天善智能
TIANSHAN SOFT

专注于商业智能BI和大数据的垂直社区平台

Spark初级课程

讲师：平常心

transformations和actions

RDD支持两种操作：

1.transformations: 通常已有的RDD创建一个新的RDD

2.actions:对RDD进行计算并将结果返回给驱动程序

说明：

1.transformations的特点是lazy,即如果一个spark应用程序中只有transformations操作，则提交该程序也不会触发执行，只是通过DAG记录了对RDD所做的操作，只有当一个action操作的时候触发执行。这种lazy特性可以进行底层优化，避免产生过多的中间结果。

2.actions的特点是会触发前面的transformations的执行，运行spark应用程序。默认情况下，actions触发的transformations都重新计算RDD。

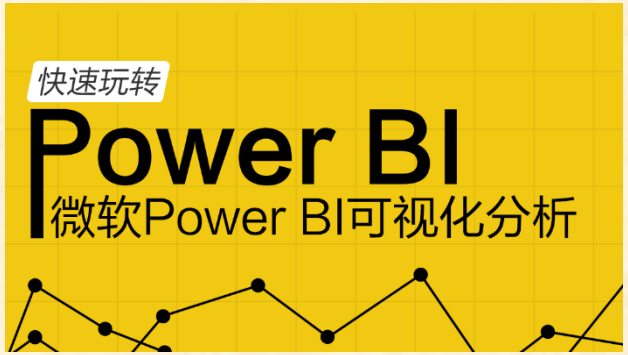
常见的transformations

名称	说明
map[U](f: (T) => U)	将传入的元素T转化为元素U 通常只改变元素类型，不改变元素数量
flatMap[U](f: (T) => TraversableOnce[U])	与map类似，将传入的元素T转化为Seq(U) 通常改变元素的数量
filter(f: (T) => Boolean): RDD[T]	对传入的元素T进行判断，如果是true则保留 通常不改变元素类型，减少元素数量
groupByKey()	对(key，value)按照key进行分组，转化为(key,Iterable<value>)
reduceByKey()	使用reduce函数合并每个key的value
sortByKey()	对key进行排序操作
join	对两个RDD的元素按照key进行连接，返回 RDD[(K, (V, W))]
cogroup	和join类似，返回RDD[(K, (Iterable[V], Iterable[W]))]

常见的actions

名称	说明
reduce(f: (T, T) => T): T	对RDD的元素进行聚合，如先是(t1+t2) => t,接下来是(t+t3) 以此类推 通常不改变元素类型，减少元素数量
collect()	返回RDD的所有元素到程序终端
take(num: Int)	获取RDD中的前num个元素到终端
first()	类似take(1),获取RDD中的第一个元素
count()	获取RDD中元素的总数
saveAsTextFile()	将RDD保存为文本文件，使用RDD中元素的toString方法按照行存储
countByKey()	对每个key进行计数，返回(key,int)对的哈希表
foreach(f: (T) => Unit)	遍历RDD中的所有元素

更多商业智能BI和大数据精品视频尽在 www.hellobi.com



BI、商业智能
数据挖掘 大数据
数据分析
R Python
机器学习
Tableau
QLIKVIEW
Hive Hadoop
BIWORK
BAO胖子 seng
贝克汉姆
曹浩

天善智能
TIANSHAN SOFT