**Spark的学习笔记（2）——部署udf函数**

————20191120：韵方

在使用pyspark的udf时由于采用的是python语言，编译引擎为pthon，而spark的架构语言是scala，使用jvm进行编译，所以有时使用python构造的udf函数会造成资源效率的降低，所以本章进行自定义udf函数的说明。

由于scalal的学习成本较高，这里我用java举例

**前提条件：**

成功完成spark学习笔记中（1）的内容，并且使用的jdk是8（不是8时maven编译jar时会报错）

**编写代码：**

在（1）的基础上，将scala，root文件夹的权限取消附加在java上，创建一个包，创建一个UDF函数，名字随意哈：

****

在UDF\_ceshi中编写如下代码，其中包名和类名自己随意，在这段代码中·我们继承了一个名为UDF的接口，并重构了它的evaluate方法（此方法就是你的自定义udf函数），需要注意的是，此方法支持多态的写法，也就是说，你定义的udf函数可以有0到多个入参，返回值也可不同（考虑到多数人的java基础不是很好，解释一下，并在下方截图举个例子）：

**package** com.bzf.csdemo;，  
  
**import** org.apache.hadoop.hive.ql.exec.UDF;  
  
**public class** UDF\_ceshi **extends** UDF {  
 **public** String evaluate(String value) {  
 **return** value+**"demo"**;  
 }  
  
 **public static void** main(String[] args) {  
 **return**;  
 }  
}



**通过maven打包jar：**

点击**view-tool Windows-maven**你会得到如下的界面:

****

依次点击clean，compile，pakage，然后你会在右边看到一个target，红色框框就是你要的jar啦：



然后就将这个jar打包去你的spark集群上就好了，后续我要写的相关操作就是我现在公司的事情了，不讲述了，你们自行研究哈。