大数据开发技术

东北林业大学

卢洋

第三章

MapReduce框架原理

3.2

InputFormat数据输入

FileInputFormat切片过程

1. 先找到数据存储的路径
2. 开始遍历处理(规划切片)目录下的每一个文件
3. 遍历第一个文件: big.dat
   (a) 获取文件大小fs.sizeOf(big.dat)
   (b) 计算切片大小
   computeSplitSize
   (Math.max(minSize, Math.max(maxSize, blockSize)))
   = blockSize = 128m

基本上是blocksize

   (c) 默认情况下，切片大小 = BlockSize

如果258m,那么一个128，一个130

   (d) 开始切分，形成第一个切片: big.dat-0:128m，第二个切片big.dat-128m:256m，第三个切片big.dat-256m:300m(每次切片时，都要判断切完剩下的部分是否大于切片大小的1.1倍，如果不大于1.1倍就不进行切分)
   (e) 将切片信息写到切片规划文件中
   (f) 整个切片文件的核心过程在getSplits()方法中完成
   (g) InputSplit只记录了切片的元数据信息，比如起始位置、长度及所在的节点列表等
4. 提交切片规划文件到Yarn上，Yarn上的MRAppMaster就可以根据切片规划文件计算开启的MapTask的数量。