链滴

spark 算子详解 ------spark 算子分类

作者：18582596683
原文链接：https://ld246.com/article/1543743495823
来源网站：链滴

1.spark算子分类

1.1Transformation算子

Transformation算子不触发提交作业,完成作业中间处理过程。Transformation算子又分为如下两类：
1.Value数据类型的Transformation算子：针对处理的数据项是Value型的数据。
2.Key-Value数据类型的Transformation算子：针对处理的数据项是Key-Value型的数据。

1.2Action算子

Action算子会触发 SparkContext 提交 Job 作业。

2.spark算子列表

2.1.Value数据类型的Transformation算子

2.1.1.输入分区与输出分区一对一类型的算子

（1）map算子
（2）flatMap算子
（3）mapPartitions算子
（4）mapPartitionsWithIndex算子
（5）glom算子
（6）randomSplit算子

2.1.2.输入分区与输出分区多对一类型的算子

（1）union算子
（2）cartesian算子

2.1.3.输入分区与输出分区多对多类型的算子

（1）groupBy算子
（2）coalesce算子
（3）repartition算子

2.1.4.输出分区为输入分区子集型的算子

（1）filter算子
（2）distinct算子
（3）intersection算子

（4）subtract算子
（5）sample算子
（6）takeSample算子

2.1.5.Cache型的算子

（1）persist算子
（2）cache算子

2.2.Key-Value数据类型的Transformation算子

2.2.1.输入分区与输出分区一对一类型的算子

（1）mapValues算子
（2）flatMapValues算子
（3）sortByKey算子
（4）sortBy算子
（5）zip算子
（6）zipPartitions算子
（7）zipWithIndex算子
（8）zipWithUniqueId算子

2.2.2.对单个RDD或两个RDD聚集的算子

单个RDD聚集
（1）combineByKey算子
（2）reduceByKey算子
（3）partitionBy算子
（4）groupByKey算子
（5）foldByKey算子
（6）reduceByKeylocally算子

两个RDD聚集
（7）Cogroup算子
（8）subtractByKey算子

2.2.3.连接类型的算子

（1）join算子
（2）leftOutJoin算子

（3）rightOutJoin算子

2.3.Action算子

2.3.1.无输出的算子

（1）foreach算子
（2）foreachPartition算子

2.3.2.输出到HDFS的算子

（1）saveAsTextFile算子
（2）saveAsObjectFile算子
（3）saveAsHadoopFile算子
（4）saveAsSequenceFile算子
（5）saveAsHadoopDataset算子
（6）saveAsNewAPIHadoopFile算子
（7）saveAsNewAPIHadoopDataset算子

2.3.3.输出scala集合和数据类型的算子

（1）first算子
（2）count算子
（3）reduce算子
（4）collect算子
（5）take算子
（6）top算子
（7）takeOrdered算子
（8）aggregate算子
（9）fold算子
（10）lookup算子
（11）countByKey算子