In simple words, ***keyBy*** is one of the specialized form of ***map***. i.e *keyBy* is derived from *map* function only.

Is answer looks too **long**?  But I assure you it will be **worth** reading. 

Firstly, we will try to understand the definitions of both functions:

**Definition of KeyBy Function:**

def keyBy(f: (T) ⇒ K): RDD[(K, T)]  
//Here keyBy function takes another function **f:(T)=>K***(f takes T(here T is original RDD’s element Type) as argument and returns K(here K is Resultant RDD’s key Type)* as argument and **returns** RDD of key-value tuples **RDD[(K,T)]**

**Definition of map Function:**

def map(f: (T) ⇒ U): RDD[U]  
//Here *map* function takes another function **f:(T)=>U***(f takes T(here T is original RDD’s element Type) as argument and returns U(here U is Resultant RDD’s element Type)* as argument and **returns**RDD of **U** Type elements **RDD[U]**

From definitions of both functions one can easily come to the conclusion that *keyBy* function returns **paired RDD** whereas *map* returns **simple RDD**. Although it is correct, you can pass a function to **map** so that it returns **paired RDD**, in fact you can achieve the same results of *keyBy* with *map* function. Look at the example below where *keyBy* and *map* both produce same results:

val rdd = sc.parallelize(List(1,2,3))

def keyByF(t:Int) = t\*t

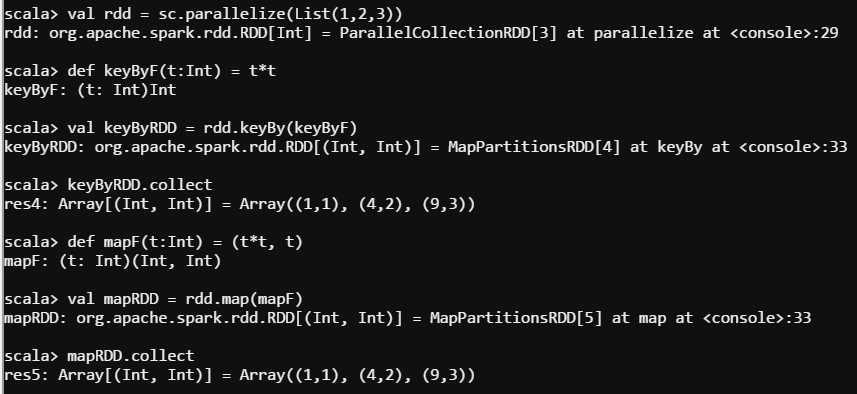
val keyByRDD = rdd.keyBy(keyByF)

keyByRDD.collect

def mapF(t:Int) = (t\*t, t)

val mapRDD = rdd.map(mapF)

mapRDD.collect

[[](http://discuss.itversity.com/uploads/default/original/1X/d07c5c8d7711f98ec3146eee28a4718c0bcf7c9d.png)](http://discuss.itversity.com/uploads/default/original/1X/d07c5c8d7711f98ec3146eee28a4718c0bcf7c9d.png" \o "itversity-2.png)

**[itversity-2.png857×394 20.5 KB](http://discuss.itversity.com/uploads/default/original/1X/d07c5c8d7711f98ec3146eee28a4718c0bcf7c9d.png" \o "itversity-2.png)**

You can find full code of **keyBy** function(I just extracted important line and provided below) in **Spark Source Code** [here] ([https://github.com/apache/spark/blob/v1.6.2/core/src/main/scala/org/apache/spark/rdd/RDD.scala#L1490 86](https://github.com/apache/spark/blob/v1.6.2/core/src/main/scala/org/apache/spark/rdd/RDD.scala#L1490)):

map(x => (f(x), x))

**Which One to Use?**

1. When you have straightforward requirement to create simple pairs then it’s better to use ***keyBy***because it is very intuitive than *map* ( expecting (key, value) pairs as output of *map* function? isn’t it ODD?)
2. Although expecting (key,value) pairs as output from ***map*** function is confusing, but it is very flexible and can be useful in many cases unlike *keyBy* .

Hope it helps!

Solution

5