Big Data Analytics (MPBA G517)

Lab Sheet - 4

Instructor: Revendranath T

Teaching Assistants: Jainil Shah & Shashank Sharma

Spark RDD Transformation Functions:

Transformation Methods	Method Usage and Description
cache()	Caches the RDD
filter()	Returns a new RDD after applying the filter function on the source dataset.
flatMap()	Returns flatten map meaning if you have a dataset with an array, it converts each element in an array as a row. In other words it returns 0 or more items in output for each element in the dataset.
map()	Applies transformation function on dataset and returns same number of elements in distributed dataset.
mapPartitions()	Similar to map, but executs transformation function on each partition, This gives better performance than map function.
mapPartitionsWithIndex()	Similar to map Partitions, but also provides func with an integer value representing the index of the partition.
randomSplit()	Splits the RDD by the weights specified in the argument.
union()	Comines elements from source dataset and the argument and returns combined dataset. This is similar to union function in Math set operations.
sample()	Returns the sample dataset.
intersection()	Returns the dataset which contains elements in both source dataset and an argument.
distinct()	Returns the dataset by eliminating all duplicated elements.
repartition()	Return a dataset with number of partition specified in the argument. This operation reshuffles the RDD randamly, It could either return lesser or more partioned RDD based on the input supplied.
coalesce()	Similar to repartition by operates better when we want to the decrease the partitions. Betterment acheives by reshuffling the

	data from fewer nodes compared with all nodes by repartition.

Problem Statement:

1. Load the textfile ‘blogtexts’. Extract first five elements of RDD.
2. Convert all words in the RDD to lower case and split the lines of a document using space.

   Take the result of this transformation in ‘RDD2’.
3. Stop words are words that do not add much value in a text and are not necessary to analyze the text. For example, “is”, “am”, “are” and “the” are few examples of stop words.

   Remove such stop words from the given RDD. Get the result of this transformation in ‘RDD3’.
4. Group each word of RDD3 based on last 3 characters.
5. Create a mapped (key,val) pair RDD of RDD3. Group all the elements based on the keys (words). Save the results in another RDD named as ‘RDD3_Mapped’
6. Calculate the frequency of the word ‘manager’.
7. Calculate the number of distinct words in RDD3.
8. Count the words ‘spark’ and ‘apache’ in RDD3 separately on each partition.
9. Create a two sample RDD (say sample1, sample2) from the “rdd3” by taking 20% sample for each. Apply a union transformation on sample1, sample2.