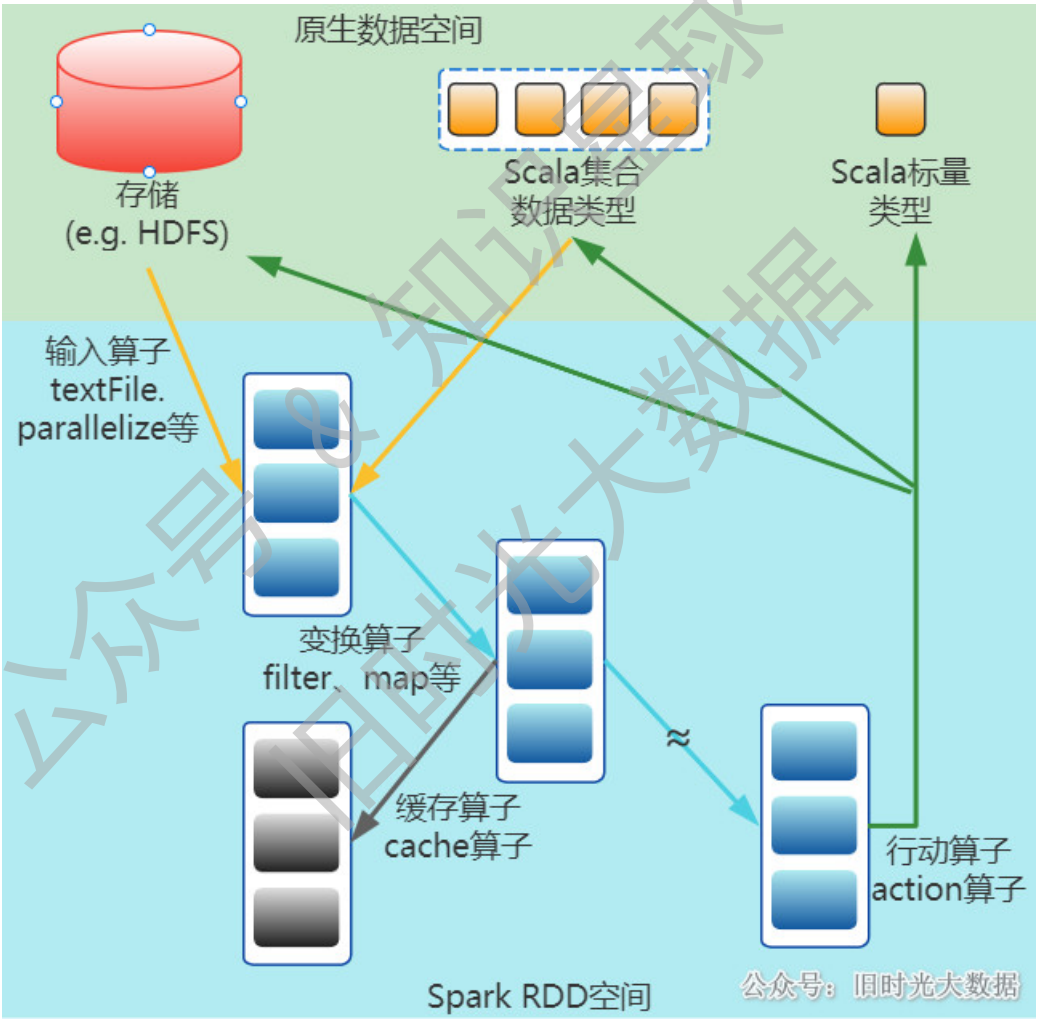
说下Spark中的Transform和Action，为什么Spark要把操作分为Transform和Action？常用的列举一些，说下算子原理

可回答：Spark常见的算子介绍一下

参考答案：

我们先来看下Spark算子的作用：

下图描述了Spark在运行转换中通过算子对RDD进行转换。 算子是RDD中定义的函数，可以对RDD中的数据进行转换和操作。



输入：在Spark程序运行中，数据从外部数据空间（如分布式存储：textFile读取HDFS等，parallelize方法输入Scala集合或数据）输入Spark，数据进入Spark运行时数据空间，转化为Spark中的数据块，通过BlockManager进行管理。

运行：在Spark数据输入形成RDD后便可以通过变换算子，如filter等，对数据进行操作并将RDD转化为新的RDD，通过Action算子，触发Spark提交作业。 如果数据需要复用，可以通过Cache算子，将数据缓存到内存。

输出：程序运行结束数据会输出Spark运行时空间，存储到分布式存储中（如saveAsTextFile输出到HDFS），或Scala数据或集合中（collect输出到Scala集合，count返回Scala int型数据）。

1、Transform和Action

Transformation是得到一个新的RDD，但并不立即执行计算，只是记录下这个操作。方式很多，比如从数据源生成一个新的RDD，从RDD生成一个新的RDD。

Action是指触发对RDD进行计算的操作，得到一个值，或者一个结果（直接将RDD cache到内存中）。

因为所有的Transformation都是采用的懒策略，就是如果只是将Transformation提交是不会执行计算的，计算只有在Action被提交的时候才被触发。这样有利于减少内存消耗，提高了执行效率。

2、算子原理

1）Transformation

map(func)：返回一个新的分布式数据集，由每个原元素经过func函数转换后组成。

filter(func)：返回一个新的数据集，由经过func函数后返回值为true的原元素组成。

flatMap(func)：类似于map，但是每一个输入元素，会被映射为0到多个输出元素（因此，func函数的返回值是一个Seq，而不是单一元素）。

union(otherDataset)：返回一个新的数据集，由原数据集和参数联合而成。

groupByKey([numTasks])：在一个由（K,V）对组成的数据集上调用，返回一个（K，Seq[V])对的数据集。注意：默认情况下，使用8个并行任务进行分组，你可以传入numTask可选参数，根据数据量设置不同数目的Task。

reduceByKey(func, [numTasks])：在一个（K，V)对的数据集上使用，返回一个（K，V）对的数据集，key相同的值，都被使用指定的reduce函数聚合到一起。和groupbykey类似，任务的个数是可以通过第二个可选参数来配置的。

join(otherDataset, [numTasks])：在类型为（K,V)和（K,W)类型的数据集上调用，返回一个（K,(V,W))对，每个key中的所有元素都在一起的数据集。

2）Action

reduce(func)：通过函数func聚集数据集中的所有元素。Func函数接受2个参数，返回一个值。这个函数必须是关联性的，确保可以被正确的并发执行。

collect()：在Driver的程序中，以数组的形式，返回数据集的所有元素。这通常会在使用filter或者其它操作后，返回一个足够小的数据子集再使用，直接将整个RDD集Collect返回，很可能会让Driver程序OOM。

count()：返回数据集的元素个数。

foreach(func): 在数据集的每一个元素上，运行函数func。这通常用于更新一个累加器变量，或者和外部存储系统做交互。

欢迎加入知识星球，获取《大数据面试题 V4.0》以及更多大数据开发学习资料

暮然 送你一张星球优惠券

「旧时光大数据」

立减

¥ 40

新人立减券

2023/12/31 12:00 后失效

知识星球

长按扫码领取优惠

公众号 & 知识星球 旧时光大数据