Reading Assignmrnt 6: Transformations and Actions

Transformations and Actions are the operations which are provides by Apache spark, which results into either one or multiple new RDD (Resilient Distributed Datasets).

Transformation: Transformation creates always new RDD rather than updating the pevious or an existing one, because RDD is not mutable. Some of the transformation functions are described below:

Map(func): Return a new distributed dataset which is formed through the function func by passing each elements of the source

filter(func): return a new dataset which contains all elements of source for which function func returns true

sample(withReplacement, fraction, speed): By using given speed, with or without replacement, sample fraction of the data

union(otherDataset): return a dataset which contains union of the elments present in source dataset and the argument

intersection(otherDataset): return a dataset which contains intersection of the elements present in source dataset and the argument

distinct([numPartition]): Return a new dataset which contains the distinct elements of the source dataset

join(otherDataset, [*numPartitions*]): Return all pairs of elements for each key. Also it supports outer joins through leftOuterJoin, rightOuterJoin, and fullOuterJoin

Actions: Actions are spark DD operations that give non-RDD values. When the action is triggered after the result, a new RDD is not formed like transformations. Some of the action functions are described below:

reduce(func): Using a function with two arguments, aggregate the elements of the dataset

collect(): Usually used after a filter to collect the elements of the dataset as an array at the driver program

count(): return a count of the elements which are present in dataset

first(): return the first element of the dataset

take(n): return an array with the first *n* elements of the dataset

saveAsTextFile(*path*): Spark will call toString on each element of the dataset to convert it to a line of text in a directory

foreach(*func*): on each element of the dataset, this function will be run