(Text File in Databricks-MapReduce)

Data set Name: mapreduce.md

Question 1: How to load a text file?

filePath="dbfs:/databricks-datasets/ mapreduce.md"

lines=sc.textFile(filePath)

Question 2: Show me first 20 lines in the file

lines.take(20)

Question 3: Show me number of lines

lines.count()

Question 4: Show me first line

lines.first()

Question 5: Show me how many times “spark” word contains

linesWithSpark = lines.filter(lambda line: "Spark" in line)

linesWithSpark.count()

Question 6: Split each line into a list of words separated by a space from the dataset

words=lines.flatMap(lambdax: x.split(' '))

words.take(10)

Question 7: Filter the list of words to exclude common stop words

stopWords= ['','a','\*','and','is','of','the','a'] filteredWords=words.filter(lambdax: x.lower() notinstopWords)

filteredWords.take(10)

Question 8: Cache the filtered dataset in memory to speed up future actions.

filteredWords.cache()

Question 9: Transform filtered words into list of (word,1) tuples for WordCount

word1Tuples=filteredWords.map(lambdax: (x, 1))

word1Tuples.take(10)

Question 10:Aggregate the (word,1) tuples into (word,count) tuples

wordCountTuples=word1Tuples.reduceByKey(lambdax, y: x+y)

wordCountTuples.take(10)

Question 11: Display the top 10 (word,count) tuples by count

sortedWordCountTuples=wordCountTuples.top(10,key=lambda (x, y): y)

fortupleinsortedWordCountTuples:

printstr(tuple)

Question 12: Create a table from the (word,count) tuples

frompyspark.sqlimportRow

wordCountRows=wordCountTuples.map(lambdap: Row(word=p[0], count=int(p[1])))

wordCountRows.toDF().createOrReplaceTempView("word\_count")