

在这个优化的版本中，我们使用SparkContext newAPIHadoopFile完成数据输入，需要着重说明一下“org.apache.hadoop.mapreduce.lib.input.CombineTextInputFormat”，这个类可以将多个小文件合并生成一个Split，而一个Split会被一个Task处理，从而减少Task的数目。这个应用的执行过程中，会产生两个Job，其中Job0包含一个Stage，一个Task；Job1包含两个Stage，每个Stage包含一个Task，也就是说，Task的总数目为3，如下图所示：