Spark函数讲解：cache

使用MEMORY_ONLY储存级别对RDD进行缓存，其内部实现是调用persist()函数的。官方文档定义：

> Persist this RDD with the default storage level (`MEMORY_ONLY`).

函数原型

```
def cache() : this.type
```

实例

```
/**
 * User: 过往记忆
 * Date: 15-03-04
 * Time: 下午06:30
 * bolg:
 * 本文地址：/archives/1274
 * 过往记忆博客，专注于hadoop、hive、spark、shark、flume的技术博客，大量的干货
 * 过往记忆博客微信公共帐号：iteblog_hadoop
 */
scala> var data = sc.parallelize(List(1,2,3,4))
data: org.apache.spark.rdd.RDD[Int] =
      ParallelCollectionRDD[44] at parallelize at <console>:12

scala> data.getStorageLevel
res65: org.apache.spark.storage.StorageLevel =
      StorageLevel(false, false, false, false, 1)

scala> data.cache
res66: org.apache.spark.rdd.RDD[Int] =
      ParallelCollectionRDD[44] at parallelize at <console>:12

scala> data.getStorageLevel
res67: org.apache.spark.storage.StorageLevel =
      StorageLevel(false, true, false, true, 1)
```

我们先是定义了一个RDD，然后通过getStorageLevel函数得到该RDD的默认存储级别，这里是NONE。然后我们调用cache函数，将RDD的存储级别改成了MEMORY_ONLY(看StorageLevel的第二个参数)。关于StorageLevel的其他的几种存储级别介绍请参照StorageLevel类进行了解，这里就不介绍了。

本博客文章除特别声明，全部都是原创！
原创文章版权归过往记忆大数据（过往记忆）所有，未经许可不得转载。
本文链接: 【】（）