Hadoop的mapper和reducer的个数如何确定？reducer的个数依据是什么？

问过的一些公司：创略科技

参考答案：

map数量

影响map个数（split个数）的主要因素有：

文件的大小。当块（dfs.block.size）为128m时，如果输入文件为128m，会被划分为1个split；当块为256m，会被划分为2个split。

文件的个数。FileInputFormat按照文件分割split，并且只会分割大文件，即那些大小超过HDFS块的大小的文件。如果HDFS中dfs.block.size设置为128m，而输入的目录中文件有100个，则划分后的split个数至少为100个。

splitSize的大小。分片是按照splitszie的大小进行分割的，一个split的大小在没有设置的情况下，默认等于hdfs block的大小。

```
splitSize=max{minSize,min{maxSize,blockSize}}
```

map数量由处理的数据分成的block数量决定default_num = total_size / split_size

reduce数量

reduce的数量job.setNumReduceTasks(x); x为reduce的数量。不设置的话默认为1