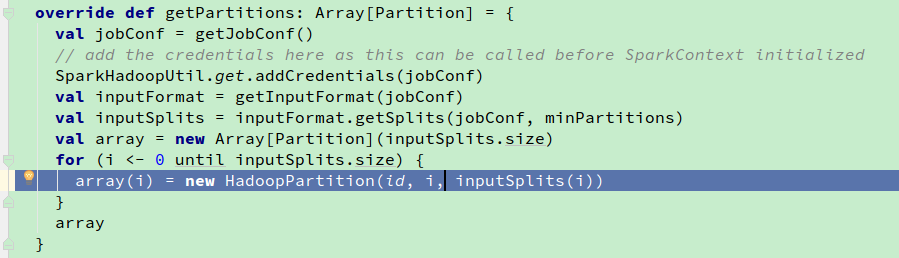
Spark HDFS Splits and Partition

1:Spark中的Partition的概念与Mapreduce中的Splits的概念是类似的

都是计算层的逻辑概念，而Block是存储层面的概念。

首先明确：多少个split 其实就对应了多少个Partition

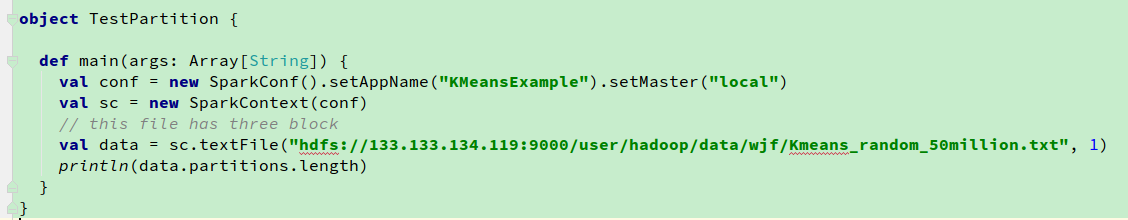


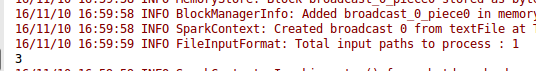
**sc.textFile(path,minPartitions)**

注意对minPartitions的理解，partition个数不可能小于

FileInputFormat.getSplits() 得到的splits个数

实验验证：

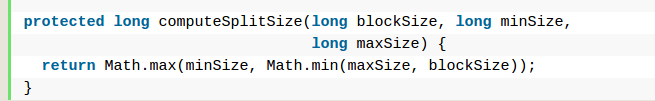




2:如果Partition 个数大于Block 个数，那么能否保证顺序，切分Block 还是如何处理？

多个输入文件，按照文件进行处理，不够一个Split 算一个Split

Split size 计算如下，



实验验证：

Situation1: with minPartition = 2

顺序是一致的

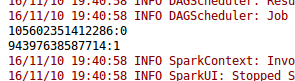


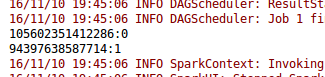




Situation 2: with minPartition = 1

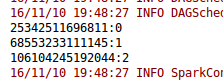
与sit1比较虽说都是两个Partition ,但是顺序并不一样。

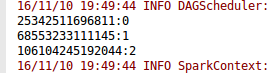


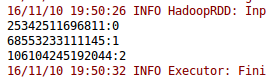


Situation 3:with minPartition = 3

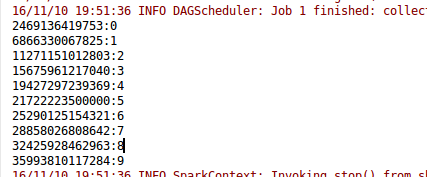
结果一致

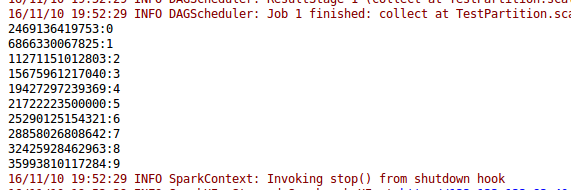


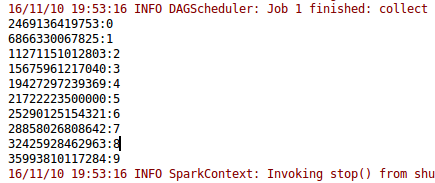




Situation 3: with minPartition = 10







结论：使用textFile读取 HDFS 文件：

* 只要设置了相同的minPartition,多次运行各个Partition结果是一致的。
* 对于minPartition 小于真实的Splits个数的情况，设置和不设置结果不一致。