mapreduce总结

# 创建工程，导入mapreduce需要的jar包

Pom文件中引入：

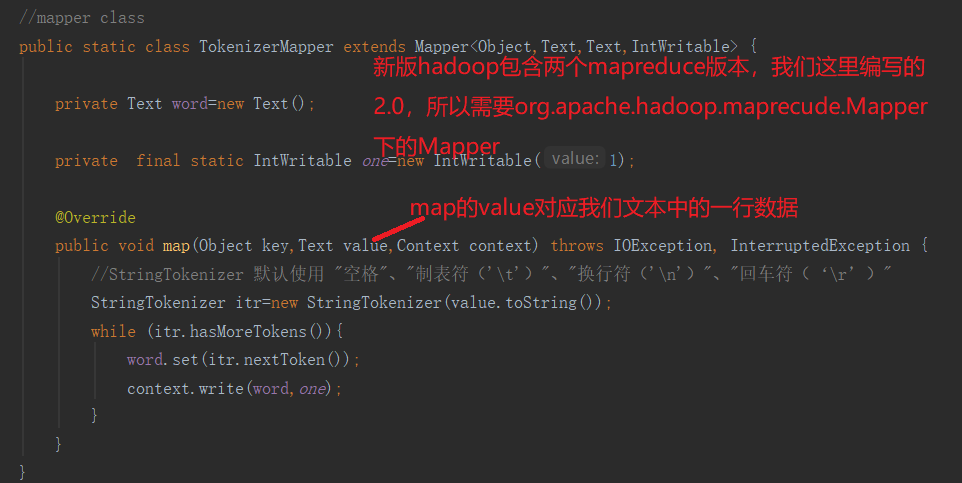
<dependency>  
 <groupId>org.apache.hadoop</groupId>  
 <artifactId>hadoop-common</artifactId>  
 <version>2.7.4</version>  
 <scope>provided</scope>  
</dependency>  
<dependency>  
 <groupId>org.apache.hadoop</groupId>  
 <artifactId>hadoop-mapreduce-client-core</artifactId>  
 <version>2.7.4</version>  
</dependency>



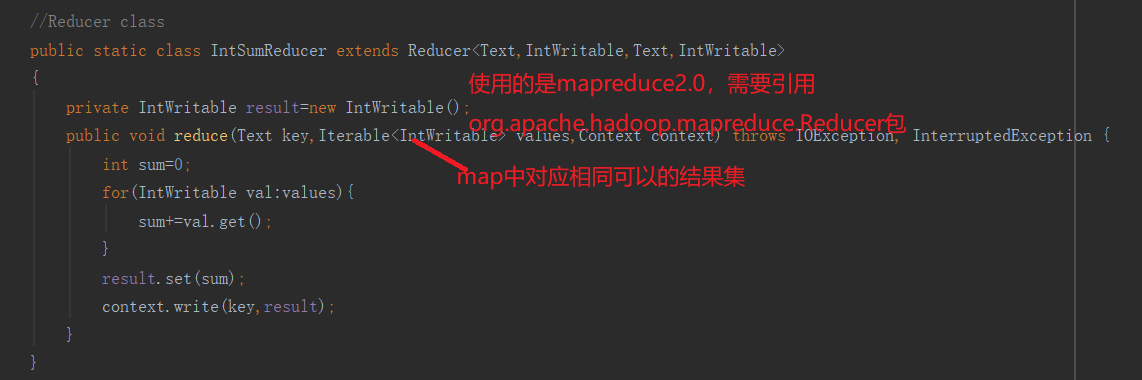
# 编写map方法

新建一个类，类名随便起，我们这里叫WordCount

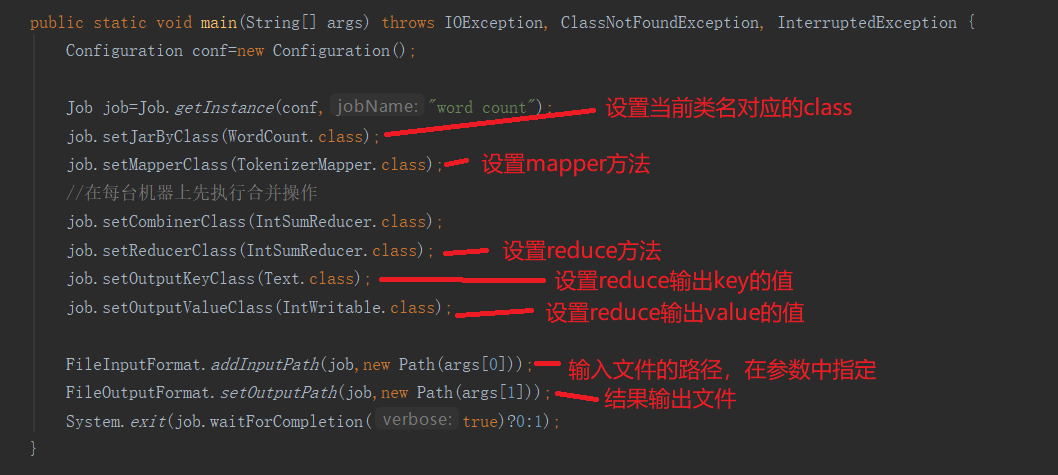




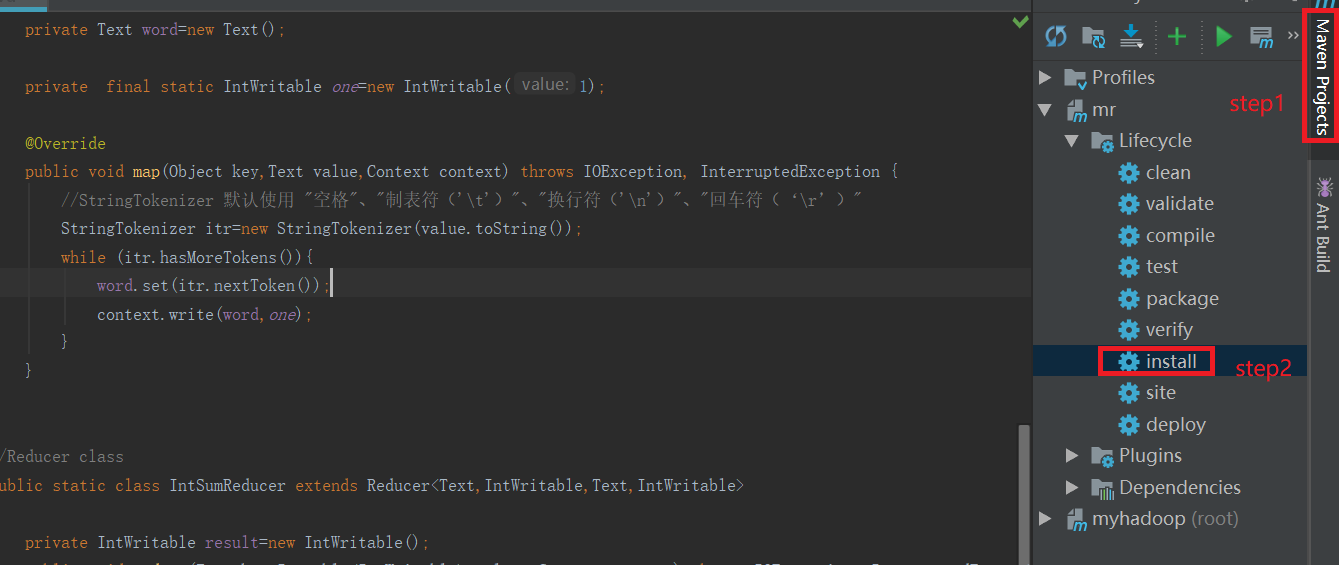
# 编写reduce方法



# 编写main方法



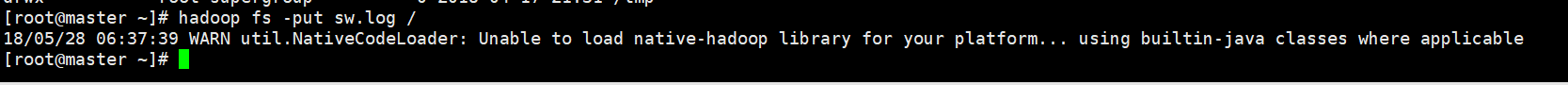
# 打包mapreduce程序



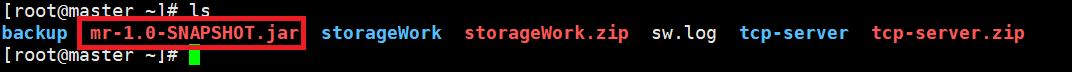
# 运行mapreduce

## 我们先将需要进行单词统计的数据上传到hdfs上





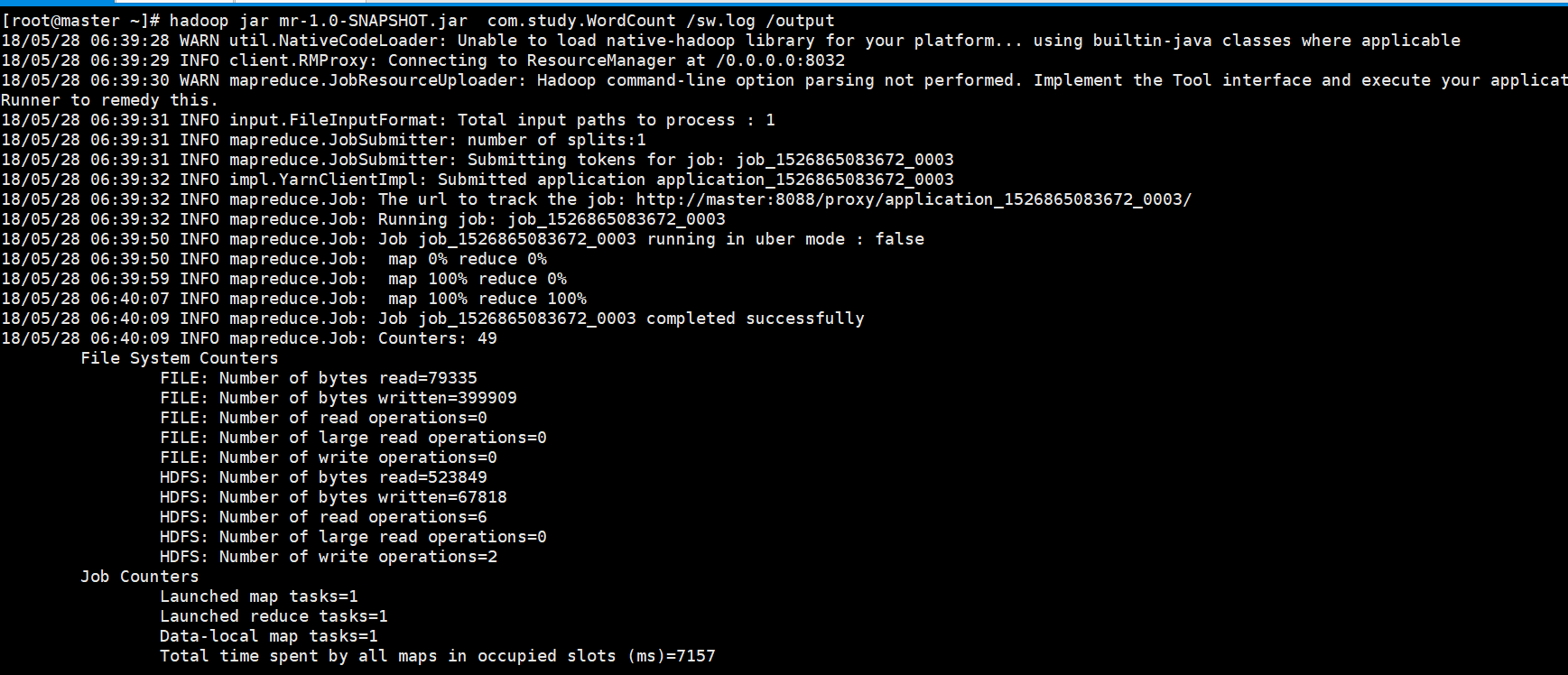
## 将我们打包好的程序上传到linux机器上



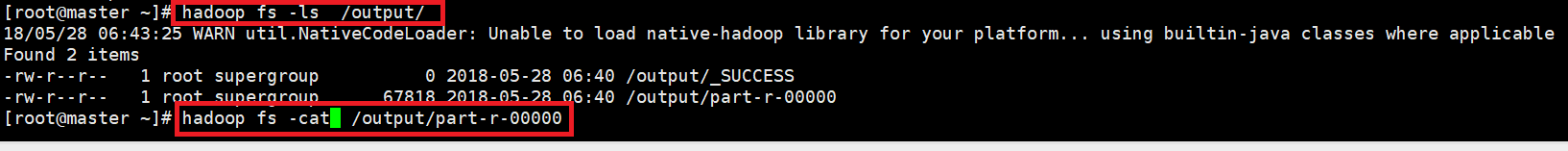
## 运行程序



Mapreduce执行过程：



## 查看执行结果



# 扩展mapreduce程序

设置扩展命令行选项“-skip 含有正则表达式的文件路径“

在主函数中解析-skip 后将文件路径存入Cachefile中，在map类的setup函数中从cachefile中取出正则数据文件路径，读取文件，解析正则表达式内容，缓存到程序中

# 运行程序



# Mapreduce版本选择

版本选择新版hadoop版本中存在两套mapreduce的api 旧版和新版；旧版在org.apache.hadoop.mapred包下；

所以在编写mapreduce程序的时候注意引入的程序包

# Mapreduce总结

* MapReduce作业通常将输入数据集拆分成独立的块
* 它们以完全并行的方式由MAP任务处理。
* (可选) combiner对map结果在本地执行处理，通常和reduce中的处理一致；尤其在统计数据过程中对本地相同key的map结果进行归并可以减少map的个数 减少网络传输
* 该框架对映射的输出进行排序，然后将这些输出输入到还原任务中。通常，作业的输入和输出都存储在文件系统中。该框架负责调度任务，监视它们并重新执行失败的任务。