mapper below is splitting the line and generating a word with its own

count i.e. 1 """

def mapper(self, \_, line):

for word in line.split():

field = line.split(',')

ratin = field[2]

yield(ratin, 1)

""" The below reducer() is aggregating the result according to their key and

producing the output in a key-value format with its total count"""

def reducer(self, ratin, counts):

yield(ratin, sum(counts))

"""the below 2 lines are ensuring the execution of mrjob, the program will not

execute without them"""

if \_\_name\_\_ == '\_\_main\_\_':

Count.run()

