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

RDD持久化

Spark重要的功能特性之一是RDD持久化存储到内存，即参与计算的节点将自己操作RDD的partition数据持久化到内存，使之后对该RDD的反复计算之间使用内存存储的partiton。优点就是针对一个RDD反复执行多个计算的场景，只要对RDD进行一次计算，而避免反复计算该RDD，提高程序性能。

说明：

1.调用cache()或persist()实现，其实cache()底层就是调用persist(MEMORY_ONLY),将数据持久化到内存，如果需要从内存释放，调用unpersist()方法。

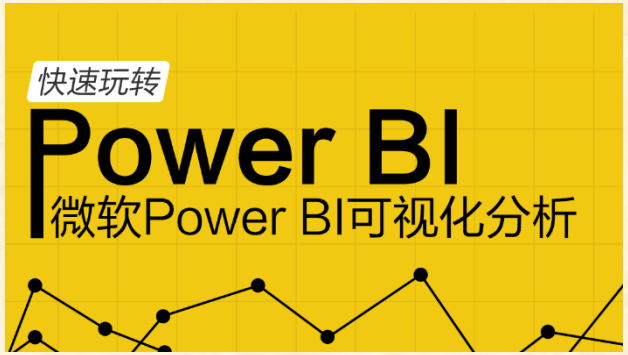
2.持久化操作是容错的，如果RDD的任何分区丢失，通过最初最初创建的transformation自动重新生成。

3.可以使用不同的存储级别存储。

存储级别

存储级别	占用空间	CPU使用	内存存储	磁盘存储	说明
MEMORY_ONLY	高	低	是	否	默认选项，RDD的（分区）数据直接以Java对象的形式存储于JVM的内存中，如果内存空间不足，某些分区的数据将不会被缓存，需要在使用的时候根据血缘重新计算。
MEMORY_AND_DISK	高	中	部分	部分	RDD的数据直接以Java对象的形式存储于JVM的内存中，如果内存空间不中，某些分区的数据会被存储至磁盘，使用的时候从磁盘读取。
MEMORY_ONLY_SER (Java and Scala)	低	高	是	否	RDD的数据（Java对象）序列化之后存储于JVM的内存中（一个分区的数据为内存中的一个字节数组），相比于MEMORY_ONLY能够有效节约内存空间（特别是使用一个快速序列化工具的情况下），但读取数据时需要更多的CPU开销；如果内存空间不足，处理方式与MEMORY_ONLY相同。
MEMORY_AND_DISK_SER (Java and Scala)	低	高	部分	部分	相比于MEMORY_ONLY_SER，在内存空间不足的情况下，将序列化之后的数据存储于磁盘。
DISK_ONLY	低	高	否	是	仅仅使用磁盘存储RDD的数据（未经序列化）。
MEMORY_ONLY_2, MEMORY_AND_DISK_2, etc.					以MEMORY_ONLY_2为例，MEMORY_ONLY_2相比于MEMORY_ONLY存储数据的方式是相同的，不同的是会将数据备份到集群中两个不同的节点，其余情况类似。
OFF_HEAP (experimental)					RDD的数据序例化之后存储至Tachyon。相比于MEMORY_ONLY_SER，OFF_HEAP能够减少垃圾回收开销、使得Spark Executor更“小”更“轻”的同时可以共享内存；而且数据存储于Tachyon中，Spark集群节点故障并不会造成数据丢失，因此这种方式在“大”内存或多并发应用的场景下是很有吸引力的。需要注意的是，Tachyon并不直接包含于Spark的体系之内，需要选择合适的版本进行部署；它的数据是以“块”为单位进行管理的，这些块可以根据一定的算法被丢弃，且不会被重建。

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