Apache Spark: Actions & Transformations

Actions and transformations are operations related to Resilient Distributed Datasets (RDDs).

Through transformation operations we can create RDDs, following are some of the examples of transformation operations.

1. ***filter(func)*** : this function returns a new dataset that is formed by selecting elements from the source on which the func returns true
2. ***map(func)*** :returns a distributed dataset that is obtained by passing the source elements into the func
3. ***sortByKey([ascending], [numPartitions])*** : It returns a dataset in the form of (key, value) pairs that is sorted in either ascending or descending order (as mentioned in the boolean argument)
4. ***pipe(command, [envVars])*** : pipes the partitions of the RDD through a shell command

Through action operations return a result to the driver program the results of which are then used to obtain a select dataset from a source. The results can also be written to a storage to kick start a computation. Following are a few examples of action operations.

1. ***countByKey()*** : it returns a mapping, where keys are mapped to an integer values where the integers are the count of the keys
2. ***first()*** : an action operation that returns the first element of a dataset
3. ***count()*** : a function that will return number of elements in a given dataset
4. ***take(n)*** : this functions will return an array of first n elements of a given dataset

Both actions and transformations operations can be applied on a given RDD, either on a single RDDs or on multiple RDDs depending on the use case to obtain desired results.