Apache Spark: Actions & Transformations

Both "Actions" and "Transformations" are Resilient Distributed Dataset (RDD) operations. They are functions that, depending on the circumstance, can be applied to a single RDD or numerous RDDs to generate the appropriate output.

Transformations are RDD activities that result in the production of another RDD. For example, using the filter () function to filter a certain name returns another RDD with the filtered data.

Here are a few examples of transformation operations:

• map(func) – This method produces a distributed dataset built by passing each element through the function func. exampleRDD=exampleRDD.map(lambda x: (x,1)) is an example command.

• filter(func) – Returns a new dataset constructed by filtering those items of the source that return true when func is called. exampleRDD=exampleRDD.filter(lambda x: "abc" in x)

• union(otherDataset) - Creates a new dataset from the union of the elements in the source dataset and the argument. Union

EXAMPLE RDDS1 = EXAMPLE RDDS 1.union(EXAMPLE RDDS 2) is an example command.

Actions are RDD operations that return a result to the driver program or write it to a storage location before starting a calculation. Example: Counting elements in a dataset using the action function count() to return the number of elements accessible.

Among the action operations are:

• take(n) – Returns an array containing the first n entries of the dataset. ExampleRDD.take is an example command (2)

• first() – This function returns the first entry of a dataset. ExampleRDD.first is an example command ()

• count() – This function returns the number of entries in the dataset. ExampleRDD.count is an example command ()

• collect() - Returns all the items of a dataset as an array to the driver.