When to use reduceByKey and when to use groupByKey

```
import org.apache.log4j.{Level, Logger}
import org.apache.spark.SparkConf
import org.apache.spark.sql.SparkSession

object ReduceByKeyandGroupByKey extends App{
 Logger.getLogger("org").setLevel(Level.ERROR)

 /**
  * create spark configuration and spark session
  */
 val sparkConf = new SparkConf()
 sparkConf.set("spark.app.Name","reduceByKey and
groupByKey").set("spark.master","local")
 val spark = SparkSession.builder().config(conf =
sparkConf).getOrCreate()

 /**
  * list of tuples with (Chars,Integers)
  */
 val listofTuple
=List(('A',2),('B',3),('A',1),('C',2),('B',4),('C',6),('A',3
))
 val rdd = spark.sparkContext.parallelize(listofTuple)

 /**
  * reduceByKey transformation is applied
  */
 val reduce_by_key_rdd = rdd.reduceByKey((x,y)=>x+y)

 /**
  * collecting the result in Array[Char, Int]
  */
 val rbk_result = reduce_by_key_rdd.collect()

 /**
  * print each element aggregation
  */
 println("reduceByKey aggregation :")
 rbk_result.foreach(println)

 /**
  * groupByKey transformation is applied
  */
 val group_by_key_rdd = rdd.groupByKey()

 /**
  * collecting result in Array[(Char,Iterable[Int])]
  */
 val gbk_result = group_by_key_rdd.collect()

 /**
  * print each element grouping
  */
 println("groupByKey grouping :")
 gbk_result.foreach(println)

 /**
  * aggregating the group value result
  */
 val gbk_agg_result =
group_by_key_rdd.mapValues(x=>x.sum).collect()

 /**
  * print each element aggregated grouping
  */
 println("groupByKey grouping and aggregation :")
 gbk_agg_result.foreach(println)

 //spark.stop()
 scala.io.StdIn.readLine()
}
```

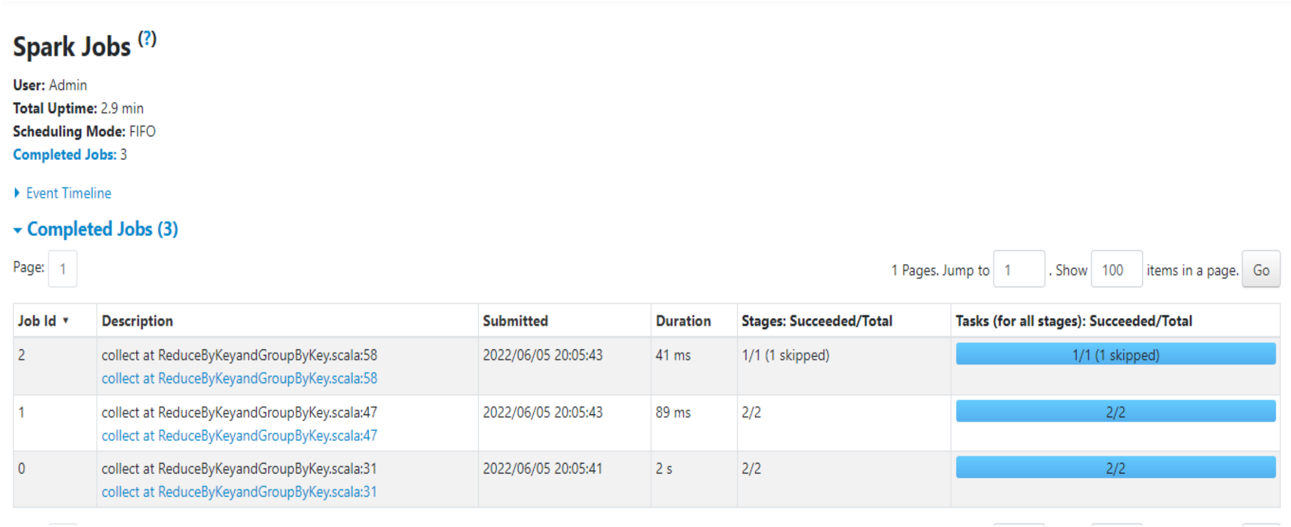
Output :

```
reduceByKey aggregation :
(B,7)
(A,6)
(C,8)

groupByKey grouping :
(B,Seq(3, 4))
(A,Seq(2, 1, 3))
(C,Seq(2, 6))

groupByKey grouping and aggregation :
(B,7)
(A,6)
(C,8)
```

Observations:

Spark UI:

In spark 3 jobs created as in code times collect() action called



Spark Jobs (?)

User: Admin
Total Uptime: 2.9 min
Scheduling Mode: FIFO
Completed Jobs: 3

▸ Event Timeline

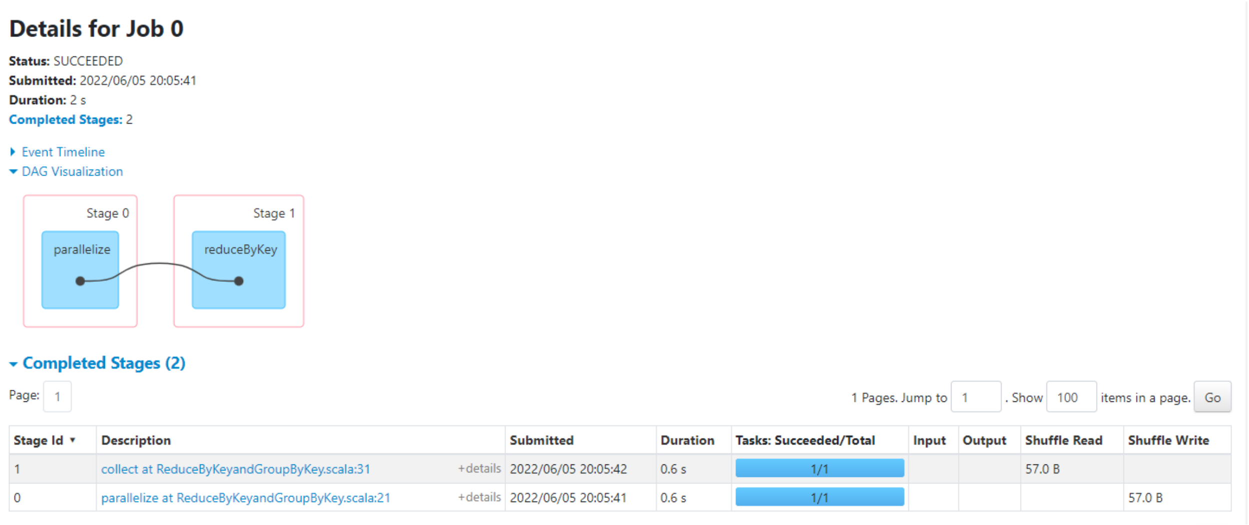
▾ Completed Jobs (3)

Page: 1   1 Pages. Jump to 1 . Show 100 items in a page. Go

Job Id ▾	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
2	collect at ReduceByKeyandGroupByKey.scala:58 collect at ReduceByKeyandGroupByKey.scala:58	2022/06/05 20:05:43	41 ms	1/1 (1 skipped)	1/1 (1 skipped)
1	collect at ReduceByKeyandGroupByKey.scala:47 collect at ReduceByKeyandGroupByKey.scala:47	2022/06/05 20:05:43	89 ms	2/2	2/2
0	collect at ReduceByKeyandGroupByKey.scala:31 collect at ReduceByKeyandGroupByKey.scala:31	2022/06/05 20:05:41	2 s	2/2	2/2

In reduceByKey is work on (key, value) pair and the aggregation is completed at the partition level first and the shuffling of data between partitions for the final aggregation

So, use When we require the final result in aggregated



Details for Job 0

Status: SUCCEEDED
Submitted: 2022/06/05 20:05:41
Duration: 2 s
Completed Stages: 2

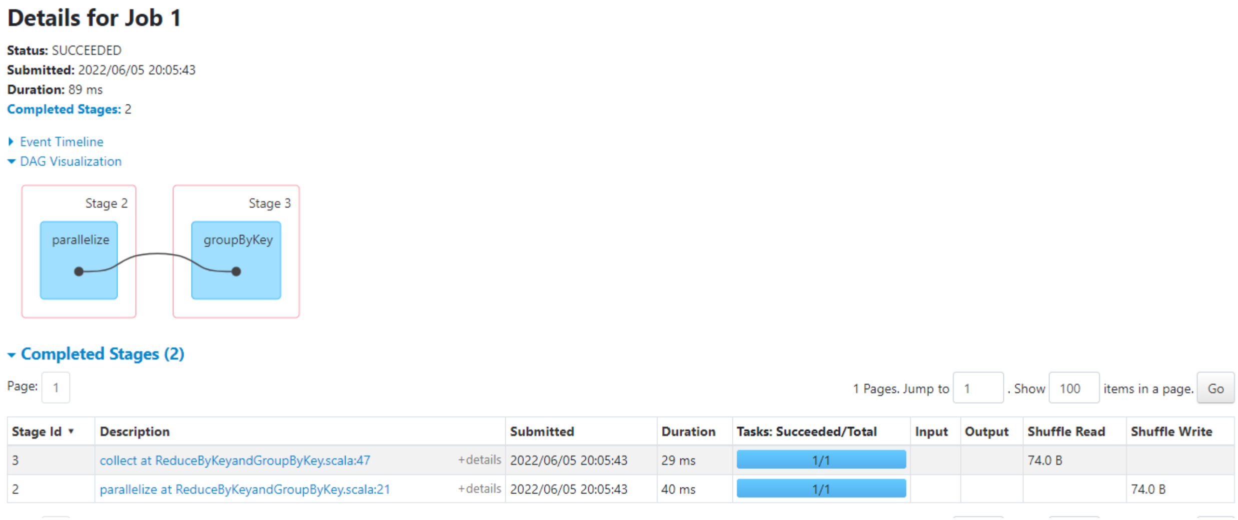
▾ Completed Stages (2)

Page: 1 1 Pages. Jump to 1 . Show 100 items in a page. Go

Stage Id ▾	Description		Submitted	Duration	Tasks: Succeeded/Total	Input	Output	Shuffle Read	Shuffle Write
1	collect at ReduceByKeyandGroupByKey.scala:31	+details	2022/06/05 20:05:42	0.6 s	1/1			57.0 B	
0	parallelize at ReduceByKeyandGroupByKey.scala:21	+details	2022/06/05 20:05:41	0.6 s	1/1				57.0 B

In groupByKey is work on (key,value) pair and the grouping is done by shuffling of data based on key between partitions for the final grouping involves more shuffling of data

So, use When we require final result value in group but not aggregated



Details for Job 1

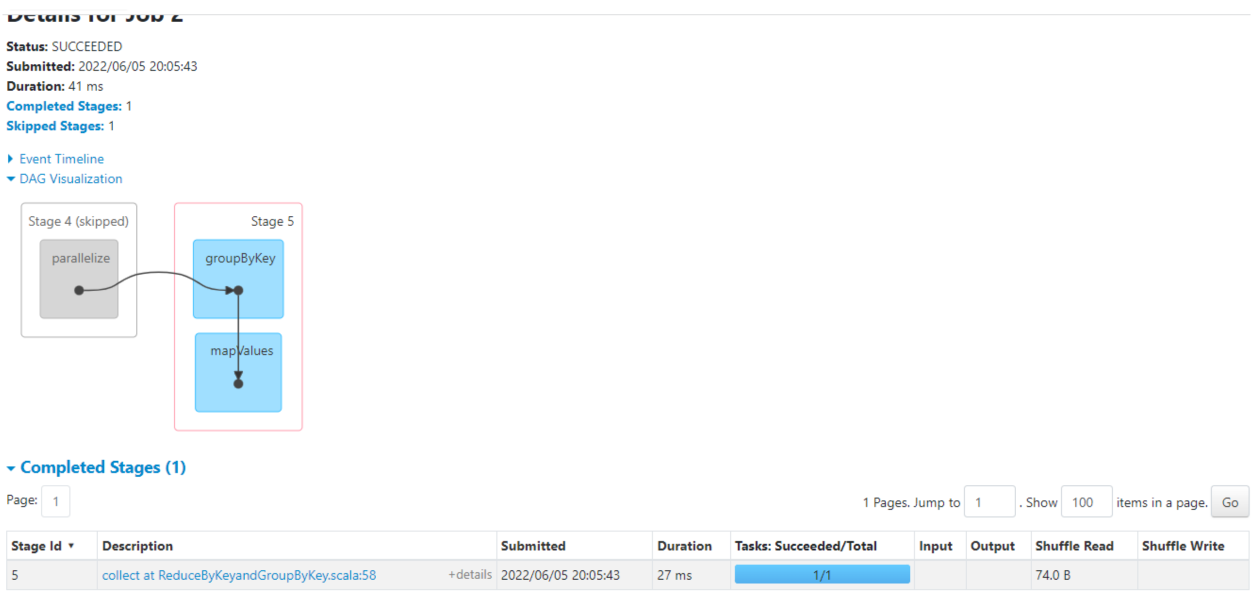
Status: SUCCEEDED
Submitted: 2022/06/05 20:05:43
Duration: 89 ms
Completed Stages: 2

▾ Completed Stages (2)

Page: 1 1 Pages. Jump to 1 . Show 100 items in a page. Go

Stage Id ▾	Description		Submitted	Duration	Tasks: Succeeded/Total	Input	Output	Shuffle Read	Shuffle Write
3	collect at ReduceByKeyandGroupByKey.scala:47	+details	2022/06/05 20:05:43	29 ms	1/1			74.0 B	
2	parallelize at ReduceByKeyandGroupByKey.scala:21	+details	2022/06/05 20:05:43	40 ms	1/1				74.0 B

If we want the final result from the groupByKey transformation then we have to apply one more transformation to aggregate the values which involve one more step of processing.



Details for Job 2

Status: SUCCEEDED
Submitted: 2022/06/05 20:05:43
Duration: 41 ms
Completed Stages: 1
Skipped Stages: 1

▸ Event Timeline
▾ DAG Visualization

▾ **Completed Stages (1)**

Page: 1

1 Pages. Jump to 1 . Show 100 items in a page. Go

Stage Id ▾	Description	Submitted	Duration	Tasks: Succeeded/Total	Input	Output	Shuffle Read	Shuffle Write
5	collect at ReduceByKeyandGroupByKey.scala:58 +details	2022/06/05 20:05:43	27 ms	1/1			74.0 B	