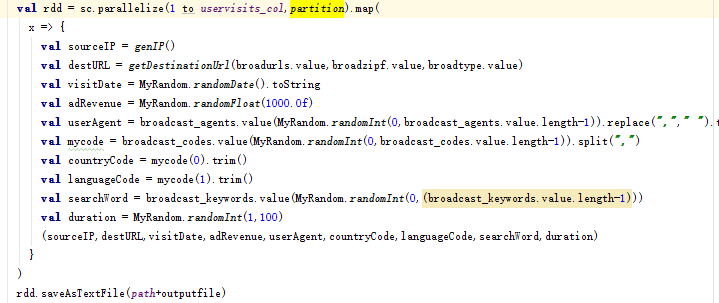
存储part数，Block 和 task的关系

在生成数据时，我们可以指定有多少个并行的线程（partition）共同生成，：



这样的话，在HDFS中会生成partition个part（part-00000，part-00001…）

此时如果不显示的指明partition数目，这个数会和在spark-submit时指定的total-executor-cores有关，比如说，

如果指定了--total-executor-cores 36, 那么一般情况下会分成36份，有些情况下会不到36份（比如有时是30个，推断是当时只有30个空闲的core分配？）

现在我们将block size 的大小设为128M，一个part的大小为a M

a <= 128M时，这一个part就只有一个block

a > 128M时，这一个part的block数量为a/128取上整

所以如果一个文件分成几个part存和一个整块的存block的数量会存在区别

（比如说9792M的文件，分一个存有77block,但是分成36份存就会出现108个block）

在实际的运行中，我们发现block的数据和task的数量是对应的（？），底层有多少个block就会对应多少个task 所以文件的不同存法会导致后续的task数量不同

这是否也是在测试中需要考虑的一个点？